生成式人工智慧和NetApp價值

NetApp artificial intelligence solutions

NetApp
December 04, 2025

目錄

生成式人工智慧和NetApp價值

對產生人工智慧 (AI) 的需求正在推動各行業的顛覆，增強商業創造力和產品創新。

抽象的

許多組織正在使用生成式人工智慧來建立新的產品功能、提高工程生產力和原型人工智慧應用程序，以提供更好的結果和消費者體驗。生成式人工智慧（例如生成式預訓練轉換器 (GPT)）使用神經網路來創建新內容，包括文字、音訊和視訊等多種內容。鑑於大型語言模型 (LLM) 涉及的極端規模和大量資料集，在公司設計 AI 解決方案之前，建立強大的 AI 基礎架構至關重要，該基礎架構可以利用內部部署、混合和多雲部署選項的強大資料儲存功能，並降低與資料移動性、資料保護和治理相關的風險。本文介紹了這些考慮因素以及相應的NetApp AI 功能，這些功能支援跨 AI 資料管道進行無縫資料管理和資料移動，以訓練、再訓練、微調和推理生成 AI 模型。

執行摘要

最近，自 2022 年 11 月推出 GPT-3 的衍生產品 ChatGPT 以來，用於根據用戶提示生成文字、程式碼、圖像甚至治療性蛋白質的新型 AI 工具獲得了極大的聲譽。這表明用戶可以使用自然語言提出請求，人工智慧將解釋和生成文本，例如反映用戶請求的新聞文章或產品描述，或使用基於現有數據訓練的演算法生成程式碼、音樂、語音、視覺效果和 3D 資產。因此，穩定擴散、幻覺、快速工程和價值一致性等短語正在人工智慧系統的設計中迅速湧現。這些自監督或半監督機器學習 (ML) 模型正作為預先訓練的基礎模型 (FM) 透過雲端服務提供者和其他 AI 公司供應商得到廣泛應用，各行各業的各種商業機構正在採用這些模型來執行廣泛的下游 NLP (自然語言處理) 任務。正如麥肯錫等研究分析公司所言——“生成式人工智慧對生產力的影響可能會為全球經濟增加數萬億美元的價值。”當企業將人工智慧重新想像為人類的思想夥伴，而設施管理者也同時拓展企業和機構利用生成性人工智慧的能力時，管理大量資料的機會將持續成長。本文檔介紹了生成式人工智慧以及與NetApp功能相關的設計概念，這些功能為NetApp客戶（包括本地和混合或多雲環境）帶來了價值。

*那麼，客戶在其 **AI** 環境中使用**NetApp**有什麼好處呢？ *

NetApp幫助組織應對快速資料和雲端成長、多雲管理以及採用 AI 等下一代技術所帶來的複雜性。NetApp將各種功能整合到智慧資料管理軟體和儲存基礎架構中，並與針對 AI 工作負載最佳化的高效能實現了良好的平衡。像 LLM 這樣的生成式 AI 解決方案需要多次讀取和處理從儲存到記憶體的來源資料集以促進智慧。

NetApp一直是邊緣到核心到雲端生態系統中資料移動性、資料治理和資料安全技術的領導者，幫助企業客戶建立大規模 AI 解決方案。NetApp憑藉著強大的合作夥伴網絡，一直致力於幫助首席資料長、AI 工程師、企業架構師和資料科學家設計自由流動的資料管道，以完成 AI 模型訓練和推理的資料準備、資料保護和策略資料管理職責，從而優化 AI/ML 生命週期的效能和可擴展性。NetApp資料技術和功能（例如用於深度學習資料管道的NetApp ONTAP AI、用於在儲存端點之間無縫高效地傳輸資料的NetApp SnapMirror以及用於在資料流從批次轉變為即時且資料工程即時發生時進行即時渲染的NetApp FlexCache）為即時生成 AI 模型的部署帶來了價值。隨著各類企業採用新的人工智慧工具，他們面臨從邊緣到資料中心再到雲端的資料挑戰，需要可擴展、負責任且可解釋的人工智慧解決方案。

作為混合和多雲領域的資料權威， NetApp致力於建立合作夥伴和聯合解決方案網絡，以幫助建立資料管道和資料湖的各個方面，以進行生成式 AI 模型訓練（預訓練）、微調、基於上下文的推理和 LLM 的模型衰減監控。

什麼是生成式人工智慧？

生成式人工智慧正在改變我們創造內容、產生新設計概念和探索新穎構圖的方式。它展示了生成對抗網路 (GAN)、變分自動編碼器 (VAE) 和生成預訓練變壓器 (GPT) 等神經網路框架，它們可以生成文字、程式碼、圖像、音訊、視訊和合成資料等新內容。 OpenAI 的 Chat-GPT、Google 的 Bard、Hugging Face 的 BLOOM 和

Meta 的 LLaMA 等基於 Transformer 的模型已成為支撐大型語言模型諸多進步的基礎技術。同樣，OpenAI 的 Dall-E、Meta 的 CM3leon 和 Google 的 Imagen 都是文本到圖像擴散模型的例子，它們為客戶提供了前所未有的照片級真實感，可以從頭開始創建新的複雜圖像，或編輯現有圖像以生成高質量的上下文感知圖像，使用數據集增強和鏈接文本和視覺語義的文本到圖像合成。數位藝術家開始將 NeRF（神經輻射場）等渲染技術與生成式人工智慧結合，將靜態 2D 影像轉換為沉浸式 3D 場景。一般來說，LLM 大致由四個參數來表徵：（1）模型的大小（通常有數十億個參數）；（2）訓練資料集的大小；（3）訓練成本；（4）訓練後的模型表現。 LLM 也主要分為三種變壓器架構。 （i）僅編碼器模型。例如 BERT（Google，2018）；（ii）編碼器-解碼器，例如 BART（Meta，2020）和（iii）僅解碼器模型。例如LLaMA（Meta，2023 年）、PaLM-E（Google，2023 年）。根據業務需求，無論公司選擇哪種架構，訓練資料集中的模型參數數量（N）和標記數量（D）通常決定訓練（預訓練）或微調 LLM 的基準成本。

企業用例與下游 **NLP** 任務

各行各業的企業都發現人工智慧有越來越多的潛力，可以從現有數據中提取並產生新的價值形式，用於業務運營、銷售、行銷和法律服務。根據 IDC（國際數據公司）關於全球生成式人工智慧用例和投資的市場情報，軟體開發和產品設計中的知識管理受到的影響最大，其次是行銷的故事情節創作和開發人員的程式碼生成。在醫療保健領域，臨床研究組織正在開闢醫學新領域。 ProteinBERT 等預訓練模型結合了基因本體 (GO) 註釋，可以快速設計藥物的蛋白質結構，這代表了藥物發現、生物資訊學和分子生物學領域的一個重要里程碑。生物技術公司已啟動生成性人工智慧藥物的人體試驗，旨在治療肺纖維化（IPF）等疾病，這是一種導致肺組織不可逆瘢痕形成的肺部疾病。

圖 1：驅動生成式人工智慧的用例

[圖 1：驅動生成式人工智慧的用例]

生成式人工智慧推動的自動化應用的增加也正在改變許多職業的工作活動的供需。根據麥肯錫的數據，美國勞動市場（下圖）經歷了快速轉型，考慮到人工智慧的影響，這種轉型可能還會持續下去。

資料來源：麥肯錫公司

[圖2：資料來源：麥肯錫公司]

儲存在生成式人工智慧中的作用

LLM 主要依賴深度學習、GPU 和運算。然而，當GPU緩衝區填滿時，資料需要快速寫入記憶體。雖然一些 AI 模型足夠小，可以在記憶體中執行，但 LLM 需要高 IOPS 和高吞吐量儲存才能快速存取大型資料集，特別是當它涉及數十億個令牌或數百萬個影像時。對於 LLM 的典型 GPU 記憶體需求，訓練具有 10 億個參數的模型所需的記憶體可能高達 80GB @32 位元全精度。在這種情況下，Meta 的 LLaMA 2（一系列 LLM，規模從 70 億到 700 億個參數）可能需要 70x80、約 5600GB 或 5.6TB 的 GPU RAM。此外，您需要的記憶體量與您想要產生的最大令牌數量成正比。例如，如果你想產生最多 512 個 token（約 380 個單字）的輸出，你需要"512MB"。這看起來似乎無關緊要——但是，如果您想運行更大的批次，它就會開始累積。因此，對於組織來說，在記憶體中訓練或微調 LLM 的成本非常高，從而使儲存成為生成 AI 的基石。

攻讀法學碩士學位的三種主要途徑

對於大多數企業而言，根據目前的趨勢，部署 LLM 的方法可以概括為 3 種基本情境。正如最近"《哈佛商業評論》"文章：（1）從頭開始訓練（預訓練）法學碩士——成本高昂，並且需要專業的 AI/ML 技能；（2）使用企業資料微調基礎模型——複雜但可行；（3）使用檢索增強生成 (RAG) 查詢包含公司資料的文件儲存庫、API 和向量資料庫。在實施過程中，每種方法都需要在工作量、迭代速度、成本效率和模型準確性之間進行權衡，以解決不同類型的問題（下圖）。

圖 3：問題類型

[圖 3：問題類型]

基礎模型

基礎模型 (FM) 也稱為基礎模型，是一種大型 AI 模型 (LLM)，使用大規模自我監督對大量未標記資料進行訓練，通常適用於廣泛的下游 NLP 任務。由於訓練資料沒有經過人工標記，因此模型是自然產生的，而不是明確編碼的。這意味著該模型無需明確編程即可產生自己的故事或敘述。因此FM的一個重要特徵是同質化，即在許多領域使用相同的方法。然而，透過個人化和微調技術，如今出現的產品中整合的 FM 不僅擅長生成文字、文字轉圖像和文字轉程式碼，而且還擅長解釋特定領域的任務或偵錯程式碼。例如，OpenAI 的 Codex 或 Meta 的 Code Llama 等 FM 可以根據程式設計任務的自然語言描述產生多種程式語言的程式碼。這些模型精通十幾種程式語言，包括 Python、C#、JavaScript、Perl、Ruby 和 SQL。它們理解使用者的意圖並產生完成所需任務的特定程式碼，這對於軟體開發、程式碼最佳化和程式設計任務的自動化很有用。

微調、領域特異性與再訓練

在資料準備和資料預處理之後進行 LLM 部署的常見做法之一是選擇已經在大型多樣化資料集上訓練過的預訓練模型。在微調的背景下，這可以是開源的大型語言模型，例如："Meta 的駱駝 2"使用 700 億個參數和 2 兆個代幣進行訓練。一旦選擇了預訓練模型，下一步就是根據特定領域的資料微調。這涉及調整模型的參數並根據新數據對其進行訓練以適應特定的領域和任務。例如，BloombergGPT 是一門專有的法學碩士課程，接受過廣泛金融數據的培訓，服務於金融業。

針對特定任務設計和訓練的領域特定模型通常在其範圍內具有更高的準確性和性能，但在其他任務或領域之間的可轉移性較低。當業務環境和資料在一段時間內發生變化時，與測試期間的表現相比，FM 的預測準確性可能會開始下降。這時，重新訓練或微調模型就變得至關重要。

傳統 AI/ML 中的模型再訓練是指使用新資料更新已部署的 ML 模型，通常是為了消除發生的兩種類型的漂移。（1）概念漂移－當輸入變數和目標變數之間的聯繫隨時間而改變時，由於我們想要預測的描述發生了變化，模型可能會產生不準確的預測。(2) 資料漂移－當輸入資料的特徵發生變化時發生，例如客戶習慣或行為隨時間變化，因此模型無法對此類變化做出反應。

類似地，再培訓也適用於 FM/LLM，但成本可能要高得多（數百萬美元），因此大多數組織可能不會考慮。它正處於積極的研究中，在 LLMOps 領域中仍然處於新興階段。因此，當微調 FM 中出現模型衰減時，企業可以選擇使用較新的資料集再次進行微調（便宜得多），而不是重新訓練。從成本角度來看，以下列出了 Azure-OpenAI 服務的模型價格表示例。對於每個任務類別，客戶可以在特定資料集上微調和評估模型。

來源：Microsoft Azure

[來源：Microsoft Azure]

快速工程和推理

即時工程是指如何與 LLM 通訊以執行所需任務而無需更新模型權重的有效方法。人工智慧模型訓練和微調對於 NLP 應用非常重要，推理也同樣重要，訓練後的模型可以回應使用者的提示。推理的系統需求通常更多地取決於 AI 儲存系統的讀取性能，該系統將資料從 LLM 輸送到 GPU，因為它需要能夠應用數十億個儲存的模型參數來產生最佳響應。

LLMOps、模型監控和向量存儲

與傳統的機器學習操作 (MLOps) 一樣，大型語言模型操作 (LLMOps) 也需要資料科學家和 DevOps 工程師的協作，並使用生產環境中 LLM 管理的工具和最佳實踐。然而，法學碩士的工作流程和技術堆疊在某些方面可能會有所不同。例如，使用 LangChain 等框架建立的 LLM 管道將多個 LLM API 呼叫串聯到外部嵌入端點（例如向量儲存或向量資料庫）。嵌入端點和向量儲存作為下游連接器（如向量資料庫）的使用代表了資料儲存和存取方

式的重大發展。與從頭開始開發的傳統 ML 模型相比，LLM 通常依賴遷移學習，因為這些模型從 FM 開始，並使用新資料進行微調以提高在更特定領域的效能。因此，LLMOps 提供風險管理和模型衰減監測功能至關重要。

生成人工智慧時代的風險與倫理

「ChatGPT——雖然很巧妙，但仍然會輸出一些無意義的信息。」——《麻省理工科技評論》。垃圾進，垃圾出，一直是計算領域的難題。生成式人工智慧的唯一區別在於，它擅長使垃圾高度可信，從而導致不準確的結果。法學碩士 (LLM) 傾向於捏造事實來適應其所建構的敘述。因此，那些將生成式人工智慧視為利用人工智慧降低成本的絕佳機會的公司需要有效地檢測深度偽造、減少偏見並降低風險，以保持系統的誠實和道德。在負責任且可解釋的生成式人工智慧模型的設計中，擁有強大人工智慧基礎設施的自由流動資料管道至關重要，該管道透過端到端加密和人工智慧護欄支援資料移動性、資料品質、資料治理和資料保護。

客戶場景和NetApp

圖 3：機器學習/大型語言模型工作流程

[圖 3：機器學習/大型語言模型工作流程]

*我們是在訓練還是微調？ *問題是（a）是否從頭開始訓練 LLM 模型、微調預先訓練的 FM，或使用 RAG 從基礎模型以外的文件儲存庫中檢索資料並增強提示，以及（b）是否利用開源 LLM（例如 Llama 2）或專有 FM（例如 ChatGPT、Bard、AWS Bedrock），對於組織來說是一個策略決策。每種方法都需要在成本效率、資料引力、操作、模型準確性和 LLM 管理之間進行權衡。

NetApp公司在其工作文化以及產品設計和工程工作方法中都採用了人工智慧。例如，NetApp 的自主勒索軟體防護是使用人工智慧和機器學習建構的。它提供檔案系統異常的早期檢測，以幫助在威脅影響操作之前識別它們。其次， NetApp將預測性 AI 用於其業務運營，例如銷售和庫存預測，並使用聊天機器人協助客戶提供呼叫中心產品支援服務、技術規格、保固、服務手冊等。第三， NetApp透過產品和解決方案為 AI 資料管道和 ML/LLM 工作流程帶來客戶價值，幫助客戶建立預測性 AI 解決方案，例如需求預測、醫學影像、情緒分析和生成性 AI 解決方案，例如用於ONTAP工業影像異常檢測和銀行及金融服務中反洗錢和詐欺檢測的 GAN ，NetApp NetApp 、 AppApp SnapMirror 、 NetApp FlexCache並

NetApp功能

聊天機器人、程式碼生成、圖像生成或基因組模型表達等生成式人工智慧應用中的資料移動和管理可以跨越邊緣、私有資料中心和混合多雲生態系統。例如，一個即時人工智慧機器人可以透過 ChatGPT 等預先訓練模型的 API 公開的終端用戶應用程式來幫助乘客將機票升級到商務艙，但由於乘客資訊並未在網路上公開，因此該機器人無法自行完成該任務。該 API 需要存取乘客的個人資訊和航空公司的機票信息，這些資訊可能存在於混合或多雲生態系統中。類似的情況可能適用於科學家透過最終用戶應用程式共享藥物分子和患者數據，該應用程式使用 LLM 完成涉及一對多生物醫學研究機構的藥物發現臨床試驗。傳遞給 FM 或 LLM 的敏感資料可能包括 PII、財務資訊、健康資訊、生物特徵資料、位置資料、通訊資料、線上行為和法律資訊。在即時渲染、快速執行和邊緣推理的情況下，資料會透過開源或專有 LLM 模型從最終用戶應用程式移動到儲存端點，再移動到內部資料中心或公有雲平台。在所有這些場景中，資料移動性和資料保護對於依賴大量訓練資料集及其移動的 LLM 的 AI 操作至關重要。

圖 4：生成式 AI - LLM 資料管道

[圖 4：生成式 AI-LLM 資料管道]

NetApp 的儲存基礎架構、資料和雲端服務產品組合由智慧資料管理軟體提供支援。

資料準備：LLM 技術棧的第一個支柱與舊的傳統 ML 棧基本沒有變化。人工智慧管道中的資料預處理是必要的

，以便在訓練或微調之前對資料進行標準化和清理。此步驟包括連接器，用於提取位於任何位置的數據，無論數據是以 Amazon S3 層的形式駐留在本地存儲系統（例如文件存儲或NetApp StorageGRID之類的對象存儲）中。

- NetApp ONTAP* 是 NetApp 在資料中心和雲端的關鍵儲存解決方案的基礎技術。 ONTAP包含各種資料管理和保護特性和功能，包括針對網路攻擊的自動勒索軟體保護、內建資料傳輸功能以及適用於本機、混合、NAS、SAN、物件和軟體定義儲存 (SDS) 等多種架構的儲存效率功能。 LLM 部署的情況。
- NetApp ONTAP AI* 用於深度學習模型訓練。對於擁有ONTAP儲存叢集和NVIDIA DGX 運算節點的NetApp客戶， NetApp ONTAP支援使用 NFS over RDMA 實作NVIDIA GPU 直接儲存。它以經濟高效的性能多次讀取和處理來自存儲的源數據集到內存中以促進智能，使組織能夠對 LLM 進行培訓、微調和擴展訪問。
- NetApp FlexCache* 是一種遠端快取功能，可簡化檔案分發並僅快取主動讀取的資料。這對於 LLM 訓練、再訓練和微調非常有用，為具有即時渲染和 LLM 推理等業務需求的客戶帶來價值。
- NetApp SnapMirror* 是ONTAP 的一項功能，可在任兩個ONTAP系統之間複製磁碟區快照。此功能可最佳地將邊緣資料傳輸到您的本機資料中心或雲端。當客戶希望使用包含企業資料的 RAG 在雲端中開發生成性 AI 時， SnapMirror可用於在本地和超大規模雲端之間安全且有效率地移動資料。它有效地僅傳輸更改，節省頻寬並加快複製速度，從而在 FM 或 LLM 的訓練、再訓練和微調操作期間帶來必要的資料移動功能。
- NetApp SnapLock* 為基於ONTAP的儲存系統帶來不可變磁碟功能，用於資料集版本控制。微核架構旨在透過 FPolicy Zero Trust 引擎保護客戶資料。當攻擊者以特別耗費資源的方式與 LLM 互動時， NetApp可透過抵禦拒絕服務 (DoS) 攻擊來確保客戶資料可用。
- NetApp Cloud Data Sense* 有助於識別、映射和分類企業資料集中的個人信息，制定政策，滿足本地或雲端的隱私要求，幫助改善安全態勢並遵守法規。
- NetApp BlueXP* 分類，由 Cloud Data Sense 提供支援。客戶可以自動掃描、分析、分類和處理資料資產中的數據，偵測安全風險，優化儲存並加速雲端部署。它透過統一的控制平面結合了儲存和資料服務，客戶可以使用 GPU 執行個體進行運算，並使用混合多雲環境進行冷儲存分層以及存檔和備份。
- NetApp檔案物件二元性*。 NetApp ONTAP支援 NFS 和 S3 的雙協定存取。透過此解決方案，客戶可以透過NetApp Cloud Volumes ONTAP的 S3 儲存桶存取來自 Amazon AWS SageMaker 筆記本的 NFS 資料。這為需要輕鬆存取異質資料來源並能夠從 NFS 和 S3 共享資料的客戶提供了靈活性。例如，在 SageMaker 上透過存取檔案物件儲存桶來微調 FM，例如 Meta 的 Llama 2 文字生成模型。
- NetApp Cloud Sync* 服務提供了一種簡單、安全的方式將資料遷移到雲端或本地的任何目標。 Cloud Sync 在本機或雲端儲存、NAS 和物件儲存之間無縫傳輸和同步資料。
- NetApp XCP* 是一款用戶端軟體，可實現快速可靠的任意到NetApp和NetApp到NetApp 的資料遷移。 XCP 還提供將大量資料從 Hadoop HDFS 檔案系統高效移動到ONTAP NFS、S3 或StorageGRID 的功能，並且 XCP 檔案分析可提供檔案系統的可見性。
- NetApp DataOps Toolkit* 是一個 Python 函式庫，它使資料科學家、DevOps 和資料工程師能夠輕鬆執行各種資料管理任務，例如近乎即時地配置、複製或快照資料磁碟區或 JupyterLab 工作區，這些任務由高效能橫向擴展NetApp儲存支援。

NetApp 的產品安全。 LLM 可能會在回答中無意中洩露機密數據，因此研究利用 LLM 的 AI 應用程式相關漏洞的 CISO 對此表示擔憂。正如 OWASP（開放式全球應用安全專案）所概述的，資料中毒、資料外洩、拒絕服務和 LLM 中的提示注入等安全性問題可能會影響企業，防止資料暴露給未經授權的攻擊者。資料儲存需求應包括結構化、半結構化和非結構化資料的完整性檢查和不可變快照。 NetApp Snapshots 和SnapLock用於資料集版本控制。它帶來嚴格的基於角色的存取控制（RBAC）、安全協定和行業標準加密，以保護靜態和傳輸中的資料。 Cloud Insights和 Cloud Data Sense 共同提供功能，協助您透過法醫手段識別威脅來源並確定需要復原的資料的優先順序。

* 搭載 **DGX BasePOD** 的**ONTAP AI ***

採用NVIDIA DGX BasePOD 的NetApp ONTAP AI 參考架構是一種適用於機器學習 (ML) 和人工智慧 (AI) 工作負

載的可擴展架構。對於 LLM 的關鍵訓練階段，資料通常會定期從資料儲存複製到訓練叢集。此階段使用的伺服器使用 GPU 來並行運算，產生龐大的資料需求。滿足原始 I/O 頻寬需求對於維持高 GPU 利用率至關重要。

* ONTAP AI 與NVIDIA AI Enterprise*

NVIDIA AI Enterprise 是一款端到端、雲端原生的 AI 和資料分析軟體套件，經過NVIDIA優化、認證和支持，可在具有NVIDIA認證系統的 VMware vSphere 上運行。該軟體有助於在現代混合雲環境中簡單、快速地部署、管理和擴展 AI 工作負載。由NetApp和 VMware 提供支援的NVIDIA AI Enterprise 以簡化、熟悉的軟體包提供企業級 AI 工作負載和資料管理。

1P雲端平台

完全託管的雲端儲存產品在 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 是一種託管的高效能檔案系統，可讓客戶在公有雲中運行高可用性 AI 工作負載並提高資料安全性，以便使用 AWS SageMaker、Azure-OpenAI Services 和 Google 的 Vertex AI 等雲端原生 ML 平台對 LLM/FM 進行微調。

NetApp合作夥伴解決方案套件

除了核心資料產品、技術和功能外， NetApp還與強大的 AI 合作夥伴網路密切合作，為客戶帶來附加價值。

- 人工智慧系統中的NVIDIA Guardrails* 作為保障措施，確保以合乎道德和負責任的方式使用人工智慧技術。AI 開發人員可以選擇定義 LLM 驅動的應用程式在特定主題上的行為，並阻止它們參與不想要的話題的討論。 Guardrails 是一個開源工具包，它能夠將 LLM 無縫安全地連接到其他服務，從而建立值得信賴、安全可靠的 LLM 對話系統。

Domino Data Lab 提供多功能的企業級工具，用於構建和產品化生成式人工智慧 - 無論您在人工智慧旅程中的哪個階段，都能快速、安全且經濟地實現。借助 Domino 的企業 MLOps 平台，資料科學家可以使用首選工具和所有數據，在任何地方輕鬆訓練和部署模型，並有效地管理風險和成本——所有這些都可以透過一個控制中心完成。

Modzy 用於 **Edge AI**。 NetApp和 Modzy 攜手合作，為任何類型的資料（包括圖像、音訊、文字和表格）提供大規模 AI。 Modzy 是一個用於部署、整合和運行 AI 模型的 MLOps 平台，為資料科學家提供模型監控、漂移偵測和可解釋性的功能，並提供無縫 LLM 推理的整合解決方案。

Run:AI 和NetApp合作展示了NetApp ONTAP AI 解決方案與 Run:AI 叢集管理平台的獨特功能，以簡化 AI 工作負載的編排。它自動分割和合併 GPU 資源，旨在將您的資料處理管道擴展到數百台機器，並為 Spark、Ray、Dask 和 Rapids 內建整合框架。

結論

只有在大量高品質資料上訓練模型時，生成式人工智慧才能產生有效的結果。雖然 LLM 已經取得了顯著的里程碑，但認識到其局限性、設計挑戰以及與資料移動性和資料品質相關的風險至關重要。 LLM 依賴來自異質資料來源的大量且不同的訓練資料集。模型產生的不準確結果或偏見的結果可能會使企業和消費者都陷入危險。這些風險可能對應於 LLM 可能因與資料品質、資料安全性和資料移動性相關的資料管理挑戰而產生的限制。NetApp幫助組織應對快速資料成長、資料移動性、多雲管理和 AI 採用所帶來的複雜性。大規模的人工智慧基礎設施和高效的數據管理對於定義生成人工智慧等人工智慧應用的成功至關重要。至關重要的是，客戶要涵蓋所有部署場景，同時又不能影響企業根據需要擴展的能力，同時還要保持成本效益、資料治理和道德的人工智慧實踐。 NetApp一直致力於幫助客戶簡化和加速他們的 AI 部署。