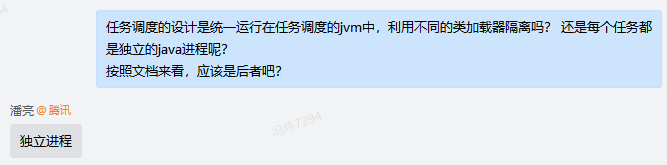
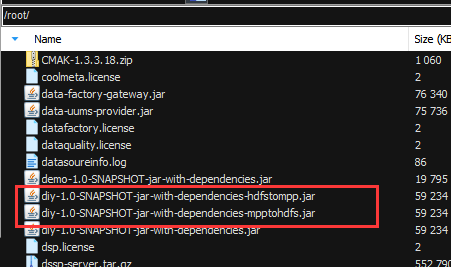
# 背景

工作提交到tbds任务调度有问题，根据潘工所说，每个任务都是一个java进程，因此将工程打包成fat-jar放到机器上测试。





项目说明：

配置问题：

项目中采用了xml和网页传入参数两种配置方式，网页参数会覆盖xml参数，因此本jar包对于本地环境和任务调度环境应是通用的。

Mpp-hdfs和hdfs-mpp 的功能切换仅需要按sample.xml替换conf.xml即可。

项目启动类：

任务调度中入口类需要继承tbds jar包中的类com.dfjx.diy.DataSyncRunner，非任务调度中入口为实际的做任务的类com.dfjx.diy.DataSync。

任务调度网页填写的主类com.dfjx.diy.DataSyncRunner所作的仅仅是调用了实际做任务的类，其他传参、结束等方法均从demo复制。

DataSync dataSync = **new** DataSync();  
 dataSync.doWork(params);

# 本地调用

## Mpp到hdfs

java -cp diy-1.0-SNAPSHOT-jar-with-dependencies-mpptohdfs.jar com.dfjx.diy.DataSync

因为hdfs写入速度较慢，因此采用读取一批写入一批的方法。

使用自定义阻塞队列

队列满时唤醒写线程

写线程阻塞在带超时的take方法上，一次将全部数据拿走

过程分解：







## Hdfs到mpp

java -cp diy-1.0-SNAPSHOT-jar-with-dependencies-hdfstompp.jar com.dfjx.diy.DataSync

