AI for the public sector

Deploy anywhere, and apply discipline with MLOps

Why Red Hat?

When your mission is critical, your IT needs to keep up.

The U.S. government demands security, stability, and reliability—core principles of Red Hat and open source. With 100% U.S. executive departments relying on Red Hat, we provide trusted cloud, virtualization, storage, and platform solutions that empower federal, state, local, and academic programs with flexibility and collaboration.[2] Supported by global expertise in training and consulting, we help agencies use open source to propel innovation within a broader, collaborative community.

The mission value of artificial intelligence

Government organizations and higher education institutions are accelerating their adoption of artificial intelligence (AI) to increase productivity, reduce manual errors, and make timely, well-informed decisions. Innovations are occurring in both predictive and generative AI (gen AI). In a FedScoop survey of government executives and IT leaders, 84% of respondents said that gen AI was critical or important to operations.[1]

Table 1. Public sector use cases for predictive and gen AI

	Predictive AI	Generative AI
Mission value	• Mitigate risk by predicting future events and trends from historical data.	• Produce, translate, or transform original content by learning from large quantities of data.
Sample public sector use cases	• Fraud detection (e.g., tax, unemployment insurance). • Predictive maintenance (e.g., critical infrastructure, fleets, shipyards, and depots). • Forecasting (e.g., disease outbreaks, and drug trafficking trends, environmental impacts on mission). • Risk assessment (e.g., disease, cybersecurity, adversary movement or disposition).	• Knowledge retrieval and semantic search. • Productivity (e.g., chatbots, voicebots, and content summarization). • Coding (e.g., refactoring existing code, translation, and modernization). • Process optimization (e.g., generation of email, contracts, proposals, document formatting, bail hearing adjudication, and healthcare protocols). • Unmanned vehicle path planning.

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

1 *Fedscoop, underwritten by Microsoft. "Gauging the impact of generative ai on government." FedScoop, 2023.*

2 *Red Hat client data, Sept. 2024.*

	Predictive AI	Generative AI
Data used for training	• Usually internal data sources.	• Internal data sources, internet, social media, partner agency data, and more. Models can learn continually from websites and vector databases by using the retrieval-augmented generation (RAG) technique.

To adopt AI at scale, public sector organizations need an AI framework that:

- **Scales AI inference.** IT operations teams need the flexibility to deploy models in the optimal location for the mission: public cloud, on-premise, or edge.
- **Enforces use of standardized processes for AI application development.** Today, many data science teams use independently designed processes, making it more difficult to enforce agency security and quality standards.
- **Automates processes such as security testing.** Automation helps teams scale to build and support a growing number of machine learning (ML) models and AI-powered applications.
- **Provides flexibility and choice.** Teams want a choice of tools, languages, and run-times.
- **Builds gen AI services within a controlled environment.** Today's widely available large language model (LLM) services introduce a high risk of data leakage and model bias, and lack attestation and data provenance.

Complete lifecycle capabilities for modeling

- Model development tooling based on JupyterLab, TensorFlow, PyTorch, CUDA, KubeFlow notebook controller
- Model serving with KServe, OpenVINO, TGIS, vLLM
- Model monitoring, including utilization and response-time metrics
- Data and model pipelines based on KubeFlow pipelines
- Distributed workloads for faster, more efficient data processing and model training
- Model inference at the edge
- Model registry

Red Hat's approach

Any model, any accelerator, any cloud

Inference, the process of running a model to generate responses, is what provides real business value for AI users. Red Hat AI Inference Server provides consistent, fast, and cost-effective inference at scale. It allows you to run any generative AI model on any hardware accelerator (NVIDIA, Intel, AMD) and in any environment (datacenter, cloud, edge)—providing flexibility and choice to meet mission requirements.

AI Inference Server is paired with an optimized model repository which offers a collection of validated and optimized models, ensuring rapid deployment with benchmarked performance. It also includes a model compression library, LLM Compressor, to optimize your models using advanced quantization and pruning techniques to improve inference speeds while maintaining prediction accuracy. Together, these components create fast, accurate, and cost-effective inferencing across a wide range of applications.

Apply DevOps-style discipline to AI model building, training, and tuning

An add-on to Red Hat® OpenShift®, Red Hat OpenShift AI is available as a self-managed cloud service. OpenShift AI brings together data scientists and developers, with oversight from IT, to develop, train, and fine-tune models, deliver AI-enabled applications, and bring models from experiments to production in less time. Data science teams have self-service access to collaboration

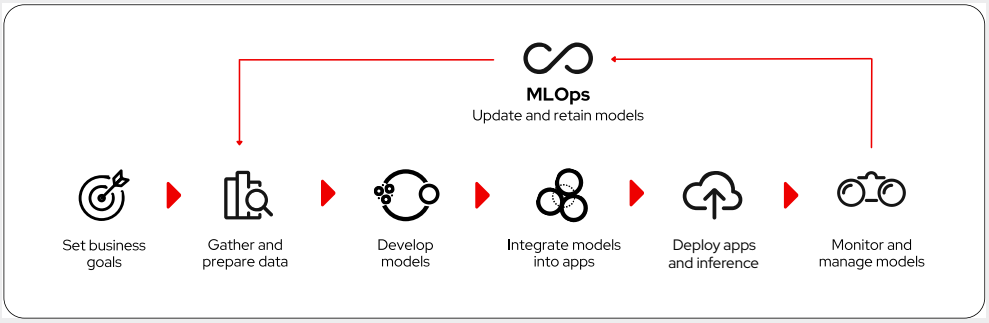

Figure 1. MLOps standardizes the approach to building and updating models

workflows and graphics processing units (GPUs), saving time for themselves and their IT teams.

OpenShift AI provides a framework for ML operations (MLOps), a set of workflow practices for efficiently deploying and maintaining ML models. Inspired by DevOps principles, MLOps helps government teams integrate ML models into their software development processes, providing continuous monitoring, retraining, and deployment to maintain model accuracy as new data is gathered. The goal of MLOps is to create ML models that are accurate, reliable (produce replicable results), and trustworthy.

Public sector data scientists can use OpenShift AI to:

- Build and tune models on any infrastructure–cloud, on-premise, or edge.
- Retrieve and harmonize data from multiple internal or external sources. To save time, consider starting with a foundation model. OpenShift AI includes IBM open source Granite family LLMs, for example, or you can bring your own LLM, such as Mistral, Llama, and others. Tune the foundation model by using RAG to search for knowledge from a specific domain, such as healthcare, law enforcement, the intelligence community, or air traffic management.
- Deploy the model anywhere in a hybrid cloud environment. Models built on OpenShift AI are in a container-ready format that can be deployed consistently on any hardware, including air-gapped and disconnected environments.
- Fine-tune the model based on results, a continuous process comparable to iterating application code.

OpenShift AI is a companion to Red Hat OpenShift. Data scientists use OpenShift AI to build, tune, and serve models, while application developers use OpenShift to code, perform quality tests, deploy, operate, and monitor AI-enabled applications.

Red Hat OpenShift AI ecosystem integrations

- Modeling and visualization: Anaconda, AI Tools from Intel
- Data engineering and versioning: Starburst and Pachyderm
- Data ingestion and storage: Red Hat Application Foundations, which includes streams for Apache Kafka and Red Hat 3scale API Management, and Amazon Simple Storage Service (S3)
- Vector database for RAG: Elastic

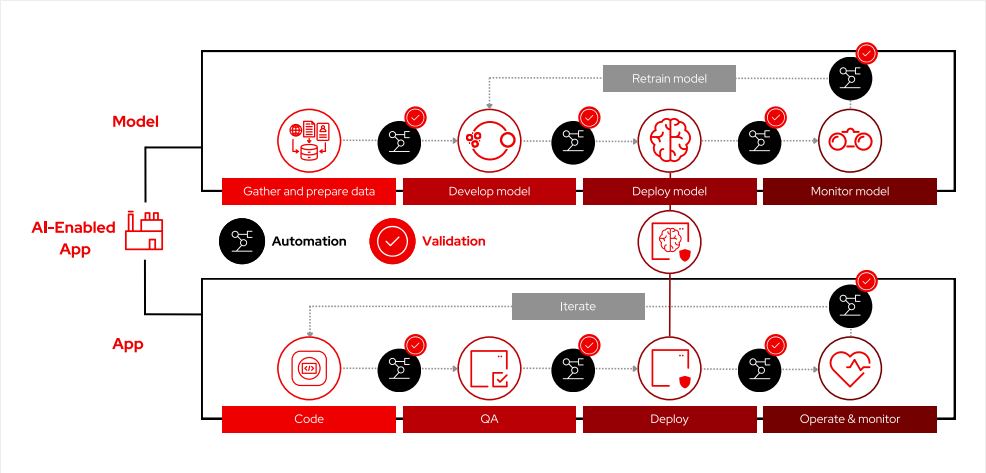

Figure 2. Create a scalable MLOps factory for ML models and AI-enabled applications

Scale AI while mitigating its risks

By applying discipline to ML model building and AI application development, OpenShift AI supports public sector missions by helping teams:

Remain flexible. Unlike AI/ML suites from cloud operators, OpenShift AI gives teams a choice of tools and infrastructure (see sidebar, "Red Hat OpenShift AI ecosystem integrations"). Build, tune, and deploy models in any location: on-premise, in a cloud, and at the edge. With Red Hat's bring your own model approach, data science teams can use any model, either developed internally or imported from a trusted third party.

Make ML experiments replicable. The MLOps framework tracks and manages changes to the code and configuration files associated with ML-enabled applications.

Bring AI-enabled applications to production in less time. Automated workflows help public sector teams start using predictive and gen AI models sooner, and to continually improve their accuracy.

Optimize and manage resources. Scale to meet workload demands of gen AI and predictive models. Focus on maintaining models–not infrastructure. Share resources, projects, and models across environments. Increase operational consistency by streamlining the process of moving models from experiments to production.

Meet governance and compliance regulations. OpenShift AI takes advantage of the security capabilities built into Red Hat OpenShift to help meet public sector compliance and regulatory requirements. Organizations with strong data privacy requirements can build and serve models on-premise or at the edge, even in disconnected environments.

In action: Department of Veterans Affairs

The electronic health record (EHR) system for the Department of Veterans Affairs contains a wealth of information that can be used to identify people at risk for conditions such as mental health issues or chronic kidney disease. Red Hat teamed with global consulting services provider Guidehouse and Philip Held, Ph.D. of Rush University Medical Center, to develop an AI/ML-based means of identifying veterans at risk for suicide ideation. Data scientists and developers used OpenShift AI managed cloud service to develop, train, and test ML models before deploying them.

Take the next step

Start a 60-day Red Hat OpenShift AI trial.

Begin experimenting and training models at a smaller scale using:

- **Red Hat Enterprise Linux® AI.** Develop, test, and deploy gen AI models using this foundation-model platform, which ships with the open source-licensed Granite family of large language models.

Talk to a Red Hatter.

Learn more about how Red Hat can help your agency or institution achieve its mission.

redhat.com
#2303831_0625