



GOTC 2023

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

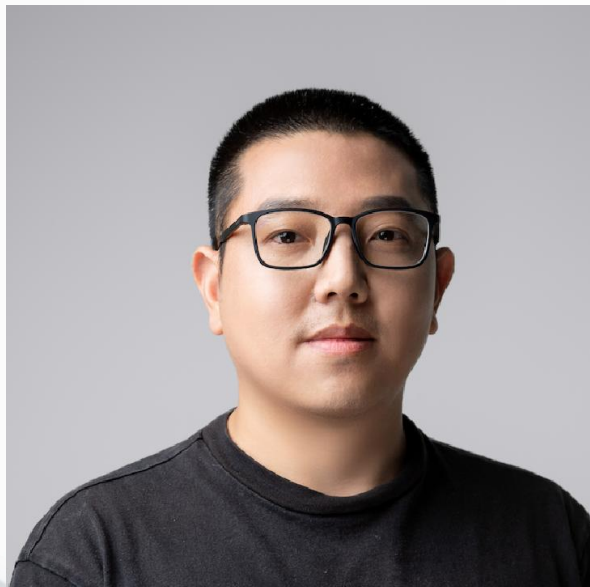
OPEN SOURCE, INTO THE FUTURE

「AI is Everywhere」专场

Xtreme1 下一代多模态开源训练数据平台



王家军
研发总监
@倍赛科技



王家军 Jagger Wang

个人技术博客: <https://blog.jaggerwang.net>

毕业学校和学历 | 华中科技大学本硕

过往经历 | 深圳腾讯、上海盛大、成都数联铭品

目前就职公司和职务 | 倍赛科技 研发总监

目前主要研究方向 | 微服务, 云原生, 大数据, DevOps

Xtreme1 是全球首个开源多模态训练数据平台，通过提供 AI 赋能的软件工具、数万项目提炼的
本体中心和丰富的数据治理特性，来加速多模态训练数据的处理效率，进而提高 AI 工程师的建模
效率，特别是在 2D & 3D 多模态融合数据方面，标注效率的提升可达 72%。Xtreme1 于 2022
年 9 月开源，同年 12 月成为 Linux Foundation AI & DATA 下的托管项目。



支持 10+ 种
多模态数据格式



2D & 3D 融合数据标注
提效 72%



用户训练数据效率
提高 87%

训练数据平台的起源

UBS Global 研究报告发现：

现如今 AI 工程师 70%-90% 的时间都花在训练数据的准备上

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE



扫码备注“进群”加入
Xtreme1 技术交流群

很多算法在实践中已非常优秀，**数据**已然成为 AI 模型开发的瓶颈



重复的数据与模型调优

AI 企业和工程师花费大量时间去定义重复或相似的问题



数据越来越复杂

硬件成本降低，AI 能力上升，造就更多更复杂的数据需求



数据质量的问题

数据的重心从数量转向了质量，并且质量的定义多元化



数据漂移

训练和生产数据的不一致性，不断催生新的数据标注需求



数据安全

大量企业或机构需要私有化部署以保障数据安全性和合规性



数据成本上升

随着人力成本、需求复杂性上升，数据成本也随之急剧上升



缺乏透明性

大量标注需求以整体外包形式完成，增加了沟通和质量控制成本



碎片化的工具

建模过程包含很多复杂的流程，每个环节都有大量工具可供使用

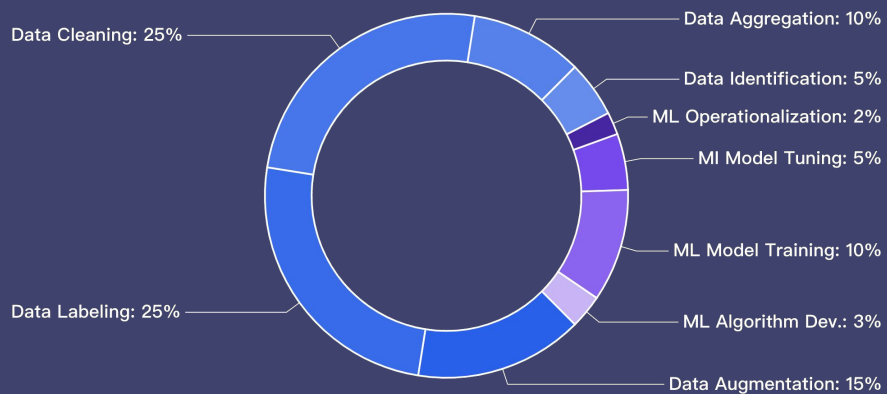
全新一代训练数据平台 Xtreme1



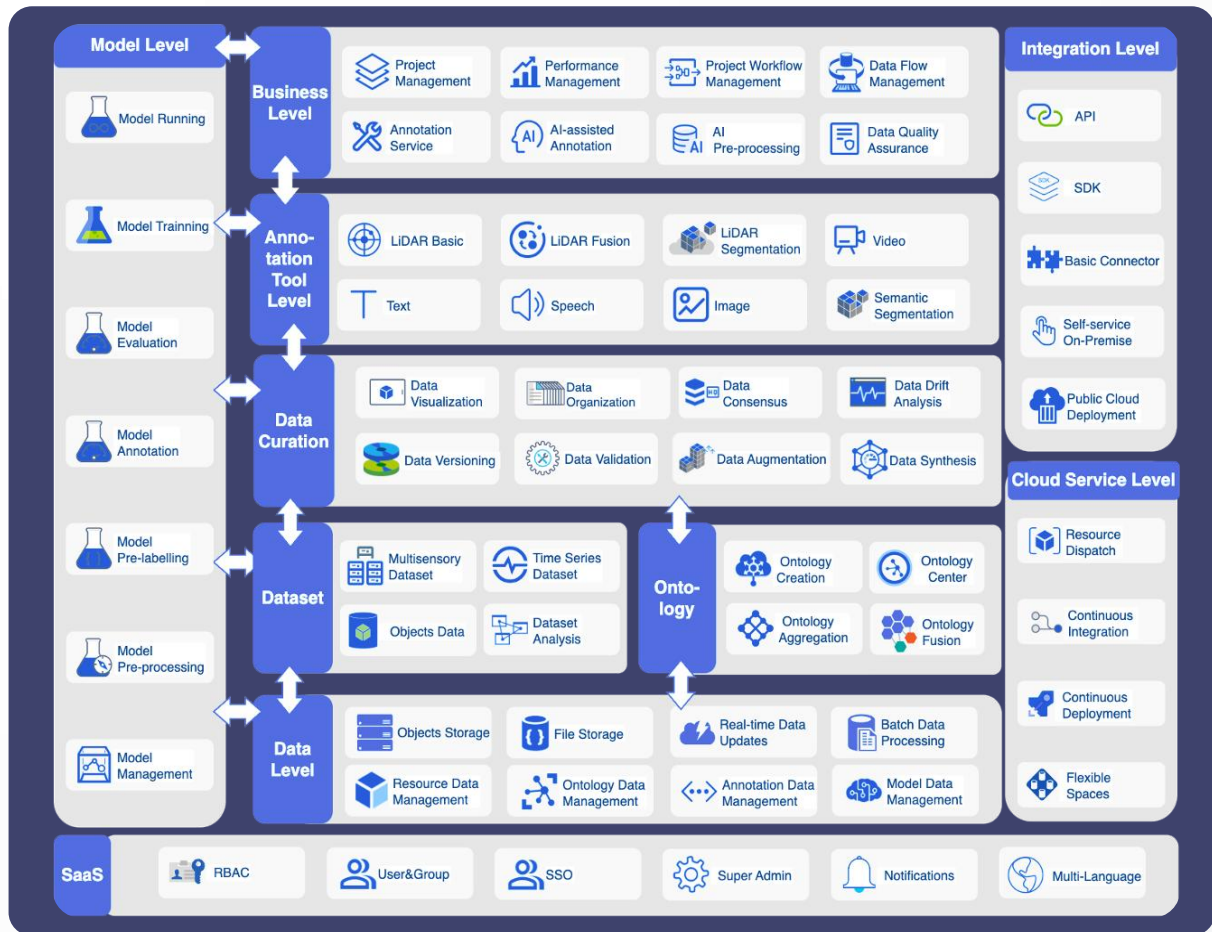
上一代 | 面向 AI 标注员标注提效

下一代 | 面向 AI 工程师建模提效

80% of time spent for Machine Learning Projects is allocated to Data related tasks



Source | Cognylitica; Factordaily



全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

平台核心功能介绍

使命：构建最容易触达的开源 Data-Centric MLOps 基础设施，以连接人、模型和数据

全球开源技术峰会

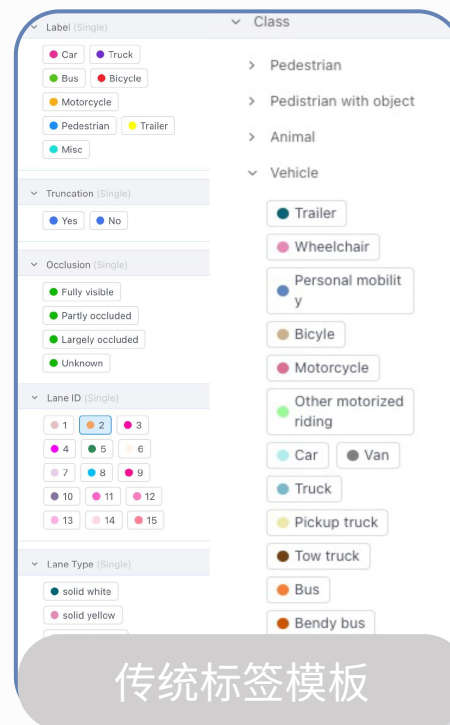
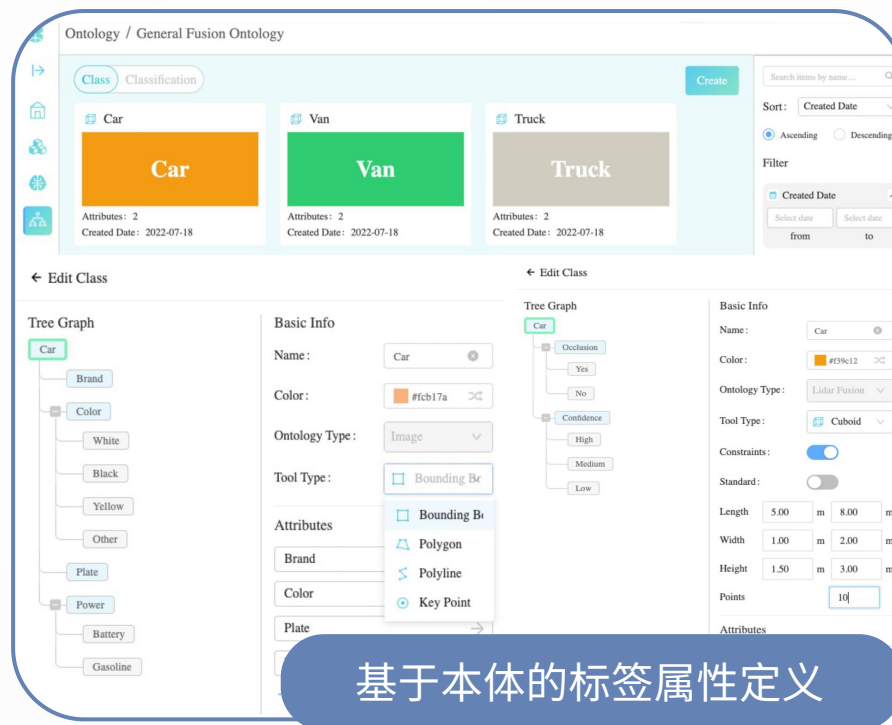
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE



扫码备注“进群”加入
Xtreme1 技术交流群

本体中心

Xtreme1 首次引入本体 Ontology 来穿透不同 AI 客户的问题抽象，以实现更易于共享、融合和推荐的子问题定义，从而加速模型的开发进程。



本体中心

Xtreme1 首次引入本体 Ontology 来穿透不同 AI 客户的问题抽象，以实现更易于共享、融合和推荐的子问题定义，从而加速模型的开发进程。

本体查询



基于本体搜索，实现本数据集或跨数据集的数据查询，可动态重构新的数据集

本体推荐



通过用户的数据或定义的部分本体，给用户推荐相似、相关的本体，以加速用户建模进程

本体关联



利用本体关联来解决单传感器多任务、多传感器融合任务的关联关系

本体融合



利用本体融合解决跨数据集同义不同名的数据管理，高效解决数据的融合

Xtreme1 标注套件

随着生产工艺的上升，AI 工业界迎来了更高分辨率、更多传感器、混合传感器、多任务的模型开发



LLM Toolset

适用于大语言模型 RLHF 训练的文本标注套件



LiDAR Basic & Fusion

基于摄像头和激光雷达的 2D & 3D 融合标注



LiDAR Segmentation

支持激光雷达的点云分割，2D & 3D 融合分割



Image

支持多种图像标注工具，矩形框、多边形、多线段、关键点、贝塞尔曲线



Task Management

灵活的任务管理、数据分发，丰富的权限控制



Performance Management

多维度的绩效考核，提升大规模生产的人力效率



Workflow Management

灵活的工作流配置，以适应各种不同的质量要求



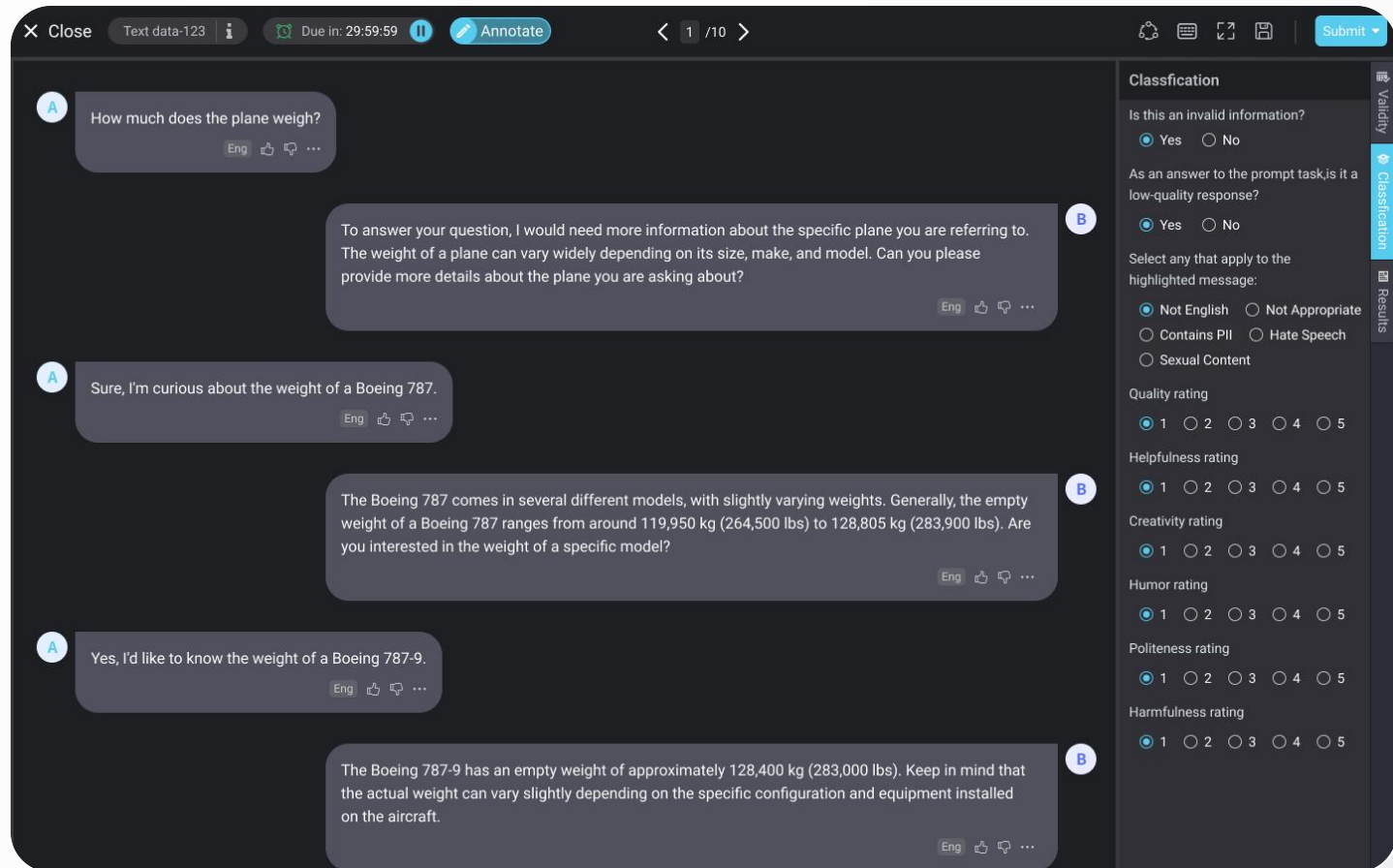
AI-assisted Annotation

引入多种 AI 能力来加速提升标注工具的效率

标注套件

LLM & RLHF 训练套件

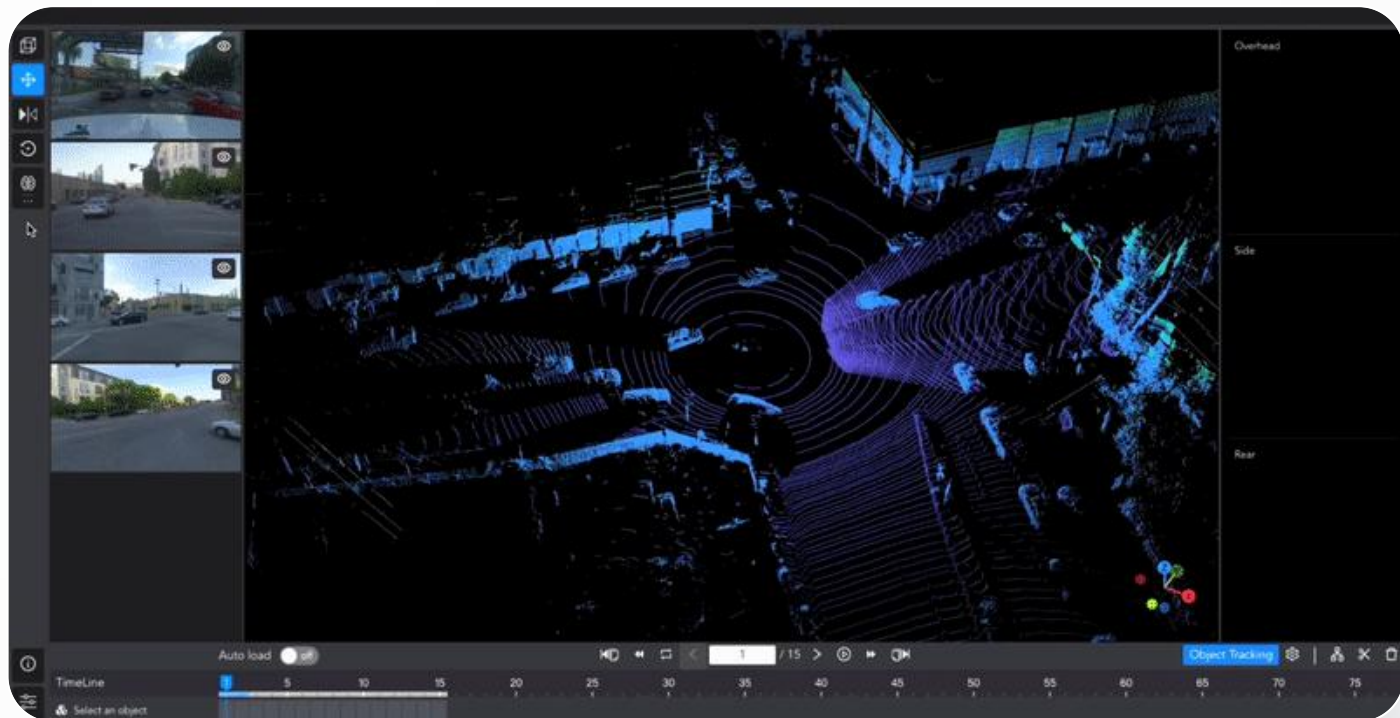
ChatGPT 的出现证明，人类反馈强化学习（RLHF）训练可以大幅提升机器学习模型生成内容的范围、有效性和准确度，Xtreme1 支持 RLHF 文本标注。



标注套件

2D & 3D 融合标注工具

全球首个支持点云模型接入的 2D & 3D 融合标注工具，预训练模型可以实现分钟级精准创建几百个 3D 框，点云数据中的标注可自动映射至对应的 2D 图中。



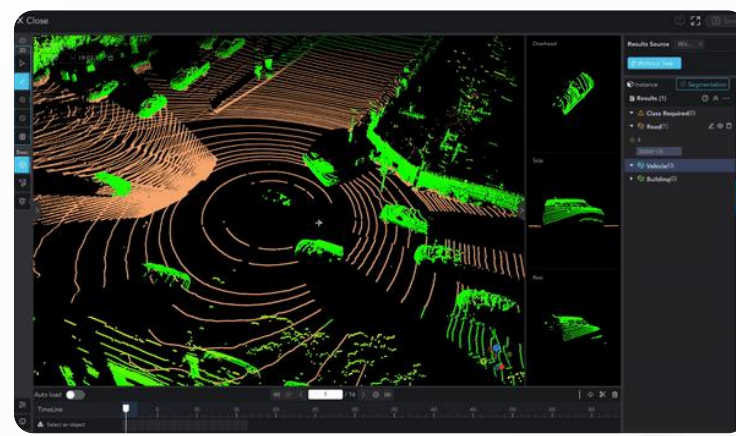
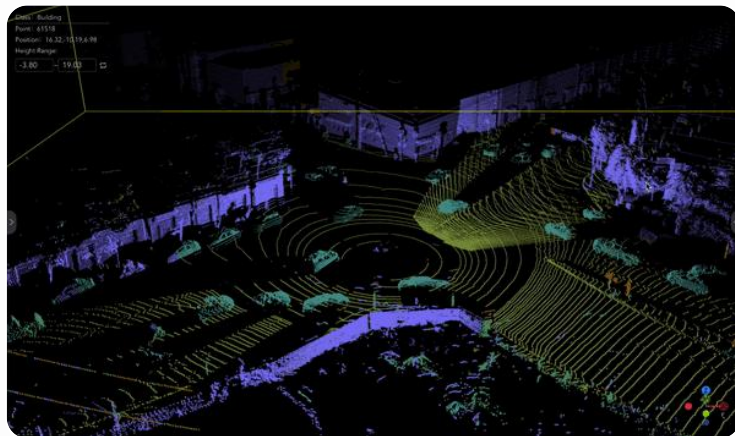
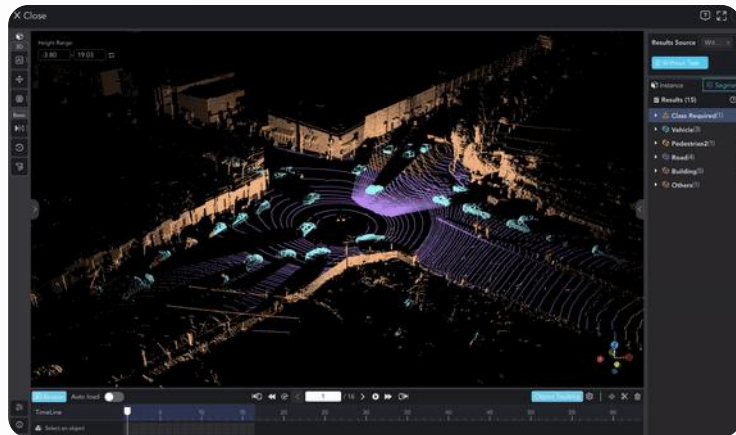
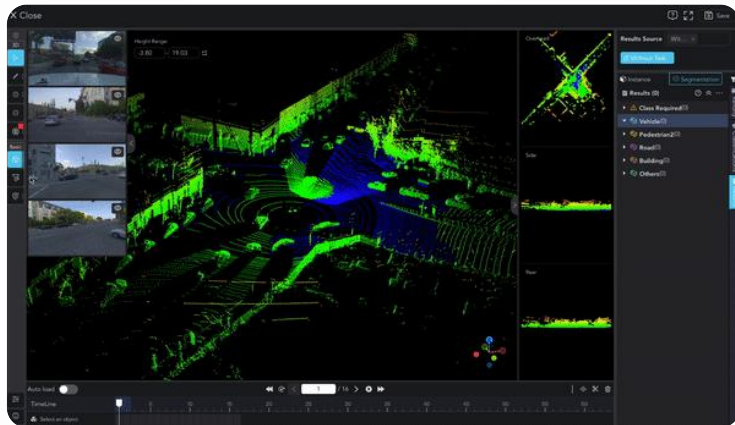
平台核心功能介绍



标注套件

LiDAR 点云分割*

点云分割套件支持单项目 1.5 亿点的分割标注，支持上千帧的连续帧标注，集成的自动分割模型提高了 AI 训练师的效率。



*3D LiDAR 点云数据分割功能目前仅在 Xtreme1 线上版支持

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

标注套件

模型调参

Xtreme1 支持灵活的模型推理参数调整，以适应项目对准确率和召回率的不同需求。



标注套件

灵活的人机交互

Xtreme1 提供更符合人机交互的方式，最大化半自动标注的效能。



全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

平台核心功能介绍

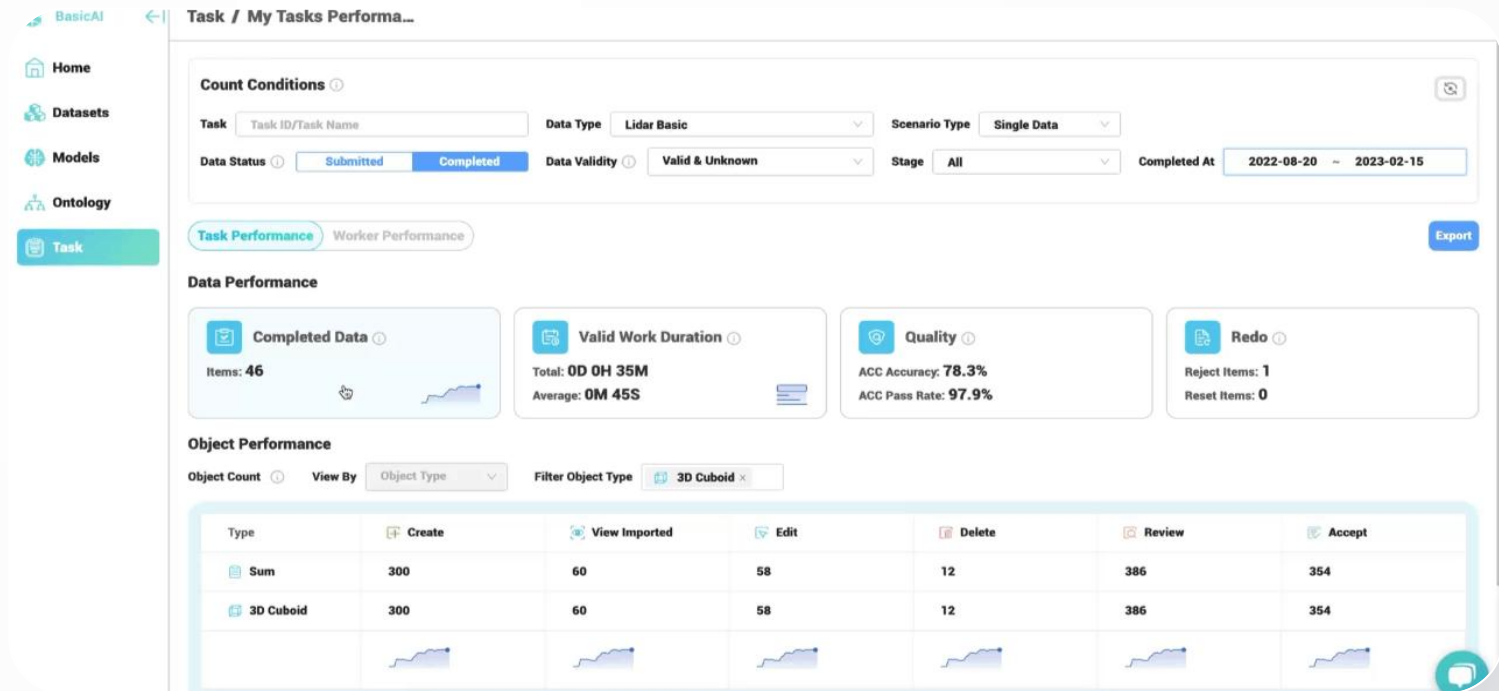
Xtreme1

GOTC

标注套件

workflow与绩效管理*

为不同规模团队、不同质量要求提供灵活的工作流配置。



*完整 workflow与绩效功能目前仅在 Xtreme1 线上版支持

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

Xtreme1 数据治理

随着 AI 对训练数据的质量要求不断上升，评价尺度也更加多样化，对数据平台提出了更高的要求

Data Visualization

多粒度、多尺度、2/3D 的数据可视化，实现高效的数据审阅，统计分布，多样化分析

Data Organization

提供多种数据组织的插件，聚类、排序、过滤、自动标签、批量操作

Data Consensus

人人对比、人机对比，验证数据一致性的问题

Data Drift Analysis

基于时间线跟踪对比不同时间、日期的数据集差异，提供可视化，统计对比，模型异常发现

Data Versioning

数据版本无处不在，不同模型、不同日期、不同批次的质检

Data Validation

提供多种策略对比模型效果，快速发现模型缺陷

Data Augmentation

提供多种主流的数据增强算法解决数据分布不均衡的问题，一站式数据增强和审查筛选

Data Synthesis

支持内置和外接合成模型解决数据的稀疏、低频、长尾问题

平台核心功能介绍

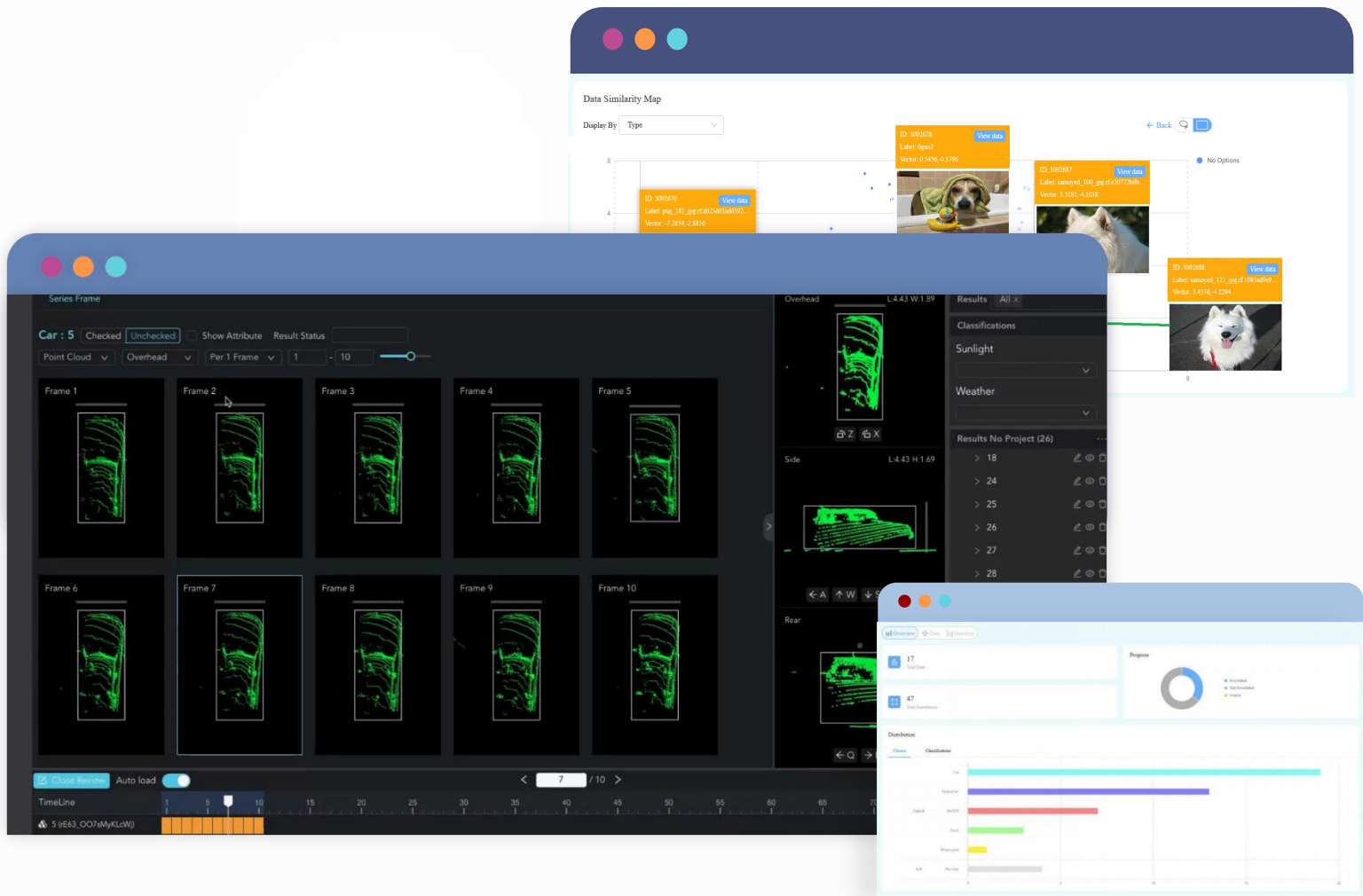
Xtreme1

GOTC

数据治理

多模态数据可视化*

Xtreme1 提供丰富的训练数据统计分析、可视化分析、多粒度的数据集搜索与管理



*连续帧标注可视化检查功能目前仅在 Xtreme1 线上版支持

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

SOTA 模型集成

下一代多模态训练数据平台大量使用 AI 能力，提供效率更高的工具和完整的解决方案

数据效能

快速标注 | AI 预标注

数据质量 | AI 数据过滤筛选

2/3D融合 | AI 传感器矫正

人力成本 | 主动学习

数据成本 | 数据增强与合成

数据效能

迁移学习 | 预训练模型准备

快速应用 | 开箱即用的特定场景
模型

快速建模 | 低代码支持，API，SDK

终身学习 | 数据管道与流数据接入

平台建设的技术挑战

Xtreme1 全面遵循云原生架构原则，以保障服务性能的可扩展性，部署规模的可弹性，以及在故障情况下的服务韧性

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE



扫码备注“进群”加入
Xtreme1 技术交流群

平台建设的技術挑战



结构化数据存储

接入层

只对外暴露少数服务，大部分内部服务对外隐藏，减少安全隐患

应用服务层

前端和后端按功能模块拆分服务，各服务可独立构建和发布

基础服务层

均采用开源软件，且都为分布式系统，支持水平扩容

容器抽象层

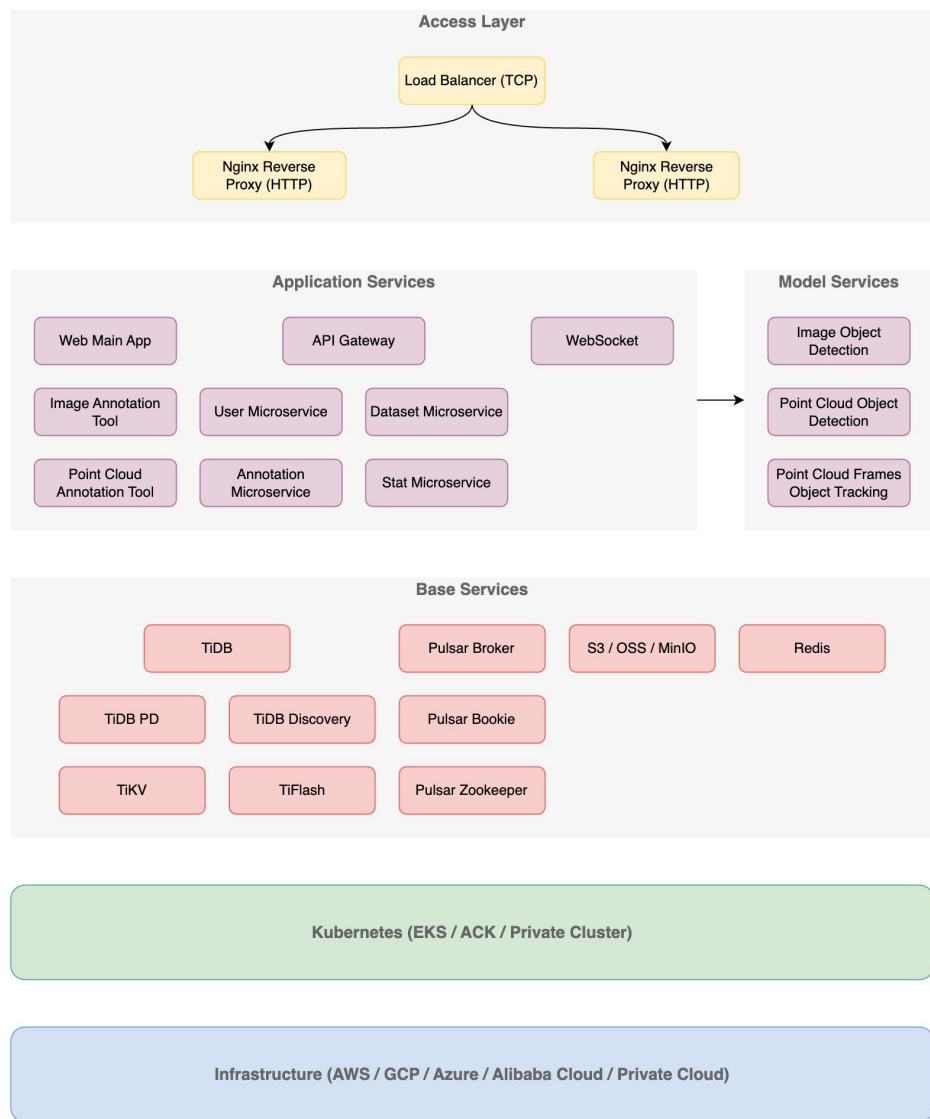
通过 Kubernetes 提供一致的运行环境，屏蔽底层基础设施差异

基础设施层

支持各大公有云、私有云，以及裸金属服务器

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE



平台建设的技術挑战



结构化数据存储

技术挑战

- 海量数据：按 1000 个租户，每个租户 100 万个数据 (Data)，每个数据标注出 30 个实例 (Instance)，每个实例 1 KB，那么总数据数量 10 亿，实例数量 300 亿，存储空间 30 TB
- 数据类型多样且关系复杂：Dataset、Data、Scene、Instance、Tracking Object、Class、Classification、Task、Team、User

解决方案

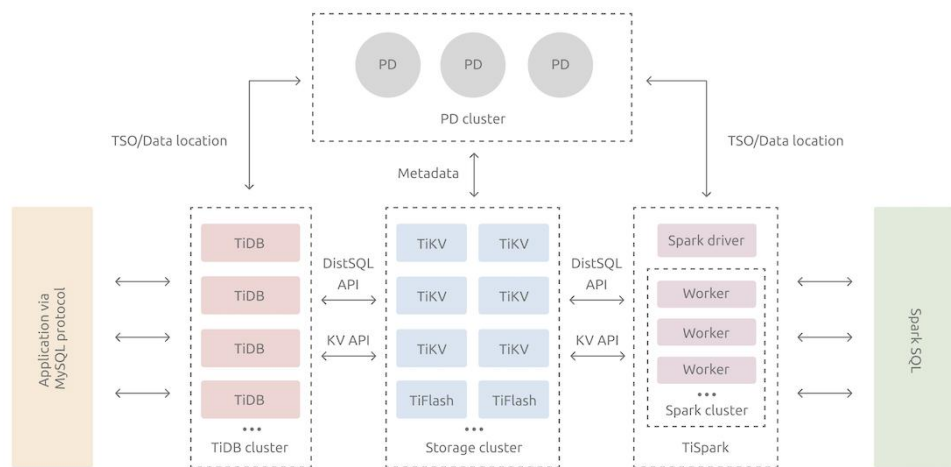
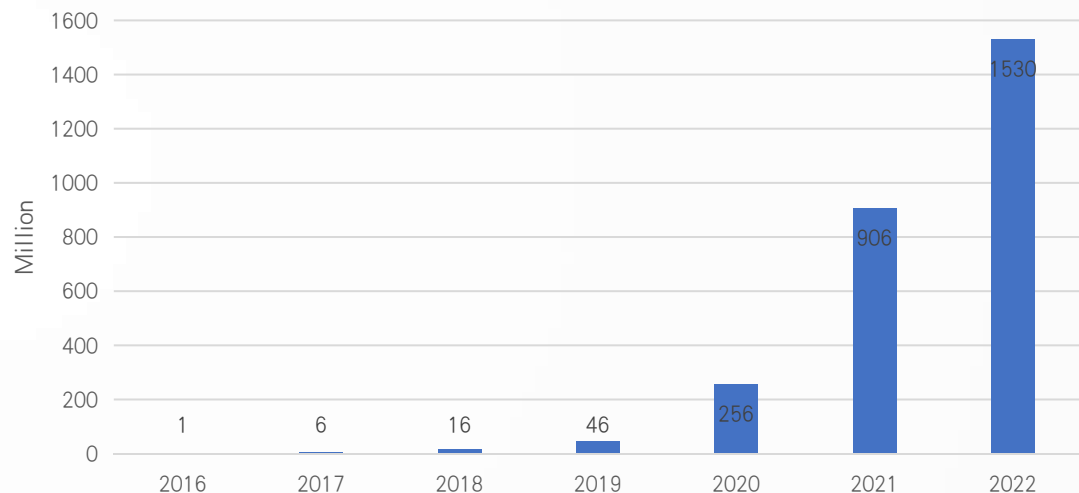
使用分布式关系型数据库 TiDB

- 纯分布式架构，拥有良好的扩展性，支持弹性扩缩容
- 兼容 MySQL 协议，大多数场景下可直接替换
- 支持高可用，少数副本失效情况下能自动修复数据和转移故障
- HTAP，同时支持 OLTP + OLAP

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

Data Amount Growth



平台建设的技術挑战

非结构化数据存储

技术挑战

- 数据类型多样：图片、点云、音频、视频、文本等
- 数据量大：单个图片从几百 KB 到几十 MB，单个点云从几 MB 到几十 MB，单个租户的数据量很容易就达到 TB 级别

解决方案

使用 MinIO 对象存储

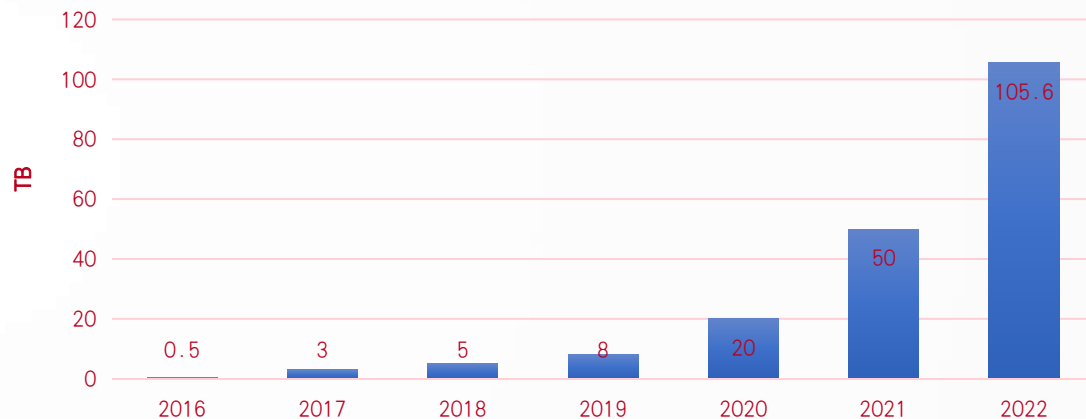
- 兼容 S3 API，可无缝替换为各大公有云的对象存储服务
- 使用可配置冗余度的纠错码来防止磁盘故障导致的数据丢失
- 支持服务端数据加密
- 持续复制能及时检测更新并快速复制，因此能支持大规模和跨数据中心场景
- 支持跨地域的联邦集群，将多个集群的资源统一在一个命名空间下

全球开源技术峰会

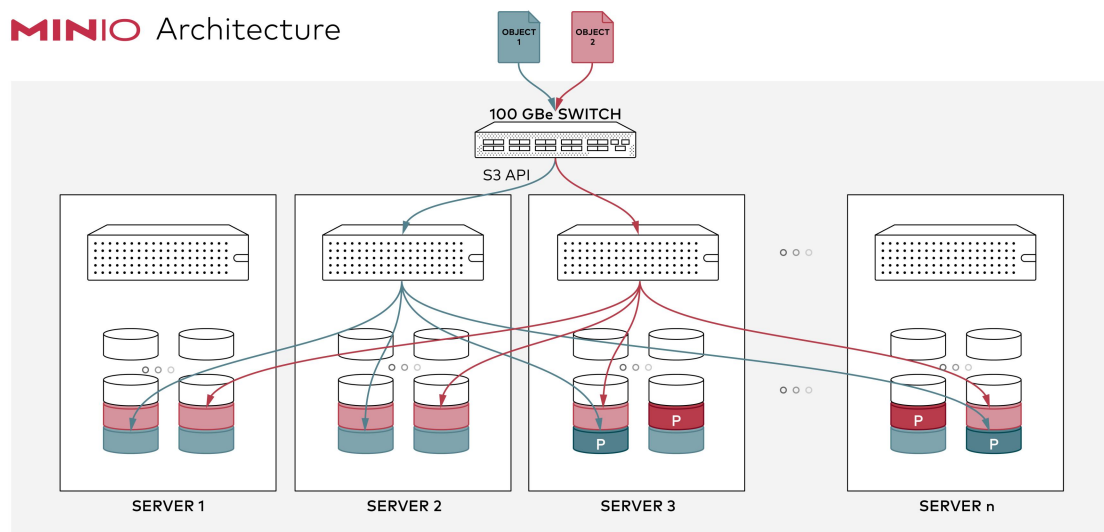
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE



Data Storage Growth



MINIO Architecture



minio server http://host{1...n}/export{1...m}

平台建设的技術挑战

异步消息通信

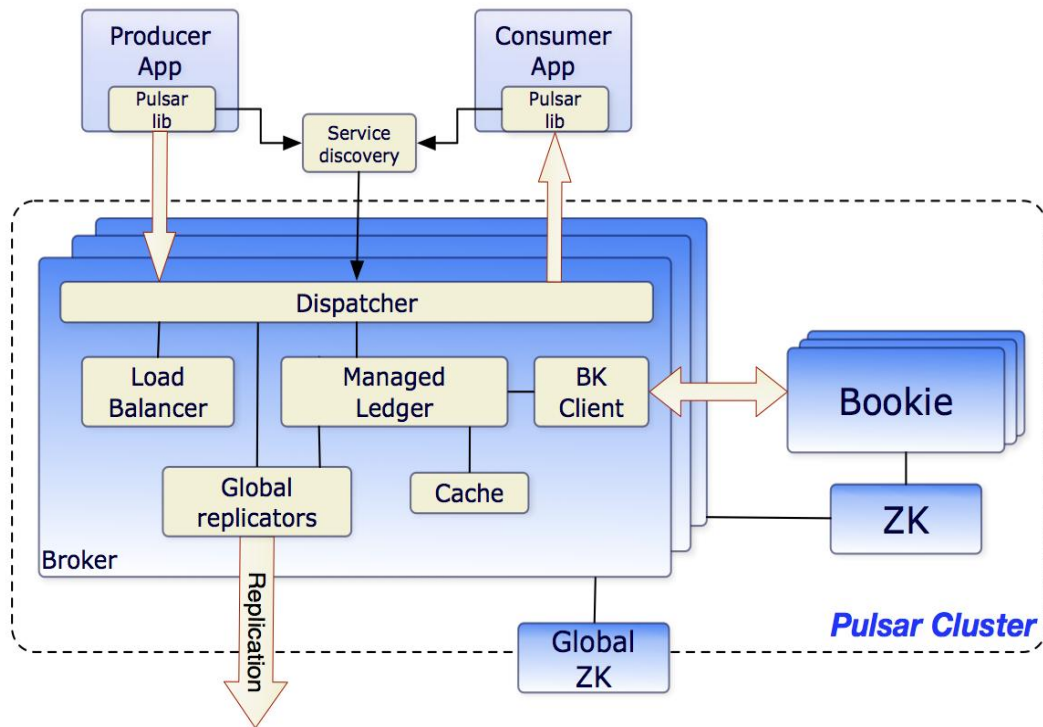
技术挑战

- 包含大量前台服务和后台任务的复杂分布式应用，除了需要同步的 API 调用通信，还需要异步的消息通信
- 耗时的数据处理需要异步通知后台任务来完成，比如缩略图生成、点云压缩、点云底图渲染、点云切分、模型预识别等
- 消息吞吐量大，单个操作可能就会产生几十上百万的消息

解决方案

使用 Pulsar 消息服务来支持异步通信和流式计算

- 云原生友好，存算分离，可独立扩展存储和计算节点
- 高性能，超低延时，轻松支持上百万 Topics
- 同时支持队列和流两种消息处理模式
- 通过轻量级的 Serverless 框架 Pulsar Functions 支持流原生的数据处理



平台建设的技术挑战



资源管理调度

技术挑战

- 多样化的部署环境，包括公有云/私有云/裸金属服务器、国内/国外、开发/测试/生产等
- 几十种服务和任务，包括前端服务、后端服务、后台任务、基础服务等，手动部署效率低且易出错
- 需要支持弹性扩缩容

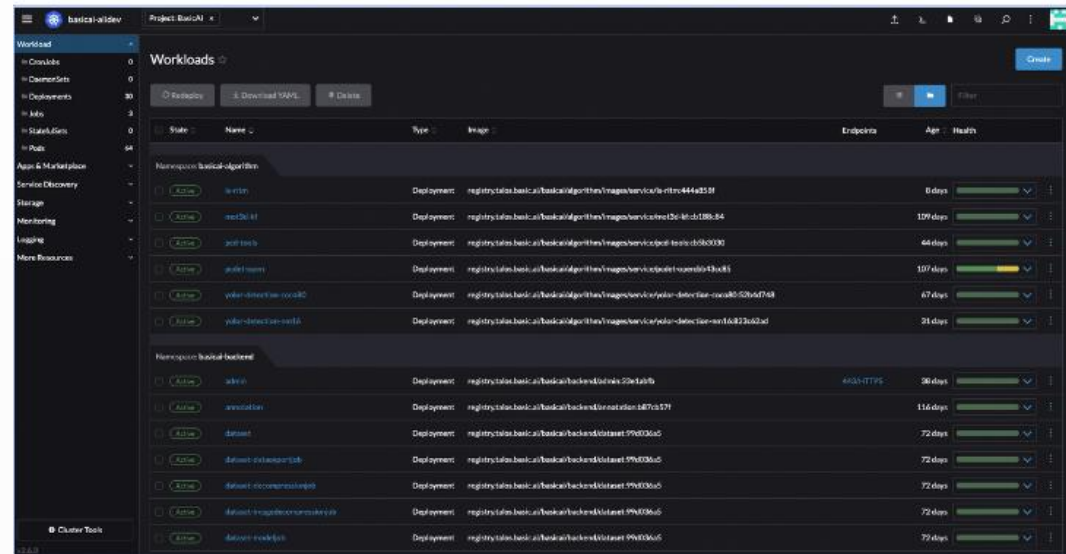
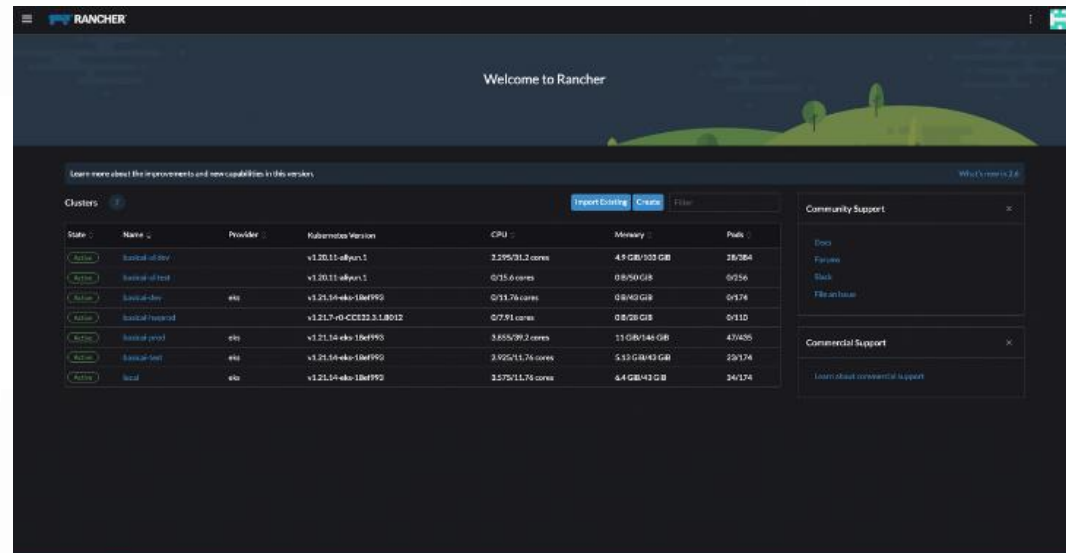
解决方案

Kubernetes + Rancher

- 使用 Kubernetes 统一部署方式，屏蔽底层基础设施差异
- 使用 Rancher 统一管理多个集群，借助于其 UI，不熟悉 Kubernetes 的研发人员也可以自行部署和管理 workload

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE



平台的未来规划和设想

Xtreme1 源于开源，拥抱开源，我们将持续向社区贡献代码，推动项目的健康发展

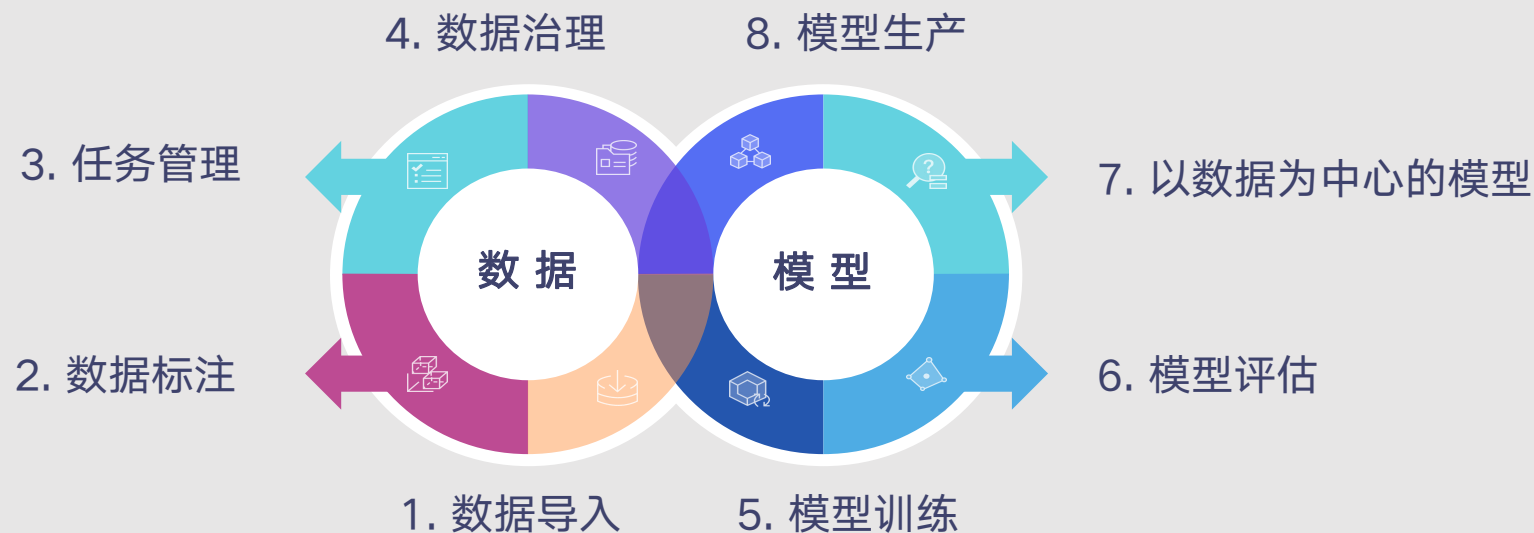
全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE



扫码备注“进群”加入
Xtreme1 技术交流群

Xtreme1 支持 Data-centric MLOps 完整生命周期



目标

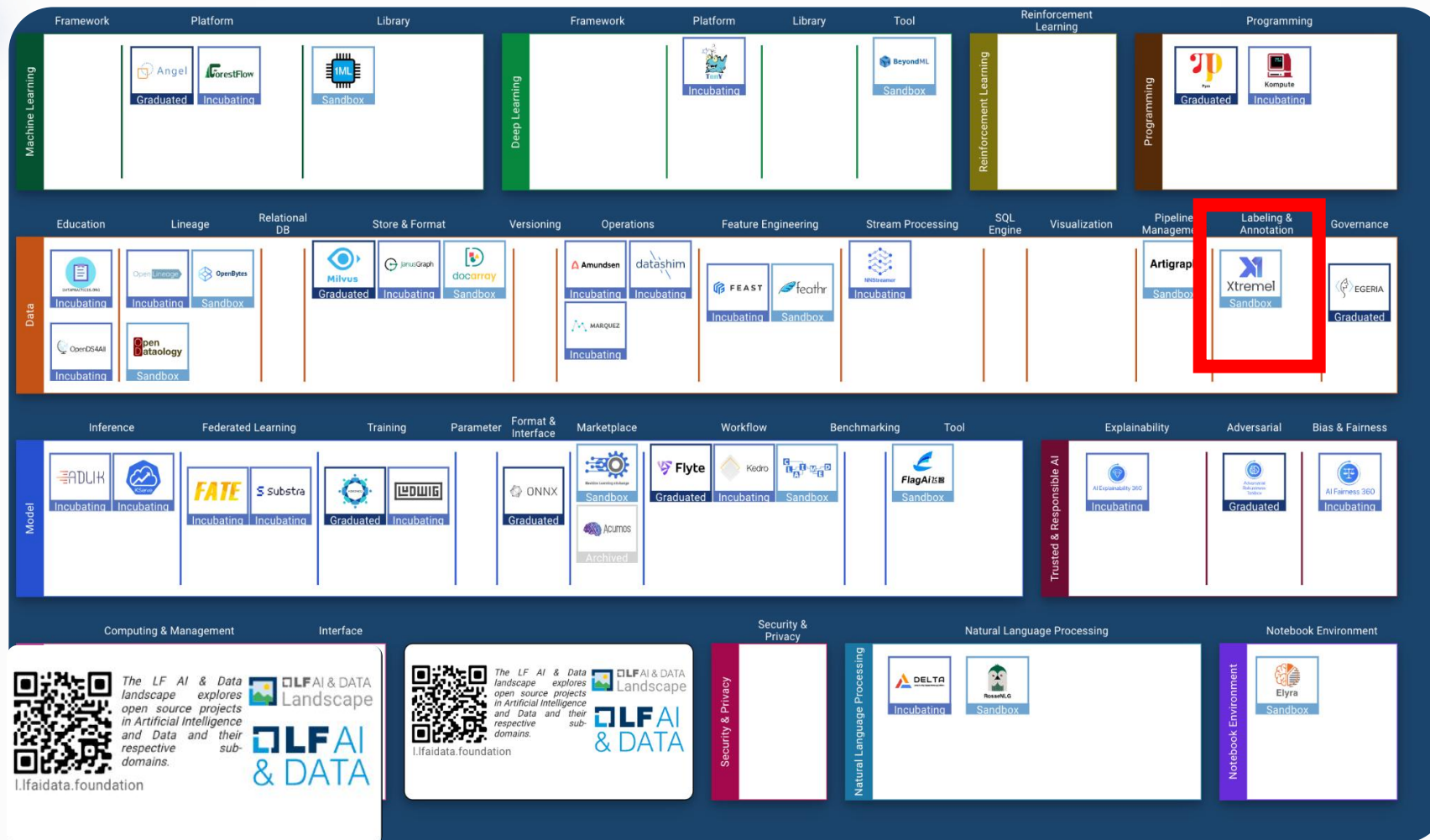
使用 AI 驱动的工具完成复杂的“训练数据”任务。

使命

构建最易于访问的 Data-centric 开源 MLOps 基础设施，以连接人、模型和数据。

信念

拥抱开源。开源是实现 Data-centric MLOps 的正确途径。

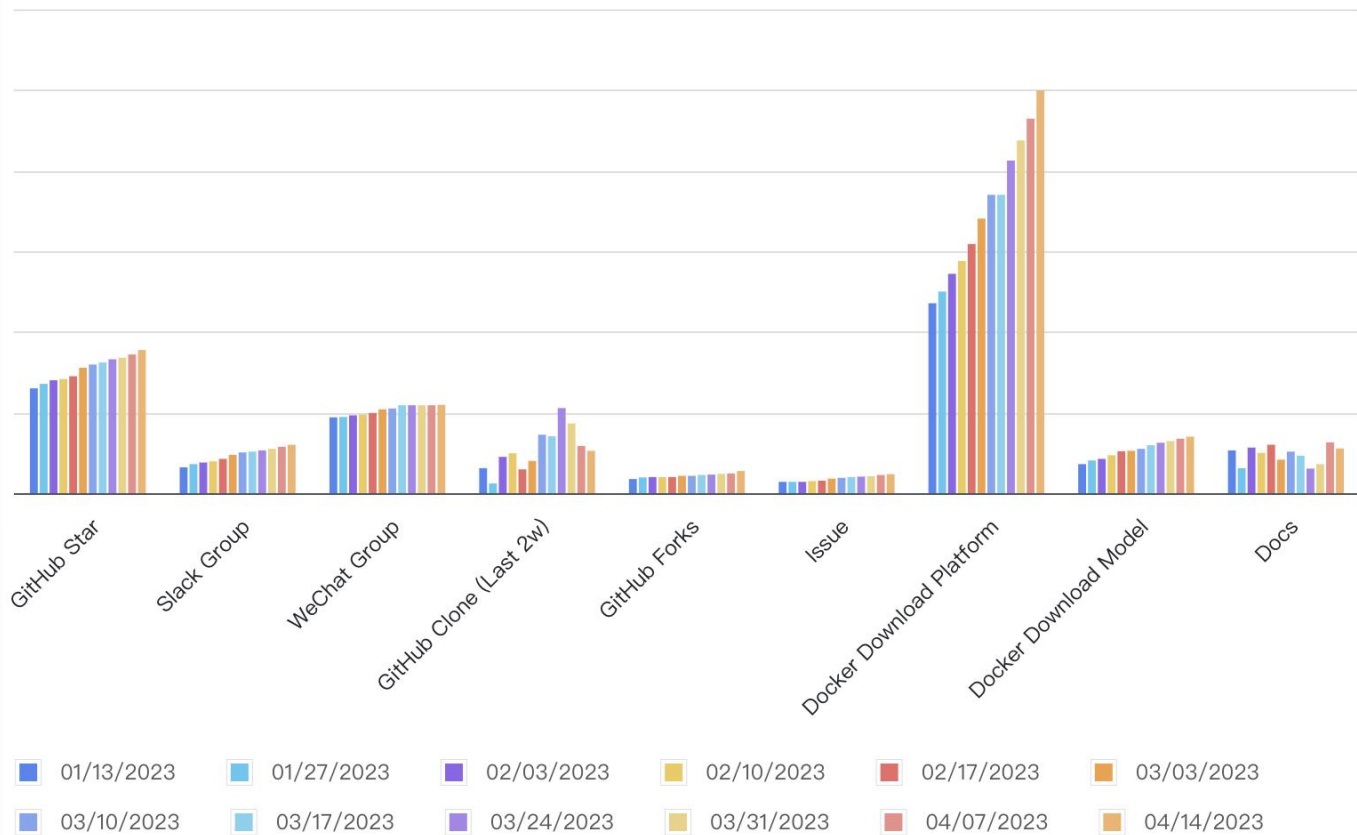


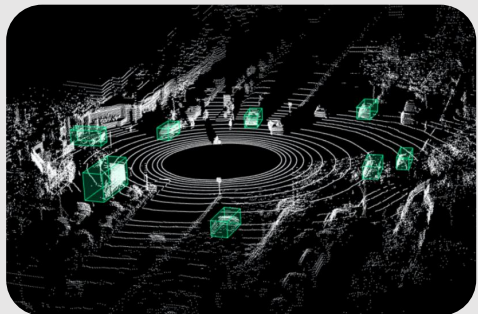


Product Roadmap 2023

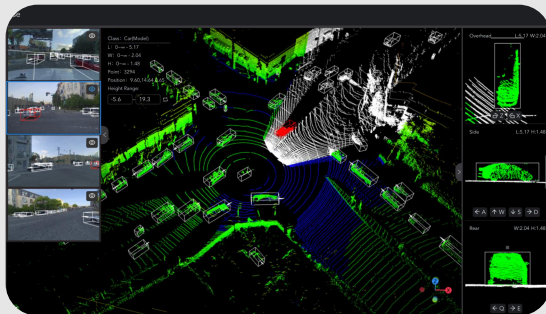
By 2023/03	Version 0.6	Model Access Point Cloud Segmentation
By 2023/04	Version 0.6.5	Meta Data Management
By 2023/05	Version 0.7	Data Versioning
By 2023/06	Version 0.7.5	Data Visualization
By 2023/07	Version 0.8	Data Organization
By 2023/08	Version 0.8.5	Data Augmentation Text Data Support
By 2023/09	Version 0.9	Audio Data Support
By 2023/10	Version 0.9.5	Support Relation Annotation
By 2023/11	Version 1.0	Ontology Fusion
By 2023/12	Version 1.0.5	Ontology Search
By 2024	Version 1.1	Ontology Recommendation

Xtreme1 (Open Source) Operation Statistics

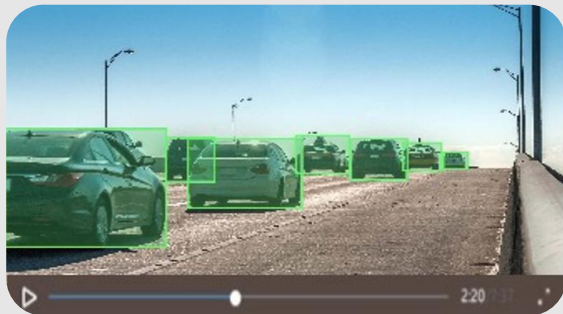




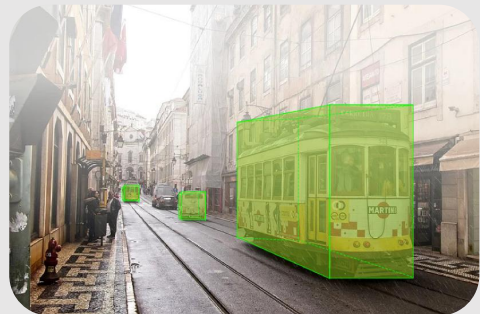
3D LiDAR Point Cloud



2D & 3D Fusion



Object Tracking*



2D / 3D Cuboid



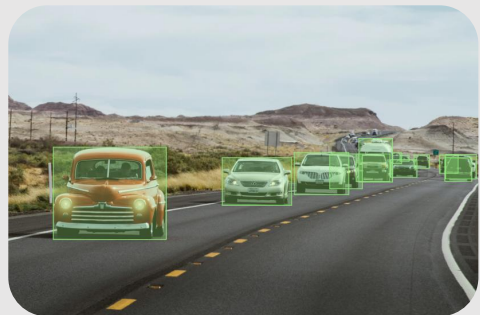
自动驾驶



智慧零售



智能家居



智慧城市

► 欢迎关注我们，获取最新动态



Xtreme1
微信公众号



Xtreme1
技术交流群

联系我们

GitHub Repo | github.com/xtreme1-io/xtreme1

文档 | docs.xtreme1.io

微信公众号 | Xtreme1-AI

官网 | xtreme1.io

Medium | xtreme1.medium.com/

YouTube | youtube.com/@xtreme1ai

Slack | xtreme1io.slack.com

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE



扫码备注“进群”加入
Xtreme1 技术交流群