

Laboratory Assignment File
for

Big Data Analytics (CS0552)

Master of Technology
in
Computer Science & Engineering

Submitted by

Yugal
(Roll no. 25903053)

Submitted to

Dr. Shveta Mahajan
(Assistant Professor, NIT Jalandhar)



Department of Computer Science & Engineering
Dr. B R Ambedkar National Institute of Technology Jalandhar
Punjab, India-144008
May, 2026

Contents

Assignment 1: Install Apache Hadoop and Setup Single Node Cluster(27-01-2026)	1
1.1 Step 1: Installing Java	1
1.2 Step 2: Setup SSH	2
1.3 Step 3: Download and Configure Hadoop	3
1.4 Step 4: Environment Variable Configuration	4
1.5 Step 5: Hadoop XML Configuration Files	5
1.6 Step 6: Format NameNode & Start Hadoop Services	10
1.7 Step 7: HDFS Operations	11
1.8 Result	14
1.9 Conclusion	14
Assignment 2: Load large datasets into HDFS and analyze block distribution. (03-02-2026)	15
2.1 Step 1: Collect large Dataset (CSV/TXT)	15
2.2 Step 2: Create HDFS directory.	15
2.3 Step 3: Upload large dataset into HDFS	15
2.4 Step 4: Verify data storage	16
2.5 Step 5: Analyze block distribution	16
2.6 Step 6: Download processed dataset from HDFS to local system	16
2.7 Step 7: Delete data from HDFS	16
2.8 Step 8: Check file block information	17
2.9 Result	17
2.10 Conclusion	17

List of Figures

1	System Update Command	1
2	Java Installation	1
3	Java Version and Path Verification	2
4	Install SSH Server	2
5	Generate SSH Key Pair	3
6	Create Hadoop User	3
7	Download and Extract Hadoop	4
8	Moving Hadoop	4
9	.bashrc Modification	5
10	hadoop-env.sh Modification	6
11	core-site.xml Modification	7
12	hdfs-site.xml Modification	8
13	Creating Folder for NameNode and DataNode	8
14	mapred-site.xml Modification	9
15	yarn-site.xml Modification	9
16	Format NameNode	10
17	Start Hadoop Services	10
18	Start YARN Services	11
19	HDFS Operation 1	12
20	HDFS Operation 2	12
21	HDFS Result 1	13
22	HDFS Result 2	13
23	Start Hadoop Services and Verify with jps	15
24	Created exceldata directory in hdfs and copied dataset	15
25	Upload large dataset into HDFS	15
26	Verify data storage in HDFS	16
27	Analyze block distribution in HDFS	16
28	Download processed dataset from HDFS to local system	16
29	Delete data from HDFS	16
30	Check file block information in HDFS	17

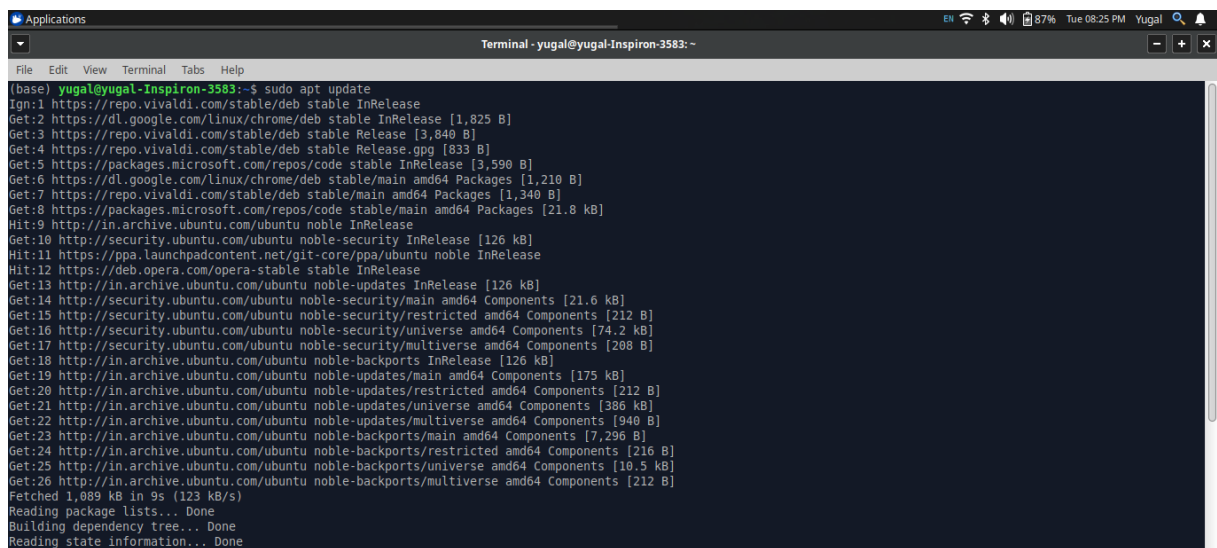
Assignment 1: Install Apache Hadoop and Setup Single Node Cluster

Objective: To install and configure Apache Hadoop in pseudo-distributed mode and perform basic HDFS operations such as upload, delete, replication check, and permission handling.

1.1 Step 1: Installing Java

Updating System

`sudo apt update`

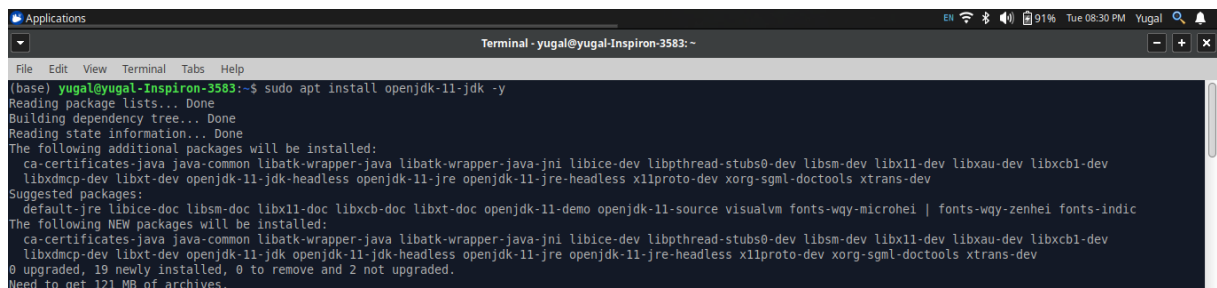


```
(base) yugal@yugal-Inspiron-3583:~$ sudo apt update
Ign:1 https://repo.vivaldi.com/stable/deb stable InRelease
Get:2 https://dl.google.com/linux/chrome/deb stable InRelease [1,825 B]
Get:3 https://repo.vivaldi.com/stable/deb stable Release [3,840 B]
Get:4 https://repo.vivaldi.com/stable/deb stable Release.gpg [833 B]
Get:5 https://packages.microsoft.com/repos/code stable InRelease [3,590 B]
Get:6 https://dl.google.com/linux/chrome/deb stable/main amd64 Packages [1,210 B]
Get:7 https://repo.vivaldi.com/stable/deb stable/main amd64 Packages [1,340 B]
Get:8 https://packages.microsoft.com/repos/code stable/main amd64 Packages [21.8 kB]
Hit:9 http://in.archive.ubuntu.com/ubuntu noble InRelease
Get:10 http://security.ubuntu.com/ubuntu noble-security InRelease [126 kB]
Hit:11 https://ppa.launchpadcontent.net/git-core/ppa/ubuntu noble InRelease
Hit:12 https://deb.opera.com/opera-stable stable InRelease
Get:13 http://in.archive.ubuntu.com/ubuntu noble-updates InRelease [126 kB]
Get:14 http://security.ubuntu.com/ubuntu noble-security/main amd64 Components [21.6 kB]
Get:15 http://security.ubuntu.com/ubuntu noble-security/restricted amd64 Components [212 B]
Get:16 http://security.ubuntu.com/ubuntu noble-security/universe amd64 Components [74.2 kB]
Get:17 http://security.ubuntu.com/ubuntu noble-security/multiverse amd64 Components [280 B]
Get:18 http://in.archive.ubuntu.com/ubuntu noble-backports InRelease [126 kB]
Get:19 http://in.archive.ubuntu.com/ubuntu noble-updates/main amd64 Components [175 kB]
Get:20 http://in.archive.ubuntu.com/ubuntu noble-updates/restricted amd64 Components [212 B]
Get:21 http://in.archive.ubuntu.com/ubuntu noble-updates/universe amd64 Components [386 kB]
Get:22 http://in.archive.ubuntu.com/ubuntu noble-updates/multiverse amd64 Components [940 B]
Get:23 http://in.archive.ubuntu.com/ubuntu noble-backports/main amd64 Components [7,296 B]
Get:24 http://in.archive.ubuntu.com/ubuntu noble-backports/restricted amd64 Components [216 B]
Get:25 http://in.archive.ubuntu.com/ubuntu noble-backports/universe amd64 Components [10.5 kB]
Get:26 http://in.archive.ubuntu.com/ubuntu noble-backports/multiverse amd64 Components [212 B]
Fetched 1,089 kB in 9s (123 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
```

Figure 1: System Update Command

Installing Java

`sudo apt install openjdk-11-jdk -y`

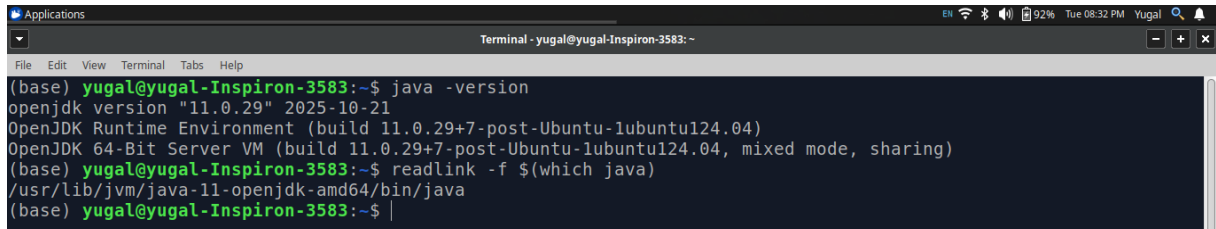


```
(base) yugal@yugal-Inspiron-3583:~$ sudo apt install openjdk-11-jdk -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java java-common libatk-wrapper-java libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev
  libxdmcp-dev libxt-dev openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
  default-jre libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-11-demo openjdk-11-source visualvm fonts-wqy-microhei | fonts-wqy-zenhei fonts-indic
The following NEW packages will be installed:
  ca-certificates-java java-common libatk-wrapper-java libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev
  libxdmcp-dev libxt-dev openjdk-11-jdk openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 19 newly installed, 0 to remove and 2 not upgraded.
Need to get 121 MB of archives.
After this operation, 479 MB of additional disk space will be used.
```

Figure 2: Java Installation

Java Version and Path Verification

```
java -version
readlink -f $(which java)
```

A terminal window titled 'Terminal - yugal@yugal-Inspiron-3583: ~' showing the output of 'java -version' and 'readlink -f \$(which java)'. The output shows OpenJDK version 11.0.29 and the path /usr/lib/jvm/java-11-openjdk-amd64/bin/java.

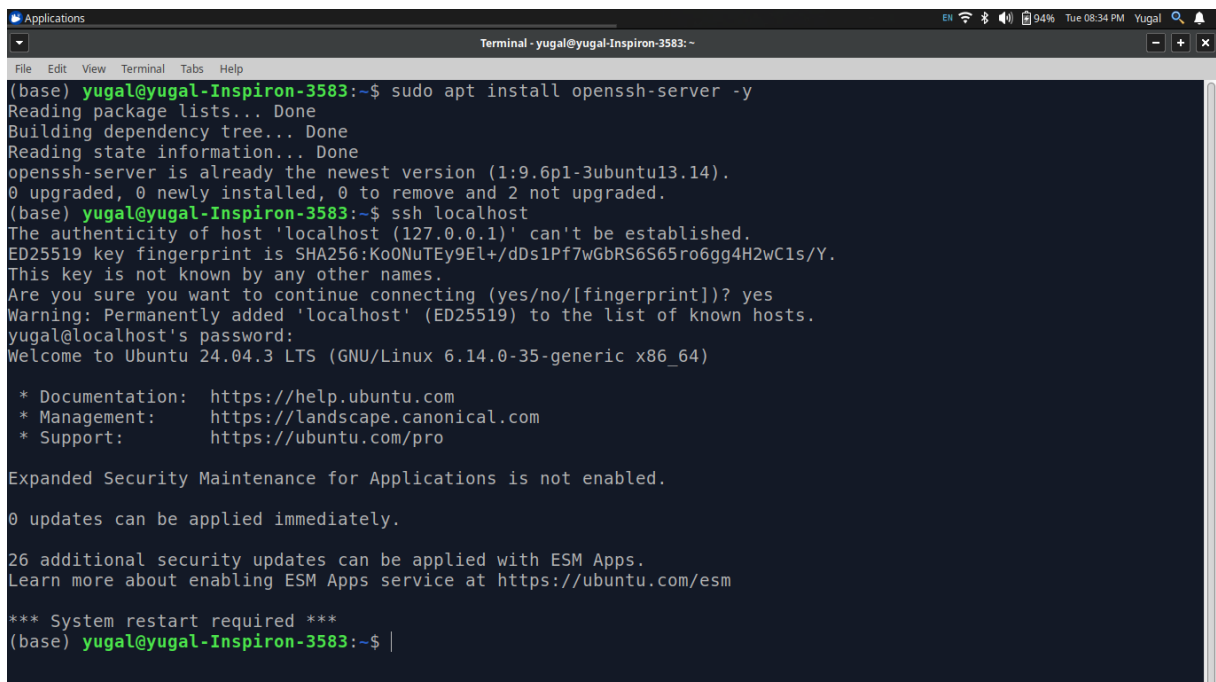
```
(base) yugal@yugal-Inspiron-3583:~$ java -version
openjdk version "11.0.29" 2025-10-21
OpenJDK Runtime Environment (build 11.0.29+7-post-Ubuntu-1ubuntu124.04)
OpenJDK 64-Bit Server VM (build 11.0.29+7-post-Ubuntu-1ubuntu124.04, mixed mode, sharing)
(base) yugal@yugal-Inspiron-3583:~$ readlink -f $(which java)
/usr/lib/jvm/java-11-openjdk-amd64/bin/java
(base) yugal@yugal-Inspiron-3583:~$
```

Figure 3: Java Version and Path Verification

1.2 Step 2: Setup SSH

Install SSH Server

```
sudo apt install openssh-server -y
ssh localhost
```

A terminal window titled 'Terminal - yugal@yugal-Inspiron-3583: ~' showing the installation of openssh-server and a successful connection to localhost. The output includes package lists, dependency tree, and a warning about permanently adding 'localhost' to the list of known hosts.

```
(base) yugal@yugal-Inspiron-3583:~$ sudo apt install openssh-server -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openssh-server is already the newest version (1:9.6p1-3ubuntu13.14).
0 upgraded, 0 newly installed, 0 to remove and 2 not upgraded.
(base) yugal@yugal-Inspiron-3583:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:Ko0NuTEy9EL+/dDs1Pf7wGbRS6S6Sro6gg4H2wC1s/Y.
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.
yugal@localhost's password:
Welcome to Ubuntu 24.04.3 LTS (GNU/Linux 6.14.0-35-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

26 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm

*** System restart required ***
(base) yugal@yugal-Inspiron-3583:~$
```

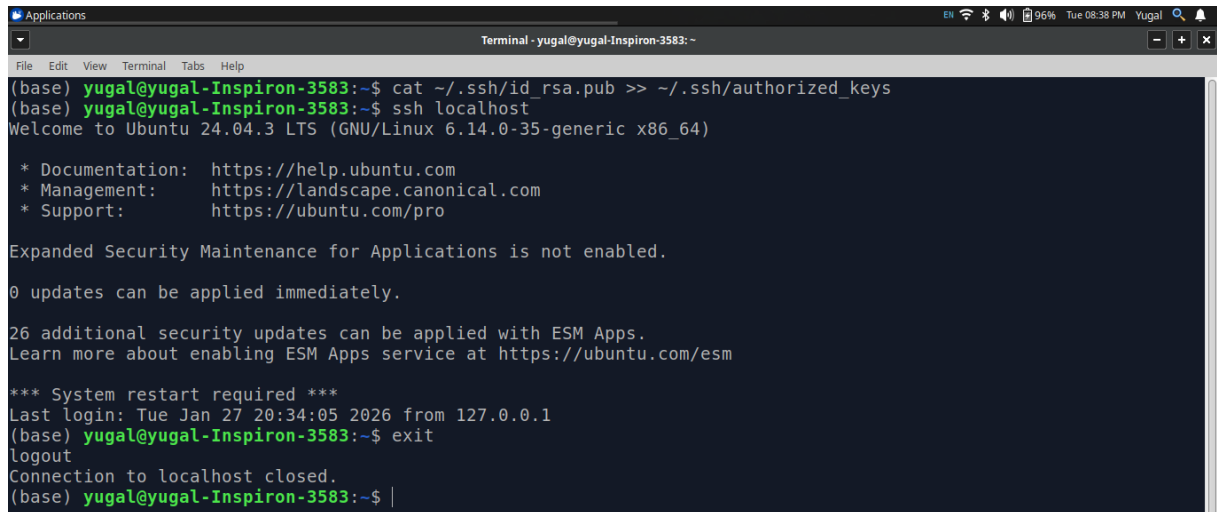
Figure 4: Install SSH Server

Generate SSH Key Pair

```
ssh-keygen -t rsa -P ""
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

```
ssh localhost
```

```
exit
```

A terminal window titled 'Terminal - yugal@yugal-Inspiron-3583: ~' showing the process of generating an SSH key pair and connecting to localhost. The user runs 'cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys', then 'ssh localhost'. The terminal displays the Ubuntu 24.04.3 LTS login banner, system updates information, and the user's login details. Finally, the user runs 'exit' and the connection to localhost is closed.

```
(base) yugal@yugal-Inspiron-3583:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
(base) yugal@yugal-Inspiron-3583:~$ ssh localhost
Welcome to Ubuntu 24.04.3 LTS (GNU/Linux 6.14.0-35-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

26 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm

*** System restart required ***
Last login: Tue Jan 27 20:34:05 2026 from 127.0.0.1
(base) yugal@yugal-Inspiron-3583:~$ exit
logout
Connection to localhost closed.
(base) yugal@yugal-Inspiron-3583:~$ |
```

Figure 5: Generate SSH Key Pair

1.3 Step 3: Download and Configure Hadoop

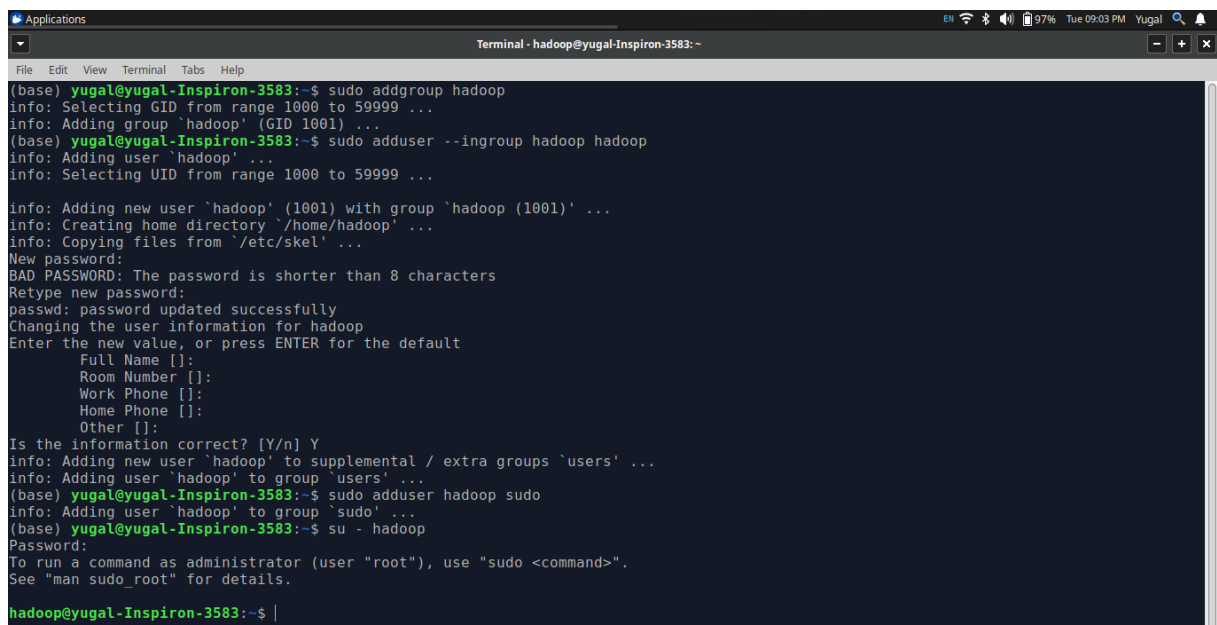
Create Hadoop User

```
sudo addgroup hadoop
```

```
sudo adduser --ingroup hadoop hadoop
```

```
sudo adduser hadoop sudo
```

```
su - hadoop
```

A terminal window titled 'Terminal - hadoop@yugal-Inspiron-3583: ~' showing the steps to create a Hadoop user. The user runs 'sudo addgroup hadoop' and 'sudo adduser --ingroup hadoop hadoop'. The terminal shows the group being added, then the user being added with a password. The user then runs 'su - hadoop' to switch to the hadoop user.

```
(base) yugal@yugal-Inspiron-3583:~$ sudo addgroup hadoop
info: Selecting GID from range 1000 to 59999 ...
info: Adding group `hadoop' (GID 1001) ...
(base) yugal@yugal-Inspiron-3583:~$ sudo adduser --ingroup hadoop hadoop
info: Adding user `hadoop' ...
info: Selecting UID from range 1000 to 59999 ...

info: Adding new user `hadoop' (1001) with group `hadoop (1001)' ...
info: Creating home directory `/home/hadoop' ...
info: Copying files from `/etc/skel' ...
New password:
BAD PASSWORD: The password is shorter than 8 characters
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
info: Adding new user `hadoop' to supplemental / extra groups `users' ...
info: Adding user `hadoop' to group `users' ...
(base) yugal@yugal-Inspiron-3583:~$ sudo adduser hadoop sudo
info: Adding user `hadoop' to group `sudo' ...
(base) yugal@yugal-Inspiron-3583:~$ su - hadoop
Password:
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

hadoop@yugal-Inspiron-3583:~$ |
```

Figure 6: Create Hadoop User

Download and Extract Hadoop

exit

```
sudo cp /home/yugal/Downloads/hadoop-3.3.0.tar.gz /home/hadoop/
```

```
sudo chown hadoop:hadoop /home/hadoop/hadoop-3.3.0.tar.gz
```

```
su - hadoop
```

```
ls
```

```
tar -xvzf hadoop-3.3.0.tar.gz
```

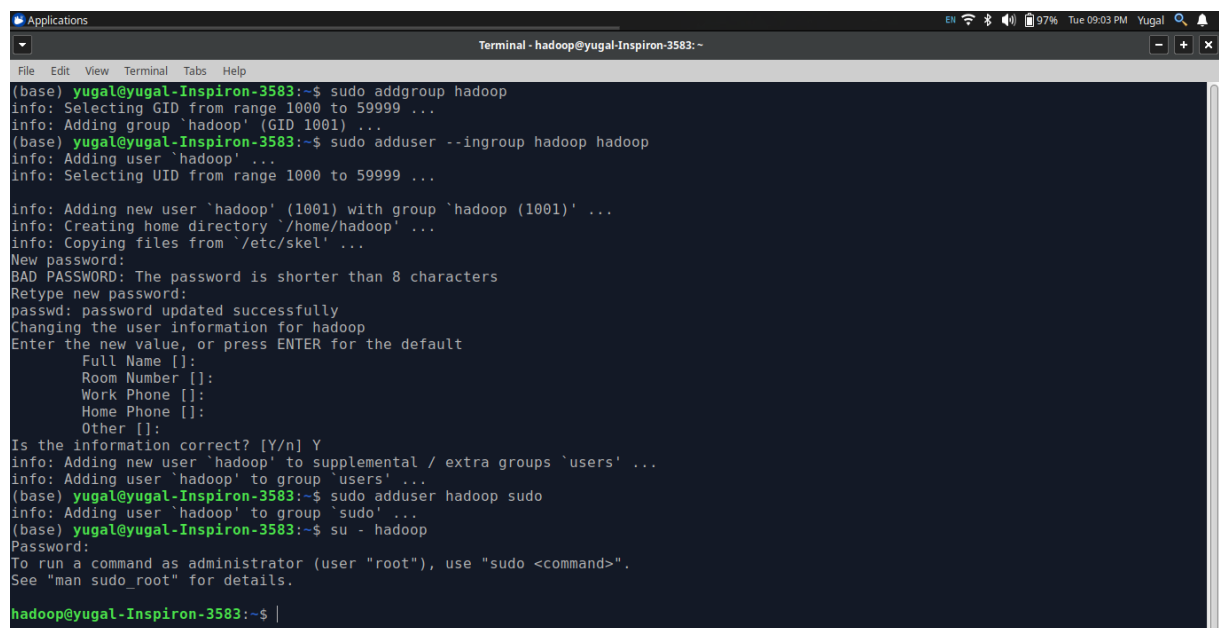
A terminal window titled 'Terminal - hadoop@yugal-Inspiron-3583: ~' showing the process of creating a new user and group. The user 'yugal@yugal-Inspiron-3583' runs 'sudo addgroup hadoop', which selects GID 1001. Then, 'sudo adduser --ingroup hadoop hadoop' is run, which prompts for a password (it fails on being too short, then succeeds on a second attempt), sets the home directory to /home/hadoop, and copies files from /etc/skel. The user is then prompted for additional information (Full Name, Room Number, Work Phone, Home Phone, Other) and confirms the information. Finally, 'sudo adduser hadoop sudo' is run to add the user to the sudo group, and 'su - hadoop' is used to switch to the hadoop user. The prompt changes to 'hadoop@yugal-Inspiron-3583:~\$'.

Figure 7: Download and Extract Hadoop

Moving Hadoop

```
mv hadoop-3.3.0 hadoop
```

```
ls
```

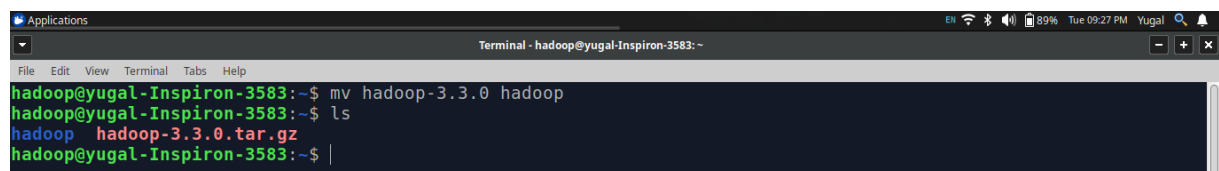
A terminal window titled 'Terminal - hadoop@yugal-Inspiron-3583: ~' showing the user 'hadoop@yugal-Inspiron-3583' running 'mv hadoop-3.3.0 hadoop'. The prompt changes to 'hadoop@yugal-Inspiron-3583:~\$'. Then, 'ls' is run, showing the directory listing: 'hadoop hadoop-3.3.0.tar.gz'. The prompt returns to 'hadoop@yugal-Inspiron-3583:~\$'.

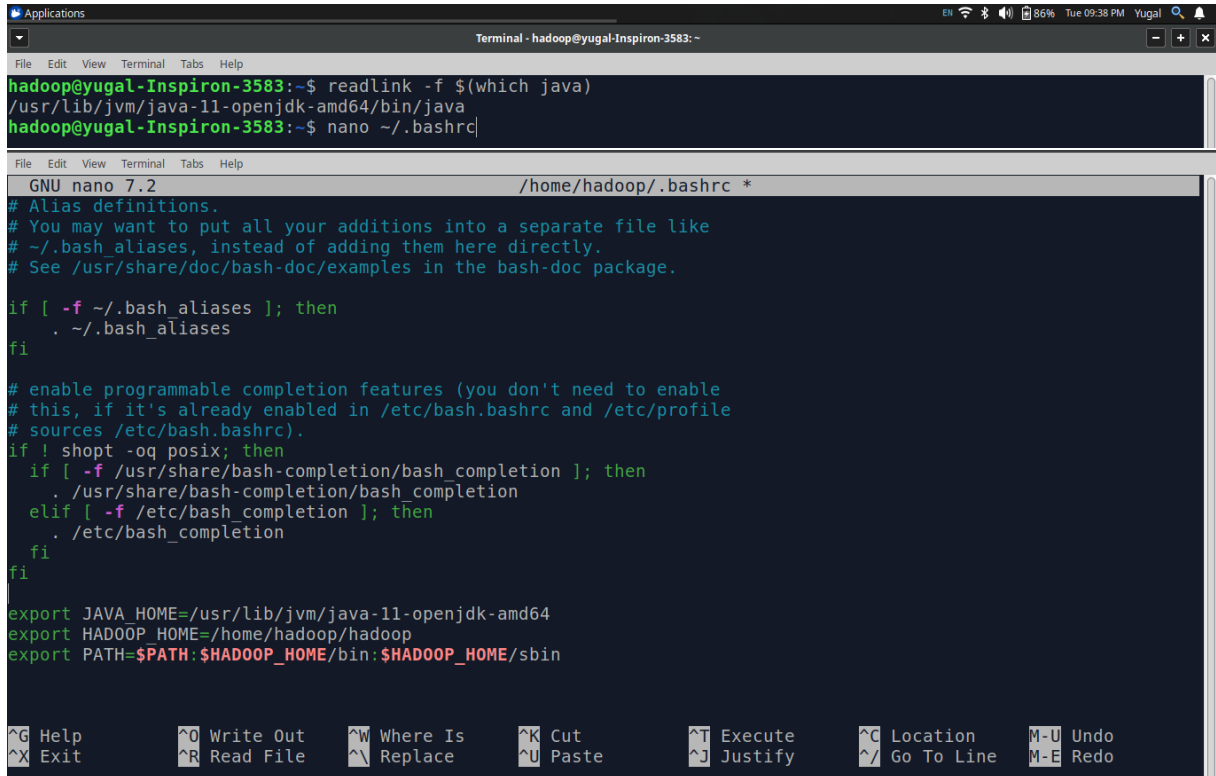
Figure 8: Moving Hadoop

1.4 Step 4: Environment Variable Configuration

.bashrc Modification

```
nano ~/.bashrc
```

```
% Add the following lines at the end of the file
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=/home/hadoop/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```



```
hadoop@yugal-Inspiron-3583:~$ readlink -f $(which java)
/usr/lib/jvm/java-11-openjdk-amd64/bin/java
hadoop@yugal-Inspiron-3583:~$ nano ~/.bashrc

GNU nano 7.2 /home/hadoop/.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

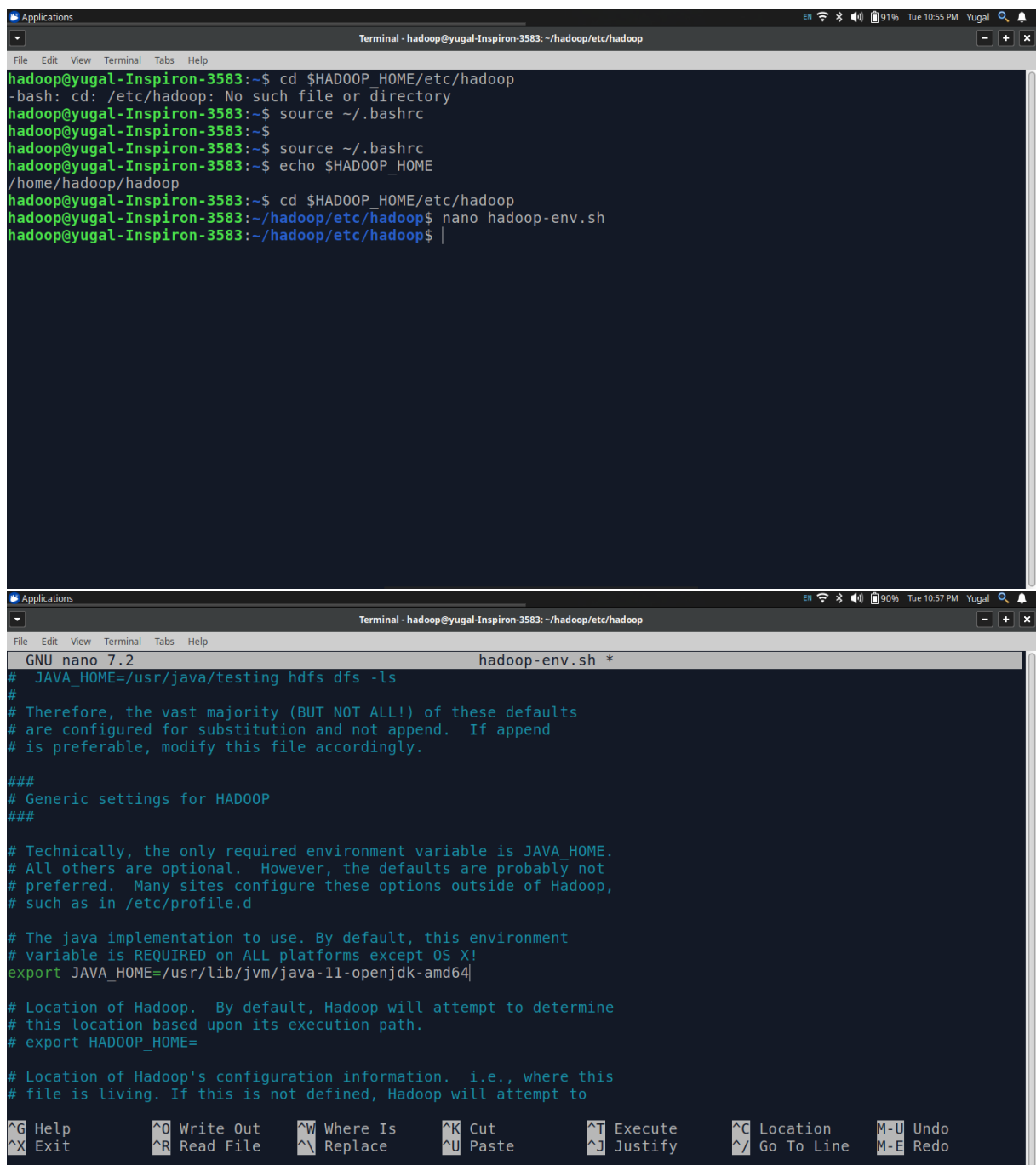
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=/home/hadoop/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

Figure 9: .bashrc Modification

1.5 Step 5: Hadoop XML Configuration Files

hadoop-env.sh Modification

```
source ~/.bashrc
echo $HADOOP_HOME
/home/hadoop/hadoop
cd $HADOOP_HOME/etc/hadoop
ls
nano hadoop-env.sh
% Add the following lines at the path
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```



```
hadoop@yugal-Inspiron-3583:~$ cd $HADOOP_HOME/etc/hadoop
-bash: cd: /etc/hadoop: No such file or directory
hadoop@yugal-Inspiron-3583:~$ source ~/.bashrc
hadoop@yugal-Inspiron-3583:~$ source ~/.bashrc
hadoop@yugal-Inspiron-3583:~$ echo $HADOOP_HOME
/home/hadoop/hadoop
hadoop@yugal-Inspiron-3583:~$ cd $HADOOP_HOME/etc/hadoop
hadoop@yugal-Inspiron-3583:~/hadoop/etc/hadoop$ nano hadoop-env.sh
hadoop@yugal-Inspiron-3583:~/hadoop/etc/hadoop$

GNU nano 7.2 hadoop-env.sh *
# JAVA_HOME=/usr/java/testing hdfs dfs -ls
#
# 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,
# 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=
#
# Location of Hadoop's configuration information. i.e., where this
# file is living. If this is not defined, Hadoop will attempt to
^G Help      ^O Write Out  ^W Where Is   ^K Cut        ^T Execute    ^C Location   M-U Undo
^X Exit      ^R Read File  ^\ Replace    ^U Paste      ^J Justify    ^_ Go To Line  M-E Redo
```

Figure 10: hadoop-env.sh Modification

core-site.xml Modification

nano core-site.xml

% Add the following property inside <configuration> tag

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

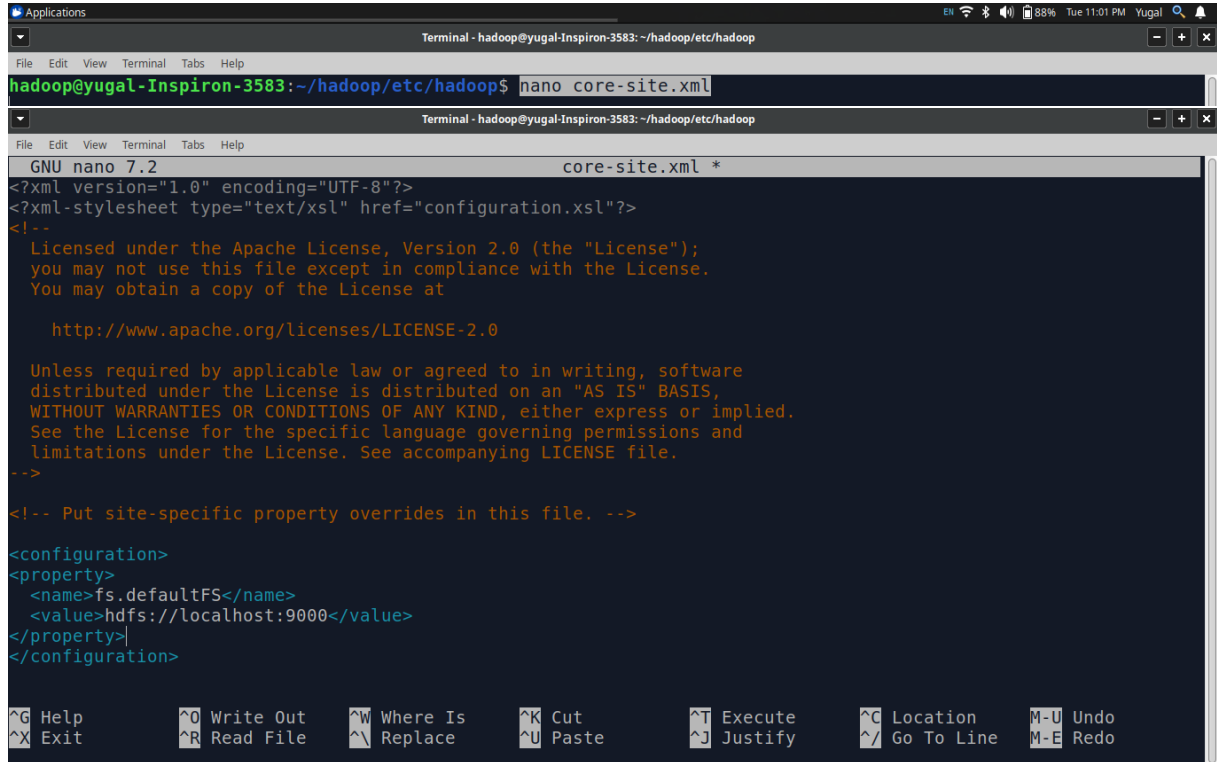


Figure 11: core-site.xml Modification

hdfs-site.xml Modification

nano hdfs-site.xml

% Add the following property inside <configuration> tag

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:///home/hadoop/hdfs/namenode</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>file:///home/hadoop/hdfs/datanode</value>

</property>

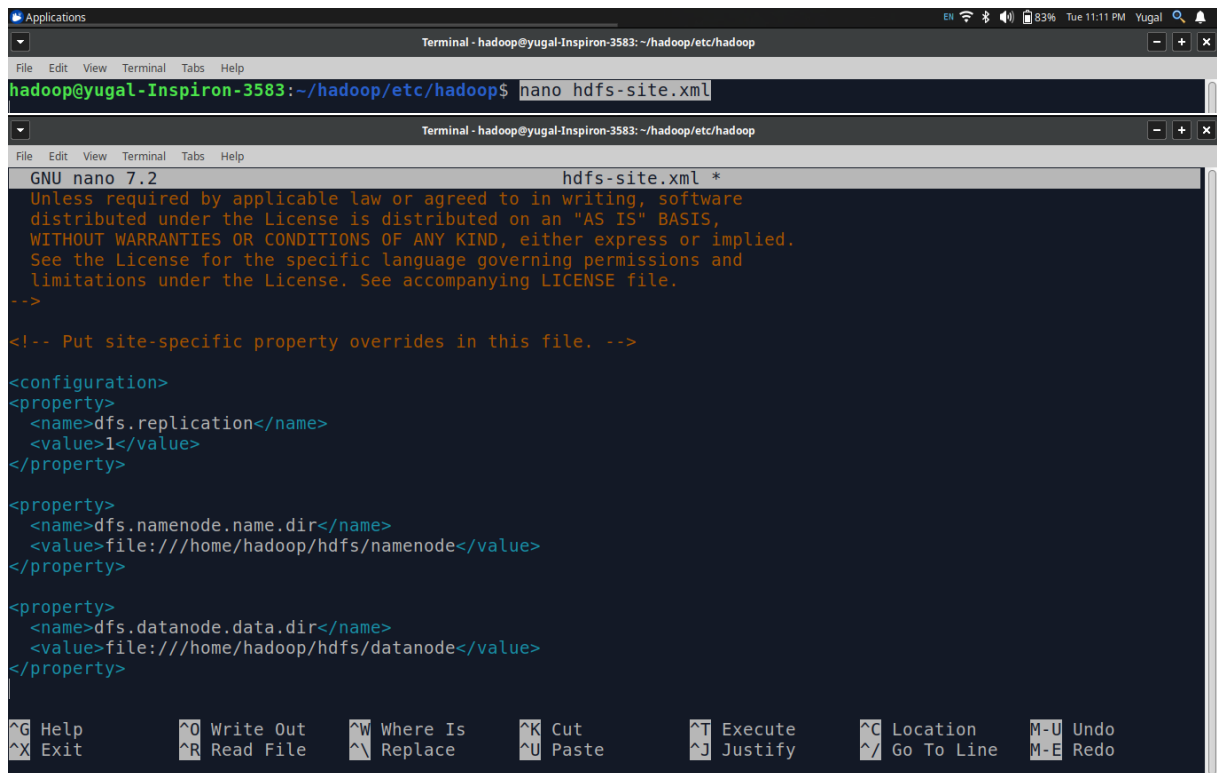


Figure 12: hdfs-site.xml Modification

Creating Folder for NameNode and DataNode

```
mkdir -p ~/hdfs/namenode
mkdir -p ~/hdfs/datanode
```

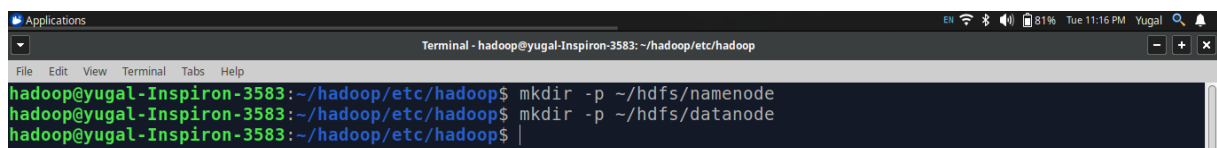
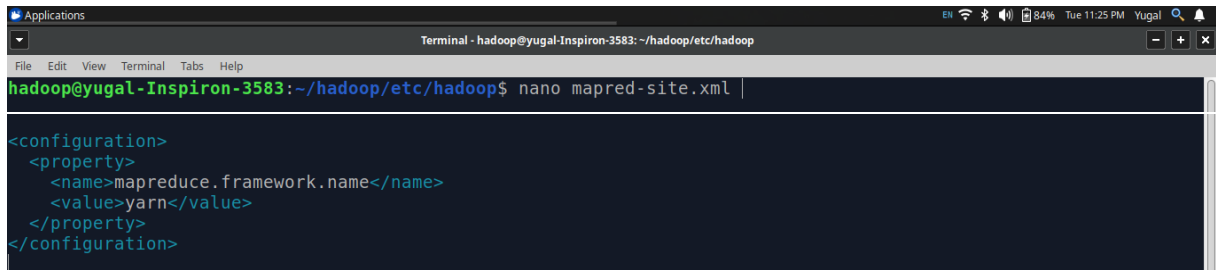


Figure 13: Creating Folder for NameNode and DataNode

mapred-site.xml Modification

```
nano mapred-site.xml
% Add the following property inside <configuration> tag
<property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
</property>
```



```
Applications
Terminal - hadoop@yugal-Inspiron-3583: ~/hadoop/etc/hadoop
File Edit View Terminal Tabs Help
hadoop@yugal-Inspiron-3583:~/hadoop/etc/hadoop$ nano mapred-site.xml

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

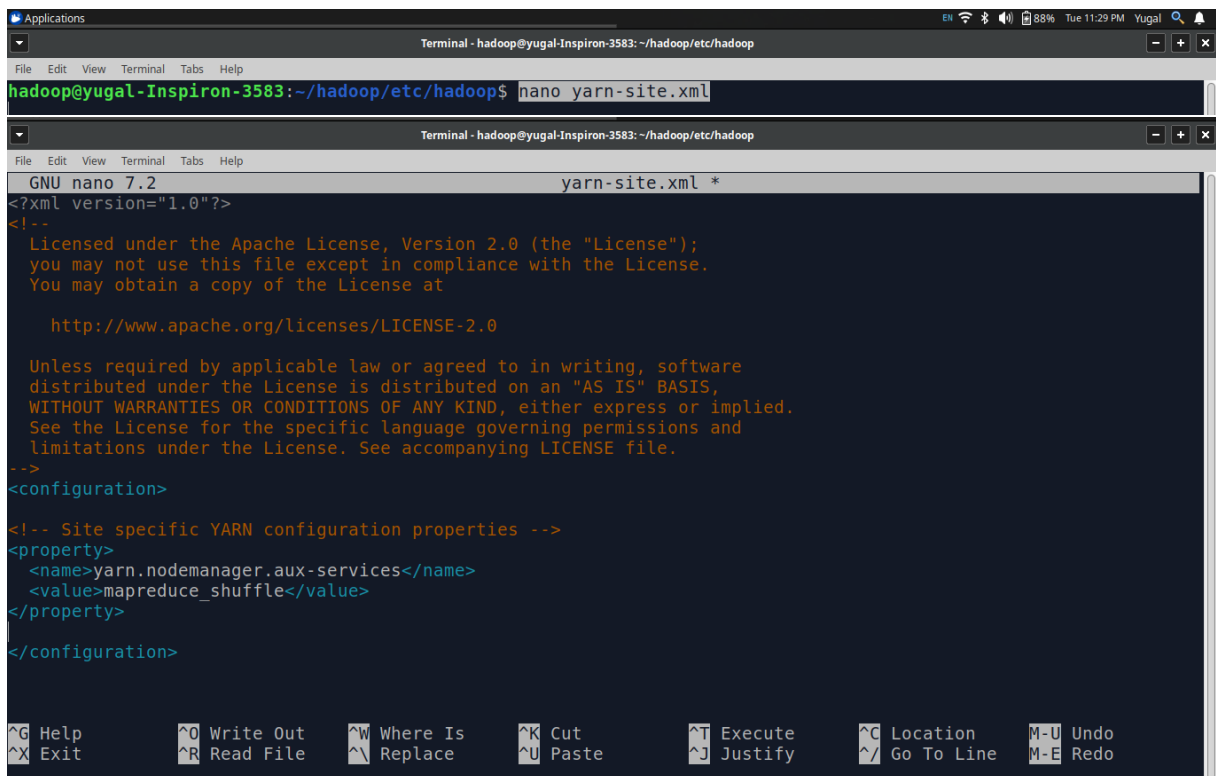
Figure 14: mapred-site.xml Modification

yarn-site.xml Modification

nano yarn-site.xml

% Add the following property inside <configuration> tag

```
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
```



```
Applications
Terminal - hadoop@yugal-Inspiron-3583: ~/hadoop/etc/hadoop
File Edit View Terminal Tabs Help
hadoop@yugal-Inspiron-3583:~/hadoop/etc/hadoop$ nano yarn-site.xml

GNU nano 7.2 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>

<!-- Site specific YARN configuration properties -->
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
</configuration>

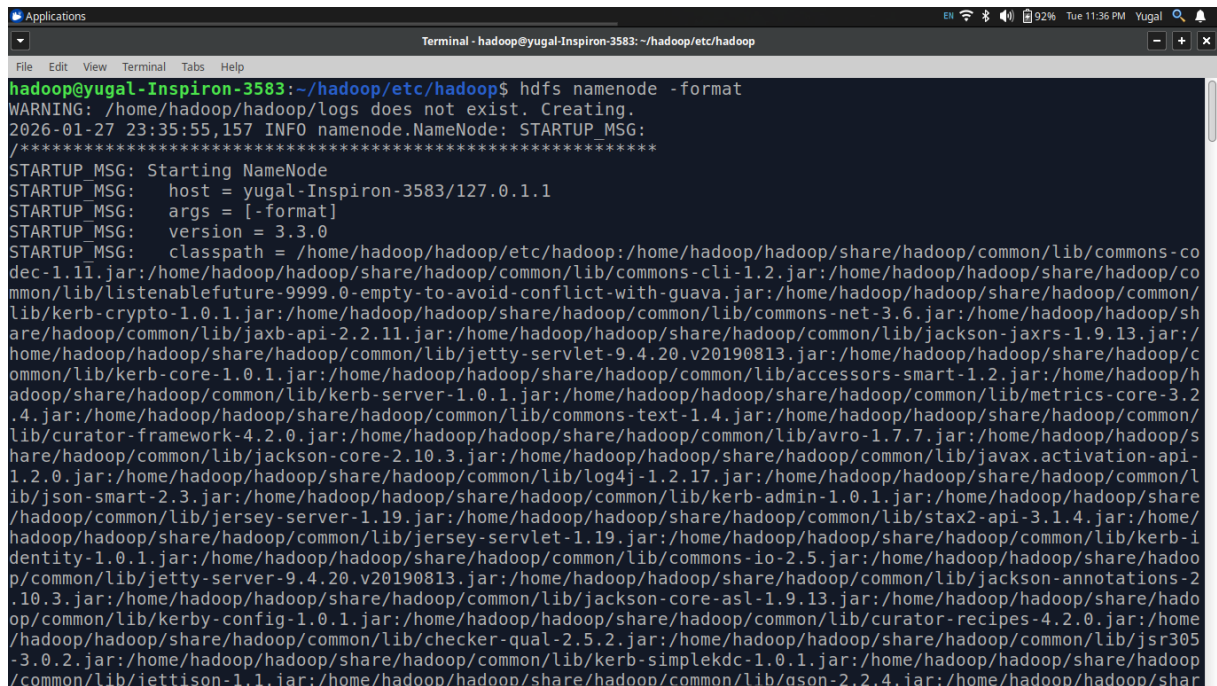
^G Help      ^O Write Out  ^W Where Is   ^K Cut        ^T Execute    ^C Location   M-U Undo
^X Exit      ^R Read File  ^_ Replace    ^U Paste      ^J Justify    ^_ Go To Line M-E Redo
```

Figure 15: yarn-site.xml Modification

1.6 Step 6: Format NameNode & Start Hadoop Services

Format NameNode

```
hdfs namenode -format
```



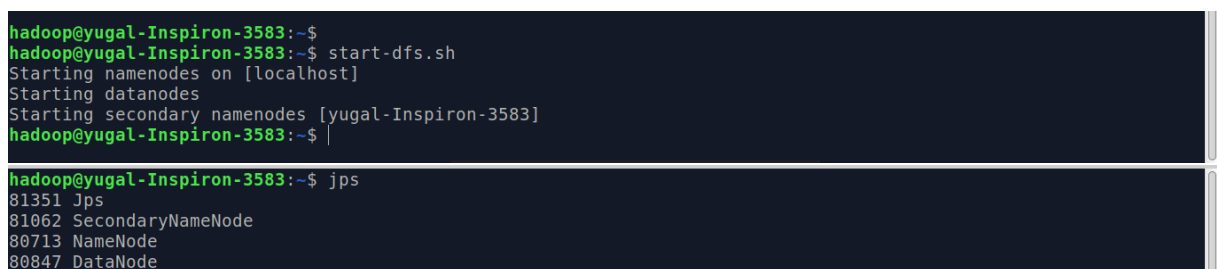
```
Applications
Terminal - hadoop@yugal-Inspiron-3583: ~/hadoop/etc/hadoop
File Edit View Terminal Tabs Help
hadoop@yugal-Inspiron-3583:~/hadoop/etc/hadoop$ hdfs namenode -format
WARNING: /home/hadoop/hadoop/logs does not exist. Creating.
2026-01-27 23:35:55,157 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = yugal-Inspiron-3583/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.0
STARTUP_MSG:   classpath = /home/hadoop/hadoop/etc/hadoop:/home/hadoop/hadoop/share/hadoop/common/lib/commons-codec-1.11.jar:/home/hadoop/hadoop/share/hadoop/common/lib/commons-cli-1.2.jar:/home/hadoop/hadoop/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerb-crypto-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/commons-net-3.6.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jaxb-api-2.2.11.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jackson-jaxrs-1.9.13.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jetty-servlet-9.4.20.v20190813.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerb-core-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/accessors-smart-1.2.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerb-server-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/metrics-core-3.2.4.jar:/home/hadoop/hadoop/share/hadoop/common/lib/commons-text-1.4.jar:/home/hadoop/hadoop/share/hadoop/common/lib/curator-framework-4.2.0.jar:/home/hadoop/hadoop/share/hadoop/common/lib/avro-1.7.7.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jackson-core-2.10.3.jar:/home/hadoop/hadoop/share/hadoop/common/lib/javax.activation-api-1.2.0.jar:/home/hadoop/hadoop/share/hadoop/common/lib/log4j-1.2.17.jar:/home/hadoop/hadoop/share/hadoop/common/lib/json-smart-2.3.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerb-admin-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jersey-server-1.19.jar:/home/hadoop/hadoop/share/hadoop/common/lib/stax2-api-3.1.4.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jersey-servlet-1.19.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerb-identity-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/commons-io-2.5.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jetty-server-9.4.20.v20190813.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jackson-annotations-2.10.3.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerby-config-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/curator-recipes-4.2.0.jar:/home/hadoop/hadoop/share/hadoop/common/lib/checker-qual-2.5.2.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jsr305-3.0.2.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerb-simplekdc-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jettison-1.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/gson-2.2.4.jar:/home/hadoop/hadoop/shar
```

Figure 16: Format NameNode

Start Hadoop Services

```
start-dfs.sh
```

```
jps
```



```
hadoop@yugal-Inspiron-3583:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [yugal-Inspiron-3583]
hadoop@yugal-Inspiron-3583:~$ jps
81351 Jps
81062 SecondaryNameNode
80713 NameNode
80847 DataNode
```

Figure 17: Start Hadoop Services

Start YARN Services

```
start-yarn.sh
```

```
jps
```

The image shows a terminal window and a web browser. The terminal window displays the following commands and output:

```
hadoop@yugal-Inspiron-3583:~$ jps
81351 Jps
81062 SecondaryNameNode
80713 NameNode
80847 DataNode
hadoop@yugal-Inspiron-3583:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@yugal-Inspiron-3583:~$ jps
81792 NodeManager
82355 Jps
81670 ResourceManager
81062 SecondaryNameNode
80713 NameNode
80847 DataNode
hadoop@yugal-Inspiron-3583:~$ |
```

The web browser shows the Hadoop Overview page for 'localhost:9000' (active). The page includes a table with the following information:

Started:	Tue Jan 27 23:49:34 +0530 2026
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled:	Tue Jul 07 00:14:00 +0530 2020 by brahma from branch-3.3.0
Cluster ID:	CID-7d9aa263-5957-4992-9da1-32c648d8341c
Block Pool ID:	BP-1685072063-127.0.1.1-1769537156943

The Summary section shows the following information:

- 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 119.44 MB of 616 MB Heap Memory. Max Heap Memory is 3.88 GB.
- Non Heap Memory used 53.86 MB of 56.5 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Figure 18: Start YARN Services

1.7 Step 7: HDFS Operations

```
% Create a test file
echo "Hello Big Data" > file.txt
% Make directory in HDFS
hdfs dfs -mkdir /data
```

```

% Upload file to HDFS
hdfs dfs -put file.txt /data
% List files
hdfs dfs -ls /data
% Check blocks & replication
hdfs fsck /data/file.txt -files -blocks
% Change permission
hdfs dfs -chmod 777 /data/file.txt
% Delete file
hdfs dfs -rm /data/file.txt
hdfs dfs -ls /data

```

```

Applications
Terminal - hadoop@yugal-Inspiron-3583: ~
File Edit View Terminal Tabs Help
hadoop@yugal-Inspiron-3583:~$ echo "Hello Big Data" > file.txt
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -mkdir /data
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -put file.txt /data
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -ls /data
Found 1 items
-rw-r--r-- 1 hadoop supergroup      15 2026-01-28 00:50 /data/file.txt
hadoop@yugal-Inspiron-3583:~$ hdfs fsck /data/file.txt -files -blocks
Connecting to namenode via http://localhost:9870/fsck?ugi=hadoop&files=1&blocks=1&path=%2Fdata%2Ffile.txt
FSCK started by hadoop (auth:SIMPLE) from /127.0.0.1 for path /data/file.txt at Wed Jan 28 00:51:19 IST 2026

/data/file.txt 15 bytes, replicated: replication=1, 1 block(s): OK
0. BP-1685072063-127.0.1.1-1769537156943:blk_1073741825_1001 len=15 Live_repl=1

Status: HEALTHY
Number of data-nodes: 1
Number of racks:      1
Total dirs:           0
Total symlinks:       0

```

Figure 19: HDFS Operation 1

```

Erasure Coded Block Groups:
Total size: 0 B
Total files: 0
Total block groups (validated): 0
Minimally erasure-coded block groups: 0
Over-erasure-coded block groups: 0
Under-erasure-coded block groups: 0
Unsatisfactory placement block groups: 0
Average block group size: 0.0
Missing block groups: 0
Corrupt block groups: 0
Missing internal blocks: 0
Blocks queued for replication: 0
FSCK ended at Wed Jan 28 00:51:19 IST 2026 in 37 milliseconds

The filesystem under path '/data/file.txt' is HEALTHY
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -chmod 777 /data/file/txt
chmod: '/data/file/txt': No such file or directory
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -rm /data/file.txt

```

Figure 20: HDFS Operation 2

Browse Directory

/data

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rwxrwxrwx	hadoop	supergroup	15 B	Jan 28 00:50	1	128 MB	file.txt

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2020.

Figure 21: HDFS Result 1

hadoop SUBMITTED Applications

Cluster

About Nodes Node Labels Applications NEW NEW_SAVING SUBMITTED ACCEPTED RUNNING FINISHED FAILED KILLED Scheduler Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used
0	0	0	0	0	0 B

Cluster Nodes Metrics

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

Scheduler Metrics

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

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus
No data available in table											

Showing 0 to 0 of 0 entries

Figure 22: HDFS Result 2

1.8 Result

Hadoop was successfully installed and configured in pseudo-distributed mode, and various HDFS operations were performed successfully.

1.9 Conclusion

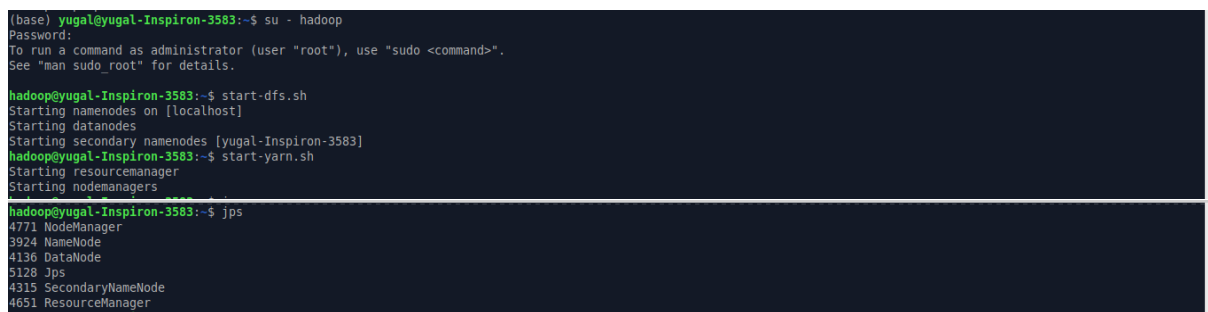
This experiment provided hands-on experience in setting up a Big Data environment using Hadoop and understanding the working of HDFS through practical command execution.

Assignment 2: Load large datasets into HDFS and analyze block distribution.

Objective: Understand how Apache Hadoop HDFS stores large files by splitting them into fixed-size blocks and placing them on DataNodes.

2.1 Step 1: Collect large Dataset (CSV/TXT)

```
su - hadoop
start-dfs.sh
start-yarn.sh
jps
```



```
(base) yugal@yugal-Inspiron-3583:~$ su - hadoop
Password:
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

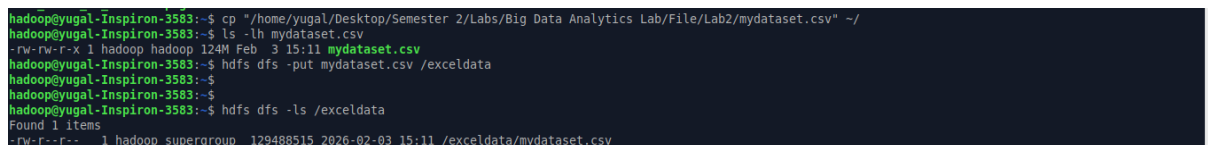
hadoop@yugal-Inspiron-3583:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [yugal-Inspiron-3583]
hadoop@yugal-Inspiron-3583:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers

hadoop@yugal-Inspiron-3583:~$ jps
4771 NodeManager
3924 NameNode
4136 DataNode
5128 Jps
4315 SecondaryNameNode
4651 ResourceManager
```

Figure 23: Start Hadoop Services and Verify with jps

2.2 Step 2: Create HDFS directory.

```
% Created 352 MB csv file using:excel and uploaded to hdfs:
% Created exceldata directory in hdfs:
hdfs dfs -mkdir /exceldata
ls -lh mydataset.csv
hdfs dfs -put mydataset.csv /exceldata
hdfs dfs -ls /exceldata
```



```
hadoop@yugal-Inspiron-3583:~$ cp "/home/yugal/Desktop/Semester 2/Labs/Big Data Analytics Lab/File/Lab2/mydataset.csv" ~/
hadoop@yugal-Inspiron-3583:~$ ls -lh mydataset.csv
-rw-rw-r--x 1 hadoop hadoop 124M Feb  3 15:11 mydataset.csv
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -put mydataset.csv /exceldata
hadoop@yugal-Inspiron-3583:~$
hadoop@yugal-Inspiron-3583:~$
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -ls /exceldata
Found 1 items
-rw-r--r-- 1 hadoop supergroup 129488515 2026-02-03 15:11 /exceldata/mydataset.csv
```

Figure 24: Created exceldata directory in hdfs and copied dataset

2.3 Step 3: Upload large dataset into HDFS

```
hdfs dfs -put mydataset.csv /exceldata
```



```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -put mydataset.csv /exceldata
```

Figure 25: Upload large dataset into HDFS

2.4 Step 4: Verify data storage

```
hdfs dfs -ls /exceldata
```

```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -ls /exceldata
Found 1 items
-rw-r--r-- 1 hadoop supergroup 129488515 2026-02-03 15:11 /exceldata/mydataset.csv
```

Figure 26: Verify data storage in HDFS

2.5 Step 5: Analyze block distribution

```
hdfs fsck /exceldata/mydataset.csv -files -blocks -locations
```

```
hadoop@yugal-Inspiron-3583:~$ hdfs fsck /exceldata/mydataset.csv -files -blocks -locations
Connecting to namenode via http://localhost:9870/fsck?ugi=hadoop&files=1&blocks=1&locations=1&path=%2Fexceldata%2Fmydataset.csv
FSCK started by hadoop (auth:SIMPLE) from /127.0.0.1 for path /exceldata/mydataset.csv at Tue Feb 03 15:18:34 IST 2026

/exceldata/mydataset.csv 369487717 bytes, replicated: replication=1, 3 block(s): OK
0. BP-1685072063-127.0.1.1-1769537156943:blk_1073741830_1006 len=134217728 Live_repl=1 [DatanodeInfoWithStorage[127.0.0.1:9866,DS-47acaae1-e863-462a-9aec-290782a597df,
DISK]]
1. BP-1685072063-127.0.1.1-1769537156943:blk_1073741831_1007 len=134217728 Live_repl=1 [DatanodeInfoWithStorage[127.0.0.1:9866,DS-47acaae1-e863-462a-9aec-290782a597df,
DISK]]
2. BP-1685072063-127.0.1.1-1769537156943:blk_1073741832_1008 len=101052261 Live_repl=1 [DatanodeInfoWithStorage[127.0.0.1:9866,DS-47acaae1-e863-462a-9aec-290782a597df,
DISK]]

Status: HEALTHY
Number of data-nodes: 1
Number of racks: 1
Total dirs: 0
Total symlinks: 0

Replicated Blocks:
Total size: 369487717 B
Total files: 1
Total blocks (validated): 3 (avg. block size 123162572 B)
Minimally replicated blocks: 3 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Missing blocks: 0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Blocks queued for replication: 0

Erasure Coded Block Groups:
Total size: 0 B
Total files: 0
Total block groups (validated): 0
Minimally erasure-coded block groups: 0
Over-erasure-coded block groups: 0
```

Figure 27: Analyze block distribution in HDFS

2.6 Step 6: Download processed dataset from HDFS to local system

```
hdfs dfs -get /exceldata/mydataset.csv mydataset_from_hdfs.csv
```

```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -get /exceldata/mydataset.csv mydataset_from_hdfs.csv
```

Figure 28: Download processed dataset from HDFS to local system

2.7 Step 7: Delete data from HDFS

```
hdfs dfs -rm -r /exceldata
```

```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -get /exceldata/mydataset.csv mydataset_from_hdfs.csv
hdfs dfs -rm -r /exceldata
get: 'mydataset_from_hdfs.csv': File exists
Deleted /exceldata
hadoop@yugal-Inspiron-3583:~$
```

Figure 29: Delete data from HDFS

2.8 Step 8: Check file block information

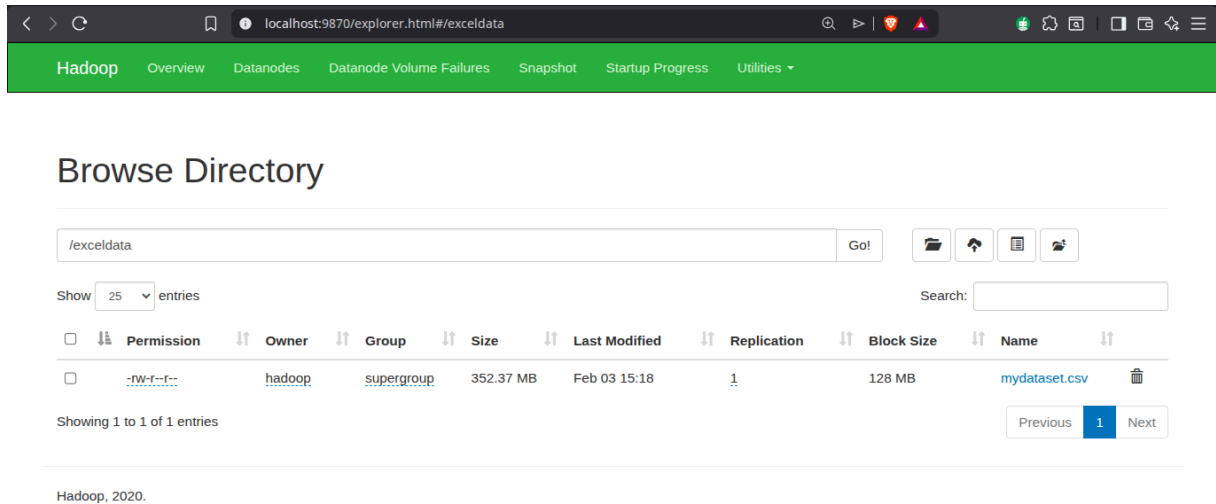


Figure 30: Check file block information in HDFS

2.9 Result

The large dataset (`mydataset.csv`) was successfully uploaded to HDFS under the `/exceldata` directory. The file was split into blocks and distributed across DataNodes, as verified by the block distribution analysis. The dataset was also downloaded back to the local system, confirming that data retrieval from HDFS works correctly. Finally, the dataset was deleted from HDFS, demonstrating proper data management and cleanup.

2.10 Conclusion

This assignment provided hands-on experience with Apache Hadoop HDFS, demonstrating how to upload large datasets, analyze block distribution, and manage data within the HDFS environment.