亚马逊云科技

中国峰会

2026年6月23日-24日 上海 · 世博中心

亚马逊云科技
中国峰会

高性能存储加速生成式 AI

Accelerating Generative AI with High-Performance Storage

Christian Smith

Amazon Web Services WW Specialist Storage

数据是生成式 AI 的关键差异化优势
Your data is the differentiator for Generative AI

存储策略对生成式 AI 至关重要

- 生成式 AI 从非结构化数据中释放新的业务价值
- 您的数据才是关键差异化优势 ——而非模型
- 无论是构建、定制还是训练模型，数据可访问性决定了结果
- 高质量数据的存储策略至关重要
- S3 是数据存储的基石：全球百万客户的数据湖构建在 S3 上

生成式 AI 生命周期：存储的关键作用
Generative AI Lifecycle: Where Storage Matters

01

训练

利用海量数据集训练基础模型（万亿级 token，PB 级数据，数千个 GPU 实例）

02

微调 / 优化

利用行业数据调整预训练模型以提升模型性能（LoRA、RLHF、全量微调）

03

智能体

使智能体能够从领域特定数据中提取洞察，从而简化复杂任务

面向生成式 AI 的存储
Storage for Generative AI

S3

Amazon S3 / S3 Express One Zone

适用于大规模数据湖和模型训练的可扩展对象存储

S3 Express One Zone 提供
个位数毫秒级延迟
大规模人工智能场景（>1000 GPU）

- 无限容量
- 10 倍速度提升（Express）
- >1000 GPU 规模

FSx

Amazon FSx for Lustre

完全托管的并行文件系统
POSIX 语义兼容

亚毫秒级延迟
高达 1 TB/s 的聚合吞吐
适用于大规模分布式训练

- 1 TB/s 吞吐量
- 亚毫秒级延迟
- POSIX 兼容

S3+

S3 Vectors / S3 Files / Mountpoint

S3 Vectors：
适用于 RAG 工作负载的向量存储

S3 Files：
通过 NFS/POSIX 访问 S3 数据

Mountpoint：
为 ML 框架提供文件访问

- 向量搜索
- NFS 访问
- 框架原生

亚马逊云科技
中国峰会

分布式模型训练和调优的存储挑战
Distributed Training: Storage Challenges

为什么存储性能对模型训练调优至关重要
Why Storage Matters

模型调优在您的领域特定数据上重新训练基础模型，存储性能决定了您昂贵的 GPU 是保持忙碌还是闲置

数据吞吐必须匹配 GPU 计算能力， 如果存储无法足够快地向 GPU 提供数据，您就在为闲置的 GPU 资源买单

错误的存储选择 = 闲置的 GPU = 浪费的支出

$32/小时

GPU 闲置小时的成本

#1

存储性能瓶颈是模型训练中的主要挑战

Amazon S3

大规模数据场景

最佳适用场景

分布式模型训练/调优， 1-16 个实例的集群规模

能力

- 流式吞吐量
- 几乎无限的容量
- 按用量付费
- 与 SageMaker 和 PyTorch 原生集成
- 适用于：LoRA/QLoRA、模型调优
- 与 S3 Files 和 S3 Mountpoints 集成

挑战

小文件工作负载（亿级小文件）
频繁的检查点保存和恢复

FSx for Lustre

大规模训练场景

最佳适用场景

分布式模型训练/调优，16-2000 个实例的集群

能力

- 跨数千个客户端的并行 I/O
- 亚毫秒级延迟，保持 GPU 满负荷运行
- 扩展至 1+ TB/s 聚合吞吐量
- POSIX 兼容（PyTorch、DeepSpeed、Megatron）
- 与 S3 集成, 支持从 S3 延迟加载，无需手动管理数据复制和同步
- LZ4 压缩 + 原生 S3 检查点
- EFA+ NVIDIA GPU Direct Storage 支持
- 适用于：多节点大规模分布式训练

挑战

数十亿小文件（元数据服务器）场景，需要针对性的部署和规划

S3 Express One Zone

超大规模训练

最佳适用场景

超大规模场景—数万个实例
适合于云原生且偏好对象存储接口的数据科学团队

能力

- 低延迟, 比标准 S3 低 10 倍以上
- 大规模并发下的一致高吞吐
- 数据并行访问，线性扩展
- 为机器学习优化的存储桶
- 适用于：超大规模集群，1000+ 或更多的 GPU 集群

挑战

需要使用 S3 API
与标准 S3 存储桶没有直接的数据集成，需要手动管理数据迁移

通用建议：从 FSx for Lustre 开始
Customers typically start with FSx for Lustre

- 简单易用：工具和数据加载器与文件系统原生集成
- 高性能：吞吐量随文件系统大小线性扩展
- 存储选择：持久化 vs. 临时: 按照工作负载生命周期来选择持久化存储或者临时存储
- 建议：从吞吐量需求开始规划，而非容量
- 成本优化：使用 Intelligent-Tiering 优化成本而不牺牲性能

Adobe 仅用九个月就发布了 Firefly 生成式 AI 模型

利用 Amazon FSx for Lustre 加速大规模模型训练

— Adobe Creative Cloud & AI

决策框架 Decision Framework

为训练选择合适的存储

< 16 个实例

就地使用您的数据

S3、EFS、EBS、FSxZ、FSxN 等 —— 无需移动数据

16 - 1,000 个实例

16-500：FSx for Lustre（除非团队或应用程序偏好对象存储）

500-1,000：FSx for Lustre 或 S3 Express OneZone

需要频繁检查点或 POSIX? → FSx for Lustre | 否则 → 团队对文件 vs 对象的偏好

> 1,000个实例

S3 Express OneZone

在此规模下，经济性通常超越其他考量因素

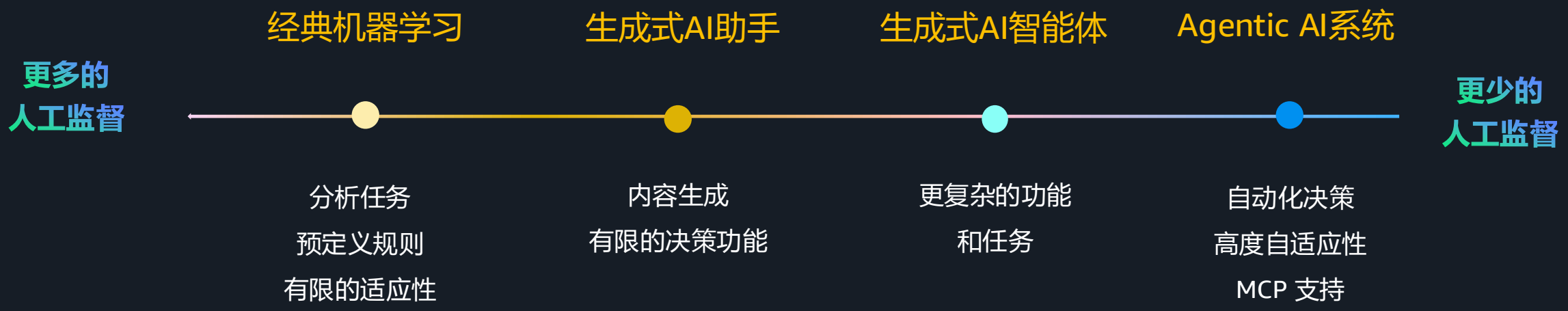
亚马逊云科技
中国峰会

面向 Agentic 工作负载的存储服务

Storage for Agentic Workloads

AI 通过 Agent 变得更加自动化

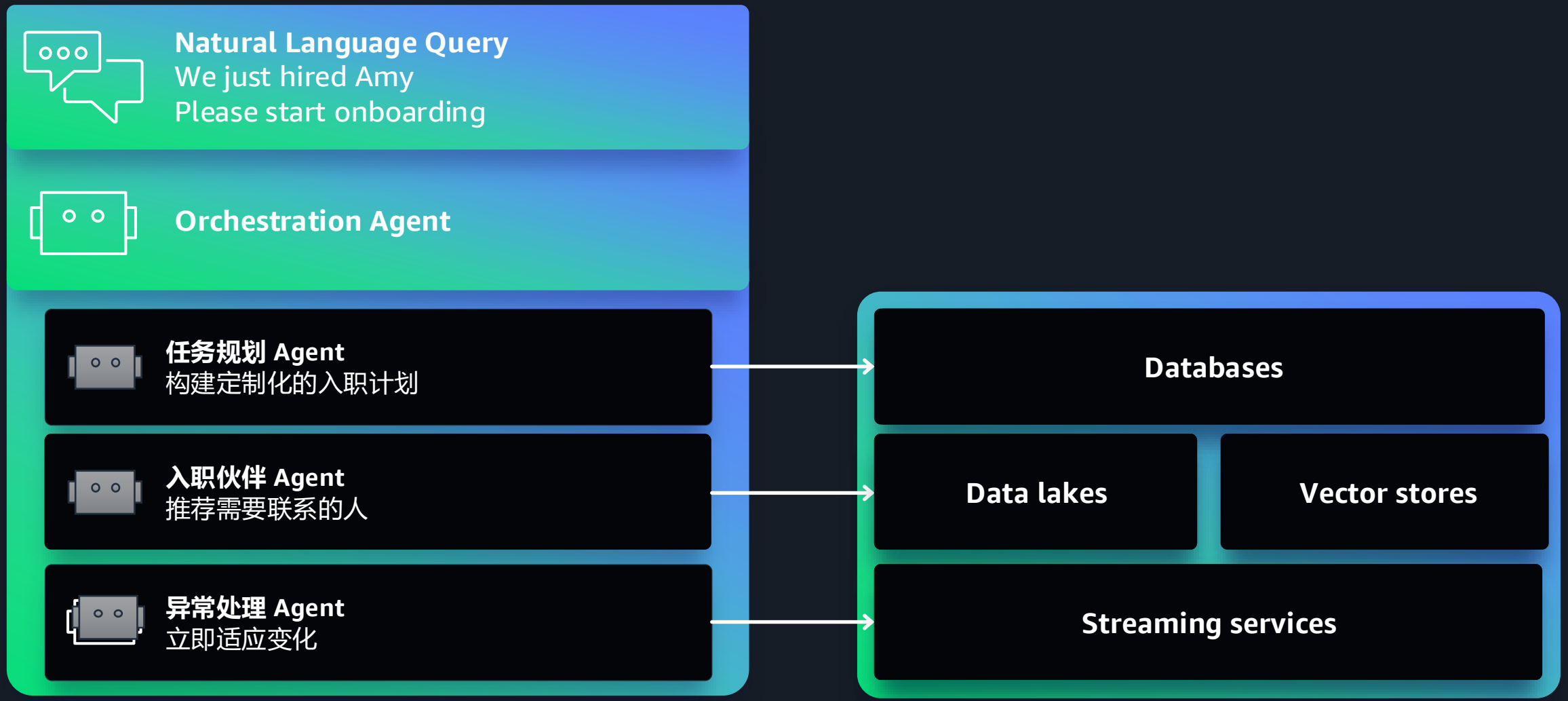
AI gets more autonomous with agents



*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

Agents 需要多种类型的数据

数据如何支持 Agentic AI 应用



Agent 记忆组件

短期 — 当前任务或当前会话（线程）

长期 — 当前用户或持久化的应用级别

Agent state	Messages	Semantic	Profiles	Episodic	Prompts
Action plan, data shared between loops, scratchpad	Conversation history, sequence of messages	Semantic context relevant for the requested task	User profiles and similar data	Historical interactions and outcomes (RL)	System prompt, instructions

通常使用
依赖 Agentic 框架
的后端数据存储

构建者缺乏控制

通常**使用工具**在向量
数据存储库和企业数据库中
进行检索

数据可访问性和质量

根据应用代码、
文件或键值存储
进行版本控制

提示词调参

不只是记住，还要智能！
Agentic memory is key

情境智能

通过情境理解和模式识别提升响应的准确性和相关性

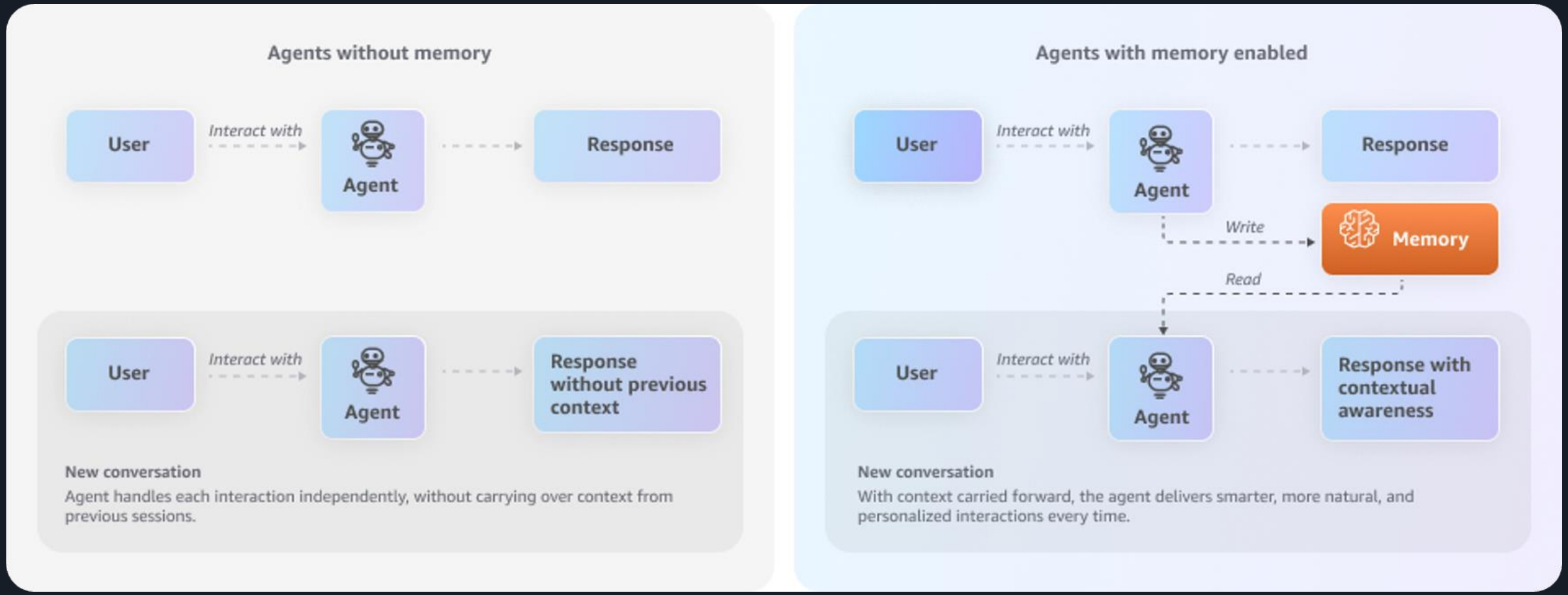
用户偏好

通过记住跨多个会话的偏好和历史对话实现个性化交互

知识留存

通过持久记忆和持续学习能力实现复杂的问题解决

两种场景：有记忆 vs. 无记忆



S3 Files像传统文件系统一样工作
Works like a traditional file system

文件和目录反映 S3 桶中的内容

- 从任何桶创建文件系统，无需数据迁移
- 智能地将数据存储在文件系统中，实现毫秒级延迟访问
- 文件系统更改立即对文件应用程序可见
- 更改在桶和文件系统之间自动同步

Amazon S3 Vectors

核心能力

- 原生支持向量的云对象存储
- 向量上传、存储和查询成本降低高达 90%
- 零基础设施管理
- 支持数十亿规模，并具备元数据过滤能力
- 与 Amazon S3 相同的持久性和可用性

典型适用场景

- 数据湖上的语义搜索
- 批量检索 pipelines
- 与向量数据库形成互补架构（冷热分层）
- 对成本敏感的大规模向量存储

延迟（P95）

>100ms

向量规模

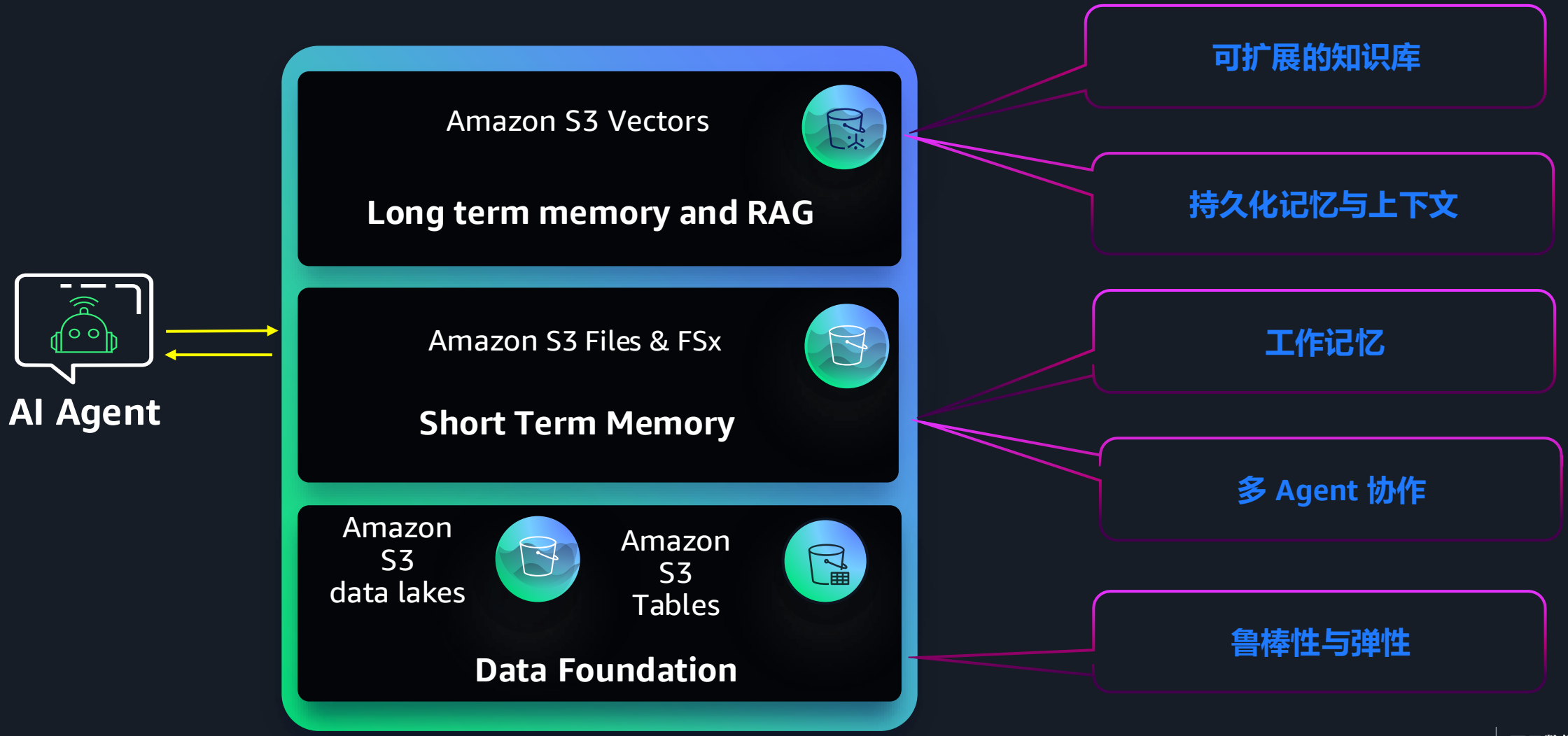
数十亿

QPS

<100s/index

存储 — AI Agent 的外部大脑

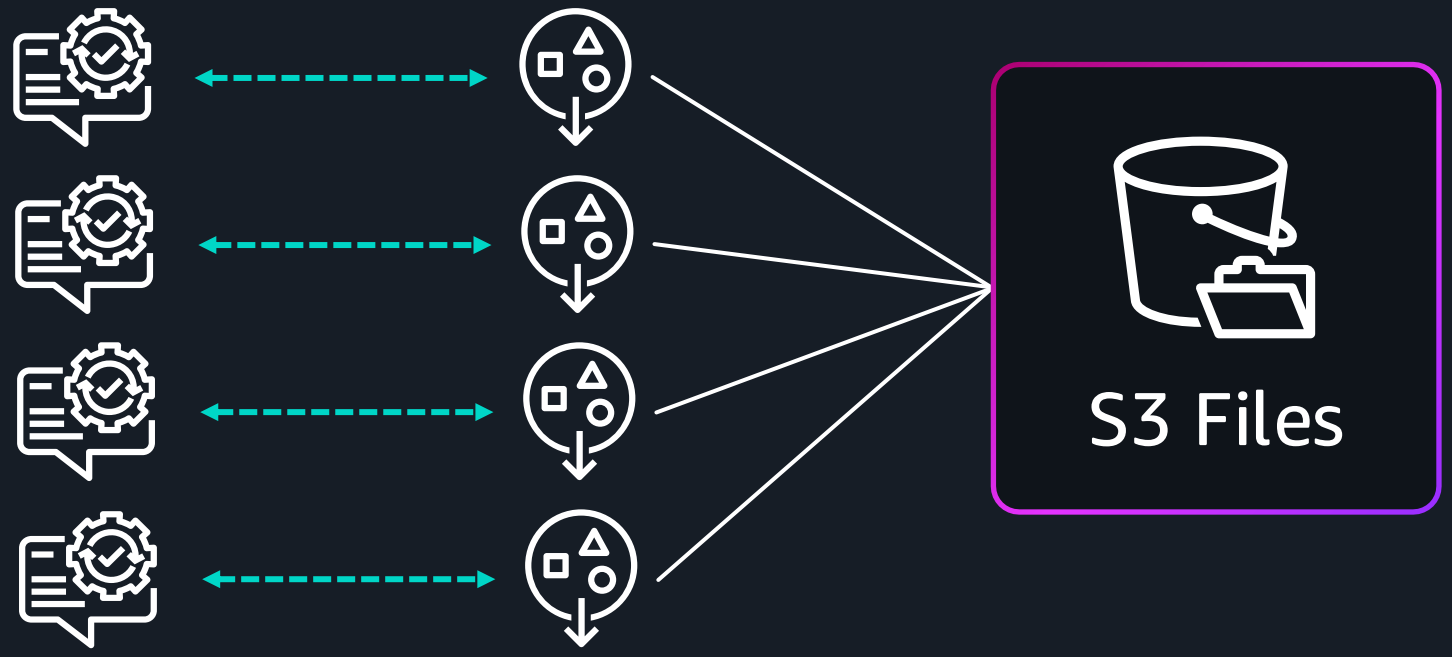
Amazon Storage - The External Brain for AI Agents



*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

S3 Files — 大规模 Agent 场景的持久化存储

For agents operating at scale



- 每个 Agent 都需要相互隔离的数据访问
- 通过 Access Points 实现数据访问隔离的 Agent 工作区，存储空间按需自动扩展
- 每个文件系统可扩展至 10,000 个 Access Points

推理与数据管道
Inference and Data Pipelines

S3 Files

为 ML 提供 S3 的文件系统访问
框架和推理管道

S3 Files 提供低延迟性能的原生NFS
挂载，无需数据复制或重新格式化

适用于模型加载和
批量推理工作负载

S3 Express One Zone

个位数毫秒级延迟
用于模型服务和实时推理

大规模并发下的一致高吞吐量
适用于需要快速读取的工作负载

适用于热模型缓存和实时服务

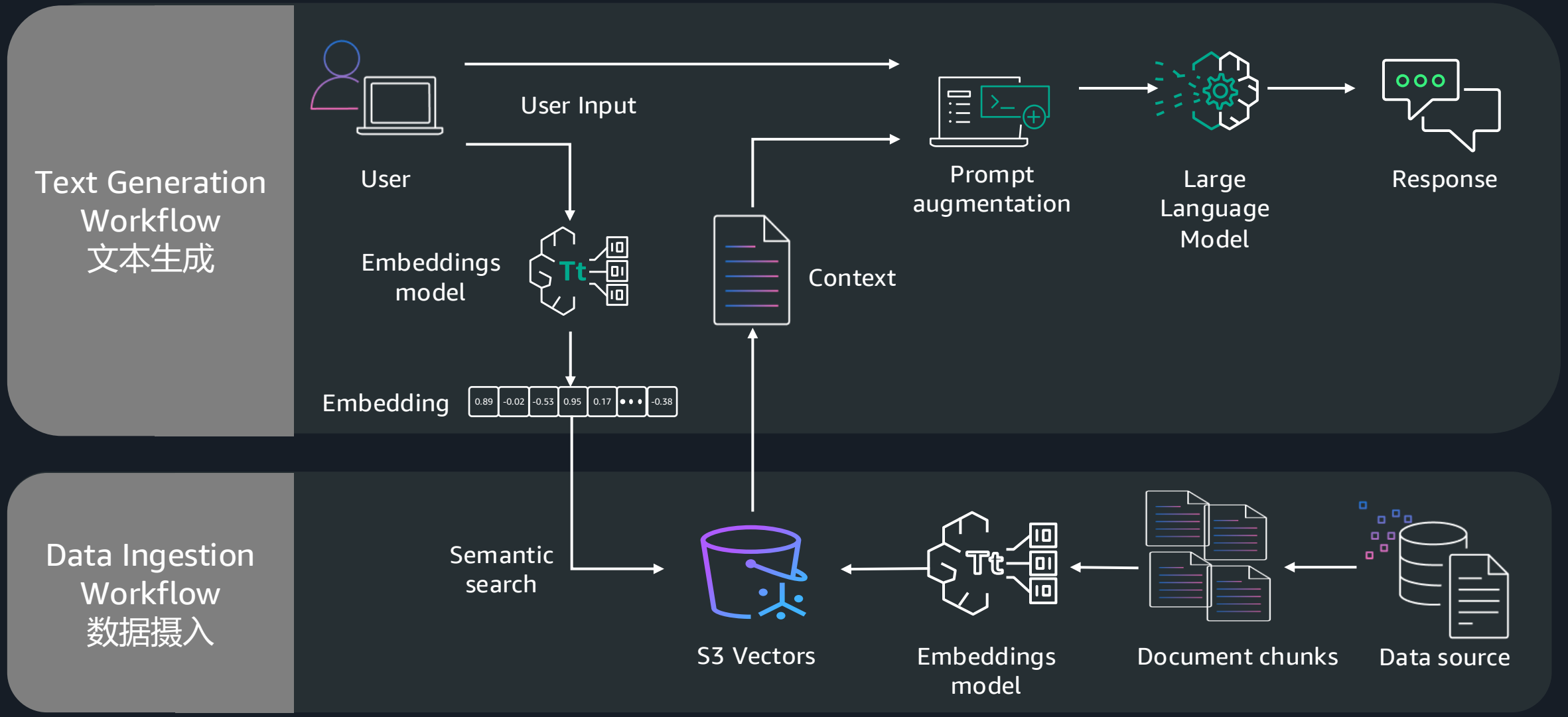
S3 Vectors

大规模存储和查询向量嵌入
用于 RAG 工作负载

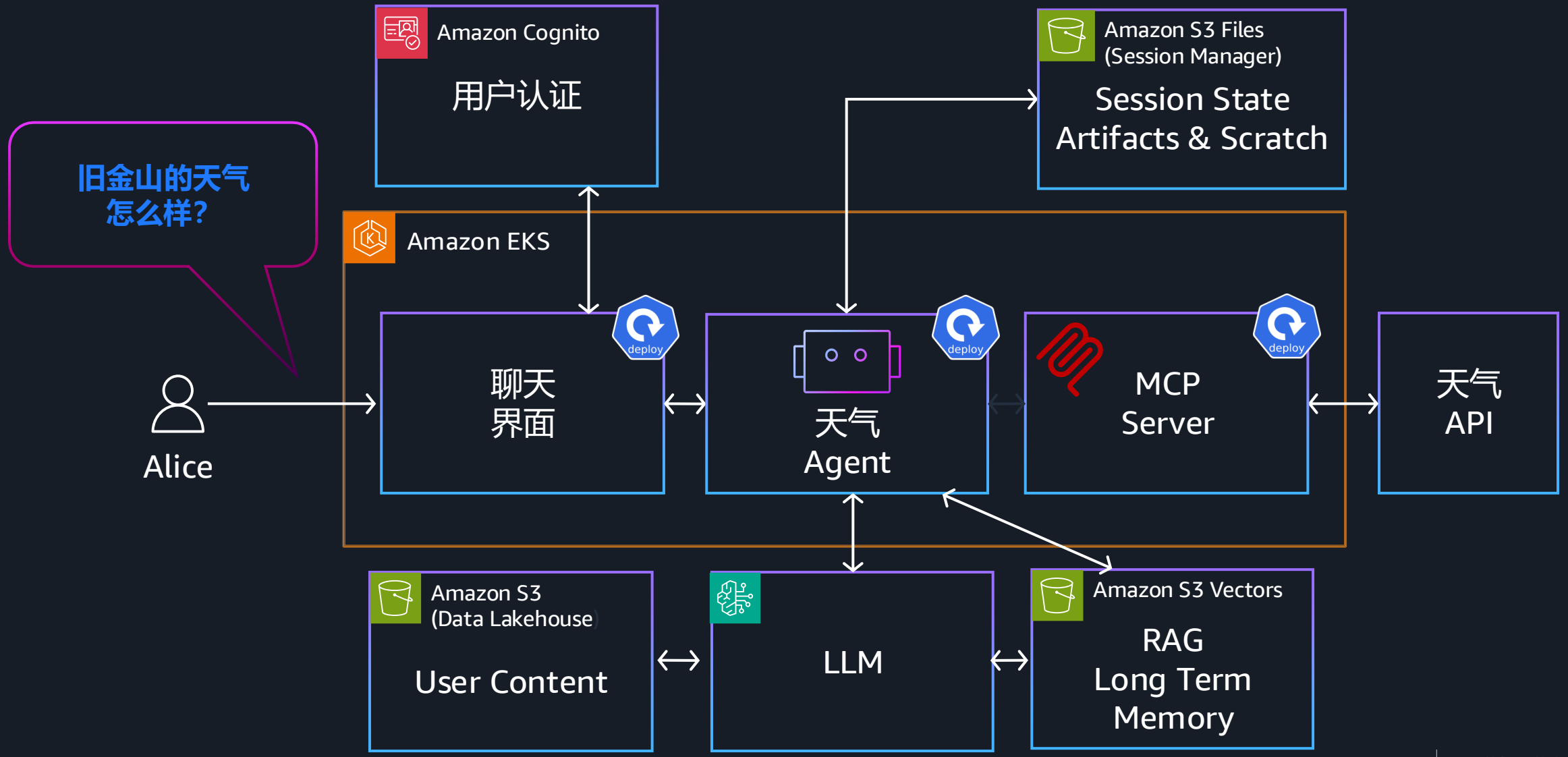
相比专用向量数据库
成本降低多达 90%
同时具备完整的 S3 持久性

适用于语义搜索和检索增强生成

RAG in Action



Agent 在 EKS 上的部署



开始使用 - Getting Started

1

识别您的工作负载

训练/微调?
将GPU数量匹配到存储层级

Agentic AI?
将数据映射到记忆类型
（短期 vs. 长期）

2

构建您的数据基础

以 S3 作为您的数据湖起步

使用 FSx for Lustre 用于训练

添加 S3 Files + S3 Vectors
用于 Agent 记忆和 RAG

3

加速投产

使用决策框架
进行训练存储选择

使用 Agent 记忆分类法
将存储原语映射
到 Agent 需求

联系您的亚马逊云科技存储专家 | aws.amazon.com/fsx | aws.amazon.com/s3

关键要点 - Key Takeaways

训练与微调

01 存储匹配集群规模

<16 GPU: S3 | 16-1000: FSx for Lustre | >1000: S3 Express

02 消除 I/O 瓶颈

闲置 GPU 每小时 $32 —— 存储吞吐必须匹配计算

03 从 FSx for Lustre 开始

POSIX、亚毫秒延迟、1 TB/s —— 覆盖大多数训练工作负载

Agentic AI

01 您的数据是差异化优势

Agent 的能力取决于它们能访问和推理的知识

02 Agent 需要记忆

短期（S3 Files）用于工作状态 + 长期（S3 Vectors）用于 RAG 和回忆

03 专用存储原语

S3 Files、S3 Vectors、S3 Tables —— 每个解决不同的 Agent 访问模式

存储是生成式 AI 的基础——从训练到生产 Agent

亚马逊云科技
中国峰会

Thank you