Accelerating AI adoption for financial services with Red Hat

Reducing time to market for AI/ML solutions with an end-to-end platform

Increased complexity of AI models equals increased adoption challenges

Financial institutions are looking to capitalize on opportunities presented by the adoption of artificial intelligence (AI). The rapid pace of development in areas such as deep learning, conversational, and generative AI has vastly increased the scope and applicability of AI solutions. At the same time, the increased complexity of the models creates new execution challenges and emphasizes existing ones. Some of the challenges include:

- **Standalone development process:** Most AI and machine learning (ML) development and training is currently done in dedicated environments and requires special resources—accelerating hardware like graphic processing units (GPUs), for example. Provisioning AI/ML environments takes a long time and is a significant blocker to rolling out new AI-based services.
- **Scaling, flexibility, and resource optimization:** AI/ML solutions require components with varying resource needs, such as the central processing unit (CPU), memory, disk and specialized hardware (GPU), tensor processing unit (TPU), and field-programmable gate array (FPGA). Scaling such solutions often requires a hybrid cloud approach.
- **Monitoring and drift:** AI/ML models require continuous monitoring and regular updating to detect and correct drift. Red Hat® OpenShift® facilitates the continuous integration of model updates by providing a standards-based monitoring infrastructure that can connect application-based drift monitoring with the AI/ML development pipeline.
- **Model supply chain security:** The AI/ML developer tool ecosystem is heavily based on open source, community-driven frameworks. Ensuring software supply chain hygiene in this environment is a growing challenge. Developers want the latest tools, but enterprises need to ensure those tools are safe, security-enhanced, and contain no vulnerable or malicious artifacts.

Benefits that significantly reduce complexities

The proposed AI/ML solution provides financial institutions benefits that include:

- An end-to-end platform for model development, training, and inference. This allows for consistency in operations across public and private clouds and reduces the friction between the different phases of the process.
- Self-serving capabilities accelerate time to value for ML environments.
- A consistent set of state-of-the art open source ML tools and libraries, together with an extensive ecosystem of open source and partner-supported technologies.

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

- Rapid development and deployment of ML models, together with monitoring and rapid iteration capabilities that ensure the deployed models remain up to date.

Case study: Large language models

An example illustrating the challenges and benefits in financial services is implementing a large language model (LLM)-based solution, like GPT-4, BLOOM, BART, DOLLY, and others. These types of solutions are used for digitizing documents as part of onboarding or know your customer (KYC) processes; analyzing environmental, social, and corporate governance data (ESG) reports; or implementing conversational solutions, such as chatbots.

In such solutions, it is common to rely on large ML models with hundreds of millions or billions of parameters. Due to the effort, complexity, and required compute power to create these models, it is common to build upon pretrained or foundation models. Because these models are typically trained on general purpose datasets, applying them to the specific context of a financial services use case requires additional domain- or firm-specific training on a smaller set of local data, via fine-tuning or transfer learning. A sample architecture for this type of solution is outlined in Figure 1.

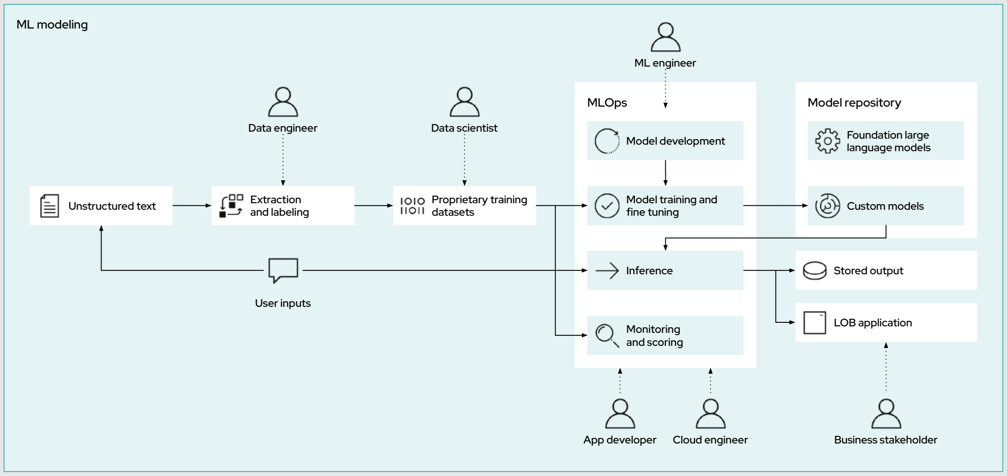

Figure 1. ML modeling overview

Capabilities overview

Solution architecture

Red Hat provides a platform that efficiently and productively hosts the full AI/ML life cycle from development to training to inference. Red Hat platform technology runs on major forms of infrastructure, from bare metal and on-premise virtualization to major public clouds. This means the same platform, with the same tools, using the same MLOps processes.

Understanding open source and the importance of safeguarding the software supply chain, Red Hat engages in upstream communities to develop great new software and trusted relationships. As part of that, Red Hat curates, supports, and certifies a large array of upstream tools that AI/ML developers require.

Let Red Hat do the work of understanding the upstream supply chain and provide your business with a product that you can rely upon and trust, with 24x7 support.

Platform components

Operating system

The foundation of Red Hat AI/ML architecture is Red Hat Enterprise Linux®, an operating system (OS) that can run on a modern deployment infrastructure on-premises or in a cloud environment, bare metal, or on virtual machines. Red Hat Enterprise Linux is certified to run on the broadest ecosystem of hardware and major cloud providers, including Amazon Web Service (AWS), Google Cloud, IBM Cloud for Financial Services, Oracle Cloud, and Microsoft Azure. The Linux platform brings security, performance, support, and world-class automation through Red Hat Ansible® Automation Platform. Finally, Red Hat Enterprise Linux provides support for specialized hardware for AI/ML model development, including GPUs and FPGAs.

Container orchestration

In addition to custom-built and commercially available applications, a vast majority of open source tools and libraries used in AI/ML processes are containerized. Pretrained or production ML models are also packaged as container images. In addition, AI/ML processes involve multiple components that must interact with each other and scale elastically. Such components include compute intensive training of new models, high throughput inference engines, and model development environments used by data scientists—all of which require a flexible and elastic platform. The industry leader for deployment and orchestration of containerized workloads is Red Hat OpenShift, a distribution of Kubernetes. This is the most popular platform for third-party and open source AI development tools, which ensures that your developer teams will have access to the AI/ML frameworks they need to accelerate time to value. Red Hat OpenShift also brings operator technologies to automate the deployment of components for self-service and reduced operational costs.

Scalable, security-hardened storage

AI/ML projects require large quantities of training data to build accurate models. This data can be historical or it can be live from sources that include market data feeds, Internet of Things (IoT), and observability. In all cases, it must be stored in a way that is user-friendly and repeatedly accessible to developers. Red Hat supports and integrates open source software-defined storage in the form of Red Hat OpenShift Data Foundation, based on Red Hat Ceph® Storage. OpenShift Data Foundation is a software-defined storage solution that integrates with Red Hat OpenShift and economically scales to Petabytes and beyond. Streaming data can be consumed with AMQ streams, based on Apache Kafka, to provide developers with replayable access to streaming data. Both OpenShift Data Foundation and AMQ streams, which are packaged in containers, can be managed with Red Hat OpenShift so that multiple developer teams can operate in a self-service fashion.

Platform capabilities

Self-service

Using Red Hat OpenShift, developer teams and projects can be onboarded on demand, and resources can be scaled up or down as needed. Additionally, costly specialized hardware, such as GPUs, can be pooled and shared. Security compliance and software supply chain safety are built-in at all levels.

Advanced monitoring and observability

Red Hat OpenShift includes open source industry standard monitoring through Prometheus and provides compatibility with third-party monitoring tools, such as Splunk. This allows the integration of MLOps pipelines with flexible, centralized infrastructure for monitoring and alerting throughout the entire pipeline. Tracking model performance can automate scaling and issue alerts for low accuracy levels.

Agility

AI/ML modeling is an iterative process. Practitioners (including both data engineers and data scientists) explore paths left by data trails, and the journey to model development is filled with starts and stops, unforeseen avenues, pitfalls, and dead ends. Typical challenges include access to quality data from different sources, such as databases, file systems, streams, application programming interfaces (APIs), and compliance with regulatory obligations and security standards. In terms of tooling, challenges include versioning across a wide array of libraries, along with updating existing tools and adopting new ones. Red Hat helps simplify the AI/ML pipeline for practitioners by empowering them with a consistent experience across a hybrid cloud environment to accelerate AI/ML projects.

One difference between traditional application development and AI/ML application development pertains to the pronounced need to update the applications themselves or AI models at the core of the applications. AI/ML techniques allow for not just initial training of a model through ML but also the continuous ability to update the model. This allows models to provide benefits not available to traditional applications, but it implies the need to periodically "close the loop" and update the models to enhance performance. Red Hat OpenShift gives application teams the ability to upscale and downscale components of the MLOps toolchain transparently. When your application requires a model update, the (more costly) training resources with GPUs or other specialized componentry can be assigned and expanded manually. When the update is complete, Red Hat OpenShift can reassign those resources to where they are most needed.

Scalability and elasticity for training and inference

The training phase of AI/ML modeling is one of the most resource-intensive operations in the MLOps pipeline. This is where the largest scale-out instances of AI/ML tools live and where the demand for specialized hardware—such as GPUs, TPUs, and FPGAs—from companies like Nvidia is highest. Individual projects and teams desire access to their own environments to perform their training. The ability for Red Hat AI/ML architecture to provide shared infrastructure offers significant advantages in efficiency and economy. Rather than hoarding dedicated resources at high cost, Red Hat OpenShift provides developers with virtual on-demand access to the entire cluster. Kubernetes orchestrates and mediates this access to ensure that such resources are delivered where and when the business needs them.

Open ecosystem

Red Hat AI/ML platform is fully open source, like all Red Hat products. The open source ecosystem of tools and technologies available to AI/ML practitioners include:

- ML libraries.
- AI/ML life cycle management.
- Data access, data quality, and metadata management.
- Bias detection and explainability.

- Pretrained models.

Due to the open nature of the ecosystem and the flexibility of the platform, these tools can be used together in various combinations, as the solutions require. Also, having an open platform supports continuous innovation by allowing emerging technologies, tools, and models to be continuously plugged into the solution.

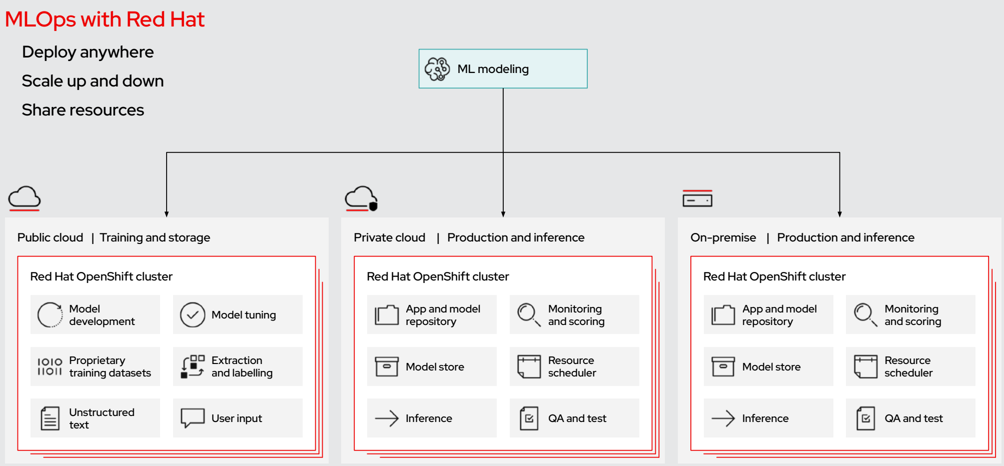

Figure 2. MLOps with Red Hat overview

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

North America	Europe, Middle East, and Africa	Asia Pacific	Latin America
1 888 REDHAT1	00800 7334 2835	+65 6490 4200	+54 11 4329 7300
www.redhat.com	europe@redhat.com	apac@redhat.com	info-latam@redhat.com