2. 两位同学组队，按照助教提供的 HDFS 文件的基本创建、读写操作的 java 代码，编程实现在 HDFS 中创建大批量小文件(文件内容随意)，分析小文件的 Block 是多大，体会使用hdfs存储小文件的缺点。

报告内容：请在报告中详细写明你的实验步骤、技术方法、实习体会等，附上相应的代码段和截图。

实验步骤

在循环中创建多个小文件，每个小文件为随机复制的多个hello，如果这些小文件已创建，则读取这些小文件。

for(int i = 0;i<10;i++)

{

try

{

String fileName = "hw1/g1/test"+i;

Configuration conf = new Configuration();

conf.set("fs.defaultFS", "hdfs://Master:9000");

conf.set("fs.hdfs.impl", "org.apache.hadoop.hdfs.DistributedFileSystem");

FileSystem fs = FileSystem.get(conf);

Path file = new Path(fileName);

if(fs.exists(file))

{

System.out.println("File exists:"+fileName);

}

else

{

System.out.println("File does not exist");

createFile(fs,file);

}

readFile(fs,file);

fs.close();

}

catch(Exception e)

{

e.printStackTrace();

}

}

技术方法

使用随机数，随机决定每个小文件中hello出现的次数

public static void createFile(FileSystem fs, Path file) throws IOException

{

// System.out.println("create file");

Random random\_obj=new Random();

int repeat\_cnt=random\_obj.nextInt(100);

byte[] buff = "Hello".getBytes();

int buff\_len=buff.length;

FSDataOutputStream os =fs.create(file);

for(int i=0;i<repeat\_cnt;i++)

{

os.write(buff,0, buff\_len);

}

// //System.out.println("Create:"+file.getName());

os.close();

}

实习体会

hdfs对于小文件的存储表现并不好，每个小文件无论多小都需要占用至少一个block造成空间的浪费，同时每个小文件都需要有对应的Namenode造成可能比文件实际空间需要成倍的存储空间。同时hdfs的分布式存储也造成较大的时间开销。

运行截图

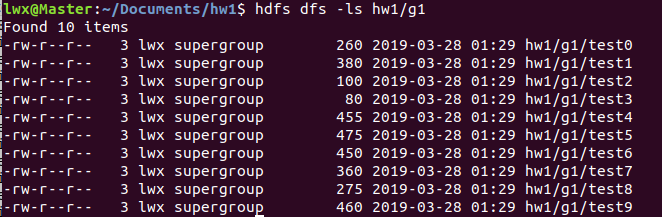
首次运行，创建随机文件并读取



再次运行，读取文件



查看hdfs中的目录



用-stat查看占用存储空间，这些小文件的所占用的存储都是一个block即128M，浪费存储空间

