
1 Hbase 和 Hive 的安装和运行

1.1 Hbase 安装

首先，在 Hbase 官网下载 1.1.4 版本的 Hbase，将其在 hadoop 用户下解压，对 Hbase 中的配置文件进行修改配置。这里，最开始配置时在网查找了很多资料，但有些资料并不适用，最后才发现官网上提供的 quick start 教程里已经讲解的非常详细，按照官网进行配置即可。首先，修改 hbase-site.xml 中的配置与 hadoop 相连，之后修改 hbase-env.sh 中 java 路径的设置。完成这两项修改之后，使用命令 ./bin/start-hbase.sh 进行启动，看到 starting zookeeper, starting master 等信息即启动成功。再使用命令 ./bin/hbase shell 开启 hbase 的 shell 模式，在里面可以用简单的语句对数据库进行测试，不出现问题即配置成功。

1.2 Hive 安装

首先，在 Hive 官网下载 2.0.0 版本的 Hive，将其在 hadoop 用户下解压，Hive 的配置相对简单，不需要修改配置文件即可使用，只需要在使用前初始化 Hive 的默认数据库 derby，初始化命令如下：./bin/schematool -dbType derby -initSchema。初始化成功后，使用命令 ./bin/Hive 即可启动 Hive。

（注：Hbase 和 Hive 启动时都需要开启 Hadoop，使其满足运行环境）

2 Hbase 中创建表格

这里没有在 hbase 中手动创建，而是将创建表格的功能在代码中进行了实现，所以具体细节将在下面进行介绍。

3 将实验 2 的输出结果保存入 Hbase

这里我们在实现时对实验 2 的代码进行了修改，代码中添加了数据库的创建，删除和插入数据的功能，每次创建数据库时会首先检查是否含有名为 ‘wuxia’ 的表，如果有则想起删除，再新建一个表，这样可以避免程序运行时报错，也减少了手动删除数据库的工作量。代码实现方面，HBaseConfiguration 类用来对 Hbase 进行配置，HbaseAdmin 类可以用来创建和删除一个表格，put 类可以用来向表格中插入数据，在了解函数功能之后，实现起来还是比较简单的。实验结果部分截图如下：

```

xiao peng@cosecdeMacBook-Pro /Users/xiao peng/Desktop/hadoop/hbase-1....
=> ["wuxia"]
[hbase(main):008:0> scan 'wuxia'
ROW 483.9223744292237 COLUMN+CELL
0 341.0 column=apptimes:times, timestamp=1462894649288, value=3.0
007 305.0 column=apptimes:times, timestamp=1462894649288, value=1.0
01 997.3333333333334 column=apptimes:times, timestamp=1462894649288, value=1.0
01\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
02\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
03\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
04 396.0 column=apptimes:times, timestamp=1462894649288, value=1.0
04\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
05 1606.0 column=apptimes:times, timestamp=1462894649288, value=1.0
05\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
06 398.3954545454545 column=apptimes:times, timestamp=1462894649288, value=1.0
06\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
07\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
08\xE5\xBC\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
08\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
09\xE7\xAB\xA0 column=apptimes:times, timestamp=1462894649288, value=1.0
0\xE4\xB8\x80 column=apptimes:times, timestamp=1462894649288, value=1.5
0\xE5\xB9\xB4 column=apptimes:times, timestamp=1462894649288, value=2.5
1 column=apptimes:times, timestamp=1462894649288, value=1.0
1. column=apptimes:times, timestamp=1462894649288, value=1.0
10 column=apptimes:times, timestamp=1462894649288, value=4.0

```

图 1 wuxia 在 Hbase 中存储结果

4 遍历表格，将表格中内容保存到本地文件

这个任务则更加简单，只需要在创建 Hbase 配置之后，利用 scan 类对表格进行扫描，将扫描得到的数据库中每一行的数据进行提取，将词语和出现次数输入到文件当中即可。结果的部分截图如下：

```

per.java x wordcount.lml x scanWuxiaTable.java x wuxia.txt x stop.txt x Daopai.txt x invertstopword.java x
007 1.0
01 1.0
01章 1.0
02章 1.0
03章 1.0
04 1.0
04章 1.0
05 1.0
05章 1.0
06 1.0
06章 1.0
07章 1.0
08张 1.0
08章 1.0
09章 1.0
0一 1.5
0年 2.5
1. 1.0
10 4.0
1000万钱 1.0
100间 1.0
102个 1.0

```

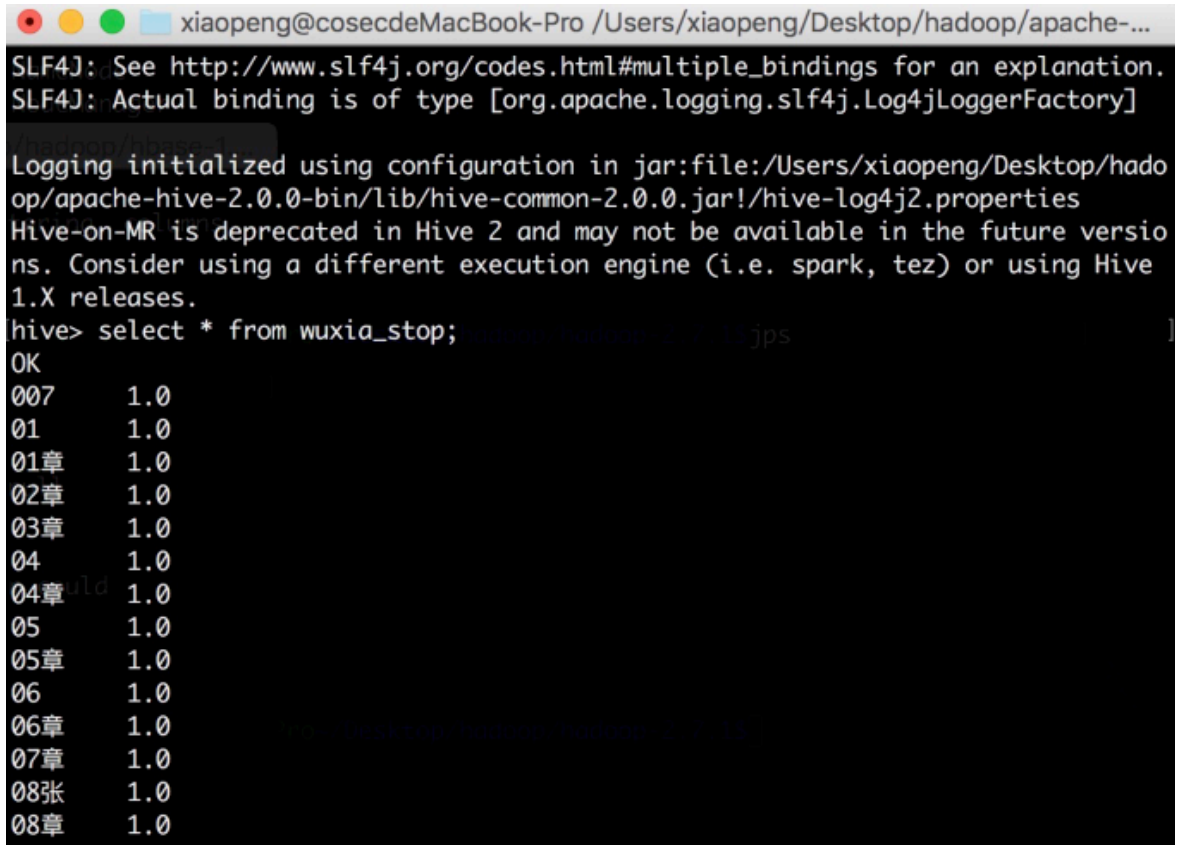
图 2 表格内容输出到文件

5 在 Hive shell 中导入上面的输出文件，并进行查询操作

在 Hive 中查询语句基本与之前使用过的 mysql 语句类似，所以在将数据导入到表格中后，可以比较容易的查询到想要的内容。

5.1 导入平均出现次数的数据

在 Hive 中利用语句：LOAD DATA LOCAL INPATH ‘本地文件路径’ INTO TABLE ‘表格名’；来进行数据的导入，导入之后可以使用“select * from 表格名”来进行查询，查询结果部分截图如下：



```
xiaopeng@cosecdeMacBook-Pro /Users/xiaopeng/Desktop/hadoop/apache-...
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/Users/xiaopeng/Desktop/hadoop/apache-hive-2.0.0-bin/lib/hive-common-2.0.0.jar!/hive-log4j2.properties
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> select * from wuxia_stop;
OK
007      1.0
01       1.0
01章     1.0
02章     1.0
03章     1.0
04       1.0
04章     1.0
05       1.0
05章     1.0
06       1.0
06章     1.0
07章     1.0
08张     1.0
08章     1.0
```

图 3 Hive 中所有数据查询结果

5.2 查询出现次数大于300的词语

```

(hive> select * from wuxia where count > 300;
OK
746.3545454545455
一个 444.1954545454545
一声 324.5136363636363
丁典 364.0
丁玲 586.5
万成 962.5
万震山 333.0
不 490.2590909090909
东方龙 544.0
两利 1324.7
中 536.1
之 387.9727272727273
乌老大 302.0
乐圣 593.0
也 1333.5181818181818
了 1560.1636363636364
人 86337.4136363636364
什么 86329.63084112149534
他 862591.1181818181817
他们 86563.4454545454546
令狐冲 861905.0
仪琳 86729.0

```

图 4 大于 300 的结果

5.3 查询前100个出现次数最多的词语

利用 DESC 语句倒序排序，可以列出数据库中前 100 的数据，结果部分截图如下：

```

OK
的 7103.027272727273
社英豪 4230.0
韦小宝 3277.0
道 3242.8636363636365
他 2591.1181818181817
你 2494.6454545454544
玮 2438.6666666666665
我 2352.05
张无忌 2338.0
令狐冲 1905.0
程小蝶 1898.5
小鱼儿 1756.25
段誉 1688.0
虚竹 1606.0
了 1560.1636363636364
芮 1528.4
百维 1416.0
狄云 1408.0
也 1333.5181818181818
两利 1324.7
是 1323.9545454545455
黄蓉 1262.5
易天行 1229.0

```

图 5 前 100 个出现次数最多的词语

6 选做部分实现

基本实现思路也是在原有代码的基础上进行添加修改,首先在 main 函数中创建一个 hbase 数据库表格,将停词表的内容存放进去,然后每一次 map 都会在停词表中进行查找,如果包含该词则不会进行下面的 reduce 工作,之后所进行的操作基本不变,将输出结果再保存到 hive 当中,供随时查询即完成任务。

下面是 hive 中的保存内容的查询结果部分截图,从图中的一些查询结果与上面的结果对比可以看出得到的结果已经不再包含停词表中的内容,任务完成。

(1) 查询所有数据:

```

[hive> select * from wuxia_stop;
OK
007      1.0
01       1.0
01章     1.0
02章     1.0
03章     1.0
04       1.0
04章     1.0
05       1.0
05章     1.0
06       1.0
06章     1.0
07章     1.0
08张     1.0
08章     1.0
09章     1.0
0一      1.5
0年      2.5
1.       1.0
10       4.0
1000万 钱      1.0
100间    1.0
102个    1.0
  
```

图 6 查询所有数据

(2) 查询出现次数大于 300 的词语：

```

xiaopeng@cosecdeMacBook-Pro /Users/xiaopeng/Desktop/hadoop/apache-...
Time taken: 0.108 seconds, Fetched: 136 row(s)
hive> select * from wuxia_stop where count > 300;
OK
一声      327.4908256880734
丁典      364.0
丁玲      586.5
万成      962.5
万震山    333.0
东方龙    544.0
两利      1471.888888888889
中        541.0183486238532
乌老大    302.0
乐圣      889.5
令狐冲    1905.0
仪琳      729.0
伍元      934.0
余沧海    378.0
余鱼同    304.0
佶        86345.6
俞        86383.962962962963
偶        86329.0
凌云凤    86415.666666666667
凤梧      86322.0
刀儿      86368.0

```

图 7 查询出现次数大于 300 的词语

(3) 查询前 100 个出现次数最多的词语

```

xiaopeng@cosecdeMacBook-Pro /Users/xiaopeng/Desktop/hadoop/apache-...
OK
杜英豪    4230.0
韦小宝    3277.0
道        3272.6146788990827
玮        2438.6666666666665
张无忌    2338.0
令狐冲    1905.0
程小蝶    1898.5
小鱼儿    1756.25
齐金蝉    1744.0
段誉      1688.0
虚竹      1606.0
芮        1528.4
两利      1471.888888888889
百维      1416.0
狄云      1408.0
黄蓉      1262.5
易天行    1229.0
岳不群    1184.0
楚天舒    1110.5
郭靖      1073.0
杨过      1021.6
袁承志    1013.3333333333334
胡斐      997.3333333333334

```

图 8 查询前 100 个出现次数最多的词语

