数据采集

产品文档

文档目录

- 产品简介
  - 产品概述
  - 产品优势
  - 应用场景
- 快速入门
  - 子帐号登录
  - 查看开通的服务
  - 创建项目
  - 进入数据采集组件
- 操作指南
  - 新建topic（用于流式采集）
    - 设计态–新建topic
    - 发布到测试
    - 发布到生产
  - 新增数据交换接口（用于批量采集）
    - 登记Bucket
    - 新建数据交换接口
    - 发布到测试
    - 发布到生产
  - 流式采集配置
  - 流式agent采集
    - 创建采集任务
    - 配置基本信息
    - 配置agent信息–流式agent采集支持采集文件、文件夹、Kafka三种类型的数据。
      - 文件采集
      - 文件夹采集
      - Kafka采集
      - 高级配置
        - 高级配置
        - 缓存配置
        - 传输配置
    - 下载并启动agent
    - 数据预览
    - 查看agent列表
  - 流式API采集

创建采集任务
配置基本信息
下载接口规范
流式数据库采集
创建采集任务
配置基本信息
配置agent信息–流式数据库采集支持采集MySQL的数据。
MySQL采集说明
下载并部署agent
特别说明
查看agent列表
批量采集配置
批量采集配置
文件推送任务配置
文件推送（Excel批量上传）配置
运行批量采集任务
文件拉取
查看批量采集运行实例
通过页面进行文件上传
最佳实践
常见问题
词汇表

产品简介

产品概述

最近更新时间: 2019-10-27 01:57:00

数据采集组件是大数据云服务的全部数据入口，为大数据云服务提供数据处理的来源。采集组件将数据采集进大数据云内后，在数据调度组件的调度下，数据集成、存储与计算等各组件协同工作，为用户提供大数据处理服务。本文档描述了数据采集组件的产品简介、快速入门、最佳实践、常见问题等信息。

产品优势

最近更新时间: 2019-10-27 01:58:44

数据采集服务，能够将云外的数据接入云内，支持多类型、多场景的全域数据采集。服务支持结构化、半结构化、非结构化的多数据格式数据采集；流式、批量，每种场景互相独立、可配置。

应用场景

最近更新时间: 2019-11-26 14:57:59

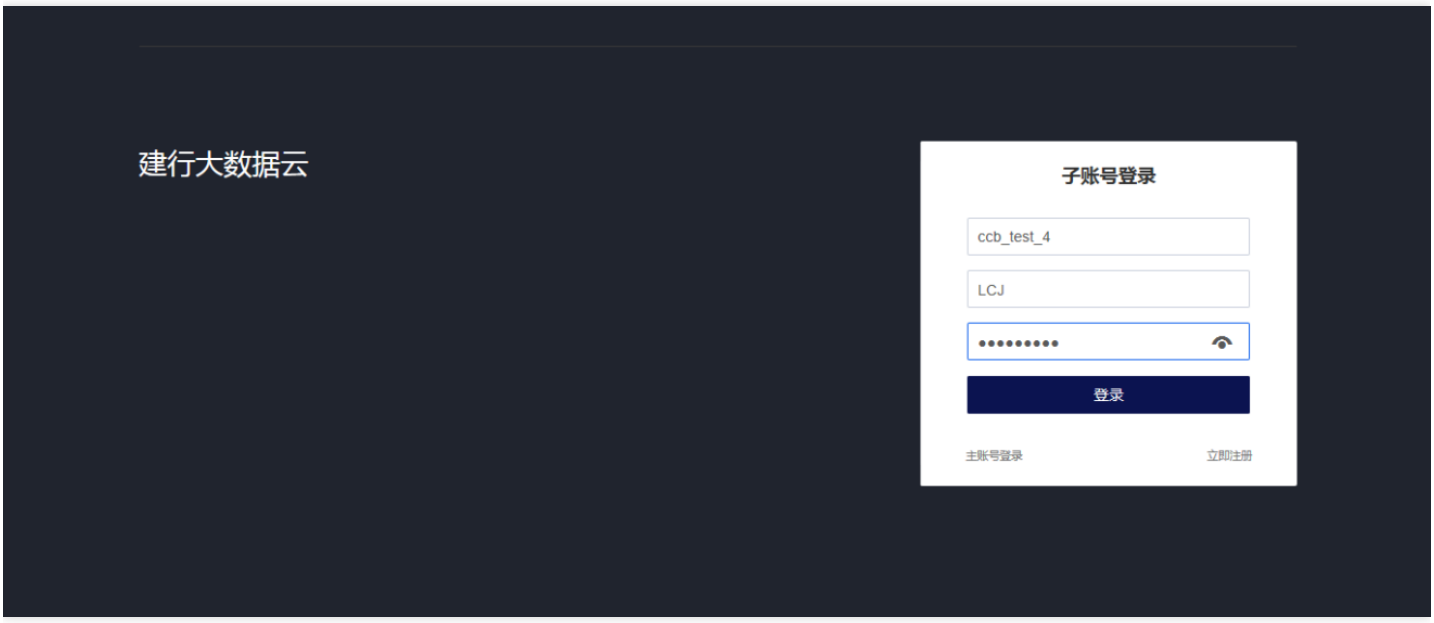
一、 流式数据采集 为流计算提供数据源。流式数据采集适用于可以直接进行数据计算，实时性要求很严格，但数据的精确度要求不太苛刻的应用场景。

二、 批量采集文件推送 为批处理提供数据源，客户将业务数据积攒整理成数据文件后，再批量推送上云进行处理。批量采集对应于先存储后计算，对实时性要求不高，同时数据的准确性、全面性更为重要的应用场景。

快速入门
子帐号登录

最近更新时间: 2019-11-11 09:22:09

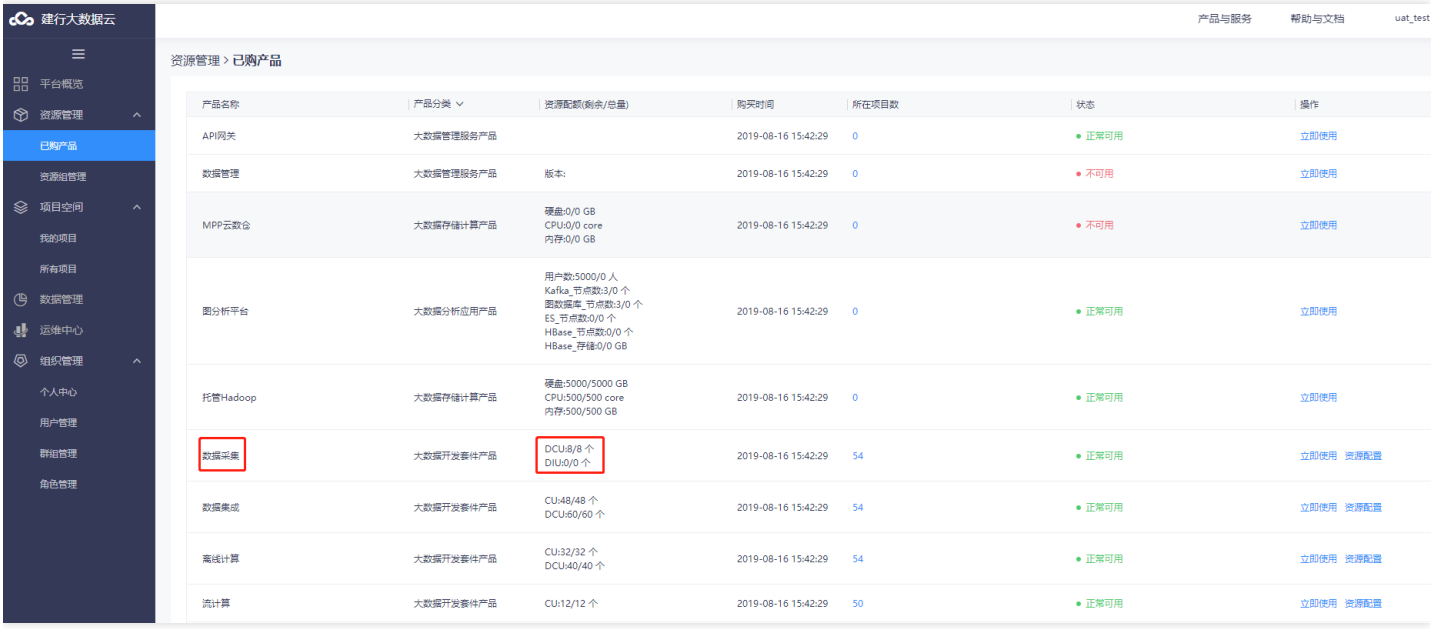
1） 进入租户控制台登录页 2） 点击【子帐号登录】按扭，显示子帐号登录页 3） 按要求填写租户名、子帐号名、子帐号密码，点击登录，即进入租户控制台。

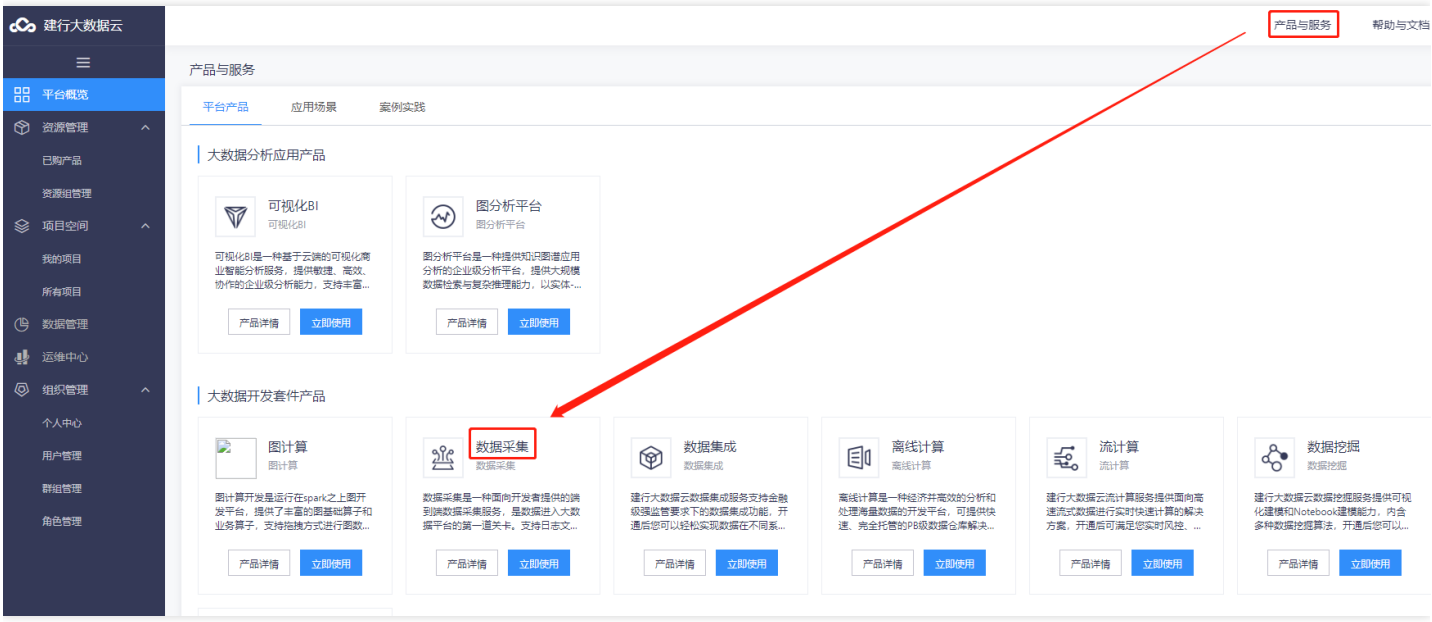


查看开通的服务

最近更新时间: 2019-11-11 09:22:09

登陆后，点击左侧的“资源管理-产品与服务”菜单，或点击上方的“产品与服务”，可查看所有开通的服务。

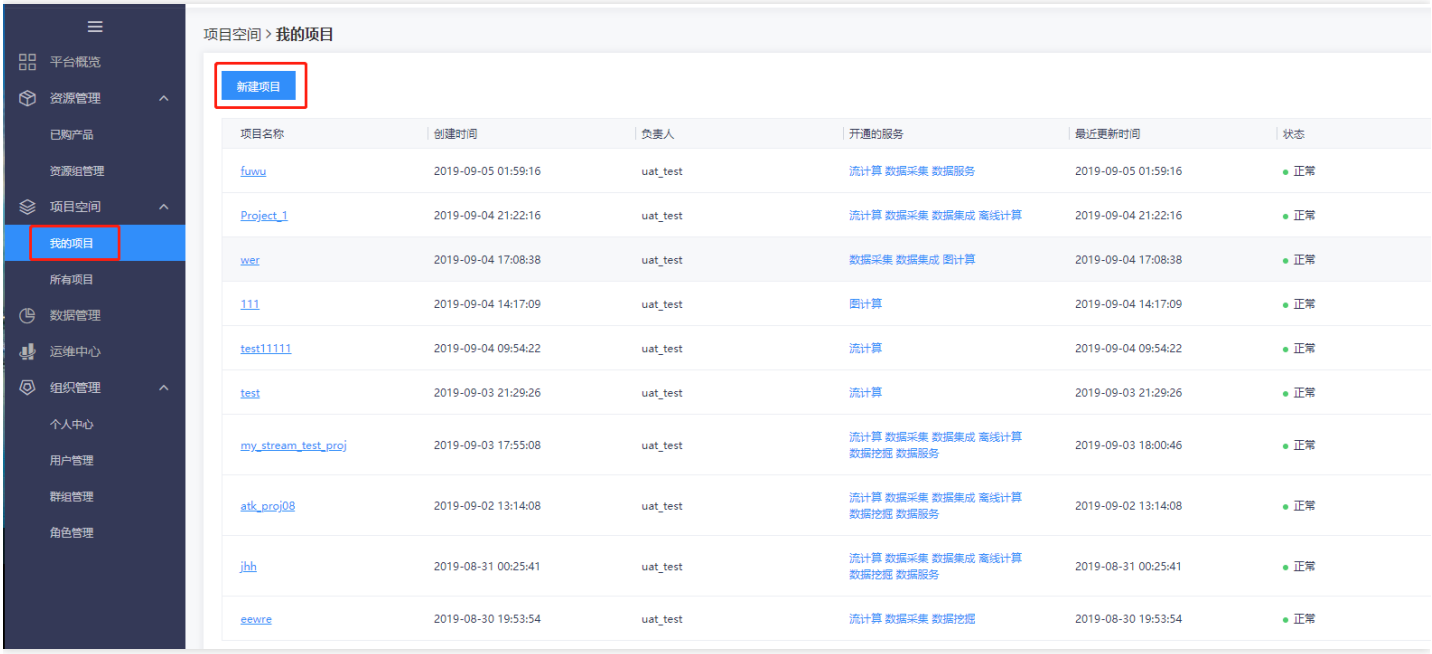




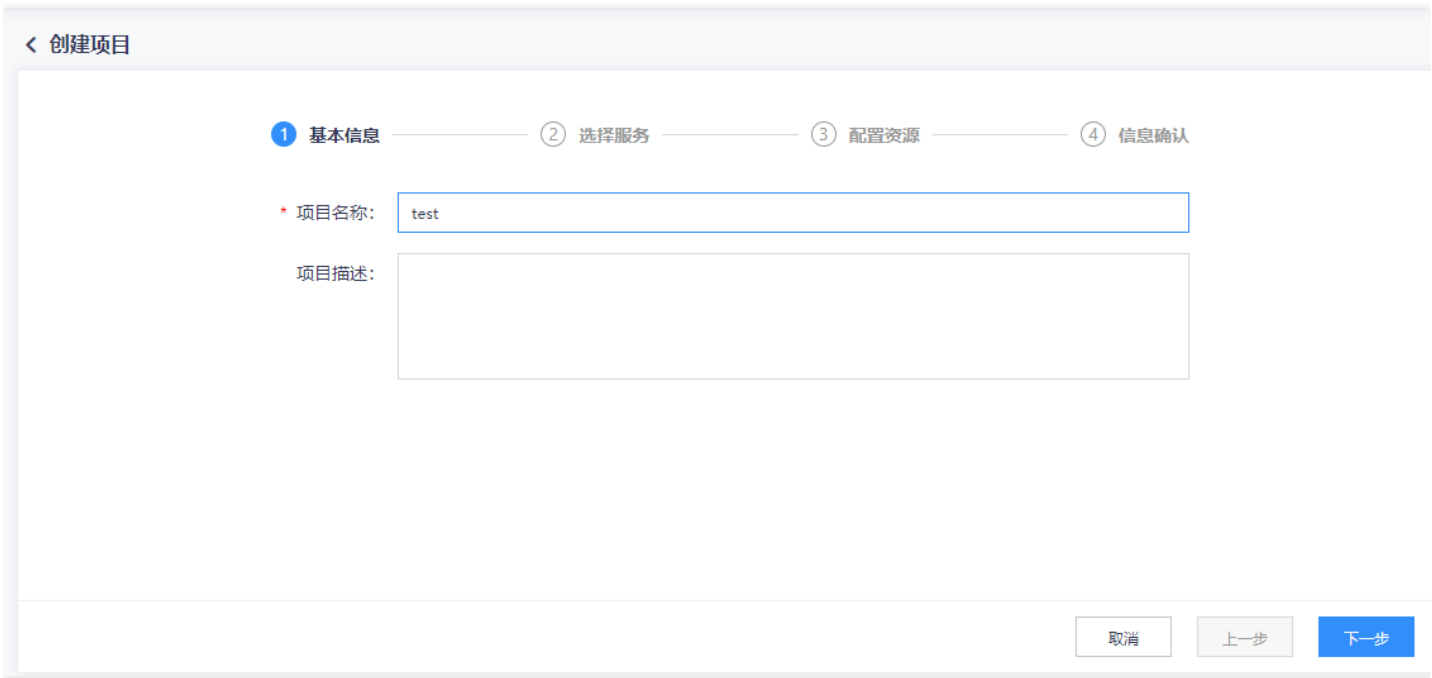
创建项目

最近更新时间: 2019-11-11 09:51:29

点击页面左侧“项目空间-我的项目”选项卡，可以看到当前已有的所有的项目列表。

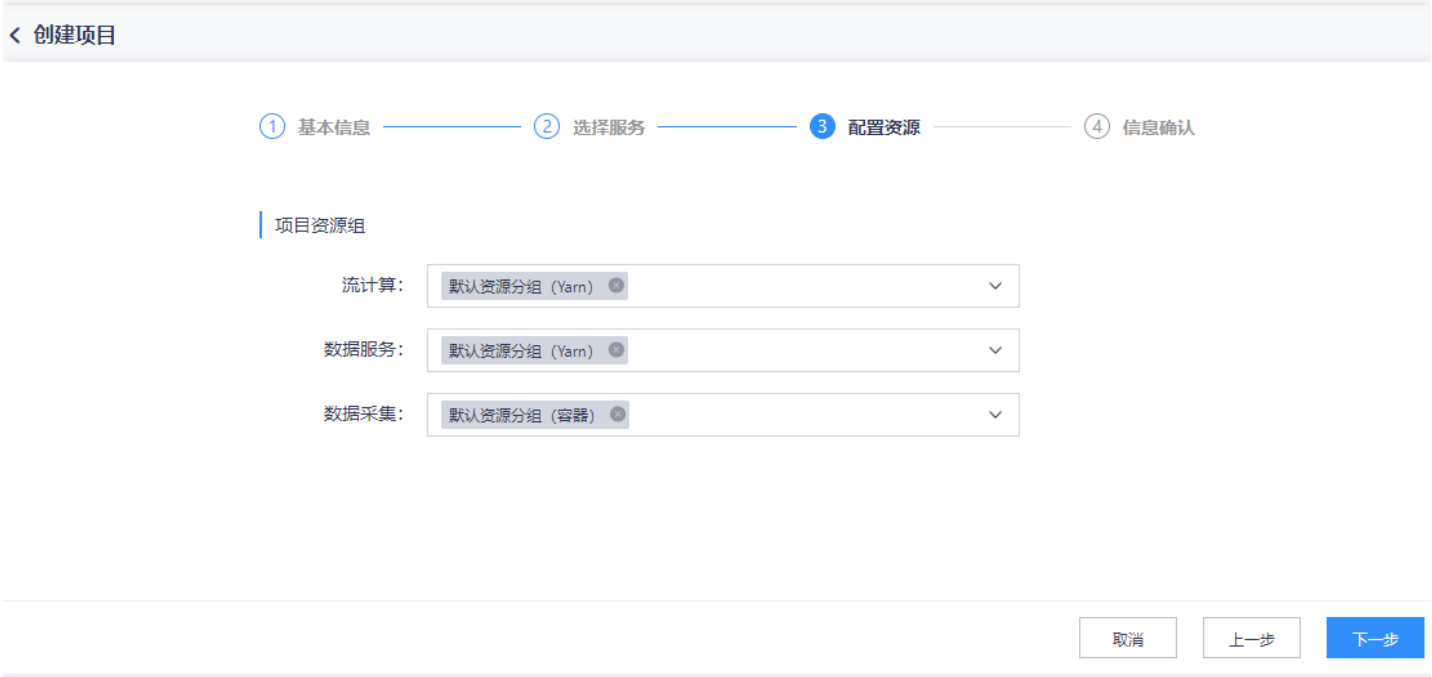


点击蓝色的“新建项目”按钮，弹出“新建项目”对话框。



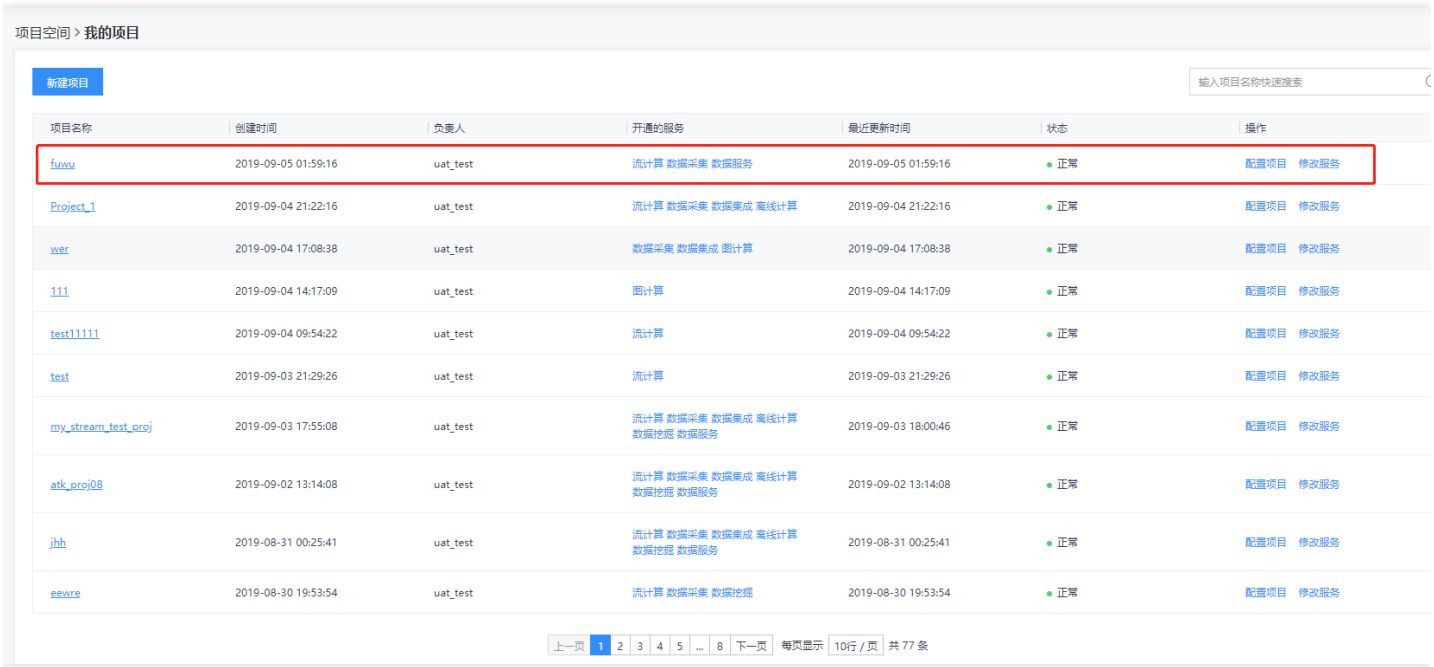
这里可以根据需要选择各项服务，本次新建项目需包含“数据采集”，其余服务按需选择。点击“下一步”配置项目的各项属性参数。

点击下一步，在项目资源组页面，通过下拉框中，为所选的服务选择资源分组。配置完成后，点击“下一步”。 备注：数据采集使用容器资源，且容器资源租户内共用（即数据采集的默认资源组），项目开通时，若选择开通数据采集服务，均需选择数据采集组件开通的“默认资源组（容器）”。



对以上配置信息进行确认，无误后点击“确定”按钮。

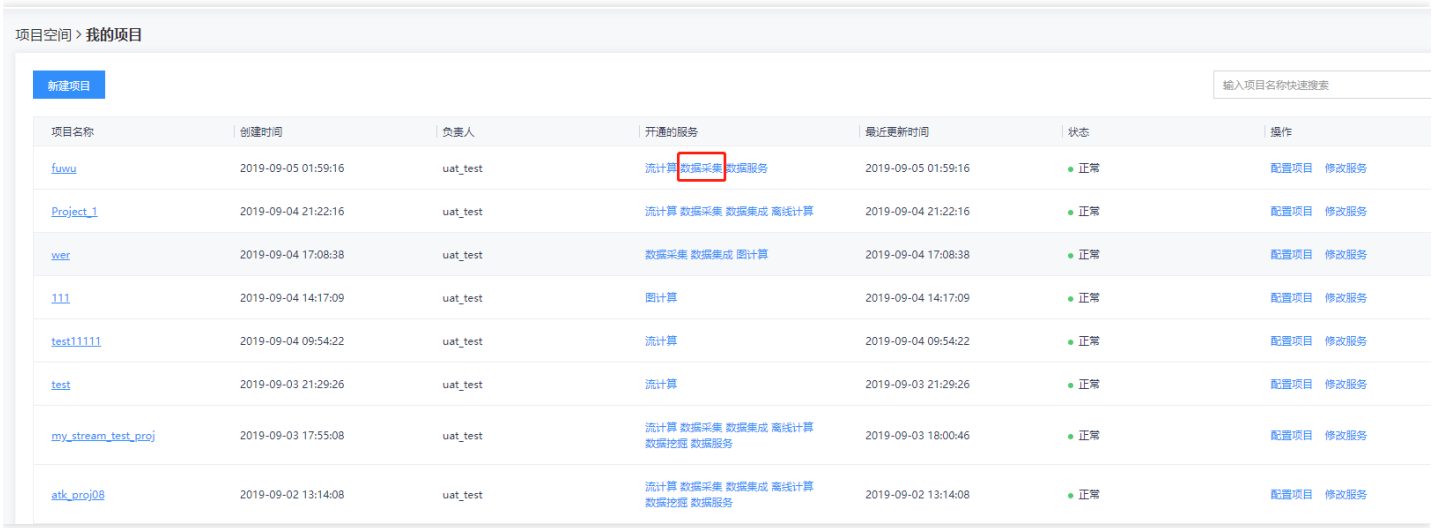
创建成功的项目会显示在项目列表的最上面。点击“开通的服务”列中的服务名称，可以对项目中的服务进行新建、修改配置、删除等操作。点击项目名称，进入项目的开发页面。

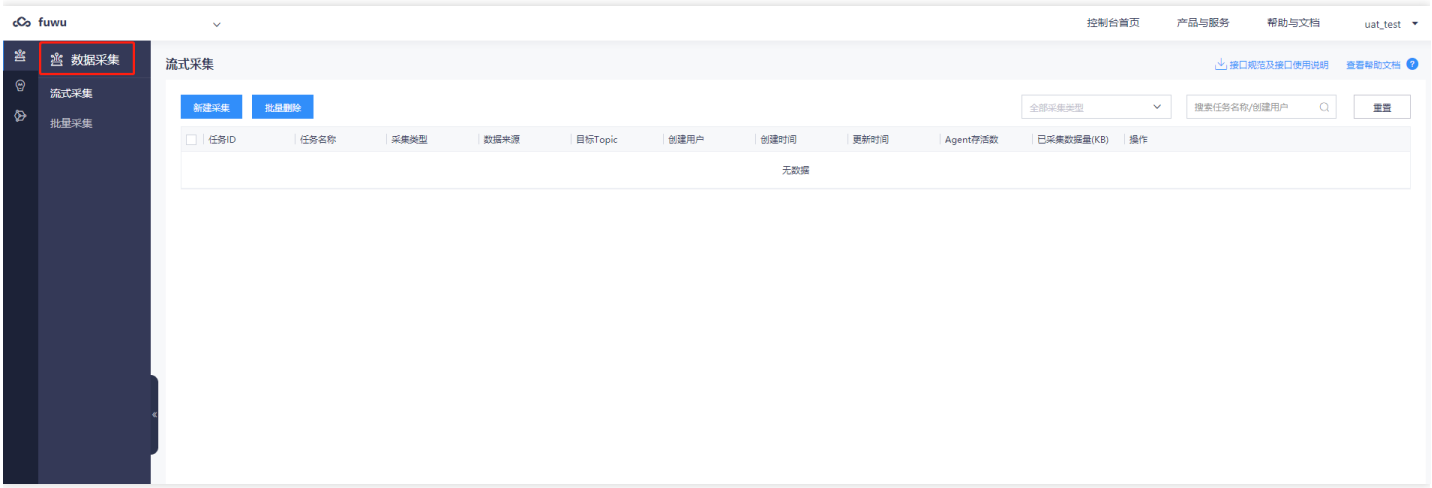


进入数据采集组件

最近更新时间: 2019-11-11 09:51:29

找到开通数据采集权限的项目，点击项目名称进入特定项目，点击【数据采集】进入数据采集开发界面（或直接点击项目列表中“开通的服务”列的“数据采集”，也可直接进入数据采集界面）。





操作指南

新建topic（用于流式采集）

设计态–新建topic

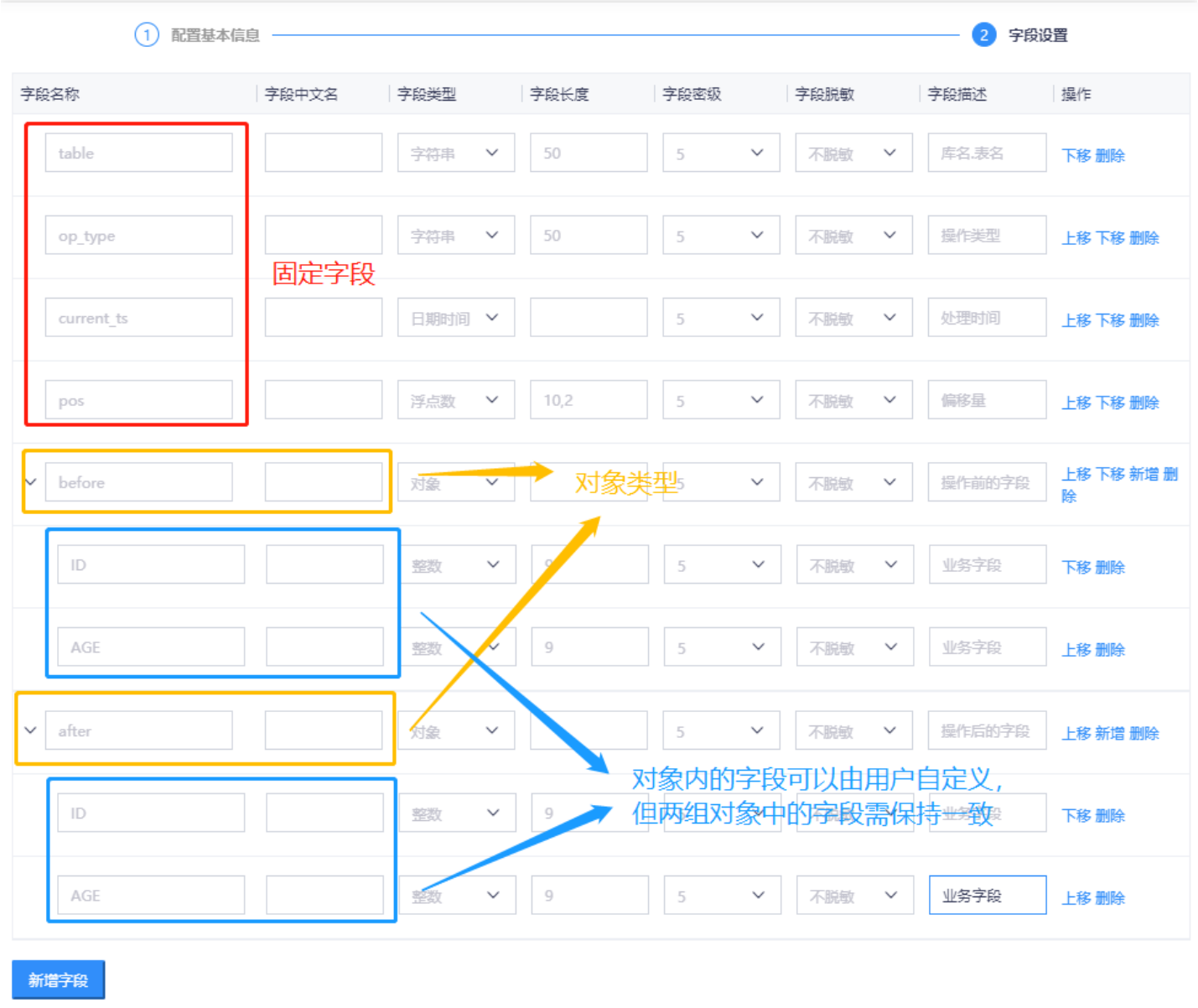
最近更新时间: 2019-11-11 09:51:29

新建topic时需指定topic归属的项目，即此topic可在归属项目下使用。

完成字段设置后，点击“完成”按钮，结束topic创建过程。 特别注意：当创建的topic用于流式数据库采集时，对topic的字段设置有特定要求，具体说明如下。

1. 当作为MySQL数据库采集的投递目标topic时，字段设置限定如下： 创建样例如下图所示：

```
{
"table": "TCLOUD.T_MySQL", //库名.表名
"op_type": "U", //操作类型 U更新 D删除 I插入
"current_ts": "2018-05-31T14:49:01.709000", //【处理时间】
"pos": "00000000000000003770", //偏移量
"before": { //object类型，操作前的字段
"XXX_A ": 1, //业务字段
"XXX_B": 20,
},
"after": { // object类型，操作后的字段
"XXX_A ": 1, //业务字段
"XXX_B": 20,
}
}
```

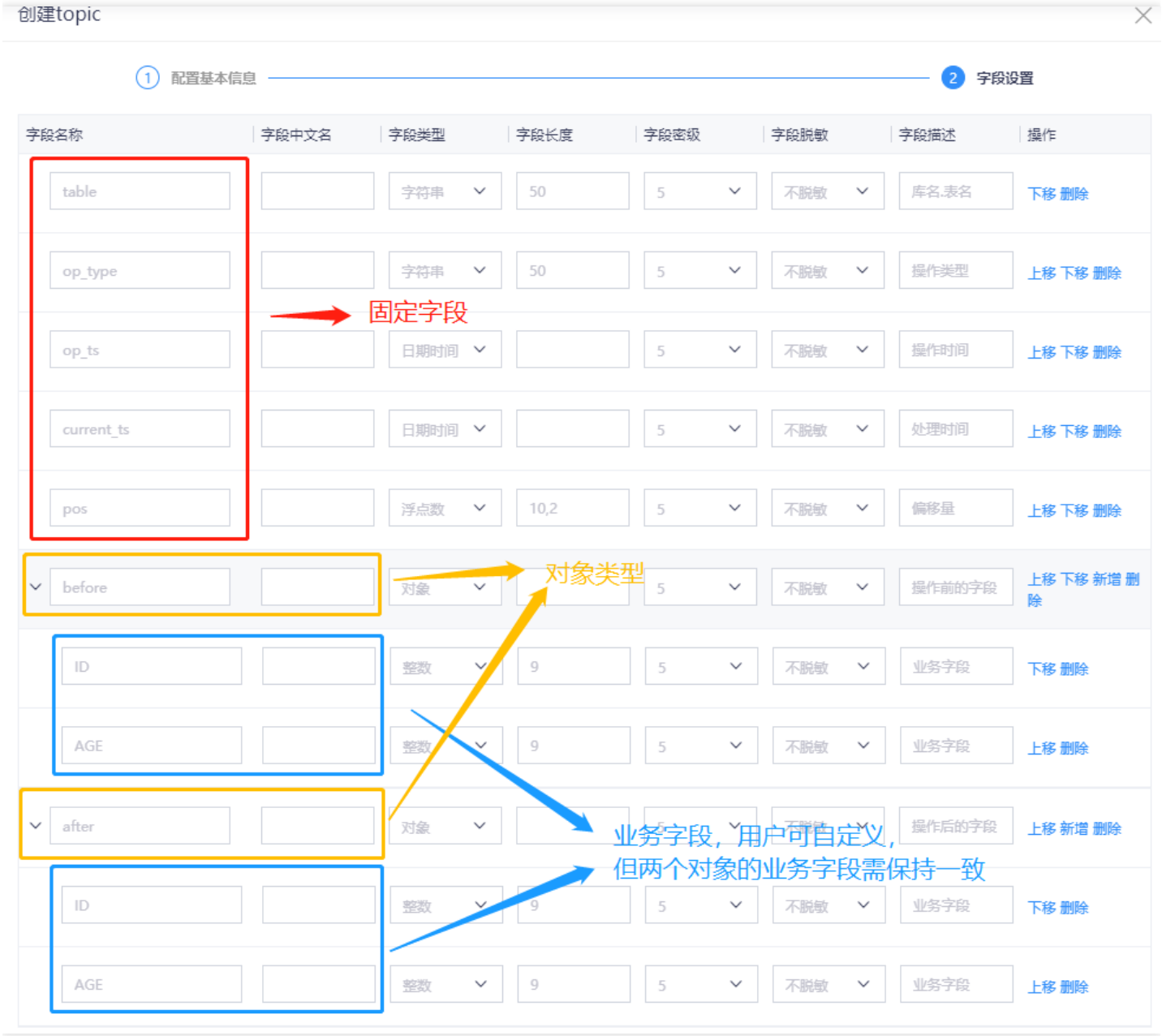


2. 当作为Oracle数据库采集的投递目标topic时

```
{
"table": "TCLOUD.T_OGG2", //库名.表名
"op_type": "U", //操作类型 U更新 D删除 I插入
"op_ts": "2018-05-31 14:48:55.630340", //操作时间
"current_ts": "2018-05-31T14:49:01.709000", //【处理时间】
"pos": "00000000000000003770", //偏移量
"before": { //object类型，操作前的字段
"XXX_A ": 1, //业务字段
"XXX_B": 20,
},
"after": { // object类型，操作后的字段
"XXX_A ": 1, //业务字段
"XXX_B": 20,
```

```
}
}
```

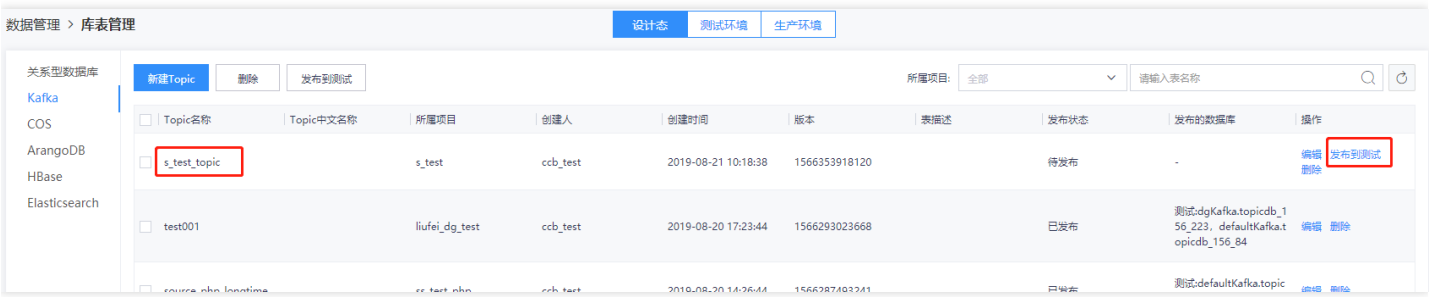
创建样例如下图所示：


创建topic
1 配置基本信息
2 字段设置
字段名称
字段中文名
字段类型
字段长度
字段密级
字段脱敏
字段描述
操作
table
op_type
op_ts
current_ts
pos
before
ID
AGE
after
ID
AGE
固定字段
对象类型
业务字段，用户可自定义，
但两个对象的业务字段需保持一致

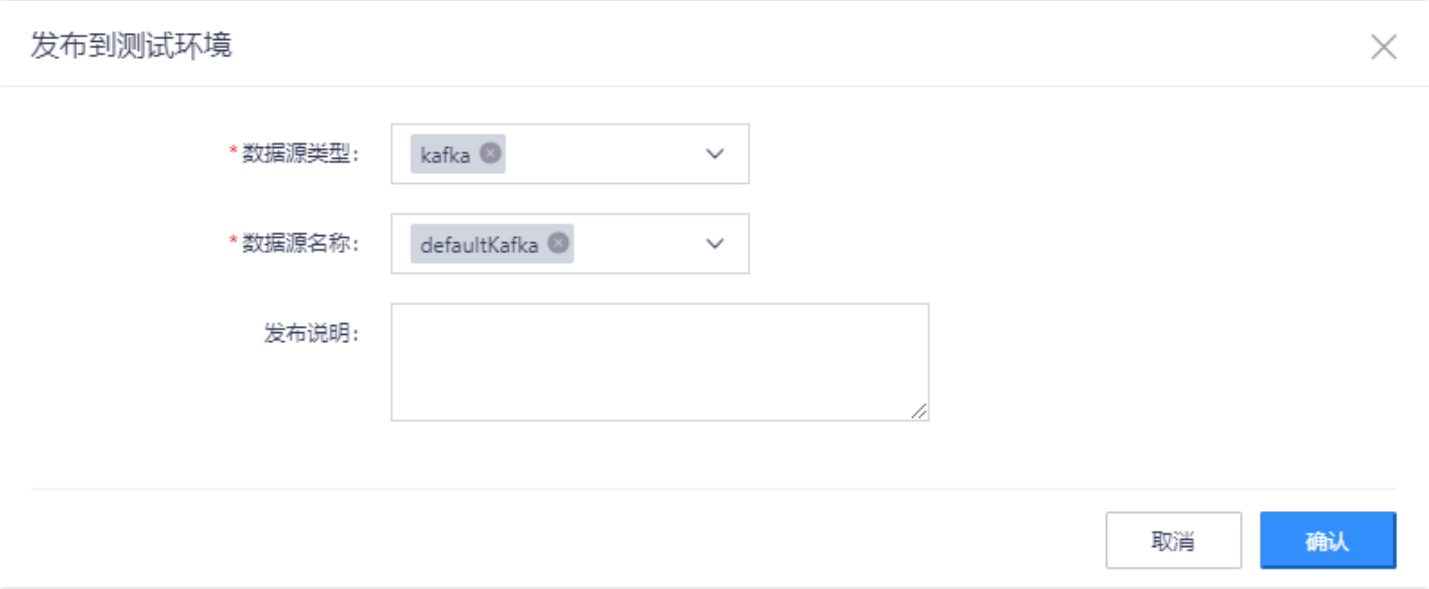
发布到测试

最近更新时间: 2019-11-11 09:51:29

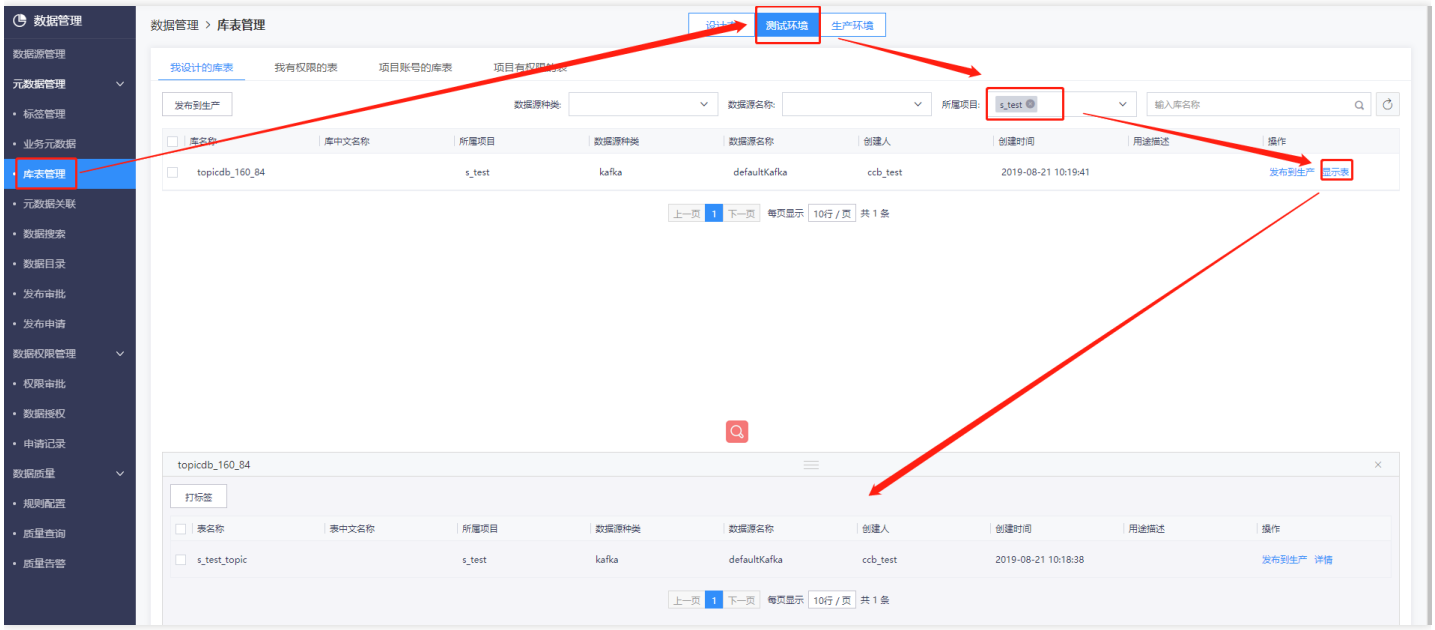
topic发布到测试环境后，可在测试环境使用



发布测试时，需指定此topic归属的数据源



发布测试完毕后，可在测试环境中筛选查看：



发布到生产

最近更新时间: 2019-11-11 09:51:29

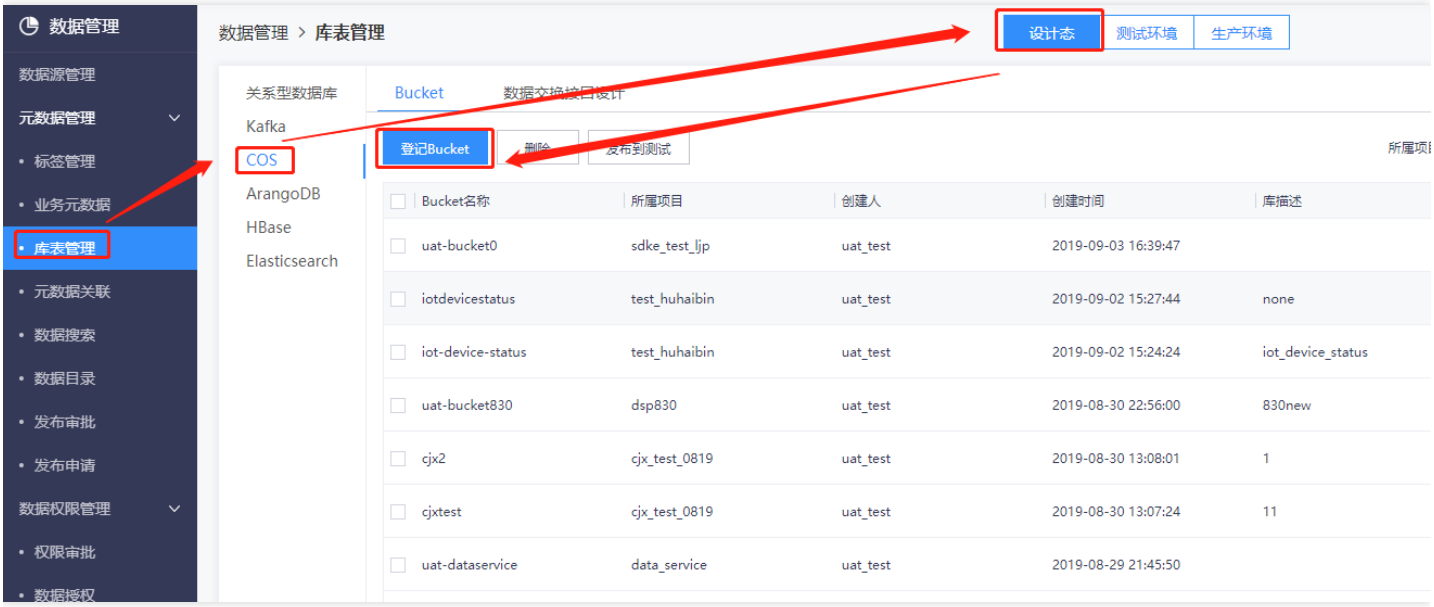
topic发布到生产后，可在生产环境使用。

新增数据交换接口（用于批量采集）

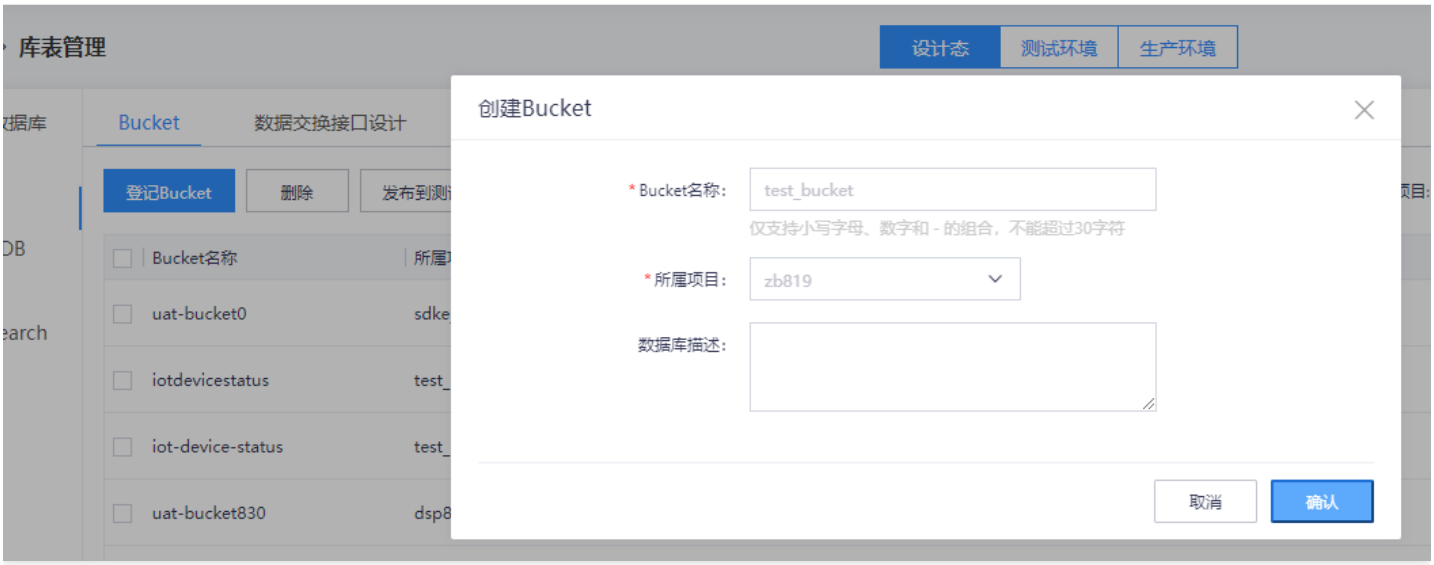
登记Bucket

最近更新时间: 2019-11-11 09:51:29

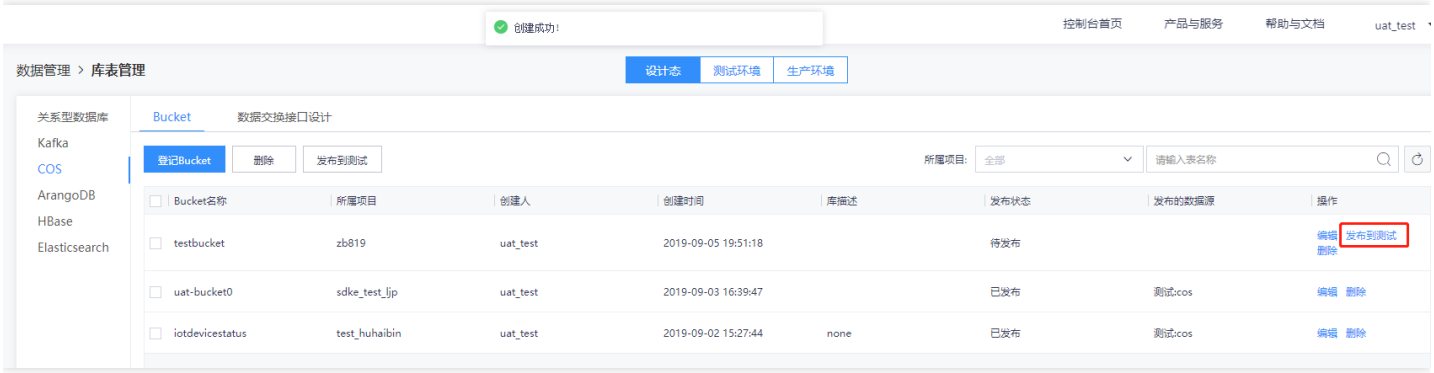
新建数据交换接口前，需要为当前项目绑定可用的Bucket。进入项目专属开发页面，在左侧依次点击“数据管理”→“元数据管理”→“库表管理”，在右侧点击“设计态”→“Bucket登记”→“登记Bucket”。



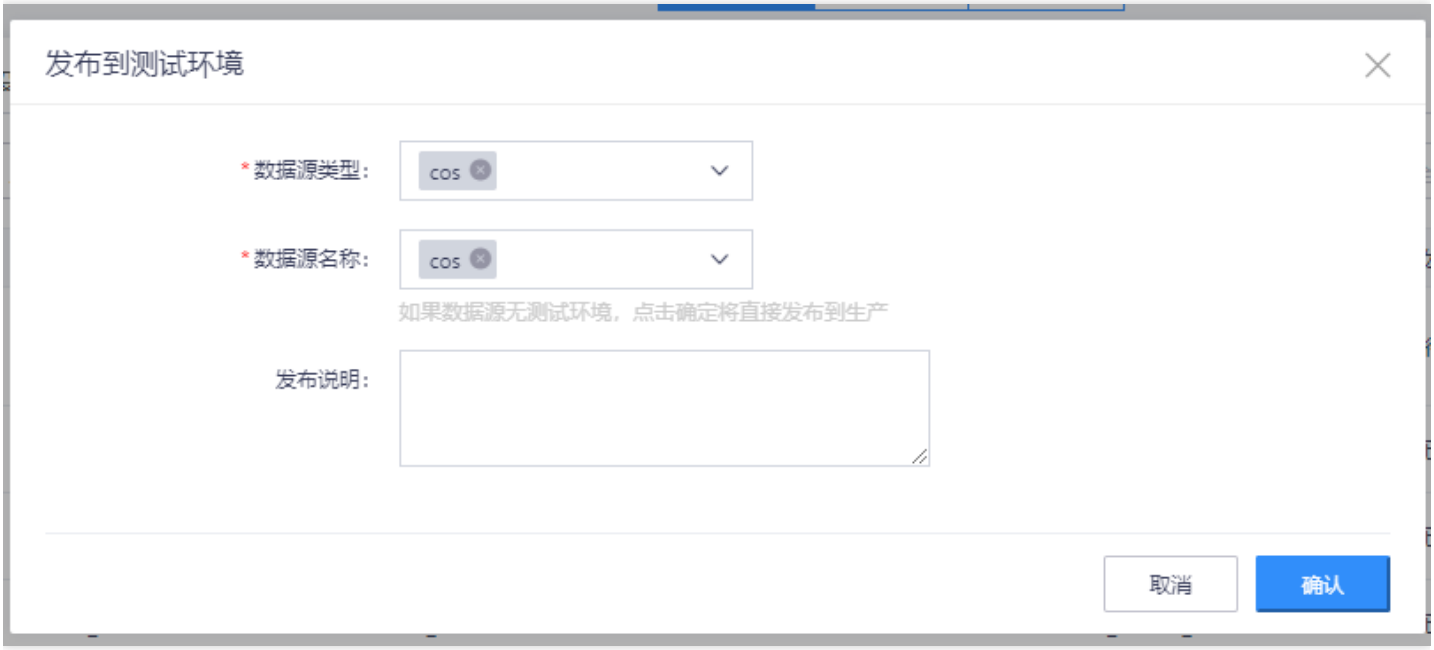
在弹出页面中，填写创建Bucket的各项属性。“Bucket模型名称”为用户自定义，填写“测试Bucket名称”与“生产Bucket名称”时，需填写COS中已存在的Bucket，“所属项目”选择新建的项目。配置完后点击确定。



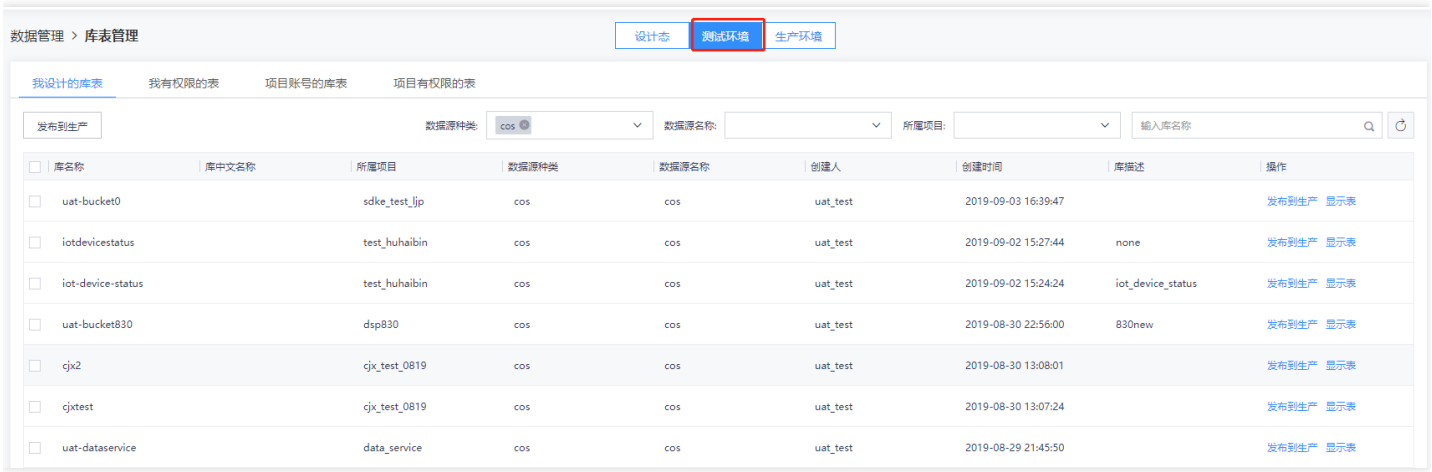
在Bucket列表中可以看到新建的Bucket。点击“操作”列发布到测试，将该Bucket发布到测试环境。



在弹出页面选择“数据源类型”为COS，选择合适的“数据源名称”，点击“确认”发布。经有审批权限的账号审批后，该发布生效。



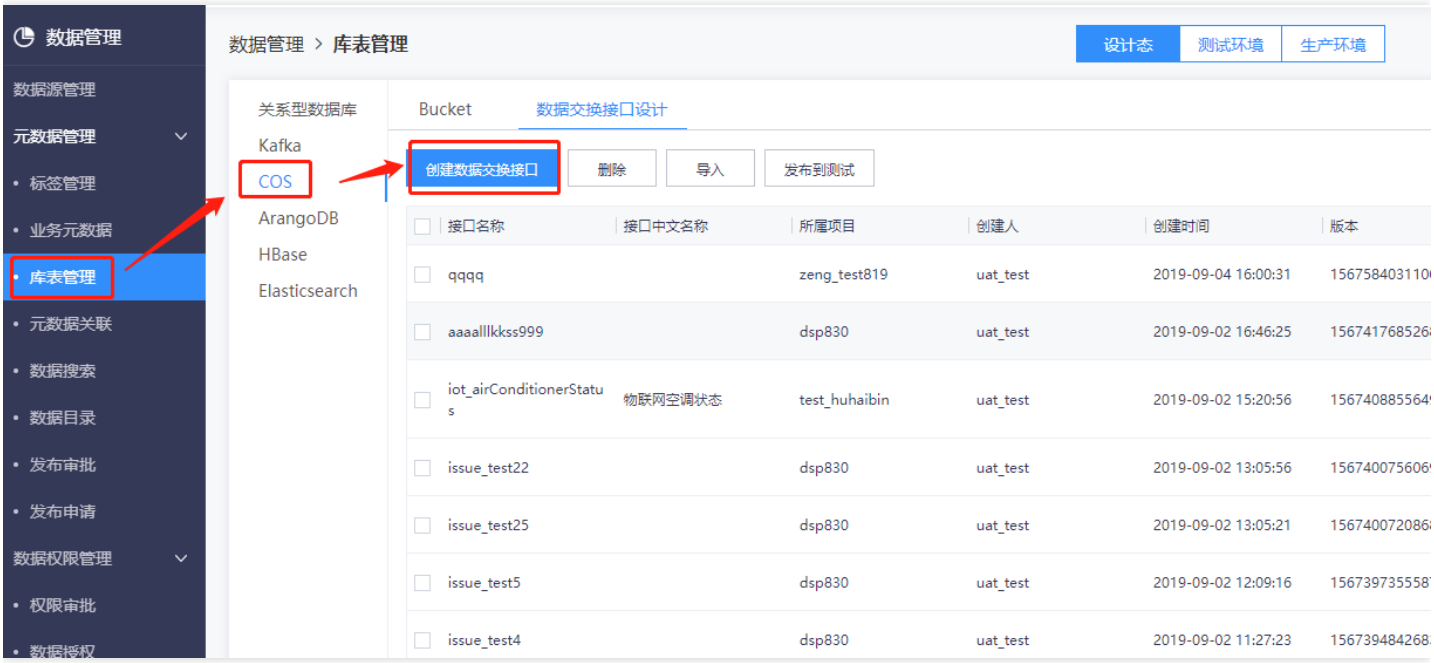
发布成功后，Bucket信息可在测试环境查看。



新建数据交换接口

最近更新时间: 2019-11-11 09:51:29

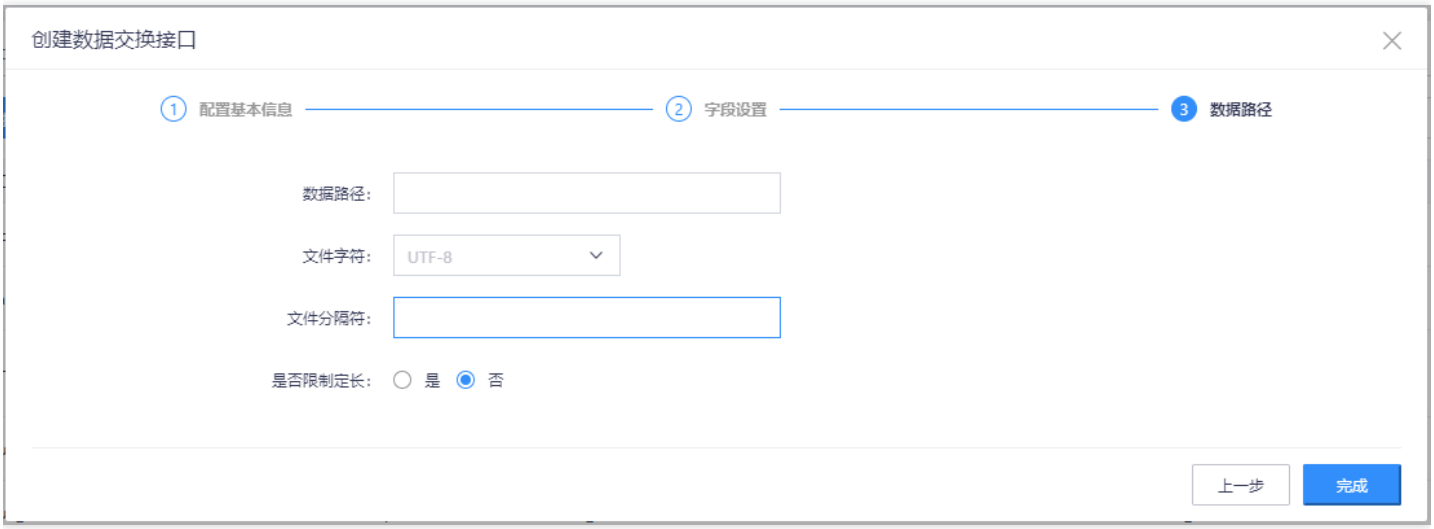
在项目专属的开发页面，也可以新建文件推送服务。与流式采集类似，创建文件推送任务前，需要先在数据管理页面新建数据交换接口。依次点击页面左侧的“数据管理”→“元数据管理”→“库表管理”，在右侧面板上依次点击“设计态”→“数据交换接口设计”→“创建数据交换接口”。



在新弹出页面填写数据交换接口的各项信息，在“所属项目”选择之前新建的项目，配置完成后点击“下一步”。

创建数据交换接口时，需要为该接口添加字段。为每个字段设置字段名称、字段类型、字段长度等属性。点击“新增字段”可以增加新的字段。设置完所有字段后点击“下一步”。

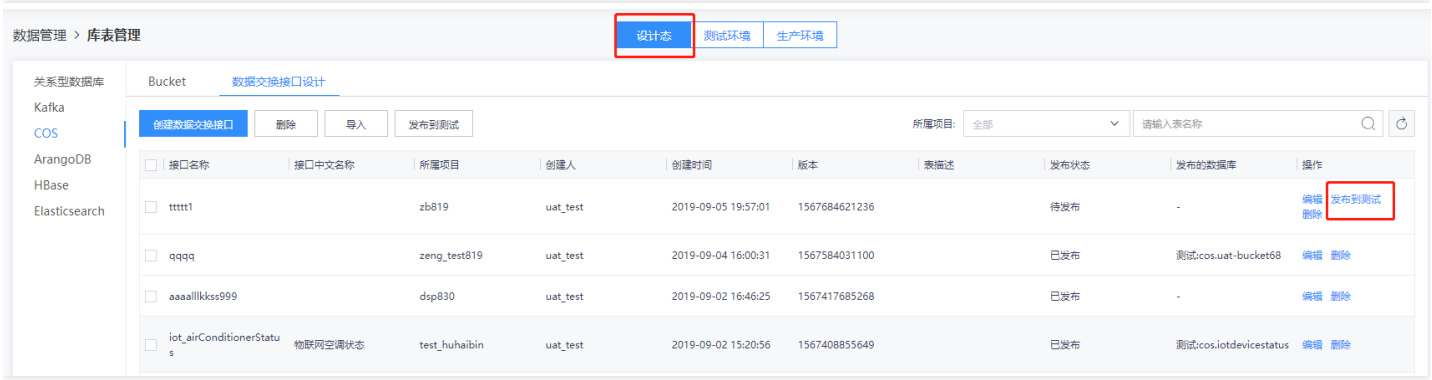
创建数据交换接口的最后一步需要为接口指定数据交换路径。由于文件推送执行时会指定路径，“数据路径”需填写但不会实际生效。“文件字符”和“文件分隔符”按需填写，“文件字符”不填时默认使用UTF-8，“文件分隔符”不填时默认使用|，点击“完成”结束配置。



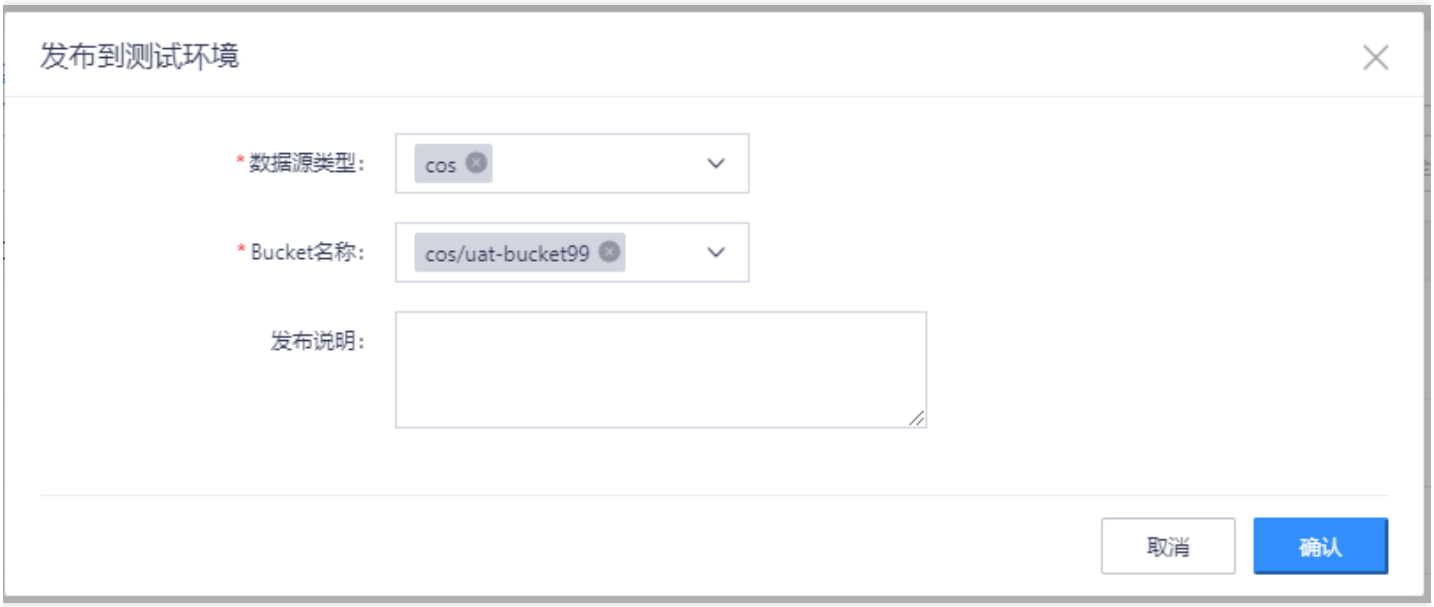
发布到测试

最近更新时间: 2019-11-11 09:57:37

此时，可以在数据交换接口列表中看到新创建的接口，其状态为待发布。点击“操作”列“发布到测试”，以将该接口发布到测试环境。

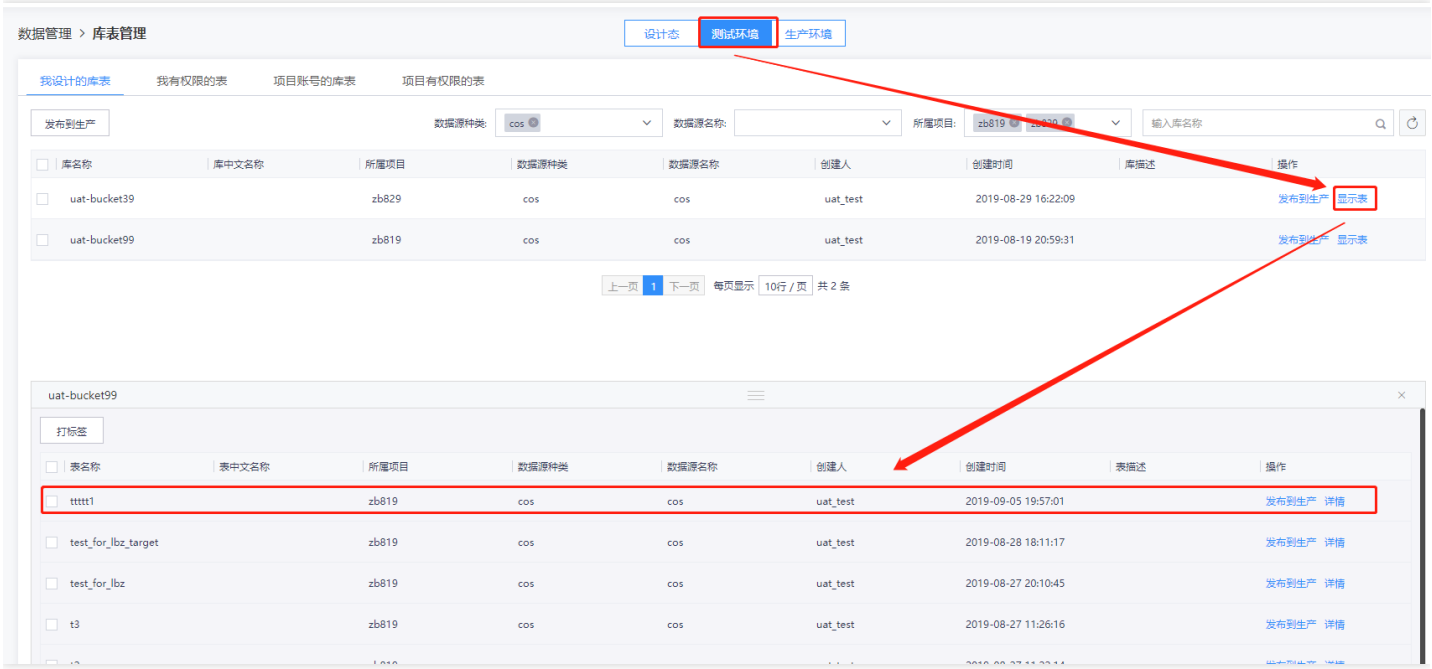


在弹出页面中，选择“数据源类型”为COS，“数据源名称”与“Bucket名称”需按照登记Bucket时填写的信息进行选择，配置完后点击“确认”发布。经有审批权限的账号审批后，该发布生效。



点击“测试环境”→“我设计的库表”，在数据库列表中，可以看到已登记的Bucket命名的数据库，点击“操作”列

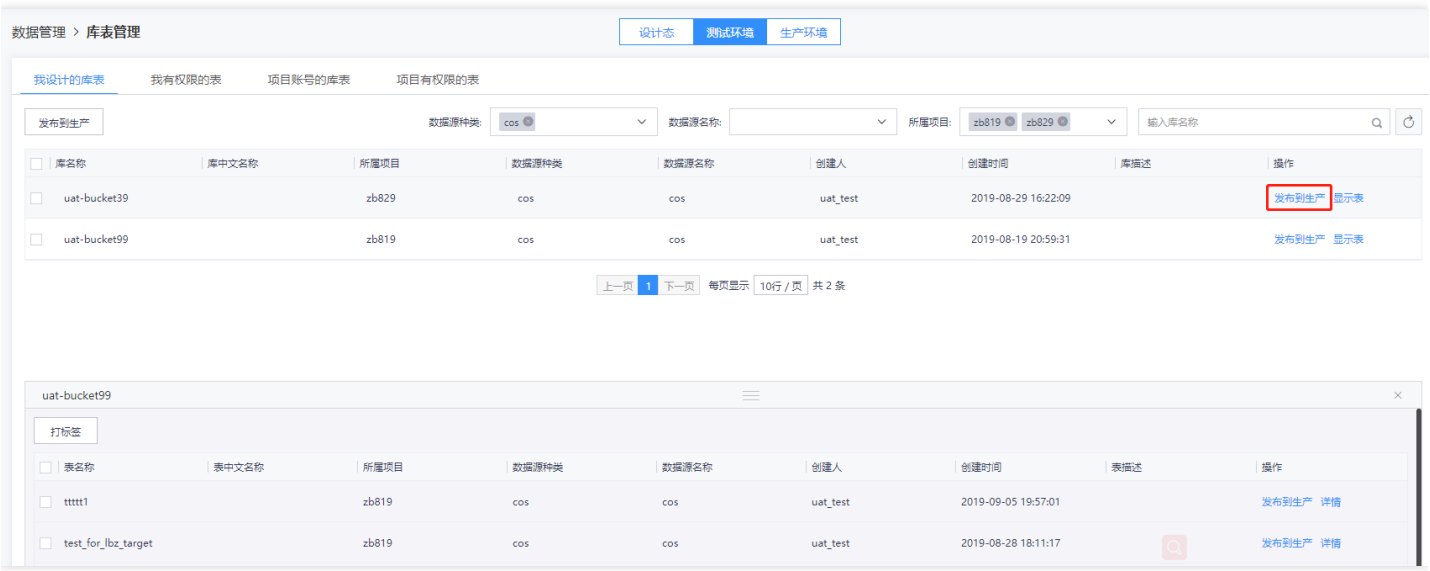
的“显示表”，可以看到已发布到测试环境的以数据交换接口命名的表。



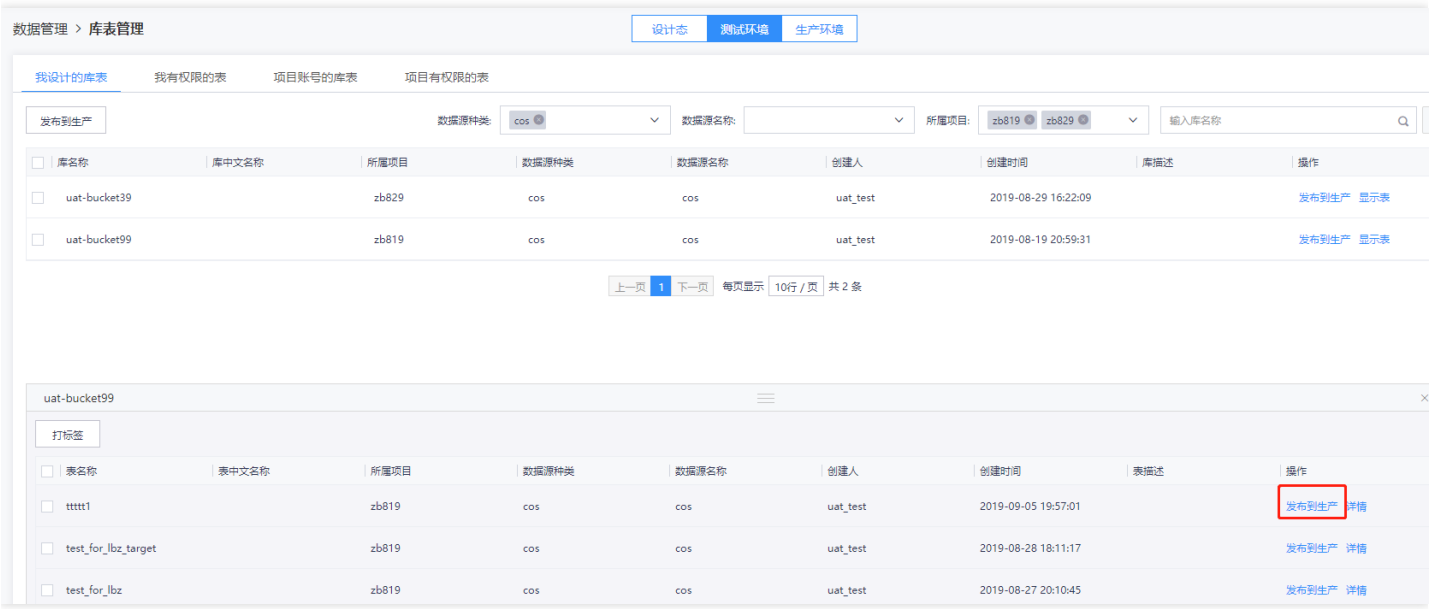
发布到生产

最近更新时间: 2019-11-11 09:57:37

点击对应的库名称“操作”列的“发布到生产”，将Bucket对应的数据库发布到生产。审批通过后，发布成功。

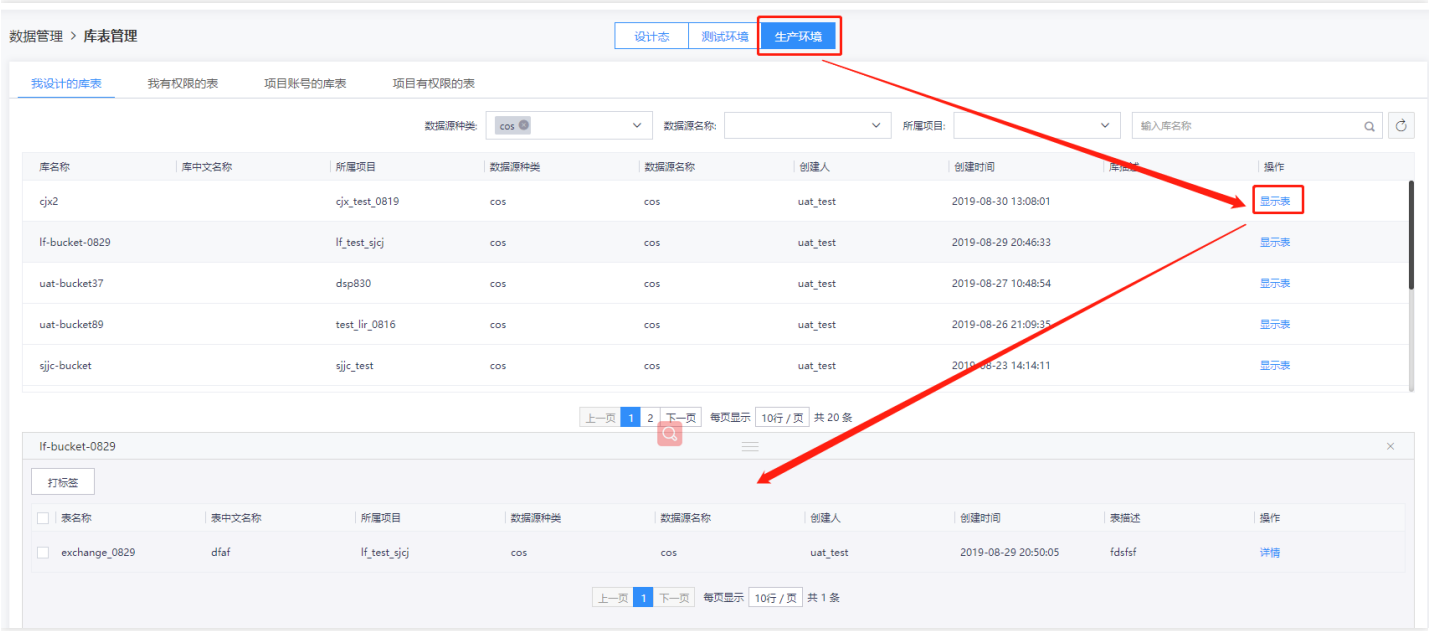


找到该数据库下对应的表，点击“发布到生产”，将数据交换接口对应的表发布到生产。审批通过后，发布成功。



依次点击“生产环境”→“我设计的库表”，就可以看到已发布到生产的数据库，点击“操作”列“显示表”，可以看到已

发布到生产的数据交换接口对应的表。

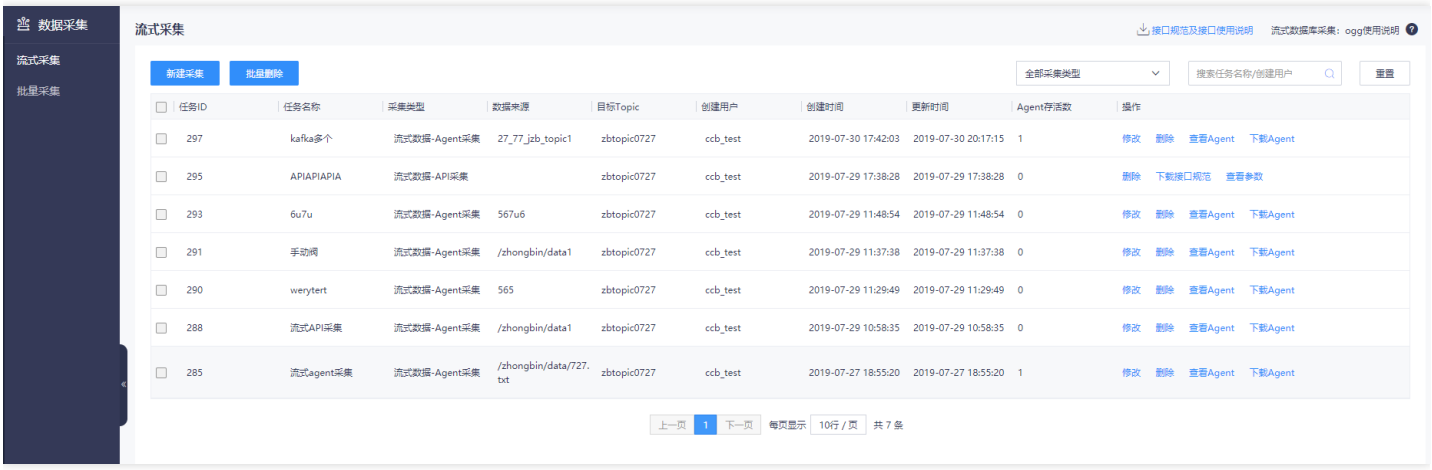


流式采集配置

最近更新时间: 2022-01-21 15:56:29

数据采集组件支持流式采集、批量采集两种方式，进入开发页面，默认展开流式采集页面。

流式采集任务可将文件、外部Kafka、文件夹、MySQL数据库等数据，实时采集至大数据云平台的Kafka中，采集页面支持：新增、修改、删除、查看agent、下载接口规范/agent等操作。



点击“新建采集”按钮，可创建采集任务，流式采集支持：① 基于agent的流式数据采集，② 基于API的流式数据采集、③ 针对MySQL数据库的流式数据库采集。

流式采集任务可通过2种方式启动运行：

运行方式	使用范围	使用方式

agent启动	流式数据采集、数据库采集	客户端部署并启动agent
API启动运行	流式数据采集	通过调用API启动流式采集任务

流式agent采集
创建采集任务

最近更新时间: 2019-11-12 03:04:33

点击页面的【新建采集】按钮，在弹出的抽屉中，点击“流式数据采集-agent采集”创建采集任务。

配置基本信息

最近更新时间: 2019-11-26 14:57:59

在弹出的窗口中填写新建采集任务的基本信息，必填参数说明如下：

- 采集名称：支持中文、英文、数字、下划线，最大50字符
- topic名称：下拉选择，可选择该项目下有权限的所有topic
- 选择topic时，支持对topic字段信息进行预览，如下图所示 配置完毕后，点击“下一步”，进行agent信息配置。

配置agent信息-流式agent采集支持采集文件、文件夹、Kafka三种类型的数据。文件采集

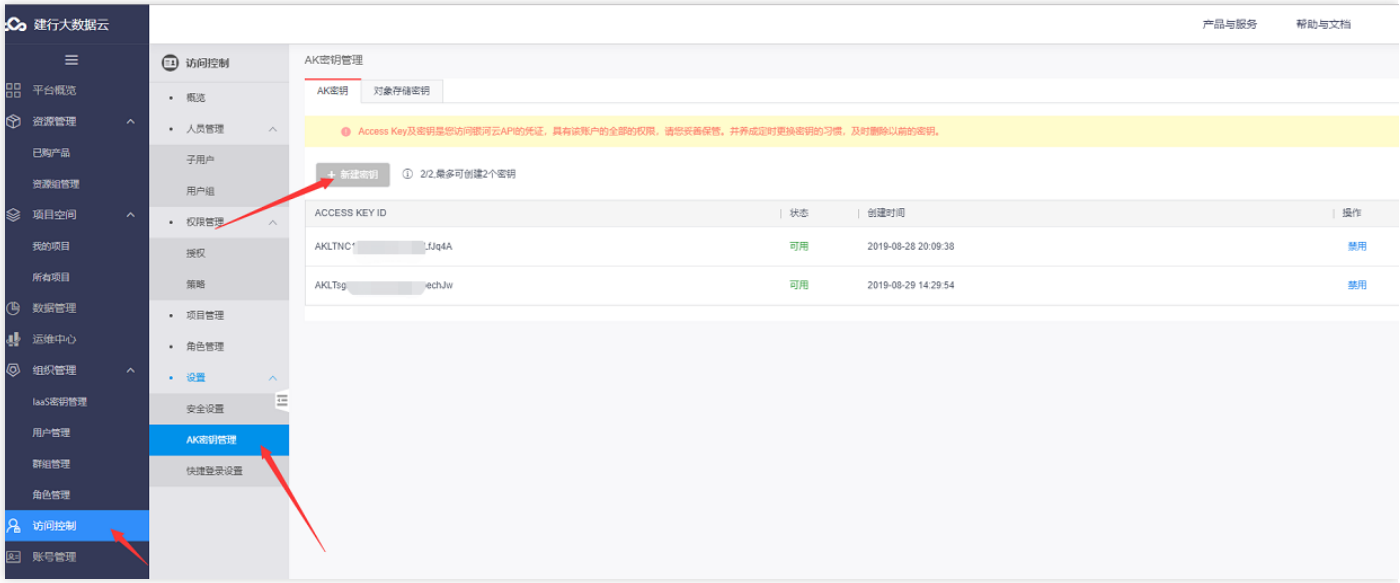
最近更新时间: 2019-11-12 03:04:33

支持同时采集多个文件，配置参数说明如下：

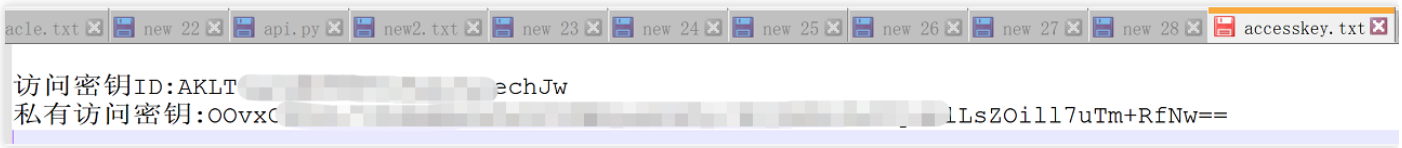
- 待采集文件：待采集的数据路径及文件名，点击页面上的“+”号，可同时采集同个文件
- AccessID：Access Key(AK)，此ID在“租户控制台-访问控制-AK密钥管理”中获取
- SK文件路径：即Secret Access key(SK)文件路径，需保证SK密钥在SK文件的首行

备注：AccessID获取方法如下：

1. 在“平台概览”–“访问控制”–“AK密钥管理”中新建密钥，



2. 创建完成之后，会提示下载SK秘钥到本地的弹窗，下载到本地后，文件里面的内容就是Secret Access key(SK)，对应访问秘钥和私有访问秘钥。



文件夹采集

最近更新时间: 2019-11-12 03:04:33

可以采集文件夹下的所有文件，配置参数说明如下：

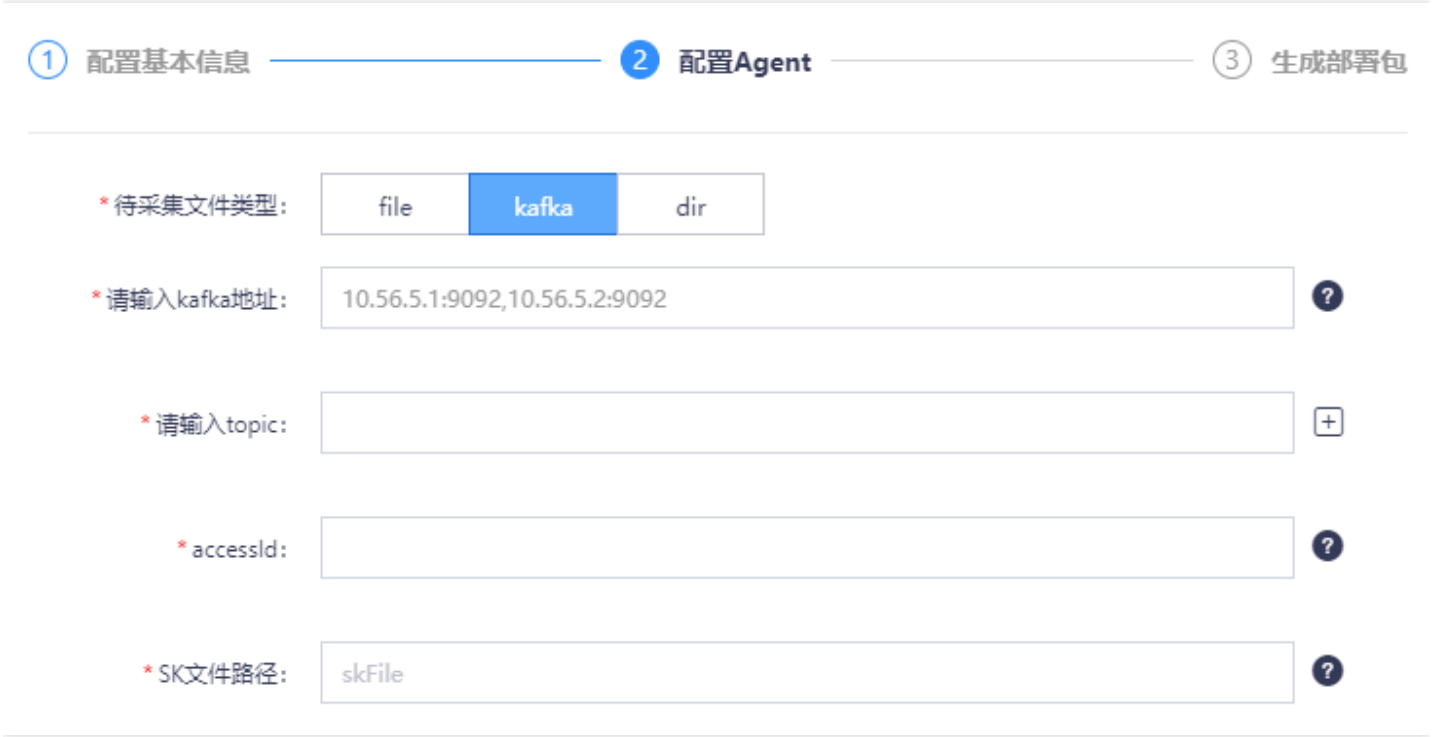
- 待采集文件夹：待采集的文件夹的完整路径
- AccessID：Access Key(AK)，此ID在“租户控制台-访问控制-AK密钥管理”中获取
- SK文件路径：即Secret Access key(SK)文件路径，需保证SK密钥在SK文件的首行

Kafka采集

最近更新时间: 2019-11-12 03:04:33

支持同时采集多个topic的数据，配置参数说明如下：

- Kafka地址：即Kafka的IP：端口号，多个时使用“，”分隔
- topic名称：支持采集多个topic，点击“+”新增
- AccessID：Access Key(AK)，此ID在“租户控制台-访问控制-AK密钥管理”中获取
- SK文件路径：即Secret Access key(SK)文件路径，需保证SK密钥在SK文件的首行



高级配置

高级配置

最近更新时间: 2019-11-12 02:56:56

高级选项，包括缓存配置和传输配置，一般情况下，使用默认配置即可，如用户有特殊需求，可以自行修改默认配置。下面对高级配置的各项属性进行说明。

缓存配置

最近更新时间: 2019-11-26 14:57:59

点击“类型选择”后面的下拉框，会弹出“memory”和“file”两个选项：

- Memory：表示agent的channel组件配置为MemoryChannel，此时agent采集的Event被缓存在内存中。
- File：表示agent的channel组件配置为File Channel，此时agent采集的Event被缓存在文件中。 根据类型选择的不同，有不同的缓存参数需要配置，下面具体说明。

1. 选择memory时，可以对“最大容量”和“事物容量”进行配置。下表对这两个配置项进行了说明：

配置项	配置项说明
最大容量	存储在channel中的event的最大数量
事务容量	从source中取得或者发送给sink时，单个事务中允许的event最大数量

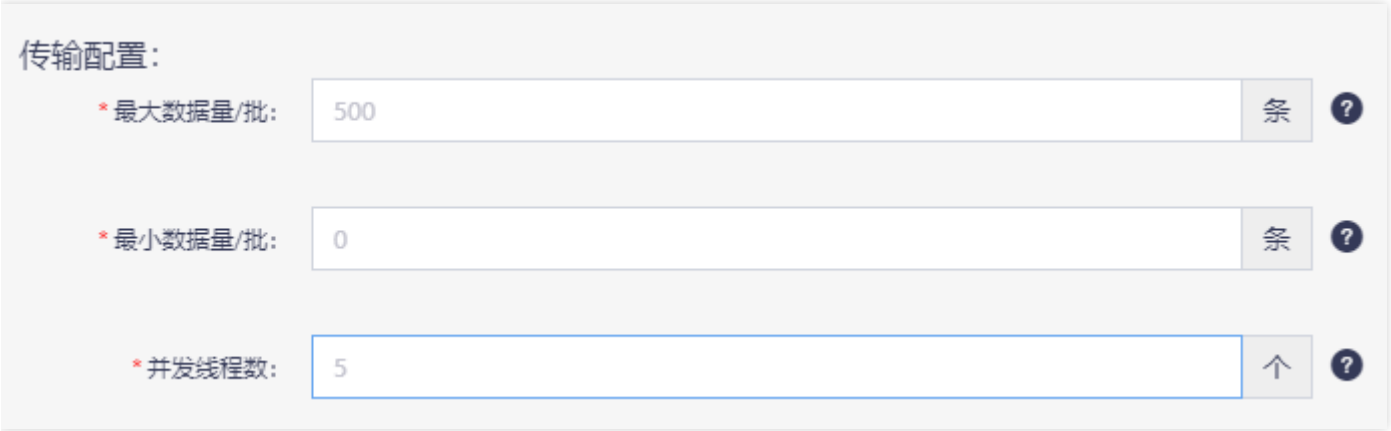
2.选择file时，可以对“最大容量”、“事务容量”、“checkpoint目录”和“缓存目录”进行配置。下表对这四个配置项进行说明：

配置项	配置项说明
最大容量	缓存在channel中的event的最大数量
事务容量	从source中取得或者发送给sink时，单个事务中的event最大数量
checkpoint目录	采集游标的存储目录，使得agent重启后仍可以从中断的位置开始采集任务
缓存目录	数据缓存在本地磁盘的目录，即File Channel的物理存储位置

传输配置

最近更新时间: 2019-11-12 02:57:02

缓存配置下方是传输配置。点击“传输配置”右侧的“展开”按钮，可以列出传输配置的配置项，如下图所示。



在传输配置中，可以设置agent上传流式数据时，每个批次的最大、最小数据量以及并发线程数量。下表对这三个配置项进行了说明：

配置项	配置项说明
最大数据量/批	传输数据按批次进行，该参数设置每个批次传输event的最大数量
最小数据量/批	每个批次传输event的最小数量
并发线程数	单个agent中sink组件的数量，每个sink组件对应一个传输数据线程

下载并启动agent

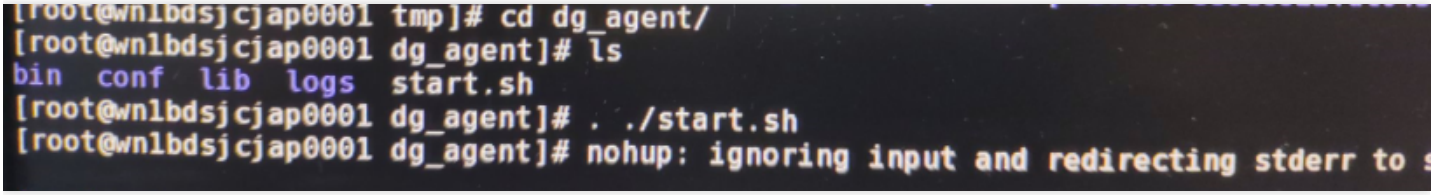
最近更新时间: 2019-11-12 02:57:02

完成agent配置后，点击下一步，即可生成agent部署包。

未采集任务创建结束的页面下载agent的，也可以在采集任务列表的：操作-下载agent处，下载采集任务的agent。

下载部署包后，部署到各个采集节点，就会开始采集流式数据，不同的部署包会将采集到的流式数据根据topic的属性发送到测试或生产环境的topic中。 将dg_agent.zip上传至客户端主机，进入解压缩后的dg_agent目录下，执行

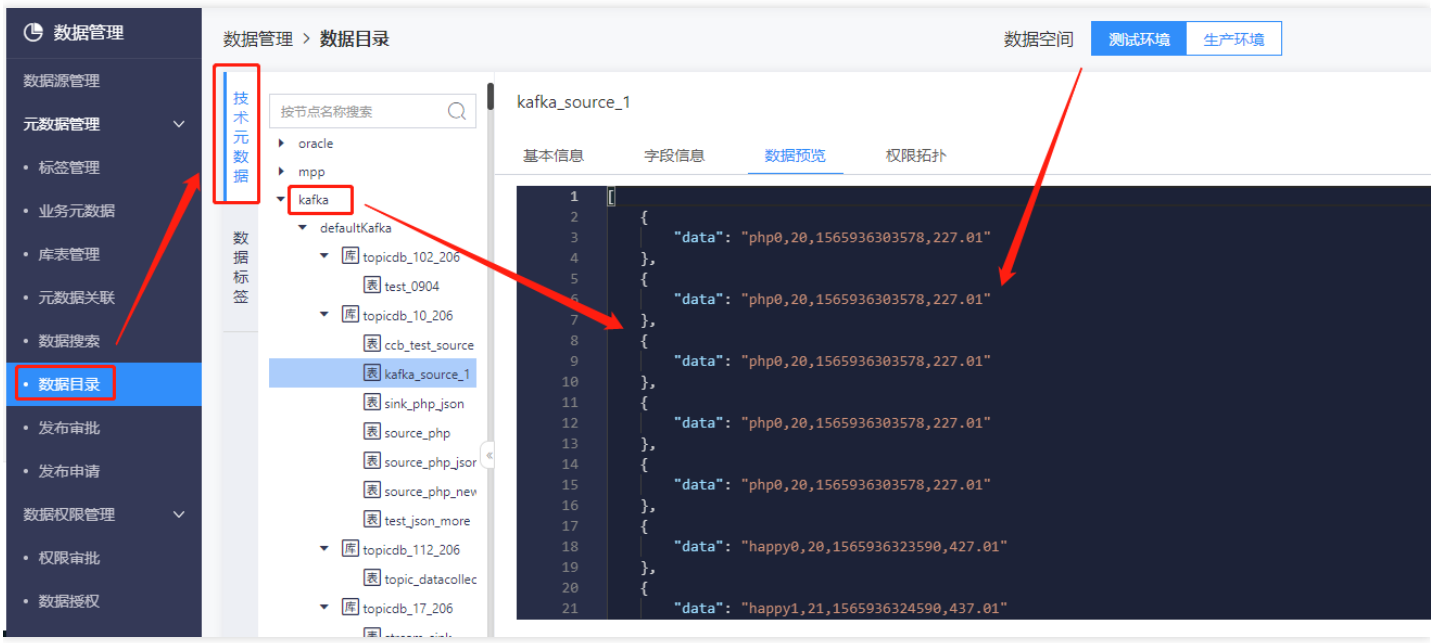
start.sh，即可在本地启动agent。



数据预览

最近更新时间: 2019-11-12 02:57:02

启动成功后，若采集任务执行正常，可在数据管理中进行数据预览（数据管理支持预览10条数据，可简单验证数据情况），验证任务执行是否成功。 点击“数据管理”→“元数据管理”→“数据目录”→“Kafka” →推送目标的具体topic，选择环境后，点击“数据预览”，可查看数据是否正常写入。 备注：待预览的Kafka和topic为创建topic时选择的数据类型和数据源。

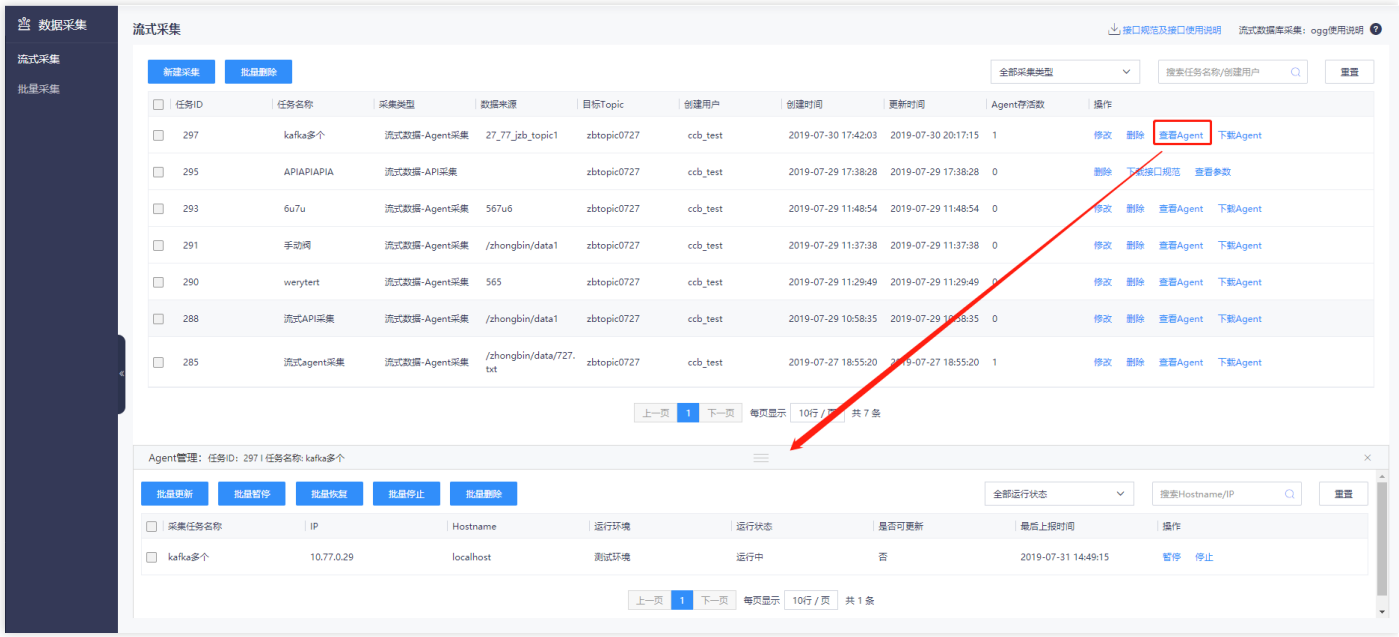


查看agent列表

最近更新时间: 2019-11-12 02:57:02

点击任务列表的“查看agent”，可查看每个agent的具体情况，并进行：暂停、恢复、停止、更新、删除等操作。

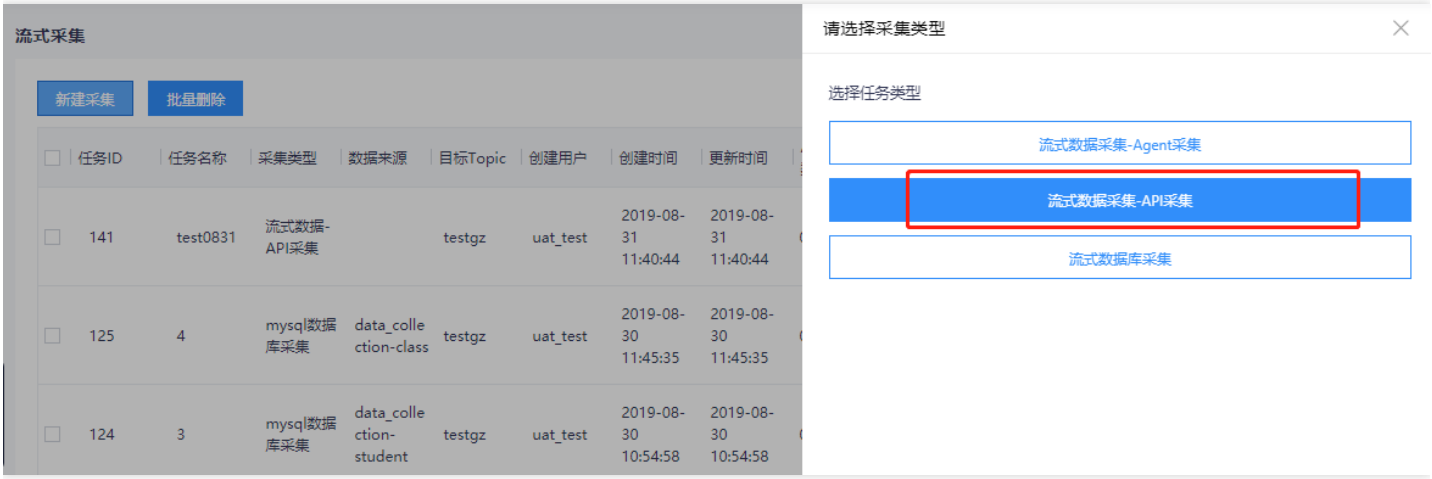
- 暂停/恢复：暂停后，采集任务暂时中断，可点击“恢复”重启采集任务
- 停止：停止后，页面无法重启任务，需通过agent重新启动
- 删除：任务停止后，可删除任务
- 更新：采集任务有更新时，可点击“更新”对agent配置文件进行更新（采集任务的agent信息有修改时，才会出现“更新”按钮并支持更新操作）



流式API采集
创建采集任务

最近更新时间: 2019-11-12 02:57:02

点击页面的【新建采集】按钮，在弹出的抽屉中，点击“流式数据采集-API采集”创建采集任务。

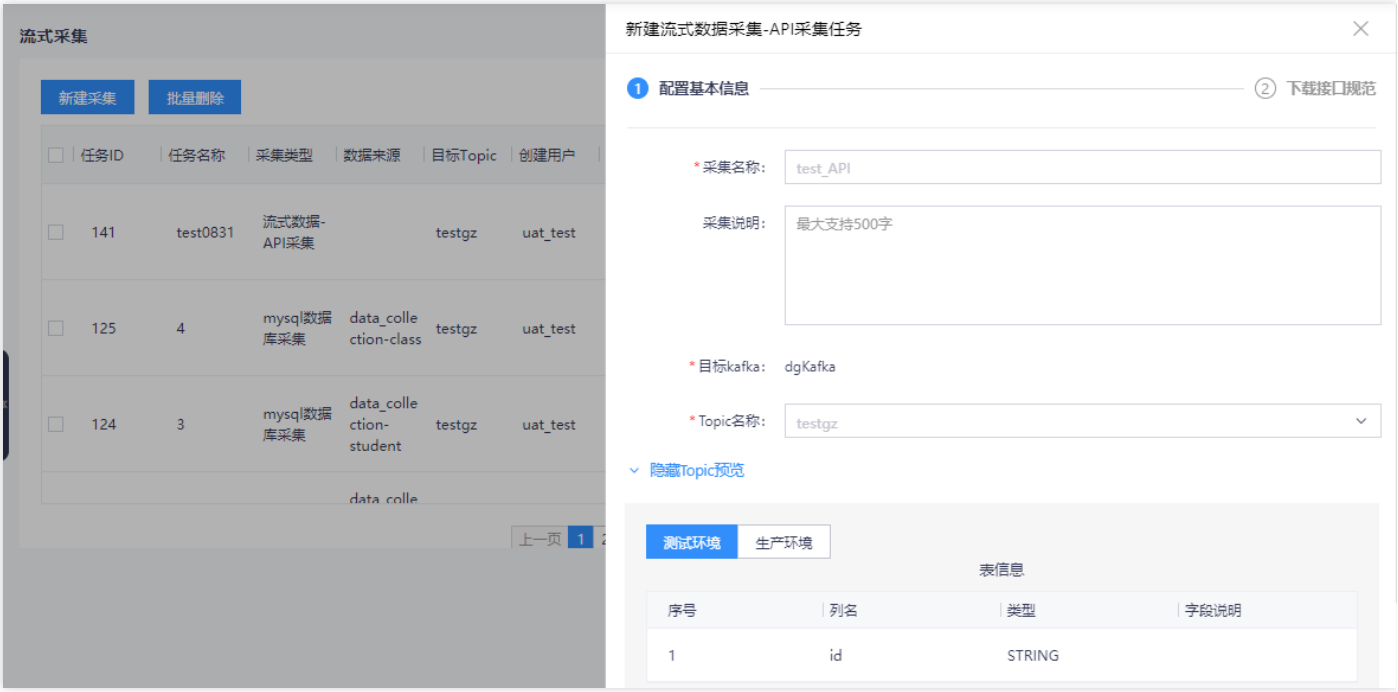


配置基本信息

最近更新时间: 2019-11-26 14:57:59

在弹出的窗口中填写新建采集任务的基本信息，必填参数说明如下：

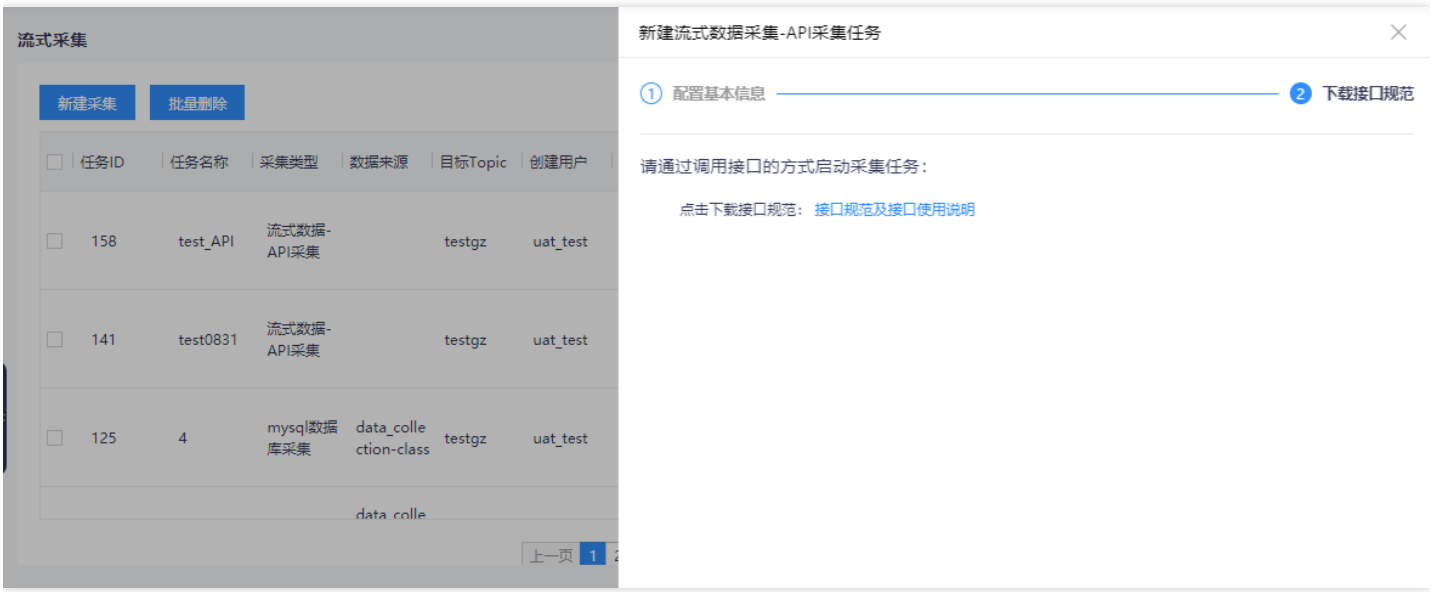
- 采集名称：支持中文、英文、数字、下划线，最大50字符
- topic名称：下拉选择，可选择该项目下有权限的所有topic
- 选择topic时，支持对topic字段信息进行预览，如下图所示 配置完毕后，点击“下一步”，完成采集任务创建。



下载接口规范

最近更新时间: 2019-11-12 02:57:02

API采集任务创建成功后，可在“下载接口规范”页面下载《接口规范及接口使用说明》。



通过API采集的任务，可参考上文《接口规范及接口使用说明》的内容通过API启动采集任务，API出入参信息可点击任务列表的“查看参数”获取，如下图所示：


新建采集
批量删除
API采集
搜索任务名称/创建用户
任务ID 任务名称 采集类型 数据来源 目标Topic 创建用户 创建时间 更新时间 Agent存活数 已采集数据量(KB) 操作
41 zhongbin_API_CSV 流式数据-API采集 topic_csv uat_test 2019-08-23 17:25:55 2019-08-23 17:25:55 0 0 删除 下载接口规范 查看参数
26 lbz_流式api采集 流式数据-API采集 lbz_test2 uat_test 2019-08-23 10:31:01 2019-08-23 10:31:01 0 0 删除 下载接口规范 查看参数

查看参数

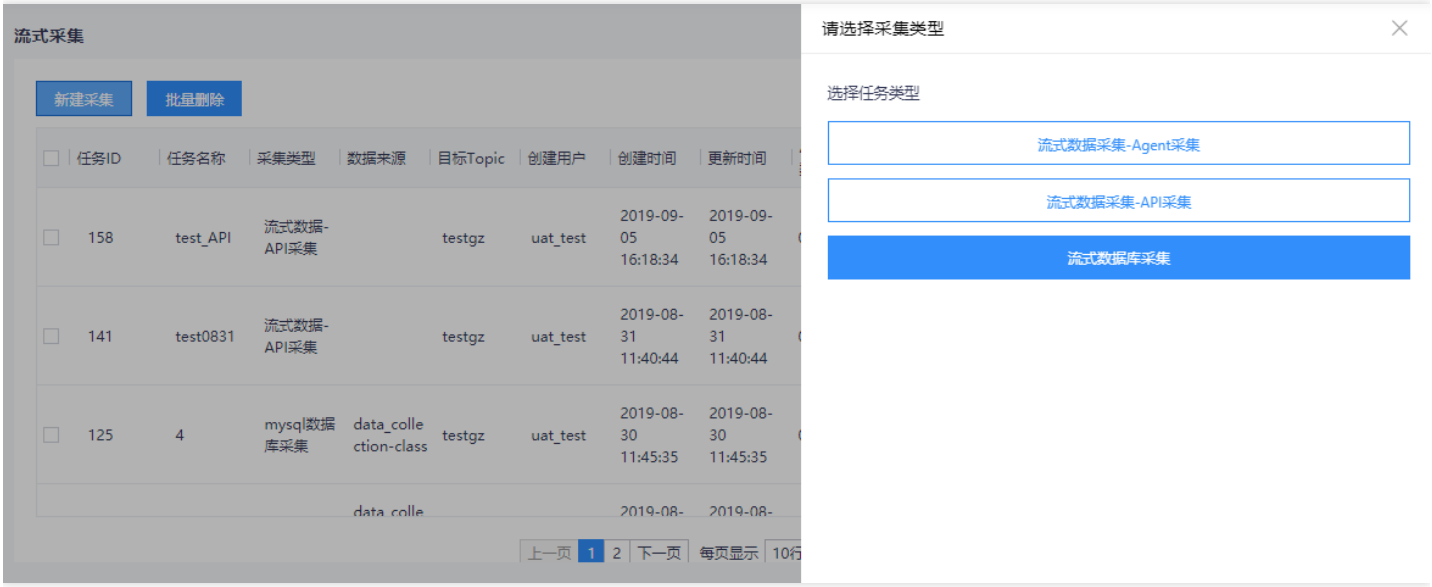
参数

参数名称	参数说明	参数值
endpoint_test	API请求地址(测试环境)	http://dg.bigdata.yun.ccb.com/dglogtest?Action=dglogtest&Version=v1&AccountId=111&tenantId=24
endpoint_online	API请求地址(生产环境)	http://dg.bigdata.yun.ccb.com/dglogonline?Action=dglogonline&Version=v1&AccountId=111&tenantId=24
tenantId	请求头参数，租户id	24
userId	请求头参数，用户id	24
env	请求头参数，kafka环境[test|online],若参数值只有test表示topic未发布到生产环境	test
projectId	请求头参数，采集任务项目id	6
ownerProjectId	请求头参数，topic所属项目id	6
kafkaSourceId	请求头参数，topic所在数据源id	218
topicName	请求头参数，数据存放的topic名称	topic_csv
requestMethod	请求头参数，请求标识固定不变	API
userData	请求体参数，采集的数据放在body，没有key	用户上报的数据

流式数据库采集
创建采集任务

最近更新时间: 2019-11-12 02:57:02

点击页面的【新建采集】按钮，在弹出的抽屉中，点击“流式数据库采集”创建采集任务。



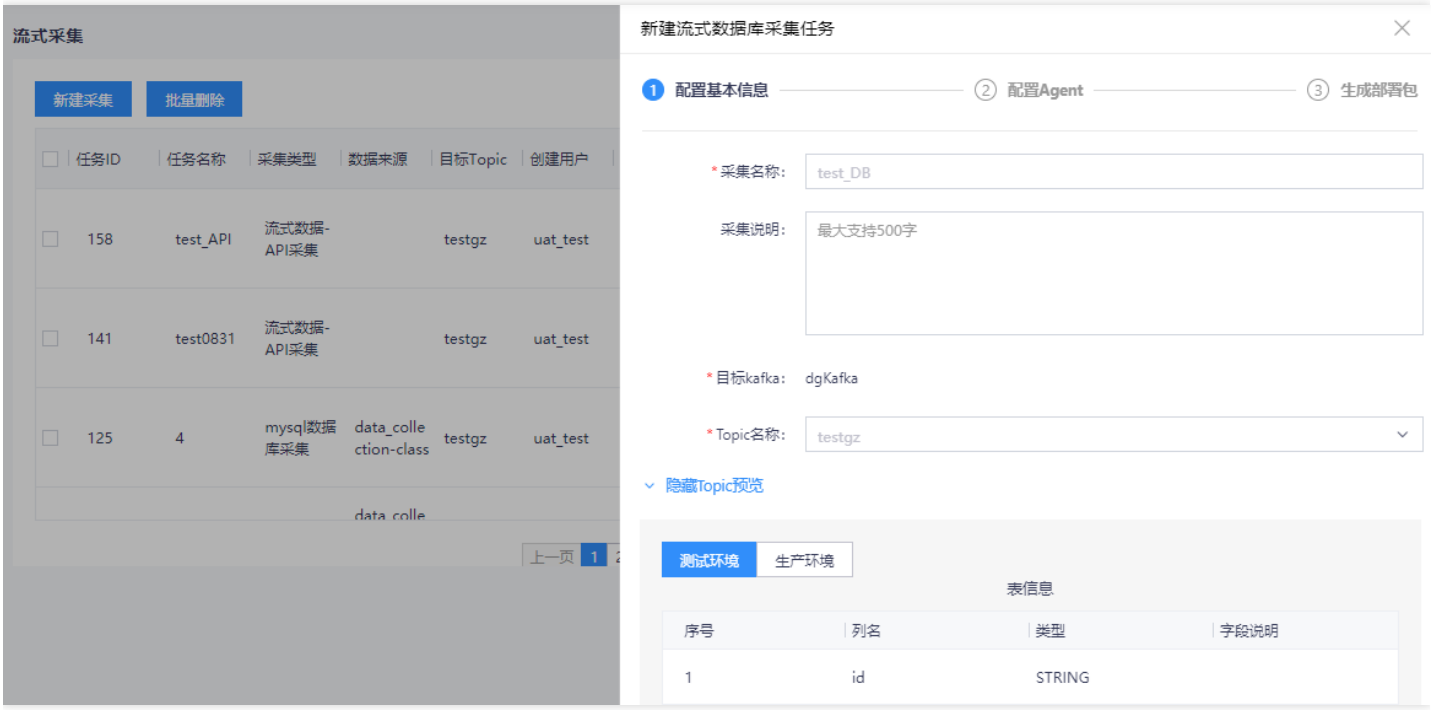
配置基本信息

最近更新时间: 2019-11-26 14:57:59

在弹出的窗口中填写新建采集任务的基本信息，必填参数说明如下：

- 采集名称：支持中文、英文、数字、下划线，最大50字符
- topic名称：下拉选择，可选择该项目下有权限的所以topic
- 选择topic时，支持对topic字段信息进行预览，如下图所示

配置完毕后，点击“下一步”，完成采集任务创建。

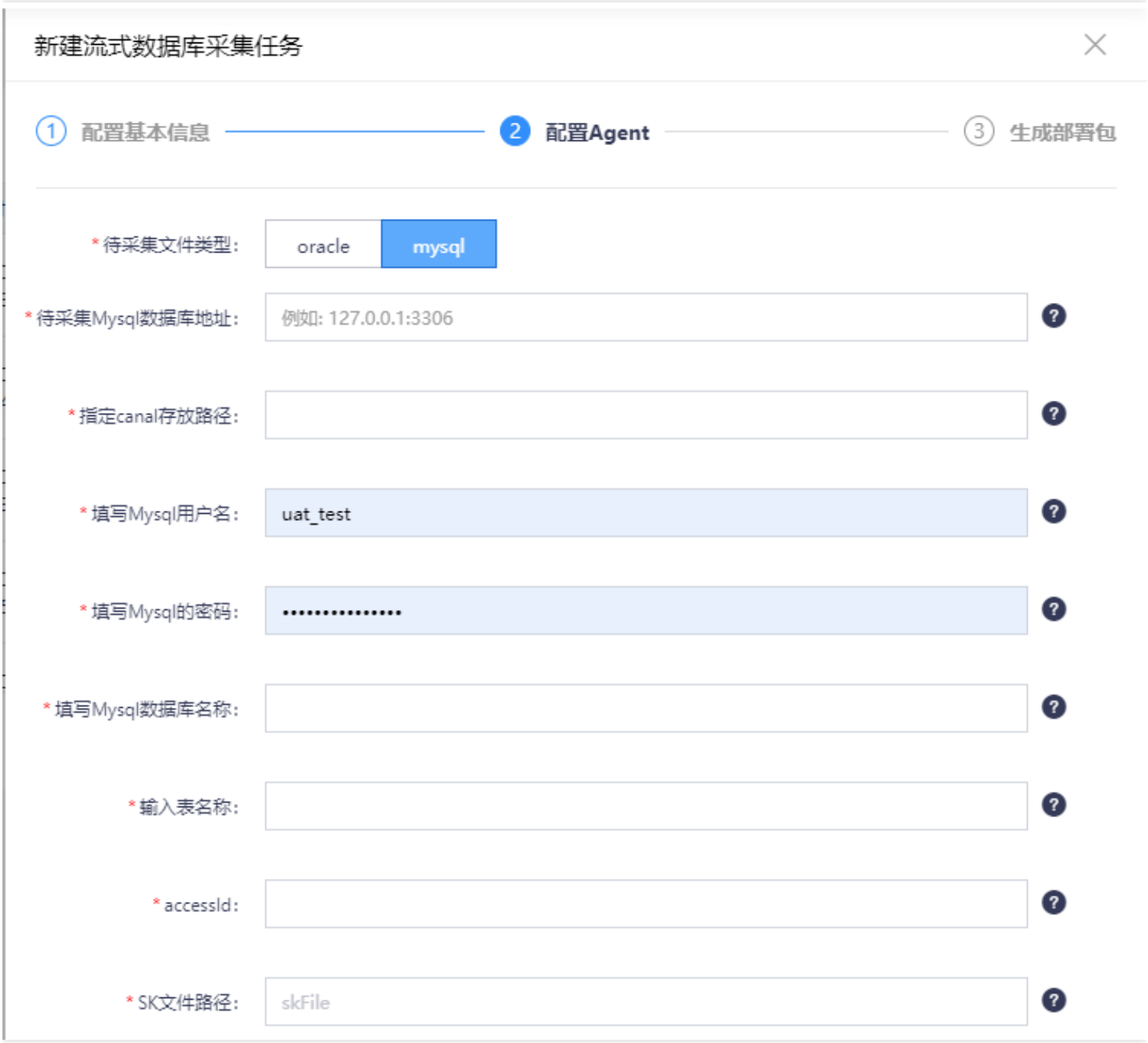


配置agent信息–流式数据库采集支持采集MySQL的数据。

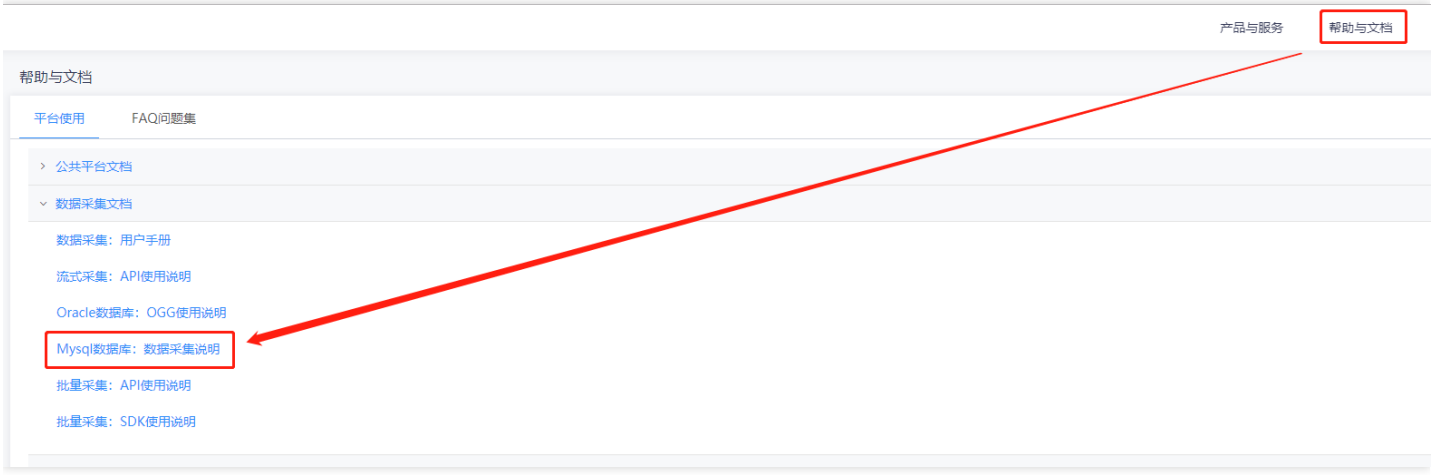
MySQL采集说明

最近更新时间: 2019–11–12 02:57:05

流式数据库MySQL采集使用的是Canal+Flume的方式采集数据，配置agent信息后，下载对应agent后，在本地部署启动。



备注：具体操作手册详见帮助与文档中心的《MySQL数据库：数据采集说明》



下载并部署agent

最近更新时间: 2019-11-12 02:57:05

MySQL数据库采集

Oracle数据库采集

特别说明

最近更新时间: 2019-11-12 02:57:05

MySQL数据库采集的投递目标topic，字段必须严格按照以下格式创建，创建方式详见“新建topic”下”设计态-新建topic”章节

```
{
"table": "TCLOUD.T_OGG2", //库名.表名
"op_type": "U", //操作类型 U更新 D删除 I插入
"current_ts": "2018-05-31T14:49:01.709000", //【处理时间】
"pos": "00000000000000003770", //偏移量
"before": { //object类型，操作前的字段
"ID": 1, //业务字段
"AGE": 20,
"IDD": "1"
},
"after": { // object类型，操作后的字段
"ID": 1,
"AGE": 1,
"IDD": "1"
}
}
```

Oracle数据库采集的投递目标topic，字段必须严格按照以下格式创建，创建方式详见“新建topic”下”设计态-新建topic”章节

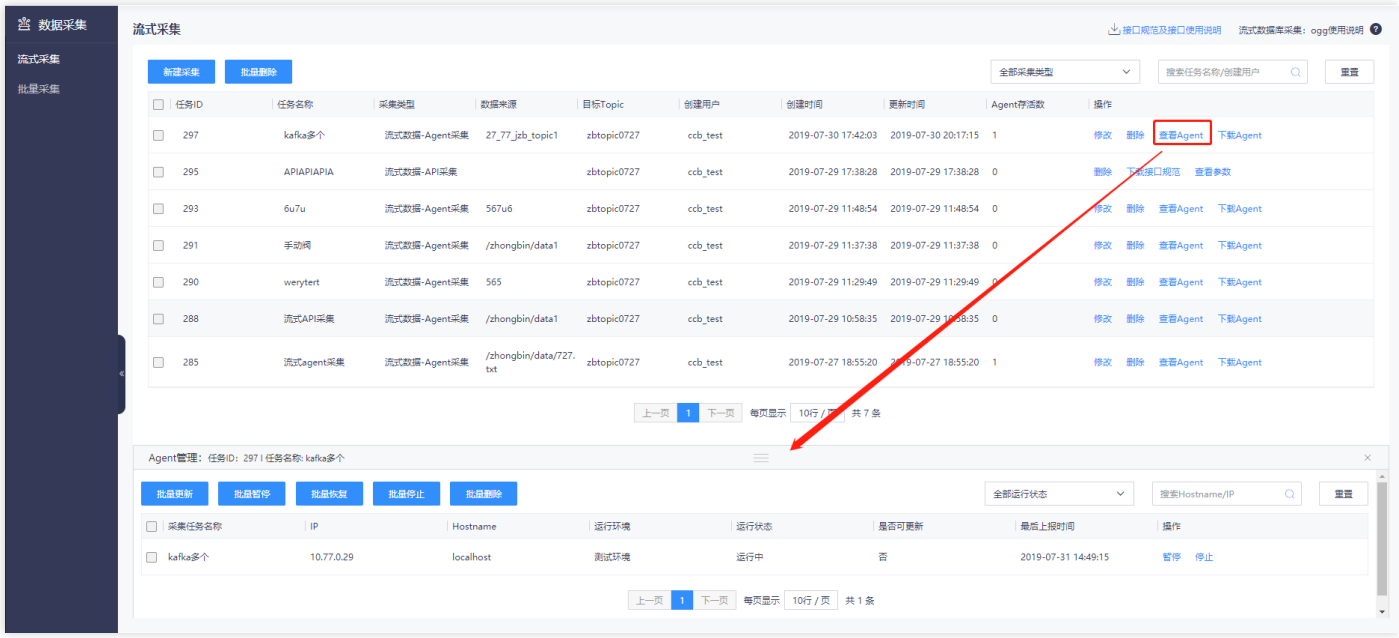
```
{
"table": "TCLOUD.T_OGG2", //库名.表名
"op_type": "U", //操作类型 U更新 D删除 I插入
"op_ts": "2018-05-31 14:48:55.630340", //操作时间
"current_ts": "2018-05-31T14:49:01.709000", //【处理时间】
"pos": "00000000000000003770", //偏移量
"before": { //object类型，操作前的字段
"ID": 1, //业务字段
"AGE": 20,
"IDD": "1"
},
"after": { // object类型，操作后的字段
"ID": 1,
"AGE": 1,
"IDD": "1"
}
}
```

查看agent列表

最近更新时间: 2019-11-26 14:57:59

点击任务列表的“查看agent”，可查看每个agent的具体情况，并进行：暂停、恢复、停止、更新、删除等操作。

- 暂停/恢复：暂停后，采集任务暂时中断，可点击“恢复”重启采集任务
- 停止：停止后，页面无法重启任务，需通过agent重新启动
- 删除：任务停止后，可删除任务
- 更新：采集任务有更新时，可点击“更新”对agent配置文件进行更新（MySQL数据库采集，不支持agent更新）

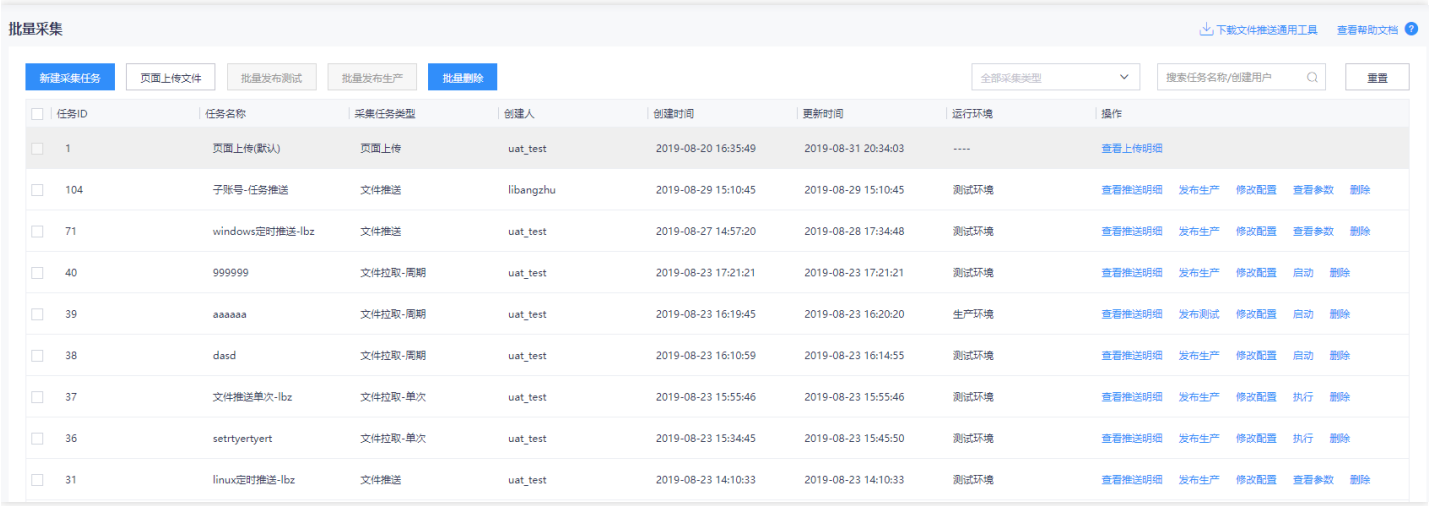


批量采集配置

批量采集配置

最近更新时间: 2019-11-12 02:39:16

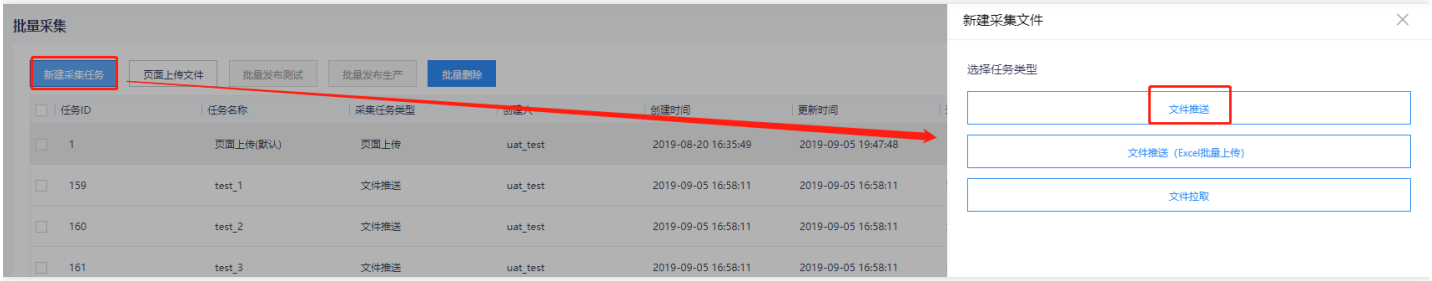
批量采集可将外部数据（支持结构化、半结构化、非结构化数据）批量推送至COS中。采集方式支持：文件推送、文件拉取、页面文件上传等方式。



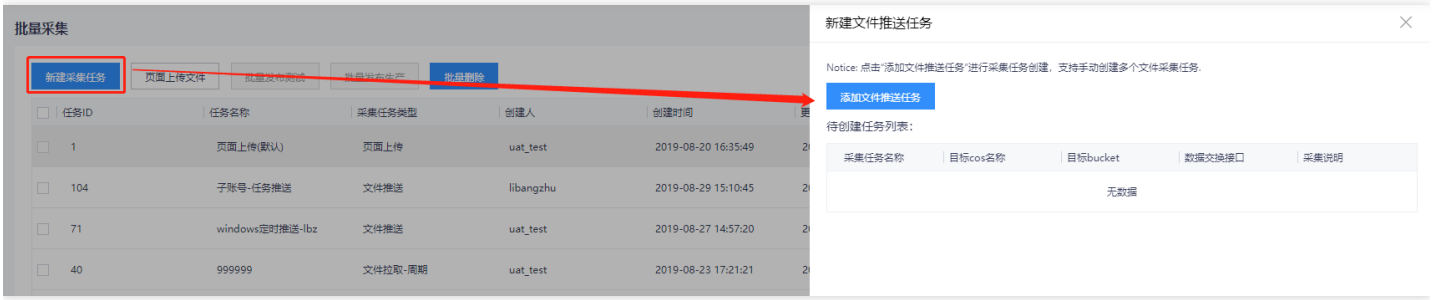
文件推送任务配置

最近更新时间: 2019-11-26 14:57:59

点击【批量采集】按钮，在弹出的窗口中选择“文件推送”，进入文件推送任务编辑页面。



在新弹出的窗口中点击“添加文件推送任务”进行采集任务创建，支持手动创建多个文件采集任务。



创建采集任务时，需要填写：

- 采集名称：支持中文、英文、数字、下划线，最大50字符
- 目标COS名称：下拉选择，需要在数据管理中预先创建
- 目标Bucket：下拉选择，需要在数据管理中预先创建，可选择项目下有权限的Bucket
- 数据交换接口：下拉选择，需要在数据管理中预先创建，可选择项目下有权限的数据交换接口

- 选择数据交换接口后，可点击下方“数据交换接口预览”查看数据交换接口的schema信息

配置完毕后，点击“下一步”，完成本条采集任务的创建，重复以上步骤，可一次性创建多个采集任务。

完成采集任务创建后，可点击下载文件上传工具包、或API、或SDK，在用户的客户端启动文件推送任务。

文件推送（Excel批量上传）配置

最近更新时间: 2019-11-12 02:34:35

在页面下载Excel模板，汇总待新增的任务信息，并按照规范填写：采集任务名称、目标COS名称、目标Bucket、数据交换接口和采集说明后，将Excel上传至大数据云平台，进行批量创建操作。

按照模板填写并上传Excel后，Excel中的内容会在页面预览出来，确认无误后，点击下一步完成批量创建。

运行批量采集任务

最近更新时间: 2019-11-12 02:32:33

批量采集任务可以通过：① 文件上传工具包，② API接口，③ SDK启动。关于文件上传工具包、API和SDK的的具体使用方法，可通过2种途径查看下载：

- 方式一：可在“批量采集 通用下载”页面下载

- 方式二：可在“帮助与文档”的跳转页面查看/下载

1、批量采集 linux工具演示demo详见下图

```
drwxr-xr-x 3 root root    4096 Jul 22 17:41 data2
-rwxr-xr-x 1 root root      63 Jul 16 17:04 file1.csv
-rwxr-xr-x 1 root root     360 Jul 16 17:05 file2.csv
-rwxr-xr-x 1 root root 7907300 Jul 16 16:47 uploadFileClient
[root@vm10-77-0-29 zhongbin]# vim a.txt
[root@vm10-77-0-29 zhongbin]# ./uploadFileClient  -id 305 -s file1.csv,file2.csv -t zhongbin/aaa
id: 305
sourcePath: file1.csv,file2.csv
targetPath: zhongbin/aaa
zhongbin/aaa/file1.csv
zhongbin/aaa/file2.csv
共上传 2 个文件,耗时 0 分 0 秒
[root@vm10-77-0-29 zhongbin]#
```

结果验证如下：

2、批量采集：SDK采集demo可参考下图

```java
public class CosDemo {
    public static void main(String[] args) throws FileNotFoundException {
        COSClient cosClient = UploadFileUtil.initParams("myqcloud.com","AKIDLVTgXKCWAze9m3afGnBTaYIgrongwtEP",
                "AAzvjNsoZ9XkpmZEVcYIrHZghLo7FCxd","ap-beijing",
                "buckettest2-test-1259417666");
        PutObjectResult putObjectResult = UploadFileUtil.putFileToCos(cosClient, "/zhongbin/log.txt", "d:/log.txt");
        System.out.println(putObjectResult.getETag());
        UploadFileUtil.closeCOSClient(cosClient);
    }
}
```

具体参数信息可在大数据云平台，点击【查看参数】获取，参见下图

查看参数

参数

参数名称	参数说明	参数值
fileServerAddress	通知调度系统接口服务器地址	http://dg.bigdata.yun.ccb.com
tenantId	租户id	24
appId	文件推送任务id	160
bucketName	当前任务对应的bucket	uat-bucket37-test-1259417666
dataInterfaceName	当前任务对应的数据交换接口名称	breast_cancer03920050804241536
Host	API方式上传文件，需要主机地址	uat-bucket37-test-1259417666.cos.ap-beijing.myqcloud.com

文件拉取

最近更新时间: 2019-11-12 02:20:08

1. 点击文件拉取按钮，进入文件拉取配置页面，文件拉取支持从FTP拉取数据投递至COS。

2. 配置拉取信息前，需先输入FTP的IP、端口号、用户名、密码信息，并验证其连通性，连通性验证通过后，点击下一步，进行具体拉取配置。

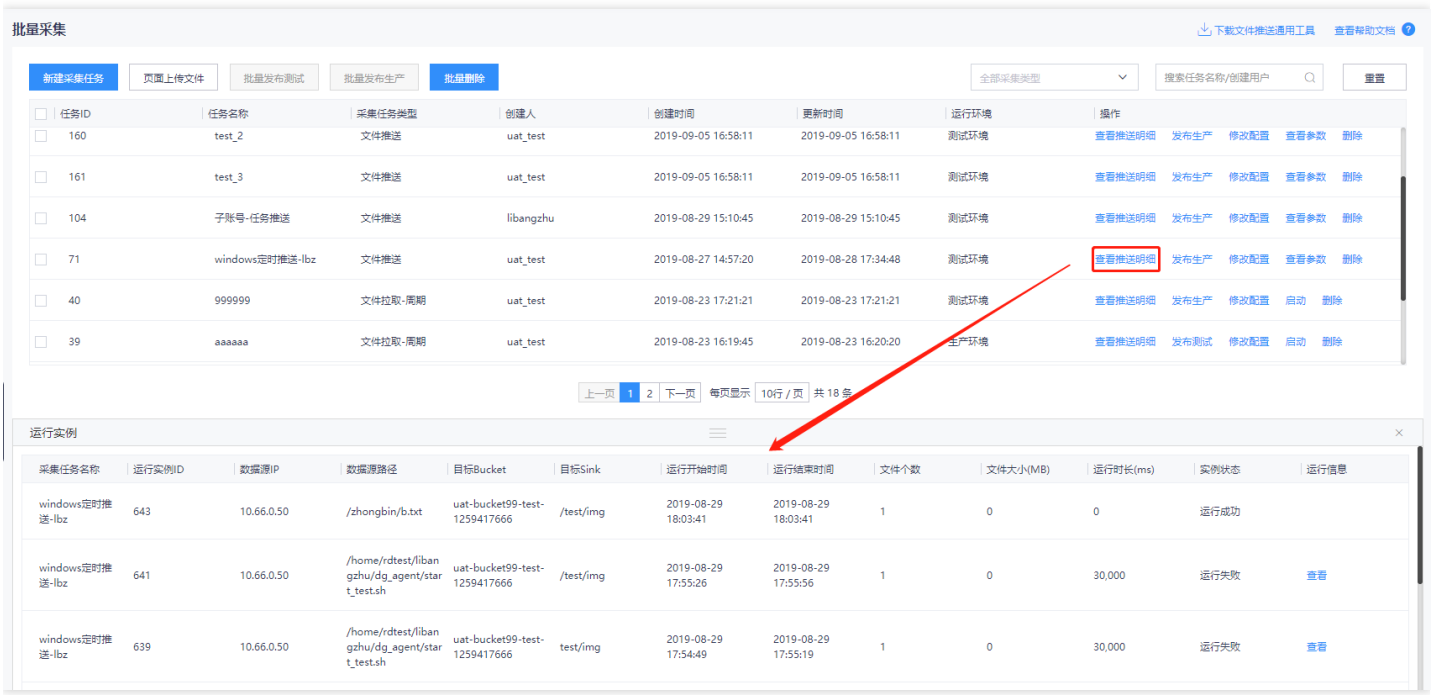
3. 文件拉取支持“周期执行”和“单次执行”2种方式，两种方式均需指定推送的目标COS、Bucket、数据交换接口和具体推送路径。其中，周期执行的任务，需要额外配置“首次拉取时间”和“拉取周期”。具体如下图所示：

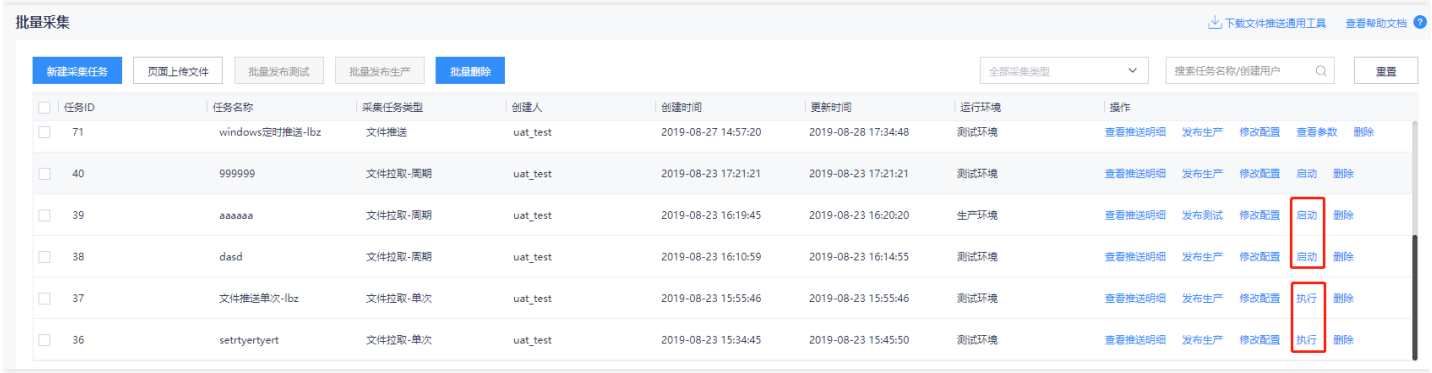

修改文件拉取任务
*拉取方式:
周期执行
单次执行
*拉取的文件路径:
/data/ftp/filename_${date}_${num}.txt
*目标cos名称：
cos
*目标Bucket：
uat-bucket99
*选择数据交换接口：
zb_exchange
*目标路径:
test/ftp
*首次拉取时间:
2019-08-23 17:22:00.0
*拉取周期:
120
秒
显示数据交换接口预览

查看批量采集运行实例

最近更新时间: 2019-11-12 02:24:13

点击【查看推送明细】按钮，会弹出表单，查看批量采集的运行实例。

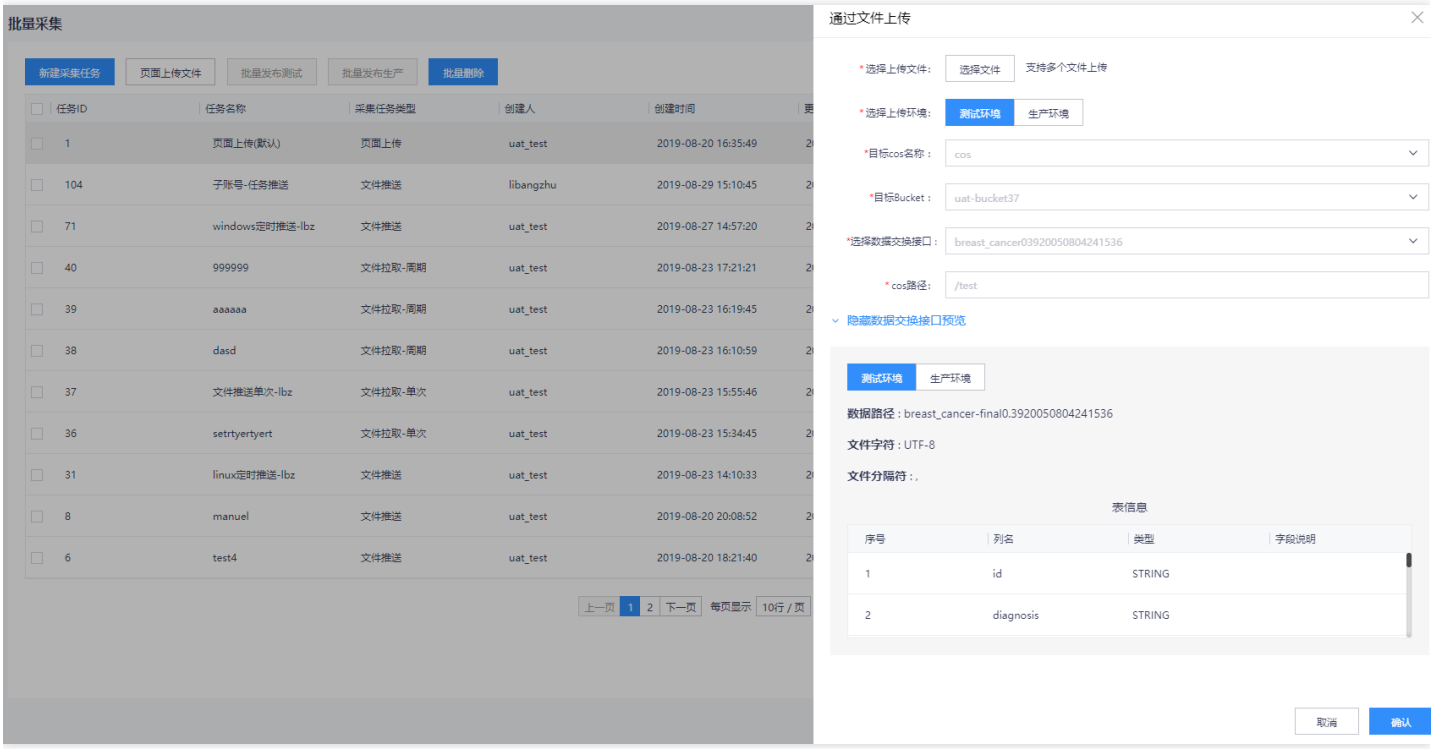




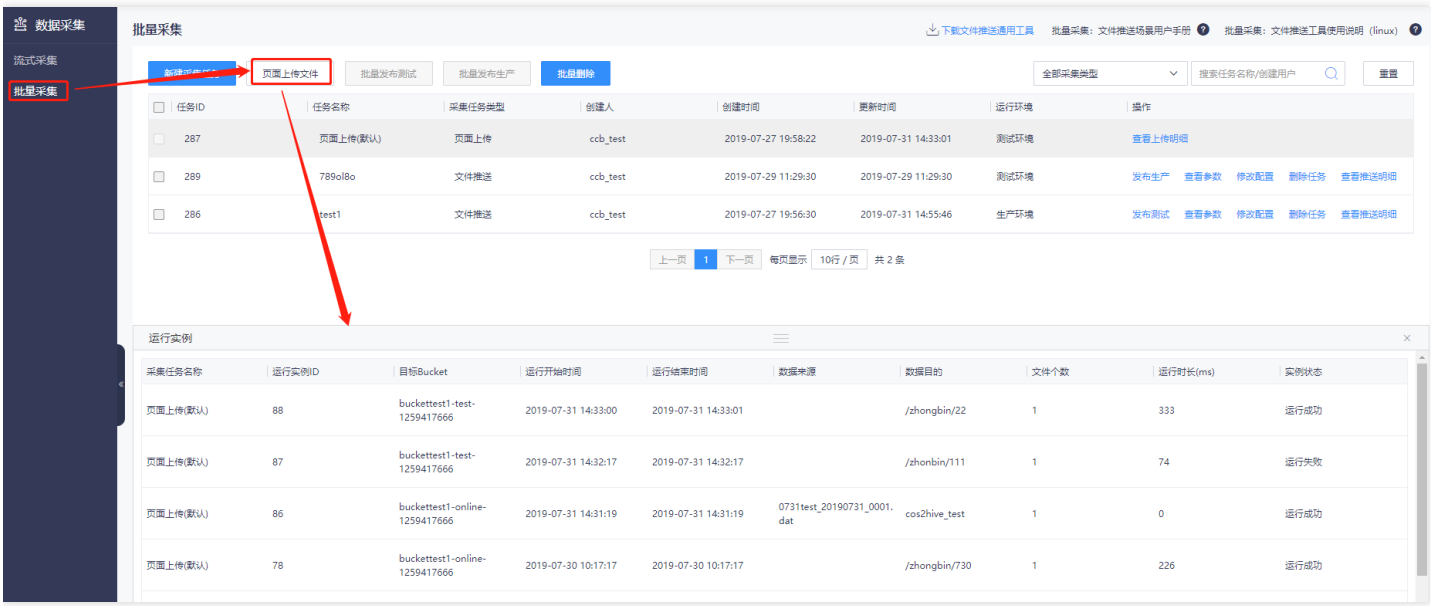
通过页面进行文件上传

最近更新时间: 2019-11-12 02:00:13

除以上方式外，对于文件数较少的临时数据采集需求，还可以通过：批量采集-页面上传文件功能，进行文件的上传。点击“页面上传文件”按钮后，选择待上传的文件（可支持多个）和上传的目标地址（COS）即可。



页面文件上传的任务默认展示在任务列表的首行，点击“操作”中的“查看上传明细”，可查看每次文件上传的信息。



最佳实践

最近更新时间: 2019-11-12 02:38:33

1） 关于批量采集大批量采集任务创建 建议使用“文件推送（Excel批量上传）”功能，在页面下载Excel模板，汇总待新增的任务信息，并按照规范填写：采集任务名称、目标COS名称、目标Bucket、数据交换接口和采集说明后，将Excel上传至大数据云平台，进行批量创建操作。 2） 关于采集任务运行情况验证

- 针对流式采集：① 可根据采集任务列表的“agent存活数”“”直观查看是否有agent运行成功；② 可点击查看agent管理列表，查看具体客户端的IP、hostname等信息； ③ 可进入“数据管理”→“元数据管理”→“数据目录”→“Kafka”→推送目标的具体topic，选择环境后，点击“数据预览”，查看数据是否正常写入。
- 针对批量采集：①可点击查看文件推送明细，查看具体的推送明细，包括运行开始时间、结束时间、运行状态，推送文件个数、文件大小等； ③ 可进入“数据管理”→“元数据管理”→“数据目录”→“COS”→推送目标的具体Bucket/数据交换接口，选择环境后，点击“数据预览”，查看数据是否正常写入。

3） 关于流式采集agent配置文件升级 为保障流式采集任务的稳定运行，对采集任务的agent配置进行编辑后，不会自动对agent的配置文件进行升级，需用户进入“agent管理”列表，暂停agent后，选择逐个/批量“更新”操作，完成指定客户端的agent升级，完成升级后，点击“恢复”/“批量恢复”可以重新启动采集动作。

常见问题

最近更新时间: 2019-11-12 02:38:33

1) 为什么新建流式数据采集任务时，topic名称处下拉框为空?

当没有为新建的项目创建topic，或者设计态的topic没有发布到测试时，直接新建流式采集的任务就会出现这种现象。新建流式数据采集任务前，请首先为当前项目创建至少一个topic，并至少将topic发布到测试。 2) 新建文件推送任务，数据交换接口下拉框处为空?

当没有为新建的项目创建数据交换接口，或者设计态的数据交换接口没有发布到测试时，直接新建文件推送任务就

会出现这种现象。新建文件推送任务前，请首先为当前项目创建至少一个数据交换接口，并至少将该接口发布到测试。 3) 为什么创建采集任务失败？ 可能原因：topic无insert权限 或 topic不存在 验证方法：在“数据管理”–“数据搜索”菜单下查看topic 权限详情，检查当前项目是否有topic的insert权限，或topic是否被删除 解决办法：

- 若topic无insert权限，在数据管理给topic添加insert权限
- 若topic不存在或topic被删除，在数据管理新建topic
- 创建topic或修复权限问题后，返回采集页面重新创建任务

4) 为什么文件推送任务发布到生产环境失败？ 可能原因：目标数据交换接口未发布到生产环境，或项目没有insert权限 验证方法：

- 在“数据管理”–“库表管理”中查看数据交换接口是否发布生产
- 在“数据管理”–“数据搜索”菜单下查看topic 权限详情，检查当前项目是否有topic的insert权限 解决办法：
- 若项目无insert权限，在数据管理中添加insert权限
- 若数据交换接口未发布生产环境，先将其发布至生产环境
- 以上问题验证修复后，返回采集页面重新进行发布生产的操作

词汇表

最近更新时间: 2019-11-26 14:57:59

名词	描述
流式数据	实时、不间断产生的数据流，如业务日志、系统日志等各类日志信息。单条日志是流式数据采集和传输的基本单位。
Flume	Flume是一个分布式、可靠、和高可用的海量日志采集、聚合和传输的系统。支持在日志系统中定制各类数据发送方，用于收集数据;同时，Flume提供对数据进行简单处理，并写到各种数据接受方(比如文本、HDFS、Hbase等)的能力 。
agent	Flume 运行的核心是 agent。Flume以agent为最小的独立运行单位。一个agent就是一个JVM。它是一个完整的数据收集工具，含有三个核心组件，分别是Source、 Channel、 Sink。通过这些组件，Event 可以从一个地方流向另一个地方
Kafka	Kafka是一种高吞吐量的分布式发布订阅消息系统，有如下特性： 通过O(1)的磁盘数据结构提供消息的持久化，这种结构对于即使数以TB的消息存储也能够保持长时间的稳定性能。 高吞吐量：即使是非常普通的硬件，Kafka也可以支持每秒数百万的消息。 支持通过Kafka服务器和消费机集群来分区消息。 支持Hadoop并行数据加载
topic	topic是Kafka对一组消息的归纳。在大数据云服务中，一个流式数据采集服务对应一个topic，单个topic可以存储一个或多个日志中的流式数据。
OGG	OGG 即Oracle Golden Gate，是一种基于日志的结构化数据复制软件。OGG 能够实现大量交易数据的实时捕捉，变换和投递，实现源数据库与目标数据库的数据同步，保持最少10ms的数据延迟
Canal	Canal是通过模拟成为MySQL 的slave的方式，监听MySQL 的binlog日志来获取数据，binlog设置为row模式以后，不仅能获取到执行的每一个增删改的脚本，同时还能获取到修改前和修改后的数据，基于这个特性，Canal就能高性能的获取到MySQL数据数据的变更。
批量数据	在大数据云服务中专指数据文件。
数据交换接口	一个批量采集的服务对应一个数据交换接口，在数据集成中，采集到的数据文件将通过数据交换接口中定义的字段映射成数据库表中的数据。