AI Inference for financial services: Speed, trust, and efficiency

Summary:

Red Hat AI Inference Server bridges the gap between model creation and actual deployment into production—capturing tangible benefits.

Bridging the gap from model development to market value

Institutions in the finance sector employ various AI models for specific use cases. They use predictive models for credit assessment, chatbots, or generative AI (gen AI) models for customer interaction, and deep learning (DL) models for fraud detection. Yet, the challenge of deploying a trained model at scale efficiently, safely, and economically still exists.

Red Hat® OpenShift® AI, powered by Red Hat AI Inference Server, is a beneficial solution that financial operations teams can use to close the gap between model creation and deployment into production. By unifying the entire machine learning operations (MLOps) lifecycle on a trusted, open hybrid cloud foundation, Red Hat can help financial institutions advance their AI projects to function at optimum production at scale. At this level, business-critical decisions are made in microseconds, compliance becomes non-negotiable, and the return on investment (ROI) reaches maximum effectiveness across the hybrid cloud landscape.

An especially valuable feature of AI Inference Server is flexibility. With the capability to use models (Chat GPT, Claude, Llama, or another open source model) and graphical processing units (GPUs), such as Nvidia and AMD, organizations can run workloads on-premise, on any choice of cloud environment, or hybrid infrastructure.

The optimized AI configuration can provide a unified application and AI platform, which helps deploy AI models and applications that work together and use agentic AI.

For firms in the financial services sector, OpenShift AI and AI Inference Server can translate directly into:

- Accelerated deployment: Reducing the time-to-market (RTTM) for new AI-led services by months, maximizing the return on data science investments.
- Mitigated model risk: Providing auditable governance, transparency, and continuous monitoring to meet stringent regulatory requirements.
- Optimized total cost of operations (TCO): Lowering the operational cost of serving massive models (especially gen AI) through advanced, open source optimization techniques like virtual large language model (vLLM) and continuous batching.

Using a strategic advantage in AI deployment

Red Hat AI consists of: OpenShift AI, Red Hat Enterprise Linux® AI, and AI Inference Server. The product suite provides a foundation for running any model, on any choice of infrastructure, with any accelerator. The product portfolio is designed to simplify architectures, provide scalable deployments, and expand existing product use across multiple use cases. The featured benefits reside in several large groupings, including:

1. **Migration from pilot projects to solutions by using AI inference and agentic components**

The ability to use AI inference as a cost-reducing measure is gaining attention. As AI adoption moves from training models to production deployment, serious attention is being dedicated to reducing cost factors, including the enablement of AI inference. Optimizing inference efficiency is becoming critical for financial services organizations to achieve long-term value and overcome cost barriers, which along with regulation and security are important industry concerns.

AI Inference Server is a component of OpenShift AI, based on the open source project vLLM. It provides a scalable foundation to optimize AI inference in production—when an AI model provides an answer based on data. The principal benefit of using Red Hat AI Inference Server is to enable large language models (LLMs) to perform calculations more efficiently and at scale, allowing optimization techniques like quantization and sparsity to reduce the expense of running AI at scale.

LLMs can be customized for a particular use case, for example, using a separate LLM to manage fraud, customer service, and risk. Routing tasks at scale via automation on a hybrid cloud helps get tasks completed effectively using the model built for running that request.

Agentic AI is a technology that banks can use to make strategic decisions and take real-time actions using AI software systems designed for autonomous, goal-oriented behavior—while interacting with data and tools that work with minimal human intervention. It helps smaller, specialized models work together and allows them to collaborate on complex tasks, delivering high-quality outcomes more efficiently.

The operational benefits of AI Inference implementation

- Reduced TCO:
  - The largest cost factor: As AI scales, inference costs increasingly dominate the TCO. Institutions that optimize AI inference can facilitate its long-term value.
  - Optimization techniques for production: Based on vLLM, techniques like quantization that use 8-bit floating point (FP8), 8-bit integer (INT8), 4-bit integer (INT4), and sparsity are supported. Model sizes can be reduced and efficiency boosted to lower AI operating expenses significantly.
- Scalability and performance:
  - Simplified architecture and scalable deployments: Red Hat provides a scalable foundation that simplifies architectures and offers efficient deployment across multiple use cases.
  - Increased accurate responses in less time: Using vLLM, helps LLMs perform calculations more efficiently and at scale, improving speed and accuracy. Risk management is strengthened and customer experience is improved.
  - High-performance for agentic AI: The combination of Red Hat solutions and appropriate hardware component accelerators can provide a reliable platform for running complex agentic AI solutions. Smaller, fit-for-purpose models can be combined to deliver high-quality outcomes efficiently using techniques like inference-time scaling.

- Flexibility, compliance, and control in hybrid cloud environments:
  - Hybrid deployment: The platform allows banks to run safe and compliant AI workloads wherever the data resides—on premise or in a cloud environment—addressing principles of data gravity and regulatory requirements for strict control over data location and use.
  - Vendor flexibility: Red Hat AI provides a vendor-neutral foundation, allowing banks to use gen AI models and accelerators in their choice of cloud environment. Flexibility gives organizations the agility to adapt and control their AI strategy more effectively.
  - Prevalidated models: Red Hat prevalidates a growing set of production-ready models such as Llama, Gemma, and Mistral via the Hugging Face open source repository. Hundreds of models give banks confidence and predictability in their inference deployments.

2. **How speed and scale directly apply in the financial services sector**

The complexity of competition in the financial services industry has changed because having a better algorithm is no longer enough. Financial firms must be able to scale reliably and deploy in less time with more security, efficiency, and flexibility. Additionally, inference latency factors are a direct measure of business performance, risk control, and customer experience.

Application uses:

- Algorithmic trading: Market-moving models must execute and respond to data streams in near-instantaneous timeframes. The ability to manage and orchestrate workloads across accelerators (NVIDIA, AMD, and Intel Gaudi) ensures performance and responsiveness.
- Fraud and financial crime: Real-time fraud detection models must score transactions with more speed than the payment network can settle. A millisecond delay can result in large losses. Red Hat delivers an engineered solution for microsecond-level latencies, supporting continuous, high-volume transactional scoring.

Scale while mitigating committed spend

The shift to gen AI introduces unprecedented operational expansion and cost challenges for financial institutions. LLMs are compute-intensive and extremely expensive to run. AI Inference Server addresses these challenges through:

- Optimized cost efficiency: Means more models can be run to serve more users, reducing the TCO for gen AI deployments. Optimization provides support for leading open source techniques—such as vLLM and LLM distributed (llm-d)—with better memory management to maximize GPU use.
- Model compression: Allows LLMs to run on fewer resources, which further cuts inference expenses. The support of quantization and sparsity techniques can shrink massive LLMs.
- Accommodating and accelerating the adoption of agentic AI: Provides the path for running complex agentic AI solutions that allow smaller, fit-for-purpose models to work together to deliver high-quality outcomes. Goal-oriented systems can perform tasks autonomously, often by orchestrating with smaller, specialized models.

3. **AI governance and risk mitigation**

In financial services, performance without compliance becomes an existential liability. Globally, regulators are tightening scrutiny on AI systems, demanding transparency and accountability. The focus on data sovereignty has amplified the need for a compliant platform that can help banks maintain control over their data location and uses, including regulations and compliance rules.

OpenShift AI and its associated components are engineered to transform these regulatory mandates into competitive advantages.

Model Risk Management and auditability

Model risk is no longer limited to quantitative models; it now necessarily encompasses complex, opaque AI systems. Red Hat can deliver the capabilities required to satisfy evolving global Model Risk Management (MRM) guidelines.

- Model lineage and audit: Maintains a complete system of record (code, data, and versions) for every model to satisfy internal audits and regulatory reviews.
- Continuous monitoring: Automates detection of model deviation and bias in production, alerting teams to deviations before they cause regulatory breaches or financial loss.
- Explainable AI (XAI): Demystifies black-box models to provide clear explanations about decisions (e.g., loan denials), ensuring regulatory compliance and consumer trust.

Security and hybrid cloud resilience and flexibility

Red Hat OpenShift, the foundation of the OpenShift AI platform, provides proven, security-focused, and operational consistency. With a unified platform and focus on security, banks can maintain resilient and adaptable hybrid cloud environments in an ever-changing landscape.

- Security for financial data: Ensures that role-based access control (RBAC), network segmentation, and encryption are applied consistently across development, training, and high-stakes inference environments.
- Hybrid cloud consistency: Accommodate workloads with data sovereignty or legacy system features using reliable safety and compliance wherever the data resides. Operational consistency is maintained across on-premise datacenters, private cloud, and multiple public clouds.
- Operational resiliency: Provide high availability and business continuity. If an AI service fails, the platform's orchestration automatically ensures rapid recovery and failover, minimizing operational disruption.

4. **The path to value: Acceleration of ROI and MLOps maturity**

The promise of AI is about the speed and potential with which it can deliver business value. As models move from pilot projects to core banking applications, banks need a unified, hybrid, and modern platform that can scale efficiently and integrate smoothly into existing systems.

Potential economic benefits

Key economic benefits for financial institutions can include:

- RTTM: Implement a unified AI architecture that can reduce the time-to-market for new use cases (e.g., churn models or ATM cash optimization)
- Efficiency gains: Realize potential operational cost savings and reduction of developer time expenditures by consolidating disparate tools onto a single platform.
- Data scientist productivity: Boost the ability to provision infrastructure instantly and dynamically. Data scientist efficiency and reduction of routine tasking can aid in the retention and acquisition of top talent.

A unified platform for enterprise MLOps

By providing a vendor-neutral, scalable foundation, OpenShift AI provides a unified experience that helps remove barriers between data scientists, application developers, and IT operations:

- Self-service AI workbenches: On-demand access to computational resources (GPUs and accelerators) dramatically speeds experimentation and model training without the need to submit manual tickets to IT.
- Production inference pipeline: Integration of hardened open source technologies provides a prescriptive path that can move trained models into high-performance, containerized production with automatic scaling, intelligent routing, and deep observability.
- Ecosystem integration: Open architecture supports integration with preferred tools commonly used in the financial industry. Ecosystem integration helps prevent vendor lock-in while maximizing existing technology investments.

The perfect AI platform for financial resilience

In an environment defined by relentless competition and increasing regulatory scrutiny, the ability to operationalize AI with speed, governance, and cost efficiency is the ultimate differentiator. With predictive AI already widely adopted and scaled to automate and streamline operations, banks are now expanding into gen AI and exploring the use of agentic AI. OpenShift AI, with the optimized performance of AI Inference Server at its core, is the right foundation for a strategic and AI-capable financial institution.

Adopting smaller, more agile, and fit-for-purpose AI models offers banks a compelling alternative to LLMs. Extremely complex LLMs can carry significant operational costs and create regulatory and compliance challenges around digital sovereignty, data privacy, and risk management. Lighter models can deploy more rapidly and operate with fewer resources, giving banks greater adaptability and control. With enhanced performance, transparency, and auditability, AI investments yield accelerated returns, empowering organizations to adapt and maintain a competitive advantage.

Next steps and resources

Discover the effect of Red Hat OpenShift AI platform and the efficiencies gained by using Red Hat AI Inference Server to manage your organization's specific goals.

Learn more about Model Risk Management (MRM) and review several Red Hat customer gen AI use cases focused on cost optimization and deployment.

Explore additional resources:

- Read the Red Hat Inference Server press release: Red Hat Unlocks Generative AI for Any Model and Any Accelerator Across the Hybrid Cloud with Red Hat AI Inference Server
- Read a technical blog: Red Hat AI Inference Server
- Watch the video about: Red Hat AI Inference Server from Red Hat executives
- Visit these webpages to learn more about:
  - Red Hat AI
  - Red Hat OpenShift AI
  - Red Hat Enterprise Linux AI

About Red Hat

Red Hat helps customers standardize across environments, develop cloud-native applications, and integrate, automate, secure, and manage complex environments with award-winning support, training, and consulting services.

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

North America
1 888 REDHAT1
www.redhat.com

Europe, Middle East, and Africa
00800 7334 2835
europe@redhat.com

Asia Pacific
+65 6490 4200
apac@redhat.com

Latin America
+54 11 4329 7300
info-latam@redhat.com

redhat.com
#3402162_0226