使用手册

SharkBuilder-V1.1.

北 京 先 进 数 通 信 息 技 术 股 份 公 司

编写说明

中文标题：**SharkBuilder-**使用手册

英文标题：

过程分类：

过 程 域：

文档类型：

版本历史：

版本	编写	日期	修改说明	审核	批准
v1.0		2019-01-13	初版		
V1.1		2019-08-13	修改		
V1.2					
V1.3					

目 录

编写说明 II
1 前言 1
1.1 目的 1
1.2 适用版本 1
1.3 名称术语 1
1.3.1 数据集成开发平台 1
1.3.2 数据对象 1
2 平台介绍 1
2.1 平台简介 1
2.2 功能总览 2
2.3 角色说明 2
2.4 业务流程 3
3 功能介绍 4
3.1 系统管理 4
3.1.1 用户管理 4
3.1.2 角色管理 5
3.1.3 机构管理 5
3.1.4 机构组管理 7
3.1.5 角色权限 7
3.1.6 机构组分配 8
3.2 项目管理 9
3.2.1 项目管理 9
3.2.2 项目版本管理 12
3.2.3 项目作业配置 14
3.2.4 开发任务管理 15
3.2.5 项目导出 18
3.2.5.1 发版包修改 20
3.2.6 项目导入 22
3.2.7 项目元数据分配 23
3.3 调度管理 24
3.3.1 调度配置 24
3.3.2 调度分配 25
3.3.3 调度日历 25
3.4 开发管理 26
3.4.1 系统信息维护 26
3.4.2 数据源 26
3.4.2.1 数据库类型 27
3.4.2.2 文件类型 29
3.4.3 公共模板 31
3.4.4 公共组件 34
3.4.5 实体表配置 36
3.5 元数据信息 36

3.5.1 数据字典.....36
3.5.2 元数据分析.....38
3.6 审批管理.....39
3.7 作业地图.....39
3.8 数据接入.....40
3.8.1 数据接入.....40
3.8.1.1 数据采集.....40
3.8.1.2 数据预处理.....41
3.8.1.2.1 字符集转码.....42
3.8.1.2.2 字段拆分.....42
3.8.1.2.3 字段筛选.....43
3.8.1.2.4 分解.....44
3.8.1.2.5 清洗检核.....44
3.8.1.2.6 码值转换.....45
3.8.1.3 数据装载.....46
3.8.1.4 批量导入.....46
3.8.2 数据归档.....46
3.8.3 转换规则.....47
3.9 数据加工.....47
3.9.1 配置开发.....48
3.10 数据供应.....51
3.10.1 文件交换.....51
3.10.2 数据供应.....52
3.11 数据质量检核.....53
3.11.1 检核规则.....53
3.11.2 检核管理.....55
3.11.3 任务监控.....56
3.11.4 内置检核规则说明.....57
3.11.4.1 当日数据不能为空.....57
3.11.4.2 字段空值检查.....57
3.11.4.3 数值类型的取值范围.....57
3.12 数据生命周期管理.....57
3.12.1 生命周期规则配置.....58
3.12.2 生命周期清理策略.....59
3.12.3 生命周期管理.....61
3.12.4 生命周期任务.....61
3.13 调研信息管理.....62
3.14 模型管理.....62
3.14.1 模型设计结果导入.....62
3.14.2 物理模型管理.....62
4 教程示例.....64
4.1 创建数据项目.....64
4.1.1 创建项目.....64
4.1.2 指定项目开发人员.....64

4.1.3 配置项目规范.....66
4.1.4 配置近源数据区.....69
4.1.5 配置历史归档数据区.....72
4.1.6 项目元数据分配.....73
4.2 数据接入.....74
4.2.1 配置调度规范.....74
4.2.1.1 调度配置.....74
4.2.1.1.1 新增计划.....75
4.2.1.1.2 新增任务.....75
4.2.1.1.3 新增作业.....76
4.2.1.2 调度分配.....76
4.2.2 新增 SCHEMA.....76
4.2.2.1 源系统配置.....79
4.2.2.2 数据源配置.....80
4.2.2.3 Schema 配置.....80
4.2.3 导入数据字典.....81
4.2.4 任务分配示例.....82
4.2.5 数据接入示例.....84
4.2.5.1 数据接入-选择作业类型.....84
4.2.5.2 数据接入-数据采集.....85
4.2.5.3 数据接入-预处理流程.....86
4.2.5.4 数据接入-数据装载.....87
4.2.6 数据归档.....88
4.2.7 批量功能.....89
4.2.7.1 数据接入.....89
4.2.7.2 数据归档.....90
4.3 数据生命周期管理.....90
4.3.1 生命周期调度配置.....90
4.3.1.1 新增计划.....91
4.3.1.2 新增任务.....91
4.3.1.3 新增作业.....92
4.3.2 调度日历配置.....92
4.3.3 生命周期规则配置.....94
4.3.4 生命周期清理策略配置.....95
4.3.5 SCHEMA 的生命周期管理.....96
4.3.6 表的生命周期管理.....97
4.3.6.1 生命周期开发任务分配.....97
4.3.6.2 生命周期任务配置.....99
4.3.7 发布作业.....99
5 参考数据.....100
5.1 宏定义说明.....100
5.2 schema 规范配置指导.....101
5.2.1 业务源系统.....101
5.2.1.1 数据库类型.....101

5.2.1.2 文件 101
5.2.2 数仓数据源.....102
5.2.2.1 RECV 区 102
5.2.2.2 MIDD 区 103
5.2.2.3 SEND 区 103
5.2.2.4 RDS 区.....104
5.2.2.5 SDS 区 104
5.2.2.6 IDS 区 104
5.2.2.7 FDS 区 104
5.2.2.8 ADS 区.....104
5.2.2.9 PDS 区 104
5.3 命名约束 104
6 其它参考 105
6.1 公共组件说明 105

1 前言

1.1 目的

本文档旨在为数据集成开发平台的使用用户提供一个全面的使用说明。通过此文档，用户可了解数据集成开发平台的功能和操作流程。

1.2 适用版本

DIDP_V1.*.*

1.3 名称术语

1.3.1 数据集成开发平台

本开发平台名称，Data Integration Development Platform（DIDP）。

1.3.2 数据对象

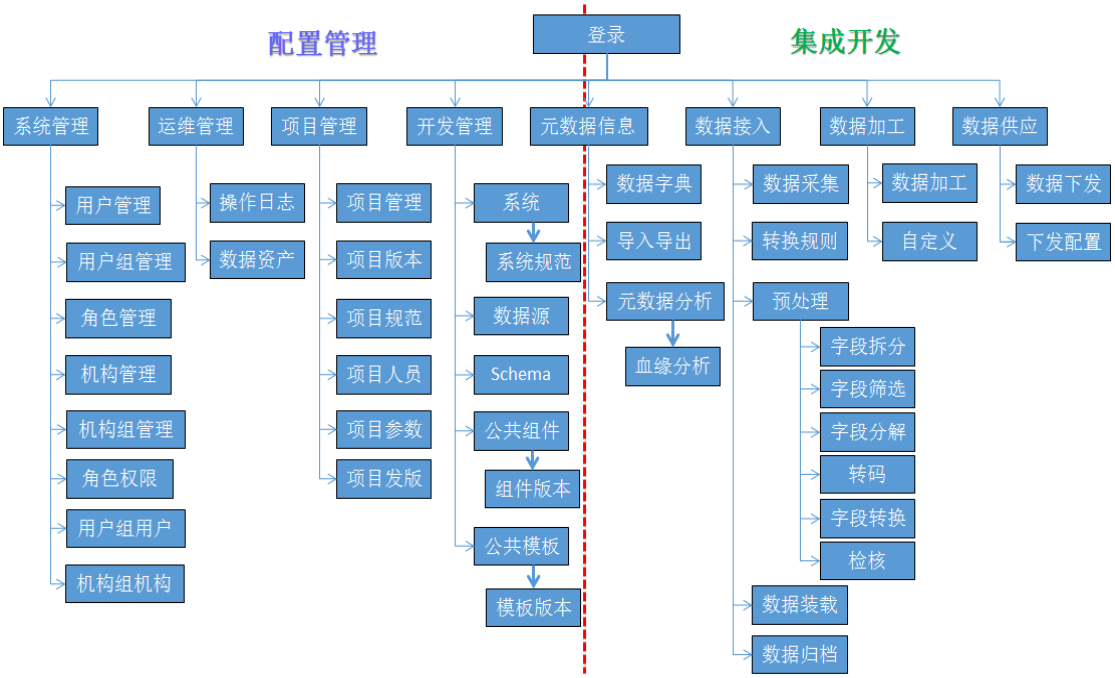
DIDP 内的数据对象的表示有四层，依次为“系统—数据源—SCHEMA—表名”，在数据字典上的以树的形式展示，如下图：

2 平台介绍

2.1 平台简介

N/A

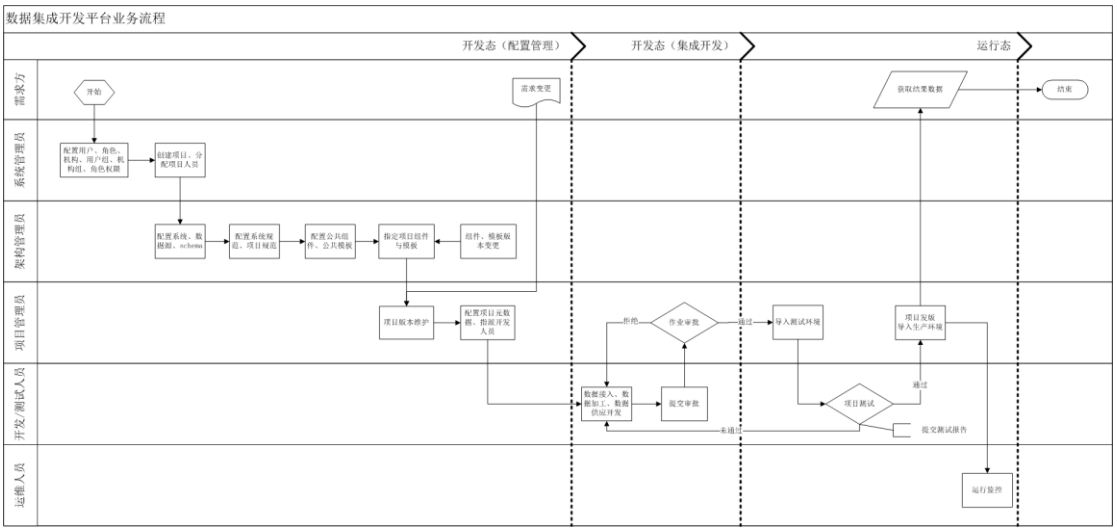
2.2 功能总览



2.3 角色说明

图表 2.3-平台角色

2.4 业务流程



图表 3-1 平台业务流程

1）需求方提出项目需求，启动项目

2）系统管理员配置用户、角色、机构、权限

3）系统管理员创建项目、分配项目人员

4）架构管理员配置系统、数据源、schema

5）架构管理员配置系统规范、项目规范

6）架构管理员配置公共组件、配置公共模板

7）架构管理员指定项目组件与模板

8）项目管理员维护项目版本

9）项目管理员配置项目元数据、指派开发人员

10）开发人员开发接入、加工、供应作业，并提交审批

11）项目管理员审批作业。审批通过则导出项目，并导入到测试环境；审批不通过，则打回给开发人员重新开发，流程跳转到“10）”

12）测试人员测试项目。测试不通过，则开发人员进行修改，流程跳转到“10）”

13）测试通过，项目管理员进行项目发版，并导入到生产管理，部署运行。

14）需求方获取结果数据

15）如果需求方进行需求变更，则流程跳转到“8）”

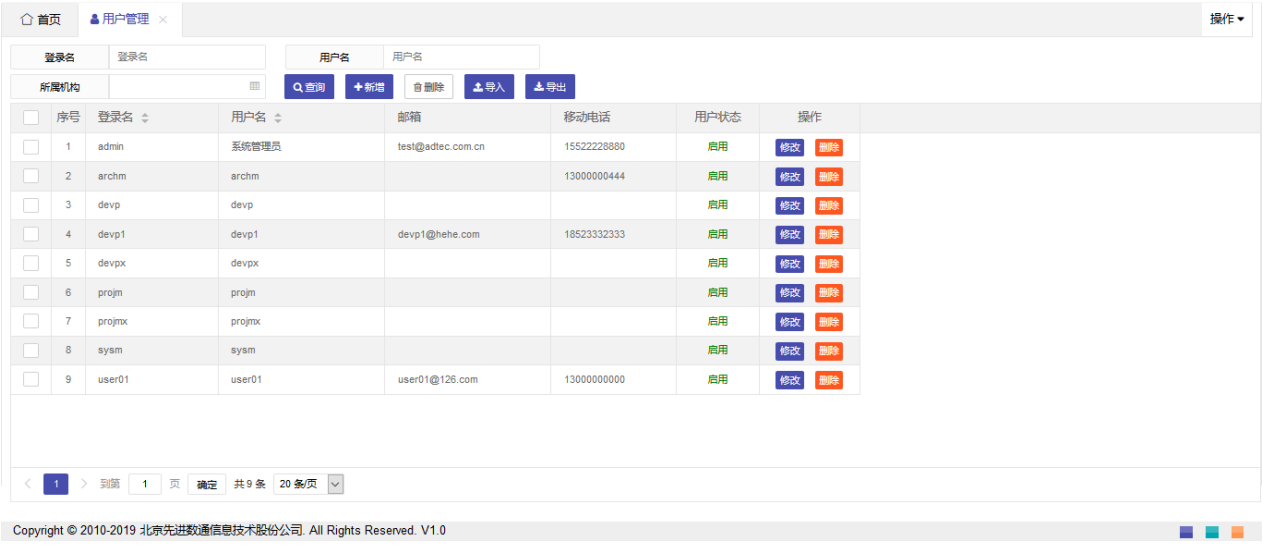
如果架构管理员进行组件、模板变更。则流程跳转到“7）”

3 功能介绍

3.1 系统管理

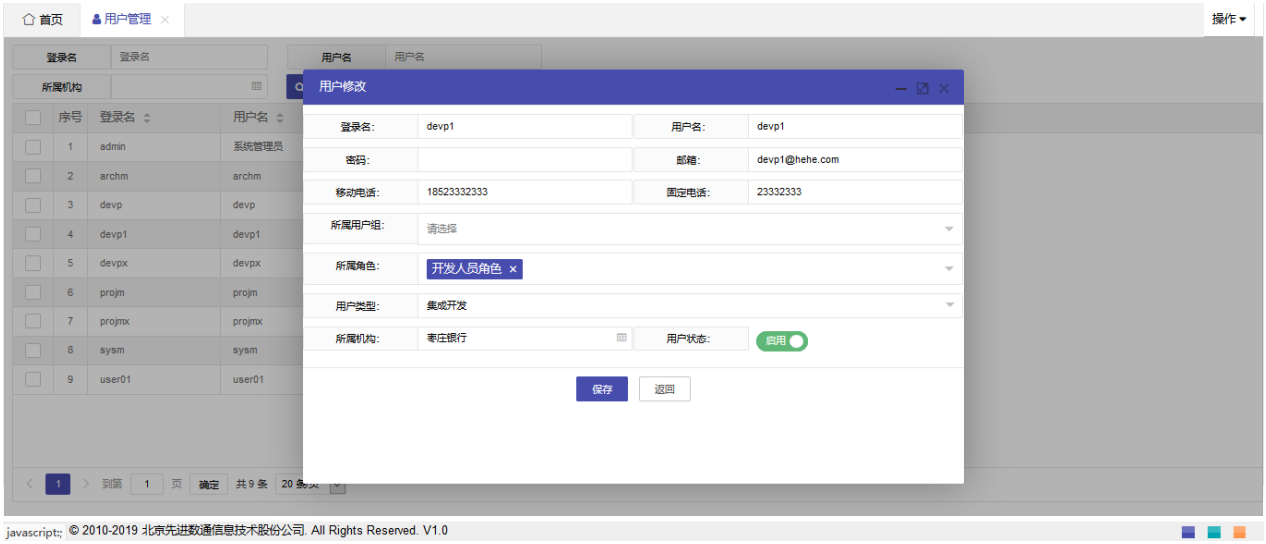
3.1.1 用户管理

用户管理提供对平台用户的配置及管理功能。



■ 用户配置项说明

- 登录名：登录用户名，只能有数字、字母、下划线组成；
- 所属用户组：来源于系统中已配置的用户组；
- 所属角色：来源于系统中已配置的角色；
- 用户类型：分为配置管理和集成开发两种，两者的区别在于用户登录是否需要选择项目进入（注：项目管理员为集成开发类型）；
- 所属机构：用户所属机构。



3.1.2 角色管理

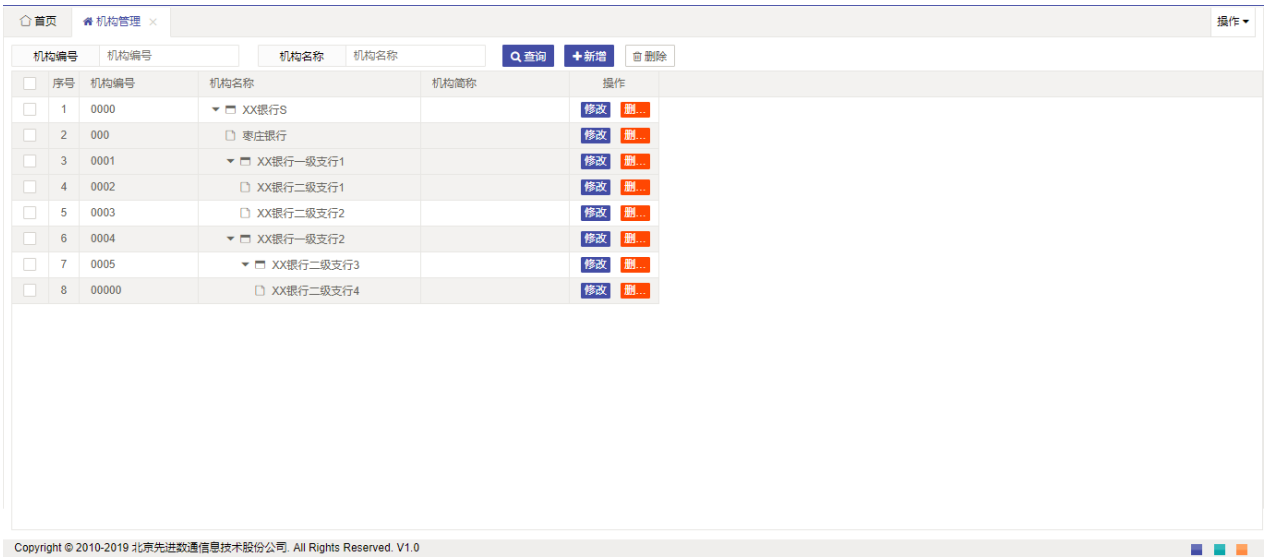
角色管理提供对平台角色的新增、修改、删除功能。平台内置系统管理员、架构管理员、项目管理员、开发人员四类角色，满足一般数据项目使用，若需要个性化角色，可自定义新增。

■ 角色配置说明

- 角色标识：角色全局唯一的标识名，建议命名规范“ROLE_角色名”；
- 角色名称：角色的中文名称；
- 角色描述：描述信息。

3.1.3 机构管理

机构管理提供对企业机构信息维护的功能，以树形结构展示机构的层级关系。



■ 机构配置说明

- 机构编号：企业法人/机构号；
- 机构名称：机构的描述名称；
- 机构简称：机构简称；
- 上级机构：若新增的机构归属于某个机构节点则从选择父机构节点，若新增的机构无上级机构则不填。

3.1.4 机构组管理

机构组管理提供对机构组标识的配置及管理功能，以列表的形式展示。

■ 机构组配置说明

- 机构组标识：机构组标识；
- 机构组名称：机构组名称；
- 机构组描述：机构组描述。

3.1.5 角色权限

角色权限提供角色的权限管理功能，包括可拥有该角色的用户/用户组权限分配、角色可访问资源（功能菜单）的配置。

■ 角色用户分配

从未分配用户中选择用户，点击“>”，则移动到已分配用户列表中，完成角色用户的分配。

- 角色用户组配置
- 角色功能菜单配置

此处会列出平台所有的菜单，勾选菜单，点击“保存”，即可完成菜单的配置功能。

3.1.6 机构组分配

机构组分配提供机构组的分配管理功能。

3.2 项目管理

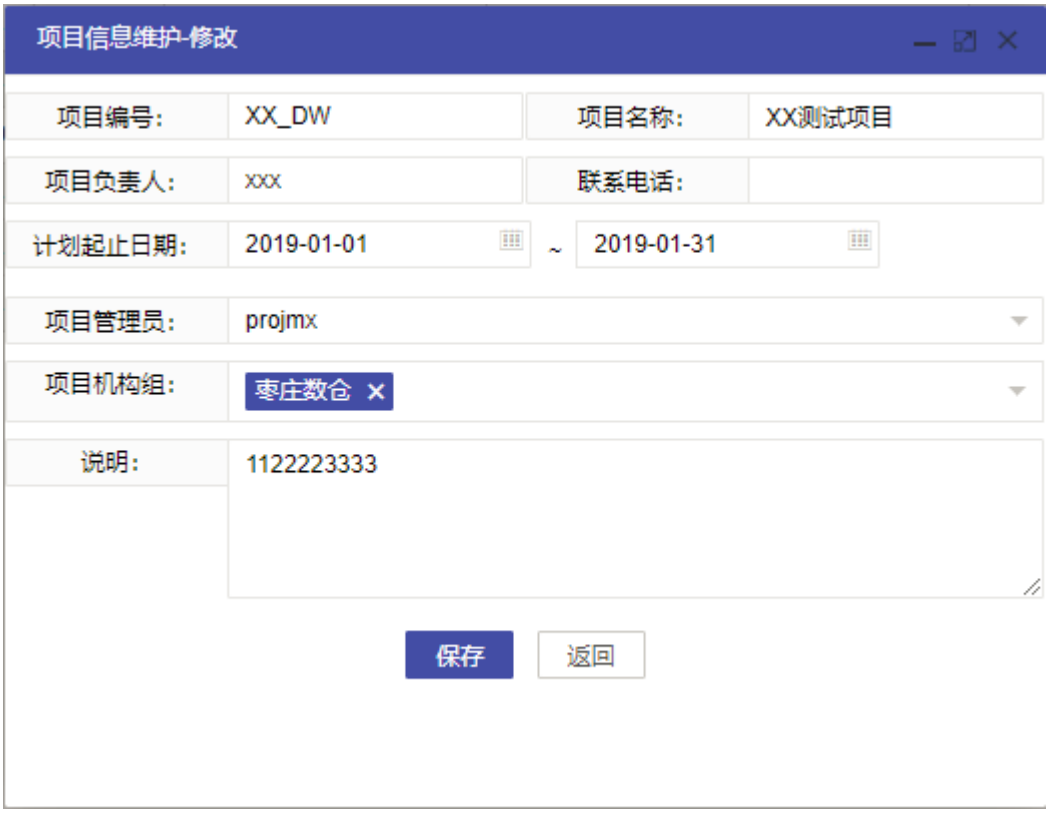
3.2.1 项目管理

项目管理提供对项目基本信息、项目可操作用户、项目规范配置以及项目复制功能。

■ 项目基本信息说明

- 项目起止日期：仅作为项目参考信息，平台当前版本在项目到期后不会将项目失效；
- 项目管理员：该项目的管理员（项目经理），从集成开发用户类型的用户中选择，该用户会用于后期开发任务的任务分配，项目版本的参数维护等功能；
- 项目机构组：从已配置的机构组中选择，用于指定项目可操作的机构范围，例如：当计划命名规范上存在#ORGCODE#宏，会根据选择的机构组，分别

生成相应机构的机构。

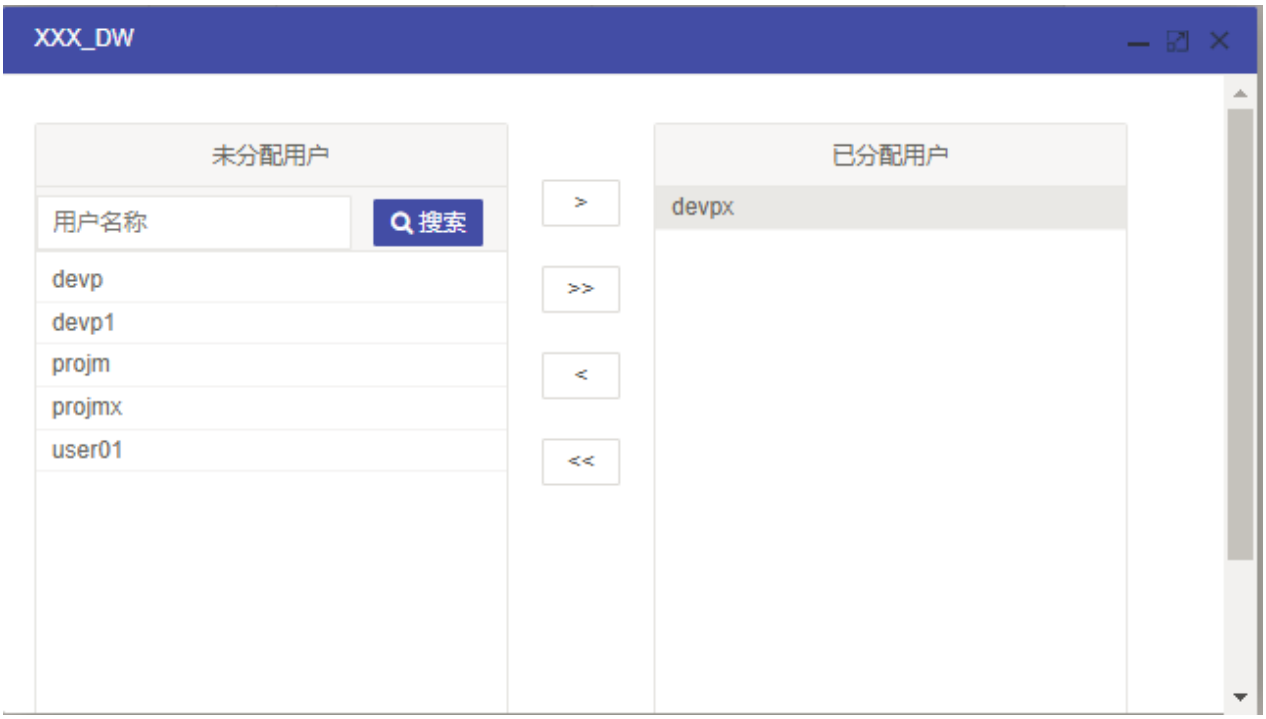


■ 项目用户分配

分配可操作该项目的用户，只有为用户分配该项目权限，用户在登录时才能选择该项目。

从未分配用户中选择用户，点击“>”，则移动到已分配用户列表中。

注：同一个用户可以分配于不同的项目下。



■ 项目配置说明：

- 字符集编码：项目内使用的字符集，支持 UTF-8 和 GBK 两种选择，对于接入的数据，项目内统一转换成该字符集的文件；
- 分隔符：项目内使用的分隔符，对于接入的数据，项目内统一转换成此处定义的分隔符；
- 换行符：项目内使用的换行符；
- 采集目标：数据采集的目标存储 schema，从已配置的 schema 中选择，此处仅提供文件类型的 schema 供选择；
- 采集目标：数据预处理的目标存储 schema，从已配置的 schema 中选择，此处仅提供文件类型的 schema 供选择。

■ 项目版本复制

提供对已发版项目版本的复制功能，复制的目标版本号自定义，建议使用数字版本号。

3.2.2 项目版本管理

项目版本管理提供对项目版本的管理功能，该功能必须使用项目管理员进行操作。

■ 项目版本复制

提供对已发版项目版本的复制功能，复制的目标版本号自定义，建议使用数字版本号。

■ 项目参数配置说明

- 调度服务 IP 地址：调度服务器的 IP 地址；
- 调度服务端口：调度服务器的端口；
- 调度服务用户名；调度服务器用户名；
- 调度服务密码：调度服务的密码；
- HDFS 认证用户：接入的装载模块需要上传数据文件到 HDFS 中，而系统内的 HDFS 采用的是 Kerberos 方式认证，此处填写的是的认证用户，默认配置（hive）；
- HDFS 中 namenode 的 IP 地址：接入的装载模块需要上传数据文件到 HDFS 中，需要连接 namenode 节点，参考格式：

http://36.33.96.208:50070;http://36.33.96.209:50070。

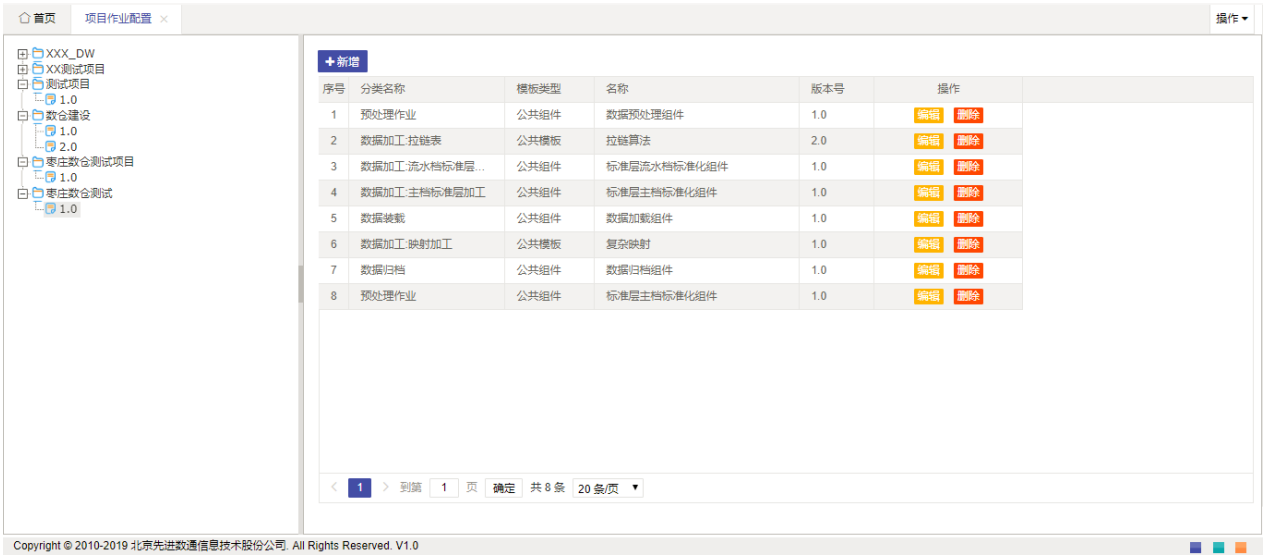
■ 发版

在项目所有的任务开发测试完成后，通过发版功能完成项目的定版上线。

注：发版后，该项目下所有资源皆不能修改。

3.2.3 项目作业配置

项目作业配置用于管理员为项目分配可用的资源，资源包括公共组件、公共模板。



■ 点击“新增”按钮，配置相应的组件，新增配置说明如下：

- 数据流程类别：选择相应的流程类别；
- 模板类型：选择公共组件或公共模板；
- 选择组件：从已配置的模板或组件中选择相应的版本。

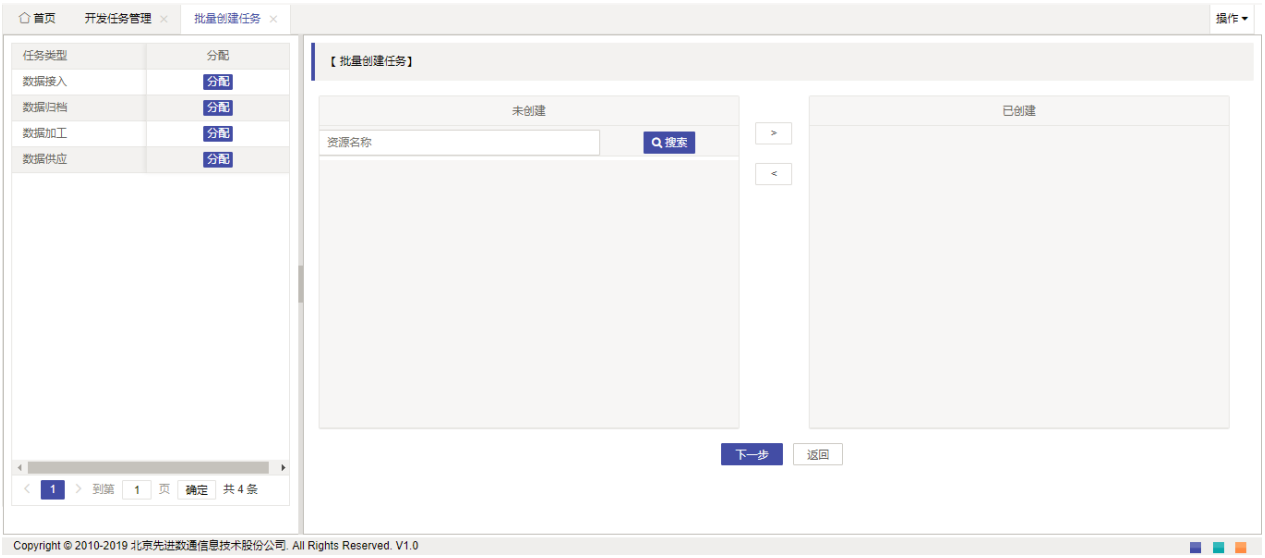
3.2.4 开发任务管理

开发任务管理用于项目管理员向开发人员进行工作任务分配，以元类型对象（SCHEMA、表）的形式进行任务分配，提供批量新增的方式进行配置，同时提供表名、任务类型、作业名、开发人员、状态多个维度进行任务的查询。

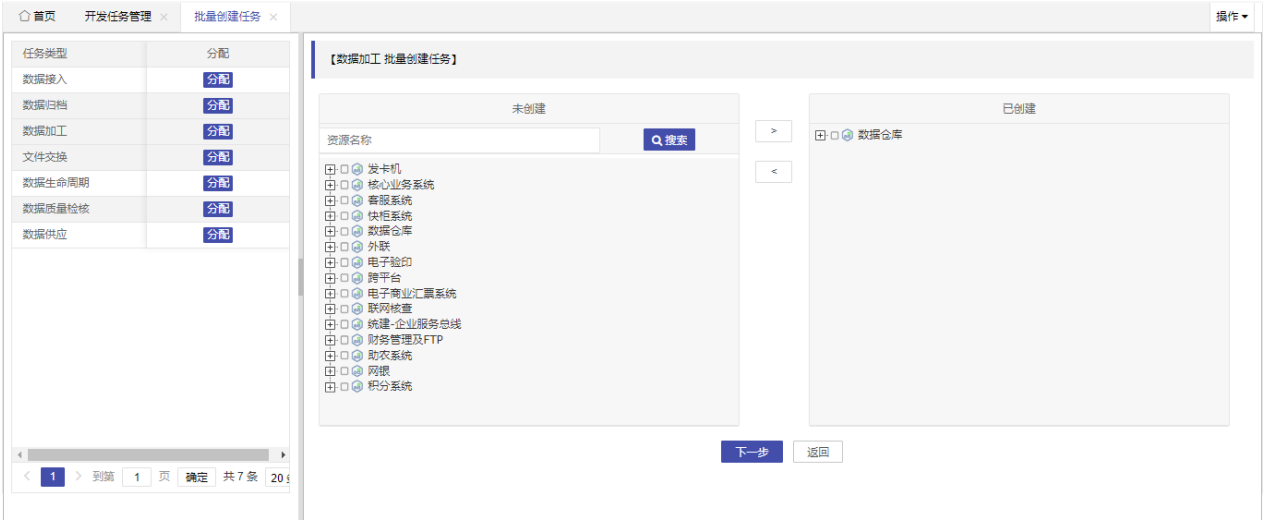
待开发的数据对象只有在任务分配后，才会在集成开发用户的开发界面中显示。

■ 任务分配流程：

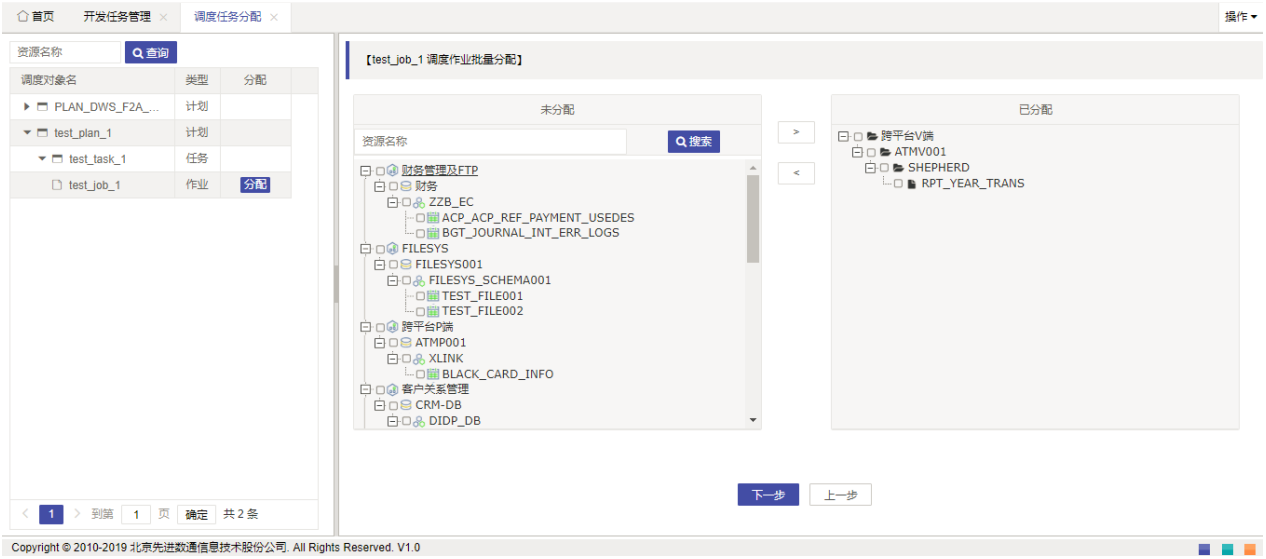
1. 点击“批量新增”，弹出任务分配页面；



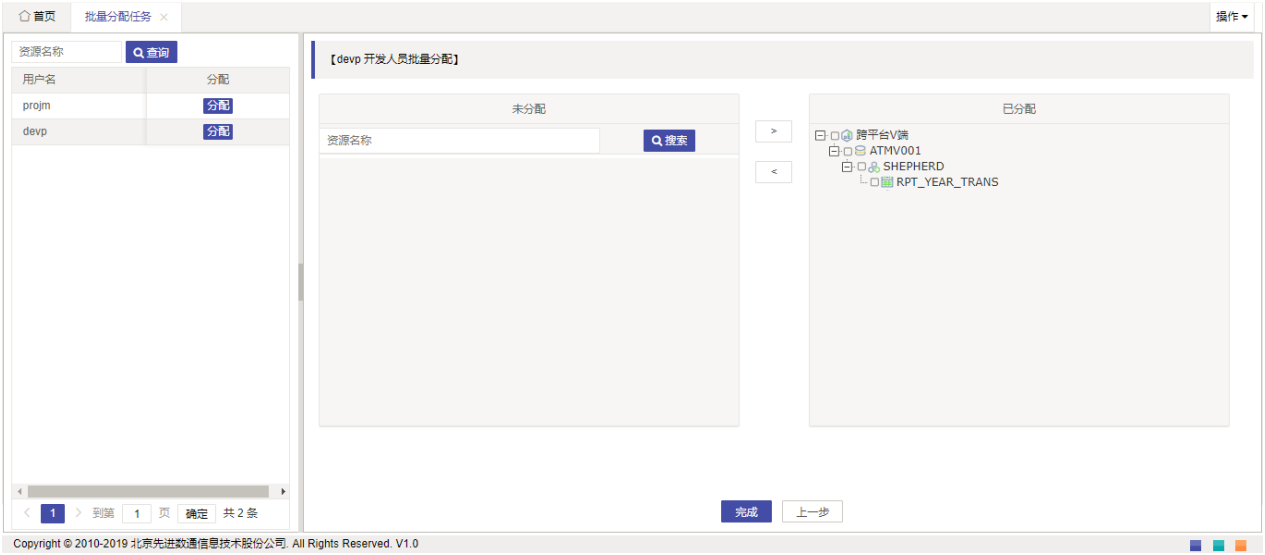
2. 选择需要分配的任务类型，点击“分配”，右侧列出选择任务分配的对象，左侧勾选，点击”>”号，移到右侧，标识选择完成；



3. 下一步，进入调度对象选择的功能页面，选择需要挂载的调度对象（作业名一级，注：需要预先在调度配置中配置，并且为该项目分配可用的调度配置），点击“分配”，在右边功能区选择相应的数据对象，点击>号，完成选择。



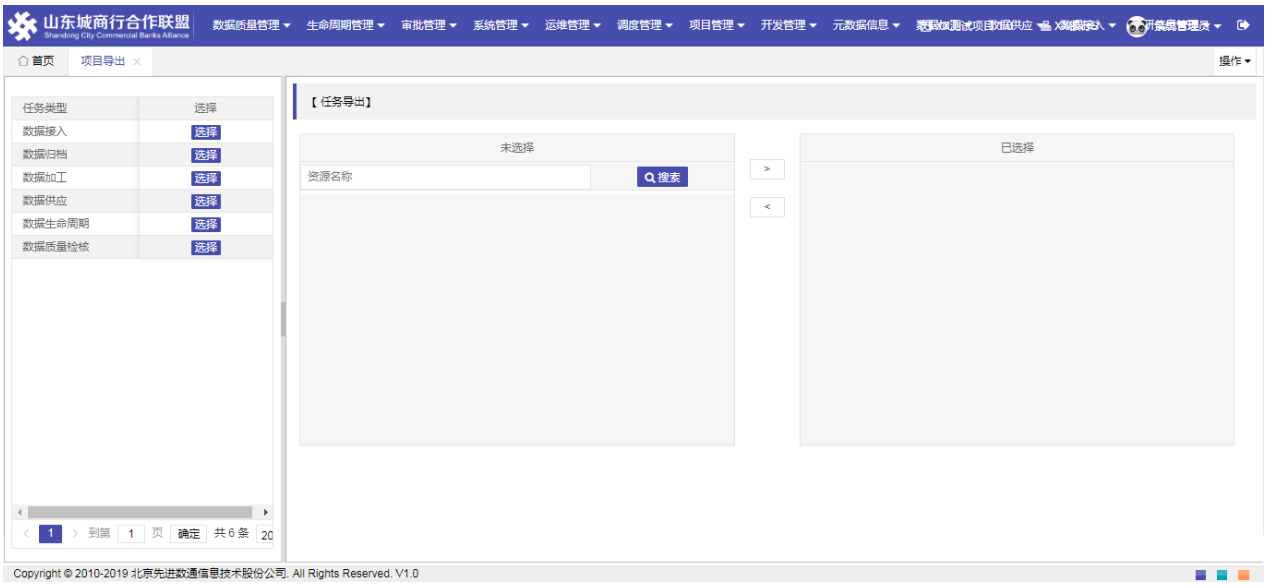
4. 下一步，分配开发用户，点击用户名后的“分配”，则完成为用户分配该数据对象的开发。



5. 点击“完成”，即保存任务分配配置，完成任务的分配功能。

3.2.5 项目导出

项目导出用于开发平台项目下已完成任务的数据导出，不同项目版本实现增量导出。提供单个任务类型和多个任务类型的导出。

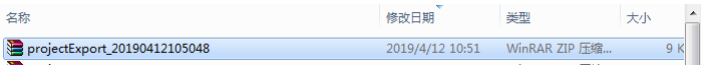


■ 项目导出流程：

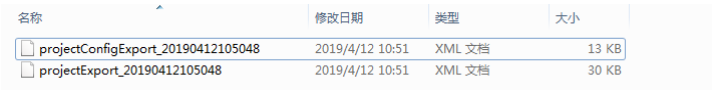
1. 点击要导出任务类型行的“选择”按钮，在任务导出区域的未选择区域的树资源下勾选已完成的表任务；

2. 点击穿梭框中间的右移按钮；

3. 点击页面的“导出”按钮，下载导出数据压缩文件。



压缩文件中包含两个文件，projectConfigExport 和 projectExport



projectExport 文件的内容为任务数据，不允许修改。

```
projectExport_20190412105048.xml
<?xml version="1.0" encoding="UTF-8"?>

<didp>
  <tasks>
    <task cqNo="002" projectVersionId="dacc0210171c4db7bca1546d84ed6090" description="" ctlObjId="f8ef0f168ee24197a3fcdcec958aacfe" taskType="
      <dataAccess dataAccessId="b466cd3774a747918fe96eaedf0b5c58" sourceType="1" tableId="1cb44779d75e4682a763894bcb290f0f" projectVersionId="
        <accesses>
          <collection dataAccessId="b466cd3774a747918fe96eaedf0b5c58" ctlFilePath="" ddlFilePath="" projectVersionId="dacc0210171c4db7bca1546c
            <process planType="INIT" processCate="001" processStatus="02" processId="4e6988cb6f244390a6307687f95ceee4" processName="" projectV
              <controlParams operLog="1" avbFlag="1" overTime="235959" projectVersionId="dacc0210171c4db7bca1546d84ed6090" weight="100" maxPic
              <processTableRels>
                <processTableRel tarTableId="86f25d0a0fb94155aaf99b374c752604" relId="f20b6db348ea4c57a7c872ad9313b36b" srcTableId="1cb44779d7
              </processTableRels>
            </process>
          </collection>
          <preProcess tarTableId="b2b34630b6b4456b86498d9537c86de5" transcodingFlag="0" processId="edbc75e59c0e40d1a4a4748c28e66e49" customize
            <preSelects>
              <preSelect colSeq="1" colName="BANK_ID" pkFlag="0" bucketFlag="0" description="" projectVersionId="dacc0210171c4db7bca1546d84ed6
              <preSelect colSeq="2" colName="BATCH_DT" pkFlag="0" bucketFlag="0" description="" projectVersionId="dacc0210171c4db7bca1546d84ec
              <preSelect colSeq="3" colName="COL_01" pkFlag="0" bucketFlag="0" description="" projectVersionId="dacc0210171c4db7bca1546d84ed60
              <preSelect colSeq="4" colName="COL_02" pkFlag="0" bucketFlag="0" description="" projectVersionId="dacc0210171c4db7bca1546d84ed60
              <preSelect colSeq="5" colName="COL_03" pkFlag="0" bucketFlag="0" description="" projectVersionId="dacc0210171c4db7bca1546d84ed60
            </preSelects>
            <process planType="INIT" processCate="002" processStatus="02" processId="edbc75e59c0e40d1a4a4748c28e66e49" processName="" projectV
              <controlParams operLog="1" avbFlag="1" overTime="240000" projectVersionId="dacc0210171c4db7bca1546d84ed6090" weight="100" maxPic
              <processTableRels>
                <processTableRel tarTableId="b2b34630b6b4456b86498d9537c86de5" relId="ee7010cd895049cf8447bb0a83c0789d" srcTableId="86f25d0a0f
              </processTableRels>
            </process>
          </preProcess>
          <load loadId="c551dde246874f5c9ae05dd42f31d4fa" processId="cafa367f995046a89ad09f628f0e3c4e" customizeFlag="0" schemaId="607f01688ab
            <process planType="INIT" processCate="003" processStatus="02" processId="cafa367f995046a89ad09f628f0e3c4e" processName="" projectV
              <controlParams operLog="1" avbFlag="1" overTime="240000" projectVersionId="dacc0210171c4db7bca1546d84ed6090" weight="100" maxPic
```


projectConfigExport 文件的内容为任务的配置数据，可以根据不同的环境修改 moia 及各种数据库的连接信息。在项目导入的时候会校验这些连接信息。


```
projectConfigExport_20190412105048.xml
<?xml version="1.0" encoding="UTF-8"?>

<didp>
  <systems>
    <system systemId="105b1e625e8e423f8a200ee6fbc56dbf" systemName="XXX_DW" systemType="2" systemShortKey="XXX_DW" lastUpdateUser="cde9b785ff2
      <dataSources>
        <dataSource sourceId="ab0007d1c26942dfb8e7eced64bda31d" sourceKey="XXX_DW_SA" systemId="105b1e625e8e423f8a200ee6fbc56dbf" port="0" dat
          <dataSourceFile separateChar="~@~" lineBreak="\n" configId="9133117f4a584b08a910fae321203ad2" serverType="1" userPwd="" lastUpdateUs
          <schemas>
            <schema sourceId="ab0007d1c26942dfb8e7eced64bda31d" schemaShortKey="RECV" schemaId="f5c817396cee41989ae6e31575c02d7a" domainName="
              <objStandard dataFileType="" zipFileName="" dataObjName="#SCHEMASHORT#.#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE
            </schema>
            <schema sourceId="ab0007d1c26942dfb8e7eced64bda31d" schemaShortKey="MIDD" schemaId="c8f02fa37eaf451d8e75213842171a1d" domainName="
              <objStandard dataFileType="" zipFileName="" dataObjName="#SCHEMASHORT#.#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE
            </schema>
          </schemas>
        </dataSource>
      </dataSources>
    </system>
    <system systemId="5bafb54b77064064bafc021282506b36" systemName="枣庄数仓" systemType="2" systemShortKey="ZZDW" lastUpdateUser="489d4f6f57
      <dataSources>
        <dataSource sourceId="9ab0b15ca0c94e92ad22c7bbab2f2666" sourceKey="XXX_DW_DB" systemId="5bafb54b77064064bafc021282506b36" port="10000"
          <dataSourceDb configId="0ee28c12aed1420084fe307744016a56" dbName="dwsrdsdb" lastUpdateUser="cde9b785ff244c2a8a6150ba6079e1d0" dbType
          <schemas>
            <schema sourceId="9ab0b15ca0c94e92ad22c7bbab2f2666" schemaShortKey="R" schemaId="607f01688ab24ebcb16577f7dfc8370d" domainName="HDS
              <objStandard dataFileType="" zipFileName="" dataObjName="#SCHEMASHORT#.#TABLENAME#" standardId="bb6a3b6ca63d4f9ab4c01aadcdf8a417
            </schema>
          </schemas>
        </dataSource>
      </dataSources>
    </system>
    <system systemId="aaeed4054fb048f59964ddd2818fe65b" systemName="TSYS" systemType="1" systemShortKey="TSYS" lastUpdateUser="f273746f4ff146c
      <dataSources>
```


3.2.5.1 发版包修改

项目导入测试、生产环境时，不同环境的连接和配置信息可能不同，需要修改后导入，修改的内容主要包括以下内容：

1. 数据源连接信息

```
<systems>
  <system systemId="8f930453e9de4fbcacc4008576ee461a" systemName="数据仓库" systemType="2" systemShortKey="EDWS" lastUpdateUser="489d4f6f57664a58b3bd7a1aa66c4261"
  description="" contactPerson="" systemKey="EDWS3" contactTel="" lastUpdateTime="2019-02-14 13:46:44">
    <dataSources>
      <dataSource sourceId="2a3ac94bf44341efa197ffabbd9f6c3f" sourceKey="DWSDB" systemId="8f930453e9de4fbcacc4008576ee461a" port="10000" dataEncoding="1"
      configId="3437a33333cc43e1bff3c605213b04f4" dataType="1" ipAddress="36.33.96.208" lastUpdateUser="640e8c7825634fd39d9f21590ffef723" description="装载R层表"
      sourceName="数据仓库" lastUpdateTime="2019-06-03 09:14:16">
        <dataSourceDb configId="3437a33333cc43e1bff3c605213b04f4" dbName="DWSRDSDB" lastUpdateUser="640e8c7825634fd39d9f21590ffef723" dbType="5" dbParams=""
        dbUrl="jdbc:hive2://36.33.96.208:10000/DWSRDSDB" lastUpdateTime="2019-06-03 09:14:16"/>
        <schemas>
          <schema sourceId="2a3ac94bf44341efa197ffabbd9f6c3f" schemaShortKey="EDWS_DFDS" schemaId="5c77726c8b274429839de365f3932a83" domainName="HDS_DOMAIN"
          userPwd="Y205dmRHaHBkbVU9" lastUpdateUser="489d4f6f57664a58b3bd7a1aa66c4261" description="" rootPath="" schemaName="DWSFDSDB" userName="hive"
          lastUpdateTime="2019-06-01 17:25:32" schemaKey="EDWS_DFDS">
            <objStandard dataFileType="0" zipFileName="" dataObjName=" #SCHEMASHORT#.#ORGCODE#_#TABLENAME#" standardId="075f3e2454e04a60a428c75601dc226e"
            ddlFileType="0" ctlFileType="0" rootPath="" dataEvtName="EVT_DIDP_#SCHEMASHORT#_#ORGCODE#_#TABLENAME#" tableName="#TABLENAME#" dataFileName=""
            okFileName="" ctlFileName="" ddlFileName="" metaObjId="5c77726c8b274429839de365f3932a83" lastUpdateUser="489d4f6f57664a58b3bd7a1aa66c4261"
            metaObjType="3" fixedLengthFlag="" verifyType="0" lastUpdateTime="2019-06-01 17:26:06" indepName=""/>
          </schema>
          <schema sourceId="2a3ac94bf44341efa197ffabbd9f6c3f" schemaShortKey="EDWS_DADS" schemaId="ceaa3a8cf27b4e6db57416f365f25838" domainName="HDS_DOMAIN"
          userPwd="Y205dmRHaHBkbVU9" lastUpdateUser="489d4f6f57664a58b3bd7a1aa66c4261" description="" rootPath="/tmp" schemaName="DWSADSDB" userName="hive"
          lastUpdateTime="2019-06-01 17:29:29" schemaKey="EDWS_DADS">
            <objStandard dataFileType="0" zipFileName="" dataObjName="#SCHEMASHORT#.#ORGCODE#_#TABLENAME#" standardId="38223dd33e7341cba1f01f1550c32618"
            ddlFileType="0" ctlFileType="0" rootPath="" dataEvtName="EVT_DIDP_#SCHEMASHORT#_#ORGCODE#_#TABLENAME#" tableName="#TABLENAME#" dataFileName=""
            okFileName="" ctlFileName="" ddlFileName="" metaObjId="ceaa3a8cf27b4e6db57416f365f25838" lastUpdateUser="489d4f6f57664a58b3bd7a1aa66c4261"
            metaObjType="3" fixedLengthFlag="" verifyType="0" lastUpdateTime="2019-06-01 17:28:52" indepName=""/>
          </schema>
        </schemas>
      </dataSource>
    </dataSources>
  </system>
</systems>
```


a. dataSource（数据源信息）：sourceKey（数据源主键）、ipAddress（数据库连接地址） 、dbName（数据库名） 、dbUrl（jdbc 连接串）

b. schemas（schema 信息）：userName（用户名）、userPwd（用户密码的加密串，联系平台管理员申请）、schemaName（schema 名称，表的所属 schema 名）

c. objStandard（schema 规范）：维持原样，不需要修改

2. 项目配置信息


```
<projectParamsConfig>
  <projectParam paramKey="inceptor.ktuser" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="HDFS认证用户" paramId="18f7763dbfc34b4ba53545b05c84d9b1" paramValue="hdfs" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="program.package.name" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283"
  projectVersionId="9a3aca26f54749f281894691985832d4" paramName="程序包名" paramId="28c35eea9dc64c5ba8a547316d2e7863" paramValue="DIDP_DATA_PROCESS_BAG"
  lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="moia.control.pwd" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="调度服务密码" paramId="2e1a97f50a2b41b3928b21ba31f07a58" paramValue="admin" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="inceptor.nnurl" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="HDFS中namenode的IP地址" paramId="6be3e2200c9346b9a5453a923f7ddb2e" paramValue="http://36.33.96.208:50070;http://36.33.96.209:50070"
  lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="moia.control.port" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="调度服务端口" paramId="86137b93649e4aa9bf1a6a92ace00f21" paramValue="22512" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="load.func.name" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="装载组件程序名" paramId="9374b0567b634de4aac98f99bc51f5c6" paramValue="DIDP_LOAD_DATA" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="moia.control.user" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="调度服务用户名" paramId="9da9ac27a13b4c288ac07c1c069fb05c" paramValue="admin" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="program.file.path" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="程序路径" paramId="a32f2d66db70488f8c0f5f0e9e162ac3" paramValue="/home/moiaagent/appjob/pythonjob" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="moia.control.ip" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="调度服务IP地址" paramId="b1f139d3a78045adb7657e3a2af2f67" paramValue="36.33.96.142" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="archive.dayFunc.name" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283"
  projectVersionId="9a3aca26f54749f281894691985832d4" paramName="日常归档组件程序名" paramId="b815acc42c5b4dc890327b94b1a6110e"
  paramValue="ARCHIVE_FROM_HIVE_TABLE" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="collection.fileFunc.name" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283"
  projectVersionId="9a3aca26f54749f281894691985832d4" paramName="文件采集组件程序名" paramId="dace783eb6b24f9f8f7edfeb39ad8980" paramValue="DIDP_EXPORT_FILE"
  lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="collection.dbFunc.name" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283"
  projectVersionId="9a3aca26f54749f281894691985832d4" paramName="数据库采集组件程序名" paramId="ede40ab781b849a384179823dff43156" paramValue="DIDP_EXPORT_DATA"
  lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="app.system.name" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="应用系统名" paramId="f02db5da8c9042a4a90e0de0fb16b92c" paramValue="DIDP_APP" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="archive.initFunc.name" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283"
  projectVersionId="9a3aca26f54749f281894691985832d4" paramName="初始化归档组件程序名" paramId="f19e4a2edbe645329c7b6fe43212b307"
  paramValue="BATCH_ARCHIVE_FROM_HIVE_TABLE" lastUpdateTime="2019-06-03 19:59:04"/>
  <projectParam paramKey="prep.func.name" description="" lastUpdateUser="2648d5965253462abf59a6d7224ab283" projectVersionId="9a3aca26f54749f281894691985832d4"
  paramName="预处理组件程序名" paramId="f846b1e3a65746e682cf3be21a4df4bf" paramValue="DIDP_PRE_PROCESS" lastUpdateTime="2019-06-03 19:59:04"/>
</projectParamsConfig>
```


a. HDFS 中 namenode 的 IP 地址：修改为对应环境实际的 namenode 连接地址信息（例如：http://36.33.96.208:50070;http://36.33.96.209:50070）

b. HDFS 认证用户

c. 调度服务 IP 地址

d. 调度服务端口

e. 调度服务用户名

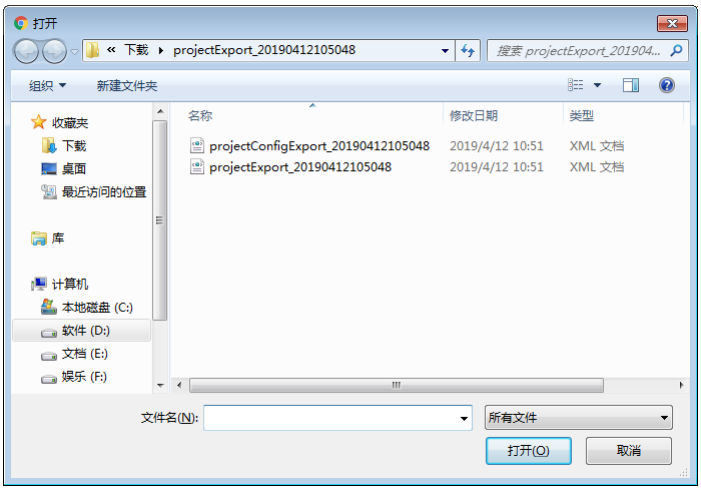
f. 调度服务密码：明文密码

g. 应用系统名：与 moia 上的应用系统配置名称一致

3.2.6 项目导入

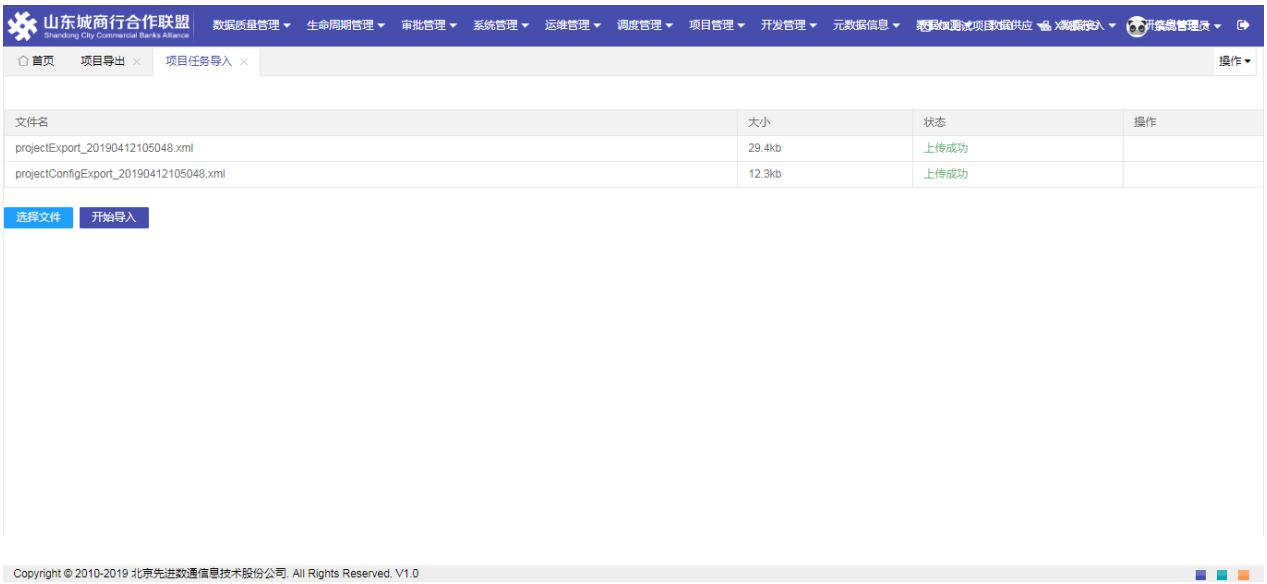
项目导入用于开发完成任务到生产环境的发布操作。

上传从项目导出中获得的两个文件，将配置文件的连接信息配置正确。



点击“开始上传”

等待上传成功的提示。



点击“开始导入”按钮，开始导入操作。

3.2.7 项目元数据分配

项目元数据分配用于架构管理员为项目配置可用的 SCHEMA 级元数据权限。权限包括控制和访问两种。

➢ 配置说明：

- 控制：对 SCHEMA 下的元数据有控制权限（可读、可写）
- 访问：对 SCHEMA 下的元数据有访问权限（可读）
- 变更控制：从仅访问权限的 SCHEMA 变更为控制权限
- 变更访问：从仅控制权限的 SCHEMA 变更为访问权限
- 移除：取消 SCHEMA 的控制权或访问权限

3.3 调度管理

3.3.1 调度配置

调度配置用于配置调度平台上的调度对象，调度配置的对象包括计划、任务、作业三类。调度配置为项目内作业开发的调度命名提供规范，配置的名称可以使用已支持的宏进行定义，在实际生成时会动态替换，宏定义参考章节《5.1.宏定义说明》。用户在配置时的顺序依次为：计划、任务、作业。

调度对象支持多节点配置，单个计划下可以配置多个任务，单个任务下可配置多个作业，在任务分配时，会选择相应的作业名挂载，则以此继承上层的任务名和计划名。

3.3.2 调度分配

调度分配用于为项目分配可用的调度对象资源，只有为项目分配调度后，在任务分配时才可被选择。

操作类型包括新增和删除两种操作。

■ 调度分配流程：左侧选择项目，选择操作类型-->选择调度的对象名称-->保存。

3.3.3 调度日历

调度日历用于配置调度的执行日历。

3.4 开发管理

3.4.1 系统信息维护

系统信息维护提供源系统配置及规范配置等功能。

规范配置提供系统层级的规范配置，例如，系统是以文件推送的形式推送过来，则需要在此描述推送的目录、文件名、控制文件格式、DDL 文件格式等信息，规范配置中的内容支持宏变量，具体可使用的宏变量参考《5.1 宏变量说明》。在 schema 下也有相应的规范配置，新建 schema 时，schema 上的规范继承于系统上的规范，schema 配置完成后，schema 规则则独立保存，schema 规范的作用域为整个 schema 下对象，建议配置 shema 规范。

- 配置说明
  - 系统标识：平台内所有系统具有唯一性；
  - 系统名称：系统的中文名称；
  - 系统简称：4 位系统简称（必填），与宏变量“#SYSSHORT#”相对应；
  - 系统类型：区分源系统和 DW；
  - 联系人：该系统的联系人；
  - 联系电话：联系电话录入。

3.4.2 数据源

数据源提供系统的连接信息配置功能。在这里数据源主要分为两种，数据库和文件。

3.4.2.1 数据库类型

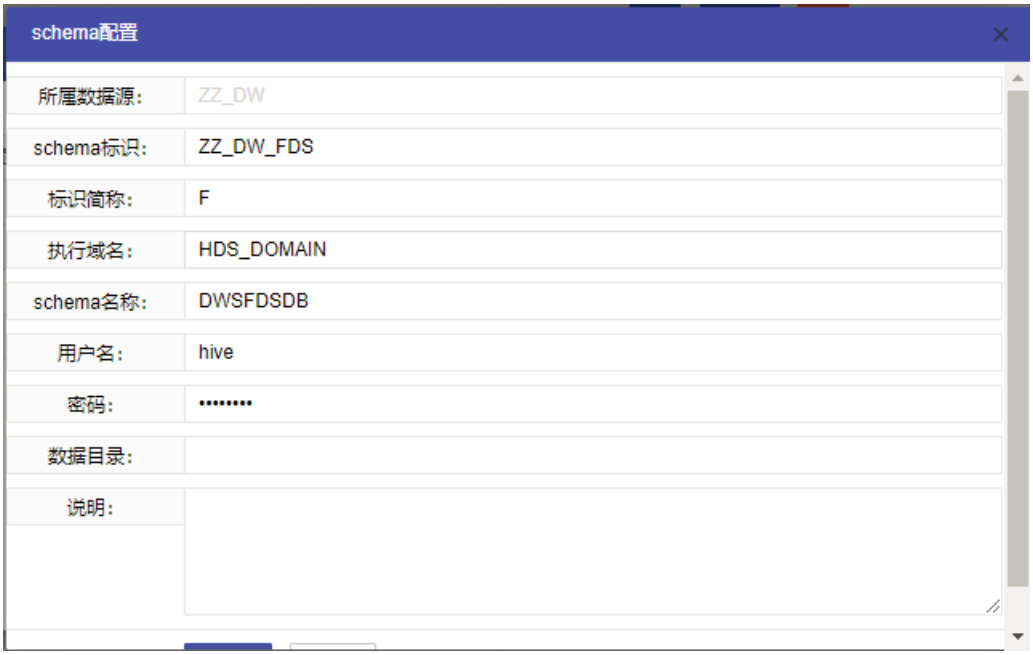
数据库类型的数据源配置根据实际信息填入。

■ 配置说明

- 数据源标识：数据源全局唯一；
- 数据源名称：通常键入中文描述信息；
- 所属系统：从已配置的系统中选择；
- URL：数据库的 jdbc 连接串，自动生成。

■ SCHEMA 配置

- SCHEMA 标识：全局唯一；
- 数据目录：若为 Inceptor 的加载 schema，则需要在此配置数据目录，装载时的外表数据文件存放目录从此处获取。



■ SCHEMA 规范

- 表名：该 schema 上的命名规范，平台内对于自动生成目标表的信息，从该规范上获取，支持宏定义组合，例如 R_[ORGCODE]_#SYSSHORT#_#TABLENAME#_#COLLTYPE#_[DATE]
- 数据对象名：全局唯一的数据对象定义，例如 #SCHEMASHORT#.#ORGCODE#_#SYSSHORT#_#TABLENAME#_#COLLTYPE#
- 数据事件名：全局唯一的数据事件名定义，例如 EVT_DIDP_#SCHEMASHORT#_#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#
- 独立资源名：调度平台内的独立资源名（暂未启用）。

3.4.2.2 文件类型

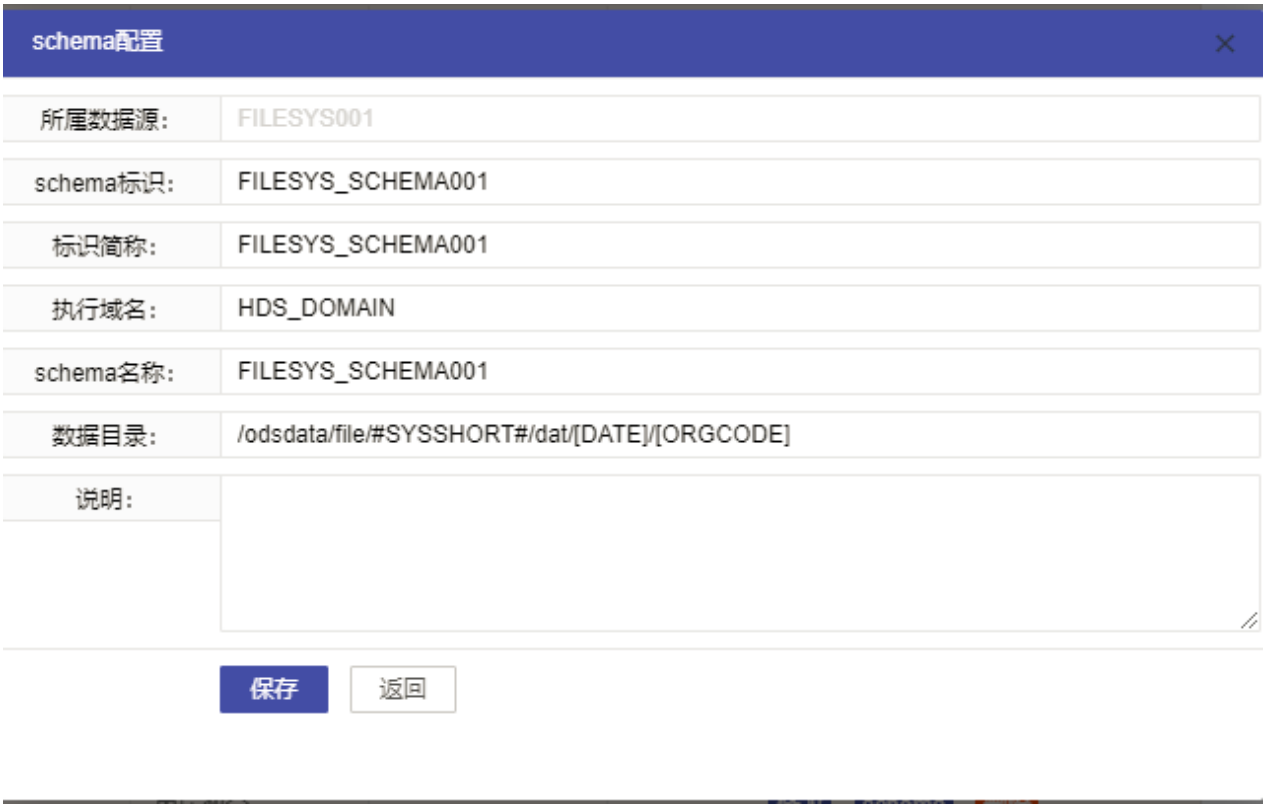
文件类型的数据源配置根据实际信息填入。

■ 配置说明

- 数据源标识：数据源全局唯一；
- 数据源名称：通常键入中文描述信息；
- 所属系统：从已配置的系统中选择；
- 服务器类型：支持本地和远程两种方式。

■ SCHEMA 配置

- SCHEMA 标识：全局唯一；
- 数据目录：数据文件的主目录。



■ SCHEMA 规范

- 主目录：作为以下文件类型的主目录；
- 数据文件名：相对主目录的一个相对路径，与主目录一起作为完整数据文件绝对路径；
- CTRL 类型：无、XML、INDICATE、PREPERTIES、TEXT；
- CTRL 文件名：相对主目录的一个相对路径，与主目录一起作为完整 CTRL 文件绝对路径；
- DDL 类型：无、XML、SQL；
- DDL 文件名：相对主目录的一个相对路径，与主目录一起作为完整 DDL 文件绝对路径；
- 压缩文件名：如果源系统推送的文件是压缩文件，则需要配置压缩文件名；
- OK 文件名：与压缩文件名对应，若有 ok 文件则配置 OK 文件名；
- 表名：#TABLENAME#；
- 数据对象名：全局唯一的数据对象定义，例如 #SCHEMASHORT#.#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#

- 数据事件名：全局唯一的数据事件名定义，例如EVT_DIDP_#SCHEMASHORT#_#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#
- 独立资源名：调度平台内的独立资源名；

注：此处的主目录+数据文件名组合起来的文件路径，原则上与 SCHEMA 配置中的主目录一致。

3.4.3 公共模板

公共模板功能用于维护平台内的公共加工模板信息。

新增分类,新增模板类型分类，点击公共模板后，在左侧空白区域右键点击新增分类。

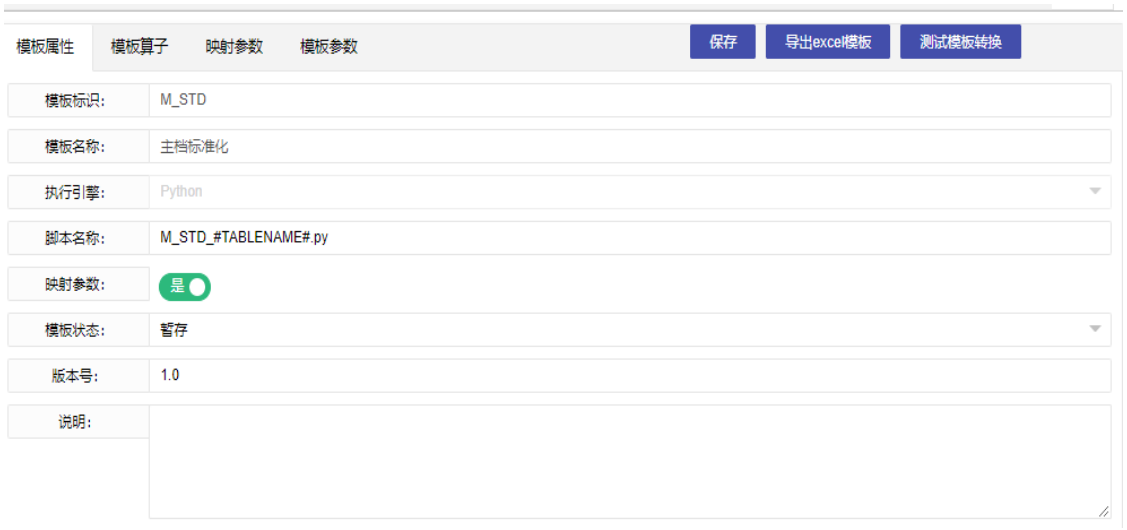
首页
公共模板
资源名称
查询
XXX
数仓加工
数据加工
新增分类

新增分类
分类名称:
数仓DEMO
保存
返回

■ 点击新增模板。

首页
公共模板
操作
资源名称
查询
XXX
数仓DEMO
数仓加工
拉链算
1.0
数据加工
新增分类
新增模板
编辑
删除
模板属性
模板算子
映射参数
模板参数
保存
导出excel模板
测试模板转换
模板标识:
模板名称:
执行引擎:
请选择
脚本名称:
映射参数:
否
模板状态:
请选择
版本号:
说明:

■ 依次配置模板的属性，模板算子，模板参数



序号	参数标识	参数名称	类型	默认值	是否必须	说明	操作
1	source_table	源表名	值		是		编辑 删除
2	mappingColumn	字段映射	列表		是		编辑 删除
3	is_std	是否标准化	值		否		编辑 删除
4	source_sys_table	接入的系统...	值		是		编辑 删除
5	type	加工类型	值		是		编辑 删除
6	source_db	源库名	值		是		编辑 删除
7	target_table	目标表名	值		是		编辑 删除
8	target_db	目标库名	值		是		编辑 删除
9	source_system	来源系统	值		是		编辑 删除

- 在模板开发填写完成后，可以对模板生成的脚本进行测试。

  步骤：

  1. 点击导出 EXCEL 模板。
  2. 填写 EXCEL 模板。
  3. 点击测试模板转换，选择填写好的 EXCEL 模板，此时会生成 python 文件，若 python 文件与预期不符，则对模板算子进行更改。

- 测试通过后，将模板的状态改为发布。并在开发作业配置种添加该公共模板，至此，公共模板就可以使用了。

3.4.4 公共组件

公共组件功能用于维护平台内的公共组件信息。

- 组件属性

  组件标识：平台内组件的唯一标识，规范由字母、数字、下划线组成；

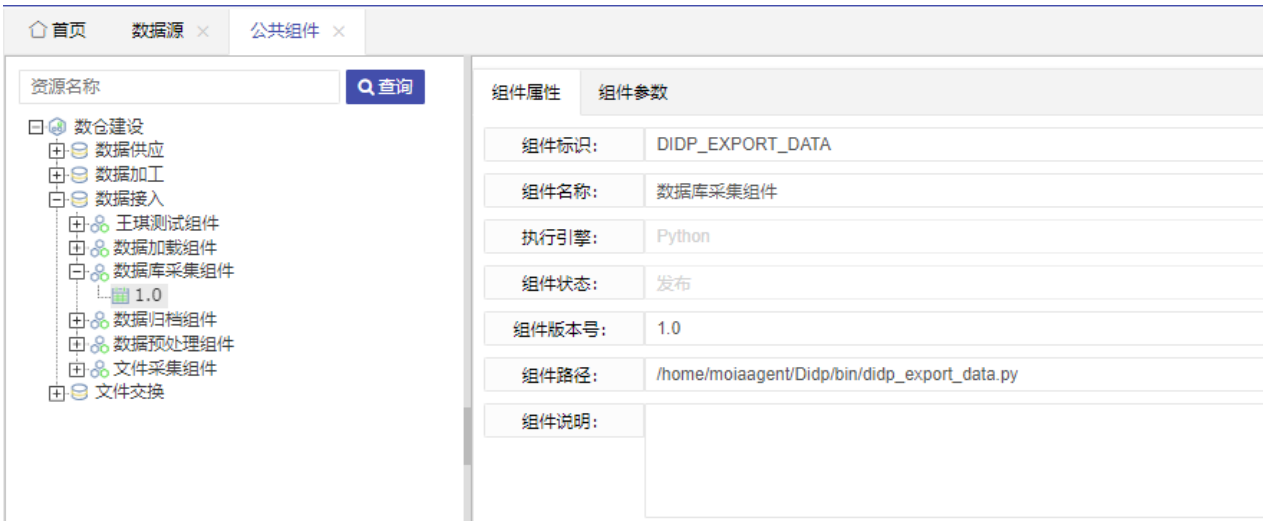
  组件名称：组件的中文名称

  执行引擎：组件程序对应的执行引擎

  组件状态：暂存、发布。发布后，才能用于项目开发使用

  组件版本号：建议采用数字编号，例如 1.0

  组件路径：组件程序的存放路径



■ 组件参数

参数名：参数名称

参数类型：参数的类型

参数前缀：执行命令中的参数名

默认值：默认参数

是否必须输入：是、否

3.4.5 实体表配置

实体表配置主要用于平台内需要导入的一些信息，提供对表和视图的维护功能。自动映射到菜单的资源树下。

■ 配置说明

菜单：实体表挂载的菜单位置，不新增菜单时从已有的菜单中选择，如果需要新增菜单，需要输入菜单名称和选择挂载的上级菜单

分组：实体表所属的分组，如果分组已存在则直接选择，不存在可新建

是否固定模板：模板下载的方式，配置固定模板时，需要用户上传模板文件

数据起始行索引：文件包含几行表头和描述信息

是否视图：如果是视图，需要配置字段的中文名称

3.5 元数据信息

3.5.1 数据字典

数据字典提供对元数据的维护功能，用户可通过页面维护元数据信息。同时提供批量导入、导出元数据信息到 Excel，为方面快速建模，提供导出 SQL 建表语句功能。

■ 导入

提供数据字典的导入功能，操作流程

1. 选择“模板下载”，下载模板填入数据字典信息。

2. 填写模板，填写的内容包括“表信息”和“字段信息”两个 sheet 表格

   注：系统标识、数据源标识、SCHEMA 标识必须在平台已配置存在，方可导入

3. 模板填写完成后，点击“选择文件”，选择完成后，点击“导入”，完成数据字典的导入

■ 导出 Excel

按照数据字典的模板导出 Excel 文件。

■ 导出 ddl

导出对应数据库类型的 DDL 建表语句文件。

3.5.2 元数据分析

元数据分析提供对平台内元数据的分析功能。

■ 分析形式

- 血缘分析：以该数据对象为出发点，向前推导出产生该数据对象的上游数据对象节点
- 影响分析：以该数据对象为出发点，向后推导出该数据对象会影响到的下游数据对象节点

■ 分析粒度（对象类型）

- 表
- 字段

对于分析出来的图，可导出 img 格式的图片文件。

■ 结果导出

- 图片导出：导出 img 格式的图片文件
- XML 导出：导出 XML 格式的文件。

3.6 审批管理

对开发人员开发完成后提交的任务进行审批发布。

- 查看

  跳转到开发人员的开发任务页面，查看开发的内容

- 审批

  项目管理员对开发的任务进行通过或拒绝操作。拒绝时，建议填写拒绝的原因，以便开发人员能及时跟进问题原因，及时修改。

- 发布

  发布生成作业运行的相关信息，包括生成调度作业信息、加工作业生成加工脚本。

3.7 作业地图

作业地图提供平台内开发的数据对象（表）与调度对象（计划/任务/作业）之前关系的查询功能。

3.8 数据接入

数据接入是相对项目而言，在逻辑上可接入数据项目需求的任何数据资源。可以从业务数据源端或者其它数据项目进行接入，接入的数据种类可以结构化、半结构化、分结构化的数据，接入的源类型可以是数据库也可以是文件类型（主动采集、被动接收）。

数据接入流程包含三个流程子节点：采集、预处理、装载。

接入的配置流程：在数据对象上右键选择，选择接入的类型，再进行后续的开发。

3.8.1 数据接入

3.8.1.1 数据采集

数据采集会根据接入源的类型进入到不同的功能配置页面，主要包括数据库和文件两种。

- 数据库采集

  采集的参数会默认生成，仅提供过滤条件可由用户自定义修改。

■ 文件采集

。

■ 调度配置

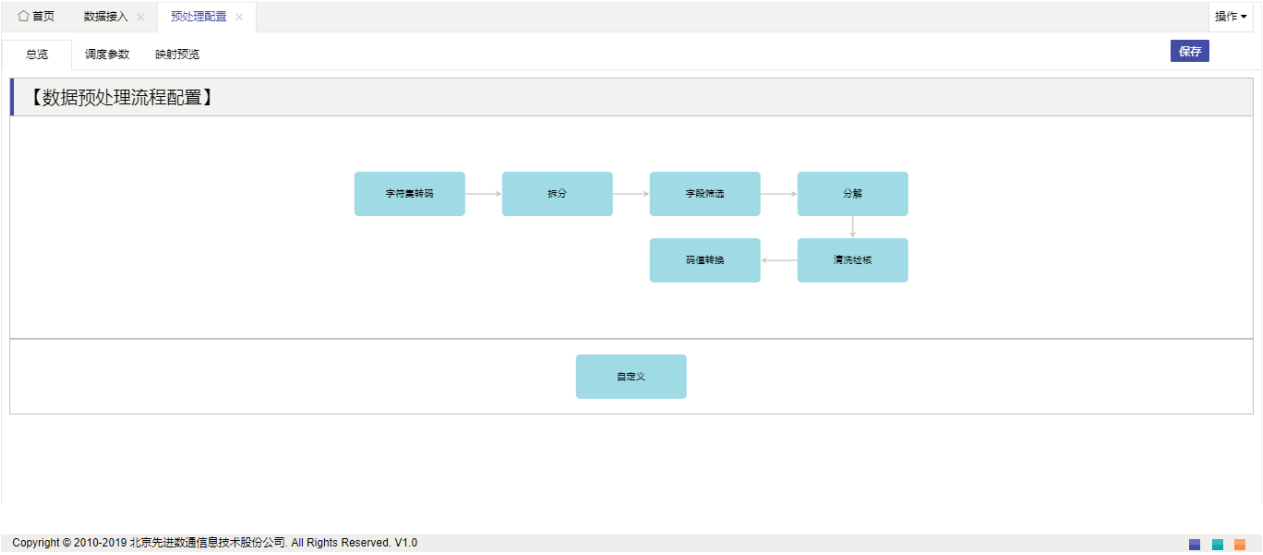
配置作业的运行的情况，此配置会有一个初始默认配置，可根据实际需要进行调整。

3.8.1.2 数据预处理

数据预处理是保证数据质量的基础，对文件形式的数据的进行清洗检核，预处理的功能包括字符集转码、字段拆分、字段筛选、数据分解、清洗检核、规则转换，各功能配置模块之前没有强制依赖，可根据实际情况进行选择，双击进入相应功能的配置。

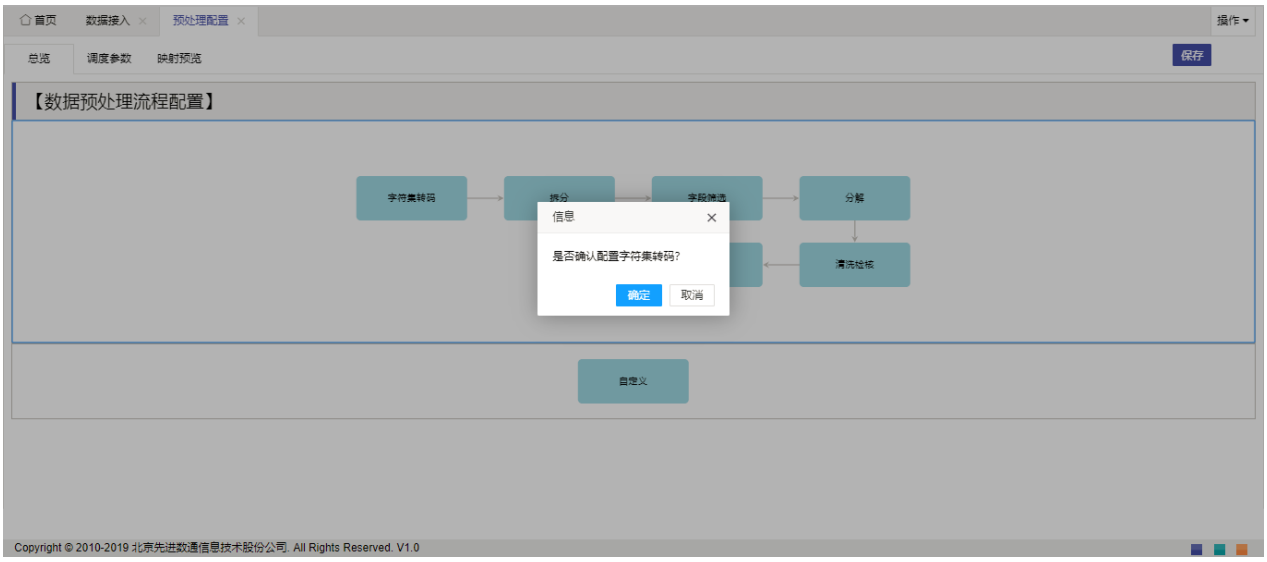
在预处理配置完成后，默认添加 BANK_ID 和 BATCH_DT 两个技术字段，且这两个

字段不允许更改。



3.8.1.2.1 字符集转码

字符集转码提供 utf8 和 gbk 编码互转功能。

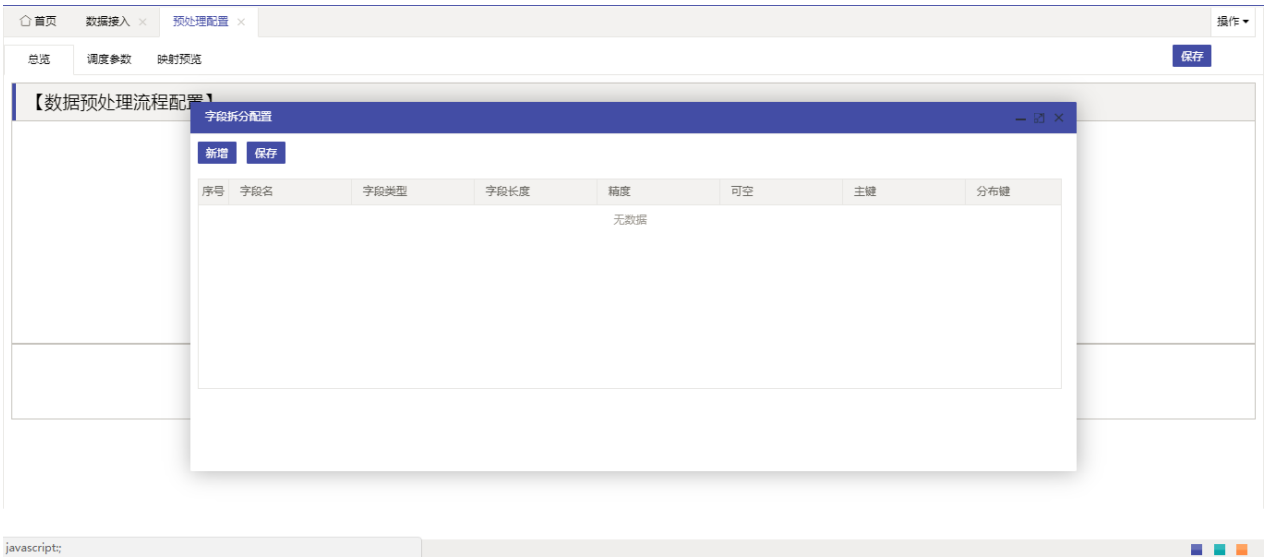


3.8.1.2.2 字段拆分

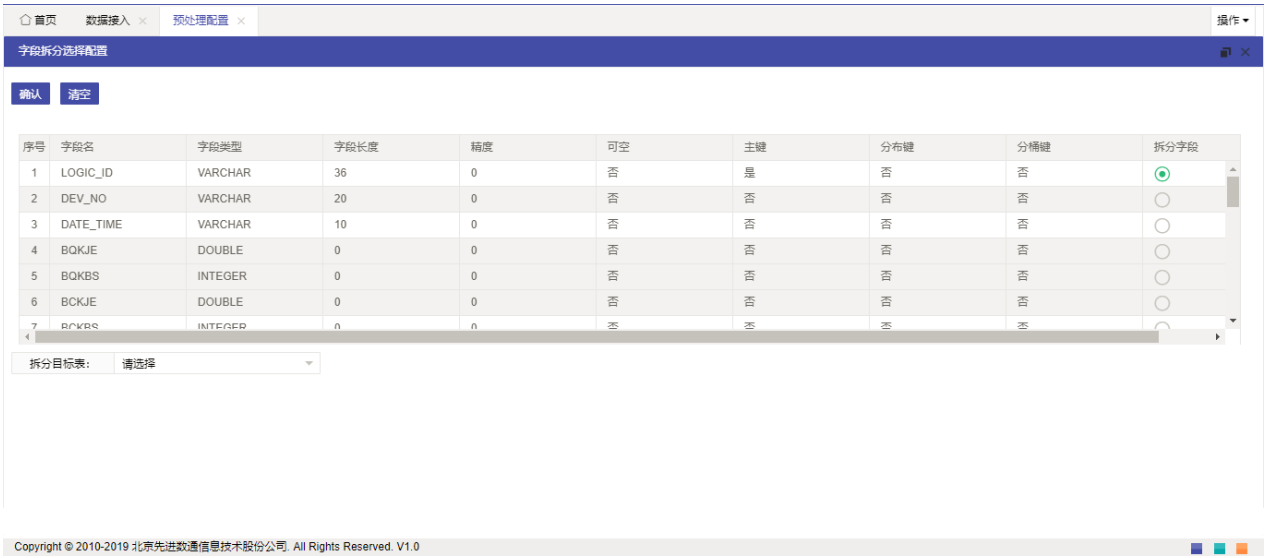
字段拆分提供将字段拆分多个字段的功能。字段拆分信息需要预先到数据字典中登记（新增表及拆分后的字段信息），表名建议“源表名_split_拆分字段名”。

■ 字段拆分操作流程：

1. 点击“拆分”功能框，弹出字段拆分配置功能页面；



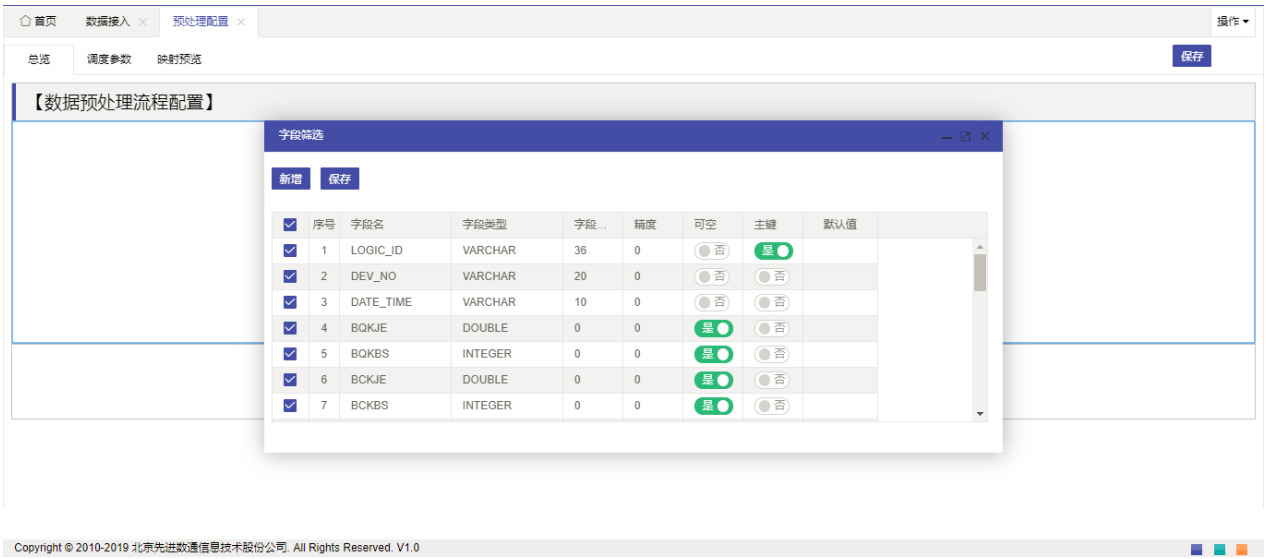
2. 点击“新增”，进行拆分功能配置；



3. 选择拆分字段，并且在下方拆分目标表中选择拆分表，完成拆分功能配置；
4. 点击“确认”，即完成字段拆分功能配置。

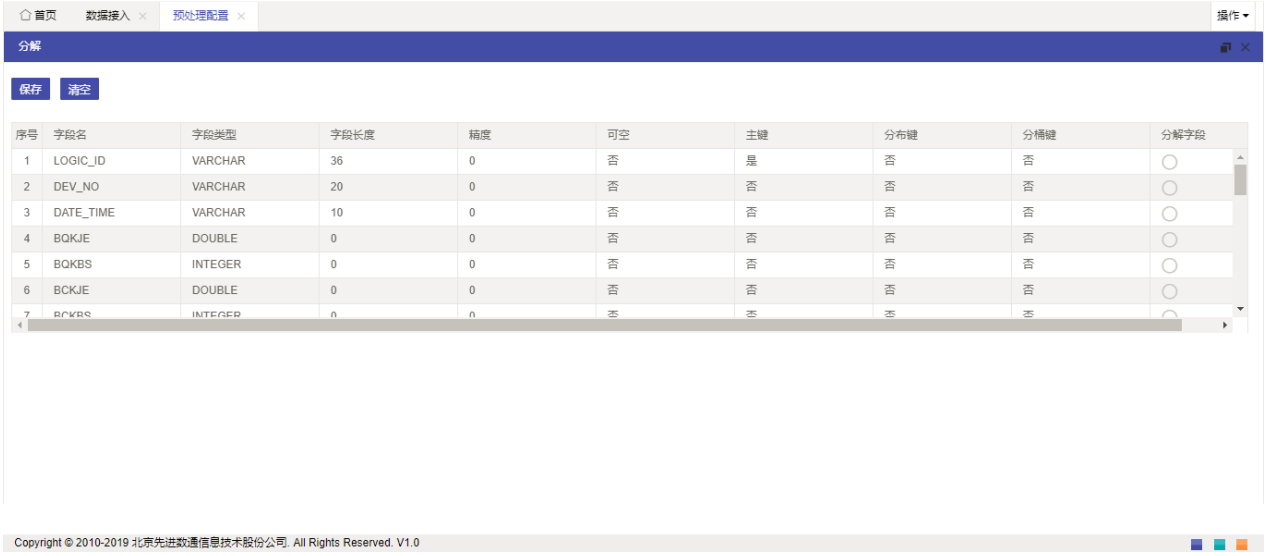
3.8.1.2.3 字段筛选

字段筛选主要提供字段筛选的功能，若不需要某些字段，点击字段行首的反选框。字段筛选功能页面，同时提供数据检核功能。



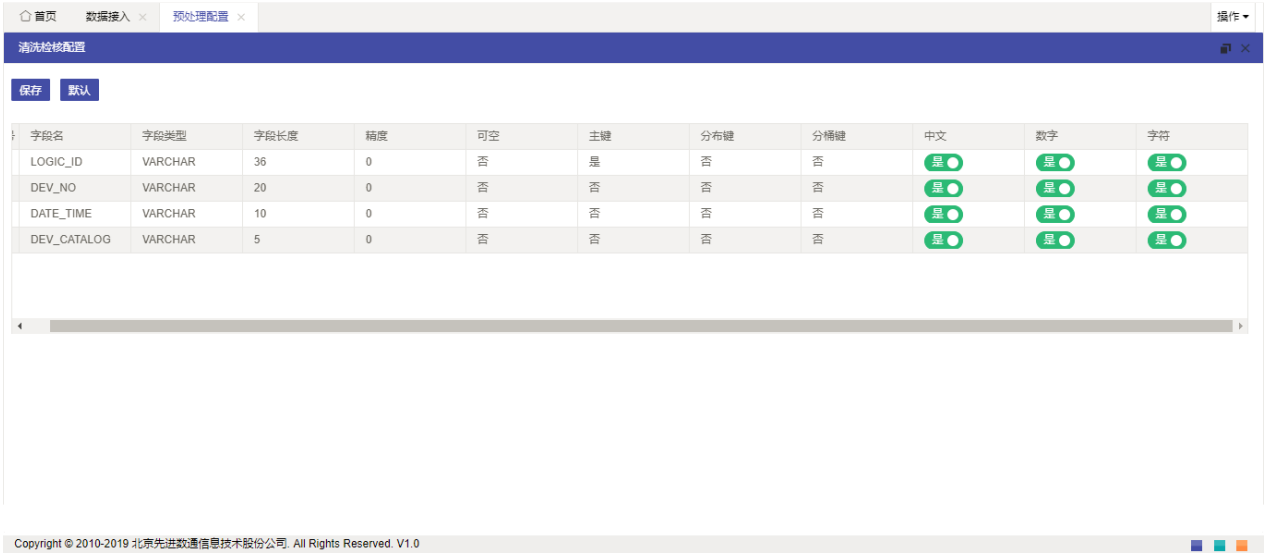
3.8.1.2.4 分解

分解是指对数据集合按特定维度拆分成多个数据集合的功能，单个文件仅支持对一个字段进行分解。分解功能使用较多的场景为总行数据按法人/机构的维度进行数据拆分。



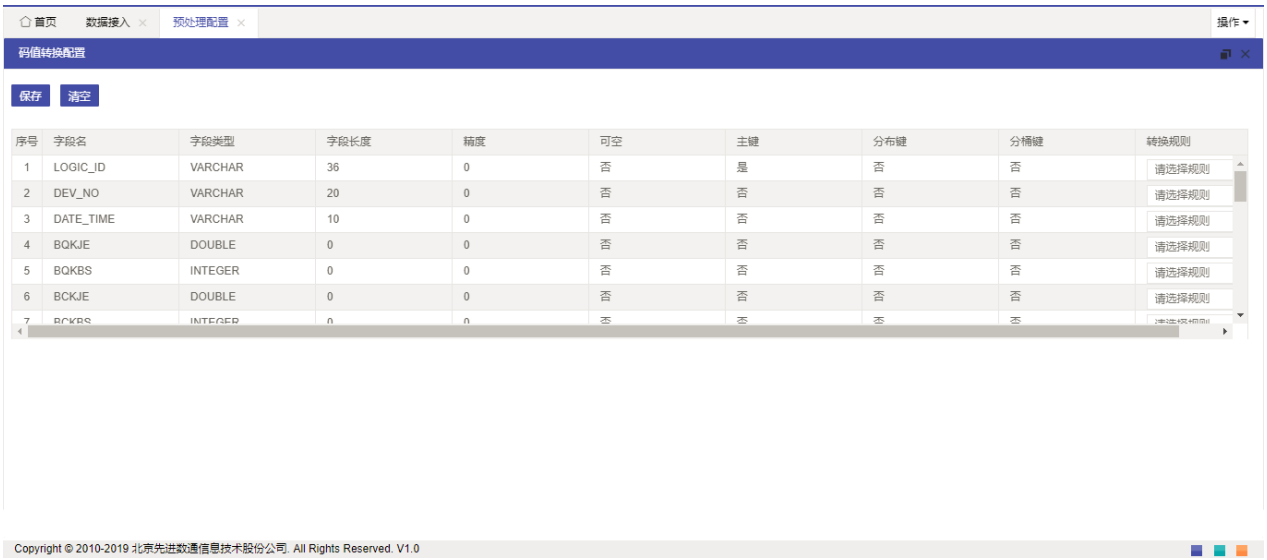
3.8.1.2.5 清洗检核

清洗检核提供对数据内容的技术性检核，分为基础检核和扩展检核。基础检核根据字段的类型及长度对字段内容的合法性进行检核；当检核的对象为字符类型，提供扩展检核功能，检核的内容包括中文、数字、字符等信息。例如，中文-否、数字-是、字符-否，表示该字段中仅包含数字，若数据中包含非数字内容，则该记录会当做非法数据拒绝掉。



3.8.1.2.6 码值转换

码值转换提供对数据的转换功能，转换功能可通过“转换规则”功能进行配置，参考章节《3.6.3 转换规则》。若需要进行转换，在字段信息最后的功能菜单中下拉选择已配置的规则。



3.8.1.3 数据装载

3.8.1.4 批量导入

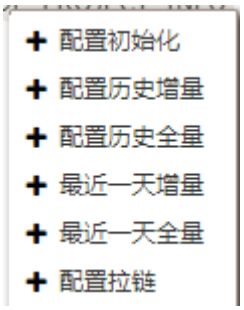
3.8.2 数据归档

数据归档提供对源系统、数仓等系统的数据归档功能，归档的数据源目前支持 Inceptor 外表和 ORC 表。

对于数据接入的归档，依赖于数据装载，只有在数据装载开发完成后，分配的对象才会在归档左侧对象资源数上显示，开发人员才能开发。

■ 归档配置说明

- 归档类型：初始化、历史增量、历史全量、最近一天增量、最近一天全量；



- 归档类型说明：

  初始化（流水档初始化）:按照给定的日期字段和日期字段格式以及分区方式将源表动态归档到目标库中。

  历史全量:保存源表所有日期的全量数据。

  历史增量:保存源表所有日期的增量数据。

  拉链：保存源表的最新截面以及历史变更信息。

  最近一天全量：保存源表中最新的一份全量数据

  最近一天增量：保存源表中最新的增量数据。

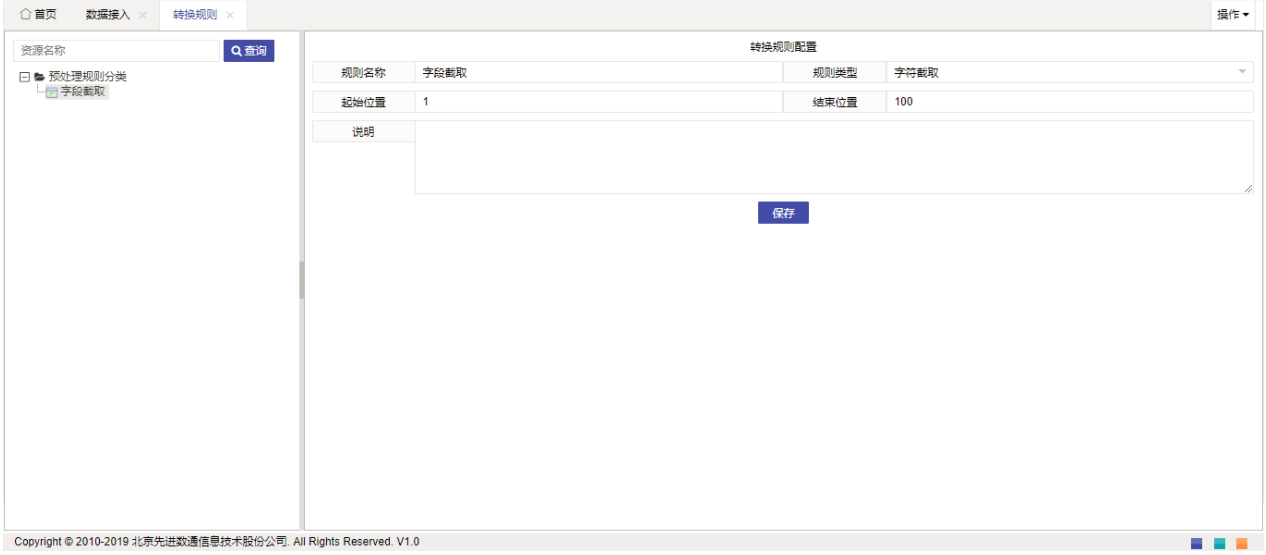
- 目标数据源：选择归档的目标数据库 SCHEMA；
- 归档表名：系统会默认根据归档类型自动生成表名，若归档表名不符合规范要求，可自定义配置；
- 日期分区方式：支持不分区、月、季、年分区配置，系统默认会增加一个日

期分区字段；

- 是否机构分区：设置是否需要增加机构分区维度；
- 分桶键：Inceptor 数据库表选择分桶键；
- 分桶数：Inceptor 数据库表的分桶个数。

3.8.3 转换规则

提供规则配置，转换规则包括字符截取、字符替换功能。



3.9 数据加工

数据加工功能为数据整合提供服务，内置多种公共算法组件、公共算法模板，满足数据整合过程中的各类需求。

数据加工是以目标加工对象的视角进行加工。

■ 作业的开发步骤：

1. 新增加工类型；
2. 选择添加源表，下载开发模板；
3. 配置开发模板；
4. 导入模板；
5. 查看并确认配置信息，并确认提交。

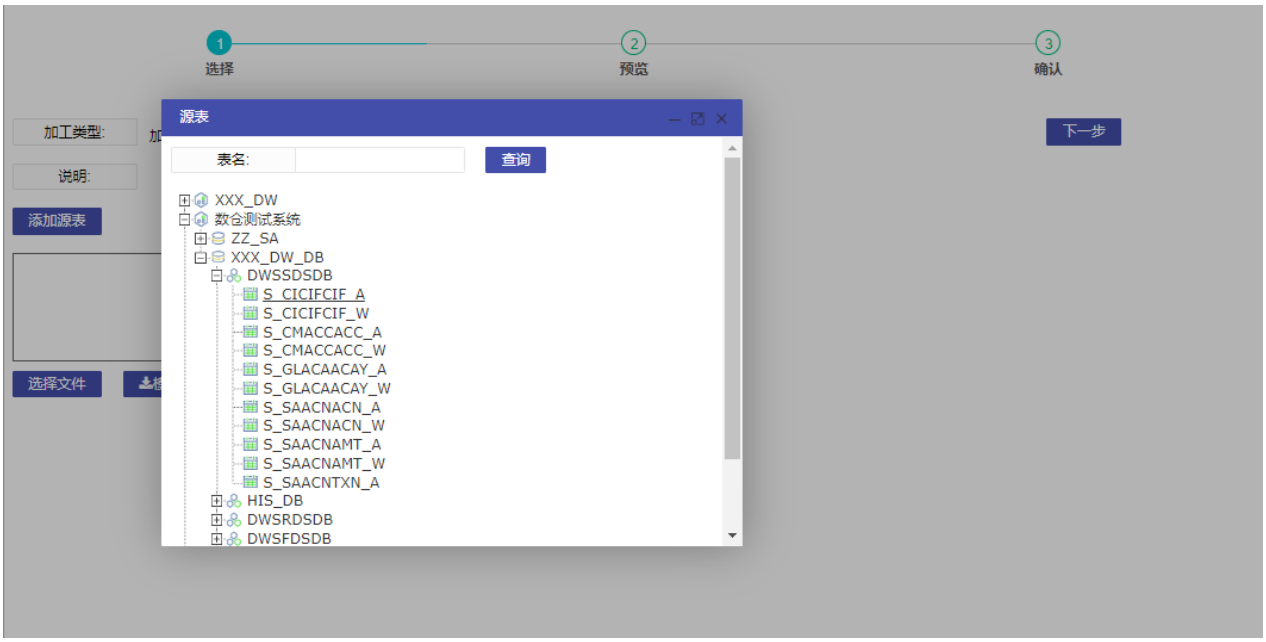
3.9.1 配置开发

■ 在待开发的目标表上，右键“新增加工类型”，进行配置。

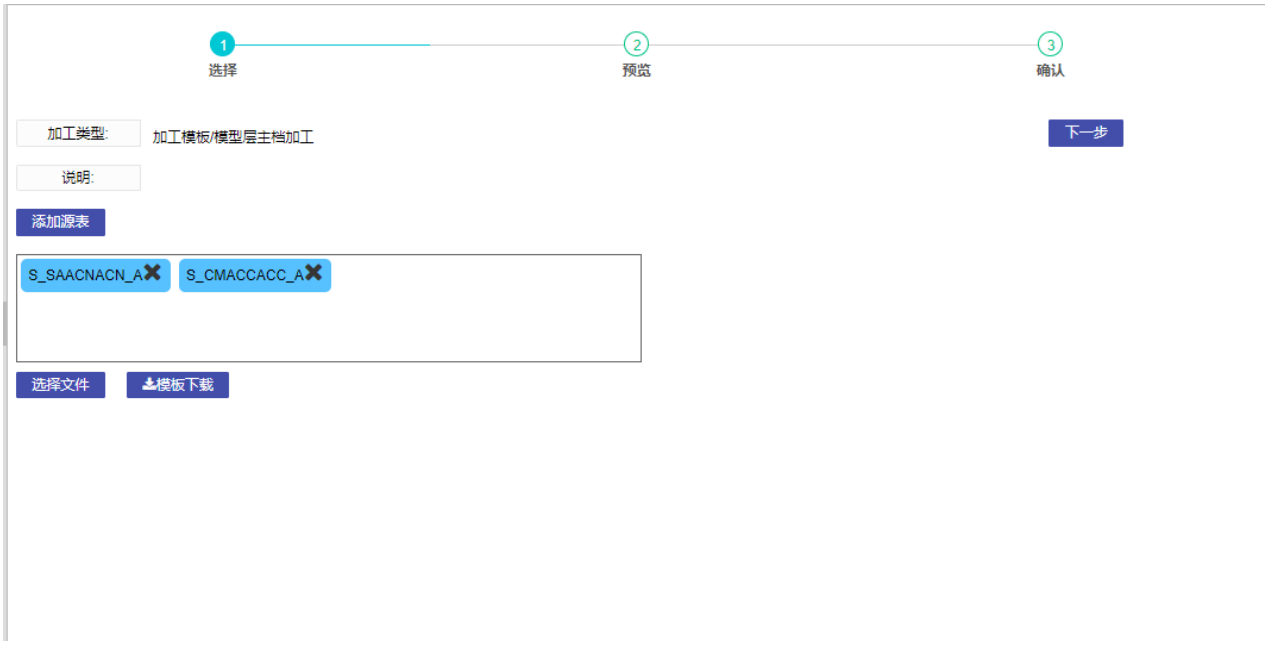
■ 配置说明

- 计划类型：日常、不定期；
- 创建类型：
  - 按类型创建：从平台内置的类型算法进行加工
  - 自定义：自定义脚本进行加工
- 加工类型：按类型创建时，从此处选择内置的加工类型
- 加工名称：此加工作业的描述说明,默认填 00，如果一个目标表由多种加工类型或者源表加工而成时，建议依次配置，方便辨识于该加工表其它的加工类型。

■ 添加加工依赖的源表



“模板下载”：下载以 Excel 文件形式的开发模板，模板内容包括待开发内容（加工参数、字段映射、表级映射）和参考信息（目标表信息、目标表结构、源表结构）



■ 配置开发模板

- 字段映射

  映射组编号：与“表级映射”中的映射组编号对应，相同的映射组编号的字段映射和表级映射为一组

目标字段名：映射关系中的目标表目标字段名

源表名：映射关系中对应源表的表名，非源表映射字段时可不填

源字段名：映射关系中对应源表字段的字段的名，非源表映射字段时可不填

映射规则：非源表字段直接映射时，需要在此处填写，填写的方式包括常量值、表达式（‘CBUS’||T1. TLR_NO）、内置标记表达式（ACCT_TX::T1.ACCT_NO)、常量标记（#IGNORE#），

- ✧ 以“ACCT_TX::”开头，表示该字段需要做账号代理键转换
- ✧ 如果填写#IGNORE#，表示该字段映射不计入生成的算法中，主要用于维护字段映射的血缘关系，例如，目标字段 AR_TM_DT 由源表 3 个字段 COLA、COLB、COLC 连接组成，配置字段映射时，第一个配置映射规则，另外两个配置成#IGNORE#：

映射组编号*	目标字段名*	目标字段中文描述	源表名	源字段名	源字段中文描述	映射规则
MAPPING001	AR_ID					'CBUS'||SA_ACCT_NO
MAPPING001	AR_TM_DT		S_CBUS_SAACNAMT_	COLA		T1.COLA||T1.COLA||T1.COLB
MAPPING001	AR_TM_DT		S_CBUS_SAACNAMT_	COLB		#IGNORE#
MAPPING001	AR_TM_DT		S_CBUS_SAACNAMT_	COLC		#IGNORE#
MAPPING001	AR_TP_CD					''

- 表级映射：

映射组编号：与“字段映射”中的映射组编号对应

schema 名称：表的 schema 名（数据库名）

表名：表名

表别名：该表的别名，同一组映射中别名不能重复

关联类型：下拉选择，包含主表、内关联左关联、右关联、外关联、全关联

关联条件：ON 的条件，填写时不带 ON 关键字，如果主表有过滤条件，填写在此处。

- 模板参数：

具体的模板参数填写值参考参数说明

模板示例：

数据模板加工示例.xls

■ 导入模板

选择文件-》导入

■ 查看并确认配置信息，并确认提交

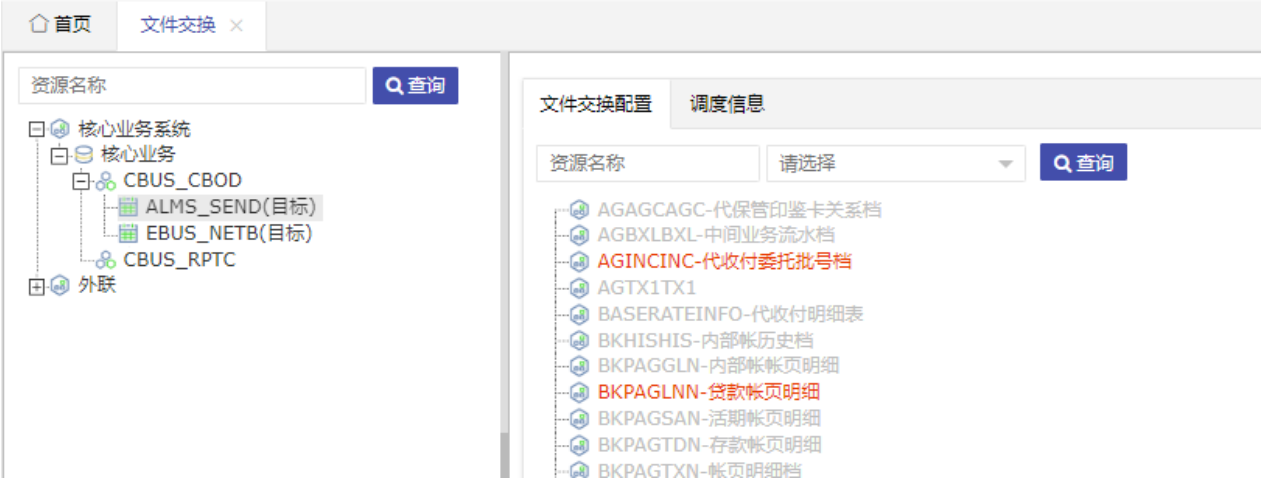
依次“下一步”，查看参数信息、映射关系图、生产的脚本内容等信息，确认无误后，点击提交完成。

3.10 数据供应

3.10.1 文件交换

文件交换提供将源系统文件直接交换到目标系统指定目录下。以 SCHEMA 为组织单位进行开发。

资源树中一级节点为文件交换的源系统，二级节点为交换的目标系统。



■ 新增交换目标 SCHEMA

文件交换类型：日常、临时。

文件交换目标系统：交换的下游系统

目标 schema：通常为数据项目的 SEND 区，只能选择有权限的文件类型的 SCHEMA

■ 选择交换的档，并配置

文件交换数据类型：全量、增量

文件交换方式：COPY、LINK

源数据文件：根据源 schema 的配置规范自动生成，可更改

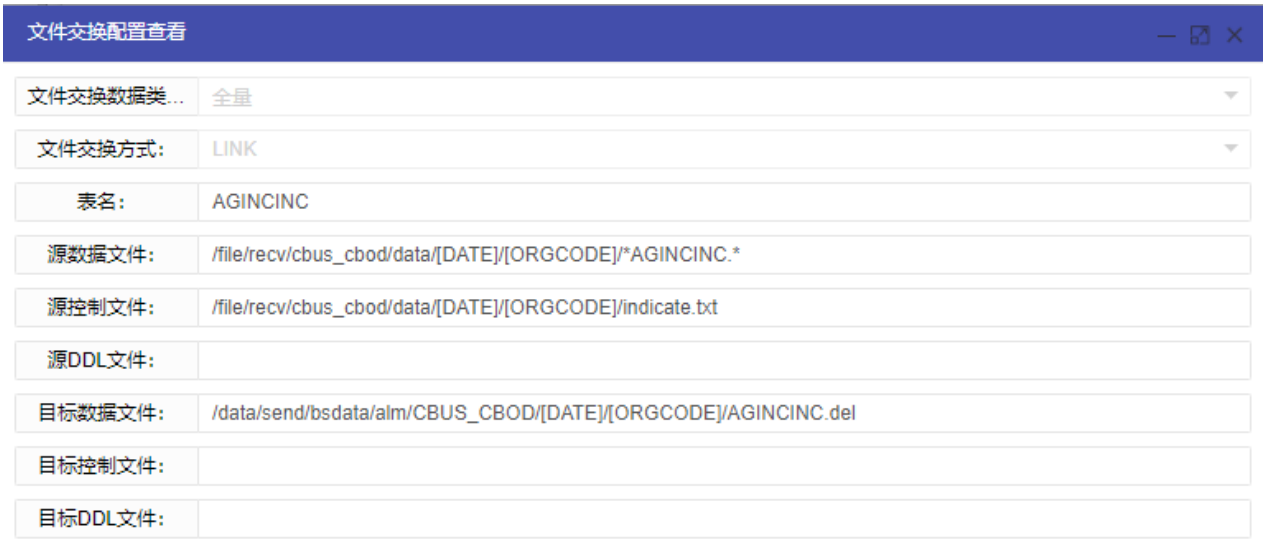
源控制文件：根据源 schema 的配置规范自动生成，可更改

源 DDL 文件：根据源 schema 的配置规范自动生成，可更改

目标数据文件：根据交换目标 schema 的配置规范自动生成，可更改，如果需要跟源文件名一致，可使用宏（#SRC_FILENAME#）

目标控制文件：根据交换目标 schema 的配置规范自动生成，可更改

目标 DDL 文件：根据交换目标 schema 的配置规范自动生成，可更改



3.10.2 数据供应

数据供应提供数仓库表对外供数的功能。

■ 供应配置

- 供应类型：日常、临时
- 供应数据类型：ADD/ALL
- 目标 schema：供应的目标 schema
- 表名：供应的目标表名、与源表一致
- 数据文件：根据供应目标 schema 规范自动生成，可更改
- 控制文件：根据供应目标 schema 规范自动生成，可更改
- DDL 文件：根据供应目标 schema 规范自动生成，可更改
- 过滤条件：从数据库进行数据供应时，可输入过滤条件

3.11 数据质量检核

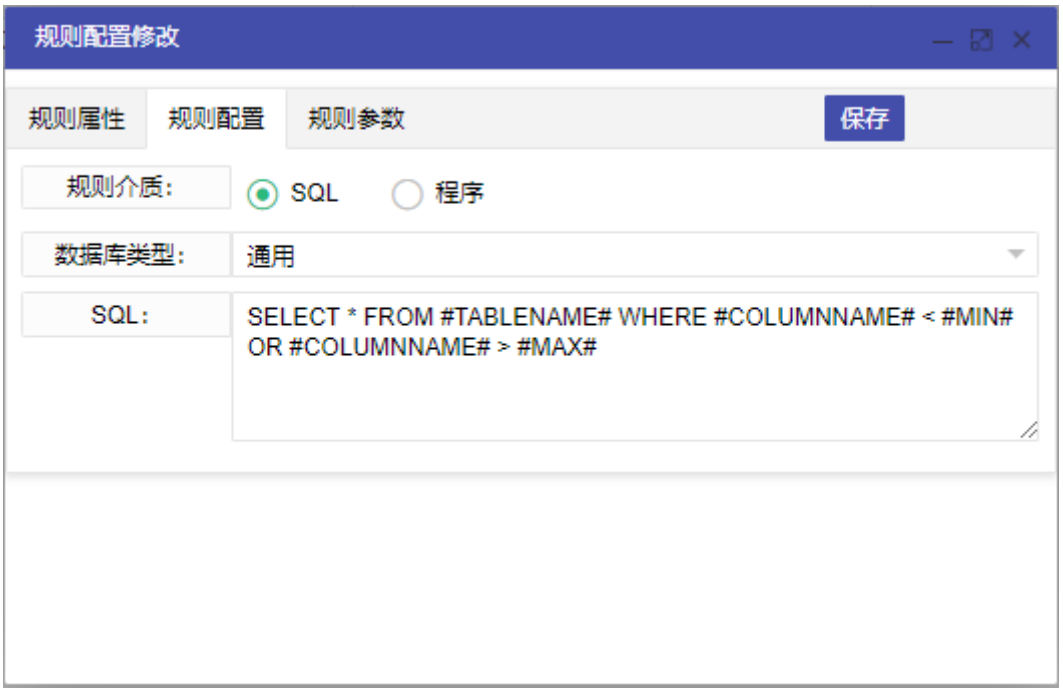
数据质量检核功能可以对数据应用各个阶段可能引发的各类数据质量问题，进行识别、度量、监控、预警等一系列管理活动，并通过改善和提高组织的管理水平使数据质量获得进一步提高。

3.11.1 检核规则

用于配置相应的检核规则。

■ 规则属性

- 检核编号：支持数字、字母、下划线
- 检核名称：检核规则的中文描述
- 检核对象：表、字段
- 适用范围：数据、文件
- 规则描述：该规则的描述信息



■ 规则配置

- 规则介质：SQL、程序
- 数据库类型：选择该规则支持的数据库类型
- SQL：规则引擎为 SQL 时，配置规则 SQL。SQL 中的参数变量格式为“#参数名#”，该功能默认提供两个宏变量：表名（#TABLENAME#）、字段名（#COLUMNNAME#），这两个不需要在参数中配置，其它参数则需要定义配置
- 执行引擎：规则介质为程序时，需要选择该引擎
- 程序路径：规则介质为程序时，需要配置程序的绝对路径

■ 规则参数

- 参数名：与规则 SQL 中的参数名一致，在规则 SQL 中用到的参数变量，皆需要在此处进行配置

3.11.2 检核管理

配置检核的任务项。点击“新增”，可增加检核的配置项。

3.11.3 任务监控

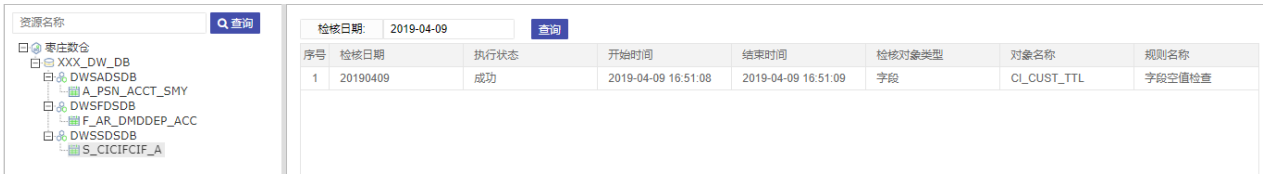
监控数据质量的运行情况。

■ 总体监控

支持按系统、检核日期两个维度，查看数据质量检核的总体运行情况。

■ 明细查看

点击“查看详情”，会跳转到检核的详细页面。



3.11.4 内置检核规则说明

3.11.4.1 当日数据不能为空

检查当日的数据记录数不能为空。

SQL：

SELECT * FROM #TABLENAME#　WHERE　#CHECK_DATE_COLUMN#='#DATE#'

参数：

CHECK_DATE_COLUMN ：　日期字段名

3.11.4.2 字段空值检查

检查字段的数据不能为空（NULL）或者空串（"）。

SQL：

SELECT * FROM #TABLENAME# WHERE #COLUMNNAME# IS NULL OR #COLUMNNAME# = ''

参数：

无

3.11.4.3 数值类型的取值范围

检查数值类型字段的取值范围。

SQL：

SELECT * FROM #TABLENAME# WHERE #COLUMNNAME# < #MIN# OR #COLUMNNAME# > #MAX#

参数：

MIN：最小值

MAX：最大值

3.12 数据生命周期管理

生命周期管理配置，用于管理仓库中的表和文件的存放周期，可根据仓库的设计规划对各个数据层的档进行生命周期管理。

生命周期管理包括架构管理员统一配置和用户开发定义配置两方面。架构管理员有权限对 schema 进行配置，配置后对 schema 下所有的数据对象有效；用户可以根据需要对数据对象进行单独配置。

配置步骤：

1. （架构管理员）生命周期调度配置（参见调度配置章节）
2. （架构管理员）调度日历配置（参见调度日历章节）
3. （开发人员）生命周期规则配置
4. （开发人员）生命周期清理策略配置
5. （开发人员）SCHEMA 的生命周期管理
6. 表的生命周期管理
   6.1（项目管理员）生命周期开发任务分配
   6.2（开发人员）生命周期任务配置
7. （开发人员）发布作业

3.12.1 生命周期规则配置

■ 新增规则

点击新增，配置规则。

规则名称：生命周期规则命名，建议根据规则的使用目标数据层命名，例如 R 层清理规则，标准层清理规则等。

规则描述：新增规则的描述信息。

■ 配置规则

在新增的规则上点击配置->右侧的界面中点击新增

日历规则：选择调度日历中配置的日历。

保存周期，周期单位：设置数据对象的保存周期，可设置为几天，几周，几月，几年，如果保存周期设置为-1 时，表示永久保存。

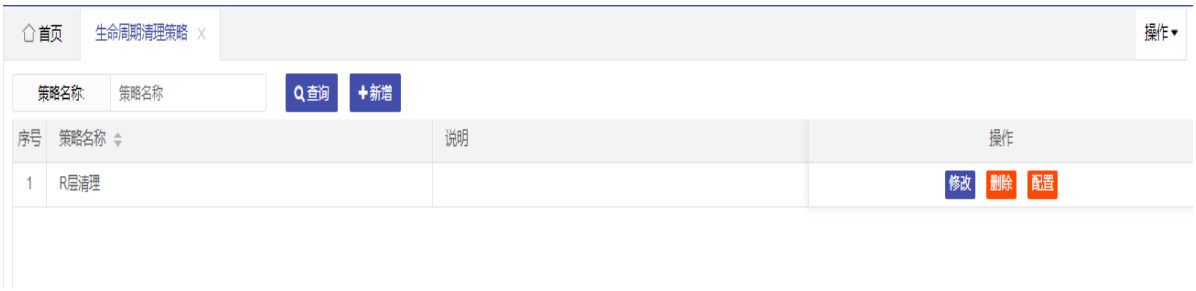
清理方式：设置规则的清理方式，有清理和归档可以选择，清理表示数据对象在周期外的将被直接清除掉，而归档表示数据对象在周期外将被归档然后清除（归档目标在生命周期管理处配置）。

注意：

i. **一个规则可以配置多个保存周期。**

ii. **最终的生命周期清理周期会根据日历和保存周期生成，例如：日历设置为每月月末执行（YYYYMMEN），保存周期设置为 1 天，则表示保存每个月月末那天的数据。**

3.12.2 生命周期清理策略



■ 新增清理策略

点击新增，设置策略名。

策略名称：设置生命周期清理策略名称，建议根据规则的使用目标数据层命名，例如 R 层清理策略，标准层清理策略等。

说明：对新增的策略进行描述。

■ 配置策略

在新增的策略上点击配置->在弹出的界面中点击新增进行策略配置。

表类型：根据待清理的数据对象类型进行选择，有当前，拉链，增量切片，全量切片，当前表日表（一天一个当前表）。

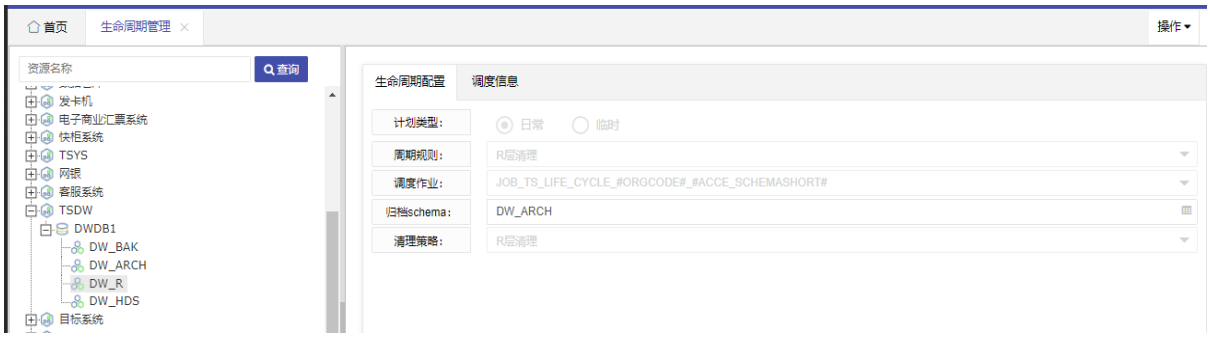
日期字段：根据表类型设置清理用的日期字段，当前表日表可以不设置，因为此

类型的表每天一张。

自定义 sql：设置清理的的 sql，设置后就不根据日期字段做清理，而是根据设置的 sql 进行清理。

3.12.3 生命周期管理

对 SCHEMA 进行生命周期设置。



■ 左侧菜单选择要设置的 SCHEMA->点击此 SCHEMA 右边菜单会显示配置界面

计划类型：设置生命周期管理程序的执行计划，日常或者临时。

周期规则：选择生命周期规则。

调度作业：选择生命周期作业（在调度配置处新增，详细操作参考调度配置）。

归档 SCHEMA：如果周期规则中设置了归档，那么此选择栏会出现，否则没有；在此选择归档的目标 SCHEMA。

清理策略：选择生命周期清理策略。

调度信息：根据需要调整。

3.12.4 生命周期任务

对特定的数据对象进行生命周期设置。

注意：配置前，数据对象所属的 SCHEMA 需要进行生命周期设置，否则无法单独配置数据对象的生命周期任务。

■ 左侧菜单选择要配置的数据对象->点击右侧出现配置菜单（默认会继承 SCHEMA

的生命周期设置）。

周期规则：生命周期规则设置。

归档 SCHEMA：如果周期规则中设置了归档，那么此选择栏会出现，否则没有；在此选择归档的目标 SCHEMA。

清理策略：选择生命周期清理策略，可以设置自定义 SQL（**如果设置了自定义 SQL，那么周期规则那里配置的将无效，会以此处配置的 SQL 为准**）。

3.13 调研信息管理

调研信息管理提供调研相关信息的导入功能。

- 导入

  提供模板内容的导入功能，导入的模板应通过模板下载的指定模板格式。导入时需指定导入 sheet 页的名称。导入时支持追加导入(分批导入)和清空导入两种。

3.14 模型管理

3.14.1 模型设计结果导入

提供模型设计结果的信息导入功能。

3.14.2 物理模型管理

提供对模型设计进行物理化的功能。

- 映射转换配置：配置模型设计表和物理

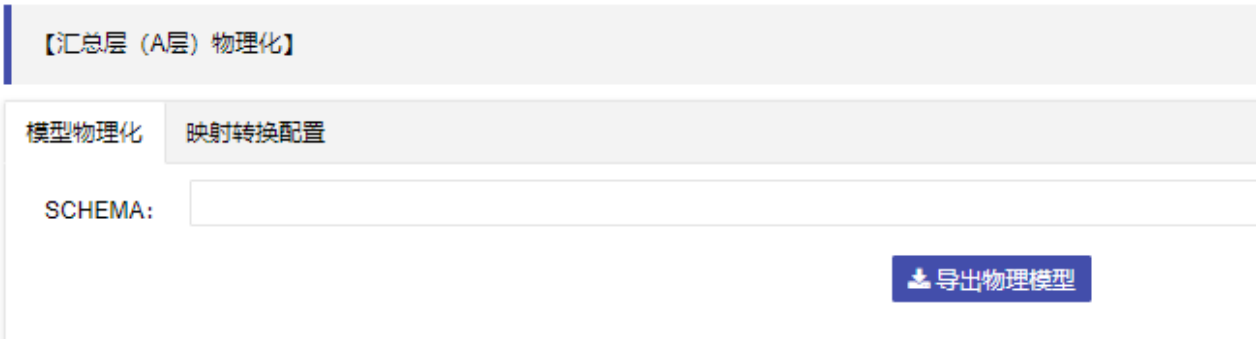
视图名称：视图的名称（为下面 SQL 定义的视图名一致）

SQL 内容：视图的定义，应包含以下字段字段

序号	字段名	字段描述	字段类型	可空	主键	默认值	备注
1	TABLE_ID	表ID	VARCHAR(50)	否	是		
2	TABLE_NAME	表名	VARCHAR(100)	否			
3	TABLE_NAME_CN	表中文名	VARCHAR(200)				
4	BUCKET_NUM	分桶数量	INT				
5	BUSI_TYPE	业务类型	VARCHAR(2)				01:主档 02:流水档
6	STORAGE_TYPE	存储类型	VARCHAR(2)				01:当前 02:拉链 03:增量切片 04:全量切片
7	DESCRIPTION	表说明	VARCHAR(500)				
8	COL_SEQ	序号	INT	否			
9	COL_NAME	字段英文名	VARCHAR(100)	否			
10	COL_DESC	字段中文描述	VARCHAR(200)				
11	COL_TYPE	字段类型	VARCHAR(50)	否			
12	COL_LENGTH	字段长度	INT	否			
13	COL_SCALE	小数位数	INT	否			
14	COL_DEFAULT	默认值	VARCHAR(200)				
15	NULL_FLAG	是否可空	VARCHAR(1)				0:否，1:是
16	PK_FLAG	是否主键	VARCHAR(1)				0:否，1:是
17	PARTITION_FLAG	是否分布键	VARCHAR(1)				0:否，1:是
18	BUCKET_FLAG	是否分桶键	VARCHAR(1)				0:否，1:是
19	DESCRIPTION	描述	VARCHAR(300)				

■ 模型物理化

选择物理化的目标 SCHEMA，可以导出平台数据字典格式的 EXCEL 文件。



4 教程示例

本章节通过一些示例，为用户的使用提供指导，在使用中若遇到问题，请联系 XXXXXX

4.1 创建数据项目

新建数据项目的流程：

1. （系统管理员）创建项目；
2. （系统管理员）指定项目人员；
3. （架构管理员）配置项目规范；
4. （架构管理员）配置数仓贴源数据区；
5. （架构管理员）配置历史归档数据区

4.1.1 创建项目

架构管理员登录创建数据项目。

➢ 操作步骤：

1. 系统管理员登录；
2. 新增项目，依次点击“项目管理 -> 项目管理 -> 新增”；
3. 点击“保存”按钮。

4.1.2 指定项目开发人员

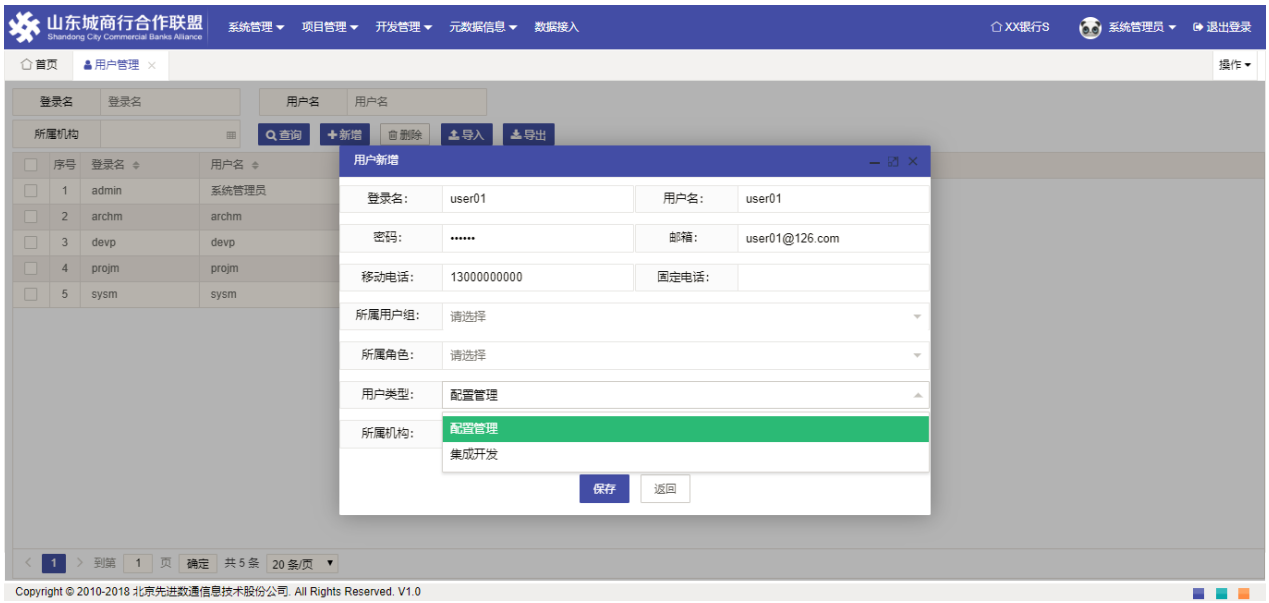
为项目配置可以操作该项目的用户。

➢ 操作步骤：

1. 架构管理员登录；
2. 指定项目权限用户，依次点击“项目管理->项目管理->权限”，若需要新增开发人员，参考下面的用户新增；
3. 关闭后保存。

➢ 创建用户操作步骤：

1. 系统管理员登录；
2. 新增用户，依次点击“系统管理->用户管理->新增”；



■ 用户类型：

配置管理：系统配置用户（系统管理员、架构管理员、项目管理员）

集成开发：需要对项目进行操作的用户（开发人员）

■ 所属机构：

选择“枣庄银行”。

4.1.3 配置项目规范

指定项目的规范，例如采集文件的字符编码、分隔符、换行符、采集目标、预处理目标位置。

采集目标和预处理目标需要先配置相应的文件类型的 schema，才能配置到该位置。

➢ 操作步骤：

1. 架构管理员登录；
2. 创建采集目标位置：创建系统（DW 系统）-> 创建数据源 -> 配置 schema：

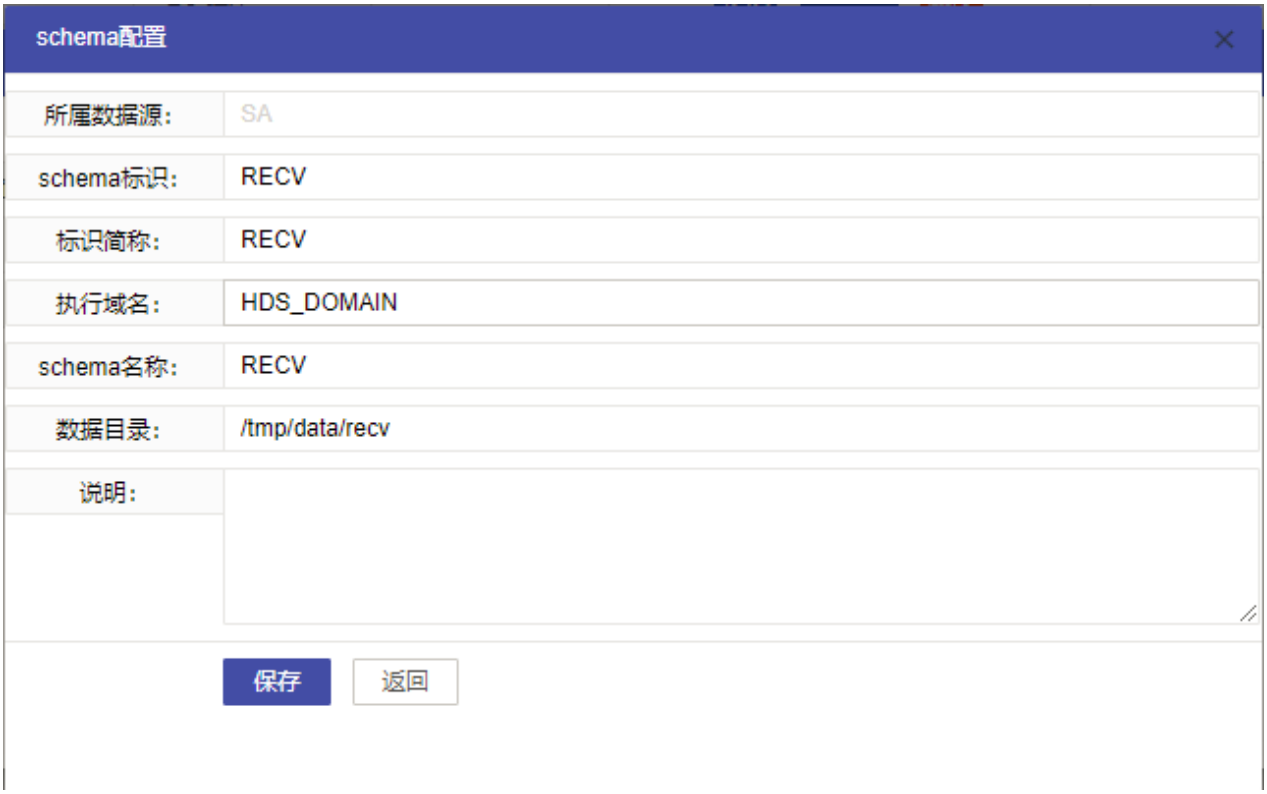
a. 创建系统：依次点击“开发管理->系统信息维护->新增”；

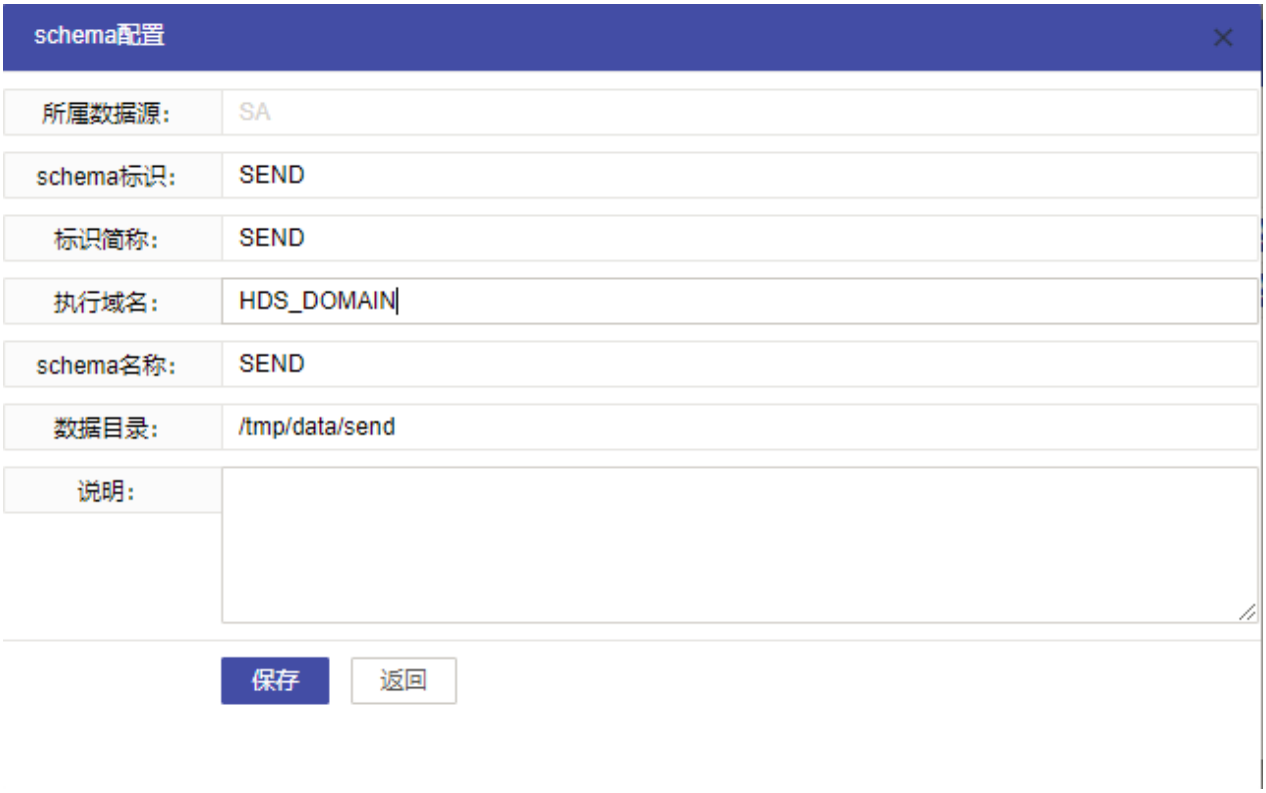
b. 配置系统规范，具体的规范根据实际情况配置：

在配置时包含平台内部宏和调度宏两种，平台内部宏在生成作业时，会动态解析生成；调度平台宏，在作业实例运行时动态替换，支持的宏参考《5.1 宏定义说明》

c. 创建数据源，依次点击“开发管理->数据源->新增”

d. 配置 schema（配置 RECV 和 MIDDLE），schema 标识名称可自定义





3. 项目元数据分配

   参考 4.1.6

4. 项目配置

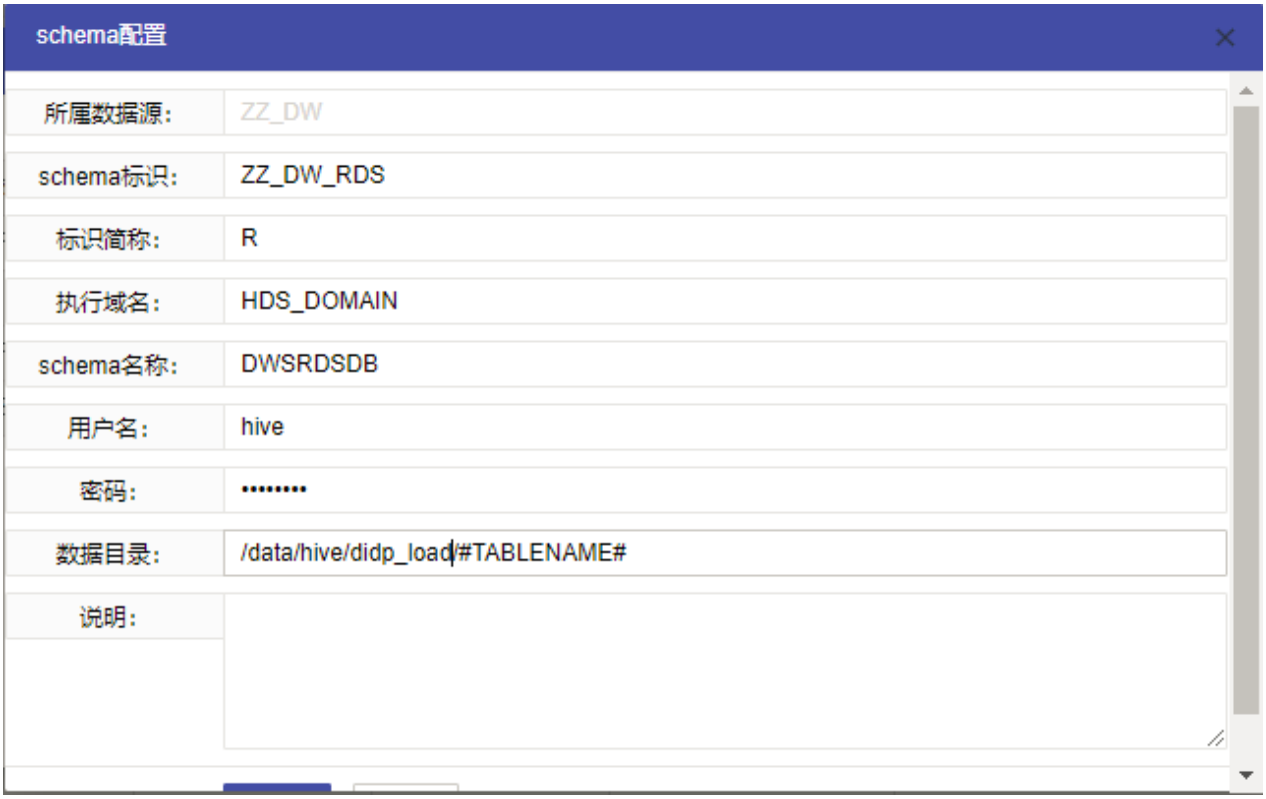
4.1.4 配置近源数据区

架构管理员登录，创建数据项目的近源数据区 schema，供开发人员在数据接入-数据加载中指定目标位置使用。

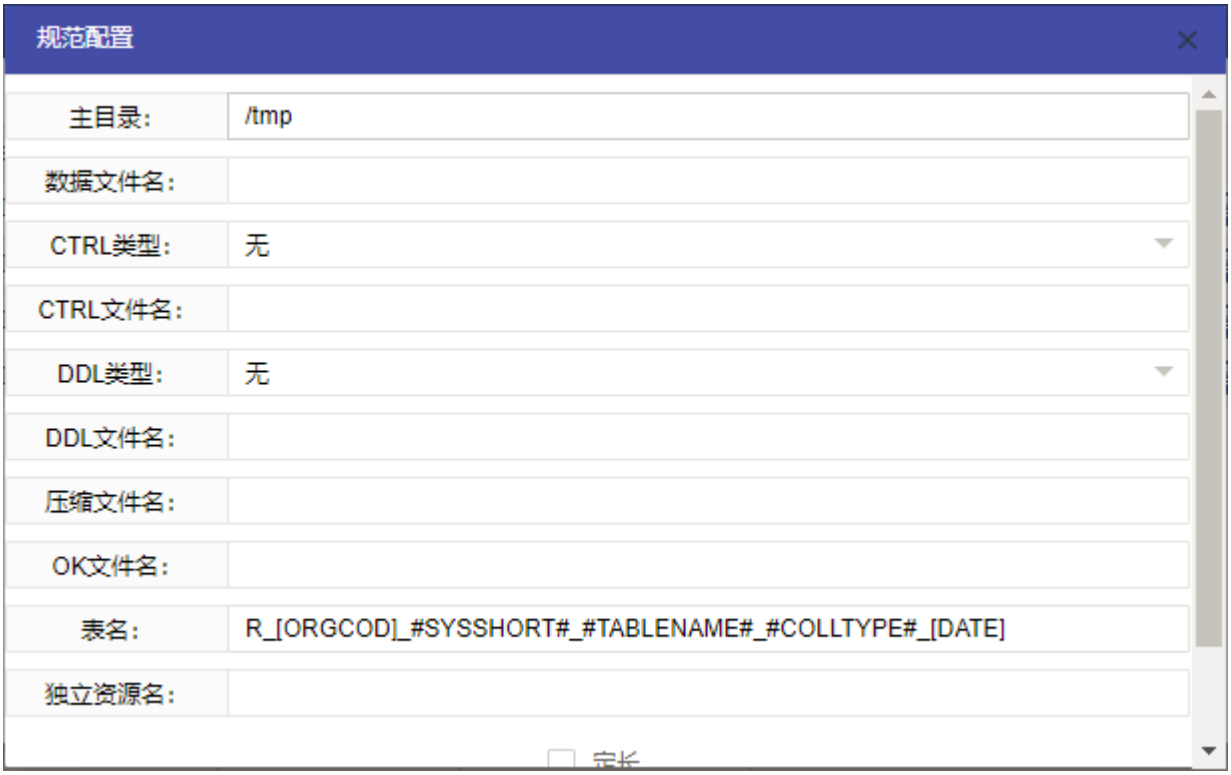
➢ 操作步骤：

1. 架构管理员登录；
2. 增加数据源，依次点击“开发管理 -> 数据源 -> 新增”；

3. 点击“保存”；
4. 配置 schema；



5. 配置 schema 规范，这里主要配置加载层的表名规范；



6. 点击“保存”。

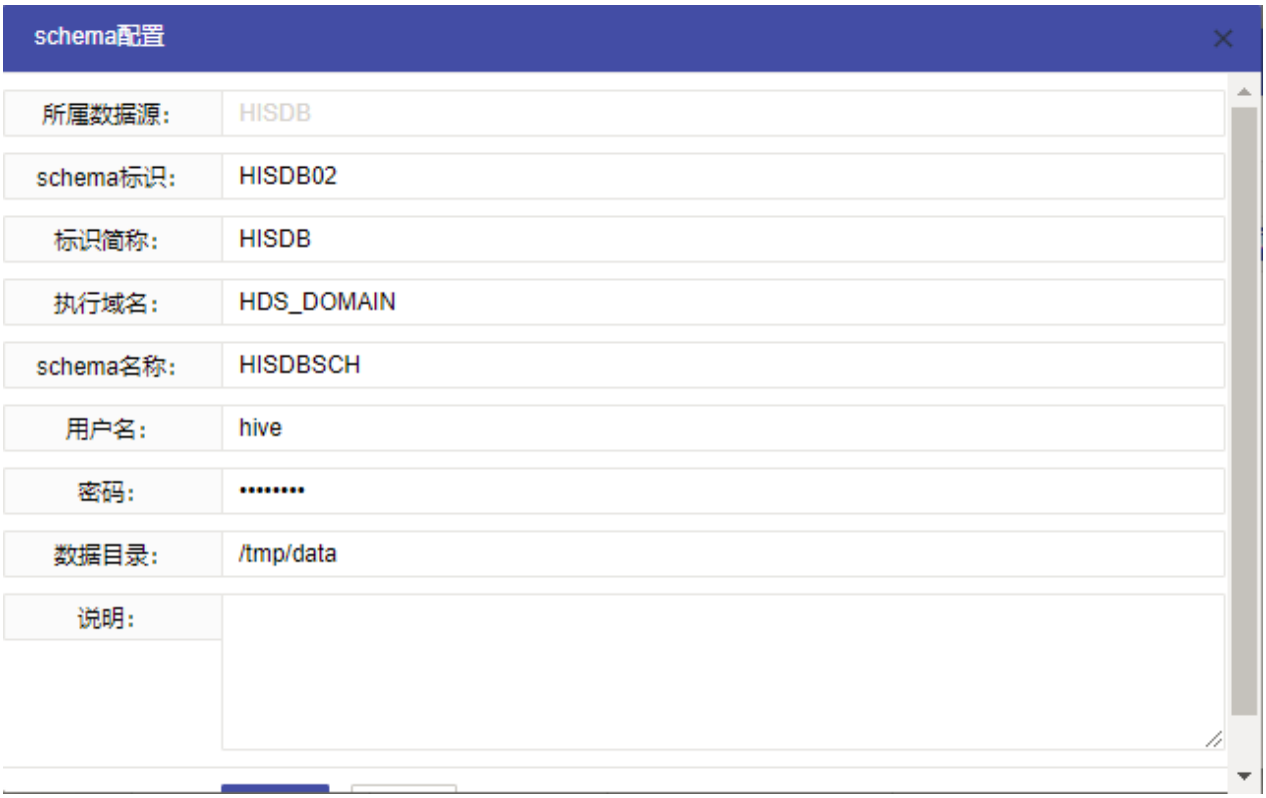
4.1.5 配置历史归档数据区

架构管理员登录，创建历史归档数据区的数据源，供开发人员在数据接入-数据归档中指定目标位置使用。

➢ 操作步骤：

1. 架构管理员登录；
2. 新增项目，依次点击“开发管理 -> 数据源 -> 新增”；
3. 点击“保存”按钮。

4. 配置 schema



5. 配置规范；

6. 点击“保存”。

4.1.6 项目元数据分配

➢ 操作步骤：

1. 架构管理员登录；
2. 进入元数据分配页面：项目管理 ->项目元数据分配；
3. 选择左侧的项目 -> 从未分配中选择（RECV、MIDD、近源数据区、归档数据区）-> “控制”

4.2 数据接入

操作流程：

1. （架构管理员）配置调度规范、为项目分配调度计划；
2. （架构管理员）创建源系统 schema（源系统、数据源、schema）；
3. （项目管理员）导入数据字典；
4. （项目管理员）分配任务；
5. （开发人员）接入（采集、预处理、加载）、归档任务开发，分为批量导入和页面两种。

4.2.1 配置调度规范

4.2.1.1 调度配置

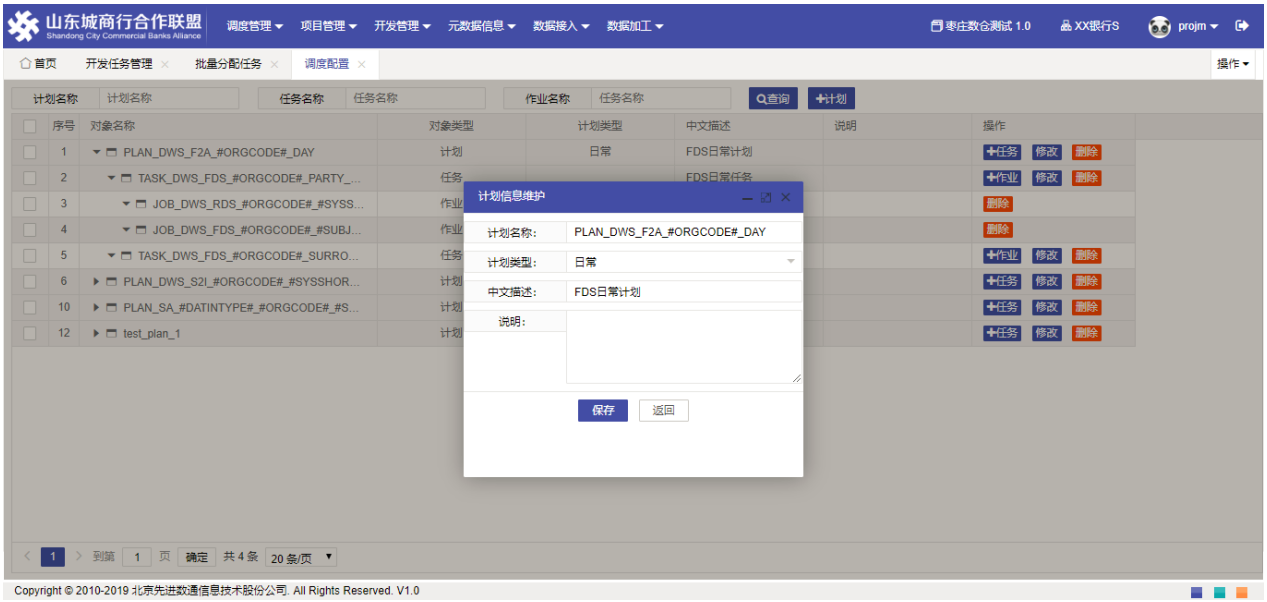
调度配置的规范中的宏可参考章节《5.1 宏定义说明》

调度配置流程：

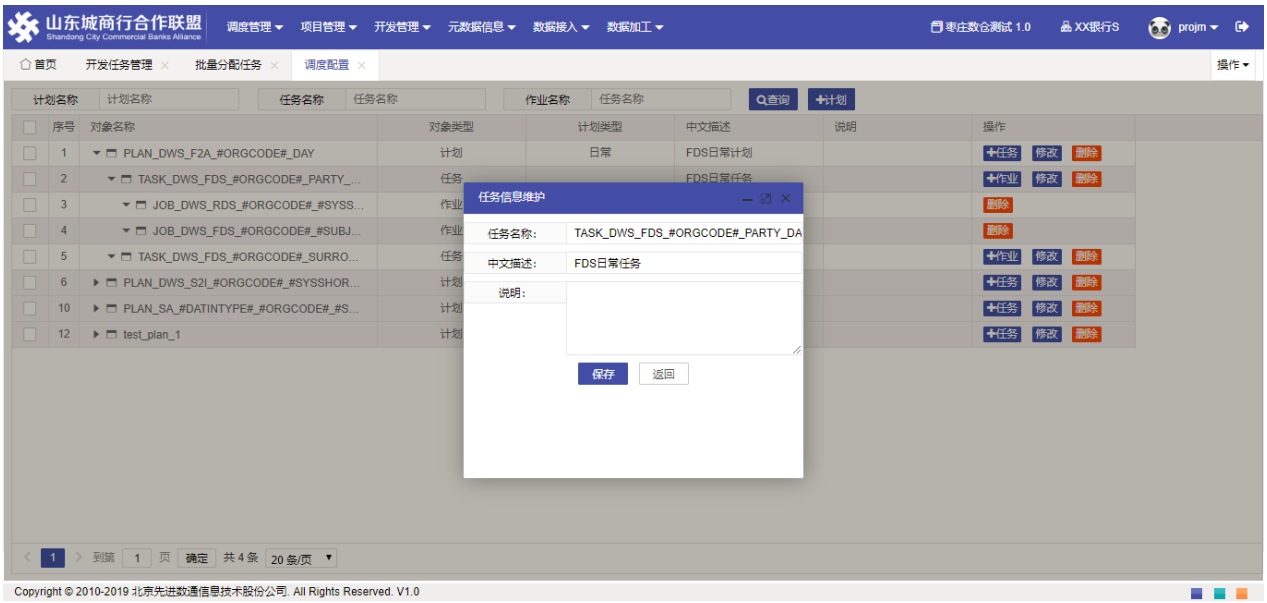
1. 新增计划
2. 新增任务
3. 新增作业

注：调度的计划、任务、作业命名按照数仓要求来进行配置。

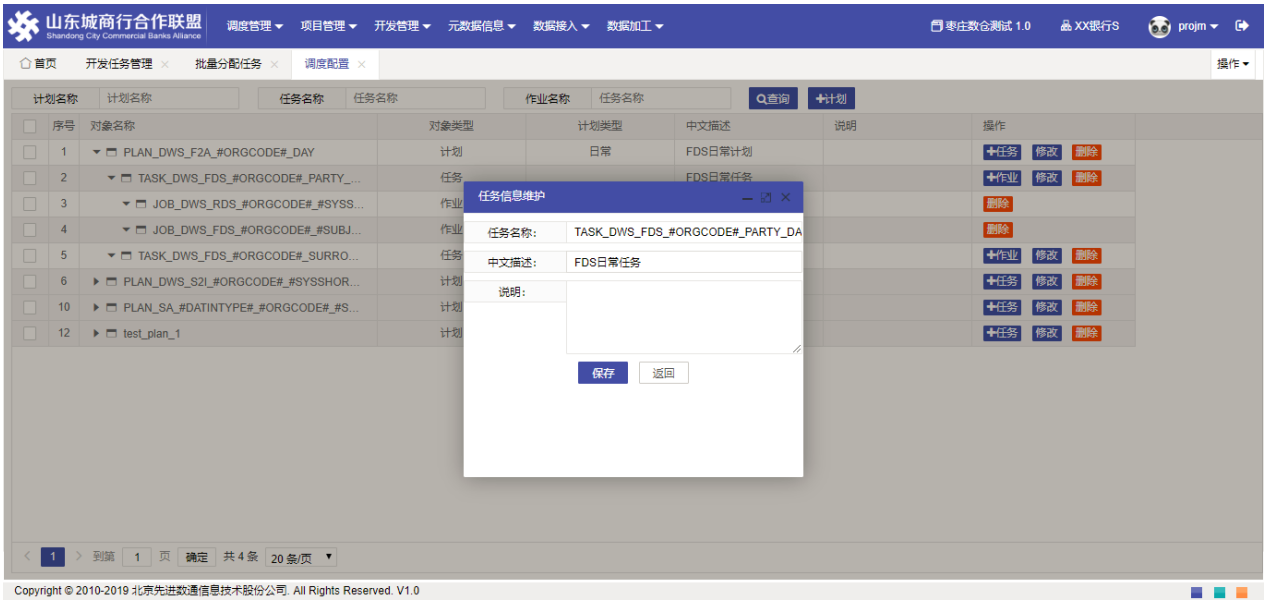
4.2.1.1.1 新增计划



4.2.1.1.2 新增任务



4.2.1.1.3 新增作业



4.2.1.2 调度分配

调度分配流程：

1. 进入项目的分配功能
2. 挂载调度对象

4.2.2 新增 SCHEMA

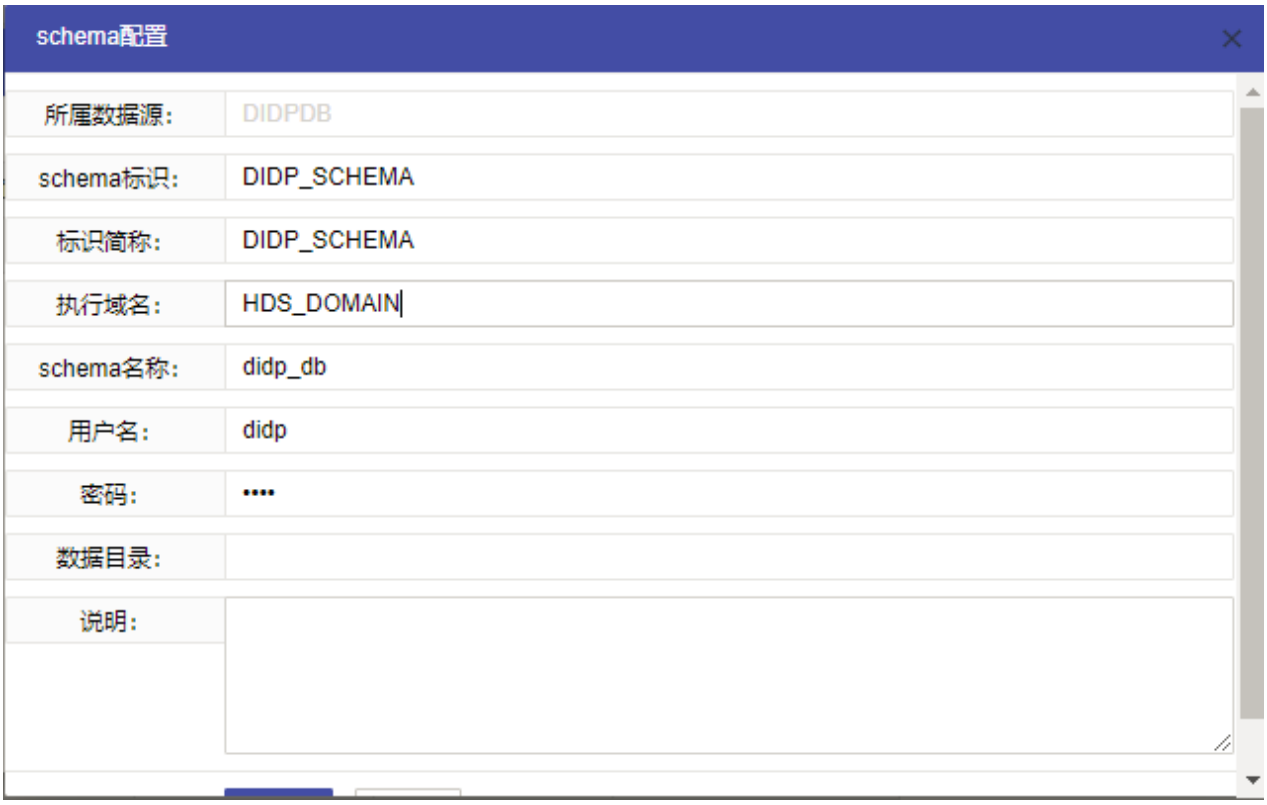
在导入元数据之前，需要新增 SCHEMA 信息，添加 SCHEMA 时需要依次添加系统、数据源、schema。

新增 Schema 流程：

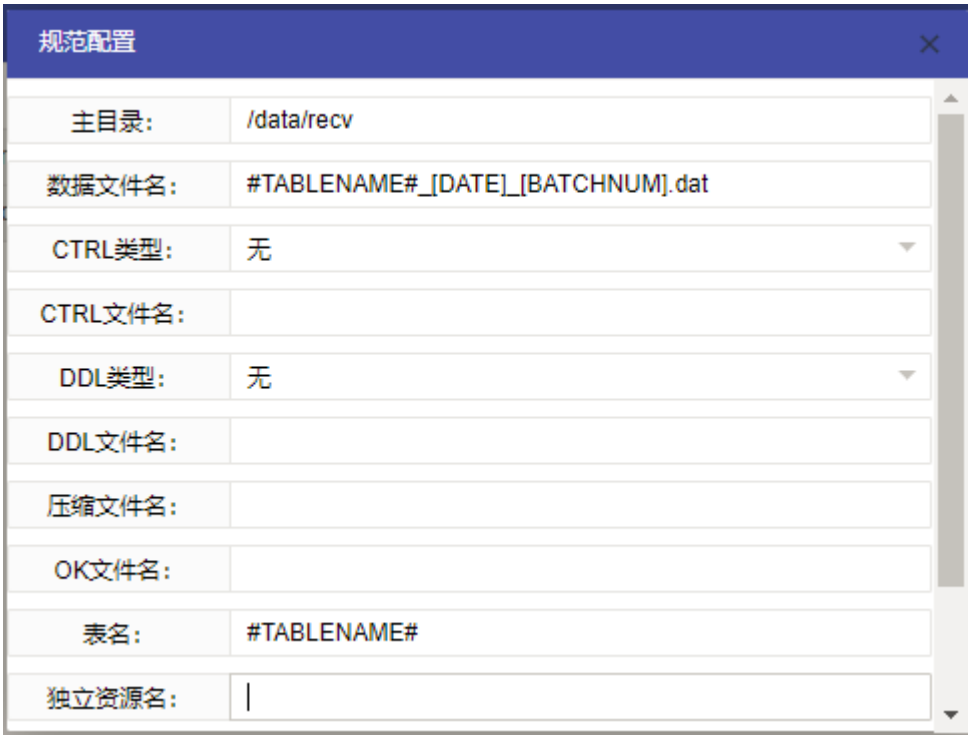
1. 创建源系统，进入“开发管理-》系统信息维护-》新增”，系统简称参考系统命名规范；

2. 创建数据源，进入“开发管理-》数据源-》新增”；

3. 创建 SCHEMA，在数据源配置行的操作列中选择“schema”；



4. 配置 schema 规范，如果来源是文件类型需要配置相应的文件命名规范、控制文件规范、DDL 文件规范、压缩文件等。



4.2.2.1 源系统配置

架构管理员登录创建接入的源系统。

➢ 操作步骤：

1. 架构管理员登录；
2. 新增项目，依次点击“开发管理 -> 系统信息维护 -> 新增”；
3. 点击“保存”按钮。

4. 配置系统规范。

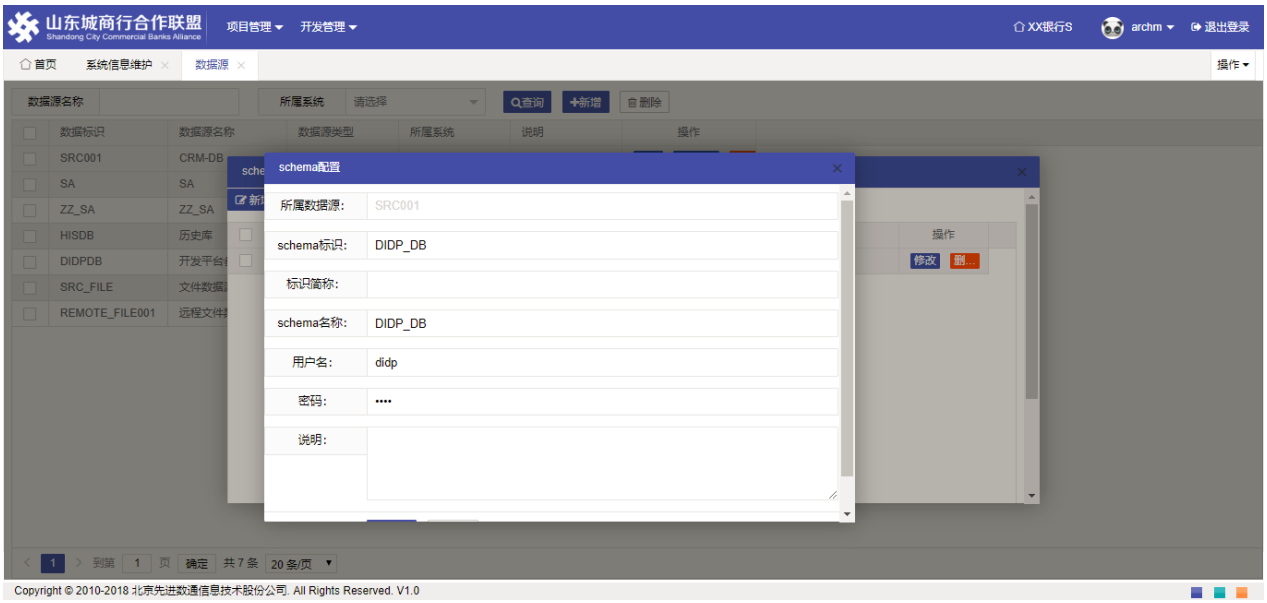
4.2.2.2 数据源配置

架构管理员登录创建源系统下的数据源。

➢ 操作步骤：

1. 架构管理员登录；
2. 新增项目，依次点击“开发管理 -> 数据源 -> 新增”；
3. 点击“保存”按钮。

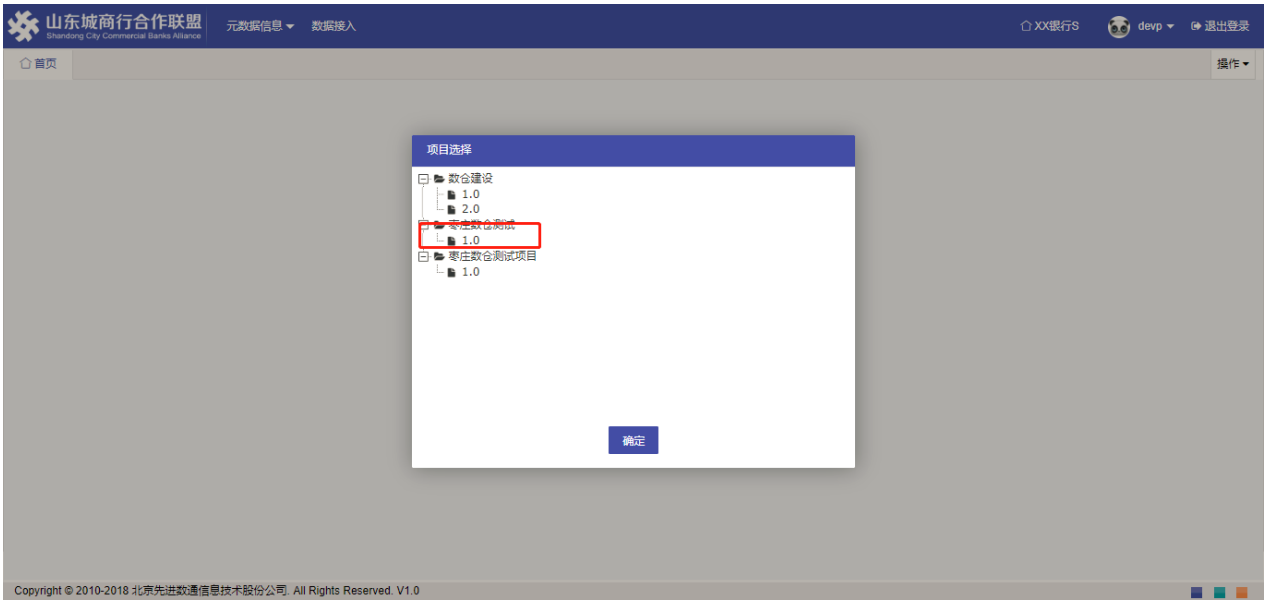
4.2.2.3 Schema 配置



4.2.3 导入数据字典

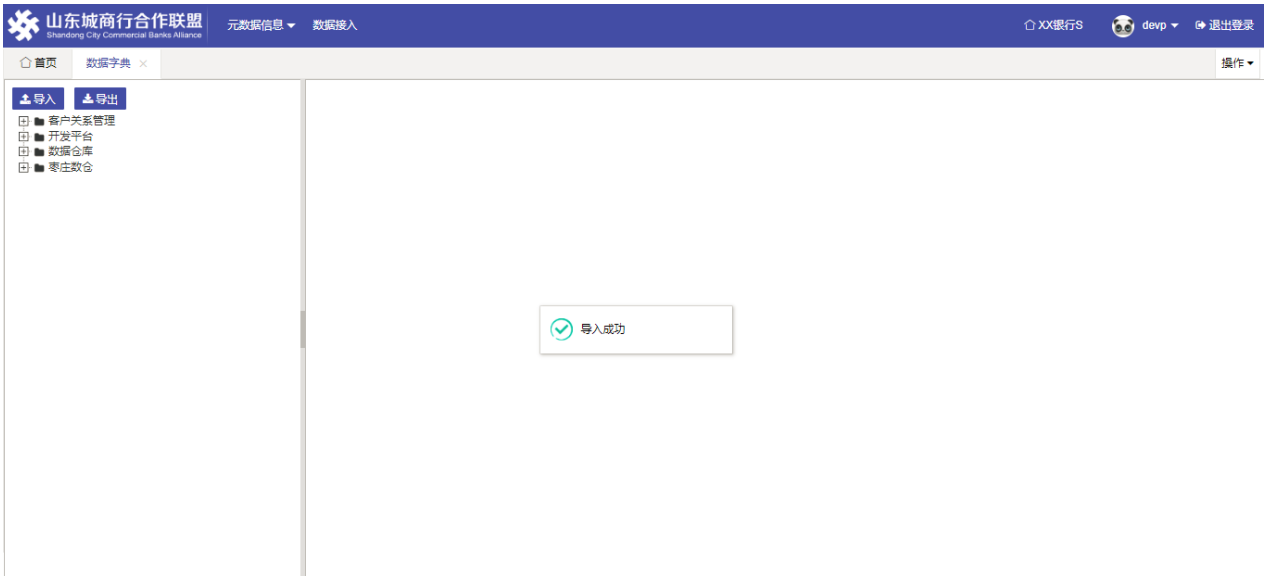
➢ 操作步骤：

1. 开发人员或项目管理员登录，选择项目及项目版本；



2. 导出模板，填写，依次点击“元数据信息 -> 数据字典 -> 导入”，然后点击“模板下载”，模板录入；

3. 导入模板，点击保存。提示“导入成功”，则表示数据字段导入成功；



参考示例：

数据字典_didp_project_version_info

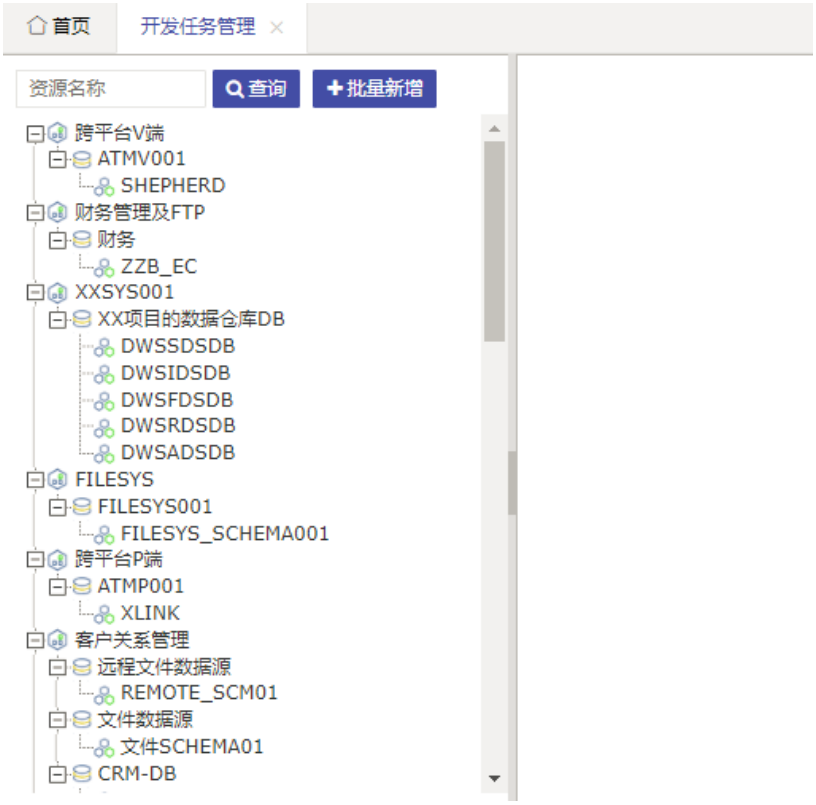
4. 可以查看刚导入档的信息。

4.2.4 任务分配示例

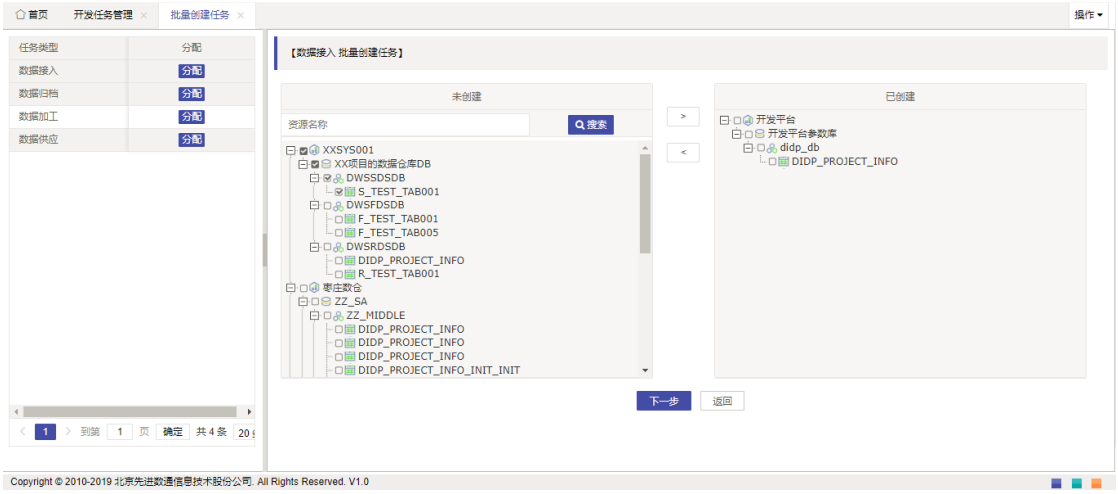
项目管理员登录，选定项目，开始分配任务。

任务分配的流程：

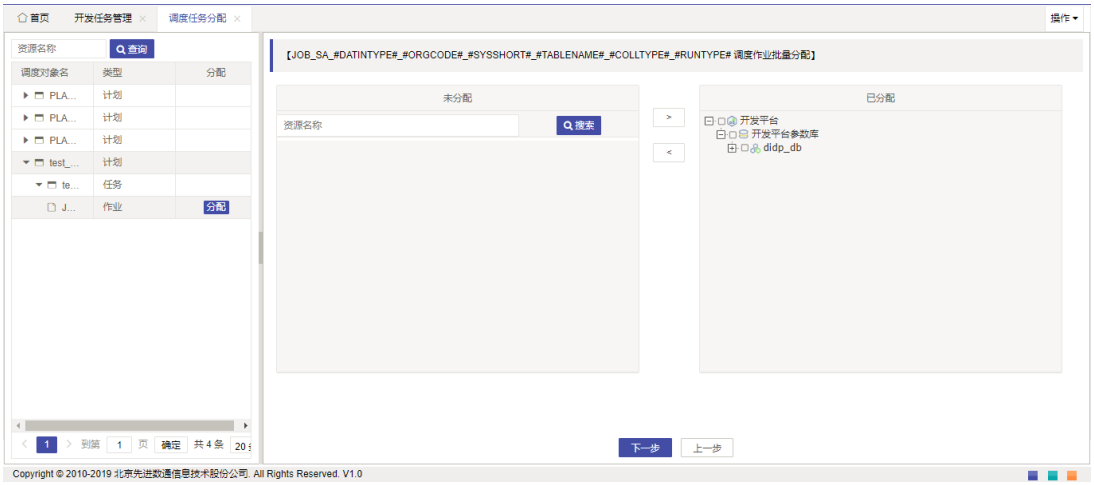
1. 依次点击“项目管理-》开发任务管理-》批量新增”



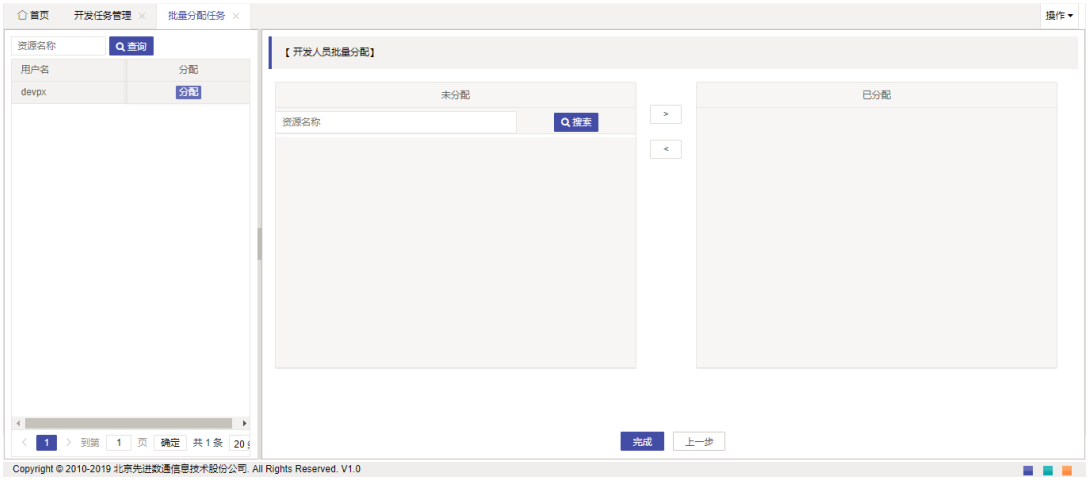
2. 在任务类型（数据接入）后，选择“分配”按钮，选择需要分配的对象（表），在左侧未创建列表选择需要分配的表，点击“>”，移动到右侧“已创建”列表中，此时选择数据对象完成；



3. 在调度对象—作业后，选择“分配”按钮，在左侧未创建列表选择需要分配的表，点击“>”，移动到右侧“已创建”列表中，此时选择数据对象完成；



4. 选择分配的开发人员



以上例子演示了分配数据接入（采集、预处理、加载）的任务，分配归档的任务依次执行 2（选择归档任务分配）、3、4。

4.2.5 数据接入示例

4.2.5.1 数据接入-选择作业类型

采集数据源的表。

➢ 操作步骤：

1. 配置接入操作，在档名上右键，配置接入类型。

4.2.5.2 数据接入-数据采集

➢ 操作步骤：

1. 配置采集条件，双击“数据采集”，弹出配置页面，数据采集可配置过滤条件。

2. 配置调度参数，域节点配置。

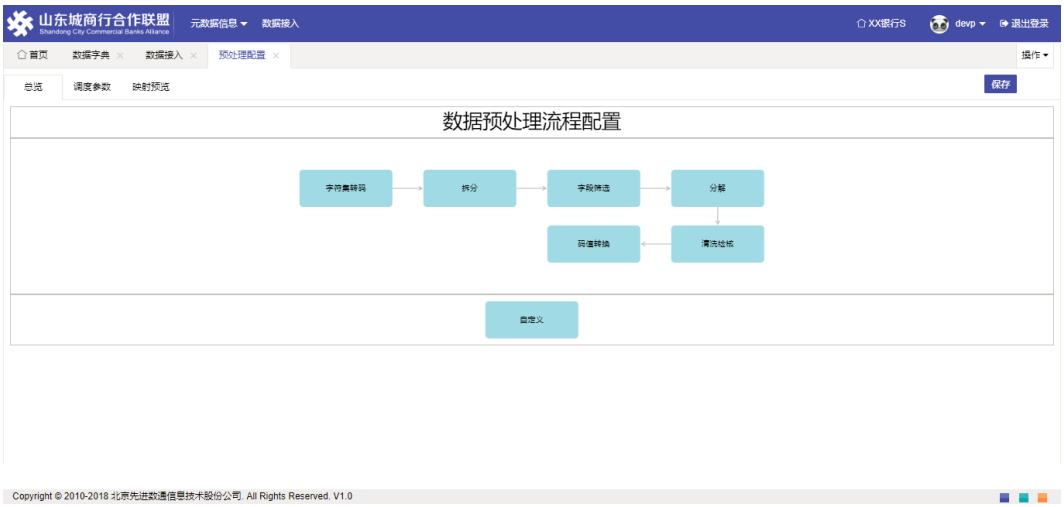
3. 保存->提交审核->发布作业。

说明：此处的提交审核是由项目管理员审核，目前暂未开发。

4.2.5.3 数据接入-预处理流程

➢ 操作步骤：

1. 进入预处理流程页面，双击“预处理流程”，跳转到预处理流程配置页面，开发人员可根据需求进行配置，详细配置参考《3.6.1.2 数据预处理》；



2. 配置调度参数，域节点配置；

3. 点击“保存”按钮，保存配置；
4. 保存->提交审核->发布作业。

说明：此处的提交审核是由项目管理员审核，目前暂未开发。

4.2.5.4 数据接入-数据装载

➢ 操作步骤：

1. 配置装载操作，双击“数据装载”，弹出配置页面，选择装载模式和目标数据 schema；

2. 配置调度参数，域节点配置。

3. 保存->提交审核->发布作业。

说明：此处的提交审核是由项目管理员审核，目前暂未开发。

4.2.6 数据归档

操作流程：

1. 在表名上右键选择归档类型

2. 归档配置，根据实际需求配置。

4.2.7 批量功能

批量操作流程：

1. 项目管理员分配任务；
2. 开发人员下载模板；
3. 开发人员开发模板；
4. 开发人员导入模板。

4.2.7.1 数据接入

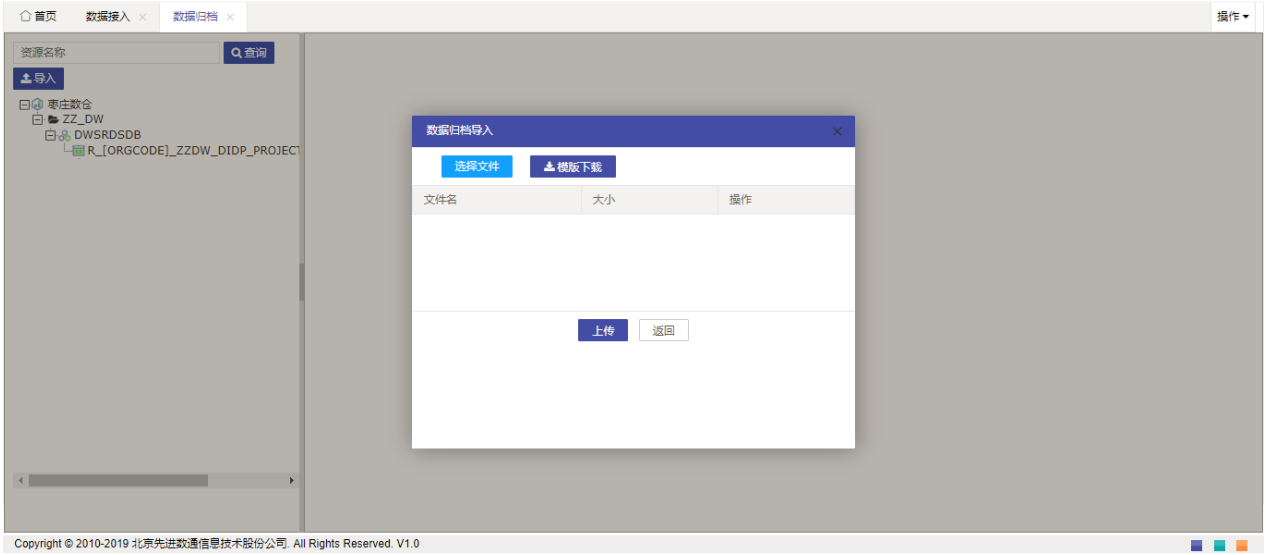
批量的操作流程：

点击“导入”，从弹出框中点“模板下载”，下载接入模板。

4.2.7.2 数据归档

批量的操作流程：

点击“导入”，从弹出框中点“模板下载”，下载接入模板。



4.3 数据生命周期管理

操作流程：

8. （架构管理员）生命周期调度配置
9. （架构管理员）调度日历配置
10. （开发人员）生命周期规则配置
11. （开发人员）生命周期清理策略配置
12. （开发人员）SCHEMA 的生命周期管理
13. 表的生命周期管理
    - 13.1 （项目管理员）生命周期开发任务分配
    - 13.2 （开发人员）生命周期任务配置
14. 开发人员）发布作业

4.3.1 生命周期调度配置

架构管理员登陆操作。

调度配置的规范中的宏可参考章节《5.1 宏定义说明》

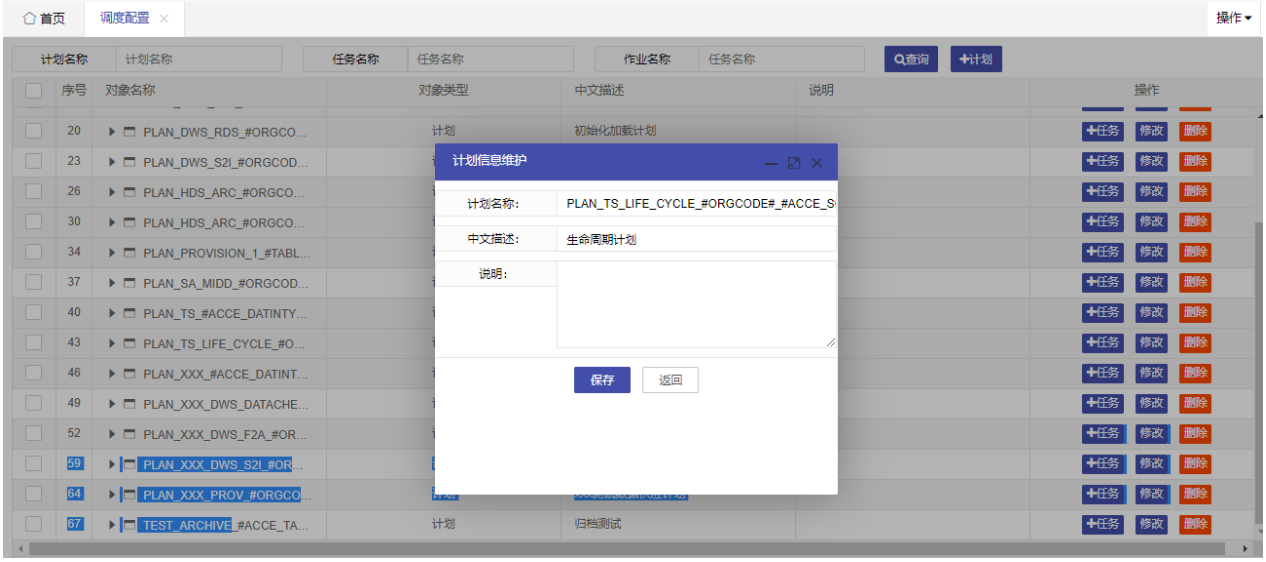
调度配置流程：

4. 新增生命周期计划

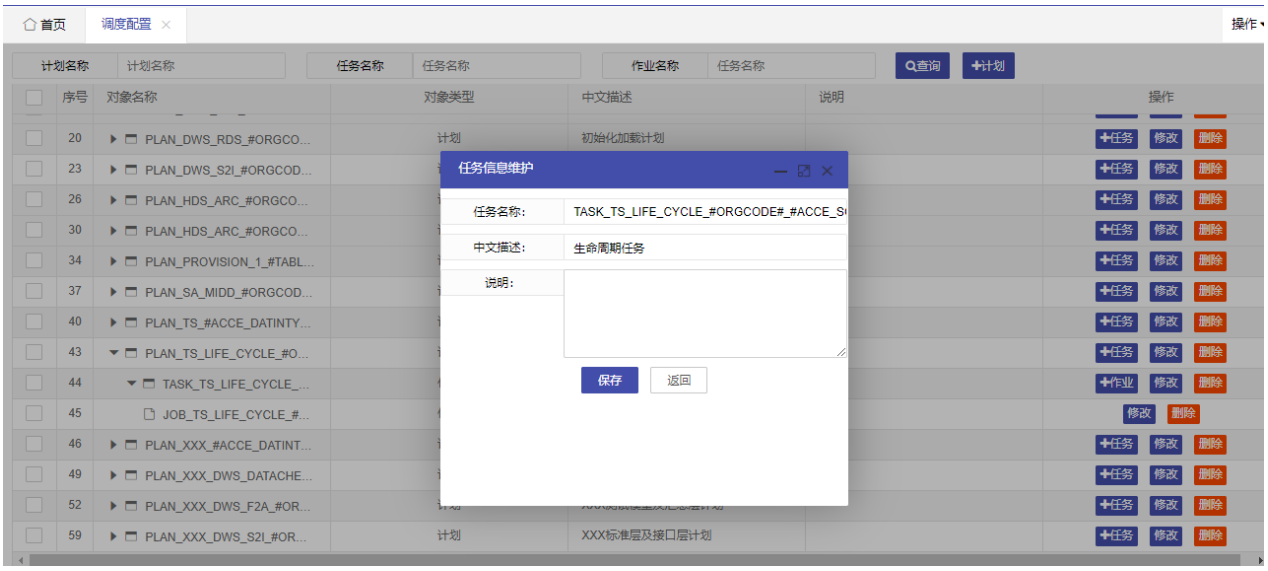
5. 新增生命周期任务
6. 新增生命周期作业

注：调度的计划、任务、作业命名按照数仓要求来进行配置。

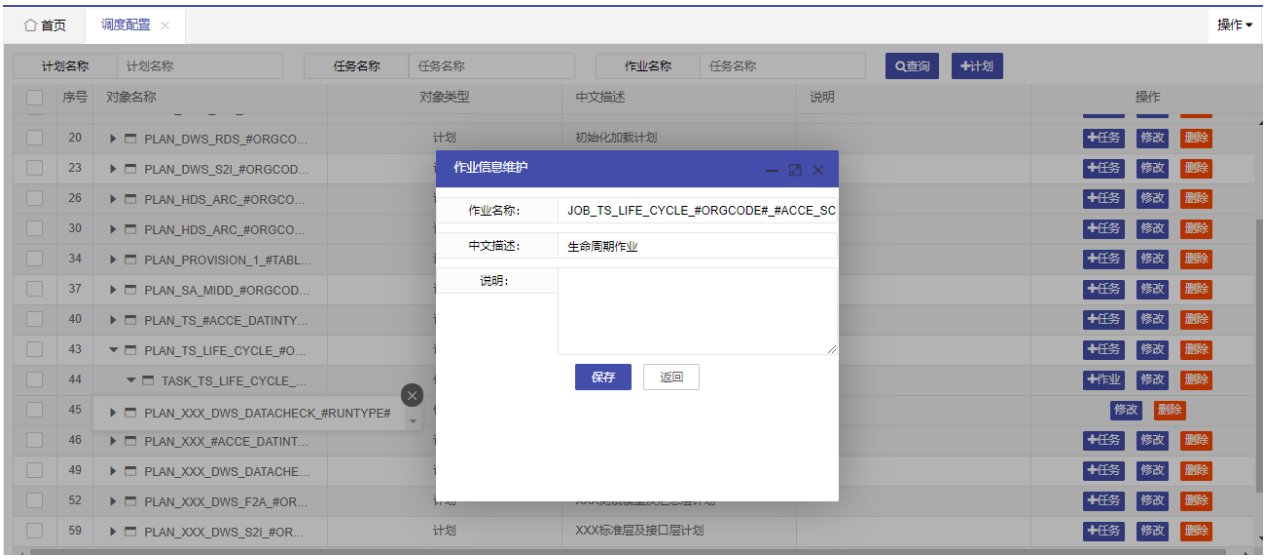
4.3.1.1 新增计划



4.3.1.2 新增任务



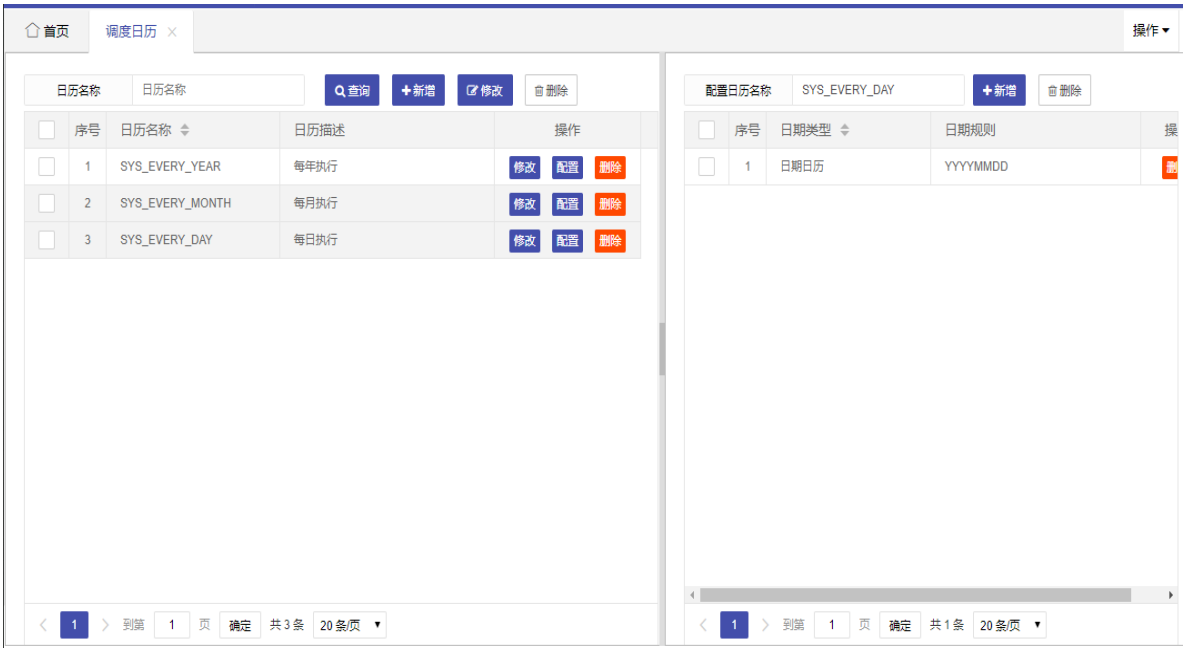
4.3.1.3 新增作业



4.3.2 调度日历配置

架构管理员登陆操作。

1. 依次点击“调度管理-》调度日历”

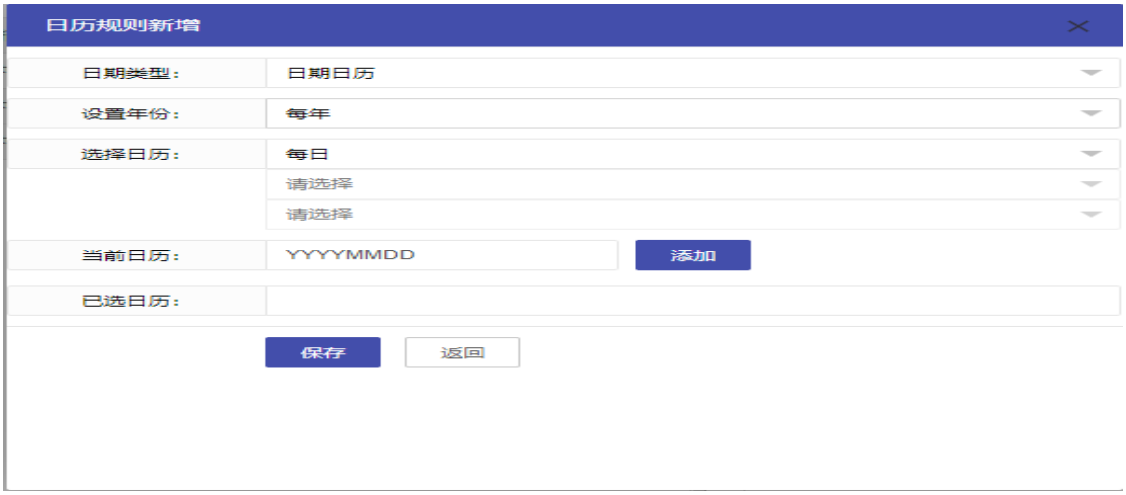


2. 日历规则新增

点击左侧新增按钮，在弹出的界面中输入日历名称和日历描述，保存规则。

3. 配置日历规则

点击上一步新增的日历规则上的配置按钮，在右侧的界面中点击新增配置日历规则。



根据需要设定日期类型，年份和选择日历。

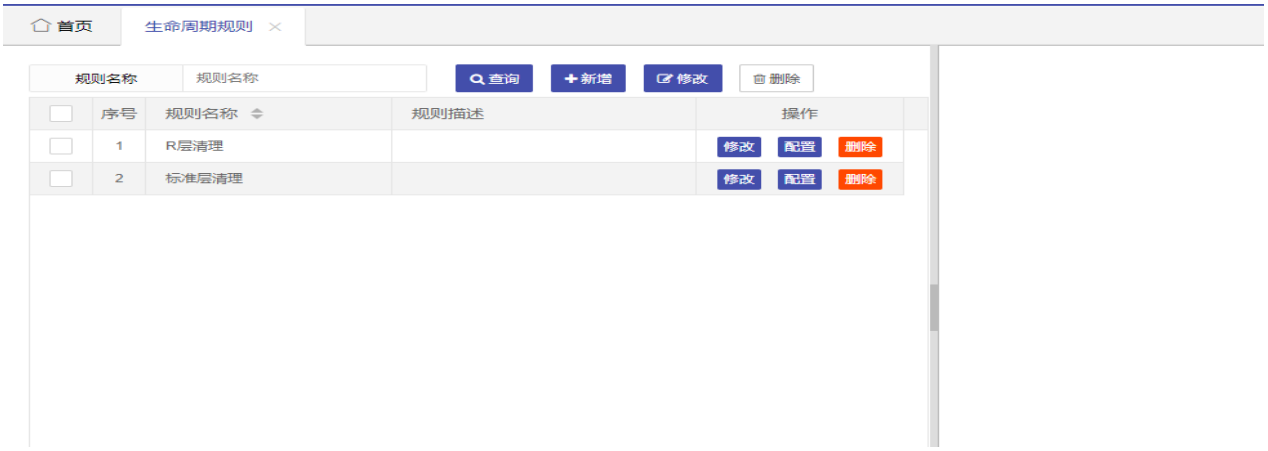
设置好日历之后点击添加，会在已选日历栏中自动生成添加的日历，然后保存日历规则。

注意：相同的日历规则不要重复添加。

4.3.3 生命周期规则配置

开发人员登陆操作。

1. 依次点击“生命周期管理-》生命周期规则-》新增”



2. 设置规则名称和描述。

3. 点击上一步新增的规则栏的配置按钮，在右侧界面中点击新增，配置生命周期规则。

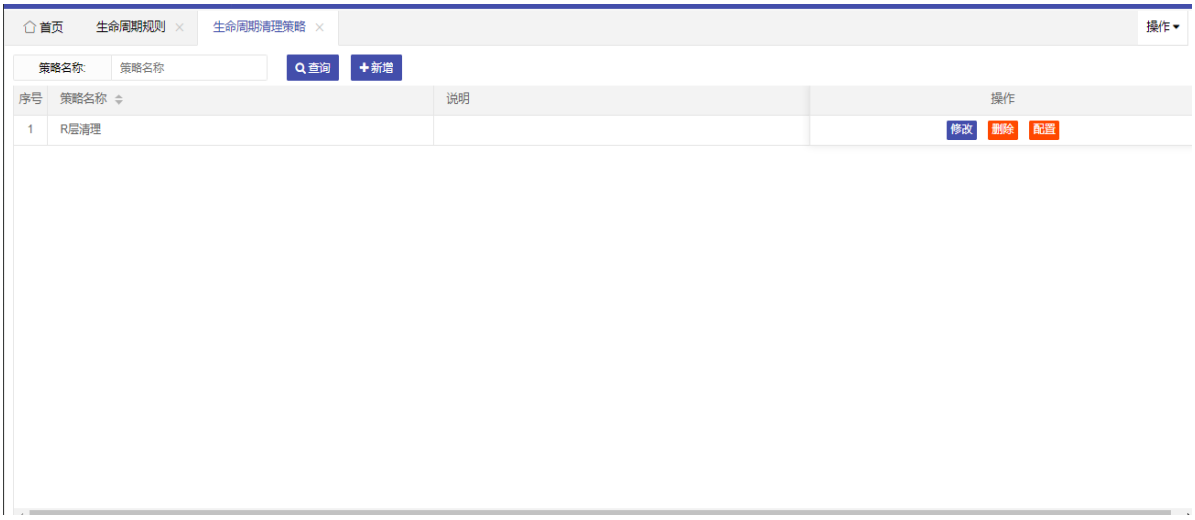
4. 在规则新增界面，设置具体规则。

注意： 可以配置多个规则，但不要重复配置。

4.3.4 生命周期清理策略配置

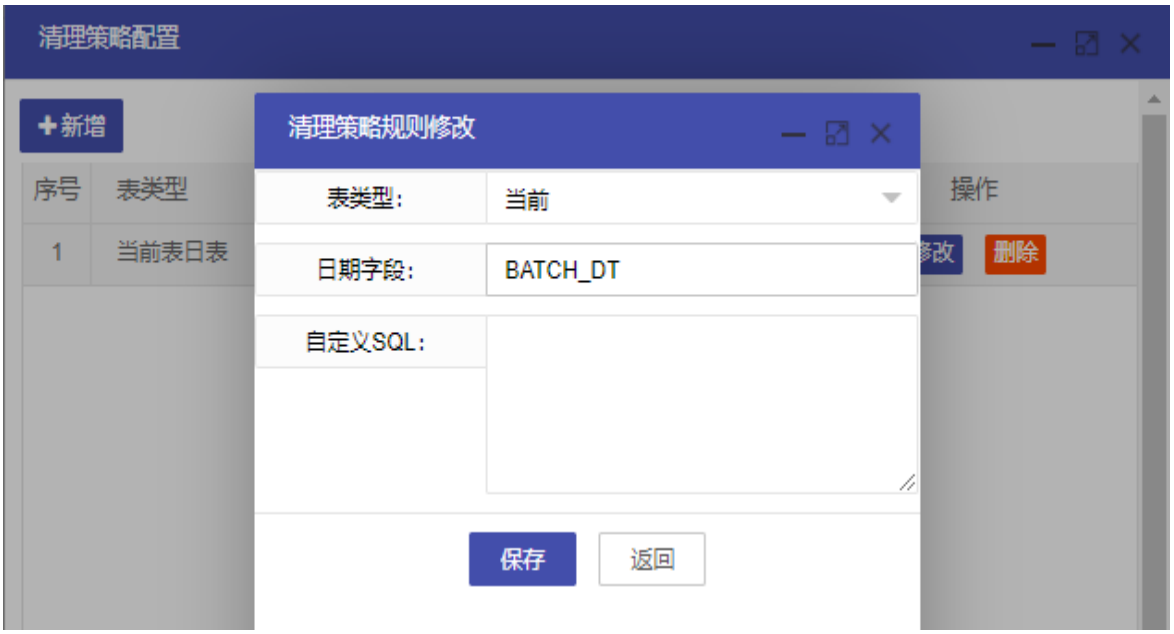
开发人员登陆操作。

1. 依次点击“生命周期管理-》生命周期清理策略-》新增”



2. 设置策略名称

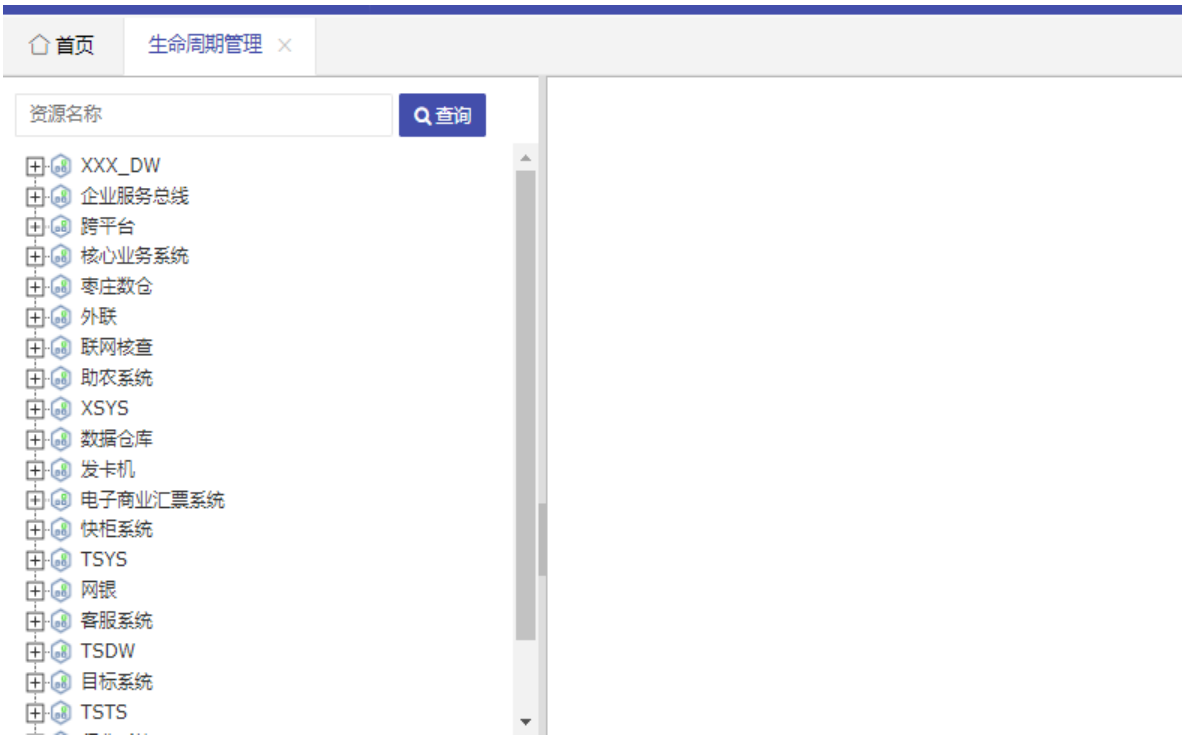
3. 点击配置新增策略规则，跟据表类型设置日期字段或者自定义 sql。



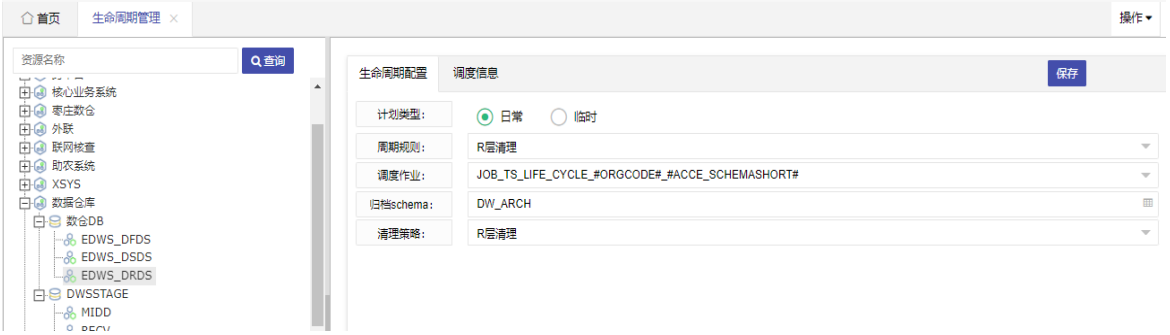
4.3.5 SCHEMA 的生命周期管理

开发人员登陆操作。

1. 依次点击“生命周期管理-》生命周期管理”



2. 在左侧菜单中选择要配置的 SCHEMA，点击后在右侧进行配置。



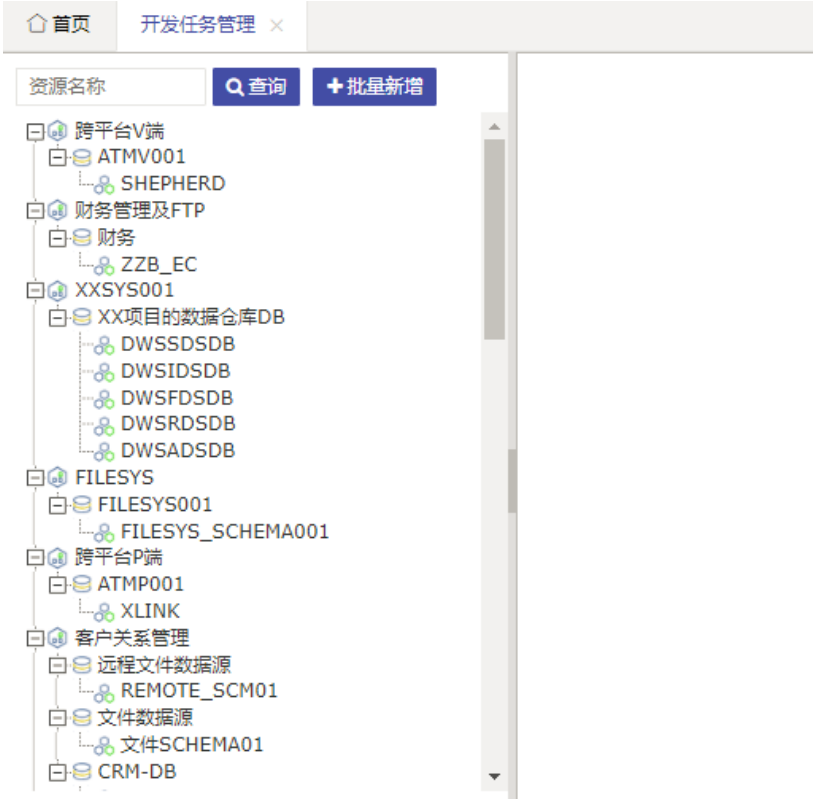
4.3.6 表的生命周期管理

单独对表的生命周期配置，必须要先配置表所属 SCHEMA 的生命周期。

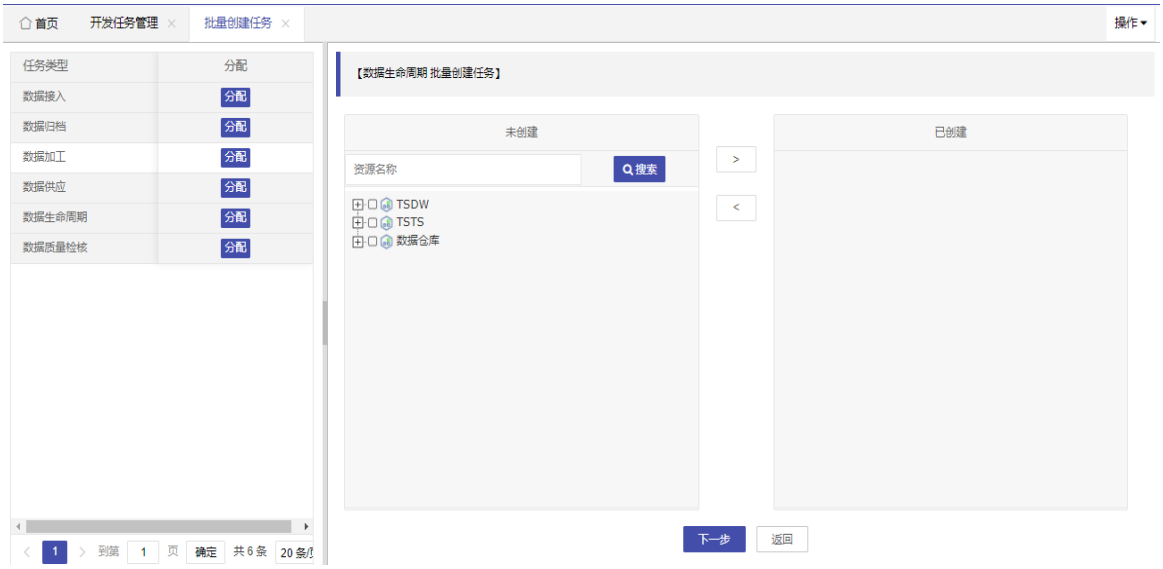
4.3.6.1 生命周期开发任务分配

项目管理员登陆操作。

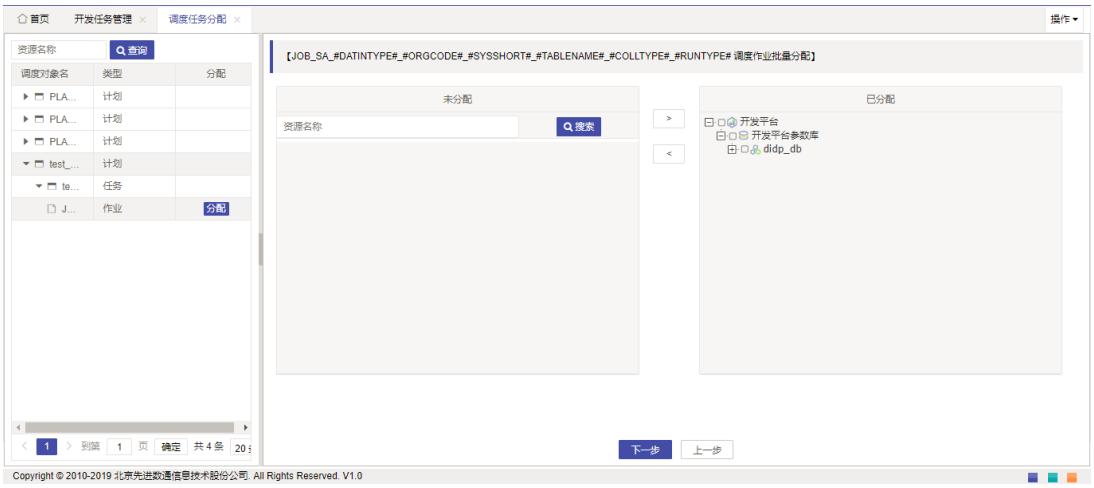
1. 依次点击“项目管理-》开发任务管理-》批量新增”



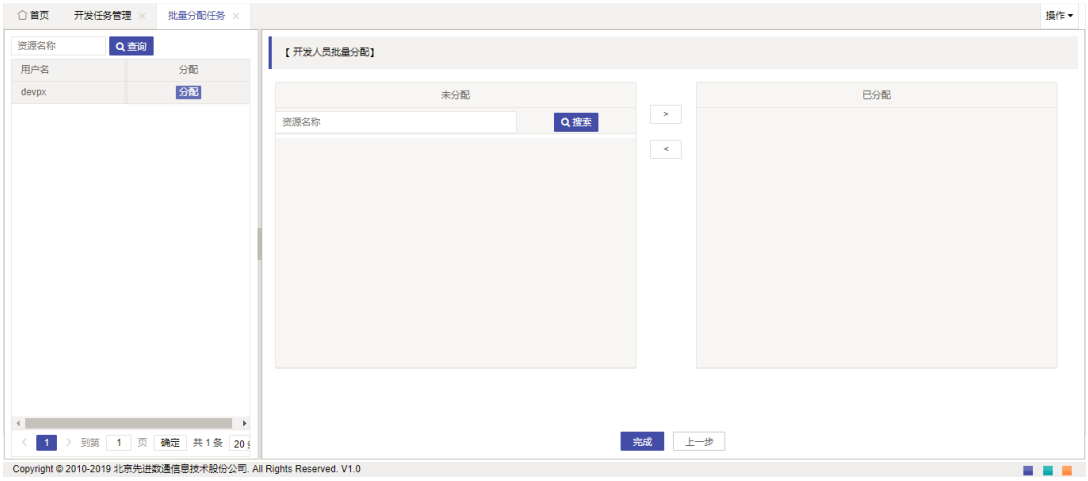
3. 在任务类型（数据生命周期）后，选择“分配”按钮，选择需要分配的对象（表），在左侧未创建列表选择需要分配的表，点击“>”，移动到右侧“已创建”列表中，此时选择数据对象完成；



4. 在调度对象—作业后，选择“分配”按钮，在左侧未创建列表选择需要分配的表，点击“>”，移动到右侧“已创建”列表中，此时选择数据对象完成；



5. 选择分配的开发人员

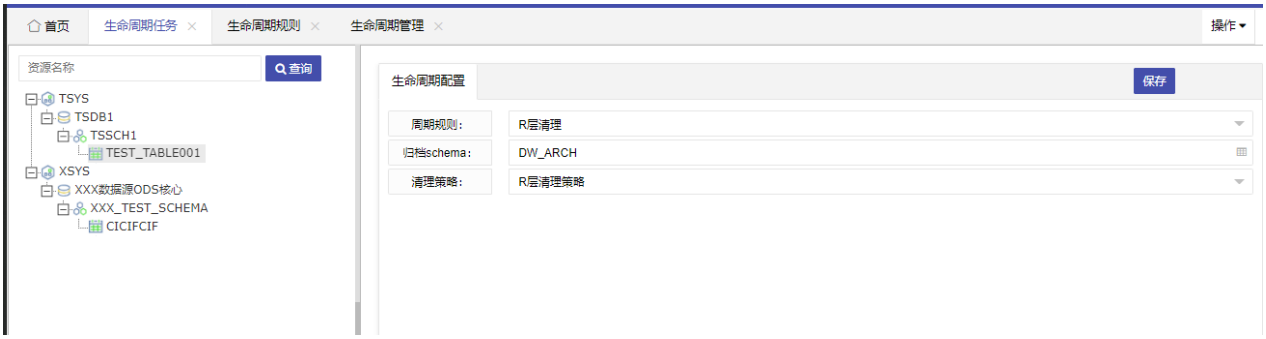


4.3.6.2 生命周期任务配置

开发人员登陆操作。

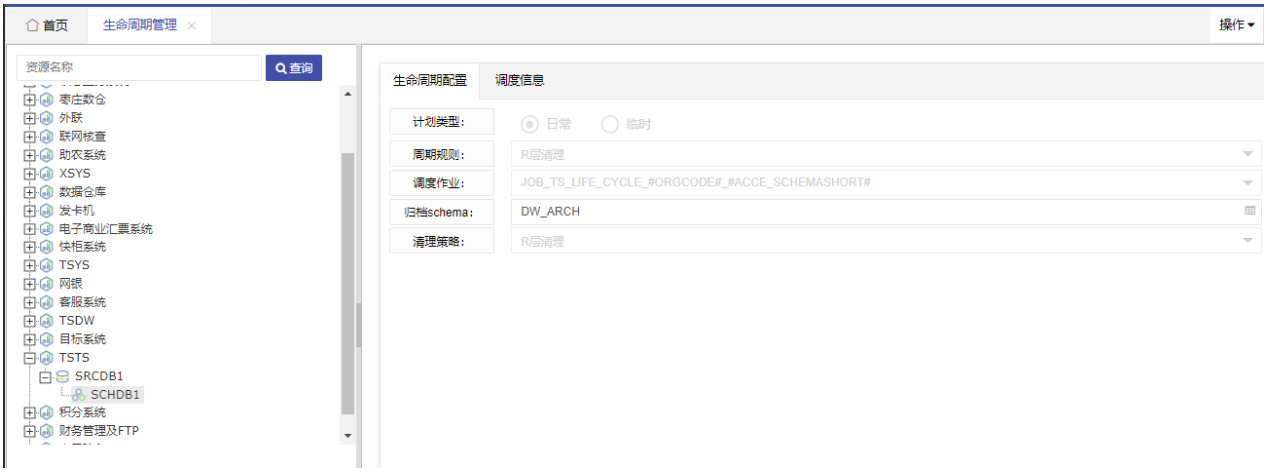
1. 依次点击“生命周期管理-》生命周期任务”

2. 在左侧菜单中选择要操作的表，点击该表在右侧的菜单中进行配置。



4.3.7 发布作业

开发人员登陆操作。



- SCHEMA 的生命周期管理配置保存后可发布作业。
- 一个 SCHEMA 对应一个生命周期管理作业。

- 表的生命周期任务配置需要在发布前配置。
- 表不能单独发布生命周期管理作业。

5 参考数据

5.1 宏定义说明

在 DIDP 上所有的宏分为平台定义宏和调度定义宏。平台提供的宏名，在配置存储时，会根据当前信息进行替换，平台宏的格式为“#名称#”；调度定义的宏在平台上不进行替换，调度平台在实例化会对其进行替换，调度宏的格式为“[名称]”。

■ 平台宏定义的说明：

#ACCE_SYSSHORT#：接入流程的源系统简称（数据接入和数据归档使用有效）

#ACCE_DSOURCESHORT#：接入流程的数据源简称（数据接入和数据归档使用有效）

#ACCE_SCHEMASHORT#：接入流程的 SCHEMA 简称（数据接入和数据归档使用有效）

#ACCE_TABLENAME#：接入流程的源表名（数据接入和数据归档使用有效）

#ACCE_LOADTYPE#：接入流程的加载类型（数据接入和数据归档使用有效，取值范围：INIT/ADD/ALL/DEL）

#ACCE_DATINTYPE#：数据接入类型（数据接入和数据归档使用有效，SA_RECV/SA_MIDD/DWS_RDS）

#ARCHTYPE#：归档类型（数据归档使用有效）

#SYSSHORT#：当前的系统简称

#DSOURCESHORT#：当前的数据源简称

#SCHEMASHORT#：当前的 SCHEMA 简称

#TABLENAME#：当前的表名

#ORGCODE#：机构号

#RUNTYPE#：运行类型

#PROCNAME#：加工名称（数据加工、数据供应使用有效）

#PROCNAME#：加工自定义名称

#SCRIPTTYPE#：脚本类型(pl/py/sh)

#TMPVERSION#：模板版本号

#TARGET_SYSSHORT#：目标系统简称（数据交换、数据供应使用有效）

#TARGET_DSOURCESHORT#：目标数据源简称（数据交换、数据供应使用有效）

#TARGET_SCHEMASHORT#：目标 SCHEMA 简称（数据交换、数据供应使用有效）

#TARGET_TABLENAME#：目标表名（数据交换、数据供应使用有效）

#SRC_FILENAME#：源文件名（数据交换使用有效）

■ MoiaControl 宏定义说明：

[ORGCODE]机构号：Moia

[DATE]跑批日期

[BATCHNUM]批次号

5.2 schema 规范配置指导

本章节为 schema 的规范配置提供指导说明。schema 的规范配置按系统的类型不同分别配置。

5.2.1 业务源系统

5.2.1.1 数据库类型

数据对象名：#SCHEMASHORT#.#ORGCODE#_#SYSSHORT#_#TABLENAME#_# ACCE_TABLENAME #

数据事件名：EVT_DIDP_#SCHEMASHORT#_#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#

表名：#ACCE_TABLENAME#

5.2.1.2 文件

主目录：/file/recv/cbus_cbod/

数据文件类型：普通文件

数据文件名：data/[DATE]/[ORGCODE]/*#ACCE_TABLENAME#.*

控制文件类型：INDICATE

控制文件名：data/[DATE]/[ORGCODE]/indicate.txt

校验码类型：CSUM

DDL 类型：无

DDL 文件名：无

压缩文件名：

压缩包到达文件名：

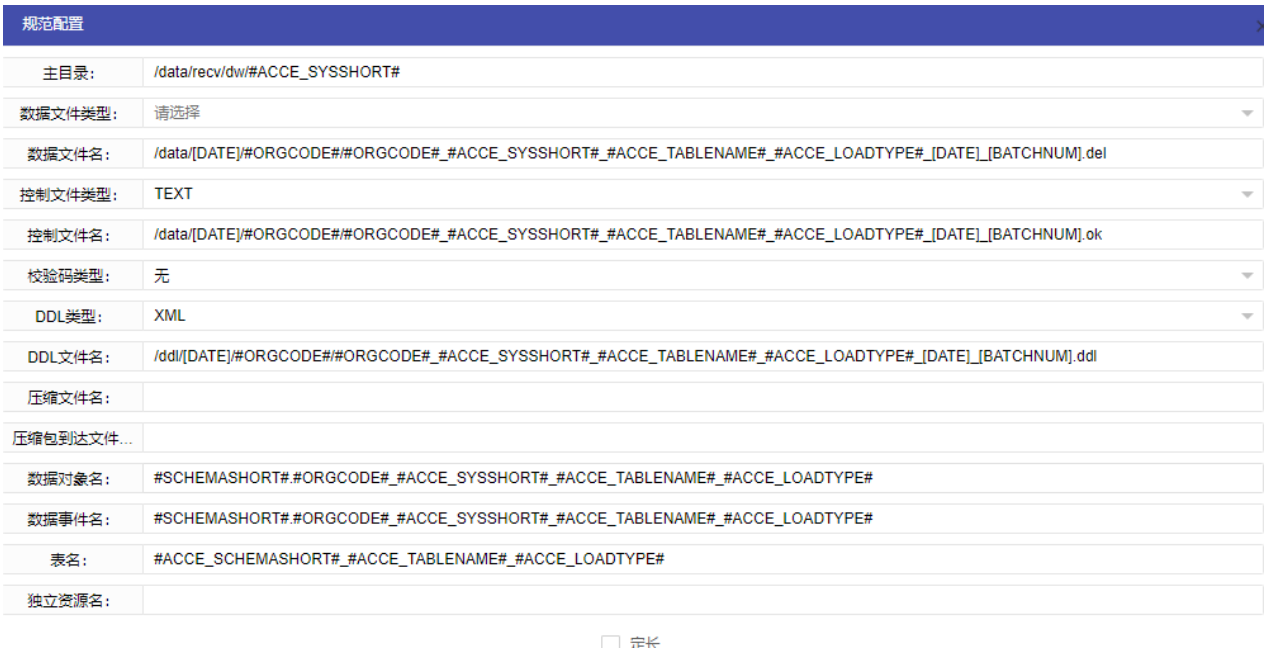
数据对象名：#SCHEMASHORT#.#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#

数据事件名：EVT_DIDP_#SCHEMASHORT#_#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#

表名：#ACCE_TABLENAME#

5.2.2 数仓数据源

5.2.2.1 RECV 区



规范配置

主目录：	/data/recv/dw/#ACCE_SYSSHORT#
数据文件类型：	请选择
数据文件名：	/data/[DATE]/#ORGCODE#/#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#_[DATE]_[BATCHNUM].del
控制文件类型：	TEXT
控制文件名：	/data/[DATE]/#ORGCODE#/#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#_[DATE]_[BATCHNUM].ok
校验码类型：	无
DDL类型：	XML
DDL文件名：	/ddl/[DATE]/#ORGCODE#/#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#_[DATE]_[BATCHNUM].ddl
压缩文件名：	
压缩包到达文件...	
数据对象名：	#SCHEMASHORT#.#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#
数据事件名：	#SCHEMASHORT#.#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#
表名：	#ACCE_SCHEMASHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#
独立资源名：	

5.2.2.2 MIDD 区

规范配置

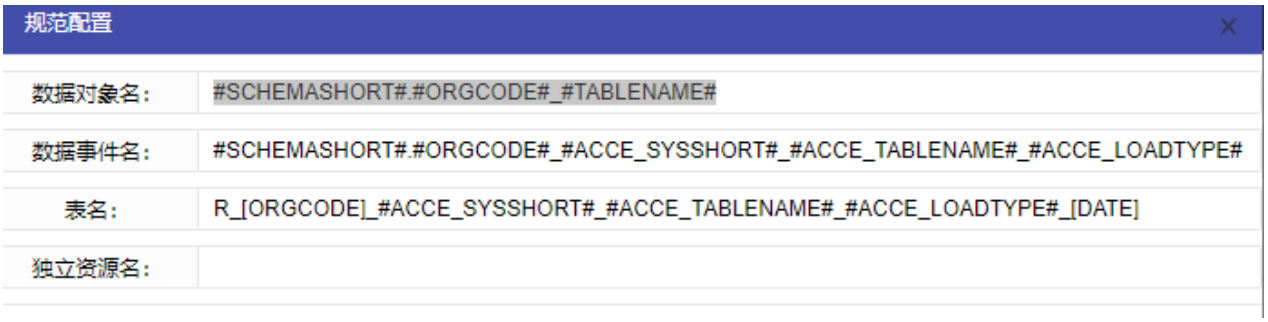
主目录：	/data/midd/dw/#ACCE_SYSSHORT#
数据文件类型：	请选择
数据文件名：	/data/[DATE]/#ORGCODE#/#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#_[DATE]_[BATCHNUM].del
控制文件类型：	TEXT
控制文件名：	/data/[DATE]/#ORGCODE#/#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#_[DATE]_[BATCHNUM].ok
校验码类型：	无
DDL类型：	XML
DDL文件名：	/ddl/[DATE]/#ORGCODE#/#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#_[DATE]_[BATCHNUM].ddl
压缩文件名：	
压缩包到达文件...	
数据对象名：	#SCHEMASHORT#.#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#
数据事件名：	#SCHEMASHORT#.#ORGCODE#_#ACCE_SYSSHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#
表名：	#ACCE_SCHEMASHORT#_#ACCE_TABLENAME#_#ACCE_LOADTYPE#
独立资源名：	

5.2.2.3 SEND 区

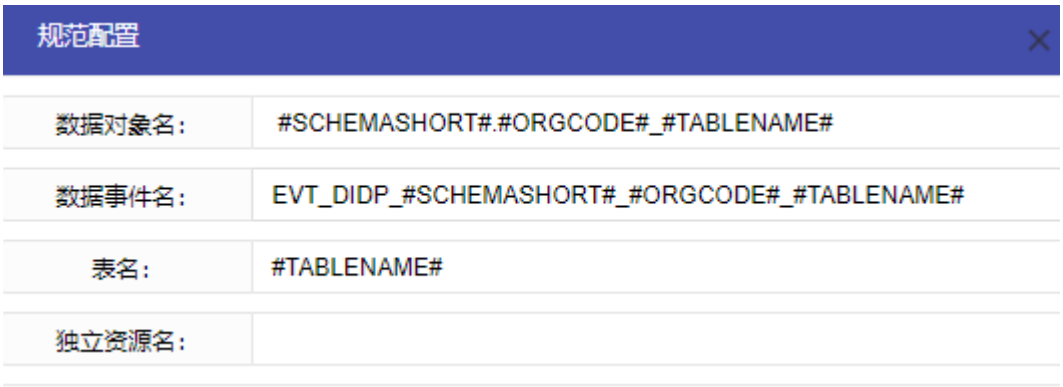
规范配置

主目录：	/data/send/#TARGET_SYSSHORT#/#SYSSHORT#/
数据文件类型：	普通文件
数据文件名：	data/[DATE]/[ORGCODE]/[ORGCODE]_#SYSSHORT#_#TABLENAME#_#LOADTYPE#_[DATE].dat
控制文件类型：	INDICATE
控制文件名：	data/[DATE]/[ORGCODE]/[ORGCODE]_#SYSSHORT#_INDICATE_[DATE].chk
校验码类型：	MD5
DDL类型：	无
DDL文件名：	
压缩文件名：	
压缩包到达文件名：	
数据对象名：	#SCHEMASHORT#.#ORGCODE#_#SYSSHORT#_#TABLENAME#_#LOADTYPE#
数据事件名：	EVT_DIDP_#SCHEMASHORT#_#ORGCODE#_#SYSSHORT#_#TABLENAME#_#LOADTYPE#
表名：	#TABLENAME#

5.2.2.4 RDS 区



5.2.2.5 SDS 区



5.2.2.6 IDS 区

参考 SDS 区配置

5.2.2.7 FDS 区

参考 SDS 区配置

5.2.2.8 ADS 区

参考 SDS 区配置

5.2.2.9 PDS 区

参考 SDS 区配置

5.3 命名约束

标识类命名：数字、字母、下划线的组合，长度大于 1。

6 其它参考

N/A

6.1 公共组件说明

组件名称	组件程序	功能描述
数据库采集组件	didp_export_data.py	从 oracle/db2/mysql/sqlserver/inceptor 等数据库中卸载数据，落地为文件
文件监控组件	didp_file_monitor.py	监控文件类型数据源的目录，判断是否就绪，触发后续文件采集作业
文件采集组件	didp_export_file.py	文件采集作业，调用联盟的拆分程序，功能包括：字段拆分、大机码文件转换等
预处理组件	didp_preprocess.py	入仓数据的清洗，格式化校验，转码等
数据装载组件	didp_load_data.py	数据装载组件
历史库账号对照表加工组件	archive_acc_pty.v1.0.py	历史库的账号对照表加工组件
流水档初始化归档组件	batch_archive_from_hive_table.py	初始化流水档归档，支持动态分区
日常数据归档组件	archive_from_hive_table.py	日常数据的归档，归档包括：历史全量、增量、拉链算法等
标准层主档加工组件	didp_m_standard.py	对入仓数据进行标准化（码值转换）处理，并求出增/全量切片表
标准层流水档加工组件	didp_f_standard.py	对入仓数据进行标准化（码值转换）处理，并求出增量切片表
数仓账号对照表加工组件	didp_acc_pty.py	数仓的账号对照表加工组件
数仓删数组件	didp_data_deleter.py	根据配置的删数条件，逻辑/物理删除数仓数据
数据供应组件	didp_data_provision.py	从数仓中进行数据供应
文件交换文件监控组件	didp_dexc_file_monitor.py	监控需要进行数据交换的文件是否就绪
文件交换组件	didp_file_exchange.py	按 schema 的粒度进行文件的批量交换

数据质量检核组件	didp_data_quality_check.py	对配置的表和字段的检核项进行检核数据量
生命周期管理组件	didp_life_cycle_manager.py	对已有的数据超过生命周期配置的数据进行归档/清理
大字段卸数组件	didp_sqoop_to_hive.py	针对大字段类型(BLOB/CLOB 等)，直接从数据源端抽取到目标 hive 数据库