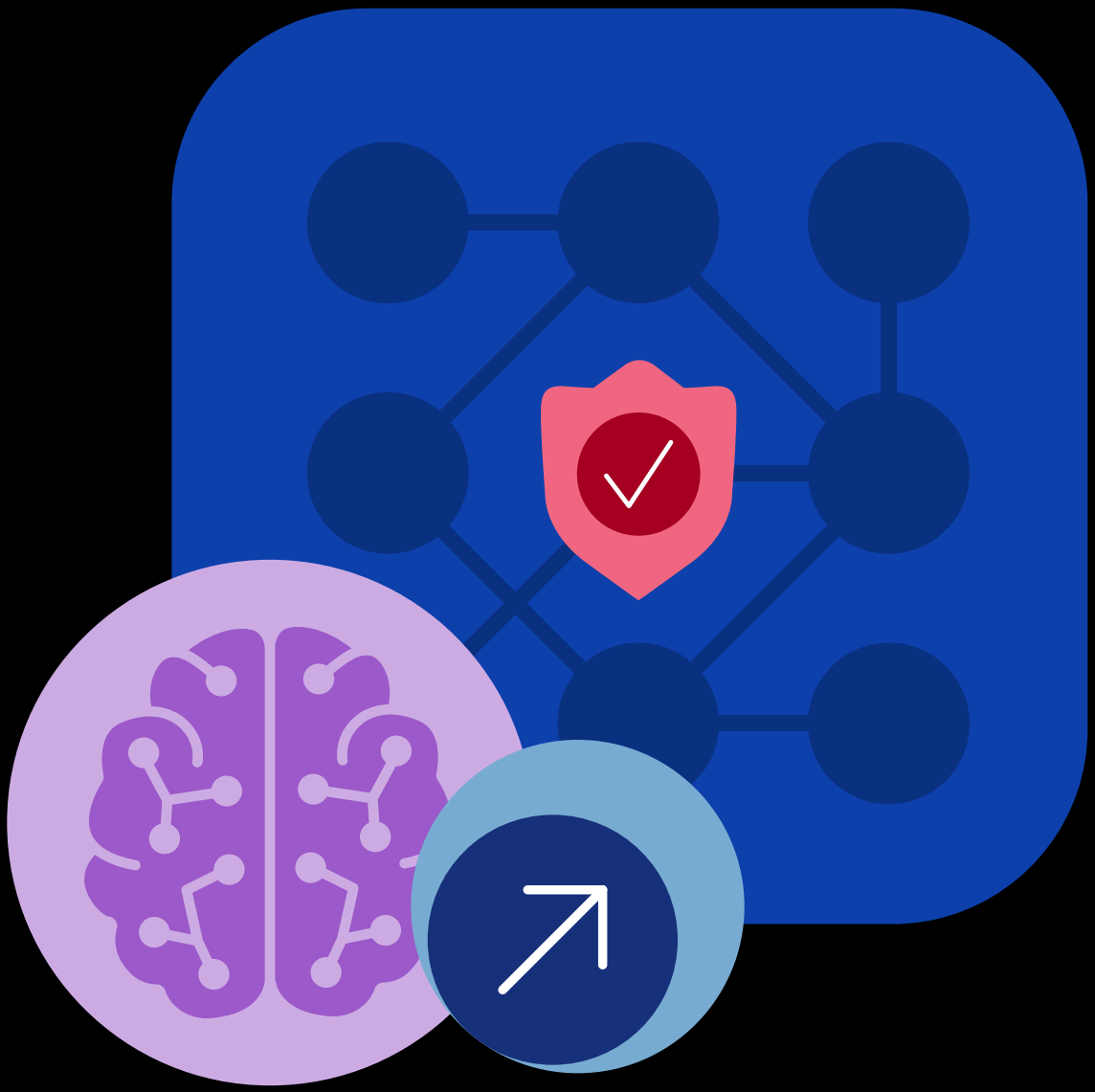
White Paper

Enhancing Red Hat OpenShift AI Application Deployments with F5



Red Hat

The F5 and Red Hat collaboration underscores a commitment to delivering cutting-edge solutions that address the complex challenges of modern AI workloads, fostering an environment where innovation and security go hand in hand.

F5 solutions enhance the scalability and security of AI workloads deployed on Red Hat OpenShift AI. This combination leverages the strengths of F5 advanced security and traffic management solutions to provide robust protection and optimized performance for AI-driven applications. By integrating the F5 comprehensive technology portfolio, including F5® Distributed Cloud API Security, F5 VELOS®, and F5 Distributed Cloud Web App and API Protection (WAAP), organizations can ensure that their AI deployments are secure, resilient, and scalable across hybrid cloud environments. F5 and Red Hat aim to deliver cutting-edge solutions that address the complex challenges of modern AI workloads, fostering an environment where innovation and security go hand in hand.

Provide Robust API Security

Red Hat OpenShift AI is a powerful platform for deploying AI-driven applications. Optimizing it with F5 Distributed Cloud API Security enhances its capabilities with advanced protection and resilience. Distributed Cloud API Security offers comprehensive defenses against a wide array of API-specific threats, protecting sensitive data and critical applications from unauthorized access and cyberattacks. By leveraging advanced features such as automated threat detection, real-time traffic monitoring, and sophisticated anomaly detection, Distributed Cloud API Security provides a fortified security framework that optimizes OpenShift's containerized environments. This allows for more secure API communication across hybrid cloud infrastructures, helping mitigate risks and enhance the overall security posture of AI deployments.

Furthermore, Distributed Cloud API Security supports the dynamic and scalable nature of Red Hat OpenShift AI hybrid environments. It provides consistent security policies and enforcement across various deployment models, whether on-premises or in the cloud. This unified approach not only simplifies security management but also helps AI applications remain compliant with industry standards and regulations. Through robust API security measures, Distributed Cloud API Security helps organizations deploy AI solutions with confidence, knowing that their hybrid cloud environments are well-protected against evolving cyber threats. This combination of F5 Distributed Cloud API Security and Red Hat OpenShift AI fosters a more secure, reliable, and scalable foundation for innovative AI-driven initiatives.

F5's dual-layered approach to security not only protects the data being processed by the AI models but also secures the communication channels through which data flows.

Accelerate Ingest and Secure Amazon S3 Compatible Workloads

In addition to enhancing API security, F5 VELOS plays an important role in load balancing and securing Amazon S3 compatible object stores within the Red Hat OpenShift AI. F5 VELOS provides advanced load balancing capabilities that enable optimal distribution of data requests across multiple object storage instances, thereby enhancing performance and availability. By intelligently managing traffic, VELOS mitigates the risk of bottlenecks and server overloads, which can be particularly critical in AI-driven environments where large volumes of data are processed. This robust load balancing functionality helps maintain high availability and reliability, enabling AI applications to have consistent access to the data they need for training and inference.

Furthermore, VELOS strengthens the security of Amazon S3 compatible object stores by incorporating comprehensive security measures that help protect against a wide range of threats. These measures include SSL/TLS encryption for more secure data transmission, advanced firewall capabilities to block unauthorized access, and sophisticated threat detection to identify and help mitigate potential security breaches. By utilizing these security features with the Red Hat OpenShift AI platform, organizations can better safeguard their sensitive data and maintain the integrity of their AI workflows. The ability of VELOS to offer both load balancing and enhanced security creates a resilient infrastructure that supports the demanding requirements of AI applications, helping them run smoothly and safely across hybrid cloud environments.

Secure Model Inference Servers

Distributed Cloud WAAP plays a valuable role in helping secure model inference within Red Hat OpenShift AI environments. By providing comprehensive security measures, Distributed Cloud WAAP helps ensure that AI models, once deployed, are protected against a variety of threats that could compromise their integrity and performance. These measures include robust input validation, which helps prevent malicious data from being used to manipulate or exploit AI models during inference. Additionally, Distributed Cloud WAAP leverages machine learning techniques to detect and respond to anomalous behavior in real time, so that any deviations from expected patterns are promptly addressed. This proactive approach to security helps safeguard the inference process, maintaining the accuracy and reliability of AI outputs.

Moreover, Distributed Cloud WAAP advanced security features extend to protecting the communication channels through which model inferences are made. By implementing SSL/TLS encryption and secure API gateways, Distributed Cloud WAAP helps data exchanged between AI applications and other systems to remain confidential and protected. This is particularly important in hybrid cloud environments where data traverses multiple networks and platforms. Distributed Cloud WAAP also provides detailed logging and reporting capabilities, enabling organizations to monitor inference activities and quickly identify any suspicious or unauthorized access attempts. This level of visibility and control is essential for maintaining the overall security posture of AI deployments and helping to ensure that model inferences are conducted in a more secure, compliant, and trustworthy manner.

Securing RAG Servers

Securing retrieval-augmented generation (RAG) workloads on the Red Hat OpenShift AI platform is crucial for maintaining the integrity and performance of AI-driven applications. F5® BIG-IP® Next™ Service Proxy for Kubernetes (SPK), when combined with Distributed Cloud WAAP, offers a comprehensive security solution tailored for these demanding workloads. BIG-IP Next SPK provides advanced traffic management capabilities, promoting efficient distribution of RAG workloads across the cluster. This optimized traffic distribution not only enhances the performance of AI applications but also helps mitigate the risks associated with server overloads and bottlenecks. By leveraging its robust security features, BIG-IP Next SPK helps safeguard RAG workloads from various threats, including malicious data injections and unauthorized access attempts. Its deep integration with Red Hat OpenShift AI's containerized environments enables seamless deployment and management of security policies, providing a more secure and compliant framework for RAG workloads.

When paired with Distributed Cloud WAAP, the security framework is further strengthened, providing an additional layer of protection for RAG workloads. Distributed Cloud WAAP's advanced API security features, including automated threat detection and real-time traffic monitoring, help ensure that any anomalies or potential threats are swiftly identified and mitigated. This dual-layered approach to security not only protects the data being processed by the AI models but also secures the communication channels through which data flows. The ability of Distributed Cloud WAAP to provide consistent security policies across hybrid cloud environments, keeping RAG workloads more secure, regardless of their deployment model. By combining the strengths of BIG-IP Next SPK and Distributed Cloud WAAP, organizations can deploy RAG workloads on Red Hat OpenShift AI with confidence, knowing that their AI applications are protected against both current and emerging cyber threats.

Secure and Scale AI Data Factories with Red Hat and F5

The optimization of the F5 comprehensive technology portfolio with the Red Hat OpenShift AI platform can significantly enhance the deployment, security, and scalability of AI-driven workloads. By incorporating Distributed Cloud API Security, organizations can better safeguard their sensitive data and critical applications from unauthorized access and cyber threats through automated threat detection and real-time traffic monitoring. This approach helps secure API communication across hybrid cloud infrastructures, aligning seamlessly with the containerized environments of Red Hat OpenShift AI and supporting dynamic, scalable deployments.

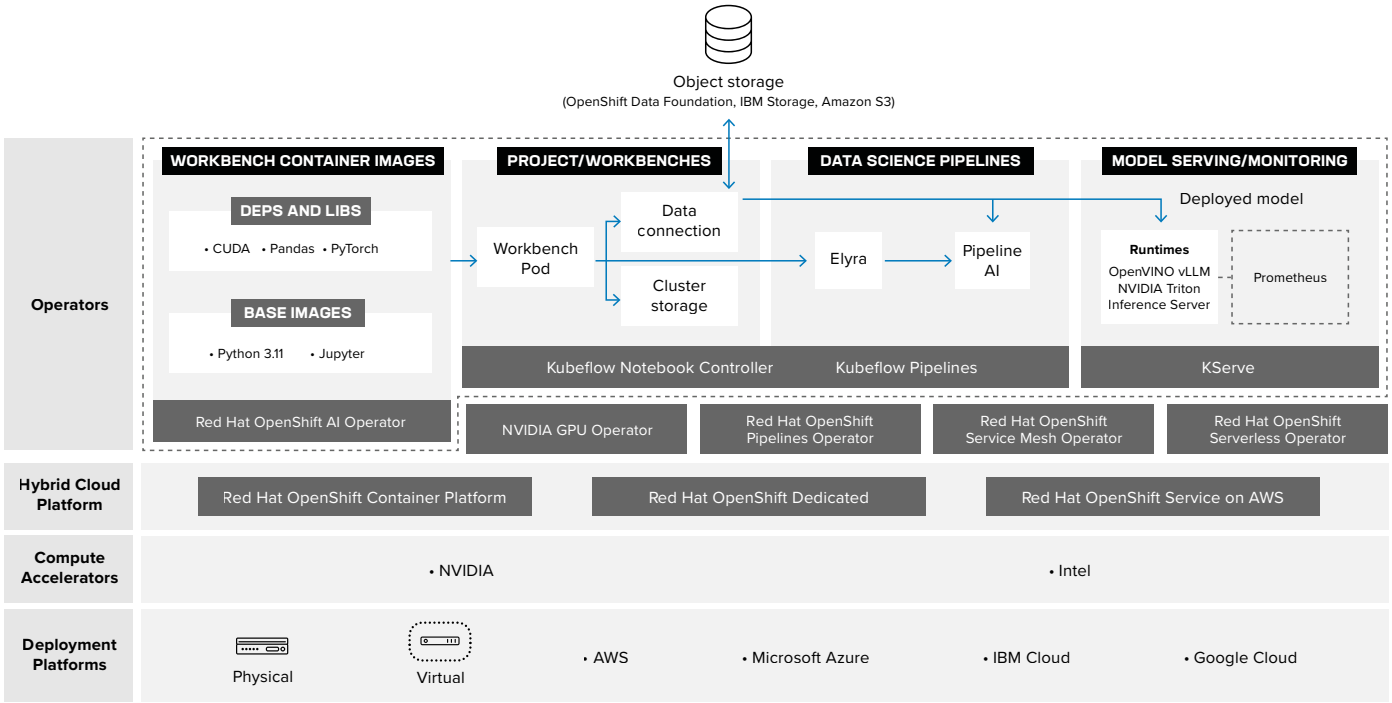


Figure 1: Architecture of OpenShift AI components and concepts.

Moreover, VELOS plays a vital role in optimizing the performance and availability of Amazon S3 compatible object stores within Red Hat OpenShift AI platforms. Its advanced load balancing capabilities helps prevent bottlenecks and server overloads, providing AI applications with consistent access to necessary data. Additionally, VELOS strengthens the security of these object stores with SSL/TLS encryption, advanced firewall capabilities, and sophisticated threat detection, thereby maintaining the integrity of AI workflows.

Distributed Cloud WAAP further secures model inference processes within Red Hat OpenShift AI environments, leveraging robust input validation and machine learning techniques to detect and respond to anomalous behavior. This helps ensure that AI models remain accurate and reliable, with more secure and tamper-resistant communication channels backed by SSL/TLS encryption and secure API gateways. The detailed logging and reporting capabilities provided by Distributed Cloud WAAP enhance visibility and control over AI deployments, for greater compliance and security.

Finally, the combination of BIG-IP Next SPK and Distributed Cloud WAAP delivers a fortified security solution for securing RAG workloads. BIG-IP Next for SPK optimizes traffic distribution across clusters, enhancing AI application performance while minimizing risks associated with server overloads. Distributed Cloud WAAP complements this by offering advanced API security features and consistent security policies across hybrid cloud environments, helping protect both the data processed by AI models and the communication channels through which it flows.

Conclusion

In conclusion, the combination of F5's advanced security and traffic management technologies and Red Hat OpenShift AI help establish a more robust, secure, and scalable foundation for AI-driven initiatives. Organizations can confidently deploy their AI applications, knowing they are protected against evolving cyber threats and capable of meeting the demanding requirements of modern AI workloads.

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

DC 02.2025 | WP-1521231652