

大数据智能体工作台 DataBuddy

数据治理

产品文档



【版权声明】

©2013–2026 腾讯云版权所有

本文档著作权归腾讯云单独所有，未经腾讯云事先书面许可，任何主体不得以任何形式复制、修改、抄袭、传播全部或部分本文档内容。

【商标声明】

及其他腾讯云服务相关的商标均为腾讯集团下的相关公司主体所有。另外，本文档涉及的第三方主体的商标，依法由权利人所有。

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

文档目录

数据治理

数据目录概述

- 统一元数据管理架构
- 三级命名空间 (Catalog / Schema / Object)
- 结构化数据 Catalog (表、视图、Schema 管理)
- 非结构化数据 Catalog (文件、Volume 管理)
- 模型 Catalog (ML 模型注册与元数据)
- 外部数据管理与基于数据源创建外部数据目录
- 数据血缘

数据发现概述

- 全局数据搜索
- 分类筛选

数据安全概述

- 数据分类分级
- 列级权限控制
- 行级权限控制 (Row Filter)
- 数据变更历史

数据质量概述

- 创建质量任务
- 质量规则配置指南
- 数据剖析 (Data Profiling)
- 异常检测 (Anomaly Detection)

语义模型概述

- 什么是语义模型
- 创建语义模型
- 定义指标与维度

治理标签概述

- 创建与管理标签体系
- 为数据资产打标签
- 基于标签的策略控制

数据治理

数据目录概述

统一元数据管理架构

最近更新时间：2026-10-09 16:26:10

概述

DataBuddy 的数据目录（Catalog）采用统一元数据管理架构，将数据湖中所有可被消费的资产实体——表、视图、卷（Volume）、模型，以及来自外部数据源的元数据——集中注册到同一个目录服务中。数据工程师可以通过单一入口发现所有数据资产，并查看它们之间的上下游血缘关系。

本文介绍数据目录的整体架构、关键组件，以及它与传统数据源管理方式的差异。

架构总览

数据目录由三层组成：

层级	角色	说明
MetaLake（元数据湖）	元数据存储	地域级元数据服务实例，每个租户在每个地域默认拥有一份 MetaLake，所有 Catalog 都注册在 MetaLake 之下
Catalog	命名空间顶层	MetaLake 下可创建多个 Catalog，每个 Catalog 拥有独立的存储路径与类型，跨工作空间可见
Schema / 资产	命名空间次层与对象	Catalog 下按 Schema 组织数据库；Schema 下挂载表、视图、卷、模型等资产对象

存储路径默认遵循下述规则，存放在腾讯云 COS：

```
cos://<bucket>/root/workspaces/<catalogN>/...
```

关键组件

- **MetaLake（元数据湖）**：地域级、租户级的元数据中枢。它维护所有 Catalog 的注册信息、Schema 与资产的元数据、权限策略、血缘关系。同一租户在每个地域只有一份 MetaLake。
- **Catalog**：可创建的元数据命名空间。Catalog 拥有：
 - **类型**：决定 Catalog 下能创建哪些资产，目前支持 `Table`（含视图）、`Volume`、`Model` 三种。
 - **存储路径**：Catalog 拥有的 COS 存储路径前缀，与计算资源解耦。

- **链接信息**：Catalog 可与外部数据源连接（如基于 Connection 创建 Foreign Catalog）。
- **Schema**：Catalog 下的命名空间，相当于传统数据库中的 Database。每个 Catalog 创建后会自动包含 `information_schema` 与 `default` 两个 Schema。
- **资产对象**：Schema 下挂载的具体资源，包括表（Table）、视图（View）、卷（Volume）、模型（Model）、函数（Function）等。每个资产由 Catalog 类型决定可承载的对象种类。

内置 Catalog

DataBuddy 在用户首次进入数据目录时，自动创建1个内置 Catalog：

内置 Catalog	用途	是否可改名 / 删除
<code>default</code>	默认目录，供首次使用时存放测试与样例数据	不可改名、不可删除

设计原则

DataBuddy 数据目录在设计上遵循以下基本原则：

1. **一个 MetaLake 多 Catalog**：同一个 MetaLake 实例下可以创建多个 Catalog，Catalog 跨工作空间可见。
2. **每个地域一份 MetaLake**：每个租户在每个地域默认只有一份 MetaLake，作为元数据的唯一事实来源；MetaLake 跨地域之间完全隔离，同名 Catalog 在不同地域会被视为不同对象。
3. **按存储划分 Catalog**：MetaLake 按 COS 存储路径进行 Catalog 划分，每个 Catalog 拥有独立的存储前缀。
4. **每个 Catalog 有明确类型**：Catalog 在创建时必须指定类型，决定其下可承载的资产种类（Table、Volume、Model）。
5. **Catalog 自带链接信息**：Catalog 可承载链接信息用于连接到外部资源（如 Connection、External Location）。
6. **Catalog 与计算资源解耦**：Catalog 仅描述元数据与存储，与计算资源无关。同一个 Catalog 可以被任意有权限的计算资源访问。
7. **统一存储路径规则**：存储路径形如 `cos://<bucket>/root/workspaces/<catalogN>/...`，由系统自动维护。

与计算引擎的关系

Catalog 仅描述元数据与存储，**不绑定具体的计算资源**。当您在 Studio 中执行 SQL、运行 Notebook，或在 Workflow 任务中读写数据时，由计算资源（DLC SQL、Spark、Ray 等）通过统一的元数据接口直接访问 Catalog：

- 元数据操作（建表、改表、删表等）通过统一接口下发，所有引擎看到的元数据一致。
- 数据读写遵循 Catalog 上的权限策略；越权访问会被引擎层拦截。
- 同一个 Catalog 可以被多个计算资源、多个工作空间同时访问（受权限和工作空间策略控制）。

工作空间与 Catalog 的关系

- Catalog 注册在 MetaLake 上，**跨工作空间可见**：在工作空间 A 中创建的 Catalog，在同一地域的其他工作空间中也能看到。
- 工作空间通过 **权限策略** 决定哪些用户在当前空间下可以访问哪些 Catalog 与资产。
- 跨工作空间的隔离主要由工作空间、用户、资源对象三元组的访问控制实现；同一 Catalog 在不同工作空间的访问范围可独立配置。

使用限制

限制项	说明
存储后端	当前仅支持腾讯云 COS 作为底层存储
自定义 COS 桶	Catalog 创建时使用平台默认存储空间，暂不支持用户自定义 COS Bucket（系统会按 <code>default</code> 存储空间分配）
MetaLake 数量	每个租户在每个地域固定一份 MetaLake，不支持创建多份
Catalog 数量上限	每个 MetaLake 下 Catalog 数量上限为 100，您可以申请扩容
资产类型范围	当前支持表、视图、卷、模型；函数、API、应用类型计划在后续版本支持

常见问题

- Q：为什么 Catalog 的列表跨工作空间可见，而表的访问还是受工作空间限制？**
- A：Catalog 在 MetaLake 上是全局可见的命名空间，便于跨空间共享元数据；但表的实际读写受 Catalog / Schema / 资产层的权限策略与工作空间策略双重控制。元数据可见 ≠ 数据可读。
- Q：Catalog 与 数据源 是不是同一个概念？**
- A：不是。在 DataBuddy 中：
- **Catalog**：用于承载元数据与底层存储的目录，存储 DataBuddy 内部托管的数据（表、卷、模型）。
 - **数据源（Connection）**：用于连接外部数据库（如 MySQL、PostgreSQL）以便接入或镜像数据。基于 Connection 创建的外部 Catalog 称为 Foreign Catalog，详见 [外部数据管理](#)。
- Q：内置的 `system` Catalog 里能看到什么？**
- A：`system` Catalog 用于存放平台侧自动维护的系统级元数据，由平台统一管理；普通用户默认对其只有只读访问权限。

相关文档

- [数据目录概述](#)
- [三级命名空间](#)
- [结构化数据 Catalog](#)

- [外部数据管理](#)
- [成员与权限管理](#)

三级命名空间（Catalog / Schema / Object）

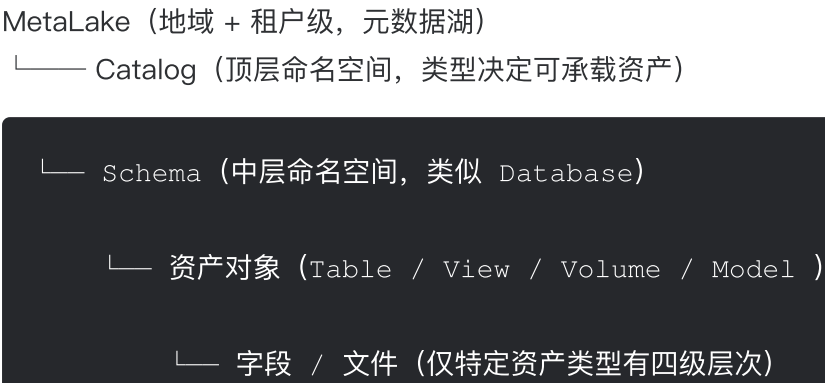
最近更新时间：2026-10-09 16:26:14

概述

DataBuddy 数据目录使用 **三级命名空间** 来组织和定位所有数据资产。任意一张表、一个视图、一个卷或一个模型，都通过形如 `catalog.schema.object` 的三段式路径唯一标识。三级命名空间为跨工作空间共享、细粒度授权和资产检索提供统一的引用方式。

本文介绍三级命名空间的层级关系、命名规则、引用方式以及在 SQL 与 Notebook 中的使用方法。

命名空间结构



各层级职责

层级	类比	职责
Catalog	数据库实例 / 数据仓库	顶层命名空间，决定底层存储位置与资产类型
Schema	数据库（Database）	按业务主题、项目或环境划分对象；相当于 SQL 中的 DATABASE
对象	表、视图、卷、模型	实际承载数据或元数据的资产

路径与引用格式

SQL 中的三段式引用

在 Studio 中的 SQL 文件、Notebook 的 Spark SQL 单元格、Workflow SQL 任务中，您可以使用三段式路径引用任意 Catalog 下的对象：

卷（Volume）路径引用

卷（Volume）是非结构化数据的容器，对其下文件的访问采用以 `/Volumes/` 为前缀的路径：

```
/Volumes/<catalog>/<schema>/<volume>/<directory>/<file>
```

例如：

```
/Volumes/ops_catalog/raw_logs/access_log/2026-06-06/access.log
```

模型 URI 引用

模型在三级命名空间之上还有版本与别名两层标识。完整的模型 URI 形如：

```
models:/<catalog>.<schema>.<model_name>/<version>
models:/<catalog>.<schema>.<model_name>@<alias>
```

例如：

```
models:/ml_catalog.recsys.deepfm_model/3
models:/ml_catalog.recsys.deepfm_model@champion
```

命名规则

命名约束

元素	规则
Catalog 名称	在同一 MetaLake 内全局唯一； <code>default</code> 为系统保留名，不可创建或修改
Schema 名称	在同一 Catalog 内唯一； <code>information_schema</code> 与 <code>default</code> 为系统在创建 Catalog 时自动创建的 Schema
对象名称（表 / 视图 / 卷 / 模型）	在同一 Schema 内唯一
字符集	详细命名约束（长度、合法字符、大小写敏感性）由底层引擎决定

 **注意：**

`default` 是系统保留 Catalog 名，您无法创建同名 Catalog；`information_schema` 与 `default` 是 Catalog 创建后自动生成的 Schema 名，您也无法手工创建或重命名。

自动初始化的对象

- 首次进入数据目录：系统自动创建 `default` 和 `system` 两个内置 Catalog。
- 创建 Catalog 后：系统自动创建 `information_schema` 与 `default` 两个 Schema。`default` Schema 的内容初始为空。

在 Studio 中浏览三级命名空间

DataBuddy Studio 在左侧导航中提供「数据目录」入口，您可以以树形结构逐层展开 Catalog → Schema → 对象 → 字段：

- 进入 [Studio](#)，在左侧导航选择 **数据目录**。
- 在树中点击 Catalog 或 Schema 即可向下展开；hover 字段或表名时，可使用插入按钮把名称插入到右侧 IDE 编辑器。
- 通过顶部搜索框可全域模糊搜索 Catalog / Schema / 资产名称（最多返回 200 条），匹配命中后目录树自动展开到对应节点。
- 通过筛选器可仅展示指定类型（数据表、模型 / Volume / 全部）的资产。

详细操作请参考 [数据目录导航](#)。

全名与跨工作空间引用

每个对象都拥有一个 **全名**（Fully Qualified Name），格式为 `catalog.schema.object`。全名是跨工作空间引用对象的标准形式：

引用场景	推荐写法
跨工作空间查询	<code><catalog>.<schema>.<table></code>
单工作空间且指定了上下文	<code><table></code> 或 <code><schema>.<table></code>
在告警 / 文档中描述资产位置	使用三段式全名，便于检索
API / SDK 调用	使用三段式全名作为标识参数

使用限制

限制项	说明
系统保留名	<code>information_schema</code> / <code>default</code> 不可作为 Schema 名

跨 Catalog 联表	跨 Catalog 联表查询能力依赖底层计算引擎；DLC Spark 已支持基础的跨 Catalog 联表查询
模型对象层级	模型在三级命名空间之上还有版本与别名两层，详见 模型 Catalog

常见问题

Q：为什么需要三级命名空间？两层（Schema.Table）不够吗？

A：三级命名空间为多业务线、多项目场景下的资产隔离与跨空间共享提供了顶层划分。例如可以为不同业务线创建独立的 Catalog（`marketing_catalog`、`finance_catalog`），每条业务线内部再按主题划分 Schema，避免命名冲突；同时便于按 Catalog 维度做权限批量授权。

Q：同名 Catalog 在不同工作空间是同一个吗？

A：是。Catalog 注册在 MetaLake 上，**跨工作空间可见**。在工作空间 A 创建的 Catalog，在同一地域的其他工作空间中能看到同名 Catalog（实际是同一对象）。具体能否访问其内部数据，取决于权限策略与工作空间访问范围。

Q：跨 Catalog 引用的语法在 Spark 和 SQL 任务中表现一致吗？

A：基本一致。`catalog.schema.object` 三段式语法在 DLC Spark SQL 中已支持；执行前请确认当前计算资源对目标 Catalog 拥有 `use catalog` 与 `use schema` 权限。

相关文档

- [统一元数据管理架构](#)
- [结构化数据 Catalog](#)
- [非结构化数据 Catalog](#)
- [模型 Catalog](#)
- [数据目录导航](#)

结构化数据 Catalog（表、视图、Schema 管理）

最近更新时间：2026-10-09 16:26:21

概述

结构化数据 Catalog（Table 类 Catalog）用于集中管理表（Table）、视图（View）和函数等结构化资产。本文介绍如何创建一个 Table 类 Catalog，并在其下管理 Schema、表和视图。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标工作空间中拥有创建 Catalog 的权限（默认主账号、控制台管理员可创建；其他用户需被授予 `create catalog` 权限）。
- 已了解 [三级命名空间](#) 的基本概念。

使用限制

限制项	说明
支持地域	全部支持 DataBuddy 的地域
存储后端	当前仅支持腾讯云 COS；存储路径在 Catalog 创建时由系统按 <code>default</code> 存储空间分配
系统保留名	<code>information_schema</code> / <code>default</code> 不可作为 Schema 名
引擎能力	表的实际读写依赖计算资源对该 Catalog 的引擎支持，DLC Spark / SQL 已支持

创建 Table 类 Catalog

步骤 1：进入数据目录

- 登录 [DataBuddy](#) 控制台，切换到目标工作空间。
- 在左侧导航栏选择 **数据目录**。

步骤 2：新建 Catalog

- 在左侧目录树顶部点击 **+ 创建数据目录**。
- 在 **创建数据目录** 弹窗中配置参数：
 - 目录名称**：必填。
 - 目录来源**：必填

- **目录类型**：必填
- **目录用途**：必填
- **描述**：可选，最多 500 字符。

1. 点击 **确定** 完成创建。

创建成功后，您会自动进入新 Catalog 的详情页，并看到系统自动生成的 Schema：`default`。

❗ 提示

若 + **创建数据目录** 按钮置灰且 hover 提示「您暂无权限」，请联系主账号或当前区域的元数据湖 admin 为您授予 `create catalog` 权限。

管理 Schema

Schema 在 Catalog 下提供二级命名空间，用于按业务主题或项目划分对象。

创建 Schema

1. 在 Catalog 详情页点击**创建 Schema**。
2. 在弹窗中输入：
 - **Schema 名称**：必填，在当前 Catalog 内唯一。
 - **描述**：可选。

1. 点击**确定** 完成创建。

创建成功后，您会自动进入 Schema 详情页。

查看 Schema 列表

在 Catalog 详情页 **概览** Tab 中：

- 通过名称搜索框可对 Schema 名称做实时模糊匹配。
- 列表展示 Schema 名称、描述、负责人（默认是创建人）、创建时间

删除 Schema

在 Schema 详情页右上角的 菜单中：

- **删除**：二次确认后删除整个 Schema 及其下属对象。

⚠ 注意

仅 Schema 的 owner、Catalog owner 和具备 `drop schema` 权限的用户可执行删除操作；其他用户对应入口置灰。

🔔 警告

删除 Schema 不可恢复，且会同时删除其下所有表、视图等对象。请在删除前确认无依赖任务。

管理表与视图

DataBuddy 中，表与视图均归属于 Table 类 Catalog 的 Schema。

创建表

DataBuddy 通过数据接入模块创建表：在 Catalog 详情或 Schema 详情页点击 **添加对象 > Table**，跳转到数据接入的建表流程。建表完成后，新表会自动出现在所在 Schema 下。

详细建表方式请参考：

- 离线数据同步：表到表同步
- 实时整库多表任务

表详情

在 Schema 详情页或左侧目录树中点击表名，进入表详情页。表与视图共用同一个详情页结构，主要内容包括：

Tab	内容
概览	表描述（可编辑）、字段列表（字段名称、字段类型、字段描述、字段标签）
数据预览	表数据采样，默认通过 DLC 引擎拉取，加载数据预览需要使用计算资源
血缘	上下游血缘关系，详见 数据血缘
权限	当前表的授权情况，详见下方 权限管理
详情	表的技术信息，eg：数据类型、表类型、存储路径、存储量、创建时间、创建人、数据变更时间、变更人等
变更历史	表的变更时间、变更类型，变更的用户，变更的操作等信息。加载变更历史需要使用计算资源
质量	关联的数据质量规则，详见 数据质量

表的右侧信息面板

表详情页右侧固定展示：

模块	字段
业务信息	标签（治理标签，可编辑）、负责人（支持转交）
技术信息	数据类型（内部表 / 外部表）、数据源、存储路径（可复制）、存储量、创建时间、数据变更时间、创建人、变更人

功能信息	数据预览、血缘、权限、详情、变更历史、质量
------	-----------------------

编辑字段元数据

在表详情页 概览 Tab 的字段列表中：

- 点击**字段描述** 旁的笔形图标可编辑字段描述。
- 点击**标签** 旁的笔形图标可为字段添加治理标签（key:value 形式）。

视图

视图与表共用同一个详情页结构。

i

提示

当前版本不支持通过 UI 创建视图（View）。如需创建视图，请使用计算引擎的 SQL 语法（如 DLC Spark SQL 的 `CREATE VIEW`）。

重命名、删除表

表详情右上方支持

- **修改表名称**：弹窗中修改表名后保存。
- **删除表**：二次确认后删除整张表。

仅表的 owner、所在 Schema 的 owner 和具备 `drop` 权限的用户可执行；其他用户对应入口置灰。

收藏与最近访问

在 Catalog 详情页右侧详情区，您可以通过以下方式快速回到常用资产：

- **全部目录**：列出当前工作空间可见的全部 Catalog。
- **最近访问**：按上次查看时间倒序展示您近期浏览过的 Catalog / Schema / 表、视图，可按资产类型筛选。
- **我的收藏**：展示您手动收藏的资产。表、视图 / Schema / Catalog 详情页右上角的星形图标可加入收藏。

在表详情页快速创建分析任务

在表详情页右上角的 **创建** 下拉菜单中，可一键创建关联当前表的开发任务：

入口	跳转目标
Notebook	跳转到 Notebook IDE ，自动带入当前 Catalog / Schema 上下文
SQL 探索	跳转到 SQL 探索 ，预填三段式表名
仪表盘	跳转到 仪表盘 创建页
Buddy	跳转到 Buddy 智能 分析空间

质量任务	跳转到 质量任务调度与执行，关联当前表创建质量规则
Power BI 连接插件	通过 Power BI 连接插件直接消费当前表数据（需先安装插件）
在线表	将当前离线特征表发布为在线特征表（详见 特征工程 SDK ）

权限管理

Catalog / Schema / 表的权限模型由三层级联控制，常用权限语义如下：

预设权限类型

预设权限	包含的权限点
读取权限	<code>use catalog</code> + <code>use schema</code> + <code>read</code> + <code>select</code> + <code>execute</code>
编辑权限	除 <code>execute</code> 与 <code>grant</code> 外的所有权限点

在不同层级授权

层级	可授予的权限点（节选）
Catalog	<code>use catalog</code> 、 <code>create catalog</code> 、 <code>alter catalog</code> 、 <code>drop catalog</code> ，以及 Schema 与对象层级的所有权限点（向下覆盖）
Schema	<code>create schema</code> 、 <code>use schema</code> 、 <code>alter</code> 、 <code>drop</code> 、 <code>select</code> 、 <code>insert</code> 、 <code>create table</code> 、 <code>create model</code> 、 <code>create volumes</code> 等
Table / View	<code>alter</code> 、 <code>drop</code> 、 <code>select</code> 、 <code>insert</code> 、 <code>all</code> 、 <code>grant</code>

提示

权限点之间存在依赖关系。例如授予表的 `select` 权限，需同时授予该表所在 Catalog 的 `use catalog` 与 Schema 的 `use schema`，否则用户无法定位到该表。授权弹窗会在确认前提示您勾选依赖项。

仅 owner 与具备 `grant` 权限的用户可授权

- 主账号、Catalog / Schema / 表的 owner 默认拥有 `grant` 权限，可对其管辖的层级授权。
- 普通用户在未被授予 `grant` 的情况下，**授权按钮置灰**，hover 提示「您无法授予权限，请联系主账号或 owner 获取权限」。

没有权限时的可见性

- 没有权限的 Catalog / Schema / 表会在目录树中置灰展示（而非隐藏），便于用户感知数据存在并申请权限。
- 用户可勾选 只展示有查询权限的数据 进行筛选。
- 进入无权限对象的详情页时，仅可查看 概览 Tab 的元数据；其他 Tab 置灰，hover 提示「您只能查看该资源对象的元数据，请联系主账号或 owner 获取权限」。

更多权限规则请参考 [成员与权限管理](#)。

业务标签与 owner 转交

添加、修改治理标签

Catalog / Schema / 表的右侧信息区均可添加治理标签：

1. 在 业务信息 > 标签 旁点击笔形图标。
2. 在弹窗中选择已有标签 key 与 value，可同时添加多个。

治理标签的整体管理请参考 [治理标签](#)。

owner 转交

每个 Catalog / Schema / 表、视图都有 owner，初始为创建人。owner 拥有该对象的全部权限，仅 owner 可发起转交：

1. 在右侧信息区 业务信息 > owner 旁点击笔形图标。
2. 选择新的 owner（仅当前对象的 owner 可见此入口）。
3. 在弹窗中确认提示文案：「转交后您不再拥有该 catalog / schema / table 的管理权限」。
4. 点击 确定 完成转交。

常见问题

Q：为什么 Catalog 详情页 / Schema 详情页中的某个表显示置灰？

A：表所在 Catalog 或 Schema 您没有 `use catalog` 或 `use schema` 权限，或者您没有该表本身的访问权限。详情页 hover 会显示具体提示，请联系 owner 或主账号申请。

Q：删除 Catalog / Schema 后能恢复吗？

A：不能。删除操作不可恢复，会连带删除所有下属对象。请确认无依赖任务后再操作。

Q：可以为表添加哪些字段级元数据？

A：当前支持字段描述与字段标签（`key:value` 形式的治理标签）。字段级权限策略由 [列级权限控制](#) 提供。

Q：表详情页的「数据变更时间」与「创建时间」有什么区别？

A：「创建时间」是对象的初次创建时间；「数据变更时间」是元数据最近一次变更的时间（如修改字段、修改描述、表数据写入等）。

相关文档

- [统一元数据管理架构](#)
- [三级命名空间](#)

- [非结构化数据 Catalog](#)
- [数据血缘](#)
- [成员与权限管理](#)

非结构化数据 Catalog（文件、Volume 管理）

最近更新时间：2026-10-09 16:26:26

概述

非结构化数据 Catalog（Volume 类 Catalog）用于集中管理日志、图像、音视频、PDF 等非结构化数据。DataBuddy 通过 **Volume（卷）** 这一资产类型，将非结构化文件纳入与表、模型相同的三级命名空间，统一进行权限治理与血缘追踪。

本文介绍如何创建 Volume 类 Catalog、在其下创建 Volume，并完成文件上传、浏览与预览。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标工作空间中拥有创建 Catalog 的权限。
- 已了解 [三级命名空间](#) 的基本概念。

使用限制

限制项	说明
支持地域	全部支持 DataBuddy 的地域
存储类型	当前仅支持腾讯云 COS；托管 Volume 由平台直接管理底层存储；外部 Volume 需基于 External Location 创建
支持的预览类型	CSV、Text、PDF（其他文件类型仅支持下载）
Volume 搜索	全局数据搜索支持搜索 Volume 名称、描述、标签，以及 Volume 下的文件夹和文件名称
文件夹创建	在 Volume 内可创建文件夹，重复文件名上传时自动覆盖

创建 Volume 类 Catalog

步骤 1：进入数据目录

- 登录 [DataBuddy 控制台](#)，切换到目标工作空间。
- 在左侧导航栏选择 **数据目录**。

步骤 2：新建 Volume 类 Catalog

- 在左侧目录树顶部点击 **+ 创建数据目录**。

2. 在 **创建数据目录** 弹窗中配置参数：

- **目录名称**：必填。
- **目录来源**：必填，选择 TCCatalog
- **目录类型**：必填，选择原生目录
- **目录用途**：必填，选择文件卷目录
- **描述**：可选，最多 500 字符。

3. 点击 **确认** 完成创建。

创建成功后，您会进入 Catalog 详情页，与 结构化数据目录 相同，包含 `default` Schema

创建 Volume

Volume 是非结构化文件的容器，挂载在 Schema 之下。

步骤 1：进入 Schema 详情页

1. 在 Catalog 详情页点击目标 Schema 名称，进入 Schema 详情页。
2. 在右上角点击 **创建数据对象** 。

步骤 2：填写 Volume 信息

在弹窗中选择 Volume 类型：

Volume 类型	说明
托管 Volume （Managed Volume）	由平台分配存储路径，您只需填写 Volume 名称与描述
外部 Volume （External Volume）	基于已注册的 External Location 创建，复用外部 COS 路径

填写以下信息：

- **Volume 名称**：必填，在当前 Schema 内唯一。
- **Volume 存储类型**：云托管、外部。
- **External Location**：仅外部 Volume 需要，下拉选择已注册的 External Location。
- **描述**：可选。

点击 **确定** 完成创建。创建成功后，您会进入 Volume 详情页。

在 Volume 中管理文件

Volume 详情页 **概览** Tab 提供文件与文件夹管理能力。

上传文件

在 Volume 详情页右上角点击 **上传数据** ，跳转到非结构化数据上传页面。详细操作请参考 [本地文件上传到卷](#) 。

提示

文件上传由前端直接调用 COS API，无需选择计算资源。重名文件会自动覆盖原文件。

创建文件夹

1. 在 Volume 详情页 **概览** Tab 点击 **创建文件夹**。
2. 在弹窗中填写文件夹名称。
3. 点击 **确定** 完成创建。

文件夹支持嵌套，路径以 `/Volumes/<catalog>/<schema>/<volume>/<directory>/...` 形式呈现，列表上方面包屑可复制完整路径。

删除文件、文件夹

- 在文件列表中勾选目标文件或文件夹，列表内点击**删除**，支持在顶部**更多**下拉选择批量删除操作。
- 在文件夹中删除非空文件夹会同时删除其下所有内容。

警告

文件与文件夹删除不可恢复。请在删除前确认无依赖任务或下游引用。

文件搜索

在文件列表上方搜索框输入文件名或文件夹名称，进行实时模糊搜索。

文件预览

在文件列表操作中点击预览图标，进入文件预览页：

- 支持类型：CSV、Text、PDF。
- 操作：下载、预览、删除
- 不支持预览的文件类型仅展示元信息，可下载到本地查看。

文件路径格式

Volume 中的文件遵循统一路径格式：

```
/Volumes/<catalog>/<schema>/<volume>/<directory>/<file>
```

例如：

```
/Volumes/ops_catalog/raw_logs/access_log/2026-06-06/access.log
```

该路径可直接被 SQL、Python 代码、Notebook 引用：

```
with open("/Volumes/ops_catalog/raw_logs/access_log/2026-06-06/access.log") as f:
    data = f.read()
```

Volume 详情页信息

展示：存储路径、创建时间、创建人、数据变更时间、变更人

权限管理

Volume 的权限模型与表、模型保持一致，可在 Catalog / Schema / Volume 任意层级单独授权。

Volume 层级权限点

权限点	含义
read	读取 Volume 中的文件（含下载、预览）
write	写入 / 更新 Volume 中的文件（含上传、覆盖）
delete	删除 Volume 中的文件
create volumes	在所属 Schema 下创建 Volume
all	授予以上所有权限
grant	可二次授权（管理以上所有权限）

提示

Volume 的权限点同样存在依赖关系：授予 `read` / `write` / `delete` 时，请确保用户已具备所属 Catalog 的 `use catalog` 与所属 Schema 的 `use schema` 权限，否则用户无法定位到该 Volume。

配置 Volume 权限

- 进入 Volume 详情页 **权限** Tab。
- 点击 **新增授权**，选择授权主体（用户/用户组）。
- 选择预设权限（读取、编辑）或勾选具体 Volume 权限点。
- 点击 **确定** 完成授权。

仅主账号、Volume 的 owner 和具有 `grant` 权限的用户可执行授权；其他用户授权按钮置灰。

转交 Volume owner

1. 在 Volume 详情页右侧 **业务信息** > **资产负责人** 旁点击笔形图标。
 2. 选择新的 owner。
 3. 在弹窗中确认提示文案：「若您非当前 Volume 创建人，转交后您不再拥有该 Volume 的管理权限」。
- 更多权限规则请参考 [成员与权限管理](#)。

业务标签

在 Volume 详情页右侧 **业务信息** > **标签** 旁点击笔形图标，可为 Volume 添加多个治理标签（`key:value` 形式）。详细规则请参考 [治理标签](#)。

常见问题

Q：托管 Volume 与外部 Volume 有什么区别？

A：

- **托管 Volume**：底层存储由平台分配（默认 `default` 存储空间），无需您预先准备 COS 路径。适合新建项目、测试场景。
- **外部 Volume**：基于已注册的 External Location 创建，复用外部 COS 桶路径，便于跨工作空间共享或对接已有数据。详见 [外部数据管理](#)。

Q：上传文件时为什么不需要选择计算资源？

A：Volume 文件上传由前端直接调用 COS API 完成，不经过计算引擎。因此上传过程不消耗计算资源。

Q：删除 Volume 会删除底层 COS 文件吗？

A：

- 托管 Volume：删除 Volume 后，平台会同时清理对应 COS 路径下的文件。
- 外部 Volume：删除 Volume 仅解除 DataBuddy 中的元数据注册，**不会删除** External Location 对应的 COS 文件。

Q：Volume 中的文件能不能被 SQL 直接读取？

A：可以。Volume 路径（`/Volumes/<catalog>/<schema>/<volume>/...`）可在 Spark SQL、Python、Notebook 中直接读取，例如使用 Spark 读取 Volume 中的 Parquet：`spark.read.parquet("/Volumes/op_s_catalog/raw_logs/event/")`。

相关文档

- [三级命名空间](#)
- [外部数据管理](#)
- [本地文件上传到卷](#)
- [成员与权限管理](#)

模型 Catalog（ML 模型注册与元数据）

最近更新时间：2026-10-09 16:26:31

概述

模型 Catalog（Model 类 Catalog）将机器学习模型纳入与表、卷一致的统一治理框架。DataBuddy 在 Catalog 中提供托管的 MLflow Model Registry，您可以将实验训练出的模型集中注册、版本化管理，并通过别名快速引用特定版本进行部署或在线推理。

本文介绍如何创建 Model 类 Catalog、注册模型、管理模型版本与元数据。模型部署与服务化推理请参考 [模型服务](#)。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标工作空间中拥有创建 Catalog 的权限，以及目标 Catalog/Schema 的 `create model` 权限。
- 已熟悉 [实验管理](#) 的基本流程，并已完成至少一次实验运行（用于 UI 注册）。
- 模型必须包含 **模型签名（Signature）**，即输入输出的字段名称与数据类型，否则无法注册。

使用限制

限制项	说明
注册前置条件	模型必须存在签名（Signature）才能注册
标签来源	模型与版本的标签必须从 治理标签 策略中选择，UI 不支持任意自由文本
别名唯一性	同一模型下的别名（Alias）不可重复
标签操作	模型标签仅支持 UI 添加 / 修改 / 删除；通过代码自定义模型创建暂不支持模型标签写入

创建 Model 类 Catalog

1. 在左侧导航栏选择[数据目录](#)。
2. 在左侧目录树顶部点击 + [创建数据目录](#)。
3. 在弹窗中配置：
 - **目录名称**：必填。
 - **目录来源**：必填，选择 TCCatalog
 - **目录类型**：必填，选择原生目录
 - **目录用途**：必填，选择模型文件目录

- **描述**：可选，最多 256 字符。

1. 点击 **确认** 完成创建。

创建成功后，您会进入 Catalog 详情页，与 结构化数据目录 相同，包含 `default` Schema

注册模型

DataBuddy 支持两种模型注册方式：

方式一：UI 注册（从实验运行）

1. 进入 [实验管理](#)，找到目标实验运行。
2. 在运行详情页点击 **注册模型**。
3. 在弹窗中：

- **注册方式**：新模型、已有模型新版本
- **模型名称**：填写模型名称，需在所选 Schema 内唯一。

1. 点击 **注册模型**。

系统会调用 Catalog 注册接口完成注册。注册成功后，您可在 Catalog 中通过 `<catalog>.<schema>.<model>` 路径找到该模型。如果同名模型已存在，新注册会作为该模型的下一个版本（版本号递增）。

方式二：Notebook / 代码注册

在 Notebook 或 Python 任务中使用 MLflow 客户端注册模型。DataBuddy 默认注册 URI 已设置为 Catalog 内的 MLflow Model Registry：

```
import mlflow
from mlflow.models import infer_signature

# 训练并打包模型
signature = infer_signature(X_train, model.predict(X_train))
mlflow.sklearn.log_model(
    sk_model=model,
    artifact_path="model",
    signature=signature,
    registered_model_name="ml_catalog.recsys.deepfm_model"
)
```

注册接口会上报实现信息（如评估指标）到 Catalog 的模型版本属性中。

提示

注册前请使用 `mlflow.models.infer_signature` 或手工构造 `Signature` 。无签名的模型在注册时会被拒绝。

模型详情页

通过左侧目录树或 Schema 详情页进入模型，详情页结构如下：

区域	内容
顶部面包屑	Catalog > Schema > 模型名称（可复制）
概览 Tab	模型描述、负责人、标签、版本列表
活动日志	模型注册时间、版本、日志
权限 Tab	模型维度的授权情况（详见 权限管理 ）

模型描述

在概览 点击添加描述 或描述区的笔形图标，弹出描述输入框：

- 模型描述与版本描述统一限制为最长256 字符。
- 支持反复编辑。

版本列表

模型概览 下方的版本列表按版本号倒序展示：

列	说明
版本号	点击进入版本详情页，可查看模型血缘等详细信息
标签	该版本的治理标签（ <code>key:value</code> ），用于标注业务属性，例如冠军版本、训练数据集等
别名	该版本绑定的别名（如 <code>champion</code> 、 <code>production</code> ）
部署状态	已部署 / 未部署
已部署服务	一个版本可部署到多个服务，多于一个时以 +N 收纳
备注	该版本的备注信息

模型所有者

右侧 业务信息 > 模型所有者 默认是注册人，点击笔形图标可发起转交：

- 仅当前所有者可见转交入口。
- 转交时弹窗提示「转交后您不再拥有该模型的管理权限」。

模型版本管理

点击版本号进入版本详情页。

版本详情页结构

区域	内容
操作区	模型描述、别名、标签
非操作区	指标、参数、活动日志、模型版本签名（输入输出字段表）

模型版本签名

签名记录模型的输入输出规范，作为部署时的输入输出强制校验依据：

列	说明
输入字段名称	例如 <code>user_id</code>
输入数据类型	例如 <code>STRING</code>
输出字段名称	例如 <code>score</code>
输出数据类型	例如 <code>DOUBLE</code>

别名

别名是指向特定版本的可变命名引用，常用于部署生产版本：

- 在版本详情页 **别名** 旁点击编辑。
- 输入别名（最长 50 字符）。
- 点击 **保存** 。

如果同一模型下别名已被占用，会提示「<别名> 这一别名已被占用，请重新输入」。

引用示例：

```
models:/ml_catalog.recsys.deepfm_model@champion
```

将别名重新分配给不同版本，可在不修改调用代码的情况下切换生产版本。

标签（版本级）

模型版本的标签同样从 [治理标签](#) 策略中选择，可添加多个：

限制	说明
----	----

数量上限	单版本最多 10 个标签
value 长度	复用治理标签管理的限制（不超过 50 字符）

版本来源

如果该版本是从其他模型版本复制而来，关于此模型版本 区域会显示：

- **复制自**：来源版本的全称（如 `ml_catalog.recsys.deepfm_model/3`）。
- **版本来源**：如果来自实验运行，会展示来源链接，可跳转实验详情。

复制模型版本

将一个版本复制到另一个模型，便于把已有版本作为新模型的初版。

1. 在版本详情页右上角点击 **复制此版本**。
2. 在弹窗中：
 - **复制到模型**：下拉选择目标模型。
 - 仅可选择您拥有权限的目标模型。
3. 点击**复制**。

复制流程的反馈：

- 触发复制后弹窗关闭，页面提示「已开始复制模型版本：请注意，此过程可能需要一些时间」。
- 成功后页面顶部显示绿色横幅「版本复制成功！新版本：[新版本全称]」，提示框右侧含 **前往新版本** 按钮，可直接跳转。
- 失败时弹窗内显示红色错误提示（例如目标 Schema 无写入权限），并保留弹窗配置，允许修改后重试。

新版本规则：

- 新版本号在目标模型中递增（若原最大版本为 5，则复制后为版本 6）。
- 新版本的状态默认为 **就绪**，与原版本状态无关。
- 关于此模型版本 自动记录 **复制自** 字段。

删除模型版本

1. 在版本详情页右上角点击**删除此版本**。
2. 在二次确认弹窗中确认文案：「确认删除 `<版本全称>` 吗？删除后不可恢复。」
3. 点击 ****删除**** 按钮 删除。



警告

删除模型版本不可恢复。删除前请确认无在线服务、无下游推理任务依赖该版本。

权限管理

模型的权限模型与表、Volume 一致，可在 Catalog / Schema / Model 任意层级授权。

Model 层级权限点

权限点	含义
<code>create model</code>	创建模型
<code>read</code>	读取模型（用于推理或调用）
<code>write</code>	更新模型
<code>delete</code>	删除模型
<code>create model version</code>	在模型基础上创建新版本
<code>use model</code>	使用该模型
<code>all</code>	授予以上所有权限
<code>grant</code>	可二次授权

 提示

权限点存在依赖关系：授予模型的 `read`、`write`、`use model` 时，请确保用户已具备所属 Catalog 的 `use catalog` 与所属 Schema 的 `use schema` 权限。授权弹窗会在保存前提示您勾选依赖项。

配置模型权限

- 进入模型详情页或版本详情页的 **权限** Tab。
- 点击 **新增授权**，选择授权主体（支持多选用户）。
- 选择需要授予的权限点。
- 点击 **确定** 完成授权。

仅主账号、模型 owner 和具有 `grant` 权限的用户可执行授权。无 `grant` 权限时，授权按钮置灰。
更多权限规则请参考 [成员与权限管理](#)。

与其他模块的关系

模块	关系
实验管理	模型注册的入口之一；实验运行可通过 UI 一键注册到 Catalog
模型服务	部署在线推理服务时，从 Catalog 选择模型版本或别名
模型管理	数据科学模块中的模型列表，与 Catalog 中的模型为同一份元数据，可双向跳转

数据血缘	模型作为血缘节点，可追踪上游训练数据与下游推理资产
治理标签	模型与版本的标签均从治理标签策略中选择

常见问题

Q：必须使用 MLflow 才能将模型注册到 Catalog 吗？

A：是。DataBuddy Catalog 中的模型注册沿用 MLflow Model Registry 协议，模型需要先以 MLflow 的格式记录（即包含模型工件、签名等）。在 Notebook 中可直接使用 `mlflow.<flavor>.log_model` API；在实验中训练的模型由 DataBuddy 自动以 MLflow 格式打包。

Q：模型注册时报错「必须存在模型签名」怎么办？

A：模型在注册前必须包含输入输出签名。请在训练时使用 `mlflow.models.infer_signature(input, output)` 自动推断，或显式构造 `Signature` 对象后传给 `log_model`。

Q：别名（Alias）和版本号（Version）的区别？

- A：
- **版本号**：在每次注册时自动递增、不可修改。版本号唯一对应一个具体的模型工件。
 - **别名**：可变的命名引用，可以重新分配到不同版本。常用于把生产版本封装为 `champion`、灰度版本封装为 `staging`，部署侧通过别名即可在不变更代码的情况下切换版本。

Q：删除模型、删除版本的影响范围？

- A：
- **删除版本**：仅删除该版本的元数据与工件。其他版本不受影响，但若该版本已被在线服务引用，相关服务调用会失败。
 - **删除模型**：删除该模型下所有版本与历史。请在删除前确认无下游推理任务依赖。

相关文档

- [三级命名空间](#)
- [实验管理](#)
- [模型管理](#)
- [模型服务](#)
- [数据血缘](#)
- [成员与权限管理](#)

外部数据管理与基于数据源创建外部数据目录

最近更新时间：2026-10-09 16:26:38

概述

DataBuddy 数据目录除了管理内部托管的资产（表、卷、模型）外，还支持将 **外部数据** 纳入治理体系。外部数据管理包含两类对象：

- **External Location（外部存储）**：将一段腾讯云 COS 路径注册为治理对象，作为创建外部 Volume 或外部表的存储入口，统一管理跨工作空间的存储授权。
- **External Catalog（外部数据目录 / Foreign Catalog）**：基于已配置的数据源连接（Connection）创建的「镜像」型 Catalog。它不存储数据，仅同步外部数据源的元数据，让用户可以像访问内部 Catalog 一样查询外部数据库。

本文介绍这两类对象的创建与管理方法。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 创建 External Location 需要 **控制台管理员** 角色。
- 创建 External Catalog 需要先在 **数据源** 中创建 Connection，并具备 Connection 的 owner 权限或被授予的相应权限。
- 已在 CAM 中配置好 TC Catalog 服务角色（ `ccatalog_QCSLinkedRoleInMetadataManagement` ），并为其分配 COS 访问权限（ `cos:*` action / `cos:*` resource ）。

使用限制

限制项	说明
存储类型	External Location 当前仅支持腾讯云 COS
创建 External Location 权限	仅控制台管理员可创建；其他用户对应入口置灰
路径唯一性	同一 COS 路径不可重复注册
External Catalog 数据源范围	当前仅支持基于 Doris、Starrocks、GaussDB、PostgreSQL、MySQL 创建外部数据目录
External Catalog 操作	Foreign Catalog 是只读镜像，不支持在 Catalog 内创建 / 修改 Schema、表；表的 DDL 与数据写入必须在源数据库中执行
External Catalog 表详情	不展示「表变更历史」「血缘」Tab；保留概览、详情、数据预览、权限、工作空间 Tab

External Location 管理

创建 External Location

步骤 1：进入外部存储页面

1. 登录 [DataBuddy 控制台](#)，切换到目标工作空间。
2. 在左侧导航栏选择 **数据目录**。
3. 在顶部点击 **外部数据**，进入外部数据管理页。
4. 切换到 **外部存储** Tab。

步骤 2：创建外部存储

1. 在外部存储页面右上角点击 **创建**。
2. 在创建页面填写参数：
 - **服务角色**：从已配置的 TC Catalog 服务角色中提取，目前固定为 `ccatalog_QCSLinkedRoleInMetadataManagement`。
 - **存储类型**：当前仅支持 **腾讯云 COS**。
 - **存储地址**：从下拉中选择目标 COS Bucket 路径地址。
 - **描述**：可选，便于团队协作。
3. 点击 **创建** 提交。

平台会调用 COS API（`ListBucket`）执行以下校验：

- Bucket 是否存在。
- 路径是否与已注册的 External Location 重复。
- 服务角色是否未被云策略拒绝访问该路径。

校验失败时会展示具体错误（例如：「平台服务角色无权访问该路径，请联系云管理员检查 CAM 策略」）。校验通过后页面顶部提示「创建成功」。

⚠ 注意：

默认策略：External Location 创建时默认对所有用户授予 `browse` 权限，对所有工作空间授予访问权限。如需限制范围，请在创建后到权限页调整。

External Location 列表页

在 **数据目录 > 外部数据 > 外部存储** 页面，您可以：

- 通过名称搜索框对名称做模糊搜索。
- 在列表中查看名称、URL、创建人、创建时间、描述。
- 点击右上角 **创建外部存储** 进入创建页（仅控制台管理员可点击；其他用户置灰，hover 提示「仅控制台管理员可创建外部存储」）。

External Location 详情页

点击列表中的 External Location 名称，进入详情页：

区域	内容
顶部面包屑	外部数据管理 > 外部存储名称
操作	重命名 、 删除
基本信息	描述（可编辑）、存储类型、URL、创建时间、创建人
权限 Tab	该 External Location 的授权情况

重命名、删除

在详情页右上角：

- **重命名**：弹窗修改名称后保存。
- **删除**：二次确认后删除注册信息。

 **警告：**
删除 External Location 仅解除 DataBuddy 的注册关系，**不会** 删除 COS 上的实际文件。但若已有外部 Volume 或外部表基于此 Location 创建，删除前请先移除依赖，否则相关资产会失效。

External Location 权限管理

权限点

权限点	含义
CREATE EXTERNAL TABLE	基于该 Location 创建 / 修改外部表
CREATE EXTERNAL VOLUME	基于该 Location 创建 / 修改外部 Volume
ALL PRIVILEGES	上述全部权限
MANAGE	可编辑、删除 Location，并可二次授权

配置权限

1. 进入 External Location 详情页 **权限 Tab**。
2. 点击 **新增授权**，选择授权主体（用户或用户组）。
3. 勾选权限点。
4. 点击 **确定** 完成授权。

仅控制台管理员、Location 创建人和被授予 `MANAGE` 的用户可发起授权。

基于 External Location 创建 Volume

当您需要复用一段已注册的 COS 路径作为非结构化资产容器时，可以基于 External Location 创建 **外部 Volume**。整个流程不会复制数据，仅在 DataBuddy 中建立元数据注册关系。

操作步骤

1. 在左侧导航栏选择 **数据目录**，进入目标 Volume 类 Catalog 的 Schema 详情页。
2. 点击右上角**添加数据对象> Volume**。
3. 在弹窗中 Volume 类型选择**外部 volume**
4. 在 **External Location** 下拉中选择已注册的目标 Location；可在该 Location 下进一步指定子路径（可选）。
5. 填写 Volume 名称与描述，点击 **确定** 完成创建。

权限要求

- 当前用户需对所选 External Location 拥有 `CREATE EXTERNAL VOLUME`（或 `ALL PRIVILEGES`）权限。
- 当前用户需对目标 Schema 拥有 `create volumes` 权限。

行为说明

- 多个外部 Volume 可指向 **同一 External Location** 下的不同子路径，相互之间隔离。
- 删除外部 Volume **仅** 解除元数据注册，不会清理 External Location 上的实际文件。
- External Location 上的文件如果被外部进程修改，外部 Volume 内容会随之变化（DataBuddy 不进行二次缓存）。

更多 Volume 操作，详见 [非结构化数据 Catalog](#)。

External Catalog（基于数据源创建外部数据目录）

External Catalog 让您像访问内部 Catalog 一样查询外部关系型数据库（如 TencentDB PostgreSQL），无需先做数据接入或 ETL，适合 **临时查询** 与 **数据探索** 场景。

如果您需要长期高性能查询外部数据，建议改用 [离线数据同步](#) 或 [实时整库多表任务](#) 把数据接入到 DataBuddy 内部 Catalog。

创建 External Catalog

DataBuddy 提供两个入口创建 External Catalog：

入口 A：从已有数据源创建

适合已经创建过数据源但还未创建 Catalog 的场景。

1. 在左侧导航栏选择 **数据目录**。

2. 顶部点击 **外部数据管理**，进入数据源列表。
3. 点击目标数据源名称进入详情页。
4. 在详情页右上角点击 **创建外部数据目录**，弹出配置页面。

入口 B：在创建数据源后引导创建

适合首次创建数据源的场景。

1. 在 [数据源管理](#) 中创建数据源。
2. 创建完成后，弹窗中选择 **创建外部数据目录**。
3. 走创建外部数据目录流程

创建外部数据目录与创建数据接入任务的区别：

选项	适用场景
创建外部数据目录	临时查询、数据探索；不复制数据
创建数据接入任务	高性能查询、数据转换；将外部数据同步到 DataBuddy 内部存储

引导规则：

- **MySQL 类数据源**：默认引导用户创建外部数据目录。
- **其他数据源**：默认引导用户进行数据接入。

填写 Catalog 配置

1. 在创建外部目录配置页填写基础信息：

参数	说明	是否必填	默认值
Catalog 名称	手动填写	是	—
目录来源	默认是直连数据源	是	直连数据源
数据源	选择数据源名称	是	—
描述	—	否	—

❗ 说明：

当前仅支持基于 Doris、Starrocks、GaussDB、PostgreSQL、MySQL 创建外部数据目录。其他数据源类型在后续版本逐步支持。

2. 点击**确认** 提交。创建成功后，自动跳转到该 Catalog 详情页，并提示「创建 Catalog 成功」。

External Catalog 详情页特点

External Catalog 与内部 Catalog 详情页结构相似，但有以下差异：

差异点	内部 Catalog	External Catalog
Connection 字段	无	显示关联 Connection 名称，可跳转到 Connection 详情页
创建 Schema / 表	支持	不支持（Foreign Catalog 仅是镜像）
修改 / 删除	支持	当前仅只读，描述、删除、修改等操作暂不支持
表详情 Tab	概览、数据预览、权限、血缘、表变更历史等	概览、详情、数据预览、权限

⚠ 注意：
External Catalog 是源数据库元数据的镜像。如需变更 Schema 与表结构，请在源数据库中执行 DDL，External Catalog 会同步反映元数据变化。

External Catalog 权限管理

External Catalog 的权限模型与普通 Catalog 一致：

- 默认对 `all account user` 授予 `browse` 权限。
- 可对授权主体进行 Catalog 层权限授权。
- 创建外部 Catalog 的权限与创建内部 Catalog 的权限点保持一致。

具体的权限判断规则：

操作	允许的角色
创建 External Catalog	Connection owner / 控制台管理员 / 被授予「创建 Catalog 权限」且具备 <code>use connection</code> 的用户或用户组
修改 Connection 描述	Connection owner / 控制台管理员 / 被授予 <code>use connection</code> 的用户或用户组
转交 Connection owner	Connection owner / 控制台管理员 / 被授予 <code>use connection</code> 的用户或用户组

元数据同步行为

External Catalog 不存储数据，但会在以下时机从源数据库拉取元数据，确保 DataBuddy 中看到的 Schema / 表结构与源数据库保持一致：

触发场景	同步方式	用户可见性
Catalog 创建完成	立即同步一次全量元数据	创建成功后即可在目录树中看到 Schema 与表
用户在目录树展开 Catalog / Schema	按需拉取（懒加载）	展开节点时实时从源库读取
进入表详情页	实时拉取该表的字段、属性等元数据	进入页面时刷新
源库 DDL 变更后再次访问	下一次访问时自动同步	用户重新打开节点 / 详情页即可看到最新结构

① 说明：

External Catalog 不维护后台定时同步任务。当源数据库元数据变更后，DataBuddy 在用户下一次访问对应资产时按需刷新；如同步过程中失败（如源库不可达、Connection 失效、网络异常等），界面会展示同步错误，并提示用户检查 Connection 与计算资源的网络可达性。

跨工作空间共享

External Location 与 External Catalog 都支持跨工作空间共享：

- **External Location**：所有有权限的工作空间均可在该 Location 上创建外部 Volume / 外部表。
- **External Catalog**：默认对所有工作空间只读可见。如需限制，可在创建时取消默认勾选并指定可访问工作空间列表。

常见问题

Q：External Location 与 External Catalog 是同一个东西吗？

A：不是。

- **External Location** 是一段被注册为治理对象的 COS 存储路径，主要用于 **创建外部 Volume / 外部表的存储位置**。
- **External Catalog** 是基于 Connection 创建的 **外部数据库元数据镜像**，用于在 DataBuddy 中查询外部关系型数据库。

两者解决的问题不同：External Location 关注外部存储治理，External Catalog 关注外部数据库元数据治理。

Q：基于 External Location 创建的 Volume 与托管 Volume 有什么区别？

A：

- 托管 Volume 由平台分配存储路径，删除 Volume 时同时清理底层文件。
- 外部 Volume 复用 External Location 上的 COS 路径，多个外部 Volume 可指向同一 Location 下的不同子路径。删除外部 Volume 仅解除元数据注册，不删除文件。

Q：External Catalog 中能不能 `INSERT INTO` 写入数据？

A：当前不支持。Foreign Catalog 是只读镜像。如需写入外部数据库，请直接在源数据库中执行，或使用离线/实时数据同步任务。

Q：External Catalog 的查询性能如何？

A：查询通过计算引擎下推到外部数据库执行，性能取决于源数据库的查询能力与网络。对于高并发或大数据量查询，建议使用数据接入任务把数据同步到 DataBuddy 内部 Catalog。

Q：删除 External Location 时它上面的外部 Volume 会怎样？

A：删除前请先移除所有依赖该 Location 的外部 Volume / 外部表，否则这些资产会失效。

相关文档

- [统一元数据管理架构](#)
- [非结构化数据 Catalog](#)
- [全局数据源配置](#)
- [离线数据同步](#)
- [实时整库多表任务配置](#)
- [成员与权限管理](#)

数据血缘

最近更新时间：2026-10-09 16:26:42

概述

数据血缘（Data Lineage）记录表、视图、模型与生产它们的任务之间的上下游关系，回答两类核心问题：

- **影响分析（向下追溯）**：如果我修改这张表，下游有哪些任务、报表、模型会受影响？
- **数据溯源（向上追溯）**：这张表、这个模型的数据来自哪些源表、由哪个任务生成？

DataBuddy 在表和模型详情页提供 **血缘** Tab，自动采集 SQL 开发、Notebook、Workflow 任务、数据集成、Dashboard 等多种执行平台上报的血缘信息，并提供 **图模式** 与 **列表模式** 两种查看方式。

本文介绍数据血缘的查看方式、节点信息、采集范围与典型用法。

前提条件

- 当前账号在目标工作空间中拥有目标资产的查看权限。
- 上下游关系依赖各执行平台的上报，仅运行过的任务才会产生血缘记录。

使用限制

限制项	说明
血缘维度	当前支持 表级血缘 ；字段级血缘在后续迭代规划
资产类型	表（含视图、推理表、特征表）、模型
任务类型	SQL 开发、Notebook、Workflow 任务、数据集成（离线 / 实时）、BI（Dashboard）
采集触发	血缘随任务运行（含测试运行）自动上报
删除联动	任务或文件被删除时，平台自动清理对应血缘关系，避免残留无效节点

查看血缘

入口

资产类型	入口
表 / 视图	表详情页 血缘 Tab
模型	模型版本详情页 血缘 Tab

进入 **血缘** Tab 后，默认展示 **列表模式**，您可以点击右上角图标切换到 **图模式**。

列表模式

列表模式以列表形式展示当前资产的上下游一层关系，适合快速查阅与按名称搜索。

表的列表模式

列	说明
名称	上下游资产名称（表 / 视图 / 推理表 / 特征表 / 模型 / SQL / Notebook / Task），支持点击跳转到详情页
方向	该项是当前表的「上游」还是「下游」
类型	表 / 视图 / 推理表 / 特征表 / 模型 / SQL / Notebook / Task

支持模糊搜索资产名称、分页浏览。

模型的列表模式

模型的列表模式仅展示一层上下游关系：

列	说明
名称	上下游资产名称
方向	上游：数据表 / 特征表；下游：数据表 / 推理表
Process 类型	上游 process：Notebook / Studio；下游 process：模型服务
最近更新	该资产最近一次更新时间

图模式

图模式以拓扑图形式展示血缘关系，便于把握整体结构。点击列表模式右上角的 **图模式** tab 切换。

节点类型

血缘图中的节点分为 **资产节点** 与 **任务节点** 。

节点类型	子类型	图标
资产	表（含视图、推理表、特征表）	表 / 视图 / 推理表 / 特征表 各自的图标
资产	模型	模型图标（在画布右下角图例中可识别）
任务	TASK（Workflow 任务）	TASK 图标

任务	SQL（Studio 中的 SQL 文件）	SQL 图标
任务	NOTEBOOK	NOTEBOOK 图标
任务	数据集成（离线 / 实时）	离线 / 实时集成各自的图标
任务	BI（Dashboard）	BI 图标

节点信息

点击图中节点查看详细信息。各节点类型展示信息如下：

资产节点（表、视图）

- 业务信息：描述、标签、创建时间、责任人。
- 技术信息：数据源类型、Catalog、Schema、存储路径、最近 DDL 变更时间、最近数据变更时间。
- 跨数据源：Kafka 节点名称展示为 `topic` 名；ES 节点名称展示为 `index` 名。

TASK 节点（Workflow 任务）

- 基本信息：任务名称（可跳转）、任务类型、描述、创建时间、更新时间、创建人、所属工作流（可跳转）。
- 业务代码：右侧展示任务代码。

SQL 节点（Studio 中的 SQL 文件）

- 基本信息：运行记录、类型、上次运行时间、来源（可跳转）
- 业务代码：右侧展示 SQL 内容。

Notebook 节点

- 基本信息：文件名称（可跳转）、创建时间、创建人。
- 业务代码：右侧展示 Notebook 内容。

模型节点

- 唯一标识：模型名称（可跳转）、版本。
- 操作：点击 [查看详情](#) 跳转到模型详情页。

数据集成节点

- 任务名称（可跳转到任务详情页）、任务类型（实时、离线）、创建人、创建时间。

BI 节点

- BI 名称（可跳转到 Dashboard）、创建人、创建时间。

画布操作

操作	说明
搜索	画布右上角检索框，按节点名称（含机器学习模型名称）模糊搜索并定位
时间过滤	按上报时间范围过滤血缘节点
缩放 / 平移	鼠标滚轮缩放、拖拽平移，便于查看复杂血缘网络

图例	画布右下角展示节点类型图例，可在多种资产 / 任务图标间快速识别
----	----------------------------------

血缘的采集与更新

自动采集

DataBuddy 自动从下列执行平台采集血缘信息：

来源	上报内容
Studio SQL 文件 / Workflow SQL 任务	由计算引擎在执行时动态解析 SQL 输入输出表，每次提交（含测试运行）都会上报
Spark Jar 任务	任务运行时上报
Notebook	Notebook 中执行的 SQL / Python 代码引用的表
Workflow 任务	周期调度运行时随任务上报
数据集成（离线 / 实时）	由数据集成模块定时上报源表与目标表关系
BI（Dashboard）	由 BI 模块定时上报 Dashboard 与数据集、源表关系

底层基于 DLC SQL 在 Spark 内核的动态血缘解析能力，对所有给引擎提交的 SQL 与作业进行统一采集。

删除时的清理

当任务或文件被删除时，平台会向血缘系统上报删除事件，自动清理对应的血缘关系，避免血缘图中残留无效节点：

对象	触发动作
任务（Workflow Task）	任务被删除
文件（Studio 中的 SQL 文件）	文件被删除

时间元数据

血缘节点会附带时间元数据，便于区分活跃与历史节点：

字段	说明
任务创建时间	任务被创建的时间，用于区分新老任务
任务最近运行时间	最近一次运行时间
SQL 最近运行时间	SQL 文件最近一次运行时间

模型血缘

模型血缘是表血缘的延伸，对应机器学习全流程的资产关系：

阶段	节点类型	关系说明
原始数据 → 特征表	上游：数据表；下游：特征表	沿用表血缘逻辑（数据处理任务）
特征表 → 训练数据	上游：特征表；下游：训练数据表	数据处理 / 特征提取任务
训练数据 → 模型	上游：表；下游：模型节点	模型训练任务（来源是 Notebook 或 Studio 任务）
模型 → 推理表	上游：模型；下游：推理表	模型服务任务

模型血缘的关键行为：

- 模型唯一标识使用 **模型名称、版本** 。
- 不存在「模型更新」的语义；新版本被视作新血缘节点。
- 任务删除时，关联血缘节点同步删除连线或节点。

典型用法

影响分析：修改前评估下游

在准备修改某张事实表前：

1. 进入表详情页 **血缘** Tab。
2. 在列表模式中查看「下游」资产清单：哪些 SQL 文件、Workflow 任务、Dashboard、模型直接依赖这张表。
3. 切换到图模式，进一步查看二级、三级下游。
4. 提前通知相关 owner，并安排回归测试。

数据溯源：定位异常源头

发现某张表数据异常时：

1. 进入表详情页 **血缘** Tab。
2. 在图模式中向上游展开，找到生产该表的任务节点（Notebook / Workflow Task / SQL）。
3. 点击任务节点查看任务代码与最近运行时间。
4. 进入任务详情页排查具体逻辑或运行日志。

模型治理：版本对比与下游评估

在模型版本切换前：

1. 进入模型详情页 **血缘** Tab。

2. 列表模式中查看「下游」推理表与模型服务。
3. 评估切换版本对线上服务的影响。

与其他模块的关系

模块	关系
Workflow	Workflow 任务运行时自动上报血缘，任务节点出现在血缘图中
Studio	SQL 开发、Notebook 在执行 SQL 时自动上报血缘
模型 Catalog	模型节点的元数据来源
数据集成	离线 / 实时同步任务作为血缘节点参与图谱
仪表盘	Dashboard 作为下游血缘节点参与图谱

常见问题

Q：为什么我刚跑完的 SQL 还没出现在血缘图中？

A：血缘采集与图模式渲染存在短暂延迟。请稍等 1-2 分钟后刷新血缘 Tab。如长时间未出现，请检查任务是否成功执行、SQL 中是否使用了三段式 `catalog.schema.table` 路径或当前 SQL 上下文是否设置正确。

Q：测试运行的 SQL 也会产生血缘吗？

A：会。引擎对所有提交执行的 SQL 与作业都会进行血缘解析，包含测试运行。如不希望测试上报，建议使用临时表或单独的测试 Schema。

Q：删除任务后血缘节点会立刻消失吗？

A：平台在任务删除时会上报清理事件，血缘节点通常在短时间内同步清理。

Q：能否查看字段级血缘？

A：当前版本仅支持表级血缘。字段级血缘已在后续迭代规划中。

Q：跨 Catalog 的血缘能否正确展示？

A：可以。血缘节点以三段式 `catalog.schema.object` 唯一标识，跨 Catalog 与跨工作空间的血缘均可在图中展示。是否对您可见，仍受当前账号的访问权限控制。

相关文档

- [结构化数据 Catalog](#)
- [模型 Catalog](#)
- [Studio 概述](#)
- [Workflow](#)
- [仪表盘概述](#)

数据发现概述

全局数据搜索

最近更新时间：2026-10-09 16:26:47

概述

全局数据搜索 是数据发现模块的核心入口，支持以**自然语言** 或**关键字** 在工作空间范围内一键定位所需资产。AI 语义检索与关键字检索并行执行，结果合并去重后按综合相关度排序，帮助您在不知道精确名称的情况下也能找到合适的表、Notebook、工作流、仪表盘、指标、智能体等。

本文介绍如何发起全局搜索、阅读搜索结果、使用 AI 语义回答等能力。资产类型 Tab、负责人、时间等筛选维度的详细说明请参见 [按标签、分类筛选](#)。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号已加入目标 Workspace（角色权限说明详见 [成员与权限管理](#)）。
- 搜索结果按 ACL 权限过滤，仅展示当前账号有权查看的资产。

使用限制

限制项	说明
搜索范围	仅限当前 Workspace 内的资产，不跨工作空间检索
检索粒度	表名、视图名、列名、对象描述、注释，以及指标 / 语义模型 / 智能体的名称与描述等
对话轮次	AI 语义回答当前仅支持单轮对话，不保留多轮上下文
默认结果数	搜索结果默认展示前 20 条，向下滚动加载更多

操作步骤

步骤 1：发起搜索

DataBuddy 提供两种发起搜索的方式：

- **顶部全局搜索框**：在控制台任意页面顶部搜索框，输入搜索内容后回车，自动跳转到全屏数据发现页。
- **左侧导航 数据发现**：点击**数据发现**，进入全屏页面后，在顶部搜索框输入内容。

搜索内容支持两种形式：

形式	示例	适用场景
----	----	------

关键字	user_info 、 olist_orders	已知精确表名、列名或对象名
自然语言	用户订单相关的表 、 最近 30 天活跃的工作流	不确定精确名称、希望按业务意图检索

后端会先判断输入是否能精准命中现有对象：能命中则直接返回精确结果，否则回退到 AI 语义检索。两路检索可能并行进行，结果合并展示。

步骤 2：阅读 AI 语义回答

当输入更接近自然语言提问时，结果区顶部会先出现 AI 语义回答：

- AI 回答以逐字流式输出的方式呈现，期间显示加载动画。
- 回答内容达到默认折叠长度时会显示 展开 按钮，点击查看完整回答。
- AI 回答下方会列出与问题相关的资产卡片，点击卡片可跳转到对应详情页。

i

提示

AI 语义回答仅在自然语言搜索时出现；输入精确关键字（如完整表名）时通常会跳过 AI 回答，直接展示精确匹配结果。

步骤 3：阅读搜索结果列表

每条搜索结果以卡片形式展示，包含的字段如下：

字段	说明
资产名称	点击跳转到对应资产详情页
资产类型 ICON	标识该资产是表 / 视图 / 工作流 / 仪表盘等
收藏标识	如该资产已被当前用户收藏，显示黄色星标
热度	暂以浏览次数表示，长期将切换为被查询次数
Owner	资产负责人
最近更新	资产最近一次更新时间，精度到天
最近浏览	当前用户最近 30 天内最后一次浏览的相对时间，例如「最近浏览 2 天前」；从未浏览过则不展示

结果默认按综合相关度排序：

排序因子	权重	说明
------	----	----

语义相关度	50%	AI 向量检索的余弦相似度
关键字匹配度	30%	表名 / 列名 / 注释的精确与模糊匹配得分
使用热度	20%	近 30 天的查询 / 访问次数

页面默认展示前 20 条结果，向下滚动可加载更多。

步骤 4：跳转到资产详情页

点击任意结果卡片，将进入对应资产的详情页，从详情页可以继续：

- 查看资产的 Schema、Owner、说明、血缘、质量监控等元信息。
- 查看预览数据（适用于表、视图、文件等）。
- 申请或管理权限。
- 收藏、取消收藏，以便后续在 **快速开始 – 收藏** 中快速返回。

不同资产的详情页字段略有差异，详情请参见各资产模块文档。

在「快速开始」中复用搜索

数据发现的搜索结果会与首页的「快速开始」联动：

- **快速开始 – 最近**：展示当前用户最近访问过的指标、语义模型、智能体、表等资产，便于继续上次的工作。
- **快速开始 – 收藏**：展示当前用户已收藏的指标、语义模型、表等资产，可在卡片上一键取消收藏。

点击「最近」或「收藏」中的资产卡片，同样跳转到对应详情页。

验证结果

完成一次搜索后，您可以通过以下方式确认搜索能力可用：

1. 在顶部输入业务关键词（例如 **订单**），结果列表中至少返回 1 条相关资产。
2. 切换到自然语言提问（例如 **有哪些订单相关的表**），AI 语义回答区域出现流式输出的回答与相关资产卡片。
3. 点击任意一条卡片，能成功跳转到对应资产详情页。

常见问题

Q：为什么搜索结果中看不到我知道存在的某张表？

A：可能原因如下：

- 当前账号没有该表的查看权限：搜索结果按 ACL 权限过滤，无权限的对象不会展示。
- 当前 Workspace 不是该表所在的 Workspace：数据发现仅在当前工作空间内检索，跨工作空间需切换后再搜。
- 该表所属的 Catalog 还未完成元数据采集：External Catalog 的元数据需要先完成同步，详见 [外部数据管理](#)。

Q：AI 语义回答与下方资产列表的关系？

A: AI 语义回答用自然语言总结当前问题相关的资产线索；下方的资产列表则是按综合相关度排序的具体资产卡片，便于您直接点击进入。两者互补，回答提供解释、列表提供入口。

Q：搜索结果排序是否可以自定义？

A：当前版本固定按综合相关度排序，暂不支持自定义排序权重。后续将持续优化排序策略。

Q：能否多轮对话式细化检索？

A：当前版本只支持单轮对话；如需多轮上下文交互，请使用 [Buddy](#)。

排序与统计口径补充说明

- **排序权重固定**：综合相关度的语义 50% / 关键字 30% / 热度 20% 在所有地域、所有用户群中保持一致，不会随地域或灰度策略动态调整。
- **使用热度统计口径**：所有资产类型的「使用热度」统一采用 **近 30 天的查询、访问次数**，与排序因子表中的描述保持一致。
- **AI 搜索建议**：首批已对外开放，输入时会即时推荐相关查询词。
- **搜索历史 & 热门搜索**：当前版本暂未对外开放，将在后续版本中逐步上线。

相关文档

- [按标签、分类筛选](#)
- [统一元数据管理架构](#)
- [Buddy 概述](#)

分类筛选

最近更新时间：2026-10-09 16:26:51

概述

在数据发现页面中，您可以在搜索结果上叠加多种**筛选条件**，进一步收敛资产范围。常用的筛选维度包括：资产类型分类（左侧 Tab）、负责人、时间范围，以及部分资产类型上额外可用的 Catalog、治理标签等维度。本文介绍如何使用这些筛选维度，以及不同资产类型支持的筛选项差异。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号已加入目标 Workspace（角色权限说明详见 [成员与权限管理](#)）。
- 如需使用治理标签维度筛选，工作空间内需先建立标签体系并对资产打标，详见 [创建与管理标签体系](#)、[为数据资产打标签](#)。

使用限制

限制项	说明
筛选范围	仅作用于当前选中的资产类型 Tab 内
筛选生效方式	多个筛选条件之间为「与」关系，需同时满足
重置	提供 重置 按钮一键清除当前 Tab 下的所有筛选条件

操作步骤

步骤 1：选择资产类型 Tab

进入数据发现页（顶部搜索框或左侧导航 **发现**），在页面左侧选择目标资产类型 Tab。每个 Tab 对应一类资产，结果列表会自动收敛到该类型。

数据发现页支持的资产类型 Tab 如下：

分组	Tab
综合	全部
数据目录	表、视图、函数、卷、模型
数据工程	文件、Notebook、SQL、文件夹、Git 文件夹、工作流
数据接入	离线接入、实时接入

数据科学	模型实验、模型服务
数据分析	仪表盘
数据治理	指标、语义模型
数据应用	智能体（Agent）

 提示

Tab 「全部」会同时展示所有类型的搜索结果，结果卡片以资产类型 ICON 标识来源；其他 Tab 仅展示对应类型。

步骤 2：使用顶部通用筛选

切换 Tab 后，顶部筛选条会展示该 Tab 支持的筛选维度。所有 Tab 都支持以下两个**通用筛选**（工作流除外）：

- **负责人**：按资产 Owner 过滤，可多选。
- **时间**：按「最近更新时间」或「创建时间」过滤，例如最近一周、最近一月、最近三月。

操作方式：

1. 点击 **全部负责人** 下拉框，勾选目标负责人后确认。
2. 点击 **全部时间** 下拉框，选择预设区间。
3. 如需清除当前 Tab 下的所有筛选条件，点击 **重置**。

步骤 3：使用资产类型专属筛选

部分资产类型在通用筛选之外，还提供额外的筛选维度。常见差异如下：

Tab	额外支持的筛选维度
表 / 视图 / 函数 / 卷	负责人、时间、Catalog、标签
模型 / 模型服务/ 指标 / 语义模型	负责人、时间
智能体（Agent）	负责人、时间、位置（文件卷位置）
文件 / Notebook / SQL / 文件夹 / Git 文件夹	负责人、时间、位置（文件卷位置）
工作流	负责人
离线接入 / 实时接入	负责人、时间
模型实验 / 仪表盘 /	负责人、时间、位置（文件卷位置）

操作方式：

1. 在目标 Tab 顶部选择对应筛选项。

2. 多个维度之间为「与」关系：例如同时设置 负责人 = alice 与 治理标签 = 含敏感数据，仅展示同时满足两个条件的资产。
3. 点击 重置 清除当前 Tab 下的全部筛选。

步骤 4：结合搜索关键字一起使用

筛选条件与搜索框可以叠加使用：

1. 在顶部搜索框输入关键字或自然语言。
2. 在左侧 Tab 切到目标资产类型。
3. 在顶部筛选条中追加负责人、时间、治理标签等条件。

最终的结果列表是「关键字检索结果 $\cap$ 当前 Tab 类型 $\cap$ 顶部所有筛选条件」的交集，按综合相关度排序展示。

筛选维度说明

资产类型分类

由 DataBuddy 内置的资产体系决定，无需额外配置。Tab 列表会随产品演进新增资产类型（例如未来的 External Volume 等）。

负责人

- 数据来源：资产创建时记录的 Owner，或通过资产详情页变更后的 Owner。
- 多选：支持同时选择多个负责人，结果取并集。

时间

- 默认按「最近更新时间」过滤。
- 支持按照最近一周、最近一月、最近三月过滤

Catalog / Schema（仅 Catalog 类资产可用）

- 来源于 [统一元数据管理架构](#) 中的 Catalog / Schema / Table 三级命名空间。
- 选择 Catalog 后，Schema 下拉会自动联动，仅展示该 Catalog 下的 Schema。

治理标签（仅 Catalog 类资产可用）

- 来源于 [治理标签](#) 模块。
- 仅当某个标签已被实际应用资产时才会出现在筛选下拉中。
- 多选时取并集（任一标签命中即纳入结果）。

验证结果

完成筛选后，您可以通过以下方式确认筛选是否生效：

1. 切换到 表 Tab，仅设置 负责人 = 当前用户，结果列表中所有卡片的 Owner 字段均为当前用户。

2. 在 表 Tab 中追加治理标签筛选（例如 `pii`），结果列表中所有表均带有 `pii` 标签。
3. 点击 重置，所有筛选条件清空，结果列表恢复为该 Tab 的默认搜索结果。

常见问题

Q：为什么我在 指标 Tab 找不到 Catalog 筛选？

A：当前版本中，指标、语义模型、智能体在数据发现中以独立资产存在，暂不依附于 Catalog 三级命名空间，因此不提供 Catalog 与治理标签筛选项。如需按业务领域分组筛选，可在对应模块（如 [语义模型](#)）内部使用领域、分组能力。

Q：治理标签筛选下拉中为什么只看到部分标签？

A：筛选下拉只展示已被实际打到当前 Tab 资产上的标签，避免出现「选了但搜不到结果」的体验。如果某个新标签尚未应用到资产，需要先在 [为数据资产打标签](#) 中完成打标。

Q：多个筛选条件是「与」还是「或」？

A：不同维度之间是「与」关系（同时满足），单个维度内的多个值是「或」关系（任一命中即可）。例如：负责人多选 `alice` 与 `bob`、治理标签选择 `pii`，则结果为「(Owner=alice 或 Owner=bob) 且 标签包含 pii」的资产。

Q：筛选条件能否在不同 Tab 间共享？

A：当前版本中，重置 仅清除当前 Tab 的筛选；切换 Tab 时会保留通用维度（负责人、时间），但不通用的维度（Catalog、标签）会按目标 Tab 的支持情况自动收敛或清除。

行为与限制补充

- **Tab 切换行为**：切换 Tab 时，负责人、时间 等通用筛选会自动保留；切到不支持 Catalog 的 Tab 时，原有 Catalog / 治理标签筛选会被自动清空。
- **时间筛选默认字段**：所有 Tab 的「时间」筛选默认对应「最近更新时间」。
- **基础筛选范围**：当前版本仅支持上文列出的基础筛选维度，不再扩展工作流的「调度状态」、仪表盘的「发布状态」等业务态字段。
- **多选无上限**：单维度筛选（如负责人）支持的多选数量无硬上限。

相关文档

- [全局数据搜索](#)
- [统一元数据管理架构](#)
- [创建与管理标签体系](#)
- [为数据资产打标签](#)

数据安全概述

数据分类分级

最近更新时间：2026-10-14 11:16:12

概述

数据分类分级用于识别 Catalog 中字段所承载的敏感信息（例如手机号、身份证号、邮箱、银行卡号等），并基于识别结果统一推动列级脱敏、行级过滤、访问审计等下游策略落地。DataBuddy 提供两类能力，配套同一套敏感标签与脱敏策略元数据：

- **敏感识别结果写入 Skill**：当 AI 在 Catalog 字段上识别出敏感数据后，由 Skill 把识别结论（敏感标签、置信度、建议脱敏策略）写回到对应字段，作为列级脱敏的依据。
- **脱敏策略管理 Skill**：用于在 Buddy 对话中按业务语义创建脱敏策略、维护用户组到脱敏方法的映射，并把策略关联到一个或多个敏感标签。

说明：

当前版本中，数据分类分级以 **Skill 接口形式** 提供，由 Buddy 智能体和 AI 识别链路调用。可视化的「分类分级管理页面」将在后续迭代上线，本文以 Skill 行为为主。脱敏策略创建后的查询体验、字段绑定方式与列级脱敏一致，详见 [列级权限控制](#)。

核心概念

概念	说明
敏感标签	描述字段中数据敏感类型的元数据标签，例如 <code>PII/手机号</code> 、 <code>PII/身份证号</code> 、 <code>Finance/银行卡号</code> 。所有标签共享同一套租户级命名空间。
脱敏策略	描述字段被读取时按用户组应用的脱敏方法集合。一个脱敏策略可关联多个敏感标签。
脱敏方法	系统内置的字段值变换方式（例如掩码、Hash、保留首/末位）。脱敏方法按用户组生效，未匹配的用户组返回明文。
敏感识别结果	AI 模型对某 Catalog 字段给出的敏感判定，包含敏感标签、置信度，以及推荐绑定的脱敏策略。
字段绑定	把脱敏策略关联到具体 Catalog → Schema → Table → Column 的动作。一个字段同一时间仅绑定一个脱敏策略。

前提条件

- 当前账号已能登录工作空间，并具备 Buddy 使用权限。

- 维护脱敏策略时，当前账号在目标 Catalog 中具备策略管理权限（默认为 Workspace 管理员、数据安全管理员）。
- 写入敏感识别结果时，调用方在目标 Table 上拥有 **Owner** 或 **Manage** 权限。权限语义参考 [成员与权限管理](#)。
- 已规划好本工作空间的敏感标签命名规范（例如统一前缀 `PII/`、`Finance/`、`Internal/`），避免相同语义出现多份标签。

使用限制

限制项	说明
适用对象	敏感标签、脱敏策略针对当前租户 + 当前地域内的所有 Catalog 共享。
字段绑定关系	一个字段同一时间仅可绑定一个脱敏策略；多次写入将以最近一次结果为准。
用户组覆盖	脱敏方法按用户组生效；未在策略中配置的用户组查询该字段时返回明文。
脱敏方法范围	当前提供系统内置脱敏方法，暂不支持自定义脱敏函数。
调用方式	当前版本仅通过 Skill 接口维护与写入分类分级结果，可视化管理页面将在后续上线。

使用方式

维护脱敏策略：脱敏策略管理 Skill

脱敏策略管理 Skill 用于在 Buddy 对话中：

- 创建、更新、删除脱敏策略。
- 维护脱敏策略的关联用户组与脱敏方法。
- 把一条脱敏策略关联到一个或多个敏感标签。

典型调用语义如下（实际 Skill 名与参数以 Buddy 实际暴露的版本为准）：

为「客户手机号」创建脱敏策略：

- 关联敏感标签：`PII/手机号`
- 用户组「客服一线」：保留首尾各 3 位、其余替换为 *
- 用户组「数据分析师」：仅返回 Hash 结果
- 其他用户组：默认返回明文

Buddy 收到该意图后，会调用脱敏策略管理 Skill 创建策略并返回策略 ID，可在后续敏感识别结果写入时引用。

说明：

策略一旦创建，与列级脱敏共用同一套语义；后续在 Catalog 表详情页的字段绑定、SQL / Notebook 查询时的运行时脱敏行为，请参考 [列级权限控制](#)。

写入敏感识别结果：敏感识别结果写入 Skill

敏感识别结果写入 Skill 用于把 AI 模型的敏感识别输出固化到 Catalog 字段元数据上。典型流程：

1. AI 识别链路对 Catalog 中的目标字段进行扫描，得到候选敏感标签和置信度。
2. 业务侧（Buddy 或外部流程）按阈值筛选可信的识别结果。
3. 调用敏感识别结果写入 Skill，将以下信息写回字段：
 - 敏感标签（如 PII/手机号）。
 - 置信度（可选，用于审计与后续复核）。
 - 推荐脱敏策略（可选，可直接引用脱敏策略管理 Skill 创建的策略 ID）。
4. 写入成功后，下游列级脱敏、行级过滤、访问审计逻辑均按最新的敏感标签、脱敏策略生效。

调用约定建议：

约定	建议做法
标签命名	统一沿用工作空间已规划的标签命名空间，避免出现 phone / mobile / PII/手机号 并存。
写入幂等性	对同一字段多次写入时，以本次写入的标签和策略为准；如需追溯历史结果，请在调用方记录调用日志。
异常处理	写入失败时由调用方做重试或人工复核，避免静默丢失敏感识别结论。

常见问题

Q：AI 识别给出的敏感标签可信度不高怎么办？

A：建议在敏感识别结果写入 Skill 之前由业务侧按置信度阈值过滤，仅写入达到阈值的标签；对低置信度结果可走人工复核流程，复核通过后再写入。

Q：字段已绑定脱敏策略，再次写入新的识别结果会怎样？

A：以最近一次写入为准，原绑定关系会被覆盖。如果业务上需要保留历史绑定记录，请在调用方自行记录变更日志。

Q：脱敏策略中未覆盖的用户组查询字段时会返回什么？

A：默认返回明文。若需要对未覆盖用户组强制脱敏，请在策略中显式追加用户组并指定脱敏方法。

Q：可视化的分类分级管理页面什么时候上线？

A：当前版本只提供 Skill 接口；可视化页面在后续迭代规划中。在管理页面上线前，建议通过 Buddy 对话或自动化流程统一调用 Skill 维护标签与策略。

相关文档

- 列级权限控制
- 行级权限控制
- 为数据资产打标签
- 基于标签的策略控制
- 成员与权限管理

列级权限控制

最近更新时间：2026-10-09 16:27:01

概述

列级权限控制基于「脱敏策略」实现：您可以在 **数据安全** 菜单中创建脱敏策略，按用户组定义不同的脱敏方法（例如管理员看明文、客服看中间 4 位掩码、开发看后 4 位），并在 Catalog 表详情页将策略关联到敏感字段。所有用户在 SQL、Notebook、仪表盘等查询链路中读取该字段时，都会按其所属用户组返回相应的脱敏结果。

说明：

本节描述的「列级权限控制」实质是基于脱敏策略的字段级数据保护，用于控制不同用户在查询同一字段时看到的内容；并非「用户能否查询某列」的访问控制。如需限制用户对表、字段的查询权限本身，请参考 [成员与权限管理](#) 中的资产 ACL 章节。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- 创建脱敏策略：当前账号已能登录工作空间。
- 为字段绑定脱敏策略：当前账号在目标表所在 Catalog 中拥有 Table 的 **Owner** 或 **Manage** 权限（权限语义详见 [成员与权限管理](#)）。
- 用户组信息已在工作空间内创建完毕。

使用限制

限制项	说明
适用对象	脱敏策略针对当前租户 + 当前地域内的所有 Catalog 共享。
字段绑定关系	一个字段同一时间仅可绑定一个脱敏策略。
用户组覆盖	脱敏方法按用户组生效；未在策略中配置的用户组在查询该字段时返回明文。
脱敏方法范围	当前提供系统内置脱敏方法，暂不支持自定义。

操作步骤

数据脱敏由两步组成：先在 **数据安全** 菜单中创建脱敏策略，再到目标表详情页将策略关联到字段。

步骤 1：进入数据安全管理工作面

1. 登录 [DataBuddy 控制台](#)。

2. 在顶部切换到目标地域和 Workspace。
3. 在左侧导航栏选择 **数据治理 > 数据安全**。
4. 默认进入 **脱敏策略** 列表页，可在搜索框中按策略名称筛选已有策略。

步骤 2：创建脱敏策略

1. 在脱敏策略列表页右上方点击 **+ 创建策略**。
2. 在弹出的创建抽屉中填写 **基本信息**：
 - **策略名称**：必填。
 - **描述**：选填，建议用于说明策略的业务含义。
 - **关联敏感标签**：选填。关联后，AI 识别到带有该标签的字段时会自动匹配此脱敏策略；不关联则保持手动绑定。
3. 在 **脱敏方法** 区域为不同用户组配置不同的脱敏方法：
 - **用户组**：从下拉中选择该条规则生效的用户组，支持多选。
 - **脱敏方法**：从下拉中选择系统内置的脱敏方法。
 - 通过列表底部的「+ 添加」按钮可继续追加用户组与脱敏方法的组合。
4. 全部信息填写完成后，点击 **确定** 保存策略，返回策略列表。

⚠ 注意：

未在脱敏策略中配置的用户组，查询该字段时仍会返回明文。请在策略中显式覆盖所有需要控制的用户组。

步骤 3：在表详情页为字段绑定策略

1. 在左侧导航栏选择 **数据目录**。
2. 进入目标 Catalog → Schema → 目标表的详情页，切换到 **概览** Tab。
3. 在字段列表中选中需要脱敏的字段，进入字段编辑模式。
4. 在「脱敏策略」字段中，从下拉中选择已创建的脱敏策略。如果列表中没有合适的策略，可在下拉中点击 **+ 创建策略**，在当前页面打开创建抽屉新建策略，创建完成后自动返回选择。
5. 保存字段配置。

步骤 4：在表详情中查看与维护已绑定的脱敏策略

1. 在表详情页的字段列表中，每个字段会展示当前已绑定的脱敏策略名称。
2. 点击策略名称，跳转到策略查看界面，可查看策略名称、描述、各用户组的脱敏方法。

步骤 5：管理脱敏策略

回到 **数据安全 > 脱敏策略** 列表页：

操作	说明
----	----

查看	在策略详情中查看用户组与脱敏方法明细。
编辑	修改策略名称、描述、关联敏感标签、用户组与脱敏方法组合。修改后查询时立即按新策略生效。
删除	删除策略。删除后该策略关联的字段会失去脱敏保护，所有用户在查询时返回明文。

参数说明

创建脱敏策略

参数	说明	是否必填	默认值
策略名称	脱敏策略的唯一名称，用于在策略列表与字段绑定下拉中识别。	是	—
描述	策略的文字说明，最长 512 字符。	否	—
关联敏感标签	选择一个敏感标签后，AI 识别到带有该标签的字段会自动匹配此脱敏策略。不关联则保持手动绑定模式。	否	不关联（手动关联）
用户组	选择当前条脱敏方法生效的用户组，支持多选。	是	—
脱敏方法	选择系统内置的脱敏方法。同一策略中可针对不同用户组合多条规则。	是	—

字段绑定脱敏策略

参数	说明	是否必填	默认值
脱敏策略	在表详情字段编辑面板中选择已创建的脱敏策略。	否	不绑定（返回明文）

验证结果

操作完成后，您可以通过以下方式确认：

- 在 **数据安全 > 脱敏策略** 列表中，可以看到刚创建的策略，关联用户组、创建人、创建时间显示正确。
- 在 Catalog 表详情页中，目标字段已显示对应的脱敏策略名称。
- 使用属于不同用户组的账号分别在 Studio 中执行 `SELECT` 语句查询该字段，结果按各自用户组对应的脱敏方法返回；策略覆盖范围之外的用户组返回明文。

常见问题

Q：删除脱敏策略后，已经绑定该策略的字段会怎么样？

A：删除策略后，已关联该策略的字段失去脱敏保护，所有用户查询时返回明文。建议在删除前先在表详情页解除绑定，并确认没有业务影响。

Q：未在策略中配置的用户组查询字段时返回什么？

A：返回字段的真实明文值。脱敏策略只对策略中显式列出的用户组生效；如需对其他用户组也实施保护，需要在策略中追加相应的用户组与脱敏方法。

Q：同一字段需要不同用户组看到不同结果，应该怎么配置？

A：在同一脱敏策略中为不同用户组分别添加一条「用户组、脱敏方法」记录即可，无需为同一字段创建多个策略（且每个字段也只能绑定一个策略）。

Q：子账号没有权限向他人授予 grant 权限怎么办？

如图，当一个子账号被授予 grant 权限后，在向其他用户授权时提示无权限：

此时，管理员可以进入元数据视图，点击导航栏数据目录，再点击 Catalog 右侧的设置按钮，选中元数据湖，再点击页面右上方的 +授权按钮，为该子账号授予 Admin：

随后再次让子账号尝试操作即可。

相关文档

[成员与权限管理](#)

行级权限控制（Row Filter）

最近更新时间：2026-10-13 11:05:02

概述

行级权限控制（Row Filter）允许您在数据表详情页对单张表配置行过滤规则，按授权主体（用户、用户组）声明 `WHERE` 过滤表达式，限制不同主体在同一张表中可见的行范围。配置后，受控表的查询请求在所有 SQL、Notebook、仪表盘等鉴权场景下都会自动追加对应的过滤条件，仅返回满足条件的行。

典型场景：一张包含全国订单的事实表中，希望华南区运营人员只能看到 `region = 'south'` 的订单，华北区运营人员只能看到 `region = 'north'` 的订单，运营总监能看到全部行——可以通过为不同用户组分别添加行过滤表达式实现。

前提条件

- 在开始本操作前，请确保满足以下条件：
- 已开通 DataBuddy 服务并完成空间初始化。
 - 当前账号在目标表上拥有 **授权（Grant）** 权限（即在「+ 授权」按钮上具备权限），权限语义详见 [成员与权限管理](#)。
 - 已确定行过滤的授权对象（用户、用户组），用户组在工作空间内已创建完毕。

使用限制

限制项	说明
适用对象	行级权限规则按表配置，作用于该表。
授权主体类型	支持 用户 与 用户组 ；同一主体出现在用户层与用户组层时，权限范围取并集。
表达式格式	行过滤表达式为标准 SQL <code>WHERE</code> 子句片段，例如 <code>year > 2010 AND country != 'US'</code> 。
表达式长度	单条行过滤表达式最大长度为 2000 字符 。
规则数量	单张表可配置的行权限规则数量无硬上限。
校验项	表达式支持一键校验，包括字段是否存在、逻辑、语法等基础检查；校验失败需修正后才能保存。
Schema 变更行为	当行过滤表达式引用的字段在表 Schema 中被重命名或删除后，相关行权限规则 自动失效 ；需重新编辑表达式才能恢复行权限管控。
用户组成员	用户组成员新增 / 移除后， 已经在执行中的查询不会实时切换 可见行集合；变更对后续新发

变更	起的查询生效。
权限继承	行权限 Tab 的可见性继承表详情页的 权限 Tab 可见性；+ 授权 > 行权限 按钮的可点击性继承 + 授权 按钮本身的权限。
配套能力范围	列权限、ABAC 策略、动态视图等高级安全能力当前版本暂未对外开放，将在后续版本中逐步上线。

操作步骤

步骤 1：进入数据表详情页

1. 登录 [DataBuddy 控制台](#)。
2. 在顶部切换到目标地域和 Workspace。
3. 在左侧导航栏选择 **数据目录**。
4. 在 Catalog → Schema 中找到目标表，进入其详情页。

步骤 2：进入行权限管理面板

1. 点击“授权”按钮，下拉中会出现 **数据表权限** 与 **行权限** 两项。
2. 点击 **行权限**，进入该表的行权限配置面板。

步骤 3：添加行权限规则

1. 在行权限管理面板点击 **+ 授权 > 行权限**，进入行级权限配置界面。
2. 选择 **授权对象**：在「用户」「用户组」之间切换，支持模糊搜索，可多选、全选。
3. 输入 **行过滤表达式**：填写 `WHERE` 子句片段，例如 `year > 2010 AND country != 'US'`。
4. 先选择计算资源，然后点击 **校验** 验证表达式是否合法（涵盖字段是否存在、逻辑与语法）。校验未通过时按提示修正后再保存。
5. 校验通过后点击**确定** 保存。

① 说明：

对同一张表为多个授权主体添加规则时，配置面板会按用户、用户组明细各保存一条；后续编辑、移除均针对单条记录单独操作。

步骤 4：编辑或移除行权限规则

1. 在行权限列表中定位到目标记录。
2. 点击 **编辑** 修改行过滤表达式，修改后再次校验并保存。
3. 点击 **移除** 后，在弹窗中确认删除；确认后该条规则立即失效。

参数说明

参数	说明	是否必填	默认值	取值范围
授权对象	选择行规则作用的主体，可在「用户」「用户组」之间切换，支持模糊搜索、多选与全选。	是	-	当前工作空间内已存在的用户 / 用户组
行过滤表达式	标准 SQL <code>WHERE</code> 子句片段。引用的字段必须为该表的真实字段。最大长度 2000 字符。	是	-	例如 <code>region = 'south'</code> 、 <code>year > 2010 AND country != 'US'</code>

验证结果

操作完成后，您可以通过以下方式确认：

1. 在表详情页 **行权限** 面板列表中，可以看到刚保存的规则；授权主体、主体类型与表达式均显示正确。
2. 使用对应主体的账号在 Studio 中执行 `SELECT * FROM <表名>`，返回的行集合仅包含表达式条件成立的行。
3. 切换到未配置行权限的主体（且具有表查询权限），仍可查询全部行；该机制保证仅显式配置过规则的主体被收敛访问范围。

常见问题

Q：同一个用户既属于一个用户组、又被单独授予行权限，最终能看到哪些行？

A：取所有规则的并集。例如用户组规则为 `region = 'south'`、用户规则为 `vip = true`，该用户最终可见的行是「`region = 'south'` 或 `vip = true`」的所有记录。

Q：行过滤表达式支持哪些语法？

A：使用标准 SQL `WHERE` 子句片段。可使用比较运算符（`=` `!=` `>` `<` `>=` `<=`）、逻辑运算符（`AND` `OR` `NOT`）、`IN` 列表、`BETWEEN`、`LIKE` 等常见语法。具体函数和算子的支持范围以底层引擎实际能力为准，建议在编辑时使用 **校验** 功能验证表达式是否合法。

Q：删除某条行权限规则会立刻生效吗？

A：移除后立即对后续查询生效。已经在执行中的查询不会被强制中断。

Q：行级权限对哪些查询入口生效？

A：完成行过滤器配置后，受控表会按行过滤表达式逻辑控制行可见性，作用于所有需要鉴权的 SQL 查询场景（包括 Studio SQL、Notebook、仪表盘、工作流中的 SQL 任务等）。

相关文档

- 列级权限控制
- [成员与权限管理](#)

数据变更历史

最近更新时间：2026-10-09 16:27:10

概述

数据表的变更历史用于回答「谁在什么时候对这张表做了什么改动」。在 Catalog 中打开任一表详情页后，**变更历史** Tab 会按版本倒序展示该表的所有数据与 DDL 操作记录，并支持按用户、操作类型、时间范围筛选。DataBuddy 同时支持在 Spark SQL 中通过 `DESCRIBE HISTORY` 直接查询表的变更历史，便于在 SQL / Notebook 链路中程式化审计。

适用场景：

- 排查表数据为何被修改、被谁修改、什么时间修改。
- 验证近期 DDL 变更（如新增列、删除分区、重写表结构）是否符合预期。
- 在事故复盘、合规审计、版本追溯时定位具体的版本号与提交记录。

前提条件

- 当前账号已能登录工作空间，并对目标 Catalog → Schema → Table 拥有 **Read** 及以上权限。
- 目标表所在 Catalog 已绑定支持读取表历史的计算引擎（详见 [使用限制](#)）。
- 在 Catalog 表详情页查看变更历史时，需要先选择一个可用的计算资源用于拉取历史数据。

使用限制

限制项	说明
引擎依赖	表变更历史读取依赖底层 Catalog 实现；不同引擎支持的操作类型与字段范围可能存在差异。
记录条数	页面最多展示前 100 条变更记录；如需更多记录，请使用 SQL 中的 <code>DESCRIBE HISTORY</code> 查询。
展示字段	当前展示版本号、时间戳、用户 ID、用户名、操作、操作参数；后续将逐步扩展（如作业、Notebook、Cluster ID、版本号、事务隔离级别、是否盲追加、操作指标、用户自定义备注、引擎信息等）。
计算资源	进入变更历史 Tab 时必须选择计算资源；未启动时会先触发启动，启动期间历史数据不可读取。

操作步骤

步骤 1：进入表详情页

1. 登录 [DataBuddy 控制台](#)。

2. 在顶部切换到目标地域和 Workspace。
3. 在左侧导航栏选择 **数据目录**。
4. 依次展开 **Catalog > Schema**，点击目标表名进入表详情页。

步骤 2：选择计算资源

1. 在表详情页切换到 **变更历史** Tab。
2. 如当前计算资源未启动，按页面提示在选择框中选取一个可用计算资源：
 - 计算资源选择框支持搜索。
 - 资源状态用颜色和图标标识：绿色为运行中，灰色为待启动，圆圈为启动中；鼠标悬停可查看完整名称和状态描述。
3. 选择后系统会提示 **计算资源正在启动中...更多信息请到计算资源查看**，可点击资源名称跳转到 **计算资源** 页面查看启动详情。
4. 启动完成后会提示 **计算资源已经启动。更多信息请到计算资源查看**，此时可以读取变更历史。

① 说明：

计算资源选择具有记忆功能：再次打开 **变更历史** Tab 时，会自动沿用您上一次的选择。

步骤 3：查看变更历史

变更历史以表格形式按版本倒序展示，常用字段如下：

字段	说明
版本	表的版本号，从大到小排列，最新版本在最上方。
时间	该版本提交的时间戳，按最新时间在前展示。
变更类型	该次变更对应的操作类型，例如 <code>INSERT</code> 、 <code>UPDATE</code> 、 <code>DELETE</code> 、 <code>MERGE</code> 、 <code>CREATE TABLE</code> 、 <code>DROP TABLE</code> 、 <code>TRUNCATE</code> 、 <code>RESTORE</code> 、 <code>ALTER TABLE</code> 等。
用户 uin	触发变更的账号 ID 与显示名。
操作 / 操作参数	引擎记录的具体操作与参数。
操作详情	详细描述，支持点击三角形展开查看。

表格支持横向滚动，向右滑动可看到完整字段。当变更记录较多时，页面底部会提示：**最多展示前 100 条变更记录**。如果需要更多变更记录，请到 **SQL 查询** 页面通过 `DESCRIBE HISTORY` 进行查询。

步骤 4：按用户或操作类型筛选

1. 在表格上方筛选：

- **用户**：默认 **全部用户**，可选择单个或多个用户进行过滤。
2. 如需按时间过滤，可在时间范围控件中选择日期。
 3. 修改筛选条件后，表格会自动刷新；可使用浏览器返回或重新选择 **全部用户 / 所有操作** 恢复默认视图。

步骤 5：在 SQL 中查询表变更历史

如需在 SQL / Notebook 中程式化读取变更历史，可在已绑定对应引擎的查询会话中执行：

```
DESCRIBE HISTORY databricks_test.logantest.`客户信息`;
```

将 `databricks_test.logantest.客户信息` 替换为您的 `catalog.schema.table` 路径。表名包含中文、空格或保留字时建议用反引号包裹。返回结果与表详情页的字段一致，可继续配合 `WHERE`、`ORDER BY`、`LIMIT` 等子句做进一步过滤。

① 说明：

`DESCRIBE HISTORY` 的具体支持范围与返回字段由底层引擎决定，请先确认会话使用的引擎已支持该语法。

常见问题

Q：变更历史中看不到某次写入或 DDL 操作？

A：请按顺序确认：

1. 当前 Catalog / 引擎是否支持记录该类型的操作（部分引擎不会记录所有 DDL）。
2. 是否被前 100 条限制截断；如截断，请改用 `DESCRIBE HISTORY` 查询。
3. 该操作是否发生在所选时间范围（最多 30 天）之外。

Q：进入 变更历史 Tab 后一直显示「请选择计算资源」？

A：说明当前会话没有可用的计算资源。请先在 **计算资源** 页面启动一个可用资源，或在选择框中选择处于「运行中」或「启动中」状态的资源。

Q：可以查看某次变更的具体数据差异吗？

A：当前版本仅展示版本号、操作、操作参数、变更日志等元信息，具体行级差异需通过 SQL 在不同版本之间做对比查询（例如借助引擎自身的版本访问语法）。

Q：能否从变更历史回滚到某个历史版本？

A：当前版本未提供页面化的回滚操作；如需回滚，请基于 `DESCRIBE HISTORY` 输出找到目标版本，再使用引擎支持的回滚或 `RESTORE` 语法在 SQL 中完成。

相关文档

- [列级权限控制](#)
- [行级权限控制](#)
- [成员与权限管理](#)

数据质量概述

创建质量任务

最近更新时间：2026-10-09 16:40:48

概述

本文介绍如何创建一个质量任务并在其中配置质量规则。一个质量任务是若干条质量规则的容器，可作用于表或字段，由 YAML 编辑器统一编辑。完成创建后可进入发布、调度与运行阶段。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 [DataBuddy](#) 服务并完成空间初始化。
- 拥有目标表的 Owner 角色，或在目标表上拥有 **Manager** 权限（角色权限说明详见 [成员与权限管理](#)）。无权限用户在创建按钮上 hover 时，会看到提示"仅表 Owner 和有表 Manager 权限的人有权限创建"。
- 已在目标 Workspace 下准备好可用的数据分析型计算资源，用于试运行与调度执行。

使用限制

限制项	说明
创建权限	仅表 Owner 与表 Manager 角色可创建
单任务监控对象	一次只能选择一张表（可对该表的多个字段配置规则）
规则编辑形态	仅支持 YAML 编辑器，UI 表单与 YAML 不并存
规则类型	系统内置模板（表级、字段级）与自定义 SQL 规则

操作步骤

步骤 1：从 Catalog 或独立菜单进入创建页

DataBuddy 提供 4 个创建入口，仅影响**监控对象**的默认值，其余步骤一致。

入口	进入路径	默认行为
表详情页质量 Tab	数据目录 > 选中表 > 质量 Tab > 新建质量任务	跳过「选择对象」，监控对象默认为当前表
数据质量独立菜单 – Catalog 层级	左侧导航 数据质量 > 选中 Catalog > 新建质量任务	进入「选择对象」，已展开当前 Catalog

数据质量独立菜单 – Schema 层级	左侧导航 数据质量 > 选中 Schema > 新建质量任务	进入「选择对象」，已展开当前 Schema
数据质量独立菜单 – Table 层级	左侧导航 数据质量 > 选中 Table > 新建质量任务	跳过「选择对象」，监控对象默认为当前表

步骤 2：选择监控对象

在选择对象面板中选择被监控的表，质量任务一次只支持一张表。选定后进入步骤 3 配置规则。

步骤 3：创建质量规则（YAML）

在对象 & 规则面板的左侧 YAML 编辑器中编写规则。详细参见 质量规则配置指南。

1. 支持 AI 辅助编写规则
2. 单击展开模板列表，选择对应模板名称后点击插入图标，编辑器会插入完整的规则 YAML 片段，包含：规则名称、规则类型、系统模板标识、检测表达式、告警设置。

YAML 规则的关键字段如下：

字段	说明	是否必填	示例
rule_name	规则名称，作为告警与运维列表中的显示名	是	订单 ID 空值检测
template	系统模板标识；自定义 SQL 时填 custom_sql	是	field_null_rate
dimension	质量维度（与 template 配套）：完整性 / 唯一性 / 及时性 / 有效性/一致性/准确性	是	completeness
target.column	字段级规则的目标字段；表级规则可省略	视模板而定	order_id
expression	检测表达式或自定义 SQL	是	value <= 0.01
filter	数据过滤条件，例如限定分区	否	dt = '\${yyyyMMdd}'
alert	规则的告警配置，enable 表示该规则触发后将发送告警通知，默认为 enable	否	-

3. 模板插入后，YAML 中需要您填写的位置会带上 # 后缀，表示此处需要您补充内容。
4. 单击模板列表上方的「帮助文档」链接可查看 YAML 规则配置说明。

步骤 4：填写基本配置

在对象 & 规则面板的右侧 **基本配置** 区域，填写以下字段：

1. **任务名称**：必填，默认值为「质量任务 _ 日期 _ 时间戳」，可手工修改。
2. **负责人**：必填，默认为当前创建人，必须是对该任务所访问的引擎数据具备操作权限的用户。
3. **任务描述**：选填，建议填写监控目的与覆盖业务范围，便于后续运维识别。

步骤 5：告警配置

支持为质量规则配置告警通知，配置完成后，当规则触发时系统将发送告警消息。

1. 通过下拉框选择需要开启告警的规则，YAML 编辑框内容与下拉选择的规则联动同步展示。
2. 为选中的告警规则配置通知渠道；尚未配置通知渠道时，可点击添加渠道，跳转至 [设置通知渠道](#) 页面完成渠道创建。

步骤 6：保存或保存并发布

页面底部提供 3 个操作按钮：

- **保存并发布**：保存当前 YAML 与基本配置，关闭抽屉返回任务列表，并立即弹出发布弹窗。
- **仅保存**：保存为草稿态版本，关闭抽屉返回任务列表。草稿态任务不会被调度，可后续在任务列表上单击发布。
- **取消**：弹出二次确认，确认后丢弃未保存的 YAML 与基本配置变更。

⚠ 注意：

草稿态任务无法在工作流中被调度。在工作流中选择质量任务时，仅会展示该任务的最新发布态版本。

参数说明

任务基本配置

参数	说明	是否必填	默认值
任务名称	任务在列表与告警中的显示名	是	质量任务_日期_时间戳
负责人	任务的所属用户，需对监控表具备操作权限	是	当前创建人
任务描述	任务的业务说明	否	—

YAML 规则字段

参考创建质量规则（YAML）

验证结果

创建任务完成后，您可进入数据质量独立菜单或表详情页 > 质量 Tab 的任务列表中，查看已创建的任务信息。单击任务名称可更新质量规则。

常见问题

Q：为什么我看不到「新建质量任务」按钮？

A：该按钮仅对表 Owner 与具备表 Manager 权限的用户开放。您可以在表详情页的权限信息中查看自己对该表的角色，必要时联系 Owner 授权。

Q：能否一次创建一个任务监控多张表？

A：本期不支持。一个质量任务的监控对象只能是一张表。如果需要批量覆盖多张表，请逐表创建任务，或在 Schema 维度开启 [异常检测](#) 获得零配置的兜底监控。

质量规则配置指南

最近更新时间：2026-10-09 16:27:17

概述

数据质量规则用于对数据表和字段进行自动化质量检测。通过配置规则，系统可以自动监控数据的完整性、准确性、唯一性、一致性、有效性和及时性，并在检测值满足触发条件时产生异常告警。

配置文件支持以下能力：

- 支持 系统模板（system_template）和 自定义 SQL（custom_sql）两种规则类型。
- 支持 表级规则（table_rules）和 字段级规则（field_rules）两种规则级别。
- 可在一个配置文件中定义 多条规则，灵活组合。

核心概念

规则类型

规则类型	标识	说明
系统模板	system_template	使用系统内置的检测模板，只需指定 template_code 即可。系统模板已内置质量维度，用户无需指定 dimension。
自定义 SQL	custom_sql	用户编写自定义 SQL 查询语句，灵活实现各种检测逻辑。需要用户指定 dimension（质量维度）和 custom_sql（SQL 语句）。

规则级别

规则级别	配置节点	说明
表级规则	table_rules	针对整张表进行检测，如行数检查、数据产出及时性检查等。
字段级规则	field_rules	针对表中的某个具体字段进行检测，如空值率、重复值、平均值等。需额外指定 field_name。

质量维度

质量维度用于对规则进行分类，仅在 custom_sql 类型下需要指定（系统模板已内置）。

维度标识	中文名	说明
completeness	完整性	检测数据是否存在缺失

accuracy	准确性	检测数据值是否在合理范围内
validity	有效性	检测数据格式、类型是否合规
uniqueness	唯一性	检测数据是否存在重复
consistency	一致性	检测不同数据源或表之间数据是否一致
timeliness	及时性	检测数据是否按时产出

系统模板参考

表级模板

模板标识	名称	说明	trigger_condition
table_row_count	表行数	检测表的总行数是否在合理范围内	用户自定义
data_timeliness	数据产出及时性	检测数据是否按时产出	系统固定无需填写。通过 filter 配置业务过滤条件

字段级模板

模板标识	名称	说明	trigger_condition
null_count	字段空值个数	统计字段中空值的个数	用户自定义
null_rate	字段空值率	统计字段中空值占比	用户自定义
not_null_count	字段非空计数	统计字段中非空值的个数	用户自定义
duplicate_count	字段重复值个数	统计字段中存在重复的值的个数	用户自定义
distinct_count	字段去重计数	统计字段中不同值的个数	用户自定义
min_value	字段最小值	获取字段的最小值	用户自定义
max_value	字段最大值	获取字段的最大值	用户自定义

sum_value	字段求和	计算字段的总和	用户自定义
avg_value	字段平均值	计算字段的平均值	用户自定义
median_value	字段中位数	计算字段的中位数	用户自定义
std_dev	字段标准差	计算字段的标准差	用户自定义
enum_range_consistency	枚举范围一致性	检测字段值是否在合法枚举范围内	系统固定无需填写。通过 params.enum_values 配置合法枚举值列表

触发条件表达式语法

trigger_condition 描述的是检测值应满足什么条件时触发异常。
核心语义：满足 trigger_condition 则触发异常。

简单比较

使用比较运算符直接与目标值进行比较。
支持的运算符：

运算符	含义
==	等于
!=	不等于
>	大于
<	小于
=	大于等于
<=	小于等于

示例：

```
trigger_condition: "> 100" # 检测值大于 100 则触发
trigger_condition: "<= 0"  # 检测值小于等于 0 则触发
trigger_condition: "== 24" # 检测值等于 24 则触发
trigger_condition: "!= 0"  # 检测值不等于 0 则触发
```

```
trigger_condition: "> 0"      # 检测值大于 0 则触发
```

区间表达式

用于检查检测值是否在或不在某个数值区间内。

语法格式： between 或 not_between 加 左括号下界逗号上界右括号

括号说明：

括号	含义
[或]	包含边界值（闭区间）
(或)	不包含边界值（开区间）

表达式说明：

表达式	含义	触发时机
between [0,150]	值落在 0到150 内	值在区间内时触发
between (0,100)	值落在 0到100 内	值在区间内时触发（不含边界）
between [0,100)	值落在 0到100 内	值在区间内时触发（含左不含右）
between (0,100]	值落在 0到100 内	值在区间内时触发（不含左含右）
not_between [0,100]	值不在 0到100 内	值超出区间时触发
not_between (0,100)	值不在 0到100 内	值超出区间时触发

实际场景中多使用 not_between，表达检测值超出正常范围则触发异常。

示例：

```
trigger_condition: "not_between [1000,1000000]"
# 表行数不在此范围内则触发
trigger_condition: "not_between [10, 10000]"
# 字段均值不在此范围内则触发
trigger_condition: "not_between [1, 50]"
# 去重计数不在此范围内则触发
```

组合条件

支持使用 AND（同时满足）和 OR（满足任一）连接多个条件。

连接词	含义
AND	所有条件同时满足时触发
OR	任一条件满足时触发

示例：

选项一

```
trigger_condition: "< 100 OR > 1000000" # 行数过少或过多均触发
trigger_condition: "> 100 AND < 200" # 值大于100且小于200时触发
```

 **注意：**
between 和 not_between 表达式中的区间不会被误拆为多个条件。

自定义 SQL 字段名前缀

当规则类型为 custom_sql 时，trigger_condition 中的每个条件必须带字段名前缀，字段名须与 SQL 返回的列名一致。

示例：

选项一

```
trigger_condition: "check_value > 0"
# SQL返回列名为check_value
trigger_condition: "null_count == 0"
# SQL返回列名为null_count
trigger_condition: "check_value > 0 AND total_count >= 100"
# 多列组合判断
trigger_condition: "check_value not_between [0, 100]"
# 区间也支持
```

模板与触发条件兼容性约束

不同模板对 trigger_condition 的配置要求不同，请严格遵循以下约束。

模板或类型	trigger_condition 配置方式	补充说明
普通系统模板（如 table_row_count、null_rate、avg_value 等）	用户自定义。支持简单比较、区间、组合条件	无
data_timeliness 模板	无需填写，触发条件由系统固定	通过 filter 字段配置业务过滤逻辑
enum_range_consistency 模板	无需填写，触发条件由系统固定	通过 params.enum_values 配置合法枚举值列表
自定义 SQL	用户自定义。支持带字段名前缀的所有表达式格式	trigger_condition 中的字段名须与 SQL 返回的列名一致

配置字段说明

通用字段

字段	类型	必填	说明
rule_name	string	是	规则名称，建议使用描述性名称
rule_type	string	是	规则类型：system_template 或 custom_sql
trigger_condition	string	视模板而定	触发条件表达式（部分模板由系统固定，无需填写）
filter	string	否	可选，附加 WHERE 过滤条件

系统模板专用字段

字段	类型	必填	说明
template_code	string	是	系统模板标识，通过 DescribeRuleTemplates 接口查询
params	object	否	部分模板需要的额外参数

自定义 SQL 专用字段

字段	类型	必填	说明
custom_sql	string	是	自定义 SQL 查询语句

dimension	string	是	质量维度：completeness、accuracy、validity、uniqueness、consistency、timeliness
-----------	--------	---	---

字段级规则专用字段

字段	类型	必填	说明
field_name	string	是	目标检测字段名称

配置示例

表级规则示例

示例 1：表行数范围检查（系统模板）

选项一

```
table_rules:
- rule_name: 表行数范围检查
  rule_type: system_template
  template_code: table_row_count
  trigger_condition: not_between [1000,1000000]
  filter: status != DELETED
```

说明：
使用 table_row_count 模板检测表行数，当行数不在 1000 到 1000000 之间时触发中等告警，并通过 filter 排除已删除数据。

示例 2：表行数组合条件检查（系统模板）

选项一

```
table_rules:
- rule_name: 表行数异常检查
  rule_type: system_template
  template_code: table_row_count
  trigger_condition: < 100 OR > 1000000
```

示例 3：数据产出及时性检查（系统模板 – 固定触发）

选项一

```
table_rules:
- rule_name: 数据产出及时性检查
  rule_type: system_template
  template_code: data_timeliness
  filter: update_time >= DATE_SUB(NOW(), INTERVAL 2 HOUR)
```

示例 4：跨表数据一致性检查（自定义 SQL）

选项一

```
table_rules:
- rule_name: 订单金额一致性检查
  rule_type: custom_sql
  dimension: consistency
  custom_sql: >
    SELECT ABS(
      (SELECT SUM(order_amount) FROM order_table WHERE dt = 20260101)
      - (SELECT SUM(detail_amount) FROM order_detail_table WHERE dt =
20260101)
    ) as check_value FROM dual
  trigger_condition: check_value != 0
```

i 说明：

自定义 SQL 计算两张表的金额差值，当差值不为 0 时说明金额不一致，触发告警。注意 trigger_condition 中的 check_value 需与 SQL 返回的列名一致。

字段级规则示例

示例 1：字段空值率检查。

对 user_id 字段进行空值率检查，作为必填字段不允许出现任何空值。

选项一

```
field_rules:
  - field_name: user_id
    rule_name: 用户ID空值率检查
    rule_type: system_template
    template_code: null_rate
    trigger_condition: "> 0"
```

示例 2：字段重复值检查

对 order_id 字段进行唯一性检查，不允许出现重复的订单 ID。

选项一

```
field_rules:
  - field_name: order_id
    rule_name: 订单ID重复值检查
    rule_type: system_template
    template_code: duplicate_count
    trigger_condition: "> 0"
```

示例 3：字段平均值范围检查

检测 order_amount 字段的平均值是否在 10 到 10000 的合理区间内。

选项一

```
field_rules:
  - field_name: order_amount
    rule_name: 订单金额均值检查
    rule_type: system_template
    template_code: avg_value
    trigger_condition: not_between [10, 10000]
```

示例 4：自定义 SQL – 字段格式校验

选项一

```
field_rules:
  - field_name: phone_number
    rule_name: 手机号格式校验
    rule_type: custom_sql
    dimension: validity
    custom_sql:
      SELECT COUNT(*) as check_value
      FROM user_info_table WHERE dt = 20260101
      AND phone_number IS NOT NULL
      AND phone_number NOT REGEXP 1[3-9][0-9]{9}
    trigger_condition: check_value > 0
```

示例 5：枚举值范围一致性校验（系统模板 – 固定触发）

选项一

```
field_rules:
  - field_name: gender
    rule_name: 性别枚举值范围一致性校验
    rule_type: system_template
    template_code: enum_range_consistency
    params:
      enum_values: M,F,Unknown
```

常见问题 FAQ

Q1：系统模板和自定义 SQL 如何选择？

建议优先使用系统模板，简洁高效且无需指定质量维度；当模板无法满足需求时，再使用自定义 SQL。

场景	推荐类型
标准的字段统计检查（空值率、重复值、均值等）	system_template
跨表数据一致性比对	custom_sql
复杂业务逻辑校验（格式、范围、关联检查等）	custom_sql
简单的表行数或及时性检查	system_template

Q2：trigger_condition 什么时候不需要填写？

以下两种模板的触发条件由系统固定，无需填写 `trigger_condition`：

- `data_timeliness`（数据产出及时性）：通过 `filter` 配置业务过滤条件。
- `enum_range_consistency`（枚举范围一致性）：通过 `params.enum_values` 配置合法枚举值。

Q3：自定义 SQL 中 `trigger_condition` 的字段名有什么要求？

`trigger_condition` 中引用的字段名必须与自定义 SQL 返回结果集中的列名完全一致。

例如 SQL 返回 `check_value` 列，则条件应写为 `check_value > 0`，而不能写成 `> 0`。

Q4：`between` 和 `not_between` 有什么区别？

- `between [a, b]`：值在区间内时触发异常（较少使用）
- `not_between [a, b]`：值不在区间内时触发异常（常用，表示超出正常范围）

实际业务中，大多数场景使用 `not_between`，表达检测值超出正常范围时触发告警。

Q5：`filter` 字段的作用是什么？

`filter` 是一个可选字段，用于在检测时附加 WHERE 过滤条件，限制参与检测的数据范围。

常见用途：

- 排除已删除或无效数据。
- 限定特定分区或地区的数据。
- 配合 `data_timeliness` 模板限定时间范围。

Q6：如何获取可用的系统模板列表？

通过调用 `DescribeRuleTemplates` 接口查询当前系统支持的所有模板信息，包括模板标识（`template_code`）、适用级别（表级或字段级）、内置质量维度等。

Q7：一个配置文件中可以定义多少条规则？

没有数量限制。可以在同一个配置文件中自由组合表级规则和字段级规则，系统将逐条执行检测。

Q8：`dimension` 字段什么时候需要填写？

- 系统模板（`system_template`）：不需要，系统已内置质量维度。
- 自定义 SQL（`custom_sql`）：必须填写，从 `completeness`、`accuracy`、`validity`、`uniqueness`、`consistency`、`timeliness` 中选择。

数据剖析（Data Profiling）

最近更新时间：2026-10-09 16:27:21

概述

数据剖析（Data Profiling）对一张表自动生成完整的字段级统计报告与漂移分析仪表盘，帮助您在使用数据前快速理解表的「长什么样」。DataBuddy 的数据剖析支持快照、时序、推理三类剖析，覆盖数值型、字符串、时间、布尔等字段的统计指标，并支持基于基线表计算分布漂移。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- **配置权限**：仅表 Owner 与拥有表 **Manager** 权限的用户可以开启或编辑数据剖析（角色权限说明详见 [成员与权限管理](#)）。
- **查看权限**：拥有表读取权限即可查看仪表盘与运行记录。
- 已准备好可用的计算资源用于执行剖析作业。
- 如果需要计算漂移指标，请提前确定基线表。

使用限制

限制项	说明
入口	Catalog 表详情页 质量 Tab 中的「数据剖析」入口
漂移产出条件	仅在配置了基线表时产生漂移指标，否则漂移指标为空
推理剖析必需列	至少配置时间戳、预测列与模型 ID 列；公平性指标需额外指定标签列、对照组、保护组
仪表盘删除	当前删除仪表盘会一并删除中间表，后续会优化为只删配置保留中间表
通知	剖析任务的告警通知本期不支持

操作步骤

步骤 1：在表详情页开启数据剖析

1. 登录 [DataBuddy 控制台](#)。
2. 进入 **数据目录** > 选中目标表 > **质量**。
3. 在**数据剖析** 区域打开开关。开关开启后会弹出配置抽屉。

步骤 2：选择剖析类型

剖析类型决定了需要填写的配置项与最终产出的指标集合。

类型	适用场景	必填配置
快照 (Snapshot)	静态表的整体分布与质量摘要	计算资源、资源模式、资源配置
时序 (TimeSeries)	按时间维度持续观察分布趋势	计算资源、资源模式、资源配置、时间戳、指标粒度
推理 (Inference)	监控 ML 模型推理日志的特征 漂移与模型质量	计算资源、资源模式、资源配置、时间戳、指标粒 度、标签列、预测列、模型 ID

步骤 3：配置剖析参数

1. 检测对象：默认展示当前表（不可修改）。
2. 计算资源：从下拉列表中选择资源组。
3. 调度方式：在 高级选项 中切换为定时调度，或保持默认手动调度。
 - 手动：首次创建时执行一次，后续通过点击 刷新仪表盘 主动触发。
 - 周期调度：首次创建时执行一次，之后按调度计划周期执行，调度逻辑复用工作流。
1. 基线表：在高级选项中选择，用于计算漂移指标。未配置时漂移指标为空。
2. 指标粒度（时序与推理类型必填）：从小时、天、周、月中选择。
3. 推理类型额外参数：

参数	是否必填	说明
问题类型	是	分类 / 回归
时间戳列	是	推理请求的时间戳字段
预测列 (prediction_column)	是	模型输出的预测值字段
模型 ID 列 (model_id_column)	是	模型标识字段，支持多版本对比
标签列 (label_column)	是	配置后可计算准确率、精确率等模型质量指标

1. 公平性与偏差参数（仅分类模型，可选）：在选择预测列与标签列后，额外指定对比列、保护组、对照组。

步骤 4：确认并启动

完成配置后确认。系统会基于当前配置启动剖析作业，并生成对应的 Profiling 仪表盘。

 提示

首次执行时，系统会对全表数据生成一个完整的统计指标快照，作为后续增量更新的基础。

步骤 5：管理仪表盘与剖析作业

开启数据剖析后，质量 Tab 的剖析区域会出现以下操作：

操作	说明
查看仪表盘	跳转到仪表盘模块，定位到该表的剖析仪表盘
刷新仪表盘	重新执行仪表盘上的查询，从中间表拉取最新数据；不会重新计算指标，也不会改变图表结构
删除仪表盘	移除当前仪表盘的所有图表与剖析数据；点击后弹出二次确认
编辑配置信息	右侧抽屉打开配置抽屉，所有字段均可再编辑，保存后按新配置更新仪表盘
运行记录	弹窗展示剖析的执行历史（执行时间、触发方式、策略、状态、耗时），可点击查看日志

 **注意**
删除仪表盘 都会覆盖或清除当前剖析结果，请确认无下游依赖后再操作。

参数说明

基础配置

参数	说明	是否必填	默认值	取值范围
剖析类型	决定指标集合与必填项	是	快照	快照 / 时序 / 推理
计算资源	执行剖析作业的计算资源	是	—	Workspace 可用计算资源
调度方式	控制剖析的触发方式	是	手动	手动 / 周期调度（高级选项内）
基线表	用于漂移对比的参照表	否	—	同 Workspace 中的表

时序与推理类型必填配置

参数	时序	推理	说明
时间戳列	是	是	用于按时间分桶聚合
指标粒度	是	是	小时 / 天 / 周 / 月
标签列	—	是	推理类型可选，配置后计算模型质量指标

预测列	—	是	推理类型必填
模型 ID 列	—	是	推理类型必填

统计指标参考

漂移指标

漂移指标基于「当前剖析结果 vs 用户指定的基线表」对比计算。

数值型字段：

指标	英文名	说明
KS 检验	Kolmogorov–Smirnov Test	检验两个分布是否来自同一总体
PSI	Population Stability Index	衡量分布偏移程度，业界最常用
Wasserstein 距离	Wasserstein Distance	两个分布之间的「搬土」距离

类别型字段：

指标	英文名	说明
卡方检验	Chi-Square Test	检验分类分布是否显著变化
JS 散度	Jensen–Shannon Divergence	衡量两个分布的对称距离
TV 距离	TV Distance	衡量两个概率分布差异的核心指标
切比雪夫距离	L^∞ Distance	衡量两个分布在单一维度上的最大差异

推理指标（推理类型）

配置 `predict` 与 `label` 列后，平台额外产出以下模型质量指标：

模型类型	指标	说明
分类	准确率（accuracy_score）	正确预测的比例
分类	混淆矩阵	仅在仪表盘展示，不在规则模板中固化
分类	精确率（precision）	$TP / (TP + FP)$
分类	召回率（recall）	$TP / (TP + FN)$
分类	F1 分数（f1_score）	精确率与召回率的调和平均

回归	MSE / RMSE / MAE / MAPE	误差类指标，越低越好
回归	决定系数 R^2	介于 0 和 1 之间，越接近 1 拟合越好
分类（公平性）	预测奇偶性、预测平等性、机会均等性、统计奇偶性	需额外指定保护组、对照组

验证结果

完成开启与首次执行后，您可以通过以下方式验证：

1. 在质量 Tab 中可以看到剖析区域的**查看仪表盘**；点击进入仪表盘模块后能正常加载图表。
2. 在「运行记录」弹窗中能看到首次执行的实例记录与状态。
3. 如配置了基线表，仪表盘中漂移指标区域有具体数值；未配置时漂移指标为空。

常见问题

Q：分区表与非分区表的剖析行为有什么差别？

A：分区表的剖析仅扫描新增分区并合并历史统计量，全局指标（如中位数）通过历史值与新增分区联合重算；非分区表每次重新扫描全表并覆盖上次结果。

Q：如果中途关闭了漂移检测，历史漂移数据会保留吗？

A：会保留。历史漂移数据保持不变，但后续剖析不再追加漂移计算。

Q：仪表盘看到的是中间表数据，重新计算指标该怎么做？

A：「刷新仪表盘」只是重新拉取已有中间表的数据，不会重新执行 Profiling。如需重新计算指标，请使用「重新生成仪表盘」。

相关文档

- [创建质量规则](#)
- [异常检测](#)
- [创建仪表盘与布局](#)

异常检测（Anomaly Detection）

最近更新时间：2026-10-09 16:27:26

概述

异常检测是 DataBuddy 提供的一项零配置数据质量兜底能力。它由系统在后台自动学习每张表的更新频率与数据量历史模式，建立基线后持续监控，当实际值偏离基线超过阈值时产生告警。用户只需在 Schema 层级开启开关，无需定义任何规则。

异常检测覆盖以下两类核心风险：

类型	检测目标	核心问题
新鲜度（Freshness）	表的更新是否及时	这张表是不是该来数据了但没来？
完整性（Completeness）	表的数据量是否正常	这张表这次来的数据量是不是不对？

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- **配置权限**：仅拥有 Schema **Manager** 权限的用户可以开启或编辑异常检测（角色权限说明详见 [成员与权限管理](#)）。
- **查看权限**：拥有 Schema 读权限即可查看检测结果。
- 已准备好用于运行检测任务的计算资源。
- 异常检测告警本期不支持

使用限制

限制项	说明
配置粒度	仅 Schema 级支持批量开启；表级自动继承，不支持单独开启或单独关闭
检测维度	本期支持新鲜度、完整性；其他维度暂不支持
学习期间行为	样本不足（新鲜度 < 7 次更新、完整性 < 14 次） 期间不告警，状态显示为「学习中」

操作步骤

步骤 1：在 Schema 层级开启异常检测

1. 登录 [DataBuddy 控制台](#)。
2. 进入 [数据目录](#) > 选中目标 Catalog > 选中目标 Schema > [详情](#)。

3. 在数据质量异常检测 点击开启入口，即可配置完成

步骤 2：查看检测结果

- 1. 在 Schema 详情页的数据质量异常检测 区域点击 查看结果 。
- 2. 支持按照 Catalog/ Schema 筛选查看结果；支持点击刷新按钮刷新列表
- 3. 列表页面展示健康表占比，健康表和不健康以及学习中的表 list。
- 4. 列表字段如下：

列名	说明
表名	表的名称
所属 Schema	表所在 Schema
状态	表的状态
最近检测时间	该表最近的检测时间
检测频率	该表检测频率
操作	查看详情，点击后跳转到表详情页面

列名	说明
表名	出现异常的表
所属 Schema	表所在 Schema
状态	表的健康状态
异常类型	完整性异常 / 新鲜度异常
上游根因	获取导致表异常的上游根因
下游影响	获取表异常后对下游的影响范围
最近检测时间	该表最新一次检测的时间
操作	查看表详情，点击后跳转到表详情页面

列名	说明
表名	表的名称
所属 Schema	表所在 Schema

状态	表的状态
最近检测时间	该表最近的检测时间
检测频率	该表检测频率
操作	查看详情，点击后跳转到表详情页面

步骤 3：在表详情页查看检测结果

1. 进入 数据目录 > 选中目标表 > 质量 Tab。
2. 在「异常检测」区域可查看该表的检测结果。

 **注意**

- 该入口只展示当前表的检测结果，**不能** 单独为表开启或关闭异常检测。
- 点击开关时，会跳转到所属 Schema 的开启入口；点击「为Schema 关闭」会跳转到 Schema 详情页的关闭入口。

步骤 4：编辑或关闭异常检测

操作	入口	说明
编辑	Schema 详情 > 数据质量监控 > 编辑	抽屉打开后可修改计算资源与告警通知
关闭	Schema 详情 > 数据质量监控 > 关闭开关	弹出二次确认「关闭后，所有表将停止智能质量检测，是否确认关闭」

 **注意**

关闭异常检测后，该 Schema 下所有表都会停止检测；告警发送也会同步停止。

关闭与重新开启的恢复逻辑

为避免反复开启时丢失已学习的样本，系统遵循以下规则：

关闭前的表状态	重新开启后的行为
学习中（样本不足）	从断点继续学习： 关闭期间的历史数据保留不清除，重新开启后继续采集，累计样本达标后转为基线就绪
基线已就绪（正常检测中）	立即恢复检测： 使用关闭前已建立的基线继续判定；如关闭超过 30 天，系统自动触发基线重建，重建期间状态临时回到「学习中」

关闭期间新增的表	自动纳入检测，从零开始学习，状态为「学习中」
----------	------------------------

关闭期间的数据处理原则：

- 关闭期间不采集、不检测、不告警；
- 已采集的历史数据与基线参数保留在存储中不清除；
- 重新开启后优先利用已有数据，避免从零开始。

参数说明

Schema 级配置

参数	说明	是否必填	默认值	取值范围
检测对象	Schema 范围	是	当前 Schema	不可修改
检测维度	当前支持的检测类型	是	新鲜度 + 完整性	不可修改
采集频率	数据采集频率	是	基于历史数据自动学习	不可修改
告警渠道	告警渠道与接收人	是	-	通知管理中已启用的渠道

常见问题

Q：一张表刚开启异常检测，多久能开始告警？

A：系统会立即开始采集历史数据。如果表的元数据中已有足够的更新记录（≥ 7 次），可以立即建立基线并开始检测；否则需等待累计到足够样本，期间状态为「学习中」、不告警。

Q：异常检测和手动质量规则有什么区别，应该用哪个？

A：两者互补。异常检测面向「未知的未知」——零配置发现新鲜度、完整性偏离；手动质量规则面向「已知的风险」——精准守护字段级业务规则。建议先开启异常检测获得兜底，再针对暴露出的高频问题创建手动规则做精细化管理。手动规则详见 [创建质量规则](#)。

Q：能否单独为某张表关闭异常检测？

A：暂不支持。检测策略以 Schema 为粒度统一开启与关闭，表级会自动继承。如需让某张表不被检测，请在 Schema 层级关闭异常检测后再有选择地启用。

相关文档

- [创建质量规则](#)
- [数据剖析](#)
- [设置通知渠道](#)

语义模型概述

什么是语义模型

最近更新时间：2026-10-09 16:27:30

概述

语义模型（Semantic Model）是 DataBuddy 在物理表与业务消费之间引入的语义层（Semantic Layer）。它把分散在数据湖各张表中的字段，按业务视角抽象为三类语义对象——**关系模型（Model）**、**指标（Metric）**和**维度（Dimension）**，为 Buddy（AI Agent）、报表与下游应用提供一份口径统一、权限可控的语义入口。

本文介绍语义模型的设计目标、整体架构与权限模型，帮助您建立对语义层的整体认知。后续操作请参阅 [创建语义模型](#)、[定义指标与维度](#) 与 [语义模型应用](#)。

为什么需要语义层

在没有语义层的传统数仓中，常见的痛点是：

- **同一指标的口径不一致**：不同部门的"GMV"可能各自指 `SUM(order_amount)`、`SUM(pay_amount)` 或 `SUM(pay_amount) WHERE order_status='paid'`，对账时反复扯皮。
- **业务分析师严重依赖数据团队**：每问一次「上海地区上周的客单价」都要重新写 SQL、申请底表权限。
- **底表权限与业务消费深度耦合**：业务方为了用一个指标，不得不申请其依赖的所有底表 Select 权限，权限授予与回收都很重。
- **AI 取数难以自动化**：自然语言问数若直接走 NL2SQL 路径，不仅幻觉率高，而且每次都要解决建表、关联、字段语义等问题。

语义层把这些问题集中在一处治理：

- 一次定义、随处消费——指标、维度成为一等公民，Buddy、报表等共用同一份口径。
- 权限独立、最小授权——业务消费指标完全不依赖底表权限。
- 行级安全可控——通过维度值范围授权实现「张三只看华东」。

设计原则

原则	说明
Owner + 权限点	每类语义资产（模型、指标）只有一个 Owner（可转交）； 模型上 授予 <code>Manage</code> / <code>CreateMetric</code> 权限点， 指标上 授予 <code>Manage</code> / <code>Query</code> 权限点
语义层独立于 Catalog	创建模型时需要底层表 Select 权限； 创建指标、查询指标数据完全不依赖 Catalog 权限
最小权限	创建指标无需看底层数据，只需引用模型；查询指标无需管理模型权限

关系模型 Model

关系模型是业务分析的最小数据单元，对应一张或多张物理表的语义封装。模型负责声明：

- 引用了哪些 Catalog 表（含 Catalog / Schema / Table 三段路径）；
- 表与表之间的关联（Join）关系；

一个模型可以被多个指标引用，模型一旦发布，其结构变更需要重新走发布流程。

指标 Metric

指标是业务上可被消费的「数字口径」，在模型之上定义。指标分为**原子指标**与**衍生指标**两大类：

- **原子指标（SIMPLE）**：指标体系的最小统计单元，对单一字段做聚合计算，如 `SUM(OrderAmount)` 表示「交易总金额」。原子指标不涉及衍生逻辑。
- **衍生指标**：在原子指标基础上按业务规则加工得到的指标，又细分为 5 种：
 - **筛选型**：在原子指标基础上按特定维度（如时间、地域、用户属性）筛选后得到，例如「线上渠道交易总金额」= `SUM(OrderAmount)` 限定 `渠道 = 线上`。
 - **累积计算型**：对原子、筛选型指标按时间窗口累加、汇总得到，例如「近 30 天线上渠道交易总金额」。
 - **同环比计算型**：以累积计算型指标为基础，与历史同期（同比）或上一周期（环比）数据对比的相对指标，例如「近 30 天线上渠道交易总金额环比增长率」。
 - **混合计算型**：结合两个及以上指标通过衍生计算规则（如加减乘除等）得到的复合指标，例如「平均交易流水金额」= `交易总金额 / 交易数量`。
 - **转化计算型**：基于不同环节的原子、筛选型指标，计算流程中「从 A 到 B」的转化效率，例如「注册-下单转化率」= `下单量 / 注册量`。

指标支持的聚合方式包括 `SUM`、`COUNT`、`COUNT(DISTINCT)`、`MAX`、`MIN`、`AVG`，以及自定义表达式。

维度 Dimension

维度用于定义分析视角。按字段的业务语义，常见维度可分为以下几类：

- **时间维度**：用于按时间窗口对指标进行切片的日期、时间类字段，支持按天、周、月、季、年等不同粒度聚合，是趋势分析（如同环比、近 N 天）的核心切片依据。
 - 示例：`order_date`（订单创建日期）、`pay_time`（支付完成时间），常按天、周、月切片。
- **字典维度**：值域固定、可枚举的维度字段，对应业务上的「类目」「品牌」等分类体系。其取值通常来自字典表或预设枚举，字段值域有限且稳定，便于做归一化与标准化管理。
 - 示例：`category_name`（商品类目）、`brand`（品牌）、`gender`（性别）。
- **普通维度**：用于描述业务实体属性的通用维度字段，没有特殊的时间或字典语义，常见为用户姓名、年龄、用户 ID 等属性。
 - 示例：`user_name`（用户名）、`age`（年龄）、`user_id`（用户 ID）。

维度是跨模型复用的全局维度（可在「维度管理」中统一定义）。

权限模型概览

语义层权限分为三大类：**功能权限**、**ACL 权限**、**数据权限**，均挂在工作空间 → 角色管理 → 语义建模菜单下：

权限类别	控制目标	配置位置
功能权限	用户能否进入「模型管理 / 指标管理 / 维度管理」 Tab，以及能否新建 / 修改 / 删除	工作空间 → 角色管理 → 语义建模菜单
ACL 权限 (资产级)	用户对单个模型的具体能力（Manage / CreateMetric）；对单个指标的具体能力（Manage / Query）	模型详情页 → 权限 Tab 指标详情页 → 权限 Tab
数据权限 (行级)	同一指标下，用户能查询到的维度值范围	指标授权时附带的「维度值范围」

模型 ACL 权限矩阵

操作	Owner 身份	Manage 权限点	CreateMetric 权限点
查看模型结构	✓	✓	✓
编辑模型结构	✓	✓	✗
引用该模型创建指标	✓	✓	✓
发布 / 下线模型	✓	✓	✗
授权他人	✓	✓	✗
转交模型 Owner	✓	✗	✗
删除模型	✓	✗	✗

指标 ACL 权限矩阵

操作	Owner 身份	Manage 权限点	Query 权限点
查看指标元数据	✓	✓	✓
查询指标数据	✓	✓	✓（可被维度值限制）
编辑指标口径	✓	✓	✗
发布 / 下线指标	✓	✓	✗
授权他人	✓	✓	✗

转交指标 Owner	✓	✗	✗
删除指标	✓	✗	✗

Catalog 与语义层的协同

语义层的设计原则是「创建需 Catalog 权限，消费不需要」：

场景	所需 Catalog 权限	所需语义层权限
创建模型	Table.Select	语义建模功能权限
管理模型（被授 Manage）	Table.Select + Table.Browse	模型的 Manage 权限点
创建指标（被授 CreateMetric）	不需要	模型的 CreateMetric 权限点
查询指标数据	不需要	指标的 Query 权限点
NL2SQL 直接查表	Table.Select + Table.Browse	不需要

❗ 提示

关于 Table.Browse：

`Browse` 是 Catalog 层对底层表的「可预览」权限，独立于 `Select` 的全量读权限。模型 Manage 用户在编辑、发布、下线等需要触达底层表结构的环节，除了 `Select` 还必须拥有 `Browse` 权限，否则会被拦截。

因此即使用户在 Catalog 中底表的 Select 权限被回收，只要用户有已发布指标的查询仍可正常执行，保障业务连续性。

相关文档

- 创建语义模型
- 定义指标与维度
- 语义模型应用

创建语义模型

最近更新时间：2026-10-09 16:27:33

概述

本文介绍如何在 DataBuddy 中通过 YAML 编辑器创建一个语义模型。语义模型是业务分析的最小数据单元，对应一张或多张物理表的语义封装，是后续定义指标的基础。完成创建后，您可以发布模型，并在其上 [定义指标与维度](#)、授权给业务方消费。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务，并在目标工作空间中拥有**本体建模**的 **Read Write** 或 **Read Write & Delete** 功能权限。无功能权限时，模型管理 Tab 不可见或仅可查看；缺权限时请联系工作空间管理员申请。
- 拥有要引用的物理表的 Catalog **Table.Select** 权限。模型 YAML 中引用的每张表都需要您当前账号具备 Select 权限，否则发布时校验不通过。

使用限制

限制项	说明
创建权限	仅具备 本体建模 的 Read Write 及以上功能权限的用户可创建
编辑形态	当前仅支持 YAML 编辑器，或使用自然语言借助 Buddy 辅助创建
负责人	模型创建者默认为 负责人，每个模型只能有一个 负责人，可转交

操作步骤

步骤 1：进入本体建模 → 模型管理

在工作空间左侧导航中单击**本体建模**，进入**关系模型**子菜单。该页面集中展示当前工作空间内所有您可见的模型。每个模型卡片展示模型名称、负责人、所含表数与指标数、模型描述、英文标识符以及状态标签（已发布、已下线）。

步骤 2：发起新建

在页面右上角单击 **+ 新建模型（YAML）**。系统弹出 YAML 编辑面板，左侧为表选择器，右侧为代码编辑器。

① 说明：

如果您不确定 YAML 写法，可以单击右上角[查看示例](#)调出官方示例片段，或借助 Hi Buddy 入口让 AI 助手协助生成模型 YAML。

步骤 3：选择引用的物理表

在左侧表选择器中选择数据来源，并可快捷选择插入要在 YAML 中引用的资源：

1. 数据目录：可选择来自 TC Catalog（原生）的数据目录、或来自直连数据源，比如 Doris、StarRocks 的 Catalog、Schema 表等。
2. 逻辑视图：可选择创建的逻辑视图以及逻辑视图内的字段
3. 关系模型：可选择关系模型、关系模型内引用的表、字段、以及关系模型内创建的指标、以及指标支持的分析维度等。

步骤 4：编写模型 YAML

在右侧编辑器中按以下结构编写 YAML：

```
# 版本：1.0=====

# 【必填】配置文件版本号，固定为 1.0
version: 1.0

# =====
# 语义模型定义
# type: DEF - 新建模型（需提供完整定义）
#         REF - 引用已有模型（只需 name + type）
# =====

model:
  # 【必填】模型英文名，1-50字符，支持字母、数字、下划线、括号
  name: ecommerce_order_model

  # 【可选】模型别名列表，支持多个，不允许含逗号
  # 为空时默认使用 name；存入数据库时以逗号分隔存储
  label:
    - 电商订单模型
    - order_model_alias

  # 【可选】模型描述
  description: "电商业务订单主模型，关联用户表和商品表"

  # 【必填】模型类型：DEF (定义新模型) | REF (引用已有模型)
  type: DEF

  # 【DEF必填】主表数据源路径，格式：catalogName.databaseName.tableName
  source: default_catalog.ecommerce_db.orders
```

```
# 【可选】关联表配置，支持嵌套 JOIN
joins:
  - name: user_table                                # 关联表别名（必填）
    source: default_catalog.ecommerce_db.users      # 关联表路径（必填）
    on:                                              # 关联条件（必填，至少一个）
      - source.user_id = user_table.id
# 嵌套关联（可选）
joins:
  - name: user_address
    source: default_catalog.ecommerce_db.addresses
    on:
      - user_table.id = user_address.user_id
  - name: product_table
    source: default_catalog.ecommerce_db.products
    on:
      - source.product_id = product_table.id
```

YAML 关键字段说明：

字段	是否必填	说明
version	是	YAML schema 版本，当前固定为 1.0
model.name	是	模型英文标识，仅支持小写字母 / 数字 / 下划线，工作空间内唯一
model.label	是	模型别名，用于列表卡片与详情页展示
description	否	模型描述，建议简述业务域与覆盖范围
model.type	是	取值：DEF（新建模型，需提供完整定义） / REF（引用已有模型，仅需 name + type）
model.source	是	模型主表的三段路径 catalog.schema.table

<code>model.join</code> <code>s</code>	多表必填	表与表的 Join 关系；当前仅在多表模型时必须声明
---	------	----------------------------

 **注意：**
模型 YAML 中引用的每一张表都必须在当前账号下具备 **Table.Select** 权限。发布时系统会逐表校验，缺权限的表会在 YAML 编辑器中以警告提示，并阻断发布流程。

步骤 5：发布模型

模型 YAML 编辑完成后，单击左下角**发布**：

1. 系统先做语法校验，校验项包括：YAML 格式、字段必填、表 Select 权限、维度字段是否存在、Join 字段是否存在等。
2. 校验通过后弹出确认弹窗。
3. 单击确认后模型进入「已发布」状态，模型卡片显示绿色 **已发布** 标签。

发布后，您可以：

- 在该模型上 [定义指标与维度](#)。
- 在模型详情页的「授权」Tab 给其他用户授予 Manage / CreateMetric 权限。
- 通过 Buddy 让 AI 协助创建指标。

步骤 6：转交 Owner（可选）

每个模型只能有一个 Owner，默认为创建者。如需转交：

1. 进入模型详情页 → 右侧抽屉「基本信息 > 负责人」→ 鼠标 hover 后，出现修改的小笔 ICON，单击后即可**转交 Owner**。
2. 选择目标用户后提交，系统弹出二次确认。
3. 转交后原 Owner 失去删除模型与转交模型的权限；如仍需操作，可由新 Owner 授予 Manage 权限。

通过 Buddy 创建模型

支持在 Buddy 中用自然语言描述需求，由 AI 协助生成模型 YAML 并完成发布。常见对话示例：

帮我用 `dlc.default_catalog.olist.orders` 和 `users` 表建一个 `Olist` 电商销售模型，关联键是 `use`
`r_id`

Agent 在执行写操作（创建、修改、发布、下线、删除）前会先校验权限：

- 若您没有「模型管理」的功能权限，Buddy 会拦截操作并提示：「您当前暂无创建语义模型的权限，可联系工作空间管理员 XXX 申请开通『语义模型 → 模型管理的读写删权限』后再操作；也可以直接提交需求，由语义层管理员 XXX 为您协助创建。」
- 若您有创建权限但缺少所需表的 Select 权限，Buddy 会列出缺权限的表名与对应表 负责人，引导您线下申请。

模型详情页

模型发布后，在「模型管理」列表页单击任一模型卡片即可进入**模型详情页**。详情页是模型上线后进行结构查看、指标与维度管理、权限配置的主入口。

页面整体结构

详情页整体分为三个区域：

- **顶部栏**：展示模型名，与状态标签（已发布 / 已下线）；右侧按当前 Tab 上下文展示操作按钮，常驻有「编辑」和「下线」，进入「权限」Tab 时还会出现「授权」按钮。
- **主内容区**：以 Tab 形式提供四个视图，依次为「实体关系」（默认）、「关联指标」、「关联维度」、「权限」。
- **右侧信息栏**（「实体关系」Tab 下展示）：包含三组信息
 - **基本信息**：别名、描述、负责人、状态。
 - **技术信息**：创建时间、创建人、最近更新时间、变更人。
 - **模型信息**：关联表数、指标数、维度数。

实体关系 Tab

默认进入的 Tab，以**画布**形式可视化展示模型内物理表之间的关联关系：

- 每张表对应一个表卡片，顶部展示表名（含 `catalog.schema.table` 三段路径），下方按字段顺序列出字段名、数据类型与中文说明；
- 表与表之间的连线对应 YAML 中声明的 `joins` 关联关系；
- 画布右上角提供放大、缩小、宽屏、全屏等视图控制，便于在多表、字段多的场景下浏览。

关联指标 Tab

以**表格**形式列出该模型上已定义的全部指标：

列	说明
指标名 / 指标别名	指标的英文标识与中文标签
指标类型	原子指标 / 衍生指标
衍生类型	当指标类型为衍生指标时显示具体子类型（筛选型 / 累积计算型 / 同环比计算型 / 混合计算型 / 转化计算型）
指标口径	指标的计算 SQL 摘要
来源表	指标所引用的物理表
负责人	指标 Owner

最近修订时间	最近一次编辑或发布的时间
操作	单条「编辑」「下线」

表格上方提供**批量授权**按钮，勾选多个指标后可统一发起授权。

关联维度 Tab

以**表格**形式列出该模型引用的全部维度：

列	说明
维度名 / 维度别名	维度的英文标识与中文标签
描述	维度的业务说明
维度类型	普通维度 / 时间维度 / 字典维度等
维度值信息	维度的取值特征，例如时间维度会显示「最小颗粒：日」
负责人	维度 Owner
关联字段	维度在底层表上对应的物理字段路径
最近修订时间	维度的最近一次修订时间

权限 Tab

以**表格**形式展示模型及其指标的授权情况：

列	说明
授权对象	被授权的用户 / 用户组
类型	用户 / 用户组
权限点	Manage / Create Metric / Query 等具体权限点
表权限	授权对象是否对模型引用的所有物理表都具备 Table.Select 权限（齐全 / 缺失）
维度	授权对象可查询的维度范围
维度值范围	授权对象可查询的维度值范围（行权限）
过滤关系	多个维度值之前的关系
授予人	发起此次授权的用户

授权时间	授权生效时间
操作	「编辑」「回收」

表格上方支持按关键字搜索用户、用户组，并可按权限范围筛选；右上角单击**授权**，可发起新的授权任务，弹出**模型授权**弹窗（详见 [模型授权](#)）；表格内单条**编辑**按钮可对已有授权进行权限点变更，**回收**按钮可撤销已授予的权限。

模型授权

单击「权限」Tab 右上角的「**授权**」按钮后，会弹出「**模型授权 | <模型名>**」弹窗，用于把模型的某项或多项权限授予指定用户、用户组。

弹窗自上而下包含三块内容：

1. 授权主体（必填）

- 可选择：用户、用户组、标签

2. 授权对象（必填）

- 如果授权主体选用户：此处展示选用户面板，支持搜索用户；至少添加 1 个
- 如果授权主体选用户组：此处展示选用户组面板，支持搜索用户组；至少添加 1 个
- 如果授权主体选标签：无需选择授权对象，仅需选下面的授权点即可

3. 授权点（必填，至少勾选 1 项）

字段标签：**授权点 ***，下方为可勾选的权限点卡片，每个权限点包含「能力说明」与「适用范围」两行说明：

权限点	能力说明	适用范围
Create Metric	可基于该模型创建新指标	允许在该模型上定义原子指标 / 衍生指标，但不能修改模型本身
Manage	可管理模型	允许编辑 / 发布 / 下线模型；删除模型仅 Owner 可操作。并要求被授权用户对模型内所有底层表均拥有 Select 和 Browse 权限，否则授权成功后该用户实际操作底层表相关动作时会被拦截
Query All Metrics	可查全部指标	快速给用户授予该模型下所有指标的查询权限。

- 三个权限点可同时勾选，也可只勾选其中之一。
- **校验规则**：若未勾选任何权限点就单击**确认授权**，字段下方会红字提示**请选择权限点**，且不会发起授权请求。

4. 授权原因（选填）

- 字段标签：`授权原因`。
- 控件：多行文本输入框，占位提示为 `请简要说明本次授权的业务背景...`。
- 用于审计与事后追溯，建议填写本次授权的业务目的或来源工单号；不填也可以提交。

5. 提交与生效

- 弹窗底部右侧提供两个按钮：**取消**（关闭弹窗，丢弃已填内容）与**确认授权**（提交授权任务）。
- 单击「确认授权」后：
 - 1.1 系统对表单进行必填校验（授权对象、授权点），校验失败时在对应字段下方给出错误提示。
 - 1.2 校验通过后写入授权记录，授权对象立即获得对应权限点；新授权会作为一条新行出现在「权限」Tab 表格中。
 - 1.3 若被授权用户对模型底层表存在 `Select` / `Browse` 权限缺失，权限 Tab 中对应行的「表权限」列会显示「缺失」，但本次授权仍会生效。

常见问题

Q1：模型创建后能修改引用的表吗？

可以。在模型详情页单击**编辑**重新进入 YAML 编辑器，修改后重新发布即可让新结构生效。

Q2：模型发布后还能下线吗？

可以。在模型详情页单击**下线**后模型状态回到「已下线」，下游对该模型的指标查询会立即失败。下线模型可重新发布。

Q3：删除模型会影响其上的指标吗？

会。删除模型前必须先删除该模型上所有指标，否则系统拦截删除操作。删除后模型不可恢复。

Q4：模型 Owner 与模型管理员（Manage）有什么区别？

Owner 拥有全部权限，包括转交 Owner 与删除模型；Manage 拥有除转交与删除外的全部权限。一个模型只能有一个 Owner，但可以同时存在多个 Manage 用户。

相关文档

- [什么是语义模型](#)
- [定义指标与维度](#)

定义指标与维度

最近更新时间：2026-10-09 16:27:37

概述

本文介绍如何在已发布的语义模型上定义指标（Metric）与维度（Dimension）。指标承载业务可消费的「数字口径」（如 GMV、客单价、DAU），维度承载分析切片视角（如地域、品类、日期）。指标与维度是后续 Buddy 取数、报表展示对外暴露的最小单元。

完成创建后请参阅 [语义模型应用](#) 完成授权与消费。

前提条件

在开始本操作前，请确保满足以下条件：

- 拥有「语义建模 → 指标管理」的 **Read Write** 或 **Read Write & Delete** 功能权限。
- 拥有目标模型的 **Manage** 或 **CreateMetric** 权限（Owner 自动包含），否则无法在该模型上创建指标。
- 目标模型已处于**已发布** 状态。
- 如需创建新维度，还需拥有「语义建模 → 维度管理」的 **Read Write** 功能权限。

使用限制

限制项	说明
创建权限	指标创建需要模型的 CreateMetric 或 Manage 权限；指标创建不需要底层表 Select 权限
编辑形态	指标与维度均通过 YAML 编辑器编辑
维度 ACL	维度本期暂不支持精细化 ACL 权限，所有有「语义建模 → 维度管理」的 Read Write 或 Read Write & Delete 功能权限的用户均可编辑、删除

指标的两种类型

指标是业务上可被消费的「数字口径」，在模型之上定义。指标分为**原子指标** 与**衍生指标** 两大类：

- **原子指标（SIMPLE）**：指标体系的最小统计单元，对单一字段做聚合计算，如 `SUM(OrderAmount)` 表示「交易总金额」。原子指标不涉及衍生逻辑。
- **衍生指标**：在原子指标基础上按业务规则加工得到的指标，又细分为 5 种：
 - **筛选型**：在原子指标基础上按特定维度（如时间、地域、用户属性）筛选后得到，例如「线上渠道交易总金额」= `SUM(OrderAmount)` 限定 `渠道 = 线上`。
 - **累积计算型**：对原子、筛选型指标按时间窗口累加、汇总得到，例如「近 30 天线上渠道交易总金额」。

- **同环比计算型**：以累积计算型指标为基础，与历史同期（同比）或上一周期（环比）数据对比的相对指标，例如「近 30 天线上渠道交易总金额环比增长率」。
- **混合计算型**：结合两个及以上指标通过衍生计算规则（如加减乘除等）得到的复合指标，例如「平均交易流水金额」= `交易总金额 / 交易数量`。
- **转化计算型**：基于不同环节的原子、筛选型指标，计算流程中「从 A 到 B」的转化效率，例如「注册-下单转化率」= `下单量 / 注册量`。

指标支持的聚合方式包括 `SUM`、`COUNT`、`COUNT(DISTINCT)`、`MAX`、`MIN`、`AVG`，以及自定义表达式。

操作步骤

步骤 1：进入指标管理

进入工作空间，选择左侧菜单中的本体建模菜单->指标管理子菜单，该页面以列表形式展示当前空间内您可见的所有指标，列表字段包括：指标名称、指标别名、指标状态、指标类型、衍生类型、口径、关联模型、来源表、负责人、最近修订时间。

步骤 2：发起新建指标

单击右上角 + 新建指标（YAML），从右侧弹出 YAML 编辑面板。面板顶部会先要求选择「数据开源」

① 提示

推荐用 Hi Buddy 协助生成指标 YAML：在左侧菜单选择 Hi Buddy 入口，用自然语言描述指标口径（例如「基于 catalogname.olist_sales_database 数据库内的表，构建销售分析相关语义模型，比如：GMV、近 30 天退货订单数、客单价等。」），AI 会自动分析制定范围的数仓元数据，并生成符合规范的语义模型（含指标、维度、关系模型）YAML 文件。

步骤 3：编写指标 YAML

① 提示

派生指标 YAML 中 `formula` 字段引用的指标必须是同一模型下已发布或同 YAML 中同时定义的指标，否则发布时校验失败。

3.1 完整 YAML 示例（模型 + 6 类指标 + 3 类维度）

下面给出一份端到端完整示例，一次性声明模型（DEF）、所有 6 类指标（`SIMPLE` / `FILTER` / `DERIVED` / `RATIO` / `CUMULATIVE` / `CONVERSION`）以及 3 类维度（`TIME` / `CATEGORICAL` / `DICT`）。可作为编写实际 YAML 时的参考模板。

```
# =====  
# WeData 语义模型 YAML 配置完整示例
```

```

# 版本: 1.0
# 说明: 通过此 YAML 可一次性创建语义模型、指标、维度
# =====

# 【必填】配置文件版本号, 固定为 1.0
version: 1.0

# =====

# 语义模型定义 (可选)
# type: DEF - 新建模型 (需提供完整定义)
#         REF - 引用已有模型 (只需 name + type)
# =====

model:
  # 【必填】模型英文名, 1-50字符, 支持字母、数字、下划线、括号
  name: ecommerce_order_model

  # 【可选】模型别名列表, 支持多个, 不允许含逗号
  # 为空时默认使用 name; 存入数据库时以逗号分隔存储
  label:
    - 电商订单模型
    - order_model_alias

  # 【可选】模型描述
  description: "电商业务订单主模型, 关联用户表和商品表"

  # 【必填】模型类型: DEF (定义新模型) | REF (引用已有模型)
  type: DEF

  # 【DEF必填】主表数据源路径, 格式: catalogName.databaseName.tableName
  source: default_catalog.ecommerce_db.orders

  # 【可选】关联表配置, 支持嵌套 JOIN
  joins:
    - name: user_table                                # 关联表别名 (必填)
      source: default_catalog.ecommerce_db.users      # 关联表路径 (必填)
      on:                                                # 关联条件 (必填, 至少一个)
        - source.user_id = user_table.id

```

```

# 嵌套关联 (可选)
joins:
  - name: user_address
    source: default_catalog.ecommerce_db.addresses
    on:
      - user_table.id = user_address.user_id
  - name: product_table
    source: default_catalog.ecommerce_db.products
    on:
      - source.product_id = product_table.id

# =====
# 指标定义列表 (可选)
# 支持类型: SIMPLE | FILTER | DERIVED | RATIO | CUMULATIVE | CONVERSION
# =====
metrics:
  # -----
  # 1. 原子指标 (SIMPLE) - 基础聚合指标
  # -----
  - name: total_order_amount
    # 【可选】指标别名列表, 支持中英文、数字、下划线、括号, 不允许含逗号
    label:
      - 订单总金额
      - GMV
    description: "统计所有有效订单的总金额"
    # 【必填】指标类型
    type: SIMPLE
    # 【必填】类型参数
    type_params:
      # 【可选】关联模型引用, 填写 model 的 name
      # 不填时默认使用当前 YAML 中定义的 model
      model_ref: ecommerce_order_model
      # 【必填】源表路径, 格式: catalogName.databaseName.tableName
      source_table: default_catalog.ecommerce_db.orders
      # 【必填】聚合表达式
      expr: "SUM(order_amount)"
      # 【必填】时间维度名称 (需在 dimensions 中定义对应的 TIME 类型维度)
      time_dimension: order_date
      # 【可选】非可加维度配置 (用于半累加指标, 如库存快照)
      # non_additive_dimension:

```

```
# dimension: snapshot_date      # 维度名称 (必填)
# window_choice: MAX            # 窗口取值: MIN | MAX (必填)
# window_groupings:            # 窗口分组 (可选)
#   - warehouse_id

# -----
# 2. 筛选型指标 (FILTER) - 在原子指标基础上加过滤条件
# -----
- name: paid_order_amount
  label:
    - 已支付订单金额
  description: "仅统计已支付状态的订单总金额"
  type: FILTER
  type_params:
    # 【必填】筛选条件表达式
    filter: "status = 'PAID'"
    # 【必填】依赖的基础指标列表 (至少一个)
    metrics:
      - name: total_order_amount

# -----
# 3. 混合计算型指标 (DERIVED) - 多指标表达式运算
# -----
- name: avg_order_amount
  label:
    - 平均客单价
  description: "订单总金额 / 订单总数"
  type: DERIVED
  type_params:
    # 【必填】计算表达式, 使用指标名称作为变量
    expr: "total_order_amount / order_count"
    # 【必填】参与计算的基础指标列表 (至少一个)
    metrics:
      - name: total_order_amount
      - name: order_count

# -----
# 4. 同环比指标 (RATIO) - 时间周期对比
# -----
- name: order_amount_yoy
```

```

label:
  - 订单金额同比
description: "订单总金额的同比增长率"
type: RATIO
type_params:
  # 【必填】计算类型: YEAR_ON_YEAR(同比) | RELATIVE_RATIO(环比)
  derived_type: YEAR_ON_YEAR
  # 【必填】基础指标列表 (至少一个)
  metrics:
    - name: total_order_amount

# -----
# 5. 累积计算型指标 (CUMULATIVE) - 时间窗口累积
# -----
- name: weekly_order_amount
label:
  - 近7日订单金额
description: "最近7天的订单金额累积值"
type: CUMULATIVE
type_params:
  # 【必填】时间窗口, 可选值:
  # MINUTE_5 | MINUTE_30 | HOUR_1 | DAY_1 | DAY_7 | DAY_30 | DAY_90
  # DAY_180
  # WTD(本周至今) | MTD(本月至今) | QTD(本季度至今) | YTD(本年至今) |
  HTD(历史至今)
  window: DAY_7
  # 【可选】额外筛选条件
  filter: "status IN ('PAID', 'SHIPPED')"
  # 【必填】基础指标列表 (至少一个)
  metrics:
    - name: total_order_amount

# -----
# 6. 转化类指标 (CONVERSION) - 漏斗转化分析
# -----
- name: view_to_purchase_rate
label:
  - 浏览到购买转化率
description: "用户从浏览商品到下单购买的转化率"
type: CONVERSION

```

```

type_params:
  # 【必填】基准指标（漏斗起点）
  base_metric:
    name: page_view_count
  # 【必填】转化指标（漏斗终点）
  conversion_metric:
    name: purchase_count
  # 【必填】转化实体类型: FROM_ENTITY | FROM_DIMENSION
  conversion_entity_type: FROM_DIMENSION
  # 【可选】转化维度（conversion_entity_type 为 FROM_DIMENSION 时使用）
  conversion_dimension: user_id
  # 【必填】计算方式: CONVERSIONS (转化值) | CONVERSIONRATE (转化率)
  calculation: CONVERSIONRATE
  # 【可选】窗口类型: WINDOW (N时间内转化) | OFFSET (第N时间转化)
  window_type: WINDOW
  # 【可选】时间窗口, 格式: {数字} {hour|day|week|month}
  window: "7 day"

# -----
# 7. 转化类指标 (CONVERSION) - FROM_ENTITY 示例
# 通过实体（主键/关联键）关联基准指标和转化指标
# -----
- name: register_to_order_conversions
  label:
    - 注册到下单转化数
  description: "用户从注册到首次下单的转化数量，通过用户实体关联"
  type: CONVERSION
  type_params:
    # 【必填】基准指标（漏斗起点）
    base_metric:
      name: user_register_count
    # 【必填】转化指标（漏斗终点）
    conversion_metric:
      name: total_order_amount
    # 【必填】转化实体类型: FROM_ENTITY 表示通过实体（主键）关联两个指标
    # FROM_ENTITY: 两个指标通过同一实体（如用户ID）自动关联，无需指定
  conversion_dimension
    # FROM_DIMENSION: 需要手动指定 conversion_dimension 字段来关联
  conversion_entity_type: FROM_ENTITY
  # 【必填】计算方式: CONVERSIONS (转化数) | CONVERSIONRATE (转化率)

```

```

    calculation: CONVERSIONS
    # 【可选】窗口类型: WINDOW (N时间内转化) | OFFSET (第N时间转化)
    window_type: OFFSET
    # 【可选】时间窗口, 格式: {数字} {hour|day|week|month}
    window: "30 day"

# =====
# 维度定义列表 (可选)
# 支持类型: TIME | CATEGORICAL | DICT
# =====
dimensions:
    # -----
    # 1. 时间维度 (TIME) - 必须指定时间精度
    # -----
    - name: order_date
      # 【可选】维度别名列表, 支持多个, 不允许含逗号
      label:
        - 下单日期
        - 订单时间
      description: "订单创建时间, 精确到天"
      # 【必填】数据源路径, 格式: catalogName.databaseName.tableName
      source: default_catalog.ecommerce_db.orders
      # 【必填】对应的数据库列名
      col_name: created_time
      # 【必填】维度类型
      type: TIME
      # 【TIME类型必填】类型参数
      type_param:
        # 时间精度, 可选值: YEAR | MONTH | DAY | HOUR | MINUTE | SECOND |
        # MILLISECOND
        time_precision: DAY

    # -----
    # 2. 普通维度 (CATEGORICAL)
    # -----
    - name: order_channel
      label:
        - 订单渠道
      description: "订单来源渠道: APP/PC/H5"
      source: default_catalog.ecommerce_db.orders

```

```

col_name: channel
type: CATEGORICAL
# CATEGORICAL 类型不需要 type_param, 不填

# -----
# 3. 字典维度 (DICT) - 关联字典表
# -----

- name: order_status
  label:
    - 订单状态
  description: "订单状态码, 关联字典表"
  source: default_catalog.ecommerce_db.orders
  col_name: status_code
  type: DICT
  # 【DICT类型必填】类型参数
  type_param:
    dict_items:
      - key: Shipped
        value: 已发货
      - key: Completed
        value: 已完成
      - key: Canceled
        value: 已取消

```

示例说明：

- `version: 1.0` 是配置文件版本号, 固定为 `1.0`。
- `model.type = DEF` 时需要提供完整模型定义 (`source` + `joins`) ; 改为 `REF` 时仅需 `name` + `type` , 表示引用已有模型。
- 6 类指标的 `type_params` 形态各不相同:
 - `SIMPLE` 通过 `expr` (聚合表达式) + `time_dimension` 定义;
 - `FILTER` 在 `metrics` 引用的原子指标上叠加 `filter` 条件;
 - `DERIVED` 用 `expr` 引用其他指标名称做四则运算;
 - `RATIO` 用 `derived_type` (`YEAR_ON_YEAR` / `RELATIVE_RATIO`) 声明同环比;
 - `CUMULATIVE` 用 `window` (如 `DAY_7` / `MTD`) 声明时间窗口累积;
 - `CONVERSION` 通过 `base_metric` / `conversion_metric` / `conversion_entity_type` / `calculation` / `window` 声明漏斗转化。
- 3 类维度的 `type_param` 形态:

- `TIME` 必填 `time_precision`（如 `DAY` / `HOURL` / `MONTH`）；
- `CATEGORICAL` 无 `type_param`；
- `DICT` 通过 `dict_items` 声明 key-value 映射。

步骤 4：发布指标

单击编辑器右上角 **发布**：

1. 系统先做校验：YAML 语法、模型权限（CreateMetric / Manage）、表达式是否在底层引擎可执行、所依赖的指标是否存在等。
2. 校验通过后，指标状态变为「已发布」，在指标管理列表可立即被消费。
3. 已发布指标会出现在指标管理列表，可在模型详情页的「关联指标」Tab 看到，并允许其他用户根据 Query 权限消费。

步骤 5：定义全局维度（可选）

维度为工作空间级全局维度，不管是 YAML 创建还是在「维度管理」tab 中创建，均为全局维度。

1. 在语义建模顶部切换到 **维度管理** Tab，单击 **+ 新建维度**。
2. 在 YAML 中声明维度的英文名、中文标签、字段类型、来源字段等。
3. 维度无须发布，保存即生效。
4. 在模型 YAML 中通过 `name` 引用全局维度，避免重复声明。

维度类型

维度用于定义分析视角。按字段的业务语义，常见维度可分为以下几类：

- **时间维度**：用于按时间窗口对指标进行切片的日期、时间类字段，支持按天、周、月、季、年等不同粒度聚合，是趋势分析（如同环比、近 N 天）的核心切片依据。
 - 示例：`order_date`（订单创建日期）、`pay_time`（支付完成时间），常按天、周、月切片。
- **字典维度**：值域固定、可枚举的维度字段，对应业务上的「类目」「品牌」等分类体系。其取值通常来自字典表或预设枚举，字段值域有限且稳定，便于做归一化与标准化管理。
 - 示例：`category_name`（商品类目）、`brand`（品牌）、`gender`（性别）。
- **普通维度**：用于描述业务实体属性的通用维度字段，没有特殊的时间或字典语义，常见为用户姓名、年龄、用户 ID 等属性。
 - 示例：`user_name`（用户名）、`age`（年龄）、`user_id`（用户 ID）。

步骤 6：转交指标 负责人（可选）

每个指标只能有一个 负责人，默认为创建者。在指标详情页 → 口径名下 → 右侧的“业务信息” → 负责人字段 hover 上去会出现修改的 ICON，点击后选择目标负责人提交即可。转交后原 负责人 失去指标管理权限。

指标详情页面

指标发布后，可在「指标管理」列表单击指标名称进入详情页，对指标进行查看、分析、血缘追溯与授权。详情页顶部展示指标基本信息与操作按钮，下方分为 4 个 Tab：口径明细 / 指标分析 / 血缘 / 权限。

顶部公共区域：

- 左侧：返回箭头、指标名、状态徽标（已发布、已下线）
- 面包屑：语义建模、指标管理、指标名
- 收藏按钮：可加入收藏，便于快速回访
- 右上角操作按钮：
 - 编辑：进入 YAML 编辑器，调整口径后重新发布
 - 下线：将指标状态置为「已下线」，停止对外消费
 - 授权：打开授权弹窗，授予其他用户 Manage / Query 权限（注意：需在指标详情页面->权限 Tab 中才展示该按钮。）

口径明细

口径明细 Tab 用于查看指标定义的全部信息，是「指标是什么」的解释页面。

- 口径描述：一段文字说明指标业务含义，例如「The average number of behaviors per session, calculated as total behavior count divided by total sessions.」
- 计算口径：
 - 指标类型：原子、筛选、混合计算、累积、同环比、转化
 - 模型：所属语义模型，可单击跳转至模型详情页面
 - 依赖指标：派生型指标会以卡片形式展示，每个依赖指标一张卡片，含名称、状态（已发布、已下线）、英文名
 - 计算逻辑：公式（如 A / B ）或文字描述，每种指标类型会根据其类型展示不同的计算逻辑信息。
 - 右上角「预览 SQL」：展示指标编译后的查询 SQL，便于排查
- 分析维度：列出该指标支持的分析维度，字段包括：维度名、维度别名、描述、维度精度（如最小颗粒度）、维度信息、负责人、关联字段、最近修订时间
- 业务信息（右侧栏）：别名、负责人
- 技术信息（右侧栏）：指标类型、指标状态、最近更新时间、最近更新人、创建时间、创建人

指标分析

指标分析 Tab 用于对当前指标做即时查询，校验口径与快速取数。顶部提供两种查询模式：

- 表单模式（适合无 SQL 背景用户）：
 - 时间区间：默认近 30 天，可自定义起止日期
 - 时间粒度：日、周、月、季、年、时、分（和每个指标所支持的最小时间粒度有关系）
 - 最大返回记录：可调，默认 30
 - 维度：+ 添加维度和维度值，按所选维度分组，可多选
 - 底部「查询、重置」按钮触发取数

- **SemQL 模式**（适合熟悉查询语法的用户）：
 - 直接编写 SemQL 语句，例如：

```
select * from query(metric=[avg_session_behavior_count], limit=30)
```

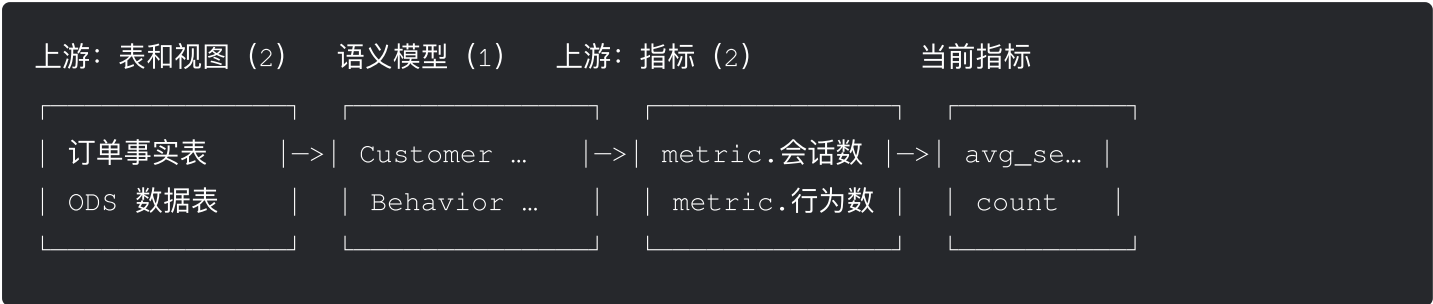
- 同样支持底部「查询、重置」

两种模式均提供右上角「预览查询 SQL」入口，用于查看底层物理 SQL。

血缘

血缘 Tab 展示指标的上下游数据链路，是「指标从何而来、流向何处」的追溯页面。提供两种视图：

- **列表模式**（默认）：以列表形式罗列上游、下游节点
- **图模式**：以图形化方式展示链路，节点之间用箭头表示数据流向。典型链路示意：



- 每个节点支持单击「详情」跳转到对应资源详情页
- 顶部操作：定位到具体节点、放大、缩小、1:1 适配、全屏
- 右侧图例：列出所有可识别的节点类型 —— 数据表、视图表、机器学习模型、指标、语义模型、数据库、NOTEBOOK、SQL、机器学习服务、工作流、离线集成、实时集成

权限

权限 Tab 用于管理指标的授权关系，对应模型层的指标级 ACL。

列表区：

- 右上角「授权」按钮：打开授权弹窗
- 搜索框：按用户筛选
- 权限类型筛选：全部权限 / Manage / Query
- 列表列：授权对象、类型（用户、所有者）、权限点（Manage / Query）、维度、维度值范围（行级权限）、授予人、授权时间、操作
- 「所有者」由创建者自动获得，且不可被撤销；列表中以「所有者」徽标标记

授权弹窗：

字段	是否必填	说明
----	------	----

授权主体	是	可选择用户、用户组、标签
授权对象	是	如果授权主体选用户：此处展示选用户面板，支持搜索用户；至少添加 1 个 如果授权主体选用户组：此处展示选用户组面板，支持搜索用户组；至少添加 1 个 如果授权主体选标签：无需选择授权对象，仅需选下面的授权点即可
授权点	是	至少勾选 1 项；未勾选时提交按钮置灰并提示「请选择权限点」。可支持设置授权对象可查询的指标行列权限。
授权原因	否	多行文本，用于审计与业务背景说明

权限点说明：

- **Manage 管理**：可编辑、发布、下线 指标，以及授权他人。
- **Query 查询**：可查询指标数据，可配置维度值范围。

通过 Buddy 创建指标和维度

Buddy 内置「自然语言创建指标和维度 Skill」，能识别常见聚合术语、字段类型并生成 YAML。

创建指标示例 Prompt

给的信息越详细，Buddy 输出结果越准确。Prompt 建议包含：指标业务含义、所属数据源、模型、聚合表达式、时间字段、过滤条件、依赖指标等关键要素。

用户输入

「在我的 `olist_sales_model` 语义模型下，新增一个 `gmv` 指标。统计逻辑是对 `dlc.default_catalog.olist.orders` 表中的 `pay_amount` 字段求和，用以对所有订单的支付金额求和（总交易额），注意统计时，剔除掉内部客户（客户类型为“internal”）下单的数据」

创建维度示例 Prompt

创建维度 Prompt 建议包含：维度英文名与中文标签、来源表与字段、数据类型、字典 key-value（如有）、所属模型、业务描述、Owner 等。

用户输入

创建「支付时间」维度，表名 `dlc.default_catalog.olist.orders`，字段名 `pay_time`，最小时间粒度按日 DAY，表示订单实际支付完成的时间点，"

注意事项

为了让 Buddy 输出符合预期的 YAML，建议在 Prompt 中尽量说明关键要素：

1. 模型定义关键要素

- 如果要指定数据源表名使用三段式：`catalog.database.table`，例如 `dlc.default_catalog.olist.orders`。
- 字段名要与底层表的实际列名完全一致，避免拼写错误。不指定表名和字段名 AI 也会根据用户输入推断，但是指定后生成的结果会更准确。

2. 指标定义要素

- **指标含义**：定义这个指标是做什么的，如线上渠道交易额等。
- **统计口径描述**：说明指标的业务含义、计算逻辑（可以是自然语言，也可以是统计 SQL）、使用场景，越详细生成的 YAML 越准确。

3. 维度定义要素

- **维度含义** 说明这个维度是什么，如支付时间、支付渠道等。
- **来源表与字段** 同样使用三段式表名、真实字段名，如果不指定，AI 也会根据输入推断，但是指定后生成的结果会更准确。

4. 权限与发布

- Buddy 在生成 YAML 前会校验当前用户是否拥有目标模型的 `CreateMetric` 权限；无权限会拒绝创建。
- 生成 YAML 后会等待用户审核确认 才会真正写入，审核时可再次调整字段后再确认。
- 提交后指标进入「已发布」状态即可在指标管理列表被消费，无需再执行额外动作。

常见问题

Q1：创建指标提示「缺少 CreateMetric 权限」？

说明您没有目标模型的 `CreateMetric` 或 `Manage` 权限。请联系模型 Owner 在模型详情页「授权」Tab 中授予您 `CreateMetric` 权限。

Q2：派生指标的 `formula` 中能引用其他模型的指标吗？

不能。派生指标只能引用同一模型下的指标。

Q3：指标发布失败，提示「表达式校验失败」？

通常原因包括：① 表达式中字段在底层表不存在；② `COUNT(DISTINCT)` 被某些引擎不直接支持，需确认平台支持的去重写法；③ 自定义 SQL 中的函数名不被引擎识别。建议先用底层表在 SQL 编辑器中验证表达式可执行后再发布。

相关文档

- [什么是语义模型](#)
- [创建语义模型](#)
- [语义模型应用](#)
- [Buddy —— 数据工程 Agent 概览](#)

治理标签概述

创建与管理标签体系

最近更新时间：2026-10-09 16:27:41

概述

治理标签（Governed Tag）是账户级、跨工作空间共享的标签字典。本文介绍如何在 **数据目录 > 标签治理** 中创建敏感标签、创建业务标签、维护业务标签的标签值列表，以及对标签进行编辑、搜索和删除等管理操作。

DataBuddy 将治理标签按用途拆分为两个 Tab：**敏感标签** 用于安全合规与脱敏管控，**业务标签** 用于业务分类与资产筛选；两者的字段、列表展示与创建表单都有差异。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号具备数据目录的访问权限（角色权限说明详见 [成员与权限管理](#)）。
- 创建、编辑、删除标签：当前账号需为主账号，或经主账号、标签 owner 授权过创建权限的用户。

使用限制

限制项	说明
标签数量	每个租户最多 1000 个治理标签
标签作用范围	单个数据实体（Catalog / Schema / Table / View / Volume / Model）最多 50 个标签；单张表最多 1000 个列标签
系统标签	系统内置敏感标签（如 <code>class.*</code> ）的名称、类型、描述（中英文）只读，仅可由有权限的用户编辑安全规范与关联脱敏策略
跨工作空间	标签定义在租户级生效，所有工作空间共享一份字典

操作步骤

步骤 1：进入标签治理页面

1. 登录 [DataBuddy 控制台](#)，选择任一空间
2. 在左侧导航栏选择 **数据目录**。
3. 点击页面顶部的 **标签管理**，进入 **标签治理** 页面。
4. 在页面顶部 Tab 中选择 **敏感标签** 或 **业务标签**。

页面默认进入 **敏感标签** Tab，列表中已经包含 DataBuddy 内置的系统敏感标签（如 `class.cn_id_car`、`class.financial_account` 等）。

步骤 2：创建敏感标签

如果您要新增一个用于敏感识别和脱敏管控的标签，按以下步骤操作。

1. 在 **敏感标签** Tab 下，点击右上角 **+ 创建敏感标签**。
2. 在弹窗中填写以下字段（详见下方参数说明）：
 - **标签名称**：必填，作为标签 Key。
 - **描述**：选填。
 - **安全规范**：在 **敏感分类配置** 中勾选适用的合规框架（PII / GDPR / HIPAA / PCI-DSS / GLBA / CCPA / DPDPA / PDPA / PIPL）。
 - **关联脱敏策略**：在下拉框中选择已有策略；如尚未创建，可暂选 **不关联**，后续在标签详情中再绑定。
3. 点击 **确定** 完成创建。新标签会出现在敏感标签列表中，**类型** 列为「自定义」。

① 说明：

敏感标签**不带标签值**。打标时仅指定标签名，所有字段共用统一定义，便于按标签触发同一份脱敏策略。

步骤 3：创建业务标签

业务标签用于按主题域、数据分层、业务线等维度对资产进行分类和筛选。

1. 切换到 **业务标签** Tab。
2. 点击右上角 **+ 创建业务标签**，在弹窗中填写：
 - **标签名称**：必填，作为标签 Key（如 `domain`、`tier`、`pii_review`）。
 - **描述**：选填。
 - **标签值**：可添加一个或多个候选值（如 `gold`、`silver`、`bronze`）；打标时由用户在候选值中单选。
3. 点击 **确定** 完成创建。

① 说明：

如果您希望打标时让用户**自由填写**标签值，可在创建时不预设任何值；打标的字段会以「其他标签」形式出现，由用户在打标弹窗中现场输入值。

步骤 4：维护业务标签的标签值

业务标签支持在创建后继续维护其候选值列表。

1. 在 **业务标签** Tab 列表中，点击目标标签的 **标签名**，进入标签详情页，可以看到该标签的描述、Key 和当前的标签值列表。
2. 在标签值列表中：
 - 点击 **+ 新增** 添加新的标签值。
 - 勾选若干已有标签值后，点击 **批量删除** 进行清理。

3. 修改完成后页面会自动保存。

⚠ 注意：
已创建的标签值**不支持直接重命名**。如需修改一个值的拼写，请先 **新增** 正确的值，再 **删除** 错误的值；同时确认该值已不再被任何资产引用，避免影响已打标的数据。

步骤 5：搜索、编辑与删除标签

1. **搜索：**在 Tab 顶部的搜索框中输入标签名关键字，列表会实时模糊匹配并过滤。敏感标签 Tab 还支持按 **类型**（系统、自定义）和 **安全规范** 筛选。
2. **编辑：**点击标签名称，进入详情弹窗或详情页修改。可编辑字段因类型而异：
 - 系统敏感标签：仅可编辑 **关联脱敏策略**；标签名、类型、描述、安全规范只读。
 - 自定义敏感标签：可编辑描述、安全规范、关联脱敏策略；标签名只读。
 - 业务标签：可编辑描述、标签值列表；标签创建成功后不支持修改。
3. **删除：**点击操作列的 **删除**，在二次确认弹窗中点击 **确定**。

⚠ 注意：
删除一个治理标签后，已经被该标签打标的资产、字段会**自动转换为「其他标签」**保留下来（标签值可继续编辑）。如果该敏感标签关联了脱敏策略，关联关系会被解除，相关字段不再继承该策略，请提前评估影响。

参数说明

敏感标签创建参数

参数	说明	是否必填	默认值
标签名称	作为标签 Key 在全平台唯一识别。	是	—
描述	标签的业务含义说明，最多 500 个字符	否	空
安全规范	标签所属的合规框架，可多选；用于在标签列表中以 Badge 形式展示	否	空
关联脱敏策略	与该敏感标签绑定的脱敏策略；字段被打上标签后自动继承该策略	否	不关联

安全规范 当前支持的取值（与列表展示一致）：

取值	含义
----	----

PII	个人身份识别信息（Personally Identifiable Information）
GDPR	欧盟《通用数据保护条例》
HIPAA	美国《健康保险流通与责任法案》
CCPA	加州《消费者隐私法案》
PCI-DSS	支付卡行业数据安全标准
GLBA	美国《金融服务现代化法案》
DPDPA	印度《数字个人数据保护法》
PDPA	新加坡 / 泰国等《个人资料保护法》
PIPL	中国《个人信息保护法》

业务标签创建参数

参数	说明	是否必填	默认值
标签名称	作为标签 Key（如 <code>domain</code> 、 <code>tier</code> ）	是	—
描述	标签的业务含义说明	否	空
标签值	一个或多个候选 Value；打标时由用户单选其中之一	否	空（不预设值）

标签列表字段

敏感标签 Tab：

列	说明
名称	标签 Key，可点击进入详情
类型	系统 / 自定义，可筛选；列表默认按「自定义在前、系统在后」+ 标签名字母序排列
描述信息	标签的描述
安全规范	以 Badge 列表形式展示已勾选的合规框架
脱敏策略	已关联：展示策略名（可点击跳转策略详情）；未关联：显示「△ 去关联」入口
创建人	系统标签为 <code>system</code> ；自定义标签为创建人账号
创建时间	支持按时间排序

业务标签 Tab：

列	说明
名称	标签 Key，可点击进入详情页
描述信息	标签的描述
标签值	该标签下的候选值，多个时折叠展示
创建人	创建人账号
创建时间	支持排序
操作	删除（二次确认）

验证结果

完成创建后，您可以通过以下方式确认：

1. 在对应 Tab 列表中可以搜索到刚创建的标签，**类型** 列为「自定义」。
2. 进入任一表的详情页，在字段标签或表标签的打标弹窗中可以选择到新创建的标签；如果是业务标签且预设了值，标签值下拉中也会出现已配置的候选值。
3. 对于敏感标签，关联脱敏策略后，将其打到任意字段上时，该字段的「脱敏策略」列会自动显示策略名 + 「标签继承」标记。

常见问题

Q：自定义敏感标签的命名是否需要遵循 `class.*` 前缀？

A：不强制。`class.*` 是 DataBuddy 内置系统标签的命名约定，自定义敏感标签可以使用更贴近业务的命名（例如 `internal_user_phone`），只需避免使用 `.`、`,`、`-`、`=`、`/`、`:` 等保留字符。

Q：业务标签可以转换为敏感标签吗？反过来呢？

A：不支持类型转换。如果需要切换用途，请新建目标类型的标签，将资产重新打标后再删除旧标签。

Q：删除标签会影响哪些资产？

A：删除后，已打该标签的资产会自动转为「其他标签」保留标签信息，不会丢失数据；但敏感标签的脱敏策略关联会被解除，原本通过标签继承策略的字段将不再脱敏，请评估后再执行。

Q：所有用户都能创建治理标签吗？

A：默认仅主账号具备完整管理权限。其他用户需经主账号或标签 owner 授权获得「创建」权限后，才能新建治理标签。具体权限模型请参考 [成员与权限管理](#)。

相关文档

- [治理标签概述](#)
- [为数据资产打标签](#)
- [基于标签的策略控制](#)
- [列级权限控制](#)
- [按标签、分类筛选](#)

为数据资产打标签

最近更新时间：2026-10-09 16:27:46

概述

打标签 是将治理标签附着到具体数据资产或字段上的动作。本文介绍如何在 Catalog、Schema、Table、View、Volume、Model 详情页以及表字段上添加治理标签、修改标签值、移除标签，并说明敏感标签与业务标签在打标交互上的差异。

通过打标，您可以让该资产被数据发现按标签筛选，也可以让敏感字段自动继承标签所关联的脱敏策略，实现「打标即治理、打标即脱敏」。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号已加入目标 Workspace，并对目标资产具备 **修改** 权限（具体权限点见 [成员与权限管理](#)）。
- 已在 [创建与管理标签体系](#) 中完成标签字典的搭建（如需打**已有的** 治理标签）。

使用限制

限制项	说明
可打标资产	Catalog、Schema、Table、View、Volume、Model 以及表字段（Column）
单实体标签上限	每个数据实体最多 50 个标签
单表列标签上限	单张表最多 1000 个列标签（所有字段累计）
计算资源	治理标签 不 用于计算资源（计算资源 / 任务使用独立的标签机制）
敏感标签的标签值	敏感标签不带标签值，打标时仅选择标签名
业务标签的标签值	业务标签若预设了候选值，打标时只能在候选值中单选；若未预设，则在「其他标签」中由用户自由填写

操作步骤

步骤 1：进入资产详情页

打标入口分散在不同资产详情页，但交互一致。常用入口如下：

资产类型	入口
Catalog	数据目录 > 选中目标 Catalog > 右侧 业务信息 > 添加标签
Schema	数据目录 > 选中 Catalog > 选中 Schema > 右侧 业务信息 > 添加标签
Table / View	Schema 详情页 > 点击表 / 视图名 > 右侧 业务信息 > 添加标签
Volume	Schema 详情页 > 点击 Volume 名 > 右侧 业务信息 > 添加标签
Model	Schema 详情页 > 点击模型名 > 右侧 业务信息 > 添加标签
表字段	Table 详情页 > 概览 Tab > 字段列表 > 字段所在行的 标签 列

进入详情页后，**标签** 区域默认展示已绑定的标签；超出展示数量的会折叠为 **+N** 。

步骤 2：添加标签

- 在**标签** 区域右侧的小笔图标处点击，打开打标弹窗。
- 在搜索框中输入标签名关键字。匹配结果按类型分组展示：
 - 治理标签**：来自标签治理中维护的字典。
 - 其他**：当前用户在打标过程中自定义的非治理标签（仅作为临时备注存在）。
- 选择目标标签后，按以下规则填写标签值：
 - 敏感标签**：无需填写值。
 - 业务标签、预设候选值**：在 Value 下拉中单选其中一个候选值。
 - 业务标签、未预设值 或其他标签**：在 Value 输入框中自定义输入。
- 点击**确定** 保存。已选标签会出现在资产的标签列表中。

i

说明：

如果搜索框中的标签名在治理标签和其他标签中都不存在，搜索结果下方会出现 **+ 创建新标签「{名称}」** 入口。点击后会**直接生成一个「其他」类型标签**（即非治理标签），并允许您自由填写标签值。该标签不会进入标签治理字典。

步骤 3：为表字段打标

字段标签是治理工作的核心场景，特别是用于敏感识别。

- 进入 Table 详情页，停留在 **概览** Tab。
- 在字段列表中找到目标字段，点击该行 **标签** 列的小笔图标。
- 在弹窗中按 **步骤 2**的方式添加治理标签。
- 保存后，字段所在行会展示已打的标签。

打上敏感标签后，如果该标签关联了脱敏策略：

- 字段的 **脱敏策略** 列会自动显示策略名，并带有 **标签继承** 标记，表示策略来源于标签而非手动配置。
- 当治理团队后续更换该敏感标签关联的脱敏策略时，所有「标签继承」的字段会自动跟随新策略；之前已**手动**指定脱敏策略的字段不受影响。

更多内容请参考 [基于标签的策略控制](#)。

步骤 4：修改或移除标签

- 在资产或字段的标签区域点击小笔图标，打开标签弹窗。
- 修改标签值：
 - 业务标签：在 Value 下拉、输入框中重新选择或输入新值。
 - 敏感标签：无标签值，跳过。
- 移除标签：在弹窗中点击该标签的 **删除** 按钮。
- 点击 **确定** 保存。

① 说明：

移除标签只解除资产与标签的绑定关系，标签字典本身仍然保留。如需彻底删除标签字典中的标签，请到 [创建与管理标签体系](#) 中执行删除操作。

步骤 5：在数据发现中按标签筛选

打标完成后，可以在数据发现页验证标签的可用性：

- 进入 [数据发现](#)，切到 **表 / 视图 / 函数 / 卷** 等 Catalog 类资产 Tab。
- 在顶部筛选条中找到 **治理标签** 维度，选择刚刚打的标签、标签值。
- 列表会收敛到所有命中该标签条件的资产。

详细筛选交互参考 [按标签、分类筛选](#)。

参数说明

打标弹窗字段

参数	说明	是否必填	默认值
标签名称	通过搜索选中已有治理标签，或现场创建「其他」类型标签	是	—
标签值	业务标签的 Value；敏感标签不需要填写	视类型	空

资产详情页中的「标签」展示

资产详情页右侧 **业务信息 > 标签** 区域：

元素	说明
----	----

标签徽章	<code>key:value</code> 形式展示，敏感标签仅展示 <code>key</code>
系统标签精简显示	形如 <code>class.location</code> 的系统敏感标签在 Catalog / 字段处仅显示 <code>.</code> 后的部分（如 <code>location</code> ），便于阅读
<code>+N</code>	超出默认展示数量的标签折叠数
小笔图标	编辑入口，点击打开打标弹窗

字段标签列字段（Table 详情 – 概览 Tab）

列	说明
字段名称	表字段名
字段类型	引擎侧元数据
字段描述	业务说明，可编辑
标签	已打的治理标签列表，可编辑
脱敏策略	该字段当前生效的脱敏策略；标签继承的策略会附 <code>标签继承</code> 标记

验证结果

操作完成后，您可以通过以下方式确认：

- 在资产或字段的 `标签` 区域，刚添加的标签徽章已经出现，鼠标悬停显示完整 `key:value` 。
- 在数据发现页按该标签筛选，能命中刚打标的资产、字段。
- 如果是带有脱敏策略的敏感标签，字段的 `脱敏策略` 列已自动显示策略名 + `标签继承` 标记。

常见问题

Q：能否对一个字段同时打多个敏感标签？

A：可以。例如同一个字段可以同时打 `class.cn_id_card` 和 `class.pii_name` 。但若多个敏感标签都关联了脱敏策略，字段最终生效的策略以该敏感标签关联的脱敏策略叠加、覆盖规则为准，详见 [基于标签的策略控制](#) 。

Q：在打标弹窗中创建的「其他」类型标签会进入治理标签字典吗？

A：不会。+ `创建新标签` 入口创建的是临时备注标签，仅附着在当前资产上，不会出现在 `标签治理` 列表中，也不会用于敏感识别和脱敏。如需作为治理标签复用，请到 [创建与管理标签体系](#) 中正式创建。

Q：批量给多个字段打同一个敏感标签的入口在哪里？

A：在 Buddy 中可以通过自然语言或一键扫描批量识别敏感字段并自动打标，详见 [Buddy](#) 章节相关内容。手动逐字段打标仍以本文步骤为准。

Q：表字段的标签数量是否单独计数？

A：单张表的字段标签累计上限是 1000 个；表本身的标签上限与其他实体一致，每个实体最多 50 个标签。

相关文档

- [创建与管理标签体系](#)
- [基于标签的策略控制](#)
- [按标签、分类筛选](#)
- [列权限控制](#)
- [数据目录概述](#)

基于标签的策略控制

最近更新时间：2026-10-09 16:27:49

概述

DataBuddy 提供「**标签 → 策略**」的联动机制：把脱敏策略关联到一个敏感治理标签上之后，所有打了该标签的字段会**自动继承**该策略，无需逐字段手动配置。本文介绍如何为敏感标签关联脱敏策略、字段如何继承策略、更换或解绑策略时的联动规则，以及如何在治理标签列表中检查未关联策略的标签。

该机制的目标是把「敏感识别」与「脱敏管控」拉到同一条数据上：**识别一次、关联一次、自动生效**。

前提条件

- 已开通 DataBuddy 服务并完成空间初始化。
- 已在 **数据安全 > 脱敏策略** 中创建好需要使用的脱敏策略，详见 [列级别权限控制](#)。
- 当前账号对目标敏感标签具备编辑权限（系统标签也允许有权限的用户更新关联策略）。
- 已了解治理标签的基本概念与打标方法，详见 [创建与管理标签体系](#) 与 [为数据资产打标签](#)。

使用限制

限制项	说明
仅敏感标签	仅 敏感标签 可关联脱敏策略；业务标签无此能力
一对一关联	一个敏感标签同一时刻只能关联一个脱敏策略
策略来源	关联的策略必须先在 脱敏策略 中存在；可在弹窗中通过 + 创建脱敏策略 → 入口跳转创建后再回到本流程
系统标签	系统敏感标签的名称、描述、类型只读，但 关联脱敏策略 仍可由有权限的用户修改
继承范围	仅当字段被实际打上该敏感标签时，才会从标签继承策略

操作步骤

步骤 1：在标签治理中关联脱敏策略

为新建的敏感标签关联策略，最便捷的入口是创建标签时直接配置；也支持事后在标签详情中绑定。

方式 A：创建敏感标签时一次性配置

1. 在 **数据目录 > 标签管理 > 敏感标签** Tab 中，点击 **+ 创建敏感标签**。
2. 在弹窗的 **敏感分类配置** 中找到 **关联脱敏策略** 下拉框：

- 在下拉中选择已有策略 → 直接关联。
- 选择 **不关联** → 后续在标签详情中再配置。
- 点击下拉底部的 **+ 创建脱敏策略** → 跳转到 **脱敏策略** Tab 创建新策略，创建完成后回到敏感标签 Tab 继续操作。

3. 点击 **确定** 完成。

方式 B：在已有标签上添加、更换关联

1. 在敏感标签列表中，找到目标标签（系统或自定义均可）。
2. 通过以下任一入口打开标签详情：
 - 点击 **名称** 列的标签名。
 - 在 **脱敏策略** 列：未关联的标签显示「△ 去关联」，点击后会打开详情弹窗并自动聚焦到 **关联脱敏策略** 区域。
3. 在详情弹窗的 **选择目标脱敏策略** 区域：
4. 点击 **保存** 完成。系统会同步刷新所有已经打了该标签的字段所继承的策略。

提示

+ **创建脱敏策略** → 入口会关闭当前详情弹窗、切到脱敏策略 Tab 并打开创建策略弹窗。新策略创建完成后，需要回到治理标签 Tab 重新打开本标签的详情弹窗，再选择刚创建的策略。

步骤 2：更换或解绑标签的关联策略

更换策略：

1. 在敏感标签详情中将关联策略切换为新策略并保存。
2. 所有「标签继承」的字段会自动切换到新策略；手动指定策略的字段不受影响。

解绑策略：

1. 在敏感标签详情的 **关联脱敏策略** 区域，点击**取消绑定**，并保存。
2. 该标签下所有「标签继承」的字段会失去脱敏策略，**脱敏策略** 列回退为空，字段不再脱敏。
3. 在治理标签列表的 **脱敏策略** 列中，该标签恢复为「△ 去关联」状态。

注意：

解绑会让字段立即失去脱敏。请提前在 **脱敏策略** 处准备好替代策略，或在解绑前对受影响字段单独配置手动策略，避免敏感数据明文返回。

步骤 3：脱敏策略侧的反向联动

在 **数据安全 > 脱敏策略** 中删除一个策略时，治理标签会同步联动：

1. 删除策略后，所有原本关联该策略的敏感标签的 **脱敏策略** 列回退为「△ 去关联」。
2. 通过这些标签继承策略的字段会同步失去脱敏。

3. 在脱敏策略详情中，您可以反向查看该策略关联了哪些敏感标签、以及这些标签影响了哪些字段，便于在删除前评估影响。

步骤 4：盘点未关联策略的敏感标签

在 敏感标签 Tab 中，您可以快速识别哪些敏感标签已经存在但还没有关联脱敏策略，从而打成「打标但不脱敏」的漏洞。

- 1. 在敏感标签列表中，按 脱敏策略 列筛选「未关联」。
- 2. 列表会展示所有 脱敏策略 列为「△ 去关联」的敏感标签。
- 3. 逐个进入详情，按上述步骤1中的方式 B 关联策略。

参数说明

敏感标签的关联脱敏策略字段

参数	说明	是否必填	默认值
关联脱敏策略	该敏感标签所绑定的脱敏策略；下拉选择已有策略，或选择「不关联」	否	不关联

标签详情弹窗中的可编辑字段

字段	系统敏感标签	自定义敏感标签
标签名	只读	只读（创建后不可改名）
类型	只读	只读
描述	只读（系统提供中英文描述）	可编辑
安全规范	只读	可编辑
关联脱敏策略	可编辑	可编辑

验证结果

操作完成后，您可以通过以下方式确认：

- 1. 在敏感标签列表中，目标标签的 脱敏策略 列已显示绑定的策略名
- 2. 进入任意一张已打该敏感标签的表，对应字段的 脱敏策略 列已显示策略名 。
- 3. 通过有权限的用户查询该字段，返回结果按策略规则脱敏（具体行为以策略定义为准）。

常见问题

Q：一个敏感标签可以关联多个脱敏策略吗？

A：不支持。一个敏感标签同一时刻只能关联一个脱敏策略。如果有多套脱敏诉求（例如对外查询和内部分析使用不同规则），建议拆分为多个敏感标签，分别关联不同策略。

Q：字段能同时被多个敏感标签关联吗？

A：不支持，一个字段只能同一时刻关联一个敏感标签。

Q：业务标签是否也能挂脱敏策略？

A：不能。脱敏策略关联是敏感标签专属能力。业务标签用于业务分类与资产筛选，不参与脱敏管控。

Q：在 Buddy 中通过自然语言打上的敏感标签，是否也会继承脱敏策略？

A：会。Buddy 触发的打标和手工打标走同一条数据通路；只要该敏感标签在治理标签中已关联脱敏策略，字段就会自动继承。

相关文档

- [创建与管理标签体系](#)
- [为数据资产打标签](#)
- [列级权限控制](#)
- [数据安全概述](#)