Exp.6 APACHE HADOOP : PROCEDURE TO SET UP THE ONE NODE CLUSTER

AIM

To find a procedure to set up the one node hadoop cluster.

PROCEDURE:

ABOUT HADOOP:

The Apache™ Hadoop project develops open-source software for relaible, scalable, distributed, computing

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming model.

It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Rather than rely on hardware to deliver high-availability,the library itself is designed to detect and and handle failures at the application layer,So delivering a high valuable service on top of a cluster of computers, each of which may prone to failures.

The project include these modules:

- Hadoop Common: The common utilities that support the other Hadoop modules.
- Hadoop Distributed File System(HDFS™): A distributed file system that provides high throwput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce:A YARN-based system for parallel processing of large data sets.

STEP 1:INSTALLING JAVA

Java is the primary requirement to setup Hadoop on any System,So make sure you have Java Installed in your system using the following command.

```
hadoop@data-HP-Notebook:~$ java -version
        java version "1.8.0_171"
        Java(TM) SE Runtime Environment (build 1.8.0_171-b11)
        Java HotSpot(TM) 64-Bit Server VM (build 25.171-b11, mixed mode)
```

If you don’t have java installed in your system,use one of the following steps to install it first.

```
$sudo add-apt-repository ppa:webupd8team/java
$sudo apt-get update
$sudo apt-get install oracle-java8-installer
```

STEP 2:CREATING HADOOP CLUSTER

We recommended creating a normal(non root) account for Hadoop working. So create a system account using the following command.

```
$ sudo adduser hadoop
$ sudo adduser hadoop sudo
```

After creating an account,it also required to set up key-based ssh to its own account.To do this execute the following commands.

```
$su - hadoop
$ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
$cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$chmod 0600 ~/.ssh/authorized_keys
```

Let’s verify key based login. Below command should not ask for the password but the first time it will prompt for adding RSA to the list of known hosts.

```
$ssh localhost
$exit
```

STEP 3:Download Hadoop 3.1

In this step, download hadoop 3.1 source archive file using below command. You can also select alternate download mirror for increasing download speed.

```
cd ~
wget http://www-eu.apache.org/dist/hadoop/common/hadoop-3.1.0/hadoop-3.1.0.tar.gz
tar xzf hadoop-3.1.0.tar.gz
mv hadoop-3.1.0 hadoop
```

STEP 4: Setup Hadoop Pseudo-Distributed Mode

Setup Hadoop Environment Variables

First, we need to set environment variable uses by Hadoop. Edit **~/.bashrc** file and append following values at end of file.

```
#HADOOP ENVIRONMENT VARIABLES

export HADOOP_HOME=/home/hadoop/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
```

Now apply the changes in the current running environment

```
$source ~/.bashrc
```

Now edit **$HADOOP_HOME/etc/hadoop/hadoop-env.sh** file and set **JAVA_HOME** environment variable. Change the JAVA path as per install on your system. This path may vary as per your operating system version and installation source. So make sure you are using correct path.

```
$export JAVA_HOME=/usr/lib/jvm/java-8-oracle
```

Setup Hadoop Configuration Files

Hadoop has many of configuration files, which need to configure as per requirements of your Hadoop infrastructure. Let’s start with the configuration with basic Hadoop single node cluster setup. first, navigate to below location

```
$cd $HADOOP_HOME/etc/hadoop
```

Edit core-site.xml

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

Edit hdfs-site.xml

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
```

Edit mapred-site.xml

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

Edit yarn-site.xml

```
<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_MAPRED_HOME</value>
    </property>
```

</configuration>

STEP 5: Format Namenode

Create folder for datanode and namenode

$mkdir -p *home*/Hadoop/hadoopdata/hdfs/namenode

$mkdir -p *home*/Hadoop/hadoopdata/hdfs/datanode

Now format the namenode using the following command, make sure that Storage directory is

$hdfs namenode -format

Sample output:

```
WARNING: /home/hadoop/hadoop/logs does not exist. Creating.
2018-05-02 17:52:09,678 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = tecadmin/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.1.0
...
...
...
2018-05-02 17:52:13,717 INFO common.Storage: Storage directory
/home/hadoop/hadoopdata/hdfs/namenode has been successfully formatted.
2018-05-02 17:52:13,806 INFO namenode.FSImageFormatProtobuf: Saving image file /
home/hadoop/hadoopdata/hdfs/namenode/current/fsimage.ckpt_0000000000000000000
using no compression
2018-05-02 17:52:14,161 INFO namenode.FSImageFormatProtobuf: Image file
/home/hadoop/hadoopdata/hdfs/namenode/current/fsimage.ckpt_0000000000000000000
of size 391 bytes saved in 0 seconds .
2018-05-02 17:52:14,224 INFO namenode.NNStorageRetentionManager: Going to retain
1 images with txid >= 0
2018-05-02 17:52:14,282 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at tecadmin/127.0.1.1
************************************************************/
```

STEP 6:Start Hadoop Cluster

Let’s start your Hadoop cluster using the scripts provides by Hadoop. Just navigate to your $HADOOP_HOME/sbin directory and execute scripts one by one.

$cd $HADOOP_HOME/sbin/

Now run **start-dfs.sh** script.

$./start-dfs.sh

Sample output:

```
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [tecadmin]
2018-05-02 18:00:32,565 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
```

Now run **start-yarn.sh** script.

```
./start-yarn.sh
```

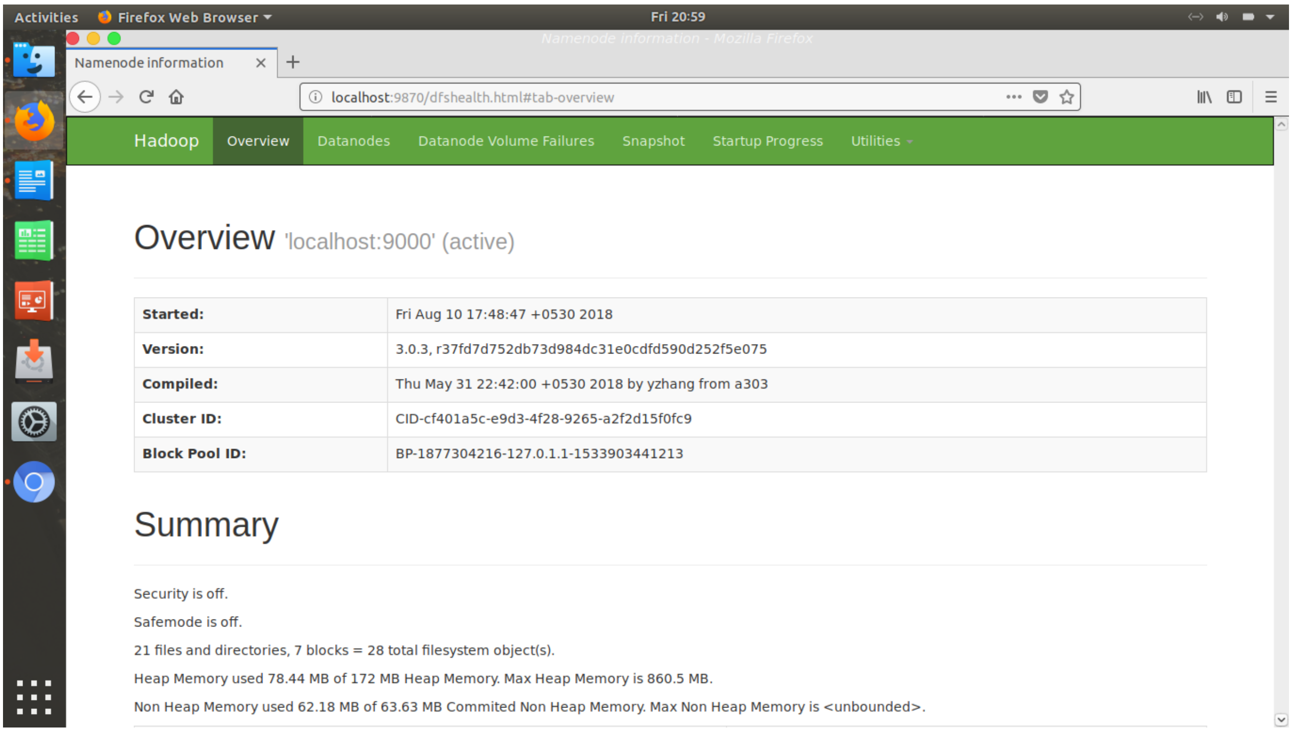
Sample output:

```
Starting resourcemanager
Starting nodemanagers
```

Access Hadoop Services in Browser

Hadoop NameNode started on port 9870 default. Access your server on port 9870 in your favorite web browser.

Localhost:9870



Overview 'localhost:9000' (active)

Started:	Fri Aug 10 17:48:47 +0530 2018
Version:	3.0.3, r37fd7d752db73d984dc31e0cdfd590d252f5e075
Compiled:	Thu May 31 22:42:00 +0530 2018 by yzhang from a303
Cluster ID:	CID-cf401a5c-e9d3-4f28-9265-a2f2d15f0fc9
Block Pool ID:	BP-1877304216-127.0.1.1-1533903441213

Summary

Security is off.

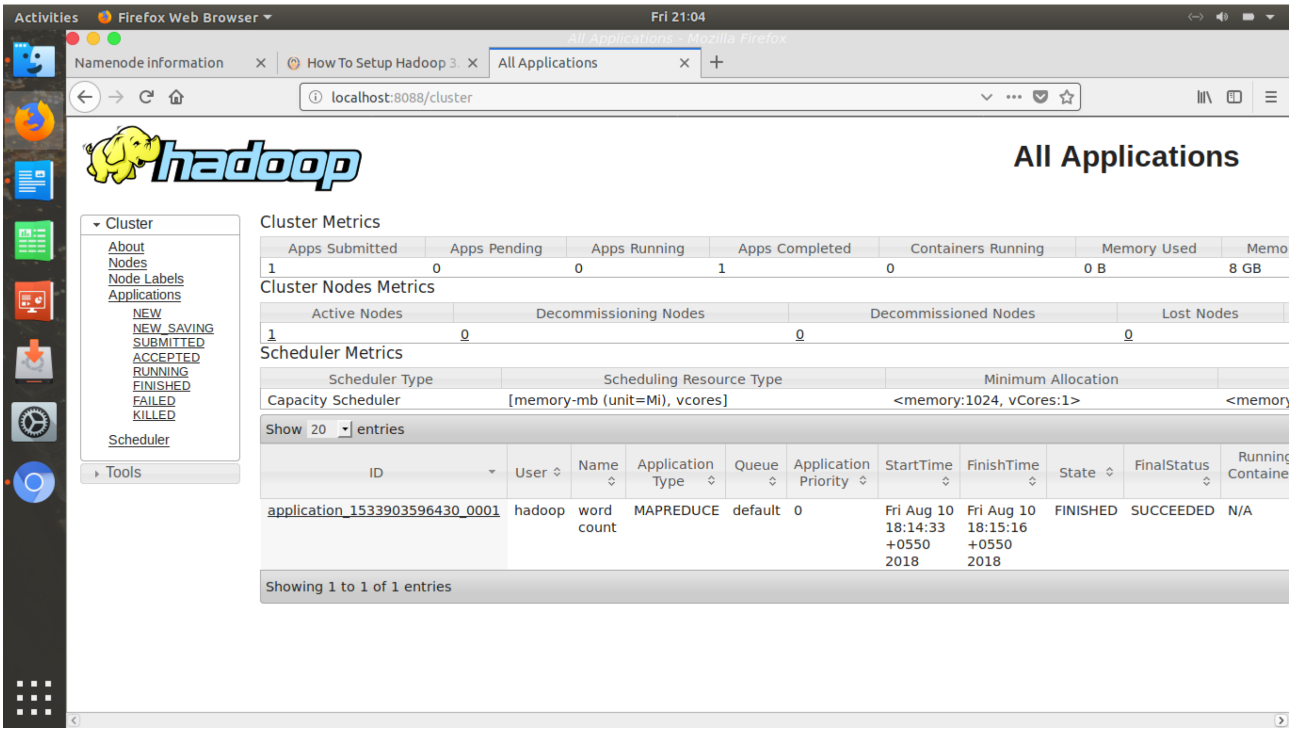
Safemode is off.

21 files and directories, 7 blocks = 28 total filesystem object(s).

Heap Memory used 78.44 MB of 172 MB Heap Memory. Max Heap Memory is 860.5 MB.

Non Heap Memory used 62.18 MB of 63.63 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Now access port 8088 getting the information about the cluster and all applications



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memo
1	0	0	1	0	0 B	8 GB

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>	<memory

Show 20 entries

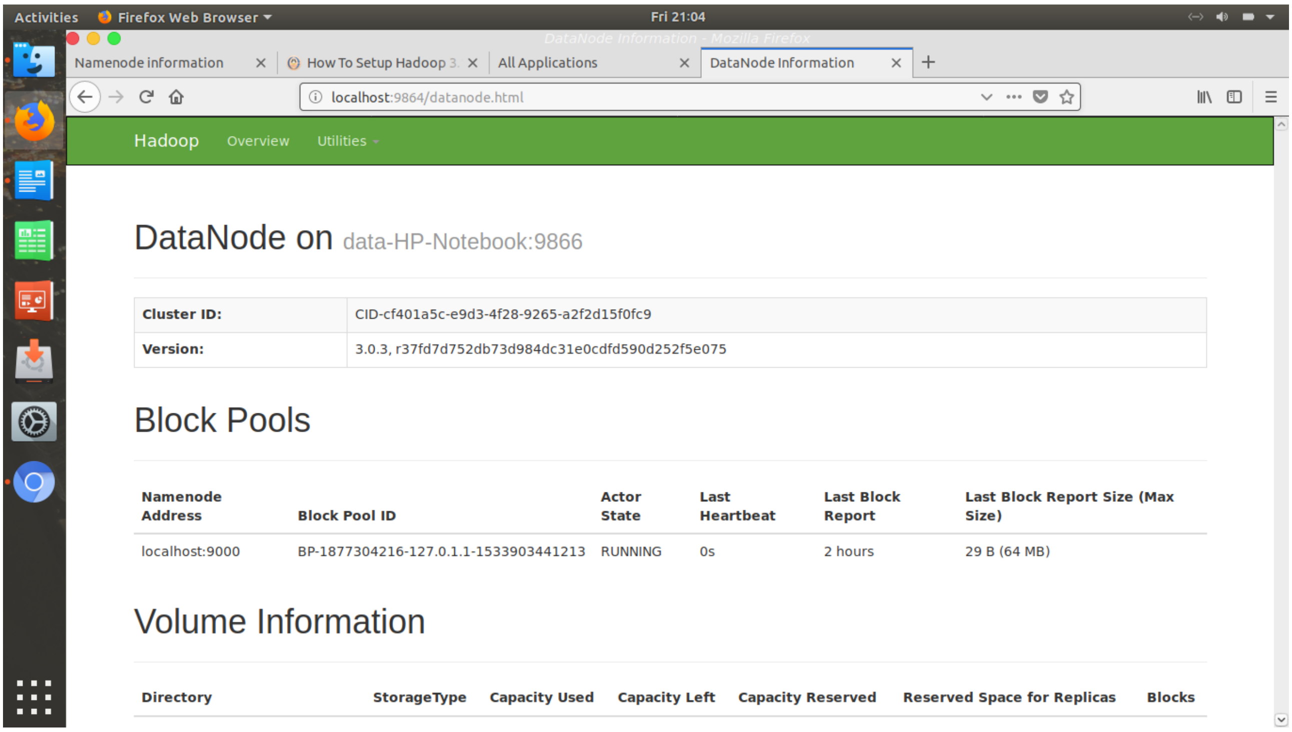
ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containe
application_1533903596430_0001	hadoop	word count	MAPREDUCE	default	0	Fri Aug 10 18:14:33 +0550 2018	Fri Aug 10 18:15:16 +0550 2018	FINISHED	SUCCEEDED	N/A

Showing 1 to 1 of 1 entries

localhost:8042

Access port 9864 to get details about your Hadoop node.

`Localhost:9864`



DataNode on data-HP-Notebook:9866

Cluster ID:	CID-cf401a5c-e9d3-4f28-9265-a2f2d15f0fc9
Version:	3.0.3, r37fd7d752db73d984dc31e0cdfd590d252f5e075

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:9000	BP-1877304216-127.0.1.1-1533903441213	RUNNING	0s	2 hours	29 B (64 MB)

Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks

RESULT:

Thus to find a procedure to set up one node Hadoop cluster was done and output was verified successfully.