Pinecone Factsheet

Pinecone is a fully managed vector database that makes it easy to add vector search to production applications. Live indexing, lightning-fast vector similarity search with filtering*, on managed and distributed infrastructure.

Vector Search Capabilities

Engine types:
- Exact (kNN): Powered by Faiss
- Approximate (ANN): Powered by proprietary engine

Distance metrics: cosine (default), dot product, euclidean

Metadata filters*, can be combined with AND and OR:
- $eq - Equal to *(number, string)*
- $ne - Not equal to *(number, string)*
- $gt - Greater than *(number)*
- $gte - Greater than or equal to *(number)*
- $lt - Less than *(number)*
- $lte - Less than or equal to *(number)*
- $in - In array *(string)*
- $nin - Not in array *(string)*

Scale and Performance

For vectors with up to 1024 dimensions, typical query latency is sub-50ms with 95–100% recall, and can be as low as 3ms with batch queries and other optimizations.

Your latency, throughput, recall, and freshness depend on data size, index type, and deployment configuration.

Example with 1.2M vectors x 200 dimensions:

Index type	Latency	QPS	Recall
Exact	100ms	10	100%
Approx	20ms	50	95.8%
Approx, 3 replicas	20ms	150	95.8%

Latencies include network overhead.
See benchmark and try with your data.

Security

- Container isolation for customer data
- Pinecone only monitors operational metrics
- Dedicated or cloud-prem deployments available
- Data encrypted in transit
- GDPR compliant
- SOC2 certification in progress

API

Use the REST API or Python, Java*, and Go* clients.
- upsert: Insert or update vectors and metadata
- delete: Delete vectors by ID
- fetch: Retrieve vectors or metadata by ID
- query: Find *k* nearest neighbors to query vector
- summarize: Stats about index contents
- list (ids): List vectors in index
- list (namespaces): List namespaces in index
- deploy: Create new index
- delete: Delete index
- list (indexes): List indexes
- status: Index liveness status
- whoami: Authenticated user/project names
- version: Server and client software versions

Monitoring

Ingest metrics into Prometheus or Prometheus- and OpenMetrics-compatible monitoring tools.
- pinecone_item_count *(gauge)*
- pinecone_request_count *(counter)*
- pinecone_request_error_count *(counter)*
- pinecone_request_latency_seconds *(histogram)*

Deployment & Pricing

Cloud
Multi-tenant environment, an API call away. Try it free.
Pricing: $0.10/node/hr (Node = 1 shard, 1GB capacity)
Estimate price per hour:
(Data size in GB, rounded up) × (# replicas, default is 1)

Dedicated Cloud
Specify regions and AZs, we'll spin up dedicated clusters in our VPC on AWS/GCP. Connect via AWS PrivateLink or Google Private Access. Contact us for pricing.

Cloud-Prem
Run Pinecone in your own AWS/GCP VPC and grant permissions to Pinecone. Contact us for pricing.

** Metadata storage, metadata filtering, Go client, and Java client are coming in September.*