{ }

Vespa Blog

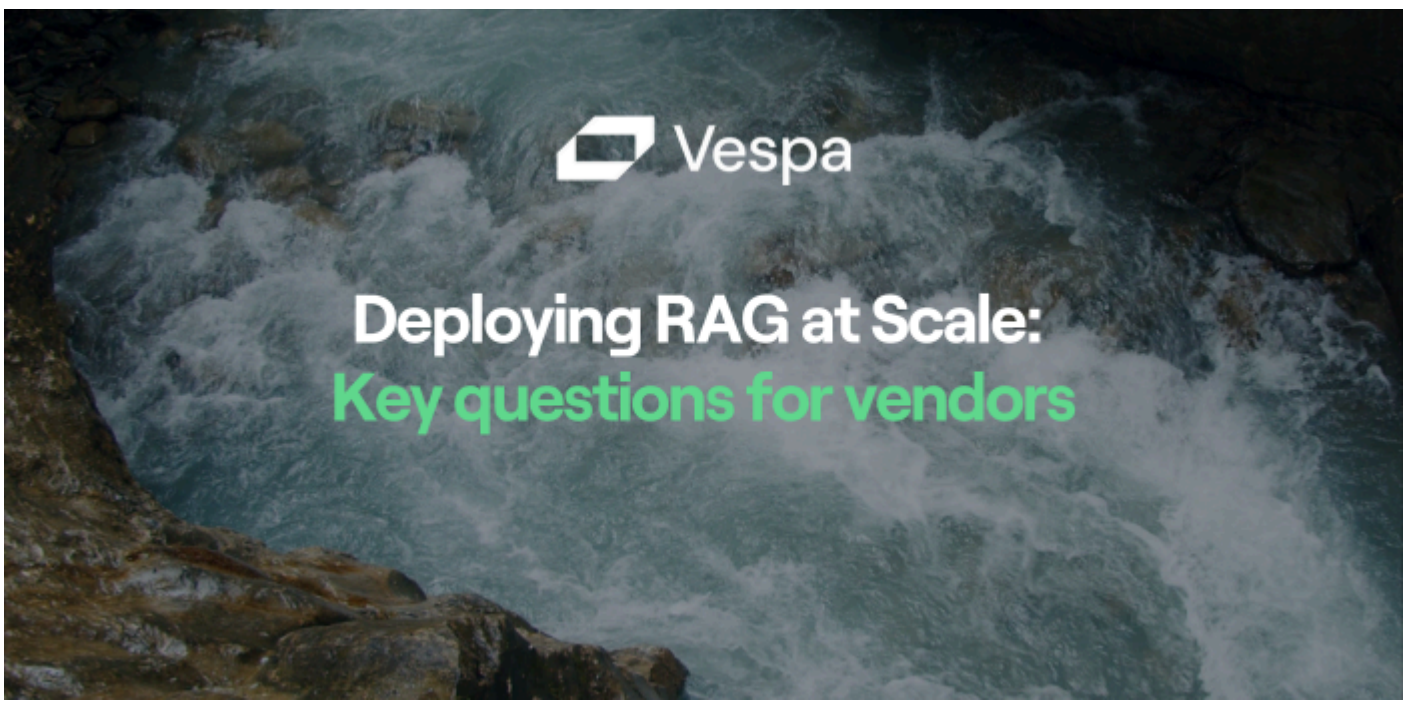
We Make AI Work

SUBSCRIBE

Tim Young
Chief Marketing Officer

28 Oct 2024

Deploying RAG at Scale: Key Questions for Vendors



Retrieval-augmented generation (RAG) has emerged as a vital technology for organizations embracing generative AI. By connecting large language models (LLMs) to corporate data in a controlled and secure manner, RAG enables AI to be deployed in specific business use cases, such as enhancing customer service through conversational AI. For those new to RAG, I recommend this BARC research note: Why and How Retrieval-Augmented Generation Improves GenAI Outcomes, available for free here.

In its recent Hype Cycle for Generative AI, Gartner identifies RAG as an early-stage technology driving innovation. However, it is approaching a peak of inflated expectations as ambitions for RAG outpace the practicalities of deploying it at scale. Vendor exuberance likely raises the bar on expectations—hype around RAG and generative AI is at a fever pitch!

Our discussions with large enterprises reveal that while generative AI pilots are proving value, scaling these solutions across the enterprise is a concern. Managers ask:

- How can we scale from concept to enterprise-wide deployment?
- Given the intensive processing demands of generative AI, how can I control costs?
- How can I ensure compliance with data privacy and security regulations?
- How can I integrate all relevant data sources beyond just vector databases?
- How can I stay current with emerging technologies and best practices?

Before answering these questions, let's introduce Vespa:

Vespa is a robust platform for developing real-time, search-based AI applications. Its large-scale distributed architecture enables efficient data processing, inference, and logic management, making it ideal for applications handling vast datasets and high volumes of concurrent queries.

From Concept to Enterprise Deployment

Proving the value of RAG in the lab is one thing, but scaling it across an entire enterprise introduces numerous challenges. These include integrating with existing data sources, ensuring strict data privacy and security, delivering required performance, and managing this complex large-scale run-time environment. Scalability is also a significant concern, as AI models must handle vast amounts of growing data and increasingly diverse use cases while maintaining high performance and reliability.

Vespa has been wrestling with these challenges since 2011—long before AI hit the mainstream. Originally developed to address Yahoo's large-scale requirements, Vespa runs 150 applications integral to the company's operations. These applications deliver personalized content across Yahoo in real-time and manage targeted advertisements within one of the world's largest ad exchanges. Collectively, these applications serve an impressive user base of nearly one billion individuals, processing 800,000 queries per second.

Vespa offers two essential components for enterprise RAG deployment:

- a comprehensive platform for developing generative AI applications
- a scalable deployment architecture to address the demands of large enterprises.

The Platform Approach

Vespa is a fully integrated platform that offers all the essential components needed to build robust AI applications. It includes a versatile vector database, hybrid search capabilities, RAG, natural language processing (NLP), machine learning, and LLM support. The platform connects easily with existing operational systems and databases through APIs and SDKs, enabling AI applications to support your specific requirements. This allows organizations to integrate existing data infrastructure easily.

Vespa's hybrid search capabilities enhance the accuracy of generative AI by combining various data types, including vectors, text, and both structured and unstructured information. Machine learning algorithms score and rank results to align with user intent, delivering precise and relevant answers. A key feature of the platform is its advanced natural language processing, which enables efficient semantic search. By understanding the meaning behind user queries rather than just matching keywords, Vespa supports vector search with embeddings and integrates custom or pre-trained machine learning models for more precise content retrieval.

Visual search is a hot topic, and Vespa offers intelligent document retrieval that combines images and text to enable detailed contextual searches. This creates a visually intuitive search experience that feels more natural and human-like.

An Execution Environment for Large Scale Enterprise Deployment

Vespa Cloud streamlines large-scale deployment, delivering high performance but simplifying performance management to ensure a seamless user experience. Applications running on Vespa dynamically adjust to fluctuating workloads, optimizing performance and cost—eliminating over-provisioning to keep costs in check and users happy.

Designed for high performance at scale, Vespa's distributed architecture ensures instant query processing and advanced data management. It offers low-latency query execution, real-time data updates, and sophisticated ranking algorithms, enabling enterprises to efficiently process and utilize data across their operations without sacrificing speed or accuracy.

The platform's robust, always-on architecture guarantees uninterrupted service. By distributing data, queries, and machine learning models across multiple nodes, Vespa achieves high availability and fault tolerance, ensuring continuous operation even under demanding conditions.

Security and compliance are core elements of Vespa's design. With computation performed close to the data and distributed across nodes, the platform reduces network bandwidth costs and minimizes latency. It adheres to data residency and security policies, with encryption at rest and secure, authenticated internal communications between nodes. This comprehensive approach provides a secure and governed environment for deploying AI applications at scale.

Future Proofing

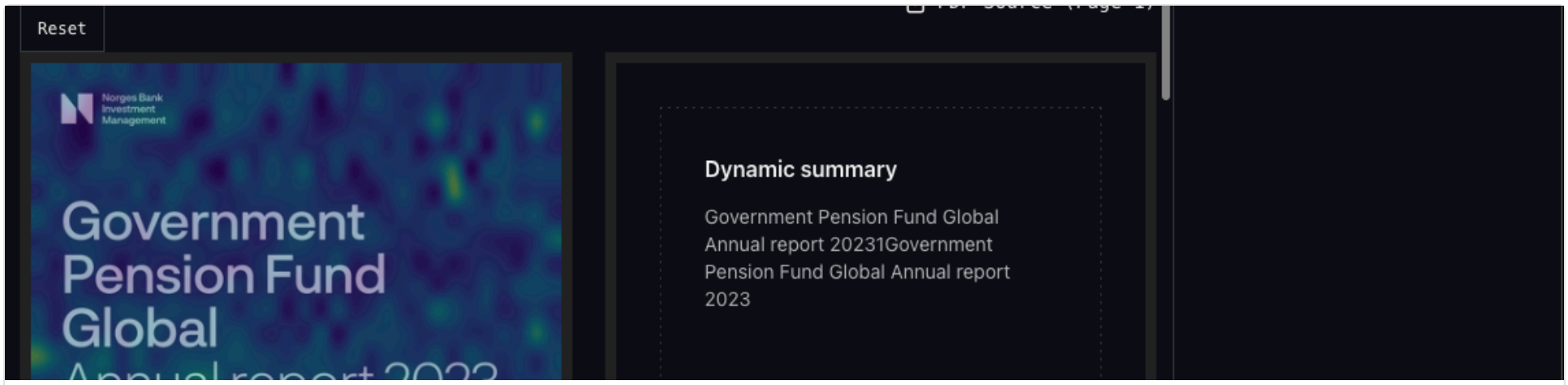
RAG deployments naturally evolve as experience and confidence grow and use case requirements become more sophisticated. What begins as a basic Q&A system for tasks like customer service chatbots can scale into dynamic, live knowledge bases that are rapidly consulted hundreds or even thousands of times, drawing on conclusions reached by the AI. Vespa supports this stepwise development, allowing for controlled, incremental rollouts that adapt to evolving needs.

Future-proofing also involves adopting the latest technologies and best practices. For example, Vespa enables visual search capabilities in eCommerce, where searches are driven by images, and supports standards like ColPali for large-scale PDF search. With Vespa Cloud, our engineers continually integrate emerging RAG best practices, ensuring your enterprise stays ahead. We incorporate this best practice so that you can focus on your AI applications.

Summary

RAG is crucial for businesses adopting generative AI. However, scaling it across an enterprise presents challenges, including integration, data privacy, infrastructure management, and performance at scale. Vespa addresses these challenges with a comprehensive platform and scalable deployment architecture. Proven by Yahoo's large-scale needs, Vespa Cloud supports AI applications with real-time, low-latency query processing, hybrid search, and advanced data processing.

Read more



Visual RAG over PDFs with Vespa - A demo application in Python

This is a technical blog post on developing an end-to-end Visual RAG application powered by Vespa. It has link to a live demo application, and will walk you through why...

Vespa

Barc Research report

Why and How Retrieval-Augmented Generation Improves GenAI Outcomes

Vespa

« Announcing support for global significance models
Vespa.ai: The “Sleeping Giant” Powering Next-Gen Search and Recommendations »