Hadoop学习笔记－伪分布式模式

信息爆炸的时代互联网上的数据正在以几何级数的速度增长。在此背景下，信息处理受到了更多关注。其中Google公司以MapReduce为基石，结合GFS、Bigtable 从TB乃至PB级数据中挖掘有用信息，对海量数据进行快捷、高效处理。Hadoop在机器学习、科学研究领域发挥重要作用，成为优秀的大数据平台。

# 配置环境：

操作系统：OS X 10.9.4；hadoop0.20.2版本

# 安装环境配置：

编辑/Applications/hadoop-0.20.2/conf/hadoop-env.sh文件:

***export JAVA\_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0\_05.jdk/Contents/Home***

即java安装根目录。

# 安装步骤：

## hadoop配置：

/Applications/hadoop-0.20.2/conf/core-site.xml，修改代码如下所示：

***<?xml version="1.0"?>***

***<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>***

***<!-- Put site-specific property overrides in this file. -->***

***<configuration>***

***<property>***

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

***<value>localhost:9000</value>***

***</property>***

***</configuration>***

/Applications/hadoop-0.20.2/conf/mapred-site.xml修改代码如下所示：

***<?xml version="1.0"?>***

***<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>***

***<!-- Put site-specific property overrides in this file. -->***

***<configuration>***

***<property>***

***<name>mapred.job.tracker</name>***

***<value>localhost:9001</value>***

***</property>***

***</configuration>***

/Applications/hadoop-0.20.2/conf/hdfs-site.xml修改代码如下所示：

***<?xml version="1.0"?>***

***<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>***

***<!-- Put site-specific property overrides in this file. -->***

***<configuration>***

***<property>***

***<name>dfs.replication</name>***

***<value>1</value>***

***</property>***

***</configuration>***

## 配置SSH:

首先在『系统偏好设置』的『共享』中，打开『远程登录』对『所有用户』有效。

打开终端输入来生成密钥对，按照默认生成密钥对：

***$ssh –keygen –t rsa***

这个命令会在.ssh文件夹下创建两个文件id\_rsa及id\_rsa.pub，这是SSH的一对私钥和公钥，类似于钥匙及锁。ssh-keygen代表生成密钥;-t(区分大小写)表示指定生成的密钥类型;rsa是rsa密钥认证的意思，即密钥类型;-P用于提供密语;-f指定生成的密钥文件。涉及SSH的一些知识。

在.ssh目录下：

***$cp id\_rsa.pub authorized\_keys***

把id\_rsa.pub(公钥)追加到授权的key里面去。在Hadoop的安装过程中，是否无密码登录是无关紧要的，但是如果不配置无密码登录，每次启动Hadoop，都需要输入密码以登录到每台机器的DataNode上，考虑到一般的Hadoop集群动辄数百台或上千台机器，因此一般来说都会配置SSH的无密码登录。

使用***$ssh localhost***测试配置是否成功。

## Hadoop运行：

进入/Applications/hadoop-0.20.2目录下:

***$bin/hadoop namenode –format***

格式化分布式文件系统。

启动hadoop守护进程：

***$bin/start-all.sh***

成功执行后启动NameNode, DataNode, JobTracker,TaskTracker, Secondary NameNode五个Java进程。

## 为Hadoop修改环境变量：

 (1)全局修改：

***$ vim /etc/paths***

 添加双引号内部分，“***/Applications/hadoop-0.20.2/bin***”

(2) 用户修改：

***$ vim ~/.profile***

添加双引号内部分，

“***export PATH=$PATH:/Applications/hadoop-0.20.2/bin***”

注：“/Applications/hadoop-0.20.2/bin”即hadoop安装目录下bin的位置。

## 问题解决：

配置hadoop-env.sh文件时，需要配置MAC中JAVA\_HOME环境变量，方法如下：

环境变量需要在etc目录下的profile文件中配置，是永久的配置。

***$   cd /etc   vi profile***

输入如下内容：

***export JAVA\_HOME=  =/Library/Java/JavaVirtualMachines/jdk1.8.0\_05.jdk/Contents/Home***

保存。然后重启或者注销，使环境变量的配置起作用。这样JAVA\_HOME的环境变量配置好了。

在基于unix/linux的操作系统下进行程序开发，使用环境变量将会方便。通过设置环境变量将可以在任意目录通过输入程序名来执行设定目录下的程序。不需要通过cd将工作目录改变到程序目录再执行程序。而且免去了输入"./"的麻烦。

具体步骤如下（在bash下）：

1、打开 应用程序 -> 实用工具 -> 终端；

2、在终端中定位到自己用户的主目录，输入： ***$cd ~*** ；

3、创建一个空文件，输入：***$touch .bash\_profile*** ；

4、编辑这个文件，输入：***$open .bash\_profile*** ；

5、添加环境变量JAVA\_HOME，可以输入：

***export JAVA\_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0\_05.jdk/Contents/Home***；

6、***$source .bash\_profile***重启终端。