Pinecone

Pinecone is a vector database that makes it easy to store and retrieve data for use in GenAI applications, such as chatbots, recommender systems, RAG pipelines, and search engines.

Index Options

An index is the highest-level organizational unit of vector data in Pinecone. A pod is a pre-configured unit of hardware for running a Pinecone service. One index is typically made up of many pods.

Note: the table below does not account for metadata

Pod Type	Max Vectors *per pod*	Optimization	QPS@k* *per pod* (1M x 768d)	Supported Vector Types	Architecture	p95 Latency*
p1	1.25M x 512-dim 1M x 768-dim 675K x 1024-dim 500k x 1536-dim	Speed & cost	30@k=10 25@k=250 20@k=1000	Dense, Sparse	Clustering + raw vectors in memory	50-100 ms
p2	1.25M x 512-dim 1.1M x 768-dim 1M x 1024-dim 550k x 1536-dim	Speed	150@k=10 50@k=250 20@k=1000	Dense	Graph + raw vectors in memory	5-50 ms
s1	8M x 512-dim 5M x 768-dim 4M x 1024-dim 2.5M x 1536-dim	Storage & Cost	10@k=10 10@k=250 10@k=1000	Dense, Sparse	Clustering + raw vectors on disk	100-200 ms

*Performance is dependent on environment; adding replicas, using multithreading/processing, increasing pod size, and/or making use of performance-optimized wrappers, like Pinecone's gRPC client, drastically increase measures like QPS.

Integrations: Haystack, Databricks, Amazon Bedrock, TruLens, Spark, Airbyte, Datadog, LangChain, LlamaIndex, Confluent

AI Models: Compatible with dense vector embeddings from any AI model or LLM (OpenAI, Anthropic, Cohere, Hugging Face, PaLM, etc.) and sparse vector embeddings (BM25, SPLADE, etc.) used in hybrid search

SDKs: Python, NodeJS, Upcoming: Java & Go

Core Engine: Rust

Clients: REST, gRPC

Search options: Vector search (semantic search) or hybrid search (keyword-aware semantic search)

Cloud: GCP, AWS, Azure deployments supported; available on AWS & GCP marketplaces

Metadata: Each vector can hold 40 KB of attached metadata; filtering by metadata fields enabled by default (<, >, <=, >=, =, !=, in, not-in)

Security: SOC2 Type II Certified; GDPR ready; HIPAA compliant; data encrypted at rest & in flight; RBAC; SSO

Consistency: Eventual (we have prioritized availability & performance over consistency)

CRUD Operations: Idempotent writes; full & partial updates; zero downtime during writes & updates

Max Dimensions: 20k (dense vectors), <=1000 non-zero values (sparse vectors)

See pricing or contact us.

October, 2023.