Apache Hadoop 2.0
Installation and Single Node Cluster
Configuration on Ubuntu

A guide to install and setup Single-Node Apache Hadoop 2.0 Cluster

edureka!
11/12/2013
APACHE HADOOP 2.0 INSTALLATION AND SINGLE NODE CLUSTER CONFIGURATION

A guide to install and setup Single-Node Apache Hadoop 2.0 Cluster

Table of Contents
Introduction .................................................................................................................. 2
1. Setting up the Ubuntu Server ............................................................................... 3
   1.1 Creating an Ubuntu VMPlayer instance......................................................... 3
      1.1.1 Download the VMWare image .................................................................. 3
      1.1.2 Open the image file ................................................................................ 3
      1.1.3 Play the Virtual Machine ......................................................................... 5
      1.1.4 Update the OS packages and their dependencies ................................. 7
      1.1.5 Install the Java for Hadoop 2.2.0 ............................................................ 7
   1.2 Download the Apache Hadoop 2.0 binaries ................................................. 7
      1.2.1 Download the Hadoop package ............................................................... 7
2. Configure the Apache Hadoop 2.0 Single Node Server ..................................... 9
   2.1 Update the Configuration files ..................................................................... 9
      2.1.1 Update “.bashrc” file for user ‘ubuntu’ .................................................... 9
   2.2 Setup the Hadoop Cluster .......................................................................... 11
      2.2.1 Configure JAVA_HOME .......................................................................... 11
      2.2.2 Create namenode and datanode directory ........................................... 12
      2.2.3 Configure the Default Filesystem ......................................................... 12
      2.2.4 Configure the HDFS ............................................................................. 13
      2.2.5 Configure YARN framework ................................................................. 14
      2.2.6 Configure MapReduce framework ......................................................... 15
      2.2.7 Start the DFS services .......................................................................... 16
      2.2.7 Perform the Health Check ...................................................................... 18
Introduction
This setup and configuration document is a guide to setup a Single-Node Apache Hadoop 2.0 cluster on an Ubuntu virtual machine on your PC. If you are new to both Ubuntu and Hadoop, this guide comes handy to quickly setup a Single-Node Apache Hadoop 2.0 Cluster on Ubuntu and start your Big Data and Hadoop learning journey.

The guide describes the whole process in two parts:

**Section 1: Setting up the Ubuntu OS for Hadoop 2.0**
This section describes step by step guide to download, configure an Ubuntu Virtual Machine image in VMPlayer, and provides steps to install pre-requisites for Hadoop Installation on Ubuntu.

**Section 2: Installing Apache Hadoop 2.0 and Setting up the Single Node Cluster**
This section explains primary Hadoop 2.0 configuration files, Single-Node cluster configuration and Hadoop daemons start and stop process in detail.
1. Setting up the Ubuntu Server
This section describes the steps to download and create an Ubuntu image on VMPlayer.

1.1 Creating an Ubuntu VMPlayer instance
The first step is to download an Ubuntu image and create an Ubuntu VMPlayer instance.

1.1.1 Download the VMware image
Access the following link and download the 12.0.4 Ubuntu image:

http://www.traffictool.net/vmware/ubuntu1204t.html

1.1.2 Open the image file
Extract the Ubuntu VM image and open it in VMware Player.

Click open virtual machine and select path where you have extracted the image.

Select the `.vmx` file and click `ok`. 
FIGURE 1-1 OPEN THE VM IMAGE

[Image: Open Virtual Machine window with files and directories]

This product is not licensed and is authorized for non-commercial use only. For commercial use, purchase a license. Buy now.
1.1.3 Play the Virtual Machine
You would see the below screen in VMware Player after the VM image creation completes.

**Figure 1-2 Play the virtual machine**

![Virtual Machine Screen in VMware Player](image-url)
Double click on the link.

You will get the home screen with the following image.

**Figure 1-3 Ubuntu Home Screen**

The user details for the Virtual instance is:

**Username** : user

**Password** : password

Open the terminal to access the file system.
1.1.4 Update the OS packages and their dependencies
The first task is to run `apt-get update` to download the package lists from the repositories and "update" them to get information on the newest versions of packages and their dependencies.

```bash
$ sudo apt-get update
```

1.1.5 Install the Java for Hadoop 2.2.0
Use `apt-get` to install the JDK 6 on the server.

```bash
$ sudo apt-get install openjdk-6-jdk
```

1.2 Download the Apache Hadoop 2.0 binaries
1.2.1 Download the Hadoop package
Download the binaries to your home directory. Use the default user 'user' for the installation.

In live production instances a dedicated Hadoop user account for running Hadoop is used. Though, it’s not mandatory to use a dedicated Hadoop user account but is recommended because this helps to separate the Hadoop installation from other software applications and user accounts running on the same machine (separating for security, permissions, backups, etc.).
$wget http://apache.mirrors.lucidnetworks.net/hadoop/common/stable2/hadoop-2.2.0.tar.gz

**Figure 1-6. Download Hadoop 2.2.0**

```bash
user@ubuntu:-$ wget http://apache.mirrors.lucidnetworks.net/hadoop/common/stable2/hadoop-2.2.0.tar.gz
Resolving apache.mirrors.lucidnetworks.net (apache.mirrors.lucidnetworks.net)... 108.166.161.136
Connecting to apache.mirrors.lucidnetworks.net (apache.mirrors.lucidnetworks.net)|108.166.161.136|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 109229873 (104M) [application/x-gzip]
Saving to: 'hadoop-2.2.0.tar.gz'
0% [ ] 805,868 177K/s eta 10m 51s
```

User@ubuntu:-$ ls
Desktop  Downloads  hadoop-2.2.0.tar.gz  Pictures  Templates
Documents examples.desktop  Music  Public  Videos

Unzip the files and review the package content and configuration files.

$tar -xvf hadoop-2.2.0.tar.gz

**Figure 1-7. Hadoop Package Content**

```bash
user@ubuntu:-$ ls
Desktop  Downloads  hadoop-2.2.0  Music  Public  Videos
Documents examples.desktop  hadoop-2.2.0.tar.gz  Pictures  Templates
user@ubuntu:-$
```
Review the Hadoop configurations files.

After creating and configuring your virtual servers, the Ubuntu instance is now ready to start installation and configuration of Apache Hadoop 2.0 Single Node Cluster. This section describes the steps in details to install Apache Hadoop 2.0 and configure a Single-Node Apache Hadoop cluster.

2. Configure the Apache Hadoop 2.0 Single Node Server
   This section explains the steps to configure the Single Node Apache Hadoop 2.0 Server on Ubuntu.

2.1 Update the Configuration files
2.1.1 Update “.bashrc” file for user ‘ubuntu’.
   Move to ‘user’ $HOME directory and edit ‘.bashrc’ file.
Update the `.bashrc` file to add important Apache Hadoop environment variables for user.

a) Change directory to home.

   $ cd

b) Edit the file

   $ vi .bashrc

-----------------------------------------------Set Hadoop environment Variables - Begin-----------------------------------------------

# Set Hadoop-related environment variables
export HADOOP_HOME=$HOME/hadoop-2.2.0
export HADOOP_CONF_DIR=$HOME/hadoop-2.2.0/etc/hadoop
export HADOOP_MAPRED_HOME=$HOME/hadoop-2.2.0
export HADOOP_COMMON_HOME=$HOME/hadoop-2.2.0
export HADOOP_HDFS_HOME=$HOME/hadoop-2.2.0
export YARN_HOME=$HOME/hadoop-2.2.0

# Set JAVA_HOME (we will also configure JAVA_HOME for Hadoop execution later on)
export JAVA_HOME=/usr/lib/jvm/java-6-openjdk-amd64
# Add Hadoop bin/ directory to PATH

```bash
export PATH=$PATH:$HOME/hadoop-2.2.0/bin
```

---------------
Set Hadoop environment Variables – End
---------------

---

**FIGURE 2-2 EDIT .bashrc**

```bash
# Set Hadoop-related environment variables
export HADOOP_HOME=$HOME/hadoop-2.2.0
export HADOOP_MAPRED_HOME=$HOME/hadoop-2.2.0
export HADOOP_COMMON_HOME=$HOME/hadoop-2.2.0
export HADOOP_HDFS_HOME=$HOME/hadoop-2.2.0
export YARN_HOME=$HOME/hadoop-2.2.0
export HADOOP_CONF_DIR=$HOME/hadoop-2.2.0/etc/hadoop

# Set JAVA_HOME (we will also configure JAVA_HOME for Hadoop execution later on)
export JAVA_HOME=/usr/lib/jvm/java-6-openjdk-amd64

# Add Hadoop bin/ directory to PATH
export PATH=$PATH:$HOME/hadoop-2.2.0/bin
```

c) Source the .bashrc file to set the hadoop environment variables without having to invoke a new shell:

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

Execute the all the steps of this section on all the remaining cluster servers.

### 2.2 Setup the Hadoop Cluster

This section describes the detail steps needed for setting up the Hadoop Cluster and configuring the core Hadoop configuration files.

#### 2.2.1 Configure JAVA_HOME

Configure JAVA_HOME in `hadoop-env.sh`. This file specifies environment variables that affect the JDK used by Apache Hadoop 2.0 daemons started by the Hadoop start-up scripts:

```bash
$cd $HADOOP_CONF_DIR
$vi hadoop-env.sh
```

*Update the JAVA_HOME to:*

```
export JAVA_HOME=/usr/lib/jvm/java-6-openjdk-amd64
```
2.2.2 Create NameNode and DataNode directory
Create DataNode and NameNode directories to store HDFS data.

$ mkdir -p $HOME/hadoop2_data/hdfs/namenode
$ mkdir -p $HOME/hadoop2_data/hdfs/datanode

2.2.3 Configure the Default File system
The 'core-site.xml' file contains the configuration settings for Apache Hadoop Core such as I/O settings that are common to HDFS, YARN and MapReduce. Configure default file-system (Parameter: fs.default.name) used by clients in core-site.xml

$cd $HADOOP_CONF_DIR
$vi core-site.xml

Add the following line in between the configuration tag:

```
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```
Where hostname and port are the machine and port on which Name Node daemon runs and listens. It also informs the Name Node as to which IP and port it should bind. The commonly used port is 9000 and you can also specify IP address rather than hostname.

**Note**
For the simplicity of understanding the cluster setup, we have updated changed only necessary parameters to start a cluster. You can research more on Apache Hadoop 2.0 page and experiment the configuration for different features.

### 2.2.4 Configure the HDFS

This file contains the configuration settings for HDFS daemons; the Name Node and the data nodes.

Configure `hdfs-site.xml` and specify default block replication, and NameNode and DataNode directories for HDFS. 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.

```
$cd $HADOOP_CONF_DIR
$vi hdfs-site.xml
```

Add the following line in between the configuration tag:

```xml
<configuration>

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

</configuration>
```
Configure the default filesystem

```xml
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/user/hadoop-2.2.0/hadoop2_data/hdfs/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/user/hadoop-2.2.0/hadoop2_data/hdfs/datanode</value>
  </property>
</configuration>
```

2.2.5 Configure YARN framework
This file contains the configuration settings for YARN; the NodeManager.

```bash
$cd $HADOOP_CONF_DIR

$vi yarn-site.xml
```

Add the following line in between the configuration tag:
2.2.6 Configure MapReduce framework
This file contains the configuration settings for MapReduce.

Configure mapred-site.xml and specify framework details.

$cd $HADOOP_CONF_DIR
$vi mapred-site.xml

Add the following line in between the configuration tag:
2.2.6 Start the DFS services

The first step in starting up your Hadoop installation is formatting the Hadoop file-system, which is implemented on top of the local file-systems of your cluster. This is required on the first time Hadoop installation. Do not format a running Hadoop file-system, this will cause all your data to be erased.

To format the file-system, run the command:

```
$ hadoop namenode -format
```

You are now all set to start the HDFS services i.e. Name Node, Resource Manager, Node Manager and Data Nodes on your Apache Hadoop Cluster.
Figure 2-8 Start the Services

Start the YARN daemons i.e. Resource Manager and Node Manager. Cross check the service start-up using JPS (Java Process Monitoring Tool).

Figure 2-9 Start the YARN Daemons

Start the History server.
2.2.7 Perform the Health Check

a) Check the NameNode status:

http://localhost:50070/dfshealth.jsp
b) JobHistory status:

http://localhost:19888/jobhistory.jsp

**FIGURE 2-12 JOBHISTORY STATUS**