**考研数据动态分析系统**

[1.引言 2](#_Toc39166565)

[2.编写目的 2](#_Toc39166566)

[3.定义 3](#_Toc39166567)

[4.参考资料 3](#_Toc39166568)

[5.设计流程 3](#_Toc39166569)

[6.效果展示 7](#_Toc39166570)

[7.小结 9](#_Toc39166571)

# 1.引言

自2016年以来，考研人数呈正比例函数增长，考研难度也是逐年递增。在此基础上，我们小组做了一款考研相关数据动态分析系统，专门用来统计21届考生对于考试学科，以及目标院校所在省市的搜索次数，从而用来研究今年的考研形势。

# 2.编写目的

该系统用于统计21届考生对考研的考试学科，考研目标院校，以及部分考研关键词的搜索次数。

**分工：**

喻梦琪（组长）：确定分析主体分配任务，分析数据并展示——统计江西省在18号，搜索哪几个关键字次数最多，取3个

熊佳斯：爬取数据，分析数据并展示——统计13号到21号，每天搜索分数线关键字次数最多的省

范姬灵：分析数据并展示——统计22号所有地区搜索关键字的次数

白佳欣：分析数据并展示——统计14号，搜索考研关键字次数最多的3个省

陈蕤：分析数据并展示——统计21号，搜索数学关键字次数最少的3个省

丁祺伟：分析数据并展示——统计所有地区在13号到22号之间，搜索关键字次数最多的5个

# 3.定义

该系统是使用python以及hive、sqoop、mysql、tomcat、hdfs、shell等技术共同编译的，用于动态抓取数据的，实时数据分析系统。

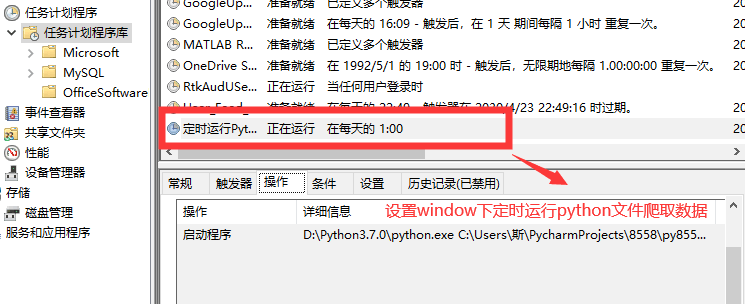
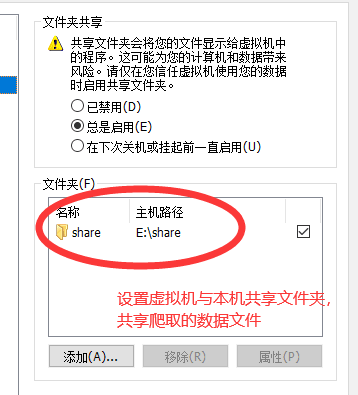
# 4.参考资料

[1] [Wes McKinney](https://book.jd.com/writer/Wes%20McKinney_1.html) 著，《Python for Data Analysis》，机械工业出版社，2014年1月，ISBN：9787111436737

# 5.设计流程

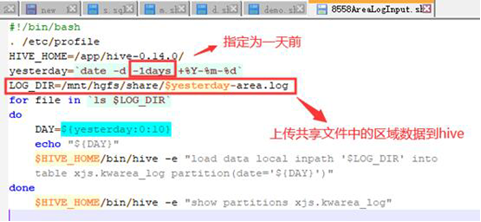
**5.1编写python代码爬取数据**

爬取百度指数上关于考研、分数线、专业、数学、英语、政治的全国搜索量和各个省份搜索量的数据



**5.2将数据上传到hive**

**指定为前一天上传共享文件中的区域数据到hive**

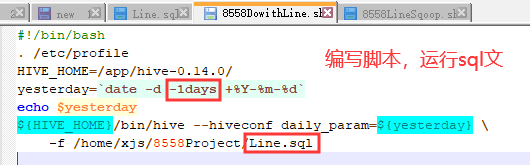


**5.3建表并编写sql脚本**

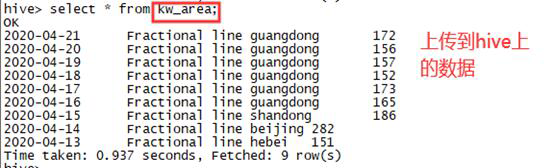
##### IMG_256



**5.4运行sql脚本**



运行完sql脚本之后我们就可以查看上传到hive上的数据结果如下图

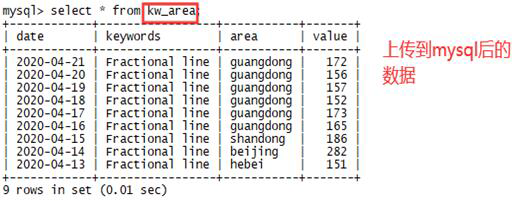


**5.5将hive中的表上传到MySQL上**

**编写脚本，利用sqoop将hive中的表上传到MySQL**

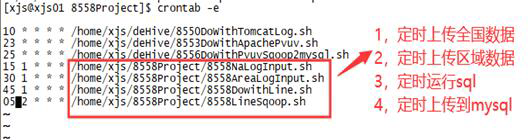


5.6查看数据



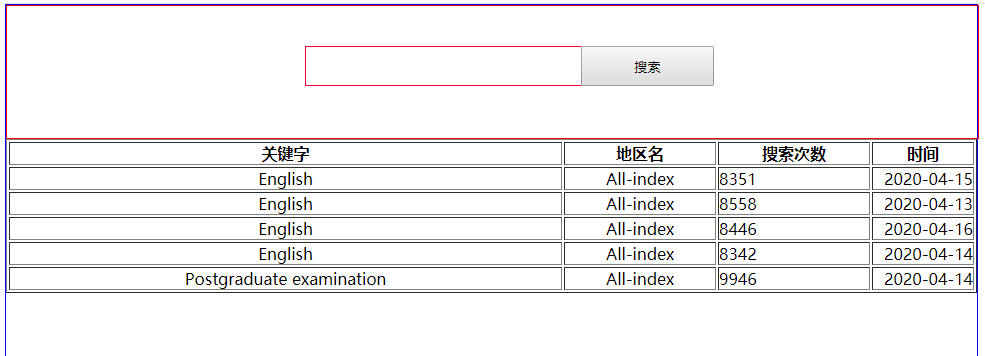
**5.7设置定时上传**

定时上传全国数据，定时上传区域数据，定时运行sql，定时上传数据到MySQL

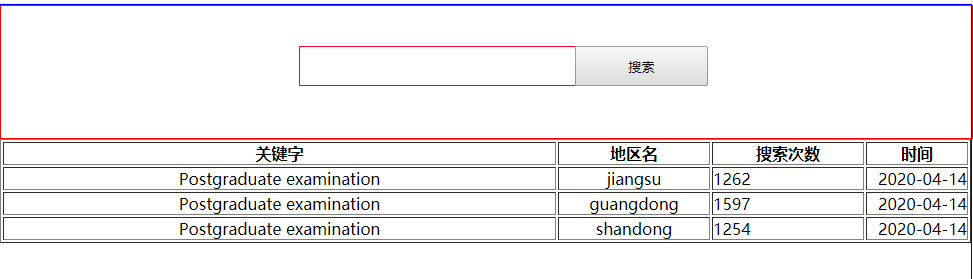


# 6.效果展示

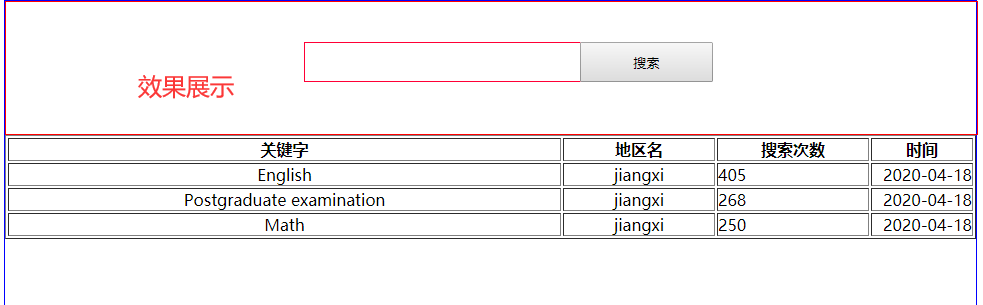
**6.1统计所有地区在13号到22号之间，搜索关键字次数最多的5个**



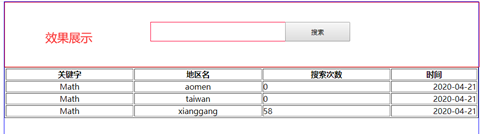
**6.2统计在14号搜索考研关键字次数最多的三个省**



**6.3统计江西省在18号搜索关键字次数最多的三个**



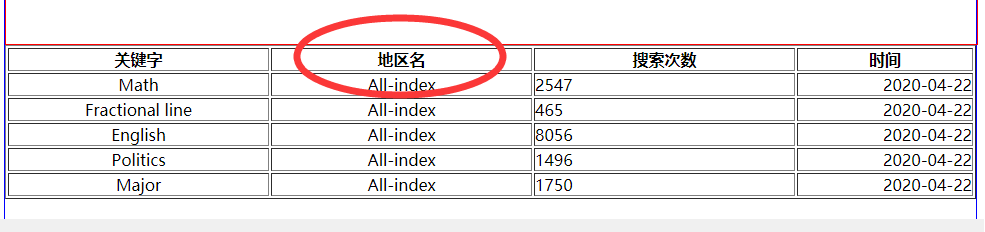
**6.4统计21号内搜索数学关键字次数最少的三个省**



**6.5统计13号到21号，每天搜索分数线关键字次数最多的省**



**6.6统计22号所有地区搜索关键字的次数**



# 7.小结

此次实验数据是使用python的爬虫功能在百度指数上爬取所得，并且会定时获得新的数据，具有实时性，客观性，便捷性。并且小组中的每个成员都编写了相当一部分代码，例如python的爬虫代码，亦或是编写sql的脚本的代码，都是由小组不同成员分工完成并且经过不断的删改整合而来。真正的做到了一个项目小组应该做到的事情，也体现了小组合作的精神。并且我们的系统也是基于大三学生的需要而设计，可以满足部分大三学生对于考研的部分消息的需要，所以说这是一个很有发展潜力的系统。当然这个系统也有弊端，例如中文优化，例如缺少更加详细更加具体的情报，所以说我们也会不断改进。