GIGASPACES

RAG-Enhanced AI Strategies for Interacting with Enterprise Databases

Table of Contents

Overcoming the Hurdles of GenAI Adoption

Across all sectors, entities are exploring ways to integrate Generative AI (GenAI) into their operations, aiming to tap into new efficiencies and capabilities. However, much like with any emerging technology, adopting GenAI is not without significant challenges.

Key obstacles include the inability to consistently access reliable, accurate data that can be trusted for decision-making. This often results in the reliance on incorrect or fabricated information, commonly referred to as “hallucinations.”

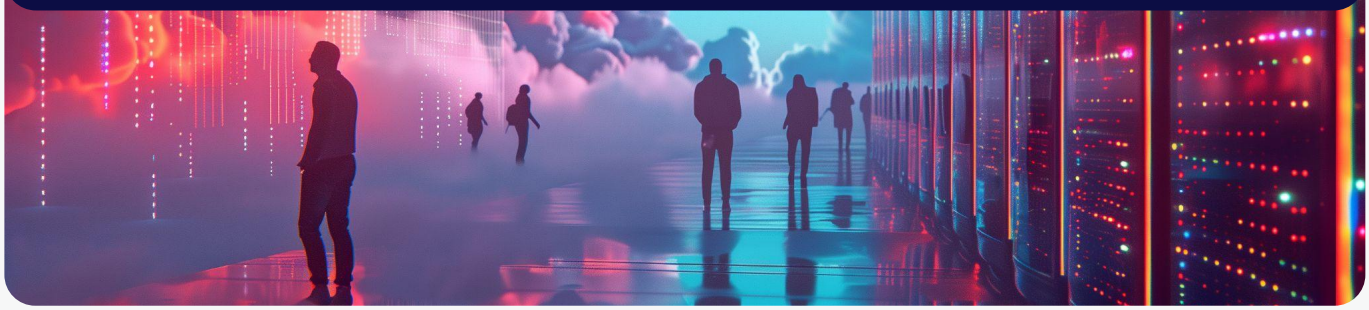
Another obstacle is the complexity and limitations of accessing data across different sources. Data is often siloed, making it difficult for teams to get the information they need without navigating through complicated systems. Lastly, many data systems are overly complicated and require specialized knowledge to be used effectively. This presents a challenge in ensuring teams can quickly adapt to and leverage data tools without a steep learning curve.

One promising solution to address these challenges is the Retrieval-Augmented Generation (RAG) model. This eBook will delve into RAG technology—examining its potential benefits and limitations—and explore its application in querying structured enterprise data—a critical pain point for many organizations today.

What is RAG? How Does it Work?

RAG stands for Retrieval-Augmented Generation, a framework combining traditional search engines' strengths with generative AI. In simple terms, it allows systems to retrieve information from external knowledge sources (such as databases or documents) and use that information to generate human-like, contextually relevant responses in natural language.

RAG blends the innovative capabilities of GenAI with advanced retrieval-based techniques



The RAG process typically involves two main steps:

Retrieval: A query is first used to retrieve relevant documents, data, or information from a knowledge base. This is the phase where the system seeks out relevant sources of information.

Augmentation and Generation: The retrieved information is then passed to a generative model, such as GPT, which augments it and generates a human-readable response. The generative model combines this information with its capabilities to provide a coherent, contextually appropriate output.

RAG systems are beneficial when handling unstructured data. For example, when querying many documents or text files, RAG can efficiently pull relevant snippets from these documents and synthesize a response that answers the query. However, this process doesn't inherently handle structured data similarly, which creates challenges.

The Challenges of Querying Structured Enterprise Data

While RAG excels at interacting with unstructured data—such as documents, emails, and multimedia—it falls short when querying structured data. Enterprise data is often stored in relational databases, spreadsheets, and other highly structured formats, where the relationships between tables, columns, and rows are critical to understanding the full context of a query.

The main challenges of querying structured enterprise data include:

- **Data Silos:** Enterprise systems often store data in disparate silos across different departments, applications, or regions, making it difficult to query and integrate data from multiple sources.
- **Complexity of Schema and Data Relationships:** The data stored in enterprise systems often follows complex schemas that can be difficult to navigate, especially for users without deep technical expertise.
- **Time-Consuming Querying:** Traditional database querying methods usually need experts to manually write complex SQL queries, which are arduous and prone to errors.
- **Data Governance and Compliance:** Enterprises are increasingly concerned with compliance, especially regarding regulations such as GDPR or HIPAA. Maintaining compliance by querying structured data adds another layer of complexity.

While RAG technology brings significant benefits when handling unstructured data, it doesn't completely address every issue associated with this. As entities strive to leverage AI in more data-intensive applications, solutions that bridge the gap between structured and unstructured data are crucial.

Solutions for Querying Structured Enterprise Data

A few specialized technologies and frameworks have emerged to address the challenges of querying structured data. These include Cache Augmented Generation (CAG), Table Augmented Generation (TAG), and Agentic RAG. Let's take a closer look at these solutions.

Cache Augmented Generation (CAG)

Cache-augmented generation (CAG) is a solution designed to improve the efficiency of RAG by incorporating a caching layer. This layer allows for retrieving commonly queried data from an in-memory cache rather than repeatedly querying the database. By doing so, CAG can significantly reduce latency and improve query performance, especially for high-demand data that is frequently accessed.

CAG addresses several key challenges:

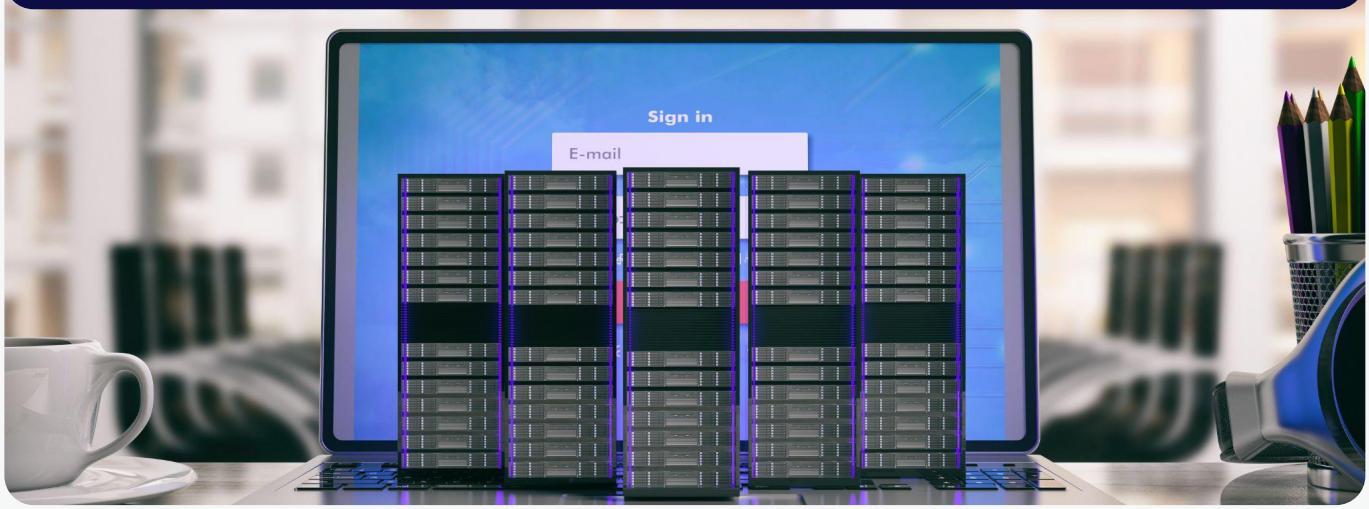
1. **Latency Reduction:** By using an in-memory cache, CAG can reduce the time it takes to retrieve frequently queried data, improving overall system responsiveness.
2. **Cost Efficiency:** Cutting the number of database queries helps lower the computational load and storage costs, making it a cost-effective solution for firms with high data requests.
3. **Scalability:** CAG enables systems to scale more efficiently, particularly in high-demand environments where data needs to be accessed rapidly.

While CAG dramatically boosts performance, it doesn't solve the more complex issues related to data understanding and retrieval from diverse sources, especially when working with structured data from different systems.

Table Augmented Generation (TAG)

Table-augmented generation (TAG) takes a more focused approach to querying structured data. It is designed to work specifically with relational databases, leveraging the power of RAG to navigate complex tables and schemas. TAG allows the system to retrieve data from tables, understand the relationships between columns and rows, and generate natural language responses based on structured data.

TAG leverages **SQL-based querying** to fetch rows and columns directly from relational tables, then augments these results with advanced AI insights—such as anomaly detection, trends, or even predictive forecasts.



TAG addresses several challenges related to structured data:

1. **Data Relationship Understanding:** TAG helps AI systems better understand the relationships between tables and columns, which leads to more accurate and relevant queries. For instance, it can help pinpoint which columns are foreign keys and how they relate to other tables.

2. **Data Integration Across Systems:** It can also be used to query data across multiple systems, which breaks down data silos and provides unified answers from varying enterprise databases.

3. **Ease of Use for Non-Experts:** With TAG, users do not need deep technical knowledge of SQL or database structures—they can query the system using natural language, which makes it accessible to a broader range of users.

While TAG improves the querying of structured data, it still has limitations in handling data from unstructured sources, and the challenge of integrating data across diverse systems remains complex.

Agentic RAG

Agentic RAG is an innovative framework to enhance the retrieval and generation of contextually relevant information. Unlike traditional RAG systems, which passively retrieve and generate information based on predefined queries, Agentic RAG incorporates autonomous decision-making into its architecture. This enables the system to proactively refine queries, adjust responses, and iteratively enhance results, creating a more dynamic and context-aware user experience.

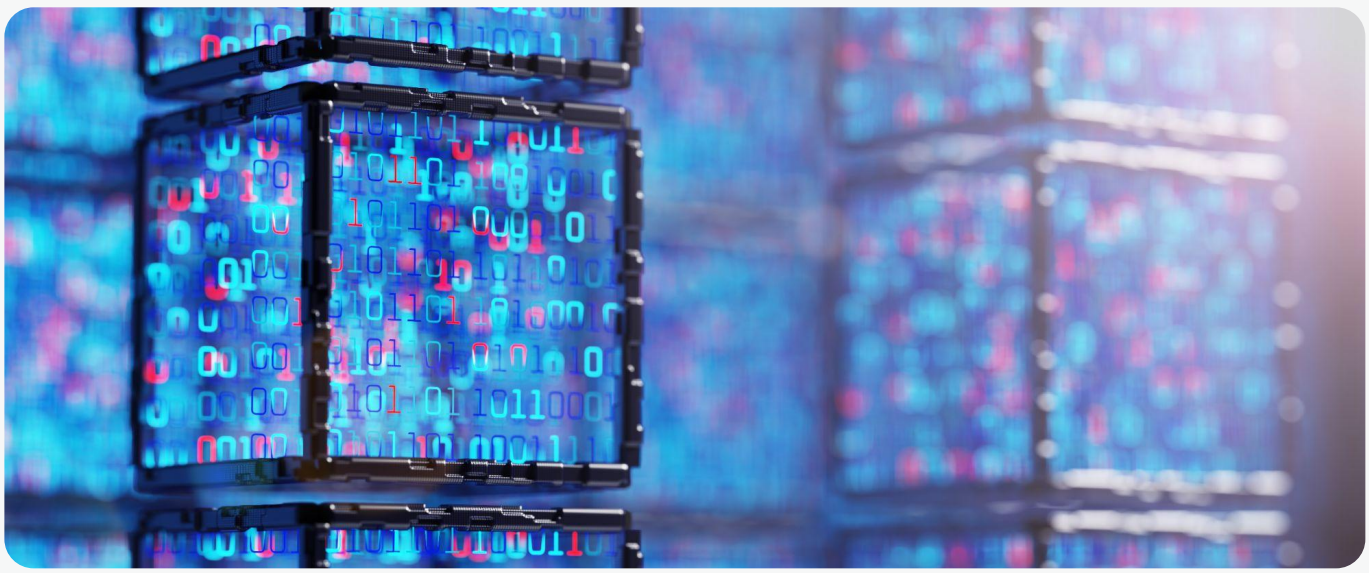
It is an advanced AI approach that moves RAG technology up a notch by adding agents to the process. These agents are AI-driven software components that help interpret the meaning of data, provide context for relationships between data points, and improve the data retrieval and generation process.

For instance, an agent could recognize that a service management system incident is related to a specific product defect recorded in another database. By bringing these agents together, Agentic RAG advanced reasoning can be carried out more efficiently, generating more accurate insights.

Agentic RAG addresses several challenges faced by organizations that work with structured data:

1. **Data Contextualization:** The agents can interpret the real meaning behind tables, columns, and relationships, adding context that traditional query methods lack.
2. **Complex Query Optimization:** Agentic RAG optimizes retrieval by intelligently identifying the most relevant data sources and relationships, improving accuracy and speed.
3. **Multi-Source Data Integration:** These agents can also integrate data from various systems and databases, making it easier to query across the enterprise and derive meaningful insights from disparate sources.

Despite these improvements, agentic RAG faces challenges, mainly when dealing with highly complex or ambiguous queries involving structured and unstructured data.



Multi-Agent RAG

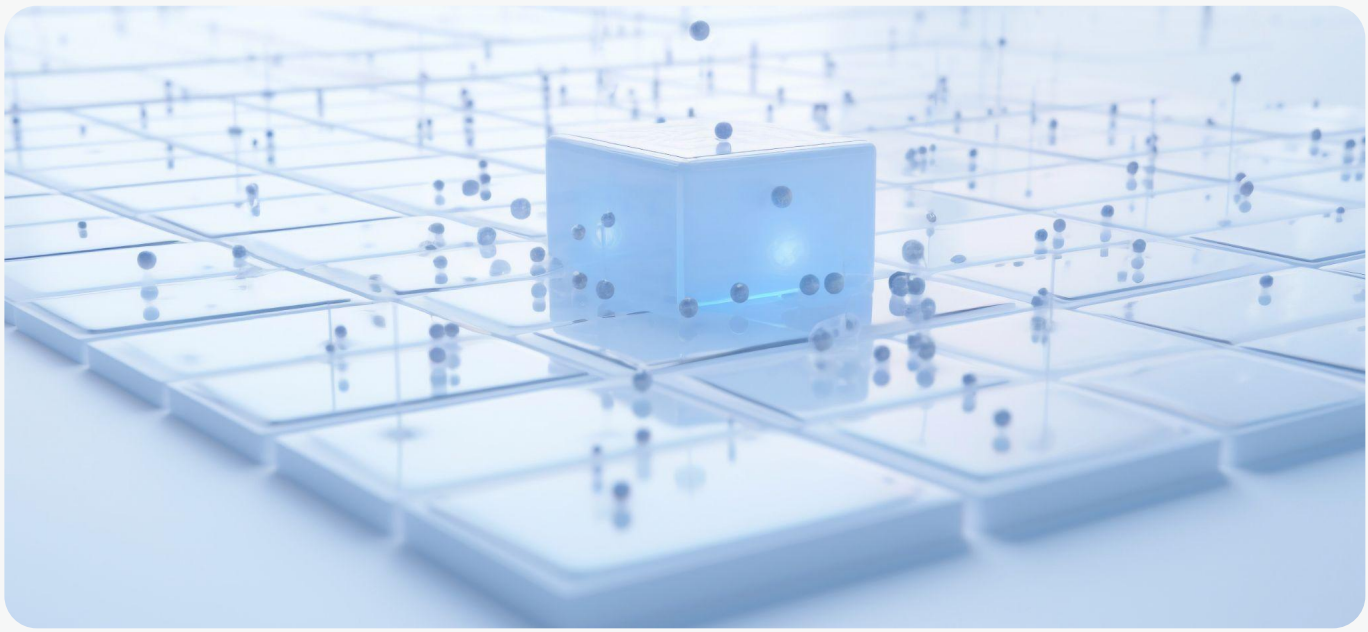
Multi-agent RAG enhances AI capabilities by using multiple agents to refine information retrieval and generation. Instead of a single agent handling both tasks, a retrieval agent sources relevant data, a generative agent synthesizes responses, and a manager agent coordinates their efforts based on user input.

Unlike traditional RAG systems, which rely on a single agent, multi-agent RAG distributes tasks among specialized agents—such as retrieval, filtering, and language generation—resulting in more accurate, efficient, and context-aware outputs. This approach is particularly valuable in complex scenarios requiring dynamic interaction with diverse data sources and large language models (LLMs).

A well-designed multi-agent RAG system includes:

- **Large Language Model (LLM):** The core processor for natural language interpretation and response generation.
- **Specialized Agents:** Each agent is assigned a specific task, optimizing efficiency and accuracy.
- **Orchestrator:** Coordinates agent interactions, ensuring seamless collaboration and workflow optimization.
- **Knowledge Bases:** Vector or graph databases store retrievable data, enabling precise, context-aware results.

By leveraging multiple agents, this system improves adaptability and performance, making it ideal for handling evolving datasets and complex AI applications.

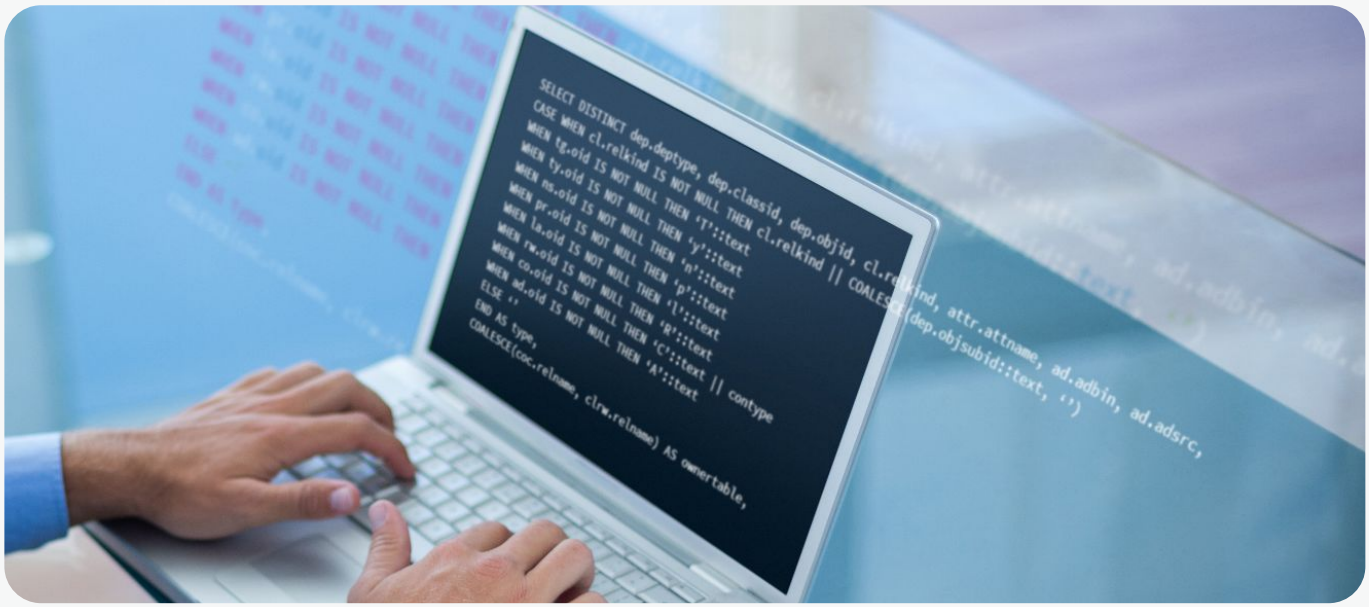


Bridging the Structured Data - LLM Gap with eRAG

A key challenge in querying structured data is deciphering the true meaning behind database tables and columns. In many instances, table and column names are cryptic, and the application logic governing the relationships between data elements is not always reflected in the database schema. Conventional querying methods often fail to capture this complexity, making it difficult for users to extract meaningful insights efficiently.

eRAG: A multi-agent RAG approach

GigaSpaces eRAG addresses the challenge of querying structured enterprise data by leveraging multi-agent RAG, enhancing natural language interaction with structured data, making complex enterprise databases accessible and understandable.



Leveraging a multi-agent RAG approach

Information you can trust

eRAG delivers reliable, accurate answers—with no hallucinations. The framework is built for enterprise data, enabling LLMs to query structured data with precision, while ensuring privacy, security, and reduced LLM consumption costs.

Accessible to all

eRAG empowers users to interact with structured data in natural language, unlocking enterprise data without technical expertise. Cross-data source integration ensures relevant information is available on demand.

Intuitive and easy to use

The framework's semantic reasoning engine understands the contextual relationships between tables, columns, and underlying data, making structured data queries as natural as asking a question.

With such an approach businesses can interact with structured data effortlessly, transforming raw data into trustworthy information that result in actionable insights—accurately, securely, and instantly.

eRAG is built on GigaSpaces DNA in enterprise data. It is built on:

- Deep semantic reasoning of operational databases
- Auto-didact learning of organizational lingo and terminology
- Continuous learning of your data context

eRAG quickly connects to multiple data sources – including Oracle, MSSQL, BigQuery, Amazon RDS, SAP and others – to unlock the hidden value of structured data across your business.

To better understand how eRAG works and the ROI, let's examine the following case study.

Case Study

Incident Management at a Utilities Asset Management Company

Client
A Utilities Asset Management Company managing over 100 sites across Europe.

Industry
Renewable and traditional energy

Business Challenge
The company struggled with inefficient incident management, particularly understanding the financial impact of service disruptions. The process of querying disparate systems took days, risking costly penalties due to SLA violations.

The GigaSpaces Solution
The utilities management company partnered with GigaSpaces to implement eRAG, enabling account managers to query multiple sites and systems in real-time, using natural language. The solution aggregates data, providing fast, ad-hoc answers to queries that pinpoint high-risk incidents and their financial impact.

Implementation

By integrating AI agents capable of generating SQL queries across multiple data sources, the solution now automatically retrieves and combines all critical necessary information. eRAG spots critical incidents and ranks them, based on their potential financial impact so that account managers can focus on the most urgent issues.

Results and Benefits

The implementation led to a **27%** reduction in penalties due to SLA violations. Incident analysis time dropped from days to minutes, enabling quicker decision-making to minimize financial losses.

Conclusion

Although RAG technology offers substantial benefits for unstructured data, querying structured enterprise data remains a challenge. Solutions like GigaSpaces' eRAG bridge this gap by augmenting language models with domain-specific information, improving efficiency and making data accessible for informed decision-making and getting everyday tasks done with the right information at hand.

As enterprises continue to explore GenAI's potential, innovations like eRAG will fuel faster, more accurate interactions with enterprise databases - and transform how organizations interact with their data.

About GigaSpaces

For over two decades, GigaSpaces has mastered the art of real-time data.

We've built platforms that power the world's most demanding systems, shaping how organizations grow their business with data-driven services. We have pioneered technologies that optimize data-driven services.

We understand every facet of operational data and are trusted by global organizations to deliver mission critical apps.

As pioneers in data-tech, we are building on this foundation to deliver cutting-edge GenAI solutions that empower businesses to unlock the full potential of their structured data, and transform how they interact with information.

www.gigaspaces.com