Red Hat

From AI prototype to production

Operationalize AI faster with Azure
Red Hat OpenShift and Red Hat OpenShift AI

Introduction

The potential of artificial intelligence for enterprise

Artificial intelligence (AI) is no longer a futuristic vision. It's a present day force reshaping industries, optimizing operations, and redefining customer engagement—and it's still growing. In fact, IDC forecasts AI infrastructure spending will reach US$223B by 2028, with servers deployed in cloud environments at 82% of the market total and accelerated servers around 74% of the total market spending.[1]

55% of leading digital organizations chose AI as a top new investment priority.[3]

Generative AI (gen AI) in particular stands out among other forms for its ability to create text, images, code, and even business insights with unprecedented speed and accuracy. Organizations across industries are exploring how to adopt AI to increase organizational efficiency, enhance decision-making, and create new revenue streams.

To bring new AI-enabled solutions to market, many organizations are adopting or expanding their use of public cloud. IDC research found that more than 50% of respondents stated that they are deploying AI applications and solutions in the public cloud.[2]

Yet, despite its promise, implementing and scaling gen AI across an enterprise presents a complex set of challenges. From navigating data privacy and governance requirements to integrating AI into existing workflows, businesses must carefully plan their approach to avoid inefficiencies, security vulnerabilities, and costly missteps.

This e-book provides a practical roadmap for enterprises looking to move beyond experimentation and operationalize gen AI with confidence. We'll explore key considerations, including infrastructure and model selection, ethics, compliance, and user adoption, to help organizations build an AI strategy that delivers measurable value.

1 IDC press release. "Artificial Intelligence Infrastructure Spending to Surpass the $200Bn USD Mark in the Next 5 years, According to IDC," 18 Feb. 2025.

2 IDC White Paper, sponsored by Red Hat. "Why open source artificial intelligence platforms help enterprise business transformation." Document #US51272823, Nov. 2023.

3 IDC perspective. "AI Transformation Is Driving Digital Business," Document #US51136823, Jun. 2024.

Chapter 1

The challenges of adopting generative AI

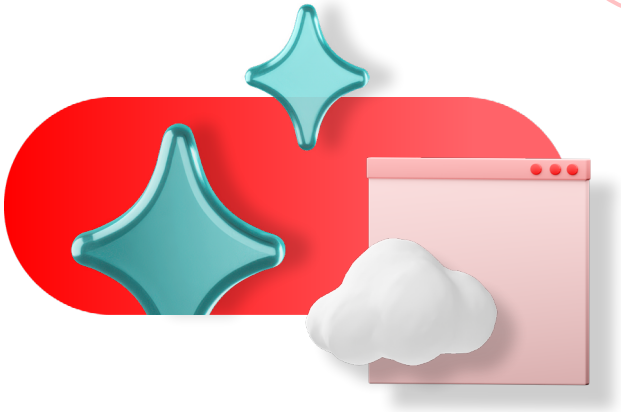
The key to a successful AI strategy isn't just adopting AI, it's understanding the challenges you are likely to encounter along the way so you can plan for them and navigate accordingly.

By identifying and addressing these pain points proactively, organizations can move beyond experimentation and unlock AI's full potential.

Here are 4 common roadblocks organizations are likely to face:

1. Cost
A considerable hurdle enterprises face when adopting gen AI is cost. Large Language Models (LLMs) provide powerful generative, predictive, and analytical functions, but can require significant computational resources and specialized infrastructure that comes with a significant price tag.

2. Complexity
Fine-tuning gen AI with enterprise-specific data is a complex task. Many organizations struggle to align LLMs to their business requirements and private data due to the complex training and tuning processes that are specifically designed for data scientists. This alignment complexity slows down AI adoption and innovation.



3. Flexibility
Enterprise AI strategies often evolve alongside business needs, requiring flexibility in how models get deployed. However, many enterprises find themselves locked into on-premise deployments or single public cloud providers, limiting their ability to move AI workloads across environments. As a result, enterprises may incur higher costs, performance and productivity issues, and compliance risks as they try to scale AI solutions across different parts of their business.

4. Operationalization
Many organizations struggle to move past AI experimentation into real-world deployment. This is largely due to a lack of tools, processes, and cross-functional collaboration between stakeholders. In the absence of a streamlined process and tight collaboration, AI initiatives often get stuck in disconnected development streams, which cannot provide value for the business.

These challenges highlight the need for a robust AI platform that provides the right automation tools and allows for clear governance frameworks to help organizations scale AI implementations successfully.

Chapter 2

Red Hat and Microsoft: A foundation for innovation

Since 2015, the Red Hat and Microsoft partnership has brought open source solutions to the cloud for enterprises across every industry. Together, Red Hat and Microsoft empower customers to create a security-focused hybrid cloud environment, with cloud-native development capabilities, that allows them to build the high-quality applications their business needs.

Central to this partnership is Microsoft Azure Red Hat® OpenShift®, a fully managed, cloud-based application platform that allows organizations to focus on building, deploying, and scaling applications efficiently, instead of managing infrastructure.

This widespread adoption highlights the reliability and performance of both platforms in powering critical workloads, AI applications, and enterprise innovation.

Learn more about Microsoft Azure Red Hat OpenShift.
Explore the video series.

Microsoft Azure and Red Hat are trusted by the world's largest organizations:

90% of Fortune 500 companies rely on Red Hat technologies.[4]
95% of Fortune 500 companies use Microsoft Azure.[2]

With Azure Red Hat OpenShift, enterprises benefit from:

- **A fully managed OpenShift environment,** reducing operational complexity.
- **Integrated security posture, compliance, and governance** features.
- **Streamlined hybrid and multicloud capabilities,** supporting workloads across cloud and on-premise environments.
- **A dedicated site reliability engineering (SRE)** team for cluster management, and increasing developer productivity.
- **A production-ready environment** with integrated tools and services to accelerate time to value.

Whether modernizing existing systems, optimizing enterprise IT operations, or deploying AI-powered applications, together Red Hat and Microsoft Azure provide the foundations for gen AI success.

4 Red Hat client data and Fortune 500 list, September 2024.

Chapter 3

Discover Red Hat AI

As you look to integrate AI into your operations, you will need a solution that is scalable, flexible, and cost-efficient so you can simplify model development and deployment across your hybrid cloud environments.

To get the most from your AI strategy and accelerate adoption, Red Hat provides Red Hat AI, a comprehensive AI portfolio that includes Red Hat Enterprise Linux® AI and Red Hat OpenShift AI.

What is Red Hat Enterprise Linux AI?

Red Hat Enterprise Linux AI is a platform for developing, testing, and running Granite family LLMs. It provides:

- A streamlined approach to gen AI with open source models.
- A foundation suited for both gen AI and predictive AI with integrated tools designed for enterprise.
- The ability to deploy AI across on-premise, public cloud, and edge environments.
- Access for developers and domain experts, even with minimal data science expertise.
- Support for training and inference on production server deployments, reducing infrastructure complexity.

What is Red Hat OpenShift AI?

An integrated MLOps platform, Red Hat OpenShift AI provides tools for managing the entire AI model lifecycle at scale. Key capabilities include:

- Support for both gen AI and predictive AI models, allowing you to bring your own models.
- Collaborative workflows and model service, allowing you to work across teams efficiently.
- Scalability for training, deploying, and monitoring AI workloads across on-premise and public cloud environments, improving flexibility and compliance.
- Ability to develop smaller, task-specific models that require fewer computational resources to tune and run.

When combined with Azure Red Hat OpenShift, Red Hat AI provides an integrated MLOps and DevOps platform that allows you to efficiently manage AI model lifecycles, optimize infrastructure, and focus on innovation instead of platform management.

Chapter 4

Augment your AI initiatives with Azure Red Hat OpenShift and Red Hat AI

As you scale your AI strategy, a security-focused, fully managed platform can simplify AI model development, help operationalize AI into your organization, and improve integrations.

Azure Red Hat OpenShift and Red Hat AI provide the necessary tools and solutions to accelerate AI adoption, reduce complexity, and optimize costs–all while maintaining compliance across your hybrid cloud environments.

The following key areas set it apart.

1. A fully managed AI-ready platform

Azure Red Hat OpenShift is a fully managed platform, which means it is designed to reduce operational burden, so your teams can provision clusters on demand with a self-service model while benefiting from:

- Ongoing management, including everything from infrastructure provisioning to daily operations.
- Automated maintenance, updates, and lifecycle management handled by a team of expert SREs.
- Built-in monitoring and proactive issue resolution, reducing downtime and streamlining AI deployments.
- A 99.95% update service level agreement (SLA).

A fully managed platform gives you:

50% improvement in operational efficiency.[5]
70% shorter development cycle, enabling teams to iterate and innovate faster.[5]
20% increase in developer productivity.[2]

2. A comprehensive application platform

More than just a Kubernetes platform, Azure Red Hat OpenShift provides fully integrated tools and services that increase developer productivity and operational efficiency. Built-in platform capabilities include:

- Built-in continuous integration and continuous delivery (CI/CD) pipelines for streamlined development.
- Monitoring, logging, and security tools that eliminate the need for do-it-yourself integration.
- An OpenShift Operator framework that lets developers choose from a variety of open source tools including service mesh, serverless, third-party tools and developer tooling that streamlines application and AI model deployment.

By reducing the complexity of tool integration, your organization can focus on delivering AI-powered applications rather than managing infrastructure.

Read more about how to transform your applications with Azure Red Hat OpenShift. Download the e-book.

5 Forrester Total Economic Impact™ (TEI) report sponsored by Red Hat. "Microsoft Azure Red Hat OpenShift Provides More Value And Support To Cloud-First Organizations", 19 Mar. 2024

3. Integration of MLOps and DevOps workflows

When you're ready to move your AI models from experimentation to production, collaboration between teams, including data scientists, ML engineers, and DevOps teams, can be a difference maker. Red Hat OpenShift AI extends Azure Red Hat OpenShift to help you connect these dots with:

- Integrated MLOps and DevOps workflows, helping pave the way for cross-team collaboration.
- A single, unified platform where AI models and applications are developed, deployed, and managed efficiently.
- Robust model training, fine tuning, serving, and monitoring capabilities, which can help accelerate time to market and allow your organization to realize the value of your AI investment sooner.

4. Hybrid cloud consistency and deployment flexibility

Having complete control over where and how your AI workloads run is essential to operationalizing AI. Azure Red Hat OpenShift supports:

- Different hardware accelerators, cloud providers, and on-premise servers.
- Flexible deployment options, including private datacenters, edge computing, and hybrid and multicloud environments.
- A bring-your-own-model (BYOM) approach, allowing your organization to tailor AI models with private enterprise data for specific business needs.

OpenShift cloud services delivers more than 450% return on investment (ROI) over 3 years with less than 6 months to payback.[5]

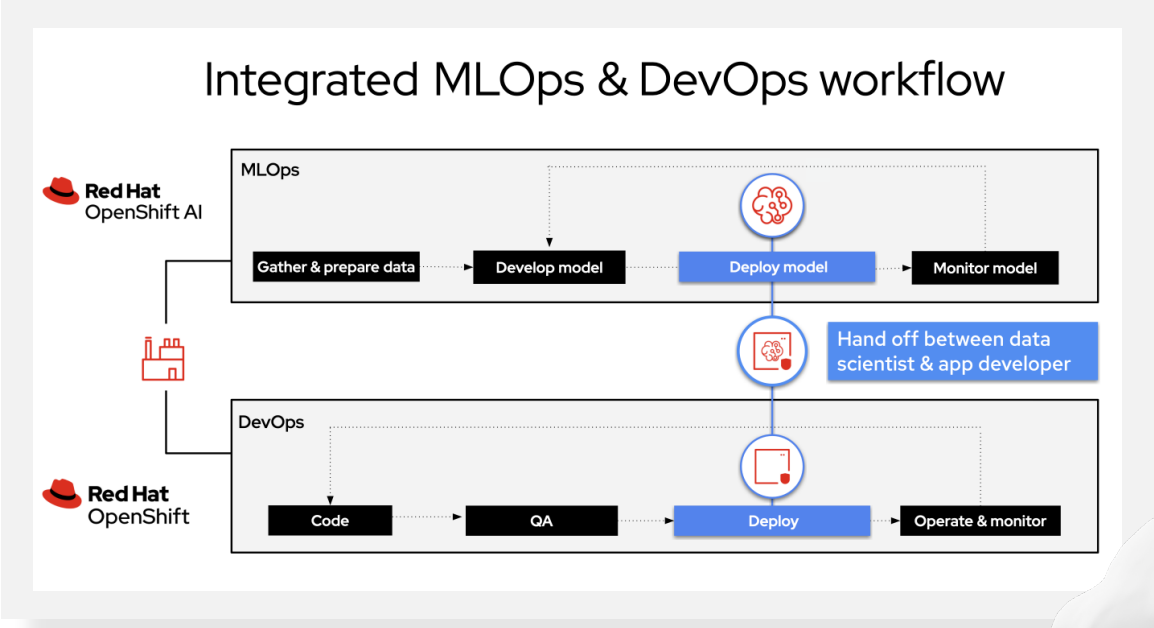


Figure 1. An overview of how to integrate MLOps and Devops workflow with Red Hat.

5. Cost optimizations for smaller AI models

Gen AI can be costly. Red Hat OpenShift AI helps to reduce expenses by supporting small, purpose-built AI models that:

- Are optimized with private enterprise data, enhancing accuracy while controlling costs.
- Reduce computational overhead, making them more cost-effective than large foundation models.
- Provide flexibility for business-specific AI tasks, improving efficiency and ROI.

6. First party, Azure-native solution

As a first-party Azure-native service jointly managed and supported by Red Hat and Microsoft, Azure Red Hat OpenShift integrates with key Azure services for application development, AI deployment, and security measures. This allows your organization to:

- Access 24/7 joint support from either Red Hat or Microsoft.
- Benefit from simplified billing through Microsoft, with the ability to draw down on Microsoft Azure Consumption Commitment (MACC) committed spend.
- Gain native integration with Azure security and compliance standards.

Explore all of the Azure integrations.

7. Benefit from an open AI ecosystem

Azure Red Hat OpenShift and Red Hat AI operate within an open source ecosystem, giving you more choice and control over your AI implementation.

- Gain a broad choice of AI hardware and software solutions tailored to your organization's needs.
- Access a community of technology providers, independent software vendors (ISVs), and open source projects.
- Experience full transparency and control over your AI models, allowing your organization to modify and refine AI systems based on your business objectives.

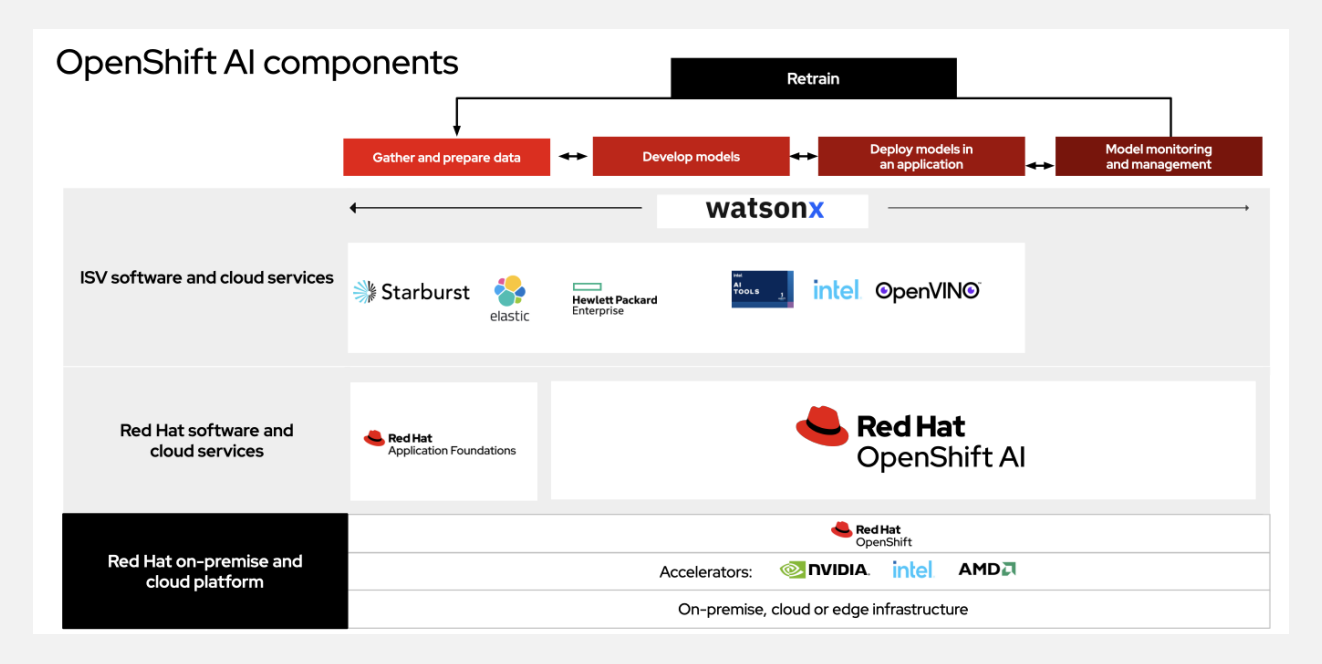


Figure 2. Red Hat OpenShift AI supported by a robust ecosystem of partners.

8. Start your AI journey wherever you are

Your organization can adopt AI at your pace, as Red Hat provides support across all stages of the AI adoption process, including:

- **Enterprises that are just starting with AI**. Red Hat AI streamlines AI deployment with prebuilt, ready-to-use models.
- **Advanced AI teams**. Red Hat AI allows for custom model training and fine-tuning with enterprise data.
- **API-accessible AI models**. AI models deployed in Red Hat AI are accessible through application platform interfaces (APIs), making them more straightforward to integrate into existing applications.

9. Built-in security focus and compliance

No AI strategy is complete without a measured and ongoing effort to safeguard valuable data and meet necessary compliance requirements. Azure Red Hat OpenShift is designed with the highest level of security standards and attention to compliance, to help protect sensitive AI workloads. It meets critical industry standards, and compliance offerings including but not limited to:

- PCI DSS
- SOC 2
- HiTrust
- ISO 27001
- FedRAMP

With Red Hat Enterprise Linux AI, your organization gains access to open source IBM Granite LLMs, fully supported and indemnified by Red Hat under the Apache 2.0 license.

Chapter 5

AI in action: A finance company delivers innovation to customers faster

Headquartered in Rotterdam, Netherlands, Ortec Finance provides technology and Software-as-a-Service (SaaS) solutions for risk and return management to help clients manage their investment decisions.

To enhance its services, Ortec Finance sought to integrate advanced AI capabilities into its solutions. However, deploying and managing AI models at scale presented a number of challenges, including infrastructure complexity, prolonged deployment times, and a need for greater collaboration between its development and operations teams.

Ortec Finance, in collaboration with HCS Company, built the Ortec Finance Cloud Application (ORCA) platform on Azure Red Hat OpenShift. This solution provided a fully managed, cloud-native environment for deploying and scaling AI-powered financial solutions. The platform streamlined software deployment, enhanced access management, and allowed for scaling across regions while ensuring compliance. It also improved collaboration between development and operations teams, accelerating the transition of AI innovations from research to production.

With Azure Red Hat OpenShift, Ortec Finance cut software delivery lead times from days to minutes, reduced operational overhead, and freed teams to focus on AI innovation. Looking ahead, the company plans to adopt Red Hat OpenShift AI to further enhance its AI/ML capabilities, ensuring scalable, compliant, and efficient AI-powered financial solutions.

"Modernizing our workloads and unifying our teams on Azure Red Hat OpenShift is already helping us be a faster-moving business and we can see how OpenShift AI on Azure Red Hat OpenShift will accelerate the transition of our AI-enhanced solutions from research and development labs into the hands of customers."

Joris Cramwinckel
Technologist, Ortec Finance

Read the complete customer story.

Learn more

Ready to get started?

Red Hat and Microsoft streamline AI adoption and application modernization with Red Hat OpenShift AI on Azure Red Hat OpenShift.

This fully managed platform provides a unified foundation for AI-powered applications, with integrated MLOps and DevOps workflows, scalable AI model management, and built-in security measures and compliance. With a vast partner ecosystem, expert support, and streamlined hybrid cloud flexibility, you can develop, deploy, and operationalize AI with less effort—accelerating time to value and innovation.

- **Learn** how to operationalize AI in less time with this checklist
- **Explore** the interactive demo
- **Try** Red Hat OpenShift AI
- **Get started** with Microsoft Azure Red Hat OpenShift