

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:

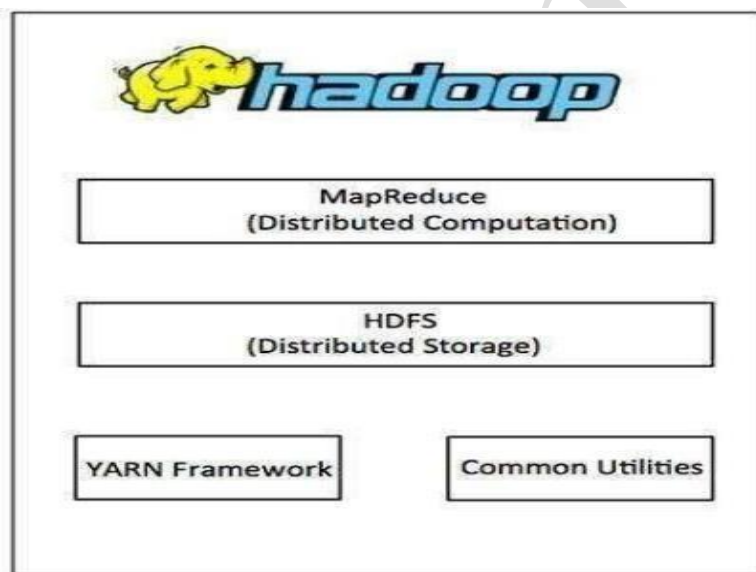
Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contain the necessary Java files and scripts required to start Hadoop.

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

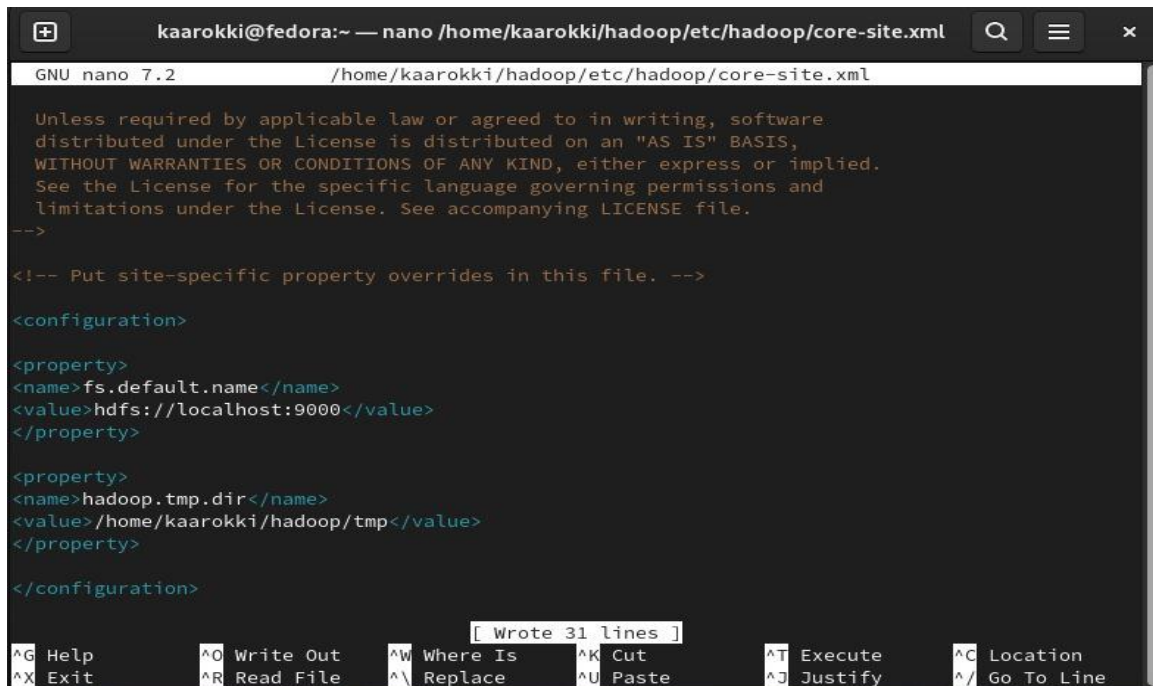
```
kaarokki@fedora:~ — nano /home/kaarokki/.bashrc
GNU nano 7.2 /home/kaarokki/.bashrc Modified
export PATH=$PATH:/usr/lib/jvm/jdk1.8.0_202/bin
export HADOOP_HOME=~/.hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$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=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export PDSH_RCMD_TYPE=ssh
export HADOOP_COMMON_NAME=~/.hadoop

#PIG settings
export PIG_HOME=/home/kaarokki/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/
export PIG_CONF_DIR=$PIG_HOME/conf
#export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
#PIG setting ends
```

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

```
kaarokki@fedora:~ — nano /home/kaarokki/hadoop/etc/hadoop/hadoop-env.sh
GNU nano 7.2 /home/kaarokki/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.
##
## 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.
[ Read 430 lines ]
```

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



The screenshot shows the nano text editor in a terminal window. The title bar indicates the user is kaarokki@fedora and the file being edited is /home/kaarokki/hadoop/etc/hadoop/core-site.xml. The editor shows the beginning of an XML configuration file. It starts with a license notice, followed by a comment to put site-specific property overrides in this file. Then, there is a <configuration> block containing two <property> elements. The first property is fs.default.name with a value of hdfs://localhost:9000. The second property is hadoop.tmp.dir with a value of /home/kaarokki/hadoop/tmp. The status bar at the bottom shows that 31 lines have been written. The bottom of the screen displays nano's command shortcuts: ^G for Help, ^O for Write Out, ^W for Where Is, ^K for Cut, ^T for Execute, ^C for Location, ^X for Exit, ^R for Read File, ^_ for Replace, ^U for Paste, ^J for Justify, and ^/ for Go To Line.

```
kaarakki@fedora:~ — nano /home/kaarokki/hadoop/etc/hadoop/core-site.xml
GNU nano 7.2 /home/kaarokki/hadoop/etc/hadoop/core-site.xml

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>

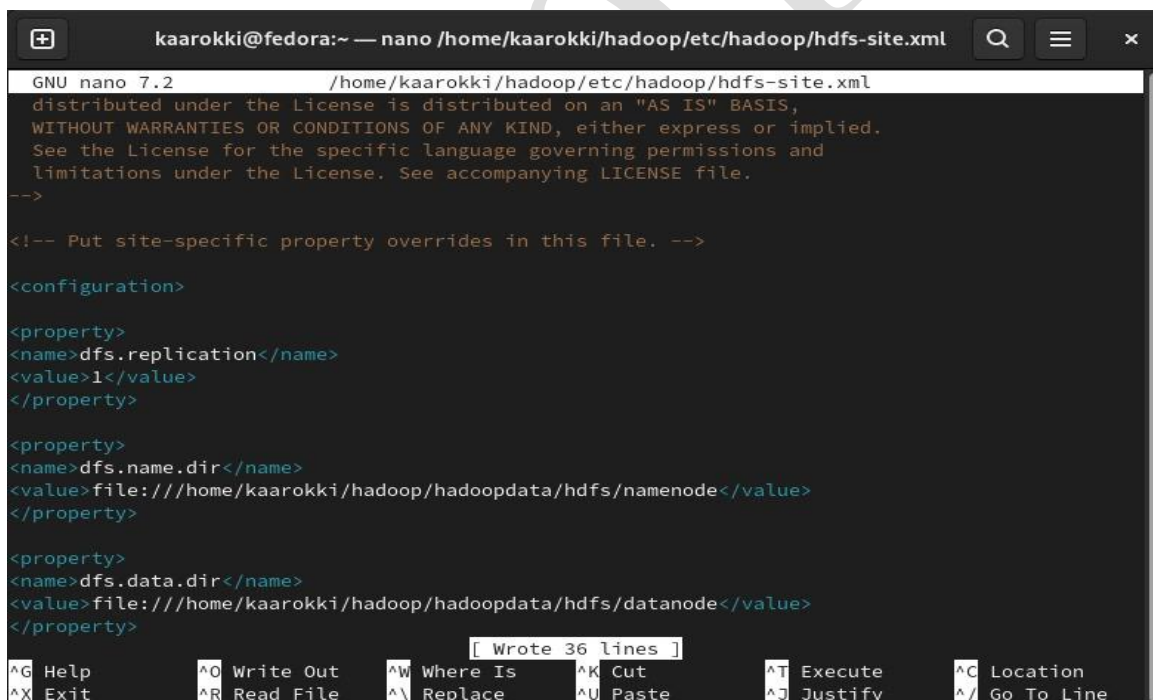
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>

<property>
<name>hadoop.tmp.dir</name>
<value>/home/kaarokki/hadoop/tmp</value>
</property>

</configuration>

[ Wrote 31 lines ]
^G Help      ^O Write Out  ^W Where Is   ^K Cut        ^T Execute    ^C Location
^X Exit      ^R Read File  ^_ Replace    ^U Paste      ^J Justify    ^/ Go To Line
```

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



The screenshot shows the nano text editor in a terminal window. The title bar indicates the user is kaarokki@fedora and the file being edited is /home/kaarokki/hadoop/etc/hadoop/hdfs-site.xml. The editor shows the beginning of an XML configuration file. It starts with a license notice, followed by a comment to put site-specific property overrides in this file. Then, there is a <configuration> block containing three <property> elements. The first property is dfs.replication with a value of 1. The second property is dfs.name.dir with a value of file:///home/kaarokki/hadoop/hadoopdata/hdfs/namenode. The third property is dfs.data.dir with a value of file:///home/kaarokki/hadoop/hadoopdata/hdfs/datanode. The status bar at the bottom shows that 36 lines have been written. The bottom of the screen displays nano's command shortcuts: ^G for Help, ^O for Write Out, ^W for Where Is, ^K for Cut, ^T for Execute, ^C for Location, ^X for Exit, ^R for Read File, ^_ for Replace, ^U for Paste, ^J for Justify, and ^/ for Go To Line.

```
kaarakki@fedora:~ — nano /home/kaarokki/hadoop/etc/hadoop/hdfs-site.xml
GNU nano 7.2 /home/kaarokki/hadoop/etc/hadoop/hdfs-site.xml

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/kaarokki/hadoop/hadoopdata/hdfs/namenode</value>
</property>

<property>
<name>dfs.data.dir</name>
<value>file:///home/kaarokki/hadoop/hadoopdata/hdfs/datanode</value>
</property>

[ Wrote 36 lines ]
^G Help      ^O Write Out  ^W Where Is   ^K Cut        ^T Execute    ^C Location
^X Exit      ^R Read File  ^_ Replace    ^U Paste      ^J Justify    ^/ Go To Line
```

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

```
kaarokki@fedora:~ — nano /home/kaarokki/hadoop/etc/hadoop/mapred-site.xml
GNU nano 7.2 /home/kaarokki/hadoop/etc/hadoop/mapred-site.xml
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>mapred.job.tracker</name>
<value>localhost:9001</value>
</property>
<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/
</property>
</configuration>

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

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

```
kaarokki@fedora:~ — nano /home/kaarokki/hadoop/etc/hadoop/yarn-site.xml
GNU nano 7.2 /home/kaarokki/hadoop/etc/hadoop/yarn-site.xml
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
/property>

/configuration>

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



```
$ start-all.sh
```

```

kaarokki@fedora:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as kaarokki in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: ssh: connect to host localhost port 22: Connection refused
Starting datanodes
localhost: ssh: connect to host localhost port 22: Connection refused
Starting secondary namenodes [fedora]
fedora: ssh: connect to host fedora port 22: Connection refused
Starting resourcemanager
Starting nodemanagers
localhost: ssh: connect to host localhost port 22: Connection refused

```

\$ jps

```
kaarokki@fedora:~$ jps
3456 ResourceManager
3750 Jps
```

localhost:9870

Overview

Version	3.3.6
Compiled	2023-06-18T08:22Z by ubuntu from (HEAD detached at release-3.3.6-RC1)
NameNode Address	localhost:9000
Started	Wed Aug 14 21:51:32 -0400 2024
Last Checkpoint	Never
Checkpoint Period	3600 seconds
Checkpoint Transactions	1000000

Checkpoint Image URI

- file:///tmp/hadoop-kali/dfs/namesecondary

Checkpoint Editlog URI

- file:///tmp/hadoop-kali/dfs/namesecondary

Hadoop, 2023.

localhost:8088

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:8 GB, vCores:8>

Cluster Nodes Metrics

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

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=MB), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Allocated GPUs
No data available in table															

Showing 0 to 0 of 0 entries

RESULT:

Thus, Hadoop has been successfully installed.