CISCO AI FACTORY MENTORED INSTALL

Service Overview:

At Netnology, we specialize in Cisco AI FACTORY enablement and implementation services to expedite the adoption of AI solutions. As part of this Mentored Install service offer, our subject matter experts (SME) partner with your team to ensure a smooth deployment and provide knowledge transfer to equip your staff with the necessary skills to configure and manage the AI-Optimized Cisco Fabric effectively.

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. Cisco AI PODs serve as the building blocks of the Secure AI Factory architecture. Workload PODs provide the infrastructure required to run AI workloads such as model training, optimization, and inference, while Services PODs deliver shared capabilities including security, observability, and data services. Together, these modular components enable enterprises to build a scalable AI infrastructure that can support multiple AI workloads while maintaining centralized control, monitoring, and security across the environment.

Service Benefits:

Netnology has a team of world class engineers who specialize in AI Solution on High-Speed Cisco Infrastructure and are passionate about customer success. This Mentored Install engagement will provide information on how to deploy, configure and integrate AI solution on Cisco Infrastructure.

Service Scope:

As part of the 10-day (up to 80 hours) engagement, Netnology will provide the following services:

- Cisco AI FACTORY
  - Configure Network Fabric, Security and Observability Infrastructure
    - Design a lossless fabric
    - Configure Underlay and Overlay
    - Set Up VLANS/VTEPs
    - Configure RoCEv2

- Cisco Secure Firewall, Hypersheild, Cisco AI Defense, Splunk

- Configure Compute infrastructure (UCS-X)
  - Initialize FI in IMM mode
  - Discover FI/Chassis/Blades
  - IP Pool Policy
  - Basic Domain Profile Setup
  - Port Policy, VLANS/vSAN, Network Connectivity policy
  - Server Profile Templates / Server Profiles
    - BIOS / Boot / LAN
    - vNIC / vHBA

- Operating System Installation on Nodes
  - Determine the compatible OS for AI workload (Red Hat).
  - Create an Image repository.
  - Perform OS installation, including setting up disk partitions and configuring basic network settings.
  - Install and configure required drivers (e.g., GPU drivers).
  - Configure firewall rules for security.

- AI Workload Deployment (LLAMA 2/LLAMA 3)
  - Obtain the LLAMA 2/3 model files and required dependencies from the official repository or source.
  - Create a virtual environment using venv or conda.
  - Install the necessary Python packages, such as transformers, torch, or other libraries.
  - Create a code to load the LLAMA 2/3 model into memory.

- Configure Data Pipeline and Inference
  - Set up the data preprocessing pipeline, including tokenization, normalization, and batching.
  - Develop or configure an inference script.

- Test and Validation
  - Solution validation by running sample inputs.

- Knowledge Transfer

Target Audience:

This service is designed for Network Architects, Network Engineers and Administrators configuring, deploying, and managing the AI Infrastructure.

Prerequisites:

- Basic knowledge of Cisco Nexus, Cisco UCS, Large Language Models (LLMs).
- Customers need to ensure that all equipment and devices are racked and stacked, cabled, and powered up prior to the kick-off. Customer also needs to acquire the necessary software licenses for the deployment of the infrastructure.

Service Deliverables:

No	Deliverable	Service Details
1.	Project Kickoff	• Project Overview • Solution Overview • Gather Customer requirements
2.	Pre-Requisite Validation	• Review/Confirm Hardware Readiness (Server and Network) • License Validation • Network Readiness (Bandwidth, latency and redundancy) • Intersight Readiness
3.	High Level Design	• Develop High Level Design (HLD) Document
4.	Configure Network Fabric Infrastructure	• Design a lossless fabric • Configure Underlay and Overlay • Set Up VLANS/VTEPs • Configure RoCEv2 • Configure Network and Security Policies • Cisco Secure Firewall, Hypersheild, Cisco AI Defense, Splunk

5.	Configure Compute infrastructure	• Initialize FI in IMM mode
		  ○ Discover FI/Chassis/Blades
		• IP Pool Policy
		• Basic Domain Profile Setup
		• Port Policy, VLANS/vSAN, Network Connectivity policy
		• Server Profile Templates / Server Profiles
		  ○ BIOS / Boot / LAN
		  ○ vNIC / vHBA

6.	Operating System Installation on Nodes	• Determine the compatible OS for AI workload (Ubuntu, Red hat) • Create an Image repository. • Perform OS installation, including setting up disk partitions and configuring basic network settings. • Install and configure required drivers (e.g., GPU drivers). • Configure firewall rules for security.
7.	AI Workload Deployment (LLAMA 2/LLAMA 3)	• Obtain the LLAMA 2/3 model files and required dependencies from the official repository or source. • Create a virtual environment using venv or conda. • Install the necessary Python packages, such as transformers, torch, or other libraries. • Create a code to load the LLAMA 2/3 model into memory.
8.	Configure Data Pipeline and Inference	• Set up the data preprocessing pipeline, including tokenization, normalization, and batching. • Develop or configure an inference script.
9.	Test and Validation	• Solution validation by running sample inputs
10.	Knowledge Transfer	• Explain how to configure and manage the deployed solution in the respective customers' environment.