火山引擎 | 多模态数据湖

Data+AI

火山引擎多模态数据湖解决方案

打破数据壁垒+加速企业 AI 落地



火山引擎数智平台出品



目 录
CONTENTS

01 趋势洞察：Data+AI 引领企业智能化变革 04

02 困境与破局：从传统数据湖到多模态智能进化 06

03 解决方案：多模态数据湖解决方案全面介绍 08

04 典型案例：六大客户案例拆解核心场景实践 11

05 生态共建：开源技术驱动开放生态体系构建 27

01

Data+AI
引领企业智能化变革



AI 时代：数据智能迎来全方位进化拐点

当前，生成式 AI 与大模型技术的爆发式发展，推动行业迈入变革关键期，数据智能也迎来以 “数据、工具、组织” 为核心的全方位进化，并逐步成为企业智能化升级的核心动力。

数据智能的三大进化维度

01 数据维度：从 “聚焦结构化” 到 “多模态融合” 进化

传统数据智能仅聚焦企业 10%–20% 的结构化数据，而 AI 时代下，图像、视频、音频、文本等非结构化数据占比已超 80%，亟需通过多模态数据融合，构建更立体的用户认知，真正释放数据资产的全面价值。

02 工具维度：从 “人类延伸” 到 “协作伙伴” 进化

过去，数据工具依赖人工完成数据提取、分析等基础流程，进化后的工具依托 MCP、A2A 等技术框架，可自主实现 “数据发现 — 分析决策 — 行动执行” 全链路自动化，将人力聚焦到创新等高价值环节。

03 组织维度：从 “人力驱动” 到 “人机协同” 进化

传统组织形态以“人力驱动”为核心，一线员工常陷入“数据搬运”、“重复执行”的低效工作，而 Data+AI 推动组织向 “人机协同” 转型，既提升个体价值，又通过人机协作破解组织管理痛点，实现组织效能提升。

Data+AI 驱动数据智能的三大维度协同进化，深刻重塑着各行业的发展逻辑：消费领域依托多模态数据实现“千人千面”的精准服务，工业领域通过人机协同优化生产流程，金融领域借助智能工具提升风险管控效率.. 无论哪个行业，能否抓住这一进化机遇，关键之一在于是否构建了适配 AI 的多模态数据基础设施——这既是数据智能落地的 “底座”，也是企业抢占竞争制高点的核心前提。

02 困境与破局：从传统数据湖到多模态智能进化

■ 传统数据湖的四大核心困境

01 数据覆盖局限

传统数据湖以结构化数据为核心，对图像、视频、音频等非结构化数据的存储、处理能力不足。例如，企业在存储视频数据时，需额外搭建独立的对象存储系统，导致数据分散存储，无法实现多模态数据的统一调用，严重制约 AI 应用的数据支撑能力。

02 成本与效能失衡

随着数据量呈指数级增长，传统数据湖采用的“存算一体”架构，需同时扩容存储与计算资源，导致存储成本年均增长 30% 以上；同时，缺乏针对多模态数据的优化处理引擎，在分析视频帧、图像特征等数据时，处理效率较 AI 应用需求低 5-10 倍，陷入“存得贵、用得慢”的困境。

03 质量与标注困境

文本、图像等非结构化数据噪声多，既有格式不规范等显性问题，也有来源不可靠等隐性缺陷，且清洗无标准流程，人力投入大却难保数据一致性，数据质量不稳定。此外，AI 时代下，监督学习高度依赖含高质量标注的数据，而大规模标注存在“成本高、周期长”的问题。

04 生态兼容性差

传统数据湖与主流 AI 框架及大模型平台的对接能力不足，数据需经过多次格式转换才能输入模型训练，导致“数据—模型—应用”链路断裂。例如，企业在利用数据湖中的图像数据训练视觉模型时，需人工完成数据格式从湖存储到模型平台的转换，整个流程耗时且易出错。

■ 破局关键：多模态数据湖的智能突破路径

火山引擎多模态数据湖解决方案并非传统数据湖的简单升级，而是围绕 AI 时代的数据智能需求，以“湖存储、湖管理、湖计算、湖生态”为支撑，实现针对性突破，构建适配多模态数据的全链路解决方案。

01 湖存储突破：高性能与低成本双优

针对传统数据湖“存得贵、用得慢”的问题，多模态数据湖通过 Lance 列式格式+存算分离架构实现多模态数据统一存储，数据读取性能提升，冷热分层降低存储成本。

02 湖计算突破：异构资源调度与效率跃迁

针对传统数据湖计算效率低、难以适配 AI 任务的痛点，多模态数据湖支持 CPU/GPU 混合调度+内置 AI 算子库，依托深度优化的 Ray/Daft 分布式引擎，使计算性能提升 2-5 倍，且支持秒级扩缩容，轻松应对高频计算需求。

03 湖管理突破：统一元数据管理

为了解决元数据与数据割裂、数据质量不稳定、数据存取性能低等问题，多模态数据湖以统一元数据管理为核心，统一存储与管理数据与元数据，统一权限管理，提供数据集版本控制与安全隔离，为数据质量稳定提供关键支撑。

04 湖生态突破：开放兼容与场景闭环

为解决传统数据湖生态兼容性差、AI 链路断裂的问题，多模态数据湖构建开放协同的生态体系，深度兼容 Lance、Ray、Daft 等开源框架，无缝对接火山方舟大模型平台，覆盖 AI 全场景应用。

03 解决方案：多模态数据湖解决方案全面介绍

方案定位

为了帮助企业更好地拥抱大模型时代，我们推出了火山引擎多模态数据湖解决方案，作为面向 AI 时代的智能数据基建，它能帮助企业实现对文本、图像、音频、视频等非结构化数据资产的统一精细化管理，兼容多种数据格式，为模型预训练、后训练和 AI 应用构建/开发等场景提供端到端的高效可靠、便捷易用的智能数据服务，帮助企业在 AI 时代构筑全面领先的数据竞争力。

核心能力

多模态存储

文本/图像/音视频等入湖存储，深度优化新一代湖格式 Lance、Iceberg

数据集管理

提供 AI 场景的数据集管理能力，包含数据查询、数据编辑、版本管理等

统一元数据

统一管理多模态数据的元数据，统一权限管理，保障数据安全

数据处理算子

提供 AI 工作流编排能力，内置数百个多模态数据处理算子、自定义算子

多模态数据处理

深度优化 Ray、PySpark 分布式引擎，支持 CPU、GPU 异构资源混合调度

调用豆包 & DeepSeek

产品内支持基于豆包家族、DeepSeek 等模型进行多模态数据处理

在线服务

提供开源模型、用户自定义镜像 Serving 功能，支持服务灵活部署

对接 AI 生态

无缝对接火山方舟、机器学习平台等，支持训练、微调、蒸馏等场景

算子展示



方案架构



方案优势

资源成本低

支持 GPU+CPU 异构计算，大幅提升模型训练中资源使用率

开源开放

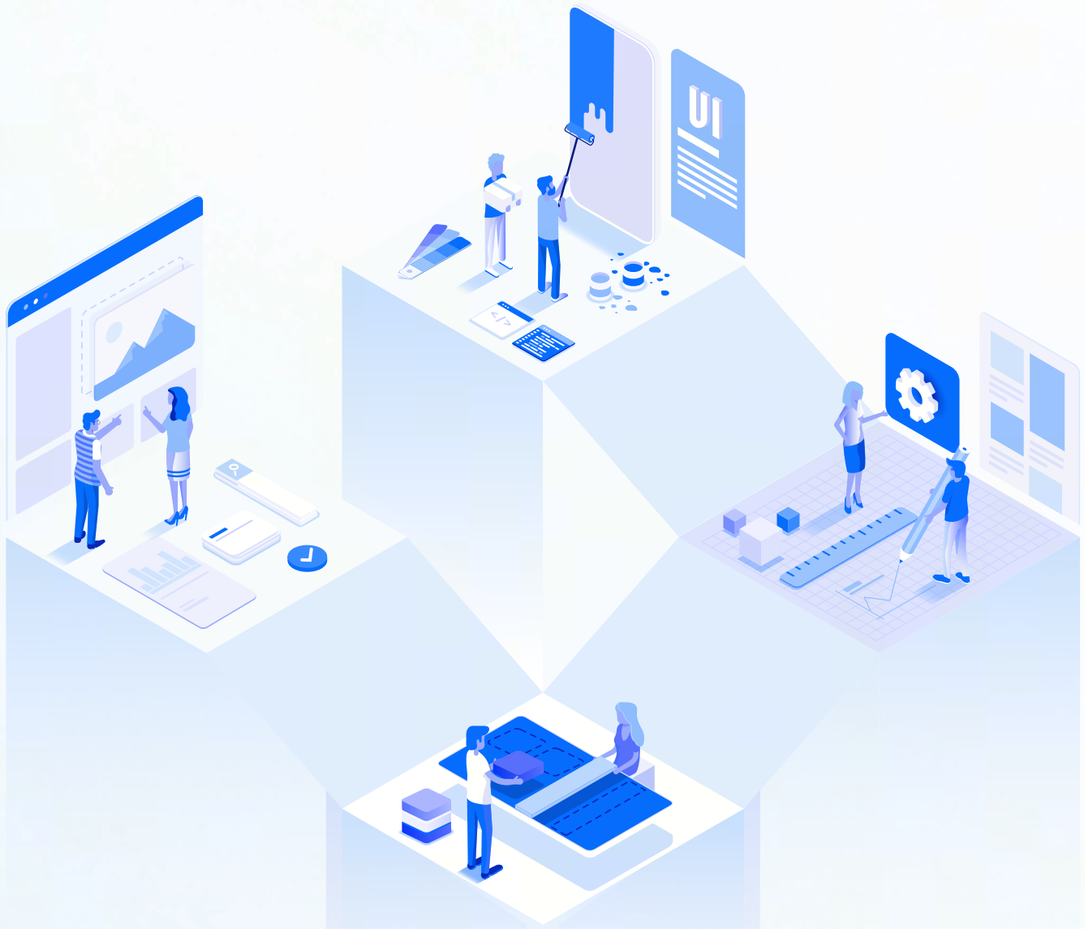
采用开源技术并提供企业级能力增强，在计算存储、管理应用等多个层面解耦，支持灵活接入，更易集成

新型技术栈

引入新型 Ray、Lance、Daft 等前沿技术栈，持续帮助企业享受领先的技术红利

产品联动性强

联动火山机器学习平台、火山方舟平台、数据智能体等，实现多场景一站式自由流动



04 | 典型案例：六大客户案例拆解核心场景实践

汽车行业 – 智驾场景

聚焦车企智驾数据处理，通过资源调度与数据管理优化，提升任务效率、GPU 利用率，降低存储成本

智能制造 – 具身智能场景

助力具身智能研发的数据需求，高效处理感知、交互类多模态数据，支撑算法训练与功能迭代

人工智能行业 – 模型训练场景

服务 AGI 企业基座模型训练，支撑模型在图像理解、语音交互等领域保持领先

创新科技行业 – 视频数据挖掘场景

高效处理海量视频数据，提取关键信息与特征，满足智能内容分析等应用需求

传媒文娱行业 – 图片数据挖掘场景

对图片类多模态数据进行精准处理与特征提取，支撑图像识别、标注分析等场景

知识服务行业 – 文本数据挖掘场景

针对多语言文献等文本数据，实现快速获取、翻译存储与深度挖掘，服务知识库搭建等场景

汽车行业

某车企：多模态数据湖在智驾场景的应用

■ 企业简介

该车企旗下拥有多个品牌，包括传统品牌、专注新能源领域的品牌以及高端智能电动汽车品牌，坚持以科技创新为核心驱动力，全面重塑企业能力、升级产业结构。近年来，该车企高度聚焦智能网联（ICV）场景，在构建“车 - 路 - 云 - 网 - 图”复杂应用场景中，面临多维度技术挑战，对海量数据实时处理与复杂算法快速迭代有极高要求。

■ 企业痛点

01 任务处理速度慢

- 无法在一个工作流里同时调用 CPU 和 GPU 资源
- CPU 算子并发需手动多进程管理
- 任务处理 PB 级别数据耗时 7 天+

02 资源成本高

- 单卡 GPU 任务无法并发，资源利用率低，整体 GPU 利用率只有 30%～60%
- 部分在线标注模型，任务完成后无法自动释放资源

03 存储冗余，数据管理效率低

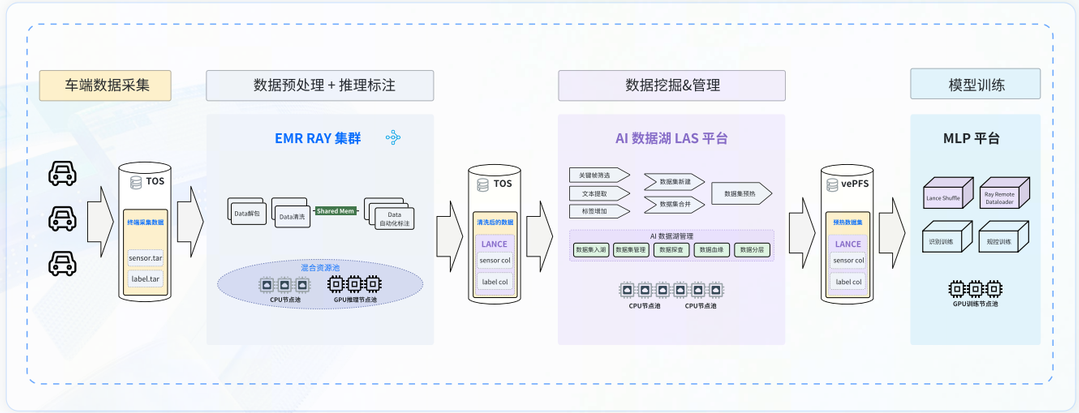
- 增加标注需全量重建数据集，资源消耗大且占存储
- 数据版本依赖手工文件，数据管理人力成本高

04 模型训练效率低，迭代时间长

- 数据加载引发 CPU 资源负荷过高
- 计算资源无法高效协同，模型迭代周期长

■ 解决方案

通过火山引擎多模态数据湖解决方案，在推理环节，EMR Ray 支持单卡 GPU 并发模式，可同时运行多个标注推理模型，充分利用计算资源实现降本；在数据挖掘与管理环节，LAS 数据预处理仅追加新增标签列、Lance 支持压缩，还具备丰富数据集管理能力，既减少 GPU 消耗与存储占用，又大幅提升数据管理效率。



■ 效果价值

01

任务效率 **提高 7 倍**

在一个工作流内实现 CPU 和 GPU 的混合调度，从单机处理变成分布式处理，极大提高数据任务处理效率

02

资源利用率 **升至 95%**

单卡 GPU 可以同时支持多个并发标注推理模型，GPU 资源利用率升至 95%，资源的自动释放

03

存储成本 **降低至 1/4**

LAS Lance 做数据预处理，无需重写全量数据，客户仅需关心新增的标签列

04

模型训练效率 **提升 1.5 倍**

EMR Ray 解决 CPU 满载的问题，将数据加载效率大大提升，加速训练进程。

智能制造行业

某 AI 创新研发企业： 多模态数据湖在具身智能场景的应用

■ 企业简介

某 AI 创新研发企业以具身机器人为战略定位，伴随业务场景持续拓展，已实现多模态数据资源规模化沉淀。为将数据资产转化为智能势能，该企业通过采集系统完成大规模图像、视频等多模态数据的自动化采集，并借助智能质检、深度清洗、精准标注等多道数据处理工序，持续提升家庭场景下具身智能机器人服务能力。

■ 企业痛点

01 数据血缘断裂

数据处理过程中版本关联信息缺失，严重影响数据质量回溯追溯与合规性审计工作

02 流程管控缺位

缺乏自动化编排机制，依赖人工操作，不仅耗时费力，错误率更是高达 15% 以上

03 处理效能偏低

随着数据量激增，单机程序难以保障数据处理时效性，无法满足业务快速推进需求

04 搜索体验割裂

在社区搜索场景中，传统架构难以整合多模态内容，且大模型缺乏高质量多模态数据支撑，无法精准理解用户意图，搜索体验割裂

■ 解决方案

通过火山引擎多模态数据湖解决方案，整合多模态算子、领先技术框架、统一湖底座及异构混合调度资源池等核心能力，通过 “数据集 + 工作流 + 开发机 + 分布式框架” 全链路，助力客户构建高性能、高效率、低成本、一站式的 AI 数据处理平台，支撑模型快速迭代与 AI 业务落地。



■ 效果价值

01

释放人力 50%

借助 AI 多模态数据湖服务提供的数据集，成功实现数据的全链路可回溯与合规可审计功能，大幅降低人为管理数据的复杂度，节省人力成本

02

开发提效 10%

依托工作流进行任务的自由化编排时，只需聚焦于原子性任务的设计，系统便能通过工作流自动完成原子任务的组装，提升开发和运维效率

03

数据处理性能提升 20%

基于 AI 数据湖服务提供的多模态算子，有效支持大模型在多机多卡环境下的推理能力；同时结合 Daft 技术架构实现数据处理，提升整体处理效率

04

成本与合规双赢

节约 30%+ 云服务器成本，助力海外市场规避跨境数据合规风险

人工智能行业

某 AGI 企业： 多模态数据湖在模型训练场景的应用

■ 企业简介

某 AGI（通用人工智能，Artificial General Intelligence）企业，为维持其核心模型在图像理解、语音交互、视频解析等领域的竞争力，需要对高达 10PB 级别的爬虫多模态数据进行预处理，转化为标准数据集，并用于模型训练，因此对数据预处理的整体时效性有极高要求。

■ 企业痛点

01 数据索引效率低

采用 WebDataset 存储多模态数据及元数据，需全量解压引发高频 IO 操作，存在分钟级别延迟

02 存储稳定性不足

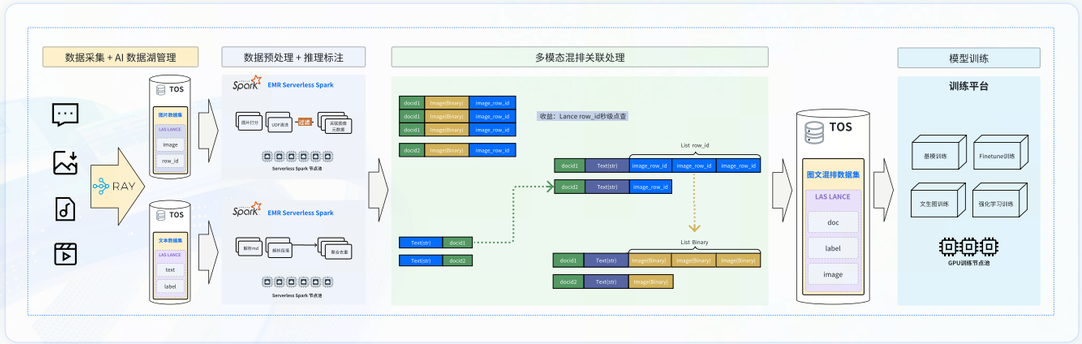
IDC 自建存储难支撑 PB 级规模，易超 QPS 阈值导致服务降级、连接拒绝，任务失败率达 30%

03 资源拓展受限

多模态数据格式异构、逻辑分块模糊，图文关联 Join 产生超大规模 Shuffle，传统处理方式无法短时间拓展至百亿级别

■ 解决方案

火山引擎多模态数据湖解决方案提供多源异构数据的统一管理、低成本高性能的多模态数据存储、高性能可拓展的数据处理和检索服务，能够高效处理多模态数据（如图像、文本、音频和视频等多种数据类型），支撑复杂数据集的 AI 场景（如多模态大模型训练）等，快速挖掘数据价值。



■ 效果价值

01 时效跃迁至秒级

借助 Lance 高效的随机读写性能，应对数亿级别数据集，秒级检索与点查。

02 存储性能大幅度提升

Lance 合并存储元数据和数据，QPS 压力减轻，降低任务失败率。

03 数据处理效率 提升 8 倍+

异构资源 CPU、GPU 混合调度并秒级扩缩容，任务成功率提升至 99.9%。

创新科技行业

某创新型科技企业： 多模态数据湖在视频数据挖掘场景的应用

■ 企业简介

某创新型科技企业专注 AI 内容领域，借数据挖掘与分析提供内容营销解决方案，可 AI 批量生成高质量视频素材、输出科学化内容策略，其服务的客户涵盖 AI、大数据、智能营销、云计算等领域，客户需搭建内容处理平台，核心场景为视频加工打标、关联投放效果数据并产出营销策略报告。

■ 企业痛点

01 工程化平台缺失，开发效率低

- 单一代码仓库，功能模块高度耦合，开发复杂度高
- 代码结构复杂，运维人力成本高
- 缺乏可视化监控，依赖人工发现异常

02 计算资源无法弹性扩展

- 分类标签任务耗时长达 10+小时
- 单机 Python 脚本缺少分布式框架
- 集群扩容周期长、维护成本高

03 技术整合升级周期长

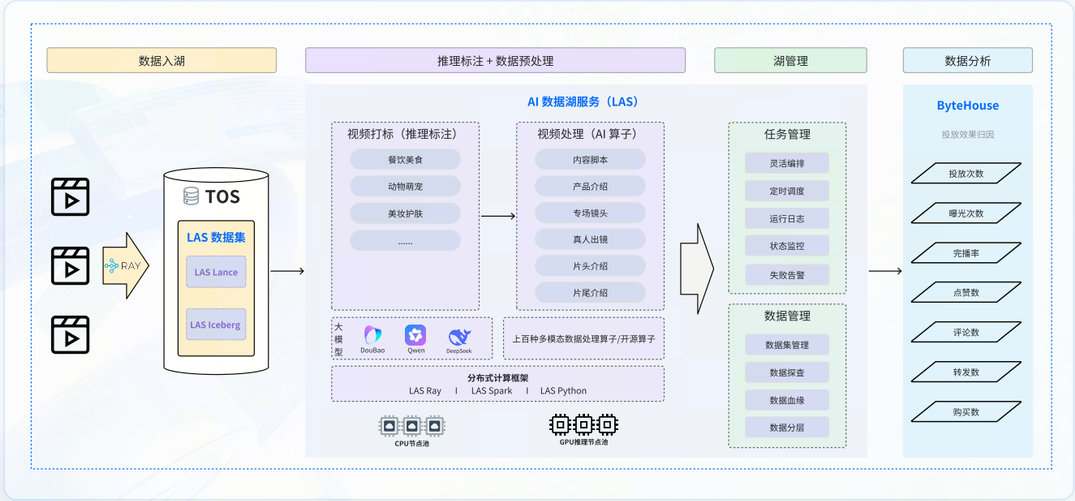
- 开源方案适配成本高，需要深度改造
- 新技术融入现有框架需消耗大量开发成本
- 平台能力缺失，制约创新业务快速落地

04 技术栈多，系统维护成本高

- 原架构涉及多异构系统，系统维护成本高
- 数据流转跨越多个组件，数据处理链路繁杂

■ 解决方案

通过火山引擎多模态数据湖解决方案，AI 数据湖 LAS 统一接入 ES、OSS、MC 等异构数据，消除孤岛；LAS 平台可视化构建工作流，集成 100+AI 算子处理多模态数据，降低开发门槛；结合 Ray 分布式计算引擎，将单机任务转为分布式处理，大幅提升效率。



■ 效果价值

01

开发效率 **提升 50%+**

标准模块化开发，无需深入理解整体架构，即可快速定位模块并进行迭代，大幅降低协作成本

02

处理效率 **提升 5 倍+**

低代码分布式化，单机任务轻松扩展；自动管理分布式集群的资源调度、弹性扩缩容，确保任务稳定运行

03

运维效率 **提升 60%**

可视化监控快速定位异常节点，智能日志分析工具使问题排查平均耗时大幅缩短，提升排障效率

04

技术栈统一化管理

一站式数据架构减少跨平台依赖，Lance 格式替代传统的"对象存储+开源数据检索引擎"方案，实现向量和视频的高效联合检索

传媒文娱行业

某媒体传播平台： 多模态数据湖在传媒行业 AI 媒资库构建场景的应用

■ 企业简介

该企业主营报刊出版、新媒体内容与广告运营等业务，同时布局了品牌活动与图书出版业务。企业拥有近 30 年的内容沉淀，积累了海量文本、图片、视频等多模态媒体资产。2025 年起，将旗下媒资平台作为数字化转型重要基础设施先行验证，旨在对多模态媒体资产统一治理和智能化管理，支撑 AI 应用场景落地，提升内容生产和运营效率。

■ 企业痛点

01 工具链路分散，工程化能力弱，交付效率低

- 多厂商工具拼接，链路跨云、跨厂商，问题定位难，运维负担重，系统稳定性差
- 跨团队协作成本高，修改周期长，同类算子重复建设，开发效率低

02 数据质量不稳定，人工成本高，业务价值难挖掘

- 复杂版面解析效果差，需大量人工二次校对，解析失败与重跑频发影响效率
- 元数据不完整不统一，关键信息缺失或格式不一，人工补录成本高昂

03 跨模态数据分散，检索能力不足，编采效率低

- 跨模态数据分散，缺少统一向量化与索引能力，难以实现关联查询
- 检索响应慢且功能单一，仅支持关键词检索，无法满足智能检索需求

■ 解决方案

通过火山引擎多模态数据湖解决方案，构建一体化媒资智能管理平台：在数据接入层，用 DataSail 对接多源数据，实现批量 / 实时数据接入；智能解析处理层借 LAS 的 100+ AI 算子（如 PDF 解析、OCR 等）处理数据，以结构化技术提升复杂版面解析效果；多模态存储层用 Lance 表统一管理多类型数据；向量检索层靠 ByteHouse 提供云原生向量索引与混合检索能力，最终在智能应用层支持多模态检索等 AI 应用，实现媒资入湖到智能应用的端到端解决方案。



■ 效果价值

01 工程交付效率 提升 60%

- 统一管控中枢，替代多厂商拼接方案
- 可视化工作流与 100 + 内置算子实现高效数据处理
- 清洗解析链路开发时间从天级降至小时级

02 元数据完备度 提升 80%

- LAS 内置 PDF 解析算子 + 结构化理解与版面建模技术
- 自动提取完整元数据，元数据完备度提升 80%
- 支持多维度检索与专题聚合，减少后续人工补录

03 内容生产和运营效率 提升 90%

- ByteHouse 向量化检索，响应从十秒级缩至毫秒级
- 支持跨模态检索（以文搜图、片段视频检索等）
- 专题线索汇聚从小时级缩至秒/分钟级，编采与运营效率提升

04 内容价值 挖掘从 0 到 1

- 构建企业级知识引擎，统一管理近30年媒体资产
- 知识图谱 + RAG 技术，支持专题生成、智能问答等
- 将沉睡的媒体资产转化为可持续产生价值的知识资产

知识服务行业

某知识服务平台：多模态数据湖在文本数据挖掘场景的应用

■ 企业简介

某知识服务平台凭借在交叉学科的深厚积累，搭建起 AI for Science 大模型体系，向能源、医疗、金融等多个行业提供智能决策支持。该企业一项核心业务是实时收集全球前沿技术文献，这些文献涵盖英语、日语、德语等多种语言。经过翻译处理后，存储到 MySQL 数据库，以服务于内部知识库搭建、智能问答，以及行业报告生成等实际应用场景 。

■ 企业痛点

01 海量异构数据开发成本高

- 10w+ 表数据
- 3+ 数据存储格式

02 运维人力投入大

- 10GB 日志数据人工检索排查故障
- 20+ 项关键指标无法实时监控

03 翻译任务成功率&准确率低

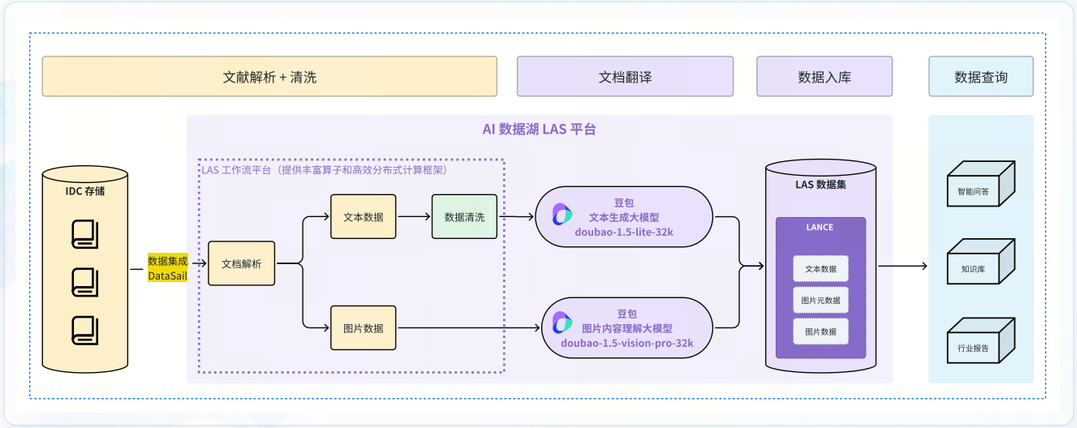
- 10w+ 表处理任务，单表失败影响整体知识库
- 40%+ 的示意图标注信息丢失率

04 图文数据分离导致数据挖掘难

- 数据存储割裂难以统一管理和级联查询
- 5 秒数据延时难支撑跨模态检索

■ 解决方案

通过火山引擎多模态数据湖解决方案，首先借助 DataSail 数据集成工具同步客户存量数据，将 IDC 中的 MySQL 与对象存储数据写入 LAS 多模态数据湖；接着在 LAS 中搭建周期性调度工作流，调用豆包大模型完成数据翻译，翻译后的数据回传并更新至数据湖，实现高质量数据集的存储与管理；最后依托数据湖的版本管理和快照能力，可查询历史数据状态，支持数据跟踪与回退。



■ 效果价值

01 开发周期 缩短 67%

员工只需通过可视化拖拉拽方式，即可 0 代码创建数据处理工作流，开发时长从 3 周缩短至 1 周

02 翻译准确率 提升 5%

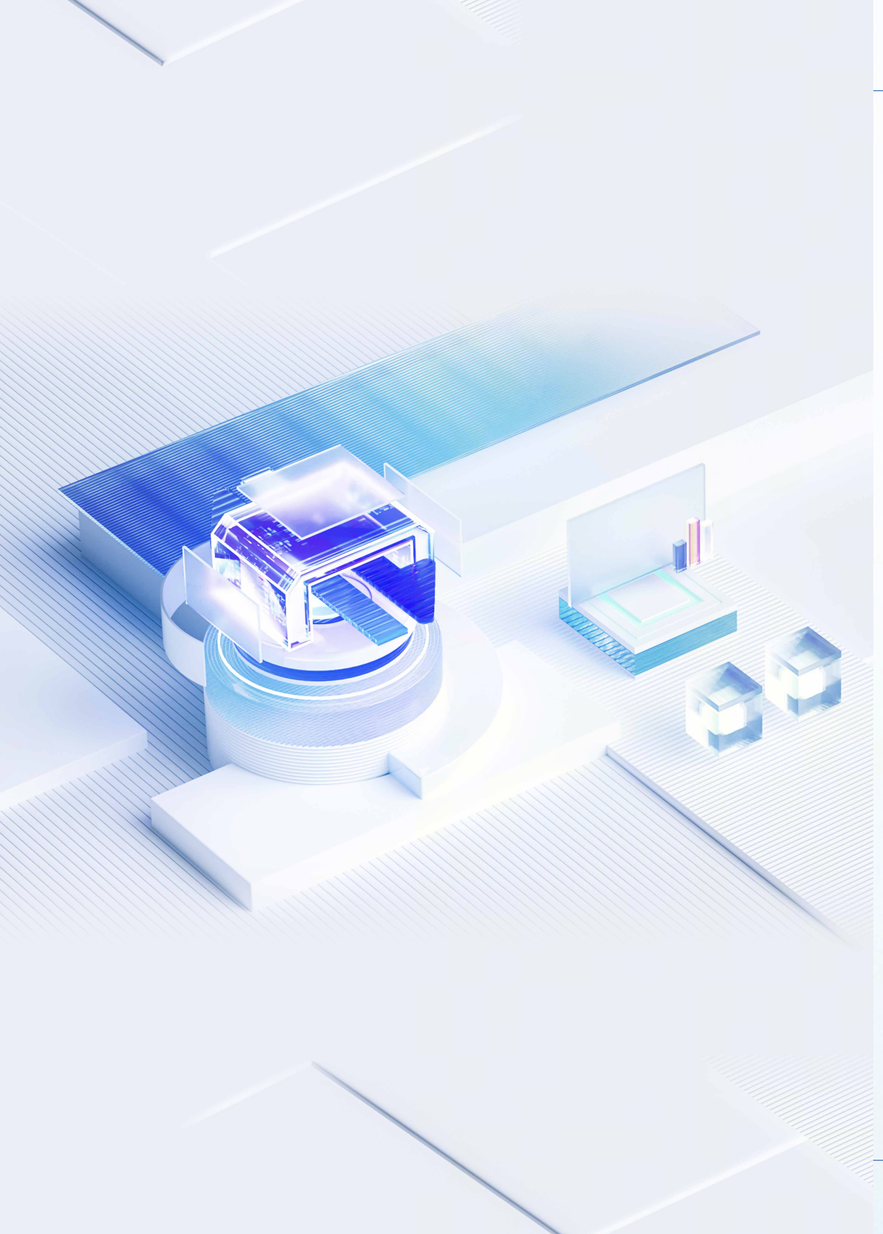
基于数据预处理+豆包模型翻译提升准确率；优先保障整体表任务执行，针对异常数据可自动保存并单独管理

03 运维 提效 50%+

员工通过 LAS 算力管理页面监测各项数据，并可快速配置告警规则，让「人力监测」转化为「规则监测」

04 图文数据 分析从 0 到 1

基于 LAS 构建多模态数据湖，企业的异构数据可以通过 Lance 架构统一存储，以数据集格式统一管理



■ 大数据基础设施产品家族



AI 数据湖服务 LAS
大模型时代 AI 数据基建

数据湖计算&存储

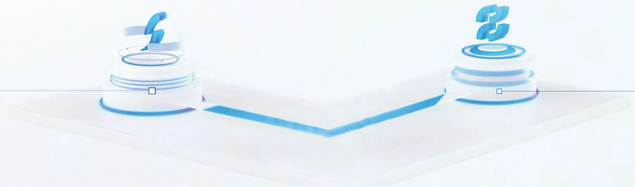


E-MapReduce
云原生开源大数据平台

流式计算 Flink 版
云原生全托管实时数据处理分析服务

ByteHouse
极致性能弹性伸缩的分析型数据库

引擎&向量化检索



大数据研发治理套件 DataLeap
一站式数据中台套件

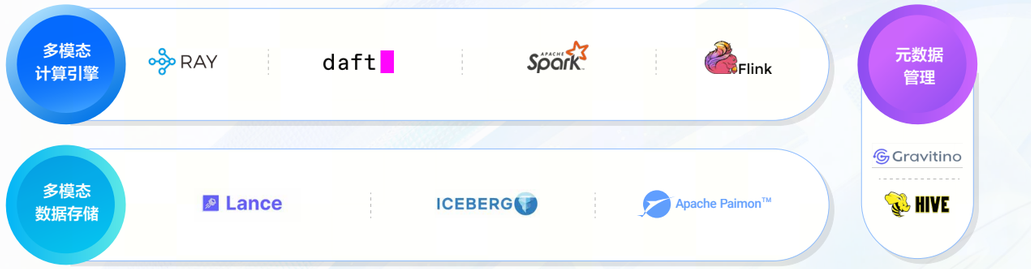
全域数据集成 DataSail
高速稳定的批量和流式数据同步引擎

数据集成&开发



05 | 生态共建：开源技术驱动开放生态体系构建

■ 技术组件的架构图



■ 核心优势解读

01 100%兼容开源，打破技术锁定，实现自由迁移

完全兼容多模态数据领域主流开源技术标准与协议，无需改变用户习惯；提供标准化迁移工具，保障多模态数据完整性与业务连续性，实现 “随时进出” 。

02 企业级增强优化，比开源更优，适配复杂场景

针对开源版本在多模态数据场景的短板定向升级，重构底层存储并智能调度计算资源以提升响应速度；强化安全与稳定性，全面超越开源基础保障，适配复杂业务。

03 生态丰富开放，一站式对接全场景，降低集成成本

深度对接多模态数据全链路主流系统，覆盖采集、存储、处理、应用环节，提供 “一站式” 方案，支持同一平台调用所有服务，大幅降低集成成本。

04 引领技术方向，以核心贡献定义行业未来

作为多模态数据领域核心开源项目的重要贡献者，搭建并运营 Daft、Lance 开源社区，推动技术落地与行业交流，主导 Lance Java SDK 开发与 Spark 生态演进，确保业务技术架构领先。

面向多模态数据处理的新型计算引擎

Daft

原生面向 AI 场景的多模态数据处理引擎

Daft 是基于Rust开发的数据处理引擎，同时提供原生 Python DataFrame&SQL 2 种使用模式，专为 AI 时代多模态数据处理设计。火山引擎AI数据湖服务 LAS 产品提供企业级优化的 Daft 引擎，作为湖计算核心组件处理多模态数据。业内首批参与开源项目，抢占开源生态先机。搭建并运营开源 Daft 中文社区，推动技术落地与交流。

扫码加入社区

Ray

高性能分布式 AI 计算框架

Ray 是一个高性能、开源的通用分布式计算框架，借助简洁的 Python API，能够高效地将计算任务分发至多个节点并行执行，具有简单易用、生态丰富、自动弹性伸缩、资源异构调度，以及与主流 AI 框架无缝集成等特点，为大规模数据处理、模型训练、离在线模型推理等场景提供了高效、灵活且易于扩展的分布式计算解决方案。

Apache Spark

企业级大数据计算引擎

Apache Spark 是基于内存的分布式计算框架，比 Hadoop 快 100倍以上。火山引擎 EMR Spark 在开源版本上实现显著优化，自研 Native Engine（Bolt）性能提升2-5倍。

Apache Flink

实时流处理计算引擎

Apache Flink 是高性能分布式流处理框架，支持毫秒级低延迟和高吞吐量处理。火山引擎在 Flink 领域持续投入核心资源，并经过企业内部多个超大规模 APP 业务场景的长期打磨，沉淀了多行业的丰富实践经验。

面向多模态数据融合的湖存储技术组件

Lance

新一代 AI 原生多模态存储格式

Lance 是专为机器学习和AI场景设计的高性能列式存储格式，相比传统 Parquet 格式提供 100 倍的随机访问速度提升。火山引擎率先引领 Lance 新型湖格式落地，在多模态检索、图文混排等场景成功服务多个客户。

- 研发规模领先：国内当前已知 Lance研发规模最大的团队,具备强劲技术攻坚实力
- 核心成员资深：团队有多名 Lance 多模块 Maintainer,深度参与核心技术维护
- 主导关键成果：主导 Lance Java SDK 开发与 Spark 生态演进,推动技术落地。
- 深度联合开发：与 LanceDB 联合共建 Namespace(Catalog) 功能,协同完善生态。
- 持续技术迭代：多个 Lance 内核级特性处于在研阶段,持续夯实技术壁垒。

扫码加入社区

Apache Iceberg

企业级数据湖表格式

Apache Iceberg 是开源 Lakehouse 表格式的标杆，提供 ACID 事务、Schema 演进、时间旅行等企业级特性。火山引擎EMR团队深度优化 Iceberg 组件，在开放兼容的基础之上提供了性能更强、成本更低的文件合并和清理等湖管理能力。

Apache Paimon

实时流式数据湖

Apache Paimon 创新性结合湖格式和 LSM 树结构，将实时流式更新引入数据湖架构。火山引擎是 Paimon 社区核心贡献者，拥有 PMC 成员和最大规模生产实践之一。

火山引擎 | 多模态数据湖



官方社群

更多火山引擎数据产品与
解决方案扫码了解

