DaoCloud

大模型时代的数据管理

主讲人：李晨露

为 数 字 世 界 寻 找 全 局 最 优 解

Powered By DaoCloud

Content

目录

Part 01

数据集基础

数据集 ？

01

数据集（Dataset）

一组样本的集合

02

样本（Example）

数据集的一行。
一个样本包含一个或多个特征，此外还可能包含一个标签。

03

标准数据集（StandardDataset）

符合一定规范的数据集

为什么要创建数据集？

- **计算机的视角**：计算机不擅长理解语言本身

规范化的输入便于学习到规律。

- **数据的视角**：现实中数据是散落的、不均衡的

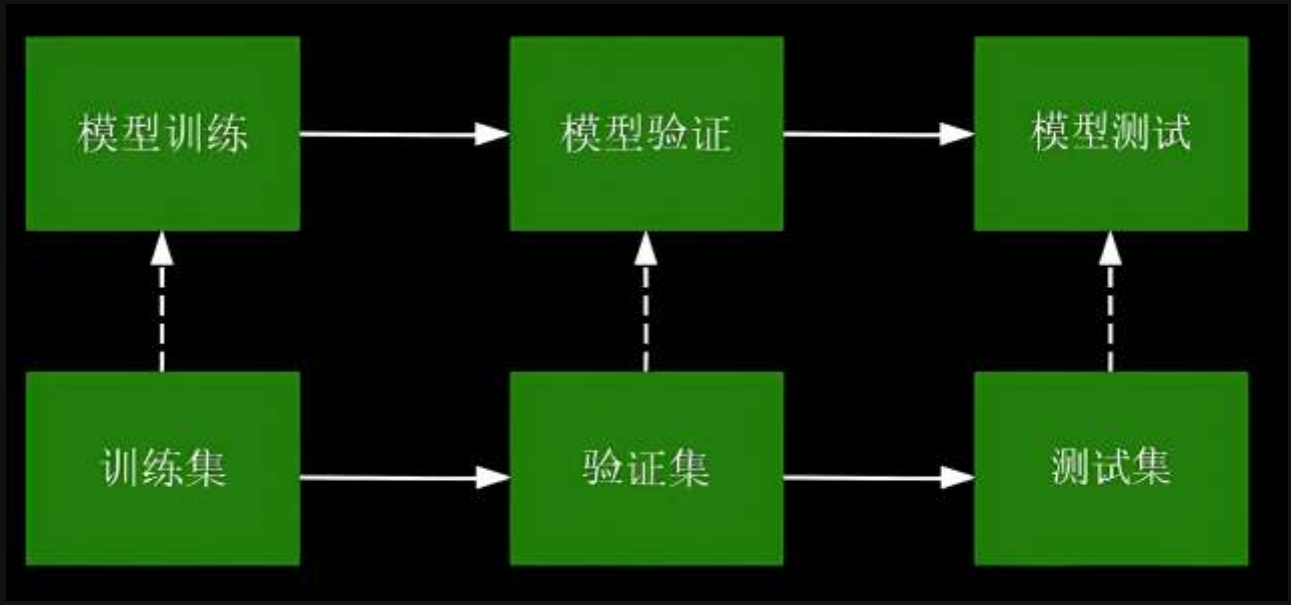
数据集中,更高效地处理。

- **人的视角**：人去量化机器学习的效果很难

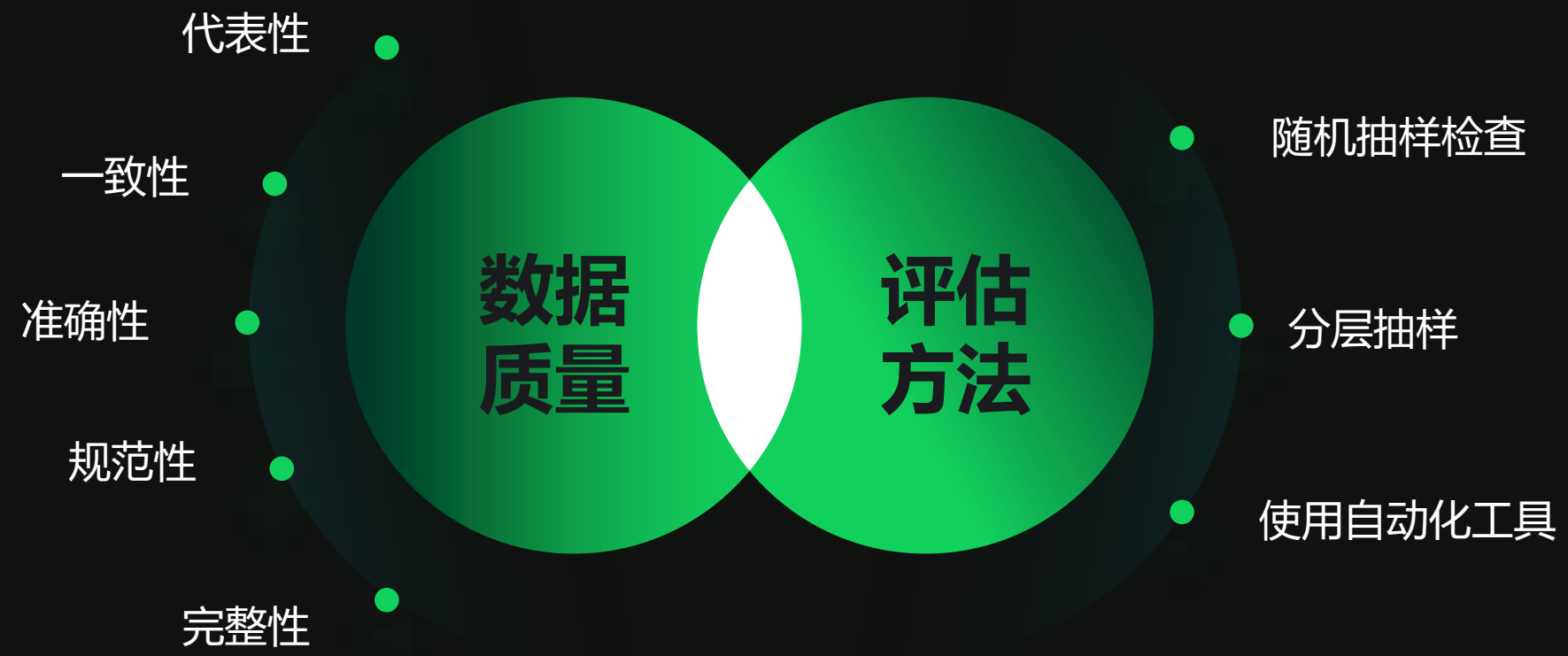
数据集便于在一定范围内评估机器学习的效果。

数据集的划分

- **训练集**：用于训练模型。
- **验证集**：用于调优模型参数和选择最佳模型。
- **测试集**：用于评估最终模型的性能。



高质量的数据集



Part 02

大模型时代数据管理的挑战

不同时代的数据差别

数据集差别

01

数据规模

- 大模型通常依赖于海量数据集
- 传统机器学习模型通常在较小规模的数据集

02

数据处理方式

- 大模型通过深度学习技术，能够自动从原始数据中提取高级特征，减少了人工特征工程的需求。
- 传统机器学习方法依赖于手动特征工程，开发者需要根据领域知识选择和设计特征

03

模型复杂性

- 大模型具有更高的计算复杂性，通常需要更多的计算资源和时间来进行训练。
- 传统机器学习算法结构相对简单，算法复杂度较低

数据集管理的挑战

- **管理复杂性**

  随着数据量增加，如何有效管理和监控存储系统成为一项挑战。

- **存储**

  传统存储架构可能无法满足分布式训练的需求，需要新的设计来优化数据访问效率

- **数据质量问题**

  大规模数据集中的噪声和不一致性可能影响模型性能，需要高效的数据清洗和融合技术。

- **隐私与合规性**

  数据隐私法规（如 GDPR）对大模型的数据使用提出了更高要求，企业需确保合规。

- **实时性需求**

  在动态环境中，如何快速更新和管理数据以支持实时决策是一个重要问题。

Part 03

统一高效的数据集管理

高效数据管理的过程

01

数据整合

02

数据质量

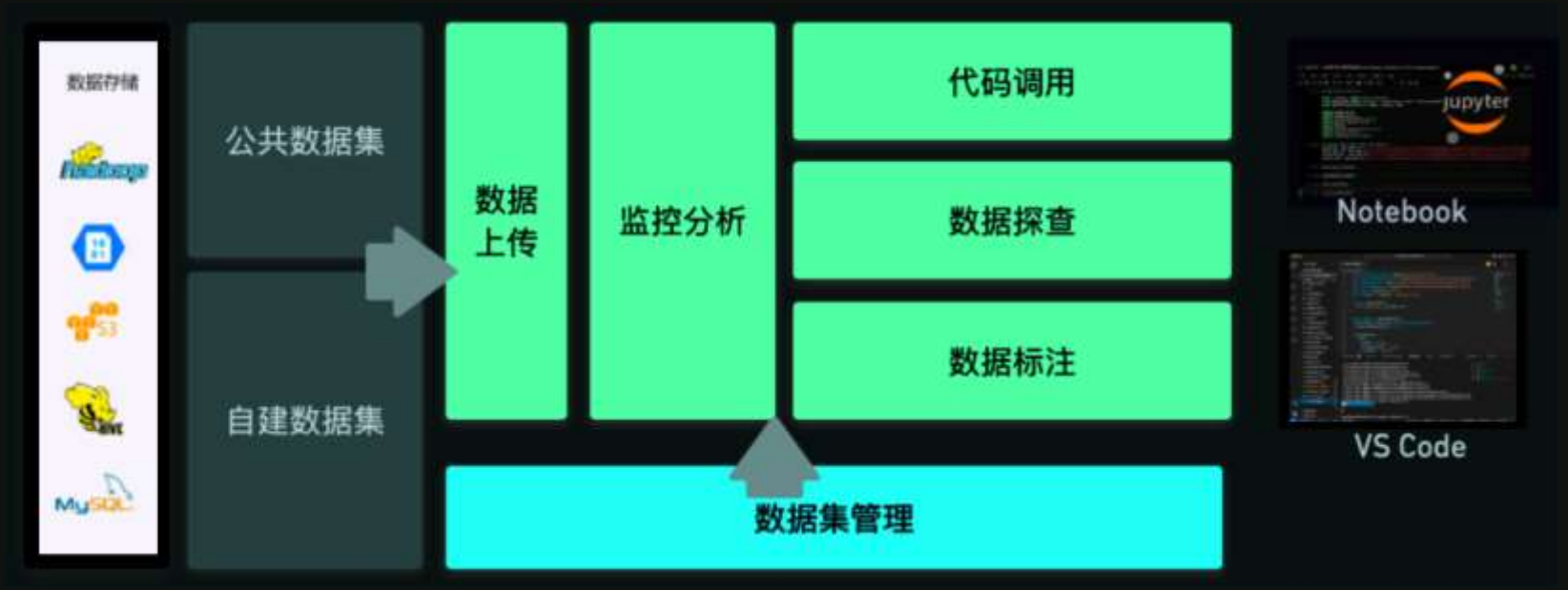
03

数据权限

数据整合

支持多种数据源的接入和整合，可以方便地将不同格式、不同来源的数据集成到一个统一的平台中，实现数据的全面分析和综合利用。

- 支持多种数据格式和文件类型
- 支持多种创建方式
- 支持多种数据源：包括：Git，NFS，S3等。
- 数据共享和协作



数据整合 - 存储需求

数据平台

数据平台

数据平台主要负责数据的接入、存储和处理，确保原始数据的有效管理和后续使用。

01

训练平台

训练平台

训练平台专注于模型的开发和训练，要求高性能计算能力以支持复杂的算法，同时需要快速访问训练数据和模型参数。

02

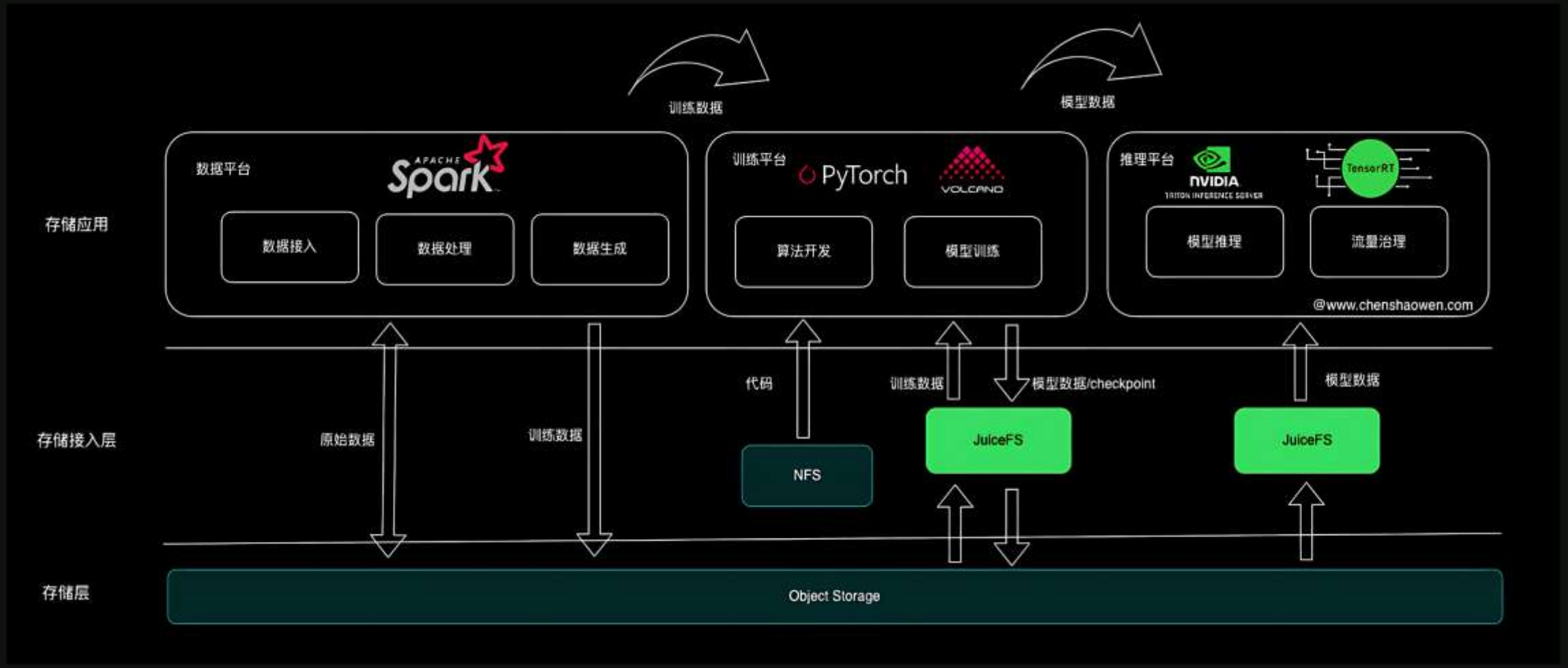
推理平台

推理平台

推理平台负责将训练好的模型部署到生产环境中，推理任务通常需要快速响应用户请求，因此对存储访问速度和系统响应时间有严格要求。

03

数据整合 - 存储


训练数据
模型数据
数据平台
APACHE Spark
数据接入
数据处理
数据生成
训练平台
PyTorch
VOLCANO
算法开发
模型训练
推理平台
NVIDIA
TRITON INFERENCE SERVER
TensorRT
模型推理
流量治理
@www.chenshaowen.com
存储应用
存储接入层
存储层
原始数据
训练数据
代码
训练数据
模型数据/checkpoint
模型数据
NFS
JuiceFS
JuiceFS
Object Storage

智能数据标注 – 数据质量

质量过滤

使用小规模高质量数据训练小体量模型也可以达到较好的效果，但是过于激进的过滤可能会导致性能下降。

去重

数据去重对提高模型性能和训练效率是有利的,常用的方法有神经网络方法以及特征空间距离比较。

无用信息过滤

过滤数据集中的有害信息可以减少模型生成有害信息的可能,但是可能对模型泛化和有害信息甄别能力有所损害,并且可能会导致模型对少数群里的偏见。

多样性

确保数据集的多样性和覆盖范围也是保证其质量的重要因素。通过各种手段（如K-Center-Greedy算法），在保证多样性的同时减少冗余数据，从而构建一个更加全面和代表性的数据集

数据质量评估

采用评分模型（如奖励模型）对数据进行打分，只有当评分超过设定阈值时，才认为该数据质量达标。这种方法能够有效筛选出高质量的数据集

数据质量 - 标注发展

人力标注 (2017年前)	平台/工具标注 (2017-2022年)	重新洗牌 (2023-2025年)	知识密集 (2025年后)
以训练任务、算法模型为导向; 简单图像标注为主。	以自动驾驶为代表的场景爆发; 人力标注无法满足数据需求,自动化标注兴起; 大量 AI 数据初创公司开始涌现。	数据质量驱动;产业链重新洗牌,更多企业参与数据标注,供应合作关系紧密; 创业门槛提高。	人机协同关系进一步耦合,人力更多承担关键决策角色; 市场竞争格局趋于稳定。

数据质量 - 标注方法

人工标注

优点

- 高质量
- 灵活性
- 情感和文化理解

缺点

- 效率低下
- 成本高昂
- 主观性强

VS

自动标注

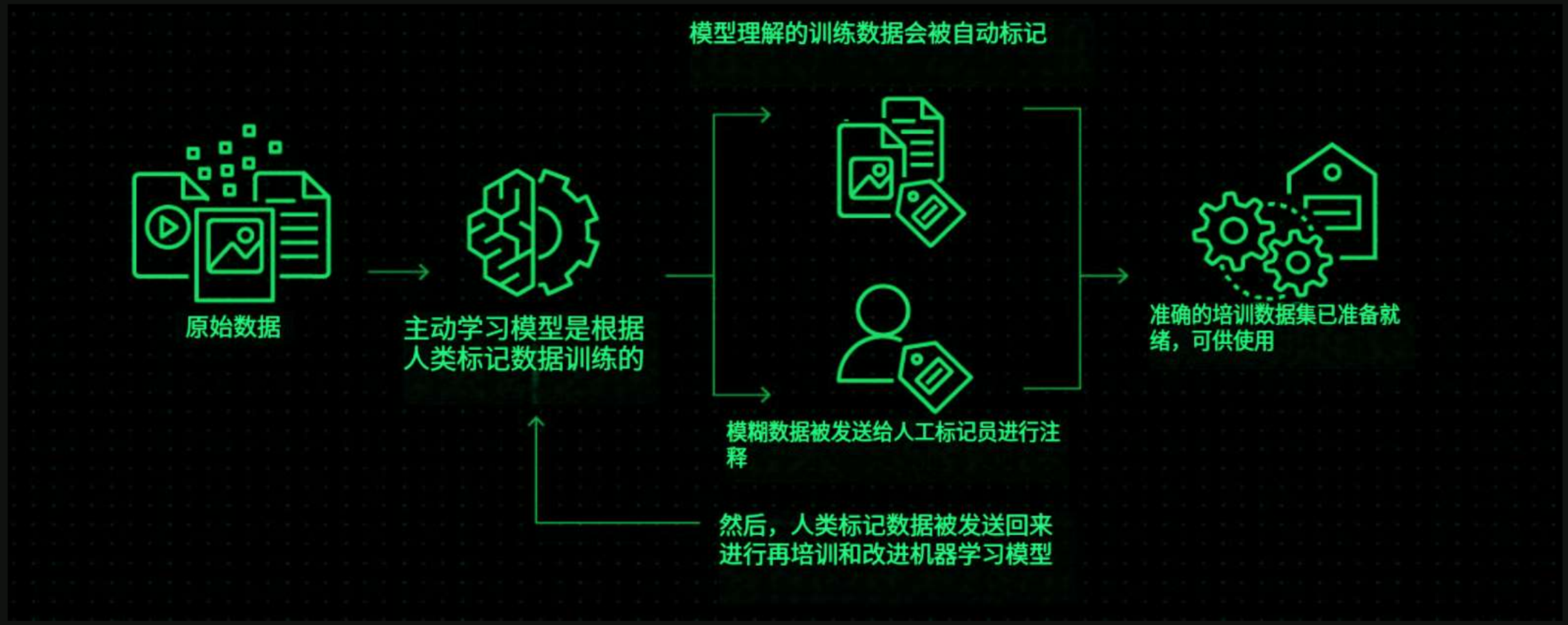
优点

- 高效率
- 一致性高
- 适应性强

缺点

- 质量不稳定
- 缺乏上下文理解
- 依赖训练数据

智能数据标注



公共数据集

大型公共数据集

- Huggingface Datasets: 用于加载、处理和共享各种数据集，支持音频、计算机视觉和自然语言处理等领域。访问链接
- AWS Public Datasets: 亚马逊云服务提供的公共数据集，适合大规模分析和机器学习应用。访问链接
- Google Cloud Public Datasets: Google云平台提供的公共数据集，包括多种类型的数据供分析使用。访问链接

数据共享平台

- data.world: 一个社交网络平台，用户可以搜索、分析和共享数据集。访问链接
- Quandl: 提供金融、经济和替代数据集的平台，部分数据免费，部分需要付费。访问链接

完善的权限管理

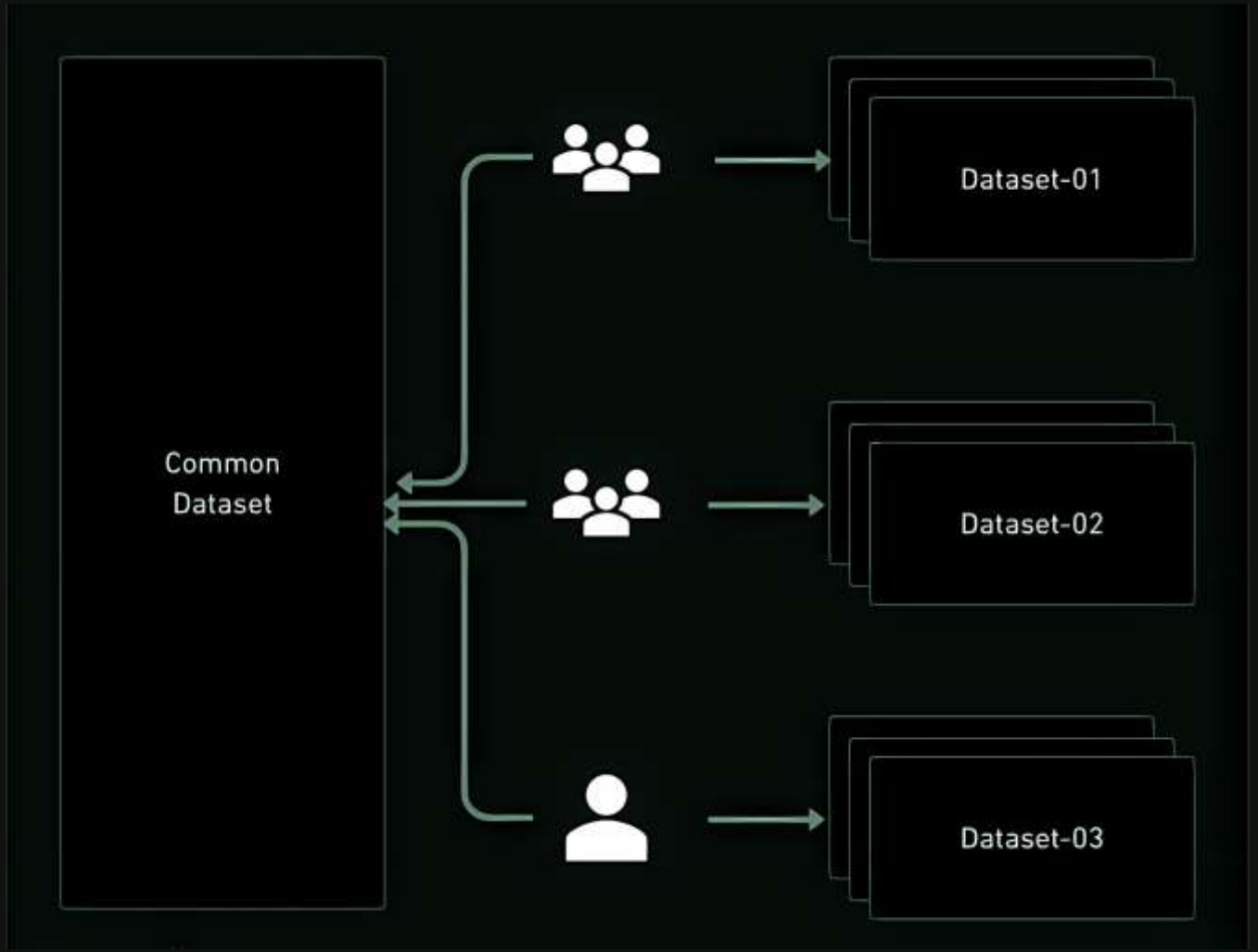
支持灵活的权限控制，可以按照团队或个人的需求，对不同的数据进行细粒度的权限设置，确保数据的安全性和隐私性。

资源隔离特性

- 根据用户、团队等来控制数据访问,确保了数据集在计算任务与数据访问层面的安全控制

公开数据集

- 在多个团队间复用，实现了单次缓存、多团队共享的高效模式



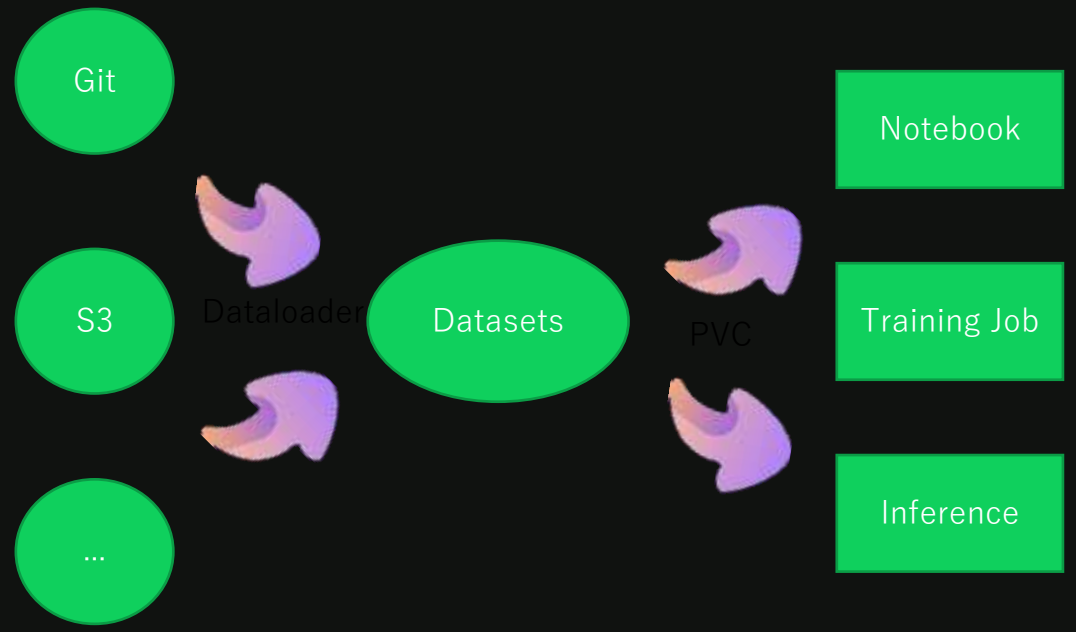
Part 04

现场 DEMO

数据集 - 一切故事开始的地方

在 Baize 中，我们提供多种数据源接入能力，包含了 Git、S3、HTTP、PVC、NFS 等。

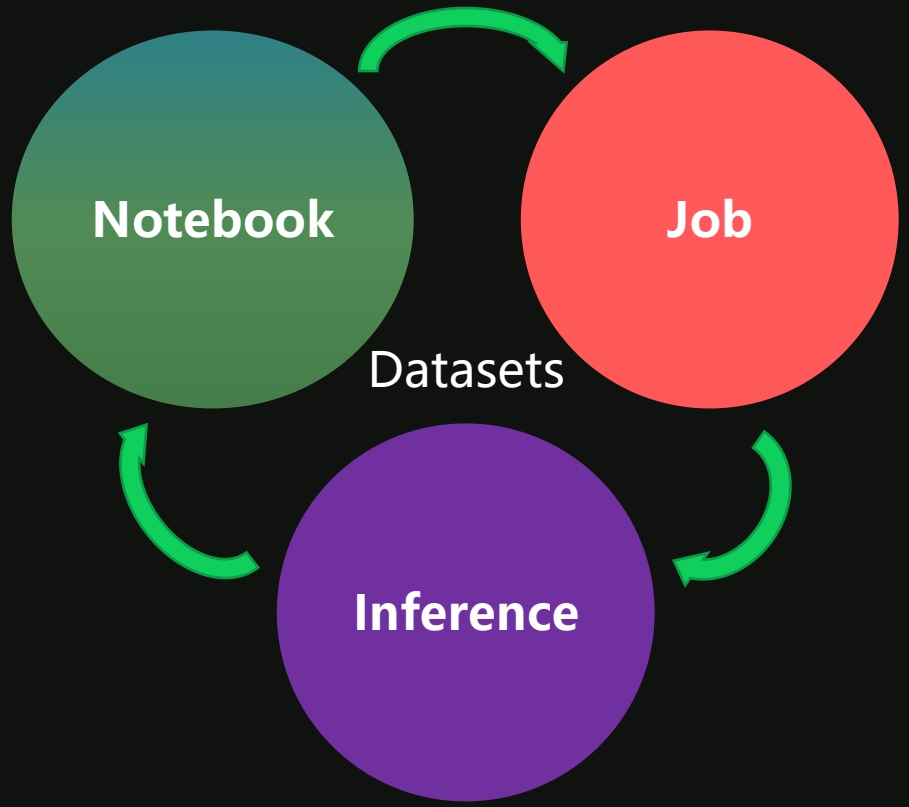
支持**实时**数据**预热**、数据源凭**证**管理、以及数据集一**键**克隆。



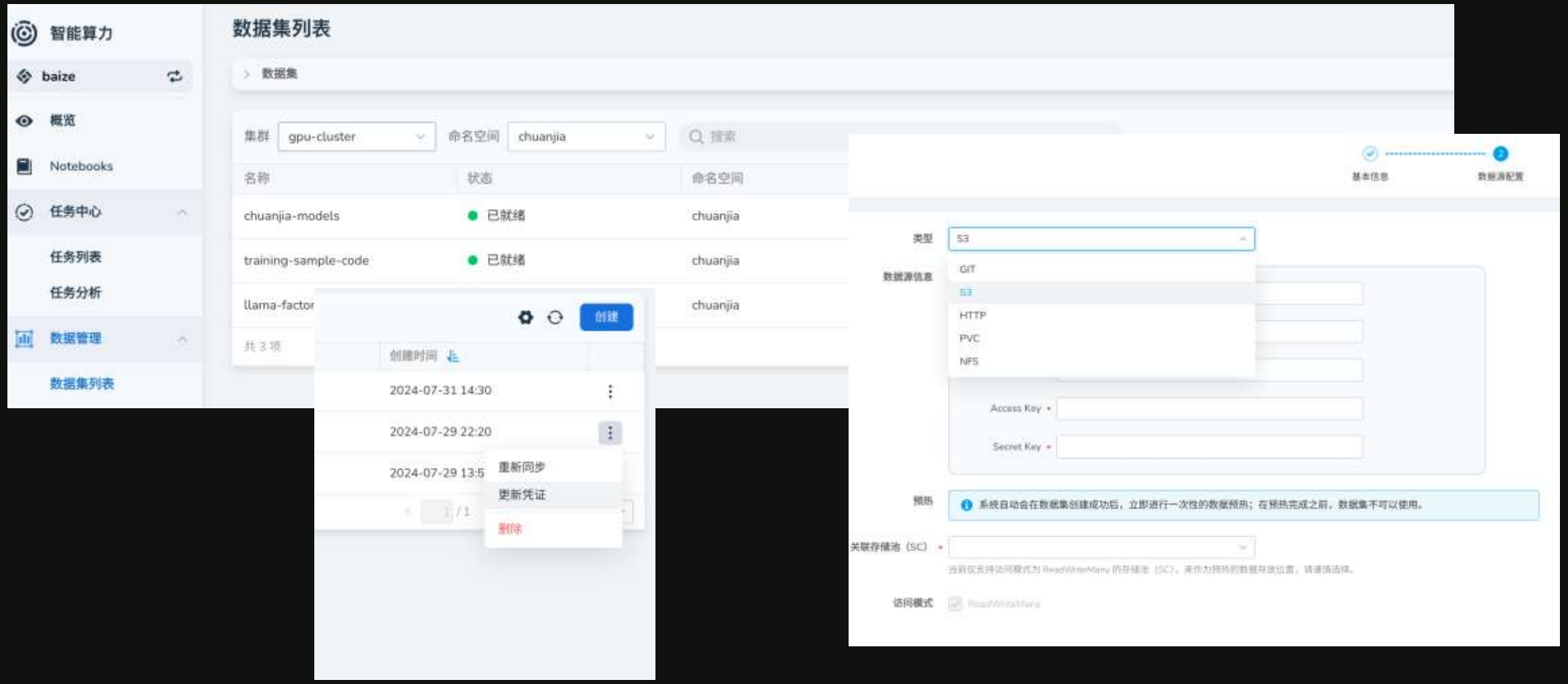
数据集常见操作

数据集 - 一切故事开始的地方

在 Notebook、训练任务、推理服务都可以使用 数据集。



数据集 - 界面一览


智能算力
baize
概览
Notebooks
任务中心
任务列表
任务分析
数据管理
数据集列表
数据集列表
数据集
集群
gpu-cluster
命名空间
chuanjia
搜索
名称
状态
命名空间
chuanjia-models
已就绪
chuanjia
training-sample-code
已就绪
chuanjia
llama-factor
chuanjia
共 3 项
创建
创建时间
2024-07-31 14:30
2024-07-29 22:20
2024-07-29 13:5
重新同步
更新凭证
删除
基本信息
数据源配置
类型
S3
数据源信息
GIT
S3
HTTP
PVC
NFS
Access Key
Secret Key
预热
系统自动会在数据集创建成功后，立即进行一次性的数据预热；在预热完成之前，数据集不可以使用。
关联存储池（SC）
访问模式

Thanks.

扫码加入课后群讨论 & 获取课件