AI时代的数据挑战与开源创新机遇

堵俊平

Datastrato

CONTENTS

01

Scaling Law引爆数据瓶颈

Scaling Law成就AI新纪元

过去几年，**AI的核心**规律是Scaling Law

从GPT、Gemini到DeepSeek，顶尖模型验证了“规模×数据×算力”的持续幂律（**Continuous Power Law**）

只要三者同步跃升，模型性能即不断提升

Scaling Law： 性能随三要素同步跃升

AI的核心瓶颈，正在从“算力”转向“数据”

数据跃升为核心瓶颈

当模型与算力持续扩张，数据能力成为新的天花板

AI能力的决定因素

你能喂给模型什么样的数据、多少数据、多快的数据、多实时的数据，将直接决定一个组织能否建立自己的AI能力

数据质量与规模
高质量、大规模的数据集是模型性能的基础

更新与实时性
数据的更新速度和实时获取能力成为关键

战略优先级
数据管线必须从支持角色升级为第一优先级

02

生成式到Agentic质变

生成式AI的数据范式

Generative AI依赖**大规模静态离线语料，其数据需求相**对单一

大规模、静态、离线

更关注“数据质量”**与**“语料多样性”，**更新**频率相对较低

典型应用场景

RAG，**模型推理**，Prompt Engineering

数据模态与治理

以单一文本模态为主，治理重点在**版**权、去重与标注

Agentic AI的数据跃迁

Agentic AI对数据提出了“**从量变到质变**”**的全新要求**

数据实时

实时特征、日志、反馈，秒/毫秒级更新。

数据多模态

文本、图像、视频、时序、空间等

数据分布式

多云、多存储、跨平台

数据动态可操作

Agent需写数据、触发流

数据统一可治理

可追踪、可管控、可授权

本质差异

Generative AI

需要“静态知识库”

重存储与只读

Agentic AI

需要“实时可交互的数据操作系统”

重实时、读写与治理

企业若继续沿用旧管线，将陷入**数据找不到、管不好、用不起来**的困境

03

多云多模治理困境

挑战1：多云+多引擎碎片化

企业数据散落在各处，缺乏真正统一的平台，导致效率低下

- **多云环境**：AWS / GCP / Azure
- **多引擎**：Lakehouse、Warehouse、Streaming、KV/Vector Store
- **多格式**：Iceberg / Delta / Hudi / Parquet / 各类API

挑战2：多模态管理难度爆炸

AI时代，80%的数据是非结构化或半结构化，传统数据系统已不堪重负

传统数据系统

仅为结构化数据设计

AI多模态数据

视频、语音、Embedding、RLHF信号等

缺乏向量索引、版本管理能力，导致存储与带宽成本指数级上升

挑战3：治理与权限碎片化

每个系统都有自己的一套规则，导致企业不得不重复实施治理，审计线索分散，合规风险高

权限模型
IAM, RBAC, ACL 各不相同

Catalog & Metadata
Schema不统一，元数据割裂

API/SDK
接口各异，增加开发成本

挑战4：实时化与成本双重压力

Agentic AI强调实时状态，但传统架构面临技术与成本的双重挑战

实时化挑战

传统数据湖偏向批处理，构建秒/毫秒级流式链路复杂，难以满足Agent对实时特征与日志的需求

成本挑战

多模态数据副本激增，存储与计算费用快速膨胀，数据爆炸式增长带来的成本压力成为新瓶颈

04

开源项目破局之道

以更加开放与专注，来推动极致创新

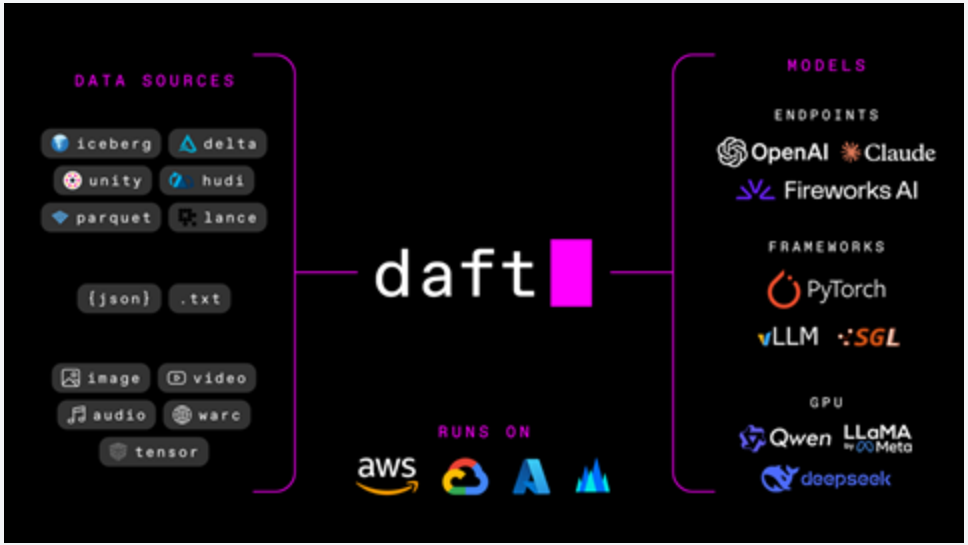
开源新机遇：LanceDB

统一多模态向量与数据格式

LanceDB通过引入Lance列存格式，将结构化与非结构化数据打包成可治理的统一对象

- 解决AI数据存储混乱问题，避免碎片文件
- 提供高性能列存，兼容向量搜索，适合RAG
- 支持高效采样与版本回退，降低存储成本

https://github.com/lancedb/lancedb



开源新机遇：Daft

AI Native的分布式Data Frame引擎

Daft是一个为AI工作流设计的分布式DataFrame引擎，让ETL真正面向多模态

- 支持视频帧、音频切片等非结构化数据处理
- 云原生执行模型，按需弹性，比Spark更轻量
- 适用于训练数据准备和RAG数据清洗，缩短准备时间

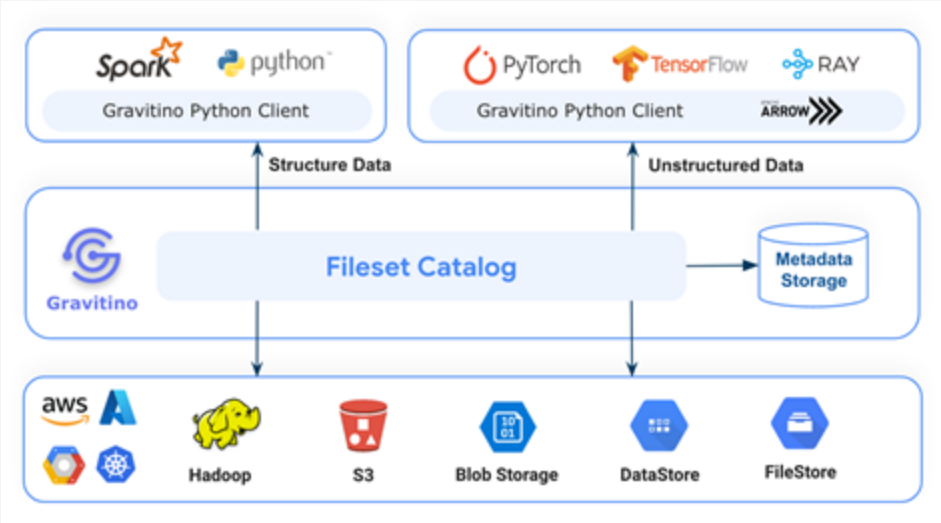
https://github.com/Eventual-Inc/Daft

开源新机遇：Apache Gravitino

Agentic AI的“统一数据大脑”

Gravitino不是存储，而是AI时代“数据资源的操作系统”

- **聚合各**类数据源Catalog**与**Metadata, 让AI**系**统像访问同一个Data Lake**一**样访问所有数据
- **解决了多云**IAM、ACL、RBAC**碎片化的**难题，让Agent **通**过MCP**即可自主**发现、授权、读写任意**来源的数据**



https://github.com/apache/gravitino

Gravitino三大核心价值

作为AI数据宇宙的智能“路由器 + 控制平面”，Gravitino提供三大核心能力

统一元数据

聚合Structured + Unstructured + Streaming + Model Metadata

统一访问层

基于REST/MCP，让Spark, Flink, Trino, AI Agent共享Catalog

统一治理

解决多云IAM, ACL, RBAC碎片化问题，实现集中管控

AI Native 开源“数据三件套”

Traditional Data Engine	Spark, Flink, ...	Multimodal Data Engine	Daft
Traditional Data Catalog	HIVE, APACHE POLARIS, ...	Multimodal Data Catalog	APACHE GRAVITINO
Traditional Table Format	ICEBERG, Apache Hudi, ...	Multimodal Data Format	Lance

05

拥抱开源赢未来

开源栈成企业基石

AI竞争已进入数据能力瓶颈期，开源方案以其独特优势成为企业构建AI**能力的最佳**选择

社区速度：
以社区速度迭代，快速跟上多模态与多云演进

避免锁定：
避免厂商锁定，降低试错成本

聚焦创新：
让企业专注于业务创新，而非底层基建

立即行动，构建AI数据能力

拥抱开源，构建面向未来的AI-Native数据平台

为下一代Agentic AI应用奠定**可持续增长的基础**

THANK YOU