**Hadoop 算法实现**

**一、实验内容**

使用 Hadoop 实现WordCount 应用

WordCount 是一个最简单的分布式应用实例，主要功能是统计输入目录中所有单词出现的总次数，如文本文件中有如下内容：

Hello world

则统计结果应为

Hello 1

world 1

WordCount 可以使用多种方式实现，本次实验内容选择使用 Hadoop 实现 WordCount 程序，并完成对应实验报告

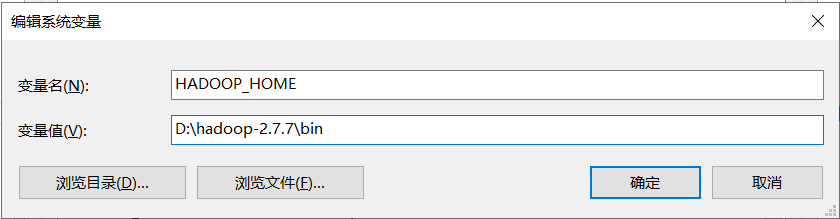
**二、实验原理**

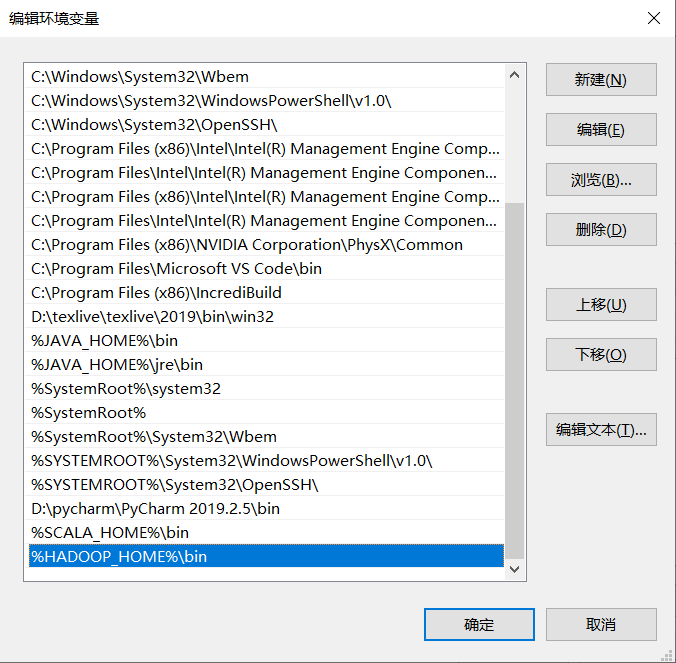
**1、配置hadoop环境**

**1）安装java1.8，并配置环境变量**

由于上学期的java课程已经安装配置过jdk1.8.0\_221，所以不需要再设置了

1. **在官网下载hadoop-2.7.7安装包并配置环境变量**





1. **使用notepad编辑器打开**

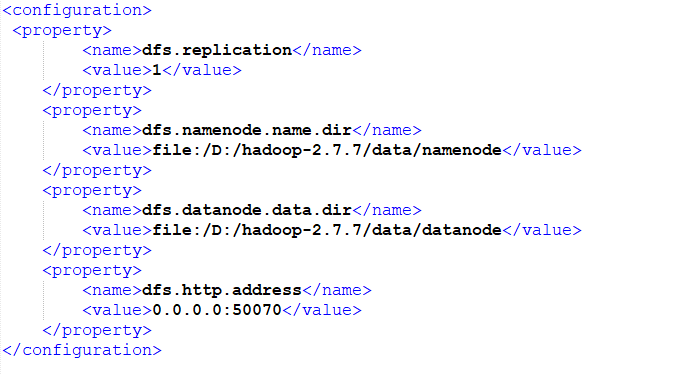
**D:\hadoop-2.7.7\etc\hadoop\hadoop-env.cmd修改JAVA\_HOME的路径，把set JAVA\_HOME改为jdk的位置**

**1591793960(1)**

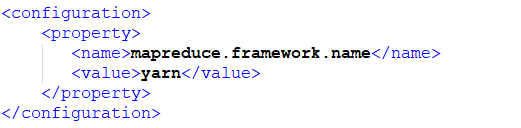
1. **新建data文件夹，并且在该文件夹下，新建两个文件夹namenode和datanode**

****

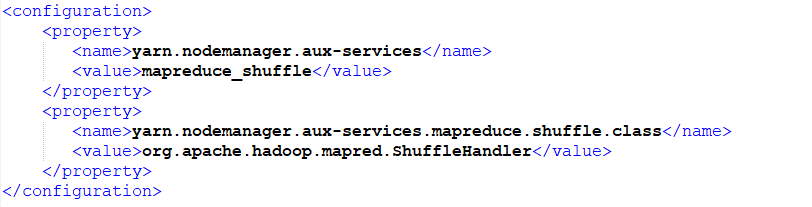
1. **打开D:/hadoop-2.7.7/etc/hadoop/hdfs-site.xml，修改路径为hadoop下的namenode和datanode**

****

1. **编辑mapred-site.xml文件**

****

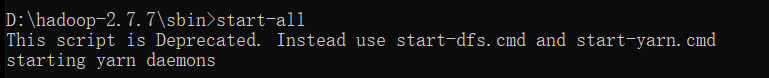
1. **编辑yarn-site.xml文件**

****

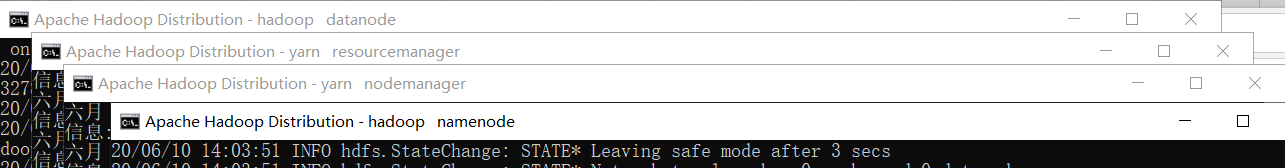
1. **在D:\Hadoop-2.7.7目录下 添加tmp文件夹**
2. **还需要把hadoop.dll（从）拷贝到 C:\Windows\System32，不然在window平台使用MapReduce测试时报错。**
3. **以管理员身份打开命令提示符输入hdfs namenode -format**

**9成功**

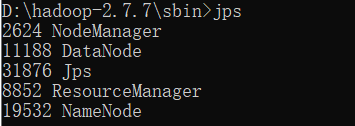
1. **转到Hadoop-2.7.7\sbin文件下输入start-all，启动hadoop集群**

****

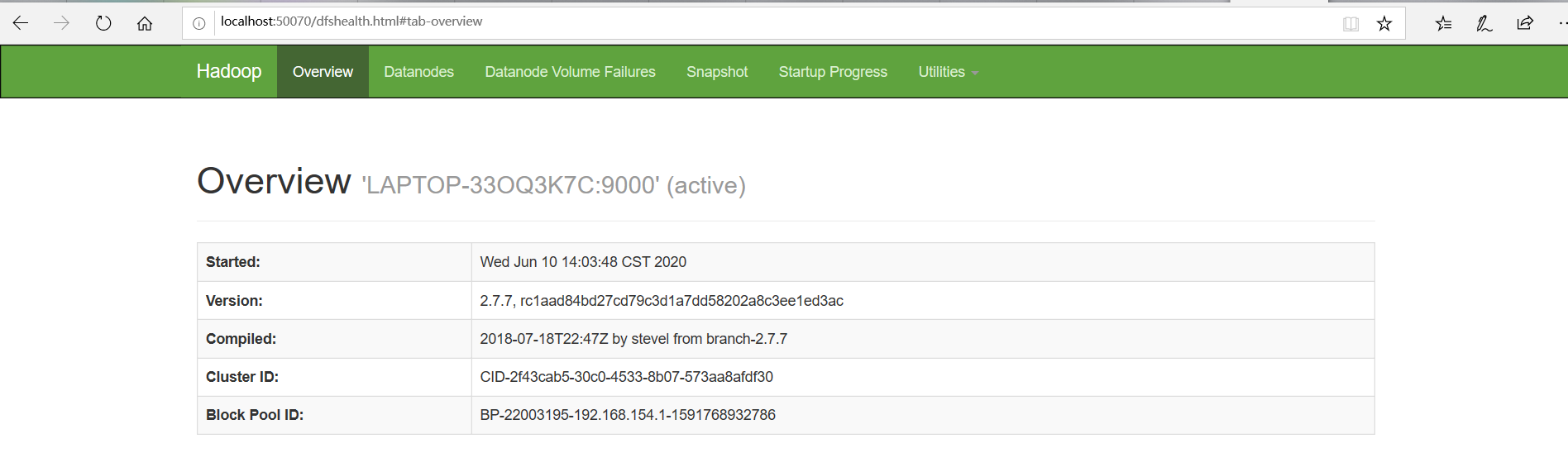
1. **出现下面四个窗口表示启动hadoop集群成功**

****

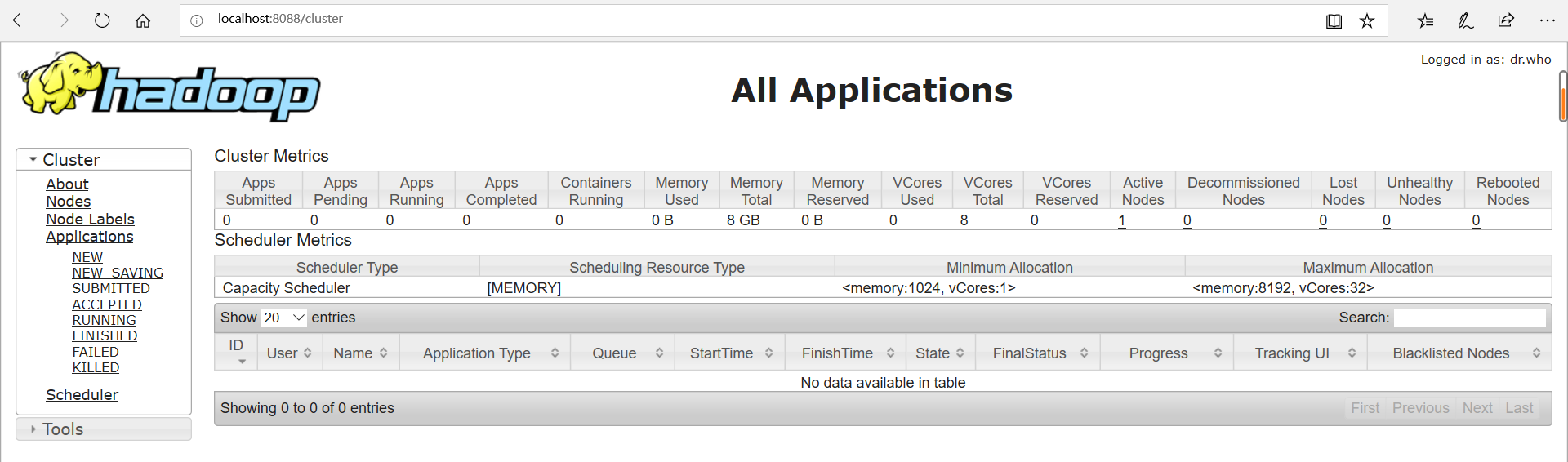
1. **输入jps - 可以查看运行的所有节点**

****

1. **访问http://localhost:50070,访问hadoop的web界面**

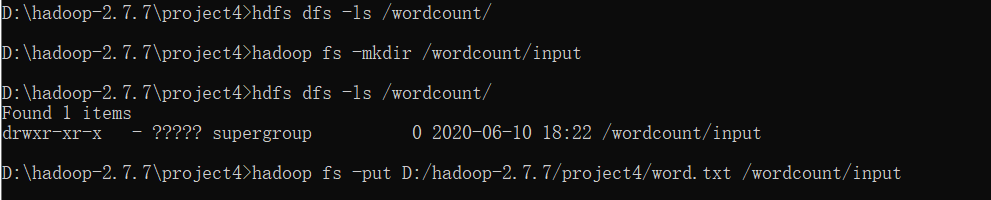
****

1. **访问http://localhost:8088,访问yarn的web界面**

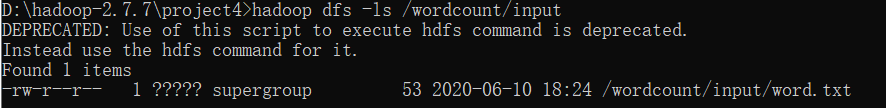
****

**三、WordCount实现**

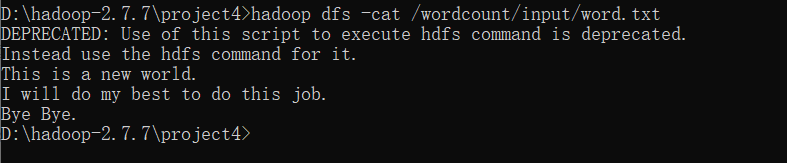
**1）首先在hdfs中创建一个文件夹，这里我起名字为wordcount，并将word文件放入hdfs中的input内**

****

1. **用下面这个命令查看文件夹状态**

****

1. **用下面这个命令查看上传的文件内容**

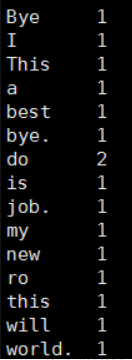
****

1. **实验结果**

运行share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar这个java程序，调用wordcount方法。/wordcount/input/word.txt是输入参数，待处理的文件 /wordcount/output是输出参数，保存处理后的数据的文件夹名字

17调用

输出结果

 实验成功

**五、实验总结**

这是第一次尝试使用hadoop，在安装与配置的时候出现了很多问题：

①启动Hadoop集群之后，发现少了namenode节点，通过查询资料发现，是因为formate超过两次导致的

②一开始没有办法打开http://localhost:50070，查询发现需要将hdfs-site.xml中localhost改成主机名称才行。

通过这次实验锻炼了我的自学能力，通过自己查询资料解决遇到的问题。提高了我对于Hadoop编程的认识。