F5 Introduces AI Gateway to Manage and Secure Surging AI Traffic and Application Demands

Nov 12, 2024 7:00 AM

- *Manages security for both users and providers of AI services, overseeing authentication and zero trust capabilities*
- *Prioritizes the consumption of AI services across teams, applications, and APIs*
- *Enables integrated automation with AI services and models from OpenAI, cloud providers, open source, and more*

SEATTLE--(BUSINESS WIRE)-- F5 (NASDAQ: FFIV) announced early access of F5 AI Gateway to streamline interactions between applications, APIs, and large language models (LLMs) driving enterprise AI adoption. This powerful containerized solution optimizes performance, observability, and protection capabilities—all leading to reduced costs. Integrated with F5's portfolio, AI Gateway gives security and operations teams a seamless path to adopting AI services through significantly improved data output quality and a superior user experience.

According to F5's State of AI Application Strategy Report, 75% of enterprises are implementing AI. Like countless modern applications, AI services are largely delivered and consumed via APIs. However, enterprises face many additional challenges in architecting and scaling AI-fluent apps and services. As an example, efficient operations require close monitoring of increasingly relevant metrics such as GPU compute costs and system responsiveness, as well as emerging regulatory compliance concerns.

"LLMs are unlocking new levels of productivity and enhanced user experiences for customers, but they also require oversight, deep inspection at inference-time, and defense against new types of threats," said Kunal Anand, Chief Innovation Officer at F5. "By addressing these new requirements and integrating with F5's trusted solutions for API traffic management, we're enabling customers to confidently and efficiently deploy AI-powered applications in a massively larger threat landscape."

Real-world AI solutions require optimized request, response, and prompt interactions across an entire data ecosystem. F5 AI Gateway observes, optimizes, and secures a vast number of user and automated variables to offer cost reductions, mitigate malicious threats, and ensure regulatory compliance.

F5 AI Gateway is designed to meet customers—and their apps—at the ideal place in their AI journey. It can be deployed in any cloud or data center and will natively integrate with F5's NGINX and BIG-IP platforms to take advantage of F5's leading app security and delivery services in traditional, multicloud, or edge deployments. In addition, the solution's open extensibility enables organizations to develop and customize programmable security and controls enforced by F5 AI Gateway. These processes can be easily updated and applied dynamically to drive instant adherence to security policies and compliance mandates.

"AI-powered applications will become a cornerstone for nearly every business and organization in the coming years," said Shari Lava, Senior Director, AI and Automation at IDC. "F5's introduction of an AI gateway to its application stack of services enables its customers to have greater flexibility in how they build their AI application structure, but still have enhanced protection and model optimization."

F5 AI Gateway:

- Delivers security and compliance policy enforcement with automated detection and remediation against the risks identified in the OWASP Top Ten for LLM Applications.
- Offloads duplicate tasks from LLMs with semantic caching, enhancing the user experience and reducing operations costs.
- Streamlines integration processes, allowing developers to focus on building out AI-powered applications rather than managing complex infrastructures.
- Optimizes load balancing, traffic routing, and rate limiting for local and third-party LLMs to maintain service availability and enhance performance.
- Provides a single API interface that developers can use to access their AI model of choice.

"F5's AI Gateway is an integral part of our AI strategy," said Austin Geraci, CTO of WorldTech IT. "With this technology, our customers are able to develop both internal and external-facing AI applications capable of handling a surge in queries without degradation to site and application performance. F5 brings leading app security and delivery capabilities to accelerate AI experiences at scale. With F5 AI Gateway, semantic caching and intelligent traffic routing alone represent significant cost savings, and the unification of F5 services will save customers hundreds of hours of integration work."

Supporting Resources

- F5 AI Gateway product page
- F5 AI Gateway overview
- F5 AI Gateway blog post

About F5

F5 is a multicloud application security and delivery company committed to bringing a better digital world to life. F5 partners with the world's largest, most advanced organizations to secure every app—on premises, in the cloud, or at the edge. F5 enables businesses to continuously stay ahead of threats while delivering exceptional, secure digital experiences for their customers. For more information, go to f5.com. (NASDAQ: FFIV)

You can also follow @F5 on X or visit us on LinkedIn and Facebook to learn about F5, its partners, and technologies. F5, BIG-IP, and NGINX are trademarks, service marks, or tradenames of F5, Inc., in the U.S. and other countries. All other product and company names herein may be trademarks of their respective owners.

Source: F5, Inc.

View source version on businesswire.com: https://www.businesswire.com/news/home/20241112358635/en/

Jenna Becker
F5
(415) 857-2864
j.becker@f5.com

Holly Lancaster
WE Communications
(415) 547-7054
hluka@we-worldwide.com

Source: F5, Inc.