Solution Brief

NetApp Solutions for the AI Pipeline and Edge Inference

NetApp provides full data mobility across the deep learning pipeline from edge to core to cloud, enabling continuous improvement of data preparation, training, and outcomes

Key Benefits

- Simple, flexible, and powerful infrastructure for a deep learning pipeline and AI edge inference workloads
- Access to data across the data fabric: edge, core, and cloud
- Easy deployment, management, and scaling
- Quality of service (QoS) for true multiworkload support

The AI Deep Learning Pipeline

For AI to work well, you need to build strong data management practices and deep learning pipelines. Cleaning, preparing, and storing data are critical jobs, as are building training and inference models for deployment and measurement. To avoid the most common pitfalls of the deep learning pipeline, you need a powerful, flexible infrastructure.

As Figure 1 shows, NetApp® infrastructure gives you data mobility across the deep learning pipeline, from the edge to the core to the cloud, enabling continuous improvement of data preparation, training, and outcomes.

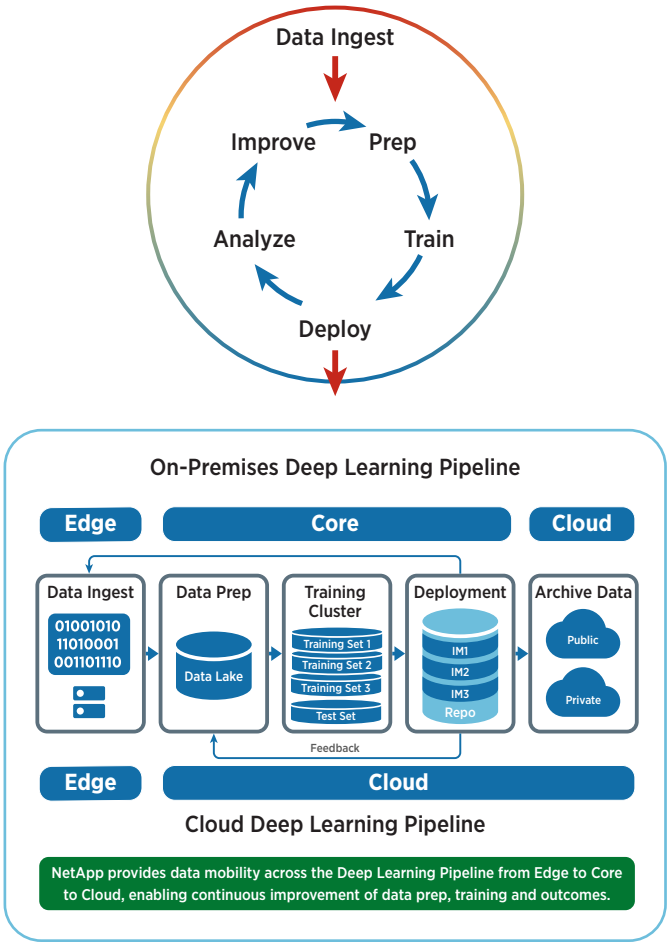

Figure 1) The NetApp powerful, flexible infrastructure helps optimize your deep learning pipeline.

NetApp®

Pitfalls of the AI Deep Learning Pipeline

The combination of new data and new workloads presents manageability and infrastructure challenges. Data scientists, researchers, and the IT operations teams who support them face common challenges in the areas where the AI deep learning pipeline can break down, affecting performance and productivity, including:

- Constrained, slow, or manual access to data and trained models
- Technical complexity of infrastructure and tools

Data is often in disparate silos. Obtaining access is cumbersome, and the technical complexity of supporting the compute, storage, and bandwidth needs of these users reduces efficiency and affects business outcomes.

The NetApp AI Pipeline Infrastructure

With the NetApp simplified AI pipeline infrastructure, you get full data mobility across the pipeline, from the edge to the core to the cloud. You can continuously improve your data preparation, training, and outcomes so that you can:

- Enable your data scientists to run more AI projects faster.
- Optimize access to premium GPU resources, whether they're on premises or in the cloud.
- Seamlessly scale performance and capacity as your AI projects grow.

AI inference servers

AI inference servers aren't new. Most of us use inference every day. Google speech recognition, image search, and spam filtering applications are all examples of inference workloads. Facebook image recognition and Amazon and Netflix recommendation engines are all based on inference. And inference systems extend well beyond Google and Facebook tools into smart cities, automotive applications, medical diagnostics, agriculture, business analytics, media and entertainment, and more.

NetApp technology in combination with the NVIDIA TensorRT Inference Server supports edge AI inference workloads such as:

- General recommender systems
- Manufacturing quality control
- Automatic video analysis
- Medical imaging analysis
- Retail and kiosk image recognition

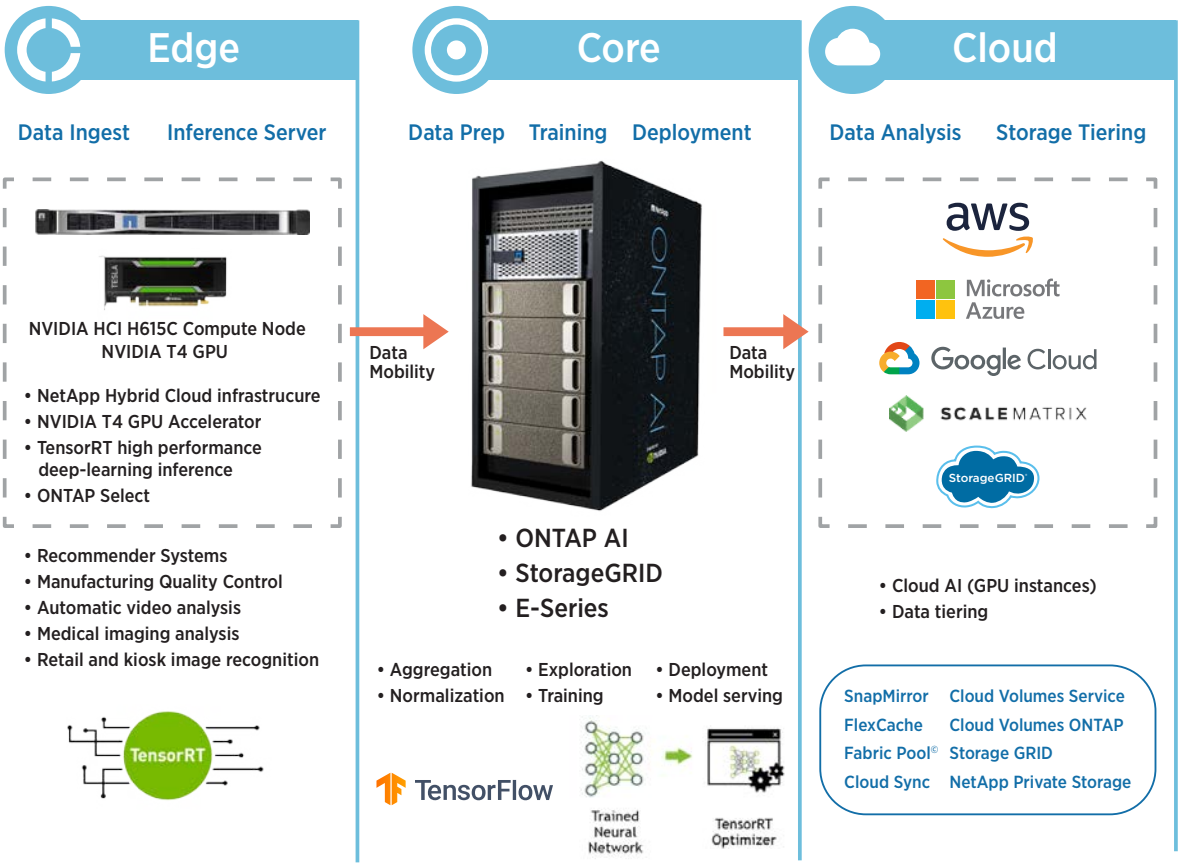

Figure 2) NetApp solutions work across the deep learning pipeline.

NetApp HCI: A Simple, Flexible Solution for AI at the Edge

With this hybrid cloud infrastructure, NetApp HCI acts as a powerful AI inference server, with ONTAP AI creating the trained model. After you train a model, you can deploy it to perform inference workloads. With NetApp FlexCache caching, your data scientists can access the trained model without exporting the full model, improving performance and ease of use. When the inference model is deployed, results can be fed back into the training model to improve deep learning.

Although you can run inference in TensorFlow itself, your applications deliver higher performance by using TensorRT Inference Server (TRTIS) on NVIDIA GPUs. TensorFlow models that are optimized with TRTIS can be run on T4 GPUs on NetApp HCI in edge inference deployments. As Figure 3 shows, NetApp recommends this approach.

In edge environments, NetApp HCI provides important benefits that a traditional infrastructure model does not, including:

- **Independent scaling.** Maximize your resources and minimize the hypervisor tax. Grow your IT based on your business needs, not architectural deficiencies.
- **Workload consolidation.** No more silos, no more unnatural workload constraints. Achieve more secure data, better QoS, dedicated performance, and guaranteed service levels.
- **Open hybrid multicloud.** No cloud lock-in. You get consistent IT consumption across public cloud, private cloud, and on-premises environments.
- **Guaranteed efficiency.** Get the industry's most effective storage efficiency guarantee, with no impact on your system performance.
- **Simplicity.** Easily support your virtualized environment as is. NetApp HCI is transparent and does not require changes to policies or procedures.
- **Proactive protection.** Monitor, troubleshoot, and optimize your entire infrastructure with NetApp Cloud Insights. Prevent issues early on and accelerate issue resolution with the NetApp Active IQ® intelligence engine.

About NetApp

NetApp is the leader in cloud data services, empowering global organizations to change their world with data. Together with our partners, we are the only ones who can help you build your unique data fabric. Simplify hybrid multicloud and securely deliver the right data, services and applications to the right people at the right time. Learn more at www.netapp.com.

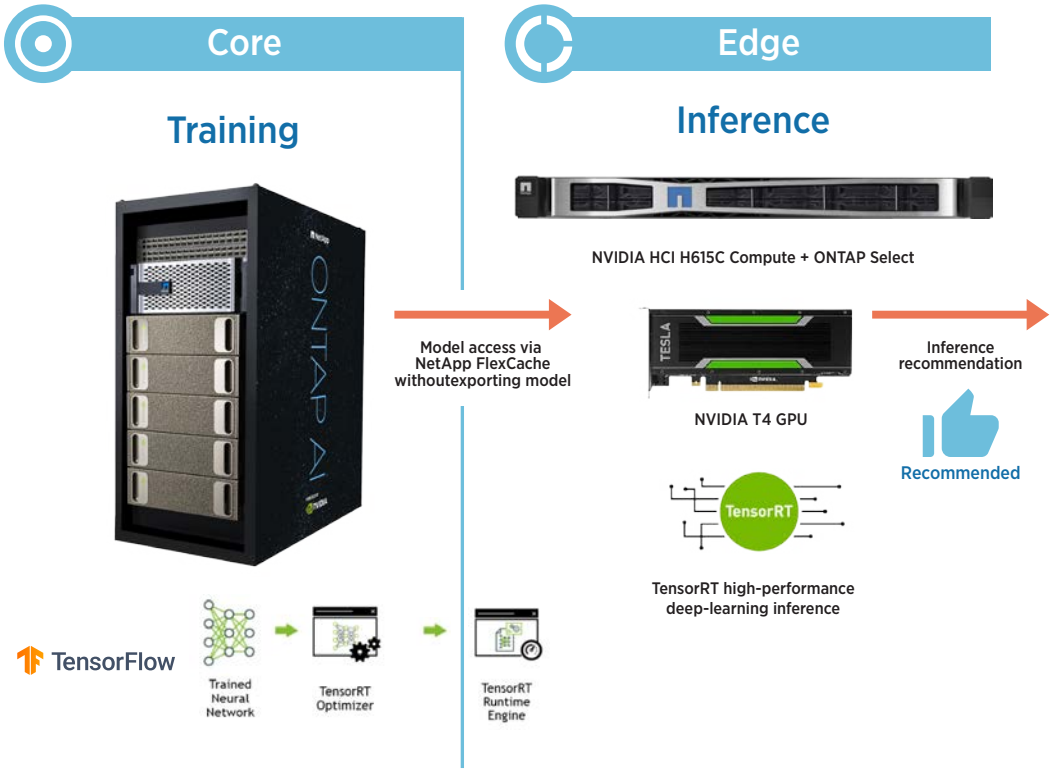

Figure 3) TensorRT Inference Server helps optimize edge inference.