FUTURIOM
FUTURE OF CLOUD TECH

Cloud Market Trend Report

RAGs to Riches: Using RAG and AI Data for Enterprise AI

July 2025

Sponsored by:

FUTURIOM.COM

Highlights and Key Findings

- **Inferencing is the heart of enterprise AI.** Enterprises will still train specialized models, but they can't reap the benefits of AI until they become experts at inference.
- **Retrieval-Augmented Generation (RAG) is a relevant way to infuse LLMs with additional data.** Thanks to larger LLM context windows, users can add quite a bit of information to a query. RAG, however, is less costly in terms of tokens and a better way to accommodate dynamic, real-time data.
- **Vector databases are having their moment.** By storing documents, images, and other data as vectors, enterprises can use RAG to perform multimedia searches. This has led major data players such as Oracle, Databricks, and Snowflake to incorporate vector support into their products.
- **Vector indexing and vector search are crucial database features.** Any database can store vectors. But with vectors numbering in the billions for individual enterprises, it's the ability to search instantly that gives vector databases their appeal.
- **RAG sprawl is an issue for early adopters.** This has led to the rise of RAG-as-a-service (RaaS) and turnkey RAG—cloud-based options that can abstract away the details of different RAG approaches.
- **Agentic AI will unlock more ambitious RAG and inference.** AI-driven agents can satisfy more complex queries that require multiple steps.
- **The Model Context Protocol (MCP) is accelerating the maturity of agentic AI**. It's an esoteric under-the-covers protocol, but developers have leapt onto MCP as a way to make AI handle sophisticated tasks.
- **Security is a major issue in all this.** That was true with RAG and goes doubly for MCP. The good news is that these issues are getting attention. Companies should strive to allay security concerns with the proper architecture and security tools.
- **Some companies highlighted in this report:** Aryaka, Amazon, Aviatrix, Chroma, Cisco, Cloudflare, Couchbase, Crunchy Data, Databricks, DDN, Dell, Deep Lake, Elastic, F5, Fortanix, Google, Hitachi Vantara, HPE, Infinidat, Juniper, LanceDB, Microsoft, MinIO, MongoDB, Neon, NetApp, Nuclia, NVIDIA, OpenSearch, Oracle, Pinecone, Pryon, Pure Storage, Qdrant, Ragie, Snowflake, TileDB, Timescale, VAST Data, Vectara, Veeam, Versa, Vultr, Weaviate, Weka, Yugabyte, and Zilliz.

Enterprise AI Delivery and Security

Solve AI deployment and security challenges to unlock emerging opportunities.

- Deliver data efficiently for AI training and inference
- Scale AI factories with optimized GPU utilization
- Secure AI deployments with end-to-end protection against app and model threats
- Connect data and applications to scale distributed AI deployments

Learn more about how we can help you with these challenges, visit f5.com/ai

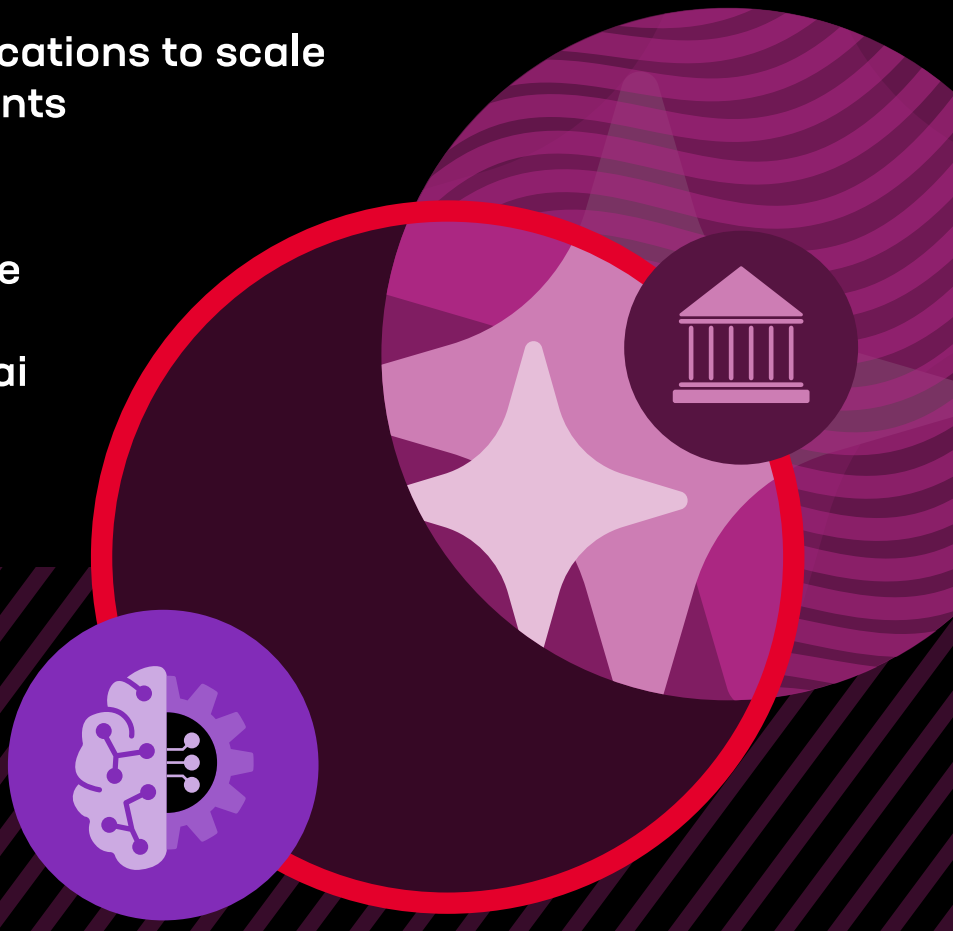

Table of Contents

1. Introduction: RAG Relevance in the Age of AI Inference

We are entering the age of AI inference, which, according to conventional wisdom, is destined to dwarf the size of the AI training market. NVIDIA's own estimates for AI's computing requirements grew roughly 100-fold in 12 months, as CEO Jensen Huang noted during his GTC 2025 keynote. That's thanks partly to the rise of agentic AI, expected to be a cornerstone of inference, he said.

Most inference does not require the massive datacenters that large language model (LLM) training does. Some will, but many if not most inference workloads will suitably run in enterprise facilities, on laptops, or even on embedded devices. That means fewer eye-popping statistics about millions of GPUs and gigawatt-sized facilities. But inference matters, because it's the way enterprises and employees wring value from AI and LLMs.

NVIDIA and AMD are both convinced that inference is a growing influence on the market. Both chipmakers re-asserted that idea during their early 2025 product launch events. Huang's 100X statement, although offered without hard-number context, is partly a reaction to the way reasoning models will iterate, conducting multiple attempts to derive an answer. AMD expects the AI inference market to grow at 80% per year "for the next few years," CEO Lisa Su said recently, although she offered neither market sizes nor a timeframe for context.

Retrieval-augmented generation (RAG) is a crucial element of inference, especially in generative AI (GenAI) use cases. It brings more information and context into consideration, adding to the knowledge already present in an LLM. RAG is in use today and will likely become more complex as AI agents become commonplace.

Visa, for example, has set up employees to use RAG routinely. They can query six different LLMs—popular foundation models such as ChatGPT and Claude—that sit behind Visa's firewall, accessed through one interface. Fraud detection is one obvious use case, but employees also use the LLMs with RAG to hunt for domain-specific knowledge. One use case highlighted by VentureBeat earlier this year was to query the LLMs about legal specifics in more than 200 countries, to verify that certain transactions are within the bounds of current policy.

This report summarizes how RAG and inferencing will grow in influence in the AI infrastructure market.

2. Key Enabling Technologies for RAG

In general terms, RAG is a framework to improve the performance of AI models, including large language models (LLMs) and small language models (SLMs), by integrating external data, making AI applications more accurate for specific use cases. This is not the same as pre-training a model by, say, teaching it the jargon specific to an industry. Training is slow and expensive, whereas RAG operates on the fly during inferencing, infusing a model with information relevant to the current query.

RAG is also useful for keeping up with changing data. An inherent problem with any trained model is that it was not trained right now. This doesn't always matter—such as when digging through years-old earnings reports—but for situations built around dynamic, real-time information, it's a shortcoming. Chances are your own interactions with LLMs have included RAG.

RAG will likely grow in importance, as more enterprises are finding that they need specific data applied to AI applications in order to ensure depth and accuracy. Instead of relying solely on their training data, RAG-based models use specific custom data to apply up-to-date knowledge to generate more accurate, relevant, and trustworthy responses.

Some of the key enabling technologies and trends that are influencing RAG include:

Inferencing approaches and architectures. RAG is a way to use the information and process of a model and augment that with external data through the inferencing process. Enterprises looking to build solid models must use the most accurate possible information, as well as build architectures to keep their data private and secure.

Interest in smaller models. Foundational AI models form the basis of RAG, enabling the AI applications to understand queries, process retrieved data, and generate responses. For specific industries or applications, LLMs can be stripped down and customized to SLMs, which can operate more economically.

Databases. External databases store information that is supplemental to the LLM's training data, providing current and specific knowledge. These systems manage and query large datasets.

Vector databases and vector search. This group of technologies includes vector databases for efficient vector storage and querying; embedding models to convert data into semantic vectors; and indexing techniques for fast retrieval.

Below we detail how some of these technologies apply to RAG as well as how it is evolving.

Vectors: The Math Behind RAG

Vectors are numerical representations of data. Unstructured data—pictures, videos, long documents—can be converted into vector form and planted into a database. This allows for multimodal search, making all types of data retrievable by a single RAG process. That's especially important for older enterprises, as much of their data exists in unstructured form.

The travel site TripAdvisor, for example, has amassed more than 1 billion user reviews and contributions and hundreds of millions of images. Travelers' behavioral data is strewn throughout the site. TripAdvisor used the Qdrant vector database to apply all that information toward a chatbot for building itineraries, finding it significantly increased the revenue-per-user compared with travelers using the old interface. The company's more ambitious goal, though, is to draw up user graphs based on those behavioral patterns, using that information to build a sophisticated recommendation engine that can be queried using natural language.

Vectors are complex beasts. The vectors you learned about in geometry class had two or three dimensions—they were arrows on a graph, remember? But vectors in an AI context can have more than 1,000 dimensions. You can drop vectors into any database you like. Doing something useful with them requires the ability to index and search those vectors, allowing you to identify the vectors that are similar to the information in the user's query.

The vectorization process goes like this:

- Chunking: An automated step that divides data into smaller pieces. This is due to the size of the LLM's context window, which is the amount of information that a model can handle at any given time. Think of the context window as a memory limit; it's a property of any machine learning model that uses the transformer architecture. Context windows have grown dramatically, reaching millions of tokens in size, but longer documents and data sources will still require chunking.
- Embedding: This is the step of converting data into vector form. It's performed by a separate AI model. Some foundational model vendors provide a spectrum of embedding models. Third-party models are available as well. For example, Anthropic, makers of the Claude LLM, refers users to external providers such as Voyage AI.
- Storage: Vectors get stored in a vector database. This can be a database built from the ground up for vectors, or a traditional database with vector search capabilities added. We'll be examining this more closely in the next section.
- Indexing: The step of preparing the vector database for searches. Ideally this happens in real time as new vectors get added to the database. Indexing lets an agent search the database

without having to search through every vector; without it, vector search, and therefore RAG, is not feasible at scale.

- Search: When RAG performs a search on that vector database, the "answer" it gets back consists of multiple vectors that are nearest neighbors to the query, as determined by well-established math operations.

This is a good point at which to remind ourselves of AI's probabilistic nature: RAG doesn't find *the* answer, but rather identifies a set of best guesses. Most search algorithms use methods such as approximate nearest neighbor (ANN), which save time over brute-forcing a search for the single best answer—but they can trade some accuracy for that performance.

Note that we skipped what might be called Step Zero: choosing the documents and data to be vectorized and made available for RAG. This can certainly be done by hand—a user feeding specific research reports into the model to enhance one specific query—but is more likely performed automatically, possibly by AI agents. MCP and its peer protocols, discussed toward the end of this report, will enable all this on a broader scale.

Long-Context Models

One reason for the "chunking" step above is that context windows used to be small—just 4,000 tokens for GPT-3.5 in 2022. That's changed rapidly. Google's Gemini Flash 2.5 Pro has a context window of 1 million tokens with 2 million coming "soon," as Google promised in April 2025 at Google Next. Meta's Llama 4 Scout, also released in April 2025, boasts a 10 million-token context window. Wider context windows will reduce the need for chunking.

Given that trend, it's tempting to proclaim that RAG is dead, since it will eventually be possible to include volumes of additional information in any query. Really, though, the goal should be in the opposite direction: Enterprises should be striving to make models and queries more streamlined, not only to control the resources AI requires but also to keep per-token charges in check. RAG is a way to keep a trove of relevant information continually accessible in digestible bites.

Databases vs. Vector Databases

Vector search has long existed as a concept. It's the advent of LLMs that made it cool.

In 2023, vector databases grabbed the spotlight as a key element of the AI ecosystem, thanks in part to a flurry of funding announcements for startups providing vector databases. Meanwhile, established vendors announced or enhanced their vector support as well; in the case of PostgreSQL databases, this is done via a well-known extension named pgvector.

Many of those announcements referenced AI training and gave only a small nod toward inference. Two years later, with inference dominating the enterprise AI discussion, vector databases are getting another wave of attention.

Vectors, being just numbers, can be stored in any database. What makes a vector database special is the capability to index those vectors for quick, thorough searching.

Vector databases can be divided into two camps: those that were built ground-up for vectors and "normal" databases that had vector search and indexing added. Interestingly, not everyone in the latter camp treated vector search as a monumental addition at first. For instance, Yugabyte issued a 2023 press release that included the debut of its vector capabilities, but the main focus was on easing lift-and-shift migration, particularly for migrations out of Oracle. As the prospects of RAG grew in the public consciousness, vector search got a stronger spotlight.

One of the arguments in favor of a vector-native database is that vector search and storage are growing more cumbersome. Vectors are growing into thousands of dimensions. With billions or eventually trillions of vectors to store, the storage and memory efficiency of a vector database can be compelling.

On the other hand, enterprises already use databases, and there's something to be said for sticking to what's familiar. This would seem especially true for longstanding Oracle customers.

Big Data Gets Modernized

A recent development is the nearly simultaneous entry of Databricks and Snowflake into the ranks of vector database providers. In the same week in June 2025, both companies announced new products. Both were based on announcements, made just weeks earlier, of agreements to acquire Postgres database startups—Neon (in the case of Databricks) and Crunchy Data (Snowflake).

For both companies, this was part of a modernization movement, acknowledging that databases can no longer be the fixed-sized monoliths that they've traditionally been. Rather, databases will need to be nimble creatures, spun up on the fly by AI agents, scaling elastically depending on the requirements of the moment and spun down when a task is completed.

Databricks introduced its vector support as part of Lakebase, a fully managed serverless database meant to sit atop a data lake. Snowflake's announcement emphasized its new database's benefits to developers, pledging a platform more enterprise-ready than off-the-shelf Postgres, including robust security and compliance controls.

VAST Data is not a traditional Databricks/Snowflake competitor but is worth noting here. The company is best known for storage products but has really been working toward a long-term vision of modernizing the handling of data. It's now branched out considerably to include things such as a vector database or the coming AgentEngine, a framework for running agentic agent pipelines. VAST sums all this up by calling itself an "operating system" for AI.

RAG-as-a-Service

Given that RAG involves multiple pieces, it only makes sense that a vendor could assemble those pieces ahead of time and present them as a service. The rationale is that while gathering those pieces could be easy—they're all available in public clouds—making them operational together is complex.

Some cloud providers offer RAG services that prepackage the necessary elements: Cloudflare offers AutoRAG for building fully managed RAG pipelines. Vultr offers what it calls Turnkey RAG, which is geared at getting enterprises started quickly with RAG.

A newer breed of startups are specialists in RAG-as-a-service (RaaS), including Elastic, Nuclia, Pryon, Ragie, and Vectara. While they can serve enterprises that are new to RAG, they also target large companies that have built multiple RAG implementations and want to simplify. Vectara, for example, describes scenarios where several groups within an enterprise each developed their own RAG models and methodologies; Vectara's role would be to work with the company to develop a unified RAG approach, giving the CIO/CTO a way to manage RAG and create consistent policies around its usage.

While enterprises have found some success building RAG workflows on their own, usually taking advantage of open-source components, the result has been RAG sprawl, Vectara CEO Amr Awadallah says. Vectara has customers who have amassed as many as 20 RAG processes. They might each work, but for the executive overseeing all 20, it's difficult to identify which methods are more prone to errors or security breaches. Moreover, LLM usage has a cost in terms of tokens, and those costs become more difficult to maintain when spread across multiple organizations.

More Complex RAG

RAG was initially conceived for a static environment, simply retrieving data from vector databases, or even just one vector database. What vendors and enterprises really envision, though, is the ability to craft multi-step AI tasks, where one query feeds into a follow-up query, or tasks that involve multiple agents coordinating on an answer.

A Google white paper from February 2025, titled “Agents Companion,” lists some of the benefits of agents. They can:

- Make multiple passes at a query to tighten results
- Execute multiple steps, possibly even handing off results of one query to other tools
- React to changes in a query, e.g., if new information calls for altering the course of a search
- Perform cross-checks to try to spot hallucinations

One hoped-for result is improved accuracy, but this is not guaranteed, as researchers from Bloomberg, the University of Maryland, and Johns Hopkins University showed in a study published in April 2025. They found that RAG can push popular LLMs toward giving unsafe answers (jumping safety guardrails by leaking private information, for example) even when fed perfectly safe documents.

Network and Security Implications

RAG depends on the network in order to connect models to data sources, including sources that could be siloed in public or private clouds or even SaaS services. All network providers tout features such as reliability and security, but the need for them is heightened in the case of RAG, particularly when we consider real-time interactions conducted by fleets of agents.

Networking vendors have integrated more and more security features into their products over the years. For those that emphasized multicloud connectivity as well, AI presents a watershed moment. Programmability and automation—not just for routing but also for policy and security—become more important as well. We discuss these and other infrastructure implications below.

3. MCP and the Agentic Age

Ambitions for highly automated AI processes run by agents will be best enabled if there is some standard way for agents and models to connect to data sources. APIs are not the full answer, because someone would need to code API calls. Developers envision a more touchless scenario, where agents independently navigate through databases and data stores.

Why MCP Is Accelerating Agentic Work

The open source MCP works like a universal port (it's often called the USB-C of AI). Going beyond the scope of APIs, MCP exposes underlying capabilities of a server. A database, for example, might provide an MCP server that tells agents (which in this scenario are MCP clients) how to navigate a GUI or even how to manipulate the database.

PayPal was among those who jumped on MCP quickly. At the MCP Developers Summit, a grass-roots gathering of 300 in San Francisco in May 2025, the company showed off some of the tools it makes available to work with Anthropic's Claude LLM, sending out agents to shop among many online catalogues based on a generic item descriptions. The capabilities are available as part of an MCP toolkit that PayPal offers to partners today.

In exposing a server's capabilities, MCP can also let an agent trigger actions. As Cloudflare's team explains in a blog post: "MCP is quickly becoming the common protocol that enables LLMs to go beyond inference and RAG, and take actions that require access beyond the AI application itself (like sending an email, deploying a code change, publishing blog posts, you name it)."

MCP was developed by Anthropic, which open-sourced it in November 2024. Momentum hit warp speed in the spring of 2025 as OpenAI and other major players embraced it. MCP's success isn't just the result of being open source, though. The protocol benefitted from having relatively contained goals, which made it relatively easy to build an MCP client or server.

Most MCP work is being done for internal use, rather than for customer-facing applications, but the impact is substantial. At the MCP Developers Summit, Angie Jones, Global VP, Developer Relations at Block, described how the company spread MCP to 15 job functions, many of them non-technical. It's inspired employee-built tools for use cases such as refining sales-lead searches in Databricks or examining long-term recurring issues hidden in PagerDuty data.

"This isn't just a side project for us anymore. It's not an indie experiment for Brad [Axen, a principal engineer]. Block is putting MCP into practice at scale."

— Angie Jones, Global VP, Developer Relations, Block

What MCP Needs Next

The MCP Developers Summit demonstrated the enthusiasm and community spirit that has helped MCP grow. But developers harbor no illusions about the work required to make agentic AI wander more freely about the web. Here are some of the major themes:

1. Security

MCP started as an attempt to solve a scaling issue. The initial specification did not address security.

Plenty of MCP exploits are out there. An MCP tool (the function that's pinging a data source to retrieve information) could establish trust with a data source but then change its behavior—or it might simply not do what it says on the tin, executing actions that go unnoticed. A tool could also persist too long, giving a bad actor an open doorway into sensitive data.

Anthropic already took the initial step of adding OAuth 2.1 to the specification and has developers working on furthering MCP's security in standardized ways. It will be up to the community, though, to fully flesh out security options for the protocol.

2. Evolved MCP

Enterprises and vendors have flooded social media with proclamations of MCP clients and servers they have written, but most of this has been for internal uses or even localized use, sticking to the confines of one's own laptop. The greater promise for MCP and RAG is to have agents navigating a wider world, tapping remote data services.

This introduces challenges MCP didn't originally support, such as asynchronous tasks or statefulness (sharing context and state with multiple agents and servers). Agents and servers will need ways to understand that a task is in progress and retain that information.

3. Additional Protocols

MCP emerged as just one of more than a dozen protocols related to agentic AI. Among them, Agent to Agent (A2A) is especially noteworthy because it came from Google and was announced after MCP had already caught fire. (The Linux Foundation took over hosting the A2A project in June 2025.) Moreover, agent-to-agent communications are a scenario MCP technically was not designed for.

Longer-term, however, it looks like AI agents, tools, and servers will all begin to blend, interacting bilaterally. As those lines blur, it is possible that MCP will start to absorb some of A2A's mission. Most of the MCP camp is being diplomatic, saying this isn't on the roadmap, but it seems to be a strong possibility.

4. What Does RAG Mean for Traditional Cloud Infrastructure?

AI and RAG are having an enormous impact on cloud and enterprise IT architectures. With at least $300 million in new infrastructure investment in AI-specific datacenters in 2025 alone (according to Futuriom and public sources), there is considerable debate about how these new AI architectures will interact with enterprise data and private enterprise IT. Data compliance, privacy, and security are key concerns, There is no doubt that this is causing consternation among CIOs, CTOs, and security architects as they are rapidly instructed by their bosses to “get out the AI quickly.”

In some greenfield deployments, AI infrastructure will be built with the most modern software architectures and components. However, in many enterprise environments, AI will need to use RAG to access existing data, including from legacy applications. Much thought and effort will have to go into how RAG is implemented in specific environments.

Below, we outline what the advancement of RAG and private enterprise AI means for existing and emerging technologies.

What RAG Means for Databases

As discussed above, the rise of AI and particularly of AI inference has made vector indexing and search into a key feature. While most database vendors added vector search after-the-fact, a crop of vector-native databases also emerged from startups and public clouds.

Vector-native	Chroma Cloudflare Vectorize Deep Lake Google Cloud's Vertex AI Vector Search LanceDB Milvus Pinecone Qdrant Weaviate
Vector capabilities added	Amazon Aurora PostgreSQL (part of Amazon Bedrock) Azure Cosmos DB Couchbase Crunchy Data (acquired by Snowflake) Elasticsearch MongoDB Atlas

	Neon (acquired by Databricks) OpenSearch Oracle TileDB Timescale VAST Data Yugabyte

What RAG Means for Storage

Those vector databases need to go somewhere. Storage vendors must ensure their platforms are scalable to accommodate sudden infusions of large vector databases. Performance is paramount, too; low latency and high throughput are necessary to avoid storage becoming a crucial bottleneck. Finally, storage will need robust access controls, as will the vector databases being stored; concerns here mirror the MCP security issues we noted above. Key vendors here include DDN, Dell, Hitachi Vantara, HPE, Infinidat, MinIO, NetApp, Pure Storage, VAST Data, Veeam, and Weka.

What RAG Means for Networking

There's certainly competition to supply connectivity within a GPU cluster, with the competitors including NVIDIA itself. RAG and inference open up broader possibilities, however, where models, data, and users are in disparate physical locations. "Fast, reliable connectivity" is a cliche, but most AI workloads push those requirements further, as RAG can involve chains of queries that rely on one another's output.

That said, GenAI may be a pivotal moment for the convergence of networking, security, and observability—an idea that has been in place for years but has new urgency given the speeds AI requires and the threats it opens up. The secure access service edge (SASE) combines networking and security, applying the latter at the edge so as to catch anomalies immediately.

Many of the SASE vendors, such as Aryaka, Cisco, and Versa Networks, have been enthusiastic about their potential role in securing data and networks for enterprise GenAI. Vendors that have focused on multicloud connectivity have a stake here as well, covering use cases where they can help maintain data sovereignty for data sources and AI models that live in different venues and geographies. Companies here also include Aviatrix, F5, and Juniper.

The speed of RAG and AI agents also presents new challenges for the network. As F5 notes, policy needs to be applied consistently but should also be programmable, to adjust to changing conditions or spontaneous requests from agents. Additionally, while networks can be

engineered to emphasize speed or cost, F5 also points out the need for context-aware routing that points traffic toward the appropriate domain expertise that a query requires. F5 is uniquely positioned here with a combination of secure networking products that can help carry secure, agentic AI traffic across networks at the application layer. This includes the F5 Application Delivery and Security Platform (ADSP), F5 BIG-IP Next for Kubernetes, and the F5 AI Gateway. Deployed in front of MCP servers, F5's BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs serves as an intelligent, high-throughput reverse proxy that secures API calls, validates requests, and enforces authentication and authorization policies.

Likewise, Aryaka recently made AI data security a key theme of the updated release of its SASE-as-a-service platform. The new winter release of the platform includes the addition of powerful AI-powered data observability and security services, as well as the capability to add worldwide dynamic points-of-presence (PoPs).

What RAG Means for Security

The implications of RAG on security are immense, as most enterprises consider their proprietary data the crown jewels of their AI efforts.

Because RAG involves tapping multiple data sources, it creates new questions around access. Is the LLM user permitted to see the data that's being retrieved? Is the arriving LLM or agent permitted to access the data at all?

Role-based access control (RBAC) is likely to become *de rigeur* for the RAG pipeline. Some startups took notice of this early; the Milvus and Qdrant vector databases both support RBAC.

Meanwhile, security companies of all stripes will look to add features and platforms to help secure AI data for RAG and inferencing. Some cloud providers, including Amazon, Google, and Microsoft, have emphasized new services to provide secure, private enclaves for RAG data and services. For example, Amazon uses Amazon Cognito to aid response generation with Amazon Bedrock, including guardrail interventions that help prevent policy violations in both inputs and outputs.

Some companies have launched entire platforms targeting private enterprise AI environments, including those using RAG. For example, Fortanix provides solutions for ensuring the privacy and security of RAG systems, particularly for GenAI applications.

Here we have only touched on some specific examples, because the impact of RAG on enterprise and cloud architectures will be far and wide.

5. Conclusion: RAG Is Emerging as a Key to More Accurate AI Inferencing

Inference is what will make AI truly beneficial to enterprises, and RAG is a crucial aspect of inference. Even as AI goes agentic, RAG will be happening behind the scenes; many of those agent interactions will involve the tasks associated with RAG.

This ties into the broader theme of data. In the age of AI, data has become more important than ever, and the tropes about data being the crown jewels of a business take on a deeper, more urgent meaning. The AI benefit that most businesses crave is the ability to wrest more value from the data they've always had. RAG is an economical way to do that, as well as an effective way to let AI tap real-time data.

Inference is about to become more complex with the rise of agents. We'll have agents tapping multiple data sources, including some outside a given organization, as well as agents talking to each other and taking complex actions, such as creating entire ephemeral databases. RAG is becoming more dynamic, and databases, data platforms, and storage vendors will all need to accommodate this world.

The MCP Developers Conference in May 2025 was a snapshot of developers' deep enthusiasm for the possibilities. MCP provides the joints that will make multi-staged AI work not only possible but also highly automated. RAG and especially MCP are still in their early stages, with most of the enterprise work so far targeting internal corporate uses. Developers are already dreaming of a more complex and ambitious world of agential inference, and the good news is that they're asking the right questions about security along the way.

AI Leader Profile

F5

Nasdaq: FFIV
Location: Seattle, Wash.

Backed by three decades of expertise, F5's Application Delivery and Security Platform (ADSP) was designed to deliver and secure every app, every API, anywhere: on-premises, in the cloud, at the edge, and across hybrid, multicloud environments. AI adds another dimension to that mission, as F5 can enable data mobility, giving enterprises a secure means to collect data from multiple silos and deliver it at the speeds required for AI processing.

https://www.f5.com/