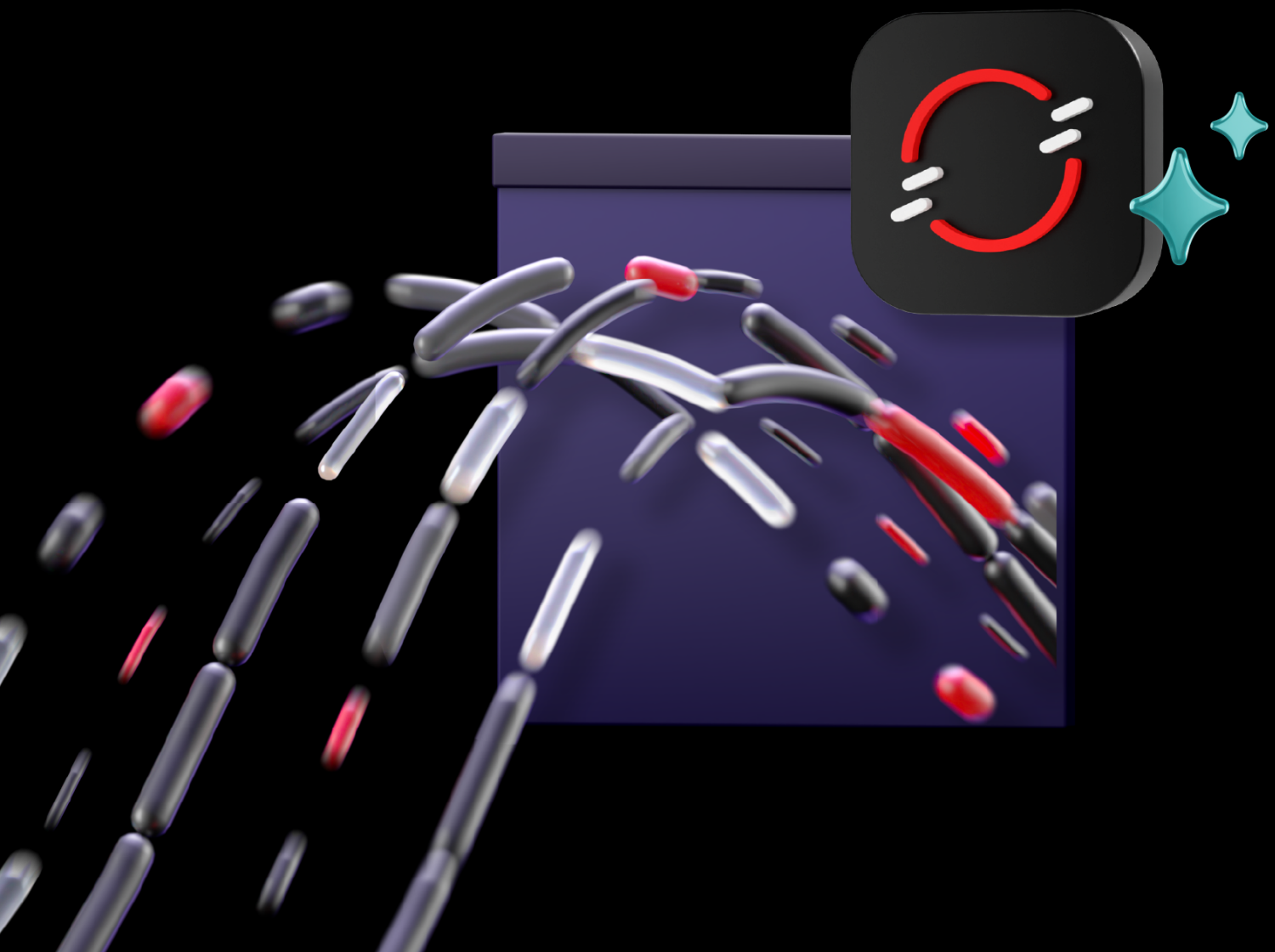
Red Hat

Operationalize AI with Red Hat and AWS

Accelerate your AI journey with Red Hat OpenShift Service on AWS and Red Hat AI



Introduction

The potential of AI for enterprise

AI has moved from the fringe of innovation to the core of enterprise strategy. Expanding beyond simple use cases and test environments, AI is now a business imperative—powering everything from predictive analytics and customer service automation to intelligent supply chains and product recommendations.

Across industries, organizations are under increasing pressure to translate AI investments into measurable value quickly, securely, and at scale. Executives and IT leaders recognize the promise of generative AI (gen AI) and machine learning (ML) models, with infrastructure spending estimated to reach US$223B by 2028,[1] but many struggle to move beyond proof of concept.

For many organizations, the challenge lies not in the ambition but in the execution. Complex infrastructure requirements, disconnected workflows between data science and IT operations (ITOps), and a lack of consistency between environments can all slow progress.

To manage the rising complexity of these environments, a unified AI application platform and support can help. In fact, according to the Forrester Total Economic Impact™ Study, organizations using application platforms saw up to a 70% reduction in application development time and 50% improvement in operational efficiency.[2]

This e-book aims to guide you through the journey of operationalizing AI with confidence. We'll delve into how the tools and technology platform can help your organization overcome common barriers, accelerate deployment, and achieve measurable outcomes.

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 Forrester Consulting, sponsored by Red Hat. "The Total Economic Impact™ Of Red Hat OpenShift Cloud Services." Feb. 2024.

Chapter 1

Make the leap to use gen AI more effectively

Across industries, AI continues to be tested and deployed to improve customer experiences, streamline operations, accelerate content and code creation, and uncover new business models. But there are a few key challenges holding teams back.

Common barriers to using gen AI more effectively include:

- **High costs.** Running large models requires significant compute power and storage, particularly when real-time responses and high availability are needed. For many teams, even pilot projects can quickly rack up unsustainable costs.
- **Rigid infrastructure.** Too often, AI development environments are built on systems not designed to easily move from model experimentation into deploying at scale. When teams can't easily shift between environments or deploy models where they're needed, momentum stalls.
- **Operational complexity.** From training to deployment to monitoring, AI workloads require coordination and expertise across multiple teams. In the absence of shared tools and infrastructure, organizations struggle to bridge the gap between data science and operations.

What begins as a bold AI strategy can soon become a patchwork of disconnected efforts and purpose-built tools, and slow time to value.

Shift toward modern platforms

To overcome these challenges, more organizations are turning to modern application platforms. These architectures are flexible, scalable, and designed to support both traditional applications and emerging AI workloads, offering the ability to provide common open source AI and application development tooling and frameworks.

At the heart of this shift is the growing adoption of Kubernetes-powered infrastructure, which allows for modular, containerized applications to be deployed and managed at scale.

A modern foundation gives teams the ability to:

- Work across hybrid or multicloud environments with consistency.
- Scale up or down automatically based on workload needs.
- Simplify deployment pipelines while reducing configuration overhead.

However, AI doesn't operate in a vacuum. It needs to be integrated with existing systems and processes. Modern platforms offer a way to embed AI workloads into the broader operations of enterprise IT.

The advantage of a unified, managed approach

Even with the right architecture, there's still the question of who manages it. Operating a modern, scalable AI platform is resource-intensive, requiring deep expertise in both infrastructure and data systems.

Many organizations are realizing that time and talent are limited, and the more teams tied up in managing infrastructure, the less they are available for building differentiated AI experiences. As a result, many organizations are ready for managed application platform services that offer flexibility without the management and maintenance burden.

A managed platform approach allows for:

- Accelerated implementation and iteration of new solutions.
- Simplified collaboration between developers, data scientists, and IT teams.
- Reduced risk through built-in security posture, compliance, and reliability.

In an environment where speed and scale are key to staying competitive, organizations need to focus on differentiated outcomes, not upkeep.

Chapter 2

Build on a solid foundation with Red Hat OpenShift Service on AWS

To successfully operationalize AI at scale takes more than tools alone—it requires the right technology foundation. It's important to keep in mind that the shift from prototyping to production is rarely a linear path.

Operationalizing AI requires coordination between teams, consistency across environments, and infrastructure that scales with demand. That's why organizations are increasingly turning to platform-based approaches, helping them unify development, deployment, and operations within a single ecosystem.

- **55% of leading digital organizations chose AI as a top new investment priority.**[2]

Discover Red Hat OpenShift Service on AWS

Red Hat® OpenShift® Service on AWS brings this vision to life as a fully managed application platform, jointly engineered and supported by Red Hat and Amazon Web Services (AWS). It delivers a powerful, comprehensive application platform with a focus on security, and deep AWS integration - without the overhead of managing the underlying infrastructure.

With Red Hat OpenShift Service on AWS, enterprises benefit from:

- **A fully managed OpenShift environment**, reducing operational complexity and day to day maintenance.
- **Integrated security posture**, compliance, and governance features.
- **Streamlined hybrid and multicloud capabilities**, supporting multiple workloads across cloud and on premise environments.
- **A dedicated site reliability engineering (SRE) team** for cluster lifecycle management and automated daily operations, to increase developer productivity.
- **A comprehensive, 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, Red Hat and AWS provide the foundation for gen AI success.

Fully managed and fully supported means:

- **99.95%** financially backed service-level agreement.
- **24x7 joint support from** Red Hat and AWS.
- **Platform automation** and **Day 2 operations by global site reliability engineers.**

Chapter 3

Discover Red Hat AI—flexible, scalable, and ready

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

Red Hat Enterprise Linux AI is a foundation model platform designed to consistently develop, test, and run large language models (LLMs).

Red Hat OpenShift AI is an AI platform for managing the lifecycle of predictive and generative AI models, at scale and across hybrid cloud environments.

Red Hat AI Inference Server is a solution that optimizes model inference across the hybrid cloud, creating faster and more cost-effective model deployments.

Building and deploying AI isn't just about choosing the right model; it's about making sure your teams have the tools, infrastructure, and support to bring those models to life.

Red Hat OpenShift AI is built with these goals in mind and provides a comprehensive and reliable platform that helps organizations move from isolated experimentation to real-world implementation.

Whether your teams are developing predictive models, implementing gen AI, or embedding machine learning into applications, Red Hat OpenShift AI is designed to support the full AI lifecycle—from data preparation to model training and tuning to deployment and monitoring.

A consistent platform for DevOps and MLOps

Collaboration between developers, data scientists, and IT operations teams depends on a consistent, reliable environment to build, test, and deploy applications and models.

A consistent application platform provides this shared foundation, allowing:

- Developers to build and integrate AI-enabled features into applications more quickly.
- Data scientists and AI engineers to train and fine-tune models without waiting for infrastructure provisioning.
- Platform operations teams to manage AI workloads and resource utilization with better visibility and control.

Integrated MLOps & DevOps workflow

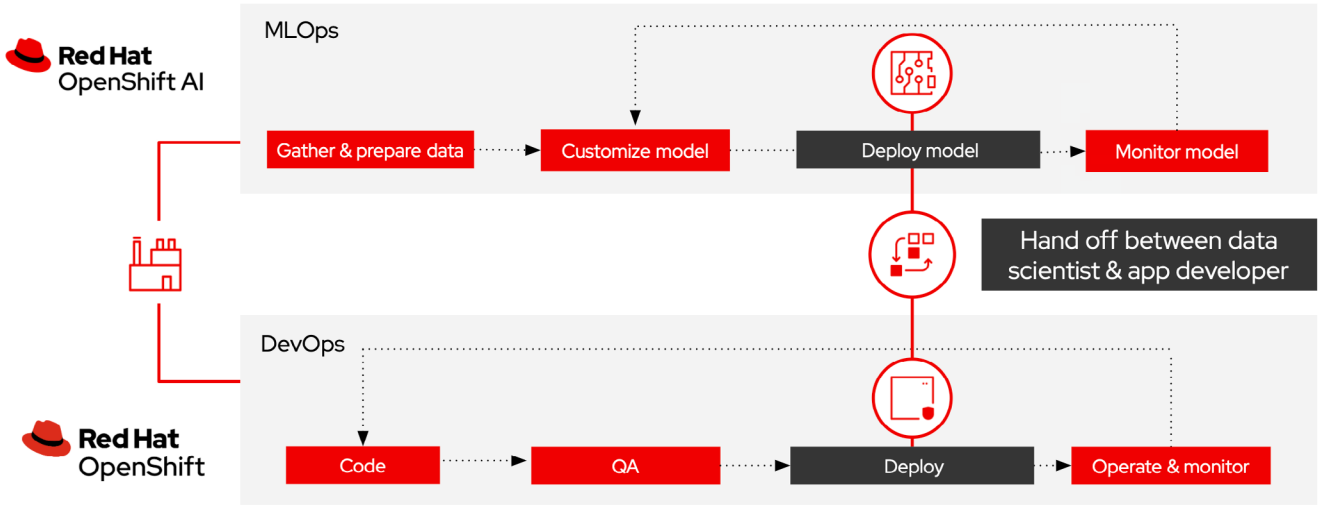


Figure 1. An overview of how to integrate MLOps and Devops workflow with Red Hat.

The convergence of DevOps and MLOps creates a more agile and efficient workflow where AI workloads can be developed, deployed, and scaled alongside traditional applications on the same infrastructure.

An integrated MLOps/LLMOps 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 or augment popular open models like Llama and DeepSeek.
- Collaborative workflows and self-service model tooling, allowing you to work across teams efficiently.
- Scalability for training, tuning, serving, and monitoring AI workloads across on-premise and public cloud environments, improving flexibility and compliance.
- Ability to develop smaller, task-specific models with optimized inferencing (vLLM) that require fewer computational resources to tune and run.

When combined with Red Hat OpenShift Service on AWS, 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.

Be the private AI provider for your organization.

Extend your AI capabilities at scale

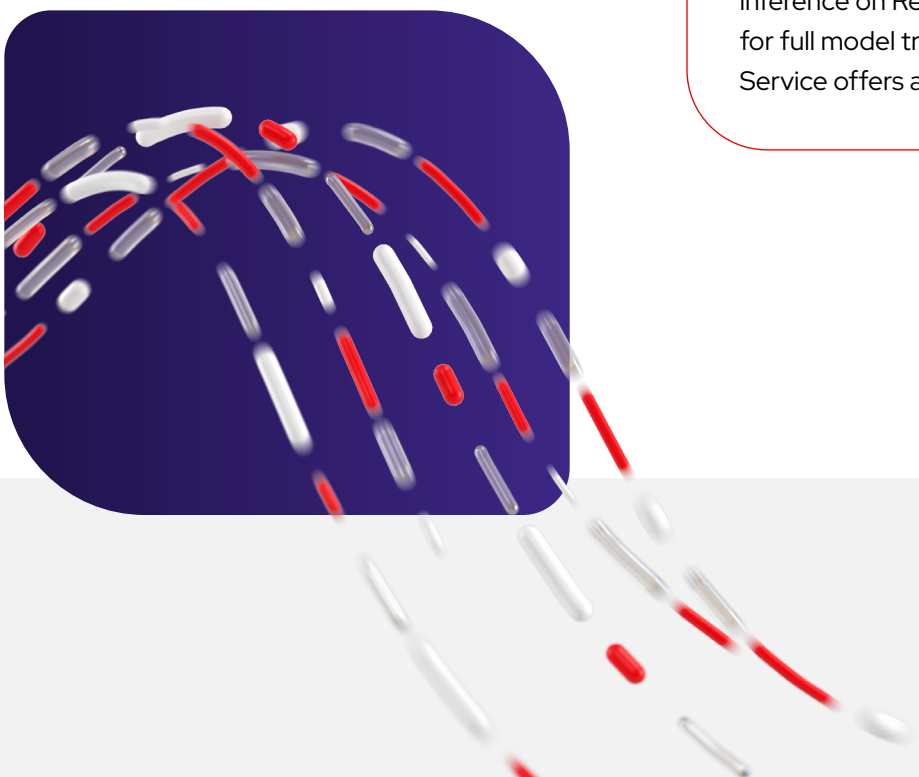
Red Hat OpenShift AI builds directly on Red Hat OpenShift, extending the platform with integrated tools and capabilities for AI model development, training, and serving. This integration allows organizations to embed AI workflows into the same platform they already use for application development—streamlining the path from experimentation to production.

This integration allows organizations to:

- **Increase efficiency at scale.** Red Hat OpenShift AI provides access to optimized, smaller models that are faster and more cost-effective to train and serve. Tools such as **virtual large language model** (vLLM) serving runtimes help manage inference workloads efficiently by reducing compute demands while maintaining high model response accuracy.
- **Reduce operational complexity.** With AI lifecycle management built in, and advanced tooling to automate and simplify operational tasks, teams can reduce complexity from model tuning to deployment. Teams can more effectively manage AI accelerators like GPUs, and data practitioners can self-serve environments tailored to their workloads.
- **Gain hybrid cloud flexibility.** Whether running in the public cloud, a private datacenter, or at the edge of the network, Red Hat OpenShift AI supports AI workflows wherever data resides. This allows organizations to meet unique regulatory and performance requirements while retaining the freedom to evolve AI strategies over time.

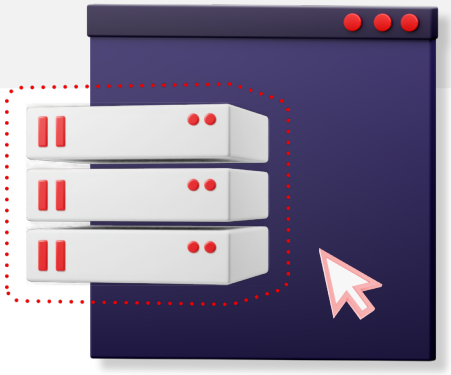
Serve and scale your AI models with Red Hat AI Inference Server

For organizations focused on running optimized AI inference on Red Hat OpenShift, without the need for full model training pipelines, Red Hat AI Inference Service offers a streamlined solution. **Learn more**



Chapter 4

Operationalize AI at scale with Red Hat OpenShift Service on AWS



As you expand your AI strategy and deploy more complex use cases, a security-focused, fully managed platform can simplify AI model development, improve integrations, and help operationalize AI into your organization.

Red Hat OpenShift Service on AWS and Red Hat AI provide the tools and solutions you need to operationalize AI at scale, reduce complexity, and optimize costs–all while maintaining compliance across your hybrid cloud environments.

Together, Red Hat OpenShift Service on AWS and Red Hat AI offer the following key benefits:

1. Flexibility and choice through hybrid capabilities

Every organization's AI journey is unique. Red Hat OpenShift Service on AWS combined with OpenShift AI gives you the freedom to deploy workloads wherever they can provide the most value—without losing consistency, control, or support. That's because it is:

- Available as either a fully managed cloud service or self-managed software product, suited for any operational need.
- Designed to support model lifecycle management at scale with an integrated MLOps/LLMOps platform and tooling that runs consistently across hybrid and multicloud environments.
- Backed by open source communities and a broad AI partner ecosystem, offering deep integration with leading technology providers and independent software vendors (ISVs).

2. Operationalize AI faster

Going from a few models in production to a dozen or more requires scalability and the infrastructure to support it. With the right tools on hand, organizations can speed up delivery while maintaining the reliability and reproducibility that DevOps principles demand, allowing teams to:

- Accelerate AI application development and deployment with tools as well as access to preoptimized LLMs ready for model training, customization, testing, serving, and monitoring.
- Eliminate time lost to integration challenges with preconfigured environments that unify data science, engineering, and operations workflows.
- Establish scalable environments that allow teams to move more quickly, without compromising governance, security focus, or reliability.
- Reduce time-to-deployment with Git integrations, built-in continuous integration and continuous delivery (CI/CD) pipelines, GPU orchestration, and accurate scheduling for AI workloads.

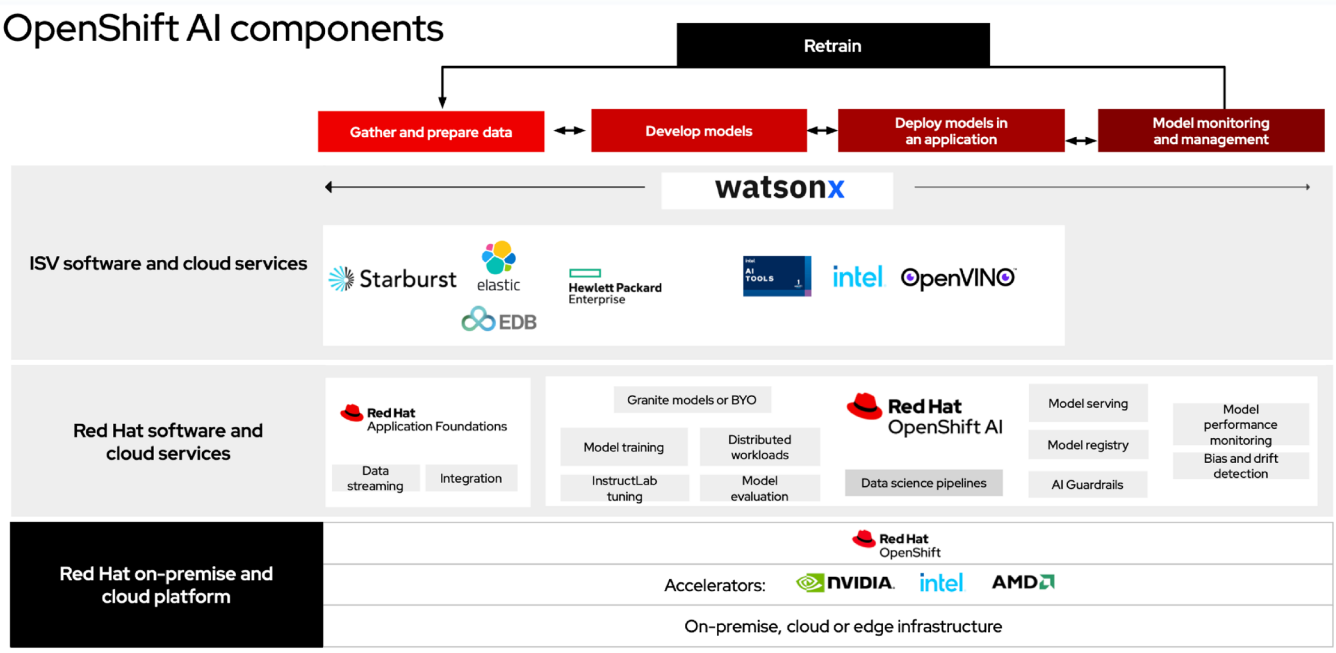


Figure 2. Red Hat OpenShift AI supported by a robust ecosystem of partners.

3. A focus on innovation

Reduced complexity means your teams can focus on what they do best. With a fully managed platform and integrated tooling, data scientists, developers, and operations teams spend less time maintaining infrastructure and more time delivering valuable outcomes. Red Hat OpenShift Service on AWS provides:

- A fully managed platform, supported by Red Hat SREs, which lets teams shift from infrastructure management to model innovation.
- A shared environment that integrates data science tools with DevOps pipelines to streamline collaboration across teams.
- A consistent platform for any workload across any environment, which helps speed innovation now and in the future.

4. Cost optimizations for smaller AI models

Red Hat OpenShift AI allows for smaller, domain-specific models to be fine-tuned with private data, reducing infrastructure costs while maintaining or even improving performance for targeted use cases, so you can:

- Create smaller models tailored to business-specific tasks which offer higher efficiency and lower training and inference costs.
- Improve performance while minimizing latency and runtime costs in production environments.
- Unlock a better return on investment by aligning model complexity with business impact—without overinvesting in oversized foundation models.

5. Built-in security posture and trustworthiness

As AI becomes more deeply integrated into business processes, security and transparency are essential. This is not just for infrastructure, but for the models and data they rely on. With built-in controls and open architectures, Red Hat AI helps you:

- Gain built in security with Red Hat OpenShift Service on AWS plus take advantage of the security features in AWS.
- Incorporate transparency into model lineage and training data with open model architectures, through the built-in model registry which is crucial for compliance and explainability.
- Reduce AI attack surfaces by deploying smaller, domain-specific models that are easier to audit and manage.
- Apply AI guardrails that filter unsafe inputs and outputs to maintain responsible, domain-appropriate use of generative models.

Learn more

Ready to get started?

Red Hat and AWS streamline AI adoption and application modernization with Red Hat OpenShift AI on Red Hat OpenShift Service on AWS. 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.

available in aws marketplace

Learn more about Red Hat OpenShift AI on Red Hat OpenShift Service on AWS.