**导入新项目大体上和导入新事件差不多**

# 批量导入

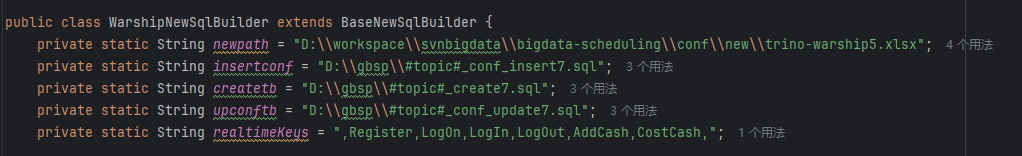
1.导入工程bigdata-scheduling，需要通过里面的程序生成建表语句和配置脚本

2.把新事件的转义加入bigdata-scheduling\conf\new下相应的excel内，注意最后一行的END

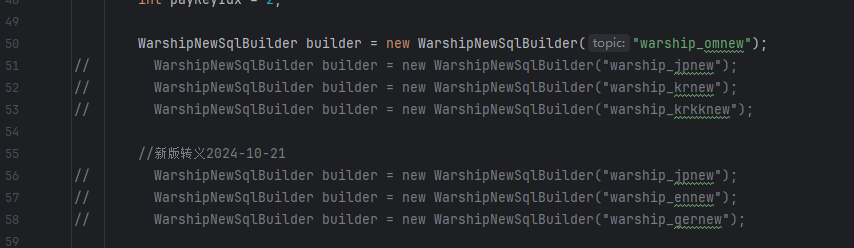


3.在工程里面相应项目的配置生成类（所有项目的都在com.sincetimes.bigdata.schedule.load）下面如果有相同项目多个的，选取类名代new的，如海战使用：com.sincetimes.bigdata.schedule.load.warship.WarshipNewSqlBuilder

，注意里面的绝对路径根据本机的项目位置修改：



同一个项目不同地区使用相同的转义，修改一下项目库名：



检查转义文档excel，看看新的事件转义：

4.运行这个类，生成脚本，这里应该生成3个文件，两个是mysql的，一个是hive的（可以搜一下，看看新事件的配置有没有在里面）：

warship\_jpnew\_conf\_insert7.sql

warship\_jpnew\_conf\_update7.sql

warship\_jpnew\_create7.sql

把这三个文件上传到114.132.169.71 的 /data/bigdata/gbsp/ 文件夹内

5.更新mysql导入配置表，连接服务器gzcloud20 ：114.132.169.71，su到bigdata用户，**注意大数据平台的所有操作都需要在bigdata用户下，除非trino的重启，需要trino用户，不要在root用户下操作大数据平台**

cd /data/bigdata/script （这下面有所有项目统计相关腾讯云mysql的连接脚本）

./gzcdb\_bi.sh

use use gbsp-bi-bigdata

在这个库里面，每个项目都会有两个表，导入配置前需要备份，防止导入错误影响其他数据录入：

conf\_log\_warship\_jpnew 新项目，需要创建此表，从其他项目copy一个相同结构的就行

conf\_log\_warship\_jptmp 新项目，需要创建此表，从其他项目copy一个相同结构的就行）

*清除上一次的备份数据（如果没有老数据可以跳过）：*

*truncate table conf\_log\_warship\_jptmp*

*备份现在使用版本数据：*

*insert into conf\_log\_warship\_jptmp select \* from conf\_log\_warship\_jpnew*

*清除现在配置数据：*

*truncate table conf\_log\_warship\_jpnew*

导入新版配置：

source /data/bigdata/gbsp/warship\_jpnew\_conf\_insert7.sql

source /data/bigdata/gbsp/warship\_jpnew\_conf\_update7.sql

检查新版本配置的合法性（下面sql不应该有结果的，如果有，就是同一个字段名有多种类型，必须修改转义后重新生成文件导入）：

select field\_code,count(distinct field\_hive\_type) as cnt from conf\_log\_warship\_jpnew group by field\_code having cnt>1

检查本次更新是否在预期（这里应该是只会输出我们新加或修改的事件的字段，不应该有其他事件出现）：

select a.keyword,a.field\_code,a.field\_idx,a.field\_hive\_type,count(1) as cnt from

(select \* from conf\_log\_warship\_jpnew union all select \* from conf\_log\_warship\_jptmp ) a

group by a.keyword,a.field\_code,a.field\_hive\_type,a.field\_idx having cnt !=2 order by a.keyword,a.field\_idx

6.配置mysql更新完成，下面创建hive表，在刚才的71服务器，

cd /data/bigdata/gbsp/

hive -f warship\_jpnew\_create7.sql

7.检查hive表，在71服务器，执行命令

hive

**注意：这个界面为hive最底层控制界面，在这个界面的所有操作不会有权限控制，做表删除操作时需要谨慎**

查看刚才的表有没有创建成功：

show create table warship\_jpnew\_r.dwd\_gserver\_live

至此，批处理任务已经能够定时导入数据了，如果需要导入历史数据，继续执行后面操作：

1.登录大数据任务平台，创建任务：

<http://gzcloud21:9090/main>

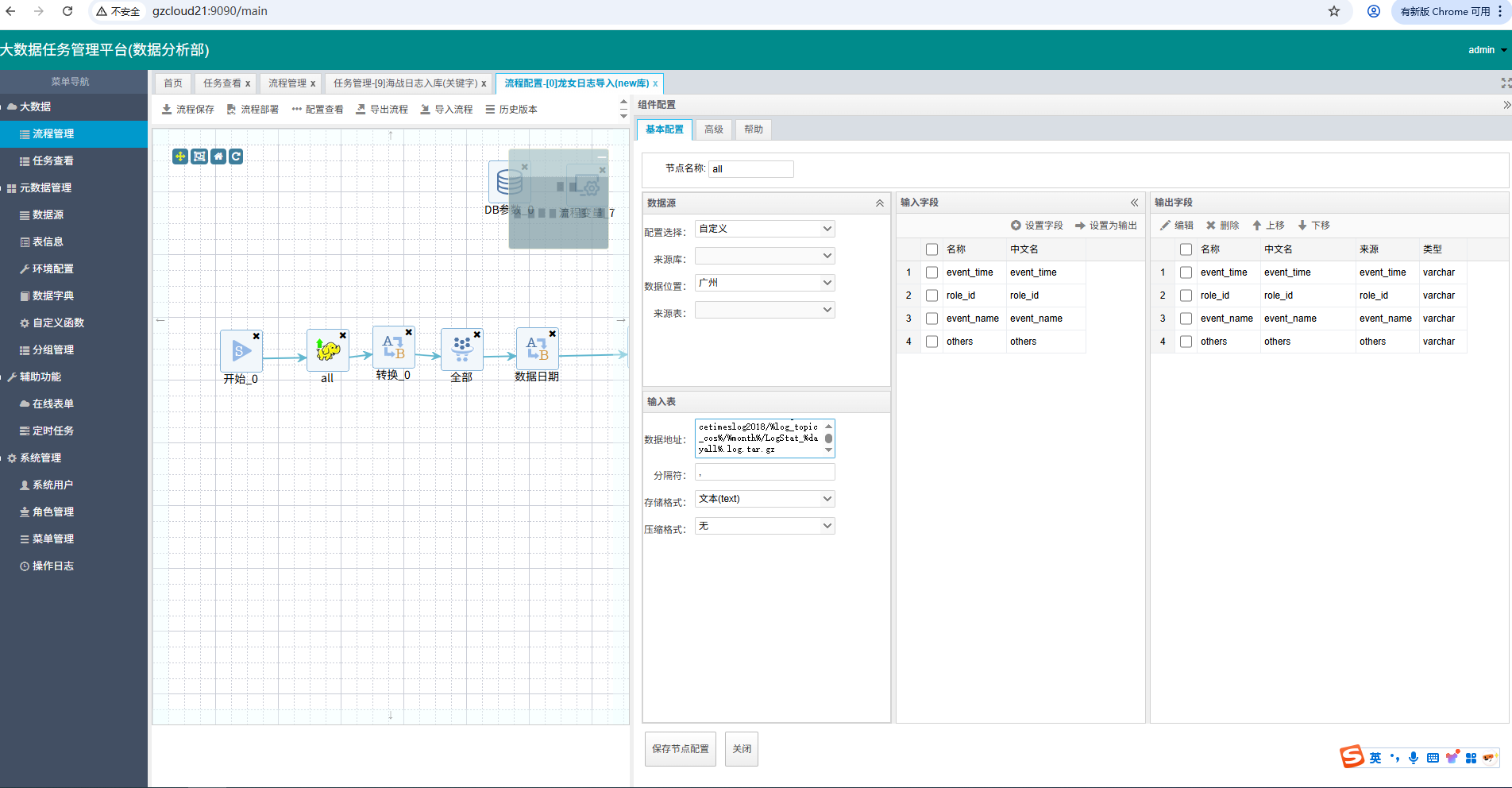
任务管理，



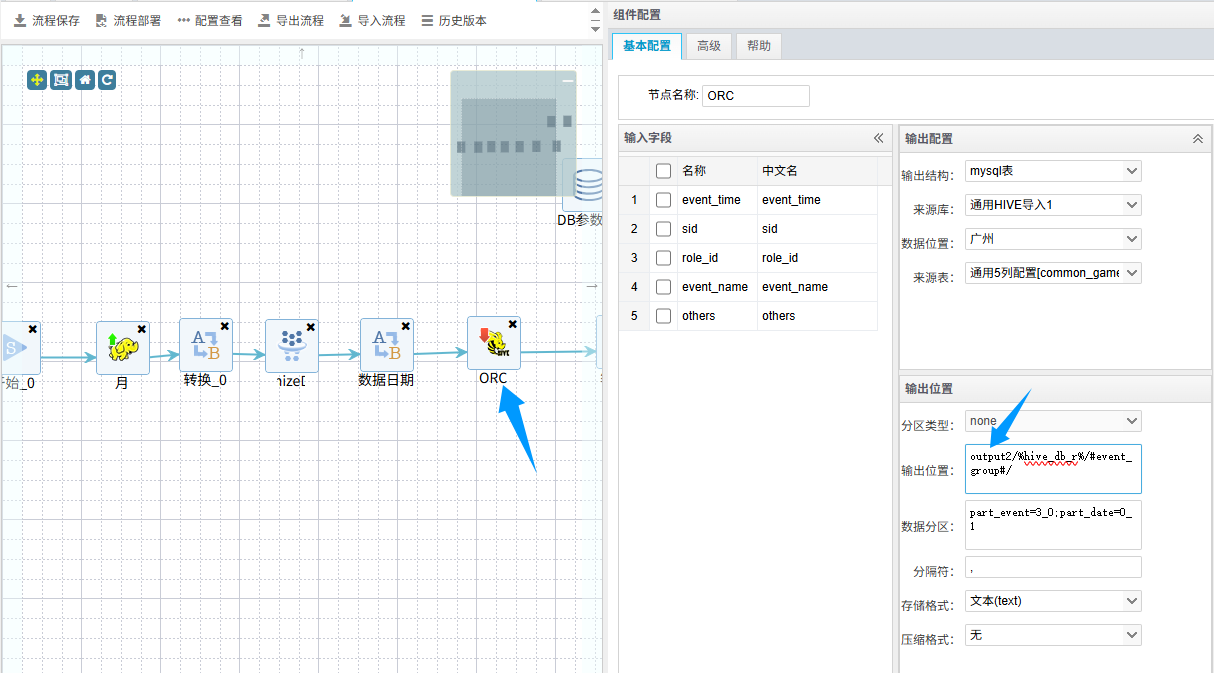


创建一个新流程，从其他天定时任务中copy一个过来。

日志一般按天定时导入，注意配置：



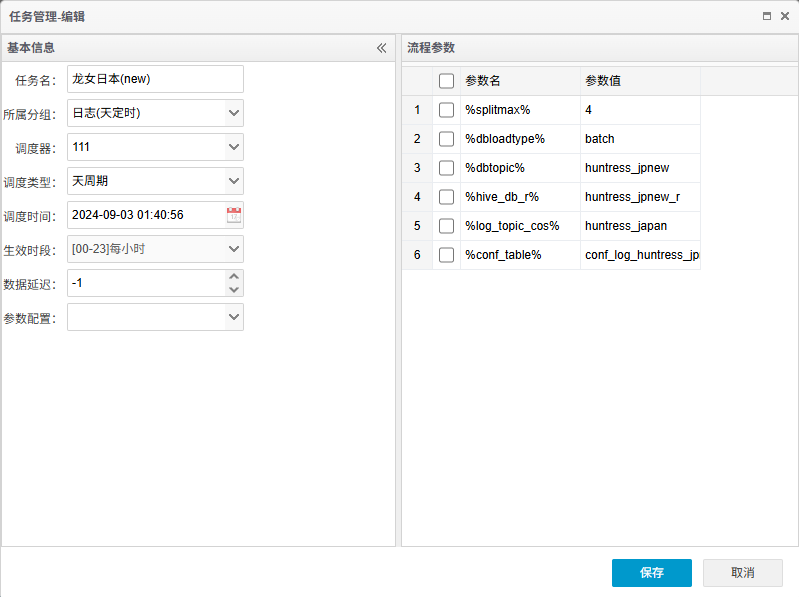
修改日志输出位置（很重要）：



**注意（很重要）：一定要改，不要输出到hive，防止出问题污染现网其他事件数据。**

**新项目可以直接输出到hive，不过最好测试的时候先输出到其他位置**

保存并部署流程，创建任务开始跑数据，创建任务注意各项配置：



调度器：gserver日志为111，adjust是另一个，各个调度器有文档说明

调度器新项目可以使用211，和111相比有一个优化，字符串为空串是hive库中会存为NULL 而不是空字符串

调度类型：按天定时

Splitmax 参数：要按照流程的reader来，不同项目可能不一样

log\_topic\_cos：数据在cos上存储目录，可以去定时调度里面找或者直接查看cos的目录结构

2.开始跑数据，可以批量跑任务，注意任务时间间隔，现在不建议晚上跑大任务，防止影响trino入库任务，**跑任务时，需要在yarn平台监控任务状态，如果因为这个任务影响到了其他任务，导致yarn任务积压，必须手动kill掉，想其他办法（比如按天或者先拉日志到本地hdfs再导入）**

Yarn任务页面：<http://gzcloud19:8088/cluster/apps/RUNNING>

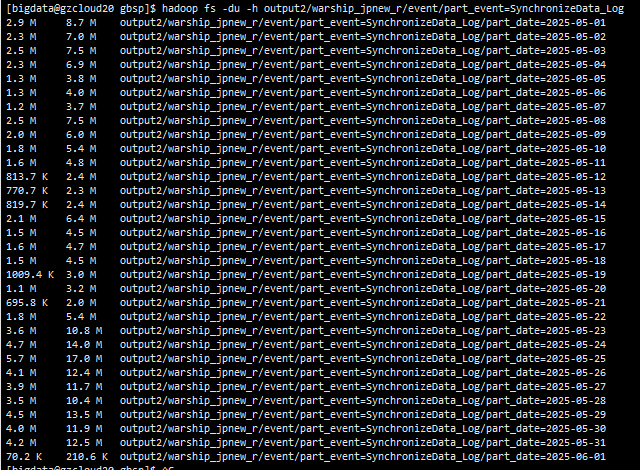


3.查看任务结果，查看hdfs上刚才配置的数据目录

hadoop fs -du -h output2/warship\_jpnew\_r/

正常下面应该生成了需要补录的日志的数据

hadoop fs -du -h output2/warship\_jpnew\_r/event/part\_event=SynchronizeData\_Log



4.查看刚才hive创建的新表分区下文件，如果是老表重新导数据，这里面会有文件，需要手动删除

查看文件

hadoop fs -du -h hive/warship\_jpnew\_r/event/part\_event=SynchronizeData\_Log

移动文件到分区下面

hadoop fs -mv output2/warship\_jpnew\_r/event/part\_event=SynchronizeData\_Log/\* hive/warship\_jpnew\_r/event/part\_event=SynchronizeData\_Log/

5.修复hive分区，在hive命令行：

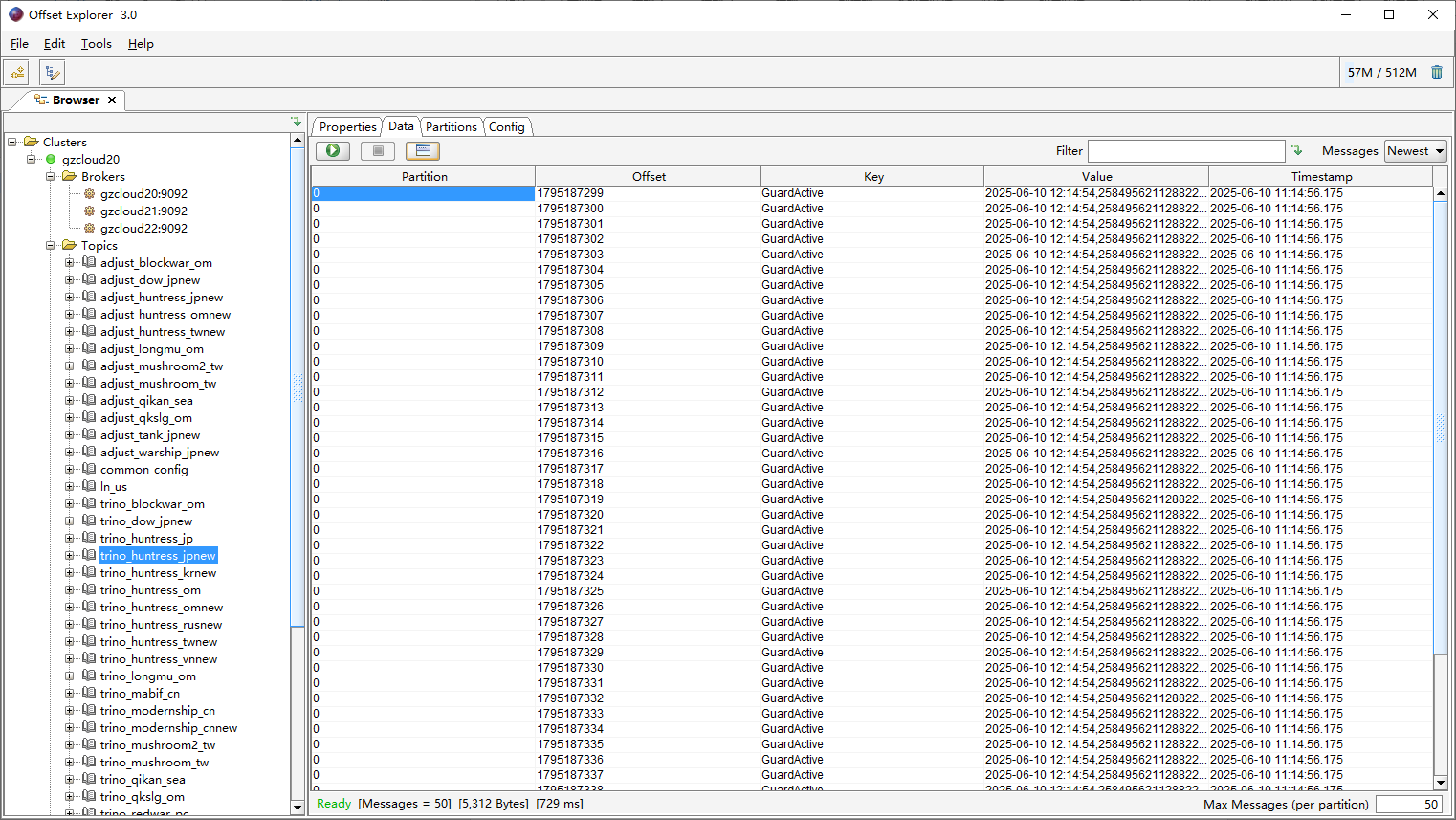
msck repair table warship\_jpnew\_r.dwd\_gserver\_synchronizedata\_log\_live

6.导入完成，已经可以在trino中查询到数据

# 实时导入

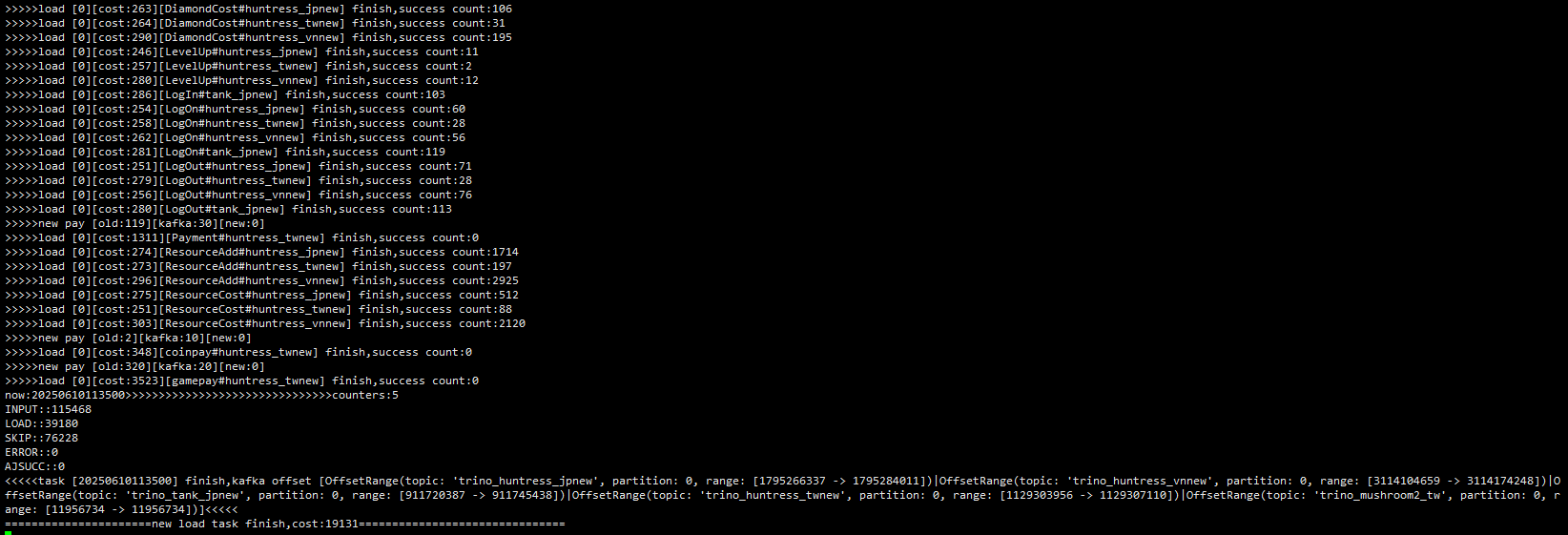
新项目批量导入配置测试完成后，下一步部署实时导入配置

1.部署flume实时采集，可以查看《3.flume数据采集》文档，部署完成后，kafka应该能查到实时数据：



2.部署实时入库流程， 步骤在 《6.日志定时导入和实时导入》文档

3.spark streaming从日志中查看实时的日志事件导入情况



4.从trino库查看日志

# 分区切换

当天需要把明天数据分区提前创建好，这样凌晨跨天是，spark streaming入明天数据完成能及时查到数据

