CISCO SECURE AI FACTORY TEST DRIVE

Service Overview:

At Netnology, we specialize in design and enablement services for Cisco Secure AI Factory with NVIDIA. Our services help organizations accelerate AI adoption on a secure, observable and well-managed AI platform.

This 2–4-hour interactive Test Drive demonstrates how a Cisco AI POD running Red Hat OpenShift is managed with Cisco Intersight, how AI applications are protected with Cisco AI Defense, how Isovalent secures and visualizes Kubernetes workloads, and how Splunk provides operational and security visibility. It also shows how AI models are sized to the right GPU infrastructure.

Delivered by Netnology subject matter experts (SMEs), this session provides a practical introduction to the Secure AI Factory platform, AI security and operational visibility.

Solution Overview:

Cisco Secure AI Factory with NVIDIA is a modular reference architecture that helps enterprises build, deploy and operate AI workloads securely. Cisco AI PODs are its pre-validated building blocks, combining Cisco UCS GPU servers, NVIDIA AI Enterprise software and Red Hat OpenShift as the enterprise Kubernetes platform for AI workloads, all managed through Cisco Intersight.

Security and observability are built into every layer. Cisco AI Defense validates models and AI applications through AI red teaming and protects them at runtime with policy-driven guardrails. Isovalent (Cilium, Hubble and Tetragon) provides identity-based network policy, service and flow visibility, and runtime security for the OpenShift workloads. Splunk brings infrastructure, GPU, Kubernetes, AI application and security telemetry together for monitoring, alerting and investigation.

Key solution capabilities include:

- Cisco Intersight: Policy-based management, monitoring and lifecycle of AI POD UCS GPU servers.
- Cisco AI POD with Red Hat OpenShift: Pre-validated AI infrastructure with an enterprise Kubernetes platform and NVIDIA GPU Operator.
- Model Sizing: Matching models, concurrency and performance targets to GPU type, GPU count and AI POD size.
- Cisco AI Defense: Policy-based runtime guardrails and AI red teaming for models and AI applications.
- Isovalent: Identity-based network policy, service graph and flow visibility with Cilium and Hubble, plus Tetragon runtime security.
- Splunk: Unified dashboards and alerting across infrastructure, Kubernetes, AI workloads and security events.

Service Benefits:

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

- AI Infrastructure Awareness: Understand the Secure AI Factory architecture and AI POD building

blocks.

- Simplified Operations: See how Intersight and OpenShift simplify AI infrastructure management.
- Right-sized Investment: Learn how model sizing translates into GPU and AI POD requirements.
- AI Application Protection: Understand how AI Defense policy, guardrails and red teaming mitigate AI-specific threats.
- Kubernetes Security: See how Isovalent policy, service graphs and flow graphs secure AI workloads.
- Integrated Visibility: Understand how Splunk correlates infrastructure, workload and AI security events.
- Accelerated AI Adoption: Identify use cases, security considerations and next steps for a proof of value (POV).

Scope of Services:

As part of the 2–4-hour Test Drive engagement, Netnology will provide the following services:

- Solution Overview
- Solution Value Proposition
- Solution Use-case Demo
  - Cisco Intersight
    - AI POD server inventory, policies and profiles
    - GPU health, power and thermal monitoring
  - Cisco AI POD with Red Hat OpenShift
    - OpenShift cluster and NVIDIA GPU Operator
    - GPU scheduling of a sample inference workload
  - AI Model Sizing
    - Sizing walkthrough for a sample model (GPU memory, concurrency and throughput)
    - Mapping requirements to AI POD configurations
  - Cisco AI Defense: Policy, Deployment and Red Teaming
    - Policy and guardrail configuration
    - Deployment pattern for AI application protection
    - Red teaming run and risk report
    - Runtime verdicts for sample prompts (e.g., prompt injection, data leakage)
  - Isovalent: Policy, Service Graph and Flow Graphs
    - Cilium identity-based network policy for AI workloads
    - Hubble service map (service graph)
    - Hubble flow graphs and policy verdicts
    - Tetragon runtime security events
  - Splunk
    - Infrastructure and GPU dashboards
    - AI Defense and Isovalent security events
- Lessons learned and next steps

Target Audience:

This service is designed for IT and security leaders, architects, and platform and security engineers evaluating Cisco Secure AI Factory. Prior Cisco compute or Kubernetes experience is helpful but not required.

Service Deliverables:

No.	Deliverable	Service Details
1.	Concepts	• Solution Overview • Environment Overview • Use-Case Overview
2.	Value-Prop	• Ease of configuration • Ease of manageability • Security and visibility across the AI stack
3.	Execution	• Intersight and AI POD with OpenShift • AI model sizing • AI Defense: policy, deployment and red teaming • Isovalent: policy, service graph and flow graphs • Splunk dashboards
4.	Wrap-up	• Lessons Learned • Q&A and next steps