**风控系统批处理开发流程指导**

目 录

[前言 1](#_Toc13325)

[术语定义 1](#_Toc1830)

[一款国外开源的ETL工具，现改名PDIPentaho Data Integration 1](#_Toc24322)

[1. 一次性作业编写 2](#_Toc4330)

[1.1. 一次性作业简介 2](#_Toc2029)

[1.2. 编写kettle文件 2](#_Toc4389)

[1.3. 编写sql文件 2](#_Toc11625)

[2. 增量作业编写 2](#_Toc26820)

[2.1. 增量作业简介 2](#_Toc7787)

[2.2. 增量作业的编写 2](#_Toc15810)

[2.2.1. 添加数据日期 2](#_Toc30859)

[2.2.2. 新建数据库连接 3](#_Toc12464)

[2.2.3. 编写kettle文件 3](#_Toc30194)

[2.2.4. 编写sql文件 3](#_Toc10124)

[3. 作业部署 3](#_Toc28493)

[3.1. 元数据拷贝 3](#_Toc8270)

[3.2. 上传kettle文件 3](#_Toc23214)

[3.3. 上传sql文件 3](#_Toc20688)

[4. 定时任务配置 4](#_Toc5717)

[4.1. 定时任务添加 4](#_Toc18170)

[4.2. 定时任务日志与问题排查 4](#_Toc1696)

版本信息

|  |  |  |  |
| --- | --- | --- | --- |
| **版本号** | **修订日期** | **修订内容** | **编辑人** |
| V1.0 | 2019.11.21 | 新建 | 李德和 |
| V2.0 | 2020.01.09 | 开发步骤详细描述 | 李德和 |
|  |  |  |  |
|  |  |  |  |

**前言**

本文档旨在指导用户完成批处理任务的开发和部署。本文假设批处理环境已经完成搭建（参考文档《批处理环境搭建.docx》），并且参与批处理开发的人员已经掌握以下技术：Linux常用命令、shell脚本的编写和调试、SparkSQL(或者Hive SQL)、HDFS原理和常用操作、Hive、kettle任务的编辑。

本文在上述假设的条件下，对批量开发作出一些指导和规范。

**术语定义**

|  |  |
| --- | --- |
| Kettle | 一款国外开源的ETL工具，现改名PDIPentaho Data Integration |
| 执行服务器 | 指脚本存放和执行的服务器 |
| 作业目录 | 执行服务器：/home/授权用户名/jobs/ |
| 脚本目录 | 执行服务器：/home/授权用户名/shell/，在该目录下按照功能另建子目录 |

# 一次性作业编写

## 一次性作业简介

一次性作业指的是指正常情况下只执行一次（区别于每日、每周等）的作业，常用于新开发一个批量业务时的建表操作。作业的名字末尾最好为Init。

## 编写kettle文件

在可视化界面中使用kettle完成待调度任务的开发，生成 kjb文件。参考案例：



## 编写sql文件

一次性作业的Sql文件用于建表，给kettle文件调用。参考案例：



# 增量作业编写

## 增量作业简介

增量作业指按照一定的时间间隔重复执行的作业，如每日、每周、每月等等。增量作业用于批量业务的具体数据处理。作业的名字末尾最好为main。

## 增量作业的编写

### 添加数据日期

登录MySQL数据库，在数据库kettle的表batch中插入记录。

由于涉及到操作数据日期，如递增、回退、跳批等，需要把数据日期保存到数据库里。内容为作业名称和数据日期。

注意：

作业名不能重复；

数据日期请填写首次执行作业要处理的数据文件的日期。

### 新建数据库连接

在kettle中新建连接，连接到MySQL服务器的kettle数据库。

### 编写kettle文件

在可视化界面中使用kettle完成待调度任务的开发，生成 kjb和ktr文件。参考案例：



### 编写sql文件

Sql文件用于处理数据，给kettle文件调用。参考案例：



# 作业部署

## 元数据拷贝

在可视化界面中使用kettle完成待调度任务的开发和测试，涉及到链接数据库时，数据库信息均保存在用户目录下的.kettle文件夹中。为了使作业到服务器上仍然能够执行，需要把该目录拷贝到服务器上用户目录下。

特别注意，每次涉及到数据库更新时，均需要重新上传该目录。

## 上传kettle文件

将开发的所有kettle文件上传到服务器的作业目录内。

## 上传sql文件

将开发的所有sql文件上传到服务器的脚本目录内。

# 定时任务配置

## 定时任务添加

登录管理台，系统监控-》定时任务，即可编辑添加定时任务。有以下几点需要注意：

1. 定时任务名称

任务名称描述任务的功能，必须是唯一的。

1. 定时任务类名

使用开发好的通用类即可：org.jeecg.modules.quartz.job.BatchJob

如果这个类不满足用户需求，请参考jeecg文档开发新的类。

1. 定时任务参数

定时任务的参数为在可视化界面编辑好的kettle文件，如xxx.kbj。需要注意的是，该文件一定要放置在作业目录。

特别要注意一次性作业和增量作业的定时配置的区别。

## 定时任务日志与问题排查

按照导航“系统日志-》日志管理-》定时任务日志”查看定时任务日志，对于执行失败的任务，可以到作业路径，日志文件名为“kjb文件名\_处理日期.log”；或者脚本路径下查看对应日志，日志文件名为“main\_数据日期.log”和“error\_数据日期.log”。异常发生时一般先查看error日志。

注意，数据处理日期指 作业运行的日期；数据日期指数据产生的日期。