

DataBuilder 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
产品动态	3
功能发布记录	3
产品介绍	3
产品简介	3
核心概念	4
快速入门	4
角色与权限	4
快速上手	7
操作指南	11
工作空间	11
元存储	12
工作区	13
元数据	14
计算资源	20
工作流	21
运行记录	22
常见问题	23

产品动态

功能发布记录

🕒 2025年3月

发布时间	功能名称	发布地域	功能描述
3-31	工作空间	保定	工作空间提供团队协作环境，支持在工作空间内完成数据管理、数据处理、数据开发的全流程工作
3-31	元数据	保定	元数据提供元数据和数据统一存储的能力，通过组织和管理数据、模型和算子资产的层级结构，提供跨工作空间的细粒度授权和数据发现功能
3-31	工作区	保定	工作区提供文件管理能力，支持文件和文件夹的增删改查
3-31	计算资源	保定	计算资源模块上线全托管的Ray计算能力，用于大数据处理和机器学习任务
3-31	工作流	保定	提供算子可视化编排和管理能力，支持算子开发调试。预置基础算子和工作流模板，便于构建和运行数据处理任务，同时提供工作流运行记录，支持对工作流进行管理，支持工作流查看、重跑、停止、暂停、恢复和删除等操作

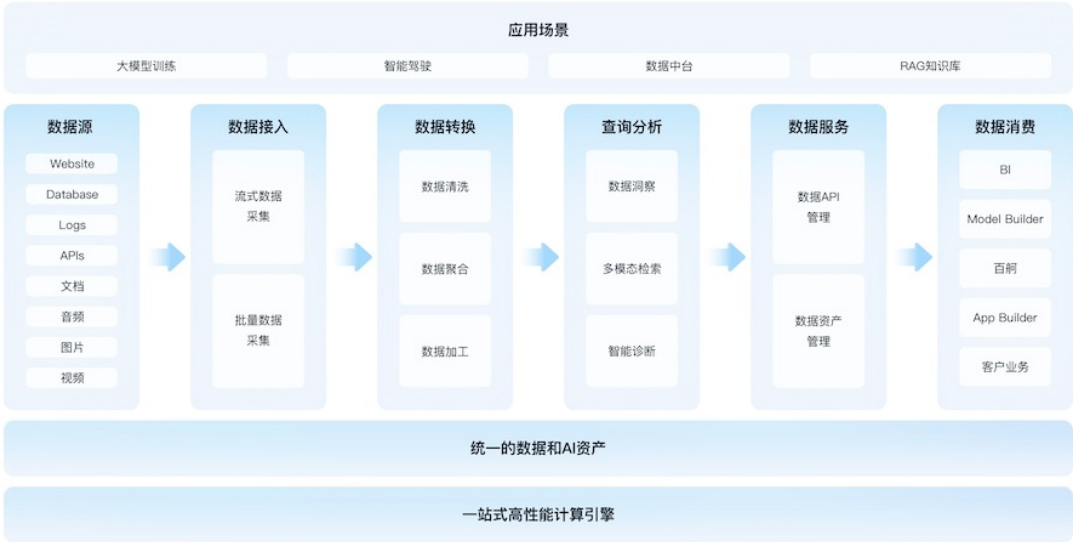
产品介绍

产品简介

🕒 产品介绍

百度智能云千帆数据智能平台DataBuilder，提供多模态数据管理、高性能计算、智能开发平台和丰富算子能力。支持用户完成一站式数据治理、数据加工和数据应用，帮助用户提升数据分析、模型训练和AI原生应用迭代效率。

🕒 产品架构



🕒 产品优势

🕒 开放易用

提供在线开发平台，面向数据工程、数据科学、机器学习与商业分析等场景。支持多团队多角色协同，敏捷完成数据任务。

🕒 智能高效

支持智能数据洞察，帮助用户解析复杂数据，并为业务决策提供支持。

提供多任务可视化调度和分布式并行能力，带给用户敏捷流畅的数据开发体验。

🔗 统一管理

通过统一元数据和lakehouse架构技术，打破数据孤岛，实现多模态数据的统一存储与集中管控。支持细粒度权限控制，全链路数据血缘追踪，让治理更简单。

核心概念

🔗 数据处理相关

🔗 工作流

通过编排数据处理任务和各种计算资源，以实现多模态数据处理的完整过程。

🔗 计算资源

包含常驻资源、Serverless资源两种类型。

🔗 数据管理相关

🔗 数据目录 (Catalog)

是组织和管理资产的层级结构（资产包括数据、模型、算子），提供跨工作空间的细粒度授权、审计和数据发现能力。

🔗 数据模式 (Schema)

用于管理资产的最小集合，类似于传统的数据库。

🔗 数据卷 (Volume)

用于存储和管理非结构化数据。

🔗 数据表 (Table)

用于存储和管理结构化数据。

🔗 算子 (Operator)

一段可以被重复调用的代码块，用于执行特定的任务或操作。

🔗 数据开发相关

🔗 工作空间

是实现多团队协同，数据管理、数据开发、数据处理的协作空间，完成数据工程、数据科学、数据分析的工作。

🔗 文件管理

类似操作系统的文件系统，有目录管理和文件编辑，用户可以新建、上传和管理工程文件和类库等。

快速入门

角色与权限

DataBuilder提供了完善的权限管理机制，支持在产品级与模块级对权限进行管控，其中产品级权限通过IAM权限体系管理，模块级权限通过RBAC权限模型管理，本文为您详细介绍 DataBuilder 的权限体系。

🔗 产品级权限

DataBuilder 产品级权限策略控制通过IAM权限策略实现，您可以为IAM用户授予系统内置的策略，实现DataBuilder管理与操作

的权限管控。

🔗 IAM基本概念

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，您可以对不同员工赋予产品的不同权限，以共享账户内的资源，完成他们的工作。当您的企业存在需要多用户协同工作、分享资源时，推荐使用多用户访问控制。更多概念请查看[IAM说明](#)。

🔗 IAM预设策略

DataBuilderFullControlAccessPolicy

完全控制管理DataBuilder的权限。

DataBuilderReadAccessPolicy

只读访问DataBuilder的权限。

🔗 模块级权限

DataBuilder根据功能使用范围分别提供空间管理员、空间普通用户、元存储管理员3类角色，不同模块的使用权限体系是基于RBAC（Role-Based Access Control）权限模型构建的，您可以直接使用这些角色给用户授权。

🔗 空间管理员

说明：向空间内添加普通用户、创建空间计算资源、创建数据目录、数据模式元数据。

对象	操作
工作空间	• 查看 • 添加普通用户
工作区	• 新建文件夹 • 上传文件
元数据	• 数据目录增删改查 • 数据模式增删改查 • 数据卷增删改查 • 数据表增删改查
计算资源	• 新建资源 • 删除资源 • 查看资源
工作流	• 新建工作流 • 删除工作流 • 修改及运行工作流 • 查看工作流
运行记录	• 删除运行记录 • 重跑运行记录 • 查看运行记录

 空间普通用户

说明：编写数据处理作业、编排工作流，对结构化、非结构化数据进行处理。

对象	操作
工作区	- 新建文件夹 - 上传文件
元数据	- 数据模式增删改查 - 数据卷增删改查 - 数据表增删改查
计算资源	查看资源
工作流	- 新建工作流 - 删除工作流 - 修改及运行工作流 - 查看工作流
运行记录	- 删除运行记录 - 重跑运行记录 - 查看运行记录

🔗 元存储管理员

说明：管理该租户下全局元数据，仅针对元数据有特殊权限。而在空间内，用户权限则取决于其是空间管理员或空间普通用户。

对象	操作
元数据	- 数据目录增删改查 - 数据模式增删改查 - 数据卷增删改查 - 数据表增删改查

快速上手

DataBuilder提供多模态数据管理、高性能计算、智能开发平台和丰富算力能力，支持用户完成一站式数据治理、数据加工和数据应用，本文以DataBuilder的部分核心功能为例，指导您使用DataBuilder接入非结构化数据并进行数据处理、写入元数据、查看处理结果。

🔗 入门简介

通过本快速入门，您可以快速完成以下操作。

1. 多模态数据管理：在DataBuilder的元数据管理模块，创建数据卷，将本地视频上传至平台。
2. 计算资源：在DataBuilder的计算资源模块，创建Ray计算资源，并支持GPU规格。

3. 数据处理：在工作流模块，利用预置模版，创建多模态数据处理工作流，对数据卷中视频内容进行总结生成文本。

🔗 前提条件

在公有云官网已[申请DataBuilder白名单](#)，推荐使用具有 DataBuilderFullControlAccessPolicy权限的 DataBuilder 用户。

🔗 准备工作

1. 开通百度智能云账号

使用手机号，[开通百度智能云账号](#)。

2. 开通DataBuilder

本教程以 保定 地域为例，介绍DataBuilder快速入门，您需要[登录DataBuilder管理控制台](#)，切换至 保定 地域，查看该地域是否开通DataBuilder，如未开通请先开通。

3. 在IAM添加子用户

单击多用户访问控制



创建子用户并授予以下3个策略：DataBuilderFullControlAccessPolicy、IAMReadAccessPolicy、BOSLISTANDReadAccessPolicy。

4. 登录子用户

利用子用户[登录百度智能云](#)，并进行下一步操作。

5. 创建BOS Bucket

为多模态数据、产品任务运行的日志等指定存储位置，登录[BOS管理控制台](#)，在 保定 地域，[创建Bucket](#)用于DataBuilder数据存储的。

6. 创建私有网络

为常驻计算资源绑定VPC资源，登录[私有网络VPC控制台](#)，在 保定 地域，[创建私有网络](#)。并[创建私有网络子网](#)。

🔗 操作步骤

🔗 第一步 创建空间

创建用于多模态数据管理及处理的工作空间，指定空间作业及日志的存储位置（已经创建的BOS Bucket）。

创建工作空间

×

空间名称 *

请输入空间中文名称0/50

存储配置 * ?

bos://

如果没有bucket, 请前往 [bos创建](#)

描述

请输入空间描述0/200

取消

确定

第二步 创建元存储

创建该租户下唯一的元存储，并指定全局数据默认的存储位置（已经创建的BOS Bucket）。

元数据/元存储

元存储名称 *

请输入元存储名称0/50

托管数据存储位置 * ?

bos://

如果没有bucket, 请前往 [bos创建](#)

管理员

请选择管理员

创建

取消

第三步 创建计算资源

进入工作空间，创建Ray计算资源，选择已经创建的VPC网络，指定GPU规格。

地域*

华北 - 保定

网络及可用区

网络*

default | 172.16.0.0/16

▼

C

如需创建新的私有网络，您可以到[私有网络](#)创建

可用区与子网*

可用区B

▼

系统预定义子网B

▼

C

如需创建新的子网，您可以到[私有网络-子网](#)创建

集群信息

集群名称*

请输入

0/64

由大小写字母、数字、中文以及-/_/特殊字符，必须以字母或者中文开头，长度1-64

集群引擎

Ray

镜像版本

V1.0 (Ray2.35.0, Python 3.9.19)

节点配置

节点类型

☐ CPU

☒ GPU

节点规格*

▼

购买数量*

-

2

+

第四步 创建数据卷

1. 创建数据卷

进入工作空间，创建两个数据卷，一个用于源端数据存储，另一个用于目标端数据存储，在元数据下逐级创建目录 -> 库 -> 数据卷。

2. 上传视频

在已经创建的源端数据存储数据卷下，上传用户本地的一个小视频。



🔗 第五步 数据处理

1. 创建视频生文工作流

利用产品预置【视频处理工作流】模版，创建工作流，修改 source 输入端节点路径为上传视频的数据卷，修改 sink 输出端节点路径为目标端数据卷，修改计算资源 clusterId 参数为已创建的计算资源ID，修改完成后运行任务。

2. 查看处理结果

待任务运行成功后，在目标端存储数据卷查看视频数据解析结果。

操作指南

工作空间

🔗 前提条件

- 已开通百度智能云 DataBuilder 服务。
- 拥有系统管理员策略 DataBuilderFullControlAccessPolicy。

🔗 创建工作空间

1. 登录百度智能云 DataBuilder 控制台，在侧边导航选中工作空间模块。
2. 在工作空间页面右上角单击**创建工作空间**按钮，在弹出页面填写创建工作空间配置项。
3. 填写完毕后单击确定按钮，工作空间创建完成。

表一 创建工作空间配置项说明

配置项名称	说明
空间名称	填写空间名称。支持中文、英文、数字、中划线、下划线，长度为1~50个字符，不能以中划线、下划线开头。注意：空间名称不可重复。
存储配置	选择存储配置，如果没有 bucket，请单击 bos创建去创建新的bucket。云存储位置用于存储工作空间内的资产，例如：日志、作业、任务执行记录等。
描述	输入空间描述，不超过200字。

🔗 工作空间列表

1. 创建成功的工作空间以列表形式展示，在列表中可以查看工作空间的具体信息。包括空间名称、状态、存储地址、描述、创建人和创建时间。
2. 列表支持通过状态（运行中或不可用）进行筛选。
3. 列表操作：
 - 打开：单击打开按钮，进入工作空间内。在工作空间内可以进行工作区的创建、元数据的创建等。
 - 详情：单击详情按钮，可查看工作空间的详细信息，包括空间名称、存储路径等。
 - 删除：单击删除按钮，删除工作空间。注意：工作空间删除后无法恢复，请您谨慎操作。

🔗 工作空间详情

1. 在工作空间列表页面单击工作空间名称或列表操作列，进入工作空间详情，查看工作空间详细信息。通过单击tab页可以切换查看空间详情或权限信息。
2. 工作空间详细信息包括空间名称、存储路径（BOS存储路径，单击复制按钮可进行复制操作）、空间状态（运行中或不可用）、元存储、创建人（创建工作空间的用户名称）、创建时间（创建工作空间的时间和描述）、描述（支持在详情中修改）。
3. **打开工作空间**：详情页面右上角可以打开工作空间，进入工作空间内部。
4. **删除工作空间**：右上角删除工作空间，注意：工作空间删除后无法恢复，请您谨慎操作。

🔗 工作空间权限

1. 在工作空间详情页面通过切换tab页进入权限页面，支持通过名称搜索用户。
2. 拥有工作空间权限的用户以列表形式展示，列表可通过用户类型进行筛选。
3. **添加成员**：单击列表上方**添加成员**按钮，在用户/用户组栏输入名称进行搜索，选择用户/用户组后划分角色（管理员/普通用户），完成后单击确定，成员添加成功。
4. **删除成员**：在添加成员页面单击删除按钮，系统进行再次确认，确认无误后单击确定，成员删除完成。

元存储

🔗 前提条件

- 已开通百度智能云 DataBuilder 服务。
- 拥有系统管理员策略 DataBuilderFullControlAccessPolicy。

🔗 创建元存储

1. 登录DataBuilder控制台，侧边导航选择**元数据**。

2. 填写相关配置项。注意：在没有元存储的情况下，才支持创建元存储。

表一 创建元存储配置项说明

配置项名称	说明
名称	填写元存储名称。支持中文、英文、数字、中划线、下划线，不能以中划线、下划线开头，长度为1~50个字符。
存储配置	选择存储配置，用于存储本租户下Catalog、Schema、Table、Volume所托管的数据。路径长度限制：1~500字节。
描述	输入元存储描述。字数在200字以内。

3. **编辑元存储**：创建完成后元存储信息会展示在当前页面，可以对当前元存储管理员进行修改。
4. **删除元存储**：在元存储页面单击删除按钮，系统弹出确认页面，确认无误后单击删除，元存储删除完成。注意：元存储删除后无法恢复，请谨慎操作。

工作区

 前提条件

- 具备空间管理员或空间普通用户角色

 创建工作区文件夹

1. 登录百度智能云 DataBuilder控制台，在选中的工作空间操作列单击打开按钮，进入空间内。
2. 侧边导航选择工作区，在工作区右侧上方单击**创建**按钮，创建文件夹。
3. 填写文件夹名称，名称支持中文、英文、数字、中划线、下划线、.特殊字符，不能以中划线、下划线开头，长度为1~128个字符。填写完毕后单击确定，文件夹创建完毕。

 工作区列表

工作区以目录形式展示在工作空间，工作区内容以列表展示。支持按照名称进行搜索文件或文件夹。

“全部”目录下展示所有用户工作区，单击具体用户名称可查看用户工作区文件夹。

工作区列表右上方更多图标支持对**导入文件**和**存储路径**的快捷操作。

表一 工作区列表项说明

列表项名称	说明
名称	文件或文件夹的名称，单击文件夹名称可进入文件夹。
类型	类型分为文件或文件夹。
创建人	上传文件或创建文件夹的用户。
创建时间	创建文件夹或上传文件的时间。
操作项	单击不同操作项，执行不同操作。 在新窗口打开：单击在新窗口打开按钮，在新的网页页签打开当前内容，保留顶部、左侧导航。 新建文件夹：单击新建文件夹按钮，在当前文件夹新建一个子文件夹。 导入文件：通过单击操作项导入文件按钮可上传文件到该文件夹。单次最多可上传100个文件，单个文件不超过100M，文件总大小不超过500M。 复制：单击复制按钮，输入新的文件或文件夹的名称，名称不可重复。系统会复制一个相同的文件或文件夹展示在列表。 复制路径：单击复制按钮可以复制当前文件/文件夹的完整路径+当前文件名。 重命名：单击重命名按钮，系统弹出重命名弹窗，修改文件的名称。 移动：单击移动按钮，系统弹出移动窗口，选择移动的位置后确定。 删除：单击删除按钮，可删除当前文件或文件夹。注意：文件或文件夹被删除后，会被移动至回收站，默认保留30天后，请谨慎操作。

🔗 回收站

删除后的文件或文件夹会被移入回收站，默认保存30天。回收站支持按照文件或文件夹名称进行搜索。

- 恢复：若要恢复文件或文件夹，需在30天内进入回收站，在操作列中单击恢复按钮，选择恢复的目标目录后，即可完成恢复操作。
- 永久删除：单击永久删除按钮，可对文件或文件夹进行永久性删除。永久删除的文件或文件夹数据将被清空，不可恢复，请谨慎操作。

元数据

DataBuilder元数据，支持用户组织和管理AI和数据资产。元数据支持三层结构：目录-库-数据表/数据卷/算子/模型，目录是第一层，目录之下可以创建库，库中可以包含数据表、数据卷、算子和模型等。本文将介绍如何使用元数据进行数据管理

🔗 前提条件

- 已配置好 DataBuilder 元存储。
- 已创建 DataBuilder 工作空间。

🔗 使用目录

目录为元数据为第一层，支持在目录中创建库。

🔗 创建目录

1. 登录 DataBuilder 控制台，在选中的工作空间操作列单击 **打开** 按钮，进入空间内。
2. 侧边导航选择元数据，进入元数据后单击添加按钮新建目录。
3. 填写新建目录配置项，填写完成后单击确定，目录创建完成。
4. 创建完成的目录在目录树展示，单击目录名称可在右侧查看目录详细信息。

表一 新建目录配置项说明

配置项名称	说明
目录名称	填写目录名称，在同一个元存储中不可重复。必填项，支持输入大小写字母、数字、下划线，1-64字符。注意：在创建目录时若提示“存在同名目录”时，目录无法创建成功。
描述	输入描述。支持输入大小写字母、数字、下划线等各种特殊字符，0-150字符。

🔗 查看目录

在目录树单击目录名称可在右侧查看目录详细信息，详情处可以对描述进行修改。基本信息包括存储路径、创建时间和创建人等。

每个工作空间预置一个system 目录，提供高质量数据和AI资产，支持用户使用。

表二 目录详情说明

基本信息名称	说明
目录名称	目录的名称，用来识别目录的唯一性。
存储路径	BOS存储路径。
创建时间	创建目录的时间。
创建人	创建目录的用户的用户名。
修改时间	最后一次修改目录时间。
最近修改人	最后一次修改目录的用户。
数据表个数	数据表的数量。
数据卷个数	数据卷的数量。
算子个数	算子的数量。

🔗 修改目录名称

目录的名称支持修改，在目录详情界面右上角更多按钮展开后单击**重命名目录**按钮，输入新的目录名称后确定即可。

注意：

- 新修改的目录名称同样不可重复。
- system目录为系统内置，不支持修改名称。

🔗 删除目录

目录支持删除，但是目录只有在没有数据表或数据卷的情况可以删除。

在目录详情界面右上角更多按钮展开后单击 **删除目录** 按钮，系统会弹出提示框，确认无误后单击确定，目录删除成功。

注意：

- 删除后的目录无法恢复，请您谨慎操作。
- system目录为系统内置，不支持删除。

🔗 库

库为元数据中第二层，支持在目录中创建库，在库中创建数据表和数据卷。

🔗 创建库

1. 登录 DataBuilder 控制台，在选中的工作空间操作列单击**打开按钮**，进入空间内。
2. 侧边导航选择元数据，进入元数据后在目录树上单击要创建库的目录名称，在详情页面右上方单击创建库按钮。
3. 在创建库弹窗填写相关配置项后单击确定按钮，库创建完成。

表三 创建库配置项说明

配置项名称	说明
库名称	输入库名称，必填。支持输入大小写字母、数字、下划线，1-64字符，库名称在目录中不可重复。
描述	输入库描述。支持输入大小写字母、数字、下划线等各种特殊字符，0-150字符。

🔗 查看库

在目录树单击库名称，可在右侧查看库详细信息，详情处可以对描述进行修改。基本信息包括存储路径、创建时间和创建人等。

每个目录创建之后默认带有一个名称为default的库，用户可以对default库进行修改和删除。

表四 查看库信息说明

基本信息名称	说明
目录名称	目录的名称，用来识别目录的唯一性。
库名称	库的名称，用来识别库的唯一性。
存储路径	BOS存储路径。
创建时间	创建库的时间。
创建人	创建库的用户名。
修改时间	最新一次库的修改时间。
最近修改人	最新一次库的修改用户名。
数据表个数	数据表的数量。
数据卷个数	数据卷的数量。
算子个数	算子数量。

🔗 修改库名称

库的名称支持修改，在库详情界面右上角更多按钮展开后单击 **重命名库** 按钮，输入新的库名称后确定即可。

注意：

- 新名称在目录中不可重复。
- system 目录中的库不可修改名称。

🔗 删除库

库支持删除。在库详情界面右上角更多按钮展开后单击**删除库** 按钮，系统会弹出提示框，确认无误后单击确定，库删除成功。

注意：

- 如果库中有数据表、数据卷或算子，则库不可删除。
- 删除后的库无法恢复，请您谨慎操作。

🔗 数据卷

数据卷为元数据中第三层，支持在库中创建数据卷。用数据卷来管理非结构化数据。

🔗 创建数据卷

1. 登录 DataBuilder 控制台，在选中的工作空间操作列单击 **打开** 按钮，进入工作空间。
2. 侧边导航选择元数据，进入元数据后在目录树上单击要创建数据卷的库名称，在详情页面右上方单击立即创建 > 创建数据卷按钮。
3. 系统跳出创建数据卷页面，在创建页面填写对应配置项，填写完成确定。
4. 创建完成的数据卷在左侧目录树展示，单击名称可查看数据卷详情。

表五 创建数据卷配置项说明

配置项名称	说明
数据卷名称	输入数据卷名称，必填。支持输入 大小写字母、数字、下划线，1-64字符。 注意：数据卷名称在库中不可重复。
数据卷类型	两种类型可选，Managed 数据卷 或 External 数据卷。 Managed 数据卷 会在元存储对应的路径下统一管理；External 数据卷为挂载形式，支持用户将其他BOS路径上的数据挂载到Databuilder。
Bos路径	在选择 External 数据卷的需要填写对应的BOS存储路径。 注意：BOS路径需要和工作空间在同一个地域。
AccessKey	在选择 External 数据卷的需要填写 AccessKey。
SecretKey	在选择 External 数据卷的需要填写 SecretKey 。
描述	填写数据卷描述。支持输入大小写字母、数字、下划线等各种特殊字符，0-150字符。

🔗 上传文件到数据卷

1. 在数据卷页面右上方单击**上传数据到数据卷** 按钮，将文件夹或多个文件拖到对应区域或者单击点击上传按钮后，上传对应文件。
2. 在上传过程中，在页面右下方可查看文件上传情况和进度。
3. 上传成功的文件以列表形式展示，可以对文件进行复制路径、删除或下载操作。
 - 复制路径：单击按钮复制文件在 DataBuilder 的文件路径。
 - 删除：单击删除按钮系统再次提示，确认无误后单击确认按钮，数据卷删除成功。注意：删除的数据卷无法恢复，请您谨慎操作。
 - 下载：单击下载按钮后文件可以通过浏览器下载到本地。

注意：

- 目标目录下如果存在同名文件，将被新上传的文件覆盖。
- 每次上传的总文件大小不超过500MB，单个文件不超过100MB；每次最多上传100个文件。

查看数据卷

左侧目录树选择数据卷后，右侧展示数据卷的详细内容，可以通过单击切换查看概览或详情。

概览

在概览页面可以查看数据卷描述、数据卷的文件路径和上传文件列表。单击描述旁边修改按钮可以对数据卷的描述进行修改。

详情

详情页面可查看关于数据卷的详细信息，包括创建人、创建信息等。

修改数据卷名称

数据卷的名称支持修改，在数据卷详情界面右上角更多按钮展开后单击**重命名数据卷**按钮，输入新的数据卷名称后确定即可。
注意：新名称在库中不可重复。

删除数据卷

在数据卷详情界面右上角更多按钮展开后单击**删除数据卷**按钮，系统会弹出提示框，确认无误后单击确定，数据卷删除成功。

注意：

- 删除后的数据卷无法恢复，请您谨慎操作。

算子

在元数据中可以查看内置算子和在工作流中使用内置算子。

查看内置算子

在system 目录中内置多个算子，单击名为system的目录展开可查看当前内置算子。单击具体算子名称可查看对应的算子概览和算子详情。

概览 通过tab切换可查看算子概览信息。包括描述、使用说明和版本列表。

- 描述：关于函数能力的介绍。
- 使用说明：包括函数详细介绍、使用方式说明和数据处理示例。
- 版本列表：介绍算子的版本信息，以列表形式展示。列表可查看版本的详细信息，包括版本号、ID和创建时间等，单击版本号可展开查看版本详情。

表六 版本列表项说明

列表项名称	说明
版本号	对应算子的版本号。单击版本号可展开查看版本详情。
ID	算子唯一ID。
创建时间	创建版本的时间点。
修改时间	更新该版本算子的时间。
版本描述	对于版本的描述信息。支持大小写字母、数字、下划线等各种特殊字符，0-150字符。默认查看少量字符，展开后显示全部描述。

表七 版本详情说明

详情项名称	说明
版本号	对应算子的版本号。
版本ID	版本的唯一ID。
版本描述	当前算子版本的描述。展开可查看全部字符。
创建时间	当前版本创建时间。
创建人	创建算子的用户。
修改时间	更新该版本算子的时间。
最近修改人	修改算子的用户。
代码语言	python。
算子代码路径	算子代码路径。
算子类型	类型包含extract、transform、filter、dedup、embedding、others。
运行环境	当前算子的运行环境。
输入参数	包括3个字段，key、类型和value。key默认为input，不同算子有不同的类型和value。
输出参数	包括3个字段，key、类型和value。key默认为output，不同算子有不同的类型和value。
运行参数	包括4个字段，key、默认值、类型和required。默认值 = 5；类型：Interger；required = true。
支持引擎	支持运行该算子的引擎，包括Ray、Spark。
资源配置	支持运行该算子的资源，包括GPU、CPU。

表八 算子说明

算子名称	说明
databuilder_location_loader	加载数据源路径的数据，目前支持BOS和数据卷方式。BOS路径格式为： bos://bucket_name/path/to/data；数据卷路径格式为：/数据卷s/test/default/data_vol/。
exporter	将数据输出到指定路径，目前支持BOS和数据卷方式。BOS路径格式为： bos://bucket_name/path/to/data；数据卷路径格式为：/数据卷s/test/default/data_vol/。
video_info_extractor	提取视频信息（帧数、高度、宽度、长宽比、帧率、帧率和像素值）。
column_filter	过滤指定数值列中满足条件的行，设置过滤数值列名、表达式运算符、阈值。eg:过滤帧数小于100的视频，filtercol值为num_frames、filter算子值为<、filter_value值为100。
video_scene_detector	检测视频场景边界，输出边界时间列表。
video_scene_cutter	根据场景时间戳分割视频。
aesthetic_evaluation	对视频的美学质量进行打分，采用improved aesthetic predictor模型，其中美学质量越高分数越高。
optical_flow_evaluation	对视频的运动进行打分，采用UniMatch模型，计算视频的光流分数，其中画面运动越多分数越高。
ocr_evaluation	对视频中的文字进行检测，采用DBNet++模型，计算OCR分数。文字越密集的视频，分数越高。
pllava_caption	采用pllava模型对视频进行caption，生成文字描述视频的内容。
matching_score	采用CLIP计算文本和视频之间的匹配分数。
video_scanner	给定一个bos视频路径，可以生成这个路径下所有的视频信息，包括文件名和bos地址。

详情

通过tab切换可查看算子详细信息。包括目录名称、库名称和创建时间等。

表九 算子详情信息说明

基本信息项名称	说明
目录名称	当前算子的目录名称。
库名称	当前算子的库名称。
算子名称	当前算子的算子名称。
算子别名	算子其他名称。
使用说明	当前算子概览的使用说明。
创建时间	创建当前算子的时间。
创建人	创建算子的用户。
修改时间	所有版本的最后修改时间。
最近修改人	最后修改版本的用户。
最新版本 ID	算子最新版本的ID名。
最新版本号	算子最新版本号名。

计算资源

DataBuilder支持计算资源创建和管理功能，计算资源可用于工作流的调试和运行。下面将详细说明计算资源功能。

前提条件

- 具备空间管理员权限。

创建计算集群

- 登录DataBuilder控制台，在选中的工作空间操作列单击 打开 按钮，进入空间内。
- 侧边导航选择**计算资源**，进入计算资源后在页面右上方单击**创建集群**按钮。
- 进入创建集群页面，填写相关配置项。

表一 配置项说明

配置项名称	说明
付费方式	付费方式有按量付费和包年包月，目前DataBuilder仅支持按量付费。 - 按量付费：也称后付费。先使用资源，后续生成账单进行结算。 - 包年包月：也称预付费，在新建集群时先支付费用，后使用资源。
区域	显示当前区域。
网络	下拉选择已有私有网络（VPC）。也可单击 私有网络 去创建新的私有网络。
可用区与子网	下拉选择可用区和子网。可单击 私有网络-子网 创建新的子网。
集群名称	输入集群名称。由大小写字母、数字、中文以及-/_/.特殊字符，必须以字母或者中文开头，长度1-64。
引擎类型	类型：Ray
镜像版本	版本：V1.0 (Ray2.35.0, Python 3.9.19)
资源类型	两种资源类型：GPU和CPU。不同的资源类型对应不同的节点规格。
节点规格	选择集群节点规格。
节点数量	设置节点数量，最少1个，最大50个。

4. 填写完成后，可以通过右侧集群明细确认集群配置信息，确认无误后单击立即创建，集群创建完成。

查看计算集群

创建成功的集群以列表形式展示。在列表可以查看集群的ID、集群名称、运行状态、引擎类型、创建时间和付费方式。列表可以通过状态进行筛选。

删除计算集群

集群支持删除，单击集群列表操作列删除按钮，系统弹出确认页确认无误后单击确定，集群删除成功。

注意：

- 集群删除后不可恢复，请您谨慎操作。
- 集群删除会影响关联的工作流任务的运行，在线上业务中，请确认工作流运行状态后再进行集群删除操作。

工作流

前提条件

- 拥有工作流管理员权限

创建工作流

- 登录百度 DataBuilder控制台，在选中的工作空间操作列单击打开按钮，进入空间内。
- 侧边导航选择**工作流**，在工作流列表右上方单击**创建工作流**。
- 填写新建空白工作流相关配置项后确定，工作流创建完成。

表一 创建工作流配置项说明

配置项名称	配置项说明
工作流名称	填写工作流名称，名称不可重复。支持中文、英文、数字、中划线(-)、下划线(_)，不超过256字符。
描述	填写工作流描述，不超过500字符。

4. 创建成功的工作流以列表形式展示，单击名称可进入工作流详情。

表二 工作流列表操作项说明

操作项名称	说明
运行	单击操作列运行按钮，执行工作流作业并且生成一条工作流作业运行记录。
编辑	单击操作列编辑按钮跳转至工作流详情进行 Json 作业的编辑。
运行记录	单击操作列运行记录按钮，进入列表查看运行记录。
导出	单击操作列导出按钮，系统会根据工作流生成一份Json文件。
复制	单击操作列复制按钮，系统生成一个新的工作流作业在列表页，详情的工作流作业内容为复制的工作流的json。
删除	单击删除按钮，删除工作流。注意：正在运行的工作流不可删除，工作流删除后运行数据会被清空，请谨慎删除。

🔗 预置工作流模板

1. 支持通过模板创建工作流作业。预置工作流模版在工作流管理中，通过切换tab进入预置作业模板页面。
2. 在模板处单击使用模板按钮，进入工作流编辑。工作流名称为模板名称，可以通过页面右边详情按钮修改工作流名称，修改完成后在右边工作流编辑区上方单击保存按钮。

🔗 工作流详情

通过单击工作流列表中工作流名称可以进入工作流详情，在工作流详情编辑区域可以进行 Json 作业。

表二 工作流详情页面操作说明

操作项	操作项说明
编辑	在详情页面单击编辑按钮进行 Json 作业的编辑。
导入工作流	编辑状态下，可以进行工作流的导入。
	单击工作流导入按钮，在导入页面可以通过选择模板导入或者选择自定义导入，自定义导入需单击上传文件按钮进行文件上传。选择导入完成后确定。
	注意：导入新工作流后，原有的工作流将会被清空，请谨慎操作。
运行	编辑结束后，单击运行按钮，执行工作流作业并且生成一条工作流作业运行记录。
查看工作流信息	在详情页面右边单击作业基本信息图标，可展开查看工作流信息并对工作流名称和描述进行编辑。
删除	单击详情页面右上角删除工作流按钮，可删除没有正在运行的工作流作业。
	注意：工作流删除后，相关运行数据将被清空，请谨慎操作。
可视化预览	可视化预览在详情页面的右侧，单击可视化预览按钮，可预览当前工作流作业。

🔗 工作流运行记录

单击运行按钮，会执行工作流作业并生成一条工作流作业运行记录。在详情页通过tab切换可查看运行记录，运行记录以列表形式展示。支持按照状态进行筛选，还可以通过开始时间或结束时间进行排序。

运行记录

运行记录用于展示本工作空间下全部工作流的运行记录。

🔗 运行记录列表

1. 登录DataBuilder控制台，在选中的工作空间操作列单击打开按钮，进入空间内。
2. 侧边导航选择**运行记录**，查看运行记录列表。
3. 运行工作流作业生成的记录以列表形式展示，在列表中可查看所有工作流运行的详细信息。
4. 列表支持按照运行id、工作流、用户和时间进行筛选。还可以通过开始或结束时间进行列表排序。

表一 列表项说明

列表项名称	说明
运行ID	工作流每次运行的唯一标识，用于区分和追踪特定运行实例。
工作流名称	工作流的名称，点击可进入工作流编辑页面。
工作流ID	工作流的唯一编码。
状态	状态分别为运行中、成功、失败等。
调度时间	工作流按计划应启动运行的时间。
开始时间	工作流实际启动执行的时间点。
结束时间	工作流运行终止的时间点。
运行时长	工作流运行从开始到结束所经历的时长。
运行用户	执行该工作流运行操作的用户。
操作	单击操作列不同按钮，执行不同操作。 查看：单击查看按钮，查看运行记录详情。在详情页面单击具体运行记录可以查看运行任务详细信息，右侧可选择单击查看任务参数、任务结果和任务日志。 重跑：在运行成功和失败时支持重跑任务，可选择全流程重跑或从失败节点重跑，任务状态为成功的时候只能选择全流程跑。 停止：停止当前运行任务，只能停止正在运行中的工作流。 暂停：暂停当前运行任务，只能对运行中的任务暂停。 恢复：恢复运行任务，只对暂停中的任务恢复。 删除：删除当前运行记录，不可删除正在运行的任务。注意：任务删除后数据将清空，请谨慎操作。

常见问题

🔗 一、工作空间

1. 空间管理员是否可以创建空间？

否，工作空间由系统管理员创建，并且空间管理员由系统管理员指定，空间管理员主要向当前空间添加空间普通用户。

🔗 二、元数据

1. DataBuilder目前支持几种湖格式？分别是什么

目前支持1种，湖格式为Iceberg。

2. 元数据和工作空间是什么关系？

元数据和工作空间是绑定关系，1个Catalog可以绑定到多个工作空间。

☞ 三、计算资源

1. 计算资源支持哪几种类型？

目前支持常驻资源，后续会逐步支持Serverless资源。

☞ 四、工作流

1. 工作流可以编排哪些类型任务？

Ray任务、算子，其中Ray任务目前主要是离线上传的工程文件。