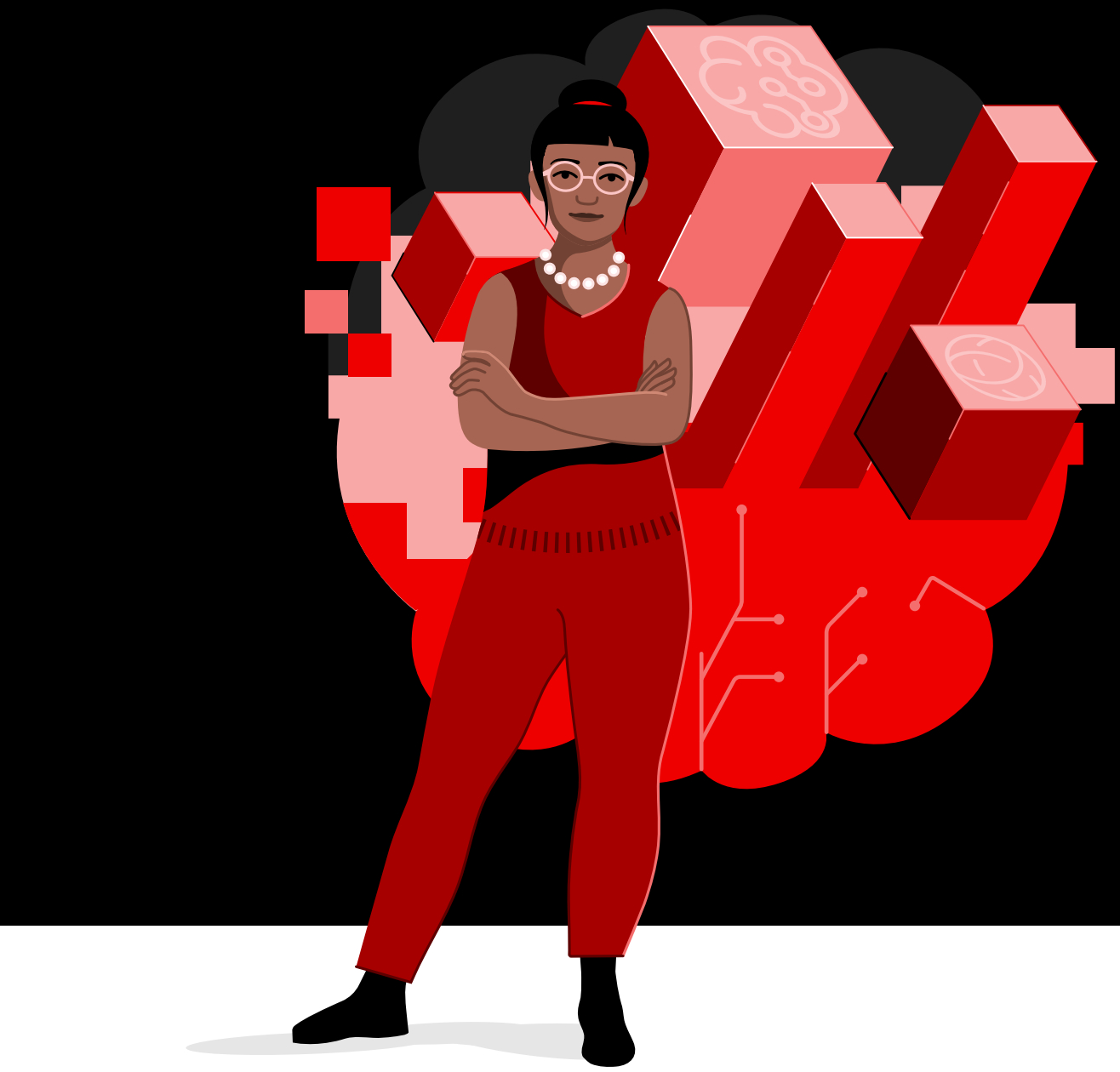
Red Hat



打造生成式 AI 基礎的

主要考量

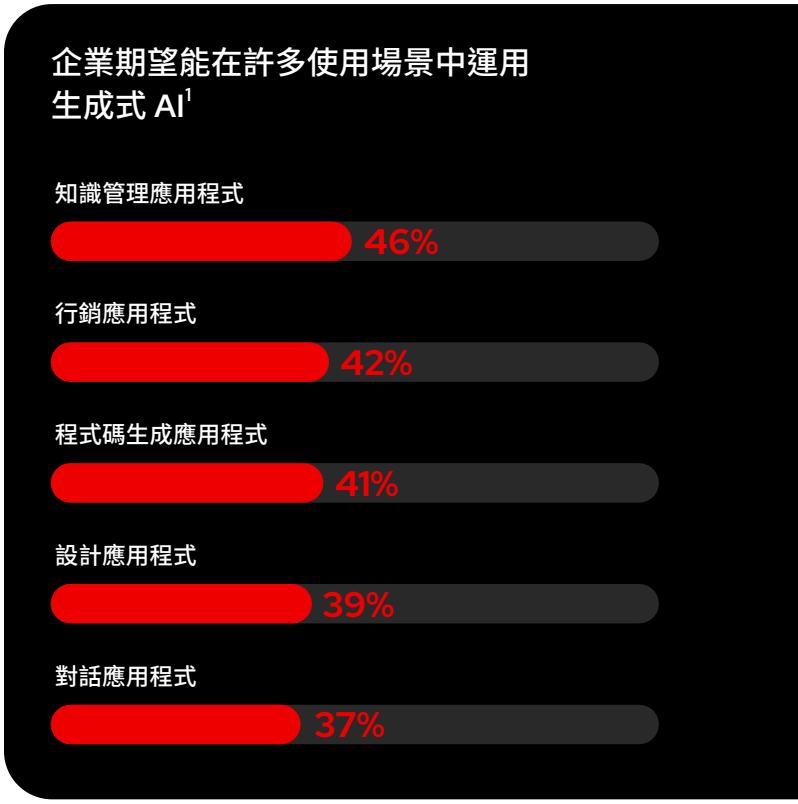
目錄

探索商業創新的全新可能性

生成式人工智慧 (AI) 是十分強大的工具，可以協助企業組織打造創新產品、將流程最佳化，並在快速變遷的市場中取得競爭優勢。這項技術奠基於深度學習和神經網路方面的進展，用途比預測型 AI 功能更廣泛，不僅能處理資料，還能產生全新的原創內容。生成式 AI 正在重塑人機協作的方式，催生解決問題的全新方法，並為各個產業帶來顯著的商業利益。

現在，全球各家企業紛紛開始利用生成式 AI 技術打造全新的創新應用程式。事實上，39% 的企業目前正在投資生成式 AI 技術，還有 37% 的企業在探索可能的應用情境。[1]當今生成式 AI 應用情境眾多，以下列出幾個例子：

- **針對複雜情境產生預測。**生成式 AI 可以分析歷史資料、辨識模式，並做出準確的預測，來輔助策略規劃與風險管理。
- **設計個人化行銷。**生成式 AI 可經由分析資料掌握客戶偏好與行為，進而製作出能達到最佳互動率和轉換率的個人化行銷素材，包括電子郵件、廣告和促銷活動。
- **將客戶服務自動化並個人化。**生成式 AI 是智慧聊天機器人和虛擬助理的基礎，可自動回應客戶問題並與客戶互動，提供個人化且有效率的客戶服務。



1 IDC Web Conference Proceeding，《Unlocking Business Success with Generative AI》 (運用生成式 AI 取得商業成果)，文件編號 US50789223，2023 年 6 月。

生成式 AI 帶來新的顧慮

儘管生成式 AI 的優點和缺點尚未完全確定，許多企業仍想要立刻開始投資這類新技術。瞭解與生成式 AI 相關的問題，能協助企業建立清楚的倫理指引和開發架構、遵守政府與產業規範，以及發現並修正潛在問題。

- **資料隱私權。**若使用敏感或個人資料來訓練生成式 AI 模型，就會產生與隱私權相關的疑慮，並引發涉及保護個人隱私權的問題。
- **資料所有權。**使用專有模型 (或是以專有資料訓練的模型) 則會面臨資料所有權問題，而這可能會引發訴訟。
- **偏見與公平性。**生成式 AI 工具產生的回應經證實會反映人類的偏見，包括有害的刻板印象和仇恨言論。
- **符合倫理的用途。**生成式 AI 模型可能會產生合成內容和深度偽造內容，並且會用於執行侵犯隱私權和假資訊作戰等惡意活動。
- **可解釋性與可詮釋性。**生成式 AI 工具不夠透明，因此難以詮釋、理解和解釋模型的產出結果，導致無法對錯誤或造假資訊問責。
- **意料之外的後果。**生成式 AI 具有自主性，可能導致意料之外的後果，對人員和企業造成實質傷害。
- **法規挑戰。**生成式 AI 技術的快速發展可能超越法規框架，以至於難以建立和落實相關指引，來確保以負責任且符合倫理的方式使用。
- **能源消耗。**訓練 AI 模型是運算密集型工作，需要耗費大量電力，會引發有關環境影響和永續發展的疑慮。

本電子書詳細分析在打造可信賴的基礎架構，以支援生成式 AI 計畫時，所需要考量的重要事項。

為生成式 AI 做準備

在《用生成式 AI 取得商業成果》 中，IDC 建議企業在推動生成式 AI 計畫時採取以下行動來做好準備。[2]

- **打造合適的環境，**以針對符合企業需求的優先應用情境靈活進行實驗。
- **研擬企業政策，**確保以負責任的方式使用，防範惡意行為。
- **評估生成式 AI 對員工的影響，**並主動進行變更管理。
- **與信賴的技術廠商**和服務供應商合作，建立您的 AI 基礎架構。
- 透過聘僱、訓練或專業服務支援，**確保員工具備合適的工程技能。**

2 IDC Web Conference Proceeding，《Unlocking Business Success with Generative AI》 (運用生成式 AI 取得商業成果)，文件編號 US50789223，2023 年 6 月。

打造生成式 AI 基礎的考量

您為生成式 AI 計畫所選擇的技術基礎，會決定採用 AI 技術的難易度以及整體成效。本章要討論生成式 AI 基礎的重要考量。

考量 1：運用經認可的工具組進行開發

以生成式 AI 模型為基礎開發應用程式是很複雜的工作。合適工具組的語言、架構和執行時間都應該以開放原始碼專案和商用解決方案為基礎，才能加速模型微調，並簡化應用程式開發和部署作業。

請選擇可提供您偏好工具組的 AI 基礎，以迅速且有效率地開發創新 AI 解決方案。透過互動式介面輔助探索式資料科學、訓練與微調，可以簡化協作。預先整合的工具組和自助服務功能有助於簡化 IT 作業，同時維持不同環境間的可攜性與一致性。

考量 2：迅速微調模型

由於訓練生成式 AI 模型的流程昂貴且耗時，大多數企業會使用基礎模型 (也就是以一般用途資料預先訓練的模型) 打造 AI 解決方案。接著，資料科學家會利用多種特定領域的資料來調整基礎模型，以執行專業工作。然而，微調仍是運算密集的工作，需要功能強大的處理器和分散式混合雲基礎架構。

您選擇的 AI 平台應該要具備分散式工作負載管理和協調功能，能在混合雲環境中部署任何模型規模、資料量或執行時間的多次訓練。選擇在現場資料中心微調基礎模型，可簡化遵循受管制模型相關技術和法規要求的作業。批次訓練功能讓您可以預先佔用微調工作負載，以更輕鬆地共用和管理資源。

微調模型的替代方案

研究人員正在尋找更快且更有效率的方式調整基礎模型。**檢索增強生成 (RAG)** 是一種 AI 架構，可以從外部資源 (如內部資料庫、企業內部網路或網際網路) 檢索事實資料，以便向生成式 AI 模型提供最準確的最新資訊。

在**提示調整**中，AI 模型會收到提示或前端提問 (例如額外的字詞或 AI 產生的數字)，引導模型做出理想決定，讓企業能運用有限資料，為極為專門的工作量身打造基礎模型。

考量 3：有效率地提供模型

對於 IT 營運團隊而言，以生成式 AI 解決方案提供絕佳的使用者體驗並不容易。為滿足多種應用程式需求，會需要可擴充的基礎架構和自動化管理。想要有效率的部署模型，需要具備監控效能的能力，還需具備迅速回復至先前版本的能力。由於 AI 解決方案會需要處理大量資料，因此有必要在各個環境中落實嚴格的安全性標準。

請優先考慮選用可以跨混合雲部署和擴充生成式 AI 模型與應用程式的平台，包括現場基礎架構、公有雲資源以及邊緣裝置。選擇從現場或獨立環境提供生成式 AI 模型，確保專有資料不會被用來重新訓練公開使用的模型。此外，支援金絲雀式推出 (canary rollout) 作法，並提供可解釋性工具，也有助於提升模型回應的一致性與可靠性。

考量 4：將生命週期管理自動化

持續整合/持續交付 (CI/CD) 流程可以自動部署和管理生成式 AI 解決方案。透過快速的漸進式變更來重新訓練和更新模型與應用程式，有助於加速開發模型，並提升模型效能。然而，AI 流程比標準 CI/CD 工作流程更複雜，因為其中經常涉及額外階段，如資料擷取、訓練、微調、驗證和重新訓練。

您選擇的基礎應該要讓您能以 Tekton 和 Jenkins 等 CI/CD 工具建立 AI 流程，並整合至現有的 DevOps 工作流程，以迅速且有效率地開發、訓練、監控和重新訓練生成式 AI 模型。**GitOps** 持續交付工具 (如 ArgoCD) 讓您能以程式碼的形式定義複雜的 AI 解決方案部署，並加以自動化，確保維持交付一致的模型和應用程式。

生成式 AI 容器

容器和 **Kubernetes** 技術提供靈活的部署、管理和可擴充性，加速生成式 AI 解決方案的雲端原生開發作業，也可在現場資料中心、公有雲和邊緣裝置佈建所需的環境。在實體和虛擬基礎架構自動建立、部署、擴充和管理容器執行個體，並將來自可靠的開源和商業供應商生態系的元件與資料儲存體，整合至生成式 AI 解決方案。深入瞭解 **AI 容器的優勢**。

考量 5：持續監控模型

生成式 AI 模型會對人員和企業造成大規模實質影響。經由追蹤模型行為，您可以分析決策和理由，辨識出不佳效能，並立即回報有問題的行為。根據上述的資訊執行有效的模型治理，有助於確保模型能夠使用沒有偏見、公平且正確的資訊，在生產環境中進行回應。

探索 AI 基礎時應採用集中式的監控功能。其中的偏見與資料偏移指標、異常偵測以及每點可解釋性，可協助您研究、維護和修正生成式 AI 模型。在生產環境中持續自動監控，可促進企業模型治理標準合規性。此外，易於使用的工具介面以及人類可判讀的非技術性報告，也有助於以負責任的方式使用和維護模型。

生成式 AI 模型重要概念

- **偏見**指的是模型行為有特定的模式，影響到產生結果的公平性、包容性和倫理，包括偏袒特定族群，或產生符合刻板印象的回應。
- **資料偏移**會在訓練資料的統計性質隨時間產生變化時發生，導致模型效能降低，以及產生較不準確或不相關的回應。
- **異常偵測**指的是辨識模型行為是否異於或偏離訓練期間的範例，並加以回報的流程。
- **每點可解釋性 (Pre-point explainability)** 指的是釐清為何模型產生特定結果的能力，有助於掌握須顧及透明度的應用程式。

考量 6：善用合作夥伴生態系

生成式 AI 解決方案需要多種經過整合的元件，才能順利提供創新的使用者體驗。適當結合來自可信賴廠商的協作生態系技術後，您可以加速開發應用程式、解決偏見與資料偏移的難題，並且確保整個解決方案的一致性和可靠性。

合作的平台廠商應該要具備龐大且經認證的合作夥伴生態系，有能力提供完善的解決方案以開發和部署生成式 AI 模型與應用程式。從資料整合和準備到模型訓練與提供模型，有多種元件可協助您更迅速和有效率地開發並部署 AI 解決方案。只要選擇經證實有互通性的解決方案，就能減少 IT 支援要求並提升生產力。

考量 7：與平台專家合作

要有效部署和管理生成式 AI 解決方案，需要專業知識與經驗。對可擴充性的要求、對可靠性的顧慮，以及整合至現有系統等因素，都會使得生產部署變得更加複雜。若運算資源使用效率不彰，可能會造成不必要的成本。如果不遵循安全性標準、隱私權政策以及 AI 法規框架，則可能導致意外的後果。

因此請選擇擁有專家團隊的廠商，才能獲得打造生成式 AI 解決方案的全方位支援和指引。例如，專屬工程師可以運用工具、資源和知識來支援整個平台，加速您的 AI 專案。專家顧問可以解決部署時碰到的挑戰，將基礎架構效率最佳化，並確保整個 AI 解決方案的互通性。此外，專業訓練服務可協助您獲得相關知識和專業技能，以更迅速地開始推動新的生成式 AI 專案。

生成式 AI 需要協作

建立具備多樣能力的團隊，是順利推動生成式 AI 專案的關鍵。[3]

- **企業高層**代表的是會使用解決方案或受其影響的對象。
- **AI 專家**負責調整、維護和更新生成式 AI 模型。
- **資料科學家**負責預先處理資料，並使用正確且無偏見的資料來訓練模型。
- **倫理與合規主管**負責確保生成式 AI 計畫符合法規。
- **IT 營運專家**負責將解決方案整合至現有的基礎架構，並落實安全性原則。

3 Kearney，《Standing up tiger teams to tackle generative AI complexity》(組成臨時任務小組以解決生成式 AI 的複雜問題)，2023 年 11 月。

運用彈性且開放的基礎迅速創新

Red Hat 提供完整的技術產品組合、經實證的專業知識以及策略性合作計畫，可協助您實現生成式 AI 目標。我們提供可用於開發和部署生成式 AI 模型與應用程式的基礎，也提供相關服務與訓練，協助您快速採用 AI 技術。

Red Hat® OpenShift® 是雲端原生創新應用程式統一平台，可供企業立即使用。隨需運算資源、硬體加速支援，以及現場、公有雲與邊緣環境的一致性，可以讓您的團隊擁有成功所需的速度與彈性。採用 Red Hat OpenShift 後，您就可以建立自助服務平台，供資料科學家、資料工程師與開發人員使用，以迅速開發智慧應用程式。協作功能讓團隊能以一致的方式建立容器化的建模結果，並與同儕和開發人員共用。

Red Hat OpenShift AI 是建立在 Red Hat OpenShift 之上，提供全方位的平台，可用來建立、訓練、微調、部署和監控模型與應用程式，也能滿足現代生成式 AI 解決方案的工作負載與效能要求。團隊可以在一致的協作環境中迅速從實驗階段進入生產階段。該協作環境整合了由 NVIDIA、Intel、Starburst、Anaconda、IBM、Run:ai 和 Pachyderm 等合作夥伴所提供，經認證的關鍵服務。Red Hat OpenShift AI 結合我們的技術生態系，可提供各項元件和功能，讓您在混合雲加速開發並部署創新生成式 AI 解決方案。

IBM watsonx.ai AI studio 提供一系列包含生成式 AI 功能的模型和部署選項，可滿足您的智慧應用程式需求。無論是開放原始碼、第三方還是 IBM 開發的基礎模型，您都可以在工作負載所在之處部署模型，以提升 AI 解決方案的效能與效率。採用由 IBM 開發，並以符合企業需求之資料進行訓練的 **IBM 基礎模型**，生成式 AI 解決方案就能理解您所屬商業領域的細節，讓您掌握競爭優勢。

Red Hat Ansible® Lightspeed with IBM watsonx Code Assistant 是一項生成式 AI 服務，可協助團隊更有效率地建立、採用和維護自動化內容。連結 IBM watsonx Code Assistant 的 Red Hat Ansible Lightspeed 可協助您使用自然語言提示，將自動化構想改寫為 Ansible 程式碼。透過這項服務，您不僅能提升生產力，還能讓自動化功能在企業組織內更加普及。

準備好開始使用生成式 AI 了嗎？

生成式 AI 是一種功能強大的工具，不僅能建立原始內容，還能改變我們與應用程式和技術互動的方式。

透過技術、專業知識和合作夥伴關係，Red Hat 可為您的團隊提供通用基礎，以建立並部署透明且易於控管的 AI 應用程式與 ML 模型。事實上，我們甚至會使用自家的 AI 工具和平台，來提升其他開放原始碼軟體的實用性。此外，我們的合作夥伴集成能讓您使用各種可信賴的 AI 工具，並與 Red Hat OpenShift AI 等開放原始碼平台共同作業。

瞭解詳情並免費試用 Red Hat OpenShift AI。

Red Hat 諮詢服務能協助您更快開始使用

與 Red Hat 專家合作，迅速展開您的 AI/ML 專案。Red Hat 提供諮詢和訓練服務，可協助企業在更短的時間內採用 AI/ML。

- 深入瞭解 AI/ML 服務：red.ht/aiml-consulting
- 與我們的專家免費諮詢：redhat.com/consulting

1057715_0324_KVM

Red Hat