適用於生成式人工智慧的 VMware 基準參考架構

主要元件

- Hugging Face Transformers Library · PyTorch
- Ray · Kubeflow
- VMware Cloud Foundation 5.0

摘要

VMware 生成式人工智慧基準參考架構旨在提供最快速的途徑，以協助企業從機器學習專案邁向生產作業。這款平台可運用領先業界的虛擬化和雲端技術來兼顧效率和安全性，藉由搭配尖端的訓練和推斷工具及 GPU 支援，運用簡單且效率十足的方式大規模管理生成式人工智慧工作負載，以加速推動人工智慧轉型。這項通過驗證的生成式人工智慧技術，提供一款運用開放原始碼元件的整合式解決方案，不僅可輕鬆操作，更內建 VMware 平台的自動化生命週期管理功能，能大幅簡化人工智慧 / 機器學習工作負載的實際運用。

主要優勢

- **加速實現成果：**消弭系統設計、測試、啟動、設定和佈建程序的複雜性，以縮短產品上市時間。
- **備受信賴的部署：**奠基於標準化的 VMware Validated 架構，能進行快速、可重複且安全的部署。
- **提高 ROI：**使用通用的基礎平台來充分利用資源，並縮短支援不同部門的所需時間。
- **彈性基礎架構：**可在通用的地端和雲端基礎架構上，執行所有人工智慧工作負載，包括預測式和生成式人工智慧。

VMware 的生成式人工智慧參考架構可提供全方位的解決方案，整合人工智慧和企業工作負載，藉此提升營運效率，並快速推動基礎架構變更。這個架構涵蓋建議的軟硬體、整合和支援，無需集結不同的項目，就能簡化實作和部署作業。當中會運用 VMware 虛擬化技術，以保護專屬虛擬環境中的敏感資料和應用程式，從而確保資料隱私權，並防範未經授權的存取。管理員可全面掌控資源配置和網路設定，藉此強化運算資源的管理和自訂作業，以滿足特定業務需求。

適用於生成式人工智慧的 VMware 解決方案經過驗證，將可減少建置和管理 LLM 基礎架構和 LLM 軟體堆疊的所需時間、心力和成本，進而縮短實現價值時間，並協助企業在兼顧安全性和資料隱私權的前提下，著重實現生成式人工智慧式應用程式的價值。IT 團隊可運用 VMware 管理工具，在 GPU 加速的軟體定義雲端上，快速部署、管理和擴充人工智慧工作負載。

適用於 LLM 自訂和應用程式部署的 VMware 生成式人工智慧平台

客戶可將這款解決方案運用在兩種主要 LLM 工作流程上，也就是大規模的自訂 (微調、提示調整和其他作業) 和推斷。對 LLM 而言，這兩種工作流程講求的運算能力，都高於傳統機器學習或深度學習工作負載。

- 開放原始碼 LLM 的自訂作業 (涵蓋數十億到 150 億個以上的參數)，需要在跨越多個伺服器的多個 GPU 上持續進行分散式訓練。
- 推斷也需運用 GPU 資源，且大幅取決於應用程式需求和並行使用者數量。

使用 vSphere 功能，即可視運算和服務層級協定需求而定，輕鬆運用環境中未充分運用的 GPU 資源來進行模型或推斷的分散式訓練，進而提高基礎架構利用率，並改善機器學習工作流程的整體生產力。

只要運用經 VMware 整合且驗證的解決方案設定，企業幾乎可立即著手實現解決方案的種種優勢，繼而縮短讓生成式人工智慧模型投入生產的所需時間，並實現其投資價值。

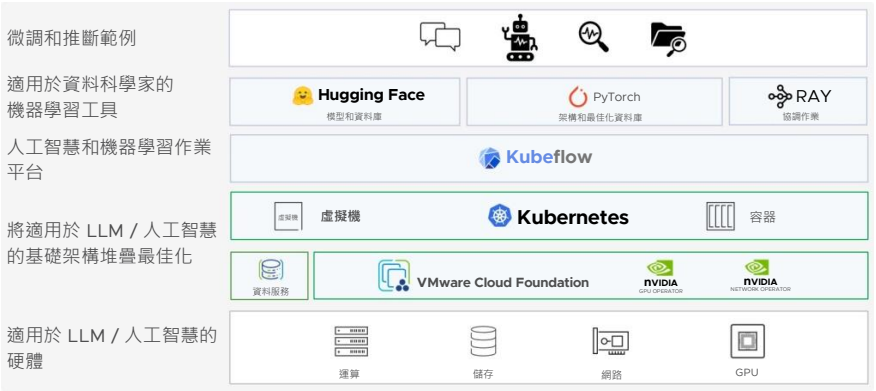
重要 vSphere 功能

- NVIDIA NVSwitch 支援
- 裝置群組
- 使用裝置群組簡化的硬體使用
- 異質 vGPU 設定檔
- 提高安全性

LLM 堆疊概觀

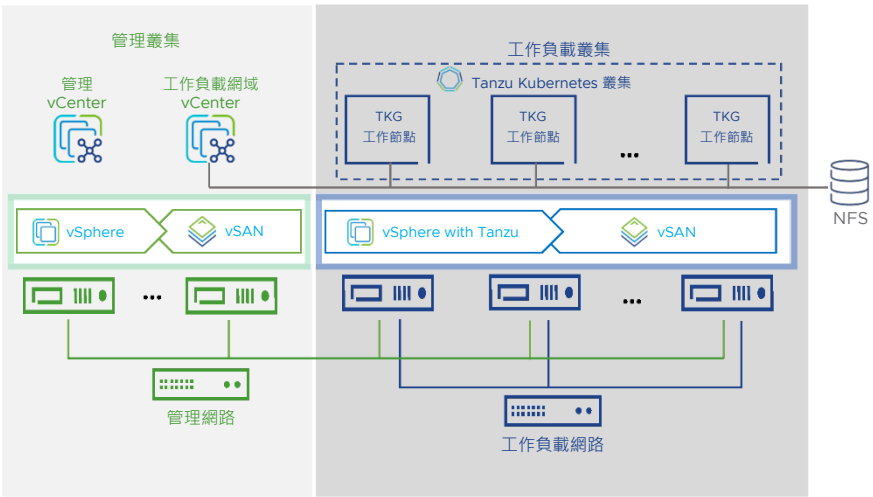
下方提供自基礎架構元件至 LLM 應用程式層的高層級解決方案概觀。應用程式層涵蓋下列開放原始碼機器學習元件：

Ray - 這款先進的開放原始碼分散式運算架構經過精心打造，可滿足現代化資料導向和採用人工智慧技術支援的應用程式需求。Hugging Face Transformers 資料庫 - 可與運用 PyTorch 深度學習架構且預先訓練的轉換器模型搭配運用，以提供專為生成式人工智慧工作流程精心打造的廣大資料庫和工具商業網路。企業可運用 Kubeflow，以大規模簡化雲原生環境的人工智慧模型建置和部署。



VMware Cloud 上的 LLM 參考架構

VMware Cloud 為新一代的多雲基礎架構即服務軟體，可協助 IT 作業為每個應用程式提供適當的基礎架構。這款解決方案能透過精心設計的架構，提供完整堆疊的企業級運算、網路、儲存、管理和安全服務。這款經過驗證且最佳化的基礎架構堆疊會運用 Tanzu Kubernetes Grid Service (TKGS)，提供適用於雲原生人工智慧和機器學習作業平台元件的 Kubernetes 基礎架構。TKG 能提供一致且可延展的 Kubernetes 體驗，藉此簡化容器協調作業的複雜性，並讓企業著重運用自身的人工智慧 / 機器學習工作負載來實現價值。



Copyright © 2023 VMware, Inc. 版權所有。VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com
11073 台北市信義區松仁路 100 號 台北南山廣場辦公大樓 38 樓之一 電話 +886-2-3725-7000 傳真 +886-2-3725-7010 www.vmware.com/tw
VMware 和 VMware 標誌是 VMware, Inc. 及其子公司在美國和其他管轄區域的註冊商標或商標。此處所提及的所有其他標誌和名稱，可能分別為其相關公司的商標。VMware 產品係受列於 http://www.vmware.com/tw/download/patents.html 之一或多項專利所保障。
文件編號：vmw-ds-temp-a4-word-2023 1/23

vSphere 人工智慧功能

使用最新發行的 VMware vSphere® 8 Update 1 企業工作負載平台，企業就能受惠於強化的管理員營運效率、適用於高階人工智慧 / 機器學習工作負載的強力效能提升，以及涵蓋整體環境的強化安全性。

vSphere 的每部主機最多可容納 8 個 GPU，現在能部署 NVIDIA NVSwitch 技術，以大幅提高大型人工智慧 / 機器學習工作負載的效能。這 8 個 GPU 或其子集都可配置給單一虛擬機。

裝置群組可讓虛擬機透過更簡單的方式，在 vSphere 8 中使用額外的硬體裝置。vSphere 8 正式版可支援網路卡和 GPU 裝置，但需採用相容的廠商裝置驅動程式，且取決於廠商版發行版本。NVIDIA® 將成為首個支援裝置群組的合作夥伴，將於近期提供相容的驅動程式。

裝置群組需使用現有的新增 PCI 裝置工作流程，以新增至虛擬機。vSphere DRS and vSphere HA 會感知裝置群組，並適時配置虛擬機，以滿足裝置群組的要求。

vSphere 可透過提高 GPU 利用率的方式，達到降低成本的目的，並透過在相同 GPU 上新增異質 vGPU 設定檔支援，減少 GPU 中的工作負載片斷化。有了這項功能，即可在 GPU 上部署不同類型的工作負載，例如 VDI 應用程式、運算應用程式和圖形應用程式等。

當中新增了適用於 Okta 識別身分同盟的支援、適用於 TPM 2.0 晶片伺服器的快速開機支援，以及使用虛擬 TPM 的虛擬機容錯，可望提高安全性。

Copyright © 2023 VMware, Inc. 版權所有。VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com
11073 台北市信義區松仁路 100 號 台北南山廣場辦公大樓 38 樓之一 電話 +886-2-3725-7000 傳真 +886-2-3725-7010 www.vmware.com/tw
VMware 和 VMware 標誌是 VMware, Inc. 及其子公司在美國和其他管轄區域的註冊商標或商標。此處所提及的所有其他標誌和名稱，可能分別為其相關公司的商標。VMware 產品係受列於 http://www.vmware.com/tw/download/patents.html 之一或多項專利所保障。
文件編號：vmw-ds-temp-a4-word-2023 1/23