Introduction

NVIDIA is helping enterprises build AI factories that are cost effective, scalable and high-performing — equipping them to meet the next industrial revolution. AI factory solutions are becoming available through NVIDIA's global ecosystem of enterprise partners with NVIDIA-Certified Systems, NVIDIA-Certified Storage, & NVIDIA Networking. These partners offer top-tier hardware, software and data center expertise to mitigate risks and enhance ROI in AI projects. This whitepaper presents the necessary components, including integrations from our ecosystem partners, automation tools, and deployment strategies. This design can be used by our enterprise partners for integrating accelerated computing, high-performance networking, and AI software for successfully building single tenant enterprise ready AI factories.

Terms and Definitions

Control Plane	The container orchestration layer that exposes the API and interfaces to define, deploy and manage of containers.
Worker Host	A bare metal server that is building the foundation for the control plane as well as worker hosts.
Worker Node	A compute resource on a physical host and is the resource that is used by AI Developers to execute workloads.
Cluster	A set of worker nodes that run containerized applications. Every cluster has at least one worker noc
AI Agent	A system of LLMs that work together to reason about a problem with data and act upon it.

Scope

The scope of this white paper is to provide guidance for building AI Factories. Architectural best practices leveraging ecosystem partners along with NVIDIA hardware and software are described to provide a starting point for building AI Factories. A broad range of ecosystem partners that provide enterprise commercial offerings are presented.

User Personas

The following user personas will interact with, administer, or use a system implemented by this design guide.

Ecosystem Partners	This is a third-party partner of hardware, software or system integrator that is NVIDIA certif integrated into the AI factory reference architecture.
OEMs	Server OEM (Original Equipment Manufacturer) companies that design, manufacture and se hardware and components.
ISVs	Independent Software Vendor (ISV) that develops and sells software products.
AI Developer	A person who will use the system and work on implementing AI-driven features in applicatio integrating LLMs and writing code to deploy AI functionality in software.

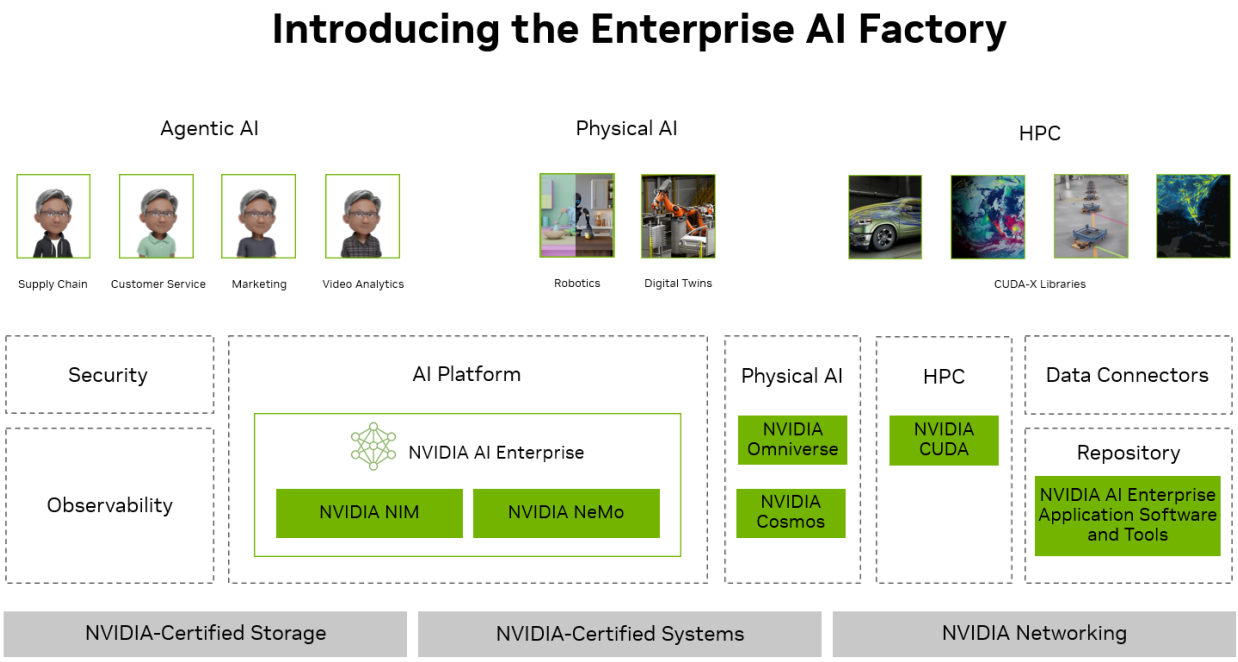
MLOps	A person who will use the system and seeks to increase automation and improve the quality o AI/ML lifecycle from development to deployment to monitoring.
IT Admin	A person who will use the system and helps design, procure and manage the underlying infrastructure, including server, operating systems and storage.
Network Administrator	A person who will use the system and manage network infrastructure, including router, switc firewalls and security protocols.

Platform Requirements

- Data centers with sufficient power, cooling, space, and network connectivity for high-density GPU systems.
  - Refer to the NVIDIA HGX B200 8-GPU and NVIDIA RTX(™) Pro Server Edition Sections of the NVIDIA Enterprise Reference Architecture
- Skilled personnel or trained system integrator
  - Base hardware installation and software platform provisioning
  - Support for ongoing operation and management

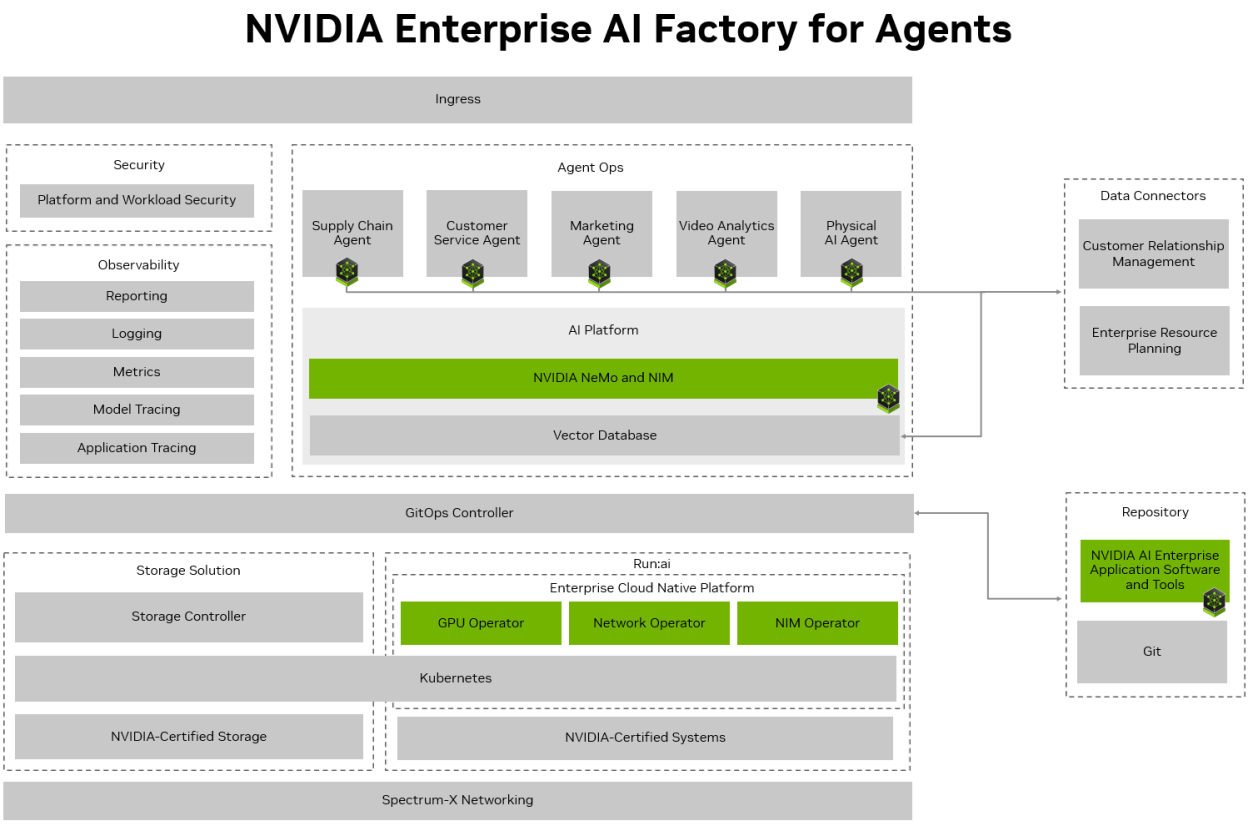
Enterprise AI Factory Overview

The design for an Enterprise AI Factory integrates seamlessly with enterprise systems, data sources, and security infrastructure through NVIDIA's partner solutions. It utilizes advanced NVIDIA hardware in conjunction with software tools and solutions from NVIDIA AI Enterprise to ensure optimal performance adhering to best practices. An ecosystem of partners is included, with specified versions and integrations, providing a comprehensive, high-performance solution for modern AI projects, including Physical AI & HPC with a focus on Agentic AI workloads. It provides a standardized on-premises platform so that Enterprise IT stakeholders can work with their chosen vendors. The following is a high-level description of the design components.



Agentic AI in the Factory

The following sections describe the implementation of Agentic AI within an Enterprise AI Factory. Information is provided for each of the Enterprise AI Factory components (AI Platform, Data Connectors, Artifact Repository, Observability, Security, & Hardware) in relation to Agentic AI.



Enterprise Cloud Native Platform

NVIDIA-Certified systems offer a reliable platform for deploying high-performance accelerated software, including NVIDIA AI Enterprise. They ensure consistent performance and allow enterprises to deploy optimized platforms for AI, Data Analytics, HPC, high-density VDI, and other data center workloads.

The Enterprise Cloud Native Platform, with Kubernetes at its core, provides agility, scalability, and resilience for an Enterprise AI Factory focused on developing and deploying sophisticated AI agents. Kubernetes embodies cloud-native principles by orchestrating containers (like those from NVIDIA AI Enterprise), managing microservice-based agent architectures, and enabling dynamic automation. This includes automated deployment of new agent versions, scaling based on demand (important for both training and inference on NVIDIA-Certified Systems), self-healing to support high availability, and resource management, particularly for GPU resources.

These cloud-native capabilities are relevant for an AI Factory. The ability to independently develop, update, and scale microservice-based agents, coupled with automated CI/CD pipelines managed via Kubernetes, allows for iteration and deployment. Kubernetes handles the significant and often burstable compute demands for training AI models and scales inference services for deployed agents based on real-time needs. This automation and resource packing on NVIDIA-Certified Systems also contribute to reducing operational burden and optimizing costs, which is a

consideration when dealing with complex AI/ML environments, especially those involving GPUs.

In this context, Kubernetes functions as a foundational platform for the AI Factory. It unifies the management of a complex stack—including NVIDIA Operators, AI software suites like NVIDIA AI Enterprise, storage, networking, and observability tools—onto a single platform. Enterprise Kubernetes distributions and validated architectures can further simplify this by providing secure, supported, and pre-integrated environments. This orchestration supports the process of efficiently building, deploying, and managing a diverse and evolving suite of AI agents on high-performance infrastructure.

Storage Solution

A key component of the Enterprise AI Factory is its storage solution. Given the continuous and high-volume data flow required throughout the AI development and deployment lifecycle, the storage infrastructure can potentially become a significant bottleneck if not architected correctly to meet these intensive demands. Therefore, the solution must possess several essential features to handle demanding AI environments. These include the scalability to manage exponentially growing datasets and model sizes, and the flexibility to support diverse data types and access patterns, ranging from high-throughput sequential reads for training to low-latency random access for inference and vector databases. Furthermore, robust data protection mechanisms like snapshots, replication, and disaster recovery are critical, as are comprehensive security features such as encryption at rest and in transit to safeguard sensitive information.

To meet these diverse workload requirements, the Enterprise AI Factory's storage solution utilizes a tiered storage architecture from various vendors. A crucial element of this architecture is NVIDIA-Certified Storage, which adheres to stringent performance and reliability standards specifically for AI tasks. This certification ensures efficient data access, which is vital for handling large model weights, managing Vector Database I/O for Retrieval Augmented Generation (RAG), and supporting knowledgebases for AI Agents. Having been vetted for these crucial characteristics, the certified storage provides a dependable, high-performance, secure, and scalable infrastructure, thereby enhancing the overall efficiency and stability of the AI Factory. This empowers partners and customers to build AI factories that efficiently leverage massive amounts of data, leading to faster, more accurate, and reliable AI models.

The NVIDIA-Certified Storage program offers two levels of certification: Foundation and Enterprise. These storage certifications integrate seamlessly with corresponding NVIDIA Enterprise RAs to ensure that storage systems possess the necessary performance to support North-South networking and effectively feed data to compute nodes. The Foundation-level storage certification certifies storage partners for PCIe-optimized reference configurations specifically for the NVIDIA RTX PRO 6000 Blackwell Server Edition. The larger-scale Enterprise-level storage certification validates storage partners for HGX reference configurations, particularly for the NVIDIA HGX B200.

Artifact Repository

The AI Factory incorporates a dedicated artifact repository designed to handle software components, especially for on-premises setups that follow GitOps principles. This repository serves as a secure, version-controlled local hub for essential NVIDIA AI Enterprise artifacts, such as containerized NVIDIA NIM microservices, AI models, libraries, and tools. In the GitOps workflow, git maintains the declarative state by linking to specific versions of these NVIDIA artifacts stored in the repository. Then, the GitOps Controller fetches these verified artifacts for deployment onto the Kubernetes platform. For on-premises environments, managing local artifacts enables essential operational practices such as scanning NVIDIA containers and other artifacts for security vulnerabilities, ensuring reliable and rapid access without relying on public registries, managing dependencies, and assuring reproducible deployments using specific, approved versions of NIMs and AI models.

GitOps Controller

A GitOps controller is a software component that continuously monitors the desired state of infrastructure and application configurations stored in a Git repository and ensures that the actual state of a system, such as a Kubernetes cluster, matches this declared state. Working in close collaboration with the artifact repository, it operates by regularly comparing the live state of resources in the environment with the version-controlled configurations in Git. If any differences are detected, the controller automatically reconciles them by applying the necessary changes to bring the system back in sync with what is defined in the repository. This approach leverages Git as the single source of truth, enabling automated, auditable, and reproducible deployments, and is typically implemented as a Kubernetes controller that runs a reconciliation loop to maintain consistency between Git and the cluster.

Observability

Ensuring the reliability, performance, and trustworthiness of an AI Agent Platform requires a robust observability strategy. Observability provides deep insights into the system's state and behavior, enabling teams to proactively identify issues, debug complex interactions, and optimize performance. This is achieved through a combination of comprehensive logging, continuous monitoring of key metrics, detailed model tracing, thorough application tracing (often as part of distributed tracing), and consolidated reporting to understand the flow of operations and overall health of the platform.

Centralized logging forms a foundational aspect of this strategy, capturing detailed events from all layers of the platform. This includes logs from the underlying infrastructure, the container platform, core AI software components, and the AI agents themselves. These logs are invaluable, offering heuristics for debugging agent behavior, supporting security analysis by providing event records, and creating comprehensive audit trails. Such trails are essential for ensuring operational reliability, building trust in the AI system's outputs, and meeting compliance requirements.

Continuous monitoring of metrics is implemented to track both the health of the infrastructure and critical Key Performance Indicators (KPIs) specific to the AI agents. Key metrics, often collected using OpenTelemetry (OTEL) instrumentation, Application Performance Management

(APM) tools, or directly from application endpoints and infrastructure tools or NVIDIA Data Center GPU Manager (DCGM), provide a real-time view of system performance and agent effectiveness. These metrics typically fall into the following categories:

- **Latency**:
  - *Time To First Token (TTFT)*: The delay before the agent produces its initial response token after receiving a request.
  - *Tokens Per Second (TPS) / Output Throughput*: The rate at which the agent generates response tokens over time.
  - *End-to-End Latency*: The total time elapsed from the user's request to the completion of the agent's full response.
  - *Component Latency*: The duration of individual processing steps, such as:
    - Plan Generation
    - Reasoning
    - Tool Calls
    - Database Queries (including vector database lookups)
    - Retriever Calls
- **Accuracy and Faithfulness**:
  - *Task Completion Rate*: The percentage of assigned tasks that the agent successfully completes.
  - *Accuracy/Relevance*: The correctness and relevance of responses that rely on retrieved information (RAG), with specific metrics for retriever performance including precision, recall, and F1-score.
  - *Faithfulness*: How well the agent's responses adhere to the provided source information, particularly in RAG scenarios.
  - *Correctness*: The validity of outputs from individual reasoning steps or executed tools.
- **Resource Utilization**:
  - Consumption of GPU, CPU, and memory during agent operation.
- **Errors and Faults**:
  - *Fault Rate*: The frequency of errors or failures within specific agent components or workflows (e.g., plan generation, tool calls, database access).
  - *Timeout Rate*: The number of operations that exceed their allocated time limit, categorized by component (e.g., tool call timeouts, retriever timeouts).

To understand operational flows and decision-making pathways, detailed model and application tracing is important. This involves mapping the journey of requests and the sequence of actions across the AI Agent Platform's distributed services. This includes tracing interactions between agents, RAG components (specifically retriever calls), various database calls, inference endpoints, and individual tool calls an agent makes. Contemporary tracing methods, often facilitated by tools like OpenTelemetry (OTEL) and visualized in Application Performance Management (APM) systems, should capture the inputs, outputs, and duration of an agent's plan generation step, as well as the logical processes in each reasoning step, offering insight into intermediate processes where feasible. This level of detailed tracing is useful for identifying performance bottlenecks, pinpointing latency sources within complex agent interactions—such as those arising from specific tool, database, or retriever calls—and providing important context by correlating traces with other metrics and logs.

To make the collected observability data useful, a consolidated reporting mechanism is beneficial. This involves presenting aggregated data from logging, metrics, and tracing in dashboards and reports. Such presentations offer a view of the AI Agent Platform's health, performance, and accuracy, which can be tailored for different user roles, including IT operations, AI developers, and business stakeholders.

By extending the observability focus to these detailed aspects of AI agent operations—such as the specifics of planning and reasoning steps, the performance and accuracy of tool and data retrieval calls, and detailed fault and timeout analysis—enterprises can develop a more thorough understanding and improve control over their AI solutions. This increased visibility contributes to building and maintaining AI systems that are powerful, robust, efficient, and trustworthy.

Security

A robust security posture for the platform is achieved through a multi-layered strategy. This approach uses core component capabilities and integrates with enterprise security frameworks to safeguard operations and data comprehensively, from the network perimeter down to individual data elements.

The first line of defense is established at the network level. This layer employs defense-in-depth strategies, primarily utilizing network policies native to the underlying container orchestration platform. These policies control traffic flow between services and pods, isolating workloads and restricting communication to only authorized pathways, thereby minimizing the attack surface. To further bolster this, dedicated communication infrastructure enforces fine-grained traffic control policies at the application layer and automatically encrypts all traffic between services, ensuring secure and authenticated communication channels throughout the platform.

Building upon network controls, this layer focuses on verifying user and service identities and their entitlements. Authentication and authorization mechanisms are typically integrated directly with the platform's own access systems. Crucially, these are seamlessly tied to broader enterprise Identity and Access Management (IAM) solutions, such as corporate directory services. This ensures consistent identity management and allows for centralized control over user access based on established enterprise credentials and policies.

Once identity is established, access to platform resources is governed by Role-Based Access Control (RBAC). This is implemented at multiple levels:

- **Orchestration Platform RBAC**: The container orchestration platform itself employs RBAC to control permissions for managing and interacting with cluster resources (e.g., deploying applications, accessing logs, configuring services).
- **Integrated Platform RBAC**: AI/ML platforms integrated within the ecosystem also commonly feature their own RBAC systems. These ensure that access to platform-specific functionalities and resources is restricted based on predefined user or service roles.

The most granular layer of security focuses on protecting the data itself within specialized data services, including various database systems. These services are often further protected by their

internal RBAC mechanisms. These controls manage fine-grained access to data elements—such as specific data sets, tables, or collections—ensuring that read and write permissions are granted exclusively to authenticated and appropriately authorized applications and users, adhering to the principle of least privilege.

Clear roles and responsibilities are defined for effective management and operation of the complex on-premise, self-hosted AI/ML platform. This document outlines a consolidated mapping of key enterprise organizational roles to their typical access levels or administrative duties within logical groupings of platform tools. The goal is to provide a concise overview that facilitates understanding of how different teams interact with the various components of the technology stack, ensuring security, efficiency, and clear accountability in a self-managed environment.

	IT Admin	Network Administrator	AI Developer	MLOp
Enterprise cloud native platform (e.g., Kubernetes/ OpenShift, Base OS, Compute/GPU Management)	**Platform Admin & Provisioning**: Manages OS, hardware, orchestration lifecycle, compute (GPUs). Ensures stability & resource availability.	**Network Infra Mgmt**: Configures platform networking (SDN, ingress, egress), physical network. Collaborates with IT Admin on cluster networking.	**Platform User**: Accesses platform for dev tools, logs, allocated compute (GPUs) for AI tasks.	**App Deployn Ops**: Admin/ in project nam for CI/CD age apps, services Manages com inference.
Storage Solution (Core & AI-specific storage, e.g., for datasets, models, VectorDBs)	**Core Storage Admin**: Manages core storage infrastructure, backups, and base provisioning. Ensures availability for platform layers.	**Network for Storage**: Ensures reliable network connectivity and segmentation for storage systems. Troubleshoots storage network issues.	**Data Consumer**: Utilizes provisioned storage for datasets, model artifacts, and vector database access.	**Storage for A** Manages pers storage claims applications a models in pro Monitors stor performance f deployed ager
Artifact Repository (e.g., for container images, packages, models)	**Infra Support**: Provides/maintains underlying infra (OS, VMs, K8s) for the repository. Core infra install/patch.	**Network Access**: Ensures repository has necessary network access and is accessible by CI/CD tools and platform.	**Artifact User/Publisher**: Manages and versions data/model artifacts, packages, and notebooks within the repository. Pulls base images.	**Artifact Life Mgmt**: Mana CI/CD integra publishing an consuming application/m artifacts and c images.
GitOps Controller (e.g., ArgoCD)	**Infra Support**: Maintains underlying infra (OS, VMs, K8s) for the GitOps controller.	**Network Access**: Ensures GitOps controller can reach Git repositories and the Kubernetes API.	**User (Indirect)**: Benefits from GitOps for consistent environments defined	**GitOps Auto Lead**: Defines/mana application an infrastructure

	IT Admin	Network Administrator	AI Developer	MLOp
			by MLOps/Platform teams.	configuration Manages GitO controller for deployments.
Observability (Monitoring, Logging, Tracing, Reporting)	**Infra Support**: Provides/maintains underlying infra for the observability stack. Core infra install/patch.	**Network Support**: Ensures monitoring tools reach targets & telemetry flows to central systems.	**Dev/Experiment Monitoring**: Creates/views dashboards for experiments, data profiles, model dev metrics. Accesses logs for debugging.	**Prod AI Perf Monitoring**: Admin/Editor dashboards/al prod AI app performance, drift, resource lifecycle qual
Security (Endpoint, Network, Identity, Data Security)	**Infra Security & IdP Support**: Manages server security tools, IdP infra. Secures base platform. Collabs on infra firewall rules.	**Network Security Impl.**: Manages firewalls, network security policies, IDS/IPS. Configures network aspects of IdP & security tools. Collabs on security posture.	**Authenticated User**: Leverages federated identity for authorized access to tools, platforms, and data. Follows security best practices.	**Secure Deplo** Implements s CI/CD practic Manages secr deployed appl Uses federate identity for to access.
Data Connectors (e.g., to ERP, CRM, other enterprise systems)	**Infra & Network Support:** Ensures underlying infrastructure and network paths are available for data connectors.	**Network Connectivity**: Ensures secure and reliable network connectivity for data connectors to source/target systems.	**Data User**: Utilizes configured data connectors to ingest data for AI model development and RAG.	**Operational Monitoring**: the health and performance c connectors us production AI
AI Platform (e.g., AI/ML dev environments, Training/Fine-tuning services, Model Registries)	**Infra Support**: Provides/maintains underlying infra (OS, K8s, GPU access) for the AI platform components.	**Network Connectivity**: Ensures AI platform components have necessary network access for data, inter-service communication, and user access.	**Primary User**: Creates projects, prepares data, builds, trains, tunes, and registers models. Uses platform tools for experimentation.	**Model Lifecy Mgmt**: Mana CI/CD integra models from t platform to production. M resource usag platform by M tools.
Agent Ops (Deployment, management, and operation of AI Agents)	**Resource Provisioning**: Ensures sufficient compute, storage, and network resources are allocated	**Network Services for Agents**: Configures network routes, load balancing, and	**Agent Logic Developer**: Develops the core logic, AI model integration, and specific functionalities of the	**Agent Deploy Production M** Deploys, scal monitors, and manages the l of AI agents i

	IT Admin	Network Administrator	AI Developer	MLO
	for deployed AI agents.	access policies for AI agents.	AI agents. Tests agent behavior.	production. Implements C agents.

For secrets management, secure storage is provided using Kubernetes Secrets or specialized solutions that adhere to security procedures. Image security is reinforced by integrating container image scanning tools within artifact repositories, embedded within CI/CD pipelines. This process follows industry standard security gates to ensure the integrity of container images.

Endpoint and workload security are bolstered by real-time threat detection and response mechanisms configured in line with validated policies. For AI agent and model security, tools such as NVIDIA NeMo Guardrails and partner solutions are employed. These tools ensure input validation, output filtering, and secure execution, adhering to the best practices that have been internally validated.

Auditing is also a critical component, with comprehensive audit logging configured within associated applications. These logs are forwarded to Information and Event Management (SIEM) systems, following validated logging standards to ensure thorough and efficient monitoring of system activities. Together, these measures create a robust and secure environment to support AI workloads and platform services effectively.

A structured approach to patching and upgrades across the entire AI platform—including operating systems, container platforms, the AI software suite (e.g., NVIDIA AI Enterprise), and partner components—is crucial for security, stability, and performance. This requires rigorous testing, coordination with hardware and software vendors (leveraging the NVIDIA ecosystem and reference designs where applicable), and scheduled deployments to minimize operational disruption. Regular maintenance and updates ensure access to the latest features and security for AI agents.

Data Connectors

For AI agents to function effectively within the platform, they need secure access to diverse sources of enterprise data. This is achieved through connectors and API endpoints that link internal systems like customer relationship management, enterprise reporting platforms and point of sale systems. The data ingestion/retrieval system ensures security, scalability, and reliability. Ingested data is transformed into embeddings and stored in a vector database for efficient semantic searches in RAG workflows. Emerging standards, such as Model Context Protocol (MCP), aim to provide structured ways for AI agents to discover and interact with external data sources and tools. Toolkits like NVIDIA's open-source Agent Intelligence toolkit help developers build, connect and optimize the AI agents using retrieved enterprise data for complex reasoning, planning, and multi-step task execution.

AI Platform

An AI platform is an integrated suite of technologies that provides the infrastructure, tools, and services needed to build, train, customize, deploy, and manage machine learning and generative AI models and AI Agents at scale. Such platforms streamline the end-to-end AI development lifecycle, offering capabilities for data preparation, model training, fine-tuning, deployment, monitoring, and governance. The AI platform natively integrates frameworks like NVIDIA NeMo and NIM, enabling organizations to efficiently develop and customize large language models (LLMs) and other generative AI systems. NVIDIA NeMo provides a cloud-native, end-to-end framework for building, training, and deploying LLMs and other AI models, while NIM offers standardized microservices and APIs for seamless model deployment across cloud, on-premises, or edge environments. By leveraging these integrated tools, an AI platform empowers enterprises to create domain-specific AI solutions, accelerate innovation, and maintain secure, scalable, and high-performance AI operations.

Agent Ops

Enterprises can use NVIDIA AI Blueprints for accelerated development of agentic AI systems in supply chain, marketing, and customer service. These workflows combine NVIDIA NIM™ microservices with GPU components for large-scale agent deployments and include frameworks, pretrained models, Helm Charts, and Jupyter Notebooks.

The Mega Omniverse Blueprint simulates warehouse operations for supply chain optimization using physics-informed digital twins and reinforcement learning. The Digital Human Blueprint uses avatar animation, speech AI, and multimodal reasoning for virtual customer service assistants. The RAG Blueprint enhances marketing applications with a hybrid vector search and multimodal extraction pipeline for enterprise data. These Blueprints utilize NVIDIA's Agent Intelligence toolkit to connect, profile, and optimize AI agent teams across complex workflows.

Ingress

Ingress is a mechanism that manages and controls external access to applications and services running within private infrastructure or a cluster, such as Kubernetes. It acts as a gateway, routing HTTP and HTTPS traffic from outside the organization's network to the appropriate internal services based on configurable rules. Ingress enables features like URL-based routing, load balancing, SSL/TLS termination, and name-based virtual hosting, allowing multiple applications to be securely exposed through a single entry point. This approach simplifies network management, centralizes configuration, and enhances security by consolidating how external clients reach internal resources.

Ecosystem Architecture

This section provides an overview of the hardware and software solutions in the enterprise ecosystem that leverage NVIDIA technology to form an NVIDIA Enterprise AI Factory. Additionally, it contains information regarding our various ecosystem partners

who offer solutions for components of the AI Factory, including Enterprise Kubernetes, storage, observability, security and developer tools.

Hardware Infrastructure

The hardware design for the Enterprise AI Factory prioritizes scalability and elasticity, facilitating horizontal scaling of compute with NVIDIA Blackwell GPUs, networking with Spectrum-X, and services using Enterprise ready Kubernetes Platform. This state-of- art hardware ensures performance for achieving the necessary latency and throughput for real-time inference and complex agent interactions. GPU resource optimization is achieved by leveraging effective scheduling, utilization, and management of high-density GPU resources.

The hardware design follows the NVIDIA Enterprise Reference Architecture (Enterprise RA) guidance which is tailored for enterprise-class deployments, ranging up to 256 GPUs. Depending on the base technology, they include configurations for 4 up to 32 nodes, complete with the appropriate networking topology, switching, and allocations for storage and control plane nodes. Enterprise RAs are right-sized for enterprise-scale deployments, it provides deployment guides, cluster characterization, provisioning automation using BCM, and sizing guides for common enterprise AI implementations. NVIDIA Enterprise RAs are designed to support a diverse range of workloads, including AI pre-training, post-training, long thinking inference, HPC, and data analytics. These designs provide a versatile foundation for enterprise AI with a focus on on-premises, single-tenant, Ethernet-based environments.

For more details on these prescriptive Blackwell design patterns and components for building Enterprise AI Factories—as well as the NVIDIA-Certified system used in the Enterprise AI Factory—please refer to the NVIDIA Enterprise Reference Architecture white paper.

When selecting hardware components for an AI platform, several factors come into play to ensure the system meets the demanding needs of AI workloads.

Accelerated Computing Platform

Enterprise AI, particularly for complex agentic systems, demands substantial computational resources that challenge traditional data center capabilities. Agentic systems execute diverse workloads, from sequential task processing and logical reasoning to parallel data analysis and model inference. This operational diversity requires a balanced computing architecture that effectively utilizes CPUs for control, orchestration, and serial tasks, alongside GPUs for massively parallel computations inherent in AI model training, inference, and complex data manipulation.

Accelerated computing platforms, integrating powerful GPUs, CPUs, alongside high-speed networking, all optimized via specialized software stacks, provide the necessary performance and efficiency for these demanding workloads. This approach yields significant improvements in processing speed and energy efficiency over CPU-centric computing.

With the following NVIDIA Blackwell accelerated computing platforms in the NVIDIA Enterprise AI Factory Design Guide, enterprises can unlock the full potential of AI in their data center infrastructure, from accelerating simulations and data analysis to enabling real-time generative design and visualization.

- **The NVIDIA RTX PRO™ Server Edition** is the ultimate data center GPU for AI and visual computing, delivering breakthrough acceleration for the most demanding enterprise workloads, from multimodal AI inference and physical AI to scientific computing, graphics, and video applications. Optimized for workloads requiring the compute density and scale that deploying in the data center offers, the RTX PRO 6000 features a passively cooled thermal design and 96 GB of ultra-fast GDDR7 memory. Enterprises can configure up to eight NVIDIA RTX PRO 6000 GPUs in a server to deliver unmatched levels of compute power, memory capacity, and throughput to power mission-critical AI-enabled applications and accelerate use cases across industries—from healthcare, manufacturing, and geoscience to retail, media, and live broadcast.
- **The NVIDIA HGX™ B200** propels the data center into a new era of accelerated computing and generative AI, integrating NVIDIA Blackwell Tensor Core GPUs with a high-speed interconnect to accelerate AI performance at scale. Configurations of eight GPUs deliver unparalleled generative AI acceleration alongside a remarkable 1.4 terabytes (TB) of GPU memory and 64 terabytes per second (TB/s) of memory bandwidth for 15X faster real-time trillion-parameter-model inference, 12X lower cost, and 12X less energy. This extraordinary combination positions HGX B200 as a premier accelerated x86 scale-up platform designed for the most demanding generative AI, data analytics, and high-performance computing (HPC) workloads.

As detailed in NVIDIA Enterprise RA's, these systems are built on NVIDIA-Certified System servers, designed for optimal performance. For inference-focused platforms, selection criteria prioritize these characteristics:

- **Inference Performance**: High efficiency at various precisions (e.g., FP16, INT8, and newer formats like FP4/FP6 for Blackwell) delivers low-latency and high-throughput model serving.
- **GPU Memory (VRAM)**: Sufficient GPU Memory (VRAM) capacity and high bandwidth are paramount for accommodating the large language models (LLMs) prevalent in Retrieval Augmented Generation (RAG) applications, handling large batch sizes during inference, and supporting the extensive context windows often required by these sophisticated AI agents. Modern NVIDIA GPUs, such as the NVIDIA RTX™ PRO Server Edition with its flexible design and substantial VRAM, or the compute focused NVIDIA B200 Tensor Core GPUs which offer exceptionally large memory footprints, are designed to meet these demands. Reference configurations for AI platforms frequently specify significant memory per GPU to ensure that these complex models and their data can be efficiently processed, enabling low-latency responses and high-throughput performance for AI factory operations.
- **Scalability and Interconnects**: While massive multi-node training setups might be less of a focus, efficient GPU-to-GPU communication via technologies like NVIDIA NVLink can still be beneficial for certain inference scenarios (e.g., model or pipeline parallelism for very large models) and for accelerating the data processing stages in RAG. Server

configurations like PCIe Optimized or HGX systems cater to different scales and performance needs.

Networking

Low-latency networking facilitates efficient data exchange, which is valuable for AI inference, especially in multi-node scenarios. For user-facing applications like AI agents, low latency directly reduces the perceived delay, improving key metrics like Time-To-First-Token (TTFT) and overall response time for a better user experience. When large models are split across multiple GPUs or nodes– using techniques like pipeline or tensor parallelism for inference– each stage of computation depends on timely communication between devices. In these scenarios, even microsecond-scale delays can accumulate, and tail latency—the slowest portion of the communication distribution—can significantly degrade overall performance. Reducing both average and tail latency is essential for ensuring consistent, fast, and responsive AI services at scale.

The NVIDIA Spectrum-X Networking Platform is purpose-built for AI Factories, delivering advanced transport offloads that accelerate collective operations combined with congestion-aware routing and hardware-based scheduling, directly addressing tail-latency issues that bottleneck multi-node AI workloads.

NVIDIA BlueField data processing units (DPUs) are essential for creating high-performance, secure, and efficient AI factories. They offload and accelerate critical tasks such as software-defined networking, storage, and security, freeing up CPU and GPU resources to focus on AI computation. Leveraging purpose-built hardware accelerators and dedicated Arm cores, BlueField supports faster, more secure cloud deployment, zero-trust multi-tenancy, accelerated data access, and real-time threat detection. This enables enterprises to build AI systems that are more scalable, resilient, and optimized for modern, cloud-native infrastructure.

AI Enterprise Infrastructure Software

The NVIDIA AI Enterprise Infrastructure software encompasses all necessary components for managing and optimizing infrastructure along with AI workloads. NVIDIA provides Release Branches to meet organizational needs. The NVIDIA Kubernetes Operators facilitate a standardized management of NVIDIA GPUs, AI models, and network resources within Kubernetes environments. The following table outlines the components and versions of the NVIDIA AI Enterprise Infrastructure software.

Component	Software	Version	Notes
GPU Driver	NVIDIA Linux Driver	570.133.20+	Supported by the GPU Operator 25.03+
GPU Management	NVIDIA GPU Operator	25.03+	Simplifies the deployment of NVIDIA AI Enterprise by auton the management of all NVIDIA software components needed provision GPUs in Kubernetes (drivers, toolkit, DCGM).

Component	Software	Version	Notes
Network Management (Hardware)	NVIDIA Network Operator	v25.1.0+	Simplifies the provisioning and management of NVIDIA netw resources in a Kubernetes cluster (NVIDIA NICs, integrates w NetQ.)
Network Management (Software)	NVIDIA NetQ	Required	Validated Ethernet fabric management.
AI Workload and GPU Orchestration	Run:ai	v2.21+	Dynamic scheduling and orchestration to accelerate AI workl throughput and maximize GPU utilization

AI Enterprise Application Software and Tools

NVIDIA distributes AI and data science tools via NGC container images from its private registry. Release Branches are provided to meet organizational needs. Each image includes the necessary user-space software (i.e. CUDA libraries, cuDNN, TensorRT, and the framework). These NVIDIA AI Enterprise container images, such as the core AI stack and NeMo, are deployed on Kubernetes for easy management, upgrades, and deployment with zero downtime.

Component	Software	Version	Notes
Core AI Stack	NVIDIA AI Enterprise	Latest Supported Version	Includes CUDA, cuDNN, TensorR Triton, etc.
Inference Serving	NVIDIA NIM Operator	v1.0.1+	Manages deployment of NIM microservices.
RAG / Data Processing	NVIDIA NeMo Retriever	Latest Supported Version	Component of NeMo framework fo
RAG / Data Processing	NVIDIA NeMo Customizer	Latest Supported Version	For fine-tuning models.
RAG / Data Processing	NVIDIA NeMo Curator	Latest Supported Version	For data curation pipelines.
RAG / Guardrails	NVIDIA NeMo Guardrails	Latest Supported Version	For adding safety layers to LLM ap
Model Evaluation	NVIDIA NeMo Evaluator	Latest Supported Version	Tools for evaluating LLM performa
Vector Search Acceleration	NVIDIA cuVS	Latest Supported Version	GPU-accelerated vector search libr
Use Case Examples	NVIDIA Blueprints (e.g., AI-Q)	Latest Supported Version	Reference implementations for com cases.

Note

A key component of the Agentic AI blueprints is a foundational Retrieval Augmented Generation (RAG) pipeline. NVIDIA NeMo tools are included for implementing RAG pipelines.

Software Partner Integrations

NVIDIA's comprehensive ecosystem of technology experts include ISVs that bring advanced skills to design, build, and deliver the AI-accelerated computing solutions by integrating NVIDIA AI Enterprise libraries and developer tools into their platforms. A tight collaboration between NVIDIA and our software partner developers and engineers ensures highly optimized and reliable performance over the lifetime of supported application releases. The following software partners have enterprise product offerings that provide components for building Enterprise AI Factories and are categorized as follows:

Enterprise Kubernetes Platform

- Canonical Kubernetes - Canonical provides a few offerings for on-premise Kubernetes solutions with full-lifecycle automation and long term support. Each integrates with the NVIDIA GPU Operator for leveraging NVIDIA hardware acceleration. They support the deployment of NVIDIA AI Enterprise, enabling AI workloads with NIM and accelerated libraries. Canonical's focus on open-source, model-driven operations and ease of use offers enterprises flexible options for building their AI Factory on NVIDIA-accelerated infrastructure
- Nutanix Kubernetes Platform (NKP) - As part of its on-premise Nutanix Cloud Platform (NCP), the Nutanix Kubernetes Platform (NKP) simplifies enterprise Kubernetes management by reducing operational complexity and ensuring consistent, secure deployment across hybrid multicloud environments. It provides centralized fleet management, policy enforcement, and AI-driven observability to streamline Day 2 operations. For AI workloads, NKP can run NVIDIA AI Enterprise , including NVIDIA NIM and NeMo, enabling enterprises to deploy and scale agentic AI applications efficiently. This collaboration allows IT teams to leverage optimized AI models, GPU-accelerated infrastructure, and secure endpoints while maintaining control over data privacy and costs.
- Red Hat OpenShift - OpenShift is an enterprise Kubernetes platform due to its comprehensive, production-grade features that extend beyond standard Kubernetes. It offers enhanced security capabilities out-of-the-box (like Security Context Constraints - SCCs, integrated container registry with security scanning), robust developer and operator tools, integrated CI/CD pipelines, and enterprise-level support. For an AI Factory, its ability to manage complex, stateful AI workloads, provide multi-tenancy, and integrate seamlessly with a wide range of hardware (especially GPUs via operators) and software makes it a strong foundation for AI . Its focus on a consistent operational experience across hybrid cloud environments is also a key advantage for enterprises.
- VMware Tanzu Platform - VMware Tanzu is an application platform that enables enterprises to modernize infrastructure and streamline Kubernetes management, allowing IT teams to deploy scalable, containerized workloads across hybrid environments. Tanzu optimizes AI/ML workloads by leveraging NVIDIA GPU-accelerated Tanzu Kubernetes clusters, with NVIDIA Operators for seamless resource provisioning. This combination supports NVIDIA AI Enterprise and Agentic AI tooling like NVIDIA NIM and NeMo.

Storage Solution

- DDN - DDN provides high-performance, on-premise storage solutions (e.g., EXAScaler) frequently used in NVIDIA DGX SuperPOD and other large-scale AI/HPC deployments. They are NVIDIA-Certified and offer strong support for NVIDIA GPUDirect Storage, ensuring maximum data throughput to NVIDIA GPUs. DDN's focus on massive parallelism and scalability makes them ideal for the most demanding AI training workloads and data-intensive tasks within an AI Factory utilizing NVIDIA NIM and accelerated libraries. They provide CSI drivers for Kubernetes integration.
- Dell - Dell offers on-premise scale-out NAS (PowerScale) and object storage (ECS), frequently part of NVIDIA DGX POD and NVIDIA-Certified Systems. These solutions are optimized for NVIDIA AI workloads and provide CSI drivers for Kubernetes, enabling dynamic storage provisioning and accelerated data access for applications using NVIDIA libraries and NIM.
- Hitachi Vantara - Hitachi Vantara's on-premise enterprise storage integrates into NVIDIA-powered AI infrastructures and can support NVIDIA GPUDirect Storage for NVIDIA GPUs. They provide CSI drivers for Kubernetes, allowing for automated provisioning and management of persistent storage, enhancing data throughput for AI tasks utilizing NVIDIA accelerated libraries and NIM on NVIDIA hardware.
- HPE - HPE offers a range of on-premise enterprise storage solutions, including Alletra for mission-critical workloads and HPE GreenLake for File Storage, which are designed to support AI/ML data pipelines. These solutions can be part of NVIDIA-Certified configurations and support technologies like NVIDIA GPUDirect Storage. With CSI drivers for Kubernetes, HPE storage provides a scalable and resilient foundation for AI Factory data, supporting applications using NVIDIA NIM and accelerated libraries on NVIDIA hardware.
- IBM Storage Scale is built for AI, high-performance computing, and analytics supporting the full range of NVIDIA technologies. IBM Storage Scale provides the flexibility of software-defined global data platform, extensible metadata, multi-tenancy, container native/CSI, and high-throughput object storage, while Storage Scale System, is optimized for clustered low-latency, scalable performance for the most demanding enterprise deployments.
- Netapp - NetApp provides on-premise, NVIDIA-Certified storage solutions optimized for GPU-accelerated workloads on NVIDIA hardware. Through its Astra Trident CSI driver, it offers seamless and dynamic storage provisioning for Kubernetes, supporting high-performance access for applications using NVIDIA accelerated libraries and NIM. This ensures efficient data handling with robust data management features for AI.
- Nutanix Unified Storage - As part of its on-premise Nutanix Cloud Platform (NCP), Nutanix Unified Storage offers integrated file, object, and block storage solutions. These are designed to support AI workloads running on the HCI platform, which itself supports NVIDIA AI Enterprise and vGPU. The storage is provisioned and managed within the Nutanix ecosystem, providing a simplified and scalable data foundation for AI applications, including those using NVIDIA NIM and accelerated libraries on NVIDIA hardware, with CSI driver support for Kubernetes.
- Pure Storage - Pure Storage provides on-premise all-flash solutions like FlashBlade and FlashArray, optimized for NVIDIA GPU-direct technologies and NVIDIA-Certified Systems. They offer robust Kubernetes integration through their Pure Service Orchestrator (CSI driver) and Portworx by Pure Storage for cloud-native storage and data

management, ensuring high IOPS and low latency for AI workloads using NVIDIA accelerated libraries and NIM on NVIDIA hardware.

- Vast - Vast Data's on-premise platform is designed for high-throughput, low-latency access, often integrating with NVIDIA GPUDirect Storage to accelerate AI/ML workloads (using NIM and accelerated libraries) on NVIDIA GPUs. VAST InsightEngine eliminates the bottlenecks of traditional AI architectures, enabling real-time, event-driven AI decision-making. Its CSI driver enables dynamic provisioning and simplified storage management for containerized AI applications within Kubernetes environments, ideal for massive datasets and vector databases.
- Weka - Weka offers a high-performance, on-premise parallel file system (WEKApod often built on NVIDIA-Certified Systems) specifically engineered for AI/ML and HPC workloads. It provides exceptional throughput and low latency, supports NVIDIA GPUDirect Storage, and is frequently chosen for large-scale NVIDIA GPU deployments. Its robust CSI driver ensures seamless integration with Kubernetes for demanding AI training and inference tasks utilizing NVIDIA NIM and accelerated libraries.

Agentic AI Developer Partner Tools

- Accenture AI Refinery - AI Refinery is an Enterprise Gen AI / Agentic AI platform designed to help companies turn raw AI technology into useful business solutions. Built on NVIDIA technology and NVIDIA AI Enterprise software, AI Refinery supports the entire life cycle of the enterprise Generative and Agentic AI – from model customization & serving, agent building & evaluation, knowledge & data processing to governance & observability. Leveraging AI Factory, AI Refinery can deploy pre-built industry solutions on-premise at an accelerated rate. AI Refinery platform enables orchestration of agents from ecosystem providers through Accenture's proprietary trusted agent huddle.
- CrewAI - CrewAI is an open-source framework for orchestrating role-playing, autonomous AI agents. Deployable on-premise, it allows developers to build sophisticated multi-agent systems that can collaborate to solve complex tasks. In an NVIDIA AI Factory, CrewAI can leverage LLMs deployed as NVIDIA NIMs for agent reasoning and decision-making, with the underlying computations accelerated by NVIDIA hardware. This enables the creation of powerful, customized agentic workflows that benefit from NVIDIA's accelerated computing and AI software stack.
- Dataiku - Dataiku is available as an on-premise enterprise AI platform that integrates with Kubernetes environments running NVIDIA GPUs. It allows for the development and operationalization of models that can leverage NVIDIA accelerated libraries and supports workflows that may incorporate NVIDIA NIM for inference, all accelerated by NVIDIA hardware. It also offers both a low-code and advanced workflow building platform.
- DataStax - DataStax provides enterprise solutions that integrate NVIDIA technologies built on Apache Cassandra® and with OpenSearch, deployable on-premise via Kubernetes operators like KubeStax (or the open-source K8ssandra). These offerings deliver a highly scalable NoSQL database with integrated vector search capabilities, making them well-suited as a foundational data layer for an AI Factory, particularly for powering real-time Generative AI and RAG applications. The ability to handle massive datasets and provide low-latency vector search is critical for AI agents developed with frameworks like Langflow that rely on retrieving context for LLMs like NVIDIA NIM.

By serving as a robust backend for contextual data and vector embeddings, DataStax's on-premise solutions support AI agents accelerated by NVIDIA hardware and leveraging NVIDIA's accelerated libraries. DataStax products include Hyper Converged Database (on premise database), AI Platform (powered by NVIDIA AI Enterprise), Langflow (with NIM integrations) and Hybrid Search (semantic and vector search powered by NeMo Retriever)

- DataRobot - DataRobot significantly accelerates the AI development lifecycle by automating many of the complex and time-consuming tasks involved in building, training, deploying, and managing and agents, including embedded support for deploying NVIDIA AI Enterprise and NVIDIA NIM . For an AI Factory focused on agents, DataRobot can help rapidly prototype and deploy the resources that might power the intelligence of these agents, allowing developers to focus more on the agentic logic and integration rather than model tuning from scratch. Its features also ensure models are monitored and managed effectively in production.
- Deloitte Zora AI - Zora AI is an Enterprise Agentic AI platform that simplifies operations, boosts productivity and efficiency, and drives more confident decision-making in enterprises. Zora AI agents deliver industry-specific solutions, augmented with extensive industry knowledge and reasoning capabilities leveraging NVIDIA NIM and NeMo. Zora AI enacts Deloitte's Trustworthy AITM principles, including a human feedback loop, to establish transparency and trust with users. With the NVIDIA Enterprise AI Factory, Zora AI can help its customers in regulated industries deploy AI systems on-premises quickly, offering strong data security guarantees and flexibility in technology options within a trusted ecosystem.
- Domino Data Lab - Domino Data Lab, an NVIDIA AI Accelerated program partner, offers an on-premise MLOps platform integrating with NVIDIA AI Enterprise. It scales GPU infrastructure efficiently on NVIDIA hardware and governs models, supporting AI agent development using NVIDIA accelerated libraries and NIM.
- HPE Private Cloud AI - HPE provides enterprise-grade solutions for AI, with HPE Ezmeral software enabling the deployment and management of containerized AI/ML workloads on HPE's infrastructure, which includes NVIDIA-Certified Systems. HPE Private Cloud AI (PC AI), co-developed with NVIDIA, offers a turnkey solution that integrates NVIDIA AI computing, networking, and software (like NVIDIA AI Enterprise) with HPE's Ezmeral software and infrastructure, aiming to simplify and accelerate AI adoption for enterprises.
- Elastic - The Elastic Stack (Elasticsearch, Kibana, Beats, Logstash) provides on-premise solutions for search, observability, and security. For AI Factories, Elasticsearch's vector search is essential for RAG applications (storing embeddings generated by models on NVIDIA GPUs and queried by NIM-powered agents). The broader stack enables log aggregation, metrics monitoring (including NVIDIA GPU metrics via DCGM exporters), and visualization (Kibana) of the entire AI platform, supporting workloads using NVIDIA accelerated libraries and NIM.
- EY.ai Agentic Platform – EY.ai Agentic Platform is designed to deliver secure and scalable AI solutions for organizations, starting with tax, risk, and finance. It automates processes and enhances decision-making for better business outcomes, while maintaining data privacy and regulatory compliance. The platform will be built on NVIDIA AI Enterprise software, NIM microservices, and the NeMo Framework, trained on EY's

curated data and shaped by EY's deep domain expertise. EY.ai Agentic Platform with NVIDIA Enterprise AI Factory delivers data security, control and low latency to the enterprises that need them most in their transformation journey.

- H2O.ai - H2O.ai offers an on-premise AI platform (H2O AI Cloud, including Enterprise h2oGPTe and Driverless AI) designed for building and deploying both predictive and generative AI models, including agentic AI applications. Their software is optimized for NVIDIA GPUs (leveraging NVIDIA RAPIDS and accelerated libraries) and supports Kubernetes for scalability. This enables enterprises to develop and operationalize AI agents that can utilize NVIDIA NIM for inference, all within their own data centers on NVIDIA hardware.
- JFrog Artifactory - Artifactory is a universal artifact repository manager. In an AI Factory, this is crucial for managing the lifecycle of all binaries, including container images from NGC for AI applications and agents, Python packages, model files, and other dependencies. It provides a single source of truth for all build artifacts, supports versioning, and integrates with CI/CD tools to ensure reproducible and reliable builds and deployments. Its security features, particularly when combined with JFrog Xray, provide deep artifact analysis, vulnerability scanning, and license compliance, which are critical for maintaining a secure software supply chain. JFrog integrates with NVIDIA NIM by embedding NIM microservices and models into Artifactory's unified artifact management framework, enabling centralized governance, secure distribution, and streamlined DevSecOps workflows post-organizational approval.
- Nutanix Enterprise AI - Nutanix Enterprise AI provides a full-stack, on-premise AI software platform built on the Nutanix Cloud Platform (NCP), often incorporating their "GPT-in-a-Box" concept. Beyond core HCI, it offers integrated MLOps capabilities, tools for managing large language models, and simplified deployment of AI workloads. It's designed in partnership with NVIDIA to run NVIDIA AI Enterprise software, including NIMs and accelerated libraries, on NVIDIA-Certified Systems or systems with NVIDIA GPUs, providing a streamlined path for developers to build and deploy AI agents.
- OpenShift AI - OpenShift AI extends the core OpenShift Container Platform with a dedicated platform for on-premise AI/ML development and deployment. It provides data scientists and developers with integrated tools for the entire model lifecycle, including Jupyter notebooks, model training services, model serving capabilities (integrating with NVIDIA NIM), and monitoring tools. Its value lies in streamlining AI workflows and providing a consistent open source environment for developing AI agents on NVIDIA-accelerated infrastructure, leveraging NVIDIA AI Enterprise and its libraries.
- SuperAnnotate - SuperAnnotate provides a comprehensive data annotation platform that supports on-premise data storage and workflows, crucial for AI data preparation. It integrates with NVIDIA technologies like NVIDIA NeMo Evaluator, allowing AI teams to incorporate both human and AI-assisted (LLM-as-a-judge) evaluation for data quality and model assessment. This supports the development of high-quality datasets for training models that will be accelerated by NVIDIA hardware and potentially served via NIM.
- Unstructured.io - Through its integration with NVIDIA NeMo Retriever Extraction, Unstructured enables high-performance processing of multimodal content—such as text, tables, and charts—from large-scale, complex documents like enterprise PDFs. This

collaboration empowers enterprises to prepare vast and varied data efficiently for AI agents and RAG systems, leveraging the speed and accuracy of NVIDIA-accelerated libraries and hardware to meet the demands of scalable, high-performance AI deployments.

- VMware Private AI Foundation with NVIDIA - VMware Private AI Foundation with NVIDIA allows enterprises to leverage their existing VMware infrastructure (vSphere, vCenter, Tanzu) to run AI/ML workloads using NVIDIA AI Enterprise and NVIDIA GPUs (including vGPU capabilities). This solution provides a familiar operational model for IT teams, enabling them to manage and scale AI applications alongside traditional enterprise workloads, with a focus on governance, security, and efficient resource utilization in virtualized environments.

Observability Partner Tools

- Arize AI - Arize AI's engineering platform offers AI observability and evaluation, integrating with NVIDIA NeMo microservices to enable enterprises to build reliable agentic AI. This collaboration creates an AI data flywheel, enhancing LLM performance by combining Arize's evaluation tools with NVIDIA NeMo's capabilities for model training, evaluation, and safety guardrails. Consequently, enterprises can automatically identify LLM failure modes, route complex cases for human feedback, and continuously refine their models through targeted fine-tuning. This process facilitates the development and deployment of accurate agentic AI systems, supported by Arize's self-hosted deployment option for on-premise AI application management.
- Datadog - Datadog is a comprehensive observability platform providing AI application monitoring and security insights. It offers broad visibility across infrastructure, including NVIDIA GPUs, applications, and logs. For an NVIDIA Enterprise AI Factory, Datadog can monitor the performance of deployed AI agents, track custom business metrics, provide distributed tracing to understand request flows through complex agent interactions, and offer security monitoring for production workloads. Its ability to correlate data from various sources, including from OTEL, helps in quickly identifying and resolving issues.
- Dynatrace - Dynatrace delivers full-stack observability, and AI-powered analytics to enterprises. It automatically discovers, maps, and monitors complex, dynamic cloud environments, providing real-time visibility into applications, infrastructure, user experience, and key business metrics. With its proprietary AI engine, Davis®, Dynatrace enables DevOps, SRE, and security teams to proactively detect anomalies, accelerate issue resolution, and optimize performance at scale. As an NVIDIA partner, Dynatrace also offers on-premise observability for Kubernetes clusters running NVIDIA GPUs, AI workloads and applications leveraging NVIDIA AI Enterprise, NIM, and accelerated libraries.
- Fiddler AI - Fiddler AI is an AI Observability and Model Performance Management (MPM) platform designed to help MLOps and data science teams monitor, explain, analyze, and improve production AI models. For an AI Factory deploying numerous agents powered by potentially complex models (including LLMs), Fiddler AI provides crucial capabilities for detecting model drift, data integrity issues, performance degradation, and biases. Its explainability features help understand model predictions, which is vital for debugging agent behavior and ensuring responsible AI

practices. Integration with NVIDIA platforms can ensure that models running on NVIDIA hardware are effectively monitored for operational health and predictive performance.

- Weights & Biases - Weights & Biases, an NVIDIA partner provides powerful tools for visualizing, debugging, and iterating on AI/ML models, with W&B Weave being a primary component for these tasks. Weave allows AI Developers to create dynamic, interactive dashboards and reports to deeply analyze model outputs, track predictions, compare model versions, and understand complex datasets. While W&B also supports experiment tracking (logging metrics, hyperparameters, and artifacts), its strength with Weave in providing rich, shareable insights into model behavior and data makes it invaluable for collaboration on complex AI agent development, debugging, and performance management within the NVIDIA Enterprise AI Factory.

Security Partner Tools

- ActiveFence - ActiveFence provides a platform that connects to on-premise solutions for Trust & Safety, integrating NVIDIA NeMo Guardrails, specializing in detecting and mitigating harmful content (e.g., hate speech, disinformation, CSAM) in online platforms. For an AI Factory deploying agents that interact with user-generated content or generate content themselves, ActiveFence helps ensure agent outputs are safe and compliant, protecting the integrity of applications running on NVIDIA hardware.
- CrowdStrike - CrowdStrike provides robust endpoint detection and response (EDR) and threat intelligence capabilities. In an AI Factory, CrowdStrike secures the hardware, containers, and AI application endpoints. Its cloud-native platform provides real-time visibility and threat hunting capabilities leveraging NVIDIA AI Enterprise, which are essential for securing the valuable IP (models, data) and the operational integrity of the AI platform.
- Galileo - Galileo's AI reliability platform enables enterprise-scale evaluation, iteration, monitoring, and protection of generative AI applications, with on-premises deployment. Its integration with NVIDIA NeMo microservices facilitates AI data flywheels for continuous optimization and high accuracy in agentic AI. This is achieved through comprehensive evaluation with NVIDIA NeMo Evaluator, assessing agent reasoning and awareness; real-time observability for production insights feeding the flywheel; and Galileo Protect with NVIDIA NeMo Guardrails for robust, low-latency safety measures against hallucinations and malicious inputs while ensuring compliance.
- Securiti.ai - Securiti.ai offers an AI-powered Data+AI Security & Governance platform that can be deployed on-premise. For an AI Factory, it provides critical capabilities for discovering, classifying, and securing sensitive data used in AI model training and RAG pipelines, including those leveraging NVIDIA NIM and accelerated libraries. Its ability to enforce data privacy and governance policies across the AI lifecycle is crucial for responsible AI development on NVIDIA hardware.
- Trend Micro - Trend Micro provides comprehensive cybersecurity solutions that can be deployed on-premise, offering protection for servers, containers, and networks within the AI Factory. Their solutions leverage NVIDIA AI and accelerated computing to help secure the underlying infrastructure (including NVIDIA-Certified systems) and workloads running NVIDIA AI Enterprise, NIM, and accelerated libraries from malware,

vulnerabilities, and other threats, contributing to the overall security posture of the NVIDIA hardware-accelerated environment.

Deployment Strategies

To ensure operational efficiency, the AI Factory emphasizes automation. NVIDIA IT and leading enterprises we've worked with have verified and utilized a combination of tools—Infrastructure as Code (IaC), Helm, and Ansible Playbooks—for software installation.

- IaC is employed for repeatable and consistent environment provisioning.
- Ansible playbooks automate the installation of the specified software stack (NVIDIA AI Enterprise, Kubernetes Operators, Partner Integrations) onto pre-provisioned AI Factory hardware and OS.
- Helm charts deploy AI workloads (NIMs) and platform services.
- Certified Kubernetes Operators are used whenever possible

Application and platform configurations are centralized using Helm value files and secure secret management practices, managed via IaC and JFrog Artifactory. Initial configuration parameters can be supplied to the provided Ansible playbooks.

Note

GitOps practices are strongly recommended as part of the approach for managing application configurations beyond the initial installation.

A logical separation of Kubernetes Namespaces, Resource Quotas, and Network Policies within OpenShift is employed to create distinct environments (Development, Testing, Production). This deployment strategy also utilizes standardized Kubernetes patterns, including Operators; NVIDIA Kubernetes Operators (GPU and NIM Operator) as well as Partner Operators.

Future Directions & Ecosystem

The NVIDIA AI Factory Pattern is an evolving architecture updated with operational insights and advancements from NVIDIA and partners. Future improvements may include:

- Incorporation of operating the Data Flywheel required for Enterprise AI
- Integration of next-gen NVIDIA hardware accelerators.
- Adoption of advanced multi-agent system orchestration frameworks.

- Enhanced federated learning and privacy-preserving techniques.
- Expanded library of integrations for MLOps, data management, and security.
- Refined blueprints based on successful internal use cases.
- Guidance on evaluation and model safety

This innovation-driven model ensures the platform stays cutting-edge, offering enterprises, system integrators, and resellers a reliable foundation to reduce time to market and effectively deploy AI capabilities.