

Exp No: 9

Date:

HADOOP

SET UP A SINGLE HADOOP CLUSTER AND SHOW THE PROCESS USING WEB UI

AIM:

To set-up one node Hadoop cluster.

PROCEDURE:

1. System Update
2. Install Java
3. Add a dedicated Hadoop user
4. Install SSH and setup SSH certificates
5. Check if SSH works
6. Install Hadoop
7. Modify Hadoop config files
8. Format Hadoop filesystem
9. Start Hadoop
10. Check Hadoop through web UI
11. Stop Hadoop

THEORY

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from a single server to thousands of machines, each offering local computation and storage.

HADOOP ARCHITECTURE

Hadoop framework includes following four modules:

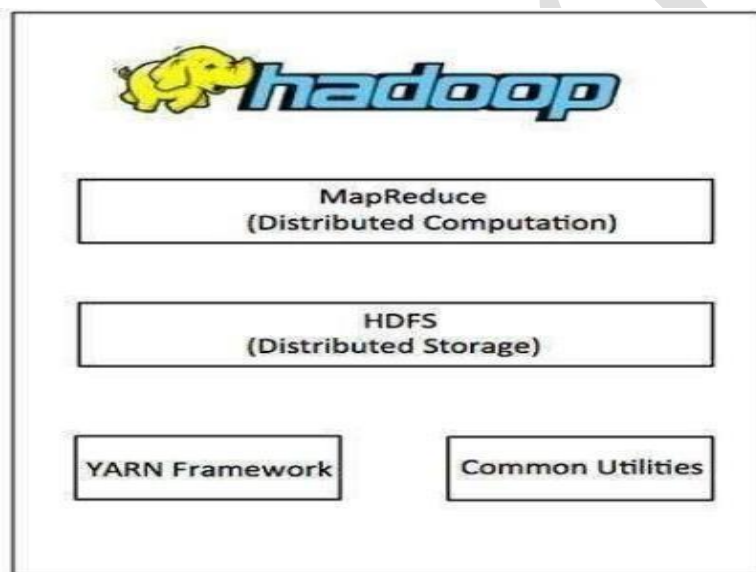
Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contain the necessary Java files and scripts required to start Hadoop.

Hadoop YARN: This is a framework for job scheduling and cluster resource management.

Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.

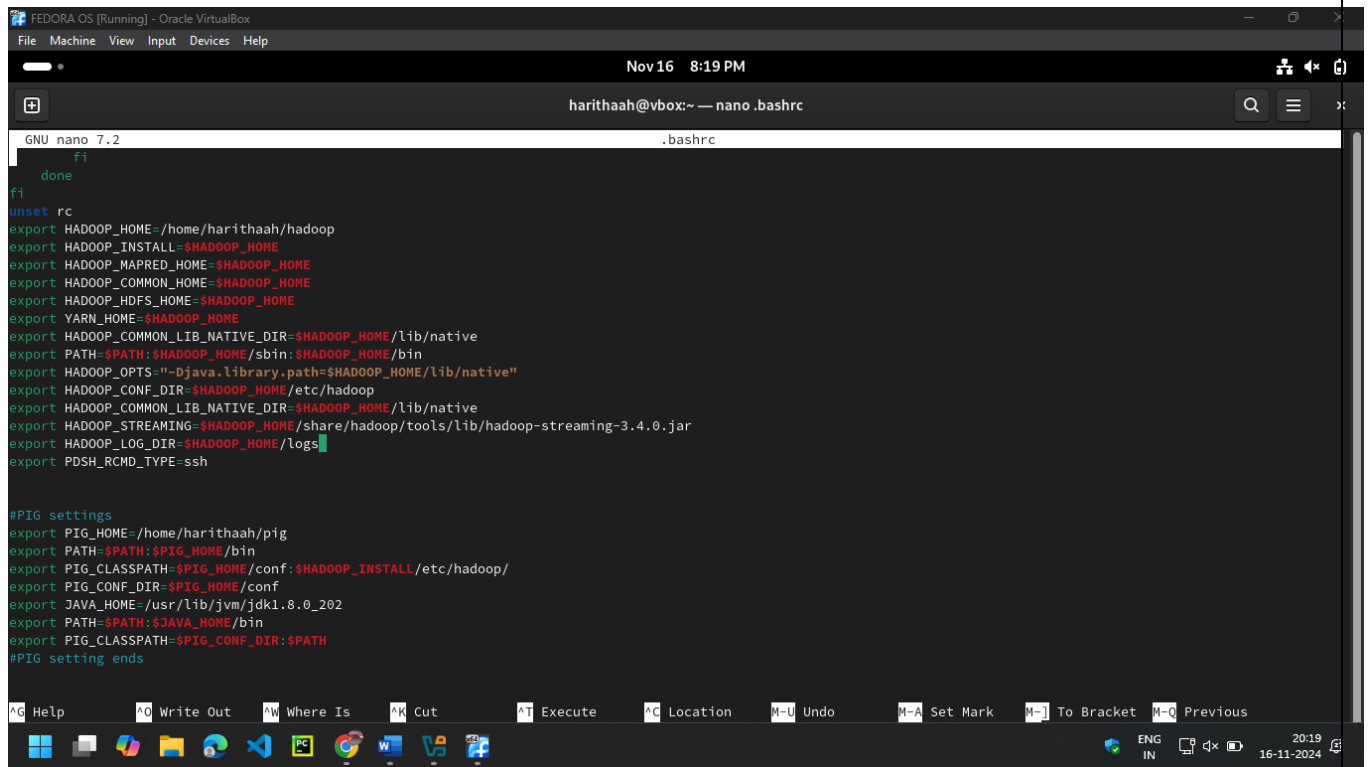
Hadoop MapReduce: This is a YARN-based system for parallel processing of large data sets.

We can use following diagram to depict these four components available in Hadoop framework.



PROCEDURE

\$ nano ~/.bashrc



```
FEDORA OS [Running] - Oracle VirtualBox
File Machine View Input Devices Help

Nov 16 8:19 PM

harithaah@vbox:~ — nano .bashrc

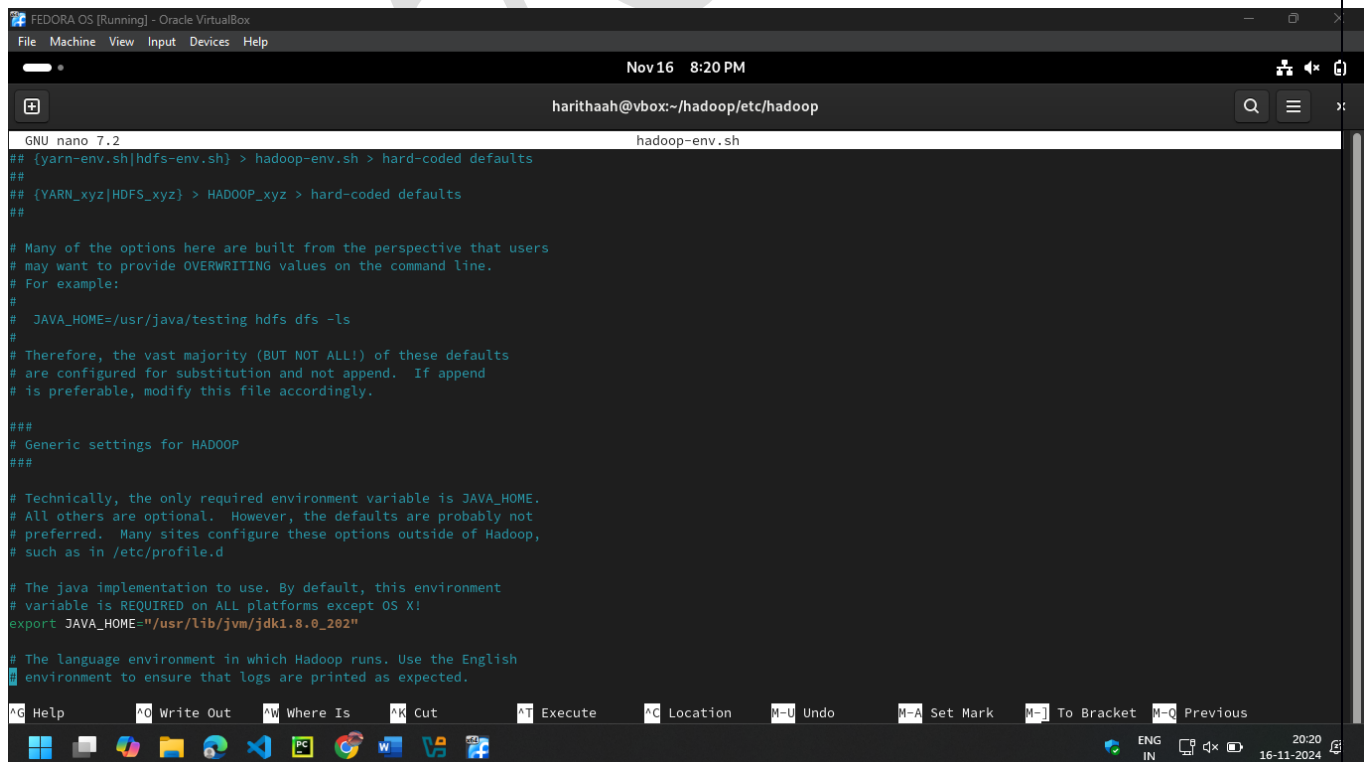
GNU nano 7.2 .bashrc

done
fi
unset rc
export HADOOP_HOME=/home/harithaah/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH:$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_STREAMING=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export PDSH_RCMD_TYPE=ssh

#Pig settings
export PIG_HOME /home/harithaah/pig
export PATH:$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/
export PIG_CONF_DIR=$PIG_HOME/conf
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_202
export PATH:$PATH:$JAVA_HOME/bin
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
#Pig setting ends

^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location ^U Undo ^A Set Mark ^] To Bracket ^Q Previous
```

\$ nano \$HADOOP_HOME/etc/hadoop/hadoop-env.sh



```
FEDORA OS [Running] - Oracle VirtualBox
File Machine View Input Devices Help

Nov 16 8:20 PM

harithaah@vbox:~/hadoop/etc/hadoop

GNU nano 7.2 hadoop-env.sh

## {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/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.

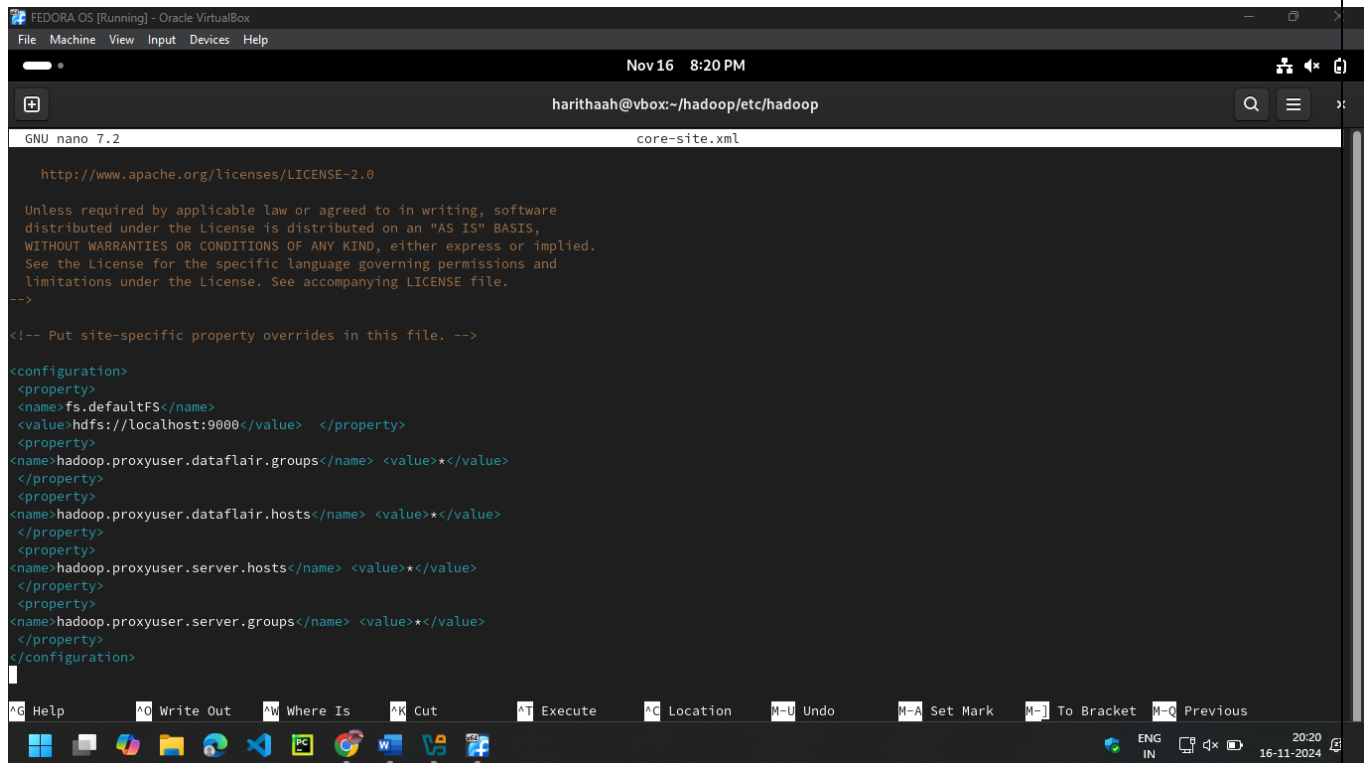
###
# Generic settings for HADOOP
###

# 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/jdk1.8.0_202"

# The language environment in which Hadoop runs. Use the English
# environment to ensure that logs are printed as expected.
```

\$nano \$HADOOP_HOME/etc/hadoop/core-site.xml



```
FEDORA OS [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Nov 16 8:20 PM
harithaah@vbox:~/hadoop/etc/hadoop
GNU nano 7.2 core-site.xml

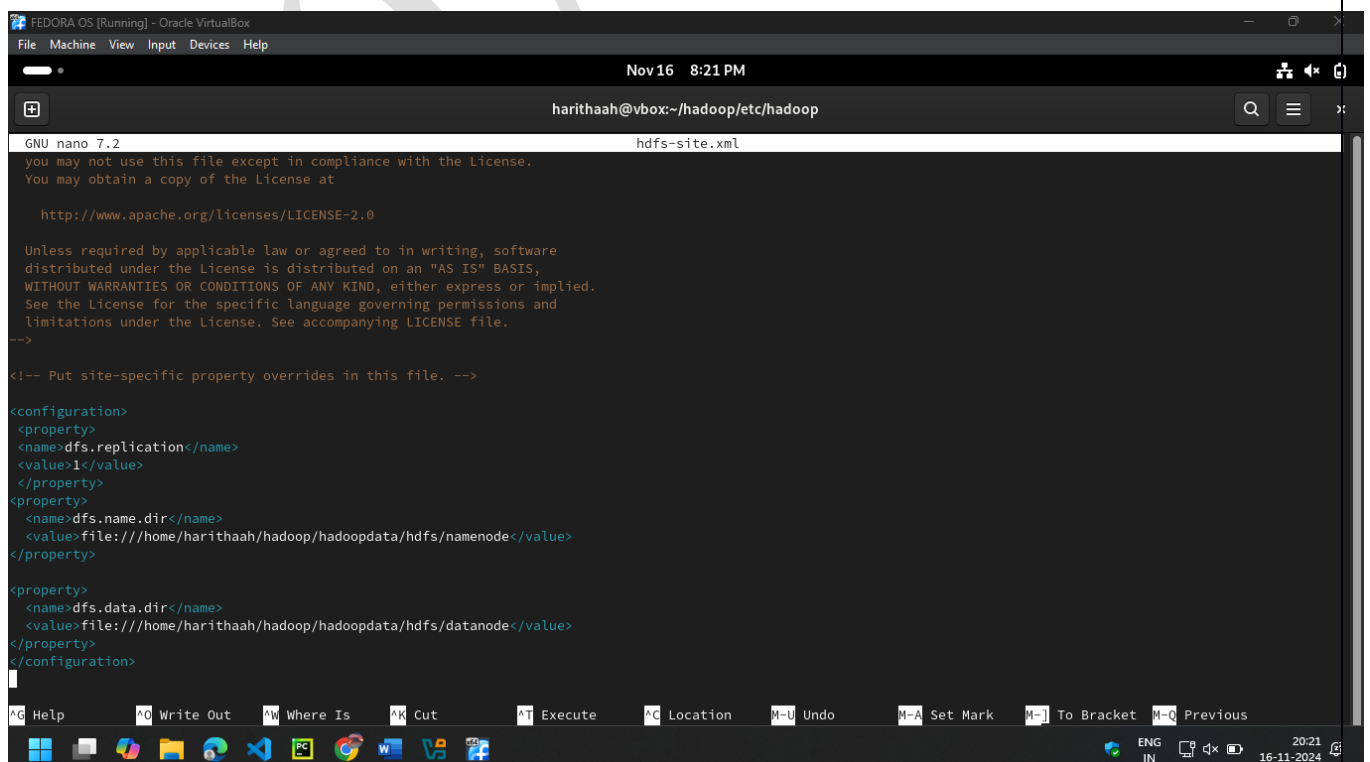
http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>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>
```

\$nano \$HADOOP_HOME/etc/hadoop/hdfs-site.xml



```
FEDORA OS [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Nov 16 8:21 PM
harithaah@vbox:~/hadoop/etc/hadoop
GNU nano 7.2 hdfs-site.xml

you may not use this file except in compliance with the License.
You may obtain a copy of the License at

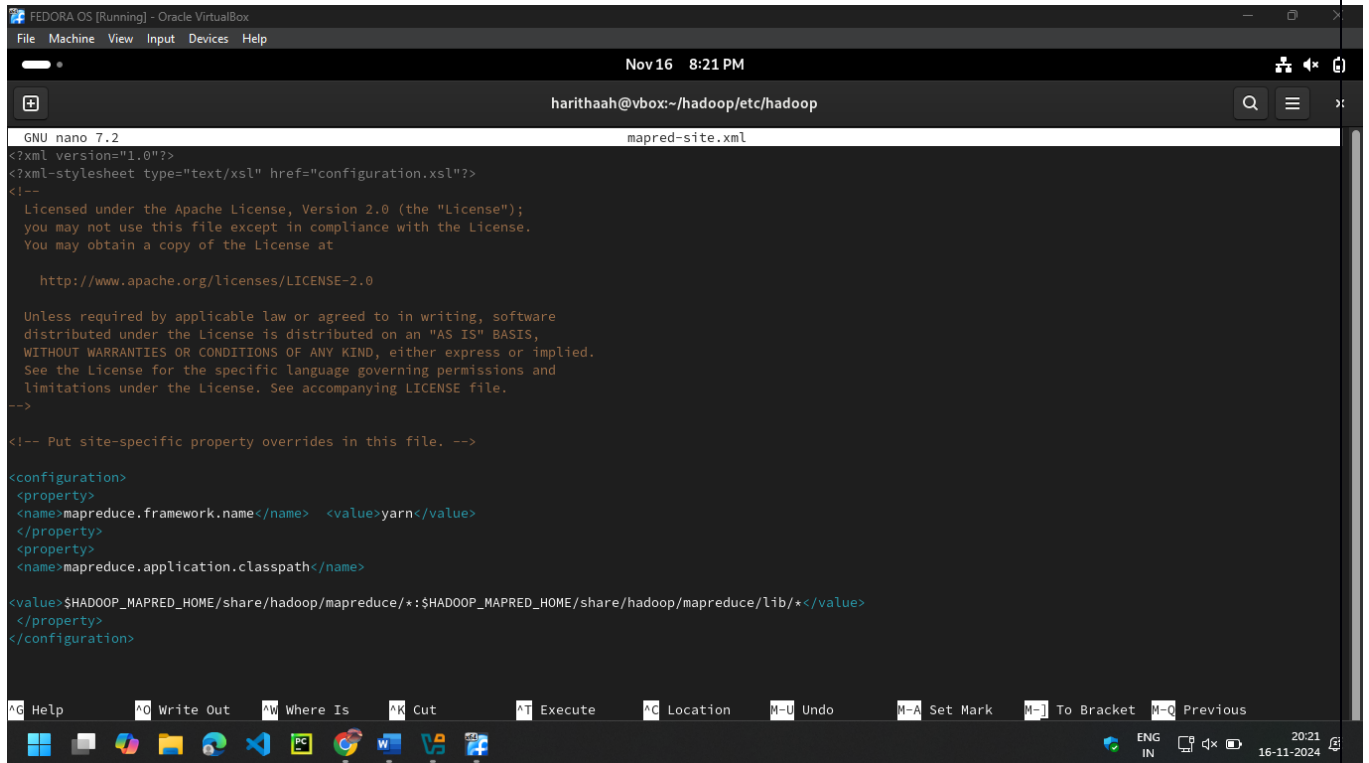
http://www.apache.org/licenses/LICENSE-2.0

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>
  <property>
    <name>dfs.name.dir</name>
    <value>file:///home/harithaah/hadoop/hadoopdata/hdfs/namenode</value>
  </property>
  <property>
    <name>dfs.data.dir</name>
    <value>file:///home/harithaah/hadoop/hadoopdata/hdfs/datanode</value>
  </property>
</configuration>
```

\$nano \$HADOOP_HOME/etc/hadoop/mapred-site.xml



```
FEDORA OS [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Nov 16 8:21 PM
harithaah@vbox:~/hadoop/etc/hadoop

GNU nano 7.2 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

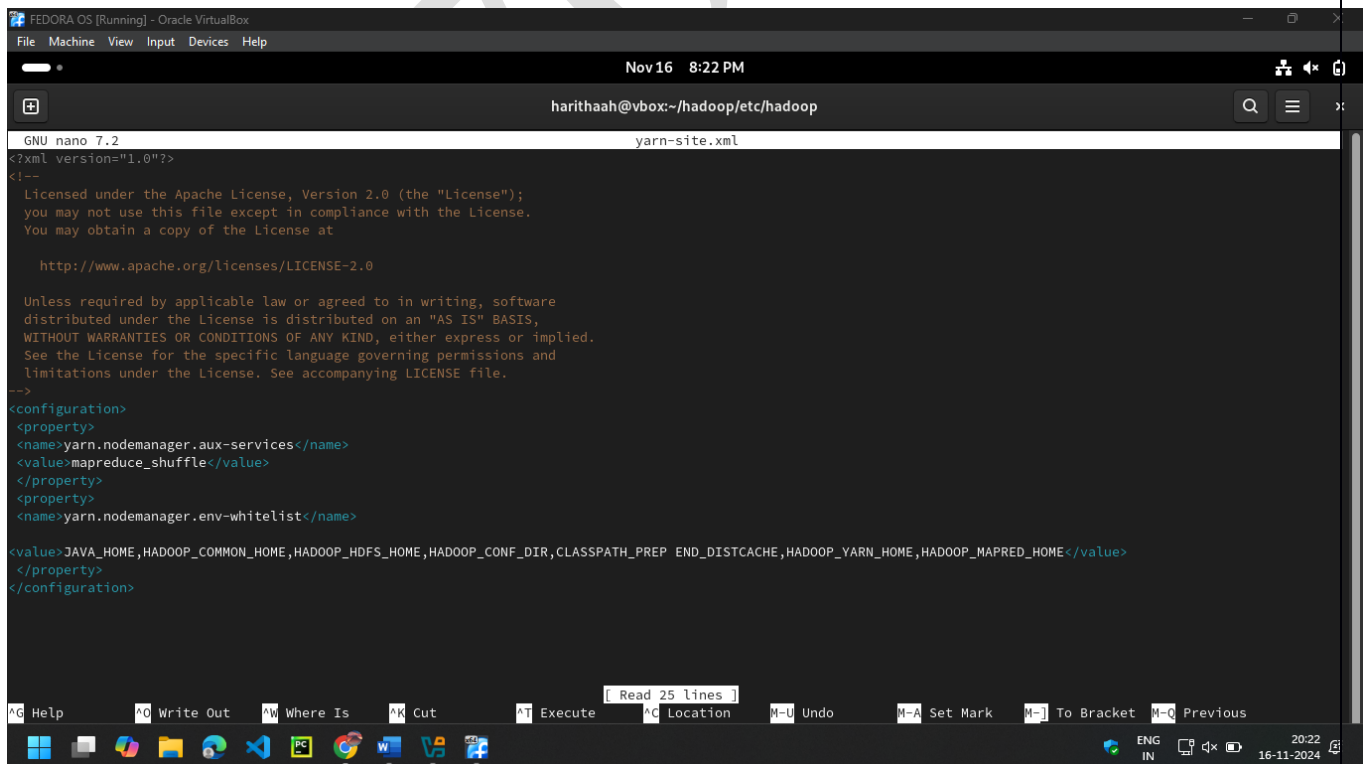
http://www.apache.org/licenses/LICENSE-2.0

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/share/hadoop/mapreduce/lib/*</value>
  </property>
</configuration>
```

\$nano \$HADOOP_HOME/etc/hadoop/yarn-site.xml



```
FEDORA OS [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Nov 16 8:22 PM
harithaah@vbox:~/hadoop/etc/hadoop

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>
  <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_PREP END_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
  </property>
</configuration>
```

\$ start-all.sh

```

harithaah@vbox:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as harithaah 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 [vbox]
vbox: Warning: Permanently added 'vbox' (ED25519) to the list of known hosts.
Starting resourcemanager
Starting nodemanagers

```

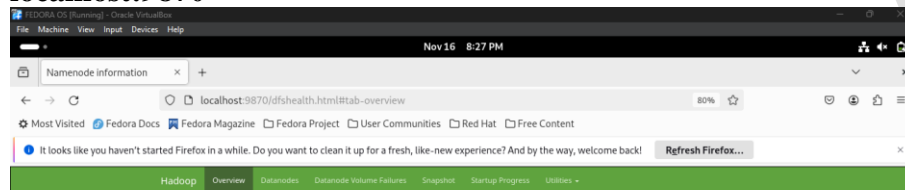
\$ jps

```

harithaah@vbox:~$ jps
5746 DataNode
5572 NameNode
6009 SecondaryNameNode
6955 Jps
6316 ResourceManager
6509 NodeManager

```

localhost:9870



Overview 'localhost:9000' (✓active)

Started:	Sat Nov 16 20:24:02 +0530 2024
Version:	3.4.0, rtd8b77f938f266b7791783192ee7a5d4aee760
Compiled:	Mon Mar 04 12:05:00 +0530 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID:6221cde8-6c43-4a26-8c9b-a624bab4fa7a
Block Pool ID:	BP-1378615733-10.0.2.15-1728572581706

Summary

Security is off.
 Safemode is off.
 61 files and directories, 32 blocks (32 replicated blocks, 0 erasure coded block groups) = 93 total filesystem objects.

localhost:8088

RESULT:

Thus, Hadoop has been successfully installed.