

CISCO AI FACTORY TEST DRIVE

Service Overview:

At Netnology, we specialize in AI-optimized infrastructure design and enablement services built on Cisco's AI FACTORY architecture. These services are designed to help organizations accelerate the adoption of AI by leveraging high-performance networking, GPU-accelerated compute, and integrated AI software platforms.

This up to 4-hour interactive session demonstrates how AI workloads operate on Cisco AI FACTORY infrastructure, showcasing the integration of Cisco UCS GPU servers, high-performance Cisco Nexus networking, Cisco Secure Suits and AI software frameworks.

This test drive is built and delivered by subject matter experts (SMEs) who have extensive experience working with AI Solution.

Solution Overview:

Cisco Secure AI Factory with NVIDIA is a modular reference architecture designed to help enterprises build, deploy, and operate secure and scalable AI infrastructure. The solution integrates high-performance compute, accelerated GPU platforms, high-bandwidth networking, storage, security, and observability into a unified architecture that enables organizations to develop and run AI applications at enterprise scale. By combining Cisco infrastructure with NVIDIA accelerated computing and AI software, the platform provides a validated foundation for running AI workloads such as model training, fine-tuning, and inference in enterprise environments.

The architecture addresses the challenges enterprises face when operationalizing AI infrastructure by providing a modular and validated design that integrates AI software, compute platforms, networking fabric, data services, security capabilities, and Kubernetes-based platforms into a cohesive solution. This approach simplifies deployment, reduces operational complexity, and enables organizations to accelerate the delivery of trusted AI applications while maintaining performance, scalability, and security across the entire AI stack.

- GPU Acceleration: Facilitates the creation and training of deep learning models using GPU clusters.
- High-Performance Networking: High-speed interfaces ensure robust connectivity and meet demanding network requirements.
- Cisco Nexus 9000 Series: Provides low latency, advanced congestion management, and telemetry for AI/ML workloads.
- Versatile Deployment: Supports AI inference solutions in both centralized data centers and at the network edge with Cisco UCS X-Series and rack servers.
- Advanced Framework Support: Integrates with popular frameworks such as PyTorch, TensorFlow, and JAX for streamlined model development.

Service Benefits:

Netnology has a team of world class engineers who specialize in AI Solution and are passionate about customer success. Netnology will partner with you to provide guidance on:

- Expert AI Strategy: Crafting tailored AI strategies to meet your specific business needs and objectives.
- Seamless Integration: Ensuring smooth integration of AI/ML technologies with your existing infrastructure.
- Optimized Performance: Fine-tuning GPU-accelerated systems and networks for maximum efficiency and performance.

Scope of Services:

As part of the up to 4 hour Test Drive engagement, Netnology will provide the following services:

- Solution Overview
- Solution Value Proposition
- Solution Use-case Demo
 - Cisco Network, Compute Fabric, Security and Observability
 - Fabric Path Validation, Fabric Underlay and Overlay
 - RoCEv2, IMM Policy and Profiles
 - Cisco AI Defense, Secure Firewall, Hypershield
 - Cisco Splunk
 - AI Operating System
 - Redhat Openshift
 - System updates, drivers, and software repositories
 - Package Management
 - LLM Workload
 - Environment Setup
 - Model Loading
 - Text Generation
 - Conversational AI
 - Interactive Q&A
 - AI-driven analytics
- Lessons learned and enhancements

Target Audience:

This service is designed for candidates with little Cisco Network and Compute experience or prior knowledge.

Service Deliverables:

| No. | Deliverable | Service Details |
|-----|-------------|---|
| 1. | Concepts | <ul style="list-style-type: none"> • Solution Overview • Environment Overview • Use-Case Overview |
| 2. | Value-Prop | <ul style="list-style-type: none"> • Ease of configuration • Ease of manageability |
| 3. | Execution | <ul style="list-style-type: none"> • Cisco Network and Compute Fabric • AI Operating System • LLM workload Execution |
| 4. | Wrap-up | <ul style="list-style-type: none"> • Lessons Learned • Q&A |