Designing an Enterprise AI Factory: Architecture for Agentic Applications

Technical White Paper

Part Number: H04604
October 2025

DELL Technologies

Notes, cautions, and warnings

NOTE: A NOTE indicates important information that helps you make better use of your product.

CAUTION: A CAUTION indicates either potential damage to hardware or loss of data and tells you how to avoid the problem.

WARNING: A WARNING indicates a potential for property damage, personal injury, or death.

Copyright © 2025 Dell Inc. All Rights Reserved. Dell Technologies, Dell, and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.

The information in this publication is provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose, and noninfringement. In no event shall the authors or copyright holders be liable for any claim, damages, or other liability, whether in an action of contract, tort, or otherwise, arising from, out of, or in connection with the publication or the use or other dealings in the publication.

The use, copying, and distribution of any software that is described in this publication requires an applicable software license.

THIRD-PARTY PRODUCTS DISCLAIMER. This solution is used with products, services, or other items that are provided by a third-party manufacturer or supplier and are not "Dell" or "Dell EMC" branded (collectively, "Third-Party Products"). Notwithstanding any other provisions: (1) such Third-Party Products are subject to the standard license, services, warranty, indemnity, support and other terms of the third-party manufacturer/supplier/community (or an applicable direct agreement between you and such manufacturer/supplier), which shall take priority; and (2) any claims we have acknowledged against Dell in relation to such Third-Party Products are expressly disclaimed and excluded.

Contents

1

Introduction

Topics:

Executive summary

The enterprise AI landscape is undergoing a rapid transformation, driven by the rise of agentic AI and the increasing demand for intelligent, scalable solutions. As organizations seek to operationalize AI across business functions, they face challenges such as architectural complexity, integration overhead, and the need for secure, governed deployments.

To address these challenges, Dell Technologies and NVIDIA have partnered to deliver a prescriptive reference architecture to accelerate enterprise adoption of agentic and generative AI. This reference architecture provides clear guidance on production-ready components across the full AI life cycle. It consolidates essential building blocks of accelerated compute into a layered, modular design that supports secure and governed deployments, while enabling flexibility in managing dependencies and evaluating alternative solutions.

This approach has the following key benefits:

- Accelerated time-to-value: By recommending validated components and configurations, the architecture reduces deployment complexity and speeds up the transition from experimentation to production – leveraging NVIDIA's advanced accelerated compute infrastructure, software tools, third party ISVs, and the Dell AI Factory with NVIDIA.
- Enterprise-grade readiness for agentic AI: The design supports both inference and AI application development, with a forward-looking capability to extend into agent-based systems and emerging use cases.
- Modular, scalable, and adaptable foundation: While this technical white paper provides a robust foundation that goes beyond the Dell AI Factory with NVIDIA design guide, it represents just one variation Dell Technologies plan to deliver. Expanded recommendations and new, detailed capabilities to meet diverse enterprise needs are expected to be incorporated. This ensures organizations can tailor implementations and remain agile as technologies and business requirements evolve.
- Expert Guidance plus hands-on implementation and operational support from Dell Professional Services fast-tracks value creation from enterprise solutions. From strategic planning and solution design to accelerated production ready-solutions plus full-stack managed operations, we simplify AI complexity and ensure business-aligned outcomes.

The architecture is based on the NVIDIA Enterprise AI Factory Design Guide. One implementation can be found in the Dell AI Factory with NVIDIA, which offers a modular and scalable foundation for deploying agentic and generative AI workloads that leverages the Dell Automation Platform. It also highlights the role of ISV (Independent Software Vendor) partners who provide production-ready and optimized AI enabling platforms that are designed with performance, consistency, and security in mind.

About this document

This guide defines the reference architecture for building enterprise AI factories using Dell Technologies infrastructure, NVIDIA accelerated compute, AI tools, and the essential ISV software. It focuses on identifying the essential components of a scalable AI stack—compute, storage, networking, orchestration, and AI services—and how they integrate to support agentic and generative AI workloads.

It also highlights the role of Dell Automation Platform (DAP) in streamlining deployment and operations.

The goal is to provide a clear but flexible framework for enterprise teams to accelerate AI adoption, reduce integration complexity, and ensure performance, security, and interoperability across the full life cycle of AI deployment.

Audience

This document is intended for CIOs, IT architects, data scientists, MLOps engineers, infrastructure administrators, and business decision-makers involved in designing, deploying, or managing enterprise AI solutions. It assumes familiarity with cloud-native platforms, AI/ML workflows, and enterprise infrastructure operations.

Revision history

The following table shows the revision history of this document.

Table 1. Revision history

Date	Version	Change summary
October 2025	1.0	Initial release

2

Architecture Overview

Topics:

Enterprise AI challenges

As AI adoption accelerates, enterprises face pressure to support increasingly complex workloads—ranging from traditional machine learning (ML) workloads to scalable inference using large language models (LLMs) and working with Retrieval-Augmented Generation (RAG) pipelines. These workloads demand infrastructure capable of handling massive data volumes, low-latency inference, and dynamic resource allocation. Additionally, most organizations lack a clear focus on highest value opportunities, which can stall adoption and value creation. It's critical to identify business goals and prioritize use cases accordingly to ensure the needs of the business fit the plans of the technology.

The challenge lies in integrating compute, storage, networking, and software while maintaining performance, security, and manageability. Fragmented tooling, inconsistent deployment models, and lack of standardization often slow down progress and increase operational overhead.

This white paper addresses these challenges by suggesting a modular AI stack that simplifies integration and accelerates deployment—built on Dell Technologies infrastructure, NVIDIA's Enterprise Reference Architecture for AI, and a rich ecosystem of ISV partners. It also demonstrates how guidance from Dell Professional Services defines the right foundation to achieve defined business objectives.

Use cases

This architecture supports a wide range of enterprise AI applications, with a particular focus on inference on LLMs and agentic AI systems for reasoning, decision-making, and task execution. Typical use cases include content creation, assistant agents like intelligent customer service agents, supply chain optimization, marketing automation, with the ability to work with multimodal data including video and connect to enterprise systems and edge deployments.

The framework that is introduced in this document reflects practical insights gained from NVIDIA's internal enterprise IT initiatives and real-world implementations, combined with Dell Technologies strategic partnerships and deep experience in enterprise infrastructure. This foundation enables advanced agentic and generative AI capabilities empowering organizations to build high-performance, scalable, and context-aware AI solutions tailored to their operational needs.

Architecture for agentic applications

The Enterprise AI Factory with NVIDIA is built on a modular architecture that brings together infrastructure, platform services, AI enablement, and governance into a unified stack. This structure is designed to support LLM inference and agentic AI systems —AI agents capable of reasoning, planning, and acting autonomously—while ensuring enterprise-grade scalability, security, and manageability.

At a high level, the architecture is organized into layers:

- The bottom layer provides the infrastructure foundation, including compute, storage, networking, orchestration, and automation tools. It ensures performance, reliability, and integration with enterprise IT environments.
- The top layer focuses on AI enablement, including AI development and operations for models and agents with full support for observability, security, and data connectivity. This is where AI services are exposed, monitored, and controlled.

The architecture builds on top of enterprise cloud-native platforms with Kubernetes, incorporating GitOps practices and specialized modules to enable AI development, including vector databases, artifact repositories, and the NVIDIA AI Enterprise software with NVIDIA NIM and NeMo microservices.

The following figure illustrates the high-level architecture. Traffic reaches the ingress point and activates agent functions that rely on models developed and managed by the AI software platform, and hosted on the underlying AI infrastructure. Each layer is modular and interoperable, enabling optionality, flexible deployment, and life cycle management across on-premises environments.

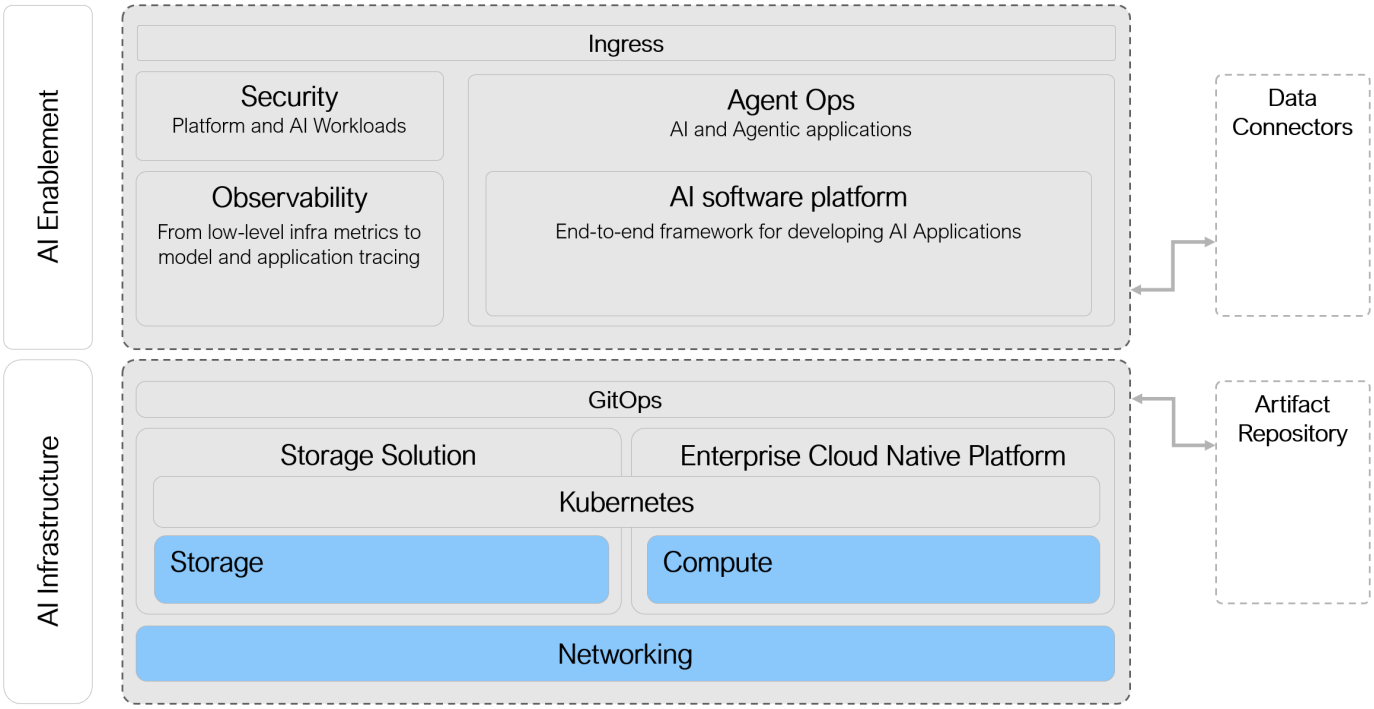

Figure 1. Enterprise AI factory architecture for agentic applications

3

AI Enablement Layer

Topics:

Enabling generative AI and agentic applications

For organizations looking to build intelligent, secure, and observable AI applications—where agility and scale are key - the AI enablement layer becomes a strategic asset. Think of it as the intelligent backbone that exposes, secures, and manages applications and services—making it easier for teams to deploy enterprise-grade AI applications and agents.

This layer brings together critical capabilities like AI development and agent operations, observability, security, data connectors, and ingress traffic management, all designed to work seamlessly across dynamic, containerized environments. The result? Faster time-to-value for AI initiatives and enterprise-grade AI applications, reduced operational overhead, and a production-focused infrastructure that can adapt to change in real time and seamlessly integrate with enterprise systems.

The architecture diagram in Figure 2 illustrates the functionalities and connectors of the AI enablement layer. Note that this represents a detailed version of the top layer that is presented in Figure 1. This chapter provides a description for each key component.

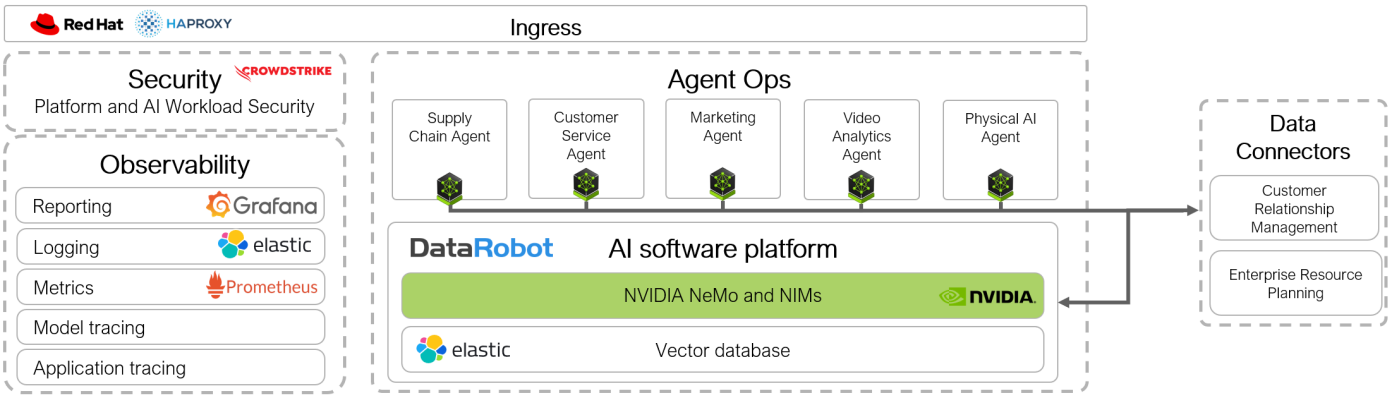

Figure 2. AI enablement functional architecture

Deployment and operation

As AI applications and agents begin to deliver real business outcomes in production, the need for robust operational infrastructure becomes crucial. Without effective deployment and operationalization, the ROI of AI initiatives remains theoretical. The framework that this guide describes goes beyond development support to enable Agent Ops—the discipline of managing agent operations across deployment, scaling, monitoring, and life cycle management. This includes provisioning and orchestrating compute resources, integrating CI/CD pipelines, and validating agent behavior across development and production environments to bridge the gap between experimentation and enterprise-grade execution.

To accelerate this process, Dell Technologies integrates NVIDIA blueprints: prebuilt, production-ready workflows that combine NVIDIA NIM and NeMo microservices and orchestration tools. These blueprints enable enterprises to deploy agents that can reason, plan, and act autonomously, supporting use cases such as document summarization, voice interaction, and structured report generation.

NVIDIA NIM and NeMo microservices

To enable the blueprints and deliver HW-optimized AI inference, NeMo provides a framework for building and customizing LLMs and workflows, while NIM delivers optimized microservices for scalable inference. Together, they enable fast, secure deployment of agentic capabilities across environments. The NeMo framework also provides guardrail functionalities as a safety and control mechanism to ensure that LLMs behave reliably, ethically, and securely.

AI software platform

An AI software platform enables developers to build, develop, evaluate, deploy, and monitor machine learning models, LLMs, and AI applications. Beyond supporting standard MLOps or LLMOps functionalities, it also supports RAG and agentic workflow development and life cycle management with built-in governance and automation.

Dell Technologies partners with DataRobot which serves as a comprehensive AI software platform providing model and application governance and transparency with easy interaction to enterprise systems such as SAP, in a low-code/no-code interface. It offers deep integration with NVIDIA AI Enterprise, enabling the developers to rely on NVIDIA NIM models, NeMo microservices, and AI blueprints, all optimized for accelerated computing on NVIDIA GPUs. For more information, see DataRobot.

Vector database

To enable RAG workflows and semantic search, a vector database stores data that could be pre-processed before embedding (especially multimodal data requiring OCR) for LLMs and agents to retrieve and interact with. Dell Technologies incorporates Elasticsearch, a fast, scalable vector database and context-engineering platform for structured, unstructured, and vector data, powering hybrid search, real-time analytics, and exceptional relevance through a single flexible API. Hybrid search with Elasticsearch combines keyword precision and semantic understanding to deliver highly relevant results—more accurate than either approach alone, even at extreme scale. As an upcoming feature, it will also enable NVIDIA GPU-accelerated vector search and retrieval in the near future. For more information about how Elasticsearch also integrates with Dell AI Data Platform, see this Elastic blog.

Observability

Observability provides comprehensive visibility into the reliability, performance and behavior of AI agents, AI applications and platform services. It includes centralized logging, metrics, tracing, and reporting—each integrated to support enterprise-grade monitoring, diagnostics, and governance. While providing deep insights into the system's behavior, observability tools enable teams to optimize performance, proactively identify issues, and debug complex interactions. On top of standard practices, AI applications present unique new challenges. Machine learning models and LLMs require detailed tracing to track model behavior and performance, which can be combined with application tracing and consolidated reporting to understand the overall health of the platform and the often-complex flow of operations. These are the main observability areas:

- Metrics: To enable proactive, continuous monitoring and resource optimization, fine-grained metrics are required from the operating system and underlying hardware. Enhanced GPU metrics are collected by NVIDIA Data Center GPU Manager Exporter (DCGM Exporter). Besides the hardware centric metrics, Key Performance Indicators (KPIs) relevant to the AI applications and AI agents help monitor workload efficiency, model behavior and health at each level.
- Logging: System events, log data is collected and aggregated from infrastructure components, containerized environments, and AI services—from data ingestion to training, inferencing and deployment—to enable effective debugging, maintain audit trails, and support security analysis.
- Reporting: Real-time dashboards present visualizations of system health, infrastructure metrics, and usage for IT and MLOps/LLMOps teams. The same consolidated reporting can also be tailored to show AI agent performance, health, and accuracy, along with operational metrics, model insights, security and compliance indicators, and business impact—providing role-specific visibility for IT operations, AI developers, and business stakeholders.
- Model tracing: Comprehensive insights into model operations provide deep visibility into precise lineage tracking, debugging, and performance optimization. With this, teams can continuously evaluate results to ensure accuracy, faithfulness, and reliability.
- Application tracing: AI applications work with agent workflows, tool calls, and service interactions. Tracing these functions can help in enabling precise monitoring, debugging, and system optimization which help identifying issues, removing bottlenecks, and optimizing execution paths.

Metrics can be exposed through Prometheus, an open-source monitoring system that enables the collection and querying of time-series data. For logging, Elastic can be used to ingest and index logs from both infrastructure components—such as

containers, various services and orchestration layers—and inference endpoints or agentic applications, capturing execution traces, decision steps, and inter-component communications. It supports scalable ingestion of structured and unstructured logs, enabling centralized, searchable log storage that supports traceability, error tracking, and performance auditing across distributed components. For more information about how Elasticsearch supports ingesting data from Prometheus, see also this web page.

To support reporting capabilities, Grafana aggregates metrics from the AI infrastructure, inference services, and agentic applications into real-time, customizable dashboards, offering rich visualization and alerting features for performance tracking, anomaly detection and system oversight.

Beyond providing key AI software platform capabilities, DataRobot also offers model and application tracing capabilities that are tailored for LLMs and AI agents, with native support for OpenTelemetry (OTel) to ensure standardized instrumentation. This enables consistent tracking of agent activities, decision paths, and performance metrics. Users can monitor and trace workflows and components of an AI pipeline including prompts, vector databases, LLMs, and predictive models.

Security

AI requires a truly multi-layered strategy for security that protects infrastructure, workloads, and data across the entire AI life cycle. This begins on the network layer, primarily using network policies to control traffic flows between different services and pods to isolate workloads and restricting communication to minimize the attack surface.

On top of the network layer, authentication and authorization mechanisms that are integrated directly with the platform's own access systems help to verify user and service entities and their entitlements. Integration with broader enterprise Identity and Access Management (IAM) solutions ensure consistency and provide centralized control over user access. Role-based Access Control (RBAC) implemented on various levels of the platforms and in the ecosystem of components restricts access to platform-specific functionalities and resources based on predefined roles. For AI-specific environments, zone-based roles, separation of duties, and regular audits, including for autonomous AI agents, mitigate privileges accumulation or misuse.

Encryption is provided across all layers for both data in transit and at rest, ensuring confidentiality and integrity while aligning with enterprise security frameworks. Security telemetry, including event data and audit logs, are forwarded continuously to enterprise Security Information and Event Management (SIEM) systems, enabling real-time centralized monitoring, anomaly detection, threat detection, and incident response. Monitoring capabilities can be expanded beyond infrastructure to include model output auditing, application API logs, and memory integrity checks to catch novel attacks and ethical violations

As the AI environments grow in complexity, deep integration with observability and behavioral analytics becomes increasingly critical. This enables detection of anomalous network patterns, resource usage spikes, model drift, suspicious API interactions, and unforeseen dependencies. Automate security testing throughout the CI/CD pipeline using adversarial and vulnerability scans, and require third-party model and dataset vetting to defend against supply chain and shadow AI risks. More importantly, with the ability to correlate signals across these layers, security teams can detect and respond to sophisticated, multi-stage threats—such as lateral movement from infrastructure to model manipulation, or unauthorized data access initiated by compromised service identities.

Together, these capabilities ensure that AI services are not only intelligent and scalable, but also secure, observable, and resilient—ready to meet the demands of modern enterprise environments.

To support platform-level protection, Dell Technologies integrates CrowdStrike technologies into its service delivery, providing real-time threat detection across the AI stack for infrastructure, data, and container visibility. This secures the underlying infrastructure and CI/CD pipelines, ensuring that AI services are deployed and monitored with state-of-the-art endpoint and behavioral defenses. For more information, see the Dell Technologies and CrowdStrike partnership announcement. See also the joint Dell Technologies and CrowdStrike offering to boost cyber defense.

For AI workload safety and model security, NeMo Guardrails and partner solutions enforce policy controls, filter outputs, and validate agent behavior. These guardrails help prevent hallucinations, enforce compliance with data policies and regulations, and ensure that agents only operate within predefined boundaries.

For more information, the Dell AI Platforms security best practices guide describes some of the potential threats to an AI Platform and offers guidance on how to mitigate them. It also describes some common security configuration procedures, followed by guidance related to categories of potential threats and their respective mitigations. The recommendations in the security best practices guide are not a complete security evaluation of the Dell AI Platform in its entirety or its individual components, nor a complete threat and risk analysis. You can find additional information at the Dell Security and Trust Center.

Data connectors

AI agents require secure and scalable access to enterprise data to deliver meaningful outcomes. The functional software architecture for AI requires connectivity to systems like Customer Relationship Management (CRM), Enterprise Resource

Planning (ERP), and Point of Sale (POS), enabling agents to retrieve structured and unstructured data for reasoning and decision-making.

To power RAG workflows, the Dell AI Factory with NVIDIA supports embedding generation and vector database storage, allowing agents to perform semantic search and context-aware responses.

This architecture is designed with flexibility to accommodate emerging standards like the Model Context Protocol (MCP) and tools such as the NVIDIA NeMo Agent Toolkit and NVIDIA NeMo Retriever, which help agents discover, connect to, and interact with enterprise data sources securely and efficiently.

To extend the connections to optimized storage, query and data engines, the Dell AI Data Platform is a comprehensive solution designed to support the entire data life cycle for analytics and AI applications. It provides a unified, modern approach to managing complex data workflows at scale, empowering teams to turn raw data into strategic assets and accelerate innovation. For more information, see Dell AI Data Platform.

These connectors form the bridge between enterprise systems and AI logic, enabling agents to operate with real-time, relevant information while maintaining data governance and security.

Traffic ingress

Traffic ingress is responsible for exposing AI applications and agent services to both internal and external users and enterprise systems in a secure and scalable manner. The framework proposed in this document supports routing and load balancing for endpoints using proven technologies such as HAProxy. This open-source software is integrated with Red Hat OpenShift and provides a high availability load balancer and proxy server for TCP and HTTP-based applications. For more information, see the https://www.haproxy.org/ and the HAProxy chapter in the Red Hat Enterprise Linux documentation.

These ingress solutions enable flexible service exposure across environments, supporting SSL/TLS termination, path-based routing, and multi-tenant isolation. They ensure that agent endpoints are reachable, performant, and protected—whether deployed internally or externally.

As part of the overall architecture, ingress components integrate with the orchestration and security layers to maintain consistent access policies and observability across the AI stack.

4

AI Infrastructure Stack

Topics:

Core infrastructure overview

To support the demands of agentic and generative AI workloads, the Dell AI Factory is built on a robust, scalable infrastructure foundation that is compliant with the NVIDIA Enterprise Reference Architecture (ERA). This infrastructure is designed to deliver high performance, low latency, and enterprise-grade reliability. The infrastructure relies on:

- NVIDIA-certified systems: Dell platforms are equipped with the latest NVIDIA GPUs, optimized for large model inference, and Spectrum-X networking, purpose-built for AI workloads with congestion-aware routing and tail-latency mitigation.
- Scalable reference architecture: The infrastructure supports up to 256 GPUs across 32 nodes, enabling horizontal scaling for compute-intensive tasks. This modular design ensures flexibility for both small-scale deployments and large enterprise AI factories.

The following figure illustrates the foundational infrastructure stack of the Dell AI Factory. It highlights the integration of compute, networking, and storage components, orchestrated through a cloud-native platform and automated via GitOps workflows.

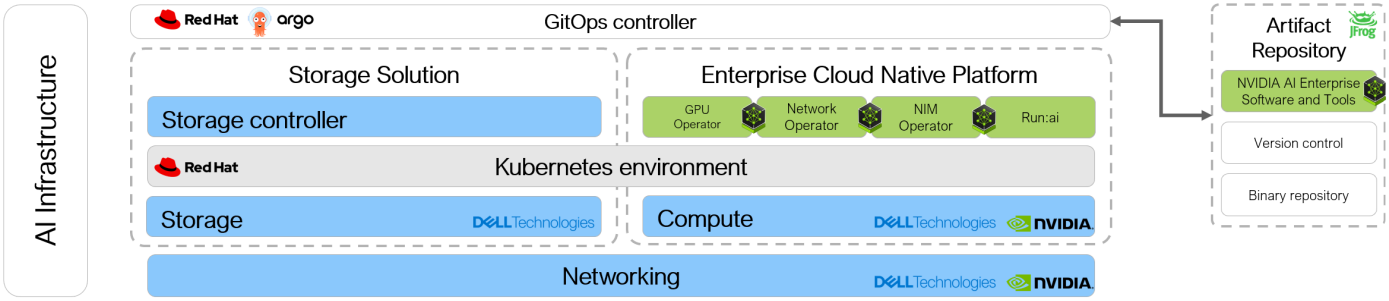

Figure 3. AI infrastructure stack functional architecture

The figure sets the stage for the detailed breakdown in the following sections:

- **Networking**: Spectrum-X switches and BlueField DPUs for secure, high-throughput connectivity.
- **Compute**: Dell PowerEdge servers with NVIDIA GPUs.
- **Storage**: Tiered architecture supporting RAG pipelines, vector databases, and model weights.
- **Platform Services**: Kubernetes-based orchestration, GitOps controllers, and artifact repositories.

Networking

AI workloads—especially those involving LLMs, generative inference, and multi-agent systems—place demanding requirements on networking infrastructure. These include high bandwidth, low latency, and predictable performance to ensure efficient communication between GPUs across nodes. Traditional Ethernet fabrics often struggle with congestion and inconsistent throughput, which can hinder distributed training and inference. To address this, Dell AI Factory integrates a networking stack that is optimized for AI scalability and reliability.

At the core of this stack are the NVIDIA Spectrum-X switches and the BlueField-3 SuperNICs. Spectrum-X switches deliver ultra-low latency and congestion-aware routing, while BlueField DPUs offload networking tasks such as traffic control and encryption—freeing up CPU and GPU resources and enhancing platform security. Together, these components form a high-throughput, lossless fabric that is designed to support demanding AI workloads with consistent performance. Alternatively, customers can choose to purchase their network infrastructure using Dell PowerSwitch options. For more information, see Data Center Network Solutions.

The architecture includes three distinct network layers: a frontend network for management, storage, and client/server traffic; a backend network for GPU-to-GPU communication during training and inference; and an out-of-band network for server administration. In larger deployments, these can be consolidated into a converged fabric, simplifying infrastructure while maintaining performance isolation. This layered approach ensures that AI workloads are supported by a robust and scalable networking foundation.

Compute

The AI infrastructure is designed around a modular compute architecture where GPUs handle the core AI processing tasks—training, inference, and data parallelism. A typical deployment starts with a control plane of four nodes to manage orchestration, scheduling, and system health. The number of worker nodes is determined by workload intensity and user demand. As AI use cases expand and concurrent users increase, the need for GPU memory (vRAM) grows, requiring additional GPUs and corresponding worker nodes. These deployments leverage Dell PowerEdge servers optimized for AI workloads, offering configurations tailored to performance, density, and bandwidth requirements. For more details, see PowerEdge AI Servers .

Dell supports multiple NVIDIA GPU types—such as H200, L40S, and RTX series—to address the varying demands of AI workloads. These GPU types differ in architectural design, memory capacity, and performance characteristics, evolving with each generation to align infrastructure with the specific needs of AI development stages while maintaining efficiency and cost-effectiveness.

Storage

This architecture uses Dell PowerScale storage for datasets, model customization, versioning, and ensemble management. It supports the large volumes of data required for training and deploying AI models.

PowerScale integrates with Kubernetes through the Container Storage Interface (CSI), enabling dynamic provisioning of persistent volumes for containerized AI workloads. This simplifies storage management and ensures seamless access to data across distributed compute environments.

Powered by the OneFS operating system, PowerScale includes built-in features such as inline compression, deduplication, and multi-level data protection to ensure performance, efficiency, and data integrity. Common configurations include models like the PowerScale F710, with clusters starting at three nodes. For more information, see PowerScale.

Cloud-Native Platform

The Dell AI Factory with NVIDIA leverages a Kubernetes-based orchestration platform to unify infrastructure and AI services into a resilient environment. This cloud-native foundation enables dynamic resource allocation, automated deployment, and life cycle management of AI workloads across on-premises infrastructure.

Among the available Kubernetes distributions, Red Hat Enterprise Linux CoreOS is highlighted as the foundation for Red Hat OpenShift, which builds on top of Kubernetes to deliver enterprise-grade capabilities, including integrated security and compliance features. OpenShift enhances operational consistency and developer productivity. For more information, see the OpenShift Container Platform documentation.

The platform uses Kubernetes Operators to automate the deployment and management of complex services. Operators extend Kubernetes' capabilities by encapsulating operational knowledge into code, enabling consistent and reliable management of infrastructure components.

Key operators include:

- GPU Operator: Automates the provisioning and configuration of NVIDIA GPUs, including driver installation and monitoring.
- Network Operator: Manages high-performance networking components such as RDMA and NVIDIA ConnectX SmartNICs, ensuring optimal data throughput and low latency.
- NIM Operator: Facilitates the deployment and life cycle management of NVIDIA Inference Microservices, enabling rapid rollout of pre-trained AI models.

- Run:ai: To optimize resource scheduling and utilization, the Dell AI Factory integrates Run:ai, which provides a virtualization layer for GPUs. It enables dynamic allocation, prioritization, and quota enforcement across teams and workloads, improving efficiency and visibility in multi-user environments. For more information, see NVIDIA Run:ai .

As illustrated in the architecture diagram, Kubernetes orchestrate containerized services, integrate with GitOps workflows, and manage the full stack—ensuring consistency and performance.

Platform Services

Building on the foundational layers of compute, storage, networking, and cloud-native orchestration, the Dell AI Factory incorporates platform services to support the development and operations of AI workloads. This layer includes GitOps-based automation and repositories.

GitOps controllers

The framework for AI in the enterprise uses GitOps controllers to automate and standardize the deployment and management of workloads. These controllers maintain alignment between the desired state defined in Git repositories and the actual state of the Kubernetes environment, supporting version control, rollback, and policy enforcement while reducing manual effort and operational risk.

One option for implementing GitOps in OpenShift environments is Argo CD, a declarative continuous delivery tool that automates synchronization, enforces access controls, and provides visual tracking of application status. For more information, see Argo CD.

As shown in the architecture diagram, GitOps integrates tightly with the artifact repository and orchestration layer, forming a closed-loop system for secure and scalable AI operations.

Repositories

The Dell AI Factory includes a local artifact repository that serves as a secure, version-controlled hub for storing and managing essential components such as AI models and container images. This repository is integrated with the GitOps workflow, enabling automated deployment and consistent configuration of AI workloads across the Kubernetes platform. By maintaining artifacts locally, enterprises gain control over software provenance, reduce reliance on public registries, and could enhance operational security through vulnerability scanning and policy enforcement. The repository plays a central role in ensuring reproducibility, traceability, and governance throughout the AI life cycle.

Key repository components include:

- NVIDIA AI Enterprise: A curated software suite that provides access to validated AI frameworks, pre-trained models, and optimized containers for NVIDIA GPUs. It supports enterprise-grade as well as government-ready deployment of AI workloads with compatibility across major platforms, including OpenShift and Kubernetes. The suite also includes access to foundational models like Nemotron, designed to accelerate the development of custom LLMs for enterprise use cases. For more information, see NVIDIA AI Enterprise.
- Version control system: Stores infrastructure and application configurations as code, enabling versioning, auditability, and collaboration. It serves as the single source of truth for GitOps workflows, ensuring consistent and trackable changes across environments.
- Binary repository: Hosts container images, model files, and other build artifacts. It supports access control, vulnerability scanning, and integration with CI/CD pipelines, ensuring that only verified and approved components are deployed into production.
- Dell Enterprise Hub: An external, collaborative portal developed with Hugging Face, offering a curated catalog of open-source models and AI applications optimized for Dell Technologies infrastructure, including AI servers and PCs. It provides ready-to-deploy containers with both Helm charts, and Docker scripts for training and inference, supporting on-premises deployment of generative AI solutions through pre-validated configurations and a simplified DIY approach. For more information, see Dell Enterprise Hub by Hugging Face.

As an implementation option, JFrog Artifactory serves as a universal repository manager that streamlines storage, versioning, and distribution of machine learning models, datasets, and containerized environments. By centralizing artifacts across the AI development life cycle, Artifactory ensures reproducibility, accelerates CI/CD pipelines, and enhances collaboration between data scientists and DevOps teams, see JFrog Artifactory.

5

Implementation with Dell AI Factory

Topics:

Overview of implementation strategies

Dell Technologies recognizes that AI adoption is not a one-size-fits-all journey. Each organization has unique goals, constraints, and starting points. That is why flexibility and optionality are built into every layer of the Dell AI Factory architecture. Customers can begin with core infrastructure—servers, storage, and networking—or choose from validated AI platform configurations tailored to enterprise use cases. And they can rely on trusted expertise from Dell Services to streamline and quickly realize production-ready values from AI solutions.

This modular approach extends to software, orchestration, and workload deployment. Organizations can integrate their preferred ISV solutions or adopt Dell pre-validated AI workloads. They can use their own tools and scripts or streamline operations with the Dell Automation Platform. For day 0, day 1, and day 2 operations, customers may choose to manage internally or engage Dell Professional Services. From assessing the needs of the environment and designing tailored solutions to deploying and seamlessly integrating hardware, software, tools, frameworks, and applications, Dell Services provides end-to-end support, including the option for fully managed operations. Dell Managed Services offer comprehensive full-stack management across hardware, platforms, and workloads to deliver optimal performance and reliability.

By offering multiple entry points and deployment paths, Dell enables organizations to adopt AI at their own pace—whether starting with a proof of concept or scaling to full production. The following sections detail the platforms, workloads, automation tools, and services that support this flexible implementation model.

Dell AI Platforms

The Dell AI Factory with NVIDIA is built on a portfolio of validated platforms designed to support a wide range of AI workloads. These include PowerEdge XE7740, XE7745, XE9680, and R760xa servers, configured with up to eight NVIDIA GPUs per node—including H200 NVL, H200 SXM, RTX Pro 6000, H100 NVL, and L40S. These platforms are optimized for training, fine-tuning, and inference of LLMs and multimodal applications.

Storage is provided by Dell PowerScale, and networking is powered by NVIDIA Spectrum SN5600 switches for low-latency, congestion-aware routing. The platform includes Linux, Kubernetes, and NVIDIA AI Enterprise, ensuring compatibility with NVIDIA NIM and NeMo microservices and a broad range of ISV workloads.

Customers can tailor deployments by selecting Dell switches, using Red Hat OpenShift for container orchestration, or opting for open-source software—variations documented in dedicated design guides. The platforms scale from a small cluster to up to 32 nodes and 256 GPUs, offering modular hardware, software stack options, and integration patterns to meet specific performance, security, and operational needs. The example shown below is a small size Dell AI Platform with NVIDIA configuration using Red Hat OpenShift. For more information, see the design guide.

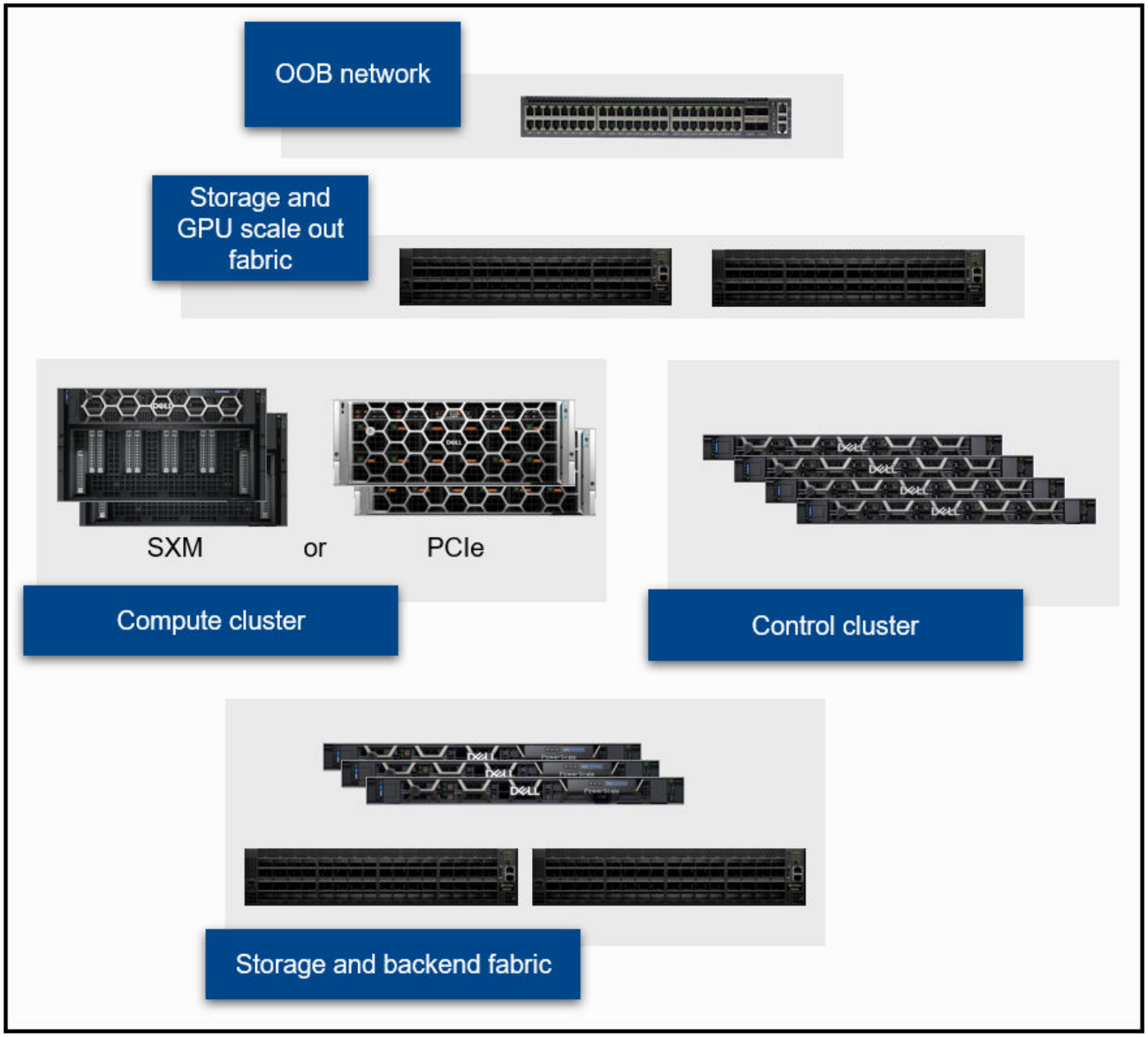

Figure 4. Example of Dell AI Platform configuration

AI workloads

The Dell AI Factory supports a broad range of AI workloads for enterprise use cases, from content creation and code generation to digital assistants and beyond. These workloads include AI software platforms and tools that address the entire AI life cycle, covering model development, fine-tuning, RAG, multi-agent orchestration, and more.

While Dell Technologies infrastructure ensures consistent throughput and low latency, these ISV integrations add ready outcomes and AI capabilities. The supported AI workloads include building blocks like NVIDIA NeMo, NIM microservices, and leading ISVs including Cohere, Tabnine or DataRobot.

By leveraging the Dell Automaton Platform, users can get access to a self-service catalog to explore, deploy, and manage AI workloads for their expected outcomes. AI workload solution guides provide practical, workload-specific guidance on architecture, infrastructure design, and operational best practices. For more information, see the AI workloads catalog.

Dell Automation Platform

Automation is a core enabler of scalability and operational efficiency for AI in the enterprise. The Dell Automation Platform (DAP) delivers a vision for full-stack automation, spanning infrastructure provisioning, platform services, and AI workload deployment.

At the heart of DAP is a catalog-based system, where blueprints ensure auditable, repeatable deployments and support full life cycle management—including updates, rollbacks, and policy enforcement across the stack.

From a single, unified interface, DAP supports automation across multiple Dell offerings:

- AI Solutions: Accelerates deployment of agentic and generative AI workloads through blueprint-driven workflows tailored for enterprise outcomes.
- Dell Private Cloud: Enables seamless orchestration and life cycle management across private cloud environments.
- NativeEdge: Extends automation to edge locations, supporting real-time AI workloads with low-latency responsiveness.

In the context of AI, Dell Professional Services can leverage DAP for turnkey deployments of Dell AI Platforms. Services experts help further maximize the value of Dell AI Solutions delivered by Dell Automation Platform, tailoring solutions to meet the specific needs of the business and environment and improving adoption of validated workloads. DAP also empowers enterprise teams to deploy and manage AI workloads independently, through a self-service catalog.

By integrating DAP into the Dell AI Factory, Dell Technologies simplifies the deployment of complex AI systems, reduces operational overhead, and accelerates time-to-value for enterprise teams across industries.

For more information, see the Dell Automation Platform portal.

Dell Professional Services

Dell Professional Services provides end-to-end guidance and hands-on work to overcome the challenges of AI complexity and accelerate measurable value. Trusted experts improve adoption throughout the life cycle so that you can deploy confidently, operate efficiently, and scale quickly.

Dell Technologies enables organizations to start small to win big—with easy points of entry that provide the clarity and efficiency required to realize meaningful outcomes at the speed the market demands. Dell offers facilitated workshops that align business and IT leaders on the right priorities and path forward, to rapid, cost-effective pilots that demonstrate real use case outcomes to validate and de-risk investments. Dell paves the way to a clear, proven, and seamless transition to scalable solutions in production, optimized to your unique needs and driving value from the start.

The path to successful AI adoption begins with a clear strategy. Rely on proven methodology to build a solid foundation that gains alignment between business and IT, prioritizing high-impact use cases and creating a detailed roadmap to achieving outcomes. Dell Professional Services can help you establish robust data practices to ensure your data is effectively leveraged and managed, while ensuring data security, reliability, and access to fully power your AI initiatives.

Dell Professional Services expertise extends far beyond hardware deployment, implementing, and integrating complete full-stack AI solutions. They streamline the setup of complex software, tools, and frameworks required for agentic AI, including model development, data pipelines, and RAG workflows. Dell Professional Services focus not only on new architecture, but they also ensure that solutions are seamlessly integrated into your existing environment and workflows. And their experts work hand-in-hand to empower your teams with new skills to easily transition and adopt new systems and processes to maximize ROI.

Adopting advanced AI technologies introduces new operational demands. Combined with knowledge gaps and stretched resources, many organizations are at risk of seeing stalled progress and limited or misaligned outcomes. Dell Services removes the burden to simplify AI, up-leveling capabilities and freeing internal resources from routine activities. Dell Managed Services offer comprehensive, full-stack management and seamless integration across hardware, platforms, and workloads—ensuring optimal performance and reliability.

By partnering with Dell Professional Services, you gain a trusted advisor to turn AI ambitions into enterprise impact. Backed by extensive global expertise and firsthand experience with a vast number of organizations and industries, they know what it takes to succeed and will partner with you to ensure confident pursuit of AI.

6

Conclusion

Topics:

Summary and next steps

The enterprise race to operationalize agentic and generative AI is intensifying. To meet this demand, Dell Technologies—backed by NVIDIA and a growing ecosystem of ISV partners—introduces a modular framework that is designed to accelerate deployment, reduce integration overhead, and enforce a security-first posture. Furthermore, Dell Technologies ensures that the needs of the business are aligned to the plans of the technology, providing expert guidance and end-to-end support

This document introduces a framework for building an enterprise AI factory architecture for agentic applications, with a first set of recommendations to enable enterprise businesses to deploy LLMs and build and run agents while satisfying the business and operational needs for observability, security, and governance. By evolving this framework in collaboration with NVIDIA, Dell Technologies aim to integrate technological advancements, operational best practices, and enabling a growing spectrum of enterprise cases – establishing a robust foundation that accelerates time-to-value and enables scalable deployment of AI capabilities.

The framework is composed of two core layers: AI infrastructure and AI enablement. The enablement layer delivers the software backbone for production AI. It includes Agent Ops powered by NVIDIA AI Blueprints, AI software platforms for managing the life cycle of models and AI applications, vector databases for semantic search, observability for full-stack telemetry, and embedded security controls. These functions connect to enterprise systems through data connectors and the Dell AI Data Platform. A subset of these functions is implemented by AI Workloads—a curated, validated and continuously expanded list of ISV solutions to address evolving use cases.

The infrastructure layer consolidates compute, networking, and storage under Kubernetes orchestration, with GitOps and artifact repositories ensuring consistency and traceability. Its core components are pre-integrated and validated across multiple configurations of Dell AI Factory with NVIDIA—commercially available systems engineered for enterprise-grade AI.

Dell Automation Platform serves as the operational engine, enabling blueprint-driven deployment and life cycle automation. It supports turnkey implementations of Dell AI Platforms via Dell Professional Services, as well as self-service AI workload deployment by enterprise teams.

By combining Dell Technologies infrastructure and automation platforms with NVIDIA's AI frameworks, tools, and ISV workloads, this reference architecture empowers organizations to accelerate AI adoption, simplify integration, and ensure performance, security, and interoperability across the full life cycle of AI deployment. This document provides a foundation for operationalizing AI at scale—with the flexibility to support diverse use cases and the reliability to meet enterprise standards.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document and the information that it contains. Contact the Dell Technologies Solutions team by email.

Authors: Fabio Souza (AI Solutions Technical Marketing Engineering, Dell Technologies), Norbert Purger (AI Solutions Product Management, Dell Technologies)

Contributor: Scott Powers (AI Solutions Technical Marketing Engineering, Dell Technologies), Justin King (Enterprise Product Marketing, NVIDIA)

7

References

Topics:

Dell Technologies documentation

- Dell Technologies Info Hub for AI Solutions
- Security Best Practices for Generative AI in the Enterprise
- Data Center Network Solutions
- PowerEdge AI Servers
- PowerScale
- Dell Enterprise Hub by Hugging Face
- Dell AI Data Platform
- Generative AI in the Enterprise with NVIDIA GPUs, Networking, and Software Stack, and Red Hat OpenShift design guide
- Dell Automation Platform portal

NVIDIA documentation

- NVIDIA Enterprise Reference Architecture
- NVIDIA Enterprise AI Factory Design Guide
- NVIDIA Run:ai
- NVIDIA AI Enterprise

Red Hat ecosystem documentation

- OpenShift Container Platform documentation
- Argo CD in Red Hat documentation
- HAProxy chapter in Red Hat Enterprise Linux documentation

AI Workloads documentation

- DataRobot
- Elastic blog
- Elasticsearch and Prometheus monitoring
- JFrog Artifactory
- Dell Technologies and CrowdStrike partnership announcement
- Joint Dell Technologies and CrowdStrike offering to boost cyber defense