



Lenovo Validated Design for IBM watsonx

Last update: **01 May 2025**

Version 1.0

**Pre-tested, scalable AI
platform optimized for IBM
watsonx**

**Enterprise-Grade AI
Performance & Security with
Lenovo ThinkSystem, NVIDIA
GPUs & networking**

**Quickly and safely, introduce
GenAI solutions with
transparency and cost control**

**Factory rack integrated
solution delivered through
EveryScale**

Hapsara Sukasdadi (Lenovo)
Pierce Beary (Lenovo)
Fahim Shaikh (IBM)



Table of Contents

1	Introduction.....	1
2	Business problem and business value	2
2.1	Business problem	2
2.2	Business value.....	2
3	Requirements.....	3
3.1	Functional requirements	3
3.2	Non-functional requirements.....	4
4	Watsonx Architectural overview.....	6
4.1	High Level Architecture	6
4.1.1	Architecture Components	6
5	Hardware and Software Components	10
5.1	Hardware Components	10
5.2	Software Components	12
5.2.1	IBM watsonx	12
5.2.2	Red Hat OpenShift Container Platform	14
6	Dimensioning.....	17
6.1	Scaling the IBM Watsonx Platform.....	17
6.1.1	Single Rack Deployment (8-24 GPUs).....	17
6.1.2	Scalable Unit Deployment	17
6.1.3	Scalability of IBM watsonx	19
7	Validation Process and Results.....	20
7.1	Solution Validation Results.....	20
7.1.1	Use Case 1: Comparing Large Language Models	20
7.1.2	Use Case 2: Wealth manager	21
8	Appendix: Lenovo Bill of materials	23
8.1	Single Rack Deployment.....	23

8.1.1	ThinkSystem SR675 V3	23
8.1.2	ThinkSystem SR635 V3	24
8.1.3	NVIDIA SN3700 Switch	24
8.1.4	NVIDIA SN2201 Switch	24
8.2	Scalable Unit (SU) Deployment	25
8.2.1	ThinkSystem SR675 V3	25
8.2.2	ThinkSystem SR655 V3	26
8.2.3	NVIDIA SN5600 Switch	26
8.2.4	NVIDIA SN2201 Switch	26
8.2.5	XClarity Software	27
Resources.....		28
Document history.....		29
Trademarks and special notices		30

1 Introduction

This Lenovo Validated Design (LVD) presents a detailed architectural blueprint for integrating IBM WatsonX with Lenovo Hybrid AI 285 using Red Hat OpenShift. The document targets technical architects, infrastructure engineers, and IT stakeholders planning to deploy enterprise-grade generative AI solutions.

This document serves as a comprehensive guide for the configuration and deployment of watsonx with Lenovo Infrastructure, detailing its architecture, key components, performance benchmarks, and deployment recommendations.

Intended Audience:

- Enterprise AI Architects & Data Scientists: Seeking an optimized and scalable AI infrastructure for handling structured and unstructured data, including video, audio, and text.
- IT and Data Engineering Teams: Responsible for deploying and managing AI workloads in data centers, edge environments, and hybrid cloud infrastructures.
- Lenovo and IBM Partners: Engaged in designing, integrating, and scaling AI solutions with a validated and repeatable deployment model. With a focus on simplifying AI deployment, Lenovo's pre-validated configurations for IBM watsonx ensure that customers can efficiently manage AI workloads, accelerate time-to-insight, and drive innovation across multiple industries.

2 Business problem and business value

2.1 Business problem

Organizations face critical challenges adopting Generative AI, including unpredictable infrastructure costs, complex deployment scenarios, inadequate scalability, and unclear return on investment. These challenges frequently hinder AI deployment efforts, causing delays, increased costs, and suboptimal utilization of resources.

- Quickly and safely, introduce GenAI solutions with transparency and cost control
- Reduce risk, transparent costs and transparent AI, simplifying user journey
 - Currently, in the AI market, there are real issues around properly sizing the investment necessary to run Gen AI at scale. This leads many customers to make significant investments without having a clear view of the economics behind the solutions. Even those confident about the software skills might struggle to translate that into efficient infrastructure.
 - The ability to bring all of it under a pre-defined solution size allows customers to feel confident about their Gen AI deployments, with measurable RoI, lower risk and better TTV.

Lenovo's watsonx solutions provides customers with advanced tools to train, validate, and deploy AI models. watsonx, with its multi-model flexibility, robust client protection, end-to-end AI governance, and hybrid multi-cloud deployments, promises to revolutionize the AI landscape. By incorporating watsonx, Lenovo aims to accelerate AI adoption, offering clients unparalleled AI capabilities and a transformative impact on their business operations.

IBM and Lenovo are partnering to deploy Gen AI on a number of different server platforms, to give clients the option for whatever best fits their needs. This includes the already launched watsonx on IBM Fusion Hyperconverged Infrastructure, powered by Lenovo.

2.2 Business value

This tested solution from Lenovo and IBM addresses these challenges by providing a predictable, secure, scalable, and highly manageable platform. WatsonX's comprehensive suite of AI tools accelerates model development, governance, and deployment processes. By leveraging Red Hat OpenShift, businesses can rapidly deploy scalable and flexible AI solutions, significantly accelerating their return on investment.

3 Requirements

The functional and non-functional requirements for this reference architecture are described below.

3.1 Functional requirements

- **Agentic workflows** - We are living through a technology shift where assistants, agents, and intelligent automation are coming together to drive enterprise productivity. Granite models combined with watsonx.ai agentic tooling supports the development, deployment, and monitoring of agents for tasks such as HR, customer service, sales, procurement, and more.
- **Build language-based tasks** such as retrieval-augmented generation or RAG – RAG retrieves data from outside a foundation model and augments prompts by adding the relevant data such as a document repositories, databases, or application programming interfaces (APIs). For example, RAG with Granite can help build a broader knowledge base for customer service assistance and ultimately improve your customer experience.
- **Use Granite for code** which will speed up application modernization efforts and generate automation to rapidly scale IT environments. With Granite powering assistants such as watsonx Code Assistant developers can create high-quality code using AI-generated recommendations based on natural language requests or existing source code.
- **Build time series forecasts** - Granite can not only incorporate historical sales data but also near-real-time data from advertising campaigns, prices, local weather forecasts, and other sources to make informed decisions. Example: Enable better inventory allocation, assortment, and replenishment planning for retailers.
- **Uncover patterns and trends in geographic data** - IBM and NASA partnered to create targeted forecasts based on local observations, detecting and predicting severe weather patterns, improving the spatial resolution of global climate simulations, and improving how physical processes are represented in numerical weather and climate models.
- **Use Granite to safeguard AI** - How do you ensure the models being used can be trusted, can scale easily and keep operating costs low? For highly regulated industries such as financial services, healthcare, and government, Granite ensures transparency of training data, within built guardrails with Granite Guardian to drive responsible AI and meet stringent compliance and regulatory requirements.

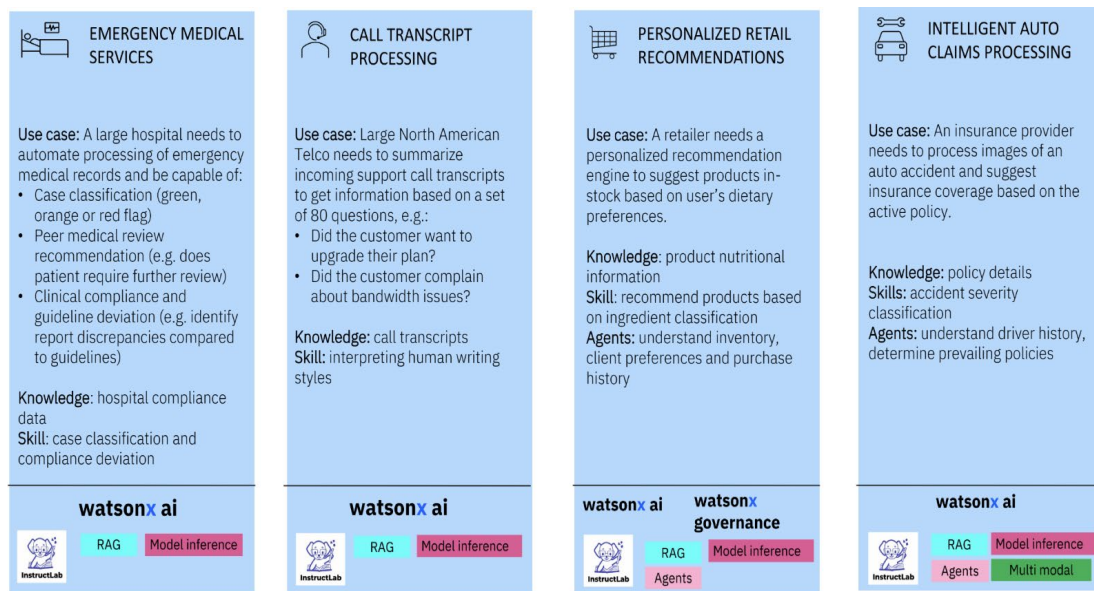


Figure 1: IBM watsonx

3.2 Non-functional requirements

Non-functional requirements define the operational characteristics of a system, ensuring it meets performance, scalability, reliability, security, and manageability standards.

- **Performance & Scalability**
 - **Inference Speed:** The system should provide low-latency inference, with target response times defined based on specific use cases.
 - **Throughput:** The system should be able to handle a high volume of inference requests concurrently, ensuring efficient processing of multiple video streams or user interactions.
 - **Scalability:** The platform should be able to scale horizontally to accommodate increasing workloads and data volumes. This includes scaling the number of GPU nodes, microservices, and database capacity.
- **Availability & Reliability**
 - **Availability:** The system should be highly available, with minimal downtime. Redundancy mechanisms and fault-tolerance strategies should be implemented to ensure continuous operation.
 - **Data Integrity:** The platform should ensure the integrity and consistency of data throughout the AI pipeline, from ingestion to processing and storage.
- **Manageability & Orchestration**
 - o **Orchestration:** The system should be deployable on Kubernetes platforms, such as Red Hat OpenShift, for automated deployment, scaling, and management.
 - o **Monitoring and Logging:** The platform should provide comprehensive monitoring capabilities, including real-time dashboards and alerts, to track system health, performance, and resource

utilization. Detailed logs should be generated to facilitate troubleshooting, performance analysis, and security auditing.

- o **Administration:** The system should provide tools and interfaces for managing and administering the platform, including configuration, deployment, and updates.
- o **Integrated:** Enterprise IT infrastructure is complex and varies by organization. Generative AI products must seamlessly integrate into existing infrastructure and adapt to various ecosystems. This gives organizations choice of where their data lives and which tools they use to drive productivity and scale. Watsonx integrates with your existing tech stack to avoid disruptions to your business when building and scaling your AI. Bring your own data, models and infrastructure and customize our AI for your specific enterprise use case.
- o **Open:** Open source is a rich source of innovation and is a rich source of many FMs (such as those available on Hugging Face). A generative AI platform needs to be open – or it will restrict what is available to its practitioners and lock them in. One model will not rule them all, which is why watsonx is fundamentally open, allowing access to data in any format, across any cloud, and open, customizable models.

- **Security**

- **Data Security:** The platform should protect sensitive data, including video streams, user data, and AI models, through encryption, access controls, and secure storage.
- **Network Security:** Secure network configurations and protocols should be implemented to prevent unauthorized access and protect against cyberattacks.
- **Authentication and Authorization:** Strong authentication and authorization mechanisms should be in place to control access to the system and its resources.
- **Trusted:** There are already well-known issues with generative AI platforms like ChatGPT or applications like Google Bard. Generative AI models are known for hallucination (making up answers). Generative AI platforms need to be trusted, offering security and data and model protection. They also need to be architected with governance in mind from the start, not an afterthought and provide transparency, and explainability for their models to support increasing regulatory compliance demands. Watsonx.governance applies throughout the AI model lifecycle. We also offer IP indemnification of our models and share the data and training methods used to build them.

4 Watsonx Architectural overview

4.1 High Level Architecture

The diagram describes the layered high level architecture, from the infrastructure layer, platform layer, up to the application layer that includes model development, serving, and monitoring. The architecture is based on the Lenovo Hybrid AI 285 platform and highlighted an end-to-end, cloud-native approach in deploying AI workloads, from data science pipelines to production grade generative AI applications. The AI workloads are deployed using Kubernetes container orchestration platform.

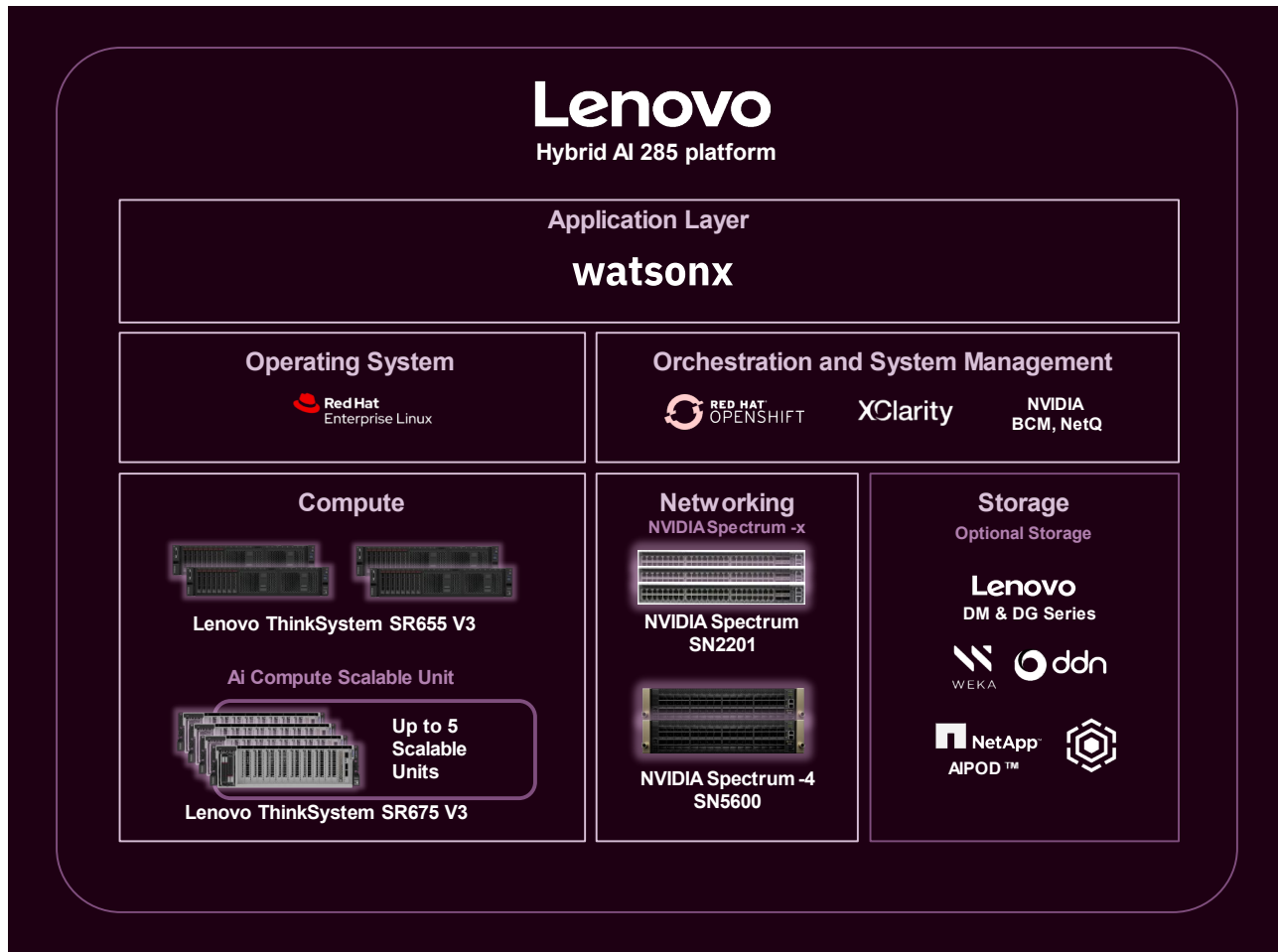


Figure 2: High Level Architecture

4.1.1 Architecture Components

This section provides a more detailed description of the high-level software architecture.

- **Compute:**
 - This represents the physical compute resources and processing power of the system. Lenovo ThinkSystem servers are the hardware foundation for running and managing the AI workloads. Key aspects of compute are Central Processing Unit (CPU), Random Access Memory (RAM), and GPU acceleration. Compute consists of Service Nodes and AI Compute Nodes.

- **Networking:**

- A high-performance, low latency network is critical for AI systems, which often involve massive data transfers between compute nodes and storage. Key requirements for network are high bandwidth to enable fast data transfer rate, low latency to minimize delays in data communication, and connectivity between other components in the architecture.
- NVIDIA Spectrum-x compliant. Features a high-performance converged network fabric in a full non-blocking fat-tree topology, this RDMA-compliant design optimizes multi-GPU communication within and across nodes.
- The high-level networking design is shown below:

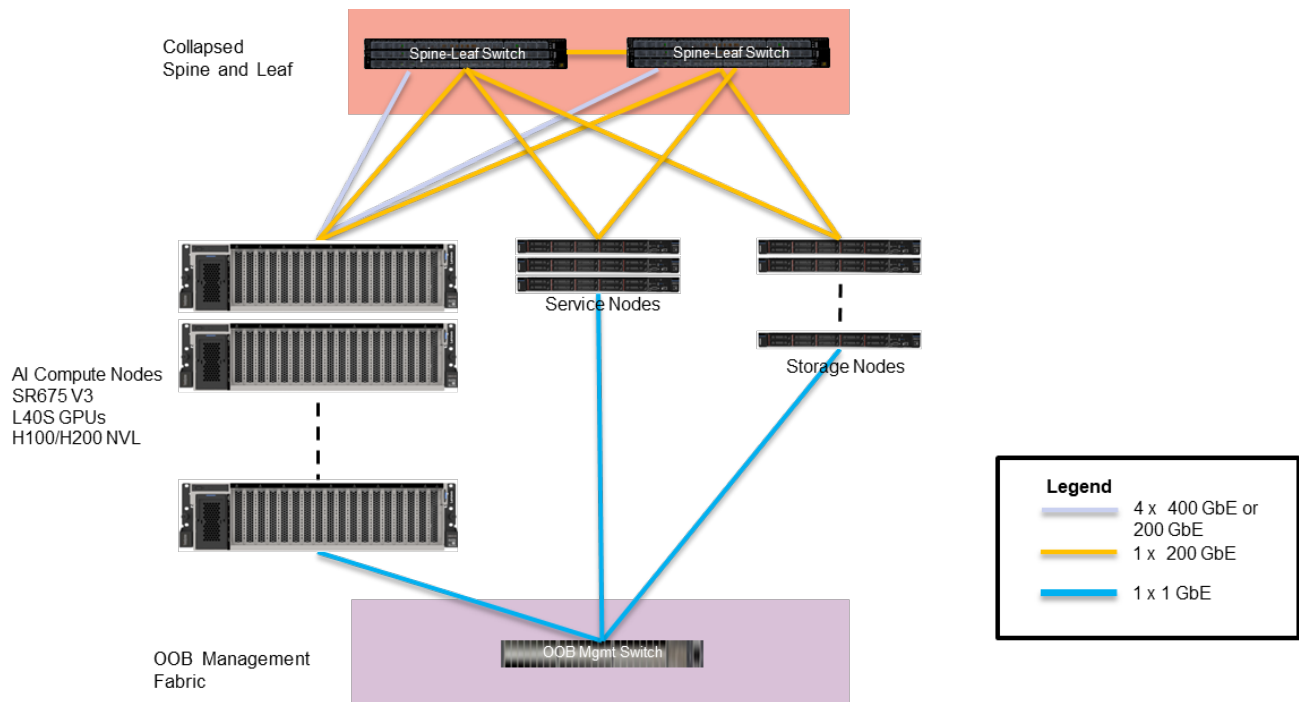


Figure 3: High Level Network Design

- **Compute Fabric (East-West):** Optimized for inter-GPU connectivity within a server and across a rail. This fabric uses 400 GbE connections for Scalable Unite (SU) deployment or 200 GbE connections for Single Rack deployment.
- **Converged Fabric (North-South):** Supports storage and in-band management with 200 GbE high-speed connections between GPU Worker Nodes, Service Nodes, and External Storage to spine-leaf switches.
- **Out-Of-Band (OOB) Management Fabric:** Connects all BMC ports and isolated management devices, including 1 GbE switch management ports.

- **Storage:**

- The architecture incorporates an optional high-performance external storage solution, seamlessly integrated into the network fabric to support demanding AI workloads, such as video stream storage and real-time data processing. This storage infrastructure is designed to deliver the high input/output operations per second (IOPS) and throughput necessary to keep pace with intensive compute demands, ensuring low-latency data access for AI inference and training.
 - Additionally, the storage solution is scalable and supports data lake integration, enabling efficient data management and analytics across AI-driven workflows. It is fully compatible with Lenovo DM & DG series storage solutions, and other storage solutions such as DDN, NetApp, VAST, and WEKA, providing enterprise-grade reliability and performance.
 - For this LVD, storage nodes are not within the scope of test and validation, but the architecture is designed to accommodate future enhancements and enterprise storage integration as needed.
- **Operating System:**
 - **Red Hat Enterprise Linux:** Red Hat Enterprise Linux (RHEL) is a stable, secure, and high-performing Linux operating system specifically designed for enterprise environments. Renowned for its reliability and long-term support, RHEL provides a robust foundation for running critical business applications and workloads. Its strong security features, extensive ecosystem of certified hardware and software, and enterprise-grade management capabilities make it a trusted choice for organizations demanding stability, security, and performance from their operating system infrastructure.
- **Orchestration and System Management:**
 - **Red Hat OpenShift:** Red Hat OpenShift is a robust, enterprise-grade container orchestration platform built upon the foundation of Kubernetes. It simplifies the deployment, management, and scaling of containerized applications, offering a developer-friendly experience with integrated tools and workflows. Designed for hybrid cloud environments, OpenShift provides enhanced security, automation, and operational efficiency, making it a powerful platform for modern application development and deployment across diverse infrastructures.
 - **XClarity:** Lenovo XClarity Administrator is a centralized, resource-management solution that simplifies infrastructure management, speeds responses, and enhances the availability of Lenovo server systems and solutions. It runs as a virtual appliance that automates discovery, inventory, tracking, monitoring, and provisioning for server, network, and storage hardware in a secure environment.
 - **NVIDIA Base Command Manager (Optional):** NVIDIA Base Command Manager (BCM) is a powerful management tool designed to optimize the utilization of NVIDIA GPU resources in high-performance computing environments. It provides administrators with centralized control over GPU allocation, monitoring, and health management, ensuring efficient resource sharing among users and workloads. By enabling features like job scheduling and integration with system

management tools, BCM maximizes GPU utilization and contributes to the stability and performance of demanding AI and data science applications.

- **NVIDIA NetQ (Optional):** NVIDIA NetQ is a powerful network telemetry and management tool designed for modern data center fabrics, particularly those built with NVIDIA Spectrum Ethernet switches. It provides real-time visibility into network performance, enabling proactive monitoring, troubleshooting, and analysis of network behavior. By offering comprehensive insights into network health, traffic patterns, and potential bottlenecks, NetQ helps optimize network efficiency, reduce downtime, and ensure the reliable operation of high-performance applications, including AI and machine learning workloads,

- **AI Application Layer:**

- **IBM Watsonx:** IBM Watsonx is an enterprise-focused AI and data platform designed to empower organizations to build, deploy, and scale AI applications across their business. It integrates a comprehensive suite of tools and services, including foundation models, machine learning capabilities, and data governance features, all within a unified environment. By providing access to cutting-edge AI technologies and streamlining the AI lifecycle, Watsonx aims to accelerate AI adoption, drive innovation, and unlock valuable insights from data for businesses of all sizes.

5 Hardware and Software Components

5.1 Hardware Components

This section provides a detailed overview of the compute, GPU, network, and storage hardware components integrated into this LVD, ensuring optimal performance, scalability, and reliability for the intended use case.

Table 1: Hardware Components





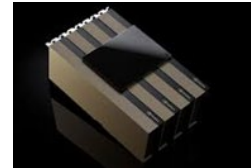





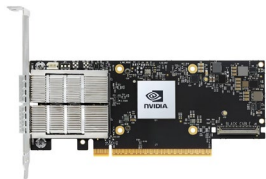
Image	Lenovo Server Model	CPU	GPUs/DPUs	Role In Design
	ThinkSystem SR675 V3 3U GPU-rich rack server	5th Gen AMD EPYC 9005 'Turin'	Supports up to 8 GPUs (NVIDIA H200 NVL, L40S) with NVLink and Lenovo Neptune cooling	AI compute nodes Worker nodes equipped with GPUs
	ThinkSystem SR655 V3 2U single-socket server	5th Gen AMD EPYC 9005 'Turin'	Equipped with NVIDIA BlueField-3 B3220 DPU for larger clusters and high-performance storage	Service nodes General-purpose servers handle resource allocation, monitoring, orchestration, and security management.
	ThinkSystem SR635 V3 1U single-socket server	5th Gen AMD EPYC 9005 'Turin'	Equipped with NVIDIA ConnectX-7 SmartNIC in single rack deployments	Service nodes General-purpose servers handle resource allocation, monitoring, orchestration, and security management.

Table 2: GPU and Networking

Image	Category	Component	Description
	GPU	NVIDIA L40S GPU	Built on Ada Lovelace architecture, optimized for AI, graphics, and video workloads.
	GPU	NVIDIA H200 NVL GPU	Accelerates AI and HPC with 141GB HBM3e memory and 4.8 TB/s bandwidth. Includes a 5-year NVIDIA AI Enterprise license.

	Networking	NVIDIA SN5600 Switch	64 ports at 800 Gbps, with flexible breakout options. Used in larger clusters with NVIDIA Spectrum-X networking.
	Networking	NVIDIA SN2201 Switch	48 ports for out-of-band (OOB) management, centralizing compute, storage, and network connectivity.
	Networking	NVIDIA SN3700V Switch	32 ports at 200 Gbps. Used in smaller clusters or when cost optimization is needed.
	SmartNIC/DPU	NVIDIA BlueField-3 B3220 DPU	Dual 200 GbE ports, required for optimal performance in Scalable Unit (SU) deployments.
	SmartNIC/DPU	NVIDIA BlueField-3 B3140H SuperNIC	Single 400 GbE QSFP112 port, used in Spectrum-X compute fabric for high-performance and large clusters.
	SmartNIC/DPU	NVIDIA ConnectX-7 NDR200 InfiniBand/200GbE QSFP112 Adapters	Dual 200 GbE ports, required for single rack deployments.

5.2 Software Components

5.2.1 IBM watsonx

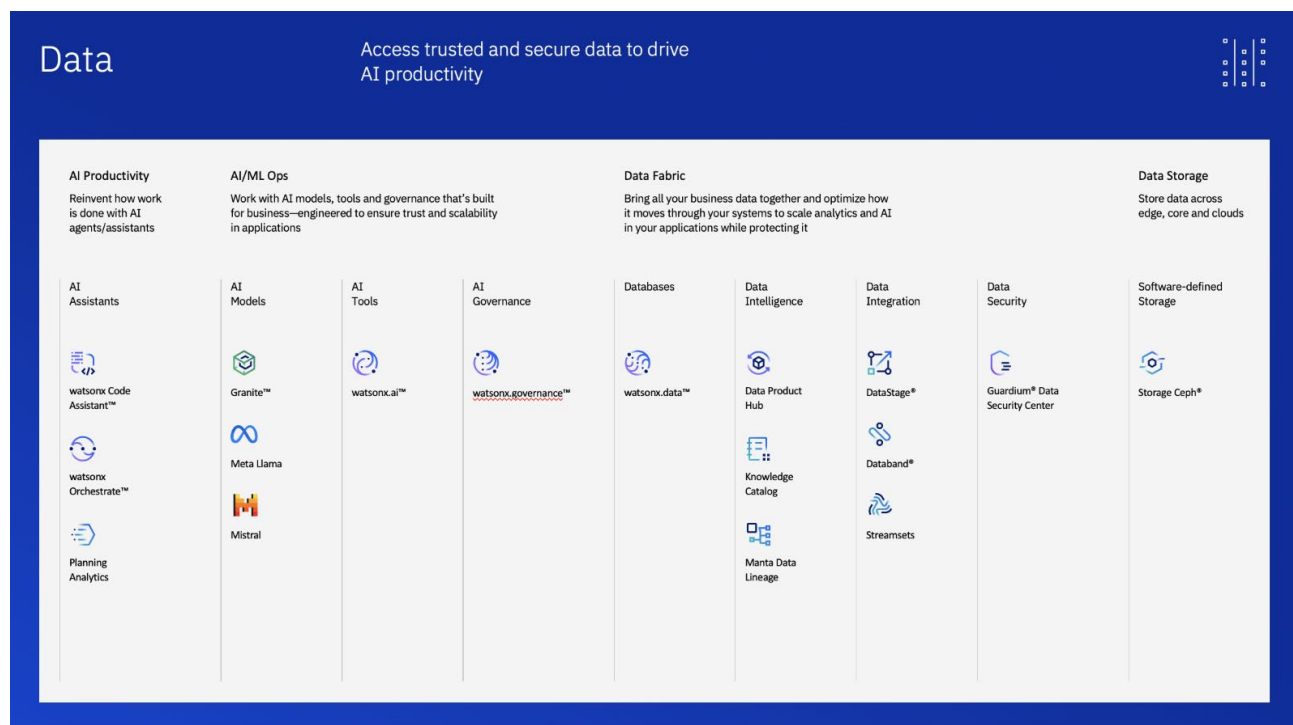


Figure 4: IBM watsonx 5.1.2

IBM watsonx is a powerful AI and data platform that enables businesses to **train, validate, tune, and deploy AI models at scale**—all with **governance, transparency, and control** at its core. In this LVD, IBM watsonx 5.1.2 is used. IBM watsonx 5.1.2 comes with all the components listed on the figure above.

It consists of five main layers:

1. Data Layer

- **IBM watsonx.data:** A data lakehouse platform that enables scalable analytics and AI workloads by combining the flexibility of data lakes with the performance of data warehouses.
- **Cloud Pak for Data:** An integrated data and AI platform built on Red Hat OpenShift, allowing organizations to collect, organize, and analyse data to infuse AI throughout their business.

2. Model Development Layer

This layer is where AI models are designed, trained, and fine-tuned.

- **IBM watsonx.ai:** An enterprise-grade AI studio that provides tools for training, validating, tuning, and deploying AI models, including foundation models and machine learning capabilities.
- **Open-Source Frameworks:** IBM supports various open-source tools such as TensorFlow, PyTorch,

and Scikit-learn for model development.

3. Model Deployment Layer

This layer involves deploying trained models into production environments.

- **Red Hat OpenShift AI:** A platform that provides tools for deploying and managing AI models in a scalable and secure manner.
- **KServe and KubeFlow:** Open-source platforms integrated into IBM's stack for serving and managing machine learning models.

4. Application Layer

This is the user-facing layer where AI models are integrated into applications.

- **watsonx Assistant:** A conversational AI platform that enables the creation of chatbots and virtual assistants for customer service and other applications.
- **watsonx Orchestrate:** A tool that uses natural language processing to automate tasks by orchestrating workflows based on user input.

5. Observability and Governance Layer

Ensures that AI systems are transparent, accountable, and compliant with regulations.

- **IBM watsonx.governance:** Provides tools for managing risk, compliance, and the AI lifecycle, including model monitoring, bias detection, and explainability.

Table 3: Stack Layer Tools and Reference Links

Stack Layer	Tool	Link
Data	watsonx.data	ibm.com/products/watsonx-data
Data	Cloud Pak for Data	ibm.com/products/cloud-pak-for-data
Model Development	watsonx.ai	ibm.com/products/watsonx-ai
Deployment	OpenShift AI Deployments	ibm.com/cloud/watsonx/openshift
Applications	watsonx Assistant	ibm.com/products/watsonx-assistant
Applications	watsonx Orchestrate	ibm.com/products/watsonx-orchestrate
Governance	watsonx.governance	ibm.com/products/watsonx-governance

Key Advantages of IBM watsonx

Enterprise Trust & Governance

- Ensures **AI transparency and compliance** with regulations (e.g., EU AI Act).
- Integrated tools for **model lineage, bias detection, risk assessment**, and auditability.
- Custom policies to meet specific industry requirements.
- Access 1000+ models in watsonx.ai from a mix of platform-provided models and custom foundation models available for import

Open & Flexible

- Built on open standards and supports **open-source models** (Meta's LLaMA, Mistral, Hugging Face models, etc.).
- Compatible with multiple languages, frameworks, and environments.

Accelerated AI Development

- Pre-trained foundation models reduce time-to-market.
- Features like **prompt tuning, parameter-efficient tuning**, and **synthetic data generation** for rapid customization.
- Integration with Red Hat OpenShift for container orchestration and scalability.

Unified Data for AI

- watsonx.data combines **data warehouse performance** with **data lake economics**.
- Connects structured and unstructured data sources into a governed platform.
- Optimized for high-performance queries and AI model training.

Build Once, Deploy Anywhere

- Models can be trained once and deployed across environments, including edge and private clouds.
- Supports REST APIs and inference endpoints for seamless integration with business apps.

Use Cases Across Industries

- **Finance**: Detect fraud, automate loan processing, and provide AI-powered insights.
- **Healthcare**: Accelerate clinical research, automate records processing, and support diagnostics.
- **Retail**: Enhance customer service with chatbots, personalize recommendations, and optimize supply chains.
- **Manufacturing**: Predict equipment failures, optimize quality control, and streamline operations.

5.2.2 Red Hat OpenShift Container Platform

OpenShift provides a robust and flexible platform for deploying and managing applications, offering benefits

such as simplified deployment, scalability, automation, security, portability, integration with DevOps practices makes it empowers organizations to build and operate applications effectively in diverse environments. And seamlessly integrates with NVIDIA GPU technology, unlocking unparalleled performance and efficiency for containerized workloads across diverse industries.

Here are some key benefits of using OpenShift:

Simplified Application Deployment: OpenShift simplifies the deployment process by leveraging containerization technology, such as Docker and Kubernetes. It allows developers to package applications into lightweight and portable containers, making them easy to deploy across different environments consistently.

Scalability and Elasticity: OpenShift enables organizations to scale applications horizontally and vertically based on demand. It can automatically adjust the number of container instances to handle increased traffic or workload, ensuring optimal performance and resource utilization.

Automation and Orchestration: OpenShift automates many aspects of application management, such as provisioning, scaling, and load balancing. It provides powerful orchestration capabilities through Kubernetes, allowing administrators to define complex deployment workflows and easily manage application lifecycles.

Multi-Platform Support: OpenShift offers flexibility in deploying applications across various cloud providers, including public, private, and hybrid cloud environments and on-premises. It provides a consistent development and deployment experience, allowing applications to run seamlessly across different infrastructure setups.

Enhanced Security: OpenShift incorporates built-in security features, such as isolation between containers, resource constraints, and role-based access control (RBAC).

DevOps Integration: OpenShift promotes DevOps practices by facilitating collaboration between development and operations teams. It provides tooling and features for continuous integration and delivery (CI/CD), enabling streamlined application development, testing, and deployment processes.

Application Portability: With OpenShift, applications built and deployed in one environment can be easily moved to another. The use of containerization ensures that applications have consistent behavior across different platforms and eliminates dependency issues, making migration and maintenance more efficient.

Extensive Ecosystem: OpenShift benefits from a large and vibrant ecosystem of applications, services, and plugins. It integrates with various tools and technologies, enabling developers to leverage a wide range of capabilities and services, such as databases, messaging systems, monitoring, and logging solutions.

Multi-tenancy and Isolation: OpenShift supports multi-tenancy, enabling the isolation of resources and applications between different teams or projects. This is crucial in enterprise environments where multiple teams might be working on diverse projects.

GPU Support: OpenShift with NVIDIA GPU support allows organizations to harness the power of NVIDIA

GPUs for accelerating compute-intensive workloads, such as machine learning, deep learning, data analytics, and scientific simulations.

NVIDIA GPU Operator Support: The NVIDIA GPU Operator is a key component that facilitates the integration of NVIDIA GPUs into OpenShift clusters. It simplifies the deployment and management of GPU resources, making it easier for administrators to configure and scale GPU-accelerated workloads.

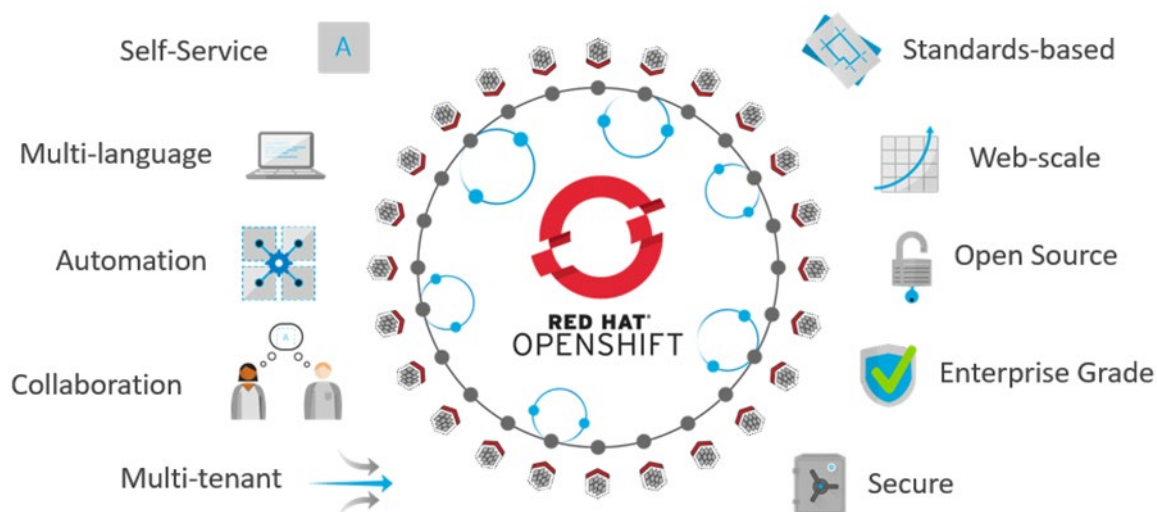


Figure 5: Red Hat OpenShift

6 Dimensioning

6.1 Scaling the IBM Watsonx Platform

Offering flexible deployment options, the IBM Watsonx Platform can be configured as single rack deployment between 8 to 24 GPUs or expanded to a multi-rack setup using the Scalable Unit deployment. Both support the RAG, Secure Code Assistant, and Document Understanding applications.

6.1.1 Single Rack Deployment (8-24 GPUs)

Single rack deployments consist of one to three AI compute nodes, outfitted with up to 24 DW PCIe NVIDIA GPUs (L40s, H200 NVL, H100 NVL, or RTX Pro 6000). It is ideal for deployments that have no plan to grow beyond 24 GPUs, allowing reduced hardware costs and complexity.

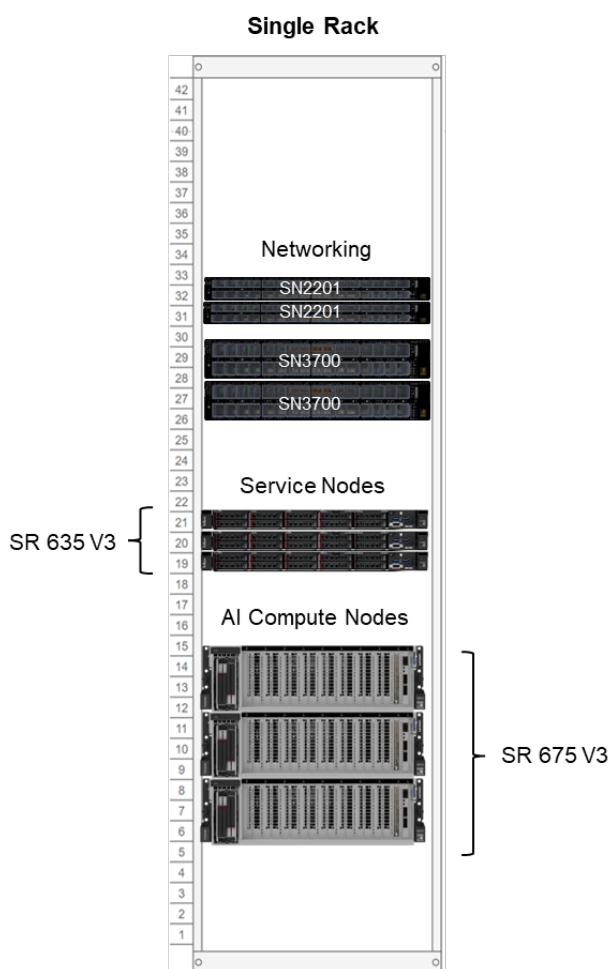


Figure 6: Single Rack Deployment

6.1.2 Scalable Unit Deployment

Scalable Unit deployment supports AI workloads across multiple racks, starting with one Scalable Unit (SU). The first SU consists of up to 4 worker nodes, minimum 3 service nodes, and networking switches. The first SU will support up to 32 GPUs. When additional nodes are required, additional SUs of 4 nodes can be added.

When additional nodes are required, additional SUs of 4 nodes can be added. An example bill of material for a single SU deployment is provided in section Appendix: Bill of material.

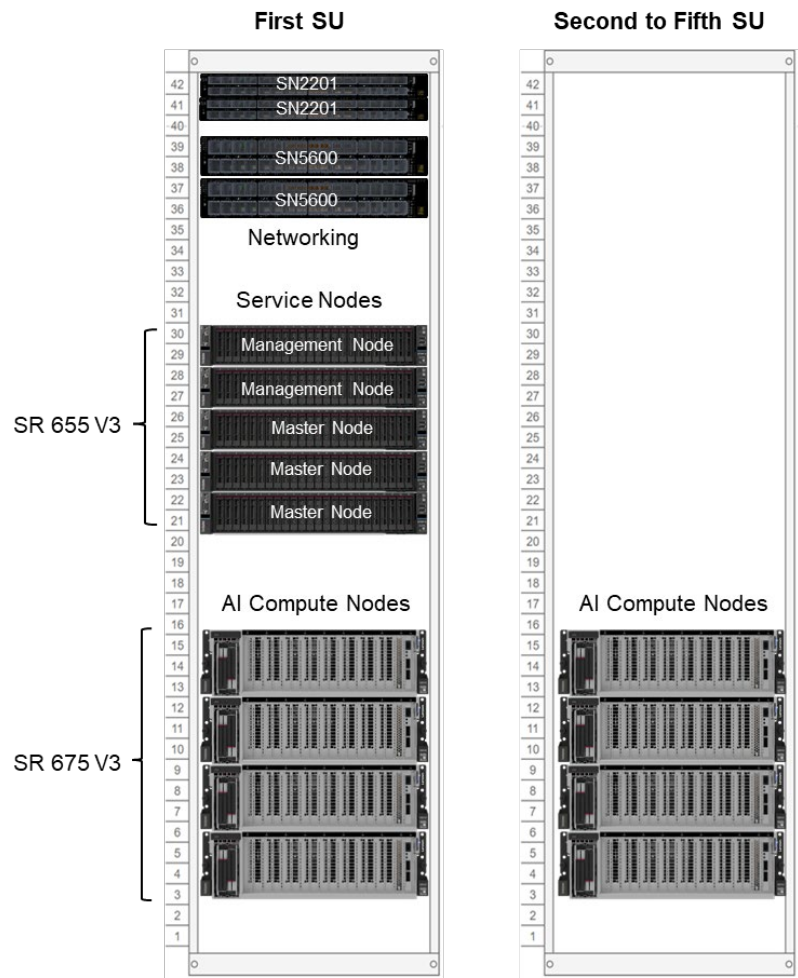


Figure 7: Scalable Unit (SU) Deployment

For both single rack and Scalable Unit deployments, Red Hat OpenShift provides the Kubernetes orchestration layer, requiring at least three service nodes as master nodes (Kubernetes services, Etc, and api), and three service nodes as infra nodes (Route, Registry, Logging, and Monitoring). One master node will be temporarily used as OpenShift bootstrap, then, once the installation of the OpenShift cluster is completed, will be added as one of the master nodes. Optional management nodes for NVIDIA BCM and NetQ can also be added for management and orchestration of the cluster..

6.1.3 Scalability of IBM watsonx

Table 4: IBM watsonx dimensioning examples

Model	Configuration	8 GPUs	16 GPUs	24 GPUs
granite-3-8b-instruct Granite models are designed to be used for a wide range of generative and non-generative tasks with appropriate prompt engineering. They employ a GPT-style decoder-only architecture, with additional innovations from IBM Research and the open community.	2 CPU 96 GB RAM 20 GB Storage	Infrastructure: Sized to support a single Granite-3-8b-instruct instance using 8 NVIDIA A100 or H100 80GB GPUs. This setup is ideal for smaller-scale development, testing, or light inference workloads. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Small Granite-3-8b-instruct instance (8 GPUs) (Total: 8 GPUs used for model serving) User Capacity: Supports up to 25 concurrent users in the Prompt Lab 1–2 active model deployments 1–2 active JupyterLab or notebook environments Use Cases: Internal chatbot for IT helpdesk (e.g., password reset, VPN issues) Fine-tuning on internal documentation to create a private support assistant Lightweight email response drafting tool for sales teams	Infrastructure: Sized to support a single Granite-3-8b-instruct instance using 16 NVIDIA A100 or H100 80GB GPUs. This configuration is suited for moderate production inference workloads and higher concurrency. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Medium Granite-3-8b-instruct instance (16 GPUs) (Total: 16 GPUs used for model serving) User Capacity: Supports up to 100 concurrent users in the Prompt Lab 3–4 active model deployments 3 active JupyterLab or notebook environments Use Cases: Legal document summarization for compliance teams Conversational interface over Confluence or SharePoint content Interactive support assistant trained on a company's product manuals	Infrastructure: Sized to support a single Granite-3-8b-instruct instance using 24 NVIDIA A100 or H100 80GB GPUs. Ideal for enterprise-grade, high-throughput serving or large-scale tuning and inference. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Large Granite-3-8b-instruct instance (24 GPUs) (Total: 24 GPUs used for model serving) User Capacity: Supports up to 250 concurrent users in the Prompt Lab 5 active model deployments 5 active JupyterLab or notebook environments Use Cases: AI-powered customer support chatbot deployed on public-facing site Multilingual employee onboarding assistant trained on HR documents Scalable internal LLM platform serving 5–10 departments concurrently
	NVIDIA Multi-Instance GPU support: Yes			
llama-3-3-70b-instruct A state-of-the-art refresh of the Llama 3.1 70B Instruct model by using the latest advancements in post training techniques.	3 CPU 96 GB RAM 75 GB Storage	Infrastructure: Sized to support a single LLaMA 3 70B Instruct instance using 8 NVIDIA A100 or H100 80GB GPUs. Best for experimentation, validation, and limited inference use cases. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Small LLaMA 3 70B Instruct instance (8 GPUs) (Total: 8 GPUs used for model serving) User Capacity: Supports up to 10–15 concurrent users in the Prompt Lab 1 active model deployment 1 active JupyterLab or notebook environment Use Cases: PoC of LLaMA 3-based virtual research assistant for a medical research lab Running LLaMA 3 in 4-bit quantization for lightweight Q&A over PDFs Testing safety alignment behaviour on domain-specific prompts	Infrastructure: Sized to support a single LLaMA 3 70B Instruct instance using 16 NVIDIA A100 or H100 80GB GPUs. Suitable for steady-state production inference with moderate concurrency. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Medium LLaMA 3 70B Instruct instance (16 GPUs) (Total: 16 GPUs used for model serving) User Capacity: Supports up to 40–60 concurrent users in the Prompt Lab 2–3 active model deployments 2–3 active JupyterLab or notebook environments Use Cases: Real-time RAG system answering questions on large legal corpora Training adapters for a healthcare chatbot that respects privacy constraints Interactive assistant for investment analysts trained on financial briefings	Infrastructure: Sized to support a single LLaMA 3 70B Instruct instance using 24 NVIDIA A100 or H100 80GB GPUs. Ideal for high-performance production inference or large-scale batch processing. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Large LLaMA 3 70B Instruct instance (24 GPUs) (Total: 24 GPUs used for model serving) User Capacity: Supports up to 150–200 concurrent users in the Prompt Lab 4–5 active model deployments 4–5 active JupyterLab or notebook environments Use Cases: Customer support LLM for a global SaaS platform (multi-language, high concurrency) AI co-pilot for developers with integration into GitHub and Jira Secure enterprise deployment for internal LLM-as-a-service with SSO
	NVIDIA Multi-Instance GPU support: No			
mistral-large The most advanced Large Language Model (LLM) developed by Mistral AI with state-of-the-art reasoning capabilities that can be applied to any language-based task, including the most sophisticated ones.	16 CPU 246 GB RAM 240 GB Storage	Infrastructure: Sized to support a single Mistral Large instance using 8 NVIDIA A100 or H100 80GB GPUs. Best suited for development, lightweight tuning, or low-volume inference use cases. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Small Mistral Large instance (8 GPUs) (Total: 8 GPUs used for model serving) User Capacity: Supports up to 30–40 concurrent users in the Prompt Lab 1–2 active model deployments 1–2 active JupyterLab or notebook environments Use Cases: Answer bot for internal Slack workspace using documentation embedding Automated blog draft generator for marketing teams Proof-of-concept for internal tool that summarizes Zendesk tickets	Infrastructure: Sized to support a single Mistral Large instance using 16 NVIDIA A100 or H100 80GB GPUs. This configuration supports steady production inference with moderate throughput. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Medium Mistral Large instance (16 GPUs) (Total: 16 GPUs used for model serving) User Capacity: Supports up to 80–100 concurrent users in the Prompt Lab 3–4 active model deployments 3 active JupyterLab or notebook environments Use Cases: Support assistant that answers customer queries from CRM and knowledge base LLM tool that parses and summarizes meeting transcripts for project managers Generating compliance briefs from legal templates and emails	Infrastructure: Sized to support a single Mistral Large instance using 24 NVIDIA A100 or H100 80GB GPUs. Ideal for high-throughput inference, concurrent user loads, and larger tuning jobs. Prompt Tuning: 1 GPU is reserved for prompt tuning workloads via the Tuning Studio. Model Serving: 1 Large Mistral Large instance (24 GPUs) (Total: 24 GPUs used for model serving) User Capacity: Supports up to 200–250 concurrent users in the Prompt Lab 5 active model deployments 5 active JupyterLab or notebook environments Use Cases: Multilingual assistant for a multinational call center operation AI summarizer for 10,000+ daily incoming customer feedback entries Live in-product chatbot for user onboarding in a B2B SaaS dashboard
	NVIDIA Multi-Instance GPU support: No			

7 Validation Process and Results

7.1 Solution Validation Results

7.1.1 Use Case 1: Comparing Large Language Models

Business Challenges

- Selecting the right LLM, chunking techniques, VectorDB, embeddings etc. for a specific use case without compromising in the accuracy of response. And ROI

Value Proposition:

- Compare LLMs on watsonx.ai or even external LLM and showing accuracy, cost and performance all together at one place for each GenAI tasks, use case for different domains
- Developers can compare different chunking techniques, embeddings, VectoreDB and output from different LLMs for a specific use case. Business users can compare the accuracy and ROI on the same page.

Benefits:

- One page to compare all aspects of Generative AI tasks, be it saving and fetching context, different LLMs performance, accuracy and cost of running it on production.

Validation Results:

- Validation results are captured below.

Add/ Update Expected Response	Retrieve Context	Generate
-------------------------------------	------------------	----------

ibm/granite-13b-chat-v2

Generated Response:

Secondary Market refers to a market where securities are traded after being initially offered to the public in the primary market and/or listed on the Stock Exchange. Majority of the trading is done in the secondary market. Secondary market comprises of equity markets and the debt market

Tokens: 576.5 input + output =
Time: 1.26 seconds

gpt-4

Generated Response:

Secondary Market refers to a market where securities are traded after being initially offered to the public in the primary market and/or listed on the Stock Exchange. Majority of the trading is done in the secondary market. Secondary market comprises of equity markets and the debt markets. For the general investor, the secondary market provides an efficient platform for trading of his securities. For the management of the company, Secondary equity markets serve as a monitoring and control conduit.

Tokens: 576.5 input + 87 output = 663.5
Time: 4.13 seconds

Figure 8: Comparing two LLMs

ibm/granite-13b-chat-v2

Input Token Price: \$0.6 / 1 million tokens
Output Token Price: \$0.6 / 1 million tokens

Cost of Infrencing:

Per Day: \$2.33163
Per Month: \$69.9487
Per Annum: \$851.043

gpt-4

Input Token Price: \$30 / 1 million tokens
Output Token Price: \$60 / 1 million tokens

Cost of Infrencing:

Per Day: \$22.515
Per Month: \$675.45
Per Annum: **\$8,217.975**

Evaluation Metrics Comparison (Accuracy)

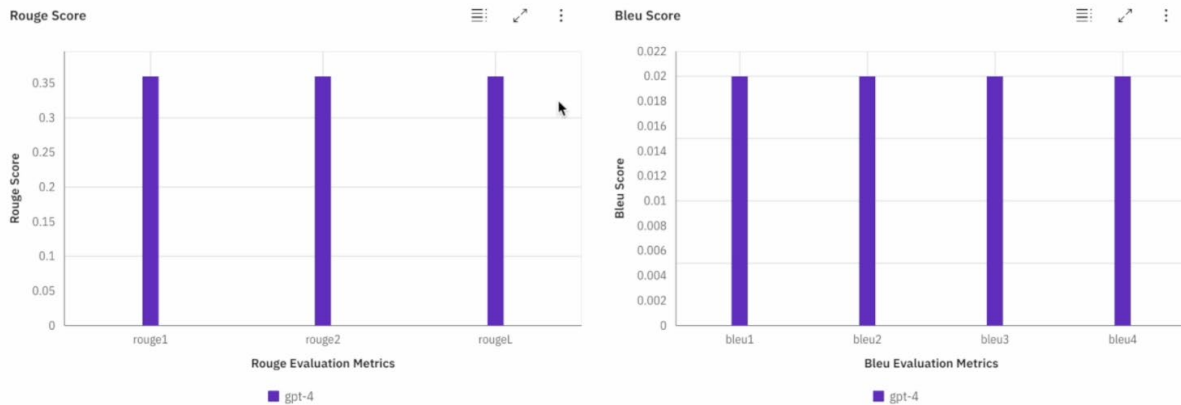


Figure 9: Comparison result

7.1.2 Use Case 2: Wealth manager

Business Challenges:

- A wealth manager must go through various data points before meeting a client for financial advisory.
- The wealth manager must spend time after the meeting with the client to summarize the key decisions that were suggested during the meeting

Value Proposition:

- Pre-meeting report generation: The wealth manager agent can generate a detailed report of client's portfolio.
- Post-meeting MoM summarization: The wealth manager agent can summarize meeting transcripts and draft an email.
- Market & Financial Insights: The wealth manager agent can also provide PoV from Goldman reports (RAG) or get financial insights (Web Search).

Benefits

- At least 2 hours of manual efforts of a wealth manager are saved by generating detailed report for client's portfolio
- At least 1 hour of manual efforts of a wealth manager are saved by generating an email draft.

Validation Results:

- Validation results are captured below.

Wealth Manager Agent

powered by IBM® watsonx™.ai



Hi 👋, I am a wealth manager agent. I am here to provide financial insights, generate stock investment reports, and assist with meeting follow-ups. Here are my capabilities:

- **Market & Financial Insights** Retrieve relevant information on market trends, stock performance, and financial topics.
- **Stock Investment Reports** Generate detailed reports on stock portfolios, including performance summaries, news analysis, and expert insights.
- **Meeting Follow-ups** Summarize recent discussions and draft professional follow-up emails.

How can I help you today?

Give me a report on John Doe's stock investment portfolio

Figure 10: Wealth Manager Agent prompt

Portfolio Summary

The portfolio consists of the following stocks:

ID	Security Name	Market Value (USD)	Y2Y %	Industry Sector
1	S&P 500	8500000	15	Index Fund
2	MSCI AC World	6000000	12	Index Fund
3	TOPIX	4200000	13	Index Fund
4	MSCI AP ex-Japan	5700000	9	Index Fund
5	STOXX 600	3500000	5	Index Fund
6	MSCI EM	2800000	3	Index Fund

Articles Summary

S&P 500 Performance and News

- As of February 14, 2025, the S&P 500 index was at 6,143.50, up by 0.13%.
- The top nine companies by weighting in the S&P 500 as of September 30, 2024, were Apple, Microsoft, Nvidia, Amazon.com, Meta Platforms, Alphabet, Berkshire Hathaway, Broadcom, and Tesla.
- On February 14, 2025, the S&P 500 index closed at 6,025.99, marking a 0.95% increase.
- Recent market discussions have focused on inflation fears, tariff concerns, and the unveiling of new AI models.

MSCI AC World Performance and News

- The iShares MSCI ACWI ETF (ACWI) has shown steady growth, aligning with broader trends in global equity markets.

Ask the Wealth Manager Agent...

Figure 11: Wealth Manager Agent result

8 Appendix: Lenovo Bill of materials

8.1 Single Rack Deployment

Below is an example BoM of Single Rack deployment. For cost optimization, alternative networking options include NVIDIA SN3700 switches and NVIDIA ConnectX-7 NDR200/200GbE NICs instead of NVIDIA BlueField-3 NICs. Additionally, copper cables can replace optical cables to reduce costs. This example BoM includes:

- 1 × Lenovo ThinkSystem SR675 V3 with 8 × NVIDIA H200 NVL GPUs per server
- 3 × Lenovo ThinkSystem SR635 V3
- 2 × NVIDIA SN3700 Switches
- 2 × NVIDIA SN2201 Switches
- Storage is optional and not included in this BoM.

8.1.1 ThinkSystem SR675 V3

Part Number	Product Description	Qty per System	Total Qty
7D9RCTOLWW	ThinkSystem SR675 V3 3yr Warranty - HPC&AI with Controlled GPU		1
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	1	1
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	2	2
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 Adapter	5	5
BTMB	ThinkSystem 1x4 E3.S Backplane	1	1
B5XJ	ThinkSystem M.2 SATA/NVMe 2-Bay Adapter	1	1
BVEK	ThinkSystem E3.S PM1743 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2	2
BE0D	N+1 Redundancy With Over-Subscription	1	1
C3V0	ThinkSystem NVIDIA 4-way bridge for H200 NVL	2	2
BKTJ	ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supply	4	4
C3V3	ThinkSystem NVIDIA H200 NVL 141GB PCIe GPU Gen5 Passive GPU	8	8
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	1	1
C2AL	ThinkSystem AMD EPYC 9535 64C 300W 2.4GHz Processor	2	2
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2	2
C0CK	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM-A	24	24

8.1.2 ThinkSystem SR635 V3

Part Number	Product Description	Qty per System	Total Qty
7D9GCTOLWW	ThinkSystem SR635 V3 - 3yr Warranty - HPC&AI		3
BLK7	ThinkSystem SR635 V3/SR645 V3 x16 PCIe Gen5 Riser 1	1	3
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1	3
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 Adapter	1	3
C0ZT	ThinkSystem 2.5" U.2 VA 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2	6
BE0C	N+1 Redundancy Without Over-Subscription	1	3
BNFH	ThinkSystem 1100W 230V/115V Platinum Hot-Swap Gen2 Power Supply v3	2	6
BQ26	ThinkSystem SR645 V3/SR635 V3 1U High Performance Heatsink	1	3
BLK4	ThinkSystem V3 1U 10x2.5" Chassis	1	3
C2AQ	ThinkSystem AMD EPYC 9335 32C 210W 3.0GHz Processor	1	3
C0CJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM-A	12	36
BH9M	ThinkSystem V3 1U Performance Fan Option Kit v2	6	18

8.1.3 NVIDIA SN3700 Switch

Part Number	Product Description	Qty per System	Total Qty
7D5FCTOCWW	Mellanox SN3700V 200GbE Managed Switch with Cumulus (oPSE)		1
BJ5U	Mellanox SN3700V 200GbE Managed Switch with Cumulus (oPSE)	1	1

8.1.4 NVIDIA SN2201 Switch

Part Number	Product Description	Qty per System	Total Qty
7D5FCTOGWW	Nvidia SN2201 1GbE Managed Switch with Cumulus (PSE)		2
BPC7	Nvidia SN2201 1GbE Managed Switch with Cumulus (PSE)	1	2

8.2 Scalable Unit (SU) Deployment

Below is an example BoM of one SU deployment with NVIDIA Spectrum-X. For cost optimization, alternative networking options include NVIDIA SN3700 switches and NVIDIA ConnectX-7 NDR200/200GbE NICs instead of NVIDIA BlueField-3 NICs. Additionally, copper cables can replace optical cables to reduce costs. This example BoM includes:

- 4 × Lenovo ThinkSystem SR675 V3 with 8 × NVIDIA H200 NVL GPUs per server
 - 5 × Lenovo ThinkSystem SR655 V3
 - 2 × NVIDIA SN5600 Switches
 - 2 × NVIDIA SN2201 Switches
 - Storage is optional and not included in this BoM.
 - Networking connectivity is based on the standard Lenovo Hybrid AI 285 Reference Architecture.
- Additional NICs may be required depending on customer networking requirements.

8.2.1 ThinkSystem SR675 V3

Part Number	Product Description	Qty per System	Total Qty
7D9RCTOLWW	ThinkSystem SR675 V3 3yr Warranty - HPC&AI with Controlled GPU		4
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	1	4
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	2	8
C0Q4	ThinkSystem NVIDIA BlueField-3 B3140H VPI QSFP112 1P 400G PCIe Gen5 x16 Adapter	4	16
BVBG	ThinkSystem NVIDIA BlueField-3 B3220 VPI QSFP112 2P 200G PCIe Gen5 x16 Adapter	1	4
BTMB	ThinkSystem 1x4 E3.S Backplane	1	4
B5XJ	ThinkSystem M.2 SATA/NVMe 2-Bay Adapter	1	4
BVEK	ThinkSystem E3.S PM1743 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	4	16
C3V0	ThinkSystem NVIDIA 4-way bridge for H200 NVL	2	8
BKTJ	ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supply	4	16
C3V3	ThinkSystem NVIDIA H200 NVL 141GB PCIe GPU Gen5 Passive GPU	8	32
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	1	4
C2AL	ThinkSystem AMD EPYC 9535 64C 300W 2.4GHz Processor	2	8
BKSS	ThinkSystem M.2 7450 PRO 1.92TB Read Intensive Entry NVMe PCIe 4.0 x4 NHS SSD	2	8
C0CJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM-A	24	96

8.2.2 ThinkSystem SR655 V3

Part Number	Product Description	Qty per System	Total Qty
7D9E-CTOLWW	ThinkSystem SR655 V3		5
BLKK	ThinkSystem V3 2U 24x2.5" Chassis	1	5
C2AC	ThinkSystem SR655 V3 MB w/IO+PIB+FB,2U	1	5
C2AQ	ThinkSystem AMD EPYC 9335 32C 210W 3.0GHz Processor	1	5
C0CJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM-A	12	60
BPQV	ThinkSystem V3 2U x16/x16/E PCIe Gen5 Riser1 or 2	1	5
BVBG	ThinkSystem NVIDIA BlueField-3 B3220 VPI QSFP112 2P 200G PCIe Gen5 x16 Adapter	1	5
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Adapter	1	5
BXMH	ThinkSystem M.2 PM9A3 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2	10
BS7Y	ThinkSystem V3 2U 8x2.5" NVMe Gen5 Backplane	1	5
C0ZU	ThinkSystem 2.5" U.2 Multi Vendor 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2	10
BLL6	ThinkSystem 2U V3 Performance Fan Module	6	30
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2	10
BLKH	ThinkSystem 1100W 230V Titanium Hot-Swap Gen2 Power Supply	2	10
B8LA	ThinkSystem Toolless Slide Rail Kit v2	1	5
C1PT	ThinkSystem SR635 V3/SR655 V3 Root of Trust Module Low Voltage-RoW V2	1	5
BQQ6	ThinkSystem 2U V3 EIA right with FIO	1	5
5PS7B08760	3Yr Premier NBD Resp + KYD SR655 V3	1	5

8.2.3 NVIDIA SN5600 Switch

Part Number	Product Description	Qty per System	Total Qty
7D5FCTONWW	NVIDIA SN5600 800GbE Managed Switch with Cumulus (PSE)		2
C0Q5	NVIDIA SN5600 800GbE Managed Switch with Cumulus (PSE)	1	2

8.2.4 NVIDIA SN2201 Switch

Part Number	Product Description	Qty per System	Total Qty
-------------	---------------------	----------------	-----------

7D5FCTOGWW	Nvidia SN2201 1GbE Managed Switch with Cumulus (PSE)		2
BPC7	Nvidia SN2201 1GbE Managed Switch with Cumulus (PSE)	1	2

8.2.5 XClarity Software

Part Number	Product Description	Qty per System	Total Qty
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)		3
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr software S&S		5

Resources

Resources	Links
IBM watsonx	IBM watsonx Website
Red Hat OpenShift	Red Hat OpenShift Website
Red Hat OpenShift AI	Red Hat OpenShift AI Website
SR675 V3	Lenovo ThinkSystem SR675 V3 Server Product Guide > Lenovo Press
SR635 V3	Lenovo ThinkSystem SR635 V3 Server Product Guide > Lenovo Press
SR655 V3	Lenovo ThinkSystem SR655 V3 Server Product Guide > Lenovo Press
XClarity	XClarity Systems Management
L40S GPU	NVIDIA L40S GPU Product Guide > Lenovo Press
H200 NVL GPU	NVIDIA H200 NVL GPU > NVIDIA Data Center
SN3700V Switch	NVIDIA SN3700V Switch > NVIDIA Datasheet
SN5600 Switch	NVIDIA SN5600 Switch > NVIDIA Datasheet
SN2201 Switch	NVIDIA SN2201 Switch > NVIDIA Datasheet
ConnectX-7 SmartNIC	NVIDIA ConnectX-7 SmartNIC > NVIDIA Datasheet
BlueField-3 B3220 DPU	NVIDIA BlueField-3 B3220 DPU > NVIDIA Datasheet
BlueField-3 B3140H SuperNIC	NVIDIA BlueField-3 B3140H SuperNIC > NVIDIA Datasheet

Document history

Version 1.0 April 2025

First version includes a scalable AI infrastructure for IBM watsonx, leveraging Lenovo SR675 V3, SR655 V3, and SR635 V3, NVIDIA GPUs, Red Hat OpenShift, and IBM watsonx 5.1.2 software.

Trademarks and special notices

© Copyright Lenovo 2025.

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at <https://www.lenovo.com/us/en/legal/copytrade/>.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®

Neptune®

ThinkSystem®

XClarity®

The following terms are trademarks of other companies:

AMD and AMD EPYC™ are trademarks of Advanced Micro Devices, Inc.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

IBM®, Code Assistant™, Granite®, IBM watsonx®, Orchestrate®, Power®, ibm.com®, watsonx Assistant®, watsonx Code Assistant™, watsonx Orchestrate™, watsonx.ai®, watsonx.data®, watsonx.governance®, and watsonx® are trademarks of IBM in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo's current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.