# 配置 Spark 集群的点点滴滴

## 1. 集群拓扑环境

| Hostname | IP | Role |
| --- | --- | --- |
| buaa | 192.168.108.211 | NameNode/DataNode/SecondaryNode<br>ResourceManager/NodeManger<br>Master/Worker |

## 2. 设置静态 IP 和主机名

## 2.1设置静态 IP

以 CentOS7 为例。

1. 编辑 IP 的配置文件
   vim /etc/sysconfig/network-scripts/ifcfg-enp0s3
2. 将配置文件更改如下：
   BOOTPROTO=”static”  #将 IP 改为静态配置
   IPADDR=”IP 地址”  #查看 windows 下连接的网段
   GATEWAY=”网关”
   NETMASK=”子网掩码”
   DNS1=”网关”
   DNS2=”180.76.76.76”  #百度的 DNS 解析服务器
3. 重启网络服务
   service network restart
4. 查看网络配置
   ip addr

## 2.2配置主机名

1. 编辑本主机名文件
   vim /etc/hostname
2. 将文件中的主机名更改为：
   buaa
3. 编辑主机名文件
   vim /etc/hosts
4. 将文件中的内容更改为：
   IP buaa  #这里的 IP 是 2.1 中配置的静态 IP

# 3. 配置 SSh 免登陆

1. 生成密钥对
   ssh-keygen –t rsa
2. 默认密钥存储文件
   回车（Enter）
3. 输入 SSH 登录密码
   回车（Enter）　　#免登陆，不输入密码
   回车（Enter）　　#确定密码
4. 将公钥保存到认证文件中
   cat ~/.ssh/id_rsa.pub >> authorized_keys
5. 设置.ssh 目录权限
   chmod 700 ~/.ssh
6. 设置私钥文件的权限
   chmod 600 ~/.ssh/id_rsa
7. 设置其余文件权限
   chmod 644 ~/.ssh/id_rsa.pub
   chmod 644 ~/.ssh/authorized_keys

# 4. 安装所需软件

## 4.1Java 配置

### 4.1.1 删除 openJDK

1. 查询已安装的 openJDK
   rpm –qa | grep jdk
2. 删除已安装的 openJDK
   rpm –e –nodeps xxxx　　#这里的“xxx”是上一条命令查询到的结果

### 4.1.2　下载 JDK

1. 下载 JDK 免安装版的压缩包：
   http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
2. 将压缩包(jdk-8u77-linux-x64.tar.gz)解压：
   tar –zxf jdk-8u77-linux-x64.tar.gz　　#解压可以得到 jdk1.8.0_77 文件目录
3. 将 jdk1.8.0_77 目录移至用户软件安装空间（将所有自己的软件放在一个文件夹下）
   mv jdk1.8.0_77 otherDir

## 4.2下载 Scala

1. 下载 Scala 免安装版的压缩包
2. http://www.scala-lang.org/download/2.11.8.html
3. 将压缩包（scala-2.11.8.tgz）解压：
   tar –zxf scala-2.11.8.tgz　　#解压可以得到 scala-2.11.8 文件目录
4. 将 scala-2.11.8 目录移至用户软件安装空间
   mv scala-2.11.8 otherDir

## 4.3下载 Hadoop

1. 选择 Hadoop 版本下载：这里使用 2.5.2 版本
   http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.5.2/hadoop-2.5.2.tar.gz
2. 将压缩包（hadoop-2.5.2.tar.gz）解压：
   tar –zxf hadoop-2.5.2.tar.gz　　#解压可以得到 hadoop-2.5.2 文件目录
3. 将 hadoop-2.5.2 目录移至用户软件安装空间

## 4.4下载 Spark

1. 选择 Spark 版本下载：这里使用 1.2.1 版本
   http://archive.apache.org/dist/spark/spark-1.2.1/spark-1.2.1-bin-hadoop2.4.tgz
2. 将压缩包（spark-1.2.1-bin-hadoop2.4.tgz）解压：
   tar –zxf spark-1.2.1-bin-hadoop2.4　　#解压可以得到 spark-1.2.1-bin-hadoop2.4 文件目录
3. 将 spark-1.2.1-bin-hadoop2.4 目录移至用户软件安装空间

# 5. 配置软件

## 5.1 软件路径配置

1. 编辑.bashrc 文件

vim ~/.bashrc

2. 向.bashrc 文件尾添加一下信息

export JAVA_HOME=JDK 的安装目录
export CLASSPATH=.:${JAVA_HOME}/jre/rt.jar:${JAVA_HOME}/lib/dt.jar:${JAVA_HOME}/lib/tools.jar
export SCALA_HOME=Scla 的安装目录
export HADOOP2_HOME=Hadoop 的安装目录
export SPARK1_HOME=Spark 的安装目录
export PATH=${PATH}:${JAVA_HOME}/bin:${SCALA_HOME}/bin:${HADOOP2_HOME}/bin:${HADOOP2

_HOME}/sbin:${SPARK1_HOME}/bin:${SPARK1_HOME}/sbin

3. 重新读取并执行.bashrc 中的命令

   source ~/.bashrc

4. 检测各软件路径是否配置正确

   java –version

   scala –version

   hadoop version

# 5.2配置并操作 Hadoop 分布式集群

## 5.2.1 创建 hdfs 的数据存储目录

1. 创建 hdfs 的存储目录

   mkdir /home/…/Hadoop 的安装目录/storage

2. 创建存放名称数据的目录

   mkdir /home/…/Hadoop 的安装目录/storage/name

3. 创建存放应用数据的目录

4. mkdir /home/…/Hadoop 的安装目录/storage/data

## 5.2.2 配置 core-site.xml

注：当前工作目录 pwd 为/…/Hadoop 安装目录/etc/hadoop

1. 编辑 core-size.xml

   vim ./core-site.xml

2. 向 core-site.xml 文件添加以下内容

```
<configuration>
    <!-- 设置 HDFS 的节点名称：就是告诉集群，HDFS 节点在哪 -->
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://buaa:8020</value>
        <final>true</final>
    </property>
    <!-- 设置临时文件夹 -->
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/home/yangel/Softwares/hadoop-2.5.2/tmp/hadoop-${user.name}</value>
    </property>
    <property>
</property>
```

## 5.2.3 配置 hdfs-site.xml

注：当前工作目录 pwd 为/…/Hadoop 安装目录/etc/hadoop

1. 编辑 hdfs-size.xml

   vim ./hdfs-site.xml

2. 向 hdfs-site.xml 文件添加以下内容

```
<configuration>
    <!—设置第二名称节点的地址：<主机名/IP：端口> -->
    <property>
        <name>hdfs.namenode.secondary.http-address</name>
        <value>buaa:50090</value>
    </property>
    <!—设置名称节点中元信息（名称数据）的存储位置 -->
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/home/yangel/Softwares/hadoop-2.5.2/storage/name</value>
    </property>
    <!-- 设置数据节点中数据的存储位置 -->
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/home/yangel/Softwares/hadoop-2.5.2/storage/data</value>
    </property>
    <!-- 设置是否启动 Web 访问 HDFS-->
    <property>
        <name>dfs.webhdfs.enabled</name>
        <value>true</value>
    </property>
    <!-- 设置是否启动 dfs 访问权限控制 -->
    <property>
        <name>dfs.permission</name>
        <value>false</value>
    </property>
    <!-- 设置数据副本数 -->
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
```

### 5.2.4 配置 mapred-site.xml

注：当前工作目录 pwd 为/…/Hadoop 安装目录/etc/hadoop

1. 从 mapred-site.xml.template 创建 mapred-size.xml
   cp ./mapred-site.xml.template ./mapred-site.xml
2. 编辑 mapred-site.xml
   vim mapred-site.xml
3. 向 mapred-site.xml 文件添加以下内容

```
<configuration>
    <!—设置资源调度框架-->
    <property>
        <name>mapreduce.framework.name</name>
```

```
            <value>yarn</value>
        </property>
        <!—设置历史作业：就是访问历史作业数据的位置：<主机名/IP:端口号> -->
        <property>
            <name>mapreduce.jobhistory.address</name>
            <value>buaa:10020</value>
        </property>
        <!-- 设置历史作业的 webapp 地址 -->
        <property>
            <name>mapreduce.jobhistory.webapp.address</name>
            <value>buaa:10021</value>
        </property>
        <property>
            <name>mapreduce.jobhistory.intermediate-done-dir</name>
            <value>/home/yangel/Softwares/hadoop-2.5.2/tmp</value>
        </property>
        <property>
            <name>mapreduce.jobhistory.done-dir</name>
            <value>/home/yangel/Softwares/hadoop-2.5.2/tmp</value>
        </property>
</configuration>
```

### 5.2.5 配置 yarn-site.xml

注：当前工作目录 pwd 为/…/Hadoop 安装目录/etc/hadoop

1. 编辑 yarn-size.xml

   vim ./yarn-site.xml

2. 向 yanr-site.xml 文件添加以下内容

```
<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>
        <!—设置资源管理器的地址-->
        <property>
            <name>yarn.resourcemanager.address</name>
            <value>buaa:8030</value>
        </property>
        <!—设置调度器的地址->
        <property>
            <name>yarn.resourcemanager.scheduler.address</name>
```

```
        <value>buaa:8031</value>
    </property>
    <!—设置资源跟踪器的地址-->
    <property>
        <name>yarn.resourcemanager.resource-tracker.address</name>
        <value>buaa:8032</value>
    </property>
    <!—设置资源管理器的管理员地址-->
    <property>
        <name>yarn.resourcemanager.admin.address</name>
        <value>buaa:8033</value>
    </property>
    <!—设置资源管理器的 web 访问地址-->
    <property>
        <name>yarn.resourcemanager.webapp.address</name>
        <value>buaa:8034</value>
    </property>
</configuration>
```

### 5.2.6 配置主节点文件 masters

注：当前工作目录 pwd 为/…/Hadoop 安装目录/etc/hadoop

1. 创建 masters 文件

   touch masters

2. 编辑 masters 文件添加以下内容

   vim masters

```
buaa    # 当前节点是主节点
```

### 5.2.7 配置从节点文件

注：当前工作目录 pwd 为/…/Hadoop 安装目录/etc/hadoop

1. 编辑 slaves 文件并添加以下内容

```
buaa    # 当前节点也是从节点
```

### 5.2.8 格式化 HDFS

1. 执行以下命令格式化 HDFS

   hdfs namenode -format

### 5.2.9 启动 Hadoop 集群

1. 启动 HDFS

   start-dfs.sh

2. 检测 HDFS 是否启动成功

   jps  #执行 jps 命令，如果出现 NameNode/DataNode/SecondaryNamNode 即为成功

3. 启动 YARN

   start-yarn.sh

4. 检测 YARN 是否启动成功

```
jps      #执行 jps 命令，如果出现 ResourceManager,NodeManager 即为成功
```

### 5.2.10 操作 Hadoop 集群

1. 查看 Hadoop 分布式文件系的概要
   http://buaa:50070/dfshealth.html#tab-overview
2. 查看 Hadoop 的历史作业记录
   http://buaa:8034/cluster

### 5.2.11 关闭 Hadoop 集群

1. 关闭 HDFS
   stop-dfs.sh
2. 关闭 YARN
   stop-yarn.sh

## 5.3配置并操作 Spark 集群

### 5.3.1 配置 slaves 文件

注：当前工作目录 pwd 为/.../Spark 安装目录/conf

4. 从 slaves.template 创建 slaves 文件
   cp ./slaves.template slaves
5. 编辑 slaves 文件
   vim slaves
6. 向 slaves 文件添加以下内容

```
buaa    #Spark 的 worker 节点
```

### 5.3.2 配置 spark-env.sh 文件

注：当前工作目录 pwd 为/.../Spark 安装目录/conf

1. 从 spark-env.template 创建 spark-env 文件
   cp ./ spark-env.template spark-env
2. 编辑 spark-env 文件
   vim spark-env
3. 向 spark-env 文件添加以下内容

```
export JAVA_HOME=JDK 的安装目录
export SCALA_HOME=Scala 的安装目录
export SPARK_MASTER_IP=主节点的 IP 地址
export SPARK_WORKER_MEMORY=1g        #工作节点的内存大小
export HADOOP_CONF_DIR=/../Hadoop 安装目录/etc/hadoop    #hadoop 的配置文件目录
```

### 5.3.3 启动 Spark 集群

1. 启动主节点，在主节点上执行以下命令
   start-master.sh
2. 启动从节点，在从节点上执行以下命令
   start-slave.sh

### 5.3.4 操作 Spark 集群

1. 查看 Spark 主节点的信息
   http://192.168.108.211:8080/

### 5.3.5 关闭 Spark 集群

1. 关闭从节点，在从节点上执行以下命令
   stop-slave.sh
2. 关闭主节点，在主节点上执行以下命令
   stop-master.sh