Hewlett Packard Enterprise

HPE Private Cloud AI

Vlajková loď HPE pro on premise AI řešení

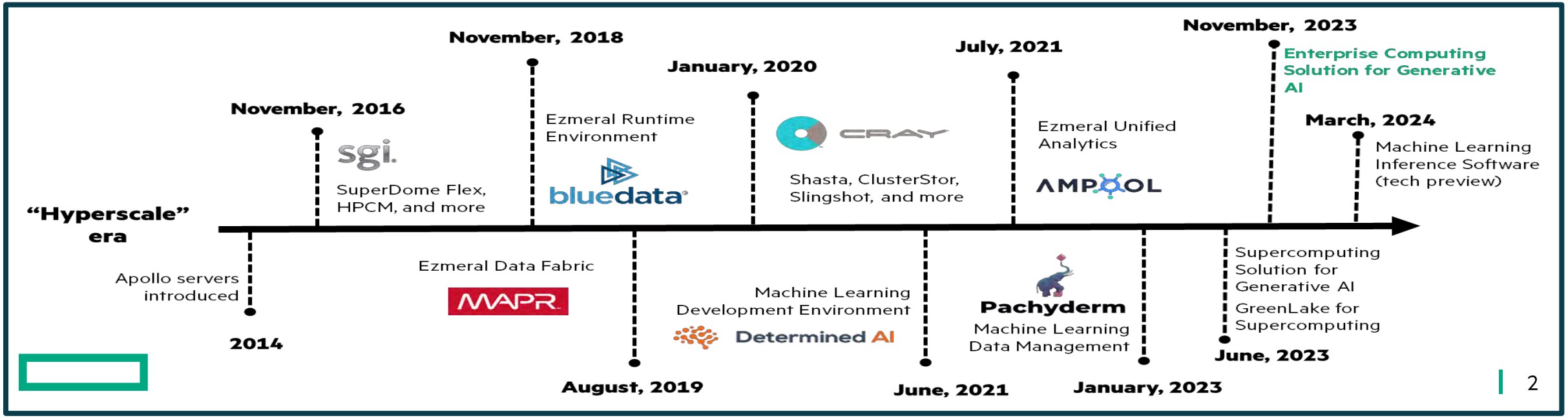
Ladislav Pecen

M Talks 17.4.2025

HPE je AI firma? Ano! Investice začaly řadu let před „AI boomem“

- HPE je leader v HPC řešeních, masivně využívaných v AI
- HPE dlouhodobě roste v oblasti technologií pro AI – ucelená řešení i jednotlivé HW či SW prvky
- Strategické akvizice i vlastní vývoj v oblasti AI

- Spolupráce s předními hráči v oblasti AI – např. NVIDIA
- Vlastní infrastrukturní řešení – compute i storage
- **HPE GreenLake** = „cloud-like zkušenost“



Proč HPE Private Cloud AI vznikl a co řeší?

HPE Private Cloud AI

Začít produkčně provozovat reálný AI usecase není snadné

Realita adopce Enteprise AI technologií

Pouze

33%

podniků má rozmyšlenu AI strategii

Pouze

20%

AI projektů jde do produkce

7+

měsíců od pilotu k produkci

Je nutné řešit řadu výzev

- Doba do zprovoznění
- Ochrana dat a řízení provozu
- Přístupnost dat
- Držet krok s vývojem v AI

Produkčního provozu se nedočká 9 z 10 AI projektů v pilotním provozu

Doba do uvedení do produkce

Trvá dlouho, než AI a IT ops týmy úspěšně vytvoří podmínky k přesunu AI pilotů do produkčního provozu, který bude realizovat předpokládané obchodní přínosy

Ochrana osobních údajů a dat

Zkonfigurovat a provozovat prostředí bezpečně a v souladu s předpisy a nařízeními není jednoduché

Rychlost AI inovací/ škálovatelnost

Modely a nástroje umělé inteligence se vyvíjejí rychleji než týmy, které je mohou využívat.

Reputační rizika a poškození značky

Obavy o bezpečnost a ochranu dat

Nedostatek specializovaných znalostí

Problémy s infrastrukturou, PoC vs Produkce

Postavit fungující a provozně efektivní platformu pro AI řešení není snadné

Je nutno uvážit množství komponent

Infrastruktura

Bare metal, VM's, kontejnery

Compute, storage, network

+

SW platforma

Data Management

Vývoj, trénování

Deployment, inference

+

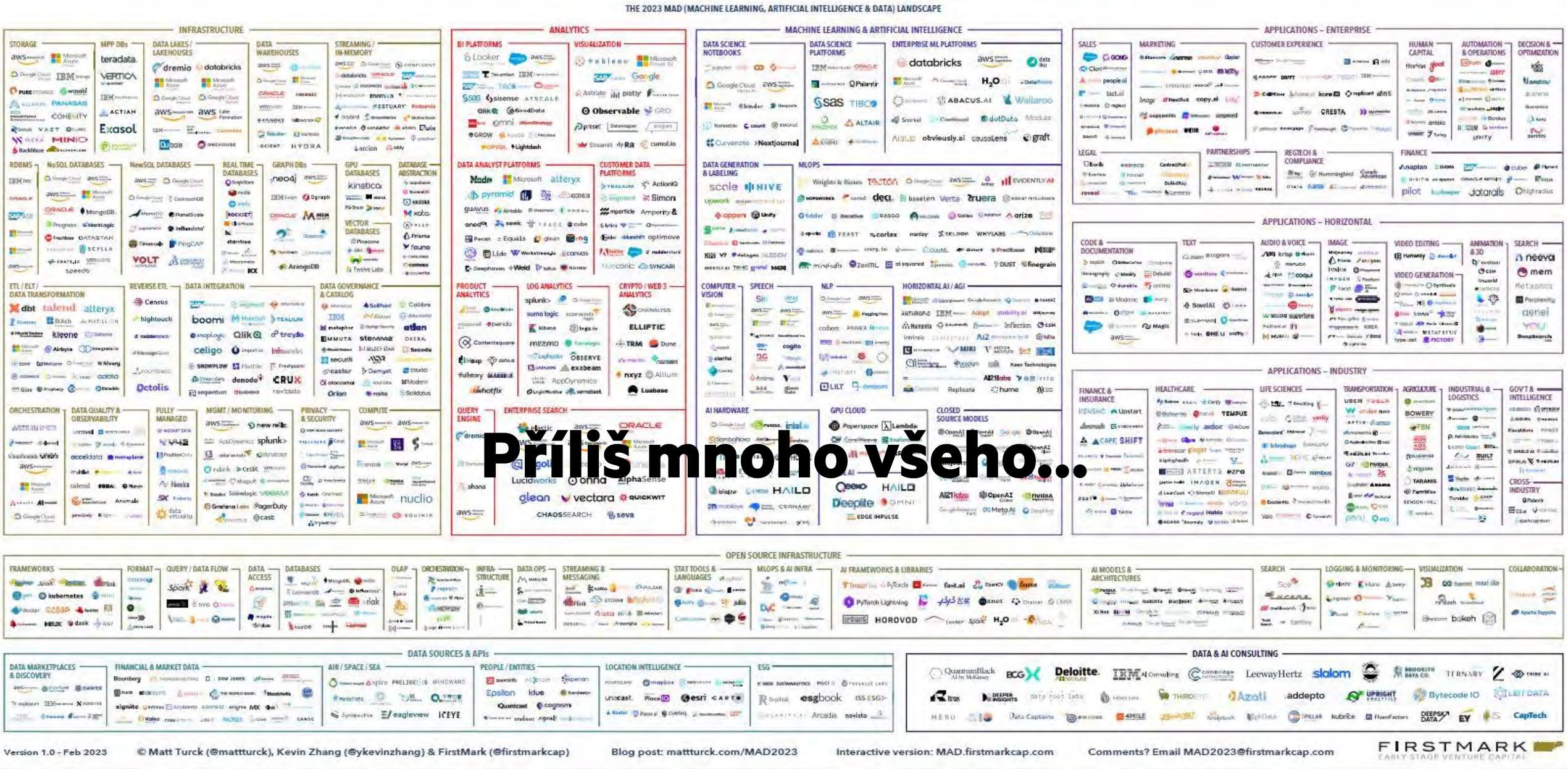
Modely

Open-source modely

Komerční platformy

Oblasti se vzájemně ovlivňují

Hewlett Packard Enterprise | NVIDIA



Hewlett Packard Enterprise | NVIDIA

DYI platforma pro AI? Jistě, ale s HPE to jde provést i lépe!

Technologický výstup

Klasický přístup „DYI“

- Bare metal, VM, kontejnery
- Compute, storage, network, služby
- Softwarové komponenty, typicky opensource. Community support

Vše hotovo – vyřešeno „na klíč!“

- Přístup hotová „appliance“
- Full-stack platorma připravená k provozu
- Management, observabilita, jednotná podpora

Když je potřeba ještě víc...

- Vystavěno na míru, HPC technologie
- Multitenance pro poskytovatele služeb
- Referenční architektury, škálovatelnost

Hewlett Packard Enterprise | NVIDIA

HPE Private Cloud AI

full-stack platforma privátního AI cloudu pro (gen)AI aplikace

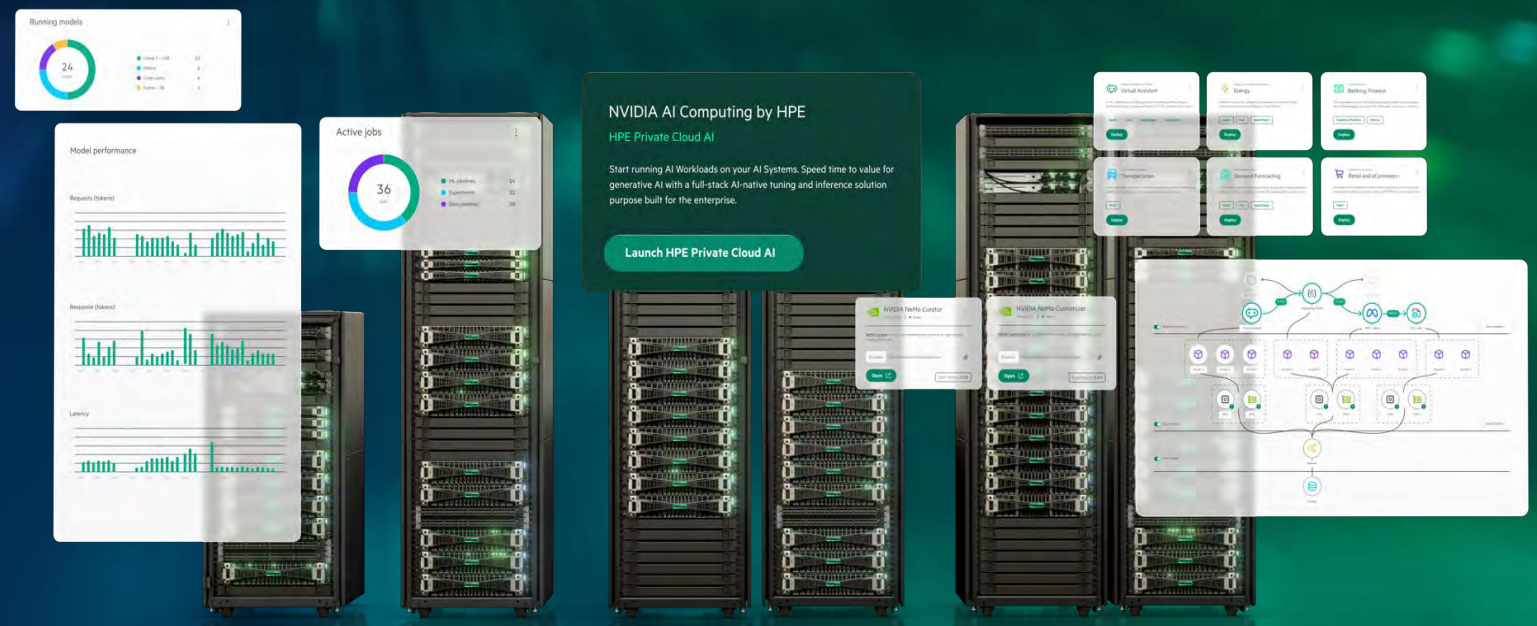
- AI modely
- AI software
- AI infrastruktura
- AI služby

Okamžitá produktivita

Bezpečný a jednotný přístup k datům

Plná kontrola nad daty, modely, aplikacemi

Cloudová zkušenost na onprem platformě



HPE Private Cloud AI v číslech

Deployment

HPE nainstaluje ve vašem DC nebo v kolokaci **<8 hodin**

Připraveno k nasazení **3 kliky**

Deployment AI workloadu se solution akcelerátorem **1 klik**

Zprovozněno za **jeden den**

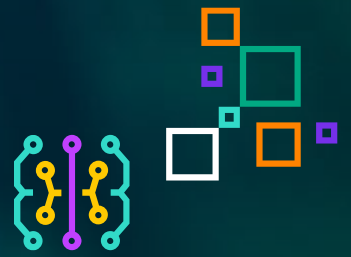
Produktivita

Až **5x** lepší výkon modelu s NVIDIA NIM

2x Zvýšení produktivity AI vývojářů

Až **63%** nižší TCO než v public cloudu

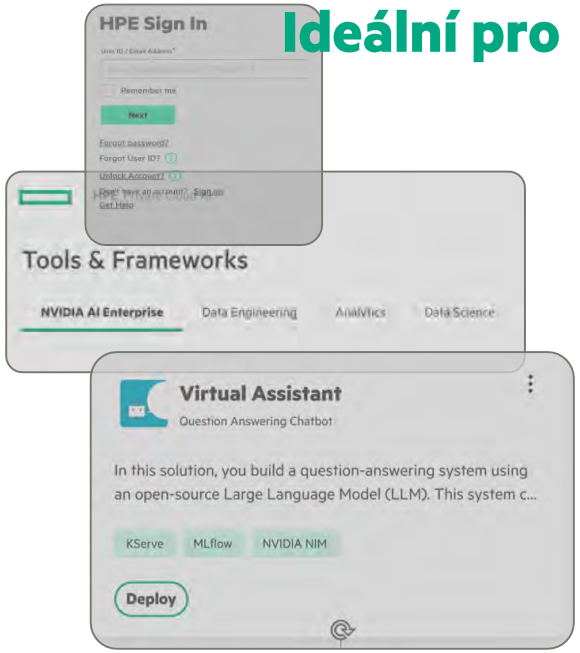
4x rychlejší vývoj



AI produktivita v **minutách** ne měsících

Source: HPE internal reports. Comparison between using GPT-4 via OpenAI API vs. self-hosted Llama3, assuming an enterprise account with 5,000 users, 5 chat sessions per day, 8,000 tokens per chat

HPE Private Cloud AI – vyladěná platforma s compute, storage, network, SW



Ideální pro

NOVINKA!

	AI Sandbox	Inferencing	Inferencing + RAG	Inferencing + RAG + Fine-tuning
	 Developer System	 Small	 Medium	 Large
Compute	2 NVIDIA H100 NVL GPU's	4 or 8 NVIDIA L40S GPUs	8 or 16 NVIDIA L40S GPUs	16 or 32 NVIDIA H100 NVL GPUs
Storage	32 TB Integrated	109 TB	217 TB	670 TB
Networking	Customer Network	100GbE NVIDIA Networking	200GbE NVIDIA Networking	400GbE NVIDIA Networking
Power	Up to 2.2 kW	up to 8 kW rack	up to 17.7 kW	up to 16.5 kW x 2

Jednotná cloudová zkušenost prostřednictvím HPE GreenLake cloud

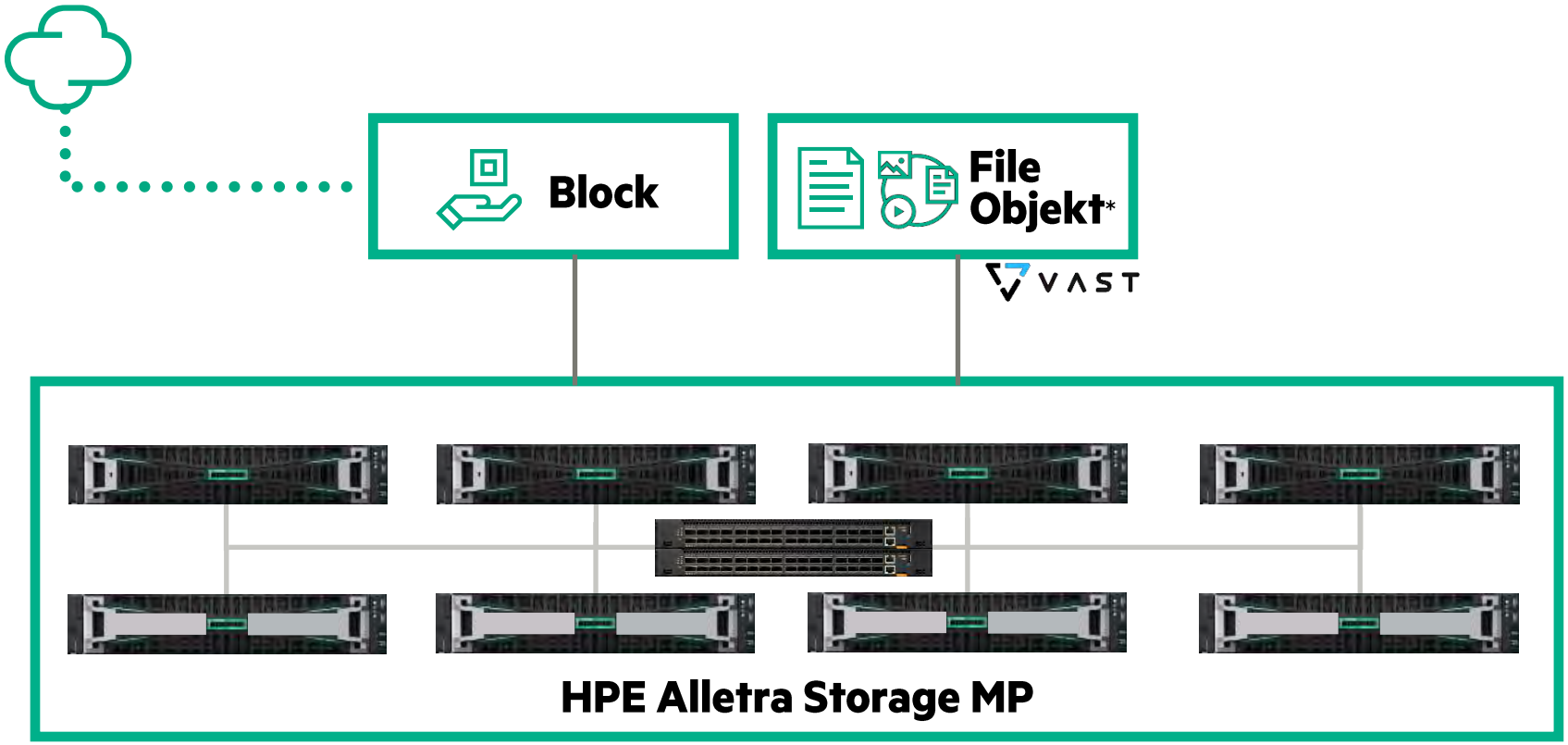
HPE PCAI obsahuje továrně předkonfigurované všechny nezbytné vrstvy

Management experience	HPE Private Cloud AI *via* HPE GreenLake cloud
Control Nodes	Virtualization • HPE VM Essentials Data Services Connector (DSC) VM HPE AI Essentials
AI Worker Nodes	Red Hat Enterprise Linux HPE AI Essentials with NVAIE
Storage	HPE GreenLake for File Storage (+Object) NFS, S3 storage for Control Nodes and AI Worker Nodes

HPE GreenLake for File jako výkonný storage základ Private Cloud AI

HPE GreenLake platform
- control plane, ovládání

Jedna HW platforma,
různé typy úložišť –
blokové nebo souborové
nebo objektové



* Funkcionalita typu objekt umožněna pouze pro vybrané workloady

NVIDIA AI Computing by HPE

HPE Private Cloud AI

AI Modely	Community + Partner + Custom NVIDIA NIM	
AI Software	NVIDIA AI Enterprise	HPE AI Essentials
AI Infrastruktura	NVIDIA GPU/GPU NVIDIA Spectrum-X	HPE AI Servers HPE File Storage
AI Služby	HPE Provided	3rd Party Partners

HPE Private Cloud Control Plane
Setup, inventory, IT and Asset Lifecycle Management
Workload provisioning, orchestration and management
VM runtimes and Kubernetes cluster management
Monitoring, observability and E2E security for apps and infrastructure

Developer edition | **Small** | **Medium** | **Large**

Konfigurace podle předpokládaných typů AI workloadů

HPE GreenLake cloud

Hewlett Packard Enterprise | NVIDIA

HPE Private Cloud AI přináší správný mix vlastností, který umožní přivést (gen)AI scénář k produkčnímu životu

Bezpečné využití firemních dat bez nutnosti odeslání pryč

Rychlý vývoj s frameworky, modely a workflow

Jednoduché škálování od pilotu do produkce

Trasování, měření, monitoring a **správa**

Hewlett Packard Enterprise | NVIDIA

A jaké vlastně AI usecasy jde na platformě provozovat?

HPE Private Cloud AI

Od ML až po Agentic AI!

Code generation

"Co-pilot" for SW engineers to generate code faster–on prem with own codebase and coding standards

Document creation

Automate form-filling—insurance, prescriptions and more; create marketing briefs, parts manuals and more

Search engine

Enhance search results with internal documents and information

Q & A chat

Create an internal support chatbot with your own data for an enhanced knowledge base and faster resolutions

Customer service

Boost productivity for customer service teams and enhance customer experiences with optimized call centers

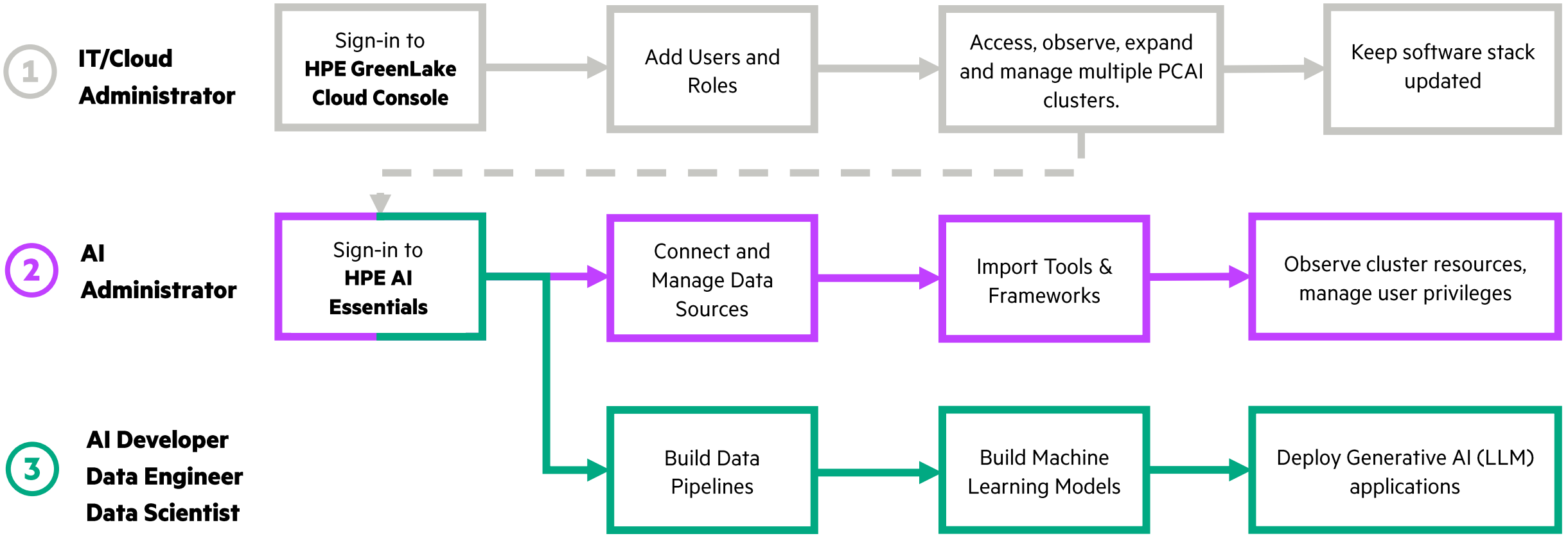
Hewlett Packard Enterprise | NVIDIA

Pojďme trochu hlouběji!

HPE Private Cloud AI

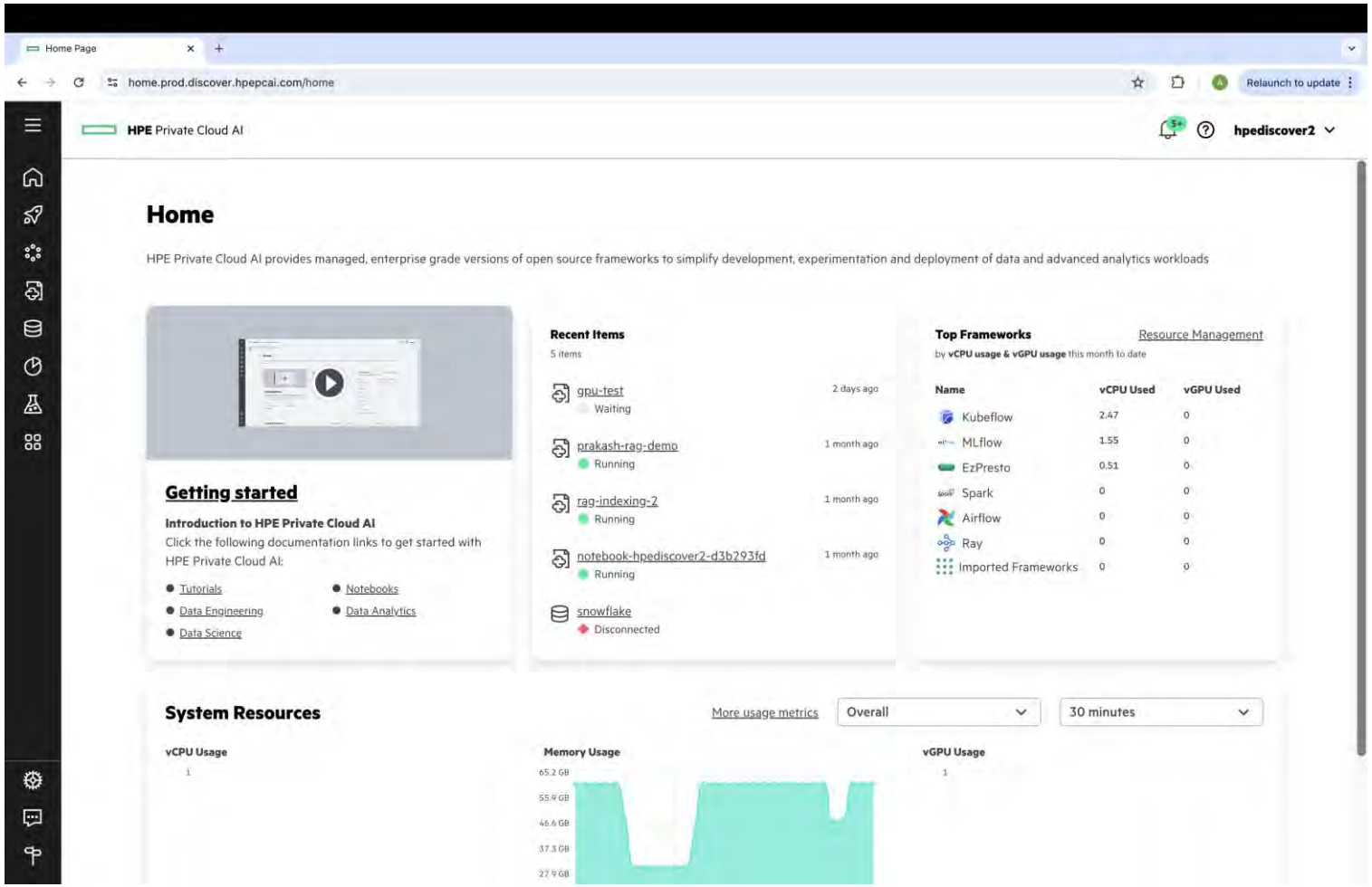
HPE PCAI – jedno zařízení pro různé uživatelské role.

Na konci je uživatel hotové AI aplikace



HPE PCAI – jedno uživatelské rozhraní

Na konci je uživatel hotové AI aplikace



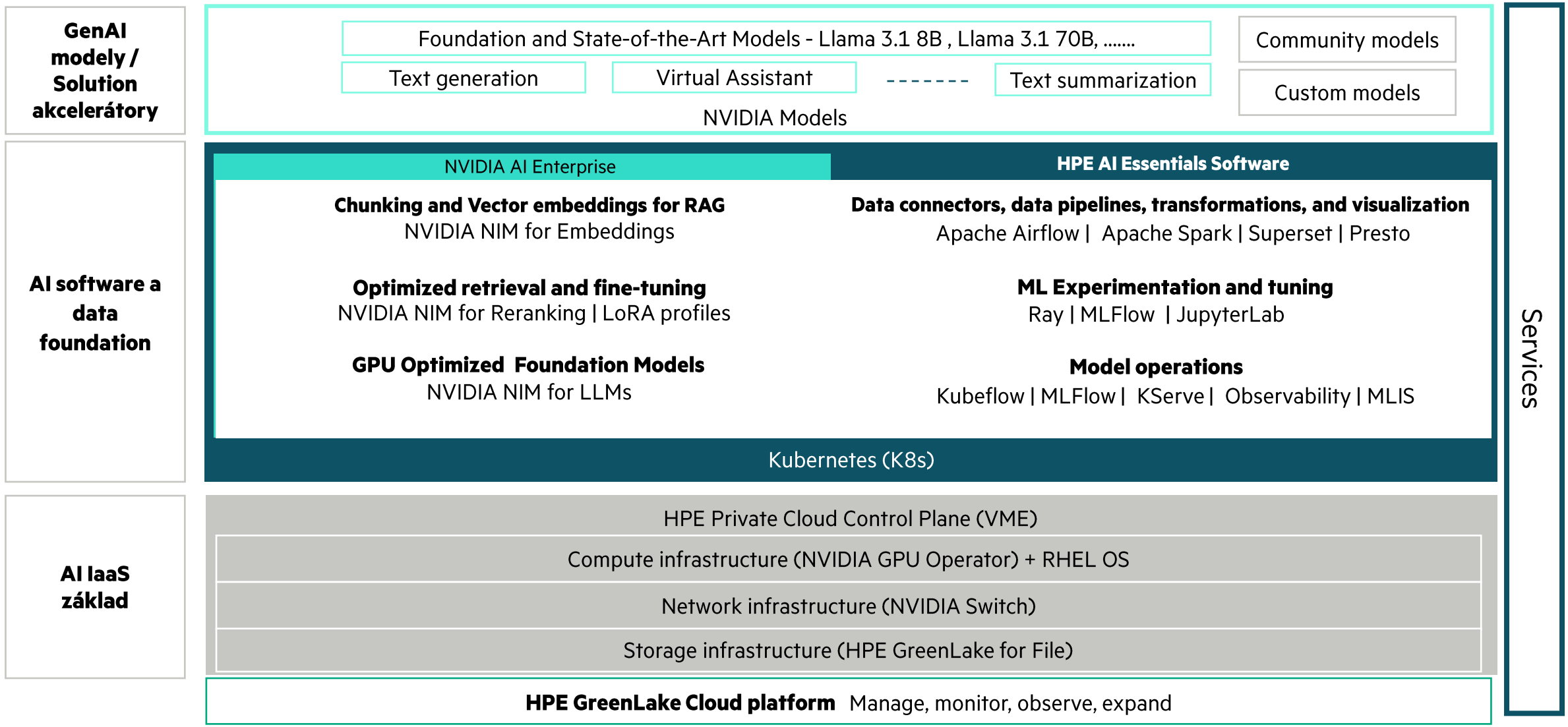
Jedno uživatelské rozhraní pro různé role:

IT/Cloud administrátor se stará o nezbytnou AI Infrastrukturu.
Rozložení, stohování, správa zdrojů, dohled infrastruktury pro umělou inteligenci na několik kliknutí.

AI administrátor řeší potřebný AI software.
Spravuje datové zdroje, nástroje, frameworky a GPU resources, sleduje využití jednotlivými uživateli a skupinami. Vše z jednoho GUI.

AI vývojář vyvíjí AI inovace. Neopravuje K8s.
Vytváření datových kanálů, DAGů, transformací, modelů a AI aplikací, bez starosti o nástroje nebo infrastrukturu.

Společné GUI schovává K8s motor i užitečné aplikace



HPE PCAI však přináší mnohem víc, než je vidět v GUI

Plně integrované platformní služby, security funkce, nebo třeba integrované NVIDIA NIM modely

HPE AI Essentials Services

ML Lifecycle

- **Experimentation and tuning**
  - Ray, MLFlow, JupyterLab
- **Model operations**
  - Kubeflow, MLFlow, Kserve

Data Analytics

- **Data Connectors, pipelines**
  - Apache Airflow, Apache Spark
- **Transformations and visualizations**
  - Superset, Presto

Platform

- **Telemetry, Metrics, and Event Management**
  - OTEL, Prometheus
- **Container service mesh**
  - Istio
- **GPU/network operations**
- **Kubernetes policy management**
  - Kyverno

Security

- **User and Certificate management**
  - cert-manager
- **SSO**
  - token/role management, Federated ID broker
- **Workload identity management**
  - Spire

Ukázaná platí. Pojďme se zalogovat!

HPE Private Cloud AI

HPE Private Cloud AI – dnes a zítra

Řešte AI, ne fungování platformy

Připraveno k okamžitému provozu

Fully managed, předintegrované s HW a SW technologiemi NVIDIA aHPE

Škálování do budoucna

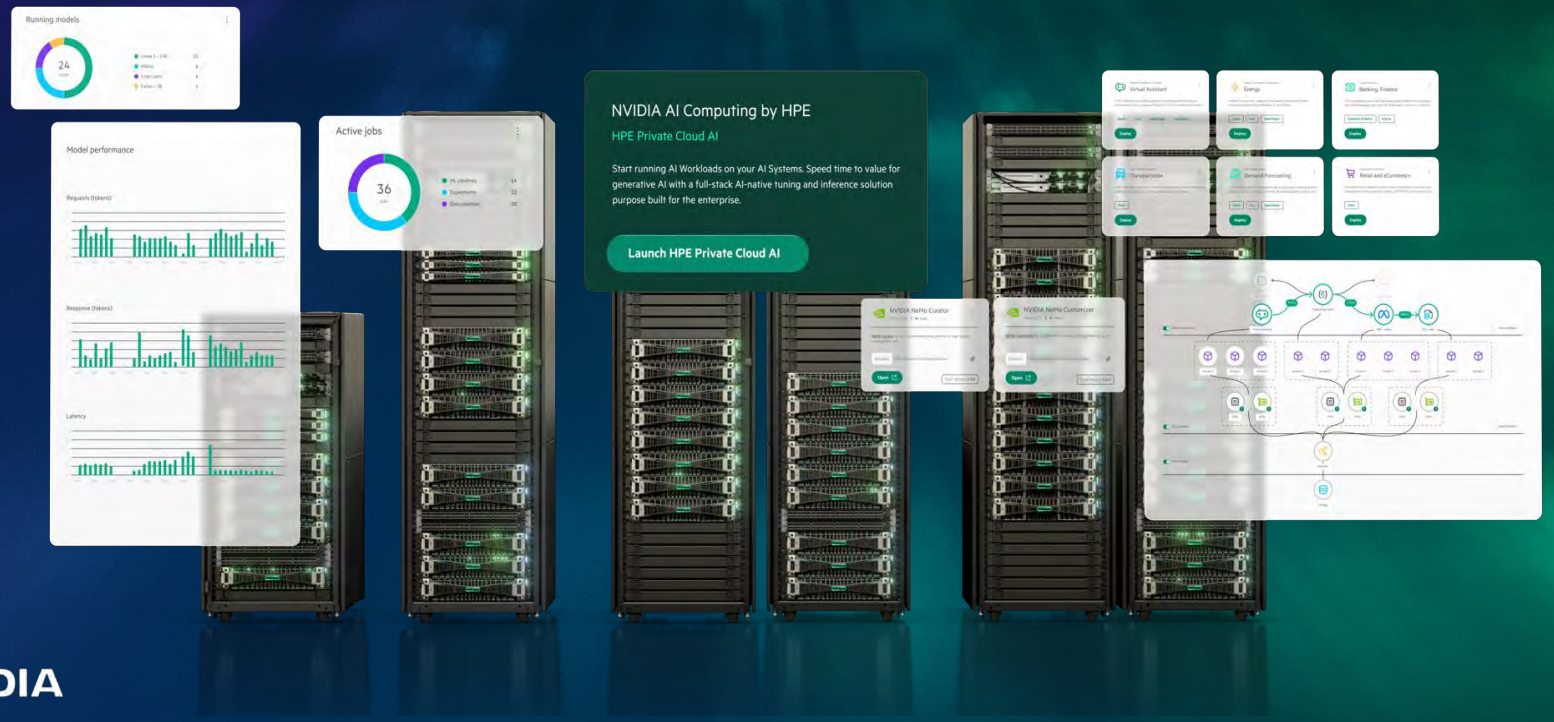
Modulární konigurace nasizované pro workloady typu inference, RAG-based aplikace, model customization, případně fine tuning

Data i modely v bezpečí

Privátní data a modely pod kontrolou, jednotná technická podpora na celé řešení, nástroje pro rychlé inovace AI aplikací, souladu s moderními principy AI vysvětlitelnosti, etiky a bezpečnosti

Životní cyklus AI

Rychlé AI inovace s přístupem k nejnovějším AI modelům, vývojovému softwaru, blueprintům a frameworkům



Hewlett Packard Enterprise | NVIDIA