

Steps to install Hadoop 2.6 Installing on Ubuntu 15.04 (Single-Node Cluster)

For **Hadoop installation on the UNIX environment** you need

1. Java Installation
2. SSH installation
3. Hadoop Installation and File Configuration

1. Java Installation

Step 1. Type "**java -version**" in prompt to find if the java is installed or not. If not then **download java** from **<http://www.oracle.com/technetwork/java/javase/downloads/jdk7-downloads-1880260.html>**

The **tar file jdk-7u71-linux-x64.tar.gz** will be downloaded to your system.

Step 2. Extract the file using the below command

```
tar xzf jdk-7u71-linux-x64.tar.gz
```

Step 3. To make java available for all the users of UNIX move the file to /usr/local and set the path. In the prompt switch to root user and then type the command below to move the jdk to /usr/lib.

```
mv jdk1.7.0_71 /usr/lib/
```

Now **in ~/.bashrc file** add the following commands to set up the path.

Command for open .bashrc file :

```
gedit ~/.bashrc
```

Add the below lines:

```
export JAVA_HOME=/usr/lib/jdk1.7.0_71
export PATH=PATH:$JAVA_HOME/bin
```

Now, you can check the installation by typing "java -version" in the prompt.

2.SSH Installation

SSH is used to interact with the master and slaves computer without any prompt for password. First of all create a Hadoop user on the master and slave systems

Adding a dedicated Hadoop user

```
dell@dell-Inspiron-3542: ~
dell@dell-Inspiron-3542:~$ sudo addgroup hadoop
[sudo] password for dell:
Adding group `hadoop' (GID 1001) ...
Done.
dell@dell-Inspiron-3542:~$ sudo adduser --ingroup hadoop hduser
Adding user `hduser' ...
Adding new user `hduser' (1001) with group `hadoop' ...
Creating home directory `/home/hduser' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hduser
Enter the new value, or press ENTER for the default
    Full Name []:
    Room Number []:
    Work Phone []:
    Home Phone []:
    Other []:
Is the information correct? [Y/n] y
```

ssh has two main components:

1. **ssh** : The command we use to connect to remote machines - the client.
2. **sshd** : The daemon that is running on the server and allows clients to connect to the server.

The **ssh** is pre-enabled on Linux, but in order to start **sshd** daemon, we need to install **ssh** first. Use this command to do that :

```
dell@dell-Inspiron-3542: ~
Home Phone []:
Other []:
Is the information correct? [Y/n] y
dell@dell-Inspiron-3542:~$ sudo apt-get install ssh
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following extra packages will be installed:
  libck-connector0 ncurses-term openssh-client openssh-server
  openssh-sftp-server ssh-import-id
Suggested packages:
  libpam-ssh keychain monkeysphere rssh molly-guard
The following NEW packages will be installed:
  libck-connector0 ncurses-term openssh-server openssh-sftp-server ssh
  ssh-import-id
The following packages will be upgraded:
  openssh-client
1 upgraded, 6 newly installed, 0 to remove and 319 not upgraded.
Need to get 1,231 kB of archives.
After this operation, 3,614 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
WARNING: The following packages cannot be authenticated!
  libck-connector0 ncurses-term ssh-import-id
Install these packages without verification? [y/N] y
```

This will install ssh on our machine. If we get something similar to the following, we can think it is setup properly:

```
dell@dell-Inspiron-3542:~$ which ssh
/usr/bin/ssh
dell@dell-Inspiron-3542:~$ which sshd
/usr/sbin/sshd
dell@dell-Inspiron-3542:~$
```

Create and Setup SSH Certificates

Hadoop requires SSH access to manage its nodes, i.e. remote machines plus our local machine. For our single-node setup of Hadoop, we therefore need to configure SSH access to localhost.

So, we need to have SSH up and running on our machine and configured it to allow SSH public key authentication.

Hadoop uses SSH (to access its nodes) which would normally require the user to enter a password. However, this requirement can be eliminated by creating and setting up SSH certificates using the following commands. If asked for a filename just leave it blank and press the enter key to continue.

```
dell@dell-Inspiron-3542:~$ su huser
Password:
huser@dell-Inspiron-3542:/home/dell$
```

```
dell@dell-Inspiron-3542:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/dell/.ssh/id_rsa):
Created directory '/home/dell/.ssh'.
Your identification has been saved in /home/dell/.ssh/id_rsa.
Your public key has been saved in /home/dell/.ssh/id_rsa.pub.
The key fingerprint is:
3a:fe:67:d7:ce:b8:d8:9b:ca:53:21:c2:e4:e2:b5:3d dell@dell-Inspiron-3542
The key's randomart image is:
+---[RSA 2048]-----+
|
|      .
|     +
|    . = . .
|   .oS+ . .
|  ... E .
|   o   o .
|  . . .+o.+
|  ...o++*+o
|
+-----+
dell@dell-Inspiron-3542:~$
```

```
dell@dell-Inspiron-3542:~$ su huser
Password:
```

```
huser@dell-Inspiron-3542:/$ /home/k$ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

The second command adds the newly created key to the list of authorized keys so that Hadoop can use ssh without prompting for a password.

We can check if ssh works:

```
huser@dell-Inspiron-3542:/home/dell$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is 21:c5:a3:5b:0c:65:33:cc:6d:4d:e6:f1:10:06:a6:a8.
Are you sure you want to continue connecting (yes/no)? yes
```


3 Hadoop Installation

Now install the freely available hadoop version from any site (I downloaded 2.6.0)

Download the Hadoop 2.6.0 version from the mirror downloads

`sudo wget http://mirrors.sonic.net/apache/hadoop/common/hadoop-2.6.0/hadoop-2.6.0.tar.gz`

Untar the downloaded package using the command: `$ tar xvfz hadoop-2.6.0.tar.gz`

```
dell@dell-Inspiron-3542:~$ su hduser
Password:
hduser@dell-Inspiron-3542:/home/dell$ cd hadoop-2.6.2
hduser@dell-Inspiron-3542:/home/dell/hadoop-2.6.2$ sudo mv * /usr/local/hadoop
```

Now change directory to this folder using: `$ cd hadoop-2.6.0`

Now move all content of this directory to the `usr/local/hadoop`

If we got "**hduser is not in the sudoers file. This incident will be reported.**" Then do this

`hduser@dell-Inspiron-3542~/hadoop-2.6.0$ su dell`

Password:

`dell@l dell-Inspiron-3542~/home/hduser$ sudo adduser hduser sudo`

[sudo] password for dell:

Adding user `hduser' to group `sudo' ...

Adding user hduser to group sudo

Done.

Now, the **hduser** has root privilege, we can move the Hadoop installation to the `/usr/local/hadoop` directory without any problem.

```
dell@dell-Inspiron-3542:~$ su hduser
Password:
hduser@dell-Inspiron-3542:/home/dell$ cd hadoop-2.6.2
hduser@dell-Inspiron-3542:/home/dell/hadoop-2.6.2$ sudo mv * /usr/local/hadoop
```

```
hduser@dell-Inspiron-3542:/home/dell/hadoop-2.6.2$ sudo chown -R hduser:hadoop /usr/local/hadoop
hduser@dell-Inspiron-3542:/home/dell/hadoop-2.6.2$
```

Setup Configuration Files

The following files will have to be modified to complete the Hadoop setup:

1. ~/.bashrc
2. /usr/local/hadoop/etc/hadoop/hadoop-env.sh
3. /usr/local/hadoop/etc/hadoop/core-site.xml
4. /usr/local/hadoop/etc/hadoop/mapred-site.xml
5. /usr/local/hadoop/etc/hadoop/hdfs-site.xml

1. ~/.bashrc

```
hduser@dell-Inspiron-3542:/$ sudo gedit ~/.bashrc
```

Append the following to the end of ~/.bashrc:

```
#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#HADOOP VARIABLES END
```

2. /usr/local/hadoop/etc/hadoop/hadoop-env.sh

We need to set **JAVA_HOME** by modifying **hadoop-env.sh** file.

```
hduser@dell-Inspiron-3542:/$ sudo gedit /usr/local/hadoop/etc/hadoop/hadoop-env.sh
```

```
# The java implementation to use.
export JAVA_HOME=${JAVA_HOME}
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
```

Adding the above statement in the **hadoop-env.sh** file ensures that the value of **JAVA_HOME** variable will be available to Hadoop whenever it is started up.

3. /usr/local/hadoop/etc/hadoop/core-site.xml

The /usr/local/hadoop/etc/hadoop/core-site.xml file contains configuration properties that Hadoop uses when starting up.

This file can be used to override the default settings that Hadoop starts with.

```
hduser@dell-Inspiron-3542:~$ sudo mkdir -p /app/hadoop/tmp
```

```
hduser@dell-Inspiron-3542:~$ sudo chown hduser:hadoop /app/hadoop/tmp
```

Open the file and enter the following in between the <configuration></configuration> tag:

```
hduser@dell-Inspiron-3542:/$ sudo gedit /usr/local/hadoop/etc/hadoop/core-site.xml
```

```
core-site.xml x
<?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>hadoop.tmp.dir</name>
  <value>/app/hadoop/tmp</value>
  <description>A base for other temporary directories.</description>
</property>

<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:54310</value>
  <description>The name of the default file system. A URI whose
scheme and authority determine the FileSystem implementation. The
uri's scheme determines the config property (fs.SCHEME.impl) naming
the FileSystem implementation class. The uri's authority is used to
determine the host, port, etc. for a filesystem.</description>
</property>
</configuration>
```

4. /usr/local/hadoop/etc/hadoop/mapred-site.xml

By default, the /usr/local/hadoop/etc/hadoop/ folder contains

/usr/local/hadoop/etc/hadoop/mapred-site.xml.template

file which has to be renamed/copied with the name **mapred-site.xml**:

```
hduser@dell-Inspiron-3542:~$ cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template
/usr/local/hadoop/etc/hadoop/mapred-site.xml
```

```
hduser@dell-Inspiron-3542:/$ sudo gedit /usr/local/hadoop/etc/hadoop/mapred-site.xml
```

The **mapred-site.xml** file is used to specify which framework is being used for MapReduce.

We need to enter the following content in between the `<configuration></configuration>` tag:

```
mapred-site.xml x
<?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>mapred.job.tracker</name>
  <value>localhost:54311</value>
  <description>The host and port that the MapReduce job tracker runs
at. If "local", then jobs are run in-process as a single map
and reduce task.
</description>
</property>
</configuration>
```

5. /usr/local/hadoop/etc/hadoop/hdfs-site.xml

The `/usr/local/hadoop/etc/hadoop/hdfs-site.xml` file needs to be configured for each host in the cluster that is being used.

It is used to specify the directories which will be used as the namenode and the datanode on that host.

Before editing this file, we need to create two directories which will contain the namenode and the datanode for this Hadoop installation.

This can be done using the following commands:

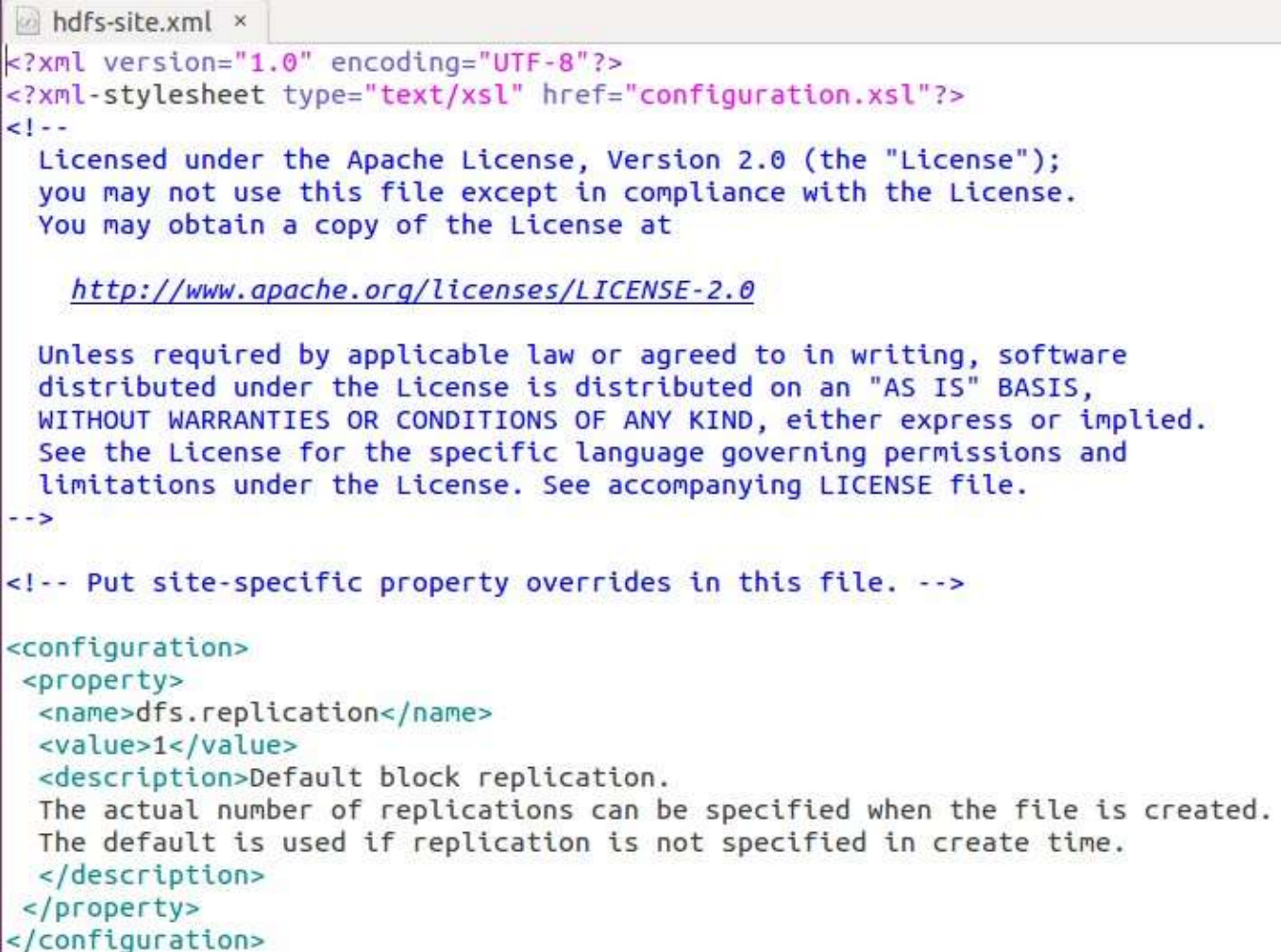
```
hduser@dell-Inspiron-3542:~$ sudo mkdir -p /usr/local/hadoop_store/hdfs/namenode
```

```
hduser@dell-Inspiron-3542:~$ sudo mkdir -p /usr/local/hadoop_store/hdfs/datanode
```


hduser@dell-Inspiron-3542:~\$ `sudo chown -R hduser:hadoop /usr/local/hadoop_store`

Open the file and enter the following content in between the <configuration></configuration> tag:

```
hduser@dell-Inspiron-3542:/$ sudo gedit /usr/local/hadoop/etc/hadoop/hdfs-site.xml
```



```
hdfs-site.xml x
<?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>
    <description>Default block replication.
      The actual number of replications can be specified when the file is created.
      The default is used if replication is not specified in create time.
    </description>
  </property>
</configuration>
```

Format the New Hadoop Filesystem

Now, the Hadoop file system needs to be formatted so that we can start to use it. The format command should be issued with write permission since it creates **current** directory under `/usr/local/hadoop_store/hdfs/namenode` folder:

`hduser@dell-Inspiron-3542:~$ hadoop namenode -format`

Note that **hadoop namenode -format** command should be executed once before we start using Hadoop. If this command is executed again after Hadoop has been used, it'll destroy all the data on the Hadoop file system.

Starting Hadoop

Now it's time to start the newly installed single node cluster.
We can use **start-all.sh** or (**start-dfs.sh** and **start-yarn.sh**)

```
hduser@dell-Inspiron-3542:~$ -cd /usr/local/hadoop/sbin
```

```
hduser@dell-Inspiron-3542:/usr/local/hadoop/sbin$ sudo su hduser
```

```
hduser@dell-Inspiron-3542:/usr/local/hadoop/sbin$ start-all.sh
```

```
hduser@dell-Inspiron-3542:/home/dell$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
16/01/31 02:34:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-dell-Inspiron-3542.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-dell-Inspiron-3542.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hduser-secondarynamenode-dell-Inspiron-3542.out
16/01/31 02:34:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-dell-Inspiron-3542.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-dell-Inspiron-3542.out
```

We can check if it's really up and running:

```
hduser@dell-Inspiron-3542:/home/dell$ jps
27723 SecondaryNameNode
27534 DataNode
27878 ResourceManager
27382 NameNode
28006 NodeManager
28303 Jps
hduser@dell-Inspiron-3542:/home/dell$
```

For Stopping Hadoop

```
hduser@dell-Inspiron-3542:/usr/local/hadoop/sbin$ stop-all.sh
```

If by chance your datanode or namenode is not starting, then you have to erase the contents of the folder /app/hadoop/tmp

STEP 1 stop hadoop

```
hduser@ dell-Inspiron-3542$ stop-all.sh
```

STEP 2 remove tmp folder

```
hduser@ dell-Inspiron-3542$ sudo rm -rf /app/hadoop/tmp/
```

STEP 3 create /app/hadoop/tmp/

```
hduser@ dell-Inspiron-3542$ sudo mkdir -p /app/hadoop/tmp
```

```
hduser@ dell-Inspiron-3542$ sudo chown hduser:hadoop /app/hadoop/tmp
```

```
hduser@ dell-Inspiron-3542$ sudo chmod 750 /app/hadoop/tmp
```

STEP 4 format namenode

```
hduser@ dell-Inspiron-3542$ hdfs namenode -format
```

STEP 5 start dfs

```
hduser@ dell-Inspiron-3542$ start-all.sh
```

STEP 6 check jps

```
hduser@ dell-Inspiron-3542$ $ jps
```

```
11342 Jps
```

```
10804 DataNode
```

```
11110 SecondaryNameNode
```

```
10558 NameNode
```

Hadoop Web Interfaces

Let's start the Hadoop again and see its Web UI:

<http://localhost:50070/> - web UI of the NameNode daemon

Overview 'localhost:54310' (active)

Started:	Mon Feb 01 00:34:37 IST 2016
Version:	2.6.2, r0cfd050febe4a30bIee1551dcc527589509fb681
Compiled:	2015-10-22T00:42Z by jenkins from (detached from 0cfd050)
Cluster ID:	CID-001dff18-0a4b-473a-8467-7f2a4ebf2605
Block Pool ID:	BP-103052493-127.0.1.1-1450598739659

Summary

Security is off.
Safemode is off.
23 files and directories, 12 blocks = 35 total filesystem object(s).
Heap Memory used 83.66 MB of 164.5 MB Heap Memory. Max Heap Memory is 889 MB.
Non Heap Memory used 38.82 MB of 39.94 MB Committed Non Heap Memory. Max Non Heap Memory is 214 MB.

Configured Capacity:	29.91 GB
-----------------------------	----------