数据准备_ETL作业帮助文档_V1.0

1. 概述

1.1 版本

报表服务器版本	JAR包版本	插件版本
10.0	2021-06-01	V1.2.1

1.2 应用场景

1. BI、FR客户均可能面临的问题
   a. 数据分散：业务数据分散在各类信息系统中，包括线上、线下等，结构不统一，汇总分析困难；
   b. 数据口径不一致：字段命名规范在公司内多个口径，命名不规范和逻辑不统一的字段，造成认知歧义。
2. FR客户面临的特有问题
   a. 无法跨库取数：报表数据来源多个数据库，虽然FR支持通过单元格过滤进行关联，但是影响报表展示速度，同时不支持分组汇总等功能；
   b. 复杂sql导致开发成本高、取数慢：报表数据集内使用大量复杂sql进行数据处理，sql开发和维护成本都很高，导致sql取数很慢，影响报表展示速度；
   c. 报表数据集无法被引用关联：报表内已有数据集无法被新数据集引用，无法和其他数据集关联。

1.3 功能简介

1. 数据开发：提供数据集成和sql脚本开发等数据处理功能；
   a. 离线同步：当需要进行跨数据库的数据抽取时，需要使用离线同步组件完成；
   b. SQL脚本：当需要对某数据库内的数据表进行创建、更新、删除、读取等操作时，需要使用SQL脚本组件。
2. 任务运维：提供工程内所有任务的整体概览，每个任务的调度配置、每次运行的详细信息等内容。

2. 插件安装

1. 插件下载
   a. 联系帆软产品运营Bernard获取（Tel：15076078933），帆软工作人员会发送插件安装压缩包，如下图所示：
   

   b.
2. 设计器安装
   a. 在FineReport设计器中点击服务器>插件管理，点击从本地安装按钮，选择已下载的插件.zip压缩包，如下图所示：（注：请勿解压上文得到的压缩包，直接安装即可。）

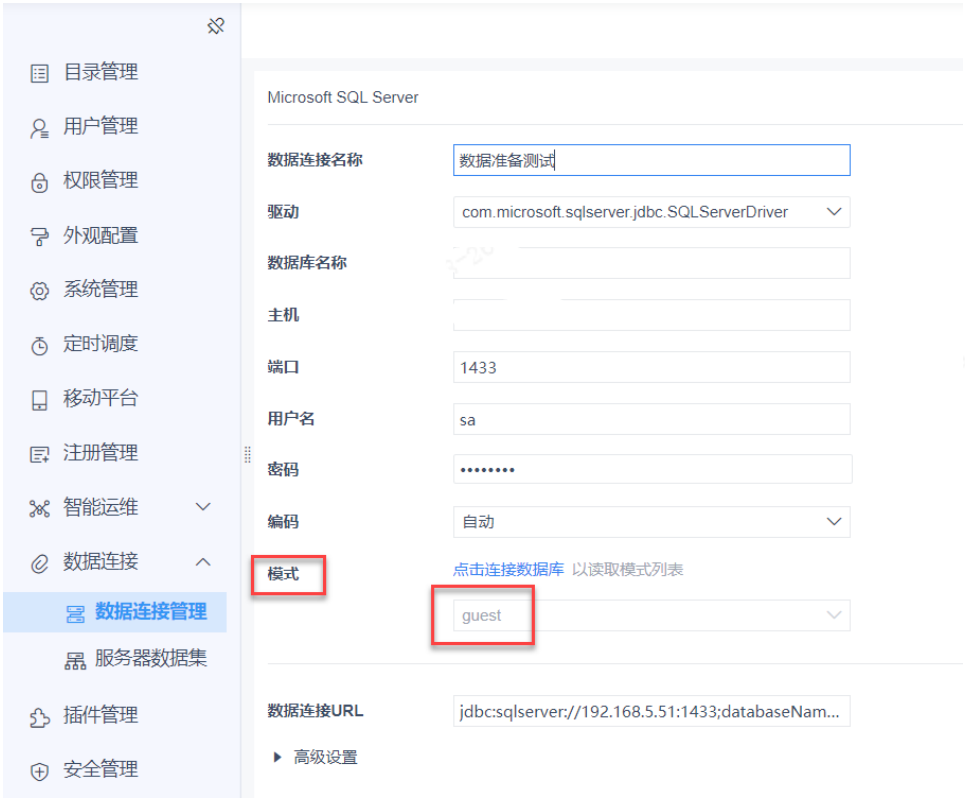
b.

3. 服务器安装
   a. 以管理员身份进入数据决策系统，点击管理系统>插件管理>从本地安装，选择.zip压缩包，如下图所示：（注：请勿解压上文得到的压缩包，直接安装即可。）
   b.

3. 插件使用

3.1 目标数据库配置

1. 在ETL作业使用之前，需要在「管理系统>数据连接」中绑定一个新的数据库，建议数据库存储空间和报表工程业务数据库相近；
   a. **当前版本插件支持MySQL、SQLserver、Oracle、PostgreSQL，4种数据库；**
   b. 由于ETL作业涉及从来源库抽取数据至目标库，所以需要指定目标库的字符编码、解码格式；
      - 以Mysql数据库举例，我们需要在数据连接的url后增加参数，useUnicode=true&characterEncoding=UTF-8，添加参数后的url样式：jdbc:mysql://localhost:3306/prep?useUnicode=true&characterEncoding=UTF-8
   c. 因为ETL作业需要读取目标数据库表结构，所以**除MySQL类型数据库外，其他类型数据库均需要指定模式**，样式如下：



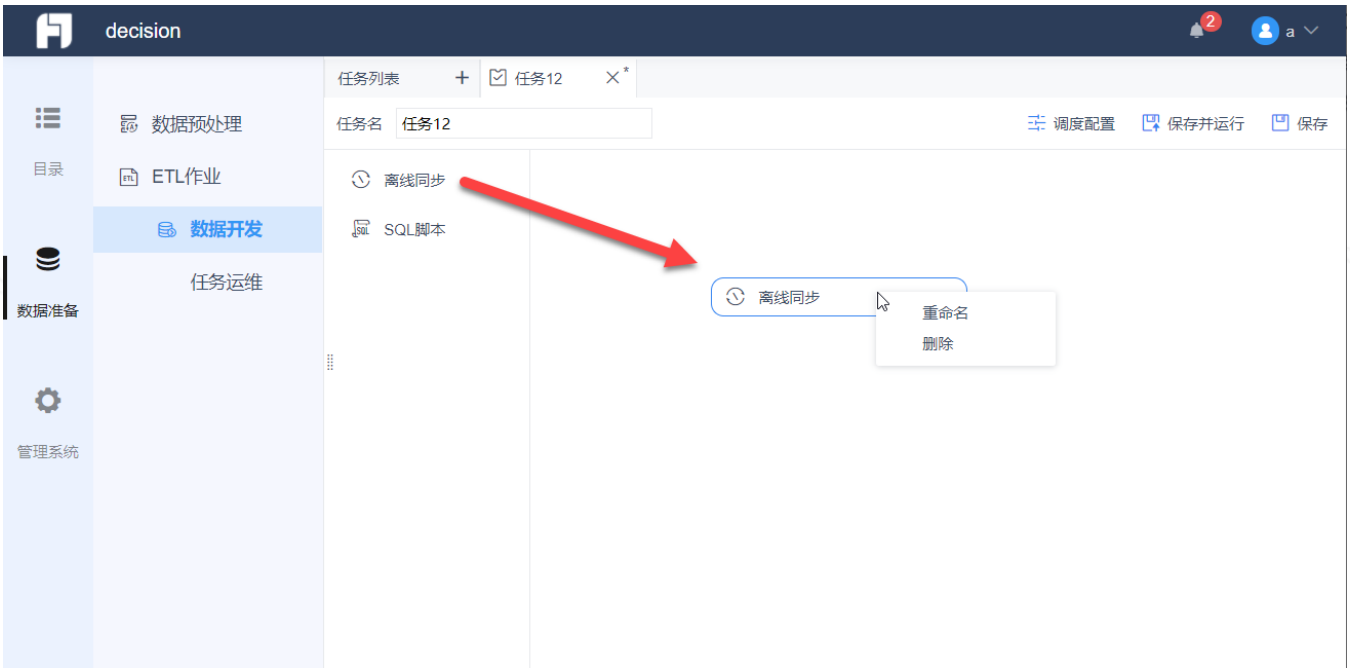
3.2 数据开发

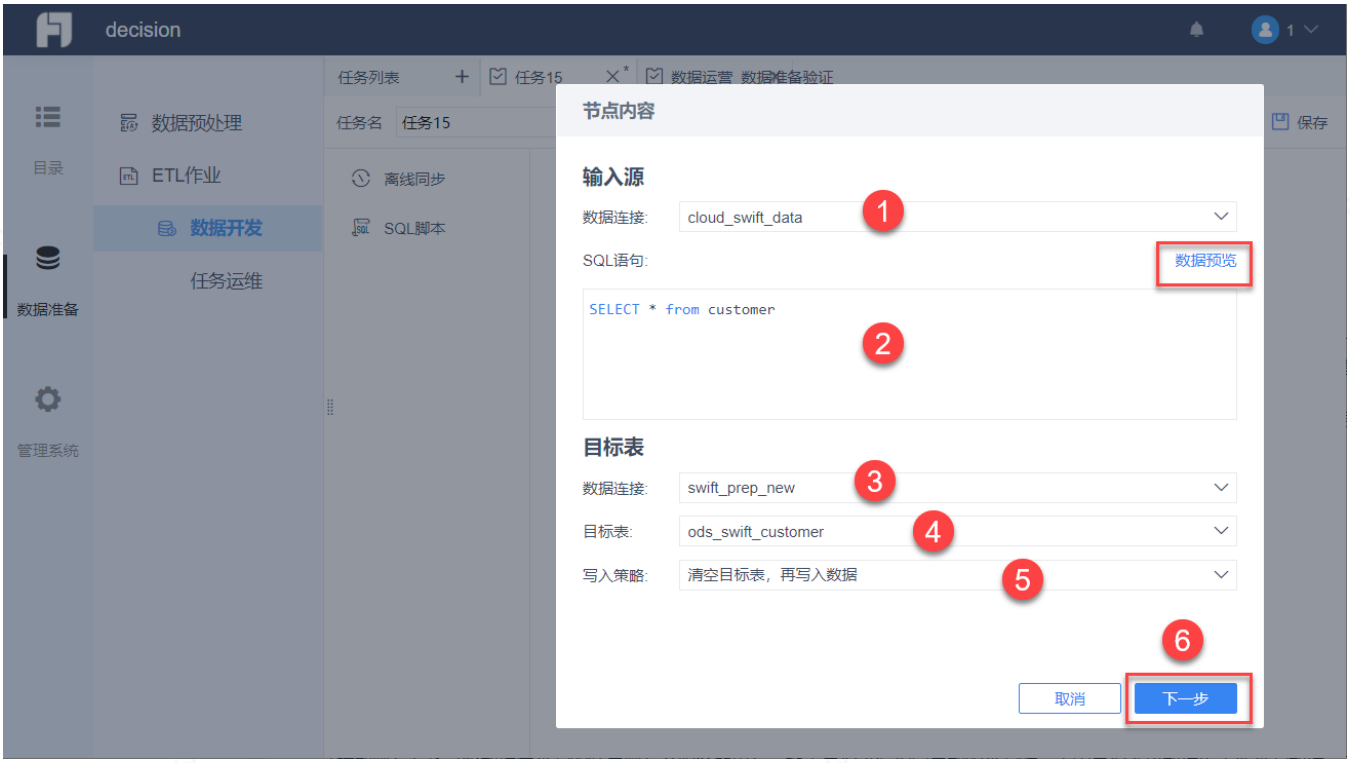
3.2.1 任务列表

1. 「数据准备」插件安装后，会在数据决策系统内增加「数据准备」目录，当管理员进入数据决策系统，点击「数据准备>ETL作业>数据开发」，界面如下图所示；
2. 点击任务列表旁①加号或页面中心②加号，会新建一个数据开发任务；
3. 任务列表中，点击任务名旁③的符号，再点击删除按钮，即可删除已有任务。

3.2.2 离线同步

1. 进入任务编辑页面后，页面左侧会显示2个组件类型供数据处理人员选择，离线同步、SQL脚本，拖动组件至页面中心区域，即可开始编辑；
2. 当鼠标焦点在组件上方，右键组件，可以选择对组件进行重命名或删除操作，显示如下图1。



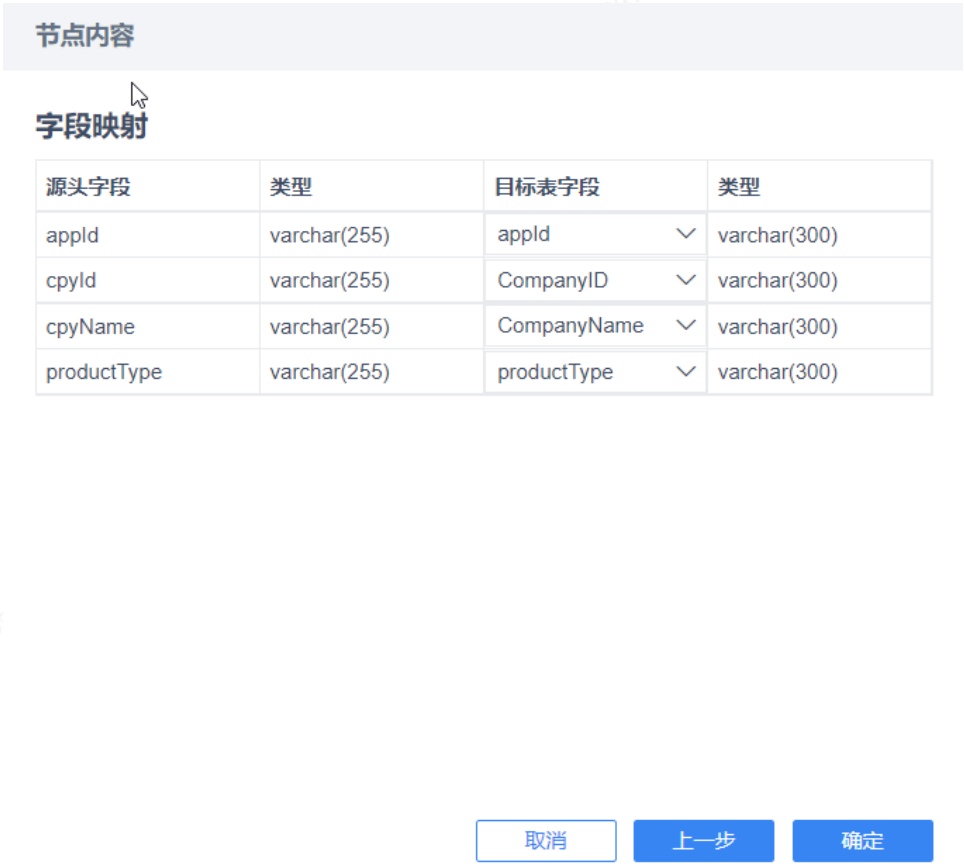


1. 点击组件，进入内容编辑区域
2. 输入源
   a. 数据连接：点击上图①处下拉列表，用户可选择来源库的数据连接名，**下拉列表仅显示当前版本插件支持的数据库类型对应的数据连接名称，当前版本插件支持MySQL、SQLserver、Oracle、PostgreSQL，4种数据库。**
   b. SQL语句：SQL语句输入内容完成后，点击数据预览按钮，可以查看当前语句对应的数据表内容，预览数据行数不超过5行，与实际数据可能会存在差异。
3. 目标表
   a. 数据连接：点击上图②处下拉列表，用户可选择目标库的数据连接名；
   b. 目标表：点击上图④处下拉列表，选择目标表的数据表名；
   c. 写入策略：

策略	含义
清空目标表，再写入数据	执行数据同步任务时，先将目标表中的数据清空，然后再将数据写入目标表中。

遇主键冲突，忽略输入源的相同主键数据	遇到写入的数据和目标表中的数据主键冲突时，忽略主键相同的数据行，插入主键不相同的数据。
遇主键冲突，更新目标表的相同主键数据	遇到写入的数据和目标表中的数据主键冲突时，将写入的数据覆盖目标库中主键相同的数据。
遇主键冲突，停止写入并报错	遇到写入的数据和目标表中的数据主键冲突时，报错并停止任务执行。

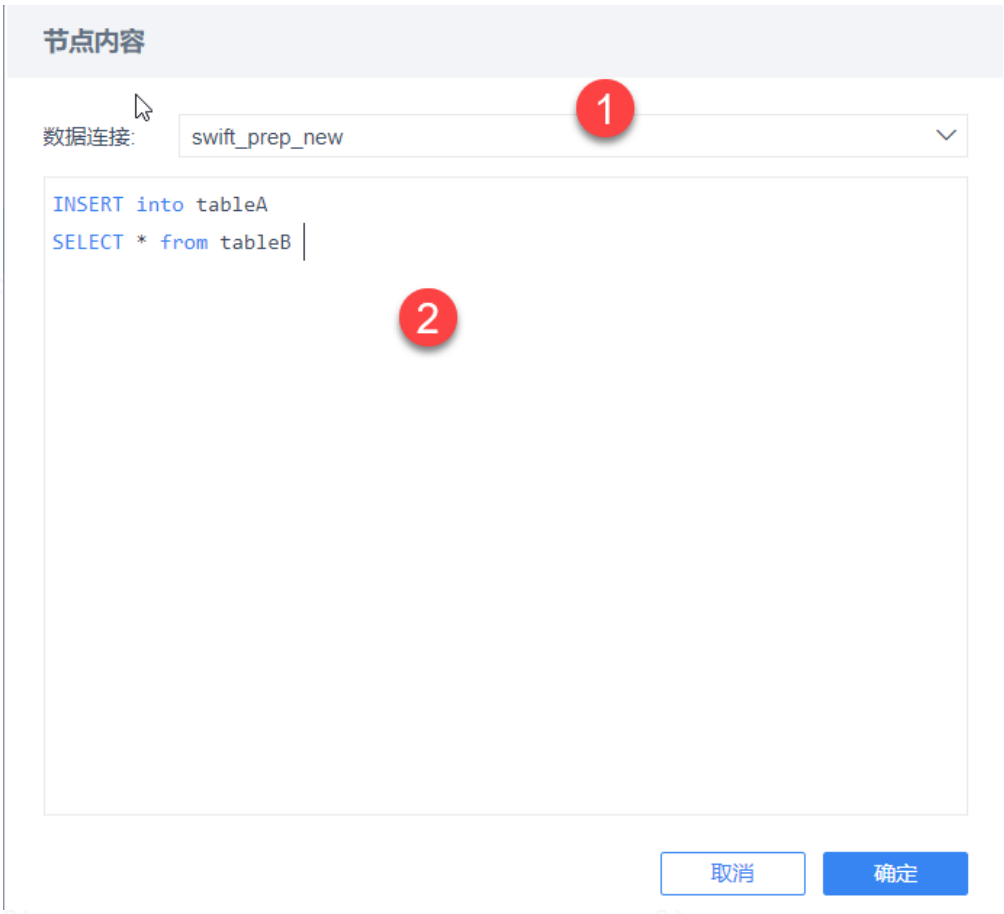
4. 字段对应关系：输入源和目标表确定后，点击上图⑥的下一步按钮，根据来源表的字段和类型，分别设置正确的目标表字段，设置完成后，点击确定按钮。



a.

3.2.3 SQL脚本

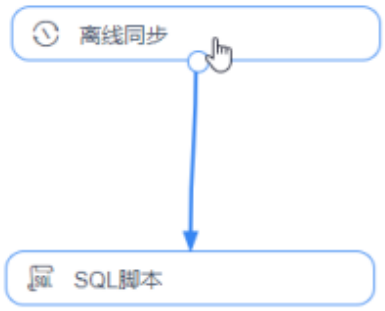
1. 数据连接：点击下图1处数据连接，选择需要处理的数据表对应的数据连接，下拉列表仅显示当前版本插件支持的数据库类型对应的数据连接名称；
2. SQL语句：目前支持insert into select、delete、create table等各类sql主流语法。



3.2.4 任务运行

1）组件依赖绑定

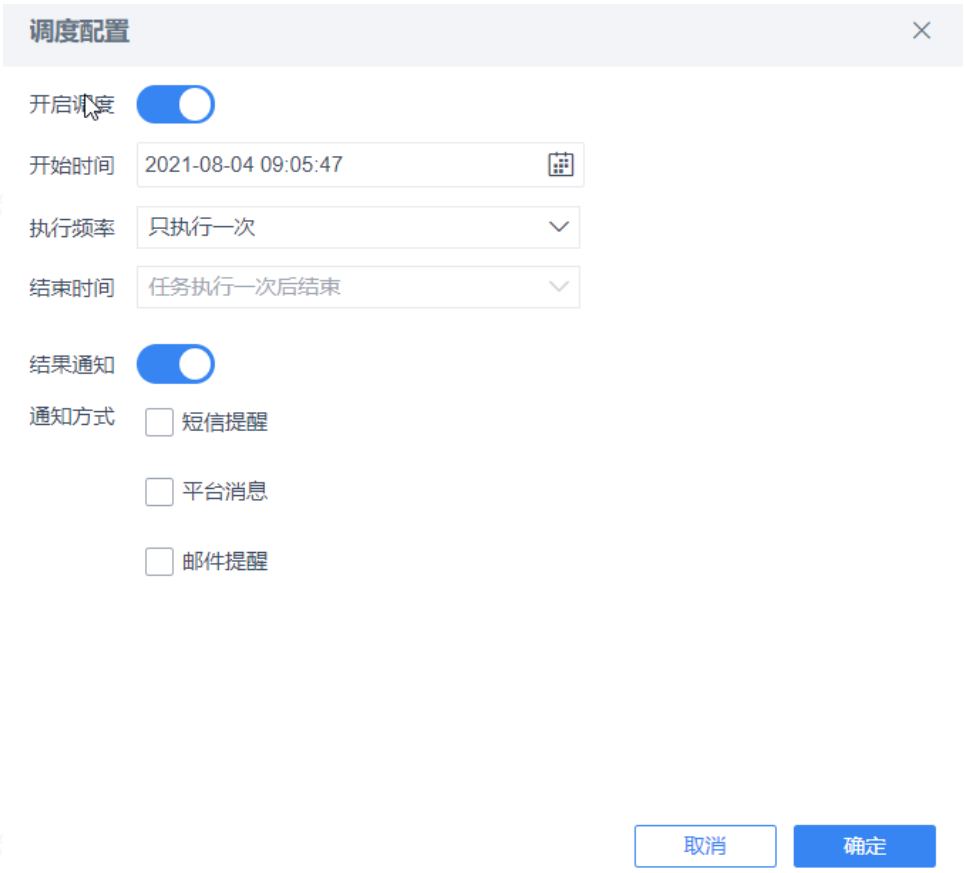
1. 离线同步、SQL脚本组件内容编辑完成后，可以通过绑定连接线的方式设置组件的执行先后顺序，先执行的组件指向后执行的组件，例如：点击A组件正下方区域，可连接至B组件正下方区域，显示效果如下图：



a.

2）调度配置

1. 设计完成的开发任务，用户可以为任务配置调度周期，点击任务编辑区域右上角【调度配置】按钮
    a. 开始时间和结束时间：指任务开始和结束的时间，并不是指更新的时长。开始和结束的时间粒度可细化到时分秒。点击  可选择开始时间。
    b. 执行频率：「只执行一次」、「简单重复执行」、「明细频率设置」、「表达式设定」四种执行频率方式。
    c. 结果通知：可选择通知方式。



d.

2. 「执行频率」以及对应的「结束时间」设置如下所示：

a.

执行频率	说明	对应结束时间
只执行一次	在设置的开始时间开始后，该任务只执行一次更新	无需设置结束时间
简单重复执行	可设置四种时间粒度，分别是：「分钟、小时、天、周」，可进行简单的间隔时间设置。例如：每隔1天执行一次	执行频率选择「简单重复执行」后，结束时间有三种选项： • 无限期：无结束时间，定时更新任务会一直定时执行 • 设定结束时间：和开始时间设定方法一致，当设定的结束时间到了后，将不再执行该定时更新的任务 • 额外重复执行次数：在设置的开始时间执行了定时任务之后会在间隔时间达到后再执行任务的次数，可手动设置次数
明细频率设置	可进行细化的间隔时间设置。 • 执行时间：细化到分钟，小时需要输入 0-23 内的整数；分钟需要输入 0-59 内的整数 • 执行日：可选每天、每周和每月。其中每周可在周一至周日之间单选或多选，每月可在 1 号到 31 号之间单选或多选 • 执行月：可在一年中的 1 月到 12 月单选或多选	执行频率选择「明细频率设置」后，任务结束时间只能选择两种：「无限期」或「设定结束时间」： • 无限期：无结束时间，定时更新任务会一直定时执行 • 设定结束时间：和开始时间设定方法一致，当设定的结束时间到了后，将不再执行该定时更新的任务
表达式设定	指设定一些「特定时间」的更新频率。 • 系统中的预置表达式包含五种 • 自定义表达式设置 填入时间频率的 Cron 表达式，需填写正确的表达式才能进行定时更新	执行频率选择「自定义表达式」后，任务结束时间只能选择两种：「无限期」或「设定结束时间」，可参考本章「明细频率设置」 注：Cron表达式的详细使用方法可参考定时更新-https://help.fanruan.com/finebi/doc-view-89.html 5. Cron表达式

3）任务运行

- 配置好调度周期的任务，可以通过点击页面左上角的【保存】按钮进行保存，任务保存成功后，**会在工程目录：\webapps\webroot\WEB-INF\dpworks 下生成对应文件。**
- 同时我们可以通过点击页面右上角的【保存并运行】按钮对任务进行运行测试，查看任务是否可跑通。

3.2.5 运行日志

- 运行中的任务，会在数据开发任务的页面下方输出运行日志，用于辅助用户判断任务是否可运行，运行失败的具体原因是什么，用户便可以针对性地调整任务设置或数据库表，日志显示效果如下：

3.3 任务运维

3.3.1 任务概览

任务概览展示工程内所有任务的整体情况，包括：

- 任务总数：所有任务的数量总和。
- 处于调度周期内任务数：筛选汇总有调度配置，且调度结束时间大于当前时间的任务数。
- 最近一次运行成功任务数：筛选汇总最近一次运行成功的任务数。
- 最近一次运行失败任务数：筛选汇总最近一次运行失败的任务数。

3.3.2 任务管理

- 任务管理会展示工程内每个任务的最近一次运行结果、调度配置和调度结束时间，点击运行记录列的查看详情按钮，即可查看对应任务的历史运行日志；
- 点击任务编辑按钮，即可跳转至任务的设计界面；
- 点击任务删除按钮，会进行弹窗确认，确认删除后会执行删除动作。

4. 注意事项

1. 若从A工程的任务文件夹dpworks，直接拷贝任务文件至B工程，因为目前没有进行针对性适配，所以B工程内的任务会存在未知的问题，因此不建议进行工程间的任务拷贝，预计后续会在产品功能上支持此诉求；
2. 暂时不支持在任务文件夹dpworks，直接删除某任务文件，若执行此操作，数据准备工具前端会出现未知的报错问题；
3. finedb中fine_dp_conf_entity表记录任务配置信息，fine_dp_execute_record表记录任务执行信息，其他数据准备工具相关的表包括：fine_dp_conf_entity_value、fine_dp_data_slice、fine_dp_dateset，注意不要操作以上表的数据或者误删表，避免影响数据准备工具的使用。