DELL Technologies

MYELINTEK 邁爾凌科技

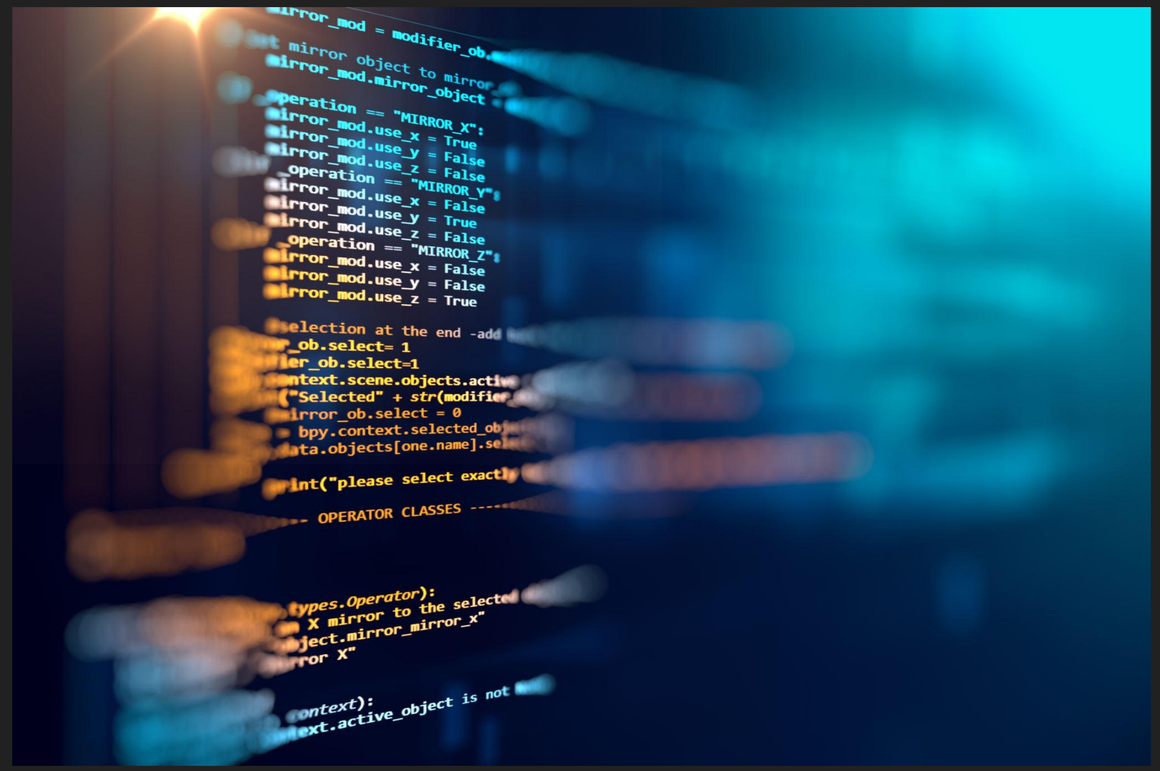
You Focus on Your AI Missions
We Handle the AI Environment

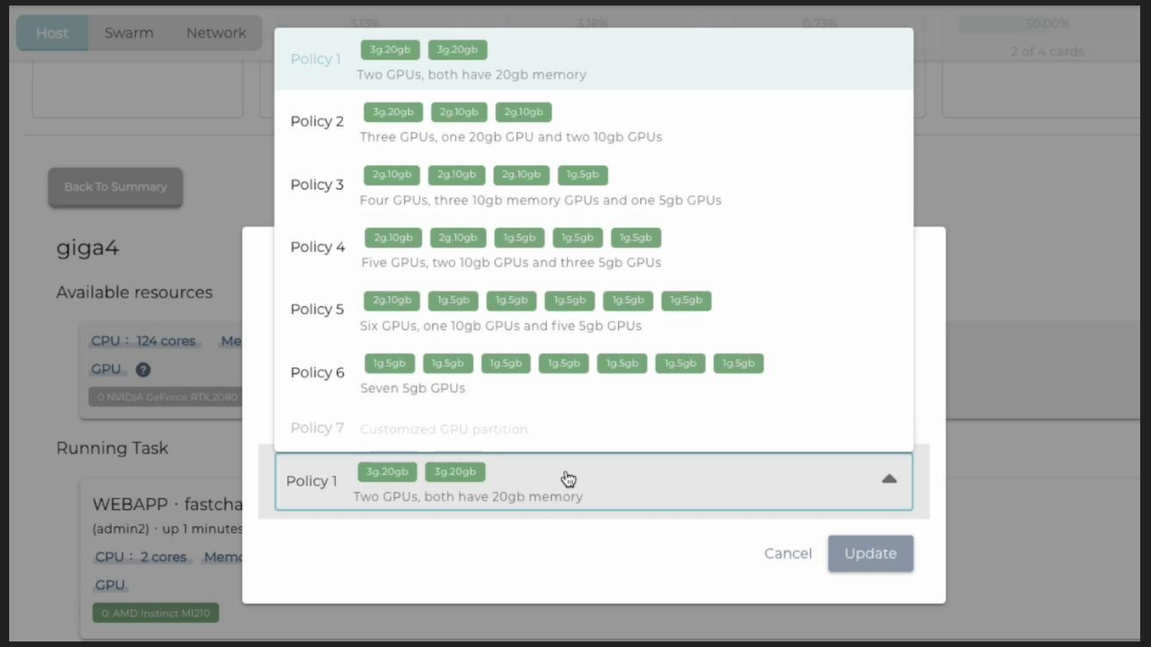
從硬體到模型
一站到底的AI行動方案

BDM Brad Huang
業務發展經理 黃丞毅

前言

前言

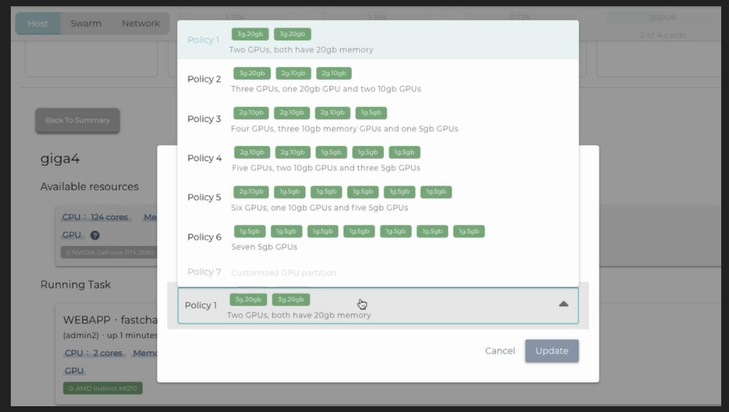

mirror_mod = modifier_ob.
mirror object to mirror
mirror_mod.mirror_object
operation == "MIRROR_X":
irror_mod.use_x = True
irror_mod.use_y = False
irror_mod.use_z = False
operation == "MIRROR_Y":
irror_mod.use_x = False
irror_mod.use_y = True
irror_mod.use_z = False
operation == "MIRROR_Z":
irror_mod.use_x = False
irror_mod.use_y = False
irror_mod.use_z = True
selection at the end -add
_ob.select= 1
er_ob.select=1
ntext.scene.objects.activ
("Selected" + str(modifier
irror_ob.select = 0
bpy.context.selected_obj
ata.objects[one.name].sel
int("please select exactly
--- OPERATOR CLASSES ---
types.Operator):
X mirror to the selected
ject.mirror_mirror_x"
rror X"
context):
xt.active_object is not


Host
Swarm
Network
Back To Summary
giga4
Available resources
CPU : 124 cores
GPU
0: NVIDIA GeForce RTX 2080
Running Task
WEBAPP · fastcha
(admin2) · up 1 minutes
CPU : 2 cores
GPU
0: AMD Instinct MI210
Policy 1
3g.20gb
3g.20gb
Two GPUs, both have 20gb memory
Policy 2
3g.20gb
2g.10gb
2g.10gb
Three GPUs, one 20gb GPU and two 10gb GPUs
Policy 3
2g.10gb
2g.10gb
2g.10gb
1g.5gb
Four GPUs, three 10gb memory GPUs and one 5gb GPUs
Policy 4
2g.10gb
2g.10gb
1g.5gb
1g.5gb
1g.5gb
Five GPUs, two 10gb GPUs and three 5gb GPUs
Policy 5
2g.10gb
1g.5gb
1g.5gb
1g.5gb
1g.5gb
1g.5gb
Six GPUs, one 10gb GPUs and five 5gb GPUs
Policy 6
1g.5gb
1g.5gb
1g.5gb
1g.5gb
1g.5gb
1g.5gb
1g.5gb
Seven 5gb GPUs
Policy 7
Customized GPU partition
Policy 1
3g.20gb
3g.20gb
Two GPUs, both have 20gb memory
Cancel
Update

前言

前言

時效性



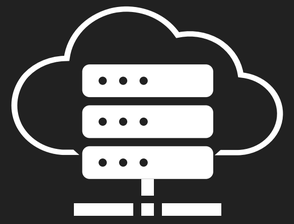
操作直觀性

整合性

開發者的運算服務需求

Bare Metal as a Service

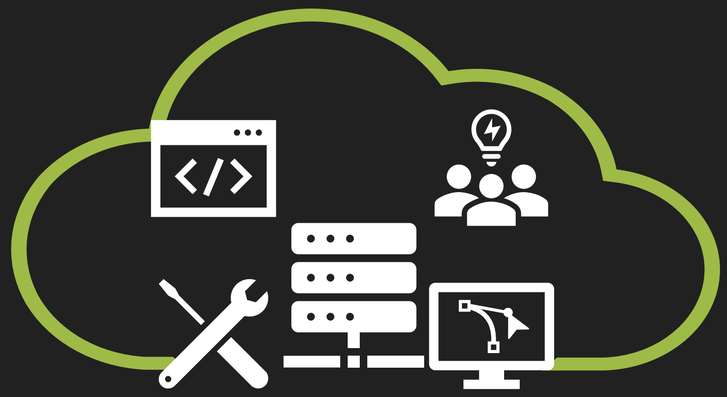
(or Infrastructure as a Service / IaaS)



以提供硬體算力
服務為主

AI as a Service

(or Software as a Service / SaaS)



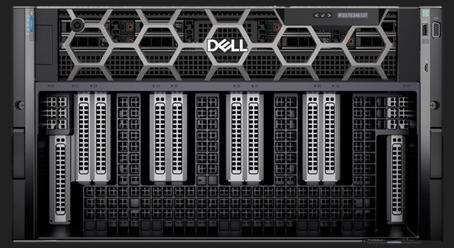
- 硬體算力服務
- AI 開發IDE與各式工具
- AI 模型樣板
- AI 開發環境
- 團隊協同開發合作

AI開發需求內容

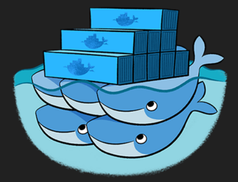
DELL Technologies

管理節點

運算節點



容器化管理

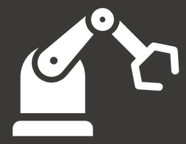


AI / MLOps 開發平台

管理

整合

介接



儲存節點

組織既有

- 帳號管理系統
- 計費管理系統
- 通訊管理系統

從硬體資源到軟體整合
從模型開發到推論服務
一站式．完整開發環境

AI開發平台需求對象

IT / MIS團隊

- [x] 系統叢集與備援
- [x] GPU算力分割
- [x] 硬體資源限制
- [x] 訓練排程規劃
- [x] 開發者權限管控

模型開發團隊

- [x] 自動標註功能
- [x] 開發IDE與標註工具整合於平台
- [x] 支援NVAIE / NIM與開源模型樣板及框架
- [x] 資料集與模型版本管控
- [x] 開發流程標準化

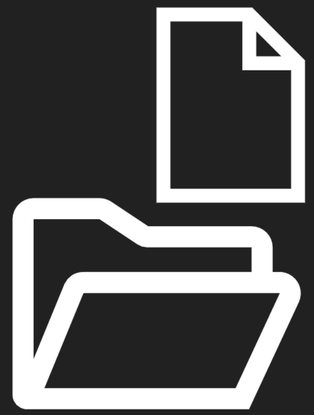
AI維運團隊

- [x] 可於平台直接啟動模型推論服務
- [x] 模型加密機制
- [x] 自動啟動模型推論服務功能
- [x] 模型效能監控

平台在模型開發過程所支援的各項功能

平台管理

- 平台將管理、運算、儲存節點叢集納管，並進行監控、分配，支援跨節點運算
- 可透過平台GUI進行GPU分割
- 平台可自主設定運算資源限制，亦可限制GPU使用數量及時間
- 平台支援開發者權限、資料集、模型版本管控，設定使用排程並記錄使用時間



資料處理

- 完整資料處理功能支援資料視覺化預覽
- 透過介面將桌面資料拖曳上傳至平台

資料標註

- 內建常用資料標註工具
- 資料自動標註功能

模型開發

- 內建常用模型樣板與框架
- 平台介面直接下載NGC及開源資源，如NVAIE、NIM、Huggingface等

模型訓練

- 訓練參數、過程與產出物版本控制
- 可於平台同時進行多組參數訓練
- 即時顯示資源使用率

模型評估

- 平台介面訓練數值可視化並可比較多筆實驗結果
- 將驗證資料集透過平台餵進程式碼產生驗證結果

流程標準化

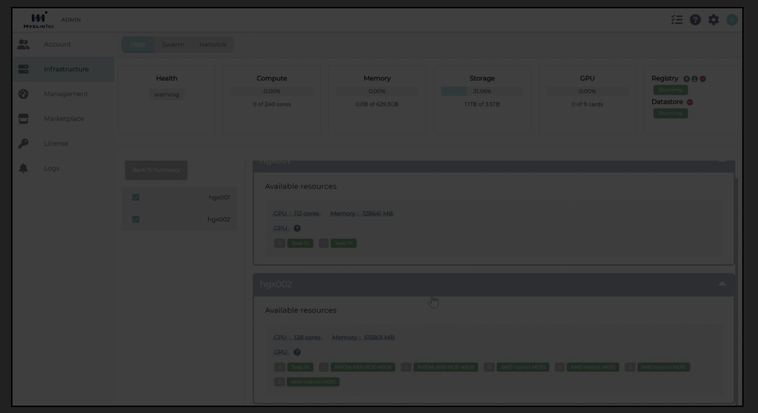
- 使用平台Pipeline功能建構模型產出標準化流程，以利模型優化與模型再利用

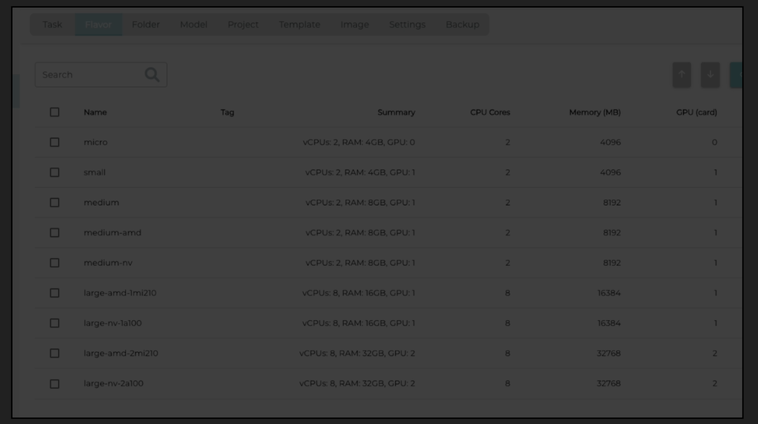
模型佈署

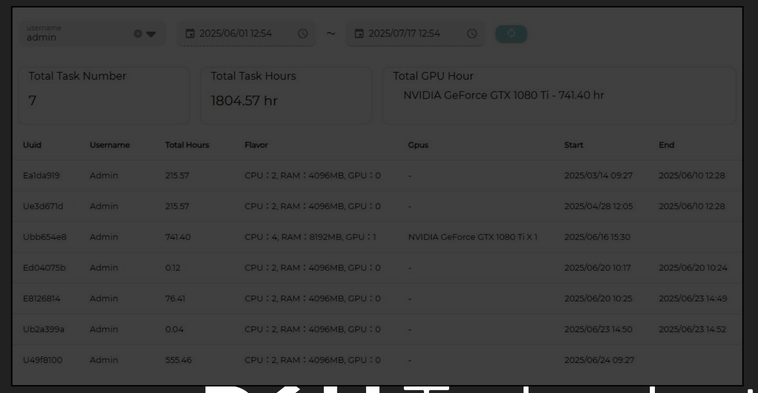
- 啟用WebApp功能於平台以API、IP方式介接前端系統，或將模型打包成image
- 模型加密機制，提升模型安全性

管理 – 硬體資源監控與管理

- 平台支援DELL全系列伺服器，可納管運算、儲存、管理節點建構資料中心等級的伺服器叢集
- 管理者可自定義資源使用限制，以供各種模型之硬體配置所需
- 視覺化界面方便辨識既有硬體資源空間與模型專案的使用情況，並可直接關閉容器服務以回收相應的硬體資源
- 可以呈現使用者所佔用的資源以及其所佔用的時間，並產出對應的報表供管理者參考。 使用者在的資源使用率可至監控介面個別查詢





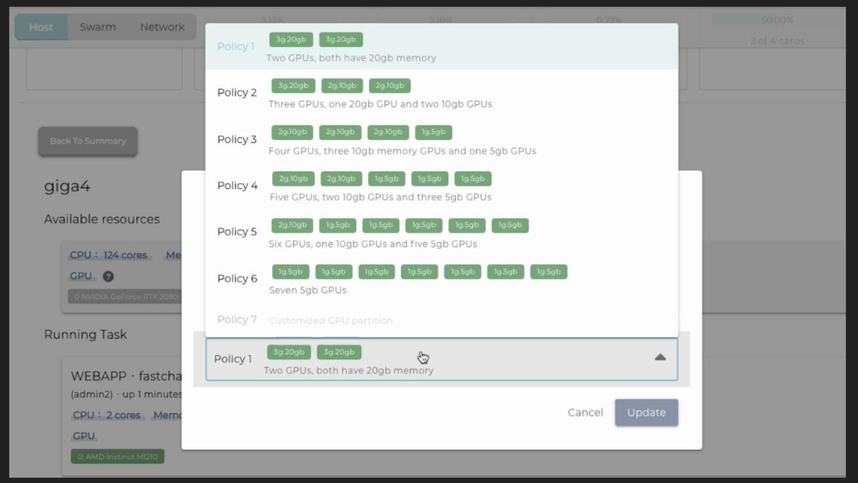


MYELINTEK 邁爾凌科技

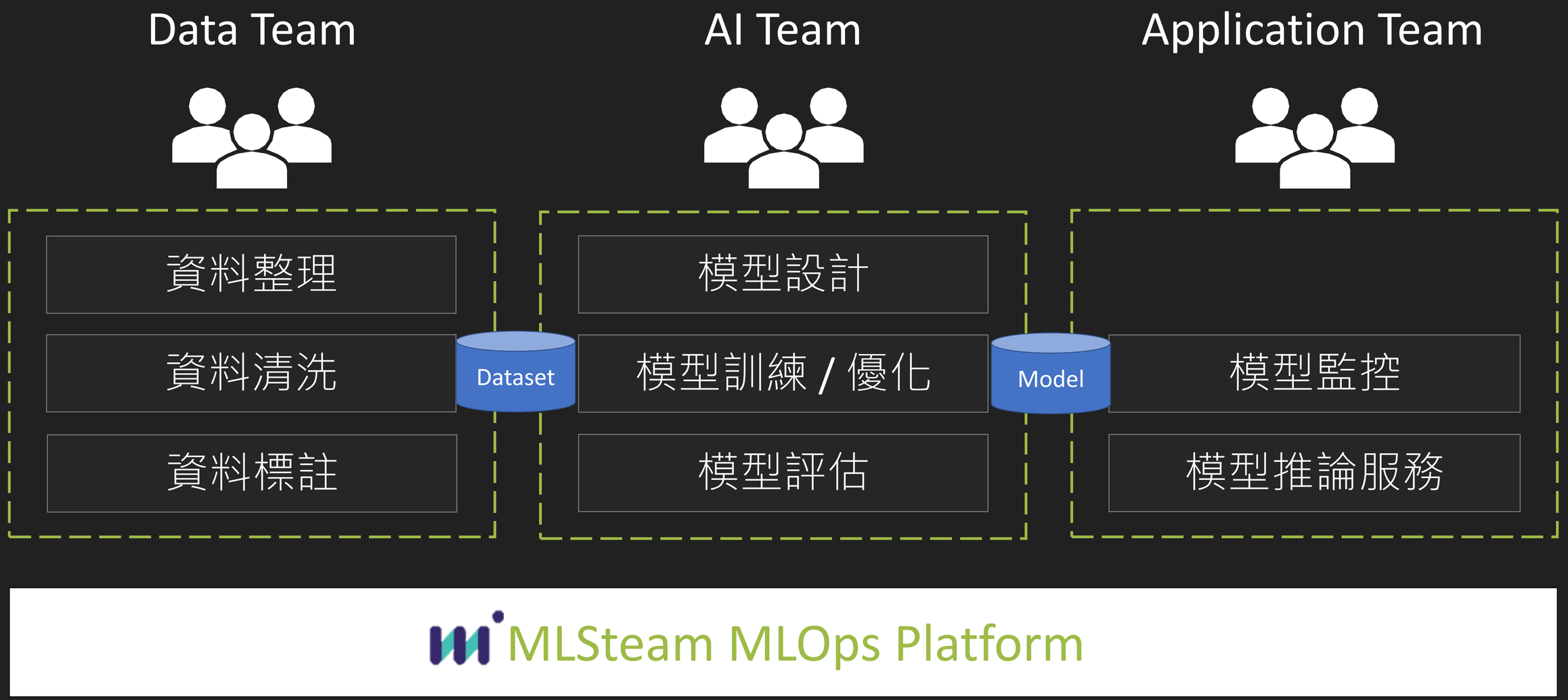
DELL Technologies

管理 – GPU運算管理

- 將NVIDIA MIG功能以下拉式選單操作，快速進行配置
- 平台支援NVIDIA GPU Fraction管理功能，提高每單位運算產能
- 平台提供預留GPU運算資源功能
- 平台可記錄GPU運算使用額度，並產生報表

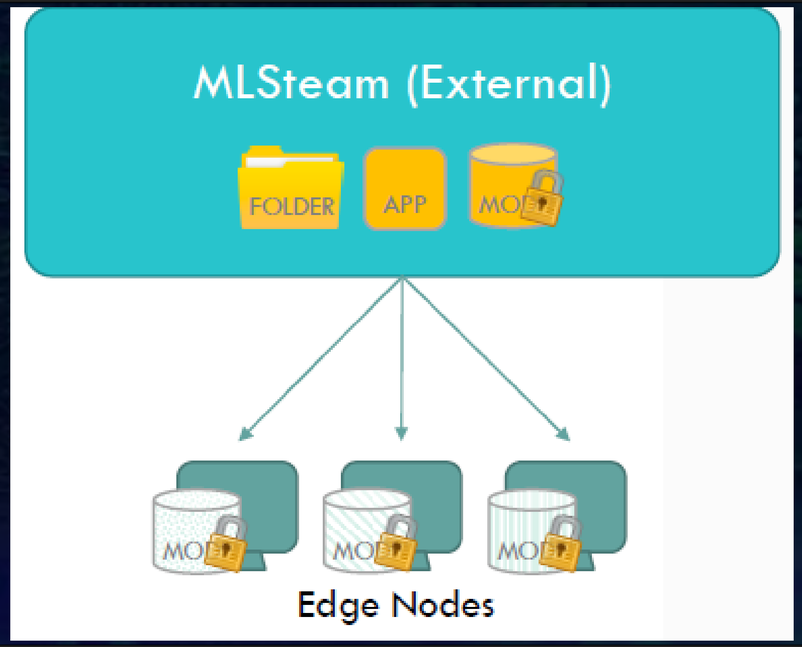


管理 – 專案協同開發


Data Team
AI Team
Application Team
資料整理
資料清洗
資料標註
Dataset
模型設計
模型訓練 / 優化
模型評估
Model
模型監控
模型推論服務
MLSteam MLOps Platform

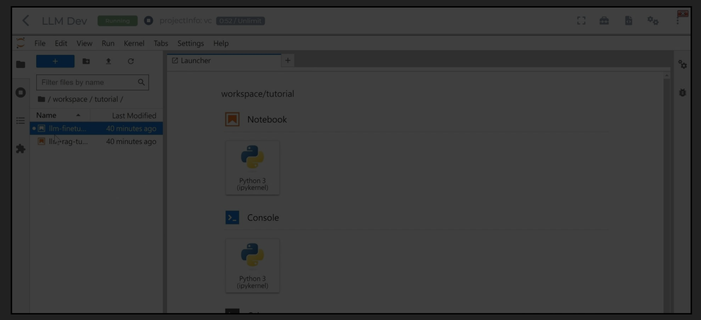
管理 – 平台安全機制

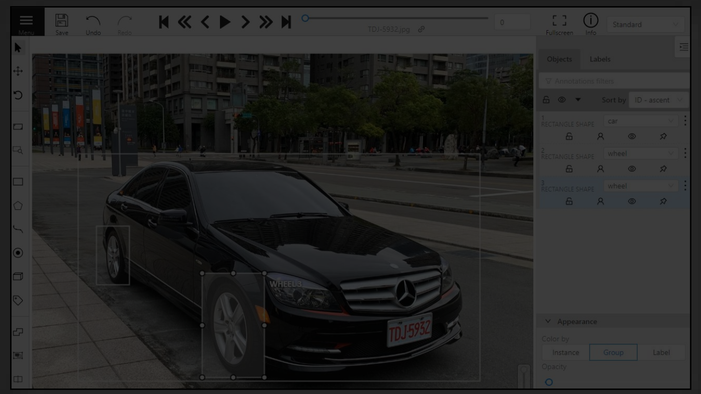
- 模型開發環境與平台介面共用網路埠對外提供服務，無須額外開放網路埠
- 各專案間的獨立作業環境 (Isolation Execution Environment)
- 資料集權限控制管理 (Dataset permission control)
- 模型部署保護機制 (Model deployment protection Mechanism)

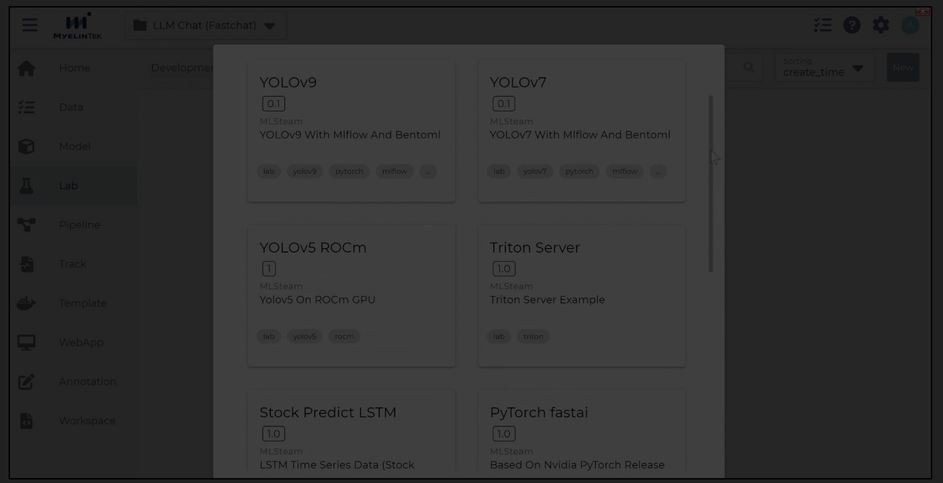


整合 – 建構即用的開發環境

- 提供JupyterNotebook與Terminal介面的模型開發環境
- 平台支援PyTorch、TensorFlow等深度學習框架
- 平台可支援NVIDIA CUDA、AMD ROCm、Intel OneAPI等異質運算
- 平台內建CVAT、LabelStudio等標註工具
- 預建Llama、YOLO等常用之模型樣板，其版本隨平台進行更新
- 容器啟動可透過GUI設定SSH port，不用編輯yaml

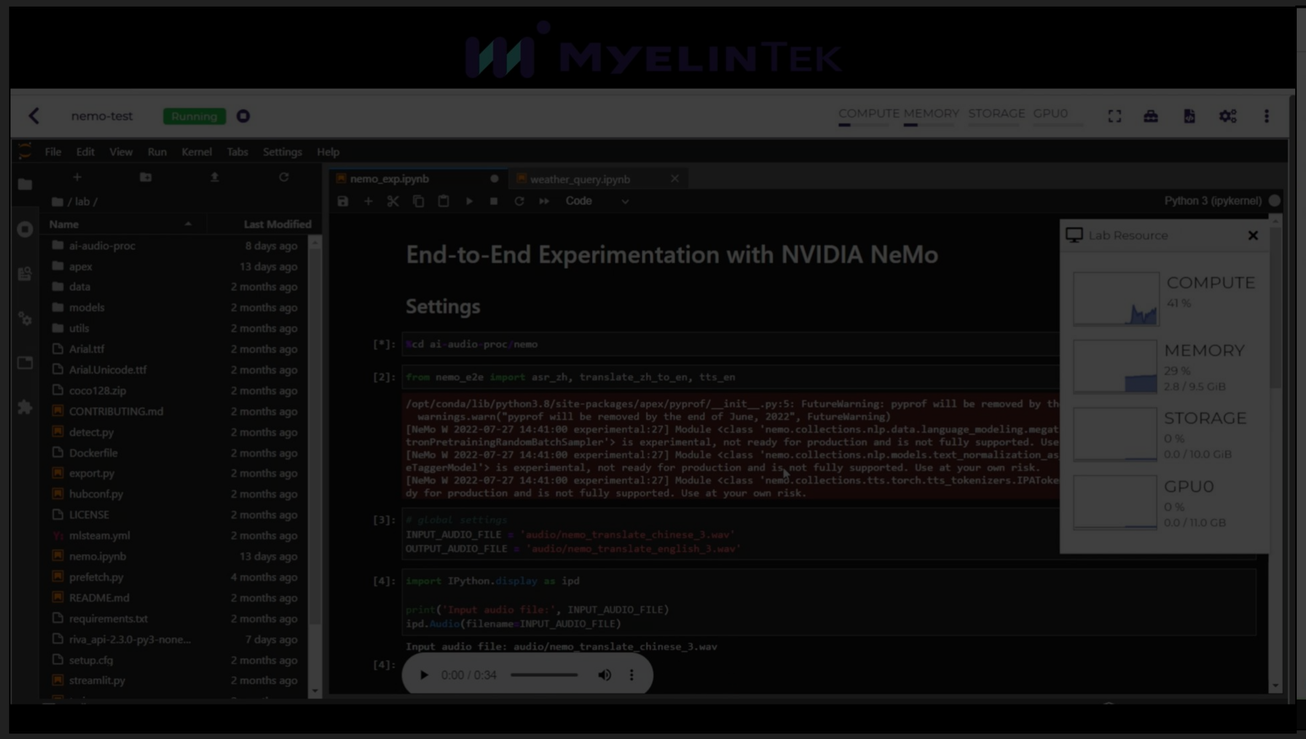


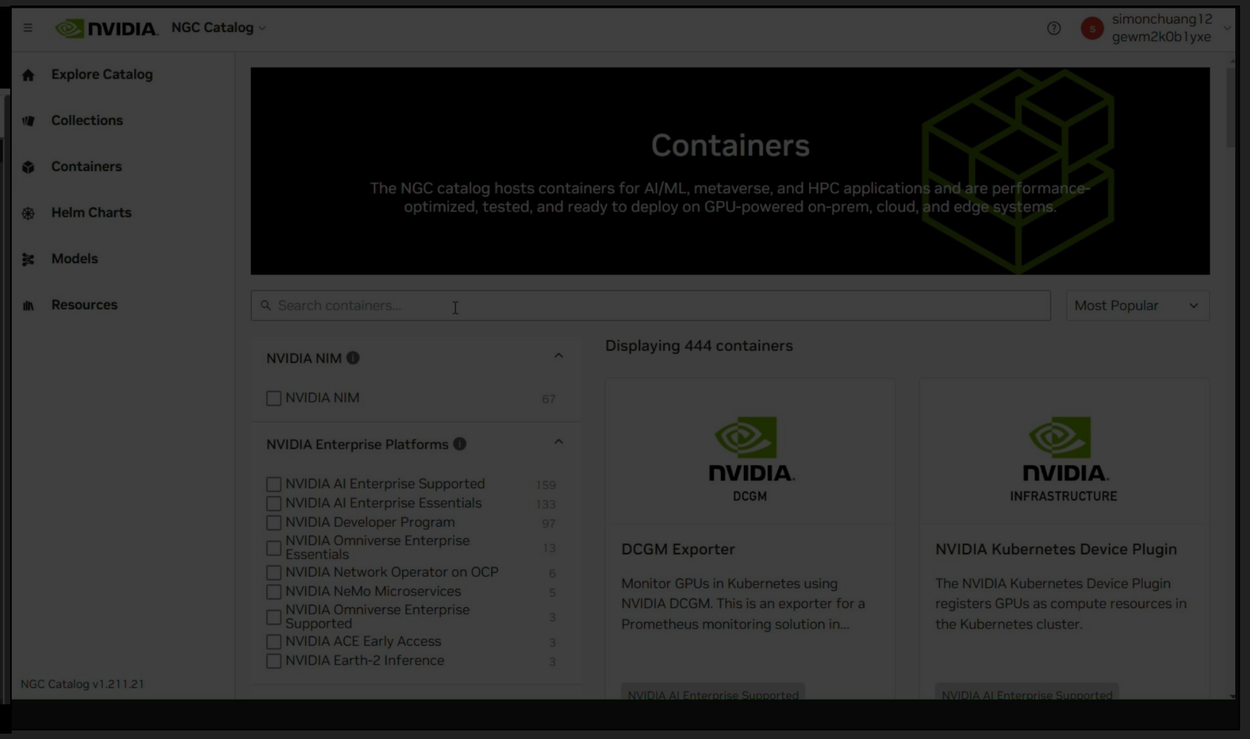




整合 – 支援NVAIE與NIM

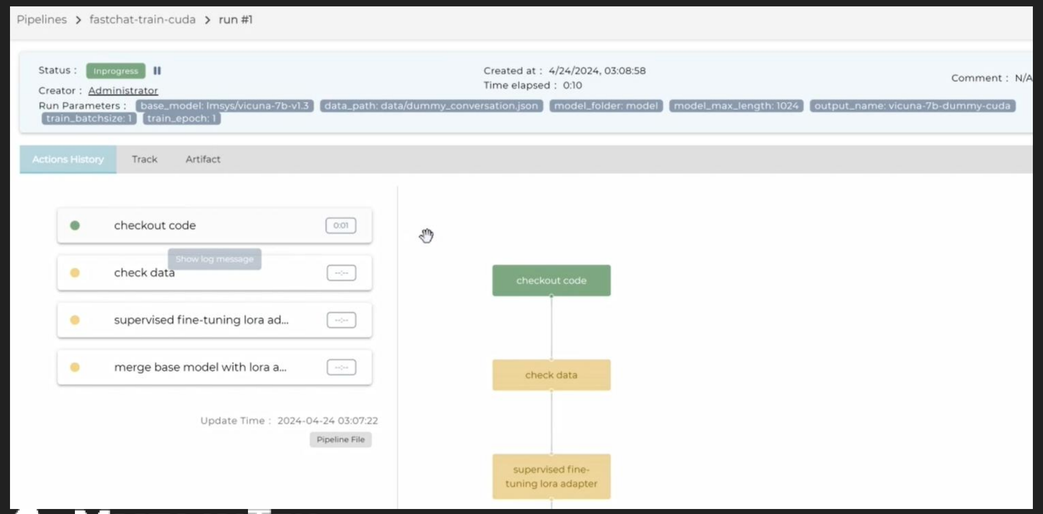
- 平台支援NVAIE模型開發工具及框架，如：NeMo、RIVA、MONAI…
- 透過平台介面，易於使用NVIDIA NIM(微服務)





整合 – Pipeline/MLFlow功能

- CI / CD整合，助於持續性的模型訓練與優化
- 直觀且視覺化的進度顯示與日誌紀錄
- 依照資源需求進行調度，將硬體資源利用最大化
- 使用YAML檔案輕鬆定義執行相依性 (DAG)
- 透過Pipeline定義執行內容，並安排排程讓服務在指定期間啟動，執行完畢後自動關閉將資源釋放，以進行批次推論服務



介接

- 平台支援LDAP等管理系統
- 平台可與儲存系統進行介接
- 管理者可透過平台的quota管理功能與既有計費系統進行介接
- 平台可介接端點系統 (edge伺服器、AIoT)，進行模型部署、版本管理、資料收集等
- 可於平台啟動WebApp進行模型推論服務

總結 – 邁爾凌AI Platform

- 工研院spin-off的擁有多年AI開發團隊
- 安裝於地端的一站式AI開發平台
- 成功案例：電信商、國防、智慧交通、輔助醫療等…
- 平台所提供三大項目：
  - 管理 → 使用者、硬體、資料集、模型版本
  - 整合 → 開發所需工具，如標註工具、開源模型、框架並支援NVIDIA NVAIE、NIM
  - 介接 → 可與儲存及管理系統介接

DELL Technologies

MYELINTEK 邁爾凌科技

AIction Now

Brad Huang 黃丞毅

brad@myelintek.com

+886-933-306-785