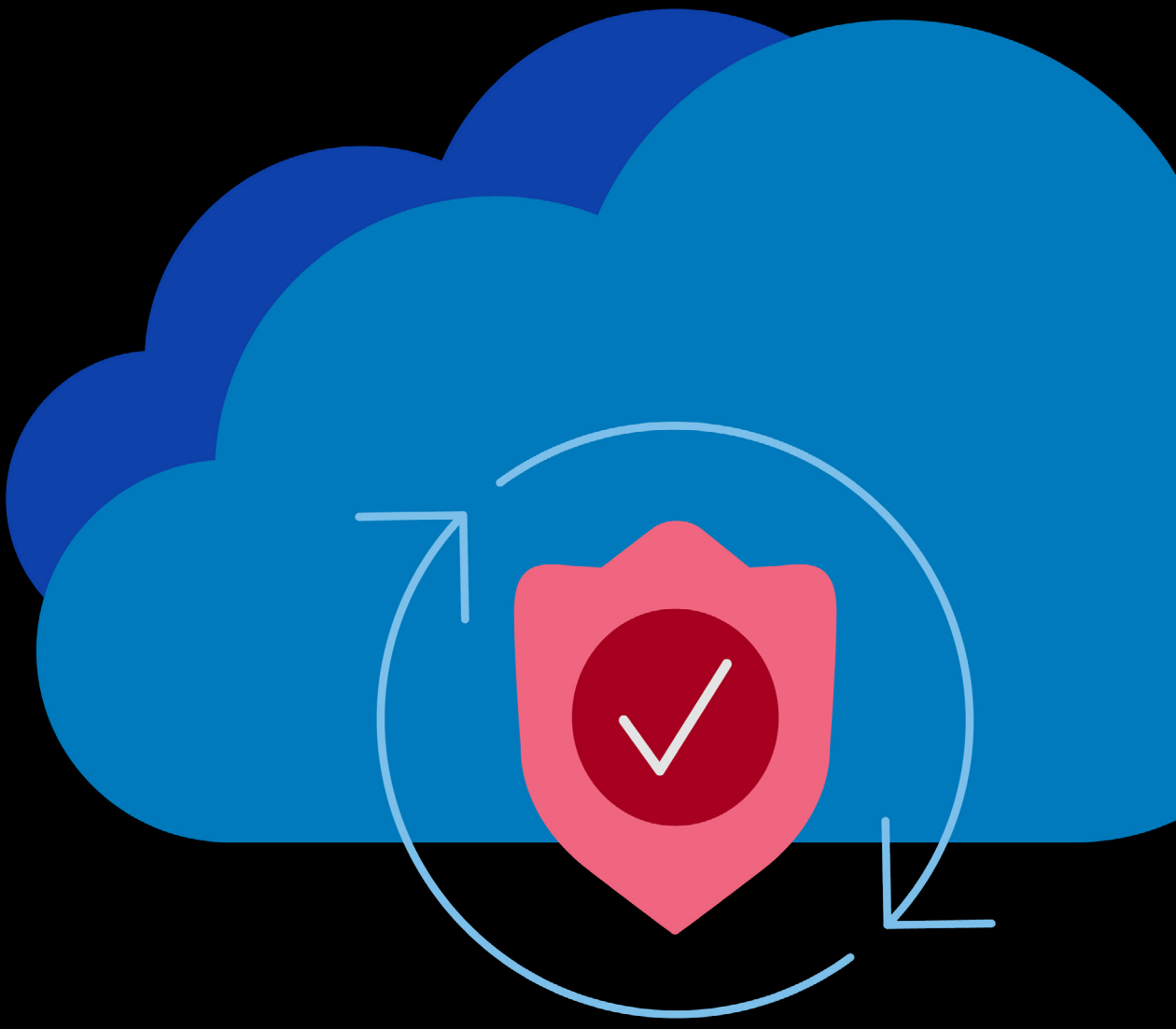
Solution Overview

F5 and NetApp for AI RAG

F5 Distributed Cloud Services greatly simplify implementation and operational complexity across proprietary RAG local data silos.



NetApp

Key Benefits

Simplified multicloud operations
Reduce complexity and simplify networking and security operations from data center to cloud to edge.

Built-in security
Securely expose private data to LLMs, no matter where it resides.

Accelerate transfers
Exercise more control over performance and transfer speeds with private network.

Rationalized tools
Move to a unified platform for network and security, which allows you to consolidate tools and retire redundant point tools.

Most publicly available LLMs are static, with knowledge cutoff dates that prevent your generative AI (GenAI) solutions from creating responses grounded in up-to-date information. Generic LLMs also lack company-specific knowledge; outputs should be tailored to differentiate your results and fit your unique use cases.

To combat these limitations, retrieval-augmented generation, or RAG, is used to optimize LLMs with needed context. RAG is a search and retrieve process that pulls in current data from various sources to augment a GenAI request before the LLM creates a response that is sent to the user—all without having to train your own model (although fine tuning with custom data has obvious benefits). This allows for more refined, accurate, and personalized responses.

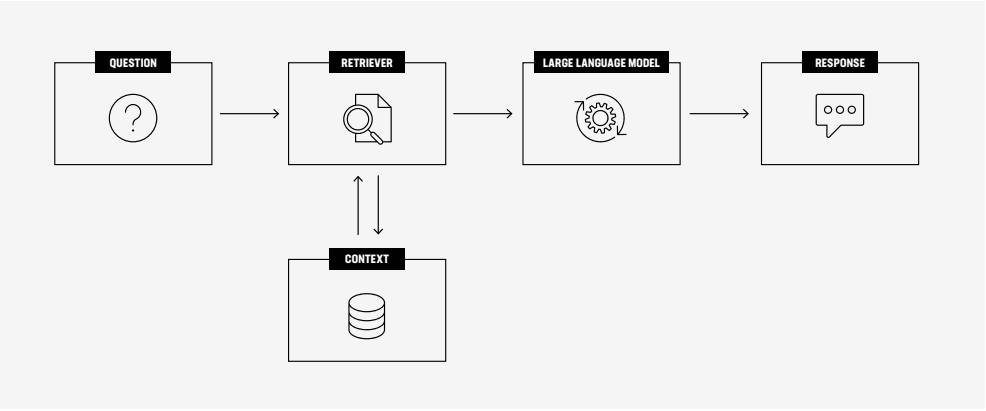


Figure 1: Basic retrieval-augmented generation flow.

However, RAG is not necessarily an easy fix. Challenges include:

- Securely connecting to dispersed data sources
- Ensuring optimal performance across the data pipeline
- Controlling access to and exposure of sensitive data
- Scaling for large data sets

Simplify RAG with F5® Distributed Cloud Services for Secure Multicloud Networking and NetApp ONTAP

The contextual data needed for RAG can be stored across the environment in dispersed, siloed data sources. This complexity can make it challenging to retrieve data quickly and securely. F5® Distributed Cloud Services for secure multicloud networking provide a unified platform for secure data access across hybrid and multicloud environments. This private, secure connectivity enables fast, safe delivery of data between distributed NetApp ONTAP locations and the LLMs, wherever they are running.

Key Features

Private connectivity
Encrypted traffic travels over a global private backbone via the F5 Global Network.

Built-in app and API security
This includes WAF, bot defenses, and API protection.

Controlled access
Local data is only accessible from known LLMs; private endpoints maintain control in the cloud.

Marketplace availability
F5 Distributed Cloud solutions are available from Azure, AWS, and Google Cloud marketplaces.

Enabling RAG with a Self-Hosted LLM

LLMs benefit from cloud compute and GPU access, something often in short supply with on-premises production rollouts. A popular solution for self-hosted LLMs is to utilize public cloud for the compute side of AI projects, while leaving documents and data needed to feed an effective RAG solution on premises (typically for security, governance, or regulatory reasons).

Distributed Cloud Services for secure multicloud networking provide secure interconnectivity between disparate physical sites and heterogeneous cloud instances such as Azure Virtual Network (VNET) or AWS Virtual Private Cloud (VPC). For RAG, this showcases two key F5 features:

- Layer 3 encrypted interconnectivity from on-premises NetApp Volumes (network-attached storage, or NAS) or LUNs (block) containing critical documents for ingestion into RAG.
- Secure access to the LLM itself, for users, with HTTPS load balancers that include web application firewall, anti-bot mechanisms, and API automatic rate limiting to mitigate abusive prompt sources.

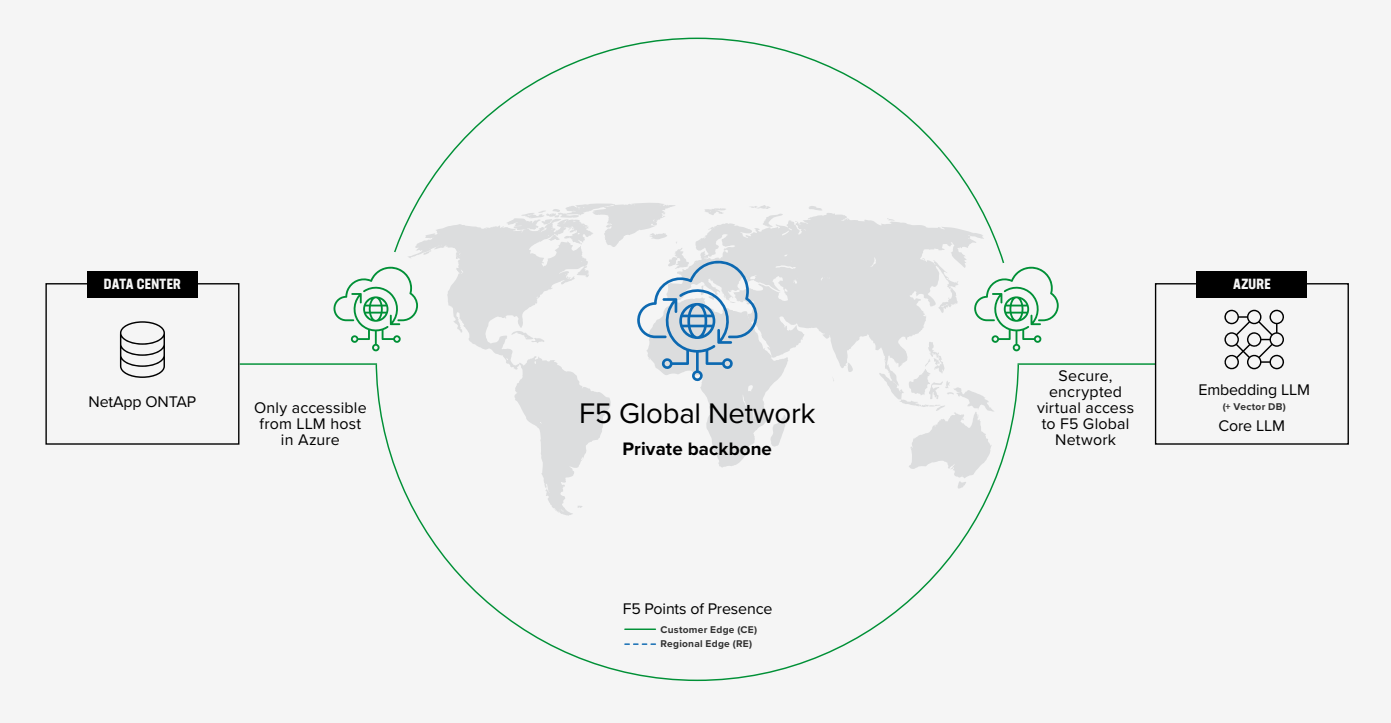


Figure 2: Secure connectivity between data-center hosted documents and LLM running in Azure–AWS and Google Cloud are also supported.

See our tech article on DevCentral to learn more about this use case.

Enabling AI RAG with Azure OpenAI Service

One of the easiest and most scalable ways to harness LLMs is to consume a leading service, such as Azure OpenAI Service. Access to hardware, such as GPUs, and the significant skillset needed to run LLMs on your own become non-issues. Consumption becomes an API call away.

But your data—AI prompts, sensitive corporate data, the requests, the responses—are exposed in this model, with everything traveling across the Internet.

Distributed Cloud Services for secure multicloud networking can be leveraged with privacy, where prompts are carried over secured, encrypted tunnels to and from the private F5 network, serving to connect on-premises enterprise locations and the enterprise's Azure VNET. The Azure OpenAI models are then exclusively exposed as private endpoints within that VNET.

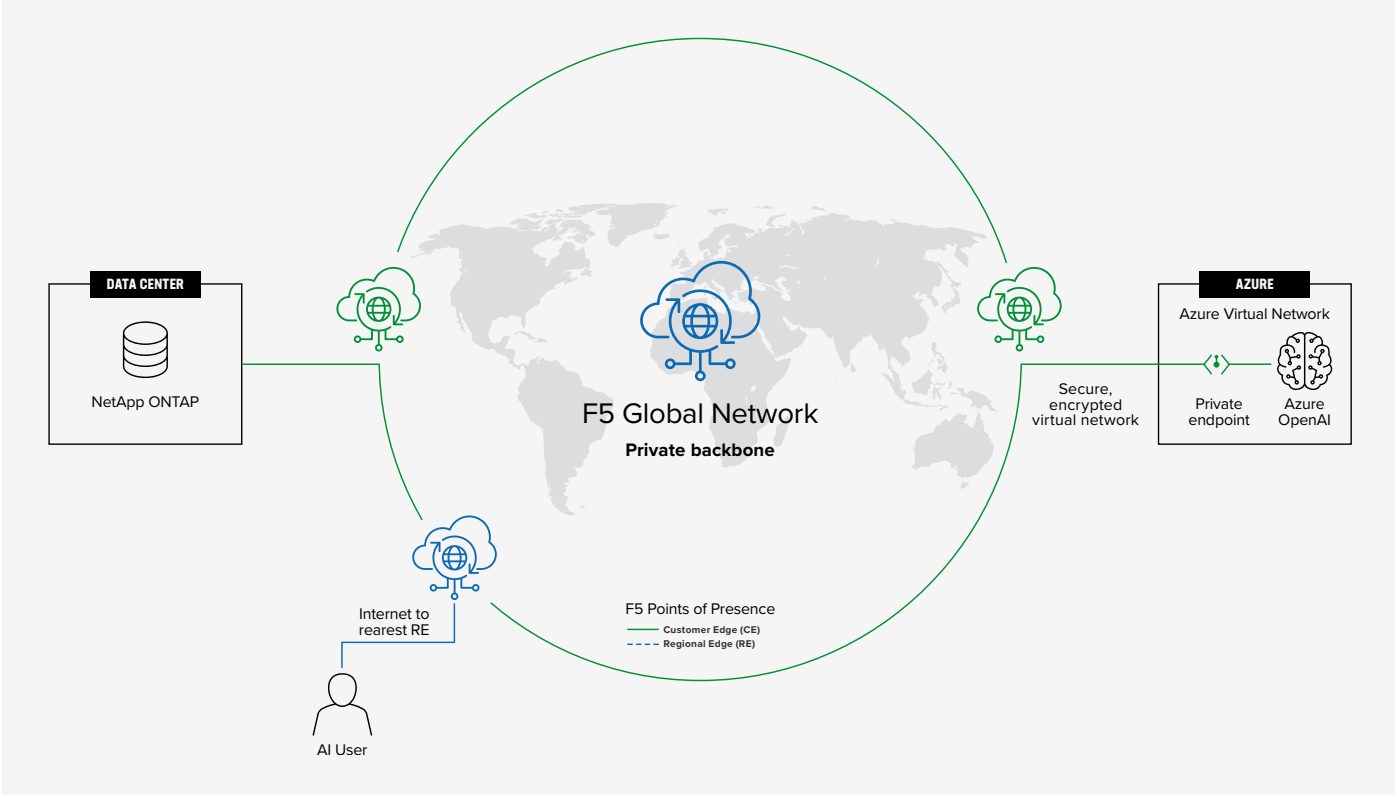


Figure 3: Secure connectivity between data-center hosted documents and Azure OpenAI.

See our tech article on DevCentral to learn more about this use case.

Learn more about F5 Distributed Cloud Services at f5.com/cloud.