

统一 Catalog 服务 操作指南



腾讯云

【 版权声明 】

©2013–2025 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【 商标声明 】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。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。

【 联系我们 】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或 95716。

文档目录

操作指南

数据目录

导入外部数据源

非结构化数据管理

操作指南

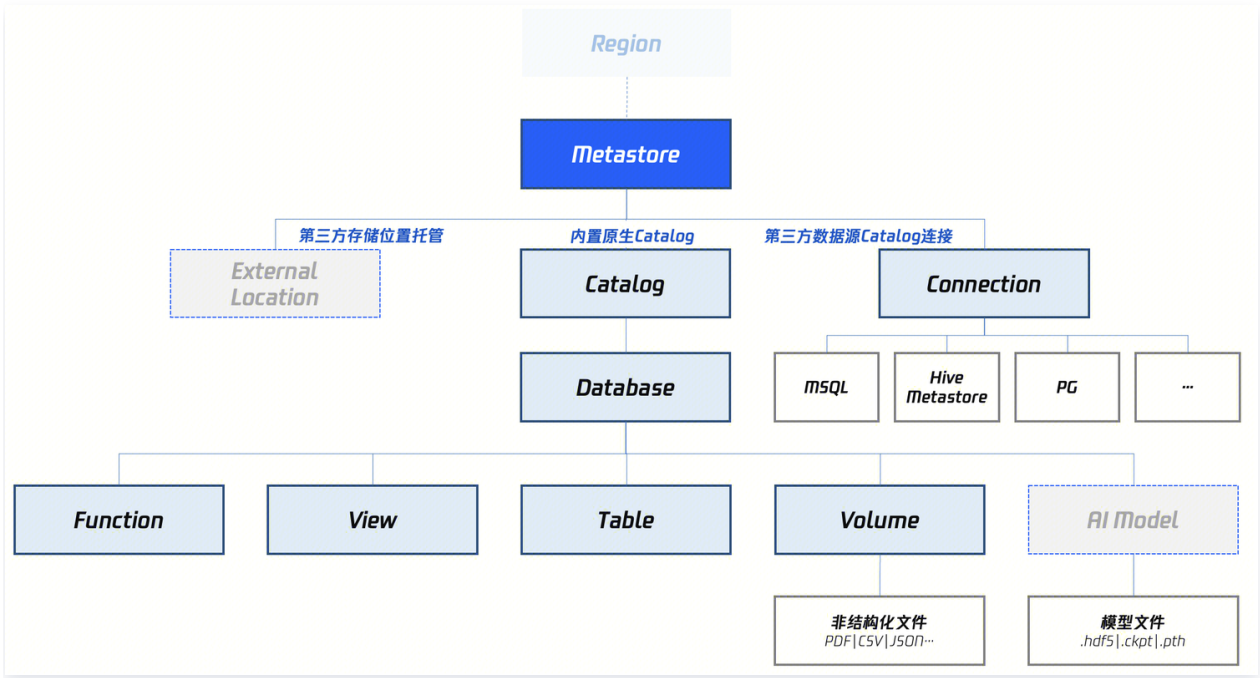
数据目录

最近更新时间：2025-04-18 15:09:23

数据目录（Catalog）是统一 Catalog 服务的元数据最上层实体，它可以包含多个 Schema。主要适用于元数据隔离的场景。本文为您介绍数据目录（Catalog）的基本操作。

统一 Catalog 层级模型

在统一 Catalog 中，所有元数据都在一个元数据存储（Metastore）中进行注册和存储，Metastore 本身对用户不可见。任何统一目录的元数据对象层级结构分为三个级别，当您引用表、数据卷、模型和函数时，会以三级命名空间（Catalog.Schema.Table 等）的形式呈现。



第一级：Catalog

数据目录用于组织不同形式的数据资产，目前腾讯云统一 Catalog 中数据目录分为以下类型。

大类	小类	功能
内置数据目录	Hive Catalog	完全兼容 Hive Metastore 的内置数据目录，可为 EMR 等产品提供 Metastore Service，替代 MySQL 等 MetaDB（本部分功能开发中）。
	Volume Catalog	面向非结构化数据的数据卷（Volume）目录，可将对象存储等文件系统中的图片、视频、音频关联到 Volume 中，实现对非结构化数据的元信息统一管理。

	Model Catalog	面向 ML 模型文件的内置数据目录，可将 MLFlow 等 MLOps 框架中训练好的 ML 模型注册到 TC-Catalog 中，实现 ML 模型文件的全生命周期管理（本部分功能开发中）。
外部数据目录	MySQL、EMR、DLC、TCHouse 等	指通过 JDBC 等方式建立与如 MySQL 等外部数据源的连接，实时获取到该数据源的元数据信息。

第二级：Scehma

Scehma（也可称为数据库）是数据目录下的二级对下，可根据具体的数据目录类型包含表、视图、数据卷、ML 模型和函数等具象化的实体数据资源。Scehma 将数据和 AI 资产组织成比目录更细化的逻辑类别。

第三级：具体数据资源

在统一 Catalog 数据目录模型的第三级，根据具体数据目录类型不同可为 Table、Volume、Model 等具象化的实体数据资源。

表（Table）/视图（View）

表是托管到统一 Catalog 服务中、按行和列组织数据集合的具体库表。视图是针对一个或多个表保存的查询。

数据卷（Volume）

数据卷是用于关联对象存储或 HDFS 等文件系统中非结构化数据的逻辑卷。例如，将一个对象存储路径

`examplebucket.cos.ap-guangzhou.myqcloud.com/folder/` 下的文件（假设有 `a.jpg/b.csv` 两个文件）关联到 `MyCatalog.MyScehma.MyVolume` 后，计算引擎通过 `MyCatalog.MyScehma.MyVolume/a.jpg` 即可直接访问到图片 `a`。

⚠ 注意：

数据卷（Volume）仅可在类型为 Volume 的数据目录中创建。

模型（Model）

模型是从 MLflow 注册到统一 Catalog 服务中的 ML 模型，本功能正在开发中。

⚠ 注意：

模型仅可在类型为 Model 的数据目录中创建。

函数（Function）

函数是用户保存到统一 Catalog 服务的自定义函数（UDF），可返回如一个标量值或一组行（本功能规划中）。

新建数据目录

1. 登录 [统一 Catalog 服务控制台](#)。
2. 在数据目录列表页面，单击**创建数据目录**。
3. 在弹出对话框中输入以下内容：
 - **数据目录名称**：输入数据目录名称，不可重名。
 - **数据目录类型**：目前支持**内置数据目录**和**外部数据目录**，内置数据目录的元数据信息内置存储于腾讯云统一 Catalog 服务中，支持兼容 Hive Metastore 的元数据服务（开发中）和支持费结构化数据统一管理的数据卷（Volume）数据目录类型。外部数据目录是指通过 JDBC 等方式建立与如 MySQL 等外部数据源的连接，实时获取到该数据源的元数据信息。

数据目录类型	说明	元数据位置	支持数据类型	适用场景
内置数据目录	TC-Catalog 提供的标准 Metastore Service，支持托管 EMR 中的元数据和提供非结构化数据 Volume 管理	统一 Catalog 服务中	● 兼容 Hive Metastore 的标准数据目录（本部分功能开发中） ● 数据卷（Volume）	● EMR 集群元数据托管（开发中） ● 非结构化数据统一管理
外部数据目录	指通过 JDBC 等方式建立与如 MySQL 等外部数据源的连接，实时获取到该数据源的元数据信息	存储于数据源，通过连接实时拉取元数据	MySQL、EMR-Hive、DLC、TCHouse等	跨源联合分析

- **描述**：可选，输入描述信息。

查看数据目录

1. 在统一 Catalog 服务左侧菜单栏，选择**数据目录**，在树状数据目录浏览器可选择对应数据目录以及旗下如 Schema、Table 等层级进行查看。
2. 选择一个数据目录后，在右侧表单上方的**详情**中，可查看数据目录的基本信息，如希望关注某个目录或 Table，可点击上方**星号收藏**。



编辑数据目录

1. 在数据目录列表页面，找到需要修改的数据目录，在右侧操作选项中选择**编辑**。
2. 在弹出对话框中修改对应数据目录的设置信息。

删除数据目录

在数据目录列表页面，找到需要修改的数据目录，在右侧操作选项中选择**删除**。

警告：

在同一 Catalog 中删除数据目录为元数据层面操作，对应数据物理文件不会被删除。

导入外部数据源

最近更新时间：2025-01-21 17:37:02

在统一 Catalog 服务中，您可通过创建外部数据目录连接如 MySQL 等其他数据源。创建后统一 Catalog 服务会自动同步数据源下的 Database 和 Table 等元数据信息，以便让上层计算引擎快速访问这些外部数据。

创建外部数据目录

1. 登录 [统一 Catalog 服务控制台](#)，在数据目录列表页面单击**创建数据目录**。
2. 数据目录类型选择**外部数据目录**，连接类型可根据需要连接的数据源选择，当前支持 MySQL、EMR、DLC、TCHouse-D。
3. 在实例下拉框中，选择需要连接的具体数据源实例，注意如选择的是 DLC 实例则为选择 DLC 下需要连接的 Catalog。
4. 如选择 MySQL/TCHouse-D，需填写该实例 JDBC 连接串、用户名密码等信息，其中 JDBC 连接串会被自动填充。
5. 确认自动填充的实例 VPC 信息后，单击确定则可创建外部数据目录。

⚠ 注意：

1. 当前仅支持连接到统一 Catalog 服务所在地域的数据源。
2. 为确保统一 Catalog 可访问上述资源，请先确认资源所在的 VPC，提交后系统将会自动打通与该 VPC 的网络连接。

连通性测试

完成外部数据目录创建后，通常要等待1~2分钟完成网络打通，过程中您可刷新页面查看完成进度。完成网络打通后，可通过单击右侧**测试**发起连通性测试，如连通性测试不通过，可单击编辑确认实例相关信息，如修改 JDBC 连接串或用户名密码登信息。

数据目录名称	连接类型	创建时间	状态	所有者	操作
test_inner_1 	VOLUME	2024-11-29 16:56:...	连接成功		编辑 删除 测试

📌 说明：

统一 Catalog 服务通过 JDBC 等连接方法从数据源中自动同步元数据信息，目前并不会存储相关元数据信息。

外部数据目录

1. 在数据目录列表页面，找到需要修改的数据目录，在右侧操作选项中选择编辑。
2. 在弹出对话框中修改对应数据目录的设置信息。

非结构化数据管理

最近更新时间：2025-04-18 15:09:23

在统一 Catalog 服务中，您可通过数据卷（Volume）统一管理各类实际存储在对象存储或 HDFS 文件系统中的非结构化数据。

什么是数据卷（Volume）

数据卷是统一 Catalog 服务中非结构化文件对象的集合，您可通过 Volume 对非表格数据集进行统一治理。数据卷代表对象存储某一位置中的一个具体逻辑存储卷，通过统一 Catalog 服务的 Volume 虚拟路径，可赋予上层计算引擎快速访问这些文件并理解其元信息的能力。

数据卷优势

在传统的非结构化数据管理实践中，往往需直接使用文件系统的物理存储路径和 Credential 信息（如对象存储 Access Key/Seret Key），而在多人协作的数据科学开发中 Credential 分发无疑将造成潜在风险，不规范的路径也会提示交付成本和出错概率。

通过数据卷功能，可将复杂的多云物理路径映射为逻辑统一、权限统一的 volume 架构，构建具备业务分类意义的非结构化数据目录体系。核心价值：

1. 统一管理：构建逻辑统一、具备业务分类意义的目录。
2. 便于迁移：算法和存储介质解耦，支持无感数据迁移。
3. 效率提升：减少不规范的路径的交付成本和出错概率。

创建数据卷

第一步：创建 Volume Catalog

1. 登录 [统一 Catalog 服务控制台](#)，在数据目录列表页面单击**创建数据目录**。
2. 数据目录类型选择**内部数据目录**，存储数据类型选择数据卷（Volume）。
3. 单击确认，即创建了一个可关联非结构化文件的数据卷目录。

第二步：创建 Volume

1. 在创建 Volume 之前，需要先在数据目录下创建一个 Schema。
2. Schema 创建成功后，在 Schema 下**单击创建数据卷**。
3. 在弹出对话框，输入 Volume 名称，随后选择一个需要关联的对象存储路径。

⚠ 注意：

Volume 将默认关联该路径 folder 下所有的文件到 Volume 中。

4. 单击确认，即创建一个映射到上述对象存储路径的 Volume。

第三步：查看 Volume

在统一 Catalog 服务左侧菜单栏，选择**数据目录**，在树状数据目录浏览器可选择对应数据目录以及旗下 Volume 等层级进行查看。