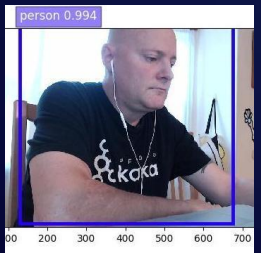
zilliz

AICamp | NEW YORK
IN PERSON MEETUP

AI, LLMs, ML and Data Meetup

Unstructured Data and Vector Databases

June 20, 2024



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
https://github.com/milvus-io/milvus

Agenda

01

Introduction

Unstructured data, vector databases, traditional databases, similarity search

02

Vectors

Where, What, How, Why Vectors? We'll cover a Vector Database Architecture

03

Introducing Milvus

What drives Milvus' Emergence as the most widely adopted vector database

Introduction

 |

zilliz

Why Vector Databases?

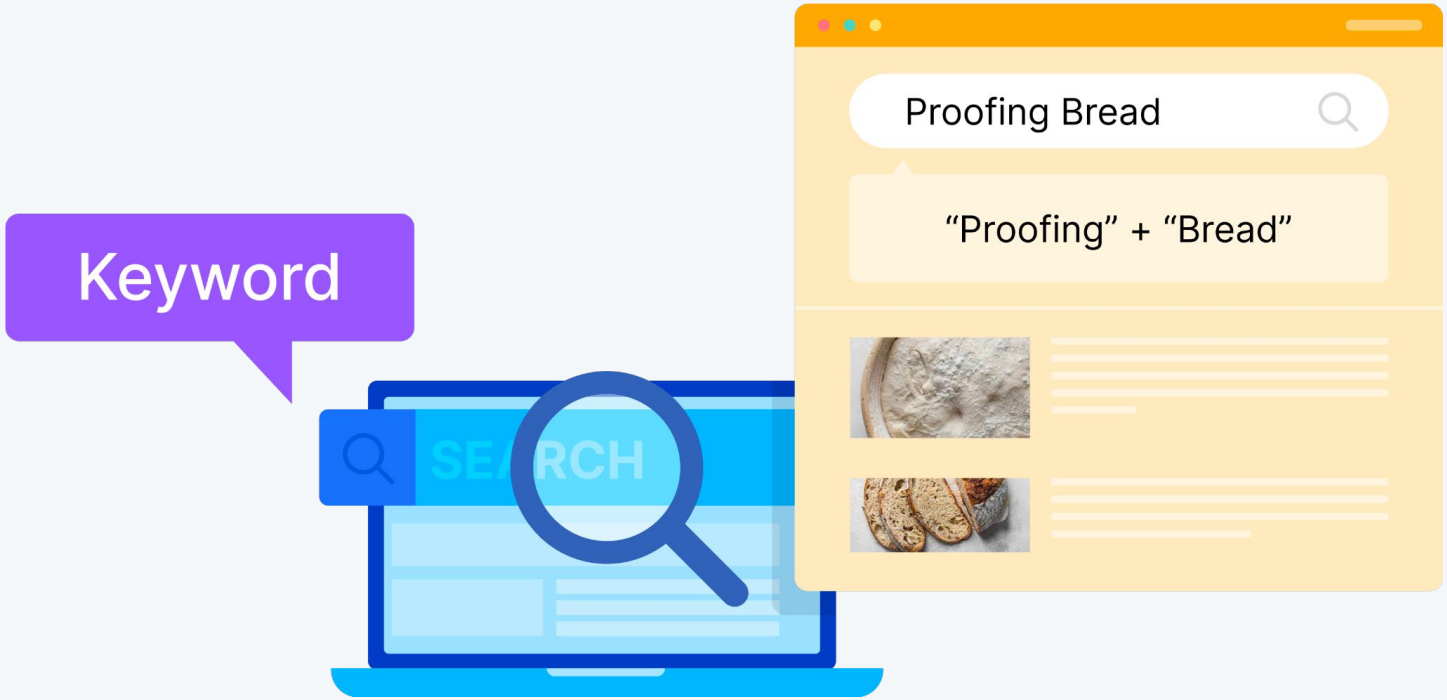
- Unstructured Data is 80% of data

- Vector Databases are the only type of database that can work with unstructured data

- Examples of Unstructured Data include text, images, videos, audio, etc

zilliz

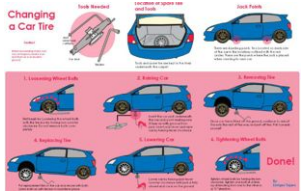
Traditional databases were built on exact search



 |

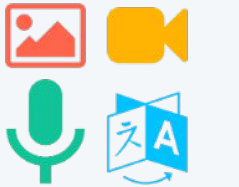
zilliz

...which misses context, semantic meaning, and user intent

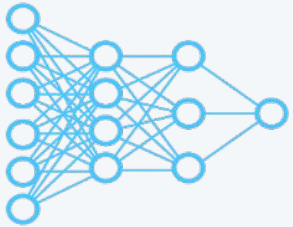
Apple	Rising dough	Change car tire
	Rising Dough ✔	
VS.	*VS.*	*VS.*
	Proofing Bread ❌	

zilliz

Where do Vectors Come From?



Knowledge Base (Documents)



Deep Learning Models

$$\begin{bmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,n} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,n} \\ V_{3,1} & V_{3,2} & \cdots & V_{3,n} \\ \vdots & \vdots & \vdots & \vdots \\ V_{n,1} & V_{n,2} & \cdots & V_{n,n} \end{bmatrix}$$

Vectors

zilliz

milvus

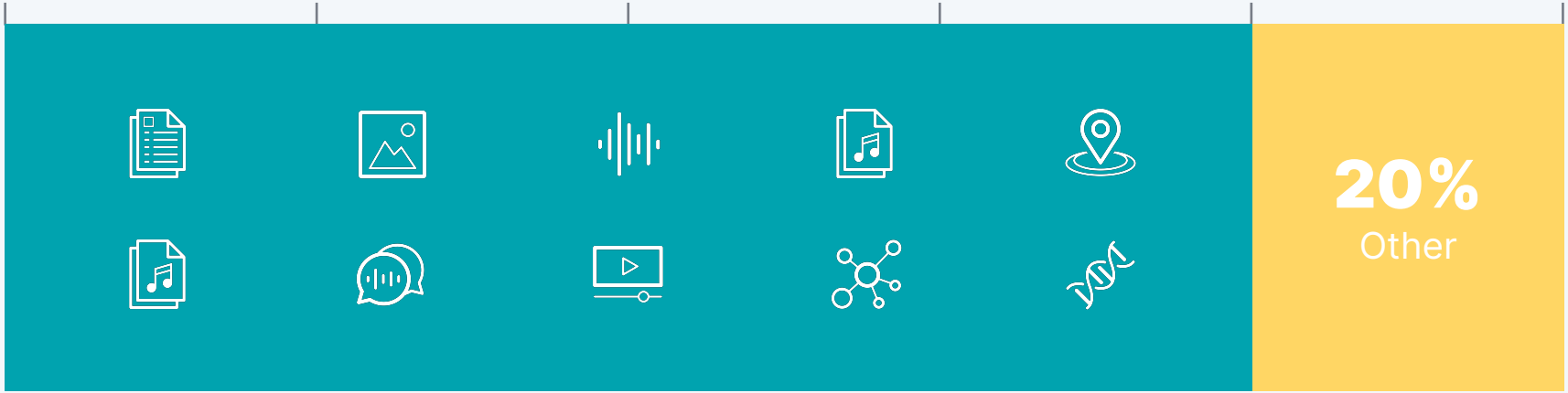
Vector Databases

 |

zilliz

…and cannot process increasingly growing unstructured data

80% **newly generated data in 2025 will be unstructured data**



*Data Source: The Digitization of the World by *IDC*

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

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

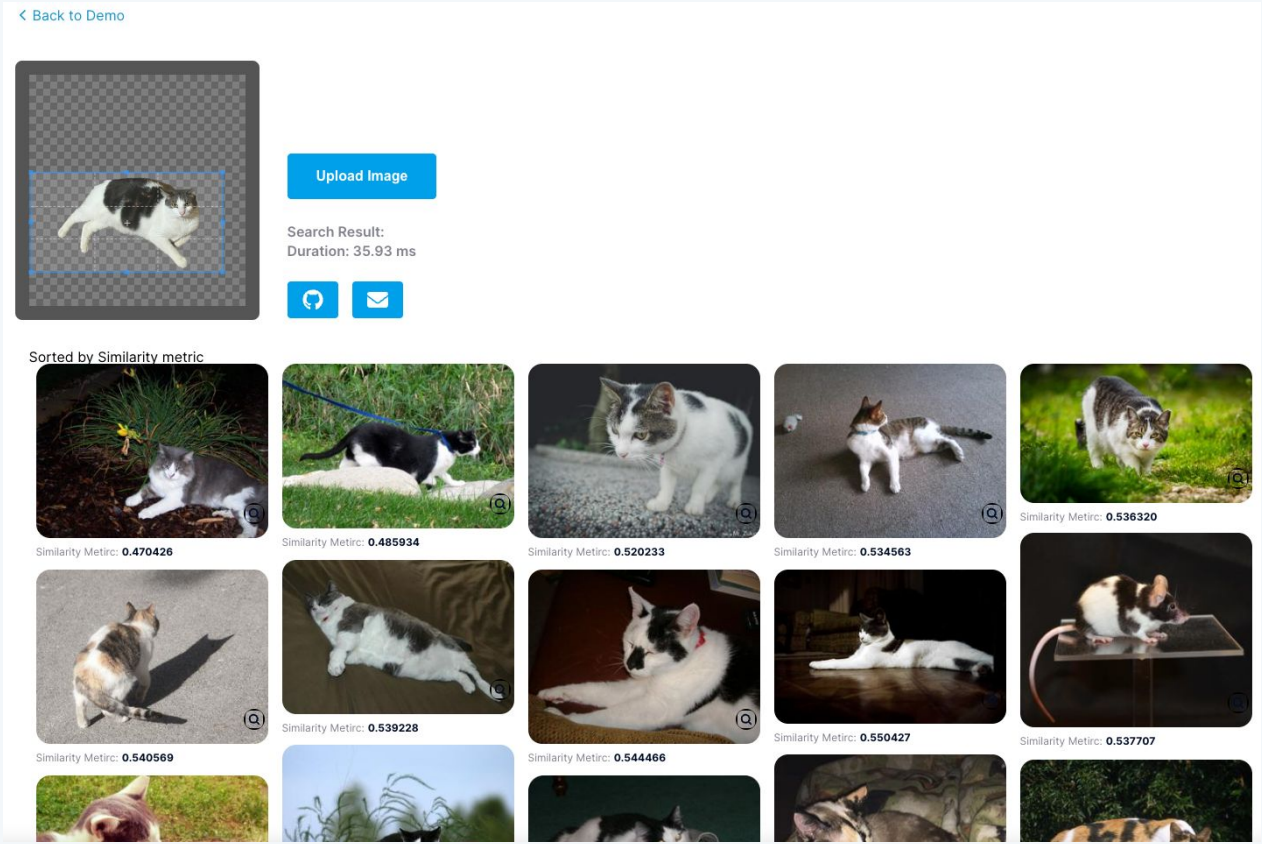
Highly performant

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

 |

zilliz

Show Me



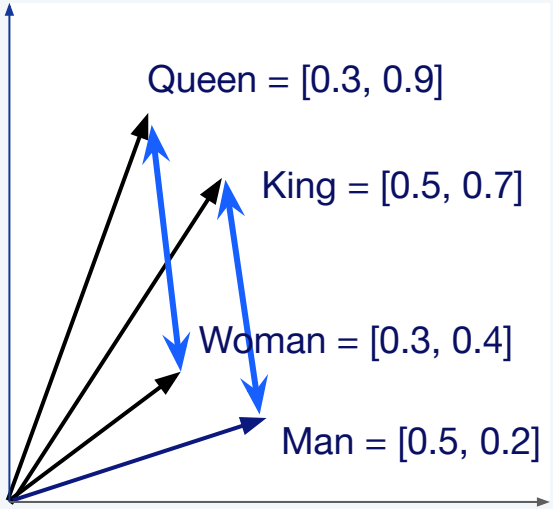
https://milvus.io/milvus-demos/reverse-image-search

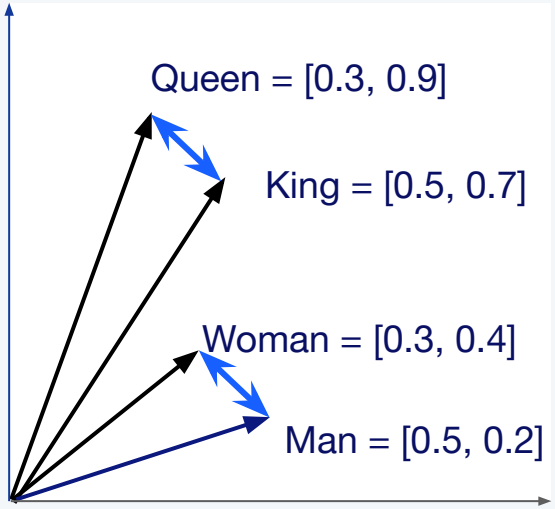
zilliz

03

How do Vector Databases Work?

Semantic Similarity





Queen - Woman + Man = King

Queen = [0.3, 0.9]
\- Woman = [0.3, 0.4]

[0.0, 0.5]
\+ Man = [0.5, 0.2]

King = [0.5, 0.7]

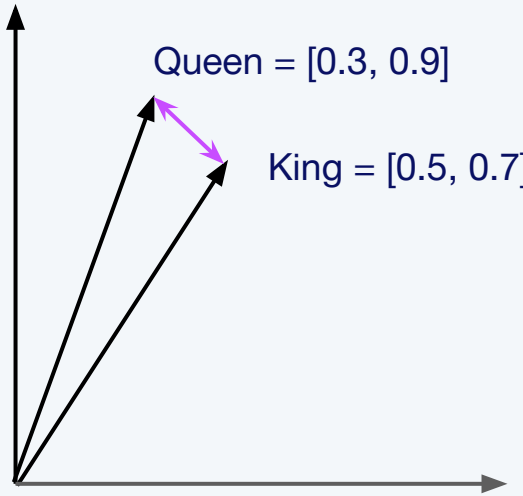
Image from Sutor et al

 |

zilliz

Vector Similarity Measures: L2 (Euclidean)

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$



d(Queen, King) = $\sqrt{(0.3-0.5)^2 + (0.9-0.7)^2}$

= $\sqrt{(0.2)^2 + (0.2)^2}$

= $\sqrt{0.04 + 0.04}$

= $\sqrt{0.08} \cong 0.28$

 |

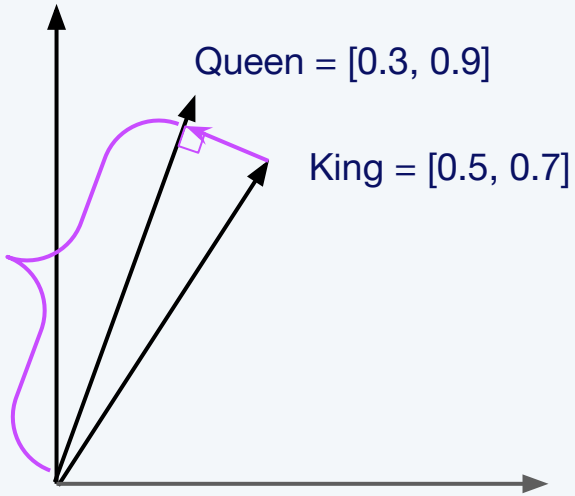
zilliz

Vector Similarity Measures: Inner Product (IP)

$$a \cdot b = \sum_{i=1}^{n} a_i b_i$$

Queen · King = (0.3*0.5) + (0.9*0.7)

= 0.15 + 0.63 = 0.78

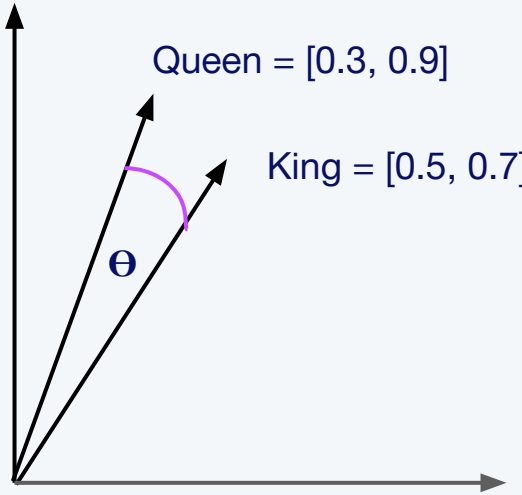


 |

zilliz

Vector Similarity Measures: Cosine

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$



$$\cos(\text{Queen, King}) = \frac{(0.3*0.5)+(0.9*0.7)}{\sqrt{0.3^2+0.9^2} * \sqrt{0.5^2+0.7^2}}$$

$$= \frac{0.15+0.63}{\sqrt{0.9} * \sqrt{0.74}}$$

$$= \frac{0.78}{\sqrt{0.666}}$$

$$\cong 0.03$$

 |

zilliz

Introducing Milvus

 |

zilliz

Why Not Use a SQL/NoSQL Database?

- Inefficiency in High-dimensional spaces
- Suboptimal Indexing
- Inadequate query support
- Lack of scalability
- Limited analytics capabilities
- Data conversion issues

TL;DR: Vector operations are **too computationally intensive** for traditional database infrastructures

zilliz

Why Not Use a Vector Search Library?

- Have to manually implement filtering
- Not optimized to take advantage of the latest hardware
- Unable to handle large scale data
- Lack of lifecycle management
- Inefficient indexing capabilities
- No built in safety mechanisms

TL;DR: Vector search libraries **lack the infrastructure** to help you **scale**, **deploy**, and **manage** your apps in production.

zilliz

What is Milvus ideal for?

Purpose-built to store, index and query vector embeddings from unstructured data **at scale**.

- Advanced filtering
- Hybrid search
- Durability and backups
- Replications/High Availability
- Sharding
- Aggregations
- Lifecycle management
- Multi-tenancy

- High query load
- High insertion/deletion
- Full precision/recall
- Accelerator support (GPU, FPGA)
- Billion-scale storage

 |

zilliz

We've built technologies for various types of use cases

Index Types

Offer a wide range of **15 indexes** support, including popular ones like HNSW, PQ, Binary, Sparse, DiskANN and GPU index

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

Search Types

Support multiple types such as **top-K ANN, Range ANN, sparse & dense, multi-vector, grouping,** and metadata **filtering**

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

Multi-tenancy

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

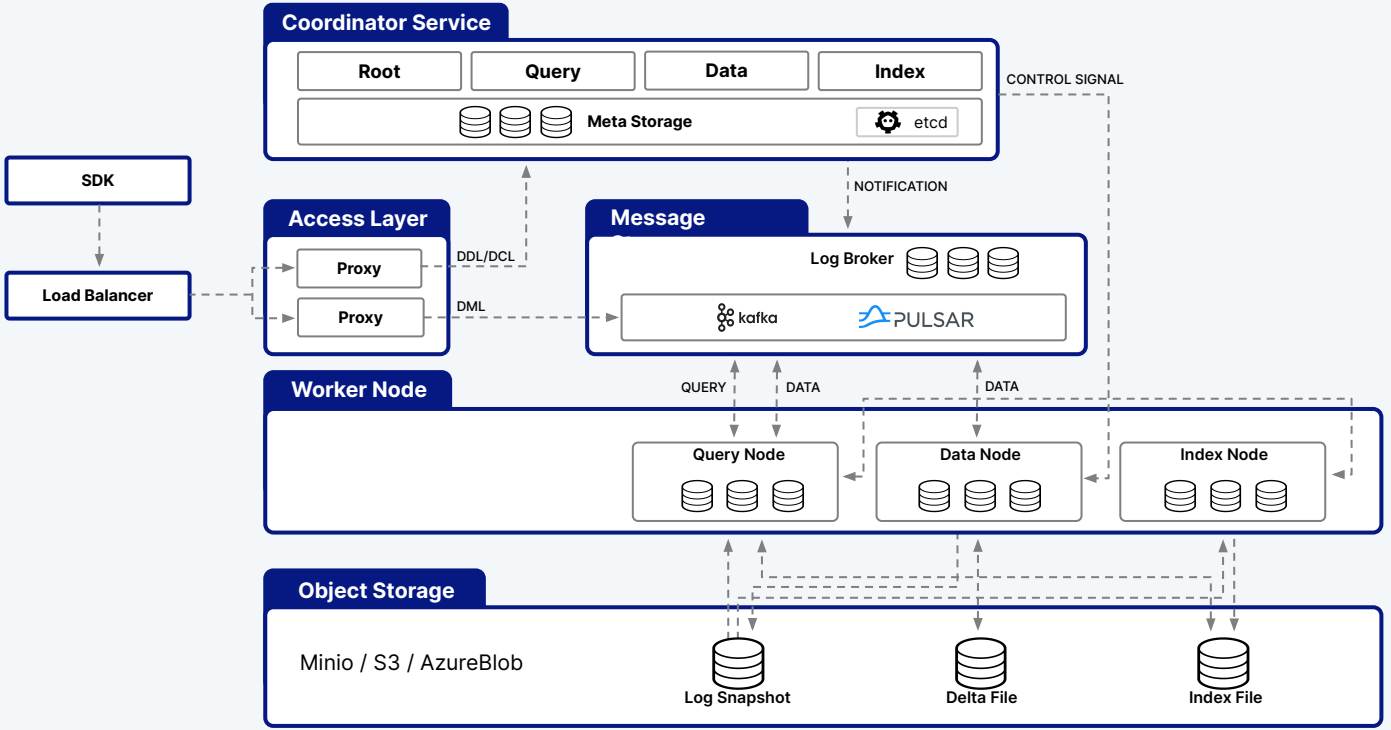
Allow for efficient resource utilization and customizable data segregation, ensuring secure and isolated data handling for each tenant

Compute Types

Designed for various compute powers, such as **AVX512, Neon for SIMD, quantization cache-aware optimization** and **GPU**

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

Milvus' fully distributed architecture is designed scalability and performance



 |

Milvus: From Dev to Prod
AI Powered Search made easy

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.

contributors | stars | downloads | forks

 |

...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.

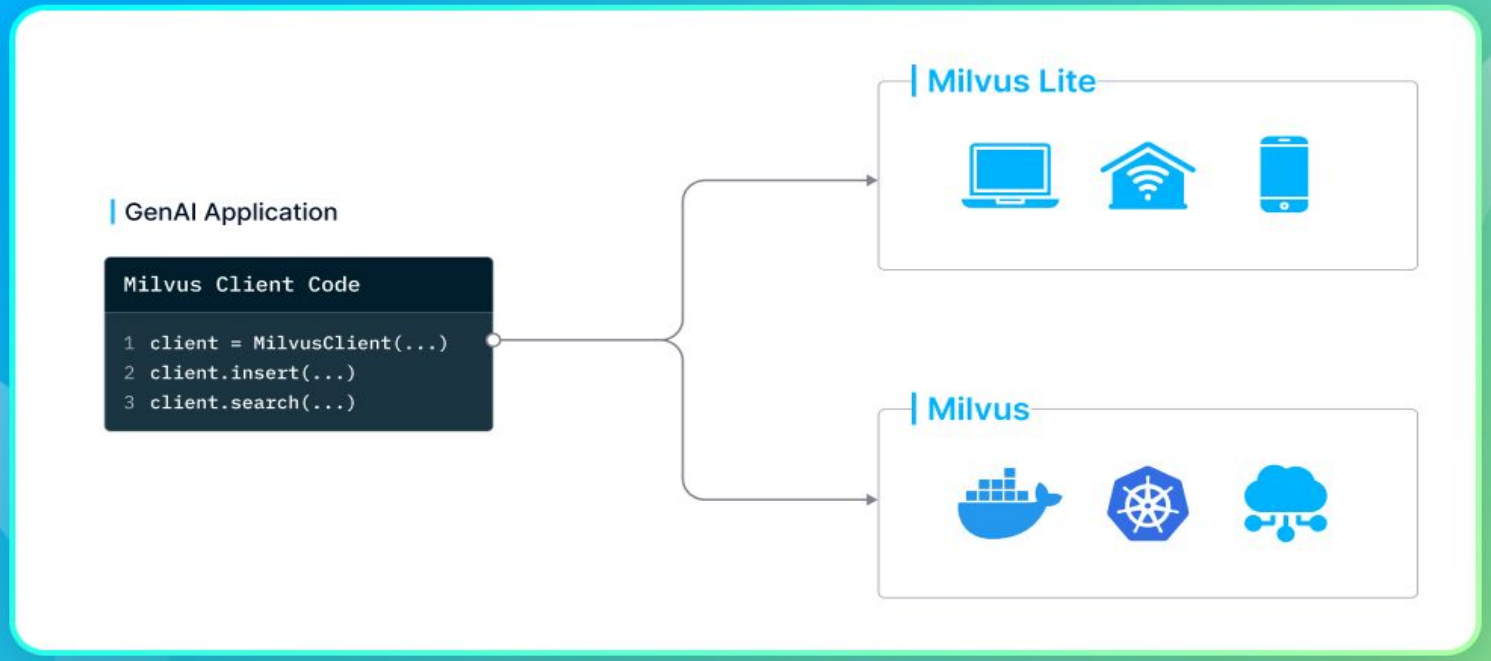
Anomaly Detection

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

Build Once
Deploy Anywhere



milvus

 |

zilliz

Up to 100 billion vectors with K8s!

You can install Milvus Operator in either of the following ways:

- With Helm
- With kubectl

Install with Helm

Run the following command to install Milvus Operator with Helm.

```
$ helm install milvus-operator \
  -n milvus-operator --create-namespace \
  --wait --wait-for-jobs \
  https://github.com/zilliztech/milvus-operator/releases/download/v0.9.15/milvus-operator-0.9.15.tg
```

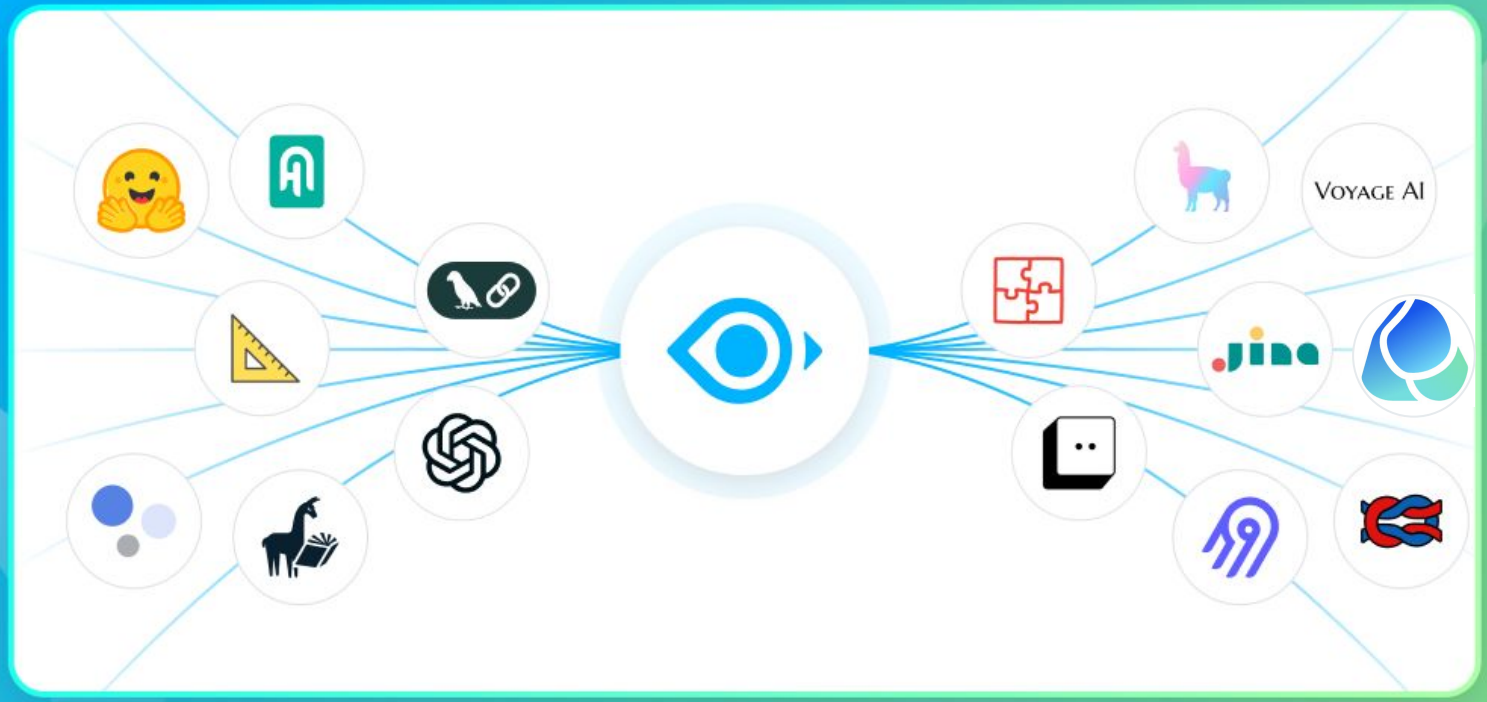
You will see the output similar to the following after the installation process ends.

```
NAME: milvus-operator
LAST DEPLOYED: Thu Jul  7 13:18:40 2022
NAMESPACE: milvus-operator
STATUS: deployed
REVISION: 1
TEST SUITE: None
NOTES:
Milvus Operator Is Starting, use `kubectl get -n milvus-operator deploy/milvus-operator` to check i
If Operator not started successfully, check the checker's log with `kubectl -n milvus-operator logs
Full Installation doc can be found in https://github.com/zilliztech/milvus-operator/blob/main/docs/
Quick start with `kubectl apply -f https://raw.githubusercontent.com/zilliztech/milvus-operator/mai
More samples can be found in https://github.com/zilliztech/milvus-operator/tree/main/config/samples
CRD Documentation can be found in https://github.com/zilliztech/milvus-operator/tree/main/docs/CRD
```

 |

zilliz

Seamless integration with all popular AI toolkits



milvus

Vector Database Resources

Give Milvus a Star!

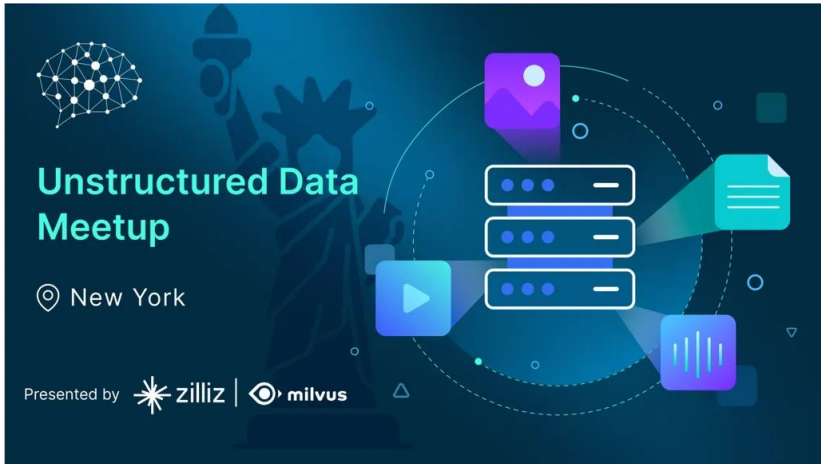
Chat with me on Discord!

https://github.com/milvus-io/milvus

 |

zilliz

Unstructured Data Meetup



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

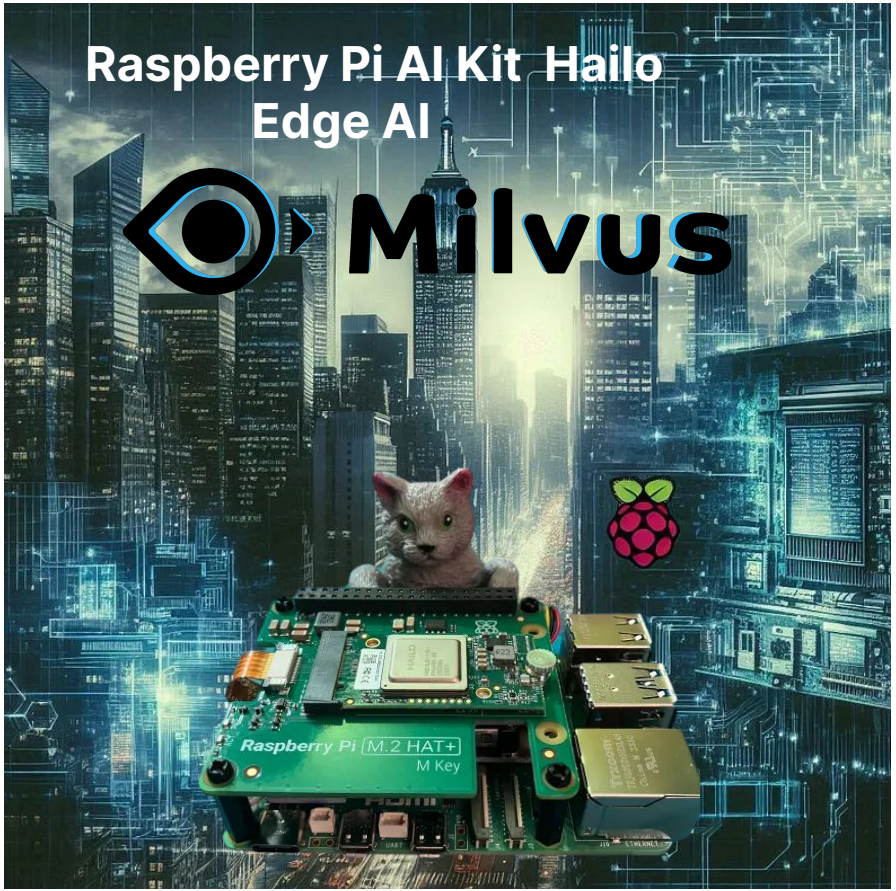
This meetup is for people working in unstructured data. Speakers will come present about related topics such as vector databases, LLMs, and managing data at scale. The intended audience of this group includes roles like machine learning engineers, data scientists, data engineers, software engineers,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

This meetup was formerly Milvus Meetup, and is sponsored by Zilliz maintainers of Milvus.

RESOURCES

zilliz

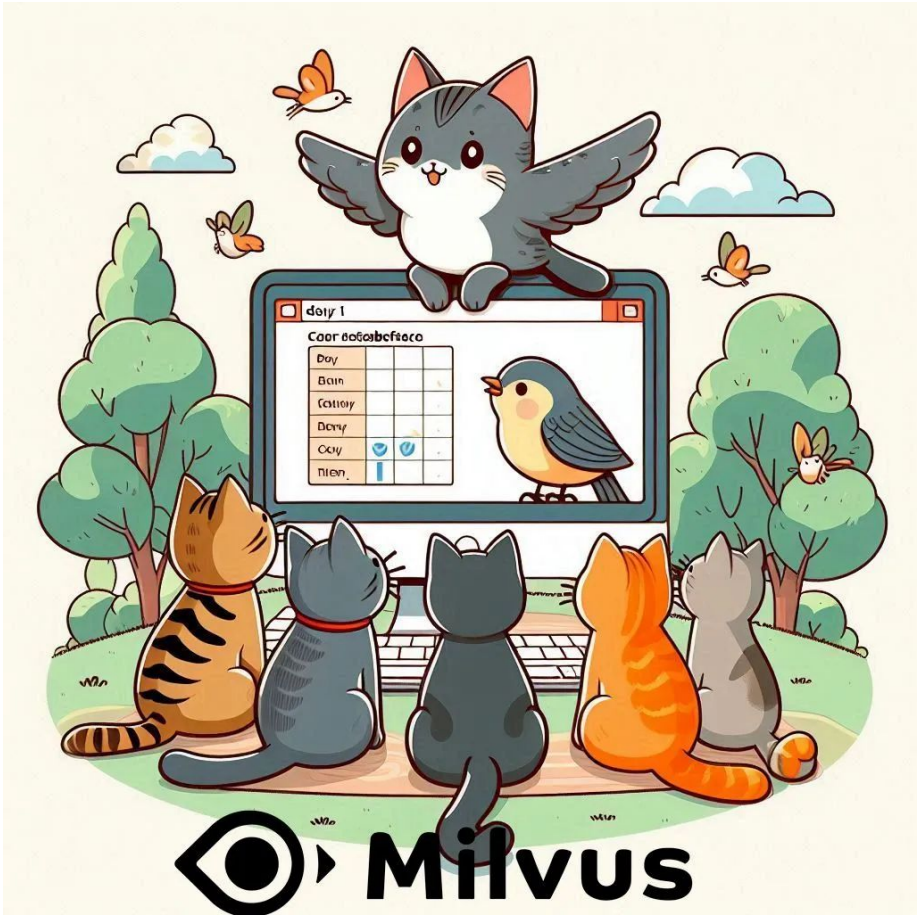

Learning Vector



https://medium.com/@tspann/unstructured-data-processing-with-a-raspberry-pi-ai-kit-c959dd7fff47

https://medium.com/@tspann/unstructured-street-data-in-new-york-8d3cde0a1e5b

https://medium.com/@tspann/not-every-field-is-just-text-numbers-or-vectors-976231e90e4d



https://medium.com/@tspann/shining-some-light-on-the-new-milvus-lite-5a0565eb5dd9

Extracting Value from Unstructured Data

Example

- A company has 100,000s+ pages of proprietary documentation to enable their staff to service customers.

Problem

- Searching can be slow, inefficient, or lack context.

Solution

- Create internal chatbot with ChatGPT and a vector database enriched with company documentation to provide direction and support to employees and customers.

https://osschat.io/chat

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

BRING YOUR OWN CLOUD

Zilliz BYOC

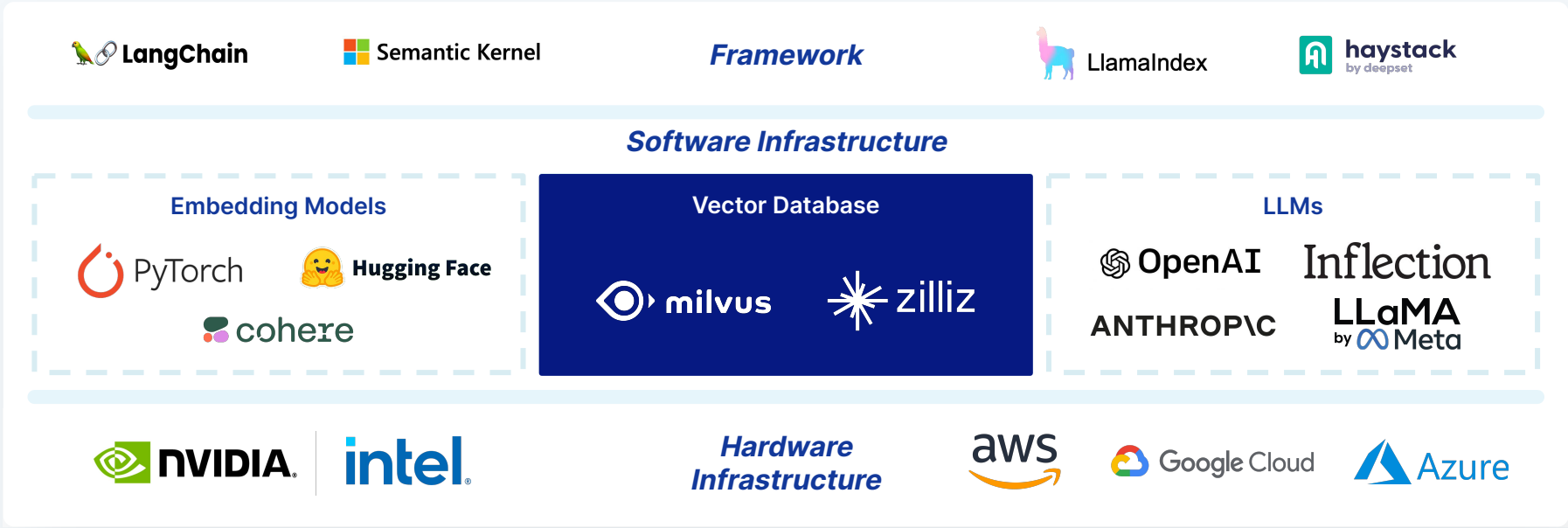
Enterprise-ready Milvus for Private VPCs

Deploy in your virtual private cloud

Coming Soon!

Coming Soon!

Well-connected in LLM infrastructure to enable RAG use cases



 |

zilliz

THANK YOU

Milvus Dependencies

https://zilliz.com/blog/Milvus-server-docker-installation-and-packaging-dependencies

🧱 Main Dependencies:

- FAISS 🔍 (vector search)
- etcd 🔑 (metadata store)
- Pulsar/Kafka 📢 (messaging)
- Tantivy 🔎 (text search)
- RocksDB 💾 (storage)
- Object Storage 🗄️ (Minio/S3/GCS/Azure Blob Storage)
- Kubernetes 🐳 (containerization)
- StorageClass & Persistent Volumes 🗄️(Storage Management for etcd and Pulsar)
- Prometheus & Grafana 📈 (monitoring)

📦 Docker Image Size: ~500MB

🆕 Release Frequency: ~1x per month, with frequent minor releases

🛠️ SDKs Available: Python 🐍, Node 🌐, Go 🐹, C# 💻, Java ☕, Ruby 💎

💻 Python SDK Installation: `pip install pymilvus`

✅ Version Compatibility: Ensure SDK and Milvus server versions match (major.minor)

 |

zilliz

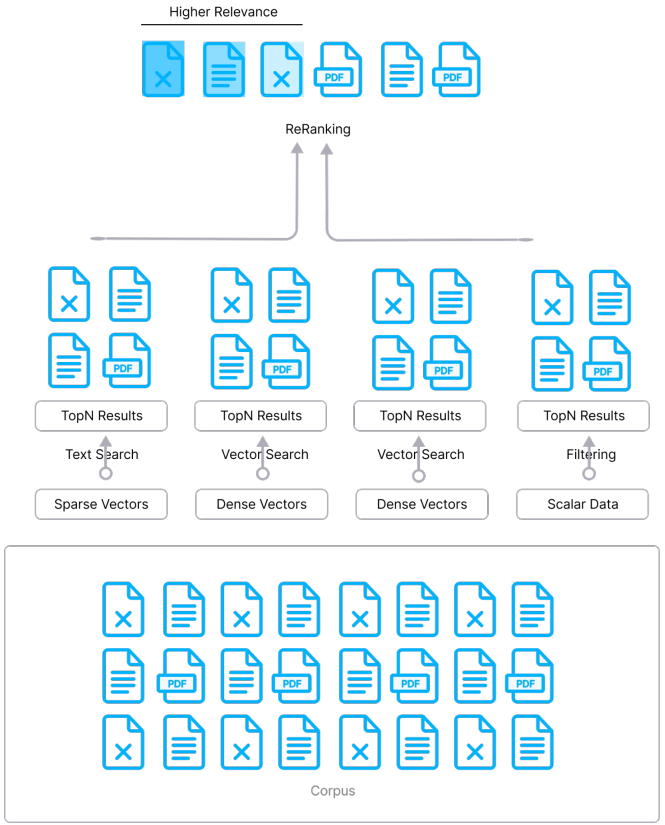
Unstructured Data Meetup

New York

Presented by zilliz | milvus

…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]



Why Not Vector Search Libraries?

- Search Quality - Hybrid Search? Filtering?
- Scalability - Billions of vectors?
- Multi tenancy - Isolating Multi-Tenant data
- Cost - Memory, disk, S3?
- Security - Data Safety and Privacy

TL;DR: Vector search libraries lack the infrastructure to help you scale, deploy, and manage your apps in production.

zilliz

Why Not Use a SQL/NoSQL Database?

- Inefficiency in High-dimensional spaces
- Suboptimal Indexing
- Inadequate query support
- Lack of scalability
- Limited analytics capabilities
- Data conversion issues

TL;DR: Vector operations are too computationally intensive for traditional database infrastructures

zilliz

What is Milvus/Zilliz ideal for?

Purpose-built to store, index and query vector embeddings from unstructured data **at scale**.

- Advanced filtering
- Hybrid search
- Multi-vector Search
- Durability and backups
- Replications/High Availability
- Sharding
- Aggregations
- Lifecycle management
- Multi-tenancy

- High query load
- High insertion/deletion
- Full precision/recall
- Accelerator support (GPU, FPGA)
- Billion-scale storage

 |

zilliz

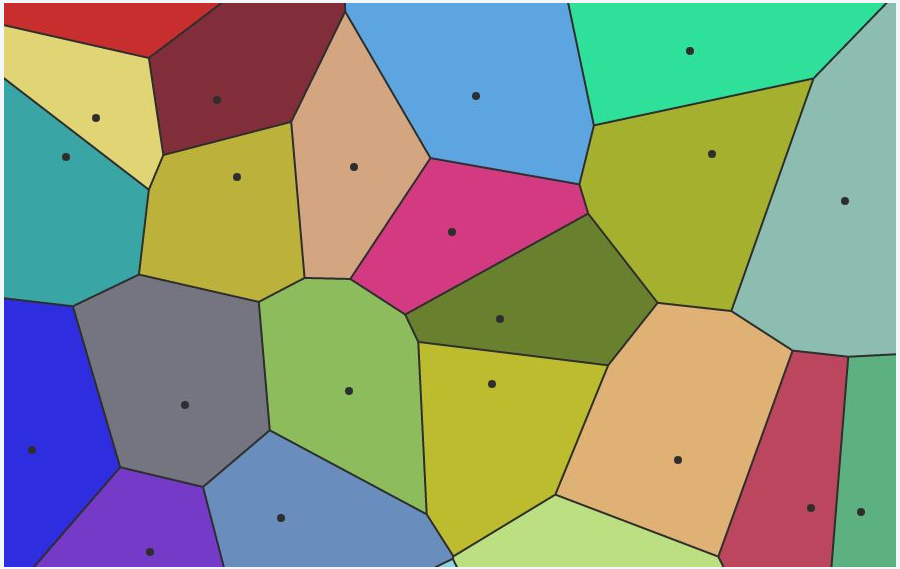
Takeaway:

Vector Databases are **purpose-built** to handle **indexing, storing, and querying** vector data.

Milvus & Zilliz are specifically designed for high performance and **billion+** scale use cases.

zilliz

Inverted File Index

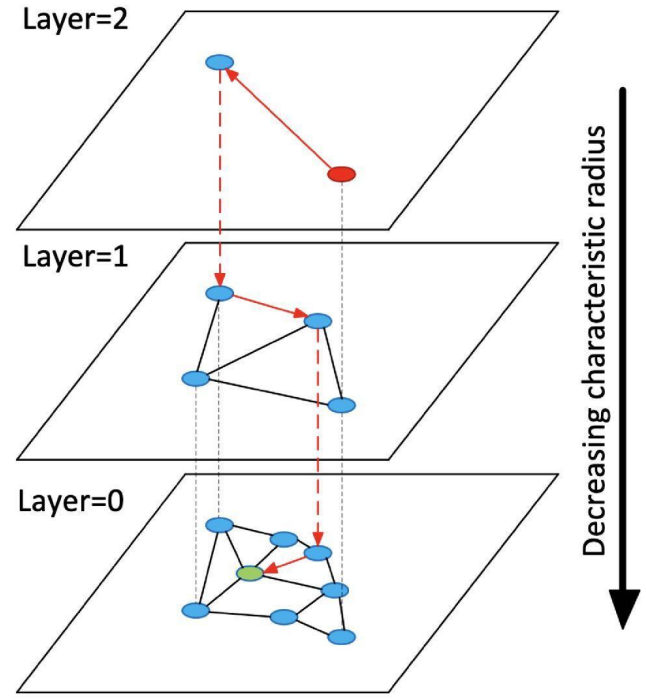


Source:
https://towardsdatascience.com/similarity-search-with-ivfpq-9c6348fd4db3

 |

zilliz

HNSW

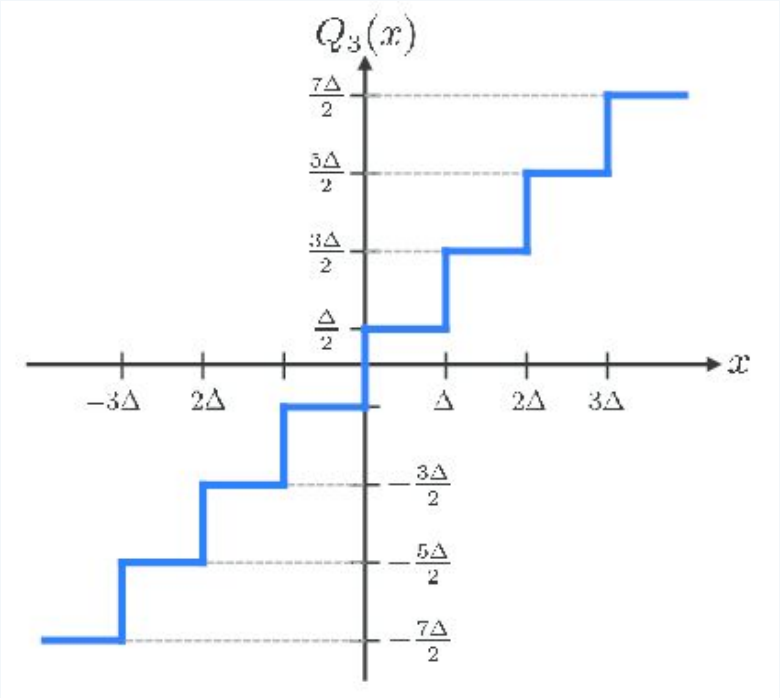


Source:
https://arxiv.org/ftp/arxiv/papers/1603/1603.09320.pdf

 |

zilliz

SQ



 |

zilliz