CISCO SECURE AI FACTORY MENTORED INSTALL

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 Install service, our subject matter experts (SMEs) partner with your team to deploy and integrate the AI platform, security and observability layers of the Secure AI Factory, and provide the knowledge transfer your staff need to operate them effectively.

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 Install engagement delivers:

- Right-sized AI infrastructure: model sizing aligns GPU and AI POD capacity with your models and performance targets.
- Consistent, policy-driven operations: AI POD servers managed end-to-end through Cisco Intersight.
- Secure AI applications: Cisco AI Defense guardrails and red teaming identify and block AI-specific threats.
- Zero-trust workload security: Isovalent identity-based policy, service graphs and flow visibility for OpenShift workloads.
- Unified visibility: Splunk Observability Cloud dashboards and detectors across infrastructure, GPU and AI workloads.

Service Scope:

As part of the 10-day (up to 80 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 AI POD software/firmware compatibility against the agreed design.
  - Record dependencies, risks, constraints and out-of-scope items before implementation.

- **AI Model Sizing**
  - Capture target models, use cases (inference, RAG, fine-tuning), user concurrency, context length and latency/throughput targets.
  - Estimate GPU memory per model (model weights, precision/quantization and KV cache) and the GPU count required.
  - Map the requirements to Cisco AI POD configuration, GPU type and server count, including headroom for growth.

- **Cisco Intersight (up to 6 nodes)**
  - Claim Cisco UCS GPU servers (and Fabric Interconnects, where used) into Cisco Intersight.
  - Create server policies and profile templates (BIOS, boot, LAN/vNIC, management) for AI POD nodes.
  - Deploy server profiles and validate firmware and hardware compatibility (HCL) compliance.
  - Validate GPU inventory, health, power and thermal monitoring in Intersight.

- **Cisco AI POD with Red Hat OpenShift (up to 6 nodes)**
  - Deploy Red Hat OpenShift on the AI POD nodes using the agreed installation method.
  - Install and configure the NVIDIA GPU Operator and Node Feature Discovery Operator.
  - Configure storage classes, projects/namespaces, RBAC and GPU resource quotas.
  - Validate GPU scheduling with a customer-provided or reference inference endpoint.

- **Cisco AI Defense: Policy, Deployment and Red Teaming (up to 10 policies)**
  - Policy: define AI Defense policies and guardrails (prompt injection, jailbreak, sensitive data/PII leakage, harmful content) aligned to the customer use cases.
  - Deployment: integrate runtime protection with the agreed AI application using the supported pattern (inspection API, gateway or connector).
  - Red Teaming: run AI Defense red teaming against the selected model or application endpoint and review the risk report.
  - Validate allow/block verdicts and alerting with agreed test prompts.

- **Isovalent: Policy, Service Graph and Flow Graphs (up to 10 policies)**
  - Deploy Isovalent Enterprise for Cilium on OpenShift, with Hubble and Tetragon.
  - Policy: create identity-based (L3-L7) network policies for AI namespaces, inference services and egress to external model/API endpoints.

  - Service Graph: use the Hubble service map to visualize service dependencies across the AI application.
  - Flow Graphs: use Hubble flow visibility to validate allowed and dropped flows and policy verdicts.
  - Enable Tetragon runtime security events for selected workloads.

- **Splunk Observability Cloud**
  - Deploy the Splunk OpenTelemetry Collector (agent and cluster receiver) on OpenShift and send host, Kubernetes and GPU (NVIDIA DCGM) metrics to Splunk Observability Cloud, where licensed.
  - Instrument the customer's AI application or inference endpoint for APM tracing, including the AI Defense guardrail call where supported.
  - Collect Isovalent (Hubble) network metrics where exposed.
  - Build dashboards and detectors with email alerting for GPU health, node and pod health, and AI application latency and errors.

- **Test and Validation**
  - Execute agreed tests across Intersight, OpenShift, AI Defense, Isovalent and Splunk.
  - Demonstrate an end-to-end request flow: user prompt, AI Defense guardrail, inference service on OpenShift, observed in Hubble and as a trace in Splunk Observability Cloud.
  - Capture results, evidence, issues and remediation recommendations.

- **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.
- 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 and engagement plan • Solution overview • Confirm success criteria and gather customer requirements
2.	Pre-Requisite and Readiness Validation	• Confirm AI POD hardware readiness • License and subscription validation • Intersight, OpenShift and SaaS connectivity readiness
3.	AI Model Sizing	• Capture model and workload requirements • Estimate GPU memory and GPU count per model • Map requirements to AI POD configuration • Model sizing summary
4.	Cisco Intersight – (up to 6 nodes)	• Claim GPU servers into Intersight • Server policies, profile templates and profiles • Firmware and HCL compliance • GPU health, power and thermal monitoring
5.	Cisco AI POD with Red Hat OpenShift - (up to 6 nodes)	• Deploy Red Hat OpenShift on AI POD nodes • NVIDIA GPU Operator configuration • Namespaces, RBAC and GPU quotas • GPU scheduling validation
6.	Cisco AI Defense - (up to 10 policies)	• AI Defense policies and guardrails • Runtime protection deployment for the agreed AI application • Red teaming run and risk report review • Verdict and alert validation
7.	Isovalent - (up to 10 policies)	• Isovalent Enterprise for Cilium, Hubble and Tetragon deployment • Identity-based network policies • Service graph (Hubble service map) • Flow graphs and policy verdicts • Tetragon runtime events

No	Deliverable	Service Details
8.	Splunk Observability Cloud	• OpenShift, Kubernetes and GPU telemetry via OpenTelemetry • APM tracing of one AI application • Dashboards and detectors
9.	Test and Validation	• Execution of agreed test cases • End-to-end AI request flow demonstration • Test results report
10.	Knowledge Transfer	• As-built configuration document • Explain how to configure, operate and troubleshoot the deployed solution in the customer's environment