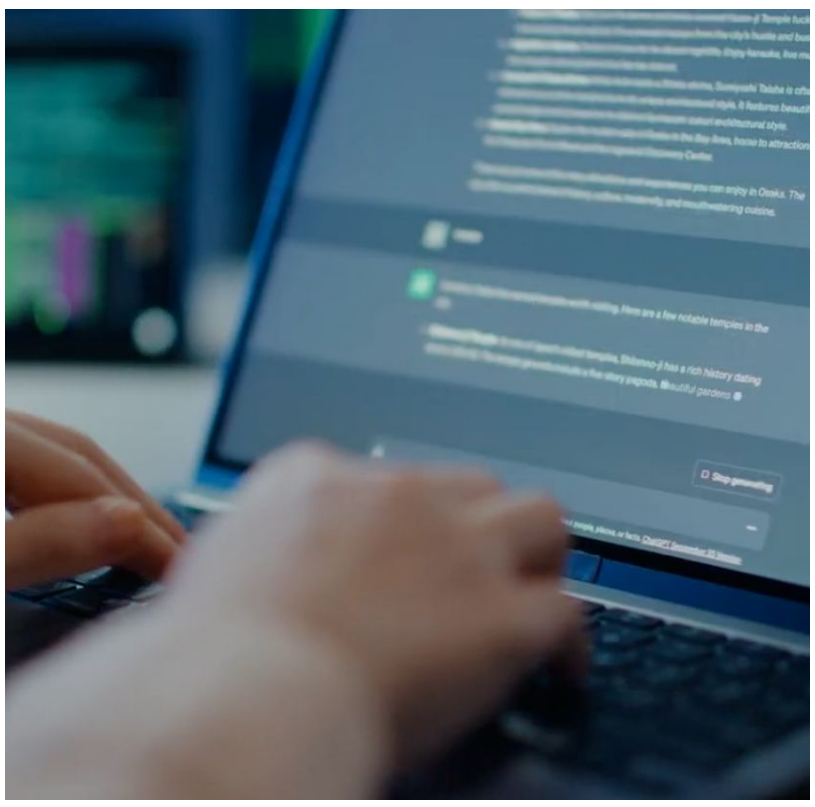
Vespa.ai

# Powering AI Applications with Real-Time Retrieval on AWS

## AI Search for Customer-Facing Use Cases

Developers building large-scale, robust search, RAG, and recommendation systems face a key challenge: retrieving and operationalizing data in near real time. Information is scattered across different sources and stored in multiple formats, including PDFs, free text, and semi-structured data, making it difficult to unify, index, and serve efficiently to AI models. Without the right infrastructure, applications become slow, brittle, and costly to scale; for many teams, it's not the models that hold them back but the retrieval workflow itself. Vespa removes that barrier, unifying structured, unstructured, and vector data in a single platform and delivering results at millisecond latency. Running on AWS, Vespa provides the scalability, reliability, and automation needed to power real-time search and retrieval applications that enhance customer experiences and drive measurable business growth.



## Benefits

### Performance and Efficiency

Reduce latency and network overhead with co-located data and computation for fast, resource-efficient retrieval at any scale.

### Relevance and Accuracy

Deliver contextual, high-quality results using hybrid retrieval and distributed ML ranking across structured, unstructured, and vector data.

### Elastic Scalability

Scale clusters up or down in real time with simple configuration changes—maintaining low latency and uptime at any scale.

## Accelerating AI Search Innovation Through the AWS ISV Accelerate Program

Vespa.ai is a member of the AWS ISV Accelerate Program, a co-sell initiative connecting leading software providers with AWS to help customers adopt modern, cloud-native solutions. This partnership reflects Vespa's alignment with AWS infrastructure and commitment to delivering scalable, high-performance AI search. Through the program, Vespa collaborates with AWS to optimize deployments, adopt new technologies such as Graviton processors, and support customers via AWS Marketplace—enabling faster deployment and time to value.

Vespa.ai

# Why Deploy Vespa on AWS

Running Vespa Cloud on AWS eliminates the operational overhead of managing your own infrastructure. Provisioning, scaling, upgrades, monitoring, and security are fully automated, freeing developers to focus on building applications instead of maintaining systems. Continuous deployment pipelines, built-in security, and 24/7 system oversight deliver the reliability and performance needed for production AI search.

With Vespa Cloud on AWS, you retain the flexibility and control of a self-hosted setup but gain managed scalability, proactive tuning, autoscaling, and fault recovery supported directly by the Vespa engineering team. The result is faster iteration, lower operational risk, and a production-ready environment for large-scale RAG, search, and recommendation workloads.

## Resources

**Vespa Overview**

The AI Search Platform behind Perplexity, Spotify, and Yahoo. Vespa.ai unifies search, personalization, and recommendations with the accuracy and performance needed for generative AI at scale.

**Vespa Cloud Subscription in AWS Marketplace**

Whether you just need to quickly deploy some experiments, or run an always available world-wide production system handling thousands of requests per second, the Vespa Cloud will fit your needs.

## About Vespa.ai

Vespa is an AI search platform purpose-built to power RAG, search, and personalization applications that demand speed and scale. It unifies vectors, tensors, full text, and structured data in one system and performs ranking and inference directly within the engine. Trusted in production by organizations including Perplexity, Spotify, Yahoo, Otto, and OkCupid, Vespa delivers on the four pillars that matter most to developers: performance, accuracy, scalability, and flexibility.

Learn more at Vespa.ai

Vespa.ai Norway AS, Prinsens Gate 49,
Trondheim, 7011, Norway




aws partner network | isv accelerate