在Windows 10上安装HDFS和HBase

## 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:\Program Files\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）
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. 编辑文件C:\Hadoop\etc\hadoop\hadoop-env.cmd，将语句：

“set JAVA\_HOME=%JAVA\_HOME%”

修改为“set JAVA\_HOME=C:\PROGRA~1\Java\jdk1.8.0\_192”。

（注意：此处有坑。如果将hadoop-env.cmd中的JAVA\_HOME设置为“C:\Program Files\Java\jdk1.8.0\_192”，将会出错，因为路径中不能含有空格。）

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13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“hdfs namenode –format”，先对HDFS分布式文件系统进行格式化。
2. 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）
3. 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．设定MapReduce运行框架

1. 将如下内容粘贴到C:\Hadoop\etc\hadoop\mapred-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>mapreduce.framework.name</name>**  **<value>yarn</value>**  **</property>**  **</configuration>** |

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

|  |
| --- |
| **<configuration>**  **<property>**  **<name>yarn.nodemanager.aux-services</name>**  **<value>mapreduce\_shuffle</value>**  **</property>**  **<property>**  **<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>**  **<value>org.apache.hadoop.mapred.ShuffleHandler</value>**  **</property>**  **</configuration>** |

**四．运行MapReduce任务**

注意：以下操作必须以**管理员身份**执行。（变量HADOOP\_HOME作为系统变量设置正确。）

1. 进入“C:\Hadoop\sbin”目录运行命令“start-all”。如果一切正常，将会启动一个“hdfs namenode”进程、一个“hdfs datanode”进程、一个“yarn resourcemanager”进程，一个“yarn nodemanager”进程。
2. 打开浏览器，输入URL：“http://localhost:8088/”，可以看到MapReduce框架的资源调度管理中心节点软件resourcemanager提供的Web监控页面。通过该页面可以查看现在正在运行的MapReduce任务的运行状态，可以查看成功或失败了的MapReduce任务信息。
3. 用“hadoop fs –mkdir /input”命令在HDFS中创建“/input”子目录。
4. 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“yarn 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。

## 五．安装HBase

在已经将HDFS安装成功的情况下，可以按照如下步骤进一步安装HBase。

1. 到HBase的官网（https://hbase.apache.org/）下载已编译好（bin格式）的HBase2.2.4（教育网可从镜像<https://mirror.bit.edu.cn/apache/hbase/2.2.4/hbase-2.2.4-bin.tar.gz>下载）。
2. 将下载的hbase-2.2.4-bin.tar.gz解压至C:\HBase子目录（可以根据自己喜好更改HBase安装目录）
3. 新增环境变量HBASE\_HOME（作为系统变量），并将其值设为“C:\HBase”；新增环境变量HBASE\_BIN\_PATH，并将其值设为“C:\HBase\bin”。（另外，HBase运行还依赖环境变量HADOOP\_HOME的正确设置。可以用“echo % HADOOP\_HOME%”命令查看设置是否正确。）
4. 修改配置文件C:\HBase\conf\hbase-env.cmd，为HBase设置正确的JDK主目录参数。具体方法：在语句“@rem set JAVA\_HOME=c:\apps\java”的下一行增加语句“set JAVA\_HOME=C:\PROGRA~1\Java\jdk1.8.0\_192”。注意这里JAVA\_HOME的具体设置要和自己计算机安装JDK的路径相匹配。
5. 将C:\HBase\conf\ hbase-site.xml修改为如下内容：

|  |
| --- |
| **<configuration>**  **<property>**  **<name>hbase.tmp.dir</name>**  **<value>tmp</value>**  **</property>**  **<property>**  **<name>hbase.unsafe.stream.capability.enforce</name>**  **<value>false</value>**  **</property>**  **</configuration>** |

1. 在C:\HBase目录下新建“C:\HBase\tmp”目录。（HBase内嵌的ZooKeeper会用到该目录。）
2. 将下载的jansi-1.18.jar文件拷贝到C:\HBase\lib目录。（否则，在运行“hbase shell”命令时会报错，但该错误不会影响hbase shell的正常功能。）

## 六．运行HBase

1. 在命令行方式下先用“start-dfs”命令启动HDFS系统。
2. 等待HDFS启动成功之后，将当前目录切换到“C:\HBase\bin”目录。下面的 命令都假定当前目录为“C:\HBase\bin”。
3. 运行命令“start-hbase”启动HBase系统。（耐心等待，正常情况下会弹出另外一个窗口，该窗口对应HBase的HMaster进程）。
4. 运行jps命令，发现多了一个HMaster进程。
5. 等待HBase启动成功之后（运行jps命令，发现多了），运行“hbase shell”命令进入HBase系统提供的shell工具。
6. 在hbase shell工具交互模式下，进行如下创建数据表、插入记录、查询记录和删除数据表的操作。

|  |  |
| --- | --- |
| **create 'test', 'cf'** | 创建数据表test，该表包含一个叫cf的列族。 |
| **list** | 查看系统里所有的数据表 |
| **list 'test'** | 查看系统里是否存在叫test的表。 |
| **describe 'test'** | 查看test表的属性 |
| **put 'test', 'row1', 'cf:a', 'value1'** | 向test表中插入一条记录，该记录行键为“row1”，字段“cf:a”的取值为“value1”。 |
| **scan 'test'** | 扫描数据表test，返回所有记录 |
| **get 'test', 'row1'** | 查询数据表test中行键为‘row1’的记录 |
| **disable 'test'** | 禁用数据表test |
| **drop 'test'** | 删除数据表test |

1. 在浏览器中输入URL: <http://localhost:16010/>可以基于Web方式监控HBase系统的运行。
2. 利用Hadoop HDFS的“hadoop fs –ls /user/llh/tmp/hbase”命令观察HBase在HDFS文件系统中存放的数据文件。（注意llh是用户名，你的系统中是你的登录用户名）。
3. 从hbase shell工具退出后，利用“stop-hbase”命令关闭HBase系统。