PCAI

Modell tanítás és Inference

Wahl László – laszlo.wahl@hpe.com

2025 november

Az AI hatása minden iparágra kiterjed

Pénzügy

Csalásfelismerés
Személyre szabott banki szolgáltatások
Befektetési betekintés

Egészségügy

Molekula szimuláció
Gyógyszerkutatás
Klinikai vizsgálati adatok elemzése

Kereskedelem

Személyre szabott vásárlás
Automatizált katalógusleírások
Automatikus áróptimalizálás

Telekommunikáció

AI virtuális asszisztensek
Hálózati teljesítmény optimalizálás
Távoli támogatási lehetőségek

Média és szórakozás

Karakterfejlesztés
Videószerkesztés és képalkotás
Stílusjavítás

Gyártás

Digitális iker
Terméktervezés
Prediktív karbantartás

Oktatás

Dokumentumok összefoglalása
HPC szimuláció
AI virtuális oktató

Energia

Tudásbázis Kérdések és válaszok
Prediktív karbantartás
Ügyfélszolgálat

“Az A-G(eneral)-I-t használni nem kell félnetek jó lessz”

Gépi Tanulási Algoritmusok

Field	Task	Leading Methods
CV	Semantic Segmentation	HRNet-OCR | Efficient-Net-L2 | ResNeSt-269 | VMVF
CV	Image Classification	FixEfficientNet | BiT-L | Wide-ResNet-101 | Branching CNN
CV	Object Detection	Efficient-Det-D7x | Rodeo | Patch Refinement | IterDet
NLP	Sentiment Analysis	BERT | T5–3B | NB-weighted-BON + dv-cosine
NLP	Language Modeling	Megatron-LM | GPT-3 | GPT-2
NLP	Text Classification	XLNet | USE_T + CNN | SGC
NLP	Question Answering	T5–11B | SA-Net on Albert | TANDA-RoBERTa
NLP	Machine Translation	Efficient-Det-D7x | Rodeo | Patch Refinement | IterDet
RS	Recommender System	Bayesian time SVD++ // flipped w/ Ordered Probit Reg | EASE | H+Vamp Gated
SR	Speech Recognition	ContextNet + Noisy Student | ResNet + BiLSTMs | LiGRU | Large-10h-LV-60k

Non-ML domains that are still AI:

RPA (robotic process automation, chatbot)

“Advanced Analytics” (BI, reporting, typically with SQL)

Miért csak 1 valósul meg a 10 AI pilotból?

Mire is használjuk?

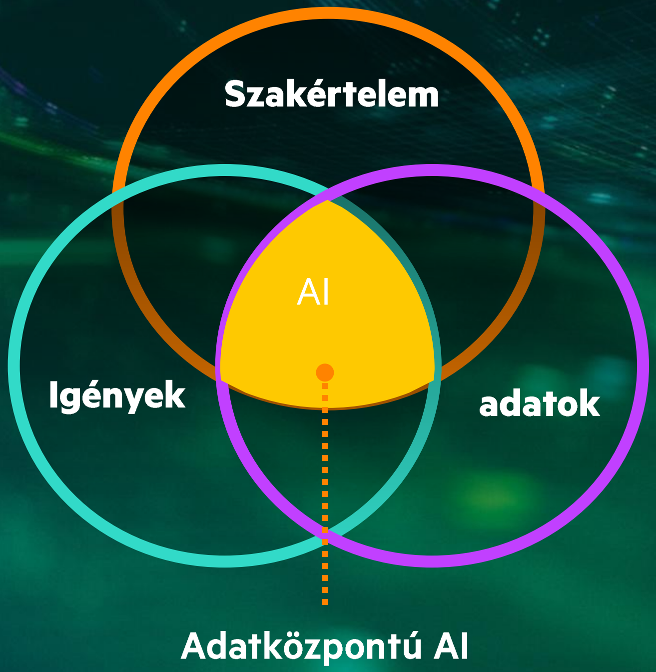
Nehéz olyan üzleti értéket kimutatni, ami igazolná a mégoly sikeres AI projektek bevezetésének IT és üzemeltetés oldali költségét és erőforrásait

Adatvédelem, elérhetőség és adatkezelés

Sok adathoz kellene biztonságosan és menedzselhetően hozzáférni, miközben a változó szabályokat be kell tartani

AI innováció gyorsasága/stabilitás

Az AI modellek és eszközök gyorsabban fejlődnek, mint a belső teamek



Márkaérték és hírnév kockázata | **Biztonsági és adat-védelmi aggályok** | **Speciális szaktudás hiánya** | **Infrastruktúra kihívások PoC -> éles üzem**

HPE AI portfólió

AI technológia

Modell előállító

AI generalista

Modell fogyasztó

HPE AI Factory Cloud Platform

- Egyedileg összeállított
- Multi-tenant, air-gapped
- Skálázható

HPE Private Cloud for AI

- Eszközalapú megoldás (appliance)
- Kulcsrakész
- Integrált menedzsment

HPE AI Pod

- BareMetal/virtuális
- Standard infra
- Kézi menedzsment
- Inference fókusz

Dobozos megoldások

Testre szabott megoldások

Egyedi megoldások

Tipikus használati eset

HPE Private Cloud AI – partnerek számára készült

AI models

AI software

AI infrastructure

AI services

HPE Private Cloud AI

Közösen fejlesztett megoldások a vállalati mesterséges intelligencia egyszerűsítésére

Hewlett Packard Enterprise | NVIDIA

NVIDIA AI Computing a HPE-től

HPE Private Cloud AI

AI Models
- Community + Partner + Custom NVIDIA NIM

AI Software
- NVIDIA AI Enterprise
- HPE AI Essentials

AI Infrastructure
- NVIDIA GPU/GPU NVIDIA Spectrum-X
- HPE AI Servers HPE File Storage

AI Services
- HPE Provided
- 3rd Party Partners

HPE Private Cloud Control Plane
- Setup, inventory, IT and Asset Lifecycle Management
- Munkafolyamatok előkészítése, összehangolása és kezelése
- VM runtime és Kubernetes cluster menedzsment
- Monitoring, Observability és E2E biztonság alkalmazások és infrastruktúra számára

Small | **Medium** | **Large** | **X Large**

A vállalati AI-feladatokhoz igazított konfigurációk

HPE GreenLake cloud

Cloud Administrator

Software (HPE AI Essentials)

- **Minden eszköz minden lépéshez.**
  A HPE AI Essentials teljes csomagot biztosít, a legjobb adatfeldolgozási, elemzési és AI/ML nyílt forráskódú eszközökkel.

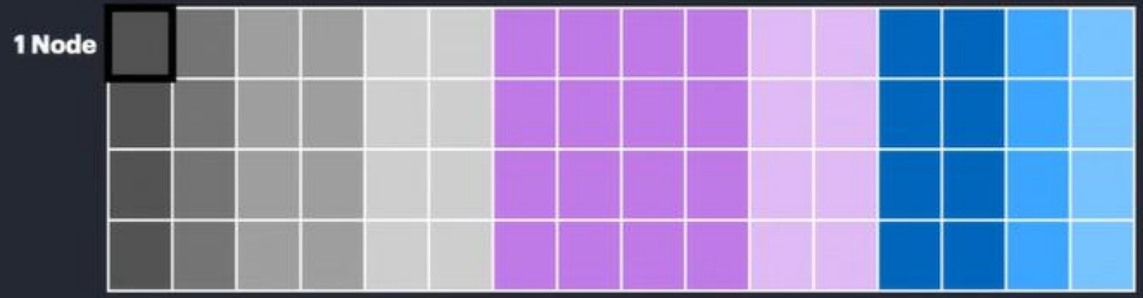
- **Az alkalmazáskezelés fájdalmainak megszüntetése.**
  Az összes eszköz, amelyre az adat- és AI-csapatoknak szükségük van, könnyen telepíthető, konfigurálható és skálázható egyetlen felületen keresztül.

- **Könnyű telepítés a HPE Private Cloud AI-ra.**
  Telepítse az összes szükséges eszközt tetszőleges számú csomópontra - legyen az Edge, a nyilvános felhő, vagy telephely - egyetlen telepítőprogramból.

HPE AI Essentials

NVIDIA GPU

HGX B200
AI Factory Workloads



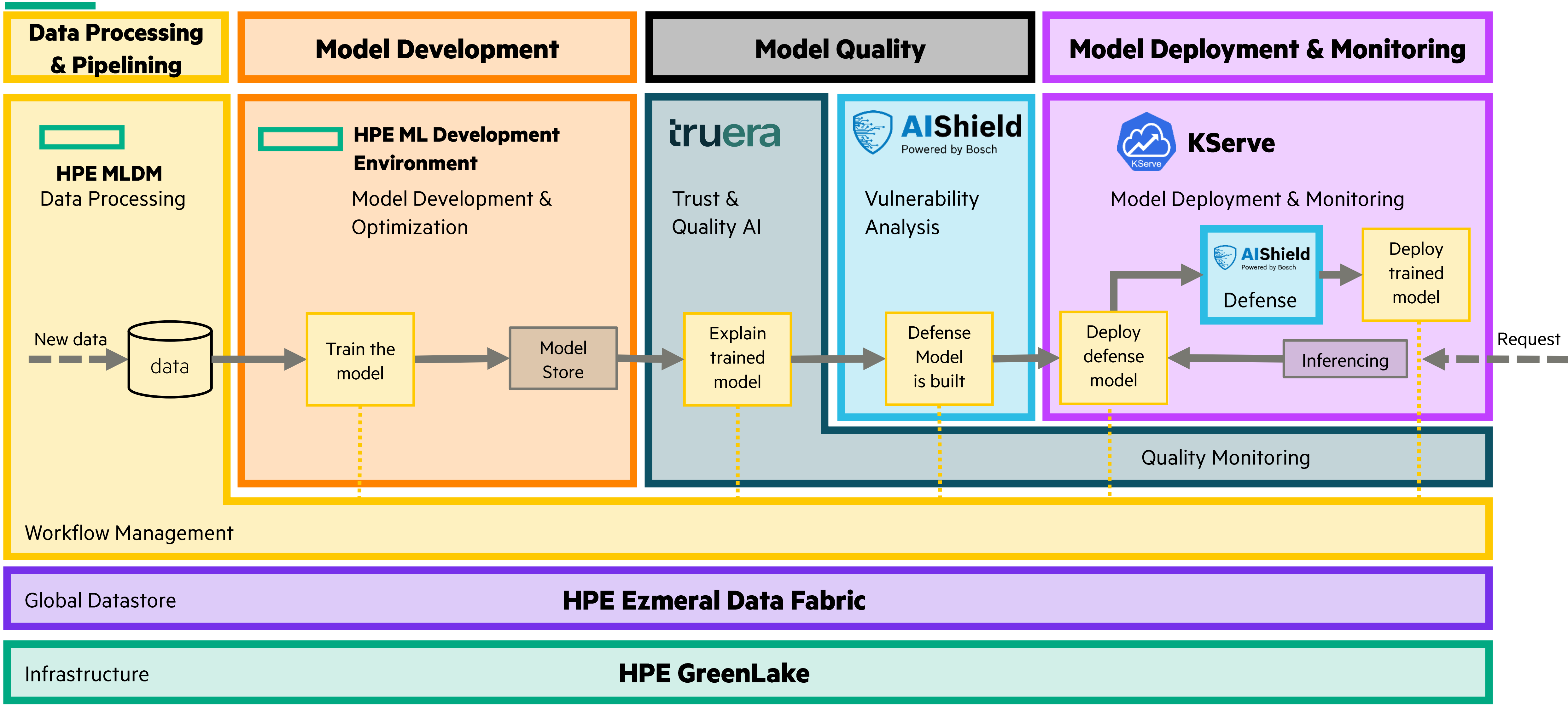
Heavy Model Training and Inference Workloads

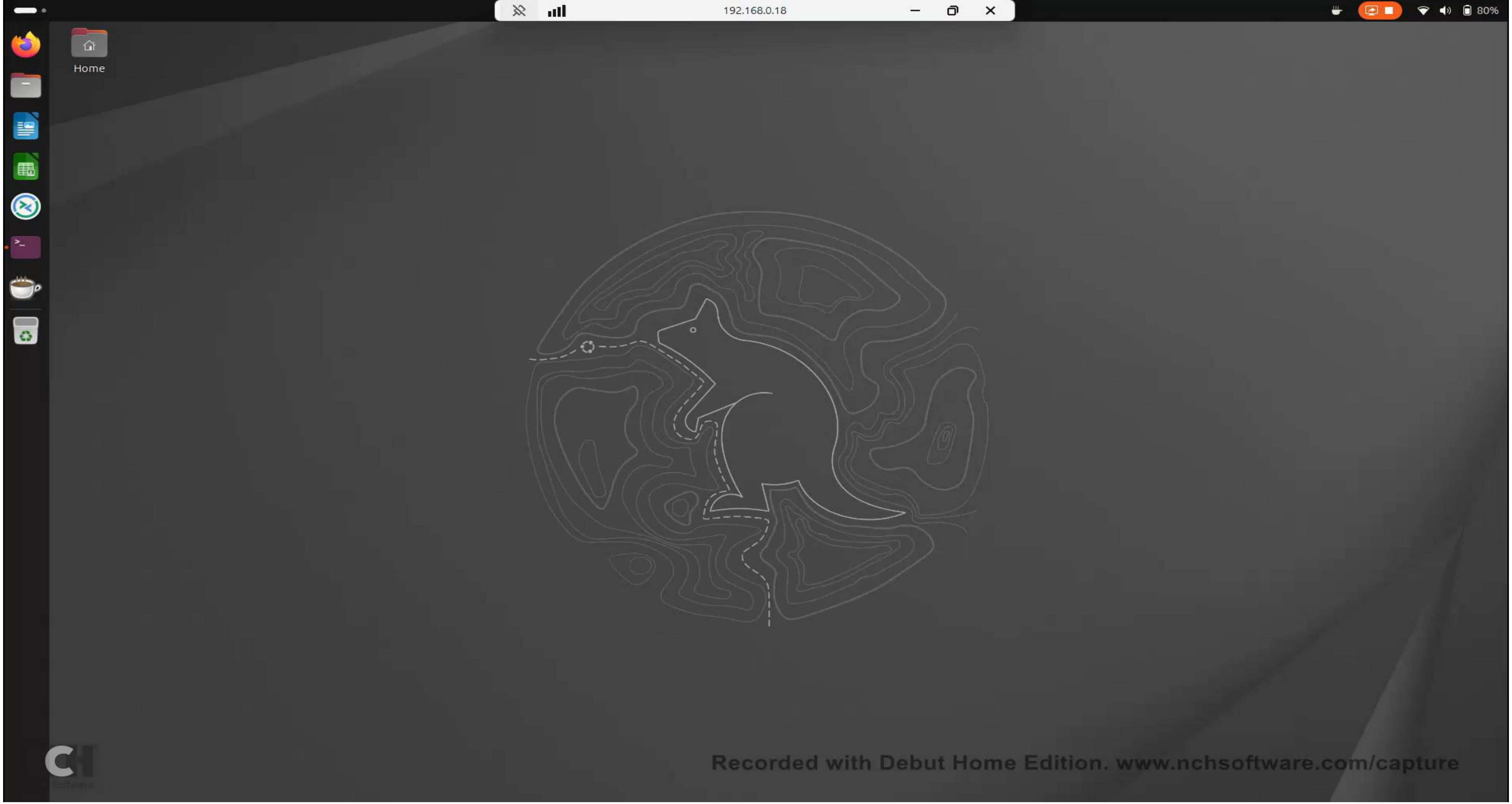
RTX PRO Server
AI Factory Workloads

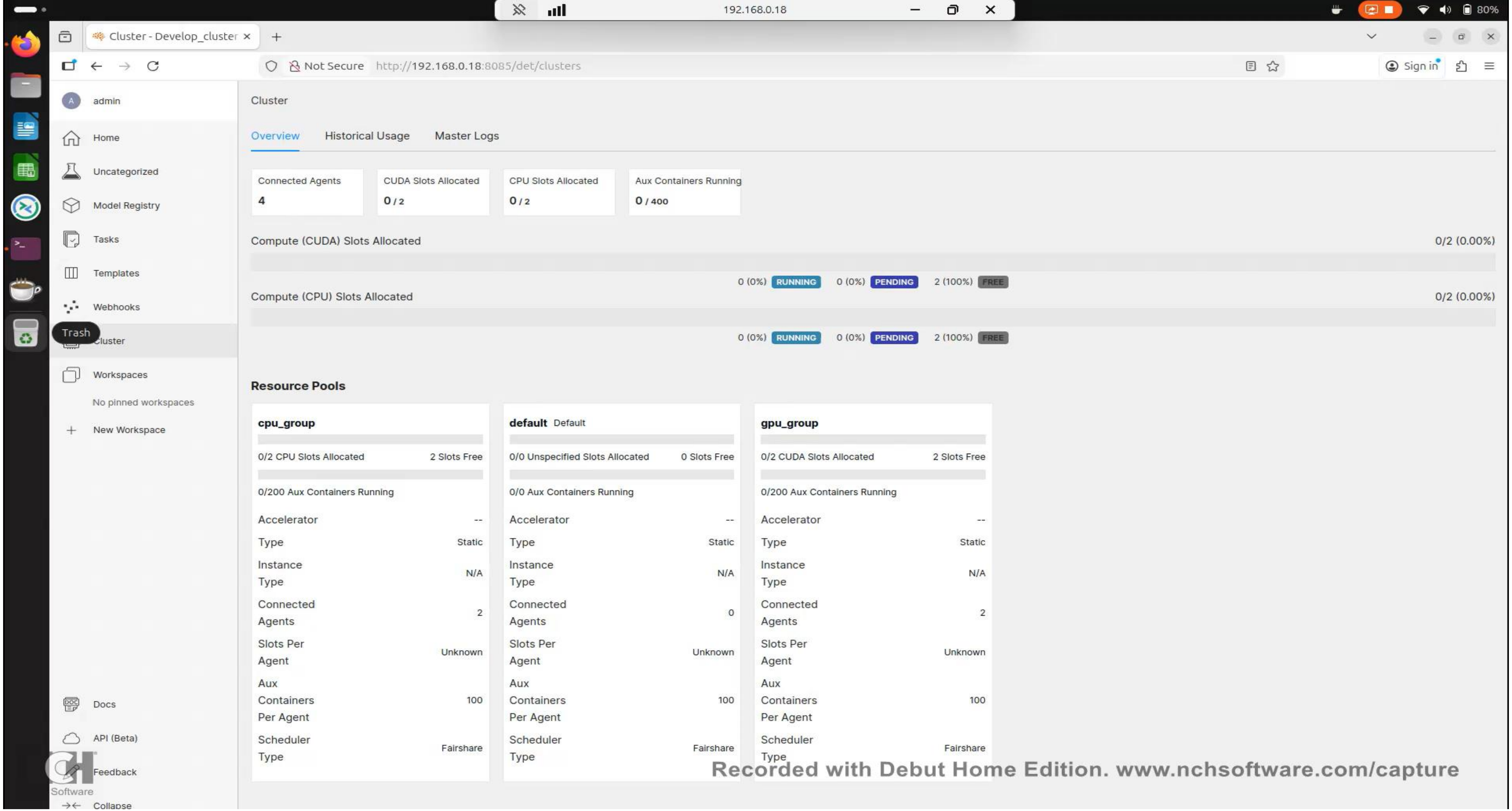
Inference and Mixed Enterprise Workloads

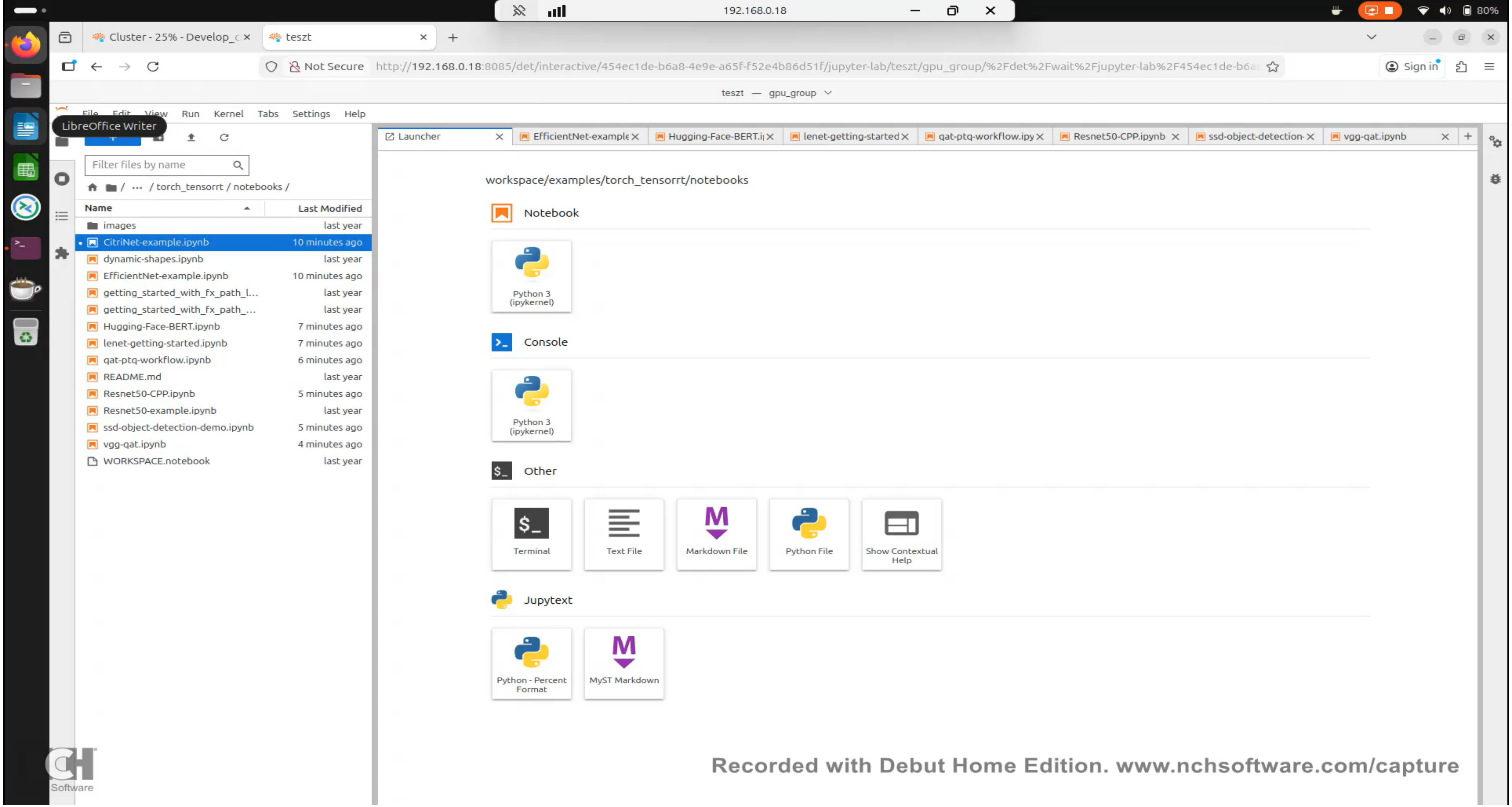
Training	Fine Tuning	Inference	RAG/Embedding	Data Analytics	Simulation	Visual Computing
Model Size: Medium-→ Small	Model Size: Medium-→ Small	AI Agents, SDG, VSS, CV, ReccSys	Multimodal Retrieval Pipelines	Spark RAPIDS, Data Science	CAE, CFD, Scientific Simulation	Digital Twins, CAD, Rendering, VDI

Trustworthy AI End-to-End Development Architecture



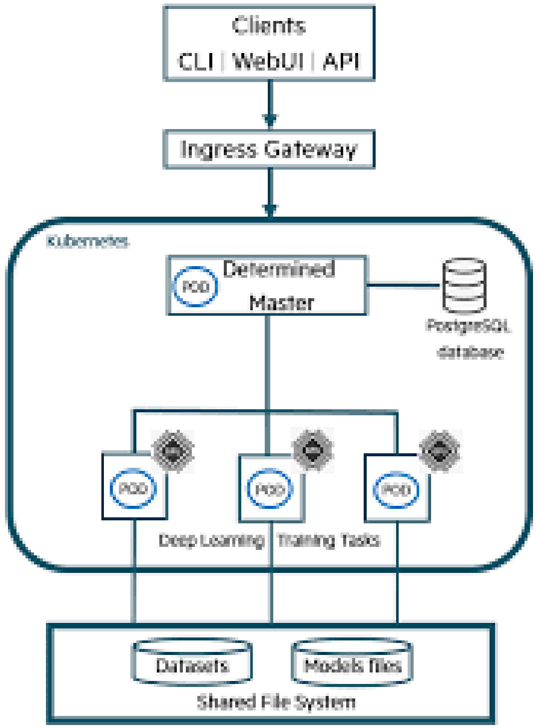

192.168.0.18
Home
80%
Recorded with Debut Home Edition. www.nchsoftware.com/capture

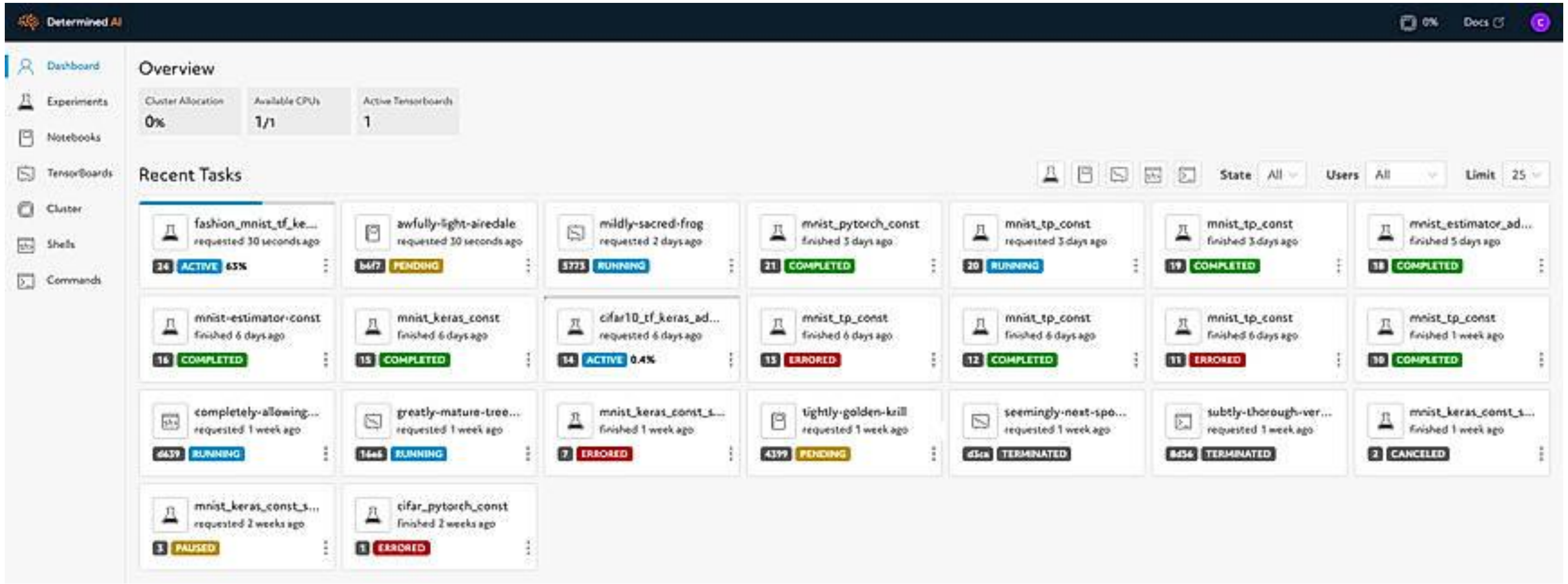

192.168.0.18
Cluster - Develop_cluster
Not Secure http://192.168.0.18:8085/det/clusters
Sign in
admin
Home
Uncategorized
Model Registry
Tasks
Templates
Webhooks
Trash
Cluster
Workspaces
No pinned workspaces
New Workspace
Docs
API (Beta)
Feedback
Collapse
Cluster
Overview
Historical Usage
Master Logs
Connected Agents 4
CUDA Slots Allocated 0 / 2
CPU Slots Allocated 0 / 2
Aux Containers Running 0 / 400
Compute (CUDA) Slots Allocated 0/2 (0.00%)
0 (0%) RUNNING 0 (0%) PENDING 2 (100%) FREE
Compute (CPU) Slots Allocated 0/2 (0.00%)
0 (0%) RUNNING 0 (0%) PENDING 2 (100%) FREE
Resource Pools
cpu_group
0/2 CPU Slots Allocated 2 Slots Free
0/200 Aux Containers Running
Accelerator --
Type Static
Instance Type N/A
Connected Agents 2
Slots Per Agent Unknown
Aux Containers Per Agent 100
Scheduler Type Fairshare
default Default
0/0 Unspecified Slots Allocated 0 Slots Free
0/0 Aux Containers Running
Accelerator --
Type Static
Instance Type N/A
Connected Agents 0
Slots Per Agent Unknown
Aux Containers Per Agent 100
Scheduler Type Fairshare
gpu_group
0/2 CUDA Slots Allocated 2 Slots Free
0/200 Aux Containers Running
Accelerator --
Type Static
Instance Type N/A
Connected Agents 2
Slots Per Agent Unknown
Aux Containers Per Agent 100
Scheduler Type Fairshare
Recorded with Debut Home Edition. www.nchsoftware.com/capture
Software


192.168.0.18
Cluster - 25% - Develop_c
teszt
Not Secure
http://192.168.0.18:8085/det/interactive/454ec1de-b6a8-4e9e-a65f-f52e4b86d51f/jupyter-lab/teszt/gpu_group/%2Fdet%2Fwait%2Fjupyter-lab%2F454ec1de-b6a
Sign in
teszt — gpu_group
File Edit View Run Kernel Tabs Settings Help
LibreOffice Writer
Launcher
EfficientNet-example
Hugging-Face-BERT.i
lenet-getting-started
qat-ptq-workflow.ipy
Resnet50-CPP.ipynb
ssd-object-detection-
vgg-qat.ipynb
Filter files by name
/ … / torch_tensorrt / notebooks /
Name
Last Modified
images last year
CitriNet-example.ipynb 10 minutes ago
dynamic-shapes.ipynb last year
EfficientNet-example.ipynb 10 minutes ago
getting_started_with_fx_path_l… last year
getting_started_with_fx_path_… last year
Hugging-Face-BERT.ipynb 7 minutes ago
lenet-getting-started.ipynb 7 minutes ago
qat-ptq-workflow.ipynb 6 minutes ago
README.md last year
Resnet50-CPP.ipynb 5 minutes ago
Resnet50-example.ipynb last year
ssd-object-detection-demo.ipynb 5 minutes ago
vgg-qat.ipynb 4 minutes ago
WORKSPACE.notebook last year
workspace/examples/torch_tensorrt/notebooks
Notebook
Python 3 (ipykernel)
Console
Python 3 (ipykernel)
Other
Terminal
Text File
Markdown File
Python File
Show Contextual Help
Jupytext
Python - Percent Format
MyST Markdown
Recorded with Debut Home Edition. www.nchsoftware.com/capture

HPE DETERMINED.AI

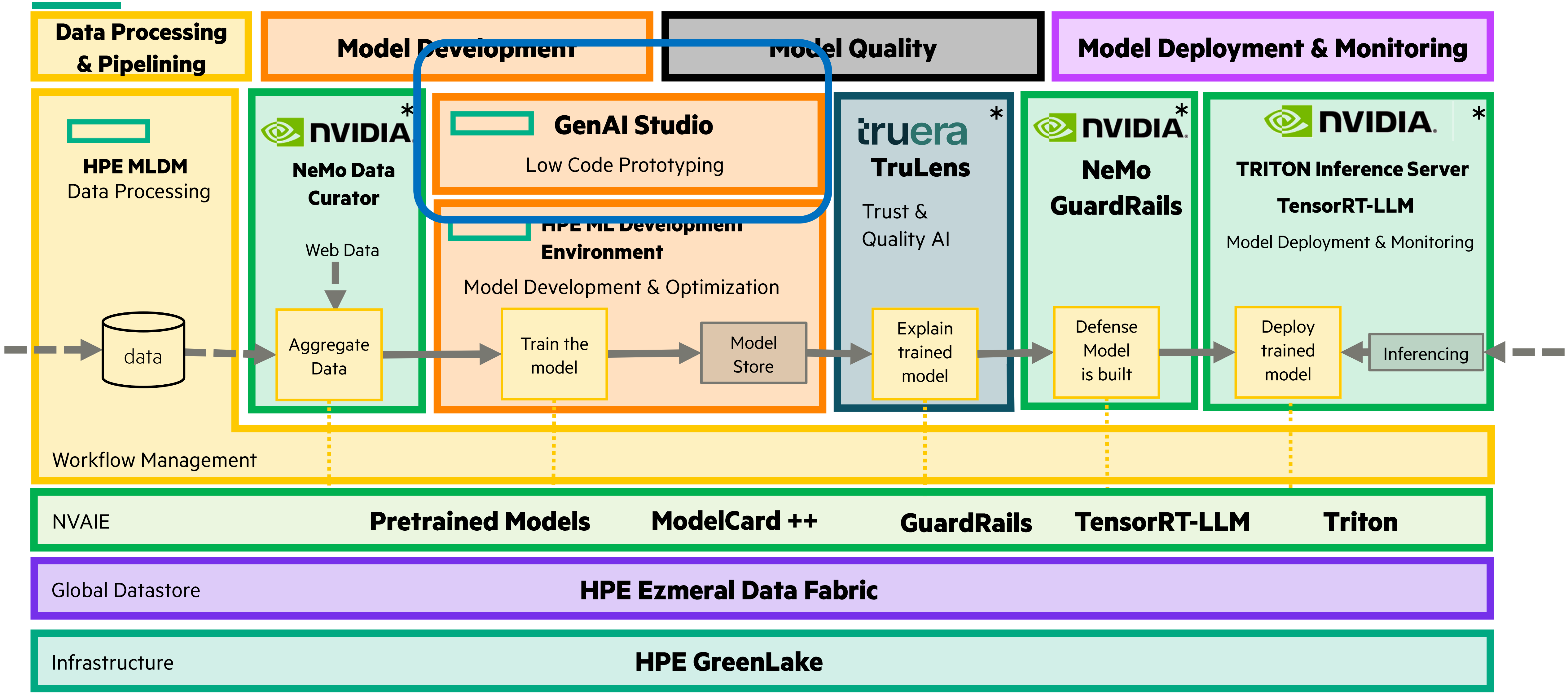
Determined AI + kubernetes



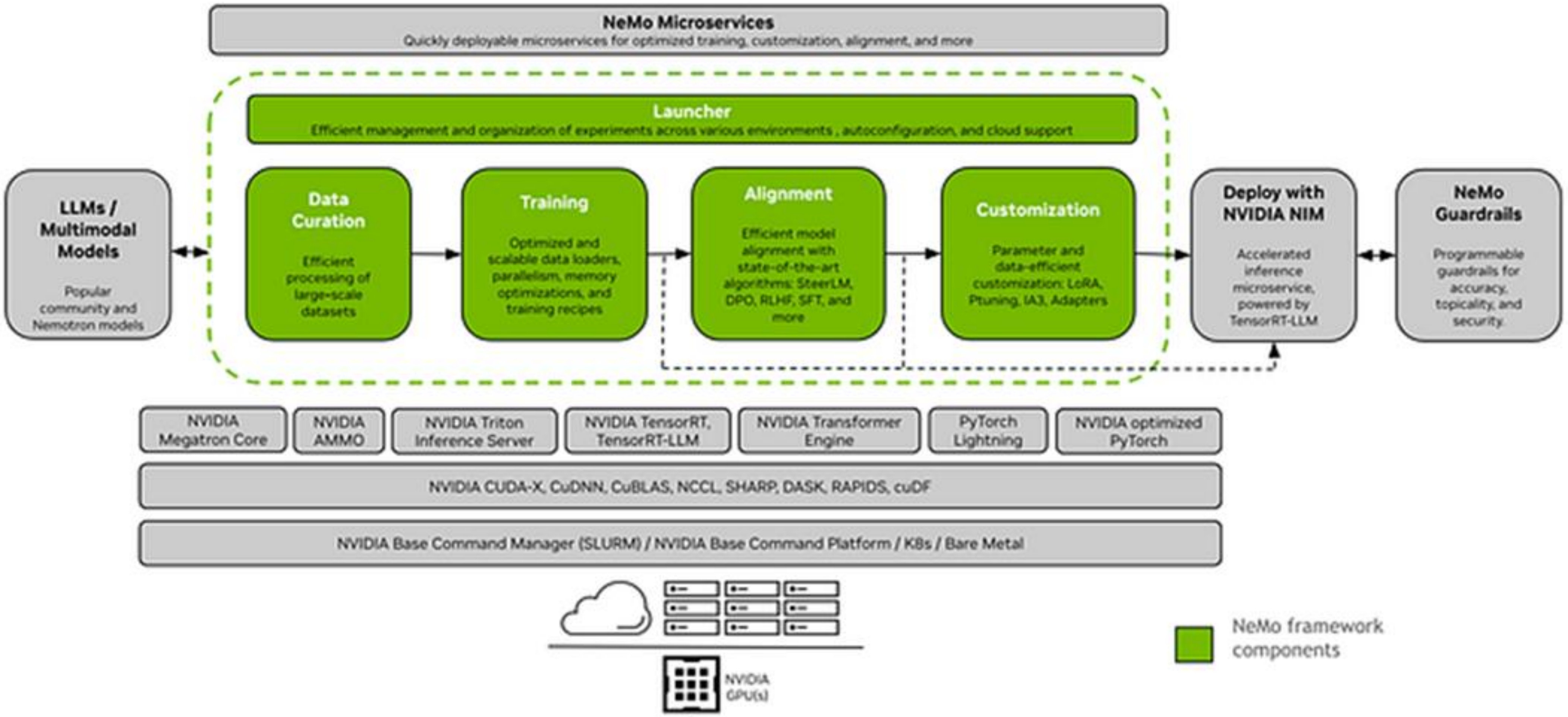


Low-Rank Adaptation (LoRA) és **Quantized LoRA (QLoRA)**

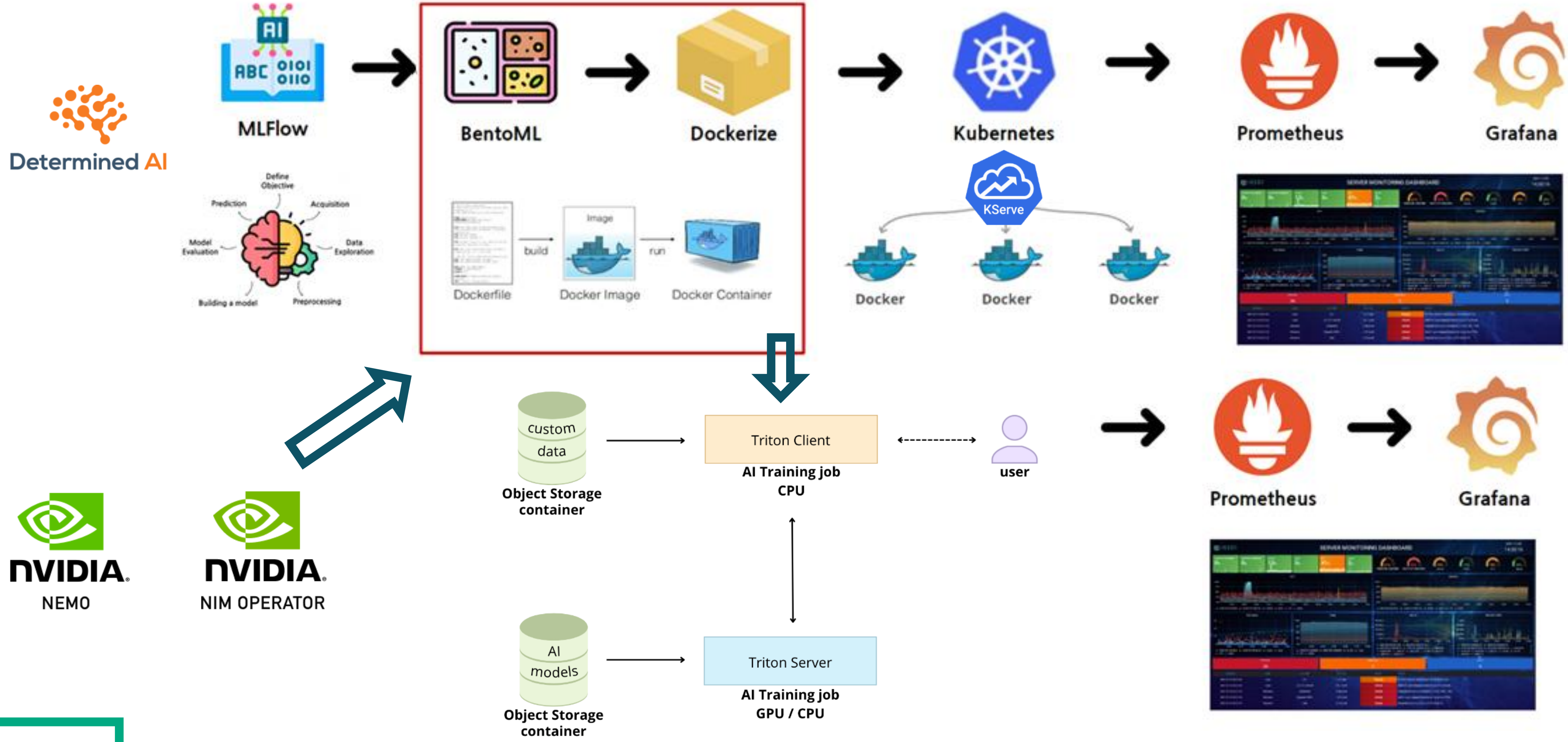
Trustworthy AI End-to-End Development Architecture – Alternative



NVIDIA NeMo framework – LLM and Multimodal



AI Inference feladatok



HPE Machine Learning Inference Software (MLIS)

A HPE MLIS kiegészülve az NVIDIA AI Enterprise szoftvert lehetőségeivel, egyszerűbb, hatékonyabb és biztonságosabb AI-telepítési folyamatot biztosít, racionalizálva a naeeow-AI / ML / LLM műveleteket.

Hatékonyság és méretezhetőség

- Széles modellválaszték, pl. NVIDIA NGC™, Hugging Face, ollama, egyedi modellek stb.
- Beépített optimalizálás az NVIDIA NIM-ekből
- Automatikus méretezhetőség a valós idejű forgalmi igényekhez. Bármely kompatibilis K8s környezetben

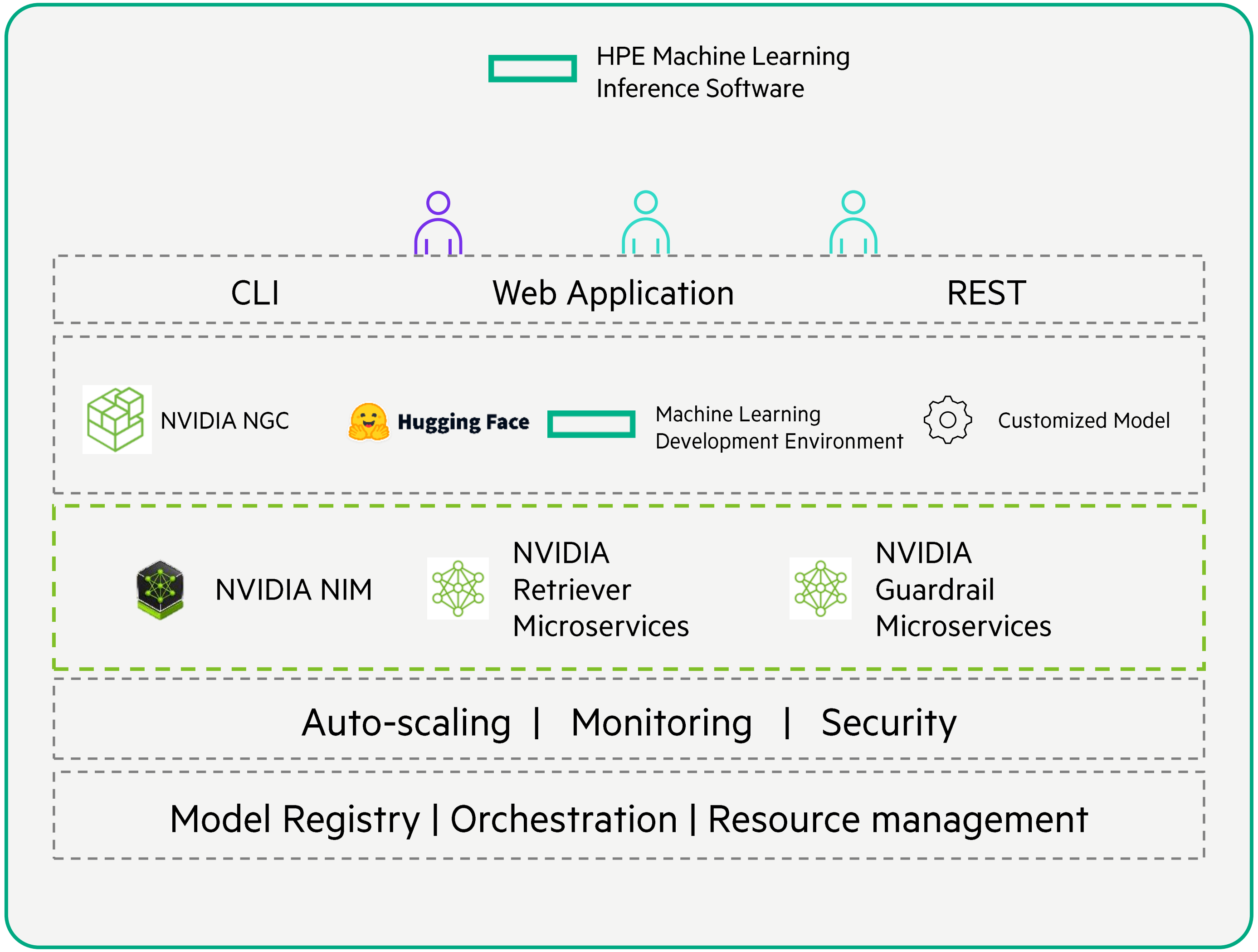
Az LLM működésének racionalizálása

- Valós idejű felügyelet és ellenőrzés
- Smoke és Canary bevezetések
- CI/CD automatizált munkafolyamat
- Modell megfigyelhetőség eltérések/kiugró értékek észleléséhez

Ellenőrzés és megfelelés

- Adatok és modellek helyben vagy saját felhőben
- Hitelesítés
- Többrétegű biztonsági megközelítés
- Védőkorlátok

HPE Machine Learning Inference Software (MLIS)



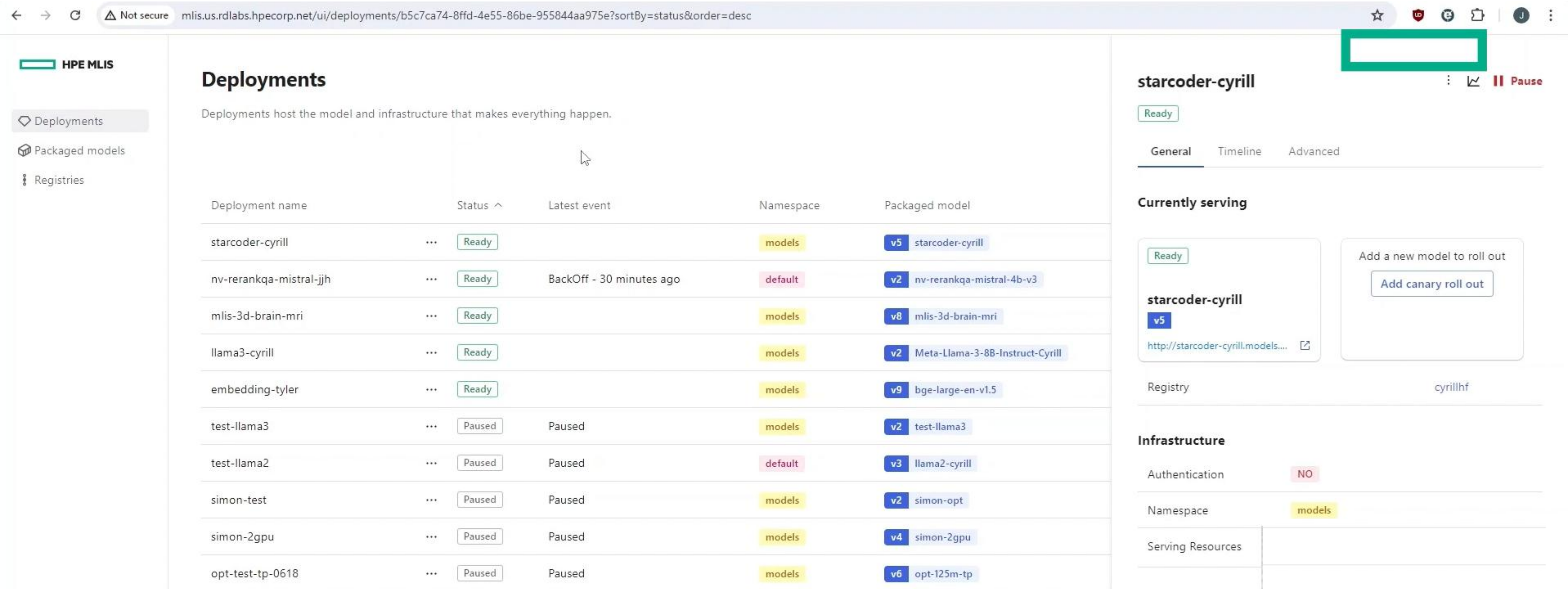
NVIDIA

- Minden modellhez és hardverkonfigurációhoz optimalizált inference motor
- API felület
- Biztonság (*konténeren belül is*)

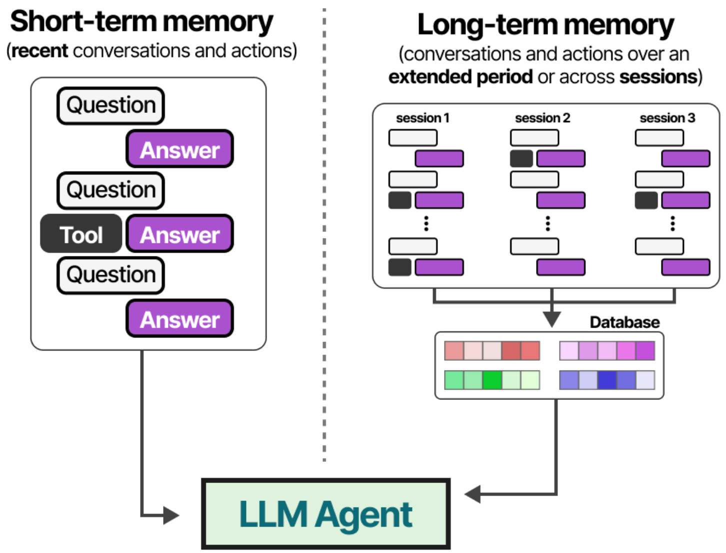
HPE Machine Learning Inference Software

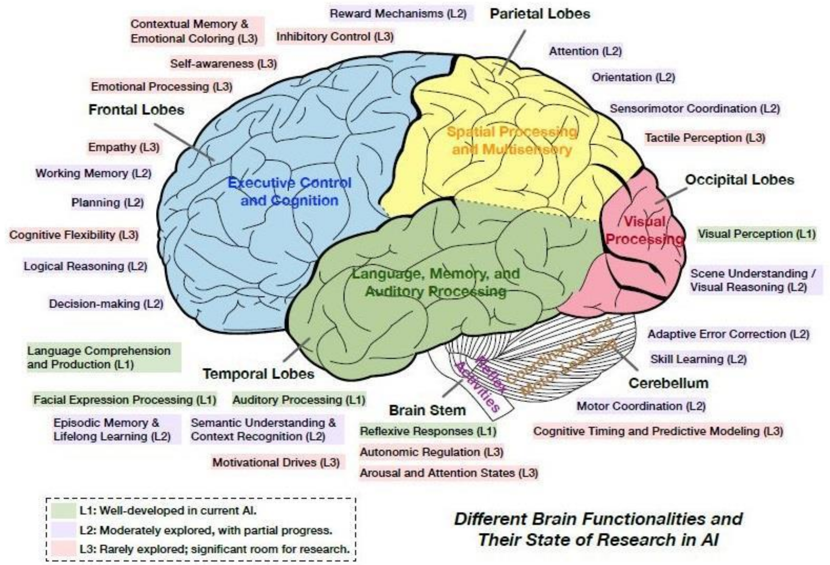
- Intuitív fejlesztői élmény GUI vagy CLI segítségével
- Egyszerű teleptési workflowk
- Valós idejű monitorozás és log-olás

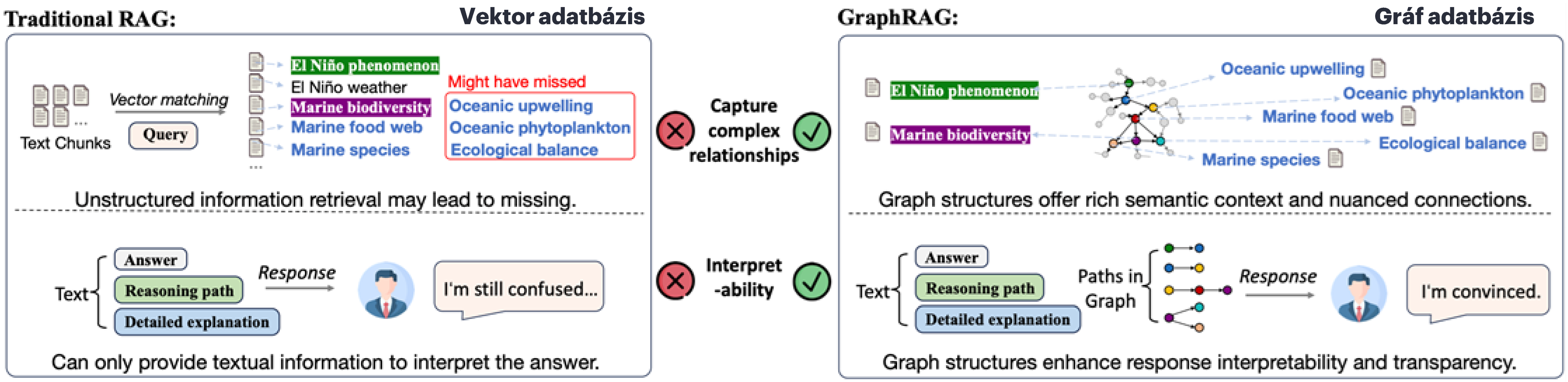
HPE Machine Learning Inference Software (MLIS)



AI Evolúció

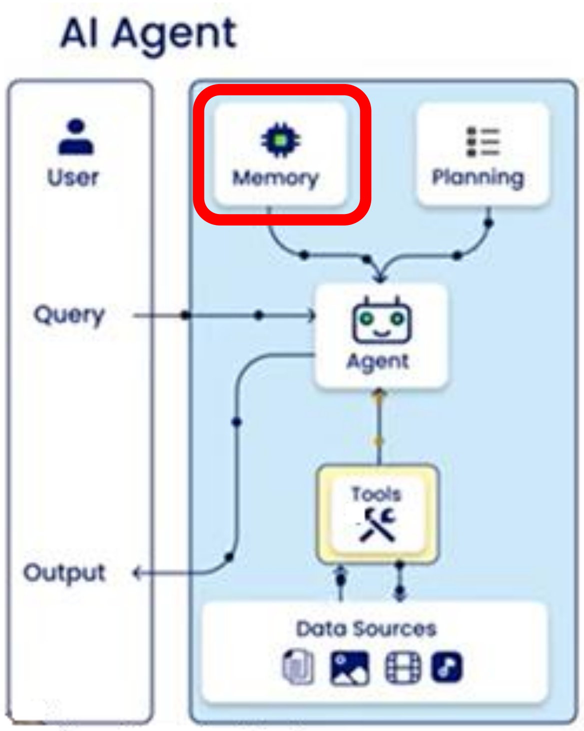




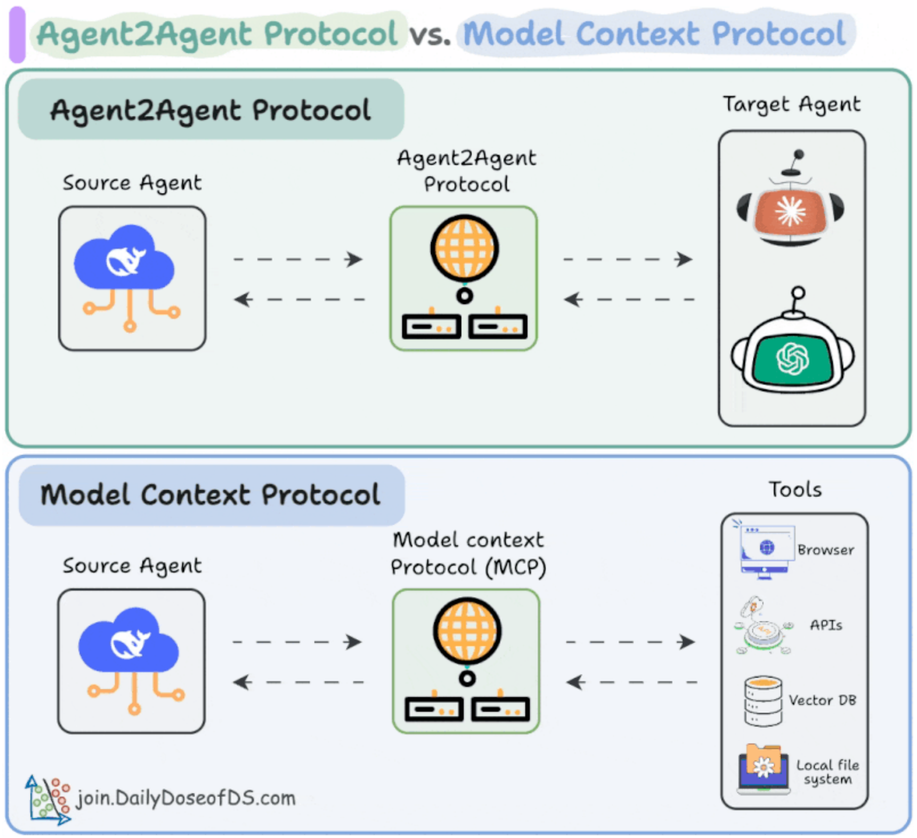


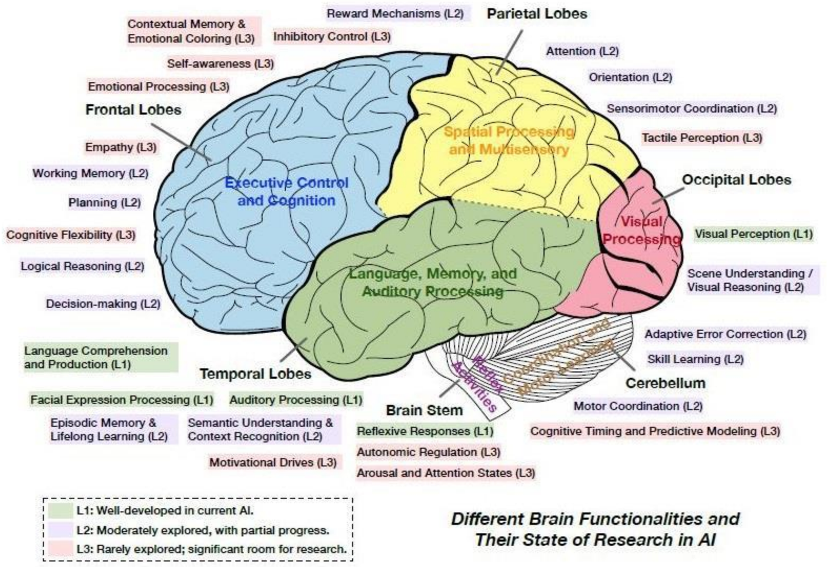
Source: The Hong Kong Polytechnic University, Hong Kong SAR, China, Jilin University, Changchun, China

AI Evolúció



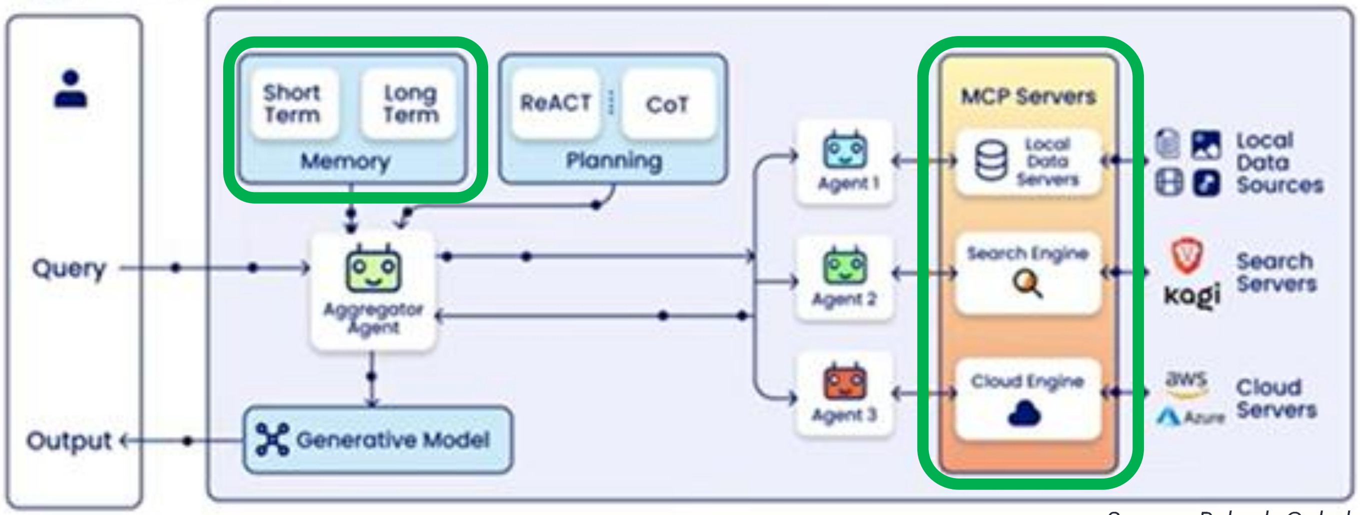
Source: Rakesh Gohel



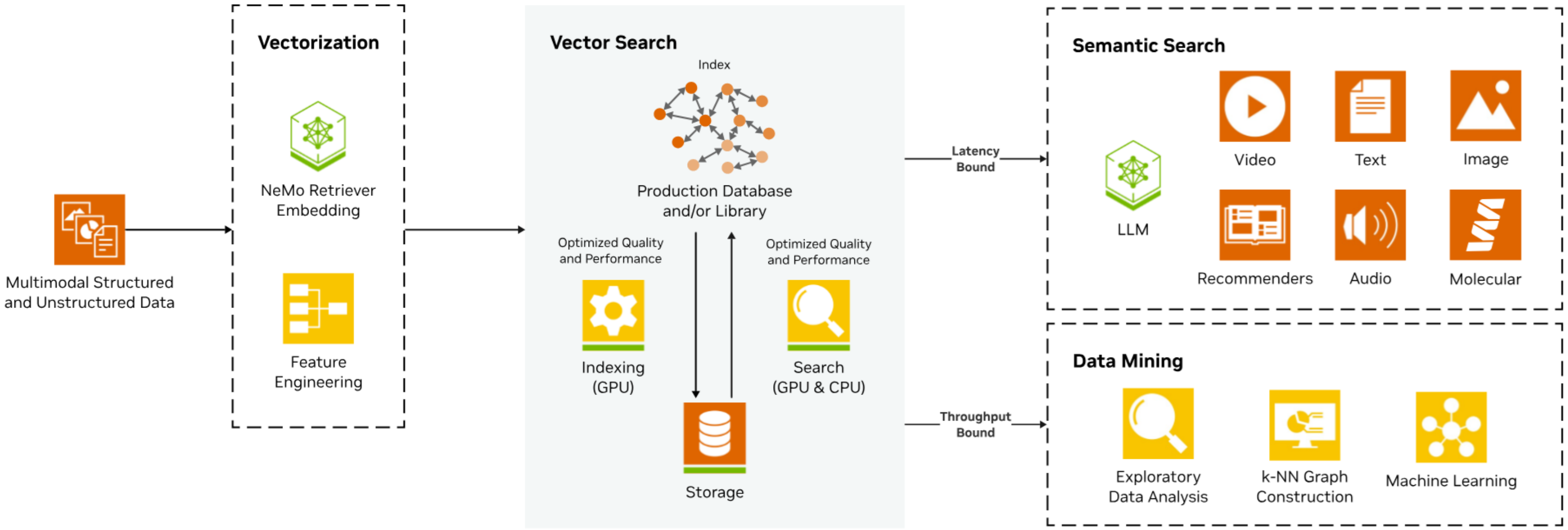


AI Evolúció

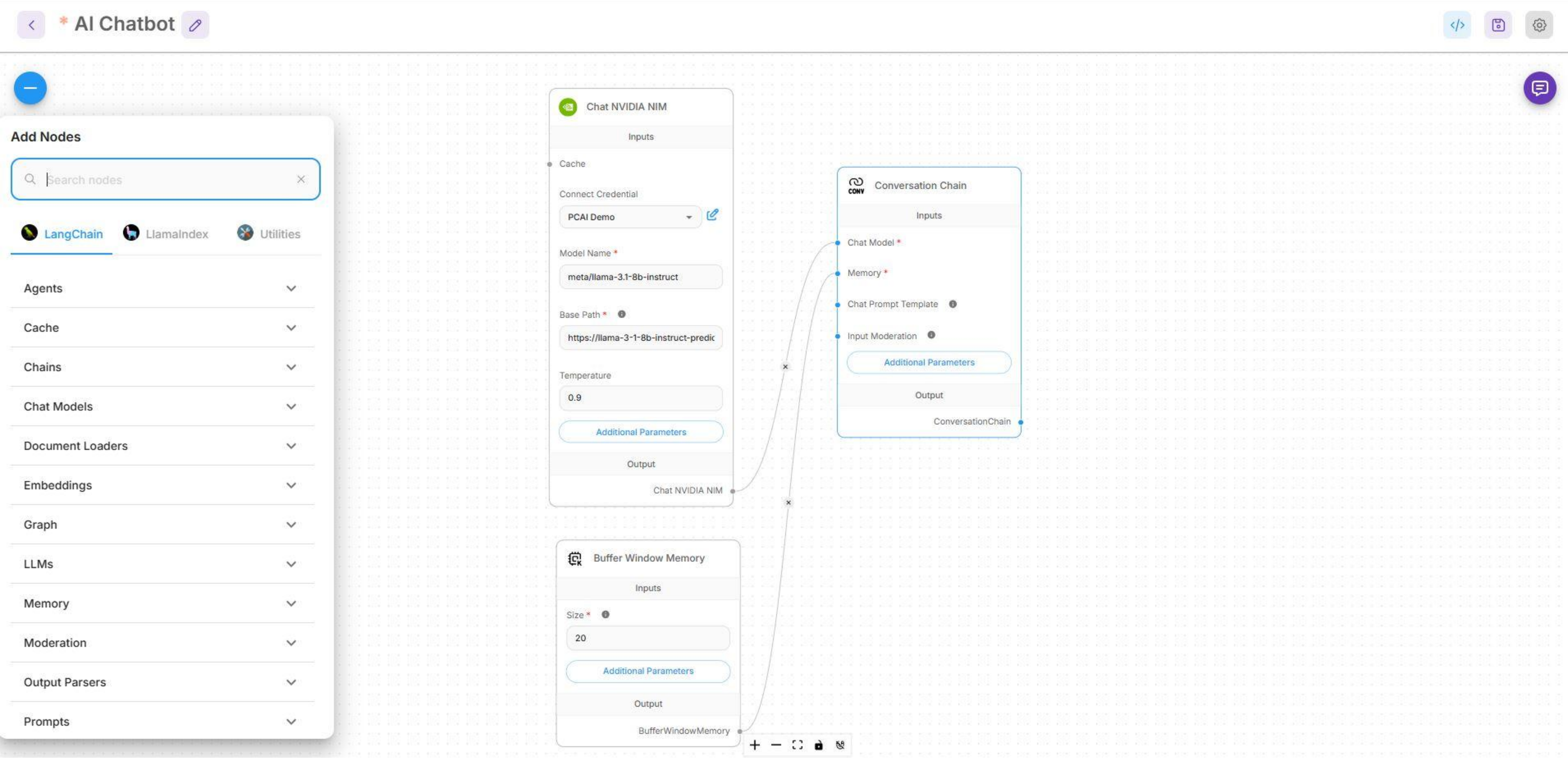
Agentic RAG



Source: Rakesh Gohel



PCAI - Flowise



HPE Private Cloud AI developer system

Accelerate AI Development

Instant AI development environment

- Ready to run out of the box
- Prove and Validate AI projects faster than ever
- Quickly deploy low-mid parameter models

Big AI power in a small package

- 2x Nvidia RTX Pro 6000 - 96GB GPUs
- 32 TB of integrated storage
- Includes fully integrated control node
- End to end AI software platform

Scalable and predictable

- Predictable and lower TCO than private cloud
- Unified experience across all HPE PCAI solutions
- Add production environments as you scale

Designed for developers and Data Engineers

- Pre trained models for rapid prototyping
- Easy data integration across public and private sources
- Deep insights into platform performance, operations, and resource management

Köszönöm!