Explore AI at the edge with Red Hat

Enhance real-time decision-making with Red Hat's edge solutions

The growing need for real-time AI processing at the edge

Three key trends are transforming how organizations process and act on data: the proliferation of edge computing, advancements in artificial intelligence (AI), and the integration of open source platforms.

Edge devices—including internet of things (IoT) sensors, industrial controllers, and retail automation systems—are generating more data than ever.[1] This data explosion makes analyzing massive amounts of information increasingly complex. Organizations must therefore employ advanced analytical tools and AI models that can identify meaningful patterns and insights in diverse, messy, and distributed datasets.

AI models are becoming more efficient and adaptable, enabling real-time analytics, anomaly detection, and automation directly at data generation points. IDC projects spending on AI solutions to reach US$632 billion with a compound annual growth rate (CAGR) of 29.0% from 2023 to 2028.[2] Yet organizations adopting these AI models at the edge face challenges:

- **Governance of AI workloads at edge locations.** Managing workloads across thousands of distributed locations demands robust governance mechanisms. Organizations must adopt platforms that have a security focus and govern AI workloads consistently across the hybrid cloud environment.
- **Operationalizing models in distributed, disconnected environments.** The operationalization of AI models is frequently cited as the most challenging phase of AI implementation. This challenge is magnified at the edge, where models must function reliably across distributed, often disconnected environments.
- **Lack of MLOps engineers at the edge.** Edge environments often lack sufficient local IT resources, complicating the management and troubleshooting of AI models. Without adequate personnel at the edge, issues can require models to be relocated to centralized locations for diagnosis, increasing downtime and costs.
- **Inadequate data pipelines.** Reliance on closed source, proprietary systems restricts scalability and adaptability to evolving market demands.

To fully use AI at the edge, organizations must embrace a modern, open approach that encompasses:

- AI inferencing at the edge for real-time decision-making.
- Scalable and security-focused deployment across cloud, datacenters, and edge locations.
- Interoperable platforms unifying IT and operational technology (OT) environments.

[1] "Number of Connected IoT Devices 2024." IoT Analytics, Sept. 2024.

[2] IDC. "IDC FutureScape: Worldwide Artificial Intelligence and Automation 2025 Predictions." Document # US51666724, Oct. 2024.

Key benefits of AI at the edge

Accelerated decision-making with AI inferencing at the source

The use of AI models at the edge allows organizations to process data where it is generated rather than sending it to a centralized cloud or datacenter for analysis. This reduces latency, allowing for real-time decision-making in industries where milliseconds matter.

- **Retail example:** AI-driven inventory management can analyze in-store purchasing trends in real time, automatically adjusting stock levels and notifying suppliers of demand shifts.
- **Healthcare example:** AI-powered medical applications can analyze patient data locally, providing customized dosage suggestions.
- **Manufacturing example:** AI-driven predictive maintenance can detect equipment anomalies earlier, preventing failures before they disrupt operations.

Lower operational costs and reduced bandwidth consumption

The use of AI at the edge reduces the cost and complexity of data transmission by minimizing the need to send large datasets to central cloud servers. Instead, AI models filter and process data locally, transmitting only actionable insights.

- **Retail example:** Security cameras in large stores generate massive amounts of footage. AI at the source can process video feeds in real time, detecting suspicious activity locally and transmitting only relevant clips for further analysis.
- **Industrial example:** High-frequency sensor data in manufacturing can be analyzed at the edge, reducing the amount of low-value data sent to core systems while preserving bandwidth for critical operations.

Enhanced security and data privacy

Sensitive data often cannot be transferred to cloud environments due to regulatory and security concerns. The use of AI at the edge keeps data localized, reducing exposure to potential breaches and following compliance with industry regulations.

- **Healthcare example:** AI-powered diagnostics can analyze patient scans on-site, making sure that personal health data never leaves the clinic.
- **Manufacturing example:** Proprietary industrial data can remain within factory premises, protecting intellectual property and trade secrets.

Scalability and flexibility across IT and OT environments

As data volumes and workload complexity grow, the use of AI at the edge helps organizations to dynamically scale their capabilities. Deploying intelligent edge devices and servers across multiple locations, and automating their management, allows businesses to quickly adapt to changing demands.

- **Manufacturing example:** A global manufacturer deploys AI-driven predictive maintenance across factories, automatically scaling the solution as new plants come online.
- **Retail example:** A major retailer rapidly expands loss prevention AI to hundreds of new store locations, managing the entire deployment from a centralized platform.

Red Hat's open platform for AI at the edge

Red Hat helps organizations across industries—including finance, healthcare, and manufacturing—to use their existing hybrid IT infrastructure for edge AI deployment.

Our robust hybrid cloud platform, based on Red Hat® Enterprise Linux® and Red Hat OpenShift®, provides the foundation needed to confidently manage AI workloads across diverse environments.

Red Hat streamlines operationalizing machine learning (ML) at scale through integrated machine learning operations (MLOps), automating model management and accelerating time-to-value.

Open, modular architecture helps organizations avoid vendor lock-in, using community-driven innovation and supporting various hardware and cloud environments. This collaborative, flexible approach improves operational efficiency, reduces complexity, and bridges skill gaps, helping data scientists, IT teams, and developers to jointly accelerate AI implementation and continuously enhance model accuracy across all edge locations.

Red Hat's platform for AI at the edge is not a standalone solution, but a portfolio of Red Hat technologies designed to support AI workloads across cloud, on-premise, near edge, and far edge environments. Red Hat's approach allows AI models to be trained centrally and deployed efficiently at the edge (or the reverse), helping organizations to process data in real time, optimize operations, and maintain a security focus and regulatory compliance.

AI in constrained environments: Red Hat Device Edge and MicroShift

AI at the edge must often run on limited hardware in remote locations that lack the computing power of a traditional datacenter. Many industries require AI inferencing on small form factor devices, such as industrial sensors, security cameras, and IoT gateways. To support these needs, Red Hat offers **Red Hat Device Edge** and **MicroShift**, 2 technologies that bring Kubernetes-based AI inferencing to resource-constrained environments.

The Red Hat build of MicroShift is a lightweight, optimized derivative of Red Hat OpenShift, a leading container platform based on Kubernetes, designed specifically for edge computing. It allows organizations to deploy and manage AI models close to where data is generated, reducing latency and enabling real-time decision-making. The primary reason for using MicroShift with AI model serving is embedding the model into a complex solution that requires orchestrating multiple components, such as front-end interfaces, backend systems, data processing pipelines, and the AI model itself—all delivered as containerized microservices. This localized orchestration not only speeds up interactions between components, but also reduces the complexity and costs associated with managing distributed solutions.

For environments that do not require full Kubernetes orchestration, Red Hat Device Edge provides a more streamlined alternative. It allows AI workloads to be deployed as standalone containerized applications on single-board computers, industrial controllers, and other low-power devices. This flexibility makes it ideal for deployments in retail, logistics, and energy infrastructure, where AI models can analyze video feeds, optimize energy consumption, or detect security threats—all while operating independently from the cloud.

Managing AI across the full lifecycle: Red Hat OpenShift AI

AI at the edge is not just about deploying models; it requires a continuous lifecycle of training, deployment, monitoring, and retraining. Red Hat OpenShift AI provides a scalable platform that integrates AI into DevOps pipelines, helping organizations to manage their AI models efficiently across cloud, on-premise, and edge environments. When AI/ML lifecycle management is integrated into DevOps principles, it is called MLOps.

With OpenShift AI, companies can train or tune AI models centrally in a core datacenter or cloud environment and then deploy them to edge locations for inferencing. This approach is particularly useful for industries that require constant model updates, such as retail loss prevention systems that must adapt to evolving theft tactics or predictive maintenance systems that learn from new sensor data in industrial equipment.

A key advantage of OpenShift AI is its model serving capabilities, which allow organizations to automate the deployment of AI models at the edge. For example, a smart grid system can use AI to balance energy distribution in real time, processing inputs from thousands of sensors and making adjustments dynamically. OpenShift AI allows for newly trained models to be quickly deployed to edge locations, improving accuracy and adaptability.

MLOps is a set of workflow practices, inspired by DevOps and GitOps, that aims to streamline the process of deploying and maintaining ML models. By integrating MLOps principles, OpenShift AI also helps automate model monitoring and retraining. Organizations can track model performance, detect drift in AI predictions, and update models without manual intervention. This is critical in industries such as healthcare, where personalized medicine AI applications must continuously refine their recommendations based on new patient data, all while maintaining compliance with strict data privacy regulations.

Ready to accelerate your AI initiatives at the edge?

Download your product trial or contact your Red Hat account executive to get started.

About Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. A trusted adviser to the Fortune 500, Red Hat provides award-winning support, training, and consulting services that bring the benefits of open innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.

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

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