Vespa.ai

VESPA ENTERPRISE PLAN

Maximize Performance and Scale without Limits

Accelerate innovation, improve reliability, and lower TCO for production-scale Vespa deployments while maintaining full control over security and data privacy.

Benefits of Enterprise

Vespa Enterprise removes the burden of running your own infrastructure so your team can stay focused on building, shipping, and improving applications.

Optimize frequently without disruption, resize infrastructure as schemas and workloads evolve, and use committed spend options to reduce cost even further.

The result is faster iteration, lower operational risk, and a production-ready environment for large-scale RAG, search, and recommendation workloads.

- Managed Infrastructure
- Continuous Deployment
- Built-In Security
- 24/7 Support
- Early Access

Trusted By

Spotify · perplexity · yahoo! · DuckDuckGo · Taboola

Onyx.app reduced infrastructure costs by 24.5% using Vespa's automated resource optimization features. The platform analyzed their workload patterns and identified CPU overprovisioning, recommending a cluster reconfiguration from 120 nodes to 60 nodes, with zero downtime, requiring only a three-line configuration change.

Yuhong Sun, Cofounder Onyx.app

onyx

For Onyx.app, Vespa's automated CI/CD, seamless node migrations, and fault-tolerant operations turned continuous optimization into a safe, push-button experience. Tasks like provisioning, upgrades, node replacements, and load-balancing disappeared overnight.

With Vespa's real-time monitoring and intelligent resource suggestions, Onyx instantly identified opportunities to right-size their cluster and **cut costs by 24.5%, without any downtime**.

The result: a faster, leaner, more reliable AI platform and a team freed to innovate at full speed.

+ 2.5x faster
after switching from Elasticsearch to Vespa

Vinted

"Vespa let us collapse 3 retrieval tiers into 1 engine we can actually reason about. Every playlist now feels like it was handcrafted for you because, under the hood, it basically was."

Daniel Doro,
Director of Search Engineering,
Spotify

Spotify

Precision at Scale

Always-on, real-time personalization.

yahoo!

- 150+ online, real-time applications
- 800 queries per second (QPS) from across the globe
- 100B+ multi-modal documents

Internet-scale RAG at consumer search speed.

perplexity

- 10B-document global RAG index
- 120M monthly users (2M+ daily)
- 10X traffic increase with zero impact on latency

Recommendations at global scale.

Spotify

- 696M monthly active users
- 100M+ songs and 7M+ podcasts
- 32% YoY catalog growth without increasing infra costs