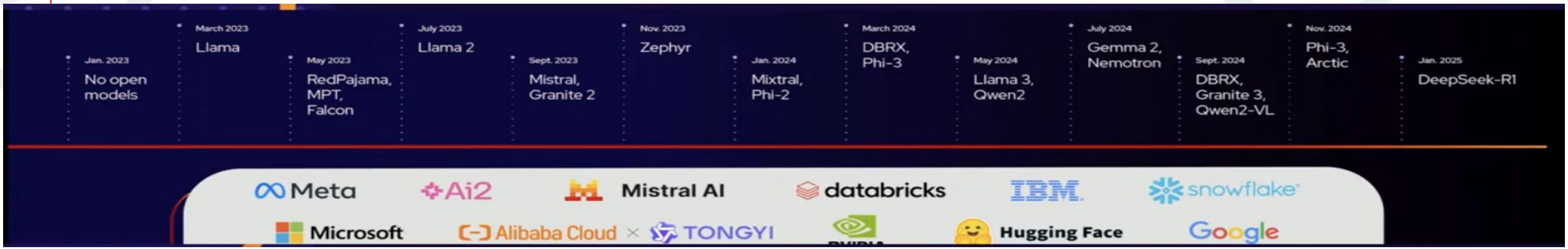
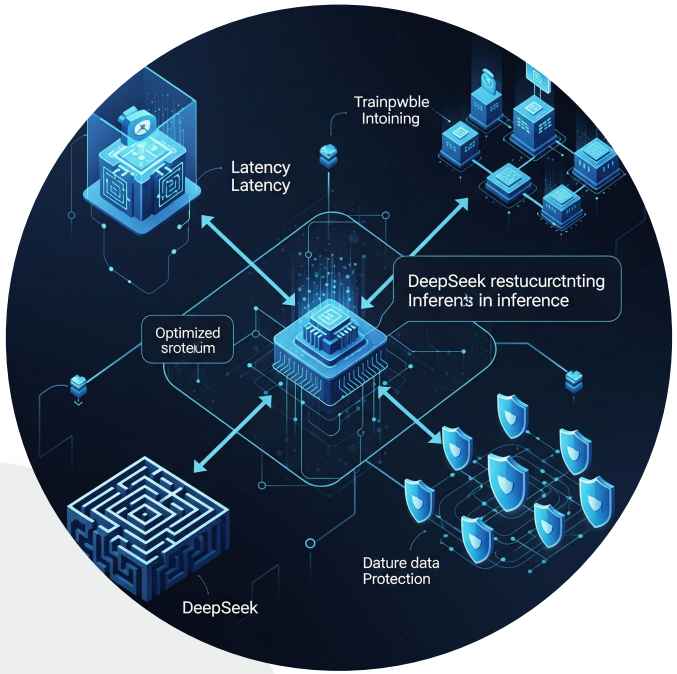
從教室到產業,
平台思維的世代交替:

與Red Hat攜手共育次世代數位人才

Ato Lin | 林嘉彥
Sr. Solution Architect
Red Hat

2025:「推理之年」- AI 從訓練競賽轉向推理革命





面臨核心痛點:

- **推理效能低**:算力緊張、推理速度慢、延遲高、影響使用者體驗和業務效率
- **企業級安全與合規保障困難**:企業對模型私域部署有強烈需求,但在保障資料安全、隱私合規方面面臨挑戰
- **雲端化部署與彈性不足**:企業希望AI應用具備雲端化與彈性伸縮能力,但現方案在彈性與彈性方面不足

Source:
What is AI Inference?, Jan 2025, Red Hat
AI Inference Market Size, Share & Trends, Feb 2025, Markets and Markets
Amazon Found Every 100ms of Latency Cost them 1% in Sales, July 2023, Gigaspaces

Red Hat

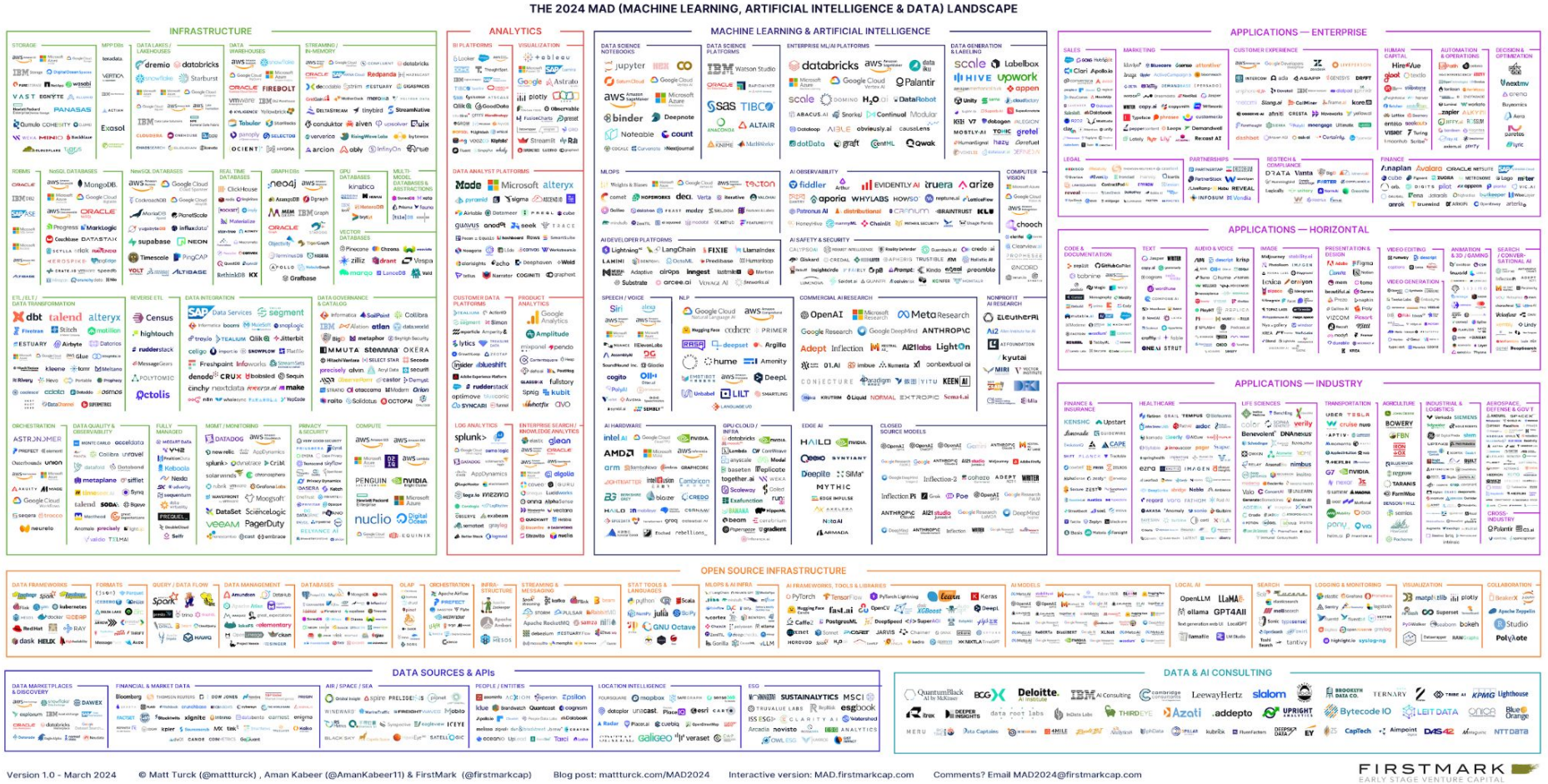
GPU 的使用狀況

- GPU 資源分配不均
  - 實驗室？各自採購？
- GPU 使用率低
  - 暑假, 不好申請, 管理？
- RTX 到高階卡 難以整合
  - 預算, 不同時期採購

Ref : GPUnion: Autonomous GPU Sharing on Campus

人工智慧佈局

上游動態複雜性



智慧應用

AI 工作負載

MLOps | LLMOps | Inference (推論)

企業級Kubernetes

AI 加速器

優化的作業系統

基礎架構

Red Hat

符合您需求的人工智慧基礎架構

協調、優化和規模化您的整個AI 生命週期

適合用途的模型

運用開源模型

使用小型語言模型進行最佳化

最大化 GPU 資源效率，
提升成本效益

降低複雜性

保持選擇彈性，同時不損掌控力

部署統一的基礎架構，以實現效能
規模化和優化成本

透過整合工具和自動化工作流程，
簡化 MLOps

提高靈活性

一次建置 AI，隨處部署

將人工智慧融入您的數據

降低供應鏈風險

人工智慧平台

晶片多樣性（GPU、TPU、NPU 等）

混合雲功能

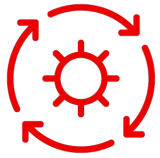
Red Hat

使用紅帽建立人工智慧解決方案的價 值

以自己的方式運用 AI

自帶模型

Red Hat AI 提供經過驗證與量化，並附帶基準測試數據的第三方模型



任何 GPU 上皆可運行 實現高性價比的 Gen AI

vLLM 提供了更有效率的資源利用，這相當於需要更少的 GPU 來處理 LLM，並且對硬體有廣泛的支援。

部署靈活性

透過統一的 AI 平台，以一套技能即可在本地端、任何雲端及邊緣環境，運用數據並部署模型。

優化使用 & 自助服務

共享資源以最佳化 GPU 等硬體的使用率，並讓資料科學家能透過自助服務快速佈建實例。

Google · Microsoft · NVIDIA · IBM · Ai2

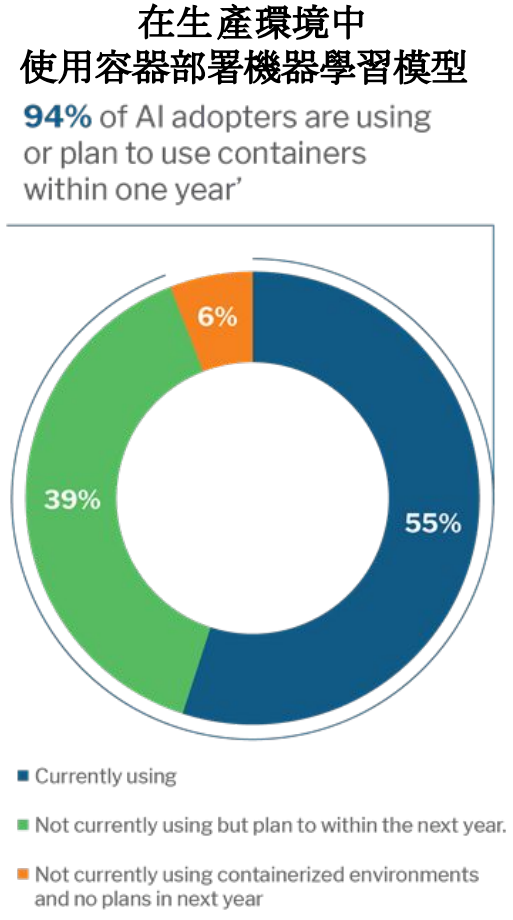
vLLM · llm-d · AMD · NVIDIA · IBM · Google · intel · aws

aws · Azure · 裸機Bare Metal · Alibaba Cloud · Red Hat OpenStack Platform · IBM Cloud · Google Cloud · NUTANIX · Oracle Cloud · vmware vSphere

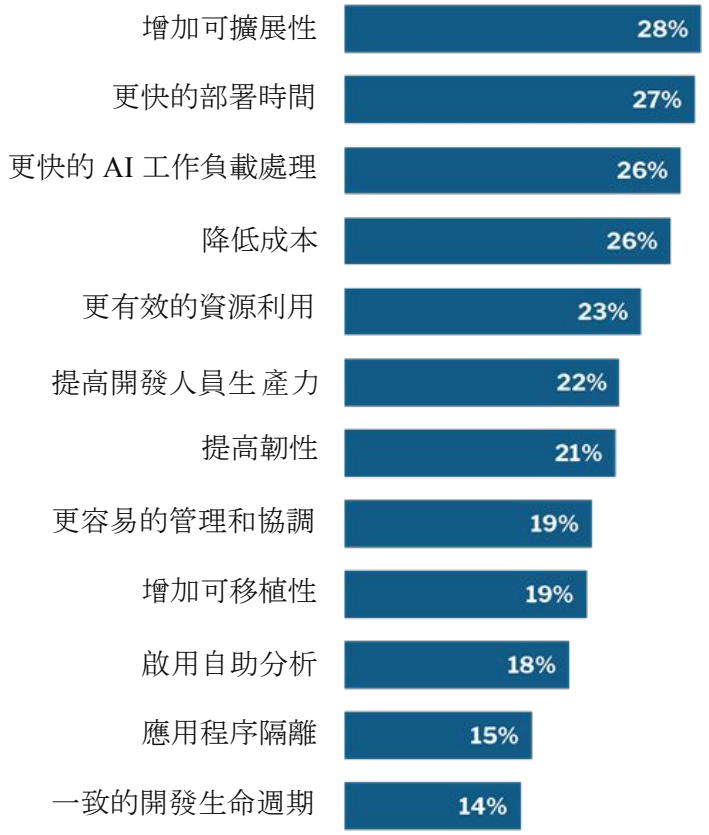
TensorFlow · PyTorch · python · Kubeflow · argo · jupyter

AI 採用者熱衷於使用容器來改善 AI 工作負載

在生產環境中
使用容器部署機器學習模型



容器對 AI 採用者的好處
Benefits of Containers



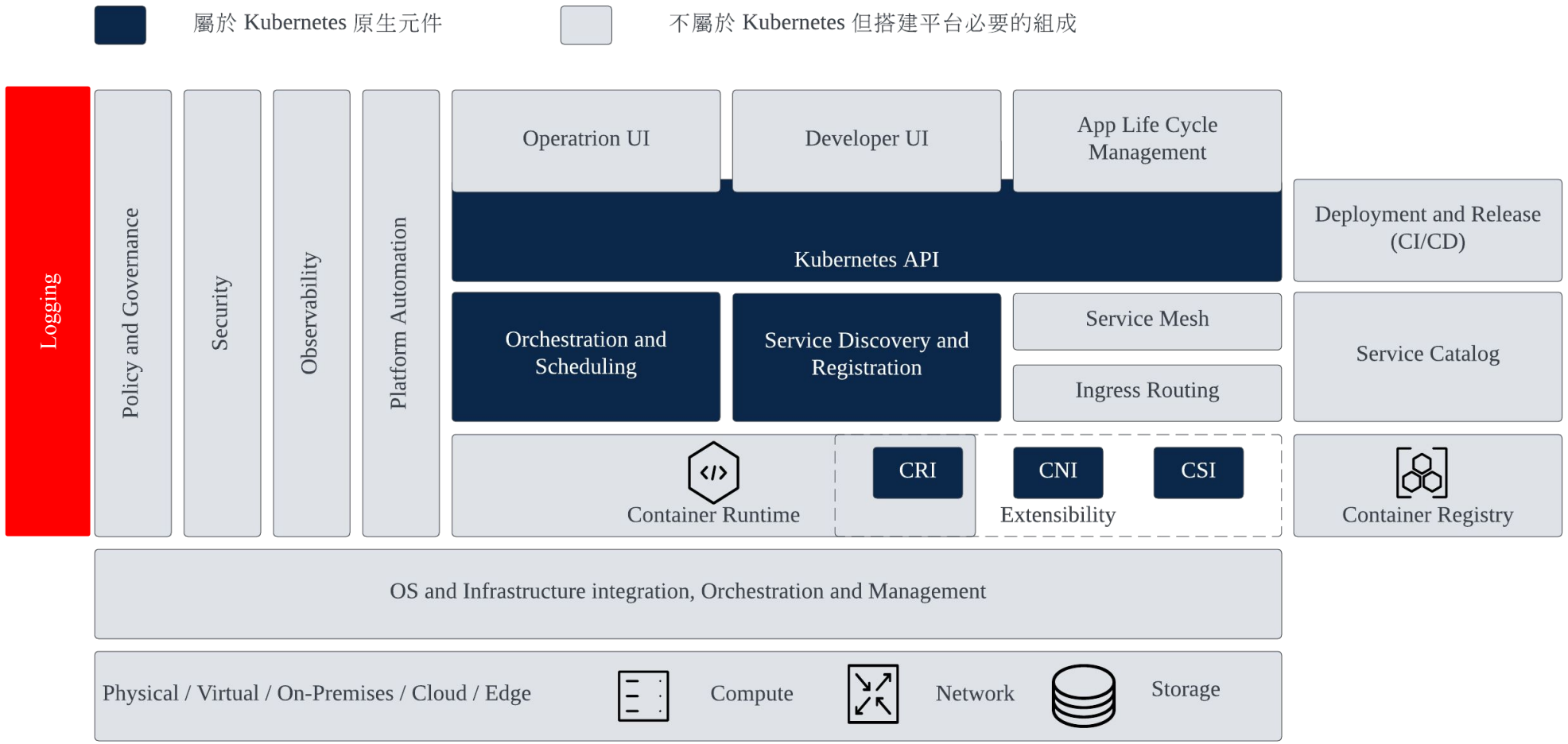
94%

的AI採用者正在用或
者計畫在一年內使用容器

Q: Are you using a containerized environment to deploy machine learning models to production? (n=493)
Q: What benefits [do or will] containerized environments bring to your AI initiatives? Please select up to 3. (n=638)
Base: ML is in production or proof of concept, or there are plans to use ML in next 12 months.1
Source: 451 Research, part of S&P Global Market Intelligence – Voice of the Enterprise: AI & Machine Learning, Infrastructure 2020

Red Hat

Kubernetes 元件 (Gartner 建議 k8s 平台需具備的元件)



Source : Gartner ID G00758623
Designing and Operating DevOps Workflows to Deploy Containerized Applications With Kubernetes

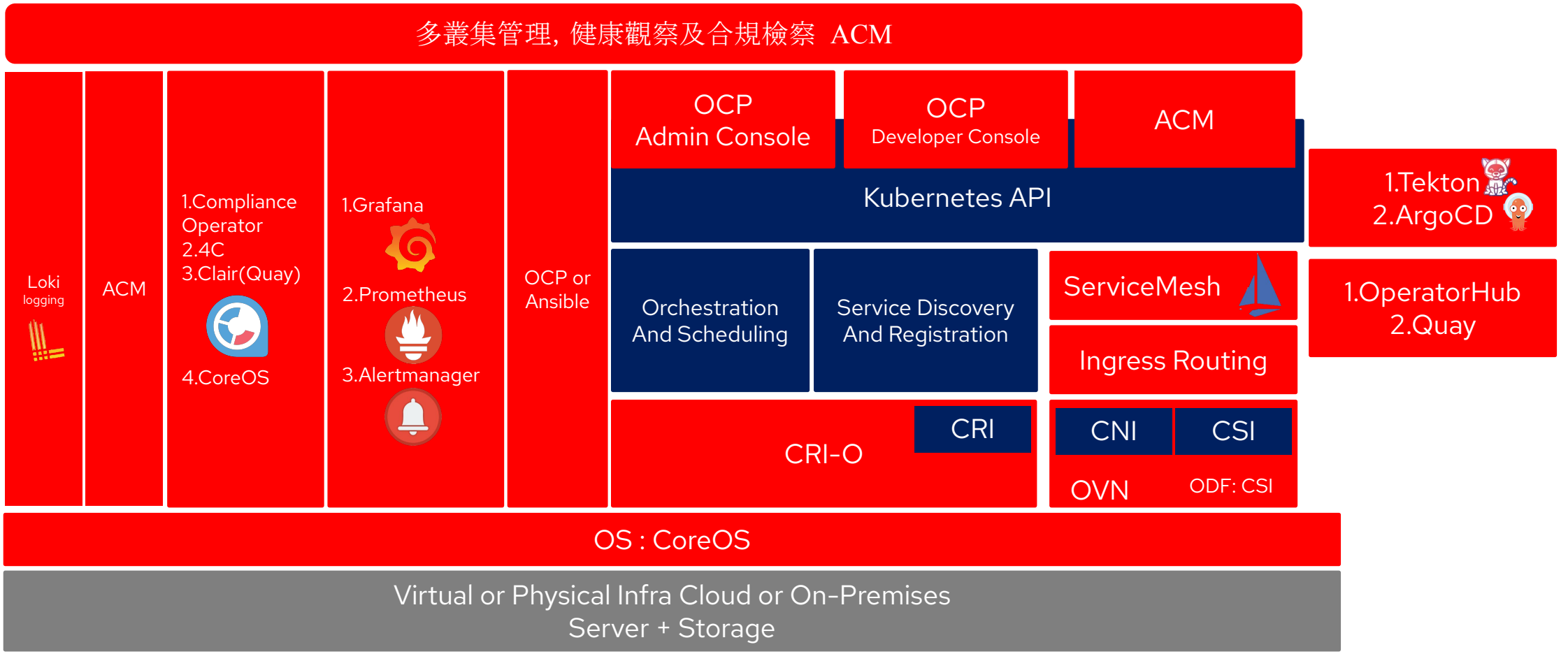
Red Hat

OpenShift 元件說明（紅帽提供的元件）

Red Hat 提供

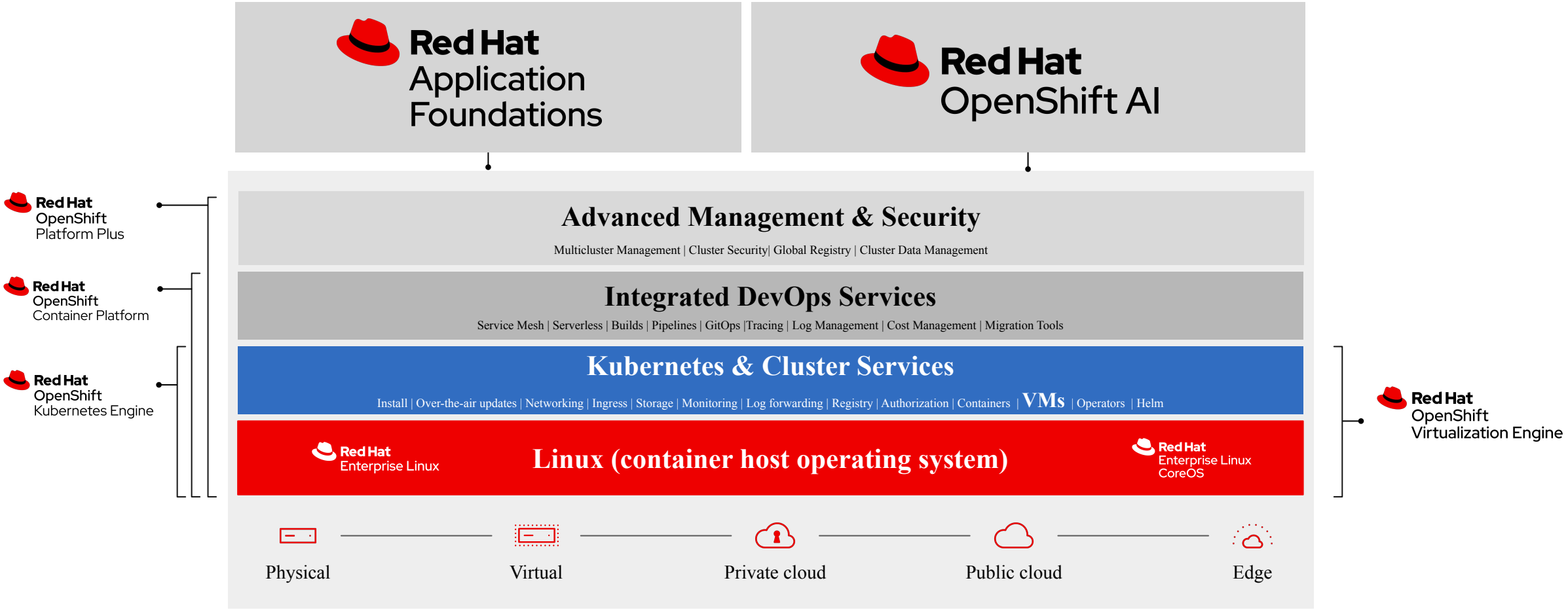
Kubernetes 原生

硬體廠商提供



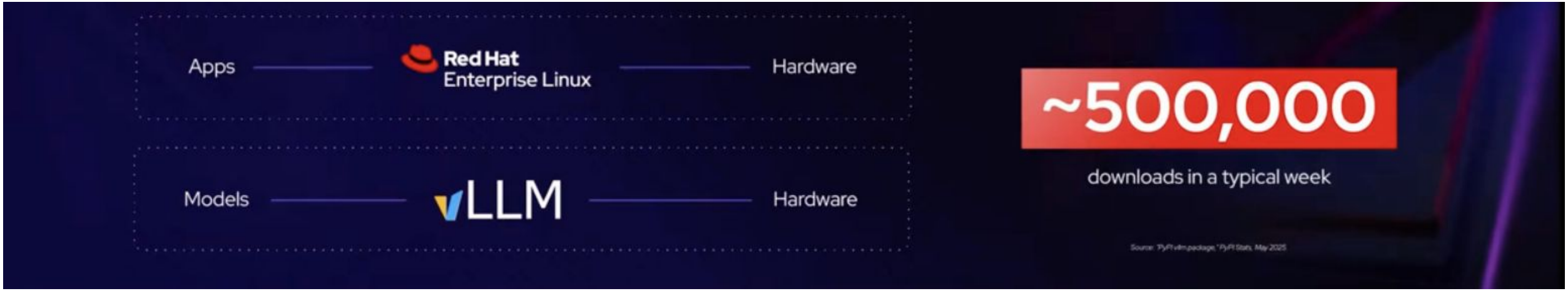
Red Hat

一站式企業級平台解決方案



Red Hat

vLLM: AI 推理的 Linux



Red Hat
OpenShift

生成式和預測式人工智慧的統一平台

Red Hat AI

Red Hat AI
Inference Server

Red Hat
Enterprise Linux AI

Red Hat
OpenShift AI

值得信賴、一致且全面的基礎

NVIDIA AMD intel 硬體加速 aws IBM Google

Physical — Virtual — Private cloud — Public cloud — Edge

NVIDIA, AMD, Intel, Google TPU supported in Red Hat AI.
AWS Inferentia/Neuron IBM AIU are on our roadmap

Red Hat's AI/ML engineering is 100% open source

Red Hat

靈活高效的推理

- 分散式推理 (llm-d)
- 新的經過驗證和最佳化的模型
- vLLM 增強功能
- LLM 壓縮器

將模型連接到數據

- 模組化且可擴展的方法：資料收集、合成資料產生、調優、評估。
- RAG 增強功能和合作夥伴集成
- Feature Store

Agentic AI

- AI 體驗：AI Hub 和 Gen AI Studio
- Gen AI Studio 中的 MCP 支援
- Llama Stack API 整合

AI Platform

- 模型目錄和模型倉儲
- Model as a Service 增強功能和 API 管理整合
- GPU as a Service 功能增強

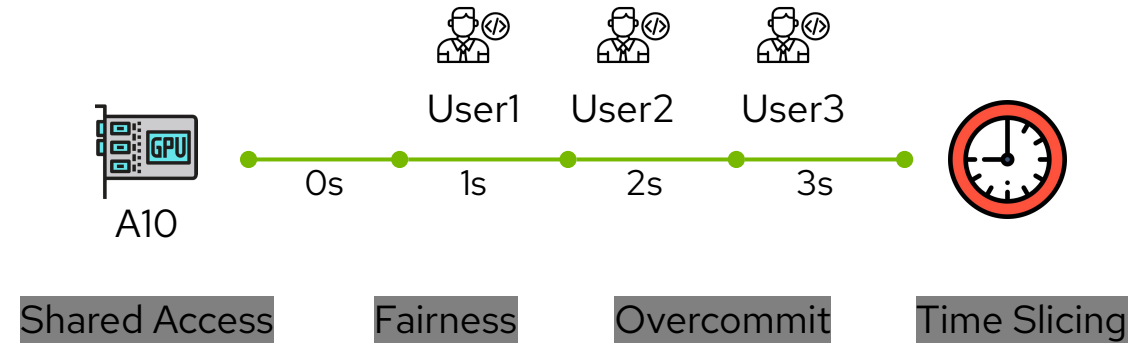
單一平台可在任何加速器、任何雲端上運行任何模型

Red Hat

Use cases (Time Slicing)

Ensuring efficient GPU sharing across users

Scenario / Use case	Common Challenges	Solution
Early stage of development (Limited computing resources)	One team's workload monopolizes GPU memory, affecting others.	**Time Slicing & Multi Tenancy**

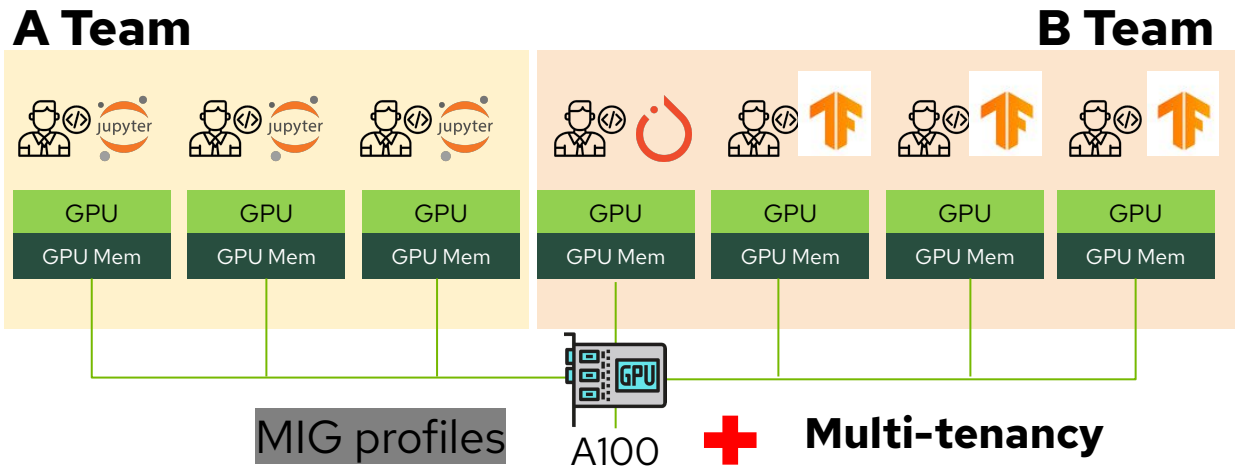


Red Hat

Use cases (MIG)

Ensuring GPU resource isolation in a multi-team environment

Scenario / Use case	Common Challenges	Solution
Multiple internal teams within a company (e.g., AI, Cloud, MLOps)	One team's workload monopolizes GPU memory, affecting others.	**Multi Instance GPU & Multi Tenancy**



GPU Isolation

Fair sharing

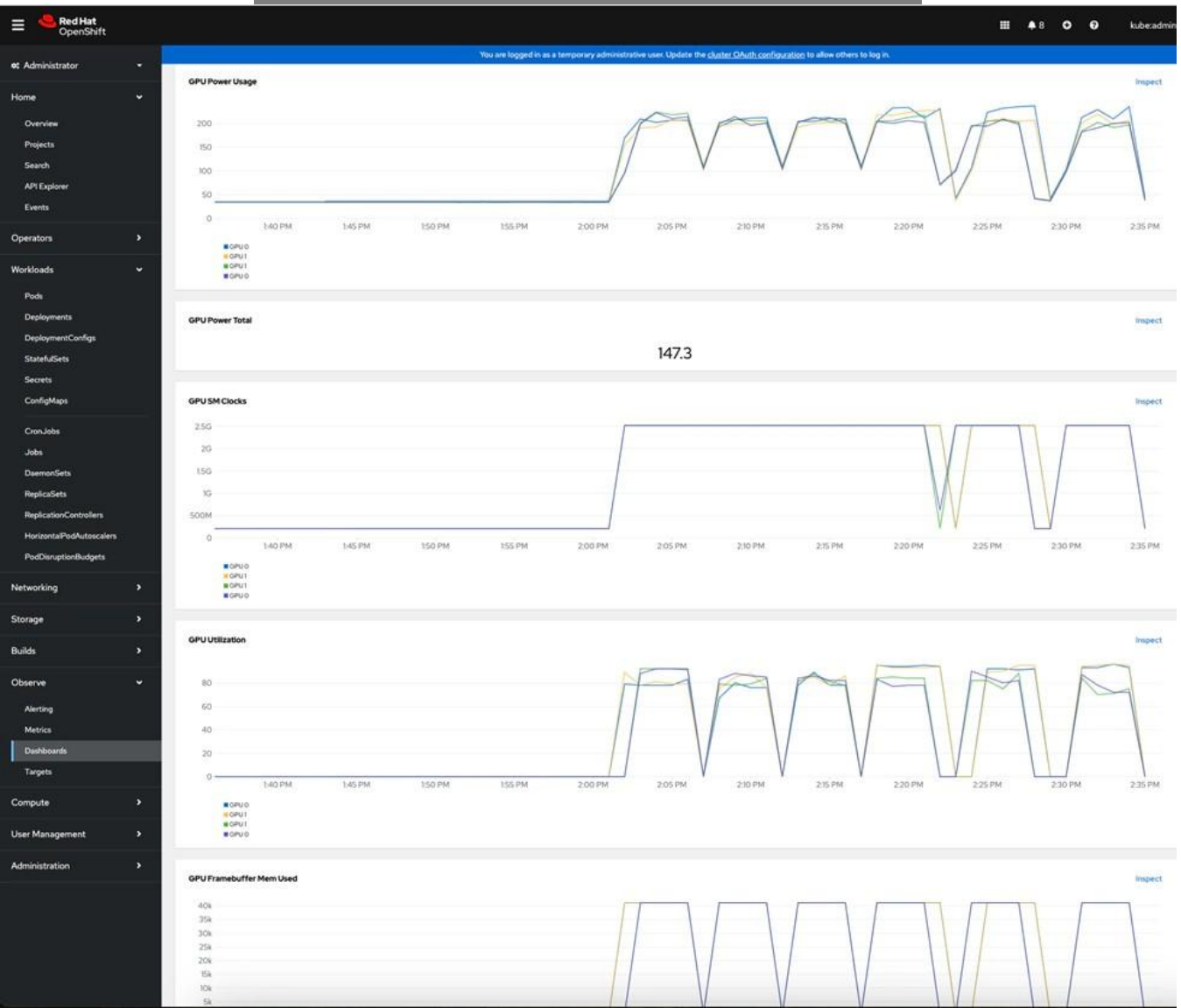
MIG profiles

Multi-tenancy

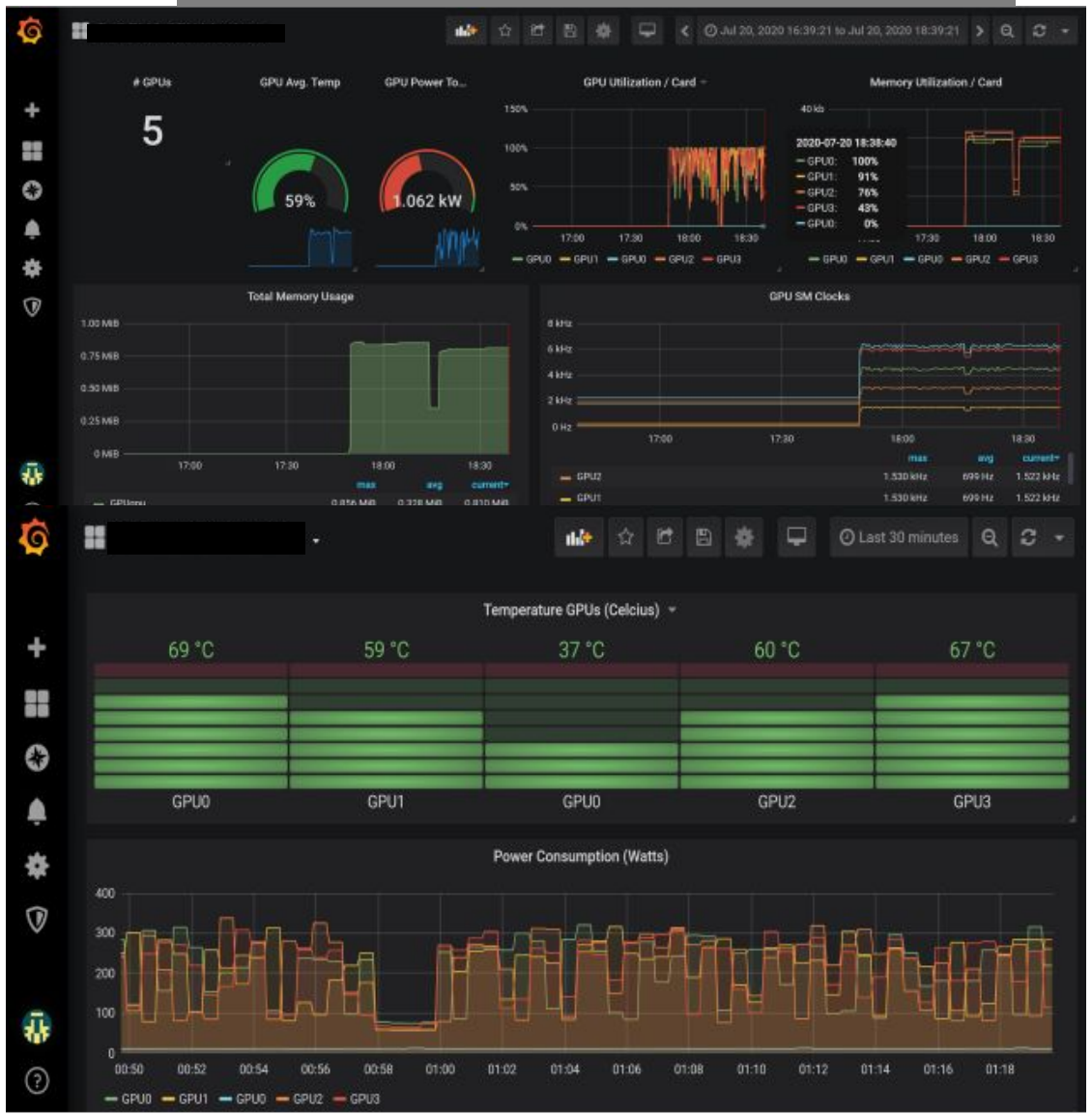
Red Hat

OpenShift GPU 資源監控

整合 NVIDIA DCGM Operator

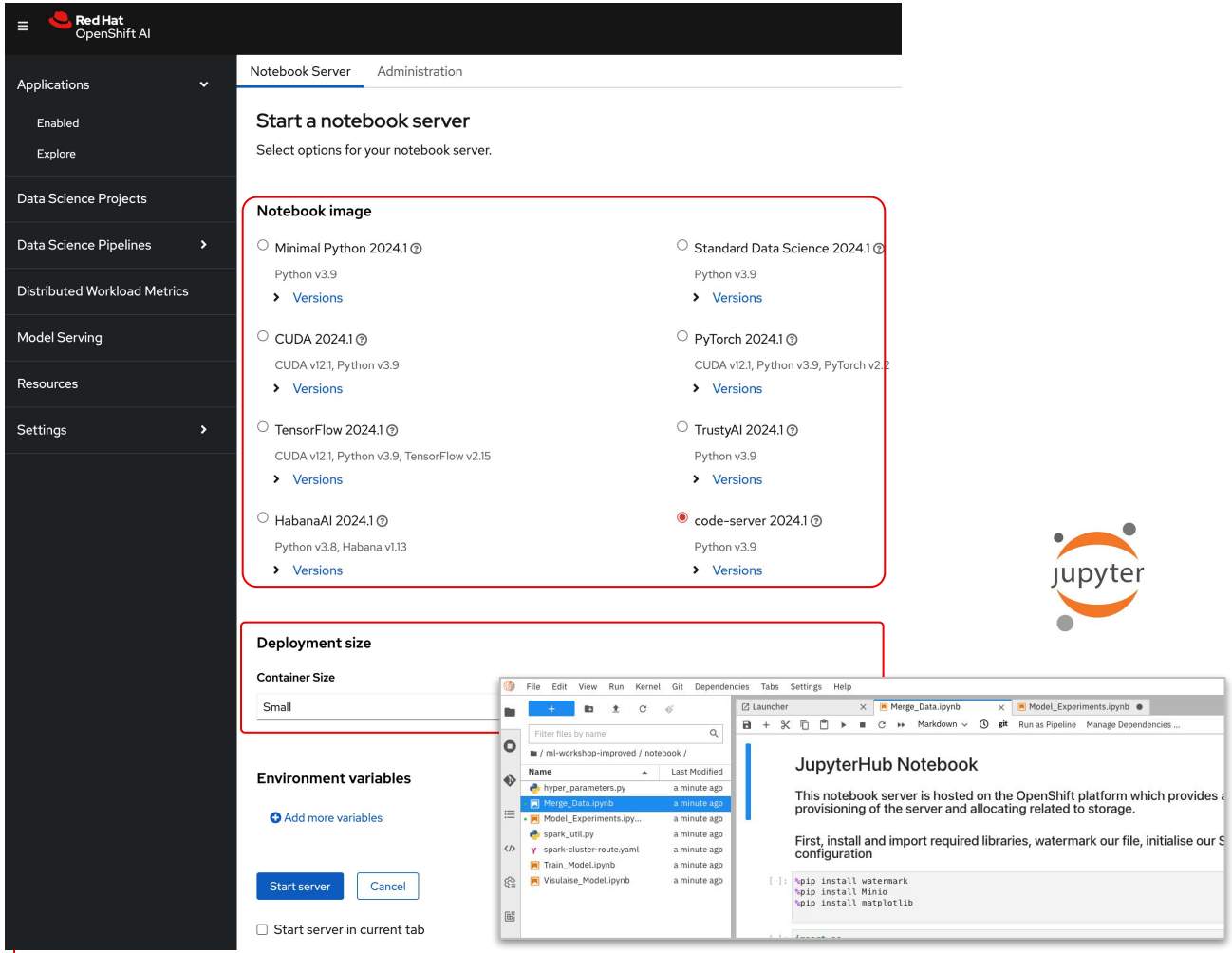


整合客戶自有 Grafana Dashboard (預設支援 NVIDIA DCGM Exporter)



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)。

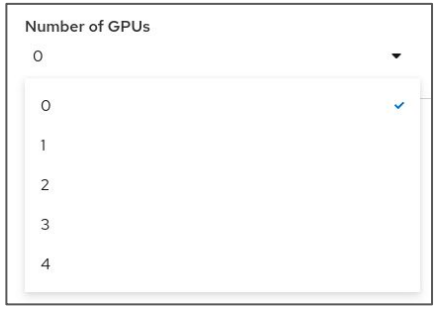
Deployment size

指定所需的資源(CPU/Memory) 來啟動容器

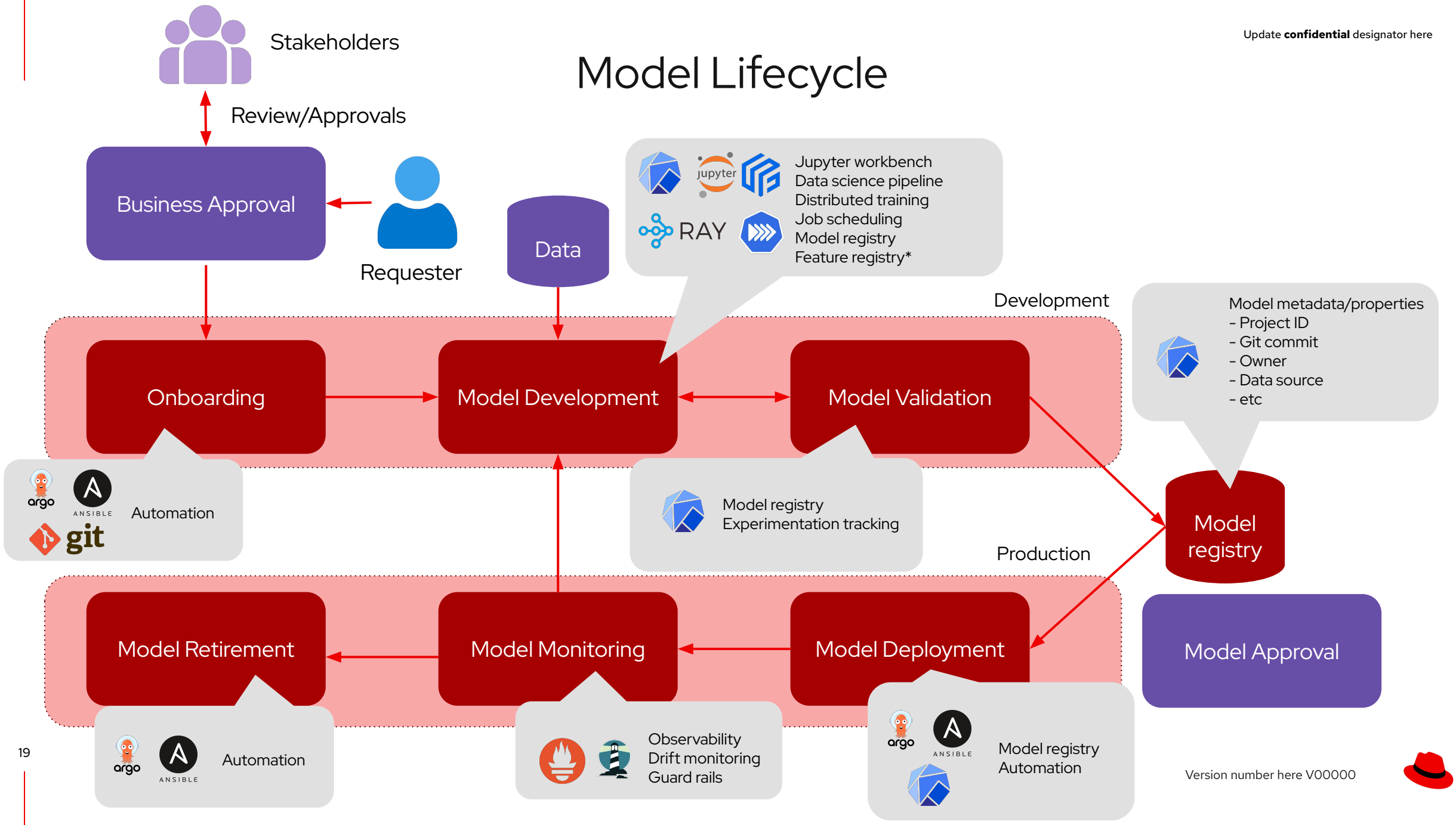
透過設定預設的停止時間，可以自動停止非活躍的環境並回收資源

Number of GPUs

加入指定 GPU 數量資源



Model Lifecycle



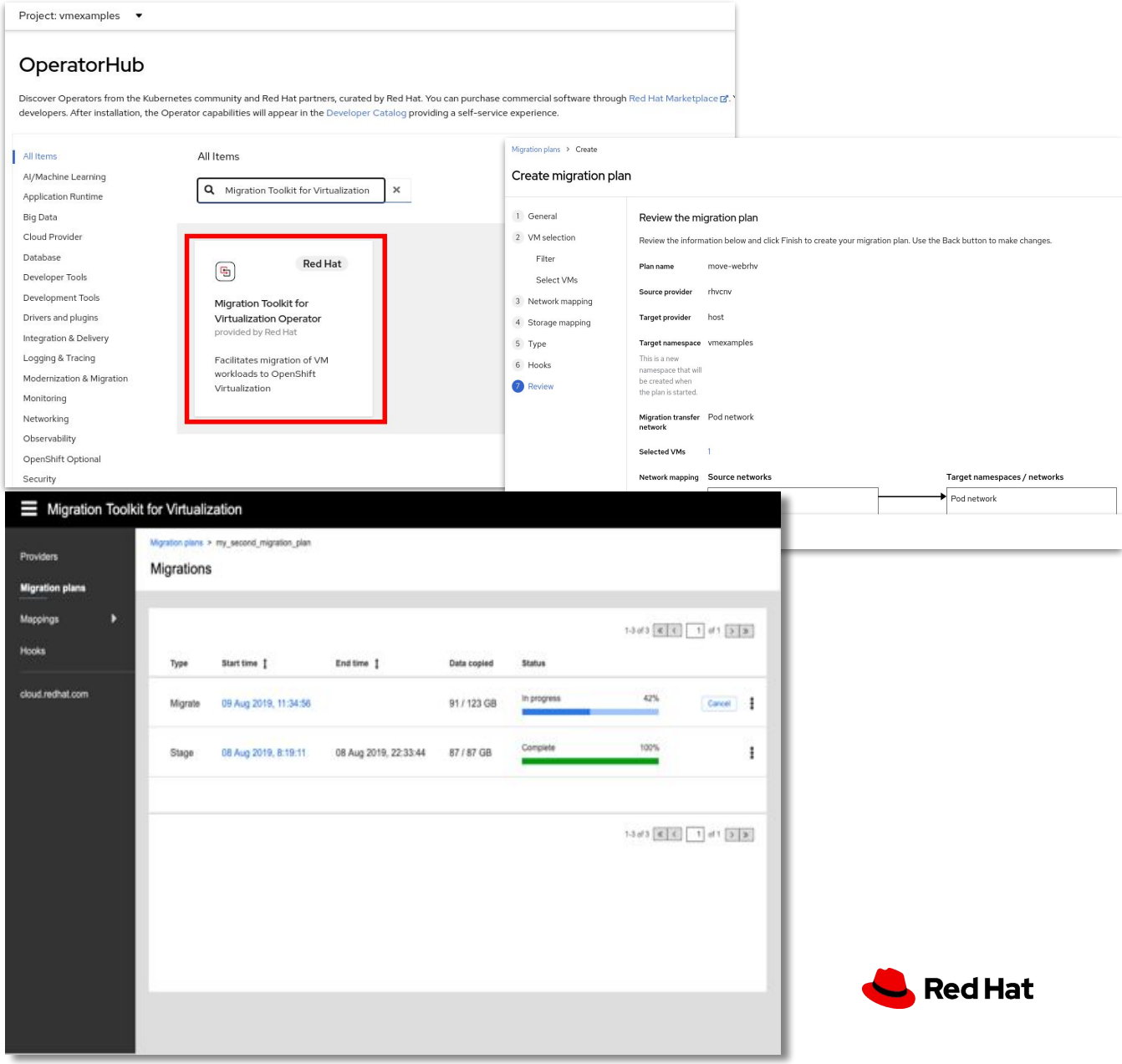
將傳統 VM 遷移進 OpenShift Virtualization

虛擬機遷移

只需幾個簡單的步驟建立轉移計畫(Migration Plan)，即可將虛擬機器大規模遷移到 OpenShift Virtualization。

目前支援來源端的虛擬平台有：

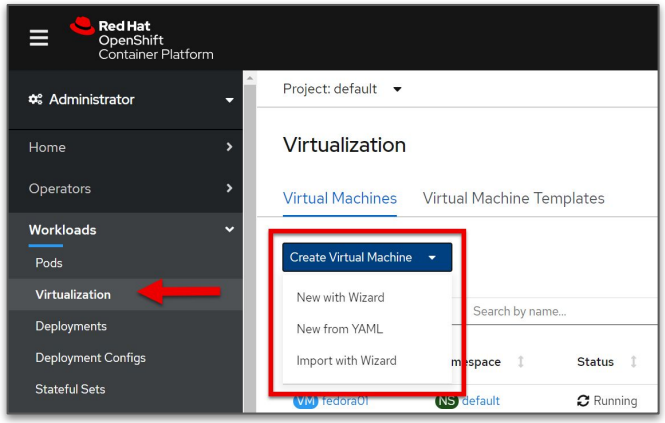
- Red Hat Virtualization
- KVM
- OpenStack
- VMware vSphere
- OVA File
- OpenShift Virtualization

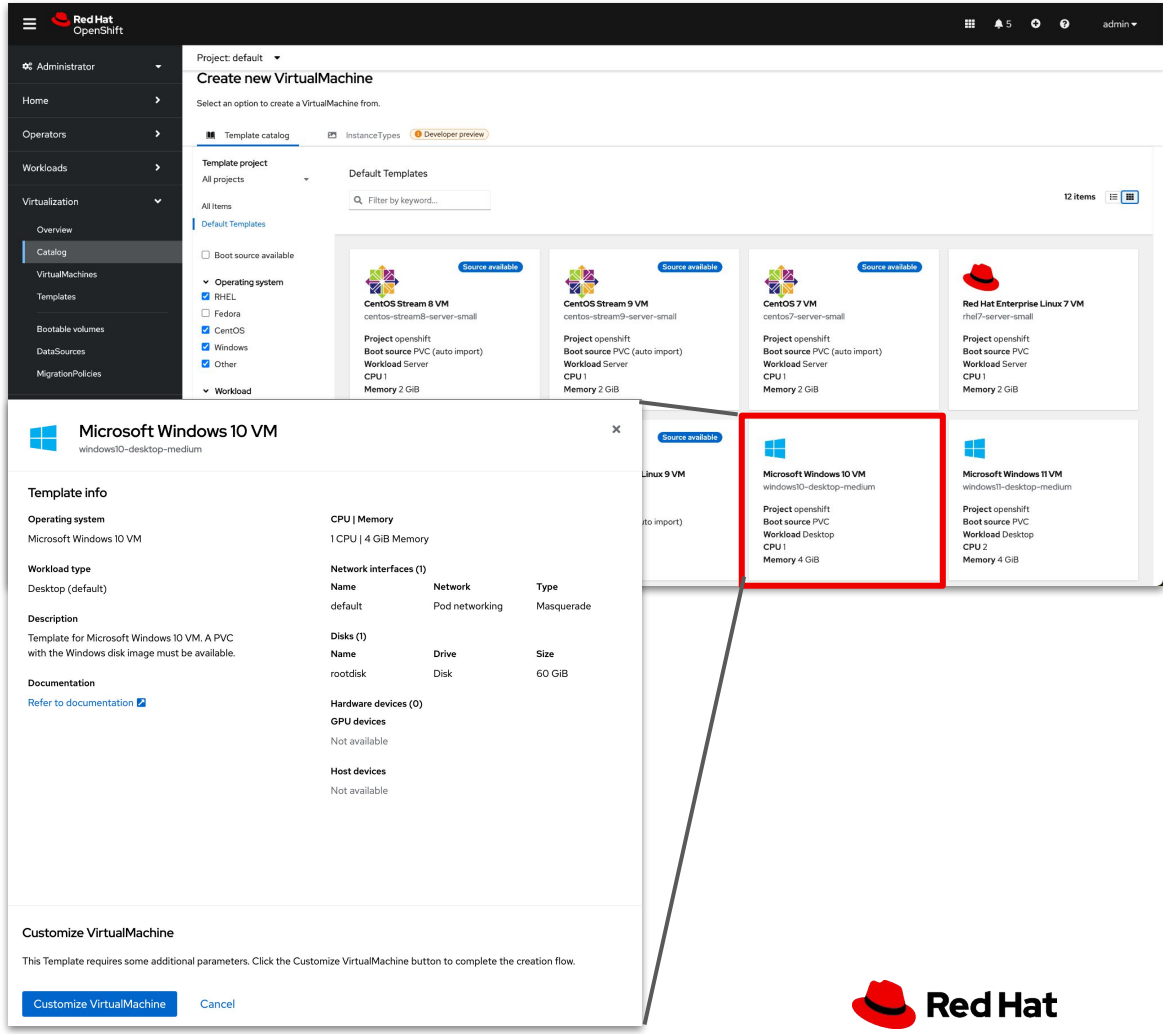


OpenShift 次世代容器平台提供運行虛擬機的能力

OpenShift Virtualization虛擬化功能

- 容器平台內建功能
- 可透過使用者介面(GUI)簡化並創建 VM 虛擬機
- 介面整合系統資源、網路和儲存配置選項
- 支持 Linux 和 Windows 虛擬機器作業系統
- 可從 VMware vSphere 遷移虛擬機器





Red Hat

FSI

台灣規模 Top 3 金控集團之一

客戶挑戰

2024 年被公認為人工智慧元年。該集團已確認利用生成式 AI 技術為客戶提供更全面的金融服務，包括金融詐騙防範、AI 智能客服、貸款/承保智能協助等。但因人員技能尚未完備、故不願意花時間建構和維護 AI 相關的基礎設施和平台工具，而是期待一個開箱即用的企業級 AI/ML 解決方案，提供穩定且有效率的 GPU 資源管理和跨團隊 AI 專案協作平台。

解決方案

提供 Red Hat OpenShift AI 和 GPS 團隊實施的諮詢服務，以協助該集團標準化 AI 專案的基礎執行環境，並集中現有 OpenShift 叢集上的 AI 周邊應用程式管理。

Why Red Hat

該集團的痛點是缺乏完整的 AI/ML 解決方案、和與 AI 相關的技能知識，這意味著開發人員無法專注於服務創新，從長遠來看將減緩業務成長。
這兩個問題都可以透過 Red Hat RHOAI 和專業服務來解決，與競爭對手相比，它具備企業級支持、且運行於經市場驗證過穩定性和安全性的 OpenShift 平台。

效益

Red Hat 協助該集團建構了集團 內第一個完整的 AI/ML 平台，並提供教育和培訓，使 IT 團隊掌握維運技能。這使得開發團隊從維護 AI 工具中解放出來，並使他們能 夠專注於應用服務開發。
1. AI 項目必要環境建置時間減少 70%。
2. 將 AI 服務開發、測試和發佈時間縮短 30%。

軟體與服務

Red Hat OpenShift Platform Plus
Red Hat OpenShift AI

Red Hat Consulting

Red Hat

簡化 AI 模型服務部署，搭配專業在地技術支援，加速創新應用落地，有效降低導入成本與複雜度。

客戶挑戰

現有 AI 環境多為分散建置，缺乏統一管理，導致維運複雜、效率不彰。部分單位選擇台製品牌導入時，常出現穩定性不足或無法使用的情況，需一個高效穩定且容易管理的平台，來簡化 AI 導入流程、提升效能，確保政策順利落實。

解決方案

完整的 AI 平台和專業服務，協助單位快速建置可支援多種 AI 模型（如 TAIDE、Mistral、Gemma3）的統一應用平台，並維持企業級的技術支援。解決快速部署與穩定性的需求，使 AI 應用得以穩定且迅速上線，顯著縮短導入時程，並簡化日後維運管理，大幅提升系統整體穩定性與運行效率。

Why Red Hat

紅帽解決方案的關鍵在於其穩定、安全與可持續的開源創新能力。支援多種主流與在地 AI 模型，且具備完善的技術服務。平台設計便於管理與擴展，協助組織有效推動 AI 應用落地，降低部署與營運上的複雜度。

效益

- 單一平台整合：提供多元且彈性擴展的 AI 模型服務，快速滿足各種應用需求。
- AI模型服務：透過簡易網頁介面部署與操作，降低 AI模型服務部署技術導入門檻。
- 專業服務在地團隊：提供完整技術支援，簡化後續管理的複雜度。

軟體與服務

Red Hat OpenShift Platform Plus
Red Hat OpenShift AI

Red Hat Consulting
AI Innovation LAB

Red Hat

2025 Market Ranking

Gartner June 2025
Container management

Figure 1: Magic Quadrant for Container Management

Gartner

Forrester Wave Q3 2025
Multicloud Container Platforms

https://www.redhat.com/en/engage/gartner-magic-quadrant-container-management-analyst-report
https://reprint.forrester.com/reports/the-forrester-wavetm-multicloud-container-platforms-q3-2025-f3367e46/index.html

Red Hat

完整的教育訓練, 從 Linux 到 容器平台 到 AI

- **Red Hat 系統管理 I** | RH124 • 5 days
- **Red Hat 系統管理 II** | RH134 • 5 days
- **Red Hat Certified System Administration exam** | EX200 • 3 hours
- **Red Hat Certified System Administrator** | RHCSA
- **利用 Ansible 實現紅帽企業 Linux 自動化** | AU294 • 4 days
- **Red Hat Certified Engineer exam** | EX294 • 4 hours
- **Red Hat Certified Engineer** | RHCE

- **Red Hat Certified System Administrator** | RHCSA(建議要有 Linux 技能再上以下課程)
- **Red Hat OpenShift I: Containers & Kubernetes** | DO180 • 4 days
- **Red Hat OpenShift 管理 II: 操作生產級 Kubernetes 集群** | DO280 • 4 days
- **Red Hat Certified Specialist in OpenShift Administration exam** | EX280 • 3 hours
- **Red Hat OpenShift管理 III: 在企業中擴展 Kubernetes部署** | DO380 • 4 days
- **Red Hat Certified Specialist in OpenShift Automation and Integration exam** | EX380 • 3 hours

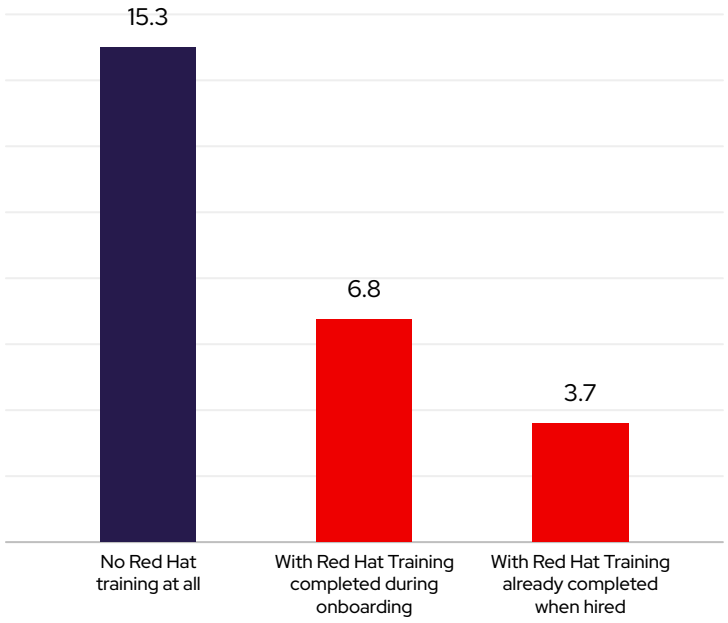
認證 | 課程 | 考試

為學生準備就職就緒程度及留任率

工作準備就緒速度提高 76%(上任之前受訓)
when a new hire had already completed Red Hat training prior to onboarding.

工作準備就緒速度提高 55%(新人訓受訓)
where new team members completed Red Hat training as part of the onboarding process.

任期延長 8%
Red Hat-trained employees have 8% longer tenure on average than untrained staff members.



達到完全生產力之時間(週數)

IDC

Source: IDC Business Value White Paper, sponsored by Red Hat, The Business Value of Red Hat Training, Doc. #US46999720, December 2020

Red Hat

Demo

快速部署一個模型服務

Red Hat

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat

Red Hat