**第6章  HDFS文件系统**

**1.实验步骤**

1.1  **启动Hadoop集群**

**步骤1.**  **启动Hadoop**

打开一个终端模拟器，通过下列命令启动Hadoop：

start-all.sh

**步骤6.**  **验证Hadoop是否启动成功**

通过下列命令，查看相应的JVM进程确定Hadoop是否启动成功：

jps

1.2  **HDFS的命令行操作**

**步骤1.**  **查看HDFS命令参数**

通过下列命令可以查看HDFS文件系统操作的一些基本参数：

hadoop fs -help

**步骤2.**  **列出HDFS文件**

通过hadoop fs -ls命令指定绝对路径，

**步骤3.**  **列出HDFS子目录下的文件**

过下面的命令浏览HDFS下/user子目录的文件：

hadoop fs -ls /user

**步骤4.**  **在HDFS中创建子目录**

1.   使用hadoop fs -mkdir命令可以在HDFS中创建子目录

hadoop fs -mkdir /test

2.   通过下列命令列出HDFS文件系统根目录（/）下的所有文件：

hadoop fs -ls /

**步骤5.**  **上传文件到HDFS**

通过hadoop fs -put命令可以向HDFS中上传文件，第一个参数代表需要上传到HDFS中的本地文件的绝对路径，第二个参数代表HDFS中对应的指定目录。

1.   通过下面的命令创建需要上传的数据文件：

echo "this is my name" > /home/input.txt

2.   通过下面的命令查看数据文件是否创建成功：

cat /home/input.txt

3.将本地的/home/input.txt文件上传到HDFS文件系统的/test子目录下：

hadoop fs -put /home/input.txt /test

4.浏览HDFS下/test子目录的文件：

hadoop fs -ls /test

**步骤6.**  **将HDFS中文件复制到本地系统中**

通过hadoop fs -get命令可以获取HDFS中的文件。第一个参数代表需要获取的文件在HDFS中的绝对路径，第二个参数代表将文件存储到本地的指定接收目录。

1.将HDFS中的/test/input.txt文件复制到本地文件系统的/usr目录下：

hadoop fs -get /test/input.txt /usr/

2.   通过下面的命令可以查看文件是否复制成功：

cat /usr/input.txt

**步骤7.**  **查看HDFS下某个文件**

通过hadoop fs -cat命令可以查看HDFS中指定文件的内容

hadoop fs -cat /test/input.txt

**步骤8.**  **删除HDFS下的文档**

1.通过hadoop fs -rmr命令可以删除HDFS下指定的目录或文件，下面的命令将会删除HDFS中的/test子目录：

hadoop fs -rmr /test

2.   通过下列命令列出HDFS文件系统根目录（/）下的所有文件：

hadoop fs -ls /

1.3  **使用Hadoop归档文件**

**步骤1.**  **创建工作目录**

1.   通过下列命令在HDFS中创建/archive工作目录：

hadoop fs -mkdir /archive

2.   通过下列命令在/archive目录下创建子目录file1：

hadoop fs -mkdir /archive/file1

3.   通过下列命令在/archive目录下创建子目录file2：

hadoop fs -mkdir /archive/file2

**步骤2.**  **上传数据文件**

1.   通过下面的命令创建需要上传的数据文件test\_file1.txt：

echo "this is test\_file1" > /home/test\_file1.txt

2.   通过下面的命令创建需要上传的数据文件test\_file6.txt：

echo "this is test\_file2" > /home/test\_file6.txt

3.   通过下面的命令将本地的/home/test\_file1.txt文件上传到HDFS文件系统的/archive/file1子目录下：

hadoop fs -put /home/test\_file1.txt /archive/file1

4.   通过下面的命令测试文件是否上传成功：

hadoop fs -cat /archive/file1/test\_file1.txt

如果文件上传成功，将会输出下列内容：

[root@master ~]# hadoop fs -cat /archive/file1/test\_file1.txt

18/11/08 16:02:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

this is test\_file1

[root@master ~]#

5.   通过下面的命令将本地的/home/test\_file6.txt文件上传到HDFS文件系统的/archive/file2子目录下：

hadoop fs -put /home/test\_file6.txt /archive/file2

6.   通过下面的命令测试文件是否上传成功：

hadoop fs -cat /archive/file2/test\_file6.txt

**步骤3.**  **创建归档文件**

通过下列命令将HDFS文件系统/archive目录下的所有文件进行归档，命名为files.har并存储到HDFS文件系统的根目录（/）下：

hadoop archive -archiveName files.har -p /archive /

**步骤4.**  **归档文件测试**

1.   通过下列命令查看HDFS根目录确定归档文件是否成功生成：

hadoop fs -ls /

2.   通过下列命令可以进行归档文件子内容的查看：

hadoop fs -ls /files.har

3.   通过下列命令可以进一步查看归档文件的内容：

hadoop fs -cat /files.har/part-0

1.4  **通过distcp进行并行复制**

HDFS提供了一个非常实用的复制命令distcp，通过distcp命令可以快速的在两个HDFS集群间并行的进行数据传送。具体实现时，distcp操作会被解析为一个MapReduce操作来执行，当没有Reducer操作时，复制被作为Map操作并行地在集群节点中运行，因此每个文件都可以被当成一个Map操作来执行复制。

使用distcp命令通常需要传入2个参数：第一个参数设置输入源文件的完整访问路径hdfs://输入源集群NameNode主机名:端口号/输入源目录；第二个参数设置目标位置的完整路径hdfs://目标集群NameNode主机名:端口号/输出目录。

为了操作方便，我们在同一个集群中演示distcp命令的使用，因此我们的输入源为hdfs://master:9000/，对应的输出源也为hdfs://master:9000/。

**步骤1.**  **创建目标路径**

1.   通过下面的命令在HDFS中创建目标路径（此过程不会有输出）：

hadoop fs -mkdir /distcpdir

2.   通过下列命令查看HDFS根目录确定目录是否成功生成：

hadoop fs -ls /

**步骤2.**  **文件复制**

通过下面的命令将HDFS中/archive下的文件复制到/distcpdir目录下（在上文中已经向/archive目录中上传过文件）：

hadoop distcp hdfs://master:9000/archive  hdfs://master:9000/distcpdir

**步骤3.**  **复制结果查看**

1.   通过下列命令可以进行目录内容的查看：

hadoop fs -ls /distcpdir

6.   通过下列命令可以进一步查看子目录的内容：

hadoop fs -ls /distcpdir/archive

**步骤4.**  **使用HFTP进行数据复制**

当使用distcp进行HDFS集群间的复制时，任务需要在目标集群中执行，所以HDFS的RPC服务版本需要匹配，当HDFS运行在不同的Hadoop版本之上，复制将会因为RPC版本的不匹配而失败，此时可以使用基于HTTP的HFTP进行数据操作。需要注意的是，输入源的URI中要定义NameNode的网络接口，这个接口会通过dfs.hftp.address的属性值设定，默认值为50070。

1.   通过下面的命令可以实现将HDFS中/archive下的文件以HFTP模式复制到/distcpdir1目录下（在上文中已经向/archive目录中上传过文件）：

hadoop distcp hftp://master:50070/archive  hdfs://master:9000/distcpdir1

2.   通过下列命令可以进行目录内容的查看：

hadoop fs -ls /distcpdir1