

Experiment No 1

Aim: Hadoop Installation and basic commands

Requirements: Linus/Windows OS, zip extractor (windows) to extract tar file, Open TCP Ports.

Theory:

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is part of the Apache Hadoop Core project.

Installation:

1. Download Hadoop from <https://dlcdn.apache.org/hadoop/common/> compatible version (my version – 2.10)

```
slowgamer@adnan-System-Product-Name:~$ wget https://dlcdn.apache.org/hadoop/common/current2/hadoop-2.10.2.tar.gz
--2022-08-07 18:25:41-- https://dlcdn.apache.org/hadoop/common/current2/hadoop-2.10.2.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... connected.
HTTP request sent, awaiting response... 200 OK
Length: 414624228 (395M) [application/x-gzip]
Saving to: 'hadoop-2.10.2.tar.gz'

hadoop-2.10.2.tar.gz 100%[=====>] 395.42M 1.87MB/s in 4m 32s

2022-08-07 18:30:33 (1.46 MB/s) - 'hadoop-2.10.2.tar.gz' saved [414624228/414624228]
```

2. Install ssh and pdsh for creating asymmetric keys with hdfs

```
slowgamer@adnan-System-Product-Name:~$ sudo apt-get install --reinstall ssh
Reading package lists... Done
Building dependency tree
Reading state information... Done
0 upgraded, 0 newly installed, 1 reinstalled, 0 to remove and 1 not upgraded.
Need to get 0 B/5,084 B of archives.
After this operation, 0 B of additional disk space will be used.
(Reading database ... 209581 files and directories currently installed.)
Preparing to unpack .../ssh_1%3a8.2p1-4ubuntu0.5_all.deb ...
Unpacking ssh (1:8.2p1-4ubuntu0.5) over (1:8.2p1-4ubuntu0.5) ...
Setting up ssh (1:8.2p1-4ubuntu0.5) ...

slowgamer@adnan-System-Product-Name:~$ sudo apt-get install --reinstall pdsh
Reading package lists... Done
Building dependency tree
Reading state information... Done
0 upgraded, 0 newly installed, 1 reinstalled, 0 to remove and 1 not upgraded.
Need to get 0 B/108 kB of archives.
After this operation, 0 B of additional disk space will be used.
Preconfiguring packages ...
(Reading database ... 209581 files and directories currently installed.)
Preparing to unpack .../pdsh_2.31-3build2_amd64.deb ...
Unpacking pdsh (2.31-3build2) over (2.31-3build2) ...
Setting up pdsh (2.31-3build2) ...
Processing triggers for man-db (2.9.1-1) ...
```

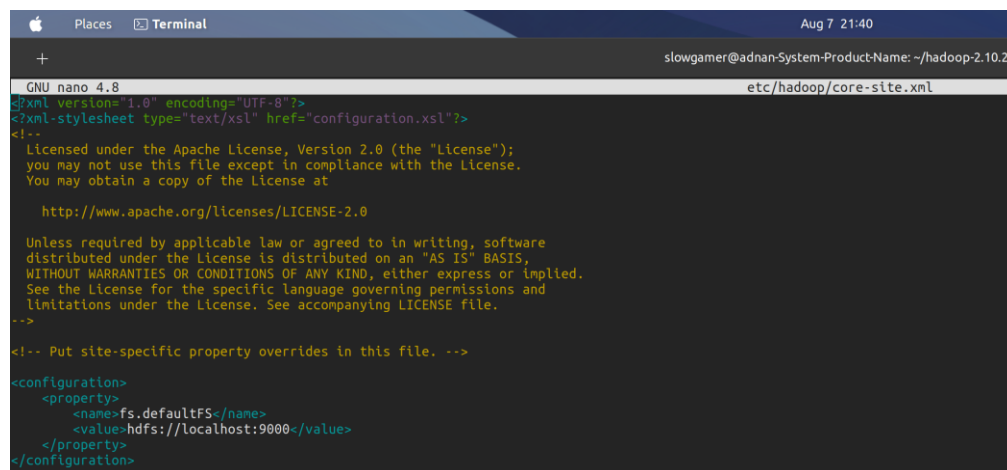
3. Extract the tar file and cd into extracted hadoop-version file and check in bin for runnable file hadoop if it is working

```
slowgamer@adnan-System-Product-Name:~$ cd hadoop-2.10.2
slowgamer@adnan-System-Product-Name:~/hadoop-2.10.2$ bin/hadoop
Usage: hadoop [--config confdir] [COMMAND | CLASSNAME]
  CLASSNAME                run the class named CLASSNAME
  or
  where COMMAND is one of:
    fs                      run a generic filesystem user client
    version                print the version
    jar <jar>              run a jar file
                           note: please use "yarn jar" to launch
                           YARN applications, not this command.
    checknative [-a|-h]    check native hadoop and compression libraries availability
    distcp <srcurl> <desturl> copy file or directories recursively
    archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
    classpath              prints the class path needed to get the
                           Hadoop jar and the required libraries
    credential             interact with credential providers
    daemonlog              get/set the log level for each daemon
    trace                  view and modify Hadoop tracing settings

Most commands print help when invoked w/o parameters.
```

4. Change settings of core-site.xml and hdfs-site.xml present in etc/hadoop as given by hadoop docs

```
slowgamer@adnan-System-Product-Name:~/hadoop-2.10.2$ sudo nano etc/hadoop/core-site.xml
slowgamer@adnan-System-Product-Name:~/hadoop-2.10.2$ sudo nano etc/hadoop/hdfs-site.xml
```



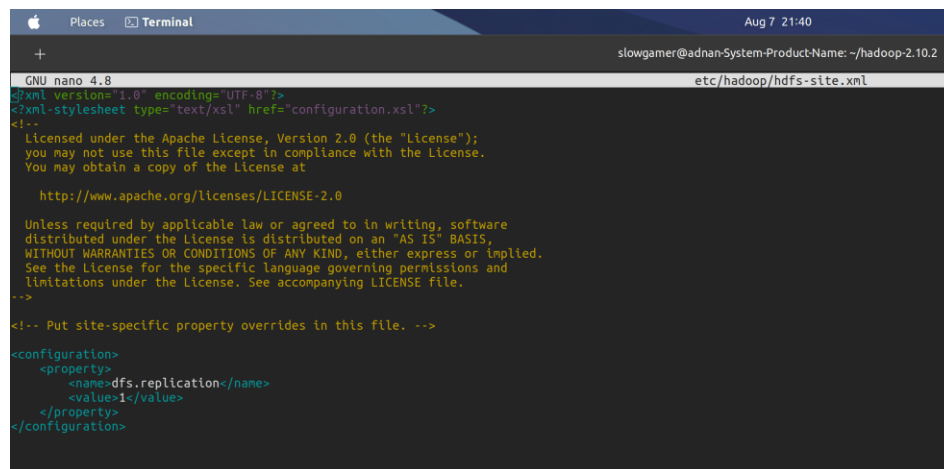
```
GNU nano 4.8
etc/hadoop/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the 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>
</configuration>
```



```
GNU nano 4.8
etc/hadoop/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

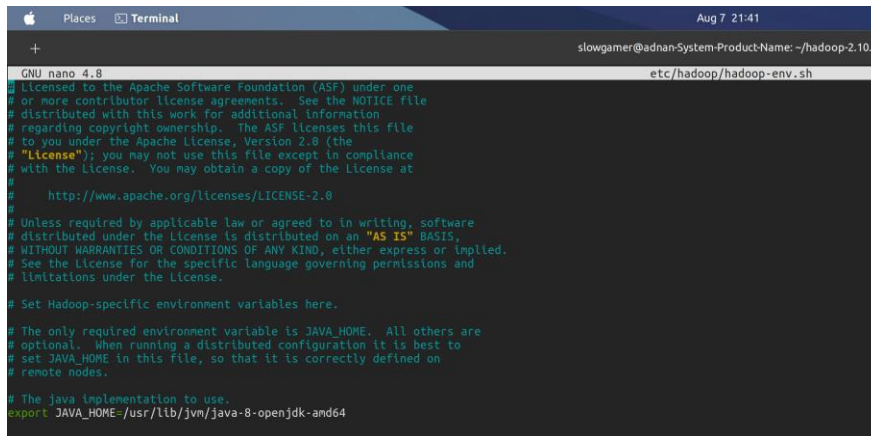
    http://www.apache.org/licenses/LICENSE-2.0

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>
```

5. Set JAVA_HOME in etc/hadoop/hadoop-env.sh to jdk 8 folder

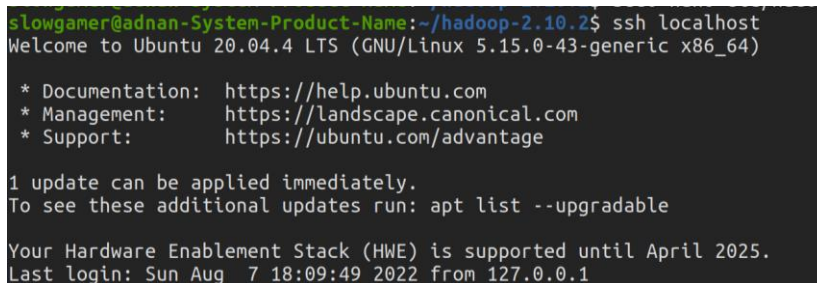


```

GNU nano 4.8 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-8-openjdk-and64

```

6. Connect ssh to localhost



```

slowgamer@adnan-System-Product-Name:~/hadoop-2.10.2$ ssh localhost
Welcome to Ubuntu 20.04.4 LTS (GNU/Linux 5.15.0-43-generic x86_64)

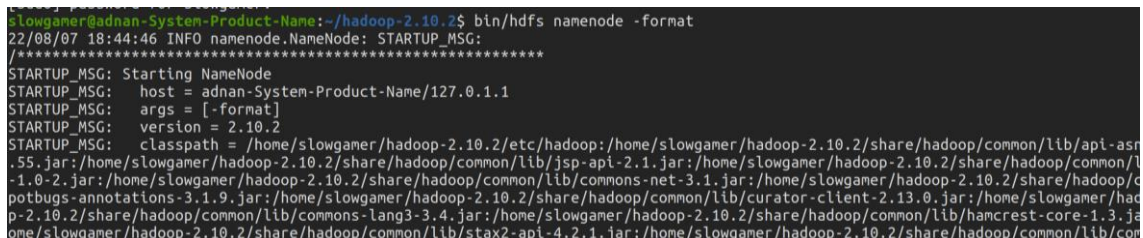
 * Documentation:  https://help.ubuntu.com
 * Management:   https://landscape.canonical.com
 * Support:      https://ubuntu.com/advantage

1 update can be applied immediately.
To see these additional updates run: apt list --upgradable

Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Sun Aug  7 18:09:49 2022 from 127.0.0.1

```

7. Format Name node

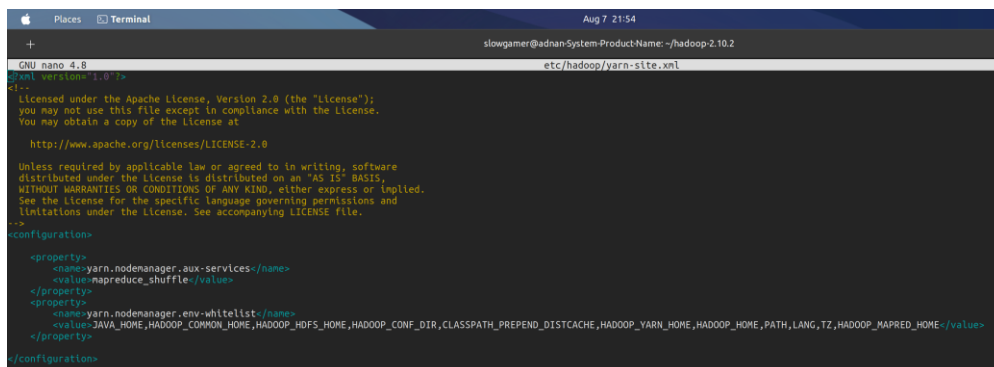


```

slowgamer@adnan-System-Product-Name:~/hadoop-2.10.2$ bin/hdfs namenode -format
22/08/07 18:44:46 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = adnan-System-Product-Name/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 2.10.2
STARTUP_MSG: classpath = /home/slowgamer/hadoop-2.10.2/etc/hadoop:/home/slowgamer/hadoop-2.10.2/share/hadoop/common/lib/api-asn
.55.jar:/home/slowgamer/hadoop-2.10.2/share/hadoop/common/lib/jsp-api-2.1.jar:/home/slowgamer/hadoop-2.10.2/share/hadoop/common/l
-1.0-2.jar:/home/slowgamer/hadoop-2.10.2/share/hadoop/common/lib/commons-net-3.1.jar:/home/slowgamer/hadoop-2.10.2/share/hadoop/c
potbugs-annotations-3.1.9.jar:/home/slowgamer/hadoop-2.10.2/share/hadoop/common/lib/curator-client-2.13.0.jar:/home/slowgamer/had
p-2.10.2/share/hadoop/common/lib/commons-lang3-3.4.jar:/home/slowgamer/hadoop-2.10.2/share/hadoop/common/lib/hamcrest-core-1.3.ja
one/slowgamer/hadoop-2.10.2/share/hadoop/common/lib/stax2-api-4.2.1.jar:/home/slowgamer/hadoop-2.10.2/share/hadoop/common/lib/com

```

8. Change settings of yarn-site.xml and mapred-site.xml present in etc/hadoop as given by hadoop docs



```

GNU nano 4.8 etc/hadoop/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,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_HOME,PATH,LANG,TZ,HADOOP_MAPRED_HOME</value>
  </property>
</configuration>

```

```
slowgamer@adnan-System-Product-Name: ~/hadoop-2.10.2
etc/hadoop/mapred-site.xml
<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}/share/hadoop/mapreduce/lib/*</value>
  </property>
</configuration>
```

9. Start dfs and yarn services using following commands and check if they’re working on ports 8088 and 50075

```
slowgamer@adnan-System-Product-Name:~/hadoop$ sbin/start-yarn.sh 8088
starting yarn daemons
starting resourcemanager, logging to /home/slowgamer/hadoop/logs/yarn-slowgamer-resourcemanager-adnan-System-Product-Name.out
localhost: starting nodemanager, logging to /home/slowgamer/hadoop/logs/yarn-slowgamer-nodemanager-adnan-System-Product-Name.out
slowgamer@adnan-System-Product-Name:~/hadoop$ sbin/start-dfs.sh 9870
Usage: start-dfs.sh [-upgrade|-rollback] [other options such as -clusterId]
slowgamer@adnan-System-Product-Name:~/hadoop$ sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/slowgamer/hadoop/logs/hadoop-slowgamer-namenode-adnan-System-Product-Name.out
localhost: starting datanode, logging to /home/slowgamer/hadoop/logs/hadoop-slowgamer-datanode-adnan-System-Product-Name.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/slowgamer/hadoop/logs/hadoop-slowgamer-secondarynamenode-adnan-System-Product-Name.out
slowgamer@adnan-System-Product-Name:~/hadoop$
```

hadoop

Cluster

Tools

All Applications

Cluster Metrics

Apps Submitted	0	Apps Pending	0	Apps Running	0	Apps Completed	0	Containers Running	0	Used Resources	<memory:0 B, vCores:0>	Total Resources	<memory:8 GB, vCores:8>	Reserved Resources	<memory:0 B, vCores:0>	Physical Mem Used %	27	Physical VCores	0
----------------	---	--------------	---	--------------	---	----------------	---	--------------------	---	----------------	------------------------	-----------------	-------------------------	--------------------	------------------------	---------------------	----	-----------------	---

Cluster Nodes Metrics

Active Nodes	1	Decommissioning Nodes	0	Decommissioned Nodes	0	Lost Nodes	0	Unhealthy Nodes	0	Rebooted Nodes	0	Shutdown Nodes	0
--------------	---	-----------------------	---	----------------------	---	------------	---	-----------------	---	----------------	---	----------------	---

Scheduler Metrics

Scheduler Type	Capacity Scheduler	Scheduling Resource Type	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]	Minimum Allocation	<memory:1024, vCores:1>	Maximum Allocation	<memory:8192, vCores:4>	Maximum Cluster Applications	0
----------------	--------------------	--------------------------	---	--------------------	-------------------------	--------------------	-------------------------	------------------------------	---

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Allocated GPUs	Reserved CPU Vcores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI
No data available in table																					

Showing 0 to 0 of 0 entries

hadoop

Overview

Utilities

HDFS basic commands:

```
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -ls /
Found 1 items
drwxr-xr-x  - slowgamer supergroup          0 2022-08-19 22:38 /user
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -ls /user
Found 1 items
drwxr-xr-x  - slowgamer supergroup          0 2022-08-19 22:38 /user/slowgamer
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -ls -r /user
Found 1 items
drwxr-xr-x  - slowgamer supergroup          0 2022-08-19 22:38 /user/slowgamer
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -ls
Found 1 items
drwxr-xr-x  - slowgamer supergroup          0 2022-08-19 22:38 testing
```

```
slowgamer@adnan-System-Product-Name:~/hadoop$ sudo nano ~/Desktop/forHdfs.txt
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop dfs -put testing/ ~/Desktop/forHdfs.txt
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

put: '/home/slowgamer/Desktop/forHdfs.txt': No such file or directory: 'hdfs://localhost:9000/home/slowgamer/Desktop/forHdfs.txt'
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop dfs -put ~/Desktop/forHdfs.txt testing/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -ls testing
Found 1 items
-rw-r--r--  1 slowgamer supergroup          43 2022-08-19 22:50 testing/forHdfs.txt
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -cat testing/forHdfs.txt
It is just a testing File
Hello From Adnan
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop dfs -mkdir testing2
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -cp testing/forHdfs.txt testin2
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -cp testing/forHdfs.txt testing2
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -ls
Found 3 items
-rw-r--r--  1 slowgamer supergroup          43 2022-08-19 22:55 testin2
drwxr-xr-x  - slowgamer supergroup          0 2022-08-19 22:50 testing
drwxr-xr-x  - slowgamer supergroup          0 2022-08-19 22:55 testing2
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -cat testing2/forHdfs2.txt
cat: 'testing2/forHdfs2.txt': No such file or directory
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -cat testin2/forHdfs2.txt
cat: 'testin2/forHdfs2.txt': No such file or directory
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -cat testin2/forHdfs.txt
cat: 'testin2/forHdfs.txt': No such file or directory
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -cat testing2/forHdfs.txt
It is just a testing File
Hello From Adnan
```

```
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop dfs -rm -r -f testing2
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Deleted testing2
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop fs -ls
Found 2 items
-rw-r--r--  1 slowgamer supergroup          43 2022-08-19 22:55 testin2
drwxr-xr-x  - slowgamer supergroup          0 2022-08-19 22:50 testing
slowgamer@adnan-System-Product-Name:~/hadoop$ bin/hadoop dfs -get testing/forHdfs.txt ~/Desktop/practiceprogram
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

slowgamer@adnan-System-Product-Name:~/hadoop$ sudo cat ~/Desktop/practiceprogram/forHdfs.txt
It is just a testing File
Hello From Adnan
slowgamer@adnan-System-Product-Name:~/hadoop$
```

Conclusion: We have successfully installed Hadoop and implemented the Hadoop basic commands