在Windows 10上安装HDFS和MapReduce框架

## 一 安装Hadoop HDFS

1. 从<http://archive.apache.org/dist/hadoop/core/> 下载Hadoop-3.0.0，下载二进制压缩包文件：hadoop-3.0.0.tar.gz
2. 从https://github.com/steveloughran/winutils下载针对Windows环境的Hadoop修正程序winutils（Hadoop-3.0.0对应路径<https://github.com/steveloughran> /winutils /tree /master/hadoop-3.0.0）。如果不会用Git工具下载，可以直接下载打包文件<https://github.com/steveloughran/winutils/archive/master.zip>，然后解压。
3. 确保自己的计算机上已经正确安装了JDK 1.8以上的Java开发和运行环境。（可以通过在命令行方式下运行java –version命令进行验证）
4. 将hadoop-3.0.0.tar.gz解压至**C:\Hadoop**子目录（可以根据自己喜好更改Hadoop安装目录）
5. 新增环境变量HADOOP\_HOME（**作为系统变量**），并将其值设为“C:\Hadoop”。（设置方法：进入“控制面板\系统和安全\系统”，然后点击“高级系统设置”，然后点击“环境变量(N)..”按钮）（可以在命令行窗口中利用命令“echo %HADOOP\_HOME%”查看设置是否正确）
6. 观察JAVA\_HOME变量设置是否正确。（本人计算机上的JAVA\_HOME被设置为C:\Progra~1\Java\jdk1.8.0\_192）（可以在命令行窗口中利用命令“echo % JAVA\_HOME%”查看设置是否正确）（此处有坑：我的JDK安装在路径“C:\Program Files\Java\jdk1.8.0\_192”中，但该路径包含空格，因此用“C:\PROGRA~1”代替“C:\Program Files”，即将JAVA\_HOME设置为C:\Progra~1\Java\jdk1.8.0\_192）
7. 将“C:\Hadoop\bin”和“C:\Hadoop\sbin”路径添加到Path环境变量（作为系统变量）。
8. 将如下内容粘贴到C:\Hadoop\etc\hadoop\core-site.xml文件：

注意是将两个configuration之间的内容粘贴到原文件的两个configuration之间

|  |
| --- |
| **<configuration>**  **<property>**  **<name>fs.defaultFS</name>**  **<value>hdfs://localhost:9000</value>**  **</property>**  **</configuration>** |

1. 在C:\Hadoop路径下创建“data”子目录；

在C:\Hadoop\data路径下创建“namenode”子目录；

在C:\Hadoop\data路径下创建“datanode”子目录。

1. 将如下内容粘贴到C:\Hadoop\etc\hadoop\hdfs-site.xml文件：

|  |
| --- |
| **<configuration>**  **<property>**  **<name>dfs.replication</name>**  **<value>1</value>**  **</property>**  **<property>**  **<name>dfs.namenode.name.dir</name>**  **<value>file:///C:/Hadoop/data/namenode</value>**  **</property>**  **<property>**  **<name>dfs.datanode.data.dir</name>**  **<value>file:///C:/Hadoop/data/datanode</value>**  **</property>**  **</configuration>** |

1. C:\Hadoop\bin目录下的内容删除。
2. 将第2步下载的“winutils-master.zip”解压，然后将解压文件中“..\winutils-master\hadoop-3.0.0\bin”目录下的内容拷贝到“C:\Hadoop\bin”目录。

## 二 检验Hadoop是否安装成功

经过前面12步，Hadoop安装结束。

运行如下命令测试Hadoop是否可以正常运行：

1. 运行命令“hadoop -version”看是否能正常输出。
2. 运行命令“hdfs namenode –format”，先对HDFS分布式文件系统进行格式化。
3. 进入“C:\Hadoop\sbin”目录运行命令“start-dfs”。如果一切正常，将会启动一个“hdfs namenode”进程和一个“hdfs datanode”进程，构成了只有1个主节点和1个从节点的“HDFS分布式文件系统集群”。可以通过“http://localhost:9870”监控HDFS系统。（用jps命令可以查看所有jvm相关的进程）
4. HDFS分布式文件系正常启动之后，可以用“hadoop fs”或“hdfs dfs”命令在分布式文件系统中实现“浏览目录”、“创建子目录”、“删除子目录”、“创建文件”、“拷贝文件”、“移动子目录或文件”、“查看文件内容”、“删除文件”、“上传本地文件”等操作。

|  |  |
| --- | --- |
| **hadoop fs –ls /** | **显示根目录下的所有文件和目录** |
| **hadoop fs –mkdir /test** | **创建子目录/test，创建多级目录 加上 –p** |
| **hadoop fs -rm /test1.txt** | **删除文件** |
| **hadoop fs -rm -r /test** | **删除子目录（要加-r参数）** |
| **hadoop fs -put C:\tmp\test.txt \test** | **将本地文件上传到HDFS分布式文件系统** |
| **hadoop fs -cat \test\test.txt** | **查看文件内容** |
| **hadoop fs -cp URI [URI …] <dest>** | **cp 复制系统内文件** |
| **hadoop fs -get[-ignorecrc] [-crc] <src> <localdst>** | **下载文件到本地** |
| **hadoop fs -mv URI [URI …] <dest>** | **将文件从源路径移动到目标路径** |
| **hadoop fs -du URI [URI …]** | **显示文件大小** |

1. <http://localhost:9870>工具中在“utilities”菜单下有一个“Browse the File System”工具，可以浏览、操作HDFS分布式文件系统。
2. 进入“C:\Hadoop\sbin”目录运行命令“stop-dfs”，关闭HDFS分布式文件系统。

## 三．运行MapReduce任务

1. 用“hadoop fs –mkdir /input”命令在HDFS中创建“/input”子目录。
2. 用“hadoop fs –put c:\source\input\_file.txt /input”命令把本地文件系统中的“c:\source\input\_file.txt”文件上传到HDFS的“/input”子目录。input\_file.txt文件将作为WordCount MapReduce任务的输入文件。

input\_file.txt文件内容如下：

|  |
| --- |
| aaa  bbb  aaa  bbb  ccc  ddd  ccc  aaa  bbb  bbb |

1. 用命令“hadoop jar c:\source\mrtest.jar WordCount /input /output”命令启动MapReduce任务。任务执行成功后到“/output”目录下查找执行结果文件。
2. 用命令“hadoop fs –cat /output/part-r-00000”查看WordCount任务的执行结果。