GIGAOM

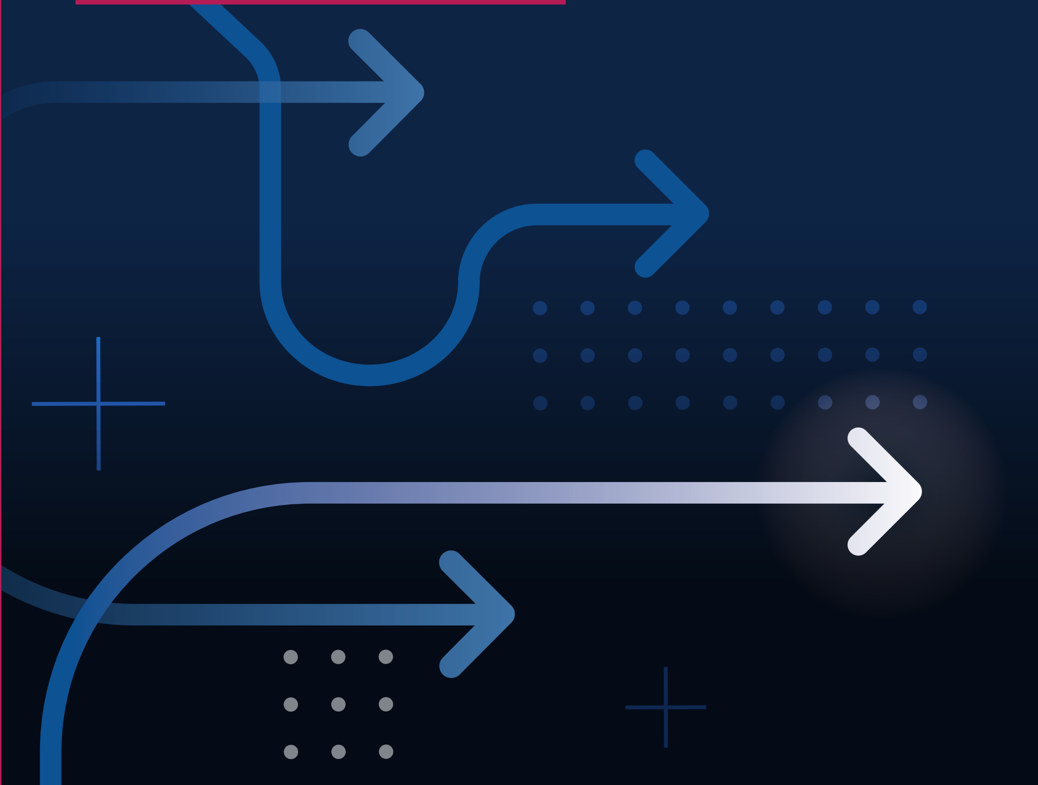
GIGAOM CxO BRIEF

COMMISSIONED BY:

Vespa

Migrating to AI-Native Search and Data Serving Platforms

DATA, ANALYTICS & AI



GIGA
OM

GigaOm CxO Decision Brief: Migrating to AI-Native Search and Data Serving Platforms

Solution Overview

Vespa.ai provides an advanced data platform for powering modern, intelligent applications. Unlike many traditional tools, Vespa allows businesses to analyze incoming information and run AI calculations instantly, directly on the freshest data. This capability enables highly responsive customer experiences, smarter operational decisions, and a faster path to innovative, AI-driven products.

Benefits

Enables accelerated AI innovation and superior application performance. Organizations report reduced TCO through infrastructure savings and enhanced search relevance, driving better user engagement.

Urgency

Urgent consideration is needed for organizations in competitive sectors (e-commerce, finance, media) or where real-time AI, RAG, and dynamic personalization are strategic imperatives.

Impact

Adoption impacts processes beyond IT, enabling faster product cycles and new marketing strategies. Requires broader training (product, support) and fosters a cultural shift toward real-time, AI-driven operations.

Risk

Primary risks involve the investment needed for specialized skills due to Vespa's distinct technology and learning curve. Underestimating training/ recruitment effort can delay realizing value.

01 Solution Value

MODERN DIGITAL BUSINESSES NOW CONFRONT A CRITICAL CHALLENGE: AI-driven applications demand near-instantaneous processing of massive data volumes while delivering precise, relevant results. Many search platforms struggle with these new workloads—especially those built on Apache Lucene (like Elasticsearch and Apache Solr), which were architected primarily for sparse, term-based retrieval rather than the dense vector operations needed for modern AI applications. This architectural mismatch results in escalating infrastructure costs, complex performance bottlenecks, and innovation delays. As enterprises implement capabilities like retrieval-augmented generation (RAG) and semantic search, conventional platforms increasingly become competitive liabilities rather than assets.

Vespa.ai solves these challenges with its unified real-time data and AI computation platform. Its architecture—unlike systems that separate data management from AI processing—was specifically designed to perform complex AI operations directly on data where it is stored. This enables developers to build applications that combine vector search, structured data filtering, keyword matching, and tensor-based scoring in a single query path, returning relevant results in milliseconds. Vespa.ai has published benchmark comparisons suggesting performance advantages in throughput and latency over alternatives; readers can consult Vespa's materials for details. Furthermore, organizations using Vespa, including Vinted, Spotify, and Perplexity, have reported achieving significant infrastructure savings while boosting accuracy following migration. These companies successfully deployed Vespa to create responsive, intelligent applications ranging from personalized recommendation engines to complex RAG systems, achieving measurable business value while simultaneously reducing technical debt.

02 Urgency & Risk

COMPETITIVE PRESSURES AND RISING USER EXPECTATIONS for instant, AI-powered, reliable experiences create urgency for modernizing data platforms. Sticking with legacy systems or less efficient alternatives introduces operational risks: mounting technical debt, escalating infrastructure costs, limitations in delivering real-time features, and slower innovation cycles. Organizations that fail to adopt AI-native platforms like Vespa will increasingly fall behind competitors that can deliver sub-second personalization, process queries in real time, and deploy sophisticated AI models at the edge. These capabilities directly translate to measurable business outcomes, including higher conversion rates, reduced customer churn, and accelerated time to market for AI-driven features.

Urgency

The urgency to adopt platforms like Vespa.ai arises from the rapid shift toward real-time, AI-driven user experiences, where falling behind competitors risks market relevance. Delaying modernization invites performance bottlenecks, unsustainable AI infrastructure costs, and an inability to leverage instant data updates—ultimately hindering innovation and degrading customer satisfaction. This need is most acute in sectors like e-commerce, finance, and media, or where real-time recommendations, RAG systems, and dynamic personalization are strategic imperatives.

Risk

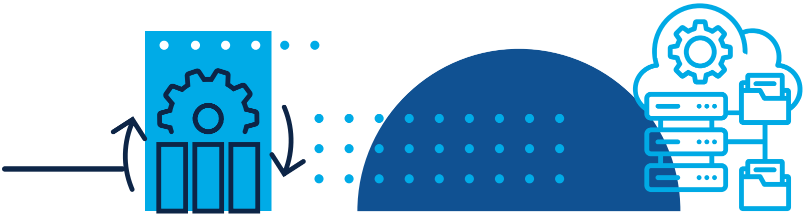
While Vespa.ai offers compelling advantages, executives must acknowledge the risks associated with its deployment and adoption. The primary risk lies in the required investment in specialized expertise. Vespa's architecture and concepts present a steeper learning curve than more established platforms, and the talent pool with deep Vespa experience is smaller. Underestimating the training, onboarding, and potential recruitment effort can lead to project delays and prevent realizing the platform's full value. Additionally, while potentially offering lower long-term TCO through efficiency, there are upfront migration costs and the risk that benchmarked performance gains may not fully materialize in every specific production environment. Organizations heavily reliant on the extensive plugin ecosystem of alternatives like Elasticsearch, or those lacking resources to invest in specialized Java/systems/ML skills, face heightened adoption risk—though Vespa is actively addressing these challenges through investments in service partners and application templates that can accelerate implementation.

03 Benefits

ADOPTING VESPA.AI TRANSLATES ITS ARCHITECTURAL ADVANTAGES into significant business benefits, enabling organizations to innovate faster, enhance user satisfaction, and operate more efficiently. Key benefits include:

- **Accelerated AI innovation:** Native AI/ML integration and real-time data handling enable development of sophisticated features, driving business outcomes. Users experience noticeable improvements in data visibility compared to traditional near real-time systems.
- **Superior performance efficiency:** Organizations achieve dramatically higher throughput and lower latency after migrating to Vespa, enabling the processing of massive data volumes while maintaining consistent performance under demanding loads.
- **Reduced total cost of ownership:** Vespa's resource optimization can deliver significant infrastructure savings, evidenced by dramatic reductions in server counts reported in case studies, contributing to lower TCO.
- **Enhanced search relevance:** Vespa's architecture allows for more sophisticated ranking that drives business results. This allows for considering vastly more candidate items per query, which improves result relevance and positively impacts user engagement.

> "Users experience noticeable improvements in data visibility compared to traditional near real-time systems."



04 Best Practices

MAXIMIZING THE BENEFITS OF VESPA.AI while mitigating adoption risks requires careful planning and proactive investment in team capabilities. A strategic approach focusing on specific use cases and leveraging the platform's strengths is crucial for success. Key recommendations include:

Start focused, scale intelligently: Initiate with well-defined pilot projects or proofs of concept to validate Vespa’s value for specific high-impact use cases before committing to large-scale migration or deployment. Implement phased rollouts.

Prioritize expertise development: Acknowledge Vespa’s distinct architecture and proactively invest in training internal teams on Java component development, schema definition, ranking profiles, and operational management, or secure external expertise.

Design for Vespa's strengths: Architect applications to fully utilize Vespa’s real-time update capabilities, integrated ML inference for ranking, and native hybrid search features rather than simply replicating patterns from previous platforms.

Plan for monitoring: Establish clear monitoring practices for Vespa’s operational metrics early in the deployment process, integrating with existing observability tools where possible.

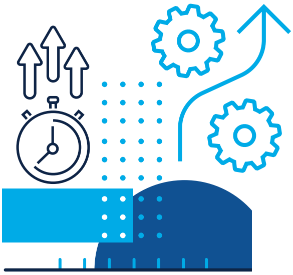
05 Organizational Impact

IMPLEMENTING VESPA.AI EXTENDS BEYOND THE TECHNICAL TEAMS, creating ripple effects across the organization and fundamentally enhancing the customer experience. Its ability to deliver real-time updates and integrated AI enables more dynamic, personalized, and relevant interactions, directly impacting customer satisfaction and loyalty. This shift necessitates adjustments in internal processes. Product development cycles can accelerate, leveraging faster feedback loops and the ability to deploy AI features more readily. Marketing teams may need new strategies to capitalize on real-time personalization capabilities. Just as a broad change like multifactor authentication requires widespread understanding, adopting Vespa's advanced potential requires more than just engineering skills. Product managers must grasp the new possibilities for feature development, support teams may need training on troubleshooting AI-influenced results or real-time data issues, and data governance practices might need updating to accommodate the platform's dynamic nature. Successfully harnessing Vespa encourages a cultural shift toward embracing real-time data insights and fostering closer alignment between technical, product, and business units to drive continuous innovation.

People Impact

To accurately assess the people impact, understanding Vespa.ai requires recognizing its specific technological identity and deployment options. Originating from Yahoo, Vespa exists both as an open-source engine and a managed service (Vespa Cloud). Crucially, Vespa is not built on Apache Lucene technology, meaning its core operations—from configuration and real-time data handling to how computation is performed—differ from other search platforms. This necessitates investment in specific skill sets, including proficiency with Vespa's configuration approach, understanding its real-time processing model, developing custom logic

> "[Vespa's] ability to deliver real-time updates and integrated AI enables more dynamic, personalized, and relevant interactions, directly impacting customer satisfaction and loyalty."



05 Organizational Impact

> **“Vespa’s operational efficiency and automation potential can lead to leaner operational staffing needs or enable reallocation of valuable personnel toward innovation and feature development, positively impacting overall TCO.”**

using its Java component framework, and integrating ML models via its compute capabilities. The talent pool specifically versed in Vespa is growing but smaller, which can potentially extend recruitment timelines. Fostering cross-functional teams that tightly integrate platform engineering with ML/data science expertise can unlock significant value from Vespa’s capabilities. Although initial budget impacts include training and potentially higher recruitment costs, these are often offset by long-term gains. Vespa’s operational efficiency and automation potential can lead to leaner operational staffing needs or enable reallocation of valuable personnel toward innovation and feature development, positively impacting overall TCO.

Investment Outlook

The investment outlook for Vespa.ai focuses on potential long-term TCO reduction, primarily achieved through significant infrastructure efficiency. While specific spend depends on deployment scale, reports indicate substantially lower infrastructure costs are possible for equivalent workloads, especially those involving AI. Initial investments include migration efforts and necessary team training. Vespa’s approach to licensing and pricing contrasts sharply with typical enterprise software models that often involve complex, feature-limited tiers. The core Vespa engine is open source (Apache 2.0), eliminating software license fees and providing access to all features for self-hosted deployments. The managed Vespa Cloud service offers similar simplicity via transparent, consumption-based pricing directly tied to allocated resources (compute, storage), not confusing tiers. Financial risk thus shifts from navigating license complexity to managing resource usage or self-hosted operational costs. Vespa Cloud, as a SaaS offering, allows for rapid provisioning, while self-hosted timelines depend on internal infrastructure capabilities.

06 Solution Timeline

THE TIMELINE FOR IMPLEMENTING VESPA.AI VARIES based on project scope, team expertise, and deployment strategy, ranging from weeks for focused proofs of concept to several months for complex migrations or full-scale rollouts. Key factors impacting the schedule include the team's familiarity with Vespa's operation and architecture, as well as the necessary investment in training and ramp-up time. Migrating from existing systems requires dedicated effort for schema translation, query logic adaptation, and data re-ingestion. Utilizing Vespa Cloud can significantly accelerate infrastructure provisioning compared to self-hosted deployments, which depend on internal infrastructure readiness. Initial implementation timelines are often shorter when following best practices like starting with a limited-scope pilot project before broader adoption. Integrating Vespa with existing data pipelines and application front ends also requires planned development time.

Future Considerations

Future data platforms must natively integrate AI capabilities like vector search and RAG to power competitive applications, exposing the technical debt in older systems. Over the next three years, the ability to perform complex ranking and compute efficiently on real-time data is critical. Relying on platforms not designed for these tasks limits responsiveness, increases AI implementation complexity, and hinders innovation due to architectural constraints. Vespa.ai's development explicitly targets these needs, prioritizing advancements in its integrated AI inference, semantic search/RAG features, and core performance. Users should anticipate continued evolution supporting demanding, real-time AI workloads. Selecting a platform fundamentally built for these modern computational patterns, like Vespa, provides a strategic opportunity to overcome legacy limitations.

"Utilizing Vespa Cloud can significantly accelerate infrastructure provisioning compared to self-hosted deployments, which depend on internal infrastructure readiness."

07 Analyst's Take

THE DATA SERVING SECTOR IS PIVOTING toward platforms that natively embed AI and handle real-time computation, driven by the need for superior user experiences and operational efficiency. Vespa.ai stands out as a technically advanced and rational solution within this evolving landscape. Its architecture inherently supports low-latency query execution combined with real-time data updates and integrated AI inference, aligning it directly with key market drivers. This positioning is validated by Vespa's role powering the RAG architecture for Perplexity, where it delivers sub-100ms response times while processing over 100 million queries weekly for more than 15 million users across billions of indexed documents. For organizations building modern search, recommendation, or RAG-enabled systems where real-time AI performance is paramount, Vespa warrants serious consideration and should be on the evaluation short list. However, successful adoption requires commitment. Key risks involve underestimating the specialized skill development needed and navigating a smaller ecosystem. Mitigation involves investing in training, focused pilot projects, assessing tooling needs, and potentially leveraging Vespa Cloud to reduce operational complexity.

Report Methodology | Vespa

THIS GIGAOM CXO DECISION BRIEF ANALYZES a specific technology and related solution to provide executive decision-makers with the information they need to drive successful IT strategies that align with the business. The report is focused on large impact zones that are often overlooked in technical research, yielding enhanced insight and mitigating risk. We work closely with vendors to identify the value and benefits of specific solutions, and to lay out best practices that enable organizations to drive a successful decision process.

About Whit Walters

WHIT WALTERS IS A PIONEERING TECHNOLOGY EXECUTIVE with over 25 years of experience driving enterprise innovation. As Field CTO at GigaOm, he brings deep expertise in data platforms, cloud architecture, and AI/ML solutions. A multiple-time CTO, Whit is a frequent speaker on artificial intelligence and technology strategy, helping organizations navigate digital transformation.

GIGAOM

About GigaOm

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization to apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.