Top considerations for building a production-ready **AI environment**

Red Hat

Contents

Data is a critical business asset
Page 3

Build a production-ready AI platform
Page 6

Containers and container orchestration
Page 9

Application management and genAIOps
Page 11

Hybrid cloud platform
Page 13

Model customization and alignment
Page 14

AI inference at scale
Page 15

AI safety
Page 17

Build an open, flexible foundation for AI
Page 18

Gain choice and flexibility with a certified AI/ML partner ecosystem
Page 20

Partner highlights
Page 21

Success in action
Page 22

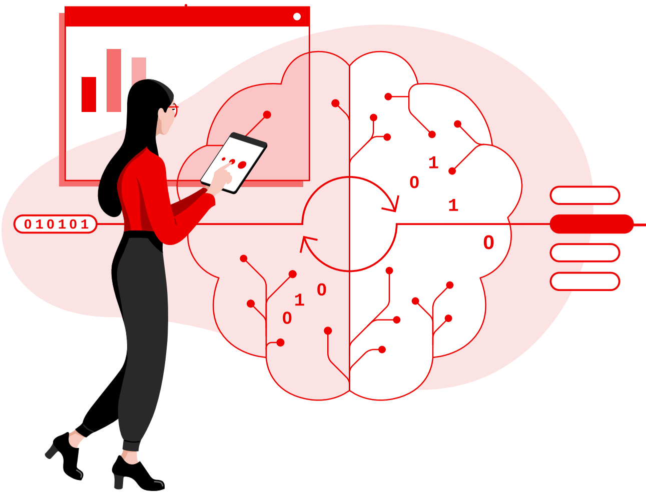
Ready to get more from your data?
Page 23

Data is a critical business asset

The status of the enterprise AI market

Generative artificial intelligence (gen AI) has moved from experiment to everyday tool for many organizations.

Teams use it to summarize content, assist with code and content creation, and interact with data in more natural ways. At an enterprise scale, leaders expect gen AI to help them improve outcomes for customers, employees, and across operations, not just answer ad hoc questions or create funny memes.

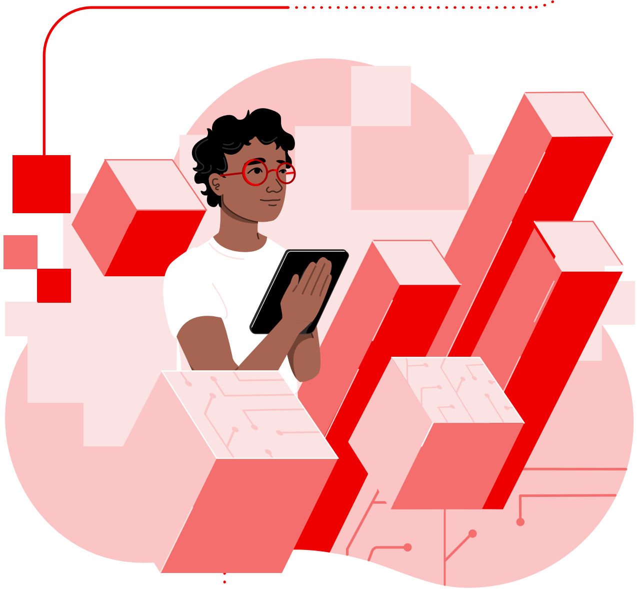


Building on their existing data and applications, gen AI can help organizations:

- Turn large volumes of unstructured content into searchable, reusable knowledge.
- Assist developers, analysts, and writers to create and refine code, reports, and content faster.
- Personalize digital experiences for customers and employees across channels.
- Automate routine decisions and workflows that follow clear policies.
- Improve productivity for development, operations, and business teams.

Recent industry research shows that this shift is already underway. IDC reports that more than half of surveyed organizations already run several gen AI-enhanced applications or services in production, and expects year-over-year AI spending between 2025 and 2029 to grow by roughly one-third, reaching around US$1.3 trillion by 2029.[1] For most enterprises, gen AI is becoming part of core products and services.

[1] IDC White Paper. "Agentic AI to Dominate IT Budget Expansion Over Next Five Years, Exceeding 26% of Worldwide IT Spending, and $1.3 Trillion in 2029, According to IDC." 26 Aug. 2025.



At the same time, organizations are looking ahead to the next step: agentic AI. Instead of treating gen AI as a single chatbot or assistant, agentic AI uses AI agents that can call tools, interact with applications, and coordinate multistep tasks. In practice, this approach can change how you build and operate software, from customer self-service and IT operations to complex business workflows.

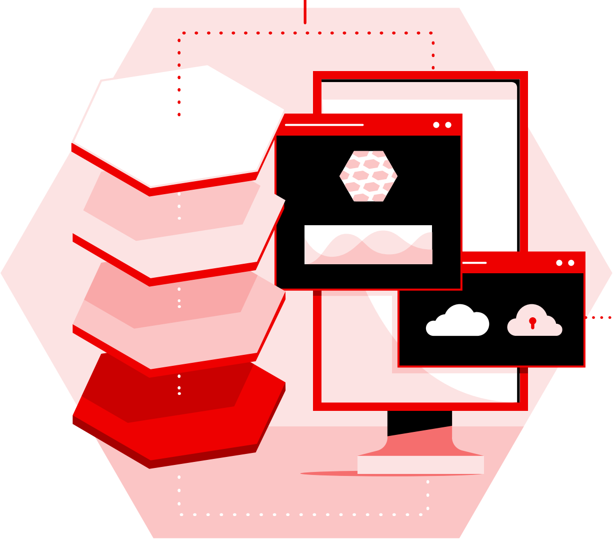
IDC reports that more than half of organizations already run proofs of concept or early use cases for agentic AI and that nearly one-third of AI-enabled applications will rely on it by the end of 2026.[2] Enterprises are now treating it as a strategic path forward.

To capture this value, you need flexibility in how and where you run AI.

Many organizations now plan for hybrid AI infrastructure, combining public clouds with dedicated on-premise environments. IDC notes that a hybrid mix of public cloud and on-premise infrastructure has become the most common digital infrastructure strategy, and that most decision makers believe their AI workloads require hybrid deployment.[3]

A hybrid, open platform lets organizations:

- Keep sensitive data and models under their control.
- Meet data privacy and sovereignty requirements.
- Choose from a range of hardware options.
- Choose from a wide range of open source models.
- Still take advantage of cloud scale when they need it.



This e-book will go over core steps to building a production-ready AI platform, key considerations an organization will face in the process, and how Red Hat® AI Enterprise offers a unified solution to help build it.

[2] IDC White Paper. "Agentic AI Impact on Digital Infrastructure Strategies." Document # US53418526, Oct. 2025. (purchase required)

[3] IDC White Paper. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Document #US53418426, Oct. 2025. (purchase required)

AI use cases across industries

Healthcare

- Increase clinical efficiency.
- Boost diagnosis speed and accuracy.
- Improve patient outcomes.

Telecommunications

- Gain insight into customer behavior.
- Enhance customer experiences.
- Optimize 5G network performance.

Insurance

- Automate claims processing.
- Deliver usage-based insurance services.
- Assist in risk calculation.

Financial services

- Personalize customer services.
- Improve risk analysis.
- Detect fraud and money laundering.

Automotive

- Support autonomous driving.
- Predict maintenance needs.
- Improve supply chains.

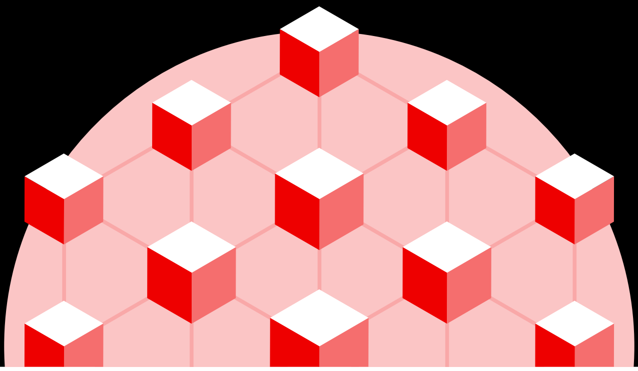
Energy

- Forecast maintenance requirements.
- Optimize field operations and safety.
- Faster reservoir simulation and prediction.

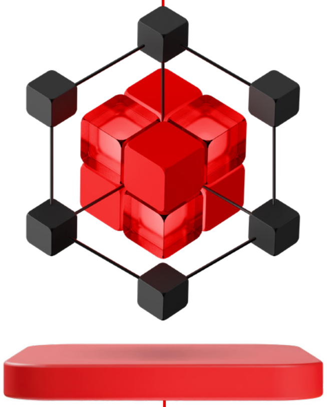
Building blocks of enterprise AI

This e-book looks at how different types of AI work together in an enterprise architecture.

- **Gen AI** uses large language models (LLMs) to generate text, code, and other content from data and prompts, so teams can work faster and experiment more easily.
- **Predictive AI** uses historical and real-time data to estimate future outcomes, such as demand, risk, or equipment health, so organizations can act earlier and with more confidence.
- **Agentic AI** uses AI agents that can call tools, connect to applications, and coordinate multistep workflows toward a goal, not just answer a single question.
- **AI inference** is the production runtime phase when models apply what they have learned to new, real-world data to return predictions, recommendations, or actions. Inference can run across a hybrid environment: on premise, in the cloud, or at the edge.



Build a production-ready AI platform



Building gen AI-powered applications and AI agents is an iterative process that extends beyond simply creating AI models. The main steps in the AI lifecycle are:

1. Define your use case, set business goals for your AI initiative, and get buy-in from stakeholders and leadership.
2. Choose where you want your model experimentation and deployment platforms to run: on premise or in the cloud.
3. Choose the AI model that best fits your needs. Avoid lock-in by choosing open models.
4. Customize or align your chosen models with your proprietary data using retrieval-augmented generation (RAG).
5. Deploy your model in an inference server.
6. Build gen AI-powered applications or workloads.
7. Once you have a working environment in place, extend and automate the workflow through agentic AI.
8. Monitor and manage models in a security-focused manner and at scale.

An open, adaptable AI architecture will help you execute this process more effectively. This architecture requires several key technologies and capabilities:

- **Access to frontier open-weight models** provide organizations with a starting point.
- **GenAIOps and DevOps tools** allow AI engineers, data scientists, machine learning (ML) engineers, and application developers to create, deploy, and manage AI models, AI agents and AI-powered applications.
- **Access to model tuning tools such as fine tuning and RAG capabilities** for customizing models with private enterprise data and aligning to domain specific use cases.
- **Inference runtimes** that allow you to deliver the best performance, throughput, and latency.
- **Foundational components for AI agents** to manage, govern and secure their implementation in production.
- **Compute, storage, and network accelerators** to speed data preparation, model customization, and inferencing tasks.
- **Infrastructure endpoints** provide resources across on-site, virtual, edge, and private, public, and hybrid cloud environments for all stages of AI operations.

This e-book reviews key considerations for building an effective AI architecture.

> Inferencing is the production runtime for AI. A model doesn't do something useful for you until it's got an API and it's serving content. That content is served through inferencing.
>
> **Chris Wright**
> Red Hat CTO[4]

[4] Miller, Ron. "Red Hat's CTO sees AI as next step for company's open approach." Fastforward, 11 Nov. 2025.

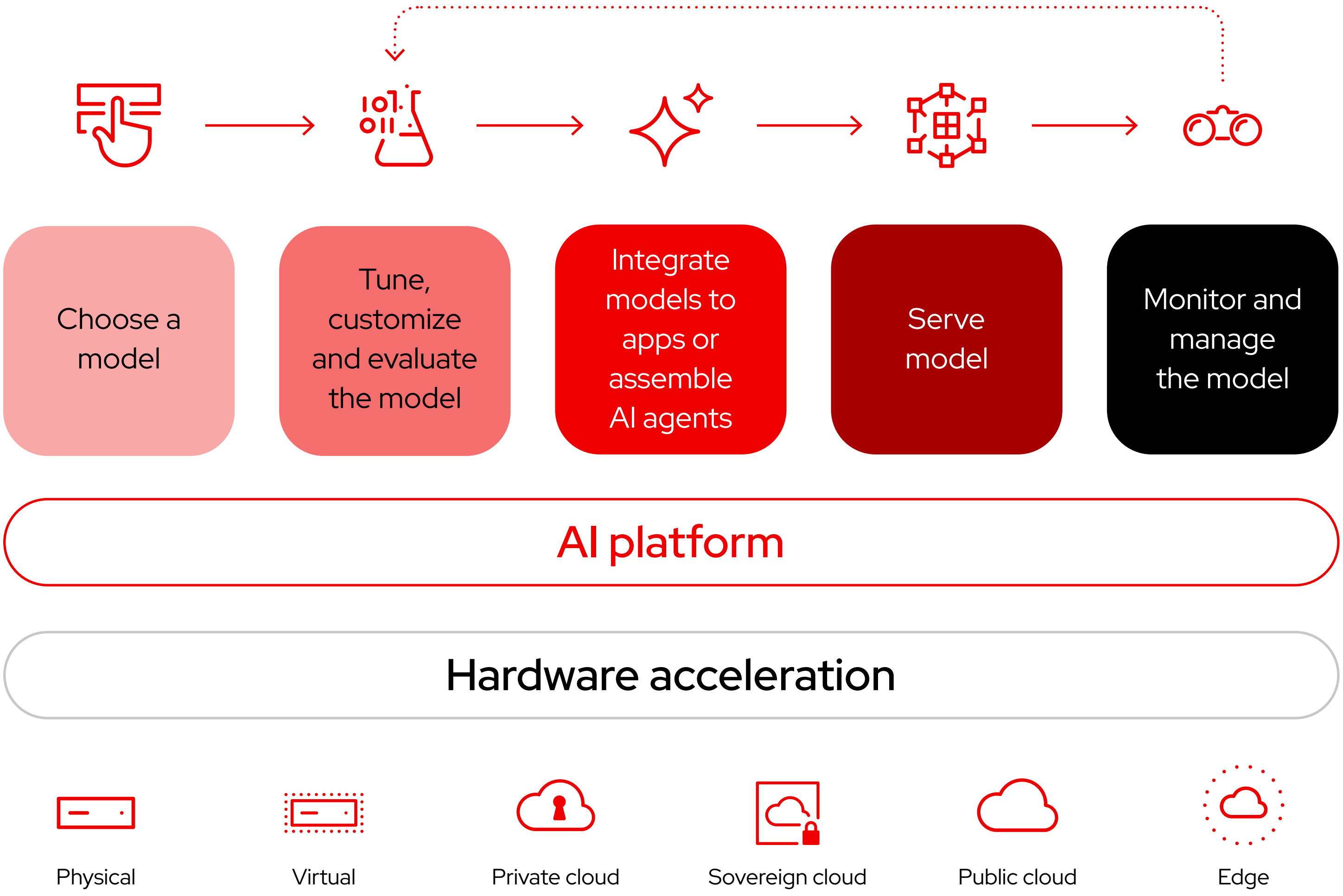


Figure 1. *The components of an AI architecture.*

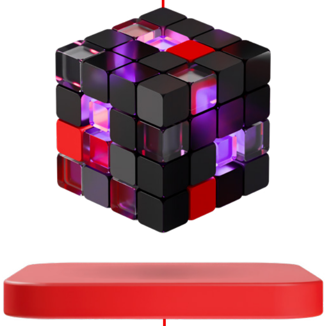
AI deployment challenges

Enterprise organizations are under pressure to choose, build, and deliver AI solutions that provide a competitive advantage. Several challenges stand in the way of operationalizing and scaling AI deployments:

- **Model cost.** Running large models and inference at scale can be expensive. Organizations must optimize models and inference to contain compute costs while still delivering accurate, responsive applications.
- **Alignment complexity.** Model training and tuning, and creating RAG pipelines are complex and graphics processing unit (GPU)-intensive. Organizations can simplify the customization of enterprise data and involve subject matter experts and developers to move from experiments to production faster.
- **Control and consistency.** Prepackaged AI services limit control over hardware, data, and governance. Choose a hybrid approach so you can select models and infrastructure while keeping ownership of data, lifecycle, and scale of your deployments.

Addressing these challenges calls for an open, hybrid AI platform that provides consistent tools for model optimization, customization, and governance across environments.

Containers and container orchestration



Containers

A container is a basic unit of software that packages applications with all of their dependencies. Containers simplify application build processes and allow applications to be deployed across different environments without change.

Why are they important for AI?

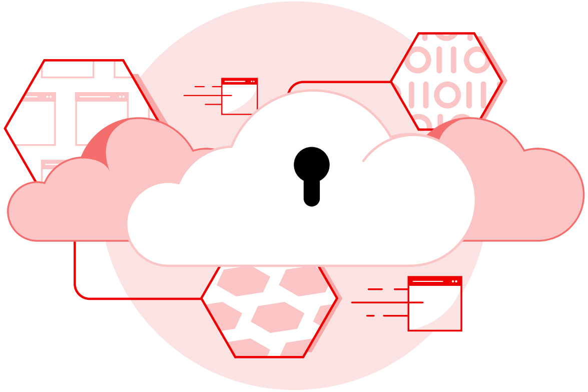
AI engineers and application developers need access to their preferred tools and resources to be most productive. At the same time, IT operations teams need to ensure that resources are up to date, in compliance, and used in a secure manner.

Containers are often the best option for deploying LLMs and gen AI-powered applications because they package model servers, dependencies, and configuration into a repeatable unit that makes production rollout, scaling, and updates more manageable.

Containers let you deploy a broad selection of AI tools across hybrid environments in a consistent way. Teams can iteratively modify and share container images with versioning capabilities that track changes for transparency. Meanwhile, process isolation and resource control improve protection from threats.

Best practices and recommendations

Look for a flexible, highly available container platform that includes integrated security features and streamlines how you deploy, manage, and move containers across your environment. Choose an open source platform that integrates with a broad set of technologies to gain more flexibility and choice.



Container orchestration

Container orchestration involves managing the creation, deployment, and lifecycle of containers across your environment.

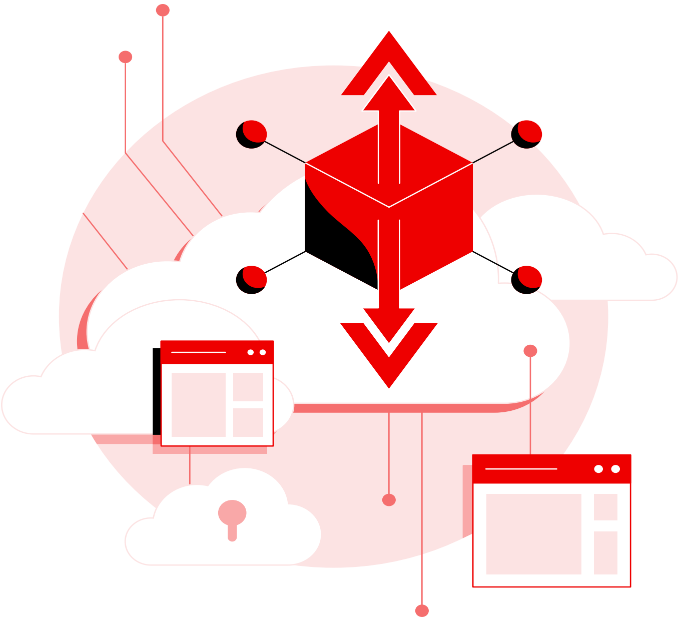
Why is it important for AI?

Once you adopt containers, you need a way to deploy, manage, and scale them efficiently. A container orchestration engine lets you administer the lifecycle of your containers in a consistent way. These tools typically centralize access to compute, storage, and networking resources across on-site, edge, and cloud environments. They also provide unified workload scheduling, multitenancy controls, and quota enforcement.

Best practices and recommendations

Select a Kubernetes-based container orchestration environment to build on a leading open source technology and avoid proprietary cloud lock-in. Look for a platform that offers strong multitenancy controls, role-based access, and policy management so you can govern AI workloads consistently. Prioritize options with a broad ecosystem of operators and integrations so you can standardize how you deploy, scale, and manage AI services across hybrid environments.

By 2027, more than 75% of all AI deployments are projected to use container technology as the underlying compute environment, up from less than 50% in 2024.[5]



[5] Gartner. "Magic Quadrant for Container Management," 10 Sept. 2024.

Application management and genAIOps

AI workload lifecycle management

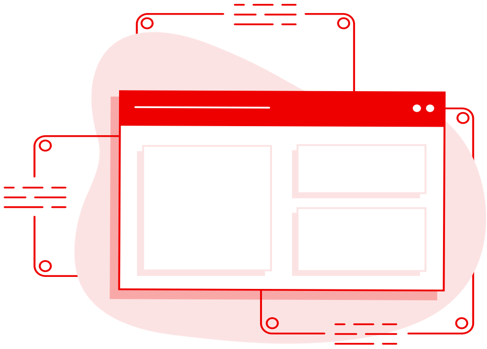
AI workload lifecycle management focuses on how you deploy, scale, and administer the tools and services that power your AI use cases.

Why is it important for AI?

AI environments are inherently complex. AI workload lifecycle management components—such as notebooks, workbenches, pipelines, and model serving endpoints—should be containerized to allow for simple control and management. IT operations teams can automate common lifecycle tasks such as configuration, provisioning, and updates to improve accuracy and reduce manual effort. Data scientists, AI engineers, and application developers can request preapproved AI environments from a catalog without opening tickets with IT. Automation also shifts staff time away from repetitive tasks toward higher-value strategic activities.

Best practices and recommendations

Effective AI workload lifecycle management starts with curated AI workbench and notebook images that include commonly used AI and ML libraries so teams start from a secure, supported baseline instead of ad hoc environments. Organizations should provide browser-based notebook environments with Git integration so teams can collaborate on experiments and track code and model changes over time.



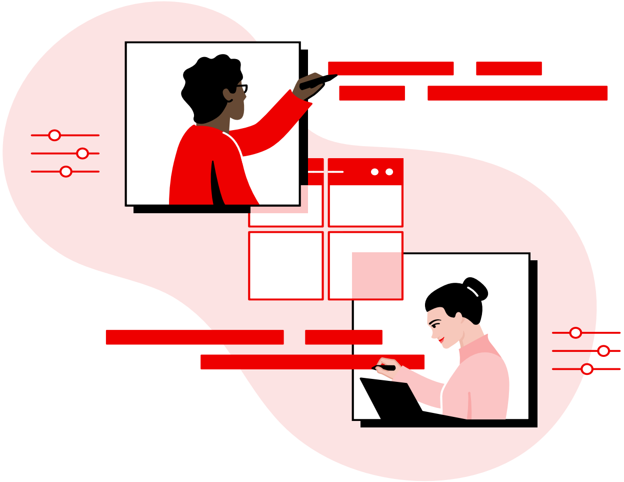
GenAIOps and MLOps practices

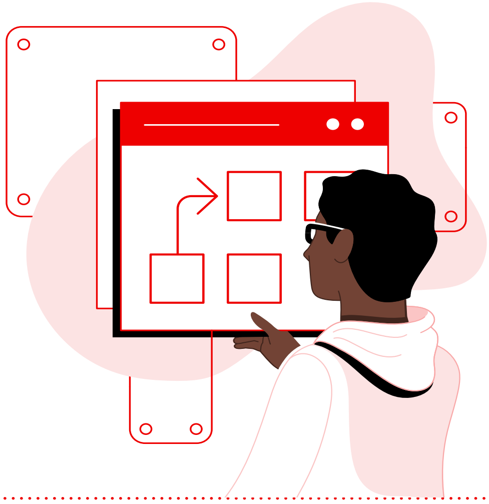
GenAIOps and MLOps practices bring together the tools, platforms, and processes needed to operationalize AI at scale.

Why is it important for AI?

Organizations need to develop and deploy AI models—and the applications that use them—rapidly and efficiently. Collaboration across teams is critical for success in these efforts.

Similar to DevOps, genAIOps and MLOps approaches foster collaboration between AI engineers, application developers, and IT operations to accelerate the creation, training, deployment, and management of gen AI models, AI agents, and AI-powered applications. Automation, often in the form of continuous integration/continuous delivery (CI/CD) pipelines, makes rapid, incremental, and iterative change possible for faster model and application development lifecycles.





GenAIOps and MLOps practices

GenAIOps and MLOps are not just about technology, people and processes also play key roles. Apply genAIOps and MLOps practices to your entire AI lifecycle. Use automation in your platforms and tools, along with open source technologies such as Kubeflow, to create CI/CD pipelines and workflows.

Hybrid cloud platform

A hybrid cloud platform provides a foundation for developing, deploying, and managing AI across on-site, edge, and cloud environments. It also gives you a way to design for sovereign AI and private AI from the start, so you can decide which workloads run in public clouds and which stay in on-premise or private cloud environments you control.

Why is it important for AI?

AI models, agents, software, and applications require scalable infrastructure for development and deployment. A consistent hybrid cloud platform allows you to develop, tune, test, deploy, and manage AI models and applications in the same manner across all parts of your infrastructure, giving you more flexibility.

It also supports sovereign AI and private AI strategies by letting you keep sensitive data and models in specific regions or even disconnected environments to meet data residency, privacy, and compliance requirements, while still connecting to public cloud services when it makes sense. Self-service capabilities can speed resource delivery while maintaining IT control.

Finally, a consistent platform supplies a foundation for technology integrations from third-party vendors, open source communities, and any custom tools you use.

Best practices and recommendations

Select a security-focused platform that supports hardware acceleration, a broad ecosystem of AI and application development tools, and integrated genAIOps and operations management capabilities.

Look for strong policy controls for data locality, model placement, and access so you can run sovereign and private AI workloads on premise or in private clouds while connecting to public clouds when needed.

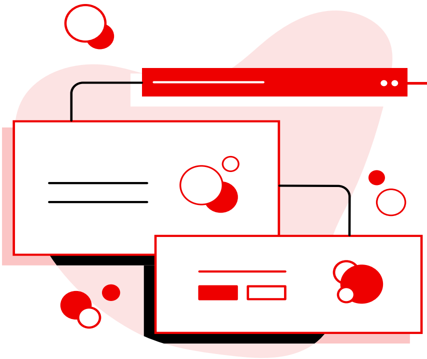
Choosing an open source platform can provide more integration opportunities and flexibility, fostering rapid innovation through community-driven development, as well as self-service capabilities to speed resource delivery while maintaining IT control.

A hybrid mix of public cloud and dedicated on-premise infrastructure is the most common digital infrastructure architectural strategy.[3]

[3] IDC White Paper. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Document #US53418426, Oct. 2025. (purchase required)

Model customization and alignment

Modern AI-powered applications require models that reflect an organization's specific data, workflows, and business constraints. Aligning a frontier or open model with your proprietary information is how you move from general-purpose responses to accurate, domain-aware outcomes.



Why is it important for AI?

Gen AI and agentic AI depend on models that understand your terminology, data, and real-world context.

Alignment helps maintain accuracy and relevance by grounding models in your private data. It improves efficiency by reducing inference cost and avoiding unnecessary oversizing. It strengthens governance and control by letting you implement business logic, safety rules, and compliance requirements directly into model behavior. It also supports scalability by giving you consistent processes to update, retrain, and version models as your data evolves.

Customization also supports sovereign AI and private AI strategies, allowing organizations to train and serve models entirely within controlled environments to meet data residency, privacy, and regulatory requirements.

Best practices and recommendations

Adopt modular workflows that start with RAG, fine-tuning, prompt engineering, and policy layers based on your needs rather than relying on a single method. Use open models to avoid lock-in and maintain the ability to fine-tune, quantize, and evaluate models transparently. Include subject matter experts to make sure models reflect real business context and data accuracy. Optimize your model for inference early by applying techniques like quantization, distillation, and efficient runtimes to control cost and latency. Furthermore, maintain strong governance with version datasets, training runs, model weights, and evaluation metrics to maintain reproducibility and strong governance.

AI platform considerations

AI inference at scale

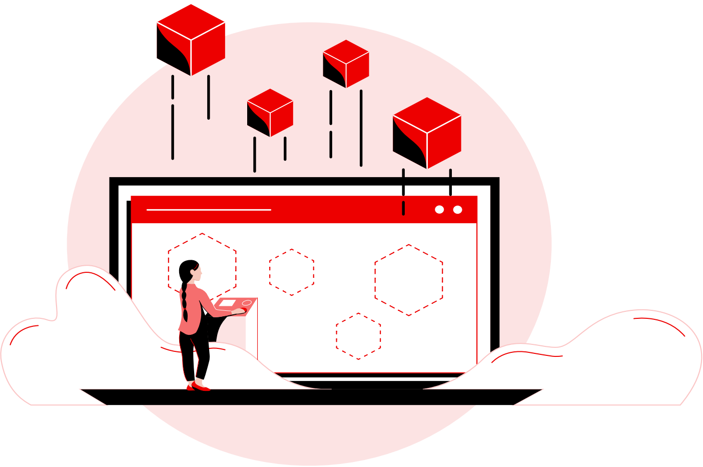
Running AI in production depends on fast, efficient, and reliable inference. Once models are trained or aligned, inference is the phase where they process new data, return predictions, generate content, or trigger actions inside an application or workflow.

As organizations adopt gen AI and agentic AI, inference becomes a critical cost and performance factor, especially as applications shift from single-query interactions to continuous, multistep tasks executed by AI agents.

Why is it important for AI?

Inference directly shapes user experience, application performance, and operational cost. Gen AI and agentic AI workloads often require rapid responses, parallel requests, and consistent throughput across many environments, from datacenters to public cloud to edge sites.

Efficient inference runtimes help reduce GPU and central processing unit (CPU) cost, improve latency for interactive tasks, and support the scaling needs of AI agents that call tools, use application programming interfaces (APIs), and coordinate multistep workflows. Optimizing inference also supports sovereign AI and private AI strategies by letting organizations run inference close to sensitive data, on premise or in private clouds, while maintaining predictable performance.

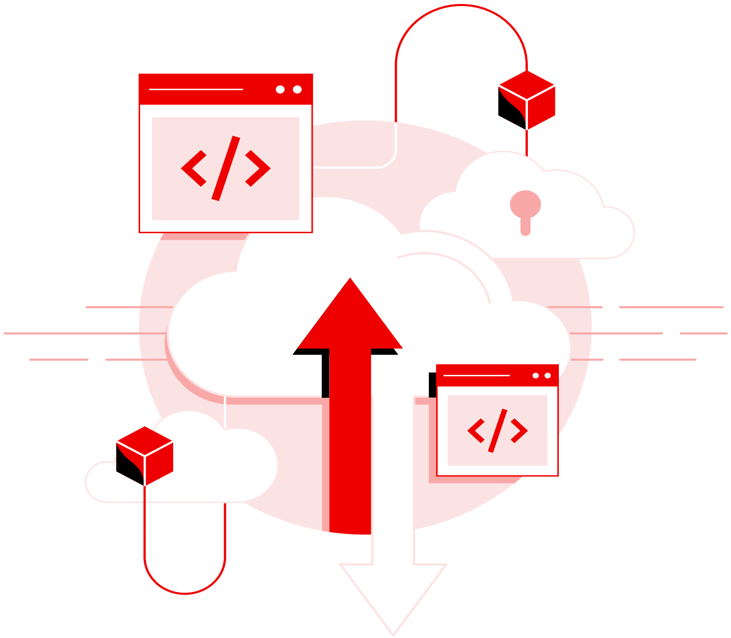


Best practices and recommendations

Select optimized inference runtimes that suit your model type and deployment environment, whether for LLMs, multimodal models, predictive models, or agentic workloads. Prioritize runtimes and infrastructure that support dynamic scaling, both horizontal and vertical, to meet the unpredictable demands of interactive LLM and agent-based inference.

Use techniques, or partner with vendors who have expertise, in approaches such as quantization, distillation, and model optimization to reduce cost and improve latency. Pair these optimizations with widely adopted technologies such as vLLM for high-throughput LLM inference and emerging distributed inference frameworks such as llm-d, which disaggregate the inference process to scale each phase independently.

Deploy inference inside containers to package dependencies and scale consistently across hybrid environments. Place inference endpoints where your data and applications live to reduce data movement and maintain control, especially for sovereign and private AI scenarios. Finally, monitor model performance over time and update versions as data distributions shift to maintain accuracy and reliability at scale.



90%

of decision makers believe AI will be an important driver of their digital infrastructure budget and technology choices through 2026.[3]

[3] IDC White Paper. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Document #US53418426, Oct. 2025. (purchase required)

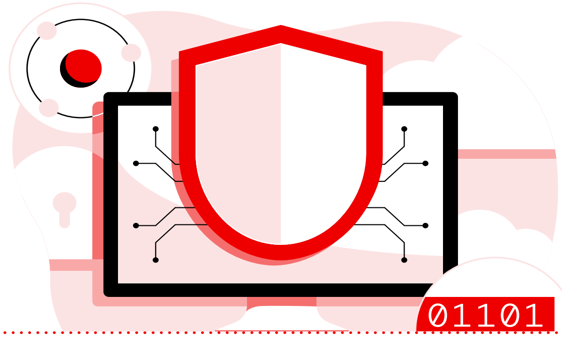
AI platform considerations

AI safety

AI systems must behave reliably, predictably, and in alignment with organizational policies. As enterprises move from experimentation to production, AI safety becomes a core requirement, especially when deploying gen AI, agentic AI, and autonomous workflows that can perform actions, not just provide suggestions.

Why is it important for AI?

Safety is focused on keeping AI models and agents within defined boundaries to uphold business, legal, and ethical requirements. Inaccurate outputs, model drift, insecure data handling, or unintended actions can create real operational risk. Gen AI and agentic systems also introduce new safety considerations, such as hallucinations, unapproved tool execution, privilege escalation, and inconsistent reasoning across multistep tasks. Strong safety practices help you maintain trust, safeguard sensitive data, and prevent harmful or irreversible actions. In regulated industries, safety controls are essential for compliance and audit readiness across hybrid and on-premise environments.



Best practices and recommendations

Adopt a layered safety approach that includes policy-based guardrails, content filters, and tool-execution controls for agentic workflows. Validate and test models regularly to detect drift or accuracy degradation. Run sensitive workloads in private or on-premise environments to maintain control over data exposure and model behavior—aligning with sovereign AI and private AI strategies. Use model-evaluation frameworks to monitor bias, robustness, and reliability. Look to augment your models and data through tools that store them in standard container (OCI) compliant registries and provide secure supply chains. Widely adopted technologies such as vLLM for LLM inference and emerging distributed technologies such as llm-d can help reduce costs and scale your AI project deployment. Finally, version and document your models, datasets, and policies so you can trace decisions and manage consistent governance across the full AI lifecycle.

Build an open, flexible foundation for AI

Red Hat AI Enterprise is an integrated AI platform for developing and deploying efficient and cost-effective AI models, agents and applications across hybrid cloud environments and is part of the Red Hat AI portfolio.

It unifies AI model and application lifecycles to increase operational efficiency, accelerate delivery, and mitigate risk by providing a ready-to-use development environment with enterprise-grade capabilities.

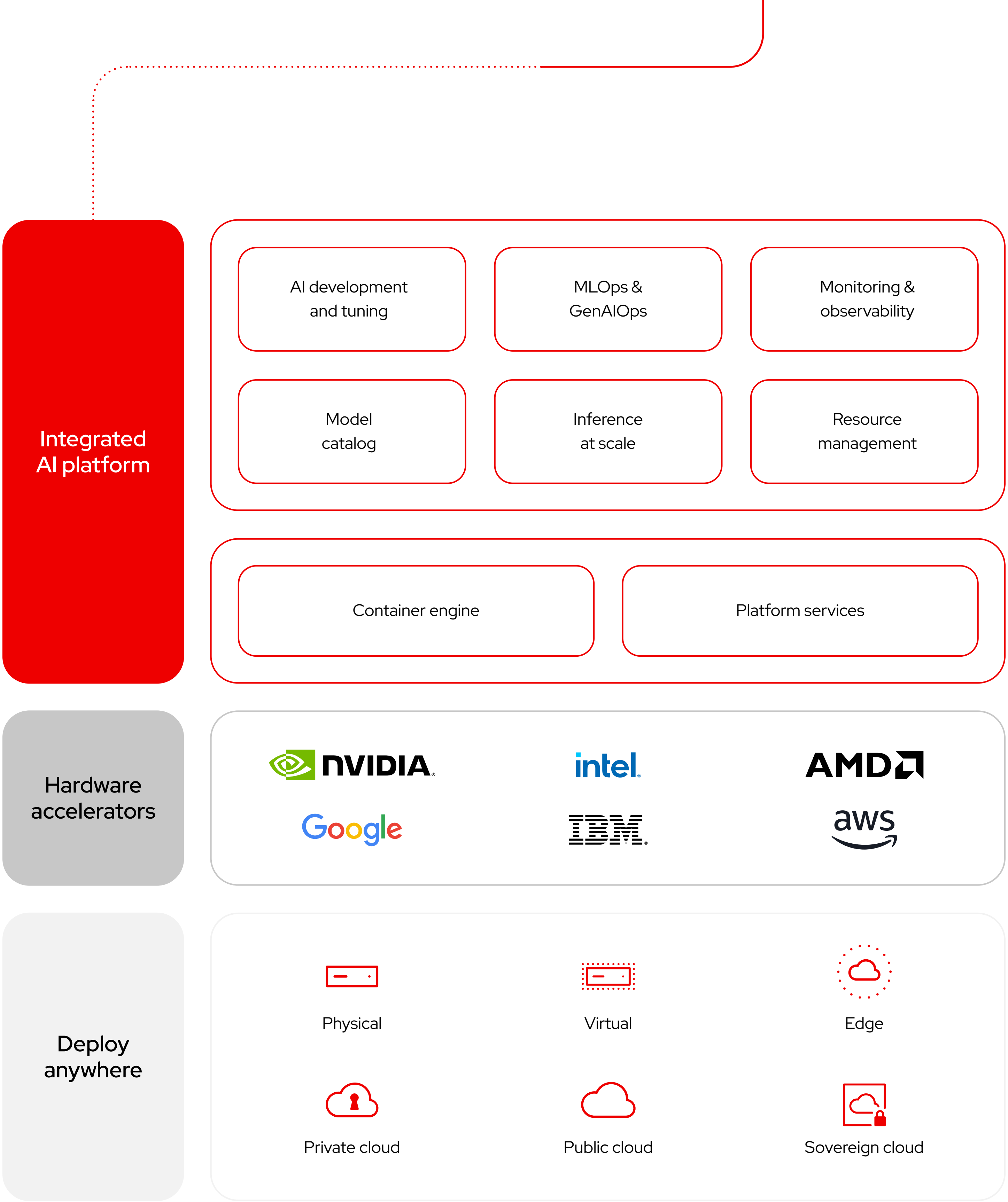
This platform is a tested, supported full AI stack, powered by Red Hat OpenShift, that enhances interoperability and ensures business continuity. It includes core capabilities such as model tuning, high-performance inference, and agentic AI workflow management. This gives the flexibility to support any model, use any hardware, and deploy anywhere while meeting data location requirements. Red Hat AI Enterprise is supported across hybrid environments so teams can plan capacity, GPUs, and future AI projects with confidence.

Red Hat AI Enterprise includes technology from the open source llm-d project, launched by Red Hat with collaborators such as IBM, NVIDIA, Google, AMD, and others. Llm-d improves cost efficiency by separating the prefill and decode phases of inference so each can scale differently. Its inference-aware load balancer routes requests based on token queues, improving response times and, in some cases, directing prefill workloads to CPUs.

Accelerate time to value.
Deploy an enterprise-ready AI stack on your infrastructure of choice, with preconfigured tooling, automated deployments, and built-in observability. This means developers and AI engineers can focus on building and delivering cloud-native AI-powered and agentic applications.

Increase operational efficiency.
Streamline and automate workflows, from code commits to establishing AI pipeline workflows through model deployment. This means IT operations can deliver consistent performance and get more value from existing infrastructure with intelligent resource allocation and integrated lifecycle management.

Mitigate risk.
Mitigate the risk of enterprise AI adoption with an integrated, tested, and fully supported AI stack that enhances interoperability across any model, any hardware, and hybrid cloud environments. Use this foundation to address data residency and regulatory requirements so you can scale AI with confidence.



Integrated AI platform

AI development and tuning

MLOps & GenAIOps

Monitoring & observability

Model catalog

Inference at scale

Resource management

Container engine

Platform services

Hardware accelerators

NVIDIA

intel

AMD

Google

IBM

aws

Deploy anywhere

Physical

Virtual

Edge

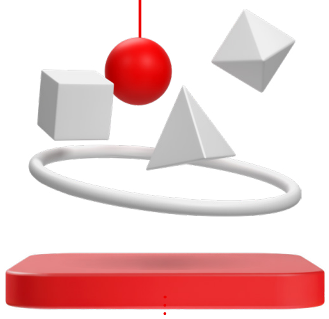
Private cloud

Public cloud

Sovereign cloud

Figure 2. *The components of an integrated AI platform.*

Gain choice and flexibility with a certified AI/ML partner ecosystem



The AI tool and technology landscape continues to evolve rapidly, making it challenging to keep up with advances while maintaining stability and reliability within your IT environment.

Through partnerships with NVIDIA, AMD, Intel, and AI technology partners, Red Hat AI Enterprise provides an end-to-end enterprise AI platform that scales in the hybrid cloud, offering faster deployment, improved efficiency, and hybrid cloud support. Red Hat's validation and certification programs ensure hardware is fully utilized and optimized workload management ensures efficient GPU usage, maximizing performance and value for customers

The Red Hat presence in the Hugging Face ecosystem and the Model Context Protocol (MCP) servers catalog gives customers access to a growing library of validated models and preintegrated tools that run consistently with Red Hat AI Enterprise. At the same time, partnerships with multiple accelerator providers help organizations take advantage of GPUs and specialized hardware across hybrid environments. You can confidently choose the partners, models, tools, and technologies that best fit your needs, knowing they will work reliably together and are backed by expert services, support, and training to help you build and scale AI workflows successfully.

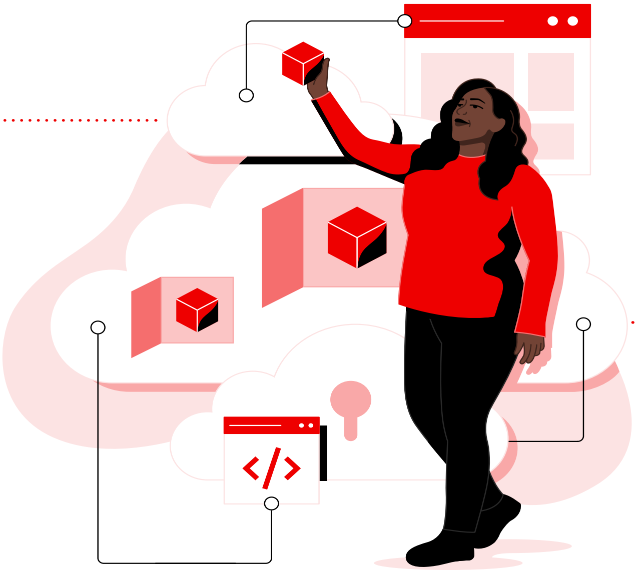
Partner highlights

NVIDIA

NVIDIA is a global leader in AI technology and provides the leading GPU architecture and platform for AI and accelerated computing, delivering leading performance and energy efficiency for training and deploying large AI models. NVIDIA and Red Hat have a strategic partnership focused on delivering optimized, cloud-native infrastructure and software for AI. Pairing with Red Hat AI Enterprise, customers are provided with a complete, optimized, cloud-native suite of AI and data analytics software and the hardware to run it on. Red Hat AI Enterprise, and NVIDIA DGX systems deliver IT manageability for AI infrastructure. The NVIDIA GPU Operator automates the management of all NVIDIA software components needed to provision GPUs.

AMD

AMD and Red Hat work together to expand choice for AI and virtualization across hybrid cloud environments. The collaboration brings AMD EPYC processors and AMD Instinct accelerators to Red Hat AI Enterprise, giving organizations more flexibility for running AI training, tuning, and inference workloads on cost-efficient, high-performance hardware. Joint engineering and certification efforts make sure that AMD-powered systems run consistently with Red Hat AI Enterprise, including support for optimized container images, Kubernetes operators, and validated performance for GPU-accelerated AI pipelines.



intel.

Intel and Red Hat work together to offer software defined infrastructure and industry-standard platforms that improve datacenter agility and flexibility. Intel's distribution of the OpenVINO toolkit optimizes and converts AI models into high-performance inference engines that can automatically scale to thousands of nodes on Red Hat AI Enterprise.

Success in action

Turkish Airlines

Turkish Airlines uses Red Hat AI to modernize operations and pioneer AI-powered innovation across aviation. By standardizing on an open, scalable AI platform, the airline accelerates model development, enhances passenger services, and streamlines operational decision-making—showing how hybrid AI can transform one of the world's largest airline networks.

Read more

Denizbank

Denizbank uses Red Hat AI to accelerate AI innovation across its digital banking ecosystem. By modernizing its AI infrastructure with an open, scalable platform, the bank speeds experimentation, improves model reliability, and delivers smarter customer experiences, demonstrating how hybrid AI helps financial institutions move faster while maintaining strict security posture and governance.

Read more

AGESIC

AGESIC, Uruguay's digital government agency, uses Red Hat AI to standardize and scale AI across more than 180 public entities. The hybrid AI platform supports MLOps practices, strengthens security, and helps teams build, deploy, and govern AI applications that improve services for citizens.

Read more

Ready to get more from your data?

AI is transforming nearly every aspect of business. Red Hat can help you build a production-ready AI environment that speeds development and delivery of intelligent applications to support your business goals.

Find out more about how Red Hat AI Enterprise can help you build a unified platform for AI.