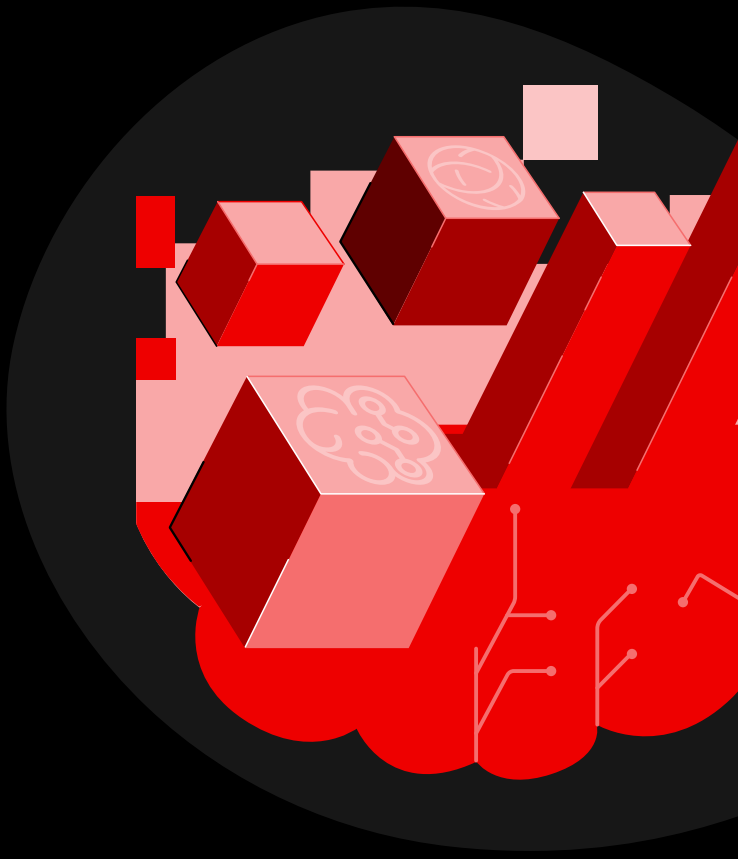
Red Hat

Top considerations for

building a foundation for generative AI

Contents



1 Explore new possibilities for business innovation

2 Considerations for building a foundation for generative AI

- 2.1 Development toolsets
- 2.2 Model tuning
- 2.3 Model serving
- 2.4 Life cycle management
- 2.5 Model monitoring
- 2.6 Partner ecosystems
- 2.7 Platform expertise

3 Innovate rapidly with a flexible, open foundation

4 Ready to get started with generative AI?

Explore new possibilities for business innovation

Generative artificial intelligence (AI) is a powerful tool for organizations that want to create innovative products, optimize processes, and gain competitive advantages in rapidly changing markets. Based on advancements in deep learning and neural networks, it goes beyond predictive AI capabilities by not only processing data, but generating new, original content. Generative AI is reshaping human-machine collaboration, inspiring new approaches to problem solving, and delivering significant business gains across industries.

Worldwide, organizations are building new, innovative applications using generative AI technologies. In fact, 39% are currently investing in generative AI technologies, while another 37% are exploring potential use cases.[1] Here are a few of the many use cases for generative AI today:

- **Generate forecasts for complex scenarios.** Generative AI can analyze historical data, identify patterns, and develop accurate forecasts to aid in strategic planning and risk management.
- **Develop personalized marketing.** By analyzing data to understand customer preferences and behaviors, generative AI can create personalized marketing materials—including emails, advertisements, and promotions—that maximize engagement and conversion rates.
- **Automate and personalize customer service.** As the foundation for intelligent chatbots and virtual assistants, generative AI can automatically respond to customer inquiries and interactions, providing personalized, efficient customer service.

1 IDC Web Conference Proceeding. "Unlocking Business Success with Generative AI." Document #US50789223. June 2023.

Generative AI brings new concerns

Although the benefits and drawbacks of generative AI are still emerging, many organizations want to invest in these new technologies now. Understanding the issues related to generative AI can help businesses establish clear ethics guidelines and development frameworks, comply with government and industry regulations, and detect and correct potential issues.

- **Data privacy.** Privacy concerns arise when generative AI models are trained using sensitive or personal data, leading to questions about the protection of individuals' privacy.
- **Data ownership.** Using proprietary models—or models pre-trained using proprietary data—introduces data ownership issues that may lead to litigation.
- **Bias and fairness.** Responses from generative AI tools have been shown to reflect human biases, including harmful stereotypes and hate speech.
- **Ethical use.** Generative AI models can create synthetic content and deep fakes used for malicious activities like privacy infringement and misinformation campaigns.
- **Explainability and interpretability.** A lack of transparency in generative AI tools makes it difficult to interpret, understand, and explain model outputs, resulting in a lack of accountability for incorrect or made-up information.
- **Unintended consequences.** The autonomous nature of generative AI can lead to unintended consequences that can cause real harm to people and organizations.
- **Regulatory challenges.** Rapid advancements in generative AI technologies may outpace regulatory frameworks, making it difficult to create and enforce guidelines that ensure responsible and ethical use.
- **Energy consumption.** AI model training is computationally intensive with large energy demands, raising concerns about environmental impacts and sustainability.

This e-book reviews key considerations for building a trusted infrastructure foundation to support generative AI initiatives.

Prepare for generative AI

In "Unlocking Business Success with Generative AI," IDC recommends these actions to prepare your organization for generative AI initiatives.[2]

- **Create an environment of agile experimentation** for prioritized use cases that meet your business needs.
- **Develop corporate policies** for responsible use that discourage malicious behaviors.
- **Assess workforce impacts** of generative AI and engage in proactive change management.
- **Partner with trusted technology vendors** and service providers for your AI infrastructure.
- **Secure the right engineering skills** through hiring, training, or professional services support.

2 IDC Web Conference Proceeding. "Unlocking Business Success with Generative AI." Document #US50789223. June 2023.

Considerations for building a foundation for generative AI

The technology foundation you choose for your generative AI initiatives can greatly impact your ease of adoption and overall success. This chapter discusses key considerations for your generative AI foundation.

Consideration 1: Build with a proven toolset

Developing applications based on generative AI models can be a complex task. The right toolset—with languages, frameworks, and runtimes based on open source projects and commercial solutions—can speed model tuning and simplify application development and deployment.

Choose an AI foundation that delivers your preferred toolsets for developing innovative AI solutions quickly and efficiently. Support for exploratory data science, training, and tuning through interactive interfaces can simplify collaboration. Preintegrated toolsets and self-service capabilities help you streamline IT operations while maintaining portability and consistency across environments.

Consideration 2: Rapidly fine tune models

Because training generative AI models is an expensive, time-consuming process, most organizations build AI solutions using foundation models that are pretrained on general-purpose data. Data scientists then use diverse, domain-specific data to adjust foundation models to perform specialized tasks. Fine tuning, however, can still be computationally intensive, requiring powerful processors and distributed hybrid cloud infrastructure.

Look for AI platforms with distributed workload management and orchestration capabilities that deploy training runs—of any model size, data volume, or duration—across hybrid cloud environments. Options for fine tuning foundation models in on-site datacenters simplify compliance with technical and regulatory requirements for restricted models. Batch training features let you preempt fine tuning workloads and make it easier to share and manage resources.

Alternatives to fine tuning models

Researchers are investigating ways to tune foundation models faster and more efficiently. **Retrieval-augmented generation (RAG)** is an AI framework for retrieving facts from external sources—like internal databases, corporate intranets, or the internet—to provide generative AI models with the most accurate, up-to-date information.

In **prompt tuning**, AI models receive cues or front-end prompts—including extra words or AI-generated numbers—that guide models toward a desired decision, allowing organizations with limited data to tailor a foundation model to a narrow task.

Consideration 3: Serve models efficiently

Delivering exceptional user experiences from generative AI solutions can be challenging for IT operations teams. Variable application demand calls for scalable infrastructure and automated management. Efficient model deployment requires capabilities to monitor performance and quickly revert to previous versions. And because AI solutions process vast amounts of data, enforcement of strict security standards across environments is critical.

Consider platforms that can deploy and scale generative AI models and applications across hybrid clouds—including on-site infrastructure, public cloud resources, and edge devices. Options to serve generative AI models from on-site or isolated environments ensure that proprietary data is not used to retrain publicly available models. And support for canary rollouts and explainability tools helps increase the consistency and reliability of model responses.

Consideration 4: Automate life cycle management

Continuous integration/continuous delivery (CI/CD) pipelines can automatically deploy and manage generative AI solutions. By retraining and updating models and applications through rapid, incremental changes, you can speed development and increase model performance. However, AI pipelines are more complex than standard CI/CD workflows since they often include additional stages like data extraction, training, fine tuning, validation, and retraining.

Choose a foundation that lets you create and integrate AI pipelines—based on CI/CD tools like Tekton and Jenkins—into existing DevOps workflows to quickly and efficiently develop, train, monitor, and retrain generative AI models. **GitOps** continuous delivery tools like ArgoCD let you define and automate complex AI solution deployments as code for consistent model and application delivery.

Containers for generative AI

Container and **Kubernetes** technologies provide agile deployment, management, and scalability to accelerate cloud-native development of generative AI solutions. Provision environments on demand across on-site datacenters, public clouds, and edge devices. Automatically create, deploy, scale, and manage container instances on physical and virtual infrastructure. And integrate components and data stores from a robust ecosystem of open source and commercial suppliers into generative AI solutions. Learn more about the **benefits of containers for AI**.

Consideration 5: Monitor models consistently

Generative AI models can have real, substantial impacts on people and businesses. By tracking model behavior, you can analyze decisions and justifications, identify poor performance, and report problematic behaviors immediately. Effective model governance based on this information helps ensure that models respond with unbiased, fair, and correct information in production environments.

Explore AI foundations with centralized monitoring capabilities that provide bias and data drift metrics, anomaly detection, and per-point explainability to help you investigate, maintain, and correct generative AI models. Continuous, automatic monitoring in production environments increases compliance with corporate model governance standards. And user-friendly tool interfaces and human-readable, non-technical reports encourage responsible model use and maintenance.

Key generative AI model concepts

- **Bias** is the presence of patterns in model behavior that impact the fairness, inclusivity, and ethics of generated outputs, including favoring certain groups or producing responses that align with stereotypes.
- **Data drift** occurs when the statistical properties of the training data change over time, resulting in a decrease in model performance and the generation of less accurate or relevant responses.
- **Anomaly detection** is the process of identifying and reporting model behaviors that are uncommon or divergent from examples seen during training.
- **Per-point explainability** is the ability to understand why models generate specific outputs, providing visibility for applications where transparency is critical.

Consideration 6: Take advantage of partner ecosystems

Generative AI solutions require multiple, integrated components to successfully deliver innovative user experiences. With the right combination of technologies from a collaborative ecosystem of trusted vendors, you can speed application development, address bias and data drift challenges, and ensure consistent, reliable performance for your entire solution.

Look for platform vendors with extensive, certified partner ecosystems that offer complete solutions for developing and deploying generative AI models and applications. A large selection of components—from data integration and preparation to model training and serving—help you develop and deploy AI solutions faster and more efficiently. And by choosing certified solutions with proven interoperability, you can reduce IT support requests and increase productivity.

Consideration 7: Work with platform experts

Effectively deploying and managing generative AI solutions requires specialized knowledge and experience. Scalability requirements, reliability concerns, and integration with existing systems can complicate production deployments. Inefficient use of compute resources can result in unnecessary costs. And non-compliance with security standards, privacy policies, and AI regulatory frameworks can lead to unintended consequences.

Choose vendors with teams of experts who provide comprehensive support and guidance for building generative AI solutions. For example, dedicated engineers may support the entire platform with the tools, resources, and knowledge to speed your AI projects. Expert consultants can solve deployment challenges, optimize infrastructure efficiency, and ensure interoperability across your AI solution. And professional training services can help you gain the knowledge and expertise to get started on new generative AI projects faster.

Generative AI requires collaboration

Building a team with a range of capabilities is key for successful generative AI projects.[3]

- **Business leaders** represent the people who use or are impacted by the solution.
- **AI specialists** tune, maintain, and update generative AI models.
- **Data scientists** preprocess and provide correct, unbiased training data for models.
- **Ethics and compliance officers** ensure that generative AI initiatives comply with regulations.
- **IT operations specialists** integrate solutions with existing infrastructure and enforce security policies.

3 *Kearney. "Standing up tiger teams to tackle generative AI complexity," November 2023.*

Innovate rapidly with a flexible, open foundation

Red Hat provides a complete technology portfolio, proven expertise, and strategic partnerships to help you achieve your generative AI goals. We deliver a foundation for developing and deploying generative AI models and applications, as well as 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 acceleration, and consistency across on-site, public cloud, and edge environments provide the speed and flexibility teams need to succeed. With Red Hat OpenShift, you can create a self-service platform for data scientists, data engineers, and developers to rapidly develop intelligent applications. Collaboration features let teams create and share containerized modeling results with peers and developers in a consistent manner.

Red Hat OpenShift AI builds on Red Hat OpenShift to provide a comprehensive platform for building, training, fine tuning, deploying, and monitoring models and applications, while meeting the workload and performance demands of modern generative AI solutions. Teams can quickly move from experiment to production in a collaborative, consistent environment that integrates key certified offerings from partners like NVIDIA, Intel, Starburst, Anaconda, IBM, Run:ai, and Pachyderm. Together with our technology ecosystem, Red Hat OpenShift AI provides components and capabilities that speed development and deployment of innovative, generative AI solutions across hybrid clouds.

IBM watsonx.ai AI studio provides a selection of models and deployment options with the generative AI capabilities that your intelligent applications need. Deploy models—including open source, third-party, and IBM-developed foundation models—wherever your workload resides to increase the performance and efficiency of your AI solutions. And with **IBM-developed foundation models** trained on enterprise-relevant data, your generative AI solutions understand the nuances of your business domain to give you a competitive edge.

Red Hat Ansible® Lightspeed with IBM watsonx Code Assistant is a generative AI service engineered to help teams create, adopt, and maintain automation content more efficiently. Connected to IBM watsonx Code Assistant, Red Hat Ansible Lightspeed helps you turn your automation ideas into Ansible code with natural language prompts. With it, you can enhance productivity and make automation more accessible across your organization.

Ready to get started with generative AI?

Generative AI is a powerful tool for creating original content and changing the way we interact with applications and technology.

Through technology, expertise, and partnerships, Red Hat provides the common foundations for your teams to build and deploy AI applications and ML models with transparency and control. In fact, we even use our own AI tools and platforms to improve the utility of other open source software. And our partner integrations connect you to an ecosystem of trusted AI tools built to work with open source platforms like Red Hat OpenShift AI.

Learn more and try Red Hat OpenShift AI for free.

Get started faster 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 in less time.

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

1057715_0324_KVM

Red Hat