## 配置HADOOP与JDK （严格遵守hadoop的分布式部署方法）

1. 准备工作

* hadoop安装包（版本2.4.0），要与spark相应的版本相对应；
* jdk1.7安装包（注意与系统的位数一致）；
* linux系统环境，由于都是基于java虚拟机，故并不依赖具体的发型linux版本， Ubuntu，centos等都可以。

1. 安装hadoop与JDK

* 将hadoop解压到相应目录：tar xvf hadoop-2.4.0.tar.gz –C *目地目录*；
* 将jdk解压到相应目录：tar xvf jdk-7u72-linux-i586.gz –C*目地目录*；

1. 配置环境变量

在$ HOME/.bashrc文件中进行配置，涉及的环境变量有：HADOOP\_HOME，JAVA\_HOME，CLASSPATH，JRE\_HOME，此外，如果虚拟机的内存小与1GB的情况下建议通过JAVA\_OPTS设置虚拟机的内存情况，实验样机的具体配置如下：

* *export JAVA\_HOME=/home/jomie/extra/jdk1.7*
* *export CLASSPATH=$JAVA\_HOME/lib/dt.jar:$JAVA\_HOME/lib/tools.jar*
* *export JRE\_HOME=$JAVA\_HOME/jre*
* *export HADOOP\_HOME=/home/jomie/extra/hadoop-2.4.0*
* *export JAVA\_OPTS="-Xms256m -Xmx512m"*
* *export PATH=$JAVA\_HOME/bin: $HADOOP\_HOME/bin*

1. 配置hadoop伪分布式

此处需要注意的是，由于hadoop不断更新，它的配置文件在几个大的版本中放置的目录和文件都有点不同，如果你手中的Hadoop版本不是2.4.0，建议参考Apache官网说明。

Hadoop2.4.0的配置文件放置在{$HADOOP}/etc/hadoop中，其中需要要配置的文件如下：

* **core-site.xml**

<configuration>

<property>

<name>fs.defaultFS</name>

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

</property>

</configuration>

* **hdfs-site.xml**

<configuration>

<property>

<name>dfs.data.dir</name> <!—配置hdfs存放数据目录—>

<value>/home/jomie/extra/hadoop-2.4.0/data</value>

</property>

<property>

<name>dfs.replication</name> <!—hdfs的备份策略 -->

<value>1</value>

</property>

<property>

<name>dfs.permissions</name>

<value>false</value>

</property>

<property>

<name>dfs.name.dir</name> <!—hdfs 元数据存放目录-->

<value>/home/jomie/extra/hadoop-2.4.0/metadata</value>

</property>

<property>

<name>dfs.webhdfs.enabled</name> <!—hdfs 是否支持web访问-->

<value>true</value>

<description>

Enable WebHDFS (REST API) in Namenodes and Datanodes.

</description>

</property>

</configuration>

* **yarn-site.xml （分布式资源管理）**

<configuration>

<!-- Site specific YARN configuration properties -->

<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>

</configuration>

* **mapred-site.xml**

<configuration>

<property>

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

<value>yarn</value>

</property>

</configuration>

1. 设置集群间无密码登陆方式

## 编译spark

1. 准备工作

* 安装Scala语言包（最好为2.10.4以上版本），并配置相应的环境变量，具体如下：

export SCALA\_HOME=/home/jomie/extra/scala-2.11.4

export SBT\_OPTS="-Xms256M -Xmx512M -Xss1M -XX:+CMSClassUnloadingEnabled -XX:MaxPermSize=128M" （设置虚拟机内存大小，可根据具体的系统硬件配置进行设置）

export PATH=$SCALA\_HOME/bin:$PATH

export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop (hadoop 配置文件目录)

* 安装maven编译工具，版本最好为3.1以上版本，并配置相应环境变量，具体如下：

export MAVEN\_HOME=/home/jomie/extra/apache-maven-3.2.3

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

1. 编译spark