Vespa.ai
Executive Overview

Vespa

Vespa Executive Overview

The Race to Operationalize GenAI and RAG

The surge in generative AI (GenAI) adoption is already showcasing tangible benefits. Companies that have embraced GenAI technologies report significant advantages: 58% have seen an enhancement in customer experiences, 53% have noted improved operational efficiency, 52% have witnessed advanced product features, and 47% have enjoyed cost reductions. It's no wonder that 44% of organizations regard GenAI as crucial to their strategic planning[1].

While GenAI is powerful, it can make false statements and has not been trained on your internal data. Retrieval Augmented Generation (RAG) enhances GenAI models by providing access to internal, private and up to date contextual information. For example, in healthcare, RAG combines GenAI with the retrieval of personal health data.

However, enterprise data search and retrieval needs are not limited to RAG. For example, online shopping websites need to match products to customers' imprecise search term wordings, content providers need to recommend personalized content in real time to visiting consumers, and finance and legal analysts need to be able to search and summaries vest quantities of data with high precision.

Integrating these advanced technologies has its challenges. Organizations seeking to use large amounts of text data and metadata are worried about privacy, accuracy, reliability, security, and scalability.

Nearly half of AI experts identify infrastructure as the primary obstacle in deploying GenAI systems effectively[1].

[1] ML Insider Results 2023

Vectors Elicit the Meaning of Your Data

Vector embeddings are increasingly crucial for generative AI, recommendation systems, and search functions. These vectors provide a numerical representation of unstructured data, like text or images, facilitating semantic search over the data.

When you search a vector database, the search phrase is converted to a vector. Then a vector search, a nearest neighbor search, finds the data with the vectors that lie closest to that search vector, which gives you search at the semantic level. While this is very useful, vector searches alone are rarely sufficient; they return imprecise results at high cost, and are problematic to scale. If you want to surface the most valuable information to your users, you need more.

Andrew Brust, Analyst, GigaOm:
“Interest in the vector database market is surging as industries seek advanced AI to manage vast unstructured data[2].”

Best Practices in Data Discovery

How can you reliably find the data you need? In real-life use cases, vectors must be combined with structured data and text search to deliver quality results. For example, in legal research, it is important to combine semantic searches on a topic with metadata filtering on geographic regions and text search for precise legal terminology. This combination of search methods will reliably outperform any individual search method.

Vespa is a platform engineered for developing, deploying, and operating enterprise-scale data and AI applications. It supports precise hybrid search, ranking and inference that combines data types, including multiple vectors, text search, and structured metadata such as labels and numbers.

Data discovery accuracy depends upon machine learning models which optimally combine information from multiple vectors, text matching and metadata to rank candidates and select the most useful information. Since these models must be executed over many candidate data items per query, they must be executed locally within the engine to meet the low latency and high scale demands of online use cases.

Jungwon Byun, COO & Cofounder, Elicit:
"Vespa is a battle-tested platform that allows us to integrate keyword and vector search seamlessly. It forms a key part of our AI research solution, guaranteeing precision and rapidity in streamlining research processes. We highly recommend Vespa for its reliability and efficiency."

[2] Gigaom Report 2024

Enterprise Ready

These data discovery capabilities cannot be achieved by combining separate components solving separate subproblems. Effective integration and automated orchestration are essential; otherwise, bandwidth and computational requirements won’t allow you to scale. Vespa uniquely enables the integration of vectors, unstructured text, and tabular data with machine learning, delivering results at any scale with latency suitable for online applications. It is an integrated platform designed to optimize computing and data handling, allowing it to scale seamlessly with increases in traffic and data volume.

Vespa Cloud has time-proven enterprise scalability, managing over 100 billion documents and processing up to 800 queries per second while maintaining low latency.

Running scalable highly available applications sustainably and with high quality is difficult and expensive; let the world's leading experts do that for you. Vespa Cloud is a managed SaaS solution that offers serverless operations for straightforward application management and deployment, supported by world-class Vespa experts. This developer-centric platform eases operational tasks, allowing teams to concentrate on innovation rather than infrastructure management. The continuous deployment capabilities facilitate efficient updates and iterations, promoting a culture of rapid development and innovation. Vespa Cloud is available on major cloud providers, including AWS, GCP, and (from mid 2024) Azure.

By switching to Vespa Cloud, Yahoo reduced its server count by over 10,000 and significantly reduced labor costs.

Case Study: Singapore Government

In Singapore, the Hansard, an exhaustive archive of parliamentary debates since 1955, is a vital policy-making resource for government officials, citizens, and technological systems. Despite its value, reliance on keyword-based searches led to suboptimal results, highlighting documents based on the frequency of keyword appearance rather than meaningful relevance to the search query.

The Pair Search system has employed Vespa to index all the parliament debates since 1955, and has demonstrated significant enhancements in the quality of search results. Pair Search employs a hybrid search approach, using multiple vectors and full text to achieve state-of-the-art quality. This type of search precision is unique to Vespa.

The Singapore Prime Minister has acknowledged the Pair Search tool’s utility in parliament, highlighting various agencies' broader interest in adopting this technology for their data.

Prime Minister Lee Hsien Loong:
"The Hansard is an open book. We can all refer to it readily, and soon, we will be able to do a generative AI search on it."

Conclusion

Whether it be for GenAI, recommendation, or search, vector embeddings are becoming an essential tool. However, in real-life use cases, vectors need to be combined with structured data and text search to deliver quality results.

Vespa is the only platform that lets you combine vectors, unstructured text and tabular data with in-database machine learning to deliver results at any scale with latency suitable for online use cases. It's known for its powerful technology, ability to scale, and built-in AI features. Vespa Cloud takes things further by making it easier to manage and secure applications, offering a range of deployment options and the ability to customize security measures.

Vespa is pleased to be recognized as a Leader and Forward Mover in the latest GigaOm Sonar Report for Vector Databases. Download a courtesy copy of this report at: Gigaom Report 2024.

Get started with Vespa Cloud today. To learn more about how Vespa is the data discovery platform solution for your business, contact us at info@vespa.ai for an obligation-free demo.