生成式人工智能和NetApp价值

NetApp artificial intelligence solutions

NetApp
August 18, 2025

This PDF was generated from https://docs.netapp.com/zh-cn/netapp-solutions-ai/gen-ai/ai-wp-genai.html on August 18, 2025. Always check docs.netapp.com for the latest.

目录

生成式人工智能和NetApp价值 . . . 1
- 摘要 . . . 1
- 内容提要 . . . 1
- 什么是生成式人工智能? . . . 1
  - 企业用例和下游 NLP 任务 . . . 2
  - 存储在生成式人工智能中的作用 . . . 2
  - 攻读法学硕士学位的三种主要途径 . . . 2
  - 基础模型 . . . 3
  - 微调、领域特异性和再训练 . . . 3
  - 快速工程和推理 . . . 3
  - LLMOps、模型监控和向量存储 . . . 3
  - 生成人工智能时代的风险与伦理 . . . 4
- 客户场景和NetApp . . . 4
- NetApp功能 . . . 4
  - * 搭载 DGX BasePOD 的ONTAP AI * . . . 5
  - * ONTAP AI 与NVIDIA AI Enterprise* . . . 6
  - **1P**云平台 . . . 6
- NetApp合作伙伴解决方案套件 . . . 6
- 结束语 . . . 6

生成式人工智能和**NetApp**价值

对生成人工智能 (AI) 的需求正在推动各个行业的颠覆，增强商业创造力和产品创新。

摘要

许多组织正在使用生成式人工智能来构建新的产品功能、提高工程生产力和原型人工智能应用程序，以提供更好的结果和消费者体验。生成式人工智能（例如生成式预训练变压器 (GPT)）使用神经网络来创建新内容，包括文本、音频和视频等多种内容。鉴于大型语言模型 (LLM) 涉及的极端规模和海量数据集，在公司设计 AI 解决方案之前，构建一个强大的 AI 基础架构至关重要，该基础架构可以利用内部部署、混合和多云部署选项的强大数据存储功能，并降低与数据移动性、数据保护和治理相关的风险。本文介绍了这些考虑因素以及相应的NetApp AI 功能，这些功能支持跨 AI 数据管道进行无缝数据管理和数据移动，以训练、再训练、微调和推理生成 AI 模型。

内容提要

最近，自 2022 年 11 月推出 GPT-3 的衍生产品 ChatGPT 以来，用于根据用户提示生成文本、代码、图像甚至治疗性蛋白质的新型 AI 工具获得了极大的声誉。这表明用户可以使用自然语言提出请求，人工智能将解释和生成文本，例如反映用户请求的新闻文章或产品描述，或使用基于现有数据训练的算法生成代码、音乐、语音、视觉效果和 3D 资产。因此，稳定扩散、幻觉、快速工程和价值一致性等短语正在人工智能系统的设计中迅速涌现。这些自监督或半监督机器学习 (ML) 模型正作为预先训练的基础模型 (FM) 通过云服务提供商和其他 AI 公司供应商得到广泛应用，各行各业的各种商业机构正在采用这些模型来执行广泛的下游 NLP (自然语言处理) 任务。正如麦肯锡等研究分析公司所言——“生成式人工智能对生产力的影响可能会为全球经济增加数万亿美元的价值。”当企业将人工智能重新想象为人类的思想伙伴，而设施管理者也同时拓展企业和机构利用生成性人工智能的能力时，管理海量数据的机会将继续增长。本文档介绍了生成式人工智能以及与NetApp功能相关的设计概念，这些功能为NetApp客户（包括本地和混合或多云环境）带来了价值。

那么，客户在其 **AI** 环境中使用**NetApp**有什么好处呢？ NetApp帮助组织应对快速数据和云增长、多云管理以及采用 AI 等下一代技术所带来的复杂性。 NetApp将各种功能整合到智能数据管理软件和存储基础架构中，并与针对 AI 工作负载优化的高性能实现了良好的平衡。像 LLM 这样的生成式 AI 解决方案需要多次读取和处理从存储到内存的源数据集以促进智能。 NetApp一直是边缘到核心到云生态系统中数据移动性、数据治理和数据安全技术的领导者，帮助企业客户构建大规模 AI 解决方案。 NetApp凭借强大的合作伙伴网络，一直致力于帮助首席数据官、AI 工程师、企业架构师和数据科学家设计自由流动的数据管道，以完成 AI 模型训练和推理的数据准备、数据保护和战略数据管理职责，从而优化 AI/ML 生命周期的性能和可扩展性。 NetApp数据技术和功能（例如用于深度学习数据管道的NetApp ONTAP AI、用于在存储端点之间无缝高效地传输数据的NetApp SnapMirror 以及用于在数据流从批量转变为实时且数据工程即时发生时进行实时渲染的NetApp FlexCache）为实时生成 AI 模型的部署带来了价值。随着各类企业采用新的人工智能工具，他们面临着从边缘到数据中心再到云端的数据挑战，需要可扩展、负责任和可解释的人工智能解决方案。作为混合和多云领域的数据权威， NetApp致力于构建合作伙伴和联合解决方案网络，以帮助构建数据管道和数据湖的各个方面，以进行生成式 AI 模型训练（预训练）、微调、基于上下文的推理和 LLM 的模型衰减监控。

什么是生成式人工智能?

生成式人工智能正在改变我们创造内容、产生新设计概念和探索新颖构图的方式。它展示了生成对抗网络 (GAN)、变分自动编码器 (VAE) 和生成预训练变压器 (GPT) 等神经网络框架，它们可以生成文本、代码、图像、音频、视频和合成数据等新内容。 OpenAI 的 Chat-GPT、Google 的 Bard、Hugging Face 的 BLOOM 和 Meta 的 LLaMA 等基于 Transformer 的模型已成为支撑大型语言模型诸多进步的基础技术。同样，OpenAI 的 Dall-E、Meta 的 CM3leon 和 Google 的 Imagen 都是文本到图像扩散模型的例子，它们为客户提供了前所未有的照片级真实感，可以从头开始创建新的复杂图像，或编辑现有图像以生成高质量的上下文感知图像，使用数据集增强和链接文本和视觉语义的文本到图像合成。数字艺术家开始将 NeRF（神经辐射场）等渲染技术与生成式人工智能相结合，将静态 2D 图像转换为沉浸式 3D 场景。一般来说，LLM 大致由四个参数表征：（1）模型的

大小（通常有数十亿个参数）；（2）训练数据集的大小；（3）训练成本；（4）训练后的模型性能。 LLM 也主要分为三种变压器架构。 （i）仅编码器模型。例如 BERT（Google，2018）；（ii）编码器-解码器，例如 BART（Meta，2020）和（iii）仅解码器模型。例如LLaMA（Meta，2023 年）、PaLM-E（Google，2023 年）。根据业务需求，无论公司选择哪种架构，训练数据集中的模型参数数量（N）和标记数量（D）通常决定训练（预训练）或微调 LLM 的基准成本。

企业用例和下游 **NLP** 任务

各行各业的企业都发现人工智能有越来越多的潜力，可以从现有数据中提取并产生新的价值形式，用于业务运营、销售、营销和法律服务。根据 IDC（国际数据公司）关于全球生成式人工智能用例和投资的市场情报，软件开发和产品设计中的知识管理受到的影响最大，其次是营销的故事情节创作和开发人员的代码生成。在医疗保健领域，临床研究组织正在开辟医学新领域。 ProteinBERT 等预训练模型结合了基因本体 (GO) 注释，可以快速设计药物的蛋白质结构，这代表了药物发现、生物信息学和分子生物学领域的一个重要里程碑。生物技术公司已启动生成性人工智能药物的人体试验，旨在治疗肺纤维化（IPF）等疾病，这是一种导致肺组织不可逆瘢痕形成的肺部疾病。

图 1：驱动生成式人工智能的用例

[图 1：驱动生成式人工智能的用例]

生成式人工智能推动的自动化应用的增加也正在改变许多职业的工作活动的供需。根据麦肯锡的数据，美国劳动力市场（下图）经历了快速转型，考虑到人工智能的影响，这种转型可能还会持续下去。

资料来源：麦肯锡公司

[图2：资料来源：麦肯锡公司]

存储在生成式人工智能中的作用

LLM 主要依赖于深度学习、GPU 和计算。然而，当GPU缓冲区填满时，数据需要快速写入存储器。虽然一些 AI 模型足够小，可以在内存中执行，但 LLM 需要高 IOPS 和高吞吐量存储才能快速访问大型数据集，特别是当它涉及数十亿个令牌或数百万个图像时。对于 LLM 的典型 GPU 内存要求，训练具有 10 亿个参数的模型所需的内存可能高达 80GB @32 位全精度。在这种情况下，Meta 的 LLaMA 2（一系列 LLM，规模从 70 亿到 700 亿个参数）可能需要 70x80、约 5600GB 或 5.6TB 的 GPU RAM。此外，您需要的内存量与您想要生成的最大令牌数量成正比。例如，如果你想生成最多 512 个 token（约 380 个单词）的输出，你需要"512 MB"。这看起来似乎无关紧要——但是，如果您想运行更大的批次，它就会开始累积。因此，对于组织来说，在内存中训练或微调 LLM 的成本非常高，从而使存储成为生成 AI 的基石。

攻读法学硕士学位的三种主要途径

对于大多数企业而言，根据目前的趋势，部署 LLM 的方法可以概括为 3 种基本场景。正如最近"《哈佛商业评论》"文章：（1）从头开始训练（预训练）法学硕士——成本高昂，并且需要专业的 AI/ML 技能；（2）使用企业数据微调基础模型——复杂但可行；（3）使用检索增强生成 (RAG) 查询包含公司数据的文档存储库、API 和矢量数据库。在实施过程中，每种方法都需要在工作量、迭代速度、成本效率和模型准确性之间进行权衡，用于解决不同类型的问题（下图）。

图 3：问题类型

[图 3：问题类型]

基础模型

基础模型 (FM) 也称为基础模型，是一种大型 AI 模型 (LLM)，使用大规模自我监督对大量未标记数据进行训练，通常适用于广泛的下游 NLP 任务。由于训练数据没有经过人工标记，因此模型是自然产生的，而不是明确编码的。这意味着该模型无需明确编程即可生成自己的故事或叙述。因此FM的一个重要特点就是同质化，即在许多领域使用相同的方法。然而，通过个性化和微调技术，如今出现的产品中集成的 FM 不仅擅长生成文本、文本转图像和文本转代码，而且还擅长解释特定领域的任务或调试代码。例如，OpenAI 的 Codex 或 Meta 的 Code Llama 等 FM 可以根据编程任务的自然语言描述生成多种编程语言的代码。这些模型精通十几种编程语言，包括 Python、C#、JavaScript、Perl、Ruby 和 SQL。它们理解用户的意图并生成完成所需任务的特定代码，这对于软件开发、代码优化和编程任务的自动化很有用。

微调、领域特异性和再训练

在数据准备和数据预处理之后进行 LLM 部署的常见做法之一是选择已经在大型多样化数据集上训练过的预训练模型。在微调的背景下，这可以是一个开源的大型语言模型，例如"Meta 的骆驼 2"对 700 亿个参数和 2 万亿个标记进行训练。一旦选择了预训练模型，下一步就是根据特定领域的数据对其进行微调。这涉及调整模型的参数并根据新数据对其进行训练以适应特定的领域和任务。例如，BloombergGPT 是一门专有的法学硕士课程，接受过广泛金融数据的培训，服务于金融行业。针对特定任务设计和训练的领域特定模型通常在其范围内具有更高的准确性和性能，但在其他任务或领域之间的可转移性较低。当业务环境和数据在一段时间内发生变化时，与测试期间的表现相比，FM 的预测准确性可能会开始下降。这时，重新训练或微调模型就变得至关重要。传统 AI/ML 中的模型再训练是指使用新数据更新已部署的 ML 模型，通常是为了消除发生的两种类型的漂移。 （1）概念漂移——当输入变量和目标变量之间的联系随时间而改变时，由于我们想要预测的描述发生了变化，模型可能会产生不准确的预测。 (2) 数据漂移——当输入数据的特征发生变化时发生，例如客户习惯或行为随时间发生变化，因此模型无法对此类变化做出反应。类似地，再培训也适用于 FM/LLM，但是成本可能要高得多（数百万美元），因此大多数组织可能不会考虑。它正处于积极的研究中，在 LLMOps 领域中仍然处于新兴阶段。因此，当微调 FM 中出现模型衰减时，企业可以选择使用较新的数据集再次进行微调（便宜得多），而不是重新训练。从成本角度来看，下面列出了 Azure-OpenAI 服务的模型价格表示例。对于每个任务类别，客户可以在特定数据集上微调和评估模型。

来源：Microsoft Azure

[来源：Microsoft Azure]

快速工程和推理

即时工程是指如何与 LLM 通信以执行所需任务而无需更新模型权重的有效方法。人工智能模型训练和微调对于 NLP 应用非常重要，推理也同样重要，训练后的模型可以响应用户的提示。推理的系统要求通常更多地取决于 AI 存储系统的读取性能，该系统将数据从 LLM 输送到 GPU，因为它需要能够应用数十亿个存储的模型参数来产生最佳响应。

LLMOps、模型监控和向量存储

与传统的机器学习操作 (MLOps) 一样，大型语言模型操作 (LLMOps) 也需要数据科学家和 DevOps 工程师的协作，并使用工具和最佳实践来管理生产环境中的 LLM。然而，法学硕士的工作流程和技术堆栈在某些方面可能会有所不同。例如，使用 LangChain 等框架构建的 LLM 管道将多个 LLM API 调用串联到外部嵌入端点（例如矢量存储或矢量数据库）。嵌入端点和矢量存储作为下游连接器（如矢量数据库）的使用代表了数据存储和访问方式的重大发展。与从头开始开发的传统 ML 模型相比，LLM 通常依赖于迁移学习，因为这些模型从 FM 开始，并使用新数据进行微调以提高在更特定领域的性能。因此，LLMOps 提供风险管理和模型衰减监测功能至关重要。

生成人工智能时代的风险与伦理

“ChatGPT——虽然很巧妙，但仍然会输出一些无意义的信息。”——《麻省理工科技评论》。垃圾进，垃圾出，一直是计算领域的难题。生成式人工智能的唯一区别在于，它擅长使垃圾变得高度可信，从而导致不准确的结果。法学硕士 (LLM) 倾向于捏造事实来适应其所构建的叙述。因此，那些将生成式人工智能视为利用人工智能降低成本的绝佳机会的公司需要有效地检测深度伪造、减少偏见并降低风险，以保持系统的诚实和道德。在负责任且可解释的生成式人工智能模型的设计中，拥有强大人工智能基础设施的自由流动数据管道至关重要，该管道通过端到端加密和人工智能护栏支持数据移动性、数据质量、数据治理和数据保护。

客户场景和NetApp

图 3：机器学习/大型语言模型工作流程

[图 3：机器学习/大型语言模型工作流程]

*我们是在训练还是在微调？*问题是（a）是否从头开始训练 LLM 模型、微调预先训练的 FM，或使用 RAG 从基础模型之外的文档存储库中检索数据并增强提示，以及（b）是否利用开源 LLM（例如 Llama 2）或专有 FM（例如 ChatGPT、Bard、AWS Bedrock），对于组织来说是一个战略决策。每种方法都需要在成本效率、数据引力、操作、模型准确性和 LLM 管理之间进行权衡。

NetApp公司在其工作文化以及产品设计和工程工作方法中都采用了人工智能。例如，NetApp 的自主勒索软件防护是使用人工智能和机器学习构建的。它提供文件系统异常的早期检测，以帮助在威胁影响操作之前识别它们。其次， NetApp将预测性 AI 用于其业务运营，例如销售和库存预测，并使用聊天机器人协助客户提供呼叫中心产品支持服务、技术规格、保修、服务手册等。第三， NetApp通过产品和解决方案为 AI 数据管道和 ML/LLM 工作流带来客户价值，帮助客户构建预测性 AI 解决方案，例如需求预测、医学成像、情绪分析和生成性 AI 解决方案，例如用于制造业工业图像异常检测和银行及金融服务中反洗钱和欺诈检测的 GAN ，NetApp NetApp ONTAP 、 NetApp SnapMirror和NetApp FlexCache。

NetApp功能

聊天机器人、代码生成、图像生成或基因组模型表达等生成式人工智能应用中的数据移动和管理可以跨越边缘、私有数据中心和混合多云生态系统。例如，一个实时人工智能机器人可以通过 ChatGPT 等预先训练模型的 API 公开的终端用户应用程序帮助乘客将机票升级为商务舱，但由于乘客信息并未在互联网上公开，因此该机器人无法自行完成该任务。该 API 需要访问乘客的个人信息和航空公司的机票信息，这些信息可能存在于混合或多云生态系统中。类似的情况可能适用于科学家通过最终用户应用程序共享药物分子和患者数据，该应用程序使用 LLM 完成涉及一对多生物医学研究机构的药物发现临床试验。传递给 FM 或 LLM 的敏感数据可能包括 PII、财务信息、健康信息、生物特征数据、位置数据、通信数据、在线行为和法律信息。在实时渲染、快速执行和边缘推理的情况下，数据会通过开源或专有 LLM 模型从最终用户应用程序移动到存储端点，再移动到内部数据中心或公共云平台上。在所有这些场景中，数据移动性和数据保护对于依赖大量训练数据集及其移动的 LLM 的 AI 操作至关重要。

图 4：生成式 AI - LLM 数据管道

[图 4：生成式 AI-LLM 数据管道]

NetApp 的存储基础设施、数据和云服务产品组合由智能数据管理软件提供支持。

数据准备：LLM 技术栈的第一个支柱与旧的传统 ML 栈基本没有变化。人工智能管道中的数据预处理是必要的，以便在训练或微调之前对数据进行规范化和清理。此步骤包括连接器，用于提取位于任何位置的数据，无论数据是以 Amazon S3 层的形式驻留在本地存储系统（例如文件存储或NetApp StorageGRID之类的对象存储）中。

- NetApp ONTAP* 是 NetApp 在数据中心和云端的关键存储解决方案的基础技术。 ONTAP包含各种数据管理和保护特性和功能，包括针对网络攻击的自动勒索软件保护、内置数据传输功能以及适用于本地、混合、NAS、SAN、对象和软件定义存储 (SDS) 等多种架构的存储效率功能。LLM 部署的情况。
- NetApp ONTAP AI* 用于深度学习模型训练。对于拥有ONTAP存储集群和NVIDIA DGX 计算节点的NetApp客户， NetApp ONTAP支持使用 NFS over RDMA 实现NVIDIA GPU 直接存储。它以经济高效的性能多次读取和处理来自存储的源数据集到内存中以促进智能，使组织能够对 LLM 进行培训、微调和扩展访问。
- NetApp FlexCache* 是一种远程缓存功能，可简化文件分发并仅缓存主动读取的数据。这对于 LLM 训练、再训练和微调非常有用，为具有实时渲染和 LLM 推理等业务需求的客户带来价值。
- NetApp SnapMirror* 是ONTAP 的一项功能，可在任意两个ONTAP系统之间复制卷快照。此功能可以最佳地将边缘数据传输到您的本地数据中心或云端。当客户希望使用包含企业数据的 RAG 在云中开发生成性 AI 时， SnapMirror可用于在本地和超大规模云之间安全高效地移动数据。它有效地仅传输更改，节省带宽并加快复制速度，从而在 FM 或 LLM 的训练、再训练和微调操作期间带来必要的数据移动功能。
- NetApp SnapLock* 为基于ONTAP的存储系统带来不可变磁盘功能，用于数据集版本控制。微核架构旨在通过 FPolicy Zero Trust 引擎保护客户数据。当攻击者以特别耗费资源的方式与 LLM 交互时， NetApp可通过抵御拒绝服务 (DoS) 攻击来确保客户数据可用。
- NetApp Cloud Data Sense* 有助于识别、映射和分类企业数据集中的个人信息，制定政策，满足本地或云端的隐私要求，帮助改善安全态势并遵守法规。
- NetApp BlueXP* 分类，由 Cloud Data Sense 提供支持。客户可以自动扫描、分析、分类和处理数据资产中的数据，检测安全风险，优化存储并加速云部署。它通过统一的控制平面结合了存储和数据服务，客户可以使用 GPU 实例进行计算，并使用混合多云环境进行冷存储分层以及存档和备份。
- NetApp文件对象二元性*。 NetApp ONTAP支持 NFS 和 S3 的双协议访问。通过此解决方案，客户可以通过NetApp Cloud Volumes ONTAP的 S3 存储桶访问来自 Amazon AWS SageMaker 笔记本的 NFS 数据。这为需要轻松访问异构数据源并能够从 NFS 和 S3 共享数据的客户提供了灵活性。例如，在 SageMaker 上通过访问文件对象存储桶来微调 FM，例如 Meta 的 Llama 2 文本生成模型。
- NetApp Cloud Sync* 服务提供了一种简单、安全的方式将数据迁移到云端或本地的任何目标。 Cloud Sync在本地或云存储、NAS 和对象存储之间无缝传输和同步数据。
- NetApp XCP* 是一款客户端软件，可实现快速可靠的任意到NetApp和NetApp到NetApp 的数据迁移。 XCP还提供将批量数据从 Hadoop HDFS 文件系统高效移动到ONTAP NFS、S3 或StorageGRID 的功能，并且XCP 文件分析可提供文件系统的可见性。
- NetApp DataOps Toolkit* 是一个 Python 库，它使数据科学家、DevOps 和数据工程师能够轻松执行各种数据管理任务，例如近乎即时地配置、克隆或快照数据卷或 JupyterLab 工作区，这些任务由高性能横向扩展NetApp存储支持。

NetApp 的产品安全。 LLM 可能会在回答中无意中泄露机密数据，因此研究利用 LLM 的 AI 应用程序相关漏洞的 CISO 对此表示担忧。正如 OWASP（开放式全球应用安全项目）所概述的，数据中毒、数据泄露、拒绝服务和 LLM 中的提示注入等安全问题可能会影响企业，防止数据暴露给未经授权的攻击者。数据存储要求应包括结构化、半结构化和非结构化数据的完整性检查和不可变快照。 NetApp Snapshots 和SnapLock用于数据集版本控制。它带来严格的基于角色的访问控制（RBAC）、安全协议和行业标准加密，以保护静态和传输中的数据。Cloud Insights和 Cloud Data Sense 共同提供功能，帮助您通过法医手段识别威胁来源并确定需要恢复的数据的优先级。

* 搭载 DGX BasePOD 的ONTAP AI *

采用NVIDIA DGX BasePOD 的NetApp ONTAP AI 参考架构是一种适用于机器学习 (ML) 和人工智能 (AI) 工作负载的可扩展架构。对于 LLM 的关键训练阶段，数据通常会定期从数据存储复制到训练集群中。此阶段使用的服务器使用 GPU 来并行计算，从而产生巨大的数据需求。满足原始 I/O 带宽需求对于维持高 GPU 利用率至关重要。

* ONTAP AI 与NVIDIA AI Enterprise*

NVIDIA AI Enterprise 是一款端到端、云原生的 AI 和数据分析软件套件，经过NVIDIA优化、认证和支持，可在具有NVIDIA认证系统的 VMware vSphere 上运行。该软件有助于在现代混合云环境中简单、快速地部署、管理和扩展 AI 工作负载。由NetApp和 VMware 提供支持的NVIDIA AI Enterprise 以简化、熟悉的软件包提供企业级 AI 工作负载和数据管理。

1P云平台

完全托管的云存储产品在 Microsoft Azure 上以Azure NetApp Files (ANF) 的形式原生提供，在 AWS 上以Amazon FSx for NetApp ONTAP (FSx ONTAP) 的形式提供，在 Google 上以Google Cloud NetApp Volumes (GNCV) 的形式提供。 1P 是一种托管的高性能文件系统，使客户能够在公共云中运行高可用性 AI 工作负载并提高数据安全性，以便使用 AWS SageMaker、Azure-OpenAI Services 和 Google 的 Vertex AI 等云原生 ML 平台对 LLM/FM 进行微调。

NetApp合作伙伴解决方案套件

除了核心数据产品、技术和功能外， NetApp还与强大的 AI 合作伙伴网络密切合作，为客户带来附加值。

- 人工智能系统中的NVIDIA Guardrails* 作为保障措施，确保以合乎道德和负责任的方式使用人工智能技术。 AI 开发人员可以选择定义 LLM 驱动的应用程序在特定主题上的行为，并阻止它们参与不想要的话题的讨论。 Guardrails 是一个开源工具包，它能够将 LLM 无缝安全地连接到其他服务，从而构建值得信赖、安全且有保障的 LLM 对话系统。

Domino Data Lab 提供多功能的企业级工具，用于构建和产品化生成式人工智能 - 无论您在人工智能之旅中的哪个阶段，都能快速、安全且经济地实现。借助 Domino 的企业 MLOps 平台，数据科学家可以使用首选工具和所有数据，在任何地方轻松训练和部署模型，并有效地管理风险和成本——所有这些都可以通过一个控制中心完成。

Modzy 用于 **Edge AI**。 NetApp和 Modzy 携手合作，为任何类型的数据（包括图像、音频、文本和表格）提供大规模 AI。 Modzy 是一个用于部署、集成和运行 AI 模型的 MLOps 平台，为数据科学家提供模型监控、漂移检测和可解释性的功能，并提供无缝 LLM 推理的集成解决方案。

Run:AI 和NetApp合作展示了NetApp ONTAP AI 解决方案与 Run:AI 集群管理平台的独特功能，以简化 AI 工作负载的编排。它自动分割和合并 GPU 资源，旨在将您的数据处理管道扩展到数百台机器，并为 Spark、Ray 、Dask 和 Rapids 内置集成框架。

结束语

只有在大量高质量数据上训练模型时，生成式人工智能才能产生有效的结果。虽然 LLM 已经取得了显著的里程碑，但认识到其局限性、设计挑战以及与数据移动性和数据质量相关的风险至关重要。 LLM 依赖于来自异构数据源的大量且不同的训练数据集。模型产生的不准确结果或有偏见的结果可能会使企业和消费者都陷入危险。这些风险可能对应于 LLM 可能因与数据质量、数据安全和数据移动性相关的数据管理挑战而产生的限制。 NetApp帮助组织应对快速数据增长、数据移动性、多云管理和 AI 采用所带来的复杂性。大规模的人工智能基础设施和高效的数据管理对于定义生成人工智能等人工智能应用的成功至关重要。至关重要的是，客户要覆盖所有部署场景，同时又不能影响企业根据需要扩展的能力，同时还要保持成本效益、数据治理和道德的人工智能实践。 NetApp一直致力于帮助客户简化和加速他们的 AI 部署。

版权信息

版权所有 © 2025 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。 美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

商标信息

NetApp、NetApp 标识和 http://www.netapp.com/TM 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。