



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。

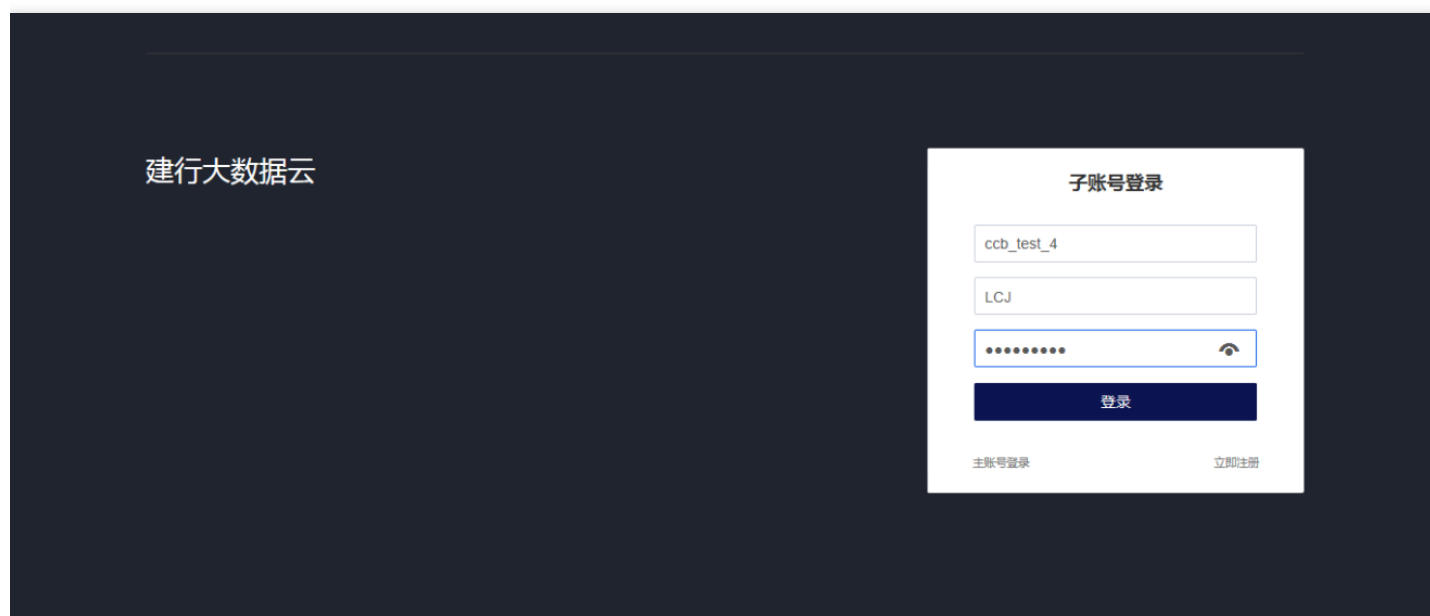
二、批量采集文件推送 为批处理提供数据源，客户将业务数据积攒整理成数据文件后，再批量推送上云进行处理。批量采集对应于先存储后计算，对实时性要求不高，同时数据的准确性、全面性更为重要的应用场景。

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】按钮，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。

产品名称	产品分类	资源配额(剩余/总量)	购买时间	所在项目数	状态	操作
API网关	大数据管理服务产品		2019-08-16 15:42:29	0	● 正常可用	立即使用
数据管理	大数据管理服务产品	版本:	2019-08-16 15:42:29	0	● 不可用	立即使用
MPP云数仓	大数据存储计算产品	硬盘:0/0 GB CPU:0/0 core 内存:0/0 GB	2019-08-16 15:42:29	0	● 不可用	立即使用
图分析平台	大数据分析应用产品	用户数:5000/0 人 Kafka_节点数:3/0 个 图数据库_节点数:0/0 个 ES_节点数:0/0 个 HBase_存储:0/0 GB	2019-08-16 15:42:29	0	● 正常可用	立即使用
托管Hadoop	大数据存储计算产品	硬盘:5000/5000 GB CPU:500/500 core 内存:500/500 GB	2019-08-16 15:42:29	0	● 正常可用	立即使用
数据采集	大数据开发套件产品	DCU:8/8 个 DIU:0/0 个	2019-08-16 15:42:29	54	● 正常可用	立即使用 资源配置
数据集成	大数据开发套件产品	CU:48/48 个 DCU:60/60 个	2019-08-16 15:42:29	54	● 正常可用	立即使用 资源配置
离线计算	大数据开发套件产品	CU:32/32 个 DCU:40/40 个	2019-08-16 15:42:29	54	● 正常可用	立即使用 资源配置
流计算	大数据开发套件产品	CU:12/12 个	2019-08-16 15:42:29	50	● 正常可用	立即使用 资源配置

产品与服务

平台产品 应用场景 案例实践

大数据分析应用产品

可视化BI
可视化BI
可视化BI是一种基于云端的可视化商业智能分析服务, 提供敏捷、高效、协作的企业级分析能力, 支持丰富...

产品详情 立即使用

图分析平台
图分析平台
图分析平台是一种提供知识图谱应用分析的企业级分析平台, 提供大规模数据检索与复杂推理能力, 以实体...

产品详情 立即使用

大数据开发套件产品

图计算
图计算
图计算开发是运行在spark之上开发平台, 提供了丰富的图基础算子和业务算子, 支持拖拽方式进行图数...

产品详情 立即使用

数据采集
数据采集
数据采集是一种面向开发者提供的端到端数据采集服务, 是数据进入大数据平台的第一道关卡, 支持日志文...

产品详情 立即使用

数据集成
数据集成
建行大数据云数据集成服务支持金融级强监管要求下的数据集成功能, 开通后您可以轻松实现数据在不同系...

产品详情 立即使用

离线计算
离线计算
离线计算是一种经济并高效的分析和处理海量数据的开发平台, 可提供快速、完全托管的PB级数据海量解决...

产品详情 立即使用

流计算
流计算
建行大数据云流计算服务提供面向高速流式数据进行实时快速计算的解决方案, 开通后可满足您实时风控、...

产品详情 立即使用

数据挖掘
数据挖掘
建行大数据云数据挖掘服务提供可视化建模和Notebook建模能力, 内含多种数据挖掘算法, 开通后您可以...

产品详情 立即使用

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。

项目名称	创建时间	负责人	开通的服务	最近更新时间	状态
fuwu	2019-09-05 01:59:16	uat_test	流计算 数据采集 数据服务	2019-09-05 01:59:16	● 正常
Project_1	2019-09-04 21:22:16	uat_test	流计算 数据采集 数据集成 离线计算	2019-09-04 21:22:16	● 正常
wer	2019-09-04 17:08:38	uat_test	数据采集 数据集成 图计算	2019-09-04 17:08:38	● 正常
111	2019-09-04 14:17:09	uat_test	图计算	2019-09-04 14:17:09	● 正常
test11111	2019-09-04 09:54:22	uat_test	流计算	2019-09-04 09:54:22	● 正常
test	2019-09-03 21:29:26	uat_test	流计算	2019-09-03 21:29:26	● 正常
my_stream_test_proj	2019-09-03 17:55:08	uat_test	流计算 数据采集 数据集成 离线计算 数据挖掘 数据服务	2019-09-03 18:00:46	● 正常
atk_proj08	2019-09-02 13:14:08	uat_test	流计算 数据采集 数据集成 离线计算 数据挖掘 数据服务	2019-09-02 13:14:08	● 正常
jhh	2019-08-31 00:25:41	uat_test	流计算 数据采集 数据集成 离线计算 数据挖掘 数据服务	2019-08-31 00:25:41	● 正常
eevre	2019-08-30 19:53:54	uat_test	流计算 数据采集 数据挖掘	2019-08-30 19:53:54	● 正常

点击蓝色的“新建项目”按钮，弹出“新建项目”对话框。

< 创建项目

1 基本信息 — 2 选择服务 — 3 配置资源 — 4 信息确认

* 项目名称:

项目描述:

取消 上一步 下一步

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。



请选择服务（至少选择一个）

 图计算 已购买

图计算开发是运行在spark之上图开发平台，提供了丰富的图基础算子和业务算子，支持拖拽方式进行图数据开发，数据输入输出仅支持hive。

 数据采集 已购买

数据采集是一种面向开发者提供的端到端数据采集服务，是数据进入大数据平台的第一道关卡。支持日志文件、数据库、报文接口等多种数据源的的流式、批量采集，提供向导式的采集配置、在线Agent管理、实时采集任务监控等服务能力。

 数据集成 已购买

建行大数据云数据集成服务支持金融级强监管要求下的数据集成功能，开通后您可以轻松实现数据在不同系统间的整合和流通，实现向业务快速供数。

 离线计算 已购买

离线计算是一种经济并高效的分析和处理海量数据的开发平台，可提供快速、完全托管的PB级数据仓库解决方案，支持以SQL代码、shell脚本、拖拽式等多种开发模式构建金融企业级数仓。

 流计算 已购买

建行大数据云流计算服务提供面向高速流式数据进行实时快速计算的解决方案，开通后可满足您实时风控、实时推荐等业务场景数据开发。

 数据挖掘 已购买

建行大数据云数据挖掘服务提供可视化建模和Notebook建模能力，内含多种数据挖掘算法，开通后您可以轻松构建算法模型，快速服务风险控制、精准营销等业务。

 数据服务 已购买

建行大数据云数据服务产品是构建数据中台的基础，可提供统一的数据访问能力，提供可视化的API开发和服务治理能力。

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）”。

< 创建项目

① 基本信息 ————— ② 选择服务 ————— ③ 配置资源 ————— ④ 信息确认

项目资源组

流计算：

数据服务：

数据采集：

取消

上一步

下一步



对以上配置信息进行确认，无误后点击“确定”按钮。

< 创建项目

① 基本信息 — ② 选择服务 — ③ 配置资源 — ④ 信息确认

基本信息

项目名称: fuwu

项目描述:

选择的服务

数据服务 数据采集 流计算

配置资源

数据服务资源组: 默认资源分组

数据采集资源组: 默认资源分组

流计算资源组: 默认资源分组

取消

上一步

确定

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。

项目空间 > 我的项目

新建项目

项目名称	创建时间	负责人	开通的服务	最近更新	状态	操作
fuwu	2019-09-05 01:59:16	uat_test	流计算 数据采集 数据服务	2019-09-05 01:59:16	● 正常	配置项目 修改服务
Project_1	2019-09-04 21:22:16	uat_test	流计算 数据采集 数据集成 离线计算	2019-09-04 21:22:16	● 正常	配置项目 修改服务
wer	2019-09-04 17:08:38	uat_test	数据采集 数据集成 图计算	2019-09-04 17:08:38	● 正常	配置项目 修改服务
111	2019-09-04 14:17:09	uat_test	图计算	2019-09-04 14:17:09	● 正常	配置项目 修改服务
test11111	2019-09-04 09:54:22	uat_test	流计算	2019-09-04 09:54:22	● 正常	配置项目 修改服务
test	2019-09-03 21:29:26	uat_test	流计算	2019-09-03 21:29:26	● 正常	配置项目 修改服务
my_stream_test_proj	2019-09-03 17:55:08	uat_test	流计算 数据采集 数据集成 离线计算 数据挖掘 数据服务	2019-09-03 18:00:46	● 正常	配置项目 修改服务
atk_proj08	2019-09-02 13:14:08	uat_test	流计算 数据采集 数据集成 离线计算 数据挖掘 数据服务	2019-09-02 13:14:08	● 正常	配置项目 修改服务
jhh	2019-08-31 00:25:41	uat_test	流计算 数据采集 数据集成 离线计算 数据挖掘 数据服务	2019-08-31 00:25:41	● 正常	配置项目 修改服务
eeure	2019-08-30 19:53:54	uat_test	流计算 数据采集 数据挖掘	2019-08-30 19:53:54	● 正常	配置项目 修改服务

上一页 1 2 3 4 5 ... 8 下一页 每页显示 10行/页 共 77 条



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）。

项目空间 > 我的项目

新建项目

项目名称	创建时间	负责人	开通的服务	最近更新时间	状态	操作
fuwu	2019-09-05 01:59:16	uat_test	流计算 数据采集 数据服务	2019-09-05 01:59:16	● 正常	配置项目 修改服务
Project_1	2019-09-04 21:22:16	uat_test	流计算 数据采集 数据集成 离线计算	2019-09-04 21:22:16	● 正常	配置项目 修改服务
wer	2019-09-04 17:08:38	uat_test	数据采集 数据集成 图计算	2019-09-04 17:08:38	● 正常	配置项目 修改服务
111	2019-09-04 14:17:09	uat_test	图计算	2019-09-04 14:17:09	● 正常	配置项目 修改服务
test11111	2019-09-04 09:54:22	uat_test	流计算	2019-09-04 09:54:22	● 正常	配置项目 修改服务
test	2019-09-03 21:29:26	uat_test	流计算	2019-09-03 21:29:26	● 正常	配置项目 修改服务
my_stream_test_proj	2019-09-03 17:55:08	uat_test	流计算 数据采集 数据集成 离线计算 数据挖掘 数据服务	2019-09-03 18:00:46	● 正常	配置项目 修改服务
atk_proj08	2019-09-02 13:14:08	uat_test	流计算 数据采集 数据集成 离线计算 数据挖掘 数据服务	2019-09-02 13:14:08	● 正常	配置项目 修改服务

fuwu 控制台首页 产品与服务 帮助与文档 uat_test

数据采集

流式采集

新建采集 批量删除

全部采集类型 搜索任务名称/创建用户 重置

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
无数据										

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。

创建topic

1 配置基本信息

* Topic名称:
仅支持英文、数字、下划线，且不能以数字和下划线开头，最大100字符

Topic中文名:

* 所属项目:

* Topic数据格式: json CSV

* Topic分区数:

生命周期:

Topic描述:



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
id		字符串	50	5	不脱敏		下移 删除
name		字符串	50	5	不脱敏		上移 删除

新增字段

上一步

完成

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": "TCLLOUD.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": "000000000000000003770", //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,
  }
}
```

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		字符串	50	5	不脱敏	库名.表名	下移 删除
op_type		字符串	50	5	不脱敏	操作类型	上移 下移 删除
current_ts		日期时间		5	不脱敏	处理时间	上移 下移 删除
pos		浮点数	10,2	5	不脱敏	偏移量	上移 下移 删除
before		对象		5	不脱敏	操作前的字段	上移 下移 新增 删除
ID		整数	9	5	不脱敏	业务字段	下移 删除
AGE		整数	9	5	不脱敏	业务字段	上移 删除
after		对象		5	不脱敏	操作后的字段	上移 新增 删除
ID		整数	9	5	不脱敏	业务字段	下移 删除
AGE		整数	9	5	不脱敏	业务字段	上移 删除

固定字段

对象类型

对象内的字段可以由用户自定义，
但两组对象中的字段需保持一致

新增字段

2. 当作为Oracle数据库采集的投递目标topic时

```

{
  "table": "TCLLOUD.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": "000000000000000003770", //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		字符串	50	5	不脱敏	库名.表名	下移 删除
op_type		字符串	50	5	不脱敏	操作类型	上移 下移 删除
op_ts		日期时间		5	不脱敏	操作时间	上移 下移 删除
current_ts		日期时间		5	不脱敏	处理时间	上移 下移 删除
pos		浮点数	10,2	5	不脱敏	偏移量	上移 下移 删除
before		对象		5	不脱敏	操作前的字段	上移 下移 新增 删除
ID		整数	9	5	不脱敏	业务字段	下移 删除
AGE		整数	9	5	不脱敏	业务字段	上移 删除
after		对象		5	不脱敏	操作后的字段	上移 新增 删除
ID		整数	9	5	不脱敏	业务字段	下移 删除
AGE		整数	9	5	不脱敏	业务字段	上移 删除

固定字段 (指向 table, op_type, op_ts, current_ts, pos)

对象类型 (指向 before, after)

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 (指向 ID, AGE in both before and after)



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

数据管理 > 库表管理

设计态 测试环境 生产环境

关系型数据库
Kafka
COS
ArangoDB
HBase
Elasticsearch

所属项目: 全部 请输入表名称

Topic名称	Topic中文名称	所属项目	创建人	创建时间	版本	表描述	发布状态	发布的数据源	操作
s_test_topic		s_test	ccb_test	2019-08-21 10:18:38	1566353918120		待发布	-	编辑 删除 发布到测试
test001		liufei_dg_test	ccb_test	2019-08-20 17:23:44	1566293023668		已发布	测试:dgKafka.topicdb_156_223, defaultKafka.topicdb_156_84	编辑 删除
source_cho_longitude		cc_test_cho	ccb_test	2019-08-20 14:36:44	1566287493741		已发布	测试:defaultKafka.topic	编辑 删除

发布测试时，需指定此topic归属的数据源

发布到测试环境

*数据源类型: kafka

*数据源名称: defaultKafka

发布说明:

取消 确认



发布测试完毕后，可在测试环境中筛选查看：

The screenshot shows the 'Database Management' (库表管理) interface. The main table lists databases with columns for name, Chinese name, project, data type, source name, creator, creation time, and description. A red box highlights the 's_test' project in the '所属项目' column. A red arrow points from this box to the '测试环境' (Test Environment) button at the top. Another red arrow points from the '发布到生产' (Publish to Production) button in the '操作' column to the '发布到生产' button in the detailed view below. The detailed view shows a table with columns for name, Chinese name, project, data type, source name, creator, creation time, and description, with a red box around the 's_test_topic' entry.

表名称	库中文名称	所属项目	数据类型	数据源名称	创建人	创建时间	用途描述	操作
topicdb_160_84		s_test	kaFka	defaultKafka	ccb_test	2019-08-21 10:19:41		发布到生产 去三表

表名称	库中文名称	所属项目	数据类型	数据源名称	创建人	创建时间	用途描述	操作
s_test_topic		s_test	kaFka	defaultKafka	ccb_test	2019-08-21 10:18:38		发布到生产 详情

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。

数据管理 > 库表管理

设计态 测试环境 生产环境

我设计的库表 我有权限的表 项目账号的库表 项目有权限的表

发布到生产 数据源种类: [v] 数据源名称: [v] 所属项目: s_test [v] 输入库名称: [input] [search] [refresh]

库名称	库中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	用途描述	操作
topicdb_160_84		s_test	kafka	defaultKafka	ccb_test	2019-08-21 10:19:41		发布到生产 显示表

上一页 1 下一页 每页显示 10行 / 页 共 1 条

数据管理 > 库表管理

设计态 测试环境 生产环境

我设计的库表 我有权限的表 项目账号的库表 项目有权限的表

数据源种类: [v] 数据源名称: [v] 所属项目: s_test [v] 输入库名称: [input] [search] [refresh]

库名称	库中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	用途描述	操作
topicdb_160_84		s_test	kafka	defaultKafka	ccb_test	2019-08-21 10:19:41		显示表

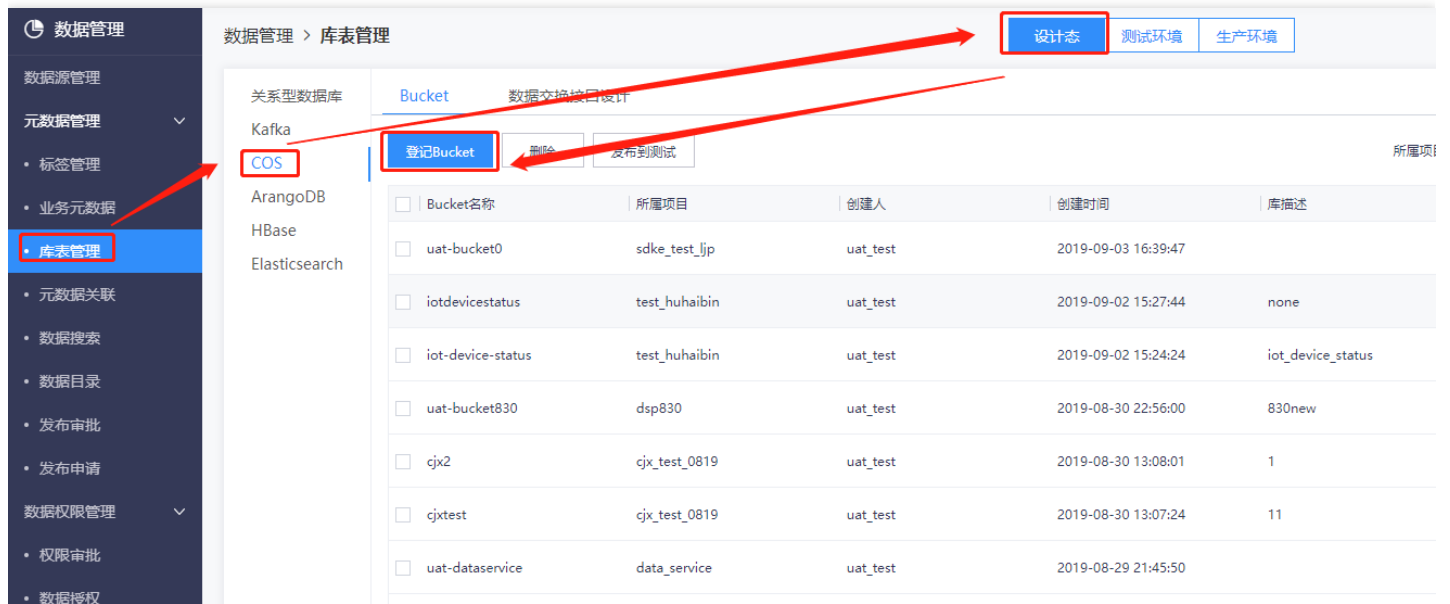
上一页 1 下一页 每页显示 10行 / 页 共 1 条

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。

控制台首页 产品与服务 帮助与文档 uat_test

数据管理 > 库表管理

设计态 测试环境 生产环境

Bucket 数据交换接口设计

关系型数据库 Kafka COS ArangoDB HBase Elasticsearch

Bucket名称 删除 发布到测试

所属项目: 全部 请输入表名称

Bucket名称	所属项目	创建人	创建时间	库描述	发布状态	发布的数据源	操作
testbucket	zb819	uat_test	2019-09-05 19:51:18		待发布		编辑 发布到测试 删除
uat-bucket0	sdke_test_ljp	uat_test	2019-09-03 16:39:47		已发布	测试cos	编辑 删除
iotdevicestatus	test_huhaibin	uat_test	2019-09-02 15:27:44	none	已发布	测试cos	编辑 删除

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。

发布到测试环境

* 数据源类型: COS

* 数据源名称: COS

如果数据源无测试环境，点击确定将直接发布到生产

发布说明:

取消 确认

发布成功后，Bucket信息可在测试环境查看。

数据管理 > 库表管理

设计态 测试环境 生产环境

我设计的库表 我有权限的表 项目账号的库表 项目有权限的表

发布到生产

数据源种类: cos 数据源名称: 所属项目: 输入库名称

库名称	库中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	库描述	操作
uat-bucket0		sdke_test_ljp	cos	cos	uat_test	2019-09-03 16:39:47		发布到生产 显示表
iotdevicestatus		test_huhaibin	cos	cos	uat_test	2019-09-02 15:27:44	none	发布到生产 显示表
iot-device-status		test_huhaibin	cos	cos	uat_test	2019-09-02 15:24:24	iot_device_status	发布到生产 显示表
uat-bucket830		dsp830	cos	cos	uat_test	2019-08-30 22:56:00	830new	发布到生产 显示表
cjx2		cjx_test_0819	cos	cos	uat_test	2019-08-30 13:08:01		发布到生产 显示表
cjxtest		cjx_test_0819	cos	cos	uat_test	2019-08-30 13:07:24		发布到生产 显示表
uat-dataservice		data_service	cos	cos	uat_test	2019-08-29 21:45:50		发布到生产 显示表

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”。

接口名称	接口中文名称	所属项目	创建人	创建时间	版本
qqqq		zeng_test819	uat_test	2019-09-04 16:00:31	156758403110
aaaallkks999		dsp830	uat_test	2019-09-02 16:46:25	156741768526
iot_airConditionerStatus	物联网空调状态	test_huhaibin	uat_test	2019-09-02 15:20:56	156740885564
issue_test22		dsp830	uat_test	2019-09-02 13:05:56	156740075606
issue_test25		dsp830	uat_test	2019-09-02 13:05:21	156740072086
issue_test5		dsp830	uat_test	2019-09-02 12:09:16	156739735558
issue_test4		dsp830	uat_test	2019-09-02 11:27:23	156739484268

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

1 配置基本信息 2 字段设置 3 数据路径

* 接口名称:
仅支持英文、数字、下划线，且不能以数字和下划线开头，最大100字符

接口中文名:

* 所属项目:

接口描述:

* 字段设置: json(半结构化) 文本(结构化) 非结构化

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

① 配置基本信息 ② 字段设置 ③ 数据路径

字段名称	字段中文名称	字段类型	字段长度	是否为主键	是否必填	默认值	字段长度
id		整数	9	否	否		5
name		字符串	50	否	否		5
age		整数	9	否	否		5

新增字段

上一步 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。

创建数据交换接口

① 配置基本信息 ② 字段设置 ③ 数据路径

数据路径:

文件字符: UTF-8

文件分隔符:

是否限制定长: 是 否

上一步 完成

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。

数据管理 > 库表管理

设计态 | 测试环境 | 生产环境

关系型数据库: Bucket | 数据交换接口设计

创建数据交换接口 | 删除 | 导入 | 发布到测试

所属项目: 全部 | 请输入表名称

接口名称	接口中文名称	所属项目	创建人	创建时间	版本	表描述	发布状态	发布的数据库	操作
ttttt1		zb819	uat_test	2019-09-05 19:57:01	1567684621236		待发布	-	编辑 删除 发布到测试
qqqq		zeng_test819	uat_test	2019-09-04 16:00:31	1567584031100		已发布	测试.cos.uat-bucket68	编辑 删除
aaaallkks999		dsp830	uat_test	2019-09-02 16:46:25	1567417685268		已发布	-	编辑 删除
iot_airConditionerStatus	物联网空调状态	test_huhaibin	uat_test	2019-09-02 15:20:56	1567408855649		已发布	测试.cos.iotdevicestatus	编辑 删除

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。

发布到测试环境

* 数据源类型: cos

* Bucket名称: cos/uat-bucket99

发布说明:

取消 | 确认

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。

数据管理 > 库表管理

设计态 **测试环境** 生产环境

我设计的库表 我有权限的表 项目账号的库表 项目有权限的表

发布到生产 数据源种类: cos 数据源名称: 所属项目: zb819 zb819 输入库名称: [搜索] [刷新]

库名称	库中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	库描述	操作
uat-bucket39		zb829	cos	cos	uat_test	2019-08-29 16:22:09		发布到生产 显示表
uat-bucket99		zb819	cos	cos	uat_test	2019-08-19 20:59:31		发布到生产 显示表

上一页 1 下一页 每页显示 10行 / 页 共 2 条

uat-bucket99

打印

表名称	表中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	表描述	操作
ttttt1		zb819	cos	cos	uat_test	2019-09-05 19:57:01		发布到生产 详情
test_for_lbz_target		zb819	cos	cos	uat_test	2019-08-28 18:11:17		发布到生产 详情
test_for_lbz		zb819	cos	cos	uat_test	2019-08-27 20:10:45		发布到生产 详情
t3		zb819	cos	cos	uat_test	2019-08-27 11:26:16		发布到生产 详情

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。

数据管理 > 库表管理

设计态 测试环境 生产环境

我设计的库表 我有权限的表 项目账号的库表 项目有权限的表

发布到生产 数据源种类: cos 数据源名称: 所属项目: zb819 zb829 输入库名称

库名称	库中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	库描述	操作
uat_bucket39		zb829	cos	cos	uat_test	2019-08-29 16:22:09		发布到生产 显示表
uat_bucket99		zb819	cos	cos	uat_test	2019-08-19 20:59:31		发布到生产 显示表

上一页 1 下一页 每页显示 10行 / 页 共 2 条

uat_bucket99

打标签

表名称	表中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	表描述	操作
ttttt1		zb819	cos	cos	uat_test	2019-09-05 19:57:01		发布到生产 详情
test_for_lbz_target		zb819	cos	cos	uat_test	2019-08-28 18:11:17		发布到生产 详情

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。

数据管理 > 库表管理

设计态 测试环境 生产环境

我设计的库表 我有权限的表 项目账号的库表 项目有权限的表

发布到生产 数据源种类: cos 数据源名称: 所属项目: zb819 zb829 输入库名称

库名称	库中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	库描述	操作
uat_bucket39		zb829	cos	cos	uat_test	2019-08-29 16:22:09		发布到生产 显示表
uat_bucket99		zb819	cos	cos	uat_test	2019-08-19 20:59:31		发布到生产 显示表

上一页 1 下一页 每页显示 10行 / 页 共 2 条

uat_bucket99

打标签

表名称	表中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	表描述	操作
ttttt1		zb819	cos	cos	uat_test	2019-09-05 19:57:01		发布到生产 详情
test_for_lbz_target		zb819	cos	cos	uat_test	2019-08-28 18:11:17		发布到生产 详情
test_for_lbz		zb819	cos	cos	uat_test	2019-08-27 20:10:45		发布到生产 详情

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。

数据管理 > 库表管理

设计态 测试环境 **生产环境**

我设计的库表 我有权限的表 项目账号的库表 项目有权限的表

数据源种类: cos 数据源名称: 所属项目: 输入库名称: [搜索] [刷新]

库名称	库中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	库描述	操作
cjx2		cjx_test_0819	cos	cos	uat_test	2019-08-30 13:08:01		显示表
lf-bucket-0829		lf_test_sjgj	cos	cos	uat_test	2019-08-29 20:46:33		显示表
uat-bucket37		dsp830	cos	cos	uat_test	2019-08-27 10:48:54		显示表
uat-bucket89		test_lir_0816	cos	cos	uat_test	2019-08-26 21:09:35		显示表
sjjc-bucket		sjjc_test	cos	cos	uat_test	2019-08-23 14:14:11		显示表

上一页 1 2 下一页 每页显示 10行 / 页 共 20 条

lf-bucket-0829

打标签

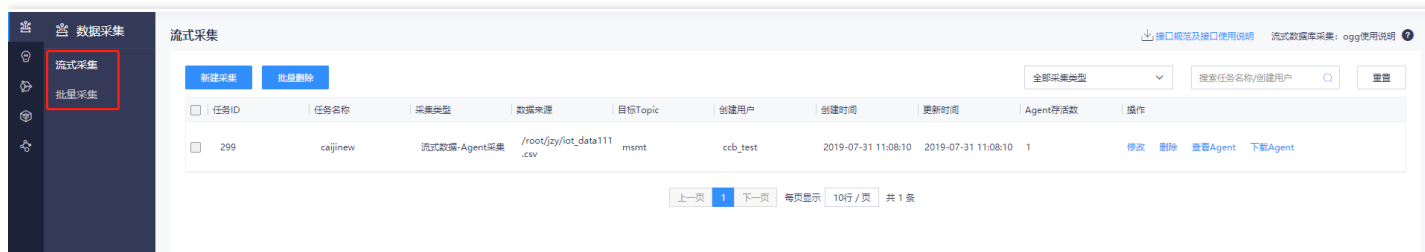
表名称	表中文名称	所属项目	数据源种类	数据源名称	创建人	创建时间	表描述	操作
exchange_0829	dfaf	lf_test_sjgj	cos	cos	uat_test	2019-08-29 20:50:05	fdsfdf	详情

上一页 1 下一页 每页显示 10行 / 页 共 1 条

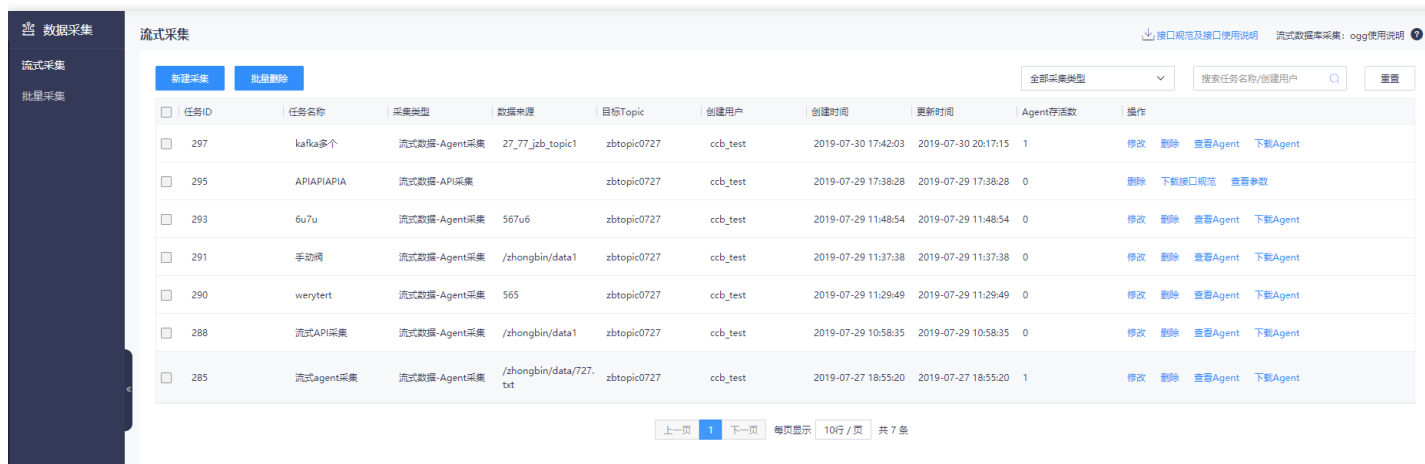
流式采集配置

最近更新: 2022-01-21 15:56:29

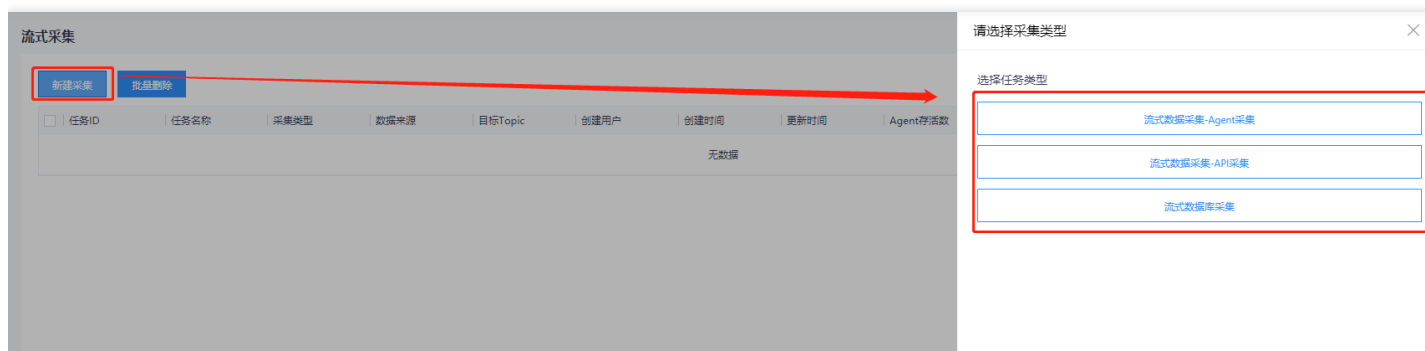
数据采集组件支持流式采集、批量采集两种方式，进入开发页面，默认展开流式采集页面。



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。

流式采集

新建采集 批量删除

任务ID	任务名称	采集类型	数据来源	目标Topic	创建用户
141	test0831	流式数据-API采集		testgz	uat_test
125	4	mysql数据库采集	data_collection-class	testgz	uat_test
124	3	mysql数据库采集	data_collection-student	testgz	uat_test
			data_colle		

上一页 1 2

新建流式数据采集-Agent采集任务

1 配置基本信息 2 配置Agent 3 生成部署包

* 采集名称: 仅支持中文、英文、数字、下划线、连字符，最大50字符

采集说明: 最大支持500字

* 目标kafka: dgKafka

* Topic名称: testgz

隐藏Topic预览

测试环境 生产环境

表信息

序号	列名	类型	字段说明
1	id	STRING	

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

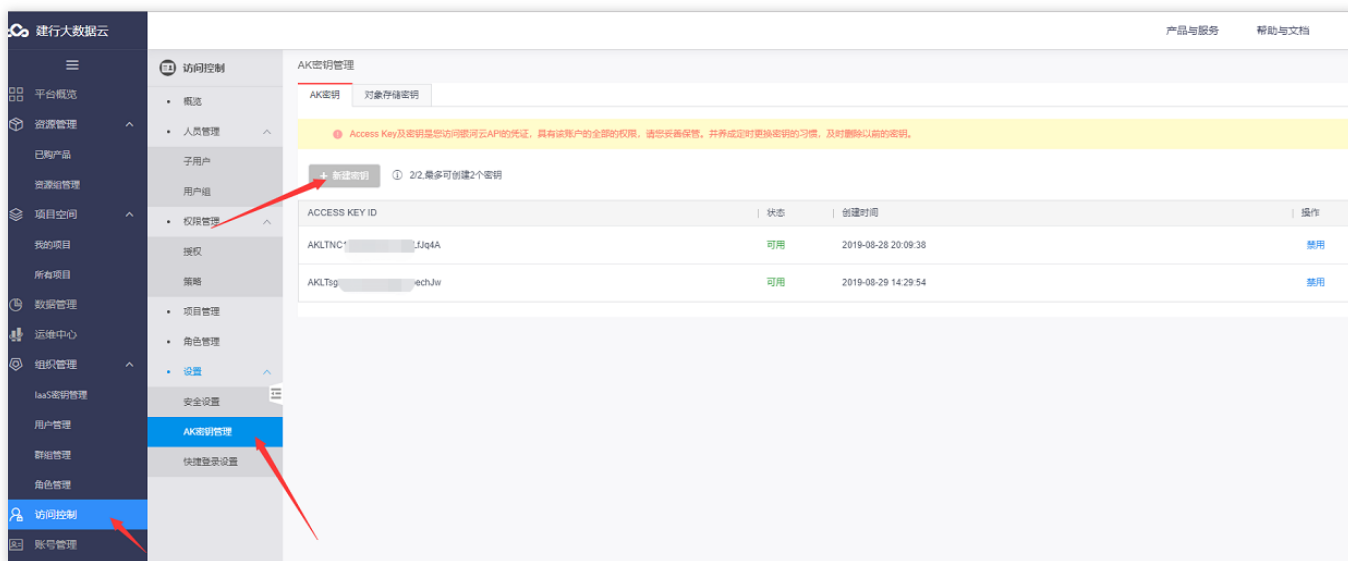


配置agent信息-文件采集的表单截图，包含以下字段：

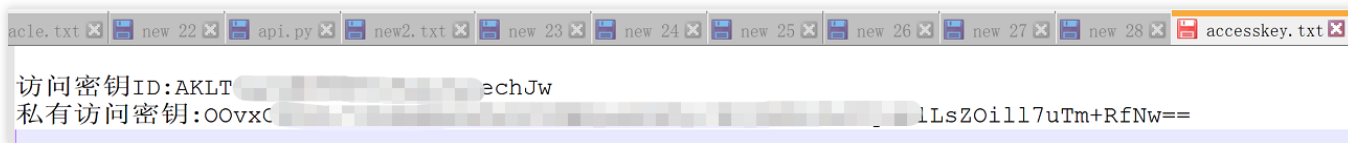
- * 待采集文件类型：file (选中)、kafka、dir
- * 待采集文件：输入框，右侧有“+”号
- * accessId：输入框，右侧有“?”号
- * SK文件路径：skFile，右侧有“?”号

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密钥到本地的弹窗，下载到本地后，文件里面的内容就是Secret Access key(SK)，对应访问密钥和私有访问密钥。





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

① 配置基本信息 ② 配置Agent ③ 生成部署包

* 待采集文件类型：

file

kafka

dir

* 待采集文件夹：



* accessId：



* SK文件路径：

skFile



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

① 配置基本信息

② 配置Agent

③ 生成部署包

* 待采集文件类型:

file

kafka

dir

* 请输入kafka地址:

10.56.5.1:9092,10.56.5.2:9092



* 请输入topic:



* accessId:



* SK文件路径:

skFile



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。

√ 隐藏高级设置


缓存配置：


类型选择： 


* 最大容量： 条 

* 事务容量： 条 

传输配置：

* 最大数据量/批： 条 

* 最小数据量/批： 条 

* 并发线程数： 个 

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：


缓存配置：

类型选择： ▼

*最大容量： 条 ?

*事务容量： 条 ?

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。

传输配置:

*最大数据量/批: 条 ?

*最小数据量/批: 条 ?

*并发线程数: 个 ?

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。

下载Agent及OGG工具



① 配置基本信息

② 配置Agent

③ 生成部署包

方式一、下载Agent启动数据采集任务

Agent类型	下载地址
Agent_for_mysql(Linux)	下载Agent及使用说明

说明: 请按需下载以上文件, 并部署到各个采集节点

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。

流式采集 [接口规范及接口使用说明](#) [查看帮助文档](#)

新建采集 批量删除 全部采集类型 搜索任务名称/创建用户 重置

<input type="checkbox"/>	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
<input type="checkbox"/>	141	test0831	流式数据-API采集		testgz	uat_test	2019-08-31 11:40:44	2019-08-31 11:40:44	0	0	删除 下载接口规范 查看参数
<input type="checkbox"/>	125	4	mysql数据库采集	data_collection-class	testgz	uat_test	2019-08-30 11:45:35	2019-08-30 11:45:35	0	0	Agent列表 查看 删除 下载Agent
<input type="checkbox"/>	124	3	mysql数据库采集	data_collection-student	testgz	uat_test	2019-08-30 10:54:58	2019-08-30 10:54:58	0	0	Agent列表 查看 删除 下载Agent
<input type="checkbox"/>	123	2	mysql数据库采集	data_collection-student	testgz	uat_test	2019-08-30 10:45:08	2019-08-30 10:45:08	0	0	Agent列表 查看 删除 下载Agent
<input type="checkbox"/>	95	gztest	Agent-文件采集	/data/b.txt	testgz	uat_test	2019-08-28 23:40:45	2019-08-30 17:56:40	0	15.14	Agent列表 修改 删除 下载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。

```
[root@wnlbdsjcjap0001 tmp]# cd dg_agent/  
[root@wnlbdsjcjap0001 dg_agent]# ls  
bin  conf  lib  logs  start.sh  
[root@wnlbdsjcjap0001 dg_agent]# . ./start.sh  
[root@wnlbdsjcjap0001 dg_agent]# nohup: ignoring input and redirecting stderr to s
```

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。

The screenshot displays the '数据管理' (Data Management) interface. On the left sidebar, '数据目录' (Data Directory) is highlighted. The main area shows a tree view of data sources under 'kafka', with 'kafka_source_1' selected. The '数据预览' (Data Preview) tab is active, showing a list of data records in a table format. The records are as follows:

Line No.	Data
1	{ "data": "php0,20,1565936303578,227.01" }
2	
3	{ "data": "php0,20,1565936303578,227.01" }
4	
5	{ "data": "php0,20,1565936303578,227.01" }
6	
7	{ "data": "php0,20,1565936303578,227.01" }
8	
9	{ "data": "php0,20,1565936303578,227.01" }
10	
11	{ "data": "php0,20,1565936303578,227.01" }
12	
13	{ "data": "php0,20,1565936303578,227.01" }
14	
15	{ "data": "php0,20,1565936303578,227.01" }
16	
17	{ "data": "happy0,20,1565936323590,427.01" }
18	
19	{ "data": "happy0,20,1565936323590,427.01" }
20	
21	{ "data": "happy1,21,1565936324590,437.01" }

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）

The screenshot displays the 'Stream Collection' (流式采集) management interface. The main table lists tasks with columns for Task ID, Task Name, Collection Type, Data Source, Target Topic, Created User, Creation Time, Update Time, Agent Count, and Actions. The 'View Agent' button for task ID 297 is highlighted with a red box. A red arrow points from this button to the 'Agent Management' modal window below, which shows details for task ID 297, including IP, Hostname, Environment, Status, and Updateability.

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	操作
297	kaika多个	流式数据-Agent采集	27_77_jib_topic1	zbtopic0727	ccb_test	2019-07-30 17:42:03	2019-07-30 20:17:15	1	修改 删除 查看Agent 下载Agent
295	APIAPIAIA	流式数据-API采集		zbtopic0727	ccb_test	2019-07-29 17:38:28	2019-07-29 17:38:28	0	删除 下载接口规范 查看参数
293	6u7u	流式数据-Agent采集	567u6	zbtopic0727	ccb_test	2019-07-29 11:48:54	2019-07-29 11:48:54	0	修改 删除 查看Agent 下载Agent
291	手动调	流式数据-Agent采集	/zhongbin/data1	zbtopic0727	ccb_test	2019-07-29 11:37:38	2019-07-29 11:37:38	0	修改 删除 查看Agent 下载Agent
290	werytert	流式数据-Agent采集	565	zbtopic0727	ccb_test	2019-07-29 11:29:49	2019-07-29 11:29:49	0	修改 删除 查看Agent 下载Agent
288	流式API采集	流式数据-Agent采集	/zhongbin/data1	zbtopic0727	ccb_test	2019-07-29 10:58:35	2019-07-29 10:58:35	0	修改 删除 查看Agent 下载Agent
285	流式Agent采集	流式数据-Agent采集	/zhongbin/data/727.txt	zbtopic0727	ccb_test	2019-07-27 18:55:20	2019-07-27 18:55:20	1	修改 删除 查看Agent 下载Agent

采集任务名称	IP	Hostname	运行环境	运行状态	是否可更新	最后上报时间	操作
kaika多个	10.77.0.29	localhost	测试环境	运行中	否	2019-07-31 14:49:15	暂停 停止

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。

The screenshot displays the 'Stream Collection' (流式采集) interface. On the left, there is a table of existing tasks with columns for Task ID, Task Name, Collection Type, Data Source, Target Topic, Created User, Created Time, and Updated Time. A modal window titled 'Please select collection type' (请选择采集类型) is open on the right, showing three options: 'Stream Data Collection-Agent Collection' (流式数据采集-Agent采集), 'Stream Data Collection-API Collection' (流式数据采集-API采集), and 'Stream Database Collection' (流式数据库采集). The 'Stream Data Collection-API Collection' option is highlighted with a red box.

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
141	test0831	流式数据-API采集		testgz	uat_test	2019-08-31 11:40:44	2019-08-31 11:40:44
125	4	mysql数据库采集	data_collection-class	testgz	uat_test	2019-08-30 11:45:35	2019-08-30 11:45:35
124	3	mysql数据库采集	data_collection-student	testgz	uat_test	2019-08-30 10:54:58	2019-08-30 10:54:58

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。

流式采集

新建采集 批量删除

任务ID	任务名称	采集类型	数据来源	目标Topic	创建用户
141	test0831	流式数据-API采集		testgz	uat_test
125	4	mysql数据库采集	data_collection-class	testgz	uat_test
124	3	mysql数据库采集	data_collection-student	testgz	uat_test

新一页 1 2

新建流式数据采集-API采集任务

1 配置基本信息 2 下载接口规范

* 采集名称: test_API

采集说明: 最大支持500字

* 目标kafka: dgKafka

* Topic名称: testgz

隐藏Topic预览

测试环境 生产环境

表信息

序号	列名	类型	字段说明
1	id	STRING	

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》。

任务ID	任务名称	采集类型	数据来源	目标Topic	创建用户
158	test_API	流式数据-API采集		testgz	uat_test
141	test0831	流式数据-API采集		testgz	uat_test
125	4	mysql数据库采集	data_collection-class	testgz	uat_test

新建流式数据采集-API采集任务

① 配置基本信息 | ② 下载接口规范

请通过调用接口的方式启动采集任务：

点击下载接口规范：[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	用户上传的数据



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。

The screenshot shows the 'Stream Data Collection' (流式采集) interface. On the left, there is a table with columns: 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, and 更新时间. The table contains three rows of data. Below the table are navigation buttons: '新建采集' (New Collection) and '批量删除' (Batch Delete). On the right, a modal titled '请选择采集类型' (Please select collection type) is open, showing three options: '流式数据采集-Agent采集' (Stream Data Collection-Agent Collection), '流式数据采集-API采集' (Stream Data Collection-API Collection), and '流式数据库采集' (Stream Database Collection), which is highlighted in blue.

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
158	test_API	流式数据-API采集		testgz	uat_test	2019-09-05 16:18:34	2019-09-05 16:18:34
141	test0831	流式数据-API采集		testgz	uat_test	2019-08-31 11:40:44	2019-08-31 11:40:44
125	4	mysql数据库采集	data_collection-class	testgz	uat_test	2019-08-30 11:45:35	2019-08-30 11:45:35

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。

流式采集

新建采集 批量删除

任务ID	任务名称	采集类型	数据来源	目标Topic	创建用户
158	test_API	流式数据-API采集		testgz	uat_test
141	test0831	流式数据-API采集		testgz	uat_test
125	4	mysql数据库采集	data_collection-class	testgz	uat_test

新一页 1 2

新建流式数据库采集任务

1 配置基本信息 2 配置Agent 3 生成部署包

* 采集名称: test_DB

采集说明: 最大支持500字

* 目标kafka: dgKafka

* Topic名称: testgz

隐藏Topic预览

测试环境 生产环境

表信息

序号	列名	类型	字段说明
1	id	STRING	

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。

新建流式数据库采集任务

1 配置基本信息 2 配置Agent 3 生成部署包

- * 待采集文件类型: oracle mysql
- * 待采集Mysql数据库地址: ?
- * 指定canal存放路径: ?
- * 填写Mysql用户名: ?
- * 填写Mysql的密码: ?
- * 填写Mysql数据库名称: ?
- * 输入表名称: ?
- * accessId: ?
- * SK文件路径: ?



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》

产品与服务 **帮助与文档**

帮助与文档

平台使用 FAQ问题集

- > 公共平台文档
- √ 数据采集文档
 - 数据采集：用户手册
 - 流式采集：API使用说明
 - Oracle数据库：OGG使用说明
 - Mysql数据库：数据采集说明**
 - 批量采集：API使用说明
 - 批量采集：SDK使用说明



下载并部署agent

最近更新时间: 2019-11-12 02:57:05

MySQL数据库采集

- ① 配置基本信息 ② 配置Agent ③ 生成部署包

方式一、下载Agent启动数据采集任务

Agent类型	下载地址
Agent_for_mysql(Linux)	下载Agent及使用说明

说明: 请按需下载以上文件, 并部署到各个采集节点



Oracle数据库采集

下载Agent及OGG工具



① 配置基本信息

② 配置Agent

③ 生成部署包

步骤一、下载Agent启动数据采集任务

Agent类型	下载地址
Agent_for_oracle(Linux & Window)	下载Agent及使用说明

说明：请按需下载以上文件，并部署到各个采集节点

步骤二、下载通用OGG下载包及使用说明（每个客户端仅下载安装一次即可）：

Oracle源端OGG：[Oracle源端OGG下载包](#)目标端OGG：[目标端OGG下载包](#)

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": "TCLLOUD.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": "000000000000000003770", //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": "TCLLOUD.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": "000000000000000003770", //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）

The screenshot displays the '数据收集' (Data Collection) interface. The main area shows a table of tasks with columns for ID, name, type, source, topic, user, creation time, update time, and status. A red box highlights the '查看Agent' (View Agent) button in the '操作' (Action) column for task ID 297. Below the table, an 'Agent管理' (Agent Management) pop-up window is open for task ID 297, showing a table of agents with columns for name, IP, hostname, environment, status, updateable, and last report time. The '查看Agent' button in the main table is also highlighted with a red arrow pointing to the pop-up window.

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	操作
297	kaika多个	流式数据-Agent采集	27_77_jib_topic1	zbtopic0727	ccb_test	2019-07-30 17:42:03	2019-07-30 20:17:15	1	修改 删除 查看Agent 下载Agent
295	APIAPIAPIA	流式数据-API采集		zbtopic0727	ccb_test	2019-07-29 17:38:28	2019-07-29 17:38:28	0	删除 下载接口规范 查看参数
293	6u7u	流式数据-Agent采集	567u6	zbtopic0727	ccb_test	2019-07-29 11:48:54	2019-07-29 11:48:54	0	修改 删除 查看Agent 下载Agent
291	手动调	流式数据-Agent采集	/zhongbin/data1	zbtopic0727	ccb_test	2019-07-29 11:37:38	2019-07-29 11:37:38	0	修改 删除 查看Agent 下载Agent
290	werytert	流式数据-Agent采集	565	zbtopic0727	ccb_test	2019-07-29 11:29:49	2019-07-29 11:29:49	0	修改 删除 查看Agent 下载Agent
288	流式API采集	流式数据-Agent采集	/zhongbin/data1	zbtopic0727	ccb_test	2019-07-29 10:58:35	2019-07-29 11:38:35	0	修改 删除 查看Agent 下载Agent
285	流式Agent采集	流式数据-Agent采集	/zhongbin/data/727.txt	zbtopic0727	ccb_test	2019-07-27 18:55:20	2019-07-27 18:55:20	1	修改 删除 查看Agent 下载Agent

采集任务名称	IP	Hostname	运行环境	运行状态	是否可更新	最后上报时间	操作
kaika多个	10.77.0.29	localhost	测试环境	运行中	否	2019-07-31 14:49:15	暂停 停止



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。

批量采集 [下载文件推送通用工具](#) [查看帮助文档](#)

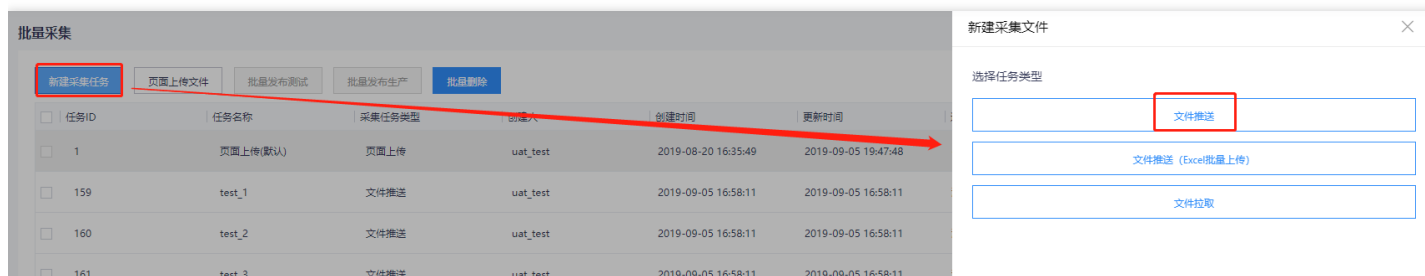
新建采集任务
页面上传文件
批量发布测试
批量发布生产
批量删除
全部采集类型
搜索任务名称/创建用户
重置

任务ID	任务名称	采集任务类型	创建人	创建时间	更新时间	运行环境	操作
1	页面上传(默认)	页面上传	uat_test	2019-08-20 16:35:49	2019-08-31 20:34:03	----	查看上传明细
104	子账号-任务推送	文件推送	libangzhu	2019-08-29 15:10:45	2019-08-29 15:10:45	测试环境	查看推送明细 发布生产 修改配置 查看参数 删除
71	windows定时推送-lbz	文件推送	uat_test	2019-08-27 14:57:20	2019-08-28 17:34:48	测试环境	查看推送明细 发布生产 修改配置 查看参数 删除
40	999999	文件拉取-周期	uat_test	2019-08-23 17:21:21	2019-08-23 17:21:21	测试环境	查看推送明细 发布生产 修改配置 启动 删除
39	aaaaaa	文件拉取-周期	uat_test	2019-08-23 16:19:45	2019-08-23 16:20:20	生产环境	查看推送明细 发布测试 修改配置 启动 删除
38	dasd	文件拉取-周期	uat_test	2019-08-23 16:10:59	2019-08-23 16:14:55	测试环境	查看推送明细 发布生产 修改配置 启动 删除
37	文件推送单次-lbz	文件拉取-单次	uat_test	2019-08-23 15:55:46	2019-08-23 15:55:46	测试环境	查看推送明细 发布生产 修改配置 执行 删除
36	setrtyertyert	文件拉取-单次	uat_test	2019-08-23 15:34:45	2019-08-23 15:45:50	测试环境	查看推送明细 发布生产 修改配置 执行 删除
31	linux定时推送-lbz	文件推送	uat_test	2019-08-23 14:10:33	2019-08-23 14:10:33	测试环境	查看推送明细 发布生产 修改配置 查看参数 删除

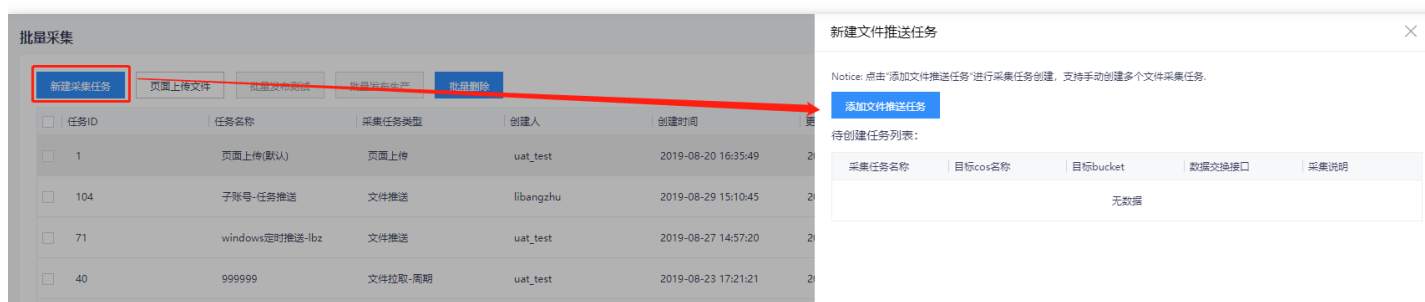
文件推送任务配置

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

添加文件推送任务 ✕

*采集任务名称：

采集任务说明：

*目标cos名称：

*目标Bucket：

*选择数据交换接口：

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。

新建文件推送任务 ✕

Notice: 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。

[添加文件推送任务](#)

待创建任务列表：

采集任务名称	目标cos名称	目标bucket	数据交换接口	采集说明
test_1	cos	uat-bucket37	breast_cancer0392005 0804241536	
test_2	cos	uat-bucket37	breast_cancer0392005 0804241536	
test_3	cos	uat-bucket37	breast_cancer0392005 0804241536	



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。

批量采集 通用下载



下载文件推送工具

Agent类型	下载地址
文件上传工具包 (Linux)	下载：文件上传工具包 (Linux) 及工具使用说明
文件上传工具包 (Windows)	下载：文件上传工具包 (Windows) 及工具使用说明
接口规范	下载：接口规范及接口使用说明
SDK	下载：SDK (JAVA) 及SDK使用说明

说明：API及SDK均支持单个文件/多个文件的推送。

创建采集任务后，可按需通过工具包 / 接口 / SDK 进行文件推送。

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。

新建采集文件

选择任务类型

文件推送

文件推送 (Excel批量上传)

文件拉取

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。

新建文件推送（Excel批量上传）任务

下载模板：[点击下载Excel模板](#)

* 上传批量任务列表：[选择文件](#)

请务必按照模板excel样式上传任务列表。

采集任务名称	目标cos名称	目标bucket	数据交换接口	采集说明
无数据				

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

运行环境	操作
2019-05 19:47:48	查看上传明细
2019-05 16:58:11 测试环境	查看推送明细 发布生产 修改配置 查看参数 删除
2019-05 16:58:11 测试环境	查看推送明细 发布生产 修改配置 查看参数 删除

Agent类型	下载地址
文件上传工具包 (Linux)	下载：文件上传工具包 (Linux) 及工具使用说明
文件上传工具包 (Windows)	下载：文件上传工具包 (Windows) 及工具使用说明
接口规范	下载：接口规范及接口使用说明
SDK	下载：SDK (JAVA) 及SDK使用说明

说明：API及SDK均支持单个文件/多个文件的推送。
创建采集任务后，可按需通过工具包 / 接口 / SDK 进行文件推送。



- 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：

The screenshot shows a web interface for a cloud storage bucket. The bucket name is 'buckettest2-test-1259417666'. The current path is 'buckettest2-test-1259417666 / zhongbin / aaa /'. There are two files listed:

文件名	大小	存储类型	更新时间	操作
file1.csv	63B	标准存储	2019-08-01 17:36:30	下载 详情 删除
file2.csv	360B	标准存储	2019-08-01 17:36:30	下载 详情 删除

2、批量采集：SDK采集demo可参考下图

```
0
1 public class CosDemo {
2     public static void main(String[] args) throws FileNotFoundException {
3         COSClient cosClient = UploadFileUtil.initParams("myqcloud.com", "AKIDLVTgXKCWAze9m3afGnBTaYIgrongwtEP",
4             "AAZvjNsoZ9XkpmZEVcYIrHZghLo7FCxd", "ap-beijing",
5             "buckettest2-test-1259417666");
6         PutObjectResult putObjectResult = UploadFileUtil.putFileToCos(cosClient, "/zhongbin/log.txt", "d:/log.txt");
7         System.out.println(putObjectResult.getETag());
8         UploadFileUtil.closeCOSClient(cosClient);
9     }
10 }
11
```

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



测试环境	查看推送明细	发布生产	修改配置	查看参数	删除
测试环境	查看推送明细	发布生产	修改配置	查看参数	删除
测试环境	查看推送明细	发布生产	修改配置	查看参数	删除
测试环境	查看推送明细	发布生产	修改配置	查看参数	删除
测试环境	查看推送明细	发布生产	修改配置	查看参数	删除

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。

新建文件拉取任务 ✕

* 采集任务名称:

采集任务说明:

拉取协议: FTP

* IP地址:

* 端口:

* 用户名:

* 密码:

[连通性测试](#)

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：

* 目标cos名称：

* 目标Bucket：

* 选择数据交换接口：

* 目标路径：

* 首次拉取时间：

* 拉取周期：

[^ 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。

批量采集

新建采集任务 页面上传文件 批量发布测试 批量发布生产 批量删除

全部采集类型 搜索任务名称/创建用户 重置

任务ID	任务名称	采集任务类型	创建人	创建时间	更新时间	运行环境	操作
160	test_2	文件推送	uat_test	2019-09-05 16:58:11	2019-09-05 16:58:11	测试环境	查看推送明细 发布生产 修改配置 查看参数 删除
161	test_3	文件推送	uat_test	2019-09-05 16:58:11	2019-09-05 16:58:11	测试环境	查看推送明细 发布生产 修改配置 查看参数 删除
104	子账号-任务推送	文件推送	libangzhu	2019-08-29 15:10:45	2019-08-29 15:10:45	测试环境	查看推送明细 发布生产 修改配置 查看参数 删除
71	windows定时推送-lbz	文件推送	uat_test	2019-08-27 14:57:20	2019-08-28 17:34:48	测试环境	查看推送明细 发布生产 修改配置 查看参数 删除
40	999999	文件拉取-周期	uat_test	2019-08-23 17:21:21	2019-08-23 17:21:21	测试环境	查看推送明细 发布生产 修改配置 启动 删除
39	aaaaaa	文件拉取-周期	uat_test	2019-08-23 16:19:45	2019-08-23 16:20:20	生产环境	查看推送明细 发布测试 修改配置 启动 删除

运行实例

采集任务名称	运行实例ID	数据源IP	数据源路径	目标Bucket	目标Sink	运行开始时间	运行结束时间	文件个数	文件大小(MB)	运行时长(ms)	实例状态	运行信息
windows定时推送-lbz	643	10.66.0.50	/zhongbin/b.txt	uat-bucket99-test-1259417666	/test/img	2019-08-29 18:03:41	2019-08-29 18:03:41	1	0	0	运行成功	
windows定时推送-lbz	641	10.66.0.50	/home/rdtest/libangzhu/dg_agent/start_test.sh	uat-bucket99-test-1259417666	/test/img	2019-08-29 17:55:26	2019-08-29 17:55:56	1	0	30,000	运行失败	查看
windows定时推送-lbz	639	10.66.0.50	/home/rdtest/libangzhu/dg_agent/start_test.sh	uat-bucket99-test-1259417666	test/img	2019-08-29 17:54:49	2019-08-29 17:55:19	1	0	30,000	运行失败	查看

批量采集

新建采集任务 页面上传文件 批量发布测试 批量发布生产 批量删除

全部采集类型 搜索任务名称/创建用户 重置

任务ID	任务名称	采集任务类型	创建人	创建时间	更新时间	运行环境	操作
71	windows定时推送-lbz	文件推送	uat_test	2019-08-27 14:57:20	2019-08-28 17:34:48	测试环境	查看推送明细 发布生产 修改配置 查看参数 删除
40	999999	文件拉取-周期	uat_test	2019-08-23 17:21:21	2019-08-23 17:21:21	测试环境	查看推送明细 发布生产 修改配置 启动 删除
39	aaaaaa	文件拉取-周期	uat_test	2019-08-23 16:19:45	2019-08-23 16:20:20	生产环境	查看推送明细 发布测试 修改配置 启动 删除
38	dasd	文件拉取-周期	uat_test	2019-08-23 16:10:59	2019-08-23 16:14:55	测试环境	查看推送明细 发布生产 修改配置 启动 删除
37	文件推送单次-lbz	文件拉取-单次	uat_test	2019-08-23 15:55:46	2019-08-23 15:55:46	测试环境	查看推送明细 发布生产 修改配置 执行 删除
36	setrjertyert	文件拉取-单次	uat_test	2019-08-23 15:34:45	2019-08-23 15:45:50	测试环境	查看推送明细 发布生产 修改配置 执行 删除

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。

The screenshot displays the 'Batch Collection' (批量采集) interface. On the left, a table lists various tasks with columns for Task ID, Task Name, Collection Type, Creator, and Creation Time. The 'Upload File' (页面上传文件) button is highlighted. On the right, a modal window titled 'Upload File' (通过文件上传) is open, showing configuration options for the upload process.

批量采集

新建采集任务 | 页面上传文件 | 批量发布测试 | 批量发布生产 | 批量删除

任务ID	任务名称	采集任务类型	创建人	创建时间	要
1	页面上传(默认)	页面上传	uat_test	2019-08-20 16:35:49	2
104	子账号-任务推送	文件推送	libangzhu	2019-08-29 15:10:45	2
71	windows定时推送-ibz	文件推送	uat_test	2019-08-27 14:57:20	2
40	999999	文件拉取-周期	uat_test	2019-08-23 17:21:21	2
39	aaaaaa	文件拉取-周期	uat_test	2019-08-23 16:19:45	2
38	dasd	文件拉取-周期	uat_test	2019-08-23 16:10:59	2
37	文件推送单次-ibz	文件拉取-单次	uat_test	2019-08-23 15:55:46	2
36	setryertyert	文件拉取-单次	uat_test	2019-08-23 15:34:45	2
31	linux定时推送-ibz	文件推送	uat_test	2019-08-23 14:10:33	2
8	manuel	文件推送	uat_test	2019-08-20 20:08:52	2
6	test4	文件推送	uat_test	2019-08-20 18:21:40	2

通过文件上传

* 选择上传文件: 选择文件 支持多个文件上传

* 选择上传环境: 测试环境 生产环境

* 目标cos名称: cos

* 目标Bucket: uat-bucket37

* 选择数据交换接口: breast_cancer03920050804241536

* cos路径: /test

隐藏数据交换接口预览

测试环境 生产环境

数据路径: breast_cancer-final0.3920050804241536

文件字符: UTF-8

文件分隔符: .

表信息

序号	列名	类型	字段说明
1	id	STRING	
2	diagnosis	STRING	

取消 确认

通过文件上传



* 选择上传文件:

blance.pdf



运维中心.rar



支持上传多个文件

选择文件

支持多个文件上传

* 选择上传环境:

测试环境

生产环境

* 目标cos名称:

请选择COS



* 目标Bucket:

请选择Bucket



* 选择数据交换接口:

请选择数据交换接口



* cos路径:

请输入目标cos路径

[显示数据交换接口预览](#)

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。

The screenshot shows the 'Batch Collection' (批量采集) interface. On the left, a sidebar contains 'Data Collection' (数据采集) with sub-items 'Stream Collection' (流式采集) and 'Batch Collection' (批量采集). The main area has a 'Batch Collection' (批量采集) header with buttons for 'New Task' (新建任务), 'Page Upload' (页面上传文件), 'Batch Release Test' (批量发布测试), 'Batch Release Production' (批量发布生产), and 'Batch Delete' (批量删除). Below this is a table of tasks:

任务ID	任务名称	采集任务类型	创建人	创建时间	更新时间	运行环境	操作
287	页面上传(默认)	页面上传	ccb_test	2019-07-27 19:58:22	2019-07-31 14:33:01	测试环境	查看上传明细
289	789oI8o	文件推送	ccb_test	2019-07-29 11:29:30	2019-07-29 11:29:30	测试环境	发布生产 查看参数 修改配置 删除任务 查看推送明细
286	test1	文件推送	ccb_test	2019-07-27 19:56:30	2019-07-31 14:55:46	生产环境	发布测试 查看参数 修改配置 删除任务 查看推送明细

Below the task list is a 'Run Instance' (运行实例) table:

采集任务名称	运行实例ID	目标Bucket	运行开始时间	运行结束时间	数据来源	数据目的	文件个数	运行时长(ms)	实例状态
页面上传(默认)	88	buckettest1-test-1259417666	2019-07-31 14:33:00	2019-07-31 14:33:01		/zhongbin/22	1	333	运行成功
页面上传(默认)	87	buckettest1-test-1259417666	2019-07-31 14:32:17	2019-07-31 14:32:17		/zhonbin/111	1	74	运行失败
页面上传(默认)	86	buckettest1-online-1259417666	2019-07-31 14:31:19	2019-07-31 14:31:19	0731test_20190731_0001.dat	cos2hive_test	1	0	运行成功
页面上传(默认)	78	buckettest1-online-1259417666	2019-07-30 10:17:17	2019-07-30 10:17:17		/zhongbin/730	1	226	运行成功

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？



* 采集名称: 仅支持中文、英文、数字、下划线、连字符，最大50字符

采集说明: 最大支持500字

* 目标kafka: dgKafka

* Topic名称: 请选择

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？



* 采集任务名称: 仅支持中文、英文、数字、下划线、连字符，最大50字符

采集任务说明: 最大支持500字

* 目标cos名称: cos

* 目标Bucket: uat-bucket13

* 选择数据交换接口: 请选择数据交换接口

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。