CISCO SECURE AI FACTORY MENTORED POV PLUS

Service Overview:

At Netnology, we specialize in Cisco Secure AI Factory enablement and implementation services that help organizations adopt AI securely. As part of this Mentored POV Plus service, our subject matter experts (SMEs) partner with your team to prove the value of the Secure AI Factory platform, security and observability layers in your environment, and provide hands-on knowledge transfer along the way.

The engagement is delivered as a mentored proof of value (POV), not as a turnkey production rollout. Activities are performed against the customer's agreed architecture, available hardware/software, supported versions and licensing within the five-day engagement.

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 Observability Cloud brings infrastructure, GPU, Kubernetes, AI application telemetry together for monitoring, alerting and tracing.

This engagement focuses on the platform, security and observability layers of the Secure AI Factory, together with model sizing. The underlying data center network fabric (Nexus switching, VXLAN EVPN, RoCEv2 and lossless Ethernet) is expected to be in place and is covered by Netnology's separate Cisco AI-Optimized Fabric services.

Service Benefits:

Netnology has a team of world-class engineers who specialize in Cisco AI infrastructure and security and are passionate about customer success. This Mentored POV Plus engagement delivers:

- Evidence-based decisions: measurable results against your own success criteria.
- Right-sized investment: model sizing that links your AI use cases to GPU and AI POD requirements.
- Proven AI security: AI Defense guardrails and red teaming, plus Isovalent policy and visibility, tested on your workloads.
- Production blueprint: POV findings feed directly into the production design.

Service Scope:

As part of the 5-day (up to 40 hours) engagement, Netnology will provide the following services:

- **Discovery and Readiness**
  - Validate required Cisco, NVIDIA, Red Hat, AI Defense, Isovalent and Splunk Observability Cloud licenses and subscriptions.
  - Confirm Intersight, OpenShift and SaaS connectivity readiness.

- **AI Model Sizing**
  - Capture the POV model(s), use case, concurrency and latency/throughput targets.
  - Estimate GPU memory and GPU count and map them to the AI POD configuration.

- **Cisco Intersight (up to 3 nodes)**
  - Claim the POV GPU servers into Cisco Intersight and apply server policies and profiles.
  - Validate GPU inventory, health, power and thermal monitoring.

- **Cisco AI POD with Red Hat OpenShift (up to 3 nodes)**
  - Deploy or validate the Red Hat OpenShift cluster on the POV nodes.
  - Install the NVIDIA GPU Operator and validate GPU scheduling with a reference inference endpoint.

- **Cisco AI Defense: Policy, Deployment and Red Teaming (up to 5 policies)**
  - Policy: configure guardrails for prompt injection, jailbreak, sensitive data leakage and harmful content.
  - Deployment: protect one agreed AI application using the supported pattern (inspection API, gateway or connector).
  - Red Teaming: run AI Defense red teaming against the POV endpoint and review the risk report.

- **Isovalent: Policy, Service Graph and Flow Graphs (up to 5 policies)**
  - Policy: apply identity-based network policies to the AI namespace and inference service.
  - Service Graph: visualize application dependencies with the Hubble service map.
  - Flow Graphs: review allowed/dropped flows and policy verdicts; observe selected Tetragon runtime events.

- **Splunk Observability Cloud**
  - Deploy the Splunk OpenTelemetry Collector on the OpenShift cluster and send infrastructure, Kubernetes and GPU (NVIDIA DCGM) metrics to Splunk Observability Cloud, where licensed.
  - Instrument one AI application or inference endpoint for APM tracing, including the AI Defense guardrail call where supported.
  - Build a POV dashboard for GPU health and AI workload performance.
  - Create one GPU detector with email alerting and demonstrate it with a controlled load test.

- **Knowledge Transfer**

Target Audience:

This service is designed for Infrastructure, Platform (Kubernetes/OpenShift), Security and Operations teams responsible for deploying, securing and operating the AI Factory.

Prerequisites:

- Working knowledge of Cisco UCS and Cisco Intersight, Kubernetes/Red Hat OpenShift and basic AI/ML concepts.
- AI POD hardware is racked, cabled and powered on, and the data center network fabric is configured and reachable before kickoff.
- Required subscriptions and licenses are available: Cisco Intersight, NVIDIA AI Enterprise, Red Hat OpenShift, Cisco AI Defense, Isovalent Enterprise and Splunk Observability Cloud.
- Customer provides administrative access to Intersight, OpenShift, AI Defense, Isovalent and Splunk Observability Cloud (organization admin and an ingest access token).
- Customer supplies IP plans, DNS/NTP, certificates, storage details and the proxy/firewall rules needed for SaaS connectivity (Intersight, AI Defense, Splunk Observability Cloud).
- Customer provides target models and use cases, an AI application or inference endpoint to protect, approved test data and success criteria.
- Customer assigns technical contacts for compute, platform, security and observability, and is available for decisions during the engagement.

Out of Scope:

- Data center network fabric configuration (Nexus switching, VXLAN EVPN, RoCEv2, PFC/ECN), available as a separate Netnology Cisco AI-Optimized Fabric service.
- Model training, fine-tuning, data pipelines and AI application development.
- Cisco Hypershield and Cisco Secure Firewall configuration.
- Hardware rack, stack and cabling.

Service Deliverables:

No	Deliverable	Service Details
1.	Project Kickoff	• Project overview • Solution overview • Agree POV success criteria and gather requirements
2.	Pre-Requisite and Readiness Assessment	• Confirm AI POD hardware readiness • License and subscription validation • Intersight, OpenShift and SaaS connectivity readiness
3.	AI Model Sizing	• POV model and workload requirements • GPU memory and GPU count estimate • AI POD mapping

No	Deliverable	Service Details
4.	Cisco Intersight – (up to 3 nodes)	• Server onboarding, policies and profiles • GPU health and monitoring
5.	Cisco AI POD with Red Hat OpenShift - (up to 3 nodes)	• OpenShift cluster deployment or validation • NVIDIA GPU Operator and GPU scheduling
6.	Cisco AI Defense- (up to 5 policies)	• Policy and guardrails • Runtime protection for one AI application • Red teaming run and risk report
7.	Isovalent - (up to 5 policies)	• Identity-based network policy • Service graph (Hubble service map) • Flow graphs and Tetragon runtime events
8.	Splunk Observability Cloud	• OpenShift, Kubernetes and GPU metrics via OpenTelemetry • APM tracing of one AI application • POV dashboard • GPU detector and email alert
9.	Knowledge Transfer	• Explain how to configure and manage the deployed solution in the customer's environment