

QCon
全球软件开发大会

元数据驱动：构建下一代智能数据架构的探索与实践

邵赛赛

Datastrato联合创始及CTO



About me

- Co-founder & CTO of Datastrato.
- The original creator of Apache Gravitino.
- The committer and PMC member of Apache Spark.
- Apache Member.



📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播

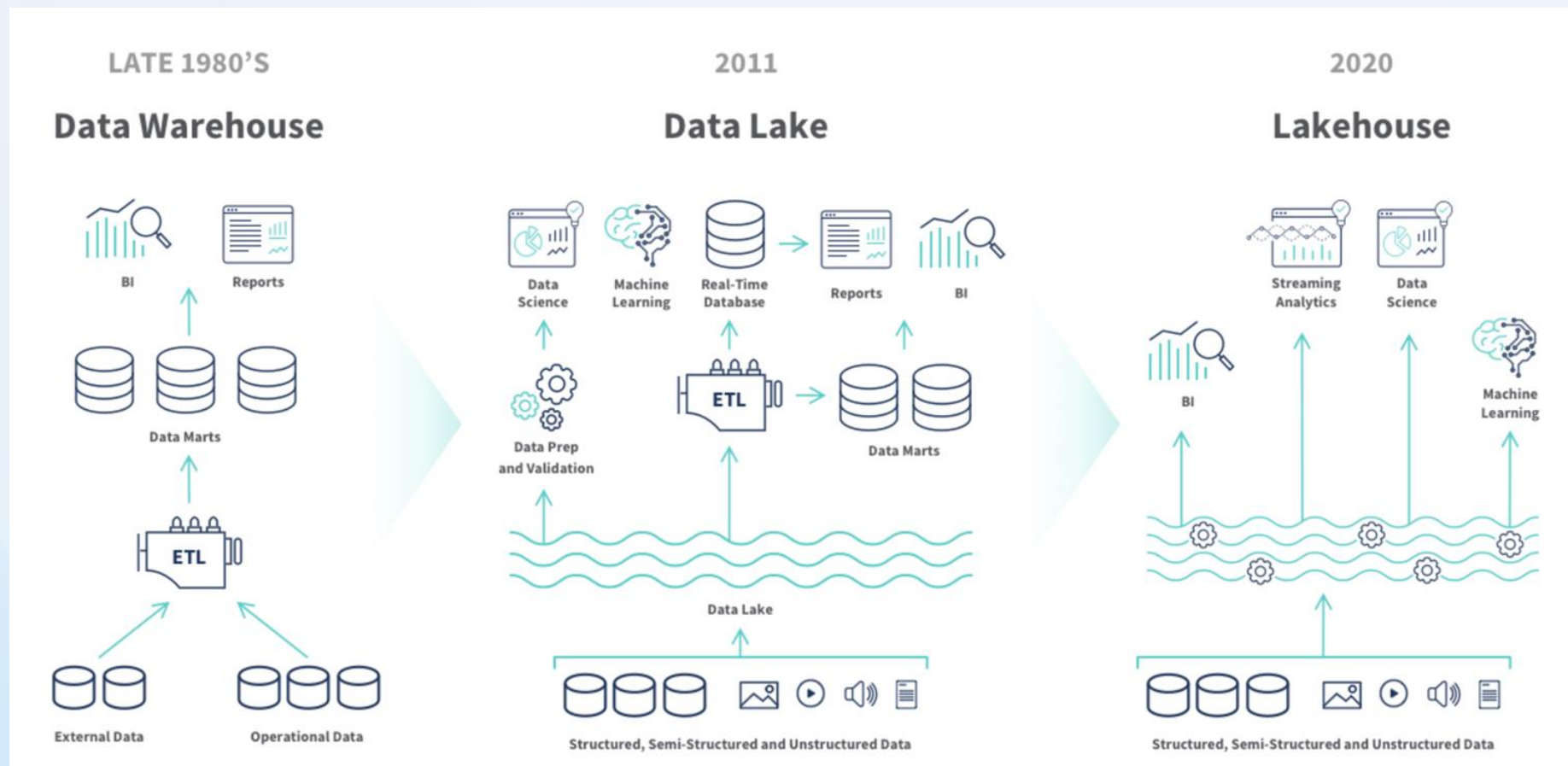


参会咨询



查看会议

数据架构的历史演变

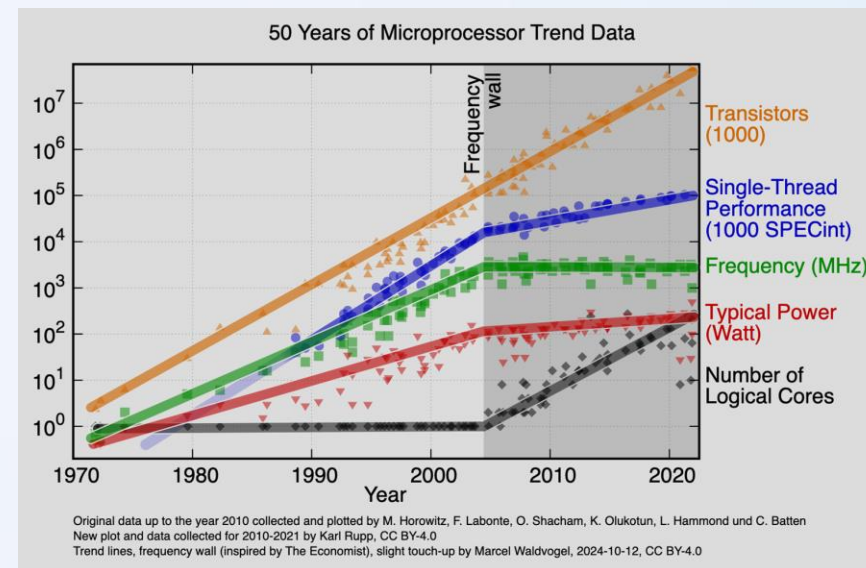
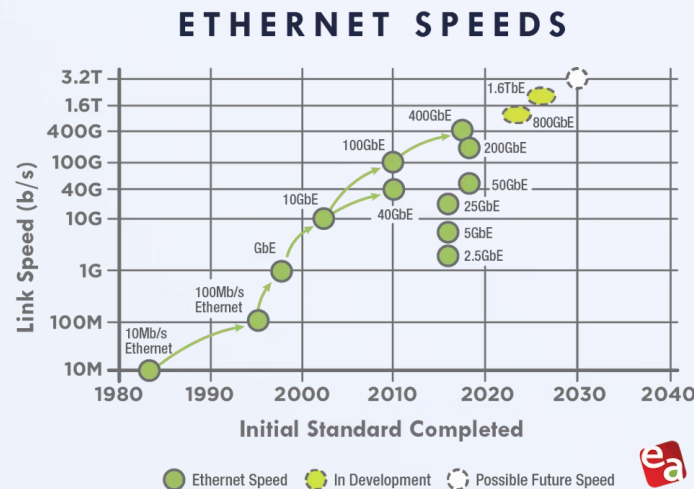
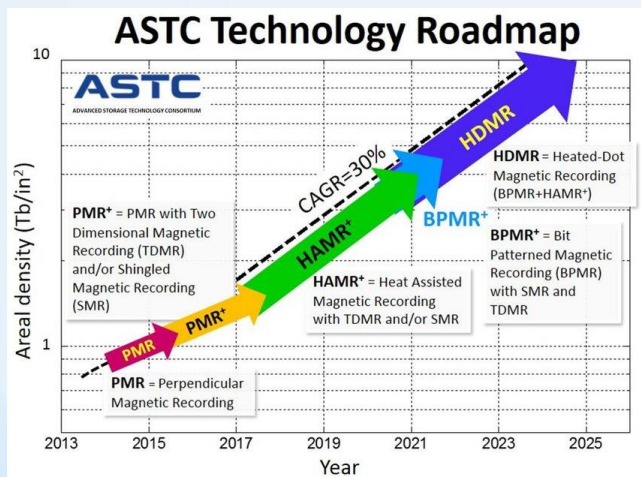


Goal

3Vs - Volume, Velocity, and Variety.

现有数据架构的问题和挑战

● 现状 1 – 追求数据架构3V的回报越来越低



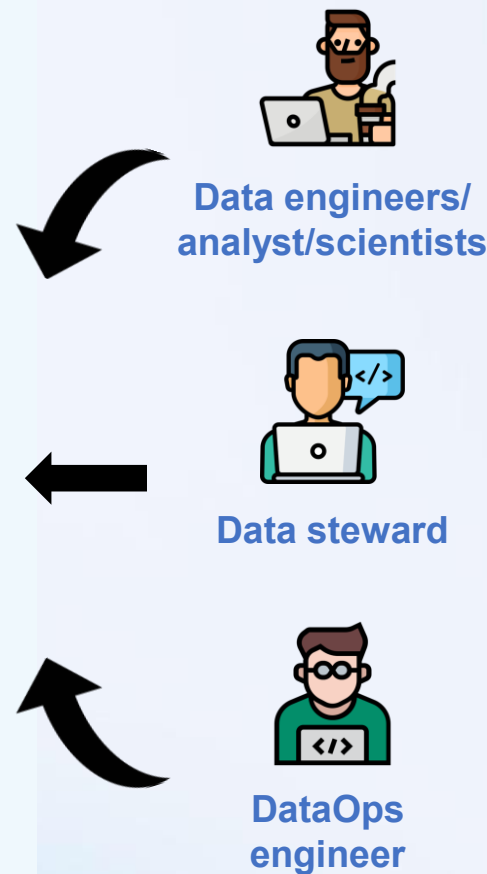
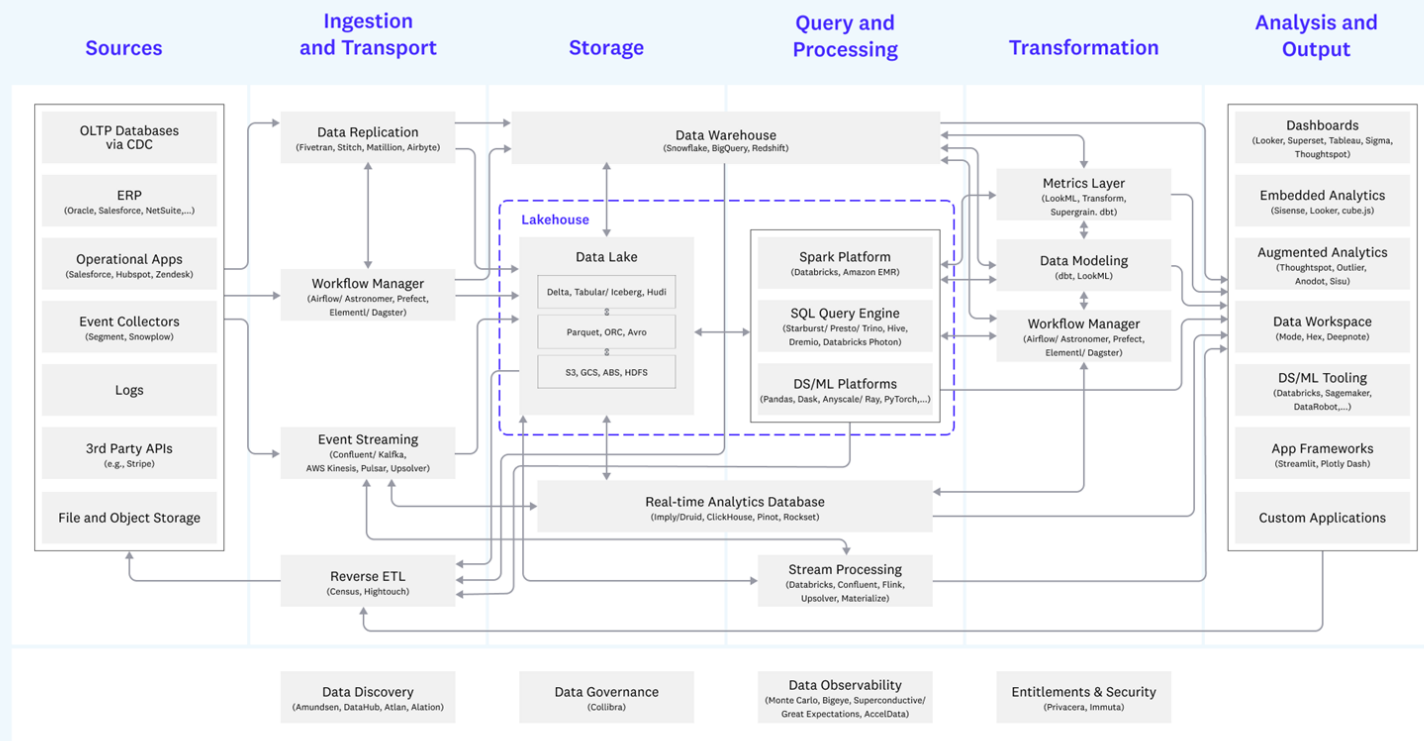
存储密度、网络带宽的快速增长，构建大规模的高密度存储架构变得更为简单

随着摩尔定律的终结，榨取CPU的性能价值变得越来越难

- https://www.researchgate.net/figure/Advanced-Storage-Technology-Consortium-ASTC-roadmap-for-the-future-hard-disk-drive_fig3_310594004
- <https://www.lightwaveonline.com/directory/components/optical-switches/article/14301154/the-world-runs-on-ethernet-the-future-of-higher-speeds>
- <https://netfuture.ch/2024/10/50-years-of-microprocessor-trend-data/>

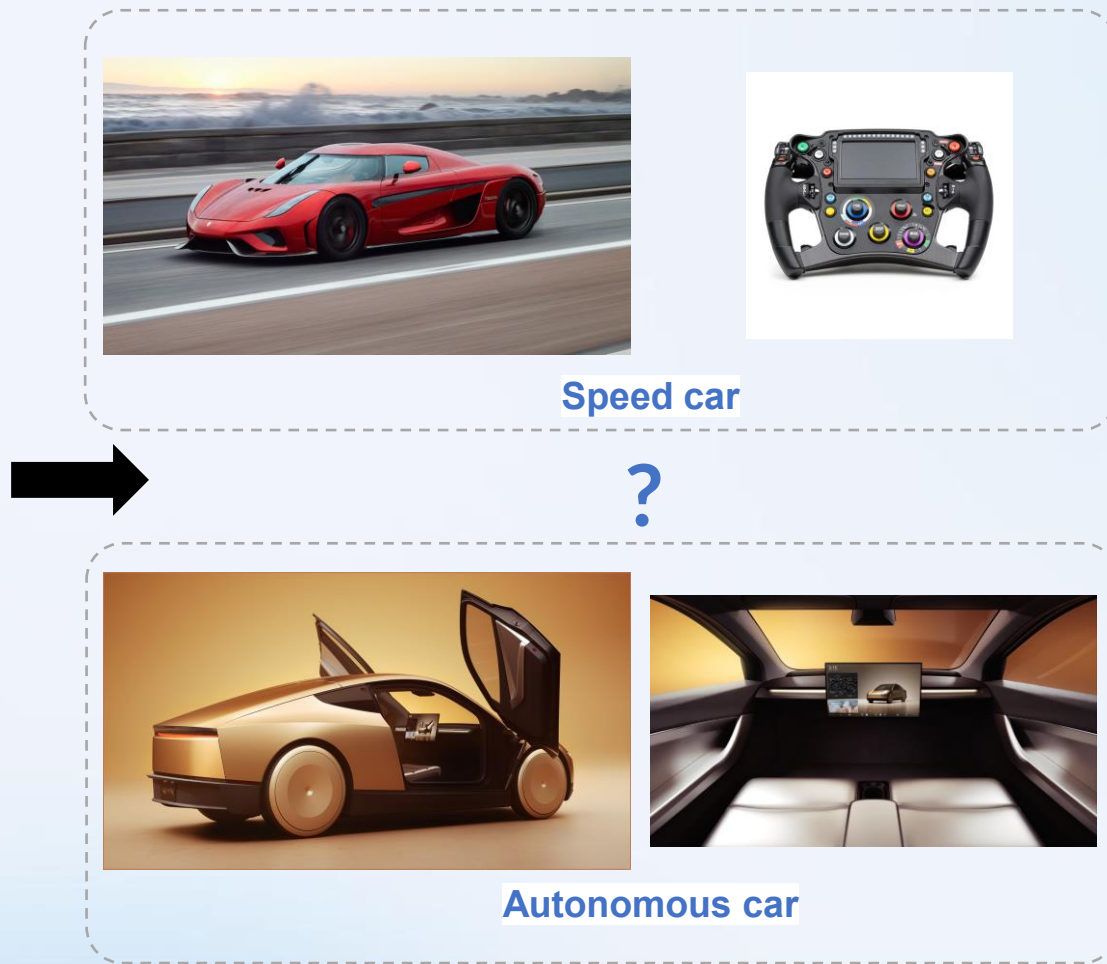
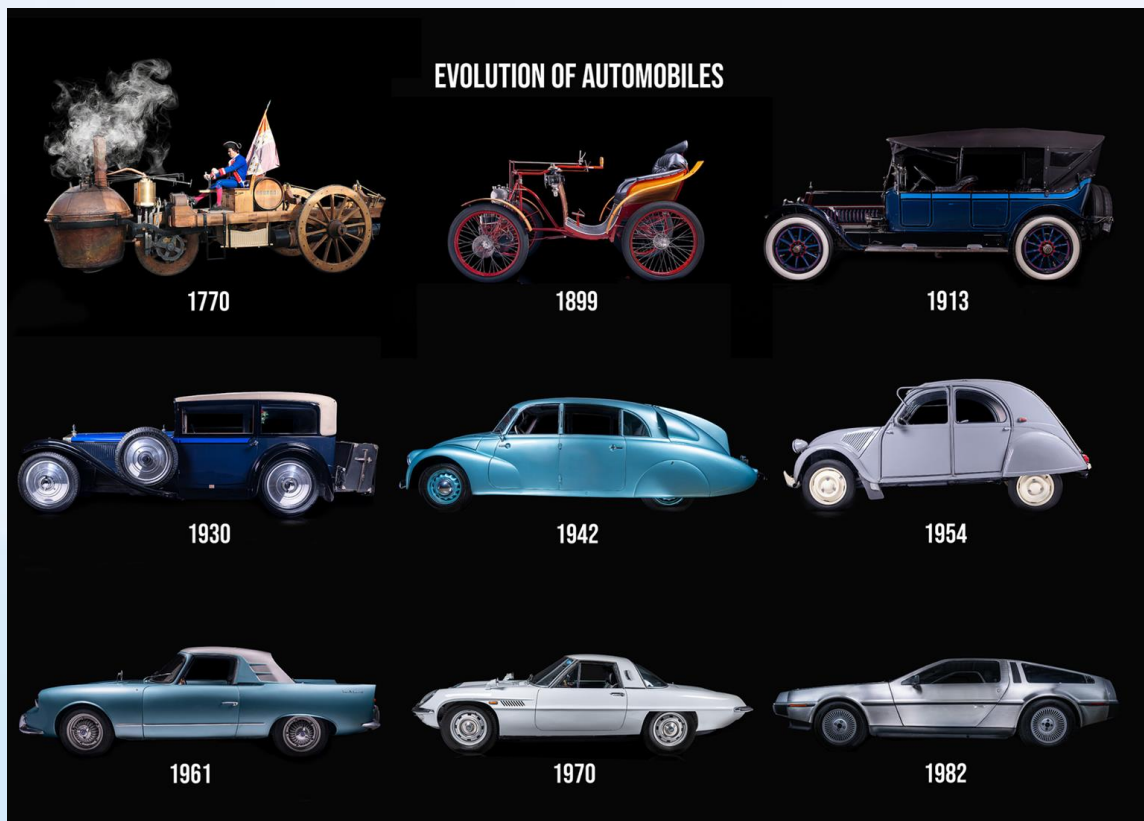
● 现状 2 – 使用和管理数据栈的复杂度还在持续上升

Unified Data Infrastructure (2.0)

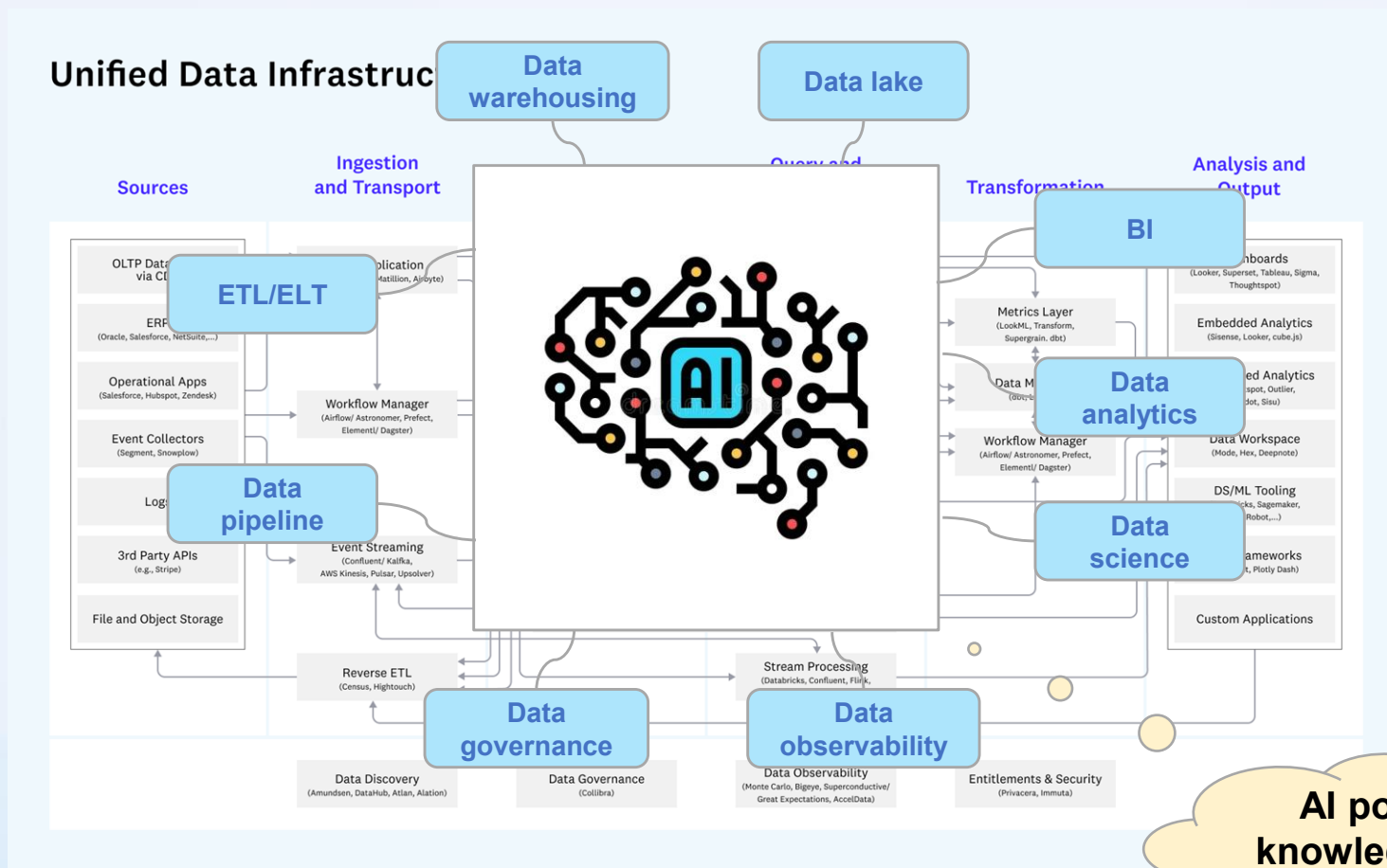


● 现状 3 – 迈向智能化是事物发展的基本规律

以汽车工业举例



围绕AI构建的下一代数据架构



在AGI时代如何重塑我们的数据架构

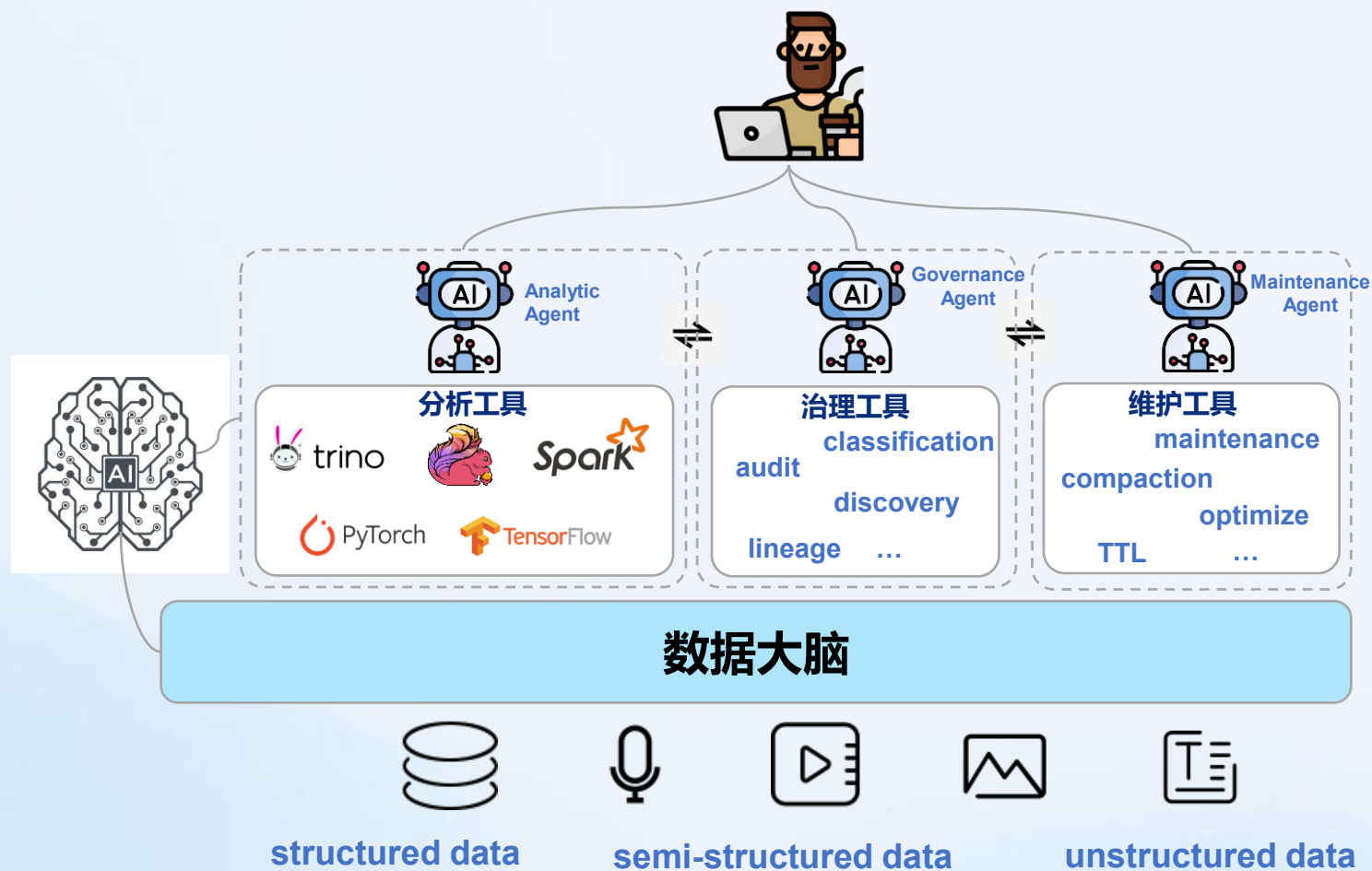
使用元数据构建数据架构的大脑

元数据是“关于数据的数据”，即描述、解释或为其他数据提供上下文的结构化信息。

- **上下文理解与发现：**元数据充当“数据字典”，描述了数据集的内容、结构和内部关系，帮助用户理解和发现数据资源。
- **协调与治理：**元数据在不同系统与团队之间架起了沟通的桥梁，是实现数据有效协调与治理的基础。通过记录数据的来源、格式、访问权限等管理性信息和技术细节，元数据确保了数据在流转过程中的一致性和可控性。
- **决策支持与自动化：**元数据为智能自动化提供了关键指引，它能够指导工具如何正确地处理数据，从而支撑高效的决策和自动化流程。
- **组织与管理：**除了上述作用，元数据还通过提供分类、索引和权限等信息，帮助对大量数据进行高效的组织和管理，确保数据的可靠性和可用性。

使用元数据和大模型结合构建出数据架构的大脑

使用MCP连接器构建数据栈的左膀右臂



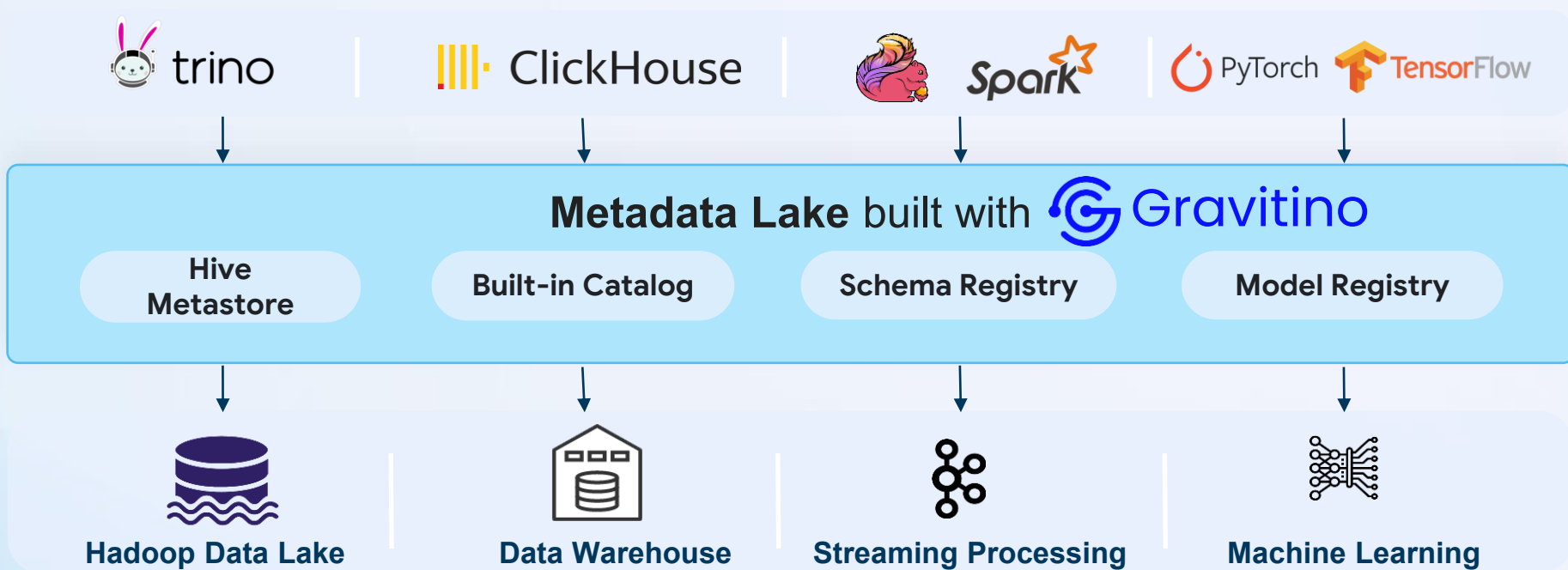
架构优势:

- 依托元数据及大模型构建的数据大脑进行任务决策。
- 将不同的工具栈MCP化, 使得工具栈可以使用数据大脑驱动。

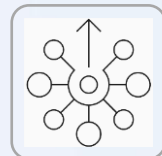
如何使用Apache Gravitino构建智能数据架构

什么是Apache Gravitino™

Apache Gravitino - Catalog of Catalogs (Metadata Lake)



Key features:



SSOT



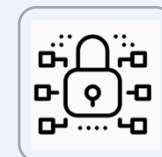
AI + Data Catalog



Geo-distributed Arch

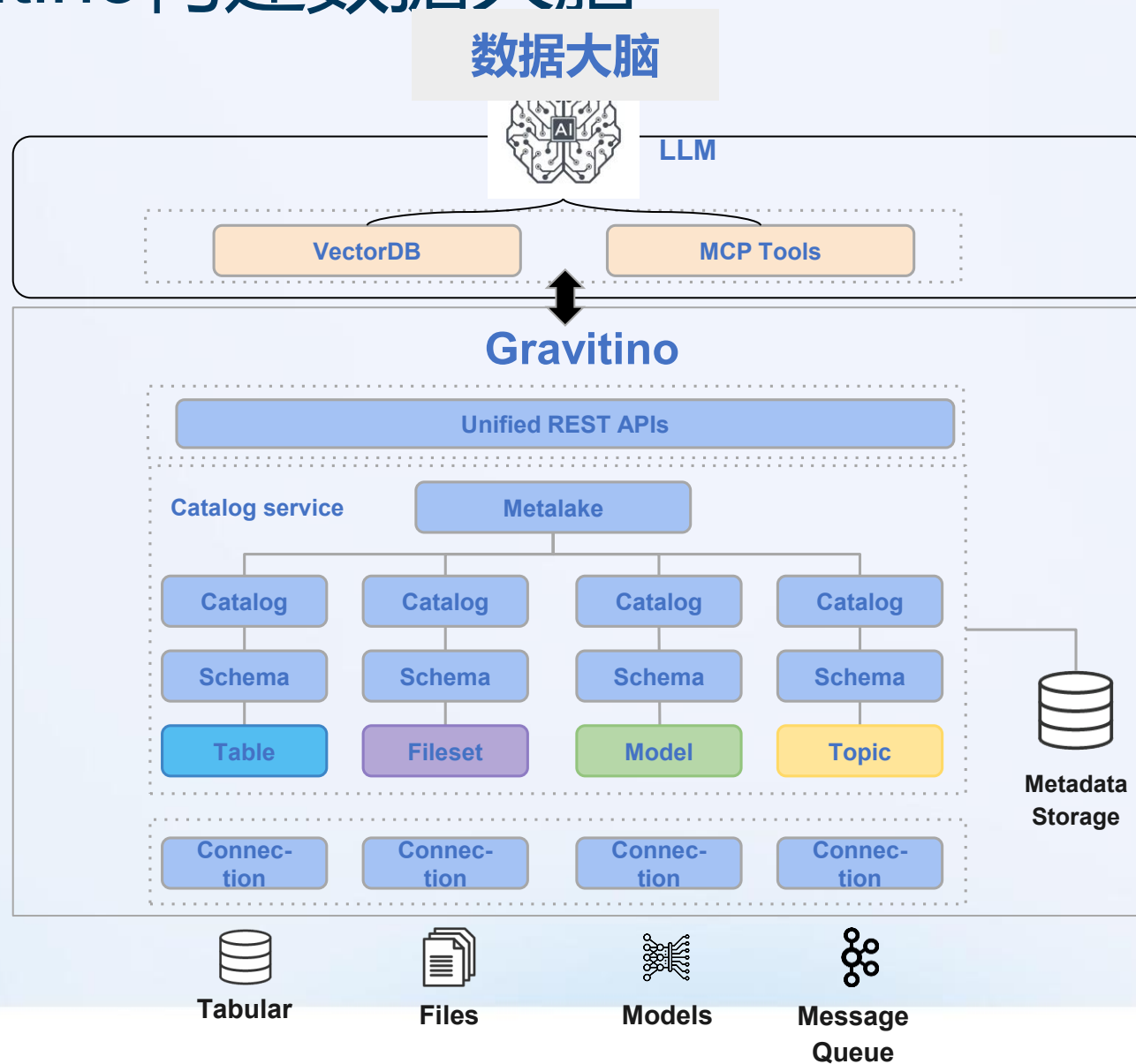


Catalog + Governance

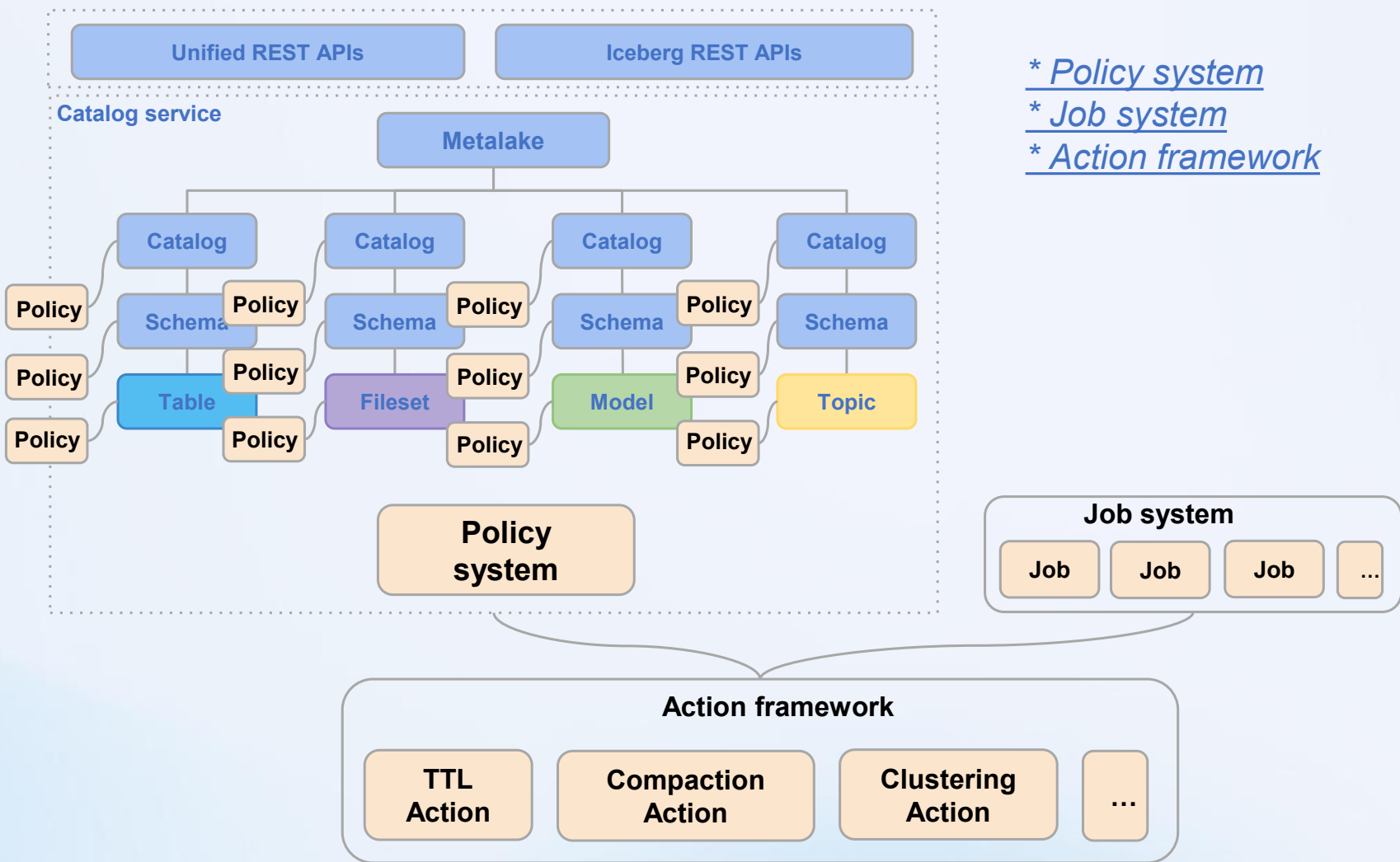


Security in One place

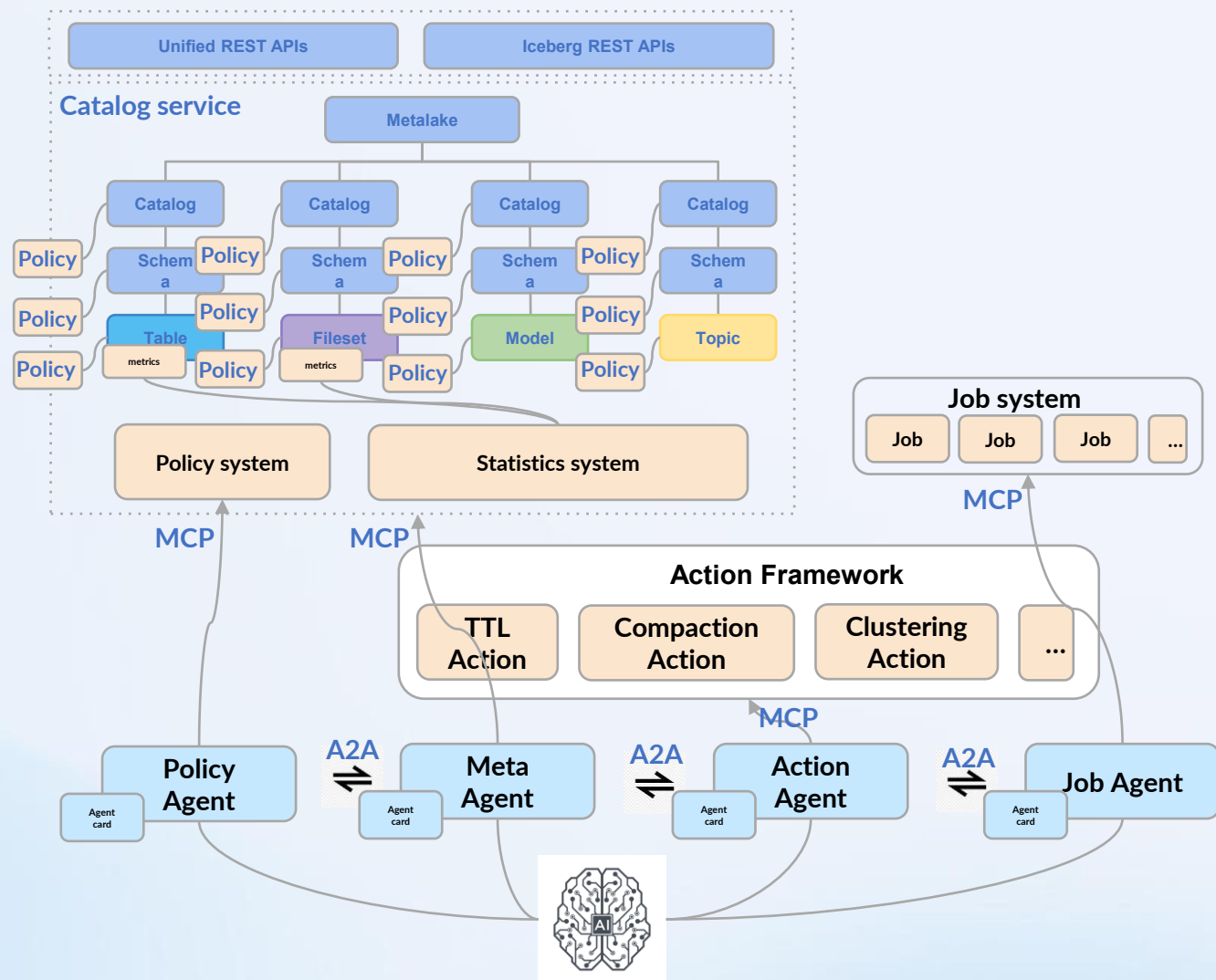
使用Gravitino构建数据大脑



使用Gravitino构建数据执行器



串联整个架构



* *Meta agent* - data discovery and context understanding

* *Policy agent* - check and build policies on metadata

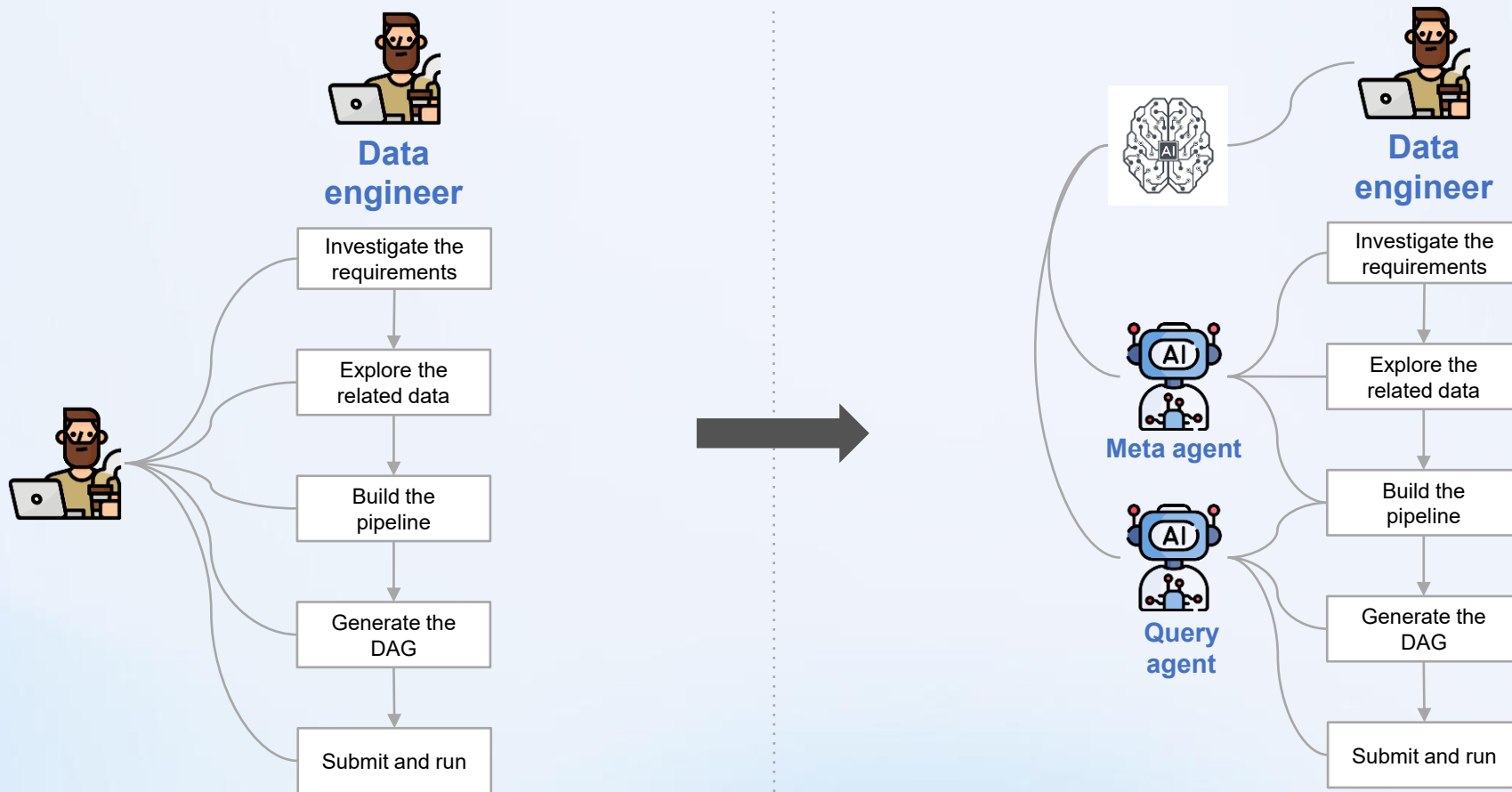
* *Action agent* - generate a job for the specific action

* *Job agent* - submit a job to the job system

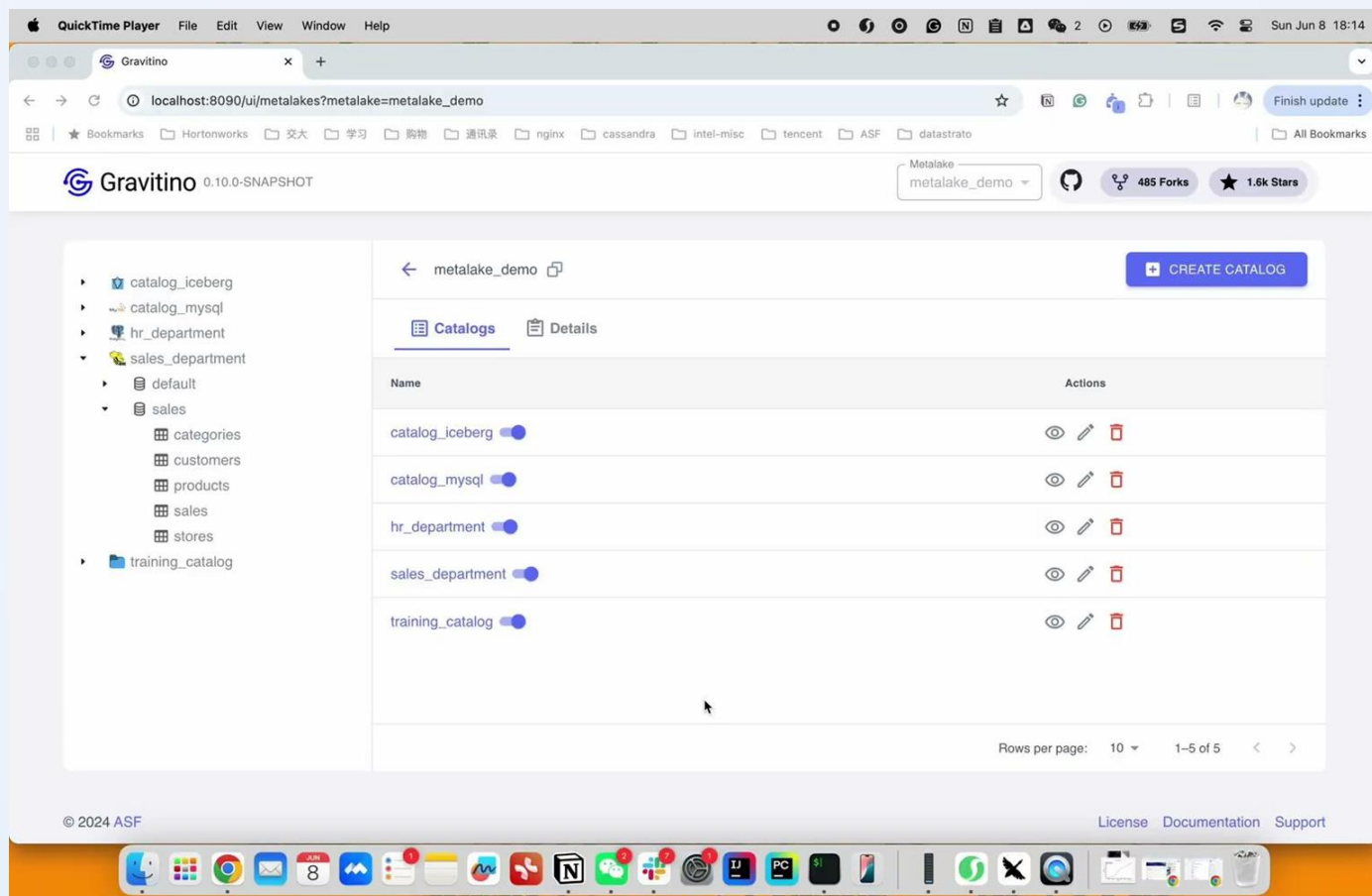
Demo

● 使用Gravitino构建智能数据架构

Building data pipelines using agents (Data engineering)

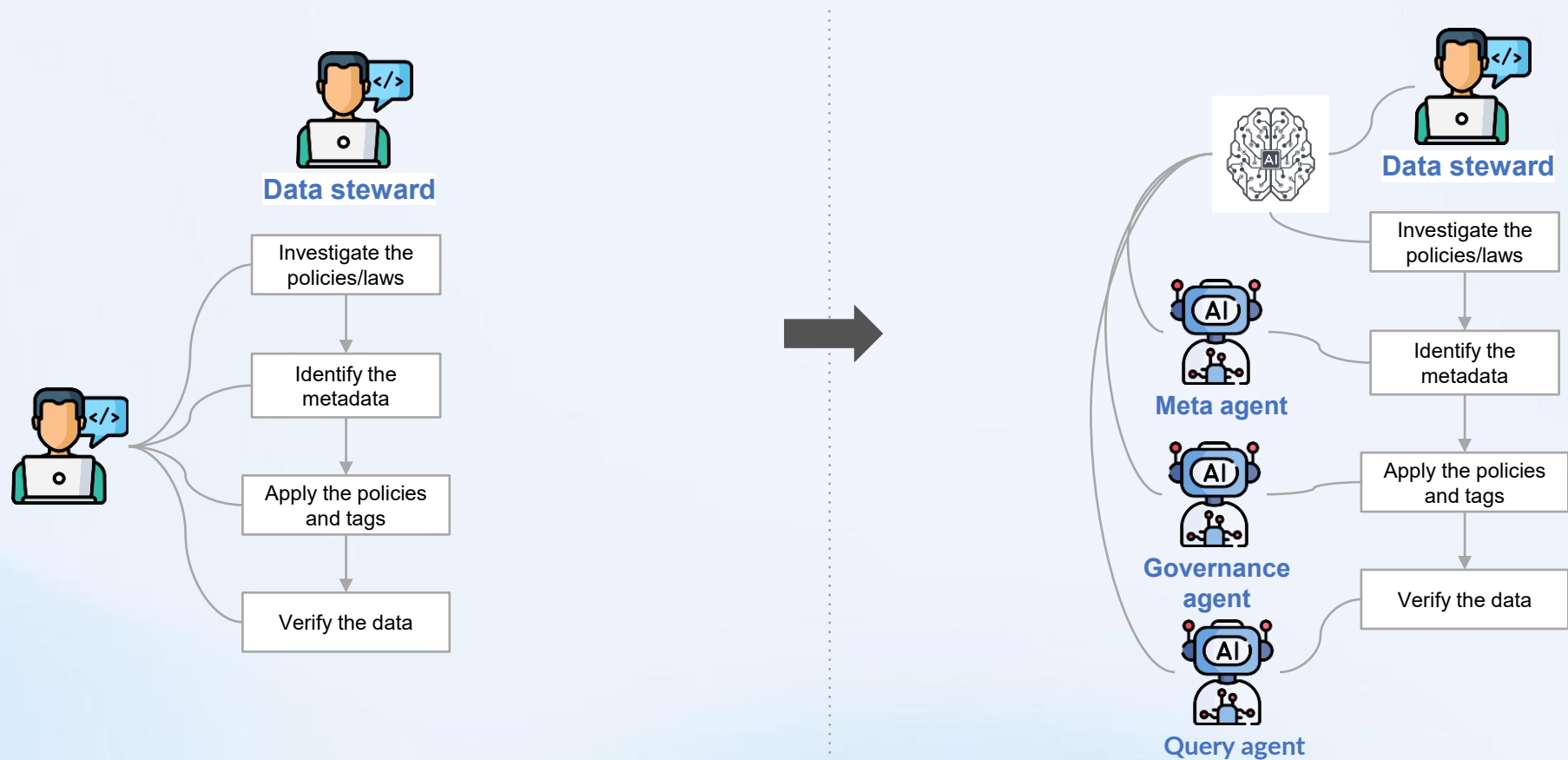


Data Engineering

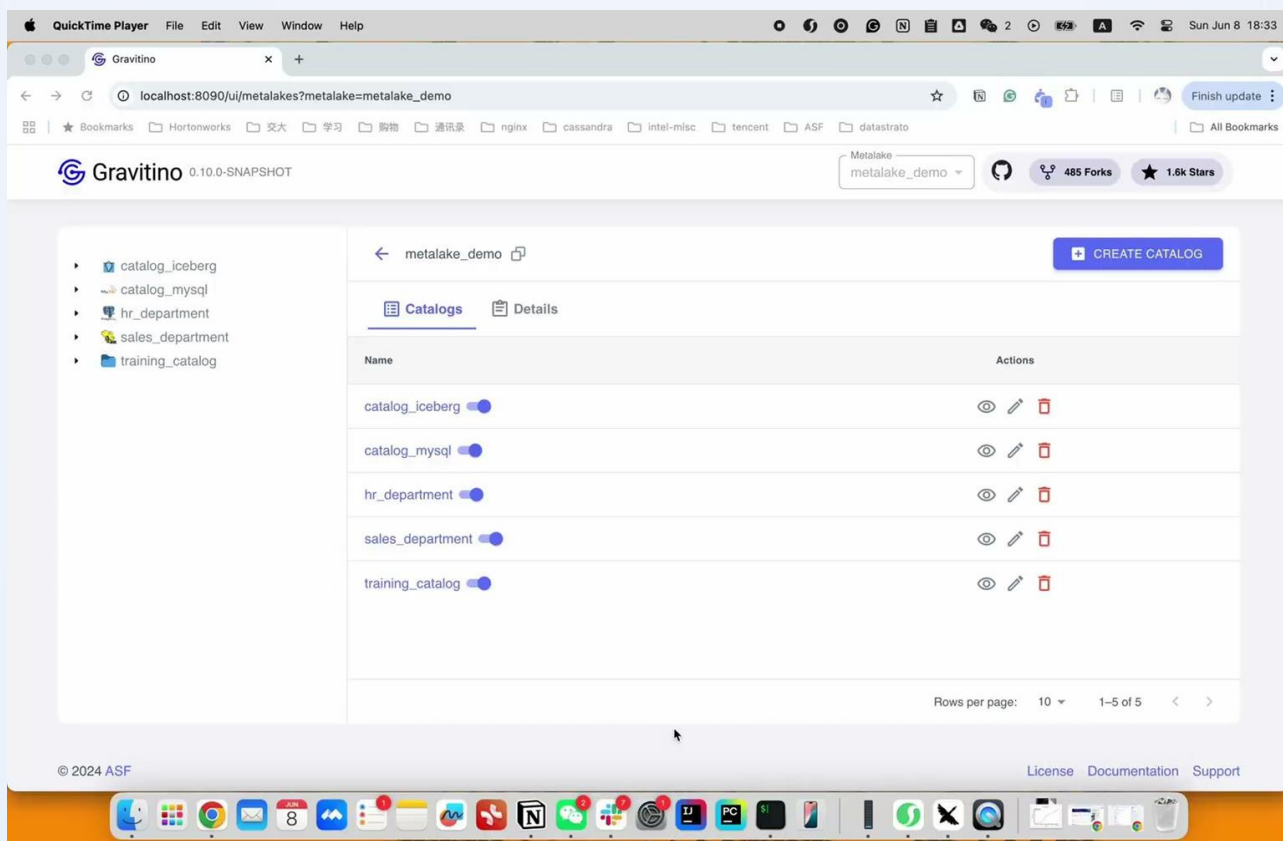


● 使用Gravitino构建智能数据架构

Data classification and labeling (Data governance)



Data governance



总结

总结和展望

- 数据架构向智能化迈进是一个不可逆的趋势
- 元数据在构建智能化数据架构中起到了不可替代的作用
- 结合大模型构建数据大脑，并由数据大脑驱动数据工作大幅解放数据工作

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

QCon
全球软件开发大会

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

