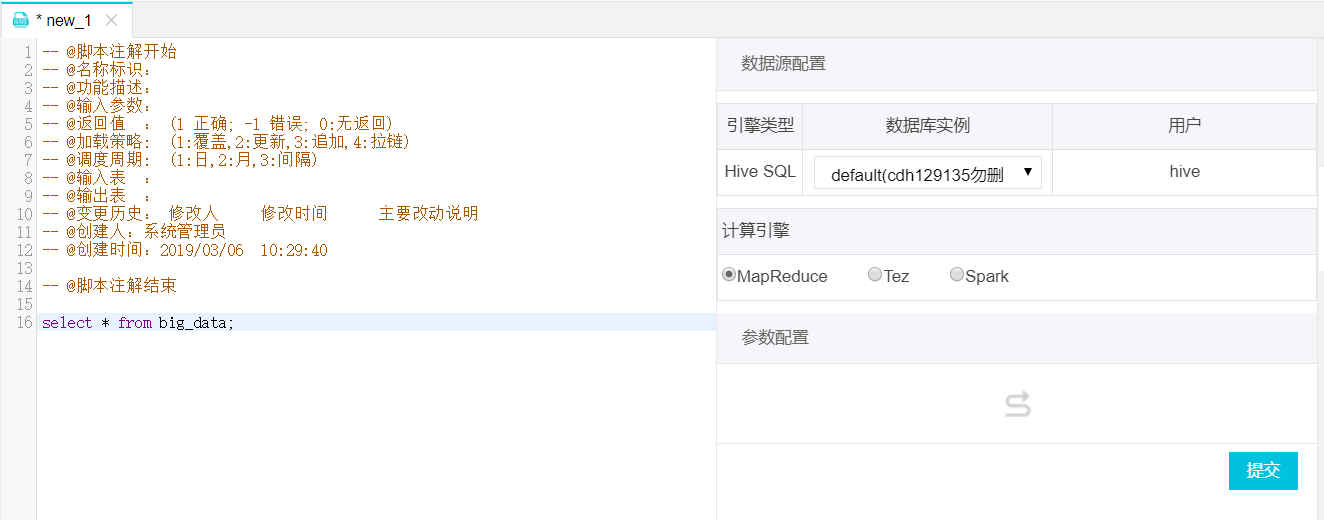
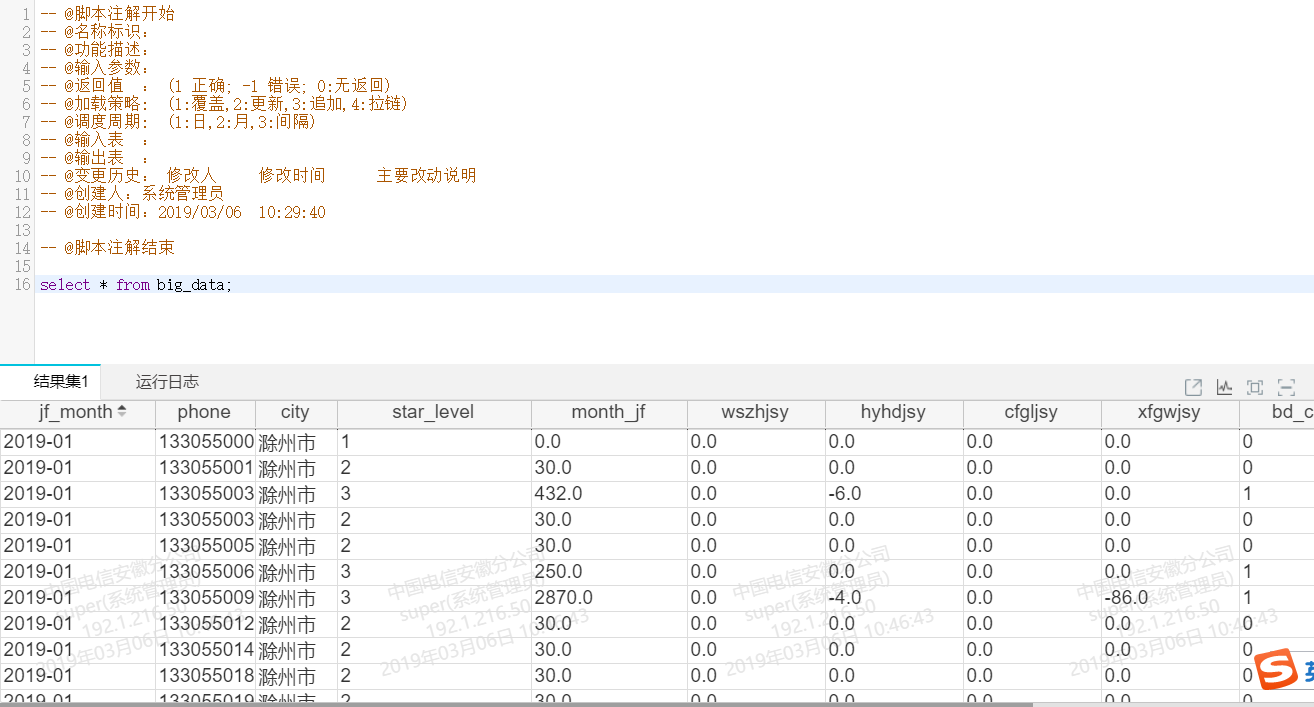
使用安徽电信ddf生产环境1.8g文件为测试对象，事先在116测试环境中建表和导入1.8g数据。查询语句和参数设置如下：





分别使用以下两种方式进行测试。

导出方式：

1. 查询结果集生成文件：

无法生成，查看日志发现执行数小时后报堆内存溢出异常；

1. 使用hdfs生成和合并文件：

正常生成，查看日志显示花费5分半钟左右，列出具体步骤所耗时间：

1）、查询表头和生成表头文件耗时：0.293秒，查询表头使用了在sql后追加limit 0的方式来减少查询时间，生成的表头文件以0为名称，后面介绍原因；

2）、执行sql: insert overwrite directory '/tmp/super/1551839243145' select \* from big\_data耗时267.046秒；

3）、复制表头文件到hdfs：hadoop fs -put /ddf/jar/116/super/1551839243145/0 /tmp/super/1551839243145，耗时：5.72秒；

4）、hdfs合并文件到本地：hadoop fs -getmerge /tmp/super/1551839243145 /data/static/file/ddf/116/out/一点八G文档.csv，耗时：38.026秒；

5）、将hive表分隔符转换为逗号分隔符，执行命令：cat /data/static/file/ddf/116/out/一点八G文档.csv | tr '\001' ',' > /data/static/file/ddf/116/out/一点八G文档\_1551839554234，耗时：20.062秒；

6）、删除旧文件一点八G文档.csv，将新文件一点八G文档\_1551839554234改名为一点八G文档.csv，耗时4.927秒。

后面删除hdfs目录等操作使用了创建线程，异步处理，不包括在导出耗时中。

几个问题：

1. 为什么将表头查询结果生成文件，是为了上传到hdfs上，和查询结果放在同一个目录下，方便合并文件。避免了在合并完成文件后又去插入表头的耗时工作。
2. 为什么将表头文件以0命名，是因为hdfs查询的结果都是000000\_0、000000\_1、000000\_2……形式，所以表头文件名称0可以排在目录首位，这样就能把表头合并到文件首行。常规的插入表头方法是解析文件插入表头，该方法步骤繁琐且耗时；其次还有使用linux命令：sed插入表头，经测试在1.8g文件首行插入表头耗时80多秒。这两种方案最终被废弃，使用了现在的方法。