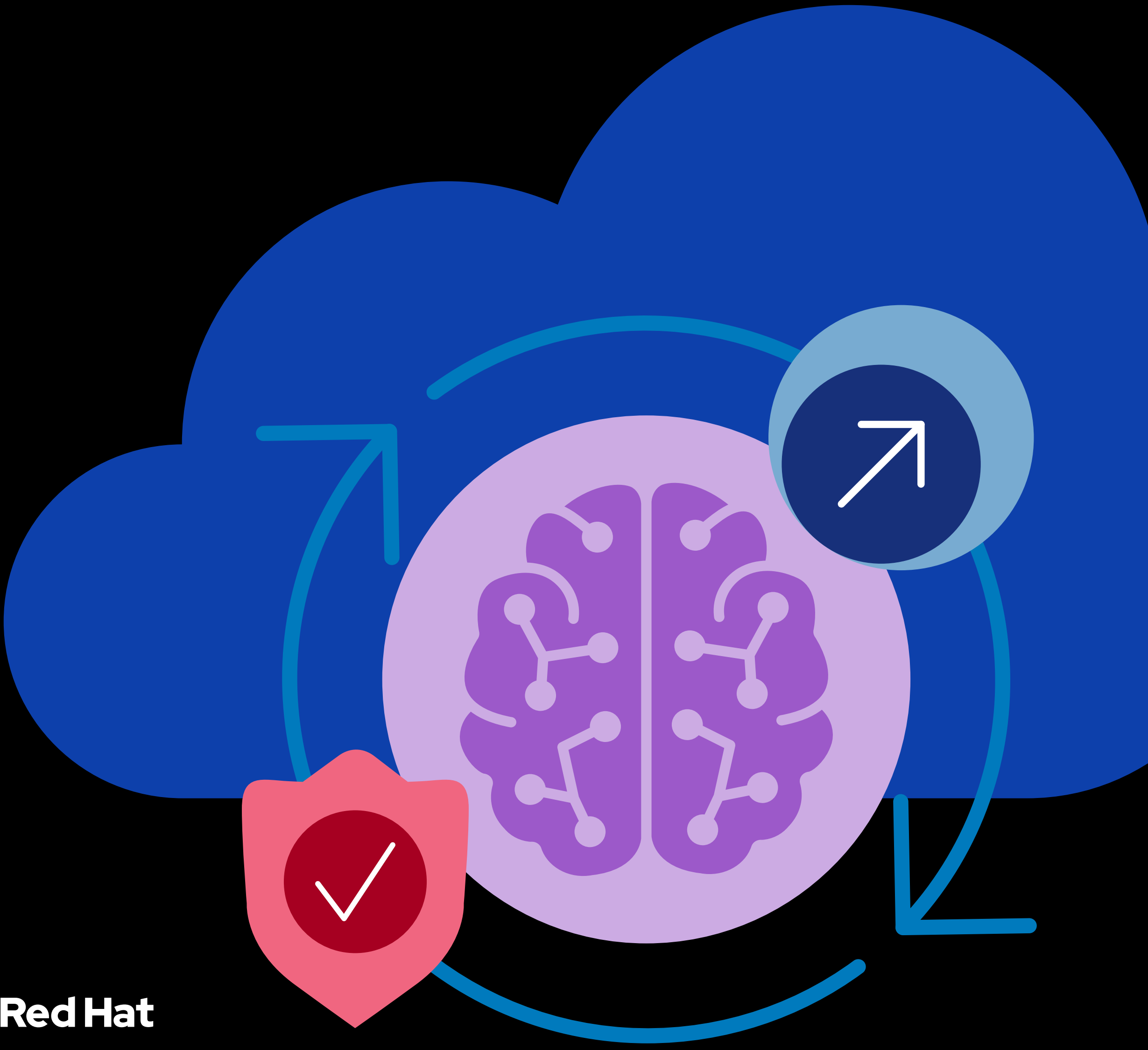
Solution Overview

How F5 Helps Secure and Scale Red Hat OpenShift AI for RAG Use Cases

F5 Distributed Cloud Services and BIG-IP Next Service Proxy for Kubernetes deliver the essential security and scalability to support Red Hat OpenShift AI for RAG use cases.



Key Benefits

Secure and optimize your RAG workloads

Comprehensive security
F5 Distributed Cloud Services provide a suite of features to protect AI workloads, including advanced API security, detailed logging and reporting, and consistent policies across environments.

Advanced traffic management
BIG-IP Next Service Proxy for Kubernetes optimizes traffic distribution across clusters, enabling AI applications to handle dynamic workloads without performance degradation.

Scalability and reliability
The combination of Distributed Cloud Services and BIG-IP Next SPK enhances AI apps' performance and reliability and efficiently manages increased demand and dynamic workloads.

In the rapidly evolving landscape of artificial intelligence (AI), Retrieval-Augmented Generation (RAG) has emerged as a powerful machine learning technique to enhance the accuracy and relevance of AI models.

Red Hat® OpenShift® combined with Red Hat OpenShift AI provides a robust foundation that supports RAG-powered models for deploying AI-driven applications.

However, maintaining the security and scalability of these applications is critical. This is where F5® Distributed Cloud Services and F5® BIG-IP® Next™ Service Proxy for Kubernetes (SPK) come into play, delivering advanced security and traffic management solutions to protect and optimize AI workloads.

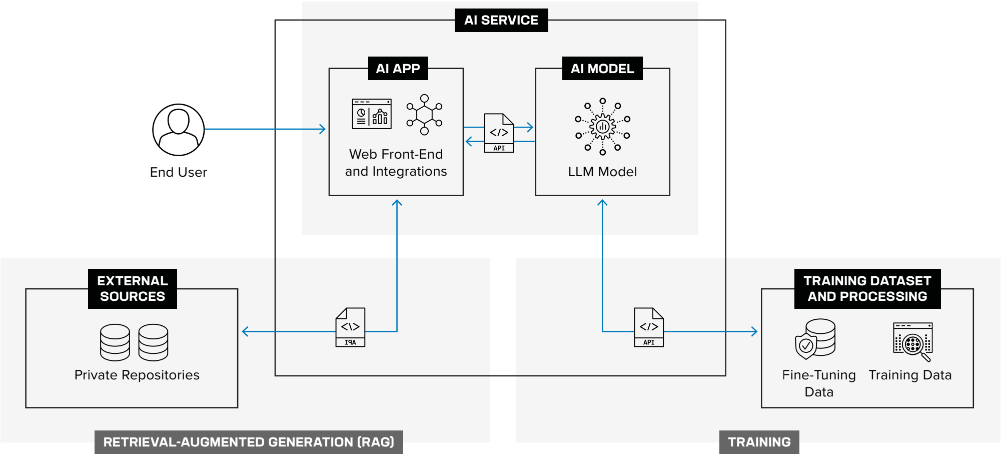


Figure 1: RAG is a machine learning technique that can improve the accuracy and relevance of your AI models.

Why Red Hat OpenShift AI?

Red Hat OpenShift AI builds on Red Hat OpenShift to deliver a consistent, streamlined, and automated experience for handling the workload and performance demands of AI/ML projects.

Key features of Red Hat OpenShift AI include:

- **Scalability:** Red Hat OpenShift AI supports the scaling of AI workloads across hybrid cloud environments, assuring that applications can handle increased demand without compromising performance.
- **Integration:** The platform integrates with various AI tools and frameworks, enabling organizations to utilize existing investments in AI technologies.

Red Hat OpenShift AI is a comprehensive platform designed to facilitate the deployment, management, and scaling of AI workloads.

- **Security:** Red Hat OpenShift AI leverages the security capabilities of the underlying OpenShift platform, including features such as role-based access control (RBAC), network policies, and integration with encryption and compliance tools, to help safeguard AI models and data.
- **Flexibility:** It supports a wide range of generative and predictive AI use cases, from machine learning and deep learning to natural language processing and computer vision.

SECURING AI WORKLOADS WITH F5

F5 Distributed Cloud Services provide a comprehensive suite of security features designed to protect AI workloads in hybrid cloud environments.

Key capabilities include:

- **API security:** Distributed Cloud Services provide API security features that help protect communication channels between AI models and external systems from potential threats and tampering. This is achieved through SSL/TLS encryption and by protecting API gateways.
- **Detailed logging and reporting:** Enhanced visibility and control over AI deployments are provided through detailed logging and reporting capabilities, ensuring compliance and security.
- **Consistent security policies:** Distributed Cloud Services enforce consistent security policies across hybrid cloud environments, protecting both the data processed by AI models and the communication channels through which it flows.

SCALING AI WORKLOADS WITH BIG-IP NEXT SPK

F5 BIG-IP Next SPK optimizes traffic distribution across clusters, enhancing AI application performance while mitigating risks associated with server overloads.

Key capabilities include:

- **Traffic management:** BIG-IP Next SPK optimizes traffic distribution, ensuring that AI applications can handle dynamic workloads without performance degradation.
- **Load balancing:** By providing advanced load balancing capabilities, BIG-IP Next SPK enables AI applications to remain responsive and reliable, even under heavy traffic conditions.
- **Integration with OpenShift:** BIG-IP Next SPK seamlessly fuses with Red Hat OpenShift, leveraging Kubernetes ingress resources to manage communications to and from clusters.

Key Features

Reap F5’s most advanced solutions

Industry-leading web app and API protection for distributed clouds
F5 Distributed Cloud Services protect and power AI-based apps from the data center to the edge. Distributed Cloud Services include:

- WAF (Web App Firewall)
- DDOS Mitigation
- API Security
- Bot Defense
- Network Connect
- App Connect
- and much more.

Central point for transitioning from complex networks to cloud-native infrastructure
BIG-IP Next SPK aids this transition by providing a central point for ingress/egress networking control and security for the Kubernetes cluster, and securely managing traffic to and from multiple networks.

F5 Distributed Cloud Services and BIG-IP Next SPK deliver the essential security and scalability to support Red Hat OpenShift AI for RAG use cases.

ADVANCING AI PERFORMANCE AND RELIABILITY

The combination of F5 Distributed Cloud Services and BIG-IP Next SPK creates a fortified security solution for protecting RAG workloads. The synergy between F5's advanced security and traffic management technologies and Red Hat OpenShift results in a robust, protected, and scalable foundation for AI-driven initiatives.

Conclusion

In summary, F5 Distributed Cloud Services and BIG-IP Next SPK deliver the essential security and scalability to support Red Hat OpenShift AI for RAG use cases. By leveraging these F5 solutions, organizations can help safeguard their AI applications, maintain reliability, and efficiently manage dynamic workloads, leading to improved outcomes and a better user experience.

See how F5 Distributed Cloud Services and/or BIG-IP Next SPK works with a free trial.

Try Red Hat OpenShift in the no-cost Developer Sandbox.