

**Experiment No. 01**

Semester	B.E. Semester VII – Computer Engineering
Subject	Big Data Analysis
Subject Professor In-charge	Prof. Pankaj Vanvari
Lab Professor In-charge	Dr. Umesh Kulkarni
Academic Year	2024-25
Student Name	Deep Salunkhe
Roll Number	21102A0014

**Title:** Set up a Hadoop cluster and verify its functionality

**Output:**

```
deepsalunkhee@LAPTOP-UGI x + v - □ ×
deepsalunkhee@LAPTOP-UGU6DF82:~$ java -version
openjdk version "1.8.0_412"
OpenJDK Runtime Environment (build 1.8.0_412-8u412-ga-1~22.04.1-b08)
OpenJDK 64-Bit Server VM (build 25.412-b08, mixed mode)
deepsalunkhee@LAPTOP-UGU6DF82:~$ |
```

```
GNU nano 6.2 .bashrc

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

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

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:/usr/lib/jvm/java-8-openjdk-amd64/bin
export HADOOP_HOME=~/.hadoop-3.2.3/
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_STREAMING=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export PDSH_RCMD_TYPE=ssh

^G Help      ^O Write Out  ^W Where Is   ^K Cut        ^T Execute
^X Exit      ^R Read File  ^\ Replace    ^U Paste      ^J Justify
```

```
deepsalunkhee@LAPTOP-UGI x + v - □ X
deepsalunkhee@LAPTOP-UGU6DF82:~$ java -version
openjdk version "1.8.0_412"
OpenJDK Runtime Environment (build 1.8.0_412-8u412-ga-1~22.04.1-b08)
OpenJDK 64-Bit Server VM (build 25.412-b08, mixed mode)
deepsalunkhee@LAPTOP-UGU6DF82:~$ sudo nano .bashrc
deepsalunkhee@LAPTOP-UGU6DF82:~$ sudo nano .bashrc
deepsalunkhee@LAPTOP-UGU6DF82:~$ sudo apt-get install ssh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  libwrap0 ncurses-term openssh-server openssh-sftp-server ssh-import-id
Suggested packages:
  molly-guard monkeysphere ssh-askpass
The following NEW packages will be installed:
  libwrap0 ncurses-term openssh-server openssh-sftp-server ssh
  ssh-import-id
0 upgraded, 6 newly installed, 0 to remove and 2 not upgraded.
Need to get 804 kB of archives.
After this operation, 6291 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-sft
p-server amd64 1:8.9p1-3ubuntu0.10 [38.9 kB]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 libwrap0 amd64 7.6.
q-31build2 [47.9 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-ser
ver amd64 1:8.9p1-3ubuntu0.10 [435 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 ssh all 1:8
.9p1-3ubuntu0.10 [4850 B]
Get:5 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 ncurses-ter
m all 6.3-2ubuntu0.1 [267 kB]
66% [5 ncurses-term 615 B/267 kB 0%]
```

```
deepsalunkhee@LAPTOP-UGI x + v - □ ×  
deepsalunkhee@LAPTOP-UGU6DF82:~$ wget https://dlcdn.apache.org/hadoop/commo  
n/hadoop-3.4.0/hadoop-3.4.0.tar.gz  
--2024-07-21 20:14:28-- https://dlcdn.apache.org/hadoop/common/hadoop-3.4.  
0/hadoop-3.4.0.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... con  
nected.  
HTTP request sent, awaiting response... 200 OK  
Length: 965537117 (921M) [application/x-gzip]  
Saving to: 'hadoop-3.4.0.tar.gz'  
  
hadoop-3.4.0.tar.g 23%[==> ] 216.44M 3.71MB/s eta 3m 15s |
```

```
deepsalunkhee@LAPTOP-UGI x + v - □ X
.sql
hadoop-3.4.0/sbin/FederationStateStore/SQLServer/dropUser.sql
hadoop-3.4.0/sbin/FederationStateStore/SQLServer/FederationStateStoreDatabase.sql
hadoop-3.4.0/sbin/start-dfs.cmd
hadoop-3.4.0/sbin/kms.sh
hadoop-3.4.0/sbin/yarn-daemon.sh
hadoop-3.4.0/sbin/workers.sh
hadoop-3.4.0/sbin/stop-all.cmd
hadoop-3.4.0/sbin/stop-all.sh
hadoop-3.4.0/sbin/stop-dfs.cmd
hadoop-3.4.0/sbin/hadoop-daemon.sh
hadoop-3.4.0/sbin/stop-secure-dns.sh
hadoop-3.4.0/sbin/httpfs.sh
hadoop-3.4.0/sbin/start-dfs.sh
hadoop-3.4.0/sbin/start-all.cmd
hadoop-3.4.0/sbin/hadoop-daemons.sh
hadoop-3.4.0/sbin/refresh-namenodes.sh
hadoop-3.4.0/sbin/start-balancer.sh
hadoop-3.4.0/sbin/start-all.sh
deepsalunkhee@LAPTOP-UGU6DF82:~$ cd hadoop-3.4.0/
deepsalunkhee@LAPTOP-UGU6DF82:~/hadoop-3.4.0$ cd etc/hadoop/
deepsalunkhee@LAPTOP-UGU6DF82:~/hadoop-3.4.0/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                yarn-site.xml
kms-env.sh                  yarnservice-log4j.properties
deepsalunkhee@LAPTOP-UGU6DF82:~/hadoop-3.4.0/etc/hadoop$ |
```



```
deepsalunkhee@LAPTOP-UGI x + v - □ ×
GNU nano 6.2 hadoop-env.sh
# 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.

# Set Hadoop-specific environment variables here.

##
## 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:
#
# 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
###

# Technically, the only required environment variable is JAVA_HOME.
# All others are optional. However, the defaults are probably not
# preferred. Many sites configure these options outside of Hadoop,
# [ Wrote 434 lines ]
^G Help      ^O Write Out ^W Where Is  ^K Cut       ^T Execute
^X Exit      ^R Read File ^\ Replace   ^U Paste     ^J Justify
```

```
deepsalunkhee@LAPTOP-UGI  ×  +  ▾  -  □  ×
GNU nano 6.2  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://localhost:9000</value>  </property>
  <property>
    <name>hadoop.proxyuser.dataflair.groups</name> <value>*</value>
  </property>
  <property>
    <name>hadoop.proxyuser.dataflair.hosts</name> <value>*</value>
  </property>
  <property>
    <name>hadoop.proxyuser.server.hosts</name> <value>*</value>
  </property>
  <property>
    <name>hadoop.proxyuser.server.groups</name> <value>*</value>
  </property>
</configuration>
```

```
deepsalunkhee@LAPTOP-UGI x + v - □ X
GNU nano 6.2 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>
</configuration>

File Name to Write: hdfs-site.xml
^G Help          M-D DOS Format   M-A Append       M-B Backup File
^C Cancel        M-M Mac Format   M-P Prepend      ^T Browse
```



```
deepsalunkhee@LAPTOP-UGI x + v - □ X
GNU nano 6.2 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>
  <property>
    <name>mapreduce.application.classpath</name>
    <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/sh
  </property>
</configuration>

[ Wrote 28 lines ]
^G Help      ^O Write Out  ^W Where Is   ^K Cut        ^T Execute
^X Exit      ^R Read File  ^\ Replace    ^U Paste      ^J Justify
```

```
deepsalunkhee@LAPTOP-UGI x + v - □ ×
GNU nano 6.2 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>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>
    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASS
  </property>
</configuration>

File Name to Write: yarn-site.xml
^G Help      M-D DOS Format  M-A Append    M-B Backup File
^C Cancel    M-M Mac Format  M-P Prepend   ^T Browse
```

```

yarn-env.sh                                yarnservice-log4j.properties
deepsalunkhee@LAPTOP-UGU6DF82:~/hadoop-3.4.0/etc/hadoop$ sudo nano yarn-site.xml
deepsalunkhee@LAPTOP-UGU6DF82:~/hadoop-3.4.0/etc/hadoop$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:8hCH3F4qn1MafbgSnF8xt0oJdjQnt07AGnN5XFkBuFY.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? y
Please type 'yes', 'no' or the fingerprint: yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts
.
deepsalunkhee@localhost's password:
Welcome to Ubuntu 22.04.4 LTS (GNU/Linux 5.15.133.1-microsoft-standard-WSL2 x86_64)

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

Last login: Sun Jul 21 20:02:04 2024
deepsalunkhee@LAPTOP-UGU6DF82:~$ |

```

```

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

Last login: Sun Jul 21 20:02:04 2024
deepsalunkhee@LAPTOP-UGU6DF82:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
Generating public/private rsa key pair.
Your identification has been saved in /home/deepsalunkhee/.ssh/id_rsa
Your public key has been saved in /home/deepsalunkhee/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:a4YXwjYGD03ePb9RadyaU1iqUkP41jfchNxpA53yJqA deepsalunkhee@LAPTOP-UGU6DF82
The key's randomart image is:
+---[RSA 3072]---+
|      .      .oo+ |
|    + . .o oo%o |
|  o o . +.+.%.= |
|    =  E +o*.Oo |
|    B S ..+ *.. |
|  o + o . o .   |
|    . = .       |
|      +         |
+---[SHA256]-----+
deepsalunkhee@LAPTOP-UGU6DF82:~$ |

```

```
deepsalunkhee@LAPTOP-UGI  X  +  v  -  □  X

2024-07-21 21:13:19,871 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.t
op.num.users = 10
2024-07-21 21:13:19,871 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.t
op.windows.minutes = 1,5,25
2024-07-21 21:13:19,876 INFO namenode.FSNamesystem: Retry cache on namenode
is enabled
2024-07-21 21:13:19,876 INFO namenode.FSNamesystem: Retry cache will use 0.
03 of total heap and retry cache entry expiry time is 600000 millis
2024-07-21 21:13:19,878 INFO util.GSet: Computing capacity for map NameNode
RetryCache
2024-07-21 21:13:19,878 INFO util.GSet: VM type          = 64-bit
2024-07-21 21:13:19,878 INFO util.GSet: 0.029999999329447746% max memory 1.
7 GB = 534.7 KB
2024-07-21 21:13:19,878 INFO util.GSet: capacity          = 2^16 = 65536 entrie
s
2024-07-21 21:13:19,904 INFO namenode.FSImage: Allocated new BlockPoolId: B
P-83941552-127.0.1.1-1721576599895
2024-07-21 21:13:19,947 INFO common.Storage: Storage directory /tmp/hadoop-
deepsalunkhee/dfs/name has been successfully formatted.
2024-07-21 21:13:20,070 INFO namenode.FSImageFormatProtobuf: Saving image f
ile /tmp/hadoop-deepsalunkhee/dfs/name/current/fsimage.ckpt_0000000000000000
0000 using no compression
2024-07-21 21:13:20,203 INFO namenode.FSImageFormatProtobuf: Image file /tm
p/hadoop-deepsalunkhee/dfs/name/current/fsimage.ckpt_00000000000000000000 of
size 408 bytes saved in 0 seconds .
2024-07-21 21:13:20,232 INFO namenode.NNStorageRetentionManager: Going to r
etain 1 images with txid >= 0
2024-07-21 21:13:20,237 INFO blockmanagement.DatanodeManager: Slow peers co
llection thread shutdown
2024-07-21 21:13:20,253 INFO namenode.FSNamesystem: Stopping services start
ed for active state
2024-07-21 21:13:20,254 INFO namenode.FSNamesystem: Stopping services start
ed for standby state
2024-07-21 21:13:20,258 INFO namenode.FSImage: FSImageSaver clean checkpoin
t: txid=0 when meet shutdown.
2024-07-21 21:13:20,259 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at LAPTOP-UGU6DF82/127.0.1.1
*****/
deepsalunkhee@LAPTOP-UGU6DF82:~$

*****/
deepsalunkhee@LAPTOP-UGU6DF82:~$ export PDSH_RCMD_TYPE=ssh
deepsalunkhee@LAPTOP-UGU6DF82:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as deepsalunkhee 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 [LAPTOP-UGU6DF82]
LAPTOP-UGU6DF82: Warning: Permanently added 'laptop-ugu6df82' (ED25519) to
the list of known hosts.
Starting resourcemanager
Starting nodemanagers
deepsalunkhee@LAPTOP-UGU6DF82:~$ |
```



localhost:9870/dfshealth.html#tab-overview

VPN

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

# Overview

'localhost:9000' (active)

Started:	Sun Jul 21 21:15:32 +0530 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaee760
Compiled:	Mon Mar 04 12:05:00 +0530 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-45a54ff5-1c7c-477c-91c0-c93812071a74
Block Pool ID:	BP-83941552-127.0.1.1-1721576599895

## 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).



```
deepsalunkhee@LAPTOP-UGI x + v - □ ×

deepsalunkhee@LAPTOP-UGU6DF82:~$ hadoop
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
or hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
where CLASSNAME is a user-provided Java class

OPTIONS is none or any of:

--config dir          Hadoop config directory
--debug              turn on shell script debug mode
--help              usage information
buildpaths          attempt to add class files from build
                    tree
hostnames list[,of,host,names] hosts to use in worker mode
hosts filename      list of hosts to use in worker mode
loglevel level      set the log4j level for this command
workers            turn on worker mode

SUBCOMMAND is one of:

Admin Commands:

daemonlog          get/set the log level for each daemon

Client Commands:

archive            create a Hadoop archive
checknative        check native Hadoop and compression libraries availability
classpath          prints the class path needed to get the Hadoop jar and the
                    required libraries
confest           validate configuration XML files
credential         interact with credential providers
distch            distributed metadata changer
distcp            copy file or directories recursively
dtutil            operations related to delegation tokens
envvars           display computed Hadoop environment variables
fedbalance        balance data between sub-clusters
fs                run a generic filesystem user client
gridmix           submit a mix of synthetic job, modeling a profiled from
                    production load
jar <jar>         run a jar file. NOTE: please use "yarn jar" to launch YARN
                    applications, not this command.
jnipath           prints the java.library.path
kdiag            Diagnose Kerberos Problems
kerbname          show auth_to_local principal conversion
key              manage keys via the KeyProvider
rbfbalance        move directories and files across router-based federation
                    namespaces

Daemon Commands:

kms                run KMS, the Key Management Server
registrydns       run the registry DNS server

SUBCOMMAND may print help when invoked w/o parameters or with -h.
deepsalunkhee@LAPTOP-UGU6DF82:~$ hadoop fs -ls
ls: '.': No such file or directory
deepsalunkhee@LAPTOP-UGU6DF82:~$ hadoop fs -ls /
Found 1 items
drwxr-xr-x - deepsalunkhee supergroup          0 2024-07-21 21:20 /users
deepsalunkhee@LAPTOP-UGU6DF82:~$ |
```

```
SUBCOMMAND may print help when invoked w/o parameters or with -h.
deepsalunkhee@LAPTOP-UGU6DF82:~$ hadoop fs -ls
ls: `.`: No such file or directory
deepsalunkhee@LAPTOP-UGU6DF82:~$ hadoop fs -ls /
Found 1 items
drwxr-xr-x - deepsalunkhee supergroup          0 2024-07-21 21:20 /users
deepsalunkhee@LAPTOP-UGU6DF82:~$ stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as deepsalunkhee in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [LAPTOP-UGU6DF82]
Stopping nodemanagers
Stopping resourcemanager
deepsalunkhee@LAPTOP-UGU6DF82:~$ |
```