实验题目

日志解析及计算：MR

实验要求

1、在公共集群中创建以自己用户名相同的文件夹

2、从bigdata4服务器中的tmp目录中找到log0605.txt上传到公共集群中自己创建的目录

3、通过mr程序解析日志文件

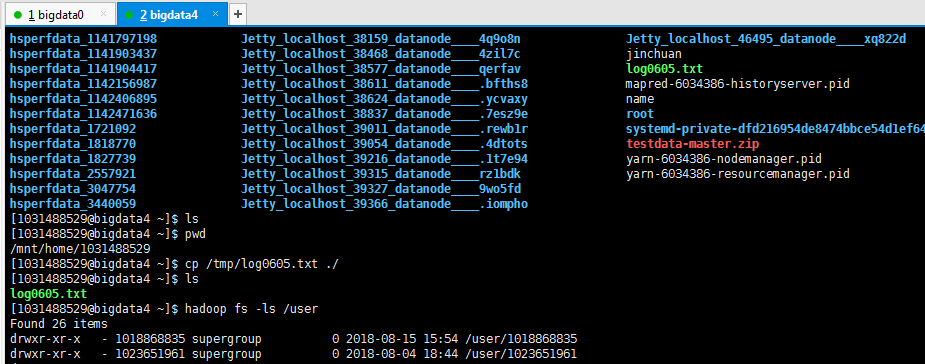
4、计算日志文件中日志条数及用户id去重后的个数，即：计算不重复的用户id个数（如果能在一个mr中完成更好）

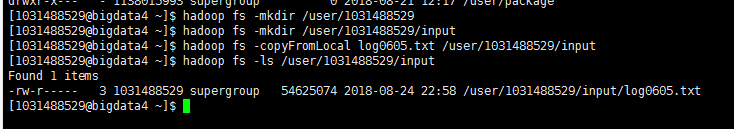
5、将操作过程及截屏写入实验报告文档中，并将代码带包一起以附件形式上传

注：实验报告文档模板已上传至QQ群，同学们请前往群文件自行下载哦

操作步骤

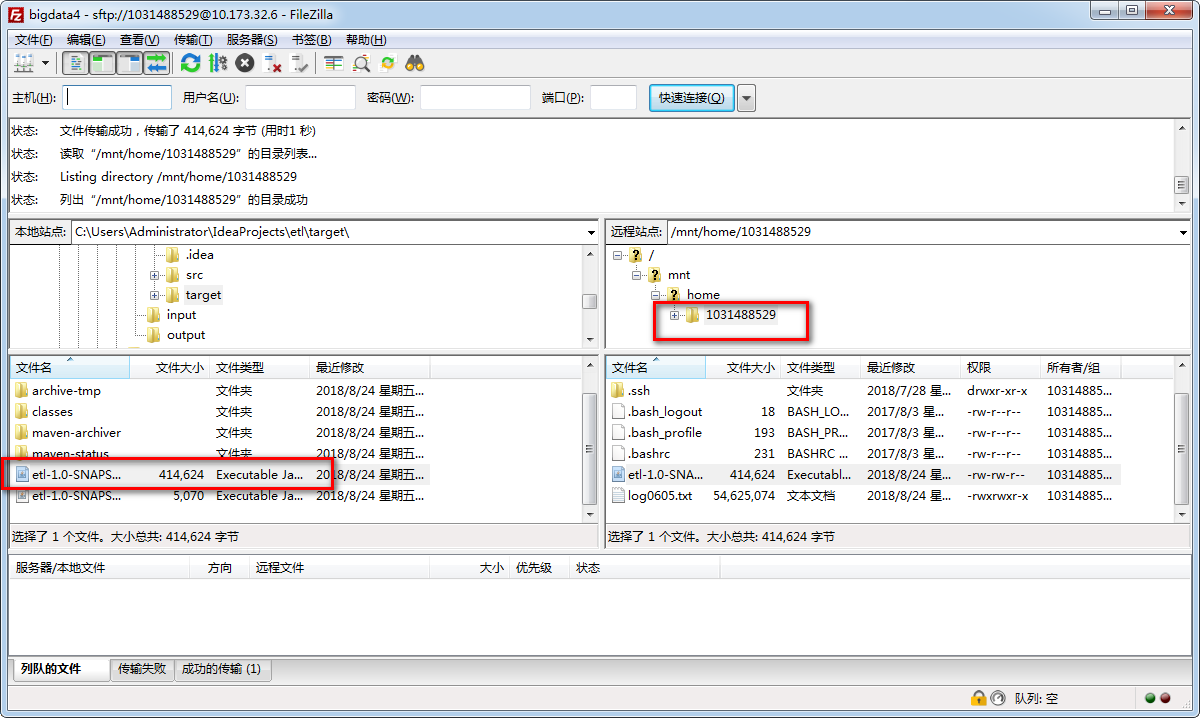
两个步骤1、2截图





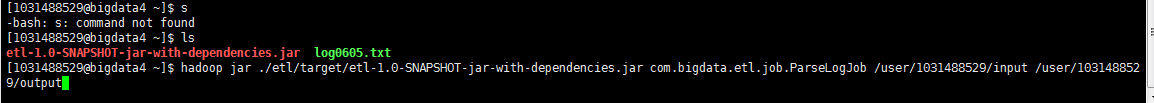
步骤3：

编写程序，打包，用FileZilla上传



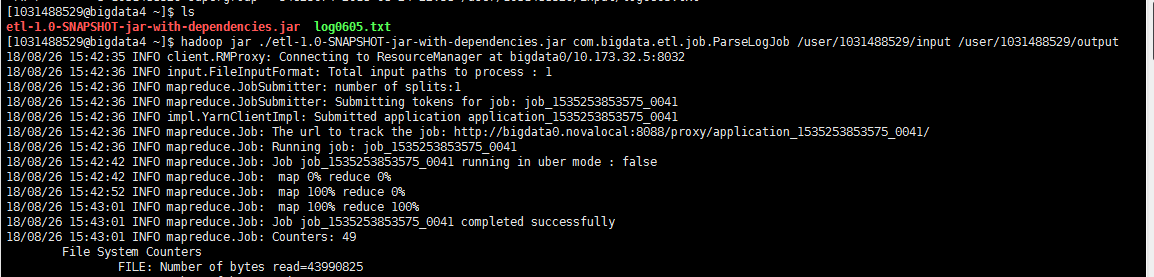
集群bigdata4上执行jar包命令：

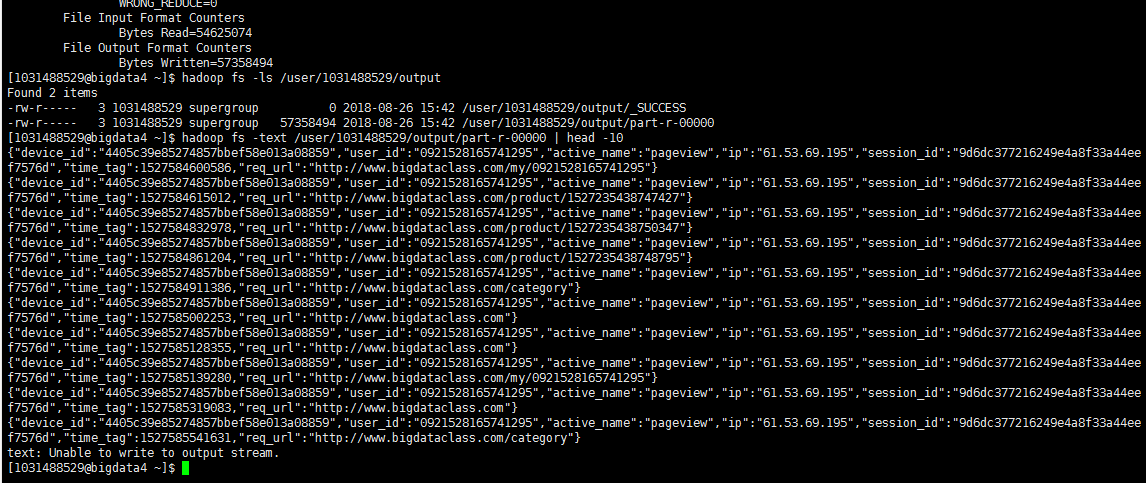
hadoop jar etl-1.0-SNAPSHOT-jar-with-dependencies.jar com.bigdata.etl.job.ParseLogJob /user/1031488529/input /user/1031488529/output



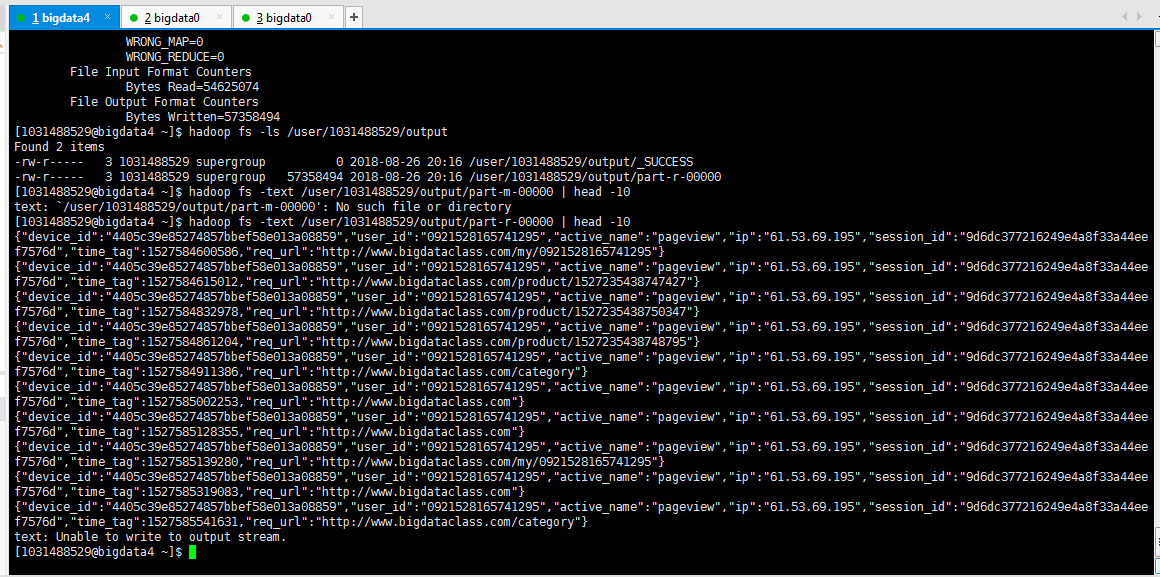


使用LogBeanWritable改造后：

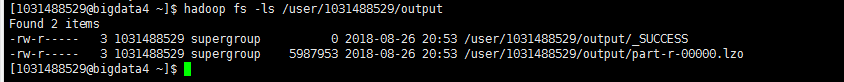




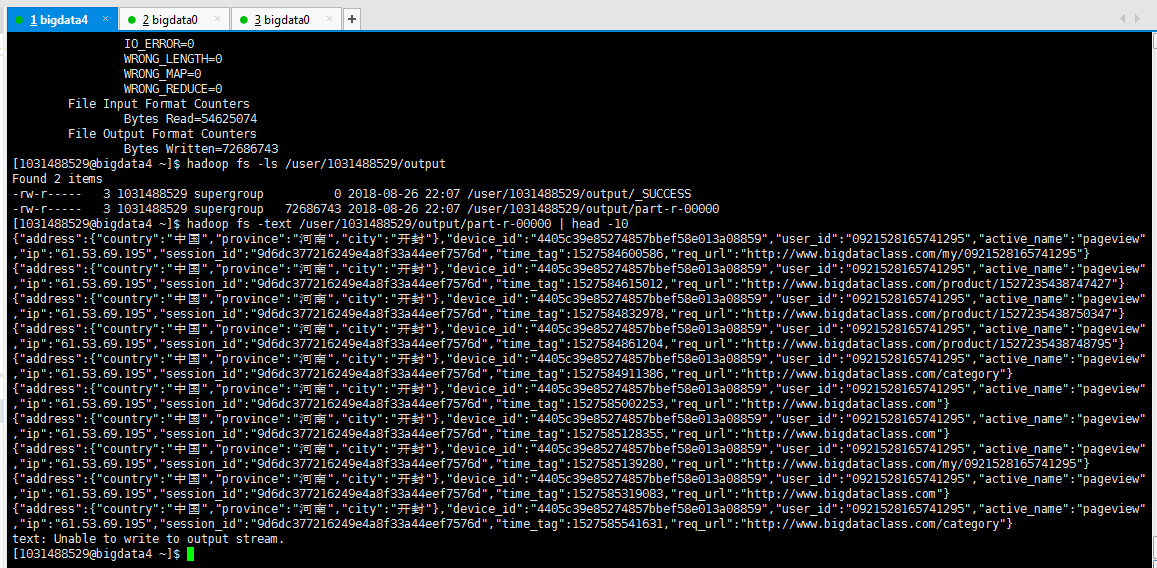
用LogGenericWritable改造后：



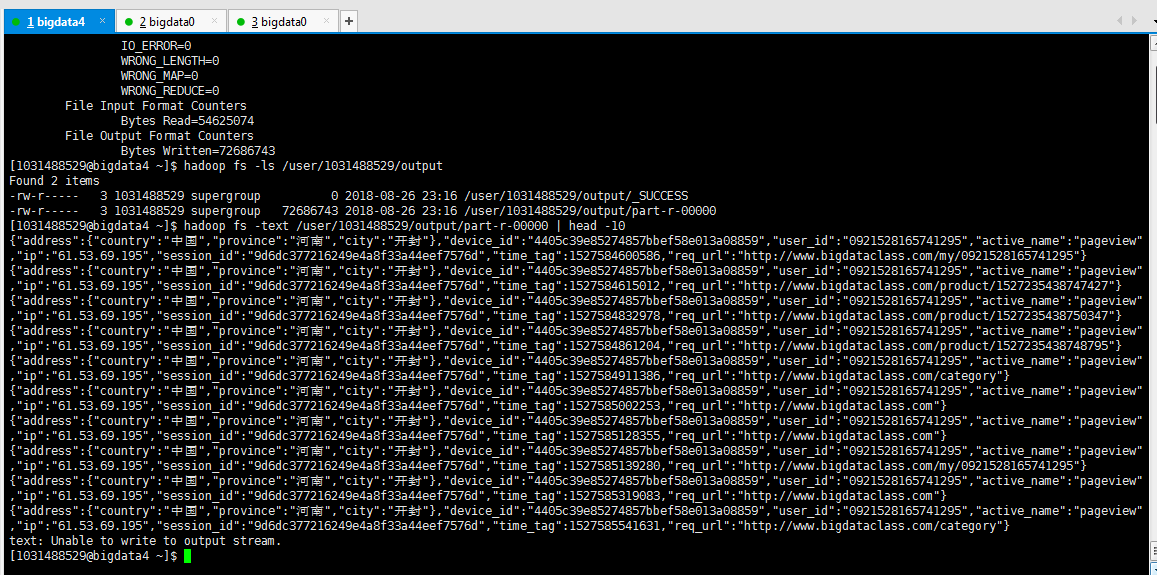
启用压缩（之后会取消压缩）：



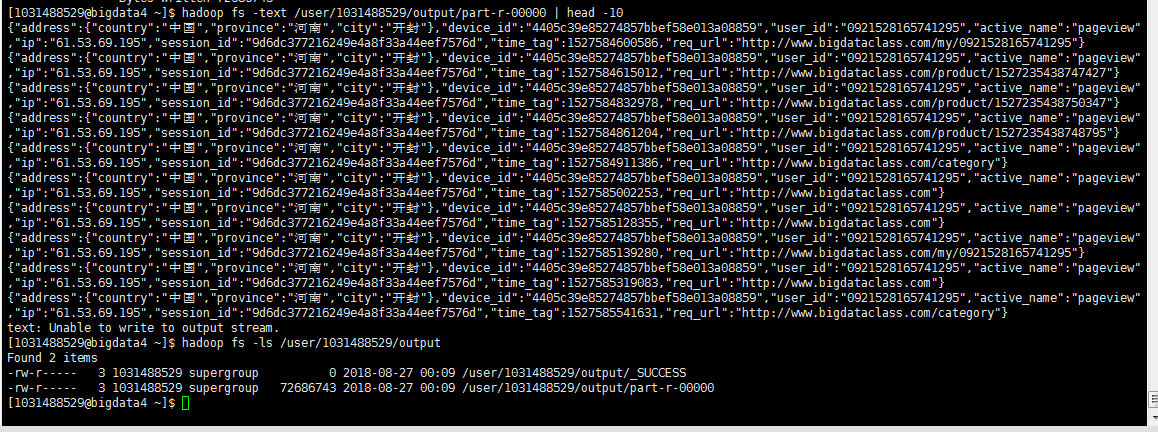
取消压缩、解析IP



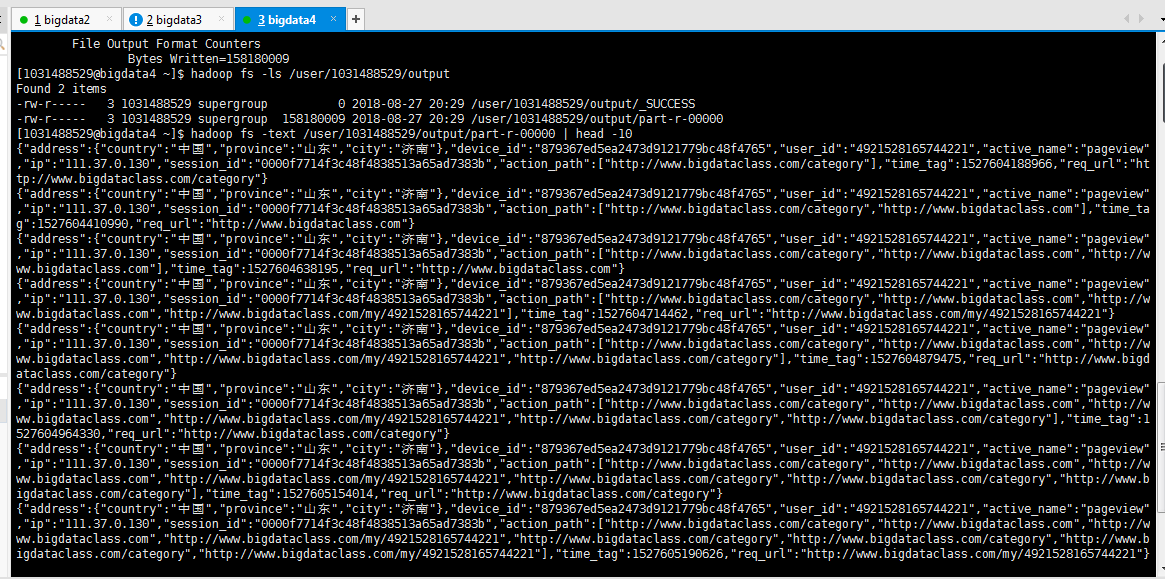
Ip库文件传参：



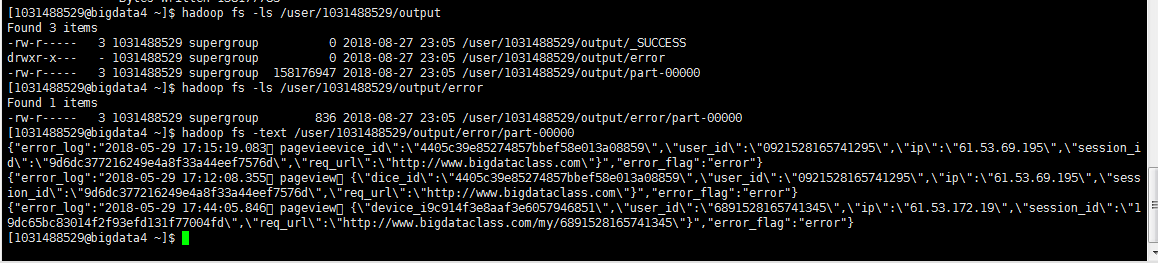
分布式URI



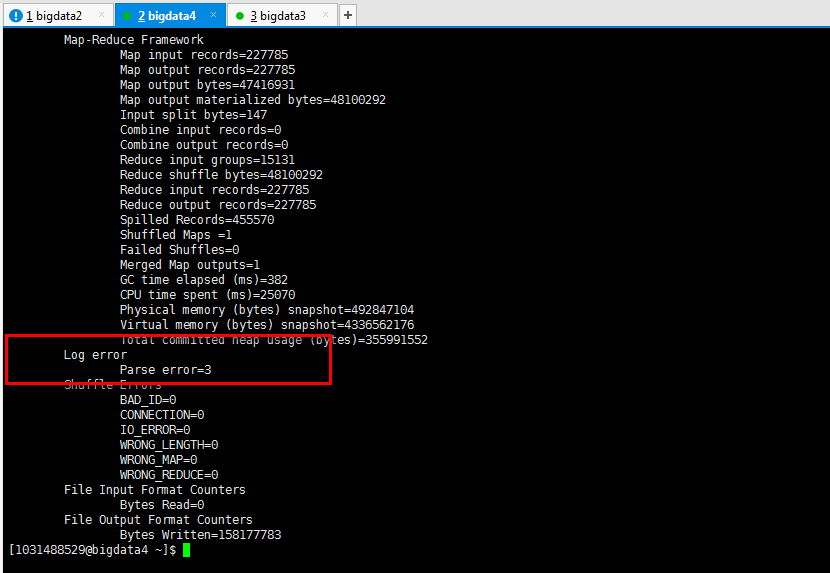
访问路径二次排序：



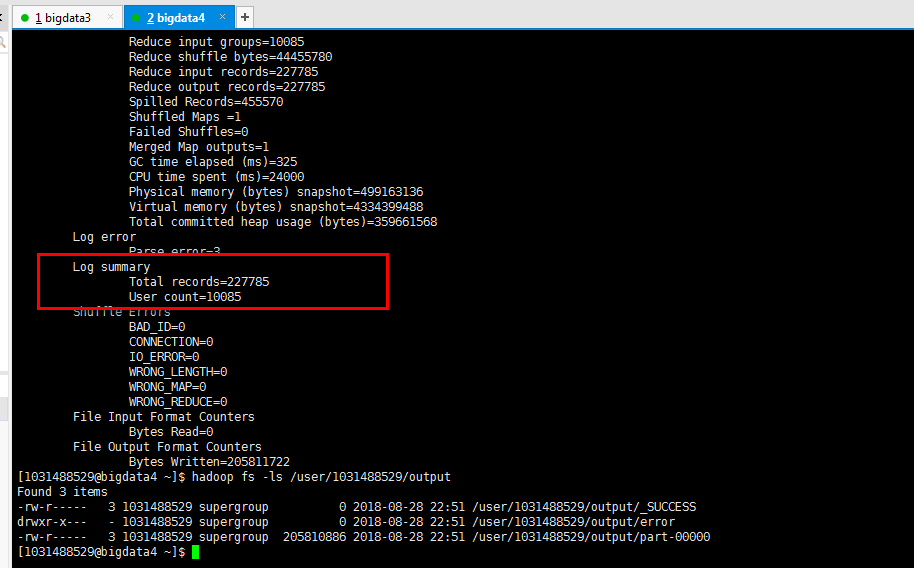
分开输出异常日志：



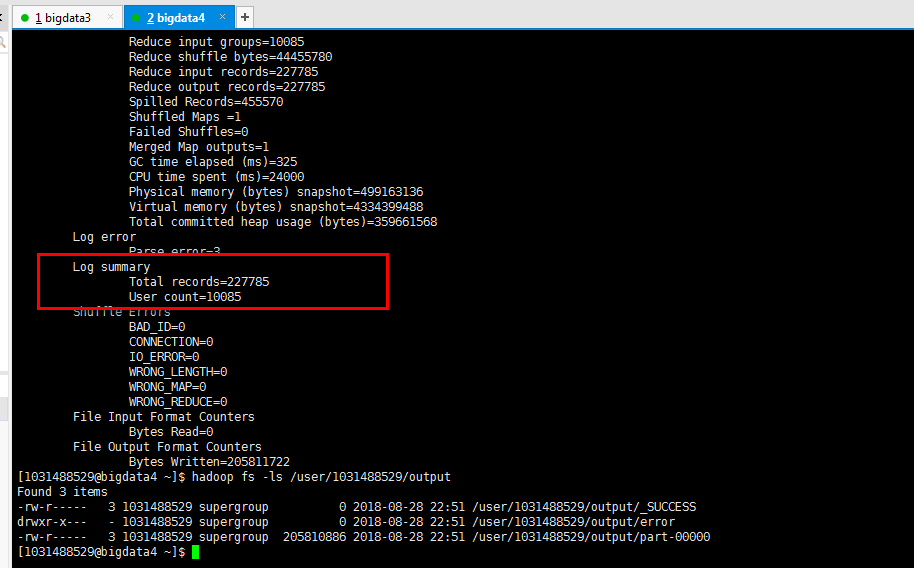
记录异常的计数器：



4、计算日志文件中日志条数及用户id去重后的个数，即：计算不重复的用户id个数（如果能在一个mr中完成更好）



实验结果



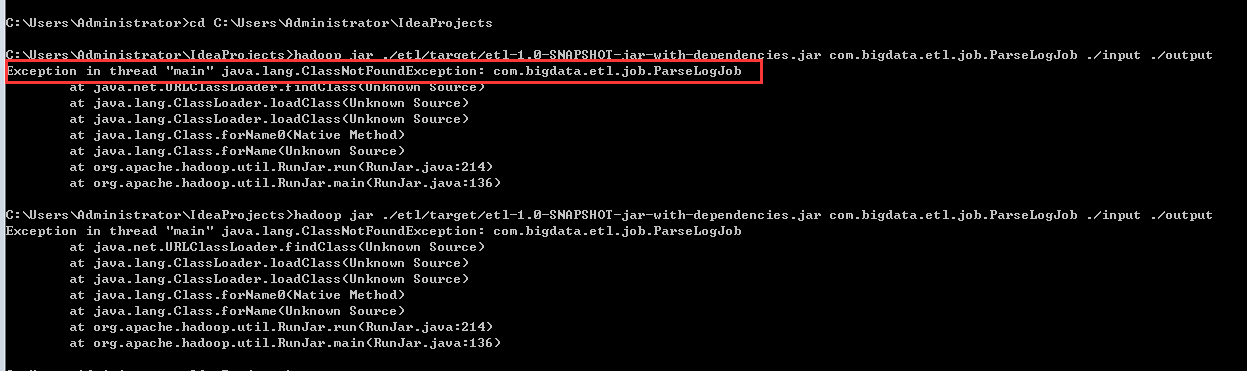
总结反思

1，编写ParseLogJob的时候有些类容易引用错，需要仔细确认，把错的类重新import；

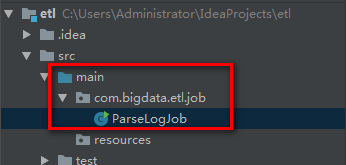
2，本地window系统配置完hadoop，执行hadoop命令报错，JAVA\_HOME is incorrectly set. Please update D:\hadoop-2.7.6\conf\hadoop-env.cmd

把文件里面这行改为：set JAVA\_HOME=C:\java\jdk1.8.0\_181，问题解决；

3，按照视频编写完，执行时报错ClassNotFoundException



通过把com.bigdata.etl.job移到main下面，删除空的java包，问题解决



4，需要反复解读课程代码，理解MR的原理，才能编写自己的逻辑；