Technical Disclosure Commons Technical
Disclosure Commons

Defensive Publications Series

2023年2月

# 机器学习数据集管理平台机器学习数据集管理平台

Ze Mao

Yang Xu

Erick Suarez

请在此处阅读本作品及其他作品：https://www.tdcommons.org/dpubs_series

推荐引用 推荐引用Mao, Ze；Xu, Yang；和 Suarez, Erick，《机器学习数据集管理平台》，
Technical Disclosure Commons，（2023年2月21日）
https://www.tdcommons.org/dpubs_series/5690



本作品采用知识共享署名 4.0 许可协议许可。本文章由 Technical Disclosure Commons 提供，供公众免费开放获取。经 Technical Disclosure Commons 授权管理员批准，本文已被纳入 Defensive Publications Series。

# 机器学习数据集管理平台


## 摘要

数据集中的数据质量会显著影响使用该数据集进行训练和/或评估的机器学习模型的性能。有效的数据集管理，包括数据清理、版本管理、访问控制、数据集转换、自动化、完整性和安全性等任务，有助于提升机器学习流程的效率和速度。目前，工程师需要投入大量人工精力和时间来管理数据集版本，或为机器学习任务准备数据集。本文介绍了一个用于有效管理和使用数据集的平台。这些技术将数据集管理与数据集转换机制相结合。文中介绍了一种存储引擎，作为所有数据的事实来源，并负责处理版本管理、访问控制等。数据集转换机制是生成数据集（快照）以满足不同用途的关键部分。所述技术可支持不同的工作流、流水线或数据编排需求，例如机器学习模型的训练和/或评估。



## 关键词

- 数据集管理
- 数据集转换
- 训练数据
- 数据版本管理
- 数据流水线
- 数据编排
- 机器学习

## 背景

在机器学习中，数据集是用于训练和评估模型的数据集合。数据集中的数据质量会对机器学习模型的性能产生重大影响。有效的数据集管理至关重要，其中包括数据清理、版本管理、访问控制、数据集转换、自动化、完整性和安全性等任务。这也有助于机器学习研究人员和工程师提升机器学习流程的效率和速度。然而，目前缺乏管理数据集（例如管理数据集版本、设置数据集访问控制等）或转换数据集的优质工具和流程。工程师需要投入大量人工和时间来管理数据集版本，或为机器学习任务准备数据集。

虽然存在一些相关产品，但它们都存在不足。DVC [1] 是一种数据版本控制系统，但它专注于模型开发和部署，而非数据集管理或数据转换。Deeplake [2, 3] 支持管理数据集，但不支持大规模数据转换。TFDS [4] 不具备版本管理和访问控制功能。git [5] 是一种广为人知的版本控制解决方案，适用于代码或文本，但不适用于大型对象。因此，git 不适用于机器学习数据集，因为这类数据集通常非常庞大，而且往往包含非文本数据。

## 说明

本公开描述了一种用于有效管理和使用数据集的平台。这些技术将数据集管理（版本、访问控制等）与数据集转换机制相结合。本文还描述了一种存储引擎，它充当所有数据的事实来源，并处理版本管理、访问控制等。数据集转换机制是一个关键

部分，用于生成数据集（快照），以满足不同目的。所述技术可支持不同的工作流、流水线或数据编排需求。

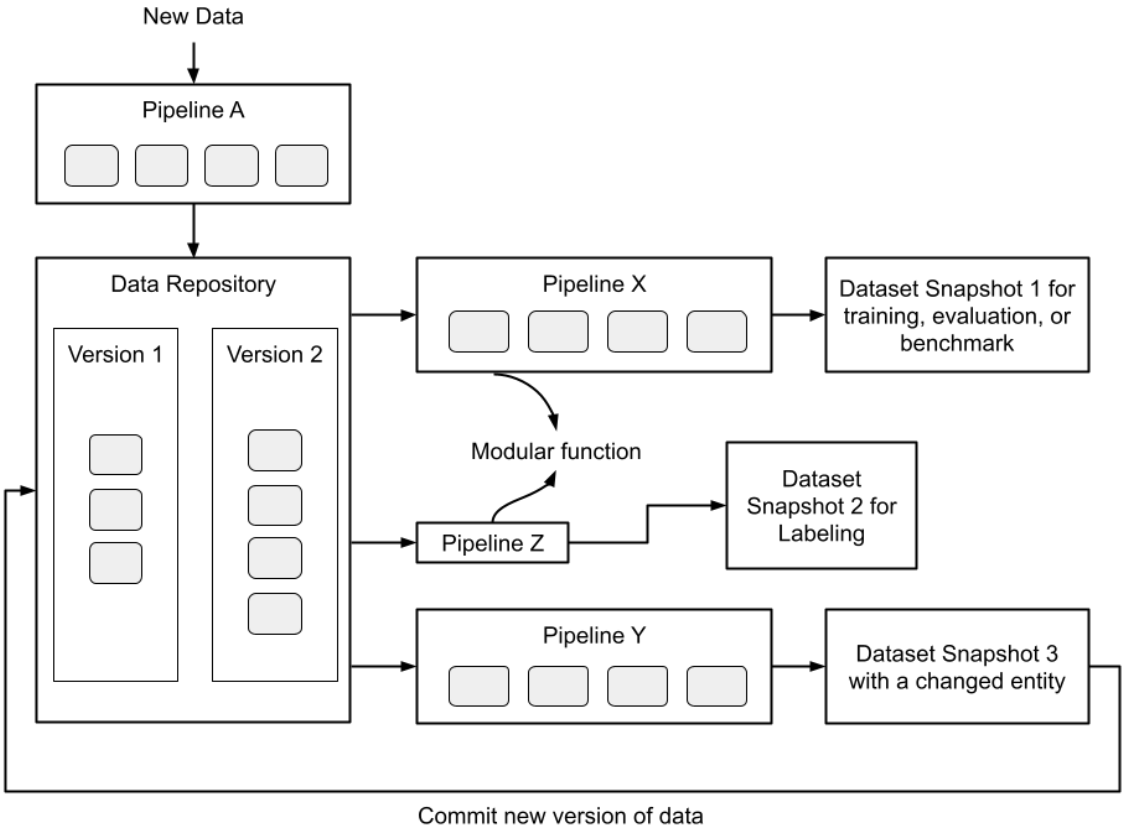



图1：数据存储库和数据管道

如图1所示，管道A、管道X、管道Y和管道Z是用于数据集转换的工作流示例。新数据通过管道A导入，并添加到数据存储库中。管道X、管道Y和管道Z分别生成用于不同用途的数据集（快照）。数据工程师或其他用户可以根据具体需求创建不同的管道。例如，管道X生成的快照1可用于训练、评估或基准测试模型，而管道Z生成的快照2可用于标注。管道Y创建的快照3包含至少一个已更改的实体。如图1所示，快照3中的数据新版本会提交到数据存储库，以供将来使用。这些管道可以手动触发，也可以

根据特定事件自动触发。这些管道易于维护，可用于生产环境，并支持数据血缘跟踪。

流水线中的各个模块（如图1中的小灰色方块所示）可共享、可复用，也可串联。流水线的运行方式类似于大数据应用中常见的抽取-转换-加载（ETL）流水线，但更适用于机器学习用例。流水线易于轻量化实现（例如，只需几行 Python 代码即可实现），支持快速迭代，而且易于运行。

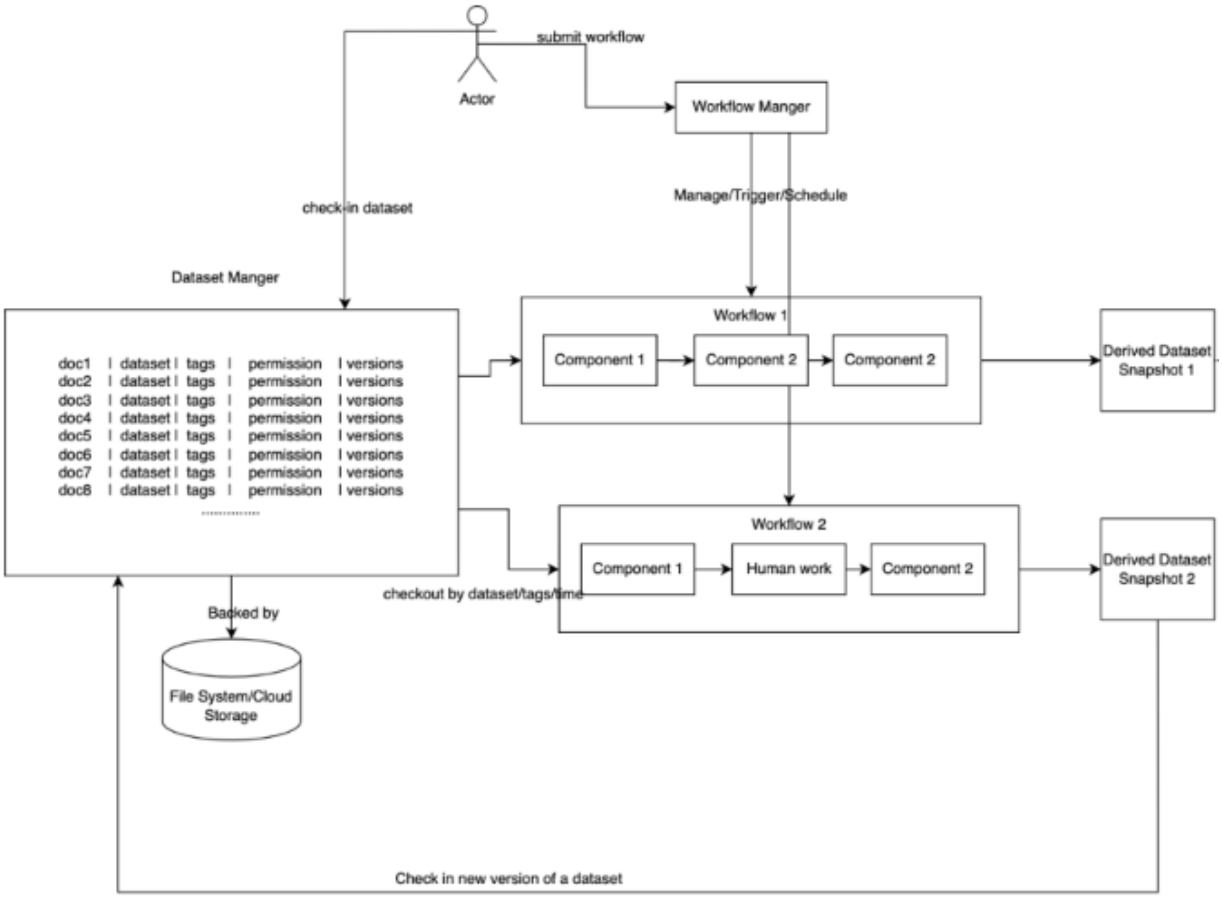



图2：数据管理平台架构

图 2 展示了根据本公开的技术实现的数据管理平台的一个示例架构。参与者（例如用户，或充当触发器的自动化程序）可以

将数据集登记到平台中，并向平台提交工作流。该平台包含两个核心模块：数据集管理器和工作流管理器。

● 数据集管理器：数据集管理器用于存储数据集、管理版本、控制访问权限以及检出数据集，并提供数据集签入/检出功能。用户可以通过命令行界面（CLI）或其他用户界面（例如图形用户界面）签入数据。数据或数据集可以添加一个或多个标签。数据集管理器会在数据签入/检出时强制执行访问控制和权限管理。它还提供查询功能，例如按标签、数据集名称或其他属性查询数据集。用户或工作流可以通过指定查询条件来检出数据。存储的数据类型不受限制，用户可以签入任何格式的数据。数据的底层存储可以采用任何合适的机制，例如文件系统或云存储。

● 工作流管理器：使用数据管理平台的用户可以向工作流管理器注册其工作流。工作流管理器负责分配资源、安排运行并报告结果。在此上下文中，工作流是用户定义的一系列组件（流水线），用于转换数据集。工作流的输入是一组满足特定条件的已检出数据。工作流的输出是一组适用于模型训练或评估的数据，或可作为新数据集签入数据集管理器的数据。执行工作流后生成的数据集快照可用于训练和/或评估机器学习模型。

工作流中的每个组件都有特定功能。组件分为两种类型：基于程序的数据处理单元和基于人工工作的数据处理单元。工作流管理器会为工作流中的计算组件分配计算资源，以支持

大规模数据处理。同时还会追踪数据血缘。

平台的主要功能包括：

- 数据集读取/写入/删除
- 数据集版本管理
  - 版本控制和版本差异
- 数据集访问控制和安全性
  - 访问控制/权限
- 数据集转换
  - 转换原始数据，生成数据集的衍生版本。
- 工作流自动化
  - 将多项转换和人工操作编排为一个工作流。
  - 通过事件触发工作流（新数据集版本或其他事件）
  - 按时间安排工作流
- 数据沿袭
  - 按版本、衍生和工作流追踪数据沿袭

The 所述数据管理平台可用于多种场景，例如：

- 对新接收的数据执行清理操作
- 执行数据转换，以满足模型训练/评估需求
- 基于数据，为编程工作和人工工作提供自动化支持与连接

- 支持大规模数据处理
- 数据撤销

## 结论

本公开描述了一种用于有效管理和使用数据集的平台。这些技术整合了数据集管理（版本、访问控制等）和数据集转换机制。文中描述了一种存储引擎，它作为所有数据的事实来源，并负责处理版本管理、访问控制等。数据集转换机制是生成数据集（快照）以满足不同用途的关键部分。所述技术可支持不同的工作流、流水线或数据编排需求，例如机器学习模型的训练和/或评估。

## 参考文献

1. “数据版本控制”可在线访问：https://dvc.org/2. “GitHub - activeloopai/deeplake：面向深度学习的数据湖。构建、管理、查询、版本控制和可视化数据集。实时向 PyTorch/TensorFlow 传输数据。”可在线访问：https://github.com/activeloopai/deeplake3. “Activeloop”可在线访问：https://www.activeloop.ai/4. “模块：tfds | TensorFlow 数据集”可在线访问：https://www.tensorflow.org/datasets/api_docs/python/tfds5. “Git --fast-version-control”可在线访问：https://git-scm.com/6. “文档人工智能 | Google Cloud”可在线访问：https://cloud.google.com/document-ai