

EDAP 文档



【版权声明】

版权所有©百度在线网络技术（北京）有限公司、北京百度网讯科技有限公司。 未经本公司书面许可，任何单位和个人不得擅自摘抄、复制、传播本文档内容，否则本公司有权依法追究法律责任。

【商标声明】



和其他百度系商标，均为百度在线网络技术（北京）有限公司、北京百度网讯科技有限公司的商标。 本文档涉及的第三方商标，依法由相关权利人所有。未经商标权利人书面许可，不得擅自对其商标进行使用、复制、修改、传播等行为。

【免责声明】

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导。 如您购买本文档介绍的产品、服务，您的权利与义务将依据百度智能云产品服务合同条款予以具体约定。本文档内容不作任何明示或暗示的保证。

目录	
目录	2
功能发布记录	5
功能发布记录（2024年）	5
2024年12月	5
功能发布记录（2023年及之前）	8
产品简介	8
升级公告	8
产品介绍	9
产品优势	9
产品架构	9
应用场景	10
核心概念	10
功能特性	11
DataBuilder 与 EasyDAP 产品关系说明	12
产品计费	13
计费概述	13
版本购买说明	13
资源购买说明	15
入门指南	16
示例场景说明	16
准备工作	17
规范设计	19
数据开发	20
数据质量	21
数据同步	22
数据服务	23
操作指南	24
准备工作	24
多用户访问控制	24
数据湖配置	33
文件管理	36
源连接管理	38
资源管理	41
行业最佳实践	42
项目管理	45
项目参数	45
概览	45
数据湖	46
湖管理	46
元数据	47

数据加工	55
数据集成	55
数据开发	69
数据分析	90
数据运维	92
数据架构	104
主题设计	104
数据建模	105
数据指标	114
数据标准	117
数据治理	121
数据地图	121
数据血缘	125
数据质量	129
数据安全	142
数据服务	147
概述	147
API创建	149
API应用	154
API调用管控	155
管理中心	155
资源管理	155
数据源管理	158
权限管理	161
审批中心	164
私有化部署	167
私有化部署说明	167
API参考	167
API概述	167
API列表	170
数据开发API	171
作业开发	171
任务运维	182
元数据API	189
统计信息	189
分区	195
函数	204
表	210
库	220
视频专区	224

Baidu 百度智能云文档

功能发布记录

产品介绍	224
功能模块介绍及最佳实践	225
常见问题	226
服务等级协议SLA	226
EDAP服务等级协议SLA	226

功能发布记录

功能发布记录（2024年）

2024年12月

序号	功能名称	功能描述	使用文档
1	作业组支持千帆ModelBuilder推理服务	【作业组新增千帆ModelBuilder推理服务节点】 作业组新增千帆ModelBuilder推理服务节点，创建AI workflow工作流，对数据利用python作业清洗、处理后，调用批量推理节点，实现一站式、自动化大模型推理服务数据流。	ModelBuilder批量预测节点
2	数据地图	【新增数据地图功能】 数据搜索：支持元数据多维度搜索（如表名、字段名、描述、标签等），提高平台元数据获取效率。 元数据采集：支持HIVE、DORIS、MYSQL、ORACLE等多种数据源的元数据采集，满足多源元数据采集及分析需求。	数据地图
		【数据湖能力提升】 新增湖概览：展示数据湖存储统计，用户可直观了解数据存储与使用情况，提升数据湖管理效率。 新增元数据发现：支持ORC、Parquet、Hudi、DeltaLake等格式文件的元数据发现，提供灵活的任务配置，帮助用户高效管理湖上数	

3	数据湖	据。 湖概览 元数据发现 扩展数据湖格式：元数据、数据集成、数据开发支持Iceberg、Delta Lake湖格式的创建和查询，完善数据湖体系。 数据湖分析，增加直连分析功能：提供Doris、Hive数据源的直连分析功能，丰富数据分析方式，提高数据分析效率。	
---	-----	--	--

📅 2024年9月

序号	功能名称	功能描述	使用文档
1	数据服务	数据服务支持负载均衡 BLB，提高了安全性、可用性。	API创建

📅 2024年8月

序号	功能名称	功能描述	使用文档
1	数据集成	实时同步能力提升，满足实时任务的灵活运维需求 - 增加全量重跑、指定点位重跑功能 - 运行过的任务可以新增、删除单表 - 目标库表增加逻辑删除能力。	实时同步

📅 2024年6月

序号	功能名称	功能描述	使用文档
1	产品功能逻辑重构	打造数据加工、架构、治理、服务的湖仓一体化产品，全面提升产品体验	-
2	首页能力升级	提供常用项目、典型场景、最佳实践、产品动态、常见问题等快捷入口，打造个人产品使用工作台	-
3	数据开发	数据湖分析体验升级：将该功能迁入项目内，您的历史分析，会统一迁移至“数据湖分析历史记录”项目中，提升数据分析效率	数据湖分析
4	权限体系升级	面向管理员、架构师、开发者等角色提供只读、读写、无权限的全功能自定义权限配置，提升权限管理易用性。	多用户控制访问
5	数据集成	源端新增TDengine类型：增加TDengine类数据源，覆盖工业时序数据的入湖需求。	数据集成 离线同步
6	数据开发	- 任务运维：作业组下线并删除，拆分为下线、删除两个操作，提升任务运维效率 - 作业开发：工作台日志支持分级检索，提升作业开发效率	作业组 任务运维

🕒 2024年3月

序号	功能名称	功能描述	使用文档
1	数据架构	维度建模支持Doris物化、逆向：维度建模支持面向Doris数据源的物化、逆向，提升面向应用端的查询效率。	数据建模 模型物化
2	数据开发	- 作业开发-默认关联最新版本：作业组间作业默认关联最新版本作业，减少作业由于发布导致的运维问题，提升运维效率 - 作业开发-新增全局参数配置：增加项目级全局参数配置，保证一处配置，多处引用，减少每个作业重复配置相同参数工作量	项目参数 全局参数配置

🕒 2024年1月

序号	功能名称	功能描述	使用文档
1	数据治理	● 新增元数据版本管理：针对数据源类型为EDAPDataLake的数据源，提供对元数据变更的追溯功能，提升元数据管理能力 ● 数据质量数据源类型拓展：支持针对ClickHouse、PG、GP、HANA、TiDB、DM的数据质量检测，提升数据质量源端类型覆盖度 ● 自定义规则参数限制由2张表、5个字段增加至5张表、50个字段，提升质量作业易用性 ● 数据质量数据源类型拓展：支持针对ClickHouse、PG、GP、HANA、TiDB、DM的数据质量检测，提升数据质量源端类型覆盖度	数据质量
2	数据架构	● 作支持ODS层建模：ODS层建模支持面向EDAPDataLake、Hive数据源的逆向、物化功能，解决数据建模在ODS层的缺失问题 ● 维度表支持查看关联关系，在维度表由于引用无法废弃时，便于梳理取消引用关系	数据建模
3	数据集成	● 离线同步支持前置检查，增加草稿状态，未执行前支持配置项修改，提升集成作业创建效率数据集成 ● 实时同步能力提升：支持源端PostgreSQL、MySql、SQL Server、Oracle实时同步至Doris、EDAPDataLake，实时同步能力全面提升	数据集成·离线同步 数据集成·实时同步
4	数据开发	修数据与作业调度关系拆分、作业实例增加依赖链路图展示，提高作业运维效率。	任务运维·离线作业

功能发布记录（2023年及之前）

产品简介

升级公告

尊敬的EDAP用户：

为了给您提供更加优质的产品体验，百度数据湖管理与分析平台EasyDAP计划于**2024年6月28日**进行重大升级，升级后产品将核心能力聚合为加工、架构、治理、服务、管理5大模块，提供更加统一的湖仓一体产品方案。

🔗 发布信息：

发布范围：北京、广州、保定

发布时间：**2024-06-28 13:50 ~ 14:30**

🔗 核心特性：

- 整体功能分布重新调整，划分为：加工、架构、治理、服务、管理5大模块，具体子项功能操作保持不变；
- 产品首页全新升级，提供常用项目、典型场景、最佳实践、产品动态、常见问题等快捷入口；
- 数据湖分析交互体验升级，并将该功能迁入项目内，您的历史分析，会统一迁移至“数据湖分析历史记录”项目中。

🔗 您需注意：

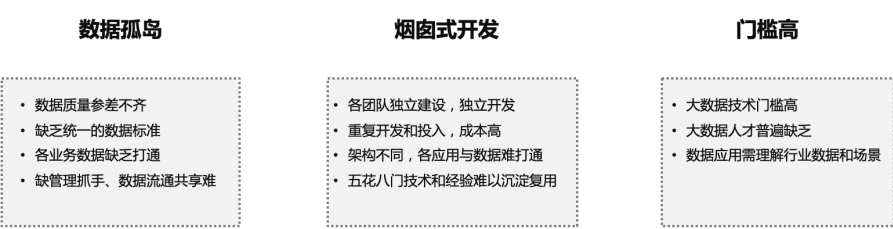
- 您运行中的数据开发任务，不受影响；
- 您正在开发中的作业，如果未做保存，可能会由于升级期间页面刷新导致数据丢失；
- 权限体系，随本次升级进行了刷新，如您有自定义角，使用期间若遇到报错问题，请提交工单或如流群联系技术人员解决。

感谢您的支持！

产品介绍

百度数据湖管理与分析平台EasyDAP（Easy - Data Lake Management and Analysis Platform，简称EDAP）是一款面向企业的全场景、低门槛、开放灵活的大数据管理与分析平台，可一站式完成数据集成、数据治理、数据开发、数据分析、数据服务，并与云计算&存储生态打通，采、存、管、用一体化，帮助企业实现数据价值释放。

企业大数据应用，普遍有数据治理管理难、开发效率低、分析应用技术门槛高的困难：



百度数据湖管理与分析平台EasyDAP，围绕以数据为核心，封装数据从汇聚、开发、管理到服务的一整套完整的产品设计体系，屏蔽了大数据相关技术并提供简易的可视化开发操作界面，降低使用门槛；无缝对接主流大数据存储计算平台，兼容企业原有IT系统、技术栈，并与百度云产品生态打通，降低数据中台建设成本。此外，安全的资源隔离、权限控制、审计日志等，提高数据安全性。

围绕企业大数据开发应用全生命周期，EDAP提供了包括数据集成、数据湖构建、数据开发Studio、数据分析等端到端完备功能体系：



产品优势

全场景

- 围绕数据全生命周期，封装数据从汇聚、开发、管理到服务的一整套体系。
- 数据集成、数据治理、数据开发、数据分析、数据服务应用一站完成。

低门槛

- 封装复杂的大数据相关技术和组件，提供简易的操作界面。
- 0代码数据集成、批&流可视化开发Studio、拖拽式作业编排、一键多源联邦分析。

开放兼容

- 无缝对接主流大数据存储计算平台，适配兼容性高。
- 兼容企业原有IT系统、技术栈，降低数据中台建设成本。
- 元数据、作业管理开放Api,方便灵活定制。

安全可靠

- 安全的资源隔离、行列级数据权限管控。
- 审计日志等。

产品架构

产品架构将介绍EasyDAP的产品架构，帮助用户理解EasyDAP重要功能和能力。主架构贯穿数据集成、数据管理、数据开发、任务调度、数据治理、数据分析、数据服务等数据全生命周期，底层支持对接大数据存储计算平台和企业原有IT系统，上层对接DAMP、SugarBI等数据应用，开放OpenApi支持数据应用灵活定制开发。



应用场景

数据集成、交换场景

通过EasyDAP实现异构数据源的数据集成、交换。

数据治理场景

通过EasyDAP从元数据角度对企业数据进行统一数据目录管理、同时保证数据质量、数据安全、提供数据萃取服务。

数仓建设场景

通过EasyDAP进行数据集成、ETL 入仓PALO，高效构建企业级数据仓库。

- 推荐搭配使用：EasyDAP +BMR+ PALO

数据中台建设场景

通过EasyDAP快速实现数据资产构建，数据治理管理， 高效进行数据开发、分析、可视化应用等，助力业务增长和成本优化。

- 推荐搭配使用：EasyDAP+BMR+PALO+SUGAR

离线分析场景

基于EasyDAP批量作业可视化作业开发Studio ,低门槛实现离线数据分析，无缝对接多上下游数据源类型。

- 推荐搭配使用：EasyDAP+BMR

实时流分析场景

基于EasyDAP实时流可视化作业开发Studio ,低门槛实现实时流数据分析，无缝对接多上下游数据源类型。

- 推荐搭配使用：EasyDAP+BMR

数据融合分析场景

通过EasyDAP统一元数据管理和交互式分析能力，实现多源异构数据的融合和联合分析洞察。

- 推荐搭配使用：EasyDAP+BMR

核心概念

🔗 数据仓库 VS 数据湖

随着大数据、人工智能、云计算、物联网等数字化技术的普及和广泛应用，移动互联网和物联网时代，产生了大量的网站数据，社交媒体数据，物联网设备数据等非结构化数据。数据仓库无法满足这些多元化的数据结构的存储和查询，以及非结构化和结构化数据的交叉分析。数据湖，可以容纳大量的原始数据的存储库和处理系统。

特性	数据仓库	数据湖
存储数据类型	历史的、结构化的、预先定义数据模型、一般来源 事务系统、运营数据、业务应用程序	所有类型数据，结构化、非结构化、半结构化数据等，数据的类型遵循数据源系统的原始数据格式，IOT设备、网站、移动程序、社交媒体和企业应用程序关系和非关系数据
数据处理方式	高度结构化的架构，数据清洗转换之后加载到数据仓库，即写时模式（Schema - On -Write），用户从数据仓库获取的是处理后的数据。	数据直接加载到数据湖中，然后根据分析的需求处理数据。即 读时模式（Schema - On Read），提供强大的计算能力处理和分析所有类型的数据，分析后的数据存储起来供用户使用。
成本	更快的查询结构，存储成本高	存储海量数据，采用分布式存储系统、对象存储系统 降低存储成本
数据分析应用	多维分析和可视化报表	跨源融合分析、批&流出数据处理、机器学习、预测分析、数据发现和分析

🔗 数据管理

元数据是对数据的描述，采集企业环境中的各类元数据并统一存储，通过分析元数据，根据业务维度、系统维度等不同维度对数据分类，并梳理出数据和数据之间的关系，将从多种视角360°展示出企业的数据资产视图，让企业不同的员工都能方便地看到自己关心的数据。EDAP的元数据管理通过主题、库、表 三级以实现数据的高效组织和管理。

- 数据湖：数据湖是按原始数据格式存储，任何数据可以以最原始的形态储存，可是结构化或者非结构化数据，以确保数据在使用时可以不丢失任何细节，所有的实时数据和批量数据，都汇总到数据湖当中，然后从湖中获取相关数据用于数据分析或者机器学习。
- (数据湖)存储路径：进行数据湖存储的存储位置。基于常见数据资产类型，EasyDAP支持的数据湖存储类型为：HDFS分布式文件系统、BOS对象存储。
- 源链接：企业数据源系统的链接信息。基于企业场景的数据源系统，EasyDAP目前已支持管理 MySQL、ORACLE、SQLSERVER等DB类，PALO、GP等MPP类，HBASE 等NOSQL类型的数据源系统。
- 元数据，EasyDAP基于数据湖存储建物理表、基于源连接各类型数据源建映射表，对企业全域数据进行统一元数据管理。并基于元数据表，支持用户高效进行数据集成、批&流数据开发、多源交互式查询分析等数据处理和分析。
- 主题：较高层次上将企业的数据进行综合、归类。主题下可以有0个、1个或者多个库。
- 库：表的组织，库下可以有0个、1个或者多个表。库命名全局唯一。
- 表：元数据表。EasyDAP 包括 物理表、映射表。
- 物理表：物理表指数据存储于接入EasyDAP的数据湖存储系统中(主要是BOS、HDFS，针对HIVE数仓，可以接入HIVE的HDFS存储路径作为数据湖存储)，存储的路径由EasyDAP进行管理。EasyDAP管理物理表的读、写操作，当表被删除时，数据也会被一同删除。
- 映射表：针对企业的各类数据源系统，如MySQL，ORACLE，SQLSERVR，HBASE、PALO、GP等，EasyDAP针对各类数据源系统通过建外表的形式建立元数据表，在EasyDAP实现元数据的统一管理，并支持数据的读取，支持多源联合查询分析。

🔗 数据开发

- 作业：一个数据流处理pipeline。
- 作业实例：一个作业的一次执行生成一个作业实例。
- 作业组：一个或多个pipeline作业以及用户自定义程序包编排形成作业组。
- 作业组实例：一个作业组的一次执行生成一个作业组实例。
- 作业编排：作业编排将作业形成一个有向无环图，按照图的顺序依次调用。

功能特性

0代码数据集成

多源异构数据的分布式集成工具

- 离线数据同步，数据库、MPP数仓、Nosql、对象存储、HDFS分布式文件系统等异构数据源的定时、周期性同步。
- 实时流采集，支持http ,kafka 实时流数据同步到 DB、MPP数仓、Nosql、对象存储、HDFS分布式文件系统等数据源。
- 数据库准实时，基于CDC 实时捕获同步发生变化的数据。
- Http、Ftp等数据类型的采集同步。
- 数据处理组件，映射、过滤、格式化等数据处理组件，并且支持自定义代码处理。
- 数据源配置组件，源端&目的端、取数据模式可视化配置。
- 可视化拖拽式，拖拽式建立数据集成流水线。
- 传输管理，同步并发数控制等。

数据管理

- 数据湖存储管理：支持接入并管理BOS、HDFS等数据湖存储location。
- 数据源链接管理：支持管理DB、MPP数仓、Nosql等数据源链接。
- 元数据管理：基于物理表、映射表管理各类型数据源系统数据进行统一元数据管理，数据高效查找。
- 数据标签主题：自定义数据主题。
- 数据权限：库、表、列细粒度数据权限管控。

可视化开发Studio

- 离线开发，集成MR、Spark离线数据处理引擎，提供Sparksql、Hsql、python、scala多语言支持。
- 实时流开发，集成Spark-streaming等流计算引擎，支持简易的在线SQL、自定义Jar程序等开发模式，高吞吐低延迟的实时计算，全方位多角度的监控告警机制等
- 可视化开发，拖拽式插件式开发流水线作业，作业组编排同步支持可视化拖拽。

作业调度

- 可视化作业编排，可视化拖拽将作业形成DAG，按照图的顺序依次执行。
- 多任务类型，支持shell、python、jar包、HSQL等多种类型。
- 自定义任务类型，支持自定义程序包。
- 周期性调度，月、周、天、小时、分 级别周期性调度。
- 单次执行，立即执行、停止。
- 跳过、失败重试等执行策略设置。
- 支持重跑、添加依赖等高级功能。

交互式分析

- 支持PB级数据高并发实时查询分析，完全兼容PostgreSQL协议。
- 与大数据生态（如：BMR、PALO、BOS）无缝打通，支持海量实时和离线数据的实时分析。
- 多源联邦分析，直接分析RDS、PALO、BMR-HBASE、BOS、BMR-HIVE、BMR-HDFS 等数据仓库、大数据存储的数据。

开放Api

为了灵活的满足企业级用户个性化需求，百度数据湖管理与分析平台EasyDAP开放Api，支持用户在数据应用上层灵活定制开发，支持用户对作业进行高效运维管理。

- 元数据开放Api，基于技术元数据的增、删、改、查。
- 项目作业开放Api，项目作业、作业组执行状态查询等。

云生态打通

- 与BMR无缝对接，用户购买弹性、成本低的BMR计算资源， EasyDAP作业提交到BMR计算资源上降低客户成本。此外，支持BMR数据的离线、实时分析。
- 与BOS存储无缝对接，用户购买存储成本低廉的BOS存储作为EasyDAP的数据湖存储，降低客户成本。此外，支持BOS数据的离线、实时分析。
- 与SUGAR数据可视化无缝对接，将EasyDAP管理、开发分析的数据高效实现数据可视化。
- 与RDS、DRDS、FusionDB、PALO 等数据源系统打通，实现异构数据源的数据集成、交换、分析。

DataBuilder 与 EasyDAP 产品关系说明

百度智能云千帆数据智能平台 **DataBuilder**，面向模型训练、AI应用及数据中台升级等场景，支持用户进行多模态数据加工、治理和应用，致力于帮助用户全面提升数据价值与迭代效率。

百度智能云数据湖管理和分析平台 EasyDAP，专注于结构化数据的采集、建设、治理和应用，覆盖数据全生命周期管理，支持企业构建数据中台。

DataBuilder 与 EasyDAP 关系

随着企业对统一管理结构化与非结构化数据的需求日益增强，DataBuilder 与 EasyDAP 产品已启动融合升级。融合后 DataBuilder 将提供：

- 全面兼容 EasyDAP 的结构化数据加工与处理能力；
- 实现统一的权限与角色管理机制；

- 基于数据湖架构统一管理结构化与非结构化数据；
- 提供更强大的多模态数据开发、分析与应用能力。

后续 EasyDAP 用户可无缝升级至 DataBuilder，体验更智能、流畅的**端到端数据智能解决方案**，全面提升数据研发效率和业务支撑能力。

🔗 DataBuilder 核心能力

- 高效数据迭代 提供在线开发平台，提供灵活任务编排和优质算力能力，内置 AI 智能开发工具，全面提升数据迭代效率。
- 多模态数据管理

统一管理多模态数据，结合大模型进行数据质量分析，提供数据加密等企业级数据安全能力。

- 一站式高性能计算

多引擎协同计算，支持 CPU/GPU 异构算力混合调度，保障大规模并发任务高效稳定运行。

- 智能数据分析

提供智能搜索能力，支持搜索文本、图片等多模态数据，支持数据分析洞察，大模型助力提供数据见解。

千帆数据智能平台 DataBuilder 产品已开放试用，欢迎用户试用体验！[点此试用](#)

产品计费

计费概述

数据湖管理与分析平台 EasyDAP 计费体系由**功能费用**与**资源费用**两部分组成，**功能费用**的计费模式为**包年包月**，**资源费用**的计费模式为**按量付费**



如上图所示EDAP的计费组成为功能费用和资源费用两部分，关联产品单独计费，其中：

- **EDAP产品计费**
 - 免费部分：您可开通**体验版**，免费体验开发、集成、数据湖管理及分析等功能。
 - 付费部分
 - 功能费用：您可以在**体验版**的基础上，开通功能更全面的**标准版**、**专业版**、**企业版**的高级版本EDAP，这类版本的计费类型为**包年包月**，您在使用前需根据业务需求规划好选择包年还是包月模式，付费后即可使用。
 - 资源费用：如您需要运行数据开发、分析、服务的任务，可在产品内资源管理下，增加数据开发、数据分析、数据服务资源组，这类资源组计费类型为**按量付费**，资源构建后即可使用。
- **关联产品计费**
 - 关联计算、存储产品费用不包含在EDAP费用中，例如BMR、Palo、BOS等产品，请参考对应产品的计费逻辑，在其产品官网购买后，即关联至 EDAP 使用。

版本购买说明

EDAP产品包括体验版、标准版、专业版、企业版，其中体验版可免费体验试用，标准版、专业版、企业版采用包年包月的计费模式，版本越高提供的功能越丰富，您可以参考各版本适用场景，结合团队的业务需求选择合适版本使用。

🔗 注意事项

- 每个区域同一时间段仅可购买一个产品版本，不同区域间不通用，购买后可进行版本升级。
- 版本到期后会提供7天宽限期，7天后仍未续费，则自动降为体验版，高级版本专有功能将立即停止使用。

🔗 EDAP各版本适用场景及计费说明

产品版本	适用场景	计费模式	计费说明
体验版	• 适用于数据平台部分场景验证 • 面向个人开发者	包年包月	版本计费说明
标准版	• 提供大数据集成、开发、服务能力，处于数据湖项目建设初期 • 面向小型团队		
专业版	• 较标准版，增加治理能力，确保数据安全，增强数据价值 • 面向中小企业		
企业版	• 提供完整的数仓体系建设能力，覆盖集成、开发、建模、治理、服务 • 面向中大型企业		

🔗 版本功能对比明细

一级菜单	二级菜单	体验版	标准版	专业版	企业版
数据源	数据湖管理	✓	✓	✓	✓
	数据源工具	✓	✓	✓	✓
数据湖	数据湖存储	✓	✓	✓	✓
	元数据管理	✓	✓	✓	✓
	数据湖分析	✓	✓	✓	✓
规范设计	主题设计	-	-	-	✓
	数据标准	-	-	-	✓
	数据建模	-	-	-	✓（模型数1,000个）
	数据指标	-	-	-	✓（指标数1,000个）
数据加工	数据开发	✓（任务数50个）	✓（任务数100,000个）	✓（任务数100,000个）	✓（任务数100,000个）
	数据集成	✓（任务数10个）	✓（任务数1,000个）	✓（任务数1,000个）	✓（任务数1,000个）
	作业运维	✓（调度100 次/天）	✓（调度1,000 次/天）	✓（调度5,000次/天）	✓（调度5,000次/天）
	项目设置	✓	✓	✓	✓
数据治理	数据质量	-	-	✓	✓
	数据血缘	-	-	✓	✓
	数据标签	-	-	✓	✓
数据安全	数据权限	✓	✓	✓	✓
	隐私数据保护	-	-	✓	✓
数据服务	API创建	✓	✓	✓	✓
运维中心	作业运维	✓	✓	✓	✓
系统中心	资源管理	✓	✓	✓	✓
	系统监控	✓	✓	✓	✓
	文件管理	✓	✓	✓	✓

🔗 购买方式

从产品官网进入，点击「立即使用」，选择某个区域开通后，即可根据需求购买相关版本产品。

🔗 版本升级

产品支持版本升级操作，升级后版本到期时间不变，仅需支付当前计费周期剩余时间差价，升级完成后即时生效。

🔗 续费说明

功能版本到期可以进行续费，EDAP会提供7天的宽限期，若7天后仍未续费，则自动降为体验版，高级版本专有功能将立即停止使用，在购买时候您可配置“自动续费”，以避免功能版本续费不及时。

🔗 退费说明

需提工单进行退费申请，退款金额 = 当前有效订单金额 + 未开始订单金额 - 资源已使用价值

资源购买说明

EDAP目前已提供数据开发、数据分析、数据服务资源组，采用按量付费的计费模式，本文将为您介绍资源组计费说明。

🔗 资源组规格说明

- 数据开发资源组：动态资源规格，根据资源实际使用量进行计费。
- 数据分析资源组：常驻资源规格，根据规格数和运行时长进行计费。
- 数据服务资源组：常驻资源规格，根据规格数和运行时长进行计费。

🔗 购买方式

进入产品后，在资源管理功能下配置资源组规格，完成后会自动生成订单。

🔗 计费规则

- 计费项：CU
- 计费方式：按时长按量后付费
- 计费周期：按分钟计费，不足1分钟按1分钟计

- 账单周期：按小时扣费，即北京时间整点扣费并生成账单，出账时间是当前计费周期结束后1小时内，例如：10-11点的账单会在12:00之前生成，具体以系统出账时间为准。

产品定价

资源	规格	单价
常驻资源/动态资源	1CU(1核4G)	0.0058元/CU/分钟

计费公式

费用 = CU数量 * 单价 * 资源运行时长

计费示例

例如您创建了20CU的资源组，共运行了29分钟30秒，计费按30分钟计算，总费用为：20 * 0.0058 * 30 = 3.48元。

规格调整说明

您可以在资源管理功能下对资源组进行规格调整，调整完成后即按实际规格进行计费。

启停说明

您在界面上对资源进行启动和禁用操作，仅影响计算资源的可用性，资源计费将持续进行。

欠费说明

购买资源组到期前，会发送邮件、短信通知，提醒您及时续费。到期后未续费将停止服务，EDAP会提供7天的宽限期，宽限期内续费服务将恢复运行，否则将对资源进行销毁且无法恢复。

入门指南

示例场景说明

本示例是一个EasyDAP全场景入门教程，介绍如何在EasyDAP中完成数据开发、数据治理、数据分析、数据服务全流程，帮助用户对产品能力建立快速了解。本示例基于一张客户信用评分表，选择EDAPDataLake作为数据湖底座，选择BMR作为计算引擎，使用EasyDAP进行数据开发、治理，最后通过数据服务实现数据开放。

1.1 示例数据简介

客户信用评分表 test_score：

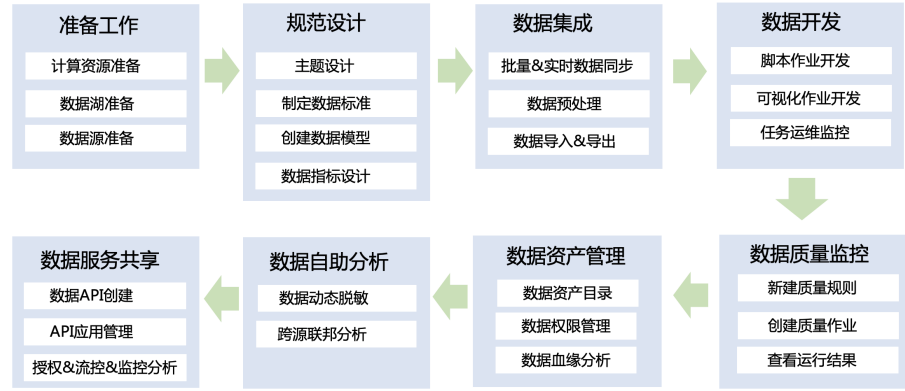
字段名称	字段描述	数据类型
id	编号	Int
name	姓名	String
email	邮箱	String
score	信用评分	Float

示例数据：

```
insert into test_db.test_score values(1, 'name1', 'e-name1@test.com', 90);
insert into test_db.test_score values(1, 'name2', 'e-name1atetest.com', 90);
insert into test_db.test_score values(2, 'name2', 'e-name2@test.com', 96);
insert into test_db.test_score values(3, 'name3', 'e-name3@test.com', 78);
insert into test_db.test_score values(4, 'name4', 'e-name4@test.com', 69);
insert into test_db.test_score values(5, 'name5', 'e-name5#test.com', -10);
insert into test_db.test_score values(6, 'name6', 'e-name6@test.com', 100);
insert into test_db.test_score values(7, 'name7', 'e-name7@test.com', 1000);
insert into test_db.test_score values(8, '', '', null);
insert into test_db.test_score values(9, null, null, null);
insert into test_db.test_score values(10, null, null, null);
insert into test_db.test_score values(null, null, null, 90);
insert into test_db.test_score values(null, null, null, 90);
```

1.2 流程简介

EasyDAP典型的一站式数据治理开发流程如下图所示：



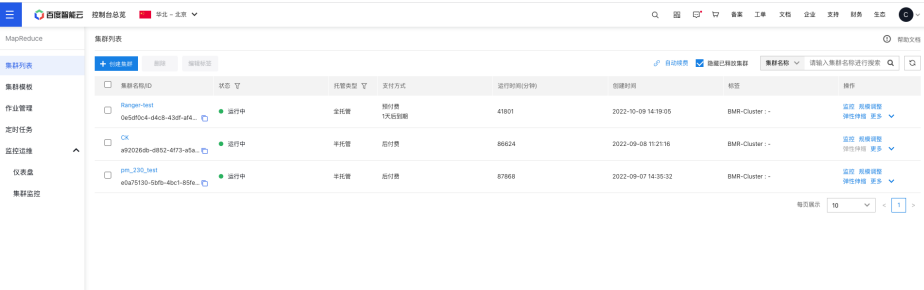
准备工作

为了完成EsayDAP全流程体验，您需要具有EDAPFullControl的系统管理员权限，以及其他关联产品：计算引擎BMR、对象存储BOS、云数据库RDS、API网关的所有权限。

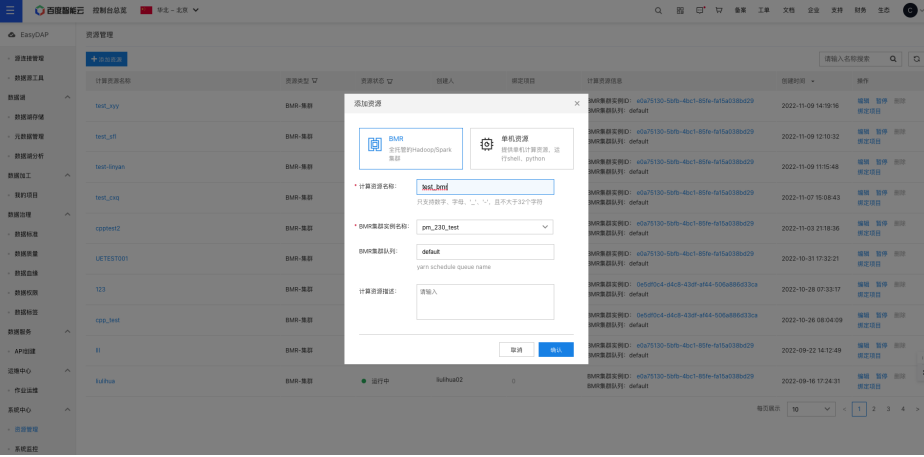
计算资源准备

本示例选择BMR的Hadoop集群作为数据湖的计算引擎，支持在EasyDAP中完成数据的离线和实时开发，以及数据查询分析。

1、通过Console登录BMR产品，新建一个计算资源集群pm_230_test。为完成本次产品体验，除必选组件外，需要增加选择Hive、Spark、Tez。



2、进入“系统中心-资源管理”，点击“添加资源”，新增一个计算资源test_bmr。在弹出的窗口中选择已创建的BMR实例，完成属性配置。

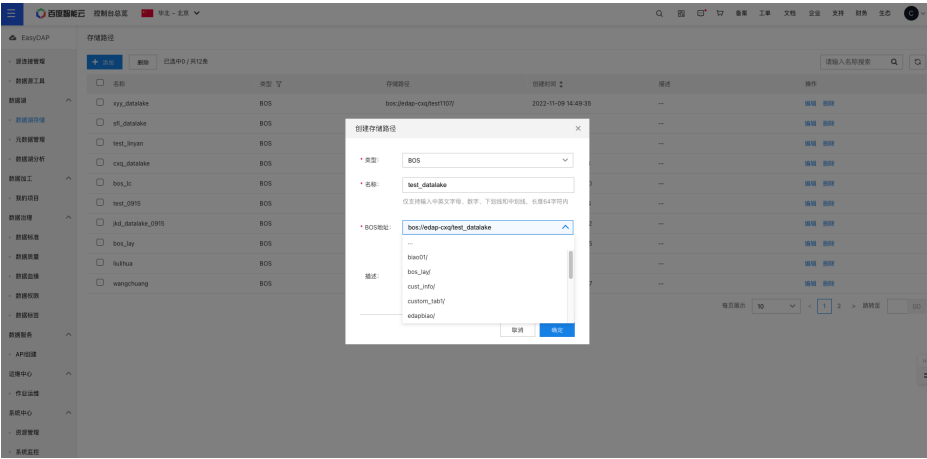


数据湖准备

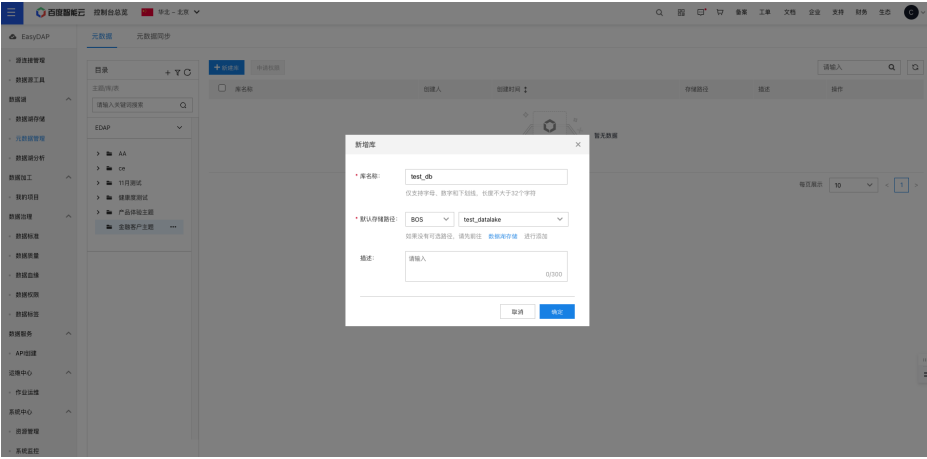
本示例选择EDAPDataLake作为数据湖底座，同时也支持选择BMR的Hive组件作为数据湖选型，EasyDAP将基于数据湖进行一站式的数据入湖、数据湖开发、数据湖治理、数据湖分析操作。

1、进入“数据湖-数据湖存储”菜单，点击“添加”，创建一个存储路径test_datalake。在弹出的窗口中，设置数据湖存储的类型、名称、地址。

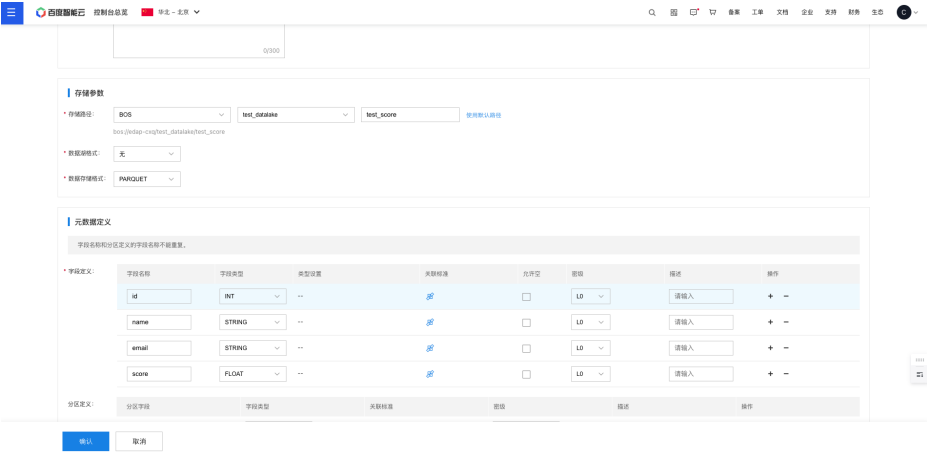
注意：需要先到BOS创建Bucket。



2、进入“数据湖-元数据”菜单，创建主题、数据库、数据表。
本示例新建数据库test_db，选中已创建的数据湖存储test_datalake作为默认存储路径。



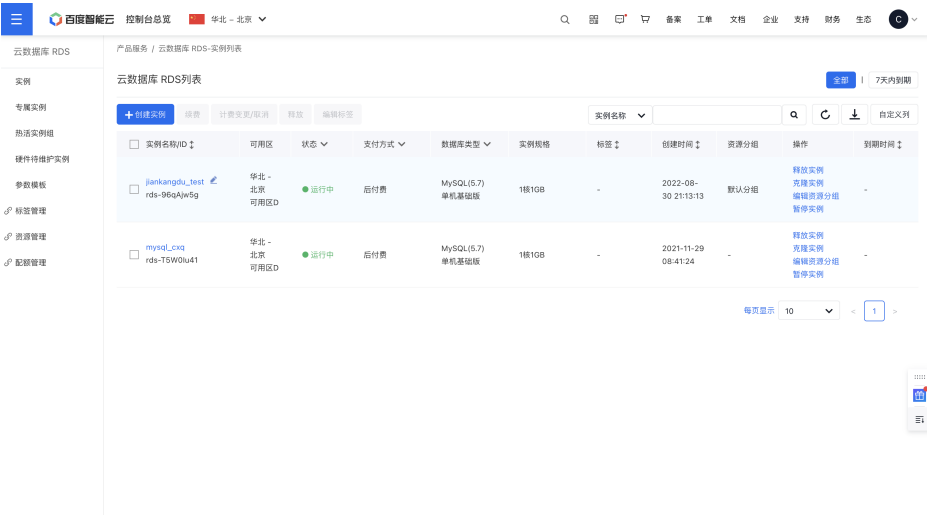
按照1.1示例数据模型，创建数据表test_score，表结构如下图所示：



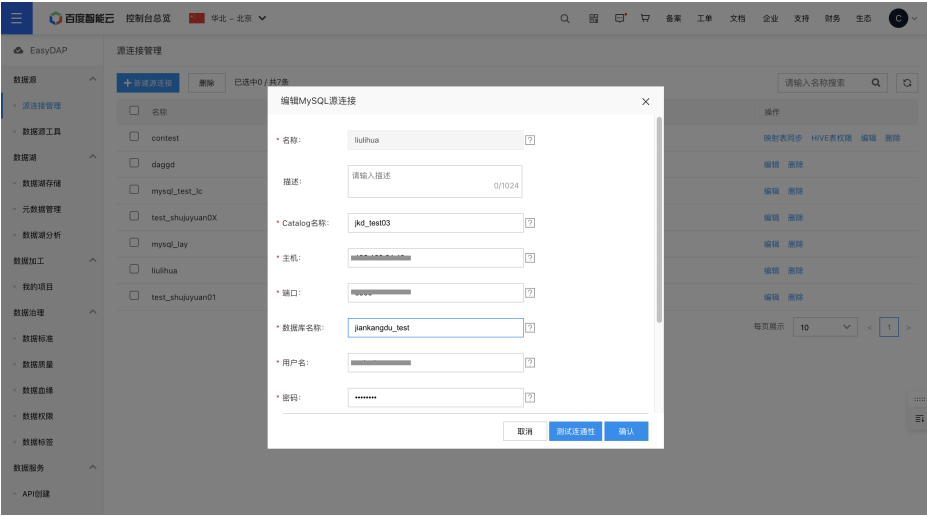
数据源准备

在准备好数据湖之后，在EasyDAP数据源模块创建数据源链接，本示例选择RDS作为外部数据源，后续将把数据湖数据导出到RDS中，以便对接数据服务。

1、通过Console登录云数据库产品，并创建一个云数据库实例。



2、进入EasyDap“数据源”模块，创建MySQL的数据源链接，在弹出的窗口中进行数据源属性配置。



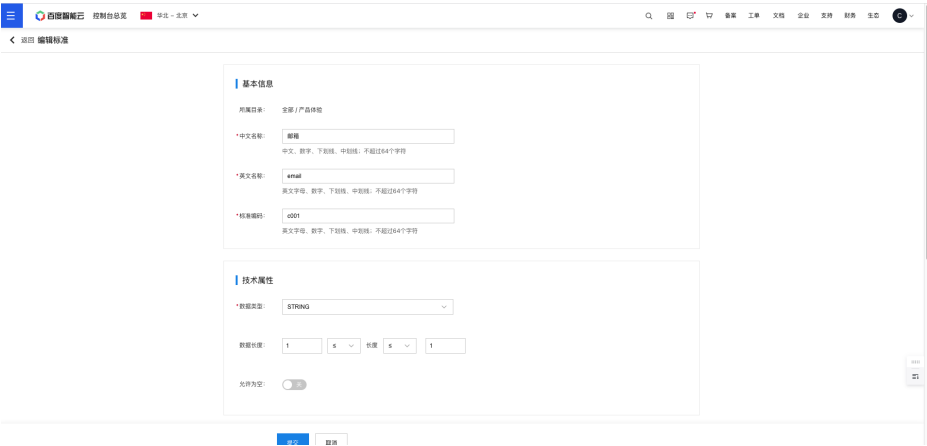
3、点击“测试连通性”，测试通过则完成EasyDAP与外部数据源的链接，可通过EasyDAP对数据源进行访问和读写操作。

规范设计

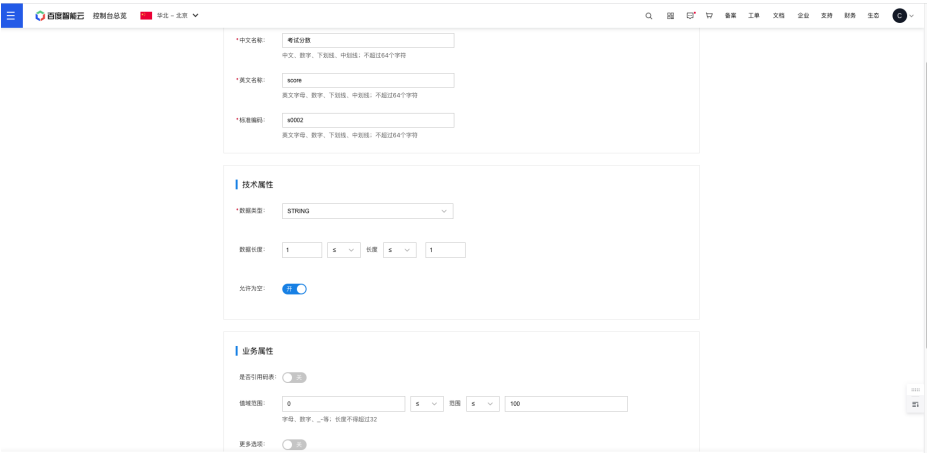
EasyDAP引入事前治理过程，通过制定数据标准，设计数据模型并引用标准，对数据表的开发进行规范，从而避免后期产生数据质量问题。本示例以创建邮箱、评分标准为例，进行产品流程贯通。

制定数据标准

1、点击“数据标准-标准管理”菜单，创建一个目录，并且新建数据标准。邮箱数据标准：注意邮箱不允许为空。



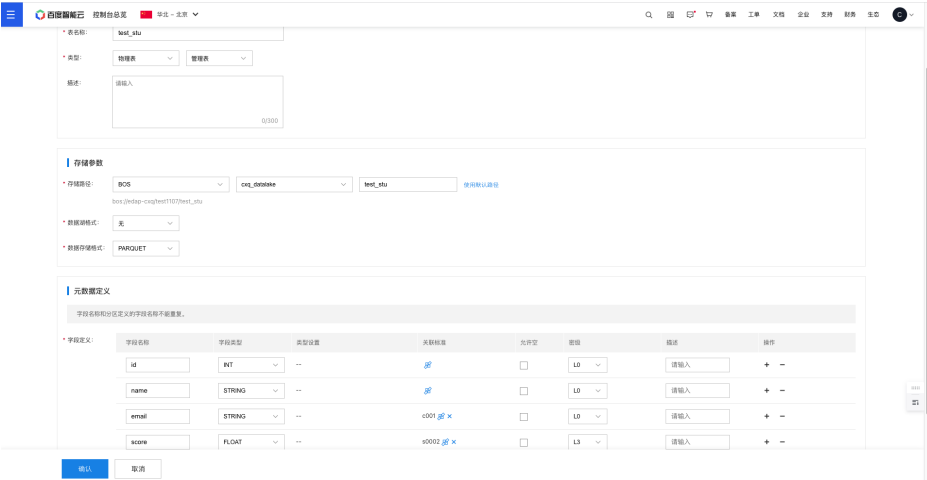
评分数据标准：注意值域范围在0-100之间。



2、点击“数据标准-标准审核”菜单，选择待审核的两个标准，审核通过的数据标准即为成功发布。

设计数据模型

1、进入“数据湖-元数据管理”菜单，选择已创建的数据库，点击“新建表”，按照1.1的数据模型，创建表test_score。



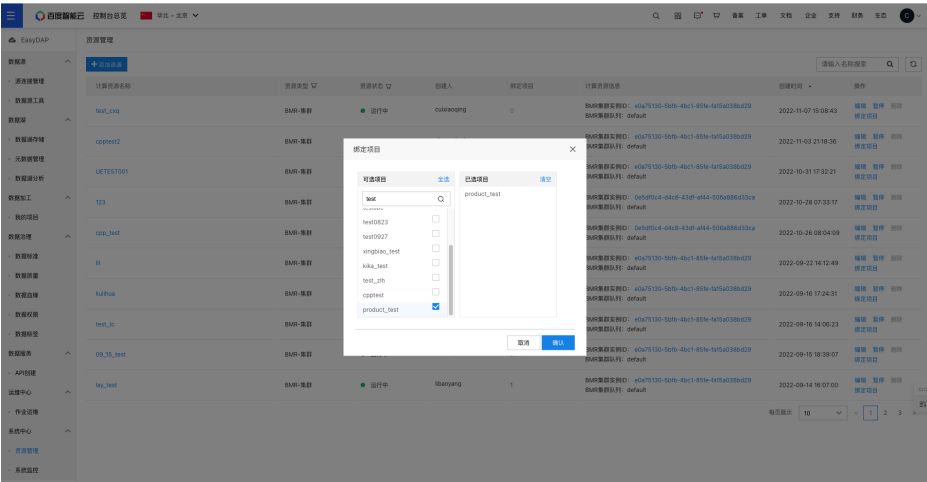
2、字段打标

字段email、score引用数据标准。

数据开发

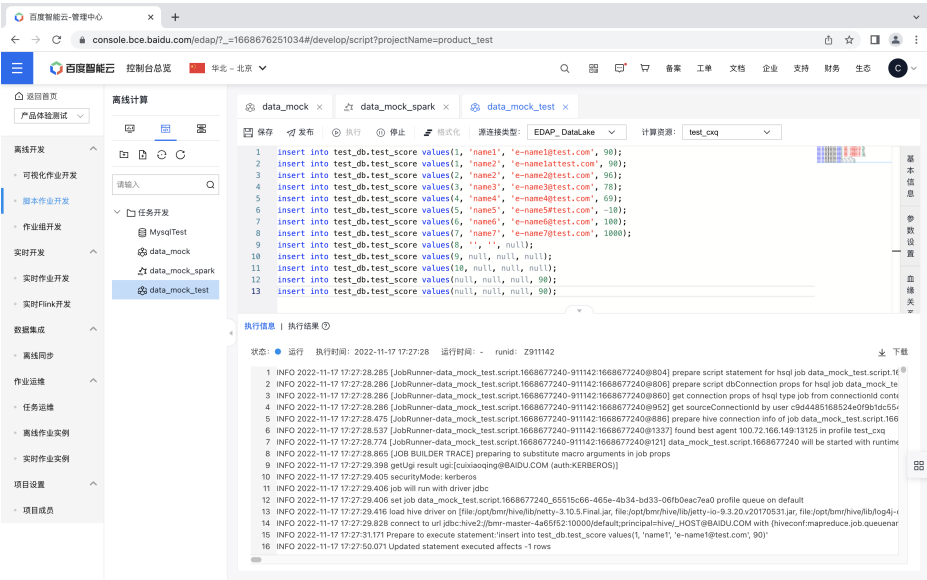
新建项目

进入“数据加工-我的项目”菜单，创建项目“产品体验测试”，并为项目绑定计算资源。



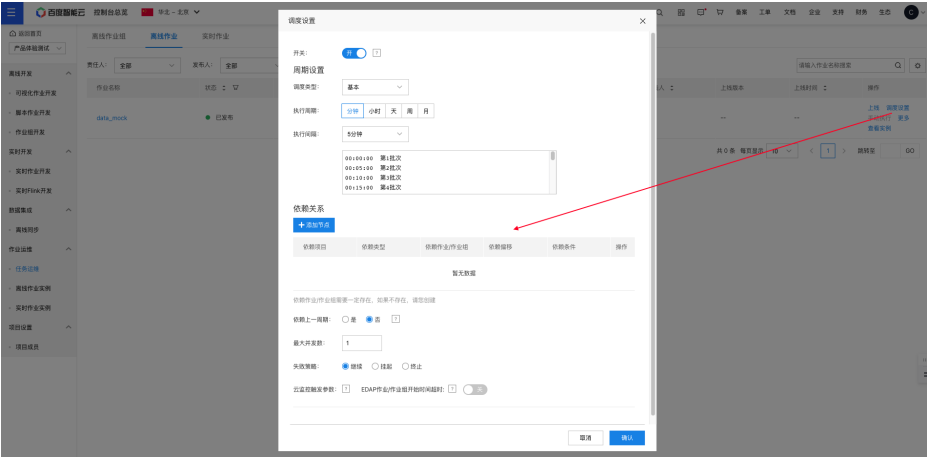
数据开发

进入项目“产品体验测试”，创建一个脚本作业，脚本类型HiveSQL、SparkSQL均可。将1.1中的示例数据写入表test_stu，并点击执行。



作业运维

执行通过的作业，点击“发布”后将进入运维阶段，在“任务运维”中，可以对该作业进行调度配置、查询运行实例等。



数据质量

数据汇聚到平台后，需要对数据的完整性、唯一性、有效性、准确性、一致性、及时性进行探查，清洗脏数据，以确保数据的存储和共享的质量。

创建质量规则

根据字段标准规范，为数据表test_score添加如下质量规则：

字段	规范	质量规则
id	唯一性	字段唯一值
email	邮箱格式	邮箱校验
	不允许为空	字段空值
score	值域范围在0-100之间	字段最小值、字段最大值

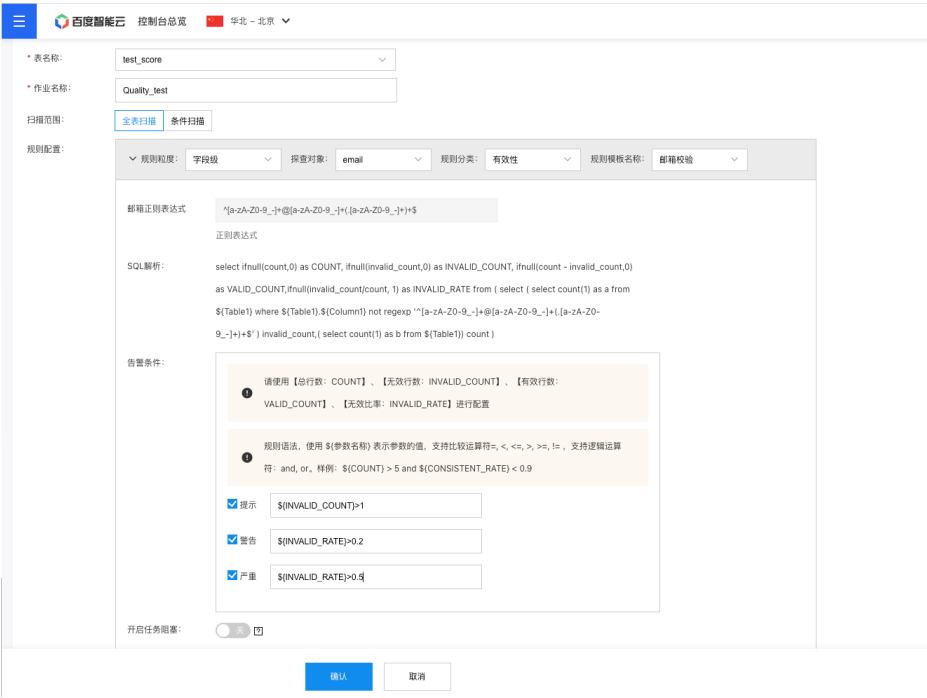
备注：以上质量规则均在系统规则模板中，可直接使用。

创建质量任务

进入“数据治理-数据质量”菜单，点击“新建质量作业”，对数据表进行监控配置。

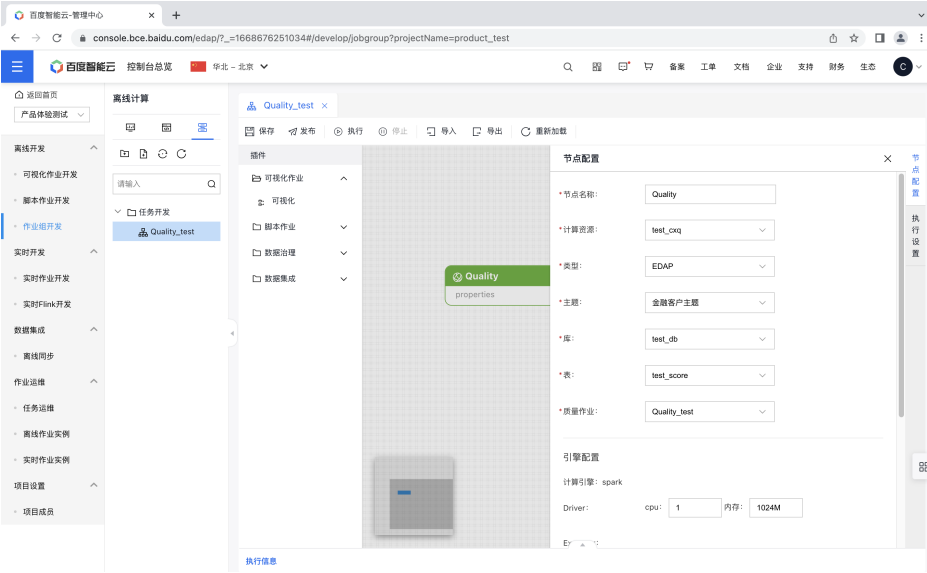
例如对email字段添加邮箱格式校验规则，并且配置了三个等级的告警条件：

- 无效行数>1时，进行提示；
- 无效比率>0.2时，进行警告；
- 无效比率>0.5时，进行严重警告。

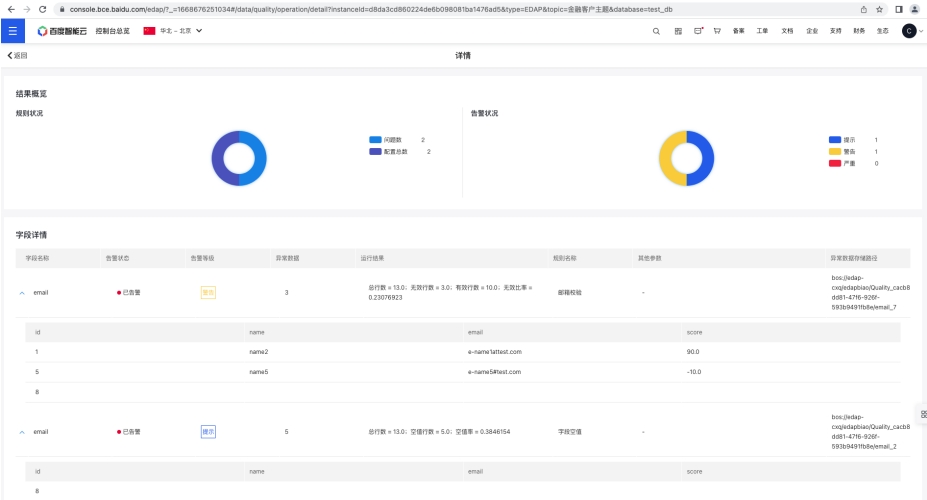


运行并查看结果

1、进入“项目-作业组开发”，拖拽Quality节点，在节点配置中选中创建好的质量任务，点击执行。



2、运行成功后，可进入“数据质量-运维管理”中查看质量作业的探查结果。包括告警状态、告警登记、异常数据明细等。用户可针对问题数据进行清洗、整改。

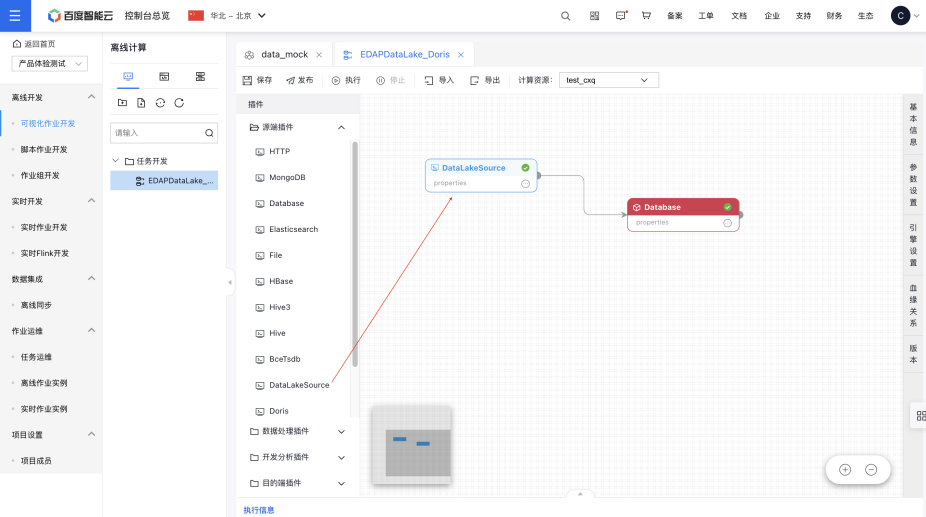


数据同步

为了方便数据发布生成数据API，需要将数据湖中的数据进行导出，本示例通过可视化ETL作业，将数据从EDAP数据湖同步到RDS。

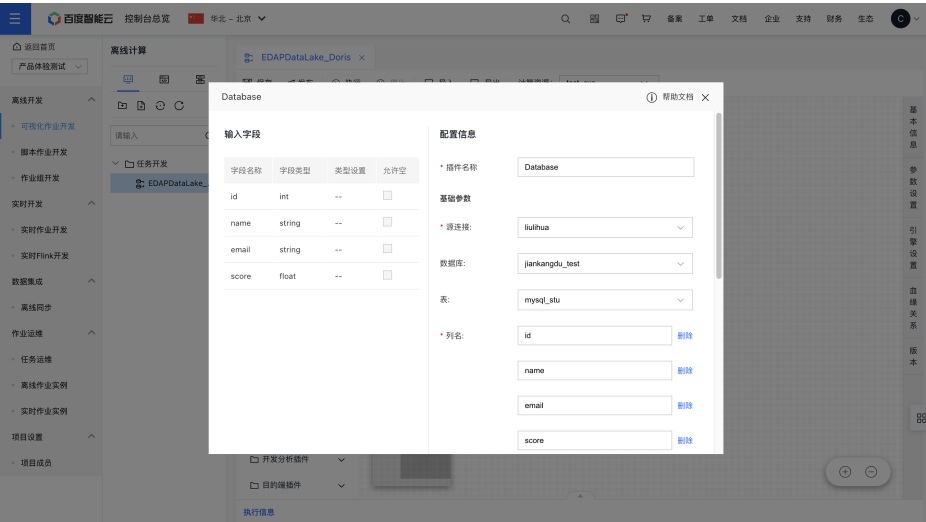
数据湖数据导出到RDS

1、进入项目“产品体验测试”，选择可视化作业，在画布中添加源端插件DataLakeSource、目标端插件DataBase。

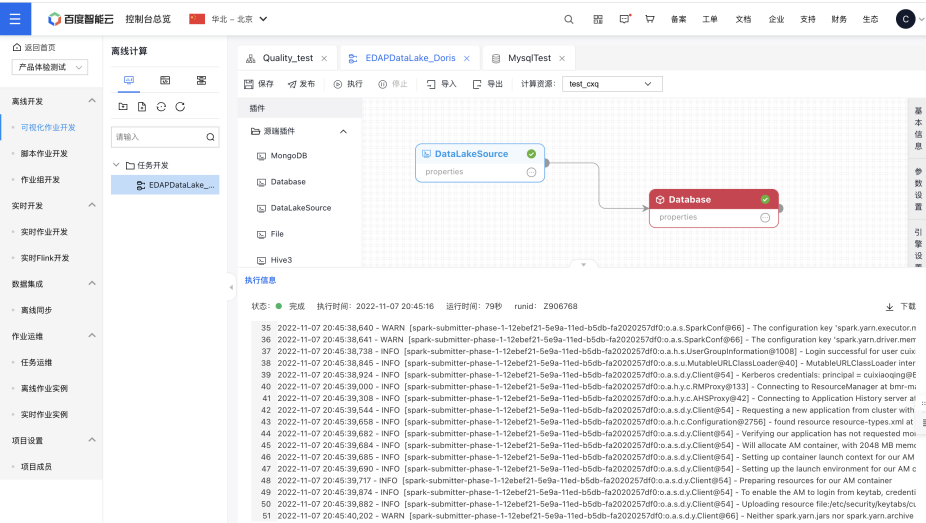


2、点击插件进行属性配置，源端数据表为test_stu，目标端数据表为mysql_stu。

注意：该表需要提前在RDS中创建。



3、点击执行，完成数据同步。



数据服务

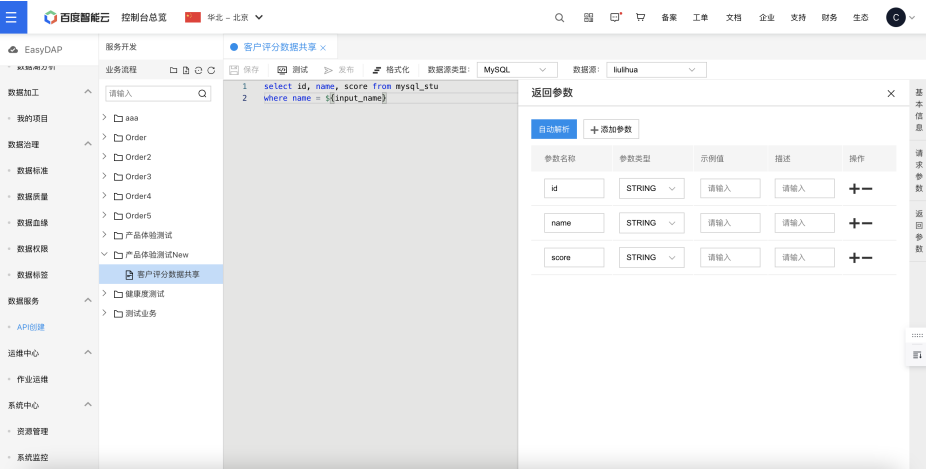
本示例将客户信用评分数据以API的方式进行共享，以便支持数据大屏分析等外部应用，本示例API的请求参数为客户姓名，返回参数为客户编号、姓名、信用评分。

API创建

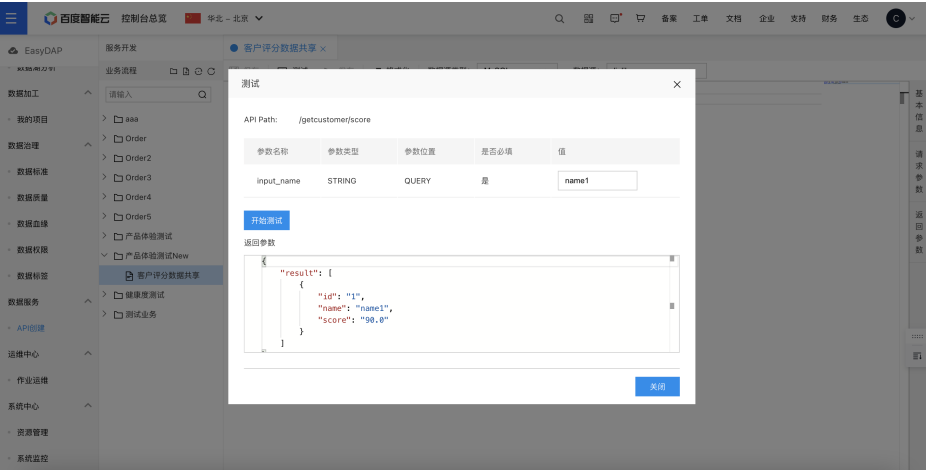
1、进入“数据服务-API创建”菜单，并创建一个业务流程，在弹出的窗口选择一个网关分组，后续API发布则会发布到对应的分组中。

注意：需要开通API网关并创建一个网关分组。

2、新建一个API“客户评分数据共享”，通过SQL的方式进行API开发，点击“自动解析”请求参数和返回参数。



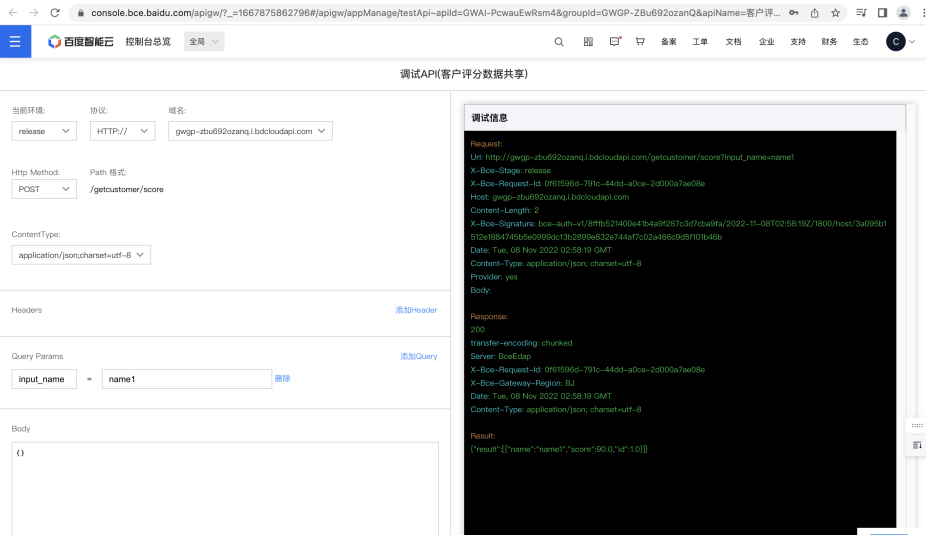
3、点击API测试，输入参数后点击“开始测试”，在下方返回参数框验证返回结果。



4、测试通过后，点击“发布”，用户即可在API网关的API列表中查看到对应API。

API调用

通过Console跳转到API网关为API绑定后端密钥，授权到应用，进行数据调用。详细操作步骤详见：[数据服务操作手册](#)



操作指南

准备工作

多用户访问控制

多用户访问控制

一、多用户访问控制描述

多用户访问控制(Identity and Access Management, IAM)，主要用于百度智能云的身份管理和访问控制，解决云账户的集中授权与管理、资源分享与多用户协同工作等问题。多用户访问控制适用于企业内的不同职能角色，你可以对不同员工赋予产品的不同权限，以共享你账户内的资源，完成他们的工作。当你的企业存在需要多用户协同工作、分享资源时，推荐你使用多用户访问控制。

以下是多用户访问控制适用的典型场景：

- 中大型企业客户：对公司内多个部门的不同员工进行集中资源和权限管理；
- 独立软件服务商(ISV)或SaaS平台商：对代理客户进行集中的资源和权限管理；
- 中小开发者或小企业：添加项目成员或协作者，进行资源和权限管理。

二、EDAP权限策略说明

EDAP支持系统策略和用户自定义策略两种。

- 系统策略：百度智能云系统为管理资源而预定义的权限集，这类策略可直接为子用户授权，用户只能使用而不能修改。
- 自定义策略：由用户自己创建，更细化的管理资源的权限集，可以针对单个实例配置权限，更加灵活的满足账户对不同用户的差异化权限管理。

系统策略

EDAP系统策略包含EDAPFullControl、EDAPDataManage两种。子用户注册后拥有默认用户权限，可以通过IAM为其授予EDAPFullControl策略，使其拥有完全控制EDAP的权限，或为其授予EDAPDataManage策略，使其拥有EDAP数据管理员权限。具体的权限范围如下：

多用户访问控制

策略名称	权限说明	权限范围
EDAPFullControl	EDAP系统管理员权限	• 首页：查看版本信息，进行升级、续费操作，查看配额信息，查看常用项目列表、项目详情，查看典型场景，一键导入最佳实践 • 我的项目：查看所有项目列表、项目详情、新建项目，删除作为项目管理员的项目 • 数据湖： • 元数据：查看、编辑、删除所有元数据；查看、创建和管理元数据发现任务 • 数据湖：查看数据湖存储列表；新建数据湖存储、编辑数据湖存储、删除数据湖存储

-

数据架构：

-

主题设计：查看主题、主题域列表，创建和管理主题、主题域

-

标准设计：查看数据标准，创建和管理数据标准

-

查看数据指标，创建和管理数据指标

-

查看数仓规划、查看数据模型列表、编辑数仓规划；创建和管理数据模型

-

数据治理：

-

数据地图：搜索数据、创建和管理元数据采集任务，查看、创建和管理数据标签

-

数据质量：查看质量总览，创建和管理质量规则、规则模版、质量作业，查看和修改规则配置

		- 数据血缘：查看、创建和管理数据血缘 - 数据安全：查看、新建、编辑、删除脱敏规则；查看、编辑密级信息；查看、新建、编辑、删除脱敏任务 - 数据服务：查看、新建业务流程，新建、编辑、删除、发布数据服务 - 管理中心： - 资源管理：查看、新建、编辑、删除计算资源，查看、新建、编辑、删除数据源链接，查看文件目录、创建目录、创建函数，上传文件、新建文件夹，查看资源监控 - 权限管理：元数据权限授予、存储路径权限授予、数据源权限授予、查看授予记录 - 审批中心：查看标准列表、审批标准，查看模型列表、审批模型，查看指标列表、审批指标，进行权限审批，查看记录列表、查看审批记录
EDAPDataManager	EDAP数据管理员权限	

•

首页：查看版本信息，查看配额信息，查看常用项目列表、项目详情，查看典型场景、最佳实践

•

我的项目：查看所有项目列表、项目详情、新建项目，删除作为项目管理员的项目

•

数据湖：

•

元数据：查看、编辑、删除所有元数据；

•

数据湖：查看数据湖存储列表；新建数据湖存储、编辑数据湖存储、删除数据湖存储

•

数据架构：

•

主题设计：查看主题、主题域列表，创建和管理主题、主题域

•

标准设计：查看数据标准，创建和管理数据标准

•

查看数据指标，创建和管理数据指标

-

查看数仓规划、查看数据模型列表、编辑数仓规划；创建和管理数据模型

-

数据治理：

-

数据地图：搜索数据，查看、创建和管理数据标签

-

数据质量：查看质量总览，创建和管理质量规则、规则模版、质量作业，查看和修改规则配置

-

数据血缘：查看、创建和管理数据血缘

-

数据安全：查看、新建、编辑、删除脱敏规则；查看、编辑密级信息；查看、新建、编辑、删除脱敏任务

-

数据服务：查看、新建业务流程，新建、编辑、删除、发布数据服务

-

管理中心：

-

		资源管理：查看、新建、编辑、删除数据源链接 权限管理：元数据权限授予、存储路径权限授予、数据源权限授予、查看授予记录 审批中心：查看标准列表、审批标准，查看模型列表、审批模型，查看指标列表、审批指标，进行权限审批，查看记录列表、查看审批记录
无	默认用户权限	首页：查看版本信息，查看配额信息，查看常用项目列表、项目详情，查看典型场景、最佳实践 我的项目：查看有权限的项目 数据湖： 元数据：查看、编辑、删除所有元数据 数据治理：

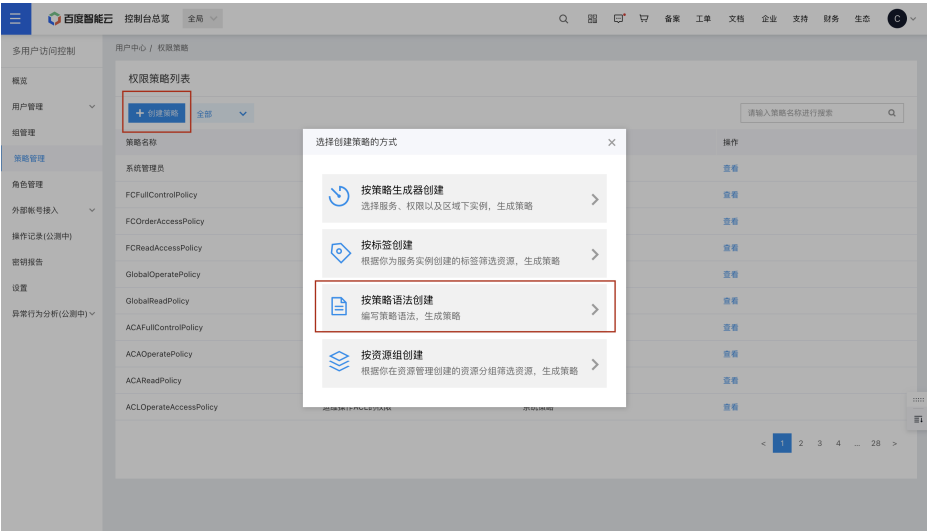
		• 数据地图：搜索数据
		• 数据质量：查看质量总览，创建和管理质量规则、规则模版、质量作业，查看和修改规则配置
		• 数据血缘：查看、创建和管理数据血缘
		• 数据安全：申请数据源、元数据、存储路径权限，查看申请记录
		• 数据服务：查看、新建业务流程，新建、编辑、删除、发布数据服务
		• 管理中心：
		• 资源管理：查看、编辑、删除有权限的数据源链接，查看、使用有权限的计算资源

自定义策略

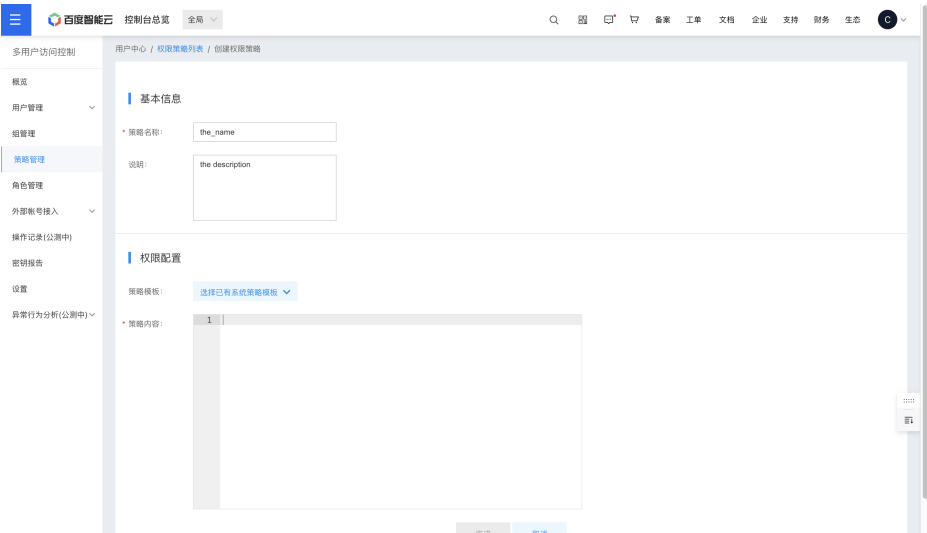
主用户可以通过点击“策略管理—>创建策略”添加自定义策略来进行实例级权限控制，自定义策略的添加有“按策略生成器创建”、“按标签创建”和“按策略语法创

建"三种方式，目前EDAP仅支持"按策略语法创建"。

1. 点击创建策略，选择创建策略的方式“按策略语法创建”。



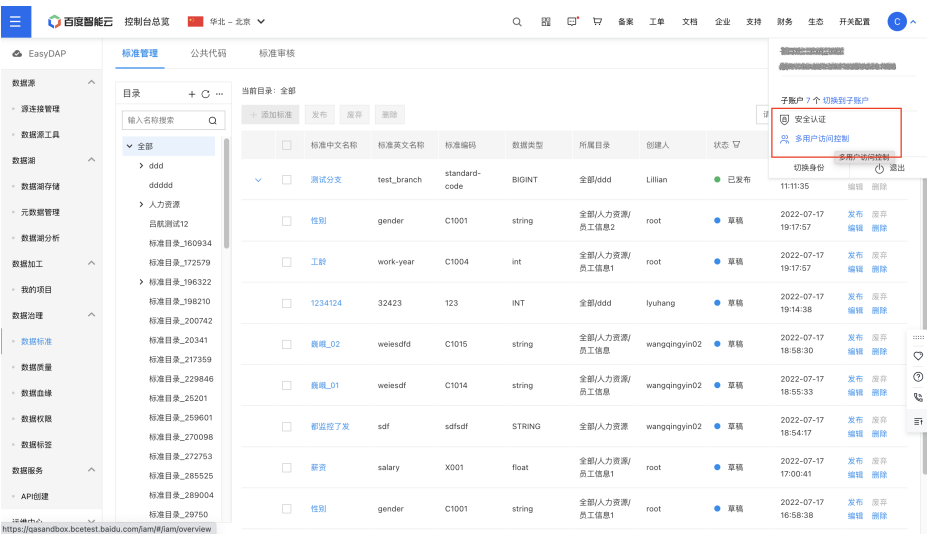
2. 输入自定义的策略名称和策略内容，策略内容需要您编辑 ACL 语句，详细的语法可参考文档策略语法（<https://cloud.baidu.com/doc/IAM/s/8jwwyc3ai>）。



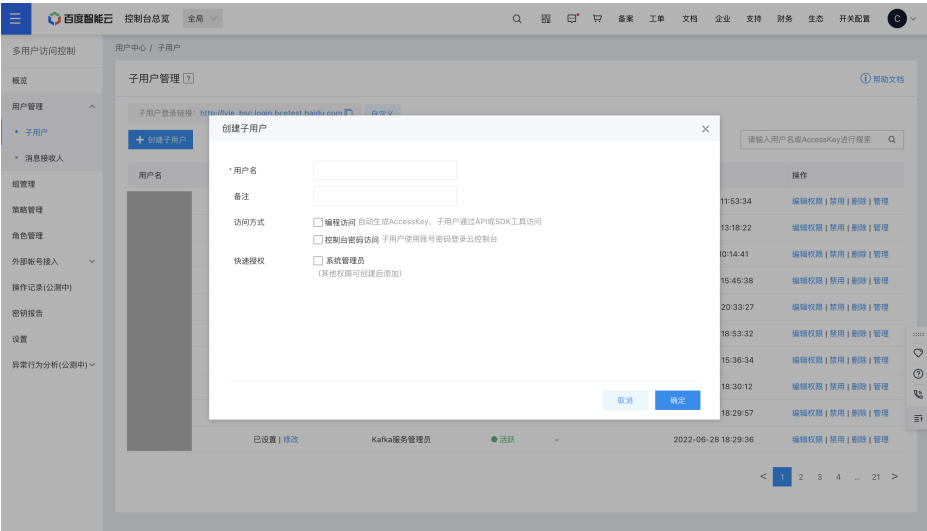
三、授权流程说明

创建子用户

1. 主账号用户登录后在控制台选择“多用户访问控制”进入用户管理页面。



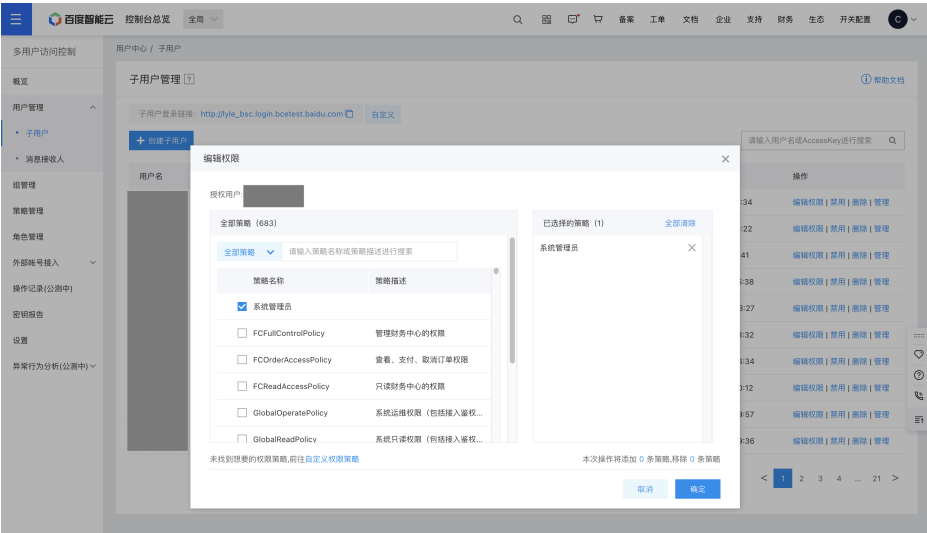
2. 在左侧导航栏点击“用户管理->子用户”，在子用户列表页，点击“创建子用户”。



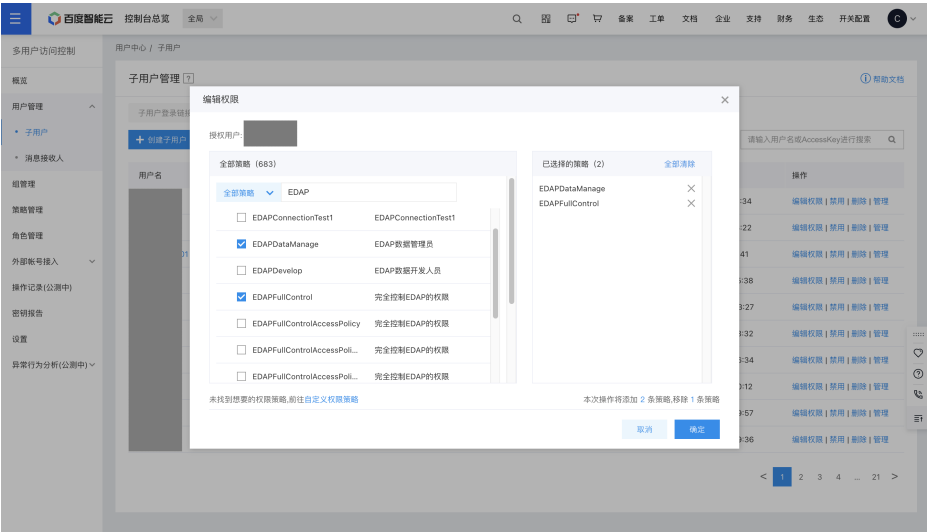
3. 在弹出的“创建子用户”对话框中，填写“用户名”和其他配置信息，点击确认后，即可在“子用户管理列表”中查看到创建的子用户。

子用户授权 当您创建完子用户后，即可为该子用户授予策略。授予后，子用户在访问主用户资源时，将按照策略中已配置的权限范围进行访问。

1. 在“用户管理>子用户”列表页中选择希望授权的子用户，点击“操作”栏中的“编辑权限”。



2. 在授权页面中，选择希望授予该子用户的权限策略（您可以在搜索框中搜索EDAP相关的策略名称），勾选后点击确认完成授权。



数据湖配置

概述

数据仓库vs数据湖

随着大数据、人工智能、云计算、物联网等数字化技术的普及和广泛应用，移动互联网和物联网时代，产生了大量的网站数据，社交媒体数据，物联网设备数据等非结构化数据。数据仓库无法满足这些多元化的数据结构的存储和查询，以及非结构化和结构化数据的交叉分析。数据湖，可以容纳大量的原始数据的存储库和

处理系统。

特性	数据仓库	数据湖
存储数据类型	历史的、结构化的、预先定义数据模型、一般来源 事务系统、运营数据、业务应用程序	所有类型数据，结构化、非结构化、半结构化数据等，数据的类型遵循数据源系统的原始数据格式，IOT设备、网站、移动程序、社交媒体和企业应用程序关系和非关系数据
数据处理方式	高度结构化的架构，数据清洗转换之后加载到数据仓库，即写时模式（Schema - On -Write），用户从数据仓库获取的是处理后的数据。	数据直接加载到数据湖中，然后根据分析的需求处理数据。即 读时模式（Schema - On Read），提供强大的计算能力处理和分析所有类型的数据，分析后的数据存储起来供用户使用。
成本	更快的查询结构，存储成本高	存储海量数据，采用分布式存储系统、对象存储系统 降低存储成本
数据分析应用	多维分析和可视化报表	跨源融合分析、批&流出数据处理、机器学习、预测分析、数据发现和分析

核心概念

- 数据湖：EDAP支持选择EDAPDataLake作为数据湖底座，也支持选择BMR集群的Hive组件，当选择Hive时需要开启数据湖元数据配置。
- 数据湖存储：进行数据湖存储的存储位置。EasyDAP支持的数据湖存储类型为：HDFS分布式文件系统、BOS对象存储。
- 元数据：EasyDAP基于数据湖存储建物理表、基于源连接各类型数据源建映射表，对企业全域数据进行统一元数据管理。并基于元数据表，支持用户高效进行数据集成、批&流数据开发、多源交互式查询分析等数据处理和分析。
- 数据库：表的组织，库下可以有0个、1个或者多个表。库命名全局唯一。
- 数据表：元数据表，EasyDAP表类型包括物理表、映射表。

角色、职责和权限

角色	职责描述	功能权限
系统管理员	拥有EDAP最高权限	全部功能权限
数据管理人员	负责EDAP平台数据资源管理，拥有所有数据的最高权限，进行数据权限管理	· 创建数据湖存储路径 · 查看、管理所有数据湖存储路径 · 创建元数据库、表 · 查看、管理所有元数据库、表
普通用户	在权限范围内，创建数据库、表，进行数据开发分析	· 在权限范围进行存储路径操作 · 在权限范围内进行数据库、表操作

数据湖存储

创建存储路径

操作步骤

- 1.点击左侧导航栏“数据湖存储”，进入EDAP数据湖存储目录；
- 2.点击左上角的“新建”按钮，在弹窗中选择所需要创建的存储路径的类型，并填写名称。

选中【HDFS】类型：填写存储路径名称，下拉选择集群名称，仅支持BMR集群，且状态为运行中。填写HDFS地址，填写存储路径描述（非必填）。完成后点击“确定”按钮，完成创建。

	含义	输入规范	默认值
类型	存储类型	下拉选择，支持HDFS	BOS
名称	所要新建的存储路径名称	仅支持中文、英文字母、数字、下划线、中划线，长度不超过32字符。名称具有唯一性	请输入
集群名称	存储路径所在的资源集群	下拉选择，仅支持选择BMR的资源集群，状态为运行中	请选择
HDFS地址	所要创建的具体文件系统地址	待确认	hdfs://
描述	存储路径的描述信息	支持输入中文、英文字母、数字、特殊字符等，长度不超过200字符	请输入

存储路径管理

操作步骤

1. 编辑：在EDAP数据湖存储目录界面点击“编辑”，在弹窗中可重新选择认证类型，修改用户和密码以及描述。完成后点击“确认”按钮，保存编辑。
2. 删除：

方法1：在EDAP数据湖存储目录界面点击“删除”，在弹窗中点击“确认”，完成删除操作。

方法2：在名称前面的勾选框中选中需要删除的存储路径，点击“删除”按钮，可完成单个或批量存储路径删除操作。

3. 添加：

在EDAP数据湖存储界面点击“添加”，创建存储路径，选择类型为HDFS，填写相关参数，如下图所示。参数定义如表1所示。

表1

参数名称	含义	输入规范	默认值
类型	存储类型	下拉选择，支持BOS、HDFS	BOS
名称	所要新建的存储路径名称	仅支持中文、英文字母、数字、下划线、中划线，长度不超过32字符；名称具有唯一性	--
集群名称	存储路径所在的资源集群	下拉选择，仅支持选择BMR的资源集群，状态为运行中	--
HDFS地址	所要创建的具体文件系统地址		hdfs://
描述	存储路径的描述信息	支持输入中文、英文字母、数字、特殊字符等，长度不超过200字符	--

元数据管理

创建库

操作步骤：

选择默认存储路径：

- 1.点击左侧导航栏“元数据管理”，进入EDAP元数据存储目录；
- 2.点击左上角“新建库”按钮，在弹窗中填写库名称，下拉单选框选择默认存储路径后，再在级连单选框中选择该类别对应的具体路径，填写描述（非必填），完成后点击“确认”按钮，完成新增操作。

备注：在数据库中建表时，表会创建在数据库的默认存储路径下。

创建表 EDAP提供可视化建表和DDL建表两种模式。

可视化模式

操作步骤：

- 1.点击左侧导航栏“元数据管理”，进入EDAP元数据存储目录，点击库名称列的蓝色字体，进入对应库的管理界面；
- 2.点击“新建表”按钮，进入新建界面；

配置项说明：

	配置项	配置项含义	支持范围
基本信息	表名称	数据表的名称	以英文开头，支持英文字母、数字、下划线，不超过128个字符
	表类型	数据表的类型支持三种：物理表-管理表、物理表-外部表、映射表	
存储参数	存储路径	指定数据湖存储路径，用于数据表数据存储	支持选择BOS和HDFS，默认存储路径为：库的默认存储路径/表名称
	数据湖格式	湖仓一体场景下，数据的组织格式	支持选择无、hudi
	数据存储格式	底层的数据存储格式	支持选择TEXTFILE、ORC、PARQUET
字段定义	字段名称	数据表字段的名称	只支持字母开头，支持字母、数字、下划线，不超过128个字符
	字段类型	字段的数据类型	支持TINY、SMALLINT、VARCHAR、CHAR、ARRAY、MAP、STRING、INT、BIGINT、DOUBLE、DECIMAL、FLOAT、BOOLEAN、BYTES、DATE、TIMESTAMP
	类型设置	字段类型选择DECIMAL时，配置数据长度和精度	(precision, scale)格式，p为数字总长度，范围为[1,38]， s为小数点后位数，范围为[0,18]，且s<p。
	关联标准	字段关联的数据标准	
	允许空	是否允许为空	
	密级	敏感数据等级	支持选择L0-L4（L0最低，L4最高），默认为L0
	描述	描述信息	支持中文、英文、数字、特殊字符等，不超过300个字符。

3. 完成表单页面所有内容填写后，点击页面右下角“确认”按钮，完成表的新建。

- 4. 点击操作栏中的“删除”，出现弹窗提示，点击右下角“确认”按钮，可删除表。

DDL模式

操作步骤：

- 1. 点击左上角“DDL模式”按钮，在弹窗中输入SQL语句，可点击查看语法规范。

点击“生成表结构”按钮，完成DDL语句解析。

示例说明

- 1. 创建一张物理表：表类型选择“物理表”，在级连选框中选择细分类型“管理表”或“外部表”。 在存储路径中选择类型并填写路径地址，可点击蓝色字体自动补齐默认路径。

选择数据存储格式，若选中“TEXTFILE”，则需另外填写分隔符。

hudi表类型		使用场景	使用说明
COW - Copy On Write		在只能overwrite的文件系统上实现update操作，写入性能慢，适用于批量写入的场景。	
MOR - Merge On Read	原表	历史数据存放在base文件中，更新数据写入一个.log文件，写入速度非常快，适用于实时写入的场景。	可勾选自动创建读优化表和实时表
	读优化表 - RO表	为提升查询性能，只读取base文件，可能不是最新数据。	表类型必须为外部表
	实时表 - RT表	在读数据的时候，可以实时合并base文件和最新的.log文件，从而读取最新的数据。	表类型必须为外部表

数据湖格式为“hudi”，数据存储格式只支持“PARQUET”。

- 1. 创建一张“映射表”：选择需要创建的映射表的数据源连接类型、数据源连接名称；

例如选择Doris后，填写数据库表名（可点击“补全字段”自动补齐）、分区字段、分区个数、分区下限。

元数据权限管理 数据开发人员申请权限：

- 1.点击左侧导航栏“元数据管理”，进入EDAP元数据存储目录，点击库名称列的蓝色字体，进入对应库的管理界面，在管理界面的右上角点击“申请权限”按钮；
- 2.在弹窗中选中申请的权限内容和有效期，点击“确认”按钮，完成权限申请，等待管理员审核。

数据管理员授予权限：

详情参见“数据分析场景 - 数据权限管理”章节。

元数据版本管理 操作步骤

- 1.点击左侧导航栏“元数据管理”，进入EDAP元数据存储目录，点击库中的表名称，进入对应库的表详情界面；点击“版本管理”，进入元数据版本管理列表页；
- 2.可选择更新时间，查询特定时间区间内的版本更新列表；
- 3.在版本列表中点击“操作”下的蓝字，可将当前版本与历史版本进行对比，修改内容以橙色高亮显示。

文件管理

概述

EDAP提供文件管理功能，用户在数据开发场景中依赖很多脚本、Jar包等文件，可将其上传、更新、删除、文件夹分类等管理，同时离线开发、实时开发中可通过节点配置引用此文件，方便用户对文件的管理使用。

功能介绍

文件管理

在后续数据开发过程中需要使用到第三方的 jar 或者用户自定义脚本、文件的情况，可以通过文件管理页面完成文件上传、管理等相关操作。

提供文件上传、管理等功能：

新建文件夹 新建文件夹，用于对文件分类管理。

上传文件 点击“上传文件”进行文件上传，支持单个/多个（最多10个）文件上传。

文件上传相关限制：

- 1. 支持多个文件上传，最多十个文件；
- 2. 支持单文件最大20M；
- 3. 同文件夹下同名文件直接覆盖。

下载文件 文件列表支持对已上传的文件进行下载操作，文件夹不支持下载。

1. 支持单个/多个文件下载；
2. 判断是否勾选文件夹，勾选文件后“下载”按钮无法点击。

删除文件 支持对已上传的文件进行删除操作：

1. 支持单个/多个文件删除；
2. 文件所有操作调用HDFS侧接口；
3. 判断要删除的文件夹下是否有文件，不允许删除非空文件夹。

文件管理列表 文件管理列表展示目前已上传的所有文件详情：

表单项说明：

表单项	说明
文件名	上传的文件名称
大小	显示文件大小，文件夹不显示大小
创建人	文件上传人
更新时间	文件上传创建时间，文件夹不显示时间
文件路径	上传文件所在全路径
操作	删除：对文件进行删除操作，下载：对文件进行下载操作，不支持文件夹下载

函数管理

函数管理模块支持函数注册、管理。

函数列表

- 支持通过「创建人」进行列表信息过滤检索，支持模糊搜索
- 支持通过函数名称进行列表信息过滤检索，支持模糊搜索，支持刷新
- 点击创建函数即可创建一个新的函数

列表项说明：

序号	字段	说明
1	文件名	上传的文件名称
2	大小	显示文件大小，文件夹不显示大小
3	创建人	文件上传人
4	更新时间	文件上传创建时间，文件夹不显示时间
5	文件路径	上传文件所在全路径
6	操作	使用该UDF的相关示例 • 删除：对文件进行删除操作 • 下载：对文件进行下载操作，不支持文件夹下载
7	操作	对当前UDF进行相关操作：删除、编辑

函数创建、发布

点击创建函数开始创建，基于用户上传的文件，进行UDF创建和发布：

序号	字段	说明
1	数据库	该UDF函数存储数据库
2	函数名称	用户自定义填写UDF名称必填，支持字母、数字、'_'、限50个字符
3	类名	UDF函数的类名必填，支持字母、数字、'_'、'.'、限128个字符
4	文件资源	输入文件资源的路径，支持添加多个
5	描述	使用该UDF的示例代码，非必填，限1024个字符

UDF创建完成后点击【发布】按钮对UDF进行发布，发布成功后当前UDF才能在后续开发中使用。

函数编辑

对已发布、不可用的UDF可进行编辑操作，点击【编辑】按钮操作：

UDF名称不可修改，其余项均可修改

函数删除

点击【删除】按钮执行删除操作：

UDF唯一性限制：统一数据库下只能有一个同名UDF

函数使用

后续数据开发中使用UDF时，需要在对应的库下使用，或者使用库名.UDF名称进行使用。

目前UDF仅注册在EDAP数据库上，可以使用UDF的功能如下：

- （1）数据湖分析，数据源类型为EDAPDataLake时
- （2）数据开发，可视化作业开发、脚本作业开发、实时flink开发数据源类型为EDAPDatalake时，并且sql执行引擎为hsql、jdbcsql、sparksql、flinksql时，可以使用udf

功能	类别	UDF使用方式
数据湖-数据湖分析	EDAP	select edap.databasename.udfname from tableName（说明：蓝色需替换为实际值）
离线开发-脚本作业开发	HIVE SQL	set hive.allow.udf.load.on.demand=true;select [databasename].udfname from tableName

文件引用

在以下场景涉及脚本文件引用时，仅需将文件上传至“文件管理”中，即可以下拉列表的方式引用。

离线开发

- （1）脚本作业开发：Shell、Python、SparkJar
- （2）作业组开发-脚本作业：Spark、Shell、Python

实时开发

- （1）Flink作业

源连接管理

概述

功能介绍

EDPA支持接入十余种数据源类型，包括常见的关系型数据库、非关系型数据库、大数据存储、半结构化存储等。用户可以在源连接管理界面创建与管理源连接，以便后续进行数据同步，实现企业多源异构数据源的数据汇聚。

支持的数据源类型：

数据源类型	数据源类型名称
关系型数据库	MySQL、Oracle、SQLServer、Hana、PostgreSQL、Doris、Greenplum、TiDB
非关系型数据库	Redis、MongoDB、ElasticSearch、Memcached
大数据存储	Hive、HBase、ClickHouse
半结构化存储	FTP
其他	腾讯TSDB、百度云TSDB

角色、职责和权限 角色 | 职责描述 | 功能权限 --- | --- | --- 系统管理员 | 拥有EDAP最高权限。 | 全部功能权限 数据管理人员 | 负责EDAP平台数据源管理，拥有所有数据源的最高权限，进行数据源权限管理。 | 创建数据源链接 ·查看、管理所有数据源连接 | 普通用户 | 在权限范围内，访问数据源进行数据集成、数据开发相关操作。 | 在权限范围进行数据源访问、操作

源连接的管理

新增数据源

创建源连接是数据接入的第一步，所有接入的数据源都将在「源连接管理」中进行统一管理。源连接中存储了所有建立数据源连接的信息，通过提供正确的源连接名称，用户可以找到相应的数据源连接。

操作步骤：

1. 点击“数据源管理 - 源链接管理”，进入源链接管理界面；

- 2. 点击“新建源链接”，选择所需创建的数据源类型；
- 3. 填写表单项，完成数据源连接配置；
- 4. 点击“测试连通性”，测试通过后，点击“确认”完成数据源的新建。

创建ORACLE数据源

- 1. 进入源连接管理页面。
- 2. 在源连接管理页面，单击左上角的新增源连接。
- 3. 在新增数据源弹框中，选择关系性数据库类型中的Oracle。
- 4. 在新增Oracle数据源弹框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
catalog名称	数据源用户数据湖分析时的唯一标识符，以字母、数字、下划线组合，且不能以数字和下划线开头。
主机	数据库对用的主机名
端口	数据库对应端口号
SID	Oracle数据库的对应的oracle system ID。
用户名	数据库对应的用户名。
密码	数据库对应的密码。
数据源版本	默认支持oracle11g，可下拉选择oracle11g、oracle12c、oracle18c、oracle19c。
VPC	私有网络(Virtual private Cloud，VPC) 是一个用户能够自定义的虚拟网络，能够帮助用户构建属于自己的网络环境。

创建HBASE数据源

- 1. 进入源连接管理页面。
- 2. 在源连接管理页面，单击左上角的新增源连接。
- 3. 在新增数据源弹框中，选择大数据存储类型中的Hbase。
- 4. 在新增Hbase数据源弹框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
catalog名称	数据源用户数据湖分析时的唯一标识符，以字母、数字、下划线组合，且不能以数字和下划线开头。
zookeeper的url	hbase所注册的zookeeper地址，例master-78d2977:2181
zNode根路径	hbase注册zk中的节点名，可从hbase-site.xml配置中获取，默认为“/hbase”。
认证类型	支持SIMPLE和KERNEROS两种模式
kdc host	密钥分发中心主机地址
参与者	Kerberos系统中的唯一身份
keytab文件	Kerberos Keytab 文件包含 Kerberos 主体名称和DES加密密钥之间的映射，这些密钥源自用于登录Kerberos密钥分发中心 (KDC) 的密码。
VPC	私有网络(Virtual private Cloud，VPC) 是一个用户能够自定义的虚拟网络，能够帮助用户构建属于自己的网络环境。

创建HIVE数据源

- 1. 进入源连接管理页面。
- 2. 在源连接管理页面，单击左上角的新增源连接。
- 3. 在新增数据源弹框中，选择大数据存储类型中的Hive。
- 4. 在新增Hive数据源弹框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
catalog名称	数据源用户数据湖分析时的唯一标识符，以字母、数字、下划线组合，且不能以数字和下划线开头。
jdbcUrl	使用JDBC连接到HiveServer2的地址，格式为jdbc:hive2://:/
hive集群	hive集群所在资源，即edap资源管理中心中对应的资源名
安全模式	hive集群是否开启了Kerberos认证
metastoreUris	hive服务设为使用远程 metastore 服务器的URI的格式为：thrift://host:port，可从配置项hive.metastore.uris中获取
数据湖启用Hive元数据	是否将该hive源的数据纳入edap数据湖
VPC	私有网络(Virtual private Cloud，VPC) 是一个用户能够自定义的虚拟网络，能够帮助用户构建属于自己的网络环境。

创建Elasticsearch数据源

1. 进入源连接管理页面。
2. 在源连接管理页面，单击左上角的新增源连接。
3. 在新增数据源弹框中，选择大数据存储类型中的Elasticsearch。
4. 在新增Elasticsearch数据源弹框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
catalog名称	数据源用户数据湖分析时的唯一标识符，以字母、数字、下划线组合，且不能以数字和下划线开头。
endpoint	格式为 http://host:port 或 https://host:port 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。
安全模式	hive集群是否开启了Kerberos认证
是否开启HTTPS认证	即访问方式是网络协议为https
keyStore路径	keystore file的路径，可上传至bos的某个路径下。 HTTPS需要 SSL证书。当您生成 SSL 证书时，您正在创建一个密钥库文件(keystore file)和一个密钥库密码(keystore password)，以便连接时使用。
keyStore密码	当您生成 SSL 证书时，您正在创建一个密钥库文件(keystore file)和一个密钥库密码(keystore password)，以便连接时使用。
trustStore路径	trustStore file路径，可上传至bos的某个路径下 TrustStore仅包含客户端信任的证书。
trustStore密码	访问trustStore的密码
VPC	私有网络(Virtual private Cloud，VPC) 是一个用户能够自定义的虚拟网络，能够帮助用户构建属于自己的网络环境。

管理数据源

编辑数据源

在源连接界面的每一个条目的右侧可见编辑按钮，通过点击编辑按钮可在界面弹窗中可对本数据源进行编辑。

源连接删除

在源连接界面的每一个条目的右侧可见删除按钮，通过点击删除按钮即可对单个源连接进行删除。

同时也可以通过勾选每个源连接条目，后点击左上角的删除按钮进行批量删除。 注：针对hive数据源类型的源连接，若源连接关联一个元数据同步任务时将无法进行删除。

🔗 网络联通性保障

测试连通性目的为测试每种数据源的连通性，以保证数据源信息的正确性，以及后续数据源工具和数据集成任务可以正常使用与运行。

网络通联时，界面右上方会弹出相应的反馈框：

反之，当网络不通畅是会针对在右上角弹窗给出失败原因。

🔗 数据源权限管理

EDAP对于源连接提供了权限管控，避免数据源数据越级访问造成数据泄露。下图可见源连接权限位于数据权限页面中：

通过点击左上角授权即可出现授权弹窗，可见可对于不同的用户和用户组进行权限列所列举的权限项进行永久和自定义时间授权。

权限项	说明
查看	被授权用户具有查看源连接信息的权限，即可在源连接管理中可以看到本条目
访问	被授权用户具有连接访问本源连接所指向的数据源的能力
管理	被授权用户可以编辑，删除该条目数据源以及给其他用户进行授权

资源管理

EDAP中用户可以将资源注册进入资源管理，统一进行资源的管控。

概述

EDAP平台的计算资源管理是整个平台资源的入口。用户可以添加BMR资源、单机资源、Serverless资源组作为后续作业开发、运行的中需要依赖的资源。

添加计算资源

在【系统中心】中点击【资源管理】，显示现有资源列表。

点击【添加资源】显示添加资源弹窗页面。

目前EDAP支持添加BMR实例、单机服务器、资源组三类资源：

资源类型	引用配置	资源节点	调度引擎
BMR实例	指定bmrId + yarn queue	全bmr节点	yarn
单机服务器	交互提供Vpc + List + user	单机节点组	NA
资源组	无	cce容器引擎	k8s

BMR实例

EDAP支持添加百度大数据平台BMR作为底层计算资源，添加BMR资源之前需要在BMR产品页面创建BMR集群（可参考[创建集群](#)），BMR集群创建完成后，可选择【BMR】资源，根据要求填写相关信息：

- 计算资源名称：根据实际业务需求可自定义填写当前计算资源名称；
- BMR集群实例名称：下拉选择已经创建的BMR集群实例；
- BMR集群队列：下拉选择已创建的BMR集群队列；
- 计算资源描述：自定义填写计算资源描述。

BMR资源授权

针对BMR资源，需要对当前用户在BMR进行授权。

进入BMR集群管理，点击进入授权使用的BMR集群。

在BMR集群点击【用户管理】，添加BMR账号。

单机资源

EDAP支持基于单机服务器资源运行自己编写的shell、python脚本，选择【单机资源】，根据要求填写相关信息：

计算资源名称：根据实际业务需求可自定义填写当前计算资源名称；

计算资源描述：自定义填写计算资源描述；

配置信息：

- IP地址：填写具体服务器IP地址；

SSH登录认证信息：

- SSH端口号：填写服务器SSH端口号（一般为22）；
- 访问用户：填写服务器访问用户。

环境设置：

- 环境参数：
 - 标准键：JAVA_HOME，填写value值
 - 标准键：python2，填写 python 2.x bin执行文件路径
 - 标准键：python3，填写 python 3.x bin执行文件路径

资源组 EDAP支持Serverless资源组, 用于数据开发、数据湖分析、数据服务。

字段名称	是否必填	输入规范
计算资源名称	必填, 计算资源名称	只支持数字、字母、'_'、'-' , 且不大于32个字符
网络环境	必填, 标识资源网络环境	无
计算资源描述	非必填, 计算资源描述	无
规格设置	必填, 支持配置常驻资源规格和动态资源规格	host校验
网络设置	必填, 配置VPC、子网和安全组	无

🔗 资源管理

计算资源详情查看

点击计算资源名称, 可查看当前资源配置详情。

资源类型 资源类型支持选择全部、BMR-集群、独享资源组、单机资源

资源状态

具体状态说明：

- 创建中：后台创建IaaS资源过程（仅Serverless资源组存在该状态）；
- 初始化：资源注册/创建后进入此状态, 后台进行客户端部署、远程服务部署，部署/探测必要引擎组件, 标注应用、环境标签；
- 创建失败：超过配置的最长初始化时间后, 仍然无可用节点, 进入此状态；
- 运行中：当前资源可用状态；
- 规格调整中：Serverless资源大小调整中（仅Serverless资源组存在该状态）；
- 不可用：当前资源不可用；
- 禁用中：当前资源禁用中；
- 释放中, 清理数据, 销毁RA、远程服务、引擎组件, 释放资源实体节点；
- 已释放, 接入的资源已释放或删除成功。

用户操作 对已有资源用户可进行相关操作：

- 编辑：对资源已有配置进行编辑、修改；
- 禁用：对当前资源进行禁用，禁用后资源不可用；
- 删除：删除当前资源，仅禁用中、创建失败、已释放状态的资源可进行删除操作；

🔗 项目绑定

资源创建之后, 需要将资源绑定到具体的项目, 只有绑定到具体项目之后, 在项目内才能够使用相应的资源。

点击【资源列表】中操作内的【绑定项目】。

显示资源和项目的绑定弹窗。在弹窗内选择需要绑定的项目名称。

绑定后, 在资源页面能够显示, 此资源绑定的项目个数。

点击【绑定项目】的数字, 能够查看具体绑定的项目名称。

行业最佳实践

🔗 电商场景DEMO

DEMO简介

1	内容	备注
2	计算资源	需要用户手动选择一个资源类型
3	数据消费存储	基于用户选择的资源属性, 自动创建存储路径
4	数据源	基于EDAP内置MySQL自动创建示例数据库

内置电商行业最佳实践DEMO, 支持用户一键导入, 用于产品教学体验。

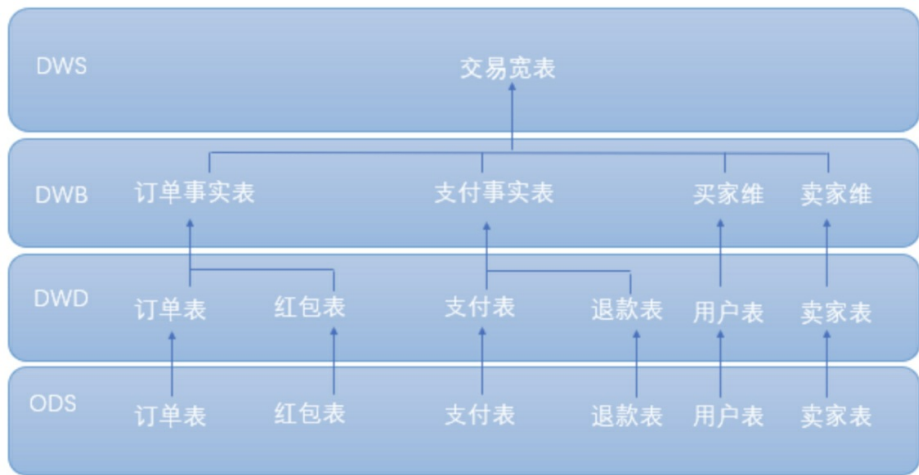
存储路径

- 类型：HDFS
- 名称：电子商务示例存储路径
- 集群名称：所选BMR资源集群
- HDFS地址：需根据建模选择demo_ecommerce_db

数据源

- 名称：电子商务示例数据库
- Catalog名称：demo_ecommerce_connection
- 主机、端口：默认填充
- 数据库名称：demo_ecommerce_db
- 用户名、密码：默认填充
- 数据源版本：默认选择

规范设计



1.主题设计

主题域名称：电子商务示例主题域，主题域编码：demo_ecommerce_subject

主题名称：电子商务示例主题，主题编码：demo_ecommerce_topic

2.数据标准

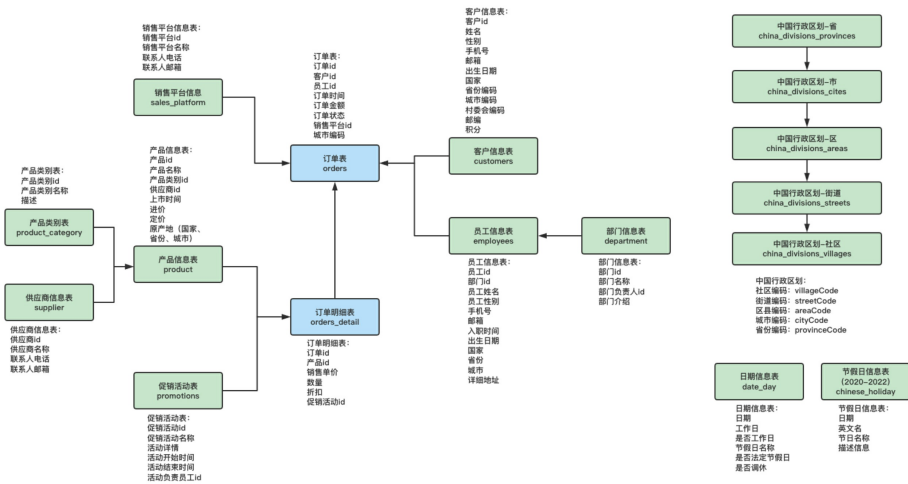
- 目录名称：一级目录与主题域名称相同，二级目录与主题名称相同
- 编码格式统一为：demoecommerce_standard/code{四位数字编号}

数据落标需要自动生成质量任务。

3.数据模型

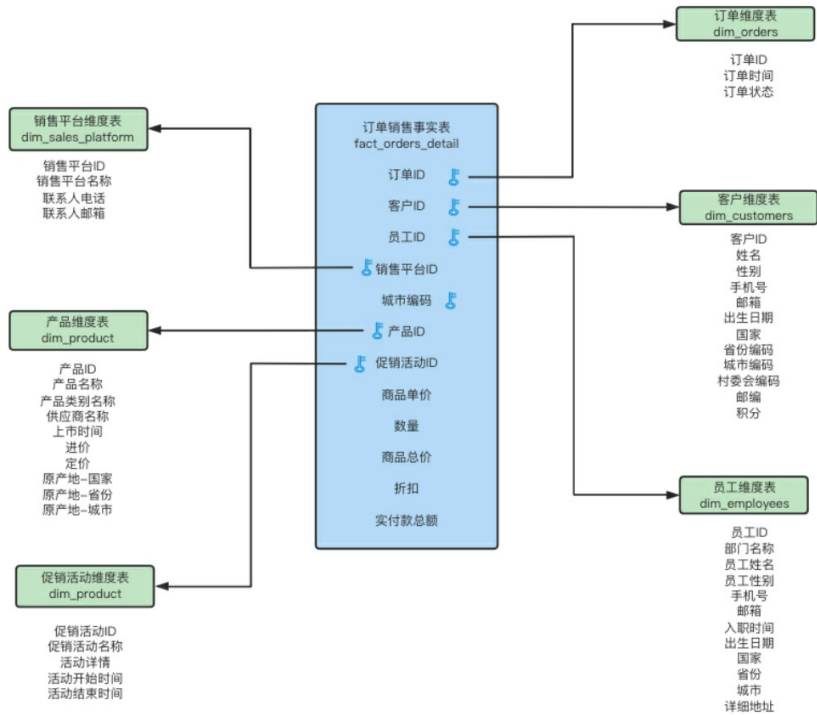
需要包含关系模型、维度模型，默认为已发布、已同步状态。

关系模型（参考）



维度模型（参考）

星型模型



模型物化

- 物化数据库：EDAPDataLake
 - 物化数据库：demo_ecommerce_db
- 4.数据指标

包含原子指标、衍生指标、复合指标的设计，默认为已发布状态。

Demo导入

最佳实践DEMO导入者为项目管理员角色，如果其他用户想访问项目，需要项目管理员将其添加到项目中。

我的项目

左侧导航栏选择“数据加工-我的项目”，可查看导入的“电子商务示例项目”。

数据集成

离线数据同步：Mysql——>EDAPDataLake ODS层，每天进行T-1增量更新；同步的数据库、表范围为关系建模中所有的模型。

数据开发

脚本作业开发：脚本作业类型为HiveSQL，内容为DWR、DWD、DWS、ADS层数据ETL逻辑

作业组开发：需包含数据同步任务、数据开发任务、质量探查任务

作业运维

配置天级别调度，每天0点0分开始执行；导入时先手动执行一次，生成当天任务实例。

Demo暂不支持数据分析和数据服务。

项目管理

项目创建

大数据平台在数据开发过程中以项目为主体，用户需要首先创建一个项目，在项目中进行任务开发、作业组的管理等。

1. 用户在【项目空间】下点击【新建项目】，在弹出框中输入项目名称、项目ID、描述等项目信息。
2. 点击【确认】后，在项目列表中显示新建项目。

项目操作

点击项目列表中的【修改】，可以对项目进行编辑修改，目前允许用户进行的编辑修改包括项目名称和项目描述。

点击项目列表中的【删除】，可以对项目进行删除操作，删除操作需要谨慎操作，目前不支持已删除项目的恢复。

当点击项目名称时，进入项目的开发界面。在项目的开发界面中，用户可以进行，批量作业、实时作业、数据分析等操作。具体操作方式，参考【批量作业开发】、【实时作业开发】、【数据分析】等模块。

项目参数

🔗 项目参数

1. 通过左侧导航栏的项目设置-项目参数进入全局参数配置页面。
2. 单击【添加参数】，可新增参数。支持按名称搜索，支持参数编辑、保存和取消，支持批量删除。可设置参数上限为100行。

- 填写说明：

填写项	说明
参数名称	最大64个字符，包含：英文、数字、下划线
参数值	最大256个字符

- 调用说明：

作业中调用方式：\${param}

概览

概览是进入EasyDAP默认打开的界面，本章为您介绍关于概览的详细操作，快速访问等功能。

🔗 功能模块

登录百度智能云控制台，选择数据湖管理与分析平台EasyDAP进入概览。在此界面可以看到EasyDAP产品的功能快速入口、配额容量、常用项目、典型场景和最佳实践。另外在页面的右边部分可查看用户的账户信息及快速续费，还有对于产品动态和常见问题的展示。

功能名称	描述
快速访问	- 数据加工：支持数据集成、开发，任务统一运维监控 - 数据架构：数仓分层建设，统一模型、标准、指标体系 - 数据治理：湖仓一体建设，保障数据安全及质量 - 数据服务：高价值数据服务化，支撑应用高效调用
配额容量	您可以查看当前版本的配额及配额用量，确保相关能力被有效使用 - 作业节点调度次数：每天执行的作业调度次数之和，作业类型包含数据开发作业、数据质量作业、数据脱敏作业等 - 数据集成任务数（个）：创建的离线数据同步和实时数据同步任务之和 - 数据开发任务数（个）：创建的离线开发和实时开发作业之和 - 数据模型数（个）：规范设计中数据模型的数量 - 数据指标数（个）：规范设计中数据指标的数量
常用项目	展示您最近访问过的项目，最多展示3个项目。可对常用项目进行快速操作： 集成、开发、分析、运维、删除
典型场景	展示产品提供的4个典型场景，覆盖数据集成、开发、架构、服务，以及该场景下的核心操作流程，点击每一步可进入具体功能模块进行操作。
最佳实践	默认向您提供基于本产品的电子商务行业最佳实践，帮助您快速上手相关功能。
产品动态	提供版本信息和产品动态，帮助您快速了解产品相关信息。
账号信息	展示当前账户id，可点击续费按钮进入续费管理。
常见问题	展示在使用产品中常见的问题，单击查看更多可查看更多常见问题并对应相关文档。

数据湖
湖管理

🔗 存储路径管理

进入EasyDAP，单击侧边导航**数据治理>数据湖**进入存储路径管理。

存储路径管理列表

在列表展示已创建的存储路径，可在列表查看存储路径的详细信息。

1. 单击添加按钮创建存储路径，填写相关配置项后单击确定可创建完成。

表一 列表展示项说明

展示项名称	描述
名称	存储路径的名称，创建存储路径时填写的名称。
类型	BOS和HDFS
存储路径	用户已选择的bos路径，默认格式为bos://bucket/库存路径/表名称，或bos://bucket/库存路径/表名称/分区。
创建时间	创建存储路径的时间。
描述	创建存储路径时填写的描述，如没有展示“”。
操作	- 编辑：单击编辑操作，可对具体存储路径进行修改，修改完毕单击确定可保存修改。 - 删除：单击操作列删除按钮，可删除该条存储路径。另外勾选存储路径在列表上方也可以选择删除。

表二 存储路径配置项说明

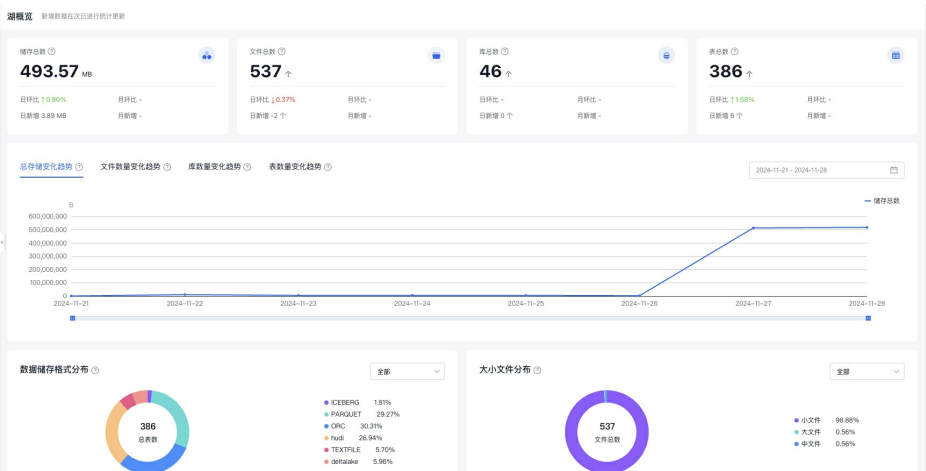
配置项名称	说明
类型	选择类型。类型支持HDFS和BOS。
名称	输入存储路径名称。仅支持输入中英文字母、数字、下划线和中划线，长度64字符内。
BOS地址	推荐 bos://bucket/库存路径/表名称，或bos://bucket/库存路径/表名称/分区 路径格式。如果没有bucket，请 前往BOS添加 。
描述	输入关于存储路径的描述。选填。

🔗 湖概览

1. 进入百度数据湖管理与分析平台EasyDAP控制台，单击页面中数据湖模块进入湖概览界面，或者单击侧边导航**数据湖>湖管理>湖概览**。

2. 湖概览是对数据湖统计的展示，展示存储信息。新增数据在次日统计更新。

图一 湖概览界面



表一 展示指标说明

指标名称	指标说明
存储总数	元数据管理下表的存储总量（仅含BOS）
存储总量-日环比	日环比 = （当日存储总量 - 昨日存储总量 ）/昨日存储总量 *100%
存储总量-月环比	月环比 = （当月存储总量 -上月存储总量 ）/上月存储总量 *100%
存储总量-日新增	日新增=当日存储总量 - 昨日存储总量
存储总量-月新增	月新增=当月最后一日存储总量 -上月最后一日存储总量
文件总数	元数据管理下表的文件总数（仅含HDFS）
文件总数-日环比	日环比=（当日文件总数-昨日文件总数）/昨日文件总数*100%
文件总数-月环比	月环比=（当月文件总数-上月文件总数）/上月文件总数*100%
文件总数-日新增	日新增=当日文件总数-昨日文件总数
文件总数-月新增	月新增=当月文件总数-上月文件总数
库总数	元数据管理中的总库数量
库总数-日环比	日环比=（当日库总数-昨日库总数）/昨日库总数*100%
库总数-月环比	月环比=（当月库总数-上月库总数）/上月库总数*100%
库总数-日新增	日新增=当日库总数-昨日库总数
库总数-月新增	月新增=当月最库总数-上月库总数
表总数	元数据管理中的总表数量
表总数-日环比	日环比=（当日表总数-昨日表总数）/昨日表总数*100%
表总数-月环比	月环比=（当月表总数-上月表总数）/上月表总数*100%
表总数-日新增	日新增=当日表总数-昨日表总数
表总数-月新增	月新增=当月表总数-上月表总数
总存储变化趋势	历史存储总量变化趋势。时间筛选默认：近7天、近1月、近3月
文件数量变化趋势	历史文件数量变化趋势。时间筛选默认：近7天、近1月、近3月
库数量变化趋势	历史数据库数量变化趋势。时间筛选默认：近7天、近1月、近3月
数据存储格式分布统计	元数据管理中的表存储格式占比。支持根据数据库进行筛选，默认全部库。具体类型有Iceberg、PARQUET、ORC、Hudi、TEXTFILE、Delta Lake
大小文件分布	当前目录下文件总数、以及文件分布。小文件：小于等于1MB中文件：小于等于128MB大文件：大于128MB

元数据

元数据管理

元数据管理是构建数据湖十分重要的一环，通过有效集中式的元数据管理，有效提升数据资产价值。您可以使用该功能管理元数据库和元数据表。

元数据列表 支持查看用户有权限的元数据列表，对元数据进行主题、库、表级别的管理，支持库的新增和删除、表的新增和删除、及库表的权限管理。

对于系统管理员和数据管理员角色，该列表展示同步至平台的所有表，并且可以进行新建、编辑等管理操作。

对于其他角色，该列表仅展示用户有权限的表。

库管理

新建库 选中目的数据源，右侧展示当前数据源下的所有数据库，数据库信息详情：

字段名称	字段描述
库名称	展示数据库名称
创建人	展示数据库的创建人
存储路径	展示数据库的存储路径信息
描述	展示数据库描述信息
操作	支持有权限的用户进行编辑、删除数据库。支持编辑数据库的名称、存储路径以及描述信息。

点击右侧更多按钮，点击新建库。

字段名称	字段描述
库名称	填写数据库名称
存储路径	填写数据库的存储路径信息
描述	填写数据库描述信息

编辑库 支持编辑有权限的数据库。

字段名称	字段描述
库名称	编辑数据库名称
存储路径	编辑数据库的存储路径信息
描述	编辑数据库描述信息

删除库 支持删除有权限的数据库，点击删除弹出二次确认弹框，点击确认删除数据库。

库详情 点击库名称，支持查看库详情，详情包含数据库下所有的数据表。

字段名称	字段描述
表名称	展示表名称
类型	包括管理表、外部表、映射表三种类型，对应通过不同方式创建的表
创建人	展示创建人
创建时间	展示创建时间
描述	展示表描述

表管理 支持查看有权限的表的信息，提供快捷查看模型信息，点击【查看关联模型信息】可以跳转至当前表对应的数据模型页面，查看建模信息。**新建表** 选中目的数据库，右侧展示当前数据库下的所有数据表，点击右侧更多按钮，点击新建表，进入新建表页面。

可视化建表

1.基本信息

字段名称	字段描述
表名称	填写表名称，以英文开头, 支持英文、数字、下划线，不超过32个字符
类型	包括管理表、外部表、映射表三种类型： ● 物理表指数据存储在EasyDAP的数据湖中的表，由EDAP进行统一的存储路径管理和元数据管理，包括外部表和管理表。管理表可以删除元数据和数据，外部表只能删除元数据，不能删除数据。 ● 映射表是指针对各类数据源系统通过建外表的形式建立的元数据表，用于在EasyDAP实现数据的查询分析。
描述	填写表描述

2.存储参数

字段名称	字段描述
存储路径	填写数据表的存储路径。点击【使用默认路径】，自动填充表名作为最后一级默认路径
数据湖格式	支持选择无湖格式、Hudi、Delta Lake、Iceberg类型创建表
数据存储格式	支持选择数据存储格式：TEXTFILE、ORC、PARQUET
高级参数配置	支持填写高级参数配置，属于文件/查询引擎的参数，可以指定字段分隔符、字段换行符等，范围比较广，给开发人员使用

3.元数据定义

支持通过界面配置新增字段，字段名称和分区定义的字段名称不能重复。

元数据定义

字段名称和分区定义的字段名称不能重复。

* 字段定义:

字段名称	字段类型	类型设置	关联标准	允许空	密级	描述	操作
请输入	STRING	-	g2	☐	L0	请输入	+ -

分区定义:

字段名称	字段类型	类型设置	关联标准	允许空	密级	描述	操作
请输入	STRING	-	g2	☐	L0	请输入	+ -

DDL建表

用户可以通过点击DDL模式弹窗中的“生成表结构”按钮，并点击确定按钮，以生成新的表。

DDL模式

DDL语句示例:

create external table table_name (column_name1 data_type,column_name2 data_type, ...)
partitioned by (partition_name1 data_type, partition_name2 data_type, ...)
stored as file_format
location 'edap://locationId/relativePath';

[点击查看语法规范](#)

1

取消

生成表结构

编辑表 支持编辑表信息，包括表名称、表描述、存储路径、高级参数，数据存储格式、添加字段。不支持修改表类型、数据湖格式、添加分区。

表详情 提供详细的数据表查看功能，展示表的元数据信息。 **基本信息**

字段名	字段说明
表名称	展示表名称
创建时间	记录表的创建时间
类型	展示表的类型，包括物理表（管理表、外部表）、映射表
数据源地址	展示数据源地址
存储路径	展示存储路径
数据湖格式	展示已选择的数据湖格式
表存储大小	当前表中存储的所有数据大小
表文件总数	当前表中存储的所有文件数量总和
表行数	当前表中数据总行数

注：首次进入无数据，用户需根据界面提示执行特定语句后，返回页面查看统计数据。

分区表不能统计存储大小、文件总数、行数

湖格式的表（Iceberg、Hudi、Delta Lake）不能统计存储大小、文件总数、行数

字段信息 展示表的字段信息和分区信息，包括字段类型、关联标准、是否允许为空、密级配置、描述等信息。

字段定义

字段名称	字段类型	类型设置	关联标准	允许空	密级	描述
id	INT	-	-	是	L9	-
name	STRING	-	-	是	L0	-

版本管理

每次保存元数据信息新增一个版本。

字段名	字段说明
版本名称	自动生成版本号，点击版本号支持查看版本详情
操作人	记录操作人名称
最后更新时间	记录表结构的最后更新时间
操作	支持对比上一版本，对比当前版本

版本对比

类型：MANAGED

数据源地址：_bosdata

数据存储格式：ORC

元数据信息

字段定义

字段名称	字段类型	类型设置	关联标准	主键	外键
id	INT	-	-	-	是
name	STRING	-	-	-	是

分区定义

字段名称	字段类型	关联标准	主键	允许空	密级
------	------	------	----	-----	----

类型：MANAGED

数据源地址：_bosdata

数据存储格式：TEXTFILE

元数据信息

字段定义

字段名称	字段类型	类型设置	关联标准	主键	外键
id	INT	-	-	-	是
name	STRING	-	-	-	是

分区定义

字段名称	字段类型	关联标准	主键	允许空	密级
------	------	------	----	-----	----

确认

脱敏规则

展示当前表已添加的脱敏规则，支持为字段添加脱敏规则。

添加脱敏规则

×

* 脱敏列：

?

id

* 脱敏列密级：

L9

* 脱敏规则：

请选择

没有可选项？[前往创建](#)

* 脱敏后密级：

L0

描述：

请输入

取消

确认

点击“[+ 脱敏规则](#)”，为当前表配置脱敏规则。

字段名	字段说明
脱敏列	选择待脱敏的列，只支持选择密级在L0以上的列
脱敏列密级	回显已选择的脱敏列密级
脱敏规则	选择脱敏规则中已创建的脱敏规则，若未创建可点击“前往创建”新建。
脱敏后密级	配置字段的脱敏后密级，若脱敏列当前密级为L9，则脱敏后可选密级为“L8~L0”
描述	填写描述信息，支持维护
数据湖格式	展示已选择的数据湖格式

数据血缘

在选择表后，以该表为中心，显示上下游血缘关系。

1. 选择显示粒度

血缘图谱支持表级、字段级两种显示粒度，默认显示表级。

2. 选择显示范围

默认直系父子，即上游1层、下游1层，如果没有上游或下游，则显示0层。

支持用户自定义上下游层级，可下拉选择：0层、1层、2层、3层。

在图谱中点击节点左侧的“+”、“-”可以展开或折叠上一层级，末端节点无展开/折叠按钮。

3. 查看节点信息

点击节点右侧“更多”，选择“查看节点信息”，右侧显示节点信息：表名称、数据源类型、数据源链接、数据库、更新时间、schema信息。对于数据湖的表，点击“查看表详情”，可跳转到元数据详情页面。

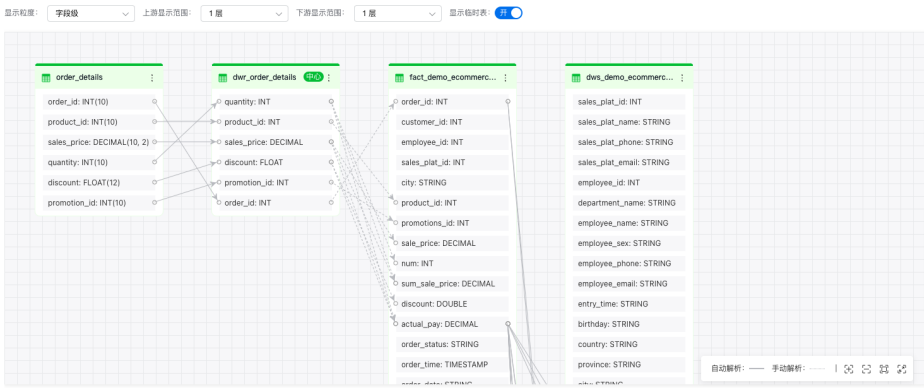
4. 查看任务信息

点击连线上的任务节点，右侧显示产生该血缘关系的任务信息，包括：来源任务名称、任务类型、创建人、创建时间、运行时间、任务内容等。点击“查看任务详情”可以跳转到具体作业。

5. 切换中心节点

双击其他节点，或点击节点右侧“更多”，选择“设为中心表”，可以进行中心节点的切换。

6.显示粒度选择字段级，查看字段级血缘



分区信息

分区信息管理，提供快捷操作，支持用户对分区进行新增编辑、删除、检索。

新增分区

1. 单击新增分区按钮，进入新建分区界面。填写配置项后单击确认分区创建成功。
2. 分区字段类型，当前支持的分区字段类型有：
 - TINYINT, SMALLINT, INT, BIGINT, BOOLEAN, FLOAT, DOUBLE, VARCHAR, CHAR, STRING, BYTES, DATE, TIMESTAMP, DECIMAL;
3. 分区路径：默认路径生成规则：路径类型+数据库路径+表路径+分区路径。

路径说明：

路径	对应含义
bos://	路径类型
/apps/spark/warehouse/icebergdb.db	库路径
ddl	表路径
id_bucket=4/category=asd/	分区路径

4. 新增后的分区以列表形式展示，单击分区信息标签页可查看分区信息。支持选择分区名称精确搜索。

列表字段说明：

字段名称	字段描述
分区字段名称	当前表创建的分区字段
分区创建时间	数据被存储到相应分区中的时间
分区路径	每个分区的路径。填写分区值后，自动生成分区路径。支持手动调整，分区路径必须是当前表路径下的子路径。
文件数	每个分区下的文件总数
分区行数	每个分区下的数据总行数
分区大小	每个分区文件的存储大小
操作	编辑：单击编辑按钮，可编辑分区值。只可编辑分区路径。 删除：单击删除按钮，删除分区。系统会再次确定是否要删除该分区信息，确认无误后单击确认可删除分区信息。

🔗 元数据发现

当数据存储在BOS中，而在EDAP还未与对应的元数据关联时，可以通过元数据发现，来构造这些数据对应的元数据信息，从而支撑用户的查询与分析，将半结构化数据转化为结构化数据。

任务管理

创建元数据发现任务

1. 进入百度数据湖管理与分析平台EasyDAP控制台，单击页面中数据湖模块进入湖概览界面，或者单击侧边导航数据湖>元数据>任务管理（侧边导航）。
2. 单击创建任务按钮，填写创建元数据发现任务配置项，填写完成后单击保存并执行任务创建完毕。

表一 创建元数据发现任务配置项说明

配置项名称	配置项说明
任务名称	填写创建的任务名称
存储路径名称	类型为BOS，只能选择【数据湖-存储路径管理】中已创建的路径。
解析格式	自动解析、Parquet、ORC、CSV、Hudi、Delta Lake
目标数据源	支持选择用户在EDAP-数据源管理中已创建的数据源
目标数据库	- 选择该数据源下的数据库 - 支持刷新数据库
目标数据前缀	输入数据前缀，为发现的表表名增加前缀。以英文开头, 支持英文、数字、下划线，不超过16个字符
执行资源	选择当前项目下的执行资源，包括BMR、serverless
发现策略	全量发现，每次执行全量扫描路径下的文件
执行策略	支持用户配置周期调度、手动执行
表重名规则	覆盖目标表、忽略该表。选择忽略表时对应表创建失败，后续执行也不同步该表
字段更新策略	仅新增列，不删除已有列、更新表结构。选择更新表结构后，元数据发现结果会覆盖原有重名的表，原表可能丢失或不可用，请谨慎配置
表删除规则	删除对应的元数据、忽略删除。 检测到某个元数据对应的存储路径不存在时： - 若您选择「删除对应的元数据」，执行元数据发现任务，将删除该元数据。存储路径被删除，对应的表就被删除。 - 若您选择「忽略删除」，执行元数据发现任务，将保留该元数据。 ● 例如：表 a 的存储路径为 location_a ，存储路径 location_a 被删除，若选择「删除对应的元数据」，执行元数据发现任务，表 a 被删除；若选择「忽略删除」，执行元数据发现任务，表 a 未删除。 注意：选择该项后，将删除已发现的元数据，删除的元数据无法恢复，请谨慎配置。

元数据发现任务列表

创建成功的元数据发现任务以列表形式展示。

表二 元数据发现任务列表字段说明

字段名称	字段描述
任务名称	元数据发现任务名称，单击查看任务详情。
上次开始时间	最近一次任务执行的开始时间，单击箭头可进行排序。
上次执行状态	最近一次任务的执行状态。支持列表按照状态进行筛选。
目标数据库	任务配置中选择的目标数据库。
存储路径	任务配置中选择的目标数据库。
执行策略	周期调度或手动执行。
创建用户	创建该任务的用户。
创建时间	任务创建时间。支持按照创建时间进行排序。
操作	启动、暂停、编辑、删除、执行记录。

元数据发现任务详情

在元数据发现任务列表界面单击任务名称进入任务详情界面，用户可以在详情界面展示任务信息、数据源信息、目标信息和高级配置。

表三 元数据发现任务详情字段详情

展示项	具体字段
任务信息	任务名称、创建时间、创建人、上次运行时间、上次运行状态
数据源信息	存储路径、解析格式
目标信息	目标数据源、目标数据库、目标数据前缀
高级配置	执行资源、执行频率、日志存储路径、目标表路径不存在、目标表已存在、字段更新策略

上次执行状态对应的操作

- 一次性任务

上次执行状态	说明	启动	停止	编辑	删除	执行记录
未运行	任务创建未运行，可修改所有配置项	✓	✗	✓	✓	✗
待触发	任务已提交，等待资源执行	✗	✓	✗	✗	✓
运行中	任务正在运行中	✗	✓	✗	✗	✓
成功	任务所有表采集成功	✓	✓	✗	✗	✓
部分成功	任务部分表采集成功	✓	✓	✗	✗	✓
失败	任务运行失败，所有表采集失败	✓	✓	✗	✗	✓
停止中	任务停止中	✗	✗	✗	✗	✓
已停止	任务已停止，停止正在运行的任务	✓	✗	✓	✓	✓

- 调度任务

上次执行状态	说明	启动	停止	编辑	删除	执行记录	手动执行
未运行	任务创建未运行，可修改所有配置项	✓	✗	✓	✓	✗	✗
待触发	任务已提交，等待资源执行	✗	✓	✗	✗	✓	✗
运行中	任务正在运行中	✗	✓	✗	✗	✓	✗
成功	任务所有表采集成功	✓	✓	✗	✗	✓	✗
部分成功	任务部分表采集成功，与成功状态一致，任务部分成功不阻塞后续调度。	✓	✓	✗	✗	✓	✓
失败	任务运行失败，所有表采集失败。 任务失败时任务调度停止，需要用户点击启动重新启动任务调度	✓	✓	✗	✗	✓	✓
停止中	任务停止中	✗	✗	✗	✗	✓	✗
已停止	任务已停止，停止正在运行的任务	✓	✗	✓	✓	✓	✗

执行记录

执行记录列表

- 进入百度数据湖管理与分析平台EasyDAP控制台，单击页面中数据湖模块进入湖概览界面，或者单击侧边导航数据湖>元数据>执行记录（侧边导航）。
- 查看执行记录可通过运行开始时间进行筛选。
- 支持保存每个任务近一月的执行记录

表四 执行记录列表字段说明

字段名称	字段描述
实例ID	执行实例ID
任务名称	执行任务名称，单击跳转至发现任务详情
任务id	实例所属的任务ID
执行状态	草稿、待触发、成功、部分成功、失败、运行中、已暂停、暂停中。支持按照执行状态进行筛选
开始时间	开始执行时间，支持按照任务开始时间进行排序
结束时间	任务执行完成时间，支持按照任务结束时间进行排序
耗时	执行任务花费的时间，结束时间减开始时间
操作	结果详情：跳转至新页查看运行结果 查看日志：弹窗展示任务日志。注意：草稿、待触发、运行中、已暂停、暂停中状态的任务不能查看结果详情

查看运行结果详情

单击执行记录操作列结果详情按钮，进入运行结果详情界面，查看结果详情和运行结果。

- 结果详情：展示开始时间、结束时间、执行耗时、解析格式、存储路径、执行策略、目标数据源、目标数据库、目标数据前缀、发现策略、表重名规则、字段更新策略、表删除规则、创建人、创建时间、执行资源。
- 运行结果：运行结果以列表形式展示，支持按照表名称进行搜索。

表五 运行结果列表字段说明

字段名称	字段描述
表名称	元数据发现任务生成的结果表名称
目录	当前表所在的子路径名称，作为采集的表名展示
变更类型	新增表、更新表、删除表、无变化、未创建。支持按照变更类型进行筛选
执行状态	成功、失败、全部。支持按照执行状态筛选
错误详情	可以查看运行失败原因。任务可能失败原因： ● 抽取路径下，没有支持的文件类型； ● 表名无效，表名格式不正确或者表名包含不支持的字符； ● 文件中不包含字段名

执行记录日志

单击执行记录操作列查看日志按钮，查看日志信息。所有状态都可单击查看日志。

表、分区发现策略

表发现策略

1. 支持发现普通表/分区表，支持三级分区表。若分区字段发现失败，表结构创建成功，执行记录-结果详情-单表日志中提示“表创建成功，分区创建失败”。
 - 分区发现失败的情况：分区超过三级或分区字段和分区数据不一致导致分区创建失败
2. 发现普通表/分区表：（不满足则发现表任务失败）
 - 文件中不包含字段名；
 - 路径格式要求为库/表/文件或库/表/分区/文件；
 - 表名所在的路径名需要字母开头，仅支持数字字母下划线组合，否则任务失败。
- 3.文件扫描策略： 仅读取路径下最新创建的一个文件。

推断数据结构

1. CSV文件：
 - 至少两列: CSV 文件必须至少包含两列数据。
 - 至少两行数据: 必须有至少两行数据，其中第一行做为标题行，第二行及之后做为数据行。
 - 若标题行不符合建字段规则，默认将列标题设置为col1, col2, col3 等
 - 默认分隔符：逗号 (,)、竖线 (|)、制表符 (\t)、分号 (;)、Ctrl-A (\u0001)

2. 字段发现策略：可以准确识别的字段类型，按照识别结果展示，不能识别字段类型默认为string。
3. 字段创建要求：以英文开头，支持英文、数字、下划线，不超过128个字符，不能输入空格。

分区数据发现

支持发现表的分区数据，根据发现的分区结构，生成每个分区的对应数据，在元数据管理-表详情-分区信息中展示。

数据加工

数据集成

☞ 离线同步

离线同步作为离线数据迁移的主要产品，承担了离线数据上云等数据迁移的重要工作，为整个离线数据开发提供数据，起到了数据准备的作用。目前版本数据集成在数据源端支持 MySQL、SQL Server、Oracle、PostgreSQL、Hana 数据库，在目的端支持 Hive、Doris 以及 EDAP。

基础条件

资源配置

资源配置在EDAP首页侧边栏系统中心>资源管理中，点击**添加资源配置**需要的数据源。

进入离线同步

1. 没有项目的情况下，选择**数据加工-数据集成**系统会自动跳转新建项目界面。点击**新建项目**，填入您的项目相关属性后点击提交项目即创建成功。

注：项目中文名和英文名是必填项，且项目英文名创建后不允许修改。

2. 已有项目，进入离线同步。在概览中选择**数据加工-数据集成（顶部导航）-离线同步**，或者选择侧边导航**数据加工-数据集成（顶部导航）-离线同步**。

操作步骤

任务创建 --> 任务配置 --> 任务运维

任务创建

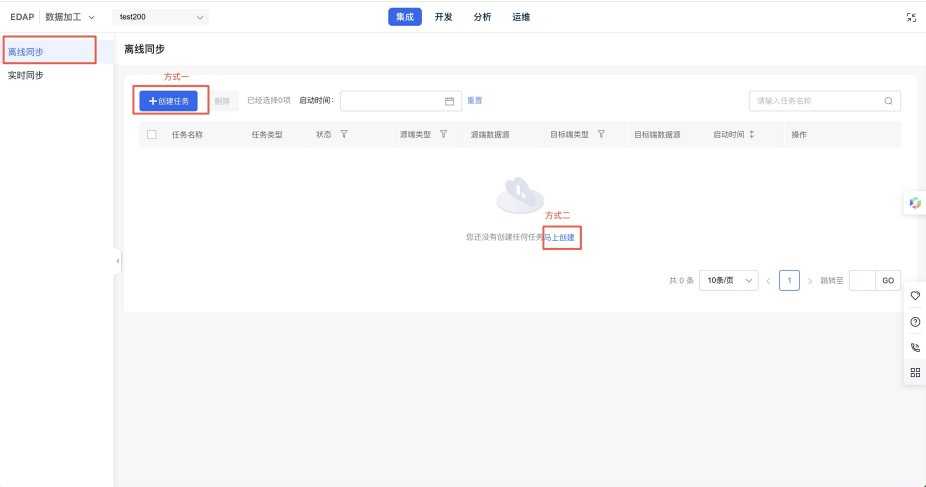
在开始特定数据集成任务的配置之前，必须在平台内创建任务。无论所涉及的数据源或目标的类型如何，都需要执行此预备步骤。

如果没有任务，目前支持两种创建任务方式：

1. 进入离线同步界面，单击**创建任务**按钮。
2. 在离线同步界面中，点击列表页面中间的**马上创建**。

以上两种方式创建任务后，跳转任务配置界面。

图一 创建任务界面



任务配置

目标端数据源类型为Doris、EDAP，配置一个数据集成任务分为“设置同步来源与目标”、“读取设置”、“批量设置”、“写入设置”四个步骤。其中，“设置同步来源与目标”、“读取设置”属于公共配置，“写入设置”在目标端上有所不同。

本章节首先以EDAP为例分析公共配置，然后按照目标端类型分别介绍“写入设置”的操作说明。

设置同步来源与目标

表一 相关配置项说明

类别	中文名称
基本信息	
集成任务名称	只能包含英文字母、数字、中划线和下划线，且以英文字母开头，不大于64个字符
源端设置	
类型	支持MySQL、SQL Server、Oracle、PostgreSQL、Hana五种类型
数据源	在源端设置配置数据类型和选择已配置好的数据源，数据源配置在EDAP主页配置栏。
目标端设置	
类型	目前支持EDAP、Hive、Doris类型
数据源	根据选择的数据类型展示对应的数据源
数据库	根据选择的数据源展示对应的数据库
任务配置	
资源配置	资源配置需要在 EDAP 界面上配置资源，选资源管理，找到合适的计算资源，点击计算资源右侧的绑定项目来绑定到该 EDAP 项目中
读取方式	批量读取
读取频率	定时读取和同步一次。默认为同步一次，定时读取可设置定时触发的周期性任务。值得注意的是，定时任务的数据集成支持“全量”和“增量”两种方式，在后续的读取设置中，配置where语句为空就是“全量”方式进行数据集成，配置where语句过滤某些特定字段就可实现“增量”方式进行数据集成，如过滤时间相关字段或单调递增字段等。建议定时读取配置的作业执行间隔时间大于作业的运行时间。定时读取可配置 Cron 表达式；
读取速率限制	默认5000行/s
数据目的地设置	
同步前，允许清空目标表	默认开启。开启后，写入目标表前进行清空。针对分区表，不论是否开启，对同一分区的重复写入，均会先清空分区数据再重写该分区
数据源变化设置	
数据源删除正在同步的表	选择继续同步其他表或者暂停数据任务。已存在表被重命名,系统会将任务表删除,重命名的表识别为新表
同步的表字段被删除	选择继续同步，目的地字段传空值或者暂停数据服务。重命名字段,系统会将任务原始字段删除, 重命名为新增字段
同步的表发现新增字段	选择继续同步新增字段或忽略新增字段。

读取设置

1.

点击读取设置旁边添加按钮，在读取列表中可以在单个同步任务中批量选择多个数据库和表进行读取，还可以在编辑阶段向任务添加更多表。
2.

读取设置中，左上方默认为关闭**批量设置**的状态，点击左边的表，针对每个单表进行配置，通过下方的where语句过滤单表的数据，目前仅支持 logicStartTime 宏。

注：LST宏表达式格式： \${logicalStartTime(yyyy.?MM.?dd.? HH.?mm.?ss.?(SSS)?(.[+]?[0-9]*[yMdhms])?)}

批量设置

可批量设置所有表的部分写入设置。可打开左上方“批量设置”按钮，通过where语句批量过滤不需要的数据，在执行中是将此where语句应用在每个单表上，分别过滤每个所选择的表中的数据。在读取设置结束后，进入批量设置。填入副本数和需要的表名设置后点击下一步即设置成功。

批量设置仅在目标端建表方式为**自动建表**时生效

EDAP类型

EDAP批量设置界面，可批量配置表的存储格式、表名的前后缀、新增字段、静态分区。

目标表设置

表二 目标表说明

配置项	默认值	说明
数据湖格式	无	必选，无、Hudi、Delta Lake、Iceberg
表类型	无	必选，管理表或外部表
存储格式	Parquet	必选，当选择数据湖格式为“无”的时候存储格式为ORC、Parquet；选择数据湖格式为Hudi、Iceberg、Delta Lake时，存储格式只能选Parquet
表名设置	无	无前后缀、增加前后缀、增加前缀、增加后缀
前缀	无	必填

新增字段设置

可以批量增加目标端表字段，同时支持固定值及时间宏函数变量：

表三 新增字段说明

配置项	默认值	说明
字段名称	无	新增字段名称，若新增字段名称与表已有字段冲突，则单表配置或任务完成时进行右上角报错提示：新增字段与表已有字段冲突！
字段类型	String	枚举值：String、INT、BIGINT、DOUBLE、DECIMAL
默认值	无	可填写固定值，如，1可填写时间宏函数
字段注释	无	输入注释
类型设置	无	主要是针对前面一列的类型做特定的配置，当前decimal， varchar, char类型会用到，设置精度和长度
操作	无	新增一行/删除

分区设置

表四 分区字段说明

配置项	默认值	说明
字段名称	无	新增字段名称
字段类型	String	枚举值：String、INT、BIGINT、DOUBLE、DECIMAL
分区值	无	填写分区值。1、\${timeFormat([column],format)}；其中column为源表时间类型(Date、Timestamp 、DateTime)字段名，例如：\${timeFormat(inspection_time,'yyyy-MM-dd')} 2、\${logicalStartTime(format,[offset])}；即任务执行时间；其中offset为y-年、M-月、d-日、h-小时、m-分钟、s-秒 例如：\${logicalStartTime(yyyy-MM-dd-HH,-1d)} 3、常量填充
操作	无	新增一行/删除

Doris类型

Doris批量设置界面，可批量配置表的存储格式、表名的前后缀、新增字段

目标表设置

表五 目标表设置字段说明

配置项	默认值	说明
Doris表格式	Duplicate明细表	下拉，必选枚举值：Duplicate明细表、Unique唯一表实时仅支持唯一表。选择Unique表并且源端没有主键或者唯一索引时，默认建立Duplicate表
表名设置	无	无前后缀、增加前后缀、增加前缀、增加后缀
副本数	3	默认副本数为3。如果 BE 节点数量小于3，则需指定副本数小于等于 BE 节点数量。

新增字段设置

可以批量增加目标端表字段，同时支持固定值及时间宏函数变量：

表六 新增字段说明

配置项	默认值	说明
字段名称	无	新增字段名称，若新增字段名称与表已有字段冲突，则单表配置或任务完成时进行右上角报错提示：新增字段与表已有字段冲突！
字段类型	String	枚举值：String、INT、BIGINT、DOUBLE、DECIMAL
默认值	无	可填写固定值，如，1可填写时间宏函数
字段注释	无	输入注释
类型设置	无	主要是针对前面一列的类型做特定的配置，当前decimal， varchar，char类型会用到，设置精度和长度
操作	无	新增一行/删除

HIVE类型

EDAP批量设置界面，可批量配置表的存储格式、表名的前后缀、新增字段、静态分区。

目标表设置

表七 目标表说明

配置项	默认值	说明
数据糊格式	无	必选，无、Hudi、Delta Lake、Iceberg
表类型	无	必选，管理表或外部表
存储格式	Parquet	必选，当选择数据湖格式为“无”的时候存储格式为ORC、Parquet；选择数据湖格式为Hudi、Iceberg、Delta Lake时，存储格式只能选Parquet
表名设置	无	无前后缀、增加前后缀、增加前缀、增加后缀
前缀	无	必填，

新增字段设置

可以批量增加目标端表字段，同时支持固定值及时间宏函数变量：

表八 新增字段说明

配置项	默认值	说明
字段名称	无	新增字段名称，若新增字段名称与表已有字段冲突，则单表配置或任务完成时进行右上角报错提示：新增字段与表已有字段冲突！
字段类型	String	枚举值：String、INT、BIGINT、DOUBLE、DECIMAL
默认值	无	可填写固定值，如，1可填写时间宏函数
字段注释	无	输入注释
类型设置	无	主要是针对前面一列的类型做特定的配置，当前decimal， varchar，char类型会用到，设置精度和长度
操作	无	新增一行/删除

分区设置

表九 分区字段说明

配置项	默认值	说明
字段名称	无	新增字段名称
字段类型	String	枚举值：String、INT、BIGINT、DOUBLE、DECIMAL
分区值	无	填写分区值。1、\${timeFormat([column],format)}； 其中column为源表时间类型(Date、Timestamp 、DateTime)字段名， 例如：\${timeFormat(inspection_time,'yyyy-MM-dd')} 2、\${logicalStartTime(format,[offset])}； 即任务执行时间；其中offset为y-年、M-月、d-日、h-小时、m-分钟、s-秒 例如：\${logicalStartTime(yyyy-MM-dd-HH,-1d)} 3、常量填充
操作	无	新增一行/删除

写入设置

HIVE类型

表类型、存储方式等均在上一步做了批量设置，此步可在左侧目录树中指定表，进行建表方式、表达式、分区等个性化设置，接下来展开说明：

1. HIVE支持自动建表和选择已有表两种方式。

- 当建表方式选择自动建表时，数据集成会在目标端自动创建对应表，字段名称和类型默认与数据源端一致，页面中的“箭头”表示数据源和目的端表的字段映射关系。用户可根据需要更改目的端表名、字段名称、字段类型、默认值。

- 此外，用户可在目的表点击每个字段最右侧的删除符号将此字段删除，表示在数据集成过程中过滤此字段，不同步源端的此字段。然后，在右侧列表最下方可以添加普通字段或分区字段。
2. HIVE表类型支持Hudi表，Hudi表类型默认为MOR。

表十 HUDI表说明

配置项	默认值	说明
EDAP表类型	Hudi	枚举值：Hudi
存储格式	Parquet	Hudi表：Parquet
Hudi表类型	MOR	hudi表枚举值：MOR、COW，默认MOR

3. HIVE类型支持批量配置，单击**批量配置**按钮弹出批量设置弹窗，详见批量设置。

表十一 批量设置字段属性说明

配置项	默认值	说明
数据湖格式	无	无、Hudi、Delta Lake、Iceberg
存储格式	Parquet	必选，当选择数据湖格式为“无”的时候存储格式为ORC、Parquet；选择数据湖格式为Hudi、Iceberg、Delta Lake时，存储格式只能选Parquet
Hudi表类型	MOR	Hudi表枚举值：MOR、COW，默认MOR
Hudi表主键	EDAP_UUID	字段名允许修改，提示说明：若不指定目标端主键，Hudi表默认生成EDAP_UUID字段作为表主键，字段值为随机UUID字符串

4. 当建表方式设置为选择现有表时，需要指定目标表名称。您可以使用“同名映射”或“同行映射”来批量确定数据源端和目的端表的字段映射方式。

此外，还可以在可视化界面中绘制手动连接以自定义字段映射。

5. 打开表达式设置时，

表十二 目前支持的transform函数列表

类别	中文名称	英文名称及参数	返回值	详细功能描述
数学函数	绝对值	abs()	T	T是数值类型，包括int、long、float、double、decimal
字符串函数	连接字符串	concat(String str1, String str2)	String	连接str1和str2两个字符串
	去掉字符串前后的空格	trim()		去掉操作字符串前后的空格
	翻转字符串	reverse()		翻转操作字符串
	返回大写字符	toupper()		返回操作字符串的大写形式
	返回小写字符	tolower()		返回操作字符串的小写形式
	字符串替换	replace(String old, String new)		搜索操作字符串中的old子串，如果能够找到则使用new子串来替换
功能函数	条件过滤	filter(String operator, String criteria)	boolean	operator是支持的运算符，criteria是过滤值，其支持的运算符如表2所示

表十三 条件过滤的支持运算符类型关系如下

序号	操作符	boolean	int	long	float	double	bytes	string	date	decimal	timestamp
1	=	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	!=	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	!=	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	>		✓	✓	✓	✓		✓	✓	✓	✓
5	>=		✓	✓	✓	✓		✓	✓	✓	✓
6	<		✓	✓	✓	✓		✓	✓	✓	✓
7	<=		✓	✓	✓	✓		✓	✓	✓	✓
8	contains							✓			
9	notContain							✓			
10	startsWith							✓			
11	endsWith							✓			
12	notStartWith							✓			
13	notEndWith							✓			
14	matchesRegex							✓			
15	notMatchRegex							✓			

Doris类型

选择数据集成目的端的类型为Doris时，相关配置操作大部分与HIVE的配置操作类似，本小节重点说明Doris类型相关配置操作的不同之处，其他操作可参考上小节。

1. 数据源变化设置中，以Doris为目的端的配置只支持“数据源删除正在同步的表”设置和“同步的表字段被删除”设置，不支持“同步的表发现新增字段”设置；
2. 写入设置中，建表方式支持自动建表/选择已有表。自动建表情况下，表的类型支持2种：Duplicate明细表和Unique唯一表。

EDAP类型

选择数据集成目的端的类型为EDAP时，相关配置操作大部分与HIVE的配置操作类似，本小节重点说明EDAP类型相关配置操作的不同之处，其他操作HIVE类型。

1. 在设置同步来源与目标时，选择目的端类型为hive，需指定数据库与数据表的存储路径，存储路径只有在写入设置中选择自动建表才会生效。

前置检查

完成任务配置后，转到离线任务创建列表并选择【前置检查】。单击开始检查以验证任务配置是否可以正常运行。

表十四 检查结果说明

检查结果	结果说明
通过	如果检查通过，则任务可以正常运行。
失败	如果检查失败，检查内容会列出具体的错误原因。您需要根据这些错误修改任务配置并再次执行检查，直至成功通过。

任务运维

任务操作

任务运维列表显示任务名称、类型、来源和目标等详细信息。您可以运行、暂停或删除任务。点击运行，任务即开始运行，一次性任务在一次运行后结束；而定时任务则继续，刚创建的任务处于“已发布”状态，可以根据用户需求更改为“正在运行”或“暂停”状态。如果任务失败，则任务状态显示“失败”；如果运行成功，任务状态为“成功”。

任务概况

点击任务名称，可具体查看任务概况。概况内容包括任务的具体详细信息，且在监控界面右上角的编辑可编辑作业配置。

注：更新作业配置时任务的状态需要提前“暂停”，失败或者正在运行等状态的任务无法编辑。

表十五 任务概况说明

任务概况名称	描述
概览	概览可看见任务目前正在执行情况，包含读写数据量等，并且可以单独重试失败的单表任务和查看操作日志。
集成基本信息	在集成基本信息部分，您可以详细的查看创建任务时设置的基本信息。包括源端设置、目标端设置、任务配置等。
读取设置和写入设置	点击读取设置或写入设置，然后从左侧选择库或表，即可查看任务创建过程中设置的过滤器配置。

实时同步

实时同步即将源端数据实时同步至目标端，确保目标数据持续更新以反映源端的变化。此过程对于需要最新数据的应用程序至关重要，例如实时分析和运营监控等。本文主要介绍了实时同步的基础条件、进入实时同步的步骤、操作步骤和任务配置等内容，旨在帮助读者理解和掌握实时同步的功能和操作方法。

基础条件

资源配置

资源配置在EDAP首页侧边栏系统中心>资源管理中，点击**+添加资源配置**需要的数据源。

源端数据源配置说明

目前，数据集成源端类型支持 MySQL、SQL Server、Oracle、Hana、PostgreSQL 版本，下面是关于版本和源端类型具体说明。

表一 目前数据集成源端类型支持的版本

数据源	支持版本
MySQL	支持MySQL5.5, 5.6, 5.7, 8.0
SQLServer	SQLServer2016, 2014, 2012
Oracle	支持Oracle11, 12.1, 19
PostgreSQL	支持PostgreSQL8.2及以上

MySQL

1.开启mysql binlog功能 判断mysql是否开启binlog，并且binlog_format配置为ROW

执行：

```
##### OFF：关闭， ON：开启
SHOW VARIABLES LIKE 'log_bin';
#####
show variables like '%binlog_format%';
```

- 如果log_bin是ON，且则表明mysql已开启binlog；binlog_format值需为ROW
- 如果前置配置不满足按照2步骤配置

2.开启cdc配置

```
##### 第一种方式:
##### 开启binlog日志
log_bin=ON
##### binlog日志的基本文件名
log_bin_basename=/var/lib/mysql/mysql-bin
##### binlog文件的索引文件，管理所有binlog文件
log_bin_index=/var/lib/mysql/mysql-bin.index
##### 配置serverid
server-id=1

##### 第二种方式: #此一行等同于上面log_bin三行
log-bin=/var/lib/mysql/mysql-bin
##### 配置serverid
server-id=1

##### binlog_format默认值是ROW，如果配置文件中修改成其他值，需要改回ROW
binlog_format = ROW
```

具体的log-bin文件位置根据需要进行修改

配置完毕后，重启mysql

再次执行SHOW VARIABLES LIKE 'log_bin';查看是否开启。

3.用户需要的权限

表二 对应权限说明

权限	权限说明
SELECT	用于从数据库中读取数据，在snapshot阶段使用
RELOAD	用于执行flush语句来刷新数据库内部缓存，在snapshot阶段使用
SHOW DATABASES	用于显示database名字，在snapshot阶段使用
REPLICATION SLAVE	允许connector连接并读取mysql binlog
REPLICATION CLIENT	运行connector使用SHOW MASTER STATUS，SHOW SLAVE STATUS，SHOW BINARY LOGS命令

SQL Server

1.开启数据库cdc

```
-- 选择要开启cdc的数据库
USE ${database}
GO
-- 执行procedure
EXEC sys.sp_cdc_enable_db
GO
```

开启数据表cdc

前提条件：

- 数据库开启了cdc
- sql server agent处于运行中
- 当前用户属于db_owner

```
USE ${database}
GO

EXEC sys.sp_cdc_enable_table
@source_schema = N'${schema}',
@source_name  = N'${tableName}',
@role_name    = N'${roleName}',
@filegroup_name = N'${filegroupName}',
@supports_net_changes = 0
GO
```

- 注意：1. roleName表示要给予select权限的用户，如果设置为Null表示只允许在sysadmin或db_owner的角色有完全权限去读取cdc数据
2. cdc文件的文件组名最好不要和源表的文件组名一致

2.检查用户权限

表三 SQL Server权限说明

数据库级别的角色名称	说明
db_owner	db_owner固定数据库角色的成员可以执行数据库的所有配置和维护活动，还可以删除数据库。
db_securityadmin	db_securityadmin固定数据库角色的成员可以修改角色成员身份和管理权限。向此角色中添加主体可能会导致意外的权限升级。
db_accessadmin	db_accessadmin固定数据库角色的成员可以为Windows登录名、Windows组和SQLServer登录名添加或删除数据库访问权限
db_backupoperator	db_backupoperator固定数据库角色的成员可以备份数据库。
db_ddladmin	db_ddladmin固定数据库角色的成员可以在数据库中运行任何数据定义语言(DDL)命令。
db_datawriter	db_datawriter固定数据库角色的成员可以在所有用户表中添加、删除或更改数据。
db_datareader	db_datareader固定数据库角色的成员可以从所有用户表中读取所有数据。
db_denydatawriter	db_denydatawriter固定数据库角色的成员不能添加、修改或删除数据库内用户表中的任何数据。
db_denydatareader	db_denydatareader固定数据库角色的成员不能读取数据库内用户表中的任何数据。

您可以通过以下命令检查：

```
USE ${database};通过
GO
EXEC sys.sp_cdc_help_change_data_capture
GO
```

如果返回结果为空，说明该用户没有相应权限。

Oracle

1.开启数据表cdc

开启oracle logminer

```
ORACLE_SID=ORACLCDB dbz_oracle sqlplus /nolog

-- 配置归档日志大小和地址，重启oracle
CONNECT sys/top_secret AS SYSDBA
alter system set db_recovery_file_dest_size = 10G;
alter system set db_recovery_file_dest = '/opt/oracle/oradata/recovery_area' scope=spfile;
shutdown immediate
startup mount
-- 开启归档日志
alter database archivelog;
alter database open;

-- 查看归档日志，如果打印“Database log mode: Archive Mode”，则表明文件归档功能已经开启
archive log list

exit;
```

开启logminer的时候需要重新启动

开启supplemental logging（辅助 logminer）

```
-- 给表开启增强日志功能
ALTER TABLE FLINKUSER.CUSTOMERS ADD SUPPLEMENTAL LOG DATA(ALL)COLUMNS;

-- 开给数据库开启增强日志功能
ALTER DATABASE ADD SUPPLEMENTAL LOG DATA;
```

2.用户权限

表四 权限和权限说明

权限	权限说明	补充说明
CREATE SESSION	用于连接oracle	
SET CONTAINER	让connector可以在pluggable databases中进行切换，仅限于开启了CDB（container database support）模式	版本大于oracle11g才需要该权限
SELECT ON V_\$DATABASE	读取V\$DATABASE表（一种视图）	
SELECT ANY TABLE	允许读取各个表	
SELECT_CATALOG_ROLE	允许读取data dictionary，主要用于LogMiner阶段	
EXECUTE_CATALOG_ROLE	允许把data dictionary写入redo log中，主要用于监控schema变化	
SELECT ANY TRANSACTION	允许在snapshot阶段，对每个事务进行flashback snapshot。当FLASHBACK ANY TABLE开启后，该权限也需要开启	
LOGMINING	在新版本的oracle中，给予访问oracle的LogMiner和相应包的权限	版本大于oracle11g才需要该权限
CREATE TABLE	允许connector在default namespace中创建flush table，flush table允许connector显示得把LGWR缓存写入到磁盘中	
ALTER ANY TABLE	修改任何表	
CREATE SEQUENCE	创建序列	
EXECUTE ON DBMS_LOGMNR	允许connector在DBMS_LOGMNR package中运行方法，如果要和Oracle LogMiner进行交互，这个权限是必须的	
SELECT ON V\$LOG TO \${user} SELECT ON V\$LOGHISTORY TO \${user} SELECT ON V\$LOGMNRLOGS TO \${user} SELECT ON V\$LOGMNRCONTENTS TO \${user} SELECT ON V\$LOGMNRPARAMETERS TO \${user} SELECT ON V\$LOGFILE TO \${user} SELECT ON V\$ARCHIVED_LOG TO \${user} SELECT ON V\$ARCHIVEDEST_STATUS TO \${user} SELECT ON V\$TRANSACTION TO \${user}	允许connector读取这些表，主要是让connector读取oracle的redo和archive日志和当前事务状态。	

PostgreSQL

1.开启cdc

编辑postgresql.conf使用replication slots方式针对pg的wals日志实现logical decoding

```
##### Use logical decoding with the write-ahead log.
wal_level=logical
##### maximum of one separate process for processing WAL changes.
max_wal_senders=10
##### Allow a maximum of one replication slot to be created for streaming WAL changes.
##### 每个cdc_job会创建一个独立的slot，该参数需要根据实际情况进行设计，任务失效时会自动drop
max_replication_slots=10
```

以上三个配置修改后需要重启pg

检查方式：

```
show wal_level; #返回logical则正确
show server_version; #查看版本
```

2.用户权限

表五 权限和权限说明

权限	权限说明	检查方式
REPLICATION	用户需要有RELICATION权限 新建用户：CREATE USER cdc_user WITH PASSWORD '****' REPLICATION; 增加权限：alter user cdc_user REPLICATION;	select rolreplication from PG_ROLES where rolname='cdc_user' 返回结果是t有权限，f无权限
USAGE	用户需要schema的USAGE权限 grant USAGE on SCHEMA di_test to cdc_user;	select * from pg_namespace where nspname='di_test'; 返回的nspacl字段中是否包含cdc_user，cdc_userd的权限是否包含U
SELECT	用户需要对同步表的SELECT权限 grant select on TABLE di_test.company to cdc_user;	SELECT * FROM information_schema.table_privileges WHERE GRANTEE='cdc_user' and privilege_type='SELECT' and table in ('t1', t2) ; 会返回cdc_user有权限的表列表

进入实时同步

创建实时同步任务的入口点位于EDAP首页数据加工>我的项目中数据集成模块内。要启动创建过程，您需要进入数据集成模块页面。

启动创建过程：

- 启动方式：

进入数据集成模块页面。

- 创建实时同步任务入口

数据加工>我的项目中数据集成模块内。

1.如没有项目，需先新建项目。

点击我的项目>【新建项目】，填入您的项目相关属性后点击提交项目即创建成功。

项目创建步骤：

- 进入界面：点击我的项目
- 开始新建：选择【新建项目】
- 填写信息：填入项目相关属性
- 完成创建：点击提交项目

注：项目中文名和英文名是必填项，且项目英文名创建后不允许修改。

2.已有项目，进入实时同步界面

通过选择导航栏中的【数据加工】，然后选择“我的项目”来访问实时同步界面。选择对应项目后单击项目名称，查找【数据集成】使用实时同步功能。

访问实时同步界面：

- 通过选择导航栏中的【数据加工】。
- 然后选择“我的项目”。

使用实时同步功能：

- 单击项目名称选择对应项目。
- 查找【数据集成】使用实时同步功能。

操作步骤

任务创建 --> 任务配置 --> 任务运维

任务创建

在开始特定数据集成任务的配置之前，必须在平台内创建任务。无论所涉及的数据源或目标的类型如何，都需要执行此预备步骤。

以下是两种创建任务的概述：

- 进入实时同步界面，单击+创建任务图标。
- 在实时同步界面中，点击列表页面中央的马上创建。

以上两种方式创建任务后，跳转任务配置界面。

图一 创建任务界面



任务配置

目标端数据源类型为Doris、EDAP，配置一个数据集成任务分为“设置同步来源与目标”、“读取设置”、“批量设置”、“写入设置”四个步骤。其中，“设置同步来源与目标”、“读取设置”属于公共配置，“写入设置”在目标端上有所不同。

本章节首先以EDAP为例分析公共配置，然后按照目标端类型分别介绍“写入设置”的操作说明。

设置同步来源与目标

基本信息

配置任务基本名称

注意命名规范要求：只能包含英文字母、数字、中划线和下划线，且以英文字母开头，不大于64个字符），同一个项目中的同步任务名不能重复

源端设置

在源端设置配置数据类型和选择已配置好的数据源，数据源配置在EDAP主页配置栏，详见基础条件-源端配置说明。

目标端设置

在目标端设置目标类型、数据源和数据库，目前类型支持Doris、EDAP 数据湖。

迁移类型

迁移类型针对所有目标端类型，目前可支持的迁移类型：全量及增量迁移

其中，目标端如果为EDAP类型，需增加资源配置选项，可以选择计算资源。

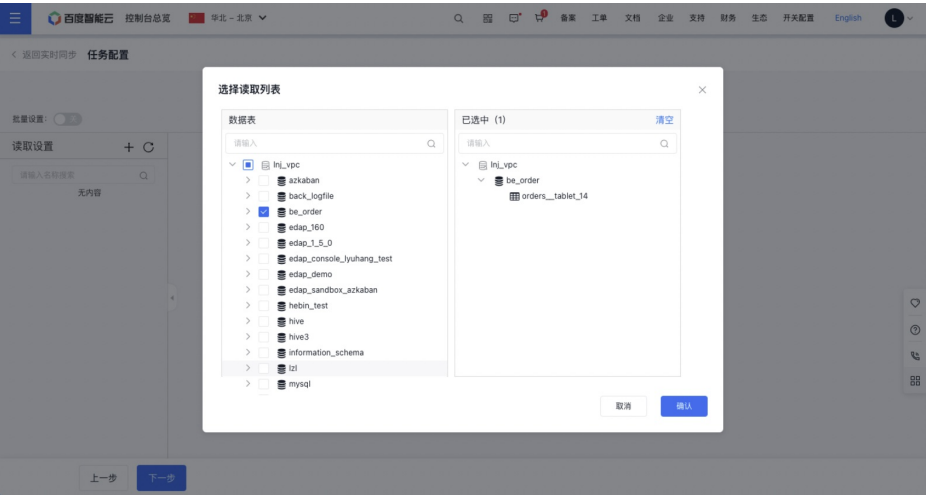
资源配置

资源配置需要在 EDAP 界面上配置资源，选择【资源管理】，找到合适的计算资源，点击计算资源右侧的【绑定项目】来绑定到该 EDAP 项目中。

读取设置

1.点击读取设置旁边“+”按钮，在读取列表中可以在单个同步任务中批量选择多个数据库和表进行读取，还可以在编辑阶段向任务添加更多表。

图一 选择读取列表界面



2.读取设置中，左上方默认为关闭“批量设置”的状态，点击左边的表，针对每个单表进行配置，通过下方的where语句过滤单表的数据，目前仅支持 logicStartTime 宏。

注：LST宏表达式格式： \${logicalStartTime(yyyy.?MM.?dd.? HH.?mm.?ss.?(SSS)?{,[+]?[0-9]*[yMdhms]})? }

批量设置

可批量设置所有表的部分写入设置。可打开左上方“批量设置”按钮，通过where语句批量过滤不需要的数据，在执行中是将此where语句应用在每个单表上，分别过滤每个所选择的表中的数据。在读取设置结束后，进入批量设置。填入副本数和需要的表名设置后点击下一步即设置成功。

批量设置仅在目标端建表方式为“自动建表”时生效

EDAP类型

EDAP批量设置界面，可批量配置表的存储格式、表名的前后缀、新增字段、静态分区。

目标表设置

配置项	默认值	说明
数据糊格式	无	必选，枚举值：无、hudi
存储格式	Parquet	必选无：Parquet、ORChudi：Parquet
hudi表类型	默认无此项	hudi表枚举值：MOR、COW，默认MOR
hudi表主键	默认无此项	字段名允许修改， 提示说明：可指定目标端主键名称，若不指定目标端主键，默认生成EDAP_UUID字段作为表主键，字段值为随机UUID字符串
表名设置	无	无前后缀、增加前后缀、增加前缀、增加后缀

新增字段设置

可以批量增加目标端表字段，同时支持固定值及时间宏函数变量：

配置项	默认值	说明
字段名称	无	新增字段名称，若新增字段名称与表已有字段冲突，则单表配置或任务完成时进行右上角报错提示：新增字段与表已有字段冲突！
字段类型	String	枚举值：String、INT、BIGINT、DOUBLE、DECIMAL
默认值	无	可填写固定值，如，1可填写时间宏函数
操作	无	增加/删除

分区设置

配置项	默认值	说明
字段名称	无	新增字段名称
字段类型	String	枚举值：String、INT、BIGINT、DOUBLE、DECIMAL
分区值	无	填写分区值
操作	无	增加/删除

Doris类型

Doris批量设置界面，可批量配置表的存储格式、表名的前后缀、新增字段

目标表设置

配置项	默认值	说明
Doris表格式	Duplicate明细表	下拉，必选枚举值：Duplicate明细表、Unique唯一表实时仅支持唯一表
表名设置	无	无前后缀、增加前后缀、增加前缀、增加后缀
副本数	3	默认副本数为3。如果 BE 节点数量小于3，则需指定副本数小于等于 BE 节点数量。
后缀	无	输入后缀值

写入设置

EDAP类型

表类型、存储方式等均在上一步做了批量设置，此步可在左侧目录树中指定表，进行建表方式、表达式、分区等个性化设置，接下来展开说明：

1. EDAP支持“自动建表”和“选择已有表”两种方式。

当【建表方式】选择“自动建表”时，数据集成会在目标端自动创建对应表，字段名称和类型默认与数据源端一致，页面中的“箭头”表示数据源和目的端表的字段映射关系。用户可根据需要更改目的端表名、字段名称、字段类型、默认值。

此外，用户可在目的表点击每个字段最右侧的删除符号将此字段删除，表示在数据集成过程中过滤此字段，不同步源端的此字段。然后，在右侧列表最下方可以添加普通字段或分区字段；

2. EDAP表类型支持hudi表，hudi表类型默认为MOR

表六 hudi表说明

配置项	默认值	说明
EDAP表类型	hudi	枚举值：hudi
存储格式	Parquet	Hudi表：Parquet
Hudi表类型	MOR	hudi表枚举值：MOR、COW，默认MOR

3.EDAP类型支持批量配置，点击批量配置按钮弹出批量设置弹窗，详见[批量设置](#)

表七 批量设置字段属性说明

配置项	默认值	说明
数据湖格式	hudi	枚举值：hudi
存储格式	Parquet	hudi：Parquet
hudi表类型	MOR	hudi表枚举值：MOR、COW，默认MOR
hudi表主键	EDAP_UUID	字段名允许修改，提示说明：若不指定目标端主键，Hudi表默认生成EDAP_UUID字段作为表主键，字段值为随机UUID字符串

4.当建表方式设置为【选择现有表】时，需要指定目标表名称。您可以使用“同名映射”或“同行映射”来批量确定数据源端和目的端表的字段映射方式。

此外，还可以在可视化界面中绘制手动连接以自定义字段映射。

5.打开表达式设置时，

表八 目前支持的transform函数列表

类别	中文名称	英文名称及参数	返回值	详细功能描述
数学函数	绝对值	abs()	T	T是数值类型，包括int、long、float、double、decimal
字符串函数	连接字符串	concat(String str1, String str2)	String	连接str1和str2两个字符串
	去掉字符串前后的空格	trim()		去掉操作字符串前后的空格
	翻转字符串	reverse()		翻转操作字符串
	返回大写字符	toupper()		返回操作字符串的大写形式
	返回小写字符	tolower()		返回操作字符串的小写形式
	字符串替换	replace(String old, String new)		搜索操作字符串中的old子串，如果能够找到则使用new子串来替换
功能函数	条件过滤	filter(String operator, String criteria)	boolean	operator是支持的运算符，criteria是过滤值，其支持的运算符如表2所示

表九 条件过滤的支持运算符类型关系如下

序号	操作符	boolean	int	long	float	double	bytes	string	date	decimal	timestamp
1	=	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	!=	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	!=	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	>		✓	✓	✓	✓		✓	✓	✓	✓
5	>=		✓	✓	✓	✓		✓	✓	✓	✓
6	<		✓	✓	✓	✓		✓	✓	✓	✓
7	<=		✓	✓	✓	✓		✓	✓	✓	✓
8	contains							✓			
9	notContain							✓			
10	startsWith							✓			
11	endsWith							✓			
12	notStartWith							✓			
13	notEndWith							✓			
14	matchesRegex							✓			
15	notMatchRegex							✓			

Doris类型

选择数据集成目的端的类型为Doris时，相关配置操作大部分与EDAP的配置操作类似，本小节重点说明Doris类型相关配置操作的不同之处，其他操作可参考上小节。

1. 【数据源变化设置】中，以Doris为目的端的配置只支持“数据源删除正在同步的表”设置和“同步的表字段被删除”设置，不支持“同步的表发现新增字段”设置；
2. 【写入设置】中，建表方式支持自动建表/选择已有表。自动建表情况下，表的类型支持2种：Duplicate明细表和Unique唯一表。

前置检查

完成任务配置后，转到实时任务创建列表并选择【前置检查】。单击开始检查以验证任务配置是否可以正常运行。

检查结果：

- 通过：如果检查通过，则任务可以正常运行。
- 失败：如果检查失败，检查内容会列出具体的错误原因。您需要根据这些错误修改任务配置并再次执行检查，直至成功通过。

任务运维

任务操作

任务运维列表显示任务名称、类型、来源和目标等详细信息。您可以运行、暂停或删除任务。点击运行，任务即开始运行，一次性任务在一次运行后结束；而定时任务则继续，刚创建的任务处于“已发布”状态，可以根据用户需求更改为“正在运行”或“暂停”状态。如果任务失败，则任务状态显示“失败”；如果运行成功，任务状态为“成功”。

任务概况

点击任务名称，可具体查看任务概况。概况内容包括任务的具体详细信息，且在监控界面右上角的【编辑】可编辑作业配置。

注：更新作业配置时任务的状态需要提前“暂停”，失败或者正在运行等状态的任务无法编辑。

表十 任务概况说明

任务概况内容	说明
概览	概览可看见任务目前正在执行情况，包含读写数据量等，并且可以单独重试失败的单表任务和查看操作日志。
集成基本信息	在“集成基本信息”部分，您可以详细的查看创建任务时设置的基本信息。包括源端设置、目标端设置、任务配置等。
读取设置和写入设置	单击读取设置或写入设置，然后从左侧选择库或表，即可查看任务创建过程中设置的过滤器配置。

数据开发

🔗 离线开发

🔗 概述

概述

EDAP支持离线作业开发模式，主要面向离线数据处理场景，同时提供了可视化作业开发、脚本作业开发、作业组开发三种开发方式，满足不同客户开发需求。

进入项目

在【我的项目】中的项目列表点击具体某个项目，进入项目内部。

离线开发

项目内，选择【离线开发】，进入离线作业的开发和编排界面。离线开发分为三类：可视化作业开发、脚本作业开发、作业组开发。

- 可视化作业开发：使用拖拽的形式，降低用户开发、使用门槛，快速实现离线任务的加工处理；
- 脚本作业开发：以脚本的形式进行数据的开发，目前支持的脚本类型包括：HIVESQL、JDBCSQL、Spark SQL、PySpark、Spark Scala、Shell、Python等脚本类型；
- 作业组开发：将可视化作业开发、脚本作业开发等开发完成的作业，或者用户线下创建好的shell、python、spark等作业进行引用。实现作业组的可视化编排，并配置调度，实现离线作业组的周期性运行。

作业批量导入导出

仅支持顶级【任务开发】目录下的文件夹/作业的导入导出（涵盖可视化、脚本、作业组）。

导入

点击【导入】按钮，页面弹窗，选择用户本地文件，支持单文件、且zip类型文件导入，如果选择文件非zip类型，会提示：「请选择zip压缩包」；

如果提示“导入文件解析失败，请检查json文件”，则可能是单个json文件或zip压缩包内某个json文件不满足edap标准。

导入完成后，弹窗展示**导入失败**的文件名。

注意：同名脚本/作业进行覆盖。

导出

点击【导出】按钮，弹窗提示“是否确认导出该文件夹下所有作业？”

点击确定后，导出的文件会被打包为一个zip压缩包，压缩包名为：**project_categ_datetime.zip**（datetime为时间戳，到秒级）。

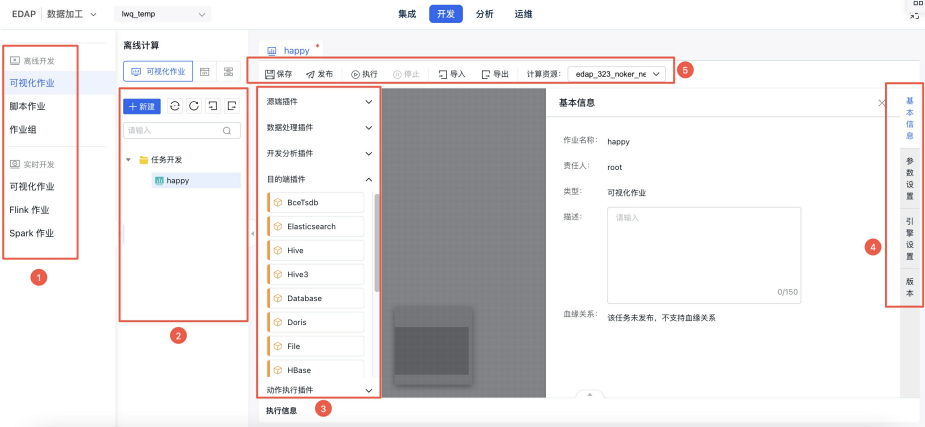
压缩包内包含文件夹及文件夹内所有的作业，作业格式仍为json格式，json包含内容与单作业导出内容保持一致（仅包含作业元信息，不包含引用的jar、文件）。

可视化作业

概述

EDAP支持离线作业开发模式，主要面向离线数据处理场景，同时提供了可视化作业开发、脚本作业开发、作业组开发三种开发方式，满足不同客户开发需求。本章节将详细的为您介绍可视化作业开发流程。

图一 离线开发可视化作业概览



表一 可视化作业概览描述

功能区域	描述
选择作业	区域1用于快速切换不同作业
任务开发列表	区域2是任务开发列表。 系统默认作业文件是任务开发。在此区域可以新建作业或新建文件夹。支持按照文件名称搜索作业，并且可以定位作业。同时支持导入/导出文件（文件大小不超过200M的zip文件）。
插件列表	区域3展示作业对应插件或者开发面板。
作业配置	区域4是作业的不同配置，包括基本设置、参数设置、引擎设置和版本
作业操作	区域5可对作业进行不同操作

名词解释

针对文档中频繁出现的名词，下面将进行详细描述：

表二 名词解释

名词	说明
作业	用户通过离线开发、实时开发功能编写的可视化、脚本、Jar包等统称为作业
作业组	离线开发下多个作业的组合被称为作业组
任务	作业发布后形成可被手动或定时调度的任务
实例	任务每次执行会形成一个实例，实例状态分为成功、失败、暂停等

新增文件夹 可视化作业支持按照文件夹进行分组管理，新建作业前可以选择新增文件夹。

1. 进入可视化作业需要新建项目。没有项目选择侧边导航或者快速访问中**数据加工>开发（顶部导航）>离线开发>可视化作业**后，系统自动跳转新建项目，完善项目信息即可创建成功。
2. 系统默认项目文件夹。可以直接选择侧边导航或者快速访问中**数据加工>开发（顶部导航）>离线开发>可视化作业**后进入可视化作业。
3. 单击**新建**按钮，填写弹窗中配置项后点击确定即可创建成功。

新增作业

文件夹新增完毕，单击**新建**按钮选择**新增作业**。填写新建作业弹窗中配置项点击确定后新增完成。支持选择存放文件夹。在列表中显示任务名称，并且在插件中展示可视化插件表示新增作业成功。

可视化作业插件

离线可视化开发作业支持五种类型插件。源端插件、数据处理组件、开发分析组件、目标端插件以及动作执行插件。用户可以将插件节点拖拽到可视化开发面板中，进行可视化作业的开发，下面对具体插件详细描述。

- 1.选择需要的插件单击或者将其拖入可视化开发面板中，然后单击插件系统回跳出关于组件的配置界面。填写相关配置项单击确定即配置成功。
- 2.插件间支持相互连接。

表三 插件说明

插件类型	插件说明
源端插件	作为整个可视化作业的数据输入端，源端插件只能作为开始节点，下游可以连接数据处理组件、开发分析组件或者直接连接目标端插件
数据处理组件	对上游源端插件的数据进行数据处理，用户可以使用脚本类插件对数据进行处理
数据开发组件	对上游源端插件的数据进行数据开发，用户可以使用SQL抽象插件进行数据处理
目标端插件	将上游处理好的数据写入到目标端插件。且目标端插件需要作为可视化作业的最后节点
动作执行插件	目前动作执行插件主要支持jdbcAction插件，支持对不同类型数据进行SQL执行相关操作，如truncate、delete等

可视化作业配置

可视化作业配置位于整个工作台的右边，单击不同配置项可以根据需要对作业进行配置。不同设置之间可进行点击切换。

表四 配置说明

作业配置名称	配置说明
基本信息	单击右侧配置信息-基本信息，可展开查看可视化作业的基本信息，包括作业名称、责任人、作业类型和描述。支持对描述进行修改
参数设置	单击参数设置，填写弹窗内参数内配置项。 - 参数名称：自定义参数名称 - 参数值：填写参数值，支持固定值及时间宏函数 - 添加/删除：添加一行参数或删除一行参数
引擎设置	1.单击引擎设置，可以根据计算需求选择mapreduce或spark引擎，并可以针对不同引擎设置Driver和Executor的CPU及内存大小。 2.引擎也支持参数设置，具体描述参照参数设置即可。
执行设置	作业独立运行时生效，被作业组引用时忽略该设置。打开失败重试按钮，填写重试和时间间隔参数。支持配置重试次数（最小1次，最大100次）时间间隔（时间间隔范围：最小100ms，最大为一天）。
版本	在作业发布后，可在此处查看历史版本、执行版本回滚操作。

可视化作业操作

在作业运行过程中，可对作业进行不同操作，包括但不限于保存、发布和执行。

表五 可视化作业操作说明

操作项	操作说明
作业保存	单击面板上方保存按钮
发布作业	单击面板上方发布按钮，可将当前作业提交为新的版本。发布完成后可在面板右侧版本中查看历史版本、执行版本回滚操作，同时将当前作业发布到任务运维中心。
作业停止	单击面板上方停止按钮，可停止当前作业。
执行作业	单击页面上方执行，进行可视化作业测试运行，且在执行信息中弹出执行日志信息。
作业导出/入	可视化作业支持导入导出，单击导出，将当前作业导出为json文件
计算资源	点击页面上方计算资源弹出作业资源设置框，对脚本作业进行资源设置。

脚本作业

在线进行离线大数据脚本任务的统一协同开发，支持多种类型脚本的在线开发、编辑、测试。

新建脚本作业

1.任务开发列表中新建按钮，单击新建按钮选择新建作业，或者在工作台上方单击新建脚本作业。填写新建作业的相关配置项即创建作业成功。

2.配置项说明

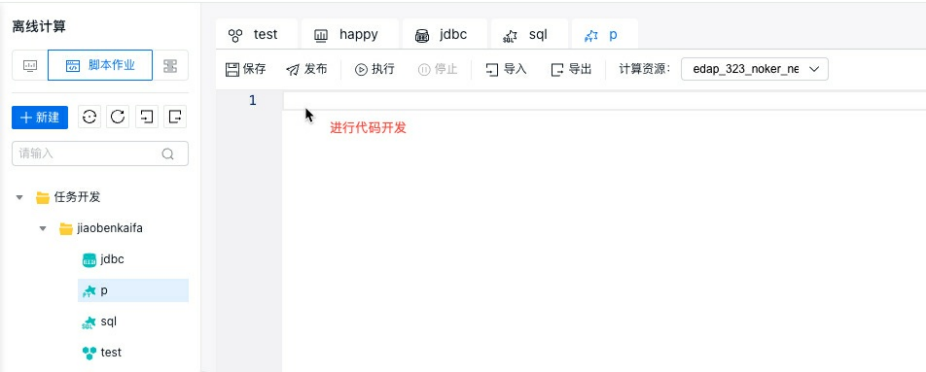
表一 配置项说明

配置项名称	配置项说明
作业名称	输入作业名称，用于区分不同作业。中文、英文、数字、中划线、下划线，不能以中划线、下划线开头，不超过50个字符。
脚本文件类型	选择脚本文件类型，不同类型的脚本支持不同的作业配置。以下是脚本文件类型配置项详情： - HIVESQL:基本信息、参数设置、版本 - JDBCSQL：基本信息、参数设置、版本 - Spark SQL：基本信息、参数设置、引擎设置、版本 - PySpark：基本信息、参数设置、引擎设置、版本 - Spark Scala：基本信息、参数设置、引擎设置、版本 - Shell：基本信息、脚本设置、版本 - Python：基本信息、脚本设置、版本
存放文件夹	默认为任务开发文件夹。也可以单击新建选择新建文件夹，填写相关配置项即可创建成功。
描述	输入对作业的描述

脚本作业开发编辑

新建脚本作业完成后，可在编辑框中进行代码编辑开发。

图一 代码开发界面



脚本作业开发配置

不同类型的脚本支持不同的作业配置，单击编辑框右侧配置项名称，可查看配置项详情。 表二 配置项说明

配置项名称	配置项说明
基本信息	显示脚本作业的基本信息，并能够进行描述修改。
参数设置	针对HIVESQL、JDBCSQL脚本类型，支持参数设置填写。 单击右侧参数设置，弹出脚本作业的参数设置。设置的参数可以在脚本作业中进行引用。 - 参数名称:自定义参数名称 - 参数值：填写参数值，支持固定值及时间宏函数 - 参数引用：以\${参数名称}方式在脚本作业中进行引用
执行设置	作业独立运行时生效，被作业组引用时忽略该设置。打开失败重试按钮，填写重试和时间间隔参数。支持配置重试次数（最小1次，最大100次）时间间隔（时间间隔范围：最小100ms，最大为一天）。
版本	在作业发布后，可在此处查看历史版本、执行版本回滚操作。

脚本作业开发操作

表三 脚本开发操作说明

操作项名称	操作项说明
作业保存	单击面板上方保存按钮可保存当前作业，直到右上角显示作业保存成功。
作业发布	单击面板上方发布，可将当前作业提交为一新的版本，并可在右侧配置中选择版本浏览历史版本、执行版本回滚操作，同时将当前作业发布到任务运维中心。
作业执行	单击页面上方执行，进行脚本作业测试运行，且在执行信息中弹出执行日志信息。
作业导出/导入	- 导出：单击导出按钮，弹窗提示“是否确认导出该作业？”后单击确认，会保存json文件至用户本地。文件名为：当前作业名.json。文件包含的内容：当前作业名、脚本内容、基本信息、参数配置。 - 导入：单击导入按钮，选择用户本地文件，只支持json类型文件导入，若用户选择非json类型文件，则提示“请选择json文件”。导入的脚本作业默认覆盖原有工作台的草稿。若导入的脚本作业类型与现有的脚本类型不一致，则提示“导入作业的脚本类型与所建作业类型不一致！”
计算资源	点击页面上方计算资源弹出作业资源设置框，对脚本作业进行资源设置。
数据源	根据数据源类型展示不同的数据源，单击下拉选择。
数据源类型	单击下拉框选择不同的数据源类型。

作业组

新增文件夹/作业

作业组支持按照文件夹进行分组管理，新建作业前可以选择新增文件夹。

1. 选择侧边导航或者快速访问中**数据加工>开发（顶部导航）>作业组**后进入作业组。
2. 新建文件夹。单击**新建**按钮选择**新建文件夹**，填写相关配置项即可新建成功。
3. 新建作业组。单击**新建**按钮选择**新建作业**，文件夹支持自行选择存放，填写完毕点击确定新建成功。

作业组配置

作业组的节点目前支持七种类型：可视化作业、脚本作业、作业组、数据治理、数据集成、数据安全、AI模型。具体类型节点包括：

表一 节点类型说明

节点类型	具体节点
可视化作业	可视化
脚本作业	Shell、Python、Spark、SparkSQL、HIVESQL、JDBCSQL、PySpark、Scala
作业组	作业组
数据治理	Quality
数据集成	离线同步
数据安全	数据脱敏
AI模型	模型训练、批量预测

执行设置

单击具体节点用户可以对节点进行执行相关设置。

表二 执行设置说明

配置项名称	配置项说明
最长执行时间	作业运行的时间上限，当超过这个时间上限后，作业失败。
失败重试	作业执行失败后重试的次数，以及每次重试之间的时间间隔，配置重试之后，作业节点会在配置时间间隔后，再次运行，直到超过重试次数之后，作业节点才算真正失败。
设置依赖节点	本节点依赖其他任务节点。只有当被依赖节点执行成功之后，此节点才会被执行。
优先级设置	作业节点的优先级。平台一共分为5级优先级。1-5，数字越高优先级越高。当在同一层级的任务，优先执行优先级高的作业。

表三 依赖节点参数配置

表达式	说明	所依赖节点生效范围	说明
节点无关语义(5种)			
one_success	对于本节点直接依赖的上游节点，只要有一个节点执行成功，就能满足条件	与该节点具有直接依赖关系的节点，表达式中的 jobName 失效，但是需要用"*"作为占位符填充进表达式。	支持"=="
one_failed	对于本节点直接依赖的上游节点，只要有一个节点执行失败，就能满足条件。	与该节点具有直接依赖关系的节点，表达式中的 jobName 失效，但是需要用"*"作为占位符填充进表达式。	支持"=="
all_success	对于本节点直接依赖的上游节点，所有节点都要执行成功，才能满足条件	与该节点具有直接依赖关系的节点，表达式中的 jobName 失效，但是需要用"*"作为占位符填充进表达式。	支持"=="
all_failed	对于本节点直接依赖的上游节点，所有节点都要执行失败，才能满足条件	与该节点具有直接依赖关系的节点，表达式中的 jobName 失效，但是需要用"*"作为占位符填充进表达式。	支持"=="
all_done	对于本节点直接依赖的上游节点，所有节点都要执行完成，才能满足条件。	与该节点具有直接依赖关系的节点，表达式中的 jobName 失效，但是需要用"*"作为占位符填充进表达式。	支持"=="
节点相关语义(2种)			
success	对于本节点直接依赖的上游节点，其中某个节点执行成功，即满足条件。	与该节点具有依赖关系(直接或间接)的节点，表达式中需准确指定 jobName。	支持"=="
failed	对于本节点直接依赖的上游节点，其中某个节点执行失败，即满足条件。	与该节点具有依赖关系(直接或间接)的节点，表达式中需准确指定 jobName。	支持"=="

作业组开发配置

作业组配置位于整个工作台的右边，单击不同配置项可以根据需要对作业进行配置。不同设置之间可进行点击切换。

表四 配置项说明

配置项名称	配置项说明
基本信息	显示作业组的基本信息，并能够进行描述修改。
参数设置	支持参数设置填写，如果作业组和作业设置了同名的参数，则作业中设置的参数优先级高于作业组中设置的参数。 单击右侧参数设置，弹出脚本作业的参数设置。设置的参数可以在作业组中的作业进行引用。 - 参数名称:自定义参数名称 - 参数值：填写参数值，支持固定值及时间宏函数 - 参数引用：以\${(参数名称)}方式在作业组中进行引用。如果作业组和作业设置了同名的参数，则作业中设置的参数优先级高于作业组中设置的参数。
版本	在作业发布后，可在此处查看历史版本、执行版本回滚操作。

表五 spark参数配置说明

参数名称	默认值	操作	是否必填
spark.driver.memory	1g	+ -	否
spark.driver.cores	1	+ -	否
spark.executor.instances	1	+ -	否
spark.executor.memory	1g	+ -	否
spark.executor.cores	1	+ -	否

作业组开发操作

在作业开发运行过程中，可对作业进行不同操作，包括但不限于保存、发布和执行。

表六 脚本开发操作说明

操作项名称	操作项说明
作业保存	单击面板上方保存按钮可保存当前作业，直到右上角显示作业保存成功。
作业发布	单击面板上方发布，可将当前作业提交为一新的版本，并可在右侧配置中选择版本浏览历史版本、执行版本回滚操作，同时将当前作业发布到任务运维中心。
作业执行	单击页面上方执行，进行作业组测试运行，且在执行信息中弹出执行日志信息。
作业导出/导入	- 导出：单击导出按钮，弹窗提示“是否确认导出该作业？”后单击确认，会保存json文件至用户本地。文件名为：当前作业名.json。文件包含的内容：当前作业名、脚本内容、基本信息、参数配置。 - 导入：单击导入按钮，选择用户本地文件，只支持json类型文件导入，若用户选择非json类型文件，则提示“请选择json文件”。导入的脚本作业默认覆盖原有工作台的草稿。若导入的脚本作业类型与现有的脚本类型不一致，则提示“导入作业的脚本类型与所建作业类型不一致！”。
重新加载	重新加载覆盖当前内容。

🔗 AI 模型

🔗 BML模型训练

BML 全功能 AI 开发平台是一个面向企业和个人开发者的机器学习集成开发环境,帮助用户更快的构建、训练和部署模型。EDAP为您提供BML节点,方便您通过算法节点载入BML workflow，实现数据在EDAP、BML间的周期性任务调度。

前提条件

- 授权EDAP作业执行的用户为BML的项目管理员，授权操作详见：<https://cloud.baidu.com/doc/BML/s/Y1q2akb8w>
- 配置AIBase访问API的权限，授权操作详见：<https://ai.baidu.com/ai-doc/REFERENCE/Lkru0zoz4>
- 在BML中作业建模下，已创建相关作业

步骤一：节点配置

- 进入EDAP
登陆EDAP控制台，进入数据加工点击左上角切换至目标项目

2. 进入离线开发下作业组开发

新建作业组，点击AI模型，新建**模型训练**节点

3. 配置模型训练相关参数

选择在BML中被授权的项目列表，训练类型选择**作业建模**，在**训练作业**中选择对应的作业，同时配置ak、sk、输入输出路径等参数 **步骤二：作业发布及调度** 点击编辑框左上角**发布作业**，该作业发布后即可在**运维功能**下对发布的**作业组**配置调度，进行定期执行 **步骤三：任务运维** 在数据加工下，可通过运维大屏，查看任务的运行状态，同时对失败的实例进行重新执行

BML批量预测

BML 全功能 AI 开发平台是一个面向企业和个人开发者的机器学习集成开发环境,帮助用户更快的构建、训练和部署模型。EDAP为您提供BML节点,方便您通过算法节点载入BML workflow，实现数据在EDAP、BML间的周期性任务调度。

前提条件

- 授权EDAP作业执行的用户为BML的项目管理员，授权操作详见：<https://cloud.baidu.com/doc/BML/s/Y1q2akb8w>
- 配置AIBase访问API的权限，授权操作详见：<https://ai.baidu.com/ai-doc/REFERENCE/Lkru0zoz4>
- 在BML中批量预测下，已创建相关作业

步骤一：节点配置

1. 进入EDAP

登陆EDAP控制台，进入**数据加工**点击左上角切换至目标项目

2. 进入离线开发下作业组开发

新建作业组，点击AI模型，新建**批量预测**节点

3. 配置批量预测相关参数

选择在BML中被授权的项目列表，在**作业名称**中选择对应的作业，同时配置ak、sk、输入输出路径等参数 **步骤二：作业发布及调度** 点击编辑框左上角**发布作业**，该作业发布后即可在**运维功能**下对发布的**作业组**配置调度，进行定期执行 **步骤三：任务运维** 在数据加工下，可通过运维大屏，查看任务的运行状态，同时对失败的实例进行重新执行

ModelBuilder批量预测

前提条件

需要当前账号拥有千帆ModelBuilder相关权限（权限分配可参考[账号创建与权限分配](#)）、EasyDAP数据湖管理与分析平台相关权限（权限分配可参考[多用户访问控制](#)）、BOS相关权限

- 需具有ModelBuilder以下任一权限：
 - 完全控制千帆大模型平台的权限：QianfanFullControlAccessPolicy
 - 运维操作千帆大模型平台预测服务的权限：QianfanServiceOperateAccessPolicy
- 需要具有EDAP以下任一权限：
 - EDAP管理员权限：EDAPFullControl
 - EDAP数据管理员权限：EDAPDataManage
- 需具有BOSFullControlAccessPolicy权限

节点配置

配置项	说明
节点名称	填写节点名称。中文、英文、数字、中划线、下划线，不能以中划线、下划线开头，不超过50个字符。
选择模型	选择预置模型或我的模型，系统会自动加载当前用户在ModelBuilder上拥有权限的模型列表，选择模型后进一步选择对应的模型版本。
输入数据路径	输入BOS路径。数据输入bos目录，BOS服务说明请参考BOS介绍，说明： (1) 建议bos目录下不要嵌套目录 (2) 数据尽可能均分到目录下的每个文件，单个任务文件数量100（含子目录） (3) 全部文件累计大小不超过300MB (4) 文件格式要求，示例请参考本文文件源数据格式说明： - 文件内容必须符合JsonLine格式要求，文件内 id唯一，query为请求数据 - 文件格式不强制要求是.jsonl，可以是.txt，但推荐.jsonl后缀名 - 示例输入文件路径：bos:/bucket-name/文件夹名称/ - 路径中支持填写时间宏函数\${logicalStartTime(yyyy-MM-dd HH:mm:ss:SSS)}，其中yyyy-MM-dd HH:mm:ss:SSS字符间的分隔符' '、':'、'-'以及保留具体某个或某几个时间，可以根据需要自行组织，例如：\${logicalStartTime(yyyyMMdd)}、\${logicalStartTime(dd)}，示例输入文件路径：bos:/bucket-name/文件夹名称/\${logicalStartTime(yyyyMMddHH)}
输出数据路径	输入BOS路径。结果输出bos目录，BOS服务说明请参考BOS介绍，说明： (1) 固定输出格式为jsonl (2) 示例输出文件：bos:/bucket-name/文件夹名称/ - 路径中支持填写时间宏函数，具体参考输入路径参数示例说明 - 输出文件数量与输入文件数量一一对应
参数设置	单击添加参数按钮，填写参数配置，可自定义请求ModelBuilder推理服务时的参数。 - 参数名称:自定义参数名称 - 参数值：填写参数值，支持固定值及时间宏函数 - 添加/删除：添加一行参数或删除一行参数

执行设置

1. 配置

打开最长执行时间按钮，填写执行时间。当超过这个时间上限后，任务状态为失败。

2. 执行依赖

打开失败重试按钮，填写相关参数。具体说明如下：

配置项	配置项说明
重试次数	最小1次，最大100次
时间间隔	最小1ms，最大为一天
设置依赖节点	本节点依赖其他任务节点。只有当被依赖节点执行成功之后，此节点才会被执行。
优先级设置	作业节点的优先级。平台一共分为5级优先级。1-5，数字越高优先级越高。当在同一层级的任务，优先执行优先级高的作业。

注意事项 任务数量限制：一个账号处于排队中（Queuing）、运行中（Running）、运行终止中（Stopping）状态的批量任务不能超过5个。否则提交任务会报错TaskRunningNumberExceedLimit；在任务调度过程中，可以通过在执行设置处配置重试次数解决该问题。

实时开发

Spark作业

新增文件夹/作业

Spark作业支持按照文件夹进行分组管理，新建作业前可以选择新增文件夹。

1. 选择侧边导航或者快速访问中数据加工>开发（顶部导航）>实时开发（侧边导航）>Spark作业后进入作业组。
2. 新建文件夹。单击新建按钮选择新建文件夹，填写相关配置项即可新建成功。
3. 新建作业。单击新建按钮选择新建作业，文件夹支持自行选择存放，填写完毕点击确定新建成功。新建Scala作业，呈现Scala代码编写面板，支持进行Scala代码的编写。新建Spark自定义作业需要对作业进行配置。

表一 新建作业配置项说明

配置项名称	配置项说明
作业类型	选择Spark Scala作业或Spark自定义作业。
执行引擎	Spark Streaming。
作业名称	只能包含英文字母、数字、中划线和下划线，且以英文字母开头，不大于50个字符。
存放文件夹	选择存放的文件夹位置。
描述	填写对作业的描述。

表二 Spark自定义作业参数配置

参数名称	说明
主程序包路径	选择主程序文件路径，如HDFS没有对应文件请上传文件，点击上传文件弹窗显示上传文件操作框。
主程序包名称	填写程序包名称。
主类名称	填写主类名称。
Spark Main参数	填写相关参数，Spark 主类参数，支持多参数填写。

Spark作业配置

在Flink作业面板右侧可以查看作业配置，配置间可点击切换。

表三 Spark配置说明

作业配置名称	配置说明
基本信息	单击右侧配置信息>基本信息，可展开查看可视化作业的基本信息，包括作业名称、责任人、作业类型和描述。支持对描述进行修改。
参数设置	单击参数设置，填写弹窗内参数内配置项。 - 参数名称：自定义参数名称 - 参数值：填写参数值，支持固定值及时间宏函数 - 添加/删除：添加一行参数或删除一行参数
引擎设置	1.单击引擎设置，可以根据计算需求选择不同引擎配置，可手动填写。 2.引擎也支持参数设置，具体描述参照参数设置即可。
版本	在作业发布后，可在此处查看历史版本、执行版本回滚操作。

表四 引擎设置说明

参数名称	默认参数值	操作	默认提示	参数名称-输入规则限制	默认参数值-输入规则限制
spark.driver.cores	1	+ -		支持字母(a-z)和小数点 (.) 256字符以内	支持字母(a-z)和数字(0-9)，长度24个字符
spark.driver.memory	1g	+ -			
spark.executor.cores	1	+ -			
spark.executor.instances	1	+ -			
spark.executor.memory	1g	+ -			

作业操作 在作业运行过程中，可对作业进行不同操作，包括但不限于保存、发布和执行。

表五 实时Flink作业操作说明

操作项	操作说明
作业保存	单击面板上方保存按钮。
发布作业	单击面板上方发布按钮，可将当前作业提交为新的版本。发布完成后可在面板右侧版本中查看历史版本、执行版本回滚操作，同时将当前作业发布到任务运维中心。
作业停止	单击面板上方停止按钮，可停止当前作业。
执行作业	单击页面上方执行，进行可视化作业测试运行，且在执行信息中弹出执行日志信息。
作业导出/入	可视化作业支持导入导出，单击导出，将当前作业导出为json文件。
计算资源	点击页面上方计算资源弹出作业资源设置框，对脚本作业进行资源设置。

Flink作业

实时Flink文件夹/作业创建

1. 实时Flink作业支持按照文件夹进行分组管理，新建作业前可以选择新增文件夹。
 2. 选择侧边导航或者快速访问中**数据加工>开发（顶部导航）>实时作业（侧边）>Flink作业**后进入Flink作业。
 3. 新建文件夹。单击新建按钮选择新建文件夹，填写相关配置项即可新建成功。
 4. 新建作业。单击新建按钮选择新建作业，文件夹支持自行选择存放作业，填写完毕点击确定新建成功。
- FlinkSQL：用户通过写SQL的方式就能够进行实时Flink作业的开发。
 - Flink自定义：用户需要在线下编辑好Flink作业包，上传到平台之上。

实时Flink作业编辑

FlinkSQL作业

创建的FlinkSQL作业中，用户能够通过类SQL脚本进行任务开发，编写相应的FlinkSQL脚本。SQL脚本模板如下：

```
--创建kafka表
drop table if EXISTS kafka_run_log;
create table kafka_run_log(
  carId int,
  fuel_consumption int,
  traffic string,
  speed int,
  latitude double,
  longitude double,
  create_time TIMESTAMP(3)
) with (
  'connector.type' = 'kafka',
  'connector.version' = '0.10',
  'connector.topic' = 'vehicle_run_info',
  'connector.properties.bootstrap.servers' = 'XX.XX.XX.XX:9093',
  'connector.startup-mode' = 'latest-offset',
  'connector.properties.enable.auto.commit' = 'false',
  'format.type' = 'csv'
);
/*
创建mysql表语句
*/
drop table if EXISTS mysql_run_log;
CREATE TABLE mysql_run_log (
  carId int,
  fuel_consumption INT,
  traffic varchar(50),
  speed int,
  latitude double,
  longitude double,
  create_time TIMESTAMP(3)
) WITH (
  'connector.type' = 'jdbc',
  'connector.url' = 'jdbc:mysql://XX.XX.XX.XX:3306/car_run_log?useSSL=false',
  'connector.table' = 'mysql_run_log',
  'connector.username' = 'edap_test',
  'connector.password' = 'edap_test@1'
);
insert into
  mysql_run_log
select
  carId,
  fuel_consumption,
  traffic,
  speed,
  latitude,
  longitude,
  create_time
from
  kafka_run_log;
select
  count(carId)
from
  mysql_run_log;
```

FlinkJAR作业

用户可以在开发界面中调整jar包的存储路径。

实时Flink作业配置

在Flink作业开发完成之后，可以在面板右侧可以查看作业配置，配置间可点击切换。

表三 配置说明

作业配置名称	配置说明
基本信息	单击右侧配置信息>基本信息，可展开查看Flink作业的基本信息，包括作业名称、责任人、作业类型、执行引擎和描述。支持对描述进行修改。
参数设置	单击参数设置，填写弹窗内参数内配置项。 - 参数名称：自定义参数名称 - 参数值：填写参数值，支持固定值及时间宏函数 - 添加/删除：添加一行参数或删除一行参数
引擎设置	1.单击引擎设置，可以根据计算需求选择不同引擎配置，可手动填写。 2.引擎也支持参数设置，具体描述参照参数设置即可。
版本	在作业发布后，可在此处查看历史版本、执行版本回滚操作。

实时Flink作业操作

在作业运行过程中，可对作业进行不同操作，包括但不限于保存、发布和执行。

表四 实时Flink作业操作说明

操作项	操作说明
作业保存	单击面板上方保存按钮。
发布作业	单击面板上方发布按钮，可将当前作业提交为新的版本。发布完成后可在面板右侧 版本 中查看历史版本、执行版本回滚操作，同时将当前作业发布到 任务运维 中心。
作业停止	单击面板上方停止按钮，可停止当前作业。
执行作业	单击页面上方 执行 ，进行可视化作业测试运行，且在 执行信息 中弹出执行日志信息。
作业导出/入	Flink作业支持导入导出，单击 导出 ，将当前作业导出为json文件。
计算资源	点击页面上方计算资源弹出作业资源设置框，对脚本作业进行资源设置。

可视化作业

新建文件夹/作业

可视化作业支持按照文件夹进行分组管理，新建作业前可以选择新增文件夹。

1. 选择侧边导航或者快速访问中**数据加工>开发（顶部导航）>实时开发>可视化作业**后进入作业组。
2. 新建文件夹。单击**新建**按钮选择**新建文件夹**，填写相关配置项即可新建成功。
3. 新建作业。单击**新建**按钮选择**新建作业**，文件夹支持自行选择存放，填写完毕点击**确定**新建成功。

可视化作业编辑

可视化作业支持四种类型插件。源端插件、数据处理组件、开发分析组件和目标端插件。用户可以返回将源端插件拖拽到开发画布中，作为实时数据的源端。具体插件的使用说明，可以点击右上角**帮助文档**查看。

表一 插件类型说明

插件类型	插件说明	操作说明
源端插件	作为整个作业的数据输入端，源端插件只能作为流水线的开始节点，下游可以连接数据处理组件、开发分析组件或者直接连接目标端插件。	目前源端支持【HTTP】、【Kafka Consumer】、【RabbitMQ Consumer】三种。用户可以返回将源端插件拖拽到开发画布中，作为实时数据的源端。
数据处理组件	对上游源端插件的数据进行数据处理，用户可以使用脚本类插件对数据进行处理。	数据处理组件拖拽到开发面板之后，将上游的源端插件连接到数据处理组件中。
数据开发组件	对上游源端插件的数据进行数据开发，用户可以使用SQL抽象插件进行数据处理。	开发分析组件和数据处理组件使用流程相同，在源端插件之后，进行数据处理。
目标端插件	将上游处理好的数据写入到目标端插件。且目标端插件需要作为流水线的最后节点。	实时作业支持目标端。用户可以将实时作业，处理加工的数据流接入到目标端。目标端可以是另一个kafka消息队列，也可以是Hive等数据库。

可视化作业配置

可视化作业配置位于整个页面右侧，单击不同的配置可查看具体配置详情。

表二 可视化作业配置项说明

可视化作业配置名称	配置说明
基本信息	单击右侧配置信息-基本信息，可展开查看可视化作业的基本信息，包括作业名称、责任人、作业类型和描述。支持对描述进行修改
参数设置	单击参数设置，填写弹窗内参数内配置项。 - 参数名称：自定义参数名称 - 参数值：填写参数值，支持固定值及时间宏函数 - 添加/删除：添加一行参数或删除一行参数
引擎设置	1.单击引擎设置，可以根据计算需求选择mapreduce或spark引擎，并可以针对不同引擎设置Driver和Executor的CPU及内存大小。 2.引擎也支持参数设置，具体描述参照参数设置即可。
版本	在作业发布后，可在此处查看历史版本、执行版本回滚操作。

可视化作业操作

在作业运行过程中，可对作业进行不同操作，包括但不限于保存、发布和执行。

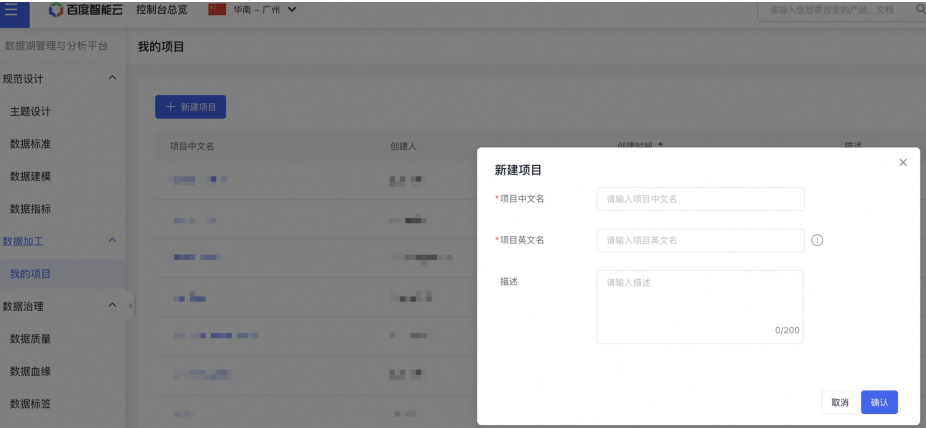
表四 可视化作业操作说明

操作项	操作说明
作业保存	单击面板上方保存按钮
发布作业	单击面板上方发布按钮，可将当前作业提交为新的版本。发布完成后可在面板右侧版本中查看历史版本、执行版本回滚操作，同时将当前作业发布到任务运维中心。
作业停止	单击面板上方停止按钮，可停止当前作业。
执行作业	单击页面上方执行，进行可视化作业测试运行，且在执行信息中弹出执行日志信息。
作业导出/入	可视化作业支持导入导出，单击导出，将当前作业导出为json文件
计算资源	点击页面上方计算资源弹出作业资源设置框，对脚本作业进行资源设置。

场景案例

EDAP Spark作业运行TPC-DS Benchmark

在【我的项目】中新建项目，用于后续在其中开发Spark作业。



在BMR控制台创建BMR集群后，在EDAP界面【资源管理】中添加资源，选择【BMR实例】、选择刚刚创建的BMR集群，用于后续作业中使用该【BMR实例】向其对应的BMR集群提交作业。然后在【绑定项目】中绑定刚刚新建的项目，使得新建项目中的作业有权使用使用【BMR实例】。

添加资源

×

计算资源配置

*资源类型:

BMR实例

金托管的Hadoop/Spark集群

单机服务器

提供单机计算资源,运行shell、python

资源组

Serverless资源组, 用于数据开发、数据分析

*计算资源名称:

只支持数字、字母、'_'、'-'，且不大于32个字符

*BMR集群实例名称:

BMR集群队列 ⓘ:

default

yarn schedule queue name

计算资源描述:

请输入

0/256

取消

确认

绑定项目

绑定项目

×

可选项目

全选

请输入

Q

tpcds-test-2

☒

tpcds_test

☒

数据集成测试

☐

lwq_test

☐

yhy_234_hudi_edap

☐

guoxinlei_test

☐

hebin

☐

已选项目

清空

tpcds-test-2

×

tpcds_test

×

取消

确认

在【数据源管理】中创建Hive类型的数据源，数据源【配置模式】选择【BMR实例模式】、hive集群选择刚刚创建的【BMR实例】资源，使得后续Spark作业可以使用该源连接与其对应的BMR集群交互。

编辑Hive源连接

×

*配置模式:

☒ BMR实例模式

☐ 连接串模式

基础信息配置

*名称:

bmr324-qa-perf

?

描述:

请输入描述

0/1024

*Catalog名称:

bmr324_qa_perf

?

*hive集群:

bmr324-qa-perf

数据湖启用Hive元数据:

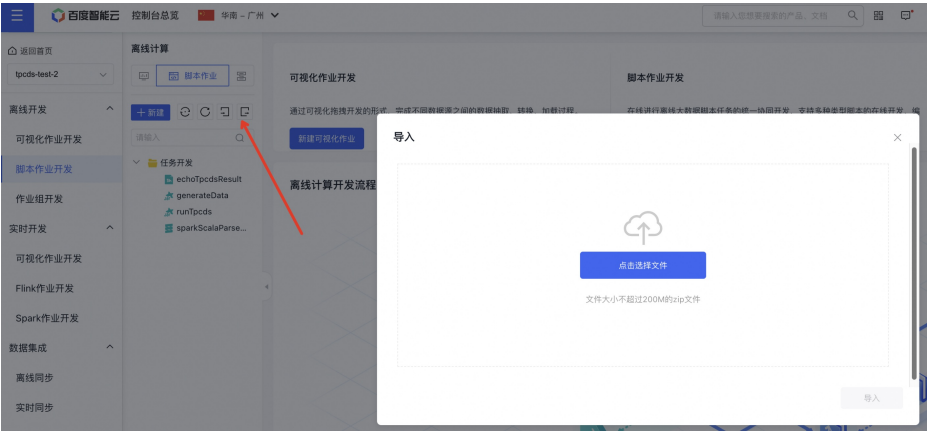
☒

?

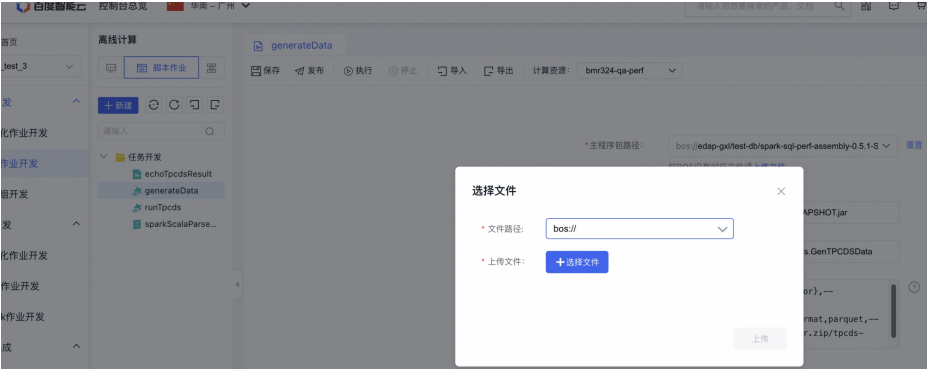
取消

确认

在【脚本作业开发】中导入Spark TPCDS测试所用的SparkJar、SparkScala、Shell作业。



在generateData作业【主程序包路径】中上传文件<https://poc-resources.bj.bcebos.com/spark-sql-perf-assembly-0.5.1-SNAPSHOT.jar> 至bos bucket，使得【主程序包路径】为刚刚上传文件的bos路径，runTpcds作业中修改【主程序包路径】为刚刚上传文件的bos路径。然后将各个作业点击【发布】，使得可以在【作业组】中引用刚刚发布的作业。



然后在【作业组开发】中导入该作业组。修改各个节点引用刚刚发布的作业的最新版本，并修改各个节点使用的【计算资源】为刚刚创建的【BMR实例】。



在bos控制台上传文件<https://poc-resources.bj.bcebos.com/tpcds-kit-master.zip>，将该文件bos路径填入作业组【参数设置】的sparkYarnDistArchives参数的值中。

"sparkYarnDistArchives"为生成数据需要使用的TPC-DS kit编译产出文件路径。

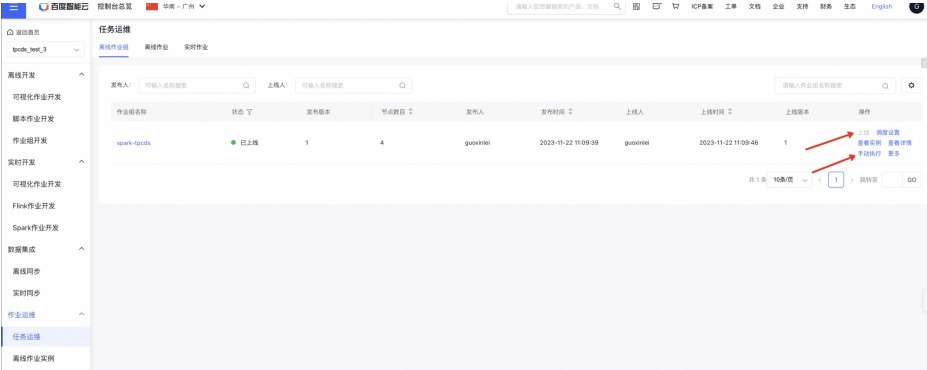
"dbPrefix"为生成数据后在BMR集群Hive中创建的database name前缀。

"location"为生成数据存储的路径。

"scaleFactor"为生成数据的规模，1相当于1 GB，100相当于100 GB。

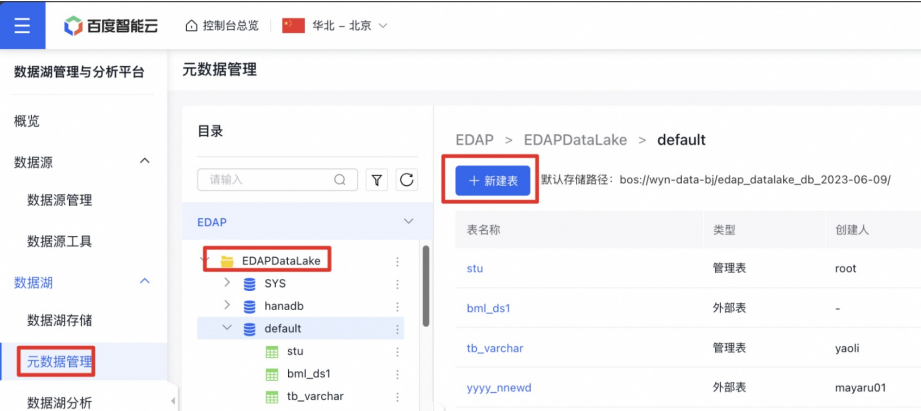
"iteration"为一次运行TPCDS测试的语句执行轮数。

然后发布该作业组，在【任务运维】中上线、手动执行该作业组，即可启动TPCDS生成数据、运行测试、解析结果、展示结果的流程。



EDAP资源组SparkSQL作业读取BOS写入RDS

首先，在【元数据管理】中选择【EDAPDataLake】中的某个database新建表。



假设BOS文件内容如下，每行为一个json字

```
{"key1":"value1","key2":"value2","properties":{"temperature":15}}
{"key1":"value1+","key2":"value2+","properties":{"temperature":18}}
{"key1":"value1++","key2":"value2++","properties":{"temperature":27}}
```

表类型选择【物理表】【外部表】，存储路径指定为BOS路径。

基本信息

* 表名称: test_json_tb

以英文开头, 支持英文、数字、下划线, 不超过128个字符

* 类型: 物理表 映射表 外部表

物理表指数据存储在EasyDAP的数据湖中的表, 由EDAP进行统一的存储路径管理和元数据管理。

描述: 请输入

0/300

存储参数

* 存储路径: BOS test test_json_tb 使用默认路径

bos://edap-auto-test/datalake-db/test_json_tb

* 数据湖格式: 无

【数据存储格式】选择【TEXTFILE】，【分

隔符】输入"\n", 添加value字段, date分区。

存储参数

* 存储路径: BOS test test_json_tb 使用默认路径

bos://edap-auto-test/datalake-db/test_json_tb

* 数据湖格式: 无

* 数据存储格式: TEXTFILE

* 分隔符: \n

高级参数配置: 0/100

元数据定义

字段名称和分区定义的字段名称不能重复。

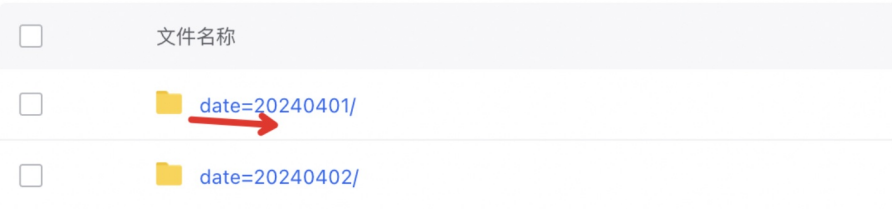
字段定义	字段名称	字段类型	类型设置	关联标准	允许空	密级	描述	操作
value	value	STRING	-	ge	☐	LO	请输入	+ -

分区定义	字段名称	字段类型	类型设置	关联标准	允许空	密级	描述	操作
date	date	STRING	-	ge	☐	LO	请输入	+ -

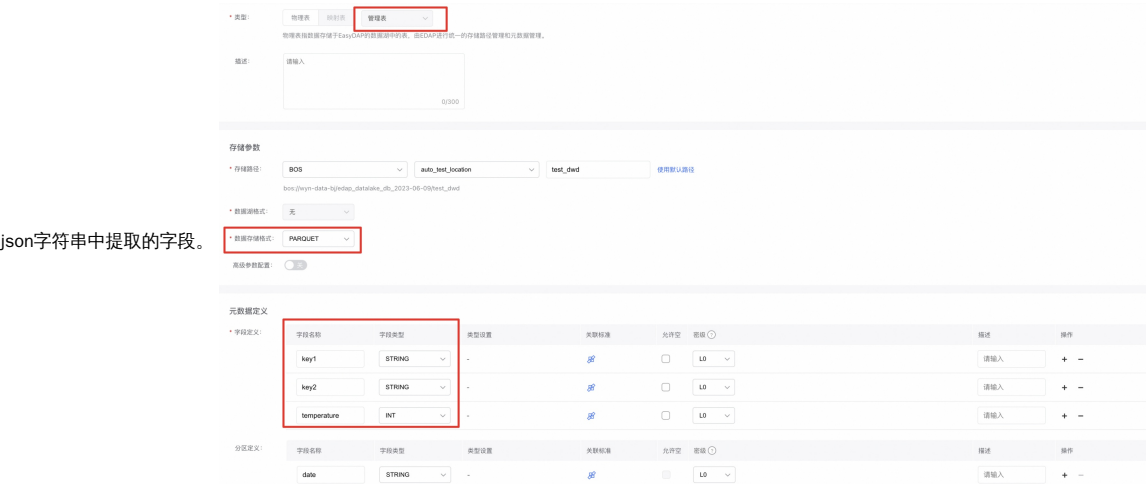
注意：要保证表bos路径下的子目录

以"date="为前缀，即刚刚创建表时指定的分区字段名称。

返回上级 | edap-auto-test / datalake-db / test_json_tb



同时，新建test_dwd中间表用于存储从原始



json字符串中提取的字段。

接下来需要在

【我的项目】界面创建项目，【资源管理】界面创建资源组，并用该资源组绑定刚刚创建的项目。



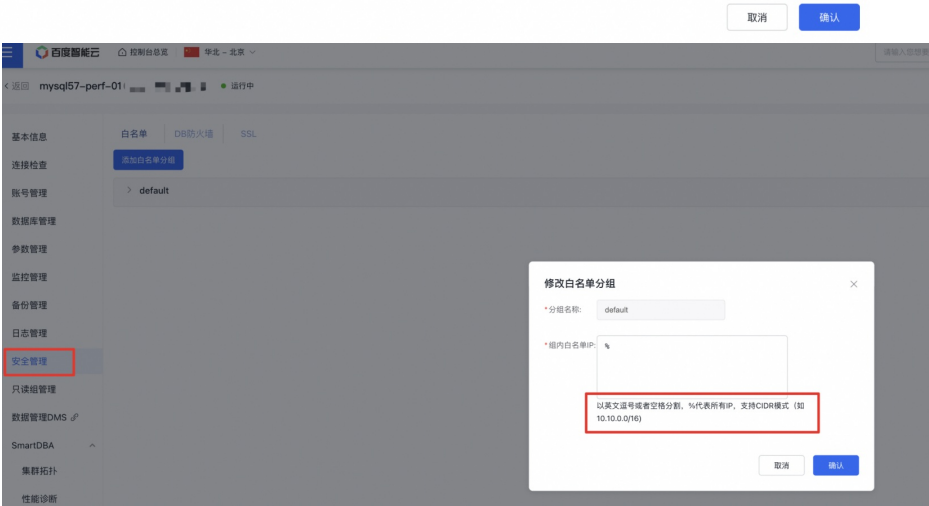
需要保证RDS服务与资源组属于相同VPC，

且资源组与RDS服务的安全组配置允许资源组访问RDS服务，资源组的VPC、安全组配置如下图。



RDS服务【安全管理】的【白名单】配置如

下图。



从【我的项目】界面进入刚刚创建的

项目，在【脚本作业开发】中创建SparkSQL作业，【数据源类型】选择EDAPDataLake，【计算资源】选择刚刚创建的资源组，使用如下SQL可读取BOS文件数据写入中间表。

sparksql-test

sparksql-test_copy

保存

发布

执行

停止

导入

导出

格式化工具

数据源类型: EDAPDataLake

计算资源: cu-g4.xl

1

— 在分布式存储中直接追加新分区数据时, 表的元数据存储不会自动感知到新分区。此时需要运行 MSCK REPAIR TABLE 命令来同步元数据并修复

2

MSCK REPAIR TABLE test_json_tb;

3

INSERT OVERWRITE TABLE test_dwd PARTITION(date='20240402')

4

SELECT

5

get_json_object(value, '\$.key1') AS key1,

6

get_json_object(value, '\$.key2') AS key2,

7

cast(get_json_object(value, '\$.properties.temperature') AS int) AS temperature

8

FROM test_json_tb

9

where date='20240402';

10

select * from test_dwd;

执行结果

key1

key2

temperature

date

value1++

value2++

19

20240401

value1++

value2++

27

20240402

value1

value2

15

20240401

value1+

value2+

18

20240401

value1

value2

15

20240402

value1+

value2+

18

20240402

务中执行如下语句准备好database、table。

执行如下SQL可将BOS数据经过聚合计算后写入RDS服务。注意：url字段填写RDS服务的内网IP。

```
mysql> select * from test_ads;
```

min_temperature	max_temperature	avg_temperature	date
15	27	20	20240402

```
1 row in set (0.00 sec)
```

首先，在【元数据管理】界面创建如下图的EDAPDataLake表，用来作为待写入的flink作业sink端



在FlinkSQL作业中写入如下SQL，实现读取Kafka维表关联MySQL写入EDAPDataLake表。需要保证Kafka、MySQL服务与资源组网络联通，例如均处于同一个VPC内。

```
CREATE TABLE kafka_source (  
  log_time STRING,  
  log_ts AS TO_TIMESTAMP(log_time,'yyyy-MM-dd HH:mm:ss.SSS'),  
  user_id INT,  
  behavior STRING,  
  proctime as PROCTIME(),  
  WATERMARK FOR log_ts AS log_ts - INTERVAL '10' SECOND  
) WITH (  
  'connector' = 'kafka',  
  'topic' = 'eventLogJson',  
  'properties.bootstrap.servers' = 'kafka_nodeip:9093',  
  'properties.group.id' = 'testGroup',  
  'scan.startup.mode' = 'latest-offset',  
  'format' = 'json',  
  'properties.security.protocol' = 'SASL_PLAINTEXT',  
  'properties.sasl.mechanism' = 'SCRAM-SHA-512',  
  'properties.sasl.jaas.config' = 'org.apache.flink.kafka.shaded.org.apache.kafka.common.security.scram.ScramLoginModule required username="xxx" password="***";',  
)  
);  
CREATE TABLE mysql_user_profile (  
  user_id INT,  
  city STRING  
) WITH (  
  'connector' = 'jdbc',  
  'url' = 'jdbc:mysql://mysql_ip:3306/biz_online?useSSL=false',  
  'table-name' = 'user_profile',  
  'username' = 'xxx',  
  'password' = '***',  
  'lookup.cache' = 'PARTIAL',  
  'lookup.cache.max-rows' = '1000',  
  'lookup.partial-cache.expire-after-write' = '60 min'  
)  
);  
CREATE CATALOG edap_catalog WITH ('type' = 'edap');  
INSERT INTO edap_catalog.`default`.test_hive_sink_bos_datalake  
/*+OPTIONS('partition.time-extractor.timestamp-pattern'='$day $hour:00:00',  
'sink.partition-commit.trigger'='partition-time',  
'sink.partition-commit.delay'='5 min',  
'sink.partition-commit.policy.kind'='metastore,success-file')*/  
SELECT ods.user_id, behavior, dim.city, DATE_FORMAT(log_ts, 'yyyy-MM-dd'), DATE_FORMAT(log_ts, 'HH')  
FROM kafka_source as ods  
LEFT JOIN mysql_user_profile FOR SYSTEM_TIME AS OF ods.proctime as dim ON ods.user_id = dim.user_id;
```

同时，如下图需要在【引擎设置】中配置"state.checkpoints.dir"="bos://<bucket name>/xx/xx"，使得flink作业生成持久化的checkpoint文件。flink作业重启时可

引擎设置

引擎配置

JobManager:

cpu:

1

内存:

1024

MB

TaskManger:

cpu:

1

内存:

1024

MB

parallelism:

1

slot数量:

1

用于状态恢复。

引擎参数

+ 添加参数

参数名称	参数值	操作
execution.checkpointing.inte	10s	+ -
state.checkpoints.dir	bos://edap-gxl/flink-checkpoi	+ -

在kafka topic中写入该数据后 {"log_time":"2023-01-05 13:08:23.867","user_id":"123","behavior":"matchV2"} {"log_time":"2023-01-05 13:05:23.867","user_id":"678","behavior":"match"}，同时mysql维表中包含下图数据

```
mysql> select * from user_profile;
+-----+-----+
| user_id | city      |
+-----+-----+
| 123     | beijing   |
| 678     | shanghai  |
+-----+-----+
2 rows in set (0.00 sec)
```

经过flinksql作业lookup join维表关联写入EDAPDataLake表后，可在EDAPDataLake表中查询到如下图结果数据。

执行信息 | 执行结果 ②

user_id	behavior	city	day	hour
123	matchV2	beijing	2023-01-05	13
678	match	shanghai	2023-01-05	13

flink作业运行过程中，如下图可在BOS控制台查看到配置的checkpoint bos路径中，在不断生成新的checkpoint文件。在修改了flink作业sql逻辑后，如果期望flink作业重启时不使用checkpoint文件进行状态恢复，在BOS控制台删除checkpoint bos路径中的所有文件即可。



基于非结构化文件在数据湖表中存储的元信息，使用PySpark任务批量拉取文件服务器中的pdf入湖。本例中数据湖选用（DeltaLake）

EDAP > EDAPDataLake > sanling > pdf_files

EDAP > EDAPDataLake > sanling > pdf_files

基本信息	字段信息	版本管理	数据规则	数据血缘
------	----------------------	------	------	------

字段定义

字段名称	字段类型	类型设置	关联标准	允许空	密级	描述
id	BIGINT	-	-	否	L0	-
name	STRING	-	-	否	L0	-
url	STRING	-	-	否	L0	-
location	STRING	-	-	否	L0	-

分区定义

字段名称	字段类型	类型设置	关联标准	允许空	密级	描述
------	------	------	------	-----	----	----

PySpark

```

import requests
from baidubce.bce_client_configuration import BceClientConfiguration
from baidubce.auth.bce_credentials import BceCredentials
from baidubce import exception
from baidubce.services import bos
from baidubce.services.bos import canned_acl
from baidubce.services.bos.bos_client import BosClient

**bos**
bos_host = "bj.bcebos.com"
access_key_id = "${access_key}"
secret_access_key = "${secret_key}"
config = BceClientConfiguration(credentials=BceCredentials(access_key_id, secret_access_key), endpoint = bos_host)
bos_client = BosClient(config)
bucket_name = "${bucket_name}"
object_prefix = "${object_prefix}"    # example: /upload/

def process_row(row):
    url = row.url
    name = row.name
    local_filename = name + ".pdf"

    response = requests.get(url, stream=True)
    response.raise_for_status()
    with open(local_filename, 'wb') as file_handle:
        for chunk in response.iter_content(chunk_size=8192):
            file_handle.write(chunk)

    object_key = object_prefix + local_filename
    data = open(local_filename, 'rb')
    bos_client.put_object_from_file(bucket_name, object_key, local_filename)
    bos_location = f"bos://{bucket_name}/{object_key}"
    return (row.id, row.name, row.url, bos_location)

df = spark.sql("select * from sanling.pdf_files")
processed_rdd = df.rdd.map(process_row)
processed_df = spark.createDataFrame(processed_rdd, df.schema)

processed_df.write.mode("overwrite").insertInto("sanling.pdf_files")

spark.sql("select * from sanling.pdf_files").show()

```

引擎设置

```

spark.hadoop.edap.endpoint=http://edap.bj.baidubce.com
spark.hadoop.fs.AbstractFileSystem.bos.impl=org.apache.hadoop.fs.bos.BaiduBosFileSystemAdapter
spark.hadoop.hive.metastore.client.class=com.baidubce.edap.catalog.metastore.EdapCatalogHiveClient
spark.sql.catalog.spark_catalog=org.apache.spark.sql.delta.catalog.DeltaCatalog
spark.sql.extensions=io.delta.sql.DeltaSparkSessionExtension

```

数据分析

概述

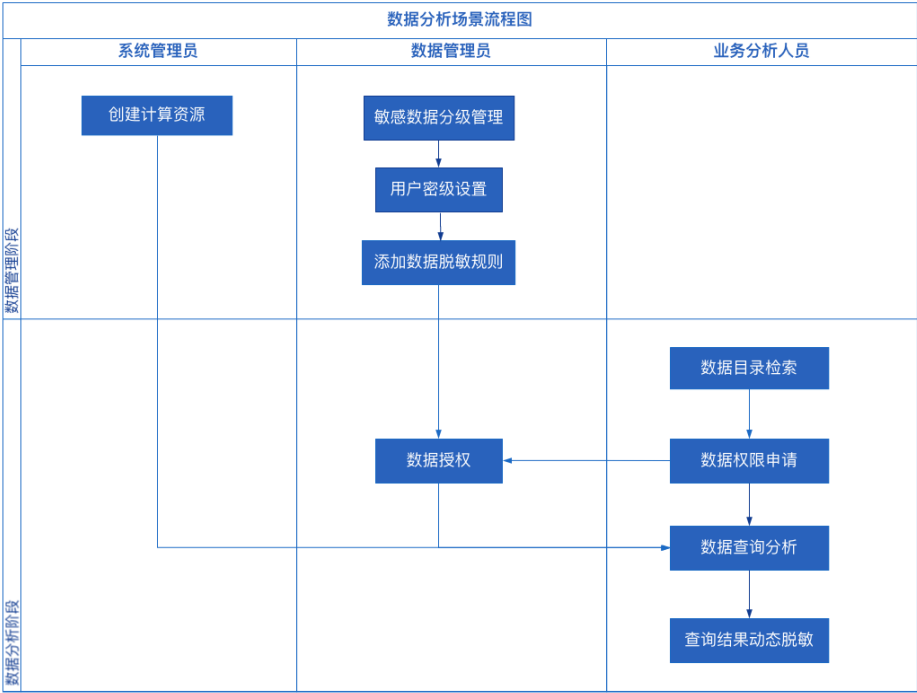
功能介绍

EDAP提供交互查询能力，支持查询SQL在线编写、运行、结果查看及结果下载，兼容标准SQL语法，支持查询历史保存。企业的业务分析人员和数据分析师可以通过EDAP自由地取数、分析，最大化发挥数据价值，赋能业务增长。数据管理人员可以使用EDAP进行细粒度的数据权限管理，同时支持多种脱敏规则，数据查询分析过程中进行动态脱敏，从而为数据的自由流动保驾护航。

产品优势

- 跨源联邦：借助计算引擎的能力，可以从不同数据源获取数据进行查询和分析。
- 查询性能：高效、海量数据查询分析，数据响应时间可达亚秒级。
- 数据安全：数据脱敏、细粒度权限管理，可以控制分析和分享数据的权限。

角色、职责和权限



角色	职责描述	功能权限
系统管理员	负责数据分析所需计算资源的创建和管理。	创建、管理计算资源集群
数据管理人员	负责EDAP平台数据管理，拥有所有数据的最高权限，进行数据权限管理。 负责EDAP平台数据安全保障，敏感数据分级管理，数据脱敏管理。	·数据授权 ·设置数据密级 ·设置用户密级 ·添加数据脱敏规则
业务分析人员	可以使用EDAP的数据湖分析、数据源查询功能进行取数、分析，形成统计分析结果用于业务决策。	·数据目录检索 ·申请数据权限 ·新建、编辑、删除数据查询任务 ·查询结果下载

🔗 数据湖分析

准备工作 数据源配置

EDAP平台提供跨数据源的交互式分析功能，支持10余种数据源的跨源联邦，包括：EDAP、Hive、Doris、Mysql、Oracle、SQLServer、Greenplum、Hana、Postgresql、Clickhouse等。在使用数据湖分析之前，需要在数据源配置界面为数据源填写Catalog名称，则数据湖分析的元数据目录将展示数据源对应的Catalog名称。

计算资源配置

为实现数据的跨源联邦分析，数据湖分析模块需要使用Trino计算引擎，在使用前需创建一个Serverless资源组集群，并且确保该集群与所查询数据源在同一个VPC下。

语法规范

查询语法遵循标准的SQL语法规范。注意：在查询数据表时，表名称需要完整写出Catalog名.库名.表名，如果不加Catalog名称，默认查询数据源为EDAP。

操作步骤 创建查询任务 操作步骤：

1. 点击左侧导航栏“数据湖分析”，新建一个查询任务。
2. 在输入框中编写SQL语句，双击左侧数据目录中的表名称，可自动生成查询语句。
3. 选择计算资源后，点击执行，将在下方显示执行语句的运行状态、耗时以及对应查询的结果。

备注：每一条执行语句，将产生一条执行结果，可点击【结果n】进行对应查看。

4. 查询结果最多显示200条数据，点击下载可将完整数据下载成excel表格。

格式化 点击【格式化】选项，即可对输入框中已有内容进行自动编排格式，按照语法规范产生对应空格。

保存查询任务 操作步骤：

1. 点击【另存为】选项，在弹出的“查询另存为”界面，编辑查询名称及其描述。

配置项名称	配置项含义	使用限制
查询名称	自定义查询名称，必填项。	只能包含英文字母、数字、中划线和下划线，且以英文字母开头，不大于50个字符。不可与已有查询名称重复。
查询描述	自定义查询描述信息，选填项。	不可超过256个字符。

2. 点击【确认】后即可将该查询结果保存，【总览】页面中可根据对应的查询名称找到该结果记录；
3. 点击查询名称，可进入查询任务详情页，点击【编辑】可再次打开上图弹窗，重新编辑信息，点击【删除】即可完成删除操作；
4. 在【运行历史】中也可查看到此前该查询运行的详细情况记录。

数据运维

运维大屏

进入运维大屏

登录EasyDAP,进入EasyDAP首页。选择侧边导航**数据加工>运维>运维大屏**，进入运维大屏中心。可以在运维大屏查看数据开发和数据集成的运行状态和实例完成情况。

数据开发运维概况

表一 数据开发运维概况说明

运维概况内容	描述
运行状态	数据每30分钟更新一次。支持通过日期筛选(今天/昨天/前天/近7天)，也支持通过任务执行人进行筛选(全部实例/我的实例)。
实例完成情况	数据每小时更新一次。每个时刻点通知的是当前小时的实例完成情况。支持根据任务状态和任务类型进行筛选。
任务数/生效调度数	数据6小时更新一次。
任务总数	数据6小时更新一次。使用饼图统计当前项目下，任务的总数量。鼠标移入饼图不同颜色表示不同作业类型。
近7天运行失败排行	数据每天更新，仅展示前20条。
昨日运行时长排行	数据每天更新，仅展示前20条。
近7天任务平均资源使用排行	数据每天更新，仅展示前20条。

数据集成运维概况

表二 数据开发运维概况说明

运维概况内容	描述
任务状态	任务状态数据实时更新。当前项目下，任务总数及各种类型任务分布情况。鼠标移入不同色块，可看到当前任务运行状态下有几个任务。
数据同步统计	同步数据统计每小时更新一次。
任务执行时长列表	查看按照执行时长进行排序。

实例运维

离线作业

实例运维分为三部分：离线作业组、离线作业和实时作业。

离线作业列表

用户在开发界面开发好离线作业组、离线作业或者实时作业之后，进行发布；发布的任务将展示到任务运维列表中，在任务运维中心可对已发布的作业进行上线、设置定时调度、手动执行、修数据、任务下线及查看实例等操作。

1. 单击**数据加工>运维（顶部导航）>实例运维（侧边导航）>离线作业**进入离线作业列表，列表显示离线作业组实例，开始时间默认选中近一小时。
2. 列表形式展示离线作业组实例显示。支持按照开始时间、结束时间和调度时间进行筛选，并且可以按照实例名称、责任人或作业名称搜索。以下是列表展示项说明：

表一 离线作业实例说明

展示项名称	展示项说明
实例ID	显示作业实例ID信息。
作业名称	显示作业的名称。
作业类型	显示作业的类型。支持按照作业类型进行筛选，包括可视化、Shell、Spark、Python、Hive SQL、JDBC SQL、Spark SQL、Spark Scala、PySpark、离线同步、数据质量。
作业状态	显示当前作业实例的运行状态。作业实例状态参考作业实例状态说明。
实例来源	周期性调度、手动执行、修数据、重跑。
调度时间	作业实例运行的业务日期。
责任人	当前作业的创建人。
开始时间	作业实例的运行开始时间。
结束时间	作业实例的运行结束时间。
运行时间	作业实例的运行时长。
操作	对离线作业组实例进行相关操作。

离线作业实例操作

可对离线作业实例进行相关操作，如：查看实例详情、上游依赖。

表二 离线作业实例操作说明

操作名称	操作描述
查看实例详情	选择具体作业，单击操作列可进入实例组详情界面，查看作业日志。
上游依赖	单击上游依赖进行跳转，展示当前离线作业实例所依赖的上游任务。

离线作业批量操作

在离线作业列表中列表项上方可单击对作业进行批量操作包括重跑、终止和忽略。

表三 离线作业批量操作说明

批量操作名称	描述
重跑	勾选作业实例，可进行实例重跑。支持重新运行全部作业或者仅运行费成功态的作业。
终止	勾选正在运行中的作业实例，进行作业实例终止操作。
忽略	针对离线作业实例状态为失败、停止、未触发的作业，支持忽略操作。

离线作业实例状态说明

作业实例运行状态可分为以下类型，以下是对作业实例运行状态的说明。

图一 作业实例状态说明

分类	名称	说明
准备态	等待中	作业节点进入等待分配执行资源的待运行状态。
	禁用	作业节点还未运行但处于准备跳过的状态(用户点击跳过节点后，在作业组 DAG 中还未轮到该节点运行)
运行态	运行中	作业节点正在运行的状态
完成态	跳过	作业节点被跳过状态
	已停止	作业节点停止状态
	已取消	因上游节点状态等不满足该节点执行条件而无法触发该节点执行的状态
	运行成功	执行成功状态
	运行失败	执行失败状态

🔗 作业组

作业组列表

1. 单击数据加工>运维（顶部导航）>实例运维（侧边导航）>作业组进入离线作业列表，列表显示离线作业组实例，开始时间默认选中近一小时。

2. 列表展示已发布的作业组，列表展示项包括：

表一 作业组展示项说明

展示项名称	展示项说明
实例ID	显示作业组实例ID信息。
作业组名称	显示作业组的名称。
状态	显示当前作业组实例的运行状态。作业组实例状态参考作业实例状态说明。
实例来源	周期性调度、手动执行、修数据、重跑。
调度时间	作业组实例运行的业务日期。
责任人	当前作业组的创建人。
开始时间	作业组实例的运行开始时间。
结束时间	作业组实例的运行结束时间。
运行时间	作业组实例的运行时长。
更新时间	最近一次更新的时间。
操作	对离线作业组实例进行相关操作。

作业组操作

单击列表后操作项，可对作业组实例进行相关操作，如：查看实例详情、上游依赖和日志详情等。

表二 离线作业实例操作说明

操作名称	操作描述
查看实例详情	选择具体作业组，单击操作列查看实例详情可进入实例组详情界面，查看作业组的DAG图列表，可对作业组进行重跑操作。
上游依赖	单击上游依赖进行跳转，展示当前离线作业实例所依赖的上游任务。
日志详情	单击操作列日志详情可跳转详情界面，查看作业组日志详情。
执行耗时	单击操作列执行耗时可跳转详情界面，查看作业节点执行耗时详情。
工作组任务	单击操作列工作组任务可跳转详情界面，查看作业节点工作组任务详情，并且可以对作业进行终止操作。

作业组批量操作

在作业组列表中列表项上方可单击对作业组进行批量操作包括重跑、终止和忽略。

表三 离线作业批量操作说明

批量操作名称	描述
重跑	勾选作业组，可进行实例重跑。支持重新运行全部作业组或者仅运行费成功态的作业组。
终止	勾选正在运行中的作业组，进行作业组终止操作。
忽略	针对作业组状态为失败、停止、未触发的，支持忽略操作。

实时作业

实时作业实例列表

列表形式展示实时作业实例，可选择展示项进行展示。支持按照实例ID、责任人和任务名称进行筛选。

表一 具体列表展示项说明

展示项说明	展示项说明
实例ID	显示实时作业实例ID。
作业名称	显示实时作业名称。
运行状态	显示当前实时作业实例的运行状态。支持按照运行状态筛选。运行状态详情参照运行状态说明。
作业类型	展示当前作业类型。支持按照作业类型进行筛选，作业类型包括实时流水线、FlinkSQL作业、Flink自定义作业、SparkStreaming Scala、SparkStreaming自定义。
责任人	当前作业组的创建人。
开始时间	实时作业实例的运行开始时间。
结束时间	实时作业实例的运行结束时间。
运行时间	实时作业实例的运行时长。
操作	终止、作业详情、日志和执行分析。

实时作业实例操作

在实时作业列表操作项，对实时作业实例能够进行的操作。

表二 实时作业操作说明

操作项	操作说明
终止	单击终止，将正在运行中的实时实例终止运行。终止运行的实例，需要找到对应的实时任务，再进行重新启动。
作业详情	单击作业详情，显示实时流水线作业实例详情信息。
日志	单击日志，显示实时作业实例日志信息。
执行分析	单击执行分析，显示实时流水线作业实例执行分析信息。

🔗 任务运维

🔗 作业组

作业组列表

列表形式展示“已发布”的离线作业组，支持按照发布人、上线人和作业组名称进行筛选。列表展示项包括：

表一 展示项说明

展示项名称	描述
作业名称	显示实时作业的名称。
状态	显示当前任务的发布/上线状态。 - 已发布：任务第一次发布时，状态为【已发布】状态，此时任务没有真正在运维环境中上线，不能进行任务启动运行。 - 已上线：针对已发布的任务进行上线操作。此时任务状态为【已上线】。已上线实时任务能够进行启动运行。 - 已上线有新版本：当已上线的任务，再次进行编辑发布后，此时对应的任务状态为【已上线有新版本】。在线上周期性运行的版本仍旧是老版本。此时，再进行【上线】操作，任务状态变为【已上线】，新版本任务覆盖老版本任务。
发布版本	已发布实时作业的最新版本。
作业类型	实时流水线、FlinkSQL作业、Flink自定义作业、SparkStreaming Scala、SparkStreaming自定义。
发布人	本次实时作业的发布人。
最近一次发布时间	实时作业最近一次的发布时间。
上线人	本次实时作业的上线操作人。
最近一次上线时间	实时作业最近一次的上线操作时间。
上线版本	已上线实时作业的最新版本。
操作	上线、启动、下线、删除、查看实例、查看详情。

作业组操作

可对当前作业组进行相关操作，包括上线、手动执行、调度设置、调度计划、修数据、下线并删除、查看实例、查看详情。

表二 离线作业操作说明

操作项名称	描述
上线	当作业状态为已发布、已上线有新版本时，能够进行上线操作。点击上线之后，作业状态变更为已上线。
手动执行	单击手动执行，手动运行单次运行任务。运行时需要选择一个业务日期，确定本次运行哪天的业务数据。
修数据	针对周期性运行失败的情况，或作业初次上线需要对历史数据进行补充的情况。需要对作业进行修数据操作。修数据任务目前需要配置三个条件： - 业务时间范围：用户选择一个时间段，在时间段内作业组会按照调度周期设置，运行相应的实例。如果对应的时间点上已经有对应实例，则不会重新生成新实例，会将历史已有实例重新运行，如果对应的时间点上没有对应实例，则产生新的实例。 - 运行设置：运行全部实例则是指在对应的业务时间范围内，运行全部的实例，不管实例是什么状态。仅运行未成功执行实例则跳过成功作业，仅运行未成功的实例。 - 是否触发下游依赖作业：选择是则将会下游的所有依赖触发重新运行。选择否则仅运行当前作业，不会拉起下游作业。
调度设置	点击操作栏中调度设置对作业进行调度设置。设置好周期性调度之后，当作业发布之后，调度即可使用周期。详见调度设置详情。
调度计划	单击操作列更多-调度计划可查看当前任务的历史修数据任务及历史调度实例信息。
下线	对当前已上线的作业进行下线当前作业历史作业实例，但在任务运维中下线的作业在开发作业中不会被删除，仍可进行作业发布，发布后的作业看展示历史作业实例。
查看实例	单击操作列查看实例，可跳转至离线作业实例列表，查看当前作业生成的历史作业实例。
查看详情	单击操作列查看详情，可跳转至离线作业DAG配置页面，对配置内容进行修改。
删除	单击操作列删除，可删除掉当前离线作业。

调度设置

调度类型

调度类型：分为基本设置和crontab设置。基本设置以可视化的形式设置周期性调度，cronTab已表达式的形式来进行功能设置。

表三 执行周期说明

执行周期	说明
分钟	选择分钟级别的周期，分钟级别的执行间隔为5分钟、10分钟、30分钟三种间隔类型。在配置好执行周期之后，可以在预览中，显示按照选中的执行周期一天内，在哪些时间点运行哪些批次。
小时	选择小时级别的周期，执行间隔从1小时-23小时。开始时间可以从0分-59分任意时间点。
天	选择天级别的周期，执行间隔从1天-31天。开始时间可以0点0分-23点59分任意时间点。
周	选择周级别的执行周期，执行间隔从周一到周日，可以进行多选。开始时间可以0点0分-23点59分任意时间点。
月	选择月级别的执行周期，执行间隔从1天-31天。开始时间可以0点0分-23点59分任意时间点。

表四 调度设置cron说明（字符整体不支持l，w，c、#）

字段	取值范围	特殊字符
分钟	[0, 59]	*, - /
小时	[0, 23]	*, - /
天	[1, 31]	? *, - /
月	[1, 12]或[JAN, DEC]	*, - /
周	支持数字【1~6】代表周一~周六、【0 7】代表周日、英文【SUN-SAT】。	? *, - /

表五 特殊字符说明

特殊字符	含义	示例
,	列出枚举值。	在字段分钟中，5,20表示分别在5分钟和20分钟触发一次。
-	指定范围。	在字段分钟中，5-20表示从5分钟到20分钟之间每隔一分钟触发一次。
/	指定数值的增量。	在字段分钟中，0/15表示从第0分钟开始，每15分钟。在字段分钟中3/20表示从第3分钟开始，每20分钟。
?	不指定值，仅用于日期和星期。	当字段日期或星期其中之一被指定了值以后，为了避免冲突，需要将另一个字段的值设为?。
*	匹配任意值。	在字段月中，*表示每个月。

依赖关系

依赖关系设置是一种任务依赖上的一种触发。当上游作业成功或者完成之后，才会触发下游的作业。

此处添加的节点即可视为上当前作业的上游。

相关依赖限制：

在选择当前作业依赖作业时，需获取当前作业所有被依赖的作业（包括当前作业本身以多层依赖），均不可再被选择依赖，具体场景如下：

- 1) 当前作业不允许依赖其本身，不允许下拉选择当前作业；
- 2) 各类作业间不允许彼此循环依赖，如：当前作业A依赖作业B，则不允许在作业B中再依赖作业A；
- 3) 多个作业间不允许形成循环闭环依赖，如：作业A依赖作业B，作业B依赖作业C，作业C依赖项包括作业C本身、作业B、作业A，因此作业C不可依赖作业C、作业B、作业A。

可以选择依赖项目和依赖作业，在依赖条件上分为全部成功和完成两类。如果是全部成功，则表示需要作业能够全部都运行成功，才能触发下游。选择完成，则表示只需要作业是完成态则能够触发下游运行。

匹配类型

在配置任务依赖后，当前任务执行之前需要对上游依赖任务的执行情况进行判断。判断逻辑如下：

- 1. 当选择精确匹配时，任务执行前会校验上游依赖任务是否在指定时间内执行了特定数量的任务。如果执行时间或任务数量与调度计划不一致，则下游任务触发失败。
- 2. 当选择模糊匹配时，当任务数量一致，但执行时间与调度计划有轻微偏差时，校验通过，触发下游任务。

依赖偏移

依赖偏移是描述哪些实例完成或者成功才能触发下游的，依赖偏移的配置分为三种情况：空、区间偏移、枚举偏移。

- 1.空：如果依赖偏移处为空，则系统默认按照当前作业的执行周期，向前偏移一个执行间隔单元作为起始时间点，到当前时间点，作为一个偏移时间段。

以在此时间段内的依赖作业是否有对应的实例及实例的状态是否和依赖条件一致，来作为是否触发当前作业的条件。如果一致则触发当前作业，如果不一致则保持“待触发”状态，在等待超过了执行间隔后，变为“未触发”状态。

不同类型的起始默认偏移开始时间点和结束时间点列表如下：

	执行周期	执行间隔	偏移开始时间点 (不包含此点)	偏移结束时间点 (包含此点)
1				
2	分钟	5分钟	当前时间-5分钟	当前时间
3	分钟	10分钟	当前时间-10分钟	当前时间
4	分钟	30分钟	当前时间-30分钟	当前时间
5	小时	1小时	当前时间-1小时	当前时间
6	小时	23小时	当前时间-23小时	当前时间
7	天	1天	当前时间-1天	当前时间
8	天	31天	当前时间-31天	当前时间
9	周	均为7天	当前时间-7天	当前时间
10	月	均为30天	当前时间-30天	当前时间

- 2.区间偏移：依赖偏移可以支持区间偏移关系的填写。基本格式为(-7day,0]。以当前时间前推第一个时间点为开始时间点，以当前时间前推第二个时间点为结束时间点。查看在此时间段内的依赖作业组实例的状态。

表六 依赖偏移参数

参数名称	详细说明
支持的表达式类型	点、集合和区间表达式。建议使用区间表达式，以降低下游任务失败的风险。
默认值	若不填写，默认为区间表达式，表示当前作业组的一个调度周期。例如，若作业组每天3点执行，则默认的检查时间范围是昨天3点到今天3点之间([-24h,0])。
集合表达式	例如：{-1min, 2hour}，使用大括号，表示检查两个时间点的任务。一个是检查本作业组调度时间开始1分钟前的依赖作业执行情况，另一个是检查本作业组调度时间开始2小时后的依赖作业执行情况。
区间表达式	支持左右开闭四种组合的数学区间写法。例如：作业组调度时间为每日7:00，填写[-20min,10min)，表示检查业务时间在每日6:40至7:10之间的依赖作业执行完成情况。
支持的单位	分钟（min）、小时（hour）、天（day）、周（week）、月（month）。
检查结果	如果依赖任务的检查时间点/集合/区间内所有任务成功完成，则下游任务将继续执行；否则将不会被触发。
点表达式	例如：-1min，表示检查本作业组调度时间开始1分钟前的依赖作业执行情况。

表七 区间偏移说明

字段	说明
区间表示符号	支持下括号和中括号。小括号指不包含此时间点，中括号指包含此时间点。如：()、[]、()、[]这四种形式。
支持的时间单位	分钟：min 小时：hour 天：day 星期：week 月：month
数字限制	数字需均为非正数的整数。且第一个数值需要小于第二个数值。

表八 常用区间偏移示例

区间偏移	依赖解释
(-7day,0]	以当前时间点前推7天为开始时间点，到当前时间点为结束点，在此范围内，来看相应的作业的实例状态。
(-24hour,-12hour]	以当前时间点前推12小时为开始时间点，到当前时间点前推12小时为结束点，在此范围内，来看相应的作业的实例状态。
(-60min,0]	以当前时间点前推60分钟为开始时间点，到当前时间点为结束点，在此范围内，来看相应的作业的实例状态。

3.枚举偏移：依赖偏移可以支持枚举偏移关系的填写。基本格式为{-24hour,-12hour}。已当前时间点前推每一个时间点，查看各个时间点上的是否有对应实例及对应实例的状态。

表九 枚举偏移说明

字段	说明
区间表示符号	支持大括号{}
支持的时间单位	分钟：min 小时：hour 天：day 星期：week 月：month
数字限制	数字需均为非正数的整数。

表十 常用枚举偏移示例

区间迁移	依赖解释
{-24hour,-12hour}	以当前时间点前推24小时和前推12小时，查看这两个时间点的作业的实例状态。
{-10min,-5min}	以当前时间点前10分钟和前推5分钟，查看这两个时间点的作业的实例状态。

依赖串并行

支持作业设置依赖的串并行功能。依赖上一周期，设置为是则表示作业为串行模式，设置为否则表示作业为并行模式，需要用户设置最大并行数。

并行设置如果为2，则表示同一个时刻只允许有两个作业实例同时运行。如果并行设置为1，则效果和串行一致，同一时刻只允许一个作业实例同时运行。

失败策略

当作业内某个作业节点失败时，指定失败策略。

- 继续：一个节点失败的时候，当前在运行节点继续运行直至完成。
- 终止：一个节点失败的时候，当前在运行节点停止。

实时作业

实时作业列表

列表形式展示实时作业，支持按照**发布人**、**上线人**和**作业名称**进行搜索。列表上方**设置**按钮可自行选择需要展示项，列表展示项包括：

表一 实时作业列表展示项说明

展示项名称	描述
作业名称	显示实时作业的名称。
状态	显示当前任务的发布/上线状态。 - 已发布：任务第一次发布时，状态为【已发布】状态，此时任务没有真正在运维环境中上线，不能进行任务启动运行。 - 已上线：针对已发布的任务进行上线操作。此时任务状态为【已上线】。已上线实时任务能够进行启动运行。 - 已上线有新版本：当已上线的任务，再次进行编辑发布后，此时对应的任务状态为【已上线有新版本】。在线上周期性运行的版本仍旧是老版本。此时，再进行【上线】操作，任务状态变为【已上线】，新版本任务覆盖老版本任务。
发布版本	已发布实时作业的最新版本。
作业类型	实时流水线、FlinkSQL作业、Flink自定义作业、SparkStreaming Scala、SparkStreaming自定义。
发布人	本次实时作业的发布人。
最近一次发布时间	实时作业最近一次的发布时间。
上线人	本次实时作业的上线操作人。
最近一次上线时间	实时作业最近一次的上线操作时间。
上线版本	已上线实时作业的最新版本。
操作	上线、启动、下线、删除、查看实例、查看详情。

实时作业操作

可对当前实时作业进行相关操作，包括上线、启动、下线并删除、查看实例、查看详情。

操作项名称	描述
上线	当实时作业状态为已发布、已上线有新版本时，能够进行上线操作。点击上线之后，实时作业状态变更为已上线。
启动	单击操作列启动，将已上线的实时作业进行任务启动。
下线	对当前已上线的实时作业进行下线操作时会一并删除当前实时作业实例，但在任务运维中下线的实时作业在开发作业中不会被删除，仍可进行作业发布，发布后的作业看展示历史实时作业实例。
删除	可对下线状态的作业进行删除。
查看实例	单击操作列 查看实例 ，可跳转至实时作业实例列表，查看当前实时作业生成的历史作业实例。
查看详情	单击操作列 查看详情 ，跳转作业详情界面，可对作业进行操作。

离线作业

任务运维分为三部分：作业组、离线作业以及实时作业。

用户在开发界面开发好离线作业组、离线作业或者实时作业之后，进行发布；发布的任务将展示到任务运维列表中，在任务运维中心可对已发布的作业进行上线、设置定时调度、手动执行、修数据、任务下线及查看实例等操作。

离线作业列表

列表形式展示已发布的离线作业，列表展示项包括：

表一 展示项说明

展示项名称	描述
作业名称	显示作业的名称。
状态	显示当前任务的发布/上线状态。状态包括三个枚举值： - 已发布：任务第一次发布时，状态为已发布状态，此时任务没有真正在运维环境中上线，不能进行周期性运行。 - 已上线：针对已发布的任务进行上线操作。此时任务状态为已上线。已上线任务能够按照调度设置进行周期性运行。 - 已上线有新版本：当已上线的任务，再次进行编辑发布后，此时对应的任务状态为已上线有新版本。在线上周期性运行的版本仍旧是老版本。此时，再进行上线操作，任务状态变为已上线，新版本任务覆盖老版本任务。
发布版本	已发布作业的最新版本。
发布人	本次作业的发布人。
发布时间	本次作业的发布时间。
上线人	本次作业的上线操作人。
上线时间	本次作业的上线操作时间。
上线版本	已上线作业的最新版本。
操作	可对当前作业进行相关操作，包括上线、手动执行、调度设置、调度计划、修数据、下线并删除、查看实例、查看详情。

离线作业操作

可对当前作业进行相关操作，包括上线、手动执行、调度设置、调度计划、修数据、下线并删除、查看实例、查看详情。

表二 离线作业操作说明

操作项名称	描述
上线	当作业状态为已发布、已上线有新版本时，能够进行上线操作。点击上线之后，作业状态变更为已上线。
手动执行	单击手动执行，手动运行单次运行任务。运行时需要选择一个业务日期，确定本次运行哪天的业务数据。
修数据	针对周期性运行失败的情况，或作业初次上线需要对历史数据进行补充的情况。需要对作业进行修数据操作。修数据任务目前需要配置三个条件： - 业务时间范围：用户选择一个时间段，在时间段内作业组会按照调度周期设置，运行相应的实例。如果对应的时间点上已经有对应实例，则不会重新生成新实例，会将历史已有实例重新运行，如果对应的时间点上没有对应实例，则产生新的实例。 - 运行设置：运行全部实例则是指在对应的业务时间范围内，运行全部的实例，不管实例是什么状态。仅运行未成功执行实例则跳过成功作业，仅运行未成功的实例。 - 是否触发下游依赖作业：选择是则将会下游的所有依赖触发重新运行。选择否则仅运行当前作业，不会拉起下游作业。 - 修数周期：选择自定义，不影响作业的调度周期配置信息。 - 运行中实例处理策略：修数据任务执行过程中，遇到触发下游的作业节点正在被调度任务调起并处于运行态，指定一种处理策略，忽略或重新执行当前处于运行态的实例。 - 最大并发数：默认：继承作业的最大并发数，若作业未配置，默认为1。 自定义：单独设置本次修数据任务的最大并发数，不影响作业原有配置。
调度设置	点击操作栏中调度设置对作业进行调度设置。设置好周期性调度之后，当作业发布之后，调度即可使用周期。详见调度设置详情。
调度计划	单击操作列 更多-调度计划 可查看当前任务的历史修数据任务及历史调度实例信息。
下线	转为不可用的状态，不能调度、修数据、手动执行。
删除	单击操作列 删除 ，可删除掉当前离线作业。
查看实例	单击操作列查看实例，可跳转至离线作业实例列表，查看当前作业生成的历史作业实例。
查看详情	单击操作列 查看详情 ，可跳转至离线作业DAG配置页面，对配置内容进行修改。

调度设置

调度类型

调度类型：分为基本设置和crontab设置。基本设置以可视化的形式设置周期性调度，cronTab已表达式的形式来进行功能设置。设置对应的周期之后，作业就会按照不同的周期进行运行。

表三 执行周期说明

执行周期	说明
分钟	选择分钟级别的周期，分钟级别的执行间隔为5分钟、10分钟、30分钟三种间隔类型。在配置好执行周期之后，可以在预览中，显示按照选中的执行周期一天内，在哪些时间点运行哪些批次。
小时	选择小时级别的周期，执行间隔从1小时-23小时。开始时间可以从0分-59分任意时间点。
天	选择天级别的周期，执行间隔从1天-31天。开始时间可以0点0分-23点59分任意时间点。
周	选择周级别的执行周期，执行间隔从周一到周日，可以进行多选。开始时间可以0点0分-23点59分任意时间点。
月	选择月级别的执行周期，执行间隔从1天-31天。开始时间可以0点0分-23点59分任意时间点。

表四 调度设置cron说明（字符整体不支持l，w，c、#）

字段	取值范围	特殊字符
分钟	[0, 59]	*, - /
小时	[0, 23]	*, - /
天	[1, 31]	? *, - /
月	[1, 12]或[JAN, DEC]	*, - /
周	支持数字【1~6】代表周一~周六、【0 7】代表周日、英文【SUN-SAT】。	? *, - /

表五 特殊字符说明

特殊字符	含义	示例
,	列出枚举值。	在字段分钟中，5,20表示分别在5分钟和20分钟触发一次。
-	指定范围。	在字段分钟中，5-20表示从5分钟到20分钟之间每隔一分钟触发一次。
/	指定数值的增量。	在字段分钟中，0/15表示从第0分钟开始，每15分钟。在字段分钟中3/20表示从第3分钟开始，每20分钟。
?	不指定值，仅用于日期和星期。	当字段日期或星期其中之一被指定了值以后，为了避免冲突，需要将另一个字段的值设为?。
*	匹配任意值。	在字段月中，*表示每个月。

依赖关系

依赖关系设置是一种任务依赖上的一种触发。当上游作业成功或者完成之后，才会触发下游的作业。

相关依赖限制：

在选择当前作业依赖作业时，需获取当前作业所有被依赖的作业（包括当前作业本身以多层依赖），均不可再被选择依赖，具体场景如下：

- 1) 当前作业不允许依赖其本身，不允许下拉选择当前作业；
- 2) 各类作业间不允许彼此循环依赖，如：当前作业A依赖作业B，则不允许在作业B中再依赖作业A；
- 3) 多个作业间不允许形成循环闭环依赖，如：作业A依赖作业B，作业B依赖作业C，作业C依赖项包括作业C本身、作业B、作业A，因此作业C不可依赖作业C、作业B、作业A。

可以选择依赖项目和依赖作业，在依赖条件上分为全部成功和完成两类。如果是全部成功，则表示需要作业能够全部都运行成功，才能触发下游。选择完成，则表示只需要作业是完成态则能够触发下游运行。

匹配类型

在配置任务依赖后，当前任务执行之前需要对上游依赖任务的执行情况进行判断。判断逻辑如下：

- 1. 当选择精确匹配时，任务执行前会校验上游依赖任务是否在指定时间内执行了特定数量的任务。如果执行时间或任务数量与调度计划不一致，则下游任务触发失败。
- 2. 当选择模糊匹配时，当任务数量一致，但执行时间与调度计划有轻微偏差时，校验通过，触发下游任务。

依赖偏移

依赖偏移是描述哪些实例完成或者成功才能触发下游的，依赖偏移的配置分为三种情况：空、区间偏移、枚举偏移。

- 1. 空：如果依赖偏移处为空，则系统默认按照当前作业的执行周期，向前偏移一个执行间隔单元作为起始时间点，到当前时间点，作为一个偏移时间段。

在此时间段内的依赖作业是否有对应的实例及实例的状态是否和依赖条件一致，来作为是否触发当前作业的条件。如果一致则触发当前作业，如果不一致则保持“待触发”状态，在等待超过了执行间隔后，变为“未触发”状态。

不同类型的起始默认偏移开始时间点和结束时间点列表如下：

	执行周期	执行间隔	偏移开始时间点 (不包含此点)	偏移结束时间点 (包含此点)
1				
2	分钟	5分钟	当前时间-5分钟	当前时间
3	分钟	10分钟	当前时间-10分钟	当前时间
4	分钟	30分钟	当前时间-30分钟	当前时间
5	小时	1小时	当前时间-1小时	当前时间
6	小时	23小时	当前时间-23小时	当前时间
7	天	1天	当前时间-1天	当前时间
8	天	31天	当前时间-31天	当前时间
9	周	均为7天	当前时间-7天	当前时间
10	月	均为30天	当前时间-30天	当前时间

2.区间偏移：依赖偏移可以支持区间偏移关系的填写。基本格式为(-7day,0]。以当前时间前推第一个时间点为开始时间点，以当前时间前推第二个时间点为结束时间点。查看在此时间段内的依赖作业组实例的状态。

表六 依赖偏移参数

参数名称	详细说明
支持的表达式类型	点、集合和区间表达式。建议使用区间表达式，以降低下游任务失败的风险。
默认值	若不填写，默认为区间表达式，表示当前作业组的一个调度周期。例如，若作业组每天3点执行，则默认的检查时间范围是昨天3点到今天3点之间([-24h,0])。
集合表达式	例如：{-1min, 2hour}，使用大括号，表示检查两个时间点的任务。一个是检查本作业组调度时间开始1分钟前的依赖作业执行情况，另一个是检查本作业组调度时间开始2小时后的依赖作业执行情况。
区间表达式	支持左右开闭四种组合的数学区间写法。例如：作业组调度时间为每日7:00，填写[-20min,10min)，表示检查业务时间在每日6:40至7:10之间的依赖作业执行完成情况。
支持的单位	分钟（min）、小时（hour）、天（day）、周（week）、月（month）。
检查结果	如果依赖任务的检查时间点/集合/区间内所有任务成功完成，则下游任务将继续执行；否则将不会被触发。
点表达式	例如：-1min，表示检查本作业组调度时间开始1分钟前的依赖作业执行情况。

表七 区间偏移说明

字段	说明
区间表示符号	支持下括号和中括号。小括号指不包含此时间点，中括号指包含此时间点。如：()、[]、()、[]这四种形式。
支持的时间单位	分钟：min 小时：hour 天：day 星期：week 月：month
数字限制	数字需均为非正数的整数。且第一个数值需要小于第二个数值。

表八 常用区间偏移示例

区间偏移	依赖解释
(-7day,0]	以当前时间点前推7天为开始时间点，到当前时间点为结束点，在此范围内，来看相应的作业的实例状态。
(-24hour,-12hour]	以当前时间点前推12小时为开始时间点，到当前时间点前推12小时为结束点，在此范围内，来看相应的作业的实例状态。
(-60min,0]	以当前时间点前推60分钟为开始时间点，到当前时间点为结束点，在此范围内，来看相应的作业的实例状态。

3.枚举偏移：依赖偏移可以支持枚举偏移关系的填写。基本格式为{-24hour,-12hour}。已当前时间点前推每一个时间点，查看各个时间点上的是否有对应实例及对应实例的状态。

表九 枚举偏移说明

字段	说明
区间表示符号	支持大括号{}
支持的时间单位	分钟：min 小时：hour 天：day 星期：week 月：month
数字限制	数字需均为非正数的整数。

表十 常用枚举偏移示例

区间迁移	依赖解释
{-24hour,-12hour}	以当前时间点前推24小时和前推12小时，查看这两个时间点的作业的实例状态。
{-10min,-5min}	以当前时间点前10分钟和前推5分钟，查看这两个时间点的作业的实例状态。

依赖串并行

支持作业设置依赖的串并行功能。依赖上一周期，设置为是则表示作业为串行模式，设置为否则表示作业为并行模式，需要用户设置最大并行数。

并行设置如果为2，则表示同一个时刻只允许有两个作业实例同时运行。如果并行设置为1，则效果和串行一致，同一时刻只允许一个作业实例同时运行。

失败策略

当作业内某个作业节点失败时，指定失败策略。

继续：一个节点失败的时候，当前在运行节点继续运行直至完成。

终止：一个节点失败的时候，当前在运行节点停止。

数据架构

主题设计

主题设计是通过简单的层级结构将繁杂的平台数据按照业务主题进行合理划分，帮助厘清业务数据，做好数据的分类整理与聚合。

功能介绍

主题设计可以对数据进行分类管理，支持查询、新建、编辑、删除等功能。以下只对主题设计功能进行简要介绍，具体操作介绍详见主题域/主题管理。

表一 功能介绍详细说明

功能名称	说明
查询	在搜索框内，支持搜索主题域/主题中包含的关键词，方便快捷地查找数据所在的主题。
新建	支持多个创建主题的入口，方便创建主题域/主题。
编辑	支持对主题域/主题的基本信息进行修订。
删除	支持对没有数据的主题域/主题进行删除。

核心概念 在主题设计中，系统默认设计两个层级对数据进行划分，分别为“主题域”和“主题”。

主题域：主题域是对同一性质或同一类型的主题进行进一步抽象划分的域，包含若干个联系紧密的主题。

主题：主题是根据业务场景或数据所表达业务的中心概括。

举例：对于“销售线上营业额”、“销售线下营业额”、“销售现金流”等数据，可以设置名为“销售营业额”的主题。当主题较多不易管理时，可以将多个相关的主题归类为高一级的主题域。比如：当对于“材料成本费用”、“研发费用”、“员工工资费用”等相关的主题又可以抽象归类为“财务管理”主题域。

应用场景 主题设计主要针对于庞大且复杂的数据场景，用户可通过主题设计对业务场景数据进行中心概括整理，帮助用户清晰地梳理数据资产；并按照简单层级对数据进行整理分类，快速建立数据资产目录树，实现数据资产一目了然。

主题域/主题管理

从左侧导航栏进入数据架构>主题，默认只有全部节点，新建的主题域都在全部节点下。

新建主题域

单击主题域右侧图标，选择“新建主题域”。配置主题域的相关参数，相关参数描述如表一所示。填写完毕后点击确认，完成主题域的创建。

表一 主题域参数说明

参数	描述
主题域名称	必填。主题域的名称，仅支持中文、英文、数字、下划线、中划线，长度不能超过64个字符。
主题域编码	必填。主题域的编码，仅支持英文、数字、下划线、中划线，长度不能超过64个字符。
主题域描述	主题域的编码，仅支持英文、数字、下划线、中划线，长度不能超过64个字符。

新建主题

新建主题域之后，可创建主题域下的主题。系统提供三种创建途径。一是点击主题域右侧图标；二是点击具体主题域右侧（三点）图标中的“新建主题”。三是点击主题域名称，主题域详情页面右上角的“新建”按钮。后两种方式的默认主题域为当前的选择的主题域。

1. 单击**新建主题**，配置主题的相关参数，主题的相关参数说明如表二所示。填写完毕后点击确认，完成主题的创建。

表二 主题参数说明

参数	描述
主题名称	必填。主题的名称，仅支持中文、英文、数字、下划线、中划线，长度不能超过64个字符。
主题编码	必填。主题的编码，仅支持英文、数字、下划线、中划线，长度不能超过64个字符。
所属主题域	必填。当前主题的所属主题域名称，仅可选择已创建好的主题域。
主题描述	主题的业务描述，长度不能超过500个字符。

🔗 主题/主题域操作

- 查找：支持通过关键词对主题域/主题名称进行筛选，您可通过左上角的输入框输入关键词进行搜索。
- 编辑：支持对主题域/主题的相关参数信息进行修改，您可点击对应的主题域/主题，点击右侧（三点）或者详情页面右上角的“编辑”按钮，对其进行修改。
- 删除：支持对主题域/主题进行删除，您可点击对应的主题域/主题，点击右侧（三点）或者详情页面右上角的“删除”按钮，对其进行删除。

注意： 1.当主题域下拥有主题时，不允许删除主题域。
2.当主题下拥有数据模型时，不允许删除主题。数据模型相关介绍，详见（数据建模）。

数据建模

🔗 概述

数据建模是一种分析与设计方法，用于定义分析数据需求，设计满足需求的数据结构，是反应数据需求和设计的数据说明与相关模型图的集合。大数据的数仓建模正是通过建模的方法，更好的组织、存储数据，以便在性能、成本、效率和数据质量之间找到最佳平衡点。

数据建模的好处：

- 企业业务庞大，数据结构复杂，事前规划可以使数据管理更加规范、有序，便于后期进行问题定位；
- 强化事前治理，为数据标准落地提供抓手，保证平台的建表和数据的产出，符合数据标准和数据质量的要求。

功能介绍 数据建模支持数仓规划设计，通过维度建模（维度表、事实表、汇总表、应用表），构建出业务的数据模型，通过数据标准和指标对模型结构进行统一的描述和约束。根据不同的存储引擎与模型定义生成DDL，对建模进行物化生成实际的物理表，利用模型之间的关系以及指标定义自动生成DML用于数据生产，并且根据数据标准生成数据质量作业对物理表的数据进行质量监控。

除此之外，还将建模数据同步给数据资产和数据血缘用于生成相关的统计信息。

数据建模各部分模块功能介绍如表一所示。

表一 数据建模模块功能介绍

名称	描述
数仓规划	数仓规划支持根据业务场景或业务数据设计数据的分层，使得数据的整个生命周期层次清晰，依赖直观。综合考虑当前应用广泛的数仓分层策略，EDAP默认将数仓划分为以下五层： · 数据引入层 ODS （Operational Data Store） · 数据关系层DWR （Data Warehouse Relationship） · 明细数据层 DWD （Data Warehouse Detail · 汇总数据层 DWS （Data Warehouse Summary） · 应用数据层 ADS （Application Data Service）
关系建模	关系建模支持将各个系统中的数据以整个企业角度按主题进行相似性组合和合并，并进行一致性处理，为数据分析决策服务。采用范式，降低数据冗余，保证数据的一致性。 · 第一范式：属性不可切割； · 第二范式：不能存在部分函数依赖； · 第三范式：不能存在传递函数依赖。
维度建模	第一范式：属性不可切割； 第二范式：不能存在部分函数依赖； 第三范式：不能存在传递函数依赖。维度建模支持根据数仓分层建立对应的数据模型，通过简单易上手的界面设计，对数据进行精确建模逻辑模型，通过物化与作业开发落地物理模型。根据数仓分层设计，EDAP将维度建模分为以下四种表： · 维度表 （Dimension Table） · 事实表 （Fact Table） · 汇总表 （Summary Table） · 应用表 （Application Table）
模型审核	模型审核支持通过审核流程，提升数据模型、数据指标创建、变更、废弃的严肃性。数据模型的发布状态分为以下三种： · 未发布 · 已发布 · 已废弃

核心概念

表二 核心概念说明

名称	概念说明
数据模型	数据模型是数据特征的抽象，它从抽象层次上描述了系统的静态特征、动态行为和约束条件，为数据库系统的信息表示与操作提供一个抽象的框架。数据模型是指现实世界数据特征的抽象，用于描述一组数据的概念和定义。数据模型管理是指在信息系统设计时，参考逻辑模型，使用标准化用语、单词等数据要素设计数据模型，并在信息系统建设和运行维护过程中，严格按照数据模型管理制度，审核和管理新建和存量的数据模型。数据模型所描述的内容有三部分，分别是数据操作、数据约束和数据结构。
数据操作	数据模型中数据操作主要描述在相应的数据结构上的操作类型和操作方式。
数据约束	数据模型中的数据约束主要描述数据结构内数据间的语法、词义联系、它们之间的制约和依存关系，以及数据动态变化的规则，以保证数据的正确、有效和相容。
数据结构	数据模型中的数据结构主要描述数据的类型、内容、性质以及数据间的联系等。数据结构是数据模型的基础，数据操作和约束都建立在数据结构上。
数仓分层	数仓分层是根据业务数据的场景，结合现有应用广泛的数仓理论，将数据按照层次进行划分，使得数据之间的依赖关系更加清晰，数据的生产与加工有迹可循。
关系建模	关系建模用实体加关系描述的数据模型描述企业业务架构，在范式理论上符合3NF，是站在企业角度，而不是针对某个具体业务流程，它更多是面向数据的整合和一致性治理。建模过程先设计概念模型，再进一步细化设计出逻辑模型，最后设计物理模型。
维度建模	维度建模以分析决策的需求为出发点构建模型，一般有较好的大规模复杂查询的响应性能，更直接面向业务。因此，它重点解决用户如何更快速完成分析需求，同时还有较好的大规模复杂查询的响应性能。

🔗 数仓规划

随着数据的增长，数据治理的难度不断增大。大多数情况下，数据体系是依赖复杂、层级混乱的。如何使得数据能够有秩序地流转，数据的整个生命周期能够清晰明确被感知到。数仓规划是其中重要的一环，通过数仓进行合理科学分层，能够较为有效地梳理数据之间的依赖关系，使得数据清晰明了。EDAP中数仓分层详细信息如表三所示。

表三 数仓分层详细信息

分层类型	中文名称	英文名称	分层描述	模型类型	命名前缀
ODS	数据引入层	ODS	存放未经处理的原始数据，表结构与源系统保持一致。	贴源表	ods_
DWR	数据关系层	DWR	对ODS层的数据进行清洗，可以采用关系建模的方式维护数据关联关系。	逻辑实体表	dwr_
DWD	明细数据层	DWD	对ods层的数据进行清洗，采用维度退化的方式对数据重新建模。	维度表	dim_
				事实表	fact_
DWS	汇总数据层	DWS	按照业务划分和指标需求，对dwd层的数据进行汇总。	汇总表	dws_
ADS	数据应用层	ADS	存放个性化的统计指标，用于报表输出。	应用表	ads_

ODS：保存最原始的数据，起到数据备份的作用，可以采用parquet等压缩格式；

DWR：维护原始数据间的关联关系，可以使用关系建模的方法重新建模；

DWD：代表的是业务的最小粒度层，对业务数据采用维度模型的方式重新建模（维度退化），为后续的层做准备；

DWS：关于各个主题的加工和使用，是事实表的聚合值，例如统计各个主题对象7天、30天、90天的行为，例如把主题对象单日的行为聚合起来组成一张多列宽表，关联维度信息后，从不同角度进行统计分析；

ADS：为各种统计报表提供数据。

数仓分层

1. 单击导航栏中的**数据架构-数据建模-数仓规划**，进入数仓规划界面。

数仓规划分为五层，分别是数据引入层、数据关系层、明细数据层、汇总数据层以及数据应用层。统计属性介绍如下：

- 数据表：当前分层下所有表的数量
- 字段：当前分层下所有表的字段数量总和
- 落标率：引用了数据标准的字段/字段总数（枚举维度落标率为100%）

2. 单击右侧的编辑按钮，可以对数仓分层信息进行修改。

数仓分层的参数信息如表四所示。

表四 数仓分层参数信息

参数	描述
中文名称	数仓分层的中文名称，仅支持中文、数字、下划线、中划线，长度不能超过64个字符。
英文名称	数仓分层的英文名称，仅支持英文、数字、下划线、中划线，长度不能超过64个字符。
描述	数仓分层的描述，长度不能超过500个字符。
模型定义	每个数仓分层都有对应的数据模型，支持对模型的命名前缀进行修改。

3. 单击新建表按钮，可以跳转到对应的维度建模界面创建模型。

🔗 关系建模

数仓的ODS层和DWR层均支持关系建模。

新建逻辑实体

- 单击导航栏中的**数据架构-数据建模-关系建模**，目前支持贴源表和关系表两项列表界面。其中，贴源表(ODS)列表存放ODS层的逻辑实体，关系表(DWR)列表存放DWR层的逻辑实体。
- 选择对应的主题域以及主题。
- 单击左上角的“新建逻辑实体”按钮进入新建页，以DWR层下新建逻辑实体为例。
- 填写基本信息部分：“逻辑实体名称”、“表中文名称”和“模型责任人”为必填项，也可根据实际需求选填“描述”项填写框。

表五 基本信息说明

字段名称	说明
数仓分层	选择关系模型所属数仓层级，不可编辑，默认为数仓分层。
所属主题领域	选选择关系模型归属的主题域，不可编辑，默认为所选择的主题域名称/主题。
逻辑实体名称	逻辑实体的名称，默认为请输入，支持英文、数字、下划线，不超过128个字符。
表中文名称	数据表中文名称，默认为请输入，支持中文、英文、数字、中划线、下划线，不超过128个字符。
模型责任人	选择关系模型的责任人，默认为请选择，下拉选择负责用户。
描述	填写逻辑实体的描述信息，默认为请输入，支持中文、英文、数字、常见的特殊字符，不超过500个字符。

- 填写属性定义部分，“字段定义”中除“描述”外所有项均为必填项。填写“字段名称”、“中文名称”，选择“字段类型”后，可点击蓝色关联按钮以关联标准，可在勾选框选择是否将该字段设为主键和是否允许为空，选择密级别并选填“描述”项。

若将某字段设为主键，则该字段不可勾选“允许为空”。

表六 字段定义

字段名称	说明
字段名称	默认为请输入，支持英文、数字、下划线，不超过128个字符。
中文名称	默认为请输入，支持中文、英文、数字、中划线、下划线，不超过128个字符。
字段类型	默认为STRING，支持多种数据类型。
类型设置	默认为空。
关联标准	列关联的数据标准，默认不关联。
允许为空	列是否允许为空，默认不勾选为不允许，勾选为允许。
密级别	列的密级，默认为LO，下拉选择可选的密级。
描述	默认为请输入，支持中文、英文、数字、常见的特殊字符，不超过500个字符。
操作	可对列执行的操作，支持删除。

- 可在分区定义部分，单击蓝色按钮添加分区定义，必填项内容与字段定义中相同。

表七 分区定义

字段名称	说明
字段名称	默认为请输入，支持英文、数字、下划线，不超过128个字符。
中文名称	默认为请输入，支持中文、英文、数字、中划线、下划线，不超过128个字符。
字段类型	默认为STRING，支持多种数据类型。
类型设置	默认为空。
关联标准	列关联的数据标准，默认不关联。
设为主键	默认不勾选为否，勾选为是。
允许为空	默认不勾选为不允许，勾选为允许。
密级别	默认为LO，下拉选择可选的密级。
描述	默认为请输入，支持中文、英文、数字、常见的特殊字符，不超过500个字符。
操作	可对列执行的操作，支持删除。

7. 在关系定义部分中，单击**添加关系**按钮，打开添加逻辑关系弹窗页。在弹窗页内先选择“关联逻辑实体”，若选中的关联逻辑实体没有主键字段，则无法完成关联，需重新选择关联逻辑实体。选中合适的关联逻辑实体后，其主键字段回自动匹配至下方。随后，在弹窗页下拉选框选择“关联关系”，选中的关联关系与关联逻辑实体的主键字段一一对应。

注：子逻辑实体和父逻辑实体相同的关系不可重复填报。

8. 添加关联关系后，可在“关系定义”部分查看到示意的关联图标。

9. 点击页面左下角确认按钮，完成逻辑实体的新建操作。

查看E-R图

点击查看E-R图即可查看实体-关系图。

注：ODS层不支持查看E-R图。

维度建模

创建维度表

- 1. 单击导航栏中的**数据架构-数据建模-维度建模-维度表**。
- 2. 选择对应的主题域以及主题。
- 3. 单击左上角的“新建表”按钮创建维度表。
- 4. 根据新建维度表界面的提示填写参数，单击确认完成维度表的创建。维度表参数如表八所示。
- 5. 在维度表详情界面，单击查看关联关系，查看所有跟当前维度表有关联的事实表。

表八 维度表参数

参数	描述
表英文名称	维度表英文名称，支持英文、数字、下划线，长度不能超过128个字符。
表中文名称	维度表中文名称，支持中文、英文、数字、特殊字符等，长度不能超过128个字符。
维度类型	维度类型，支持普通维度和枚举维度。
模型责任人	模型责任人名称。
描述	维度表的描述，长度不能超过500个字符。
字段定义	维度表字段参数定义，详见表九。
分区定义	维度表分区参数定义，详见表九。

表九 维度表字段定义与分区定义参数

参数	描述
字段名称	字段的名称，以英文开头，支持英文、数字、下划线，以字母开头，长度不能超过128个字符。
中文名称	字段的中文名称，支持中文、英文、数字、常见的特殊字符等，不超过128个字符。
数据类型	字段的数据类型，支持TINY、SMALLINT、VARCHAR、CHAR、ARRAY、MAP、STRING、INT、BIGINT、DOUBLE、DECIMAL、FLOAT、BOOLEAN、BYTES、DATE、TIMESTAMP。
类型设置	当数据类型为DECIMAL时，需要填写类型设置，用于精度确认。
关联标准	字段支持关联数据标准，用于规范字段。
代理键	表示维度的唯一键，作用是连接维度表和事实表，为避免数据库结构发生变化带来影响。
允许空	是否允许为空。
密级	字段保密程度等级，支持L0-L4。
描述	字段描述，长度不能超过64个字符。

创建事实表

- 1.单击导航栏中的**数据架构>模型>维度建模>事实表**。
- 2.选择对应的主题域以及主题。
- 3.单击左上角的**“新建表”**按钮创建事实表。
4. 根据新建事实表界面的提示填写参数，单击确认完成事实表的创建。事实表参数如表十所示。

表十 事实表参数

参数	描述
表英文名称	表英文名称，支持英文、数字、下划线，长度不能超过128个字符。
表中文名称	表中文名称，支持中文、英文、数字、特殊字符等，长度不能超过128个字符。
模型责任人	模型责任人名称。
描述	表的描述，长度不能超过500个字符。
字段定义	表字段参数定义，详见表十一。
分区定义	表分区参数定义，详见表十一。

表十一 事实表字段定义与分区定义参数

参数	描述
字段类型	字段的类型，支持维度、度量。
字段名称	字段的英文名称，以英文开头，支持英文、数字、下划线，长度不能超过128个字符。
中文名称	字段的中文名称，支持中文、英文、数字、常见的特殊字符等，不超过128个字符。
数据类型	字段的数据类型，支持TINY、SMALLINT、VARCHAR、CHAR、ARRAY、MAP、STRING、INT、BIGINT、DOUBLE、DECIMAL、FLOAT、BOOLEAN、BYTES、DATE、TIMESTAMP。
类型设置	当数据类型为DECIMAL时，需要填写类型设置，用于精度确认。
关联标准	字段支持关联数据标准，用于规范字段。
关联维度	字段支持关联维度，维度表的代理键字段与当前字段进行关联。
代理键	表示维度的唯一键，作用是连接维度表和事实表，为避免数据库结构发生变化带来影响。
允许空	是否允许为空。
密级	字段保密程度等级，支持L0-L4。
描述	字段描述，长度不能超过64个字符。

创建汇总表

- 1.单击导航栏中的**数据架构>建模>维度建模>汇总表**。
- 2.选择对应的主题域以及主题。
- 3.单击左上角的“新建表”按钮创建汇总表。
- 4.根据新建汇总表界面的提示填写参数，汇总表参数如表十二所示。
- 5.单击元数据定义右上角的**从指标导入**（须选择维度后可操作），通过导入的方式填写字段定义与分区定义。
- 6.最后点击确认完成汇总表的创建。

表十二 汇总表参数

参数	描述
表英文名称	表英文名称，支持英文、数字、下划线，长度不能超过128个字符。
表中文名称	表中文名称，支持中文、英文、数字、特殊字符等，长度不能超过128个字符。
统计维度	所有“已发布”的衍生指标的所有维度组合，例如：员工维度、公司维度、员工和公司维度。
模型责任人	模型责任人名称。
描述	表的描述，长度不能超过500个字符。
字段定义	表字段参数定义，详见表十三。
分区定义	表分区参数定义，详见表十三。

表十三 汇总表字段定义与分区定义参数

参数	描述
字段名称	字段的英文名称，以英文开头，支持英文、数字、下划线，长度不能超过128个字符。
中文名称	字段的中文名称，支持中文、英文、数字、常见的特殊字符等，不超过128个字符。
数据类型	字段的数据类型，支持TINY、SMALLINT、VARCHAR、CHAR、ARRAY、MAP、STRING、INT、BIGINT、DOUBLE、DECIMAL、FLOAT、BOOLEAN、BYTES、DATE、TIMESTAMP。
类型设置	当数据类型为DECIMAL时，需要填写类型设置，用于精度确认。
关联标准	字段支持关联数据标准，用于规范字段。
关联维度/指标	字段若是维度，则关联维度；字段若是度量指标，则关联指标。
代理键	表示维度的唯一键，作用是连接维度表和事实表，为避免数据库结构发生变化带来影响。
允许空	是否允许为空。
密级	字段保密程度等级，支持LO-L4。
描述	字段描述，长度不能超过64个字符。

关联维度/指标

选择了统计维度才可配置。

- 关联类型：统计维度、衍生指标、复合指标；
- 关联对象：如果是统计维度，先选维度，后选字段；如果是指标，直接选择指标即可。

创建应用表

1. 单击导航栏中的**数据架构>建模>维度建模>应用表**。
2. 选择对应的主题域以及主题。单击左上角的**新建表**按钮创建应用表。
3. 根据新建应用表界面的提示填写参数。应用表参数与汇总表保持一致。
4. 单击元数据定义右上角的**从指标导入**，通过导入的方式填写字段定义与分区定义。
5. 最后单击确认完成汇总表的创建。具体过程可参考**创建汇总表**，本节不进行赘述。

通用操作

发布与管理表 数据模型生命周期分为三个状态，分别是**未发布**、**已发布**、**已废弃**。

- **未发布**：表示数据模型处于草稿状态，未发布的数据模型无法被使用，用户可以对模型进行编辑、发布、删除操作。
- **已发布**：表示数据模型处于已发布状态，已发布的数据模型可以被使用，用户可以对模型进行编辑、物化、废弃操作。
- **已废弃**：表示数据模型处于已废弃状态，已废弃的数据模型无法被使用，用户可以对模型进行删除操作。

发布 新建的数据模型处于**草稿**状态，需要发布之后才可以被正式使用。在右侧的操作栏单击“发布”按钮，即可进行发布操作。发布后，数据模型变更为**已发布**状态。**废弃** 对于**已发布**的数据模型，如果不再使用，则可以进行**废弃**。单击右侧操作栏中的“废弃”按钮，可以对指标进行**废弃**操作。

当模型被其他模型或指标引用时，不支持废弃操作。

编辑

- 1. 当数据模型处于“未发布”状态时，单击“编辑”按钮，可以对数据模型的信息进行编辑修改，单击确认之后，保存用户修改之后的模型。
- 2. 当数据模型处于“已发布”状态时，单击“编辑”按钮，可以对数据模型的信息进行编辑修改，单击确认之后，会创建当前模型的子版本。子版本存储用户修改之后的模型，并处于“未发布”状态，支持发布、编辑、发布操作。当前父版本的信息不会改变，且父版本不支持任何操作。

当子版本发布之后，会覆盖父版本的模型，并处于“已发布”状态。

删除 当模型处于“未发布”或“已废弃”状态时，单击“删除”按钮，可以对模型进行删除。

模型物化 当模型处于“已发布”状态时，支持对模型进行物化操作，为数据模型创建同名物理表。

批量物化

- 1.选中“已发布”状态的一个或多个模型，单击“批量物化”按钮展开下拉选项，选择“批量物化”，并配置物化参数，可对模型进行批量物化。

如果初次物化，所选库中已有同名的数据表，则物化失败。

用户可根据需要配置物化信息，物化信息详细介绍如表十四所示。

表十四 物化配置信息介绍

参数	描述	交互	是否必填
数据源类型	数据源的类型，支持EDAP、Hive和Doris	下拉单选	是
数据源	取自数据源管理列表	下拉单选	是
数据库	取自数据源管理列表	下拉单选	是
表类型	仅展示		
副本数	默认值为3	文本框输入	是
物化类型	增量发布和删除重建	单选	是

- 2.完成物化配置之后，单击下一步，可展示物化结果。
- 3.当发布状态为“成功”时，表示物化成功。当发布状态为“失败”时，可单击“重试”按钮或者修改相关配置重新物化；支持右下角单击全部重试按钮对“失败”模型批量重试。右下角单击确定按钮结束物化操作。
- 4.物化完成后，“同步状态”会更新为“同步成功”，鼠标焦点放在“同步成功”后面的眼睛图标上时，可点击气泡中“查看元数据”蓝字按钮，跳转至元数据管理页面查看详情（数据源类型为Doris，不显示“查看元数据”）。
- 5、单击“逻辑实体名称”蓝字按钮，跳转至模型详情页，详情页内展示逻辑实体相关信息，包括：基本信息、元数据信息、物化表信息、关系定义信息、发布历史、审核历史。

上次物化结果

- 1.选中“已发布”状态的一个或多个模型，单击“批量物化”按钮展开下拉选项，选择“上次物化结果”，可查看上次物化结果。

物化场景

表十五 物化场景说明

EDAP Type类型	Doris Type类型	处理措施
TINYINT	TINYINT	
SMALLINT	SMALLINT	
INT	INT	
BIGINT	BIGINT	
不支持	LARGEINT	
BOOLEAN	BOOLEAN	
FLOAT	FLOAT	
DOUBLE	DOUBLE	
STRING	STRING	
CHAR	CHAR	
VARCHAR	VARCHAR	
BYTES	STRING	
TIMESTAMP	DATETIME	
DECIMAL(precision, scale)	DECIMAL(precision,scale)	
DATE	DATE	
ARRAY	ARRAY	doris1.2版本开始支持，仅Duplicate模型
MAP	不支持	doris2.0开始支持，仅Duplicate、Unique模型；遇到低版本，则报错提示，提示内容为Doris原始错误提示
STRUCT	不支持	doris2.0开始支持，仅Duplicate模型；遇到低版本，则报错提示，提示内容为Doris原始错误提示
不支持	HLL	
不支持	BITMAP	
不支持	JSON	
不支持	AGG_STATE	

逆向数据库

关系建模关系表（DWR）

- 1.选中任意主题，鼠标悬浮“逆向操作”按钮，展开“逆向数据库”和“上次逆向结果”选项。单击“逆向数据库”按钮，打开操作弹窗页。
- 2.单击“逆向数据库”按钮，打开操作弹窗页。首先选择数据源：选择“所属主题域”、“所属主题”、“资源类型”和“数据源链接”。

表十六 当前支持的数据源类型和版本

数据库类型	版本	对应客户的业务域
MySQL	8.0.26	标准规范
MySQL	5.7.38-log	知识产权、供应链
MySQL	8.0.25	主数据
Oracle	11.2.0.1.0	海外营销、市场域
Oracle	11.2.0.4.0	合约域
Doris	-	-

注：关系建模中，当前DWR层支持的数据源类型为MySQL和Oracle。

- 3.选择完成后可单击“立即测试”按钮，测试其连通性。测试结果无论通过与否都不影响提交。单击“下一步”按钮完成数据源的选择。
- 4.在数据表目录中勾选需要逆向操作的数据表，选择完成后单击“下一步”按钮。
- 5.进入“逆向建模”流程后，等待自动建模完成即可。完成后自动流转进入“建模结果”，可在表单中查看建模结果总览，点击“查看”蓝字按钮可跳转至逻辑实体详情页。
- 6.在“逆向操作”中单击“上次逆向结果”按钮，查看“上次逆向数据库”的主题及实体等。
- 7.同步成功的模型可以单击查看元数据。

关系建模贴源表(ODS)

1. 选中任意主题，鼠标悬浮“逆向操作”按钮，展开“逆向数据库”和“上次逆向结果”选项。单击“逆向数据库”按钮，打开操作弹窗页。
2. 单击“逆向数据库”按钮，打开操作弹窗页。首先选择数据源：选择“所属主题域”、“所属主题”、“数据源类型”和“数据源链接”。

注：关系建模中，ODS层支持的数据源类型为Hive和EDAP。

3. 选择完成后可单击“立即测试”按钮，测试其连通性。测试结果无论通过与否都不影响提交。单击“下一步”按钮完成数据源的选择。
4. 在数据表目录中勾选需要逆向操作的数据表，选择完成后单击“下一步”按钮。
5. 进行前缀替换配置，选中库/表后，加载不符合表名称规范的前缀，需要选择和配置；选中的前缀将会被替换为规范前缀，其余将默认向前追加前缀。

选中表名称后，可以对表名称进行快速修改，修改后必须满足前缀规范，否则报错。

6. 进入“逆向建模”流程后，等待自动建模完成即可。完成后自动流转进入“建模结果”，可在表单中查看建模结果总览，点击“查看”蓝字按钮可跳转至模型详情页。
7. 在“逆向操作”中点击“上次逆向结果”按钮，查看“上次逆向数据库”的主题及实体等。
8. 同步成功的模型可以点击查看元数据。

维度建模

关系建模中，ODS层支持的数据源类型为Hive、EDAP和Doris。

表类型：逆向时，后端自动识别模型类型为「Unique 唯一表」、「Aggregate 聚合表」、「Duplicate 明细表」。

逆向详情页

1. 首次逆向成功：详情页面显示逆向表信息，不显示物化表信息。
2. 模型发布之后：详情页面显示物化表信息、逆向表信息。

逆向状态说明

1. 首次逆向成功：
 - 发布状态：未发布；
 - 同步状态：未同步。
2. 模型发布之后：
 - 发布状态：已发布；
 - 同步状态：如果跟底层表名称相同，显示同步成功；如果不相同，显示未同步，可以重新物化进行同步；都可以看到“眼睛”，即都有底层物理表。

逆向场景

表十七 逆向场景说明

Doris Type类型	EDAP Type类型	处理措施
TINYINT	TINYINT	
SMALLINT	SMALLINT	
INT	INT	
BIGINT	BIGINT	
LARGEINT	STRING	
BOOLEAN	BOOLEAN	
FLOAT	FLOAT	
DOUBLE	DOUBLE	
STRING	STRING	
CHAR	CHAR	
VARCHAR	VARCHAR	
不支持	BYTES	
DATETIME	TIMESTAMP	
DECIMAL(precision, scale)	DECIMAL(precision,scale)	
DATE	DATE	
ARRAY	STRING	默认处理为STRING
MAP	STRING	
STRUCT	STRING	
HLL	STRING	
BITMAP	STRING	
JSON	STRING	
AGG_STATE	STRING	

模型开发

汇总表、应用表支持对“已发布”状态且已成功物化的模型进行作业开发，创建脚本开发作业，通过脚本作业导入数据，提高开发效率。

- 1. 选择“已发布”的汇总表或应用表，单击“开发”按钮，进入模型开发-作业配置界面。

表十八 作业开发配置信息介绍

参数	描述
作业开发类型	作业开发的类型，支持HiveSQL、SparkSQL。
存在项目	选择平台已创建的项目，用于作业开发。
存放文件夹	选择作业开发项目下的文件夹。

- 2. 完成作业配置之后，单击下一步，可展示作业开发结果。
- 3. 当创建状态为“SUCCESS”时，表示脚本作业创建成功，单击确定完成创建作业开发任务。用户可根据作业的名称到对应的作业开发任务查看详情。当发布状态为“FAIL”时，可点击“重试”按钮或者修改相关配置重新创建作业。

数据指标

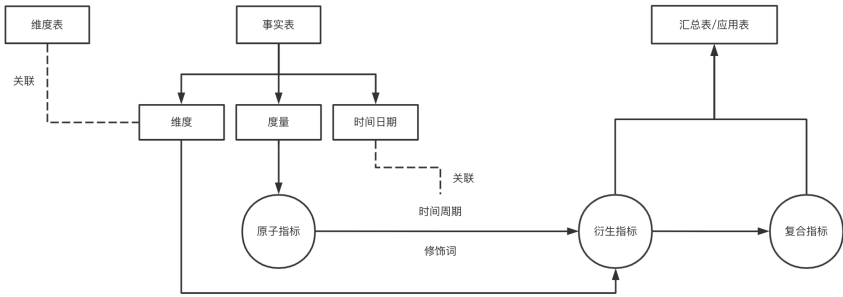
概述

功能介绍 EDAP提供数据指标功能，帮助企业建立统一的指标体系，对企业数据指标进行集中管理，避免多方操作导致指标混乱、口径不一致等问题。

核心概念

- 原子指标：原子指标中的度量和属性来源于多维模型中的维度表和事实表，与多维模型所属的业务对象保持一致，与多维模型中的最细数据粒度保持一致。用于明确业务的统计口径和计算逻辑，是基于用户的业务活动创建的，用于统计业务活动中某一业务状况的数值。例如，用户的业务活动为购买，则原子指标就可以指定为支付金额。
- 衍生指标：由原子指标、时间周期、修饰词构成，用于反映企业某一业务活动在指定时间周期及目标范围中的业务状况。例如，某企业近一周上海地域的销售金额。
- 复合指标：由一个或多个衍生指标叠加计算而成，其中的维度、限定均继承于衍生指标。

图一 模型和指标间的关系



📄

🔗 数据指标管理

新建原子指标

1. 在EDAP平台首页概览的左侧导航栏中选择**数据架构>指标**，进入后在标签页中选择「原子指标」，在左侧主题域的树形控件中选择主题后，单击**新建**按钮，配置原子指标的相关参数。
2. 在填写完计算逻辑后，可单击**校验**按钮对填写内容进行校验，如果校验通过，可以提交保存。如果校验不通过，则需要根据校验失败的提示信息调整计算逻辑。

表一 原子指标参数说明（*表示必填）

参数		描述
基础信息	*指标中文名称	支持中文、英文、数字、下划线，长度不能超过128个字符。
	*指标英文名称	以英文开头，支持英文、数字、下划线，长度不能超过128个字符。
	*指标编码	支持英文、数字、下划线、中划线，长度不能超过128个字符。
	业务定义	指标的描述信息，支持中文、英文、数字、特殊字符等，长度不能超过500个字符。
计算逻辑	*指标来源	单选，可选项为所有已发布的事实表。
	*计算公式	根据情况进行填写，支持HiveSQL中的函数及运算符，计算公式中的字段需要是指标来源中的度量字段。
	*数据类型	下拉选择：TINY、SMALLINT、VARCHAR、CHAR、ARRAY、MAP、STRING、INT、BIGINT、DOUBLE、DECIMAL、FLOAT、BOOLEAN、BYTES、DATE、TIMESTAMP 选中DECIMAL时可以配置长度和精度。
	*计量单位	支持中文、英文、数字、特殊字符等，长度不能超过64个字符。
管理属性	指标责任人	下拉选择用户。
	指标责任部门	支持中文、英文、数字、特殊字符等，长度不能超过64个字符。
*自定义属性	*自定义属性	每个自定义属性包括属性名称和内容，支持中文、英文、数字、特殊字符等，长度不能超过64个字符。

新建衍生指标

1. 进入数据指标模块，在侧边导航中选择**衍生指标**，在左侧主题域的树形控件中选择主题后，单击**新建**按钮，即可新建衍生指标。配置衍生指标的相关参数。
2. 在填写好计算逻辑之后，可以点击左上角的「指标试运行」校验填写的计算逻辑是否正确，并将会生成相应的SQL语句用于参考。如果计算逻辑填写错误，将会报错，用户可根据错误信息修改计算逻辑。
3. 各项指标参数填写完成后，单击**确认**，即可创建成功。

表二 衍生指标参数说明

参数		是否必填	描述
基础信息	指标中文名称	是	支持中文、英文、数字、下划线，长度不能超过128个字符。
	指标英文名称	是	以英文开头，支持英文、数字、下划线，长度不能超过128个字符。
	指标编码	是	支持英文、数字、下划线、中划线，长度不能超过128个字符。
	业务定义	否	指标的描述信息，支持中文、英文、数字、特殊字符等，长度不能超过500个字符。
计算逻辑	原子指标	是	单选，可选项为所有已发布的原子指标。
	时间周期	是	下拉选项：近1天、近7天、近30天、近180天、当天、当月、当年、自定义。 如果选择自定义，可以使用宏函数填写时间范围。
	关联字段	是	可选项为原子指标所在事实表里的维度、度量类型字段。
	统计维度	否	多选，可选项为原子指标所在事实表里的维度字段，支持搜索。
	修饰词	否	用来限定统计数据业务范围的，支持中文、英文、数字、特殊字符等，长度不能超过64个字符。
	修饰词表达式	否	修饰词的表达式，相当于SQL语句中的WHERE部分。
管理属性	指标责任人	否	下拉选择用户。
	指标责任部门	否	支持中文、英文、数字、特殊字符等，长度不能超过64个字符。
自定义属性	自定义属性	是	每个自定义属性包括属性名称和内容，支持中文、英文、数字、特殊字符等，长度不能超过64个字符。

新建复合指标

1. 进入数据指标模块，在侧边导航中选择**复合指标**，在左侧主题域的树形控件中选择主题后，单击**新建**按钮，即可新建复合指标。
2. 在填写好计算逻辑之后，可以点击左上角的**指标试运行**校验填写的计算逻辑是否正确，并将会生成相应的SQL语句用于参考。如果计算逻辑填写错误，将会报错，用户可根据错误信息修改计算逻辑。
3. 各项指标参数填写完成后，单击**确认**，即可创建成功。

表三 复合指标参数说明（*表示必填）

参数		描述
基础信息	*指标中文名称	支持中文、英文、数字、下划线，长度不能超过128个字符。
	*指标英文名称	以英文开头，支持英文、数字、下划线，长度不能超过128个字符。
	*指标编码	支持英文、数字、下划线、中划线，长度不能超过128个字符。
	业务定义	指标的描述信息，支持中文、英文、数字、特殊字符等，长度不能超过500个字符。
计算逻辑	*统计维度	单选，可选项为全局维度、所有主题的统计维度；选择全局维度时，衍生指标列表可选项为未关联维度的指标。选择非全局维度时，则必须选一个衍生指标。
	计算公式	支持选择当前维度下的衍生指标和所有原子指标。根据实际情况进行填写，公式的字段为衍生指标的英文名称，公式支持运算符包括： +、-、、/、（、）。
	*数据类型	下拉选择：TINY、SMALLINT、VARCHAR、CHAR、ARRAY、MAP、STRING、INT、BIGINT、DOUBLE、DECIMAL、FLOAT、BOOLEAN、BYTES、DATE、TIMESTAMP 选中DECIMAL时可以配置长度和精度。
	*计量单位	支持中文、英文、数字、特殊字符等，长度不能超过64个字符。
管理属性	指标责任人	下拉选择用户。
	指标责任部门	支持中文、英文、数字、特殊字符等，长度不能超过64个字符。
*自定义属性	*自定义属性	每个自定义属性包括属性名称和内容，支持中文、英文、数字、特殊字符等，长度不能超过64个字符。

导入/导出数据指标

将鼠标悬浮在主题域右侧的「…」处，即可弹出数据指标的**导入/导出**菜单，将会根据所选标签页决定**导入/导出**的数据指标类型。

1. 单击**导入指标**，展示导入指标的功能弹窗。
2. 单击**下载导入模版**，可以下载数据指标导入的 Excel 模版，不同类型的数据指标的导入模版也不同。
3. 按照模版的格式，填写导入的原子指标数据，将编辑好的 Excel 文件上传，单击选择文件，选择已经填写好的表格文件，即可进行导入，导入完成后，页面会展示导入结果。
4. **指标导出**：单击**指标导出**，可以导出指定主题下的所有数据指标。导出的数据指标为 Excel 格式。

🔗 指标状态管理

数据指标的生命周期包括以下状态：草稿、已发布、已废弃。通过发布、修订和废弃这三种操作，可以修改它们的状态。

各个状态的流转方式如下：

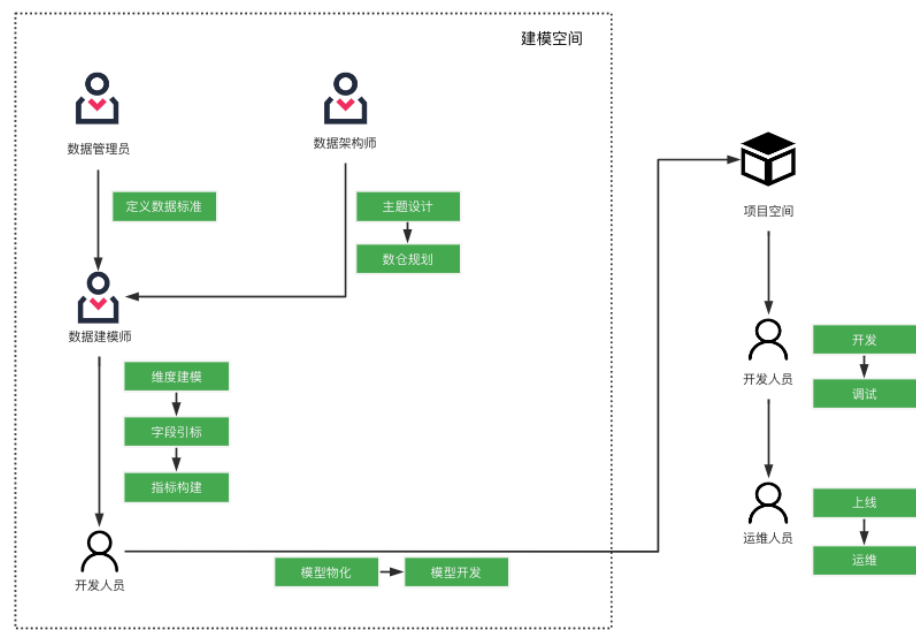
流转方式	描述
发布	新建的数据指标处于草稿状态，需要发布之后才可以被正式使用。在右侧的操作栏单击「发布」按钮，即可进行发布操作。发布后，指标变更为已发布状态。
修订	对于已发布的数据指标，单击右侧操作栏中的「编辑」按钮，可以对其进行修订操作。此时会跳转到数据指标的编辑页面，对数据指标的属性进行修改之后，单击提交，将会创建一个新的子版本。此时，在数据指标的左侧会出现一个可以点击的下三角按钮，单击该按钮，弹出该数据指标的子版本。
废弃	对于已发布的数据指标，如果不再使用，则可以进行废弃。单击右侧操作栏中的「废弃」按钮，可以对指标进行废弃操作。注意：只有未被引用的数据指标才可以被废弃，如果原子指标被衍生指标引用、衍生指标被复合指标引用、指标被数据模型引用，都需要解除引用关系之后才可以进行废弃。
删除	对于已废弃的数据指标，可以进行删除。单击右侧操作栏中的「删除」按钮，可以对指标进行删除操作。

数据标准

EDAP支持在数据建模前规划制定企业数据标准，包括基础数据标准和代码数据标准两类，提供灵活可配置的向导模式，以及标准的批量导入导出，同时提供数据标准的发布、废弃、审核、删除、修订的全生命周期管理能力。用户可在后续的数据建模过程中对数据标准进行引用，保障后续开发过程中数据口径的一致性，从源头上避免数据质量问题的产生。

- 数据标准：数据标准是和业务相关的、用于描述企业层面需共同遵守的数据含义和业务规则。通过数据标准，可以指定其相关的字段类型、取值范围、度量单位等内容。在业务数据系统中，不同表的多个字段可以与同一个数据标准进行关联，从而约束它们遵守相同的规范。
- 公共代码：公共代码是数据标准的取值范围，在公共代码中可设置某一数据标准可选择的数据的内容以及范围。例如性别数据标准的公共代码码内容应该为男或女。

应用场景



角色、职责与权限

角色	职责	权限说明
系统管理员	负责数据标准发布、废弃的审核	数据标准审核
数据管理员	负责数据标准的定义、发布、管理	数据标准新建 数据标准发布、废弃、修订、删除
普通用户	根据已发布的数据标准，进行数据标准落标引用	数据标准引用

🔗 目录管理

数据标准与公共代码均需要在创建在指定的目录下，两者的目录功能相同，但数据彼此独立。

默认情况下，只有「全部」一个根目录，该目录下不允许创建数据标准/公共代码。在创建数据标准/公共代码前，首先需要新建子目录。对于子目录来说，还可以创建下级子目录。

注意：

- 目录的名称具有局部唯一性，即同一个父目录下不能有名称相同的子目录
- 目录最多不能超过 5 层

新建目录 单击目录侧边栏上方的添加按钮，即可新建目录。

表一 目录参数说明

参数	描述
目录名称	必填。目录的名称，仅支持中文、英文、数字、下划线、中划线，长度不能超过64个字符
目录路径	必填。目录所属的路径，层级不能超过5层

重命名目录

右键点击指定的目录，可以对目录进行重命名。

删除目录

右键点击指定的目录，可以对目录进行删除操作。只有当目录为空时，目录才允许删除，否则，删除选项置灰。

目录为空指的是该目录下即没有数据，也没有子目录。

🔗 数据标准管理

新建数据标准

1. 进入**数据标准**模块，在上方的标签页中选择**标准管理**，选中目录后，即可在该目录下创建数据标准（选中“全部”根目录时除外，该目录下不允许创建数据标准，此时**添加标准**按钮置灰。
2. 在编辑页面中，配置相关参数，参数说明如表二。

表二 数据标准参数说明（*表示必填）

参数		描述
基础信息	*中文名称	标准中文名称，支持中文、英文、数字、特殊字符等，长度不能超过128个字符
	*英文名称	标准英文名称，支持英文、数字、下划线、中划线，长度不能超过128个字符
	*标准编码	标准英文编码，支持英文、数字、下划线、中划线，长度不能超过128个字符
技术属性	*数据类型	下拉选择：TINY、SMALLINT、VARCHAR、CHAR、ARRAY、MAP、STRING、INT、BIGINT、DOUBLE、DECIMAL、FLOAT、BOOLEAN、BYTES、DATE、TIMESTAMP 选中DECIMAL时可以配置长度和精度
	数据长度	可以选择固定值或范围，输入值为正整数
	允许为空	是否允许为空
	是否引用码表	是否引用公共代码，如果是，则可以下拉选择引用的公共代码
业务属性	值域范围	数据标准的取值范围，取值为一个区间，可以选择是开区间或者闭区间，左值或右值如果为空，则表示负无穷或正无穷
	业务定义	支持中文、英文、数字、常见的特殊字符等，长度不能超过500个字符
	业务规则	支持中文、英文、数字、常见的特殊字符等，长度不能超过500个字符
	引用文档	支持中文、英文、数字、常见的特殊字符等，长度不能超过500个字符
	描述	支持中文、英文、数字、特殊字符等，长度不能超过500个字符
管理属性	责任人	数据标准的责任人，下拉选择用户
	责任部门	数据标准的责任部门，支持中文、英文、数字、特殊字符等，长度不能超过64个字符
自定义属性	自定义属性	用户最多可填写三个自定义属性，每个自定义属性包括属性名称和内容，支持中文、英文、数字、特殊字符等，长度不能超过64个字符

查看数据标准详情

1. 单击标准中文名称，可以跳转到标准详情页面。
2. 标准详情页面展示包括标准的基本信息、引用记录、发布历史、审核历史等数据。

编辑数据标准

- 只有处于草稿、已驳回和已发布状态的数据标准，才支持编辑操作。

- 对处于草稿的数据标准，点击编辑按钮，会跳转到编辑页面，修改数据标准的属性之后点击提交，即可完成编辑，编辑后的数据标准仍然为草稿状态。
- 对处于已驳回的数据标准，点击编辑按钮，会跳转到编辑页面，修改数据标准的属性之后点击提交，即可完成编辑，编辑后的数据标准变化为草稿状态。
- 对处于「已发布」状态的数据标准进行编辑，见修订。

删除数据标准

- 只有处于「草稿」和「已驳回」状态的数据标准，才支持删除操作。
- 对处于「草稿」和「已驳回」状态的数据标准，单击删除按钮，可以对数据标准进行删除。

说明：删除数据标准时，该标准的发布记录和审核记录将会同步删除。

导入/导出数据标准

将鼠标悬浮在主题域右侧的「...」处，即可弹出数据指标的导入/导出菜单，将会根据所选标签页决定导入/导出的数据指标类型。

1. 单击导入指标，展示导入指标的功能弹窗。
2. 单击下载导入模版，可以下载数据指标导入的 Excel 模版，不同类型的数据指标的导入模版也不同。
3. 按照模版的格式，填写导入的原子指标数据，将编辑好的 Excel 文件上传，单击点击选择文件，选择已经填写好的表格文件，即可进行导入，导入完成后，页面会展示导入结果。
4. 指标导出：单击指标导出，可以导出指定主题下的所有数据指标。导出的数据指标为 Excel 格式。

公共代码管理

新建公共代码

1. 进入「数据标准」模块，在上方的标签页中选择「公共代码」，选中目录后，即可在该目录下创建公共代码（选中“全部”根目录时除外，该目录下不允许创建数据标准，此时「添加代码」按钮置灰）：
2. 在编辑页面中，配置如下内容。

表三 公共代码参数说明（*表示必填）

参数		描述
基本信息	*代码中文名称	支持中文、英文、数字、特殊字符等，长度不能超过128个字符。
	*代码英文名称	支持英文、数字、下划线、中划线，长度不能超过128个字符。
	*代码编码	仅支持英文、数字、下划线、中划线，长度不能超过128个字符。
	描述	支持中文、英文、数字、特殊字符等，长度不能超过500个字符。
代码表配置	*字段名称	代码字段的名称，以英文开头，支持英文、数字、下划线，长度不能超过128个字符。
	*字段中文名称	代码字段的中文名称，支持中文、英文、数字、中划线、下划线，长度不能超过128个字符。
	*字段类型	代码字段的数据类型，下拉选择：TINY、SMALLINT、VARCHAR、CHAR、ARRAY、MAP、STRING、INT、BIGINT、DOUBLE、DECIMAL、FLOAT、BOOLEAN、BYTES、DATE、TIMESTAMP。
	字段描述	代码字段的描述信息，支持中文、英文、数字、特殊字符等，长度不能超过500个字符。

填写数值

1. 创建公共代码之后，点击公共代码右侧操作栏的「填写数值」按钮，可以对公共代码的数值进行编辑。填写好数值之后，点击确定，就可以保存公共代码的数值。
2. 如果想更新数值，只需再次点击「填写数值」，对其中的数据进行编辑，再次点击确定，就可以更新公共代码的数值。

查看公共代码详情

1. 单击代码中文名称，可以跳转到代码详情页面。
2. 代码详情页面展示包括公共代码的基本信息、代码表字段、引用记录、发布历史、审核历史等数据。

编辑公共代码

- 只有处于「草稿」、「已驳回」和「已发布」状态的公共代码，才支持编辑操作。
- 对处于「草稿」的公共代码，点击「编辑」按钮，会跳转到编辑页面，修改公共代码的属性之后点击「提交」，即可完成编辑，编辑后的公共代码仍然为「草

稿」状态。

- 对处于「已驳回」的公共代码，点击「编辑」按钮，会跳转到编辑页面，修改公共代码的属性之后点击「提交」，即可完成编辑，编辑后的公共代码变化为「草稿」状态。
- 对处于「已发布」状态的公共代码进行编辑，见「修订」。

删除公共代码

只有处于「草稿」和「已驳回」状态的公共代码，才支持删除操作。

对处于「草稿」和「已驳回」状态的公共代码，点击「删除」按钮，可以对公共代码进行删除。

说明：删除公共代码时，该公共代码的发布记录和审核记录将会同步删除。

导入/导出公共代码

将鼠标悬浮在主题域右侧的「...」处，即可弹出公共代码的导入/导出菜单。

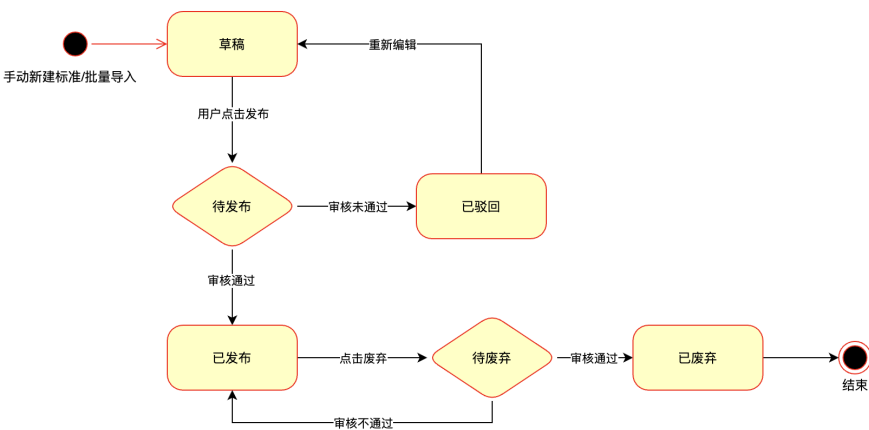
1. 单击导入代码，展示导入代码的功能弹窗。
2. 单击下载导入模版，可以下载公共代码导入的 Excel 模版，不同类型公共代码的导入模版也不同。
3. 按照模版的格式，填写导入的公共代码数据，将编辑好的 Excel 文件上传，单击点击选择文件，选择已经填写好的表格文件，即可进行导入，导入完成后，页面会展示导入结果。
4. 代码导出：单击代码导出，可以导出指定主题下的所有公共代码。导出的公共代码为 Excel 格式。

状态管理

生命周期

数据标准/公共代码的生命周期包括以下状态：草稿、待发布、已发布、已驳回、待废弃、已废弃。通过发布、修订和废弃这三种操作，可以修改它们的状态。

各个状态的流转方式如下：



表四 操作说明

操作项名称	说明
发布	新建的数据标准/公共代码处于草稿状态，需要发布之后才可以被正式使用。在右侧的操作栏点击「发布」按钮，即可进行发布操作。点击「发布」之后，该数据标准的状态变化为待发布。 另外在标准审核页面，会出现一条待审核的发布申请，具有“标准审核”权限的用户可以对该条申请进行审核，审核操作见“审核”小节。 如果审核通过，则该标准的状态变更为已发布，如果审核不通过，则该标准的状态变为已驳回，此时，可以根据审核意见，对标准进行修改之后再次提交发布申请。
修订	对于已发布的数据标准/公共代码，点击右侧操作栏中的「编辑」按钮，可以对其进行修订操作。此时会跳转到数据标准的编辑页面，对数据标准的属性进行修改之后，点击提交，将会创建一个新的子版本。此时，在数据标准的左侧会出现一个可以点击的下三角按钮。 点击该按钮，弹出该数据标准的子版本，可以对该子版本进行发布、编辑、删除操作，点击「发布」按钮，则会创建一个修订申请，由数据管理员对该申请进行审核（见“审核”小节）。如果审核通过，则该子版本将会替代原标准，成为新的数据标准；如果审核不通过，则该子版本的状态变更为已驳回，此时可对子版本进行编辑后再次发布，或者进行删除。
废弃	对于已发布的数据标准/公共代码，如果不再使用，则可以进行废弃。点击右侧操作栏中的「废弃」按钮，可以对标准进行废弃操作。 点击「废弃」按钮之后，此时数据标准的状态变更为待废弃。 另外在标准审核页面，会出现一条待审核的发布申请，具有“标准审核”权限的用户可以对该条申请进行审核，审核操作见“审核”小节。 如果审核通过，则该标准的状态变更为已废弃，如果审核不通过，则该标准的状态重新变为已发布。

数据落标检查

设置落标规则

1. 在**数据架构>数据标准>标准管理**页面，单击操作列的**更多**按钮，在展开的选项中单击设置落标规则项，进入设置界面。

或者点击任一标准中文名称的蓝字按钮，进入标准详情页，单击右上角“设置落标规则”按钮，进入设置界面。

2. 在“设置落标规则”操作界面，可查看选中标准对应的标准信息、自动落标规则以及手动落标规则。

若在标准中配置了以下属性，则会自动生成对应的落标规则。

表五 规则说明

	标准属性	质量规则	规则类型
技术属性	数据类型	数据类型校验	元数据校验
	允许为空	字段空值	完整性
业务属性	引用码表	枚举值校验	准确性
	值域范围	值域范围校验	准确性

可在页面最下方勾选是否要自动创建落标任务。

3. 单击“手动落标规则”板块中的“+新增落标规则”，可在弹出的新增弹窗页中进行新增操作。

无论是自动生成的“自动落标规则”还是手动新增的“手动落标规则”，初始状态都是“已停用”，单击打开开关按钮可修改状态至“已启用”。

4. 在“新增质量规则”的弹窗页中可选择配置模式，默认关闭高级模式，单击打开开关可进入高级模式的新增配置页。

在默认的配置模式下，选择“规则模板名称”和“告警级别”，选填“关联描述”。

某些规则模版需要填写相应信息以完成完整的质量规则新增配置。

5. 若打开按钮进入高级模式，则有更多规则模板可供选择。根据选择规则模板的不同，规则配置项也会有对应差别。

告警条件中的条件规则可参照样例填写。完成所有规则配置项的填写或选择后，选填“关联描述”，单击“确认”按钮，完成质量规则新增。

6. 完成所有规则新增后可在“设置落标规则”界面确认是否启用，单击操作页蓝字按钮“编辑”、“删除”完成对应操作。单击页面最下方“确认”按钮，完成落标规则设置。

引用数据标准

1. 在「元数据」中新建表，在字段定义中可以关联数据标准。

2. 单击「关联标准」的 icon 后，在弹出的页面中会展示所有状态为**已发布**和**待废弃**的数据标准。

3. 关联数据标准后，会对字段定义的类型、是否允许为空、密级等属性进行校验，如果与数据标准定义的不一致，则会进行提示，且此时元数据表无法进行提交。

4. 将字段定义的属性修改为与标准定义一致，直到没有校验提示信息，才可以进行提交。

5. 在数据标准的详情页面，可以看到当前引用该标准的元数据表及字段。

执行落标任务

1. 在“数据标准”的“设置落标规则”页面最下方勾选“自动创建落标任务”功能后，即可自动创建落标任务并进行落标检查。

2. 进入**数据治理>数据质量>质量作业**，也可在页面内找到作业类型为**落标检查任务**的相应作业。

3. 单击我的项目>作业组开发，进入离线计算页面，新建或选中任一任务组，在**数据治理**中拖拽出一个“Quality”节点。点击节点，在右侧打开的“节点配置”页面的“质量作业”项下拉选框，选择落标任务从而完成其执行操作，完成后点击“保持”和“发布”按钮。

查看落标检查报告

1. 单击**数据架构>数据标准>落标报告**，查看落标任务的统计数据总览。

2. 找到“评估标准”板块，单击操作列**查看结果**，可跳转至标准对应的详情界面，查看更具体的结果概览和字段详情信息。

数据治理

数据地图

数据地图基于元数据采集提供各类数据检索、数据血缘、数据质量、版本信息查看等功能，帮助用户更加方便快捷的找数、用数。

🔍 数据搜索

进入数据搜索需要通过首页侧边导航选择数据治理>数据地图>数据搜索。搜索框支持按照表名称、表描述、字段名称和字段描述进行搜索。侧边通过数据源类型和筛选条件（数据源、数据库和标签）刷新列表。单击表进入详情可以查看表的详细信息。

支持的数据源类型包括HIVE、DORIS、MYSQL、ORACLE、SQLSERVER、CLICKHOUSE、POSTGRESQL、GREENPLUM、HANA、TIDB、DM。

基本信息

用户可以在表详情页查看表基本信息和技术信息。基本信息包括表名称、数据库等。

表一 表详情基础信息说明

字段名称	字段名称
表名称	表英文名
表中文名	采集表描述，作为表中文名
数据库	表所属数据库
创建时间	元数据创建时间
更新时间	元数据采集任务更新时间
类型	元数据的表类型。包括：物理表、映射表、外部表
创建人	采集任务的创建人
湖格式	元数据表所属的湖格式
数据标签	普通用户可以使用标签管理中已创建的标签对数据表打标签。

字段信息

详情表字段信息展示表的字段定义和分区定义，单击去查询，跳转至数据分析，支持用户查询表数据。

数据血缘

详细表信息汇总数据血缘通过可视化图表查看数据血缘图谱，您可以切换表视图、字段视图，灵活选择上下游层级，其中表视图是查看表的上下游血缘图，字段视图可查看表中的字段的上下游血缘图，支持查看节点详情和产生指定血缘关系链路的任务详情。

数据质量

数据汇聚到平台后，需要对数据的完整性、唯一性、有效性、准确性、一致性、及时性进行探查，清洗脏数据，以确保数据的存储和共享的质量。

- 表质量得分：统计表质量得分，展示质量得分，健康状态展示合格/不合格。
- 数据质量：可查看表上次运行的实例中的质量规则配置，以及告警情况，查看任务上次实例的运行时间。
- 历史得分趋势：支持查询历史情况；统计表历史得分趋势。

版本管理

版本信息查看历史的元数据版本，支持按照更新时间筛选列表。版本管理查看版本名称、操作人、最后更新时间和对比上一版本操作。每次采集元数据有变化，就生成一个版本，没有变化就不会生成版本。

分区信息

进入表详情分区信息可以查看、搜索表的所有分区。支持选择分区字段，对单个分区字段进行精确查询。

表二 分区信息字段说明

字段名称	字段描述
分区字段名称	当前表创建的分区字段
分区创建时间	数据被存储到相应分区中的时间
分区路径	每个分区的路径
文件数	每个分区下的文件总数
分区行数	每个分区下的数据总行数
分区大小	每个分区文件的存储大小

元数据采集

元数据采集统一采集和汇总不同系统中的元数据，支持用户在数据地图中查看汇集而来的元数据信息。

创建采集任务

1. 登陆百度数据湖管理与分析平台EasyDAP进入首页，单击侧边导航数据治理>数据地图>元数据采集，进入元数据采集界面。
2. 单击创建任务按钮，在创建页面填写相关配置项后单击保存并执行按钮后采集任务创建完毕。

- 采集数据源类型：

HIVE、DORIS、MYSQL、ORACLE、SQLSERVER、CLICKHOUSE、POSTGRESQL、GREENPLUM、HANA、TIDB、DM

- 采集对象：

表三 采集对象字段说明

字段名称	字段说明
采集任务名称	输入采集任务名称。
数据源名称	选择数据源管理中配置的对应类型的数据源链接。
测试连通性	测试数据源连通性。
库范围	选择“ALL”会按照数据源粒度采集。已采集的库不能重复采集。
表范围	全部表（包含后续新增）、自定义。采集粒度支持到表，支持用户选择全部表（包含后续新增）、自定义
标签管理	批量添加标签。用户可以使用标签管理中已创建的标签对数据表批量打标。

- 运行配置：

表四 运行配置字段说明

字段名称	字段说明
执行资源	用户选择任务执行所需的资源。
执行频率	调度执行、手动执行。
目标表被删除	删除对应的元数据、忽略删除。
字段更新策略	更新元数据、忽略更新。

采集任务列表

采集任务列表入口点位在首页>数据治理>数据地图>元数据采集。创建的采集任务会以列表的形式展示，列表可以根据运行时间范围进行筛选。

表五 列表字段说明

字段名称	字段描述
任务名称	采集任务名称
数据源类型	任务选择的数据源类型
数据源	任务选择的数据源链接
采集库	任务选择的采集库
任务类型	一次性任务、周期任务
上次执行状态	上次任务执行状态（草稿、待触发、运行中、成功、部分成功、失败、已暂停、暂停中）
上次执行时间	上次任务执行时间
描述	任务创建时填写的任务描述
创建人	任务创建人
创建时间	任务创建时间
操作	- 启动：运行中不能重复启动，一次性任务启动实例运行，调度任务启动调度，新增一条待触发实例，实例按照调度时间运行。 - 暂停：运行中可以暂停，停止正在运行的实例和调度。 - 编辑：运行中状态不能编辑。 - 删除：运行中状态不能删除。 - 调度配置：调整任务的调度配置。 - 查看实例：查看实例详情。

任务执行状态说明：

- 一次性任务

上次执行状态	说明	启动	停止	编辑	删除	调度配置	查看实例
未运行（草稿）	任务创建未运行，可修改所有配置项	✓	✗	✓	✓	✓	✗
待触发	任务已提交，等待资源执行	✗	✓	✗	✗	✗	✓
运行中	任务正在运行中	✗	✓	✗	✗	✗	✓
成功	任务所有表采集成功	✓	✓	✗	✗	✗	✓
部分成功	任务部分表采集成功	✓	✓	✗	✗	✗	✓
失败	任务运行失败，所有表采集失败	✓	✓	✗	✗	✗	✓
停止中	任务停止中	✗	✗	✗	✗	✗	✓
已停止	任务已停止	✓	✗	✓	✓	✓	✓

- 调度任务

上次执行状态	说明	启动	暂停	编辑	删除	调度配置	查看实例	手动执行
未运行（草稿）	任务创建未运行，可修改所有配置项	✓	✗	✓	✓	✓	✗	✗
待触发	任务已提交，等待资源执行	✗	✓	✗	✗	✗	✓	✗
运行中	任务正在运行中	✗	✓	✗	✗	✗	✓	✗
成功	任务所有表采集成功	✗	✓	✗	✗	✗	✓	✗
部分成功	任务部分表采集成功	✗	✓	✗	✗	✗	✓	✓
失败	任务运行失败，所有表采集失败	✓	✓	✗	✗	✗	✓	✓
已停止	任务已停止。 已停止正在运行的任务、停止周期调度	✓	✗	✓	✓	✓	✓	✗
停止中	任务停止中	✗	✗	✗	✗	✗	✓	✗

采集任务操作

编辑采集任务

在采集任务列表可以对采集任务进行编辑操作编辑时，不能修改采集类型（数据源类型）、采集对象（数据源名称、库范围），只能编辑任务名称、表范围、表描述、调度周期、执行资源。

- 库范围为：ALL或多个库时，表范围只能选择「全部表」，不可编辑；
- 库范围为：单个库时，表范围可以选择「全部表」、「自定义」，可编辑。

删除采集任务

- 元数据采集任务删除，对应实例也全部删除，并删除任务采集下所有已采集的元数据。
- 若有正在运行的实例，不能删除任务，提示已有正在运行的实例，请停止后删除。

查看实例详情

单击查看实例按钮进入实例详情界面，展示当前任务所有实例页面，实例保存一个月。

注意：不能查看采集结果的状态：草稿、待触发、运行中、已暂停、暂停中

表六 实例详情列表字段说明

字段名称	字段描述
实例ID	任务执行的实例ID
执行状态	采集任务的执行状态：草稿、待触发、运行中、成功、部分成功、失败、已暂停、暂停中
实例类型	周期任务、一次性任务、手动执行
开始时间	采集任务开始时间
结束时间	采集任务结束时间
执行耗时	任务执行耗时
采集结果	单击采集结果查看已采集数/总数。 - 表类型：变更表、未变更表和失败表（可查看失败原因） - 变更类型：新增表、删除表和更新表

1. 登陆百度数据湖管理与分析平台EasyDAP进入首页，单击侧边导航数据治理>数据地图>标签管理，进入标签管理界面。
2. 单击创建标签按钮，进入创建页面，填写标签名称和描述后单击确认按钮标签创建成功。
3. 标签创建完毕后会以列表形式展示，支持按照标签名称进行搜索。列表展示创建时填写的标签名称和描述，还展示创建人和创建时间。
4. 对标签可以进行编辑和删除操作，标签支持进行批量删除。

数据血缘

🔗 数据血缘概述

功能介绍

EDAP数据血缘，监控元数据和相关任务变化状态，进行血缘解析，提供数据在生产流转过程中的血缘关系。旨在帮助您分析元数据的上下游关联关系，定位影响层级和任务依赖关系，可快速发现和回溯对其他元数据对象的影响。

本功能支持通过可视化图表查看数据血缘图谱，支持的数据源类型包括EDAP、Postgresql、Oracle、Hana、Doris、SQLServer、Greenplum、TiDB、MySQL、Hive、ClickHouse。您可以切换表视图、字段视图，灵活选择上下游层级，其中表视图是查看表的上下游血缘图，字段视图可查看表中的字段的上下游血缘图，支持查看节点详情和产生指定血缘关系链路的任务详情。EDAP平台提供了多种作业类型的血缘关系自动解析服务，针对部分无法解析的使用场景，可以通过手动填报入口进行血缘关系的填报。

名词解释

数据血缘

数据血缘是元数据管理、数据治理、数据质量的重要一环，追踪数据的来源、处理、出处，对数据价值评估提供依据，描述源数据流程、表之间的流向关系，表与表的依赖关系、表与平台作业，计算引擎之间的依赖关系。

活跃血缘

血缘图谱当前仅展示活跃状态的血缘。活跃血缘指的是已上线作业产出的血缘，且作业执行时间间隔在35d内。已上线作业包括作业组调度、单节点调度、离线同步作业调度。

静默血缘

血缘图谱当前不展示活跃状态的血缘。静默血缘指的是未来可能不会更新或即将失效的血缘，包括开发模式运行、线上调度已运行过但是已取消调度超过35d、线上模式逾期35d执行。

应用场景

场景一：帮助用户快速找到想要的数​​据，并且了解数据表的详情和上下游血缘关系。

场景二：帮助开发人员高效地定位问题，在问题排查时能评估故障影响范围。

- 归因分析（上游）：采用向上追溯的方式查找数据来源于哪里，经过了哪些加工和处理。常用于在发现数据问题时，追溯上游表、字段，快速定位和找到数据问题的原因。
- 影响分析（下游）：影响分析是向下游追踪，用来查询和定位数据去了哪里。常用于当元数据发生变更时，分析和评估变更对下游业务的影响。

🔗 血缘关系查询

EDAP提供两个数据血缘关系查询入口，分别为数据血缘-血缘图谱和元数据-数据血缘，用户可根据使用场景进行选择。

血缘图谱查询

操作步骤：

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。
2. 在左侧导航栏，单击数据治理-数据血缘。
3. 单击上方血缘图谱览即可进入血缘查询页面。

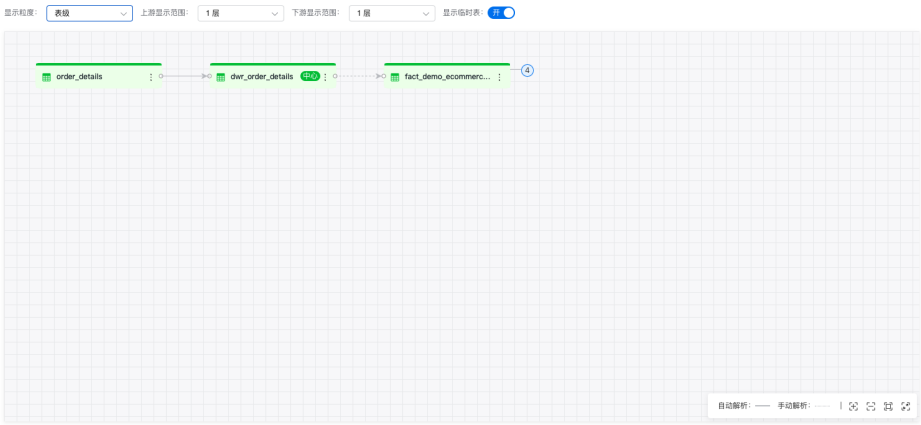
元数据检索

操作步骤：

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。
2. 在左侧导航栏，单击数据治理-数据血缘，进入后，默认在血缘图谱页面。
3. 选择你需要查看的表，从左侧依次选择数据源类型、数据库、表，查看该表的数据血缘。

表级血缘

在选择表后，以该表为中心，显示上下游血缘关系。



操作步骤：

1. 选择显示粒度

血缘图谱支持表级、字段级两种显示粒度，默认显示表级。

2. 选择显示范围

默认直系父子，即上游1层、下游1层，如果没有上游或下游，则显示0层。

支持用户自定义上下游层级，可下拉选择：0层、1层、2层、3层。

在图谱中点击节点左侧的“+”、“-”可以展开或折叠上一层级，末端节点无展开/折叠按钮。

3. 查看节点信息

点击节点右侧“更多”，选择“查看节点信息”，右侧显示节点信息：表名称、数据源类型、数据源链接、数据库、更新时间、schema信息。对于数据湖的表，点击“查看表情”，可跳转到元数据详情页面。

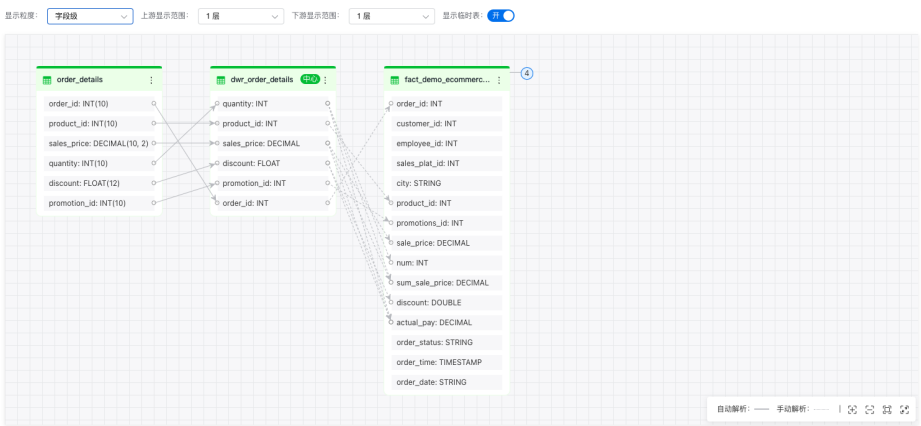
4. 查看任务信息

点击连线上的任务节点，右侧显示产生该血缘关系的任务信息，包括：来源任务名称、任务类型、创建人、创建时间、运行时间、任务内容等。点击“查看任务详情”可以跳转到具体作业。

5. 切换中心节点

双击其他节点，或点击节点右侧“更多”，选择“设为中心表”，可以进行中心节点的切换。

字段血缘



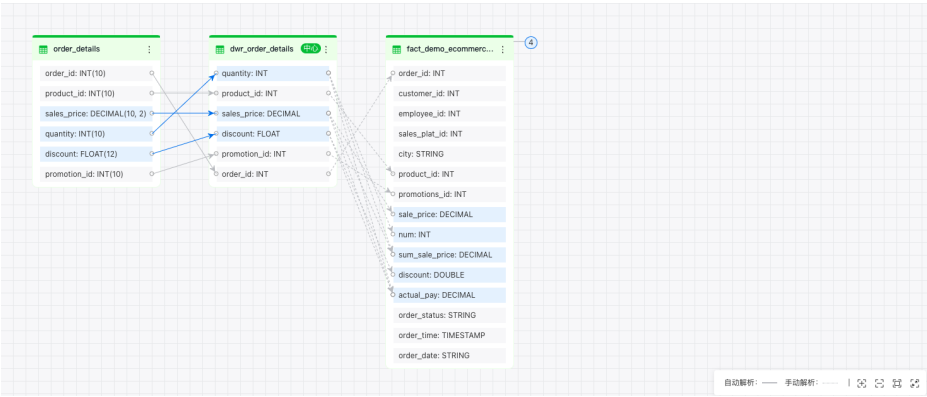
操作步骤：

1. 切换显示粒度

下拉“显示粒度”选择字段级，血缘图谱会展开数据表的所有字段，并显示字段间的血缘关系连线。

2. 指定字段查询

选中某一个字段，图谱上高亮显示该字段的上下游字段。



血缘关系解析

自动解析

EDAP平台支持数据集成作业、大部分脚本开发作业、特定场景可视化ETL作业的血缘关系自动解析，支持范围详情如下。

数据集成

源端	目标端	血缘获取方式
mysql、sqlserver、oracle、hana、postgresql	hive、doris、edapdatalake	系统解析

数据集成完全支持字段级血缘系统解析。

脚本作业

脚本作业类型	数据源类型	血缘获取方式
HIVE SQL	hive、edapdatalake	系统解析
JDBC SQL	mysql、sqlserver、clickhouse、oracle、doris、hana	系统解析
Spark SQL	hive、edapdatalake	系统解析
PySpark		手动配置
Scala		手动配置
Shell		手动配置
Python		手动配置

SQL类脚本作业支持字段级血缘系统解析。

- Hive SQL
 - 支持Create table as等DDL操作产生的【Hive表/数据湖表】之间的血缘。
 - 支持Insert into/overwrite等DML操作产生的【Hive表/数据湖表】表之间的血缘。
- JDBC SQL
 - 支持Create table as等DDL操作产生的【数据源表】之间的血缘。
 - 支持Insert into等DML操作产生的【数据源表】之间的血缘。
- Spark SQL
 - 支持Create table as等DDL操作产生的【Hive表/数据湖表】之间的血缘。
 - 支持Insert into/overwrite等DML操作产生的【Hive表/数据湖表】表之间的血缘。

对于程序类脚本作业，支持“手动配置”的方式，补齐血缘。

可视化ETL

插件类型	数据源类型	血缘获取方式
Hive、Hive3	hive	自动解析
DataLakeSource DataLakeSink	edapdatalake	自动解析
Doris	doris	自动解析
Database	mysql、tidb、oracle、sqlserver、hana、postgresql、greenplum	自动解析
其它		手动配置

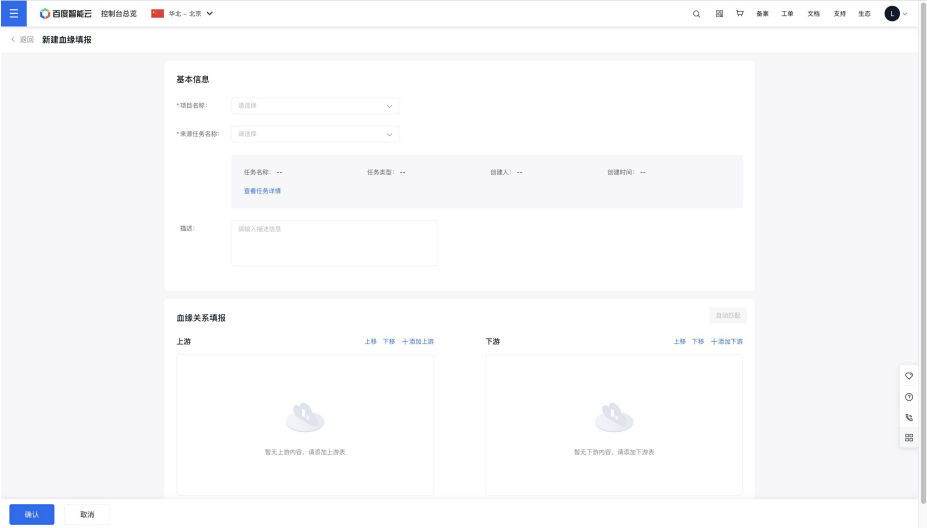
支持的场景：数据同步，即源端插件——>目标端插件，数据处理插件、开发分析插件不支持

插件范围包括：Hive、Hive3、DataLakeSource、DataLakeSink、Doris。对于暂未支持的插件，支持“手动配置”的方式，补齐血缘。

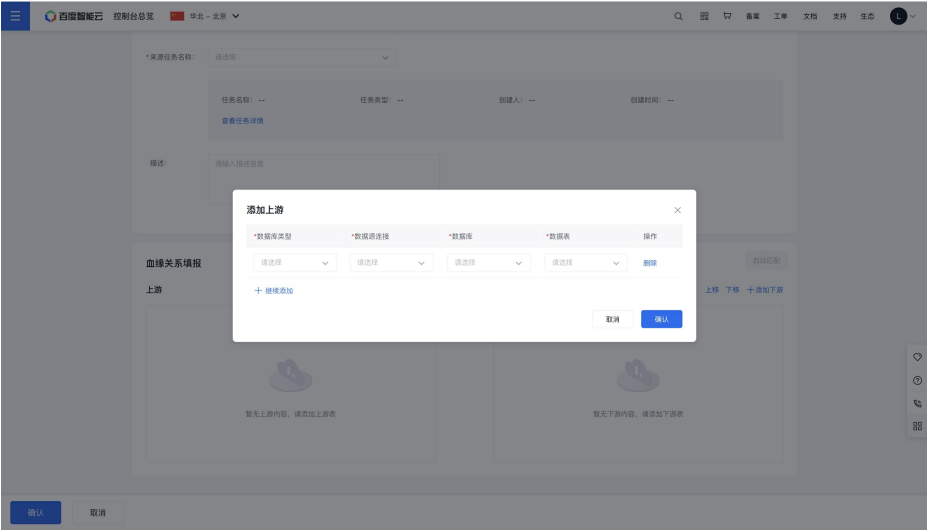
手动填报

血缘手工填报

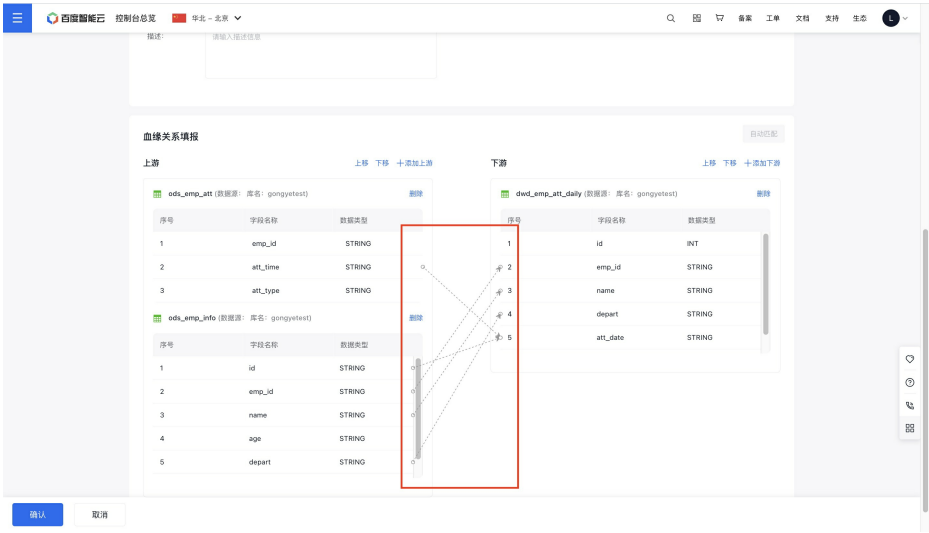
- 1、点击“数据治理——数据血缘”进入该界面，在当前界面内切换进入“血缘填报”tab页，点击“+新建血缘填报”按钮，进入新建页。
- 2、在“新建血缘填报”界面选择“项目名称”和“来源任务名称”，选填“描述”。选中来源任务后，可在选框下的灰色注释框内点击“查看任务详情”按钮，跳转至任务的脚本开发界面。
- 3、在“血缘关系填报”部分可为选中的任务新增或编辑上下游关系。所选任务内已有的填报记录会同步展现在此处，若对其重新编辑并点击确认，则新的填报内容将会覆盖。



- 4、点击“血缘关系填报”部分的蓝字按钮可对其进行相应操作。点击“+添加上游”，打开添加弹窗页，选择“数据库类型”、“数据源连接”、“数据库”和“数据表”，选择完成后点击确认按钮。添加下游操作同理。



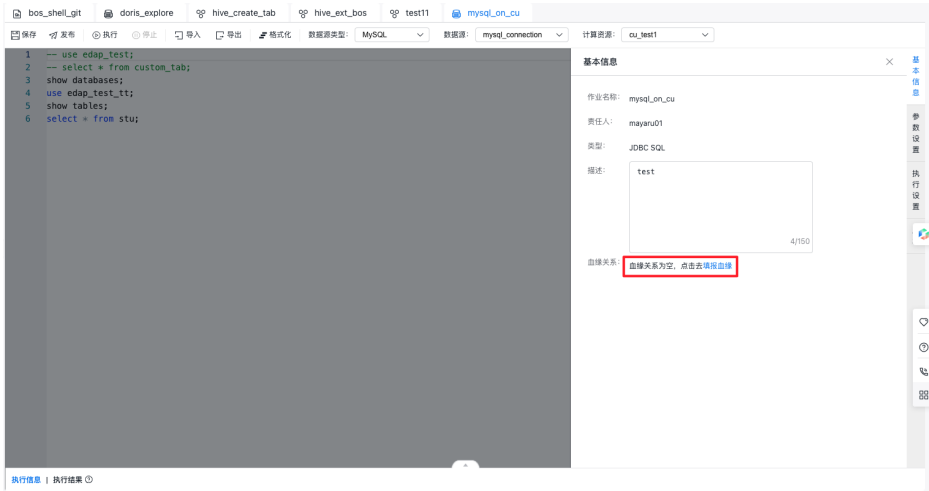
- 5、在上游表和下游表之间，可以分别在其末端和首端手动拖拽箭头进行血缘关系的连线；也可点击“自动匹配”按钮，快速连接上下游同名字段。若想清除某条线的连接，点击箭头并拖拽离开下游表单首端即可。



6、完成后点击页面下方“确认”按钮，完成手工填报操作。

查看项目内血缘关系 1、在左侧导航栏，单击“数据加工-我的项目”，进入后再点击“离线开发-可视化作业开发”或“离线开发-脚本作业开发”。

2、选中一个任务后，单击右侧“基本信息”，可查看该作业是否产生元数据血缘。



任务基本信息下方显示：

- 1. 该任务支持/暂不支持血缘关系自动解析。
- 2. 存在血缘关系手动填报记录，点击“查看”蓝字按钮，点击后在新页面打开“血缘填报详情”页面。
- 3. 暂无血缘关系手动填报记录，点击“新建”蓝字按钮，点击后在新页面打开“血缘填报新建”页面，自动填充项目、任务名称。

数据质量

概述

数量质量概述

数据质量支持对常见的离线数据源（包括EDAP数据湖、Hive、Doris、Mysql、Oracle、SqlServer等）进行质量校验，从数据的完整性、唯一性、有效性、准确性、一致性、及时性六大维度进行质量监控，提供从质量规则制定、质量作业创建、重大问题告警、问题分析整改的全流程能力。内置45个质量规则，支持通过规则模板、自定义SQL模板快速创建质量任务，支持将质量任务与作业组进行关联调度，及时产生告警和阻塞链路，避免问题数据污染扩散。从多维度质量分析报表和质量绩效评分，提供作业实例查询、异常数据查看功能，帮助您迅速定位质量问题，完成质量管理闭环。

功能介绍

数据质量各个功能模块介绍如下：

名称	描述
质量总览	数据质量总览页面为您展示数据质量全局配置和告警统计情况。包括： - 支持的所有数据源的表级规则设置率。 - 按规则维度分类的质量规则异常占比。 - 按告警等级分类的历史告警趋势图。
规则引擎	数据质量支持自定义规则模板库，并对自定义规则进行统一管理。 也可以查看系统内置规则模板详情。
质量作业	数据质量支持按库表配置质量监控规则。
运维管理	在运维管理页面您可以通过库表查看历史校验记录和校验结果详情。
质量报告	数据质量报告页面为您展示库表的规则告警和评分情况。包括： - 所选日期的主题/源连接、库得分。 - 所选日期范围的主题/源连接、库历史得分趋势图。 - 所选日期的主题/源连接下库得分或库下表得分详情。
配置管理	数据质量配置管理页面您可以对异常数据存储路径进行配置。

名词解释

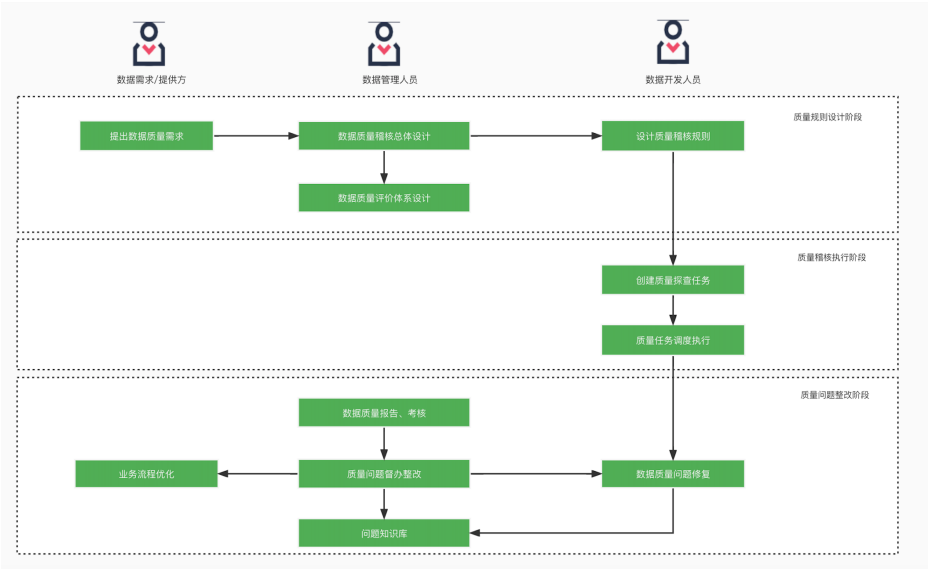
规则模板 数据质量的数据校验方法基于规则模板，规则模板定义了从特定维度和粒度来校验数据是否符合预期。

维度 为从不同角度评估质量问题，将规则模板分为六大维度，即完整性、唯一性、有效性、准确性、一致性、及时性。

粒度 从数据表角度看，将规则模板划分为三大粒度，即表级、字段级、跨字段级。
质量作业 数据质量的具体执行依赖质量作业配置，质量作业对数据表按规则粒度定义规则模板和触发告警条件。
告警 对于给定数据和对应规则模板，如果数据结果不符合规则模板预期，则触发告警，分为L1提示、L2警告、L3严重三级。

角色、职责与权限

数据质量管理流程：



角色	职责描述	功能权限
数据管理员	- 负责根据业务方需求、数据标准要求，进行数据质量稽核总体设计。 - 负责制定数据质量评价体系，定期开展数据质量考核，对未达标数据进行督办整改。	- 查看质量总览； - 创建并管理质量规则； - 拥有平台全部数据权限，可对所有数据表创建质量任务，并查看作业运行实例； - 查看并下载质量报告等。
数据开发人员	根据数据质量需求，设计质量校验规则，配置数据稽核任务，并对相关数据质量问题进行处理。	- 查看数据总览； - 创建并管理质量规则； - 根据数据管理员授权，对有权限的数据表创建质量任务，并查看作业运行实例； - 查看并下载质量报告等。

🔗 总览

数据质量总览页面为您展示数据质量全局配置和告警统计情况。

进入质量总览

- 登录并进入百度智能云数据湖管理与分析EasyDAP，单击侧边导航**数据治理>数据质量（顶部）>总览（侧边）**。
- 总览包括规则设置率、质量异常占比和告警趋势。

表一 总览说明

名称	说明
规则设置率	数据质量支持的离线数据源（包括EDAP数据湖、Hive、Doris、Mysql、Oracle、SqlServer）下所有表的规则设置率统计。一个表如果设置了一个或多个质量规则，该表计为已设置，反之计为未设置。
质量异常占比	按照规则维度分类，统计各维度下质量任务规则结果触发告警（异常）的比例。
告警趋势	各告警类型（L1 提示、L2 警告、L3 严重）数量趋势，最多保存记录30天数据。支持按照时间维度（近1天/近7天/近30天）选择查看，同时可选择时间范围进行筛选。

🔗 规则管理

数据质量支持自定义规则模板库，并对自定义规则进行统一管理，也可以查看系统内置规则模板详情。规则维度包括：完整性、唯一性、有效性、准确性、一致性、及时性。规则粒度包括：表级、字段级、跨字段级。

表一 规则维度说明

规则维度	描述
完整性	数据在创建、传递过程中无缺失和遗漏，包括实体完整、属性完整、记录完整和字段完整四个方面。
一致性	遵循统一的数据标准记录和传递数据信息，主要体现在数据记录是否规范、数据是否符合逻辑。
准确性	真实、准确地记录原始数据，无续建数据及信息。数据准确反映其所建模的“真实世界”实体。
唯一性	同一数据只能有唯一的标识符。体现在一个数据集中，一个实体只出现一次，并且每个唯一实体有一个键值 且该键值只指向该实体。
有效性	数据的值、格式和展现形式符合数据定义和业务定义的要求。
及时性	及时记录和传递相关数据，满足业务对信息获取的时间要求。数据交付及时，抽取要及时，展现要及时，数据交付时间过长可能导致分析结论失去参考意义。

创建规则模板

- 登录并进入百度智能云数据湖管理与分析EasyDAP。在左侧导航栏选择**数据治理>数据质量>规则模板**。
- 单击**新建规则模板**，填写规则模板的基本配置和关系定义。完成后确认即完成创建。

表二 创建规则模板配置项说明

配置项名称	描述	使用限制
规则模板名称	自定义规则模板名称。	只能包含英文字母、数字、中划线和下划线，且以英文字母开头，不大于50个字符。
规则粒度	下拉列表选择表级、字段级、跨字段级。	-
规则分类	下拉列表选择完整性、唯一性、有效性、准确性、一致性、及时性。	-
描述	自定义规则描述信息。	长度不大于100个字符。
规则表达式	自定义SQL校验逻辑，支持添加库表、字段和输入参数。最多添加2张表、5个字段参数、2个输入参数；如果有多行输出，只取第一行数据，且数据必须为数值类型。	不支持"select * "语法。库表、字段和输入参数请使用添加参数按钮进行添加，修改参数序号以保证逻辑正确性。
规则输出	输出可为最大值、最小值、合法行数、合法比率等。	如有多个输出值，输出值个数需和规则表达式输出个数一致，用英文逗号","进行分隔。

内置规则列表

1. 登录并进入百度智能云数据湖管理与分析EasyDAP，在左侧导航栏，选择数据治理>数据质量。
2. 单击左侧导航规则管理进入规则模板管理页面，即可查看规则列表。可对规则进行编辑和删除。
- 编辑：单击对应规则操作栏的编辑按钮，进入规则编辑页面。如果规则已经被质量作业引用，则无法编辑。

● 删除：单击对应规则操作栏的删除按钮，即可删除。如果规则已经被质量作业引用，则无法删除。

表一 内置规则列表表级规则

规则维度	规则名称	描述	输出值	输入参数
完整性	字段空值	指定字段为空或空字符串，则为1个空值行。	空值行数、总行数、空值率	
唯一性	字段唯一值	指定字段是唯一的，则为1个唯一值行。包含NULL值（NULL值计入唯一值行数）。	唯一值行数、总行数、唯一值率	
	字段重复值	指定字段有重复值，则为n个重复值行，n为重复次数。不包含NULL值（NULL值不计入重复值行数）。	重复值行数、总行数、重复值率	
有效性	正则表达式校验	判断指定字段是否满足一定的格式，通过自定义正则表达式进行匹配。不校验NULL值（NULL值计入有效行数），下面正则类型规则相同。	无效行数、有效行数、总行数、无效比率	正则表达式
	邮箱校验	匹配邮箱格式，正则表达式为：^[a-zA-Z0-9_-]+@[a-zA-Z0-9_-]+(\.[a-zA-Z0-9_-]+)+\$	无效行数、有效行数、总行数、无效比率	
	身份证校验	匹配身份证号码格式，正则表达式为：(^([1-9]\d{5}(18 19 20)\d{2}((0[1-9]) (10 11 12))((0-2) ([1-9]) 10 20 30 31)\d{3}[0-9Xx]\$)((^([1-9]\d{5}\d{2}((0[1-9]) (10 11 12))((0-2) ([1-9]) 10 20 30 31)\d{3}\$	无效行数、有效行数、总行数、无效比率	
	IP地址校验	匹配IP地址格式，正则表达式为：^(?:(?:1[0-9] [0-9-]) (?:(?:2[0-4] 0-9) (?:(?:25[0-5] (?:(?:1-9) [0-9-])) 3(?:?:1[0-9] 0-9)) (?:(?:2[0-4] 0-9) (?:(?:25[0-5] (?:(?:1-9) [0-9-])) 3(?:?:1[0-9] 0-9)))\$	无效行数、有效行数、总行数、无效比率	
	电话校验	匹配手机号码号码格式，正则表达式为：^([1][3,4,5,6,7,8,9]\d{9})\$	无效行数、有效行数、总行数、无效比率	
	邮编校验	匹配邮编格式，正则表达式为：^[1-9]\d{5}\$	无效行数、有效行数、总行数、无效比率	
	日期校验	匹配日期格式（yyyy-MM-dd），正则表达式为：^[1-	无效行数、有效行数、总行	

		9]\d{3}-(0[1-9] [10-2])-(0[1-9] [1-2][0-9] 3[0-1])\$	数、无效比率	
准确性	字段平均值	指定字段的平均值是否准确。	平均值	
	字段平均值1天波动率	与1天前的执行结果做比较。	1天波动率	
	字段平均值7天波动率	与7天前的执行结果做比较。	7天波动率	
	字段平均值30天波动率	与30天前的执行结果做比较。	30天波动率	
	字段平均值7天平均值波动率	与前7天的平均值做比较。	7天平均值波动率	
	字段平均值30天平均值波动率	与前30天的平均值做比较。	30天平均值波动率	
	字段汇总值	指定字段的汇总值是否准确。	汇总值	
	字段汇总值1天波动率	与1天前的执行结果做比较。	1天波动率	
	字段汇总值7天波动率	与7天前的执行结果做比较。	7天波动率	
	字段汇总值30天波动率	与30天前的执行结果做比较。	30天波动率	
	字段汇总值7天平均值波动率	与前7天的平均值做比较。	7天平均值波动率	
	字段汇总值30天平均值波动率	与前30天的平均值做比较。	30天平均值波动率	
	字段最大值	指定字段的最大值是否准确。	最大值	
	字段最大值1天波动率	与1天前的执行结果做比较。	1天波动率	
	字段最大值7天波动率	与7天前的执行结果做比较。	7天波动率	
	字段最大值30天波动率	与30天前的执行结果做比较。	30天波动率	
	字段最大值7天平均值波动率	与前7天的平均值做比较。	7天平均值波动率	
	字段最大值30天平均值波动率	与前30天的平均值做比较。	30天平均值波动率	
	字段最小值	指定字段的最小值是否准确。	最小值	
	字段最小值1天波动率	与1天前的执行结果做比较。	1天波动率	
	字段最小值7天波动率	与7天前的执行结果做比较。	7天波动率	
	字段最小值30天波动率	与30天前的执行结果做比较。	30天波动率	
	字段最小值7天平均值波动率	与前7天的平均值做比较。	7天平均值波动率	
	字段最小值30天平均值波动率	与前30天的平均值做比较。	30天平均值波动率	

	离群值	检测指定字段的数据与其他数据差异较大的，3倍标准差。	离群值行数、总行数、离群值比率	
	值域范围	指定字段的取值范围是否在规定范围内。	无效行数、有效行数、总行数、无效比率	值域范围
	枚举值校验	指定字段的取值范围是否在枚举值中。	无效行数、有效行数、总行数、无效比率	枚举值

表二 字段级规则

规则维度	规则名称	描述	输出值	输入参数
完整性	字段空值	指定字段为空或空字符串，则为1个空值行。	空值行数、总行数、空值率	
唯一性	字段唯一值	指定字段是唯一的，则为1个唯一值行。包含NULL值（NULL值计入唯一值行数）。	唯一值行数、总行数、唯一值率	
	字段重复值	指定字段有重复值，则为n个重复值行，n为重复次数。不包含NULL值（NULL值不计入重复值行数）。	重复值行数、总行数、重复值率	
有效性	正则表达式校验	判断指定字段是否满足一定的格式，通过自定义正则表达式进行匹配。不校验NULL值（NULL值计入有效行数），下面正则类型规则相同。	无效行数、有效行数、总行数、无效比率	正则表达式
	邮箱校验	匹配邮箱格式，正则表达式为：^[a-zA-Z0-9_-]+@[a-zA-Z0-9_-]+(\.[a-zA-Z0-9_-]+)+\$	无效行数、有效行数、总行数、无效比率	
	身份证校验	匹配身份证号码格式，正则表达式为：(^([1-9]\d{5}(18 19 20)\d{2}((0[1-9]) (10 11 12))(((0-2)[1-9]) 10 20 30 31)\d{3}[0-9Xx]\$)((^[1-9]\d{5}\d{2}((0[1-9]) (10 11 12))(((0-2)[1-9]) 10 20 30 31)\d{3}\$)	无效行数、有效行数、总行数、无效比率	
	IP地址校验	匹配IP地址格式，正则表达式为：^(?:(?:1[0-9] (?:2[0-4] 0-9).)(?:25[0-5].) (?:[1-9] 0-9).){3}(?:1[0-9] (?:2[0-4] 0-9).)(?:25[0-5].)(?:[1-9] 0-9))\$	无效行数、有效行数、总行数、无效比率	
	电话校验	匹配手机电话号码格式，正则表达式为：^([1][3,4,5,6,7,8,9]\d{9})\$	无效行数、有效行数、总行数、无效比率	
	邮编校验	匹配邮编格式，正则表达式为：^[1-9]\d{5}\$	无效行数、有效行数、总行数、无效比率	
	日期校验	匹配日期格式（yyyy-MM-dd），正则表达式为：^[1-9]\d{3}-(0[1-9] 1[0-2])-(0[1-9] 1-2[0-9] 3[0-1])\$	无效行数、有效行数、总行数、无效比率	
准确性	字段平均值	指定字段的平均值是否准确。	平均值	
	字段平均值1天波动率	与1天前的执行结果做比较。	1天波动率	
	字段平均值7天波动率	与7天前的执行结果做比较。	7天波动率	
	字段平均值30天波动率	与30天前的执行结果做比较。	30天波动率	
	字段平均值7天平均值波动率	与前7天的平均值做比较。	7天平均值波动率	

	字段平均值30天平均值波动率	与前30天的平均值做比较。	30天平均值波动率	
	字段汇总值	指定字段的汇总值是否准确。	汇总值	
	字段汇总值1天波动率	与1天前的执行结果做比较。	1天波动率	
	字段汇总值7天波动率	与7天前的执行结果做比较。	7天波动率	
	字段汇总值30天波动率	与30天前的执行结果做比较。	30天波动率	
	字段汇总值7天平均值波动率	与前7天的平均值做比较。	7天平均值波动率	
	字段汇总值30天平均值波动率	与前30天的平均值做比较。	30天平均值波动率	
	字段最大值	指定字段的最大值是否准确。	最大值	
	字段最大值1天波动率	与1天前的执行结果做比较。	1天波动率	
	字段最大值7天波动率	与7天前的执行结果做比较。	7天波动率	
	字段最大值30天波动率	与30天前的执行结果做比较。	30天波动率	
	字段最大值7天平均值波动率	与前7天的平均值做比较。	7天平均值波动率	
	字段最大值30天平均值波动率	与前30天的平均值做比较。	30天平均值波动率	
	字段最小值	指定字段的最小值是否准确。	最小值	
	字段最小值1天波动率	与1天前的执行结果做比较。	1天波动率	
	字段最小值7天波动率	与7天前的执行结果做比较。	7天波动率	
	字段最小值30天波动率	与30天前的执行结果做比较。	30天波动率	
	字段最小值7天平均值波动率	与前7天的平均值做比较。	7天平均值波动率	
	字段最小值30天平均值波动率	与前30天的平均值做比较。	30天平均值波动率	
	离群值	检测指定字段的数据与其他数据差异较大的，3倍标准差。	离群值行数、总行数、离群值比率	
	值域范围	指定字段的取值范围是否在规定范围内。	无效行数、有效行数、总行数、无效比率	值域范围
	枚举值校验	指定字段的取值范围是否在枚举值中。	无效行数、有效行数、总行数、无效比率	枚举值

表三 跨字段级规则

规则维度	规则名称	描述	输出值	输入参数
一致性	字段一致性校验	指定字段的取值在参考字段的取值中存在（以参考字段为基准）。	一致行数、不一致行数、总行数、一致比率。	

举例说明

示例1：查询满足自定义过滤条件的表行数 规则模板名称：test_sql_1

- 规则粒度：表级
- 规则分类：准确性
- 规则表达式：select count(*) from \${Table1} where \${Table1}.\${Column1} > \${Input1} and \${Table1}.\${Column2} is not null
- 规则输出：表行数

说明：该自定义规则查询表行数，且需满足条件字段1 \${Table1}.\${Column1} 大于自定义输入值 \${Input1}、字段2 \${Table1}.\${Column2} 不为NULL，自定义输入值 \${Input1} 需在配置质量作业字段规则时填入。规则输出为“表行数”，对应规则表达式的“count(*)”查询结果。

示例2：查询表的相关自定义统计信息 规则模板名称：test_sql_2

- 规则粒度：字段级
- 规则分类：准确性
- 规则表达式：select count(*),max(\${Table1}.\${Column1}),sum(\${Table1}.\${Column2}),avg(\${Table1}.\${Column3}) from \${Table1}
- 规则输出：表行数,字段1最大值,字段2汇总值,字段3平均值

说明：该自定义规则查询表行数、字段1 \${Table1}.\${Column1} 最大值、字段2 \${Table1}.\${Column2} 汇总值、字段3 \${Table1}.\${Column3} 平均值。规则输出一一对应，用英文分号“,”分割。

示例3：查询两表按指定字段join关联且满足过滤条件的行数 规则模板名称：test_sql_3

- 规则粒度：跨字段级
- 规则分类：准确性
- 规则表达式：SELECT count(1) FROM \${Table1} LEFT JOIN \${Table2} ON \${Table1}.\${Column1} = \${Table2}.\${Column1} WHERE \${Table1}.\${Column2} > 0 AND \${Table2}.\${Column2} > 0 AND \${Table2}.\${Column3} > 0
- 规则输出：表行数

说明：该自定义规则查询表1 \${Table1} 和表2 \${Table2} 按照指定字段left join并且满足对应过滤条件的表行数。

质量报告

查看技术报告

- 登录并进入百度智能云数据湖管理与分析EasyDAP。在左侧导航栏，单击数据治理>数据质量>质量报告。
- 在左侧目录中选技术报告，从左侧各数据源类型中选择一个，可查看对应主题/源连接或库的评分。
- 单击库/表评分部分中，告警状况列下的蓝字按钮，可跳转至运维管理列表页。
- 导出报告：单击库评分或表评分下的导出报告，可以导出当前评分详情为word格式。
- 报告更新周期：质量报告按周期定时更新，默认更新周期为4小时。

极端情况说明：

- 如果库下没有一个表设置质量规则，则评分全部为默认25分。
- 对于设置了质量规则的表评分，使用最近一次运行完的实例结果计算字段评分。如果运行一次后没有再运行该表相关的质量任务，该表的评分会一直为本次运行结果计算出的评分。
- 质量报告是周期性更新的，对于当天最后一个周期内执行的质量任务，可能会第二天才更新。

表一 技术报告详情说明

名称	描述
主题/源连接得分、库得分	如果选择的是一个主题/源连接，则显示当前主题/源连接的整体得分，即所有库评分的平均值。如果选择的是一个数据库，则显示当前数据库的整体得分，即库下所有表评分的平均值。
历史得分趋势	选中的主题/源连接或者库的历史得分趋势。
库评分、表评分	如果选择的是一个主题/源连接，则显示当前主题/源连接的所有库的得分详情。如果选择的是一个数据库，则显示当前数据库的所有表的得分详情。配置规则数是当天该库或表已有的所有规则总数。告警规则数和告警状态数量是仅当天的告警数据统计，不包含历史告警数量。

查看业务报告

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。在左侧导航栏，单击**数据治理>数据质量>质量报告**。
2. 在左侧目录选中业务报告，从左侧各数据源类型中选择一个，可查看对应主题/源连接或库的评分。
3. 单击主题域评分部分中，**告警状况**列下的蓝字按钮，可跳转至运维管理列表页。
- 4.单击**导出报告**按钮，即可完成业务报告的word格式导出。

🔗 质量作业

创建质量作业

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。在左侧导航栏，单击**数据治理>数据质量>质量作业**，进入质量作业界面。
2. 从左侧各数据源类型中选择一个，并且选中一个数据库。单击**新建质量作业**进入创建质量作业页面。
3. 选择一个表，输入作业名称，规则配置添加对应质量规则配置，单击**确认**按钮完成创建。

表一 配置项说明

名称		含义	使用限制
表名称		选择一个需要进行质量校验的表。	
作业名称		数据质量作业名称。	不大于30个字符。
扫描范围		选择全表扫描或条件扫描。条件扫描需要输入过滤条件（where 后语句）。	
规则配置	规则选择	根据筛选框给表或字段添加对应粒度、分类下的规则。表级规则无需选择字段，字段级和跨字段级规则需要选择一个字段作为探查对象。	
	输入值	如果所选规则有自定义输入值，则需要根据提示填入对应的值。 系统内置规则模板包括正则表达式、取值范围、枚举值等需要输入值。自定义规则则根据实际逻辑填入。	
	SQL解析	展示所选规则的sql逻辑。	
	其他字段参数	如果所选规则需要关联其他数据源，则选择其他字段参数。 系统内置规则模板中一致性校验需要选择参考数据源。自定义规则根据实际逻辑选择对应数据源。	
	告警条件	根据规则对应输出提示配置三级（L1 提示、L2 警告、L3 严重）告警条件。	
	开启任务阻塞	如果开启任务阻塞，该项规则校验产生 严重 告警时会阻塞下游任务节点。	
	生成异常数据	如果系统内置规则支持异常数据保存输出，开启生成异常数据会把对应异常数据保存在配置的异常数据输出路径中。	如果没有配置质量异常数据保存路径，请参考 七、配置管理 先配置质量异常数据保存路径，否则会导致任务执行失败。
描述		数据质量作业描述信息。	不大于100个字符。

举例说明

表二 示例说明

示例类型	
正则表达式 校验	输入值：正则表达式。需要输入自己的正则表达式，这里为 <code>^[a-zA-Z0-9_-]+@[a-zA-Z0-9_-]+(\.[a-zA-Z0-9_-]+)+\$</code>。 SQL解析：展示正则表达式的sql逻辑 其他字段参数：无。 告警条件：该规则提示“请使用【总行数：COUNT】、【无效行数：INVALID_COUNT】、【有效行数：VALID_COUNT】、【无效比率：INVALID_RATE】进行配置”。则告警条件如下： - 提示：<code>#{INVALID_COUNT} > 0</code>，表示无效行数 > 0时产生提示告警。 - 警告：<code>#{INVALID_COUNT} > 10</code>，表示无效行数 > 10时产生警告告警。 - 严重：<code>#{INVALID_COUNT} > 100</code>，表示无效行数 > 100时产生严重告警。 开启任务阻塞：开。如果无效行数 > 100时产生严重告警并阻塞任务流中下游任务执行。 生成异常数据：开。该系统规则支持保存异常数据，即该字段数据不符合输入正则表达式的数据。 - 保存数据量：全部。保存全部异常数据，也可指定保存指定行数的异常数据，如“20”。
字段一致性 校验	输入值：无。 SQL解析：展示字段一致性校验的sql逻辑。 其他字段参数：需要一个关联表和一个关联字段，用于判断数据一致性。 告警条件：该规则提示“请使用【总行数：COUNT】、【一致行数：CONSISTENT_COUNT】、【不一致行数：INCONSISTENT_COUNT】、【一致比率：CONSISTENT_RATE】进行配置”。则告警条件如下： - 提示：<code>#{INCONSISTENT_COUNT} > 0</code>，表示不一致行数 > 0 时产生提示告警。 - 警告：无。 - 严重：<code>#{CONSISTENT_RATE} < 0.9</code>，表示一致比例 < 0.9 时产生严重告警。 开启任务阻塞：开。如果一致比例 < 0.9 时产生严重告警并阻塞任务流中下游任务执行。 生成异常数据：开。该系统规则支持保存异常数据，即该字段数据不在其他字段参数所选关联表和字段的数据中的数据。 - 保存数据量：指定条数，20。
自定义规则 模板	输入值：<code>#{Input1}</code>：90。自定义模板sql中的输入表达式占位符。 SQL解析：改自定义规则模板的sql表达式。 其他字段参数：选择当前表的两个字段。结合输入值，该规则最终表达的含义为score > 90分且name不为null的数量。 告警条件：该规则提示“请使用【表行数：表行数】进行配置”。则告警条件如下： - 提示：无。 - 警告：<code>#{表行数} < 10</code>。 - 严重：无。 开启任务阻塞：关。 生成异常数据：关。自定义规则模板暂不支持保存异常数据。

调度设置

数据质量作业任务运行依赖离线作业组调度执行，需要在离线作业组开发中添加质量任务节点执行。

操作步骤：

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。
2. 在左侧导航栏，单击数据加工>我的项目，查看项目列表。
3. 点击一个项目进入具体开发页面，单击离线开发-作业组开发，新建作业或者选择一个已有作业。选择数据质量-Quality节点拖入画布中。
4. 对于多节点任务或者质量任务需要阻塞下游执行，则串联多个节点。当质量任务开启阻塞并且发生L3严重告警时会阻塞下游执行。
5. 单击Quality节点，配置质量作业节点，需要选择计算资源，然后通过类型-主题/源连接-库-表-质量作业选择一个已经配置好的质量作业。
6. 根据实际情况（该表数据量大小、集群资源），可以修改计算引擎配置参数。包括Driver的cpu个数（默认1）和内存大小（默认1024M），Executor的cpu个数（默认1）、内存大小（默认1024M）和数量（默认1）。注意内存配置需要带上单位（M，G等）。
7. 关于计算引擎参数配置：
 - 默认参数对于常见如单表数据10G以内，可以正常执行。
 - 如果数据量级在100G，建议参数Driver的cpu个数1和内存大小4G，Executor的cpu个数2、内存大小8G和数量2。
 - 对于多表关联规则，如系统内置字段一致性规则模板、自定义多表关联规则，可以适当增加资源。

注意：配置示例仅供参考，如果出现如OOM问题，需根据实际情况调整。

运维管理

数据质量运维管理模块可以查看任务运行结果、告警情况和异常数据预览。

查询作业实例

- 1. 登录并进入百度智能云数据湖管理与分析EasyDAP。在左侧导航栏，单击数据治理>数据质量>运维管理，进入运维管理界面。
- 2. 从左侧各数据源类型中选择一个，并且选中一个数据库，即可查看该数据库下所有表的质量任务运行实例。

表一 展示项说明

展示项名称	描述
表名称	该质量任务实例所属质量作业的基表名称。
质量作业名称	该质量任务实例所属质量作业名称。
实例ID	该质量任务实例ID。
状态	该质量任务实例运行状态（全部、成功、失败、告警、阻塞、运行中。支持按照状态进行筛选。
业务时间	该质量任务实例运行业务时间。单击箭头可进行排序。
运行开始时间	该质量任务实例运行开始时间。

表二 状态说明

质量作业状态包括：运行中、成功、告警、阻塞、失败五种。调度中质量节点的状态有运行中、成功、失败三种。

质量作业状态	调度中质量节点状态	说明
运行中	运行中	作业运行中。
成功、告警	成功	质量作业运行成功或产生告警（L1提示、L2警告），调度中质量节点状态为成功。
阻塞、失败	失败	质量作业发生阻塞（L3严重告警）或者失败，调度中质量节点状态为失败，如果是阻塞状态则会停止执行下游节点。

查看异常数据

- 1. 登录并进入百度智能云数据湖管理与分析EasyDAP。在左侧导航栏，单击数据治理>数据质量>运维管理，进入运维管理界面。
- 2. 从左侧各数据源类型中选择一个，并且选中一个数据库，查看该数据库下所有表的质量任务运行实例。
- 3. 单击实例列表操作栏中的结果查看，进入执行结果查看详情页面。
- 4. 在结果查看详情页面，可以查看任务每个规则的告警状态和异常数据数量、路径，如果有异常数据，单击字段名称下拉按钮可以预览最多20条异常数据。构造表结构，查看不同规则的异常数据展示。

表二 表结构说明

col_name	date_type
id	int
name	string
email	string
score	float

- 5. 示例数据：

```
insert into test_stu.test_stu values(1, 'name1', 'e-name1@test.com', 90);
insert into test_stu.test_stu values(1, 'name2', 'e-name1attest.com', 90);
insert into test_stu.test_stu values(2, 'name2', 'e-name2@test.com', 96);
insert into test_stu.test_stu values(3, 'name3', 'e-name3@test.com', 78);
insert into test_stu.test_stu values(4, 'name4', 'e-name4@test.com', 69);
insert into test_stu.test_stu values(5, 'name5', 'e-name5#test.com', -10);
insert into test_stu.test_stu values(6, 'name6', 'e-name6@test.com', 100);
insert into test_stu.test_stu values(7, 'name7', 'e-name7@test.com', 1000);
insert into test_stu.test_stu values(8, "", "", null);
insert into test_stu.test_stu values(9, null, null, null);
insert into test_stu.test_stu values(10, null, null, null);
insert into test_stu.test_stu values(null, null, null, 90);
insert into test_stu.test_stu values(null, null, null, 90);
```

完整性

- 规则粒度：字段级
- 探查对象：Name
- 规则分类：完整性
- 规则模板名称：字段空值
- 异常数据：异常数据为name字段为NULL或者空字符串的数据。

唯一性

- 规则粒度：字段级
- 探查对象：ID
- 规则分类：唯一性
- 规则模板名称：字段重复值
- 异常数据：异常数据为ID重复的数据，为id=1的两条数据。多个NULL值不算重复。

有效性

- 规则粒度：字段级
- 探查对象：email
- 规则分类：有效性
- 规则模板名称：正则表达式校验
- 异常数据：异常数据为email字段不符合邮箱格式正则表达式的数据。不包含NULL值。

准确性

- 规则粒度：字段级
- 探查对象：score
- 规则分类：准确性
- 规则模板名称：值域范围校验，[0,100]。
- 异常数据：异常数据为score字段不在[0,100]范围内的2条数据。

一致性

- 规则粒度：字段级
- 探查对象：ID
- 规则分类：一致性
- 规则模板名称：字段一致性校验
- 异常数据：异常数据为ID值不在参考表ID字段中的数据。

问题分析整改

对于有质量问题的数据，需要根据校验结果并查看对应异常数据，结合业务定位数据出现问题的原因。

比如完整性问题，一般由业务数据本身缺失导致，或者数据开发过程中如join关联出错导致空值数据问题。业务数据本身缺失导致的需要完善初始业务数据收集，保证数据完整性。数据开发过程中导致的问题需要结合实际处理逻辑定位问题原因，可以借助比如数据血缘根据数据链路排查问题。

🔗 配置管理

针对异常数据的存储路径配置和字段质量评分配置。

操作步骤

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。选择侧边导航**数据治理>数据质量>配置管理**。进入配置管理界面
2. 单击**修改**按钮后选择一个存储路径，然后单击保存即修改成功。
3. 字段质量评分配置单击修改按钮修改后单击保存即修改成功。

质量评分规则说明

默认采用百分制，满分为100分，75-100 优秀、50-75 良好、25-50 合格、0-25 不合格；

字段评分 = $\sum$ 规则设置率评分 * 评分系数 / 规则的数量；

- 规则设置率评分：未设置规则，默认得25分(可配置)，每设置一项规则，默认加25分(可配置)，加满为止(满分100分)；
- 评分系数：无告警为1，提示为0.75，告警为0.5，严重为0.25；

表评分 = $\sum$ 表所有字段的评分 / 字段的数量；

库评分= $\sum$ 库下所有表的评分 / 表的数量；

主题评分 = $\sum$ 主题下所有表的评分 / 表的数量；

主题域评分 = $\sum$ 主题域下所有主题的评分 / 主题的数量。

数据安全

🔒 数据保护

🔒 脱敏规则

内置脱敏算法 EDAP平台目前支持对String、Int、Bigint、Float、Double、Decimal等多种数据类型的字段添加脱敏规则，脱敏规则主要支持哈希算法、遮掩算法、截断算法、变换算法、加密算法。

表一 脱敏规则说明

脱敏规则名称	脱敏规则类别	适用数据类型
SHA256	哈希算法	String
HMAC-SHA256	哈希算法	String
MD5	哈希算法	String
SHA1	哈希算法	String
遮掩前m后n	遮掩算法	String
保留前m后n	遮掩算法	String
遮掩自m到n	遮掩算法	String
保留自m到n	遮掩算法	String
遮掩全部	遮掩算法	String
遮掩邮箱前缀	遮掩算法	String
数值类型截断	截断算法	Float、Double、Decimal
字符类型截断	截断算法	String
数值类型翻倍	变换算法	Int、Bigint、Float、Double、Decimal
DES算法	加密算法	String
3DES算法	加密算法	String
AES算法	加密算法	String
FPE算法	加密算法	String
SM3、SM4	加密算法	String
IDEA	加密算法	String
SM2	加密算法	String
RSA	加密算法	String
DSA	加密算法	String
ECC	加密算法	String

新建脱敏规则

脱敏规则由脱敏算法和配置项组成。

1. 在EDAP首页选择**数据治理>数据安全>脱敏规则**，单击新建按钮进入新建脱敏规则配置页面。
2. 填写对应配置项后单击确定。

表二 新建脱敏规则配置项说明

配置项名称	配置项含义	输入规范
脱敏规则名称	--	支持中文、英文、数字、特殊字符等，长度不超过32字符
脱敏算法类型	--	下拉选择，哈希、遮掩、截断、变换、加密
脱敏算法名称	--	下拉选择
盐值	哈希算法参数配置	支持中文、英文、数字、特殊字符等，长度不超过128字符
参数位置	哈希算法参数配置	下拉选择开头、末尾，默认为末尾
m	m、n为0或正整数；遮掩算法参数配置	默认为0
n	m、n为0或正整数；遮掩算法参数配置	默认为0
截断小数点前n位	n为正整数；截断算法参数配置	默认为1
截断前m后n位	m、n为0或正整数；截断算法参数配置	默认为0，即不截断
翻倍系数	任意数字，支持小数；交换算法参数配置	默认为1
密钥位数	加密、解密密钥位数	下拉选择，16bit、32bit、64bit、128bit，默认16bit
描述	脱敏规则描述信息	支持中文、英文、数字、特殊字符等，长度不超过256字符

脱敏规则操作

表三 脱敏规则操作详细说明

操作名称	描述
查看脱敏规则	单击 脱敏规则名称 ，进入 脱敏规则 详情界面。可以看到“基本信息”、“参数信息”、“引用记录”等内容。
编辑脱敏规则	单击操作列下的 编辑 按钮即可编辑脱敏规则。支持修改脱敏规则名称、描述信息。
删除脱敏规则	单击操作列下的 删除 按钮即可删除脱敏规则。单击删除后弹窗提示是否确认删除。注意：如果有字段/任务使用了该脱敏规则，则不支持删除。

数据动态脱敏

单一脱敏规则

- 单击**数据治理>元数据管理**，单击数据表名称，进入表详情页。
- 单击**添加脱敏规则**按钮，依次选择脱敏列、脱敏规则、脱敏后密级，脱敏规则如果没有可选项，单击前往创建按钮去创建。点击确定完成脱敏规则添加。

表四 配置项说明

配置项名称	配置项含义	输入规范
名称	脱敏规则名称	支持中文、英文字母、数字、中划线、下划线，不超过32个字符。同一个字段，脱敏到同一个密级，脱敏规则名称唯一。
脱敏列	待脱敏字段	下拉选择，目前仅可下拉选择String类型的字段。
脱敏列密级	添加脱敏规则前的数据密级	--
脱敏规则	--	--
脱敏后密级	添加脱敏规则后的数据密级	下拉选择，脱敏后密级必须小于脱敏前密级。
描述	脱敏规则描述信息	支持中文、英文字母、数字、特殊字符等，不超过256个字符。

多级脱敏规则

如果一个字段添加了多个脱敏规则，且脱敏后密级均小于用户密级，则用户可读取权限范围内较大的。

举例说明：假设用户密级为L2，待脱敏列email数据密级为L3，分别为该字段添加了两个脱敏规则：

- 脱敏规则1为遮掩（全部），脱敏后数据密级为L0。
- 脱敏规则2为遮掩（邮箱前缀），脱敏后数据密级为L1

则该用户可查询数据范围为遮掩（邮箱前缀）。

动态脱敏生效

完成上述配置后，用户在数据查询时，便可按照脱敏规则情况进行数据动态脱敏。

注意事项：如果使用Hive作为数据分析引擎，需要在Hive集群上进行相关配置。

数据密级设置

EDAP平台支持L0-L4五级数据密级，分别表示绝密、机密、秘密、敏感、公开，用户可根据国家、行业或企业内部要求，对敏感数据进行分级管理，默认数据密级为L0。

1. EasyDAP首页选择**数据治理>元数据**，进入元数据管理界面。
2. 单击数据表名称，进入表详情页。单击标签页**元数据信息**查看详细元数据信息。
3. 查看字段定义，单击密级列**编辑**按钮，下拉选择字段密级。单击确认，完成数据密级设置。

表一 数据密级说明

数据密级	密级含义	定义标准
L4	绝密	数据受到破坏后，会对国家安全造成特别严重损害。
L3	机密	数据受到破坏后，会对社会秩序和公共利益造成特别严重损害，或者对国家安全造成严重损害。
L2	秘密	数据受到破坏后，会对公民、法人和其他组织的合法权益产生特别严重损害，或者对社会秩序和公共利益造成严重损害，或者对国家安全造成损害。
L1	敏感	数据受到破坏后，会对公民、法人和其他组织的合法权益产生严重损害，或者对社会秩序和公共利益造成损害，但不损害国家安全。
L0	公开	数据受到破坏后，会对公民、法人和其他组织的合法权益造成损害，但不损害国家安全、社会秩序和公共利益。

用户密级设置

EDAP平台支持设置用户密级，在数据查询时，根据用户密级和数据密级的大小关系，决定用户对数据的访问权限。用户密级默认为L1。

1. EasyDAP首页侧边导航选择**管理中心>权限管理**，进入权限管理界面。
2. 在目录列选择**用户维护**，单击左侧用户/用户组，点击密级**编辑**按钮，下拉选择字段密级。选择完毕后单击确认后完成数据密级设置。

表二 用户密级设置说明

情况	数据查询结果
用户密级 >= 脱敏前密级	脱敏前数据
脱敏后密级 <= 用户密级 < 脱敏前密级	脱敏后数据
用户密级 < 脱敏后密级	无法查询数据

数据密级管理

1. EasyDAP首页侧边导航选择**数据治理>数据安全>密级管理**，进入数据密级列表界面。
2. 在列表界页面单击**编辑密级**，进入密级编辑页面，单击**新增密级**，可向所在行密级上方添加一行新密级，数据密级可自增为至L9。

🔒 静态脱敏

创建脱敏任务

1. 单击首页左侧导航栏**数据治理>数据安全>静态脱敏**，在数据静态脱敏中点击新建创建脱敏任务。
2. 填写基本信息、源端配置、脱敏规则配置和目的端配置信息。

表一 创建脱敏任务基本信息配置项说明

配置项名称	配置项含义	输入规范
任务名称	静态脱敏任务名称	支持中文、英文字母、数字、中划线，不超过32个字符。
描述	静态脱敏任务描述信息	支持中文、英文字母、数字、特殊字符等，不超过256个字符。

表二 源端配置说明

配置项名称	配置项含义	输入规范
数据源类型	源端数据源类型	下拉选择，Hive、EDAP、Doris
数据源名称	源端数据源名称	--
数据库	源端数据库	--
数据表	源端数据表	--
扫描范围	源端配置扫描范围	可选全局扫描或局部扫描，默认为全局扫描，局部扫描支持logicalStartTime语法规范

表三 脱敏规则配置说明

配置项名称	配置项含义	输入规范
数据源类型	源端数据源类型	下拉选择，Hive、EDAP、Doris
数据源名称	源端数据源名称	--
数据库	源端数据库	--
数据表	源端数据表	--
扫描范围	源端配置扫描范围	可选全局扫描或局部扫描，默认为全局扫描，局部扫描支持logicalStartTime语法规范

表四 目标端配置说明

配置项名称	配置项含义	输入规范
数据源类型	目标端数据源类型	下拉选择，Hive、EDAP、Doris
数据源名称	目标端数据源名称	--
数据库	目标端数据库	--
选择已有表	写入已有表时，字段进行同名映射，如果对应不上，执行时会报错	--
自动建表	默认填充表同名映射，支持修改，点击“同名映射”可快速填充	表名称输入规范：支持字母、数字、下划线，不超过32个字符

执行脱敏任务

1. 在概览页选择**我的项目**进入项目列表，单击项目进入项目详情。
2. 在侧边导航单击**数据加工>离线开发>作业组开发**，选中任务开发文件夹，单击**新建作业组**按钮。选中新建好的作业组，在节点列单击**数据安全>数据脱敏**，进行节点配置。
3. 单击任意节点，配置参数。

表五 配置项说明

配置项	输入规范
节点名称	与现有保持一致
计算资源	与现有保持一致
脱敏任务名称	下拉选择，支持检索

🔗 数据申请

🔗 存储路径申请

存储路径是指，数据湖指定的存储路径，用于数据表数据存储

通过“存储路径申请”功能，普通用户可以申请访问和操作特定存储路径的权限，确保数据安全并防止未经授权的访问。授权时长可根据需要灵活设置，以满足不同场景的需求。

存储路径申请

单击权限申请按钮，弹出批量申请弹窗，用户可选择特定的存储路径并为其申请读取、写入、创建和管理权限。

字段	字段描述
选择存储路径	选择待申请的存储路径，支持按照数据源名称搜索。
权限	支持权限授予用户读、写、新建、管理权限 - 读：可读取已选数据湖数据 - 写：可增加、删除和修改数据湖 - 新建：可在该数据湖建表 - 管理:可进行编辑、授权
有效期	配置申请的有效期，配置永久访问、临时授权，支持灵活控制用户访问时长
授权原因	输入申请原因，最多输入200字

查看权限 支持按照存储路径列表，查看当前存储路径已授权的用户/用户组，查看用户的访问时间和管理权限。

用户可以通过存储路径申请面查看和申请存储路径的权限：

- 存储路径列表：提供一个存储路径的链接列表，方便用户快速管理相关数据源。
- 查看已授权用户/用户组：用户可以查看哪些用户或用户组被授予了访问特定存储路径的权限。

- 查看权限有效期：可以查看用户访问存储路径的时间记录，以便于进行安全审计。
- 管理权限：数据管理员可以设置和管理用户或用户组的权限，包括读、写和管理等。

🔗 元数据申请

在元数据申请功能中，您可以申请对数据库、数据表或数据列的权限，以便进行相应的数据操作和管理。

申请库权限 选中需要申请的数据库，单击申请按钮，申请选中数据库权限：

字段	字段说明
所选数据库操作权限	支持申请读、写、新建、修改、删除、管理权限： ● 读：可读取已选数据库 ● 写：可以增加、删除和修改数据 ● 新建：可以在该库建表 ● 修改：可以修改库下所有表结构 ● 删除：可以删除库下所有表 ● 管理：以上所有，此外还可以对该库/表进行编辑、权限审批、权限授予操作
有效期	配置申请的有效期，配置永久访问、临时授权，支持灵活控制用户访问时长
授权原因	输入申请原因，最多输入200字

申请表权限 选中需要申请的数据表，单击申请按钮，申请选中数据表权限：

字段	字段说明
所选数据表操作权限	支持申请读、写、管理权限： ● 读：可读取已选数据库 ● 写：可以增加、删除和修改数据 ● 管理：以上所有，此外还可以对该库/表进行编辑、权限审批、权限授予操作
有效期	配置申请的有效期，配置永久访问、临时授权，支持灵活控制用户访问时长
授权原因	输入申请原因，最多输入200字

申请列权限 选中需要申请的数据表下的某列，单击申请按钮，申请选中数据列权限：

字段	字段说明
所选数据表操作权限	支持申请读、写权限： ● 读：可读取已选数据 ● 写：可以增加、删除和修改数据
有效期	配置申请的有效期，配置永久访问、临时授权，支持灵活控制用户访问时长
授权原因	输入申请原因，最多输入200字

🔗 数据源申请

在企业或组织中，数据的管理和使用是至关重要的，为保障数据的安全性，通常需要对数据访问权限进行严格控制。

数据源权限申请功能允许用户根据自身的业务需求，申请对特定数据源的访问权限，包括访问和管理权限。

权限申请 用户可以根据实际工作需要申请相应的权限，而不是一刀切地分配权限，提高了资源使用的灵活性。 通过申请和审批流程，可以明确数据访问的责任人，便于事后审计和责任追究。

字段	字段描述
选择数据源	选择待申请的数据源，支持按照数据源名称搜索。
权限	支持权限授予用户访问、管理权限 - 访问:可连接已选数据源 - 管理:可编辑、删除已选数据源链接
有效期	配置申请的有效期，配置永久访问、临时授权，支持灵活控制用户访问时长
授权原因	输入申请原因，最多输入200字

查看权限 支持按照数据源链接列表，查看当前数据源已授权的用户/用户组，查看用户的访问时间和管理权限。

用户可以通过数据源申请面查看和申请数据源的权限：

- 数据源链接列表：提供一个数据源的链接列表，方便用户快速管理相关数据源。
- 查看已授权用户/用户组：用户可以查看哪些用户或用户组被授予了访问特定数据源的权限。
- 查看权限有效期：可以查看用户访问数据源的时间记录，以便于进行安全审计。
- 管理权限：数据管理员可以设置和管理用户或用户组的权限，包括读取、写入和管理等。

🔗 申请记录

用户可以通过申请记录功能，查看和管理他们的申请信息。这包括但不限于源链接、元数据、存储路径和权限设置。此外，系统还提供实时的申请状态和审核进度更新，以及编辑申请内容的选项，确保用户能够轻松跟踪和管理申请流程。

申请列表 申请记录，查看当前用户申请的源链接、元数据、存储路径权限等申请详情。支持查看申请详情，例如申请状态、审核状态，支持编辑申请内容。

字段	字段说明
申请对象	展示用户申请对象，包括源链接、存储路径、数据库、数据表、数据列等
申请权限	展示用户申请的权限类型，包括访问、管理、读写等权限
申请用户	记录申请用户信息
申请时间	记录用户的申请时间
申请状态	展示当前申请记录的状态，包括审核中、通过、不通过
申请时间	记录用户的申请时间
操作	支持查看详情、重新编辑

申请详情 单击查看详情，支持用户查看申请详情：

字段	字段说明
申请对象	展示用户申请对象，包括源链接、存储路径、数据库、数据表、数据列等
申请权限	展示用户申请的权限类型，包括访问、管理、读写等权限
权限时效	永久、时间段
申请原因	记录用户的申请时间
申请状态	展示当前申请记录的申请流程

申请详情

×

申请信息：

申请对象	申请权限	权限时效
源连接: mysql_connection	访问	永久
源连接: hive_connection	访问	永久

申请原因： 申请数据

申请状态：

● 提交申请

提交时间：2024-09-18 19:31:09

申请用户：user_idaas

● 审核中

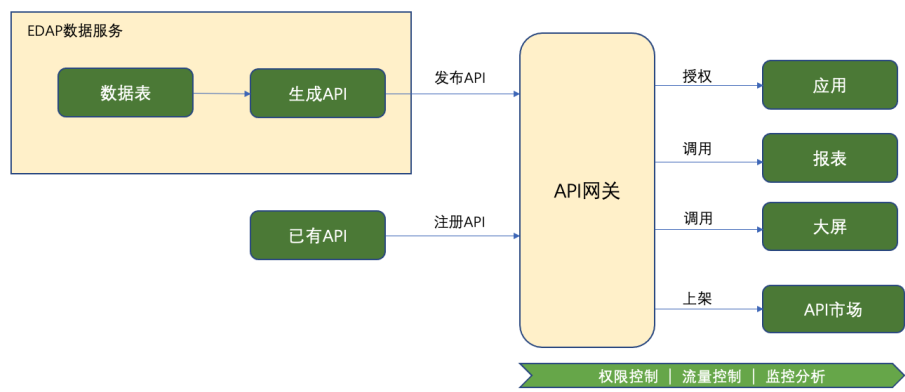
数据服务

概述

数据共享场景概述

数据共享是将企业中的数据通过服务化的方式发布出去，提供给数据使用方调用查询。EDAP 数据共享通过 SQL 方式将数据发布为 HTTP API 接口，可以一键发布 API 数据共享服务到 API 网关，API 网关提供了 HTTP 访问、维护、监控等能力。EDAP 数据共享和 API 网关共同提供了一个易用、安全、可控、可维护的数据共享场景。目前 EDAP 数据共享支持创建为数据共享 API 的数据源有 Doris、MySQL、ClickHouse、GreenPlum。

数据共享流程如下：

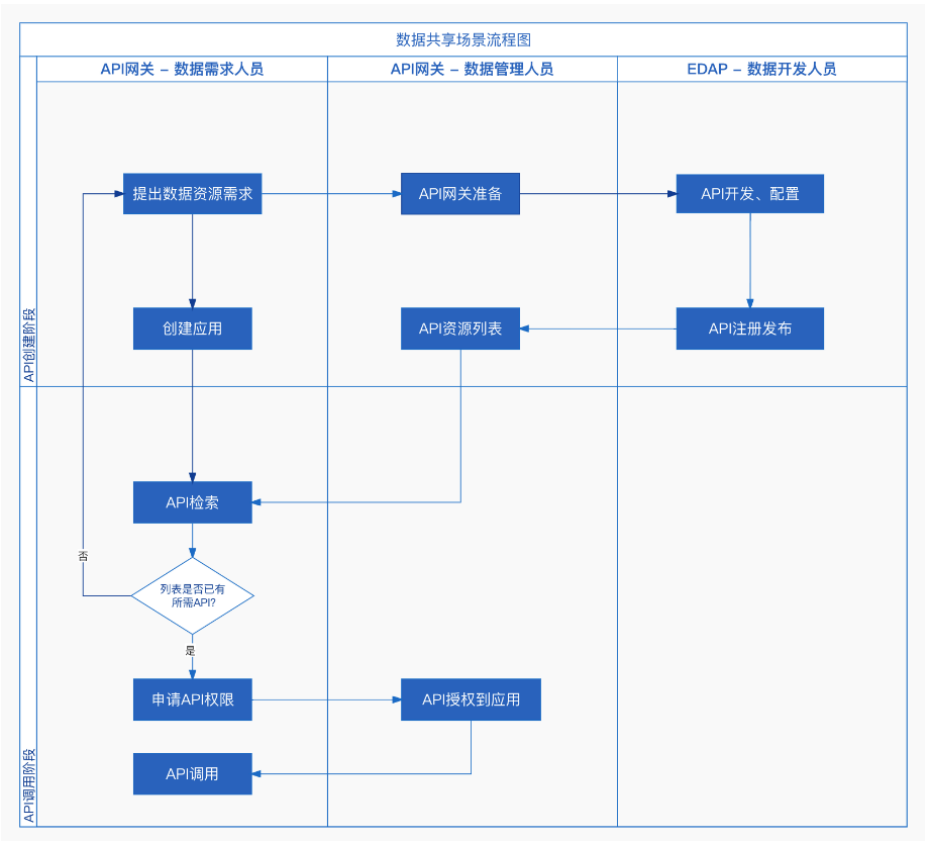


名词解释

数据共享 API：将数据表中数据根据自定义 SQL 脚本发布成 HTTP 接口，接口提供了请求参数以及返回参数的格式。通过调用该 API 接口，使用不同的请求参数值，可以查询到不同的数据。

API 网关API：API 网关是一项云上 API 托管的服务，为开发者提供 API 的创建、维护、发布、监控等整个生命周期的管理。通过 API 网关，您可以封装后端各种服务，以 API 的形式，提供给各方使用。同时 API 网关协助您完成 API 文档管理、API 测试和 SDK 生成等。

角色、职责与权限



角色	职责描述	功能权限
数据需求人员	● 根据需要提出数据资源共享需求。 ● 通过数据共享 API 进行数据调用访问，对于获取的数据资源在授权范围内进行使用。	● 创建 API 网关应用 ● 使用 API 网关应用提供的 AK/SK 调用数据共享 API
数据管理人员	● API 网关准备，创建 API 网关分组，后端密钥。 ● 绑定数据共享 API 到后端密钥。 ● 将数据共享 API 授权给数据需求方创建的数据应用。	● 创建、查看、修改业务流程 ● 创建、查看、修改后端密钥，绑定数据共享 API ● 创建、查看、修改应用，授权数据共享 API
数据开发人员	● 数据资源采集、开发和封装，根据需求方确定的共享内容，进行数据共享 API 开发、注册、发布。	● 创建、查看、修改 API 网关分组 ● 创建、查看、修改、测试、发布数据共享 API

API创建

🔗 BLB准备

1. 开通负载均衡 BLB 服务。点击[去开通](#)。
2. 在负载均衡 BLB 中创建应用型实例。详见[创建应用型BLB实例](#)。

换绑BLB应用型实例

1. 选择对应文件夹旁边操作按钮，单击修改业务流程。
2. 在应用型实例选择需要换绑的选项后确定即换绑成功。注意：换绑后，业务流程下所有API均需要重新发布，请谨慎换绑。
3. 换绑BLB应用型实例后，API内容已修改，API变为待发布状态，支持重新发布。

🔗 API网关准备

开通API网关

数据共享服务需通过[API网关](#)转发到 EDAP 的，因此需要确保当前用户或子用户已经开通了 API 网关服务的权限。目前在完成实名认证后即可开通，[开通 API网关](#)。

创建API网关分组

API 网关的分组管理功能可以高效地、便捷地管理一组具有关联的 API。

操作步骤：

1. 登录并进入 API 网关 API GW。
2. 在左侧导航栏，单击API网关-分组管理，[分组管理文档](#)。
3. 单击新建分组。
4. 填写分组名称和分组描述信息，单击确认，即可完成分组创建。
5. 在创建完API网关分组后，会默认为每个网关分组提供一个内网域名。如需使用自定义域名，可点击下方绑定域名按钮，具体操作流程[绑定自定义域名流程](#)。

表一 表单项说明

名称	描述	使用限制
分组名称	网关分组名称	必填，分组名称必须唯一，支持汉字、英文字母、数字、英文格式的下划线，必须以英文字母或汉字开头，4~50个字符
描述	网关分组描述	非必填，0-180个字符

创建后端密钥

为保证后端收到的请求来源可靠，网关提供后端密钥签名验证功能，创建密钥并将密钥绑定到具体的 API 上即可开启后端签名。在登陆用户的“安全认证”页中，有对应的 Access Key 和 Secret Key ，[获取 AK/SK文档](#)。

操作步骤：

1. 登录并进入 API 网关 API GW。

2. 在左侧导航栏，单击API网关-后端密钥管理，[后端密钥管理文档](#)。
3. 单击“创建密钥”。
4. 填写完成后端密钥的信息。例如，创建密钥 xwd_test_sc。

表二 表单项说明

名称	描述	使用限制
密钥名称	密钥名称	必填，支持汉字，英文，数字，下划线，且只能以英文或汉字开头，4~50个字符
Key	Access Key	必填，支持字母，数字，下划线，中划线，！，@，#，\$，%，字母或数字开头，16~64个字符
Secret	Secret Key	必填
确认secret	Secret Key	必填

🔗 创建EDAP数据共享API

新建业务流程

业务流程为 EDAP 数据共享中一组相同的 API 的分组集合，在新建业务流程时需要关联一个 API 网关分组。

操作步骤：

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。
2. 在左侧导航栏，单击数据服务>API创建。
3. 单击新建按钮，下拉选择新建业务流程。
4. 填写业务流程名，并选择在 API 网关中创建的网关分组。例如，这里创建名为 xwd_test_group 的业务流程，并关联 xwd_test API 网关分组。

表三 创建业务流程说明

名称	描述	使用限制
业务名称	业务流程名称	必填，最大长度限制 100 字符
API 分组	选择在 API 网关中新建的网关分组	必填
应用型BLB实例	请选择已开通的BLB应用型实例	必填
描述	业务流程描述	非必填，最大长度限制 300 字符

新建API

向导模式

操作步骤：

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。
2. 在左侧导航栏，单击数据服务-API创建。
3. 悬浮创建好的业务流程名，点击选新建API，或者单击上方的新建按钮，选择新建API。
4. 选择向导模式或脚本模式，并填写api名称业务流程等详细信息。

表四 配置项说明

名称	描述	使用限制
API名称	API 名称	必填，长度限制 4 - 50 字符
业务流程	选择配置该 API 存储的业务流程	必填
API Path	通过 API 网关访问该 API 的 HTTP 路径	必填，在同一个网关分组中 Path 需唯一。Path 必须以 / 开始，只能包含字母、数字、下划线、中划线，Path 中 / 不超过 70 个
协议	通过 API 网关访问的协议，支持 HTTP 和 HTTPS 两种方式	必填
请求方式	通过 HTTP 访问 API 的 METHOD，支持 HTTP GET 和 POST 两种方式	必填
返回类型	JSON返回参数的响应体的格式，目前支持	必填
描述	API 描述	非必填，最大长度限制 150 字符

选择表：

参数	输入规范
数据源类型	下拉选择，可选Doris/Greenplum/Clickhouse/ MySQL
数据源名称	下拉选择，与数据源管理中对应数据源类型的已有数据源目录保持一致
数据表名称	下拉选择，与选定数据源类型中的数据表名称保持一致

参数配置：

参数	输入规范
字段名称	--
字段类型	--
描述	--
设为请求参数	默认不选
设为返回参数	默认不选
添加到排序字段	默认不选

字段排序：

参数	输入规范
排序	--
字段名称	--
排序方式	下拉选择，默认升序
排序操作	--

请求参数/返回参数：

参数	输入规范
参数名称	与参数配置表中被选为请求参数的字段名称保持一致
参数显示名	--
参数类型	下拉选择，默认同步选定的数据表内参数配置对应的字段
参数位置	下拉选择：query/body，默认为query
是否必填	默认必填
操作符	下拉选择，可选列表：= / LIKE / IN / NOT IN / != / > / < / >= / <=，默认为=
示例值	默认为空
默认值	默认为空
描述	--
添加行数筛选	默认不选

返回行数筛选：

参数	输入规范
返回起始行	默认为空，仅可填写正整数
返回终止行	默认为空，仅可填写正整数
操作	--

转换为脚本模式

点击“转为脚本模式”可将向导模式API转换为脚本模式，向导模式已编辑的参数配置信息将会以代码的形式同步到脚本模式界面。

配置API

在新建完 API 之后，需要选择对应的需要查询的数据源，EDAP 提供了更为灵活的 SQL 脚本模式来查询数据并发布成数据共享服务。

操作步骤：

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。
2. 在左侧导航栏，单击数据服务-API创建。
3. 选择创建好的业务流程和API。例如，这里选择创建的 xwd_test_api_01。

表五 表单项说明

名称	描述	使用限制
数据源类型	选择需要查询的数据源类型	必填，目前支持 Doris、MySQL、ClickHouse、GreenPlum 数据源
数据源	选择数据源类型下的需要查询的数据源连接	必填
查询 SQL	用于查询数据源 SQL 语句，其中 SELECT 后字段为返回参数，SQL 语句中由 \${ } 中的字段为请求参数。	必填 支持同一数据源下的单表查询、多表关联查询和嵌套查询
	例如：select empno, ename, job, mgr, hiredate, sal, comm, deptno from emp where empno= \${param_empno}	不支持多条SQL语句
	请求参数：param_empno	只支持 SELECT 查询语句，不支持 INSERT、UPDATE 和 DELETE 等语法
	返回参数：empno, ename, job, mgr, hiredate, sal, comm, deptno	不支持 SELECT * 语法，必须明确指定查询的列 如果查询列的列名带有表名前缀或使用聚合函数（min、max、sum和count等），必须取别名作为返回参数名

4. 配置请求参数：在根据规则配置完成 SQL 模版之后，单击右侧请求参数，并单击自动解析会根据规则解析出对应的请求参数。请求参数页面表单展示分页参数，用户可通过勾选（选择是否需要开启结果分页）。

注意：

- 可通过勾选框开启结果分页，默认为关闭状态。前两行pageNum和pageSize作为分页参数，其【参数位置】、【示例值】、【默认值】均为可编辑状态，【示例值】和【默认值】默认对应填写10和1；
- 若不开启结果分页，将默认一次性返回所有数据且上限为1000；
- 开启结果分页可将API返回的数据分成多页，可以根据特定的需求获取所需的数据，而不必一次性获取所有数据；若开启结果分页，建议添加排序信息以确保每次查询返回顺序相同的结果。

表六 表单项说明

名称	描述	使用限制
参数名称	默认为 SQL 中 \${} 中的参数。 - pageSize - pageNum - totalNum	必填
参数类型	参数类型，支持 STRING、INT、LONG、FLOAT、DOUBLE、BOOLEAN 类型	必填
参数位置	指定在通过 HTTP 请求时候解析该参数的位置，支持 QUERY、BODY 两种类型	必填 - QUERY：最终的请求参数需要放在url中，如 http://xxxx/v1/data-service-api/default/queryStu?age=10" - BODY：最终的请求参数放在请求body中，使用json的方式发送，如 curl -X POST http://xxxx -d '{"age":10}'
是否必填	参数是否必填	
示例值	给调用者提供参数的示例值	非必填，示例值需满足参数类型
默认值	参数默认值	非必填，示例值需满足参数类型
描述	- pageSize:每页的数据条数，即一页中最多展示多少条数据 - pageNum:当前页码，即请求返回的是第几页的数据 - totalNum:总数据条数，即满足查询条件的数据总共有多少条。	必填

5. 配置返回参数：在根据规则配置完成 SQL 模版之后，单击右侧返回参数，并单击自动解析会根据规则解析出对应的返回参数。

表七 表单项说明

名称	描述	使用限制
参数名称	SELECT 查询的字段名或别名	必填
参数类型	参数类型，支持 STRING、INT、LONG、FLOAT、DOUBLE、BOOLEAN 类型	必填
示例值	给调用者提供参数的示例值	非必填，示例值需满足参数类型
描述	对于返回参数的描述	非必填

测试API

测试 API，可用于验证所配置的查询语句是否满足预期，且只有在测试完 API 之后才能发布 API。

操作步骤：

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。
2. 在左侧导航栏，单击**数据服务-API创建**。
3. 选择创建好的**业务流程**和**API**。
4. 单击对应的 API，进入到 API 配置界面。例如，这里测试 API xwd_test_api_01，输入参数值。
5. 在设置了对应的参数值之后，单击开始测试，可以看到对应查询出来的数据。

发布API 只有测试 API 成功了，才会允许发布 API。发布 API 即将发布到 API 网关供数据使用者调用。

操作步骤：

1. 登录并进入百度智能云数据湖管理与分析EasyDAP。
2. 在左侧导航栏，单击**数据服务-API创建**。
3. 选择创建好的**业务流程**和**API**。

- 4. 单击对应的 API，进入到 API 配置界面。
- 5. 单击**发布**按钮，发布 API。

API应用

🔗 API应用

创建应用

通过API网关创建应用，在API网关对该应用进行授权之后，可以通过应用来访问授权的API，应用有AppKey和 AppSecret的信息，在用户选用应用验证的时候，AppKey、AppSecret 用于计算签名。 **操作步骤：**

- 1. 登录并进入API网关API GW。
- 2. 在左侧导航栏，单击**API网关-应用管理**，[应用管理文档](#)。
- 3. 单击**创建应用**。
- 4. 然后填写创建应用的信息。
- 5. 在应用创建完成之后，API网关会为应用自动生成AppKey和AppSecret，用于后续调用时计算签名。

表一 表单项说明

名称	描述	使用限制
应用名称	应用名称	必填，支持汉字、英文字母、数字、英文格式的下划线，且必须以字母或汉字开始，长度限制为4~26个字符，1个中文占2个字符
描述	应用描述	

授权API给应用

对于已经发布的API，是需要通过API网关应用来进行调用的，因此需要将API授权给对应的应用。 **操作步骤：**

- 1. 登录并进入 **API 网关 API GW**。
- 2. 在左侧导航栏，单击**API网关-API列表**，[API 列表文档](#)。
- 3. 选择从 EDAP 数据共享中发布的 API。
- 4. 单击 API 后的**授权**按钮。
- 5. 选中授权环境、授权有效时间和授权应用。例如，将刚刚发布的 xwd_test_api_01 授权给应用 xwd_test_app。
- 6. 在授权给应用之后，在应用管理中选择授权的应用可以查看已经授权的API列表。

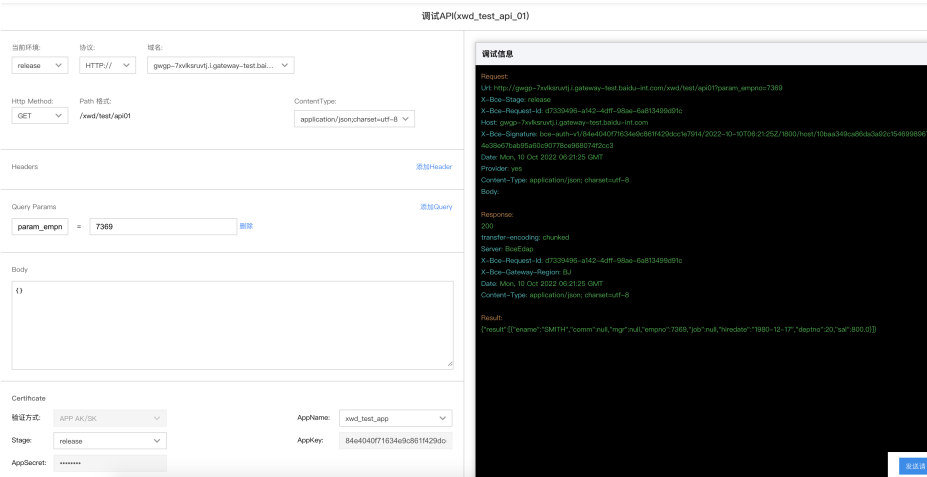
绑定后端密钥 应用通过API网关进行调用EDAP查询数据时，需要通过后端密钥进行认证，因此还需要为API绑定后端密钥。

操作步骤：

- 1. 登录并进入 API 网关 API GW。
- 2. 在左侧导航栏，单击**API网关>后端密钥管理**，[后端密钥管理文档](#)。
- 3. 选择创建好的后端密钥，单击**绑定API**。
- 4. 在对话框中选择发布的 API，并单击确定。例如，将刚刚发布的 xwd_test_api_01 API 绑定后端密钥 xwd_test_sc。
- 5. 在绑定完成之后，在后端密钥管理，单击该后端密钥，可以查询密钥下绑定的API列表。

通过API网关测试 通过 API 网关可以测试数据共享API是否已经可以成功调用。 **操作步骤：**

- 1. 登录并进入**API 网关 API GW**。
- 2. 在左侧导航栏，单击**API网关-API列表**，[API列表文档](#)。
- 3. 单击对应的 API，然后单击**调试API**。
- 4. 根据配置的参数列表，设置完参数，单击**发送请求**，可以查看查询结果。例如，这里测试了刚刚发布的 xwd_test_api_01 API。



API调用管控

访问控制

访问控制是 API 网关提供的安全组件之一，可以配置基于请求来源 IP、用户的账户ID (ACCOUNT ID) 或应用 AK (APP KEY) 的控制策略，决定是否允许调用指定的 API。详见见[API网关访问控制](#)。

流量控制

为了保护 API 开放者的后端服务不因过于频繁的调用导致负载过高，并保证后端资源合理的分配到每个 APP 或用户，API 网关提供了流量控制的功能来限制 API 的调用。详见见[API网关流控管理](#)

监控统计

监控统计，用于监控 API 的调用量、流量、访问时延、响应码的统计，可以更直观的监控 API 的调用量和响应情况。

操作步骤：

- 1. 登录并进入 API 网关 API GW。
- 2. 在左侧导航栏，单击**API网关>监控统计**。
- 3. 选择需要查询的 API 网关分组和 API。例如，这里选择 xwd_test 网关分组和 xwd_test_api_01 API。

管理中心

资源管理

计算资源

EDAP中用户可以将资源注册进入资源管理，统一进行资源的管控。

概述

EDAP平台的计算资源管理是整个平台资源的入口。用户可以添加BMR资源、单机资源、Serverless资源组作为后续作业开发、运行的中需要依赖的资源。

添加计算资源 在【管理中心】中点击【资源管理】，显示现有资源列表。

点击【添加资源】显示添加资源弹窗页面。

目前EDAP支持添加BMR实例、单机服务器、资源组三类资源：

资源类型	引用配置	资源节点	调度引擎
BMR实例	指定bmrlId + yarn queue	全bmr节点	yarn
单机服务器	交互提供Vpc + List + user	单机节点组	NA
资源组	无	cce容器引擎	k8s

BMR实例

EDAP支持添加百度大数据平台BMR作为底层计算资源，添加BMR资源之前需要在BMR产品页面创建BMR集群（可参考[创建集群](#)），BMR集群创建完成后，可选择【BMR】资源，根据要求填写相关信息：

- 计算资源名称：根据实际业务需求可自定义填写当前计算资源名称；
- BMR集群实例名称：下拉选择已经创建的BMR集群实例；

- BMR集群队列：下拉选择已创建的BMR集群队列；
- 计算资源描述：自定义填写计算资源描述。

BMR资源授权

针对BMR资源，需要对当前用户在BMR进行授权。

进入BMR集群管理，点击进入授权使用的BMR集群。

在BMR集群点击【用户管理】，添加BMR账号。

单机资源

EDAP支持基于单机服务器资源运行自己编写的shell、python脚本，选择【单机资源】，根据要求填写相关信息：

计算资源名称：根据实际业务需求可自定义填写当前计算资源名称；

计算资源描述：自定义填写计算资源描述；

配置信息：

- IP地址：填写具体服务器IP地址；

SSH登录认证信息：

- SSH端口号：填写服务器SSH端口号（一般为22）；
- 访问用户：填写服务器访问用户。

环境设置：

- 环境参数：
 - 标准键：JAVA_HOME，填写value值
 - 标准键：python2，填写 python 2.x bin执行文件路径
 - 标准键：python3，填写 python 3.x bin执行文件路径

资源组 EDAP支持Serverless资源组, 用于数据开发、数据湖分析、数据服务。

字段名称	是否必填	输入规范
计算资源名称	必填，计算资源名称	只支持数字、字母、'_'、'-'，且不大于32个字符
网络环境	必填，标识资源网络环境	无
计算资源描述	非必填，计算资源描述	无
规格设置	必填，支持配置常驻资源规格和动态资源规格	host校验
网络设置	必填，配置VPC、子网和安全组	无

资源管理

计算资源详情查看

点击计算资源名称，可查看当前资源配置详情。

资源状态

资源具体状态说明：

- 创建中：后台创建IaaS资源过程（仅Serverless资源组存在该状态）；
- 初始化：资源注册/创建后进入此状态, 后台进行客户端部署、远程服务部署，部署/探测必要引擎组件, 标注应用、环境标签；
- 创建失败：超过配置的最长初始化时间后, 仍然无可用节点，进入此状态；
- 运行中：当前资源可用状态；
- 规格调整中：Serverless资源大小调整中（仅Serverless资源组存在该状态）；
- 不可用：当前资源不可用；
- 禁用中：当前资源禁用中；
- 释放中：清理数据, 销毁RA、远程服务、引擎组件, 释放资源实体节点；
- 已释放：接入的资源已释放或删除成功。

用户操作 对已有资源用户可进行相关操作：

- 编辑：对资源已有配置进行编辑、修改；
- 禁用：对当前资源进行禁用，禁用后资源不可用；
- 删除：删除当前资源，仅禁用中、创建失败、已释放状态的资源可进行删除操作；

项目绑定 资源创建之后，需要将资源绑定到具体的项目，只有绑定到具体项目之后，在项目内才能够使用相应的资源。

- 点击【资源列表】中操作内的【绑定项目】。显示资源和项目的绑定弹窗。在弹窗内选择需要绑定的项目名称。
- 绑定后，在资源页面能够显示，此资源绑定的项目个数。点击【绑定项目】的数字，能够查看具体绑定的项目名称。

🔗 文件管理

通过EDAP前端完成对文件的上传、管理等能力

文件管理 上传文件 进入文件管理列表后，默认选择第一个BOS bucket。需要先下拉选择BOS bucket，然后展示对应bucket下文件列表清单。

点击"上传文件"按钮，支持单个/多个（最多10个）文件上传至BOS，支持文件下载。 文件上传相关限制：

- 1) 支持多个文件上传，最多十个文件；
- 2) 支持单文件最大5G；
- 3) 同文件夹下同名文件直接覆盖

删除文件 支持对已上传的文件进行删除操作：

- 1、支持单个/多个文件删除；
- 2、不支持删除非空文件夹。

文件管理列表 文件管理列表展示目前已上传的所有文件详情：

序号	字段	说明
1	文件名	上传的文件名称
2	大小	显示文件大小，文件夹不显示大小
3	创建人	文件上传人
4	更新时间	文件上传创建时间，文件夹不显示时间
5	文件路径	上传文件所在全路径
6	描述	● 删除：对文件进行删除操作 ● 下载：对文件进行下载操作，不支持文件夹下载

函数管理 自定义函数管理模块支持用户自定义函数注册、管理。**函数管理列表** 支持进行创建、编辑、删除等函数管理操作，支持通过创建人、函数名进行搜索。

序号	字段	说明
1	文件名	上传的文件名称
2	大小	显示文件大小，文件夹不显示大小
3	创建人	文件上传人
4	更新时间	文件上传创建时间，文件夹不显示时间
5	文件路径	上传文件所在全路径
6	操作	使用该UDF的相关示例 ● 删除：对文件进行删除操作 ● 下载：对文件进行下载操作，不支持文件夹下载
7	操作	对当前UDF进行相关操作：删除、编辑

新建函数

基于用户上传的文件，进行UDF创建并发布。

序号	字段	说明
1	数据库	该UDF函数存储数据库
2	函数名称	用户自定义填写UDF名称必填，支持字母、数字、'_'、限50个字符
3	类名	UDF函数的类名必填，支持字母、数字、'_'、'.'、限128个字符
4	文件资源	输入文件资源的路径，支持添加多个
5	描述	使用该UDF的示例代码，非必填，限1024个字符

🔗 系统监控

提供资源的使用分析，查看资源的使用趋势，当前资源正在执行哪些作业，帮助用户解决业务问题，数据每5分钟更新一次。

名词解释：

作业：用户通过离线开发、实时开发功能编写的可视化、脚本、Jar包等统称为作业

作业组：多个作业的组合被称为作业组，暂时仅支持离线作业的组合

任务：作业发布后形成可被手动或定时调度的任务

实例：任务每次执行会形成一个实例，实例状态分为成功、失败、暂停等 **资源使用情况** 统计资源管理中添加的资源，列表中自动展示其资源类型。支持根据资源类型、资源名称、作业开始时间进行检索，查看不同资源的使用趋势。

若资源类型选择为BMR，统计项为：

- 调度作业数
- CPU使用量
- 内存使用量
- CPU使用率
- 内存使用率

若资源类型选择为资源组，统计项为：

- 调度作业数
- CU数

作业执行列表 统计作业的执行情况，展示作业最新的实例执行状态

序号	字段名称	字段说明	备注
1	作业名称	脚本作业名称	
2	作业类型	支持作业类型包括：Shell、Spark、Python、可视化、HIVE SQL、JDBC SQL、Spark SQL、PySpark、Spark Scala、实时流水线、FlinkSQL作业、Flink自定义作业、SparkStreaming Scala、SparkStreaming。	
3	实例ID	任务执行的实例ID	
4	任务名称	脚本所属的任务名称	
5	任务类型	任务类型包括：实时作业、离线作业、离线作业组、工作台临时执行作业	
6	项目名称	任务所属项目	
7	调度时间	实例按照调度计划开始时间	
8	开始时间	实例实际开始的时间	
9	执行时长	实例执行耗时	
10	内存使用率 (使用量)	内存使用情况	
11	CPU 使用率 (使用量)	CPU使用情况	

注：其中**内存使用率（使用量）**、**CPU 使用率（使用量）**：

- 仅支持Spark SQL、PySpark、Spark Scala、FlinkSQL作业、Flink自定义作业、SparkStreaming Scala、SparkStreaming类型作业展示。
- Shell、Python、可视化、HIVE SQL、JDBC SQL、实时流水线任务不支持展示。

数据源管理

🔗 概述

功能介绍

EDPA支持接入十余种数据源类型，包括常见的关系型数据库、非关系型数据库、大数据存储、半结构化存储等。用户可以在源连接管理界面创建与管理源连接，以便后续进行数据同步，实现企业多源异构数据源的数据汇聚。

支持的数据源类型：

数据源类型	数据源类型名称
关系型数据库	MySQL、Oracle、SQLServer、Hana、PostgreSQL、Doris、Greenplum、TiDB
非关系型数据库	Redis、MongoDB、ElasticSearch、Memcached
大数据存储	Hive、HBase、ClickHouse
半结构化存储	FTP
其他	百度云TSDB

角色、职责和权限 角色 | 职责描述 | 功能权限 --- | --- | --- 系统管理员 | 拥有EDAP最高权限。 | -全部功能权限 数据管理人员 | 负责EDAP平台数据源管理，拥有所有数据源的最高权限，进行数据源权限管理。 | -创建数据源链接 -查看、管理所有数据源连接 | 普通用户 | 在权限范围内，访问数据源进行数据集成、数据开发相关操作。 | -在权限范围进行数据源访问、操作

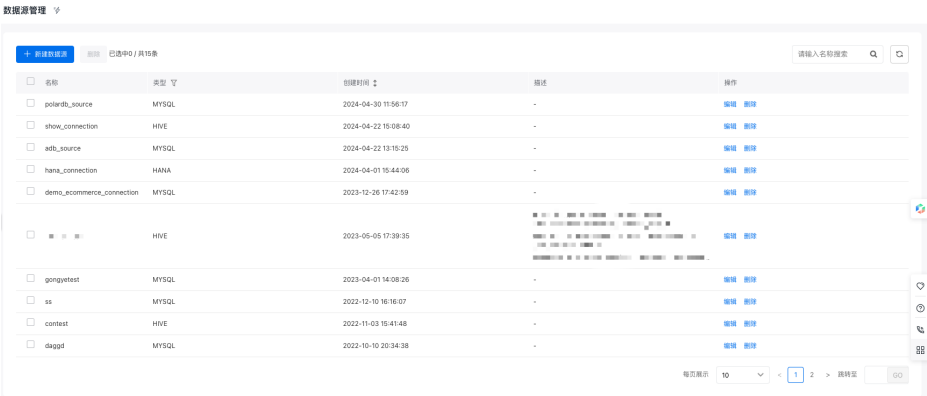
源连接的管理

新增数据源

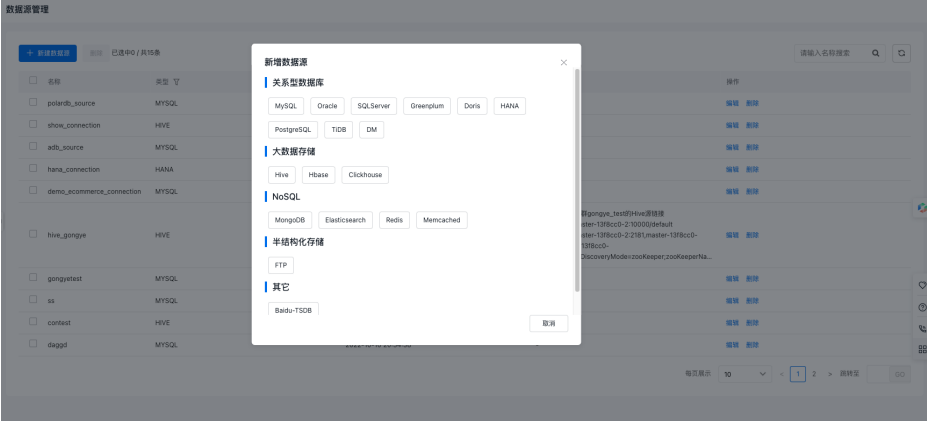
创建源连接是数据接入的第一步，所有接入的数据源都将在「源连接管理」中进行统一管理。源连接中存储了所有建立数据源连接的信息，通过提供正确的源连接名称，用户可以找到相应的数据源连接。

操作步骤：

- 1. 单击“数据源管理 - 源链接管理”，进入源链接管理界面；



- 2. 单击“新建源链接”，选择所需创建的数据源类型；



- 3. 填写表项，完成数据源连接配置；
- 4. 单击“测试连通性”，测试通过后，单击“确认”完成数据源的新建。

创建ORACLE数据源

- 1. 进入源连接管理页面。
- 2. 在源连接管理页面，单击左上角的新增源连接。
- 3. 在新增数据源弹框中，选择关系性数据库类型中的Oracle。

4. 在新增Oracle数据源弹框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
catalog名称	数据源用户数据湖分析时的唯一标识符，以字母、数字、下划线组合，且不能以数字和下划线开头。
主机	数据库对用的主机名。
端口	数据库对应端口号。
SID	Oracle数据库的对应的oracle system ID。
用户名	数据库对应的用户名。
密码	数据库对应的密码。
数据源版本	默认支持oracle11g，可下拉选择oracle11g、oracle12c、oracle18c、oracle19c。
VPC	私有网络(Virtual private Cloud，VPC) 是一个用户能够自定义的虚拟网络，能够帮助用户构建属于自己的网络环境。

创建HBASE数据源

1. 进入源连接管理页面。
2. 在源连接管理页面，单击左上角的新增源连接。
3. 在新增数据源弹框中，选择大数据存储类型中的Hbase。
4. 在新增Hbase数据源弹框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
catalog名称	数据源用户数据湖分析时的唯一标识符，以字母、数字、下划线组合，且不能以数字和下划线开头。
zookeeper的url	hbase所注册的zookeeper地址，例master-78d2977:2181。
zNode根路径	hbase注册zk中的节点名，可从hbase-site.xml配置中获取，默认为“/hbase”。
认证类型	支持SIMPLE和KERNEROS两种模式。
kdc host	密钥分发中心主机地址。
参与者	Kerberos系统中的唯一身份。
keytab文件	Kerberos Keytab 文件包含 Kerberos 主体名称和DES加密密钥之间的映射，这些密钥源自用于登录Kerberos密钥分发中心 (KDC) 的密码。
VPC	私有网络(Virtual private Cloud，VPC) 是一个用户能够自定义的虚拟网络，能够帮助用户构建属于自己的网络环境。

创建HIVE数据源

1. 进入源连接管理页面。
2. 在源连接管理页面，单击左上角的新增源连接。
3. 在新增数据源弹框中，选择大数据存储类型中的Hive。
4. 在新增Hive数据源弹框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
catalog名称	数据源用户数据湖分析时的唯一标识符，以字母、数字、下划线组合，且不能以数字和下划线开头。
jdbcUrl	使用JDBC连接到HiveServer2的地址，格式为jdbc:hive2://:/
hive集群	hive集群所在资源，即edap资源管理中心中对应的资源名。
安全模式	hive集群是否开启了Kerberos认证。
metastoreUris	hive服务设为使用远程 metastore 服务器的URI的格式为：thrift://host:port，可从配置项hive.metastore.uris中获取。
数据湖启用Hive元数据	是否将该hive源的数据纳入edap数据湖。
VPC	私有网络(Virtual private Cloud，VPC) 是一个用户能够自定义的虚拟网络，能够帮助用户构建属于自己的网络环境。

创建Elasticsearch数据源

1. 进入源连接管理页面。
2. 在源连接管理页面，单击左上角的新增源连接。
3. 在新增数据源弹框中，选择大数据存储类型中的Elasticsearch。
4. 在新增Elasticsearch数据源弹框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
catalog名称	数据源用户数据湖分析时的唯一标识符，以字母、数字、下划线组合，且不能以数字和下划线开头。
endpoint	格式为 http://host:port或https://host:port。
用户名	数据库对应的用户名。
密码	数据库对应的密码。
安全模式	hive集群是否开启了Kerberos认证。
是否开启HTTPS认证	即访问方式是网络协议为https。
keyStore路径	keystore file的路径，可上传至bos的某个路径下。HTTPS需要 SSL证书。当您生成 SSL 证书时，您正在创建一个密钥库文件(keystore file)和一个密钥库密码(keystore password)，以便连接时使用。
keyStore密码	当您生成 SSL 证书时，您正在创建一个密钥库文件(keystore file)和一个密钥库密码(keystore password)，以便连接时使用。
trustStore路径	trustStore file路径，可上传至bos的某个路径下 TrustStore仅包含客户端信任的证书。
trustStore密码	访问trustStore的密码。
VPC	私有网络(Virtual private Cloud，VPC) 是一个用户能够自定义的虚拟网络，能够帮助用户构建属于自己的网络环境。

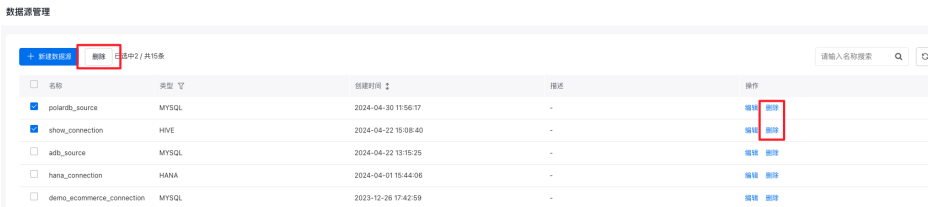
管理数据源

编辑数据源

在源连接界面的每一个条目的右侧可见编辑按钮，通过单击编辑按钮可在界面弹窗中可对本数据源进行编辑。

源连接删除

在源连接界面的每一个条目的右侧可见删除按钮，通过单击删除按钮即可对单个源连接进行删除。



同时也可以通过勾选每个源连接条目，后单击左上角的删除按钮进行批量删除。 注：针对hive数据源类型的源连接，若源连接关联一个元数据同步任务时将无法进行删除。

🔗 网络联通性保障

测试连通性目的为测试每种数据源的连通性，以保证数据源信息的正确性，以及后续数据源工具和数据集成任务可以正常使用与运行。

网络通畅时，界面右上方会弹出相应的反馈框。反之，当网络不通畅是会针对在右上角弹窗给出失败原因。

权限管理

🔗 元数据权限

元数据权限管理是数据治理的一个重要组成部分，对数据元数据（即描述数据的数据）的访问和操作权限的控制，防止未授权访问。实施有效的元数据权限管理，组织可以更好地控制数据资产，提高数据治理的效率和效果。

库表授权 支持管理员为用户或用户组授予库表权限。

支持所选数据库操作权限，支持针对库、表、字段进行授权，授权项：

字段	字段说明
读	可读取已选数据库
写	可以增加、删除和修改数据
新建	可以在该库建表
修改	可以修改库下所有表结构
删除	可以删除库下所有表
管理	以上所有，此外还可以对该库/表进行编辑、权限审批、权限授予操作

授权

* 用户或用户组:

请选择

* 所选数据库操作

权限:

☐ 读

可读取已选数据库

☐ 写

可以增加、删除和修改数据

☐ 新建

可以在该库建表

☐ 修改

可以修改库下所有表结构

☐ 删除

可以删除库下所有表

☐ 管理

以上所有，此外还可以对该库/表进行编辑、权限审批、权限授予操作

额外授权

☐ 关

* 有效期:

永久

临时授权

* 授予原因:

请输入授予原因

0/200

取消

确认

用户授权 支持按照用户维度查看元数据权限情况，修改用户的权限范围、权限有效期，调整用户密级等权限管理操作。

编辑

库表范围

quality_db

权限	有效期
☒ 修改 可以修改库下所有表结构	☒ 永久 ☐ 自定义
☐ 读 可读取已选数据库	☒ 永久 ☐ 自定义
☐ 写 可以增加、删除和修改数据	☒ 永久 ☐ 自定义
☐ 新建 可以在该库建表	☒ 永久 ☐ 自定义
☐ 删除 可以删除库下所有表	☒ 永久 ☐ 自定义

取消

确认

数据源权限

数据源权限管理提供数据源权限管控，对接外部数据源权限，授权时长灵活可控。

支持授予源链接的访问、管理权限。

批量授权

支持管理员为用户或用户组批量授予数据源链接的访问、管理权限，授权项：

字段	字段说明
用户或用户组	选择系统添加的用户或用户组进行授权
权限	● 访问：可连接已选数据源 ● 管理：可编辑、删除已选数据源链接
有效期	支持永久授权和临时授权
授予原因	输入授权原因

查看权限

支持按照数据源链接维度查看用户权限。

字段	字段说明
用户或用户组	有权限的用户或用户组名称
访问	用户的访问时间
管理	用户的使用权限类型
操作	编辑或删除用户的权限策略

🔗 存储路径权限

支持管理员管理存储路径权限，对存储路径权限的访问和操作权限的控制，防止未授权访问，授权时长灵活可控。

批量授权

支持管理员为用户或用户组批量授予存储路径的访问、管理权限，授权项：

字段	字段说明
用户或用户组	选择系统添加的用户或用户组进行授权
权限	● 读：可读取已选数据湖数据 ● 写：可增加、删除和修改数据湖 ● 新建：可在该数据湖建表 ● 管理：可编辑、删除已选数据源链接
有效期	支持永久授权和临时授权
授予原因	输入授权原因

查看权限

支持按照存储路径维度查看用户权限。

字段	字段说明
用户或用户组	有权限的用户或用户组名称
读	用户的是否有读权限
写	用户的是否有写权限
新建	用户的是否有新建权限
管理	用户的是否有管理权限
操作	编辑或删除用户的权限策略

🔗 授权记录

支持管理员查询已完成的授权操作记录，为管理员提供了全面的授权记录查看功能，以便您能够轻松追溯和审计授权操作。

- 支持查看授权记录列表，支持按照授予类型、授予用户、授予操作人、授予时间进行关联查询

字段	字段说明
授予对象	展示授权对象类型以及授权对象名称
权限性质	展示本次授权包含的权限性质，包括：授予、撤销
权限类型	展示本次授权包含的权限类型，包括：访问、管理等
授予用户	展示本次被授权的用户id
授予操作人	展示本次进行授权操作的用户id
授予时间	记录本次授权时间
操作	支持查看授权详情

查看授予详情

- 支持查看授予对象、权限性质、授权原因等内容。

授予详情

授予信息：

授予对象	权限性质	权限类型	权限时效	授予用户
原模型 - 模型，	授予	访问	永久	

授予原因：

审批中心

标准审批

标准审核

在模型审核 Tab 下，共有「待审核」、「我的审核」两个子页面。根据用户的权限不同，每个页面的展示情况也不同。

- 待审核：显示所有待审核的单据，有权限的用户可以对这些单据进行审核
- 我的审核：只有当用户具有审核模型的权限时才展示该页面，展示该用户所有审核过的单据

提交审核

在「标准管理」或者「公共代码」中，所有的「发布」、「修订」、「废弃」操作都会生成一个审核单据。

处理审核 如果用户具有审核标准的权限，在「待审核」页面中查看所有待审核的单据，点击右侧的「审核」操作，可以对其它用户提交的申请单据进行审核。审核结果可以是「通过」或者「驳回」，支持批量审核。

标准审核

待审核我的审核

批量审核

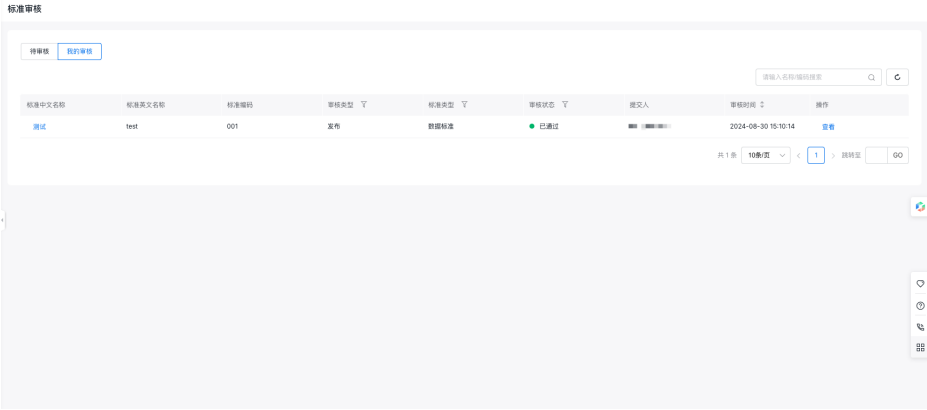
请输入名称和编码搜索

☐	标准中文名称	标准英文名称	标准编码	审核类型	Y	标准类型	Y	提交人	提交时间	操作
☐				发布		公共代码			2024-04-14 15:37:58	审核
☐				发布		数据标准			2024-04-14 15:32:46	审核

共 2 条10条/页1页筛选GO

查看审核

如果用户具有审核标准的权限，在「我的审核」页面中查看所有审核过的单据。



模型审批

模型审核

在模型审核 Tab 下，共有「待审核」、「我的审核」两个子页面。根据用户的权限不同，每个页面的展示情况也不同。

- 待审核：只有当用户具有审核模型的权限时才展示该页面，显示所有待审核的单据，可以对这些单据进行审核
- 我的审核：只有当用户具有审核模型的权限时才展示该页面，展示该用户所有审核过的单据

提交审核

在「关系建模」或者「维度建模」中，所有的「发布」、「修订」、「废弃」操作都会生成一个审核单据。

处理审核

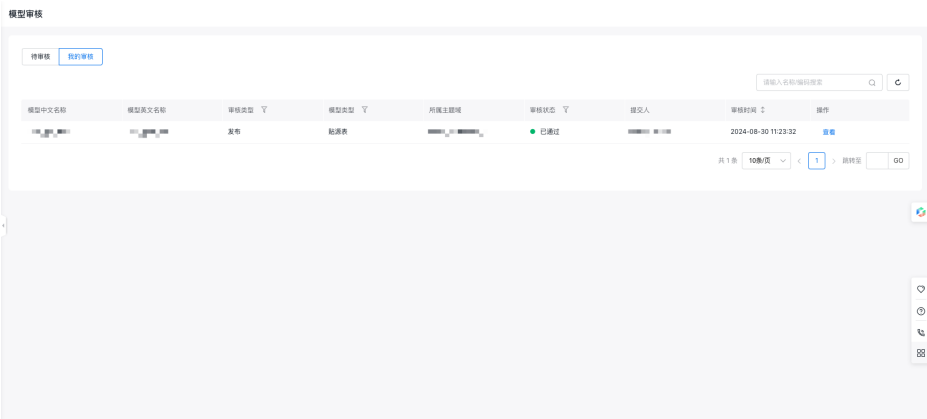
如果用户具有审核模型的权限，在「待审核」页面中查看所有待审核的单据，点击右侧的「审核」操作，可以对其它用户提交的申请单据进行审核：



审核结果包含「通过」或者「驳回」，支持批量审核。

查看审核

如果用户具有审核模型的权限，在「我的审核」页面中查看所有审核过的单据，点击右侧的「查看」，可以查看审核详情：



指标审批

指标审核 在指标审核 Tab 下，共有「待审核」、「我的审核」两个子页面。根据用户的权限不同，每个页面的展示情况也不同。

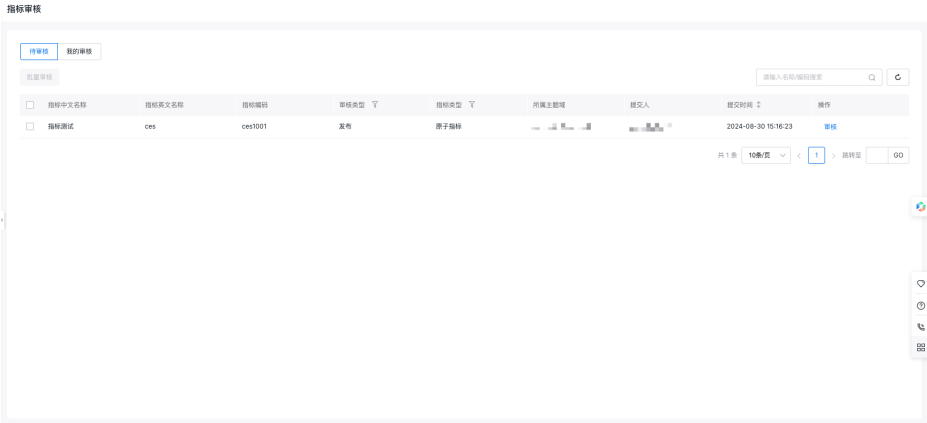
- 待审核：只有当用户具有审核模型的权限时才展示该页面，显示所有待审核的单据，可以对这些单据进行审核
- 我的审核：只有当用户具有审核模型的权限时才展示该页面，展示该用户所有审核过的单据

提交审核

在「原子指标」或者「衍生指标」或者「复合指标」中，所有的「发布」、「修订」、「废弃」操作都会生成一个审核单据。

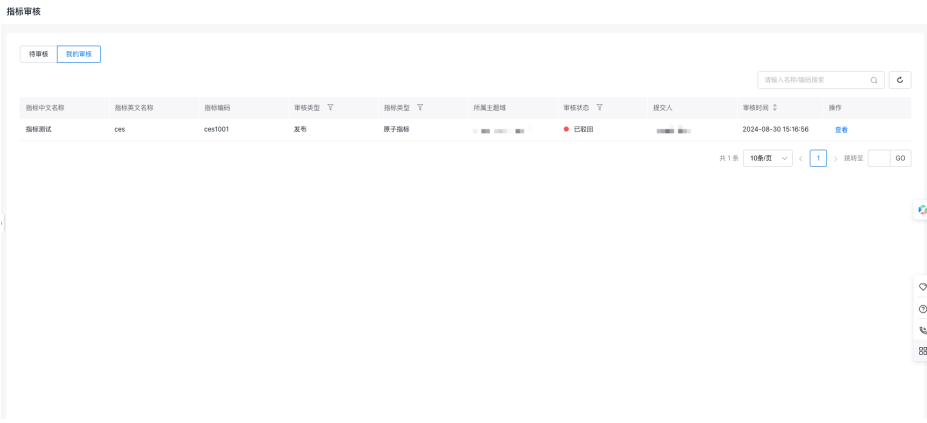
处理审核

如果用户具有审核指标的权限，在「待审核」页面中查看所有待审核的单据，点击右侧的「审核」操作，可以对其它用户提交的申请单据进行审核。审核结果可以是「通过」或者「驳回」，支持批量审核。



查看审核

如果用户具有审核指标的权限，在「我的审核」页面中查看所有审核过的单据，点击右侧的「查看」，可以查看审核详情。



权限审批

权限审批

权限审批模块，提供一个统一的界面，展示所有待审核的单据，包括数据源申请、元数据申请、存储路径申请。允许授权用户高效地处理数据源、元数据、存储路径等资源的申请请求。

提交审核 在「数据安全-数据申请」模块中，进行数据源申请、元数据申请、存储路径申请操作都会生成一个审核单据，支持申请访问、管理等不同粒度的权限。

处理审核 审批列表展示了所有待审批的资源申请，包括以下关键信息：

- 申请对象：申请的具体资源，如数据库、数据表、存储路径等。
- 申请权限：申请者请求的权限级别。
- 申请用户：提交申请的用户身份。
- 申请时间：申请提交的时间戳。
- 操作：用于执行审批操作。

权限审批

申请类型：

全部

申请用户：

全部

申请时间：

重置

查询

批量审批

☐	申请对象	申请权限	申请用户	申请时间	操作
☐	源连接: bian_234_hive	访问		2024-06-19 16:25:21	审批
☐	源连接: bian_234_hive	访问		2024-06-19 16:13:13	审批
☐	存储路径: _edap_test	管理		2024-05-23 18:04:14	审批
☐	存储路径: _edap_test	管理		2024-05-23 18:04:11	审批
☐	源连接: hive_connection	管理		2024-05-23 18:04:04	审批
☐	源连接: hive_connection	管理		2024-05-23 18:03:55	审批
☐	数据表: auto_test_db/datalake_physicl_... 数据表: auto_test_db/datalake_physicl_... 数据表: auto_test_db/datalake_physicl_... 数据列: auto_test_db/datalake_physicl_...	可读 可写 管理 可写		2024-05-23 18:03:49	审批

针对不同的申请类型提供不同的申请权限，例如数据表可申请权限包括可读、可写、管理权限。

权限审批记录

权限审批记录模块为用户提供了一个详尽的审批操作历史查看功能，支持多条件筛选查询，以便用户快速定位和追溯审批内容。

- 多条件查询：支持按照申请类型、申请用户、申请时间、审批用户、审批结果等条件进行综合查询。
- 审批记录列表：清晰展示每项审批的申请对象、申请权限、申请用户、申请时间和审批结果。

权限审批记录

申请类型：

全部

申请用户：

全部

申请时间：

重置

查询

审批用户：

全部

审批结果：

全部

审批时间：

重置

查询

申请对象	申请权限	申请用户	申请时间	审批结果	操作
存储路径: _edap_test	管理	anonymous_cPbiIBFQR67xchS6scaz	2024-08-27 18:13:27	通过	查看详情
存储路径: _edap_test	管理	anonymous_cPbiIBFQR67xchS6scaz	2024-08-27 18:13:07	不通过	查看详情
源连接: hive_connection	管理	anonymous_cPbiIBFQR67xchS6scaz	2024-08-27 18:12:47	通过	查看详情
源连接: hive_connection	管理	anonymous_cPbiIBFQR67xchS6scaz	2024-08-27 18:12:27	不通过	查看详情
数据表: auto_test_db/datalake_physicl_table_ms 数据表: auto_test_db/datalake_physicl_table_ms 数据表: auto_test_db/datalake_physicl_table_ms 数据列: auto_test_db/datalake_physicl_table_ms	可读 可写 管理 可写	anonymous_cPbiIBFQR67xchS6scaz	2024-08-27 18:12:07	通过	查看详情

私有化部署

私有化部署说明

私有化部署是指将EDAP整套服务安装部署到您的企业内网中，可以对接企业内网的数据源，并且支持对接企业的内部账号进行单点登录。

私有化部署版具有当前SaaS版本的所有功能，并且对模型数、指标数、集成任务数、开发任务数、任务调用次数不做任何限制。

您在购买之前，可通过体验版进行[产品试用](#)，同时可以进行[私有化部署咨询](#)，填写完成后会有专人与您对接。

API参考

API概述

通用说明

EDAP服务域名

区域	服务端点	协议	regionID
北京	edap.bj.baidubce.com	HTTP and HTTPS	bj
广州	edap.gz.baidubce.com	HTTP and HTTPS	gz
苏州	edap.su.baidubce.com	HTTP and HTTPS	su
保定	edap.bd.baidubce.com	HTTP and HTTPS	bd

注意：Region(区域)代表着一个独立的地域，是百度智能云中的重要概念，请参考[区域选择说明](#)。百度智能云中的服务除了极少数如账号服务全局有效之外，绝大部分服务都是区域间隔离的，每个区域的服务独立部署互不影响，服务间共享数据需要通过显式拷贝完成，在API中引用区域必须使用其ID。为了保证整体服务的高效，EDAP的服务区域会和BMR的服务区域保持一致。北京区域的EDAP服务可以纳管北京区域创建的BMR资源，以此类推。

168

认证机制

- 所有API的安全认证一律采用Access Key与请求签名机制。 Access Key由Access Key ID和Secret Access Key组成，均为字符串。 对于每个HTTP请求，使用下面所描述的算法生成一个认证字符串。提交认证字符串放在Authorization头域里。服务端根据生成算法验证认证字符串的正确性。 认证字符串的格式为 bce-auth-v{version}/{accessKeyId}/{timestamp}/{expirationPeriodInSeconds}/{signedHeaders}/{signature}
- version是正整数。
 - timestamp是生成签名时的UTC时间。
 - expirationPeriodInSeconds表示签名有效期限。
 - signedHeaders是签名算法中涉及到的头域列表。头域名之间用分号（;）分隔，如host;x-bce-date。列表按照字典序排列。（本API签名仅使用host和x-bce-date两个header）
 - signature是256位签名的十六进制表示，由64个小写字母组成。

当百度智能云接收到用户的请求后，系统将使用相同的SK和同样的认证机制生成认证字符串，并与用户请求中包含的认证字符串进行比对。如果认证字符串相同，系统认为用户拥有指定的操作权限，并执行相关操作；如果认证字符串不同，系统将忽略该操作并返回错误码。鉴权认证机制的详细内容请参见[鉴权认证机制](#)。

通信协议

支持HTTP和HTTPS两种调用方式。为了提升数据的安全性，建议通过HTTPS调用。

公共头域

请求头

头域	说明
host	详见 EDAP服务域名
Authorization	详见 鉴权认证机制
x-bce-date	该请求创建的时间，表示日期一律采用yyyy-MM-dd方式，例如2023-06-01表示2023年6月1日。如果用户使用了标准的Date域，该头域可以不填。当两者同时存在时，以x-bce-date为准。
content-type	默认为: application/json
x-bce-content-sha256	表示内容部分的SHA256签名的十六进制字符串，其中内容指HTTP Request Payload Body，即Content部分在被HTTP encode之前的原始数据。
x-bce-if-match	同If-Match语义。
x-bce-if-none-match	同If-None-Match语义。

响应头

头域	说明
x-bce-request-id	对应请求的requestId

请求结构说明

数据交换格式为JSON，所有request/response body内容均采用UTF-8编码。 请求参数包括如下4种：

参数类型	说明
URI	用于指明操作实体，如:PUT /v1/cluster/{clusterUuid}
Query参数	URL中携带的请求参数
HEADER	通过HTTP头域传入，如:x-bce-date
RequestBody	通过JSON格式组织的请求数据体

响应结构说明

响应值分为两种形式：

响应内容	说明
HTTP STATUS CODE	如200,400,403,404等
ResponseBody	JSON格式组织的响应数据体

错误码

错误格式规范

当用户访问API出现错误时，会返回给用户相应的错误码和错误信息，便于定位问题，并做出适当的处理。请求发生错误时通过Response Body返回详细错误信息，遵循如下格式：

参数名	类型	说明
code	String	表示具体错误类型。
message	String	有关该错误的详细说明。
requestId	String	导致该错误的requestId。

示例

```
{
  "code": "ProjectNotFound",
  "message": "Project test_project is not found",
  "requestId": " 81d0b05f-5ad4-1f22-8068-d5c9de60a1d7"
}
```

公共错误码

详见[百度智能云公共错误码](#)

业务错误码

HTTP状态码	错误码（code）	错误消息（message）	中文描述
403 Forbidden	ProjectNoPermission	No permission to operate project {0}	对项目{0}无操作权限
404 Not Found	ProjectNotFound	Project {0} is not found.	项目{0}不存在。
403 Forbidden	WorkbenchNoPermission	No permission to operate workbench {0}	对工作台草稿{0}无权限
400 Bad Request	InvalidCategory	Category {0} is invalid.	不合法的作业种类 {0}
400 Bad Request	InvalidWorkbenchName	Name of workbench {0} is invalid.	草稿名称{0}不合法
400 Bad Request	DuplicatedWorkbenchName	Name of workbench {0} is duplicated.	草稿名称{0}已存在
404 Not Found	WorkbenchNotFound	Workbench {0} is not found.	草稿{0}不存在
400 Bad Request	InvalidDirectoryName	Name of directory {0} is invalid.	文件夹名称{0}不合法
400 Bad Request	DuplicatedDirectoryName	Name of directory {0} is duplicated.	文件夹名称{0}已存在
400 Bad Request	DirectoryIllegalDepth	A directory can contain a maximum of 5 levels.	文件夹层级最多5层
404 Not Found	DirectoryNotFound	Directory {0} is not found	文件夹 {0} 不存在。
403 Forbidden	JobNoPermission	No permission to operate job {0}	对作业{0}无操作权限
404 Not Found	JobNotFound	Job {0} is not found.	作业 {0} 不存在。
409 Conflict	JobAlreadyExists	Job {0} is already existed	作业 {0} 已存在。
404 Not Found	ExecutionNotFound	Execution {0} is not found.	实例 {0} 不存在。
409 Conflict	ExecutionIsRunning	Execution {0} is running.	实例 {0} 已运行。
409 Conflict	ExecutionAlreadyStopped	Execution {0} was stopped.	实例 {0} 已停止。
404 Not Found	SchedulerNotFound	schedule option is not found	调度配置不存在
404 Not Found	ProfileNotFound	profile {0} is not found.	资源 {0} 不存在
409 Conflict	ProfileIsUnavailable	profile {0} is unavailable.	资源 {0} 不可用

🔗 版本号

参数	类型	参数位置	描述	是否必须
version	String	URI参数	API版本号	必须

🔗 日期和时间

日期与时间的表示有多种方式。为统一起见，除非是约定俗成或者有相应规范的，凡是HTTP标准中规定的表示日期和时间字段用GMT，其他日期时间表示的地方一律采用UTC时间，遵循ISO 8601，并做以下约束：

- 表示日期一律采用yyyy-MM-dd方式，例如2023-06-01表示2023年6月1日。
- 表示时间一律采用HH:mm:ss方式，并在最后加一个大写字母Z表示UTC时间。例如12:00:00Z表示UTC时间12点0分0秒。
- 凡涉及日期和时间合并表示时，在两者中间加大写字母T，例如2023-06-01T12:00:00Z表示UTC时间2023年6月1日21点0分0秒。

🔗 规范化字符串

通常一个字符串中可以包含任何Unicode字符。在编程中这种灵活性会带来不少困扰。因此引入“规范字符串”的概念。一个规范字符串只包含百分号编码字符以及URI（Uniform Resource Identifier）非保留字符（Unreserved Characters）。RFC 3986规定URI非保留字符包括以下字符：字母（A-Z，a-z）、数字（0-9）、连字号（-）、点号（.）、下划线（_）、波浪线（~）。将任意一个字符串转换为规范字符串的方式是：

- 将字符串转换成UTF-8编码的字节流。
- 保留所有URI非保留字符原样不变。
- 对其余字节做一次RFC 3986中规定的百分号编码（Percent-Encoding），即一个%后面跟着两个表示该字节值的十六进制字母。字母一律采用大写形式。示例：原字符串：this is an example for 测试, 对应的规范字符串：this%20is%20an%20example%20for%20%E6%B5%8B%E8%AF%95。EDAP编码规范

EDAP编码规范

- 可解析内容，所有request/response body内容目前均使用UTF-8编码，后续会支持更多encoding类型。
- 在请求时，需要对以下做UrlEncode：
 - Objectname，其中，Resource做UrlEncode的时候需要忽略“/”。
 - Querystring的Value。
 - x-bce-copy-source（忽略“/”）。
 - 自定义Meta:Meta Value只支持可见的ASCII字符，如果需要其它的字符，推荐使用UrlEncode处理。

幂等性

详见[幂等性](#)

API列表

API列表

EDAP提供符合RESTful API规范的自研接口，目前已覆盖数据开发（作业开发、作业运维），其他模块接口逐步开放中。

数据开发接口

数据开发接口说明

类型	API名称	API功能	描述	调用频率限制
作业开发	ImportWorbenchItem	作业批量导入	根据category批量导入工作台草稿，仅支持若干文件夹与json文件压缩的zip	5次/min
	ExportWorbenchItem	作业批量导出	根据category批量导出工作台草稿，导出为一个zip压缩文件	5次/min
	CreateExecution	创建实例	创建临时执行作业实例	10次/min
	RerunExecution	重跑实例	重跑临时执行实例	10次/min
	GetExecutionStatus	获取实例状态	获取执行实例状态信息	10次/min
	GetExecutionLog	获取作业实例日志	获取临时作业执行实例日志。	10次/min
	GetJobExecutionLog	获取作业组实例节点日志	获取临时执行作业组实例的作业节点日志。	10次/min
	GetExecutionJobOutput	获取作业组脚本运行结果	获取作业组中脚本作业运行的数据结果（仅对含有数据结果的脚本作业类型生效）	10次/min
任务运维	DeployJobgroup	发布最新版作业组	上线已发布的最高版本作业组	5次/min
	ListJobgroup	获取作业组列表	获取作业组列表	50次/min
	DeleteJobgroup	删除作业组	删除作业组	10次/min
	KillJobgroupExecution	停止作业组实例	停止正在执行的作业组实例	10次/min

元数据接口

元数据开发接口说明

类型	API名称	API功能	描述	调用频率限制
库	ListDatabases	获取数据库信息列表	获取数据库信息列表	无
	ListDatabaseNames	获取数据库名称列表	获取数据库名称列表	无
	GetDatabase	获取指定数据库信息	获取指定数据库信息	无
	CreateDatabase	新增数据库信息	新增数据库信息	无
	UpdateDatabase	更新数据库信息	更新数据库信息	无
	DeleteDatabase	删除指定数据库	删除数据库，仅允许删除没有表的数据库。	无
表	ListTables	获取表信息列表	获取表信息列表	无
	ListTableNames	获取表名称列表	获取表名称列表	无
	GetTable	获取表信息	获取表信息	无
	CreateTable	创建表信息	创建表信息	无
	UpdateTable	更新表信息	更新表信息	无
	DeleteTable	删除表信息	删除表信息	无
	ListTableVersions	获取表版本信息列表	获取表版本信息列表，表版本从1开始叠加，最大1000个版本，超过1000个版本，滚动删除历史版本信息	无
	GetTableByVersion	获取指定版本的表信息	获取指定版本的表信息	无
分区	ListPartitions	获取表分区列表信息	获取表分区列表信息	无
	ListPartitionNames	获取表分区名称列表	获取表分区名称列表	无
	BatchGetPartition	批量获取表指定分区信息	批量获取表指定分区信息	无
	BatchCreatePartition	批量新增分区信息	批量新增分区信息	无
	BatchAlterPartition	批量更新表分区	批量更新表分区	无
	BatchDeletePartition	批量删除表分区	批量删除表分区	无
	RenamePartition	重命名表分区	重命名表分区	无
函数	ListFunctions	获取函数信息列表	获取函数信息列表	无
	GetFunction	获取函数详情	获取函数详情	无
	CreateFunction	创建函数信息	在指定数据库下创建函数，库下唯一，不能创建同名函数	无
	UpdateFunction	更新函数信息	更新函数信息，除绑定的数据库和函数名不能更新，其余数据字段都可更新	无
	DeleteFunction	删除函数信息	删除函数信息	无
统计信息	SetStatistics	更新表字段/分区统计信息	更新表字段/分区统计信息	无
	GetStatistics	获取字段/分区统计信息	获取字段/分区统计信息	无
	DeleteStatistics	删除字段/分区统计信息	删除字段/分区统计信息	无
数据湖存储	ListLocations	获取数据湖存储信息列表	获取数据湖存储信息列表	无
	GetLocations	获取数据湖存储信息	获取数据湖存储信息	无
	CreateLocation	创建数据湖存储信息	创建数据湖存储信息	无
	UpdateLocation	更新数据湖存储信息	更新数据湖存储信息	无
	DeleteLocation	删除数据湖存储信息	删除数据湖存储信息	无

数据开发API

作业开发

ImportWorbenchItem 接口说明

根据category批量导入工作台草稿，仅支持若干文件夹与json文件压缩的zip

原型

PUT /v1/project/{projectId}/workbench/{category}?import

请求参数

名称	类型	位置	必须	描述
projectId	String	URL	是	项目空间名称
category	String	URL	是	工作台分类的枚举：pipeline、script、jobgroup、realtime、flink 和 sparkstreaming
file	binary	Body	是	仅支持zip压缩方式的包文件

压缩包文件结构

名称	类型	层级	描述
category	文件夹	1	工作台分类的枚举：pipeline、script、jobgroup、realtime、flink 和 sparkstreaming
directory	文件夹	2-5	工作台的文件夹
json	文件	2-6	仅支持*.json类型的文本文件

响应参数

以下是将您提供的表格信息转换为Markdown格式的语法表达：

名称	类型	描述
result		导入结果
totalCount	Integer	总json数量
successCount	Integer	成功json数量
failedCount	Integer	失败json数量
failedItems	List<FailedItem>	错误文件列表

FailedItem

名称	类型	描述
path	String	文件路径或文件夹路径
cause	String	错误原因

错误码

Code	message	HTTP状态码	描述
AccessDenied	No permission to operate project	403	无权限
ProjectNotFound	Project is not found.	404	项目空间不存在

请求示例

```
PUT /v1/project/pro/workbench/script?import
Host: edap.bj.baidubce.com
Content-Type: application/octet-stream
Content-Length: 102400

// real binary file content
```

压缩包示例 |—script |—dir1 |—dir2 |—aaa.json |—dir3 |—bbb.json |—ccc.json

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxx
Content-Type: application/json

{
  "result": [
    {
      "totalCount": 10,
      "successCount": 8,
      "failedCount": 2,
      "failedItems": [
        {
          "path": "dir/aaa.json",
          "cause": "wrong json format"
        },
        {
          "path": "bbb.json",
          "cause": "wrong category"
        }
      ]
    }
  ]
}

HTTP/1.1 404 Not Found
Content-Length: xxx
Content-type:json/application

{
  "code":"ProjectNotFound",
  "message":"Project pro is not found",
  "requestId":" 81d0b05f-5ad4-1f22-8068-d5c9de60a1d7"
}
```

ExportWorbenchItem 接口说明

根据categroy批量导出工作台草稿，导出为一个zip压缩文件。

原型

GET /v1/project/\${projectId}/workbench/\${category}?export

请求参数

名称	类型	位置	必须	描述
projectId	String	URL	是	项目空间名称
category	String	URL	是	工作台分类的枚举：pipeline、script、jobgroup、realtime、flink 和 sparkstreaming

响应参数

二进制的zip文件内容

错误码

Code	message	HTTP状态码	描述
AccessDenied	No permission to operate project	403	无权限
ProjectNotFound	Project is not found.	404	项目空间不存在

请求示例

```
GET /v1/project/pro/workbench/script?export
Host: edap.bj.baidubce.com
Content-Type: application/json
Content-Length: xxx
```

响应示例

```
HTTP/1.1 200 OK
Content-Type: application/octet-stream
Content-Disposition: attachment;filename=project_category_datetime.zip

// real binary file content

HTTP/1.1 404 Not Found
Content-Length: xxx
Content-type:json/application

{
  "code":"ProjectNotFound",
  "message":"Project pro is not found",
  "requestId":" 81d0b05f-5ad4-1f22-8068-d5c9de60a1d7"
}
```

🔗 执行管理

CreateExecution 接口说明 创建临时执行作业实例

原型 POST /v1/project/<projectName>/execution

请求参数 | 参数 | 类型 | 位置 | 必须 | 说明 | |-----|-----|-----|-----| | projectName | String | URL | 是 | 项目空间名称 | | content | Property | Body | 是 | 作业组执行配置 |

Jobgroup.Content对象描述

参数	类型	必须	描述
config	Map<String,Object>	是	作业组运行配置
connections	Lis<Connection>	是	作业组 DAG边连接信息，详见[Jobgroup.Content.Connection对象]
nodes	List<JobNode>	是	节点信息，详见[Jobgroup.Content.JobNode 对象信息]
runtimeArgs	Map<String,String>	否	运行时参数

Jobgroup.Content.Connection对象

描述DAG中的边（节点连接关系）

字段	类型	是否必须	描述
from	String	是	节点连接线的源端的标识
to	String	是	节点连接线的目的端的标识

Jobgroup.Content.JobNode对象

描述作业组节点配置信息

参数	类型	必须	描述
name	String	是	节点名称
type	String	是	节点类型：edap,shell,python,spark,jdbc,hsql,quality,sparkscala,sparksql,pyspark,flinksql,flinkjar,di
version	Integer	否	节点中引用的某个作业的版本信息
config	Map<String,Object>	是	作业节点运行配置，详见[Jobgroup.Content.JobNode.Property 对象]
profile	String	是	节点执行资源
controlFlags	ControlFlag	否	节点的执行控制参数，详见[Jobgroup.Content.JobNode.ControlFlag 对象]
pos	Position	是	节点在画布上的位置信息，详见[Jobgroup.Content.JobNode.Position 对象]
runtimeArgs	Map<String,String>	否	节点运行时宏参数

Jobgroup.Content.JobNode.Property 对象

作业类型	可配置参数	类型	必须	描述
edap	appsName	String	是	引用的流水线作业名称
shell	command	String	否	需要执行的 shell 脚本命令，当选择「shell 程序包」必填
	filePath	String	否	shell作业上传的包路径，当选择「shell 程序包」可选填
	scriptName	String	否	shell作业节点引用的脚本作业名称，当选择「shell 脚本」必填
	environment	Map<String,String>	否	作业节点的环境变量，用户界面采用key-value形式填写，替用户填写，该字段由用户在UI中填写

	envArgs	ng>	否	SN 脚本的环境变量，用户界面上用key-value形式填写。若用户未填写，该字段出现在 body中，用{}表示
python	cmdParams	String	否	python命令行参数，用户填写时，多个参数用','分开。当选择「python 程序包」选填
	filePath	String	否	python 作业上传的包路径，当选择「python 程序包」必填
	scriptName	String	否	python作业节点引用的脚本作业名称，当选择「python 脚本」必填
	envArgs	Map<String,String>	否	py 脚本的环境变量，用户界面上用key-value形式填写。若用户未填写，该字段出现在 body中，用{}表示
hsql	connectionId	String	是	用户勾选的源连接名称，当选择「在线编辑 SQL」为必选项
	statement	String	是	用户需要执行的 sql 语句，当选择「在线编辑 SQL」为必填项
	scriptName	String	是	作业节点引用的脚本作业名称，当选择「SQL脚本」为必填项
	hiveConf	Map<String,String>	否	hive引擎执行参数，用户界面上用key-value形式填写。若用户未填写，该字段出现在 body中，用{}表示
jdbc	connectionId	String	是	用户勾选的源连接名称，当选择「在线编辑 SQL」为必选项
	jdbcPluginName	String	是	源连接驱动类型，当选择「在线编辑 SQL」为必选项
	statement	String	是	用户需要执行的 sql 语句，当选择「在线编辑 SQL」为必填项
	scriptName	String	是	作业节点引用的脚本作业名称，当选择「SQL脚本」为必填项
spark	mainName	String	是	spark 主类所在的 jar 名称
	filePath	String	是	spark执行包路径，支持.jar 和.zip; 若为.zip 文件，这需要指定主类所在的 jar，即参数 mainName
	className	String	是	主类名称
	appParams	Map<String,String>	否	spark main 函数参数，用户填写时，多个参数用','分开。
	submitParams	Map<String,String>	否	spark 提交参数，用户界面上用key-value形式填写。若用户未填写，该字段出现在 body中，用{}表示; 详见 [spark引擎参数默认值]
	sparkConf	Map<String,String>	否	spark 运行时配置参数，用户界面上用key-value形式填写。若用户未填写，该字段出现在 body中，用{}表示
sparkscala	scriptName	String	是	作业节点引用的脚本作业名称
	sparkConf	Map<String,String>	否	spark 运行时配置参数，用户界面上用key-value形式填写。若用户未填写，该字段出现在 body中，用{}表示
pyspark	scriptName	String	是	作业节点引用的脚本作业名称
	sparkConf	Map<String,String>	否	spark 运行时配置参数，用户界面上用key-value形式填写。若用户未填写，该字段出现在 body中，用{}表示
di	jobId	String	是	数据集成Job的 Id
	sinkName	String	是	Job下的某一目标表名称
	watermark	String	是	水位线信息，包括用户填入的时间字符串和logicalStartTime宏函数： 时间字符串的格式为 'yyyy-MM-ddTHH:mm:ssZ'，如2022-02-12T14:00:32Z 宏函数的格式为\${logicalStartTime(yyyy-MM-dd hh:mm:ss, offset)}，如\${logicalStartTime(yyyy-MM-dd hh:mm:ss, -1d)}
quality	topic	String	是	质量任务所属的主题名称
	database	String	是	质量任务所属的数据库名称
	table	String	是	质量任务验证的表名称
	qualityTaskId	String	是	质量作业id
jobgroup	sourceName	String	是	节点引用的作业组名称
	jobName	String	否	作业组的唯一名称 ID
datamask	taskId	String	是	脱敏任务id

Jobgroup.Content.JobNode.ControlFlag 执行控制对象

参数	类型	必须	描述
timeoutInSeconds	Integer	是	超时时间，单位 s。默认为0(不控制超时)
retriesFlag	Boolean	是	重试标记，默认为 false
retries	Integer	否	重试次数，默认为0
retryBackoffInMilliseconds	Long	否	重试间隔，默认为0，单位 ms
maxConcurrentNum	Integer	否	作业执行最大并行度，默认为1
priority	Integer	否	作业执行优先级参数，默认为5
condition	List<CondExpr>	依赖结点参数表达式列表，默认为空，详见 [Jobgroup.Content.JobNode.ControlFlag.CondExpr 对象]	

Jobgroup.Content.JobNode.ControlFlag.CondExpr 节点表达式对象

参数	类型	必须	描述
jobName	String	是	节点名称
parameter	String	依赖参数，目前支持 STATUS	
expression	String	是	表达式，支持 ==all_success, ==success, ==failed, ==all_done

Jobgroup.Content.JobNode.Position 对象

参数	类型	必须	描述
left	Integer	是	距离屏幕左端的像素点
top	Integer	是	距离屏幕上端的像素点

- 其中，作业组中作业节点执行控制参数(带星参数对于该执行接口无效)

参数说明	参数	类型	必须	描述
作业节点执行控制参数	condition	List<ConExpr>	否	依赖节点参数表达式列表，组成一个 ConExp 对象的列表传入。若 condition 为空，则依赖节点设置开关为关闭状态
	timeoutInSeconds*	Integer	是	最长执行时间（秒）（前端需要将时间单位做一下统一，统一为秒），默认3600
	retriesFlag*	Boolean	是	是否失败重试，默认为false(是)
	retries*	Integer	是	最大失败重试次数，失败重试为 true 时生效，默认为1
	retryBackoffInMilliseconds*	Integer	是	重试时间间隔（毫秒），失败重试为 true 时生效，默认为100
	maxConcurrentNum*	Integer	是	最大并行作业数，默认为1
	priority*	Integer	是	作业优先级参数，默认为5

响应参数 | 名称 | 类型 | 描述 | |----|----|----| |execId|String|执行实例ID,带有Z前缀的为临时执行的执行ID|

错误码

Code	message	HTTP状态码	描述
AccessDenied	No permission to operate project	403	无权限
ProjectNotFound	Project is not found.	404	项目空间不存在

请求示例

```
POST /v1/project/project1/execution
Host: edap.bj.baidubce.com
Content-Length: 0
Content-Type: application/json

{
  "content":{
    "runtimeArgs":{

    },
    "connections":[
```

```
{
  "from": "Shell",
  "to": "JDBCSQL"
},
"config": {
  "profiles": [
    "aaa"
  ]
},
"nodes": [
  {
    "name": "HIVESQL",
    "type": "hsq",
    "profile": "aaa",
    "runtimeArgs": {
      "condition": [
        "timeoutInSeconds": 600,
        "maxConcurrentNum": 1,
        "retriesFlag": false,
        "retries": 1,
        "retryBackoffInMilliseconds": 100,
        "priority": 5,
        "dependsOn": ""
      ]
    },
    "config": {
      "hiveConf": {
        "connectionId": "123",
        "statement": "show tables;",
        "scriptName": ""
      }
    },
    {
      "name": "JDBCSQL",
      "type": "jdbc",
      "profile": "aaa",
      "runtimeArgs": {
        "condition": [
          "timeoutInSeconds": 600,
          "maxConcurrentNum": 1,
          "retriesFlag": false,
          "retries": 1,
          "retryBackoffInMilliseconds": 100,
          "priority": 5,
          "dependsOn": ""
        ]
      },
      "config": {
        "jdbcPluginName": "mysql",
        "connectionId": "aaa",
        "statement": "show tables;",
        "scriptName": ""
      }
    },
    {
      "name": "Shell",
      "type": "shell",
      "profile": "aaa",
      "runtimeArgs": {
        "condition": [
          "timeoutInSeconds": 600,
          "maxConcurrentNum": 1
```

```
    "maxConcurrentRuns": 1,
    "retriesFlag": false,
    "retries": 1,
    "retryBackoffInMilliseconds": 100,
    "priority": 5,
    "dependsOn": ""
  },
  "config": {
    "command": "ls /",
    "filePath": "bos://xxx",
    "envArgs": {
      // ...
    }
  }
}
```

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxxx
Content-Type: application/json

{
  "execId": "Z123"
}
```

RerunExecution 接口说明 重跑作业执行实例

原型 POST /v1/project/<projectName>/execution/<execId>?rerun

请求参数 | 参数 | 类型 | 位置 | 必须 | 说明 | |-----|-----|-----|-----| | projectName | String | URL | 是 | 项目空间名称 | | execId | String | URL | 是 | 执行实例Id | | setSkip | List<String> | Body | 否 | 重跑跳过的节点列表, <节点名称>, 节点跳过后执行状态为 SKIPPED, 作业组类型实例有效 | | setSucc | List<String> | Body | 否 | 重跑置为成功的节点列表, <节点名称>, 作业组类型实例有效 | | startNodes | List<String> | Body | 否 | 重跑起始节点列表, <节点名称>, 从起始节点开始调度作业组DAG子图, 作业组类型实例有效 | | rerunAllNodes | Boolean | Body | 否 | 是否重跑所有节点, true表示重跑所有节点, false表示仅重跑失败节点, 默认为true, , 作业组类型实例有效 | | runNodes | List<String> | Body | 否 | 需要重跑的节点列表, 其他节点状态与上次执行保持一致, 作业组类型实例有效 |

响应参数 | 名称 | 类型 | 描述 | |-----|-----|-----| | execId | String | 作业执行实例 ID | | rerunId | Integer | 作业重跑批次ID |

错误码

Code	message	HTTP状态码	描述
AccessDenied	No permission to operate project	403	无权限
ProjectNotFound	Project is not found.	404	项目空间不存在
ExecutionNotFound	Execution is not found.	404	实例不存在
ExecutionIsRunning	Execution {0} is running.	409	实例(0)正在运行中

请求示例

```
POST /v1/project/project1/execution/Z123?rerun
Host: edap.bj.baidubce.com
Content-Length: 0
Content-Type: application/json

{
  "rerunNodes":["nodeA","nodeB","NodeC"]
}
```

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxxx
Content-Type: application/json

{
  "rerunId": 1
}
```

GetExecutionStatus 接口说明 获取作业组执行实例状态信息

原型 GET /v1/project/<projectName>/execution/<execId>?status

请求参数 | 参数 | 类型 | 位置 | 必须 | 说明 | |-----|-----|-----|-----| | projectName | String | URL | 是 | 项目空间名称 | | execId | String | URL | 是 | 执行实例Id |

响应参数

当作业实例来自于 jobgroup 类型执行时

名称	类型	描述
startTime	DateTime	开始时间，格式: yyyy-MM-ddTHH:mm:ssZ
endTime	String	结束时间，格式: yyyy-MM-ddTHH:mm:ssZ
status	String	此次运行状态枚举: STARTING, RUNNING, SUCCEED, FAILED, KILLED
nodes	List<NodeRunStatus>	作业组内部各节点任务状态，详见[NodeRunStatus 对象]

NodeRunStatus 对象

名称	类型	描述
name	String	节点名称
type	String	节点类型
startTime	DateTime	开始时间，格式: yyyy-MM-ddTHH:mm:ssZ
endTime	String	结束时间，格式: yyyy-MM-ddTHH:mm:ssZ
status	String	此次运行状态枚举: READY, QUEUED, PREPARING, RUNNING, SUCCEED, FAILED, KILLED
profile	String	执行资源
in	List<String>	上游节点名称集合

错误码 | Code | message | HTTP状态码 | 描述 | | :---: | :---: | :---: | :---: | | AccessDenied | No permission to operate project | 403 | 无权限 | | ProjectNotFound | Project is not found. | 404 | 项目空间不存在 | | ExecutionNotFound | Executionis not found. | 404 | 实例不存在 |

请求示例

```
GET /v1/project/project1/execution/Z123?status
Host : edap.bj.baidubce.com
Content - Length : 0
Content - Type : application/json
```

响应示例

```
HTTP/ 1.1 200 OK
Content-Length:xxx
Content-Type:application/json

{
  "status": "FAILED",
  "startTime": "2020-02-11T09:00:21Z",
  "endTime": "2020-02-11T09:24:57Z",
  "nodes": [
    {
      "startTime": "2020-02-11T09:24:57Z",
      "endTime": "2020-02-11T09:24:57Z",
      "name": "test",
      "status": "CANCELLED",
      "type": "command",
      "profile": "bmr1",
      "in": [
        "test-foo"
      ]
    },
    {
      "startTime": "2020-02-11T09:24:57Z",
      "endTime": "2020-02-11T09:24:57Z",
      "name": "test-bar",
      "status": "SUCCEEDED",
      "type": "edap",
      "profile": "bmr1"
    },
    {
      "startTime": "2020-02-11T09:24:57Z",
      "endTime": "2020-02-11T09:24:57Z",
      "name": "test-foobar",
      "status": "FAILED",
      "type": "edap",
      "profile": "bmr1",
      "appUrl": "https://bmr.bce.baidu.com/e8fde22f-1504-4993-7125-f69c14d20809/8088/cluster/app/application_1643576560043_10161",
      "in": [
        "test-bar"
      ]
    },
    {
      "startTime": "2020-02-11T09:24:57Z",
      "endTime": "2020-02-11T09:24:57Z",
      "name": "test-foo",
      "status": "CANCELLED",
      "type": "spark",
      "profile": "bmr1",
      "appUrl": "https://bmr.bce.baidu.com/e8fde22f-1504-4993-7125-f69c14d20809/8088/cluster/app/application_1643576560043_10161",
      "in": [
        "test-foobar"
      ]
    }
  ]
}
```

GetExecutionLog 接口说明 获取执行作业组实例日志。

原型 GET /v1/project/<projectName>/execution/<execId>?log

请求参数 | 参数 | 类型 | 位置 | 必须 | 说明 | |-----|-----|-----|-----| | projectName | String | URL | 是 | 项目空间名称 | | execId | String | URL | 是 | 执行实例ID | | marker | String | QueryString | 否 | 分页标识 | | limit | Integer | QueryString | 否 | 获取的最大行数 |

响应参数 | 名称 | 类型 | 描述 | |-----|-----|-----| | marker | String | 分页标识 | | nextMarker | String | 获取下一页所需marker值 | | isTruncated | Bool | 是否还有下一页数据, true为有 | | log | List<String> | log内容 |

错误码 | Code | message | HTTP状态码 | 描述 | | :---: | :---: | :---: | :---: | | AccessDenied | No permission to operate project | 403 | 无权限 |

请求示例

```
GET /v1/project/project1/execution/Z123?log&marker=0&limit=1000
Host: edap.bj.baidubce.com
Content-Length: 0
Content-Type: application/json
```

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxx
Content-Type: application/json

{
  "marker": "0",
  "nextMaker": "1576753441596",
  "isTruncated": false,
  "log": [
    "2019-12-19 14:41:25,870 - DEBUG [provisioning-service-4:i.c.c.i.p.t.ProvisioningTask@121] - Executing PROVISION subtask REQUESTING_CREATE for program run program_run:default.file2file_v1.-SNAPSHOT.workflow.DataPipelineWorkflow.4bfd3c14-f629-11e9-a615-000000735566.\n",
    "2019-12-19 14:41:25,871 - DEBUG [provisioning-service-4:i.c.c.i.p.t.ProvisioningTask@121] - Executing PROVISION subtask REQUESTING_CREATE for program run program_run:default.file2file_v1.-SNAPSHOT.workflow.DataPipelineWorkflow.4bfd3c14-f629-11e9-a615-000000735568.\n"
  ]
}
```

GetJobExecutionLog 接口说明 获取执行实例的作业组中的作业节点日志或者某个作业实例的执行日志。

原型 GET /v1/project/<projectName>/execution/<execId>/job/<jobName>?log

请求参数 | 参数 | 类型 | 位置 | 必须 | 说明 | |-----|-----|-----|-----| | projectName | String | URL | 是 | 项目空间名称 | | execId | String | URL | 是 | 执行实例Id | | jobName | String | URL | 是 | 作业名称 | | marker | String | QueryString | 否 | 分页标识 | | limit | Integer | QueryString | 否 | 返回最大行数，默认100条 | | filter | String | QueryString | 否 | 过滤条件，如 {"INFO", "WARN", "ERROR"}，默认INFO 注：此参数目前只有pipeline和 realtime类型支持 |

响应参数

名称	类型	描述
marker	String	分页标识
nextMarker	String	获取下一页所需marker值
isTruncated	Bool	是否还有下一页数据, true为有
log	List<String>	log内容

错误码 | Code | message | HTTP状态码 | 描述 | | :---: | :---: | :---: | :---: | | AccessDenied | No permission to operate project | 403 | 无权限 |

请求示例

```
GET /v1/project/project1/execution/Z_123/job/FileTest01?log&marker=0&limit=1000
Host: edap.bj.baidubce.com
Content-Length: 0
Content-Type: application/json
```

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxx
Content-Type: application/json

{
  "marker": "0",
  "nextMaker": "1576753441596",
  "isTruncated": false,
  "log": [
    "2019-12-19 14:41:25,870 - DEBUG [provisioning-service-4:i.c.c.i.p.t.ProvisioningTask@121] - Executing PROVISION subtask REQUESTING_CREATE for program run program_run:default.file2file_v1.-SNAPSHOT.workflow.DataPipelineWorkflow.4bfd3c14-f629-11e9-a615-000000735566.\n",
    "2019-12-19 14:41:25,871 - DEBUG [provisioning-service-4:i.c.c.i.p.t.ProvisioningTask@121] - Executing PROVISION subtask REQUESTING_CREATE for program run program_run:default.file2file_v1.-SNAPSHOT.workflow.DataPipelineWorkflow.4bfd3c14-f629-11e9-a615-000000735568.\n"
  ]
}
```

GetExecutionOutput 接口描述 获取脚本作业执行实例的输出结果

原型 GET /v1/project/<projectName>/execution/<execId>/job/<jobName>?output

请求参数 | 参数 | 类型 | 位置 | 必须 | 说明 | |-----|-----|-----|-----| | projectName | String | URL | 是 | 项目空间名称 | | execId | String | URL | 是 | 执行ID | | jobName | String | URL | 是 | 作业名称 |

响应参数 | 名称 | 类型 | 描述 | |----|----|----| | columns | List<String> | 查询结果列名 | | rows | List<List<Object>> | 数据结果 |

错误码 | Code | message | HTTP状态码 | 描述 | | :---: | :---: | :---: | :---: | | AccessDenied | No permission to operate project | 403 | 无权限 |

请求示例

```
GET /v1/project/project1/execution/Z_123/job/FileTest01?output
Host: edap.bj.baidubce.com
Content-Length: xxxx
Content-Type: application/json
```

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxxx
Content-Type: application/json
{
  "columns": [
    "col1",
    "col2"
  ],
  "rows": [
    [
      "hello",
      "world"
    ],
    [
      "next",
      "time"
    ]
  ]
}
```

任务运维

作业组管理

DeployJobgroup 接口说明 上线已发布的最高版本作业组

原型 POST /v1/project/<projectName>/jobgroup/<jobgroupName>?deploy

请求参数 名称|类型|位置|必须|描述| |-----|-----|---|-----| |projectName|String|URL|是|项目空间名称| |jobgroupName|String|URL|是|作业组名称|

响应参数 除公共响应参数外，无特殊响应。 错误码 | Code | message | HTTP状态码 | 描述 | | :---: | :---: | :---: | :---: | | AccessDenied | No permission | 403 | 无权限 | | ProjectNotFound | Project is not found. | 404 | 项目空间不存在 | | JobNotFound | Job is not found | 404 | 作业不存在 |

请求示例 POST /v1/project/project1/jobgroup/my-jobgroup?deploy

响应示例

```
HTTP/1.1 200 OK
Content-Length: 0
Content-Type: application/json
```

ListJobgroups 接口说明 获取运维中心作业组列表

原型 GET /v1/project/<projectName>/jobgroup?

请求参数 名称|类型|位置|必须|描述| |---|---|---|---| |projectName|String|URL|是|项目空间名称| |owner|String|QueryString|否|责任人搜索项| |deployer|String|QueryString|否|上线人搜索项| |publisher|String|QueryString|否|发布人搜索项| |name|String|QueryString|否|作业组名称(唯一id串)搜索项| |sourceName|String|QueryString|否|作业组展示名称搜索项| |status|String|QueryString|否|任务状态过滤项(ALL, PUBLISHED(已发布), DEPLOYED(已上线), DEPLOYED_NEWVERSION(已上线有新版本)), 默认 ALL| |pageNo|Integer|QueryString|是|页码, 默认:1, 若为-1, 则不分页| |pageSize|Integer|QueryString|是|当前页最大显示数量, 默认:10, 若为-1, 则不分页。| |orderBy|String|QueryString|否|排序字段| |order|String|QueryString|否|排序方式, 枚举值{asc, desc}, 默认desc|

响应参数 名称|类型|描述| |---|---|---| |jobs|List<JobObj>|作业组描述对象列表| |pageNo|Integer|页码| |pageSize|Integer|当前页最大显示数量| |totalCount|Integer|总数|

JobObj 对象定义

名称	类型	描述
name	String	作业组名称(唯一 id 串)
sourceName	String	作业组展示名称
status	String	作业组状态
newestVersion	Integer	最新版本
onlineVersion	Integer	上线版本
nodeCount	Integer	作业组中节点数量
owner	String	责任人
deployer	String	上线人
publisher	String	发布人
publishTime	DateTime	发布时间，格式: yyyy-MM-ddTHH:mm:ssZ
deployTime	DateTime	上线时间，格式: yyyy-MM-ddTHH:mm:ssZ
isAlarmed	Boolean	是否配置告警
isScheduled	Boolean	是否配置调度
isRunning	Boolean	是否存在正在运行的实例

错误码 | Code | message | HTTP状态码 | 描述 | | :---: | :---: | :---: | :---: | | AccessDenied | No permission | 403 | 无权限 |

请求示例 GET /v1/project/project1/jobgroup?status=DEPLOYED&pageNo=1&pageSize=10

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxxx
Content-Type: application/json

{
  "jobs":[
    {
      "name":"testJobgroup-00",
      "sourceName":"testJobgroup",
      "status":"DEPLOYED",
      "nodeCount":10,
      "publishTime":"2020-01-01T00:00:00Z",
      "deployTime":"2020-01-01T00:00:00Z",
      "owner":"project1",
      "publisher":"user1",
      "deployer":"user1",
      "isRunning":false,
      "isAlarmed":false,
      "isScheduled":false,
      "newestVersion":1,
      "onlineVersion":1
    }
  ],
  "totalCount":1,
  "pageNo":1,
  "pageSize":10
}
```

DeleteJobgroup 接口说明 删除运维中心作业组

原型 DELETE /v1/project/<projectName>/jobgroup/<jobgroupName>

请求参数 |名称|类型|位置|必须|描述| |---|---|---|---| |projectName |String|URL|是|项目空间名称| |jobgroupName|String|URL|是|作业组名称|

响应参数 除公共响应参数外，无特殊响应。

错误码 | Code | message | HTTP状态码 | 描述 | | --- | --- | --- | --- | | AccessDenied | No permission | 403 | 无权限 | | ProjectNotFound | Project is not found. | 404 | 项目空间不存在 |

请求示例 DELETE /v1/project/project1/jobgroup/my-jobgroup

响应示例

```
HTTP/1.1 200 OK
Content-Length: 0

HTTP/1.1 403 NOTPERMISSION
Content-Length: xxx
Content-Type: application/json

{
  "code": "AccessDenied",
  "message": "No Permission to access project p1",
  "requestId": " 81d0b05f-5ad4-1f22-8068-d5c9de60a1d7"
}
```

KillJobgroupExecution 接口说明 停止正在执行的作业组实例

原型 POST /v1/project/<projectName>/jobgroup/<jobgroupName>/execution/<execId>?kill

请求参数 |名称|类型|位置|必须|描述| |---|---|---|---| |projectName |String|URL|是|项目空间名称| |jobgroupName|String|URL|是|作业组名称| |execId|Integer|URL|是|实例 ID|

响应参数 除公共参数外，无特殊响应参数

错误码 | Code | message | HTTP状态码 | 描述 | | :---: | :---: | :---: | :---: | | AccessDenied | No permission | 403 | 无权限 | | ProjectNotFound | Project is not found. | 404 | 项目空间不存在 | | JobNotFound | Job{0} is not found | 404 | 作业{0}不存在 | | ExecutionNotFound | Execution {0} is not found. | 404 | 实例 {0} 不存在。 |

请求示例 POST /v1/project/project1/jobgroup/my-jobgroup-00/execution/302?kill

响应示例

```
HTTP/1.1 200 OK
Content-Length: 0
Content-Type: application/json
```

ListJobgroupExecutions 接口说明 获取运维中心作业组执行实例列表

原型 GET /v1/project/<projectName>/jobgroup?execution

请求参数

名称	类型	位置	必须	描述
projectName	String	URL	是	项目空间名称
owner	String	QueryString	否	责任人搜索项
name	String	QueryString	否	作业组名称(唯一id串)搜索项
sourceName	String	QueryString	否	作业组展示名称搜索项
execId	String	QueryString	否	作业组示例 ID 过滤搜索，默认为空，显示全部
status	String	QueryString	否	任务状态过滤项 包括 ALL, RUNNING、STARTING、FAILED、SUCCEED和 KILLED，PENDING, QUEUED, READY,NOTFIRE 状态的查询 默认 ALL 多状态查询，各状态字段用,分开，如 STARTING,FAILED
startTime	DateTime	QueryString	是	起始时间过滤项 （默认为 空）format: 'yyyy-MM-ddTHH:mm:ssZ'
endTime	DateTime	QueryString	是	结束时间过滤项 （默认为 空）format: 'yyyy-MM-ddTHH:mm:ssZ'
fromLogicalStartTime	DateTime	QueryString	是	起始业务时间过滤项 (默认为空)format: 'yyyy-MM-ddTHH:mm:ssZ'
toLogicalStartTime	DateTime	QueryString	是	截止业务时间过滤项 (默认为空) format: 'yyyy-MM-ddTHH:mm:ssZ'
executionType	String	QueryString	是	指定实例类型过滤，包括 MANUAL(手动触发)/PERIOD(周期调度)，默认为空
executionCause	String	QueryString	是	指定实例调度来源类型过滤, 包括 MANUAL(手动执行)、TRIGGER(周期触发)、RERUN(重跑)、FIX(修数据)，默认为空(全部)
pageNo	Integer	QueryString	是	页码，默认:1，若为-1，则不分页
pageSize	Integer	QueryString	是	当前页最大显示数量，默认:10，若为-1，则不分页。
orderBy	String	QueryString	否	排序字段
order	String	QueryString	否	排序方式，枚举值{asc, desc}，默认desc

响应参数

名称	类型	描述
executions	List<ExecutionInfo>	作业组实例记录列表
pageNo	Integer	页码
pageSize	Integer	当前页最大显示数量
totalCount	Integer	记录总数

ExecutionInfo 对象定义

名称	类型	描述
execId	String	执行实例 ID
name	String	作业组名称(唯一 id 串)
sourceName	String	作业组展示名称
status	String	作业组实例执行状态
executionCause	String	指定实例来源
executionType	String	指定实例类型
startTime	DateTime	开始时间，格式: yyyy-MM-ddTHH:mm:ssZ
endTime	DateTime	结束时间，格式: yyyy-MM-ddTHH:mm:ssZ
updateTime	DateTime	更新时间，格式: yyyy-MM-ddTHH:mm:ssZ
logicalStartTime	DateTime	基准时间，格式: yyyy-MM-ddTHH:mm:ssZ
runtime	String	运行时间，format 'xxHxxmxxs'
rerun	Integer	重跑次数, 默认为0
owner	String	责任人

错误码

Code	message	HTTP状态码	描述
AccessDenied	No permission	403	无权限
ProjectNotFound	Project is not found.	404	项目空间不存在

请求示例

GET /v1/project/project1/jobgroup?execution&sourceName=jg1&status=ALL&pageNo=1&pageSize=10

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxx
Content-Type: application/json

{
  "executions":[
    {
      "startTime":"2020-02-13T01:23:13Z",
      "owner":"user1",
      "status":"RUNNING",
      "execId":1,
      "name":"jg1",
      "sourceName":"jg1",
      "nodeCount":10,
      "endTime":"",
      "logicalStartTime":"2020-02-13T01:23:00Z",
      "runtime":"1h2m5s",
      "rerun":0
    },
    {
      "startTime":"2020-02-13T01:23:13Z",
      "owner":"user1",
      "status":"RUNNING",
      "execId":1,
      "name":"jg1",
      "sourceName":"jg1",
      "nodeCount":10,
      "endTime":"2020-02-13T01:26:15Z",
      "logicalStartTime":"2020-02-13T01:23:00Z",
      "runtime":"3m2s",
      "rerun":0
    }
  ],
  "totalCount":2,
  "pageNo":1,
  "pageSize":10
}
```

ListJobExecutions 接口说明 获取运维中心作业执行实例列表

原型 GET /v1/project/<projectName>/job?execution

请求参数

名称	类型	位置	必须	描述
projectName	String	URL	是	项目空间名称
owner	String	QueryString	否	责任人搜索项
name	String	QueryString	否	作业名称(唯一id串)搜索项
sourceName	String	QueryString	否	作业展示名称搜索项
execId	String	QueryString	否	作业示例 ID 过滤搜索，默认为空，显示全部
type	String	QueryString	否	作业类型搜索项, 枚举值 { ALL(全部) realtime (实时流水线) flinksql(flinksql 作业) flinkjar(flink 自定义作业) streamingjar(sparkstreaming jar作业) streamingscala(sparkstreaming scala作业) pipeline(可视化作业) shell(shell 作业) python(python 作业) sparksql sparkscala pyspark spark jdbc hsqldb} 离线作业 全部 {pipeline, shell, python, spark, sparksql, pyspark, jdbc, hsqldb,pipeline,di,quality} 实时作业全部 {realtime, flinksql, flinkjar}
status	String	QueryString	否	任务状态过滤项 包括 ALL, RUNNING、STARTING、FAILED、SUCCEED和 KILLED , PENDING, QUEUED, READY,NOTFIRE 状态的查询 默认 ALL 多状态查询，各状态字段用,分开，如 STARTING,FAILED
startTime	DateTime	QueryString	是	起始时间过滤项 （默认为 空）format: 'yyyy-MM-ddTHH:mm:ssZ'
endTime	DateTime	QueryString	是	结束时间过滤项 （默认为 空）format: 'yyyy-MM-ddTHH:mm:ssZ'
fromLogicalStartTime	DateTime	QueryString	是	起始业务时间过滤项 (默认为空)format: 'yyyy-MM-ddTHH:mm:ssZ'
toLogicalStartTime	DateTime	QueryString	是	截止业务时间过滤项 (默认为空) format: 'yyyy-MM-ddTHH:mm:ssZ'
executionType	String	QueryString	是	指定实例类型过滤，包括 MANUAL(手动触发)/PERIOD(周期调度)，默认为空
executionCause	String	QueryString	是	指定实例调度来源类型过滤, 包括 MANUAL(手动执行)、TRIGGER(周期触发)、RERUN(重跑)、FIX(修数据)，默认为空(全部)
pageNo	Integer	QueryString	是	页码，默认:1，若为-1，则不分页
pageSize	Integer	QueryString	是	当前页最大显示数量，默认:10，若为-1，则不分页。
orderBy	String	QueryString	否	排序字段
order	String	QueryString	否	排序方式，枚举值{asc, desc}，默认desc

响应参数

名称	类型	描述
executions	List<ExecutionInfo>	作业组实例记录列表
pageNo	Integer	页码
pageSize	Integer	当前页最大显示数量
totalCount	Integer	记录总数

ExecutionInfo 对象定义

名称	类型	描述
execId	String	执行实例 ID
name	String	作业名称(唯一 id 串)
sourceName	String	作业展示名称
type	String	作业类型
status	String	作业组实例执行状态
executionCause	String	指定实例来源
executionType	String	指定实例类型
startTime	DateTime	开始时间，格式: yyyy-MM-ddTHH:mm:ssZ
endTime	DateTime	结束时间，格式: yyyy-MM-ddTHH:mm:ssZ
updateTime	DateTime	更新时间，格式: yyyy-MM-ddTHH:mm:ssZ
logicalStartTime	DateTime	基准时间，格式: yyyy-MM-ddTHH:mm:ssZ
runtime	String	运行时间，format 'xxHxxmxxs'
rerun	Integer	重跑次数, 默认为0
owner	String	责任人
runId	String	运行ID

错误码

Code	message	HTTP状态码	描述
AccessDenied	No permission	403	无权限
ProjectNotFound	Project is not found.	404	项目空间不存在

请求示例

GET /v1/project/proj/job?execution&type=flinksq,flinkjar,realtime&pageNo=1&pageSize=10

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxxx
Content-Type: application/json

{
  "pageNo": 1,
  "pageSize": 10,
  "totalCount": 2,
  "executions": [{
    "runId": "wqqxxx",
    "name": "test",
    "type": "realtime",
    "status": "RUNNING",
    "startTime": "2021-02-19T21:30:08Z",
    "endTime": "",
    "runtime": "1H20m33s",
    "owner": "user1"
  }, {
    "runId": "sqgxxx",
    "name": "flinksql",
    "status": "RUNNING",
    "startTime": "2021-02-19T21:30:08Z",
    "endTime": "",
    "runtime": "24H00m01s",
    "owner": "user2"
  }
]
```

元数据API

统计信息

🔗 统计信息管理

UpdateStatistics

接口说明

更新表字段/分区统计信息

请求原型

POST /v1/database/\${databaseName}/table/\${tableName}?setStatistics

请求参数

名称	类型	位置	是否必填	描述
databaseName	String	URL	是	数据库名
tableName	String	URL	是	表名
numFiles	Long	Body	否	文件个数
numRows	Long	Body	否	行数
rawDataSize	Long	Body	否	原始数据大小
totalSize	Long	Body	否	存储数据大小
statisticsDesc	StatisticsDesc	Body	是	统计信息描述信息
columnStatisticsData	List<ColumnStatisticsData>	Body	否	统计信息

StatisticsDesc

名称	类型	是否必填	描述
tableLevel	boolean	是	是否是表级别的统计信息
partitionName	String	否	分区名
lastAnalyzedTime	Long	是	上次分析时间

ColumnStatisticsData

名称	类型	是否必填	描述
columnName	String	是	列名
columnType	String	是	列类型
statisticsData	statisticsData	是	统计信息

StatisticsData

名称	类型	是否必填	描述
booleanStats	BooleanStats	否	布尔型统计信息
longStats	LongStats	否	long型统计信息
doubleStats	DoubleStats	否	double类型统计信息
stringStats	StringStats	否	string类型统计信息
binaryStats	BinaryStats	否	binary类型统计信息
decimalStats	DecimalStats	否	decimal类型统计信息
dateStats	DateStats	否	date类型统计信息

BooleanStats

名称	类型	是否必填	描述
numTrues	long	是	真值数量
numFalses	long	是	假值数量
numNulls	long	是	空值数量

LongStats

名称	类型	是否必填	描述
longLowValue	long	是	最小值
longHighValue	long	是	最大值
numNulls	long	是	空值数量

DoubleStats

名称	类型	是否必填	描述
decimalHighValue	Decimal	是	最大值
decimalLowValue	Decimal	是	最小值
numNulls	long	是	空值数量

Decimal

名称	类型	是否必填
scale	short	是
unscaled	byte[]	是

StringStats

名称	类型	是否必填	描述
maxColLen	long	是	最大长度
avgColLen	double	是	最小长度
numNulls	long	是	空值数量

BinaryStats

名称	类型	是否必填	描述
maxColLen	long	是	最大长度
avgColLen	double	是	最小长度
numNulls	long	是	空值数量

DecimalStats

名称	类型	是否必填	描述
decimalLowValue	double	是	最小值
decimalHighValue	double	是	最大值
numNulls	double	是	空值数量

DateStats

名称	类型	是否必填	描述
dateLowValue	long	是	最小值
dateHighValue	long	是	最大值
numNulls	long	是	空值数量

响应参数

除公共响应参数外，无特殊响应。

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
PartitionNotFound	Partitionnot found.	404	分区不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/db1/table/tbl1?setStatistics HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: xxxx

Content-Type: application/json
{
  "numFiles": 12,
  "numRows": 10000,
  "rawDataSize": 76546659248,
  "totalSize": 8594679398,
  "statisticsDesc": {
    "tableLevel": true,
    "lastAnalyzedTime": 123
  },
  "columnStatisticsData": [
    {
      "columnName": "n_nationkey",
      "columnType": "long",
      "statisticsData": {
        "longStats": {
          "longLowValue": 1,
          "longHighValue": 3,
          "numNulls": 2
        }
      }
    },
    {
      "columnName": "n_name",
      "columnType": "boolean",
      "statisticsData": {
        "booleanStats": {
          "numNulls": 2,
          "numTrues": 0,
          "numFalses": 0
        }
      }
    }
  ]
}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

GetStatistics

接口说明

获取列/分区统计信息

请求原型

```
POST /v1/database/${databaseName}/table/${tableName}?getStatistics
```

请求参数

名称	类型	位置	是否必填	描述
databaseName	String	URL	是	数据库名
tableName	String	URL	是	表名
columns	List<String>	Body	否	需要获取统计信息的column列表
partitionNames	List<String>	Body	否	需要获取统计信息的partition列表

响应参数

名称	类型	描述
statistics	List<Statistics>	统计信息

Statistics

名称	类型	描述
partitionName	String	分区名
columnStatisticsData	List<ColumnStatisticsData>	统计信息

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
PartitionNotFound	Partitionnot found.	404	分区不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/db1/table/tbl1?getStatistics HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: xxxx
{
  "columns": ["n_nationkey", "n_name"]
}
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "statistics": [
    {
      "columnStatisticsData": [
        {
          "columnName": "n_nationkey",
          "columnType": "long",
          "statisticsData": {
            "longStats": {
              "longLowValue": 1,
              "longHighValue": 3,
              "numNulls": 2
            }
          }
        }
      ],
      {
        "columnName": "n_nation",
        "columnType": "long",
        "statisticsData": {
          "longStats": {
            "longLowValue": 1,
            "longHighValue": 3,
            "numNulls": 2
          }
        }
      }
    ]
  }
}
```

请求示例

```
{

  "columns": ["n_nationkey", "n_name"]

  "partitionNames":["n_nationkey=100"]

}
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "statistics": [
    {
      "partitionName": "n_nationkey=100",
      "columnStatisticsData": [
        {
          "columnName": "n_nationkey",
          "columnType": "long",
          "statisticsData": {
            "longStats": {
              "longLowValue": 1,
              "longHighValue": 3,
              "numNulls": 2
            }
          }
        },
        {
          "columnName": "n_nation",
          "columnType": "long",
          "statisticsData": {
            "longStats": {
              "longLowValue": 1,
              "longHighValue": 3,
              "numNulls": 2
            }
          }
        }
      ]
    }
  ]
}
```

DeleteStatistics

接口说明

删除列统计信息

请求原型

POST /v1/database/\${databaseName}/table/\${tableName}?deleteStatistics

请求参数

名称	类型	位置	是否必填	描述
databaseName	String	URL	是	数据库名
tableName	String	URL	是	表名
column	String	Body	否	需要删除统计信息的列名
partitionName	String	Body	否	需要删除统计信息的分区名

响应参数

除公共响应参数外，无特殊响应。

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
PartitionNotFound	Partitionnot found.	404	分区不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/db_01/table/tb_01?deleteStatistics HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: xxx

{

  "column": "col_01",

  "partitionName": "n_regionkey=100"

}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

分区

分区管理

PartitionInstance结构体

名称	类型	是否必须	描述
values	List<String>	是	分区值
location	String	否	分区位置绝对路径
createTime	String	否	创建时间
properties	Map<String, String>	否	自定义附加参数
authProperties	Map<String, String>	否	安全认证信息

ListPartitions

接口说明

获取表分区列表信息

请求原型

```
GET /v1/database/${dbName}/table/${tableName}/partition
```

请求参数

字段	类型	位置	是否必须	描述
dbName	String	URL	是	数据库名
tableName	String	URL	是	表名
pageNo	Integer	QueryString	是	页码
pageSize	Integer	QueryString	是	每页查询记录数
filter	String	QueryString	否	查询条件表达式，如 ds>20201201

响应参数

名称	类型	描述
partitions	List<PartitionInstance>	PartitionInstance实例结构体的列表

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database/db_01/table/tb_01/partition?pageNo=1&pageSize=10 HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "pageNo": 1,
  "pageSize": 10,
  "totalSize": 1,
  "partitions": [
    {
      "values": [
        "1",
        "hangzhou"
      ],
      "location": "bos://bucket/prefix",
      "locationId": "loc1",
      "properties": {
        "p1": "v1"
      },
      "createTime": "2019-06-01T23:00:10Z"
    }
  ]
}
```

ListPartitionNames

接口说明

获取表分区名称列表信息

请求原型

```
GET /v1/database/${dbName}/table/${tableName}/partition?names
```

请求参数

字段	类型	位置	是否必须	描述
dbName	String	URL	是	数据库名
tableName	String	URL	是	表名
filter	String	QueryString	否	查询条件表达式，如 ds>20201201

响应参数

名称	类型	描述
names	List<String>	分区名列表

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database/db_01/table/tb_01/partition?names HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "names": [
    "id=1/age=12",
    "id=2/age=11"
  ]
}
```

BatchGetPartitions

批量获取Table的分区信息

请求原型

```
POST /v1/database/${databaseName}/table/${tableName}/partition?batchGet
```

请求头域

除公共头域外，无其它特殊头域。

请求参数

名称	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
tableName	String	URL	是	表名
needAuthProperties	Boolean	QueryString	否	是否需要返回安全认证信息, 默认false
partitionValuesList	List<List<String>>	Body	是	分区值信息

响应参数

名称	类型	描述
partitions	List<PartitionInstance>	Partition实例结构体的列表，详建PartitionInstance结构体

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
PartitionNotFound	Partitionnot found.	404	分区不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/db_01/table/tb_01/partition?batchGet&needAuthProperties=true

Host: edap.bj.baidubce.com

Content-Length: xxxx

Content-Type: application/json
{
  "partitionValuesList": [
    {
      "partitionValues": [
        "1",
        "hangzhou"
      ]
    },
    {
      "partitionValues": [
        "2",
        "shanghai"
      ]
    }
  ]
}
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "partitions": [
    {
      "values": [
        "1",
        "hangzhou"
      ],
      "location": "bos://bucket/prefix",
      "properties": {
        "p1": "v1"
      },
      "createTime": "2019-06-01T23:00:10Z",
      "authProperties": {
        "accessKey": "key1",
        "secretKey": "key2",
        "sessionTokenKey": "key3"
      }
    },
    {
      "values": [
        "1",
        "shanghai"
      ],
      "location": "bos://bucket/shanghai",
      "properties": {
        "p1": "v1"
      },
      "createTime": "2019-06-01T23:00:11Z",
      "authProperties": {
        "accessKey": "key1",
        "secretKey": "key2",
        "sessionTokenKey": "key3"
      }
    }
  ]
}
```

BatchCreatePartitions

批量创建表格分区信息

请求原型

```
POST /v1/database/<database-name>/table/<table-name>/partition?batchCreate
```

请求头域

除公共头域外，无其它特殊头域。

请求参数

名称	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
tableName	String	URL	是	表名
partitions	List<PartitionInstance>	Body	是	分区信息，详情参见PartitionInstance结构体

响应参数

无

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
PartitionAlreadyExist	Partitionalready exist.	409	分区已存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/db_01/table/tb_01/partition?batchCreate

Host: edap.bj.baidubce.com

Content-Length: xxxx

Content-Type: application/json
{
  "partitions": [
    {
      "values": [
        "1",
        "hangzhou"
      ],
      "location": "bos://bucket/hangzhou",
      "properties": {
        "p1": "v1"
      }
    },
    {
      "values": [
        "2",
        "guiyang"
      ],
      "location": "bos://bucket/guiyang",
      "properties": {
        "p1": "v1"
      }
    }
  ]
}{
  "partitions": [
    {
      "values": [
        "1",
        "hangzhou"
      ],
      "location": "bos://bucket/hangzhou",
      "properties": {
        "p1": "v1"
      }
    },
    {
      "values": [
        "2",
        "guiyang"
      ],
      "location": "bos://bucket/guiyang",
      "properties": {
        "p1": "v1"
      }
    }
  ]
}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

BatchAlterPartitions

批量修改分区信息

请求原型

```
PUT /v1/database/<databaseName>/table/<tableName>/partition?batchAlter
```

请求头域

除公共头域外，无其它特殊头域。

请求参数

名称	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
tableName	String	URL	是	表名
partitions	List<PartitionInstance>	Body	是	分区信息，详情参见PartitionInstance结构体

响应参数

无

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
PartitionNotFound	Partitionnot found.	404	分区不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
PUT /v1/database/db_01/table/tb_01/partition?batchAlter

Host: edap.bj.baidubce.com

Content-Length: xxxx

Content-Type: application/json

{
  "partitions": [
    {
      "values": [
        "1",
        "hangzhou"
      ],
      "location": "bos://bucket/hangzhou",
      "properties": {
        "p1": "v1"
      }
    },
    {
      "values": [
        "2",
        "guiyang"
      ],
      "location": "bos://bucket/guiyang",
      "properties": {
        "p1": "v1"
      }
    }
  ]
}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

BatchDropPartitions

批量删除分区

请求原型

```
POST /v1/database/${databaseName}/table/${tableName}/partition?batchDrop
```

请求头域

除公共头域外，无其它特殊头域。

请求参数

名称	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
tableName	String	URL	是	表名
partitionValuesList	List<List<String>>	Body	是	分区值信息

响应参数

无

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
PartitionNotFound	Partitionnot found.	404	分区不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/db_01/table/tb_01/partition?batchDrop

Host: edap.bj.baidubce.com

Content-Length: xxxx

Content-Type: application/json
{
  "partitionValuesList": [
    {
      "partitionValues": [
        "1",
        "hangzhou"
      ]
    },
    {
      "partitionValues": [
        "2",
        "shanghai"
      ]
    }
  ]
}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

RenamePartition

重命名表分区。

请求原型

```
POST /v1/database/<database-name>/table/<table-name>/partition?rename
```

请求头域

除公共头域外，无其它特殊头域。

请求参数

名称	类型	参数位置	是否必填	描述
database	String	URL	是	数据库名
tableName	String	URL	是	表名
partitionValues	List<String>	Body	是	旧的分区值
newPartition	PartitionInstance	Body	是	新分区

响应参数

无

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
PartitionNotFound	Partitionnot found.	404	分区不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/db_01/table/tb_01/partition?rename

Host: edap.bj.baidubce.com

Content-Length: xxxx

Content-Type: application/json
{
  "database": "db_01",
  "table": "tb_01",
  "partitionValues": [
    "2022-04-18"
  ],
  "newPartition": {
    "partitionValues": [
      "2022-04-19"
    ],
    "location": "bos://bucket/prefix",
    "properties": {
      "p1": "v1"
    }
  }
}
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
```

函数

函数管理

函数信息 | 名称 | 类型 | 是否必须 | 描述 | |-----|-----|-----|-----| | functionName | String | 是 | 函数名,长度不大于32个字符 | | databaseName | String | 是 | 函数隶属数据库名,长度不大于32个字符 | | className | String | 是 | 函数类名 | | functionType | Integer | 是 | 函数类型,(JAVA(1) | | resourceUris | List<ResourceUri> | 是 | 函数资源 | | creator | String | 否 | 创建者 | | createTime | String | 否 | 创建时间 | | modifyTime | String | 否 | 修改时间 | | description | String | 否 | 描述信息,长度不大于1024个字符 |

ResourceUri对象： | 名称 | 类型 | 是否必须 | 描述 | |-----|-----|-----|-----| | resourceType | Integer | 是 | 函数资源类型, (JAR(1),FILE(2),ARCHIVE(3) | | resourceUri | String | 否 | 函数资源路径 |

ListFunctions

接口说明

获取函数信息列表

请求原型

```
GET /v1/database?function
```

请求参数

字段	类型	位置	是否必须	描述
pageNo	Integer	URL	否	页码
pageSize	Integer	URL	否	每页查询记录数
database	String	URL	否	按数据库名过滤
namePattern	String	URL	否	用户定义函数名的正则匹配表达式
creator	String	URL	否	按函数创建者过滤

响应参数

名称	类型	描述
pageNo	Integer	当前页数
pageSize	Integer	每页条数
totalCount	Integer	总数
result	List<Function>	函数详情列表

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database?function&pageNo=1&pageSize=10 HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "pageNo": 1,
  "pageSize": 10,
  "totalCount": 1,
  "result": [
    {
      "functionName": "count",
      "database": "db_01",
      "className": "com.baidu.edap.CountUDF",
      "functionType": 1,
      "resourceUris": [
        {
          "resourceType": 1,
          "resourceUri": "bos://abc/a.jar"
        },
        {
          "resourceType": 1,
          "resourceUri": "bos://abc/a.txt"
        }
      ],
      "createTime": "2019-06-01T23:00:10Z",
      "modifyTime": "2019-07-01T23:00:10Z",
      "description": "count tow numbers",
      "creator": "root"
    }
  ]
}
```

GetFunction

接口说明

获取函数详情

请求原型

GET /v1/database/\${databaseName}/function/\${functionName}

请求参数

名称	类型	位置	必须	说明
databaseName	String	URL	是	数据库名
functionName	String	URL	是	函数名

响应参数

参见Function对象说明

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
FunctionNotFound	Function not found.	404	函数不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database/db_01/function/count HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "functionName": "count",
  "database": "db_01",
  "className": "com.baidu.edap.CountUDF",
  "functionType": 1,
  "resourceUris": [
    {
      "resourceType": 1,
      "resourceUri": "bos://abc/a.jar"
    },
    {
      "resourceType": 1,
      "resourceUri": "bos://abc/a.txt"
    }
  ],
  "createTime": "2019-06-01T23:00:10Z",
  "modifyTime": "2019-07-01T23:00:10Z",
  "description": "count tow numbers",
  "creator": "root"
}
```

CreateFunction

接口说明

在指定数据库下创建函数，库下唯一，不能创建同名函数

请求原型

```
POST /v1/database/${databaseName}/function
```

请求参数

字段	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
functionName	String	Body	是	函数名
database	String	Body	是	函数隶属数据库名
className	String	Body	是	函数类名
functionType	Integer	Body	是	函数类型(JAVA(1)
resourceUris	List<ResourceUri>	Body	是	函数资源
description	String	Body	是	描述信息 长度不大于1024个字符

响应参数

除公共响应参数外，无特殊响应。

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
FunctionAlreadyExist	Functionalready exist.	409	数据库已存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/bd_01/function HTTP/1.1

Content-Length: xxx

Content-Type: application/json
{
  "functionName": "count",
  "database": "db_01",
  "className": "com.baidu.edap.CountUDF",
  "functionType": 1,
  "resourceUris": [
    {
      "resourceType": 1,
      "resourceUri": "bos://abc/a.jar"
    },
    {
      "resourceType": 1,
      "resourceUri": "bos://abc/a.txt"
    }
  ],
  "description": "count tow numbers"
}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

UpdateFunction

接口说明

更新函数信息, 除绑定的数据库和函数名不能更新，其余数据字段都可更新

请求原型

```
PUT /v1/database/${databaseName}/function/${functionName}
```

请求参数

字段	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
functionName	String	URL	是	函数名
className	String	Body	是	函数类名
resourceUris	List	Body	是	函数资源
description	String	Body	是	描述信息 长度不大于1024个字符

响应参数

除公共响应参数外，无特殊响应。

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
FunctionNotFound	Function not found.	404	函数不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
PUT /v1/database/bd_01/function/count HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: xxxx

Content-Type: application/json
{
  "className": "com.baidu.edap.Count",
  "resourceUris": [
    {
      "resourceType": 1,
      "resourceUri": "bos://abc/ax.jar"
    },
    {
      "resourceType": 1,
      "resourceUri": "bos://abc/a.txt"
    }
  ],
  "description": "count tow numbers"
}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

DeleteFunction

接口说明

删除函数

请求原型

```
DELETE /v1/database/${databaseName}/function/${functionName}
```

请求参数

字段	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
functionName	String	URL	是	函数名

响应参数

除公共响应参数外，无特殊响应。

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
FunctionNotFound	Function not found.	404	函数不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
DELETE /v1/database/db_01/function/count HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: 0
```

响应示例

HTTP/1.1 204 No Content
Content-Length: 0

表

🔗 表管理

表格信息 | 名称 | 类型 | 是否必须 | 描述 | |-----|-----|-----|-----| | tableName | String | 是 | 名称,仅支持小写字母、数字和下划线,当前数据库下唯一,长度不大于32个字符 | | tableType | String | 是 | 表格类型,枚举值: 管理表(MANAGED), 外部表(EXTERNAL) | | creator | String | 否 | 创建者 | | createTime | String | 否 | 创建时间 | | lastModifyTime | String | 否 | 表最近更新时间 | | description | String | 否 | 描述信息,长度不大于1024个字符 | | primaryKey | List<String> | 否 | 表格主键信息,目前只支持数据湖格式为hudi,iceberg时填写,其对应iceberg中的identified | | storage | Storage | 是 | 数据存储格式 | | columns | List<Column> | 是 | 普通字段列表,【可视化字段编辑】 | | partitions | List<Partition> | 否 | 分区字段列表,【可视化字段编辑】 | | properties | Map<String, String> | 否 | 表格属性,多条key-value组成 | | version | Integer | 否 | 表版本号 | | owner | String | 否 | 负责人名称 | | ownerType | String | 否 | 负责人类型,枚举值:{USER, ROLE, GROUP} | | authProperties | Map<String, String> | 否 | 安全认证信息 |

安全认证信息authProperties数据结构 | 名称 | 类型 | 是否必须 | 描述 | |-----|-----|-----|-----| | accessKey | String | 否 | 平台访问密钥 | | secretKey | String | 否 | 平台访问私钥 | | sessionTokenKey | String | 否 | 平台会话级口令密钥 | | authType | String | 否 | 认证类型 (UGI,KERBEROS) | | user | String | 是 (UGI) | 用户,长度不超过128 | | passwd | String | 是 (UGI) | 密码,长度不超过128 | | host | String | 是 (KERBEROS) | kerberos kdc host,长度不超过128 | | principal | String | 是 (KERBEROS) | kerberos参与者,长度不超过128 | | keytab | String | 是 (KERBEROS) | keytab文件BASE64编码结果,长度不超过1M |

表格存储属性 (Storage)

管理表/外表属性说明 (tableType=MANAGED/EXTERNAL)

名称	类型	是否必须	描述
format	String	是	文件格式,枚举值见存储格式类型定义
locationId	String	是	数据湖存储Id,指定后为table独有的location, tableType=MANAGED或tableType=EXTERNAL
relativePath	String	否	相对路径,与locationId所指向的目录拼接为表实际存储路径,未指定是默认以表名生成
locationUri	String	否	表格存储路径地址
fieldDelimiter	String	是	分隔符,如: \t, format=TEXTFILE时必须填,否则不填
collectionDelimiter	String	否	集合元素分隔符,字段类型为ARRAY,MAP,STRUCT时选填,其余情况不填
mapKeyDelimiter	String	否	MAP数据类型中key值分隔符
dataLakeType	String	否	数据湖类型,目前支持iceberg,hudi,deltalake
hudiTableType	String	否	hudi表类型 枚举值: COW、MOR,当datalakeType为hudi时此项为必填项
hudiQueryTableType	String	否	hudi请求表类型 枚举值: RO(读优化表),RT(实时表),OR(原表) 当datalakeType为hudi时此项为必填项

存储格式类型定义 (format) | 枚举值 | 描述 | |-----|-----| | AVRO | AVRO | | TEXTFILE | 文本格式 | | ORC | ORC | | PARQUET | PARQUET | | RCFILE | RCFILE |

表格字段列表 (Column) | 名称 | 类型 | 是否必须 | 描述 | |-----|-----|-----|-----| | name | String | 是 | 名称,仅支持小写字母、数字和下划线,以字母开头,长度不大于32个字符 | | type | String | 是 | 类型,见字段类型枚举值描述 | | nullable | Boolean | 是 | 是否能为空,枚举值: true, false | | description | String | 否 | 描述信息,长度不大于1024个字符 | | precision | Integer | 否 | 精度 | | scale | Integer | 否 | 小数位 | | complexTypeDetail | String | 是 | 复杂数据(ARRAY,,MAP,STRUCT)类型详情 | | id | String | 否 | 数据湖格式为iceberg时,系统默认生成,其表示字段的唯一标识符 |

字段类型枚举值 | 枚举值 | 描述 | |-----|-----| | TINYINT | 微整型 | | SMALLINT | 短整型 | | INT | 整型 | | BIGINT | 长整型 | | DOUBLE | 双精度浮点数 | | FLOAT | 浮点数 | | DECIMAL | 精确数字类型 | | STRING | 字符 | | VARCHAR | 可变长度字符串 | | CHAR | 固定长度字符串 | | BOOLEAN | 布尔值 | | BYTES | 二进制数据 | | DATE | 日期,文本类型格式为: yyyy-MM-dd | | TIMESTAMP | 时间戳,文本类型格式为: yyyy-MM-dd HH:mm:ss.SSSSSSSSS | | ARRAY | 一组有序字段,ARRAY<data_type> | | MAP | 一组无序的键/值对,MAP<primitive_type, data_type> | | STRUCT | 一组命名的字段,STRUCT<col_name: data_type, ...> |

表格分区信息(Partition) | 名称 | 类型 | 是否必须 | 描述 | |-----|-----|-----|-----| | name | String | 是 | 名称,仅支持小写字母、数字和下划线,以字母开头,长度不大于32个字符 | | type | String | 是 | 类型,见分区字段类型枚举值描述 | | precision | Integer | 否 | 精度 | | scale | Integer | 否 | 小数位 | | description | String | 否 | 描述信息,长度不大于1024个字符 | | function | String | 否 | 分区函数,数据湖格式为iceberg时设置,枚举值如结构体Transform Function所示 | | functionParameter | Integer | 否 | 分区函数选择bucket和truncated时填写 |

分区字段类型枚举值 | 枚举值 | 描述 | |-----|-----| | TINYINT | 微整型 | | SMALLINT | 短整型 | | INT | 整型 | | BIGINT | 长整型 | | DOUBLE | 双精度浮点数 | | FLOAT | 浮点数 | | DECIMAL | 精确数字类型 | | STRING | 字符 | | VARCHAR | 可变长度字符串 | | CHAR | 固定长度字符串 | | BOOLEAN | 布尔值 | | BYTES | 二进制数据 | | DATE | 日期,文本类型格式为: yyyy-MM-dd | | TIMESTAMP | 时间戳,文本类型格式为: yyyy-MM-dd HH:mm:ss.SSSSSSSSS |

iceberg表格分区函数类型(Transform Function) | 枚举值 | 描述 | 支持类型 | |-----|-----|-----|-----| | year | 按年分区 | DATE,TIMESTAMP | | month | 按月分区 | DATE,TIMESTAMP | | day | 按天分区 | DATE,TIMESTAMP | | hour | 按小时分区 | TIMESTAMP | | bucket | 按哈希值modN存储桶分区,填写格式bucket(N,col) | TINYINT,SMALLINT,INT, BIGINT,DECIMAL,DATE,TIMESTAMP,STRING,BYTES | | truncate | 按值截断为L分区,填写格式truncate(L, col) | TINYINT,SMALLINT,INT,BIGINT, DECIMAL,STRING |

ListTables

接口说明

获取表信息列表

请求原型

```
GET /v1/database/${databaseName}/table
```

请求参数

字段	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
pageNo	Integer	QueryString	是	当前页数
pageSize	Integer	QueryString	是	每页条数
tableTypes	String	QueryString	否	表的类型，枚举值(MANAGED，EXTERNAL)，多个类型筛选时以逗号分隔每种类型。
namePattern	String	QueryString	否	搜索关键词

响应参数

名称	类型	描述
pageNo	Integer	当前页数
pageSize	Integer	每页条数
totalCount	Integer	总数
tables	List<TableBaseInfo>	表基本信息列表

TableBaseInfo

名称	类型	描述
tableName	String	表名称
creator	String	创建者
createTime	String	创建时间
description	String	描述信息
tableType	String	表类型

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database/db_01/table?pageNo=1&pageSize=10 HTTP/1.1
Host: edap.bj.baidubce.com
Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxx
Content-Type: application/json
{
  "pageNo": 1,
  "pageSize": 10,
  "tables": [
    {
      "createTime": "2024-02-01T09:13:17Z",
      "creator": "root",
      "description": "complex table",
      "tableName": "tab1",
      "tableType": "MANAGED"
    },
    {
      "createTime": "2024-01-25T07:01:41Z",
      "creator": "root",
      "description": "this a external table",
      "tableName": "es_tab",
      "tableType": "EXTERNAL"
    }
  ],
  "totalCount": 2
}
```

ListTableNames

接口说明

获取表名称列表

请求原型

GET /v1/database/\${databaseName}/table?name

请求参数

字段	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
pageNo	Integer	QueryString	是	当前页数
pageSize	Integer	QueryString	是	每页条数
tableTypes	String	QueryString	否	表的类型，枚举值{MANAGED，EXTERNAL，VIRTUAL_VIEW，MATERIALIZED_VIEW，MAPPING}，多个类型筛选时以逗号分隔每种类型。
namePattern	String	QueryString	否	搜索关键词

响应参数

名称	类型	描述
tableNames	List<String>	表的名称列表
pageNo	Integer	当前页数
pageSize	Integer	每页条数
totalCount	Integer	总数

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database/db_01/table?name&pageNo=1&pageSize=10 HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
```

GetTable

接口说明

获取表信息

请求原型

```
GET /v1/database/${databaseName}/table/${tableName}
```

请求参数

名称	类型	位置	必须	描述
databaseName	String	URL	是	数据库名称
tableName	String	URL	是	表名称
needAuthProperties	Boolean	QueryString	否	是否需要返回安全认证信息，默认false

响应参数

参见Table对象参数说明

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database/db_01/table/tb_01?needAuthProperties=true HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length:0
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "columns": [
    {
      "description": "id",
      "name": "id",
      "nullable": false,
      "type": "STRING"
    },
    {
      "complexTypeDetail": "<string,int>",
      "description": "map column",
      "name": "map_column",
      "nullable": false,
      "type": "MAP"
    }
  ],
  "createTime": "2024-02-01T09:13:17Z",
  "creator": "root",
  "description": "complex table",
  "lastModifyTime": "2024-02-01T09:13:17Z",
  "partitions": [
    {
      "description": "",
      "name": "par",
      "type": "STRING"
    }
  ],
  "properties": {
    "skip.header.line.count": "1"
  },
  "authProperties": {
    "accessKey": "key1",
    "secretKey": "key2",
    "sessionTokenKey": "key3"
  },
  "storage": {
    "collectionDelimiter": "|",
    "fieldDelimiter": ",",
    "format": "TEXTFILE",
    "mapKeyDelimiter": ":",
    "locationId": "loc1",
    "relativePath": "tb_01",
    "locationUri": "bos://edap/db_01/tb_01"
  },
  "tableName": "tb_01",
  "tableType": "MANAGED",
  "version": 1
}
```

CreateTable

接口说明

创建表

请求原型

POST /v1/database/\${databaseName}/table

请求参数

字段	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
tableName	String	Body	是	表名称，支持字母、数字和下划线，当前数据库下唯一，长度不大于32个字符
tableType	String	Body	是	表类型，枚举值：映射表（MAPPING）管理表(MANAGED), 外部表(EXTERNAL), 视图(VIRTUAL_VIEW), 物化视图(MATERIALIZED_VIEW)
storage	Storge	Body	是	参考Storge对象说明，不同表类型填写对应参数
columns	Column	Body	是	参考Column对象说明
partitions	Partition	Body	否	参考Partition对象说明
owner	String	Body	否	负责人名称
ownerType	String	Body	否	负责人类型，枚举值:{USER，ROLE，GROUP}
properties	Map	Body	否	定义表的参数和属性的键值对
description	String	Body	否	描述信息，长度不大于1024个字符

响应参数

除公共响应参数外，无特殊响应。

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
AccessDenied	Access Denied.	403	无权限
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
TableAlreadyExist	Table already exist.	409	表已存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
{
  "tableName": "tab1",
  "description": "this is a table",
  "tableType": "MANAGED",
  "partitions": [
    {
      "name": "c",
      "type": "STRING",
      "description": "c"
    }
  ],
  "storage": {
    "format": "PARQUET",
    "locationId": "loc",
    "relativePath": "tab1"
  },
  "columns": [
    {
      "name": "a",
      "type": "STRING",
      "description": "a",
      "nullable": false
    }
  ],
  "properties": {
    "skip.header.line.count": "1"
  }
}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

UpdateTable

接口说明

更新表格信息，可更新信息包括，表名，表描述信息，字段信息，分区信息，表参数, 数据存储格式

请求原型

```
PUT /v1/database/${dbName}/table/${tableName}
```

请求参数

字段	类型	位置	是否必须	描述
dbName	String	URL	是	数据库名
tableName	String	URL	是	表名
tableName	String	Body	是	表名称，支持字母、数字和下划线，当前数据库下唯一，长度不大于32个字符
tableType	String	Body	是	表类型，枚举值：映射表（MAPPING）管理表(MANAGED), 外部表(EXTERNAL), 视图(VIRTUAL_VIEW), 物化视图(MATERIALIZED_VIEW)
storage	Storge	Body	是	参考Storge对象说明，不同表类型填写对应参数
columns	Column	Body	是	参考Column对象说明
partitions	Partition	Body	否	参考Partition对象说明
owner	String	Body	否	负责人名称
ownerType	String	Body	否	负责人类型，枚举值:{USER，ROLE，GROUP}
properties	Map	Body	否	定义表的参数和属性的键值对
description	String	Body	否	描述信息，长度不大于1024个字符

响应参数

除公共响应参数外，无特殊响应。

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
AccessDenied	Access Denied.	403	无权限
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
TableAlreadyExist	Table already exist.	409	表已存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
PUT /v1/database/db_01/table/tb_01 HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: xxx

Content-Type: application/json
{
  "tableName": "tb_02",
  "description": "描述信息",
  "storage": {
    "format": "TEXTFILE",
    "fieldDelimiter": "\t",
    "locationId": "lc_02",
    "relativePath": "tb_02"
  },
  "properties": {
    "transactional": "false",
    "skip.header.line.count": "1"
  },
  "columns": [
    {
      "name": "a",
      "type": "STRING",
      "description": "a",
      "nullable": false
    },
    {
      "name": "height",
      "type": "decimal",
      "nullable": true,
      "description": "高度",
      "precision": 18,
      "scale": 18
    }
  ],
  "partitions": [
    {
      "name": "c",
      "type": "STRING",
      "description": "c"
    }
  ]
}
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

DeleteTable

接口说明

删除表，MANAGED类型表格在删除时，同时清空表真实数据。

请求原型

```
DELETE /v1/database/${databaseName}/table/${tableName}
```

请求参数

字段	类型	位置	是否必须	描述
databaseName	String	URL	是	数据库名
tableName	String	URL	是	表名

响应参数

除公共响应参数外，无特殊响应。

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
AccessDenied	Access Denied.	403	无权限
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
DELETE /v1/database/db_01/table/tb_01 HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: 0
```

响应示例

```
HTTP/1.1 204 No Content

Content-Length: 0
```

ListTableVersions

接口说明

获取表格版本信息列表

请求原型

```
GET /v1/database/${databaseName}/table/${tableName}?versionList
```

请求参数

名称	类型	位置	是否必须	描述
databaseName	String	URL	是	库名
tableName	String	URL	是	表名
pageNo	Integer	QueryString	是	当前页数
pageSize	Integer	QueryString	是	每页条数
startTime	String	QueryString	否	开始时间
endTime	String	QueryString	否	结束时间

响应参数

名称	类型	描述
pageNo	Integer	当前页数
pageSize	Integer	每页条数
totalCount	Integer	总数
versions	List<TableVersion>	表格版本信息

TableVersion

名称	类型	描述
version	Int	表格版本号
creator	String	创建者
createTime	String	创建时间

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database/default/table/tb?versionList&pageNo=1&pageSize=10 HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "pageNo": 1,
  "pageSize": 10,
  "totalCount": 2,
  "versions": [
    {
      "version": 0,
      "creator": "user_1",
      "createTime": "2019-06-01T23:00:10Z"
    },
    {
      "version": 1,
      "creator": "user_2",
      "createTime": "2020-06-01T23:00:10Z"
    }
  ]
}
```

GetTableByVersion

接口说明

获取指定版本的表格信息

请求原型

```
GET /v1/database/${databaseName}/table/${tableName}?version
```

请求参数

名称	类型	位置	是否必须	描述
databaseName	String	URL	是	库名
tableName	String	URL	是	表名
version	Integer	QueryString	是	表格版本号

响应参数

参见Table对象参数说明

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
DatabaseNotFound	Database not found.	404	数据库不存在
TableNotFound	Table not found.	404	表不存在
InternalServerError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
GET /v1/database/db_01/table/tb_01?version=2 HTTP/1.1

Host: edap.bj.baidubce.com

Content-Length:0
```

响应示例

```
HTTP/1.1 200 OK

Content-Length: xxxx

Content-Type: application/json
{
  "columns": [
    {
      "description": "id",
      "name": "id",
      "nullable": false,
      "type": "STRING"
    },
    {
      "complexTypeDetail": "<string,int>",
      "description": "map column",
      "name": "map_column",
      "nullable": false,
      "type": "MAP"
    }
  ],
  "createTime": "2024-02-01T09:13:17Z",
  "creator": "root",
  "description": "complex table",
  "lastModifyTime": "2024-02-01T09:13:17Z",
  "partitions": [
    {
      "description": "",
      "name": "par",
      "nullable": false,
      "type": "STRING"
    }
  ],
  "properties": {
    "skip.header.line.count": "1"
  },
  "storage": {
    "collectionDelimiter": "|",
    "fieldDelimiter": ",",
    "format": "TEXTFILE",
    "mapKeyDelimiter": ":",
    "locationId": "lc_02",
    "relativePath": "tb_01",
    "location": "bos://edap/db_01/tb_01"
  },
  "tableName": "tb_01",
  "tableType": "MANAGED",
  "version": 2
}
```

库



库管理

ListDatabases 获取数据库信息列表 请求原型

GET /v1/database

请求参数 | 名称 | 类型 | 位置 | 必须 | 描述 | | ----- | ----- | ----- | ----- | ----- | | pageNo | Integer | QueryString | 否 | 页码，默认1 | | pageSize | Integer | QueryString | 否 | 每页查询记录数，默认10 | | namePattern | String | QueryString | 否 | 搜索关键词 |

响应参数

名称	类型	描述
pageNo	Integer	当前页数
pageSize	Integer	每页条数
totalCount	Integer	总数
databases	List<Database>	数据库信息列表

Database对象

名称	类型	描述
database	String	数据库名称
creator	String	创建者
createTime	String	创建时间
description	String	描述信息

错误码 公有云标准错误返回码

Code	message	HTTP状态码	描述
InternalServerError	We encountered an internal error. Please try again.	500	服务器内部错误

请求示例

GET /v1/database?pageNo=1&pageSize=10 HTTP/1.1
Host: edap.bj.baidubce.com
Content-Length: 0

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxx
Content-Type: application/json
{
  "pageNo": 1,
  "pageSize": 10,
  "totalCount": 2,
  "databases": [
    {
      "createTime": "2023-11-22T08:21:27Z",
      "creator": "root",
      "database": "default",
      "description": "this is default db"
    },
    {
      "createTime": "2023-12-27T15:56:42Z",
      "creator": "root",
      "database": "db",
      "description": "my db"
    }
  ]
}
```

ListDatabaseNames

接口说明

获取用户数据库名称列表

请求原型

GET /v1/database?name

请求参数

名称	类型	位置	必须	描述
pageNo	Integer	QueryString	否	页码，默认1
pageSize	Integer	QueryString	否	每页查询记录数，默认10
namePattern	String	QueryString	否	搜索关键词

响应参数

名称	类型	描述
pageNo	Integer	当前页数
pageSize	Integer	每页条数
totalCount	Integer	总数
databaseNames	List<String>	数据库名称列表

错误码

公有云标准错误返回码

Code	message	HTTP状态码	描述
InternalError	We encountered an internal error. Please try again.	500	服务器内部错误

请求示例

```
GET /v1/database?name&pageNo=1&pageSize=10 HTTP/1.1
Host: edap.bj.baidubce.com
Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxx
Content-Type: application/json
{
  "pageNo": 1,
  "pageSize": 10,
  "totalCount": 2,
  "databaseNames": [
    "default",
    "dept"
  ]
}
```

GetDatabase 获取数据库信息 请求原型

```
GET /v1/database/${databaseName}
```

请求参数 | 名称 | 类型 | 位置 | 必须 | 说明 || ----- | ----- | ----- | ----- || databaseName | String | URL | 是 | 数据库名 |

响应参数 | 名称 | 类型 | 描述 || ----- | ----- | ----- || database | String | 数据库名称 || locationId | String | 数据湖存储id || locationUri | String | 库路径 || creator | String | 创建者 || createTime | String | 创建时间 || description | String | 描述信息 || owner | String | 负责人名称 || ownerType | String | 负责人类型，枚举值:{USER，ROLE，GROUP} || properties | Map | 定义数据库的参数和属性的键值对 |

错误码 公有云标准错误返回码

Code	message	HTTP状态码	描述
DatabaseNotFound	Database not found.	404	数据库不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内部错误

请求示例

```
GET /v1/database/default HTTP/1.1
Host: edap.bj.baidubce.com
Content-Length: 0
```

响应示例

```
HTTP/1.1 200 OK
Content-Length: xxxx
Content-Type: application/json
{
  "createTime": "2023-11-22T08:21:27Z",
  "creator": "root",
  "database": "default",
  "description": "this is default db",
  "locationId": "default_location",
  "locationUri": "bos://default"
  "owner": "hdfs",
  "ownerType": "USER",
  "properties": {
    "department": "big data"
  }
}
```

CreateDatabase 新增数据库 请求原型

POST /v1/database

请求参数 | 名称 | 类型 | 位置 | 是否必须 | 描述 || ----- | ----- | ----- | ----- | ----- || database | String | Body | 是 | 数据库名称，仅支持字母、数字和下划线，大小写不敏感，账户下全局唯一，长度不大于32个字符 || locationId | String | Body | 是 | 数据湖存储id || owner | String | Body | 否 | 负责人名称 || ownerType | String | Body | 否 | 负责人类型，枚举值:{USER, ROLE , GROUP} || properties | Map | Body | 否 | 定义数据库的参数和属性的键值对 || description | String | Body | 否 | 描述信息，长度不大于1024个字符 |

响应参数 除公共响应参数外，无特殊响应。

错误码 公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
AccessDenied	Access Denied.	403	无权限
DatabaseAlreadyExist	Database already exist.	409	数据库已存在
InternalError	We encountered an internal error. Please try again.	500	服务器内部错误

请求示例

```
POST /v1/database HTTP/1.1
Content-Length: xxxx
Content-Type: application/json
{
  "database": "db_01",
  "locationId": "loc_01",
  "properties": {
    "department": "applied mathematics"
  },
  "description": "applied mathematics"
}
```

响应示例

```
HTTP/1.1 204 No Content
Content-Length: 0
```

UpdateDatabase 更新数据库信息 请求原型

PUT /v1/database/\${databaseName}

请求参数 | 名称 | 类型 | 位置 | 是否必须 | 描述 || ----- | ----- | ----- | ----- | ----- || databaseName | String | URL | 是 | 原始数据库名称 || database | String | Body | 是 | 数据库名称，仅支持字母、数字和下划线，大小写不敏感，账户下全局唯一，长度不大于32个字符 || locationId | String | Body | 是 | 数据湖存储id || owner | String | Body | 否 | 负责人名称 || ownerType | String | Body | 否 | 负责人类型，枚举值:{USER, ROLE , GROUP} || properties | Map | Body | 否 | 定义数据库的参数和属性的键值对 || description | String | Body | 否 | 描述信息，长度不大于1024个字符 |

响应参数 除公共响应参数外，无特殊响应。

错误码 公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	请求体规则校验错误
AccessDenied	Access Denied.	403	无权限
DatabaseAlreadyExist	Database already exist.	409	数据库已存在
DatabaseNotFound	Database not found.	404	数据库不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内不错误

请求示例

```
POST /v1/database/db_01 HTTP/1.1
Host: edap.bj.baidubce.com
Content-Length: xxx
Content-Type: application/json
{
  "database": "db_02",
  "locationId": "loc_02",
  "properties": {
    "department": "big data"
  },
  "description": "big data"
}
```

响应示例

```
HTTP/1.1 204 No Content
Content-Length: 0
```

DeleteDatabase 删除数据库（仅允许删除没有数据库表格的数据库） 请求原型

```
DELETE /v1/database/${databaseName}
```

请求参数 | 名称 | 类型 | 位置 | 是否必须 | 描述 || ----- | ----- | ----- | ----- || databaseName | String | URL | 是 | 数据库名称 |

响应参数 除公共响应参数外，无特殊响应。

错误码 公有云标准错误返回码

Code	message	HTTP状态码	描述
BadRequest	Bad request, please check your request	400	该库下存在表无法删除
AccessDenied	Access Denied.	403	无权限
DatabaseNotFound	Database not found.	404	数据库不存在
InternalError	We encountered an internal error. Please try again.	500	服务器内部错误

请求示例

```
DELETE /v1/database/db_01 HTTP/1.1
Host: edap.bj.baidubce.com
Content-Length: 0
```

响应示例

```
HTTP/1.1 204 No Content
Content-Length: 0
```

视频专区

产品介绍

🔗 数据湖管理与分析平台EasyDAP产品介绍

- 本视频介绍百度数据湖管理与分析平台EasyDAP的产品定位、特性、架构及应用场景，以典型的大数据开发和治理流程为线索，分模块介绍EasyDAP的产品功能，为企业数据中台建设、数据治理、数据资产管理提供产品级解决方案。



功能模块介绍及最佳实践

🔗 数据建模介绍及最佳实践

- 本视频介绍EasyDAP数据建模模块的作用和定位，详细介绍建模模块的功能使用，并通过最佳实践示例的讲解，帮助用户快速上手，玩转大数据治理开发。



🔗 数据集成介绍及最佳实践

- 本视频介绍EasyDAP数据集成模块的作用和定位，详细介绍集成模块的功能使用，并通过最佳实践示例的讲解，帮助用户快速上手，玩转大数据治理开发。



🔗 数据开发介绍及最佳实践

- 本视频介绍EasyDAP数据开发模块的作用和定位，详细介绍开发模块的功能使用，并通过最佳实践示例的讲解，帮助用户快速上手，玩转大数据治理开发。



🔗 数据质量介绍及最佳实践

- 本视频介绍EasyDAP数据质量模块的作用和定位，详细介绍质量模块的功能使用，并通过最佳实践示例的讲解，帮助用户快速上手，玩转大数据治理开发。



🔗 数据服务介绍及最佳实践

- 本视频介绍EasyDAP数据服务模块的作用和定位，详细介绍服务模块的功能使用，并通过最佳实践示例的讲解，帮助用户快速上手，玩转大数据治理开发。



常见问题

一、准备工作

- 1、使用EDAP产品组合推荐
- 少任务量时：推荐使用 EDAP版本+EDAP资源组+BOS存储
- 大任务量时：推荐使用 EDAP版本+BMR计算+BOS存储

二、数据集成

- 1、实时同步支持哪些源端、目标端？
- 源端：PostgreSQL、MySQL、SQL Server、Oracle
- 目标端：EDAPDataLake、Doris
- 2、离线同步支持哪些源端、目标端？
- 源端：MySQL、SQL Server、Oracle、Hana、PostgreSQL
- 目标端：EDAPDataLake、Doris、Hive

三、数据开发

- 1、EDAP资源组支持哪些作业类型？
- FlinkSQL、FlinkJar、SparkSQL、SparkScala、PySpark、SparkJar、JDBCSQL
- 2、作业告警如何配置？
- 登陆公有云 **云监控BCM**，点击菜单 **事件监控** -> **云产品事件**，进一步点击 **添加报警策略**，在基本信息下 **产品类型**中 输入EDAP 逐步操作添加相关报警策略。

四、数据治理

- 1、EDAP数据血缘的解析策略
- 支持，针对SQL脚本作业、可视化ETL作业、数据集成等作业类型，可以进行血缘关系自动解析，针对Shell、Jar包等作业类型，提供手动填报的入口，实现数据血缘覆盖率100%。
- 2、EDAP具备哪些数据安全能力？
- 细粒度的数据权限管理，库、表、字段级别，可以添加行过滤规则； 隐私数据保护，支持哈希、遮掩、变换、加密等多种脱敏算法，支持进行数据静态和动态脱敏。
- 3、是否支持日志审计？
- 支持，公有云通过对接云审计产品，实现用户操作日志的保存和查看。

服务等级协议SLA

EDAP服务等级协议SLA

协议生效日期：2024年10月30日

本服务等级协议（Service Level Agreement，以下简称“SLA”）约定了百度智能云向客户提供的百度数据湖管理与分析平台EasyDAP（简称“EDAP”）的服务可用性等级指标及赔偿方案。

1. 定义

1.1 服务周期

一个服务周期为一个自然月。如客户开通EDAP实例不满一个自然月，按照一个自然月计算。

1.2 服务周期总分钟数

服务周期内的总天数 * 24（小时）* 60（分钟）计算。

1.3 服务不可用分钟数

当某一分钟内，客户某个指定账号在指定区域的EDAP建立连接的连续尝试均失败，则视为该分钟内该账号该区域的EDAP产品服务不可用。

1.4 月度服务费用

指在一个服务周期（即自然月）内，客户就服务周期内的EDAP服务所支付的服务费总额，如客户一次性支付多个服务周期的服务费，则将按照所购买的月数分摊计算月度服务费。

2. 服务可用性

2.1 服务可用性计算公式

服务可用性=（1-服务不可用分钟数/服务周期总分钟数） * 100%

2.2 服务可用性承诺

EDAP服务可用性不低于99.90%，如EDAP未达到前述可用性承诺，客户可以根据本协议第3条约定获得赔偿。赔偿范围不包括以下原因所导致的不可用：

- 1. 百度云预先通知客户后进行系统维护所引起的，包含但不限于割接、维修、升级、补丁和模拟故障演练等；
- 2. 客户操作（如通过控制台/API等方式对集群进行删除/停止/重启等操作，对作业/连接/驱动/等进行删除等操作）引起的不可用；
- 3. 任何由百度云所属设备以外的网络、设备故障或配置调整引起的；
- 4. 由于客户的应用程序或数据信息受到黑客攻击而引起的；
- 5. 客户维护不当或保密不当致使数据、口令、密码等丢失或泄漏所引起的；
- 6. 由于客户的疏忽或因客户授权的操作所引起的；
- 7. 任何由非EDAP所引起的，比如：由EDAP所依赖的上下游软件、硬件产品故障或配置调整引起的；由于资源组被用户手动扩容规格等变更操作，以及用户账号欠费导致的资源组不可用等场景中；数据开发模块中作业调度配置不合理导致的调度失败/未调度/作业执行超时/作业失败后不重试等情形；数据集成模块中由于源端数据库或目的端数据库自身原因导致的；数据集成模块中，由于源端数据库或者目的端数据库用户名、密码修改后，客户未及时在EDAP侧更新配置导致的等；数据服务模块中，由于数据源不可用、数据源性能瓶颈、网络环境等原因导致的服务不可用；
- 8. 因您上传的自定义任务、高危 shell（包含不限于删除重要文件、数据等），高危 SQL（包含不限于删除重要库表数据，修改权限等）等任务造成的故障、损失；
- 9. 由于客户未遵循百度云官网所展示的产品使用文档或使用建议引起的；
- 10. EDAP只赔付EDAP自身，不赔付EDAP所关联的其他云服务；
- 11. 使用EDAP体验版服务不计入服务不可用统计；
- 12. 由不可抗力以及意外事件或其他非百度智能云原因所造成的EDAP无法正常使用引起的。

3. 服务赔偿条款

3.1 赔偿标准

根据客户某一百度智能云账号下某一地域的EDAP的月度服务可用性，按照下表中的标准计算赔偿金额，赔偿方式仅限于用于购买EDAP的代金券，且赔偿总额不超过未达到服务可用性承诺的当月客户就该账号下该地域支付的月度服务费的25%。

服务可用性	赔偿代金券金额
99.00%≤服务可用性<99.90%	月度服务费的10%
服务可用性<99.00%	月度服务费的25%

3.2 赔偿申请时限

客户可以在每月第五（5）个工作日中对上个月没有达到可用性的服务提出赔偿申请。赔偿申请必须限于在消息服务没有达到可用性的相关月份结束后两（2）个月内提出。超出申请时限的赔偿申请将不被受理。

4. 其他说明

- (1) 在法律法规允许的范围内，百度智能云负责对本协议进行解释说明。
- (2) 本协议一经公布立即生效，百度智能云有权对本SLA条款作出修改。如本SLA条款有任何修改，百度智能云将提前30天以网站公示或发送邮件的方式通知您。如您不同意百度智能云对SLA所做的修改，您有权停止使用消息服务，如您继续使用消息服务，则视为您接受修改后的SLA。
- (3) 本协议项下百度智能云对于用户所有的通知均可通过网页公告、站内信、电子邮件、手机短信或其他形式等方式进行；该等通知于发送之日视为已送达收件人。因用户未及时获知百度智能云的服务变更或终止条款遭受损失的，百度智能云不承担任何责任。
- (4) 协议的订立、执行和解释及争议的解决均应适用中国法律并受中国法院管辖。如双方就本协议内容或其执行发生任何争议，双方应尽量友好协商解决；协商不

成时，任何一方均可向北京市海淀区人民法院提起诉讼。

(5) 本协议构成双方对本协议之约定事项及其他有关事宜的完整协议，除本协议规定的之外，未赋予本协议各方其他权利。

(6) 如本协议中的任何协议无论因何种原因完全或部分无效或不具有执行力，本协议的其余协议仍有效并且有约束力。

(7) 关于用户约束条款，详见 [《百度智能云用户服务协议》](#) 中的"用户的权利与义务"相关条款内容。

(8) 关于服务商免责条款，详见 [《百度智能云用户服务协议》](#) 中的"免责声明"相关条款内容。