采石矶数据质量基础系统

用户手册

文档版本：2.2

深圳计算科学研究院

目录

注意

您购买的产品、服务或特性等应受深圳计算科学研究院商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，深圳计算科学研究院对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

产品特点

在全球全面利用大数据进行现代化治理的背景下，数据日益融入到政府决策、社会治理、公共服务、生产制造、消费流通等环节，低质量无序数据存在的危害风险越来越突出。根据美国数据仓库研究所统计数据显示：数据质量问题每年造成美国工业界大约6110亿美元的经济损失，约占美国GDP的6%，同时80%的公司都能从低质量的数据中发现重大的成本改进，因此解决数据质量的需求越来越大。但目前在全球市场上销售的数据质量主力产品80%以上是以ETL系统为主，其数据质量规则依赖人工经验配置，表达能力弱，只能支持单表、单行规则，对数据质量的提升非常有限。同时AI在数据质量方面还处于探索阶段，当前成熟的机器学习模型不具备解决数据质量的所有问题的能力，主流模型比如Bert，GPT等含有非常多参数，如何处理海量数据一直是一个难点。为此我们研发了采石矶系统解决上述问题，该产品以自动管理为特征，融合逻辑规则与人工智能，支持数据规则的自动挖掘、分析和推理，提供数据错误的自动监测和纠错等功能。

产品介绍

本系统在数据质量奠基性理论指导下，通过统一逻辑框架下规则和AI的结合，实现数据内部潜在规则自动发现。支持单表单行、单表多行、跨表规则的挖掘，并通过规则执行完成数据潜在错误的发现，提供确定性的修复建议，解决数据一致性、时效性、精确性、完整性和实体的同一性问题。面向集中式/分布式数据、关系型数据，打造具有可信数据采集，规则发现、数据查错、数据纠错、实体聚类、数据剖析、模型管理、规则管理等功能的一站式数据治理（数据质量）解决方案。

产品运行环境

序号	项目	详细信息
1	硬件环境	X86、ARM
2	后台软件环境	Centos7.x、麒麟V10
3	浏览器软件	Chrome 63版本及以上、Edge12版本及以上、 Firefox18版本及以上

产品最小验证硬件要求

序号	项目	数量	详细信息
1	CPU服务器	3台	CPU:2*18C 2GHz以上; Mem: 32G Disk: 1T以上 网卡：2个1G以上
2	GPU显卡服务器	1台	Tesla V100 32G PCIe Pas或等价显卡 、gpu驱动 >=450.80.02 、CUDA>=11

关键技术介绍

规则发现

规则发现主要用于发现数据中存在的规律。一般数据量增大后一些潜在的逻辑关系会被隐藏起来。需要花费大量人工去分析才能找出，而使用规则发现功能就可以轻松解决这样的问题。 规则发现分为CR规则发现和ER规则发现： CR规则是用于处理数据冲突错误的规则；ER规则是用于处理数据实体一致性问题的规则。采石矶根据用户输入的挖掘偏好设置（用户可以根据需要选择机器模型或相似度算法），自动进行数据分析，输出规则，用户根据业务背景挑选适合实际场景的规则，便于后续的数据质量提升做准备。

数据查错

在指定数据集上基于规则（可以是规则发现输出的规则，也可以是用户自定义规则）进行查错，将不满足规则的数据识别出来，方便用户进一步分析或处理。数据查错可以在原始数据中找到数据冲突的规则或者规则集合，通过反复迭代的执行这些规则，最终发现数据中所有的冲突（包括数据一致性、完整性、准确性）。 数据经过查错规则的执行，查错结果以通知的形式反馈给用户进行查看。查错可以针对全量和增量的数据进行处理；查错的规则统称为REE规则，包含FD（函数依赖）、CFD（条件函数依赖）、MD（匹配依赖）、DC（拒绝约束）规则，同时支持机器学习模型的运行(如上图中ML谓词的推理运行)，扩展逻辑规则的能力，提供语义层的识别能力。 通过查错规则的运行，用户能够得到数据中相关于查错规则的所有冲突和错误信息，这个信息会以结果的形式标记出来，供用户参考。

数据纠错

数据纠错针对大数据质量问题中数据冲突的问题，主要解决数据的准确性问题。

在指定数据集上基于规则（可以是规则发现输出的规则，也可以是用户自定义规则）进行纠错，对不满足规则的数据进行自动修复，用户对自动修复后的数据进行错误的修改和冲突的确认后输出修复结果。通过数据纠错，用户能够得到错误和冲突被纠正后的数据。

实体聚类

实体聚类针对大数据质量问题中实体不一致的问题，主要解决不同系统中同一实体的记录如何关联的问题。

在指定数据集上，基于实体规则（可以是规则发现输出的规则，也可以是用户自定义规则）进行实体聚类，可以找出数据中属于同一实体的数据，将分散的实体信息关联到一起。

常见场景介绍

规则发现场景

XX新能源汽车有130+传感器，通过采石矶系统的规则发现功能在大量的传感器数据发现部分传感器之间的逻辑关联关系，从中提炼出数十条满足客观逻辑的规则（cr规则），在后续执行中帮助客户有效补齐缺失数据，提高了数据的完整性和正确性，得到客户的好评。xxx药协会拥有从上世纪90年代至今的所有药物数据和各个3甲医院的所开具的所有药物清单，但是由于各地写法和药物计量的不同，医院的清单和药物数据难以准确匹配，采石矶系统利用规则发现功能对数据之间的关联关系进行挖掘，根据药名和计量、价格等相关数据，发现多条实体属性规则（er规则），能够有效说明清单数据和药物数据之间的关联，能够将医院的药物清单和药物数据准确的匹配到一起，准确率达到90%以上，顺利地帮客户解决了数据一致性问题，节省了大量人力。

数据查错场景

xx运营商拥有海量的宽带签约用户地址信息和机房资源点的地址信息，但是由于数据大部分为早期手工录入，地址数据存在格式不规范和内容不准确的问题，严重影响现场客户维护和系统维护。采石矶系统先通过地址标准化功能，将地址类数据统一为同一个标准格式，再根据规则发现中发现的规则对数据进行查错，找出大量同一地点但是地址不同甚至冲突的数据，提供给客户，帮助运营商找到数据有误的用户，方便进一步确认。

数据纠错场景

xx银行，有大量账户地址信息，但是由于数据大部分为早期手工录入，地址数据存在格式不规范和内容不准确的问题，严重影响账户维护和系统维护。采石矶系统先通过地址标准化功能，将地址类数据统一为同一个标准格式，再根据规则发现中发现的规则对数据进行纠错，找出大量同一地点但是地址不同甚至冲突的数据，提供给客户，在客户进一步确认后，输出准确度更高的地址信息。

实体聚类场景

xx快递公司，有大量的月结企业客户信息，但很多客户公司的信息填写的不够准确或者甚至还有错误。客户希望借助天眼查等机构的企业标准信息数据，对客户数据进行补充和校正。采石矶系统使用实体聚类的方案，利用从数据中挖掘得到的ER规则，对客户数据和天眼查数

据进行匹配，然后将匹配结果以数据对的方式输出给用户，大大降低了客户手动比对数据的工作量，至少减少了20人-2月的工作量，显著提升了客户数据处理的效率。

最优记录场景

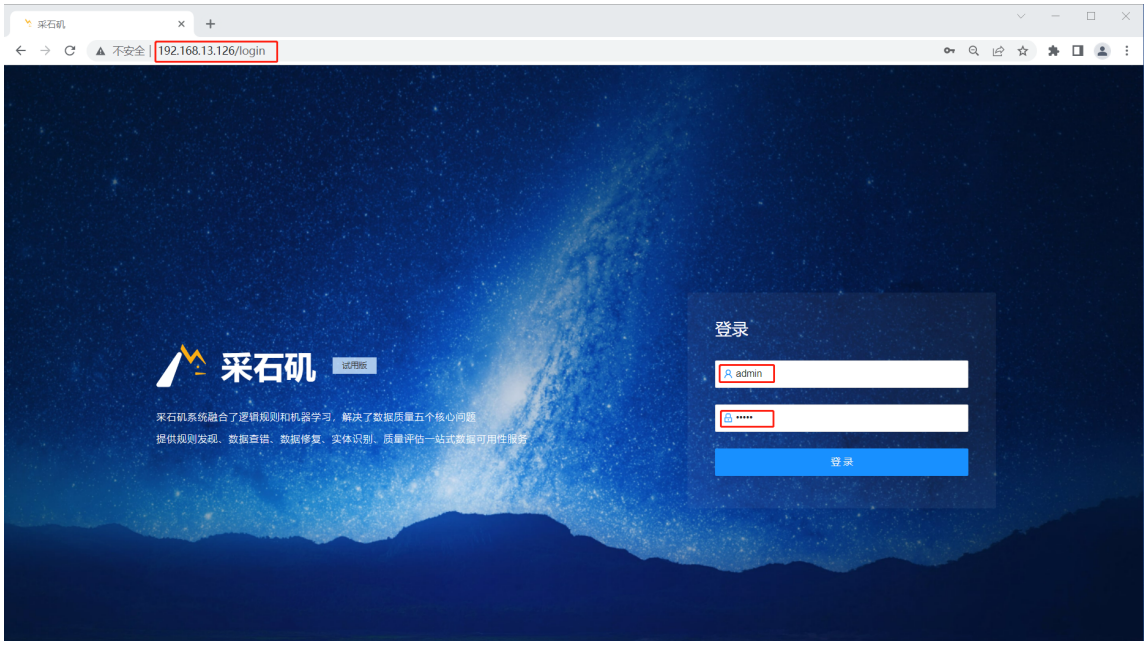
xxx政府大数据中心有大量的企业注册信息，但是由于企业经常需要更新或者注册信息，导致数据库中存有大量的过期信息，并且由于录入的不标准，很多数据存在误录入的情况，这样的情况大大的增加的数据的维护成本，同时也降低了数据管理的效率。采石矶系统的实体聚类功能根据企业名和企业的其他参照标签，成功将同一企业的数据识别成为同一实体后，再由最优记录功能，根据客户需求选择最新的数据为最优数据，推荐出最符合要求的数据，迅速高效地解决了数据冗余和过期的问题，节省了大量的人力物力。

字段匹配

xx物流公司，由于公司的发展迅速，每日数据量激增，同时由于数据管理的不完善，加上表格数据的多次复用，导致数据无法追溯血缘。由于没有数据血缘管理，出现了比较严重的数据一致性的问题。通过采石矶字段匹配功能，扫描多表的数据内容，利用算法发现出字段关联度高的数据，在海量表格中，寻找出各个表格之间的关联，并通过人工确认的方式，最终确认数据的血缘关系。协助客户解决了历史遗留的数据问题，并完成了一数一源的数据改革，提高了客户的数据管理效率，得到了客户的好评。

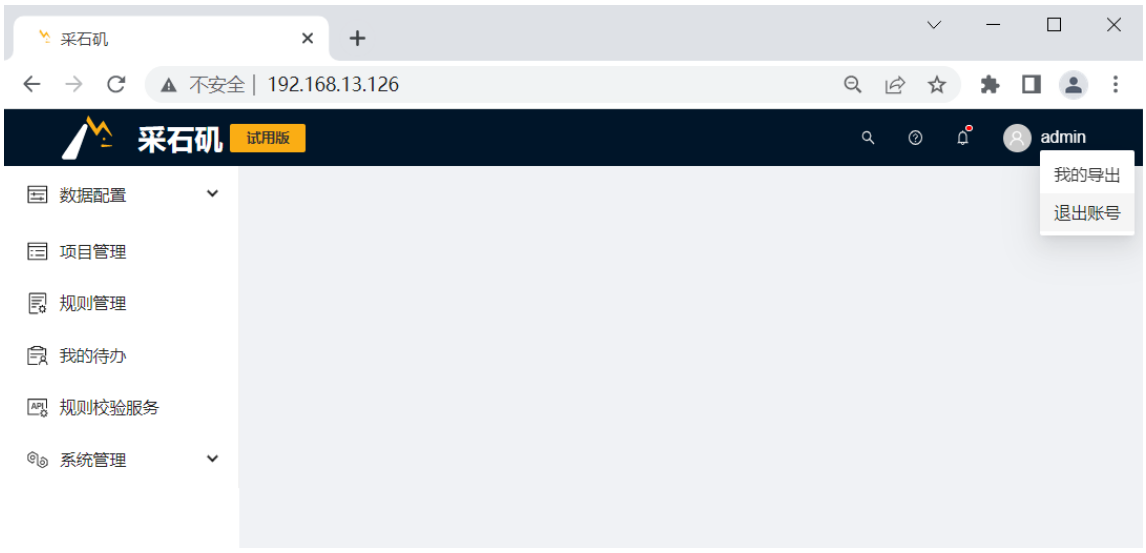
登录和退出系统

在满足版本要求的浏览器中（Chrome 63版本及以上、Edge12版本及以上、 Firefox18版本及以上），输入地址，即ubi服务所在节点ip，输入用户名和密码，默认是admin/admin，如下图所示。



登录系统

登录成功后， 退出系统界面如下图所示。



退出系统

用户指南

本文详细讲解采石矶系统的各项功能。包括”数据源“、”数据集“、”数据剖析“、”可信度管理“、”项目/工作流管理“、”规则发现“、”查错“、“数据纠错”、”实体聚类“、”最优记录“、”字段匹配“、”规则管理“，最后通过实际案例进一步说明，帮助用户学习使用采石矶系统。

数据源

本章节主要介绍采石矶系统对接外部数据库数据源和文件数据源的主要方法和流程。

本系统定义数据源为采石矶系统获取数据的源头，外部数据需要进行处理和分析之前需要先将数据导入采石矶系统内部，从而保证数据能够在采石矶系统内得到充分的分析。

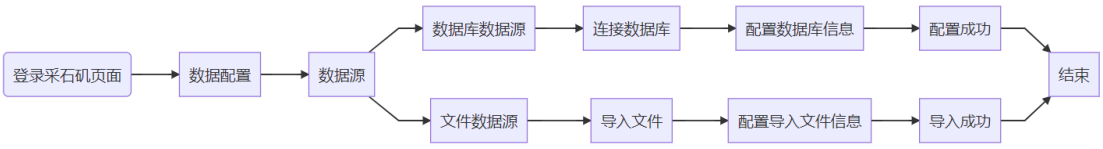
通过完成本章节步骤，可以了解到可以通过多种方式将外部数据导入到采石矶系统。

前置条件

须同时满足以下两个条件：

- 正常安装采石矶系统；
- 采石矶系统可以正常登录。

数据源操作流程图



数据源操作流程图

数据库数据源操作说明

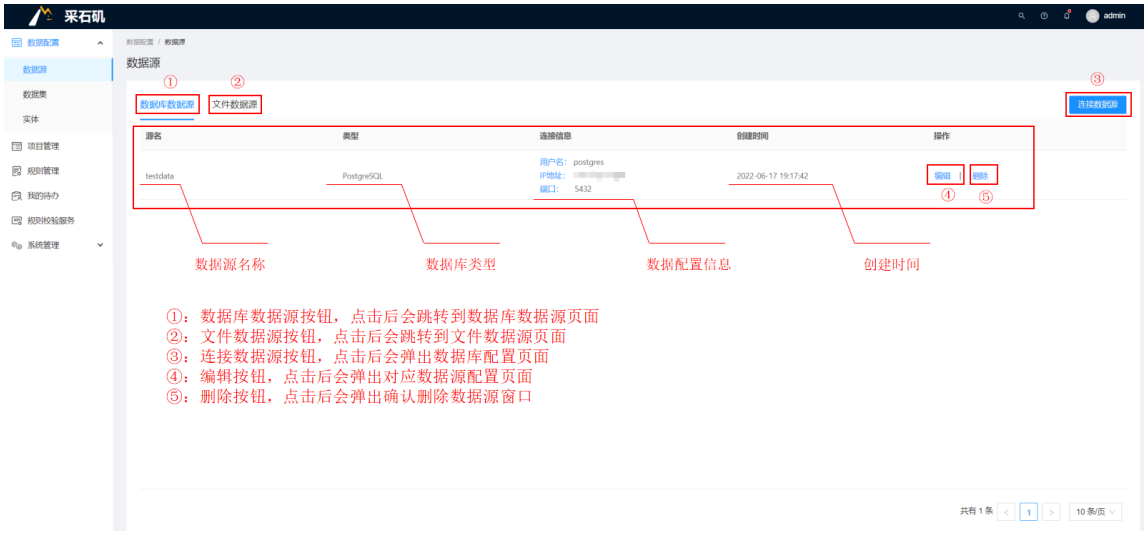
本章主要讲解数据库数据源操作说明。

数据库数据源操作流程图如下图所示。

数据库数据源操作流程图

1. 数据源页面简介

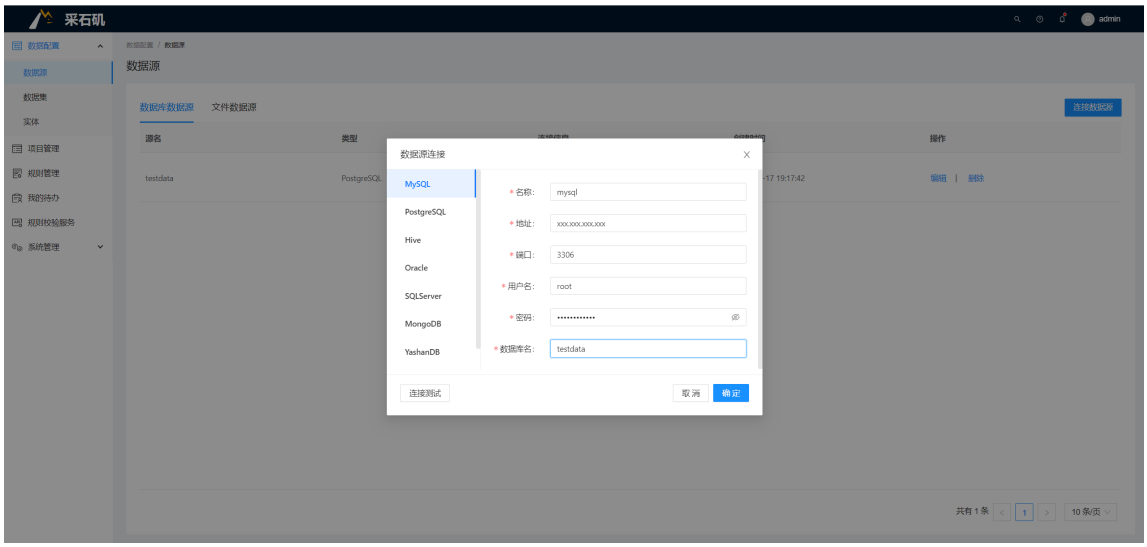
点击 `数据配置` 按钮，选择 `数据源` 按钮，会看到数据源页面，具体呈现如下图。



数据库数据源界面

2. 数据源连接

点击 连接数据源 按钮，弹出添加数据源窗口



数据源配置界面

目前采石矶系统支持如下数据库

数据库类型	数据源
关系型数据库	Mysql、PostgreSQL、Oracle、Microsoft SQL Server
大数据数仓存储	Hive、Hbase
NoSQL数据存储	MongoDB
国产数据库	YashanDB、GaussDB

数据源连接详细说明

选项	配置说明	必要
名称	自定义对应数据库的别名（仅支持中文、字母、数字、下划线）	是
地址	对应数据库的IP地址	是
端口	对应数据库的端口	是
用户名	对应数据库的用户名	是
密码	对应数据库的密码	是
数据库名/schema（hive）/服务名（oracle）	对应数据库的库名/schema/服务名	是
Zookeeper Quorum（Hbase）	对应数据库的IP和端口	是
Zookeeper Base Path（Hbase）	对应数据库的Base Path	是

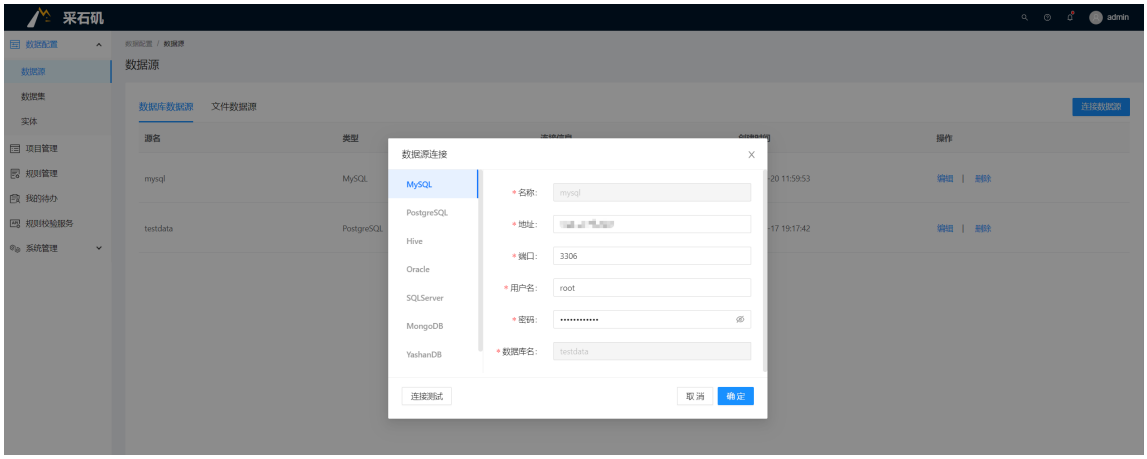
- 完成数据源信息配置以后，点击 `连接测试` 按钮，页面窗口会返回连接测试情况。如果不成功请检查数据源配置。

- 完成配置后点击 `确定` 按钮完成数据源配置，配置完成后数据源列表会新增刚刚配置的数据源。

新增数据源界面

3. 数据源编辑

如果需要修改已配置的数据源，可以点击数据源页面的 `编辑` 按钮，点击后会弹出修改数据源的窗口，修改完成后点击 `确定` 按钮即可。

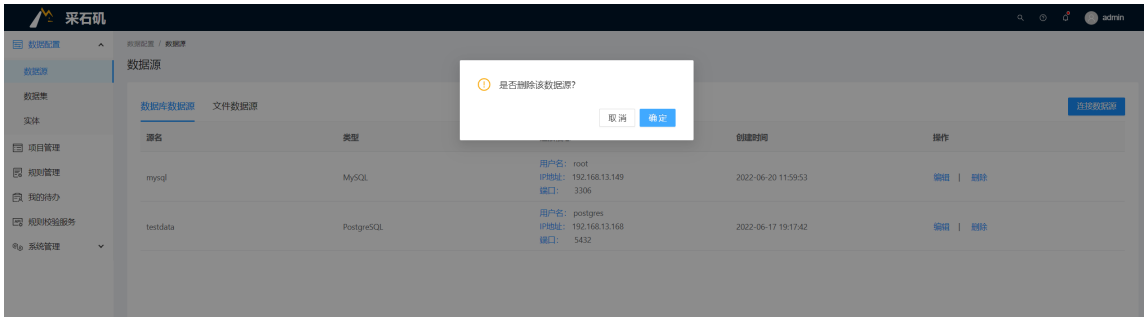


编辑数据源界面

选项	配置说明	是否可以修改
名称	自定义对应数据库的别名（仅支持中文、字母、数字、下划线）	否
地址	对应数据库的IP地址	是
端口	对应数据库的端口	是
用户名	对应数据库的用户名	是
密码	对应数据库的密码	是
数据库名/schema（hive）/服务名（oracle）	对应数据库的库名/schema/服务名	否
Zookeeper Quorum（Hbase）	对应数据库的IP和端口	是
Zookeeper Base Path（Hbase）	对应数据库的Base Path	是

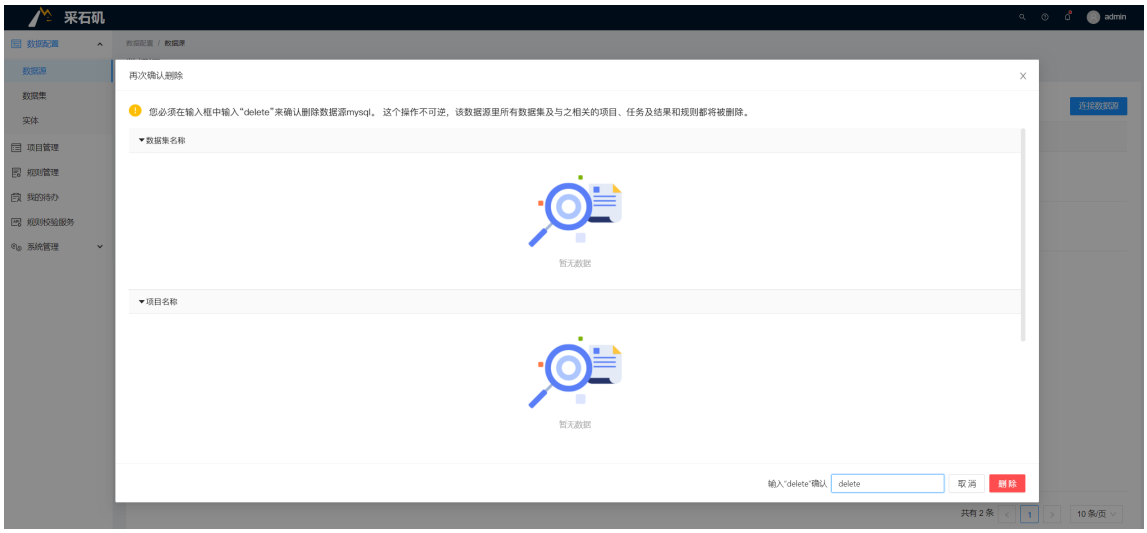
4. 数据源删除

点击数据源的 `删除` 按钮，弹出是否删除该数据源的提示窗口，点击 `确定` ，则弹出再次确认删除的弹窗，点击 `取消` 则不删除，并返回数据库数据源页面；



删除数据源界面

在再次确认删除的弹窗中可看到该数据源相关的数据集名称、项目名称以及规则内容，如数据源没有关联，则显示暂无数据。在弹窗右下角的输入框中输入 `delete` ，并点击 `删除` ，则删除数据源；在弹窗右下角点击 `取消` ，则不删除数据源，并返回到数据库数据源页面。删除数据源是不可逆操作，需谨慎操作。

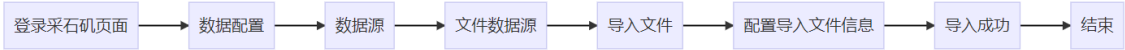


删除数据源界面

文件数据源操作说明

本章主要讲解文件数据源操作说明。

文件数据源的操作流程图如下图所示。



文件数据源操作流程图

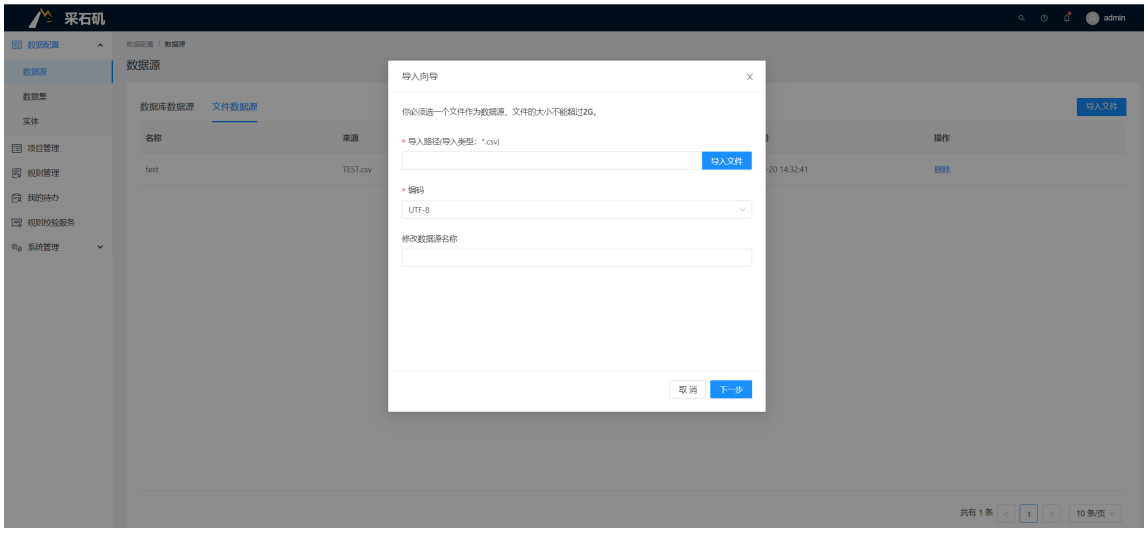
1. 文件数据源页面简介

文件数据源界面

- 在导入文件数据源时，如果没有修改数据源名称，则数据源名称与导入文件名称一致，如有修改，则显示修改数据源名称。
- 数据源来源显示导入的文件名
- 数据源状态为 `导入成功` ，则表示文件导入成功；
- 数据源状态为 `导入失败` ，则表示文件导入失败；
- 数据源创建时间为文件导入时间；

2. 文件数据源导入

a. 在文件数据源页面，点击 `导入文件` 按钮，弹出导入向导窗口。

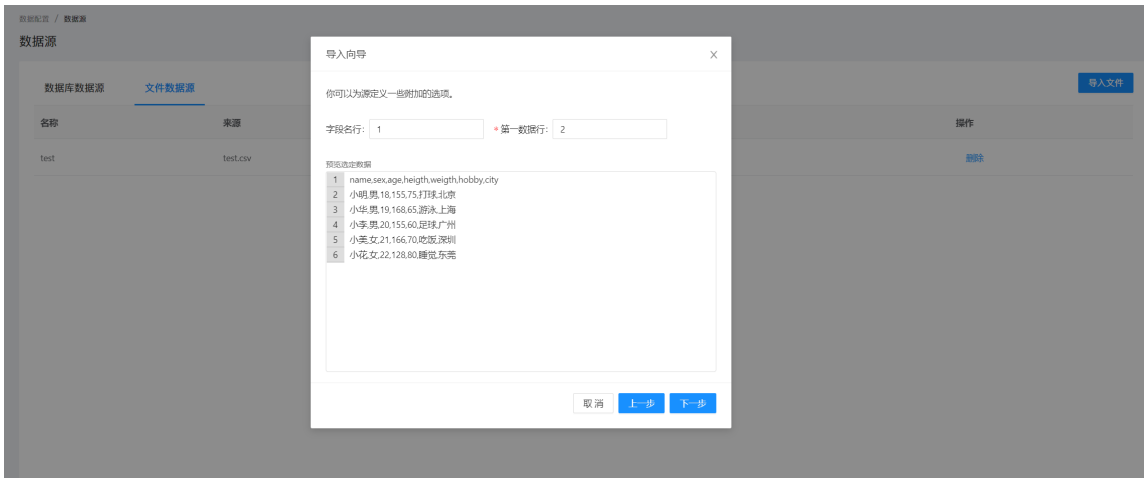


文件数据源导入向导界面

b. 点击 `导入文件` ，选择需要导入的文件。完成后点击 `下一步` ，进入选择字段名和数据行页面；点击 `取消` 则会取消文件数据源的导入。

- 当前采石矶系统支持导入的文件最大不能超过2G，且只支持csv格式的文件。
- `编码` 下方的下拉框，可以选择当前csv文件的编码格式，当前支持UTF-8、GBK、GB2312、Unicode四种编码格式，默认为UTF-8格式。
- （选填） `修改数据源名称` 输入框为用于修改导入文件的名称，不填则默认数据源名称为导入文件的文件名，文件数据源名称不能重复。

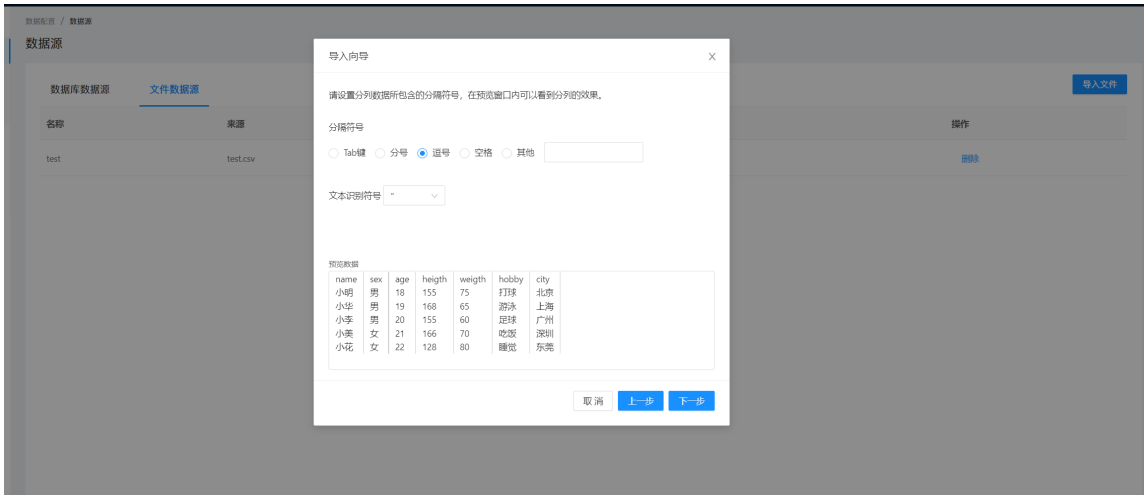
c. 在选择字段名和数据行页面，可看到预览的部分数据，根据实际情况填写后点击 `下一步` ，进入到设置分隔符页面。点击 `上一步` 则回到选择文件页面，点击 `取消` 则会取消文件数据源的导入，返回文件数据源页面。



文件数据源导入向导界面

- （选填）字段名行输入框填写导入文件的字段名所在行；
- （必填）第一数据行输入框填写导入文件的第一行数据，不包括字段名。

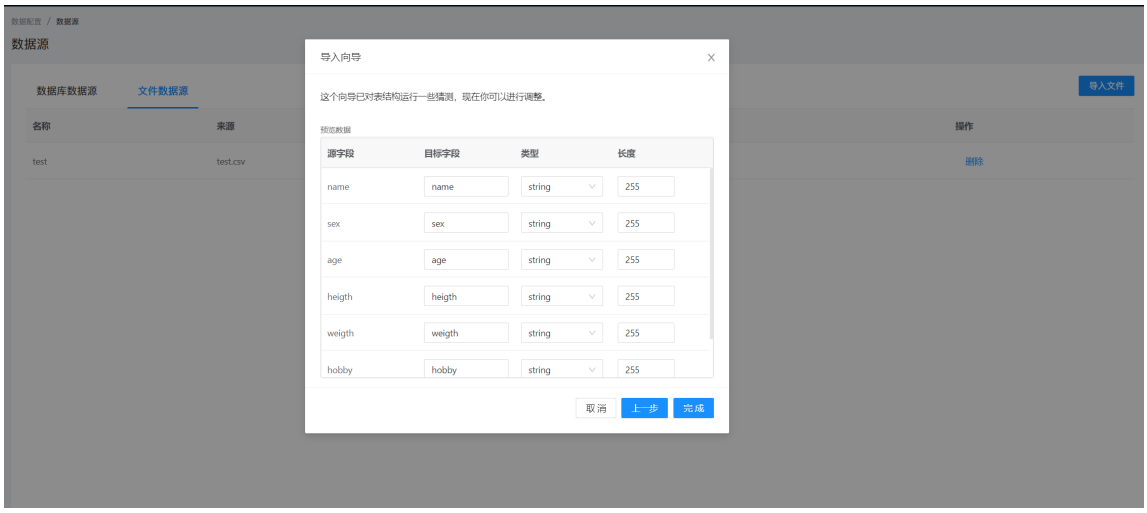
d. 在设置分隔符页面，根据实际情况选择分隔符号和文件识别符号。填写完成后，点击 `下一步` ，进入字段调整页面，点击 `上一步` 则回到选择字段名和数据行页面，点击 `取消` 则会取消文件数据源的导入，返回文件数据源页面。



文件数据源导入向导界面

e. 进入字段调整页面，字段的类型默认为string，长度默认为255，根据实际情况进行调整；调整完成后，点击 完成 ，即可完成创建；点击 上一步 则回到设置分隔符页面，点击 取消 则会取消文件数据源的导入，返回文件数据源页面。

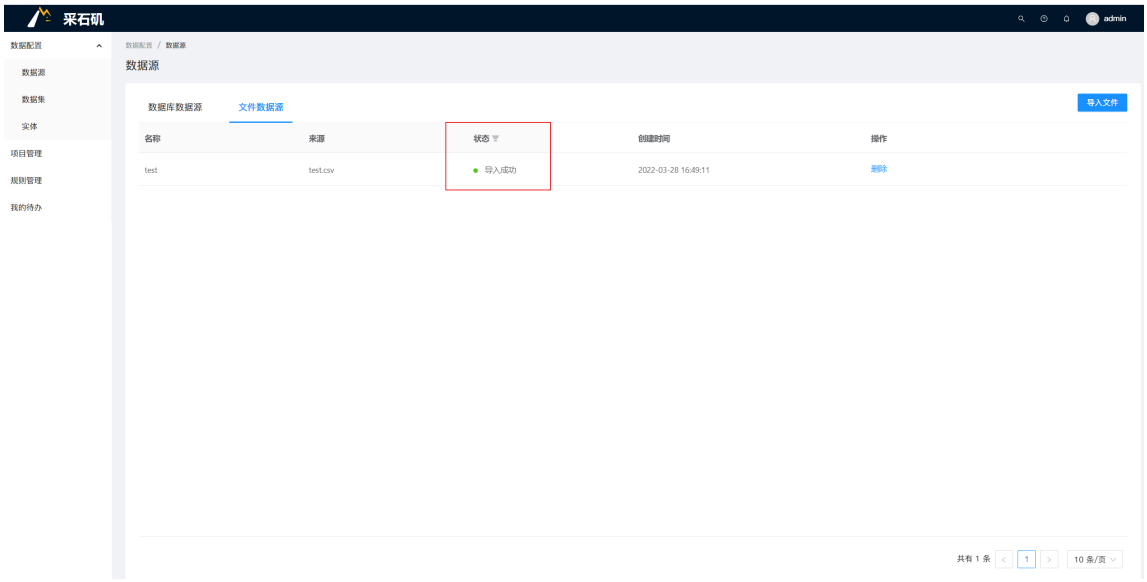
- 如果在选择字段名和数据行页面，填写了字段名行，此处的源字段为选择的字段名行中的字段，目标字段默认与源字段一致，可以进行修改，然后根据实际情况调整字段的类型和长度；



文件数据源导入向导界面

- 如果在选择字段名和数据行页面，没有填写字段名行，此处则没有源字段，需在目标字段列的输入框中，手动输入目标字段，然后根据实际情况调整字段的类型和长度。

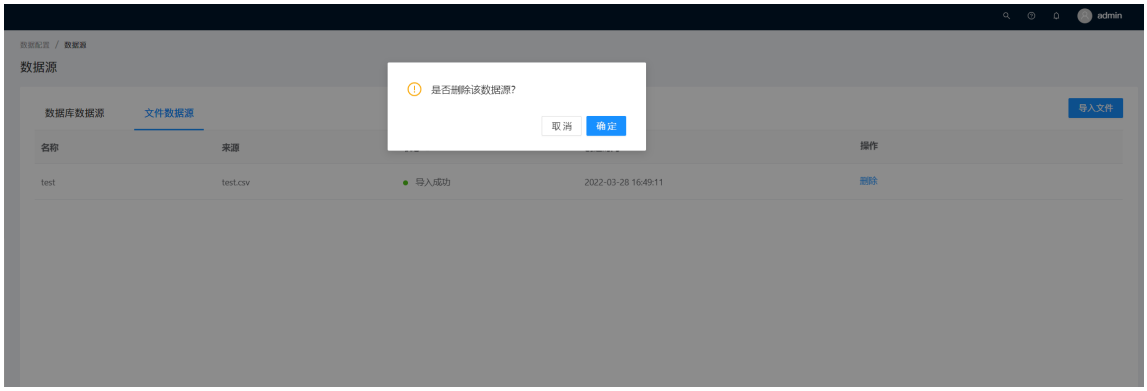
f. 完成创建后，会自动返回到文件数据源页面，在该界面上可看到新添加的文件数据源，可通过查看状态来判断文件是否导入成功。



文件数据源界面

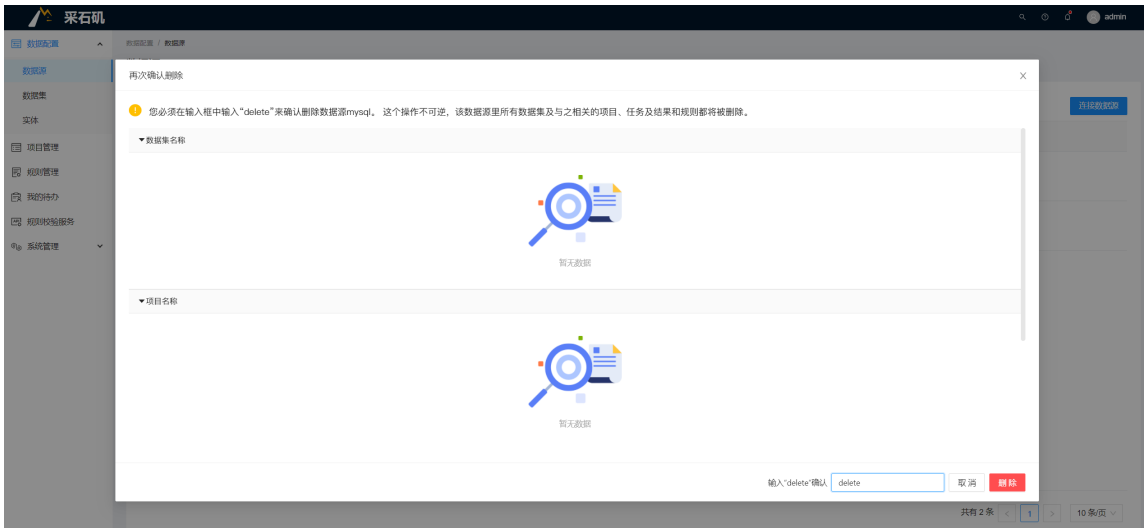
3. 数据源删除

点击数据源的 `删除` 按钮，弹出是否删除该数据源的提示窗口，点击 `确定` ，则弹出再次确认删除弹窗，点击 `取消` 则不删除，并返回文件数据源页面；



删除文件数据源界面

在再次确认删除的弹窗中可看到该数据源相关的数据集名称、项目名称以及规则内容，如数据源没有关联，则显示暂无数据。在弹窗右下角的输入框中输入 `delete` ，并点击 `删除` ，则删除数据源；在弹窗右下角点击 `取消` ，则不删除数据源，并返回到文件数据源页面。删除数据源是不可逆操作，需谨慎操作。



删除文件数据源界面

数据集

本章节主要介绍采石矶系统创建数据集的主要方法和流程。

数据集主要用于存放从数据源处同步过来的数据表，分为镜像表和维度表。镜像表是指将外部数据源中的数据表一对一的复制到采石矶系统中；维度表是指通过映射关系，将外部数据源中的数据表按一个维度重新组织，得到一个新的数据表。

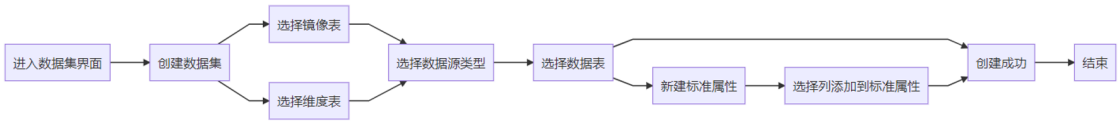
通过完成本章节的步骤，可以了解到在数据集中创建镜像表和维度表的方式。

前置条件

须同时满足以下两个条件：

- 数据库数据源正常配置且已连接成功或文件数据源导入成功；
- 数据库数据源中有一张及以上的数据表可用。

数据集操作流程图

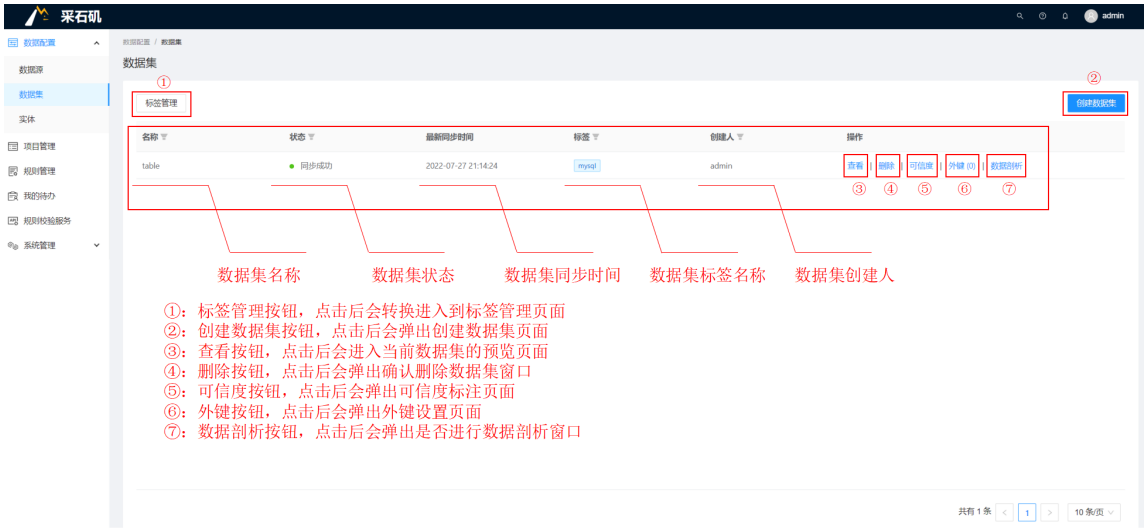


数据集操作流程图

页面说明

1. 数据集页面简介

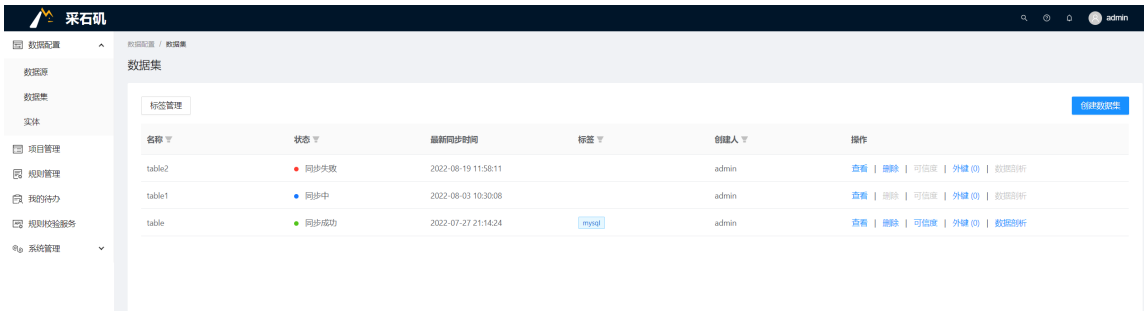
登录采石矶系统，点击 `数据配置` 按钮，选择 `数据集` 按钮，会看到数据集页面，具体呈现如下图：



数据集界面

2. 数据集状态说明

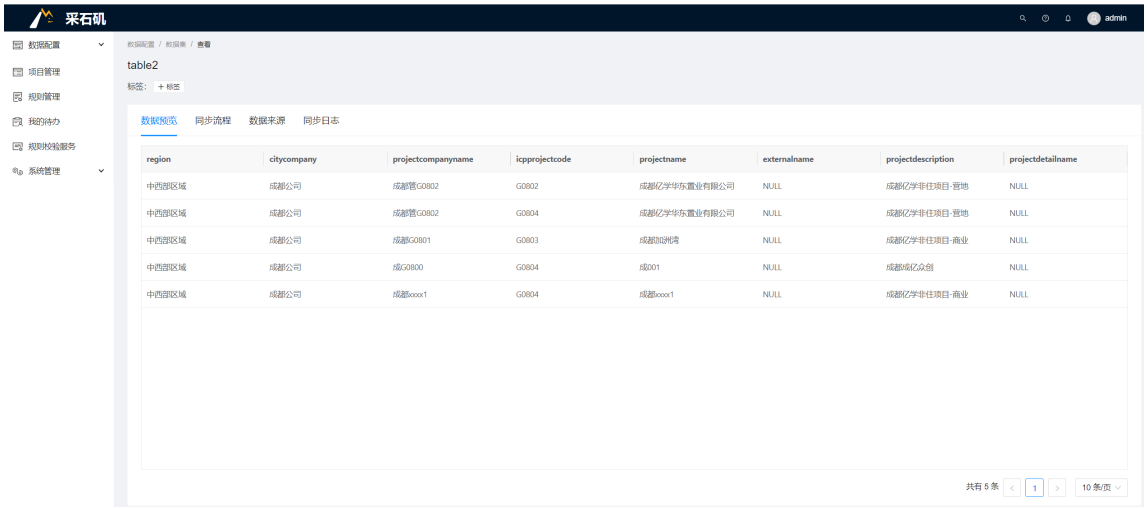
- 数据集状态为同步中时，需等待数据集同步完成；
- 数据集状态为同步失败时，说明该数据集异常，需及时处理；
- 数据集状态为同步成功时，说明该数据集正常。



数据集界面

3. 预览

点击 查看 按钮，可进行数据集的预览。

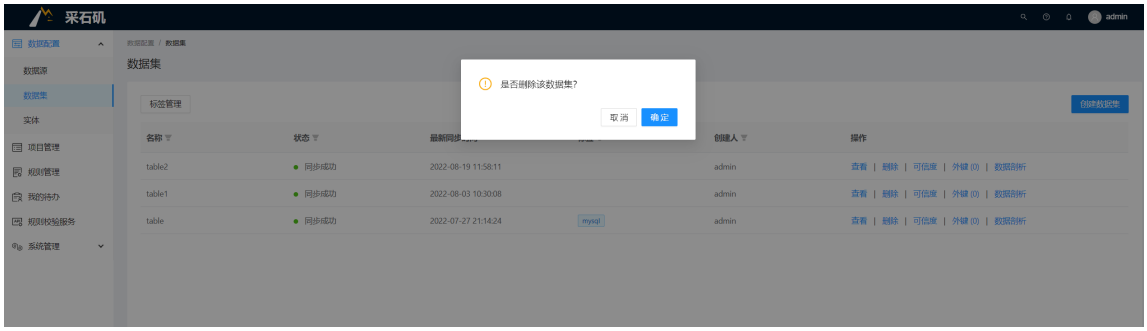


数据集预览界面

- 点击左上角的 `+标签` 按钮，可对该数据集进行添加标签。在输入框中输入需要添加的标签名，如当前已存在该标签，会进行筛选，如当前不存在该标签，可进行创建。具体可参考“标签管理”章节。
- 点击 `同步流程` 按钮，可查看同步流程。
- 点击 `数据来源` 按钮，可查看数据来源。
- 点击 `同步日志` 按钮，可查看数据集同步信息，可点击操作栏中的 `同步` 按钮，对数据集进行再次同步。

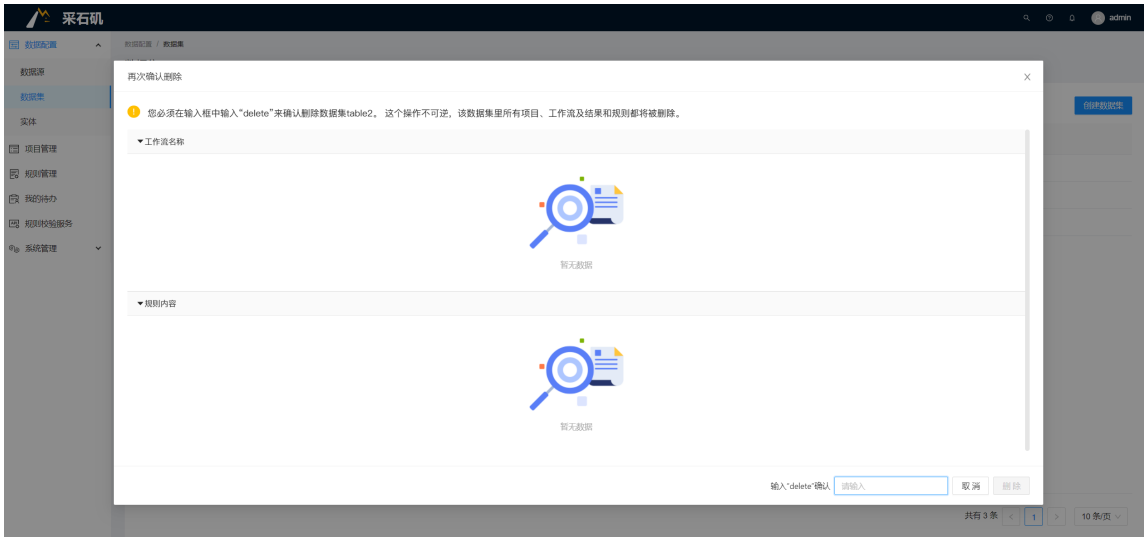
4. 删除

当数据集状态为同步成功或同步失败时，可点击 `删除` 按钮，对数据集进行删除；点击删除按钮后，弹出提示是否删除该数据集，点击 `确定` ，弹出再次确认删除窗口，点击 `取消` 则表示不删除，返回数据集页面；



数据集删除界面

在再次确认删除窗口中可看到与该数据集关联的任务和规则，如该数据集没有关联任务和规则，则显示暂无数据，在弹窗右下角输入 `delete` ，点击 `删除` ，即可删除数据集；点击 `取消` ，则表示不删除，返回数据集页面。



数据集再次确认删除界面

5. 可信度说明

当数据集状态为同步成功时，点击 `可信度` 按钮，可进入可信度标注页面。更多关于可信度的介绍参见后续“可信度管理”章节。

6. 外键说明

当数据集状态为同步成功时，点击 `外键` 按钮，可进入外键设置页面。更多关于外键的介绍参见后续“外键管理”章节。

7. 数据剖析说明

在数据集页面，点击 `数据剖析` 按钮，弹出提示窗口，点击 `确定` ，即表示将开始对该数据集进行剖析，点击 `取消` ，表示取消数据剖析，返回数据集页面。更多关于数据剖析的介绍参见后续“数据剖析”章节。

操作说明

本章主要讲解数据集的相关操作说明，包括创建镜像表、创建维度表、标签管理。

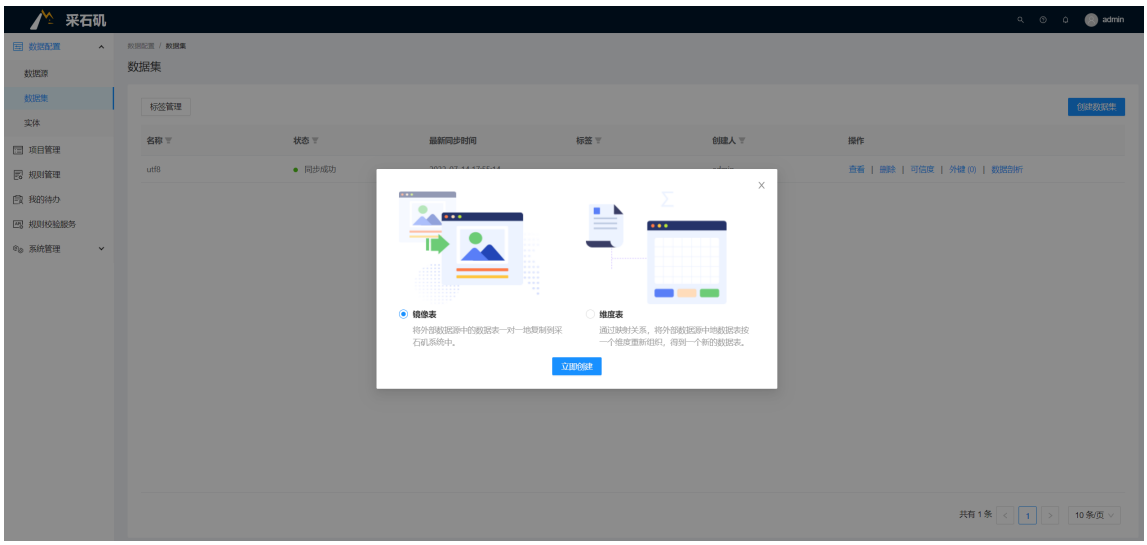
创建镜像表

镜像表的操作流程图

镜像表操作流程图

1. 创建数据集

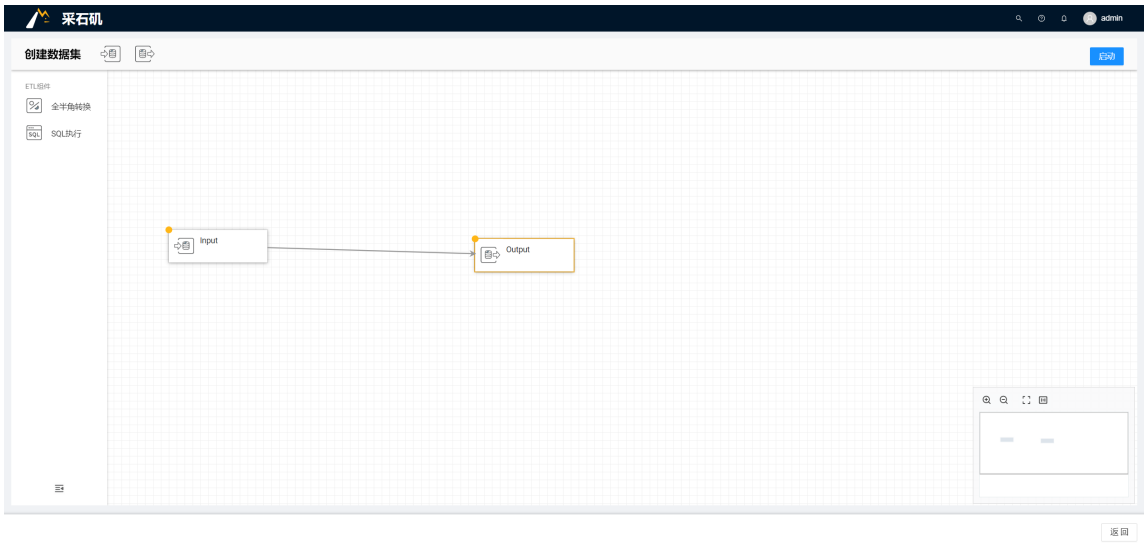
进入 数据集 页面点击 创建数据集 按钮，选择 镜像表 ，点击 立即创建 ，进入 数据配置 页面；



创建数据集界面

2. 选择数据来源

选中页面左上角的input和output两个图标，拖入到画布中，并把input组件和output组件进行连线；

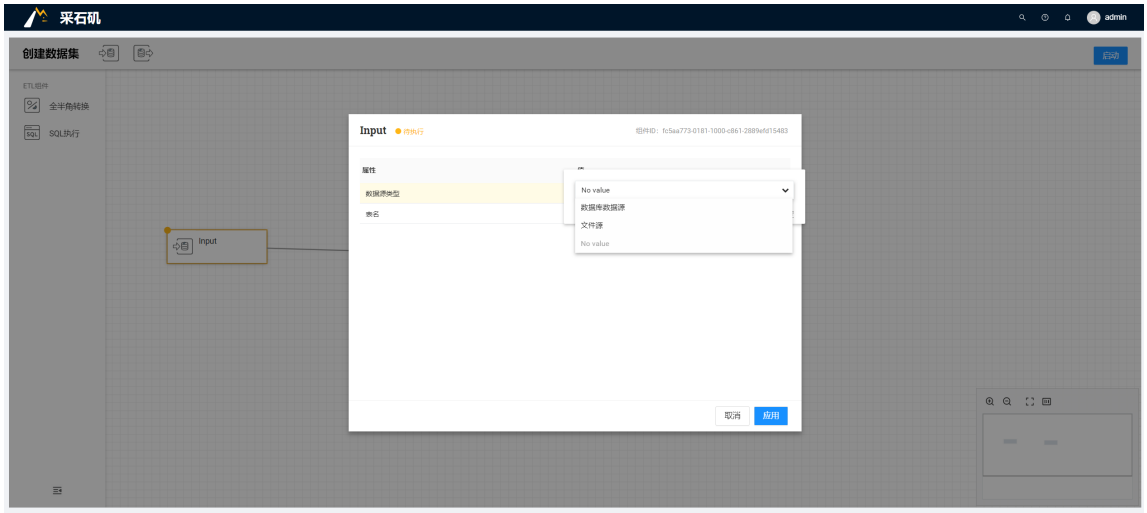


数据配置界面

- 如需要使用ETL组件，可从左侧中拖出需要使用的ETL组件到画布中，连线时把ETL组件置于中间，即：input → ETL组件 → output

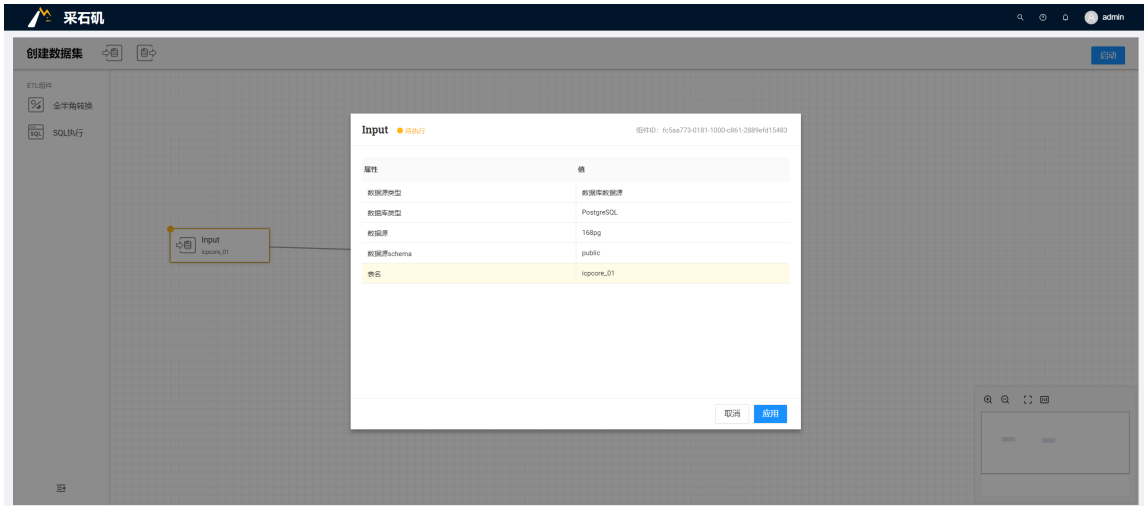
3. 选择数据表

双击打开input组件，弹出input弹窗，在弹窗中“值”的那一列，单击第一个属性对应的值，默认为No value，点击下拉列表，选中数据源类型



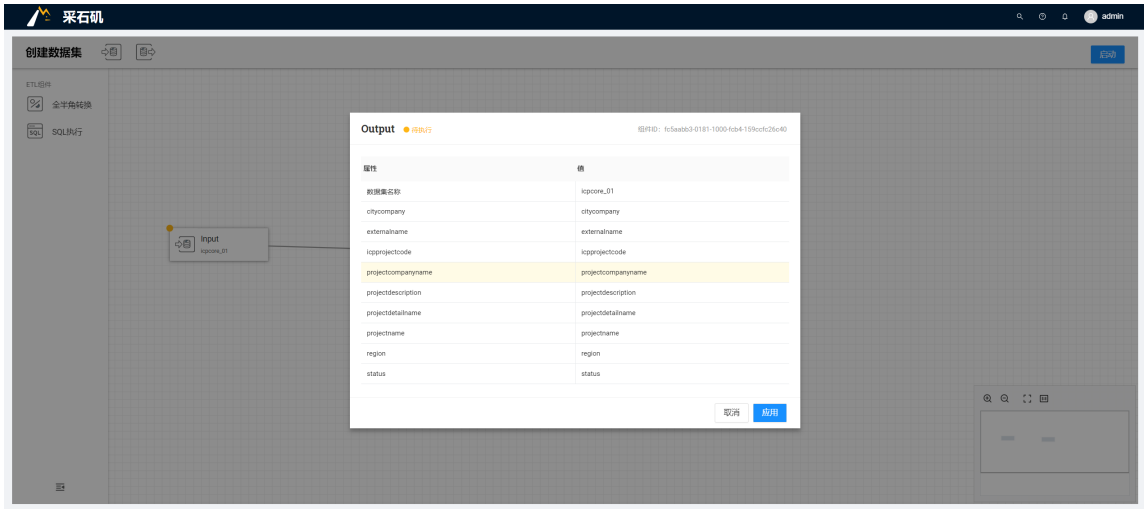
数据配置界面

- 依次对每个属性对应的值进行选择，选择完成后点击应用（如果数据库类型选的是Hbase，还需要手动填写列族名）



数据配置界面

- 双击打开output组件，弹出output弹窗，单击“属性”列中“数据集名称”对应的值，输入数据集名称，点击 确定 ，再点击 应用



数据配置界面

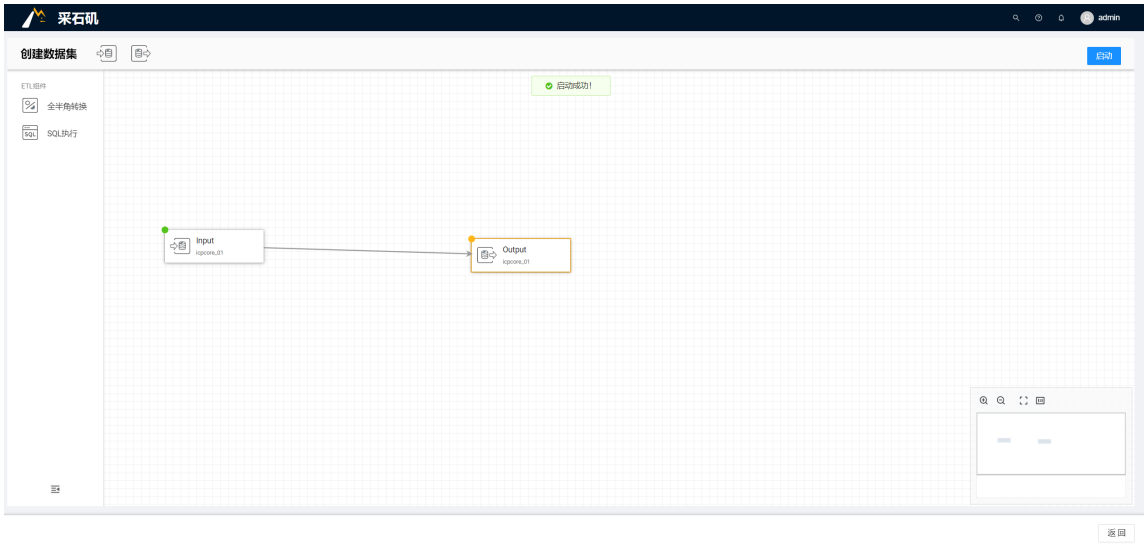
说明：

①、数据集名称不能重复

②、除数据集名称外，属性列中其他的值为数据表中的字段值，在对应的值列单击可进行修改，修改后点击 确定

③、在output中所有的操作完成后，点击 应用

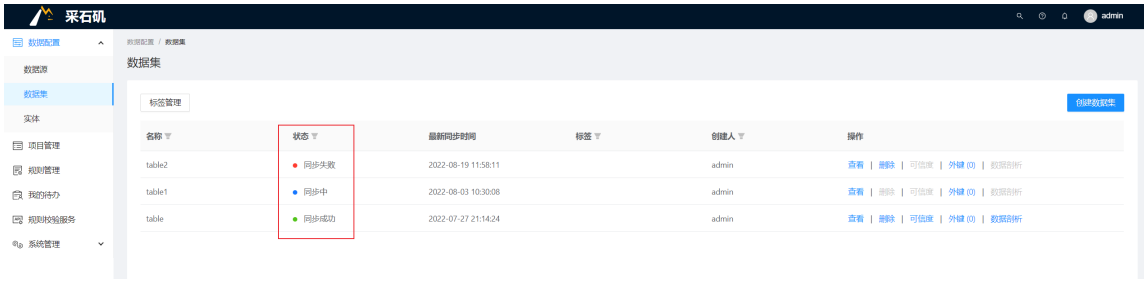
- 所有的配置完成后，在创建数据集页面的右上角点击 启动 ，即可启动成功，启动成功后进行数据同步，点击右下角的 返回 按钮，即可返回数据集页面



数据配置界面

4. 查看结果

点击 确定 完成创建后，自动返回数据集页面，可在数据集页面看到新建的数据集，可通过观察状态来观察该数据集是否同步成功。



数据集界面

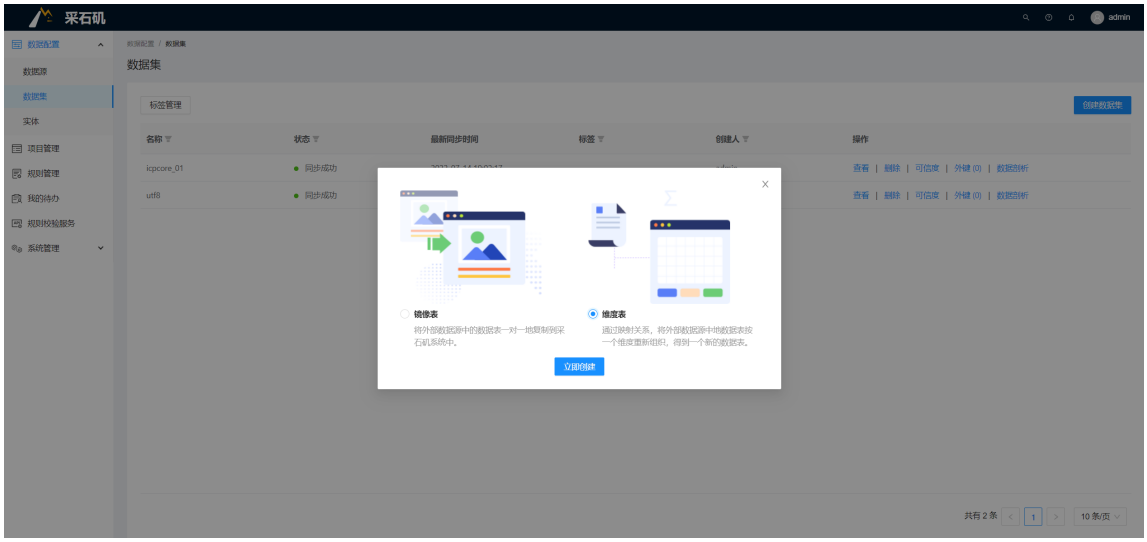
创建维度表

维度表的操作流程图

维度表操作流程图

1. 创建数据集

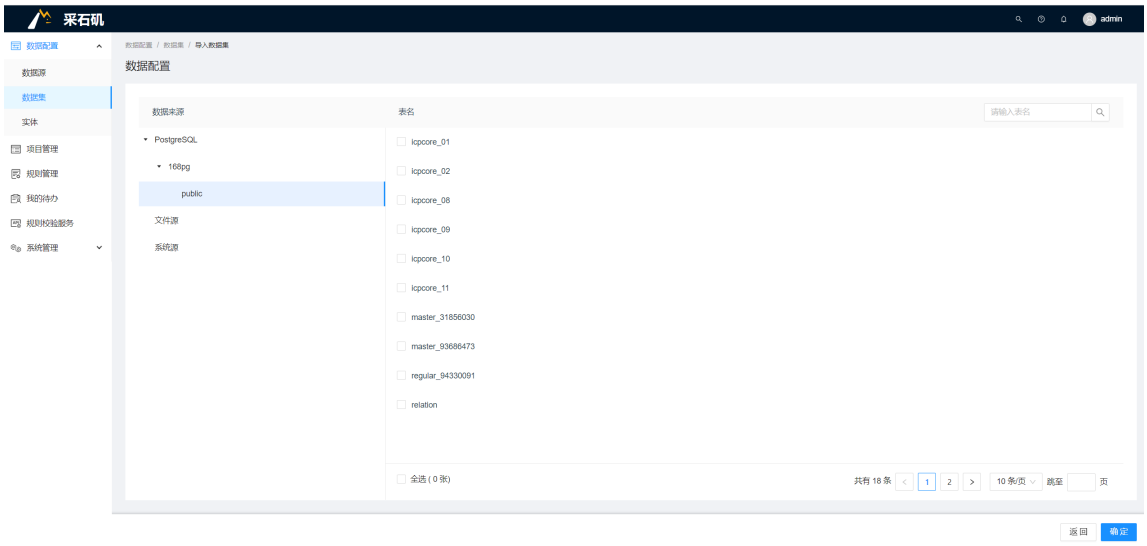
进入数据集页面点击 `创建数据集` 按钮，选择 `维度表` ，点击 `立即创建` ，进入数据配置页面；



创建数据集界面

2. 选择数据来源

页面的左侧为数据来源，可选择需要添加的数据源类型，点击数据源类型后，会下拉数据源名称，再点击需要添加的数据源名称，即可在页面右侧看到数据表；

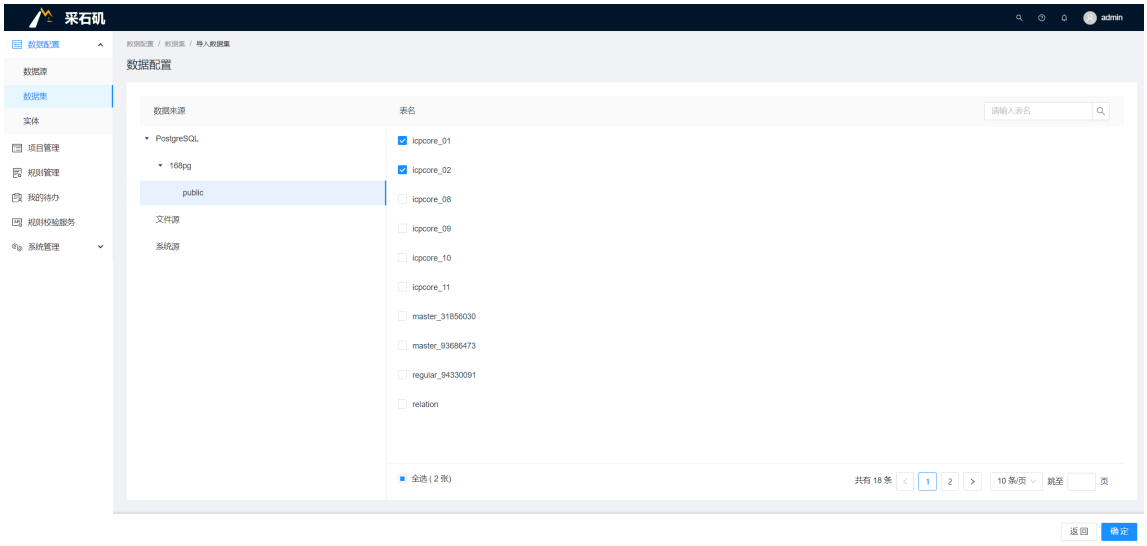


数据配置界面

- 系统源表示已经添加的数据集

3. 选择数据表

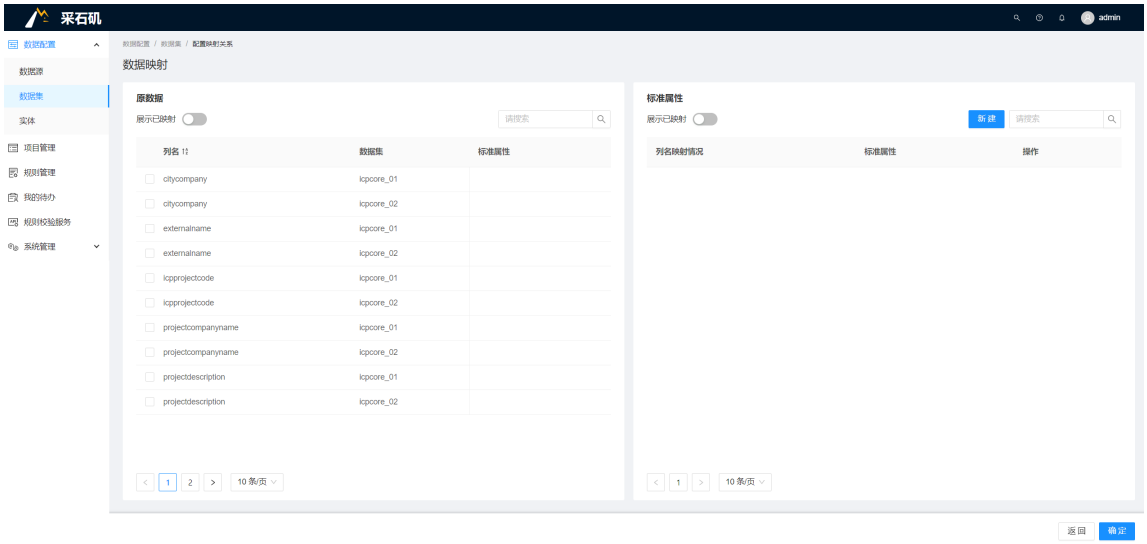
右侧页面看到数据表后，选中需要的数据表，点击 `确定` ，进入数据映射页面；



数据配置界面

4. 数据映射

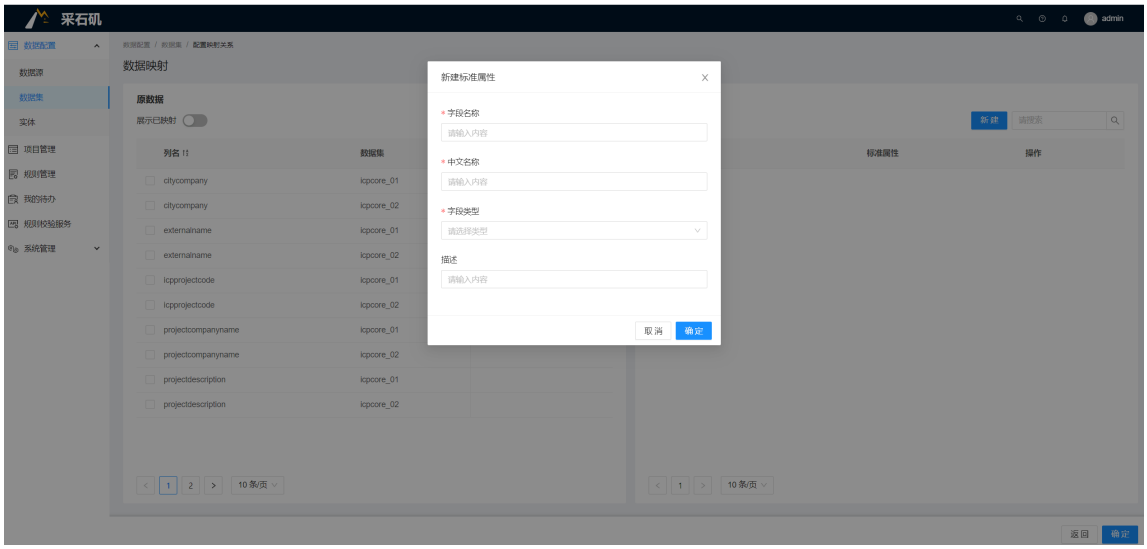
a. 在数据映射页面左侧，可看到已选数据表的列的信息，包括列名、数据集和标准属性。如当前已有标准属性，则展示在右侧页面，当前如没有标准属性，右侧页面的标准属性则显示空白，可点击右上方的 `新建` 按钮来新建标准属性；



数据映射界面

- （可选）可通过左侧页面左上方 展示已映射 开关，来筛选当前选择的数据集已映射的列；
- （可选）可通过左侧页面右上方的搜索框来所搜需要的列名；
- （可选）可通过右侧页面的左上方 展示已映射 开关，来筛选当前已映射的标准属性；
- （可选）可通过右侧页面右上方的搜索框来所搜需要的标准属性。

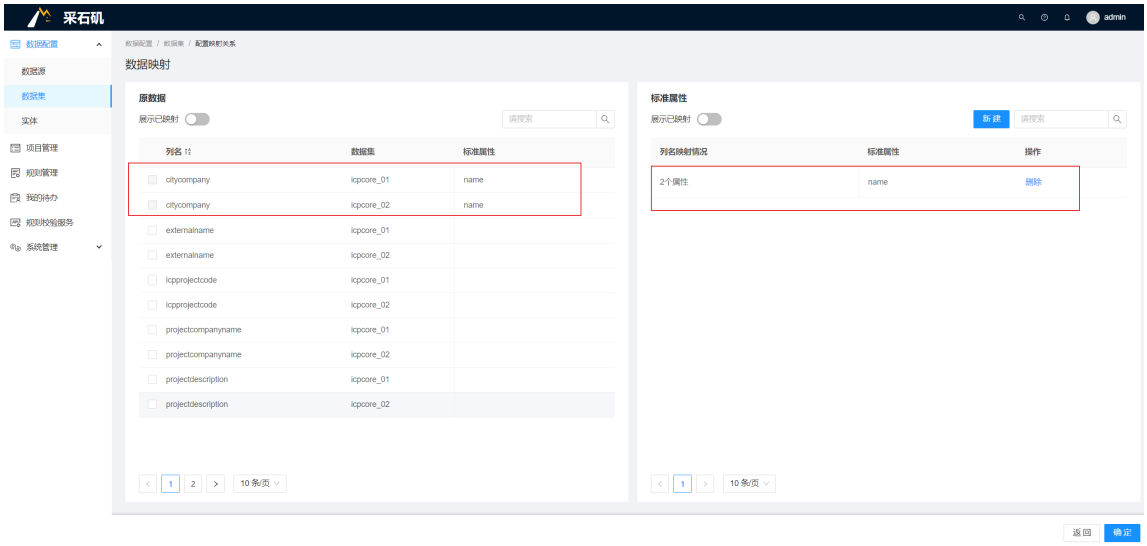
b. 在数据映射页面的右上方，点击 新建 按钮来新建标准属性；



数据映射界面

- （必选）字段名称输入框填写英文；
- （必选）中文名称输入框填写中文；
- （必选）字段类型下拉框选择合适的字段类型；
- （可选）描述输入框填写相应的描述；
- （可选）新建标准属性完成后，可通过该标准属性同一行的右侧的 删除 按钮进行删除。

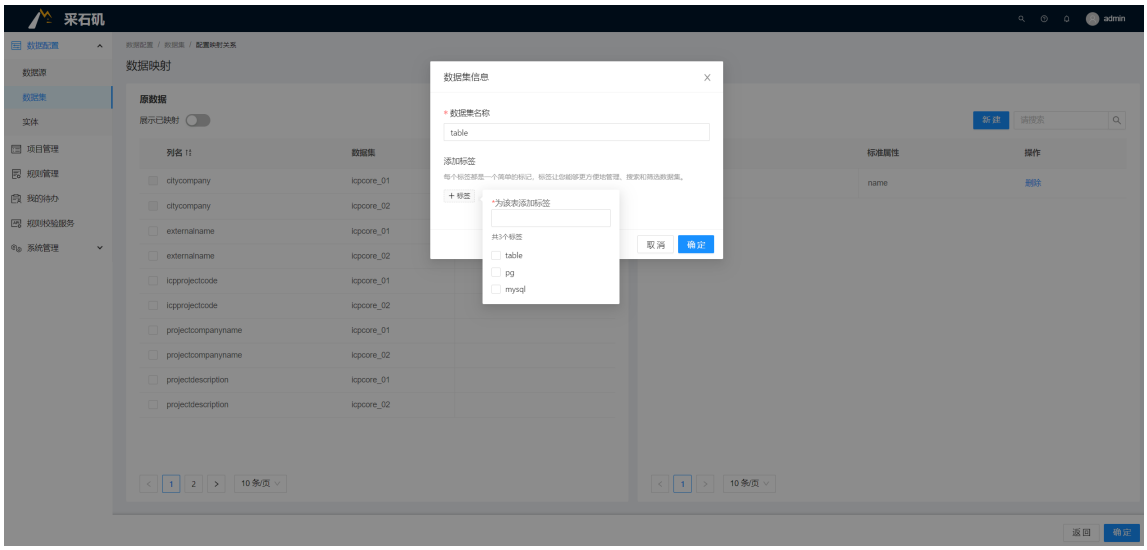
c. 标准属性新建完成后，在左侧页面选中需要进行映射的列，拖入到页面右侧相应的标准属性那一栏，即可完成映射。映射完成后，可在页面左侧标准属性列看到已映射的标准属性；



数据映射界面

- 映射完成后，在页面左侧选中已映射的列，可看到在同一行的最右边有 取消映射 按钮，可进行取消映射；
- 映射完成后，在页面右侧选中已映射的标准属性，光标放到列名映射情况那一列的值上，可看到已映射的列。

d. 映射完成后，点击 确定 ，弹出数据集信息窗口，输入数据集名称，点击 确定 ，即可完成创建；



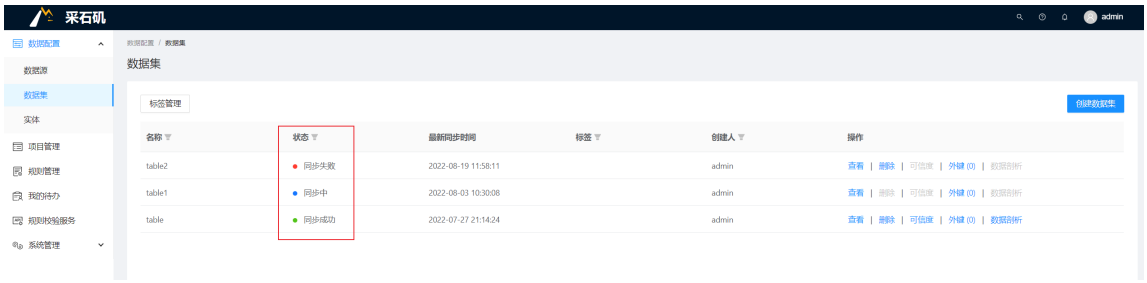
数据集信息界面

- （可选）输入数据集名称后，点击下方的 +标签 ，可对该表进行添加标签。在输入框中输入需要添加的标签名，如当前已存在该标签，会进行筛选，如当前不存在该标签，可

进行创建。具体可参考“标签管理”章节。

5. 查看结果

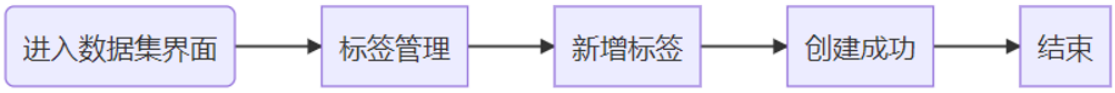
点击 确定 完成创建后，自动返回数据集页面，可在数据集页面看到新建的数据集，可通过观察状态来观察该数据集是否同步成功。



数据集界面

标签管理

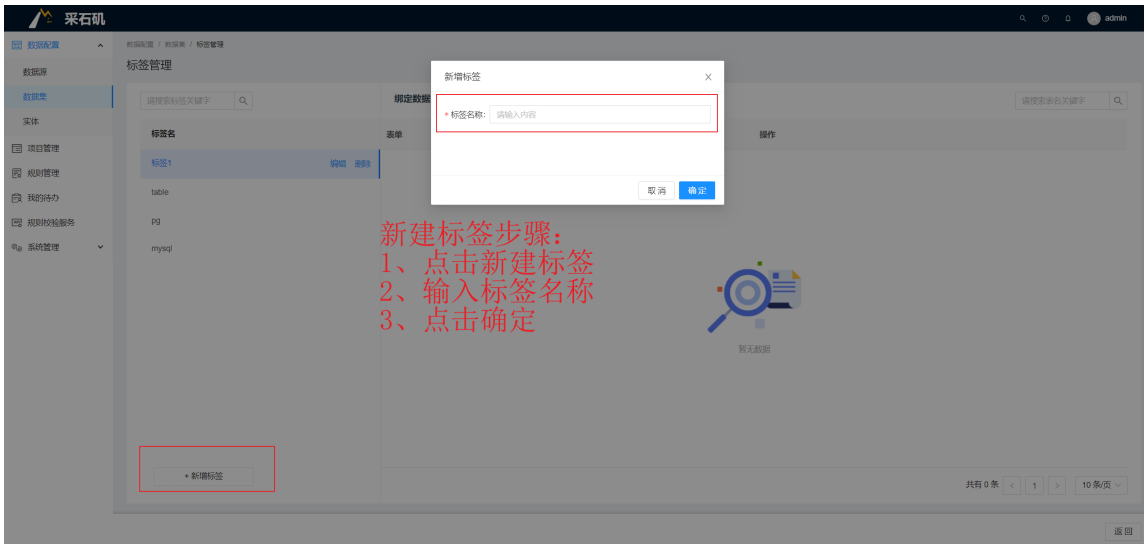
标签的操作流程图



标签操作流程图

1. 新增标签

在数据集页面点击 标签管理 ，进入标签管理页面，点击左下角 新增标签 按钮，在标签名称输入框中输入标签名称，点击 确定 即可，点击 取消 ，表示取消标签创建。标签创建完成后，点击右下角 返回 按钮，即可返回数据集页面。



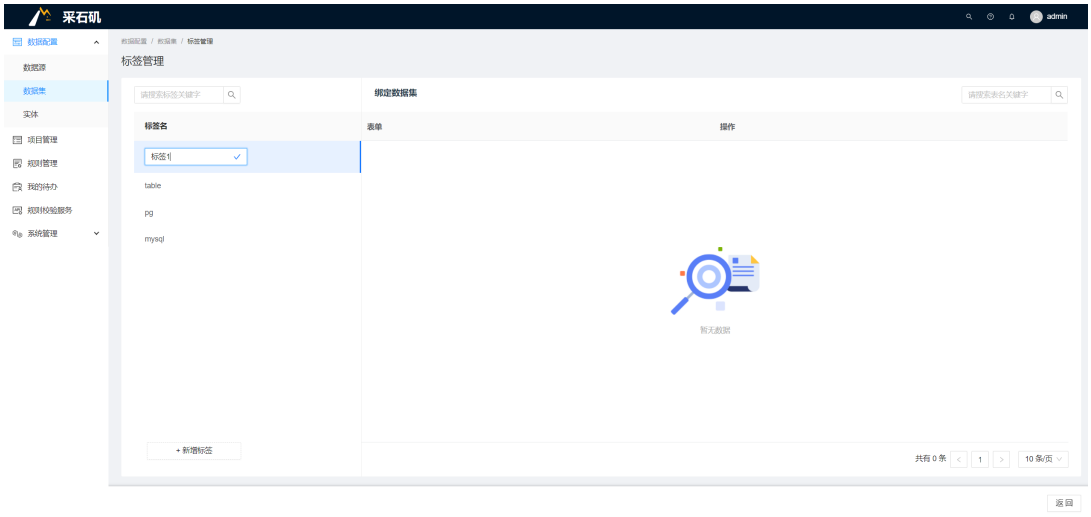
标签管理界面

2. 标签管理

标签管理页面左侧显示已存在的标签，选中已存在的标签，可进行编辑和删除；

- 标签编辑

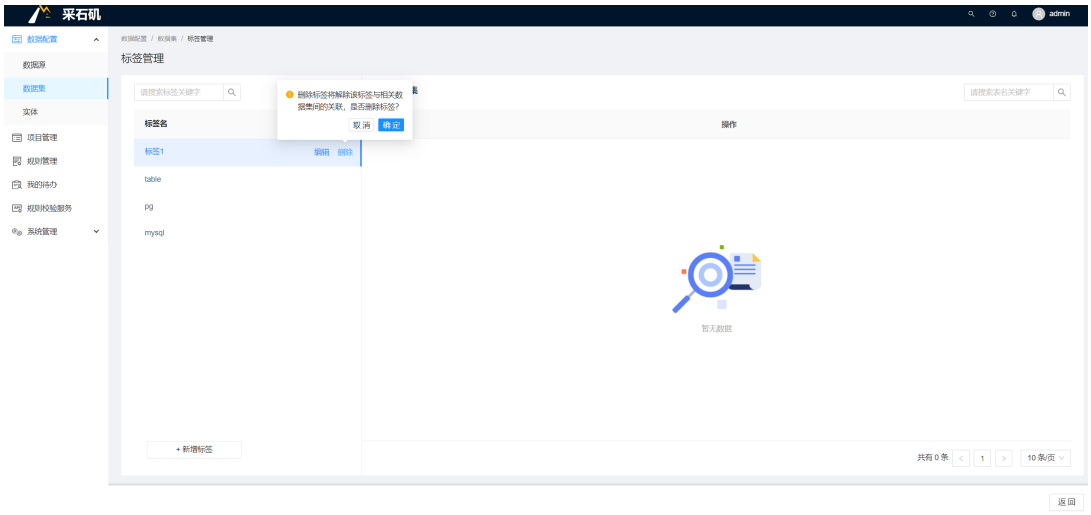
  选中标签，点击 编辑 按钮，即可进行修改，标签名修改完成后，点击 √ ，即可完成修改。



标签管理界面

- 标签删除

  选中标签，点击删除按钮，弹出提示窗口，点击 确定 即可删除，点击 取消 则不会删除。



标签管理界面

标签管理右侧显示与该标签绑定的数据集，可点击取消绑定，如该标签没有绑定数据集，则显示为空。

标签名上方的搜索框可以对标签进行搜索，输入标签关键字即可搜索；标签管理右侧右上方的搜索框，可对与该标签绑定的数据集进行搜索，输入数据集关键字即可进行搜索。

标签管理页面操作完成后，点击返回，即可返回到数据集页面。

外键管理

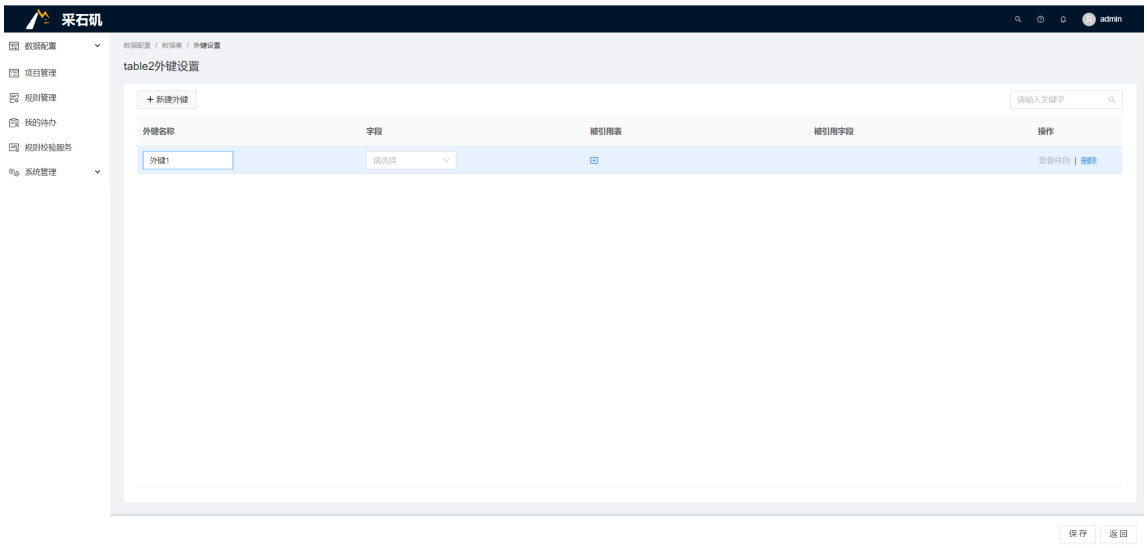
外键的操作流程图

外键操作流程图

1. 新建外键

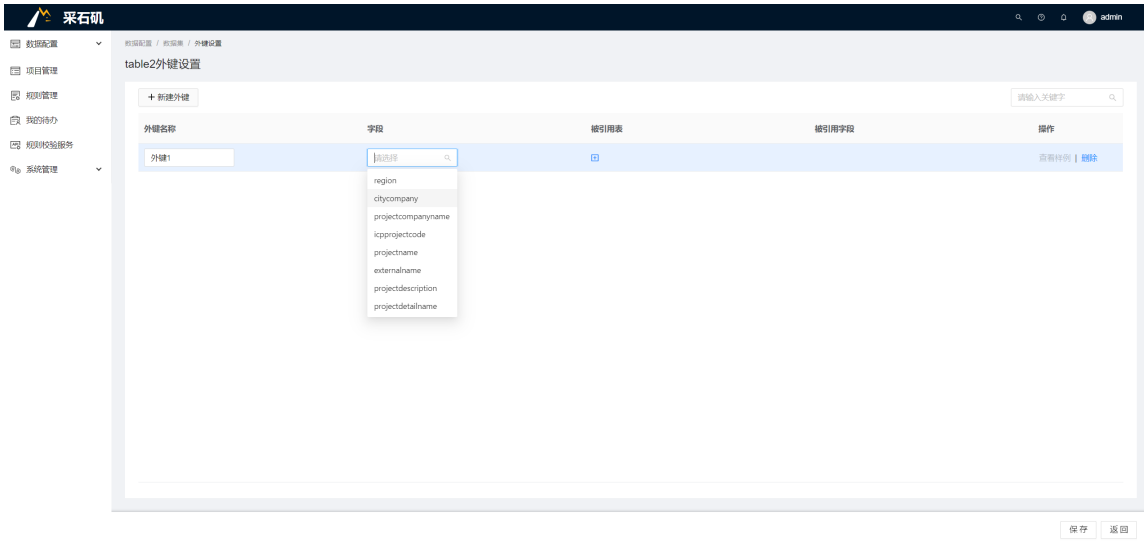
在数据集页面点击 外键 按钮，进入外键设置页面，点击左上角 新建外键 按钮，在页面上会新增一栏输入框。

a. 在外键名称那一栏的输入框中输入外键名称，名称可以自定义。



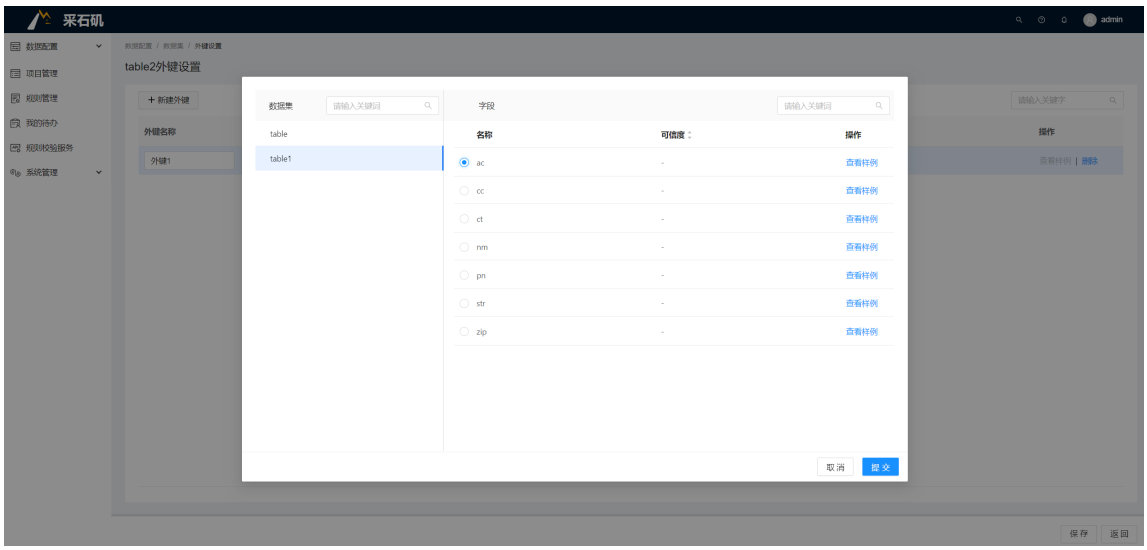
外键管理界面

b. 外键名称输入完成后，在字段栏的下拉框中选择字段，下拉框中包含的字段为该表的所有字段。



外键管理界面

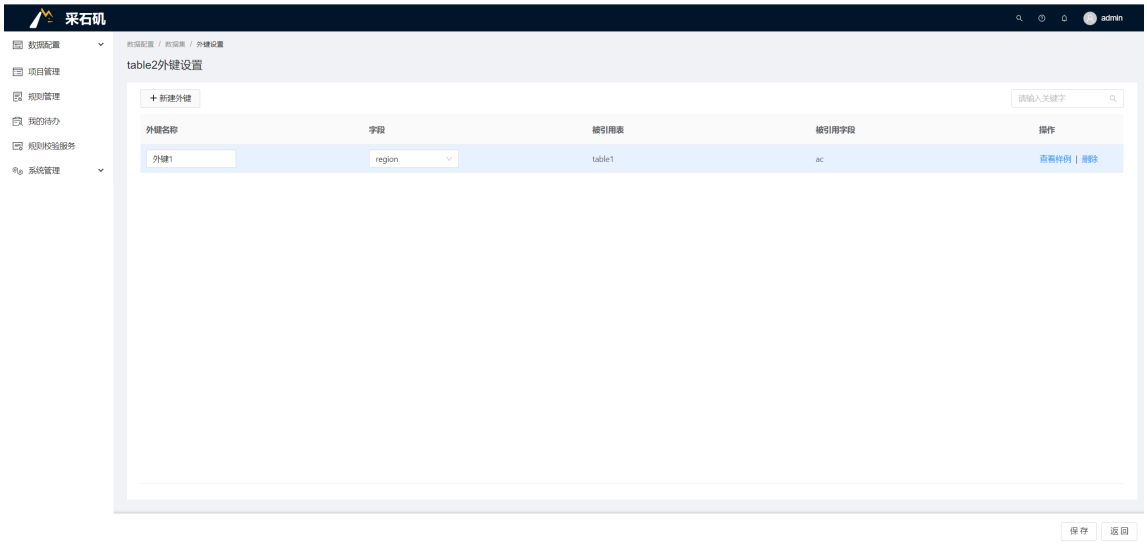
c. 在字段选择完成后，点击被引用表那一栏中的 `+`，弹出添加被引用表的窗口，弹窗左侧的数据集为已添加的所有数据集，可在弹窗左上角的搜索框中搜索目标数据集，选中目标数据集后，在弹窗右侧会显示目标数据集的所有字段，选中需要做外键的字段，点击提交即可。



外键管理界面

- 选中目标字段后，点击 `查看样例`，可以看到主表的字段列数据和对比表的字段列数据。
- 弹窗右上角的搜索框可以进行字段搜索。
- 如果添加外键的两个表之间进行过字段匹配，可信度列会显示字段匹配的值，没有进行过字段匹配则不显示。

d. 提交完成后，即可看到添加的记录，点击右下角的 `保存` 按钮，即可保存成功。此时一条外键已经添加成功，可继续点击 `新建外键` 按钮进行添加外键，也可点击左下角的 `返回` 按钮，返回数据集页面。

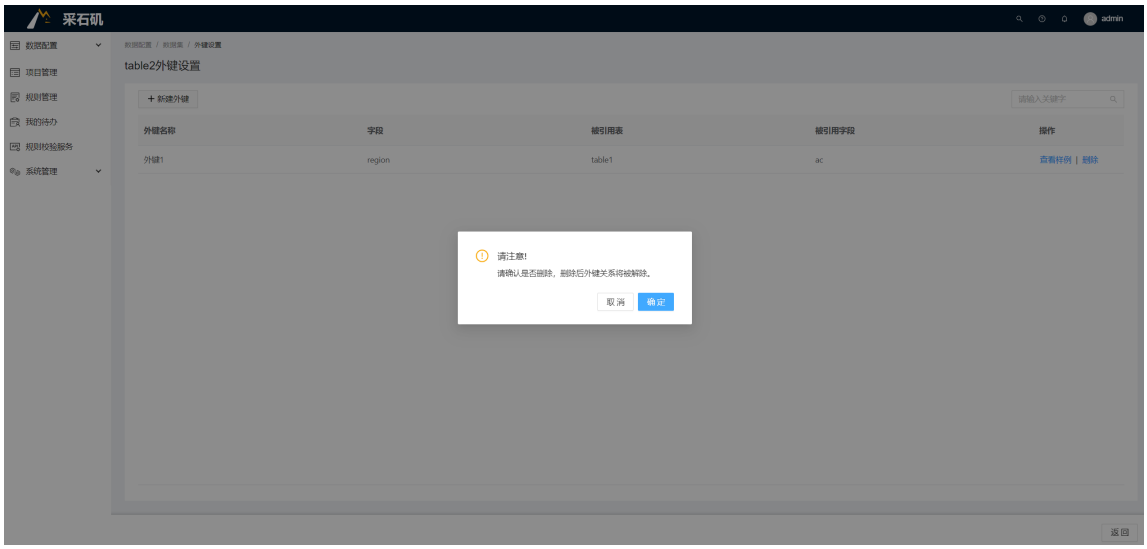


外键管理界面

- 外键添加完成后，可点击操作栏的 查看样例 按钮，再次对主表的字段列数据和对比表的字段列数据进行查看。

2. 删除外键

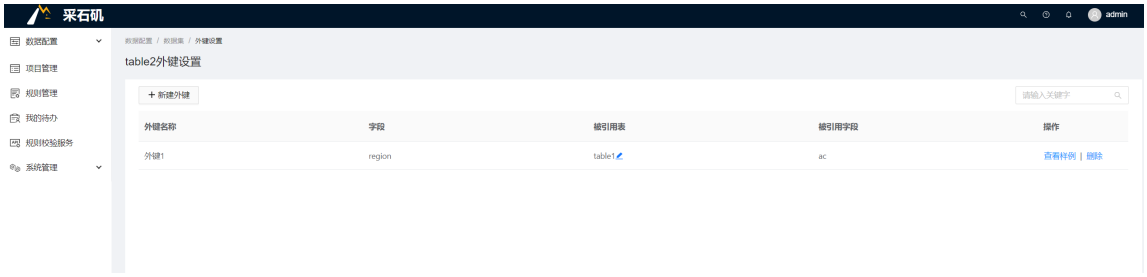
外键添加完成后，点击右侧操作栏中的 删除 按钮，弹出再次确认删除窗口，点击 确定 ，即可删除该条外键记录；点击 取消 ，则不删除，返回外键设置页面。



外键管理界面

3. 编辑外键

外键添加完成后，如需修改对比表字段，可将光标放到被引用表上，此时在表名称右侧会弹出修改的记号，点击该记号，会弹出选择被引用表窗口，此时可重新选择目标数据集或目标字段，选择完成后，点击弹窗右下角的 保存 按钮即可。


采石矶
admin
数据配置
项目管理
规则管理
我的待办
规则校验服务
系统管理
数据配置 / 数据集 / 外键设置
table2外键设置
+ 新建外键
请输入关键字
外键名称
字段
被引用表
被引用字段
操作
外键1
region
table1
ac
查看样例 | 删除

外键管理界面

数据剖析

本章节主要介绍采石矶系统数据剖析功能的操作流程以及相关含义。通过完成本章节步骤，可以对数据剖析功能有清晰的概念并了解相关操作。

前置条件

需满足：数据表是同步成功状态。

数据剖析介绍

数据剖析功能是对数据表的每列内容进行分析并通过图形等方式展示，包括字段类型，字段描述，总数、有效值、值占比等。以下是各数据类型的统计项：

数据类型	字段信息	总数统计	唯一值统计	有效值	空值统计	零值统计	数据统计	重复值TopN统计	值分布统计	值占比统计
字符串(String)	√	√	√	√	√			√		√
日期(Date)	√	√	√	√	√			√		√
整型(Integer)	√	√	√	√	√	√	√	√	√	√
浮点型(Float)	√	√	√	√	√	√	√	√	√	√
布尔(Boolean)	√	√	√	√	√			√		√

各字段释义：

字段信息：包含字段名称、字段类型、字段描述。

总数统计：等于总行数。

唯一值统计：总数去重后的个数（不包括空值），以及对总数的占比。

有效值统计：非空值。

空值统计：统计该列为NULL的个数，及与总个数的比值。

数据统计：统计该列的Min、Max、Sum、Mean、Stdev。

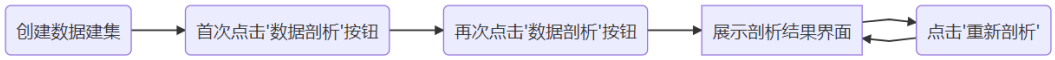
零值统计：统计该列为0/0.0的个数，及与总个数的比值。

重复值TopN统计：统计该列的数据重复次数。（界面展示10条，点击更多降序展示前2000个值）

值分布统计：该列数据的区间切分，统计各个区间数量，并以直方图方式展示。（界面展示10个区间，点击更多以表格形式展示最多20个区间）

值占比统计：统计该列的数据重复次数占比，并以饼状图方式展示。（界面展示10条，点击更多以表格形式展示前2000条数据）

数据剖析流程如下图所示。



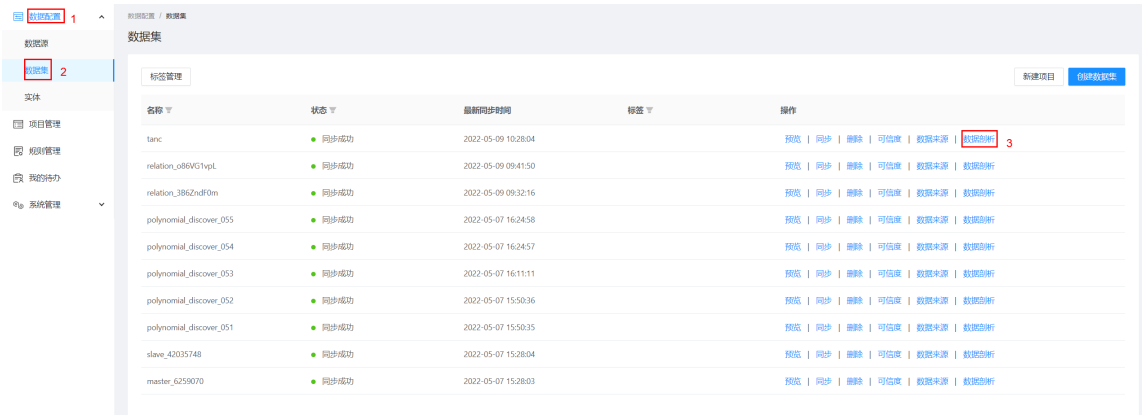
数据剖析流程图

说明：

1. 数据剖析操作的前提：数据表是同步成功状态。
2. 首次点击‘数据剖析’按钮会提示将要触发数据剖析任务，再次点击会提示任务状态，如果是已完成状态，则进入数据剖析结果展示界面。
3. 已完成的数据剖析，在数据剖析结果展示界面可选择重新剖析，重新触发数据剖析任务。

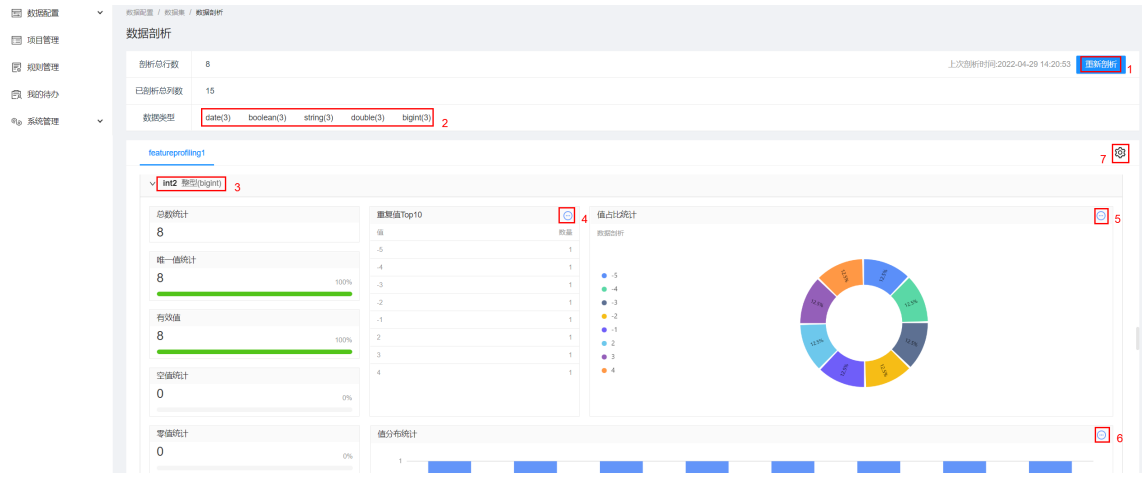
数据剖析界面介绍

- 数据剖析操作入口：



数据剖析操作入口

- 数据剖析结果界面：



数据剖析结果展示

说明：

1. 重新剖析：已经剖析过的数据表，可以重新进行数据剖析。
2. 数据类型：展示当前数据表字段中各数据类型的数量。
3. 字段信息：展示列名、字段类型、字段描述。
4. 重复值更多按钮：点击后弹出框可查看更多重复值情况。
5. 值占比更多按钮：点击后弹出框可查看更多值占比情况。
6. 值分布更多按钮：点击后弹出框可查看更多值分布情况。
7. 筛选按钮：可根据数据表列名模糊搜索，或者根据数据类型来筛选想要展示的数据。

可信度管理

本章节主要介绍采石矶系统数据集的可信度管理，可用于后续的纠错任务中。可信度管理包含数据集的可信度设置，在采石矶系统中，数据的可信程度分为可信和不可信两种，可信度又分为列可信度、单元可信度两种。

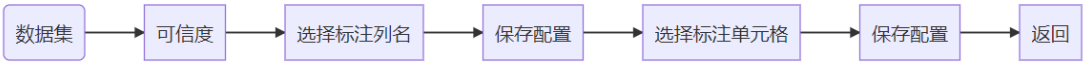
完成本章节步骤，可以了解到数据集的可信度设置流程。

前置条件

需满足以下条件：

- 系统中已有数据源。
- 系统中已有同步的数据集。
- 用户已登录。

可信度的设置流程图

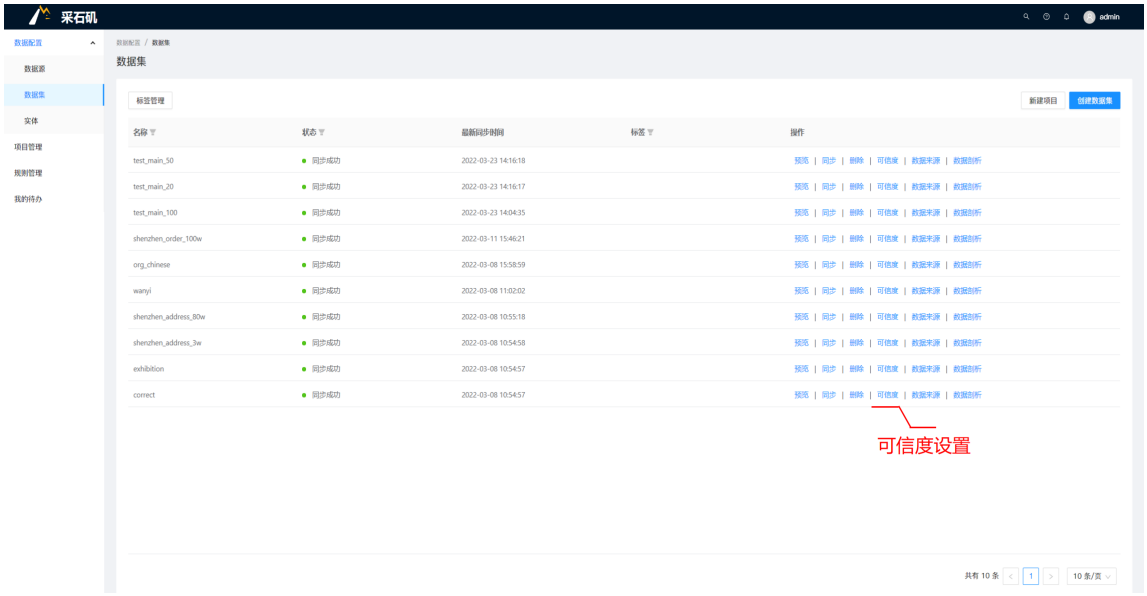


可信度的设置流程图

操作说明

1. 可信度设置入口

点击 `数据配置` 按钮，选择 `数据集` 按钮，进入到数据集页面，点击 `可信度` 按钮，进入可信度标注。

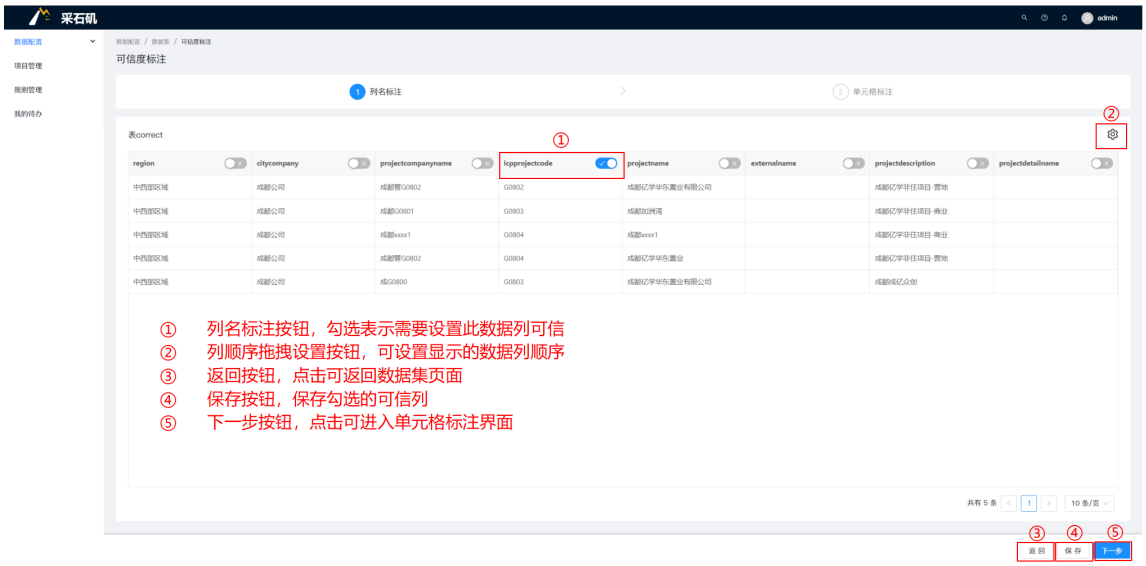


数据集界面

> 数据集正在同步中或者同步失败，可信度按钮将为不可点击状态。

2. 列名标注

在列名标注界面，勾选列名后面的按钮，点击 `保存` 按钮，列可信度配置生效，点击 `下一步` 按钮进入到单元格标注页面。

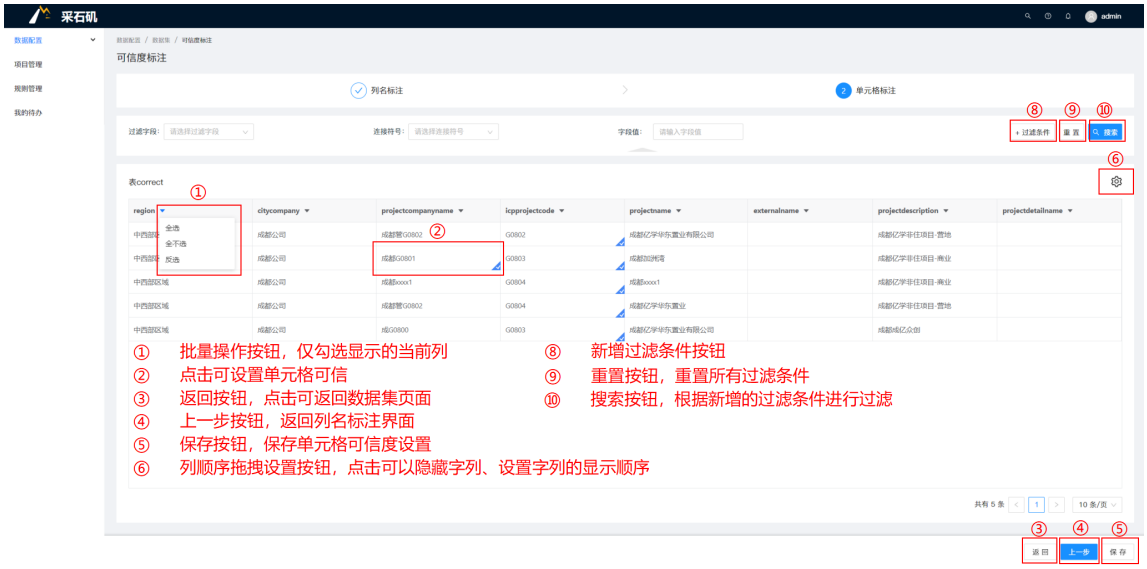


列名标注

> 1.第一次进入可信度标注时有标注弹窗提示，后续进入页面时不再弹窗提示。
>
> 2.列名标注界面勾选列后，需要点击 `保存` 按钮，配置才能生效，否则直接点击返回/下一步会弹出保存提示。

3. 单元格标注

在单元格标注界面，通过鼠标点击勾选单元格，也可以使用 全选 、 全不选 、 反选 按钮对当前显示列进行勾选操作，点击 保存 按钮，单元格可信度标注生效。



单元格标注

> 1.已标注为可信的列，单元格标注界面的整列单元格默认勾选。
>
> 2.单元格的可信度大于列的可信度。
>
> 3.未展示出来的字段列添加筛选条件后也会执行该过滤条件，但不会把该列展示出来。

过滤条件中，字段支持的连接符号如下：

字段类型	连接符号
字符串型	包含，不包含，空值，非空，等于，不等于
数值型	大于，小于，大于等于，小于等于，等于，不等于，空值，非空

项目/工作流管理

本章节主要介绍采石矶系统项目管理及工作流管理在系统中的功能作用及操作流程。

规则发现、查错、实体聚类等任务都是以工作流的形式通过项目的方式进行管理，用户可以创建项目，选择项目相关数据集后，在项目中创建工作流。

1. 前置条件

因为在项目中会关联到数据集，所以在创建项目之前要创建好对应的数据集并同步成功。

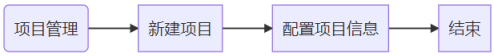
2. 项目、工作流管理总体操作流程图

项目、工作流管理总体流程图

项目管理操作说明

本章节主要讲解项目管理操作说明，包括项目管理的流程和操作。

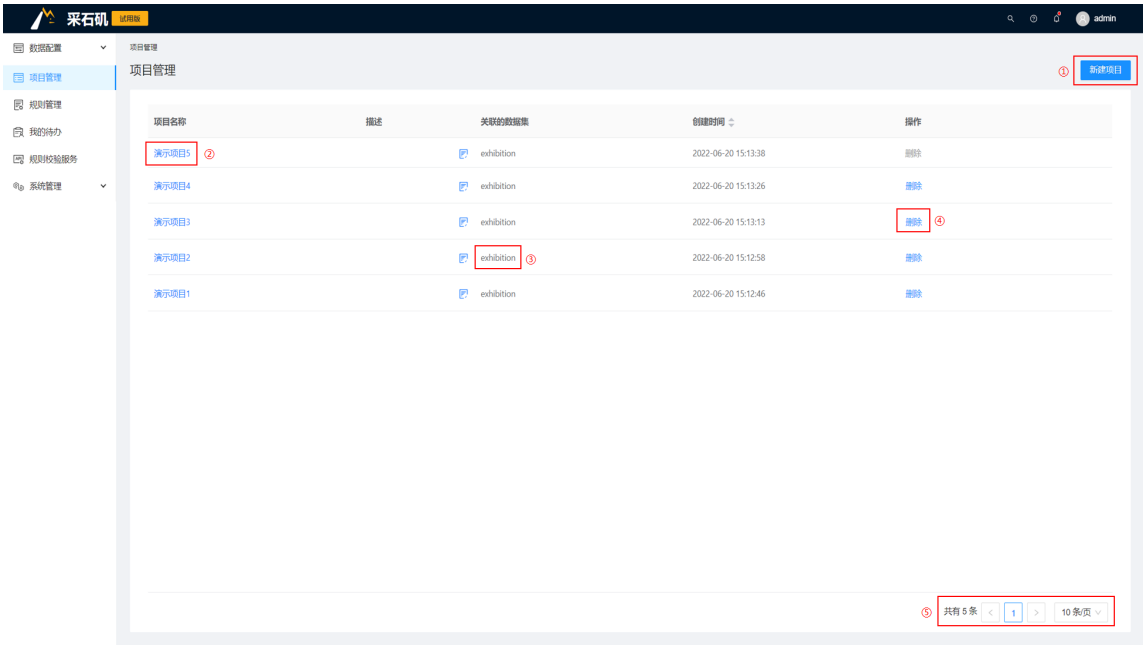
项目管理操作流程如下图所示。



项目管理流程图

1. 项目管理列表页面说明

点击 `项目管理` 菜单，会看到项目管理页面，具体呈现如下图。

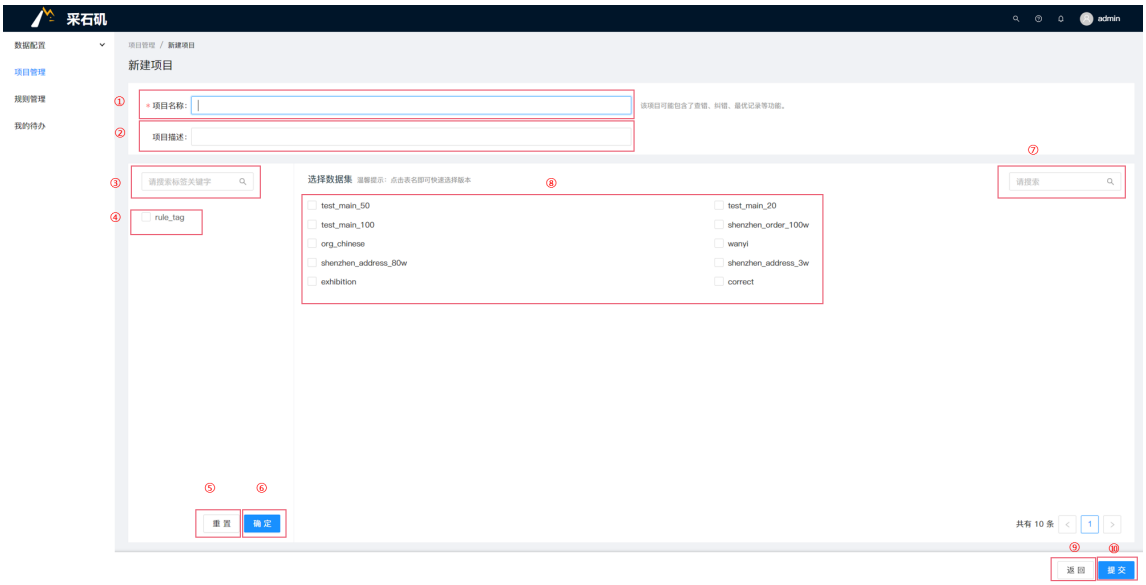


项目管理列表界面

1. 新建项目按钮，点击后会进入新建项目页面；
2. 项目名称按钮，点击后会进入项目详情页面查看对应工作流；
3. 关联数据集按钮，多个数据集时鼠标悬浮后会展示全部数据集；
4. 删除项目按钮，只有项目中的工作流都为完成状态时删除按钮才可点击，点击后会弹出二次确认弹窗，二次确认后会再弹出级联删除确认弹窗，级联删除确认后即可删除项目；
5. 分页器，可切换选择每页显示内容数量。

2. 新建项目页面说明

点击 新建项目 按钮，会进入新建项目页面，具体呈现如下图。



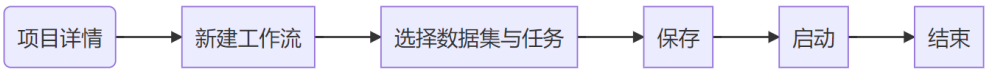
新建项目界面

1. 项目名称输入框，必填，仅支持中文、字母、数字、下划线，不能超过50个字符；
2. 项目描述输入框，非必填，可以对项目添加额外说明；
3. 标签搜索框，可对标签名称进行模糊搜索；
4. 标签选择区，可多选；
5. 标签选择重置按钮，可重置标签的选择；
6. 标签选择确定按钮，选中标签后，点击确定，可筛选出右侧与对应标签绑定的数据集；
7. 数据集搜索框，可对数据集名称进行模糊搜索；
8. 数据集选择区，必选，可多选；
9. 返回按钮，点击即不做数据保存，返回项目管理列表页；
10. 提交按钮，必填字段校验通过后，点击即创建项目成功，返回项目管理列表页。

工作流管理操作说明

本章节主要讲解工作流管理操作说明，包括操作流程图和操作指导。

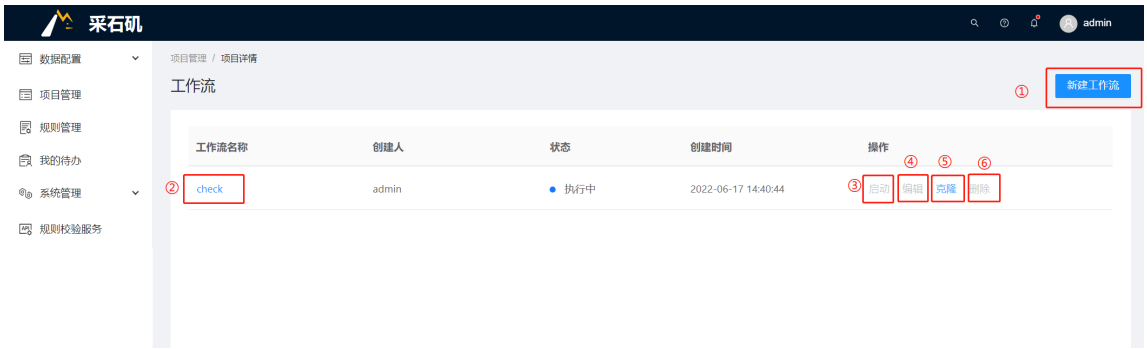
工作流管理操作流程如下图所示。



工作流管理流程图

1. 工作流管理列表页面介绍

在项目管理列表点击对应的项目名称，会进入项目详情页面，具体呈现如下图。

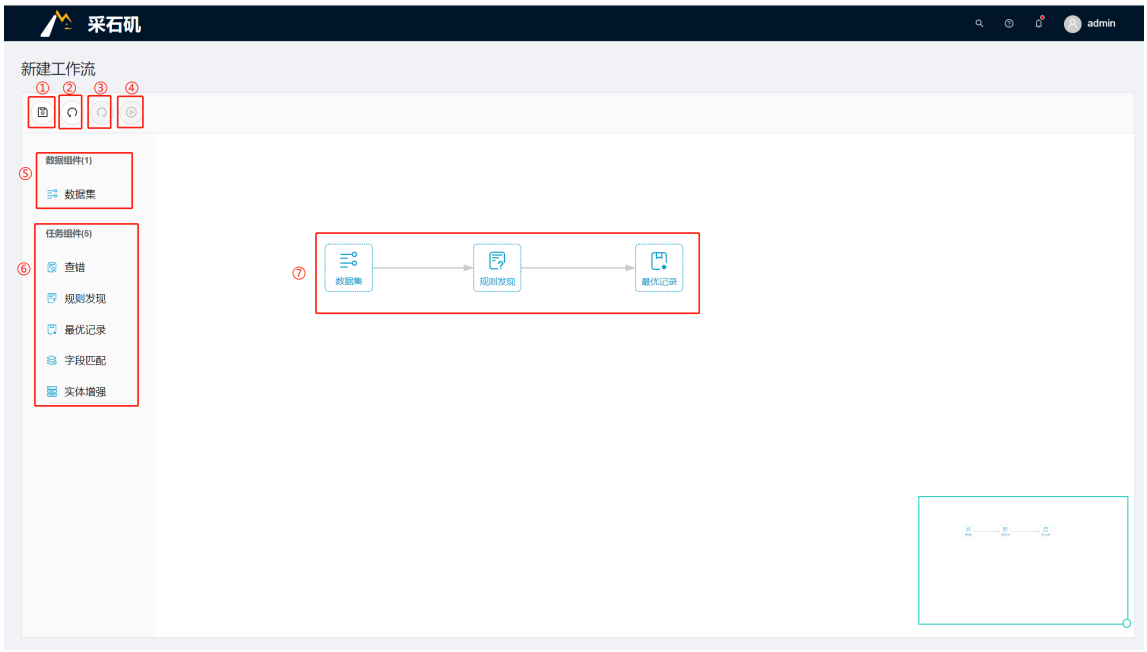


工作流管理列表界面

1. 点击后会进入到新建工作流画布；
2. 点击后可以查看工作流详情；
3. 点击 启动 按钮可以启动工作流；
4. 在工作流启动之前，可以对工作流进行编辑；
5. 点击 克隆 按钮可以克隆工作流。
6. 点击 删除 按钮可以删除工作流，只有工作流的状态为完成和待开始的时候，工作流才可以删除；

2. 新建工作流

点击 新建工作流 按钮，会进入到新建工作流页面，具体呈现如下图。



新建工作流界面

1. 点击可以保存当前工作流；
2. 点击可以撤销上一步操作；
3. 点击可以恢复上一步操作；
4. 点击可以启动工作流；
5. 数据组件仅可拖拽一次，数据集拖拽到画布后双击数据集可以选择工作流的数据集范围。
6. 任务组件可以重复拖拽；
7. 将数据组件和任务组件拖拽到画布内并连线，系统会按照工作流的顺序执行所选任务；

工作流创建并启动后，点击工作流名称会进入到工作流详情页面，点击任务可以进入到任务配置界面。因为不同的任务类型需要配置的信息不同，后续的配置操作会在对应任务类型的章节介绍，本章节不做具体介绍。可参考“规则发现”、”查错”、”数据纠错”、”实体聚类”、”最优记录”、”字段匹配”章节。

规则发现

本章节主要介绍了采石矶系统中规则发现的流程。

通过采石矶系统，用户可利用规则发现来挖掘数据中存在的规则。当前采石矶支持以下三种类型的的规则发现：

- CR规则发现：能够发现用于处理冲突数据的规则。
- ER规则发现：能够发现用于处理实体识别的规则。
- ER+CR规则发现：能够同时发现用于处理实体识别和冲突数据的规则。

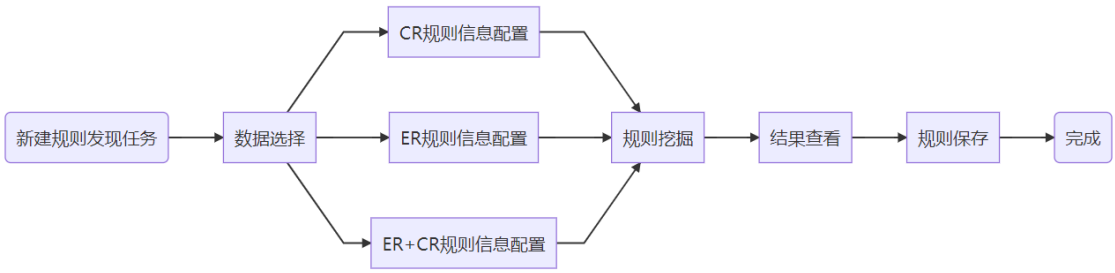
通过完成本章节的步骤，用户可以获取到数据中存在的规则。

前置条件

需满足如下条件：

- 用户已登录。
- 用户已连接数据源，并创建相关的数据集且同步成功。
- 用户已新建项目。

规则发现操作流程图



规则发现-规则发现流程图

规则发现操作说明

本章主要介绍规则发现的操作说明，包括“新建任务”、“数据选择”、“信息配置”、“规则挖掘”、“结果查看”。

新建任务

在新建工作流页面拖入数据集组件，选择数据集后，拖入规则发现任务组件并连线。保存和启动工作流后，会进入查看工作流页面。在查看工作流页面点击规则发现任务组件，配置任务信息。

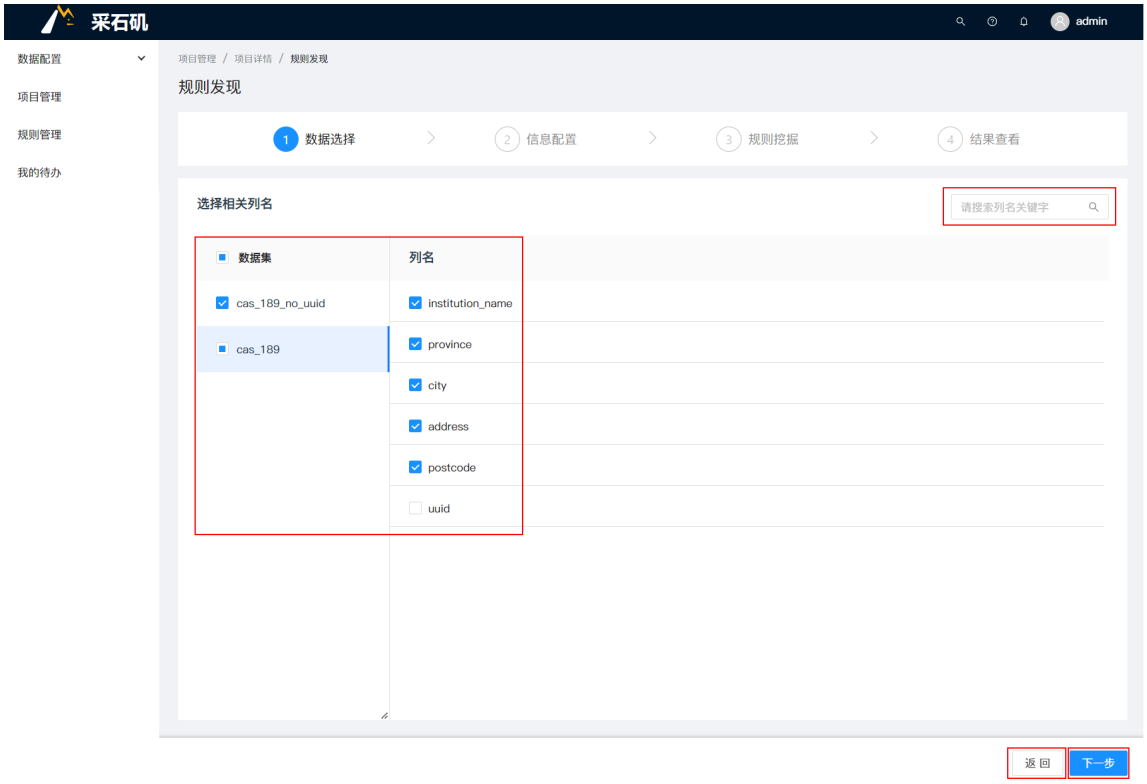
数据选择

用户可以根据需要勾选单个或多个数据集。对于数据集对应的列，也可进行勾选。 未勾选的数据集和列不会参与到规则发现过程中，即未勾选的数据集和列不会出现在规则发现的信息配置中和规则中。

当列过多时，用户可在右方搜索框中输入关键字对列进行过滤后再勾选。搜索框只针对当前数据集的列进行搜索。

用户若点击 返回 按钮，会出现是否结束任务的弹框。

当用户勾选数据集和列后，点击 下一步 按钮后，进入信息配置页面。



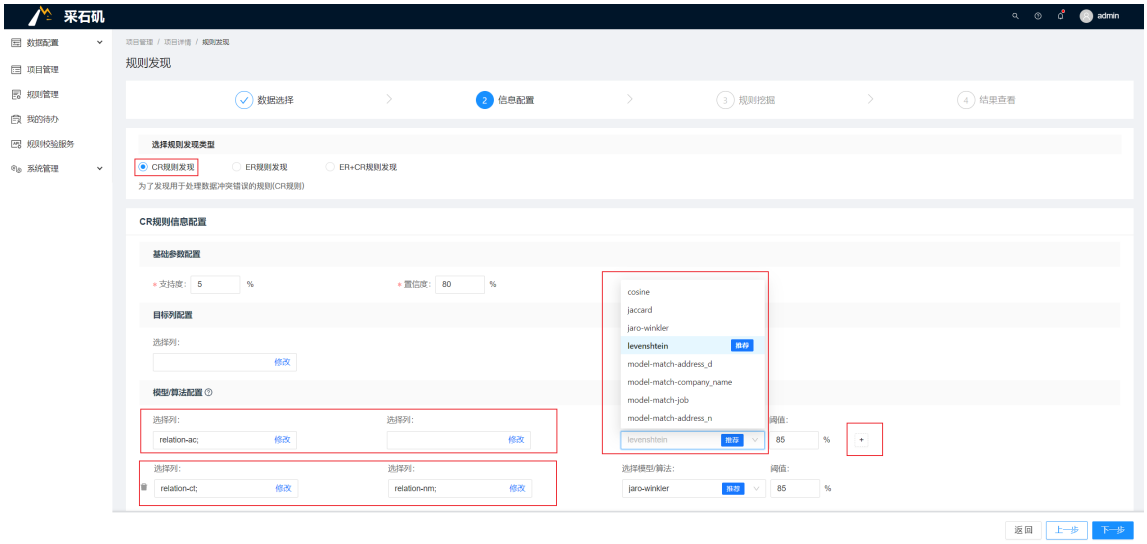
规则发现-数据选择

信息配置

信息配置包括“CR规则信息配置”、“ER规则信息配置”、“CR+ER规则信息配置”。

1. CR规则信息配置

若用户要执行CR规则发现，需选择规则发现类型为CR规则发现。CR规则信息配置如下图所示。



规则发现-CR规则信息配置

在CR规则信息配置中，有以下三种配置：基础参数配置、目标列配置和模型/算法配置。

选项	配置说明	必要
基础参数配置：支持度	满足X的数据占总数据的比例，数据范围是0~100%，默认5%	是
基础参数配置：置信度	满足X且满足Y的数据占满足X的数据的比例，数据范围是0~100%，默认80%	是
目标列配置：选择列	勾选的字段既可出现在X中也可出现在Y中，未勾选的列只出现在X中，默认全部勾选	否
模型/算法配置：选择列	可选择列添加模型或算法进行规则发现	否

> 采石矶系统中规则的样式展示为：X -> Y

- 基础参数配置

  支持度和置信度的默认值分别为5%和80%。在结果查看页面中，只展示支持度和置信度满足大于等于设置的支持度和置信度的规则，用户可根据需要修改支持度和置信度。

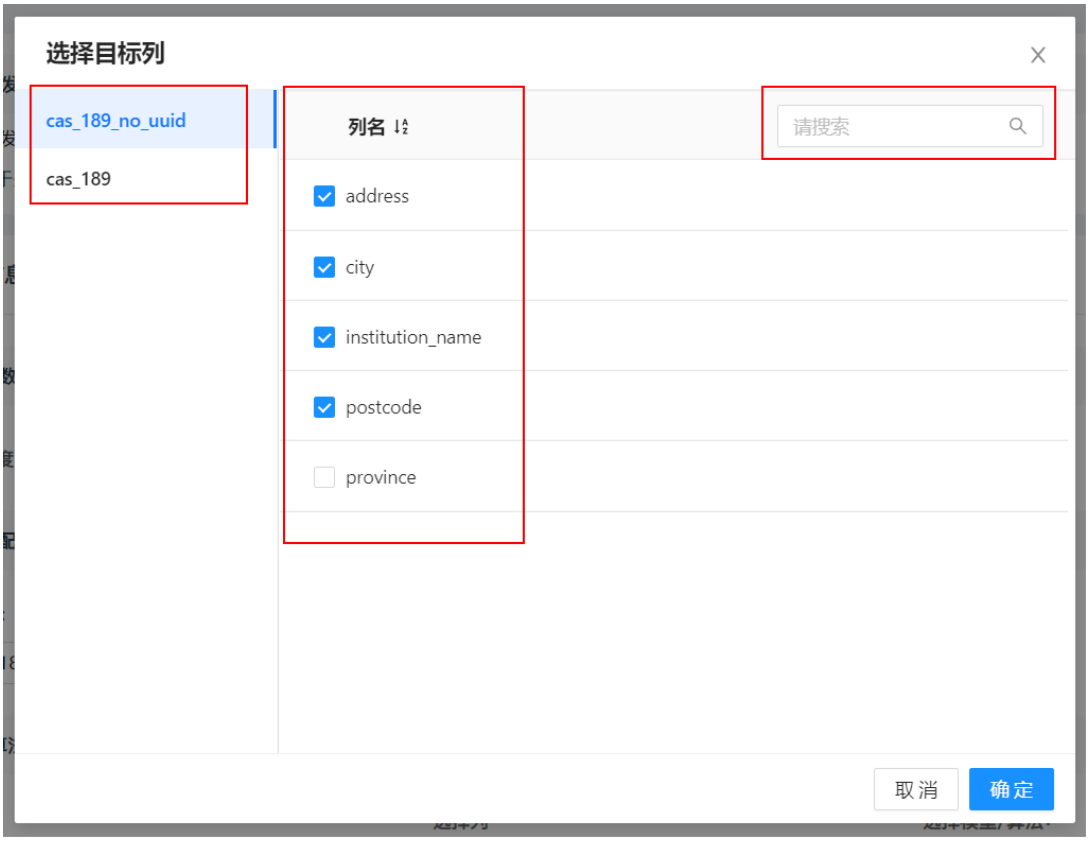
- 目标列配置

  如上图CR规则信息配置所示，当目标列不进行修改时，默认全部勾选。

用户若想修改目标列，可点击 修改 按钮，会出现选择目标列的弹框，如下图所示。

目标列可单表选列，也可跨表选列。左边可切换数据集，列名会出现数据集对应的列。

在右方搜索框中可对列进行过滤。



规则发现-修改目标列

- 模型/算法配置

  用户可根据需要，为列绑定模型/算法。模型/算法配置的详细说明如下：

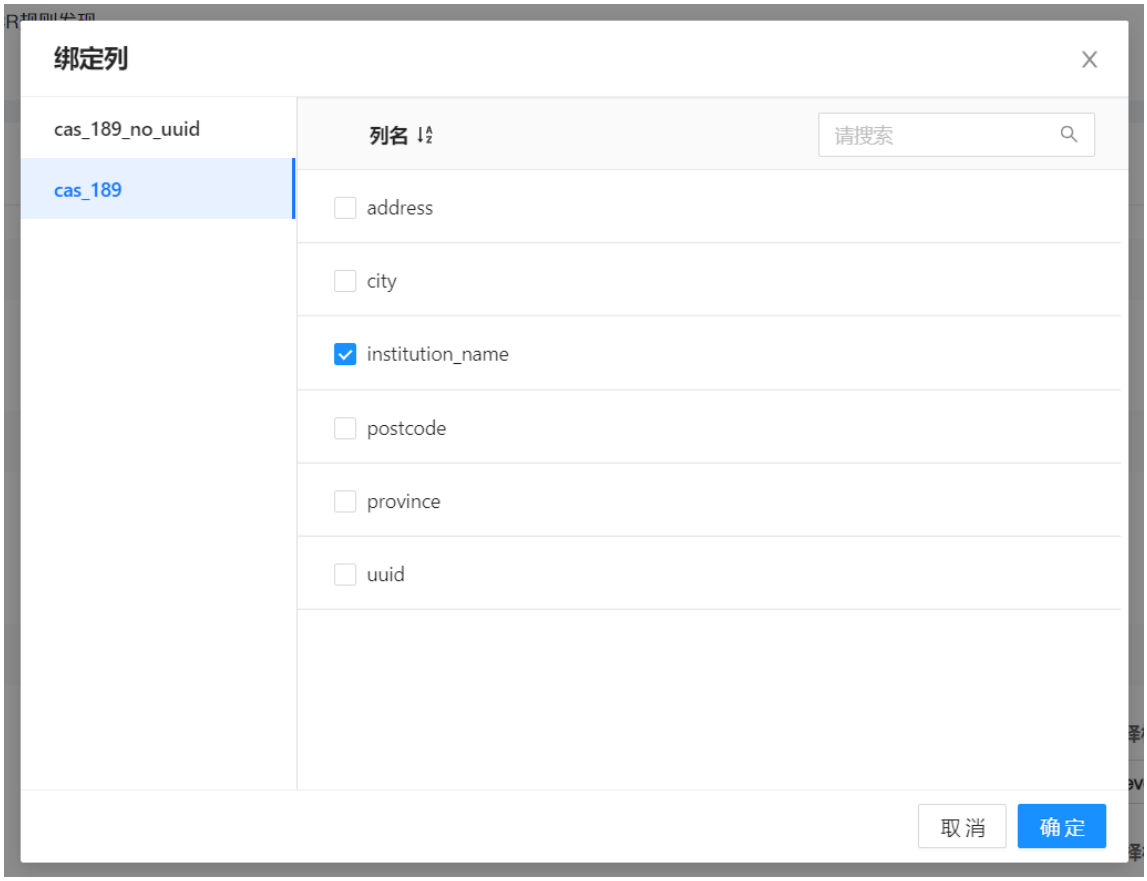
选项	配置说明	必要
选择列-左列	可单选可多选，多选时不可以跨表选列，只能在一张表中选择多个列	是
选择列-右列	当用户需要进行跨表绑定列时，则需要设置右列的选择列。若右列为空时默认是左列的值。	否
选择模型/算法	一组模型/算法配置只能选择一种模型/算法，默认使用系统推荐的模型/算法，用户也可通过下拉选择其他模型/算法	是
阈值	当用户选择相似度算法时才需要填写阈值，阈值范围为0~100%，默认85%	否
右侧”+“号	点击该按钮，可以添加多组模型/算法配置	否

如上图CR规则信息配置所示，当用户想添加多组模型/算法时，可以点击 `+` 按钮。

当用户只需对单张表中的列绑定模型/算法时，只需点击左边选择列中的 `修改` 按钮，就会出现绑定列的弹框，如下图所示。

用户绑定列时只可对单张数据集进行勾选，不可切换数据集勾选，但同一张数据集下的列可勾选单个或多个。点击 `确定` 按钮模型/算法绑定列成功。

若用户想跨表选列绑定模型/算法，则需同时对左边选择列和右边选择列进行配置。此时右列选择列的数据集是与左列选择列的数据集是不一样的。



规则发现-模型/算法配置绑定列

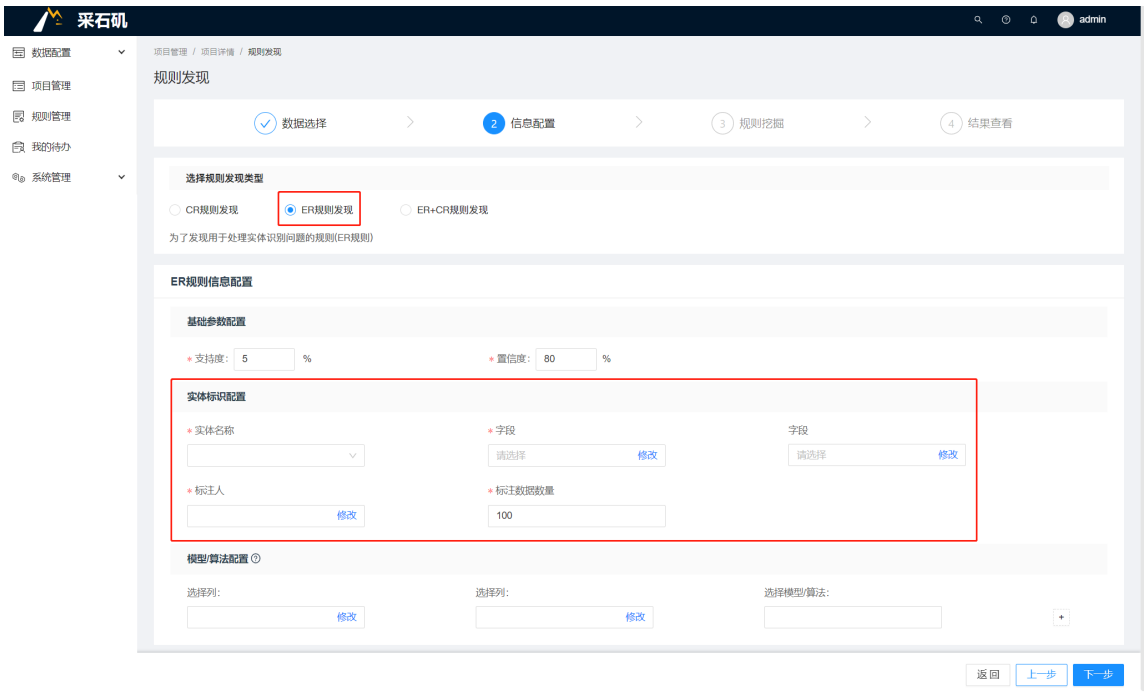
在绑定列后，系统会为列自动推荐相应的模型/算法。除了使用系统推荐的算法以外，用户也可设置其他的模型/算法。

目前规则发现支持如下模型算法：

名称	类型	阈值
cosine	相似度算法	范围0~100%，默认85%
jaccard	相似度算法	范围0~100%，默认85%
jaro-winkler	相似度算法	范围0~100%，默认85%
levenshtein	相似度算法	范围0~100%，默认85%
model-match-address_d	机器学习模型	无
model-match-company_name	机器学习模型	无
model-match-job	机器学习模型	无
model-match-address_n	机器学习模型	无

2. ER规则信息配置

选择规则发现的类型为ER规则发现。



规则发现-ER规则信息配置

ER规则信息配置有以下三种配置：基础参数配置、实体标识配置和模型/算法配置。信息配置详细说明如下。

选项	配置说明	必要
基础参数配置：支持度	满足X的数据占总数据的比例，数据范围是0~100%，默认5%	是
基础参数配置：置信度	满足X且满足Y的数据占满足X的数据的比例，数据范围是0~100%，默认80%	是
实体标识配置：实体名称	只能输入英文、数字和下划线，长度不能超过30	是
实体标识配置：字段-左边	当用户只需要创建单表实体时，只需配置左边字段。可单选可多选，只能在一张表中选择多个	是
实体标识配置：字段-右边	当用户需要创建跨表实体时，才需配置右边字段。可单选可多选，只能在一张表中选择多个列	否
实体标识配置：标注人	可单选可多选，生成的标注数据会平均分发给标注人	是
实体标识配置：标注数据数量	范围100~10000的正整数，默认值是100	是
模型/算法配置：选择列	可选择列添加模型或算法进行规则发现	否

> 采石矶系统中规则的样式展示为：X -> Y

基础参数配置和模型/算法配置与CR规则发现中的一致，在此不做赘述。

针对实体标识配置，有以下四种情况：1)选择实体，未生成标注集；2)选择实体，已生成标注集且使用已有标注集； 3)选择实体，已生成标注集但不使用已有标注集；4)新建实体。

现针对四种情况的操作做出说明。

- 选择实体，未生成标注集

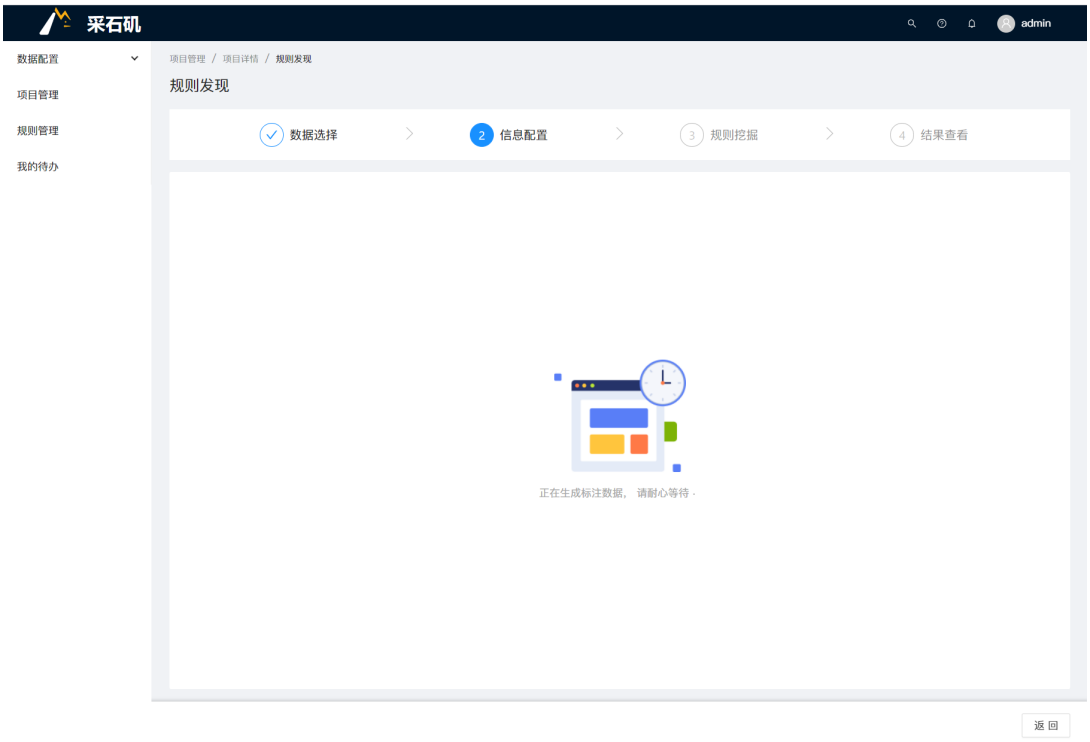
  当数据集已有关联的实体时，用户可下拉实体名称选择实体。由于选择的实体已配置字段，因此选择实体后不需配置字段。

规则发现-实体标识配置（选择实体，未生成标注集）

若要配置标注人，用户可点击 修改 按钮，勾选对应标注人后点击 确定 按钮，标注人配置成功。

规则发现-修改标注人

当ER规则信息配置完成后，点击 下一步 按钮，进入生成标注数据页面，并将标注数据平均分发给配置的标注人。



规则发现-生成标注数据

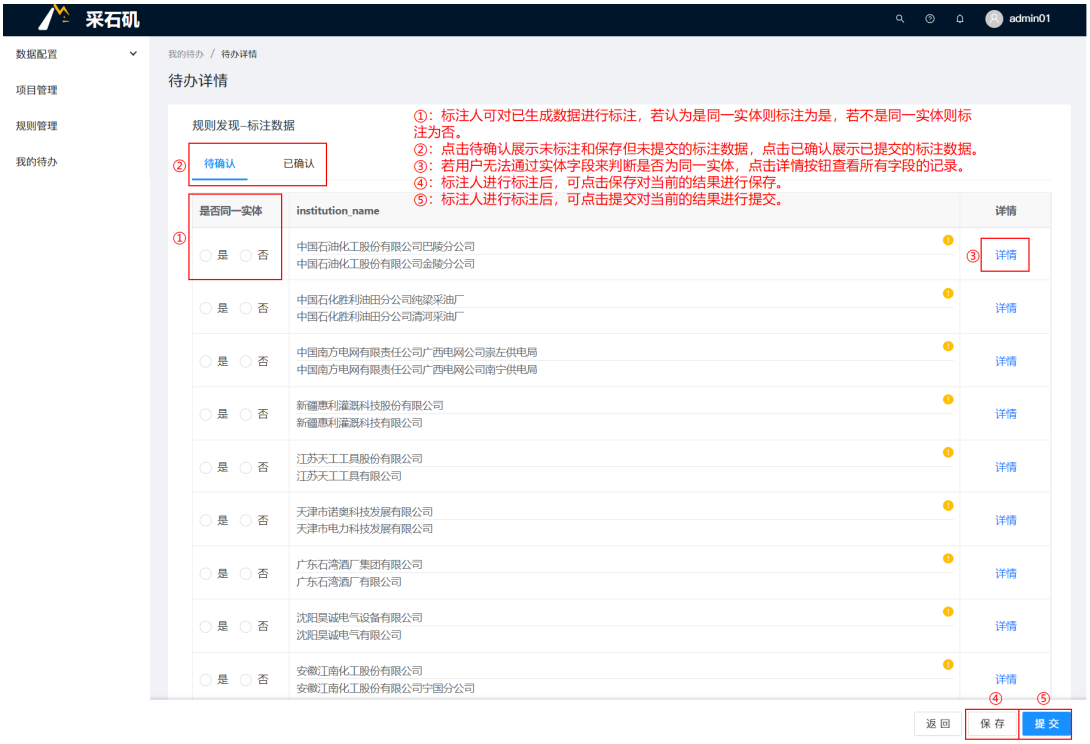
当标注数据生成后，标注人需登录进入到我的待办中对已分发的标注数据进行标注。

这里以用户admin01为例来讲述标注人标注数据的操作。

用户admin01已登录，点击 我的待办 ，即可查看待办列表。 点击对应的标注任务，即可进入到标注页面进行标注。

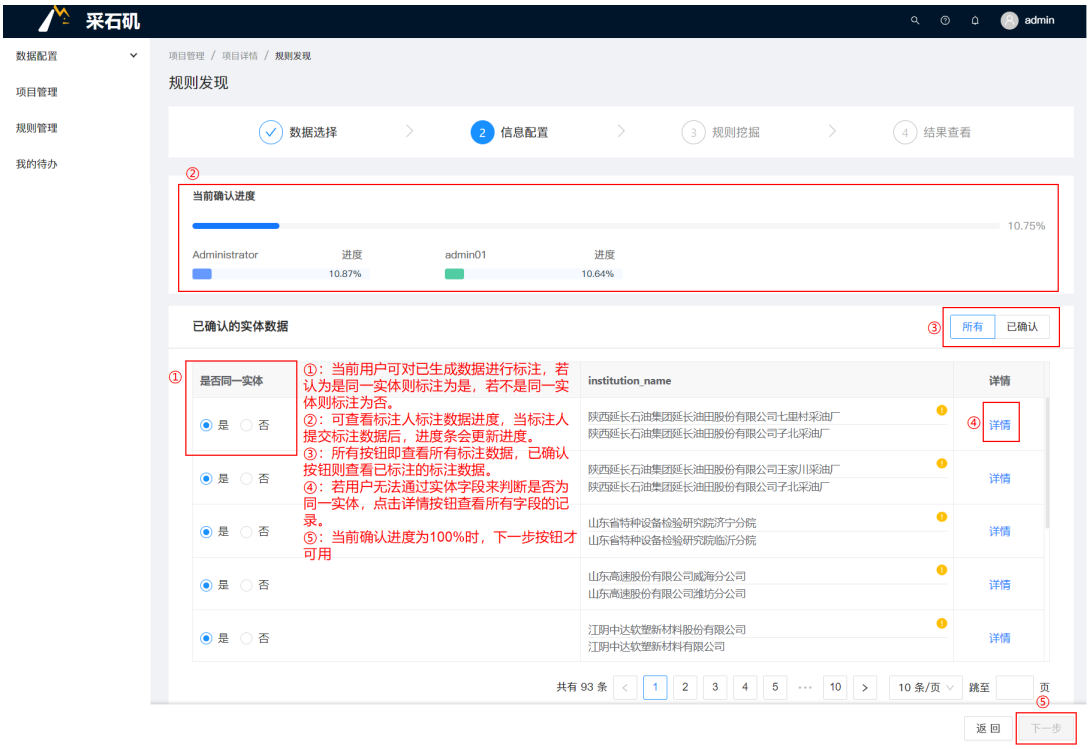
规则发现-我的待办

标注人进入到标注页面对数据进行标注。



规则发现-标注数据

当标注数据生成后，创建该ER规则发现任务的用户会进入到标注数据展示页。



规则发现-标注数据展示页

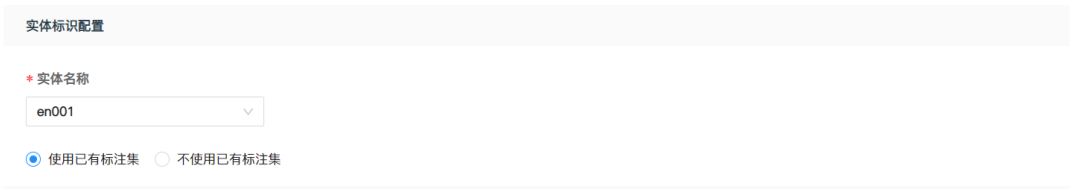
若当前用户对标注人的标注数据存疑，可对标注人标注的结果进行修改。

当标注人将所有数据标注完成后，当前用户的标注数据展示页的进度更新为100%，`下一步`按钮才从置灰变为可用。用户才能点击`下一步`按钮进入到规则挖掘页面。

- 选择实体，已生成标注集且使用已有标注集

  若选择的实体已生成标注数据，并且所有标注数据经过标注人的标注和提交，生成了对应的标注集，此时用户可选择是否使用已有标注集。

  此处选择使用已有标注集。



规则发现-使用已有标注集

由于选择使用已有标注集，此时不需要生成标注数据，因此当ER规则信息配置完成后，点击`下一步`按钮会直接进入到规则挖掘页面。

- 选择实体，已生成标注集但不使用已有标注集

  当选择不使用已有标注集时，需重新生成标注数据，因此用户要配置标注人和标注数据数量。

规则发现-不使用已有标注集

当ER规则信息配置完成后，点击`下一步`按钮会进入到生成标注数据页面，并将标注数据平均分发给配置的标注人。

生成标注数据的后续操作与选择实体，未生成标注集时的一致，在此不赘述。

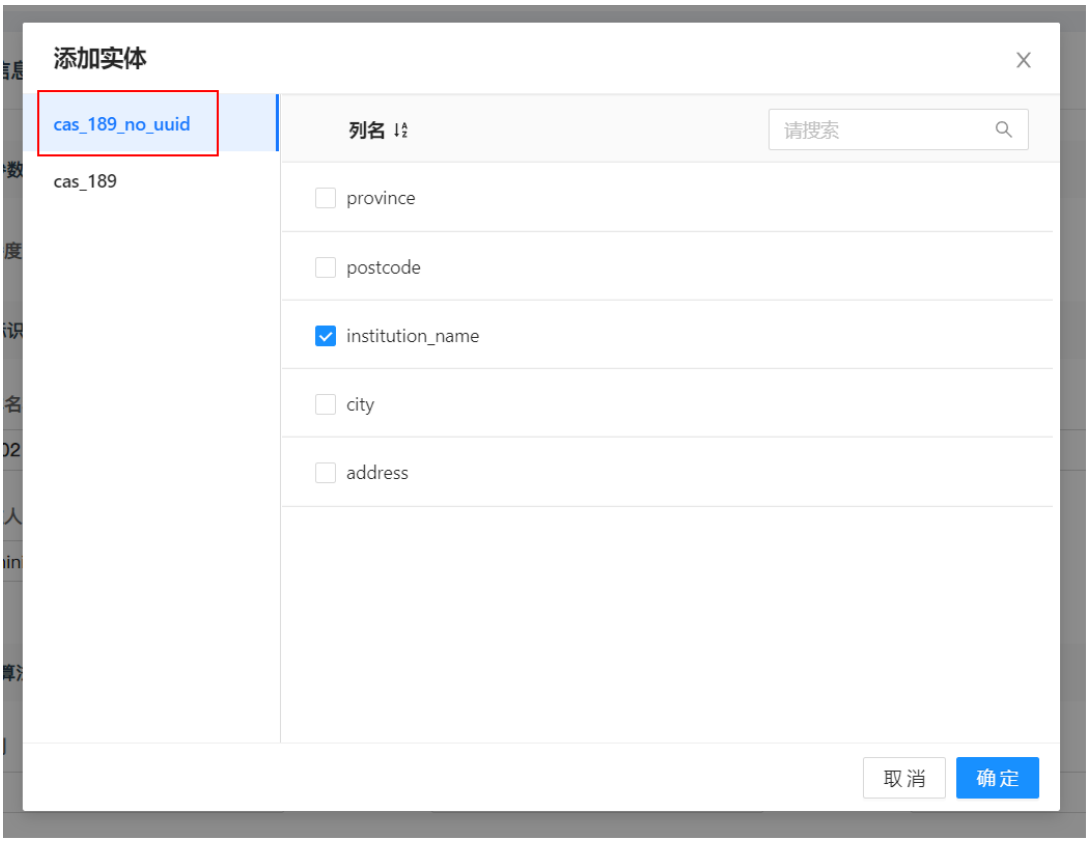
- 新建实体

  若用户不选择已有的实体，可直接在实体标识配置中新建实体。在实体名称的输入框中输入实体名称，例如这里将实体命名为"en002"。

规则发现-新建实体

新建实体需设置实体标识的字段。若用户想新建单表的实体，则只需配置左边字段，右边字段不需配置。点击左边的字段中的 `修改` 按钮就会出现添加实体弹框，勾选列后点击 `确定` 按钮即可。添加实体时列名可单选也可多选，但多选时只能对同一数据集的列进行多选，不可切换数据集勾选列名。

若用户想要生成跨表实体，则需同时配置左边字段和右边字段。此时右边字段选择的数据集与左边字段选择的数据集不同。



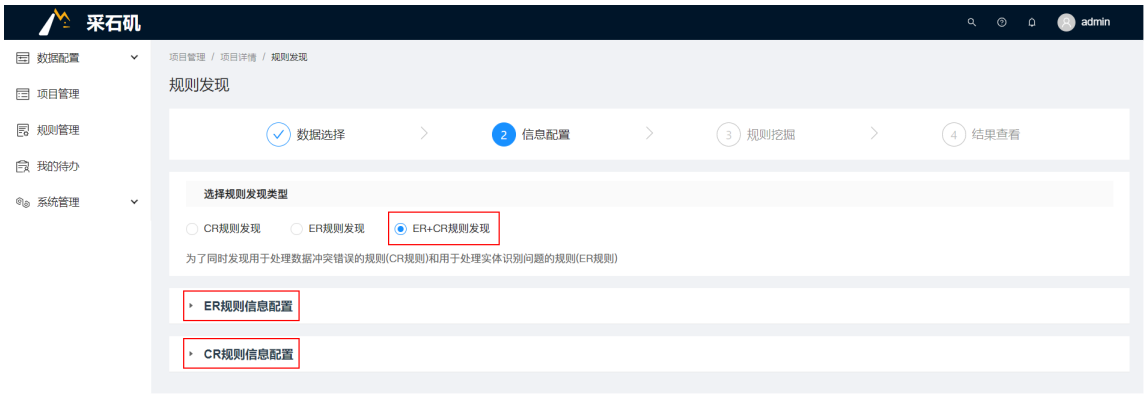
规则发现-左边字段的配置

当ER规则信息配置完成后，点击 `下一步` 按钮会进入到生成标注数据页面，并将标注数据平均分发给配置的标注人。生成标注数据的后续操作与选择实体，未生成标注集时的一致，在此不赘述。

3. ER+CR信息配置

选择规则发现的类型为ER+CR规则发现。

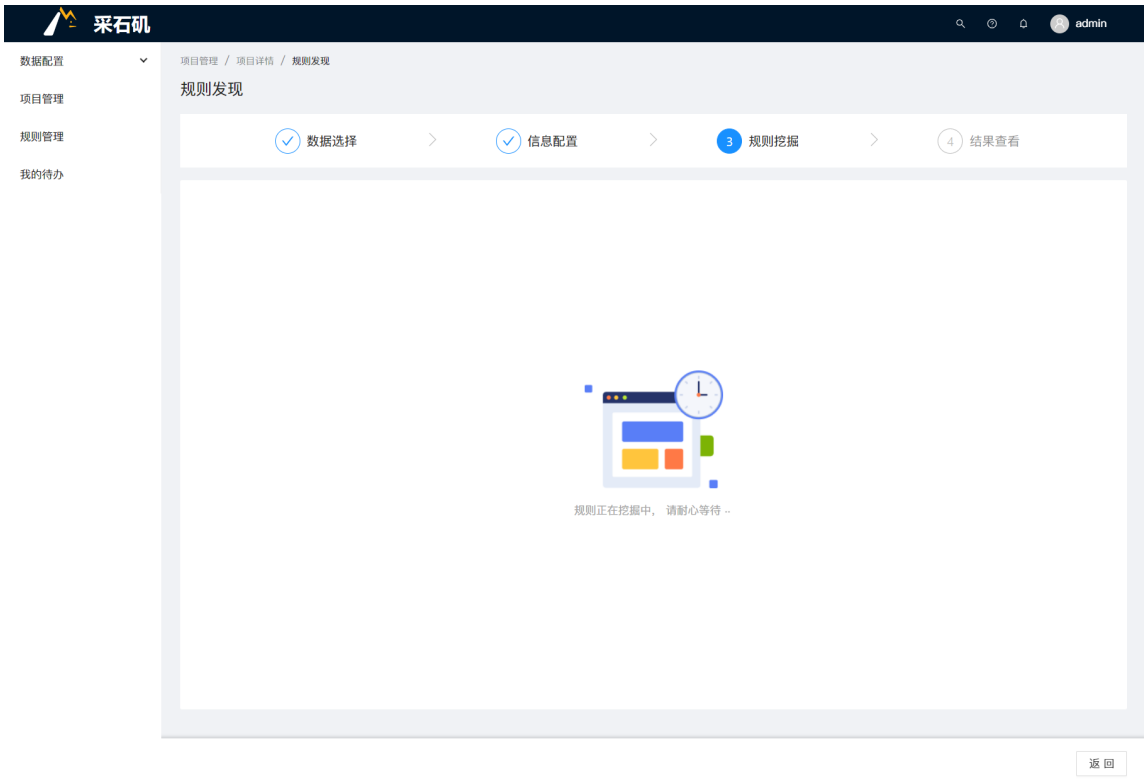
下图中的ER规则信息配置和CR规则信息配置与ER规则发现中的信息配置和CR规则发现中的信息配置是一致的，流程也是一致的，只是同时将两者结合起来，在此不做赘述。



规则发现-ER规则信息配置和CR规则信息配置

规则挖掘

当信息配置完成后，点击 下一步 按钮，进入规则挖掘等待页面，此时等待采石矶系统生成规则。



规则发现-规则挖掘

结果查看

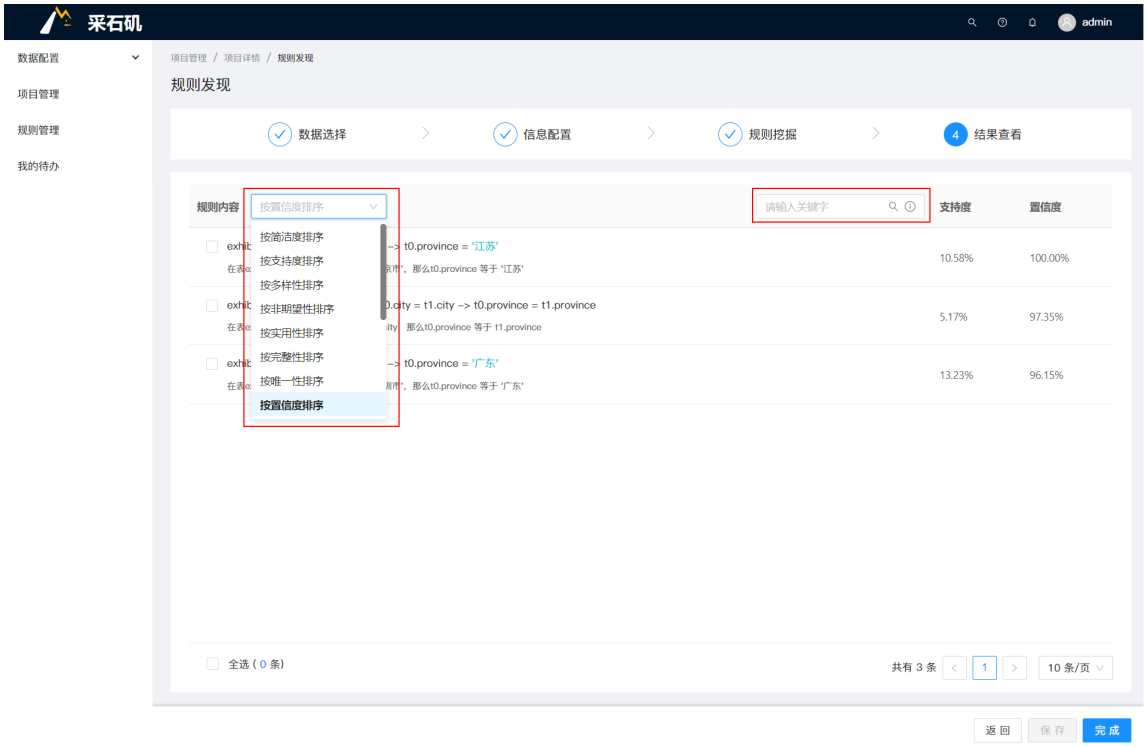
规则挖掘完成后，即进入到结果查看页面，以CR规则发现的结果展示为例，如下图所示。

在结果查看页面中，用户可根据自己的需求选择对应的排序方法。当前系统支持以下排序方法：简洁度、支持度、多样性、非期望性、实用性、完整性、唯一性、置信度、提升度、确信度、Top-100、Top-500和Top-1000。

排序方法	英文名称	描述
简洁度	Conciseness	描述规则的简洁程度。规则中谓词个数越少，分值越大。
支持度	Support	描述规则覆盖数据的程度。规则覆盖数据的个数越多，则分值越高。
多样性	Diversity	描述规则与其他规则的差异程度。规则如果与其他规则谓词差异越大，则分值越高。
非期望性	Unexpectedness	描述规则的出乎意料程度。规则中左边（LHS）和右边（RHS）关联性越小，则分值越高。
实用性	Utility	描述规则的实用程度。规则包含的实用属性列越多，则分值越高。
完整性	Completeness	描述规则所在列的数据完整程度。规则包含的属性列中空值越少，则分值越高。
唯一性	Uniqueness	描述规则所在列的数据不重复程度。规则包含的属性列中冗余数值越少，则分值越高。
置信度	Confidence	描述规则的可信程度，规则同时满足LHS和RHS的数据在只满足LHS数据的占比，比率越高，则分值也高。
提升度	Lift	描述规则中LHS和RHS的相关性。他们的相关性越高，则分值越高。
确信度	Conviction	描述规则中LHS和RHS同时出现的期望频率的占比（是Confidence的一个变体）。占比越高，则分值越高。
TopK	TopK	根据用户的喜好（结合用户主观特征和客观特征），通过学习的方式对规则进行打分。

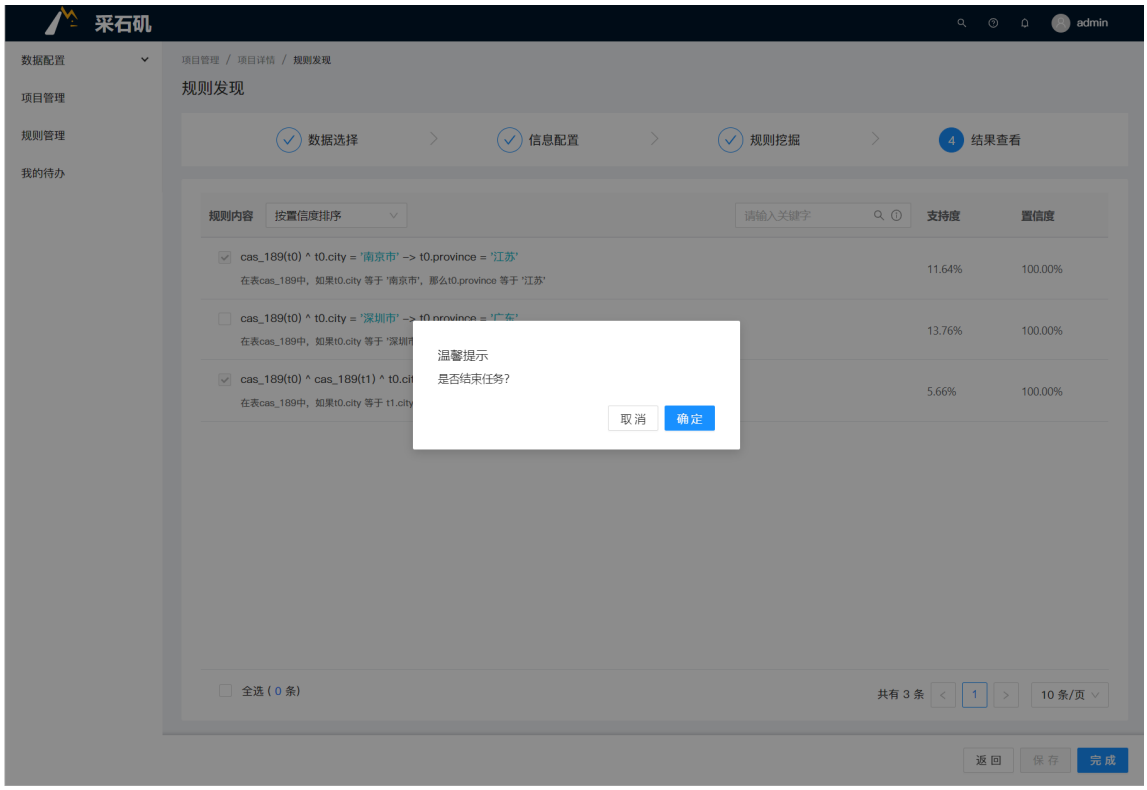
用户可在右方的搜索框输入关键字对规则进行过滤。

对于生成的规则，用户勾选自己所需的规则后，点击 保存 按钮，将规则保存至规则库中，以便后续其他任务的执行。保存后的规则不可重复保存，因此所对应的勾选框是置灰的。



规则发现-CR规则结果查看

当用户选择自己所需的规则并保存后，可点击 完成 按钮，会出现弹框提示用户是否结束任务。当点击 确定 按钮后，会有相应提示任务已完成，即当前的CR规则发现任务的整个流程已结束。对于已完成的规则发现任务，结果查看中的所有规则是不可操作的，即所有规则的勾选框是置灰的。

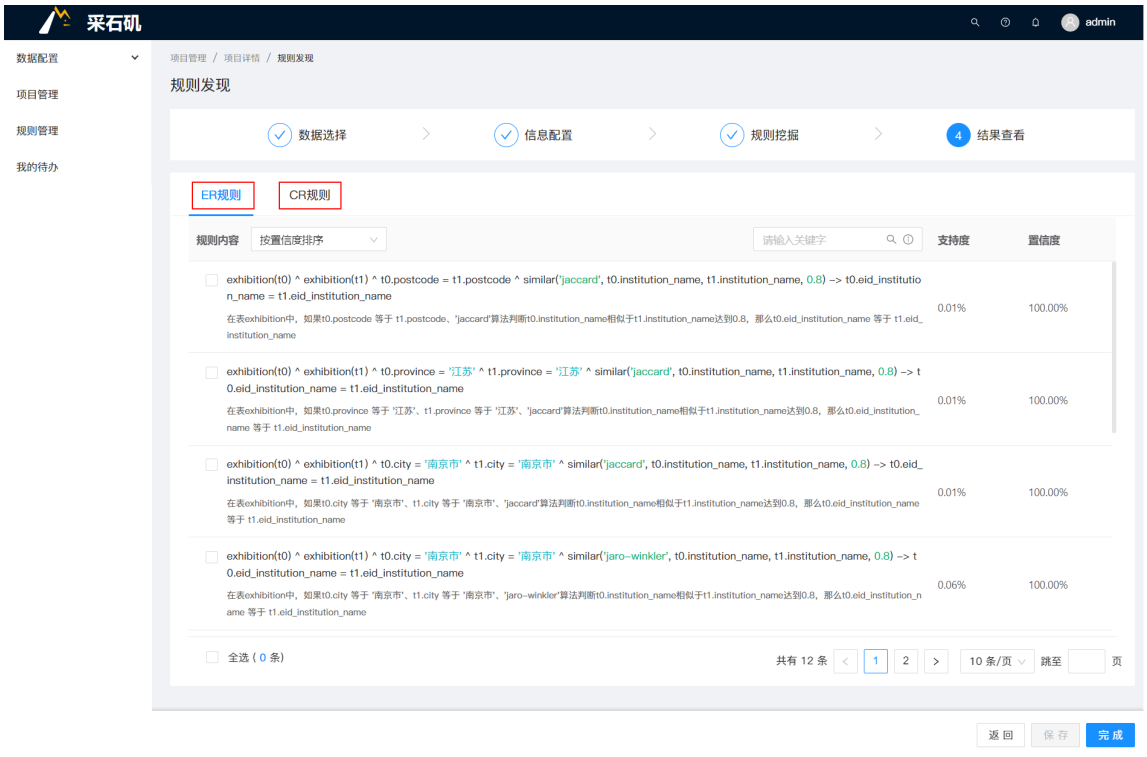


规则发现-结束任务

此外，ER规则发现的结果查看和ER+CR规则发现的结果查看与CR规则发现的结果查看基本一致，只是展示的规则内容不同。

ER规则发现只展示ER规则，而ER+CR规则发现同时展示ER规则和CR规则，用户可切换tab页进行查看。

规则发现-ER规则发现结果查看



规则发现-ER+CR规则发现结果查看

查错

本章节主要介绍采石矶系统中查错的流程与操作方法。

在我们现实生活中，存在大量的脏数据，例如：深圳所在的省份广东省被错误的写成了广西省。在大数据下，如果想找出这样的错误数据，需要耗费大量的人力。采石矶系统提供了查错功能，用户需要输入CR规则，采石矶能通过规则执行找出冲突数据，从而提升业务数据的质量。

此外，正则规则也可应用于查错，通过查错可以找出不符合正则规则的冲突数据。

前置条件

需满足如下条件：

- 用户已登录。
- 系统中已有数据集且同步成功。
- 已为数据集创建CR规则或正则规则。

查错操作总体流程图如下。

查错操作流程图

操作说明

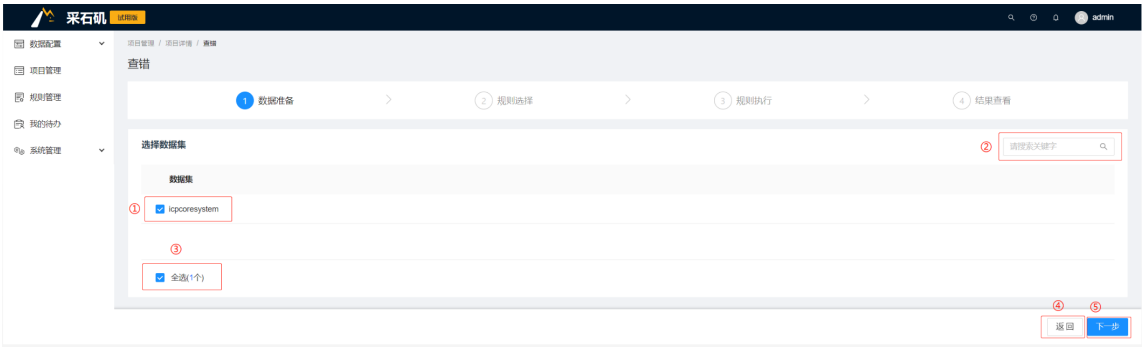
1. 新建任务

在新建工作流页面拖入数据集组件，选择数据集后，拖入查错任务组件并连线。保存和启动工作流后，会进入查看工作流页面。在查看工作流页面点击查错任务组件，配置任务信息。

查错任务共有四个阶段，分别是数据准备、规则选择、规则执行和结果查看，具体操作介绍如下：

2. 数据准备

在查看工作流页面点击查错任务组件，会进入数据准备页面，具体呈现如下图。

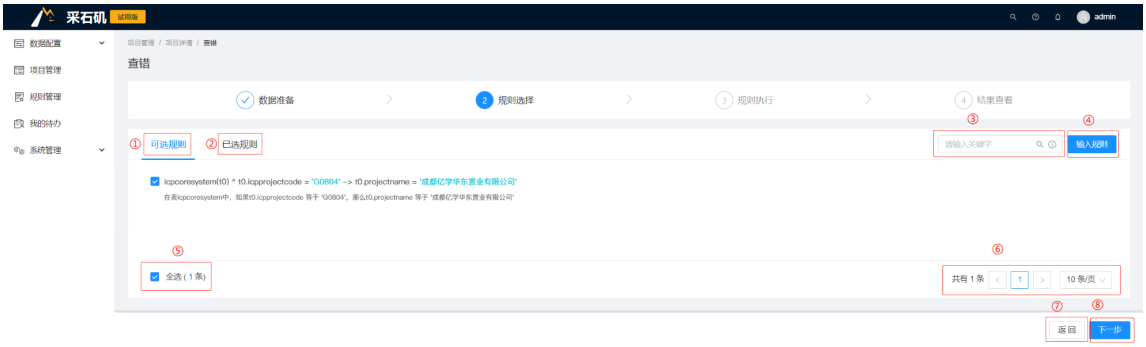


数据准备配置界面

1. 选择数据集：用户可以根据需要选择执行查错任务的数据集，数据集可以单选和多选。
2. 搜索数据集：用户可以输入搜索条件搜索已有的数据集，数据集的搜索支持精确查询和模糊查询。
3. 全选：用户可以通过全选按钮勾选当前页的所有数据集。
4. 返回：用户可以点击此按钮返回至项目详情页面。
5. 下一步：勾选完当前页的数据集后可以点击下一步按钮进入到下一步骤。

3. 规则选择

点击 `下一步` 按钮，进入选择CR规则或正则规则的页面，具体呈现如下图。



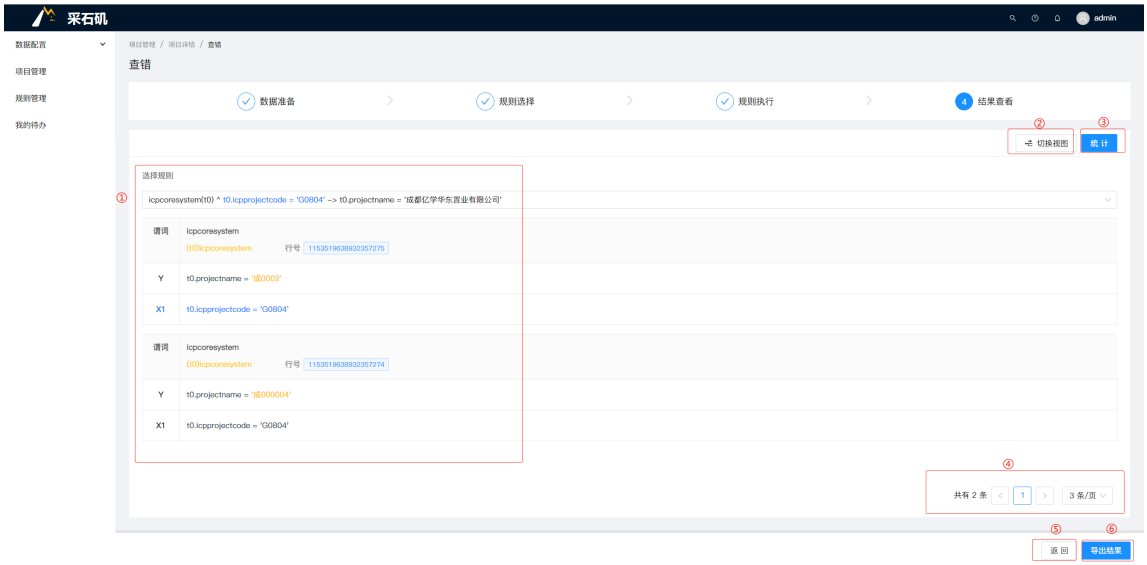
规则选择的界面

1. 可选规则：用户可以根据可选规则列表中的规则，选择想要执行的规则。
2. 已选规则：已勾选的规则，可以在已选规则列表中查看。
3. 搜索规则：用户可以输入搜索条件搜索已有的规则，规则的搜索支持精确查询和模糊查询。
4. 输入规则：当用户未创建规则或当前规则不满足需要时，用户可以点击 `输入规则` 快速跳转至规则管理页面。

5. 全选：用户可以通过全选按钮勾选当前页的所有规则。
6. 分页器：用户可以通过分页器实现快速查看规则的操作，也可以控制当前页面展示规则的数量。
7. 返回：用户可以点击此按钮返回至项目详情页面。
8. 下一步：当完成当前页的配置后可以点击 `下一步` 按钮进入到下一步骤。

4. 规则执行&结果查看

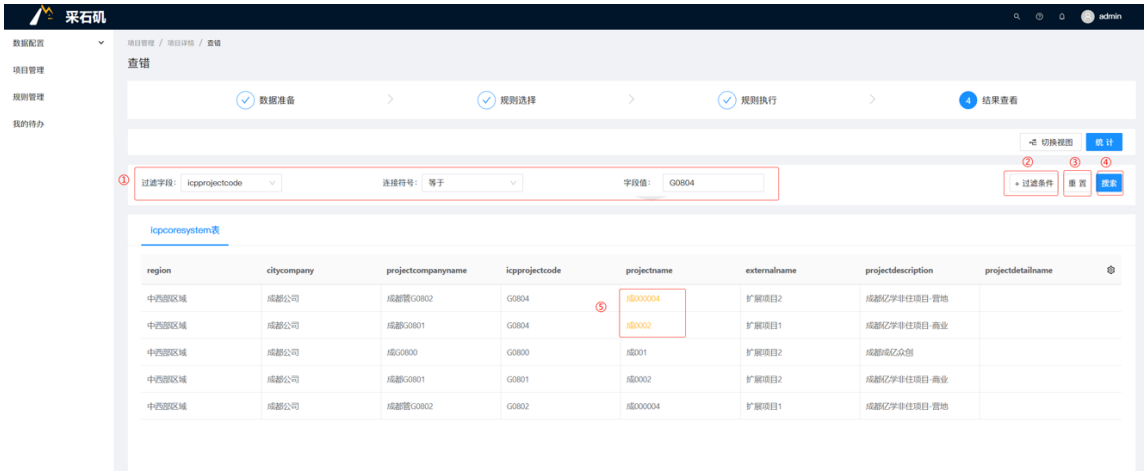
点击 `下一步` 按钮，会进入到规则执行页面，规则执行完成之后会自动进入结果查看页面，具体呈现如下图。



纠错结果查看的界面

1. 选择规则：用户可以选择规则查看此条规则执行的查错结果。
2. 切换视图：在原表数据中，通过高亮的形式展示冲突数据。
3. 统计：用户可以通过统计查看该任务中执行的规则数量、查错的数据集数量、找出的冲突数据条数。
4. 分页器：用户可以通过分页器实现快速查看冲突数据。
5. 返回：用户可以点击此按钮返回至项目详情页面。
6. 导出结果：导出一个压缩文件，在压缩文件中每一个规则对应一个csv文件。

点击 `切换视图` 按钮，具体呈现如下图。



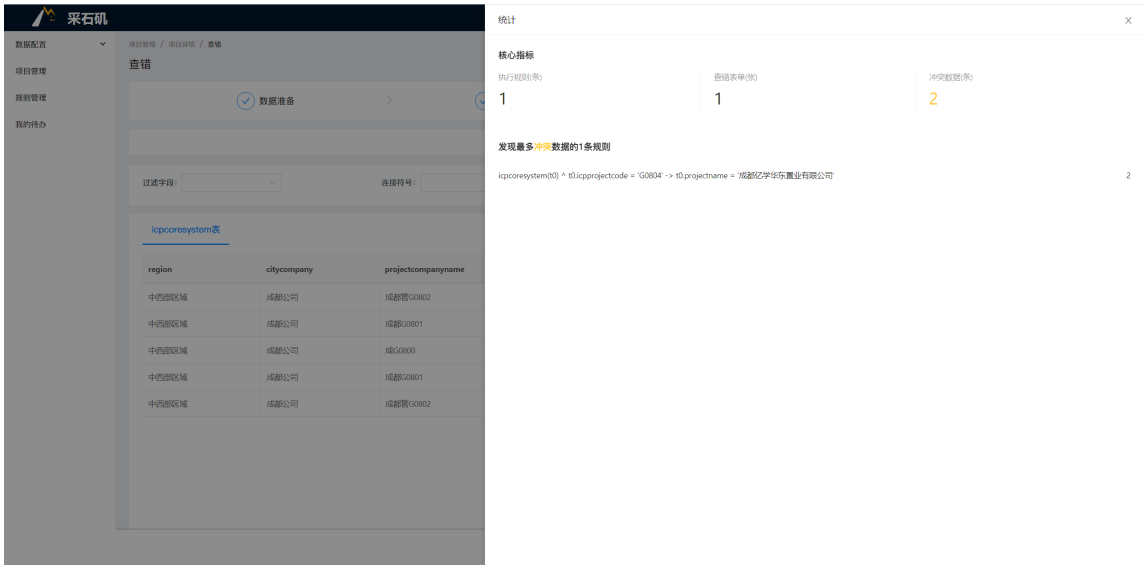
切换视图的界面

1. 在过滤条件中，添加过滤字段、连接符号、字段值。
2. 过滤条件：新增过滤条件。
3. 重置：清空已输入的过滤字段、连接符号以及字段值。
4. 搜索：按过滤条件进行搜索。
5. 高亮数据展示：冲突数据。

字段类型支持的连接符号

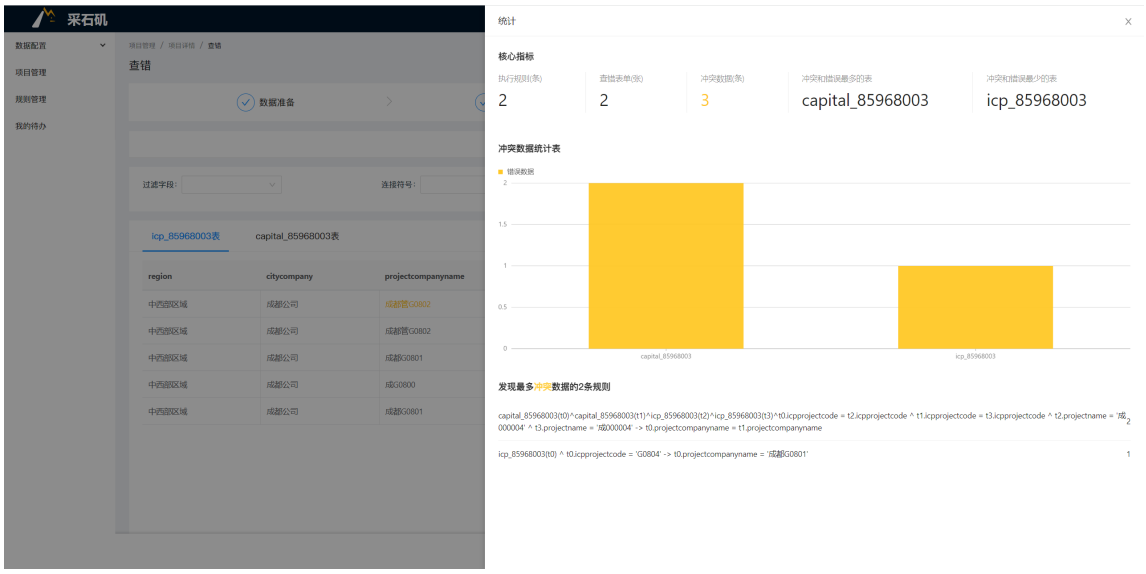
字段类型	连接符号
字符串型	包含，不包含，为空，非空，等于，不等于
数值型	大于，小于，大于等于，小于等于，等于，不等于，为空，非空

点击 统计 按钮，单表的统计呈现如下图。



统计的界面

多表的统计呈现如下图。



多表统计的界面

- 多表的查错统计相比较单表而言，新增了冲突和错误数据最多的表、冲突和错误数据最少的表、和各表的冲突数据数量统计直方图。

数据纠错

本章节主要介绍采石矶系统中数据纠错的流程与操作方法。

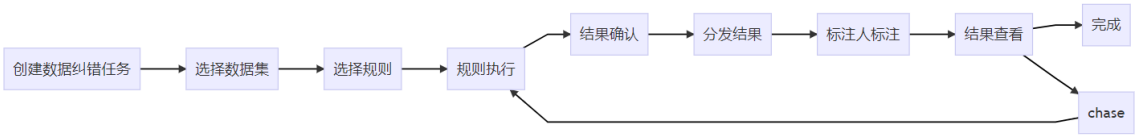
在我们日常的生活中，存在着大量的错误数据，如城市为广州，省份应该为广东省但却被错误的写成了湖北省，这样的数据想要靠人工来修复需要消耗巨大的人力和物力。采石矶系统提供了数据纠错的功能来解决这个问题，用户只需要输入CR规则，系统会自动识别出不满足规则的数据，同时也可以对不满足规则的数据进行修改，大量地减少了人力和物力。

前置条件

需满足如下条件：

- 用户已登录。
- 系统中已有数据集且同步成功。
- 已为数据集设置可信度。
- 已为数据集创建CR规则。

数据纠错总体操作流程图如下。



数据纠错操作流程图

数据纠错操作说明

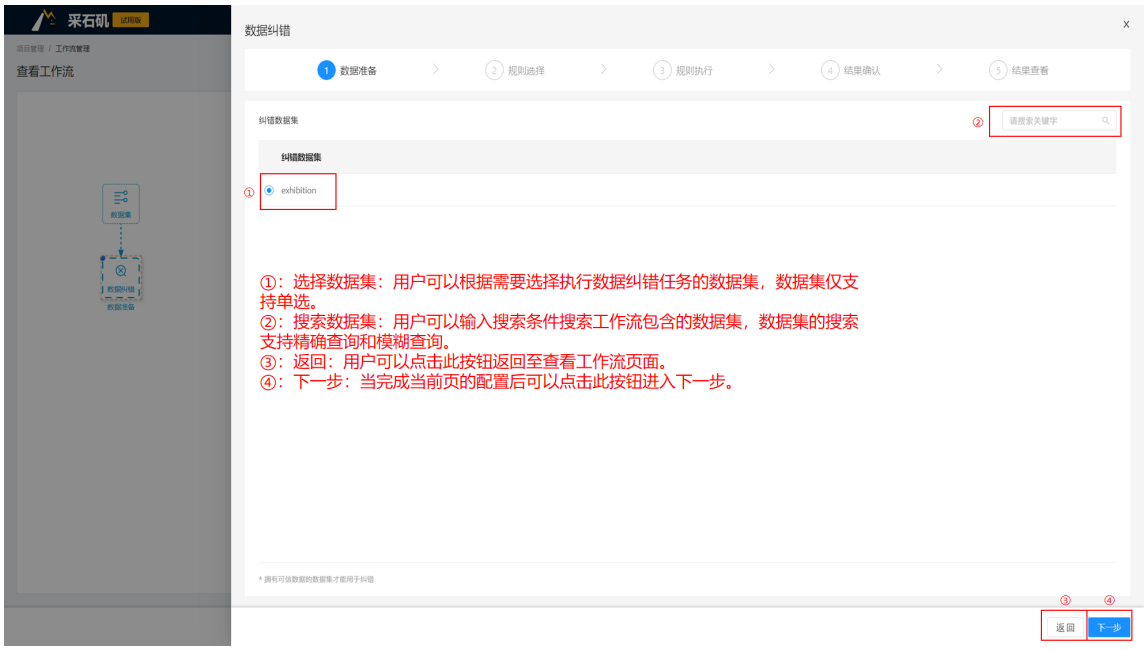
1. 新建任务

在新建工作流页面拖入数据集组件，选择数据集后，拖入数据纠错任务组件并连线。保存和启动工作流后，会进入查看工作流页面。在查看工作流页面点击数据纠错任务组件，配置任务信息。

数据纠错任务共有五个阶段，分别是数据准备、规则选择、规则执行、结果确认和结果查看，具体操作介绍如下：

2. 数据准备

在查看工作流页面点击数据纠错任务组件，会进入数据准备页面，具体呈现如下图。



数据纠错任务数据准备页面

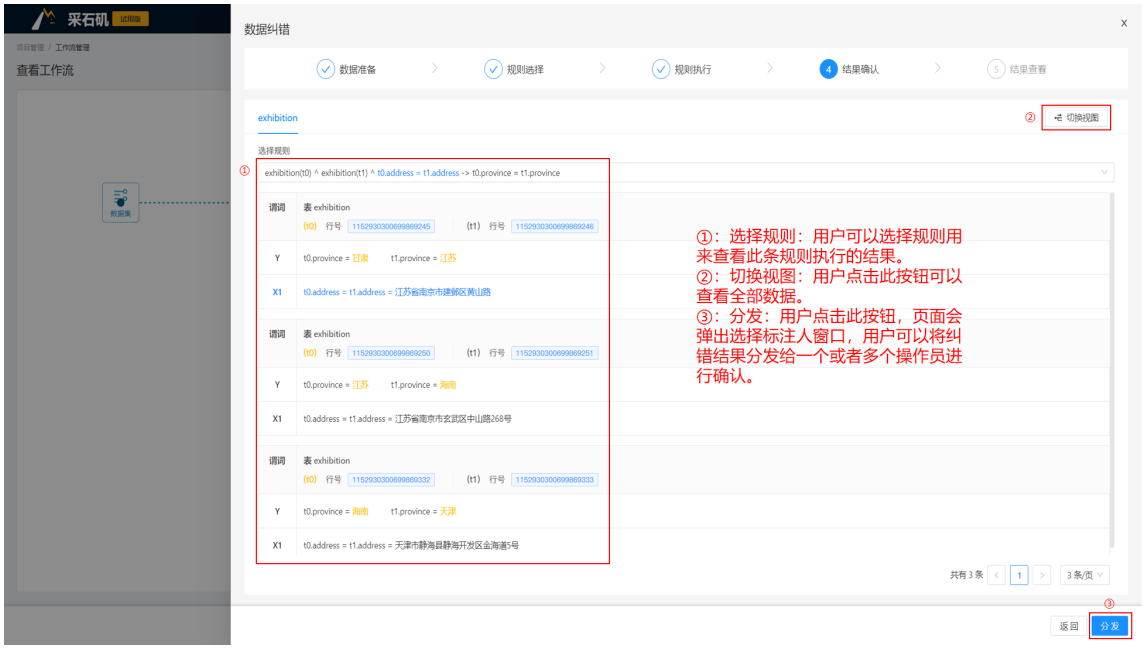
3.规则选择

点击 下一步 按钮，会进入规则选择页面，具体呈现如下图。

数据纠错任务规则选择页面

4. 结果确认

点击 下一步 按钮，会进入规则执行页面，规则执行完成后会自动进入结果确认页面，具体呈现如下图。



数据纠错任务结果确认页面

纠错结果分发给操作员后，使用操作员的账号登录采石矶系统，在我的待办页面可以查看到分发的任务，具体呈现如下图。

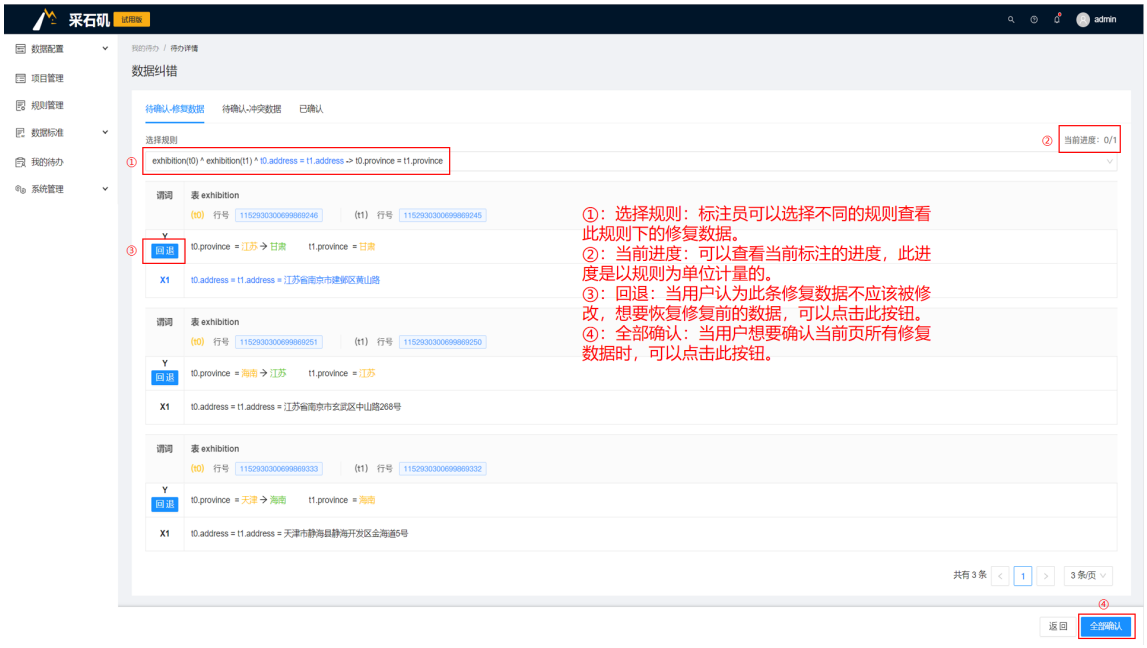
标注员我的待办页面

点击任务名可以进入到待办详情页面，该页面一共有三个页签，分别是待确认-修复数据、待确认-冲突数据和已确认数据。

修复数据、冲突数据说明：

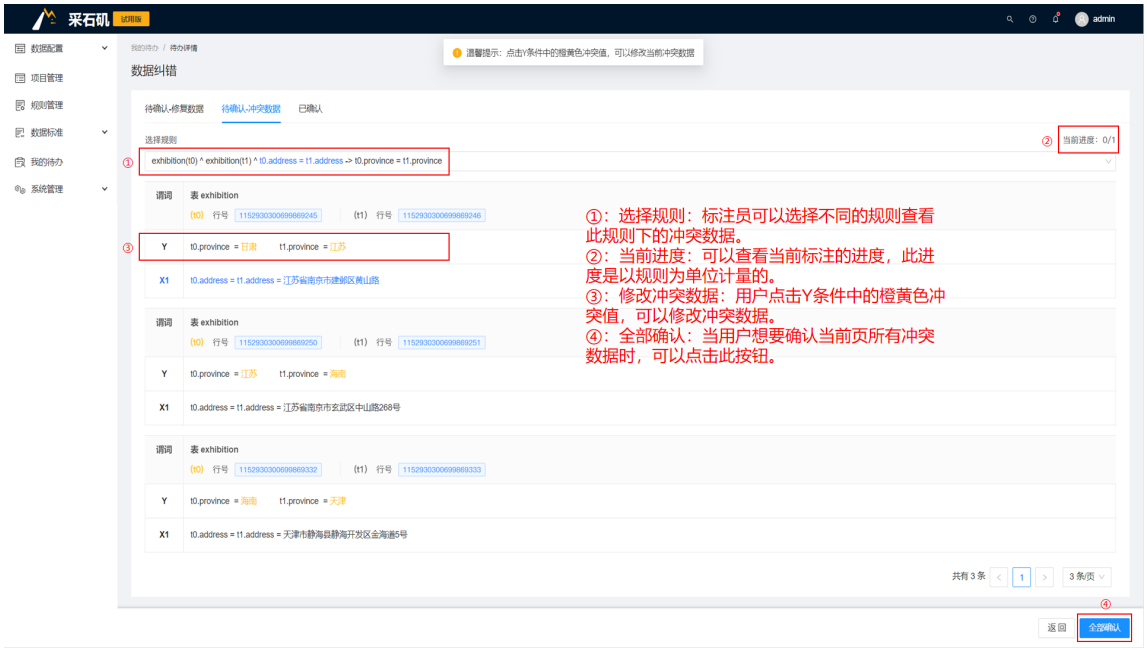
在CR规则执行过程中，系统会根据CR规则找出不满足规则的数据，然后判断数据的可信度是否满足要求。如果可信度满足要求，则系统会根据CR规则将这条数据修正，同时将被修正后的数据设为可信数据，这样的数据即为修复数据。如果可信度不满足要求，则系统不会对此条数据进行修改，这样的数据即为冲突数据。

标注员进入待办详情页，可以查看待确认的修复数据，具体呈现如下图。



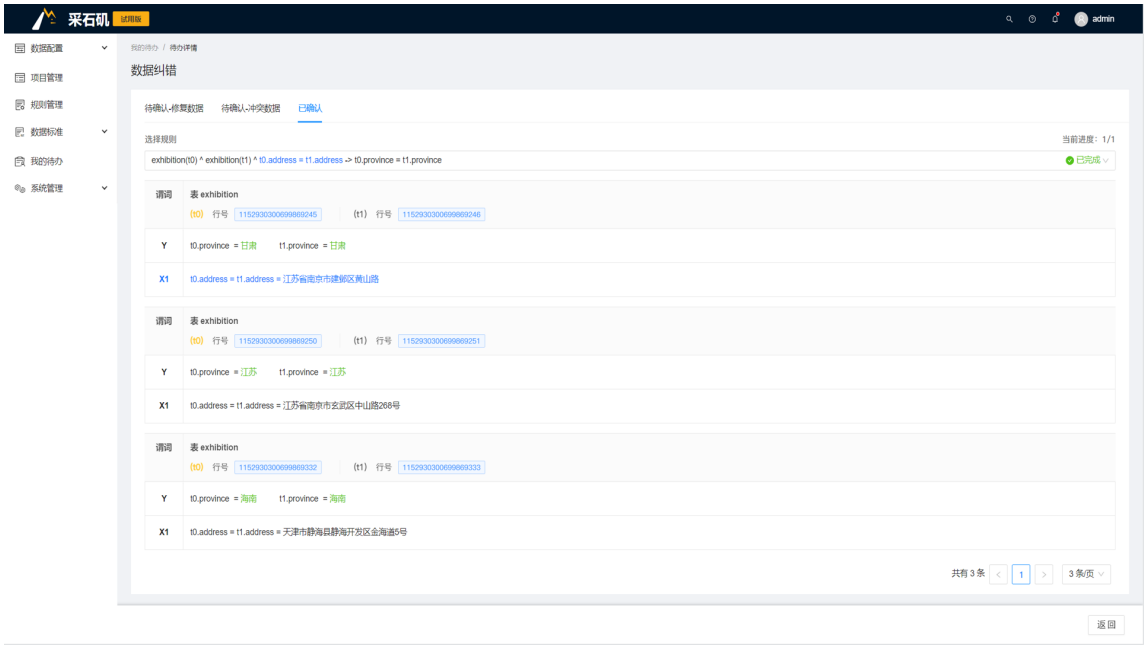
标注员我的待办-待确认-修复数据页面

标注员进入待办详情页，可以查看待确认的冲突数据，具体呈现如下图。



标注员我的待办-待确认-冲突数据页面

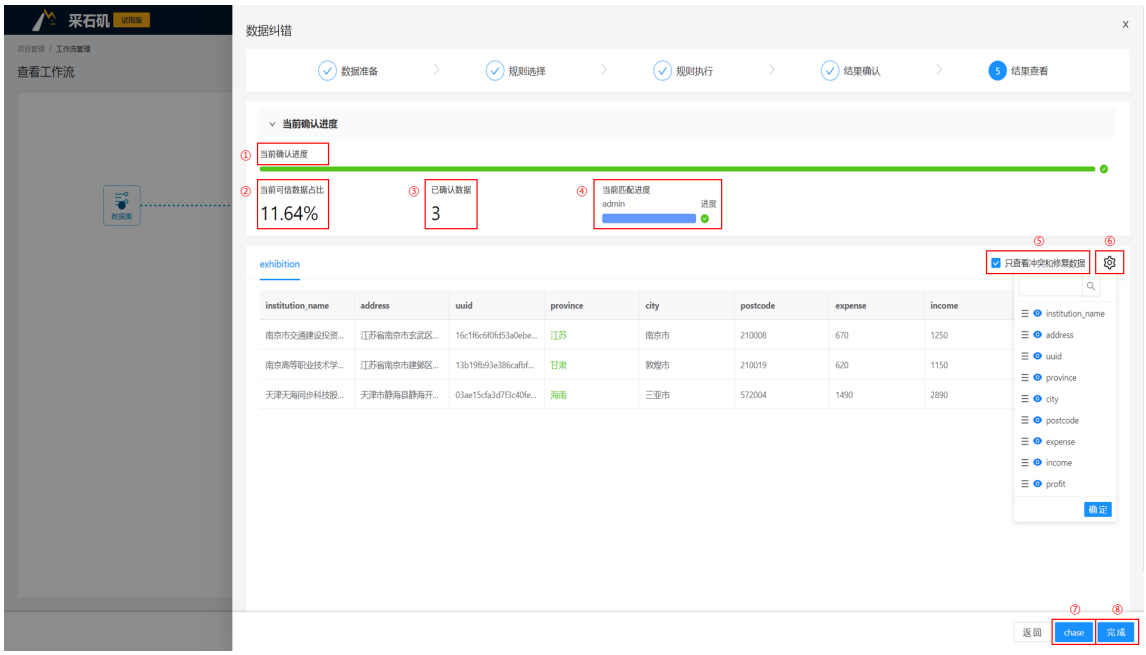
确认后的数据会移动到已确认数据中，具体呈现如下图。



标注员我的待办-已确认数据页面

5. 结果查看

标注员确认数据后，管理员可以查看当前数据的确认情况。



数据纠错任务结果查看页面

1. 当前确认进度：已确认的数据在所有待确认的数据中的占比。
2. 当前可信数据占比：当前可信的数据在所有数据中的占比。
3. 已确认数据：当前已确认的数据。
4. 标注员当前匹配进度：分发到标注员的数据中，已确认的数据在所有数据中的占比。

5. 勾选此勾选框则只查看冲突数据和修复数据，不勾选则查看完整数据集。
6. 选择需要查看的列：用户可以勾选具体的列名来展示对应的数据。
7. chase按钮：当用户已经对当前的数据进行确认后，期望能对当前确认的结果进行优化时，可以点击 `chase` 按钮，系统会回到规则执行步骤，其他步骤和上面流程一致，不再赘述。
8. 完成按钮：用户点击 `完成` 按钮后可选择导出结果，在导出结果窗口可选择导出部分字段和全部字段。

实体聚类

本章节主要介绍采石矶系统中实体聚类的流程与操作方法。

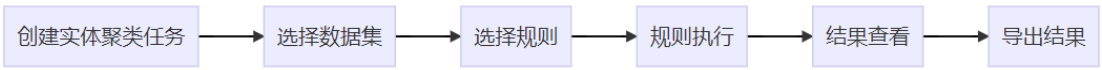
实体聚类是在已有的样本数据里，根据实体规则将具有同样数据特征的数据归为一个实体，用户只需要选择数据集并为数据集创建合适的实体规则，系统将自动识别出实体数据。

前置条件

需满足如下条件：

- 用户已登录。
- 系统中已有数据集且同步成功。
- 已为数据集创建实体。
- 已为数据集创建实体规则。

实体聚类总体操作流程图如下。



实体聚类操作流程图

实体聚类操作说明

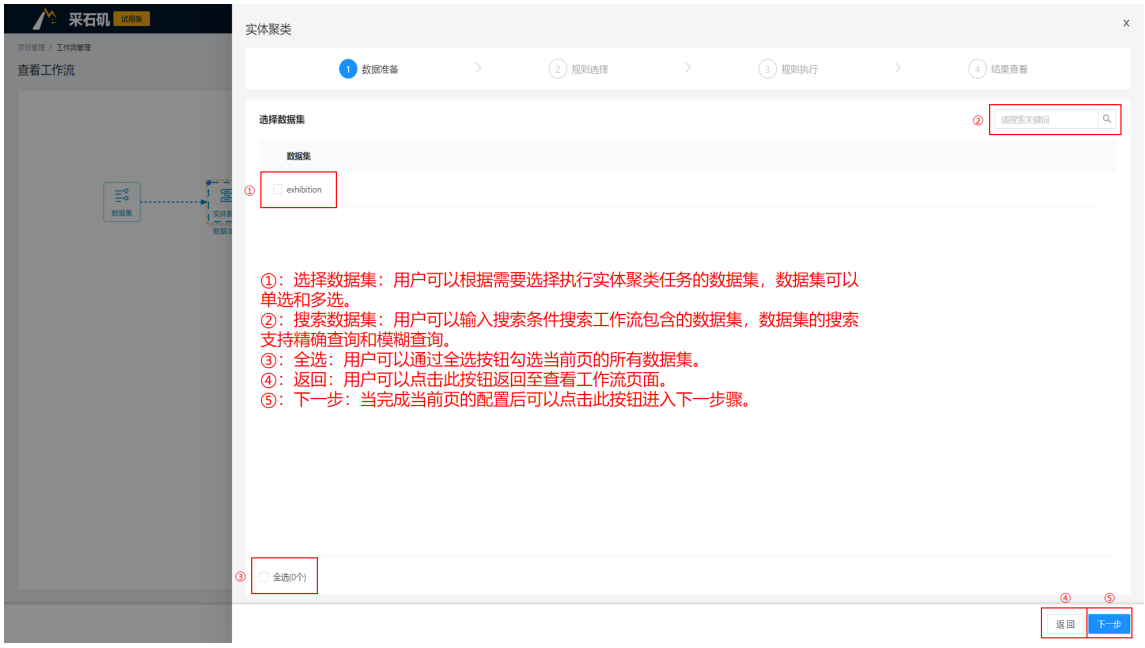
1. 新建任务

在新建工作流页面拖入数据集组件，选择数据集后，拖入实体聚类任务组件并连线。保存和启动工作流后，会进入查看工作流页面。在查看工作流页面点击实体聚类任务组件，配置任务信息。

实体聚类任务共有四个阶段，分别是数据准备、规则选择、规则执行和结果查看，具体操作介绍如下：

2. 数据准备

在查看工作流页面点击实体聚类任务组件，会进入数据准备页面，具体呈现如下图。



实体聚类任务数据准备页面

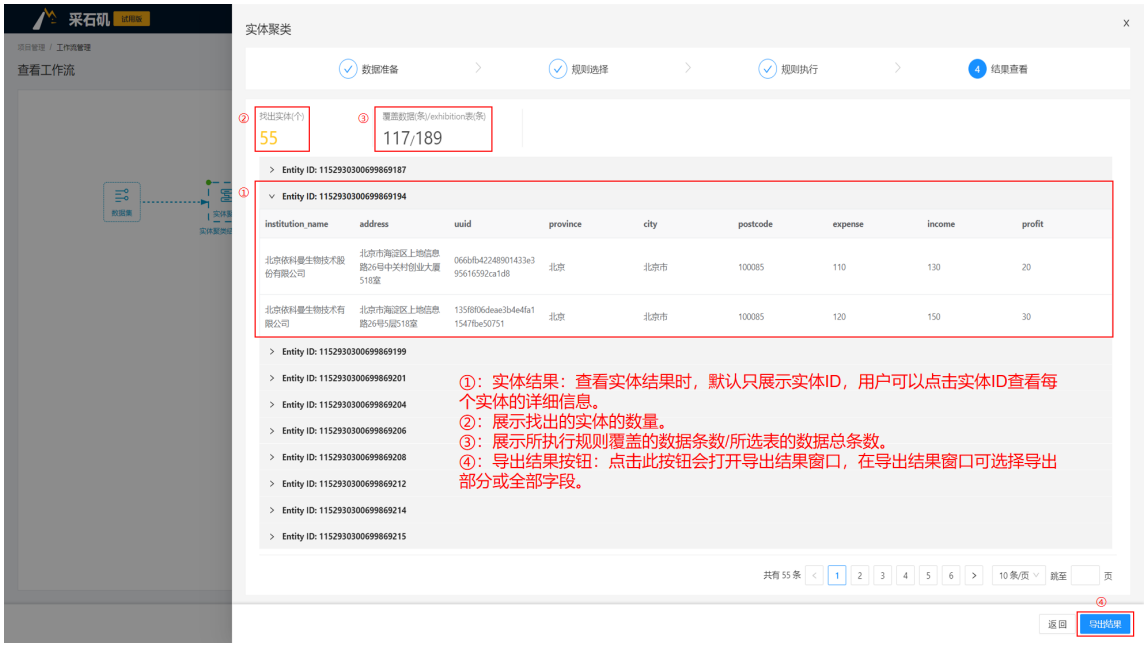
3、规则选择

点击 下一步 按钮，会进入规则选择页面，具体呈现如下图。

实体聚类任务规则选择页面

4. 结果查看

点击 下一步 按钮，会进入实体规则执行页面，实体规则执行完成后会自动进入结果查看页面，具体呈现如下图。



实体聚类任务结果查看页面

最优记录

本章节主要介绍在采石矶系统中执行最优记录任务的流程和操作方法。

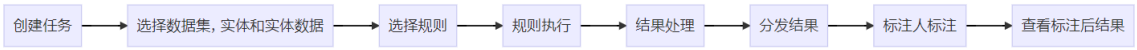
在了解最优记录任务之前，首先应了解实体聚类的概念。实体聚类是在已有的样本数据里，根据实体规则将具有同样数据特征的数据归为一个实体。最优记录功能是在同属于一个实体的数据中，通过获取每个字段的最优值从而得到实体的最优记录。用户只需要为所需字段创建合适的最优规则和CR规则，系统会对实体聚类任务输出的实体数据执行这些规则，进而获得每个实体的最优记录。

前置条件

须同时满足以下6个条件：

- 用户已经登录；
- 已经创建了一个项目，项目关联了至少一个数据集；
- 已经创建了这个数据集包含的实体；
- 已经创建了这个实体关联的实体规则；
- 已经在这个数据集上执行过实体聚类任务，并得到了实体数据；
- 已经创建了这个实体包含的列的最优规则和CR规则。

最优记录任务操作流程图



最优记录任务操作流程图

操作说明

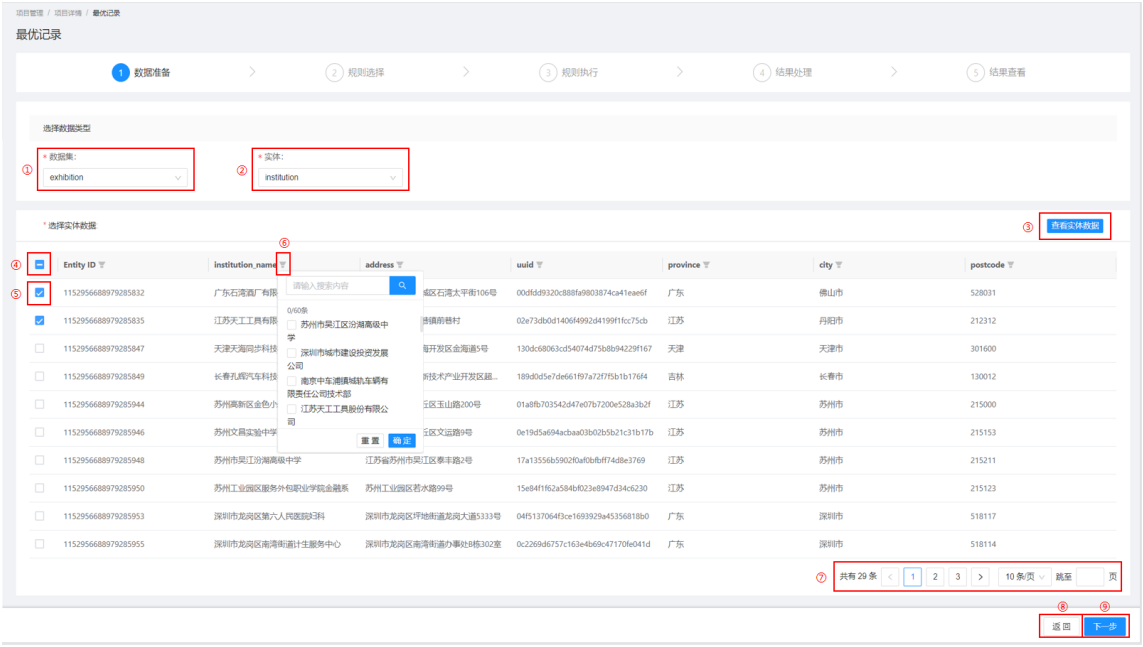
1. 新建任务

在新建工作流页面拖入数据集组件，选择数据集后，拖入最优记录任务组件并连线。保存和启动工作流后，会进入查看工作流页面。在查看工作流页面点击最优记录任务组件，配置任务信息。

最优记录任务共有五个阶段，分别是数据准备、规则选择、规则执行、结果处理和结果查看，具体操作介绍如下：

2. 数据准备

用户需要在数据准备页面选择数据集、实体和实体数据。数据准备页面具体呈现如下图。

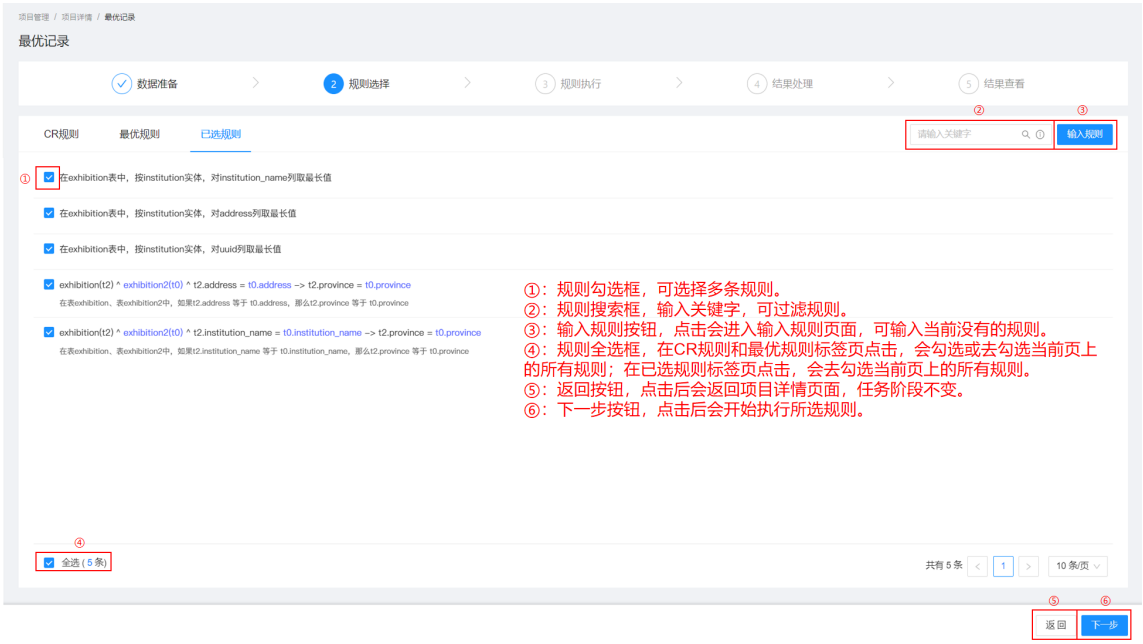


最优记录-数据准备

1. 数据集：用户可以选择执行最优记录任务的数据集，数据集的可选范围是项目选择的数据集，数据集仅支持单选。
2. 实体：用户选择了数据集之后，会获取到该数据集已执行过实体聚类任务的实体，用户可根据需要选择实体。选择实体后，此实体关联的实体记录会展示在实体列表中，如果同一个实体执行过多次实体聚类任务，那么系统会获取最新一次执行的结果。
3. 查看实体数据：一般情况下，一个实体包含多条实体记录，在实体列表中只会展示第一条实体记录，若用户想要查看实体完整的记录，可以点击 `查看实体数据` 按钮。
4. 全选框：点击会选中当前页上的所有实体数据。
5. 选择框：点击会选中该条实体数据，目前系统暂不支持实体记录超过999条的实体执行最优记录任务。因此，如果实体的记录超过999条，那么这条实体前的选择框将会被置灰，无法选择。
6. 过滤按钮：点击会打开搜索框，可以输入搜索条件，根据搜索出的结果，选择过滤条件。可以勾选一个或多个过滤条件，勾选后点击确定，过滤条件生效。页面会显示符合过滤条件的实体数据。
7. 分页器：用户可以通过分页器实现快速查看实体记录的操作，也可以控制当前页面展示的实体记录的数量。
8. `返回` 按钮，点击后会返回项目详情页面，在本页中所做的配置不会被保存。
9. `下一步` 按钮，点击后会进入规则选择页面。

3. 规则选择

在数据准备页面点击 下一步 ，会进入规则选择页面。该页面有三个标签页，分别是CR规则、最优规则和已选规则，用户至少需要选择一条规则才可以进行下一步操作，已经选中的规则会出现在已选规则标签页。具体呈现如下图。



最优记录-规则选择

在规则选择页面点击 下一步 ，采石矶系统会开始执行已选规则，并进入规则执行页面。

4. 结果处理

规则执行完成后，会根据规则执行的结果自动跳转至指定页面。如果规则执行的结果全部为推荐记录，那么用户无需对此结果进行处理，直接进入到结果查看页面。如果任务执行产生了缺失记录和执行失败的数据，那么将会进入到结果处理页面。

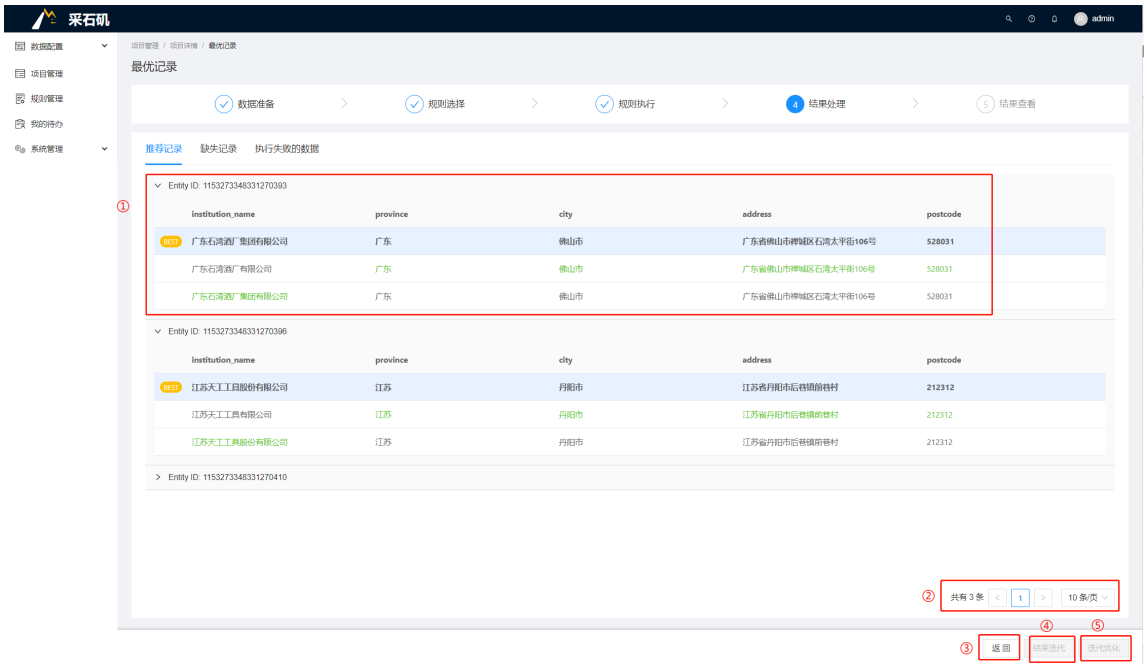
结果处理页面有三个页签，分别是推荐记录、缺失记录、执行失败的数据。

推荐记录、缺失记录、执行失败的数据说明：

1. 最优记录支持CR规则和最优规则的执行。最优规则是对现有的实体记录的值进行排序和加工，从而得到此实体的最优值。当最优规则无法满足需要时，也就是说无法从实体记录中获取此条实体的最优值，可以选择CR规则来获取实体的最优值，CR规则适用的场景是存在另外一张表作为主表，当满足一定条件时，可以用主表的值去填充某个实体的最优值。
2. 对于实体的某一列来说，只能选择CR规则或最优规则。如果此列选择的是最优规则，会根据规则判断是否能够得到最优值，如果最优规则定义的不全，系统无法得到最优值，会认为这一列的最优值是缺失的，并给出缺失值的可选值，用户需在结果处理步骤中对缺失的最优值进行保存。如果此列选择的是CR规则，当系统未找到该列对应的最优值的时候，也无法给出有意义的可选值，因此用户无需在结果处理步骤对此类值进行确认。

3. 当CR规则在执行的过程中，发现某个实体的同一列得到了两个不同的最优值，则会认为此条实体是执行失败的实体。否则则会判断实体是否存在需要保存的缺失值时，若存在，则认为此条实体是缺失记录，用户可以在结果处理步骤中进行保存，否则则认为此条实体是推荐记录。

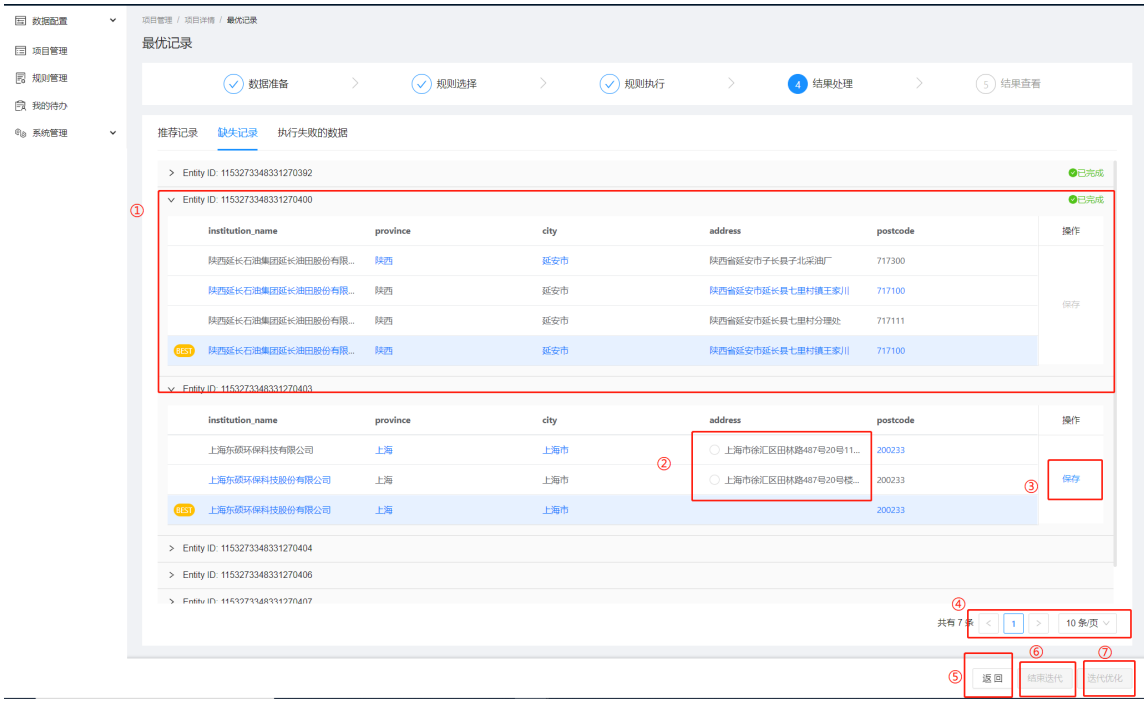
推荐记录页面具体呈现如下图。



最优记录-推荐记录

1. 最优记录执行结果：用户可以点击实体ID查看具体的实体数据，其中第一行带有BEST标记的是规则执行出的最优值。图中绿色高亮显示的是当前列最优值来源的实体值。
2. 分页器：用户可以通过分页器实现快速查看推荐记录的操作，也可以控制当前页面展示的推荐记录的数量。
3. 返回按钮：点击 返回 按钮可以返回至项目详情页面。
4. 结束迭代：用户可以点击 结束迭代 按钮结束当前迭代任务，进入到结果查看页面。应注意的是，若当前任务存在缺失记录，需要将缺失记录保存后才可以点击 结束迭代 按钮，否则按钮是置灰状态。
5. 迭代优化：若当前任务存在执行失败的数据，用户可以点击 迭代优化 按钮对执行失败的数据进行优化，点击后页面会跳转至数据准备页面，用户可以重新录入实体和规则执行最优记录任务。

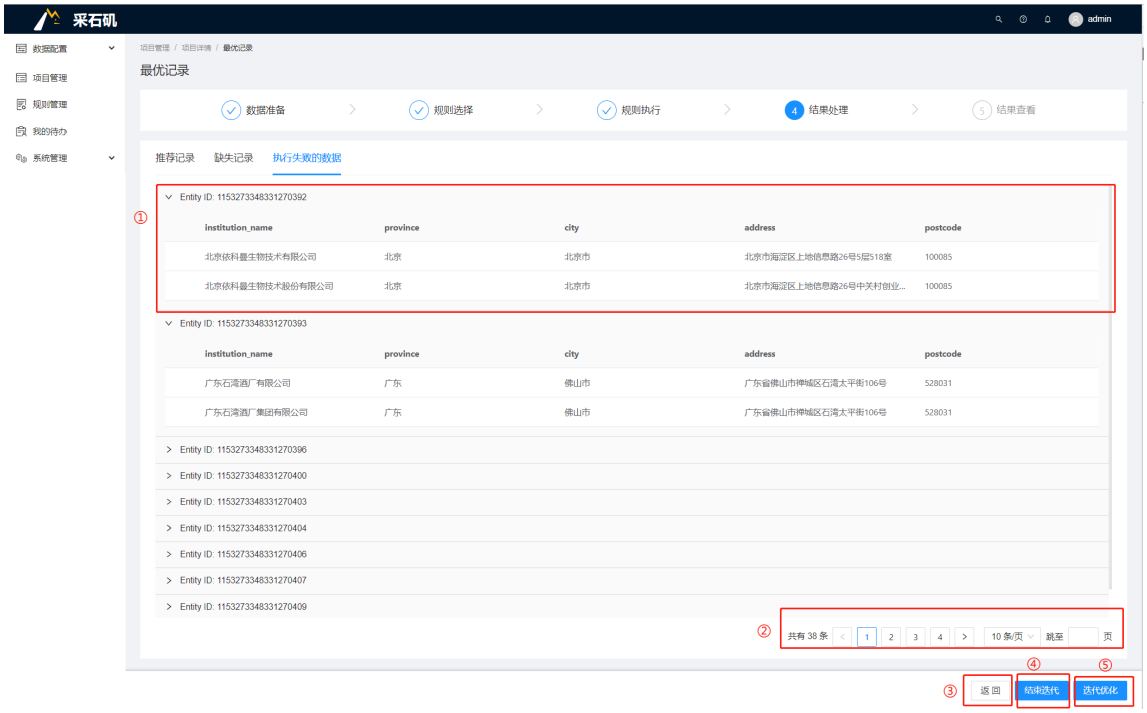
缺失记录页面具体呈现如下图。



最优记录-缺失记录

1. 已保存的缺失记录：若系统中存在缺失记录，用户可以点击实体记录前的单选按钮选择此列的最优值，点击保存后会将此实体的值存储到数据库中，保存后 `保存` 按钮将置灰并显示当前实体的状态为已完成。
2. 可选值：用户可以从可选值中选择此条实体的最优值。
3. 保存按钮：点击 `保存` 按钮会将此实体的值存储到数据库中。
4. 分页器：用户可以通过分页器实现快速查看缺失记录的操作，也可以控制当前页面展示的缺失记录的数量。
5. 返回按钮：点击 `返回` 按钮可以返回至项目详情页面。
6. 结束迭代：用户可以点击 `结束迭代` 按钮结束当前迭代任务，进入到结果查看页面。应注意的是，若当前任务存在缺失记录，需要将缺失记录保存后才可以点击 `结束迭代` 按钮，否则按钮是置灰状态。
7. 迭代优化：若当前任务存在执行失败的数据，用户可以点击 `迭代优化` 按钮对执行失败的数据进行优化，点击后页面会跳转至数据准备页面，用户可以重新录入实体和规则执行最优记录任务。

执行失败的数据页面具体呈现如下图。

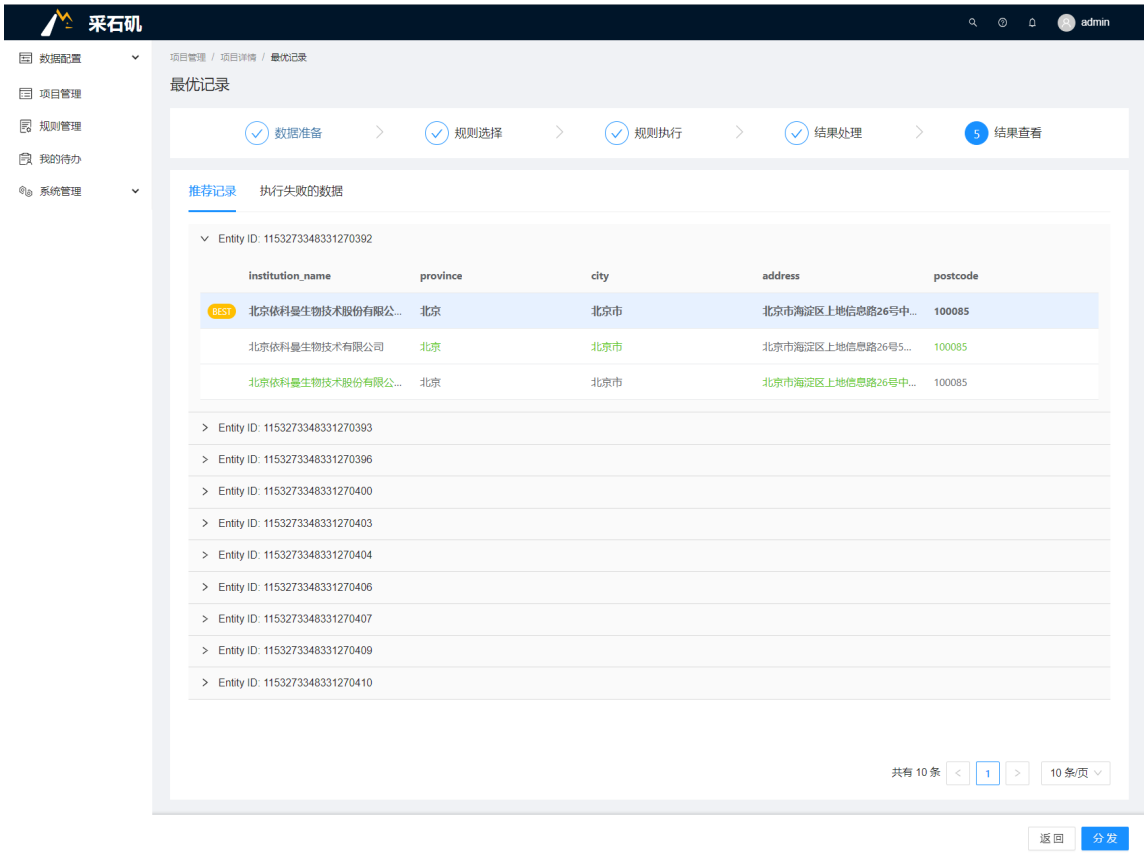


最优记录-执行失败的数据

1. 实体记录：用户可以点击实体ID查看实体记录。
2. 分页器：用户可以通过分页器实现快速查看执行失败的数据的操作，也可以控制当前页面展示的执行失败的数据的数量。
3. 返回按钮：点击 返回 按钮可以返回至项目详情页面。
4. 结束迭代：用户可以点击 结束迭代 按钮结束当前迭代任务，进入到结果查看页面。应注意的是，若当前任务存在缺失记录，需要将缺失记录保存后才可以点击 结束迭代 按钮，否则按钮是置灰状态。
5. 迭代优化：若当前任务存在执行失败的数据，用户可以点击 迭代优化 按钮对执行失败的数据进行优化，点击后页面会跳转至数据准备页面，用户可以重新录入实体和规则执行最优记录任务。

5. 结果查看

点击 结束迭代 按钮可以进入到结果查看页面，在结果查看页面可以看到系统生成的推荐记录和执行失败的数据，具体呈现如下图。

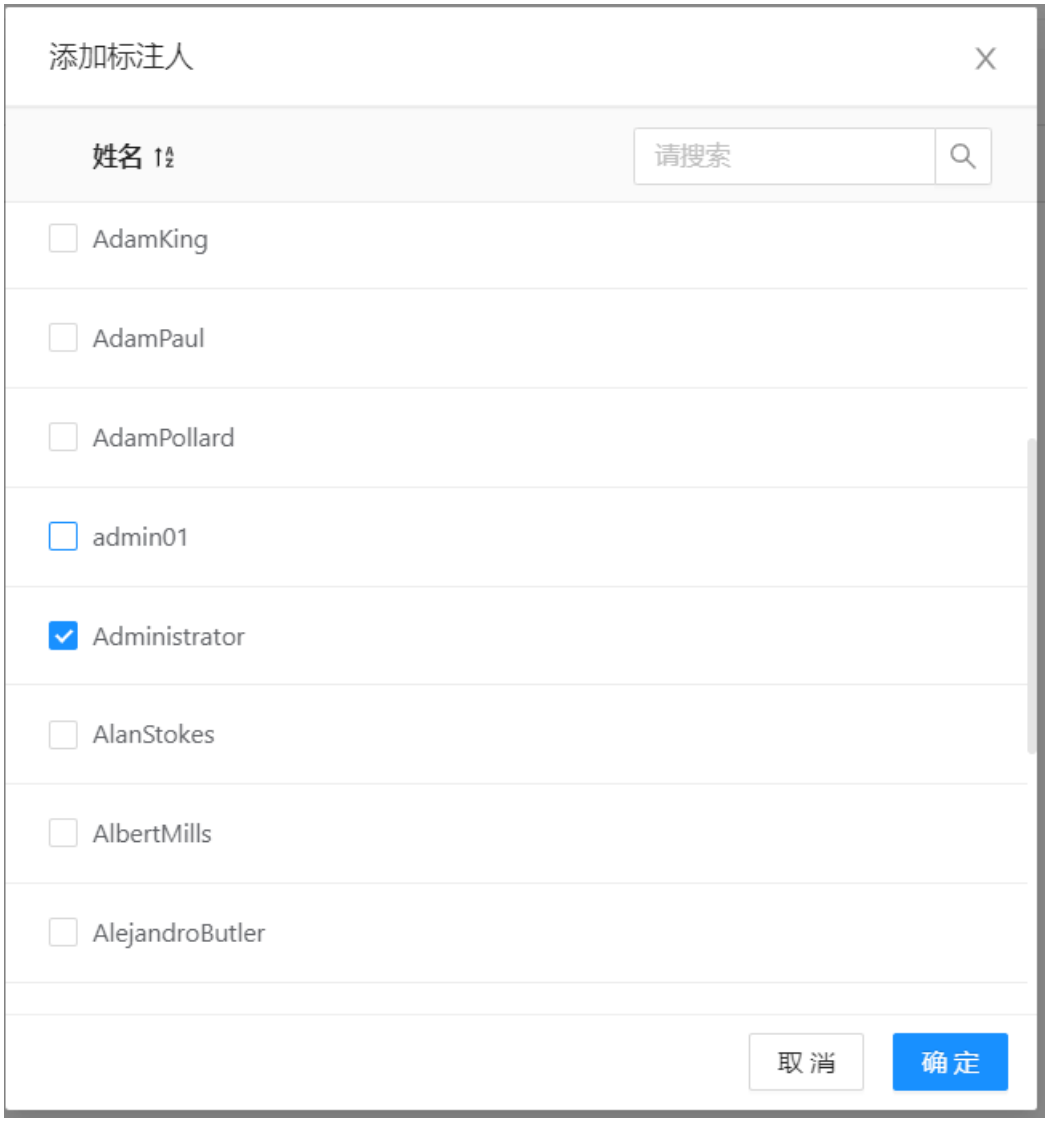


最优记录-结果查看

在此页面用户可以查看推荐记录和执行失败的数据，也可以点击 分发 按钮将推荐记录分发给操作员。

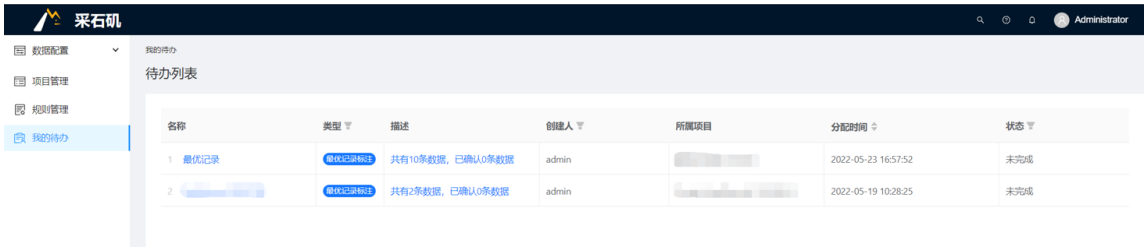
用户点击此页面的 分发 按钮，页面会弹出选择标注人窗口，用户可以将推荐记录分发给一个或者多个操作员进行确认，执行失败的数据不会分发给操作员去确认。

分发结果页面具体呈现如下图。



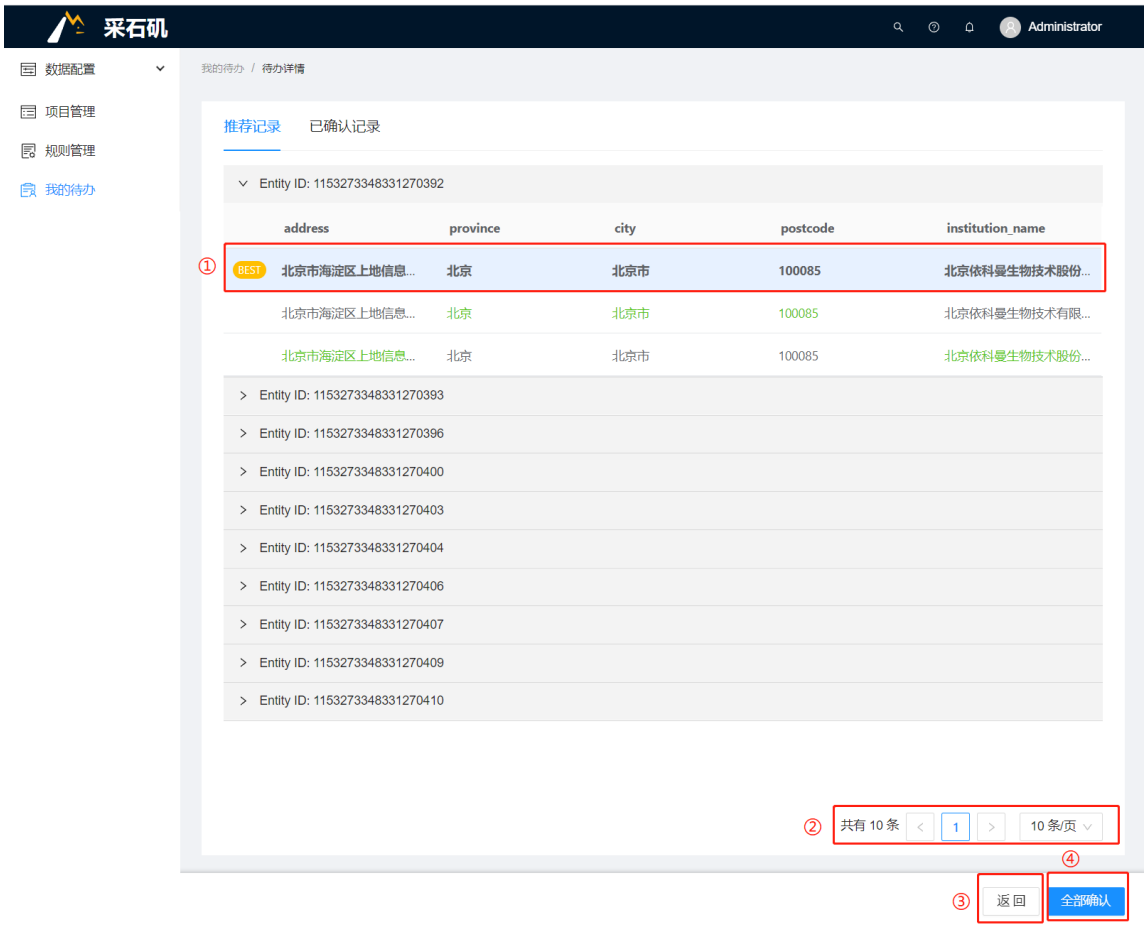
最优记录-分发结果

推荐记录分发给操作员后，使用操作员的账号登录采石矶系统，在我的待办页面可以查看到分发的任务，具体呈现如下图。



最优记录-操作员我的待办

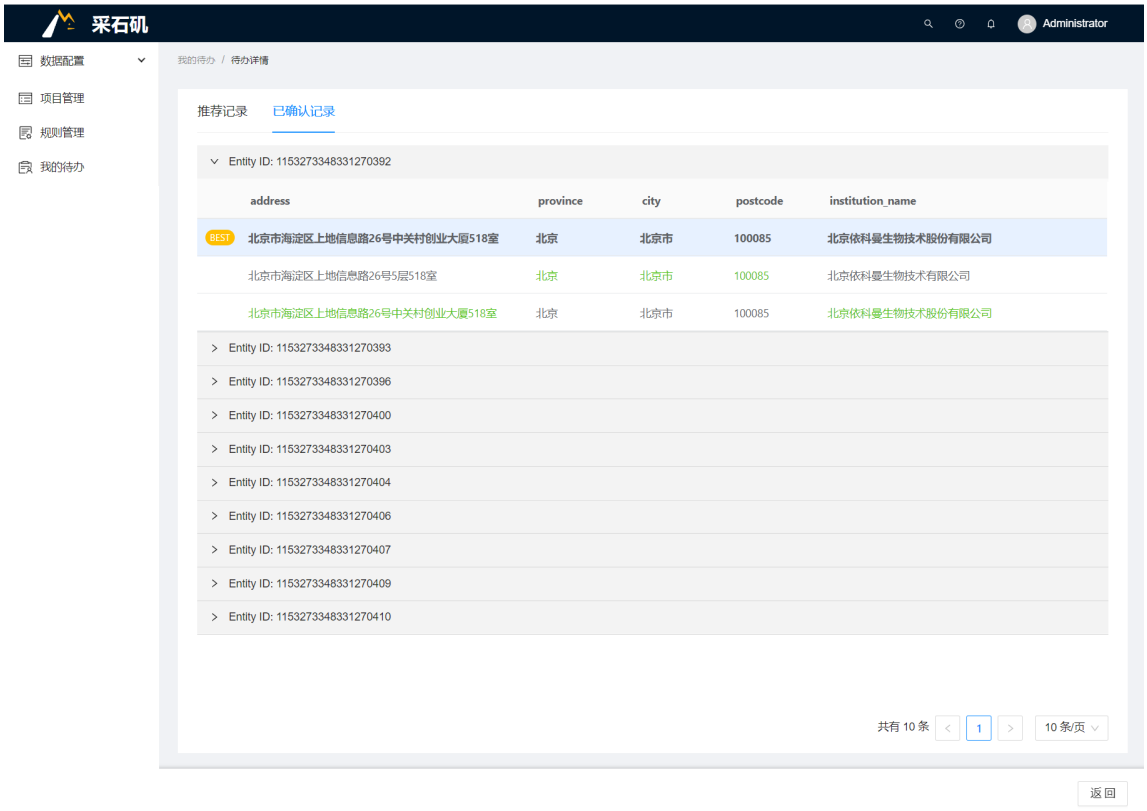
操作员进入到待办详情页，可以查看推荐记录，具体呈现如下图。



最优记录-操作员推荐记录

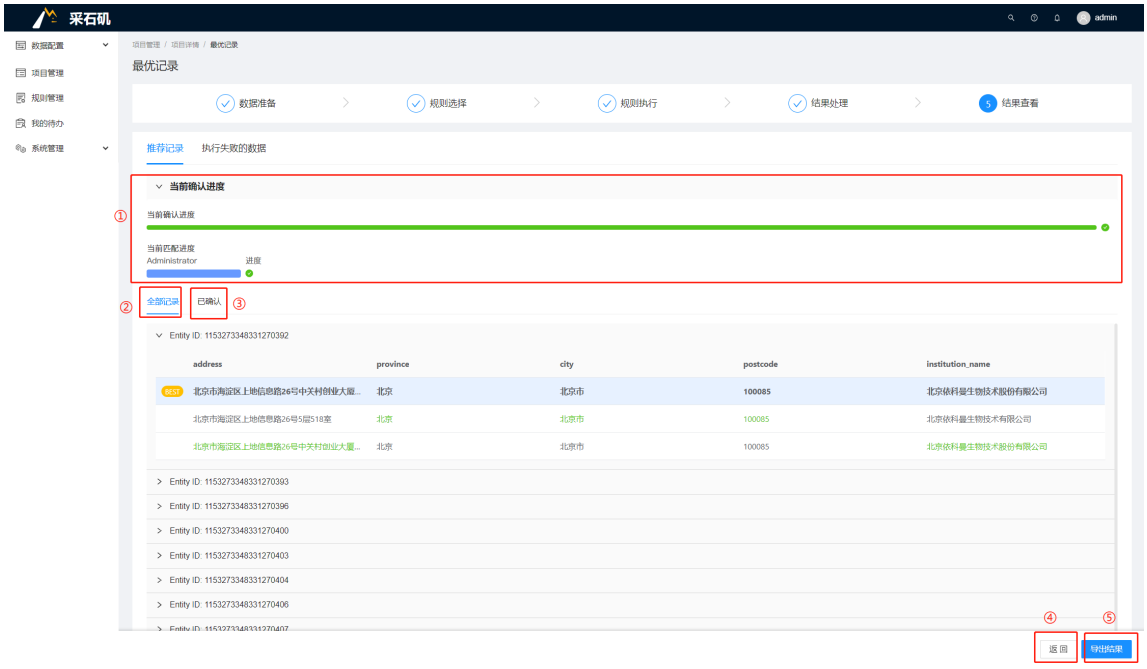
1. 最优值：若系统执行出的最优值不满足预期，用户可以点击最优值进行修改
2. 分页器：用户可以通过分页器实现快速查看推荐记录的操作，也可以控制当前页面展示的推荐记录的数量。
3. 返回按钮：点击 返回 按钮可以返回至我的待办页面。
4. 全部确认按钮：当用户想要确认当前页所有推荐记录时，可以点击 全部确认 。

确认后的数据会放到已确认数据中，具体呈现如下图。



最优记录-操作员已确认记录

操作员确认过数据后，管理员可以查看当前数据的确认情况，具体呈现如下图。



最优记录-结果查看（分发后）

1. 当前确认进度：已确认的数据在所有待确认的数据中的占比。

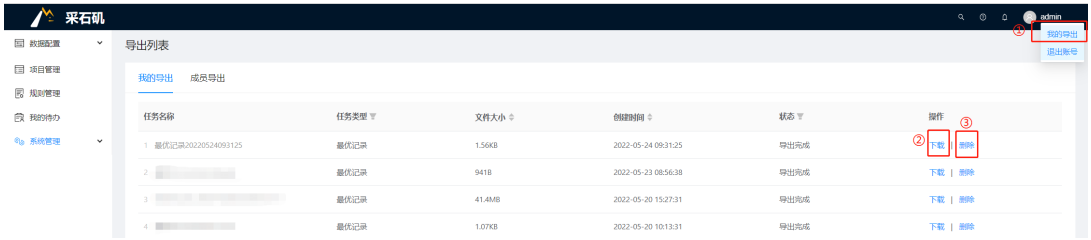
2. 全部记录：用户可以点击全部记录查看所有推荐记录。

3. 已确认：用户可以点击已确认查看已确认的记录。

4. 返回：点击 返回 按钮可以返回至项目详情页面。

5. 导出结果：点击 导出结果 按钮可以将当前任务的最优值导出。

   导出后可以在我的导出中看到导出结果，具体呈现如下图。



最优记录-我的导出

   i. 我的导出：用户点击 我的导出 可以查看导出结果列表。

   ii. 下载：用户点击 下载 按钮可以将本次导出结果下载，下载的文件为.csv文件。

   iii. 删除：用户点击 下载 按钮可以删除本次导出的记录。

字段匹配

本章节主要介绍采石矶系统的字段匹配功能。

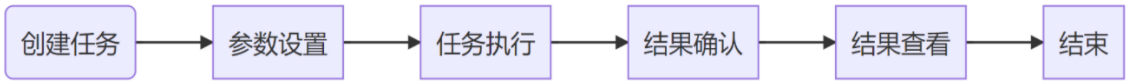
本系统支持字段匹配功能，输入一个主数据表选择需做匹配的字段，再选择一个或多个对比表；通过对字段数据进行算法分析后判断字段之间是否存在匹配关系。

通过完成本章节步骤，可以了解字段匹配功能及字段匹配任务的操作方法。

前置条件

- 用户已登录；
- 已添加数据源和数据集；
- 已创建项目。

字段匹配操作流程图

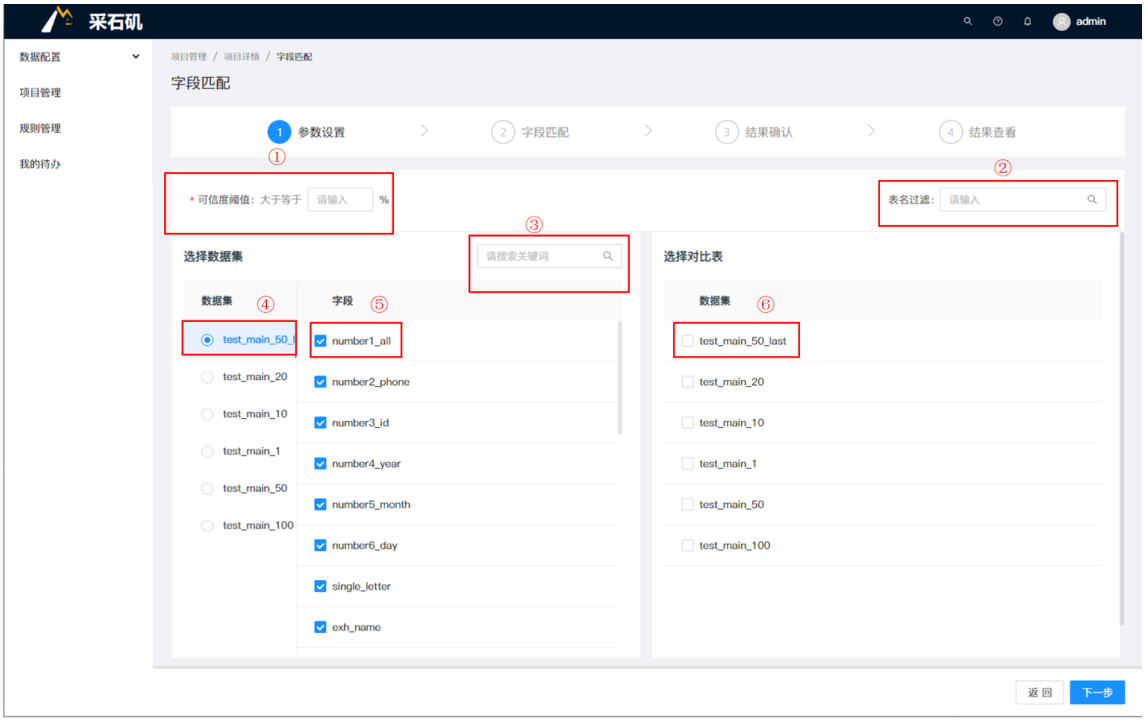


字段匹配流程图

在新建工作流页面拖入数据集组件，选择数据集后，拖入字段匹配任务组件并连线。保存和启动工作流后，会进入查看工作流页面。在查看工作流页面点击字段匹配任务组件，可以配置任务信息。

字段匹配任务共有四个阶段，分别是参数设置、字段匹配、结果确认和结果查看，具体操作介绍如下：

在查看工作流页面点击字段匹配任务组件，页面窗口会跳转到参数设置页面。



①：可信度阈值设置，即匹配度大于等于配置值会出现在结果中，配置范围0-100。
②：选择数据表时，可通过表名模糊匹配过滤出对应的表，输入字符串点击放大镜图标生效。
③：选择主表字段名时，可通过关键词匹配过滤出对应的字段名，输入字符串点击放大镜图标生效。
④：勾选主表，鼠标点击圆点选中，主表只能选择一个。
⑤：勾选字段名，鼠标点击方框选中，字段可选择一个或多个。
⑥：勾选对比表，鼠标点击放开选中，对比表可选择一个或多个。

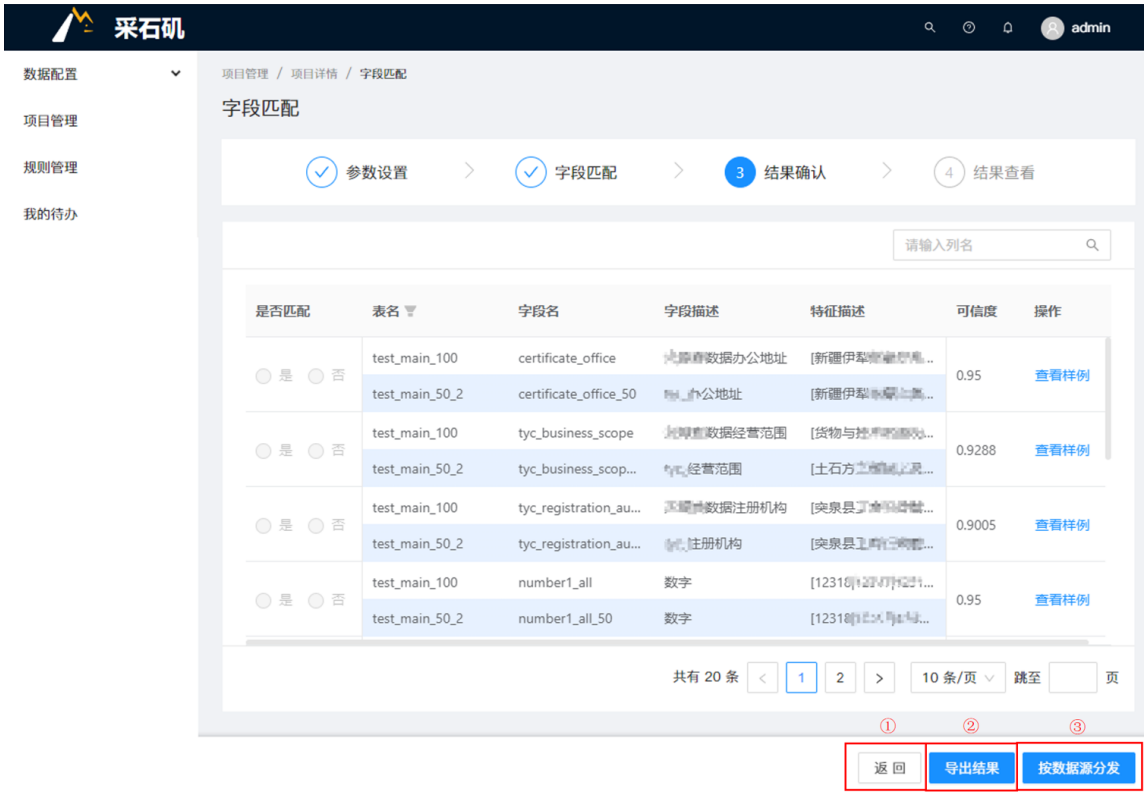
参数配置

参数设置详细说明

选项	配置说明	必要
可信度阈值	配置任务可信度阈值	是
选择数据集	需勾选数据集和字段名	是
选择对比表	需勾选对比表的数据集	是

参数设置完成后，点击 下一步 ，字段匹配任务开始执行。

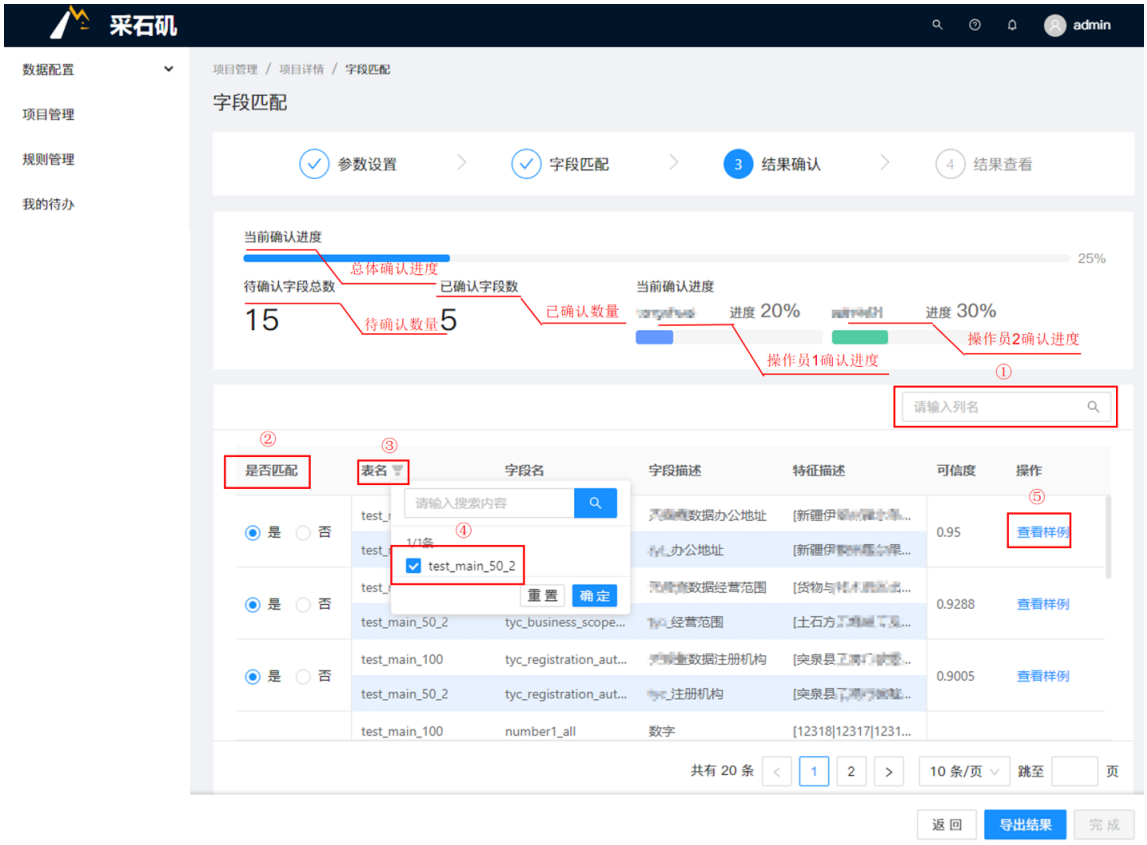
字段匹配任务完成后，自动跳转到结果查看页面。



①：返回按钮可返回项目详情页面；
②：导出按钮可导出字段匹配任务结果；
③：按数据分发按钮，可以把数据分发给数据源负责人进行结果确认。

任务结果

点击 按数据源分发 按钮后，进入到数据结果的分发确认页面，系统会将数据分发给数据源的责任人进行确认，管理员用户也可对数据进行确认操作，此时管理员页面显示数据确认进度和结果。

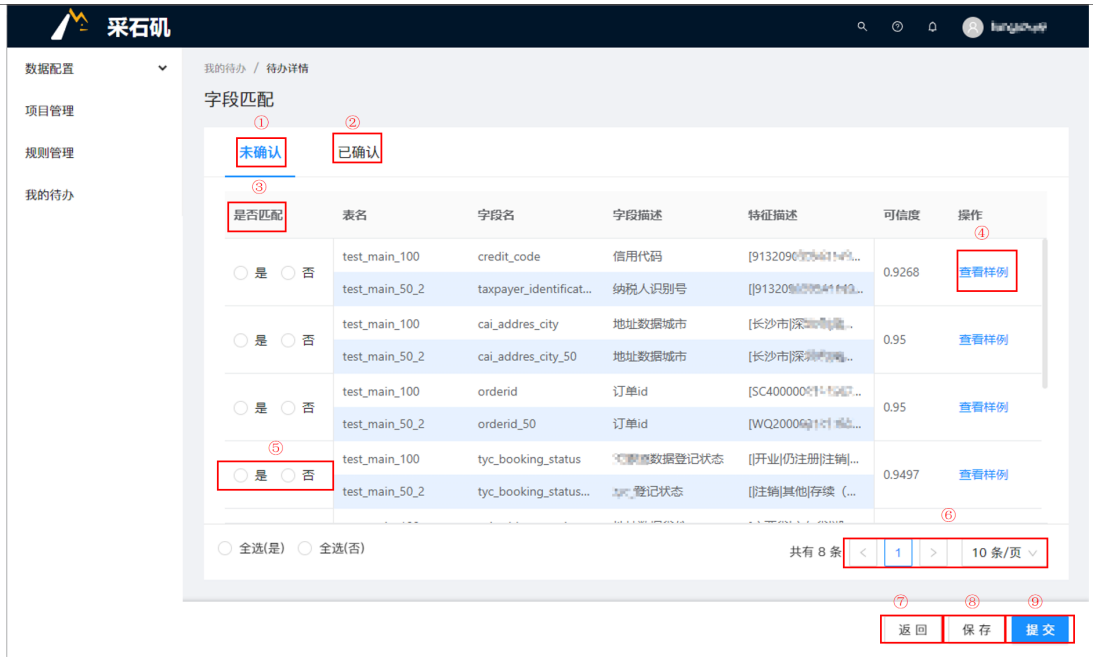


①：可输入字段名模糊匹配进行结果的过滤；
②：标注匹配结果，“是”或者“否”；
③：可通过表名进行结果过滤；
④：勾选表名过滤对应表名的结果；
⑤：可点击“查看样例”，查看样例数据。

任务结果确认中

用操作员账号进行登录，登录后点击 我的待办 菜单栏中可看到对应任务，选择对应的任务名进入到标注页面。

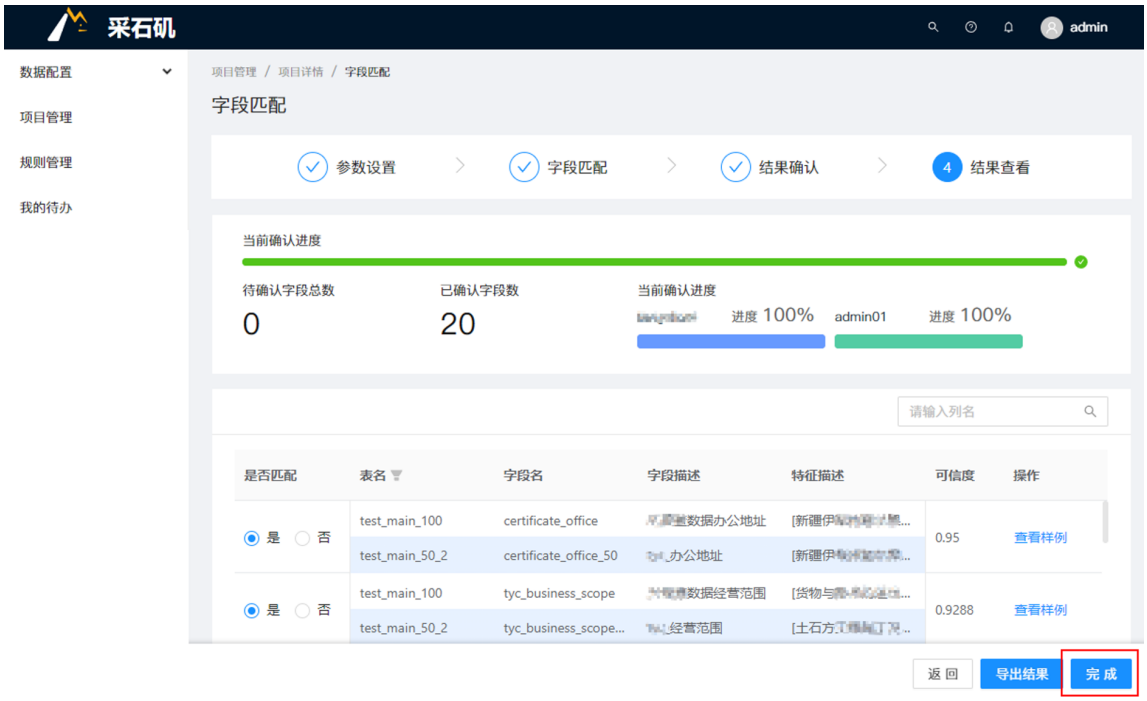
此时操作员页面可看到待办任务中存在对应数据，可对结果进行标记、保存或者提交操作。



①：操作员标注默认进入待确认页面；
②：点击‘已确认’可以切换到已确认的页面，此页面数据内容为确认提交的数据，已确认数据无法修改匹配结果；
③：对结果进行确认，标注‘是’或者‘否’；
④：数据确认时可点击‘查看样例’查看样例数据，辅助操作员对数据进行判断；
⑤：全选按钮对于批量确认的数据可进行批量标注；
⑥：对数据进行分页，可进行分页查看；
⑦：返回按钮，点击‘返回’按钮返回我的待办页面；
⑧：保存按钮，点击‘保存’按钮对当前确认的结果进行本地保存；
⑨：提交按钮，点击‘提交’按钮对当前确认数据进行提交，数据进入已确认页面中，待所有数据提交完成，标注任务结束。

操作员结果确认界面

操作员标注完成，总体进度达到100%，可点击 完成 按钮完成整个字段匹配任务。



结果确认完成

此时字段匹配任务完成，任务状态为已完成，点击对应任务名称可进行结果查看和导出，导出的结果会带上操作员标注信息。

任务完成后状态

规则管理

本章节主要介绍采石矶系统规则管理的具体操作流程。

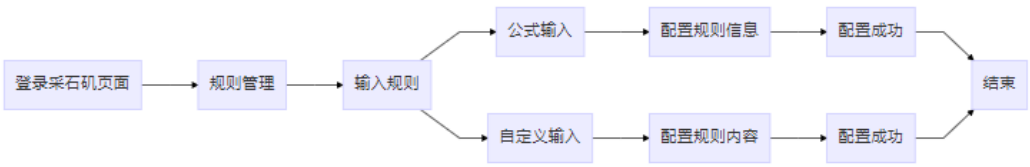
本系统定义的规则管理包括四种规则，分别是CR规则、ER规则、最优规则和正则规则。其中，CR规则是用于处理数据冲突错误问题的规则，ER规则是用于处理数据实体识别问题的规则，最优规则是用于在已有的实体记录中找到每个实体的最优值的规则，正则规则是用于通过正则表达式去查找不符合该表达式的冲突数据的规则。

需要特别说明的是，通过本系统的规则发现功能对数据进行分析，能够自动得出CR、ER规则和正则规则，其中正则规则是包含在CR规则发现中的。

前置条件

1. 用户已登录；
2. 已导入数据源，已添加数据集；
3. 已创建实体。

规则管理操作流程如下图所示。

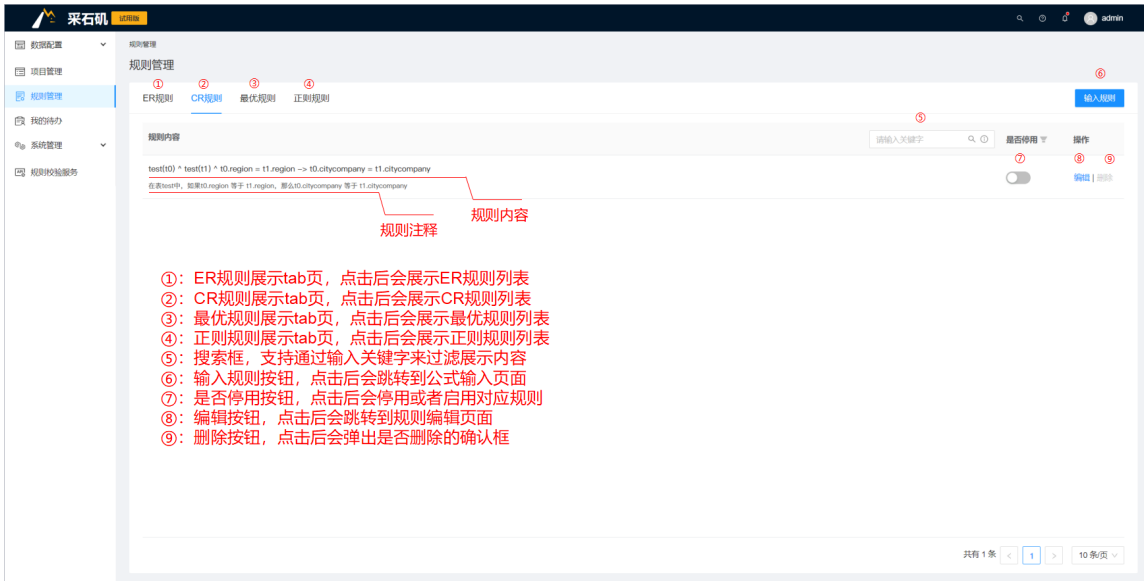


规则管理操作流程图

规则管理页面说明

本章主要介绍规则管理页面。

点击 `规则管理` 按钮，默认展示ER规则管理页面，具体呈现如下图。



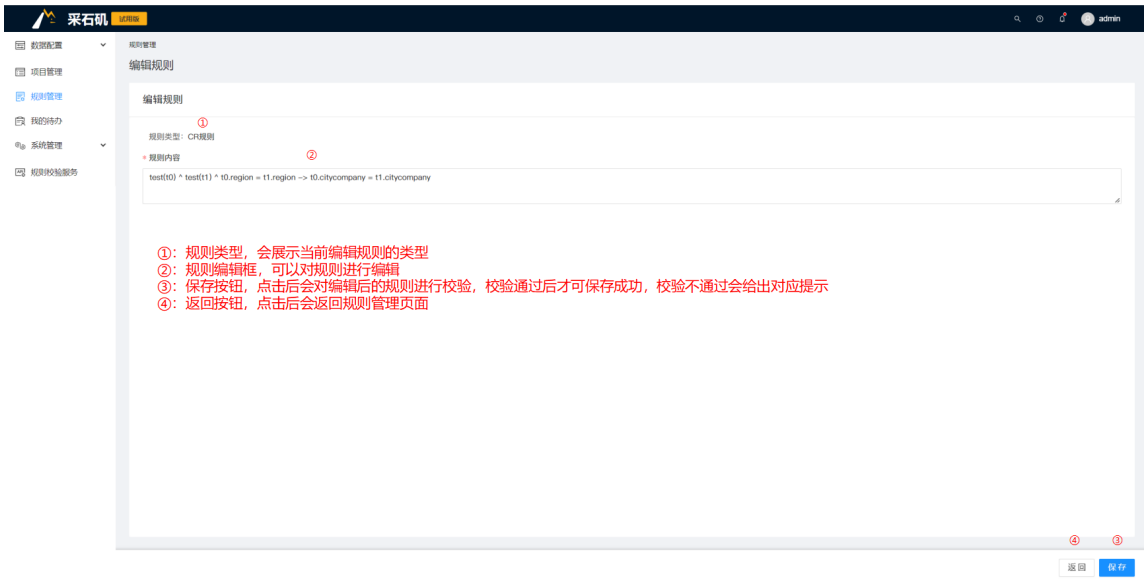
规则管理界面

- 停用和启用

点击规则右侧的 `是否停用` 按钮，页面窗口会返回状态更改成功的响应信息，对应规则不可用，能够对规则进行 `删除` 操作；再次点击 `是否停用` 按钮，页面窗口同样会返回状态更改成功的响应信息，对应规则恢复可用，对应 `删除` 按钮不可用。

- 编辑

点击 `编辑` 按钮，会跳转到编辑规则页面，可以对规则内容进行编辑，编辑完成后点击 `保存` 按钮，系统会对编辑后的规则进行校验，校验通过后才可保存，校验不通过会做出相应提示。最优规则不支持编辑。



输入规则界面

- 删除

点击 删除 按钮，会弹出"是否删除该规则"的确认框。点击 确定 ，页面窗口会返回删除成功的响应信息。

- 输入规则

点击 输入规则 按钮，会展示公式输入规则页面，页面具体操作后续有详细介绍。

点击 自定义输入 按钮，会展示自定义输入规则页面，页面具体操作后续有详细介绍。

公式输入操作说明

本章介绍公式输入操作说明，包括“CR规则操作说明”、“ER规则操作说明”、“最优规则操作说明”、“正则规则操作说明”。

1. CR规则操作说明

- 创建CR规则

在规则管理页面中点击 输入规则 按钮，默认会进入CR规则的公式输入页面。

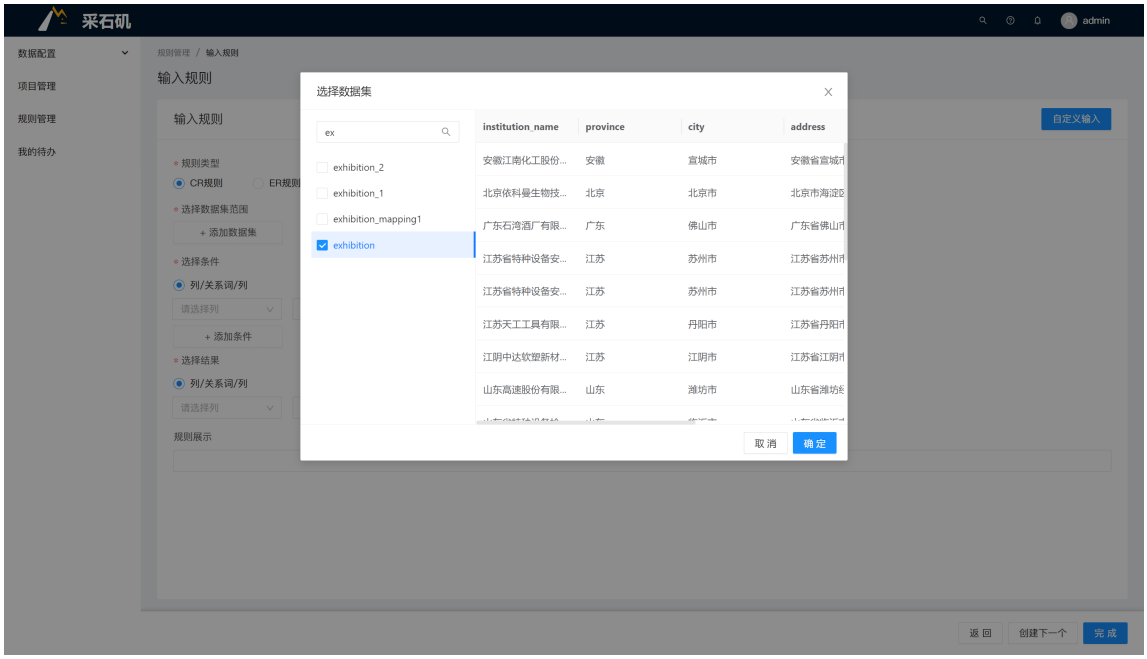
CR规则公式输入操作介绍

点击 添加数据集 ，会弹出选择数据集的窗口。

可以通过拖动滚动条查找数据集，也可以通过搜索框输入数据集名称查找，支持模糊查找。

点击数据集名称，可以预览数据。

点击数据集名称左侧的复选框，选中数据集，支持选中多个数据集。



选择数据集界面

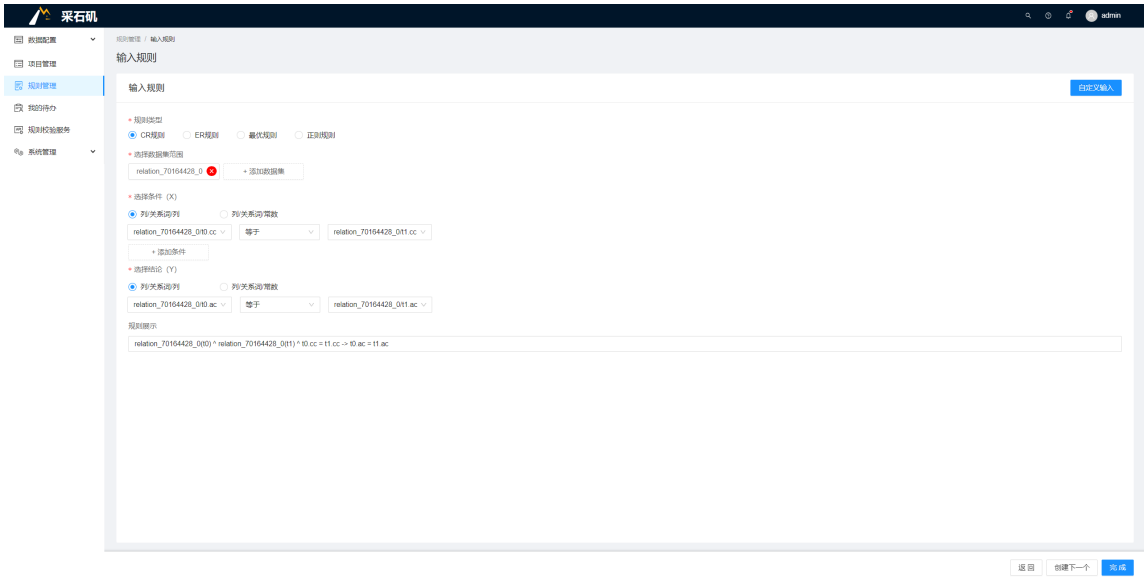
点击 `确定` 按钮，添加数据集成功，并返回输入规则界面。

选择条件

- 选择 `列/关系词/列` 时，表示两个列的关系，两个列可以是同一张表的两个列，也可以是跨表之间的两个列；
- 选择 `列/关系词/常数` 时，表示列和常数之间的关系。

选择结果

- 选择 `列/关系词/列` 时，同 `选择条件` ，表示两个列的关系，两个列可以是同一张表的两个列，也可以是跨表之间的两个列；
- 选择 `列/关系词/常数` 时，同 `选择条件` ，表示列和常数之间的关系。



CR规则公式输入界面

点击 完成 按钮，创建CR规则成功。

备注：

采石矶系统支持如下的关系词：

数据类型	关系词	举例
数值型数据	等于	t0.age = t1.age
	不等于	t0.age != t1.age
	大于	t0.age > t1.age
	大于等于	t0.age >= t1.age
	小于	t0.age < t1.age
	小于等于	t0.age <= t1.age
字符串型数据	相似于	similar('jaccard', t0.city, t1.city, 0.85)
	等于	t0.city = t1.city
	不等于	t0.city != t1.city

从上方表格可以看到，"相似于"关系词和其他关系词有不同之处，"相似于"关系词用到了相似度算法和机器学习模型。 目前采石矶系统支持如下的相似度算法：

算法	举例	说明
cosine	similar('cosine', t0.city, t1.city, 0.85)	文本相似度算法的一种，使用向量空间中两个向量夹角的余弦值作为衡量个体间差异的大小的度量，用于计算两段文本相似的程度。速度较快，准确度较差，对于中英文效果类似
jaccard	similar('jaccard', t0.city, t1.city, 0.85)	文本相似度算法的一种，使用样本交并集比值衡量样本之间的相似性与差异性，用于计算两段中长文本相似的程度。速度最快，召回率一般。对于短文本效果较差，长文本(>100)速度较快。内置中文分词工具，英文效果极差，中文效果较好
jaro-winkler	similar('jaro-winkler', t0.city, t1.city, 0.85)	文本相似度算法的一种，对相等字符的距离进行过滤作为度量计算相似度，用于计算两段短文本相似的程度。速度一般，召回率较高。对于短文本效果较好，长文本速度较慢。中英文效果都较好
levenshtein	similar('levenshtein', t0.city, t1.city, 0.85)	文本相似度算法的一种，使用两段文本转换所需的编辑操作次数作为度量计算相似度，用于计算两段短文本相似的程度。速度较慢，召回率极高。对于短文本效果较好，长文本速度较慢。中英文效果都较好

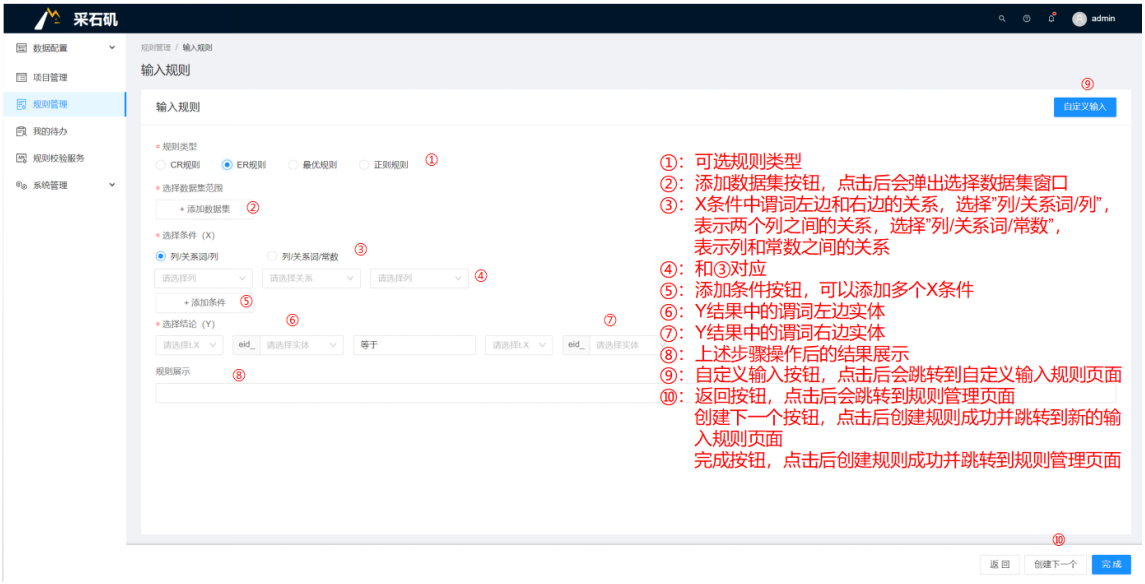
支持如下的机器学习模型：

模型	举例	说明
model-match-address_d	ml('model-match-address_d', t0.city, t1.city)	地址匹配的机器学习模型1，用于判断两个地址是不是描述了同个地理位置
model-match-company_name	ml('model-match-company_name', t0.city, t1.city)	公司机构名称的机器学习模型，用于判断两个公司名称叫法是不是同一个实体
model-match-job	ml('model-match-job', t0.address, t1.address)	职位匹配的机器学习模型，用于判断招聘职位或其它职位信息是不是描述了相同岗位
model-match-address_n	ml('model-match-address_n', t0.address, t1.address)	地址匹配的机器学习模型2，用于判断两个地址是不是描述了同个地理位置

2. ER规则操作介绍

- 创建ER规则

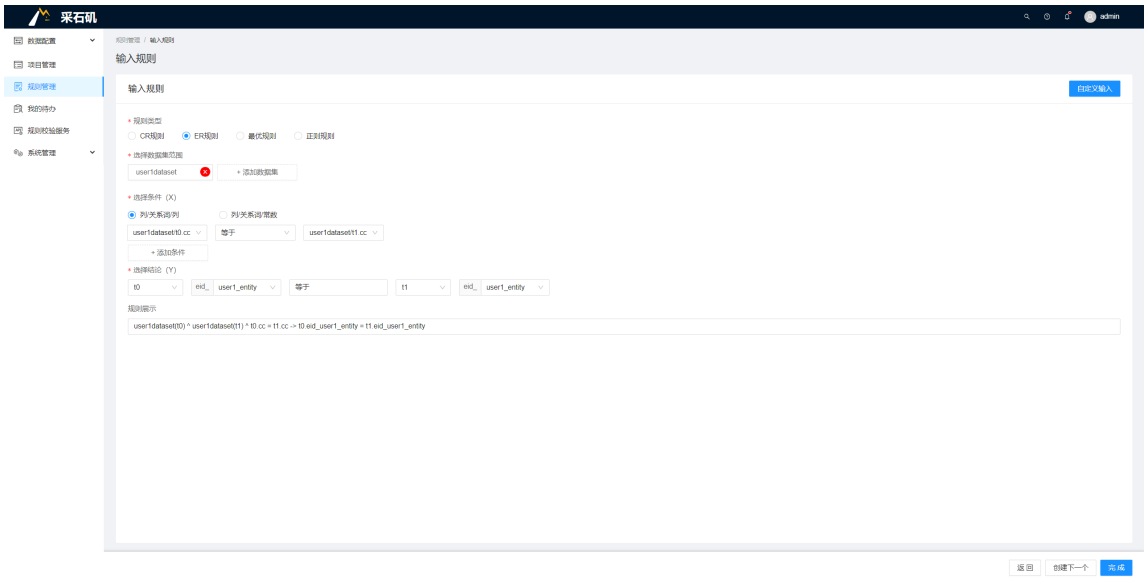
在规则管理页面中点击 输入规则 按钮，规则类型中选择 ER规则 ，会进入ER规则的公式输入页面。



ER规则公式输入操作介绍

添加数据集 和 选择条件 ，在CR规则操作介绍一章节中已作描述，此处不再赘述。

选择结果 中，选择两行，认为是同一个实体，这两行可以是同一张表的两行，也可以是跨表之间的两行。



ER规则公式输入界面

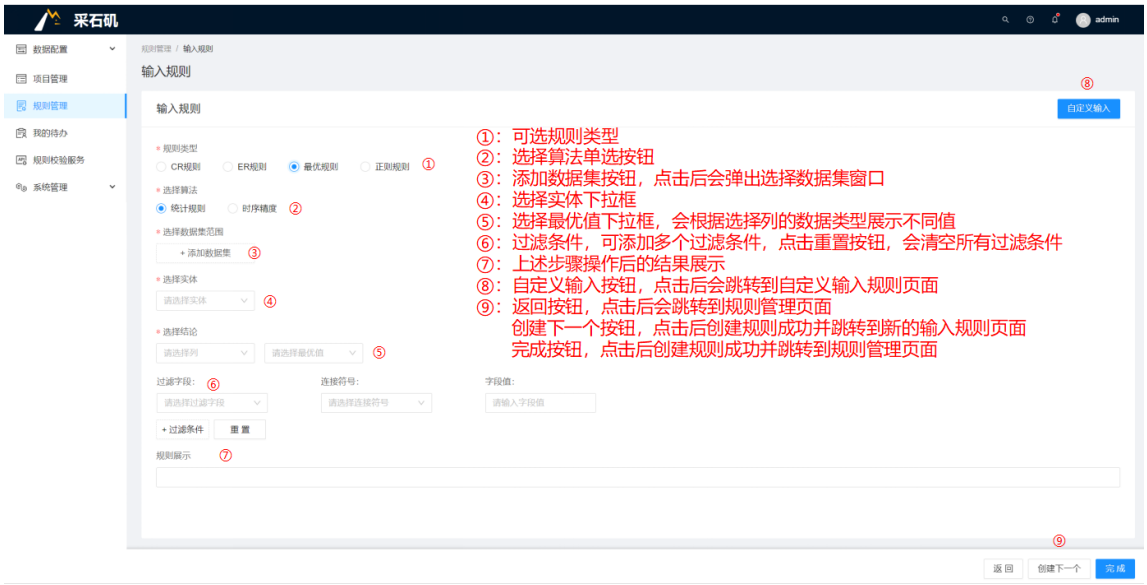
点击 完成 按钮，创建ER规则成功。

3. 最优规则操作介绍

按照算法划分，最优规则分为统计规则和时序精度两种，下面将分别介绍两种规则是如何创建的。

- 创建统计规则

在规则管理页面中点击 输入规则 按钮，规则类型中选择 最优规则 ，会进入统计规则的公式输入页面。



统计规则公式输入操作介绍

添加数据集 ，在CR规则操作介绍一章节中已作描述，此处不再赘述。

选择实体 ，点击下拉框，会展示数据集已有实体。

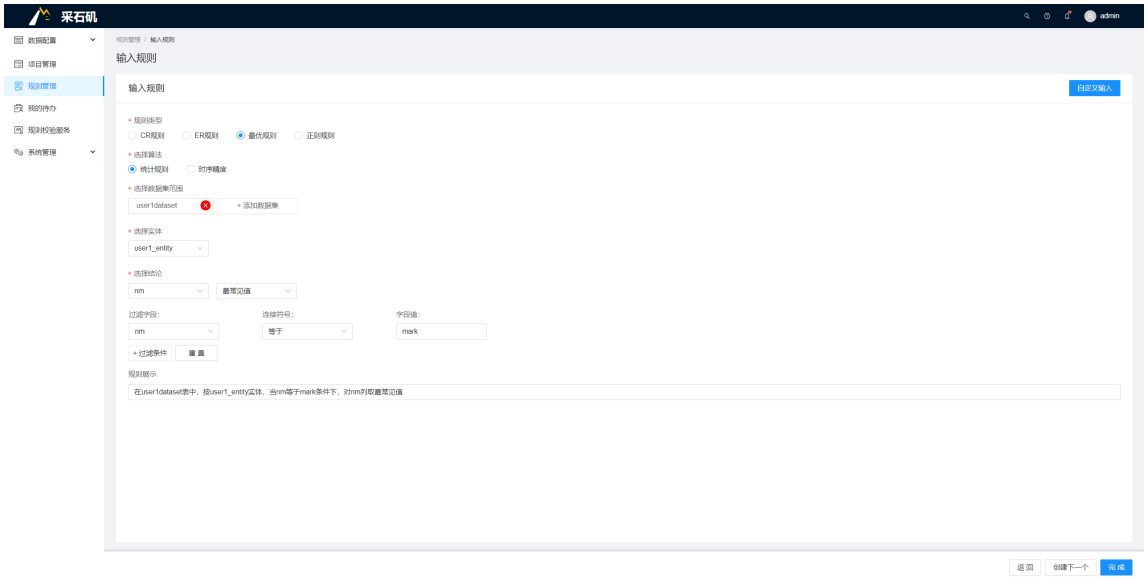
选择结论 中，点击选择列下拉框，会展示选中实体对应的列。

选择结论 中，点击选择最优值下拉框，会展示选中列对应的最优值类型。

过滤条件 中的 过滤字段 ，点击下拉框，会展示选中实体对应的列。

过滤条件 中的 连接符号 ，包括了等于、不等于、是空值、非空、大于等于、大于、小于等于、小于、包含、不包含，支持添加多个过滤条件。

过滤条件 中的 重置 ，会清空所有过滤条件。

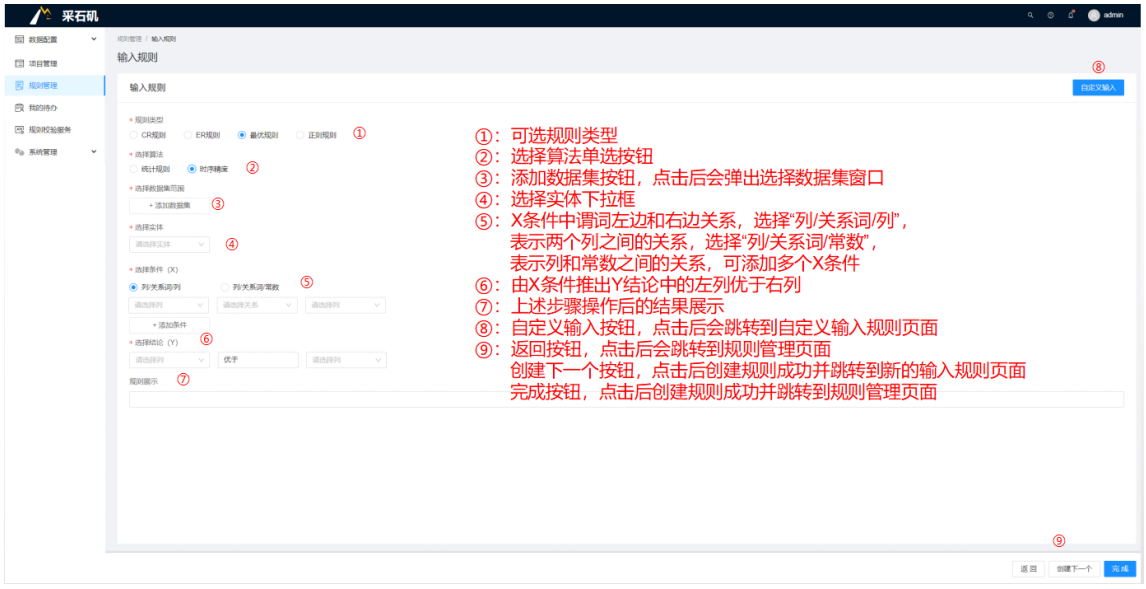


统计规则公式输入界面

点击 完成 按钮，创建统计规则成功。

- 创建时序精度规则

在规则管理页面中点击 输入规则 按钮，规则类型中选择 最优规则 ，选择算法中选择 时序精度 ，会进入时序精度规则的公式输入页面。



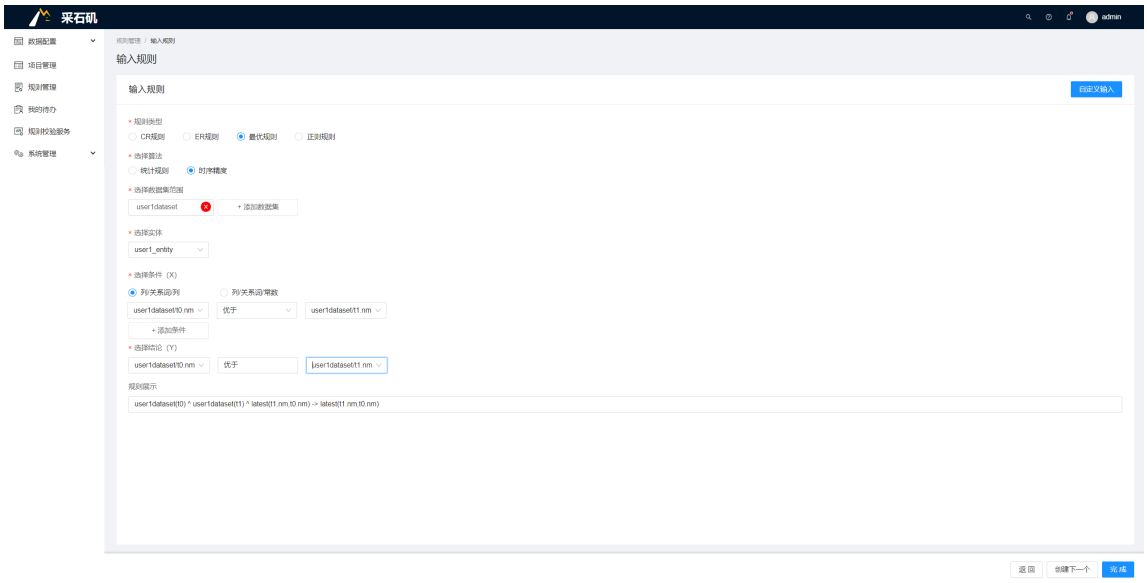
时序精度规则公式输入操作介绍

添加数据集 ，在CR规则操作介绍一章节中已作描述，此处不再赘述。

选择实体 ，点击下拉框，会展示数据集已有实体。

选择条件 ，在CR规则操作介绍一章节中已作描述，此处不再赘述。

选择结论 中，点击下拉框分别选择Y结论中的左列和右列。



时序精度规则公式输入界面

点击 完成 按钮，创建时序精度规则成功。

备注：

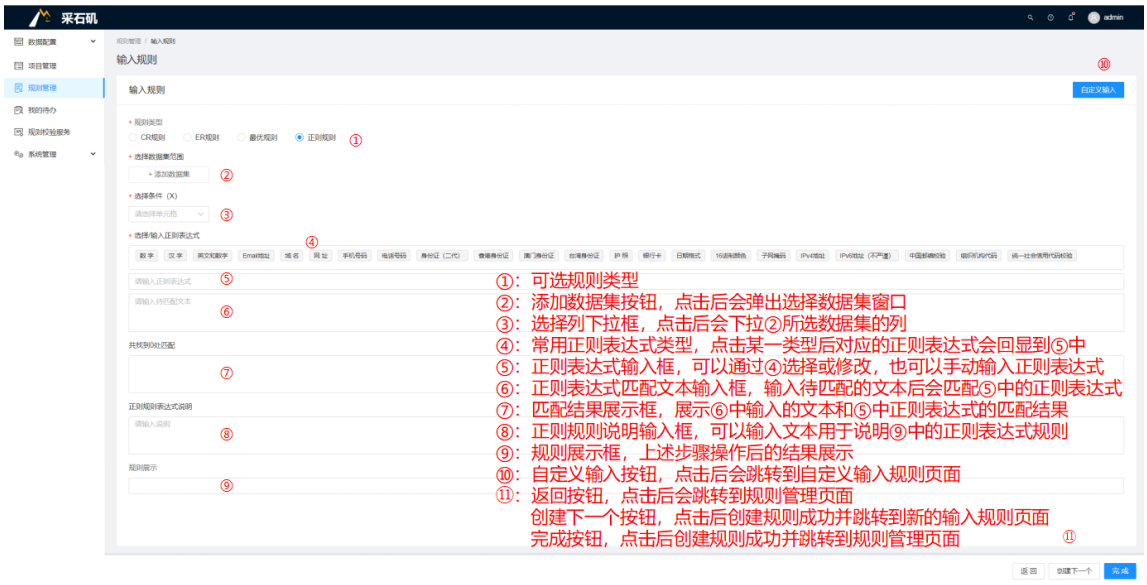
采石矶系统支持如下的最优值类型：

数据类型	最优值类型
数值型	最大值(Max)
	最小值(Min)
	平均值(Mean)
	最常见值(Most common value)
	总和(Sum)
	计数(Count)
	不同值计数(Count Distinct)
字符串型	最常见值(Most common value)
	最长值(Longest)
	最短值(Shortest)

4. 正则规则操作介绍

- 创建正则规则

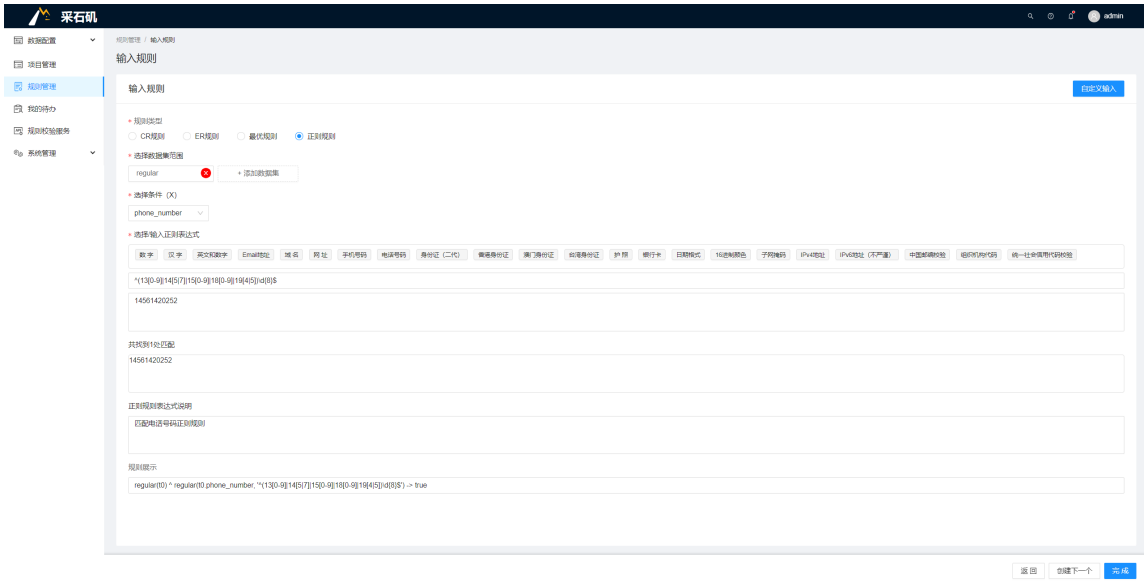
在规则管理页面中点击 输入规则 按钮，规则类型中选择 正则规则 ，会进入正则规则的公式输入页面。



正则规则公式输入操作介绍

添加数据集 ，在CR规则操作介绍一章节中已作描述，此处不再赘述，正则数据集选择和CR数据集选择不同的一点是正则只支持单个数据集。 选择条件 ，下拉选择数据集的列，只能选择单列。

选择/输入正则表达式 ，此处的正则表达式有两种方式输入，一种是选择系统提供的常用正则表达式方式，该方式不需要用户去构造正则，只需要点击所需要的类型即可生成对应的正则表达式并且可以在此基础上修改，第二种就是可以手动输入自定义正则，如果第一种方式不能满足用户需求也可以用第二种方式构造任意的正则。此外还提供正则表达式的验证功能用于验证表达式的正确性。



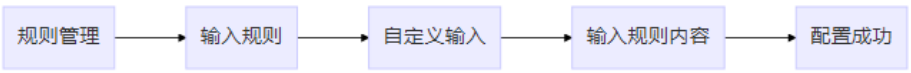
正则规则公式输入界面

点击 完成 按钮，创建正则规则成功。

自定义输入操作说明

本章介绍自定义输入操作说明，包括“CR规则操作说明”、“ER规则操作说明”。

自定义输入操作流程如下图所示。

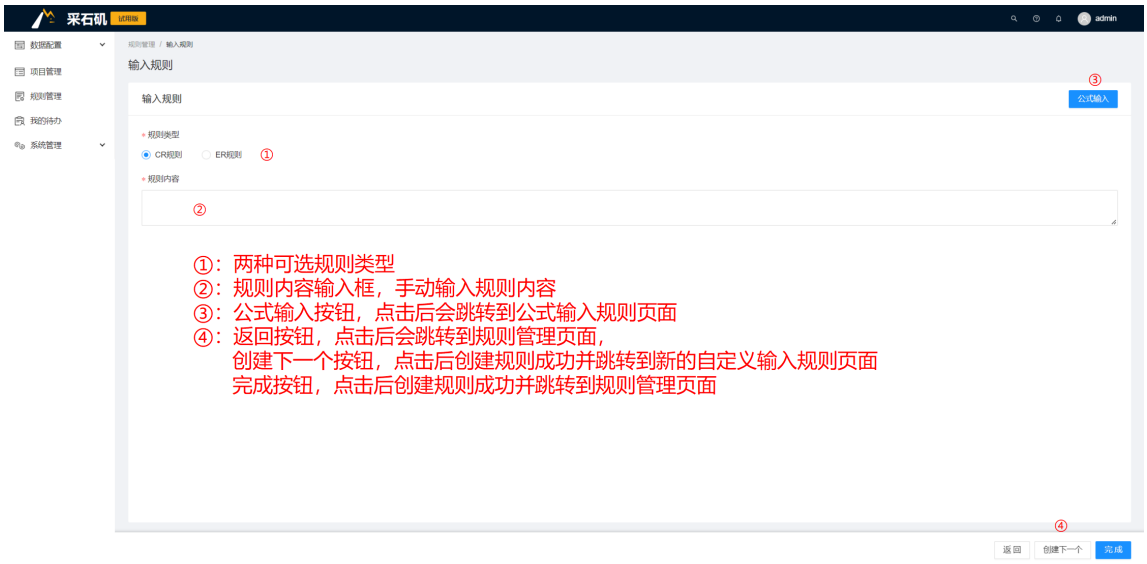


自定义输入操作流程图

1. CR规则操作介绍

- 创建CR规则

点击 输入规则 按钮，点击 自定义输入 按钮，默认会进入CR规则的自定义输入页面。



CR规则自定义输入界面

手动输入规则内容，比如：

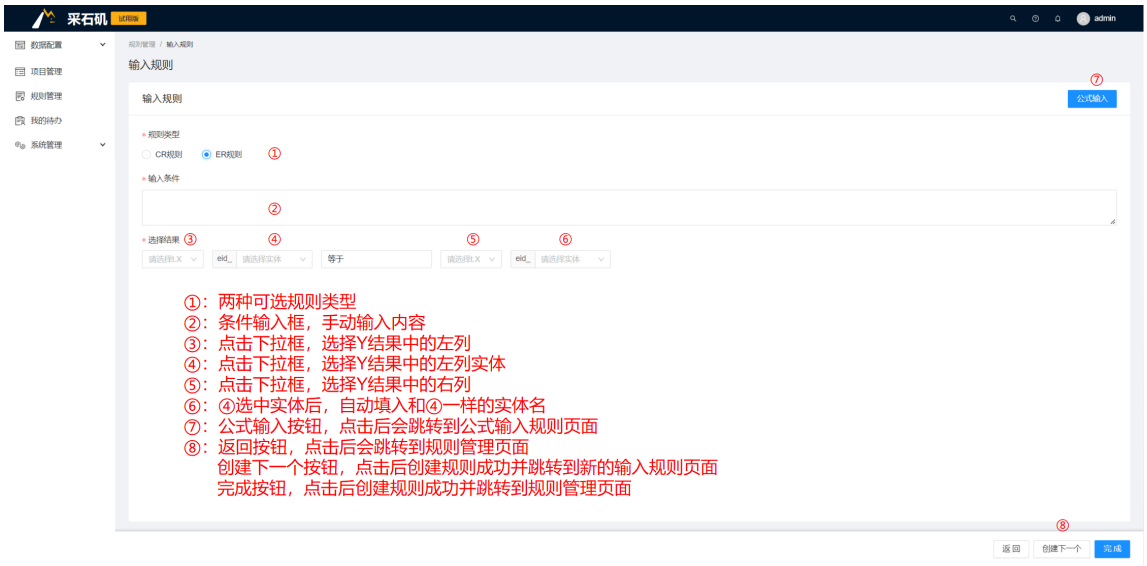
exhibition(t0) ^ exhibition_mapping(t1) ^ t0.city = t1.city -> t0.institution_name = t1.institution_name

上述规则表达的意思是在表exhibition上取一行标记为t0，在表exhibition_mapping上取一行标记为t1，并且满足 t0.city = t1.city，推出t0.institution_name = t1.institution_name。

2. ER规则操作介绍

- 创建ER规则

点击 `输入规则` 按钮，点击 `自定义输入` 按钮，选择 `ER规则` ，会进入ER规则的自定义输入页面。



ER规则自定义输入界面

手动输入规则内容，比如：

exhibition(t0) ^ exhibition(t1) ^ t0.city = t1.city -> t0.eid_exhibition_name = t1.eid_exhibition_name

上述规则表达的意思是在表exhibition上取一行标记为t0，再取一行标记为t1，并且满足 t0.city = t1.city，则推出这两行是同一实体。

用户管理

本章节主要介绍采石矶系统用户管理在系统中的功能作用及操作流程。

采石矶系统存在多用户和多角色，不同角色的用户权限不同，用户管理可以对采石矶用户进行角色授权和查看用户信息，目前系统只有3个角色：管理员、专家、管理员，其中只有admin用户是管理员角色所以也仅有admin有用户管理的权限可以进行授权。

新用户默认无任何权限，不能登录系统，要授权后才可登录系统。

前置条件

- LDAP已安装并配置好。
- 用户已导入LDAP。

用户管理操作说明

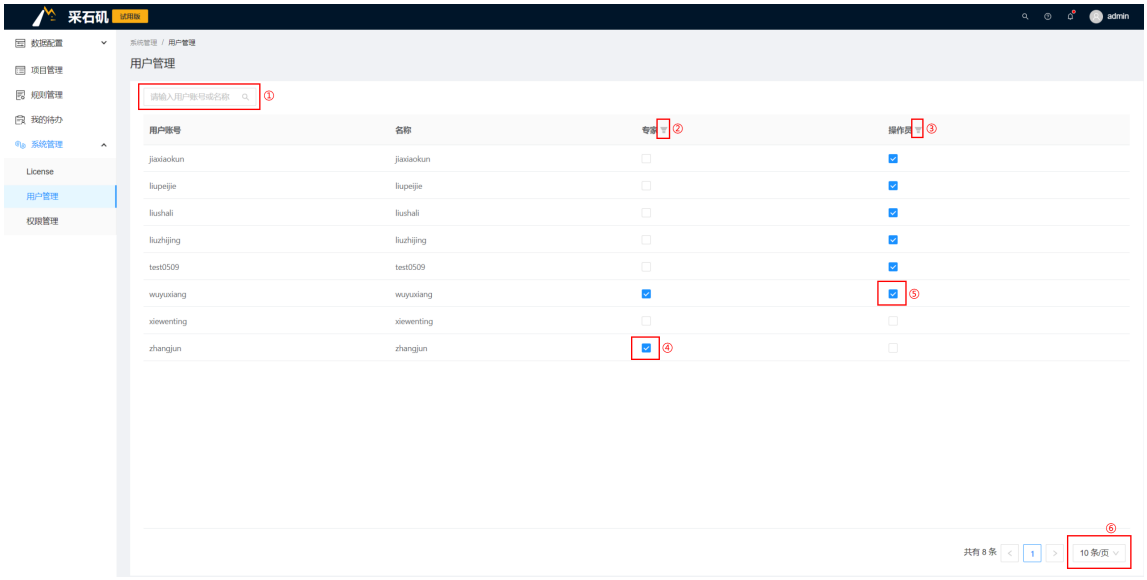
本章节主要讲解用户管理操作说明，包括用户管理的流程和操作。

用户管理操作流程如下图所示。

用户管理流程图

1. 用户管理列表页面说明

点击 `用户管理` 菜单，会看到用户管理页面，具体呈现如下图。



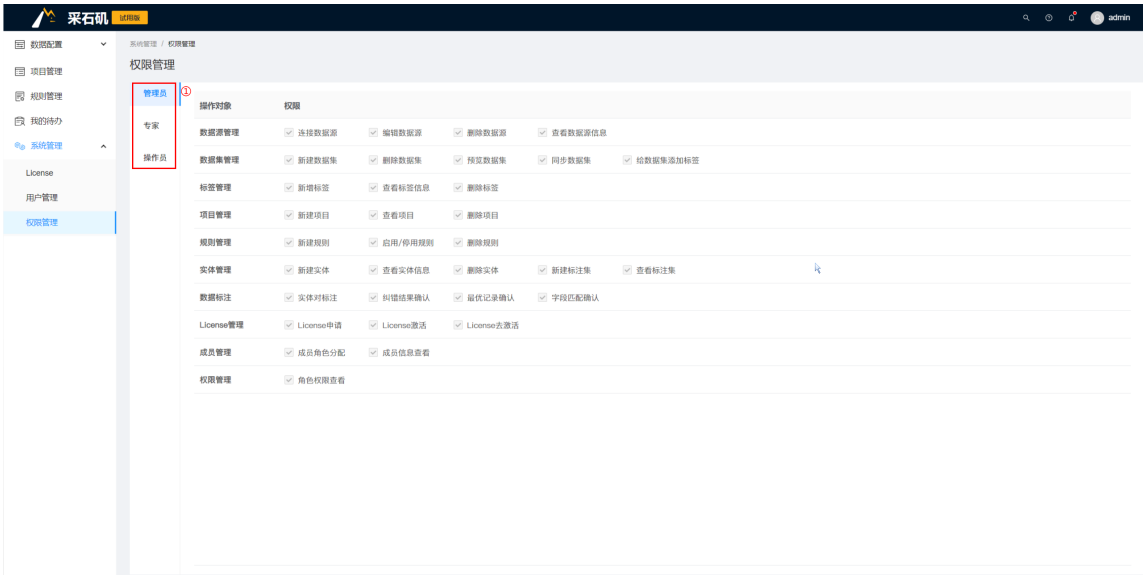
用户管理界面

> 用户管理中不能添加用户，该页面的数据同步自LDAP。

1. 用户账号或名称搜索框，可以通过账号或名称模糊搜索用户；
2. 专家角色过滤器，点击后弹出已选中、未选中过滤选项，可以过滤已授权或未授权专家角色的用户；
3. 操作员角色过滤器，点击后弹出已选中、未选中过滤选项，可以过滤已授权或未授权操作员角色的用户；
4. 专家授权复选框，选中后即授权专家角色，取消选中即取消专家角色授权；
5. 操作员授权复选框，选中后即授权操作员角色，取消选中即取消操作员角色授权；
6. 分页器，可切换选择每页显示内容数量。

2. 权限管理页面说明

点击 `权限管理` 菜单，会进入权限管理页面，具体呈现如下图。



权限管理界面

> 目前每个角色的权限是固定不可变的，所以页面中的复选框是置灰不可编辑的，仅能查看。

角色切换页签，点击可切换页签查看不同角色的权限。

典型应用场景配置方案

本章节主要介绍采石矶的典型应用场景和配置过程，让用户通过本章节可以轻松地掌握采石矶的使用过程和核心能力。

规则发现

目前大数据质量问题中存在大量的数据冲突、数据错误、实体不一致等问题，为解决这类问题必须要有针对性的规则，目前采石矶具备规则发现能力，采石矶融合了逻辑规则和机器学习，实现了跨表，表内关联关系规则挖掘，实现规则自动发现，无需手工设计规则，提高了数据分析师的工作效率，对于快速提升数据质量起到了很好的帮助，在采石矶系统中使用规则发现的步骤如下：

1、创建数据源：根据业务需要添加需要进行规则发现任务数据集所在的数据源，详细操作请参考”数据源“章节；

2、创建数据集：根据业务需要添加需要进行规则发现的数据集，详细操作请参考“数据集”章节；

3、创建项目：采石矶的所有业务都是以项目为粒度进行管理的，根据业务创建对应的项目，详细操作请参考“项目、工作流管理”章节；

4、创建任务：采石矶所有的业务都是以任务为驱动进行触发的，根据业务创建对应的任务，详细操作请参考“规则发现”章节。

经过上面的步骤后采石矶系统就能自动的进行数据分析与处理，输出对应的规则，数据分析师根据业务背景挑选适合自己的规则，应用于后续的查错、实体聚类等业务。

数据查错

数据查错是针对大数据质量问题中数据冲突问题的处理，采石矶系统作为数据质量增强系统，对于数据中的一致性和准确性等质量问题能快速方便地展现给用户，这样用户知道具体的数据问题后能快速地进行分析和处理，在采石矶系统中使用查错的步骤如下：

1、创建数据源：根据业务需要添加需要进行查错任务数据集所在的数据源，详细操作请参考”数据源“章节；

2、创建数据集：根据业务需要添加需要进行查错的数据集，详细操作请参考“数据集”章节；

3、创建项目：采石矶的所有业务都是以项目为粒度进行管理的，根据业务创建对应的项目，详细操作请参考“项目、工作流管理”章节；

4、创建任务：采石矶所有的业务都是以任务为驱动进行触发的，根据业务创建对应的任务，详细操作请参考”查错”章节。

经过上面的步骤后采石矶自动输出数据冲突，数据分析师根据冲突的展现情况能快速知道数据质量的位置和问题点。

数据纠错

数据纠错是针对大数据质量问题中数据冲突问题的处理，采石矶系统可以根据用户选定的规则对数据进行错误的修改和冲突数据的确认，提高数据的准确性，在采石矶系统中使用数据纠错的步骤如下：

1、创建数据源：根据业务需要添加需要进行数据纠错任务数据集所在的数据源，详细操作请参考”数据源“章节；

2、创建数据集：根据业务需要添加需要进行数据纠错的数据集，详细操作请参考“数据集”章节；

3、创建项目：采石矶的所有业务都是以项目为粒度进行管理的，根据业务创建对应的项目，详细操作请参考“项目、工作流管理”章节；

4、创建任务：采石矶所有的业务都是以任务为驱动进行触发的，根据业务创建对应的任务，详细操作请参考”数据纠错”章节。

经过上面的步骤后采石矶会自动进行数据修复，修复的数据经过分发确认后输出修复的结果，让用户轻松得到质量提升后的数据。

实体聚类

实体聚类是针对大数据质量问题中实体不一致的问题的处理，目的在解决不同系统中同一实体的记录如何关联的问题，采石矶系统可以根据用户指定的规则找出数据中属于同一实体的数据，将分散的实体信息关联到一起，在采石矶系统中使用实体聚类的步骤如下：

1、创建数据源：根据业务需要添加需要进行实体聚类任务数据集所在的数据源，详细操作请参考”数据源“章节；

2、创建数据集：根据业务需要添加需要进行实体聚类的数据集，详细操作请参考“数据集”章节；

3、创建项目：采石矶的所有业务都是以项目为粒度进行管理的，根据业务创建对应的项目，详细操作请参考“项目、工作流管理”章节；

4、创建任务：采石矶所有的业务都是以任务为驱动进行触发的，根据业务创建对应的任务，详细操作请参考”实体聚类”章节。

经过上面的步骤后采石矶会自动输出实体聚类的结果，让用户轻松知道哪些数据被判断为同一实体。

最优记录

最优记录针对的是大数据质量问题中一条实体下有多条记录，如何从多条记录中选出最优值的问题。为了解决这个问题，采石矶提供了最优记录的功能，将实体列表中的数据按着指定的最优规则找出与实体对应的最优记录，其使用步骤如下：

1、创建数据源：根据业务需要添加需要进行最优记录任务数据集所在的数据源，详细操作请参考”数据源“章节；

2、创建数据集：根据业务需要添加需要进行最优记录的数据集，详细操作请参考“数据集”章节；

3、创建项目：采石矶的所有业务都是以项目为粒度进行管理的，根据业务创建对应的项目，详细操作请参考“项目、工作流管理”章节；

4、创建任务：采石矶所有的业务都是以任务为驱动进行触发的，根据业务创建对应的任务，详细操作请参考”最优记录”章节。

经过上面的步骤采石矶根据规则自动输出实体下的最优记录，让用户轻松知道实体下的数据质量情况。

字段匹配

大数据质量问题现在比较凸显的是跨系统、跨表之间的数据结构拉通或者字段语义对齐问题，目前海量表中人工识别跨表的字段语义表达已经基本很难完成，所以采石矶提供了字段匹配的特性，来解决跨系统之间字段识别，其使用步骤如下：

1、创建数据源：根据业务需要添加需要进行字段匹配任务数据集所在数据源，详细操作请参考”数据源“章节；

2、创建数据集：根据业务需要添加需要进行字段匹配的数据集，详细操作请参考“数据集”章节；

3、创建项目：采石矶的所有业务都是以项目为粒度进行管理的，根据业务创建对应的项目，详细操作请参考“项目、工作流管理”章节；

4、创建任务：采石矶所有的业务都是以任务为驱动进行触发的，根据业务创建对应的任务，详细操作请参考”字段匹配”章节。

经过上面的步骤采石矶自动根据输入的数据进行抽样分析和深度学习，输出跨系统的字段匹配结果，让用户轻松知道在不同系统下定义为不同名称的字段，从语义上表达的是相同意思。