Hadoop单机伪分布式环境搭建

> 微信公众号：小康新鲜事儿

一、前提条件

Hadoop的运行依赖 JDK，需要预先安装，安装步骤见：

- Hadoop前置准备
- Linux下jdk的安装

二、Hadoop(HDFS和YARN)环境搭建

3.1 下载并解压

下载 Hadoop 安装包，这里我下载的是hadoop-2.7.7.tar.gz

```
[xiaokang@hadoop ~]$ sudo tar -zxvf hadoop-2.7.7.tar.gz -C /opt/software/
```

3.2 配置环境变量

```
[xiaokang@hadoop ~]$ sudo vim /etc/profile
```

在原来jdk基础上更新配置环境变量：

```
export JAVA_HOME=/opt/moudle/jdk1.8.0_191
export JRE_HOME=${JAVA_HOME}/jre
export HADOOP_HOME=/opt/software/hadoop-2.7.7
export CLASSPATH=.:${JAVA_HOME}/lib:${JRE_HOME}/lib
export PATH=${JAVA_HOME}/bin:${HADOOP_HOME}/bin:${HADOOP_HOME}/sbin:$PATH
```

执行 `source` 命令，使得配置的环境变量立即生效：

```
[xiaokang@hadoop ~]$ source /etc/profile
```

3.3 修改Hadoop配置

进入 `${HADOOP_HOME}/etc/hadoop/` 目录下，修改以下配置：

```
[xiaokang@hadoop ~]$ cd ${HADOOP_HOME}/etc/hadoop
```

1. hadoop-env.sh

```
#25行 export JAVA_HOME
export JAVA_HOME=/opt/moudle/jdk1.8.0_191
#33行 export HADOOP_CONF_DIR
export HADOOP_CONF_DIR=/opt/software/hadoop-2.7.7/etc/hadoop
```

2. core-site.xml

```
<configuration>
        <!--默认文件系统的名称 -->
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://hadoop:9000</value>
        </property>
        <!--指定HDFS执行时的临时目录 -->
        <property>
                <name>hadoop.tmp.dir</name>
                <value>/opt/software/hadoop-2.7.7/tmp</value>
        </property>
</configuration>
```

3. hdfs-site.xml

指定副本系数、namenode、datanode文件存放位置和hdfs操作权限：

```
<configuration>
        <property>
                <!--指定hdfs保存数据副本的数量，包括自己，默认为3-->
                <!--伪分布式模式，此值必须为1-->
                <name>dfs.replication</name>
                <value>1</value>
        </property>
         <property>
                <!--namenode节点数据（元数据）的存放位置-->
                <name>dfs.name.dir</name>
```

```
                <value>file:///opt/software/hadoop-
2.7.7/dfs/namenode_data</value>
        </property>
        <property>
                <!--datanode节点数据（元数据）的存放位置-->
                <name>dfs.data.dir</name>
                <value>file:///opt/software/hadoop-
2.7.7/dfs/datanode_data</value>
        </property>
        <property>
                <!--设置hdfs操作权限，false表示任何用户都可以在hdfs上操作文件-->
                <name>dfs.permissions</name>
                <value>false</value>
        </property>
</configuration>
```

4. mapred-site.xml

说明：在${HADOOP_HOME}/etc/hadoop的目录下，只有一个mapred-site.xml.template文件，复制一个进行更改。

```
[xiaokang@hadoop hadoop]$ sudo cp mapred-site.xml.template mapred-site.xml
```

```
<configuration>
        <property>
            <!--指定mapreduce运行在yarn上-->
            <name>mapreduce.framework.name</name>
            <value>yarn</value>
        </property>
</configuration>
```

5. yarn-site.xml

```
<configuration>
        <property>
                <!--指定yarn的老大resourcemanager的地址-->
                <name>yarn.resourcemanager.hostname</name>
                <value>hadoop</value>
        </property>
        <property>
                <!--NodeManager获取数据的方式-->
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
</configuration>
```

6. slaves

配置所有从属节点的主机名或 IP 地址，由于是单机版本，所以指定本机即可：

```
hadoop
```

3.4 关闭防火墙

不关闭防火墙可能导致无法访问 Hadoop 的 Web UI 界面：

```
# 查看防火墙状态
sudo firewall-cmd --state
# 关闭防火墙：
sudo systemctl stop firewalld.service
```

3.5 初始化

第一次启动 Hadoop 时需要进行初始化执行以下命令：

```
[xiaokang@hadoop ~]$ hdfs namenode -format
或
[xiaokang@hadoop ~]$ hadoop namenode -format
```

3.6 启动HDFS和YARN

```
[xiaokang@hadoop ~]$ start-dfs.sh
[xiaokang@hadoop ~]$ start-yarn.sh
```

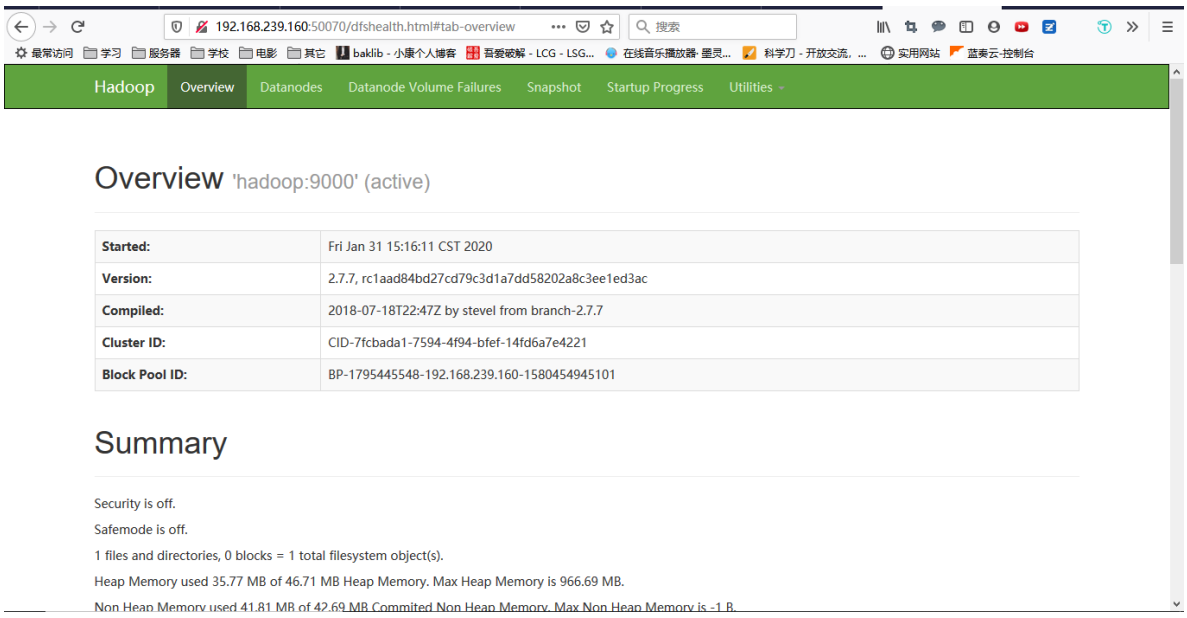
或者逐个服务启动，命令见文章最后部分

3.7 验证是否启动成功

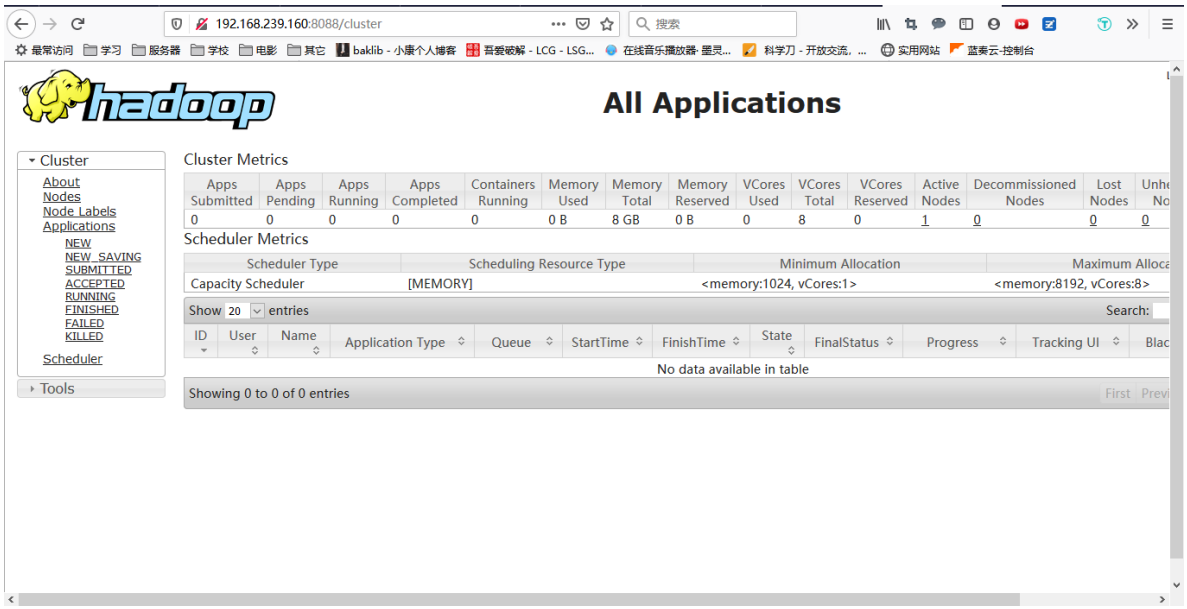
方式一：执行 `jps` 查看 `NameNode` 、 `DataNode`、`SecondaryNameNode`、`ResourceManager`、`NodeManager` 服务是否已经启动：

```
[xiaokang@hadoop ~]$ jps
11637 ResourceManager
11734 NodeManager
11241 DataNode
11146 NameNode
12075 Jps
11436 SecondaryNameNode
```

方式二：查看HDFS Web UI 界面，端口为 `50070`：



查看YARN Web UI 界面，端口号为 `8088`：

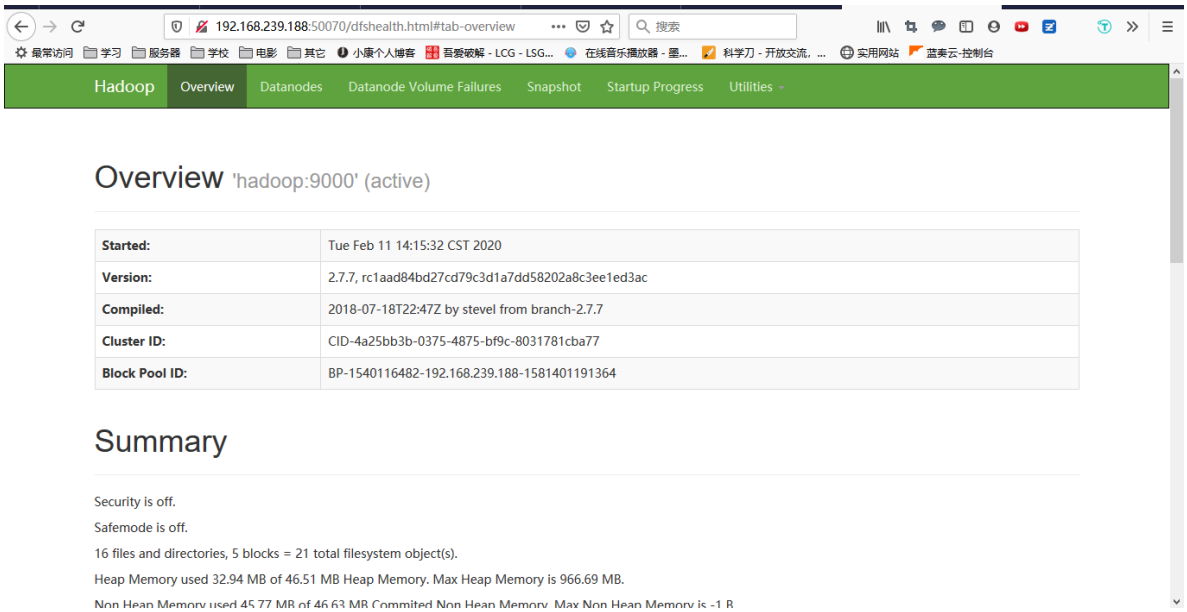


或者大家也可以使用 `hadoop-daemon.sh start/stop NN/DN/SNN`、`yarn-daemon.sh start/stop RM/NM`

启动NameNode服务和DataNode服务

```
[xiaokang@hadoop ~]$ hadoop-daemon.sh start namenode
[xiaokang@hadoop ~]$ hadoop-daemon.sh start datanode
```

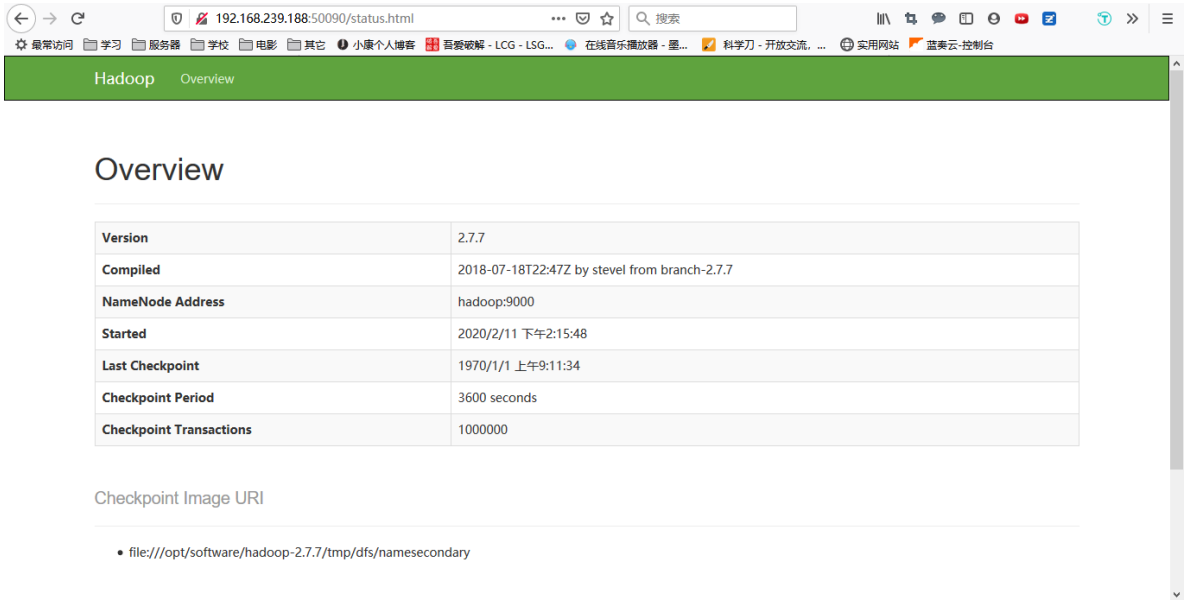
查看NameNode和DataNode Web UI 界面，端口为 `50070`：



启动SecondaryNameNode服务

```
[xiaokang@hadoop ~]$ hadoop-daemon.sh start secondarynamenode
```

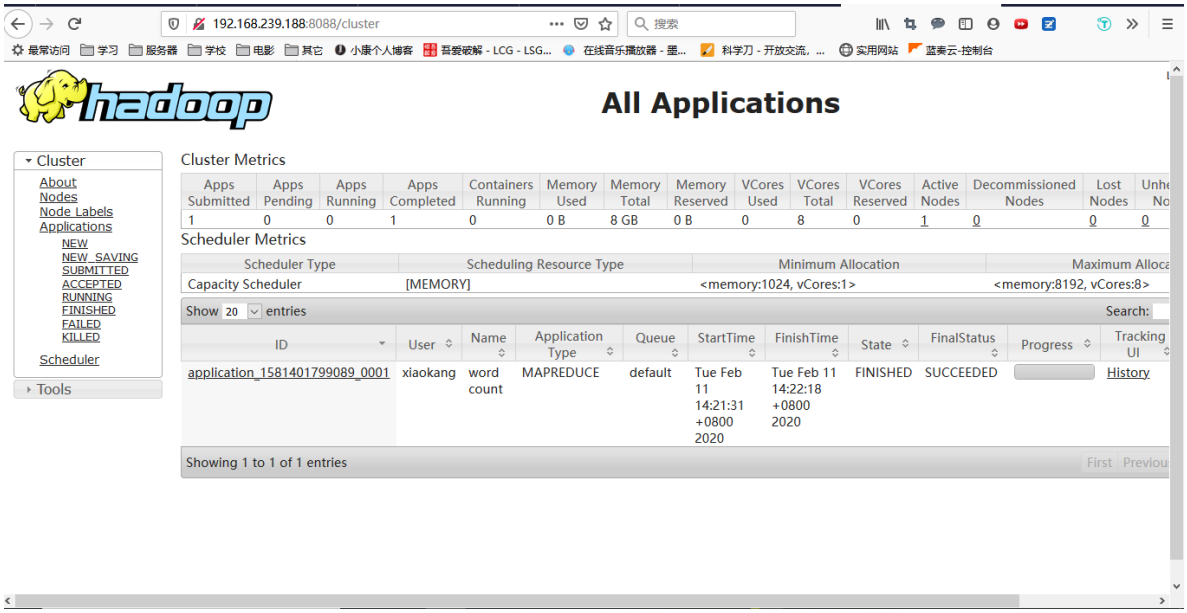
查看SecondaryNameNode Web UI 界面，端口为 `50090`：



启动ResourceManager服务

```
[xiaokang@hadoop ~]$ yarn-daemon.sh start resourcemanager
```

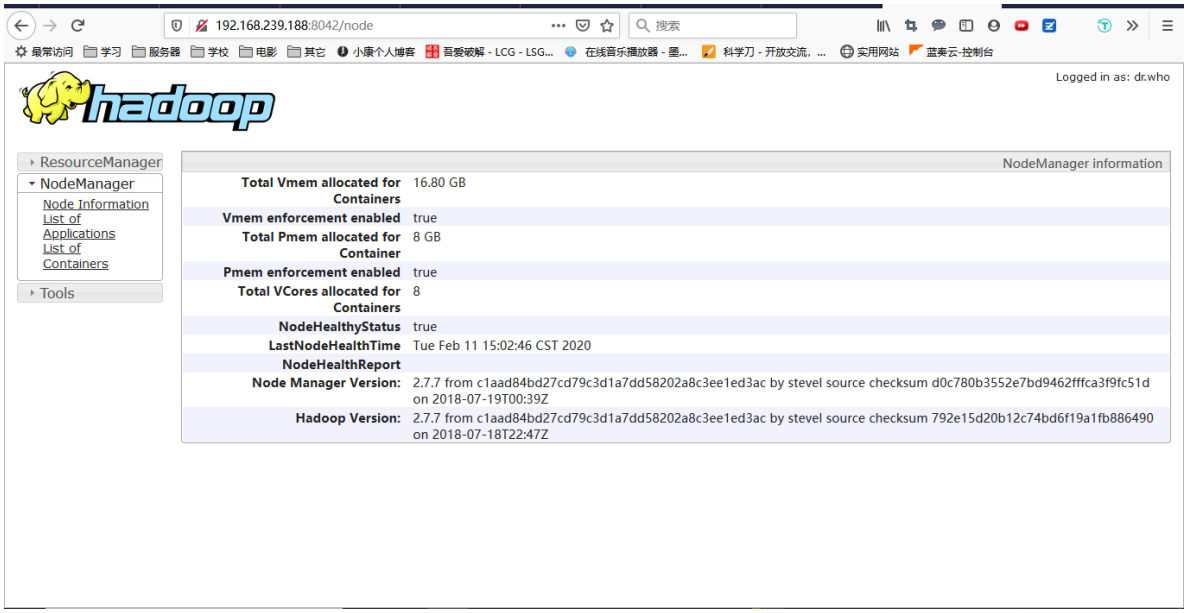
查看ResourceManager Web UI 界面，端口为`8088`：



启动NodeManager服务

```
[xiaokang@hadoop ~]$ yarn-daemon.sh start nodemanager
```

查看NodeManager Web UI 界面，端口为`8042`：



3.8 单机伪分布式官方wordcount案例测试

准备一个需要统计词频的小文件`wordcount.txt`

```
微信公众号：小康新鲜事儿 xiaokang xiaokangxxs
xiaokang
xiaokangxxs xiaokang1 xiaokang2
xiaokang2 xiaokangxxs
xiaokang6 xiaokang3 xiaokang2
小康新鲜事儿
```

将此文件上传至HDFS文件系统内（这里直接传到了根路径下。也可以自行创建目录）

```
[xiaokang@hadoop ~]$ hadoop fs -put ./wordcount.txt /
```

执行作业，测试wordcount

```
[xiaokang@hadoop ~]$ hadoop jar /opt/software/hadoop-
2.7.7/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar wordcount
/wordcount.txt /wc_result
```

查看结果

```
[xiaokang@hadoop ~]$ hadoop fs -cat /wc_result/part-r-00000
xiaokang    2
xiaokang1   1
xiaokang2   3
xiaokang3   1
xiaokang6   1
xiaokangxxs 3
小康新鲜事儿  1
微信公众号：小康新鲜事儿     1
```

三、配置任务的历史服务器

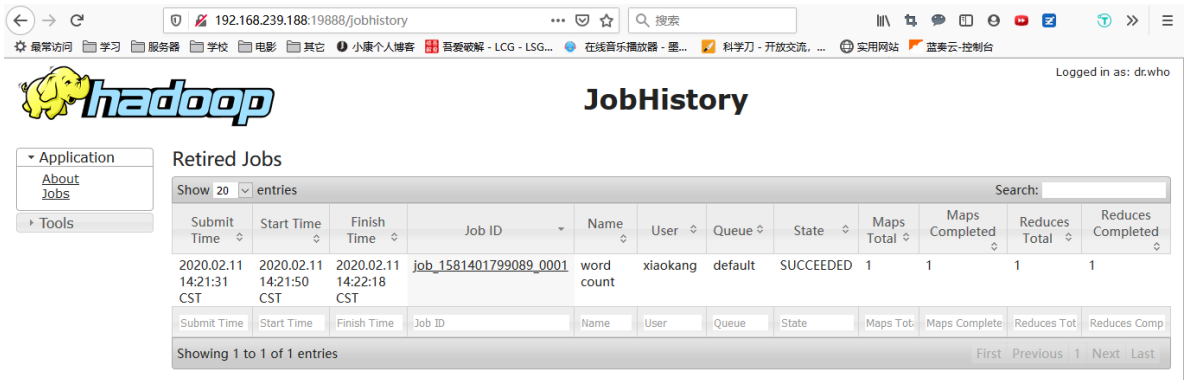
将以下内容加入到mapred-site.xml文件中

```
<property>
    <!--配置任务历史服务器IPC-->
    <name>mapreduce.jobhistory.address</name>
    <value>hadoop:10020</value>
</property>
<property>
    <!--配置任务历史服务器web-UI地址-->
    <name>mapreduce.jobhistory.webapp.address</name>
    <value>hadoop:19888</value>
</property>
```

启动任务历史服务器

```
[xiaokang@hadoop hadoop]$ mr-jobhistory-daemon.sh start historyserver
```

查看任务历史服务器 Web UI 界面，端口为 `19888`：



四、开启日志聚集功能

将以下内容追加到yarn-site.xml文件中

```
<property>
    <!--开启日志聚集功能-->
    <name>yarn.log-aggregation-enable</name>
    <value>true</value>
</property>
<property>
    <!--配置日志保留7天-->
    <name>yarn.log-aggregation.retain-seconds</name>
    <value>604800</value>
</property>
```

需要重启Yarn和任务历史服务器生效

```
[xiaokang@hadoop hadoop]$ stop-yarn.sh
[xiaokang@hadoop hadoop]$ mr-jobhistory-daemon.sh stop historyserver

[xiaokang@hadoop hadoop]$ start-yarn.sh
[xiaokang@hadoop hadoop]$ mr-jobhistory-daemon.sh start historyserver
```

五、日志文件

.log：观察服务的执行日志

.out:统计信息：简单的输出、少量的出错信息