Unlocking Enterprise Data with a Production-Ready AI Platform

IBM Fusion provides an open platform for streamlined AI infrastructure, helping organizations deploy and scale high-value AI workloads in days, not months.

Highlights

Delivers AI-optimized infrastructure for cloud-native and VM-based workloads

Provides flexibility in choosing where to tune, deploy, and run models and gen AI

Enables AI systems to deliver accurate contextual answers to plain-language queries without data migration or duplication

Unifies best-of-breed AI technologies into a turnkey platform that delivers real outcomes quickly and safely

Enterprises are at a critical decision point: AI has become a strategic imperative for driving innovation and maintaining competitive advantage, but organizations can't afford lengthy infrastructure projects before they start building and deploying AI solutions. Success depends on unlocking the value of enterprise data, much of which remains hidden in disparate data stores, inaccessible to AI tools.

In parallel, organizations face growing requirements around data residency and digital sovereignty. Business leaders around the globe now recognize the need for AI systems that can use trusted enterprise data while keeping control over where sensitive data lives and how it is accessed.

IBM Fusion delivers the answer – an integrated platform that enables organizations to rapidly deploy open, AI-optimized infrastructure. Fusion helps them accelerate data pipelines, streamline IT operations, manage inference, and enable hybrid cloud flexibility from the data center to the cloud and the edge. Fusion focuses on making data accessible, meaningful, and AI-ready while enabling modernization at their own pace on a single OpenShift platform foundation.

Fusion delivers:

- *Speed to insight* that brings together a production-ready platform foundation, integrated data services, and the tooling teams need so they can start building, testing, and deploying AI workloads in days rather than months.
- *An open platform for AI* that supports openness at every layer – choice of foundation models, choice of where workloads run, and choice of accelerators.
- *Streamlined AI infrastructure* that simplifies Day-1 deployment and Day-2 operations with built-in resilience, lifecycle management, and governance controls that help keep AI environments reliable and secure as usage grows.

Data as the foundation of trusted AI

AI outcomes depend on data quality, freshness, and governance. Fusion is designed to make enterprise data usable for AI workloads where accuracy and trust matter, including retrieval-augmented generation (RAG) and inferencing for knowledge-based assistants and agentic AI.

Fusion includes differentiated content-aware storage capabilities that use natural language processing to extract semantic meaning from unstructured data and keep RAG pipelines up to date and more efficient by moving document processing and vectorization closer to the data.

Fusion supports both unstructured and structured data pipelines and integrates natively with watsonx.data to simplify deployment of structured data services on the same OpenShift foundation. This helps organizations ensure that both pipelines contribute to trusted AI and RAG workloads.

Multiple AI Stacks and Use Cases

Fusion provides the essential platform foundation for the AI workloads and other applications your enterprise needs to build and deploy, delivering a consistent operational model so teams can scale adoption over time without reinventing the platform for every new use case.

Organizations typically start with practical, high-value workloads such as document intelligence and RAG-based assistants for customer communications, risk detection, and IT operations automation, then expand as they prove their value.

IBM Fusion for Red Hat AI – integrated AI workbench

As organizations move from AI pilot projects to full-scale implementation, they typically confront the challenge of how to build AI applications efficiently and deploy them consistently at scale.

Red Hat AI serves as the default AI development and runtime environment for Fusion. This ensures that teams have a consistent, enterprise-grade workbench regardless of whether they later deploy NVIDIA blueprints, custom models, or watsonx agents.

Every Fusion deployment can natively incorporate the Red Hat AI workbench, providing teams with consistent tooling for model development, packaging, and serving. Red Hat AI's "any model, any cloud, any hardware" approach aligns with Fusion's open architecture, allowing enterprises to build AI applications that are optimized to run on NVIDIA GPUs today while remaining free to adopt accelerators from AMD, Intel, IBM and emerging inference device providers.

Fusion amplifies the value of this approach by providing the production-ready OpenShift platform for deploying and operating those Red Hat AI applications with resilience, automation, and enterprise governance.

Fusion and Red Hat AI together provide:
- A unified application workbench for developing, packaging, and serving AI applications;
- Consistent, enterprise-ready deployment of Red Hat AI applications on Fusion's automated, resilient OpenShift foundation;
- Broad accelerator flexibility, enabling teams to choose any GPU.

This integration helps organizations move from early production wins to repeatable, team-wide AI development, while keeping their options open across models, tools, and hardware. Fusion ensures that whatever developers create in Red Hat's workbench can be deployed, scaled, and governed as part of a unified platform.

Figure 1 – IBM Fusion delivers an integrated platform that enables organizations to standardize on automated, AI-optimized infrastructure.

Fusion also simplifies standing up watsonx.data services on OpenShift by providing the platform, storage, and operational layer required to run data lakehouses, so organizations can bring structured lakehouse data directly into their AI pipelines without complex infrastructure setup.

Red Hat streamlines implementation with Red Hat Validated Patterns, which are automated, deployable code architectures designed to provide complete, end to end solutions for AI/ML and multi cloud GitOps use cases. Red Hat Validated Patterns provide tested, continuously validated, ready to-run blueprints that reduce deployment risk and accelerate implementation.

Red Hat also provides AI QuickStarts, which are ready-to-run, AI-centric demonstration projects designed to show how to build real business use cases on Red Hat OpenShift AI and the broader Red Hat AI ecosystem.

IBM Fusion as a Platform for AI Blueprints

The NVIDIA AI ecosystem has expanded beyond GPU acceleration to encompass data centric platforms that help enterprises build and operate AI factories for transforming raw enterprise data into intelligence. To accelerate adoption, NVIDIA provides a growing portfolio of AI blueprints and AI Data Platform reference designs that give organizations prescriptive architectures for data ingest, curation, vectorization, multimodal pipelines, and high-performance inference.

NVIDIA AI Blueprints are GPU-optimized workloads that run on top of the Fusion + Red Hat AI foundation and are designed to reduce time-to-value for common enterprise workloads. Organizations benefit from NVIDIA's accelerator-optimized pipelines combined with Fusion's openness across hardware and models.

NVIDIA contributes the AI blueprints: validated, accelerator-optimized recipes that include model architectures, reference pipelines, and performance-tuned guidance for building high value, production ready use cases such as RAG pipelines, document intelligence, multimodal search, and GPU-accelerated inference services.

The blueprints define what to deploy, but they still need a stable, governed environment that can run them at scale. Fusion provides the operational guardrails that NVIDIA blueprints rely on and that organizations would otherwise have to engineer themselves, including stable GPU scheduling, storage throughput tuning, and governed multi tenant operations.

Fusion complements the NVIDIA AI Data Platform, which provides data ingest, processing, and embedding services used by AI factories. Fusion's content-aware storage and unified data foundation ensure that these data pipelines run efficiently, with predictable throughput and consistent governance.

Fusion provides:
- A production ready "AI factory floor" where NVIDIA blueprint can be deployed to achieve rapid time to first inference on a stable OpenShift foundation;
- Integrated support for NVIDIA AI Data Platform, ensuring ingest, curation, vectorization, and multimodal processing pipelines benefit from throughput optimized storage and consistent operational guardrails;
- Enterprise consistent operations, with security, lifecycle automation, and reliability layered around NVIDIA toolkits and accelerators;
- Scalable GPU performance as demand grows, without platform redesign.

This gives customers immediate, production-grade AI wins built on NVIDIA's best practices while establishing Fusion as the long-term platform that can support additional AI stacks as the organization's requirements evolve.

By integrating NVIDIA's best practices, data microservices, and accelerator-tuned blueprints with Fusion's operational platform and Red Hat's unified AI workbench, enterprises get a turnkey path to build and run AI factories, accelerating high-value workloads today while preserving flexibility for future stacks, models, and hardware choices.

IBM Fusion with watsonx for agent orchestration

Once organizations have production AI running and teams are building repeatable solutions, they can then layer agentic orchestration on top to automate decisions, workflows, and multi step business processes.

Enterprises are addressing this opportunity by adding watsonx capabilities, particularly watsonx Orchestrate, to create agent-driven workflows that can execute multi-step tasks with appropriate controls and visibility, while Fusion provides the stable platform underneath.

Fusion with watsonx Orchestrate provides:
- Agentic workflows that coordinate multi-step tasks across systems and teams, moving beyond Q&A to execution;
- Orchestration that helps turn AI into repeatable business processes with appropriate controls and visibility;
- A strong fit for work that spans multiple applications (for example, handling requests, updating records, and triggering next steps);
- Support for governance and auditability needs as AI takes on more responsibility in core processes.

Why IBM Fusion

Fusion unifies best-of-breed AI technologies into a turnkey platform that delivers real outcomes quickly and safely. Red Hat AI provides the OpenShift-based workbench for building and serving applications, NVIDIA supplies the accelerated compute and validated blueprints for high performance inference, and watsonx adds the governed, agentic execution layer that drives cross-enterprise workflows.

By running NVIDIA validated blueprints on Fusion, organizations get production-grade GPU operations without custom integration, accelerating deployment of high-value AI workloads. Fusion's pre-engineered foundation eliminates the time and risk of stitching together GPUs, storage, data services, and OpenShift, enabling AI inferencing in days, not months.

Fusion also acts as the data unification layer, preparing structured lakehouse data and unstructured content for downstream workloads. This ensures both Red Hat AI applications and watsonx Orchestrate agents can reliably access the governed, high-quality enterprise data they need to operate at scale.

For more information

To learn more about IBM Fusion, contact your IBM representative or IBM Business Partner, or visit https://www.ibm.com/products/storage-fusion.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
March 2026