A cloud native database for next generation AI applications

$whoami

- Solutions Architect with Zilliz
-
- Deeply passionate about data and (Gen)AI

Zilliz's Mission:

Helping organizations make sense of unstructured data.

2017
Founded

$113M
Raised

140+
Employees

Redwood City, CA
Headquarters

Milvus: The most widely-adopted vector database

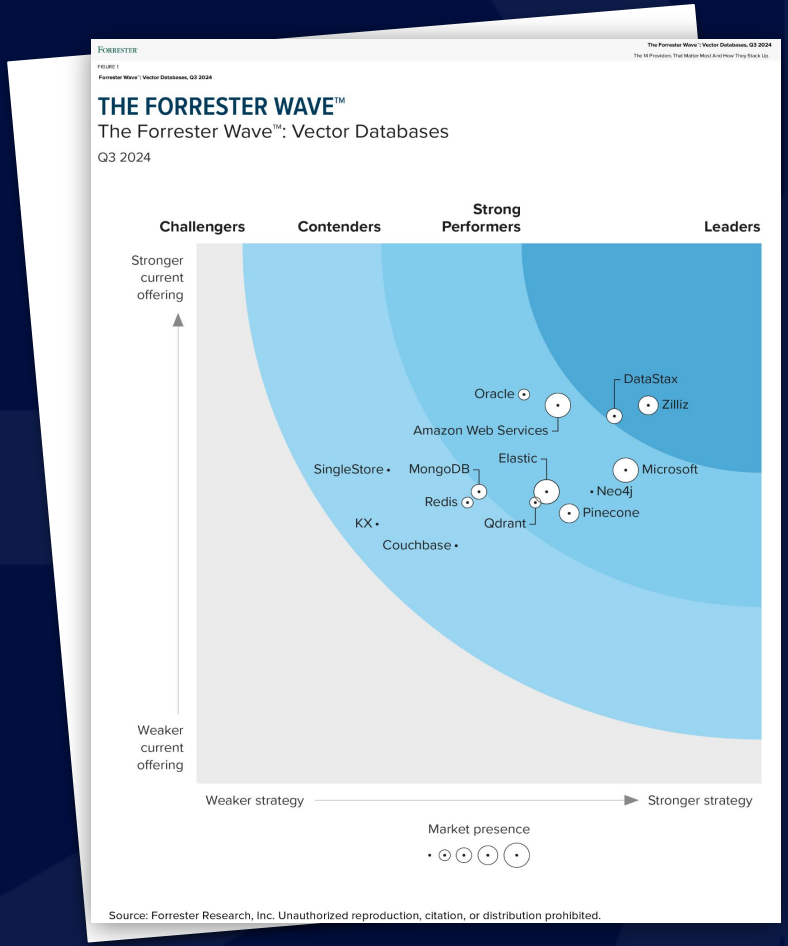
Milvus is an **Open-Source Vector Database** to **store**, **index**, **manage,** and **use** the massive number of **embedding vectors** generated by deep neural networks and LLMs.

283+
contributors

30K+
stars

67M+
docker pulls

2.8K
+
forks



The top industry analysts endorse our leading position

Specifically, we perform top at the following core vector database competency categories

- **Vector Index** - *the only vendor in the market leverage multiple types of index and benefit from the unique strengths of each*
- **Performance** - *the only vendor provides advanced performance acceleration through various hardware optimization, query tuning*
- **Scalability** - *the only vendor offer compute storage separation and auto tiered storage*

Built by database & AI experts


Milvus: A Purpose-Built Vector Data Management System


Manu: A Cloud Native Vector Database Management System

Zilliz was built by a top-tier team of **algorithm and database engineers** with a strong pedigree in developing **high-performance, scalable**, and **highly available** distributed systems, uniquely tailored for **vector search.**

What is a vector database?

What is a vector database?

What are vector embeddings?

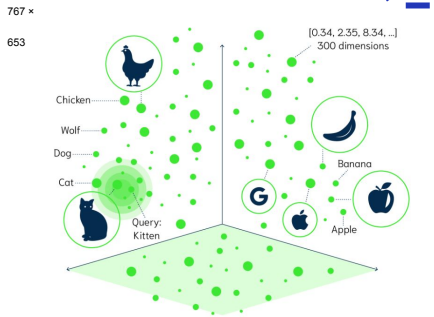
Numerical representations of unstructured data (text, images, audio, video, etc) that captures their semantics and relationships.

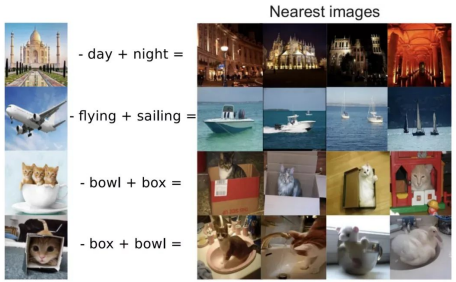
What is a vector database?

Vector database supports efficient storage, search and management of vector embeddings

How to use vector db

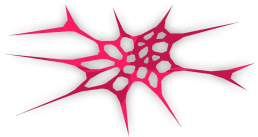
extract → insert → index → search / operations





What problems do vector databases solve?

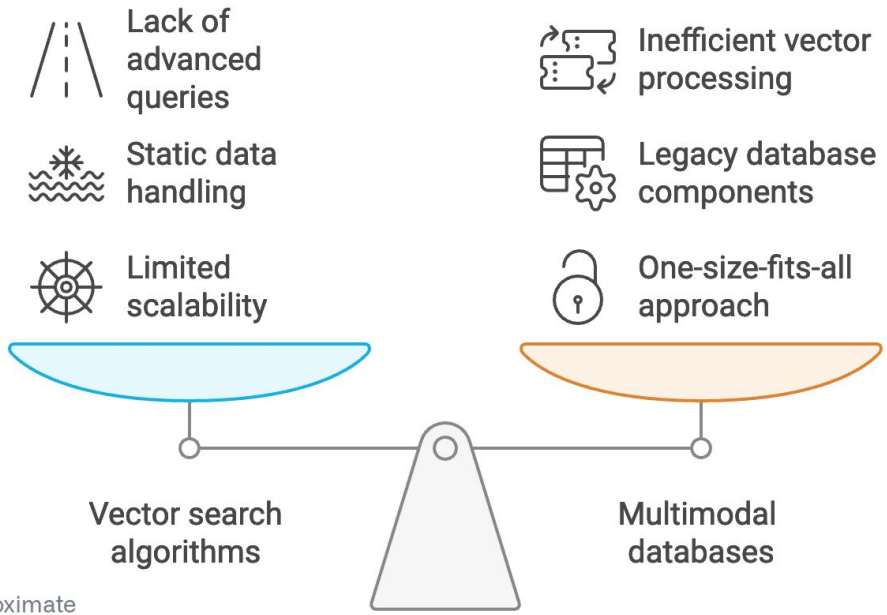
FAISS
Scalable Search With Facebook AI



ANNOY

nmslib/**hnswlib**

Header-only C++/python library for fast approximate nearest neighbors

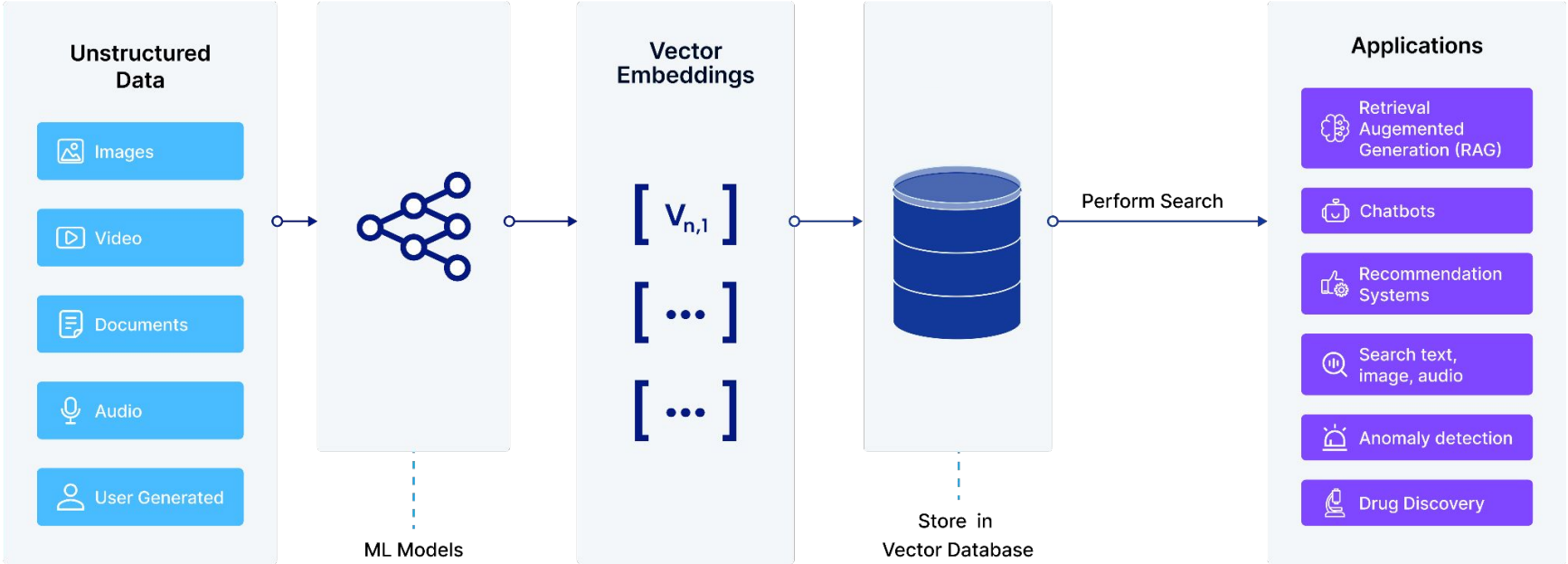


mongoDB

elasticsearch

redis

A New tool emerged. The Vector Database

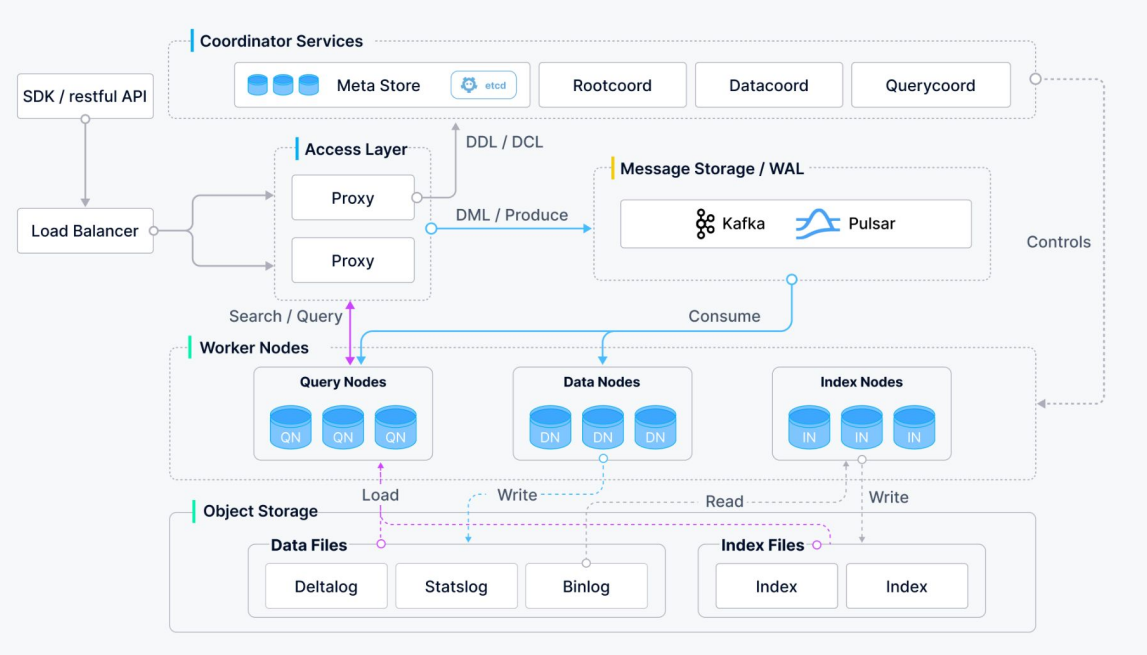


Milvus Architecture

Design Principles

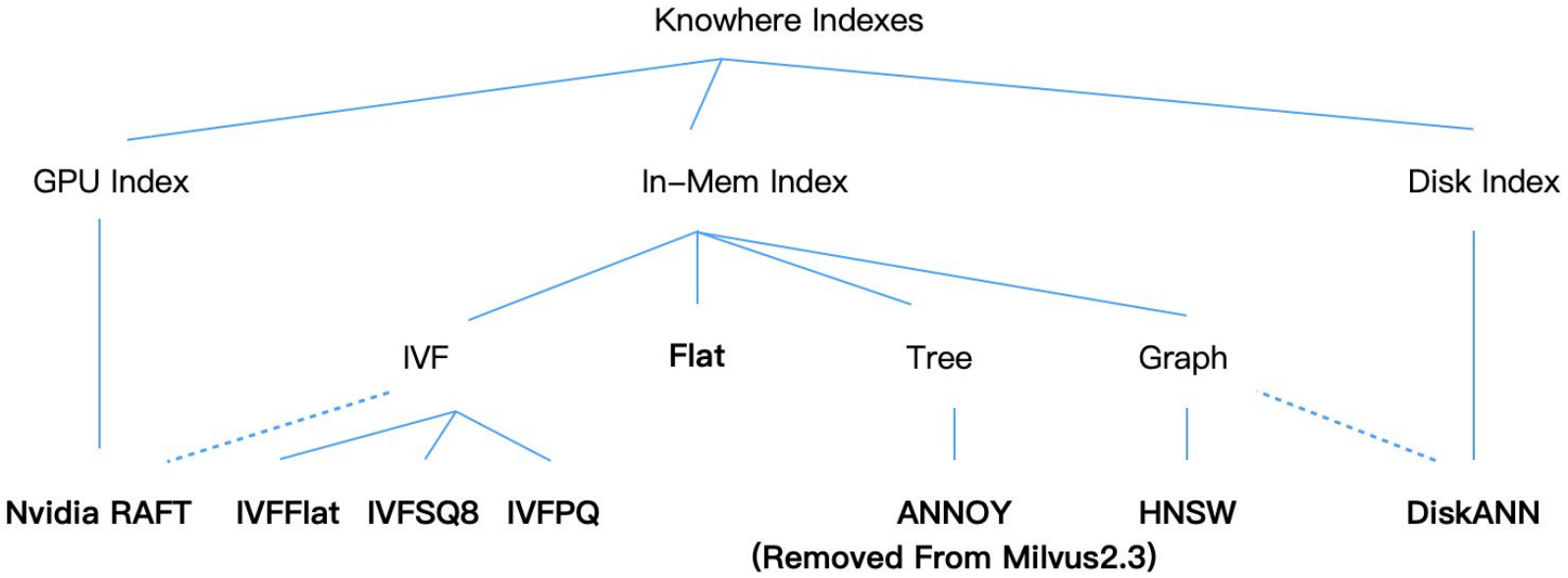
- Separation of storage and compute
- Fully depend on mature storage systems
- Microservice - scale by functionality
- Separate streaming and historical data
- Pluggable engine, storage and index
- Log as data

Fully distributed, designed for scalability



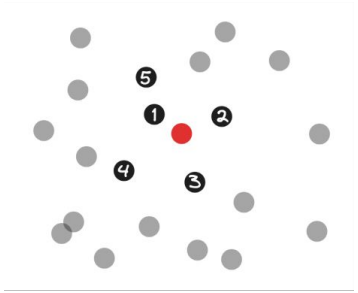
Milvus v2.4.x architecture overview

Vector Indexes supported by Milvus

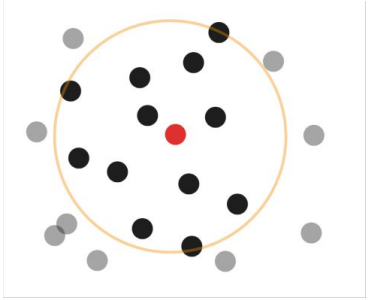


Search queries

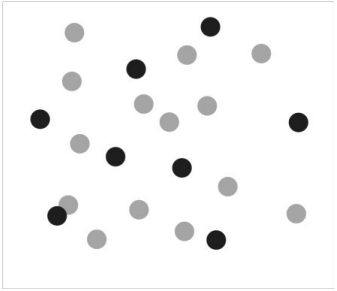
Search



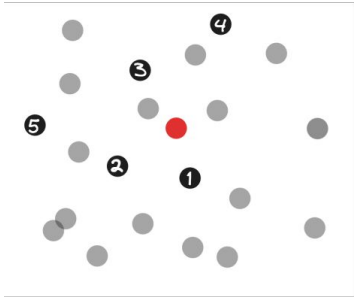
Range Search

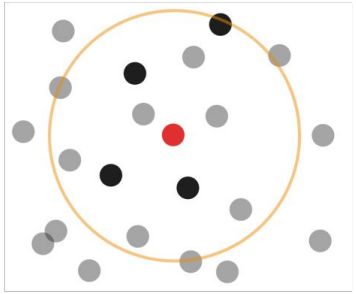


Filtered Search

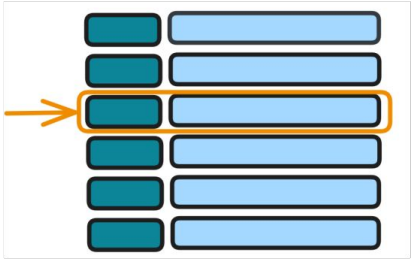


Hybrid Search

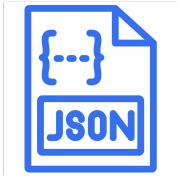




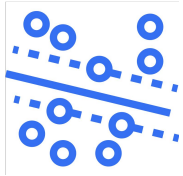
Query



Rich functionality



Dynamic Schema



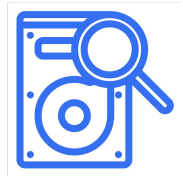
Float, Binary, & Sparse Vector

Tag+Vector Optimized Filtering

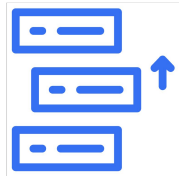
Hybrid Search Dense & Sparse

RBAC, TLS, Encryption

Million+ level tenant support



Disk Based Index

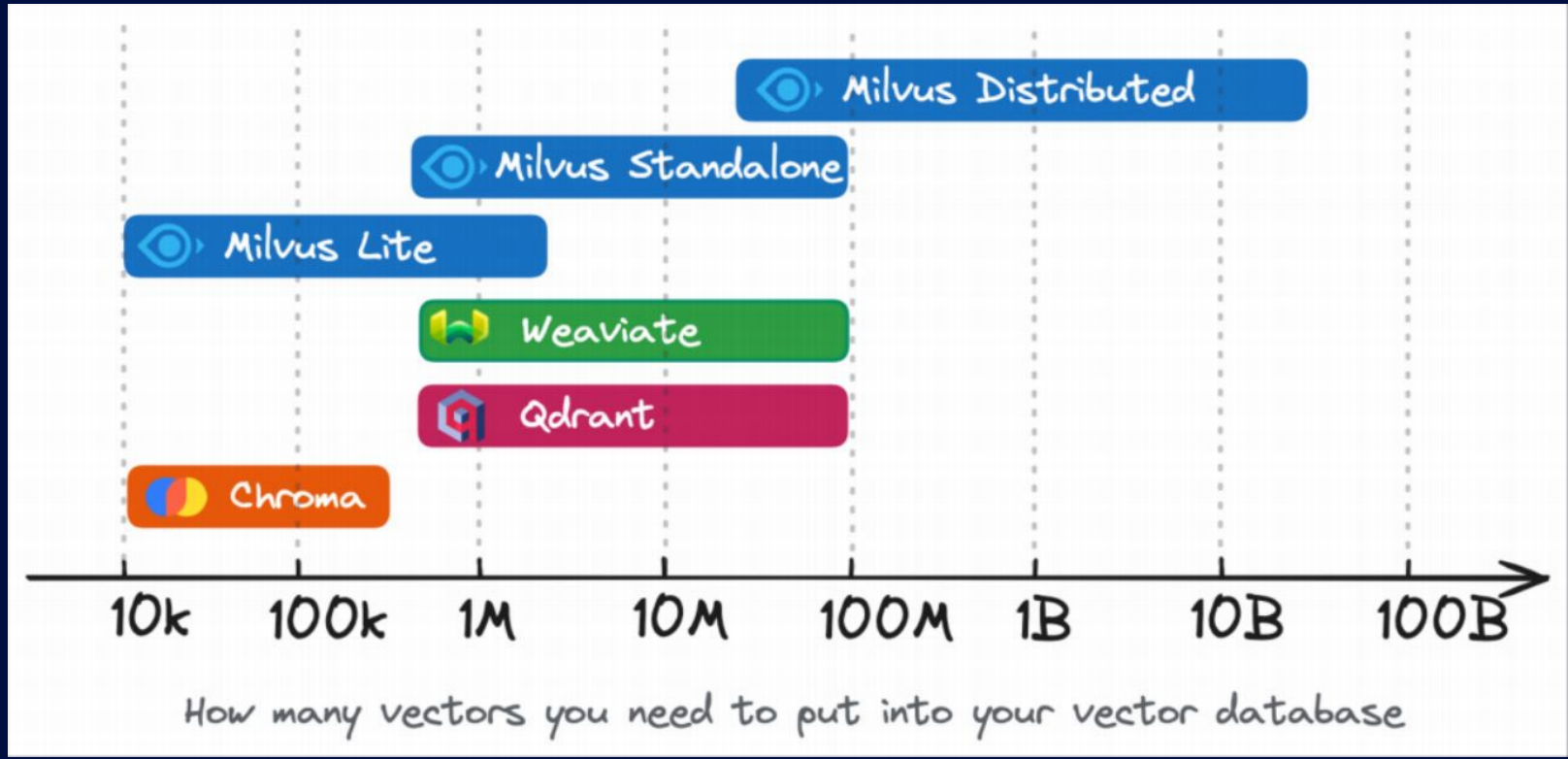


Tiered Storage

Bulk Import

GPU, Intel & ARM CPU support

Milvus vs other open source vector databases



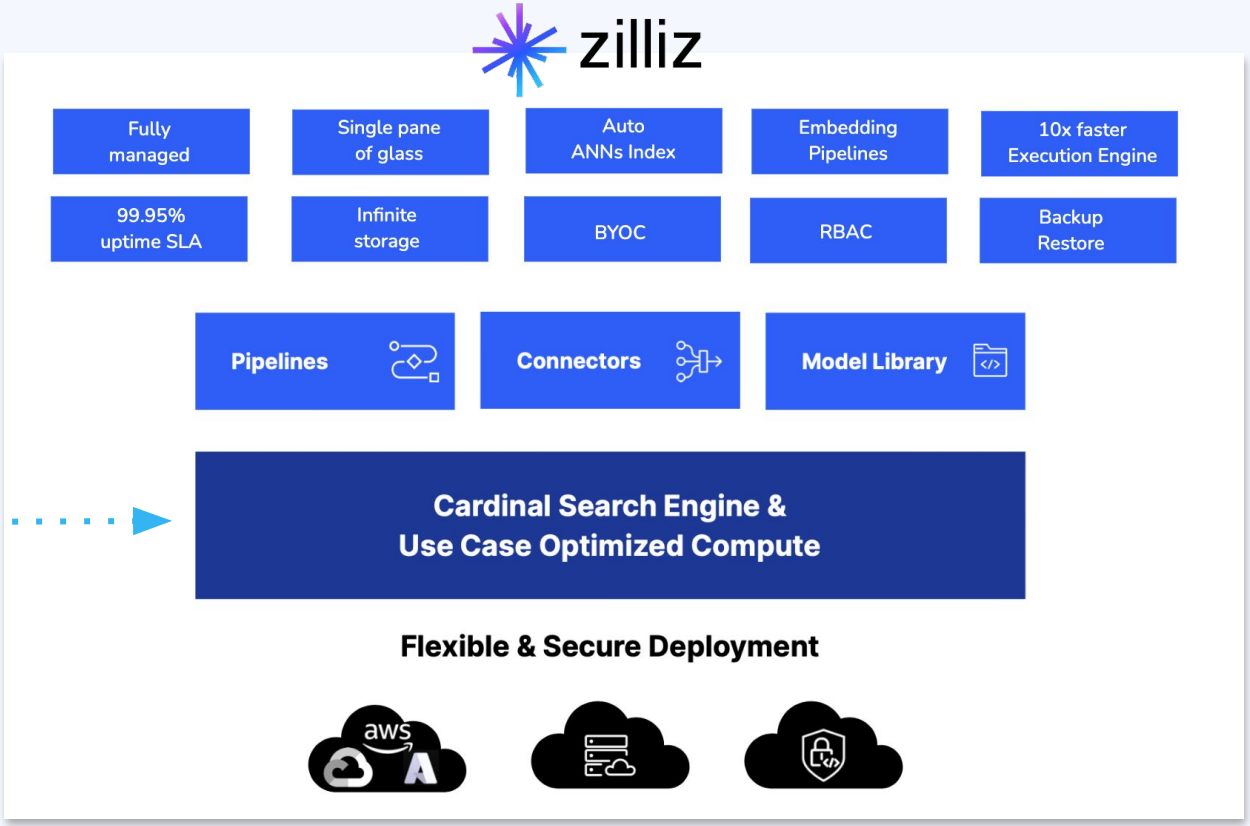
 |

zilliz

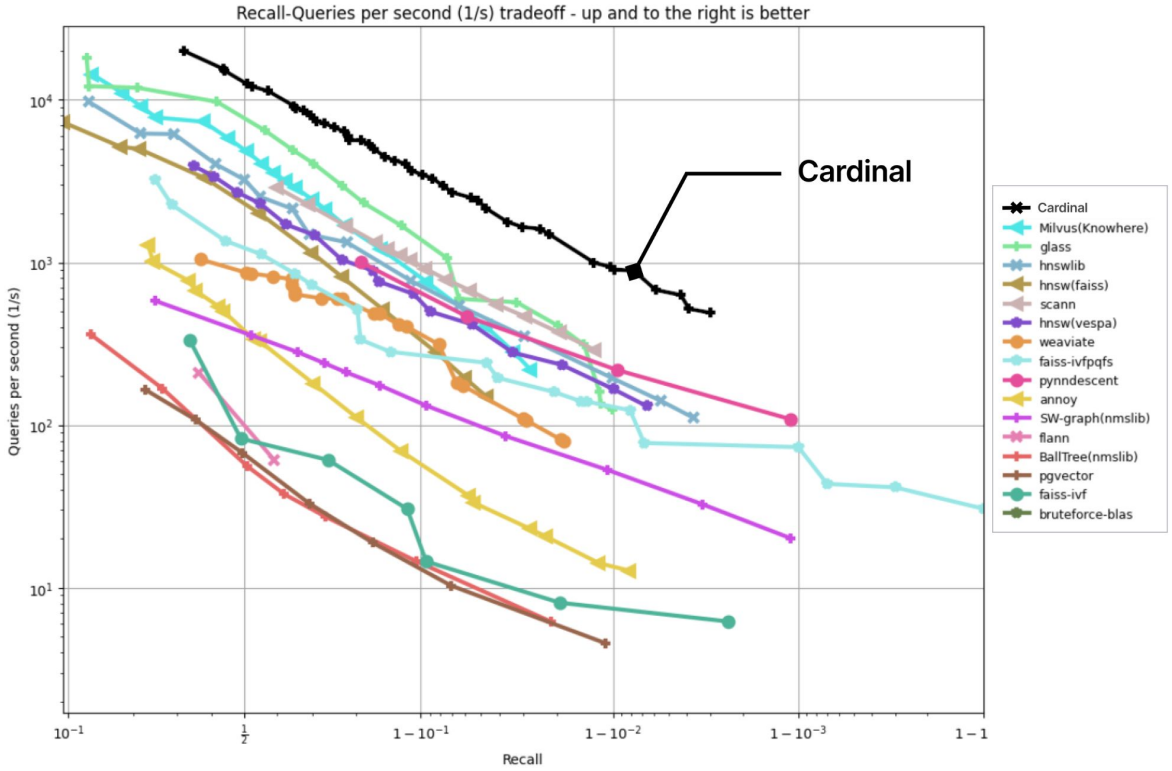
..but Zilliz Cloud is much more than OSS Milvus

Milvus

Stable Milvus versions are continuously deployed to Zilliz Cloud



ANN Benchmarks



Dataset: gist-960-euclidean

Open source vector db benchmark tool

https://zilliz.com/vector-database-benchmark-tool

Performance Ranking(QPS) | Performance Ranking(P99 Latency) | Cost Ranking

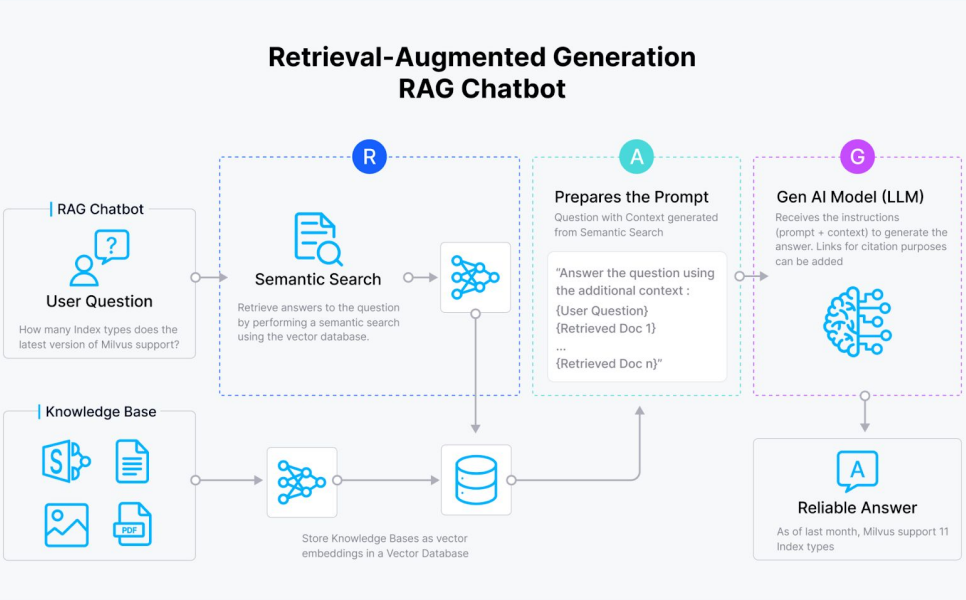
Rankings	Databases with different hardware resources	QPS Scores	QPS/Recall Medium OpenAI None Filter	QPS/Recall Medium OpenAI Low Filter	QPS/Recall Medium OpenAI High Filter	QPS/Recall Medium Cohere None Filter	QPS/Recall Medium Cohere Low Filter	QPS/Recall Medium Cohere High Filter
1	ZillizCloud-8cu-perf-(Jan-2024)	100	5115.53 / 0.947	3685.077 / 0.974	4742.162 / 0.994	6054.443 / 0.916	4104.26 / 0.951	4252.127 / 0.996
2	ZillizCloud-8cu-perf	41.9361	1871 / 0.96	1583 / 0.984	2345 / 1	2884.689 / 0.88	1689.58 / 0.949	1517.679 / 1
3	Milvus-16c64g-hnsw	23.5105	722.032 / 0.976	599.421 / 0.996	2098.211 / 1	1258.704 / 0.98	1075.878 / 0.98	1494.849 / 1
4	ZillizCloud-1cu-perf-(Jan-2024)	15.7792	633.603 / 0.919	467.58 / 0.99	1509.329 / 1	873.371 / 0.948	571.426 / 0.967	1156.29 / 0.999
5	QdrantCloud-4c16g-5node	14.7192	626.524 / 0.995	434.406 / 0.918	975.25 / 0.994	789.123 / 0.94	544.62 / 0.977	930.916 / 0.997
6	ZillizCloud-2cu-cap-(Jan-2024)	10.967	503.228 / 0.968	413.323 / 0.981	730.7 / 0.959	579.942 / 0.921	467.179 / 0.97	596.794 / 0.969
7	Pinecone-p2.x1-8node	9.0181	379.972 / 0.982	303.8 / 0.948	584 / 1	537.498 / 0.89	425.253 / 0.969	431.751 / 1
8	ZillizCloud-2cu-cap	8.8043	322.7 / 0.948	303.255 / 0.988	526.885 / 1	536.073 / 0.973	372.047 / 0.89	427.523 / 1
9	Milvus-4c16g-disk	8.7228	321.605 / 0.989	287 / 0.987	445.329 / 1	516.27 / 0.946	354.842 / 0.98	411.765 / 0.997
10	ZillizCloud-1cu-perf	8.3082	297.5 / 0.974	240.036 / 0.982	425.549 / 0.994	392.883 / 0.958	343.82 / 0.968	397.054 / 1

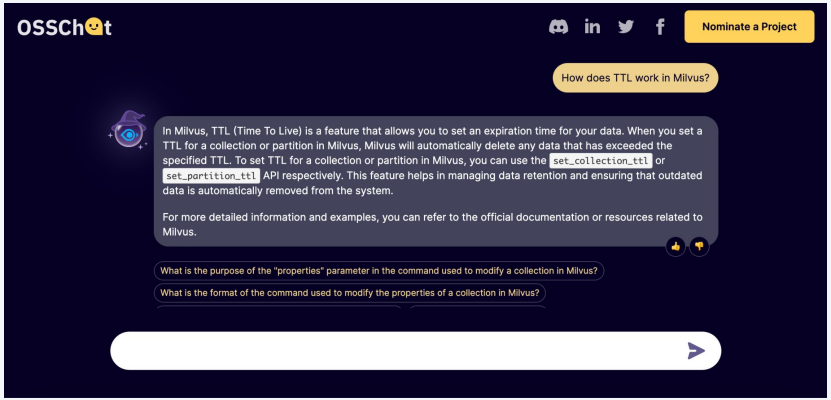
Milvus use cases

RAG

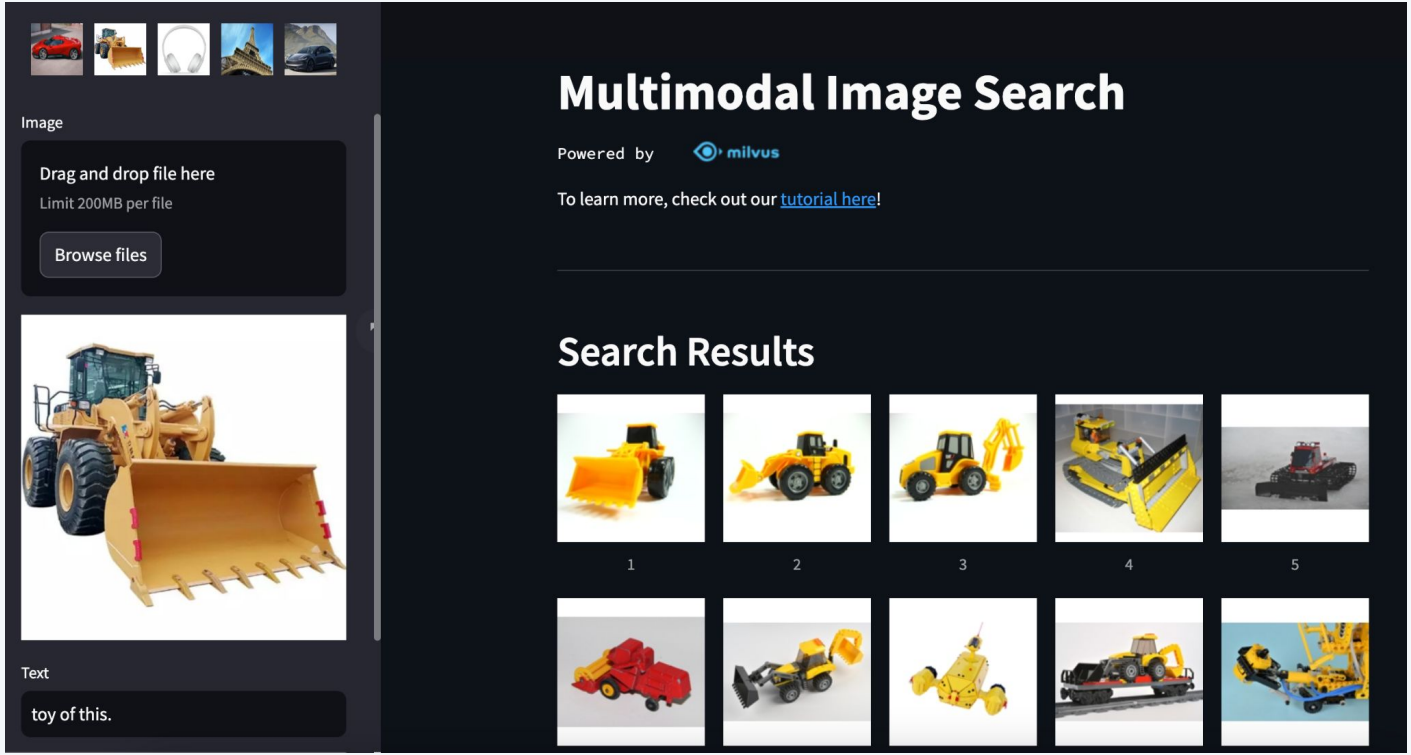
Further break-down

- Internal document Q&A
  - Query intention understanding
  - Tabel / figure processing
  - PDF parsing
- Customer Service
  - Rule based question routing
  - Observability
- Companionship
  - Memory management
  - Multi-tenancy
  - Relationship of characters
  - Cost reduction
- Marketing Campaign
  - Personalization
  - Co-pilotting
  - Quality assurance

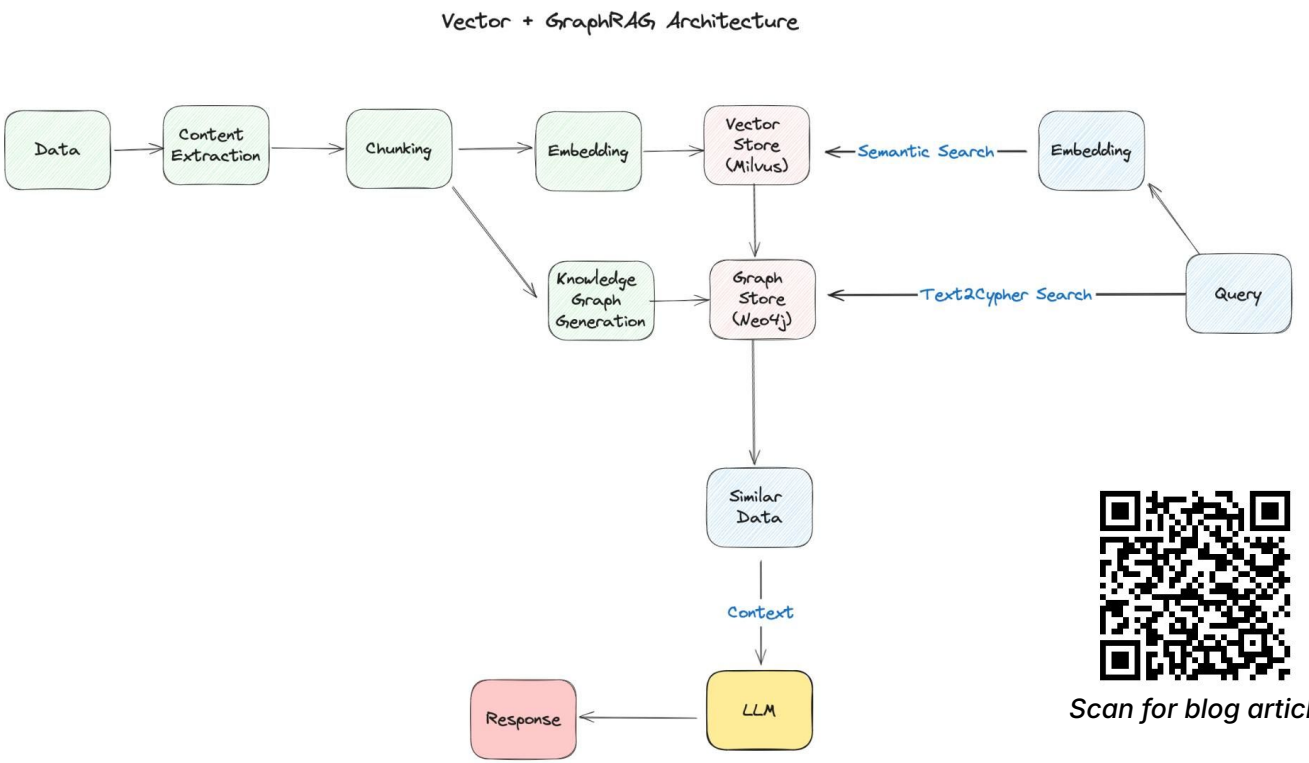




Multi-modal Retrieval / Visual Search

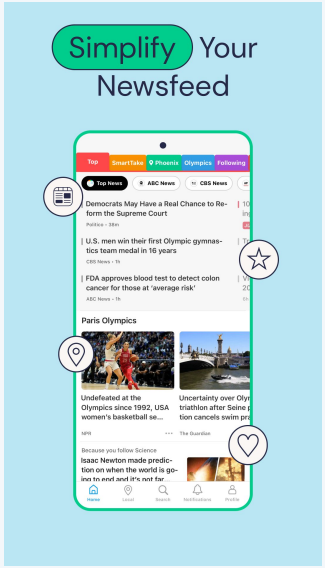


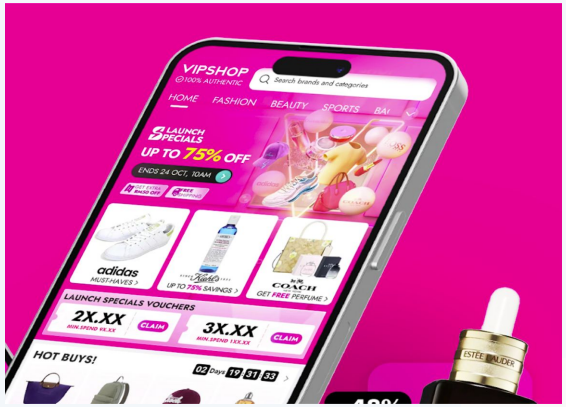
GraphAgent with Milvus and Neo4j

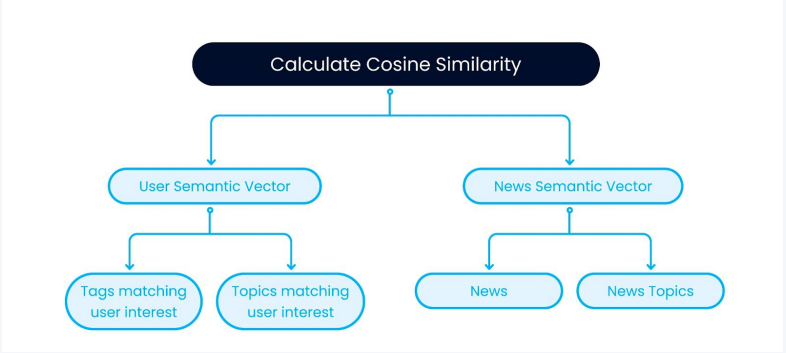


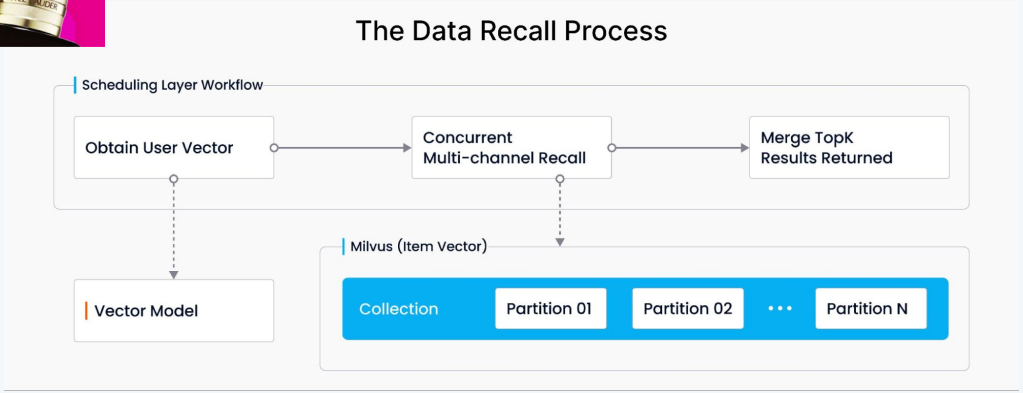
Scan for blog article

Recommender System









Demo time

Useful resources

- Forrester Wave Report: Vector Databases, Q3 2024
- Interactive Vector Database comparisons
- Zilliz Integration Hub
- Milvus demo notebooks

@MILVUSAP

Discord

GitHub

Telegram
Milvus Asia Pacific