

Accelerate Agentic AI with Cisco, NVIDIA, and VAST Data



Value statement

What if your AI infrastructure unlocked the true potential of your enterprise data? With Cisco, NVIDIA, and VAST Data, you can accelerate retrieval-augmented generation (RAG) pipelines, streamline data movement, and scale agentic and physical AI across your organization. Together, we provide a validated, enterprise-class AI data platform—so you can move beyond experimentation and deliver real business outcomes with AI.

Benefits

- **Accelerate RAG pipelines:** VAST InsightEngine and NVIDIA AI Data Platform streamline data retrieval to feed large language models with enterprise knowledge.
- **Unlock agentic and physical AI:** AI agents are able to operate autonomously with fast, accurate, and contextual data pipelines.
- **Get a fully validated architecture:** The platform is built on Cisco AI PODs, the foundation of Cisco Secure AI Factory with NVIDIA, ensuring reliability and scalability.
- **Unify your infrastructure:** Cisco UCS, Cisco Nexus® Hyperfabric networking, and VAST InsightEngine integrate into a seamless, enterprise-class AI fabric.
- **Get enterprise-ready security and scale:** The platform is built with Cisco’s Secure AI Factory principles to ensure governance, resilience, and compliance. Cisco AI Defense secures AI models and applications, Cisco Hybrid Mesh Firewall with Isovalent® protects containerized workloads, NVIDIA BlueField DPUs offload security policy enforcement to preserve GPU and CPU resources, and Splunk® Observability Cloud delivers real-time AI infrastructure and agent monitoring with integrated threat detection through Splunk Enterprise Security.

Overview

Enterprises are shifting from experimenting with AI to deploying real, production-scale agentic and physical AI systems. But success requires solving one of the hardest problems: making the right data available at the right time. Legacy storage architectures, siloed compute, and slow retrieval mechanisms limit the effectiveness of RAG and generative AI models.

Cisco, NVIDIA, and VAST Data have partnered to deliver a validated AI data platform designed specifically for enterprise-scale AI. Built on Cisco UCS® servers and Cisco AI PODs, integrated with VAST InsightEngine, and powered by NVIDIA AI Enterprise and designed by NVIDIA AI Data Platform, this solution forms the foundation of the Cisco Secure AI Factory with NVIDIA. It is the first enterprise architecture that unifies compute, fabric, and storage into a single, validated platform to accelerate RAG, retrieval, and agentic and physical AI workflows at scale.

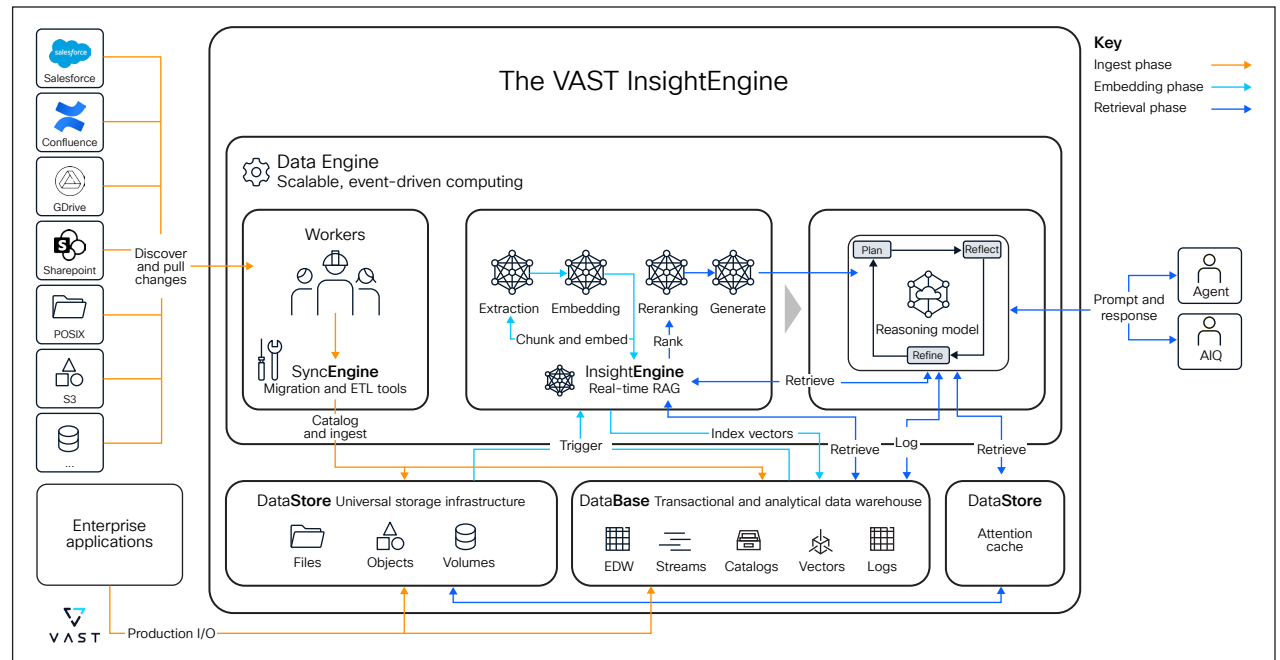


Figure 1. VAST InsightEngine

AI is moving from models to agents

The shift from LLM experimentation to **agentic AI systems** is underway. Gartner predicts that by 2026, **over 30 percent of enterprises will deploy AI agents** to drive automation and decision-making. But to succeed, enterprises must solve the **data problem**—how to make vast, unstructured, and distributed data immediately useful to AI systems.

Key challenges include:

- **Data retrieval bottlenecks:** Traditional storage cannot meet AI inference latency requirements.
- **Fragmented architectures:** Disconnected storage, compute, and fabric introduce inefficiency.
- **Scalability risks:** AI pipelines break down as workloads and data sizes expand.
- **Security and observability gaps:** Hostile AI agents use external tools, ingest diverse data, and act independently, creating new attack surfaces including prompt injection, model poisoning, data leaks, and unauthorized access that require integrated security and real-time observability across the entire AI stack.

Cisco, NVIDIA, and VAST Data together eliminate these barriers by **integrating compute, storage, and AI frameworks into a validated AI data platform** that accelerates enterprise agentic and physical AI from core to edge.

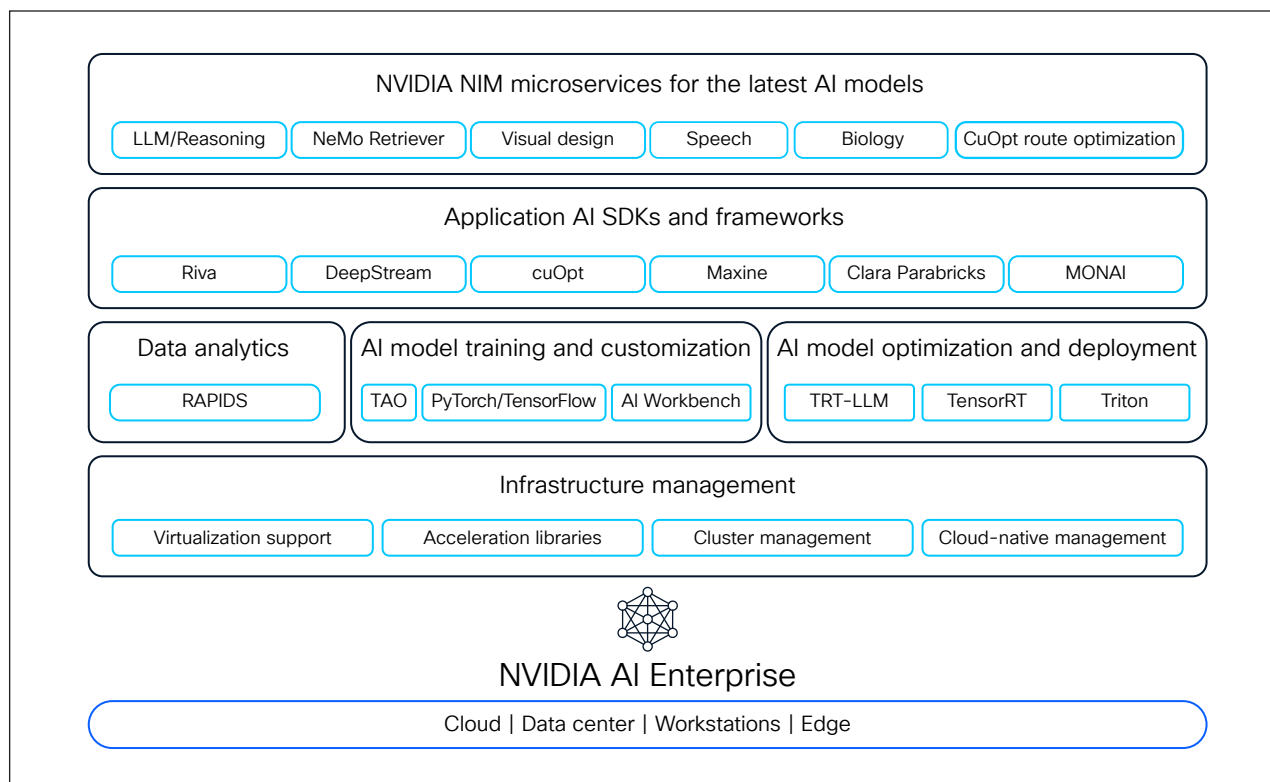


Figure 2. NVIDIA AI Enterprise – AI stack for the AI data platform

“Cisco, NVIDIA, and VAST Data deliver one of the first NVIDIA AI Data Platform validated designs that accelerates RAG and unlocks agentic AI for the enterprise.”

“With VAST InsightEngine and NVIDIA AI Data Platform running on Cisco AI PODs, enterprises can finally bring AI agents to life at scale.”

What you buy

The AI data platform is a part of Cisco Secure AI Factory with NVIDIA, providing a framework for governance, lifecycle management, and automation.

Cisco AI PODs with AI servers and networking as the validated compute and networking building blocks for NVIDIA AI Data Platform:

- **RTX PRO Server from Cisco**
(Cisco UCS C845A M8 Rack Servers, NVIDIA BlueField equipped)
- **VAST InsightEngine**
Provides a unified data engine optimized for AI retrieval and pipelines.

VAST Accelerated Core SW License:
 - Includes VAST OS and VAST InsightEngine services
 - Is fully supported and managed by VAST
 - Includes VAST container orchestration (Rancher)
- **Networking**
Cisco N9000 Series Switches with NVIDIA Spectrum-X Ethernet Technology

- **Storage and data platform**
 - VAST Data Platform on Cisco UCS (Cisco Ebox, Cisco UCS C225A M8 All-NVMe Node, NVIDIA BlueField equipped)
 - VAST core SW license:
 - Includes VAST InsightEngine and all **VAST DataBase and VAST DataEngine** functionalities
 - Fully supported and managed by VAST
 - VAST capacity software license:
 - Includes VAST OS and covers **VAST DataStore** capabilities (including NFS, S3, SMB, and NVMe/TCP)
 - **NVIDIA AI Enterprise** for AI model training, inference, and data pipeline orchestration

Key capabilities

- **Enterprise RAG acceleration:** Seamlessly retrieve and enrich enterprise data to fuel generative AI with accurate knowledge.
- **Agentic AI enablement:** Provide AI agents with high-performance data retrieval and orchestration for autonomous workflows.

- **Validated designs:** NVIDIA AI Data Platform is the validated design that connects Cisco AI PODs and VAST InsightEngine.
- **Cloud-managed Fabric:** Get Cisco Nexus Hyperfabric to ensure integrated, secure, and automated networking.
- **Security-first architecture with observability:** Cisco AI Defense protects AI models and applications; Cisco Hybrid Mesh Firewall with Isovalent secures containerized workloads; NVIDIA BlueField DPUs offload security enforcement; and Splunk Observability Cloud monitors AI infrastructure health, agent performance, and security threats across the entire stack.

Models and options

- **AI data platform reference design:** The first NVIDIA AI Data Platform reference built on VAST InsightEngine and Cisco AI PODs.
- **Turnkey AI PODs:** Cisco Secure AI Factory PODs with full-stack integration of compute, networking, and storage.

Services

Use Cisco services to operationalize your AI data platform

Cisco services can help enterprises plan, deploy, and manage VAST InsightEngine on Cisco AI PODs with NVIDIA AI Data Platform. From readiness assessments to validated design deployment, Cisco experts accelerate adoption while reducing complexity. Lifecycle services ensure that your AI infrastructure evolves with your business needs.

Cisco, NVIDIA, and VAST offer a validated solution designed to enable faster data extraction and retrieval for agentic AI workflows.

VAST InsightEngine is one of the first storage solutions to offer an **NVIDIA AI Data Platform reference design** built on Cisco AI PODs, the AI Infrastructure building block of Cisco Secure AI Factory.

Cisco Secure AI Factory with NVIDIA provides validated architecture to speed enterprise AI adoption, no matter the use case.

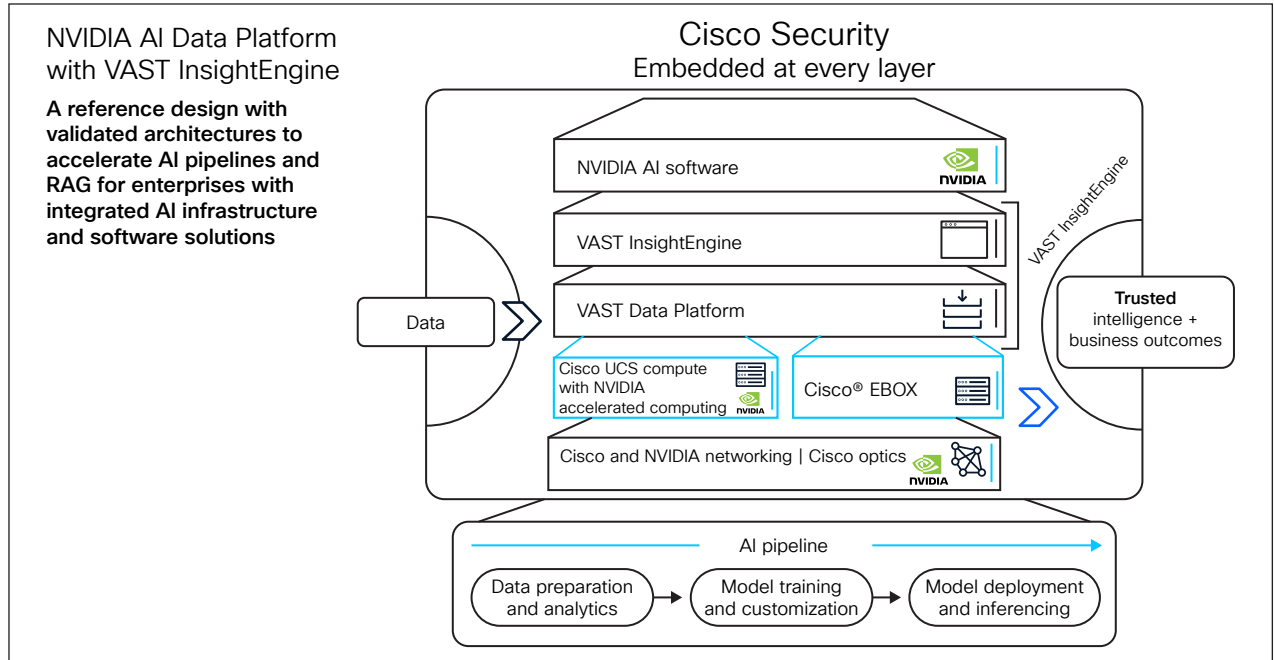


Figure 3. Cisco AI POD architecture supporting NVIDIA AI Data Platform

Table 1. Use cases

Industry	Use case
Financial services	Accelerate fraud detection with agentic AI models retrieving real-time transaction data.
Healthcare	Empower clinical assistants with AI agents accessing genomic and imaging datasets.
Retail and e-commerce	Use RAG pipelines to deliver hyper-personalized recommendations at scale.
Manufacturing	Drive predictive maintenance and AI-guided automation using agentic AI with sensor data.
Enterprise IT	Deploy enterprise-wide AI assistants that retrieve knowledge from documents, policies, and systems.



Learn more

Unlock the power of agentic and physical AI with Cisco, NVIDIA, and VAST Data.

Learn more about Cisco Secure AI Factory with NVIDIA at: <https://www.cisco.com/site/us/en/solutions/artificial-intelligence/index.html>.

Cisco Capital

Flexible payment solutions to help you achieve your objectives

Cisco Capital makes it easier to get the right technology to achieve your objectives, enable business transformation and help you stay competitive. We can help you reduce the total cost of ownership, conserve capital, and accelerate growth. In more than 100 countries, our flexible payment solutions can help you acquire hardware, software, services and complementary third-party equipment in easy, predictable payments. [Learn more.](#)

The Cisco Advantage

Cisco is uniquely positioned to operationalize agentic AI at scale. With **Cisco Secure AI Factory with NVIDIA**, enterprises gain validated AI infrastructure integrated with **VAST InsightEngine** and **NVIDIA AI Data Platform**. Only Cisco offers the end-to-end stack—from compute and fabric to storage and lifecycle services—validated with industry leaders to accelerate AI adoption securely and at scale. Unlike other AI factories, Cisco embeds security and observability at every layer: Cisco AI Defense for model and application protection, Cisco Hybrid Mesh Firewall for workload security, Splunk Observability Cloud for real-time AI infrastructure and agent monitoring, and Splunk Enterprise Security for threat detection. This security-first approach helps to ensure that enterprises can deploy trusted agentic and physical AI applications from core to edge with confidence.