BIG DATA ANALYTICS

LAB ASSIGNMENT 1

MAHESH PACHARE

FINAL YEAR B.TECH IT

191080054

Aim:

Compare different versions of Hadoop(Hadoop 1.x, Hadoop 2.x, and Hadoop 3. x). Also setup Hadoop 1.x single node cluster.

Theory:

Hadoop is an open-source programming framework that is used to store and process a large amount of data in a distributed computing environment. Hadoop has emerged as a premier choice for Big Data processing tasks.
Hadoop 1.x is built on two whitepapers published by Google, i.e,

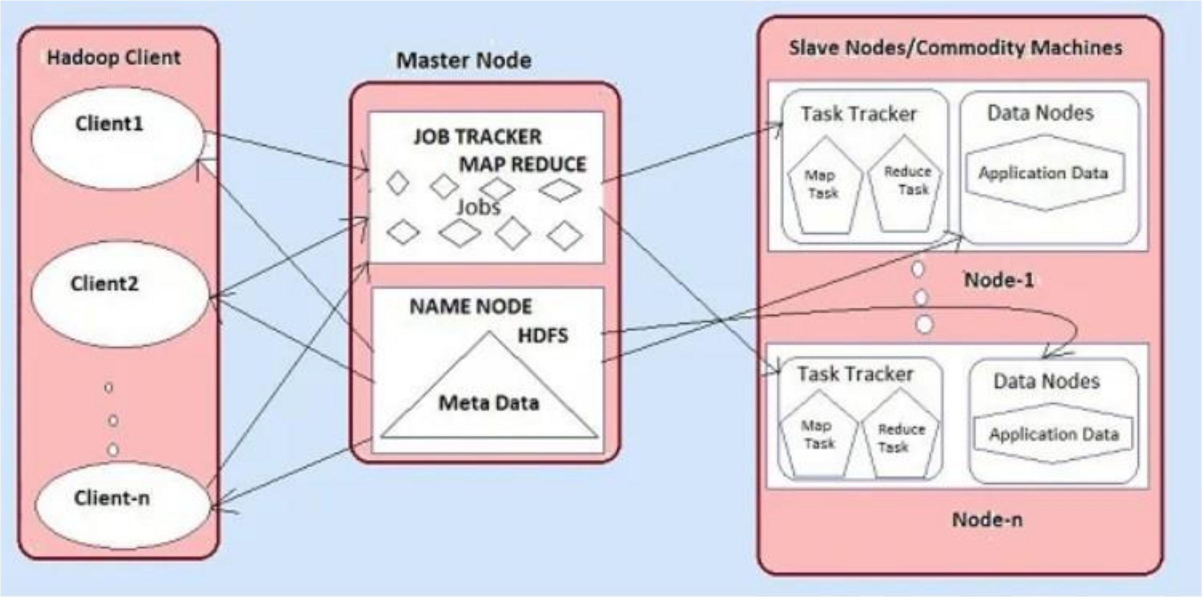
- HDFS
- Map Reduce

HDFS: Hadoop Distributed File System
It is different from the normal file system in a way that the data copied on to HDFS is split into 'n' blocks and each block is copied on to a different node in the cluster. To achieve this we use master-slave architecture

- HDFS Master => Name Node: Takes the client request and responsible for orchestrating the data copy across the cluster
- HDFS Slave => Data Node: Actually saves the block of data and coordinates with its master

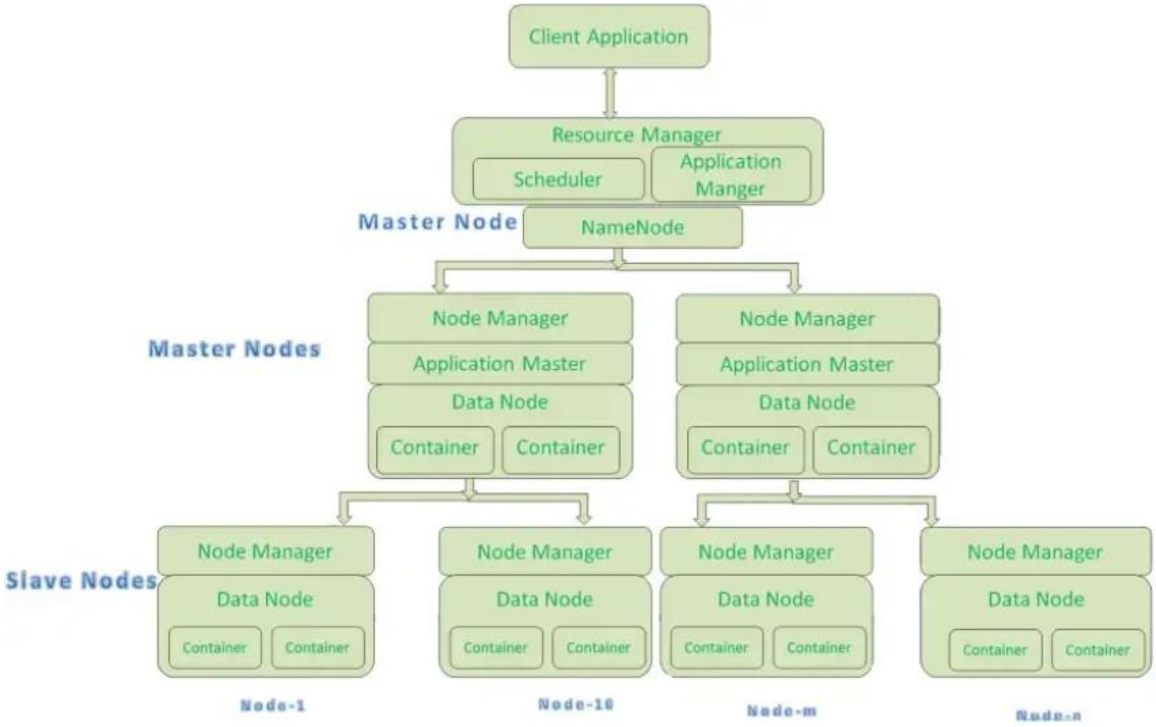
MapReduce: This is the processing engine and is also implemented in master-slave architecture.

- MR Master => Job Tracker: Takes the incoming jobs, identifies the available resources across the cluster, divides the job into tasks and submits it to the cluster
- MR Slave => Task Tracker: Actually runs the task and coordinates with its master.

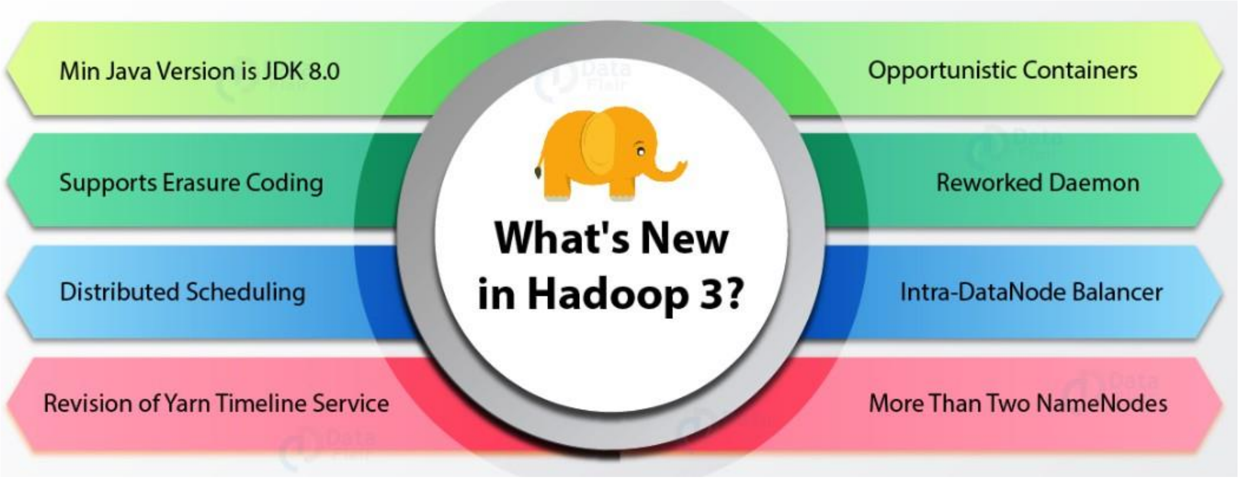


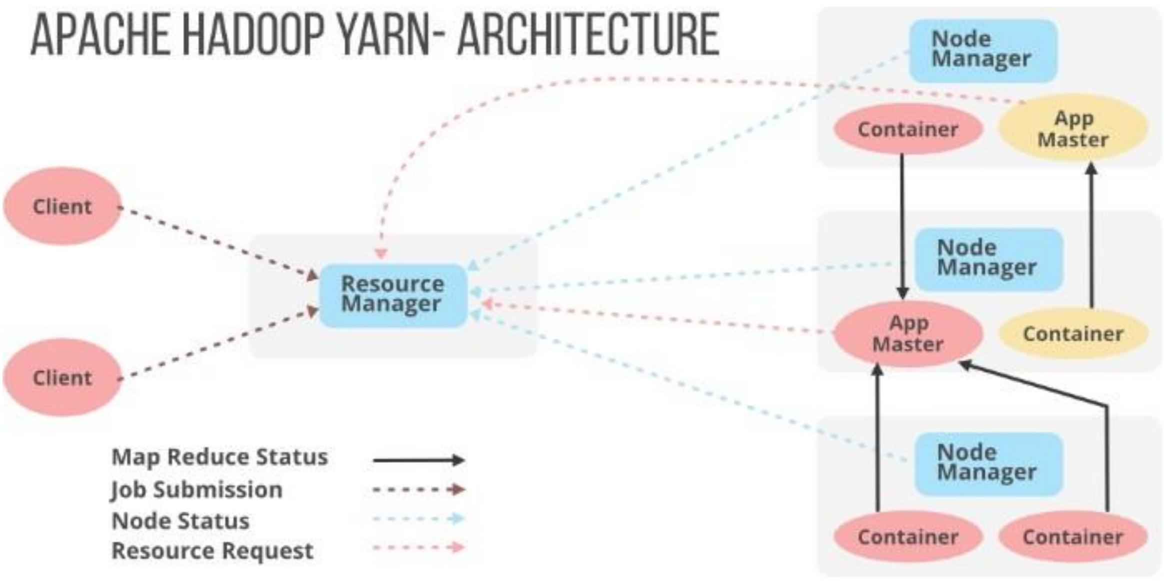
Hadoop 1.x Architecture

Hadoop 2.0 broadly consists of two components: Hadoop **Distributed File System(HDFS)** which can be used to store large volumes of data and **Yet Another Resource Negotiator(YARN)** which provides resource management and scheduling for running jobs. YARN supports various processing frameworks such as MapReduce, Spark, Storm etc.



Hadoop 2.x In-Detail Architecture





Hadoop 1.x	Hadoop 2.x
Supports **MapReduce (MR)** processing model only. Does not support non-MR tools	Allows to work in MR as well as other distributed computing models like **Spark, Hama, Giraph, Message Passing Interface) MPI & HBase coprocessors**.
MR does both processing and **cluster-resource management**.	**YARN** (Yet Another Resource Negotiator) does cluster resource management and processing is done using different processing models.
Has **limited scaling** of nodes. Limited to 4000 nodes per cluster	Has **better scalability**. Scalable up to 10000 nodes per cluster
Works on **concepts of slots** – slots can run either a Map task or a Reduce task only.	Works on **concepts of containers**. Using containers can run generic tasks.
A **single Namenode** to manage the entire	**Multiple Namenode** servers manage

namespace.	multiple namespaces.
Has **Single-Point-of-Failure (SPOF)** – because of single Namenode- and in the case of Namenode failure, needs manual intervention to overcome.	Has the feature to overcome SPOF with a standby Namenode and in the case of Namenode failure, it is configured for **automatic recovery**.
MR API is compatible with Hadoop 1.x. A program written in Hadoop1 executes in Hadoop 1.x without any additional files.	MR API **requires additional files** for a program written in Hadoop 1.x to execute in Hadoop 2.x.
Has a **limitation** to serve as a platform for event processing, streaming and real-time operations.	Can **serve** as a platform for a wide variety of data analytics-possible to run event processing, streaming and real-time operations.
A **Namenode failure** affects the **stack**.	The **Hadoop stack – Hive, Pig, HBase** etc.are all equipped to **handle Namenode failure**.
Does **not support Microsoft Windows**	Added **support for Microsoft windows**

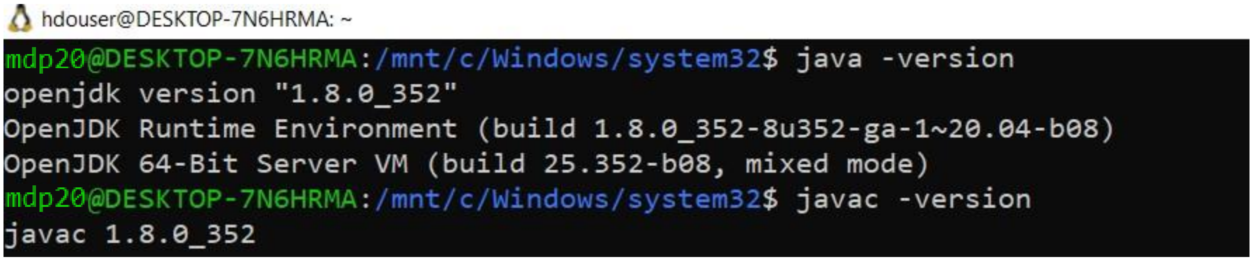
Hadoop 2.x	Hadoop 3.x
Java version 7 is the minimum requirement.	**Java version 8** is the minimum requirement. As most of the dependency library files used are from java8.
HDFS supports **replication for fault tolerance**.	HDFS support for **erasure encoding**. (Erasure coding is a technique for durably storing

	information with significant space savings compared to replication)
For data, balancing uses **HDFS balancer**.	For data, balancing uses Intra-data node balancer, which is invoked via the **HDFS Disk balancer CLI**.
YARN timeline service Introduced. Uses an old timeline service which has scalability issues.	**YARN timeline service v.2**(improved scalability and reliability). Improve the timeline service v2 and improve the scalability and reliability of timeline service.
Due to data **Node caching** we can fast access the data.	Here also through **Datanode caching** we can fast access the data.
Limited Shell scripts with **Bugs**.	Many new Unix shell API, along with **old Bug Fixed**.
Map reduce became fast due to YARN.	Map reduce became faster, particularly at **map output collector** and shuffle jobs by 30%.
Secondary namenode was introduced as standby.	Supports **more than 2 namenode**
Default ports were Conflicting in Linux port range. Which leads to failure in port reservation.	**Port range** has been optimized.
Hadoop **did not support Microsoft filesystem**.	Hadoop now **supports integration with Microsoft Azure Data Lake** as an alternative to Hadoop-compatible filesystem
A **single DataNode** manages multiple disks. Disks inside can lead to significant	New functionality **intra-DataNode balancing** is added, which is invoked via the HDFS disk

skew within a DataNode.	balancer CLI.
The host needs to set the **Heap Size** for JAVA and Hadoop task.	New methods for configuring daemon heap sizes. Notably, auto-tuning is now possible based on the memory size of the host, and the **HADOOP_HEAPSIZE** variable has been deprecated

Setting up hadoop 3.x single node cluster on wsl:

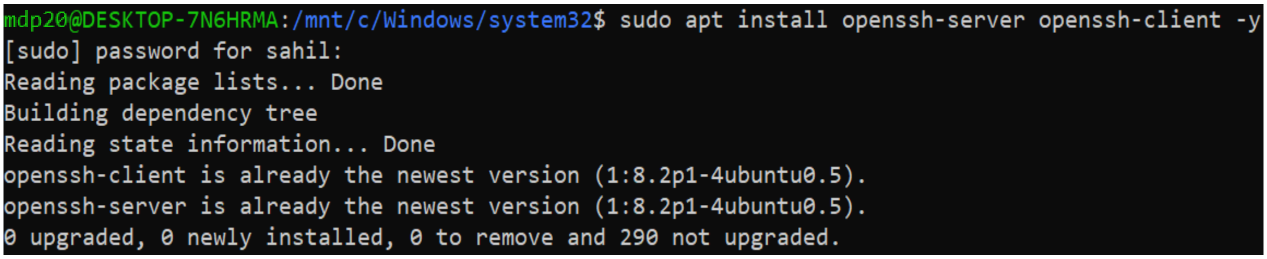
1) Installing java-jdk version 8.



hdouser@DESKTOP-7N6HRMA: ~

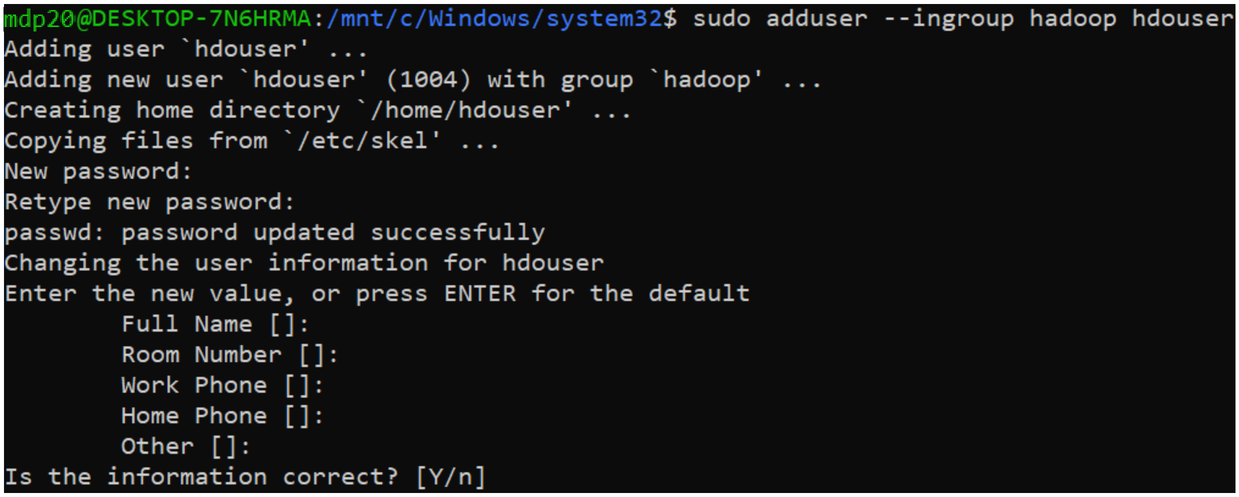
```
mdp20@DESKTOP-7N6HRMA:/mnt/c/Windows/system32$ java -version
openjdk version "1.8.0_352"
OpenJDK Runtime Environment (build 1.8.0_352-8u352-ga-1~20.04-b08)
OpenJDK 64-Bit Server VM (build 25.352-b08, mixed mode)
mdp20@DESKTOP-7N6HRMA:/mnt/c/Windows/system32$ javac -version
javac 1.8.0_352
```

2) Installing openssh client and server.



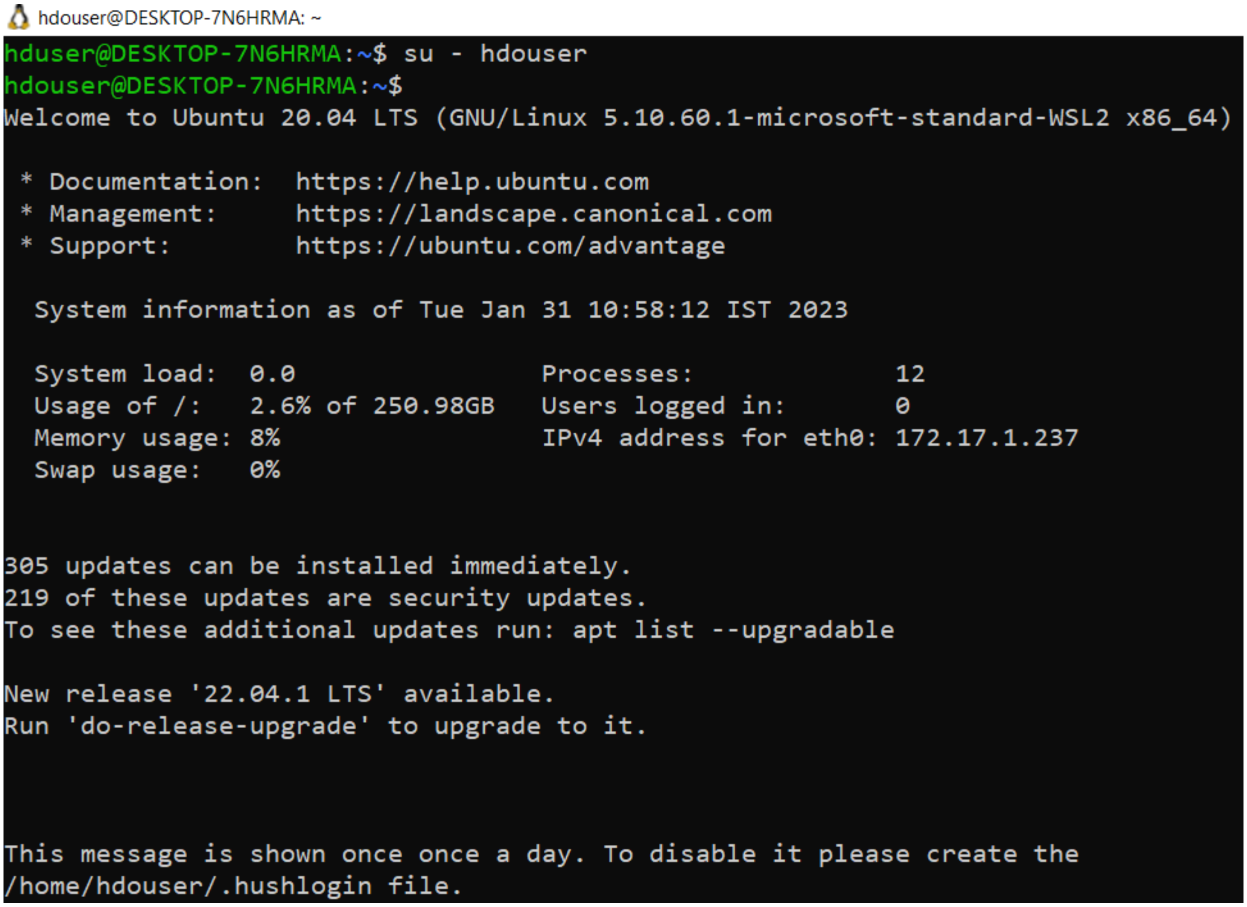
```
mdp20@DESKTOP-7N6HRMA:/mnt/c/Windows/system32$ sudo apt install openssh-server openssh-client -y
[sudo] password for sahil:
Reading package lists... Done
Building dependency tree
Reading state information... Done
openssh-client is already the newest version (1:8.2p1-4ubuntu0.5).
openssh-server is already the newest version (1:8.2p1-4ubuntu0.5).
0 upgraded, 0 newly installed, 0 to remove and 290 not upgraded.
```

3) Create a new group and new user.



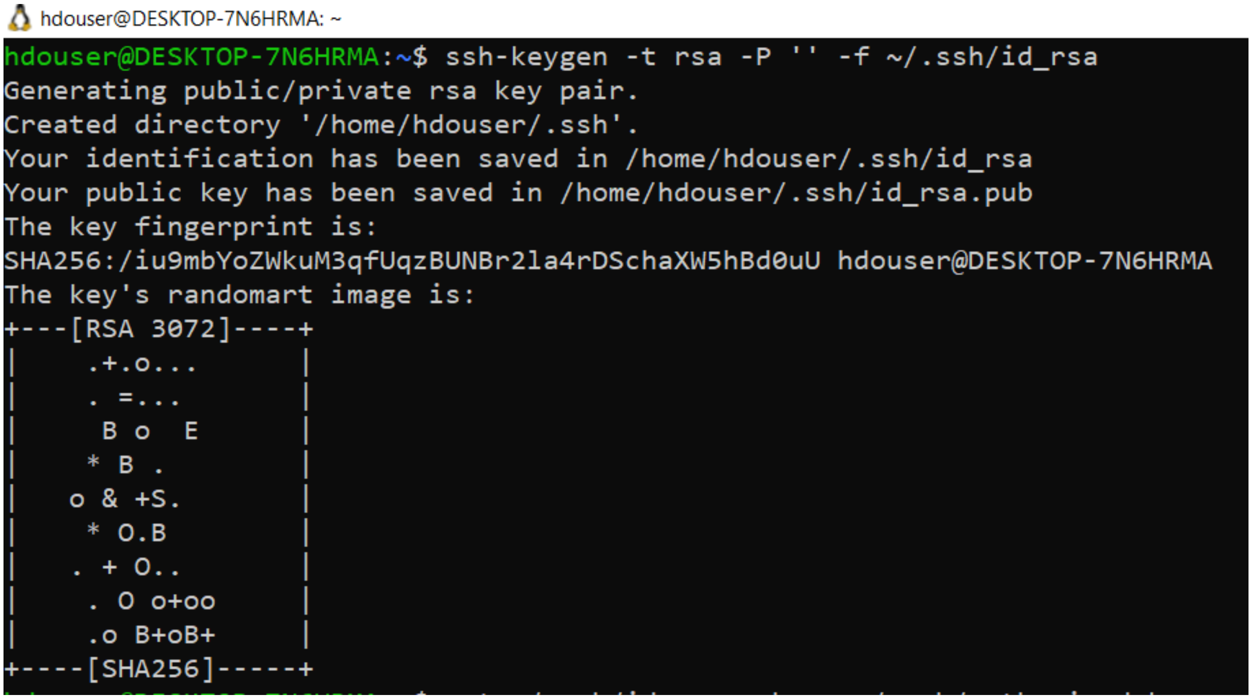
```
mdp20@DESKTOP-7N6HRMA:/mnt/c/Windows/system32$ sudo adduser --ingroup hadoop hdouser
Adding user `hdouser' ...
Adding new user `hdouser' (1004) with group `hadoop' ...
Creating home directory `/home/hdouser' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hdouser
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]
```

4) Switch to the new created user.



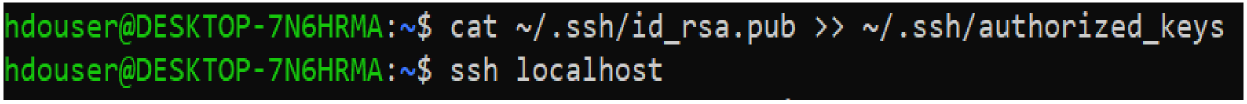
```
hduser@DESKTOP-7N6HRMA:~$ su - hdouser
hdouser@DESKTOP-7N6HRMA:~$
Welcome to Ubuntu 20.04 LTS (GNU/Linux 5.10.60.1-microsoft-standard-WSL2 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Tue Jan 31 10:58:12 IST 2023

  System load:  0.0                Processes:             12
  Usage of /:   2.6% of 250.98GB   Users logged in:       0
  Memory usage: 8%                 IPv4 address for eth0: 172.17.1.237
  Swap usage:   0%


305 updates can be installed immediately.
219 of these updates are security updates.
To see these additional updates run: apt list --upgradable

New release '22.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.



This message is shown once once a day. To disable it please create the
/home/hdouser/.hushlogin file.
```

5) Enable passwordless SSH for hadoop users and also define location where the keys to be stored.



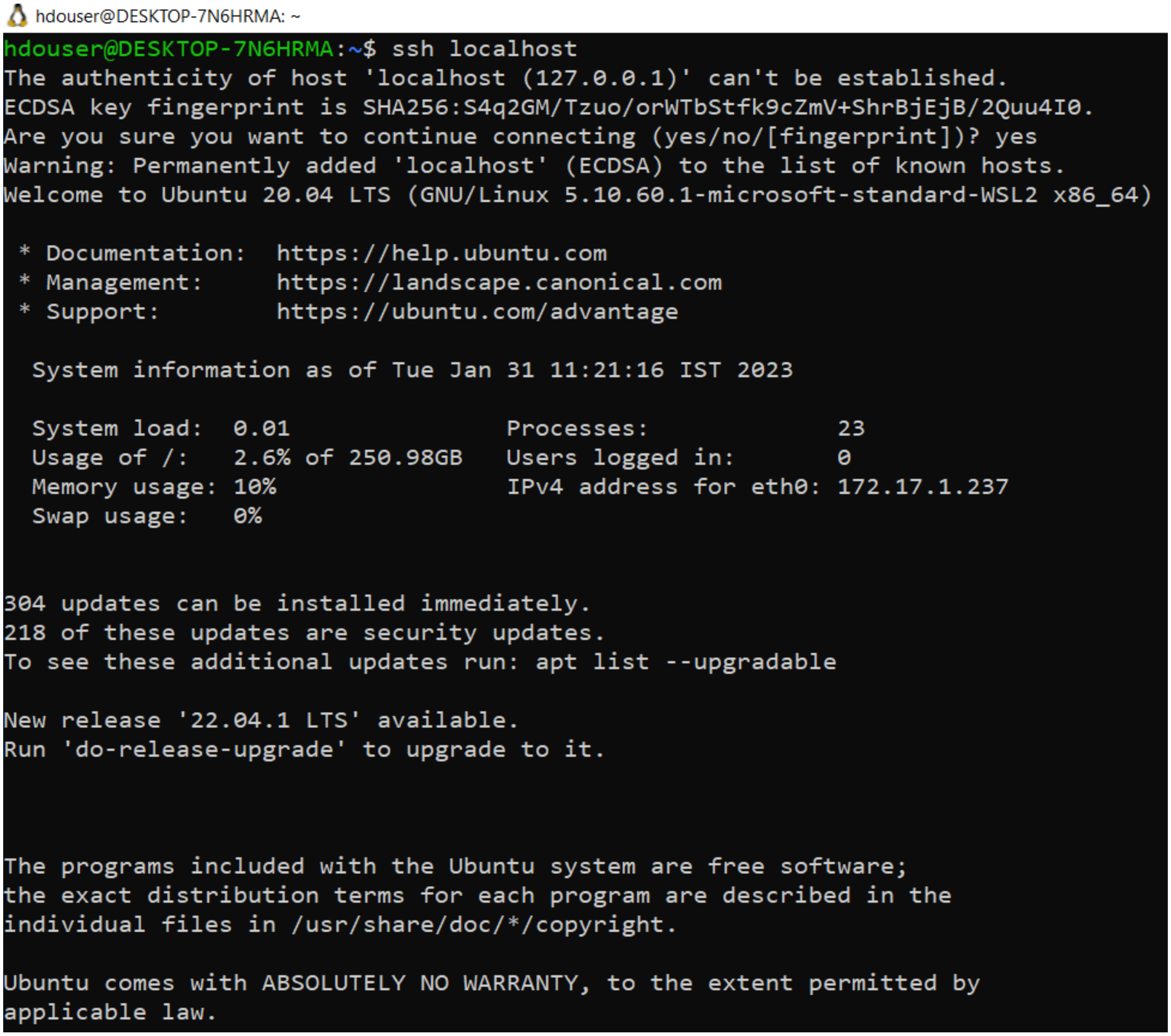
```
hdouser@DESKTOP-7N6HRMA: ~
hdouser@DESKTOP-7N6HRMA:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/hdouser/.ssh'.
Your identification has been saved in /home/hdouser/.ssh/id_rsa
Your public key has been saved in /home/hdouser/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:/iu9mbYoZWkuM3qfUqzBUNBr2la4rDSchaXW5hBd0uU hdouser@DESKTOP-7N6HRMA
The key's randomart image is:
+---[RSA 3072]----+
|    .+.o...      |
|    . =...       |
|     B o  E      |
|    * B .        |
|   o & +S.       |
|    * O.B        |
|   . + O..       |
|    . O o+oo     |
|    .o B+oB+     |
+----[SHA256]-----+
```

6) Using the cat command to store the public key as authorized_keys in ssh directory.



```
hdouser@DESKTOP-7N6HRMA:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hdouser@DESKTOP-7N6HRMA:~$ ssh localhost
```

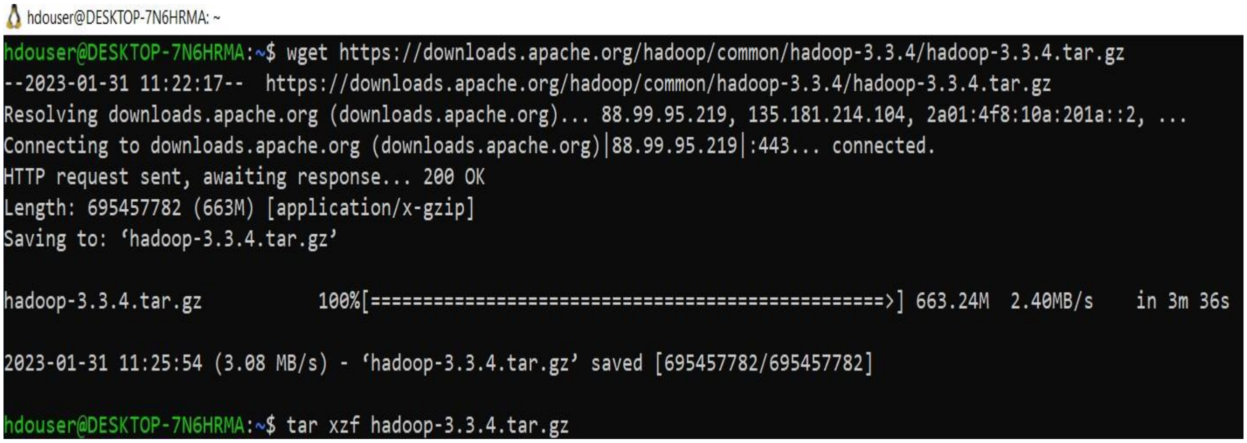
7) The new user is able to ssh without using the password every time, by using ssh localhost command.



hdouser@DESKTOP-7N6HRMA: ~

```
hdouser@DESKTOP-7N6HRMA:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:S4q2GM/Tzuo/orWTbStfk9cZmV+ShrBjEjB/2Quu4I0.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 20.04 LTS (GNU/Linux 5.10.60.1-microsoft-standard-WSL2 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Tue Jan 31 11:21:16 IST 2023

  System load:  0.01               Processes:             23
  Usage of /:   2.6% of 250.98GB   Users logged in:       0
  Memory usage: 10%                IPv4 address for eth0: 172.17.1.237
  Swap usage:   0%


304 updates can be installed immediately.
218 of these updates are security updates.
To see these additional updates run: apt list --upgradable

New release '22.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.



The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.
```

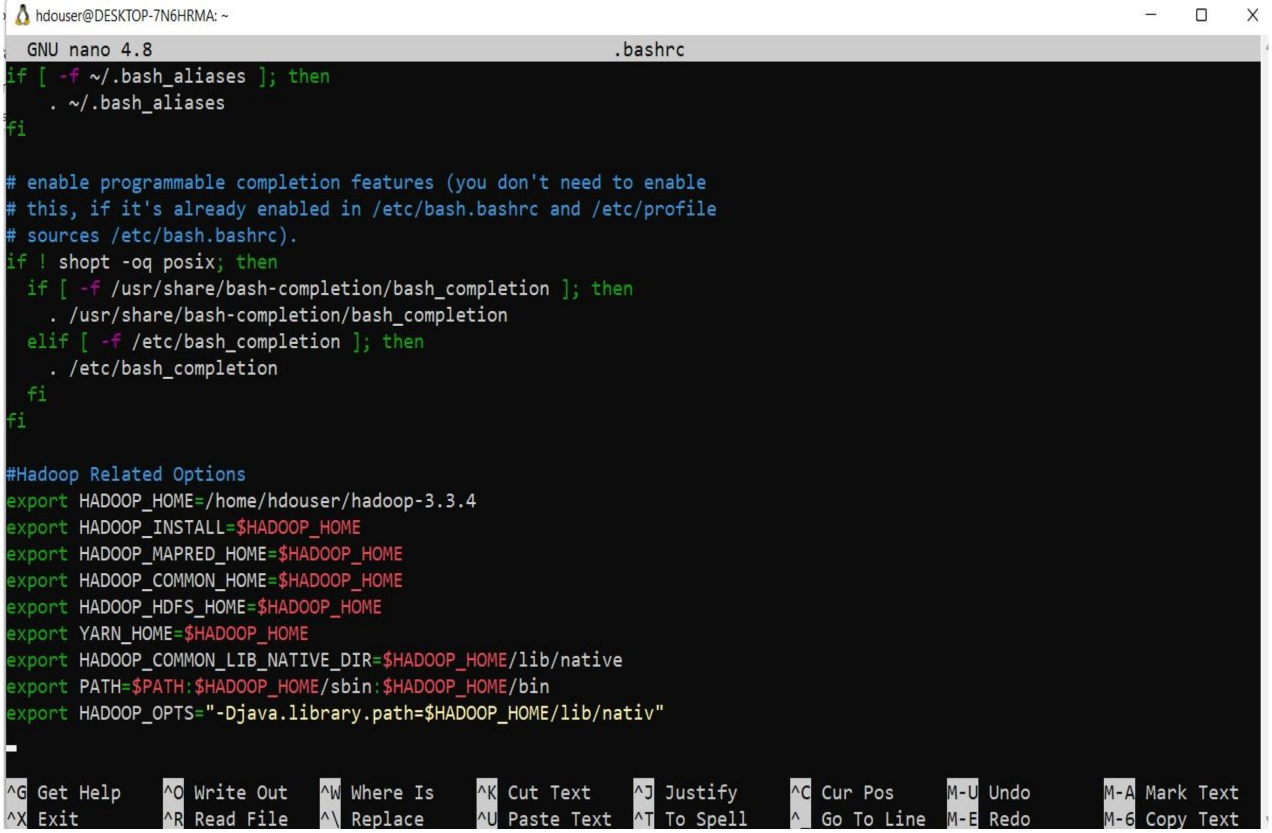
8) Downloading and extracting the hadoop package using the wget command.



```
hdouser@DESKTOP-7N6HRMA:~$ wget https://downloads.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz
--2023-01-31 11:22:17--  https://downloads.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 135.181.214.104, 2a01:4f8:10a:201a::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 695457782 (663M) [application/x-gzip]
Saving to: ‘hadoop-3.3.4.tar.gz’

hadoop-3.3.4.tar.gz          100%[===============================================>] 663.24M  2.40MB/s    in 3m 36s

2023-01-31 11:25:54 (3.08 MB/s) - ‘hadoop-3.3.4.tar.gz’ saved [695457782/695457782]

hdouser@DESKTOP-7N6HRMA:~$ tar xzf hadoop-3.3.4.tar.gz
```

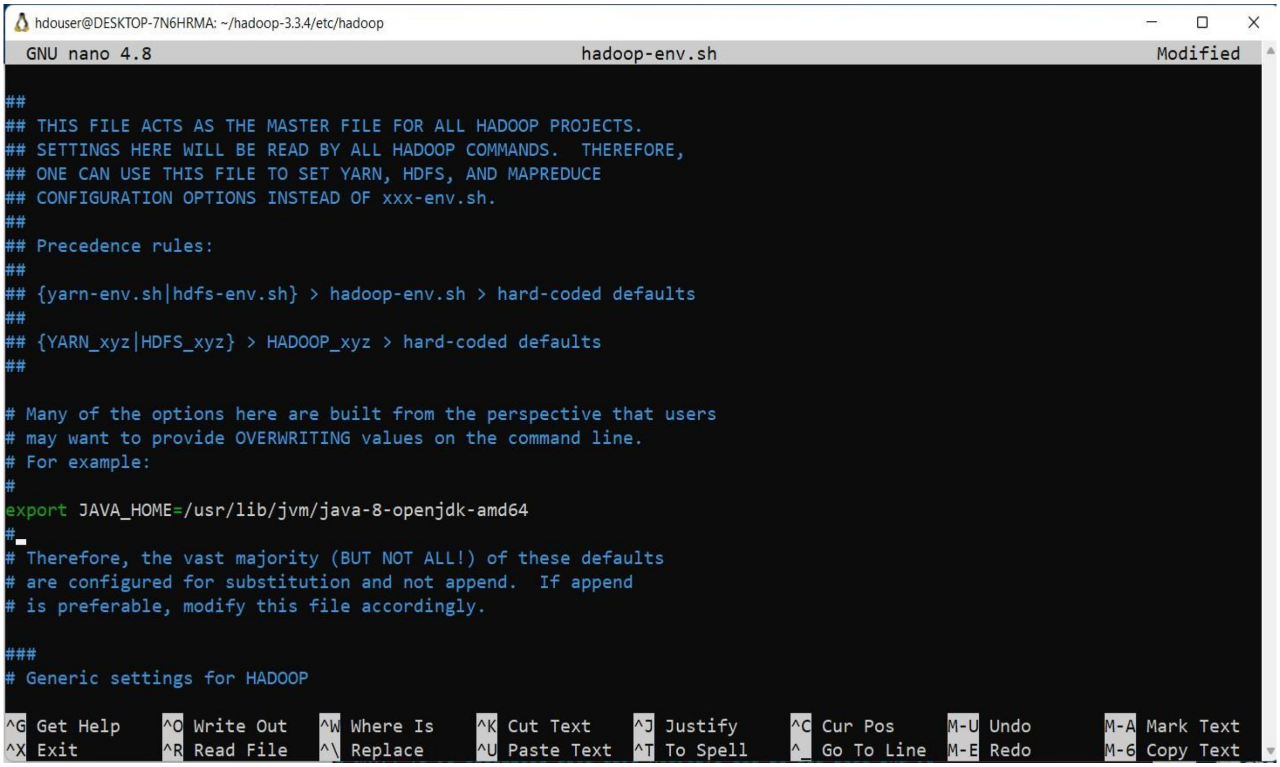
9) A Hadoop environment is configured by editing a set of configuration files:

a) bashrc:



```
if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

#Hadoop Related Options
export HADOOP_HOME=/home/hdouser/hadoop-3.3.4
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export 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/nativ"
```

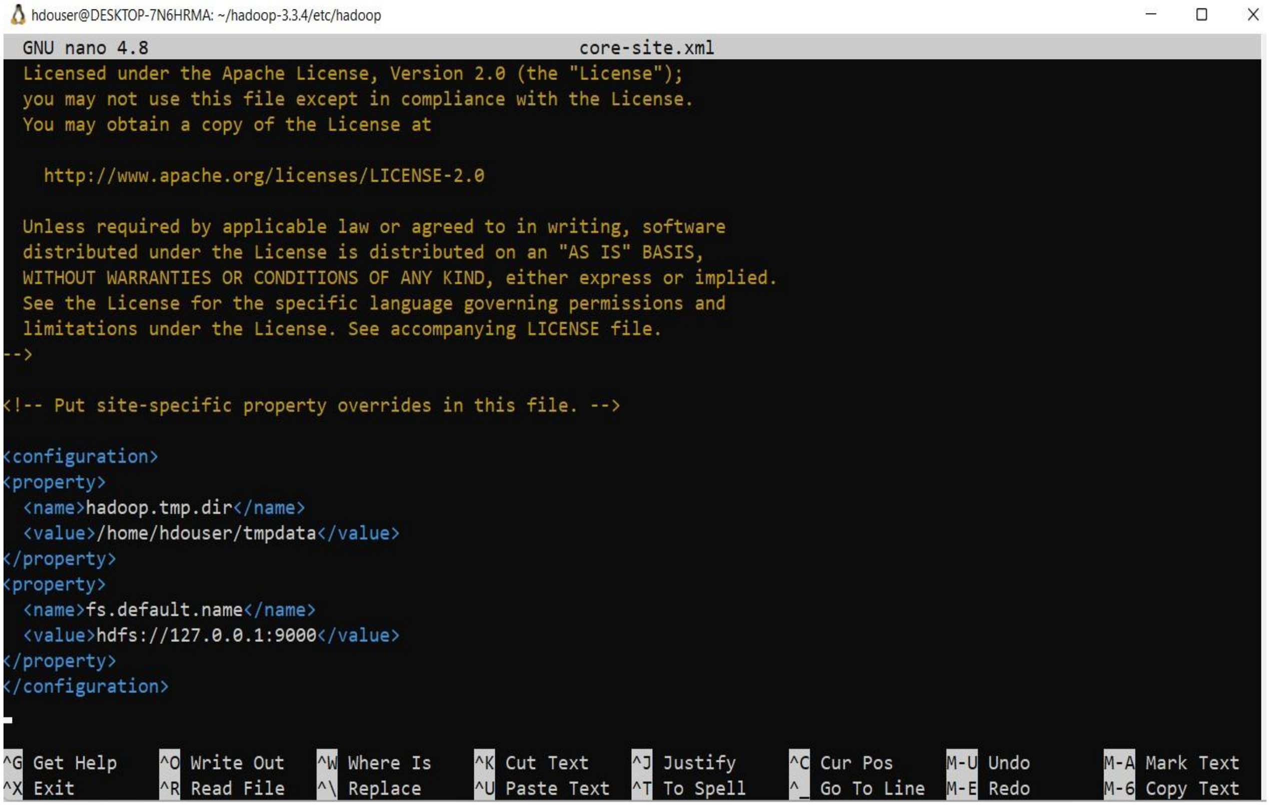
b) hadoop-env.sh:



hdouser@DESKTOP-7N6HRMA: ~/hadoop-3.3.4/etc/hadoop

```
  GNU nano 4.8                          hadoop-env.sh                          Modified

##
## THIS FILE ACTS AS THE MASTER FILE FOR ALL HADOOP PROJECTS.
## SETTINGS HERE WILL BE READ BY ALL HADOOP COMMANDS.  THEREFORE,
## ONE CAN USE THIS FILE TO SET YARN, HDFS, AND MAPREDUCE
## CONFIGURATION OPTIONS INSTEAD OF xxx-env.sh.
##
## Precedence rules:
##
## {yarn-env.sh|hdfs-env.sh} > hadoop-env.sh > hard-coded defaults
##
## {YARN_xyz|HDFS_xyz} > HADOOP_xyz > hard-coded defaults
##

# Many of the options here are built from the perspective that users
# may want to provide OVERWRITING values on the command line.
# For example:
#
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
#
# Therefore, the vast majority (BUT NOT ALL!) of these defaults
# are configured for substitution and not append.  If append
# is preferable, modify this file accordingly.

###
# Generic settings for HADOOP

^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos    M-U Undo      M-A Mark Text
^X Exit       ^R Read File  ^\ Replace    ^U Paste Text ^T To Spell   ^_ Go To Line M-E Redo      M-6 Copy Text
```

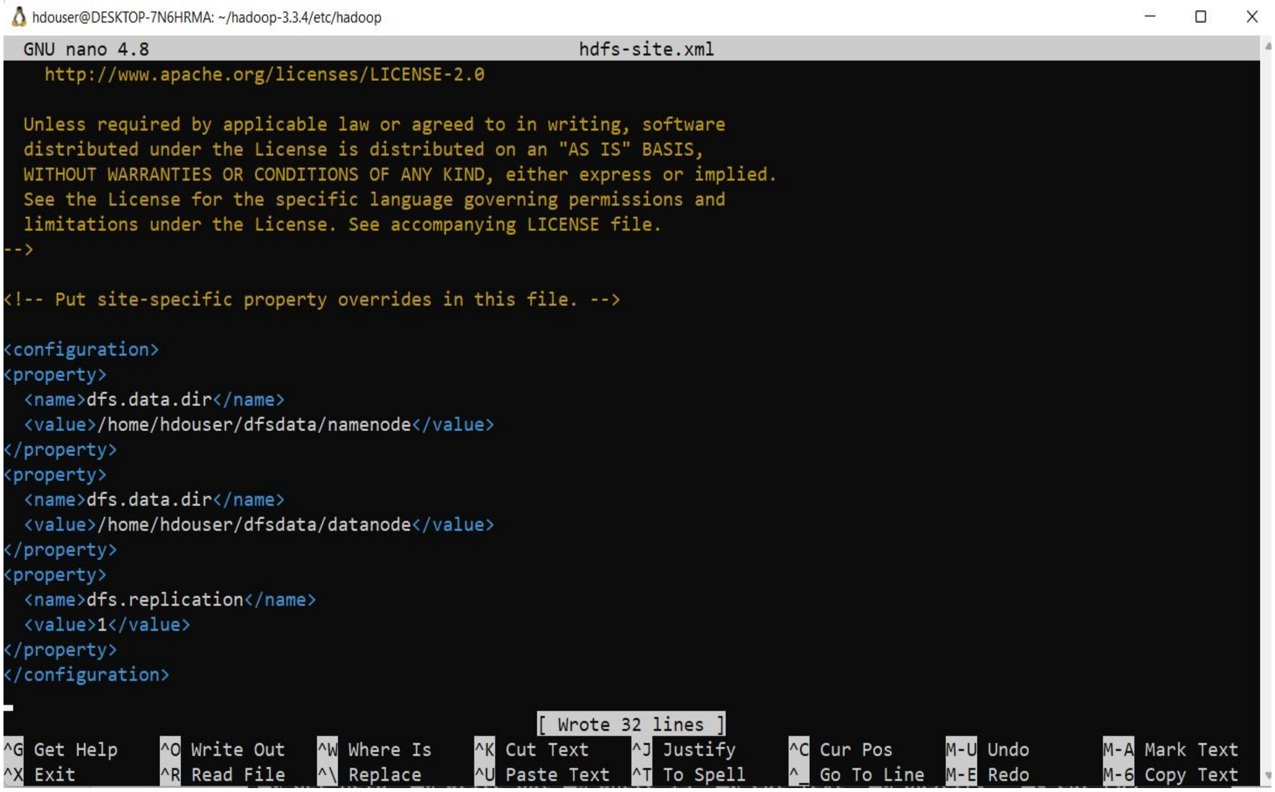
c) core-site.xml:



hdouser@DESKTOP-7N6HRMA: ~/hadoop-3.3.4/etc/hadoop

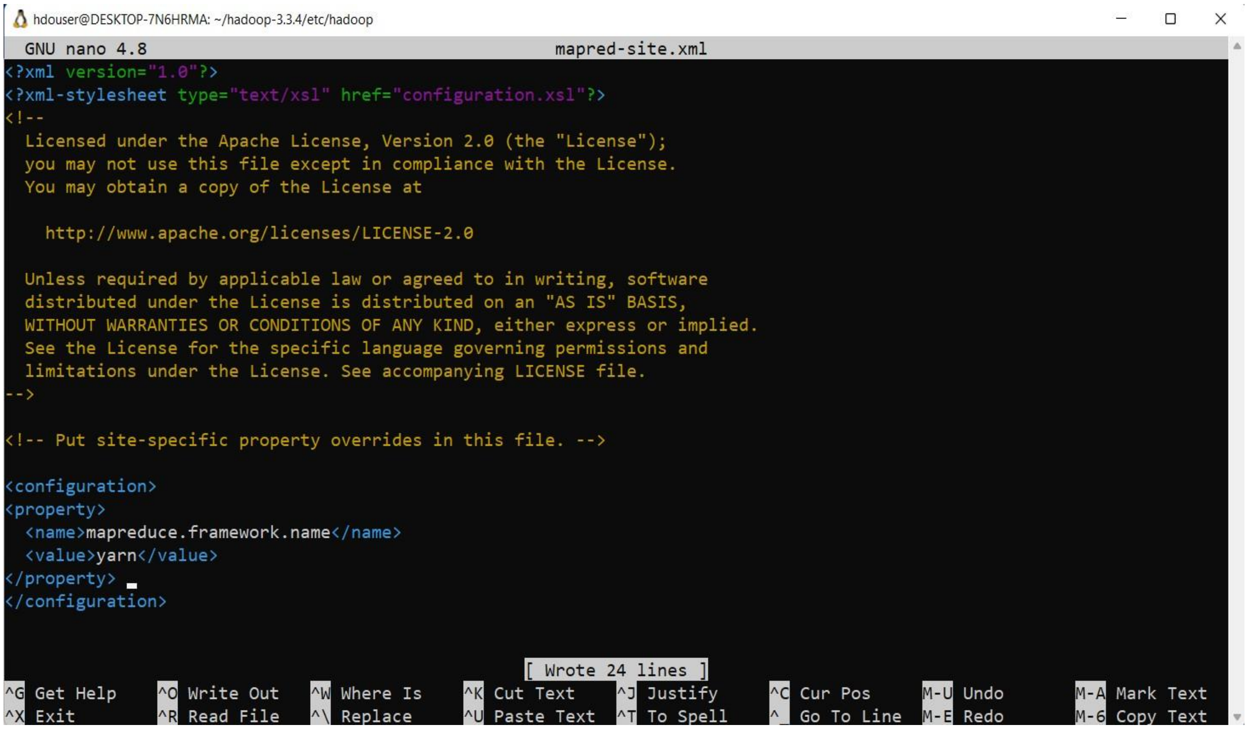
```
  GNU nano 4.8                          core-site.xml
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    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>hadoop.tmp.dir</name>
  <value>/home/hdouser/tmpdata</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>


^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos    M-U Undo      M-A Mark Text
^X Exit       ^R Read File  ^\ Replace    ^U Paste Text ^T To Spell   ^_ Go To Line M-E Redo      M-6 Copy Text
```

d) hdfs-site.xml:



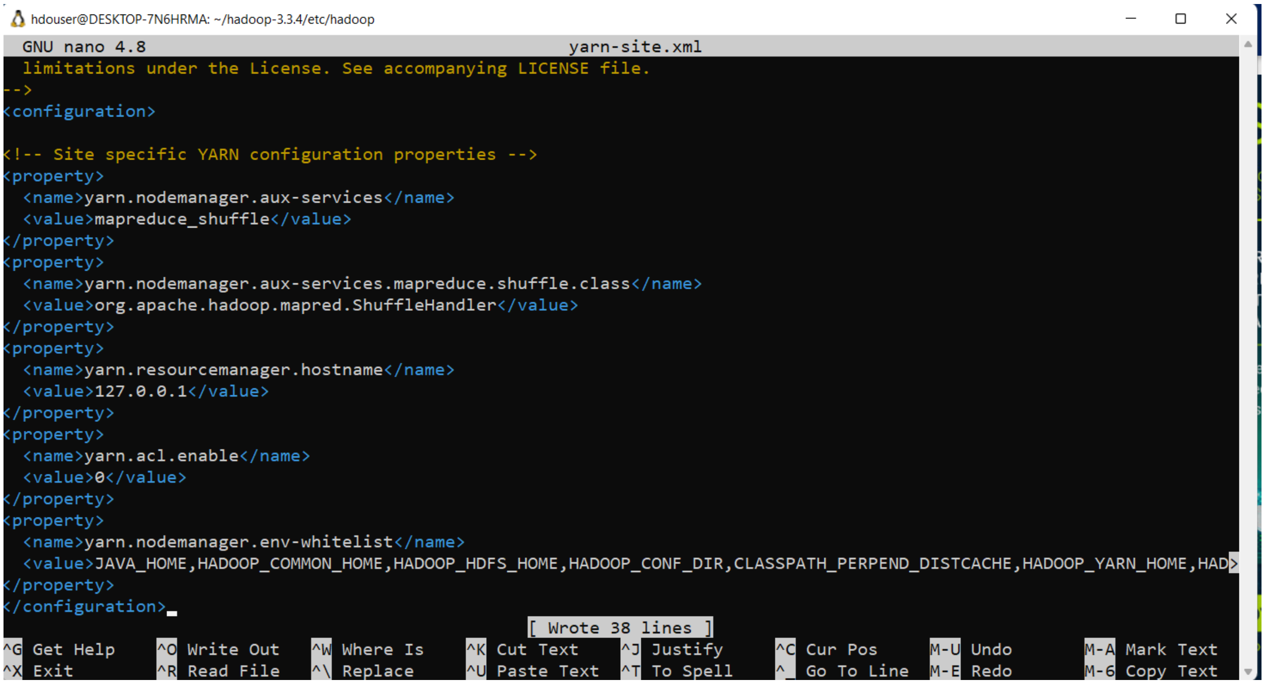
```
  GNU nano 4.8                                     hdfs-site.xml
    http://www.apache.org/licenses/LICENSE-2.0

  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.data.dir</name>
  <value>/home/hdouser/dfsdata/namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hdouser/dfsdata/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
</configuration>

                                                  [ Wrote 32 lines ]
^G Get Help     ^O Write Out    ^W Where Is     ^K Cut Text     ^J Justify      ^C Cur Pos      M-U Undo        M-A Mark Text
^X Exit         ^R Read File    ^\ Replace      ^U Paste Text   ^T To Spell     ^_ Go To Line   M-E Redo        M-6 Copy Text
```

e) mapred-site.xml:



```
  GNU nano 4.8                                     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>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
</configuration>

                                                  [ Wrote 24 lines ]
^G Get Help     ^O Write Out    ^W Where Is     ^K Cut Text     ^J Justify      ^C Cur Pos      M-U Undo        M-A Mark Text
^X Exit         ^R Read File    ^\ Replace      ^U Paste Text   ^T To Spell     ^_ Go To Line   M-E Redo        M-6 Copy Text
```

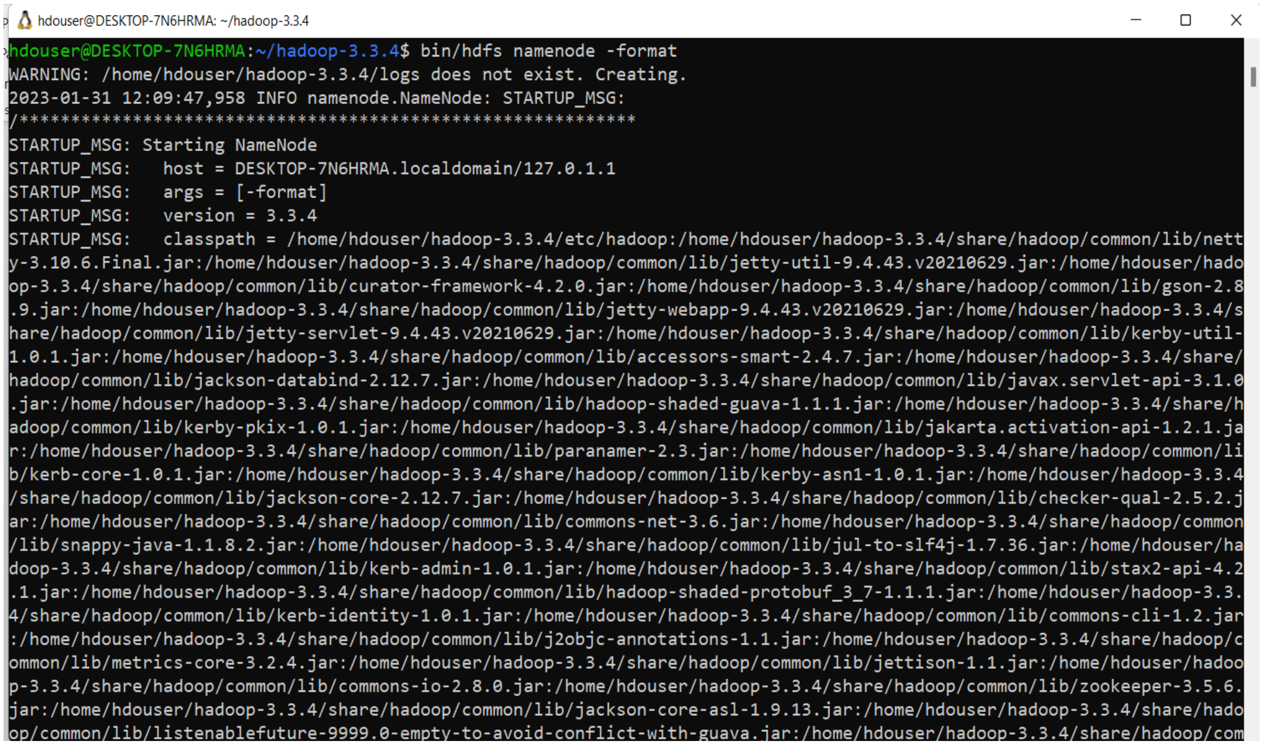
f) yarn-site.xml:



hdouser@DESKTOP-7N6HRMA: ~/hadoop-3.3.4/etc/hadoop

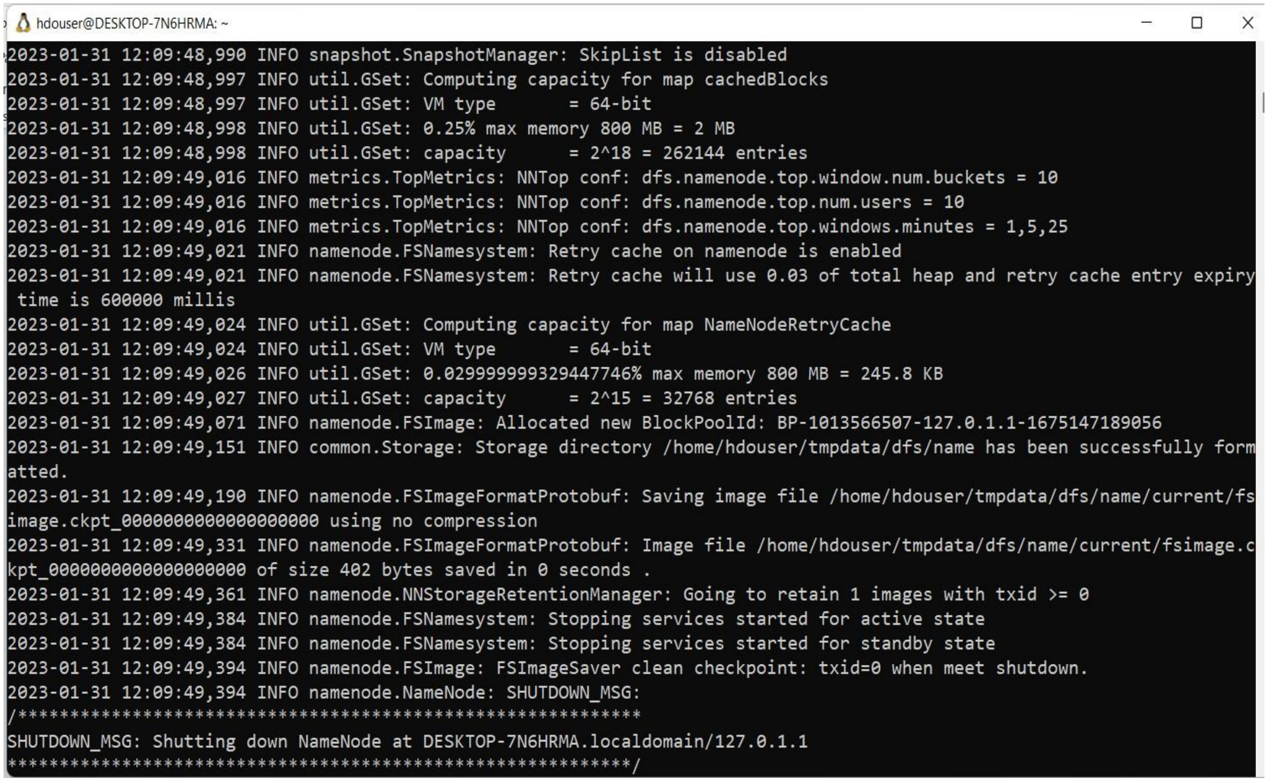
```
  GNU nano 4.8                                          yarn-site.xml
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
<property>
  <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>127.0.0.1</value>
</property>
<property>
  <name>yarn.acl.enable</name>
  <value>0</value>
</property>
<property>
  <name>yarn.nodemanager.env-whitelist</name>
  <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HAD>
</property>
</configuration>
                                             [ Wrote 38 lines ]
^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos     M-U Undo       M-A Mark Text
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line  M-E Redo       M-6 Copy Text
```

10) Format the NameNode before starting Hadoop services for the first time.



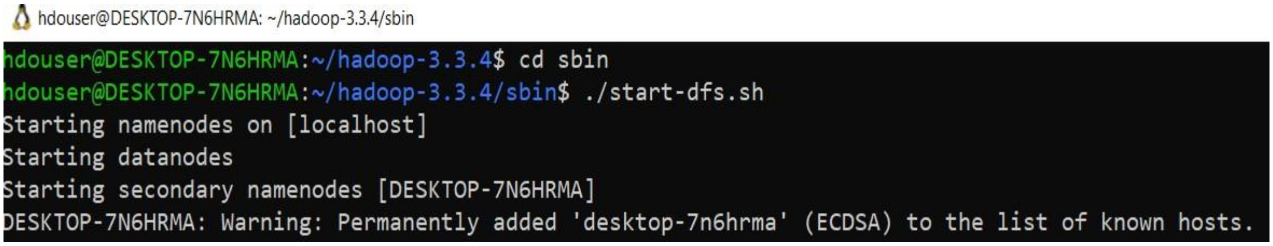
hdouser@DESKTOP-7N6HRMA: ~/hadoop-3.3.4

```
hdouser@DESKTOP-7N6HRMA:~/hadoop-3.3.4$ bin/hdfs namenode -format
WARNING: /home/hdouser/hadoop-3.3.4/logs does not exist. Creating.
2023-01-31 12:09:47,958 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = DESKTOP-7N6HRMA.localdomain/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.4
STARTUP_MSG:   classpath = /home/hdouser/hadoop-3.3.4/etc/hadoop:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/nett
y-3.10.6.Final.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/jetty-util-9.4.43.v20210629.jar:/home/hdouser/hado
op-3.3.4/share/hadoop/common/lib/curator-framework-4.2.0.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/gson-2.8
.9.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/jetty-webapp-9.4.43.v20210629.jar:/home/hdouser/hadoop-3.3.4/s
hare/hadoop/common/lib/jetty-servlet-9.4.43.v20210629.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/kerby-util-
1.0.1.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/accessors-smart-2.4.7.jar:/home/hdouser/hadoop-3.3.4/share/
hadoop/common/lib/jackson-databind-2.12.7.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/javax.servlet-api-3.1.0
.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/hadoop-shaded-guava-1.1.1.jar:/home/hdouser/hadoop-3.3.4/share/h
adoop/common/lib/kerby-pkix-1.0.1.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/jakarta.activation-api-1.2.1.ja
r:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/paranamer-2.3.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/li
b/kerb-core-1.0.1.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/kerby-asn1-1.0.1.jar:/home/hdouser/hadoop-3.3.4
/share/hadoop/common/lib/jackson-core-2.12.7.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/checker-qual-2.5.2.j
ar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/commons-net-3.6.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common
/lib/snappy-java-1.1.8.2.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/jul-to-slf4j-1.7.36.jar:/home/hdouser/ha
doop-3.3.4/share/hadoop/common/lib/kerb-admin-1.0.1.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/stax2-api-4.2
.1.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/hadoop-shaded-protobuf_3_7-1.1.1.jar:/home/hdouser/hadoop-3.3.
4/share/hadoop/common/lib/kerb-identity-1.0.1.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/commons-cli-1.2.jar
:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/j2objc-annotations-1.1.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/c
ommon/lib/metrics-core-3.2.4.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/jettison-1.1.jar:/home/hdouser/hadoo
p-3.3.4/share/hadoop/common/lib/commons-io-2.8.0.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/zookeeper-3.5.6.
jar:/home/hdouser/hadoop-3.3.4/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/home/hdouser/hadoop-3.3.4/share/hado
op/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/hdouser/hadoop-3.3.4/share/hadoop/com
```



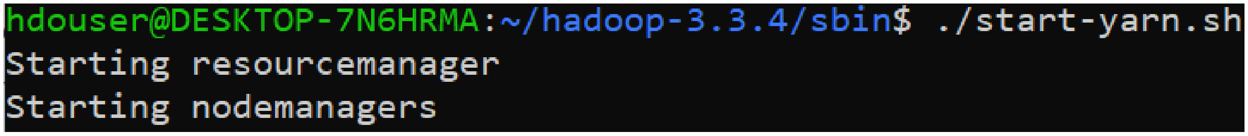
```
2023-01-31 12:09:48,990 INFO snapshot.SnapshotManager: SkipList is disabled
2023-01-31 12:09:48,997 INFO util.GSet: Computing capacity for map cachedBlocks
2023-01-31 12:09:48,997 INFO util.GSet: VM type       = 64-bit
2023-01-31 12:09:48,998 INFO util.GSet: 0.25% max memory 800 MB = 2 MB
2023-01-31 12:09:48,998 INFO util.GSet: capacity      = 2^18 = 262144 entries
2023-01-31 12:09:49,016 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2023-01-31 12:09:49,016 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2023-01-31 12:09:49,016 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2023-01-31 12:09:49,021 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2023-01-31 12:09:49,021 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry
 time is 600000 millis
2023-01-31 12:09:49,024 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2023-01-31 12:09:49,024 INFO util.GSet: VM type       = 64-bit
2023-01-31 12:09:49,026 INFO util.GSet: 0.029999999329447746% max memory 800 MB = 245.8 KB
2023-01-31 12:09:49,027 INFO util.GSet: capacity      = 2^15 = 32768 entries
2023-01-31 12:09:49,071 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1013566507-127.0.1.1-1675147189056
2023-01-31 12:09:49,151 INFO common.Storage: Storage directory /home/hdouser/tmpdata/dfs/name has been successfully form
atted.
2023-01-31 12:09:49,190 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hdouser/tmpdata/dfs/name/current/fs
image.ckpt_0000000000000000000 using no compression
2023-01-31 12:09:49,331 INFO namenode.FSImageFormatProtobuf: Image file /home/hdouser/tmpdata/dfs/name/current/fsimage.c
kpt_0000000000000000000 of size 402 bytes saved in 0 seconds .
2023-01-31 12:09:49,361 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-01-31 12:09:49,384 INFO namenode.FSNamesystem: Stopping services started for active state
2023-01-31 12:09:49,384 INFO namenode.FSNamesystem: Stopping services started for standby state
2023-01-31 12:09:49,394 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-01-31 12:09:49,394 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-7N6HRMA.localdomain/127.0.1.1
************************************************************/
```

11) Navigate to the hadoop-3.3.4/sbin directory to start the NameNode and DataNode.



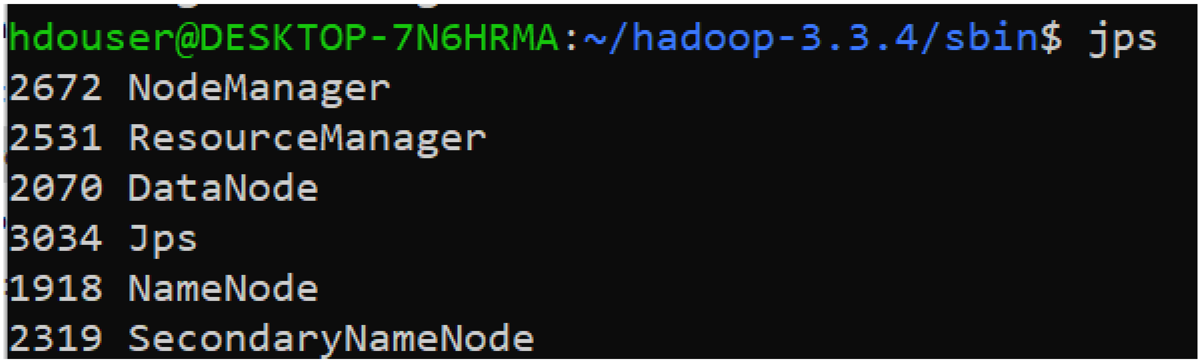
```
hdouser@DESKTOP-7N6HRMA:~/hadoop-3.3.4$ cd sbin
hdouser@DESKTOP-7N6HRMA:~/hadoop-3.3.4/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [DESKTOP-7N6HRMA]
DESKTOP-7N6HRMA: Warning: Permanently added 'desktop-7n6hrma' (ECDSA) to the list of known hosts.
```

12) Once the namenode, datanodes, and secondary namenode are up and running, start the YARN resource and nodemanagers.



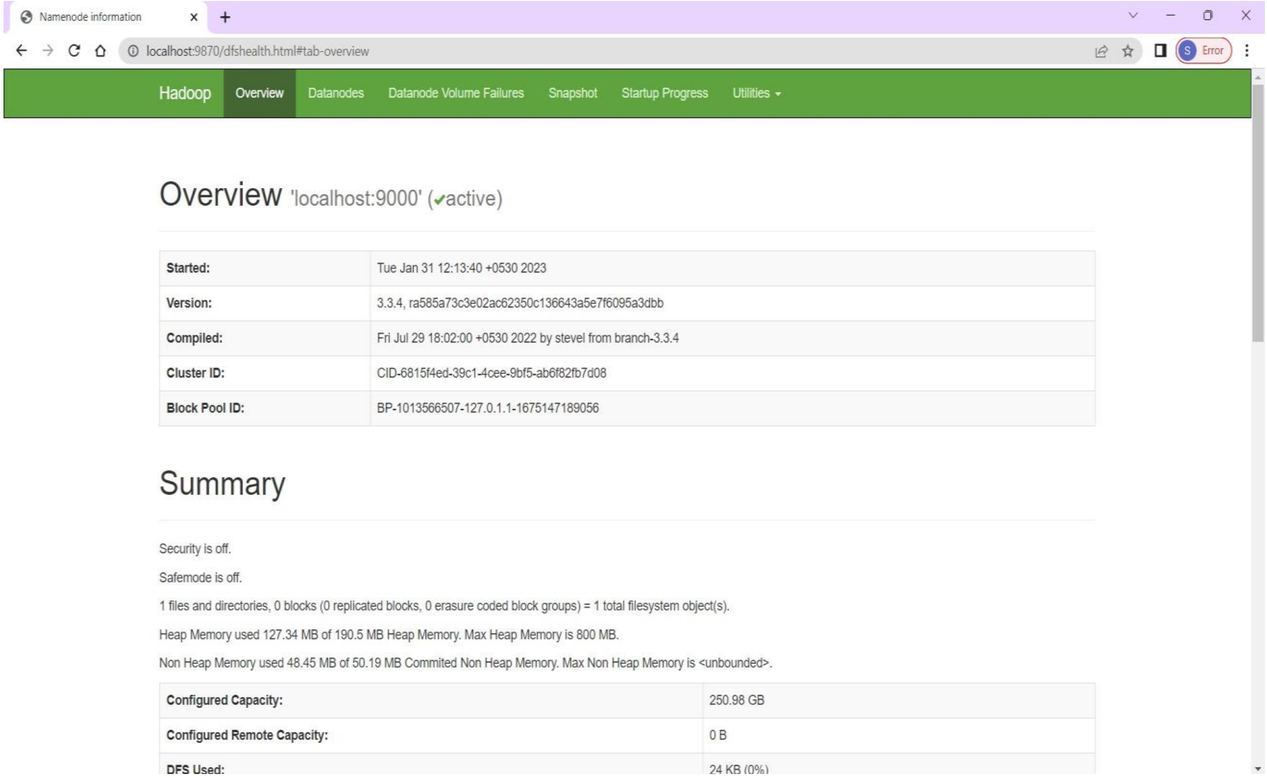
```
hdouser@DESKTOP-7N6HRMA:~/hadoop-3.3.4/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

13) Use jps command to check if all the daemons are active and running as Java processes.



```
hdouser@DESKTOP-7N6HRMA:~/hadoop-3.3.4/sbin$ jps
2672 NodeManager
2531 ResourceManager
2070 DataNode
3034 Jps
1918 NameNode
2319 SecondaryNameNode
```

14) To access hadoop namenode from the browser, we can use localhost for port 9870 by default.



localhost:9870/dfshealth.html#tab-overview

Overview 'localhost:9000' (✓active)

Started:	Tue Jan 31 12:13:40 +0530 2023
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled:	Fri Jul 29 18:02:00 +0530 2022 by stevel from branch-3.3.4
Cluster ID:	CID-6815f4ed-39c1-4cee-9bf5-ab6f82fb7d08
Block Pool ID:	BP-1013566507-127.0.1.1-1675147189056

Summary

Security is off.

Safemode is off.

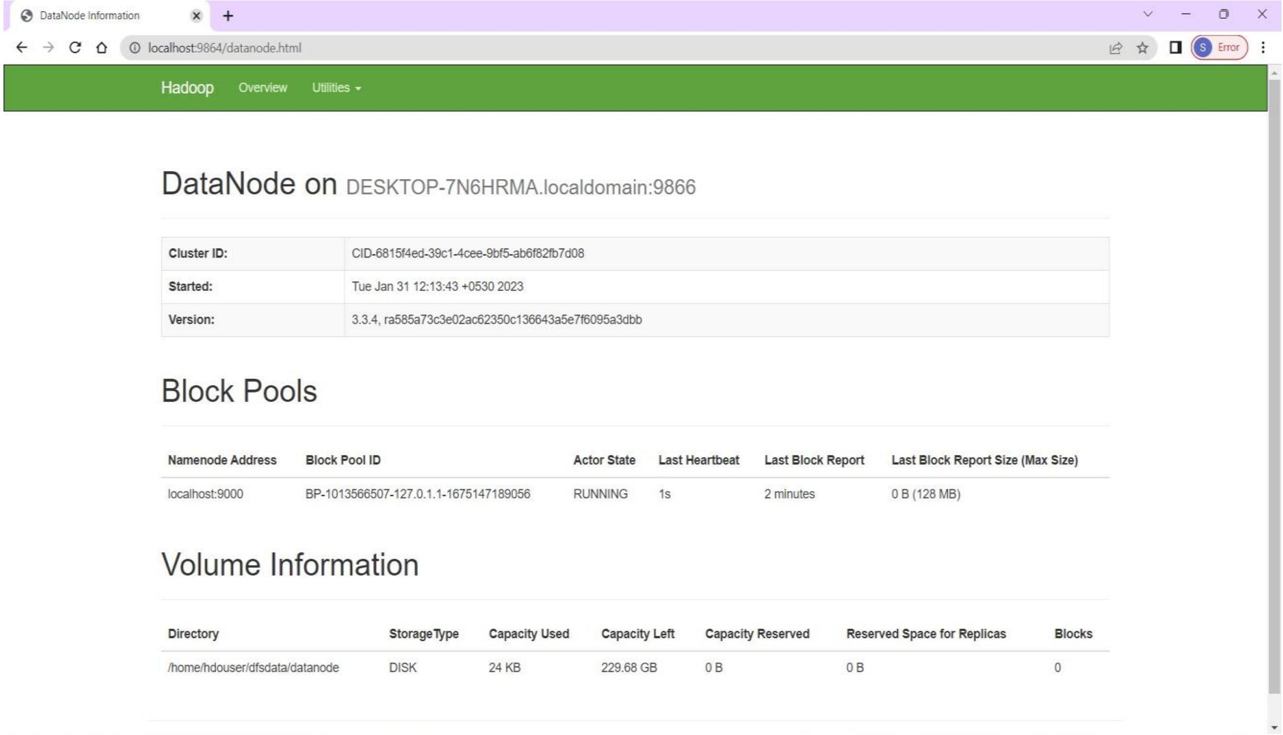
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 127.34 MB of 190.5 MB Heap Memory. Max Heap Memory is 800 MB.

Non Heap Memory used 48.45 MB of 50.19 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	250.98 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)

15) The default port 9864 is used to access individual DataNodes directly from the browser.



DataNode on DESKTOP-7N6HRMA.localdomain:9866

Cluster ID:	CID-6815f4ed-39c1-4cee-9bf5-ab6f82fb7d08
Started:	Tue Jan 31 12:13:43 +0530 2023
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb

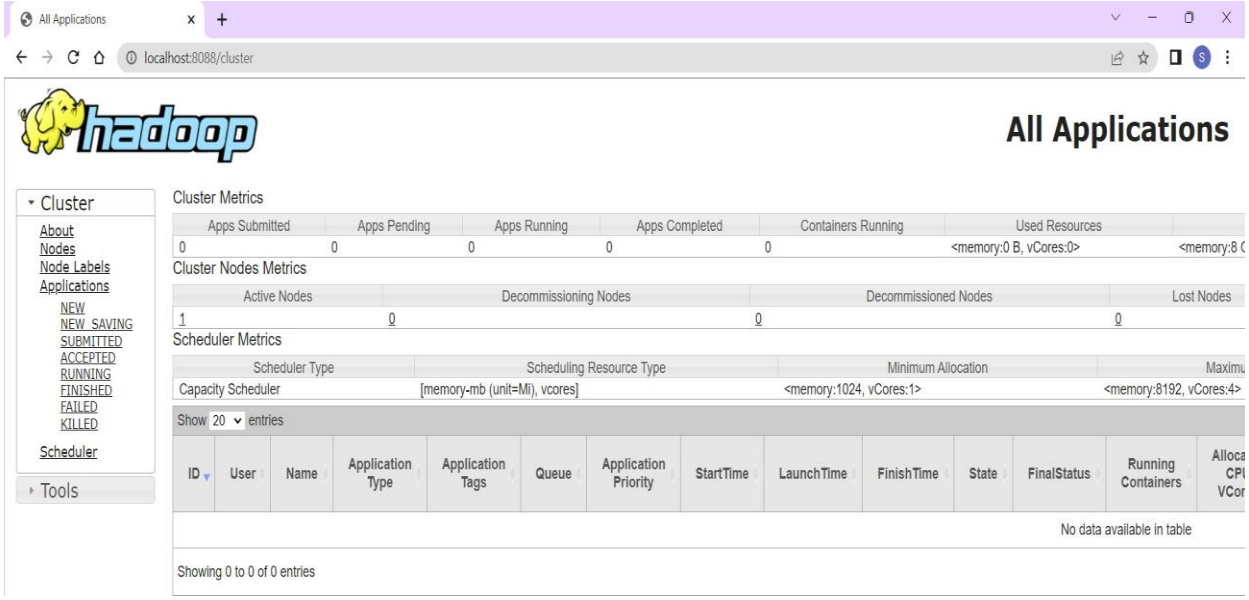
Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:9000	BP-1013566507-127.0.1.1-1675147189056	RUNNING	1s	2 minutes	0 B (128 MB)

Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
/home/hdouser/dfsdata/datanode	DISK	24 KB	229.68 GB	0 B	0 B	0

16) The YARN Resource Manager is accessible on port 8088.



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:8 C

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
1	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximu
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Alloca CPU VCor
No data available in table													

Showing 0 to 0 of 0 entries

Conclusion:

In conclusion, Hadoop is a powerful and scalable open-source big data processing framework that has evolved significantly over the years. The three main versions of Hadoop, Hadoop 1.x, Hadoop 2.x, and Hadoop 3.x, offer increasing levels of scalability, processing power, data management capabilities, and security features. Hadoop 1.x is the original version of Hadoop, which introduced the MapReduce processing engine and the Hadoop Distributed File System (HDFS) for storage management. Hadoop 2.x introduced YARN as the resource manager and improved HDFS, while Hadoop 3.x continues to build on these advancements with improved scalability, real-time processing capabilities, and data security.

In this experiment, I have set up hadoop 3.3.1 version by following the corresponding step as mentioned in the implementation part.