Cisco AI PODs for VAST InsightEngine with NVIDIA AI Data Platform



© 2026 Cisco and/or its affiliates. All rights reserved.

Benefits

- **Accelerate AI Insights:** Speed Retrieval-Augmented Generation (RAG) pipelines and unlock the full potential of agentic and physical AI with near-real-time data access.
- **Validated and Unified Infrastructure:** Deploy a pre-tested, scalable architecture that seamlessly integrates compute, networking, and storage for optimal AI performance.
- **Enterprise-Grade Security and Governance:** Ensure data integrity and compliance with built-in policy-based access, audit readiness, and Cisco's Zero-Trust AI infrastructure. The Cisco Secure AI Factory embeds security at every layer: Cisco AI Defense protects AI models and applications, Cisco Hybrid Mesh Firewall with LiveProtect® secures containerized workloads, NVIDIA BlueField DPUs offload security enforcement to preserve GPU resources, and Splunk® Observability Cloud provides real-time AI infrastructure and agent monitoring with integrated threat detection through Splunk Enterprise Security.
- **Flexible Deployment and Scaling:** Support diverse AI workloads and grow seamlessly from experimentation to production with modular, cloud-managed AI PODs. Choose from modular pre-validated AI PODs backed by CVDs, turnkey Cisco Nexus Hyperfabric™-managed stacks, or build your own secure AI factory over time with individual components.

© 2026 Cisco and/or its affiliates. All rights reserved.

Accelerate enterprise RAG with real-time data Insights

Enterprises are rapidly moving beyond AI experimentation to deploying production-scale agentic and physical AI systems. Success hinges on making the right data available at the right time, a challenge for legacy infrastructures. The NVIDIA AI Data Platform, powered by Cisco and VAST InsightEngine, provides a validated, enterprise-class foundation designed specifically for this shift. As a core data layer within the Cisco Secure AI Factory with NVIDIA, this solution is backed by a security-first architecture with end-to-end observability, enabling enterprises to operationalize trusted AI from core to edge.

Built on Cisco UCS® (including Cisco Unified Communications for RTX PRO Server for compute and Cisco UCS C225 M8 for VAST EBOX storage), integrated with VAST Data AI OS, and powered by NVIDIA AI Enterprise, this solution unifies compute, fabric, and storage into a single, secure platform. It accelerates RAG pipelines, streamlines data movement, and enables agentic and physical AI workflows at scale, moving your organization beyond experimentation to real business outcomes.

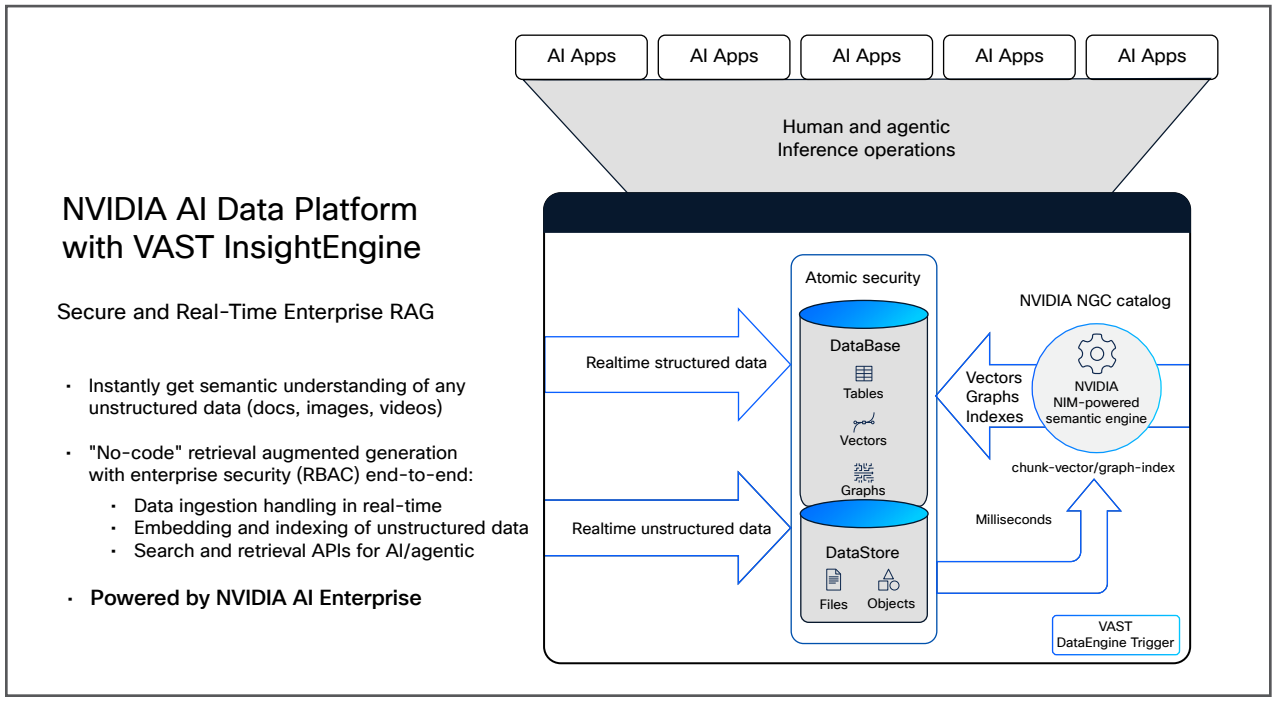


Figure 1. VAST InsightEngine architecture Framework for NVIDIA AI Data Platform

What it does

VAST Data Platform acts as an AI-native data engine that eliminates traditional data preparation bottlenecks. Its **InsightEngine** processes data in real time as it's ingested, instantly creating vector embeddings and establishing graph relationships.

This makes data immediately available and consumable for AI tasks such as Retrieval-Augmented Generation (RAG), enabling AI agents to access and reason over vast, diverse, and multimodal enterprise data without latency.

It provides a "single source of truth" for AI, supporting complex AI workloads and multimodal AI agents by transforming raw data into instant AI-ready datasets.

VAST Data Platform: The AI-native data foundation

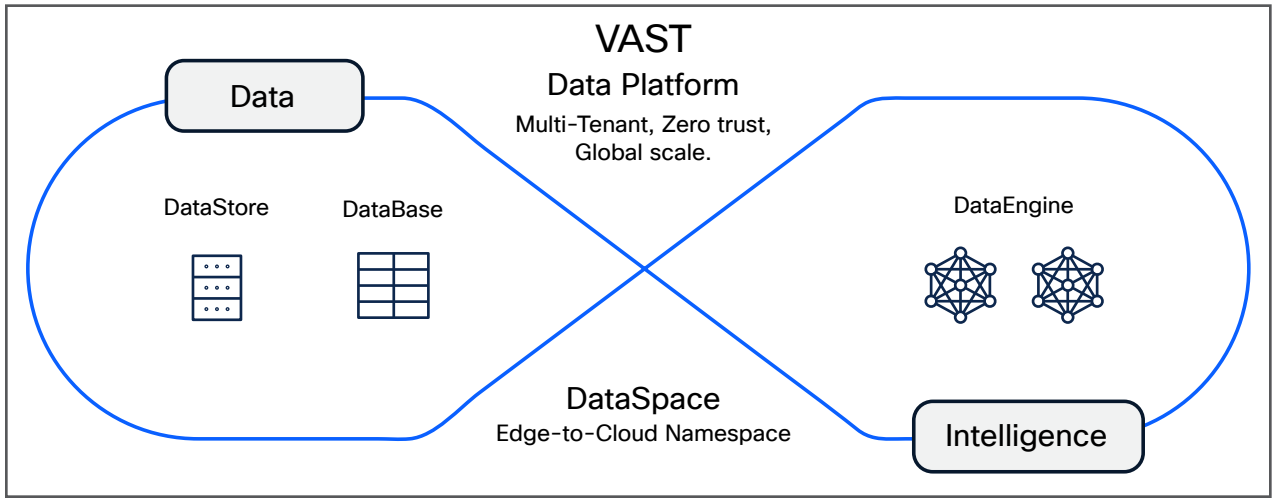


Figure 2. VAST Data Platform, foundation for agentic AI

The VAST Data Platform, particularly its AI Operating System (AI OS) with InsightEngine, represents a groundbreaking shift in enterprise data management, purpose-built for the demands of modern AI. It unifies all enterprise data types—files, objects, tables, and streams—into a single, scalable semantic database. Built on VAST's innovative DASE (Disaggregated, Shared-Everything) architecture, it delivers linear scalability, exceptional performance, and unparalleled efficiency.

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

© 2026 Cisco and/or its affiliates. All rights reserved.

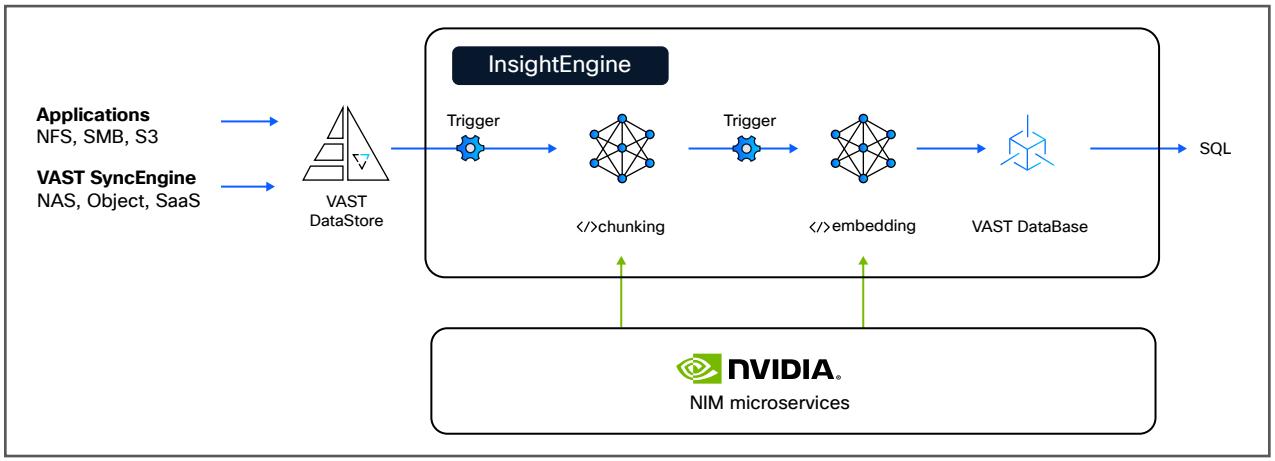


Figure 3. VAST AI OS and InsightEngine framework

Cisco AI Data Platform with VAST AI OS makes enterprise data continuously usable for AI by unifying files, objects, tables, streams, metadata, vectors and events in one software platform. Instead of copying data across storage tiers, ETL jobs, standalone vector stores, and external indexing services, VAST keeps information live, governed, and ready for retrieval, inference, analytics, and agents. VAST InsightEngine extends that foundation by continuously classifying content, extracting context, generating embeddings, and turning enterprise knowledge into searchable, AI-ready signals.

Benefits of running on Cisco C225 M8 Servers with Cisco Intersight management

Integrating the VAST Data Platform with Cisco UCS C225 M8 servers, managed by Cisco Intersight®, offers significant advantages for enterprise AI initiatives:

- **Validated Performance and Reliability:** The Cisco UCS C225 M8 servers are specifically designated as the “EBOX” (storage servers) for the VAST Data Platform within the Cisco AI POD architecture. This pre-validated integration ensures optimal performance, high throughput, and low latency for VAST’s data-intensive operations, delivering the reliability required for critical AI pipelines.
- **Seamless Integration into AI PODs:** The C225 M8 servers seamlessly integrate into the broader Cisco AI POD framework, which combines compute (NVIDIA RTX PRO Servers), networking (Cisco Nexus Hyperfabric), and storage into a cohesive, pre-tested, and scalable solution. This simplifies deployment and ensures predictable performance across the entire AI infrastructure.
- **Simplified Operations with Cisco Intersight:** Managing the VAST Data Platform on Cisco C225 M8 servers is streamlined through Cisco Intersight. This unified cloud management platform provides a single pane of glass for automated provisioning, monitoring, and lifecycle management of the entire Cisco UCS infrastructure. Intersight reduces operational overhead, minimizes complexity, and allows IT teams to focus on AI innovation rather than infrastructure management.
- **Enterprise-Grade Foundation:** Leveraging Cisco’s robust server infrastructure ensures the stability, security, and uptime necessary for demanding enterprise AI workloads, providing a solid foundation for your AI factory.

© 2026 Cisco and/or its affiliates. All rights reserved.

Learn more

Accelerate your enterprise AI journey

Unlock the full power of agentic and physical AI with the validated NVIDIA AI Data Platform on Cisco and VAST Data. Learn more about how Cisco Secure AI Factory with NVIDIA can accelerate your enterprise AI adoption securely and at scale. With security-first design, end-to-end observability powered by Splunk, and flexible deployment options from modular Cisco AI PODs to turnkey Cisco Nexus Hyperfabric stacks, the Cisco Secure AI Factory removes infrastructure and security as barriers to AI innovation.

Read VAST Data on Cisco UCS Data Sheet:

Transform data into actionable AI intelligence

This integrated solution transforms enterprise storage into an AI-ready knowledge retrieval system, making massive, unstructured data instantly usable by AI. It delivers high-performance data retrieval and orchestration for autonomous workflows, fueling generative AI with accurate, contextual knowledge. Our unique approach leverages the RTX PRO Server from Cisco, UCS C845A M8, for powerful compute and Cisco UCS C225 M8 servers as EBOX storage nodes running VAST Data Platform, all managed by Cisco Intersight for simplified operations.

Unlike GPU-focused AI stacks that lack integrated data intelligence, we provide a complete Data + Compute + Networking reference design. While storage-only solutions offer limited compute integration, our platform deeply integrates NVIDIA AI Enterprise on Cisco AI PODs. For those concerned about cloud-only offerings' security and governance challenges, we offer enterprise control with hybrid flexibility. This solution supports SDS Modernization, serves as a robust storage solution for Cisco AI PODs, and offers a turnkey approach with cloud-managed networking, integrated PODs, and automated deployment through Cisco Hyperfabric. Built within the Cisco Secure AI Factory with NVIDIA framework, the solution benefits from security engineered at every layer of the stack, including Cisco AI Defense for AI model and application protection, end-to-end observability through Splunk Observability Cloud for monitoring AI infrastructure health and agent performance, and Cisco's high-performance Silicon One® and NVIDIA Spectrum-X based networking for congestion-free, lossless AI connectivity.

© 2026 Cisco and/or its affiliates. All rights reserved. Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

C45-5415816-01 05/26