Vespa

Vespa: Scalable Infrastructure for High-Performance, Large-Scale AI Applications

About Vespa

Vespa is a platform for building real-time, production-scale search and GenAI applications. It goes beyond simple retrieval by unifying keyword and vector search, structured filtering, and model inference—all in a single, high-performance engine.

Vespa simplifies your architecture without compromising on speed or scalability. With built-in support for deploying and running machine learning models during query execution, Vespa lets you operationalize AI faster and more efficiently.

Whether you're building copilots, conversational interfaces, or domain-specific RAG systems, Vespa provides the speed, flexibility, and relevance AI demands—without the need to stitch together multiple systems.

Industries We Serve

- ✓ Technology
- ✓ Retail
- ✓ Media
- ✓ Healthcare
- ✓ Finance
- ✓ Manufacturing

Key Application Areas

GenAI Agents, Chatbots

Vespa grounds LLM responses with accurate, real-time data from your own content. Use it to build fast, domain-specific copilots and assistants that combine search, structured logic, and inline inference.

Enterprise Search

Unlock all of your organization's data—from documents and PDFs to images and videos. Vespa supports secure, multi-source search with advanced filtering, access control, and personalized ranking.

Application Search

Power rich, embedded search experiences in web and mobile applications. Vespa supports instant updates, typo-tolerant search, and customizable scoring logic—all built to scale.

Product Search

Drive conversion with smarter, faster product discovery. Vespa lets you blend text, attributes, and behavior into personalized search and merchandising strategies that update in real time.

Recommendations

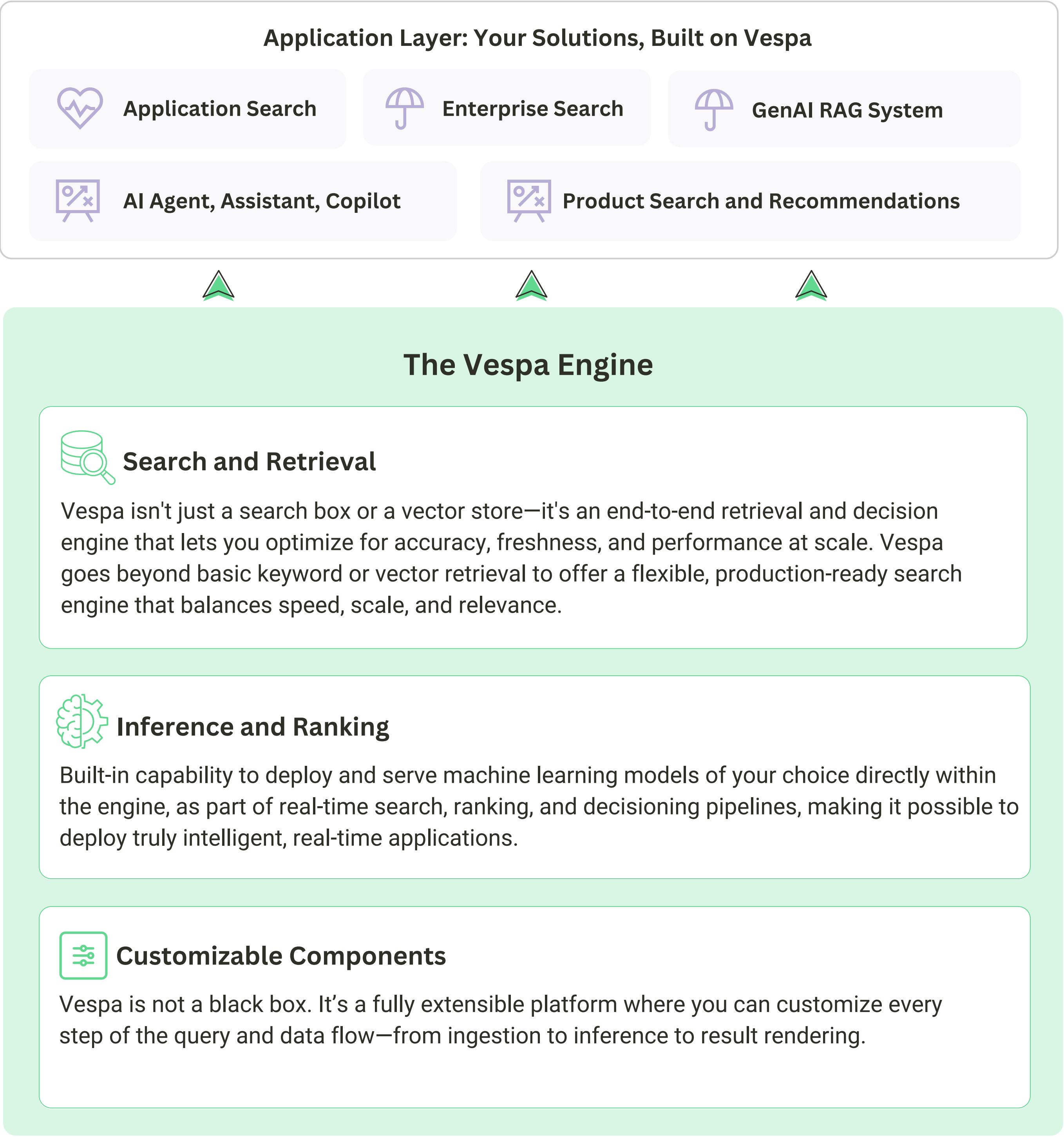
Serve real-time recommendations by combining collaborative signals, content metadata, and model-based scoring. Vespa supports both retrieval-based and reranking pipelines for highly relevant outputs.

vespa.ai/free-trial

Vespa

Overview of the Vespa Engine

Vespa is the scalable infrastructure behind high-performance, real-time AI applications at companies like Spotify, Perplexity, Vinted, and Yahoo. This document outlines the core components of the Vespa engine and how they work together to power large-scale search, recommendation, and GenAI experiences.



Application Layer: Your Solutions, Built on Vespa

- Application Search
- Enterprise Search
- GenAI RAG System
- AI Agent, Assistant, Copilot
- Product Search and Recommendations

The Vespa Engine

Search and Retrieval

Vespa isn't just a search box or a vector store—it's an end-to-end retrieval and decision engine that lets you optimize for accuracy, freshness, and performance at scale. Vespa goes beyond basic keyword or vector retrieval to offer a flexible, production-ready search engine that balances speed, scale, and relevance.

Inference and Ranking

Built-in capability to deploy and serve machine learning models of your choice directly within the engine, as part of real-time search, ranking, and decisioning pipelines, making it possible to deploy truly intelligent, real-time applications.

Customizable Components

Vespa is not a black box. It's a fully extensible platform where you can customize every step of the query and data flow—from ingestion to inference to result rendering.

Overview of the Vespa Engine:
Search and Retrieval

Vespa isn't just a search box or a vector store—it's an end-to-end retrieval and decision engine that lets you optimize for accuracy, freshness, and performance at scale. Vespa goes beyond basic keyword or vector retrieval to offer a flexible, production-ready search engine that balances speed, scale, and relevance.

Hybrid Search and Retrieval

- Use dense and sparse embeddings together for better relevance
- Combine semantic, lexical, and structured signals with precision
- Query across multiple data sources and content types, with built-in unified result logic.

Real-time Ingestion and Updates

- Index and query fresh data in milliseconds
- No need for offline indexing or batching

First-Class Vector Support with Native Tensor Framework

- Store and index dense embeddings (text, image, etc.)
- Perform ANN search with HNSW for low-latency retrieval
- Combine vector, keyword, and structured filters in one query
- Vectors are stored and manipulated as tensors, not black-box blobs
- Define custom ranking expressions using tensor math (e.g., dot product, cosine similarity, max pooling, etc.)
- Support for multi-dimensional tensors (not just 1D vectors) for advanced use cases like image and video search

Multi-Phase Ranking Pipelines

- Retrieve fast, then re-rank with precision—supporting BM25, dense vectors, ML models, and custom expressions.
- Filter and rank using metadata, numeric ranges, categories, and geolocation
- Inline inference and scoring enables you to apply models like ONNX or embedding transforms directly within ranking phase
- Execute tensor-based ranking logic with full transparency and control

Flexible Result Rendering

- Customize how results are formatted and delivered to meet product or application needs.

Overview of the Vespa Engine: AI Inference and ranking

The Vespa AI Model Hub is Vespa's built-in capability to deploy and serve machine learning models directly within the engine, as part of real-time search, ranking, and decisioning pipelines. Vespa lets you run ML models of your choice directly in the decision loop, making it possible to deploy truly intelligent, real-time applications.

ML Model Deployment

- Upload and register models (e.g., ONNX, XGBoost) directly in Vespa
- Models are versioned, managed, and executed inline during query and document processing

Real-Time Model Execution

- Evaluate models during serving time, not as a pre-processing step
- Supports stateless inference—ideal for low-latency, high-throughput applications

LLM-Ready Infrastructure

Power GenAI use cases like conversational search, tool-augmented agents, reranking, and grounded question answering—without bolting on extra layers.

Embed on the Fly

Generate dense or sparse embeddings at query time using integrated embedders—no preprocessing or third-party services needed.

End-to-End Ranking with ML

Deploy multi-stage pipelines that combine vector similarity, lexical scores, and learned models—all in a single request.

Custom Scoring Logic

Mix classic IR, vector relevance, metadata, and model predictions with fully controllable ranking expressions.

Inline Tensor Computation

Perform real-time feature calculations, transformations, and scoring using Vespa’s native tensor math engine.

Overview of the Vespa Engine: Customizable Components

Vespa isn’t a black box—it’s a fully extensible engine that lets you inject custom logic at every stage of the search, retrieval, and inference pipeline. Whether you're building a commerce search engine, a GenAI assistant, or a real-time moderation system, Vespa adapts to your application—not the other way around.

Custom Searchers

Write your own query processing, filtering, and reranking logic—tailored to your domain.

Request Handlers

Define custom HTTP APIs to serve complex user experiences or multi-tenant logic.

Processors

Add logic mid-stream in the data or query lifecycle—ideal for filtering, rewriting, or business rule injection.

Server Providers

Integrate real-time responses from external services (e.g., fraud signals, user profiles) during query execution.

Document Processors

Transform or enrich content as it’s indexed, including feature extraction or language-specific preprocessing.

Custom Application Blocks

Mix classic IR, vector relevance, metadata, and model predictions with fully controllable ranking expressions.

Search Result Renderers

Define how the result is serialized and returned to the client.

Result Page Templates

Define how search results are presented to end-users.

Real-World Use Cases

perplexity

Perplexity used Vespa to efficiently scale its RAG architecture, ensuring low-latency, real-time retrieval of relevant information from massive datasets. By building on Vespa's platform, Perplexity delivers accurate, near-real-time responses to more than 15 million monthly users, handling more than 100 million queries each week. Vespa.ai provides Perplexity with the flexibility, speed, and reliability needed to deliver best-in-class conversational experiences to millions of users worldwide.

Vinted

Vinted transitioned to Vespa to enhance the performance and scalability of their search and recommendation systems. Vespa's ability to handle complex, large-scale data and support machine learning models natively allowed Vinted to provide faster, more personalized search results. By leveraging Vespa's real-time indexing and flexible architecture, Vinted improved both the speed and relevance of its recommendations and search functionalities, offering users a more seamless experience when browsing second-hand items. This move also enabled Vinted to better scale with its growing user base while maintaining high efficiency and customization for search and discovery use cases.

Spotify®

Spotify introduced natural language search for podcast episodes, enhancing discovery by allowing users to search using conversational phrases like "podcasts about space exploration." Over time Spotify moved the entire search experience and recommendation use cases to Vespa. By leveraging Vespa, Spotify can efficiently process these complex queries, using machine learning models to understand context and meaning, and deliver highly relevant results in real-time. Vespa's ability to handle large-scale data and support advanced search capabilities ensures that users can find what they look for even if they don't use specific keywords, making the discovery process more intuitive and personalized.