Hadoop, R & RStudio Installation Tutorial

RTAIB - 2015

Department of Computer Science and Engineering,
National Institute of Technology Goa,
Farmagudi, Ponda, Goa-403401
Website: www.nitgoa.ac.in

Hadoop, R and RStudio Installation Tutorial

This tutorial explains how to install the following software on Ubuntu:

a) Java
b) ssh
c) Hadoop
d) R
e) RStudio

On your desktop, we have downloaded:
1) Hadoop binaries tar file:
http://mirror.nexcess.net/apache/hadoop/common/hadoop-2.7.1/hadoop-2.7.1.tar.gz
You may download from any mirror site at:
http://www.apache.org/dyn/closer.cgi/hadoop/common/

2) RStudio installer:
You may download it from https://www.rstudio.com/products/rstudio/download/

Conventions followed in this tutorial:

- All the commands are to be run in the Terminal application on your Ubuntu system.
- Comments are in bold and will begin with #. Please do not type comments in the Terminal window.
- Commands to be typed are in bold and DO NOT begin with #
- Expected output is in non-bold font. Your output on your system should be similar to this.

System Setup

The NIT–Goa system administrator has created a dedicated Hadoop system user account with admin privileges as follows:

```
# Creates a new group in the filesystem
sudo addgroup hadoop
Adding group `hadoop' (GID 1002) ...
Done.

# Creates a new user and adds it to the group
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

#Adds the hduser to the sudo list
sudo adduser hduser sudo
```

Java Installation

To begin the installation process, please type the following in a new Terminal window:
[Please do not type sentences beginning with #]

```
#Move to home directory
cd ~

#Update the available package list
sudo apt-get update

#Install the default java packages
sudo apt-get install default-jdk

#Check for the system version of java
java -version
java version "1.7.0_76"
Java(TM) SE Runtime Environment (build 1.7.0_76-b13)
 Java HotSpot(TM) 64-Bit Server VM (build 24.76-b04,
  mixed mode)
```

SSH Installation

```
#Install ssh
sudo apt-get install ssh

#Check if ssh is installed as expected
which ssh
/usr/bin/ssh
```

SSH Configuration

Hadoop requires SSH access to manage its nodes (remote and local machines). For our single-node setup of Hadoop, we therefore need to configure SSH access to localhost (127.0.0.1). So, we need to have SSH up and running on our machine and configured it to allow SSH public key authentication.

```
#Generate a SSH key pair with empty password
ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key
(/Users/vidya/.ssh/id_rsa):
 Your identification has been saved in
/Users/vidya/.ssh/id_rsa.
 Your public key has been saved in
/Users/vidya/.ssh/id_rsa.pub.
The key fingerprint is:
68:08:fc:ec:4e:4a:ff:dd:84:92:6c:91:2d:fc:73:69
 vidya@portugal.local
The key's randomart image is:
+--[ RSA 2048]----+
|                 |
| .               |
|  o              |
|   + o +         |
|    + B S        |
|   . o = . .     |
|  . o = + E      |
| . = . o *       |
|  . o.. . .      |
+-----------------+


#Enable SSH access to your local machine
cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

Hadoop Installation

```
#Move to ~/Desktop (or when you downloaded hadoop file)
cd ~/Desktop

#Create a folder called hadoop
sudo mkdir -p /usr/local/hadoop


#Unzip the downloaded Hadoop zip file
tar xvzf hadoop-2.7.1.tar.gz

#cd into the unzipped hadoop folder
cd hadoop

#Copies all unzipped files to /usr/local/hadoop
sudo mv * /usr/local/hadoop

#Change the ownership to hduser and group to hadoop
sudo chown -R hduser:hadoop /usr/local/hadoop
```

Hadoop is now installed on the system, however we need to modify the following system files to complete the Hadoop setup:

i. ~/.bashrc
ii. /usr/local/hadoop/etc/hadoop/hadoop-env.sh
iii. /usr/local/hadoop/etc/hadoop/core-site.xml
iv. /usr/local/hadoop/etc/hadoop/mapred-site.xml.template
v. /usr/local/hadoop/etc/hadoop/hdfs-site.xml

```
#Find the path to java installed on the system
update-alternatives --config java
There is only one alternative in link group java (providing /usr/bin/java): /usr/lib/jvm/java-7-openjdk-amd64/jre/bin/java
Nothing to configure.
```

.bashrc

```
#1.Open the .bashrc file for editing
gedit ~/.bashrc

#Append the following to the end of the .bashrc file
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"


# Re-execute the file contents in current shell
source ~/.bashrc
```

hadoop-env.sh

```
#2.Open hadoop-env.sh file
gedit /usr/local/hadoop/etc/hadoop/hadoop-env.sh

#Append the following at the end of hadoop-env.sh
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
```

core-site.xml

```
#3. Edit the Hadoop Config file at core-site.xml
gedit /usr/local/hadoop/etc/hadoop/core-site.xml

#Add the following to core-site.xml
<configuration>
 <property>
  <name>hadoop.tmp.dir</name>
  <value>/app/hadoop/tmp</value>
</property>

 <property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:54310</value>
</property>
</configuration>
```

mapred-site.xml.template

```
#4. Rename the file mapred-site.xml.template
cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template
/usr/local/hadoop/etc/hadoop/mapred-site.xml

#Open the mapred-site.xml
gedit /usr/local/hadoop/etc/hadoop/mapred-site.xml

#Add the following to mapred-site.xml
<configuration>
 <property>
  <name>mapred.job.tracker</name>
  <value>localhost:54311</value>
</property>
</configuration>
```

hdfs-site.xml

```
#5. Configure HDFS
#Create a new directory to contain namenode info
sudo mkdir -p /usr/local/hadoop_store/hdfs/namenode

#Create a new directory to contain datanode info
sudo mkdir -p /usr/local/hadoop_store/hdfs/datanode

#Change the ownership to hduser and group to hadoop
sudo chown -R hduser:hadoop /usr/local/hadoop_store

#Open the hdfs-site.xml
gedit /usr/local/hadoop/etc/hadoop/hdfs-site.xml

#Add the following to hdfs-site.xml
<configuration>
 <property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
 <property>
   <name>dfs.namenode.name.dir</name>
    <value>file:/usr/local/hadoop_store/hdfs/namenode
    </value>
 </property>
 <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/usr/local/hadoop_store/hdfs/datanode
    </value>
 </property>
</configuration>
```

R Installation

```
#Install the R packages
sudo apt-get install r-base
```

RStudio Installation

Double-Click on the RStudio Installed downloaded on your Desktop and follow the setup-instructions.

You may download it from https://www.rstudio.com/products/rstudio/download/