

✦ COMPARISON · RAG SYSTEMS

Vector Database Comparison

pgvector, Pinecone, Weaviate, Qdrant, Milvus, OpenSearch / Elasticsearch and Chroma compared on positioning, hosting, hybrid search, filtering, scale and day-two operations.

FORMAT	LEVEL	READ TIME	UPDATED
Comparison	Beginner	12 min	Feb 2026

OVERVIEW

What this comparison covers

Choosing a vector store is rarely the decision that makes or breaks a RAG system, but choosing badly creates months of avoidable operational work. This guide lays out how the main options are positioned, where each one is strong, and the questions we ask clients before recommending one. It is deliberately neutral: every option here runs production workloads somewhere.

KEY TAKEAWAYS

- Start from your existing stack: if you already run Postgres or a search cluster, those are serious contenders before any new database.
- Hybrid search (keyword + vector) and metadata filtering matter more for enterprise text than raw vector speed.
- Managed services trade control and data locality for far less operational work — decide which you value more.
- Purpose-built engines (Milvus, Qdrant, Weaviate, Pinecone) earn their place at large scale or with demanding filtering and multi-tenancy.
- Keep the vector store behind an interface so a migration is a re-index job, not a rewrite.

WHO IT'S FOR

CTOs and engineering leads choosing a RAG stack

AI / ML engineers building retrieval

Platform and DevOps teams who will run it

Architects reviewing build-vs-buy decisions

HOW TO USE IT

Read section 01 for the landscape, use the tables in 02–03 to shortlist two or three options, then run the decision checklist in section 06 against a real slice of your data.

CONTENTS

01 Three families of vector store
02 Hybrid search and filtering
03 Hosting, scale and day-two operations
04 Extend existing vs adopt dedicated
05 When to choose which
06 Decision checklist

01 / LANDSCAPE

Three families of vector store

The options fall into three groups. Knowing which family you are choosing from narrows the decision faster than comparing feature lists.

EXTEND WHAT YOU RUN

Database extensions

pgvector adds vector columns and indexes to Postgres. One database, one backup story, SQL joins and transactions.

SEARCH ENGINES

OpenSearch / Elasticsearch

Mature full-text engines with vector fields added. Strong keyword search, aggregations and an established ops ecosystem.

PURPOSE-BUILT

Dedicated vector databases

Pinecone, Weaviate, Qdrant, Milvus and Chroma are built around vectors first, with filtering and hybrid search layered on.

Positioning at a glance

OPTION	POSITIONING	LICENSE / DELIVERY	TYPICAL SWEET SPOT
pgvector	Postgres extension	Open source; on any Postgres incl. major managed services	Teams already on Postgres; small-to-medium corpora
Pinecone	Fully managed vector service	Proprietary SaaS	Teams that want zero database operations
Weaviate	Vector DB with built-in hybrid and modules	Open source; self-host or Weaviate Cloud	Hybrid search and multi-tenant SaaS products
Qdrant	Vector DB focused on filtering and efficiency	Open source (Rust); self-host or Qdrant Cloud	Filter-heavy retrieval; cost-conscious self-hosting
Milvus	Distributed vector DB for very large scale	Open source; self-host or Zilliz Cloud	Very large collections; dedicated data platform teams
OpenSearch / Elasticsearch	Search engines with vector support	OpenSearch: open source; Elastic: source-available / commercial	Organizations already running search clusters
Chroma	Developer-first, embeddable store	Open source; local or Chroma Cloud	Prototypes, notebooks, small apps

BEFORE YOU READ ON

Verify current features and pricing

This market moves quickly. Hybrid search, serverless tiers, licensing and limits have all changed in the past year. Confirm current capabilities, quotas and pricing with each vendor at the time you decide.

02 / RETRIEVAL

Hybrid search and filtering

Enterprise questions contain product codes, names and policy numbers that pure vector search misses. Filtering by tenant, permission and date is equally non-negotiable.

Retrieval capabilities

OPTION	KEYWORD / HYBRID	METADATA FILTERING	NOTES
pgvector	Combine with Postgres full-text search; fuse results yourself	Full SQL WHERE clauses and joins	Most flexible filters; hybrid needs your own fusion query
Pinecone	Sparse-dense hybrid supported	Metadata filters; namespaces for partitioning	Sparse vectors must be generated and managed
Weaviate	Native BM25 + vector hybrid with tunable weighting	Rich filters; native multi-tenancy	Hybrid works out of the box
Qdrant	Sparse vectors and query-time fusion	Strong payload filtering and indexes	Filtering is a core design focus
Milvus	Sparse vectors and full-text features in recent releases	Scalar filtering and partitions	Check version for hybrid features
OpenSearch / Elasticsearch	Best-in-class BM25; hybrid query support	Mature filters and aggregations	Keyword side is the strongest here
Chroma	Basic document text filtering	Metadata where-filters	Fine for prototypes; limited hybrid

Filtering approaches

POST-FILTERING
Retrieve top-k, then drop results the user cannot see
Can return too few or zero results
Risky for permission enforcement
Simple to implement

FILTER INSIDE THE QUERY
Permission and tenant filters applied during search
Result counts stay predictable
The only safe pattern for access control
Supported by every serious option above

✦ 03 / OPERATIONS

Hosting, scale and day-two operations

The real cost of a vector store shows up after launch: re-indexing, upgrades, backups, capacity planning and on-call. Weigh these as heavily as query features.

Operational profile

OPTION	HOSTING CHOICES	SCALING MODEL	OPS BURDEN
pgvector	Self-host or any managed Postgres	Vertical first; read replicas; partitioning	Low if you already run Postgres
Pinecone	Managed only	Handled by the service	Lowest; you manage usage and cost
Weaviate	Docker / Kubernetes or managed cloud	Sharding and replication	Moderate when self-hosted
Qdrant	Docker / Kubernetes or managed cloud	Sharding and replication; quantization	Moderate when self-hosted
Milvus	Kubernetes or managed (Zilliz)	Distributed, separate compute and storage	Highest when self-hosted; most components
OpenSearch / Elasticsearch	Self-host or managed (AWS, Elastic Cloud)	Shards and replicas	Moderate; familiar to search teams
Chroma	Embedded, single node or managed cloud	Single node in open source	Low; not designed for large clusters

Operational questions to answer before you commit

- How long does a full re-index take when you change embedding models, and can you run it alongside the live index?
- How are backups, point-in-time restore and disaster recovery handled?
- What happens to query latency during bulk ingestion?
- Where does the data physically live, and does that satisfy your residency obligations?
- Who is on call for it, and do they already know the technology?

04 / HEAD TO HEAD

Extend existing vs adopt dedicated

The most common real decision is not Pinecone versus Qdrant. It is whether to add vectors to a database you already run, or introduce a dedicated engine.

EXTEND POSTGRES OR SEARCH CLUSTER	DEDICATED VECTOR DATABASE
No new system to secure, back up and monitor	Built for vector workloads at large scale
Vectors live next to the business data they describe	Hybrid, quantization and multi-tenancy designed in
Transactions and joins for permissions and metadata	Another system and another data sync to operate
Tuning gets harder as collections grow very large	Managed options remove most of that burden

HYBRID OPTION

Use both, deliberately

A common production pattern keeps system-of-record data and permissions in Postgres and syncs embeddings to a dedicated engine for retrieval. It works well, but only with a reliable change feed and a reconciliation job so the two never drift apart.

✦ 05 / GUIDANCE

When to choose which

Short, opinionated defaults. Each one assumes you will still validate on your own data.

DEFAULT

pgvector

You already run Postgres and your corpus is modest. Fewest moving parts and the simplest permission joins.

MANAGED

Pinecone

You want no database operations and are comfortable with a SaaS-only, usage-priced service.

HYBRID-FIRST

Weaviate

You need strong keyword + vector hybrid and native multi-tenancy, self-hosted or managed.

FILTER-HEAVY

Qdrant

Retrieval depends on complex payload filters and you want an efficient open-source engine.

VERY LARGE SCALE

Milvus

Collections are very large and you have a platform team, or you choose its managed offering.

SEARCH-LED

OpenSearch / Elasticsearch

You already operate a search cluster and keyword relevance is critical to your users.

PROTOTYPE

Chroma

You are exploring an idea locally and want the shortest path to a working demo.

BOTH

Postgres + dedicated engine

Business data and ACLs stay in Postgres while a dedicated engine serves high-volume retrieval.

> The best vector database is usually the one your team can already operate at three in the morning.
>
> SPLENDENSLABS ENGINEERING PRINCIPLE

✦ 06 / DECIDE

Decision checklist

Run this with engineering, security and whoever owns the budget. Then benchmark the shortlist on a real slice of your corpus and golden questions.

FIT

- [ ] We know our expected document and chunk counts for the next 12–24 months
- [ ] We have listed the filters every query needs (tenant, permission, date, type)
- [ ] We know whether keyword matching on codes and names matters to users

OPERATIONS

- [ ] We have decided managed versus self-hosted, and who is on call
- [ ] Backup, restore and re-index procedures are documented
- [ ] Data residency requirements are confirmed against the hosting region

VALIDATION

- [ ] Shortlisted options benchmarked on our data with our golden questions
- [ ] Recall, latency and cost measured with filters enabled, not in isolation
- [ ] Current pricing, limits and license terms verified with each vendor

PORTABILITY

- [ ] Vector store sits behind a retrieval interface in our code
- [ ] Raw text and metadata are stored outside the vector index
- [ ] A migration is a re-index job we have tested at least once

✦ GET IN TOUCH

Pick once, pick well

We help teams choose, benchmark and run the retrieval layer of production RAG systems. Bring your corpus and constraints, and we will shortlist and test the right store with you.

- Vector store shortlist and benchmark on your data
- Hybrid search and permission-aware retrieval design
- Migration and re-indexing plans

BOOK AN AI STRATEGY SESSION →

EMAIL
hello@splendenslabs.com

WEB
splendenslabs.com

OFFICE
Sona Incubation Foundation, Salem