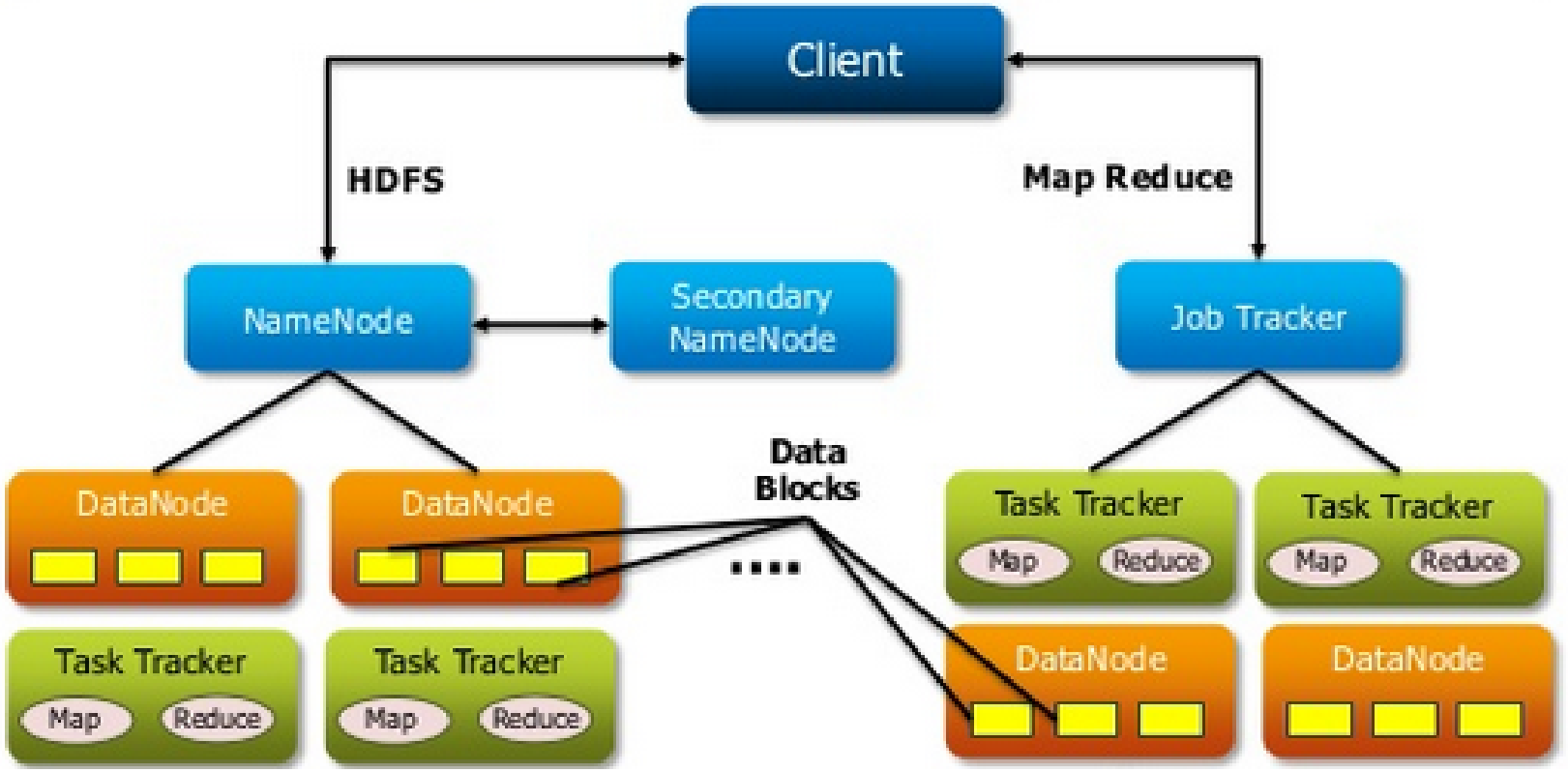
Install Hadoop: Setting up a Single Node Hadoop Cluster

Enabling technologies for Data Science

2017msbda008
Samiksha Agarwal

What is Hadoop ?

- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.



What is a Node ?

- A Node is a single independently running computer which is the part of a cluster.
- A cluster is made up of one or more nodes.
- Each node has processing as well as storage hardware.

Running Hadoop on Ubuntu (Single node cluster setup)

The presentation here will describe the required steps for setting up a single-node Hadoop cluster backed by the Hadoop Distributed File System, running on Ubuntu Linux.

SINGLE-NODE INSTALLATION

- Step 1: Make a folder (installation) in your home directory.
- Step 2: Click Here to download the Java 8 Package. Save this file in installation folder.
- Step 3: Extract the Java Tar File.
  Command: tar -xvf jdk-8u181-linux-x64.tar.gz

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ cd /home/samiksha/installtion
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion$ tar -xvf jdk-8u181-linux-x64.tar.gz
jdk1.8.0_181/
jdk1.8.0_181/javafx-src.zip
jdk1.8.0_181/bin/
jdk1.8.0_181/bin/jmc
jdk1.8.0_181/bin/serialver
jdk1.8.0_181/bin/jmc.ini
jdk1.8.0_181/bin/jstack
jdk1.8.0_181/bin/rmiregistry
jdk1.8.0_181/bin/unpack200
jdk1.8.0_181/bin/jar
```

- Step 4: Download the Hadoop 2.7.3 Package.
  Command: wget
  https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz
- Step 5:Extract the Hadoop tar File.
  Command: tar -xvf hadoop-2.7.3.tar.gz

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion$ tar -xvf hadoop-2.7.3.tar.gz
hadoop-2.7.3/
hadoop-2.7.3/bin/
hadoop-2.7.3/bin/hadoop
hadoop-2.7.3/bin/hadoop.cmd
hadoop-2.7.3/bin/rcc
hadoop-2.7.3/bin/hdfs
hadoop-2.7.3/bin/hdfs.cmd
hadoop-2.7.3/bin/container-executor
hadoop-2.7.3/bin/test-container-executor
```

- Step 6:Add the Hadoop and Java paths in the bash file(.bashrc).
  Open.bashrcfile.
  Command:nano .bashrc

Now, add Hadoop and Java Path as shown below.

```
# user specific aliases and functions

export HADOOP_HOME=$HOME/installtion/hadoop-2.7.3
export HADOOP_CONF_DIR=$HOME/installation/hadoop-2.7.3/etc/hadoop
export HADOOP_COMMON_HOME=$HOME/installation/hadoop-2.7.3
export HADOOP_MAPRED_HOME=$HOME/installation/hadoop-2.7.3
export HADOOP_HDFS_HOME=$HOME/installation/hadoop-2.7.3
export YARN_HOME=$HOME/installation/hadoop-2.7.3
export PATH=$PATH:$HOME/installation/hadoop-2.7.3/bin

#set java_home

export JAVA_HOME=/home/samiksha/installation/jdk1.8.0_181
export PATH=/home/samiksha/installation/jdk1.8.0_181/bin:$PATH
export HADOOP_CLASSPATH=/home/samiksha/installation/jdk1.8.0_181
/lib/tools.jar
```

Then, save the bash file and close it. Press (ctrl + x) $\rightarrow$ y $\rightarrow$ enter For applying all these changes to the current Terminal, execute the source command.

Command: source .bashrc

To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and hadoop version commands.

Command: java-version

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion$ java -version
java version "1.8.0_181"
Java(TM) SE Runtime Environment (build 1.8.0_181-b13)
Java HotSpot(TM) 64-Bit Server VM (build 25.181-b13, mixed mode)
```

Command: hadoop version

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion$ hadoop version
Hadoop 2.7.3
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r baa91f7c6bc9cb9
2be5982de4719c1c8af91ccff
Compiled by root on 2016-08-18T01:41Z
Compiled with protoc 2.5.0
From source with checksum 2e4ce5f957ea4db193bce3734ff29ff4
This command was run using /home/samiksha/hadoop-2.7.3/share/hadoop/common/hadoo
p-common-2.7.3.jar
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion$
```

- Step 7:Edit the Hadoop Configuration files.
  Command: cd /hadoop-2.7.3/etc/hadoop/
  Command: ls
  All the Hadoop configuration files are located in hadoop-2.7.3/etc/hadoop directory as you can see in the snapshot below:

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion$  cd hadoop-2.7.3/etc/hadoo
p/
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$ ls
capacity-scheduler.xml      httpfs-env.sh            mapred-env.sh
configuration.xsl           httpfs-log4j.properties  mapred-queues.xml.template
container-executor.cfg      httpfs-signature.secret  mapred-site.xml.template
core-site.xml               httpfs-site.xml          slaves
hadoop-env.cmd              kms-acls.xml             ssl-client.xml.example
hadoop-env.sh               kms-env.sh               ssl-server.xml.example
hadoop-metrics2.properties  kms-log4j.properties     yarn-env.cmd
hadoop-metrics.properties   kms-site.xml             yarn-env.sh
hadoop-policy.xml           log4j.properties         yarn-site.xml
hdfs-site.xml               mapred-env.cmd
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$
```

- Step 8: Open core-site.xml ,core-site.xml informs Hadoop daemon where NameNode runs in the cluster.

Command: nano core-site.xml

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$ nano core-site.xml
```

This file will be open →

```
<?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>

</configuration>
```

Edit the property mentioned below inside configuration tag:

```
< ?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

- Step 9: Edit hdfs-site.xml , hdfs-site.xml contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, Secondary NameNode). It also includes the replication factor and block size of HDFS.
  Command: nano hdfs-site.xml

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$ nano hdfs-site.xml
```

Edit the property mentioned below inside configuration tag:

```
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.permission</name>
<value>false</value>
</property>
```

- Step 10:Edit the mapred-site.xml file ,mapred-site.xmlcontains configuration settings of MapReduce application like number of JVM that can run in parallel, the size of the mapper and the reducer process, CPU cores available for a process, etc.
  In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml fileusing mapred-site.xml template.
  Command: cp mapred-site.xml.template mapred-site.xml

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$ cp mapred-site.xml.template mapred-site.xml
```

  Command: nano mapred-site.xml

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$ nano mapred-site.xml
```

  Edit the property mentioned below inside configuration tag:

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

- Step 11: Edit yarn-site.xml ,yarn-site.xml contains configuration settings of ResourceManager and NodeManager like application memory management size ,the operation needed on program and algorithm, etc.
  Command: nano yarn-site.xml

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$ nano yarn-site.xml
```

Edit the property mentioned below inside configuration tag:

```
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</nam
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
```

- Step 12: Edit hadoop-env.sh ,hadoop-env.sh contains the environment variables that are used in the script to run Hadoop like Java home path, etc.
  Command: nano hadoop−env.sh

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$ nano hadoop-env.sh
```

  add the Java Path as mentioned below:

```
#set java_home

export JAVA_HOME=/home/samiksha/installation/jdk1.8.0_181
```

- Step 13: Go to Hadoop home directory and format the NameNode.
  Command: cd
  Command: cd /home/samiksha/installtion/hadoop-2.7.3
  Command: bin/hadoop namenode -format

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/etc/hadoop$ cd
samiksha@samiksha-HP-Pavilion-Notebook:~$ cd /home/samiksha/installtion/hadoop-2.7.3
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3$  bin/hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

18/07/26 21:04:59 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = samiksha-HP-Pavilion-Notebook/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.3
STARTUP_MSG:   classpath = /home/samiksha/hadoop-2.7.3/etc/hadoop:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/htrace-core-3.1.0-incubating.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/commons-compress-1.4.1.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/commons-cli-1.2.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/junit-4.11.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/commons-io-2.4.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/netty-3.6.2.Final.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/zookeeper-3.4.6.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/avro-1.7.4.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/jersey-json-1.9.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/asm-3.2.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/httpcore-4.2.5.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/jersey-server-1.9.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/xmlenc-0.52.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/apacheds-i18n-2.0.0-M15.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/activation-1.1.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/commons-codec-1.4.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/snappy-java-1.0.4.1.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/paranamer-2.3.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/guava-11.0.2.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/stax-api-1.0-2.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/servlet-api-2.5.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/mockito-all-1.8.5.jar:/home/samiksha/hadoop-2.7.3/share/hadoop/common/lib/jackson-jaxr
```

This formats the HDFS via NameNode. This command is only executed for the first time. Formatting the file system means initializing the directory specified by the dfs.name.dir variable.

Never format, up and running Hadoop filesystem. You will lose all your data stored in the HDFS.

- Step 13: Once the NameNode is formatted, go to hadoop-2.7.3/sbin directory and start all the daemons.
  Command: cd /home/samiksha/installtion/hadoop-2.7.3/sbin

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3$ cd /home/samiksha/installtion/hadoop-2.7.3/sbin
```

Either you can start all daemons with a single command or do it individually.

Command: ./start-all.sh

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/sbin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
samiksha@localhost's password:
localhost: starting namenode, logging to /home/samiksha/installtion/hadoop-2.7.3/logs/hadoop-samiksha-namenode-samiksha-HP-Pavilion-Notebook.o
ut
samiksha@localhost's password:
localhost: starting datanode, logging to /home/samiksha/installtion/hadoop-2.7.3/logs/hadoop-samiksha-datanode-samiksha-HP-Pavilion-Notebook.o
ut
Starting secondary namenodes [0.0.0.0]
samiksha@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/samiksha/installtion/hadoop-2.7.3/logs/hadoop-samiksha-secondarynamenode-samiksha-HP-Pav
ilion-Notebook.out
starting yarn daemons
starting resourcemanager, logging to /home/samiksha/installtion/hadoop-2.7.3/logs/yarn-samiksha-resourcemanager-samiksha-HP-Pavilion-Notebook.
out
samiksha@localhost's password:
localhost: starting nodemanager, logging to /home/samiksha/installtion/hadoop-2.7.3/logs/yarn-samiksha-nodemanager-samiksha-HP-Pavilion-Notebo
ok.out
```

Or you can run all the services individually as below:

- **Start NameNode:** The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files stored in the HDFS and tracks all the file stored across the cluster.
  Command: ./hadoop-daemon.sh start namenode
  Command: jps

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/sbin$  ./hadoop-daemon.sh start namenode
starting namenode, logging to /home/samiksha/installtion/hadoop-2.7.3/logs/hadoop-samiksha-namenode-samiksha-HP-Pavilion-Notebook.out
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/sbin$ jps
7109 NameNode
7183 Jps
```

- **Start DataNode:** a DataNode connects to the Namenode and it responds to the requests from the Namenode for different operations.
  Command: ./hadoop-daemon.sh start datanode
  Command: jps

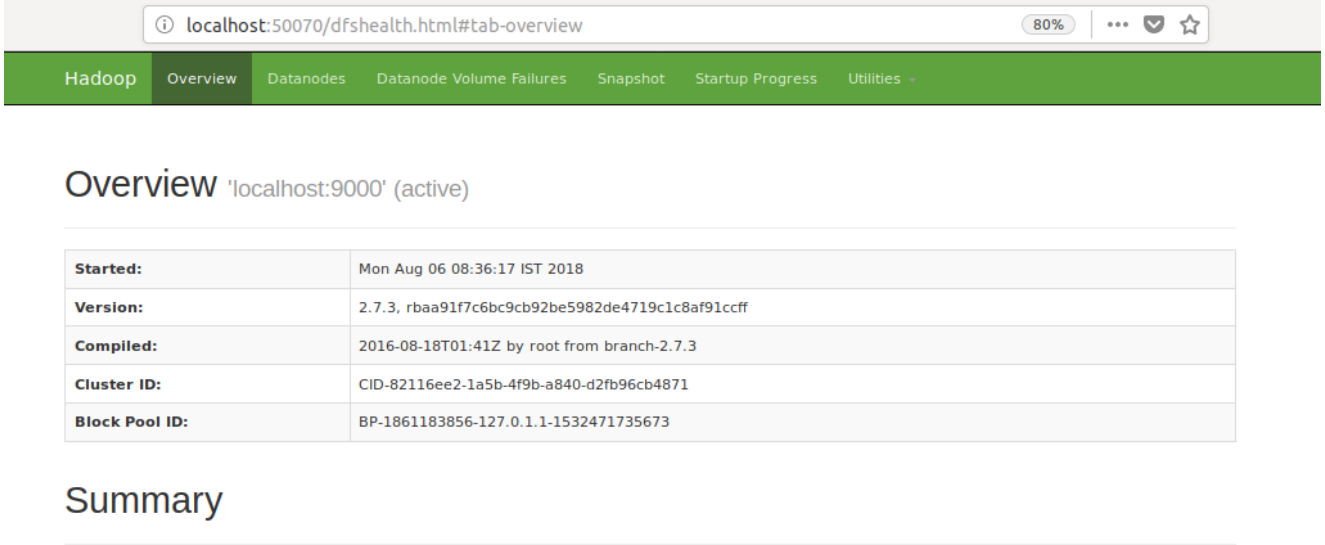
```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/sbin$ ./hadoop-daemon.sh start datanode
starting datanode, logging to /home/samiksha/installtion/hadoop-2.7.3/logs/hadoop-samiksha-datanode-samiksha-HP-Pavilion-Notebook.out
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/sbin$ jps
7216 DataNode
7297 Jps
7109 NameNode
```

- **Start ResourceManager:** Its work is to manage each NodeManagers and the each application's Application Master.
  Command: ./yarn-daemon.sh start resourcemanager
- **Start NodeManager:** The NodeManager in each machine framework is the agent which is responsible for managing containers, monitoring their resource usage and reporting the same to the ResourceManager.
  Command: ./yarn-daemon.sh start nodemanager
- **Start JobHistoryServer:** JobHistoryServer is responsible for servicing all job history related requests from client.
  Command: ./mr-jobhistory-daemon.sh start historyserver

- Step 15: To check that all the Hadoop services are up and running, run the below command.
  Command: jps

```
samiksha@samiksha-HP-Pavilion-Notebook:~/hadoop-2.7.3/sbin$ jps
9636 Jps
8981 SecondaryNameNode
9498 NodeManager
8746 DataNode
8556 NameNode
7548 JobHistoryServer
9150 ResourceManager
samiksha@samiksha-HP-Pavilion-Notebook:~/hadoop-2.7.3/sbin$
```

- Step 16: Now open the Mozilla browser and go to **localhost:50070/dfshealth.html** to check the NameNode interface.



successfully installed a single node Hadoop cluster.

now, stop all services

Command: ./stop-all.sh

```
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/sbin$ ./stop-all.sh
This script is Deprecated. Instead use stop-dfs.sh and stop-yarn.sh
Stopping namenodes on [localhost]
samiksha@localhost's password:
localhost: stopping namenode
samiksha@localhost's password:
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
samiksha@0.0.0.0's password:
0.0.0.0: stopping secondarynamenode
stopping yarn daemons
stopping resourcemanager
samiksha@localhost's password:
localhost: stopping nodemanager
no proxyserver to stop
samiksha@samiksha-HP-Pavilion-Notebook:~/installtion/hadoop-2.7.3/sbin$
```

YARN Web UI

(for more details you can follow this link Click Here)

Thank You