Hbase 和 Hive 实验

组名 strong bro 邮箱:1613541957@qq.com 联系方式：15895987843

131220044 查鹏 131220034 许金强
(南京大学 计算机科学与技术系, 南京 210046)

1 实验目的

1.1 在本地电脑上按照Hbase和hive；

1.2 修改第二次的MapReduce程序，在Reduce阶段将倒排索引的信息通过文件输出，而每个词及其平均出现次数写入Hbase的Wuxia表中。

1.3 编程实现：将1.2中保存在Wuxia表中的数据导出到本地文件中；

1.4 使用Hive，在Hive shell命令行创建表，将1.3中文件内容里的数据保存到表中，然后查询出现次数大于300的所有词汇和前100个出现次数最多的词汇。

2 实验环境和数据说明

2.1 Ubuntu环境：Ubuntu12.04.5

2.2 Java环境：JDK1.7.0

2.3 Hadoop版本：Hadoop2.7.1

2.4 Hbase:1.2.4;

2.5 Hive:1.2.1

3 实验过程

3.1 安装和运行Hbase和hive

3.1.1 Hbase 安装配置

1. 从网站下载 hbase1.1.4-bin.tar.gz,移入 Ubuntu 的 hadoop 用户目录下；
2. 使用指令 tar –xvf hbase1.1.4-bin.tar.gz 解压得到 hbase 文件夹；
3. 使用指令 vim hbase-env.sh，添加 java 路径；

```
# The java implementation to use.  Java 1.7+ required.
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-i386
```

4. 使用指令 vim hbase-site.xml，设置配置文件；

```
<configuration>
 <property>
     <name>hbase.rootdir</name>
            <value>hdfs://localhost:9000/hbase</value>
           </property>
             <property>
                 <name>hbase.zookeeper.property.dataDir</name>
                     <value>/home/hadoop/workspace/zookeeper</value>
         </property>
         <property>
           <name>hbase.cluster.distributed</name>
            <value>true</value>
            </property>
</configuration>
```

5. 使用指令 bin/start-hbase.sh 开启 hbase，没开 hadoop 之前应该有如下 4 个进程

```
hadoop@ubuntu:~/hbase$ jps
3161 HQuorumPeer
3388 HRegionServer
3259 HMaster
3471 Jps
```

6. 使用指令 bin/hbase shell 进入 hbase shell，创建表 Wuxia。

3.1.2 Hive 安装配置

1. 下载并且解压 hive 压缩包，得到 hive 文件夹；
2. 进入 bashrc 目录添加 HADOOP_HOME,JAVA_HOME,HIVE_HOME 和 HBASE_HOME
3. 使用如下指令配置 hive 运行环境：

$HADOOP_HOME/bin/hadoop fs –mkdir /tep;

$HADOOP_HOME/bin/hadoop fs –mkdir /user/hive/warehouse;

$HADOOP_HOME/bin/hadoop fs –chmod g+w /tmp;

$HADOOP_HOME/bin/hadoop fs –chmod g+w /user/hive/warehouse;

4. 使用指令 bin/hive 开启 hive，开启 hive 后界面如下：

```
hadoop@ubuntu:~/hive$ bin/hive

Logging initialized using configuration in jar:file:/home/hadoop/hive/lib/hive-
ommon-1.2.1.jar!/hive-log4j.properties
hive>
```

3.2 输出数据到Hbase

3.2.1 代码实现

1. 连接到 hbase 数据库

```
connection= ConnectionFactory.createConnection(HBaseConfiguration.create());
table=connection.getTable(TableName.valueOf("Wuxia"));
```

2. 数据导入

数据的导入就是在之前代码的 Reduce 函数循环内部加入如下的代码片段。

每一次循环会得到一个词汇和对应的出现次数，这段代码片段的作用是将得到的词汇和出现次数写入 hbase 数据库中。

```
String average=String.format("%.2f", (double)count/(double)txtCount);//获取平均出现次数
//out.insert(0,average+",");
Put p = new Put(CurrentItem.getBytes());//使用词汇作为行
p.addColumn("average".getBytes(), "count".getBytes(), average.getBytes());//出现次数
table.put(p);//数据写入Hbase表
```

3. 代码运行情况

```
16/05/12 19:31:11 INFO mapreduce.Job:  map 100% reduce 68%
16/05/12 19:31:14 INFO mapreduce.Job:  map 100% reduce 69%
16/05/12 19:31:17 INFO mapreduce.Job:  map 100% reduce 70%
16/05/12 19:31:20 INFO mapreduce.Job:  map 100% reduce 71%
16/05/12 19:31:23 INFO mapreduce.Job:  map 100% reduce 72%
16/05/12 19:31:26 INFO mapreduce.Job:  map 100% reduce 74%
16/05/12 19:31:29 INFO mapreduce.Job:  map 100% reduce 76%
16/05/12 19:31:33 INFO mapreduce.Job:  map 100% reduce 77%
16/05/12 19:31:36 INFO mapreduce.Job:  map 100% reduce 79%
16/05/12 19:31:39 INFO mapreduce.Job:  map 100% reduce 80%
16/05/12 19:31:42 INFO mapreduce.Job:  map 100% reduce 81%
16/05/12 19:31:45 INFO mapreduce.Job:  map 100% reduce 83%
16/05/12 19:31:48 INFO mapreduce.Job:  map 100% reduce 85%
16/05/12 19:31:51 INFO mapreduce.Job:  map 100% reduce 87%
16/05/12 19:31:54 INFO mapreduce.Job:  map 100% reduce 88%
16/05/12 19:31:57 INFO mapreduce.Job:  map 100% reduce 90%
16/05/12 19:32:01 INFO mapreduce.Job:  map 100% reduce 92%
16/05/12 19:32:04 INFO mapreduce.Job:  map 100% reduce 94%
16/05/12 19:32:07 INFO mapreduce.Job:  map 100% reduce 95%
16/05/12 19:32:10 INFO mapreduce.Job:  map 100% reduce 97%
16/05/12 19:32:13 INFO mapreduce.Job:  map 100% reduce 99%
16/05/12 19:32:15 INFO mapreduce.Job:  map 100% reduce 100%
16/05/12 19:32:16 INFO mapreduce.Job: Job job_1463106379142_0001 completed succe
ssfully
```

3.2.2 查看 Wuxia 表

使用指令 scan ‘Wuxia’查看 Wuxia 表中的内容，结果如下：

```
Type "exit<RETURN>" to leave the HBase Shell
Version 1.1.4, r14c0e77956f9bb4c6edf0378474264843e4a82c3, Wed Mar 16 21:18:26 PD
T 2016

hbase(main):001:0> scan 'Wuxia'
```

```
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666946, value=2.44
\x9C\xE4\xB9\x8B\xE9
\x97\xB4
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666957, value=1.00
\x9C\xE9\x97\xB4
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666960, value=1.00
\xA5
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666963, value=8.86
\xA7
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666965, value=7.31
\xA7\xE5\x8D\x8A
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666967, value=4.23
\xA7\xE5\x8F\xA3
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666968, value=5.46
\xA7\xE5\x9D\x97
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666970, value=1.88
\xA7\xE5\xA0\x86
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666974, value=5.20
\xA7\xE5\xB8\xAE
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666978, value=2.44
\xA7\xE6\x88\xAA
\xE4\xB8\x80\xE5\xA4 column=average:count, timestamp=1463106666982, value=1.57
\xA7\xE6\x89\xB9
```

由于使用 Byte 存储的，所以查询 hbase 的时候是 ASCII 码格式的。

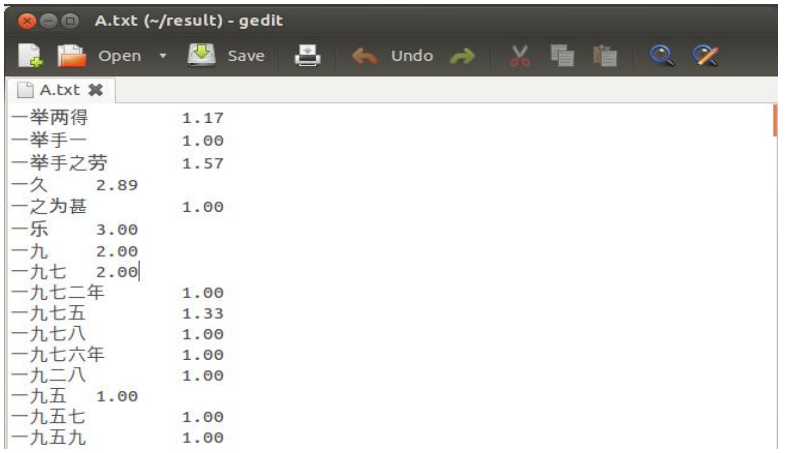
3.3 导出Hbase数据到本地文件

3.3.1 代码实现

```
Connection connection = ConnectionFactory.createConnection(HBaseConfiguration.create());
Table table = connection.getTable(TableName.valueOf("Wuxia"));
//指定输出文件目录为/home/hadoop/result/A.txt
java.io.File file=new java.io.File("/home/hadoop/result/A.txt");
java.io.PrintWriter output=new java.io.PrintWriter(file);
try {
    ResultScanner rs = table.getScanner(new Scan());
    for (Result r : rs) {
        for (KeyValue keyValue : r.raw()) {
            //获取行号：我把行号设置为词汇
            output.print(new String(r.getRow())+"\t");
            //获取列:我把列设置为平均出现次数
            output.print(new String(keyValue.getValue()));
            output.println();
        }
    }
    output.close();
```

3.3.2 查看文件

进入/home/hadoop/result/A.txt 文件，查看文件内容部分截图如下：



3.4 Hive查询操作

3.4.1 导入数据

使用指令 load data local inpath ‘/home/hadoop/result/A.txt’ into table Wuxia

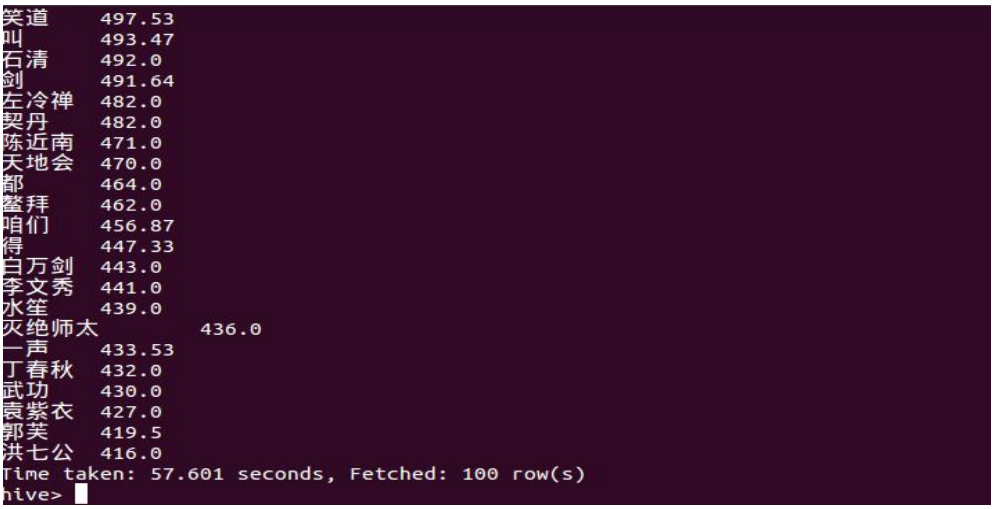
```
hive> LOAD DATA LOCAL INPATH
    > '/home/hadoop/result/A.txt'
    > INTO table PUT;
Loading data to table default.put
Table default.put stats: [numFiles=1, totalSize=974868]
OK
Time taken: 0.847 seconds
hive>
```

3.4.2　　查询出现次数大于 300 的词汇

使用指令 select * from ‘Wuxia’　where count>300，部分截图如下：

3.4.3　　按照降序排列，查询输入出现次数排名前 100 的词汇

输入指令 select *from Wuxia where order by count desc limit 100，部分截图如下：



4　实验体会

这次实验编程部分不是很难，在上一次实验的基础上添加一点代码了，通过这次实验让我们掌握了 hadoop 中实现 java 和数据库连接的方式，同时掌握了数据库的基本操作，包括创建删除表，还有数据的读入导出等等。