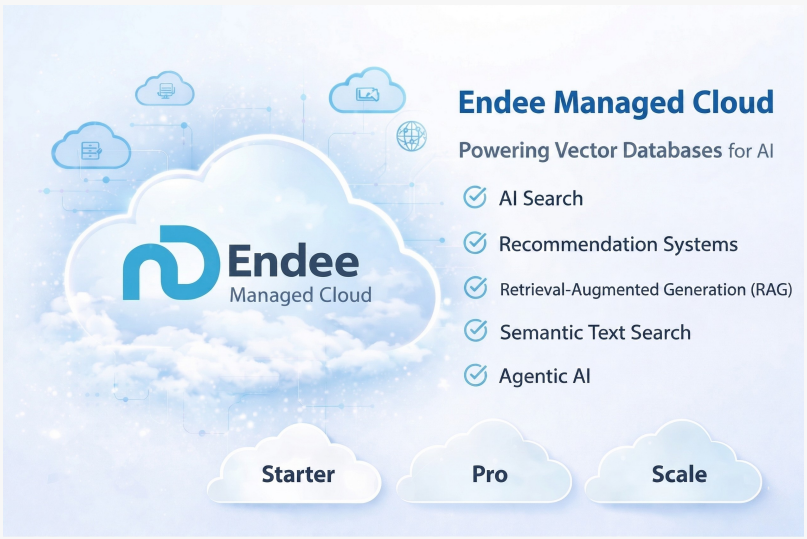
Endee Launches Managed Cloud for its Open-Source Vector Database with Generous Free Tier

The open-source vector database Endee.io, that is well known for its Ultra High performance with 10x lower Infra, is now available as a fully managed cloud



Endee cloud use cases

SAN FRANCISCO, CA, UNITED STATES, April 13, 2026 /EINPresswire.com/ -- Endee Labs today announced the launch of Endee Cloud, a fully managed, serverless vector database built for production AI workloads. Endee Cloud makes the fastest and most cost-efficient open-source vector database available as a zero-ops managed service, with a free Starter plan and paid Pro and Scale tiers for teams running mission-critical AI search, retrieval-augmented generation (RAG), and recommendation systems at scale.

Endee is an open-source, Apache 2.0-licensed vector database. On VectorDBBench and other independent vector database benchmarks, Endee delivers the highest throughput (QPS), the highest recall, the lowest P99 latency, and the lowest cost per query of any tested vector database, outperforming Pinecone, Qdrant, Milvus, Weaviate, and others across all metrics simultaneously at 10x lower infrastructure.

> Endee is the vector database for production AI. The fastest, most accurate, most cost-efficient, and now the easiest to deploy."
>
> *Vineet Dwivedi, Founder, Endee Labs*

BENCHMARK PERFORMANCE

Highest throughput (QPS): Endee delivers more queries per second than any competing vector database, making it the highest-throughput option per dollar for high-traffic AI search, real-time recommendations, and production RAG pipelines.

Highest recall (accuracy): Endee achieves the highest recall of any tested vector database. For RAG pipelines and AI agents, higher recall means fewer hallucinations and more grounded, reliable AI output.

Lowest P99 latency: Even at the 99th percentile, Endee returns results faster than any competing vector database. Sub-10ms tail latency makes it the right choice for real-time AI applications, interactive search, and latency-sensitive retrieval workflows.

Lowest cost per query: Endee's C++ core, SIMD acceleration, filter-aware HNSW indexing, and multi-precision quantization deliver dramatically better cost efficiency, running production workloads on modest hardware where other vector databases require expensive, memory-heavy clusters.

ENDEE CLOUD

"Endee is the vector database for production AI. We built it to be the fastest, most accurate, and most cost-efficient vector database available - and now we’re making it the easiest to deploy," said Vineet Dwivedi, Founder of Endee Labs. "With Endee Cloud, teams get the best performance of any vector database with zero infrastructure management. No clusters to provision, no indexes to tune, no performance to debug. Just an API key and production-grade vector search in minutes."

Endee Cloud delivers serverless, zero-ops deployment with no clusters to provision, no indexes to tune, and no infrastructure to manage. The platform auto-scales dynamically to match workload demands. The free Starter plan gives full access to the Endee API and client SDKs for Python, TypeScript, Java, and Go, designed for prototyping, evaluation, and early-stage AI products. Pro and Scale tiers unlock higher throughput, dedicated resources, priority support, and enterprise-grade SLAs for teams running production AI at scale.

Endee Cloud includes native hybrid search, combining dense vector retrieval with sparse search in a single query, with payload filtering for metadata-aware results. Queryable encryption keeps data encrypted even during search operations, a capability no other open-source vector database offers. Endee is ISO 27001, SOC 2 Type II, and GDPR certified.

Because Endee is Apache 2.0 licensed, teams can migrate between Endee Cloud and self-hosted deployments at any time with zero code changes.

USE CASES

Endee is already powering production AI systems at enterprises across e-commerce, manufacturing, industrial automation, and education. Teams use Endee for semantic search, RAG pipelines and LLM applications, AI agent memory and context retrieval, real-time recommendation engines, and enterprise knowledge retrieval. Endee's queryable encryption and

compliance certifications also make it suitable for government, healthcare, and financial services deployments where data security, compliance, and cost control are critical requirements.

AVAILABILITY

Endee Cloud is available now at https://endee.io with a free Starter plan requiring no credit card. Sign up and start querying in under five minutes. The full open-source engine is available at https://github.com/endee-io/endee. Complete documentation, integration guides, and a quick start guide are available at https://docs.endee.io.

ABOUT ENDEE LABS

Endee Labs builds the fastest and most cost-efficient vector database for production AI workloads. Endee is an open-source, Apache 2.0-licensed vector database that delivers the highest throughput, highest recall, lowest latency, and lowest cost per query of any vector database on independent benchmarks. Trusted by top enterprises across industries and backed by an active open-source community on GitHub, Endee powers semantic search, RAG pipelines, recommendation engines and AI agents.

Vineet Dwivedi
Endee Labs Pvt Ltd
email us here
Visit us on social media:
LinkedIn
X
Other
