**目录**

[配置前的准备工作 2](#_Toc37144022)

[第一步：创建Hadoop用户 2](#_Toc37144023)

[第二步：更新apt-get 2](#_Toc37144024)

[第三步：安装SSH server 2](#_Toc37144025)

[第四步：安装Java 2](#_Toc37144026)

[第五步：安装Hadoop 3](#_Toc37144027)

[伪分布式配置 3](#_Toc37144028)

[第一步：添加Java到Hadoop的环境变量中 3](#_Toc37144029)

[第二步：修改core-site.xml文件 4](#_Toc37144030)

[第三步：修改hdfs-site.xml文件 4](#_Toc37144031)

[第四步：启动Hadoop 4](#_Toc37144032)

[词频统计实例实现 5](#_Toc37144033)

[第一步：启动Hadoop 5](#_Toc37144034)

[第二步：创建Hadoop用户 5](#_Toc37144035)

[第三步：导入待统计文件 5](#_Toc37144036)

[第四步：进行词频统计 5](#_Toc37144037)

[第五步：输出统计结果 5](#_Toc37144038)

# 配置前的准备工作

## 第一步：创建Hadoop用户

目的：方便后续实验的操作，以及未来Hadoop分布式的搭建，故将用户名都命名为hadoop

1. sudo useradd -m hadoop -s /bin/bash 此步创建名为hadoop的用户
2. sudo passwd hadoop 此步为Hadoop用户创建密码
3. sudo adduser hadoop sudo 此步将Hadoop添加到管理员列表中，赋予权限

## 第二步：更新apt-get

目的：方便后续软件的安装等操作

1. sudo apt-get update

遇到困难：E: 无法获得锁 /var/lib/apt/lists/lock - open (11: 资源暂时不可用)

解决办法：sudo rm /var/lib/apt/lists/lock 此步将暴力删除锁，随后即可继续执行上述命令，之后遇到了类似的问题都是采取相同的方法解决

## 第三步：安装SSH server

目的：SSH是建立在应用层和传输层基础上，专为远程登录会话和其他网络服务提供安全性的协议，使用SSH协议可以有效防止远程管理过程中的信息泄露问题

1. sudo apt-get install openssh-server 此步安装SSH server
2. ssh localhost 此步登陆SSH server
3. cd ~/.ssh/ 此步切换目录到SSH server的安装目录下
4. ssh-keygen -t rsa 此步利用rsa算法加密后生成私钥和公钥储存到~/.ssh/目录下
5. cat ./id\_rsa.pub >> ./authorized\_keys 此步将公钥加入到授权列表中，此后登陆即可免密

## 第四步：安装Java

目的：方便后续实验的操作

1. 从甲骨文中国官网上下载jdk-8u221-linux-x64.tar.gz压缩包
2. mkdir /usr/lib/jvm 此步创建文件夹jvm存储解压后的jdk文件

遇到困难：mkdir: 无法创建目录”/usr/lib/jvm”: 权限不够

解决办法：sudo -i 此步改变了用户一些命令的使用权限，重复上述操作即可成功创建，后续可能会遇到相同的修改文件时权限不够的困难，均用相同的方式处理即可

1. sudo tar zxvf /home/yuba316/下载/jdk-8u221-linux-x64.tar.gz -C /usr/lib/jvm 此步将第一步中下载得到的压缩包解压到/usr/lib/jvm目录下，得到jdk1.8.0\_221
2. sudo apt-get install vim 此步安装vim以方便后续对环境变量的修改
3. vim ~/.bashrc 此步利用vim来编辑环境变量，将Java的安装路径添加到环境变量中：

export JAVA\_HOME=/usr/lib/jvm/jdk1.8.0\_221

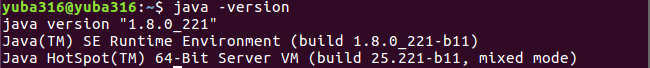
export JRE\_HOME=${JAVA\_HOME}/jre

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

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

1. source ~/.bashrc 此步更新环境变量使其生效
2. java -version 此步检查Java是否安装并配置成功

【图1】Java安装成功结果



## 第五步：安装Hadoop

目的：搭建伪分布式的环境

1. 到http://mirrors.hust.edu.cn/apache/hadoop/common/上去下载hadoop-2.9.2.tar.gz
2. sudo tar zxvf /home/yuba316/下载/ hadoop-2.9.2.tar.gz -C /usr/local 此步将上一步中下载得到的压缩包解压到/usr/local目录下，得到hadoop-2.9.2
3. vim ~/.bashrc 此步利用vim来编辑环境变量，将Hadoop的安装路径添加到环境变量中：

export HADOOP\_HOME=/usr/local/hadoop-2.9.2

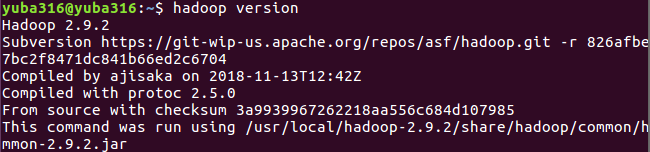
export CLASSPATH=$($HADOOP\_HOME/bin/hadoop classpath):$CLASSPATH

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export PATH=$PATH:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin

1. source ~/.bashrc 此步更新环境变量使其生效
2. hadoop version 此步检查Hadoop是否安装并配置成功

【图2】Hadoop安装成功结果



# 伪分布式配置

## 第一步：添加Java到Hadoop的环境变量中

目的：使得Hadoop的进程能够以分离的Java进程来运行

1. cd /usr/local/hadoop-2.9.2/etc/hadoop 此步进入hadoop环境文件所在路径
2. vim hadoop-env.sh 此步利用vim来编辑环境变量，将Java的安装路径添加到环境变量中：export JAVA\_HOME=/usr/lib/jvm/jdk1.8.0\_221

## 第二步：修改core-site.xml文件

目的：设置Hadoop重要临时文件的存储位置

1. gedit core-site.xml 此步修改核心配置内容

<configuration>

<property>

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

<value>file:/usr/local/hadoop-2.9.2/tmp</value>

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

</property>

<property>

<name>fs.defaultFS</name>

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

</property>

</configuration>

## 第三步：修改hdfs-site.xml文件

目的：设置dfs管理节点和数据节点的存储位置

1. get hdfs-site.xml 此步修改hdfs配置内容

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

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

<value>file:/usr/local/hadoop-2.9.2/tmp/dfs/name</value>

</property>

<property>

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

<value>file:/usr/local/hadoop-2.9.2/tmp/dfs/data</value>

</property>

</configuration>

## 第四步：启动Hadoop

1. cd /usr/local/hadoop-2.9.2 此步返回原始路径
2. ./bin/hdfs namenode -format 此步格式化管理节点，每次更改配置后都必须执行
3. ./sbin/start-dfs.sh 此步启动管理节点和数据节点进程

遇到困难：localhost: mkdir: 无法创建目录”/usr/local/hadoop-2.9.2/logs”: 权限不够

解决办法：sudo chown -R yuba316 /usr/local/hadoop-2.9.2 此步改变了用户一些命令的使用权限，重复上述操作即可

1. jps 此步查看管理节点和数据节点进程是否成功启动
2. stop-dfs.sh 此不关闭管理节点和数据节点进程

【图3】伪分布式配置成功结果



# 词频统计实例实现

## 第一步：启动Hadoop

1. start-dfs.sh 此步启动所有进程

## 第二步：创建Hadoop用户

1. hdfs dfs -mkdir -p /user/yuba316/hadoop

## 第三步：导入待统计文件

1. hdfs dfs -put ./etc/hadoop/\*.xml input 此步导入/etc/hadoop文件夹下所有的xml文件
2. hdfs dfs -ls input 此步查看导入的文件列表

## 第四步：进行词频统计

1. hdfs dfs -rm -r output 此步确定output文件夹不存在
2. hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-\*.jar wordcount input output 此步对导入的文件进行词频统计
3. hdfs dfs -cat output/\* 此步查看输出结果

## 第五步：输出统计结果

1. rm -r ./output 此步删除本地输出文件夹
2. hdfs dfs -get output ./output 此步将输出结果导出到本地输出文件夹中
3. cat ./output/\* 此步查看本地输出文件夹
4. hdfs dfs -rm -r output

rm -r ./output 此步删除输出结果

1. stop-dfs.sh 此步关闭所有进程

【图4】词频统计输出结果

