Red Hat與AMD攜手合作打造企業AI應用平台

Joe Yu
首席架構師經理
Red Hat, Taiwan

Red Hat

世界發生了什麼事？

Chat GPT

Ansible Lightspeed with IBM **Watson** Code Assistant

watsonx

GitHub Copilot

生成式 AI

一組新的模型可以從自然語言輸入中產生一些東西

特別是LLMs

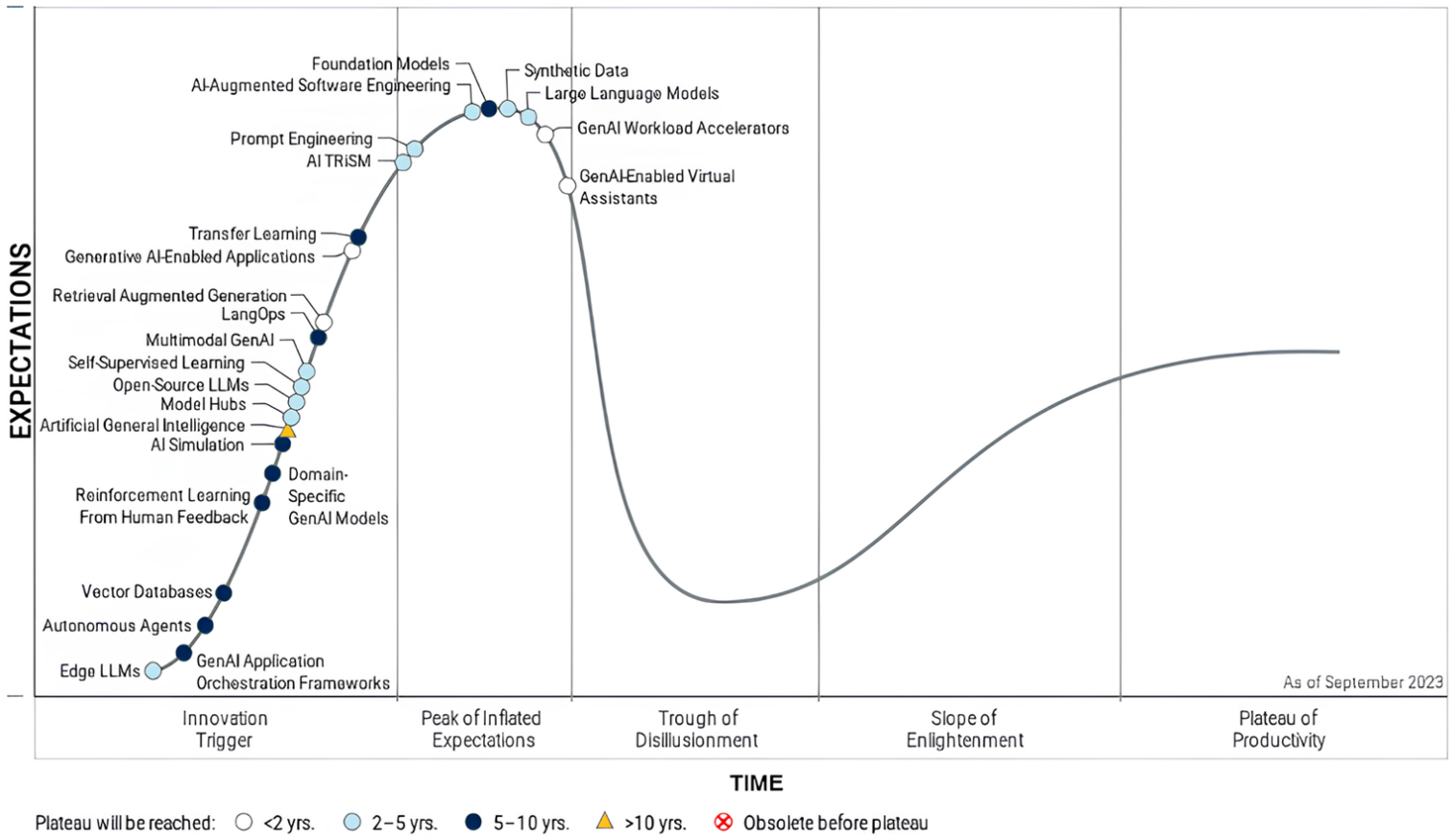
收集網路規模的文字資料集來訓練能夠令人信服地預測成文字的東西

Red Hat OpenShift AI (RHOAI)

Red Hat 新解決方案名稱

Red Hat

隨著時代的興起，AI的導入歷程



Gartner

Gartner

*“ More Than **80%** of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026. “*

https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026

Red Hat

AI/ML技術與方案組成呈現開源多樣化

AI + MLOPS 平台

應用程式平台

機器學習函式庫

開發語言與工具

資料視覺化、標籤與處理

實驗&模型生命週期

軟體定義存儲

整合

大規模容器管理

自動化軟體測試交付

容器化與資源編配

程序調配與硬體加速

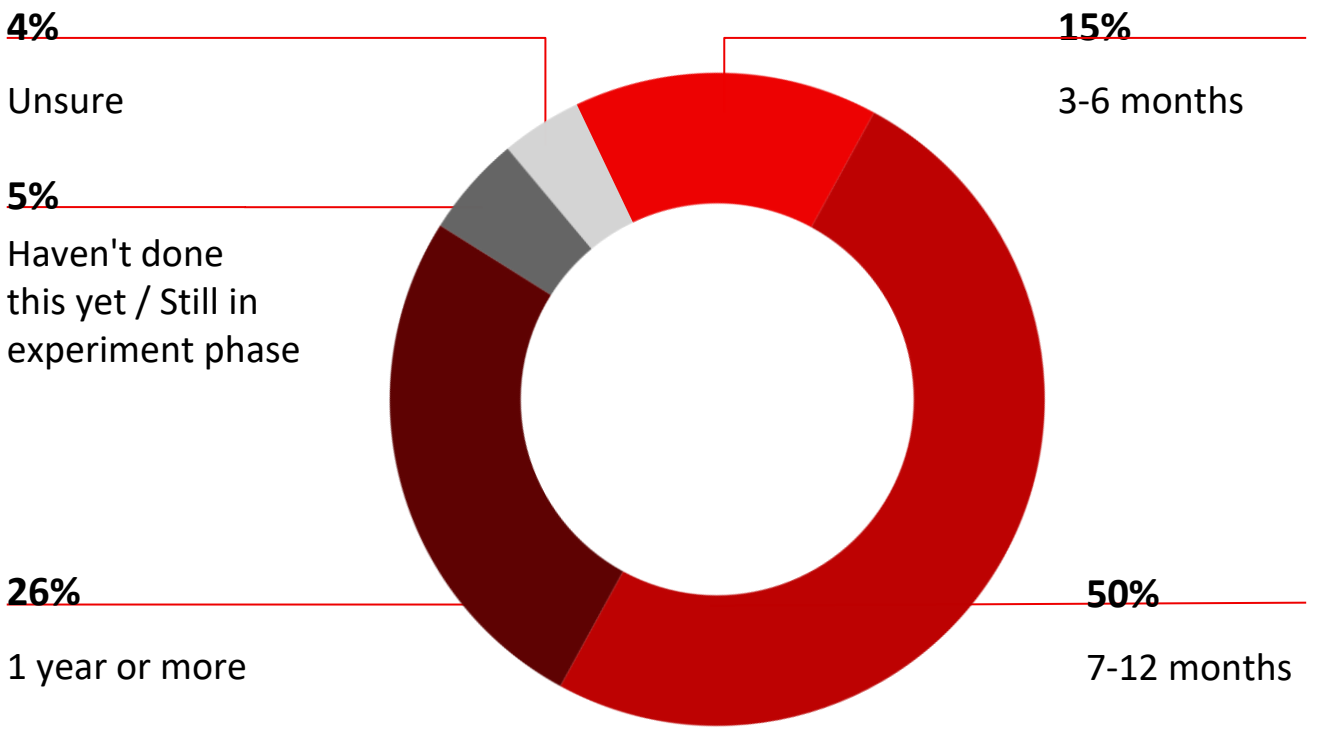
+ AI Accelerator

Red Hat

實施AI是一個充滿挑戰的過程

AI/ML 從想法到模型實施的時間平均需要多少？

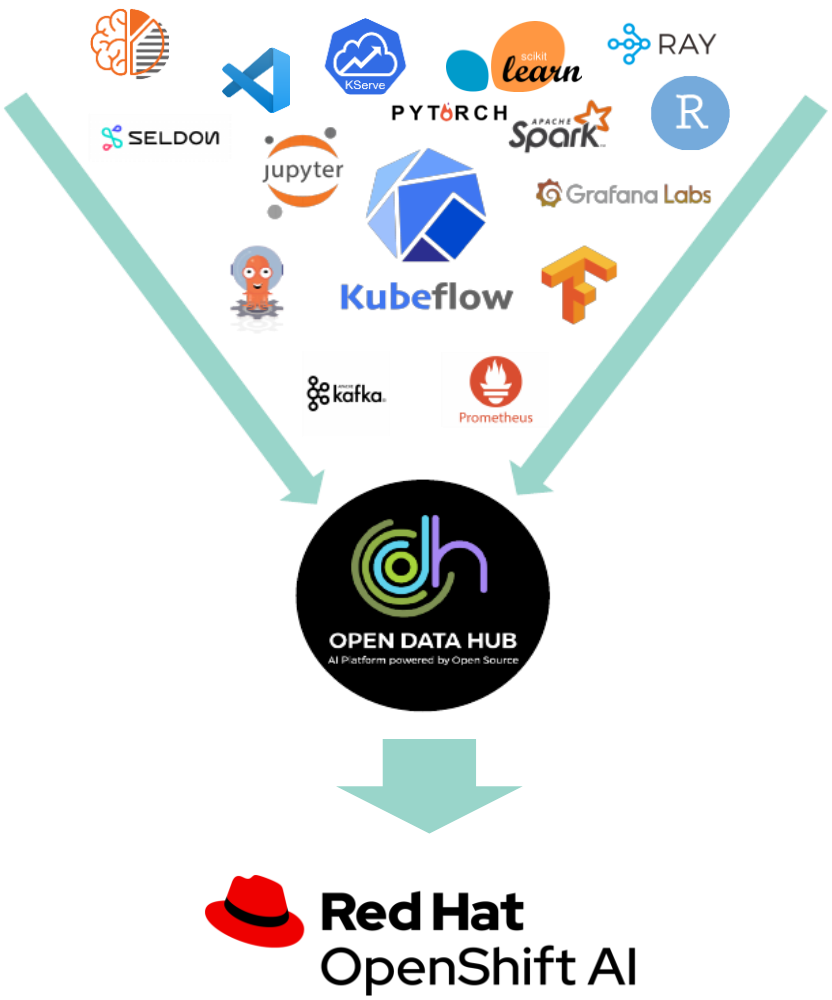
一半的受訪者 (50%) 表示，他們從構思到實施模型的平均 AI/ML 時間為 **7-12 個月**。



Source: Gartner Peer Insights, Open Source AI for Enterprise survey, 2023

Red Hat

讓企業可以專注 - Red Hat將 AI/ML的開源創新賦予企業支持



模型開發

使用Red Hat提供或自有的image，在 JupyterLab 中進行探索性實驗，並存取核心 AI/ML 程式庫和框架，包括 PyTorch 和 Tensorflow

模型服務和監控

跨任何雲端、代管與自我維運的 OpenShift 以部署和擴展模型，並集中監控其效能。

生命週期管理

建立用於模型訓練和驗證的可重複資料科學pipeline，並將其與 DevOps機制整合，以便在整個企業中交付模型。

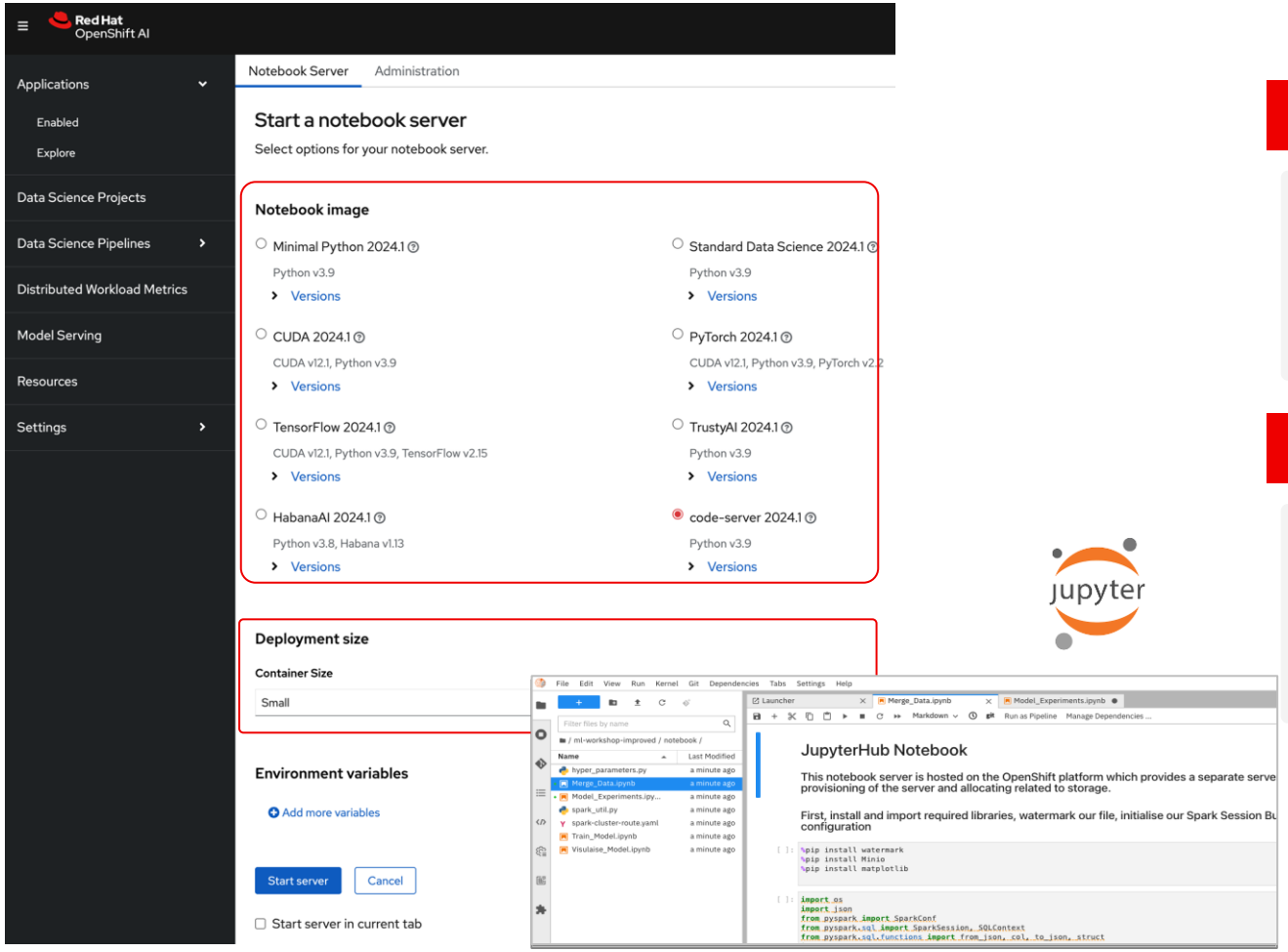
提昇能力/協作

建立專案並在團隊之間共用。結合Red Hat套件、開源以及 ISV 認證軟體。

Red Hat

RHOAI: 提供自助式的模型開發環境

透過 OpenShift AI 的平台，數據科學家自己可以申請和配置所需要的模型開發環境。



Notebook image

預設支援常用 ML 開發框架(Framework) 等容器映像檔。可透過創建自訂義映像檔*，您也可以新增所需的函式庫(Library)，或使用Jupyter Notebook 以外的編輯器 (e.g. VSCode)。

Deployment size

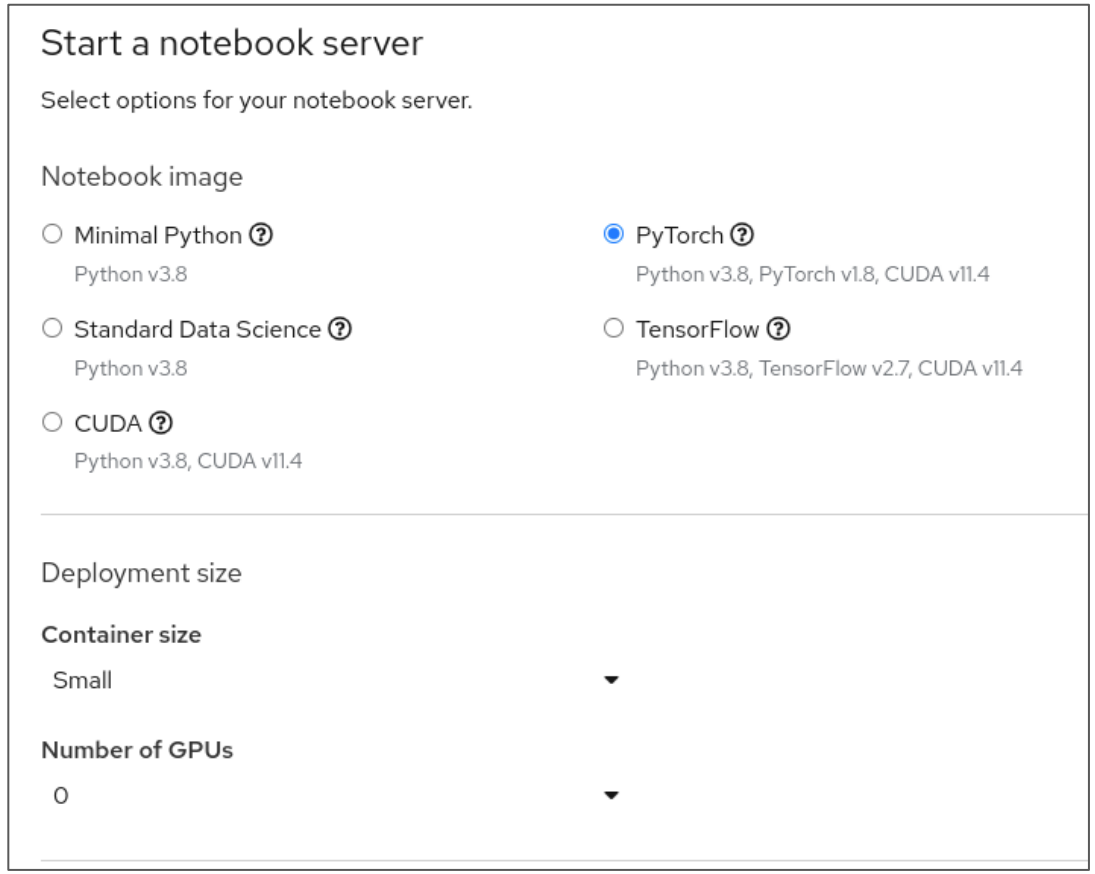
指定所需的資源(CPU/Memory) 來啟動容器

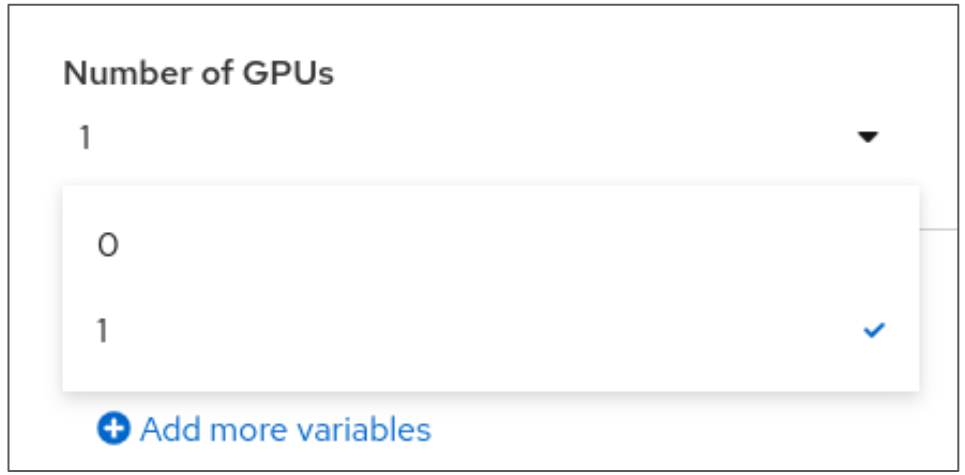
可以加入指定 GPU 數量資源

透過設定預設的停止時間，可以自動停止非活躍的環境並回收資源

*: 客製化的映像檔需要由客戶自行管理。

RHOAI: 充分使用OpenShift平台上的GPU資源

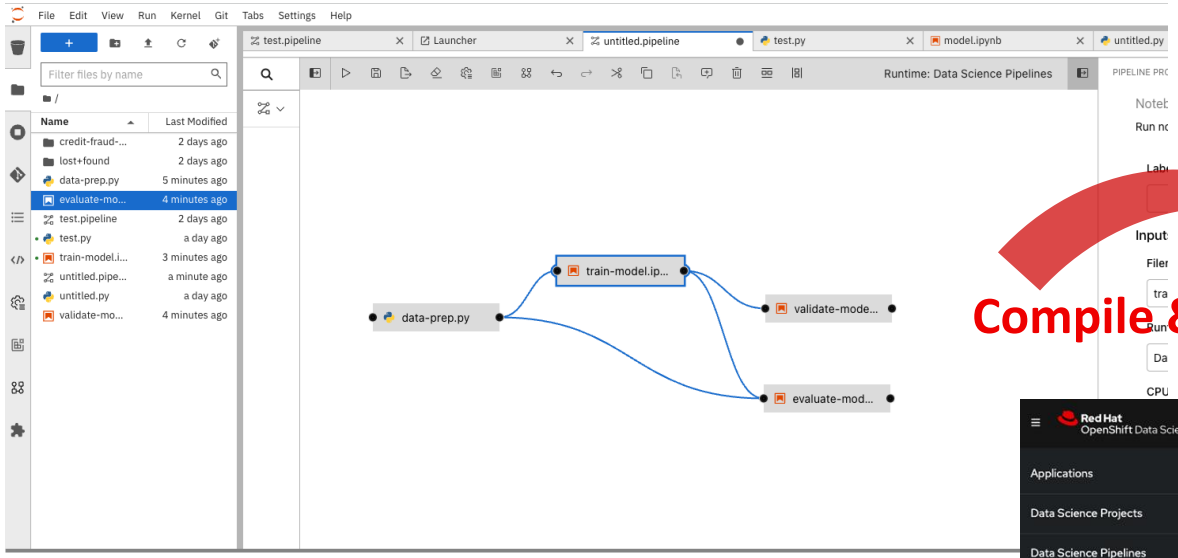




Red Hat

RHOAI: 支援透過流水線（Pipeline）實現流程自動化

把寫好的模型程式碼組合起來，建立一條流程管線（Pipeline），從而自動化多個和模型重新訓練或部署相關的步驟。



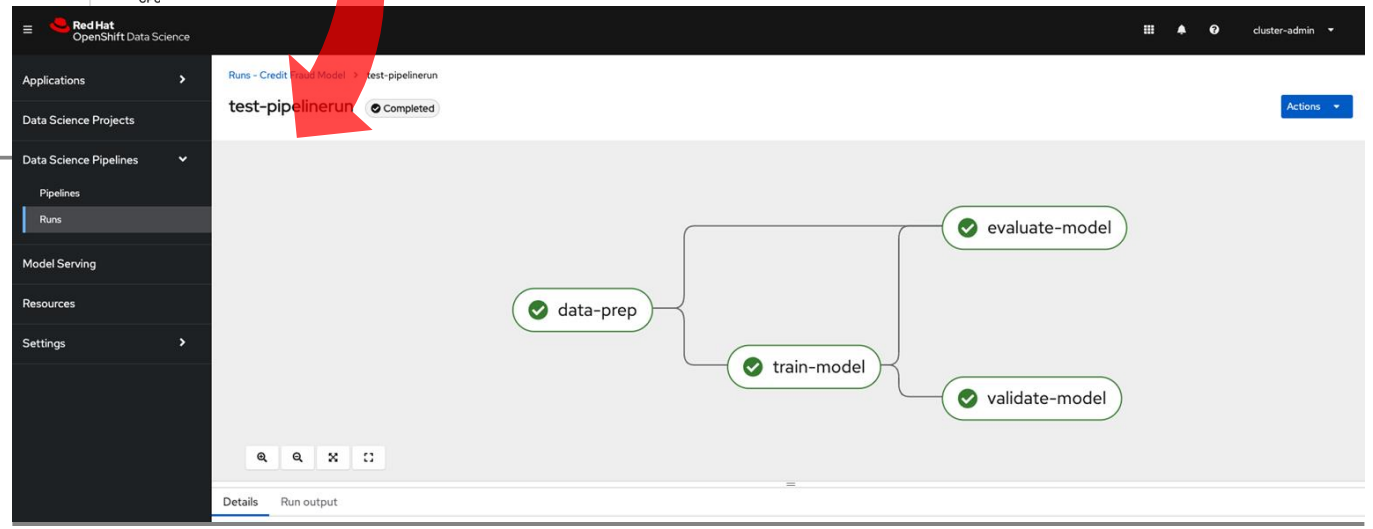
Pipeline Creation

你可以基於已有的 Python 腳本或 Notebook 檔案，透過圖形使用者介面（GUI）來組建流程管線。

Compile & Run

Process Automation

透過自動化流程管線的執行，可以極大地減少 AI/ML 模型重新訓練和部署所需的手動操作。



RHOAI: 簡化模型服務 (Mode Serving) 提供

透過 OpenShift AI Portal 網站部署已訓練的模型，我們提供能夠以API形式呼叫的終端點。

Deploy model

將儲存在物件儲存空間的已訓練模型，進行容器化後部署對應後續服務提供。

Inference endpoint

將模型轉換成 API 形式，並提供外部應用程式調用的推論(inference)端點。

Red Hat

Red Hat OpenShift提供最佳的應用程式平台支持並完整整合RHOAI

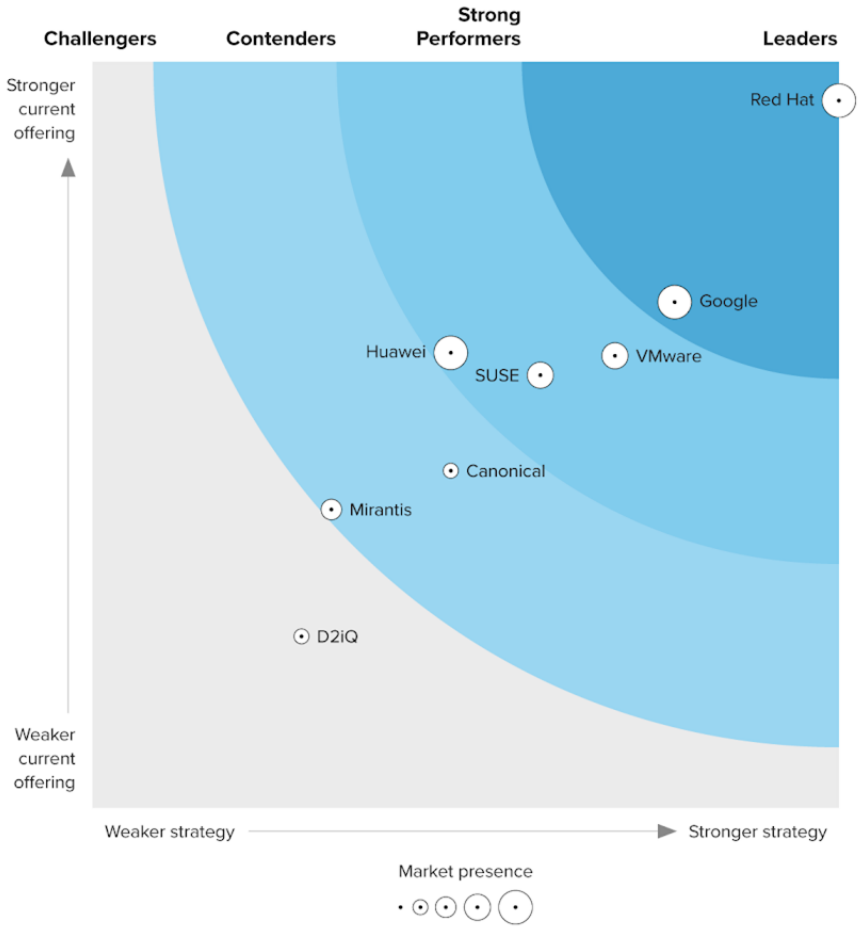
Figure 1: Magic Quadrant for Container Management

Red Hat OpenShift 在 2024 Gartner® Magic Quadrant™ for Container Management研究報告中

- 延續2023年，持續處於**領導象限**中
- 最具有願景(Vision)

The Forrester Wave™: Multicloud Container Platforms, Q4 2023

- 多雲容器平台**名列第一**

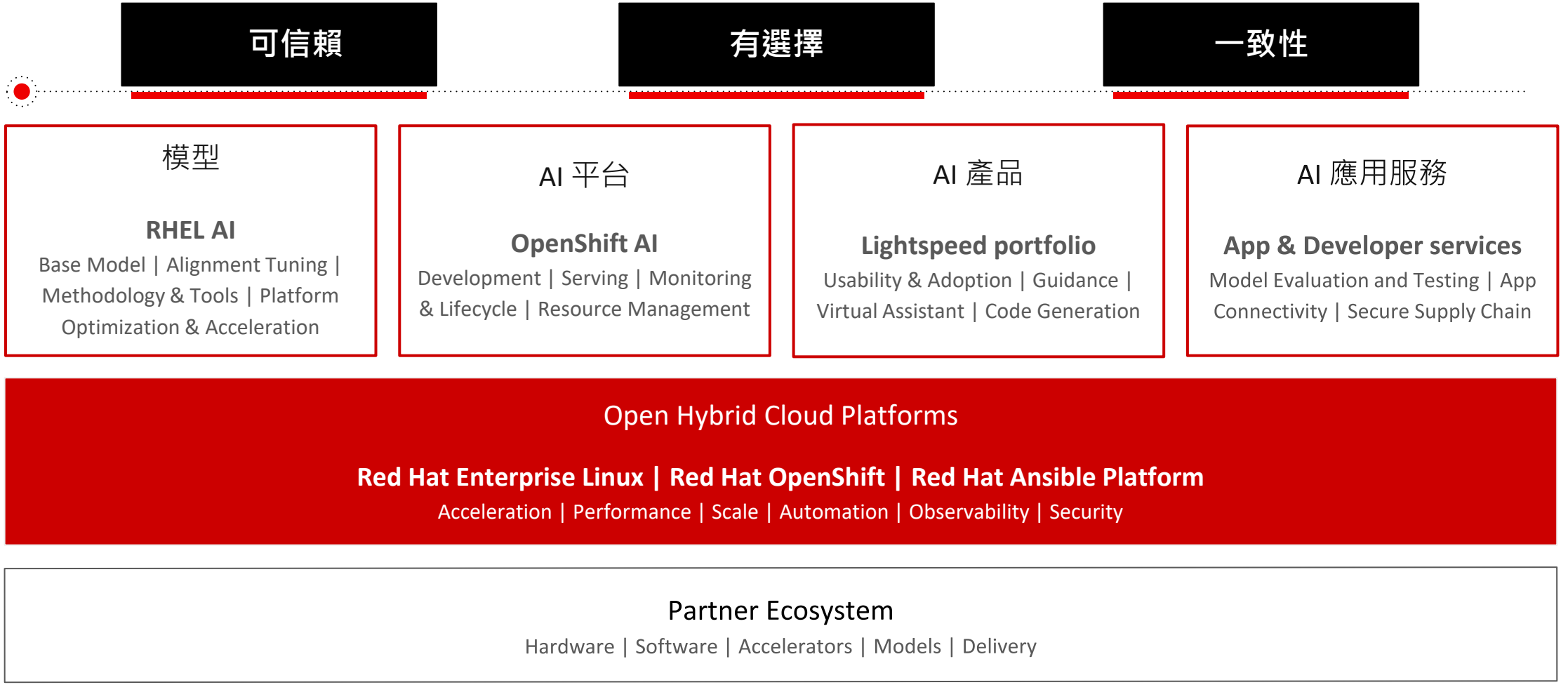


Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

The Forrester Wave™ is copyrighted by Forrester Research, Inc. Forrester and Forrester Wave™ are trademarks of Forrester Research, Inc. The Forrester Wave™ is a graphical representation of Forrester's call on a market and is plotted using a detailed spreadsheet with exposed scores, weightings, and comments. Forrester does not endorse any vendor, product, or service depicted in the Forrester Wave™. Information is based on best available resources. Opinions reflect judgment at the time and are subject to change.

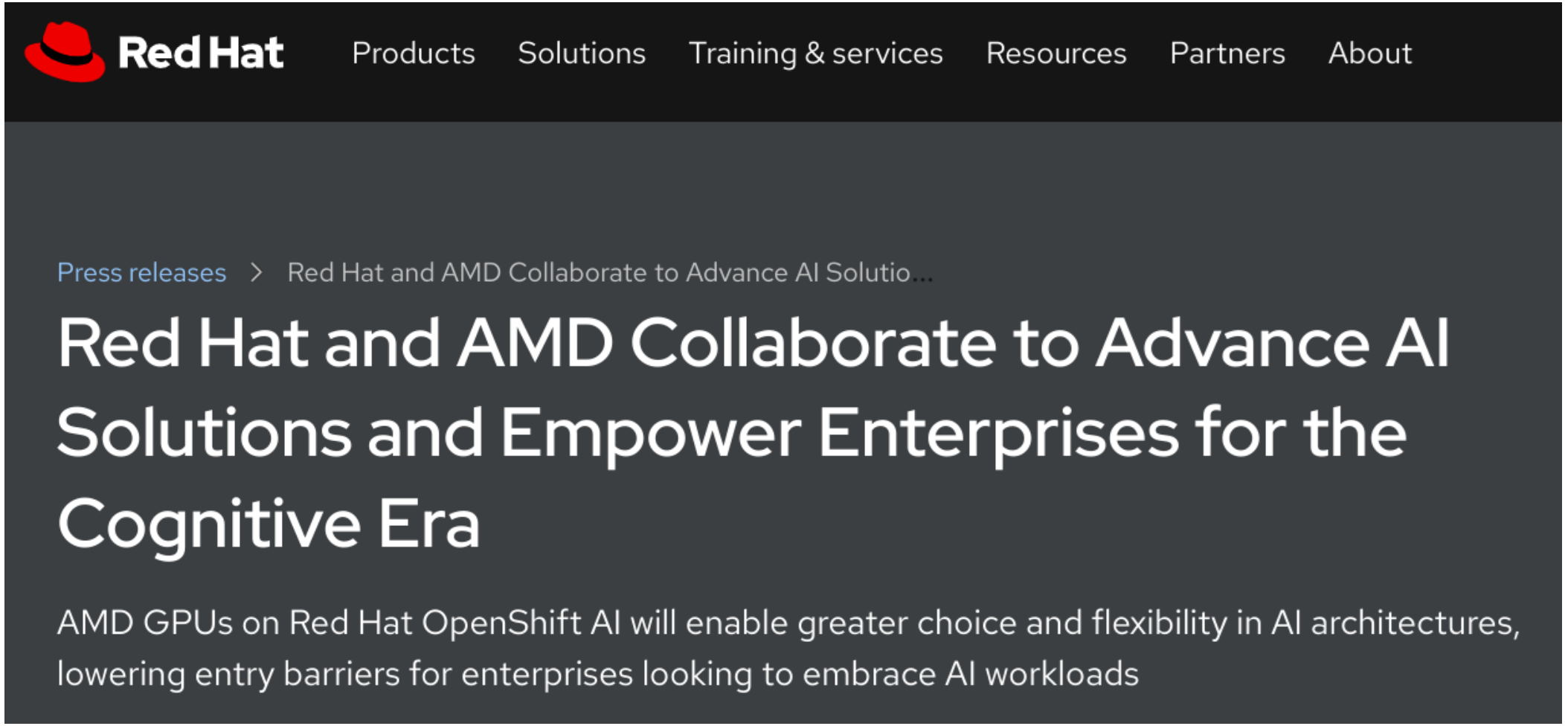
Gartner, "Magic Quadrant for Container Management," Dennis Smith, Tony Iams, Wataru Katsurashima, Michael Warrilow, Richard Watson, 10 September 2024

Red Hat AI的策略與服務 - 協助企業導入A的旅程



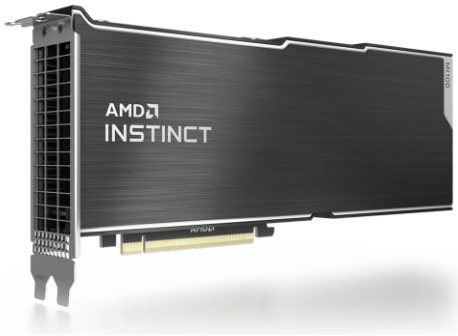
Red Hat

Red Hat與AMD攜手合作推動企業AI時代



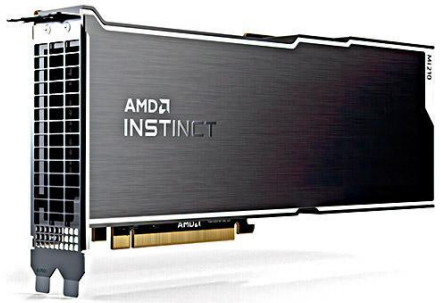
https://www.redhat.com/en/about/press-releases/red-hat-and-amd-collaborate-advance-ai-solutions-and-empower-enterprises-cognitive-era

AMD Instinct MI 系列 GPU 加速器



MI100

32GB VRAM
Arcturus GPU
7680 cores
7nm CDNA1
(2020)

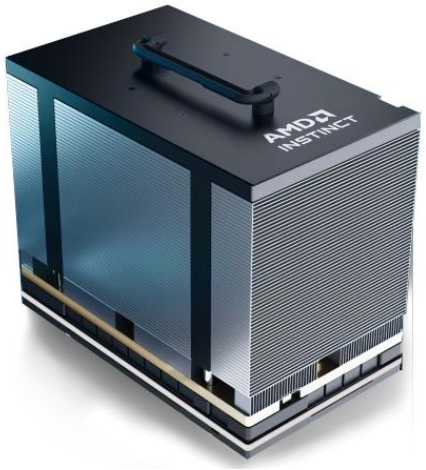


MI210

64GB VRAM
Aldebaran GPU
6656 cores
6nm CDNA2
(2022)

MI250

128GB VRAM
2 x Aldebaran GPUs
13312 cores
6nm CDNA2
(2022)



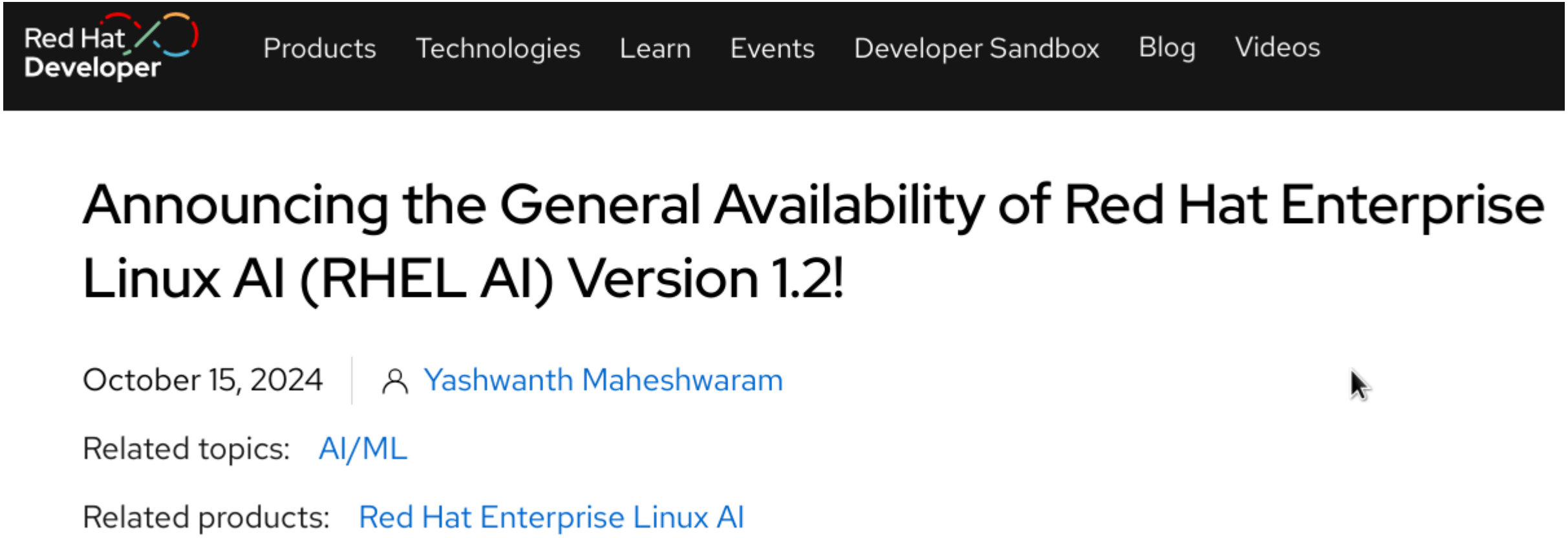
MI300X

192GB VRAM
MI300
14080 cores
5nm CDNA3
(2023)

Red Hat

RHEL AI平台支援AMD GPU

RHEL AI 1.2 現在支援 **AMD accelerators**

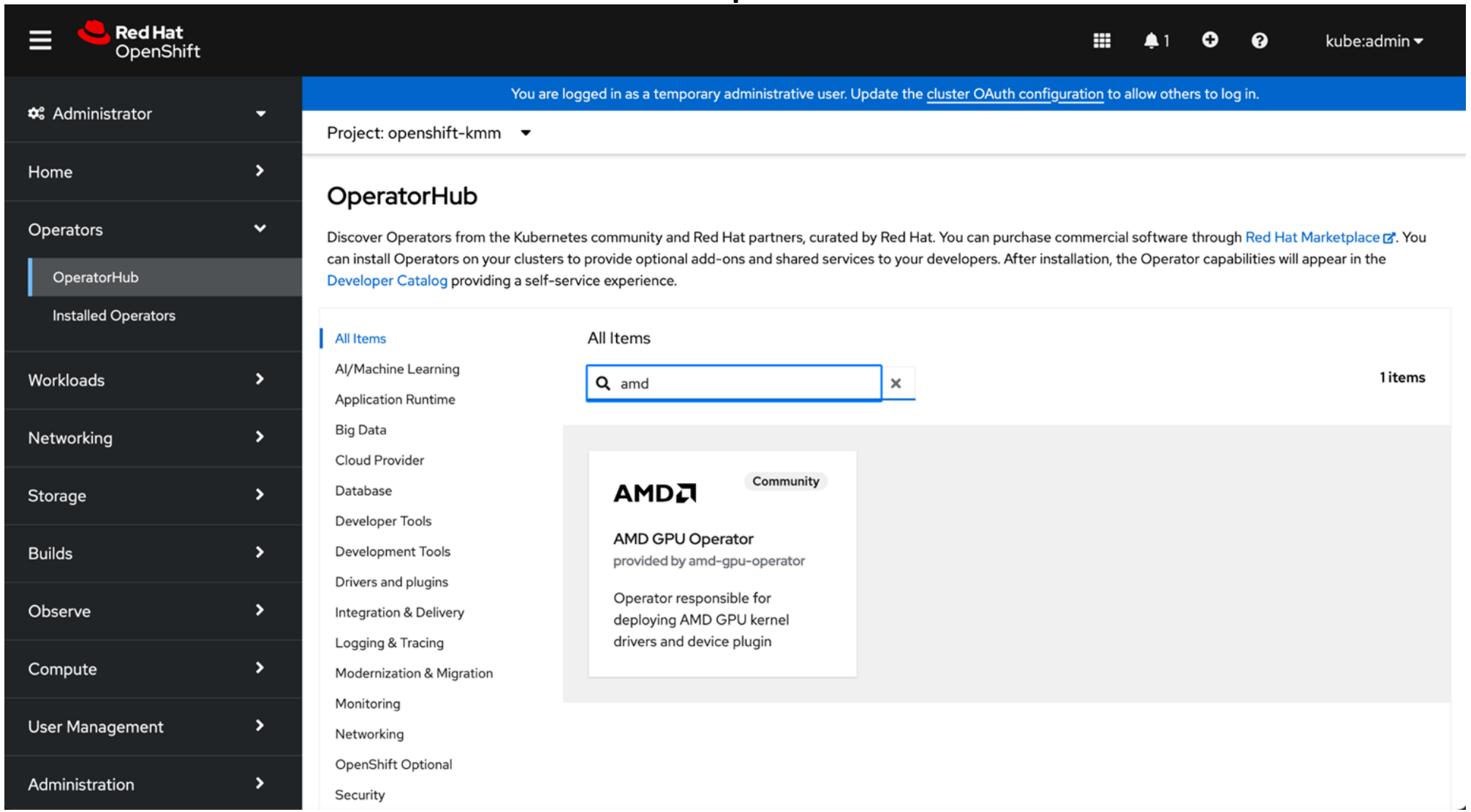


https://developers.redhat.com/blog/2024/10/15/announcing-general-availability-red-hat-enterprise-linux-ai-rhel-ai-version-12

Red Hat

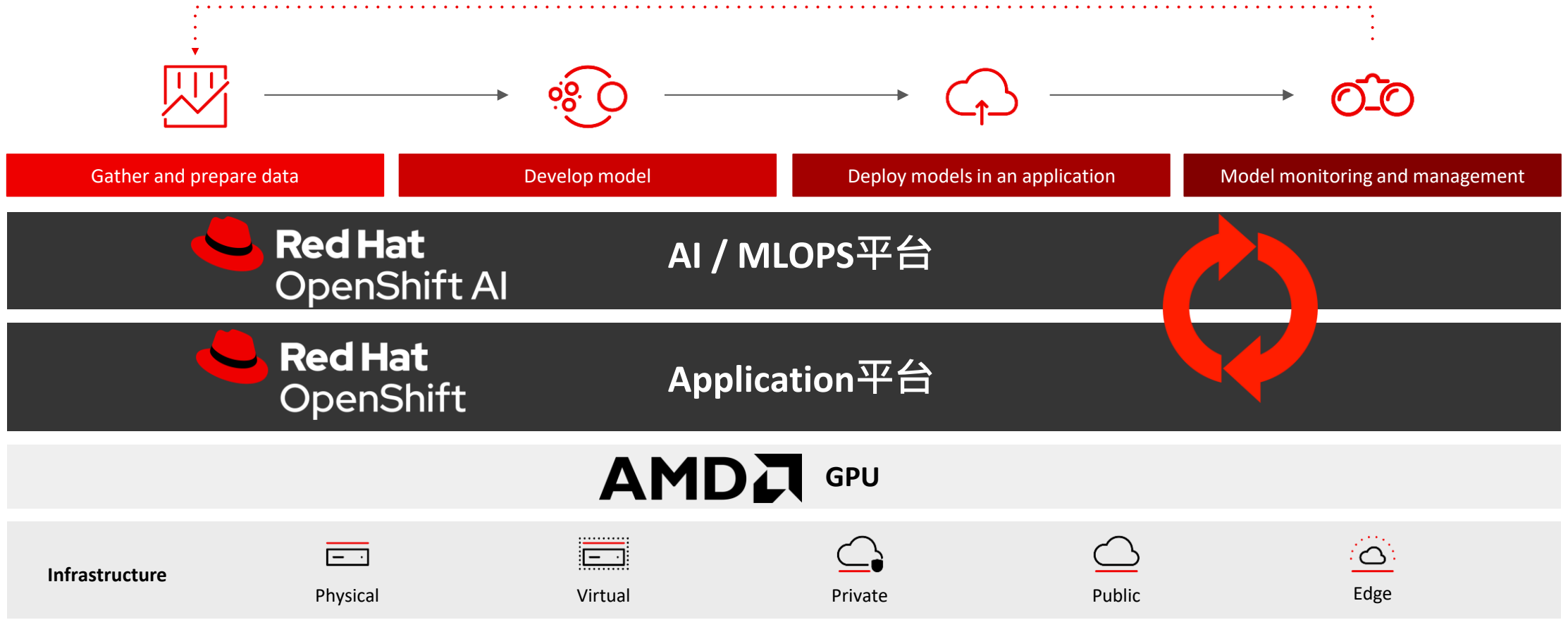
OpenShift平台支援AMD GPU

AMD GPU Operator Installation



Red Hat

Red Hat+AMD提供完整的應用程式平台+運算加速能力+AI MLOps平台



Red Hat

Thank You!

如果沒有全球合作夥伴生態系統的支持和積極協作，Red Hat就無法實現我們的AI願景。

AI並沒有“一套適合所有的標準（one size fits all）”，也沒有任何一家供應商能夠單獨滿足所有客戶的需求。

資料來源： The AI opportunity is defined by a skilled ecosystem

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat