DATAPIPELINE
实时数据融合产品
用户操作手册

1.0

版本号: 1.0
出版日期: 2020年7月

DATAPIPELINE

文档名：DataPipeline_DIP_V2.8_User_Guide_CN_R1.0.docx

第一版（2020年7月）

此版本适用于DataPipeline提供的数据融合产品以及所有后续的发布版本和修改，除非在新版本或技术新闻中另有说明。DataPipeline会定期对产品说明书的内容进行修订，任何修订均会在后续版本或技术新闻中报告。

您可以向您的DataPipeline销售人员或服务人员提出有关DataPipeline其他出版物的请求。

DataPipeline可能拥有涉及本文档主题的专利或正在申请的专利。提供本文档并不授予您使用这些专利的任何许可。

如有任何疑问与建议请联络：
product@datapipeline.com
DataPipeline产品部
北京数见科技有限公司
中国 北京市 海淀区王庄路1号
清华同方科技大厦D座东楼1801室

DATAPIPELINE

前言

PREFACE

关于本文档 About This Guide

本文档包含了DataPipeline实时数据融合产品的用户操作相关的全面说明，包含产品基本功能介绍，支持的数据源与目的地节点类型，数据节点、数据链路、数据任务的配置、管理等相关内容。

本文档以PDF与HTML两种方式提供，在开始使用本文档之前，请确保您所使用的是DataPipeline产品中心中此文档的最新版本，获取地址：www.datapipeline.com

目标受众 Who Should Read This Guide

本文档的目标受众包括利用DataPipeline实时数据融合产品构建实时数据融合平台的工程技术人员，利用DataPipeline实时数据融合产品配置管理实时数据融合任务的数据工程师及负责DataPipeline实时数据融合产品日常运行维护的专业运维人员。

硬件与软件要求 Hardware and Software Requirements

本文档假定用户正在使用DataPipeline实时数据融合产品，同时还可能在使用本文档中列出的受支持的数据源与数据目的地节点类型的数据库产品。用户应自行准备部署DataPipeline实时数据融合产品的硬件环境，具体要求请参阅《DataPipeline

实时数据融合产品安装部署手册》。用户应自行准备配合DataPipeline实时数据融合产品运行管理及故障处理可能用到的相关软件，包括但不限于操作系统、数据库管理工具、网络终端等。

相关文档 Related Documentation

DataPipeline实时数据融合产品V2.8安装部署手册

DataPipeline实时数据融合产品V2.8故障处理手册

DATAPIPELINE
实时数据融合产品

目录

图表目录

1. 产品概览及系统要求 PRODUCT OVERVIEW AND SYSTEM REQUIREMENTS

1.1. 产品概览 ABOUT DATAPIPELINE

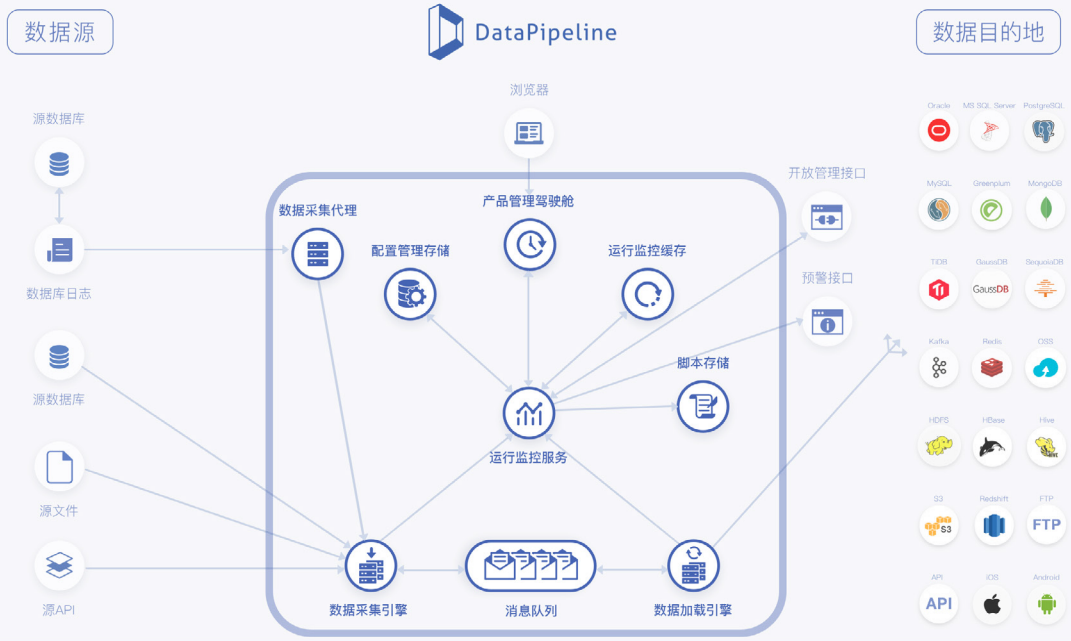
DataPipeline数据融合产品通过多年在数据融合技术领域的积累，支持Oracle、MySQL、Microsoft SQL Server及PostgreSQL等数据库的实时增量数据捕获，基于异构语义映射实现异构数据实时融合，帮助用户提升数据流转时效性，降低异构数据融合成本。在支持传统关系型数据库的基础上，对大数据平台、国产数据库、云原生数据库、API及对象存储也提供广泛的支持，并在不断扩展。

DataPipeline数据融合产品致力于为用户提供企业级数据融合解决方案，为用户提供统一的平台，同时管理异构数据节点的实时同步与批量数据处理任务，采用分布式集群化部署方式，可水平垂直线性扩展，保证数据流转稳定高效，让客户专注数据价值释放。

DataPipeline数据融合产品架构中关键组件的说明如下：

- **产品管理驾驶舱**：用户配置，监控和管理各种数据融合任务的管理驾驶舱。用户可以通过管理平台注册节点、构建链路、配置任务，也可以通过管理平台监控融合任务的执行情况，按需调整各项策略配置与限制参数，控制管理任务状态。不仅记录用户的各项配置与自定义脚本，同时，管理平台还允许用户监控各类事件消息、数据节点连接状态，数据链路级别的任务执行情况、具体到数据对象的任务执行状态、延迟情况及其他统计信息。管理平台与运行监控及融合引擎完全解耦，数据融合任务配置完毕并激活后，即使终止管理平台的服务，也不会影响数据源与数据目的地之间的数据融合任务的执行。
- **运行监控服务**：依据用户各类配置对融合引擎进行调度管理与信息收集的中间层。负责依据任务的执行配置对任务状态进行调度调整。在任务执行过程中出现数据源结构变化、错误数据、各项策略配置

图 1 DATAPIPELINE数据融合产品关键组件



给出执行预案。收集任务执行过程中产生的各类信息，依据任务状态监控、运行情况统计、日志策略及预警策略等各项配置进行信息分类、统计、记录、推送及预警。

- **数据采集引擎**：执行数据融合任务的核心引擎。负责采集不同类型数据节点实时增量数据。通过各类连接器实现各类数据源节点的日志解析、实时增量数据采集、结构检测、语义转化等工作。感知数据融合过程中的各类事件与情况，并基于自身记录的及运行监控给出的预案进行相应处理。
- **数据加载引擎**：执行数据融合任务的核心引擎。负责清洗、融合不同类型数据节点实时增量数据，并实时加载到数据目的地。针对不同的数据节点类型提供相适应的、准确的、高性能的增量数据加载。同时负责按需执行结构变化策略与错误数据的发现与处理。感知数据融合过程中的各类事件与情况，并基于自身记录的及运行监控给出的预案进行相应处理。
- **消息队列**：传输、缓存、持久化数据采集节点推送的实时增量数据并供数据加载节点使用。
- **数据采集代理**：部署安装于数据源节点的日志解析工具，负责解析日志文件采集增量数据，并将数据发送至数据采集引擎。
- **运行监控缓存**：包括任务状态缓存、运行事件缓存、错误队列缓存在内的任务执行过程中各个任务的状态信息、各类运行事件及出现的错误数据会被缓存在缓存中再交由各个相关管理模块使用，以保证任务执行的效率，减少运行监控、事件处理及错误数据处理对任务执行造成的资源消耗。
- **脚本存储**：存储用户自定义脚本的文件系统，通过代码分发与动态加载实现分布式架构下的用户自定义脚本执行。

- **配置管理存储**：

1. 系统配置存储：存储系统信息、用户信息、权限信息、节点信息、链路信息、任务信息及其他各类配置信息的数据库。
2. 监控信息存储：存储任务执行过程中的各类性能指标信息并存储运行监控要求的各类监控要求的加工计算指标。
3. 错误队列存储（支持用户按照数据链路指定外部存储）：存储错误数据与错误数据描述信息。
4. 用户日志存储（支持用户按照数据链路指定外部存储）：存储任务状态日志、任务报错日志、任务性能日志、配置变更日志、数据处理日志等用户定制的日志信息。
5. 预警信息存储（支持用户指定外部存储）：存储基于用户配置的预警规则收集的预警事件信息。

1.2. 支持的数据源与数据目的地 SUPPORTED SOURCES AND TARGETS

使用DataPipeline产品之前，您需确定当前版本支持的数据源与数据目的地满足您的需求。
下表列出了DataPipeline支持的可以作为数据源的数据节点类型及版本：

作为数据源的数据节点类型及版本

数据源节点类型	支持版本
Oracle	读取模式支持10G、11G版本及12C非多租户版本 Agent读取模式支持11G、12C，灾备架构只支持Oracle ADG，不支持只读库； LogMiner读取模式支持11.2.0.4及12C非多租户版本，不支持Oracle RAC，不支持灾备架构从库，不支持Oracle只读库。
MySQL	JDBC读取模式支持5.5、5.6、5.7、8.0； Binlog读取模式支持5.5、5.6、5.7、8.0，不支持只读库。
MS SQL Server	JDBC读取模式支持2008、2012、2014版本； Change Tracking读取模式支持2008、2012、2014版本的单实例非只读数据库。
PostgreSQL	JDBC读取模式支持9.4、9.5、9.6、10.4版本； wal2json读取模式支持9.4、9.5、9.6、10.4版本，不支持分布式架构，不支持只读库。
Apache Kafka	0.9.x、0.10.x、0.11.x、1.x.y、2.x.y

表格1 DataPipeline数据融合产品支持数据源节点类型与版本

下表列出了DataPipeline支持的可以作为数据目的地的数据节点类型及版本：

作为数据目的地的数据节点类型及版本

数据目的地节点类型	支持版本
Oracle	10G、11G版本及12C非多租户版本，不支持只读库；
MySQL	官方版5.5、5.6、5.7、8.0，不支持只读库；

MS SQL Server	2008、2012、2014，不支持只读库
Apache Kafka	0.8.x、0.9.x、0.10.x、0.11.x、1.x,y、2.x.y
Greenplum	4.x.y、5.x.y、6.x.y
TiDB	2.x.y
SequoiaDB	3.2.1
Amazon RedShift	latest

表格2 DataPipeline数据融合产品支持数据目的地节点类型与版本

1.3. 推荐部署环境要求 RECOMMENDED DEPLOYMENT ENVIRONMENT REQUIREMENTS

DataPipeline实时数据融合产品支持多种部署方式，包括物理机部署、虚拟机部署、容器部署，不同的部署方式对服务器要求不尽相同。下表是推荐的部署环境要求：

名称	要求	备注
服务器数量	集群版，三节点或以上；	支持单机部署，无法支持高可用，不推荐；
硬盘配置	8核CPU	此处为最低配置，需根据实际需要评估
	32G内存	此处为最低配置，需根据实际需要评估
	200G SSD硬盘	此处为最低配置，需根据实际需要评估
MS SQL Server	JDBC读取模式支持2008、2012、2014版本； Change Tracking读取模式支持2008、2012、2014版本的单实例非只读数据库。	此处为最低配置，需根据实际需要评估
网络/端口	管理控制台向外开放80端口； 集群间所有端口可达；	与数据源及数据目的地的特定端口可达

表格3 DataPipeline数据融合产品推荐部署环境要求

在安装DataPipeline实时数据融合产品之前，请确保您选择的系统满足操作系统、容器版本等软件要求与CPU、内存、磁盘等硬件要求。除此以外您还需要考虑诸如数据库配置要求、系统用户账户要求及网络环境要求等方面的内容，具体内容请参阅《DataPipeline实时数据融合产品部署安装手册》或联络DataPipeline工程师。

2. 管理控制台使用说明 MANAGEMENT CONSOLE INSTRUCTIONS

2.1. 管理数据节点 SETTING UP DATA NODE

数据节点是数据的原始载体。「数据节点」可以是数据库、文件系统、数据仓库、文件、应用，一切存储数据的载体都能成为「数据节点」。

在数据融合过程中，数据节点既可以做数据源节点也可以做数据目的地节点，在您通过数据节点管理注册一个数据节点时，它是没有被分配角色的，数据链路的存在才会赋予相关数据节点「数据源节点」和「数据目的地节点」的角色属性。

支持的数据节点类型与版本详见"1.2. 支持的数据源与数据目的地 Supported Sources and Targets"

◎ **数据节点基础配置**

数据节点基础配置是支持数据节点在系统中使用的基础配置，数据节点的连接信息、身份认证及个性化的连接参数，包括地址、端口、不同认证方式所需的信息等都属于基础配置。在新建节点过程中，完成基础配置即可使用该数据节点。

◎ **数据节点策略配置**

数据节点策略配置是数据融合任务在执行过程中，出现不同的运行事件及状态变化时的应对策略与管理规则。

◎ **检查点策略配置**

您可以通过数据节点策略配置中的检查点策略配置MySQL、Oracle、SQL Server、PostgreSQL四种类型数据节点的检查频率和检查范围，方便数据任务定位读取起点。

详见："2.1.1.2.1. 检查点策略配置 Setting up Check Point Policy"

◎ **语义映射策略配置**

您可以通过数据节点策略配置中的语义映射策略配置来明确数据节点与不同数据节点类型之间的数据类型、索引及特性的具体映射，以保证数据传输一致性及确定性。

详见： "2.1.1.2.2. 语义映射策略配置 Setting up Semantic Mapping Policy"

◎ **数据节点基本管理**

在数据节点列表页可以新增数据节点，配置数据节点的基础连接信息，并对数据节点进行挂起、激活状态调整操作、删除数据节点操作和设置重要节点操作。您也可以批量删除、挂起、激活数据节点。

新增数据节点

您可以通过基本配置来定义节点的名称、类型、描述、状态、重要程度、参与人以及不同类型数据节点的相关连接参数等信息。

详见："2.1.1. 新增数据节点 Adding a Data Node"

编辑数据节点

您可以在数据节点详情界面可以编辑数据节点的名称、描述、基础连接信息以及检查点策略和语义映射策略配置信息。

详见："2.1.2. 修改数据节点 Editing a Data Node"

删除数据节点

您可以数据节点列表页和数据详情界面删除数据节点。

详见："2.1.5. 删除数据节点 Deleting a Data Node"

激活数据节点

您可以在数据节点列表页和数据详情界面通过激活节点来调整节点的状态。

详见："2.1.3. 激活数据节点 Activating a Data Node"

挂起数据节点

您可以在数据节点列表页和数据详情界面通过挂起节点来调整节点的状态。

详见："2.1.4. 挂起数据节点 Deactivating a Data Node"

2.1.1. 新增数据节点 Adding a Data Node

关于此功能

新增数据节点的时候，您需要配置节点的基本信息，包括数据节点，定义节点的名称、描述、类型以及该类型节点的连接配置项等信息。点击新建节点弹窗的顶端「查看配置规则」可以查看不同数据节点的配置要求，包括数据节点的版本要求、权限要求和其他要求。

操作步骤

1. 确保您是系统管理员并拥有新建数据节点的权限。
2. 在数据节点列表页点击「新建数据节点」按钮，弹出新增数据节点弹窗。首先配置数据节点的名称、描述和类型。
   - 名称：是该节点在系统的唯一标识。
   - 描述：您可添加描述信息，例如描述节点功能便于使用。
   - 类型：在已支持的数据节点类型中选择，支持的数据节点类型与版本详见1.2支持的数据源与数据目的地 Supported Sources and Targets。
3. 当用户选择数据节点类型之后，我们会根据用户选择的不同类型数据节点显示不同的配置模板。具体信息如下表：

数据节点类型	配置项
Oracle	• 版本 • 服务器地址 • 端口 • 数据库 • 用户名 • 密码 • Agent读取模式（Kafka 地址、Kafka Topic 前缀地址） • 连接参数
MySQL	• 版本 • 服务器地址 • 端口 • 数据库 • 用户名 • 密码 • 连接参数
MS SQL Server	• 版本 • 服务器地址 • 端口 • 数据库 • Schema目录 • 用户名 • 密码 • 连接参数
PostgreSQL（仅用作数据源节点）	• 版本 • 服务器地址 • 端口 • 数据库 • 用户名 • 密码 • 连接参数
Apache Kafka	• 版本 • 服务器地址 • ZooKeeper地址 • 数据格式 • 连接参数
Greenplum（仅用作数据目的地节点）	• 版本 • 服务器地址 • 端口 • 数据库 • Schema目录 • 用户名 • 密码 • 连接参数
TiDB（仅用作数据目的地节点）	• 版本 • 服务器地址 • 端口 • 数据库 • 用户名 • 密码 • 连接参数
SequoiaDB（仅用作数据目的地节点）	• 版本 • 服务器地址 • 端口 • 数据库 • 用户名 • 密码 • 连接参数
Amazon Redshift （只能做数据目的地节点）	• Redshift配置 • 服务器地址 • 端口 • 数据库 • 用户名 • 密码 • Schema（架构） • S3配置 • Bucket • 写入目录 • 地区 • 访问权限 • 连接参数

表格4 不同类型数据节点配置信息

4. 点击新建节点弹窗的底部「连接验证」按钮，可以进行数据节点基础信息的连接和权限验证。
5. 配置完相关类型节点的连接配置项之后，点击「保存」按钮，跳转到该数据节点详情界面。

下一步

您可以在数据节点详情界面进行参与人设置、重要节点设置以及数据节点的检查点策略和语义映射配置。

2.1.1.1. 数据节点-基础配置 Basic Configuration of Data Node

2.1.1.1.1. 配置Oracle数据节点基本信息 Basic Configuration of Oracle Data Node

在进行配置之前，请务必检查您的Oracle数据节点是否符合平台要求，参考如下：

版本要求
1. 数据源节点JDBC读取模式支持10G、11G版本及12C； 2. 数据源节点Agent读取模式支持11G、12C，灾备架构只支持Oracle ADG，不支持只读库； 3. LogMiner读取模式支持11.2.0.4及12C，不支持Oracle RAC，不支持灾备架构从库，不支持Oracle只读库； 4. 数据目的地节点支持Oracle 10G、11G版本及12C，不支持Oracle只读库。
用户权限要求
SELECT ANY TRANSACTION , LOGMINING , EXECUTE ON DBMS_LOGMNR_D , EXECUTE ON DBMS_LOGMNR , EXECUTE ON DBMS_FLASHBACK , CREATE SESSION , RESOURCE , SELECT ANY DICTIONARY , FLASHCK ANY TABLE , EXECUTE_CATALOG_ROLE ,CONNECT , LOCK ANY TABLE , SELECT ANY TABLE , SELECT ANY DICTIONARY , ALTER SYSTEM , EXECUTE ON DBMS_FLASHBACK ,以及相关视图和表的 SELECT权限
其他要求
1. 提供Oracle LogMiner； 2. 12c以下版本不需要LOGMINING权限； 3. 日志补全至少需要开启primary key、all、unique级别的日志级别； 4. Archive LOG保留时间建议>=72小时； 5. 不支持Oracle只读实例； 6. 要求需要同步的表拥有增量识别字段,用于同步增量数据； 7. 选择的字段必须为可排序,例如数字或时间类型,推荐的字段类型一般为随数据更新而自增的字段,如：更新序列号（例：SequenceID），更新时间戳 （例：UpdatedAt）； 8. 授予SELECT ON V_$SESSION权限用于处理死锁导致的任务中断。

表格5 Oracle数据节点配置要求

新建弹窗界面选择了Oracle节点类型之后，显示Oracle的基本信息配置模板：

- 版本：Oracle 10G、Oracle 11G、Oracle 12C
- 服务器地址：输入连接数据节点的域名或IP地址。例如192.168.2.11或instance1.oracle.example.com
- 端口：输入连接数据节点的端口值
- 数据库名称：要求连接数据库的名称
- 用户名和密码：输入可访问该数据节点的用户名和密码

◎ Agent读取模式：

» 点击「开启」Agent读取模式后，立即显示两个输入项：Kafka地址、Kafka Topic前缀名称

» Kafka地址：代表通过Kafka获取Oracle数据源数据

・允许只写主节点

・允许写多个节点（主、从），若主节点访问失败，可自动切换从节点访问

» Kafka Topic前缀：需要用户输入Topic前缀以区分不同的数据源下的同名表

> **注意事项**
>
> Oracle用户名和密码区分大小写，创建用户时如果用户名不加双引号会被Oracle以大写的形式保存在系统的字典表中；如果创建用户名的时候在用户名上加上双引号，那么则会把双引号内的用户名直接保存在数据库字典表中。DataPipeline实时数据融合产品在使用JDBC处理输入的用户名时会遵循Oracle的处理原则将小写的用户名改写成大写，Oracle接收到用户名之后会以大写用户名为条件在系统的字典表中进行查找，这时便无法查找到使用小写的用户名，因此我们建议用户在使用Oracle创建用户时使用大写用户名。

2.1.1.1.2. 配置MySQL数据节点基本信息 Basic Configuration of MySQL Data Node

在进行配置之前，请务必检查您的MySQL数据节点是否符合平台要求，参考如下：

版本要求
1. 数据源定时模式支持MySQL官方版5.5、5.6、5.7、8.0；
2. 数据源实时模式–Binlog支持MySQL官方版5.5、5.6、5.7、8.0，不支持只读库；
3. 目的地支持MySQL官方版5.5、5.6、5.7、8.0，不支持只读库；
用户权限要求
相关视图和表的 select 权限，replication slave，replication client，create object，insert，update，delete，drop
其他要求
1. Binlog修改为row模式: mysql binlog mode = 'row' binlog_row_image = 'full'（此参数在MySQL5.6及以上版本需要设置） 2. Binlog保留时间建议>=72小时； 3. 不支持只读从库； 4. 要求需要同步的表拥有增量识别字段,用于同步增量数据。选择的字段必须为可排序,例如数字或时间类型,推荐的字段类型一般为随数据更新而自增的字段,如:更新序列号 （例:SequenceID）,更新时间戳 （例: UpdatedAt）。

表格6 MySQL数据节点配置要求

新建弹窗界面选择了MySQL节点类型之后，显示MySQL的基本信息配置模板：

- 版本：MySQL 5.5、MySQL 5.6、MySQL 5.7
- 服务器地址：输入连接数据节点的域名或IP地址。例如192.168.2.11或datapipeline.com
- 端口：输入连接数据节点的端口值
- 用户名和密码：输入可访问该数据节点的用户名和密码

> **注意事项**
>
> 1. MySQL 的实时处理模式下，暂时无法读取字段类型为 geometry 的数据，具体请参照下表，如果存在对应类型的数据，请选择定时模式进行同步。

geometry字段	
point	multipoint
linestring	multilinestring
polygon	multipolygon
geometry	geometrycollection

表格7 MySQL数据节点不支持geometry类型字段

2. 目前MySQL数据节点编码类型仅支持 UTF-8，若数据节点存在Unicode越界字符，系统会进行重编码以写入，不会导致任务报错暂停。

2.1.1.1.3. 配置MS SQL Server数据节点基本信息 Basic Configuration of MS SQL Server Data Node

在进行配置之前，请务必检查您的MS SQL Server数据节点是否符合平台要求，参考如下：

版本要求
1. 数据源定时模式支持MS SQL Server 2008、2012、2014版本;
2. 数据源实时模式-CT支持MS SQL Server 2008、2012、2014版本的单实例非只读数据库;
3. 目的地支持MS SQL Server 2008、2012、2014版本，不支持MS SQL Server只读库

用户权限要求
select on table , view change tracking on table , view change tracking on schema , select on schema , create object , insert , update , delete , drop , truncate

其他要求
1. 如果default schema为dbo,可以直接授权select和view change tracking权限: grant select on schema::dbo to user; grant view change tracking on schema::dbo to user;
2. 如果主体模式不是dbo就需要按照表为单位进行view change tracking的授权, 比如: grant view change tracking on object::rpt.表名 to user;
3. MS SQL Server Change Tracking不支持只读数据库;
4. 要求需要同步的表拥有增量识别字段,用于同步增量数据;
5. 选择的字段必须为可排序,例如数字或时间类型,推荐的字段类型一般为随数据更新而自增的字段,如:更新序列号 （例:SequenceID）,更新时间戳 （例: UpdatedAt）

表格8 MS SQL Server数据节点配置要求

新建弹窗界面选择了MS SQL Server节点类型之后，显示MS SQL Server的基本信息配置模板：

- 版本：MS SQL Server 2008、MS SQL Server 2012、MS SQL Server 2014
- 服务器地址：输入连接数据节点的域名或IP地址。例如192.168.2.11或datapipeline.com
- 端口：输入连接数据节点的端口值
- 数据库名称：要求连接数据库的名称
- Schema有两个选项

» 选项一：dbo（默认）

» 选项二：用户可输入自定义选项

» 允许输入多个Schema，多个Schema可用英文逗号分隔

- 用户名和密码：输入可访问该数据节点的用户名和密码

> **注意事项**
> 使用Change Tracking初始化数据时，全量数据会打一个快照，增量数据会根据Change Tracking同步。

2.1.1.1.4. 配置PostgreSQL数据节点基本信息 Basic Configuration of PostgreSQL Data Node

在进行配置之前，请务必检查您的PostgreSQL数据节点是否符合平台要求，参考如下：

版本要求
1. 数据源定时模式支持PostgreSQL 9.x、10.x版本； 2. 数据源实时模式–wal2json支持PostgreSQL 9.4+、10.x版本，不支持PostgreSQL只读库； 3. 目的地暂无
用户权限要求
select on table , view change tracking on table , view change tracking on schema , select on schema , create object , insert , update , delete , drop , truncate
其他要求

表格9 PostgreSQL数据节点配置要求

新建弹窗界面选择了PostgreSQL节点类型之后，显示PostgreSQL的基本信息配置模板：

- 版本：PostgreSQL 9.4、PostgreSQL 9.5、PostgreSQL 9.6、PostgreSQL 10.4
- 服务器地址：输入连接数据节点的域名或IP地址。例如192.168.2.11或datapipeline.com
- 端口：输入连接数据节点的端口值
- 数据库名称：要求连接数据库的名称
- Schema目录：要求输入数据节点读取目录地址

  » 选项一：public（默认）

  » 选项二：用户可输入自定义选项

  » 允许输入多个Schema，多个Schema可用英文逗号分隔
- 用户名和密码：输入可访问该数据节点的用户名和密码

> **注意事项**
> PostgreSQL类型的数据节点仅用作数据源节点。

2.1.1.1.5. 配置Kafka数据节点基本信息 Basic Configuration of Kafka Data Node

在进行配置之前，请务必检查您的PostgreSQL数据节点是否符合平台要求，参考如下：

版本要求
1. 数据源支持0.9.x、0.10.x、0.11.x、1.x.y、2.0.x、2.1.x、2.2.x;
2. 目的地支持0.8.x、0.9.x、0.10.x、0.11.x、1.x,y、2.0.x、 2.1.x、2.2.x；
用户权限要求
具备需要进行读取topic的消费权限； 具备需要进行写入topic的生产权限。
其他要求

表格10 Kafka数据节点配置要求

新建弹窗界面选择了Kafka节点类型之后，显示Kafka的基本信息配置模板：

- 版本：
  - » 支持Kafka 0.8.x、Kafka 0.9.x、Kafka 0.10.x、Kafka 0.11.x、Kafka 1.x.y、Kafka 2.0.x、Kafka 2.1.x、2.2.x
  - » Kafka 0.8.x 只支持数据目的地节点
- 服务器地址：输入连接数据源的域名或IP地址及端口值，允许输入多个地址，使用","分隔

  注意：填写前需要检查Kafka集群是否有别名，若有别名则需要联系DataPipeline运维人员，添加Kafka别名的映射关系，并填写别名
- ZooKeeper地址：输入连接Zookeeper的域名或IP地址及端口值，允许输入多个地址，使用","分隔

2.1.1.1.6. 配置Greenplum数据节点基本信息 Basic Configuration of Greenplum Data Node

在进行配置之前，请务必检查您的Greenplum数据节点是否符合平台要求，参考如下：

版本要求
1. 1. 目的地支持Greenplum 4.3+、5.x.y、6.x.y；
2. 2. 数据源暂无；
用户权限要求
select, create object, insert, update, delete, drop, truncate
其他要求

表格11 Greenplum数据节点配置要求

新建弹窗界面选择了Greenplum节点类型之后，显示Greenplum的基本信息配置模板：

- 版本：Greenplum 4.x.y、Greenplum 5.x.y、Greenplum 6.x.y
- 服务器地址：输入连接数据节点的域名或IP地址。例如192.168.2.11或datapipeline.com
- 端口：输入连接数据节点的端口值

- Schema名称：要求选择默认配置或自定义写入地址
- 用户名和密码：输入可访问该节点的用户名和密码

> **注意事项**
>
> Greenplum类型的数据节点仅用作数据目的地节点。

2.1.1.1.7. 配置TiDB数据节点基本信息 Basic Configuration of TiDB Data Node

在进行配置之前，请务必检查您的TiDB数据节点是否符合平台要求，参考如下：

版本要求
1. 目的地支持TiDB 2.x.y
用户权限要求
select,create object,insert,update,delete,drop,truncate,select on mysql.*
其他要求

表格12 TiDB数据节点配置要求

新建弹窗界面选择了TiDB节点类型之后，显示TiDB的基本信息配置模板：

- 版本：TiDB 2.x.y
- 数据目的地名称：是该数据目的地在DataPipeline的唯一标识。
- 服务器地址：输入连接数据目的地的域名或IP地址。IP地址如192.168.2.11；域名如 data-pipeline.cye55uthbqll.cn-north-1.redshift.amazonaws.com.cn。
- 端口：输入连接数据目的地的端口值。
- 数据库名称：要求连接数据库的名称。
- 用户名和密码：输入可访问该数据目的地的用户名和密码。

> **注意事项**
>
> TiDB类型的数据节点仅用作数据目的地节点。

2.1.1.1.8. 配置SequoiaDB数据节点基本信息 Basic Configuration of SequoiaDB Data Node

在进行配置之前，请务必检查您的SequoiaDB数据节点是否符合平台要求，参考如下：

版本要求
1. SequoiaDB 2.8.7； 2. 数据源暂无
用户权限要求
SequoiaDB对于用户的DML，DDL，DQL等操作并没有明确的权限划分，因此每个用户都具有相同权限，数据管理员（安装时创建，默认为sdbadmin）以及用户列表 SDB_LIST_USERS 中的所有用户都可以在DataPipeline中新建数据节点。

其他要求

表格13 SequoiaDB数据节点配置要求

新建弹窗界面选择了SequoiaDB节点类型之后，显示SequoiaDB的基本信息配置模板：

- 版本：SequoiaDB 2.8.7
- 服务器地址：输入连接数据目的地的域名或IP地址。IP地址如192.168.2.11 或datapipeline.com。
- 端口：输入连接数据目的地的端口值。
- 数据库名称：要求连接数据库的名称。
- 用户名和密码：输入可访问该数据目的地的用户名和密码。

> **注意事项**
>
> SequoiaDB类型的数据节点仅用作数据目的地节点。

2.1.1.1.9. 配置Amazon Redshift数据节点基本信息 Basic Configuration of Amazon Redshift Data Node

在进行配置之前，请务必检查您的 Amazon Redshift 数据节点是否符合平台要求，参考如下：

版本要求
1. AWS REDSHIFT； 2. 数据源暂无；
用户权限要求
其他要求

表格14 Amazon Redshift数据节点配置要求

新建弹窗界面选择了Amazon Redshift节点类型之后，显示Amazon Redshift的基本信息配置模板：

- 服务器地址：输入连接数据目的地的域名或IP地址。IP地址如192.168.2.11或datapipeline.com。
- 端口：输入连接数据目的地的端口值。
- 数据库名称：要求连接数据库的名称。
- 用户名和密码：输入可访问该数据目的地的用户名和密码。
- Schema名称：要求输入数据节点写入目录地址。
- S3：
  - » Bucket：输入用户使用的Bucket。
  - » 地区：选择用户的S3服务器位置。
  - » 访问权限：您可以选择获取访问S3权限的方式，其中有：
    - · Role：使用AWS角色委托授权以获取S3权限。
    - · Access Key：需要输入Access Key ID 和 Access Key。

- 写入目录：要求输入写入目录
  - » 写入目录允许为空

> **注意事项**
>
> 1. Amazon Redshift类型的数据节点仅用作数据目的地节点。
> 2. 连接RedShift需要配置S3提高同步性能。

2.1.1.2. 数据节点–策略配置 Policy Configuration of Data Node

数据节点策略配置是在节点可以成功连接的基础上，对于节点在数据链路配置、数据融合任务运行过程中出现的问题进行应对的策略配置。

2.1.1.2.1. 检查点策略配置 Setting up Check Point Policy

MySQL、Oracle、SQL Server、PostgreSQL四种类型数据节点的节点详情界面的策略配置信息模块可以配置数据节点的检查点策略配置。

关于此功能

- MySQL：当数据源没有Binlog权限时，系统无法对数据进行检查点配置；当数据源有Binlog权限时，系统会根据用户设置对数据进行检查。
- Oracle：当数据源没有LogMiner权限时，系统无法对数据进行检查点配置；当数据源有LogMiner权限时，系统会根据用户设置对数据进行检查。
- SQL Server：当数据源没有Change Tracking权限时，系统无法对数据进行检查点配置；当数据源有Change Tracking权限时，系统会根据用户设置对数据进行检查。
- PostgreSQL：当数据源没有wal2json权限时，系统无法对数据进行检查点配置；当数据源有wal2json权限时，系统会根据用户设置对数据进行检查。
- 检查点频率：
  - » 系统会根据用户设置的检查频率对数据进行检查，用户可以设置分钟、小时频率，默认为10分钟，支持cron表达式。
  - » 检查频率决定了回滚的最小粒度。假设检查范围为一天，意味着只能从一天前的检查点时间回滚；如果设置检查频率为一小时，系统就会每小时记录一个检查点位置。
- 检查范围：系统会根据用户设置的检查范围保留数据日志情况，用户可以设置小时、天，默认为3天。

操作步骤

1. 点击数据节点列→数据节点详情，进入数据节点详情页。
2. 点击策略配置按钮，切换至策略配置页面。
3. 配置检查频率和检查范围。
4. 选择开启该策略。

2.1.1.2.2. 语义映射策略配置 Setting up Semantic Mapping Policy

语义映射策略是通过界面配置的方式，将数据源节点的数据类型、索引、特性等语义与数据目的地的数据类型、索引、特性等语义关联起来的映射配置，节点的语义映射是数据链路与数据任务配置数据映射的基石。

关于此功能

语义映射功能中包括了数据的类型映射，索引映射与特性映射。

我们支持多版本的语义映射管理，您可针对不同数据目的地类型及版本，对语义映射策略进行针对性调整。

操作步骤

- 新建语义映射
  1. 在数据节点列表页面点击节点名称或查看详情按钮进入数据节点详情页面。
  2. 点击策略配置切换至数据节点策略配置Tab。
  3. 在语义映射模块中点击添加。
  4. 输入语义映射策略名称，选择节点类型，选择节点版本，保存即新建语义映射。
  5. 点击映射详情按钮修改映射详情（详见下方修改映射）。
- 修改语义映射
  1. 在数据节点列表页面点击节点名称或查看详情按钮进入数据节点详情页面。
  2. 点击策略配置切换至数据节点策略配置Tab。
  3. 在语义映射模块中点击映射详情按钮进入语义映射详情弹窗。
  4. 弹窗内容分为三个部分，类型映射、索引映射、特性映射。
- 修改类型映射
  1. 添加类型映射
     a. 在类型映射数据源部分选择字段类型。
     b. 指定数据源该字段类型的精度标度条件。
     c. 在类型映射数据源部分选择字段类型。
     d. 点击数据目的地字段类型后的输入框指定数据目的地字段类型属性。
  5. 修改类型映射
     · 同添加步骤。
  6. 删除类型映射
     · 点击删除按钮即可删除类型映射。
- 修改索引映射
  1. 在索引映射数据源部分选择索引类型。
  2. 在索引映射数据目的地部分选择目的地映射索引类型。
  3. 添加删除逻辑同修改类型映射。
  4. 当前版本仅支持将数据源的唯一索引映射成为数据目的地的主键。
- 修改特性映射
  1. 在特性映射中，当前版本支持映射表名称与字段名称的修改。
  2. 支持选项
     a. 和源相同

b. 全部大写

c. 全部小写

3. 此项配置的应用场景

a. 数据链路的数据映射配置中，通过「创建表」方式生成的数据映射，会自动根据此项配置生成目的地表名称和字段名称。

b. 在数据源结构新增字段后，如用户配置结构变化策略为同步此字段，会自动根据此项配置生成目的地表名称和字段名称。

- 删除语义映射

1. 在数据节点列表页面点击节点名称或查看详情按钮进入数据节点详情页面。
2. 点击策略配置切换至数据节点策略配置Tab。
3. 在意义映射模块点击删除进入删除确认提示。

a. 删除之后不可恢复。

b. 如系统中有数据映射对该语义映射有依赖关系，语义映射不可删除。

4. 在删除确认提示点击删除即删除此条语义映射策略，取消即取消删除。

注意事项

1. 准确的语义映射策略是保证数据可以准确同步的基石，语义映射的错误将直接导致数据传输错误，请您谨慎配置。
2. 当有任务正在同步依赖于当前需要修改的语义映射的数据映射时，语义映射规则不可修改，如需修改请暂停相关数据任务。
3. 如您不对语义映射策进行编辑，我们提供默认的语义映射策略；如您应用不能覆盖所有数据源类型的语义映射策略，系统将从默认语义映射规则中自动选择最安全的语义映射进行映射。

下一步

2.2.1. 新增数据链路 Adding a Data Pipeline

2.2.1.1. 数据链路–基础配置 Basic Configuration of Data Pipelines

2.1.2. 修改数据节点 Editing a Data Node

关于此功能

编辑数据节点是指修改已存在的数据节点的配置内容，您可以对数据节点的名称、描述以及每种节点类型对应的连接参数进行修改，您也可以修改节点的策略配置，保证数据任务稳定运行。

操作步骤

- 编辑节点的名称和描述：

1. 点击数据节点列表→查看详情，进入数据节点详情界面。
2. 点击页面顶部节点名称旁边的「编辑」按钮，弹出编辑弹窗，可以修改节点的名称和描述。

   注：如果该数据节点有相关数据任务正在运行中，则无法编辑数据节点的名称和描述。
3. 编辑完节点名称和描述，点击弹窗右下角「保存」按钮即应用配置信息并下发到节点的相关链路和相关任务

- 编辑节点的连接配置信息：

1. 点击数据节点列→查看详情，进入数据节点详情界面。
2. 点击页面底部基本配置信息模块右上角「编辑」，弹出编辑弹窗，可以修改每种节点类型对应的连接参数。

   注：如果该数据节点有相关数据任务正在运行中，则无法编辑数据节点的连接配置信息。
3. 修改配置信息后，您可以点击弹窗左下角的「连接验证」检查当前页面修改的配置信息是否能连接成功。
4. 点击「保存」即应用配置信息并下发到节点的相关数据链路和相关数据任务。

> **注意事项**
>
> 1. 初次配置数据节点的时候您需要慎重选择数据节点的类型，选择完成保存后，系统将不允许修改数据节点的类型。
> 2. 如果该数据节点有相关任务正在运行中，则无法编辑数据节点的名称和描述，如果想要编辑该节点，您需要先暂停与节点相关的数据任务。

2.1.3. 激活数据节点 Activating a Data Node

数据节点有「活动」和「挂起」两个状态，用户可以自定义管理节点状态。激活数据节点指在系统管理节点状态的层级将数据节点状态由「挂起」变为「活动」，激活后的数据节点可被链路选择。

关于此功能

激活数据节点需要数据节点状态为挂起，当您激活数据节点，系统将会提示您节点的相关数据链路和相关数据任务也可被激活。

操作步骤

1. 点击数据节点列表或者数据节点详情页，点击激活数据节点开关。
2. 弹出激活确认弹窗，显示节点的相关数据链路和数据任务也可被激活提示以及连接和权限验证结果。
3. 点击确定即激活该数据链路。

> **注意事项**
> 激活数据节点是在系统管理层级上对节点的操作，激活数据节点会做节点的连接和权限验证以提示用户，验证失败也不会影响激活节点，但会影响数据任务实际运行。

下一步

如您需要激活相关数据链路和数据任务，可以在数据节点详情–相关数据链路和相关数据任务模块找到相关链路和相关任务并激活。

2.1.4. 挂起数据节点 Deactivating a Data Node

数据节点有「活动」和「挂起」两个状态，用户可以自定义管理节点状态。激活数据节点指在系统管理节点状态的层级将数据节点状态由「活动」变为「挂起」，挂起后的数据链路在管理上呈现不可用状态。

关于此功能

挂起数据节点会同时挂起节点相关的数据链路和数据任务，挂起节点后，节点的相关数据链路和相关数据任务也被挂起，并且不可被激活。

操作步骤

1. 点击数据节点列表或者数据节点详情页，点击挂起数据节点开关。
2. 弹出挂起确认弹窗，显示节点相关的数据链路和数据任务。
3. 挂起数据节点将会同步挂起节点相关数据链路和数据任务。
4. 点击「挂起」即挂起数据链路。
5. 系统会异步执行挂起相关数据链路和数据任务的操作。

> **注意事项**
> 挂起数据节点后，相关数据链路和相关数据任务将被异步挂起，故可能出现挂起任务失败的情况；未成功挂起的任务将会继续运行，直到您手动将其暂停。

2.1.5. 删除数据节点 Deleting a Data Node

关于此功能

删除数据节点指将已经配置好的数据节点删除，删除后将不可恢复。

操作步骤

1. 点击数据节点列表界面，每一个数据节点的右侧有「删除」按钮，您也可以勾选节点进行批量删除操作。
2. 当数据节点有相关数据任务和相关数据链路时，无法删除该节点，「删除」按钮置灰，勾选按钮置灰。此时，您需要在数据节点详情界面找到数据节点的相关任务以及相关链路，并且删除相关任务、相关链路之后，才能删除该数据节点。
3. 如果「删除」按钮是可点状态，点击「删除」按钮弹出二次确认提示，在二次确认窗口点击删除，及删除数据节点。

注意事项

1. 当数据节点有相关数据任务和相关数据链路时，无法删除该节点，如果想要删除该节点，需要先删除该节点的相关数据任务和相关数据链路。
2. 已被删除的数据节点不可恢复。

2.2. 管理数据链路 SETTING UP AND CONFIGURING DATA PIPELINES

数据链路是将数据任务配置集中管理，统一配置的功能模块。完成数据链路配置后，在数据任务配置中选择数据链路，相关配置将被直接应用至数据任务。

准确的链路配置是保证数据任务稳定运行的关键，例如数据源与数据目的地配置、数据映射、结构变化策略、主键冲突策略、增量处理策略、错误队列策略等，任务将完全应用链路的配置。与此同时，为了提升数据任务管理与运维的便捷性，在数据链路配置的基础上，数据任务可以自行定义自动重启策略、预警策略、日志策略等配置选项。

丰富的配置种类为任务运行提供了稳定性保障，但配置选项过多也会对您理解并使用DataPipeline进行数据同步带来一定影响。

因此，从配置的逻辑层面我们将数据链路的配置分成了三个层级，基础配置、限制配置、策略配置。

数据链路基础配置

数据链路基础配置是数据任务可以成功运行的最小化配置，即当数据任务选择已完成基础配置的数据链路，并完成数据任务基础配置，任务即可运行。其中包括：

◎ **数据源配置**

数据源配置是对不同数据源的读取配置。

◎ **数据目的地配置**

数据目的地配置是对不同数据源的写入配置。

◎ **数据映射配置**

数据映射配置是将数据源数据与数据目的地数据通过映射关联的配置，其中包括：

· 表映射配置

数据映射–表映射配置是建立数据源表与数据目的地表映射关系的配置。

· 字段映射配置

数据映射–字段映射配置是建立数据源表中字段与数据目的地表中字段映射关系的配置。

· 读取内容限制

数据映射–读取内容限制是数据源读取内容的限制配置。

· 清洗规则配置

数据映射–清洗规则配置是对即将写入数据目的地的数据进行清洗处理的配置。

数据链路策略配置

数据链路策略配置是解决数据任务运行时可能遇到的问题的配置选项。它可以被分成两类：一类是对于任务运行过程中可能会遇到的错误提供的系统自动的应对策略，其中包括：写入主键冲突策略、结构变化策略、主键冲突策略、增量处理策略、错误队列策略、自动重启策略等。另一类是记录和反馈任务运行信息，方便处理任务错误情况提供的策略，包括日志策略、预警策略和错误队列策略中的错误数据存储等。

◎ **写入主键冲突策略**

写入主键冲突策略是在任务写入过程中，写入数据与目的地数据有主键冲突时任务执行的应对策略，我们提供覆盖数据与忽略数据的选项。

◎ **结构变化策略**

结构变化策略是当数据源数据结构发生变化时，系统将为您执行的策略，能够有效避免由于数据源结构变化使任务暂停带来的影响。

◎ **增量处理策略**

当数据源产生已同步的增量数据被删除的情况时，您可以通过配置增量处理策略来对这部分数据进行处理，保证数据一致性。

◎ **端到端一致性策略**

在任务运行过程中，可以开启端到端一致性策略来保证数据从源端到目的地端的一致性。

◎ **自动重启策略**

自动重启策略指当任务报错时，系统将自动重启的策略，以对数据任务的不同错误类型做出是否重启的调整。

◎ **错误队列策略**

开启错误队列策略时，运行中的数据任务产生错误数据时，可以不暂停数据任务，将错误数据存储于指定节点，并记录错误堆栈信息。有效避免因任务出现错误数据而暂停所带来的影响。

◎ **预警策略**

通过设置预警规则，选择预警发送组，预警策略可以帮助您实现对关注内容的预警配置，当任务出现预警超出规则限制的情况时，可以及时通过预警发送方式（包括邮件与WebHook）通知到您。

◎ **日志策略**

合理的日志策略可以有效帮助您降低查询与管理日志数据的时间投入，通过日志策略，您可以配置日志记录的类别与日志存储方式。

数据链路的基本管理

新增数据链路

详见："2.2.1. 新增数据链路 Adding a Data Pipeline"

编辑数据链路

详见："2.2.2.修改数据链路 Editing a Data Pipeline"

激活数据链路

详见："2.2.3. 激活数据链路 Activating a Data Pipeline"

挂起数据链路

详见："2.2.4. 挂起数据链路 Deactivating a Data Pipeline"

删除数据链路

详见："2.2.5. 删除数据链路 Deleting a Data Pipeline"

2.2.1. 新增数据链路 Adding a Data Pipeline

新增数据链路指添加一条新的数据链路至系统中，您需要完成链路基本配置，数据任务选用该链路后才可以正常运行。

关于此功能

新增数据链路操作通常由数据部门工作人员进行，您可以对数据链路的数据源、数据目的地、数据映射进行配置，限定链路数据映射范围，您也可以为数据链路添加策略配置，保证数据任务稳定运行。

操作步骤

1. 点击数据链路列表页–>新增数据链路；
2. 输入链路名称及描述，点击「确定」进入链路基本配置页面；
3. 选择数据源（可选多个）并选择增量/全量读取方式；
4. 选择数据目的地（可选多个）并选择写入方式；
5. 配置数据映射，操作步骤详见：
   "2.2.1.1.1. 数据链路–基础配置–关系型数据节点数据映射 Data Mapping—RDBMS"；
6. 保存基本配置；
7. 完成基本配置后，任务应用链路配置即可运行，但为保证任务运行的稳定性，我们推荐您完成链路策略配置后再通过数据任务使用链路配置。

> **注意事项**
>
> 初次配置数据链路时需慎重选择数据源的读取方式，选择完成保存后，系统将不允许修改读取方式。

下一步

- 快速配置使任务快速运行：数据任务配置—选择数据链路
- 更加关注任务运行稳定：数据链路配置–策略配置

2.2.1.1. 数据链路–基础配置 Basic Configuration of Data Pipelines

数据链路基本础配置是保证数据任务成功运行的基本配置，其中包括数据源配置、数据目的地配置与数据映射配置。

关于此功能

完成数据链路基础配置的基本功能模块：

- **数据源配置**
  1. 选择数据源
     - 新建数据源
  2. 增量读取配置
  3. 全量读取配置
- **数据目的地配置**
  1. 选择数据目的地
     - 新建数据目的地
  2. 写入配置
- **数据映射配置**
  1. 配置表映射
     - 创建目的地表
  2. 配置字段映射
  3. 限制读取范围
  4. 配置数据清洗脚本

新建数据源、数据目的地：

详见"2.1.1. 新增数据节点 Adding a Data Node"。

配置数据映射：

详见"2.2.1.1.1. 数据链路–基础配置–关系型数据节点数据映射 Data Mapping—RDBMS"。

操作步骤

1. 点击数据链路列表–>数据链路详情，进入数据链路详情页；
2. 选择数据源节点；
   - » 可选数据源节点为已被激活的数据节点。
3. 选择增量读取方式或者全量读取方式；
   - » 增量读取方式指获取增量数据的方式，对于RDBMS类型的数据源来说，增量获取方式通常包括日志增量获取与增量识别字段增量获取。
4. 选择数据目的地节点；
5. 选择写入方式；
6. 配置数据映射。

数据映射相关配置较为复杂，我们使用另一个篇幅描述：

详见"2.2.1.1.1. 数据链路–基础配置–关系型数据节点数据映射 Data Mapping—RDBMS"。

> **注意事项**
>
> 读取方式完成选择并保存后，将不允许修改，请您慎谨慎操作。

下一步

- 数据映射配置

2.2.1.1.1. 数据链路–基础配置–关系型数据节点数据映射 Data Mapping—RDBMS

数据映射配置是将数据源数据与数据目的地数据关联起来的步骤，是DataPipeline产品支持多元异构场景的核心，通过图形化的配置，您可以轻松地建立不同数据节点之间的数据映射，来为数据任务运行做准备。

关于此功能

数据映射功能分为表映射和字段映射两个部分，通过限制读取内容与数据清洗脚本，可对数据映射进行更加特异化的调整，我们将使用四个页面的篇幅介绍数据映射：

- **表映射<数据映射配置--配置表映射>**
  1. 选择同步列表；
  2. 选择语义映射规则；
  3. 创建新表；
  4. 选择已有表；
  5. 创建目的地表与刷新目的地表。
- **字段映射<数据映射配置--配置字段映射>**
  1. 创建新表--调整字段映射内容；
  2. 选择已有表--选择字段映射；
  3. 刷新目的地表结构。
- **限制读取内容<数据映射配置--限制读取内容>**
  1. 配置增量识别字段。
- **配置清洗脚本<数据映射配置--配置清洗脚本>**
  1. 编辑与保存清洗脚本；
  2. 样例数据与试运行；
  3. 脚本库；
  4. 常见场景说明。

注意事项

读取方式完成选择并保存后，将不允许修改，请您谨慎操作。

下一步

- 数据映射配置--配置表映射<数据映射配置--配置表映射>

2.2.1.1.1.1. 配置表映射 Setting up Table Mapping

配置表映射是在界面中建立数据源表与数据目的地表的映射关系。建立表映射关系后，才能对其中字段的映射进行配置，以完成数据链路映射配置。

关于此功能

配置数据表映射需要使用数据节点语义映射的配置内容，您可参照<数据节点策略配置--语义映射策略配置>完成该项配置。

操作步骤

1. 点击表映射按钮，切换至表映射Tab；
2. 点击数据源按钮，切换至您需要配置的数据源：
   - » 此处支持的数据源是上方数据源配置中选择的数据源；
3. 点击选择同步列表按钮，选择数据源中您需要同步的数据表：
   - » 选择完成后数据目的地将会展现对应行，每行均有两个可选项：创建新表、选择已有表；
4. 点击创建新表，切换至字段映射关系Tab：
   - » 系统会使用已选的语义映射策略，自动将数据源表的字段映射至数据目的地表；
   - » 您可在字段映射中修改配置内容,详见<配置字段映射>；
5. 点击选择已有表，切换至字段映射Tab：
   - » 您可在字段映射中修改配置内容,详见<配置字段映射>；
6. 在需要新建表的场景中，修改完字段映射后，点击创建目的地表以通过系统在数据目的地建表：
   - » 建表需要创建表权限；
   - » 系统提供导出建表语句功能，您可在创建目的地表弹窗点击导出建表语句按钮以导出建表语句；
   - » 通常情况下，生产环境建表需要您公司DBA创建表。

注意事项

1. 完整准确的表映射是准确进行数据同步的基础，您需要谨慎操作。
2. 新增同步列表中内容不会影响已有数据映射，与使用已有数据映射的数据任务。
3. 修改表映射关系会影响使用该映射的相关数据任务，修改后相关数据任务将被暂停，需要您手动启动任务。

下一步

- 数据映射配置--配置字段映射<数据映射配置--配置字段映射>

2.2.1.1.1.2. 配置字段映射 Setting up Field Mapping

配置字段映射是在界面中建立数据源表中字段与数据目的地表中字段的映射关系，建立字段映射关系后，数据才能被准确的从数据源同步至数据目的地。

关于此功能

因在配置数据映射时可以在数据目的地选择已有表或新建表，配置字段映射也会根据表映射中所选内容而改变。

操作步骤

1. 点击数据映射–表映射–选择已有表，跳转至字段映射tab；
2. 在每一行的数据目的地字段列中选择数据目的地表的字段；
   - » 您可以为每一个数据源字段选择一个匹配的数据目的地字段。
3. 如数据目的地没有您需要的数据表，您可以点击「创建新表」，跳转至字段映射tab；
   - » 系统将根据已选的语义映射规则自动的为您创建目的地字段与字段属性。
   - » 如您对数据目的地表进行了更改，可以点击刷新表按钮刷新目的地表结构。
   - » 您可以修改数据目的地名称与字段属性。
4. 在需要新建表的场景中，修改完字段映射后，点击创建目的地表以通过系统在数据目的地建表；
   - » 建表需要创建表权限。
   - » 系统提供导出建表语句功能，您可在创建目的地表弹窗点击导出建表语句按钮以导出建表语句。
   - » 通常情况下，生产环境建表需要您公司DBA创建表。
5. 在配置数据目的地字段映射过程中，可以为清洗脚本输出的数据预留字段。
   - » 预留字段的映射没有相应的数据源字段信息。
   - » 预留字段字段名称需与清洗脚本输出数据的scheme中的列名称保持一致。

注意事项

1. 完整准确的字段映射是准确进行数据同步的基础，您需要谨慎操作；
2. 新增同步列表中内容不会影响已有数据映射，与使用已有数据映射的数据任务；
3. 修改表映射关系会影响使用该映射的相关数据任务，修改后相关数据任务将被暂停，需要您手动启动任务。

下一步

- 数据映射配置––配置字段映射<数据映射配置––限制读取内容>

2.2.1.1.1.3. 配置清洗脚本 Setting up ETL Script

清洗脚本是系统提供的在数据同步过程中，用于复杂数据处理场景的工具，清洗脚本需要在字段映射配置中添加高级清洗字段功能一并使用。

关于此功能

我们支持使用java语言编写清洗脚本，通过自定义清洗脚本，您可以实现时间数据格式处理，添加DML标识和简单合并计算数据等功能。

操作步骤

编写与保存清洗脚本

1. 点击清洗脚本图标按钮，进入清洗脚本编辑弹窗；
2. 使用Java语言编写清洗脚本；
3. 点击存入脚本库，保存已编辑内容。

样例数据与试运行

1. 在清洗脚本编辑弹窗，点击获取样例数据，可以根据筛选条件获取当前表的样例数据；
2. 样例数据为Json格式；
3. 点击试运行即使用上方编写好的清洗脚本，对数据进行处理:
   - » 清洗后的数据将打印在试运行窗口内；
   - » 任务运行过程中，清洗脚本输出的数据会被写入至数据目的地；
   - » 我们会根据清洗脚本输出数据中的字段名称与目的地表的字段名称做对比，完全一致的会进行写入；
4. 您可以通过对比样例数据与试运行的输出数据进行对比，判断清洗脚本是否存在处理逻辑错误。

使用脚本库

1. 点击脚本库按钮进入脚本库选择弹窗；
2. 输入脚本库名称可以搜索脚本库；
3. 选择脚本库列表中的脚本或默认模板中的脚本；
4. 选择完成后脚本将被替换至清洗脚本编辑框；
5. 系统默认模板及使用方法介绍详见<数据映射配置--配置清洗脚本>。

注意事项

1. 清洗脚本处理程序会调用sink端服务器来进行数据计算，根据计算逻辑的复杂程度会占用部分服务器资源。
2. 系统中任务运行资源占用模式为争抢模式，清洗脚本的运行同样包含在任务运行的范畴内，资源配置及程序资源使用说明详见<数据任务配置-资源组配置>。

下一步

- 数据任务配置<选择数据链路>
- 数据链路配置<策略配置>

2.2.1.1.1.3.1 清洗脚本配置样例 Best Practice of Using ETL Script

为了方便用户使用清洗脚本，我们提供了丰富的清洗脚本使用样例。

一、DML字段标识脚本

脚本描述：源表数据发生变化时，在目的地表中增加相应的 DML 标识字段，包括 insert、update、delete。

```
import com.datapipeline.clients.connector.
record.DpRecordMeta;
import java.util.Map;

public  class AddDMLField {
    public  static Map process(Map record,
DpRecordMeta meta) {
    // 系统会向 dml 字段写入标
识：insert、update、delete。
    // 如果是定时读取模式，所有 dml 字段会表示为
insert，只有通过实时模式读取数据时才会识别 update
和 delete。
    record.put("dml", meta.getType());
    // 保存该脚本后，请在目的地表结构添加新字
段：dml，并把字段类型改为字符串类型（例如：varchar）
    return record;
    }
}
```

二、读取时间脚本

脚本描述：根据数据读取时间，在目的地表中增加相应的时间字段。

```
import com.datapipeline.clients.connector.
record.DpRecordMeta;
import java.util.Map;
import java.time.LocalDateTime;
import java.time.ZoneId;
import java.time.Instant;
import java.time.format.DateTimeFormatter;

public  class AddCollectTimeField {
    private  static DateTimeFormatter
DATE_TIME_FORMATTER = DateTimeFormatter.
ofPattern("yyyy-MM-dd HH:mm:ss");

    public Map process(Map record,
DpRecordMeta meta) {
        // 系统会向 collect_time 字段写入读取该数
据的时间，格式为：yyyy-MM-dd HH:mm:ss。
        record.put("collect_
time", Instant.ofEpochMilli(meta.
getSourceRecordTimestamp()).atZone(ZoneId.
of("UTC")).format(DATE_TIME_FORMATTER));
        // 保存脚本后，在目的地表结构中添加字
段：collect_time，并把字段类型改为时间类型（或字符串
类型）。
        return record;
    }
}
```

三、写入时间脚本

脚本描述：根据数据写入时间，在目的地表中增加相应的时间字段。

```
import java.util.Map;
import java.time.LocalDateTime;
import java.time.ZoneId;
import java.time.format.DateTimeFormatter;

public  class AddUpdateTimeField {
    private  static DateTimeFormatter
DATE_TIME_FORMATTER = DateTimeFormatter.
ofPattern("yyyy-MM-dd HH:mm:ss");

    public Map process(Map record) {
        // 系统会向 update_time 字段写入写入目的
地的时间，格式为：yyyy-MM-dd HH:mm:ss。
        record.put("update_time",
LocalDateTime.now(ZoneId.of("Asia/
Shanghai")).format(DATE_TIME_FORMATTER));
        // 保存脚本后，在目的地表结构中添加字
段：update_time，并把字段类型改为时间类型（或字符串
类型）。
        return record;
    }
}
```

四、日期类型转换脚本

脚本描述：将 Timestamp 类型数据转换为日期类型。

```
import java.time.LocalDateTime;
import java.time.format.DateTimeFormatter;
import java.util.Map;

public  class TimestampProcess {
    private  static  final DateTimeFormatter
DATE_FORMATTER = DateTimeFormatter.
ofPattern("yyyy-MM-dd");
    private  static  final DateTimeFormatter
TIMESTAMP_FORMATTER =
        DateTimeFormatter.ofPattern("yyyy-MM-
dd HH:mm:ss.SSS");

    public Map process(Map record) {
    // timestamp_type 类型为字符串，格式为：2019-
07-24 17:06:54.000
    final String timestampStr = (String)
record.get("timestamp_type");
    if (timestampStr != null) {
        try {
            // 将字符串转换为日期对象
            final LocalDateTime localDateTime
= LocalDateTime.parse(timestampStr,
TIMESTAMP_FORMATTER);
            // 将时间对象转换为 DATE_FORMATTER
格式的字符串，转化之后的字符串为：2019-07-24
            record.put("formatedTime", DATE_
FORMATTER.format(localDateTime));
        } catch (Exception e) {
        e.printStackTrace();
        }
    }
    return record;
    }
}
```

五、Timestamp 转换脚本

脚本描述：将日期类型转换为 Timestamp 类型数据。

```
import java.time.ZoneId;
import java.time.ZonedDateTime;
import java.time.format.DateTimeFormatter;

public class Date2TimeStampSample {

  public Map<String, Object>
process(Map<String, Object> record)  throws
Exception{
    // date = "20180101 12:0:22.123"
    String date = record.get("date");
    long timeStamp = ZonedDateTime.
parse(date, DateTimeFormatter.
ofPattern("yyyy-MM-dd HH:mm:ss.SSS").
withZone(ZoneId.of("Asia/Shanghai"))).
toInstant().toEpochMilli();
    record.put("ts", timeStamp);
    return record;
  }
}
```

六、JSON解析脚本

脚本描述：将 JSON 数据解析，并指定到目标表表结构。

```
import com.alibaba.fastjson.JSON;
import com.alibaba.fastjson.JSONArray;
import com.alibaba.fastjson.JSONObject;
import java.util.ArrayList;
import java.util.HashMap;
import java.util.List;
import java.util.Map;
import java.util.Objects;

public class Test_api_url_nameTransformEngine
{
  public List<Map<String, Object>>
process(Map<String, Object> record) {
    List<Map<String, Object>> records = new
ArrayList<>();
    if (!Objects.isNull(record)) {
      String jsonStr = (String) record.
get("content");
      JSONObject jsonNode = JSON.
parseObject(jsonStr);
      JSONArray items = jsonNode.
getJSONObject("reponseData").
getJSONArray("data");
      for (Object iterm : items) {
        JSONObject eachIterm = (JSONObject)
iterm;
        String title = String.
valueOf(eachIterm.get("title"));
        if (title != null && title.length() >
100) {
          title = title.substring(0, 100);
        }
        Map<String, Object> eachRowData = new
HashMap<>();
        eachRowData.put("_id", String.
valueOf(eachIterm.get("id")));
        eachRowData.put("tab", String.
valueOf(eachIterm.get("tab")));
        eachRowData.put("title", title);
        records.add(eachRowData);
      }
    }
    return records;
  }
}
```

七、数据脱敏

场景描述：用户想对表格中的信息进行脱敏，比如手机号或者身份证号。

例：身份证号脱敏，保留前四位和后四位，手机号码脱敏，保留前三位和后四位。

具体代码如下：

```
import java.util.Map;

public class
Sid_2382_ 05b0f75d2382f96a23824d7023829c
612382ef7ac208785d {

 public Map process(Map record) {

    record.put("id",((String)record.
get("id")).replaceAll("(\\d{4})\\d{10}
(\\d{4})","$1**********$2"));
record.put("phone",((String)record.
get("phone")).replaceAll("(\\d{3})\\d{4}
(\\d{4})","$1****$2"));

    return record;
  }
}
```

八、枚举类型转换

场景描述：用户想对表格中的类型进行转换。

例：把原表中表示性别的‘0’和‘1‘，到目的地表中，转换为对应的‘男’和‘女’。

具体代码如下：

```
import java.util.Map;

public class Sid_2383_
a8984a582383944223834f332383b7e
52383aa115c629f93 {

public Map process(Map record) {
    if (record.get("sex").equals("0"))
    {record.put("sex","男");}
    else if(record.get("sex").equals("1"))
    {record.put("sex","女");}
    else
    {record.put("sex","");}
    return record;
  }
}
```

九、字段求和

场景描述：用户想对表格中的几个字段的内的数据进行求和。

例：把原表中的两个字段内的薪水，到目的地表中，生成一个新的字段，其值为前两个字段的和。

具体代码如下：

```
import java.util.Map;
public class Sid_2384_
f049d91a2384daea23844a3a23
848fd923847cc3bc23dc6e {
public Map<String, Object>process(Map<String,
Object>record) {
    record.put("sum",(Double) record.
get("salary1")+(Double) record.
get("salary2"));
    return record;
  }
}
```

十、时间类型格式统一

场景描述：用户想对表格中字段的日期格式统一。

（2019115、20190115、2019.01.15）转换成 YYYY-MM-DD（2019-01-15）。

具体代码如下：

```
import java.text.SimpleDateFormat;
import java.util.Date;
import java.util.HashMap;
import java.util.Map;

public class ConventDateSample {

  public Map<String, Object>
process(Map<String, Object> record)  throws
Exception{
    try{
        SimpleDateFormat df = new
SimpleDateFormat("yyyyMMdd");
        SimpleDateFormat df2 = new
SimpleDateFormat("yyyy-MM-dd");
        Date datetime=df1.parse(record.
get("datetime").toString());
        String newDatetime = df2.
format(datetime);
        record.put("datetime",newDatetime);
    }catch(Exception e){
    }
    return record;
  }
}
```

十一、库存日期结合

场景描述：用户想对表格中日期和时间，结合到一个字段内。

具体代码如下：

```
import java.util.Map;
import java.util.Date;
import java.text.ParseException;
import java.text.SimpleDateFormat;

public class Sid_2516_8ffcc7ca2516201f25164ee
7251690f3251641b41d3bbd5c {

  public Map<String,
Object>process(Map<String, Object>record) {
    record.put("datetime2", record.
get("date")+" "+ record.get("time"));
    return record;
  }
}
```

十二、日期间隔计算

场景描述：用户想对表格中两个日期之间相差的天数，进行计算，再写入到一个字段内。

具体代码如下：

```
import java.text.SimpleDateFormat;
import java.util.Date;
import java.util.HashMap;
import java.util.Map;

public class Sid_2519_5017952225194a2
425194ff5251980f3251985af080e3b28 {

  public Map<String,
Object>process(Map<String, Object> record)
throws Exception{

    SimpleDateFormat df = new
SimpleDateFormat("yyyy-MM-dd");
    SimpleDateFormat df2 = new
SimpleDateFormat("yyyy-MM-dd HH:mm:ss");
    Date start=df2.parse(record.
get("startTime").toString());

    Date end=df2.parse(record.get("endTine").
toString());
    try{
    int days= (int) ((df.parse(df.
format(end)).getTime() -df.parse(df.
format(start)).getTime()) / (24 * 60 * 60 *
1000));
    record.put("days",days);
}catch(Exception e){
}
    return record;
  }
}
```

十三、非空字段判断

场景描述：判断一个字段是否为空。

```
import java.util.Map;
import java.util.Objects;
import com.datapipeline.clients.utils.
CodeEngineUtils;

public  class Example13 {
  public Map process(Map record)
    final Object orderno = record.
get("orderno");
    boolean ordernoIsNull = CodeEngineUtils.
isNull(orderno);
}
```

十四、指定数据进错误队列

场景描述：指定某一些数据进入错误队列。

```
import java.util.Map;
import java.util.Objects;
import java.lang.String;
import java.lang.Throwable;
import com.datapipeline.clients.codeengine.
CustomizedCodeEngineException;

public class Example14 {
 public Map process(Map record) throws
CustomizedCodeEngineException {
 if(Integer.parseInt(record.get("age").
toString()) <= 20){
 throw new CustomizedCodeEngineException(new
Exception("Illegal data!"));
 }
 return record;
 }
}
```

十五、时区转换

场景描述：将上海时间转化为标准时间。

```
import java.util.Map;
import java.time.*;
import java.time.format.DateTimeFormatter;


public class localTime2standardTime {
   public Map<String, Object>
process(Map<String, Object> record) {
      String crt = (String)record.
get("create_time");
      try {
        long ncrt = ZonedDateTime.
parse(crt,DateTimeFormatter.ofPattern("yyyy-
MM-dd HH:mm:ss.SSS").withZone(ZoneId.
of("Asia/Shanghai"))).toInstant().
toEpochMilli();
        record.put("timestamp",ncrt);
      }
      catch (Exception e) {
           e.printStackTrace();
      }
   return record;
   }
}
```

十六、字符集编码转换

场景描述：当数据源的数据为乱码时，可通过高级清洗转换字符集编码，处理乱码数据。

```
package com.dp.exltcsv;

import java.io.UnsupportedEncodingException;
import java.util.Map;

public class utf2gbk {
    public Map<String, Object>
process(Map<String, Object> record) throws
UnsupportedEncodingException {
        // 针对单个字段的格式
//        String key = (String)record.
get("key");
//        record.put(key,(Object)new
String(key.getBytes("utf-8"),"gbk"));
//        return record;
        // 针对所有字段
        for (String key : record.keySet()) {
            String value = (String)record.
get(key);
            record.put(key,(Object)new
String( value.getBytes("utf-8"), "gbk"));
        }

        return record;
    }
}
```

2.2.1.1.1.4. 限制读取内容 Setting up Data Reading Limitation Factor

限制读取内容是对数据源中数据读取的限制配置，通常用于增量识别字段模式指定增量数据的获取。

关于此功能

在增量识别字段模式获取增量数据时，我们提供lastmax（）函数配置的方式来指定增量数据获取。

操作步骤

1. 在数据映射--表映射页面点击限制读取内容按钮；
2. 或在数据映射--字段映射页面点击编辑读取条件，进入限制读取内容配置弹窗：
   - 您可根据需求设置 WHERE 语句，DataPipeline 提供 last_max()函数帮助用户解决定时读取增量数据的需求。

・last_max()函数：

使用该函数 DataPipeline 会取该任务下已同步数据中某一个字段的最大值，您可以使用该值作为 WHERE 语句读取条件。

使用 last_max()函数，在第一次执行该语句或对应字段暂无数值，则会忽略该函数相关的读取条件。

» 例子一：

已同步数据中，取某一个字段的最大值用于读取条件。

SELECT * FROM table1 WHERE _id>=last_max(_id)

注：每次执行批量读取时，使用 last_max 自定义函数取已同步数据中「_id」字段最大值，读取大于等于「id」已同步最大值的数据。

» 例子二：

以日期字段作为读取条件，每次只同步当前日期减一天的数据，并且只同步未曾读取的数据。

SELECT * FROM E_Commerce_Retail_Sales WHERE date >last_max(date) and date <=CURDATE()-1

每次只同步 date 字段值大于已同步数据中 date 字段最大值，并且只同步昨天的数据（date 字段值）。

注意事项

选择合适的增量识别字段有助于增量任务稳定运行，通常情况下选取日期、时间戳、数字自增id等字段作为增量识别字段配置相对容易，且计算更加稳定。

2.2.1.1.2. 数据链路-基础配置-NoSQL类型节点数据映射 Data Mapping—NoSQL

数据映射配置是将数据源数据与数据目的地数据关联起来的步骤，是DataPipeline产品支持多元异构场景的核心。通过图形化的配置，您可以轻松的建立不同数据源之间的数据映射，来为数据任务运行做准备。

关于此功能

数据映射功能分为表映射和字段映射两个部分，通过限制读取内容与数据清洗脚本，可对数据映射进行更加特异化的调整，我们将使用三个页面的篇幅介绍数据映射。

- 表映射<数据映射配置--配置表映射>
  - » 选择同步列表
  - » 选择语义映射规则
  - » 创建新的表层级结构或选择已有的表层级结构
  - » 执行创建目的地表与刷新目的地表
- 字段映射<数据映射配置--配置字段映射>
  - » 调整字段映射内容定义目的地表写入数据结构
- 配置清洗脚本<数据映射配置--配置清洗脚本>
  - » 编辑与保存清洗脚本
  - » 样例数据与试运行
  - » 校本库
  - » NoSQL类型数据读取解析与写入赋值
  - » 常见场景说明

下一步

- 数据映射配置--配置表映射<数据映射配置--配置表映射>

2.2.1.1.2.1. 配置表映射 Setting up Table Mapping

配置表映射是在界面中建立数据源表层级结构与数据目的地表层级结构的映射关系，建立表层级结构映射关系后，才能对其中具体数据内容、数据字段映射进行配置，以完成映射配置。

关于此功能

NoSQL类型数据节点作为数据源或数据目的地，通常会带有具备独立特性的数据层级结构，例如：

- Kafka：Topic
- Redis：Key-Value
- SequoiaDB：Collection Space—Collection

我们将针对上述节点的独立特性，将数据层级结构一一对应。

操作步骤

1. 点击表映射按钮，切换至表映射Tab；
2. 点击数据源按钮，切换至您需要配置的NoSQL类型数据源

此处支持的数据源是上方数据源配置中选择的数据源

3. 点击选择同步列表按钮，选择数据源中您需要同步的数据表层级结构
   - » 选择完成后数据目的地将会展现对应行，每行均有两个可选项
     - · 对于RDBMS类型目的地，有创建新表/选择已有表两个选项
     - · 对于NoSQL类型数据目的地
       a. Kafka支持选择已有Topic
       b. Redis支持创建新Key-Value
       c. SequoiaDB支持创建新Collection或选择已有Collection
4. 选择目的地表映射方式
5. 点击字段映射按钮，进入字段映射页面进行字段映射配置
6. 在需要新建表的场景中，修改完字段映射后，点击创建目的地表以通过系统在数据目的地建表：
   - » 建表需要创建表权限；
   - » 系统提供导出建表语句功能，您可在创建目的地表弹窗点击导出建表语句按钮以导出建表语句；
   - » 通常情况下，生产环境建表需要您公司DBA创建表。

注意事项

1. 完整准确的表映射是准确进行数据同步的基础，您需要谨慎操作。
2. 新增同步列表中内容不会影响已有数据映射及使用已有数据映射的数据任务。
3. 修改表映射关系会影响使用该映射的相关数据任务，修改后相关数据任务将被暂停，需要您手动启动任务。

下一步

- 数据映射配置--配置字段映射<数据映射配置--配置字段映射>

2.2.1.1.2.2. 配置字段映射 Setting up field mapping—NoSQL

配置字段映射是在界面中建立数据源表层级结构中字段与数据目的地表层级结构中的字段或数据的映射关系，建立字段或数据的映射关系后，数据才能被准确的从数据源同步至数据目的地。

关于此功能

NoSQL类型数据节点作为数据源或数据目的地时，数据往往是半结构化的，存储数据格式通常包含JSON、BSON、Avro、XML等。

DataPipeline数据采集组件读取半结构化数据后，需要对数据进行解析，提取数据元素；数据加载组件写入半结构化数据时，需要指定写入数据结构。

操作步骤

Kafka数据源字段映射配置

1. 点击数据映射-表映射-字段映射按钮，进入字段映射Tab；
2. 针对Kafka数据源，选择该Topic中Key和Value的数据反序列化器：
   - » 如选择Key和Value的数据反序列化器均为String格式，则源端展示字段为Key、Value；
   - » 如Key或Value的数据反序列化器为Avro格式，数据将被解析并展示于字段映射。
     注：Avro格式仅进行第一层级的数据解析，如您需要多层复杂结构的数据解析，请于清洗脚本中配置。

Kafka数据目的地字段映射配置

1. 点击数据映射-表映射-字段映射按钮，进入字段映射Tab；
2. 针对Kafka目的地，我们提供数据序列化器为String格式；
3. 点击结构定义按钮进行结构定义；
   - » 映射中目的地Topic的Key与Value可以分别进行结构定义；
   - » 结构定义中使用数据源字段变量，以约定形式添加至结构定义输入框。
4. 通过高级清洗向结构定义中除数据源字段变量之外的其他变量赋值；
5. 完成字段映射配置。

Redis目的地字段映射配置

1. 点击数据映射-表映射-字段映射按钮，进入字段映射Tab；
2. 针对Redis目的地，您可以选择目的地映射方式为Hash或

String；

3. 选择String模式，进入字段映射页面，将自动根据数据源生成目的地字段：
   - » 点击结构定义按钮，使用目的地字段变量与数据表信息变量对Key与Value进行结构定义。
4. 选择Hash模式，进入结构定义选择Hash类型：
   - » 选择Hash_table类型
     a. HashKey使用数据表信息变量拼接表名称；
     b. FieldKey使用主键集合变量进行拼接；
     c. FieldValue使用数据进行结构拼接定义。
   - » 选择Hash_行转列模式
     a. HashKey使用数据表信息变量与主键集合变量拼接表名称；
     b. FieldKey默认为字段名称变量；
     c. FieldValue默认为字段值变量。
4. 通过高级清洗向结构定义中除数据源字段变量之外的其他变量赋值；
5. 完成字段映射配置。

SequoiaDB目的地字段映射配置

1. 点击数据映射–表映射–字段映射按钮，进入字段映射Tab；
2. 针对SequoiaDB目的地，我们提供根据数据源表结构；自动生成单层BSON结构的功能；
3. 通过高级清洗向单层BSON结构变量赋值；
4. 完成字段映射配置。

注意事项

1. 完整准确的字段映射是准确进行数据同步的基础，您需要谨慎操作。
2. 新增同步列表中内容不会影响已有数据映射，与使用已有数据映射的数据任务。
3. 修改表映射关系会影响使用该映射的相关数据任务，修改后相关数据任务将被暂停，需要您手动启动任务。
4. NoSQL类型数据节点的数据映射高级清洗功能与RDBMS类型数据节点的数据映射高级清洗功能一致，配置可参考：<数据映射配置--RDBMS--配置清洗脚本>。

下一步

- 数据映射配置--配置字段映射<数据映射配置--清洗脚本配置>

2.2.1.2. 数据链路–策略配置 Policy Configuration of Data Pipeline

数据链路策略配置是在任务可以成功运行的基础上，对于任务运行过程中出现的问题进行应对的策略。

关于此功能

策略配置共包括8个细分策略，分别是

- 写入主键冲突策略
- 结构变化策略
- 增量处理策略
- 端到端一致性策略
- 自动重启策略
- 错误队列策略
- 预警策略
- 日志策略

操作步骤

1. 点击数据链路列表–>数据链路详情，进入数据链路详情页；
2. 点击策略配置按钮，切换至策略配置页面；
3. 您可以点击以收起/展示每条策略的详细内容：
   - » 配置写入主键冲突策略
     a. 点击以切换数据目的地节点
     b. 选择策略执行选项
   - » 配置结构变化策略
     a. 点击以切换数据目的地节点
     b. 选择策略执行选项
     策略选项中删除表与修改表结构的操作需要数据节点的相应权限，如执行过程中无相关权限，任务将报错暂停
   - » 配置增量处理策略
     a. 点击以切换数据目的地节点
     b. 选择策略执行选项

选项3中保留增量数据的处理需要配合数据映射的清洗脚本功能使用，详见<数据映射–清洗脚本–常见场景说明>

» 配置端到端一致性策略

a. 开启以保证数据端到端一致性

» 配置自动重启策略

a. 详见<数据链路–策略配置–自动重启策略>

» 配置错误队列策略

a. 详见<数据链路–策略配置–错误队列策略>

» 配置预警策略

a. 点击新增任务运行预警创建预警规则

b. 输入预警规则名称，选择预警指标，点击规则配置进行规则配置

c. 为预警规则选择预警发送组

1. 预警发送组配置，详见<系统设置–预警中心–预警发送组配置>

» 配置日志策略

a. 选择日志记录类别开关

b. 为每一类日志配置存储位置

此操作与错误队列数据存储类似，您可参照<数据链路策略配置–错误队列策略>。

注意事项

1. 任务执行过程中，修改主键冲突策略，结构变化策略，增量处理策略端到端一致性策略，任务将被暂停。
2. 选择该链路的数据任务将会遵循数据链路的策略配置，但在任务配置中，您可针对不直接影响任务运行的选项进行适应性调整。

下一步

- 数据任务配置<数据任务策略配置>

2.2.1.2.1. 数据链路策略配置–写入主键冲突策 Setting up Primary Key Conflict Policy

当任务写入过程中，写入数据与目的地数据有主键冲突，DataPipeline 将会按照您选择的策略对此种情况进行处理。

关于此功能

- 我们为不同的数据节点提供：
- 覆盖此条数据：根据主键执行update，覆盖有利于保证数据一致性；
- 忽略此条数据：不写入该条冲突数据，忽略有利于任务运行速率提升。

注意事项

1. 只有在数据目的地节点为DBMS类型节点，且数据表中有主键时，该策略才会生效。
2. 当您选择覆盖此条数据，可能存在update权限问题导致写入失败，如遇此类错误请您检查权限后再运行任务。

2.2.1.2.2. 数据链路策略配置–结构变化策略 Setting up Schema Change Policy

结构变化策略是当数据源数据结构发生变化时，系统将为您执行的策略，能够有效避免由于数据源结构变化使任务暂停带来的影响。

关于此功能

我们针对三种数据源结构变化情况提供结构变化策略，您可以为每一个节点选择自己的结构变化策略，以避免任务暂停。

情况	选项	选项说明
数据源删除正在同步的表	暂停数据任务	遇到此种情况暂停数据任务
	删除映射并删除数据目的地表和数据	删除对应数据链路中的数据映射，并删除数据目的地表（drop table）
	删除映射并保留数据目的地表和数据	删除对应数据链路中的数据映射，但不对数据目的地表中数据进行操作
	不删除映射，持续扫描并保留数据目的地表和数据	不删除对应数据链路中的数据映射，不对数据目的地表中数据进行操作，保留任务可能因为数据源表被误删后恢复同步的可能性
数据源正在同步的表新增字段	暂停数据任务	遇到此种情况暂停数据任务
	忽略数据源新增字段	数据目的地不新增该字段
	自动新增该字段，建立映射并同步数据	根据新增字段的信息自动建立数据映射，数据目的地应用新的数据映射同步数据，新增该字段（alter table）
数据源正在同步的表中字段被删除	暂停数据任务	遇到此种情况暂停数据任务
	删除映射并删除目的地表中字段和数据	删除对应数据链路中的字段映射，删除目的地表中该字段（alter table）
	删除映射但保留目的地表中字段和数据，并写入空值	删除对应数据链路中的字段映射，但不对数据目的地表结构和历史数据进行操作，后续同步向该字段写入空值

表格15 结构变化策略选项说明列表

注意事项

1. 结构变化策略需要对数据目的地有更改表结构（alter table）、删除表（drop table）权限，如无该类型权限，遇到数据源结构变化的情况，执行结构变化策略的任务将报错暂停。
2. 现版本支持结构变化策略的数据目的地有：MySQL、MS SQL Sever、Oracle、PostgreSQL。

2.2.1.2.3. 数据链路策略配置–增量处理策略 Setting up Incremental Data Policy

当数据源产生已同步的数据被删除这样的增量数据时，您可以通过配置增量处理策略来对这部分数据进行处理，保证数据一致性。

情况	选项	选项说明
数据源已同步的数据被删除	同步，删除目的地数据	将数据源已删除的数据在数据目的地删除
	忽略，保留目的地数据	不对数据目的地进行操作
	同步，保留增量数据，并按照merge模式处理	系统会将写入方式设置为update，之后按照主键将高级清洗产生的数据写入数据目的地删除标记字段中

表格16 增量处理策略选项说明列表

注意事项

1. 当您选择同步，保留增量数据，并按照merge模式处理 时，需要保证系统拥有数据目的地的update权限。
2. 该场景仅支持数据源删除的数据有主键的情况。

2.2.1.2.4. 数据链路策略配置–端到端一致性策略 Setting up End–to–End Data–consistency Policy

在任务运行过程中，可以开启端到端一致性策略来保证数据从源端到目的地端的一致性。

为了保证数据一致性，系统在数据源与数据目的地建立了完整的程序逻辑。

关于此功能

- **数据源**

系统从数据源读取数据后，会定期记录读取的进度，数据对应的进度被成功记录了，才会被允许写入到目的地。

- **目的地**

系统在每次成功执行写入操作后会记录已写入数据的进度。

  - » 如果是 JDBC 目的地，将会在目的地建立一张表进行记录；
  - » 如果是 FTP/HDFS 目的地，将会采取内部的二阶段提交协议，假如数据写入完成，进度记录失败，将会回滚(删除)已写入的目的地的数据；
  - » 如果是 Hive 目的地，将会记录进度 walog 到 hdfs，如果进度提交失败，将会回滚已写入目的地的数据；
  - » 如果是 Kafka 目的地，将会使用 Kafka 的事务功能，在进度被提交成功前，Kafka 内的数据无法消费，以此保证写入数据的数据一致性。

注意事项

当数据源为JDBC数据源时，我们将提供二阶段提交的方式进行数据写入，开启数据一致性选项，数据写入性能可能会受到影响。

2.2.1.2.5. 数据链路策略配置–自动重启策略 Setting up Auto Restart Policy

自动重启策略是在任务运行过程中当任务报错时对于系统是否重启数据任务所设置的策略。

关于此功能

我们默认任务报错需要重启，执行重启5次，时间间隔1分钟；

在自动重启列表中，系统提供了部分已验证重启无法自动解决，需人工介入的错误类型，不重启任务；

您也可以添加自定义的错误类型，当任务报错，错误堆栈包含您定义的错误类型时，任务将被暂停。

操作步骤

1. 点击策略配置按钮进入策略配置tab；
2. 开启自动重启策略开关：
   - » 开启即任务报错自动重启
   - » 关闭即所有任务报错均不会自动重启
3. 添加错误类型。
   - » 点击添加按钮，输入错误类型描述，错误堆栈片段，保存
   - » 可以添加多个错误堆栈片段，多个错误堆栈片段之间是or关系

注意事项

1. 我们将使用您提供的错误堆栈片段与任务报错的错误堆栈进行匹配，只要一个错误类型中的某一条错误堆栈片段匹配成功，该错误类型的自动重启策略将按照列表内配置的策略执行，反之则执行默认策略。

2. 数据任务中可以添加更多的错误类型以完善适应该任务的自动重启策略。
3. 错误堆栈信息是程序运行过程中，返回的错误信息，根据错误信息的特定内容可以定位错误并解决错误，使用错误堆栈片段添加错误类型，需要使用有特异性的错误堆栈信息。

2.2.1.2.6. 数据链路策略配置–错误队列策略 Setting up Data Error Queue Policy

错误队列策略是在任务运行过程中，产生错误数据时，系统为保证任务稳定运行，或数据传输准确性的策略配置。

关于此功能

开启错误队列策略，运行中的数据任务产生错误数据时，可以不暂停数据任务，将错误数据存储于指定节点，并记录错误堆栈信息。可有效避免因任务出现错误数据而暂停所带来的影响。

操作步骤

1. 点击错误队列存储，选择错误队列数据存储位置；
   - » 系统内置节点
   - » 外部节点

     · 在选择外部节点弹窗中选择新建表或选择已有表

     a. 新建表：系统根据错误队列数据存储需求帮助您建表

     b. 选择已有表：您的数据库中已经有了符合错误队列数据存储需求的数据表
2. 2. 选择是否存储错误堆栈信息；
3. 选择任务遇到错误数据后执行的策略。
   - » 根据错误数据条数多少暂停数据任务，或错误队列数据后置处理，不暂停数据任务

注意事项

存储错误队列数据时，选择内部节点与选择外部节点的优劣：

- 使用系统内置节点

  优势
  - » 系统内置节点减少配置步骤，系统代管错误队列数据，省时省心；
  - » 通常情况下系统内置节点连接稳定性较外部节点高，写入失败概率更小。

  劣势
  - » 系统节点不会对外开放，用户无法自行查询数据；
  - » 使用系统节点会占用系统部署服务器存储资源，当错误数据数据量超过1,000,000条时所有使用系统节点存储错误队列数据的数据任务将被暂停。
- 使用外部节点

  优势
  - » 外部节点存储错误队列数据，错误队列数据完全处于您的管辖范围内，查询数据、清空或备份历史数据操作都十分方便；
  - » 外部节点不受存储空间容量限制，若需要存储的错误队列数据过多，您可自行对节点进行扩容。

  劣势
  - » 使用外部节点可能存在连接不稳定的情况，无法写入错误数据，系统将会发出无法写入错误数据的报错，数据任务不会暂停。

2.2.1.2.7. 数据链路策略配置––预警策略 Setting up Alert Policy

通过配置预警策略，可以在任务出现预警指标定义的情况时，发送预警信息到指定的预警渠道。

关于此功能

支持的预警指标有：任务报错预警，任务配置变更预警，错误队列数据条数预警，数据源Kafka Topic Lag数据量预警，系统消息队列Topic Lag数据量预警。

操作步骤

1. 点击策略配置按钮进入策略配置tab；
2. 开启预警策略开关；
3. 点击添加预警策略按钮，进入添加预警策略弹窗；
4. 输入预警名称；
5. 选择预警指标；
6. 设定预警阈值；
7. 选择预警发送组；
8. 设置预警发送时间间隔；
9. 完成配置。

注意事项

1. 预警发送组可在系统配置—预警中心页面进行配置，支持Webhook与邮件渠道。
   详情见：2.4.3. 管理预警中心 Setting up and Configuring Alarm Center
2. 为了避免预警信息发送过于频繁，我们提供预警发送时间间隔设置，当您指定时间间隔为1小时，则在1小时内触发的多条相同预警只会被发送一次，此选项没有默认值。

2.2.1.2.8. 数据链路策略配置--日志策略 Setting up Log Policy

关于此功能

在日志策略中，我们提供两种数据任务运行日志配置，分别为：任务报错日志，任务配置变更日志。

◎ 任务报错日志：即任务在运行过程中出现错误导致暂停的报错信息的日志记录，包含错误堆栈信息。

◎ 任务配置变更日志：即任务配置发生变化的日志记录，包含任务配置变更的具体内容。

操作步骤

1. 点击策略配置按钮进入策略配置tab；
2. 开启日志策略策略开关；
3. 选择是否记录任务报错日志与任务配置变更日志；
4. 为两种日志分别选择存储节点；
5. 存储于系统内部节点；
6. 存储于外部节点；
7. 点击选择节点按钮，进入节点选择弹窗；
8. 选择存储节点并创建新表存储日志；
9. 如系统无创建表权限，您可点击查看建表语句按钮后，获取建表语句，手动建表后指定日志存储表。

注意事项

无论是使用系统节点或外部节点，系统默认链路中日志策略已选节点的日志存储数据量最大为100万条，如日志超过100万条，任务详情页面消息列表将进行提示，产生日志的数据任务将继续运行，但不会继续记录日志。

2.2.2. 修改数据链路 Editing a Data Pipeline

修改数据链路指修改已存在的数据链路的配置内容，您可以对数据链路的数据源、数据目的地、数据映射进行修改，以限定链路数据映射范围，您也可以修改数据链路策略配置，保证数据任务稳定运行。

关于此功能

修改数据链路配置不会检查链路状态，但是保存链路配置意味着我们需要将修改后的链路的配置下发至数据任务，故链路配置保存时是需要链路为挂起状态的。

操作步骤

1. 点击数据链路列表->查看详情，查看数据链路详情；
2. 选择数据源（可选多个）并选择增量/全量读取方式；
3. 选择数据目的地（可选多个）并选择写入方式；

4. 配置数据映射，操作步骤详见：2.2.1.1.2.2 配置字段映射 Setting up field mapping—NoSQL；
5. 保存基本配置；
6. 完成基本配置后，任务应用链路配置即可运行，但为保证任务运行的稳定性，我们推荐您完成链路策略配置后再通过数据任务使用链路配置。

注意事项

初次配置数据链路时需慎重选择数据源的读取方式，选择完成保存后，系统将不允许修改读取方式。

下一步

- 更加关注快速配置使任务快速运行：
  - » 激活数据链路
  - » 数据任务配置—选择数据链路
- 更加关注任务运行稳定：
  - » 数据链路配置–策略配置

2.2.3. 激活数据链路 Activating a Data Pipeline

激活数据链路指在系统管理链路状态的层级将数据链路激活，激活后的数据链路可被任务选择。

关于此功能

激活数据链路需要数据链路状态为挂起，当您激活数据链路，系统将会提示您链路的相关数据任务也可被激活。

操作步骤

1. 点击数据链路列表–>激活数据链路开关；
2. 弹出链路的相关数据任务也可被激活提示；
3. 点击确定即激活该数据链路。

注意事项

激活数据链路是在系统管理层级上对链路的操作，激活数据链路不会校验链路中的配置内容是否正确。

下一步

- 如您需要激活相关数据任务，可以在数据链路–相关数据任务页面批量激活相关数据任务。

2.2.4. 挂起数据链路 Deactivating a Data Pipeline

挂起数据链路指在系统管理层面将数据链路挂起，挂起后的数据链路在管理上呈现不可用状态。

关于此功能

挂起数据链路需要挂起链路相关的数据任务，挂起链路后，相关数据任务亦不可被激活。

操作步骤

1. 点击数据链路列表–>挂起数据链路开关；
2. 弹出挂起确认弹窗；
3. 挂起数据链路将会同步挂起链路相关数据任务；
4. 确定即挂起数据链路；
5. 系统会异步执行挂起相关数据任务的操作。

注意事项

挂起数据链路后，相关数据任务将被异步挂起，故可能出现挂起任务失败的情况；未成功挂起的任务将会继续运行，直到您手动将其暂停。

下一步

- 更加关注快速配置使任务快速运行：
  - » 激活数据链路
  - » 数据任务配置—选择数据链路
- 更加关注任务运行稳定：
  - » 数据链路配置–策略配置

2.2.5. 删除数据链路 Deleting a Data Pipeline

删除数据链路指将已经配置好的数据链路删除，删除后将不可恢复。

关于此功能

为了确保数据任务因操作带来的不稳定性降到最低，删除数据链路会校验链路中是否存在相关数据任务，如果还有相关数据任务，则不能删除数据链路。

操作步骤

1. 点击数据链路列表–>删除数据链路；
2. 如有链路有相关数据任务，则不能删除，删除按钮置灰；
3. 如删除按钮可点，则无相关数据任务，点击按钮弹出删除确认提示；
4. 在确认提示窗口中点击删除，即删除数据链路；
5. 已被删除的数据链路不可恢复。

> **注意事项**
>
> 对于系统稳定性来说， 删除数据链路是一项高危操作，请谨慎操作。

2.3. 管理数据任务 SETTING UP AND CONFIGURING DATA INTEGRATION

什么是数据任务

数据任务是DataPipeline同步进行数据同步的最小管理单位。数据任务支持全量同步模式与增量同步模式，满足不同的同步场景需求。

准确的任务配置是保证数据任务稳定运行的关键，运行相关的配置选项，例如数据源与数据目的地配置、数据映射、结构变化策略、主键冲突策略、增量处理策略、错误队列策略等，任务将完全应用链路的配置。与此同时，为了方便数据任务管理与运维的便捷性，在数据链路配置的基础上，数据任务可以自行定义自动重启策略、预警策略、日志策略、读取限制、写入限制、传输队列限制等配置选项。

丰富的配置种类为任务运行提供了稳定性保障，但配置选项过多也会对您理解并使用DataPipeline进行数据同步带来一定影响。

因此，从配置的逻辑层面我们将数据任务的配置分成了三个层级，基础配置、限制配置、策略配置。

◎ **数据任务基础配置**

数据任务基础配置是数据任务可以成功运行的最小配置，完成数据任务基础配置后，数据任务将即限制任务读取并发个数。

◎ **选择数据链路**

在新建数据任务过程中，您需选择数据链路，来应用数据链路的配置内容。

◎ **选择/修改同步列表**

选择同步列表是在您已经为数据任务选择数据链路后，数据任务可以获取到链路中已经配置好的全部映射内容，您需指定该任务同步的数据映射范围。

◎ **任务执行配置**

此配置指定任务执行方式，定时设置，全量初始化等配置选项。

◎ **任务资源配置**

此配置指定任务执行过程中使用的物理资源。

◎ **数据任务限制配置**

数据任务限制配置是为了保证数据任务稳定运行，对数据任务的读取、写入、传输队列进行限制的配置内容。

◎ **读取限制**

作用于数据任务对数据源的读取。

◎ **读取速率限制**

即限制任务读取速率。

◎ **读取并发限制**

即限制任务读取并发个数。

◎ **写入限制**

即限制任务读取并发个数。

◎ **写入速率限制**

即限制任务读取并发个数。

◎ **写入并发限制**

即限制任务读取并发个数。

◎ **Batch设置**

即限制任务读取并发个数。

◎ **传输队列限制**

» **传输队列最大缓存值**

即限制任务读取并发个数。

» **传输队列回收时间**

设置数据源读取和写入的传输队列回收时间，超过缓存时间的数据会被资源回收，从而造成数据丢失。

◎ **数据任务策略配置**

数据链路策略配置是解决任务运行问题的配置选项。

策略配置可以被分成两类：

» 一类是解决任务运行过程中可能会遇到的错误的系统自动的应对策略，其中包括：写入主键冲突策略、结构变化策略、主键冲突策略、增量处理策略、错误队列策略、自动重启策略等。

» 另一类是将任务运行信息记录，或将任务运行信息反馈至负责人，方便处理任务错误情况的功能，包括日志策略、预警策略，错误队列策略中的错误数据存储等。

◎ **写入主键冲突策略**

写入主键冲突是在任务写入过程中，写入数据与目的地数据有主键冲突，任务执行的应对策略，我们提供覆盖数据与忽略数据的选项。

◎ **结构变化策略**

结构变化策略是当数据源数据结构发生变化时，系统将为您执行的策略，能够有效避免由于数据源结构变化使任务暂停带来的影响。

◎ **增量处理策略**

当数据源产生已同步的数据被删除这样的增量数据时，您可以通过配置增量处理策略来对这部分数据进行处理，保证数据一致性。

◎ **端到端一致性策略**

在任务运行过程中，可以开启端到端一致性策略来保证数据从源端到目的地端的一致性。

◎ **自动重启策略**

自动重启策略指当任务报错时，系统将执行自动重启策略，以对数据任务的不同错误类型做出是否重启的调整。

◎ **错误队列策略**

开启错误队列策略，运行中的数据任务产生错误数据时，可以不暂停数据任务，将错误数据存储于指定节点，并记录错误堆栈信息。可有效避免因任务出现错误数据而暂停所带来的影响。

◎ **预警策略**

通过设置预警规则，选择预警发送组，预警策略可以帮助您实现对关注内容的预警配置，实现当任务出现预警超出规则限制的情况时，可以及时通过预警发送方式（包括邮件与WebHook）通知到您。

◎ **日志策略**

合理的日志策略可以有效帮助您降低查询与管理日志数据的时间投入，通过日志策略，您可以配置日志记录的类别与日志存储方式。

◎ **数据任务的基本管理**

新增数据任务

详见："2.3.1. 新增数据任务 Adding a Data Integration Task"

修改数据任务

详见："2.3.2. 修改数据任务 Editing a Data Integration Task"

删除数据任务

详见："2.3.3. 删除数据任务 Deleting a Data Integration Task"

激活数据任务

详见："2.3.4. 激活数据任务 Activating a Data Integration Task"

挂起数据任务

详见："2.3.5. 挂起数据任务 Deactivating a Data Integration Task"

◎ **通过项目管理数据任务**

数据任务项目分组指在数据任务首页可以对任务进行项目分组来进行统一管理，通过对数据任务的分组，您可以更清晰，更省时的进行任务管理，实现多项目任务解耦。

2.3.1. 新增数据任务 Adding a Data Integration Task

新增数据任务是指添加一项新的任务至系统中，您需要完成任务的基本设置，选择一条数据链路后完成基本配置信息可以正常运行。

关于此功能

新增数据任务操作通常由数据部门工作人员进行，您可以对数据任务选择的数据链路、数据映射进行配置，可以对任务执行配置、全量初始化、资源配置进行设置，您也可以对数据任务添加限制配置信息和策略配置信息，保证数据任务稳定运行。

操作步骤

1. 点击数据任务列表页→「新增数据任务」；
2. 输入数据任务名称及描述，点击「确定」进入数据任务基本配置页面；
3. 选择数据链路并配置数据映射；
4. 任务执行配置；
5. 全量初始化配置；
6. 资源配置；
7. 「保存」基本配置；
8. 完成基本配置后，任务即可运行，但为保证任务运行的稳定性，我们推荐您完成限制配置和策略配置。

> **注意事项**
>
> 初次配置数据任务时需谨慎选择数据链路，选择完成后，系统将不允许更改。

2.3.1.1. 数据任务-基础配置 Basic Configuration of Data Integration Task

数据任务基本配置是保证数据任务成功运行的基本配置。

关于此功能

数据任务基本配置包括选择数据链路与同步列表、任务执行配置、资源组配置。

- **选择数据链路与同步列表**

选择数据链路与同步列表是在数据任务中应用数据链路的配置选项并在已选数据链路中选择需要同步的数据映射的功能。

- **选择数据链路与同步列表**

选择数据链路与同步列表是在数据任务中应用数据链路的配置选项并在已选数据链路中选择需要同步的数据映射的功能。

- **任务执行配置**

任务执行配置是通过指定同步方式，执行时间，执行方式等配置选项指定任务执行方式的配置。

- **资源组配置**

资源组配置是任务执行过程中对数据源读取和数据目的地写入的资源管理功能。

2.3.1.1.1. 任务执行配置 Setting up Execution Configuration

任务执行配置是通过指定同步方式，执行时间，执行方式等配置选项指定任务执行方式的配置。

关于此功能

我们提供两种任务同步方式：

- 增量同步

  支持日志读取模式（CDC）；

  支持JDBC读取模式，通过增量识别字段获取增量数据。
- 全量同步
- 支持JDBC读取模式，读取全量数据。

操作步骤

一、日志读取模式（CDC）下配置增量同步

在任务详情页点击基础配置，进入基础配置Tab。

1. 选择增量同步；
2. 选择是否开启全量初始化；
   - » 开启：即增量同步之前先进行全量数据同步，将数据目的地数据覆盖

     开启后，您可在下方配置每个数据目的地的全量初始化方式
   - » 不开启：即直接进行增量数据同步，不对数据目的地数据进行操作

     · 此选项需要您指定增量同步读取起点

- 以MySQL增量同步监听模式即Binlog读取模式为例，需要填写：
  a. Binlog Position
  b. Binlog 文件名称
  c. GTID
- 当数据源节点在节点策略配置中开启检查点策略，您可通过系统指定同步起点（检查点策略详见–<数据节点策略配置–检查点策略>）

3. 指定任务首次执行时间；
4. 指定任务执行方式（日志读取模式（CDC）下仅可选监听模式）。

二、JDBC读取模式下下配置增量同步

在任务详情页点击基础配置，进入基础配置Tab。

1. 选择增量同步；
2. 选择是否开启全量初始化；
   - » 开启：即增量同步之前先进行全量数据同步，将数据目的地数据覆盖；
     开启后，您可在下方配置每个数据目的地的全量初始化方式
   - » 不开启：即直接进行增量数据同步，不对数据目的地数据进行操作
     - 此选项需要您指定增量同步读取起点
3. 指定任务首次执行时间；
4. 指定任务执行方式（JDBC读取方式下仅可选定时模式）。
   - » 指定定时频率

三、配置全量同步

在任务详情页点击基础配置，进入基础配置Tab。

1. 选择增量同步；
2. 指定任务首次执行时间；
3. 指定任务定时频率；
4. 配置数据目的地全量初始化。

注意事项

1. 任务暂停后，修改增量任务执行配置中的全量初始化开关、增量读取起点等配置，任务将会按照您修改后的执行配置方案重新执行，可能会覆盖已同步数据；
2. 增量同步–监听模式不同数据源读取起点配置项列表：

数据源	读取方式	读取起点配置
MySQL	Binlog	Binlog Position
		Binlog 文件名称
		GTID
SQL Server	Change Tracking	Change Tracking Version

数据源	读取方式	读取起点配置
PostgreSQL	wal2json	LSN
Oracle	Agent	请联系DataPipeline工程师

表格 17 数据源指定同步起点参数列表

2.3.1.1.2. 资源组配置 Setting up Resource Group

资源组配置是任务执行过程中对数据源读取和数据目的地写入的资源管理功能。

关于此功能

在部署DataPipeline时，通过修改配置文件，可以将数据源端/目的地端的服务器资源划分为多个资源组，实现业务资源组解耦。

操作步骤

1. 在部署DataPipeline时，前往路径/data/datapipeline/dpconfig/；
2. 修改配置文件resource_group_config.json；

配置项	说明
groupName	资源组名称，用于系统调用标识
displayName	资源组显示名称，用于前端选择
memory	分配给资源组的内存数量
memoryUnit	内存单位
cpu	分配给资源组的cpu核数
nodes	资源组对应节点地址列表

表格18 资源组配置文件字段说明表

3. 保存修改后重启服务，使配置文件生效；
4. 在数据任务基础配置页面，为数据源配置资源组；
5. 您可切换数据目的地节点，分别为不同的目的地节点设置资源组。

> **注意事项**
>
> 资源组隔离可以在一定程度上保证重要任务获取独立运行资源以更加稳定的运行，资源组模型详见<资源组模型>。

下一步

- 数据任务管理–激活数据任务

2.3.1.1.3.选择数据链路与同步列表 Selecting Data Pipeline and Sync List

选择数据链路与同步列表是在数据任务中应用数据链路的配置选项并在已选数据链路中选择需要同步的数据映射的功能。

关于此功能

新建数据任务时您需要选择数据链路，并在已选数据链路的数据映射中选择需要同步的数据映射。

操作步骤

1. 在新建任务过程中，输入任务名称及描述进入任务详情页面；
2. 点击选择链路，进入选择链路弹窗；
3. 选择一条已创建好的链路，或者创建一条链路，选择完成后，点击保存；
4. 回到任务详情页，基础配置，您可以选择同步列表：
   - » 选择同步列表即选择同步哪些已经配置好的数据映射；
   - » 当您选择了一个数据源的数据映射后，不允许选择其他数据源的数据映射。
5. 通过切换当前数据源对应的多个数据目的地，可以选择一对多数据同步。

> **注意事项**
>
> 1. 选择链路后，已选数据链路将被固化至数据任务，任务不可更改已选数据链路。
> 2. 通过修改同步列表，您可以调整任务同步数据的同步范围。
>   - » 任务暂停后，您可以添加任务同步的数据映射，任务启动时，将按照现有的任务执行配置执行新的数据映射同步。

下一步

- 数据任务基础配置–资源配置
- 数据任务管理–激活数据任务

2.3.1.2. 数据任务–限制配置 Restraining Configuration of Data Integration Task

数据任务基本配置是保证数据任务成功运行的基本配置。

关于此功能

数据任务的限制配置可以被分为三类：

- 读取限制
- 写入限制
- 输队列限制

操作步骤

1. 在数据任务详情页面点击限制配置按钮，切换至限制配置tab；

2. 选择任务读取速率限制方式：
   » 如选择手动限制需要输入限制值。
   · 支持的单位有：行/秒、数据大小/秒。
3. 选择任务读取速并发限制方式；
4. 选择任务写入速率限制方式；
5. 选择任务写入并发限制方式；
6. 配置写入batch限制：
   » 您可以通过条数，大小，时间来限制Batch规则，当前Batch写入成功后，下一Batch的满足条件判断归零，根据所选规则，达到任一规则上线则形成Batch写入。
7. 配置传输队列限制：
   » 配置传输队列缓存值限制
      a. 任务开始读取数据后，单个任务默认缓存10GB数据（读写数据量差），用户可自定义。
      b. 读写数据量差达到10GB（最大缓存值时），根据先进先出的原则，旧数据将会被回收。
      c. 当任务数据读写速率失衡，读写数据量差大于10GB（最大缓存值）时，将会出现部分数据被回收，未能成功写入数据目的地的情况。
   » 配置传输队列最大回收时间
      a. 任务开始读取数据后，但个任务默认缓存3天数据，用户可自定义。
      b. 缓存数据达到回收时间，旧数据将会被回收。
      c. 当任务数据读写速率失衡，超过回收时间的数据尚未被写入到目的地，将会出现部分数据被回收，未能成功写入数据目的地的情况。

> **注意事项**
> 在任务写入batch限制中，当您同时配置任务级别的Batch设置与映射级别的Batch设置，映射配置将会覆盖任务配置。

下一步

- 数据任务配置<数据任务策略配置>

2.3.1.3. 数据任务-策略配置 Policy Configuration of Data Integration Task

数据任务策略配置是在任务可以成功运行的基础上，对于任务运行过程中出现的问题进行应对的策略。

关于此功能

策略配置共包括8个细分策略，分别是：

- 写入主键冲突策略
- 结构变化策略
- 增量处理策略
- 端到端一致性策略

- 自动重启策略
- 错误队列策略
- 预警策略
- 日志策略

操作步骤

1. 通过数据任务列表/卡盘/拓扑图，进入数据链路详情；
2. 点击策略配置按钮，切换至策略配置页面；
3. 您可以点击以收起/展示每条策略的详细内容：
   - » 配置写入主键冲突策略
     a. 点击以切换数据目的地节点；
     b. 选择策略执行选项。
   - » 配置结构变化策略
     a. 点击以切换数据目的地节点；
     b. 选择策略执行选项。
     · 策略选项中删除表与修改表结构的操作需要数据节点的相应权限，如执行过程中无相关权限，任务将报错暂停。
   - » 配置增量处理策略
     a. 点击以切换数据目的地节点；
     b. 选择策略执行选项。
     · 选项3中保留增量数据的处理需要配合数据映射的清洗脚本功能使用，详见<数据映射–清洗脚本–常见场景说明>。
   - » 配置端到端一致性策略
     · 开启以保证数据端到端一致性。
   - » 配置自动重启策略
     · 详见<数据链路–策略配置–自动重启策略>。
   - » 配置错误队列策略
     · 详见<数据链路–策略配置–错误队列策略>。
   - » 配置预警策略
     a. 点击新增任务运行预警创建预警规则；
     b. 输入预警规则名称，选择预警指标，点击规则配置进行规则配置；
     c. 为预警规则选择预警发送组。
     · 预警发送组配置，详见<系统设置–预警中心–预警发送组配置>。
   - » 配置日志策略
     a. 选择日志记录类别开关；
     b. 为每一类日志配置存储位置。
     · 此操作与错误队列数据存储类似，您可参照<数据链路策略配置–错误队列策略>。

注意事项

1. 任务执行过程中，修改主键冲突策略，结构变化策略，增量处理策略端到端一致性策略，任务将被暂停。
2. 数据任务将会follow已选数据链路的策略配置，但在任务配置中，您可针对不直接影响任务运行的选项进行适应性调整，可调整内容：
   - » 自动重启策略---添加错误类型
   - » 预警策略---添加预警规则
   - » 日志策略---添加日志记录类别

下一步

- 数据任务激活<数据任务管理-激活数据任务>

2.3.1.3.1. 数据任务策略配置-主键冲突策略 Setting up Primary Key Conflict Policy

数据任务主键冲突策略将完全遵循已选数据链路的主键冲突策略。

配置数据链路主键冲突策略详见："2.2.1.2.1. 数据链路策略配置-写入主键冲突策 Setting up Primary Key Conflict Policy"

2.3.1.3.2. 数据任务策略配置-增量处理策略 Setting up Incremental Data Policy

数据任务增量处理策略将完全遵循已选数据链路的增量处理策略。

配置数据链路增量处理策略详见："2.2.1.2.3. 数据链路策略配置-增量处理策略 Setting up Incremental Data Policy"

2.3.1.3.3. 数据任务策略配置-结构变化策略 Setting up Schema Change Policy

数据任务支持结构变化策略的应用，结构变化策略及策略中用户的配置方式将完全使用已选数据链路的结构变化策略，但不在数据任务策略任务中显示。

配置数据链路结构变化策略详见："2.2.1.2.2. 数据链路策略配置-结构变化策略 Setting up Schema Change Policy"

2.3.1.3.4. 数据任务策略配置-端到端一致性策略 Setting up End-to-End Data-consistency Policy

在配置数据任务基础配置时，您需要选择数据链路，选择数据链路后，数据任务策略配置-端到端一致性策略将应用数据链路的端到端一致性策略的配置内容。

- 如数据链路要求数据任务开启端到端一致性策略，则数据任务必须开启端到端一致性策略；
- 如数据链路不要求数据任务开启端到端一致性策略，数据任务可以开启端到端一致性策略。

配置数据链路结构变化策略详见："2.2.1.2.2. 数据链路策略配置-结构变化策略 Setting up Schema Change Policy"

2.3.1.3.5. 数据任务策略配置–预警策略 Setting up Alerting Policy

关于此功能

通过配置预警策略，可以在任务出现预警指标定义的情况时，发送预警信息到指定的预警渠道。

支持的预警指标有：任务报错预警，任务配置变更预警，错误队列数据条数预警，数据源Kafka Topic Lag数据量预警，系统消息队列Topic Lag数据量预警。

操作步骤

1. 点击策略配置按钮进入策略配置tab；
2. 开启预警策略开关；
3. 点击添加预警策略按钮，进入添加预警策略弹窗：
   a. 输入预警名称
   b. 选择预警指标
   c. 设定预警阈值
   d. 选择预警发送组
   e. 设置预警发送时间间隔
4. 完成配置。

注意事项

1. 预警发送组可在系统配置—预警中心页面进行配置，支持Webhook与邮件渠道，*详情见：2.2.1.2.7 数据链路策略配置––预警策略 Setting up Alerting Policy*
2. 在配置数据任务基础配置时，您需要选择数据链路，选择数据链路后，数据任务策略配置–预警策略将自动将已选数据链路中的预警策略带入数据任务。
3. 数据链路与数据任务将分别根据各自的预警规则发送预警，预警信息可能会被重复发送。
4. 为了避免预警信息发送过于频繁，我们提供预警发送时间间隔设置，当您指定时间间隔为1小时，则在1小时内触发的多条相同预警只会被发送一次，此选项没有默认值。

2.3.1.3.6. 数据任务策略配置–日志策略 Setting up Logging Policy

在日志策略中，我们提供两种数据任务运行日志配置，分别为：任务报错日志，任务配置变更日志。

- 任务报错日志：即任务在运行过程中出现错误导致暂停的报错信息的日志记录，包含错误堆栈信息。
- 任务配置变更日志：即任务配置发生变化的日志记录，包含任务配置变更的具体内容。

操作步骤

1. 点击策略配置按钮进入策略配置tab；
2. 开启日志策略策略开关；
3. 选择是否记录任务报错日志与任务配置变更日志；
4. 为两种日志分别选择存储节点：

» 存储于系统内部节点

» 存储于外部节点

a. 点击选择节点按钮，进入节点选择弹窗；

b. 选择存储节点并创建新表存储日志。

· 如系统无创建表权限，您可点击查看建表语句按钮后，获取建表语句，手动建表后指定日志存储表。

注意事项

1. 无论是使用系统节点或外部节点，系统默认链路中日志策略已选节点的日志存储数据量最大为100万条，如日志超过100万条，任务详情页面消息列表将进行提示，产生日志的数据任务将继续运行，但不会继续记录日志。
2. 在配置数据任务基础配置时，您需要选择数据链路，选择数据链路后，数据任务策略配置–日志策略将自动应用数据链路要求的日志策略。
3. 如数据链路要求数据任务开启某项日志记录，则数据任务必须开启该日志记录；如数据链路关闭某项日志记录，数据任务可以自行开启该项日志记录。

2.3.1.3.7. 数据任务策略配置–自动重启策略 Setting up Auto Restart Policy

自动重启策略是任务运行过程中，当任务报错时，系统是否重启数据任务的策略。

关于此功能

我们默认任务报错需要重启，执行重启5次，时间间隔1分钟。

在自动重启列表中，系统提供了部分已验证重启无法自动解决，需人工介入的错误类型，不重启任务。

您也可以添加自定义的错误类型，当任务报错，错误堆栈包含您定义的错误类型时，任务将被暂停。

操作步骤

1. 点击策略配置按钮进入策略配置tab；
2. 开启自动重启策略开关；
   a. 开启即任务报错自动重启；
   b. 关闭即所有任务报错均不会自动重启。
3. 添加错误类型。
   a. 点击添加按钮，输入错误类型描述，错误堆栈片段，保存；
   b. 可以添加多个错误堆栈片段，多个错误堆栈片段之间是or关系。

注意事项

1. 我们将使用您提供的错误堆栈片段与任务报错的错误堆栈进行匹配，只要一个错误类型中的某一条错误堆栈片段匹配成功，该错误类型的自动重启策略将按照列表内配置的策略执行，反之则执行默认策略。
2. 数据任务中可以添加更多的错误类型以完善适应该任务的自动重启策略。
3. 错误堆栈信息是程序运行过程中报错，返回的错误信息，根据错误信息的特定内容可以定位错误并解决错误，使用错误堆栈片段添加错误类型，需要使用有特异性的错误堆栈信息。

2.3.1.3.8. 数据任务策略配置–错误队列策略 Setting up Error Queue Strategy

错误队列策略是在任务运行过程中，产生错误数据时，系统为保证任务稳定运行，或数据传输准确性的策略配置。

关于此功能

开启错误队列策略，运行中的数据任务产生错误数据时，可以不暂停数据任务，将错误数据存储于指定节点，并记录错误堆栈信息。可有效避免因任务出现错误数据而暂停所带来的影响。

操作步骤

1. 点击错误队列存储，选择错误队列数据存储位置：
   - 系统内置节点
   - 外部节点
     - » 在选择外部节点弹窗中选择新建表或选择已有表。
       a. 新建表：系统根据错误队列数据存储需求帮助您建表；
       b. 选择已有表：您的数据库中已经有了符合错误队列数据存储需求的数据表。
2. 选择是否存储错误堆栈信息；
3. 选择任务遇到错误数据后执行的策略。
   - » 根据错误数据条数多少暂停数据任务，或错误队列数据后置处理，不暂停数据任务。

注意事项

存储错误队列数据时，选择内部节点与选择外部节点的优劣如下。

- 使用系统内置节点：

  优势

  - » 系统内置节点减少配置步骤，系统代管错误队列数据，省时省心；
  - » 通常情况下系统内置节点连接稳定性较外部节点高，写入失败概率更小。

  劣势

  - » 系统节点不会对外开放，用户无法自行查询数据；
  - » 使用系统节点会占用系统部署服务器存储资源，当错误数据数据量超过1,000,000条时所有使用系统节点存储错误队列数据的数据任务将被暂停。
- 使用外部节点：

  优势

  - » 外部节点存储错误队列数据，错误队列数据完全处于您的管辖范围内，查询数据、清空或备份历史数据操作都十分方便；
  - » 外部节点不受存储空间容量限制，若需要存储的错误队列数据过多，您可自行对节点进行扩容。

  劣势

  - » 使用外部节点可能存在连接不稳定的情况，无法写入错误数据，系统将会发出无法写入错误数据的报错，数据任务不会暂停。

2.3.2. 修改数据任务 Editing a Data Integration Task

修改数据链路是指修改已存在的数据任务的配置内容，您可以对数据任务的数据映射配置进行修改，也可以对任务执行配置、全量初始化、资源配置进行修改，您也可以修改数据任务的限制配置信息和策略配置信息，保证数据任务稳定运行。

关于此功能

修改数据任务配置需要任务处于暂停状态，任务配置中，已选数据链路与部分链路策略配置会影响任务运行，故不可修改；其他内容开放修改。

操作步骤

1. 点击数据任务列表→「详情」，查看数据任务详情；
2. 修改数据映射配置；
3. 任务执行配置；
4. 全量初始化配置；
5. 资源配置；
6. 「保存」基本配置；
7. 完成基本配置后，任务即可运行，但为保证任务运行的稳定性，我们推荐您完成限制配置和策略配置。

> **注意事项**
>
> 初次配置数据任务时需谨慎选择数据链路，选择完成后，系统将不允许更改。

下一步

- 数据任务管理–激活数据任务

2.3.3. 删除数据任务 Deleting a Data Integration Task

删除数据任务是指将删除一条现有的数据任务，删除后将不可恢复。

关于此功能

删除数据任务要求数据任务处于非激活状态，删除数据任务不会删除已同步的数据。

操作步骤

1. 点击数据任务列表或者数据任务详情页→「删除」数据任务；
2. 任务进行中不可删除，需要暂停任务；
3. 点击「删除」需要弹窗二次确认提示；
4. 在提示窗口中点击「删除」，即删除数据任务；
5. 已被删除的数据任务不可恢复。

> **注意事项**
>
> 因DataPipeline数据任务状态管理与任务实际运行为异步操作，暂停和删除数据任务在实际执行过程中可能会产生延迟，详见<产品说明–组件模型>。

2.3.4. 激活数据任务 Activating a Data Integration Task

激活数据任务是数据任务状态管理中将数据任务从未激活状态转换成激活（待执行或执行中状态）的功能。

关于此功能

激活数据任务后，任务会根据其执行配置开始执行，进入待执行或执行中的状态。

操作步骤

1. 在数据任务列表或数据任务详情页面→点击激活任务开关；
2. 任务即被激活。

> **注意事项**
>
> 1. 在数据任务列表中，通过勾选多个数据任务，您可激活多个数据任务，同时激活多个数据任务可能带来服务器压力激增，请您谨慎操作。
> 2. 已暂停的数据任务修改配置后激活，任务将执行新的配置。
>   - 当您为暂停的任务添加数据映射，激活任务后，该映射将使用现有的任务执行配置运行：
>   a. 当任务执行配置为增量同步–监听模式，该映射将按照您初始选择的初始化配置进行初始化；
>   b. 当任务执行配置为全量同步或增量同步–定时模式，该映射将按照任务当前全量初始化配置执行。

2.3.5. 挂起数据任务 Deactivating a Data Integration Task

挂起数据任务是通过任务状态管理将正在运行的数据任务暂停的操作。

关于此功能

当数据任务报错时，任务将被自动挂起，您亦可手动挂起数据任务来暂停数据传输；任务被挂起时，您可修改任务配置，激活任务后，修改的配置选项将被执行。

操作步骤

1. 在数据任务列表或数据任务详情页点击挂起任务开关；
2. 数据任务被挂起。

> **注意事项**
>
> 1. 数据任务的状态管理模块是处理任务激活挂起等执行状态的模块，与页面管理的网站服务模块存在异步处理关系，所以任务挂起操作可能会存在失败的情况，当任务挂起失败，任务将继续运行，直到您下一次手动挂起任务。

注意事项

2. 挂起数据任务后，您可修改任务配置。
   - » 修改任务执行配置及映射相关信息后的任务执行，详情参照<数据任务管理–激活数据任务>。

2.3.6. 数据任务概览 Overview of Data Integration Task

数据任务概览是总览数据任务的页面，在页面中可以对任务进行项目分组，重要标记等操作。

关于此功能

我们通过拓扑图的形式展现全部的数据任务及项目中的数据任务的运行情况，数据源到数据目的地的唯一线条代表一个任务本身，线段粗细是任务传输的数据量抽象映射。

操作步骤

通过数据任务拓扑图进入数据任务详情：

1. 在数据任务概览页面点击数据任务拓扑图中的任务线段，弹出任务卡片；
2. 点击任务卡片，进入数据任务详情。

查看数据源/数据目的地的相关数据任务：

1. 在数据任务概览页面点击数据任务拓扑图中的数据源节点或数据目的地节点，弹出相关数据任务卡片；
2. 点击任务卡片，进入数据任务详情。

注意事项

通过项目对数据任务进行分组详见：数据任务管理–通过项目管理数据任务。

2.3.7. 数据任务监控 Monitoring Data Integration Task

在数据任务监控模块用户可以在任务主页按照重要任务、故障任务、异常任务等不同任务状态及属性直接监控重要任务运行情况、分析故障任务原因、关注异常任务及低性能任务，提升任务管理效率及问题处理效率。

关于此功能

分以下几个模块，请点击下方链接查看具体的模块详情介绍：

"2.3.7.1. 重要任务 Important Data Integration Task"

"2.3.7.2. 故障任务 Failed Data Integration Task"

"2.3.7.3. 非激活状态 Deactivated Data Integration Task"

"2.3.7.4. 性能关注 Performance Issue List"

2.3.7.1. 重要任务 Important Data Integration Task

重要任务是指用户对可以看到的（有权限可浏览、可编辑的任务）所有任务允许「添加为重要任务」。

关于此功能

- 添加星标的任务显示在数据传输页面的「重要任务」模块。
- 用户可在任务卡片中或任务详情页中添取消星标。
- 重要任务模块的排序顺序为：按照创建任务的时间倒序排序。
- 用户对任意任务「添加重要任务」后，重要任务将对所有用户可见。

操作步骤

1. 点击数据任务列表页→点击「添加重要任务」图标；
2. 任务详情页→点击「添加为重要任务」图。

> **注意事项**
>
> 用户对任意任务「添加重要任务」后，重要任务将对所有用户可见。

2.3.7.2. 故障任务 Failed Data Integration Task

重要任务是指用户对可以看到的（有权限可浏览、可编辑的任务）所有任务允许「添加为重要任务」。

关于此功能

- 故障任务主要展现数据任务收到错误通知没有被处理的任务。

操作步骤

1. 点击数据任务卡片→进入详情页；
2. 消息列表→查看具体错误。

> **注意事项**
>
> - 任务出现故障会报错并暂停运行。
> - 故障任务排序顺序为：任务创建时间。

2.3.7.3. 非激活状态 Deactivated Data Integration Task

非激活状态指显示状态为「待完善」、「未激活」、「已暂停」的任务。

关于此功能

非激活状态统一展示所有状态为「待完善」、「未激活」、「已暂停」的任务，可以对非激活状态任务进行配置。

操作步骤

- 点击任务监控→选择「非激活状态」。

下一步

- 如果您配置完需要激活数据任务，可以在数据任务列表页激活数据任务。

2.3.7.4. 性能关注 Performance Issue List

性能关注是将运行较慢的数据任务集中展示的功能。

关于此功能

增量任务–监听模式：

- 展示内容：任务名称、延迟时间、读取速率、写入速率、「查看」按钮、「忽略」按钮。
- 延迟时间：对比每一张表最近写入数据的时间减去该数据生产时间，并显示该任务所有的表中延迟时间最大的值。
- 「查看」按钮：用户点击「查看详情」，进入该任务详情页面。
- 「忽略」按钮：用户点击「忽略」，该任务要求立即在性能关注列表中隐去，在已忽略任务中显示。
- 检查范围为，用户自己创建和参与的所有实时传输数据任务（管理员则排序范围为全部实时传输任务）
- 所有忽略服务可在「已忽略任务」找到，可点击「取消忽略」把该任务放到性能关注范围中。
- 按传输速率降序排列。
- 排序的任务数量默认为10个任务。

增量同步–定时模式&全量同步：

- 展示内容：任务名称、完成时长、完成批次、传输速率、「查看」按钮。
- 完成时长：显示该任务最近一次定时传输完成时间减去开始时间的值。
- 完成批次：显示该任务被激活后已经完成的批次数量值。
- 传输速率：最近一次定时传输数据量/最近一次传输完成时长。
- 按照「传输速率」作为排序依据，「传输速率」最大的排在第一位。
- 排序的任务数量默认为10个任务。
- 检查范围为，用户自己创建和参与的所有定时传输数据任务（管理员则排序范围为全部定时传输任务。

2.3.8. 通过项目管理数据任务 Managing Data Fusion Task through Through Project

数据任务项目分组指在数据任务首页可以对任务进行项目分组来进行统一管理，通过对数据任务的分组，您可以更清晰，更省时的进行任务管理，实现多项目任务解耦。

关于此功能

通过对数据传输任务的分组，您可以更清晰，更省时的进行任务管理，实现多项目任务解耦。

操作步骤

1. 数据任务主页→左侧点击「加号」新增项目分组；
2. 弹窗输入项目名称，点击「新增」完成；
3. 点击项目卡片勾选框可以对项目内进行移动；
4. 点击「移动」，弹窗选择移动位置；
5. 弹窗点击「移动」，完成任务移动。

2.3.9. 管理任务错误数据 Managing Error Quene of Data Fusion Task

错误队列处理是用于处理任务在运行过程中产生的错误数据的功能，此功能与<数据链路策略配置–错误队列策略>相关。

关于此功能

我们提供两种处理错误数据的方式：

- 重新同步：即再次尝试写入该条数据，此选项通常作用于数据目的地映射配置不正确的情况。
- 忽略：即不在系统中处理该条错误数据，此选项需要您手动处理错误数据。

操作步骤

1. 重新同步–映射错误数据
   a. 在数据任务详情页面错误队列Tab点击重新同步该映射数据，或勾选数据映射后点击全部重新同步；
   b. 点击即开始执行。
2. 重新同步–单条错误数据
   a. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面；
   b. 点击重新同步该条错误数据，或勾选错误数据后点击全部重新同步；
   c. 点击即开始执行。
3. 忽略–映射错误数据
   a. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面；
   b. 点击忽略该条错误数据，或勾选错误数据后点击全部忽略；
   c. 错误数据详情中，错误处理状态将变更为已处理。
4. 忽略–单条错误数据
   a. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面；
   b. 点击忽略该条错误数据，或勾选错误数据后点击全部忽略；
   c. 错误数据详情中，错误处理状态将变更为已处理。
4. 清除–映射错误数据
   在数据任务详情页面错误队列Tab点击清除该映射错误数据，或勾选数据映射后点击全部清除。
5. 清除–单条错误数据
   a. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面；

2.3.9. 管理任务错误数据 Managing Error Quene of Data Fusion Task

错误队列处理是用于处理任务在运行过程中产生的错误数据的功能，此功能与<数据链路策略配置–错误队列策略>相关。

关于此功能

我们提供两种处理错误数据的方式：

- 重新同步：即再次尝试写入该条数据，此选项通常作用于数据目的地映射配置不正确的情况。
- 忽略：即不在系统中处理该条错误数据，此选项需要您手动处理错误数据。

操作步骤

1. 重新同步–映射错误数据
   a. 在数据任务详情页面错误队列Tab点击重新同步该映射数据，或勾选数据映射后点击全部重新同步；
   b. 点击即开始执行。
2. 重新同步–单条错误数据
   a. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面；
   b. 点击重新同步该条错误数据，或勾选错误数据后点击全部重新同步；
   c. 点击即开始执行。
3. 忽略–映射错误数据
   a. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面；
   b. 点击忽略该条错误数据，或勾选错误数据后点击全部忽略；
   c. 错误数据详情中，错误处理状态将变更为已处理。
4. 忽略–单条错误数据
   a. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面；
   b. 点击忽略该条错误数据，或勾选错误数据后点击全部忽略；
   c. 错误数据详情中，错误处理状态将变更为已处理。
4. 清除–映射错误数据
   在数据任务详情页面错误队列Tab点击清除该映射错误数据，或勾选数据映射后点击全部清除。
5. 清除–单条错误数据
   a. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面；
   b. 点击清除该条错误数据，或勾选错误数据后点击全部清除。
6. 批量导出错误数据
   a. 在数据任务详情页面错误队列Tab勾选数据映射后点击导出错误数据；
   b. 在数据任务详情页面错误队列Tab点击查看详情，进入该映射的错误数据详情页面，勾选错误数据后点击导出错误数据；
   c. 在导出错误数据弹窗选择导出数据格式；
      » 当前支持csv、JSON
   d. 点击确定即导出错误数据。

注意事项

在错误数据详情页面，您可以通过错误类型、错误发生时间、错误处理状态排序/筛选错误数据，筛选后的全选操作范围为被筛选的内容。

下一步

- 激活数据任务<数据任务管理–激活数据任务>

2.3.10. 重新同步映射数据 Resync Data Mapping

重新同步映射数据是在数据任务中以映射为单位的重新同步功能。

关于此功能

重新同步将重新读取数据源数据，清空该映射中数据目的地表中数据后，写入新数据。

操作步骤

1. 在数据链路详情页面同步映射详情中勾选需要重新同步的数据映射；
2. 点击重新同步，进入弹窗；
3. 在重新同步弹窗，选择清空数据目的地数据；
4. 点击重新同步则开始执行。

> **注意事项**
>
> - 任务暂停状态下不提供重新同步功能。
> - 重新同步默认会清空数据目的地表中数据，清除数据可以保证数据唯一性，且可以提升任务性能。

下一步

- 激活数据任务<数据任务管理–激活数据任务>

2.4. 管理系统设置 SETTING UP THE SYSTEM CONFIGURATION

系统设置包括用户管理、预警中心、邮件服务器配置等功能模块，可以为您提供用户管理、用户组管理、权限管理、预警发送配置、邮件服务配置、个人信息更改等功能。

◎ **用户管理**

用户设置：用户设置是用来管理系统用户的设置模块，提供新增用户、冻结用户、重置密码、更改用户分组的功能。

小组设置：小组设置是用户分组功能，提供新建分组、查看分组成员的功能。

◎ **个人设置**

个人设置是个人信息的设置功能，提供个人信息（邮箱、系统登录密码）更改、产品授权查看与更改、退出登录的功能。

◎ **预警中心**

预警中心是系统预警配置管理的功能，提供以下功能模块。

预警发送组设置：预警发送组设置是将预警发送配置集中分组的功能，预警发送渠道包括WebHook与预警邮件。

WebHook配置：WebHook配置是预警渠道配置的一种，通过定义WebHook的URL、Header、Body，您可配置一种WebHook发送渠道。

预警邮件配置：预警邮件配置是预警渠道配置的一种，通过定义收件邮箱地址、邮件标题、邮件内容，您可配置一种邮件发送渠道。

◎ **邮件服务配置**

邮件服务配置是系统邮件发送服务的配置，通过指定邮件服务器地址、端口号、认证方式、邮件服务协议、发件用户名与密码，您可配置系统发送邮件的服务配置。

2.4.1. 用户管理 User Management

重新同步映射数据是在数据任务中以映射为单位的重新同步功能。

关于此功能

- 用户设置

用户设置是用来管理系统用户的设置模块，提供新增用户、冻结用户、重置密码、更改用户分组的功能。

- 小组设置

小组设置是用户分组功能，提供新建分组、查看分组成员的功能。

操作步骤

- 添加用户
  1. 以管理员身份登录系统；
  2. 在系统设置–用户管理–用户设置页面点击「新增用户」；
  3. 填写用户名、邮箱地址、密码、用户分组；

4. 点击创建即完成添加。

- 冻结用户
  1. 以管理员身份登录系统；
  2. 在系统设置–用户管理–用户设置页面选中用户，点击「冻结用户」。
- 重置密码
  1. 以管理员身份登录系统；
  2. 在系统设置–用户管理–用户设置页面选中用户，点击「重置密码」；
  3. 输入新密码。
- 更改用户分组
  1. 以管理员身份登录系统；
  2. 在系统设置–用户管理–用户设置页面选中用户，点击「修改」；
  3. 选择新用户组。
- 添加用户组
  1. 以管理员身份登录系统；
  2. 在系统设置–用户管理–小组设置页面点击「添加小组」；
  3. 输入小组名称；
  4. 点击保存即添加完成。

注意事项

- 系统默认提供管理员组与公共组：
  - 管理员组拥有系统管理权限，可以查看系统中所有数据节点、数据链路、数据任务，可以管理用户。
  - 公共组只能查看自己创建或参与的数据节点、数据链路、数据任务。

2.4.2. 管理预警中心 Setting up and Configuring Alarm Center

预警中心是配置系统预警、系统通知、预警渠道与发送的功能模块。

关于此功能

- 预警发送组设置

预警发送组设置是将预警发送配置集中分组的功能，预警发送渠道包括WebHook与预警邮件。

- WebHook配置

WebHook配置是预警渠道配置的一种，通过定义WebHook的URL、Header、Body，您可配置一种WebHook发送渠道。

- 预警邮件配置

预警邮件配置是预警渠道配置的一种，通过定义收件邮箱地址、邮件标题、邮件内容，您可配置一种邮件发送渠道。

操作步骤

- 添加/编辑预警发送组
  1. 在系统设置–预警中心–预警发送组页面点击「添加预警发送组」；
  2. 输入预警发送组名称；
  3. 添加预警收件人；
     - » 收件人可以为邮箱或WebHook。
  4. 点击确定即完成添加/编辑。
- 删除预警发送组

  在系统设置–预警中心–预警发送组页面点击「删除预警发送组」。
- 添加/编辑Webhook预警发送渠道
  1. 在系统设置–预警中心–Webhook配置页面点击「添加Webhook」；
  2. 输入Webhook名称与URL；
  3. 添加Header；
  4. 编辑Body；
  5. 点击确定即完成添加/编辑。

  注：系统提供四种预警内容参数用来替换Body中的预警内容：

参数	说明
${预警标题}	发送预警内容的标题
${预警类型}	预警类型包括数据链路策略配置–预警策略中已支持的预警类型
${预警来源}	预警来源包括链路预警与任务预警
${预警内容}	预警内容是不同预警类型对应的系统预定义的预警内容模板

表格19　系统预警内容参数说明表

- 删除Webhook预警渠道

  在系统设置–预警中心–Webhook配置页面点击「删除Webhook」。
- 启用与停用Webhook预警渠道
- 在系统设置–预警中心–Webhook配置页面点击每项Webhook的开关。
- 发送测试Webhook预警渠道
  1. 在系统设置–预警中心–Webhook配置页面点击「发送测试」；
  2. 系统将会按照您配置好的Webhook进行发送测试，可用于Webhook调试。
- 添加/编辑预警邮件发送渠道
  1. 在系统设置–预警中心–预警邮件配置页面点击「添加邮件收件人」；
  2. 填写用户名称；
  3. 填写收件地址；
  4. 填写邮件标题；
  5. 填写邮件内容。

  注：系统提供四种预警内容参数用来替换邮件标题与邮件内容中的预警内容，详见表格19。

- 删除预警邮件发送渠道

  在系统设置–预警中心–预警邮件配置页面点击「删除」。
- 启用与停用Webhook预警渠道
- 在系统设置–预警中心–Webhook配置页面点击每项Webhook的开关。
- 发送测试Webhook预警渠道
  1. 在系统设置–预警中心–Webhook配置页面点击「发送测试」；
  2. 系统将会按照您配置好的Webhook进行发送测试，可用于Webhook调试。
- 添加/编辑预警邮件发送渠道
  1. 在系统设置–预警中心–预警邮件配置页面点击「添加邮件收件人」；
  2. 填写用户名称；
  3. 填写收件地址；
  4. 填写邮件标题；
  5. 填写邮件内容。

  注：系统提供四种预警内容参数用来替换邮件标题与邮件内容中的预警内容，详见表格19。

> **注意事项**
>
> 通过配置Webhook预警渠道功能可以支持钉钉、企业微信等IM工具的预警消息发送，详见<Webhook预警配置模板–钉钉>。

2.4.3. 配置邮件服务 Setting up Email Services

重新同步映射数据是在数据任务中以映射为单位的重新同步功能。

关于此功能

系统私有化部署于您的服务器后，可以使用指定的邮箱通过指定的邮件服务器发送通知邮件与预警邮件。

操作步骤

1. 在系统配置–邮件服务配置页面点击编辑；
2. 输入邮件服务器地址与端口号；
3. 选择认证方式；
4. 选择邮件服务协议；
5. 填写发件用户名与密码；
6. 点击保存即完成配置。

COMPANY NAME

3.管理控制台应用程序接口
MANAGEMENT CONSOLE API

DataPipeline实时数据融合产品使用Swagger管理驾驶舱应用程序接口，当您需要使用产品内部接口时，可以联系DataPipeline工程师，以开放内部应用程序接口管理页面。

如您生产环境无法开启额外的web服务，您可尝试使用测试环境开启应用程序接口管理页面，或联系DataPipeline工程师，获取PDF版本接口文档。

4. 常见问题
FREQUENTLY ASKED QUESTIONS

> 系统配置过程中与运行过程中的常见问题分析与解释。

Q: Docker安装的集群部署方式?

DataPipeline产品是采用docker容器的部署方式，支持docker集群；单机是docker compose部署，集群是docker swarm部署。

Q: DataPipeline的并发任务是线程还是进程?

线程。源端一个任务是一个source task，用线程池做并发，目的端一个任务是多个独立的sink task，每个sink task是一个线程。运行环境支持分布式部署，根据需要起一个或者多个source worker和sink worker实例，每个实例是独立的jvm进程，运行在容器里。sink端用多个consumer提升消费性能，kafka connect本身提供了这种并行能力，所以就不需要自己做线程池了。source端如果做成独立的task，task之间的协调需要额外的通信，否则简单hash分配的话容易不均衡，所以是一个任务一个source task。

Q: 生产环境配置推荐及回答?

一个32G节点是30个任务上限；生产环境建议Kafka使用至少两个副本；实际吞吐量受到上下游多重因素影响，我们只能给出的是Kafka的理论值，也就是平台的理论吞吐上限；多节点的主要优势在于高可用和同时运行的任务数多。

Q: 一般项目使用DataPipeline的服务是统一管理还是私有化部署？若是私有化部署若要升级怎么操作?

DataPipeline是私有化部署产品，升级操作可以参考最新版本的产品文档中部署相关部分。

Q: DataPipeline的Kafka如果与客户目前使用的Kafka版本不一样，是否需要适配?

DataPipeline的Kafka作为消息系列缓存在系统内部使用，与您使用的Kafka节点没有关联关系，不需要适配。

Q: 在从节点上装MySQL，对单表导入1000万数据对任务有影响吗?

在使用DataPipeline的集群上，从节点安装数据库会占用cpu、内存、磁盘、IO等资源，如果数据量比较大的情况下，最好单独找一台服务器装，以便不会对任务的性能造成影响。

Q: 数据源多个表是否可以写到目的地一张表?

允许单个或多个任务的多张表往一个数据目的地表同步数据，但要求同步的表与目的地表结构一致。多个表表结构一致时，要求每个表的主键值不同，否同时同步多张表到同一个目的表会发生「主键冲突」现象，多条主键相同数据会保留最后一条。若多个表结构不一致，如结构不一致问题属于结构变化策略处理范围，则会根据策略执行。

Q: 数据节点连接失败怎么办?

检查节点连接信息是否填写正确；检查本地网络是否可以ping通节点地址；检查节点端口是否开启；检查SSL、Kerberos等认证信息是否需要更新。

Q: DataPipeline如何应对Mysql数据库表和字段名称大小写不敏感问题?

您可在数据映射中修改MySQL目的地建表表名与字段名。

Q: 数据源Mysql的实时处理模式下，暂时无法读取哪些字段类型?

MySQL的实时处理模式下，暂时无法读取字段类型为 GEOMETRY 的数据，具体请参照下表。如果存在对应类型的数据，请选择定时模式进行同步。GEOMETRY字段：

POINT

LINESTRING

POLYGON

GEOMETRY

MULTIPOINT

MULTILINESTRING

MULTIPOLYGON

GEOMETRYCOLLECTION

Q: Mysql数据源实时处理模式下，暂不支持哪些语句操作的同步?

Mysql实时处理模式下，暂不支持下列语句：
TRUNCATE语句不会同步删除目的地数据
MySql级联删除不会删除目的地数据
Mysql级联更新不会更新目的地数据

Q: Oracle实时模式为LogMiner时，为什么还需要设置读取频率?

当数据源为Oracle时，LogMiner实时读取模式会采用定时查询方式读取增量，所以需要用户去设置一定的读取频率，用户设置读取频率越大相对来说数据延迟性也就越大，但读取频率过小时，会对系统造成一定的压力，因此，需要用户根据实际情况设置相对合理的读取频率，目前DataPipeline会默认读取频率为60秒。

Q: SQL Server数据源读取方式选择Change Tracking时需要注意什么?

您需要启动Change Tracking， 查看官方参考文档。
请在同步前开启同步表的Change Columns Updates，若您未主动开启，系统会自动开启。
请确保在选择同步表时选择存在主键的表，Change Tracking的读取方式需要基于主键捕获数据更新，若不存在主键，将无法被选择。

Q: SQL Server实时模式为Change Tracking时，为什么还需要设置读取频率?

当数据源为SQL Server时，Change Tracking实时读取模式会采用定时查询方式读取增量，所以需要用户去设置一定的读取频率，用户设置读取频率越大相对来说数据延迟性也就越大，但读取频率过小时，会对系统造成一定的压力，因此，需要用户根据实际情况设置相对合理的读取频率，目前DataPipeline会默认读取频率为60秒。

Q: 时区问题需要注意什么?

当MySQL源端有字段为timestamp，字段值为0000-00-00 00:00:00，无法同步到MySQL目的地，需要用户在清洗脚本中进行数据时区的转换。
目前DataPipeline读取源端表结构时会默认将0000-00-00 00:00:00转换成1970-01-01 00:00:00.000000000 +00:00
但是1970-01-01 00:00:00.000000000 +00:00是世界时区，需要用户根据实际的需求使用清洗脚本进行清洗，
如：转换成中国时区。

Q: 行级的物理删除，使用Change Tracking的方式，是否获取的到?DataPipeline会如何处理这类的数据?

当已同步的数据在数据源被删除（需要同步表存在主键，否则无法获取数据源数据删除信息），DataPipeline提供三种方式：
同步，删除目的地数据；
忽略，保留目的地数据；
同步，保留增量数据，并按照merge模式处理。
选择merge模式处理，需要在写入设置完成以下步骤：
各个同步表开启高级清洗。
在脚本库中选择「DML.java」脚本库。
在目的地表结构增加「dml」新字段，用于标识：insert、update、delete

Q: MySQL增量数据同步任务可以同步视图吗?

Binlog读取模式无法同步视图，JDBC读取模式可以同步视图。

Q: 各种类型的数据源同步数据到TIDB目的地时，需要注意哪些事项?

首先，系统验证TIDB不同级别的权限，因此需要验证mysql.db/mysql.user/mysql.table_privs三张表的SELECT权限。
其次，用户需要确认TIDB的同步表是否有SELECT权限，如果TIDB里表没有SELECT权限的话，同步表时，系统无法检测到重名的表，无权限的表同步到目的地会导致任务报错，因此用户在使用TIDB目的地同步表时要保证目的地的表具有SELECT权限才可以。

Q: Redshift并发数设置是50,DataPipeline对100个表并发插入的方案？对Redshift性能的影响？DataPipeline对大数据量并发插入Redshift的处理方式?

DataPipeline会根据Redshift的最优的性能来设置Redshift的并发数，没有一个固定值50个并发同时写就是最优的写入方案。
我们可以进行100张表的并发写入，单高并发对DataPipeline服务器内存、磁盘有很大的消耗，根据测试环境配置并发数量。
DataPipeline是通过Redshift 客户端写入到Redshift，大量并发的情况下会影响性能，所以在测试时需要根据源端、目的地端配置及性能对写入并发进行设置。
DataPipeline先将数据写成文件，当文件到一定size的时候load到Redshift。

Q: kafka目的地支持设置新的分区吗?

目前DataPipeline同步数据到Kafka，都是要求用户自己在Kafka目的地设置好Topic和Partition等分区逻辑。默认用的是round robin，也就是用户如果定义了10个分区，数据会均匀写入。目前暂不支持往新的Topic里面写。

Q: 多个表结构不一致的表，可以同步至kafka的同一个topic吗?

目前在DataPipeline中，如果在一个数据任务中，同步多个表结构不一致的表，到kafka目的地的一个topic中，会出现数据源变化错误，导致任务无法正常运行。

该情况有两种解决方案：

创建多个任务，每个任务包含一个将要同步的表，目的地可以是相同的kafka topic，任务会正常运行。

Q: 任务设置中读取频率的实现原理是什么样的?

DataPipeline定时频率需要指定定时时间间隔，从上一个次执行读取数据开始计时，当本次执行完成读写后才会开启下次执行。如，设置定时频率为60分钟，从1点钟开始执行任务，如果60分钟内完成了读写，则2点会开始执行下一次的读写，若没有在60分钟内完成读写，则系统会等待写入完成后才开始执行下一次，也就是系统无法按照60分钟的读取频率来实现批次读取。

Q: 采用增量同步的情况，新建同步任务时，源端的数据表有大量的存量数据，如何通过产品实现数据同步的一致性的?

如数据源增量数据获取方式为日志获取，则在任务配置中，选择开启全量初始化，并在全量初始化配置中选择清空数据目的地数据，在增量数据同步之前，会先同步全量数据，保证数据一致性。

如增量数据获取方式为增量识别字段方式，则需要您在数据映射–读取限制配置中仅设置增量识别字段限制，任务运行时，会先同步全量数据，全量数据同步完毕后会按照增量识别字段进行增量同步，保证数据一致性。

Q: 数据源端基于日志的实时模式，是源库推送还是我们做捕获?

Oracle（归档日志）、MS SQL Server（Change Tracking）是通过扫描的方式，DataPipeline主动去捕获日志信息；Mysql（binlog）、PostgreSQL（wal2json）是通过源库自带的日志机制进行主动推送。

Q: 关系型数据库，如MySQL，如果出现大量的数据修改，BinLog日志如何抓取，如何实现及时的消费?

如果大数据量会出现数据堆积的状况，消化就有一定的延迟，但是只要binlog日志不被删除，消费速度比读取速度快的话就会追上数据。

Q: 读取与写入的速率限制是按照任务还是按照表?

目前DataPipeline是按照并发进行限制的，也就是按照表做限制，每一张表的读取和写入速率不会超过用户设置的速率。

Q: 读取与写入的速率限制原理是?

速率在任务执行过程中是一个平均指标，即限制单位时间内读取或写入数据的数据量即可以理解为对平均速率的限制，故数据节点读取或写入速率可能会超过平均值，但从任务整体执行的周期内，一般情况下，不会增加数据节点负载。

Q: 我们的无侵入性是如何实现的? 是完全无侵入性，还是侵入性很小? 是否无侵入性就意味着源端服务器没有访问请求的压力，那目的端写入是否还存在压力?

DataPipeline 对源侵入性是指无需在数据源端去按照Agent进行实时数据读取，DataPipeline采用的方式都是基于数据库本身自带的能力来实现的，如：MySQL用的是binlog；Oracle用的logminer；sql server 用的 change tracking 的方式；但对于Oracle节点的读取，我们推荐使用Agent模式进行读取，可以使用相对小的源端服务器资源换取成倍的同步速率增长。

Q: 动态限速的策略是什么?

根据配置的缓存大小，当估算堆积数据 (所有数据– 已消费数据) 的数据量大于总缓存的90%(可以后台配置) 则会停止读取。根据配置的检查时间间隔，直到下一次检查时间，重新计算是否允许读取。

估算方式：单条数据大小 = [过去数据的平均字节数（已读取的数据占比/已读取的数据量）+ 140字节]*安全系数 （安全系数：也是根据占用的缓存比例计算，剩余缓存比例越小，安全系数越大)

安全系数的计算方式：占用的缓存比例在0 – 70%之间 ，固定是环境变量配置的值，默认1.5；超过70%之后，1.5 + (已占用的缓存比例 – 固定70%) * 2

Q: 如果任务激活后进行重新同步，目的地数据会清空吗?

目前用户在选择表进行重新同步时，支持用户去选择是否要清理目标表的数据，如果选择清理目标表的数据，则点击重新同步，目的地数据将被会清空后重新导入。
如果选择不清理目标表的数据，则点击重新同步，目的地数据将被不会被清空，目的地按主键去重，若无主键则会存在重复数据。

Q: 如何设置数据读取条件where语句？有哪些注意事项?

若用户设置了where语句，DataPipeline向数据源读取数据时会执行数据读取条件，在读取端直接过滤部分数据；
请根据数据源类型输入相对应的 SQL 语句作为数据读取条件。
一个数据表映射只能设置一个where语句，where语句中可以包含多个字段的条件。
当使用where语句设置增量识别字段作为增量数据获取手段时，数据任务第一次运行将会同步全量数据，全量数据同步完成后，会根据增量识别字段判断数据增量，进行同步。

Q: 暂停运行中的数据任务，修改映射，取消表后又新加入此表，DataPipeline对于此表的处理策略是什么样的?

激活任务时，将按照新增映射处理，按照当前任务执行配置，重新同步该表。

Q: 表结构中的精度和标度是什么意思?

精度与标度在不同数据节点中的表述方式不同。
MySQL节点，精度：Length ，标度：Decimals
Oracle节点，精度：Size ，标度：Scale
MS SQL Sever节点，精度：Size ，标度：Scale
PostgreSQL节点，精度：Length ，标度：Decimals

Q: 数据源端支持哪些字符集类型?

数据源读取使用jdbc driver，除非要进行手动编码转换，一般不需要特别关注数据库的字符集类型。写入方面，目前产品尚不支持定义表字符集，采用库默认字符集。如果出现字符范围不匹配，有可能出现乱码。如果源库和目标库的默认编码一致，不会出现这个问题。

Q: 选择增量识别为主键，如何保证源端和目标的数据一致性呢，如果该记录有修改，系统是怎么处理的?

增量识别字段应为自增字段或时间戳类型，这样我们可以通过增量字段来判断哪些数据进行了变更，一般update数据不会更新主键，所以不建议用主键做自增字段。

Q: 数据目的地ODS有大量无主键表，同步时DataPipeline是如何处理的?

在数据映射中，需要您指定该目的地表主键，如无法指定已有表主键，则无法进行增量数据同步。

Q: DataPipeline是否支持将不同的数据表（在不同的数据库中，但是表结构一致，同时有主键和唯一性识别的字段），导入同一个目的端表?

DataPipeline支持同步表结构相同的表到目的地，但是数据源表的主键不能重复，否则数据会被覆盖。如果有唯一性识别字段，可以把这唯一性设置为新的主键，避免出现主键冲突的问题。

Q: 哪些数据错误会进入错误队列?

您可以根据数据的错误类型名称和具体的错误信息来了解导致进入错误队列的原因，例如：

「RULE_ERROR」，当您看到这个错误，就代表着数据内容因为无法匹配清洗规则而进入了错误队列。

例如，当您设置了「Column > 0」的过滤规则，而数据源「Column」字段值为NULL，平台无法进行过滤，导致了这些所有值为NULL的数据进入错误队列。

具体的清洗规则规范您可以点击查看数据清洗。

其他类型的错误您可以具体查看下方的错误信息详情来定位具体的错误来源，如果您对这些信息存在疑问，欢迎联系我们的技术团队帮您定位问题。

Q: 错误队列里的原始数据是指源端读取的原始数据还是说经过清洗规则后的数据?

目前DataPipeline支持的错误队列里的原始数据是指清洗之前读取的原数据。当用户设置了清洗脚本时，但进入了错误队列的数据是指没有清洗过的原始数据。

Q: 部分表已读取已写入等都为0，但完成进度为100%?

这是由于这部分表为空表，已读取和已写入等数据反应了数据表真实情况。我们默认把空表视作已完成全量，若您的任务选择了同步增量数据，则后续有新增数据时，这些数据将会实时更新。

Q: 任务详情页中的数据读写量具体含义是？为什么有时候还会减少?

已读取数据量是指数据任务被激活后系统从数据源读取的数据量：
当数据任务由于断点续传机制重新读取已读取的数据时不会重新记录到已读取数据量中，只有该数据被更新时会算作新的数据来记录到已读取数据量中。
已写入数据量是指系统读取数据后已经被处理的数据量：
所以已写入数据量不仅仅包括已同步到目的地的数据，还包括错误队列的数据和用户要求过滤的数据。已写入数据量也遵循当数据任务断点续传发生的重复写入数据不会记录到已写入数据量中。
读写数据量在某一段时间内会突然减少的原因是：
DataPipeline为了在传输过程中不丢失任何数据，若需要同步的表存在主键时可以支持断点续传。当数据任务由于突发情况导致重启时，数据任务会从上一个记录的读取数据点开始重新同步数据，此时已读写的数据统计值会回到上一个记录的读写点上，由此用户会看到统计值会减少的现象。
注意：当数据任务已读写统计值频发出现减少是DataPipeline同时进行读写的数据任务过多的信号，需要增加节点的方式解决。您如果经常遇到该情况，请联系DataPipeline专业工程师帮助您解决问题。

Q: 激活任务后，已读取数据百分比为什么会发生回调，如：从99%跳到30%?

Oracle的表行数meta信息是不准的。出于性能考虑，我们在估算表大小的时候，不能用count去精确计算，就会出现，从meta信息里面拿表大小是1千万，但实际表大小是1千1百万。当实际读取超过meta信息里的值时候，就会用实际的读取数进行修正。

Q: 为什么已暂停的数据任务还会写入数据?

在DataPipeline的数据模型中，我们将数据从数据源读取出来，写入消息队列进行缓存，数据写入端来消费缓存中的数据，写入至数据目的地。当数据任务的读取速率大于写入速率时，在缓存中会产生部分已读取但尚未写入的数据，这时，暂停数据任务，数据任务需要将已读取但未写入的数据继续写入目的地，以防数据同步进度丢失。

Q: 目前进行数据任务的时候，读取速率远大于写入速率，其中，已读取且还未写入的数据会暂时存储在Kafka上，但是由于Kafka存储空间有限，超出后容易造成数据的丢失，这怎么办?

> 我们支持用户开启动态限速功能来定时检查kafka的最大存储空间，通过设置一个时间间隔来确认已读取且还未写入的数据量是否大于设定的Kafka的最大存储空间，从而判断是否需要暂停数据的读取。直到下一次确认，已读取且还未写入的数据量小于设定的Kafka的最大存储空间，则启动数据任务的剩下数据读取。

Q: 如果一条数据多次、频繁变化，在DataPipeline产品侧如何保证数据的并行和保序是如何保证的?

> 我们源端会将任务按照一定原则拆分为多个互不干扰的子任务进行并行执行。在JDBC源读取场景下，如果任务包括多张表，每个表是由一个独立线程进行顺序读取的，线程并行度可以在任务属性中进行设置。为了保证顺序写入和读取，默认每个单独子任务会创建一个独立的topic，设置一个分区，这样目的端消费的时候，同一个topic只有一个consumer在进行消费，从而保证消费的顺序性。如果可以接受非顺序消费，也可以为一个topic创建多个分区，这样目的端可以更好地利用Kafka的并行能力提高吞吐量。

Q: 什么样的实时传输任务会在性能关注中显示?

> 当该实时传输任务前一天至少有有一张表有数据流入的时候，才会显示在实时传输任务的性能关注里。

Q: 产品使用期限到期所有任务都会被暂停任务，那么如何提前获知产品使用期限是否到期以避免任务被暂停?

> Datapipeline提供提前通知服务，当产品距离到期10天/7天/3天时，datapipeline会向用户发送三次邮件，通知用户及时申请新的激活码，工作人员会及时为用户提供新的激活码。

5. 词汇表
GLOSSARY

词汇/图标	系统内释义
数据节点	数据节点是数据任务进行数据集成的原始数据载体。「数据节点」可以是数据库、文件系统、数据仓库、文件、应用，一切存储数据的载体都能成为「数据节点」。
数据节点状态	数据节点状态是系统赋予数据节点的管理状态，其中包括激活、挂起。
激活	激活是指将数据节点的管理状态设置为激活，激活后数据节点在系统内为可用状态。
挂起	挂起是指将数据节点的管理状态设置为挂起，挂起后数据节点在系统内将不可用。
连接参数	连接参数是指在系统连接数据节点时，可选的连接参数，不同类型的数据节点的连接参数不同，可能带有不同的配置含义。
连接验证	连接验证是系统连接数据节点时，提供的连接测试功能，系统将会校验是否可以成功连接与是否开通节点读取方式必要的相关权限。
数据节点基础配置	数据节点基础配置是在系统中连接使用该节点的最小化配置。
数据节点策略配置	数据节点策略配置是将数据节点的使用、配置更加易用的配置选项，统一管理节点的策略配置也有助于提升系统稳定性。
语义映射策略	语义映射策略是通过界面配置的方式，将数据源节点的数据类型、索引、特性等语义与数据目的地的数据类型、索引、特性等语义关联起来的映射配置。
检查点策略	检查点策略是系统记录作为数据源的数据节点的日志检查点位置，方便在任务执行配置中选择同步起点。
数据链路	数据链路是将数据任务配置集中管理，统一配置的功能模块。
数据链路基本配置	数据链路基本配置是任务保证数据任务成功运行的基本配置，其中包括数据源配置、数据目的地配置与数据链路配置。
数据源配置	数据源配置是数据任务对数据源读取方式的配置。

词汇/图标	系统内释义
数据目的地配置	数据目的地配置是数据任务对数据目的地读取方式的配置。
数据映射	数据映射是将数据源的数据表和字段与数据目的地的表和字段建立映射关系的功能。
表映射关系	表映射关系是数据源的数据表与数据目的地表的映射关系。
字段映射关系	字段映射关系是数据源表中字段与数据目的地表中字段的映射关系。
选择同步列表	选择需要同步的数据源中的数据。
语义映射规则	当前映射配置使用数据源的具体语义映射规则版本，仅可单选。
	数据映射--查看数据源一对多关系。
	数据映射--编辑读取限制条件。
	数据映射--查看数据目的地多对一关系。
	数据映射--编辑清洗脚本。
	数据映射--查看该表的字段映射。
	数据映射--删除该行映射规则。
创建新表	在数据目的地为数据源表选择新建表的映射方式。
创建目的地表	为当前链路中已选在数据目的地为数据源表选择新建表的映射创建目的地表，您可以选择系统创建或导出建表语句自行创建。
未创建	在数据目的地为数据源表选择新建表的映射方式，但未在数据目的地创建表。
选择已有表	在数据目的地为数据源表选择已有目的地表的映射方式。
刷新	刷新目的地数据表信息、表结构。
清洗脚本–样例数据	从数据源中获取后，经系统处理，待写入数据目的地的数据。
清洗脚本–试运行	使用已编辑的清洗脚本测试处理样例数据。
清洗脚本–运行结果	使用已编辑的清洗脚本测试处理样例数据，获取到的运行结果。
清洗脚本–脚本库	包含系统预置与用户保存的脚本的脚本库。

词汇/图标	系统内释义
写入主键冲突策略	主键冲突策略是在任务写入过程中，处理写入数据与目的地数据有主键冲突的策略。
结构变化策略	结构变化策略是当数据源数据结构发生变化时，系统将为您执行的策略，能够有效避免由于数据源结构变化使任务暂停带来的影响。
增量处理策略	当数据源产生已同步的数据被删除这样的增量数据时，您可以通过配置增量处理策略来对这部分数据进行处理，保证数据一致性。
端到端一致性策略	端到端一致性策略是在任务运行过程中保证数据从数据源端到数据目的地端一致性的策略。
自动重启策略	自动重启策略时在任务运行过程中，任务出现报错自动重启的执行策略。
错误堆栈匹配	系统在判断报错任务是否需要自动重启时，会根据已定义的错误堆栈片段与任务报错的错误堆栈做匹配。
错误队列策略	错误队列策略是任务运行过程中，出现错误数据，系统帮助您处理处理错误数据的执行策略。
错误堆栈	错误堆栈信息是任务运行过程中报错返回的错误信息，因DataPipeline所使用的编程语言是Java，故错误堆栈信息均为Java错误堆栈。
内部节点	内部节点是指使用系统内部数据节点存储错误数据、日志数据。
外部节点	外部节点是指使用外部数据节点存储错误数据、日志数据，外部数据节点需在系统节点管理中被管理。
预警策略	预警策略是监控任务运行状态、任务错误数据、数据源变化情况并及时通知用户的执行策略。
日志策略	日志策略是将任务配置变更，任务报错信息以日志形式记录，方便用户查询的策略。
数据任务	数据任务是DataPipeline进行数据同步的最小管理单位。
任务概览	任务概览是通过拓扑图的形式查看全部数据任务及数据任务之间的关系的功能。
任务监控	数据任务监控是将重要任务、故障任务、未激活状态任务、性能关注任务集中监控管理的功能模块。
项目	项目是系统提供的对任务进行逻辑分组管理的功能。
重要任务	重要任务是用户赋予数据任务的重要程度属性。

词汇/图标	系统内释义
待完善	数据任务状态为尚未配置完成。
未激活	数据任务状态为配置完成，尚未激活。
故障任务	报错暂停的数据任务。
未激活状态	创建完成后，没有被激活的数据任务。
性能关注	延迟时间较长的增量数据任务与同步速率较慢的全量数据任务。
延迟时间	最近写入数据的时间减去该数据产生时间。
未分组	尚未进行项目分组的数据任务的集合。
参与人	任务/节点/链路权限控制功能，数据任务/节点/链路仅对参与人可见，参与人可修改可见任务/节点/链路。
已读取数据量	• 指 DataPipeline 从数据源已读取的数据量。 • 当 DataPipeline 系统重启，会重新读取数据，若源端没有主键则重复读取的数据量会记录到已读取数据量里。
已写入数据量	• 指 DataPipeline 已处理的数据量，这里包括：同步到数据目的的数据量和进入到错误队列的数据量。 • 当 DataPipeline 系统重启，会根据断点续传机制从上一个写入记录点开始重新写入部分数据，但这部分数据会记录到已写入数据量里。
错误队列数据量	指已读取的数据中系统判断无法写入到数据目的地，而异步放到错误队列中的数据量。
读取速率	指任务当前对数据源的读取速率。
处理速率	指任务当前对数据目的地的写入速率，多个数据目的地，展示其速率之和。
消息列表	系统展示数据任务状态变更及部分配置变化的通知区域。
关联任务数量	该数据链路关联数据任务的数量
任务映射–传输队列设置	针对此条数据映射的传输队列设置。
任务映射–batch设置	针对此条数据映射的写入Batch拆分设置。

词汇/图标	系统内释义
同步方式	同步方式是指定数据任务同步的方式，包括增量同步与全量同步。 • 增量同步：读取数据库日志以获取数据增量或通过增量识别字段获取增量，将增量数据同步至数据目的地。 • 全量同步：将全量数据同步至数据目的地。
全量初始化	即进行增量同步之前，是否进行一次全量同步，以保证数据目的地数据与数据源一致。
同步起点	进行增量同步之前不进行全量同步，故在日志增量获取模式下，需要指定日志读取起点，以明确数据从那个snapshot开始同步。
Binlog Position	MySQL数据库Binlog日志获取的记录位置。
Binlog 文件名称	MySQL数据库Binlog日志文件名称。
GTID	MySQL数据库Binlog日志GTID。
Change Tracking Version	MS SQL Sever数据库的Change Tracking功能提供的日志记录点。
PostgreSQL LSN	PostgreSQL数据库的日志记录点。
LogMiner SCN	Oracle数据库通过LogMinner方式获取日志的记录点。
任务执行时间	任务开始执行的时间。
任务执行方式	数据任务的执行方式，与同步方式强相关。 • 监听模式：通常用于日志模式增量获取，监听日志是否有增量。 • 定时模式：通常用于全量同步，设置任务定时开关。
同步前清空目的地	同步前清空数据目的地中的数据，有助于保持数据一致性，无主键数据的唯一性。
目的地更新方式	目的地更新方式指将数据目的地清空后再进行数据写入，分为下列两种方式： • 清除数据（truncate & insert）：执行清除数据目的地表数据语句，清除后执行插入语句进行数据同步。 • 删除重建（drop & create）：执行删除目的地表语句，删除后新建数据目的地表进行数据同步。
数据源读取资源设置	数据源读取进程运行的资源组的分配设置。
数据目的地写入资源设置	数据目的地写入进程运行的资源组的分配设置。

词汇/图标	系统内释义
读取速率限制	数据任务对数据源的读取的速率限制。
动态限速	为了避免读取数据过多过快，同时写入速率过慢导致的数据传输缓存过大触发缓存清除策略导致的数据丢失的动态限速规则。
读取并发限制	数据任务对数据源读取的并发限制。
写入速率限制	数据任务对数据目的地写入的速率限制。
写入并发限制	数据任务对数据目的地写入的速率限制。
任务限制配置–Batch设置	数据任务对数据目的地写入的Batch切分设置。
任务限制配置–传输队列限制	数据任务运行过程中使用的传输队列缓存的限制。
冻结用户	用户被系统管理员冻结后，将不再能登录。
用户分组	系统进行用户分组权限控制的功能。
预警发送组	预警发送渠道的分组功能。
Webhook	发送预警信息的Webhook API配置功能。
预警邮件	发送预警邮件的邮件配置功能。
邮件服务器	不能连接外网情况下，系统发送邮件的邮件服务器配置功能。

1.0

DATAPIPELINE

北京市海淀区清华同方
科技大厦D座东楼1801

官网
www.datapipeline.com

电话
400-606-5709

邮箱
service@datapipeline.com