



**Hewlett Packard
Enterprise**

HPE Private Cloud AI

Part of the NVIDIA AI Computing by HPE Portfolio

Speaker name and title
Month DD, 20YY



Agentic AI Brings Automation and Independence

Agentic and Physical AI demand reliable platforms



Generative



Retrieval Augmented (RAG)



Agentic and Physical

Create

- Increased productivity with heavy human oversight and verification.
- Inspire, guide, and help.
- Accelerate employee or customer tasks.

Cite and Trust

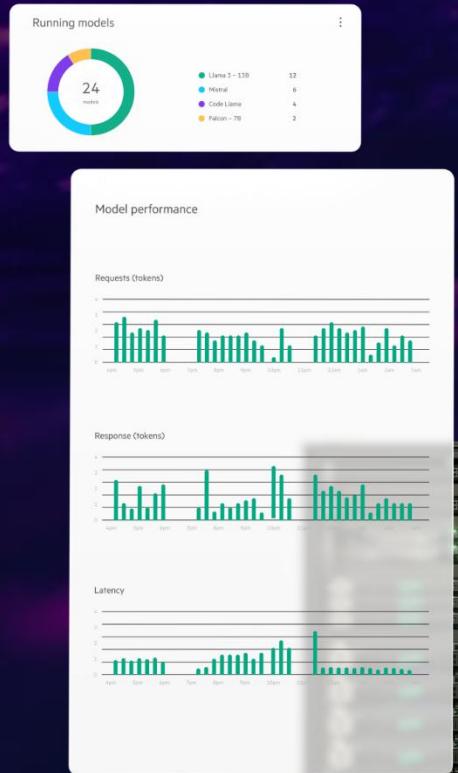
- Verifiable and more trustworthy output with citation.
- Write, build, validate.
- Improved output accuracy and expanded enterprise use cases.

Enterprise Productivity

- Automation and Integration of physical systems.
- Act, Power, Decide,
- Less human intervention, the ability to power physical objects, and increased decision making abilities demand a secure and reliable AI platform.

NVIDIA AI Computing by HPE

HPE Private Cloud AI



NVIDIA AI Computing by HPE
HPE Private Cloud AI

Start running AI Workloads on your AI Systems. Speed time to value for generative AI with a full-stack AI-native tuning and inference solution purpose built for the enterprise.

[Launch HPE Private Cloud AI](#)

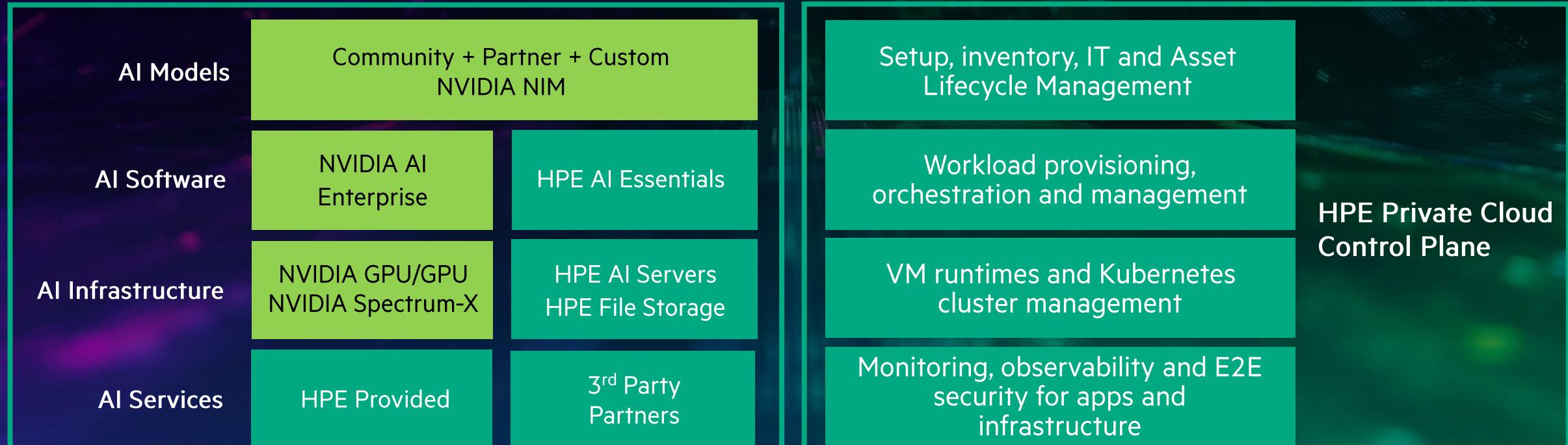


 Hewlett Packard
Enterprise

 NVIDIA

NVIDIA AI Computing by HPE

HPE Private Cloud AI



Developer System

Small

Medium

Large

X Large

Configurations aligned to enterprise AI workloads

HPE GreenLake cloud

AI optimized and scalable hardware

Start Fast



Developer System



Small



Medium



Large

Scale Fast



Extra Large
Coming 2025

2 NVIDIA H100 NVL GPU's

4 or 8 NVIDIA L40S GPUs

32 TB Integrated

109 TB

Customer Network

100GbE NVIDIA Networking

Up to 2.2 kW

up to 8 kW rack

8 or 16 NVIDIA L40S GPUs

217 TB

200GbE NVIDIA Networking

up to 17.7 kW

16 or 32 NVIDIA H100 NVL GPUs

670 TB

400GbE NVIDIA Networking

up to 16.5 kW x 2

Future Blackwell Support

TBD

800GbE NVIDIA Networking

TBD

AI optimized and scalable hardware

Small-AI inference

- 1 x HPE ProLiant DL380a Gen12 AI Optimized Node (4 x NVIDIA L40s GPUs Total)
- 3 x HPE ProLiant DL325 Gen12 Control Nodes (DataLake HPE Ezmeral)
- 109 TB HPE GreenLake for File Storage with Object enabled HPE Alletra MP
- NVIDIA SN4600cM Switches (100GbE Networking)
- 42U Rack with PDUs
- HPE AI Essentials with NVIDIA AI Enterprise Software
- 3 or 5-year subscription

Medium-AI inference and Retrieval Augmented Generation (RAG)

- 2 x HPE ProLiant DL380a Gen12 AI Optimized Node (8 x NVIDIA L40s GPUs Total)
- 3 x HPE ProLiant DL325 Gen12 Control Nodes DataLake HPE Ezmeral
- 217 TB HPE GreenLake for File Storage with Object enabled HPE Alletra MP
- NVIDIA SN4700M Switches (200GbE Networking)
- 42U Rack with PDUs
- HPE AI Essentials with NVIDIA AI Enterprise Software
- 3 or 5-year subscription

Large-AI inference, Retrieval Augmented Generation (RAG), and Model Fine Tuning

- 4 x HPE ProLiant DL380a Gen12 AI Optimized Node (16 x NVIDIA H100NVL GPUs Total)
- 3 x HPE ProLiant DL325 Gen12 Control Nodes (DataLake HPE Ezmeral)
- 670 TB HPE GreenLake for File Storage with Object enabled Alletra MP
- NVIDIA SN4700M Switches (400GbE Networking)
- 2 x 42U Racks with PDUs
- HPE AI Essentials with NVIDIA AI Enterprise Software
- 3 or 5-year subscription

Developer System

Small

Medium

Large

Extra Large Coming 2025

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	Future Blackwell Support
32 TB Integrated	109 TB	217 TB	670 TB	TBD
Customer Network	100GbE NVIDIA Networking	200GbE NVIDIA Networking	400GbE NVIDIA Networking	800GbE NVIDIA Networking
Up to 2.2 kW	up to 8 kW rack	up to 17.7 kW	up to 16.5 kW x 2	TBD

End to End AI Software Platform

Full Software Framework

Data Analytics Services

- Data Connectors
- Data Pipelines
- Data Transformations and visualization

ML Lifecycle Services

- Model Experimentation
- Model Tuning
- Model operations

Platform services

- Telemetry Metrics, and Event management
- Container service mesh
- GPU/network operations
- Kubernetes policy management

Security services

- User and Certificate management
- SSO (token/role management, Federated ID broker)
- Workload identity management

Data Engineering Tools

HPE Ops Ramp (GPU observability and Optimization)

Analytics & Visualization Tools

NVIDIA AI Enterprise

Data Science Tools

HPE Data Fabric (Data access and security)

Kubernetes (K8s)

HPE VM Essentials (Virtualization)

HPE GreenLake (Hybrid Cloud Management)

- Modular architecture with evergreen or single click upgrades
- Integrated lifecycle management with evergreen NVIDIA experience and a rich ecosystem of open-source tools and models
- End-to-end zero-trust security from GreenLake platform through all elements of PCAI Designed for enterprise grade compliance

- Enterprise-grade data lineage, verifiability, and auditability for AI pipelines
- Lo code/no code RAG deployment
- Solution accelerators for common enterprise use cases
- Rich set of models from NVIDIA, partners, and ecosystem

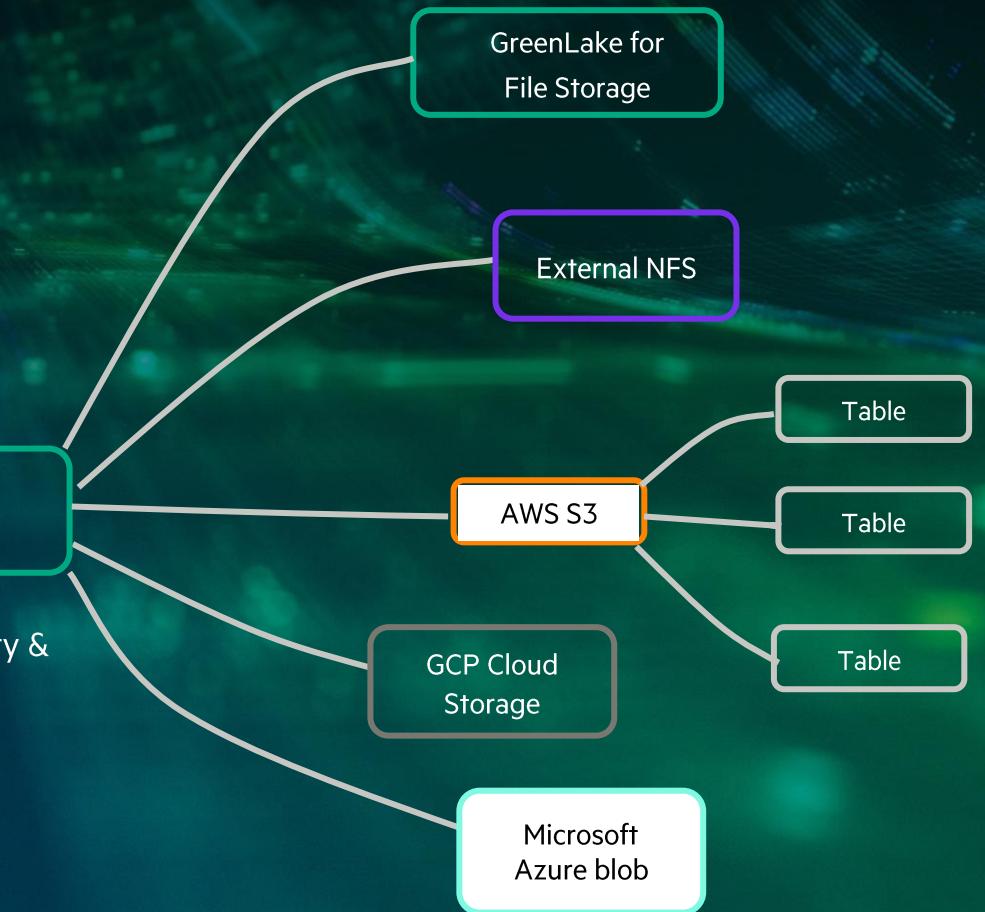
HPE Data Fabric

Unified data access and centralized security



HPE Data Lakehouse

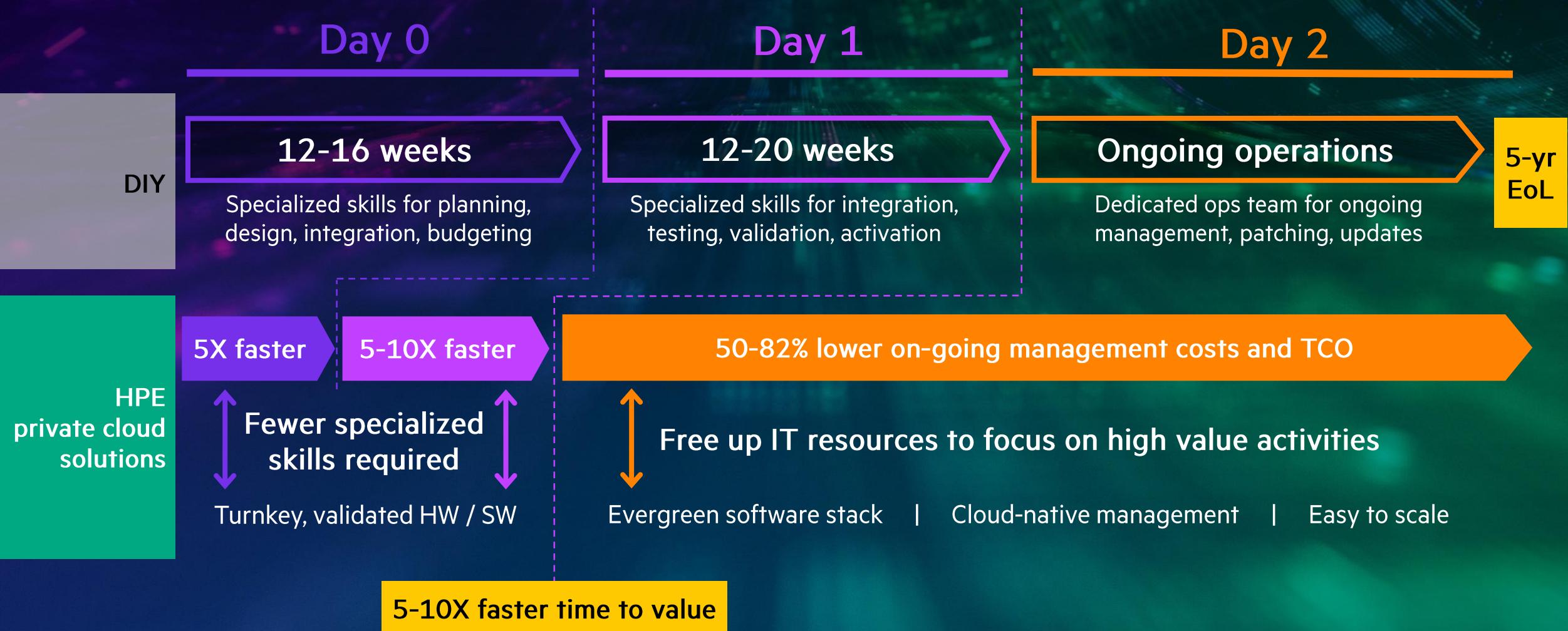
Centralized security & governance



Transparent, secure access for
AI and Analytics data

Consistent data access and
management across all data
assets and formats

Engineered systems drive speed and efficiency



What to remember about HPE Private Cloud AI

A full-stack, turnkey private cloud for production AI

Ready to run out of the box

Fully managed, pre-integrated NVIDIA accelerated computing, networking, and software with HPE compute, storage, and software

Scale as you grow

Available in modular configurations sized for inference, RAG-based applications and model customization and fine tuning

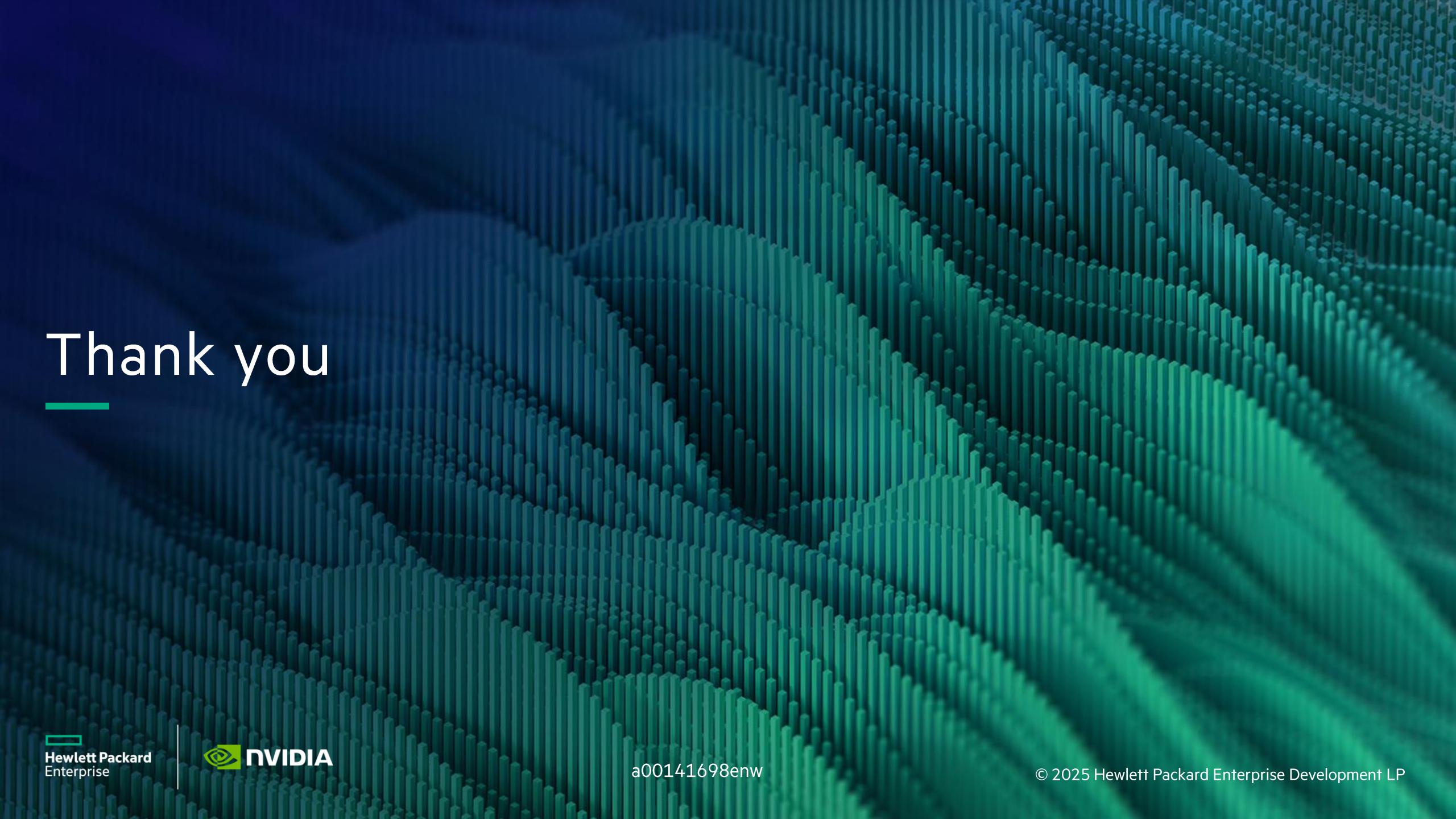
Data privacy and control

Eliminate data silos with one global namespace for seamless access to different data types, anywhere

End to End AI Software Platform

Keep up with AI innovation with access to the latest AI models, development software or NVIDIA Blueprints





Thank you

Mapa de Capacidades IA en HPE PCAI

