

AI 時代 人, 時空與金錢挑戰下的最佳解方 Nutanix 混合雲多平台

陳永恆 Eternity Chen
Nutanix 資深技術顧問

NUTANIX

企業的新選擇 - 超融合與軟體定義平台

Nutanix – 客戶信賴的核心應用系統運行平台

70%

move workloads
from SANs to
software-defined
infrastructure
when refreshing

56%

of enterprises run
mission-critical
workloads on
software-defined
infrastructure

75%

of Nutanix
Customers deploy
mission-critical
workloads on
Nutanix Cloud
Platform



Business Value White Paper, sponsored by Nutanix
OCTOBER 2022

REGION FOCUS: WORLDWIDE

The Business Value of Nutanix Cloud Platform



Harsh Singh

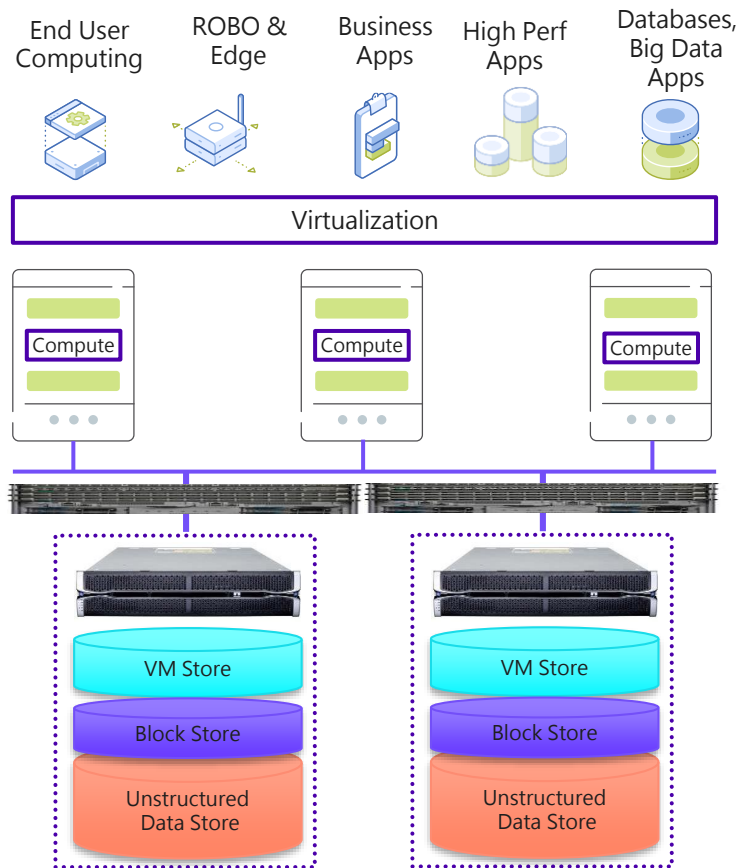


Ashish Nadkarni



Source: IDC White Paper, sponsored by Nutanix, The Business Value of Nutanix Cloud Platform,
October 2022 (#US49715622)
Nutanix Installed base survey July 2022

傳統三層式架構資料中心

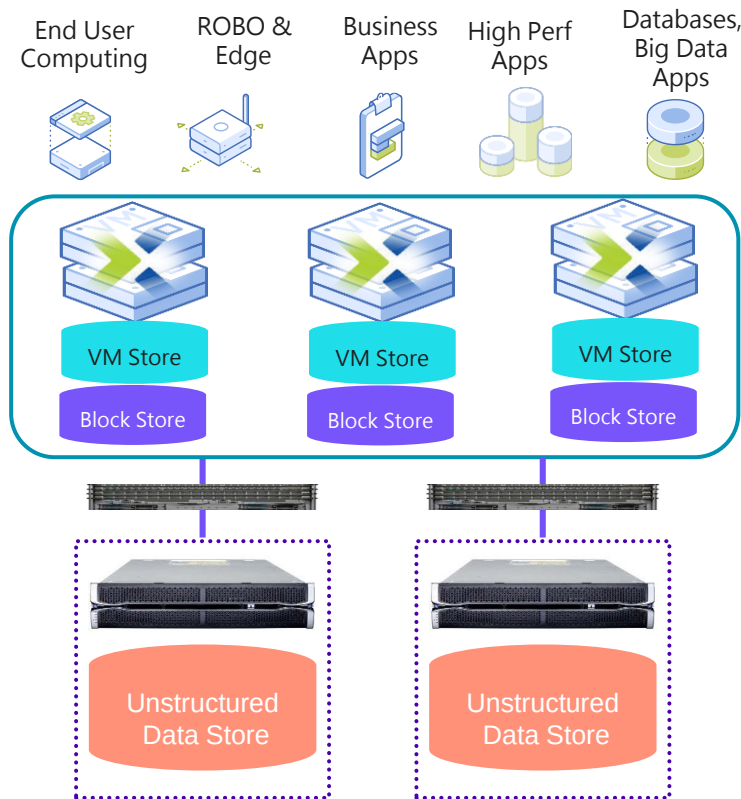


運算
與
儲存
多元件
組成

三層式架構 挑戰

- 複雜
 - 建置佈署 維運管理
 - 資源監控 效能分析
 - 擴充升級 故障排除
- 成本高昂
 - 資源孤島 過量佈署
 - 機櫃空間 電力耗用

Nutanix 超融合資料中心

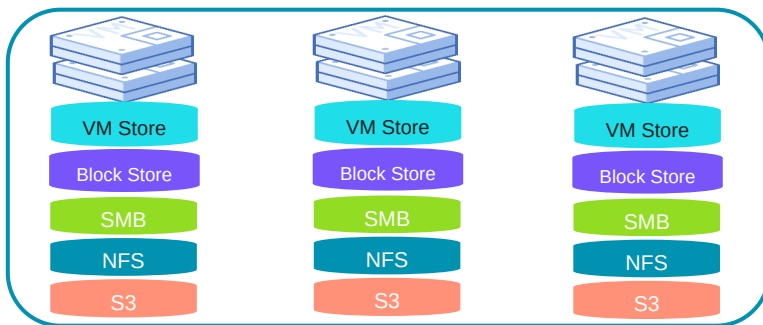


融合
運算
儲存

Nutanix HCI 效益

- **複雜 簡單**
 - 建置佈署 維運管理
 - 資源監控 效能分析
 - 擴充升級 故障排除
- **成本高昂 降低 TCO**
 - 資源孤島 過量佈署
 - 機櫃空間 電力耗用
- **提升效能 可靠度與資訊安全**
- **混合多雲架構 Ready**

啟用 Nutanix Unified Storage



融合
運算
提供
儲存



單一平台 多元應用

- 簡單
 - 建置佈署 維運管理
 - 資源監控 效能分析
 - 擴充升級 故障排除
- 降低 TCO
 - 資源孤島 過量佈署
 - 機櫃空間 電力耗用
- 提升效能 可靠度與資訊安全
- 在任何地方 運行各種服務
- 具備多元儲存服務

建構生成式 AI



Compute



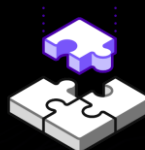
Storage



GPU



Kubernetes
Cluster



LLM
Model



Inference
Server



Model
Access
Control

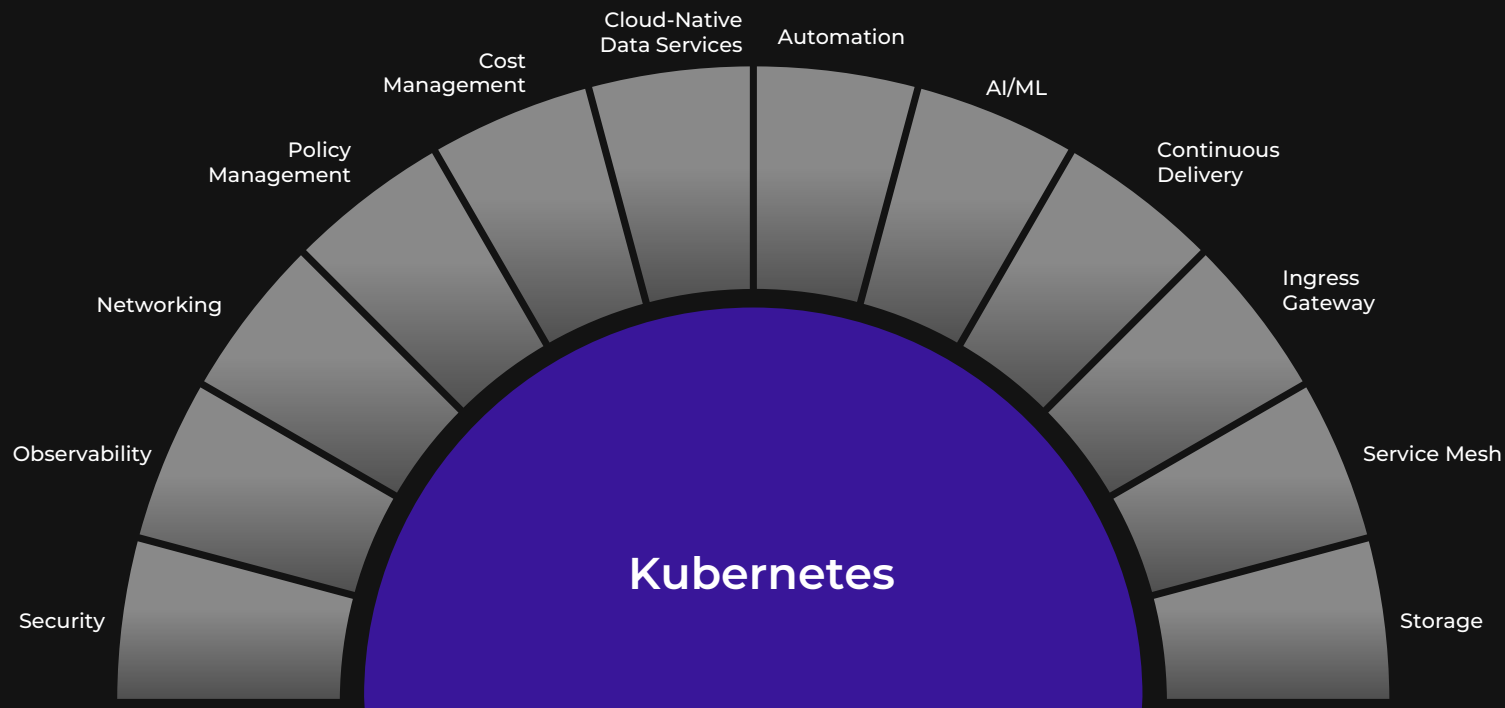


User
Access
Control



AI
Application

構建雲原生平臺需要什麼？



這不難吧？

我認為我的團隊可以自己動手

The image displays the CNCF Landscape, a comprehensive catalog of cloud-native technologies. The landscape is organized into several functional categories, each represented by a colored header bar on the left and top. The categories include:

- Application Definition and Development:** Tools like HELM, Backstage, Buildpacks.io, dapr, KubeVela, KubeVirt, and others.
- Database:** A grid of various database offerings such as 1KV, Vitess, and others.
- Continuous Integration & Delivery:** Tools like argo, flux, and others.
- Streaming & Messaging:** Tools like cloudevents, NATS, and others.
- Scheduling & Orchestration:** Tools like KEDA, and others.
- Service Mesh:** Tools like Istio, and others.
- Remote Procedure Call:** Tools like gRPC, and others.
- Security & Compliance:** Tools like Falco, and others.
- Automation & Configuration:** Tools like Juju, and others.
- Container Registry:** Tools like Harbor, and others.
- Key Management:** Tools like Spiffe, and others.
- Observability and Analysis:** Tools like fluentd, and others.
- Continuous Optimization:** Tools like bmc, and others.
- Chaos Engineering:** Tools like Chaos Mesh, and others.
- Feature Flagging:** Tools like Open Feature, and others.

The central part of the image features a large, semi-transparent text overlay that reads:

CNCF Landscape

Over 1,200 offerings

您正面臨的挑戰



開發人員管理
Kubernetes 與基礎架構
產生複雜性
並且浪費時間.



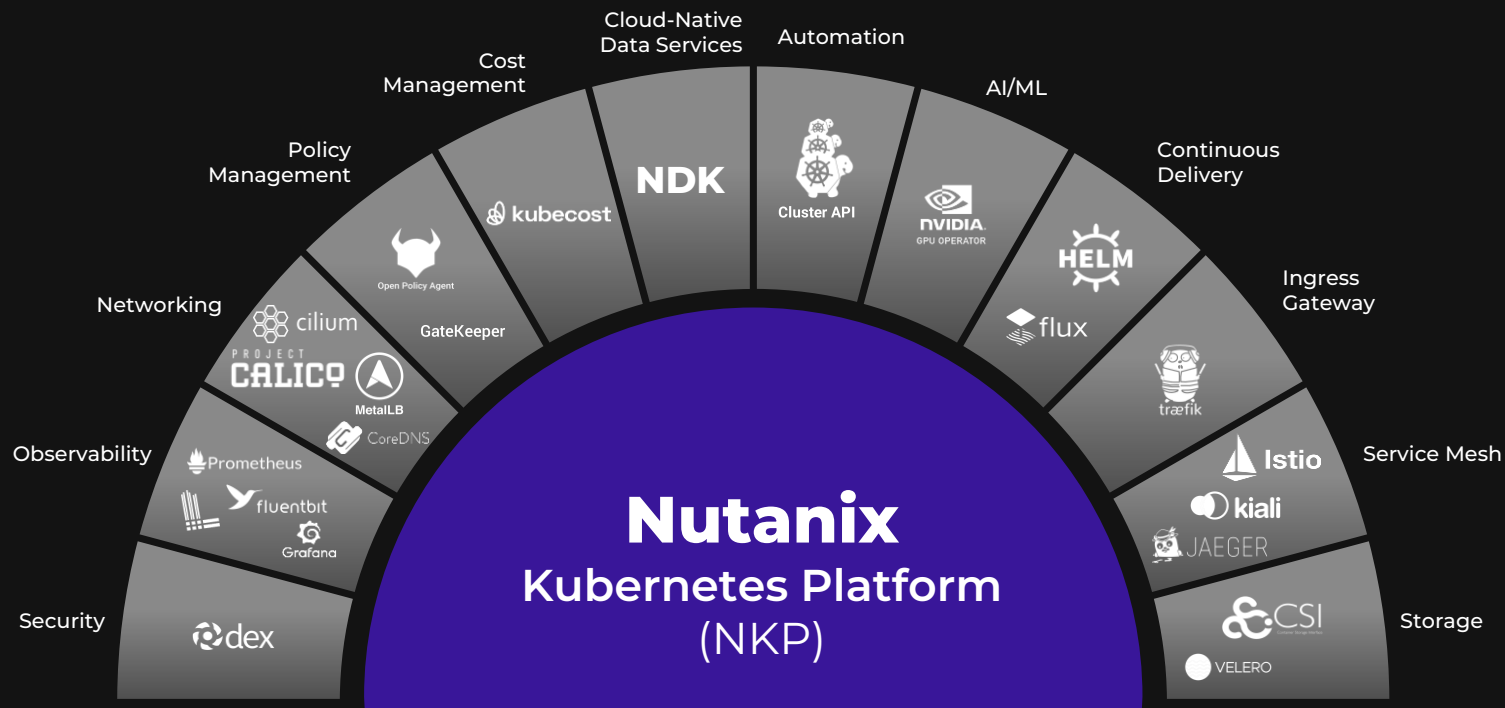
獨立運行的
VM 與 Container
造成大量 資源孤島
效率低下 成本高昂.



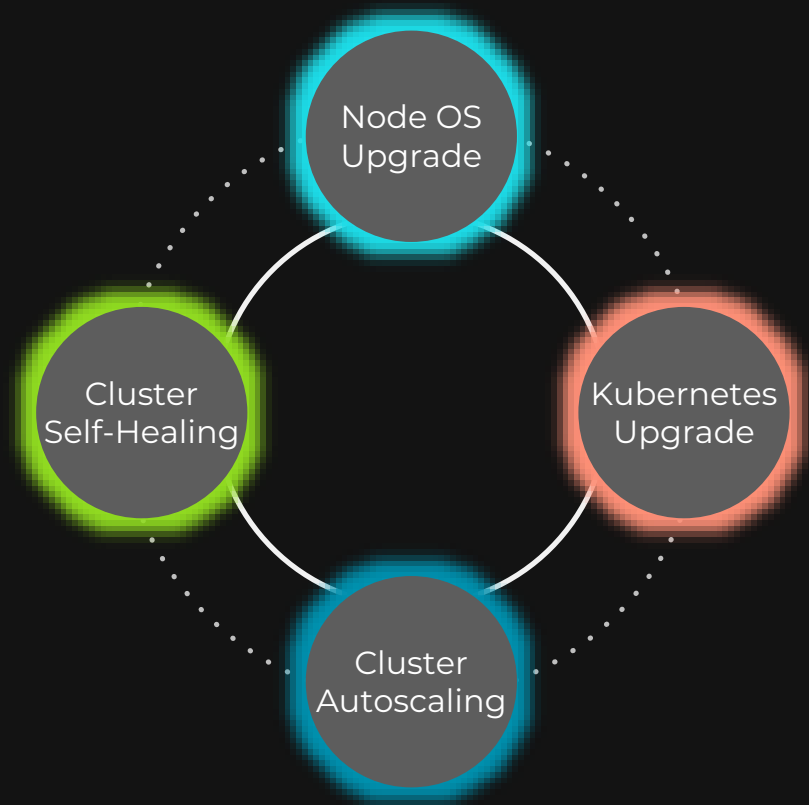
存儲數據 用於容器
通常涉及 多平臺 和
多供應商解決方案.

整合和掌握您的雲原生運營

部署 保護 管理 升級：
大規模企業級雲原生堆疊



叢集生命週期管理 Cluster LCM : 主要優點和特點



輕鬆升級叢集

NKP CLI 通過 Cluster API 使 Kubernetes 叢集的升級變得簡單，包括節點作業系統、Kubernetes 版本和核心外掛程式

自動 **Scale In/Out** 叢集

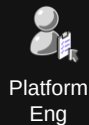
NKP 包含 Cluster Autoscaler 來自動調整 Kubernetes 叢集的大小

自我修復叢集的彈性

CAPI (Cluster API) 持續監控叢集的狀態並將其與所需的配置進行協調

一致的 運營模式

- Open-Source Innovation
- Comprehensive Platform
- Automated Workflows
- AI-Driven Operations



Cloud
Admin

Nutanix Kubernetes Platform (NKP)

Deploy

Secure

Manage

Upgrade

Choice of Kubernetes Distribution



Infrastructure
Data Services

Platform
Data Services



vmware®



NKP 具備 軍規級 DevSecOps



Pre-scanned
releases



Installation
bundle



SSO



Centralized
RBAC



Multi-tenancy



VPC
integration



Network
policies



Encrypted
comms



Policies



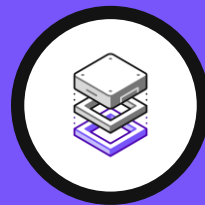
Anomaly
detection

Stateful App 的備份與異地備援



Application DR

Nutanix Data Services for
Kubernetes (NDK)
提供數據保護和管理
並支援異步複製和低 RTO



Application Backup

NKP 提供 Velero
一個用於備份和遷移
Kubernetes 資源和 Persistent
Volumes 的開源專案

存儲對於 Kubernetes 至關重要

所有企業和組織
持續生成各種數據資料

它需要具有 彈性, 可用性,
安全性和 受保護性.

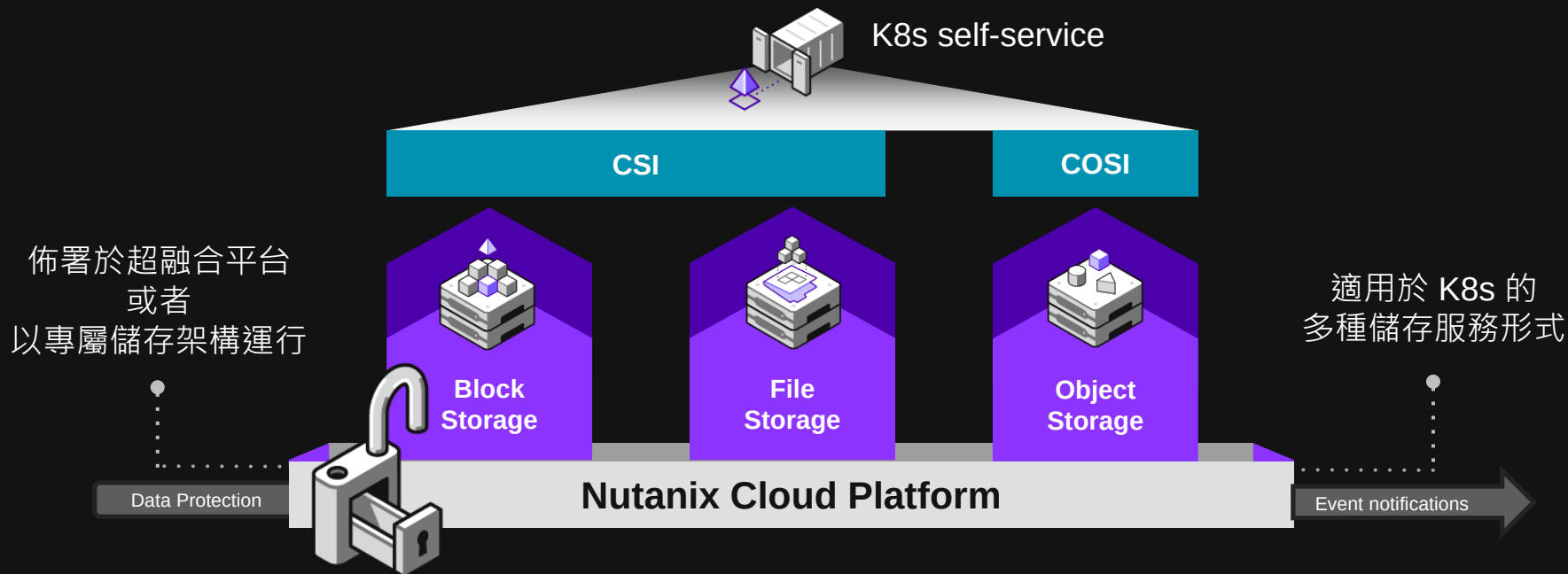
傳統基礎設施
無法跟上

具有 Kubernetes 的
動態和短暫特性.

不要忽視不可避免的事情
必須提前計劃

Stateful workloads
是不可避免的,
應儘早包含在規劃中,
以降低成本和複雜性.

Nutanix 軟體定義儲存服務



Software Defined Data Services | Security-First Design | Start Small and Scale Out | Flexible Consumption

成功 Kubernetes 轉型



Freedom of Choice

整合 Nutanix 平台
完全支援公有雲
和其他虛擬化平台



One Unified Platform

適用您
運行於任何位置的
所有應用程式



Power of Open

純粹 Open
無綁定的
Kubernetes 平台



Enterprise Class Data Services

結合 Nutanix 特有的
儲存與數據服務
完備 K8s 的部署與維運

人工智慧應用系統 受到 C-Level 關注

將 Enterprise AI 用於
企業業務？這重要嗎？

領導者點頭。

49%

的 CEO 受訪者制定了
包含 GenAI 的員工生產力計劃

Gartner

成功的 AI 部署之路是什麼？

要能成功的使用 **Enterprise AI** 需要先回答以下關鍵問題：

我們應該使用哪些
AI 模型？

如何採用 GenAI ？

如何擴展基礎架構和運營
以滿足 GenAI 需求？



NAI 的架構與運作

透過簡易的介面
部署您選擇的大型語言模型
並建立 GenAI endpoints,
使用者就可以透過 secure APIs
進行模型推理。

Sizing and integration services

Enterprise GenAI Applications

Foundational
Models

Nutanix
Enterprise AI

**Nutanix
Kubernetes Platform**

Nutanix Cloud Infrastructure

Nutanix
Unified
Storage
(Optional)

CPU Enabled Servers



Nutanix Enterprise AI 運作流程

| 23

1.



Nutanix
Enterprise AI



Nutanix
Kubernetes Platform

Red Hat OpenShift
Azure AKS | AWS EKS | GCP
Rancher | Docker

在任何 Kubernetes 平台
佈署和運行
Nutanix Enterprise AI.

2.



Hugging Face



Your Own
Model

登入 NAI 管理介面
下載並佈署您的 LLM.

3.



API-to-Model
User API



Hugging Face

Test model access with a
GenAI question.

針對您的模型建立一個
secure API key 並進行測試.

4.



API key:
B23014KJH.....

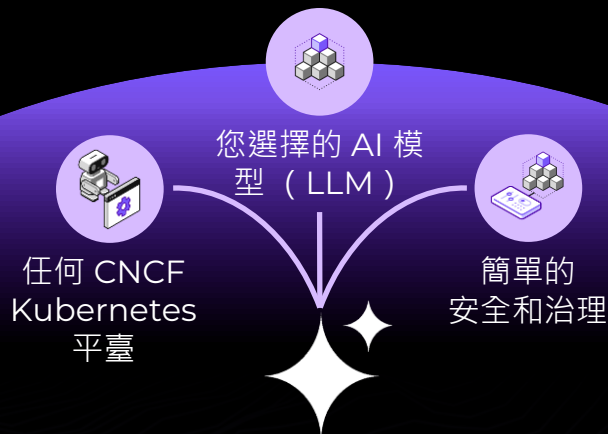


提供該 API key 給
開發者或應用系統使用者,
進行模型推理.



監控 LLMs 和 APIs 執行狀態

✦ Nutanix Enterprise AI 帶來關鍵效益



Nutanix Enterprise AI

簡單

一個簡潔、簡單的介面，用於 AI 模型選擇和部署，以運行 GenAI 應用程式。

管控

IT 管理員能夠進行 AI 部署的維運操作，控制應用程式存取、端點和 AI 模型。

成本

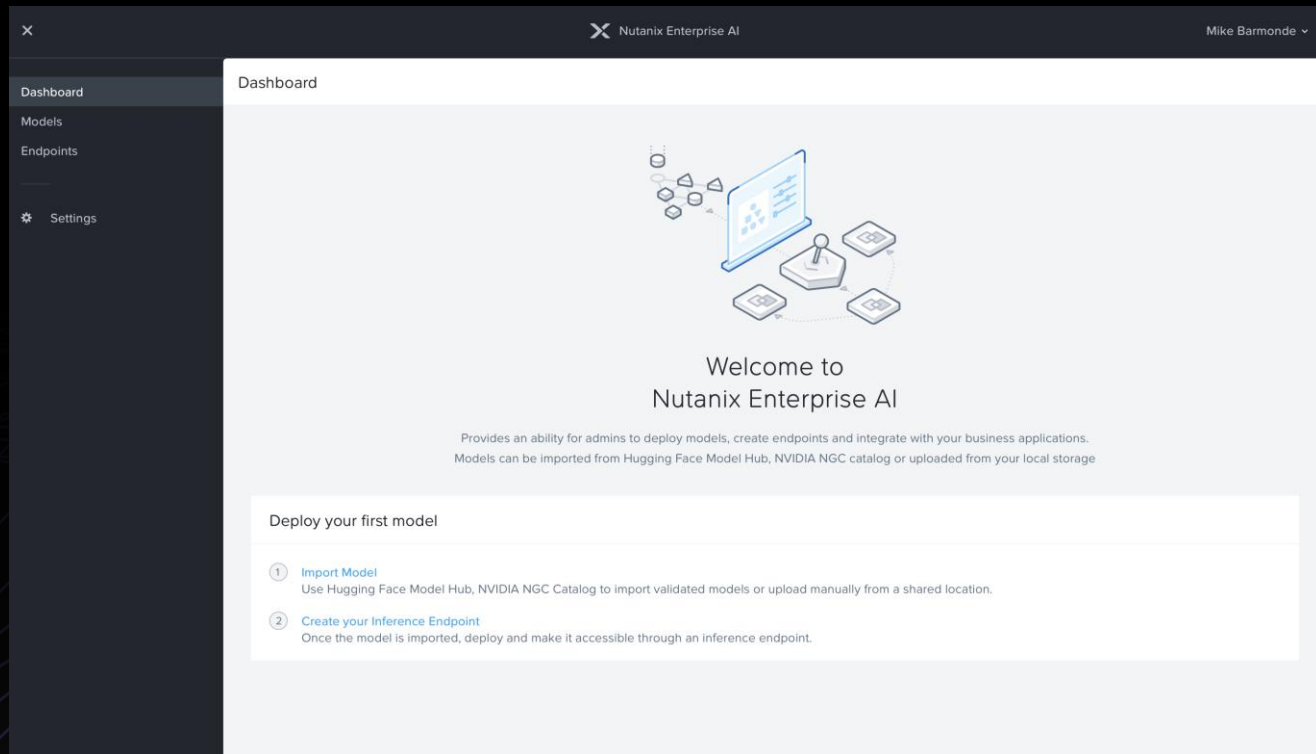
確保在本地或公有雲端運行 AI 的總體擁有成本 (TCO) 是可預測的。



Simple and Intuitive UI

| 25

Wizard-driven user experience.

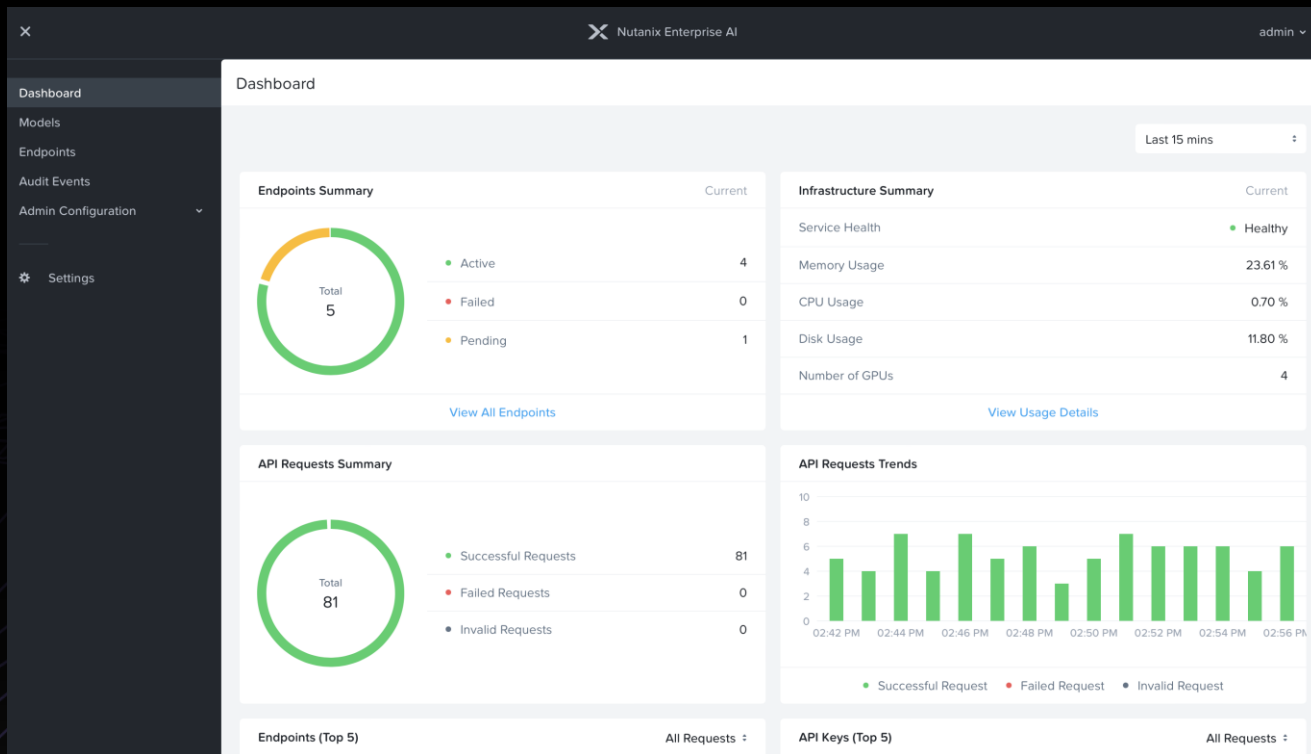




Auditing + Monitoring

| 26

Easily understand
infrastructure
health and
security.





Choice of AI Models

| 27

Choose from
validated AI
models or import
a custom model.

Nutanix Enterprise AI Mike Barmonde

Dashboard
Models
Endpoints
Settings

Models



No Models Available

Get started by importing from Hugging Face, NVIDIA NGC catalog or upload model manually. Pre-validated models can be imported from Hugging Face, NVIDIA NGC or models can be manually uploaded from a file share or a bucket.

From NVIDIA NGC Catalog

Import models which are pre-validated to run on Nutanix directly from NVIDIA NGC catalog.

[Import from NVIDIA NGC Catalog](#)

From Hugging Face Model Hub

Import models which are pre-validated to run on Nutanix directly from Hugging Face.

[Import from Hugging Face Model Hub](#)

Import Model

Package the models in a compatible format and manually upload the models from a file share or bucket.

[Upload Manually](#)





Secure Access Controls

Role-based access controls (RBAC) for endpoints.

The screenshot displays the 'Users' management page in the Nutanix Enterprise AI interface. A modal window titled 'Add User' is open, allowing the creation of a new user. The modal contains the following fields:

- Role:** A dropdown menu with 'AI / ML User' selected.
- Email Address:** A text input field containing 'Jeffery.Lebowski@thedude.biz'.
- First Name:** A text input field containing 'Jeffery'.
- Last Name:** A text input field containing 'Lebowski'.
- Username:** A text input field containing 'TheDude'.
- Password:** A password input field with a 'Show' toggle.

At the bottom of the modal are 'Cancel' and 'Add' buttons. In the background, a table lists existing users with columns for Name, Status, and Date Added. The table shows 10 rows of data, with the first row being 'nutanix.c' and the last row being 'nutanix.co'.

Name	Status	Date Added
nutanix.c	Active	07-Nov-2024
nutanix.c	Active	05-Nov-2024
@nutanix	Active	05-Nov-2024
om	Active	04-Nov-2024
nutanix.c	Active	28-Oct-2024
c.com	Active	20-Oct-2024
o.com	Active	24-Oct-2024
utanix.co	Active	23-Oct-2024
	Active	23-Oct-2024



Endpoints and APIs

Easily create and manage endpoints and APIs.

Create an Endpoint

Number of GPUs (Per Instance)

1

GPU Card

NVIDIA-L40S

☐ Use Custom Configuration for Instances

API Key Details

The key has been generated. Copy the key and store it securely. You cannot view the key again after you close this dialog box. If you lose the key, you must generate a new one.

Key Name

mykey

API Key

[REDACTED]

Copy Key

Close

The endpoints can be accessed using an API Key. New Keys can be generated or existing keys can be used.

API Keys

mykey

List of active API Keys created by you

Cancel

Create



Pre-flight Model Testing

| 30

Test your AI model using a natural language interface.

The screenshot displays the Nutanix Enterprise AI interface. On the left is a sidebar with navigation options: Dashboard, Models, Endpoints, Audit Events, Admin Configuration, Users, Compute, and Settings. The main panel shows the details for the 'nim-llama3-1' model. A 'Test Endpoint' dialog box is open in the center, featuring a 'Test' button and a 'Done' button. The dialog box contains the following information:

- Endpoint Name:** nim-llama3-1
- Model Instance:** NIM-llama-3.1-8b-instruct
- Requests:**
 - ☒ Sample Request
 - ☐ Custom Request
- Explain how Deep Neural Networks work in simple terms**
- Describe the landscapes of Mars in a few sentences**

The background shows the 'nim-llama3-1' model details, including a 'Status' of 'Succeeded' and a 'Result' of 'Mars' landscapes feature vast plains, towering volcanoes, deep valleys, and sprawling canyons. The largest canyon, Valles Marineris, stretches over 4,000 km and is among the largest in our solar system. Dust storms can last for weeks or even months, largely covering the planet's reddish-orange terrain.'

Below the dialog box, there is a table of requests with columns for 'Request ID', 'Status', and 'Latency'. The table shows three requests, all with a status of 'Succeeded' and a latency of '25 ms'.



單一平台 多元應用 混合多雲 自在悠遊



企業核心
應用系統



雲原生應用



大數據分析



各種規模
資料庫系統



生成式 AI



邊緣運算



終端使用者
運算



私有雲



公有雲

Nutanix
混合多雲
平台



Thank you.

NUTANIX™