Hadoop 2.5 完全分布式的安装、配置、运行

**安装前环境准备**

1. VM虚拟机

运行Linux操作系统（有很多选择，由于小编更熟悉ubuntu ,so 后面的所有安装均基于ubuntu 14.04）

1. 在VM虚拟机上安装ubuntu操作系统
2. 准备3个虚拟节点

这一步比较简单，如果你已经完成了第2步，那么只需将第二步的ubuntu系统完整复制2份即可(小编建议的系统的jdk安装好后在复制这样就可以省去重复安装)，为什么要安装3个Linux系统原理如下：根据hadoop集群的要求至少需要3个节点（其中一个master节点主要是用于运行hadoop程序中的namenode、secondorynamenode和jobtracker任务。用外两个结点均为slave结点，其中一个是用于冗余目的，如果没有冗余，就不能称之为hadoop了，所以模拟hadoop集群至少要有3个结点，如果电脑配置非常高，可以考虑增加一些其它的结点。slave结点主要将运行hadoop程序中的datanode和tasktracker任务）.

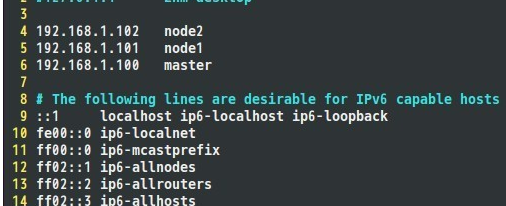
**下面将开始正式搭建hadoop集群**

1. 配置hostname和 hosts 文件。
2. 重命名主机名(vim /etc/hostname) 分别为 master node1 node2
3. 修改3个操作系统的hosts文件 vim /etc/hosts

master 192.168.10.115

node1 192.168.10.116

node2 192.168.10.117 ip地址可以根据个人需求修改



1. 建立hadoop 运行帐号

即为hadoop专门设置一个用户组和用户

sudo groupadd hadoop //添加hadoop 用户组

sudo useradd –s /bin/bash –d /home/Hadoop –m Hadoop –g hadoop –G admin //新建一个Hadoop 用户 属于hadoop 组 ，且具有admin权限。

sudo passwd Hadoop //为Hadoop 用户设置密码

su Hadoop //切换到Hadoop用户组

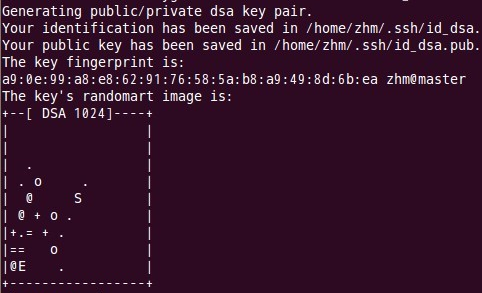
1. 配置ssh免密码链接（非常重要这一步，一定不能出错）

SSH主要是通过RSA算法来产生公钥和私钥 ，在数据传输过程中对数据进行加密处理。

1. 在每一个节点上生成公钥和私钥

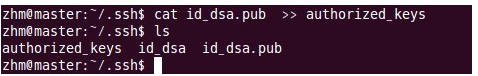
（**ubuntu**默认并没有安装**ssh**服务，如果通过ssh链接ubuntu，需要自己手动安装ssh-server。  sudo apt-get install openssh-server 安装完成后需要启动SSH: sudo /etc/init.d/ssh start）

输入命令：



以上命令是产生公私密钥，产生目录在用户主目录下的.ssh目录中，如下：

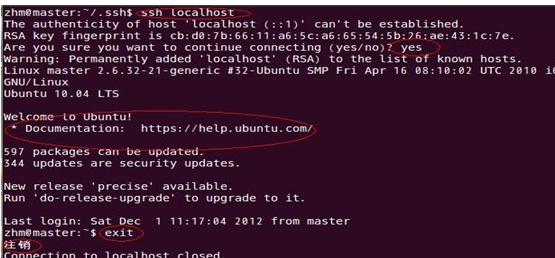
输入ls .ssh 可以看到Id\_dsa.pub为公钥，id\_dsa为私钥，紧接着将公钥文件复制成authorized\_keys文件，这个步骤是必须的，过程如下：



用上述同样的方法在剩下的两个结点中操作即可。

1. 单机回环ssh免密码登录测试

即在单机结点上用ssh进行登录，看能否登录成功。登录成功后注销退出，过程如下：



注意标红圈的指示，有以上信息表示操作成功，单点回环SSH登录及注销成功，这将为后续跨子结点SSH远程免密码登录作好准备。

用上述同样的方法在剩下的两个结点中操作即可。

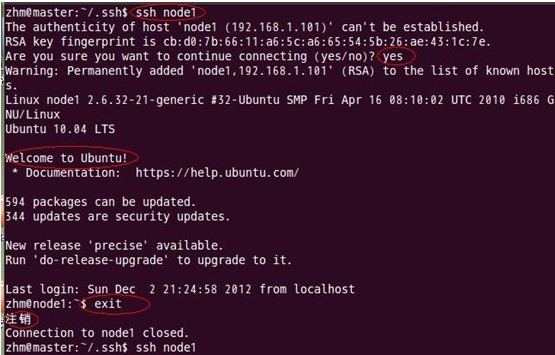
1. 让主节点master能ssh 免密码登录slave节点

为了实现这个功能，那么两个slave 节点的公钥文件中就必须包含主节点master的公钥文件，具体的实现步骤如下：

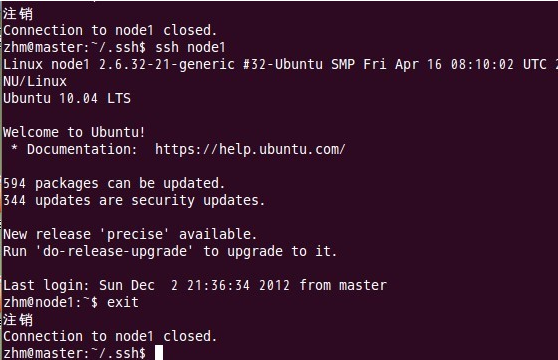
在node1 主机下：



完成上述过程后，进行如下测试：



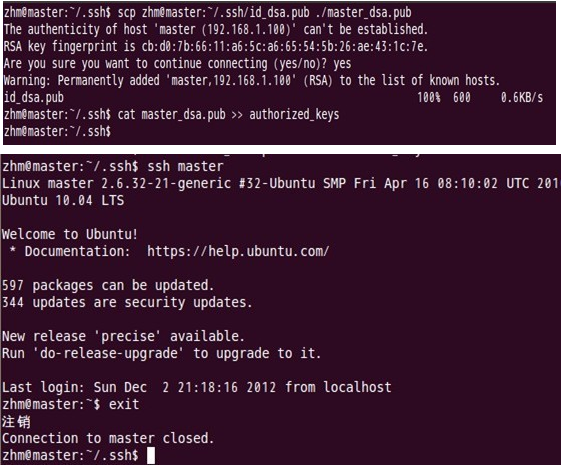
由上图可以看出，node1结点首次连接时需要，“YES”确认连接，这意味着master结点连接node1结点时需要人工询问，无法自动连接，输入yes后成功接入，紧接着注销退出至master结点。要实现ssh免密码连接至其它结点，还差一步，只需要再执行一遍ssh node1，如果没有要求你输入”yes”，就算成功了，过程如下：



上述步骤在node2节点上重复操作

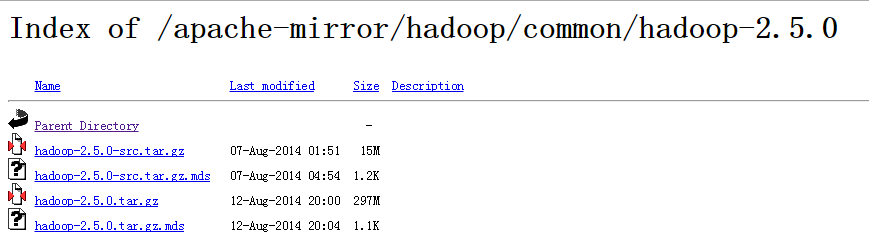
表面上看，这两个结点的ssh免密码登录已经配置成功，但是我们还需要对主结点master也要进行上面的同样工作，这一步有点让人困惑，但是这是有原因的，具体原因现在也说不太好，据说是真实物理结点时需要做这项工作，因为jobtracker有可能会分布在其它结点上，jobtracker有不存在master结点上的可能性。

对master自身进行ssh免密码登录测试工作：



1. 下载并解压hadoop （先在master节点进行安装和配置好后复制到slave节点即可）

我下载的是hadoop 2.5 地址：<http://apache.fayea.com/apache-mirror/hadoop/common/>



下载 hadoop-2.5.0.tar.gz 并解压 这里小编解压的路径是 /usr/hadoop 可根据个人需求选择

五、修改hadoop的配置文件/usr/hadoop/etc/hadoop/conf

1.在配置前先要在三个linux系统上安装JDK 这里我安装的是jdk1.8.0\_20

具体的安装在这里就不详细写出参见<http://www.aboutyun.com/thread-6750-1-1.html>

2．配置之前，需要在master本地文件系统创建以下文件夹：

~/dfs/name ~/dfs/dat ~/tmp

3.关闭系统防火墙 ufw disable (重启生效)

配置文件1:hadoop-env.sh

修改JAVA\_HOME值（export JAVA\_HOME=你的jdk安装位置）

配置文件2：yarn-env.sh

修改JAVA\_HOME值（export JAVA\_HOME=你的jdk安装位置）

配置文件3：slaves (这里面保存所有的slaves节点)

将node1 node2 写入

配置文件4：core-site.xml

<configuration>

<property>

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

<value>file:/home/hadoop/tmp</value>

<description>A base for other temporary directories</description>

</property>

<property>

<name>fs.default.name</name>

<value>hdfs://192.168.10.115:8020</value>

</property>

<property>

<name>hadoop.native.lib</name>

<value>true</value>

<description>Should native hadoop libraries, if present, be used.</description>

</property>

<property>

<name>mapreduce.jobhistory.address</name>

<value>192.168.10.115:10020</value>

</property>

</configuration>

配置文件5：hdfs-site.xml

<configuration>

<property>

<name>dfs.namenode.secondary.http-address</name>

<value>192.168.10.115:9001</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:/home/hadoop/dfs/name</value>

</property>

<property>

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

<value>file:/home/hadoop/dfs/data</value>

</property>

<property>

<name>dfs.replication</name>

<value>3</value>

</property>

<property>

<name>dfs.webhdfs.enabled</name>

<value>true</value>

</property>

<property>

<name>dfs.support.append</name>

<value>true</value>

</property>

</configuration>

配置文件6：mapred-site.xml

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

<property>

<name>mapred.system.dir</name>

<value>file:/home/hadoop/mapred/system</value>

<final>true</final>

</property>

<property>

<name>mapred.local.dir</name>

<value>file:/home/hadoop/mapred/local</value>

<final>true</final>

</property>

<property>

<name>mapreduce.jobhistory.address</name>

<value>192.168.10.115:10020</value>

</property>

<property>

<name>mapreduce.jobhistory.webapp.address</name>

<value>192.168.10.115:19888</value>

</property>

<property>

<name>mapred.remote.os</name>

<value>Linux</value>

<description>

Remote MapReduce framework's OS, can be either Linux or Windows

</description>

</property>

<property> <!—这个property在win7 eclipse下远程调用hadoop 时需要配置 具体原因不明 -->

<name>mapreduce.application.classpath</name>

<value>

/usr/hadoop/etc/hadoop,

/usr/hadoop/share/hadoop/common/\*,

/usr/hadoop/share/hadoop/common/lib/\*,

/usr/hadoop/share/hadoop/hdfs/\*,

/usr/hadoop/share/hadoop/hdfs/lib/\*,

/usr/hadoop/share/hadoop/mapreduce/\*,

/usr/hadoop/share/hadoop/mapreduce/lib/\*,

/usr/hadoop/share/hadoop/yarn/\*,

/usr/hadoop/share/hadoop/yarn/lib/\*

</value>

</property>

</configuration>

配置文件7：yarn-site.xml

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

<property>

<name>yarn.resourcemanager.address</name>

<value>192.168.10.115:8032</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.address</name>

<value>192.168.10.115:8030</value>

</property>

<property>

<name>yarn.resourcemanager.resource-tracker.address</name>

<value>192.168.10.115:8031</value>

</property>

<property>

<name>yarn.resourcemanager.admin.address</name>

<value>192.168.10.115:8033</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.address</name>

<value>192.168.10.115:8088</value>

</property>

1. 将master节点上配置好的Hadoop 复制到 slaves 节点上（记住放在相同路径下）

具体的操作如下：

在node1下

sudo scp -r /XXX/hadoop hadoop@node1:~ 先复制到node1的/home目录下

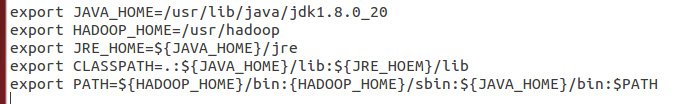
然后在移动到node1 的 /XXX/hadoop

在node2下

sudo scp -r /XXX/hadoop hadoop@node2:~ 先复制到node1的/home目录下

然后在移动到node1 的 /XXX/hadoop

1. 将hadoop 的bin 和 sbin加入系统环境变量



1. 这里是最后一步 ，启动验证

1、格式化namenode hadoop namenode –format

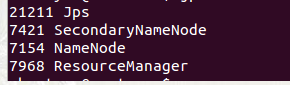
注：Hadoop偶尔会抽风 namenode datanode 会找不到了 目前我采用的是删除 所有节点~/dfs/name ~/dfs/dat ~/tmp 下的数据重新格式化 但是会清空hadoop的所有数据 真实环境不能使用 测试时可以

2、启动hadoop

start-all.sh //关闭是stop-all.sh

3、输入jps

如果看到入下 ，表示配置成功：

 namenode

 datanode