Vespa.ai

Modernizing Enterprise Search

Replacing Elasticsearch with Vespa For RAG and Vector Search

We Make AI Work

Vespa.ai

We are a platform for building and running real-time AI-driven applications for search, recommendation, personalization, and RAG. We enable enterprise-wide AI deployment by efficiently managing data, inference, and logic, handling large data volumes and over 100K queries per second. Vespa supports precise hybrid search across vectors, text, and structured metadata.

Document
Modernizing Enterprise Search - Replacing Elasticsearch with Vespa

03.06.2025

Contents

↗ Introduction 3

Modernize enterprise search by migrating from Elasticsearch to Vespa.ai

Read

↗ Vespa vs Elasticsearch 5

Real-time availability, fine-grained scalability, and efficient data handling provide significant business advantages.

Read

↗ Vespa Key Use Cases 7

two areas gaining significant attention are Retrieval-Augmented Generation (RAG) and vector databases.

Read

↗ A Performance Benchmark 9

Vespa Engineering conducted a complete benchmark.

Read

Introduction

Over the past decade, Elasticsearch has evolved from the Apache Lucene project into a widely adopted, distributed search and analytics engine. Today, it powers thousands of large enterprises, supporting use cases beyond full-text search, including real-time analytics, logging, security, and observability. However, as data volumes grow and AI-driven applications demand lower latency and greater scalability, Elasticsearch is showing its limitations. Organizations face challenges such as scaling complex configurations, ongoing performance tuning, and service disruptions during writes—issues that add operational burden and cost.

Benchmarks and real-world deployments demonstrate that Vespa offers significantly higher efficiency, handling 10x more queries per CPU core, sustaining 8.5x higher query loads, and applying real-time updates 4x faster than Elasticsearch. Companies like Vinted report faster indexing, reduced infrastructure needs, and up to 5x cost savings, making Vespa a powerful alternative for scalable, cost-effective search. Additionally, Vespa is designed for modern AI-driven applications and the latest customer engagement techniques.

This Manager's Guide outlines the key considerations for organizations looking to modernize enterprise search by migrating from Elasticsearch to Vespa.ai—a faster, more cost-effective solution to support the latest AI-driven search and recommendation techniques. By leveraging modern best practices, Vespa provides a scalable and future-ready foundation. This guide presents:

- Competitive advantages of Vespa over Elasticsearch
- A summary of key Vespa strengths in RAG and vector database
- Results from real-world benchmarking

Additional resources
Benchmark Summary:
Elasticsearch VS Vespa
Performance Comparison

“The migration [to Vespa] was a roaring success. We managed to cut the number of servers we use in half (down to 60). The consistency of search results has improved since we’re now using just one deployment (or cluster, in Vespa terms) to handle all traffic. Search latency has improved by 2.5x and indexing latency by 3x. The time it takes for a change to be visible in search has dropped from 300 seconds (Elasticsearch refresh interval) to just 5 seconds. Our search traffic is stable, the query load is deterministic, and we’re ready to scale even further.”

Vinted Engineering Blog

Vinted Search Scaling Chapter 8: Goodbye Elasticsearch, Hello Vespa Search Engine

Ernestas Poškus
Senior Search Engineer

Vespa vs Elasticsearch

Vespa's real-time availability, fine-grained scalability, and efficient data handling provide significant business advantages over Elasticsearch. Immediate data availability ensures that critical updates—such as pricing changes, inventory updates, or new content—are instantly reflected in search results. This enables better user experiences, more responsive applications, and faster decision-making. Unlike Elasticsearch, which introduces delays due to batch processing, Vespa ensures that every update is searchable the moment it is ingested, making it ideal for dynamic environments like e-commerce, financial services, and social platforms.

Vespa's granular data distribution model optimizes resource utilization, preventing performance bottlenecks and ensuring smooth scaling as data and traffic grow. Its support for individual updates and mixed mutable-immutable data structures further enhances efficiency by reducing operational overhead and eliminating unnecessary rewrites. This allows businesses to maintain real-time responsiveness at scale while lowering infrastructure costs. By seamlessly handling both structured and unstructured data with advanced search and ranking capabilities, Vespa provides a powerful foundation for AI-driven applications that demand speed, accuracy, and scalability.

Support for tensor frameworks demonstrates Vespa's superior customer engagement capability. Consider a large online marketplace that tailors product recommendations to each visitor. As users browse, their interactions—such as clicks or cart additions—should instantly refine the recommendations they see.

With Vespa's built-in tensor capabilities, user preferences and product attributes can be stored as vectors, enabling real-time similarity calculations and dynamic updates. Because these computations are processed directly within Vespa, recommendations adjust instantly, delivering a seamless, highly relevant shopping experience without delays from batch updates or external processing. In contrast, Elasticsearch requires handling recommendation logic outside the system, adding complexity and slowing response times. Vespa's ability to integrate AI-driven personalization at scale makes it a powerful choice for businesses that prioritize real-time engagement and customer satisfaction.

Vespa Strengths vs Elasticsearch

Real-time Availability

With Elasticsearch data doesn’t become searchable until the next refresh operation, which occurs asynchronously. While this allows Elasticsearch to process writes in batches, it introduces a delay between writes and visibility that can impact real-time applications. With Vespa, you get immediate availability. Data is searchable immediately after ingestion, ensuring real-time updates. Any updates, like new product listings or pricing changes, are instantly visible to users.

Scalability

Both systems scale horizontally but differ in their underlying models. Elasticsearch scales by spreading indices and shards across nodes, making data movement efficient. However, this approach can also create uneven load distribution, hot spots, and potential bottlenecks at scale. Vespa distributes data into granular buckets, offering fine-tuned load balancing and redundancy. This granular distribution allows for more efficient resource utilization and better handling of uneven workloads, making it highly scalable for large, dynamic datasets.

Individual Updates

With Elasticsearch, data enters the system through batch operations using the Bulk API. This is efficient for handling large volumes but delays the visibility of changes. Vespa supports individual data updates, making changes instantly searchable. This is perfect for dynamic environments like marketplaces or social platforms where up-to-the-minute information is essential.

Mixed Mutable and Immutable Data Structures

Elasticsearch segments are immutable, meaning every update requires writing a new document and deleting the old one, even for something as simple as incrementing a counter. Retrieving the latest version of the document adds further overhead, as it requires either looking up the transaction log or forcing a refresh, which are both costly operations. Vespa's approach offers significant advantages by utilizing a mix of mutable and immutable data structures. Updates in Vespa scale better, as it can update the counter directly within a mutable structure, avoiding expensive reads and rewrites and enabling immediate updates to be searchable.

Native Support for Tensors

Vespa has native support for tensors, allowing you to store, process, and search tensor data directly within the system. This makes it well suited for real-time personalization, recommendation engines, advanced ranking and other AI-driven applications, where calculations happen on the fly without needing an external service. Elasticsearch, on the other hand, doesn’t have built-in tensor support in the same way. With Elasticsearch, you would need to preprocess data or use external tools for complex computations, which can add extra steps and slow things down.

Vespa Key Use Cases

Vespa For RAG

As a platform for building and running search-driven AI applications, Vespa supports a wide range of use cases. However, two areas gaining significant attention are Retrieval-Augmented Generation (RAG) and vector databases. Vespa's strengths in these domains are outlined below.

The effectiveness of a RAG application depends on its ability to surface the most relevant data for a given task. Achieving this requires more than basic vector similarity—it demands advanced retrieval techniques such as hybrid search, multi-signal processing, and machine learning-driven ranking. Vespa provides a comprehensive solution, seamlessly integrating these methods to enhance accuracy and relevance. Its robust distributed architecture ensures effortless scaling, enabling organizations to handle large data volumes and high traffic efficiently and reliably.

Unmatched Speed and Scalability
Vespa delivers real-time, high-performance RAG at any scale. Vespa's distributed architecture enables seamless horizontal and vertical scaling, unlike other platforms that slow down under growing workloads. It automatically adjusts capacity to optimize resource use and keep costs low, ensuring lightning-fast responses even as data and user demands expand.

Superior Accuracy with Hybrid Search
RAG needs more than just vector search—it requires a powerful hybrid approach. Vespa integrates multiple retrieval techniques, including vector, lexical, and hybrid search, all enhanced by machine learning. Its advanced indexing and ranking systems ensure precise, relevant results for every query, helping businesses extract actionable insights with confidence.

AI Flexibility and Enterprise Security
Vespa supports any ML model, offering seamless retrieval, ranking, and inference capabilities while giving developers full control over computations. Security is built-in, with encryption, compliance with industry standards like GDPR and HIPAA, and robust protections against unauthorized access. Whether handling sensitive data or scaling billion-scale PDF applications with ColPali, Vespa ensures efficiency, accuracy, and security in every RAG deployment.

Vespa Vector Database

Additional resources
GigaOm Sonar Report for Vector Databases (2025)

Vespa has offered an open-source vector database since 2014. Designed for real-time, online use cases, Vespa handles hundreds of thousands of queries per second while seamlessly integrating vectors, tensors, unstructured text, and structured data. Its comprehensive indexing and retrieval capabilities enable users to query across diverse data types, leveraging advanced ML models for precise relevance ranking. With built-in support for lexical and hybrid search, as well as group aggregation for deriving insights from multiple signals, Vespa delivers a powerful and flexible foundation for AI-powered search.

Billion Scale Vector Search
Vespa delivers low-latency, high-throughput search by colocating data, indices, metadata, and ML inference on the same physical machine, even in distributed environments. It ensures search relevance with powerful ML models, including ONNX-based inference and Gradient Boosting Decision Trees (GBDT), making it a top choice for AI-driven applications.

Advanced, Flexible Indexing
Unlike traditional vector databases, Vespa integrates real-time vector indexing with a customized HNSW implementation while supporting true full-text search, going beyond simple bag-of-words to perform matching and ranking using positional text data. Structured data is efficiently indexed using B-Trees and hashes, while disk-based retrieval minimizes in-memory costs, providing a cost-effective and scalable solution.

Optimized Storage and Retrieval
Vespa supports binary vector indexing and quantization, enabling organizations to store compact vectors efficiently while retrieving larger vectors for ranking when needed. With built-in embedding models and the flexibility to vectorize content externally, Vespa enhances search relevance through metadata filtering, delivering superior accuracy and control over results.

Native Support for Tensors
Vespa has native support for tensors, meaning you can store, process, and search tensor data directly within the system. This makes it great for real-time personalization, recommendation engines, advanced ranking, and other AI-driven applications, as calculations happen locally without needing an external service.

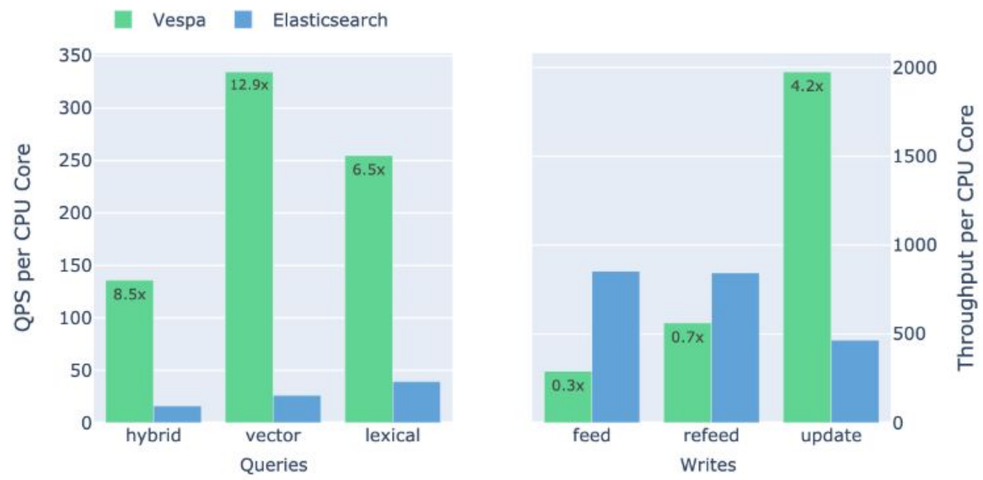
A Performance Benchmark

Enterprises should benchmark software products before purchasing to ensure they meet performance, scalability, and cost expectations in real-world conditions. Benchmarking helps validate vendor claims, identify potential bottlenecks, and assess suitability for existing infrastructure. However, this approach requires time—typically weeks to months—depending on the complexity of the evaluation. It also demands dedicated resources, including technical staff to design and run tests, hardware or cloud resources for execution, and a method for fair comparisons.

To streamline Elasticsearch vs Vespa benchmarking, Vespa Engineering conducted a complete benchmark which is fully documented in this 80-page report. A shorter 6-page Executive Summary is available here.

This benchmark presents a reproducible and comprehensive performance comparison between Vespa (8.427.7) and Elasticsearch (8.15.2) for an e-commerce search application using a dataset of 1 million products. The benchmark evaluates both write operations (document ingestion and updates) and query performance across different search strategies: lexical matching, vector similarity, and hybrid approaches. All query types are configured to return equivalent results, ensuring a fair, apples-to-apples comparison.

Vespa demonstrates significant query performance, scalability, and update efficiency advantages.

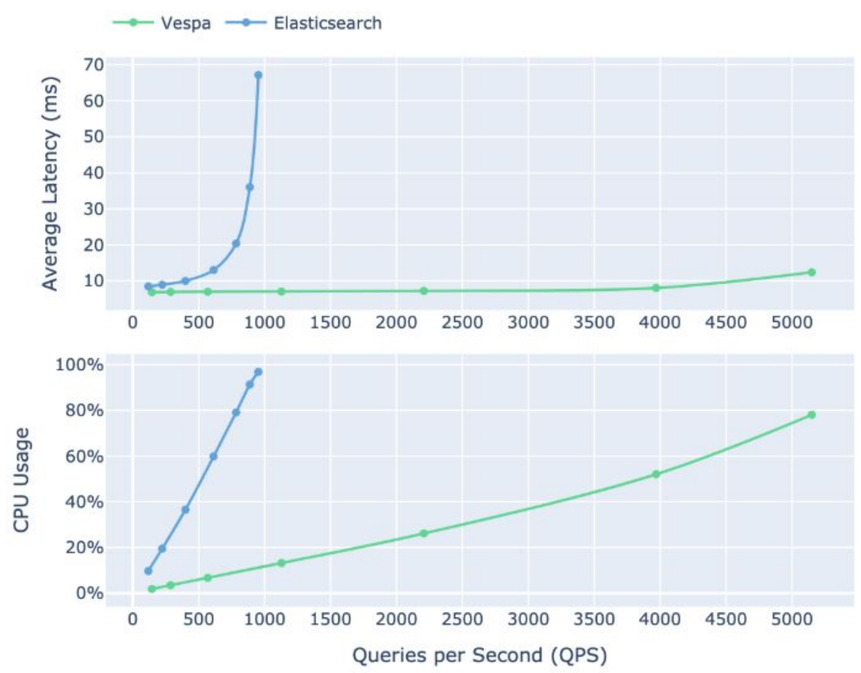


Query Efficiency Advantages

The illustration shows queries per second per CPU core for different query types and reveals Vespa's significant query efficiency advantages:

- **Hybrid Queries:** Vespa achieves 8.5x higher throughput per CPU core.
- **Vector Searches**: Vespa demonstrates 12.9x higher throughput per CPU core.
- **Lexical Searches:** Vespa yields 6.5x better throughput per CPU core.
- **Updates**: Vespa is 4x more efficient for in-place updates. While Elasticsearch showed higher efficiency during the initial write (bootstrap from 0 to 1M) phase, Vespa excels in steady-state operations, handling queries and updates more efficiently.

The following illustration compares how both systems handle hybrid queries, showing the relationship between latency, CPU usage, and query throughput as user concurrency increases.



Vespa shows higher CPU efficiency, demonstrated by a lower CPU usage gradient compared to Elasticsearch. This superior performance efficiency directly reduces infrastructure costs, as demonstrated in section 10 of the report, where the efficiency improvements yield up to 5x reduction in infrastructure costs.

About Vespa.ai

Vespa.ai is a platform for building and running real-time AI-driven applications for search, recommendation, personalization, and RAG. It enables enterprise-wide AI deployment by efficiently managing data, inference, and logic, handling large data volumes and over 100K queries per second. Vespa supports precise hybrid search across vectors, text, and structured metadata. Available as both a managed service and open source, it's trusted by organizations like Spotify, Vinted, Wix, and Yahoo. The platform offers robust APIs, SDKs for integration, comprehensive monitoring metrics, and customizable features for optimized performance.

Interested to learn more? We have many different resources and information available through our social platforms

GitHub Twitter LinkedIn YouTube

Vespa.ai