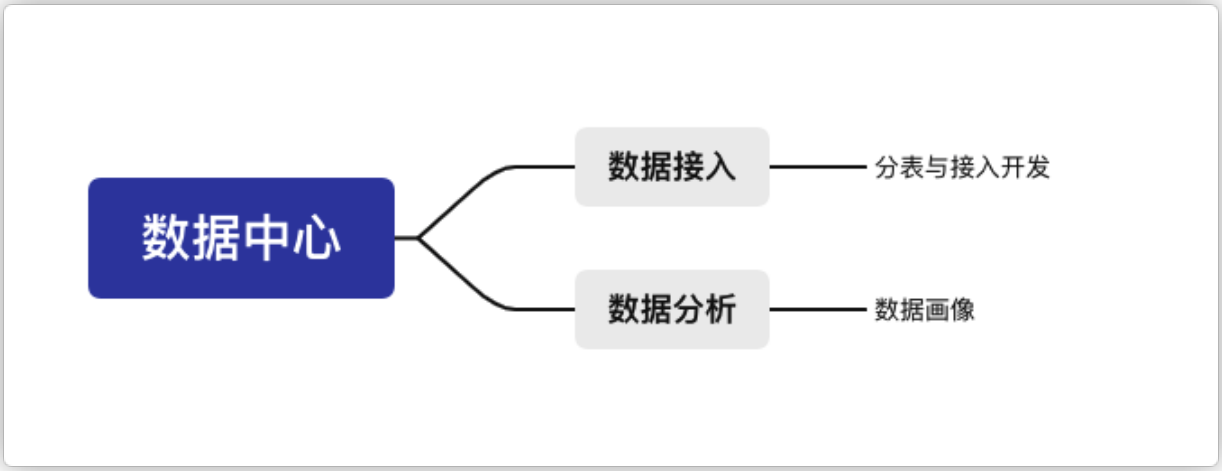
11月review

Created and last modified by 王文磊 on Dec 22, 2021

1. 需求总览



2. 需求拆解

2.1 分表与接入开发

需求背景

数据中心原有接入解决了以往需要联系数仓或平台同学并手动新建接入集群的问题，降低了沟通与任务创建的时间成本，

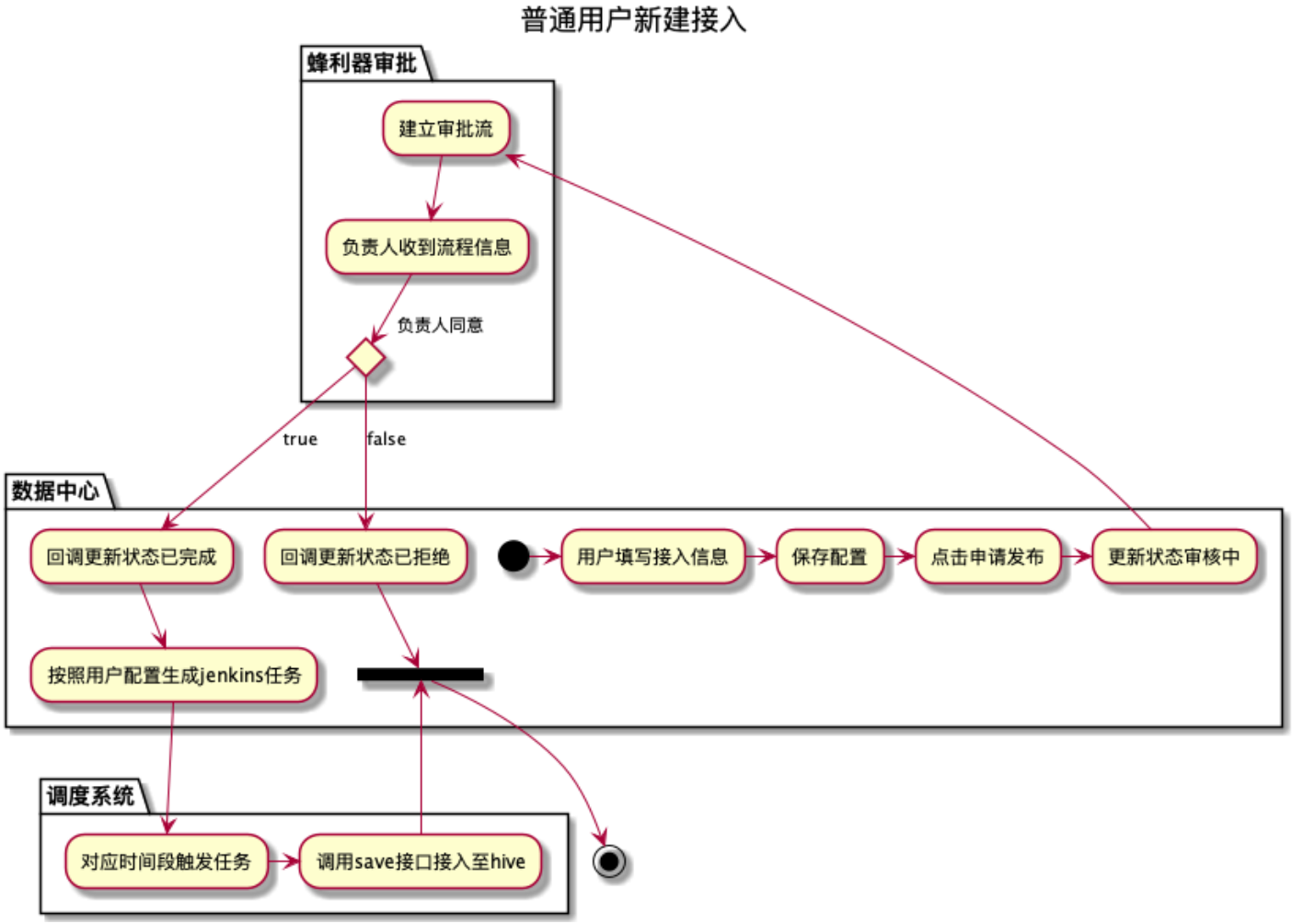
本需求目的在于扩大接入数据表的范围，并提供给用户配置更多接入信息的能力，如：

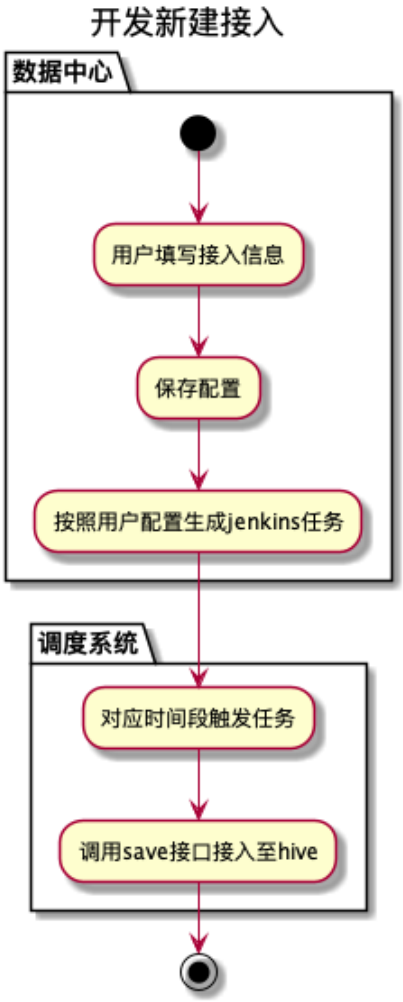
1. 支持 单/多库 分表
2. 自定义接入任务触发时间 段
3. 支持增量接入
4. 支持接入到当前时间（解决n + 1类型业务数据问题）
5. 支持调度错峰配置（通过默认值与时间 段）

其次是原dmp平台离线接入迁移（数据中心–接入开发）

整体功能

在线建立接入任务

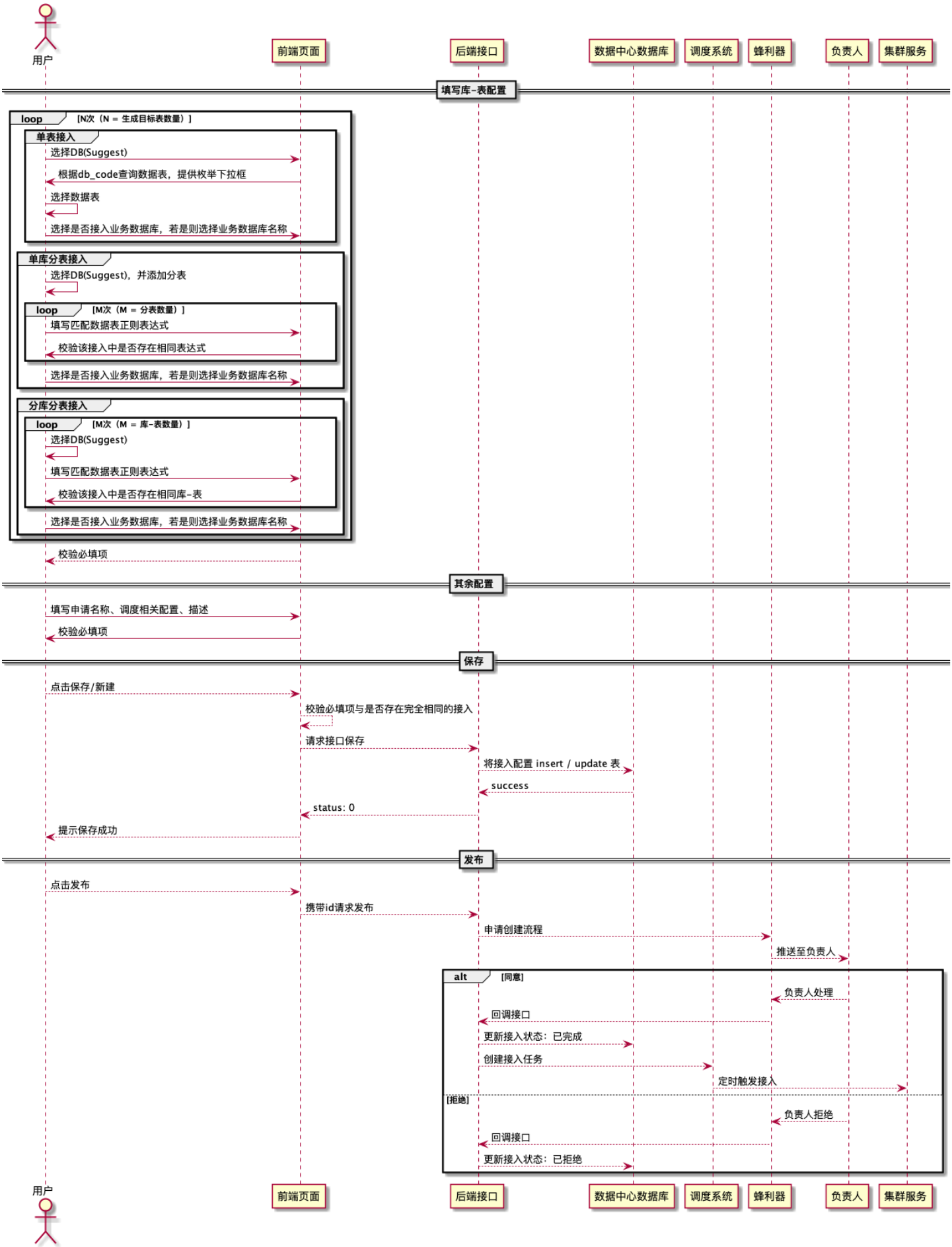

普通用户新建接入
蜂利器审批
建立审批流
负责人收到流程信息
负责人同意
true
false
数据中心
回调更新状态已完成
回调更新状态已拒绝
用户填写接入信息
保存配置
点击申请发布
更新状态审核中
按照用户配置生成jenkins任务
调度系统
对应时间段触发任务
调用save接口接入至hive


开发新建接入
数据中心
用户填写接入信息
保存配置
按照用户配置生成jenkins任务
调度系统
对应时间段触发任务
调用save接口接入至hive

核心拆解

配置接入

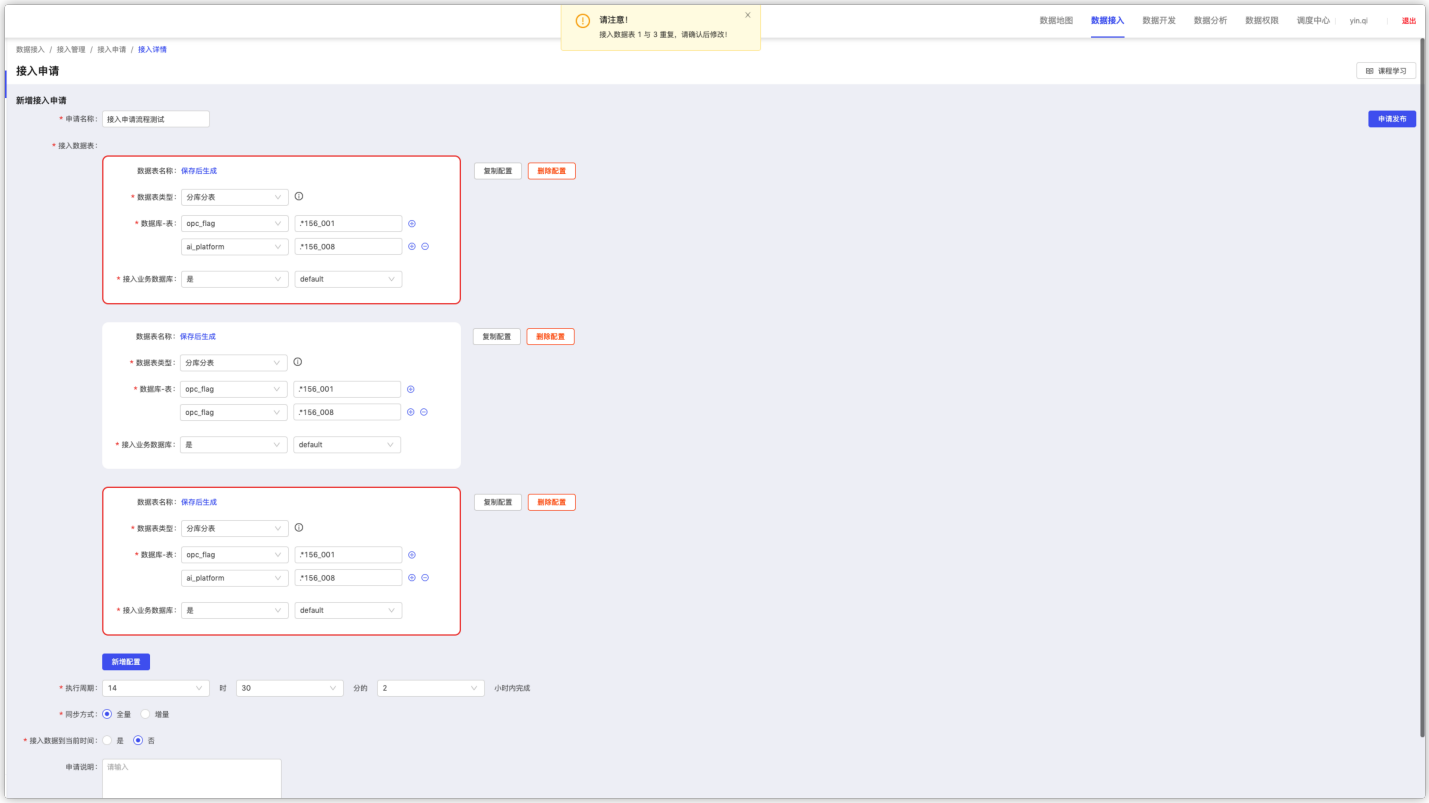
普通用户配置接入



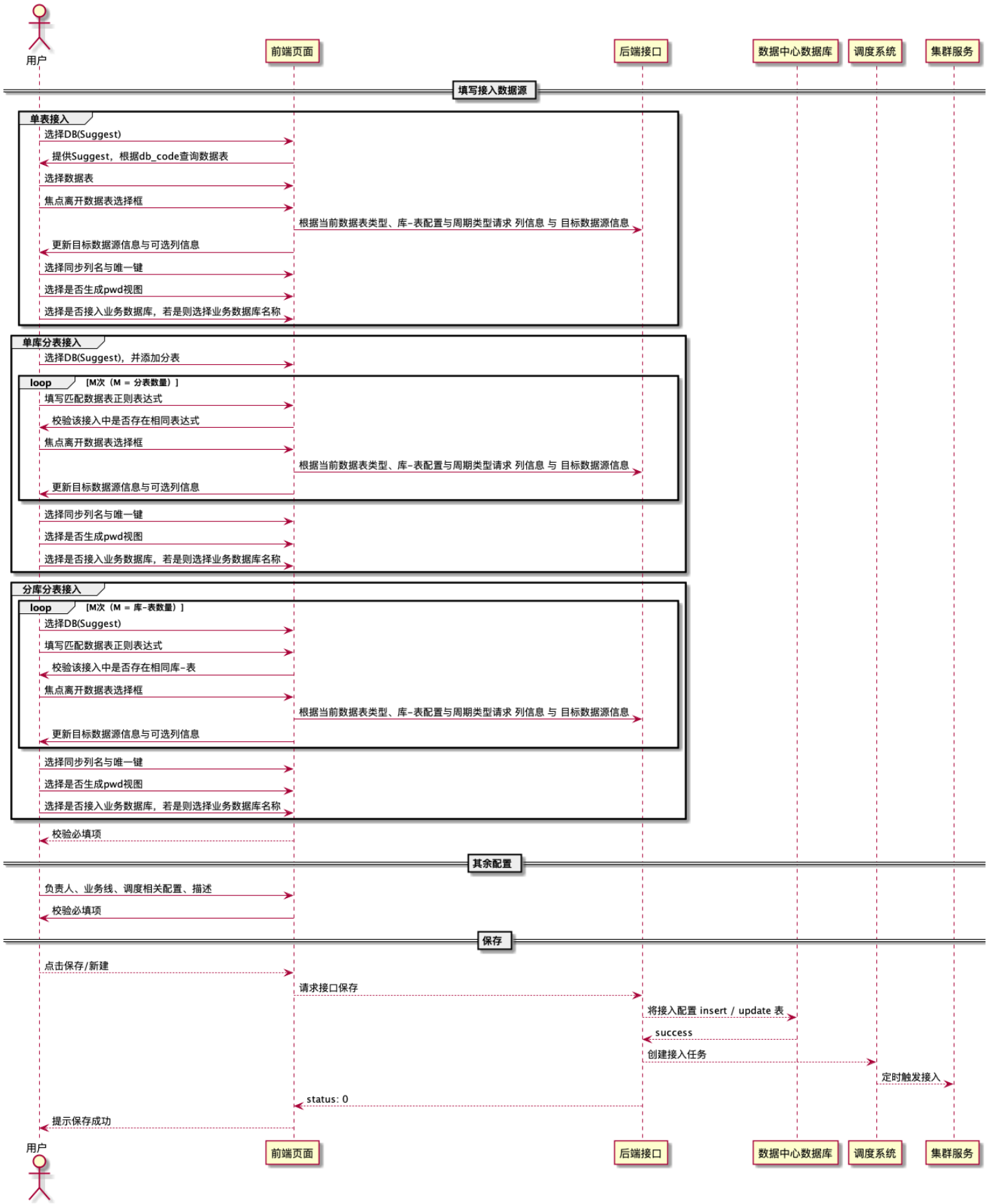
case：单库分表相同 正则表达式

case：分库分表接入相同 库-表

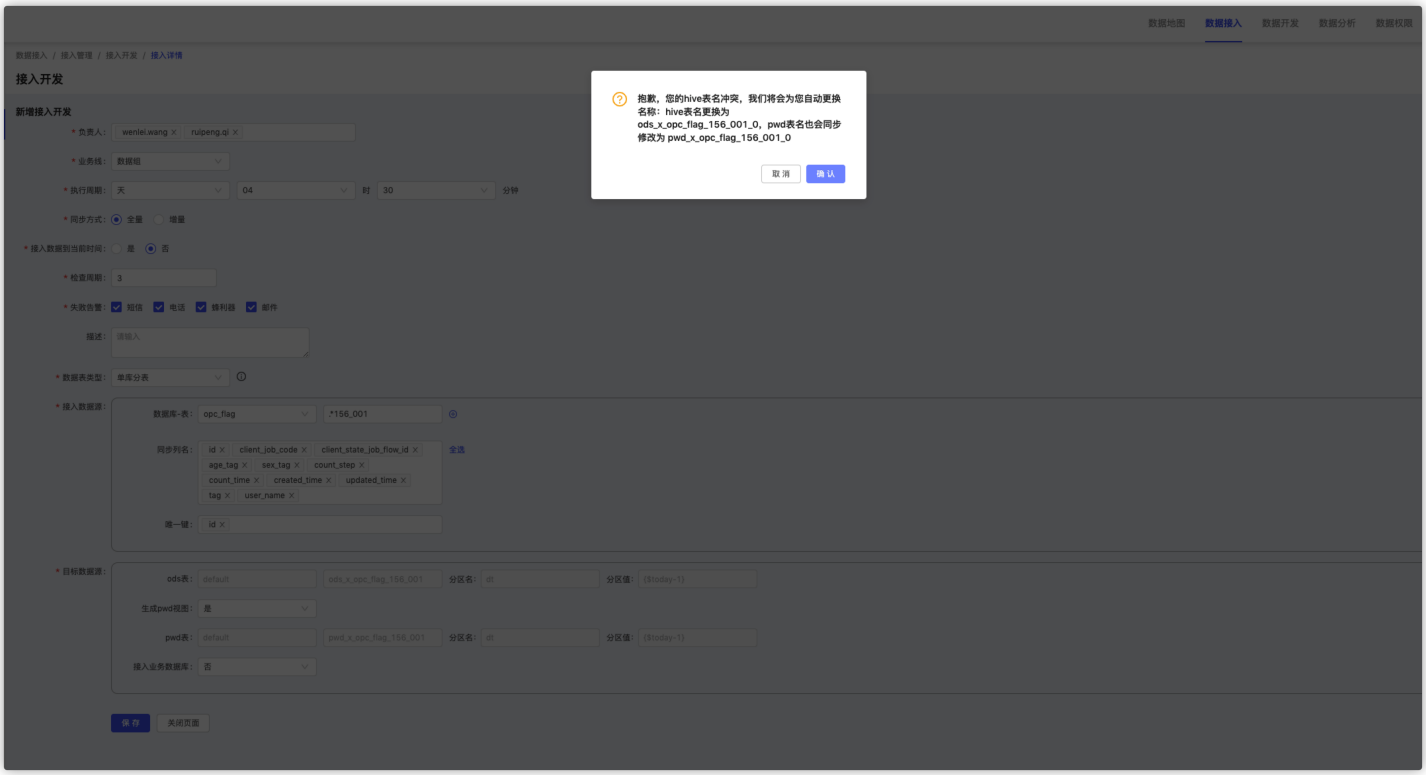
case：保存时存在 完全相同接入


请注意!
接入数据表 1 与 3 重复，请确认后修改!
数据地图
数据接入
数据开发
数据分析
数据权限
调度中心
yin.qi
退出
数据接入 / 接入管理 / 接入申请 / 接入详情
接入申请
课程学习
新增接入申请
* 申请名称：
接入申请流程测试
申请发布
* 接入数据表：
数据表名称：保存后生成
复制配置
删除配置
* 数据表类型：
分库分表
* 数据库-表：
opc_flag
*156_001
ai_platform
*156_008
* 接入业务数据库：
是
default
新增配置
* 执行周期：
14
时
30
分的
2
小时内完成
* 同步方式：
全量
增量
* 接入数据到当前时间：
是
否
申请说明：
请输入

开发配置接入



case: 接入Hive表名冲突



开发时产生的部分疑惑：

1. where to where?
2. 接入数据到当前时间作用
3. 同步列名与唯一键作用
4. 生成Pwd表与接入业务数据库实际触发动作

通过询问得到的解答：

1. 从所有数据库（包含业务数据库）接入到hive
2. 代表是否接入之前的数据，若选择是则不接入之前时间数据
3. 同步列名决定最终接入到hive里的表需要拥有哪些列，唯一键是提供给后端作为索引之后根据记录merge表的
4. 生成pwd代表在hive的default库中除了创建实际表外再添加一个pwd_开头的视图表，而接入业务数据库则是将上述视图表建立与目标业务数据库的映射关系

2.3 数据画像

需求背景

通过标签-人群，解决

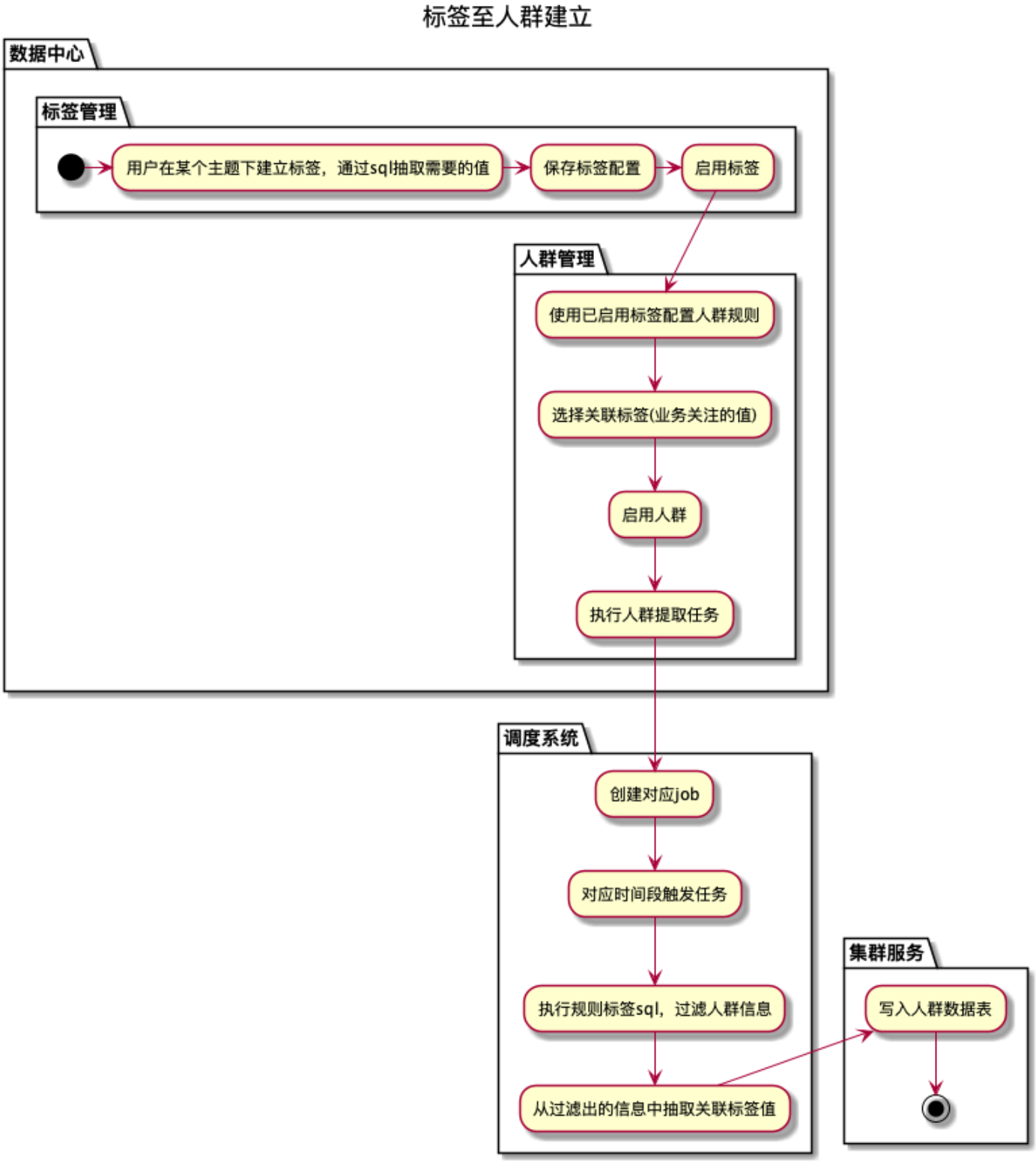
1. 用户直接使用pdw底层数据，无数据积累
2. 频繁调用明细数据进行数据查询，浪费资源

优化内容

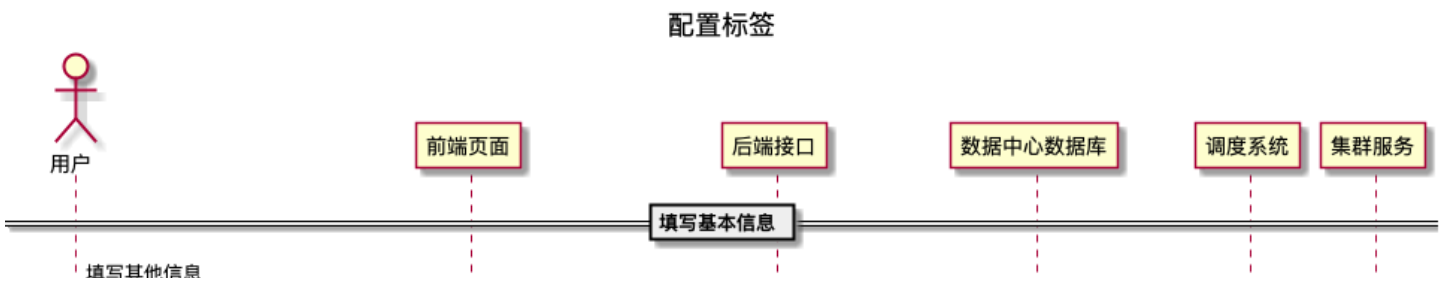
1. 通过标签配置，将指标进行细分细化
2. 通过人群，划分需要的数据范围
3. 通过附加字段，增加数据范围内的指标信息
4. 生成数据表，提供上层使用，降低生成中间数据的成本和门槛，提升使用数据的效率

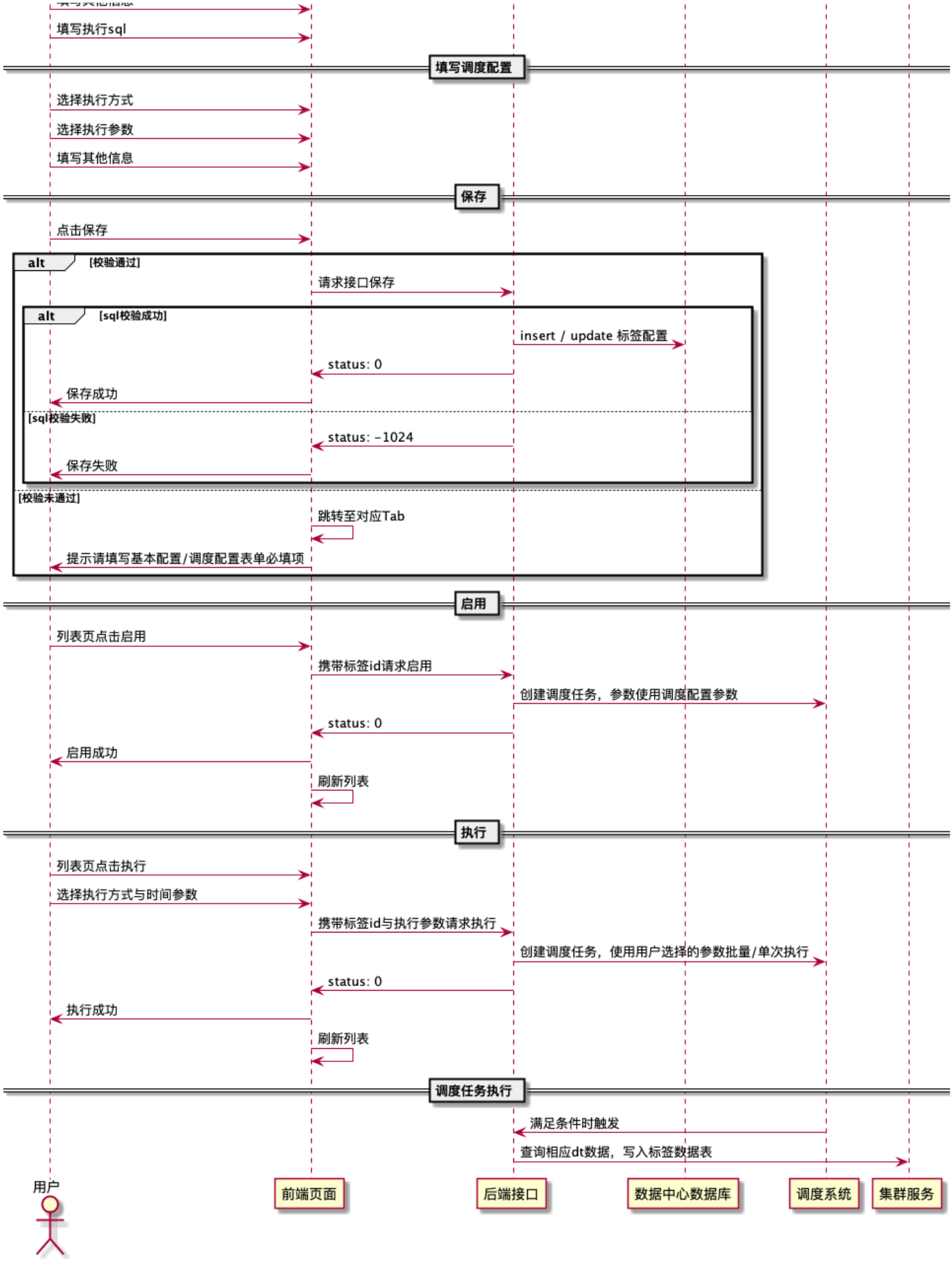
整体功能

建立标签供人群配置规则使用，输出人群数据表

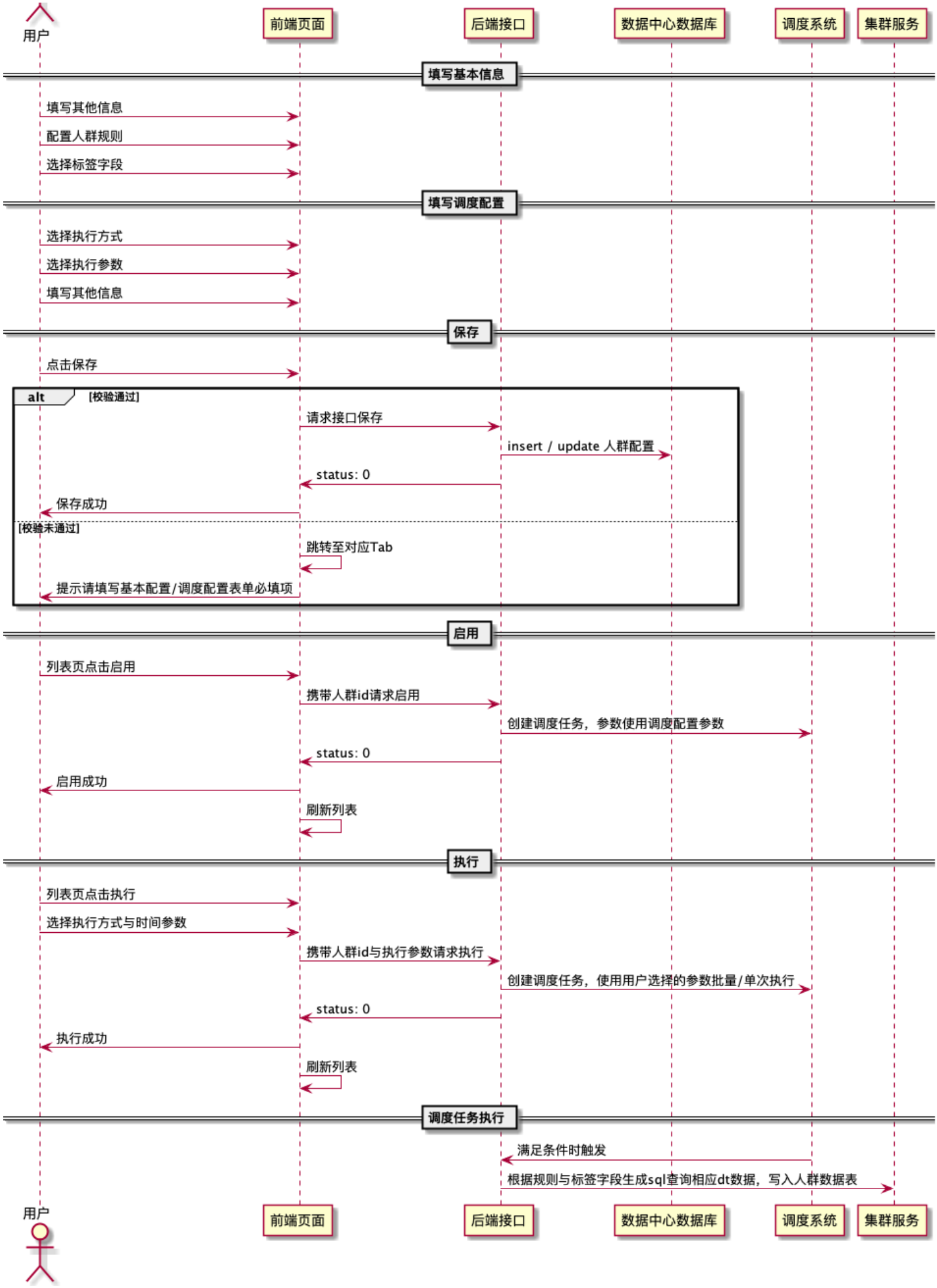

标签至人群建立
数据中心
标签管理
用户在某个主题下建立标签，通过sql抽取需要的值
保存标签配置
启用标签
人群管理
使用已启用标签配置人群规则
选择关联标签(业务关注的值)
启用人群
执行人群提取任务
调度系统
创建对应job
对应时间段触发任务
执行规则标签sql，过滤人群信息
从过滤出的信息中抽取关联标签值
集群服务
写入人群数据表

核心拆解

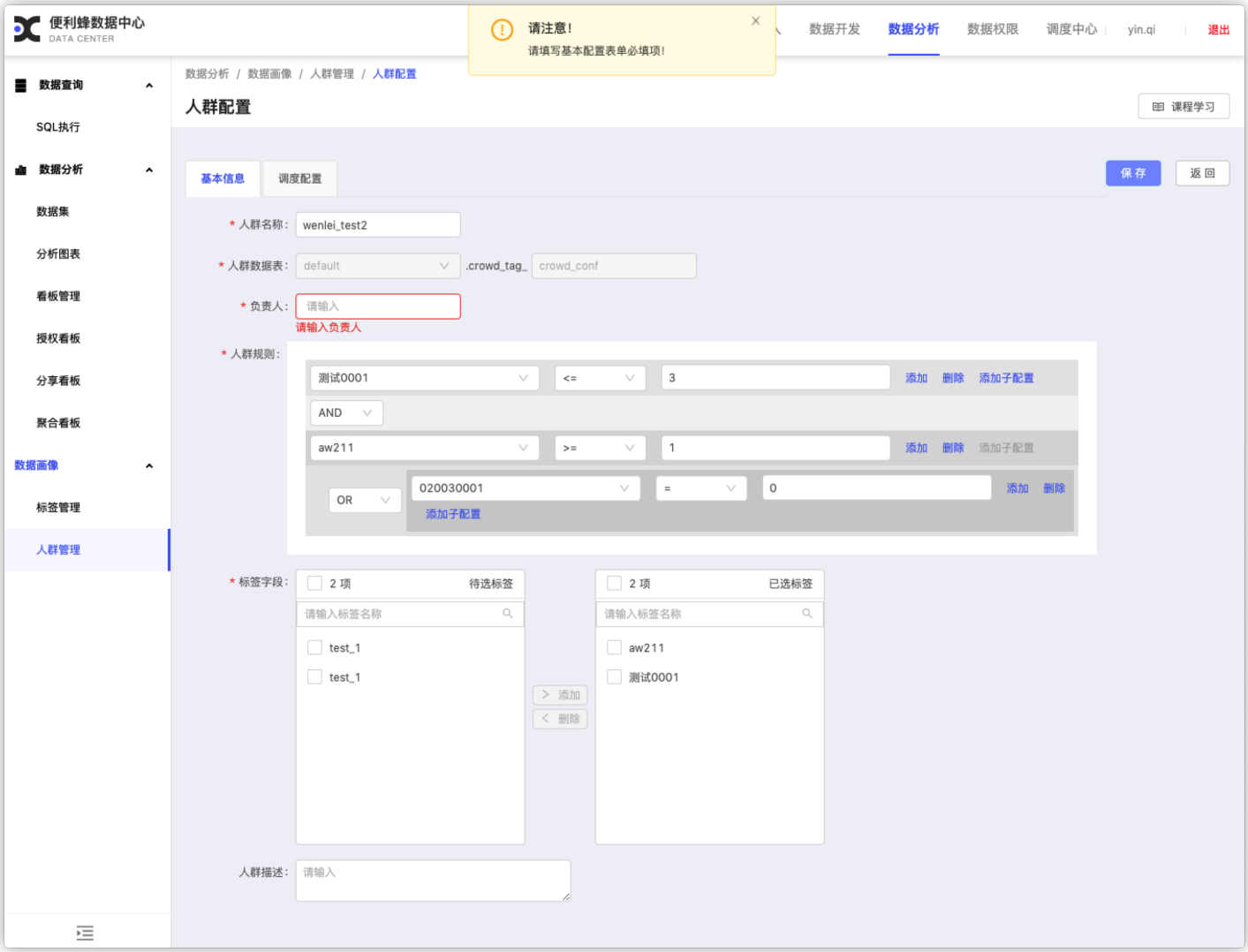

配置标签
用户
前端页面
后端接口
数据中心数据库
调度系统
集群服务
填写基本信息
填写其他信息

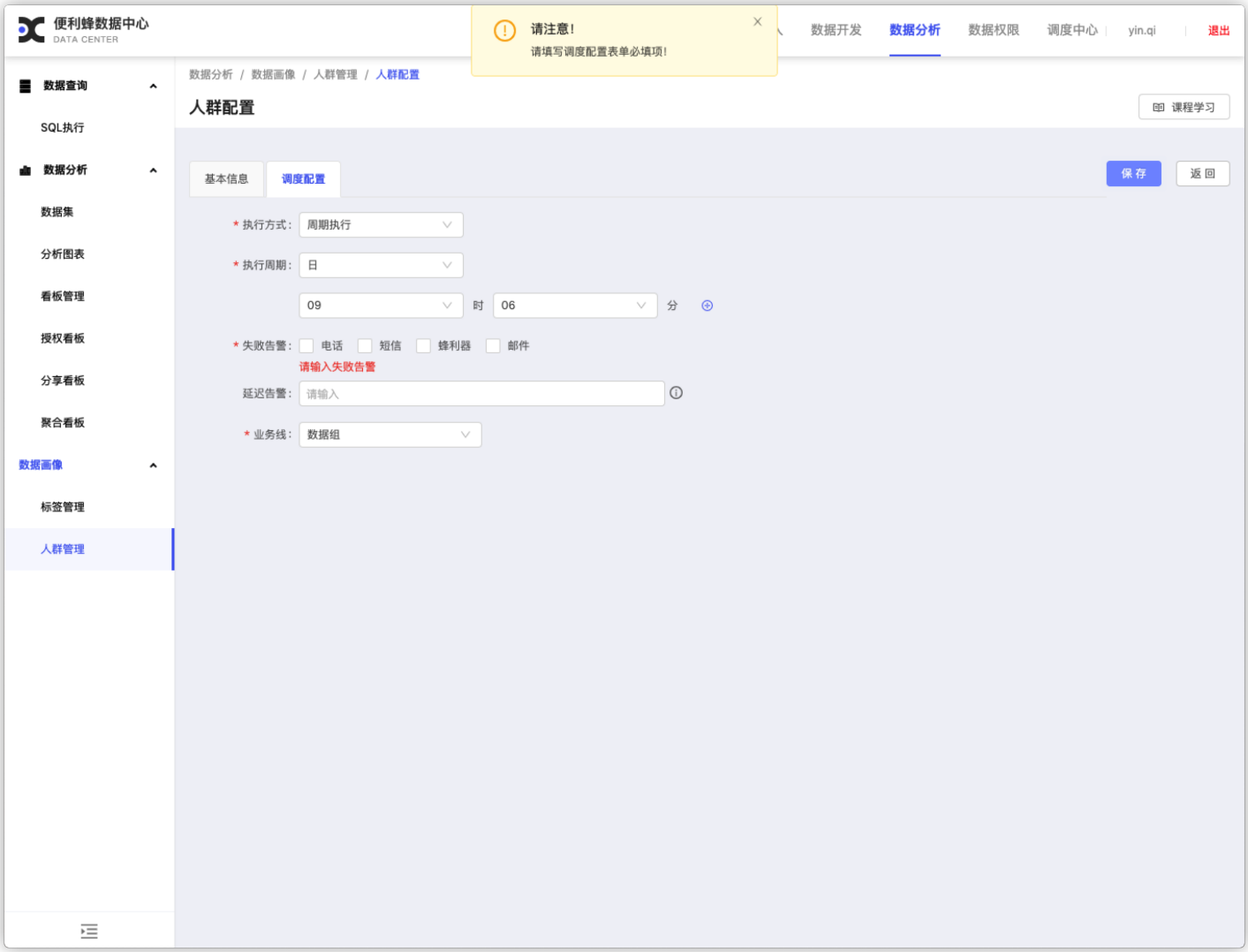

填写执行sql
填写调度配置
选择执行方式
选择执行参数
填写其他信息
保存
点击保存
alt
[校验通过]
请求接口保存
alt
[sql校验成功]
insert / update 标签配置
status: 0
保存成功
[sql校验失败]
status: -1024
保存失败
[校验未通过]
跳转至对应Tab
提示请填写基本配置/调度配置表单必填项
启用
列表页点击启用
携带标签id请求启用
创建调度任务，参数使用调度配置参数
status: 0
启用成功
刷新列表
执行
列表页点击执行
选择执行方式与时间参数
携带标签id与执行参数请求执行
创建调度任务，使用用户选择的参数批量/单次执行
status: 0
执行成功
刷新列表
调度任务执行
满足条件时触发
查询相应dt数据，写入标签数据表
用户
前端页面
后端接口
数据中心数据库
调度系统
集群服务

配置人群

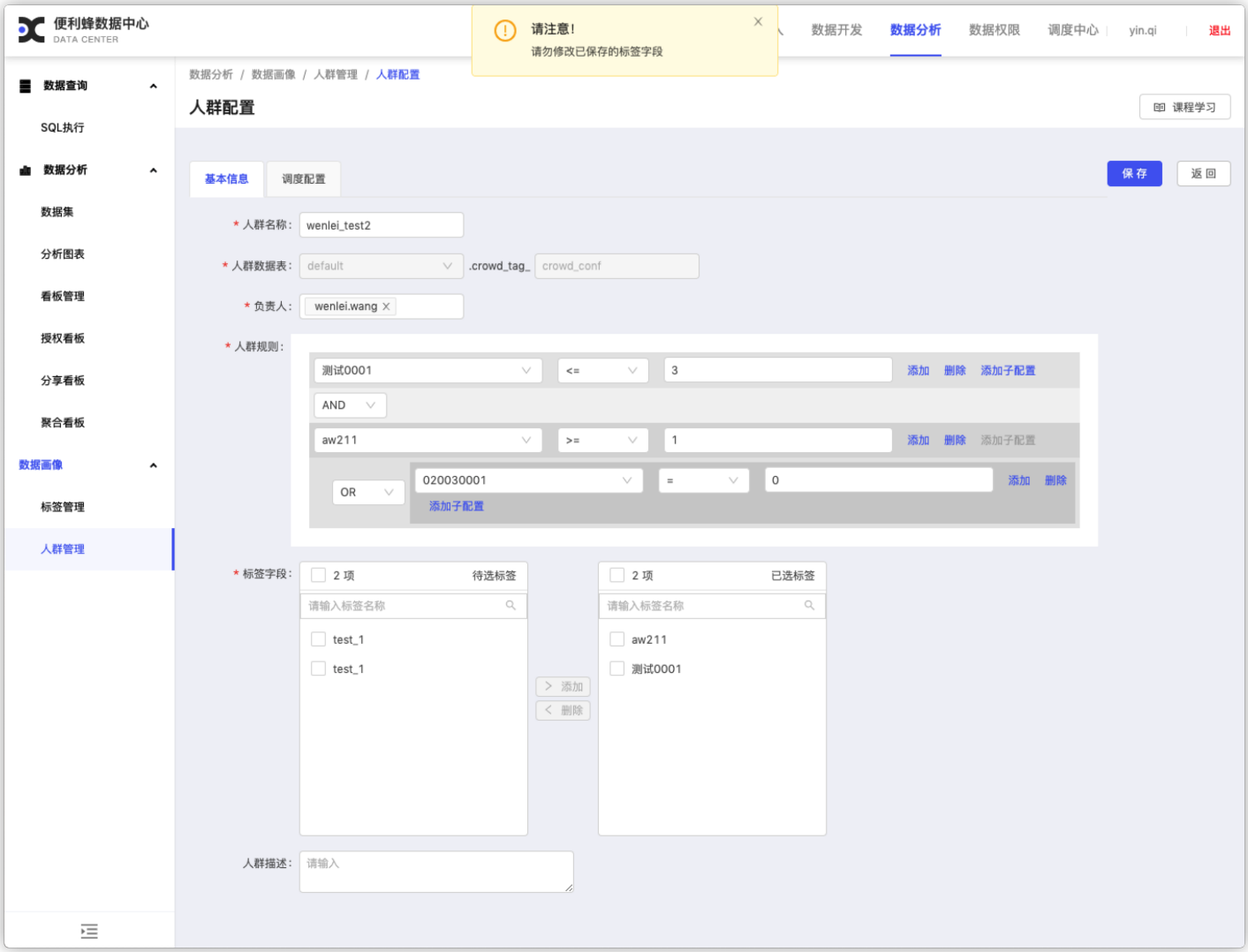


case 未完成某部分配置时点击保存 / 新建


便利蜂数据中心
DATA CENTER
请注意！
请填写基本配置表单必填项！
数据开发
数据分析
数据权限
调度中心
yin.qi
退出
数据查询
SQL执行
数据分析
数据集
分析图表
看板管理
授权看板
分享看板
聚合看板
数据画像
标签管理
人群管理
数据分析 / 数据画像 / 人群管理 / 人群配置
人群配置
课程学习
基本信息
调度配置
保存
返回
* 人群名称：
wenlei_test2
* 人群数据表：
default
.crowd_tag_
crowd_conf
* 负责人：
请输入
请输入负责人
* 人群规则：
测试0001
<=
3
添加 删除 添加子配置
AND
aw211
>=
1
添加 删除 添加子配置
OR
020030001
=
0
添加 删除
添加子配置
* 标签字段：
2 项
待选标签
请输入标签名称
test_1
test_1
添加
删除
2 项
已选标签
请输入标签名称
aw211
测试0001
人群描述：
请输入



case: 人群配置时删除已保存的标签字段



开发时产生的部分疑惑：

1. 创建的标签任务执行时会触发什么动作，是否会产生中间数据
2. 标签以及人群的调度配置只在执行时被使用吗？为什么执行还有选择时间，批量执行选择时间段和单次执行又有什么区别（调度选择周期执行的情况下）

通过询问得到的解答：

1. 执行时会触发标签sql的执行，会产生标签数据表
2. 调度配置产生的参数是调度系统构建任务时使用的，作用是指定条件执行，而单次执行与批量执行选择的时间/时间段是手动构建任务使用的，作用是立即执行，若选择批量则按天为维度构建多个立即执行的任务

3. 总结

1. 动手开发之前需要确保对需求有足够的理解，避免漏出
2. 对于case多的需求开发需要耐心，切莫顾此失彼
3. 对于不合理的实现方式需要敢于拒绝并提出解决方案

Like Be the first to like this