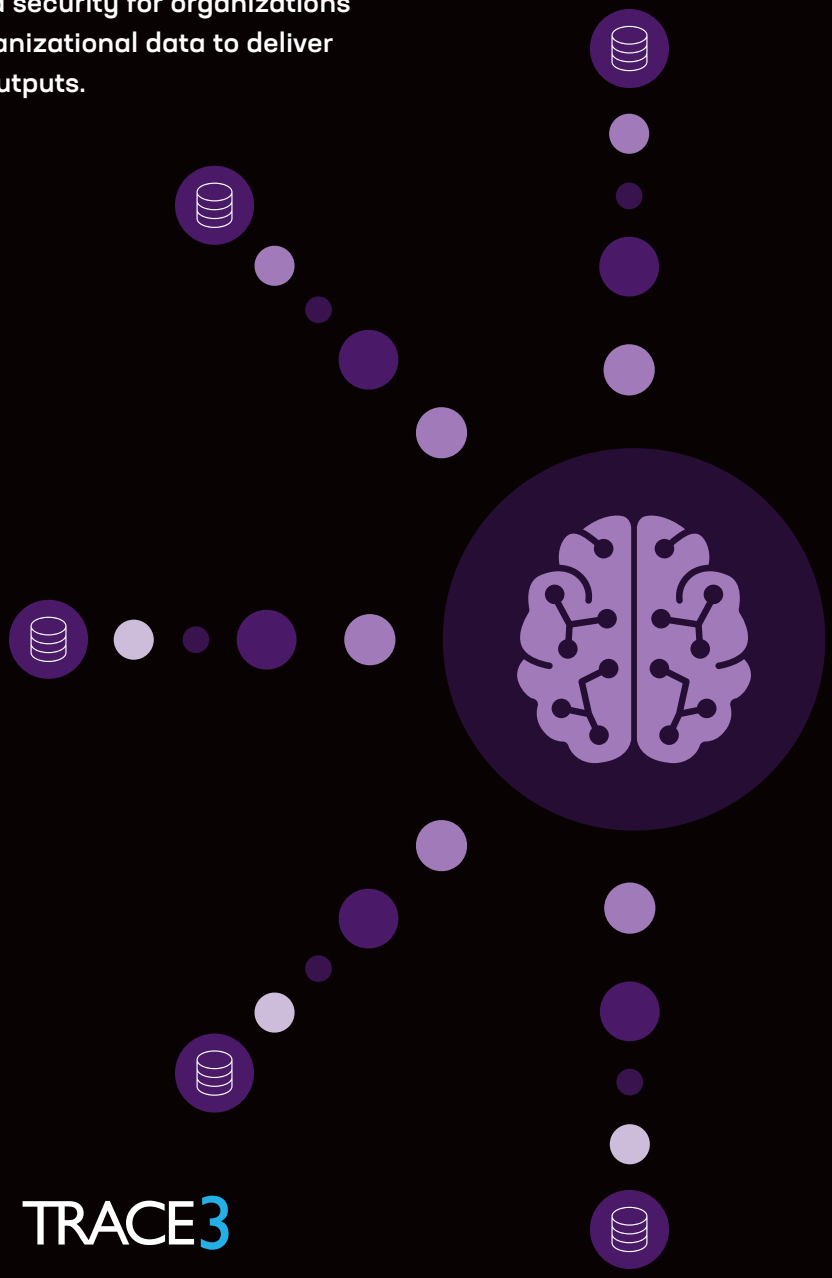
Solution Overview

Enhance AI Accuracy During Inference with RAG Powered by F5 and NetApp

Enable high-performance data mobility and security for organizations combining foundational AI models with organizational data to deliver more accurate and contextually aware AI outputs.



F5 | NetApp | TRACE3

Benefits

Enhance Accuracy and Contextual Relevance
Enable access to relevant data resulting in more accurate and context-specific AI outputs.

Simplify Data Integration
Provide unified operations for data management, security, and networking.

Secure Access
Improve data protection and compliance by providing secure, restricted access to data.

Ensure High-Performance AI Deployments
Make AI deployments efficient and secure, allowing enterprises to confidently harness AI capabilities.

Deliver Value and Maximize AI Investments with RAG

Enterprises are adopting AI to enhance decision making and operational efficiency. As they look to cultivate value from AI and further their competitive advantage, ensuring AI models offer accurate, contextually aware, and timely outputs is critical.

Organizations turn to three options when deploying AI models: using commercially available or open-source foundational models, fine-tuning an existing model, or training a model from scratch. For most enterprises, fine-tuning or training requires specific use cases and substantial investments in AI factories or large-scale AI infrastructure. The cost-prohibitive nature of these options results in organizations deploying a commercially available or open-source model.

Since commercial and open-source models are trained on public data sets and information, they do not have specific context about an organization. Even fine-tuned or newly trained models without access to the latest data sources are often outdated once training and fine-tuning is complete. Additionally, some industries like finance and healthcare are limited in their available training materials, as they cannot use health records or individual financial data for model training. In these cases, the question arises: how can AI models be contextually aware with these knowledge limitations?

To solve the knowledge gap and ensure AI delivers maximum value, companies are turning to retrieval-augmented generation (RAG), which overcomes these limitations by enabling organizations to combine pre-trained large language models (LLMs) or other generative AI models with their unique, proprietary data. RAG incorporates connected corporate data and storage sources to complement model prompts and deliver more accurate and contextually aware outputs during inferencing.

RAG overcomes AI model knowledge limitations by enabling organizations to combine pre-trained LLMs or other generative AI models with their unique, proprietary data.

However, implementing RAG in enterprise settings poses many challenges, particularly relating to delivering and consolidating data across hybrid and multicloud environments with siloed storage deployments. Organizations must navigate the complexity of securely connecting apps with data across numerous storage deployments that require timely inference outputs generated from AI models. Ensuring robust security during this process is essential to protect against unauthorized data access and to ensure data integrity. Building secure network connectivity across hybrid and multicloud networking environments offers more efficient data transit and delivery, while helping enterprises adhere to strict data privacy for internal, proprietary data. Addressing these challenges is critical to harnessing the potential of RAG without compromising networking simplicity or data security.

Benefits

Increase Data Mobility and Protection
Migrate data effortlessly across zones and regions while ensuring valuable enterprise data remains secure.

Transform Data into a Strategic Asset
Drive secure use of AI within unique business contexts by with secure connections to organizational data stored across the hybrid or multicloud environment.

Future-Proof Your AI Deployments
Unlock the full power of AI across distributed IT landscapes while maintaining security and performance.

Seamless Data Integration for RAG with F5 and NetApp

The F5 and NetApp technology alliance addresses the challenges of implementing RAG in large-scale environments while navigating the complexities of connecting data sources across multiple clouds. F5® Distributed Cloud Services enable enterprises to create secure on- and off-ramps to the private F5 Global Network. Doing so provides secure data transit between NetApp storage instances and AI models across F5's privately managed global network.

The combined solution enhances enterprise AI capabilities by ensuring secure access to private data from any location, facilitating seamless mobility and access. The F5® Distributed Cloud Platform offers native network tooling that seamlessly integrates with existing public cloud networks, eliminating the need to build custom routing and transit between different providers. Enterprises can unify their operational models for data management, security, and networking, helping reduce costs and efficiently implement RAG with siloed data and storage deployments.

F5's secure multicloud networking solutions optimize performance while enhancing data protection and access. This allows for quick, secure data mobility, reducing overhead for IT departments and enabling effortless data migration across zones and regions. NetApp appliances typically expose data volumes through NAS protocols like NFS or SMB. To support RAG, exposing files as objects via HTTPS and S3-compliant APIs is becoming more common, a preferred method for cloud-native services. 5 Distributed Cloud Services facilitates three core methods for data ingestion into RAG AI solutions from NetApp deployments. First, Distributed Cloud Network Connect enables secure Layer 3 connectivity between model inferencing and NetApp volumes. Second, Distributed Cloud App Connect provides TCP load balancing enabling NetApp volumes to connect to RAG workloads. TCP load balancing simplifies the complexities posed by overlapping IP ranges at either end of secure data transfers. Third, Distributed Cloud App Connect enables HTTPS load balancing between storage deployments, securely integrating NetApp data sources to the AI models, including support for S3 protocols. This leverages robust security from the Distributed Cloud Services platform, allowing organizations to deploy advanced security controls including web application and API protection and DDoS mitigation, empowering enterprises to enforce positive security controls for their proprietary data. As a result, enterprises can support business-critical AI applications with confidence, knowing their data is protected and efficiently managed for AI-based initiatives.

Securely Connecting Disparate Storage Deployments to AI Models

Self-hosted LLMs and AI models like Llama 3.2 leverage software deployed closer to data storage sources that connect to the F5 Distributed Cloud Platform, providing high-performance network connectivity between NetApp storage deployments across the F5 Global Network. As requests are made to the model, access to relevant data, regardless of the location, is augmented to the prompt or request for inferencing.

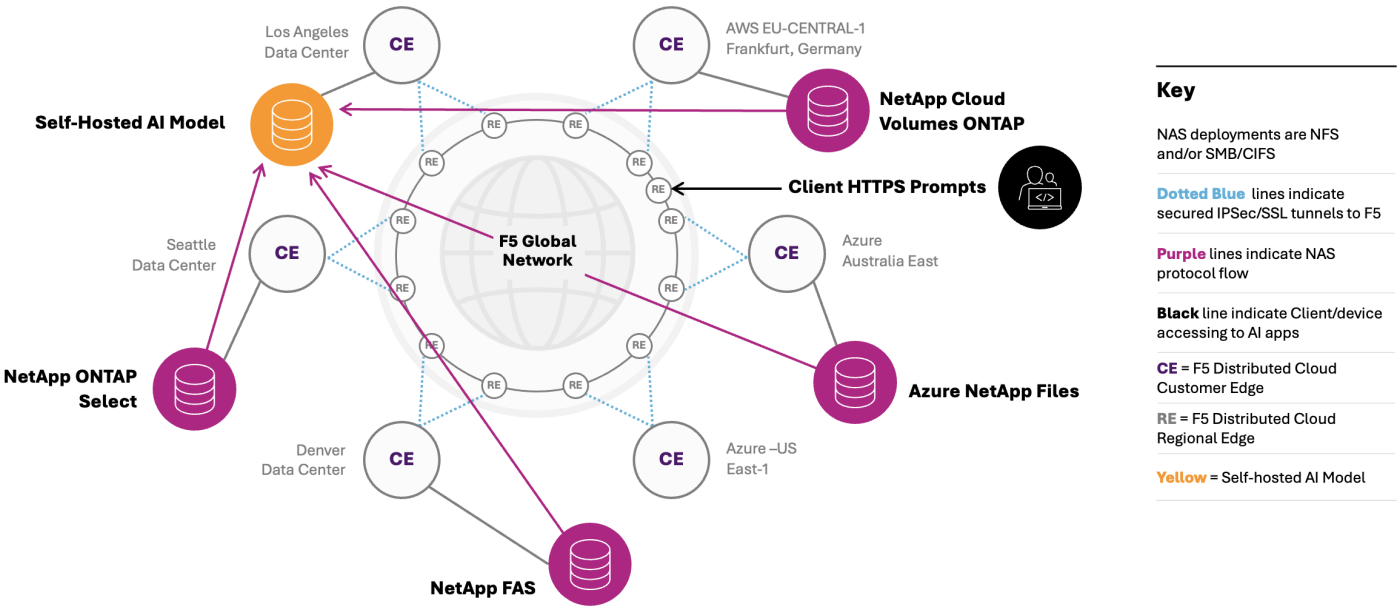


Figure 1: Securely connect LLMs to NetApp storage deployments via the F5 Global Network.

Transform data storage into a strategic investment, propelling innovation and enhancing user experiences during inferencing.

Ensure AI Infrastructure Runs at Optimal Performance

Integrating F5's high-performance secure multicloud networking capabilities with NetApp's robust data management solutions enables RAG, which reduces calculations and provides contextual responses with access to the latest information from connected data sources. By providing secure access to NetApp storage deployments, organizations realize the benefits of unparalleled performance as they harness AI within their business contexts. With storage as a strategic asset, enterprises can drive further innovation and operational efficiencies while enhancing their competitive advantage.

Additional Resources Available Online

Deploying AI models? Contact F5

Unlock the power of RAG today by contacting your F5 and NetApp account teams.

Unlocking the Magic of AI – Read the article

Combine a foundational AI model with organizational data to enhance the value of AI investments.

F5 and NetApp – Explore the partnership

Deliver inference with RAG by connecting to on-premises local NetApp storage using F5 Distributed Cloud Services.

To learn more about Trace3, NetApp, and F5, please visit trace3.com.