aws

Pinecone: The Vector Database for Knowledgeable AI

Turn your data into knowledge on AWS Serverless Infrastructure

In collaboration with

Pinecone

Table of Contents

INTRODUCTION

Making AI knowledgeable with Pinecone and AWS

The AI revolution powered by large language models (LLMs) like Claude and Amazon Titan transformed how we build applications. However, LLMs struggle with long-term memory and context windows. As a result, they are inefficient and costly for complex, multi-step workflows that require accuracy and performance at scale.

LLMs do not inherently understand the internal processes, documentation, or regulatory nuances specific to an organization. Additionally, they can't capture proprietary data critical for enterprise use cases that include search, recommendations, and agents. Meanwhile, traditional databases are not up to the task of supporting semantic searches, a key component of AI applications.

Vector databases bridge the gap

Vector databases address these challenges. They store embeddings, which are numerical representations of emails, images, logs, PDFs, and other types of unstructured content in a continuous vector space. This type of information, which comprises more than 80% of the world's data, then becomes queryable in real time.

Introducing the Pinecone and AWS partnership

Pinecone is a pioneer in the vector database category. It supports its serverless architecture with Amazon Web Services (AWS) to help solve fundamental scaling and cost challenges. Seamless integration with AWS services enables multimodal search with fresh, relevant results and zero infrastructure management.

This ebook explores how engineering leaders rely on Pinecone's AWS-native and serverless architecture and its unique features to ship production-grade AI applications faster while reducing costs.

5,000+

More than 5,000 organizations, including Gong and Vanguard, use Pinecone on AWS to power AI applications in production with hundreds of millions to billions of vectors.

The challenge of unstructured data at scale

Key data and storage demands emerge from real AI use cases: freshness for real-time updates, elasticity for traffic spikes, cost efficiency for intermittent queries, and scalability for high-dimensional unstructured data.

Traditional databases were made for structured data

Traditional databases struggle with high-dimensional data. For one, as data grows, performance degrades. Also, they aren't built for handling it. They were originally designed for structured data and exact matches, instead of unstructured data and the semantic and similarity search capabilities required by AI use cases. Therefore, using traditional databases for AI workloads converts engineering effort into infrastructure management and optimization rather than model development or business innovation.

Bolt-on vector capabilities still have traditional infrastructure

The answer to scaling for high-dimensional data isn't bolt-on vector capabilities in existing databases, however. Adding vector search as a module does not change the underlying storage, indexing, and compute assumptions of the traditional database. These bolt-on capabilities also keep large portions of the vector index in memory, which quickly becomes prohibitively expensive.

The result of bolt-on vector capabilities in existing databases

10X higher costs

100X slower queries

Open-source vector databases can struggle with performance at scale and high operational overhead

Open-source vector databases can fall short when it comes to high-dimensional data and distributed cluster management. Security and compliance requirements multiply with sensitive data, requiring reliable isolation between users and datasets that many open-source vector databases struggle to provide at scale, especially without the necessary engineering expertise.

Configuration complexity becomes a bottleneck. There are steep learning curves for approximate nearest neighbor (ANN) algorithm tuning and distributed cluster management that pull even experienced engineering teams away from core AI development. Additionally, many vector databases make architectural trade-offs that break down under production workloads. Memory requirements balloon costs. Query accuracy degrades with dataset size. Availability issues appear during traffic spikes.

The solution to this challenge is Pinecone, a vector database built from the ground up to handle enterprise requirements and scale.

Why Pinecone powers knowledgeable AI

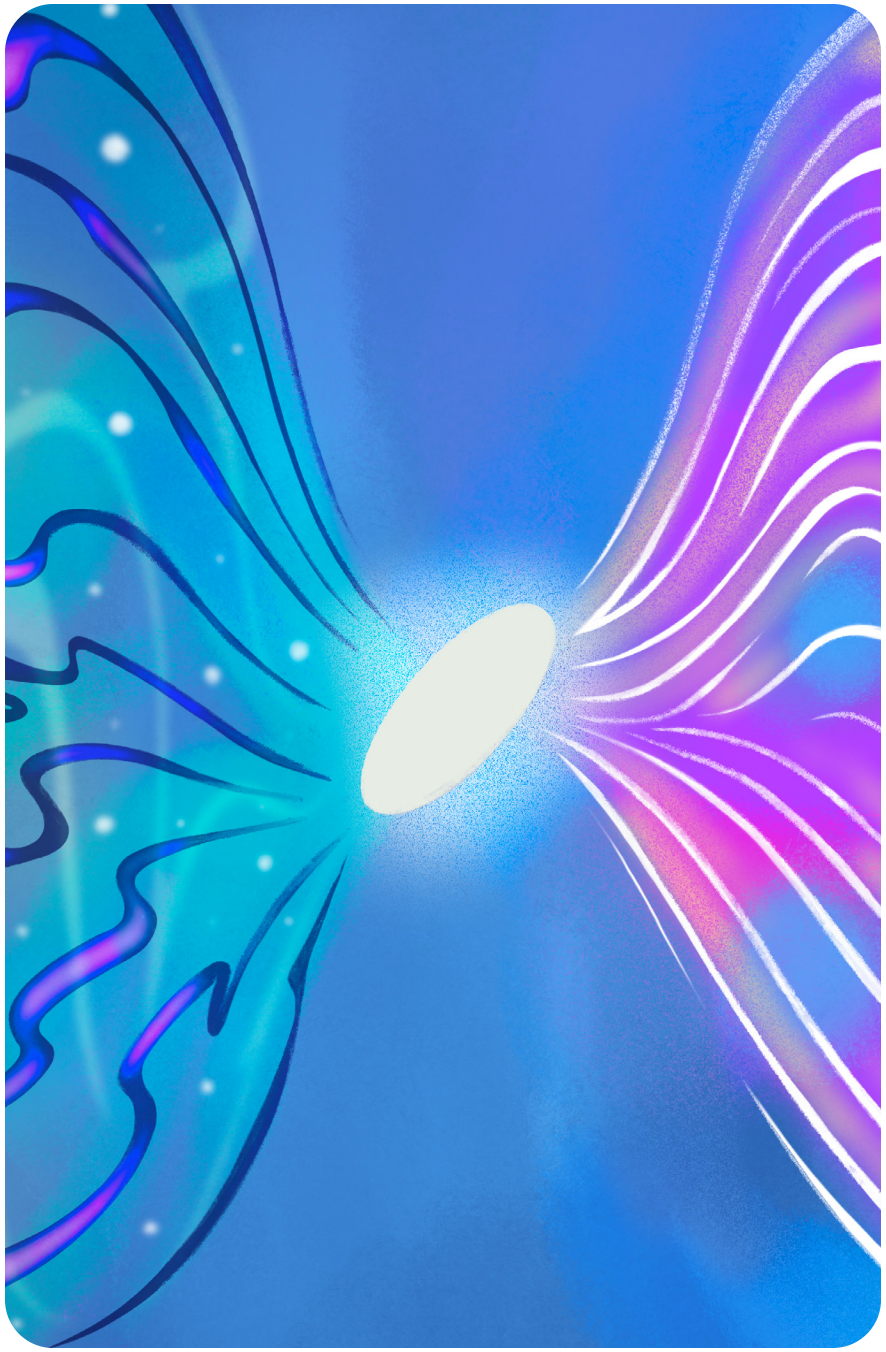
Pinecone is a pioneer in the vector database category. Others quickly understood the importance of transforming unstructured data into high-dimensional numerical vectors that enable semantic search. However, what makes Pinecone different is its ability to minimize the trade-offs of cost, scale, accuracy, performance, and management that other technologies force customers to make. Uniquely, Pinecone dynamically orchestrates ANN algorithms at scale to optimize for workload use case and outcome.

This is how Pinecone's latest serverless architecture delivers exceptional results for vastly different use cases many companies need including search, recommendations, and agents. Pinecone delivers true knowledge retrieval by understanding context and meaning, allowing AI applications to find relevant insights in real time from semantically organized data.

Context engineering made simple

Building accurate LLM applications requires context engineering, the architecting necessary to maintain relevant information across complex, multi-step workflows. Context engineering involves organizing, filtering, deleting, and processing information so that an LLM can continue to focus on the task at hand.

Pinecone's vector database provides the ideal infrastructure for retrieving context at scale. It includes options for highly-performant sparse and dense embedding models and hosted rerankers for optimizing the relevance of returned results.



Performance, accuracy, and scale without compromise

Accurate retrieval combines embedding models, reranking, and flexible query mechanisms, including vector similarity, metadata filtering, and namespaces. Pinecone's Rust-based architecture delivers low latency and high accuracy, even on billion-vector datasets.

Pinecone delivers 20ms-100ms latency on 100M to billion-vector datasets, with thousands of sustained QPS, depending on the workload configuration.

Real-time indexing ensures new data becomes searchable in seconds. A multi-tenant architecture with intelligent caching enables cost-effective access, so customers pay only for operations performed, not idle infrastructure.

Ecosystem integration

Integration with Amazon Bedrock, Amazon SageMaker, LangChain, OpenAI, Cohere, and other AI and AWS services accelerates development from prototype to production. Pinecone Assistant now functions as an MCP server, enabling seamless integration with agentic workflows and the growing ecosystem of AI tools and frameworks.

These capabilities are all made possible by Pinecone's innovative approach to infrastructure design.

The power of Pinecone's serverless infrastructure

Vector databases introduce unique complexities around data locality, search performance, and memory management. Pinecone's serverless infrastructure solves these challenges through a unique and sophisticated slab-based architecture, purpose-built on AWS.

Slab architecture: Engineered for scale and speed

Pinecone's vector indexing architecture balances freshness, scalability, and accuracy through three core mechanisms:

- **Immediate durability**
  Writes are logged durably and immediately in memory, and then written to Amazon S3 as immutable files called slabs. Writes are never blocked by reindexing or ongoing queries.
- **Multi-level compaction**
  Multi-level compaction runs continuously in the background. It merges smaller slabs into larger, more efficiently indexed slabs and prevents queries from scanning thousands of tiny files. Because slabs are distributed, the system scales elastically without resharding.
- **Adaptive indexing**
  Reads instantly fan out to search in memory and across all existing data slabs. Reads will always return the freshest results because writes happen in parallel, and the most accurate results because each slab is adaptively indexed based on data size.

Pinecone serverless and AWS under the hood

Pinecone serverless uses AWS services like Amazon Elastic Kubernetes Service (Amazon EKS), Amazon Aurora, Amazon Simple Storage Service (Amazon S3), and AWS Key Management Service (AWS KMS) to deliver a cloud-native architecture.

Amazon S3 provides durable object storage for vector indexes, and on-demand queries run and access just the parts of the index needed from S3. This is different from the vector databases that keep the full index in memory on the shards, which makes it expensive to run queries on large datasets.

Pinecone also integrates seamlessly with services like Amazon Bedrock and Amazon SageMaker for enhanced AI capabilities.

Choose your deployment for your use case: On-demand or dedicated

On-demand delivers elastic, usage-based pricing with intelligent caching. Frequently accessed slabs stay in memory, while less-used slabs are fetched from S3 on demand. This deployment is ideal for RAG applications, agentic workloads with millions of small namespaces, and bursty traffic patterns.

Dedicated Read Nodes provide predictable low latency through isolated infrastructure with data always warm in memory and local SSD. This option was built for semantic search at billion-vector scale, high-QPS recommendation engines, and mission-critical services with strict SLOs.

Bring your own cloud

The Pinecone data plane can be deployed directly into a customer's VPC while still being managed directly by Pinecone.

AWS-native architecture

Pinecone's innovative architecture decouples storage from compute, allowing each to scale independently based on actual workload demands. It also employs a tiered storage mechanism that dynamically moves relevant portions of indexes between persistent object storage and high-performance ephemeral compute resources.

This separation enables usage-based pricing that aligns costs with real consumption, automatic scaling from as few as 10 vectors to billions without manual intervention, and built-in high availability backed by a 99.9 percent uptime SLA. Namespaces provide logical data isolation within indexes, supporting multi-tenancy without infrastructure overhead, while global deployment ensures low latency and data residency compliance.

With the power of Pinecone's unique serverless architecture, developers can focus entirely on building applications that deliver real value to users.

Building production-grade AI applications with Pinecone

Pinecone supports developers throughout the entire AI application lifecycle, from rapid prototyping to deployment in production and beyond. It provides tools and capabilities designed to accelerate development while ensuring production readiness.

Rapid prototyping

Developers can create their first index in under five minutes using intuitive Python or JavaScript SDKs. Simple APIs eliminate the learning curve, allowing teams to focus on application logic rather than infrastructure complexity. Pinecone's architecture enables multimodal search capabilities out of the box, delivering fresh and relevant results across text, images, and structured data without additional configuration.

Sophisticated functionality

As applications mature and use case needs become more focused, sparse indexes allow for the seamless addition of hybrid search with existing dense indexes. This flexibility supports diverse use cases from RAG pipelines and conversational agents to recommendation systems and semantic search, with integration patterns optimized for common architectures.

Production excellence

Built-in monitoring and observability tools track query performance, index statistics, and usage patterns in real time, giving developers visibility into application behavior as it scales. Enterprise security features, including private endpoints, encryption, audit logs, and SOC 2 Type II compliance certification, ensure applications meet stringent production requirements.

Optimization at scale

As a managed service, Pinecone implements index optimization techniques, including batch operations and efficient upsert strategies that maintain performance as datasets grow. This guidance helps developers avoid common pitfalls and ensures applications remain responsive under increasing load.

From initial concept to enterprise deployment, Pinecone provides the infrastructure, tools, and expertise developers need to build AI applications that deliver real value to users. And Pinecone and AWS have the real-world stories to prove it.

Pinecone and AWS real-world success stories

Businesses across industries and around the world are using Pinecone for their AI initiatives.

GONG

Revolutionizing revenue AI: Gong's collaboration with Pinecone

Gong is a Revenue AI operating system that harnesses customer interactions to unify teams, workflows, and insights to drive consistent revenue growth. Gong's commitment to innovation led to developing proprietary AI technology to enable teams to capture, understand, and act on all customer interactions in a trusted, unified OS.

Gong selected Pinecone as the core database infrastructure for its AI Tracker agent. It plays a crucial role in storing and processing vector embeddings, enabling accurate search and classification of concepts within user conversations. Through Pinecone, Gong achieves efficient vector searches, empowering its AI agents to offer users precise and relevant examples for concept tracking in conversations.

Gong processes 7 billion vectors on Pinecone, achieving a 10x cost reduction while powering agents for sales insights.

Vanguard

Boosting customer support: Vanguard chooses Pinecone to power hybrid retrieval

Vanguard, a leading investment management company, offers a range of financial services, including retirement services, advice, and investments. Vanguard chose Pinecone as their vector database to power hybrid retrieval for Agent Assist, an AI assistant, so that customer support representatives could respond faster and more accurately to calls.

Since using Pinecone for RAG and search, Vanguard has seen tangible improvements. Hybrid retrieval improved result accuracy by over 12% compared to dense retrieval alone. Faster, more precise retrieval has significantly cut customer wait times. The team can now support peak periods (such as tax season) without additional overhead. Metadata tagging enables better traceability for audit purposes.

Vanguard now has a solution that supports hybrid search, real-time updates, enterprise scale and production, flexible metric selection, and advanced metadata filtering.

The future of knowledgeable AI

Advances in search, architectures, and integration are shaping how AI systems gain, use, and maintain knowledge, promising more accurate, adaptive, and deeply informed applications. With Pinecone at the center of these applications, updating knowledge in real time as new data arrives and maintaining relevance without manual retraining through continuous learning, the future of knowledgeable AI looks bright. Here are just a few of the advances and capabilities that can be enhanced with Pinecone.

Hybrid and semantic search

Modern search blends semantic understanding with lexical precision for more relevant results than traditional full-text search. Pinecone supports hybrid retrieval (semantic + lexical) with integrated reranking to improve retrieval accuracy. Namespaces and metadata filtering support multi-tenant workloads. Its serverless slab architecture helps keep accuracy high and latency low as data and traffic scale.

Pinecone integrates with popular frameworks like Amazon SageMaker, LangChain, Haystack, Hugging Face, and more, and data sources like Databricks, Snowflake, and Confluence. This allows it to fit smoothly into existing workflows and ecosystems for multimodal projects.

Recommendation engines

Applications from ecommerce to media and beyond rely on recommendation engines to surface “similar” items or content that are relevant to their users. At their core, these recommendation engines use vectors to provide similarity search. They often must support very high throughput and low latency to maintain a positive user experience and, therefore, are well served by Pinecone’s Dedicated Read Nodes.

RAG and agents

As model capabilities converge, your competitive advantage in AI will be your data. RAG grounds models in your data. Agents extend models with planning and tool use. An agent breaks a task into steps, selects tools (such as databases, MCPs, and APIs), calls them, often retrieving context from Pinecone, and feeds results back into the model. Pinecone delivers concise, relevant context through hybrid search with integrated reranking. It keeps your context minimized and on topic to help reduce hallucinations and costs. On AWS, teams can use Pinecone as a Knowledge Base for Amazon Bedrock, providing Bedrock with the most semantically relevant content at prompt time.

Context engineering

Context engineering is the discipline of assembling the information an LLM or agent needs into the context window so it can act correctly. This includes tools, instructions, specific pieces of content (such as docs, tickets, code, logs), memory, and intermediate outputs. Pinecone provides the retrieval backbone and hybrid search with integrated reranking returns the most relevant information. Its serverless slab architecture keeps indexes fresh and latency low as content changes, enabling better answers with smaller context windows.

Convergence with LLMs

When LLMs and the right vector database converge, it's possible to build truly knowledgeable applications that understand context and deliver accurate insights. The outcomes of integrating LLMs with Pinecone include deep learning capabilities for embedding generation with efficient vector storage and retrieval.

The era of truly intelligent applications, where AI systems can learn, remember, and reason with the full breadth of human knowledge, is upon us. Once a distant vision, it's an achievable reality with Pinecone as the foundation.

> "As we adopted a more advanced architecture, Pinecone remained the clear choice. The reliability of their product and the quality of their support reaffirmed our decision to work with them as a trusted partner."
>
> — **Alvin Alaphat, Founding Engineer at Delphi**

Start building today

Pinecone removes the complexity of AI infrastructure so you can focus on creating value from your proprietary data. Our serverless architecture with AWS means zero infrastructure management. You ship faster and pay only for what you use.

Start building knowledgeable AI applications with Pinecone and scale seamlessly as you grow.

Create and query your first Pinecone index for free →

In collaboration with

Pinecone

aws PARTNER