Vespa.ai

Building AI-Native Information Platforms

Why Retrieval Engineering Is Becoming the Next Competitive Advantage

We Make AI Work

Vespa.ai

Vespa.ai develops the Vespa AI Search Platform, which unifies retrieval, ranking, machine-learning inference, and real-time serving for large-scale search and AI-native applications.

Document
Building AI-Native Information Platforms

07.07.2026

Contents

About this eBook

Who should read this ebook?

This ebook is written for organizations that build products around proprietary information, including:

- Business Information
- Competitive Intelligence
- Financial Intelligence
- Legal Research
- Market Intelligence
- Scientific Publishing

Although these organizations serve different markets, they share many of the same AI search, Retrieval Engineering, and architectural challenges. They typically:

- Build products around proprietary knowledge
- Deliver AI-powered research and investigation
- Require continuously updated information
- Operate at large document scale
- Need explainable AI over trusted content
- Need to simplify fragmented AI architectures

Every information platform is becoming an AI platform.

For decades, data service providers have competed on the quality, depth, and timeliness of their proprietary information. Today, a new competitive battleground is emerging: how intelligently that information can be retrieved, understood, and transformed into actionable insight.

Artificial intelligence is fundamentally changing how subscribers interact with data. Instead of searching reports, dashboards, and documents, users increasingly expect AI applications that investigate, reason, and answer complex questions using trusted, proprietary knowledge. This shift creates enormous opportunities for information platforms to deliver richer customer experiences, differentiate their products, and unlock new revenue streams.

Delivering those experiences, however, requires more than adding a large language model. As AI applications evolve from conversational assistants to deep research and autonomous agents, retrieval becomes the defining engineering challenge. AI systems must retrieve, rank, and continuously update the right information across billions of documents with predictable latency, operational efficiency, and complete trust in the results.

This is where AI Search Platforms emerge. Rather than stitching together vector databases, search engines, rerankers, and inference services, they unify the entire retrieval workflow within a single architecture, reducing complexity while improving relevance, scalability, and cost efficiency.

This eBook explores how leading information platforms are evolving to meet these demands. It introduces the principles of retrieval engineering, explains why unified AI Search Platforms are replacing fragmented retrieval stacks, and shows how organizations such as AlphaSense and Perplexity are building AI-native products that combine deep domain expertise with internet-scale performance.

Introduction

Search is Entering a New Era

Information platform providers—from market intelligence and financial data to legal research, scientific publishing, and business information services—are entering a new phase of AI adoption. They are no longer competing solely on the breadth and quality of their proprietary information, but increasingly on how intelligently that information can be retrieved, understood, and transformed into actionable insight.

Users no longer want to search thousands of reports, documents, and datasets themselves. They expect AI applications that investigate, reason, summarize, and explain complex questions using trusted, proprietary knowledge. For information platform providers, retrieval is no longer a supporting capability—it is becoming a defining part of the product experience.

Why Now?

Public AI assistants have fundamentally changed user expectations. Conversational interfaces are rapidly evolving into deep research assistants and autonomous AI agents capable of performing increasingly sophisticated analytical tasks. At the same time, advances in large language models, semantic retrieval, and real-time inference have made these experiences practical at scale.

For information platform providers, this creates a unique opportunity. Unlike general-purpose AI assistants trained primarily on public information, they combine proprietary content, trusted data sources, and deep domain expertise to deliver differentiated insights that others cannot. As AI becomes the primary interface to that information, the quality of retrieval increasingly determines the quality of the product itself.

Retrieval Engineering

Prompt engineering influences how a model reasons. Retrieval engineering determines what it has to reason about.

Search engineering has always balanced competing priorities: relevance, performance, scalability, and cost. AI raises the stakes considerably. Instead of retrieving information for people to evaluate, retrieval systems increasingly assemble the context that large language models and AI agents use to reason, generate, and act. Every retrieval decision becomes part of an automated workflow where quality, latency, and freshness directly influence the outcome.

This shift has given rise to a new engineering discipline: **Retrieval Engineering**. It is the practice of designing, optimizing, and operating retrieval workflows that balance search quality, latency, freshness, scalability, and infrastructure cost. Rather than focusing on individual technologies such as vector databases, rerankers, or inference services, it treats the retrieval workflow as a single system whose components must work together efficiently.

Retrieval workflows orchestrate multiple retrieval techniques, ranking models, and relevance signals, including:

- Hybrid retrieval (keyword, semantic, and structured data)
- Structured filtering and business rules
- Query rewriting and expansion
- Personalization
- Machine-learned ranking
- Real-time updates
- Machine learning inference
- Context assembly for language models and AI agents

Each capability improves retrieval but also adds computational cost and architectural complexity. Retrieval Engineering determines where these techniques add value and how to execute them efficiently at scale. As AI applications evolve from conversational assistants to deep research systems and autonomous agents, a single request may trigger hundreds of retrieval operations, making workflow efficiency essential for accurate, responsive, and cost-effective AI applications.

But what does an effective retrieval workflow look like?

The Limits of Fragmented Architectures

Vector databases solved an important problem: making semantic retrieval practical at scale. They have become an essential building block for AI applications, allowing systems to retrieve information based on meaning rather than exact keyword matches.

But retrieval is only one stage of a much larger workflow.

As AI applications mature, they quickly outgrow semantic retrieval alone. High-quality AI systems increasingly combine keyword search, structured filtering, business rules, machine-learned ranking, real-time updates, and inference to assemble accurate context before a language model generates a response. The engineering challenge shifts from selecting the right retrieval technology to orchestrating an increasingly sophisticated retrieval workflow.

Many organizations address this by integrating specialized technologies. A vector database provides semantic retrieval. A search engine handles keyword matching. Additional services provide reranking, filtering, personalization, and machine-learning inference. This approach works well initially, but every new component introduces another network hop, another operational dependency, and another source of latency.

The problem is no longer vector search. It is engineering a fragmented retrieval architecture.

As search evolved, engineers naturally adopted best-of-breed architectures, combining specialized technologies to improve retrieval quality. While this delivered increasingly capable applications, it also introduced additional latency, operational complexity, and infrastructure cost. As AI workloads grow, those trade-offs become increasingly difficult to justify.

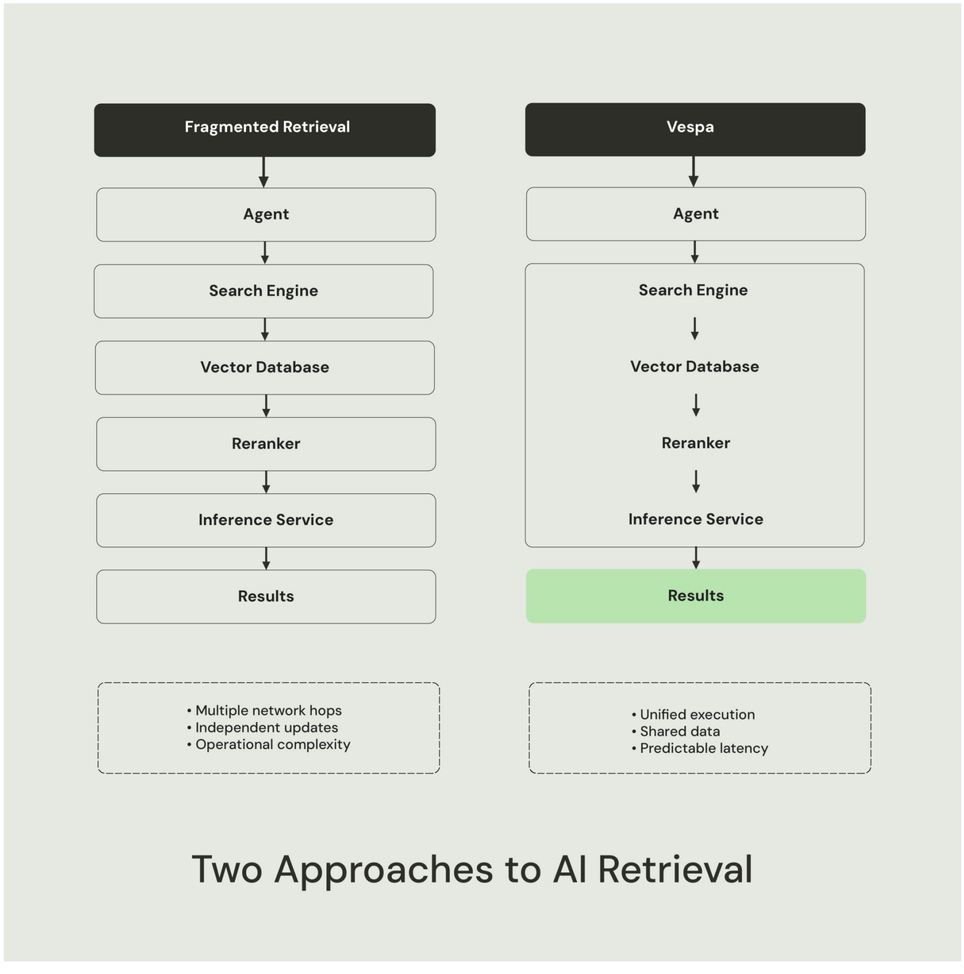
Vector databases remain an essential part of retrieval, but semantic similarity alone rarely determines the best result. High-quality AI applications combine vector similarity with exact keyword matching, structured filters, business rules, behavioral signals, freshness, authority, and machine-learned ranking to retrieve the most relevant information.

This is particularly important for domain-specific applications, where precise terminology, proprietary vocabularies, and structured business data often carry as much weight as semantic similarity. Retrieval quality depends not only on the embedding model, but on how effectively all of these signals work together.

Many vector databases including AI search architectures built around Lucene-based search engines address individual parts of this workflow extremely well. The engineering challenge is bringing those capabilities together into a retrieval architecture that remains efficient, scalable, and operationally simple as AI applications evolve.

The next evolution isn't another retrieval component. It's a platform that executes the entire retrieval workflow as a single system.

The AI Search Platform



Two Approaches to AI Retrieval

AI Search Platforms have become the execution layer for search applications—from traditional search and recommendations to answer engines and AI agents. Rather than assembling retrieval workflows from multiple specialized systems, they integrate retrieval, ranking, machine-learning inference, and real-time serving into a single architecture.

This unified approach enables organizations to execute the entire retrieval workflow as one system, reducing operational complexity while improving search quality, scalability, and infrastructure efficiency. As AI applications become more sophisticated, the platform, rather than the individual components, becomes the foundation for delivering fast, accurate, and trustworthy results.

Vespa is an AI Search Platform built for Retrieval Engineering, enabling customer-facing AI-native applications to deliver fast, trustworthy, and scalable intelligent experiences.

Building the Retrieval Workflow

Retrieval Engineering defines the principles. The AI Search Platform provides the solution.

1. Unified Retrieval

Retrieval identifies candidates.

Retrieval workflows combine keyword search, semantic retrieval, structured filtering, machine-learned ranking, real-time updates, and business logic into a single execution path. The challenge is not implementing any one of these capabilities in isolation—it's executing them together with predictable latency, operational simplicity, and infrastructure efficiency.

Rather than assembling independent services for retrieval, ranking, filtering, and inference, Vespa executes the entire retrieval workflow within a single distributed serving engine. This unified approach reduces unnecessary data movement, simplifies operations, and enables intelligent applications to scale without adding infrastructure.

Every successful AI Search Platform solves the same four engineering problems.

Retrieval is no longer simply about finding semantically similar documents. AI applications depend on retrieving the right amount of relevant context before reasoning begins. Modern retrieval workflows combine dense vector search, keyword search, structured filtering, metadata, and business rules to identify, rank, and assemble that context.

Vector databases solved an important problem by making semantic retrieval practical, but vector similarity alone rarely determines the best result. High-quality retrieval depends on combining multiple retrieval techniques within a single query.

Vespa executes hybrid retrieval natively, combining vectors, text, and structured data in a single distributed query. Because all retrieval methods execute where the data resides, complex hybrid queries avoid unnecessary network hops, maintaining predictable performance while lowering infrastructure costs.

This enables applications to combine semantic understanding with exact terminology, structured metadata, recency, authority, and other domain-specific signals to deliver more accurate retrieval.

Retrieval workflows have evolved beyond single-vector representations. Techniques such as late interaction, multi-vector retrieval, multimodal retrieval, and visual document understanding require richer representations than a single embedding can provide. Vespa's tensor-native architecture was designed for these emerging retrieval techniques, enabling sophisticated retrieval models to execute within the same distributed platform.

2. Intelligent Ranking

Ranking determines which information reaches the user or the language model.

3. Continuous Freshness

AI applications are only as trustworthy as the information they retrieve.

As AI increasingly consumes retrieved information directly, ranking becomes as important as retrieval. Rather than presenting a short list of results for a person to evaluate, AI retrieval must identify and assemble the right amount of relevant context for language models and AI agents. Ranking therefore determines not only which information is retrieved, but which context is ultimately used for reasoning, generation, and decision-making.

Applying sophisticated ranking models to every candidate would quickly become prohibitively expensive. Vespa uses multi-phase ranking to progressively refine candidate sets, applying increasingly sophisticated ranking models, including machine-learning inference, only where they improve the final result.

Because ranking executes locally where the data resides, inference scales naturally with the cluster while minimizing network overhead. The result is higher search quality without sacrificing throughput, latency, or infrastructure efficiency.

Information platforms operate in constantly changing environments where documents, customer data, embeddings, user behavior, and machine-learning models evolve continuously. Delayed updates quickly become stale retrieval, leading directly to poorer recommendations, outdated answers, and less reliable AI systems.

Vespa continuously indexes and updates structured data, text, vectors, and machine-learning models while serving live traffic. Applications remain available as data, indexes, and models evolve, eliminating disruptive rebuilds, scheduled refreshes, and maintenance windows.

This enables organizations to combine proprietary content with customer-specific information while ensuring AI applications always retrieve the latest available knowledge.

For example, Perplexity combines its indexed knowledge base with files uploaded by Pro users, allowing retrieval across both public and private information within a single workflow.

4. Internet-Scale Performance

AI dramatically changes the economics of retrieval

Traditional search applications typically perform a single retrieval for each user query. AI agents and deep research workflows may perform dozens, or even hundreds, of retrieval operations before producing a response.

Retrieval infrastructure designed for human-speed search can quickly become overwhelmed by these new workloads. As retrieval volumes increase, latency becomes harder to predict, infrastructure costs rise, and systems built from multiple specialized components become increasingly difficult to scale efficiently.

Vespa was built for internet-scale serving from the beginning. It partitions and distributes data automatically across clusters while executing retrieval, ranking, filtering, and machine-learning inference where the data resides. Nodes can be added, removed, or upgraded without interrupting queries or writes, allowing applications to scale elastically as both data volumes and AI workloads grow.

The result is predictable performance, lower infrastructure costs, and the ability to support billions of documents, thousands of concurrent queries, and continuously evolving AI applications.

Retrieval Engineering in Practice

The principles described in this guide already power some of the world's most demanding AI-native information platforms, from market intelligence and deep research to internet-scale answer engines.

AlphaSense

AI-powered market intelligence over 500 million premium business documents.

AlphaSense combines proprietary business content with AI-powered search built on Vespa's AI Search Platform to help professionals investigate, reason, and make faster decisions across finance, life sciences, and corporate strategy.

perplexity

Delivering AI answers at internet scale

Perplexity relies on Vespa to power retrieval across the public web, supporting fast, accurate answers with the performance required by millions of users.

RavenPack

Transforming proprietary financial data into actionable intelligence.

RavenPack uses Vespa to power AI search across millions of financial documents, helping AI agents investigate proprietary financial data and deliver trusted insights for investment professionals.

Summary

Ready to Build Your AI-Native Information Platform?

Whether you're modernizing an existing search architecture or building a new AI-native application, we'd be happy to discuss your retrieval architecture, explore opportunities to simplify your infrastructure, and help you deliver intelligent applications at scale.

Contact us.

Artificial intelligence is changing what customers expect from information platforms. Instead of searching for information themselves, they increasingly expect AI applications that investigate, reason, and deliver trusted insights directly into their workflow. That shift transforms retrieval from a supporting capability into one of the most important components of the product itself.

Building those experiences requires more than connecting a large language model to proprietary data. It requires engineering retrieval workflows that balance search quality, freshness, latency, scalability, and infrastructure cost. Prompt engineering influences how a model reasons. Retrieval engineering determines what it has to reason about.

Vector databases, search engines, rerankers, and inference services each solve part of the problem. The engineering challenge is bringing them together into a retrieval architecture that remains accurate, efficient, and operationally simple as AI applications evolve.

Vespa addresses that challenge by unifying retrieval, ranking, machine learning inference, and real-time serving within a single AI Search Platform. The result is a simpler architecture that delivers more accurate retrieval, lower infrastructure costs, and the predictable performance required for customer-facing AI-native applications.

The next generation of information platforms will not compete solely on the quality of their proprietary data. They will compete on how intelligently they retrieve, rank, and transform that data into actionable insight. Retrieval Engineering is becoming the discipline that makes that possible.

About Vespa.ai

Vespa.ai develops the Vespa AI Search Platform that brings retrieval, ranking, machine-learning inference, and real-time serving together within a single distributed architecture. Rather than stitching together fragmented search and AI retrieval components, Vespa executes the complete retrieval workflow close to the data, delivering high relevance, predictable latency, and operational simplicity at scale. Organizations including Yahoo, Spotify, Perplexity, and AlphaSense use Vespa to power mission-critical customer-facing applications.

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

GitHub Twitter LinkedIn YouTube

Vespa.ai