AI INDUSTRY IMPACT

The future of AI innovation starts with NetApp and NVIDIA

NetApp | NVIDIA

Agenda

June 2025
Post-GTC'25

- NVIDIA & NetApp alliance: Better together
- Why you struggle with AI projects
- Doubling down for the future
- Joint solutions: Compute, software, and storage infrastructure
- Q&A and discussion

Building the world's best AI solutions

Partners since 2018

>600 joint customers

Worldwide partner ecosystem

Joint solution development:

- Certified reference architectures for DGX SuperPOD
- NetApp® AIPod™ and FlexPod™ reference architectures
- NCP certification
- NVAIE in hyperscalers with NetApp
- NeMo/NIMS in ONTAP (soon)
- And more….

NetApp | NVIDIA

"Half of the world's files run on NetApp, and now you can securely talk to them with GenAI with NeMo."

—Jensen Huang, Keynote at GTC, March 2024

The “Agentic AI” Opportunity

Reasoning, adapting, acting independently

Perceptive AI

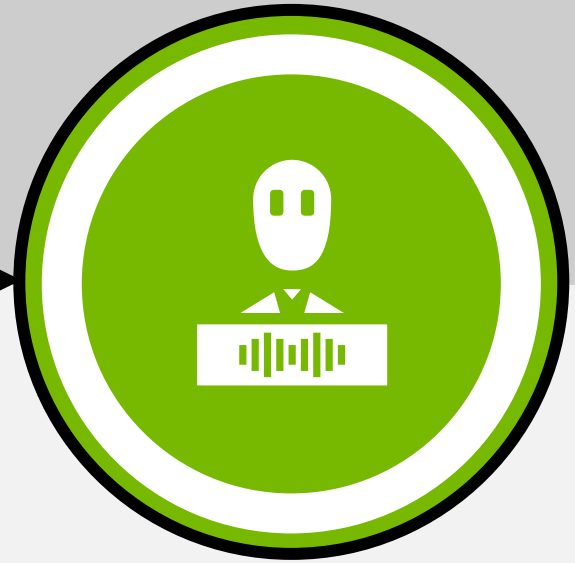
- Observe and understand inputs
- Lacks autonomous action

Ex: Computer vision, object detection.

Generative AI

- Novel outputs
- Coherent, contextually relevant
- Assist human creativity

Ex: ChatGPT, customer service chatbots



Agentic AI

- Act independently
- Memory, reasoning
- Interacting with other agents

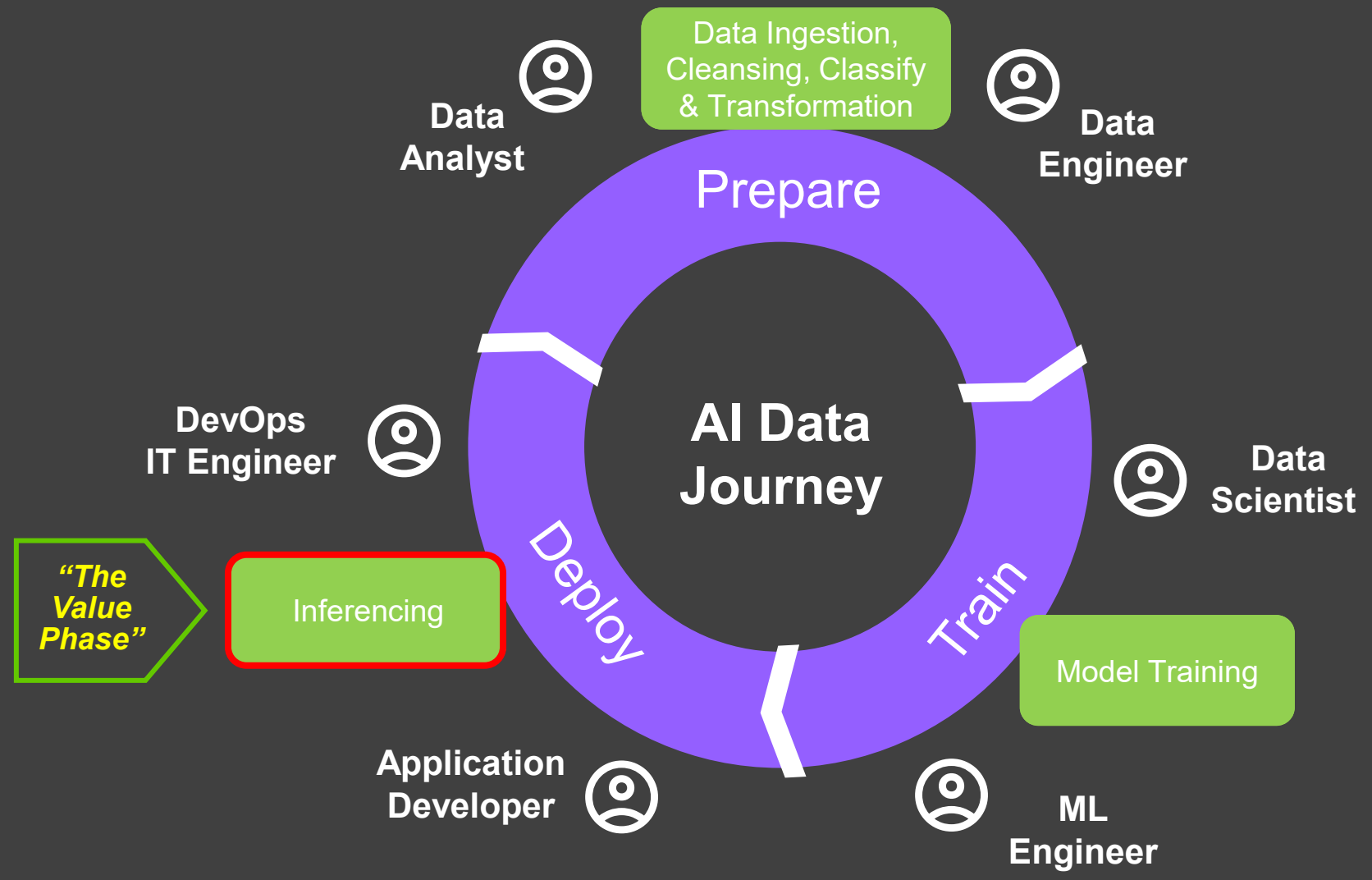
Ex: Managing schedules, making purchases, solving multi-layered problems

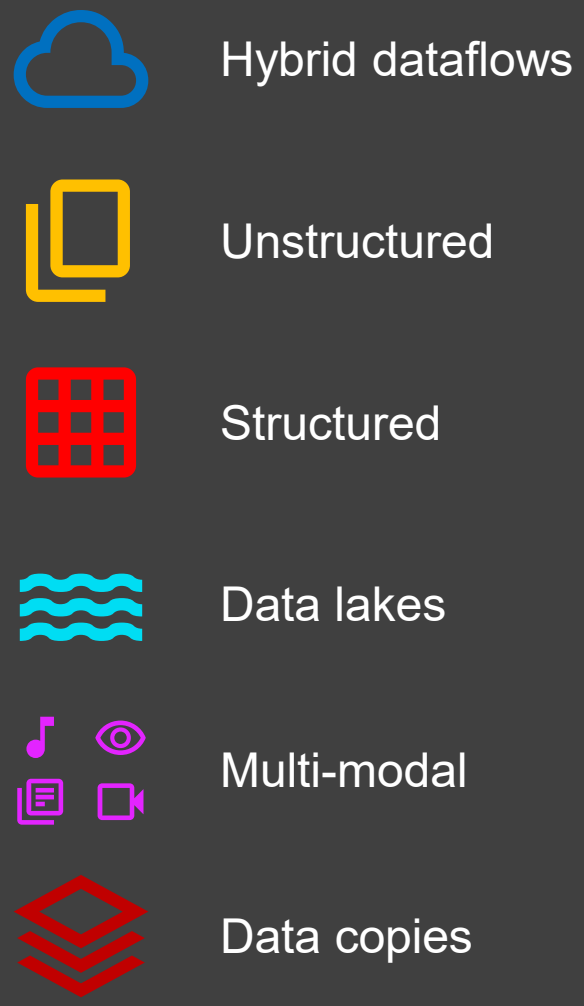
Single application use cases

AI Factories
creating agents at massive scale

Visualizing the AI data journey

Much more than training





WHY YOU STRUGGLE WITH AI PROJECTS

The challenges faced by organizations seeking to derive value from AI

Challenges Deploying Generative AI

The world's enterprises need a better platform for building Generative AI

- x Constrained developer productivity
- x Poor infrastructure utilization
- x Escalating compute demands

3 hours

Average frequency with which LLM training runs crash[1]

30%

Typical compute utilization when training advanced LLMs [3]

5X

yearly compute growth for frontier AI models[4]

[1] Anthony Garreffa. Meta's huge 16,384 NVIDIA H100 AI GPU cluster: HBM3 memory crashed half of Llama 3 training. July 2024.
[2] Paul Gillin. AI hallucinations: The 3% problem no one can fix slows the AI juggernaut. Silicon Angle. Feb 2024.
[3] Chowdery, Aakanksha, et al., "PaLM: Scaling Language Modeling with Pathways," arXiv, October 2022.
[4] Jaimie Sevilla and Edu Roldan. Training Compute of Frontier AI Models Grows by 4-5x per Year. Epoch AI. May 2024.

85%

of AI projects fail

Top reasons AI and ML projects fail

1. **Data access**
2. Model development and deployment
3. Service complexity
4. Data governance
5. Rising costs
6. AI bias

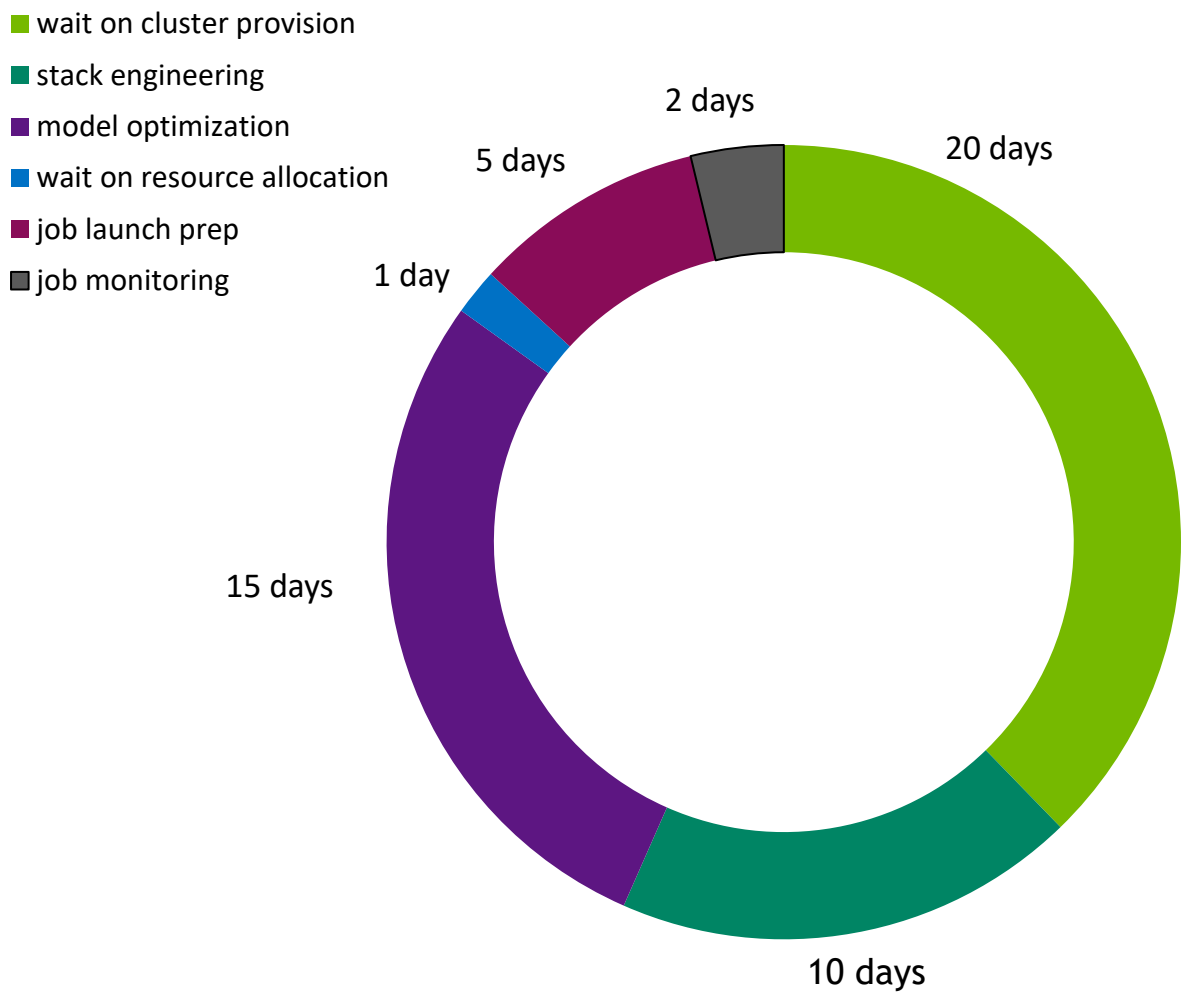
Source: OZ Digital Consulting, Why 85% of AI and ML Projects Fail—How to Ensure Yours Succeeds, February 16, 2024.

Why Are AI Initiatives Costing More Than Anticipated?

Expending effort on non-development work

"Hidden" AI challenges that drive up OpEx

- Delays in infrastructure provisioning
- Effort expended on AI code modification / adaptation
- Training job troubleshooting / resource utilization inefficiency



An AI developer could lose over 30 days on non-value add "effort"

- $30k - $50k lost productivity per developer
- Potentially $1m+ across an AI team of 30 developers
- Unaccounted when observing purely the cost of infrastructure

CHALLENGES IN AI ADOPTION

Organizations need to rethink their approach to storage and infrastructure to compete in the AI race

63%

believe that storage needs an improvement or complete overhaul for AI

20%

have mature, centralized policies for data governance and security for AI

27%

cite poor data access due to infrastructure as the top cause of AI project failure

"Forecast Analysis: AI-Optimized Servers, Worldwide", Gartner, November 2024
"AI Transformation Study", IDC, January 2024, n=1220
"Scaling AI Initiatives Responsibly: The Critical Role of an Intelligent Data Infrastructure", IDC, May 2024

A VISION FOR THE FUTURE

The power of investing in AI infrastructure with NetApp and NVIDIA

Why NetApp and NVIDIA for an AI Factory

More than a science project

1. AI is coming ***out of the lab*** and into the ***office***

2. ***Shared resources*** and ***multi-tenancy*** are the future of AI infrastructure

3. Today's AI requires ***feature-rich infrastructure*** that meets the demands of enterprise workloads.

HPC competency isn't enough

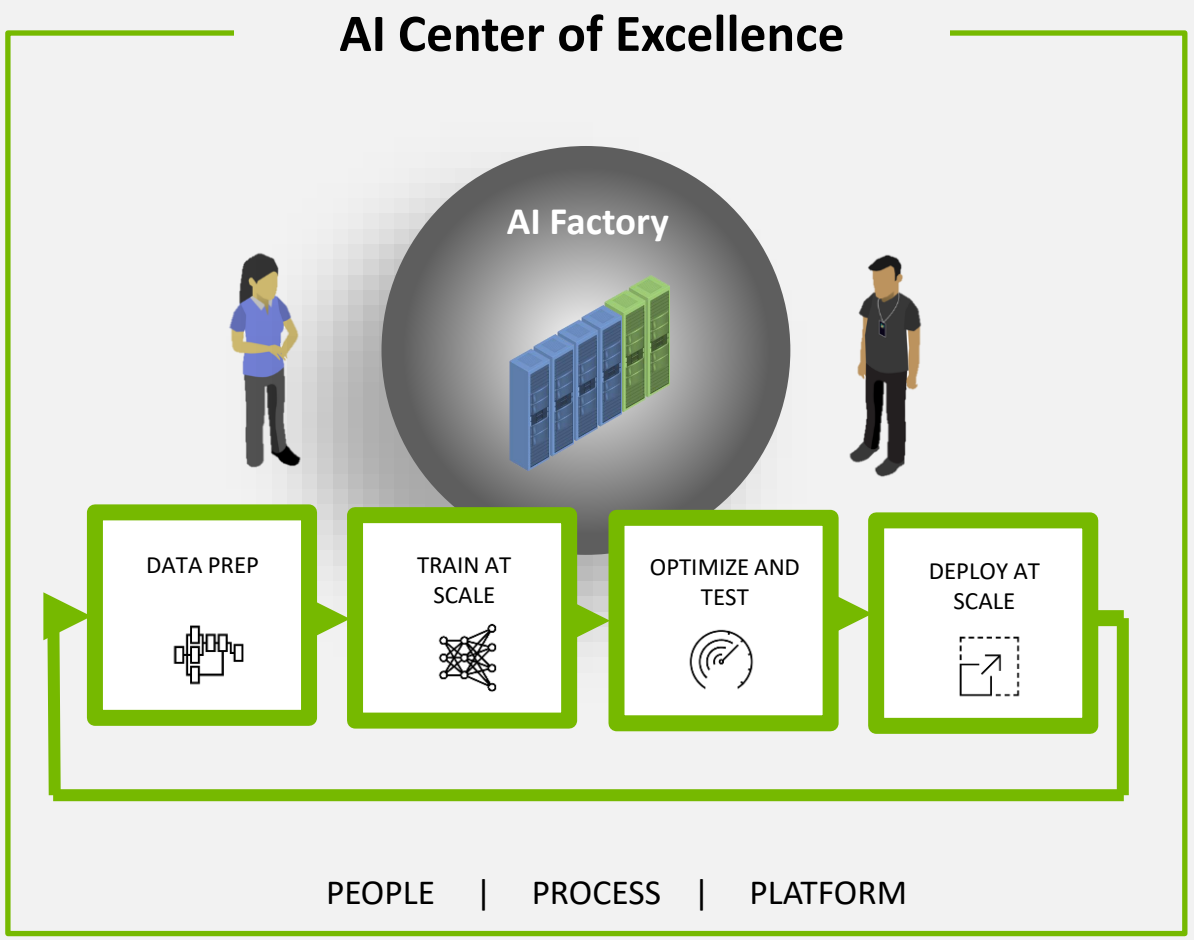
Enabling simultaneous workflows for next-generation AI factories with ONTAP

NVIDIA DGX SuperPOD + NetApp AFF A90

Five Benefits from Investing in AI Infrastructure

Your dedicated platform for turning data into intelligence

1. Improve infrastructure utilization
2. Maximize developer productivity
3. Centralized AI development flywheel
4. Reduce AI project TCO, speed ROI
5. Grow in-house AI expertise from within



AI Factories Power the Next Industrial Revolution

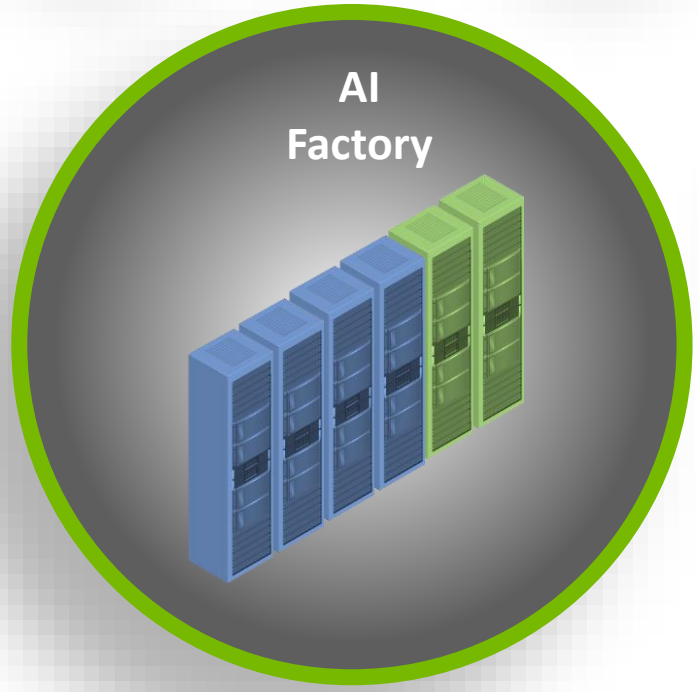
The manufacturer of “intelligence” - now the essential engine of every enterprise

Drive Innovation

Analyze vast amounts of data to identify market trends and consumer preferences

Foster Agility

Scale AI solutions quickly across functions and tasks, easily adapt to market changes



Cut Costs

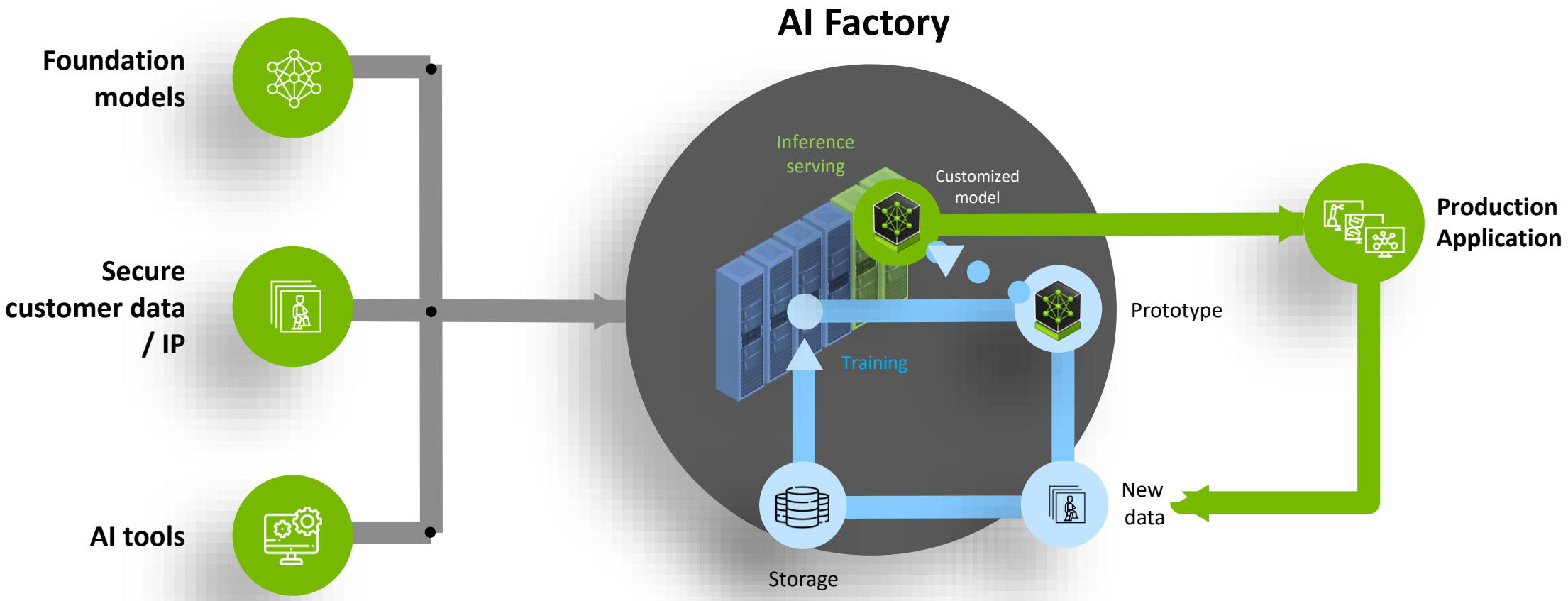
Reduce operational costs by automating processes and optimizing resource usage

Enhance Productivity

Streamline processes, reduce downtime, automate repetitive tasks

AI Factories That Manufacture Intelligence at Scale

Enabling the AI application flywheel



Your Artificial Intelligence solution: driven by NetApp

Maximize the potential of AI and secure a competitive advantage with NetApp

On-prem

Hybrid

Public Clouds

NVIDIA

NVIDIA-certified solutions

NetApp

INTELLIGENT & INTEGRATED

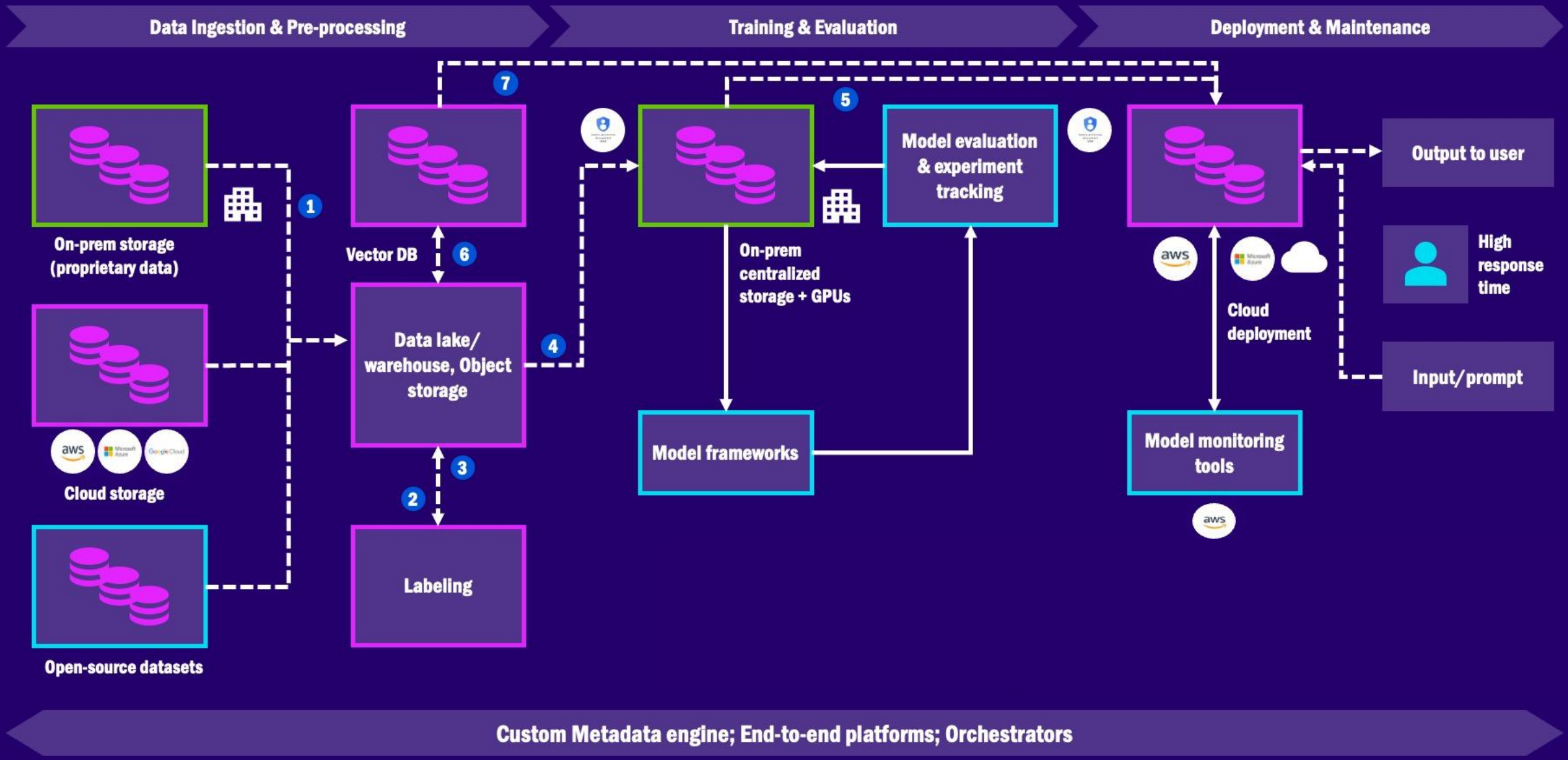
SCALABLE PERFORMANCE

TRUSTED & SECURE

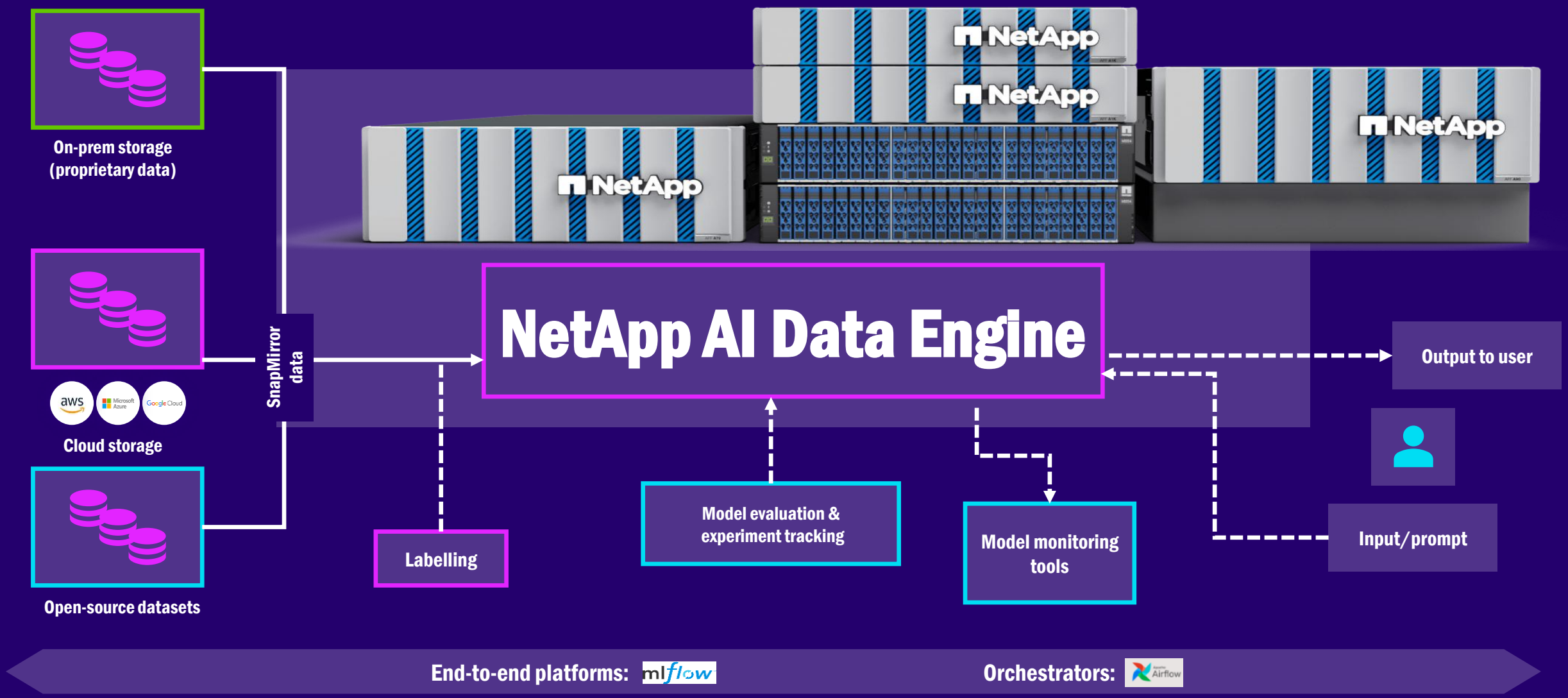
The data infrastructure foundation for enterprise AI

AI SOLUTIONS

One customer's challenge… a simplified view



Simplified customer Gen AI tech stack – WITH NetApp AI Data Engine

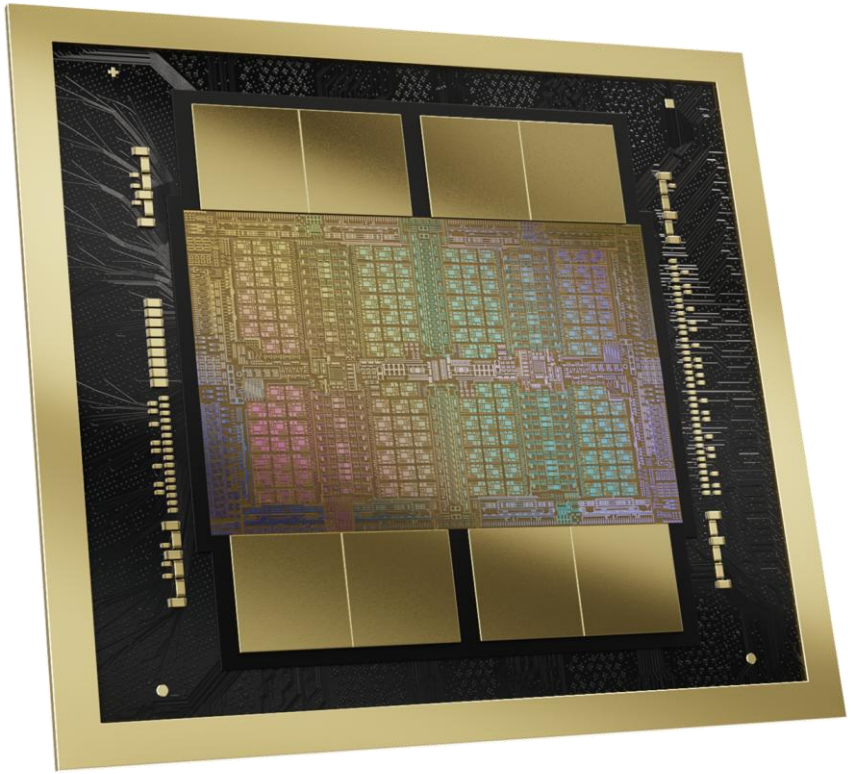


THE PORTFOLIO

How NetApp and NVIDIA are delivering success today for AI Projects

NVIDIA Blackwell – Available NOW

The Engine of Every AI Factory



Built to Democratize Trillion-Parameter AI

20 PetaFLOPS of AI performance on a single GPU

4X Training | 30X Inference | 25X Energy Efficiency & TCO

Expanding AI Datacenter Scale to beyond 100K GPUs

AI SUPERCHIP
208B Transistors

2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

5th GENERATION NVLINK
Scales to 576 GPUs

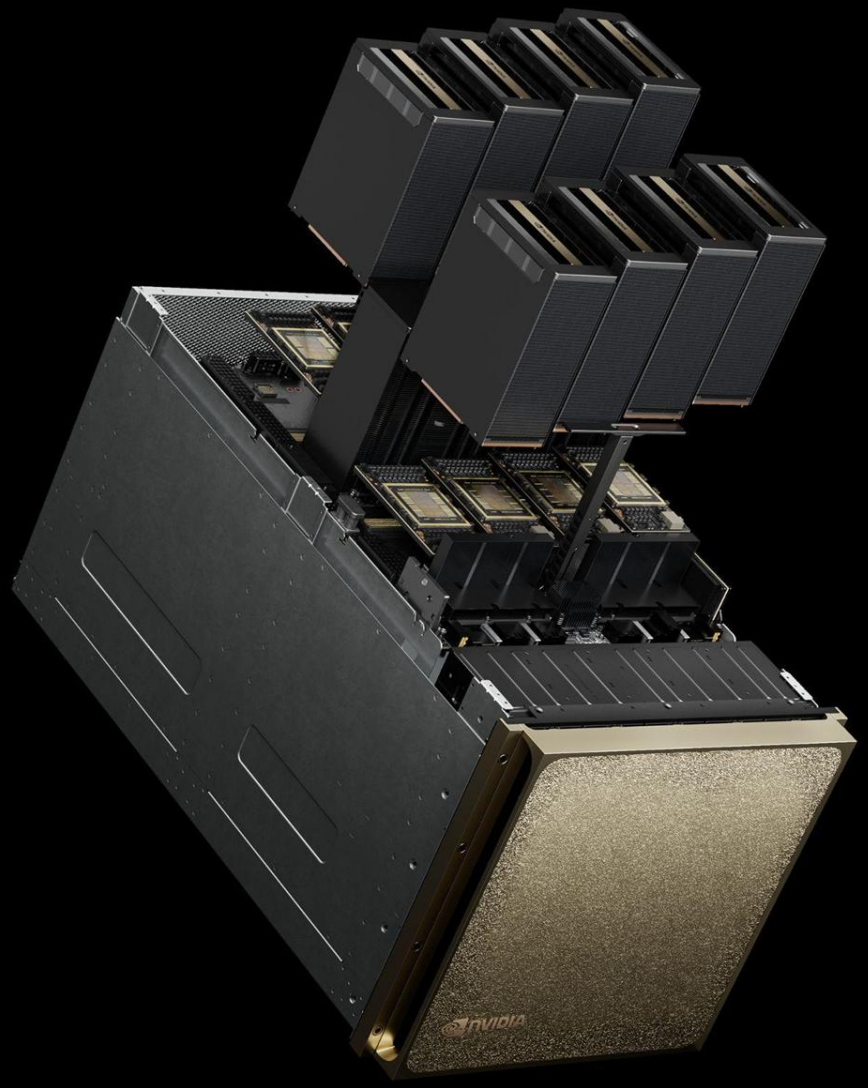
RAS ENGINE
100% In-System
Self-Test

DECOMPRESSION ENGINE
800 GB/s

NVIDIA DGX B200:
The Gold Standard for AI Infrastructure

The proven choice for enterprise AI is now even better

- 8x NVIDIA B200 GPUs with 1,440 Gigabytes of total GPU memory
  - 18x NVIDIA NVLink connections per GPU, 900 gigabytes per second of bidirectional GPU-to-GPU bandwidth
  - 64 TB/s memory bandwidth
- 2x NVIDIA NVSwitches
  - 7.2 terabytes per second of bidirectional GPU-to-GPU bandwidth
- 4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VP
  - Up to 400Gb/s InfiniBand/Ethernet
- 2x dual-port QSFP112 NVIDIA BlueField-3 DPU
  - Up to 400Gb/s InfiniBand/Ethernet
- Dual Intel® Xeon® Platinum 8570 processors (112 cores total) and 4TB system memory
  - Powerful CPUs and massive system memory for the most intensive AI jobs
- 30 terabytes NVMe SSD
  - High-speed storage for maximum performance
- 72 petaFLOPS training and 144 petaFLOPS inference



AI Factory Accelerated Infrastructure: NVIDIA DGX SuperPOD

Powered by NVIDIA Blackwell

- World's fastest commercially available AI infrastructure
- Turnkey AI data center solution
- Best-of-breed tools for developers and IT
- Integrated, optimized software that keeps getting faster
- Designed and deployed by NVIDIA, optionally managed by certified partners

The NetApp AI opportunity - Today

A broad portfolio of solutions

High Performance Model Training	Data Prep/Data Lake Modernization	Model Training & Fine Tuning	RAG & Inferencing
EF-Series with BeeGFS for NVIDIA DGX SuperPOD  **AFF A-Series** for NVIDIA DGX SuperPOD 	FSXn StorageGRID  **AFF**  **FAS** **BlueXP** data classification	FSXn  **AFF A-Series or C-Series** for NVIDIA DGX BasePOD, DGX SuperPOD, and OEM Servers with NVIDIA GPUs	FSXn **NetApp AIPod** FlexPod Lenovo instaclustr **BlueXP** data classification

NVIDIA DGX SuperPOD + NetApp AFF

ONTAP for the highest-performance AI workloads

NetApp has certified NetApp ONTAP storage on the AFF A90 system with NVIDIA DGX SuperPOD AI infrastructure (through B200 GPUs)

NVIDIA Cloud Partner

NetApp AFF A90 offers differentiated value for cloud service providers

DATA MANAGEMENT AND ACCESS

- Unlock enterprise-wide data access with unified management and control

SCALABILITY

- Scale your AI infrastructure without boundaries. Unify resources without compromise

SECURITY

- Enterprise-grade security for AI innovation: Built-in data protection, authentication, and multi-tenancy

Certified through B200 GPUs

Enabling simultaneous workflows for next-generation AI factories with ONTAP

Certified to support the NVIDIA Cloud Partner Reference Architecture

AI turnkey solutions with NetApp AIPod

Get started quickly, reducing cost and complexity

Consolidate a data center's worth of analytics, training, and inference compute into a single AI infrastructure

NetApp® AIPod™ architectures with DGX Systems

- Simplify, accelerate, and integrate your data pipeline for ML and DL with flexible, NVIDIA-validated solutions.
- Deliver the right performance and nondisruptive scalability.
- Eliminate infrastructure silos and unify AI workloads.
- Take advantage of the NVIDIA DGX BasePOD and reference designs.

AIPod for NVIDIA Enterprise Systems

NetApp's AIPod is certified to support the NVIDIA Enterprise Reference Architecture

NVIDIA Certified

Empower your business to design an architecture that suits your needs.

Choose your server and leverage NVIDIA's design for unmatched performance and confidence

	Foundation	Enterprise
	Flexibility + Performance for Enterprise AI and HPC, LLM Inference, RAG Workloads	**Highest Performance for HPC, LLM Training / Inference Workloads**
	Partner Validated	Validated in NVIDIA lab
Compute	16 node cluster with BlueField adapters	Large cluster with BlueField adapters
Networking	Spectrum 3/4/X Ethernet switches	Spectrum-X Ethernet switches
Supports	NVIDIA-Certified servers following 2-4-3 and 2-8-5 NVIDIA Enterprise Reference Architectures	NVIDIA-Certified servers following 2-4-3, 2-8-5, and 2-8-9 NVIDIA Enterprise Reference Architectures

NetApp Unlocks Exabytes of Data for Secure, Private Gen AI

NetApp

- Documents
- Spreadsheets
- Presentations
- Tech Drawings
- Images
- Meeting Recordings
- Databases
- Applications
- ERP/CRM Data

NETAPP CUSTOMER DATA

ON-PREMISES STORAGE SYSTEMS

aws Microsoft Azure Google Cloud

EQUINIX

NVIDIA NeMo Retriever

CHATBOTS

Agents

APPLICATIONS

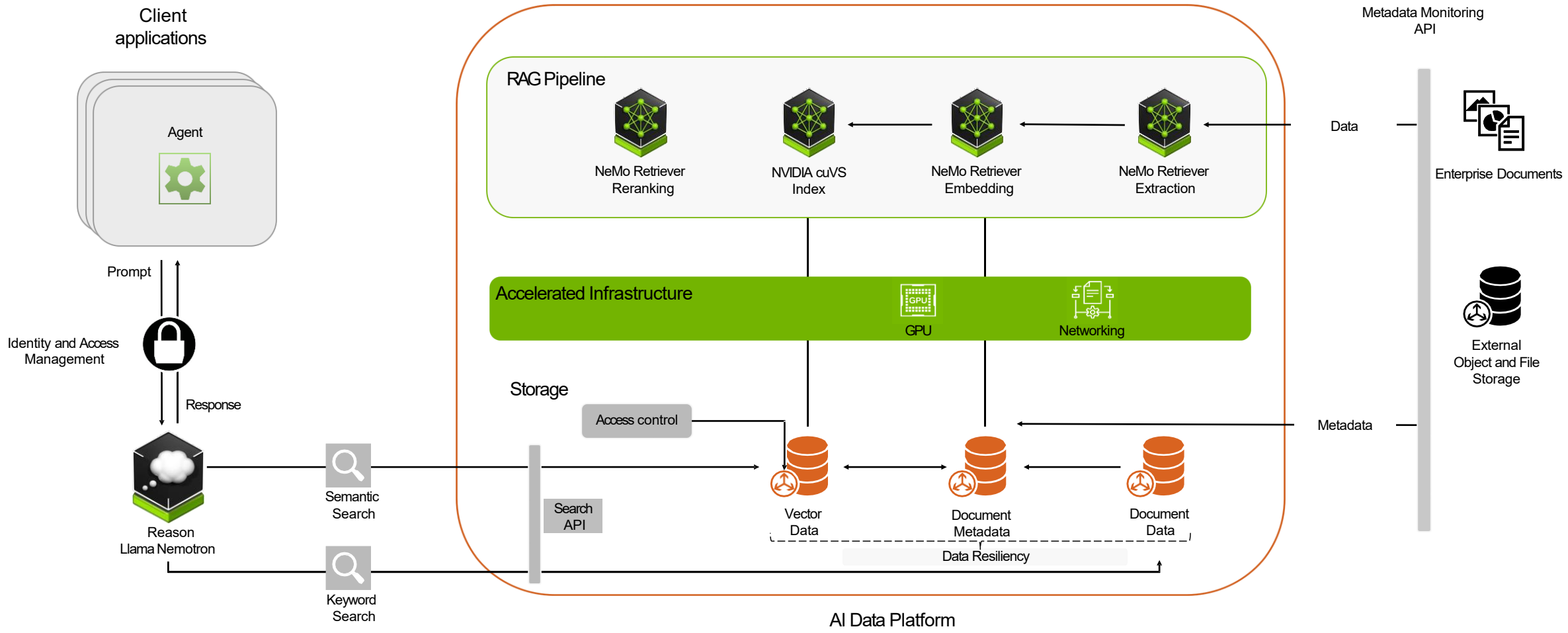
Gen AI with NetApp

- Access Control
- Data Compliance
- Disaster Recovery
- Data Security

NETAPP CUSTOMERS CAN NOW "TALK TO THEIR DATA"

AI Data Platform

Enabling Storage Partners to Bring AI to Private Enterprise Data



Turnkey

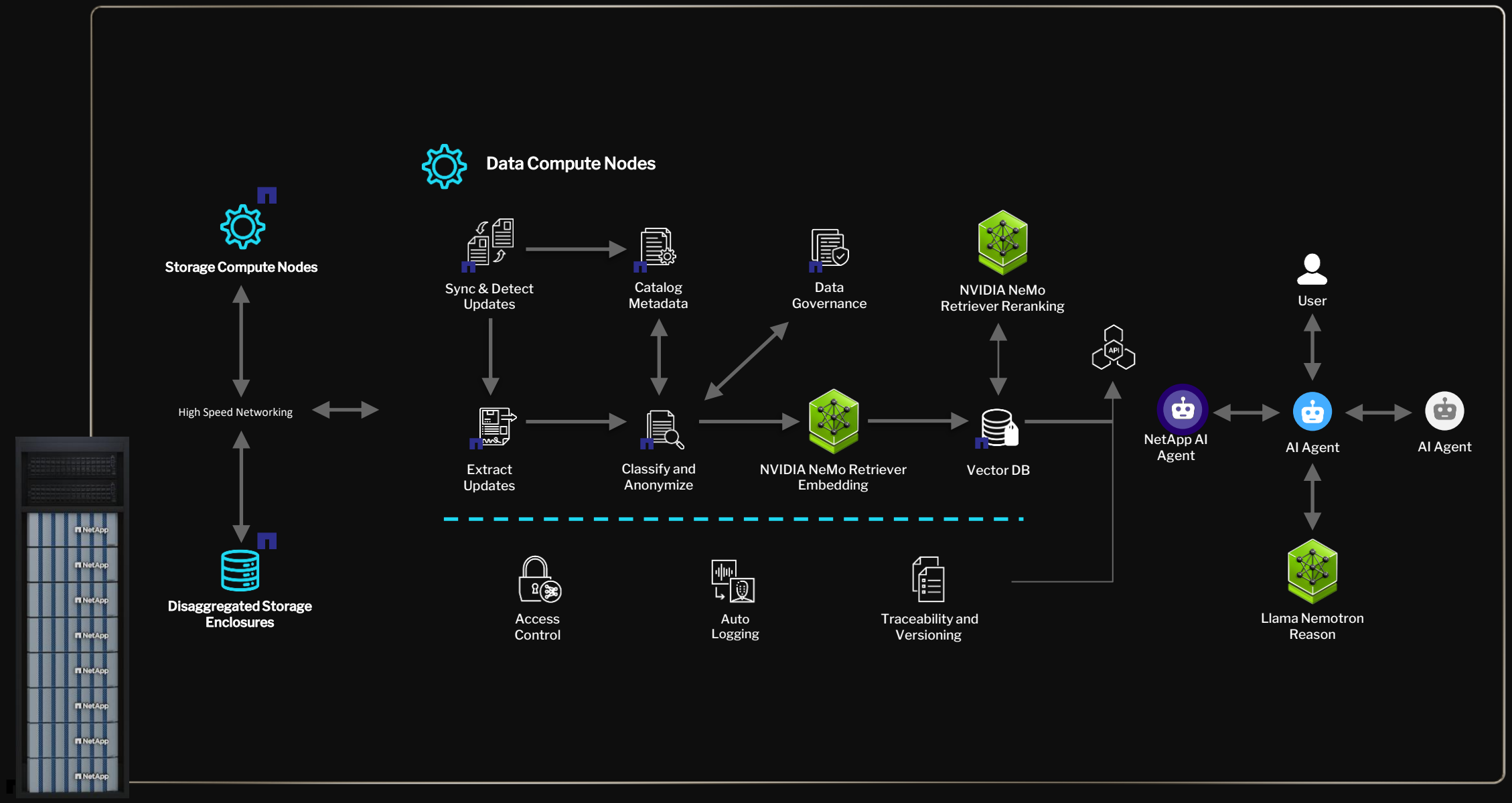
Multimodal search

Data Security

High Throughput at Scale

NetApp

NetApp

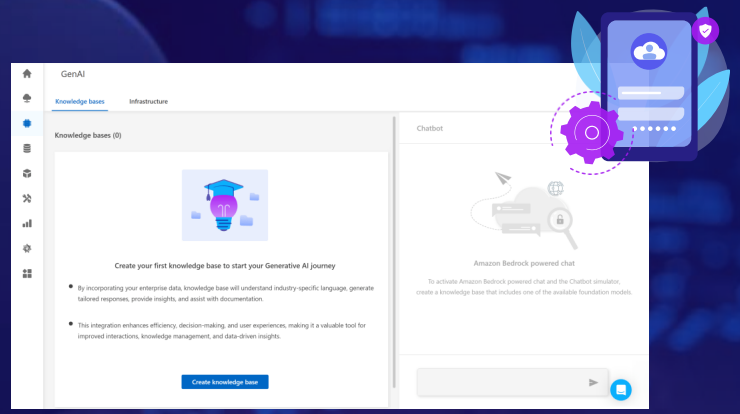

Data Compute Nodes
Storage Compute Nodes
High Speed Networking
Disaggregated Storage Enclosures
Sync & Detect Updates
Catalog Metadata
Data Governance
NVIDIA NeMo Retriever Reranking
Extract Updates
Classify and Anonymize
NVIDIA NeMo Retriever Embedding
Vector DB
API
User
NetApp AI Agent
AI Agent
AI Agent
Llama Nemotron Reason
Access Control
Auto Logging
Traceability and Versioning

NETAPP CLOUD STORAGE

AI READY

Google Cloud Vertex AI +
Google Cloud NetApp Volumes
w/ BigQuery and Agent Builder

Google Cloud



BlueXP Workload Factory
AWS Bedrock Knowledge Bases
on AWS FSx for NetApp ONTAP

aws

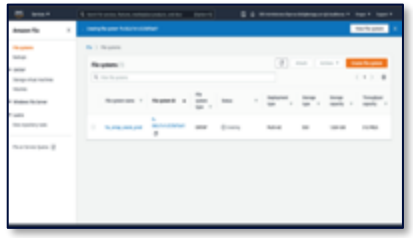
Azure NetApp Files Now
Supported with
Azure OneLake

Microsoft Azure

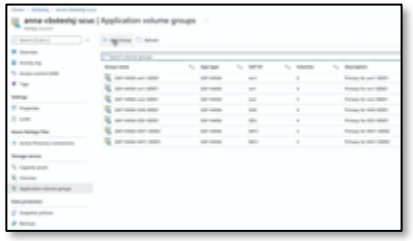
NetApp solves the multi-cloud data challenge

Unified data storage delivered directly inside the hyperscaler console

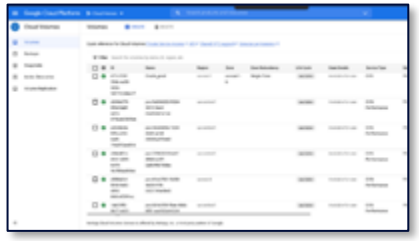
The only enterprise storage managed service available in the top 3 hyperscalers



Amazon FSx for NetApp ONTAP



Azure NetApp Files



Google Cloud NetApp Volumes

Only NetApp offers:

- 118 regions of availability in over 30 countries across 3 hyperscalers
- Fully managed service integrated directly into hyperscaler control panels
- FedRAMP, GDPR, HIPAA, and more.
- Integrated data protection, disaster recovery, and multi-zone high availability
- Hybrid cloud ready

You need trusted partners

NetApp and NVIDIA deliver the enterprise capabilities necessary for business-critical AI infrastructure

You need...	NetApp + NVIDIA deliver...
DATA ACCESS AND MOBILITY	SEAMLESS DATA MANAGEMENT
ENTERPRISE-GRADE SECURITY	ZERO-TRUST SECURITY
UNIFIED DATA ACCESS	AI-READY DATA PIPELINE
24x7x365 UPTIME	PRODUCTION-SCALE RELIABILITY
SHARED RESOURCE OPTIMIZATION	MULTI-TENANT INTELLIGENCE

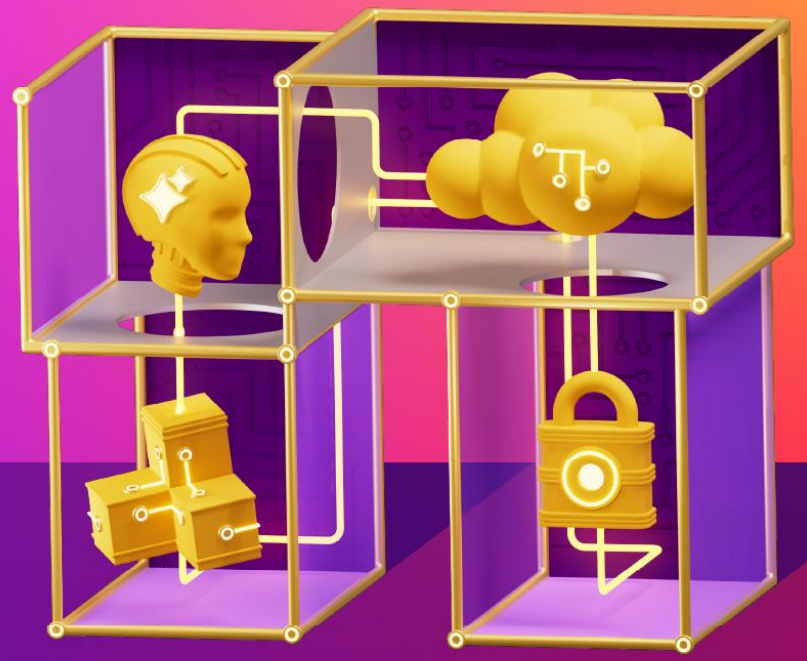
Q&A / DISCUSSION

NetApp

NetApp

INSIGHT

October 14 – 16, 2025 | MGM Grand, Las Vegas



www.netapp.com/insight

THANK YOU

NetApp