华傲数据政务大数据融合平台操作手册 V4.1.0

编写人：张玉英

审核：胡云

编写时间：2019-07-27

1. 目录

2. 名词释义

GLDM 模型:政务逻辑数据模型，结合了 5 年以来国内大数据城市建设成果突出的深圳、沈阳、贵阳等 12 个省区市不同层级政府的政务数据模型的设计和实施经验凝练形成的。国内首个指导区域数据资源化，实现跨地域、跨部门、跨业务的数据资源共享、整合、集中、开放建设的大规模数据处理的知识型产品。

基础库：指包含人口库，法人库，房屋库，地址库，车辆库六大库在内基础库，也是一个城市中最核心的库。

主题库：指包含证照库、信用库、社会关系库、视频库和事件库五大扩展库。将来可能随着客户需求增加会有新的主题库产生。

ETL：用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程

归集库：归集库存储所有的源数据，是源数据的一个备份。

贴源层：贴源层是融合平台的数据入口，其作用主要有两个：一、防止后端数据处理出错时，再次执行时反复重抽会给源系统带来不必要的麻烦；二、防止二次抽取数据时，因为源系统的更新导致已经找到不当时数据的快照。

标准层：标准层的数据保存了源系统数据的最新信息，并且在此基础上对源头数据做了标准化清洗等处理，保证信息以同一套标准来表达。

原子层：原子层主要用于以某种形式组织或归类分散在各个源表中的信息项。如果数据不按一定的要求组织在一起，则会造成冗余、缺失等情况带来的数据不一致，并且极大的增加了维护成本，信息也无法做到溯源。

整合层：整合层的数据来源于原子层，整合层是在原子层数据基础上进行数据合并。与其它数据层相比，整合层的数据解决了信息的唯一性问题，并且针对每一项信息的合并规则，都通过足够的样本验证，以确定数据的准确性。
整合层存储信息的具有唯一性，如人的婚姻信息是某种确定的状态，如未婚、初婚、再婚、复婚、离婚、丧偶等中的某一种状态，它已经在多源的原子层信息中做出了最合理的判断。也只有针对人的每种属性的状态确定了，才能支撑后续的各类应用场景。

主题层：主题层是应用基础层，它的数据往往不是直接面向某个应用，而是安装主题划分融合后的基础数据。按数据的统计粒度来细分，主题层又分为基础宽表层和共性汇总统计。
共性汇总层则主要是生成一些公共的统计指标，以减少应用层的重复计算，这种统计基础表的粒度比应用数据可能要细一下，但比基础宽表层要更粗。

应用层：面向各类个性化应用的数据服务区域，向外提供各种数据服务。背后访问的数据，可能是表，物化视图，普通视图，文件或 HDFS 文件等。同时，数据还可能来自数据库、文件系统或大数据平台，因此在对外提供数据服务的时候，需要兼容不同的存储介质。

3. 界面操作步骤

3.1. 整体流程图

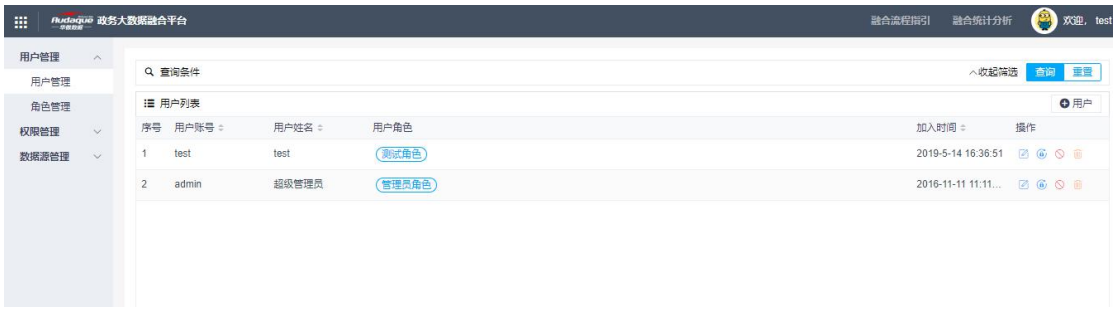
政务数据融合平台全部流程指引

1. 关系型数据源管理：配置关系型数据源的连接信息
2. 大数据来源管理：配置大数据存储源的连接信息
3. 文件源管理：配置文件数据源的连接信息
4. 归集库初始化：配置归集库地址
5. GLDM 初始化： 配置融合平台使用的大数据环境，安装一些必要的脚本和初始化数据，
6. 资源目录管理：建立接入的数据资源的目录，导入元数据信息
7. 资源库管理：管理基础库和主题库，包括对库模型的修改和扩展
8. 归集库管理：管理归集库模型。
9. Gldm 模型管理：融合平台各层表的管理，包括数据源元数据导入，贴源层的表创建
10. 标准代码管理：数据标准的管理
11. 自定义代码管理：非标准代码管理，主要是源数据使用的标准代码
12. 函数规则：定义数据清洗规则，清洗函数
13. 数据集成管理：配置数据在各层表与表之间的字段映射，字段清洗规则
14. 工作流配置：生成数据在各层表之间的处理流程
15. 流程调度配置：配置工作流运行的时间，频率等。
16. 流程调度监控：监控工作流运行状态和日志
17. 调度监控告警：对工作流调度异常问题进行告警和处理流程跟踪
18. 监控告警通知：配置工作流调度告警的通知对象等

3.2. 平台配置管理

3.2.1. 用户管理

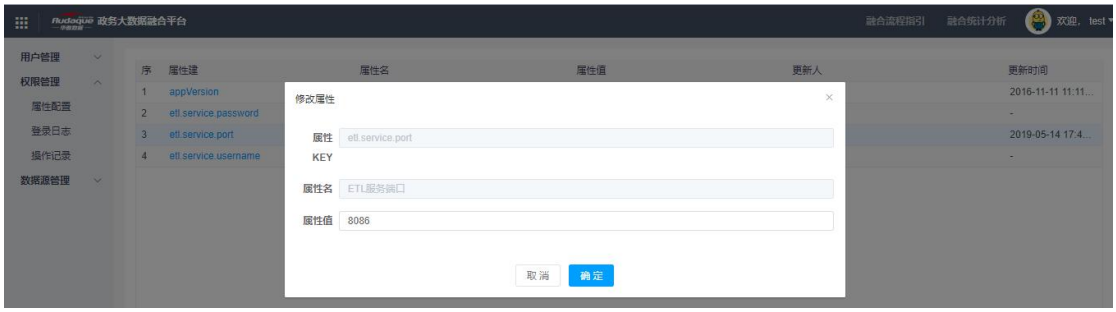
用户管理界面提供新增、修改、删除、禁用账号信息，用户系统登录账号得统一管理，新增得用户账号可赋权后登录系统操作和查看使用系统的功能，禁用账号不可登录系统。已存在的系统账号可修改编辑其基础信息。



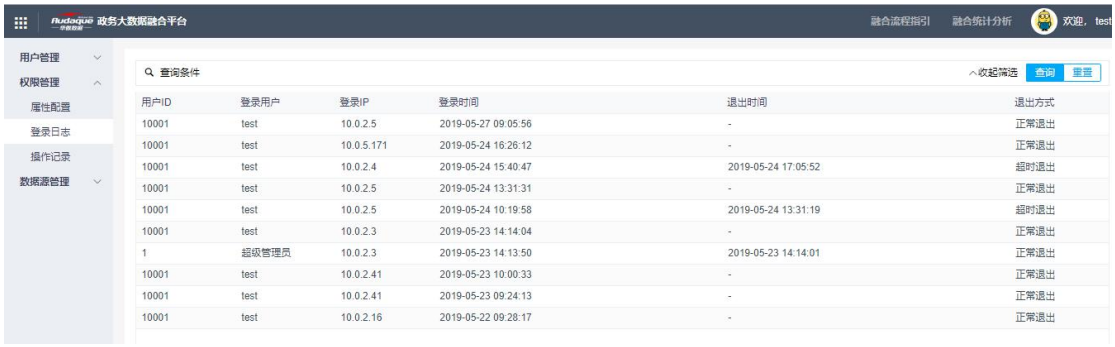
角色管理界面能新增、删除、修改角色信息，用户可以根据实际情况授权访问信息。角色将赋予该账号的权限菜单做操作权限限制，一个账号是管理员角色，那么他登录系统的操作菜单权限将就只是管理员角色的权限菜单。

3.2.2. 权限管理

属性配置，选择属性键，可以修改 ETL 服务器置相关值：



查看登录日志：



查看操作日志：

3.2.3. 数据源管理

数据源管理模块管理所有接入的数据源的连接信息，包括关系型数据源连接信息，大数据源连接信息，文件数据源访问地址。

3.3. 系统配置管理

3.3.1. 归集库初始化

归集库初始化，包括选择大数据平台，数据库类型和指定归集库唯一 ID 号

3.3.2. GLDM 初始化

GLDM 初始化步骤包括选择数据平台，选择存储数据库类型，指定数据库唯一 ID，然后选择初始化执行初始化任务

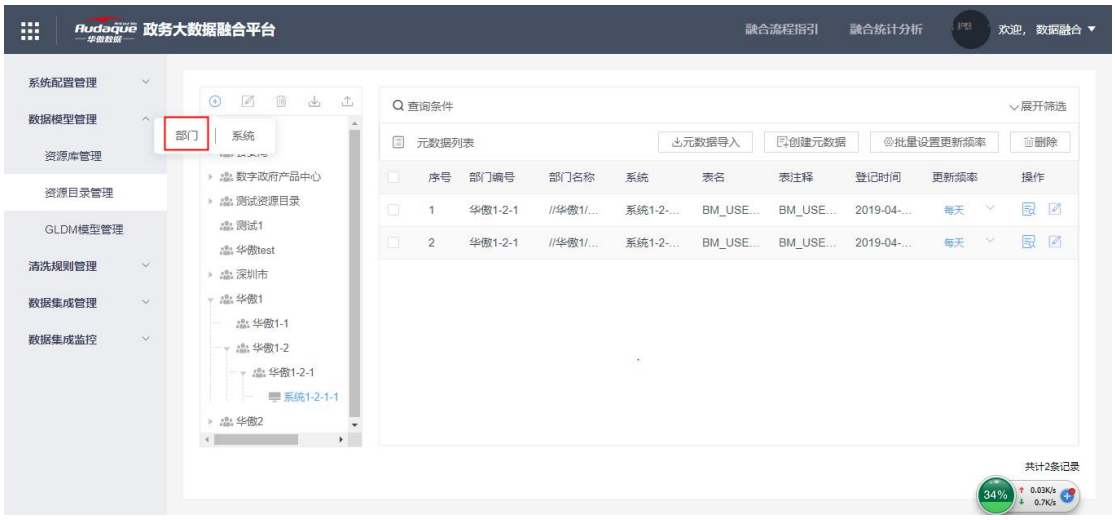
3.4. 数据模型管理

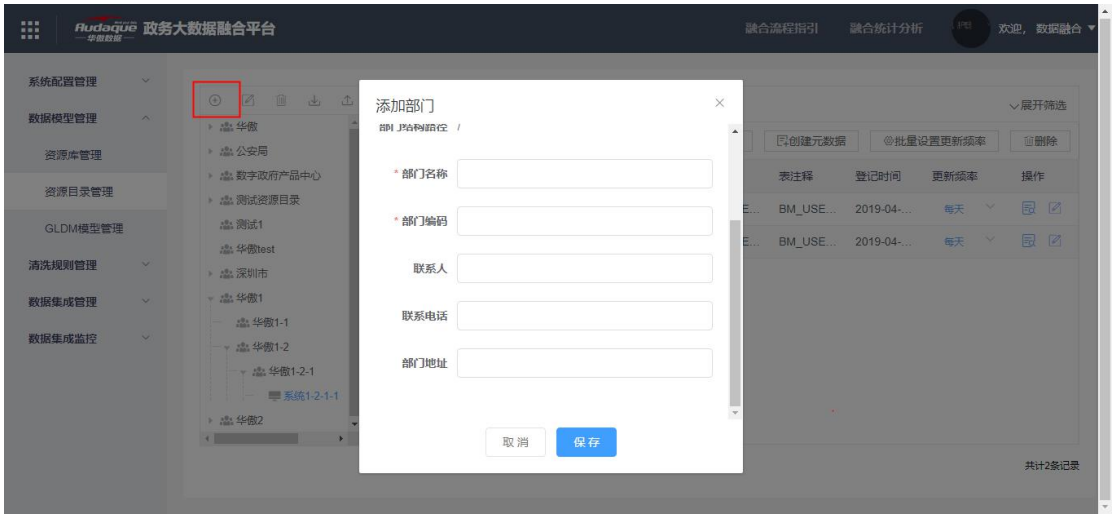
3.4.1. 资源库管理

初始化之后，选择的模块信息会展示，初始化产生的模块信息不支持删除和修改，基础库和主题库支持扩展，能新增修改和删除主题库和资源库

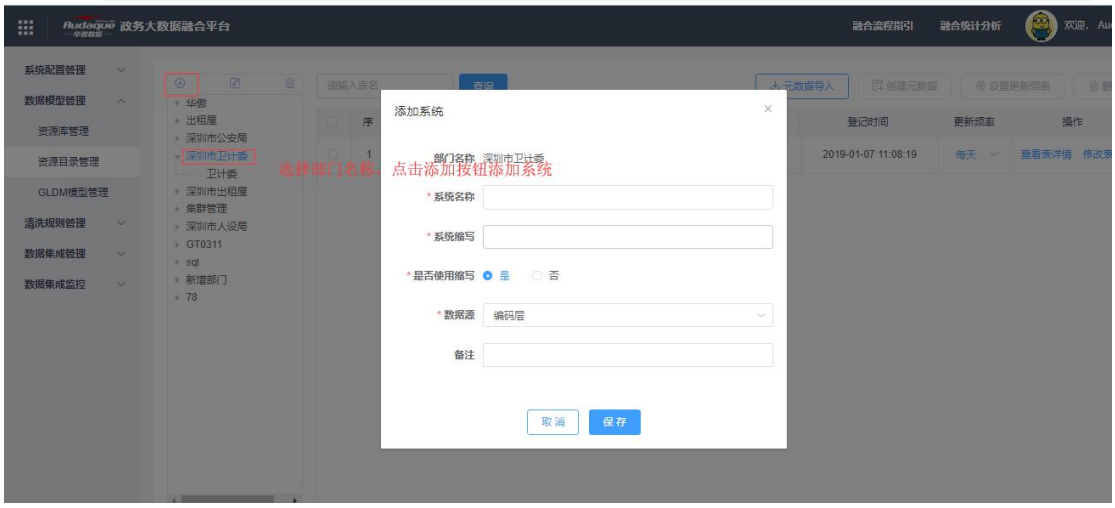
3.4.2. 资源目录管理

1.进入资源目录管理，在业务源添加部门名称，也可以对添加以后的业务源进行修改删除操作，可多级部门 ，支持 EXCEL 导入导出资源目录组织架构。主要用于共享交换平台资源目录中部门与系统，部门与部门层级关系的组织维护建设。如图

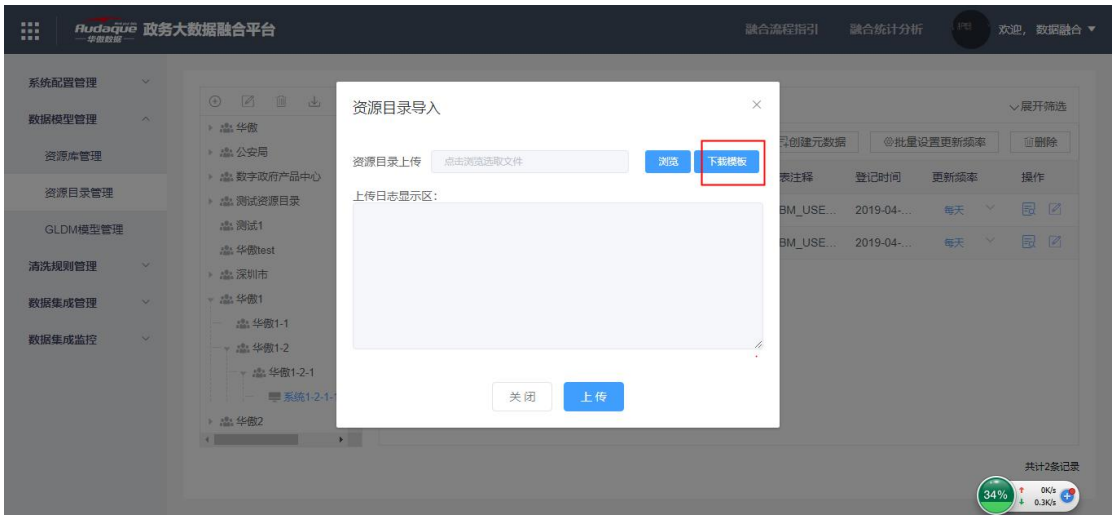




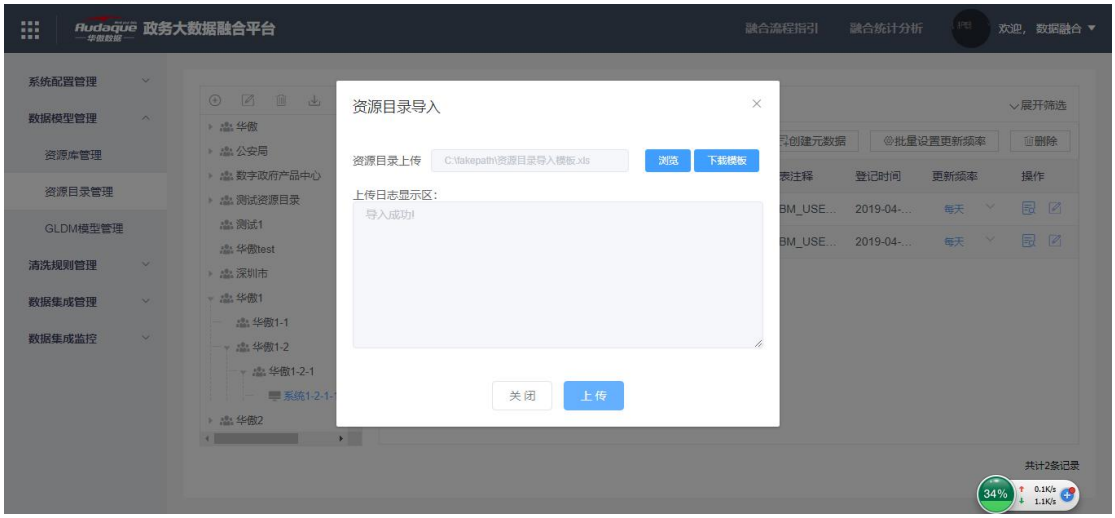
2.在部门目录下点击添加系统名称，也可以对添加以后的系统名称进行修改删除操作 如图



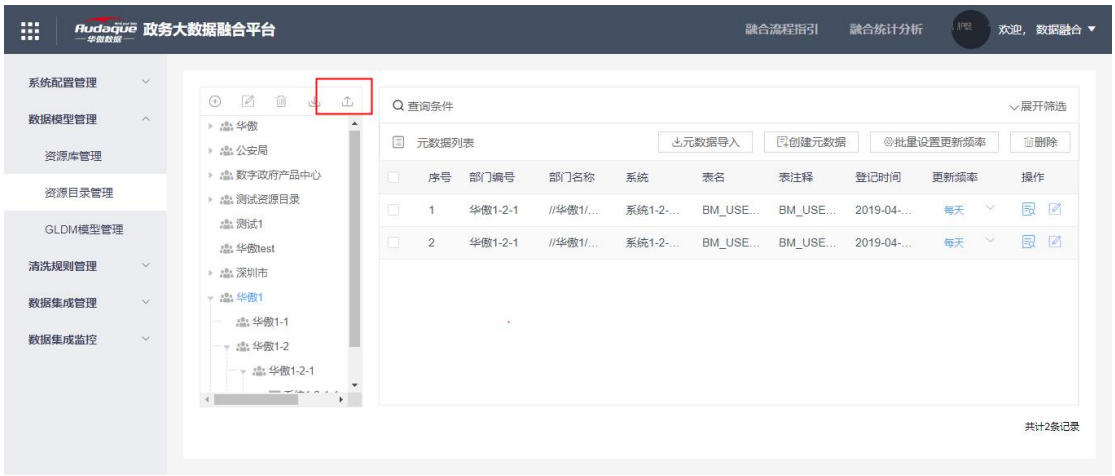
3.资源目录架构导入及导出



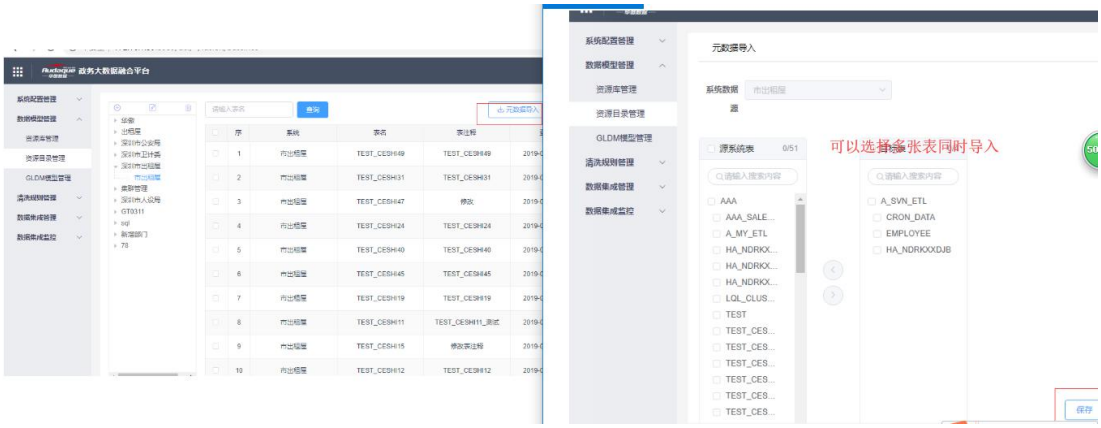
根据模板格式，填写部门或系统信息字段，点击上传，系统会将上传结果日志展示



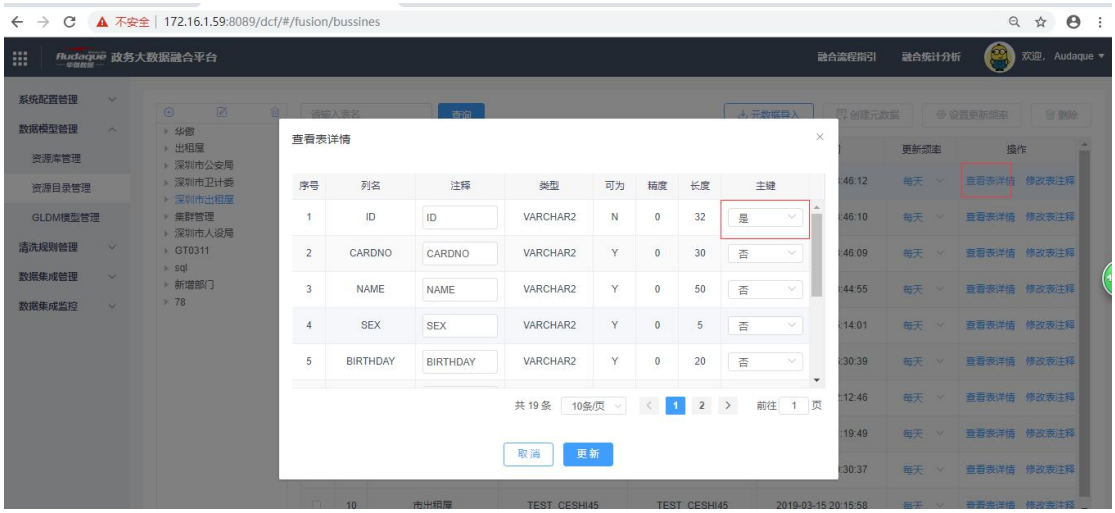
点击导出按钮，将资源目录组织架构导出 EXCEl



4.导入数据源元数据（支持一键导入功能），界面中一键导入为，选择系统即对应得业务源数据库时，一键导入将全部表全部导入。



5.元数据导入之后，没有主键时，可以手动设置主键



6.支持手动建表和批量导入，批量导入模板中一个 sheet 一个表

3.4.3. GLDM 模型管理

GLDM 模型管理功能管理整个模型的表结构的创建，修改。贴源层，标准层，原子层，整合层，主题层，应用层。

3.4.3.1. 贴源层表创建

贴源层表结构根据源数据表结构自动生成，创建过程分为元数据导入，创建并导入两个步骤

1、选择元数据导入

2、选择要创建的表，选择创建表并导入完成表的创建

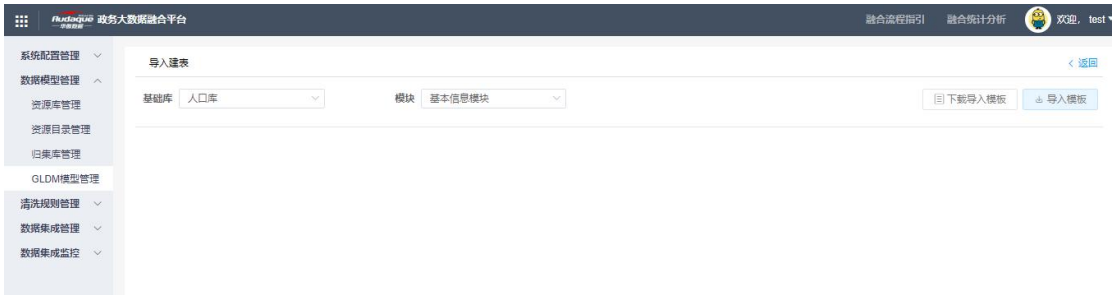
3.4.3.2. 标准层表创建

标准层表在数据集成配置中创建，在贴源层表向标准层映射配置完成后创建，表结构为贴源层表结构添加清洗字段，如下图，在完成列映射配置后，选择创建表

3.4.3.3. 其它层表创建

其他层包括原子层，整合层，主题层，应用层。如下图有手工建表，和导入建表两种模式。当需要创建大量的表时，建议选择导入建表，下载建表模板，填入表结构信息后导入模板完成创建

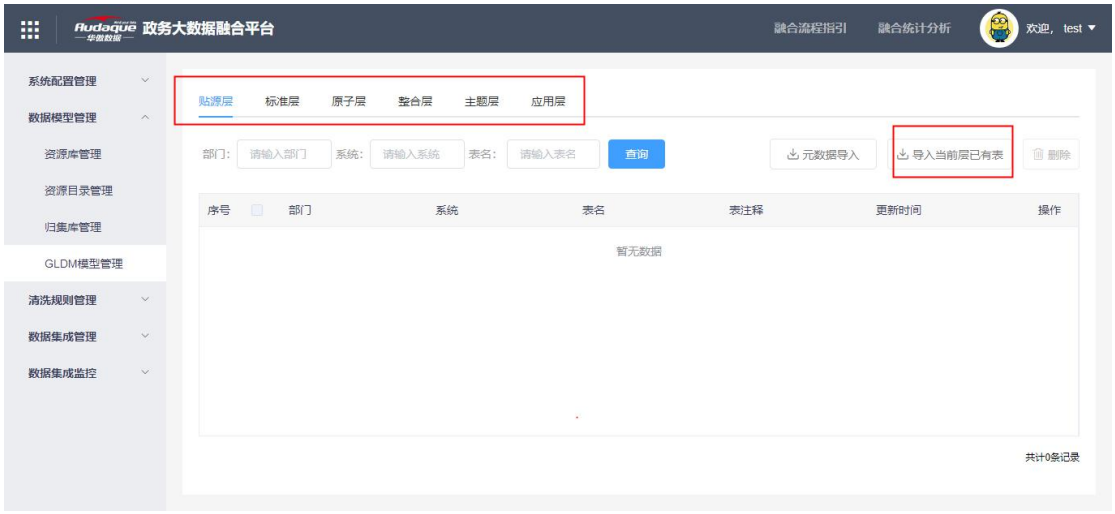
导入批量建表模板



3.4.4. 各层导入当前已存在表

在大数据集群数据仓库中已经存在的表，即通过大数据数据仓的后台建的表，可

以在此导入模型管理中，可以根据建表的表名前缀，标识是哪个层即数据区，本次版本大数据平台可以归集库 hive，数据仓为 hive，每个层的表均有特定的标识，如标准层表都是 std_前缀，应用层为 app_,主题层为 dm_等等。



3.4.5. 归集库管理

归集库管理提供对归集库表结构预览和修改功能

3.5. 清洗规则管理

3.5.1. 标准代码管理

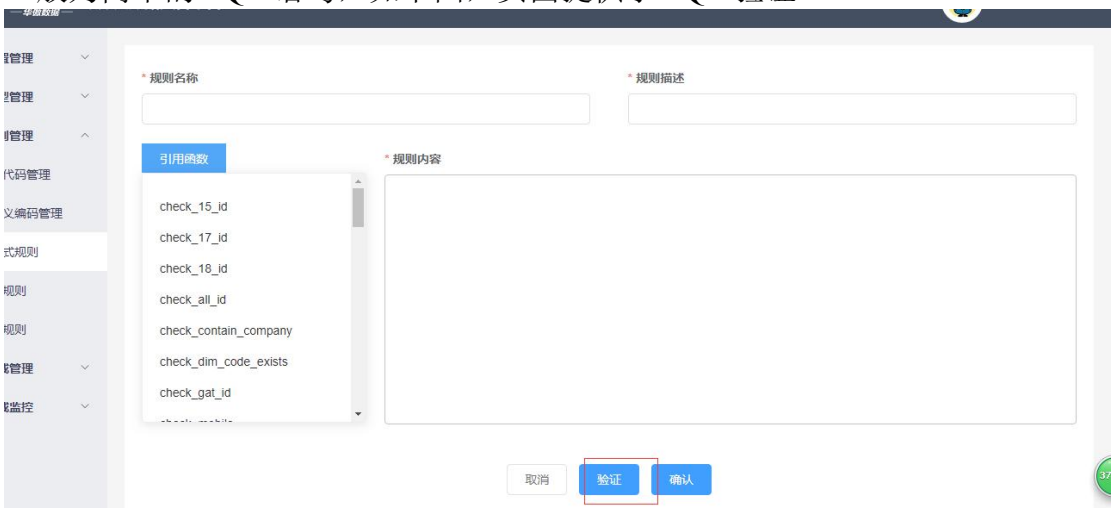
当 init.propertiesp 文件中配置的是通质量平台一起部署，该界面只能查看，如果配置文件设置的不是同治理平台一起部署，界面可以新增修改删除数据。标准代码用于数据从贴源层到标准层时，对非标准化的源数据进行标准化。

3.5.2. 自定义代码管理

自定义代码界面可以新增删除和修改，自定义代码主要用于收集整理源数据使用的非标准代码，或自定义的非标准编码集

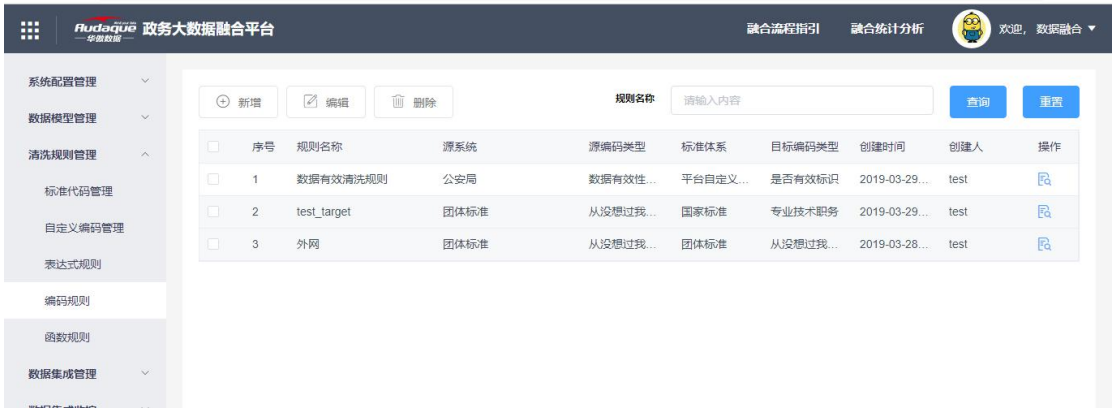
3.5.3. 表达式规则

表达式规则页面定义数据清洗转换的规则，规则符号数据库语法即可。表达式规则用于在数据从贴源层到标准层的映射时，对表字段进行数据处理。表达式一般为简单的 SQL 语句，如下图，页面提供了 SQL 验证



3.5.4. 编码规则

编码规则指源编码到目标编码的映射关系。用于数据从贴源层到标准层的映射中，对字段用编码规则进行转换，把字段转换成目标编码。目标编码一般为国家标准编码，行业标准编码……等。



3.5.5. 函数规则

函数规则页面用于定义对数据进行复杂的清洗。需在数据库端开发清洗函数，然后在函数规则页面注册成规则。用于在数据从贴源层到标准层的映射中，对字段值用函数规则进行处理

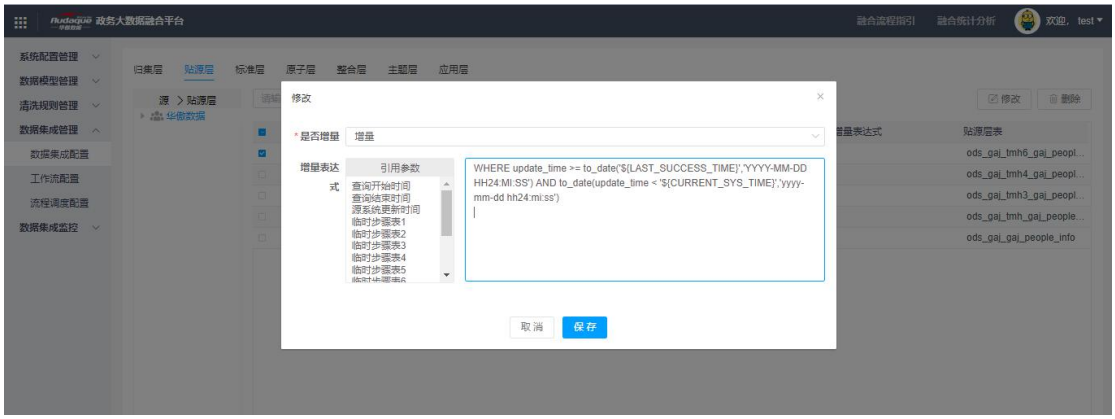
3.6. 数据集成管理

3.6.1. 数据集成配置

3.6.1.1. 贴源层

源数据到贴源层只要是在 gldm 模型管理创建表之后，就会在这一层展示，贴源层数据更新模式有增量和全量两种模式，全量模式指每次数据全抽取，全覆盖。增量模式需配置表达式，每次调度都从上一次成功调度截至时间开始获取增量数据

（注：增量表达式样式：WHERE update_time >= to_date('${LAST_SUCCESS_TIME}','YYYY-MM-DD HH24:MI:SS') AND to_date(update_time < '${CURRENT_SYS_TIME}','yyyy-mm-dd hh24:mi:ss')）



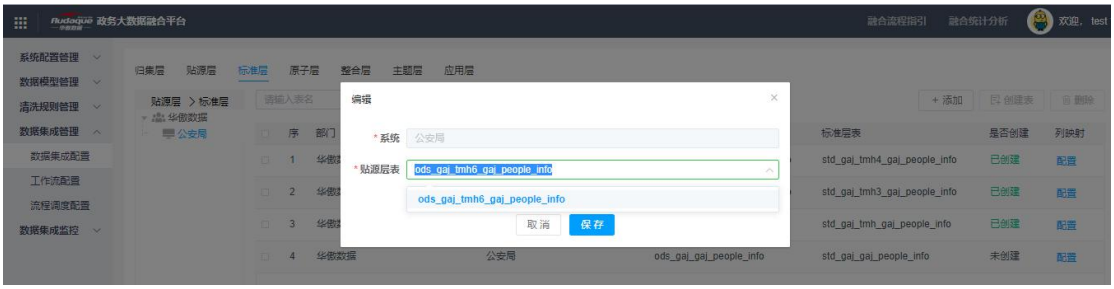
3.6.1.2. 标准层

这一层的数据来源广泛，各类来源数据标准完全不统一，因此在贴源层之上，还会通过标准化映射过程生成标准化数据，保障数据按照统一的标准进行表达。标准层的数据更新并不是追加模式，而是覆盖更新。同时，标准化后的映射列并不会覆盖原来的列，而是新增对应列来保存映射后的值。

从贴源层到标准层时，数据需要进行标准化处理，这个过程可能包含了数据清洗、转换、编码映射等过程。对于编码映射的过程，需要用到大量的数据元标准，而很多标准是可能，已经有现在的国家、地方或行业标准的，如性别、婚姻状况、学历等，有些编码，这些我们直接参考已有的标准，再将源数据的编码映射到标准编码即可。如果没有可参考的标准，则我们需要为这些数据制订标准，以便在多源数据合并时，提供统一的标准。这些都是在做标准层之前，或者过程中需要做的工作。

1. 添加标准层表映射

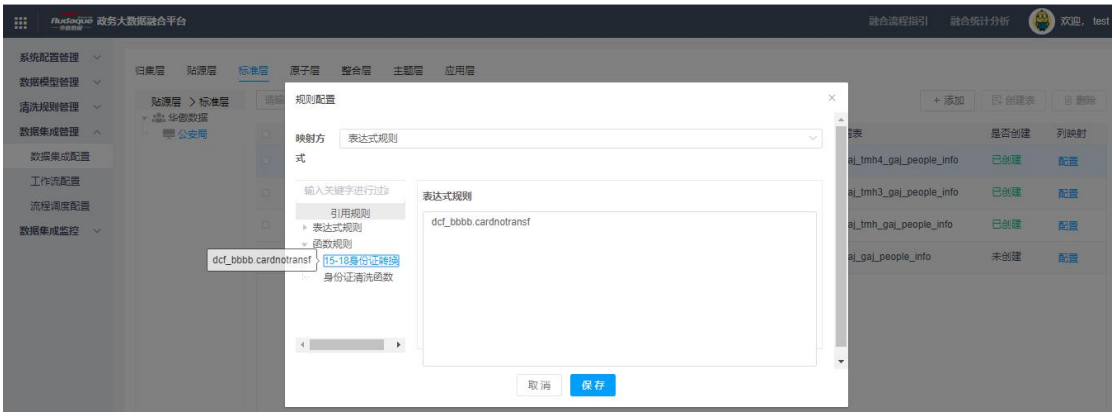
选择标准层，选择添加，挑选要标准化的表



2. 添加表成功之后，进行列映射配置，列映射配置完成之后，点击创建表成功，标准层映射配置完成

添加要映射到标准层的表：

添加规则：

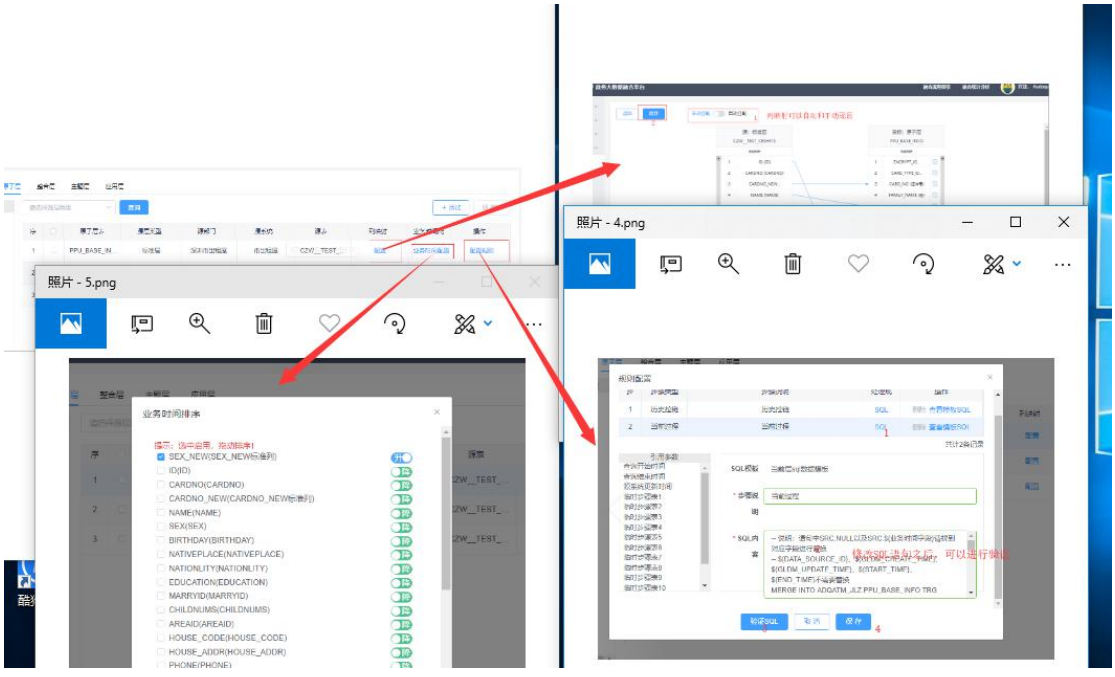


3.6.1.3. 原子层

所谓原子层，顾名思义，就是指此处数据的粒度很细，就像原子一样不可拆分。这种特性反映在数据行上，就是指数据是明细的，甚至各来源的数据都没有做任何合并处理，各源的数据保持完全独立。反映在数据列上，则是指组织这些信息的数据表包含的属性比较小，如人的信息会拆分到各个片段或阶段，如人可分为基本信息、关系信息、联系方式信息、联系地址信息（包括户籍、居住和工作地址，与地址库、房屋库可关联）、教育信息、婚姻信息、生育信息、就业信息（与法人库可关联）、保障信息、公积金信息、名下房产、名下车产、名下企业（与法人库可关联）、良好记录、不良记录、死亡信息等，即使某张源数据表全部或部分包含了这些信息，它也会分多次从这张表中提取信息。

1. 选择表，新增表映射

2. 列映射配置、规则配置、业务时间配置等



3.原子层模板语句更改
语句中 SRC.NULL 以及 SRC.${业务时间字段}请找到对应字段进行替换
${DATA_SOURCE_ID} 、 ${GLDM_CREATE_TIME} 、 ${GLDM_UPDATE_TIME}、${START_TIME}、${END_TIME}不需要替换

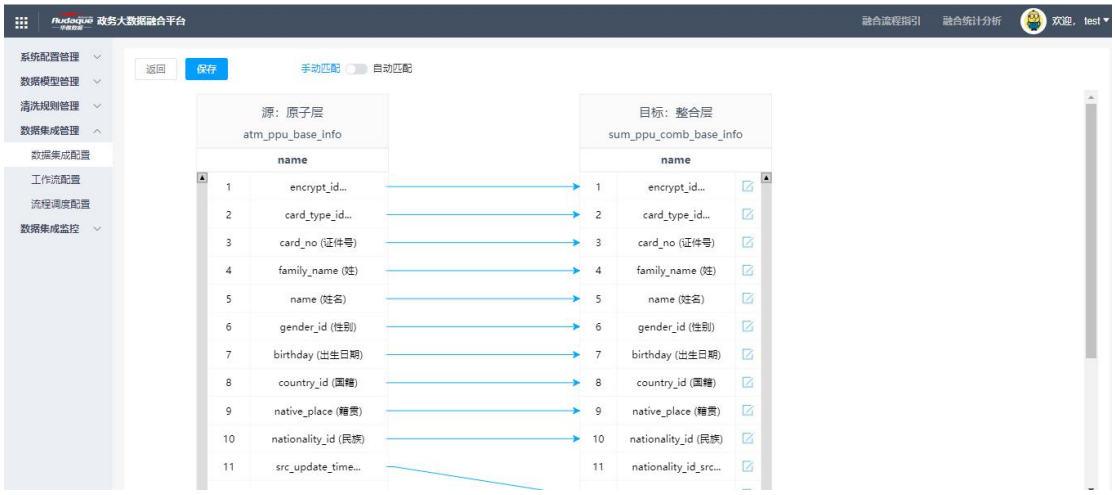
华为的 MPP 数据库需要注意下，如果目标列的字段类型是日期类型，但是没有源列，即只是给 NULL 给目标列，需要显示转换成日期类型， CAST(NULL AS TIMESTAMP)或者
NULL::timestamp,其它类似如果直接给默认值的话，都是需要显示转换下，例如某个目标列想给个默认值 1，可以 '1'::varchar(1) ,否则会报错。

3.6.1.4. 整合层

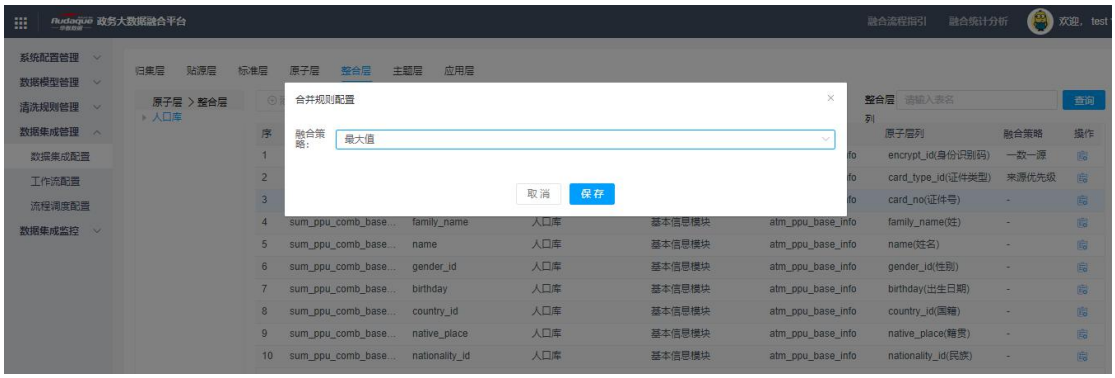
整合层相对于原子层的表模型来说，基本没有做大的变动，并且基本在从原子层到生成整合层数据的过程中是表对表的。当然，信息是做了一些扩展的，特别是在数据合并的时候，会衍生出很多标签或统计信息。如针对某人的电话号码，就可以衍生出如下指标：最早登记时间，最近登记时间，被多少个来源登记过，曾经被哪些人作为登记联系方式，在所有人中被最早登记的时间等。

整合层配置表映射之后，可以对列字段进行融合策略配置，列映射配置可以手动和自动匹配，保存之后，可以选择 SQL 进行验证，整合层 SQL 语句一般不需要修改，可以直接使用

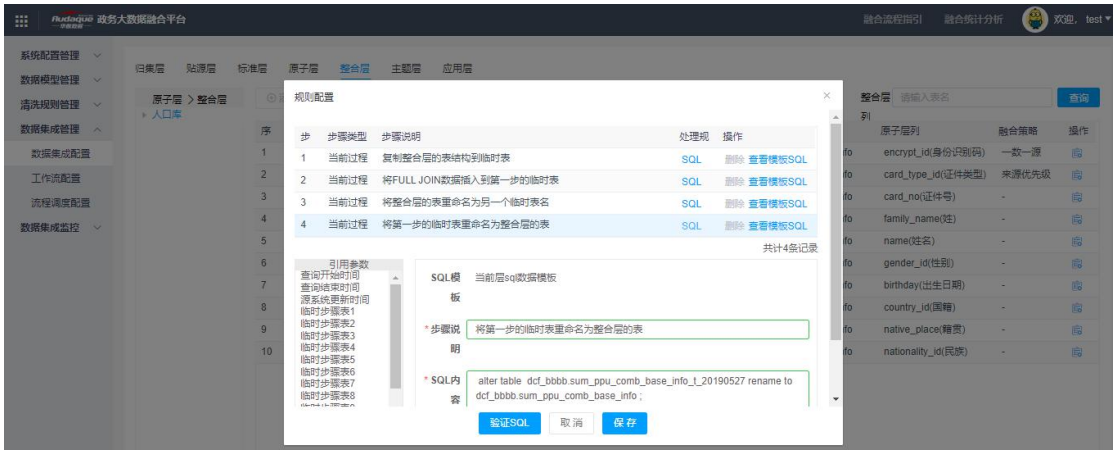
配置列映射：



选择融合策略



配置规则中验证 SQL



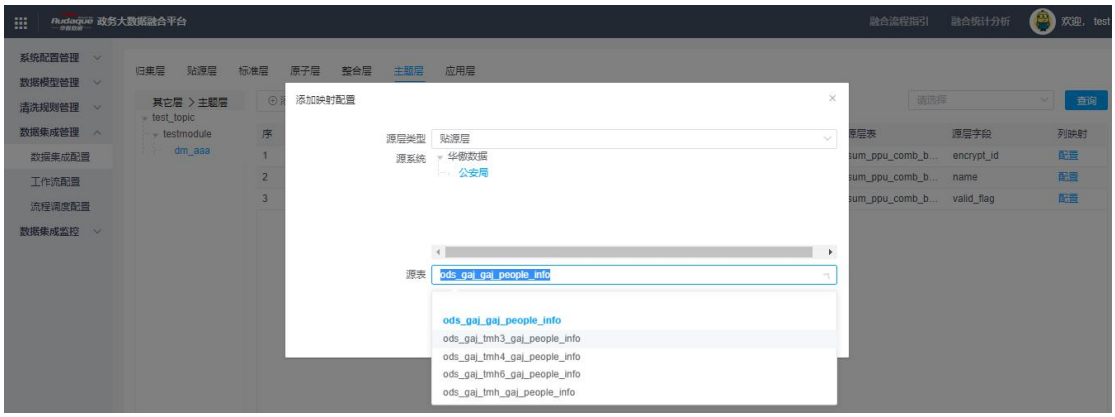
3.6.1.5. 主题层

主题层就是通过关联的方式，拼接整合层的片段表，将这些信息拼在一起形成各类应用需要的基础宽表。

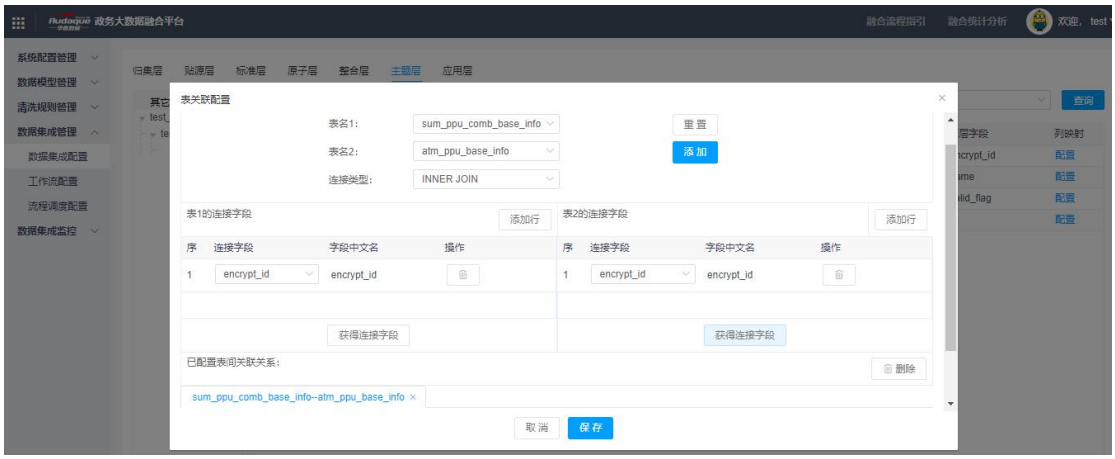
如统计分析的宽表，会将一些统计分析所需求的基础属性，如性别、年龄、籍贯、政治面貌、婚姻状况、户籍、居住区域、工作区域、五险状态、最近缴纳社保时间、公积金状态、最近缴纳公积时间等多达 50 多个属性、标签或统计指标，通过这些维度和指标和组合，可以进行多样化的统计分析应用，而挖掘基础宽表更是多达 150 多个字段，这些属性都是基于大量的应用和经验抽象出来的。如果还有更个性化的需求，则可以扩展出更多的应用基础表或宽表。

主题层新增表映射时，可以增加多个源，配置好表关联之后，进行列映射配置，列映射配置可以手动和自动关联，保存之后，修改 SQL 语句，进行 SQL 验证，SQL 语句一般修改为空的地段，给出默认的值就行，根据具体需求而定

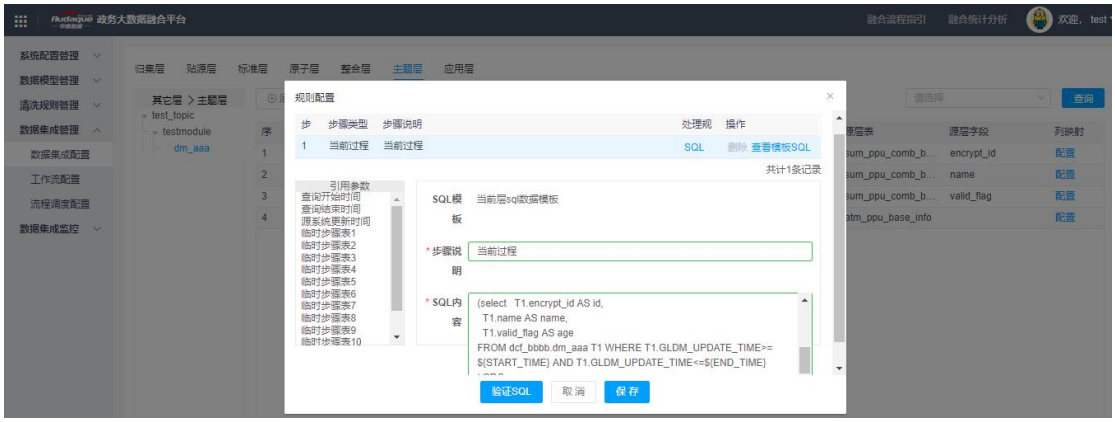
添加到主题层的映射



表关联配置：



SQL 调整

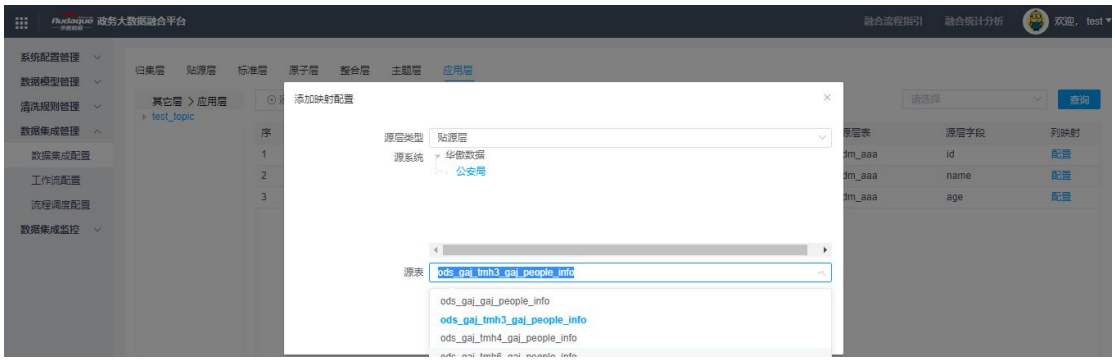


3.6.1.6. 应用层

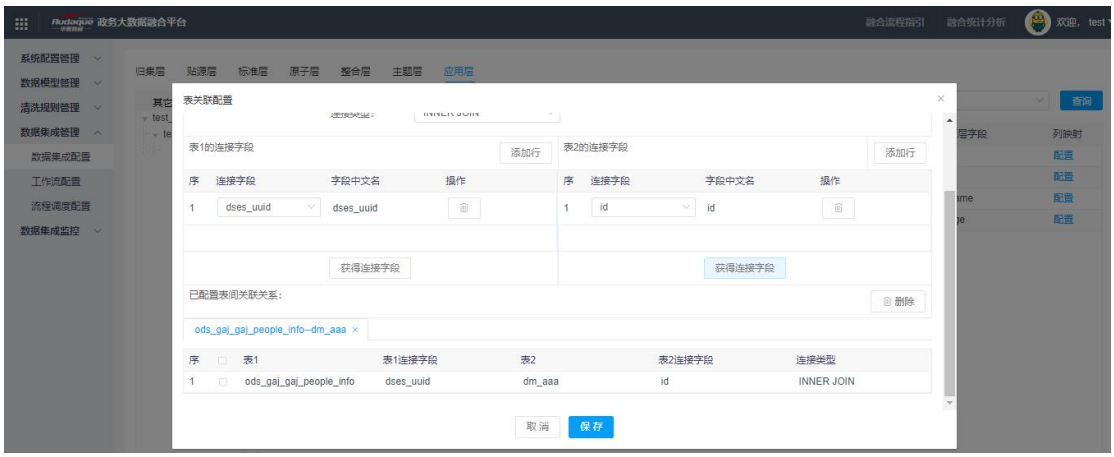
主要要兼顾应用需求的扩展性、易变性，并且要充分考虑访问的性能问题，因此，除了结构上尽量灵活（甚至会考虑用键值对的方式保存），还可能通过一些索引的手段提升性能。

应用层和主题层配置一样。

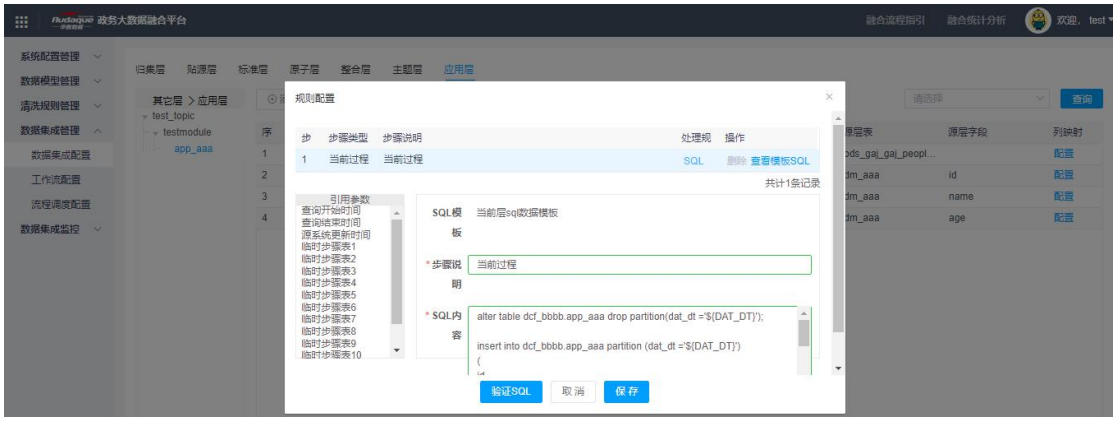
添加到应用层的表映射：



添加表关联条件：



配置规则中 SQL 调整



3.6.1.7. 工作流配置

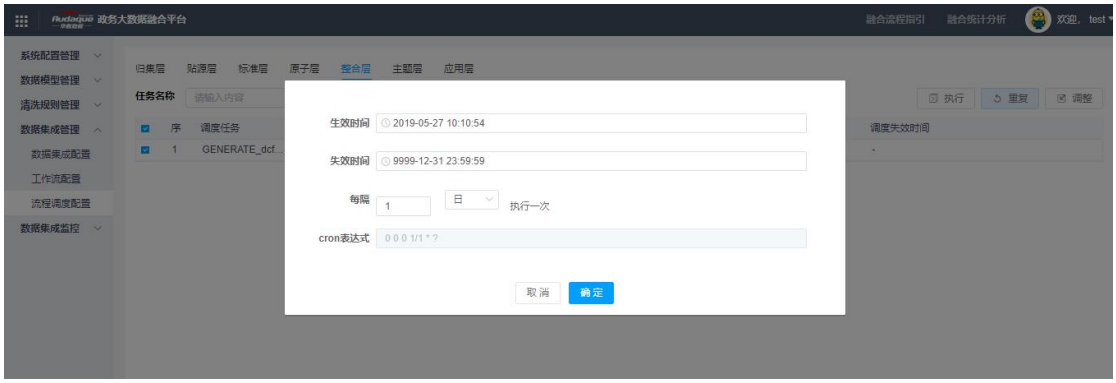
1. 添加工作流信息，每一层都可以添加工作流，可以对语句添加的工作流就行更新和删除操作，也可以查看工作流信息，包含该工作流成的生成和从源到过程，结束的整个可视化流程信息，记录了该流程的每一个节点动作的表关联追踪。

添加工作流：

查看工作流全图：

3.6.1.8. 流程调度配置

调度配置可以立即执行，也可以配置定时启动任务，支持批量设置调度时间。立即执行将对该做工作只触发该次流程执行，重复调度可配置对应的策略，包含每时，每日，每分，每秒等更复杂丰富扩展的调度策略，服务将根据配置执行流程。

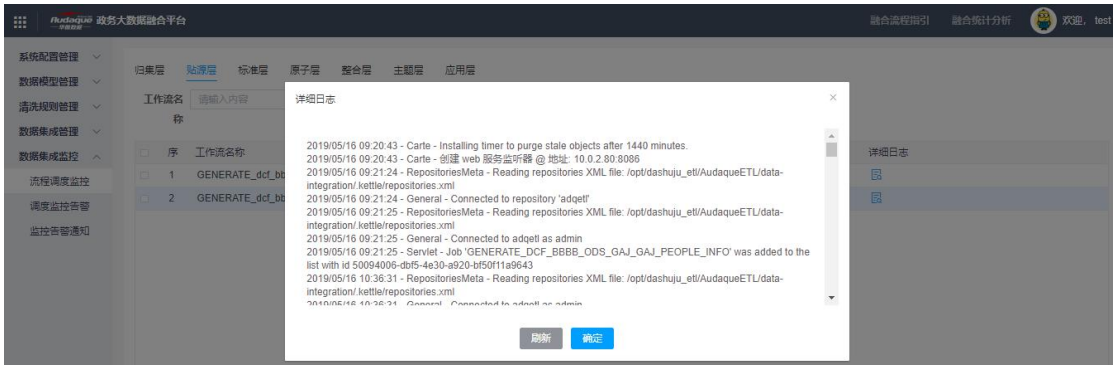


3.7. 数据集成监控

3.7.1. 流程调度监控

1.查看各层任务执行情况，可以查看节点日志信息和详细日志，节点日志中将对每个流程在不同节点步骤记录执行日志，详细日志中详细记录了流程执行调度输出的 SQL 执行数据连接等非常详细的日志记录。

查看详细日志：



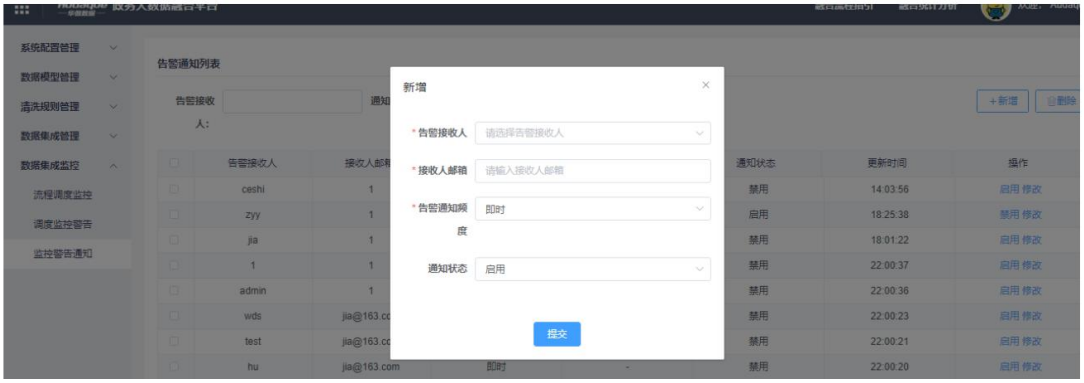
查看节点日志：

3.7.2. 调度监控警告

1.调度任务监控是查看调度任务中的错误信息，可以点击进行处理，

3.7.2.1. 监控警告通知

1.告警通知添加接收人以后，主要是提醒任务的错误信息，方便当事人连接调度情况，及时处理错误信息


系统配置管理
数据模型管理
清洗规则管理
数据集成管理
数据集成监控
流程调度监控
调度监控警告
监控警告通知
告警通知列表
告警接收人：
+新增
删除
新增
*告警接收人
请选择告警接收人
*接收人邮箱
请输入接收人邮箱
*告警通知频度
即时
通知状态
启用
提交
告警接收人
接收人邮箱
通知状态
更新时间
操作
ceshi
zyy
jia
admin
wds
test
hu
禁用
启用
14:03:56
18:25:38
18:01:22
22:00:37
22:00:36
22:00:23
22:00:21
22:00:20
启用 修改
禁用 修改