Haoop安装\_单机/伪分布式配置

# 1 支持平台

GNU/Linux是产品开发和运行的平台。 Hadoop已在有2000个节点的GNU/Linux主机组成的集群系统上得到验证。

Win32平台是作为开发平台支持的。由于分布式操作尚未在Win32平台上充分测试，所以还不作为一个生产平台被支持。

# 2 版本号

Hadoop：2.6.0

Ubuntu：16.04

# 3 准备工作

在Windows系统上安装VMware虚拟机，下载Ubuntu.iso镜像文件。在VMware中创建新虚拟机，选择“安装程序光盘镜像文件”，选择下载好的iso文件，根据提示填写相应信息，选择虚拟机安装位置，选择最大磁盘大小(默认20GB)，单击完成创建虚拟机，并开始安装Ubuntu64位和VMware Tools(非必须，但安装后可以实现更丰富的功能，比如说全屏）

# 4 具体安装与配置

(注意：在以下指令中，以#开头的代表是注释，每条指令的两个相邻命令之间有一个空格)

## 4.1 创建hadoop用户

1. ctrl+alt+t打开终端窗口，用如下命令创建新用户：

sudo useradd -m hadoop -s /bin/bash

2. 为刚创建的用户设置密码

sudo passwd hadoop

3. 为hadoop用户增加管理员权限，方便部署

sudo adduser hadoop sudo

4. 注销当前用户，选择以hadoop用户进行登录

## 4.2 更新apt

用hadoop用户登录后，要先更新一下apt，为了方便使用apt安装软件

sudo apt-get update

## 4.3 安装ssh、配置ssh无密码登录

单点、集群模式都需要ssh登录(类似于远程登录，你可以登录某台Linux主机，并在上面运行命令)。Ubuntu以默认安装了SSH client，此处还需要安装SSH server:

sudo apt-get install openssh-server

安装后，可用如下命令登录本机：

ssh localhost

但这样每次登录时需要密码的，可以配置成SSH无密码登录。先退出刚刚的ssh，然后用ssh-keygen生成密钥，并将密钥加入到授权中

eixt #退出刚才的ssh localhost

cd ~/.ssh/

ssh-keygen -t rsa #会有提示，都按回车就可以

cat ./id\_rsa.put >>./authorized\_keys #加入授权

此时再用ssh localhost命令，无需输入密码就可以直接登录了

## 4.4 安装java环境

### 4.4.1 安装JDK

下载地址：http://www.oracle.com/technetwork/java/javase/downloads/index.html

选择合适的版本，这里选择的是1.7版本

### 4.4.2 下载eclipse

下载地址：http://www.eclipse.org/downloads/

注意选择for linux版本，为了与jdk版本保持一致，此处选择的是Luna

### 4.4.3 解压

转至某个目录下，对jdk和eclipse进行解压

cd /usr/local

tar -zxvf \*\*.tar.gz

### 4.4.4 配置环境变量

打开.bashrc文件

gedit ~/.bashrc

在打开的.bashrc文件的末尾添加几行(jdk解压目录)：

export JAVA\_HOME=/usr/local/jdk1.7.0\_25

export JRE\_HOME=${JAVA\_HOME}jre

export CLASSPATH=.:{JAVA\_HOME}/lib:${JRE\_HOME}/lib

export PATH=${JAVA\_HOME}/bin:$PATH

配置好后，重启系统，在打开终端，输入java -version，可显示java版本号，则证明配置成功

## 4.5. 安装Hadoop 2

Hadoo2可以通过 <http://mirror.bit.edu.cn/apache/hadoop/common/>下载。这里下载了2.6.0版本([hadoop-2.6.0.tar.gz](http://mirror.bit.edu.cn/apache/hadoop/common/hadoop-2.6.0/hadoop-2.6.0.tar.gz) )。

下载时强烈建议也下载 **hadoop-2.x.y.tar.gz.mds** 这个文件，该文件包含了检验值可用于检查 hadoop-2.x.y.tar.gz 的完整性，否则若文件发生了损坏或下载不完整，Hadoop 将无法正常运行。

下载完成后，使用以下指令判断是否完全下载  
cat ~/下载/hadoop-2.6.0.tar.gz.mds | grep 'MD5' *# 列出md5检验值*

md5sum ~/下载/hadoop-2.6.0.tar.gz | tr "a-z" "A-Z" *# 计算md5值，并转化为大写，方便比较*

如果两个值不一样，则代表下载文件不完整

将hadoop安装至/usr/local中

sudo tar -axf ~/Downloads/hadoop-2.6.0.tar.gz -C /usr/local

cd /usr/local

sudo mv ./hadoop-2.6.0/ ./hadoop #将文件夹改名为hadoop

sudo chown -R hadoop ./hadoop #修改文件权限

输入以下命令检验Hadoop是否可用

cd /usr/local/hadoop

./bin/hadoop version

# 5 Hadoop单机配置(非分布式)

Hadoop 默认模式为非分布式模式，无需进行其他配置即可运行。非分布式即单 Java 进程，方便进行调试。

现在我们可以执行例子来感受下 Hadoop 的运行。我们选择运行 grep 例子，我们将 input 文件夹中的所有文件作为输入，筛选当中符合正则表达式 dfs[a-z.]+ 的单词并统计出现的次数，最后输出结果到 output 文件夹中。

Shell命令：

1. cd /usr/local/hadoop
2. mkdir ./input
3. cp ./etc/hadoop/\*.xml ./input *# 将配置文件作为输入文件*
4. ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-\*.jar grep ./input ./output 'dfs[a-z.]+'
5. cat ./output/\* *# 查看运行结果*

执行成功后，会输出作业的相关信息，输出的结果是符合正则的单词dfsadmin出现了一次。

执行结果：1 dfsadmin

注意，Hadoop不会默认覆盖结果文件，因此再次运行上面实例会报错。需要先将./output删除

rm -r ./output

# 6 伪分布式配置

Hadoop 可以在单节点上以伪分布式的方式运行，Hadoop 进程以分离的 Java 进程来运行，节点既作为 NameNode 也作为 DataNode，同时，读取的是 HDFS 中的文件。

Hadoop 的配置文件位于 /usr/local/hadoop/etc/hadoop/ 中，伪分布式需要修改2个配置文件 **core-site.xml** 和 **hdfs-site.xml** 。Hadoop的配置文件是 xml 格式，每个配置以声明 property 的 name 和 value 的方式来实现。

修改配置文件core-site.xml

gedit ./etc/hadoop/core-site.xml

将当中的

<configuration>

</configuration>

修改为

<configuration>

<property>

<name>hadoop.tmp.dir</name>

<value>file:/usr/local/hadoop/tmp</value>

<description>Abase for other temporary directories.</description>

</property>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

同样的，修改配置文件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/tmp/dfs/name</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>file:/usr/local/hadoop/tmp/dfs/data</value>

</property>

</configuration>

**Hadoop配置文件说明**

Hadoop 的运行方式是由配置文件决定的（运行 Hadoop 时会读取配置文件），因此如果需要从伪分布式模式切换回非分布式模式，需要删除 core-site.xml 中的配置项。

此外，伪分布式虽然只需要配置 fs.defaultFS 和 dfs.replication 就可以运行（官方教程如此），不过若没有配置 hadoop.tmp.dir 参数，则默认使用的临时目录为 /tmp/hadoo-hadoop，而这个目录在重启时有可能被系统清理掉，导致必须重新执行 format 才行。所以我们进行了设置，同时也指定 dfs.namenode.name.dir 和 dfs.datanode.data.dir，否则在接下来的步骤中可能会出错。

配置完成后，执行NameNode格式化：

./bin/hdfs namenode -format

成功的话，会看到successfully formatted和Exitting with status 0

接着开启NameNode和DataNode守护进程

./sbin/start-dfs.sh

启动完成后，可通过命令jps来判断是否启动成功。如果成功启动，会出现如何进程：

NameNode DataNode SecondaryNameNode Jps

成功启动后，可以访问 Web 界面 [http://localhost:50070](http://localhost:50070/) 查看 NameNode 和 Datanode 信息，还可以在线查看 HDFS 中的文件。

# 7 运行Hadoop伪分布式实例

上面的单机模式，grep 例子读取的是本地数据，伪分布式读取的则是 HDFS 上的数据。要使用 HDFS，首先需要在 HDFS 中创建用户目录：

1. ./bin/hdfs dfs -mkdir -p /user/hadoop

接着将 ./etc/hadoop 中的 xml 文件作为输入文件复制到分布式文件系统中，即将 /usr/local/hadoop/etc/hadoop 复制到分布式文件系统中的 /user/hadoop/input 中。我们使用的是 hadoop 用户，并且已创建相应的用户目录 /user/hadoop ，因此在命令中就可以使用相对路径如 input，其对应的绝对路径就是 /user/hadoop/input:

./bin/hdfs dfs -mkdir input

./bin/hdfs dfs -put ./etc/hadoop/\*.xml input

复制完成后，可以通过如下命令查看文件列表：

./bin/hdfs dfs -ls input

伪分布式运行 MapReduce 作业的方式跟单机模式相同，区别在于伪分布式读取的是HDFS中的文件（可以将单机步骤中创建的本地 input 文件夹，输出结果 output 文件夹都删掉来验证这一点）。

./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-\*. jar grep input output 'dfs[a-z.]+'

查看运行结果的命令（查看的是位于 HDFS 中的输出结果）：

./bin/hdfs dfs -cat output/\*

我们也可以将运行结果取回到本地：

1. rm -r ./output *# 先删除本地的 output 文件夹（如果存在）*
2. ./bin/hdfs dfs -get output ./output *# 将 HDFS 上的 output 文件夹拷贝到本机*
3. cat ./output/\*

若要关闭Hadoop，运行

./sbin/stop-dfs.sh

下次启动hadoop时，无需进行NameNode的初始化，只要运行./sbin/start-dfs.sh就可以