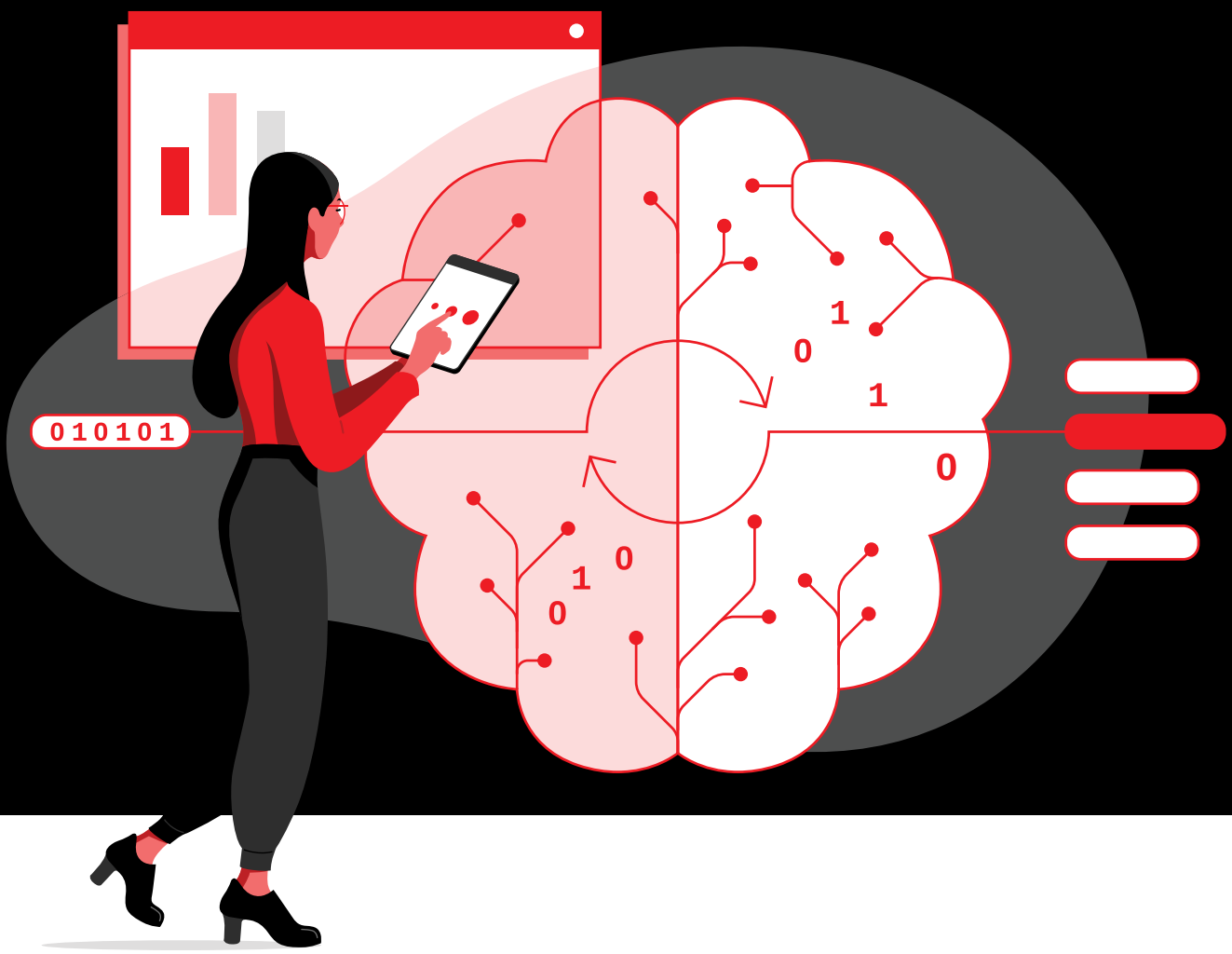
Red Hat



Top considerations for

building a production-ready AI/ML environment

Contents

Get more business value from your data

The amount of data created is expected to reach more than 221,000 exabytes by 2026.[1] In a digital world, your data can be a critical competitive advantage, but collecting data is only the starting point—how you use your data is the real differentiator.

Artificial intelligence (AI), machine learning (ML), and deep learning (DL) employ data to deliver business insights, automate tasks, and advance system capabilities. These technologies have the potential to transform all aspects of business, from customers and employees to development and operations.

Building AI/ML into your applications can help you achieve measurable business outcomes:

- Increase customer satisfaction.
- Offer differentiated digital services.
- Optimize existing business services.
- Automate business operations.
- Increase revenue.
- Improve decision-making.
- Increase efficiency and reduce costs.

Key technologies

This e-book discusses several technologies for actionable data analysis:

- **Artificial intelligence** involves machines imitating human behavior to perform tasks that typically require human intervention.
- **Machine learning** is a subset of AI that uses algorithms and statistical models to perform tasks without explicit instructions.
- **Deep learning** is a subset of ML that uses layers to progressively extract high-level features from raw input, similar to a human brain. For example, generative AI can create high-quality text, images, and other content based on trained DL models.
- **Machine learning operations (MLOps)** encompasses all of the tools, platforms, and processes needed to create, train, deploy, monitor, and continuously improve AI/ML models for use in cloud-native applications.

*1 IDC White Paper, sponsored by Dell Technologies and NVIDIA. "**High Data Growth and Modern Applications Drive New Storage Requirements in Digitally Transformed Enterprises**." Document #US49359722, July 2022.*

AI/ML use cases across industries

Across industries, AI/ML can help you deliver business outcomes faster.

Financial services

- Personalize customer services and offerings.
- Improve risk analysis.
- Detect fraud and money laundering.

Telecommunications

- Gain insight into customer behavior.
- Enhance customer experiences.
- Optimize 5G network performance.

Retail

- Optimize supply chains and inventory management.
- Improve customer insight and experiences.

Automotive

- Support autonomous driving technologies.
- Predict equipment maintenance needs.
- Improve supply chains.

Healthcare

- Increase hospital and clinic efficiency.
- Boost diagnosis speed and accuracy.
- Improve patient outcomes.

Energy

- Optimize field operations and maintenance.
- Improve worker safety.
- Streamline energy trading.

Manufacturing

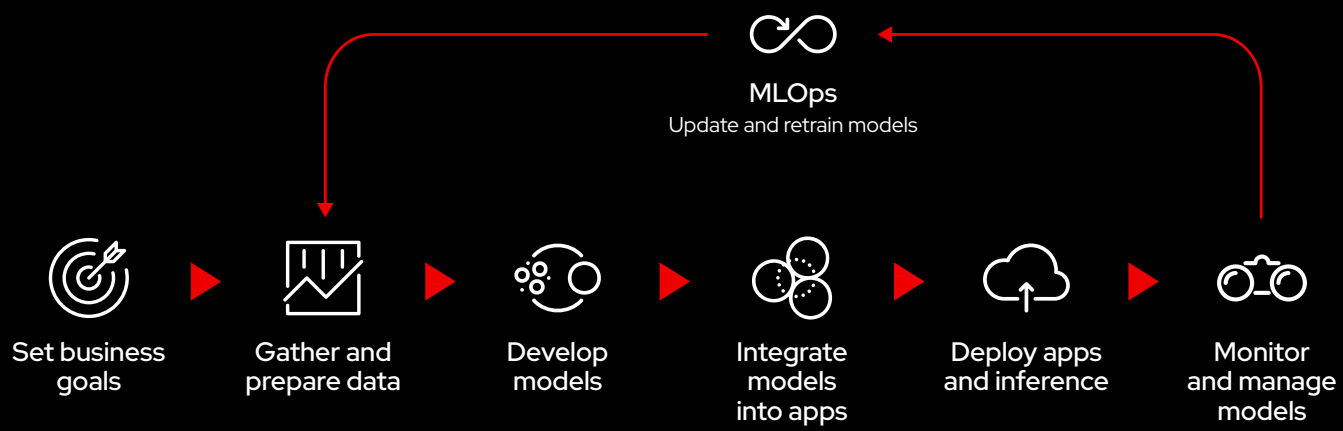
- Predict equipment failures.
- Perform preventative maintenance.
- Improve factory floor safety.

Build a production-ready AI/ML environment

Deploying AI/ML in production is an iterative process that extends beyond simply creating AI/ML models. The main steps in the AI/ML life cycle are:

1. Set business goals for your AI/ML initiative and share them with all stakeholders.
2. Gather and prepare the data required for your AI/ML initiative.
3. Develop models according to your goals.
4. Deploy models in your application development process.
5. Implement intelligent, ML-powered applications and start inferencing.
6. Monitor and manage models for accuracy over time.

AI/ML life cycle and MLOps



AI/ML deployment challenges

Organizations face several challenges when building an AI/ML environment:

- **Talent shortages.** A limited number of available AI/ML experts makes it harder to find and retain data scientists and engineers, ML engineers, software developers, and other staff with the right knowledge.
- **Lack of readily usable data.** Organizations collect large amounts of data but must find, prepare, and protect the appropriate data for each AI/ML initiative.
- **Disparate teams and technology.** Slow, manual, and disconnected operations and infrastructure can impede collaboration between teams and AI/ML deployment across your organization.
- **Delayed resource availability.** Slow infrastructure and tool delivery can hinder model development, integration, and deployment within applications.

Even so, you can overcome these challenges by applying cloud-native application development approaches to your AI/ML life cycle.

An open, adaptable architecture can help you adopt AI/ML and MLOps more effectively to achieve your business goals.

A production-ready AI/ML architecture requires several key technologies and capabilities:

- **AI/ML and MLOps tools** allow data scientists, ML engineers, and application developers to create, deploy, and manage ML models and applications.
- A **cloud platform** gives data engineers, data scientists, ML engineers, and application developers access to the resources they need to work rapidly.
- **Compute, storage, and network accelerators** speed data preparation, model development, and inferencing tasks.
- **Infrastructure endpoints** provide resources across on-site, virtual, edge, and private, public, and hybrid cloud environments for all stages of AI/ML operations.
- **Edge deployments** (optional) provide large amounts of data from devices and sensors that can be used to train models and gain insights in real time.

This e-book reviews key considerations for building an effective AI/ML architecture.

Containers

A **container** is a basic unit of software that packages applications with all of their dependencies.

Data scientists, ML engineers, and application developers need access to their preferred tools and resources to be maximally productive. At the same time, IT operations teams need to ensure that resources are up to date, in compliance, and used in a secure manner. Containers simplify application build processes and allow applications to be deployed across different environments without change. They let you deploy a broad selection of AI/ML tools across hybrid environments in a consistent way. Teams can iteratively modify and share container images with versioning capabilities that track changes for transparency. Meanwhile, process isolation and resource control improve protection from threats.

Recommendations for container solutions

Look for a robust, highly available container platform that includes integrated security features and reduces the complexity of deploying, managing, and moving containers across your environment. Choose an open source platform that integrates with a broad set of technologies to gain more flexibility and choice.

Container orchestration

Container orchestration involves managing the creation, deployment, and life cycle of containers across your environment.

Once you adopt containers, you need a way to deploy, manage, and scale them efficiently. A container orchestration tool lets you administer the life cycle of your containers in a consistent way. These tools typically centralize access to compute, storage, and networking resources across on-site, edge, and cloud environments. They also provide unified workload scheduling, multitenancy controls, and quota enforcement.

Container orchestration recommendations

Select a **Kubernetes**-based orchestration tool to take advantage of a leading open source technology.

Application life cycle management

Application life cycle management involves deploying, scaling, and administering applications that run within containers.

AI/ML environments are inherently complex. Container application life cycle management components that integrate with your container orchestration tool let you directly administer containerized applications, including AI/ML development tools. IT operations teams can automate common life cycle management tasks like configuration, provisioning, and updates to gain efficiency, speed, and accuracy. Data scientists, ML engineers, and application developers can use tools and applications from a preapproved service catalog—without needing to engage IT teams. Automation also frees staff from tedious tasks, allowing them to focus on more interesting strategic activities.

Application management recommendations

Choose container application life cycle management tools that include easy-to-use automation and integration with your preferred AI/ML tools. Popular options include **Kubernetes Operators** and **Helm Charts**.

MLOps practices

MLOps practices bring together the tools, platforms, and processes needed to operationalize AI/ML at scale.

Organizations need to develop and deploy AI/ML models—and the applications that use them—rapidly and efficiently. Collaboration across teams is critical for success in these efforts. Similar to **DevOps**, MLOps approaches foster collaboration between AI/ML teams, application developers, and IT operations to accelerate the creation, training, deployment, and management of ML models and ML-powered applications. Automation—often in the form of **continuous integration/continuous delivery (CI/CD)** pipelines—makes rapid, incremental, and iterative change possible for faster model and application development life cycles.

MLOps best practices

MLOps is not just about technology—people and processes play key roles. Apply **MLOps practices** to your entire AI/ML life cycle. Use automation in your platforms and tools, along with open source technologies like Argo, Kubeflow, Tekton, and Jenkins, to create CI/CD pipelines and workflows.

Hybrid cloud platform

A hybrid cloud platform provides a foundation for developing, deploying, and managing intelligent applications and models across on-site, edge, and cloud environments.

AI/ML models and intelligent applications require infrastructure for development and deployment. A consistent hybrid cloud platform allows you to develop, test, deploy, and manage models and applications in the same manner across all parts of your infrastructure. It gives you the portability, scalability, and flexibility to provision AI/ML environments on demand. It can also provide self-service capabilities to speed resource delivery while maintaining IT control. Finally, a consistent platform supplies a foundation for technology integrations from third-party vendors, open source communities, and any custom-developed tools you may use.

Hybrid cloud platform recommendations

Select a security-focused platform that supports hardware acceleration, a broad ecosystem of AI/ML and application development tools, and integrated DevOps and management capabilities. Open source platforms can provide more integration opportunities and flexibility.

Edge deployments

Edge deployments are decentralized environments with devices that collect data and perform functions on site, outside your core datacenter.

Edge computing can deliver insights and experiences at the exact location and moment they're needed. Sensors and devices often generate large volumes of data that can be used in AI/ML workflows for model training and runtime inferencing. This data can be difficult and costly to transmit to a central cloud in real time. For example, image recognition algorithms run more efficiently closer to the source of the data, removing the need to transfer large amounts of data for processing.

Edge best practices

Scalability, connectivity, and device management are critical for edge deployments. Look for solutions that can be managed with the same tools and processes as your datacenter and cloud infrastructure. Platforms that can handle disrupted communications and disconnected environments are crucial. Finally, solutions that support many device and hardware footprints offer more flexibility and customization.

Start with an open, flexible foundation

Red Hat provides a complete technology portfolio, proven expertise, and strategic partnerships to help you achieve your AI/ML goals. We deliver a foundation for building production-ready AI/ML environments in addition to services and training for rapid adoption.

Red Hat® OpenShift® is a unified, enterprise-ready application platform for cloud-native innovation. On-demand compute resources, support for hardware and graphics processing unit (GPU) acceleration, and consistency across on-site, public cloud, and edge environments provide the speed and flexibility that teams need to succeed. For example, you can create a self-service MLOps platform for data scientists, data engineers, and developers to rapidly build models, incorporate them into applications, and perform inferencing tasks. Collaboration features let teams create and share containerized modeling results with peers and developers in a consistent manner.

Red Hat OpenShift AI is a platform that help you train, serve, monitor and manage the life cycles of AI/ML models and applications. Included within this offering, Red Hat OpenShift AI gives data scientists and developers a powerful AI/ML platform for gathering insights and building intelligent applications. Teams can move from experiment to production in a collaborative, consistent environment that integrates key certified partner offerings from NVIDIA, Intel, Starburst, Anaconda, IBM, and Pachyderm, to name a few.

The **Red Hat Application Services** portfolio helps you create a unified environment for application development, delivery, integration, and automation. Data integration services let you build effective data pipelines while runtime services simplify application development. Process automation tools and services can access intelligent applications and ML models to automate business decisions.

Finally, Red Hat platform products—including **Red Hat Enterprise Linux®**, **Red Hat OpenStack® Platform**, and **Red Hat OpenShift Platform Plus**—provide a scalable, software-defined infrastructure.

Build through community

Red Hat actively participates in the **Kubeflow** and **Open Data Hub** open source communities. Open Data Hub is a community project that provides a blueprint for integrating common open source AI/ML tools into an OpenShift environment. Common data analytics and machine learning tools—like Ray, Ceph®, Apache Kafka, Kubeflow, TensorFlow, and Jupyter notebooks—are integrated into the reference architecture.

Gain flexibility with a certified partner ecosystem

Red Hat's certified partner ecosystem lets you integrate popular AI/ML, data analysis, management, storage, security, and development tools into this architecture. We work closely with partners to certify their software on our platforms for increased manageability, security, and support. Many partners also provide certified **Red Hat OpenShift operators** to simplify software life cycle management.

Choose your preferred products and technologies

Red Hat fosters a growing ecosystem of certified AI/ML partners, allowing you to incorporate popular products and technologies into your environment.

NVIDIA and Red Hat offer solutions that speed delivery of AI-powered intelligent applications across environments. **NVIDIA AI Enterprise with Red Hat OpenShift** provides a complete, optimized, cloud-native suite of AI and data analytics software. Red Hat Enterprise Linux, Red Hat OpenShift, and NVIDIA DGX systems deliver IT manageability for AI infrastructure. The **NVIDIA GPU Operator** automates the management of all NVIDIA software components needed to provision GPUs.

Starburst and Red Hat help you unlock insights across distributed data sources. **Starburst Enterprise** works with Red Hat OpenShift to rapidly analyze data across multiple platforms. The combination provides enterprise-grade automation, high availability, elasticity, and monitoring capabilities. With this solution, you can modernize your data; efficiently run extract, transform, and load (ETL) workloads; perform interactive data investigation; and inform business intelligence tools.

Intel and Red Hat work together to offer software-defined infrastructure and industry-standard platforms that improve datacenter agility and flexibility. Intel's distribution of the **OpenVINO toolkit** optimizes and converts DL models into high-performance inference engines that can automatically scale to thousands of nodes on Red Hat OpenShift. **Intel AI Analytics Toolkit powered by oneAPI** delivers a complete set of interoperable AI software tools for speeding and scaling ML workflows.

SAS and Red Hat collaborate to create open, hybrid cloud technologies and analytical capabilities that deliver business-level intelligence. **SAS Viya** on Red Hat OpenShift combines leading analytics, ML, and AI applications with a hybrid cloud platform to let you build once and deploy anywhere. Consistent management across infrastructures unites teams and promotes collaboration. This unified platform allows you to develop and deploy models using your preferred interfaces, languages, and infrastructures.

See success in action

With Red Hat Consulting, **Banco Galicia** built an AI-based intelligent natural language processing (NLP) solution on Red Hat OpenShift, Red Hat Integration, and Red Hat's single sign-on.

Cut corporate customer onboarding times from

20 days to minutes

while achieving 90% data analysis accuracy.

Read the **success story**.

U.S. Department of Veterans Affairs

The **U.S. Department of Veterans Affairs' Team Guidehouse** deployed Red Hat OpenShift and Red Hat OpenShift AI to use machine learning techniques in a prototype solution to prevent Veteran suicides.

Phase 2 winner

in the Mission Daybreak challenge

Read the **blog post**.

Nippon Telegraph and Telephone East Corporation (NTT East) built an edge computing data analysis service using Red Hat OpenShift.

"[...] Red Hat OpenShift has made it possible to stably develop and operate innovative video AI services through collaboration with AI developers."

Masashi Toyama
Manager, Server Infrastructure Technology Cloud Server Engineering Department, Advanced Promotion Division, Network Business Headquarters, NTT East

Read the **success story**.

Boston University uses Red Hat OpenShift AI as its main classroom platform for computer science and computer engineering systems courses.

"This effort is providing my students with a rich, full-fledged Linux experience that hides no details and yet can be easily accessed and integrated into my teaching materials and methodologies."

Jonathan Appavoo
Associate Professor at Boston University

Read the **blog post**.

Ready to begin your journey to AI/ML?

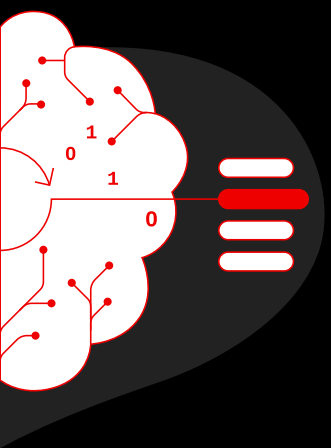
AI/ML and MLOps are transforming nearly every aspect of business.

Red Hat can help you build a production-ready AI/ML environment that accelerates development and delivery of intelligent applications to support your business goals.

Red Hat
OpenShift AI

Learn how Red Hat OpenShift can accelerate AI/ML workflows and delivery of AI-powered intelligent applications: **red.ht/openshift_ai**

Try it for free



Get started right away with Red Hat Consulting

Work with Red Hat experts to jump-start your AI/ML projects. Red Hat offers consulting and training services to help your organization adopt AI/ML faster.

- Learn about AI/ML services: **red.ht/aiml-consulting**
- Schedule a complimentary discovery session: **redhat.com/consulting**

479615_0823_KVM

Red Hat