目录

产品概述

数据采集是大数据云平台内外部数据传输的桥梁，可以将云外数据安全有序的接入到大数据云平台内部。采集模块支持对文件、文件夹、外部kafka、自定义source等进行采集，将分散在各地的数据方便快捷的采集到大数据云平台，同时可通过流计算组件进行实时消费处理。

名词解释

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。

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。

行业场景与资源

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。

产品优势

统一采集管理

支持多种采集任务统一管理，采集配置一件修改，采集agent实时监控，采集任务一键暂停与恢复。

采集方式多样

支持文本文件数据流式采集，支持Oracle/MySQL等关系型数据库流式采集，支持http方式的数据采集。

保证数据质量

采集数据自动重发机制，支持数据查重，支持数据格式校验，支持采集数据分布式存储。

数据投递

支持将采集到的数据直接投递到ES中。

产品功能

统一采集管理

支持针对不同类型的采集任务，通过统一的方式来进行管理；提供数据采集任务的创建，查询，采集任务的启动与停止等功能的服务，包括：创建数据采集任务、查询采集任务状态、采集任务启动与停止、采集状态上报等功能。

流式数据采集

流式采集任务支持将文件、外部Kafka、文件夹、自定义source、oracle数据库、MySQL数据库等数据，实时采集至大数据云

平台的Kafka中。

流式投递

支持将数据实时采集到kafka，并可实时投递到ES等数据库中。

库表创建

新建Topic（用于流式采集）

元数据管理 > 库表管理 > 设计态 > 新建Topic

 注意：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：

```
{
   "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,
              }
}
```

创建样例如下图所示：  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： 

发布到测试

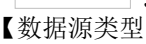
发布测试时，需指定此Topic归属的数据源。  发布测试完毕后，可在测试环境中筛选查看。

发布到生产

Topic发布到生产后，可在生产环境使用。

新建Object（用于批量采集）

登记bucket

1.新建Object前，需要为当前项目绑定可用的bucket。进入项目专属开发页面，在左侧导航栏依次点击**数据管理 > 元数据管理 > 库表管理**，随后点击**设计态 > Bucket > 登记bucket**。  2.在弹出页面中，填写创建bucket的各项属性。【bucket名称】为用户自定义，【所属项目】选择要创建在哪个项目下。配置完后点击确定。  3.在bucket列表中可以看到新建的bucket。点击操作列**发布**，将该bucket发布到测试环境。 4.在弹出页面选择【数据源类型】为ks3，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Object

1.依次点击左侧导航栏**数据管理** > **元数据管理** > **库表管理**，随后点击**设计态** > **Object设计** > **创建Object**。 2.在新弹出页面填写Object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**。 3.创建Object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**。 4.创建Object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，点击**完成**结束配置。

发布到测试

1.点击操作列**发布到测试**，以将该接口发布到测试环境。 2.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Object命名的表。

发布到生产

1.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。 2.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Object对应的表。

流式采集

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（采集开发页面的任务运行在测试环境中）。

流式采集任务可将文件、外部Kafka、文件夹、自定义source、oracle数据库、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：任务增/删/改，支持采集工具下载、采集明细查询、数据投递等操作。

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oracle、MySQL数据库的流式数据库采集（支持批量创建采集任务）。 流式采集任务大体可通过2种方式启动运行：

启动方式	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

流式采集-流式Agent采集

新建流式Agent采集

创建采集任务

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

1.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：	参数名称	说明
采集名称	支持中文、英文、数字、下划线，最大50字符。	
目标 Topic	待采集数据需要写入的Topic，支持下拉选择，可选择该项目下有权限的所有Topic。	
异常数据 Topic	当指定错误队列时，格式异常的数据会写入异常Topic下，支持下拉选择，可选择该项目下有权限的所有Topic（异常 Topic 不能和目标Topic选择同一个）（备注：异常Topic在创建时，只需要指定一个string类型的字段即可）	

2.选择 Topic 时，支持对 Topic 字段信息进行预览。 3.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、自定义source四种类型的数据。

文件采集

支持同时采集多个文件，配置参数说明如下：	参数名称	说明
待采集文件	待采集的数据路径及文件名，点击页面上的“+”号，可添加多个路径，实现多文件的采集。	

待采集文件：待采集的数据路径及文件名，点击页面上的“+”号，可添加多个路径，实现多文件的采集。

文件夹采集

可以采集文件夹下的所有文件，配置参数说明如下：	参数名称	说明
待采集文件夹	待采集的文件夹的完整路径。	

Kafka采集

支持同时采集多个 Topic 的数据，配置参数说明如下：	参数名称	说明
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。	

自定义source

自定义 source 需要在依赖线下开发的 JAR 包，通过 source 类路径名称实现采集任务和 JAR 包的关联：	参数名称	说明
Source 类路径	填写Kafka的IP：端口号，多个时使用“，”分隔。	

高级配置

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。

1.缓存配置 点击类型选择后面的下拉框，弹出memory和file两个选项。	参数名称	说明
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。

2.传输配置 缓存配置下方是传输配置。点击传输配置右侧展开按钮，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

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上方的公共部分：【下载通用采集工具及接口规范】处，下载采集任务的 Agent。 下载采集工具后，上传至各个采集节点，进入解压缩后的目录下，执行start.sh，即可在本地启动 agent，开始采集流式任务。 备注：采集开发页面的任务运行在测试环境中。

数据预览

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** > **数据地图** > **数据目录** > **技术元数据** > **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/采集明细

点击任务列表的采集明细，可查看每个 Agent 的具体情况，并进行：暂停、恢复、停止、升级、删除等操作。

操作名称	说明		
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（任务停止状态，才会出现删除按钮）。
升级	采集任务有升级/更新时，可点击“升级”对 Agent 配置文件进行升级。（采集任务的 Agent 信息有修改时，才会出现“升级”按钮并支持更新操作）。		
备注	自定义 source 类型的任务，不能进行暂停、恢复、升级操作。		

流式采集-API采集

创建采集任务

点击**新建采集** > **流式数据采集** > **API采集**，创建采集任务。

配置基本信息

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：

参数名称	说明

采集名称	支持中文、英文、数字、下划线，最大50字符。
目标 Topic	待采集数据需要写入的 Topic，支持下拉选择，可选择该项目下有权限的所有 Topic。
异常数据 Topic	当指定错误队列时，格式异常的数据会写入异常 Topic下，支持下拉选择，可选择该项目下有权限的所有 Topic（异常 Topic 不能和目标 Topic 选择同一个）（备注：异常 Topic 在创建时，只需要指定一个 string 类型的字段即可）
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

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》。 **备注：** 采集开发页面的任务运行在测试环境中。

查看 API 参数

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。

流式采集-流式数据库采集

创建采集任务

点击**新建采集 > 流式数据库采集**，创建采集任务。

配置基本信息

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：	参数名称	说明
采集名称	支持中文、英文、数字、下划线，最大 50 字符。	
目标 Topic	待采集数据需要写入的 Topic，支持下拉选择，可选择该项目下有权限的所有 Topic 。	
异常数据 Topic	当指定错误队列时，格式异常的数据会写入异常 Topic下，支持下拉选择，可选择该项目下有权限的所有 Topic（异常 Topic 不能和目标 Topic 选择同一个）（**备注：** 异常 Topic 在创建时，只需要指定一个 string 类型的字段即可）	

选择 Topic 时，支持对 Topic 字段信息进行预览。 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、Oracle 的数据。

MySQL采集说明

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。

Oracle采集说明

流式数据库 MySQL 采集使用的是 OGG+Flume 的方式采集数据，配置 Agent 信息后，下载对应 Oracle 源端 OGG、目标端 OGG、Agent 后，在本地部署启动。

特别说明

1.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。

```
{
  "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"
   }
}
```

2.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。

```
{
   "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/采集明细

点击任务列表的采集明细，可查看每个 Agent 的具体情况，并进行：暂停、恢复、停止、升级、删除等操作。	参数名称	说明
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（任务停止状态，才会出现删除按钮）。	
升级	采集任务有升级/更新时，可点击**升级**对 Agent 配置文件进行升级/更新（**备注：** （1）采集任务的 Agent 信息有修改时，才会出现【升级】按钮并支持更新操作；（2）MySQL 数据库采集，不支持 Agent 更新）。	

流式采集-批量创建流式数据库采集

点击**新建采集 > 流式数据库采集（Excel 批量创建）** 创建采集任务。

配置通用信息

批量创建数据库采集任务的通用配置信息和【新建流式数据库采集】中的 Agent 信息基本一致，但不需要配置待采集的表信息。

配置批量信息

操作步骤： 1.下载 Excel 模板，并按照模板在线下完成批量任务信息填写。 2.上传批量任务列表，上传时，会对列表信息进行校验： (1)目标 Topic 和异常 Topic 不能相同。 (2)目标 Topic 和异常 Topic 需要在数据管理中先进行创建（可匹配数据管理中的 Topic 列表）。 (3)采集任务名称、采集数据表名称、目标 Topic 必填。

下载采集工具

备注： 采集开发页面的任务运行在测试环境中。

任务上线申请与生产运行

任务上线申请

在采集任务列表点击**申请上线**，可申请将采集任务发布至生产环境。

Agent-文件采集任务上线

申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

Agent-kafka 采集任务上线

申请上线时，需要填写生产环境待采集的 kafka 地址。 申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

Agent-文件夹采集任务上线

申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

Agent-自定义 source 采集任务上

申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

Agent-文件采集任务上线

申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

Agent-Oracle 采集任务上线

申请上线时，需要填写生产环境待采集的 Oracle 数据库地址和 Agent 监听地址。 申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

Agent-Mysql 采集任务上线

申请上线时，需要填写生产环境待采集的 Mysql 数据库地址和 Mysql 用户名密码。 申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

API 采集任务上线

申请上线时，需要填写生产环境待采集的 Mysql 数据库地址和 Mysql 用户名密码。 申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

任务上线审核

点击**发布管理 > 发布审批**，可在【未审批】列表中，对申请上线的任务进行审核，审核时，可选择审核通过或审核拒绝。审批通过后，可在【已发布列表】中，查看任务信息，或进行下线操作。

任务生产运行

上线启动后，可在运维中心 > 数据采集 > 生产任务中查看任务列表。此处操作与查看Agent列表/采集明细操作一致，区别在于此处的参数只能查看，不能编辑修改。 点击任务列表的采集明细，可查看每个 Agent 的具体情况，并进行：暂停、恢复、停止、升级、删除等操作。

操作名称	说明
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（任务停止状态，才会出现删除按钮）。
升级	采集任务有升级/更新时，可点击“升级”对 Agent 配置文件进行升级/更新（**备注：** 1.采集任务的 Agent 信息有修改时，才会出现**升级**按钮并支持更新操作；2.MySQL 数据库采集，不支持 Agent 更新）。

备注：自定义 source 类型的任务，不能进行暂停、恢复、升级操作

流式投递

操作步骤

点击左侧导航栏**流式采集**，进入页面后点击操作列**投递管理**，即可进入流式投递功能模块。

新建流式投递任务

点击新建投递任务，进入投递任务的参数配置页面。 **参数配置说明如下：**

参数名称	说明
投递任务名称	支持中文、英文、数字、下划线，最大 50 字符。
待投递 Kafka/Topic	流式采集任务的目标Kafka/Topic，默认，用户无需填写。
Kafka 消费位置	支持 Earliest、Latest 两种选择，选择 Earliest 会投递 kafka 所有历史数据及最新数据，选择 Latest 只会采集最新数据。
目标数据源类型	当前只支持 Elasticsearch。
目标数据源名称	目标数据投递写入的数据源。
目标索引	结果数据投递写入的索引。
目标数据表	结果数据投递写入的表（Type）。
数据写入批次大小	每批次写入数据的最大条数阈值（数据将分批次写入），最大限制 10 万条。
写入时间间隔（ms）	数据写入目标表的最大等待时间，0 或者不填表示不启用，最大限制 1 亿条（1 天）。

测试运行

点击投递列表中的**发布测试**，并按需选择资源后，可对投递作业进行在测试环境运行。为保障投递作业正常运行，需保证DCU资源数必须大于2。

查看测试实例

发布到测试后，可在运维中心 > 数据采集 > 测试实例中查看任务列表。可进行启动、终止、发布生产、资源修改、查看Flink UI、查看日志等操作。

操作命令	说明
启动/终止	终止后，流式投递作业会终止，可点击**启动**重启流式投递作业。启动时，可选择从【当前时间启动】或从【上一次作业终止时间启动】。（上一次作业终止时间即：上一次手动停止流式投递作业的时间。）
发布生产	可将作业发不到生产环境，发布时需选择版本。
资源修改	作业终止后，可进行资源修改操作。
查看Flink UI	可通过Flink UI查看作业运行进度与状态。（查看Flink_UI前，需要提前配置部署集群的hosts，否则页面无法正常跳转。）
查看日志	查看作业日志与调度日志。（作业日志：主要用于定位作业运行中遇到的异常问题。调度日志：主要用于定位作业是否正常启动等。）

任务上线发布

在流式投递作业列表点击**发布生产**，选择要发布的版本后可申请将流式投递作业发布至生产环境。

任务上线审核

从屏幕左下角点击进入**发布管理**模块，点击**发布审批**进入页面，可在【未审批】列表中，对申请上线的任务进行审核，审核时，可选择审核通过或审核拒绝。 审批通过后，可在【已发布列表】中，查看任务信息，或进行下线操作。

任务上线启动

上线审核通过后，可在**运维中心 > 任务运维 > 数据采集任务**中查看任务列表。对于流式投递任务，点击**上线启动**，可启动。

任务上线运行

任务上线启动后，可在**运维中心 > 数据采集 > 生产任务**中查看任务列表，并进行启动/终止、资源修改、查看 Flink UI、查看日志等操作。

批量采集配置

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KS3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

点击批量采集，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：

参数名称	说明
采集名称	支持中文、英文、数字、下划线，最大 50 字符。
目标ks3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、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

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KS3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

批量采集任务可以通过： 1.文件上传工具包。 2.SDK 启动。 关于文件上传工具包、SDK 的的具体使用方法，可在【批量采集-通用下载】页面下载。

文件拉取

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KS3。 2.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】两种方式，两种方式均需指定推送的目标KS3、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【执行周期】。具体如下图所示：

查看批量采集运行实例

点击**采集明细**按钮，会弹出表单，查看批量采集的运行实例，并且可以查看目标数据路径。 备注：当离线计算的作业依赖批量采集任务（一般依赖文件拉取任务）时，【目标-路径】页面的处理事件名字段，可作为离线计算的依赖监听值。

通过页面进行文件上传

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（KS3）即可。 页面文件上传的任务默认展示在任务列表的首行，点击操作列中的**采集明细**，可查看每次文件上传的信息。

任务上线申请

文件推送任务上线

- 在批量采集任务列表点击**申请上线**，可申请将采集任务发布至生产环境。
- 申请上线后，需项目管理员在**发布管理 > 发布审批**中进行审批。

文件拉取任务上线

文件拉取需要先把任务发布到测试进行测试验证以后，才可以申请上线。 点击批量采集列表中的**发布测试**，并按需选择资源后，可对文件拉取作业进行在测试环境运行。

任务上线审核

从屏幕左下角进入【发布管理】模块，点击**发布审批**进入页面，可在【未审批】列表中，对申请上线的任务进行审核，审核时，可选择审核通过或审核拒绝。 审批通过后，可在【已发布列表】中，查看任务信息，或进行下线操作。

任务上线启动

- 文件推送任务在上线审核通过后，就已经发布到生产环境了，可直接在**运维中心** > **数据采集** > **生产任务** > **任务管理** > **批量采集**中查看。
- 文件拉取上线审核通过后，可在**运维中心** > **任务运维** > **数据采集任务**中查看任务列表。对于文件拉取任务，点击**上线启动**，可启动。

任务生产运行

上线启动后，可在**运维中心** > **数据采集** > **生产任务**中查看任务列表。

子账号数据采集权限配置

子账号获取采集配置权限

通过Agent、API或推送工具启动采集任务时，可使用子账号的AK/SK进行鉴权验证。使用子账号前， 需要为子账号配置采集相关的权限，具体操作步骤如下：

1.进入**访问控制** > **权限管理** > **策略**页面，选择**自定义策略**，点击**新建策略**。

2.在弹出的页面中，选择**可视化配置**，并点击下方的**添加策略语句**按钮。

3.在弹出的窗口中，分别选择：

- 产品服务：和采集相关的产品服务有：“大数据云数据采集”、“大数据云平台”、“大数据云运维中心”，创建策略时可分别选择。
- 操作名称：所有操作
- 资源范围：所有资源

点击下方的**确认**，创建策略语句成功。

4.创建策略语句成功后，点击**创建策略**按钮，完成策略创建，并可在自定义策略列表中查看创建的策略信息。

5.重复以上步骤，分别创建三个策略，如：

- cbr-datagather，创建策略语句时，选择“大数据云数据采集”产品服务。
- cubricks，创建策略语句时，选择“大数据云平台”产品服务。
- cbr-opscenter，创建策略语句时，选择“大数据云运维中心”产品服务。

点击**创建策略**

6.通过点击策略列表中策略的**关联对象**操作。

在弹窗的【被授权主体】列表，选择需要关联的子账号

获取子账号的AK/SK

1.从页面左侧的菜单进入【访问控制】模块，点击**子用户**进入子用户页面。

2.点击子账号名称，进入【用户详情】页面。点击**创建密钥**，可创建子账号的AK/SK，创建成功后，请务必将AK/SK的凭证下载下来，并妥善保存。

OGG插件

OGG（Oracle GoldenGate）是一个基于日志的结构化数据备份工具，一般用于Oracle数据库之间的主从备份以及Oracle数据库到其他数据库（DB2, MySQL等）的同步。下面是Oracle官方提供的一个OGG的整体架构图，从图中可以看出OGG的部署分为源端和目标端两部分组成，主要有Manager，Extract，Pump，Collector，Replicat这么一些组件。

- Manager：在源端和目标端都会有且只有一个Manager进程存在，负责管理其他进程的启停和监控等；
- Extract：负责从源端数据库表或者事务日志中捕获数据，有初始加载和增量同步两种模式可以配置，初始加载模式是直接将源表数据同步到目标端，而增量同步就是分析源端数据库的日志，将变动的记录传到目标端，本文介绍的是增量同步的模式；
- Pump：Extract从源端抽取的数据会先写到本地磁盘的Trail文件，Pump进程会负责将Trail文件的数据投递到目标端；
- Collector：目标端负责接收来自源端的数据，生成Trail文件；
- Replicat：负责读取目标端的Trail文件，转化为相应的DDL和DML语句作用到目标数据库，实现数据同步。

常见问题

1. **数据采集组件可以将数据采集到什么地方？**

   数据采集组件支持流式采集、批量采集两种方式，其中：

   - 流式采集可以将数据采集到Kafka的Topic中；
   - 批量采集可以将数据采集到KS3中；

2. **流式采集支持采集什么数据？**

流式采集任务可将文件、外部Kafka、文件夹、自定义source、oracle数据库、MySQL数据库中的数据，实时采集至大数据云平台的Kafka中。

3. **批量采集支持哪几种采集方式？**

- 批量采集可以通过“文件上传工具包”将用户客户端本地的多个文件、或选定文件夹下的所有文件推送至KS3；
- 批量采集还支持将FTP下的文件周期/单次拉取至KS3

4. **可以通过页面直接将文件上传文件至KS3中吗？**

可以，进入**批量采集** > **页面上传文件**功能页面，可以将选中的多个文件一次性上传至选定的KS3目录下。可在采集开发页面将将本地文件上传至测试环境，可在**运维中心** > **数据采集** > **生产任务** > **任务管理** > **批量采集**中，将文件上传至生产环境。

5. **流式采集Agent可以部署在多个客户端吗？**

流式采集Agent可以部署在多个客户端，部署启动后，可以在采集任务列表点击**Agent列表**查看部署Agent的客户端IP、hostname。

6. **启动流式采集任务后，如何查看哪些客户端正在进行采集任务？**

- 启动采集任务后，可在任务列表通过【Agent存活数】字段查看存活的Agent数量；
- 点击采集任务列表的【Agent列表】字段，可查看具体的客户端情况，其中运行状态为【运行中】的客户端正在进行采集任务

7. **修改流式Agent采集的Agent信息后，没什么没有生效？**

修改流式Agent采集的Agent信息后，需要在Agent列表中，选择需要升级的客户端，暂停采集任务后，点击**更新**按钮，进行更新。

8. **为什么将Agent停止后，没办法在页面重新启动？**

Agent停止后，无法在平台页面上重启采集动作，需要在客户端上通过Agent进行启动。

9. **批量采集任务如何查看采集历史？**

无论是通过页面上传的文件，还是通过上传工具上传的文件，或者从FTP拉取的文件，都可以通过批量采集任务列表中的【采集明细/拉取明细】功能，查看历史的采集/上传/拉取情况。