CENTRE OF EXCELLENCE AI

Red Hat OpenShift AI with IBM watsonx.governance on AWS

Ready-to-run platform for MLOps, with AI Governance for trustworthy models

Tech Data
A TD SYNNEX Company

Solution Overview

Use Red Hat OpenShift AI to define and implement all aspects of the MLOps lifecycle for both Predictive and Foundation Models. Use IBM watsonx.governance to enable trustworthy and reliable models that can be implemented in production with confidence.

Technology Partners

watsonx Red Hat aws

Key Benefits

- **Red Hat OpenShift AI:** Enables data acquisition and preparation, model training and fine-tuning, model serving and model monitoring, and hardware acceleration. With an open ecosystem of hardware and software partners, OpenShift AI delivers the flexibility you need for your specific use cases.

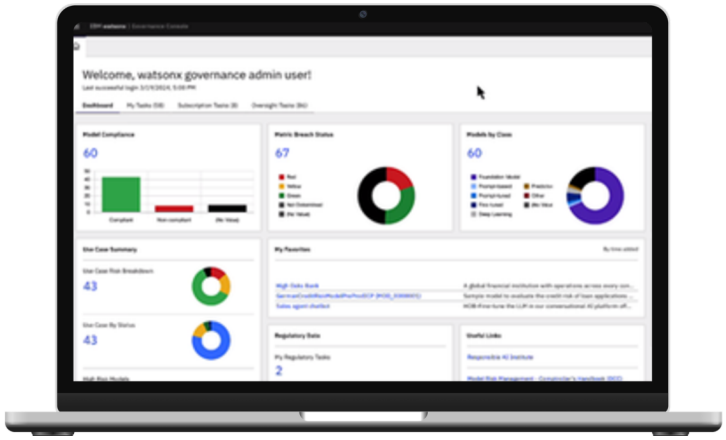
- **IBM watsonx.governance:** Ensures correct model behaviour, both as part of testing and in production.
  Enable trustworthy and reliable models that can be implemented in production with confidence. Automate and accelerate responsible AI workflows to help save time, reduce costs and comply with regulations.

Tech Data Centre of Excellence

Red Hat OpenShift AI and IBM watsonx.governance, available for Demos, PoCs and Workshops.

USE CASES

- MLOps – full lifecycle management
- Develop, test and deploy Foundational and Predictive AI models
- Model risk governance and operational risk management
- Model monitoring and evaluation for bias, drift, accuracy and explainability
- Automated deployment to AWS



CENTRE OF EXCELLENCE AI

Elastic RAG with Red Hat OpenShift AI on AWS

Vector Database implemented on an enterprise-class platform for MLOps

Solution Overview

Elasticsearch Relevance Engine (ESRE) is a comprehensive suite of developer tools for building generative AI and RAG applications. ESRE incorporates a vector database that stores embeddings for text, image, and video data. ESRE is installed on Red Hat OpenShift AI, providing the complete platform for MLOps - integrating with AI Models along with other required components to develop, test and serve the production AI solution.

Technology Partners

elastic Red Hat aws

Key Benefits

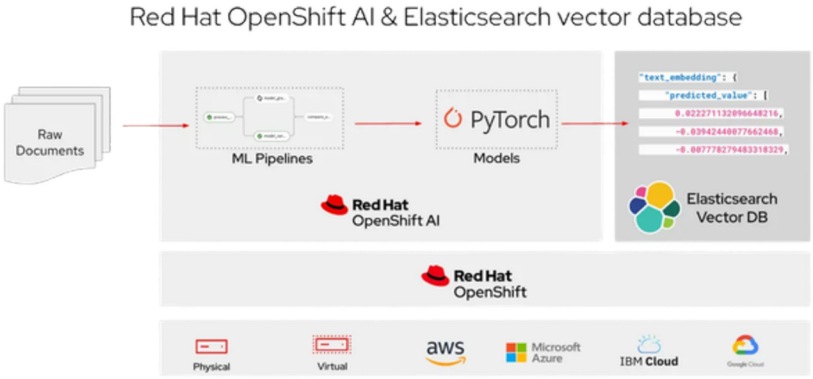
- Implement vector search and semantic search, including k-nearest neighbours (kNN) and approximate nearest neighbour (ANN) search, along with both built-in and third-party natural language processing (NLP). ESRE's native hybrid search can effectively combine results containing text, vectors, and geospatial data, with filtering, aggregations, and document-level security.
- Build on a platform with full support for the complete MLOps Lifecycle – model development, model serving, model monitoring and lifecycle management. Integrate all of the necessary components to build a complete Retrieval Augmented Generation Solution, with ESRE providing the vector database.

Tech Data Centre of Excellence

Red Hat OpenShift AI and Elasticsearch Vector Database, available for Demos, PoCs and Workshops.

USE CASES

- Vector database for RAG
- Integrate with Foundation Models
- MLOps – provide full lifecycle management
- All the tooling needed to develop, test and deploy Foundational and Predictive AI models
- Automated deployment to AWS



CENTRE OF EXCELLENCE AI

SAS Viya with Red Hat OpenShift on AWS

Unlock Advanced Analytics with Scalable Cloud Solutions

Solution Overview

Enhance your SAS Viya deployments with Red Hat OpenShift GitOps, automating the lifecycle management on the OpenShift Container Platform. Enjoy effortless deployment, updates, and management, with improved efficiency, scalability, and collaboration. Say goodbye to manual intervention and streamline your process for consistent, reliable results, allowing you to focus on maximising your SAS analytics.

Technology Partners

Key Benefits

- **Achieve** automated and streamlined SAS Viya deployment and management with Red Hat OpenShift GitOps.
- **Optimise** SAS administrators' productivity by using the SAS Deployment Operator for software lifecycle management and maintain consistent and controlled SAS Viya environments across the cluster with version-controlled manifests.
- **Scalability** Leverage the power of AWS to scale SAS Viya deployments seamlessly, ensuring performance and efficiency.

Tech Data Centre of Excellence

Red Hat OpenShift with SAS Viya on AWS, available for Demos, PoCs and Workshops.

USE CASES

- Automated Lifecycle Management
- Scalable Data Analytics
- Integration of workflows and enhanced productivity
- Automated deployment to AWS

CENTRE OF EXCELLENCE AI • AUTOMATION

Red Hat Ansible Lightspeed with IBM watsonx Code Assistant on AWS

AI-assisted playbook development and Ansible content generation

Solution Overview

Red Hat Ansible Lightspeed with IBM watsonx Code Assistant integrates Ansible and AI capabilities to translate plain-text prompts and generate Ansible Playbook content. It is built to accelerate Ansible development and provide high-quality contextual recommendations.

Technology Partners

watsonx Code Assistant

Red Hat

aws

Key Benefits

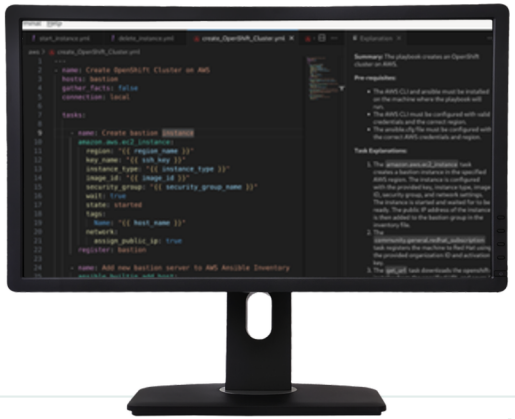
- Infuses the Ansible Automation Platform with new generative AI capabilities to enhance workflow orchestration.
- Purpose built for Ansible developers to accelerate content development.
- Content source matching and transparency.
- Meets developers where they are by integrating into the Ansible Visual Studio Code extension.
- Address skills gaps and increase developer productivity.

Tech Data Centre of Excellence

Red Hat Ansible Lightspeed with IBM watsonx Code Assistant, available for Demos, PoCs and Workshops.

USE CASES

- Build and manage resources across the IT landscape with AI-assisted Ansible Playbook generation
- Quickly and efficiently build out Ansible Playbooks with multi-task code recommendations
- Automated deployment to AWS
- Integration with AWS resources



CENTRE OF EXCELLENCE AI

Red Hat Enterprise Linux AI on AWS

Integrated Gen AI Linux server appliance for model customisation and inference

Solution Overview

Red Hat Enterprise Linux AI (RHEL AI) delivers an integrated platform for building, customising, and deploying generative AI models on enterprise-grade Linux. It combines the stability and security of RHEL with AI-focused tooling such as InstructLab for fine-tuning and vLLM for efficient inference. RHEL AI enables organisations to accelerate AI adoption while maintaining compliance, scalability, and operational consistency across on-premises and cloud deployments.

Technology Partners

Key Benefits

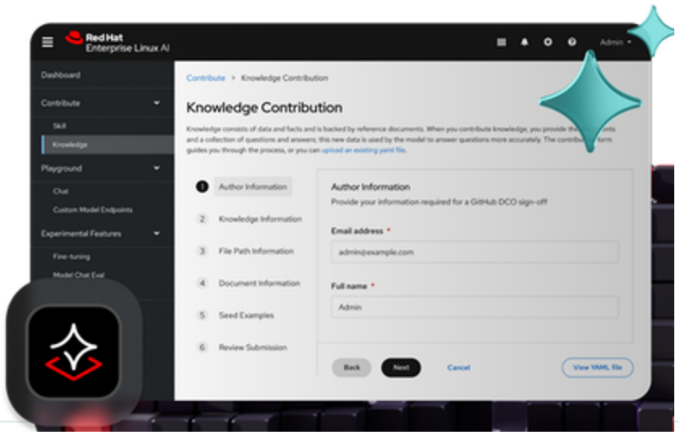
- **Enterprise-grade AI platform:** Combines the stability, security, and compliance of RHEL with AI- focused tooling for production-ready deployments.
- **Efficient inference at scale:** Leverage vLLM for high-performance, low-latency model inference across hybrid environments.
- **Customisable AI workflows:** Use InstructLab to fine-tune and adapt models to your organisation's unique requirements.
- **Hybrid cloud flexibility:** Deploy and manage AI workloads consistently across on-premises and cloud environments, including AWS.

Tech Data Centre of Excellence

RHEL AI on AWS, available for Demos, PoCs and Workshops.

USE CASES

- Efficient Inference with vLLM
- Testing with validated, quantised models
- Fine Tuning via InstructLab
- Document processing and parsing for Retrieval Augmented Generation (RAG) applications



CENTRE OF EXCELLENCE AI

Red Hat Inference Server on AWS

Generative AI model inference on RHEL / Linux or OpenShift / Kubernetes

Solution Overview

Red Hat Inference Server provides a secure, scalable platform for deploying and serving generative AI models in production. Built on RHEL or OpenShift/Kubernetes, it enables organisations to run inference workloads efficiently across hybrid environments. By leveraging AWS infrastructure, the solution delivers high-performance model serving with enterprise-grade security, operational consistency, and integration with existing cloud-native workflows.

Technology Partners

Red Hat

aws

Key Benefits

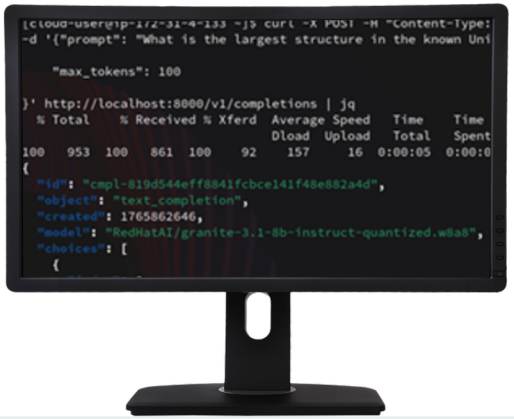
- **High-Performance Model Serving:** Deliver low-latency inference for generative AI models using optimised infrastructure on AWS.
- **Enterprise-Grade Security & Compliance:** Ensure secure deployment and governance across hybrid environments with RHEL or OpenShift foundations.
- **Operational Consistency & Scalability:** Simplify AI workload management with containerised deployments and automated scaling on AWS.
- **Integration with Cloud-Native Workflows:** Seamlessly connect to existing CI/CD pipelines, monitoring tools, and enterprise systems.

Tech Data Centre of Excellence

Red Hat Inference Server on AWS, available for Demos, PoCs and Workshops.

USE CASES

- Efficient Inference with vLLM
- Testing with validated, quantised models
- Performance evaluation with different models and hardware
- Integration with Retrieval Augmented Generation (RAG) applications



CENTRE OF EXCELLENCE AI

Elastic RAG with watsonx Assistant on Red Hat OpenShift AI on AWS

Empower scalable intelligence with Elastic RAG and watsonx Assistant on OpenShift AI

Solution Overview

Elastic Retrieval-Augmented Generation (RAG) with watsonx Assistant on Red Hat OpenShift AI on AWS combines Elasticsearch's powerful hybrid search capabilities with IBM's context-aware AI assistant. Deployed on OpenShift and AWS, it enables scalable and efficient generative AI workflows like semantic search and content summarisation.

Technology Partners

elastic watsonx Red Hat

Key Benefits

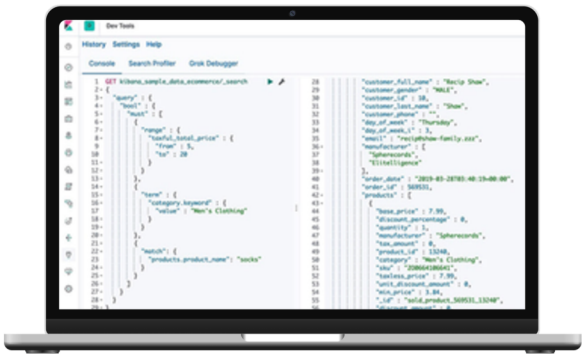
- **Precision and Relevance:** Elasticsearch Relevance Engine (ESRE) enhances watsonx Assistant's capabilities by combining advanced hybrid search with AI, delivering highly accurate and context-aware responses.
- **Scalability and Efficiency:** With Red Hat OpenShift AI and AWS, the solution ensures seamless scalability and optimised workflows for fast and reliable performance.

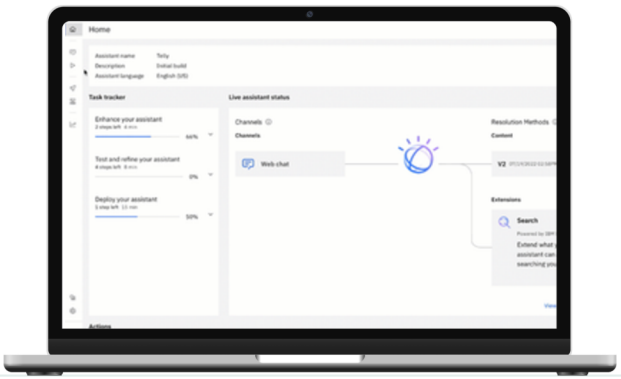
Tech Data Centre of Excellence

Red Hat OpenShift AI, Elasticsearch Relevance Engine with IBM watsonx Assistant, available for Demos, PoCs and Workshops.

USE CASES

- Vector database for RAG
- Integration with watsonx Assistant
- All the tooling needed to develop, test and deploy AI models
- Automated deployment to AWS





CENTRE OF EXCELLENCE AI

Red Hat OpenShift AI integrating with AWS Bedrock using Model Context Protocol

Providing Agentic AI capability for OpenShift Container workloads using Model Context Protocol (MCP)

Solution Overview

Integrating OpenShift AI with AWS Bedrock via an MCP server creates a streamlined way for OpenShift workloads to access Bedrock's foundation models. The MCP server runs inside OpenShift and provides a standardised interface that handles Bedrock authentication, model calls, and response formatting. This lets notebooks, pipelines, and applications in OpenShift use Bedrock models without embedding AWS-specific logic. The result is a clean, multi-cloud architecture where OpenShift AI remains the central operational platform while Bedrock supplies scalable, managed foundation models.

Key Benefits

- **Unified multi-cloud AI access:** OpenShift workloads can use AWS Bedrock models without needing AWS-specific code or credentials embedded in each application.
- **Stronger governance and security:** The MCP server centralises authentication, auditing, and traffic control, keeping Bedrock access aligned with OpenShift AI's enterprise policies.
- **Operational simplicity and scalability:** Teams keep OpenShift AI as their primary platform for orchestration and lifecycle management while seamlessly tapping into Bedrock's managed foundation models.

Technology Partners

Red Hat

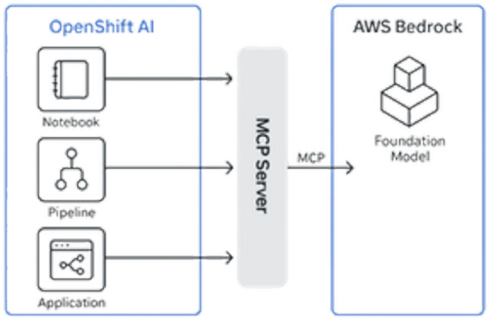
aws

Tech Data Centre of Excellence

Red Hat OpenShift AI with MCP Server and integration with AWS Bedrock, available for Demos, PoCs and Workshops.

USE CASES

- Agentic AI via MCP, integrating with AWS Bedrock model APIs and other models and APIs
- ModelOps – full lifecycle management with self-service for developers
- GPU-as-a-Service with policy and governance
- Model monitoring and evaluation for bias, drift, accuracy and explainability



CENTRE OF EXCELLENCE AUTOMATION

IBM Turbonomic with Red Hat Ansible Automation Platform on AWS

Enterprise-class Application Resource Management combined with Enterprise-wide Automation

Solution Overview

IBM Turbonomic provides the recommendations for appropriate resource allocation, and Ansible Automation Platform implements those recommendations, with integration to the wider IT landscape and the processes that govern them.

Technology Partners

Red Hat IBM Turbonomic aws

Key Benefits

- **Right-size resource allocation:** IBM Turbonomic provides comprehensive Application Resource Management in order to right-size resource allocation whilst maintaining performance.
- **Comprehensive enterprise automation:** Combining with Red Hat Ansible Automation Platform enables deeper and more comprehensive integration with the overall IT environment, such as automated integration into ITSM and updating of backup facilities.

Tech Data Centre of Excellence

Red Hat Ansible Automation Platform with Event-Driven Ansible and IBM Turbonomic, available for Demos, PoCs and Workshops.

USE CASES

- Resource re-allocation
- Seasonal resource adjustment
- Event-Driven Automation
- ITSM integration