Faculty: Ms. Sana Shaikh **Subject:** Big Data Analytics 2020-2021

Name: Alston Fernandes **Roll no:** 19 **Exp no:** 1

Topic:	Installing Hadoop in Pseudo Distributed Mode & getting familiar with Hadoop HDFS commands.
Prerequisite:	Basic knowledge of Hadoop is required.
Mapping With COs:	CSL7012.1
Objective:	• To acquire the knowledge of different components present in Hadoop Ecosystem. • To understand the Hadoop 2.x architecture. • To learn how to set up and configure Hadoop in Pseudo Distributed Mode. • To learn how to work with the Hadoop HDFS file system. • Getting familiar with Hadoop HDFS commands.
Outcome:	• Students will be able to differentiate between the local file system and HDFS. • Students will be able to install Hadoop clusters and configure it in any installation mode. • Students will be able to start and work with the Hadoop HDFS file system using various Hadoop commands.
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.
Deliverables:	**1. Explain all Hadoop installation modes.** **Local (Standalone) Mode:** • This mode is primarily used for development, debugging, and testing purposes. • In this mode, Hadoop runs on a single machine without the need for a distributed cluster. It does not take advantage of Hadoop's distributed computing capabilities. • Only the essential Hadoop components are active, such as HDFS for local file storage and MapReduce for data processing. There is no need for YARN resource management.

- It is easy to set up and suitable for small-scale tasks where distributed computing benefits are unnecessary.

Pseudo-Distributed Mode:

- This mode is often used for learning and development on a single machine that simulates a distributed cluster environment.
- In this mode, all Hadoop components run on a single machine, but they communicate as if they were part of a distributed cluster. It allows developers to test their applications in an environment that resembles a real Hadoop cluster.
- HDFS, YARN, and MapReduce are fully functional.
- Provides a realistic testing environment for Hadoop applications without the complexity of setting up a multi-node cluster.

Cluster (Fully-Distributed) Mode:

- This is the production-ready mode for deploying Hadoop in a distributed cluster, suitable for processing large-scale data.
- In this mode, Hadoop operates in a true distributed cluster environment with multiple nodes. Each node has its HDFS storage and runs various Hadoop services.
- HDFS, YARN, MapReduce, and other Hadoop ecosystem components are distributed across the cluster nodes.
- Offers scalability, fault tolerance, and efficient distributed data processing capabilities for big data workloads.

2. List down components of Hadoop Cluster.

Core Components:

1. **HDFS (Hadoop Distributed File System)**:
2. **YARN (Yet Another Resource Negotiator)**:
3. **MapReduce**:

Optional Components (Hadoop Ecosystem):

1. **Hbase**:
2. **Hive**:
3. **Pig**:
4. **Sqoop**:
5. **Flume**:

6. **Oozie**:
7. **ZooKeeper**:
8. **Mahout**:
9. **Ambari**:

3. Take a snapshot of each step of hadoop installation and for all HDFS commands with input and its output.

1. Update the system

```
dbit@complab3:~$ pwd
/home/dbit
dbit@complab3:~$ sudo apt-get update
[sudo] password for dbit:
Hit:1 http://in.archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:3 http://in.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:4 http://in.archive.ubuntu.com/ubuntu jammy-backports InRelease
Reading package lists... Done
dbit@complab3:~$
```

2. Install jdk

```
dbit@complab3: $ sudo apt-get install default-jdk
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java default-jdk-headless default-jre default-jre-headless fonts-dejavu-extra java-common libatk-
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev
  openjdk-11-jdk openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev xorg-sgml-doctools xtr
Suggested packages:
  libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-11-demo openjdk-11-source visualvm fonts-ipafont-got
  fonts-wqy-microhei | fonts-wqy-zenhei
The following NEW packages will be installed:
  ca-certificates-java default-jdk default-jdk-headless default-jre default-jre-headless fonts-dejavu-extra java-co
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev
  openjdk-11-jdk openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev xorg-sgml-doctools xtr
0 upgraded, 24 newly installed, 0 to remove and 2 not upgraded.
Need to get 122 MB of archives.
After this operation, 275 MB of additional disk space will be used.
Do you want to continue? [Y/n]
```

```
dbit@complab3: $ java --version
openjdk 11.0.20.1 2023-08-24
OpenJDK Runtime Environment (build 11.0.20.1+1-post-Ubuntu-0ubuntu122.04)
OpenJDK 64-Bit Server VM (build 11.0.20.1+1-post-Ubuntu-0ubuntu122.04, mixed mode, sharing)
dbit@complab3: $
```

3. Install ssh

```
dbit@complab3:~$ sudo apt-get install ssh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh-import-id
Suggested packages:
  molly-guard monkeysphere ssh-askpass
The following NEW packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh ssh-import-id
0 upgraded, 5 newly installed, 0 to remove and 2 not upgraded.
Need to get 755 kB of archives.
After this operation, 6,180 kB of additional disk space will be used.
Do you want to continue? [Y/n]
```

4. Generate ssh key pair

```
dbit@complab3:~$ ssh-keygen -t rsa -P ''
Generating public/private rsa key pair.
Enter file in which to save the key (/home/dbit/.ssh/id_rsa):
Your identification has been saved in /home/dbit/.ssh/id_rsa
Your public key has been saved in /home/dbit/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:LNjr3DIPlykKj12yNBKE7PezBxYvbuVUdjXH7GwHOjw dbit@complab3
The key's randomart image is:
+---[RSA 3072]----+
|              o  |
|..           o.+ |
|...         ...=. |
|..   + . o .E  +.|
| .... = S .  o. .|
|  ...+ * o       |
|  o *o@ +        |
|   B @=B         |
|  . *.++o        |
+----[SHA256]-----+
dbit@complab3:~$
```

5. Connect to localhost via ssh

```
dbit@complab3:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:fEyXPWreVyvn45GMOvwmfpP6IlOJjVwbU1z/JHcB9n8.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
dbit@localhost's password:
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 6.2.0-31-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status


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.

dbit@complab3:~$
```

6. Move hadoop package to usr folder

```
dbit@complab3:~$ sudo mv ~/Downloads/hadoop-2.7.7 /usr/local/hadoop
[sudo] password for dbit:
dbit@complab3:~$
```

7. Add environment variables to bashrc file and update it

```
 GNU nano 6.2                                    .bashrc *
#   sleep 10; alert
alias alert='notify-send --urgency=low -i "$([ $? = 0 ] && echo terminal || echo error)" "$(history|tail -n1|sed -e

# 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


#Hadoop variables
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
#end of Hadoop variable declaration
```

```
dbit@complab3:~$ nano .bashrc
dbit@complab3:~$ source .bashrc
dbit@complab3:~$
```

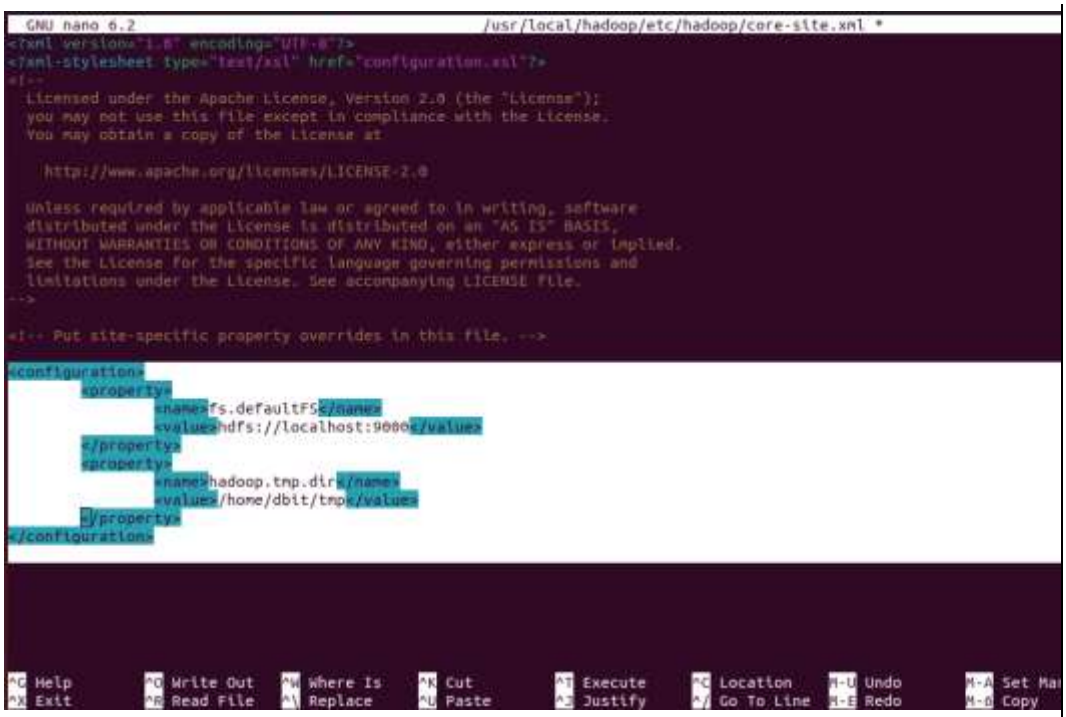
8. Check hadoop version

```
dbit@complab3:~$ hadoop version
Hadoop 2.7.7
Subversion Unknown -r c1aad84bd27cd79c3d1a7dd58202a8c3ee1ed3ac
Compiled by stevel on 2018-07-18T22:47Z
Compiled with protoc 2.5.0
From source with checksum 792e15d20b12c74bd6f19a1fb886490
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-2.7.7.jar
dbit@complab3:~$
```

9. Update hadoop environment settings

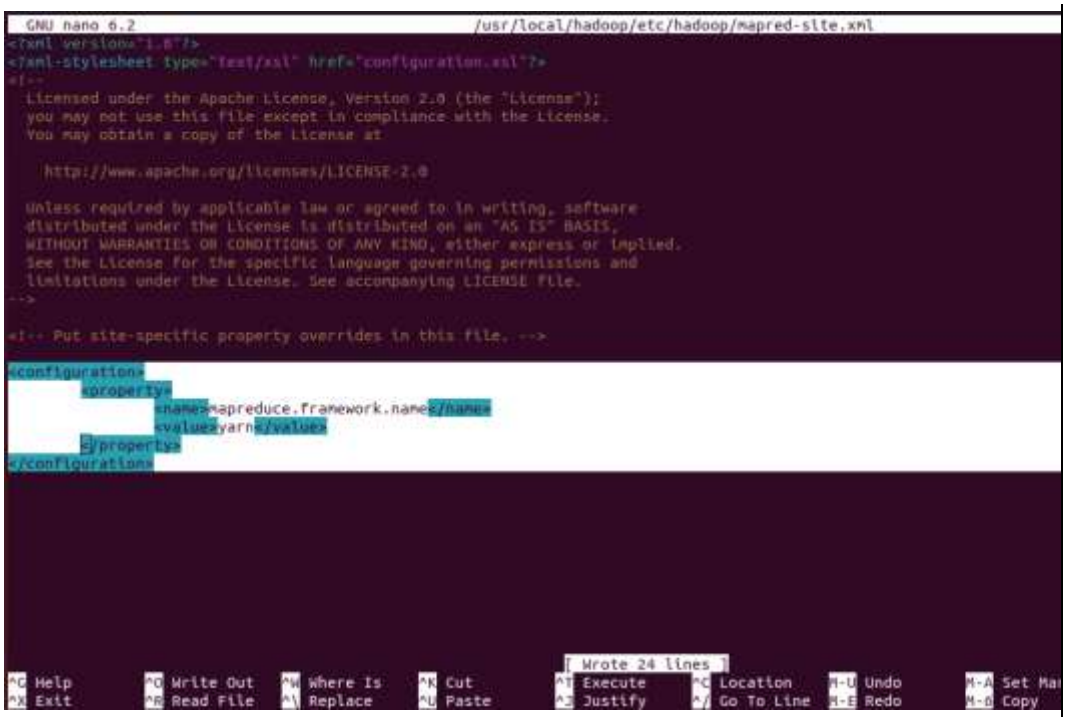
```
  GNU nano 6.2                          /usr/local/hadoop/etc/hadoop/hadoop-env.sh *
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements.  See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership.  The ASF licenses this file
# to you 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.

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME.  All others are
# optional.  When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME=usr/lib/jvm/java-1.8.0-openjdk-amd64

# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol.  Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/etc/hadoop"}

^G Help      ^O Write Out   ^W Where Is   ^K Cut       ^T Execute   ^C Location   M-U Undo      M-A Set Ma
^X Exit      ^R Read File   ^\ Replace    ^U Paste     ^J Justify   ^/ Go To Line M-E Redo      M-6 Copy
```

10. Update hadoop core file

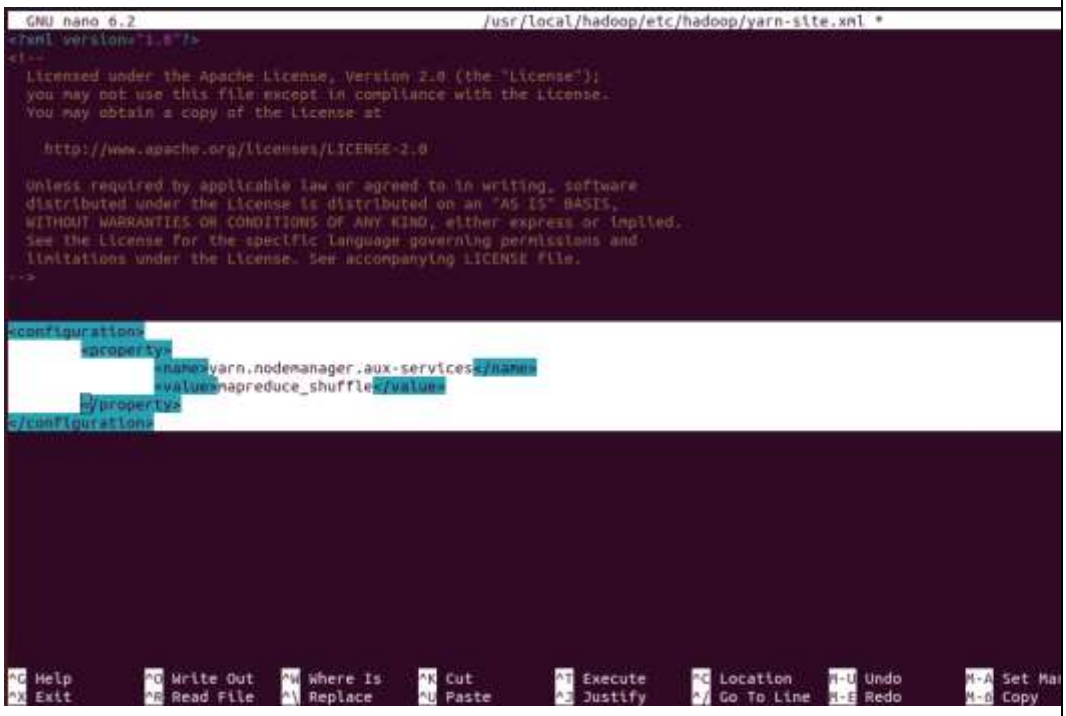


11. Update hadoop hdfs file

12. Update hadoop map reduce file



13. Update hadoop yarn file

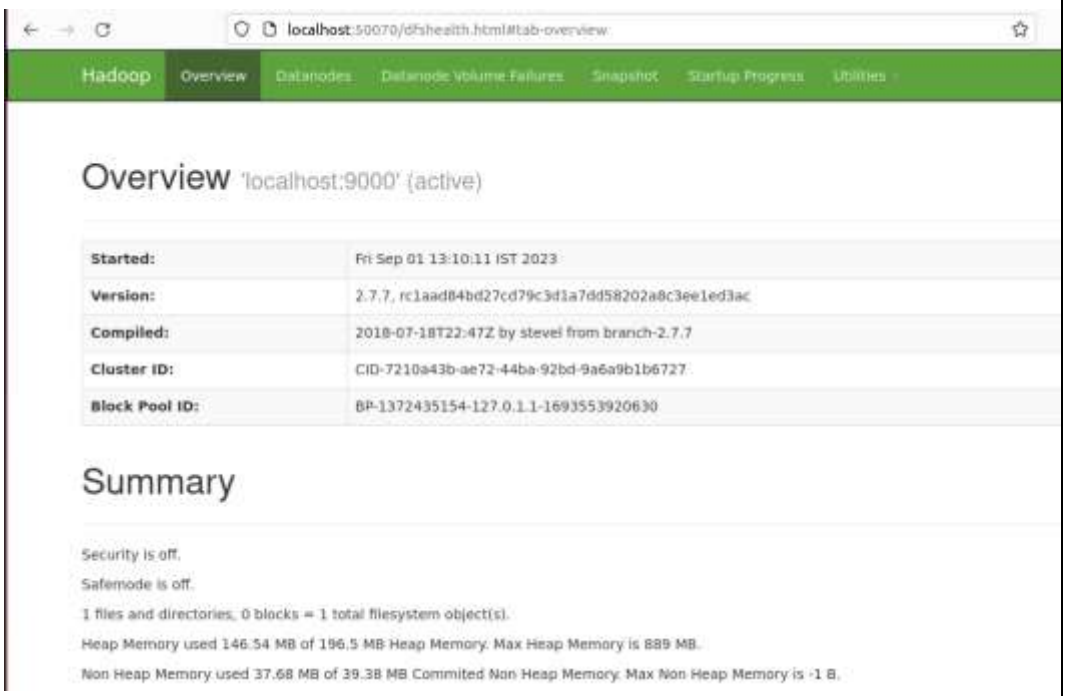


14. Starting hadoop

```
dbit@complab3:~$ start-dfs.sh
Starting namenodes on [localhost]
dbit@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-dbit-namenode-complab3.out
dbit@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-dbit-datanode-complab3.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ED25519 key fingerprint is SHA256:fEyXPWreVyvn45GMDvwmfpP6IlOJjVwbU1z/JHcB9n8.
This host key is known by the following other names/addresses:
    ~/.ssh/known_hosts:1: [hashed name]
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ED25519) to the list of known hosts.
dbit@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-dbit-secondarynamenode-complab3.out
dbit@complab3:~$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-dbit-resourcemanager-complab3.out
dbit@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-dbit-nodemanager-complab3.out
dbit@complab3:~$
```

```
dbit@complab3:~$ jps
11062 SecondaryNameNode
10742 NameNode
10889 DataNode
11660 Jps
11197 ResourceManager
11486 NodeManager
dbit@complab3:~$
```

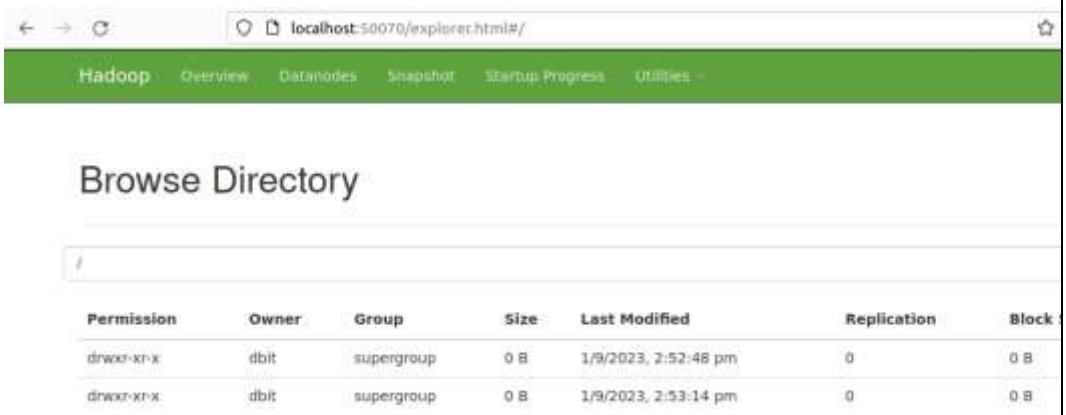
15. Hadoop dashboard



16. Basic hadoop commands

mkdir

```
dbit@complab3:~$ hadoop fs -mkdir /bda
dbit@complab3:~$ hadoop fs -mkdir /bda
mkdir: `/bda': File exists
dbit@complab3:~$ hadoop fs -mkdir /test
dbit@complab3:~$
```



copy and move files

```
dbit@complab3:~$ hadoop fs -put Downloads/ExperimentNo1.pdf /bda/
dbit@complab3:~$ hadoop fs -moveFromLocal Downloads/ExperimentNo1.pdf /test/
dbit@complab3:~$
```

ls command

```
dbit@complab3:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /test
dbit@complab3:~$ hadoop fs -ls /bda
Found 1 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /bda/ExperimentNo1.pdf
dbit@complab3:~$ hadoop fs -ls /test
Found 1 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /test/ExperimentNo1.pdf
dbit@complab3:~$
```

get command

```
dbit@complab3:~$ hadoop fs -get /test/ExperimentNo1.pdf ~/Downloads/
dbit@complab3:~$ ls ~/Downloads/
ExperimentNo1.pdf  hadoop-2.7.7.tar.gz
dbit@complab3:~$
```

touch command

```
dbit@complab3:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /test
dbit@complab3:~$ hadoop fs -touch hi.txt
-touch: Unknown command
dbit@complab3:~$ hadoop fs -touchz hi.txt
touchz: `hi.txt': No such file or directory
dbit@complab3:~$ hadoop fs -touchz /hi.txt
dbit@complab3:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
-rw-r--r--   1 dbit supergroup          0 2023-09-03 19:42 /hi.txt
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /test
dbit@complab3:~$
```

cat command

```
dbit@complab3:~$ hadoop fs -cat /test/hi1.txt
Hello
dbit@complab3:~$ hadoop fs -cat /test/hi2.txt
, World!
dbit@complab3:~$
```

cp and mv command

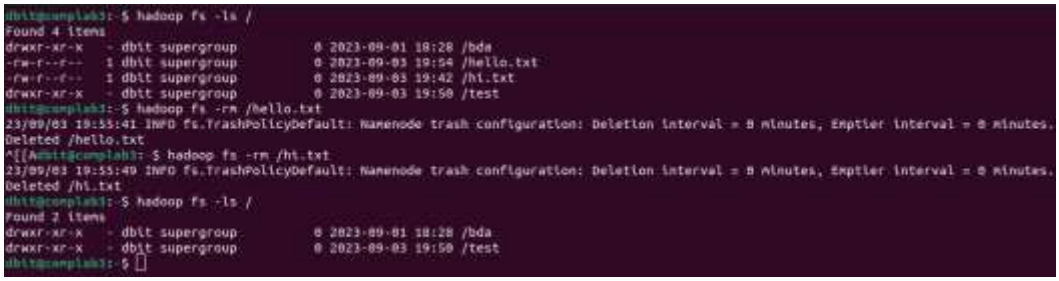
```
dbit@complab3:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:56 /hello.txt
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:50 /test
dbit@complab3:~$ hadoop fs -ls /test
Found 5 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /test/ExperimentNo1.pdf
-rw-r--r--   1 dbit supergroup          6 2023-09-03 19:50 /test/hi1.txt
-rw-r--r--   1 dbit supergroup          9 2023-09-03 19:50 /test/hi2.txt
-rw-r--r--   1 dbit supergroup         10 2023-09-03 19:50 /test/hi3.txt
-rw-r--r--   1 dbit supergroup          8 2023-09-03 19:50 /test/hi4.txt
dbit@complab3:~$ hadoop fs -ls /bda
Found 1 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /bda/ExperimentNo1.pdf
dbit@complab3:~$ hadoop fs -cp /hello.txt /bda
dbit@complab3:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:58 /bda
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:56 /hello.txt
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:50 /test
dbit@complab3:~$ hadoop fs -ls /bda
Found 2 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /bda/ExperimentNo1.pdf
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:58 /bda/hello.txt
dbit@complab3:~$ hadoop fs -mv /hello.txt /test
dbit@complab3:~$ hadoop fs -ls /test
Found 6 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /test/ExperimentNo1.pdf
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:56 /test/hello.txt
-rw-r--r--   1 dbit supergroup          6 2023-09-03 19:50 /test/hi1.txt
-rw-r--r--   1 dbit supergroup          9 2023-09-03 19:50 /test/hi2.txt
-rw-r--r--   1 dbit supergroup         10 2023-09-03 19:50 /test/hi3.txt
-rw-r--r--   1 dbit supergroup          8 2023-09-03 19:50 /test/hi4.txt
```

```
dbit@complab3:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:58 /bda
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:59 /test
dbit@complab3:~$
```

appendToFile command

```
dbit@complab3:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:50 /test
dbit@complab3:~$ hadoop fs -appendToFile hi1.txt hi2.txt hi3.txt hi4.txt /hello.txt
dbit@complab3:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:56 /hello.txt
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:50 /test
dbit@complab3:~$ hadoop fs -cat /hello.txt
Hello
, World!
 This is
Alston.
dbit@complab3:~$
```

rm command

	
Conclusion:	I was able to set up and configure Hadoop cluster in Pseudo Distributed Mode, which helps to simulate a multi node installation on a single node and also will able to work with Hadoop HDFS file system using various commands.
References:	https://data-flair.training/blogs/install-hadoop-on-ubuntu/