zilliz

Discussion on Vector Databases, Unstructured Data and AI

Tim Spann | Zilliz

Speaker

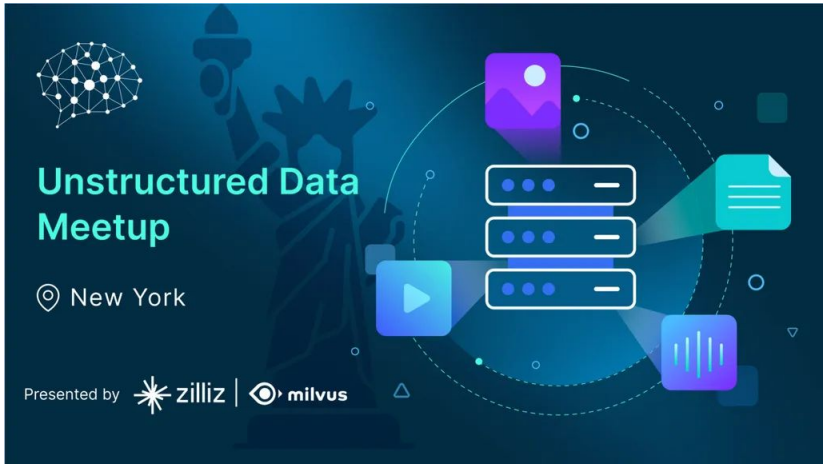
Tim Spann

Principal Developer Advocate, Zilliz

tim.spann@zilliz.com
https://www.linkedin.com/in/timothyspann/
https://x.com/paasdev
https://github.com/tspannhw

zilliz

Unstructured Data Meetup



https://www.meetup.com/unstructured-data-meetup-new-york/

This meetup is for people working in unstructured data. Speakers will come present about related topics such as vector databases, LLMs, and managing data at scale. The intended audience of this group includes roles like machine learning engineers, data scientists, data engineers, software engineers, and PMs.This meetup was formerly Milvus Meetup, and is sponsored by Zilliz maintainers of Milvus.

zilliz

milvus

LF AI & DATA

Easy Setup

Pip-install to start coding in a notebook within seconds.

Reusable Code

Write once, and deploy with one line of code into the production environment

Integration

Plug into OpenAI, Langchain, LlmaIndex, and many more

Feature-rich

Dense & sparse embeddings, filtering, reranking and beyond

Milvus is **an open-source vector database** for **GenAI** projects. `pip install` on your laptop, plug into popular AI dev tools, and push to production with a single line of code.

27.5K+
GitHub Stars

2,700+
Forks

25M+
Downloads

250+
Contributors

The evolution of AI made the semantic search of unstructured data possible

Search by Probability

Statistical analyses of common datasets established the foundation for processing unstructured data, e.g. NLP, and image classification

AI Model Breakthrough

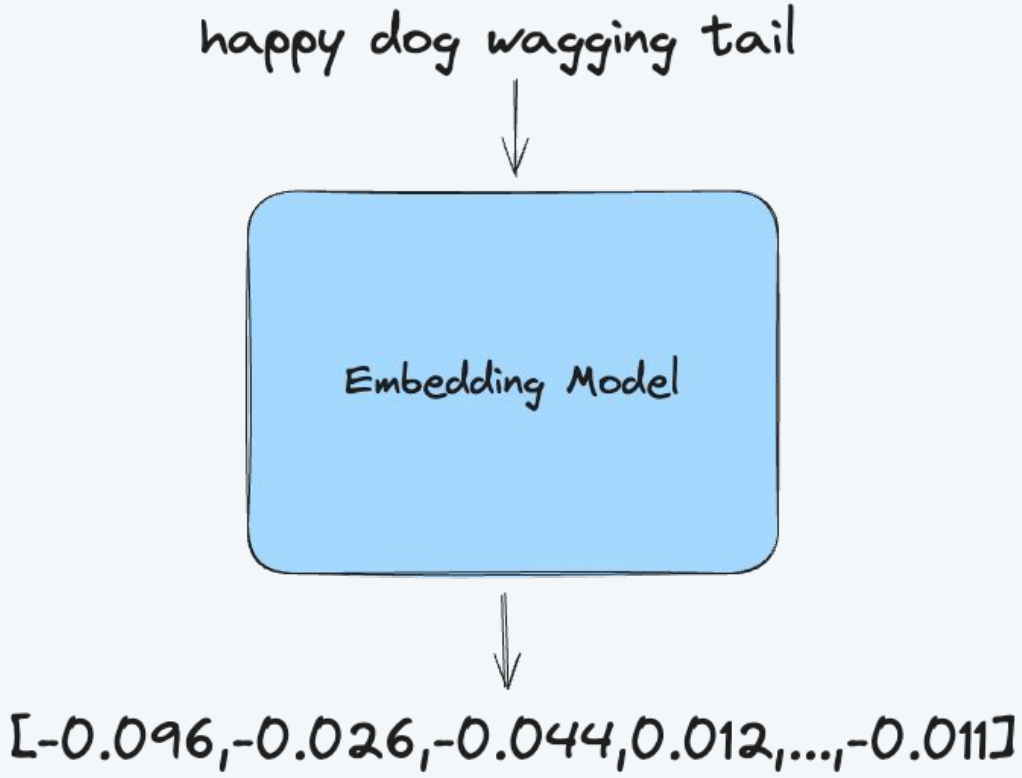
The advancements in BERT, ViT, CBT etc. have revolutionized semantic analysis across unstructured data

Vectorization

Word2Vec, CNNs, Deep Speech pioneered unstructured data embeddings, mapping the words, images, videos into high-dimensional vectors

zilliz

What your data looks like



zilliz

This new AI breakthrough requires new databases to fully unleash its potential

Support multiple use case types

Accommodate diverse data requirements, enhancing flexibility and effectiveness in varied operational contexts

Scale as needed

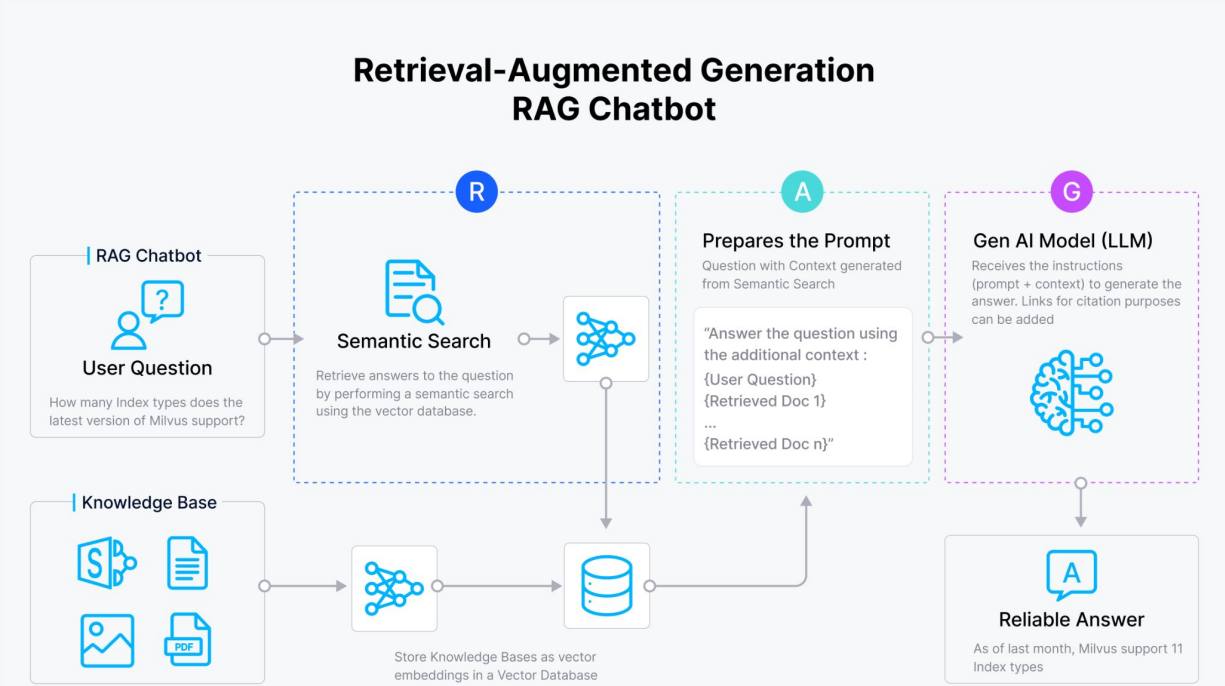
Enable robust handling of expanding data volumes and search demands

Highly performant

Ensures swift and accurate query responses, crucial for optimal user experience

zilliz

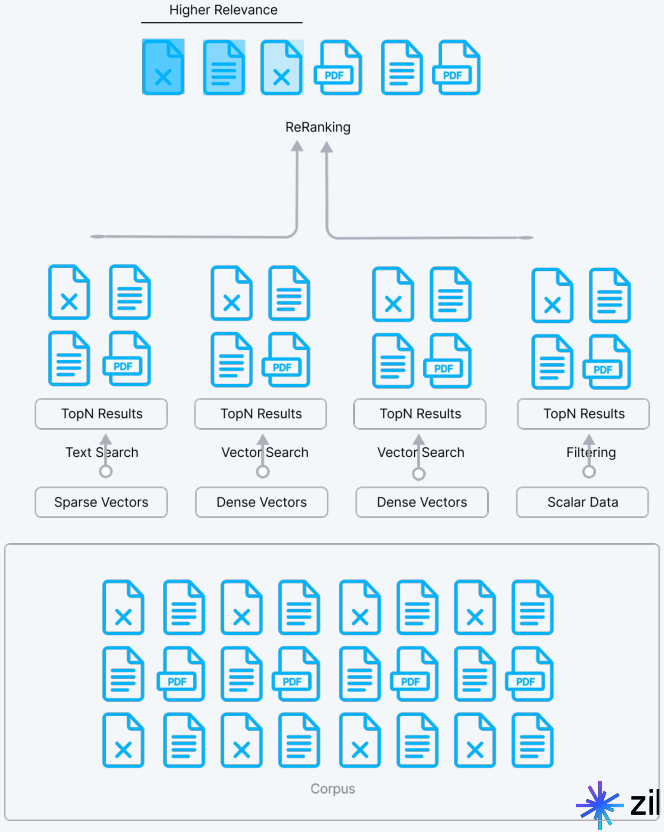
Vector Databases are core component for Retrieval Augmented Generation (RAG)



zilliz

...different types of data and schemas needs to be thoroughly planned ahead of time

Field Name	Type	Description	Example Value
chunkID	Int64	Primary key, uniquely identifies different parts of a document	123456789
userID	Int64	Partition key, data partitioning is based on userID to ensure searches occur within a single userID	987654321
docID	Int64	Unique identifier for a document, used to associate different chunks of the same document	555666777
chunkData	varchar	A part of the document, containing several hundred bytes of text	"This is a part of the document..."
dynamicParams	JSON	Stores dynamic parameters of the document, such as name, source URL, etc.	{"name": "Example Document", "source": "example.com"}
sparseVector	Specific format	Data representing a sparse vector. Specific format will have non-zero values only in certain positions to represent sparsity.	[0.1, 0, 0, 0.8, 0.4, 0]
denseVector	Specific format	Data representing a dense vector. Specific format will have a fixed number of dimensions with values in each.	[0.2, 0.3, 0.4, 0.11]



...powers searches across various types of unstructured data

Retrieval Augmented Generation (RAG)

Expand LLMs' knowledge by incorporating external data sources into LLMs and your AI applications.

Recommender System

Match user behavior or content features with other similar ones to make effective recommendations.

Text/ Semantic Search

Search for semantically similar texts across vast amounts of natural language documents.

Image Similarity Search

Identify and search for visually similar images or objects from a vast collection of image libraries.

Video Similarity Search

Search for similar videos, scenes, or objects from extensive collections of video libraries.

Audio Similarity Search

Find similar audios in large datasets for tasks like genre classification or speech recognition

Molecular Similarity Search

Search for similar substructures, superstructures, and other structures for a specific molecule.

Anomaly Detection

Detect data points, events, and observations that deviate significantly from the usual pattern

Multimodal Similarity Search

Search over multiple types of data simultaneously, e.g. text and images

We've built technologies for various types of use cases

Index Types

Offer a diverse range of **11+ index types**, including popular ones like HNSW, IVF, PQ, and GPU index

Empower developers with tailored search optimizations, catering to specific performance and accuracy needs

Search Types

Provide diverse search types such as **top-K ANN, Range ANN, hybrid ANN** and metadata filtering

Enable unparalleled query flexibility and accuracy, allowing developers to tailor their data retrieval needs

Multi-tenancy

Enable **multi-tenancy** through collection and partition management

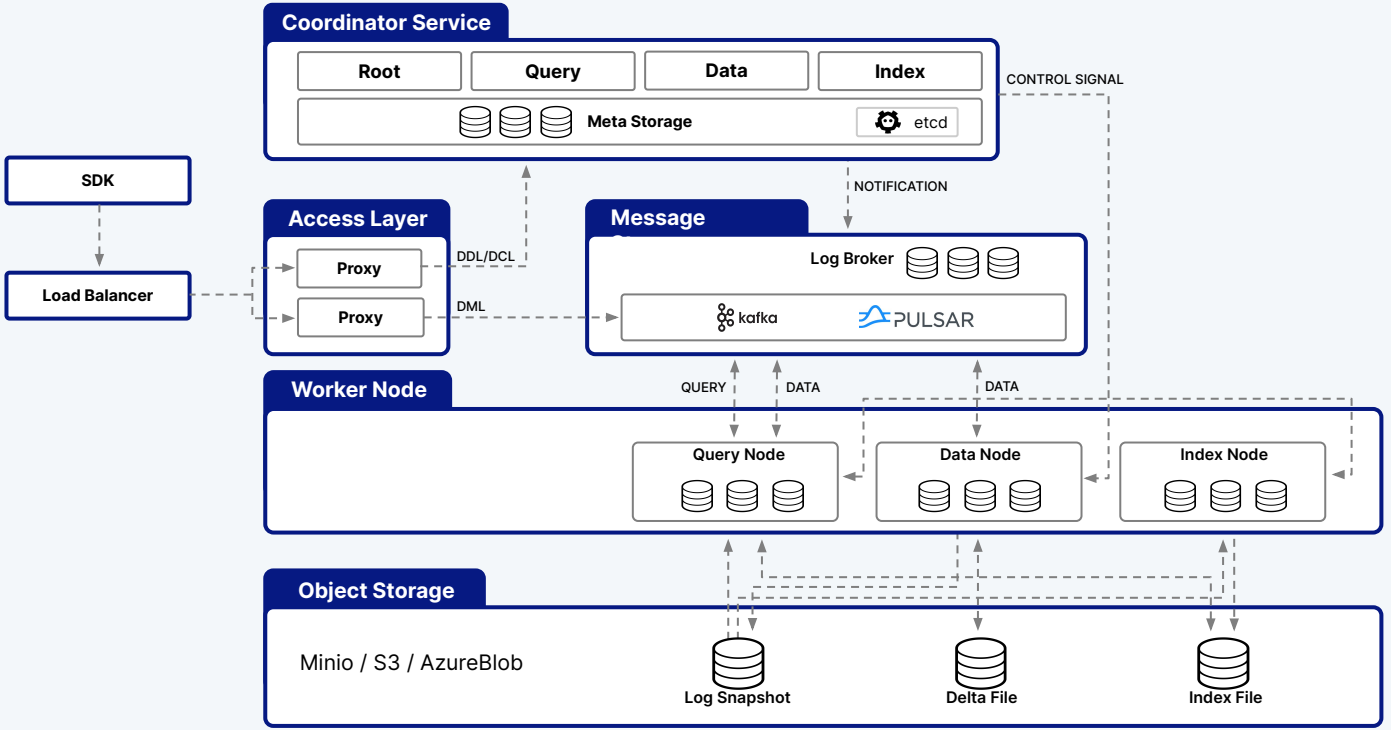
Allow for efficient resource utilization and customizable data segregation, ensuring secure and isolated data handling for each tenant

Compute Types

Support different types of compute powers, such as **AVX512, Neon for SIMD execution, quantization & cache-aware optimization, and GPU**

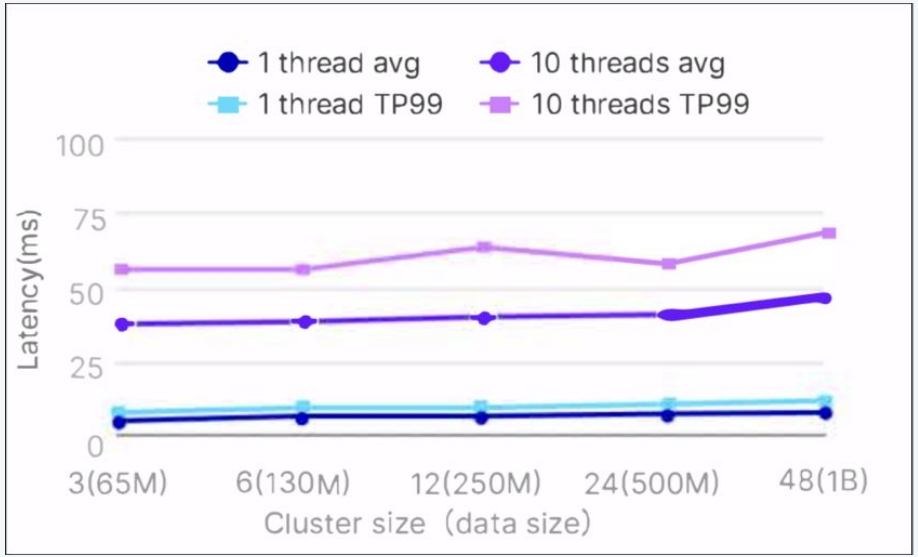
Leverage specific strengths of each hardware type efficiently, ensuring high-speed processing and cost-effective scalability for diverse application needs

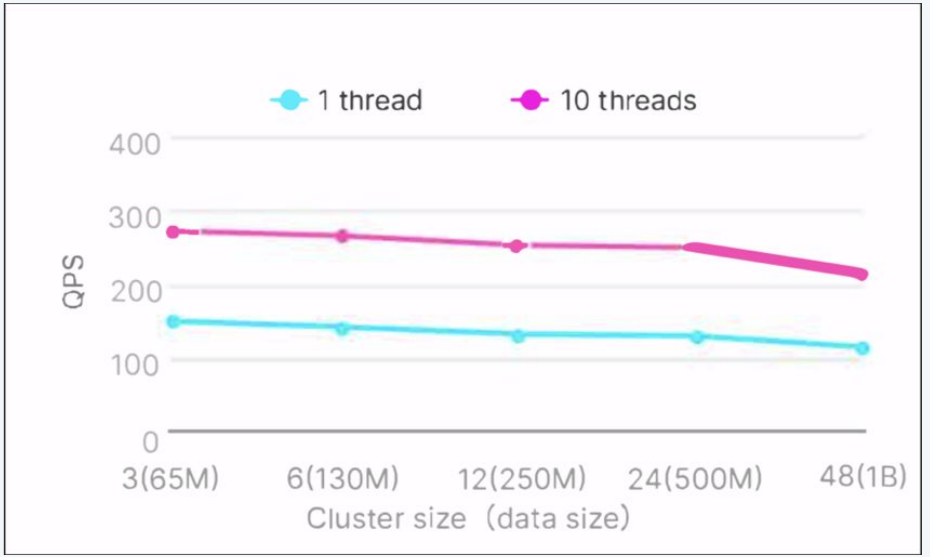
Milvus' fully distributed architecture is designed scalability and performance



 |

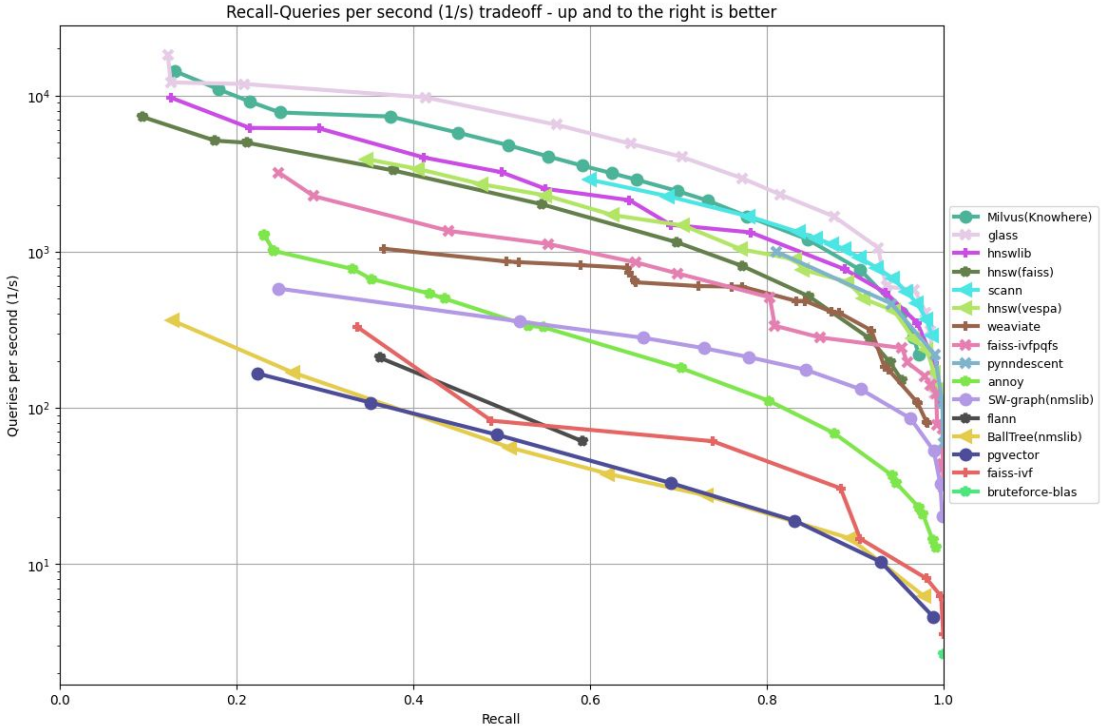
Tests shows consistent query performance when scaled from 65 million to 1 billion vectors





zilliz

ANN Benchmark has recognized Milvus as the performance leader among vector database players



 |

zilliz

We provide deployment flexibility for different operational, security and compliance requirements

SELF MANAGED SOFTWARE

Milvus

Most widely-adopted open source vector database

Self hosted on any machine with community support

Local Docker K8s

FULLY MANAGED SERVICE

Zilliz Cloud

Milvus Re-engineered for the Cloud

Available on the leading public clouds

aws Google Cloud Azure

BRING YOUR OWN CLOUD

Zilliz BYOC

Enterprise-ready Milvus for Private VPCs

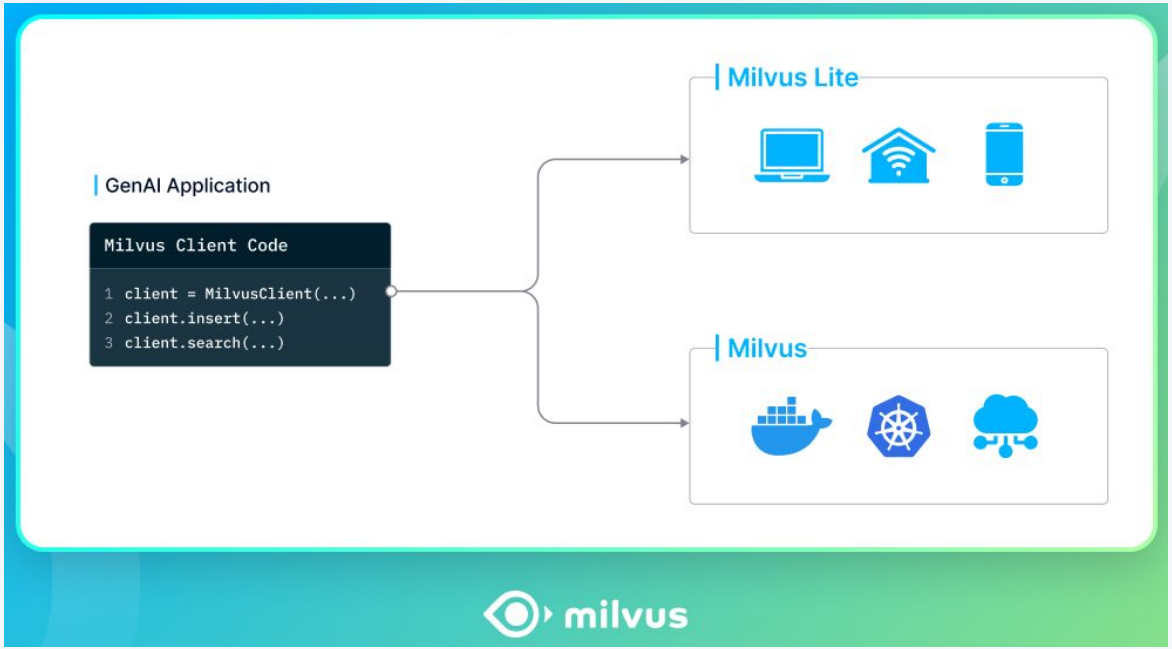
Deploy in your virtual private cloud

aws Google Cloud *Coming Soon!* Azure *Coming Soon!*

zilliz

Milvus Lite

```
pip install pymilvus
```



zilliz

Embeddings Models

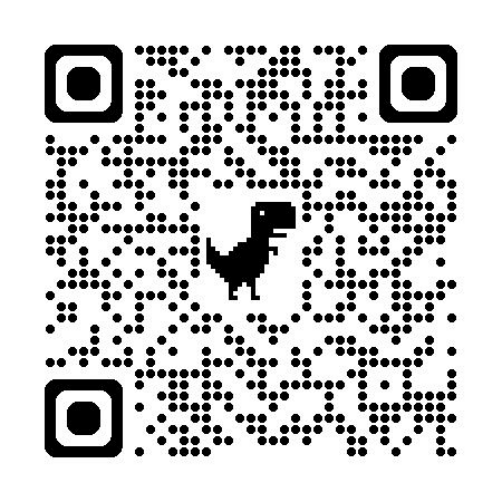
zilliz

Questions?

Give Milvus a Star!

github.com/milvus-io/

Code on Github



github.com/tspannhw

zilliz