

# 统一 Catalog 服务

## 产品简介



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

## 产品简介

产品概述

产品优势

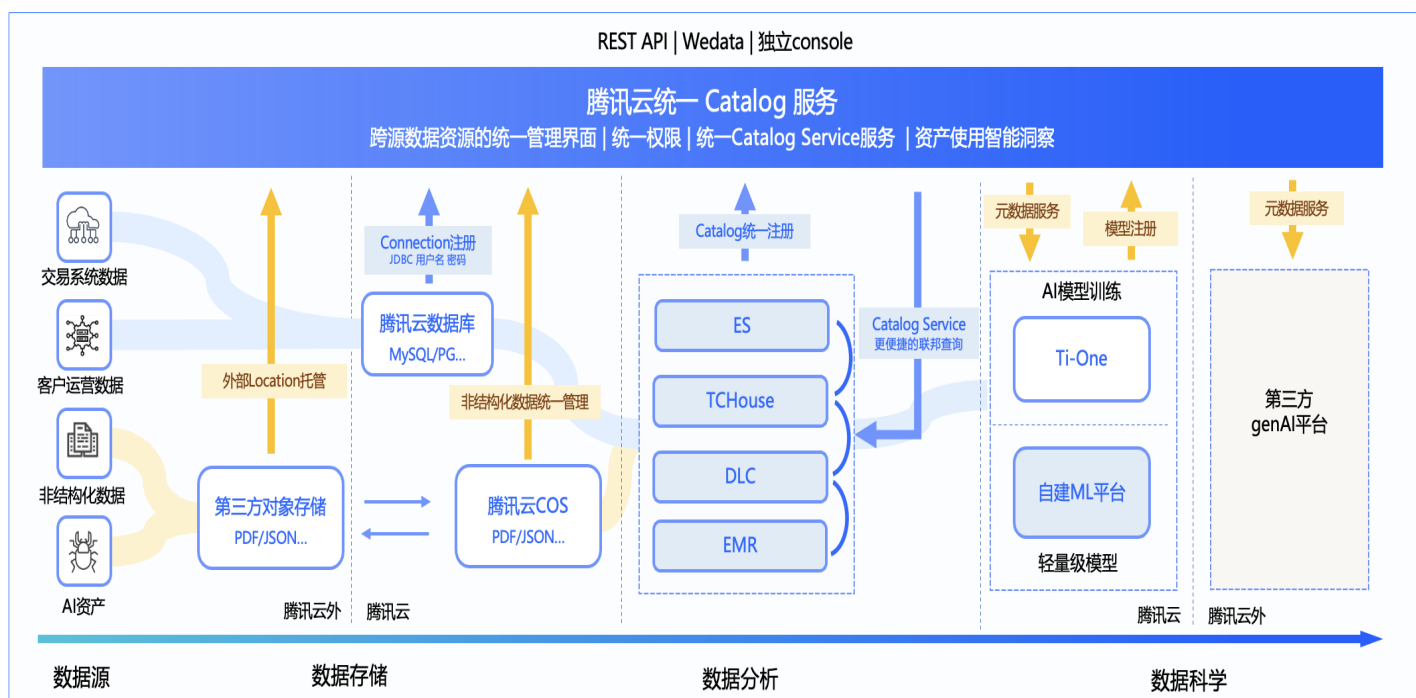
应用场景

# 产品简介

## 产品概述

最近更新时间：2025-01-21 17:37:02

统一 Catalog 服务（简称 TC-Catalog）是腾讯云新一代的多模态统一元数据管理平台，提供覆盖传统结构化数据（如 Table/View）、半/非结构化数据（如文本/音频/视频）和 AI 模型的统一元数据管理服务，并具备完整的统一权限管理功能。TC-Catalog 支持整合腾讯云多种数据源，及未来将进一步支持跨云 Location 托管，帮助用户在大数据+AI 场景中有效消除数据孤岛，实现一体化数据资产管理。同时支持对接多种大数据、搜索和 AI 引擎，实现多引擎便捷的协同计算，从而基于同一份湖存储数据快速构建 Data+AI 一体化应用。



## TC-Catalog 核心功能

### 多模态数据源统一管理

实现对腾讯云数据库、数据湖、数仓系列产品的元数据统一管理，同时支持对象存储中的 AI/ML 训练数据（如语料/图片/视频）的统一纳管，支持 AI 模型文件的生命周期管理，彻底打破结构化+非结构化数据+AI 模型管理的传统鸿沟。

### 实现大数据+AI 一体化

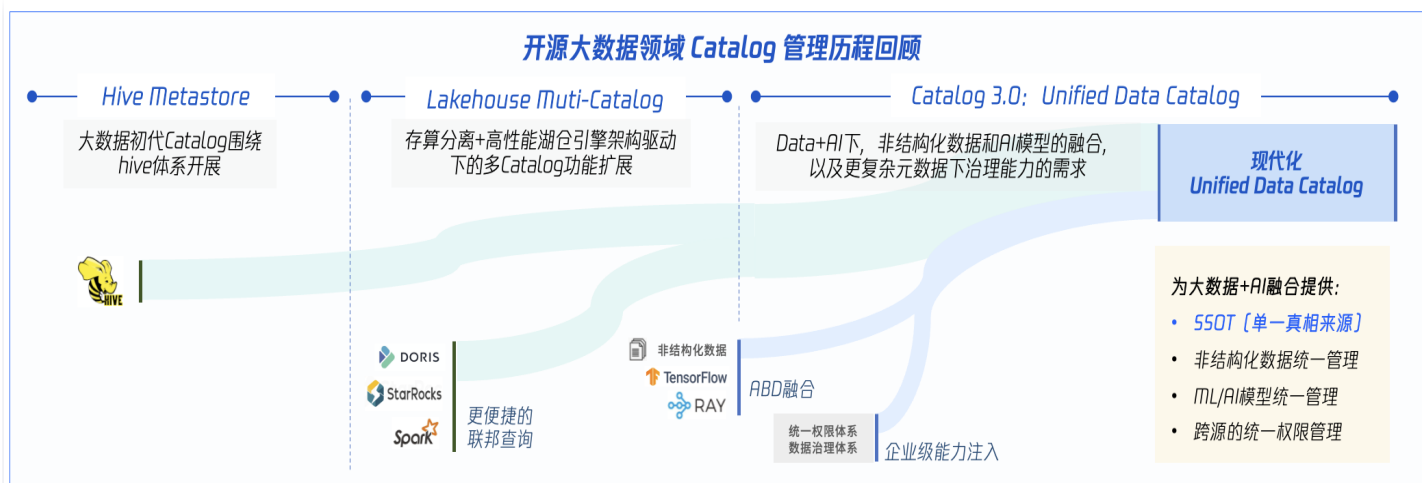
TC-Catalog 逐步支持 Spark、StarRocks、Hive 等多种大数据计算引擎 Connector，并逐步对接支持 Ray、Pandas、TensorFlow 等主流 AI 训练框架，大数据+AI 引擎可平权统一访问 Catalog 中的结构化+非结构化数据+模型文件，减少数据在大数据系统和 AI 系统的搬迁。

### 统一权限管理

通过 TC-Catalog 内置权限体系，统一管理多模态数据源的用户权限，免去维护如对象存储 AK/SK、MySQL 用户名密码、Kerberos 等多种体系的权限配置，提升效率并减少泄露风险。

## 产品定位与背景

TC-Catalog 定位于新一代的多模态 Catalog 管理平台，有别于传统主要对结构化数据进行管理的元数据产品，其产生及推出背景与 Data+AI 时代下大数据平台形态的演进息息相关。



- 在传统 Hadoop 时代，Hive Metastore 作为元数据管理的事实标准被用户广泛应用，但 Hive Metastore 整体缺乏对多 Catalog 的良好支持。随着 Lakehouse 架构兴起，在存算分离下高性能MPP引擎对多个异构数据源进行跨源查询成为了重要的应用场景，以 StarRocks、Doris 为代表的内置 Muti-Catalog 一定程度提升了跨源查询的便捷性，但整体治理功能仍偏弱，也缺乏对 ML 及非结构化数据的支持。Data+AI 时代到来后，大数据系统与 AI 系统将逐渐融合，如一个典型的 Agent RAG 应用，需同时在用户行为数据和非结构化业务知识库中召回上下文结果，大数据与 AI 系统边界变得模糊，业界逐渐往 Data+AI 充分融合的数据智能 (Data Intellegence) 平台演进。
- Data+AI 时代，需要新一代可覆盖传统大数据领域结构化数据、AI领域的非结构化数据源与模型资产的多模态 Unified Data Catalog，为 Data+AI 平台提供多模态数据源的统一元数据服务 (SSOT-”单一真相来源“)，从而有效收敛数据口径等差异带来的业务问题。同时让大数据和AI计算引擎均能平权访问与计算多种异构的数据源，构建 Data+AI 时代的新一代 Data Infra。
- 在此背景下，腾讯云统一 Catalog 服务定位于 Data+A I新一代多模态元数据管理平台，提供覆盖 AI 领域的多模态元数据管理服务，帮助用户打破大数据和 AI 系统的数据孤岛，协助构建数据强一致性的 Data Infra 底座，加速从数据洞察到 GenAI 一体化应用的进程。

# 产品优势

最近更新时间：2025-01-21 17:37:02

## 多模态数据管理

TC-Catalog 为多种类型的数据源抽象了统一的元数据模型和 API 接口，支持多种结构化数据源，如腾讯云数据库、数据湖、TCHouse 等，同时提供非结构化数据源（如腾讯云 COS、HDFS 等）的统一文件元数据管理，支持 ML 模型文件的生命周期管理。

## Data+AI 一体化

TC-Catalog 支持 Spark 等多种大数据计算引擎及 Ray、Pandas 等主流 AI 训练框架（引擎逐步支持中），大数据 + AI 引擎可平权访问 Catalog 中的结构化 + 非结构化数据资源。

## 统一权限管理

通过 TC-Catalog 内置权限体系，统一管理多模态数据源的用户权限，不再需要维护多种体系的复杂权限配置，提升效率并减少潜在数据泄露风险。

## 全托管服务

TC-Catalog 元数据统一管理服务开箱即用，无需对元数据服务进行运维，支持数据开发与治理的全周期。

# 应用场景

最近更新时间：2025-01-21 17:37:02

## Lakehouse 平台搭建

Lakehouse 架构有效整合数据湖和数据仓库的优点，提供了廉价、灵活、多样化的数据存储底座，用户在 Lakehouse 之上可基于同一份湖存储数据构建多计算引擎应用，如基于 Spark 的批处理、基于 StarRocks 的高性能实时数仓、基于 SparkML 的机器学习应用等，TC-Catalog 通过多数据源统一元数据管理，可帮助用户快速搭建 Lakehouse 架构。

## 大数据与机器学习一体化

通过 TC-Catalog 的非结构化数据管理和 ML 模型的生命周期管理，可快速搭建大数据+机器学习一体化应用平台，如将 Spark 清洗好的训练语料数据注册到 TC-Catalog 的 Volume 中，ML 训练框架可通过 TC-Catalog 的元数据信息直接读取训练数据，完成训练后模型可注册到 TC-Catalog 进行统一生命周期管理。

## RAG 和 AI Search

在 RAG 和 AI Search 应用中，TC-Catalog 可有效整合结构化与非结构化数据，确保 GenAI 模型和 Agent 可通过 TC-Catalog 获取企业实时数据服务和准确的知识支持，从而产生更准确、符合业务和用户意图的 GenAI 内容。