2025 年 4 月刊 | 電子書

利用 Nutanix 推行企業 AI

路徑越簡單，速度越快

NUTANIX

目錄

快速推動企業 AI

生成式 AI（GenAI）的出現，促使企業重新思考 AI 規劃，並加速時程。

然而，在針對 650 位 IT 決策者、DevOps 與平台工程人員的最新調查中，許多組織仍不確定如何啟動 GenAI。主要顧慮包括：

- **IT 基礎架構不足：**91% 受訪者認為必須強化基礎架構以支援 AI 工作負載。
- **確保資料安全性與韌性：**資料安全、韌性與可擴展性是企業 AI 的關鍵挑戰。
- **技能缺口：**100% 的組織表示未來 12 個月需要新增技能以支援 AI 專案。

在部署 AI 前，可先評估以下三種一般路徑：

- **完全依賴雲端：**使用雲端進行 AI 實驗、開發和製作 AI 應用程式是一種選擇，但成本可能高、需與他人競奪雲端資源，且資料風險需審慎評估。
- **在資料中心和邊緣建立客製化 AI 堆疊：**需要高專業度與時間；若缺乏內部專才或可信賴夥伴，進度恐緩慢且易踩雷。
- **選擇統包式解決方案：**即使團隊經驗有限，也能以業界頂尖硬體與 AI 軟體快速上手，並提升彈性。

本電子書聚焦團隊在 AI 規劃時的痛點，並說明 Nutanix 雲端平台和 Nutanix GPT-in-a-Box 如何在地端、雲端與邊緣簡化 AI 上線路徑，更快達成更佳成效。

AI 痛點

大多數組織(85%) 計畫採購既有模型或使用開源模型建置 AI 應用程式，僅 10% 打算自建模型。

對多數企業而言，不必從零設計與訓練模型，更有效率的是授權並擴充預訓練基礎模型，配合你的私有資料，方法包括：

- **微調：**以私有資料微調模型，貼合需求。
- **檢索增強生成(RAG)：**引用模型外部資料，輸出貼合企業情境的結果。

無論採用哪種方式，從實驗走向生產都會在以下面向遭遇挑戰。

複雜性

要善用 GenAI，需能部署容器化 LLM、進行微調或實施 RAG，並以 MLOps 反覆上線到生產。

因此你可能需要適合微調與適合推論的不同基礎架構。推論常在邊緣執行(貼近使用者)，以降低延遲、提升回應。

- 但邊緣推論帶來遠端基礎架構管理挑戰。
- 推論與訓練(無論資料中心或邊緣)都需適當搭配運算、GPU 和儲存，以避免超額佈建並在需求變動時快速擴展。

為了確保 AI 營運萬無一失，你需要智慧化工具來判斷問題根因是模型、基礎架構或 MLOps 流程弱點。

合規性

AI 團隊常獨立於 IT，著重開發解法，卻忽略關鍵韌性、資料管理和個人識別資訊(PII)的保護。

在不損害 IT 政策對於安全性、資料保護、韌性和其他日常維運要求的合規性下，部署和營運 GenAI 或其他基於 AI 的應用程式，可能需要耗費極大的精力。

所有訓練資料必須去識別與清理，避免包含姓名、地址、電話、社會安全碼、財務與信用卡等個人識別資訊。

成本

雲端雖有優勢，但在雲端上運行模型的成本可能遠高於地端。

另一方面，AI 基礎架構耗能高是眾所周知，若在資料中心或邊緣運行 AI，需納入電力、散熱、空間等日常營運的因素。

治理

AI 還伴隨著多項治理挑戰。

- **資料管理與資料主權：**在邊緣、核心和雲端落實一致、原則驅動的保護與安全並不容易。邊緣產生的資料常需回饋訓練，因此需要簡單、可重複、可自動化的跨域資料移動與管理方法。

  有時邊緣資料必須存放與被保護在可信賴之處，以符資料主權規範。若需使用該資料訓練，可能要把模型移到資料端進行訓練。

  若有多個據點，可能需採用聯邦式做法，讓模型逐點訓練。

- **MLOps：**AI 團隊必須精準追蹤各模型版本，含訓練資料集、部署時間與位置等完整資訊。

  隨訓練迭代的頻率、資料集數量與複雜度提升，其難度往往高於傳統軟體版本控管。

安全性與資料隱私

AI 還會帶來額外的安全顧慮。

- **資料外洩：**將機密或專有資訊輸入非主權雲端 AI 服務的誘惑與風險。

  這可能看似無害，例如將未公開的報告丟給 Copilot、Gemini 或 ChatGPT 產生摘要，或用 OpenAI 的 Whisper 模型轉錄內部會議。

  但模型處理後的資料去向為何？即使像 Microsoft、Google、AWS 和 OpenAI 這樣的大型雲端商與供應商值得信任，但整個生成式應用程式和服務的生態系統興起仍提高了風險。

- **資料投毒：**惡意向 LLM 注入未經授權的資料，即便模型受你控管，仍可能會導致幻覺與偏差。
- **提示注入：**以惡意提示繞過防護、操縱模型回應。
- **安全工具：**既有工具可能無法銜接 MLOps，需升級或換新。

Nutanix 雲端平台採用創新軟體和獨特的超融合基礎架構（HCI），於地端、邊緣與公有雲全面解決上述挑戰。

為什麼選擇 Nutanix 雲端平台運行 AI？

Nutanix 專注於降低基礎架構複雜度，並實現混合多雲運作。Nutanix 雲端平台 (NCP) 建基於經驗證的超融合基礎架構 (HCI)，提供敏捷且具韌性的基礎架構，滿足從邊緣、核心資料中心到公有雲的 AI 需求。

簡化你的 IT 營運

營運複雜性是另一個通往 AI 實際執行部署道路上的主要挑戰，鑑於資源限制，許多組織在雲端進行大量的 AI 實驗。

但是，如何將在雲端中完成的工作移轉到資料中心，或將其在邊緣進行操作？

NCP 消弭基礎架構的複雜性與限制，讓你的 AI 與機器學習專案快速啟動。有了 Nutanix，你可以在邊緣、資料中心和雲端環境之間，實現應用程式、工作負載和資料的無縫移動。

由於 NCP 可於各環境運行，讓團隊在任何場域都能高效率運作。

NCP 於邊緣環境

邊緣因遠端部署與管理複雜基礎架構而特別具挑戰性。NCP 採用成熟的 HCI 設計，結合運算、網路和儲存，可特別滿足邊緣部署的需求：

- 小型化佔用
- 易於遠端管理與應用程式部署
- 進階資料保護與安全性
- 不受限連線的運作能力
- 完整資料服務

Nutanix GPT-in-a-box

許多企業在導入 GenAI 時受阻，特別是因資料主權、治理與隱私而無法上公有雲的使用案例。

Nutanix GPT-in-a-Box 提供即開即用、統包式的私人 AI，可讓你微調和運行 LLM 及其他 AI 模型，同時維持完全掌控。

它可化解 GenAI 應用程式在複雜度、擴展與安全上的難題，還具備全堆疊、軟體定義、AI 就緒特性，運行於 NCP 上，讓你的 AI 專案在部署和實施都能簡化又有高度效率。

GPT-in-a-Box 提供：

- **全堆疊 AI 平台：**自選硬體、CPU/GPU、虛擬機/容器和 LLM/框架，可於任何場域部署 GenAI。
- **內建 AI 資料服務：**涵蓋檔案、區塊和物件，具一致化快照與災難復原控制。
- **可擴充的 AI 工作負載：**以一致的平台服務統一雲端運作，可在任何地方提供可擴充的 AI，從而事半功倍。
- **即用型 LLM 供部署和調整：**存取經驗證的 GenAI 模型，從邊緣到雲端快速部署，同時加快實現價值的時間。
- **部署及運行安全的 AI 應用程式 API：**輕鬆建立基於角色的安全 API，將應用程式串接至你的 AI 模型。

透過 Nutanix GPT-in-a-Box，你可以維護內部 AI 訓練資料和模型，以符合安全、隱私和合規性要求，同時最佳化 IT 成本。

GPT-in-a-Box 涵蓋啟動 GenAI 模型所需的一切，你只需提供選定的基礎模型。

彈性的 GPU 與 CPU 選項

無論是推論、訓練或兩者兼具，正確的 GPU 和 CPU 都是 AI 成功的關鍵。

GPU

Nutanix 支援全系列 NVIDIA GPU，滿足各種需求。NCP 支援 GPU 直通，能在虛擬化或容器環境中高效使用 GPU。

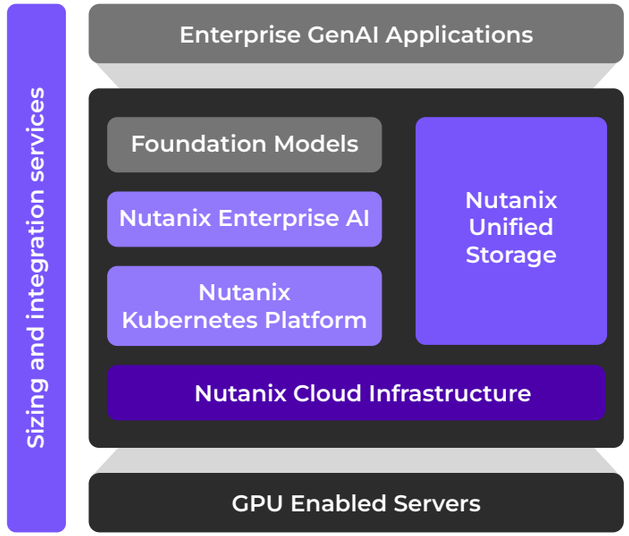
GPU 直通功能可讓在虛擬機器上運行的應用程式直接存取 GPU 資源。Nutanix 提供整個叢集的 GPU 檢視，可將可用 GPU 指派給任一虛擬機，也可以將多個 GPU 指配給單一虛擬機。採直通功能時，同一時間一顆 GPU 僅供一台虛擬機使用。

CPU

儘管多數人認為 GPU 對 AI 不可或缺，但新世代 CPU 已強化訓練與推論加速，可在無 GPU情境下支援部分 AI 作業。

合適的 CPU 也能提供必要的平行性、記憶體容量與頻寬，發揮 GPU 效益。

可提供的服務包括規劃工作坊、堆疊設計工作坊和堆疊部署。



AI 就緒服務套組，可協助容量規劃與配置，以主流開源 AI 和 MLOps 框架部署精選的 LLM 組。

產業合作夥伴

Nutanix 與業界領先的 AI 公司合作，提供全堆疊解決方案。

透過與 NVIDIA 合作，Nutanix AHV 已通過 NVIDIA AI Enterprise 認證，可運行其雲原生軟體平台，以精簡生產級 AI 開發與部署。

這涵蓋 AI 代理、電腦視覺、語音 AI 等多種情境。NVIDIA AI Enterprise 認證確保上述功能可於 NCP 上如預期運作。

NCP 可與 NVIDIA AI Enterprise 搭配使用，能形塑敏捷、高效率、可擴展的環境。與 NVIDIA GPU 運算加速器的整合，
可確保 AI 工作負載直通 GPU，降低延遲、提升效能。

除了 NVIDIA 之外，Nutanix 還與 Intel 和 AMD 深度合作，因而能緊貼這些領導廠商的 AI 技術創新，並快速在平台支援。

AI 使用案例

對於許多企業而言，諸如智慧型聊天機器人、支援 Copilots、智慧型文件處理等 GenAI 使用案例為首要。Nutanix 提供多元選擇，涵蓋廣泛 AI 應用情境。

NVIDIA AI Enterprise 和 NVIDIA NIM

憑藉 NVIDIA AI Enterprise 認證，Nutanix 支援多元模型與應用情境領域，包括電腦視覺、語音、語意理解與分子生成。

對於 NVIDIA AI Enterprise 客戶而言，GPT-in-a-Box 可讓他們輕鬆部署 NVIDIA NIM（針對 GenAI 最佳化的雲原生微服務）。

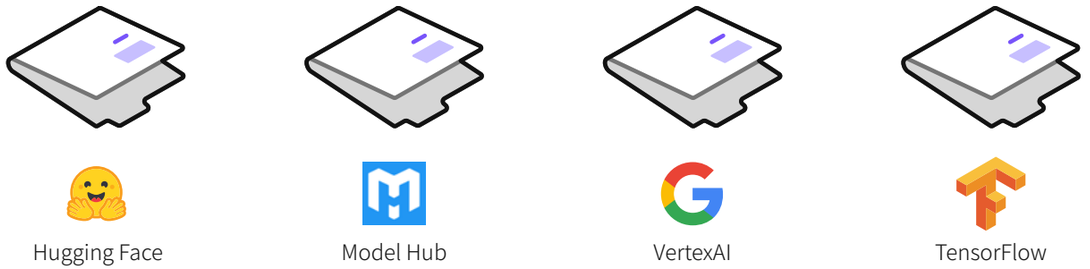
Hugging Face 和其他 Model Hub

Nutanix 與 Hugging Face 合作，快速部署模型，讓你可以輕鬆將 Hugging Face 的 LLM 與 GPT-in-a-Box 整合。提供經驗證 AI LLM 在搜尋、下載和部署的無縫工作流程與完整支援。

GPT-in-a-Box 還可讓你上傳並部署未驗證或未支援的自選模型。

Model Hub

開源模型倉庫（寬鬆授權、便於使用）



立即開始體驗 Nutanix

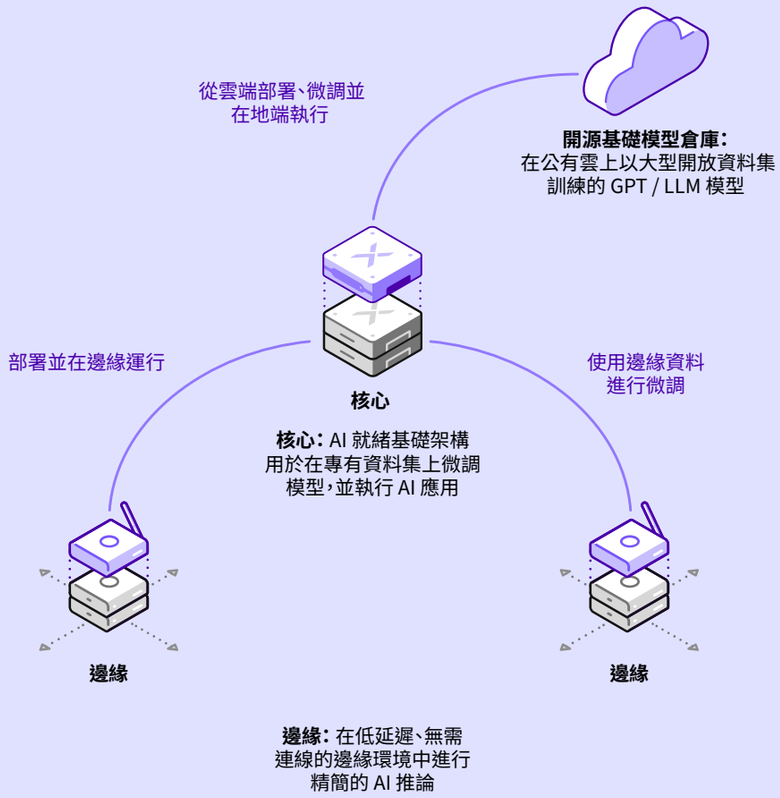
Nutanix 致力於消除痛點，讓你持續走在成功將 GenAI 等 AI 應用程式導入生產的路上。

NCP 從邊緣到核心到雲端，簡化 AI 營運與資料服務，並提供資料保護與安全，讓你安心前進。搭配 Nutanix 的即用型 GPT-in-a-Box 解決方案，導入 GenAI 更加容易。

想瞭解更多有關 Nutanix Enterprise AI 解決方案的資訊，請造訪 AI 解決方案頁面，並歡迎體驗我們的 AI 試用活動。

開始產品試用

簡化從核心、邊緣到雲端的 AI 營運



NUTANIX
info@nutanix.com | www.nutanix.com | @nutanix

檔名（含版次與日期）：AI-EnterpriseAIOnNutanix-eBook-FY25Q3-v2 04302025