Unleashing the Potential of Cloud Native Open Source Vector Databases

Tim Spann @ Zilliz

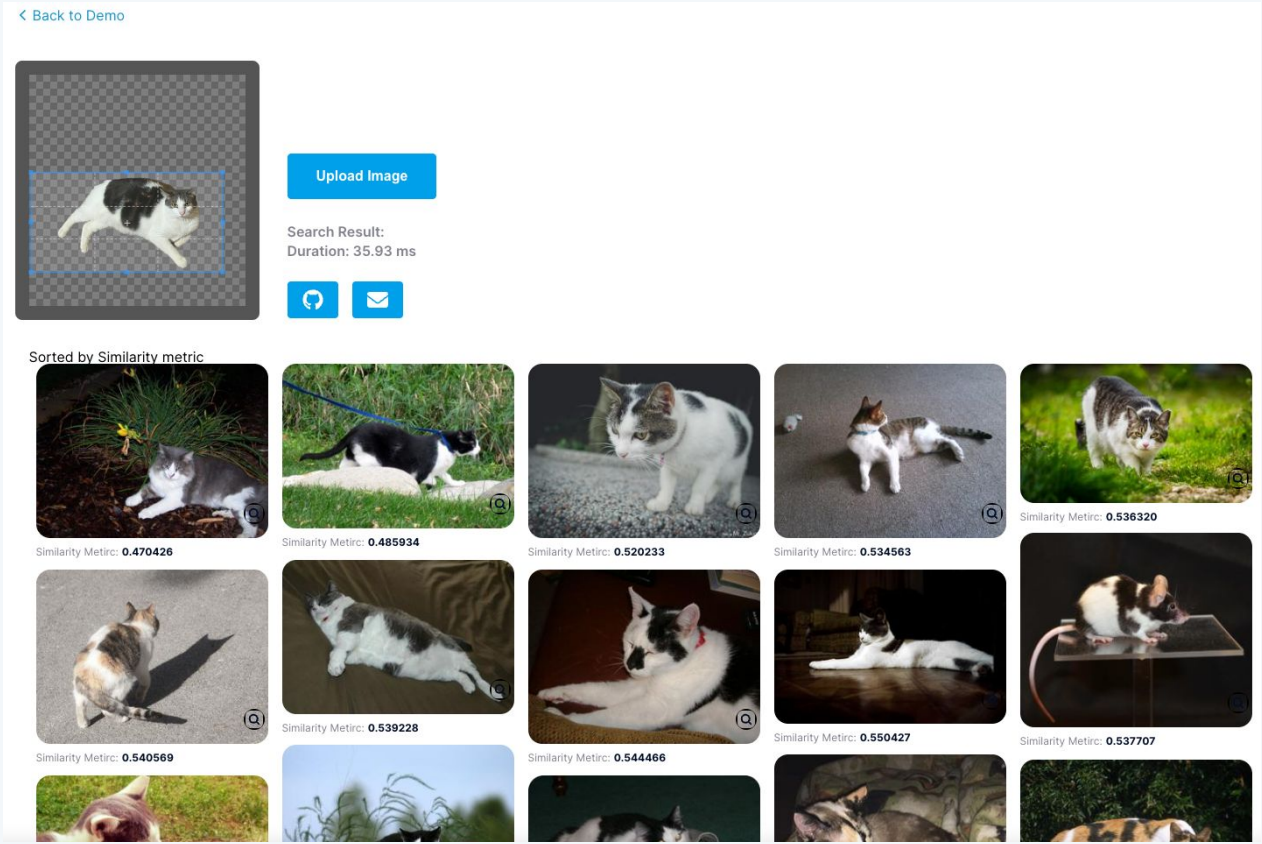
Tim Spann

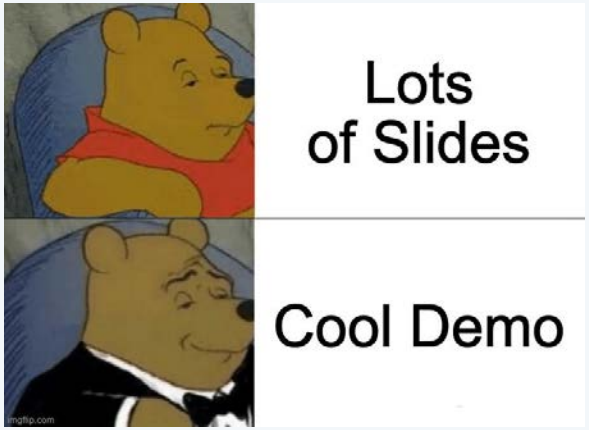
Principal Developer
Advocate, Zilliz

tim.spann@zilliz.com
https://www.linkedin.com/in/timothyspann/
https://x.com/PaaSDev

zilliz

Show Me

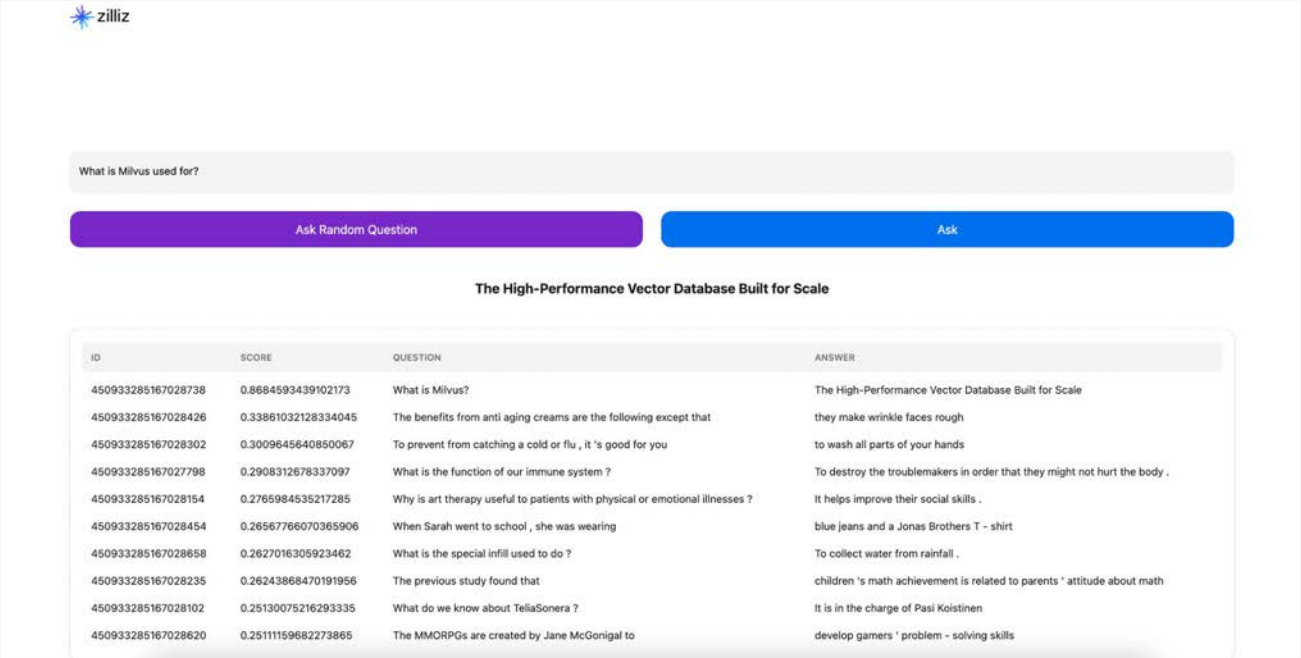




https://milvus.io/milvus-demos/reverse-image-search

zilliz

Show Me Another Demo



https://zilliz-semantic-search-example.vercel.app/

zilliz

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

zilliz

Unstructured Data is Everywhere

Unstructured data is any data that does not conform to a predefined data model.

By 2025, IDC estimates there will be 175 zettabytes of data globally (that's 175 with 21 zeros), with **80% of that data being unstructured**. Currently, 90% of unstructured data is never analyzed.

Text

Images

Videos

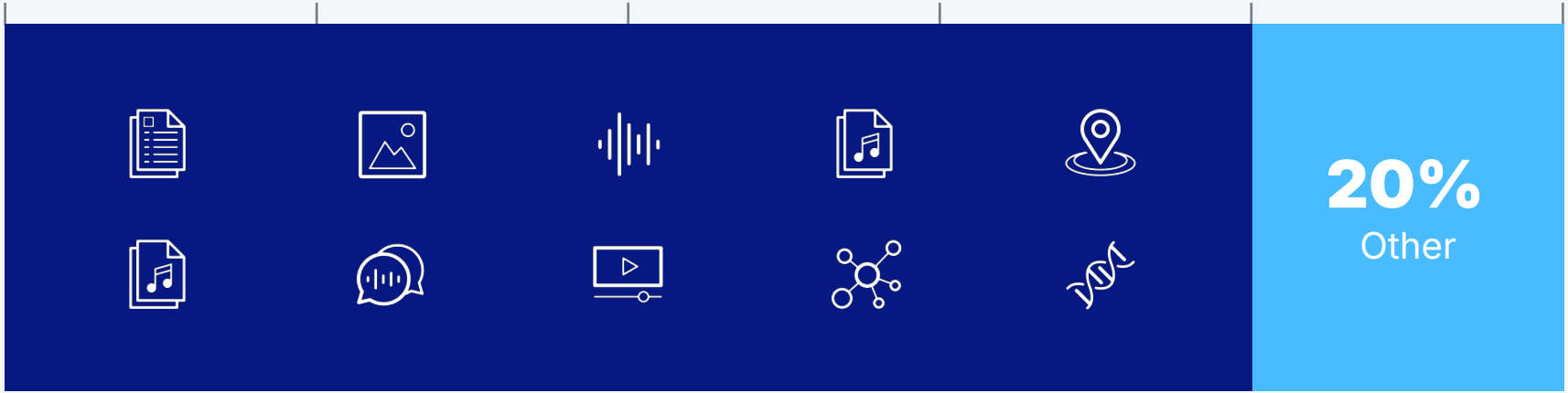
and more!

 |

zilliz

...and cannot process increasingly growing unstructured data

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

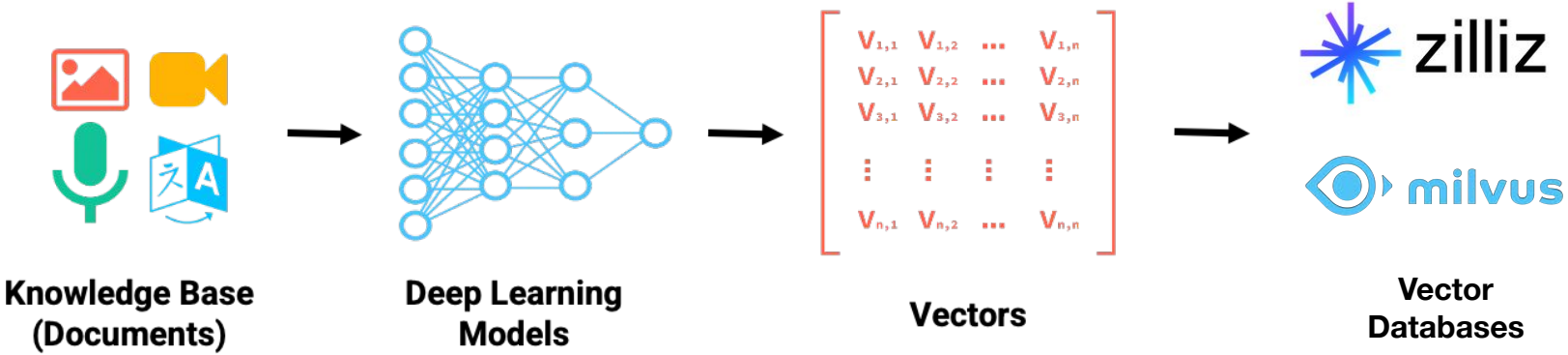


*Data Source: The Digitization of the World by *IDC*

zilliz

The challenge of unstructured data

- **Problem:** Unstructured data comes in lots of forms, no easy way to interact with it all
- **Solution:** Vector embeddings
- **How:** Neural networks e.g. embedding models



02

Overview of Vector Databases

zilliz

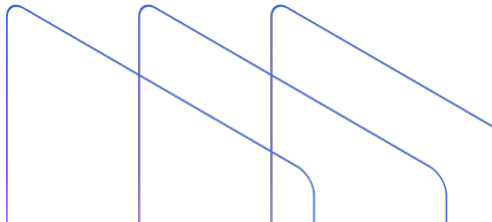
Why a Vector Database?

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

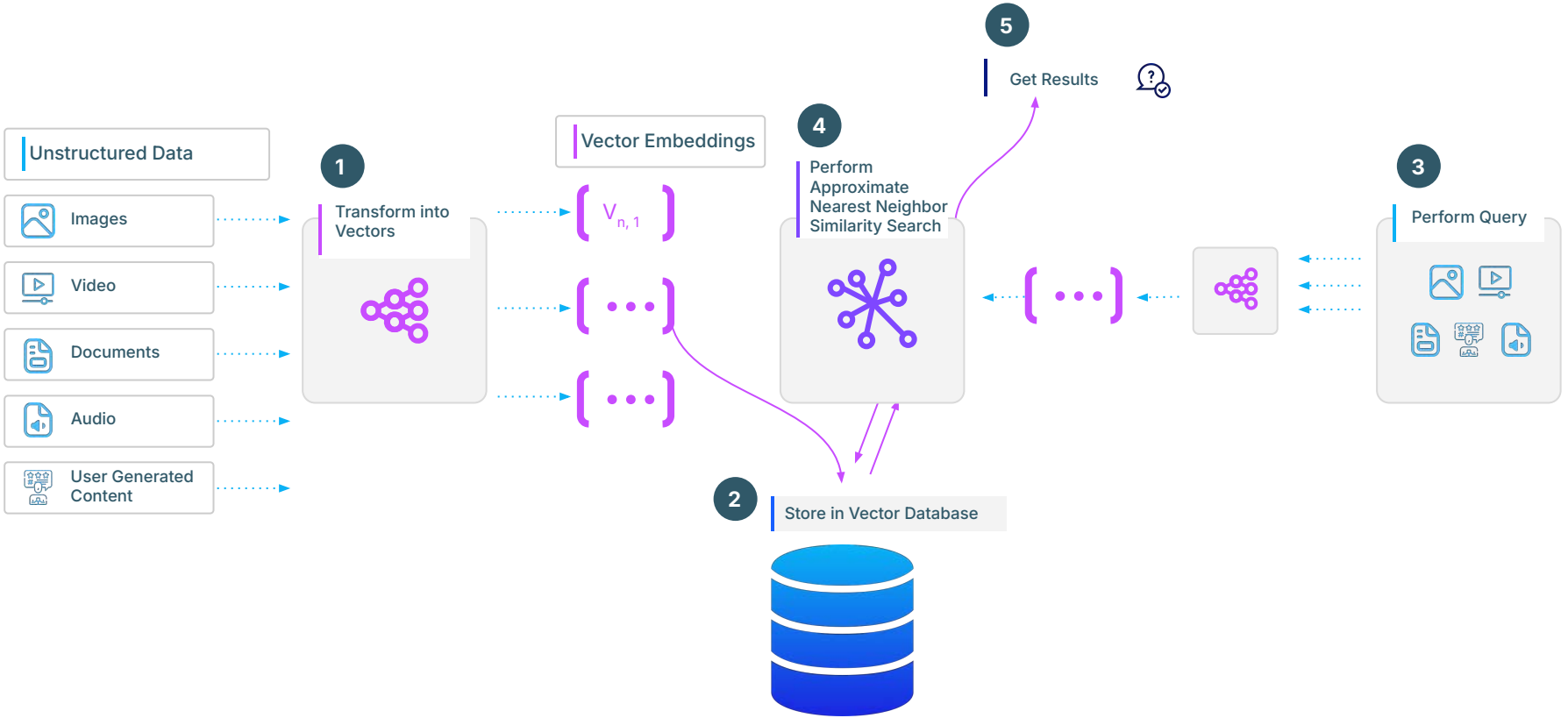
- Vector database
  - Advanced filtering (filtered vector search, chained filters)
  - Hybrid search (e.g. full text + dense vector)
  - Durability (any write in a db is durable, a library typically only supports snapshotting)
  - Replication / High Availability
  - Sharding
  - Aggregations or faceted search
  - Backups
  - Lifecycle management (CRUD, Batch delete, dropping whole indexes, reindexing)
  - Multi-tenancy

- Vector search library
  - High-performance vector search

- How do I support different applications?
  - High query load
  - High insertion/deletion
  - Full precision/recall
  - Accelerator support (GPU, FPGA)
  - Billion-scale storage

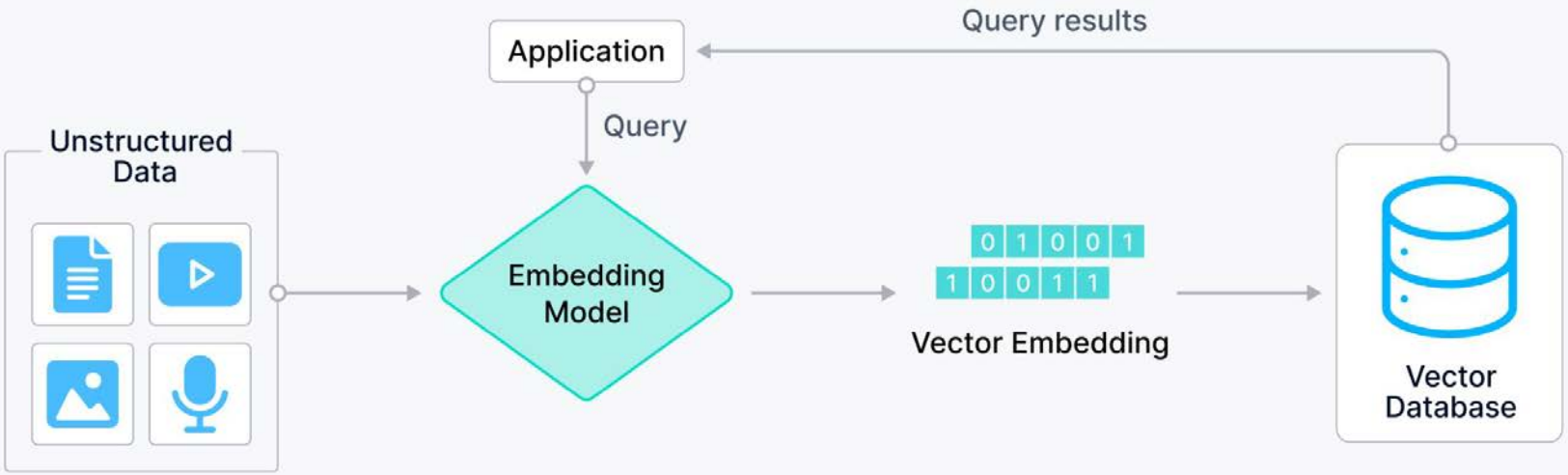


How Similarity Search Works



Vector Database: Making Sense of Unstructured Data

A vector database stores embedding vectors and allows for semantic retrieval of various types of unstructured data.



zilliz

Do you really need a Vector Database?

ANN Libraries

- FAISS, ANNOY, HNSW
- Supports 1M vectors
- Good for prototyping

Existing Solutions

- 50M~100M vectors
- PostgreSQL, ElasticSearch, Big Query, MongoDB, etc with ANNS plug-ins

Vector Databases

- Purpose-built for vectors top support the requirements and lifecycle of vectors
- Billion+ scale
- CRUD, real-time search, top-k/range/hybrid search, multi-modal, mulit-vector query, distributed
- Semantic Search is core to your business

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

 |

zilliz

03

A Quick Introduction to Milvus

About Milvus

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.

29K+
GitHub Stars

2,600+
Forks

25M+
Downloads

250+
Contributors

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

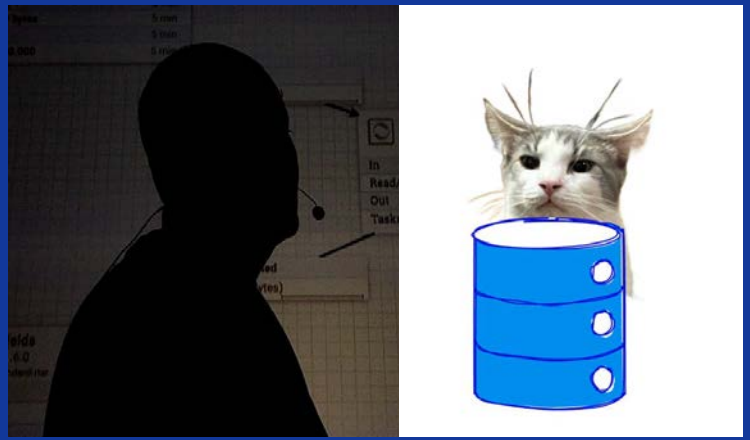
Plug into OpenAI, Langchain, LlmaIndex, and many more

Feature-rich

Dense & sparse embeddings, filtering, reranking and beyond

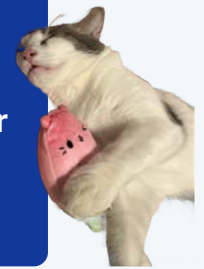
zilliz

Milvus Features



- Scalable and Elastic Architecture
- Diverse Index Support
- Versatile Search Capabilities
- Tunable Consistency

- Multi-Tenancy
- Hardware-Accelerated Compute Support
- Python, Java, Golang, NodeJS
- Milvus Lite, K8, Zilliz Cloud, Docker



 |

Technologies for various types of Use cases

Index Types

Offer a wide range of **15 indexes** support, including popular ones like Hierarchical Navigable Small Worlds (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

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

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

 |

Higher Scalability

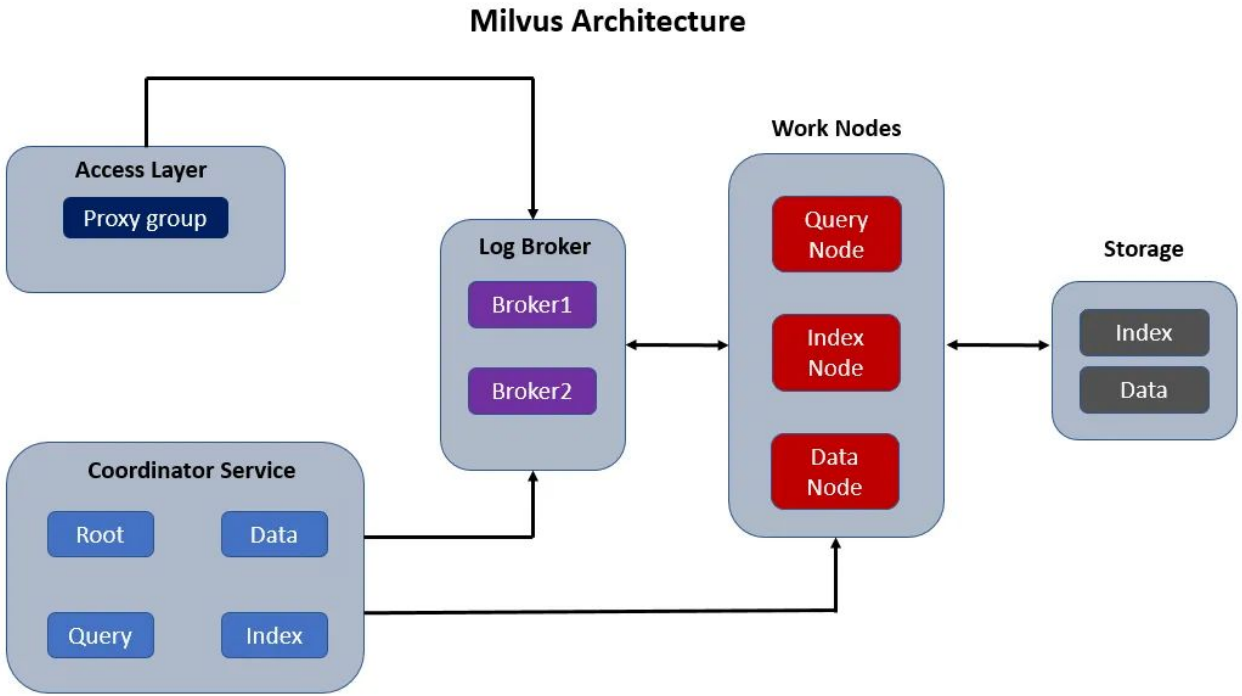
10B vectors

of 1536 dimensions
in a single Milvus/Zilliz Cloud instance

100B vectors

in one of the largest deployment

Milvus: decoupling computation and storage



Indexes

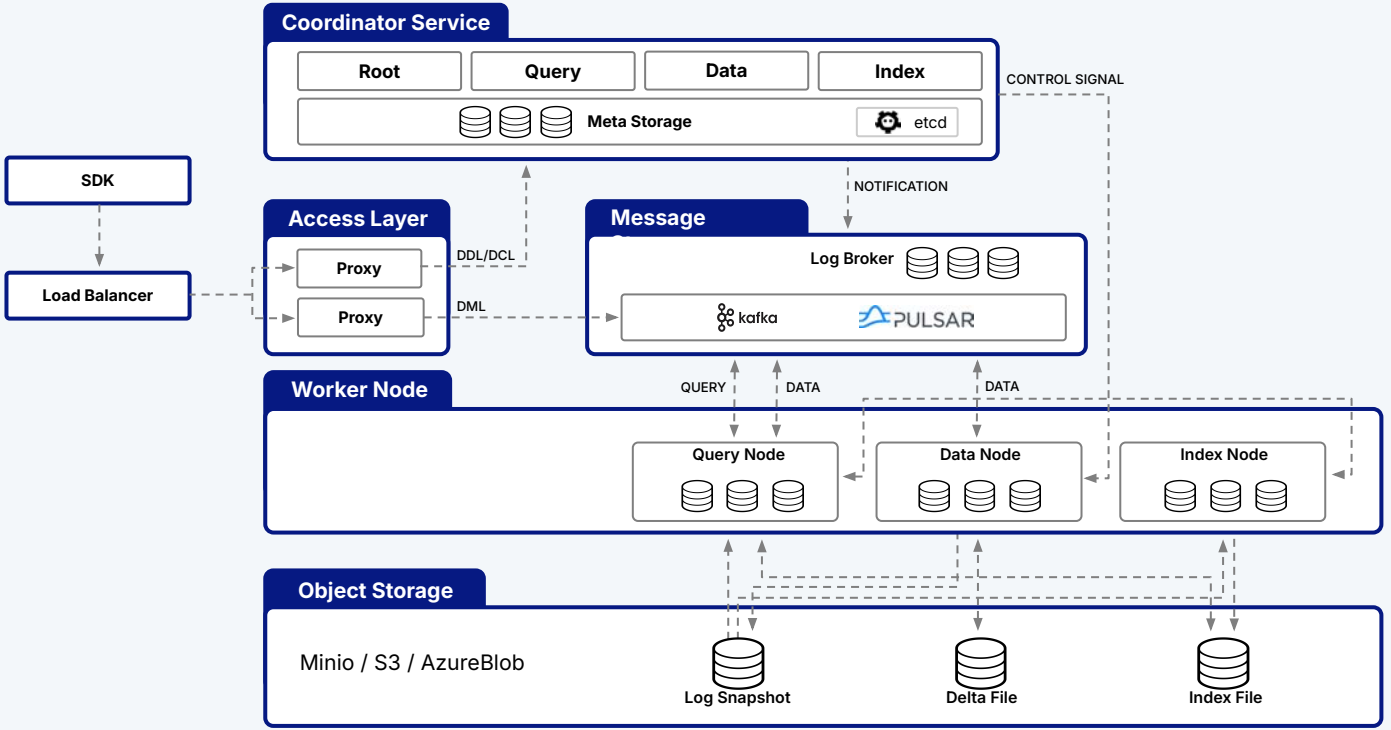
Most of the vector index types supported by Milvus use approximate nearest neighbors search (ANNS),

- **HNSW**: HNSW is a graph-based index and is best suited for scenarios that have a high demand for search efficiency. There is also a GPU version **GPU_CAGRA**, thanks to Nvidia's contribution.
- **FLAT**: FLAT is best suited for scenarios that seek perfectly accurate and exact search results on a small, million-scale dataset. There is also a GPU version **GPU_BRUTE_FORCE**.
- **IVF_FLAT**: IVF_FLAT is a quantization-based index and is best suited for scenarios that seek an ideal balance between accuracy and query speed. There is also a GPU version **GPU_IVF_FLAT**.
- **IVF_SQ8**: IVF_SQ8 is a quantization-based index and is best suited for scenarios that seek a significant reduction on disk, CPU, and GPU memory consumption as these resources are very limited.
- **IVF_PQ**: IVF_PQ is a quantization-based index and is best suited for scenarios that seek high query speed even at the cost of accuracy. There is also a GPU version **GPU_IVF_PQ**.

Indexes Continued.

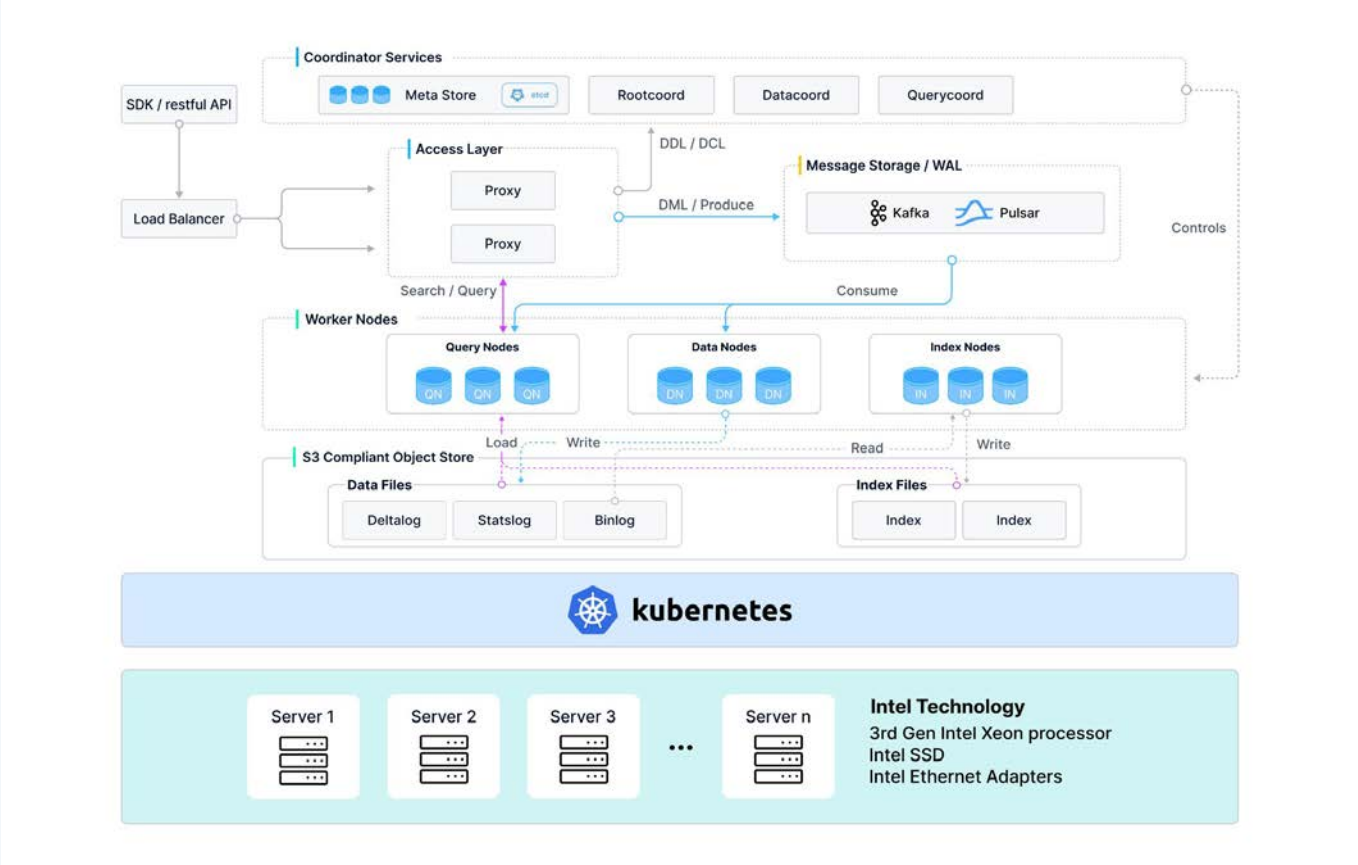
- **SCANN**: SCANN is similar to IVF_PQ in terms of vector clustering and product quantization. What makes them different lies in the implementation details of product quantization and the use of SIMD (Single-Instruction / Multi-data) for efficient calculation.
- **DiskANN**: Based on Vamana graphs, DiskANN powers efficient searches within large datasets.

Milvus' fully distributed architecture is designed scalability and performance



 |

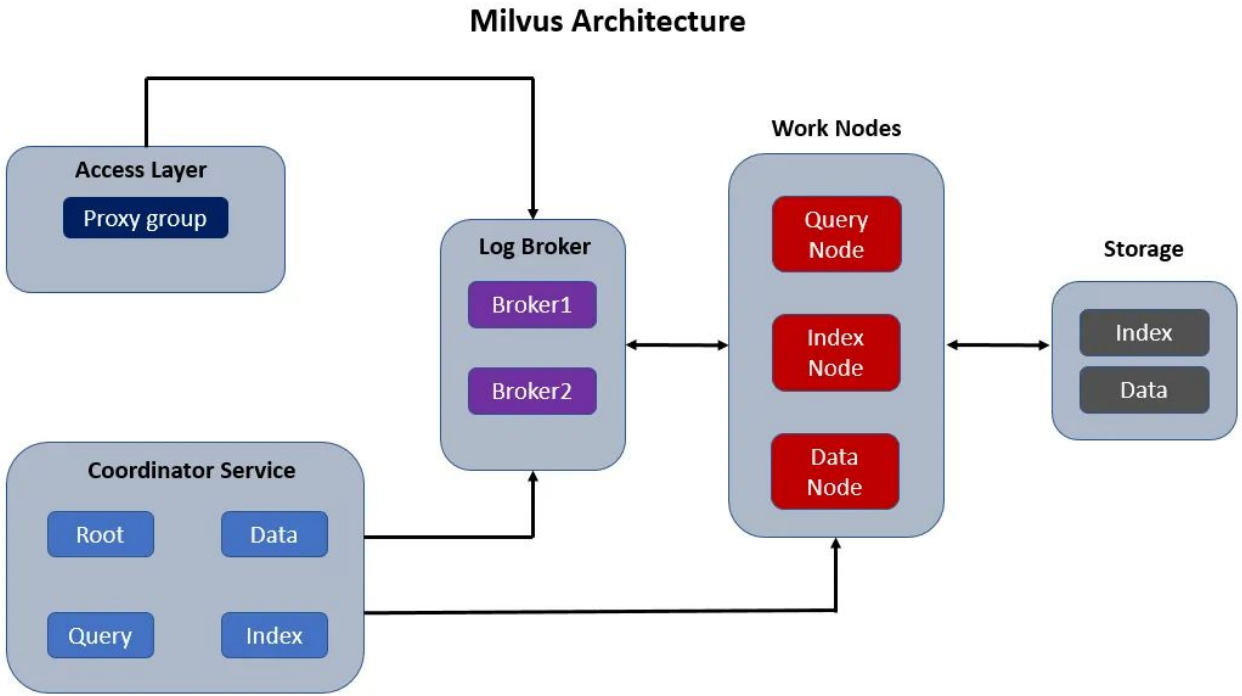
High-level Overview of Milvus' Architecture



 |

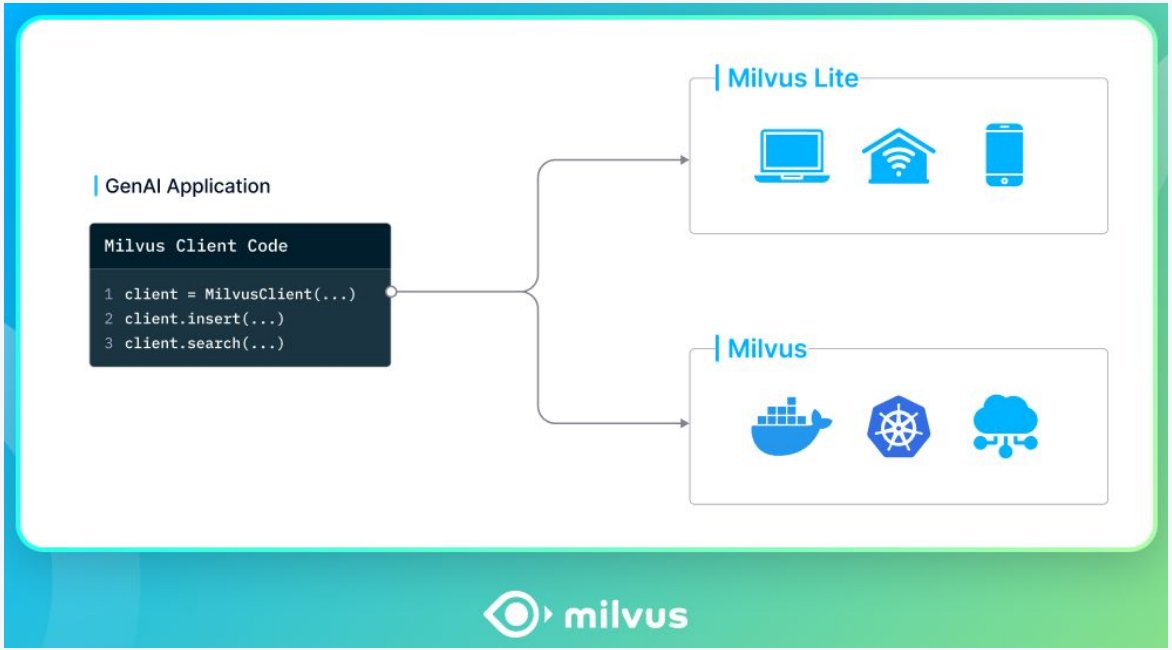
zilliz

Milvus: decoupling computation and storage



Milvus Lite

```
pip install pymilvus
```



 |

zilliz

05

Building a local RAG application

zilliz

Vector embeddings are something computers can understand

0.62	0.57	0.97	0.83	0.21	0.71	0.99	0.98	0.19	0.3	0.03	0.78
0.48	0.77	0.06	0.2	0.42	0.34	0.88	0.29	0.35	0.96	0.37	0.11
0.57	0.93	0.99	0.14	0.64	0.21	0.57	0.14	0.96	0.39	0.51	0.29
0.52	0.51	0.07	0.26	0.77	0.96	0.55	0.19	0.25	0.39	0.63	0.32
0.75	0.97	0.35	0.7	0.39	0.41	0.32	0.81	0.59	0.11	0.23	0.4
0.89	0.28	0.09	0.15	0.8	0.16	0.57	0.7	0.38	0.44	0.35	0.75
0.42	0.96	0.33	0.53	0.65	0.89	0.63	0.19	0.49	0.93	0.09	0.92
0.6	0.86	0.86	0.49	0.41	0.39	0.02	0.39	0.62	0.41	0.91	0.29
0.64	0.22	0.89	0.44	0.47	0.41	0.72	0.34	0.16	0.12	0.68	0.89
0.93	0.27	0.16	0.37	0.66	0.76	0.23	0.6	0.19	0.85	0.63	0.8
0.81	0.93	0.06	0.57	0.33	0.45	0.01	0.53	0.73	0.56	0.98	0.14
0.62	0.75	0.81	0.55	0.17	0.55	0.74	0.18	0.26	0.29	0.22	0.35
0.82	0.72	0.09	0.3	0.01	0.3	0.68	0.4	0.72	0.17	0.1	0.19
0.32	0.79	0.08	0.2	0.44	0.63	0.35	0.16	0.27	0.62	0.96	0.88
0.89	0.22	0.26	0.71	0.96	0.46	1	0.98	0.43	0.53	0.78	0.17
0.07	0.45	0.89	0.68	0.68	0.4	0.02	0.39	0.05	0.56	0.57	0.92

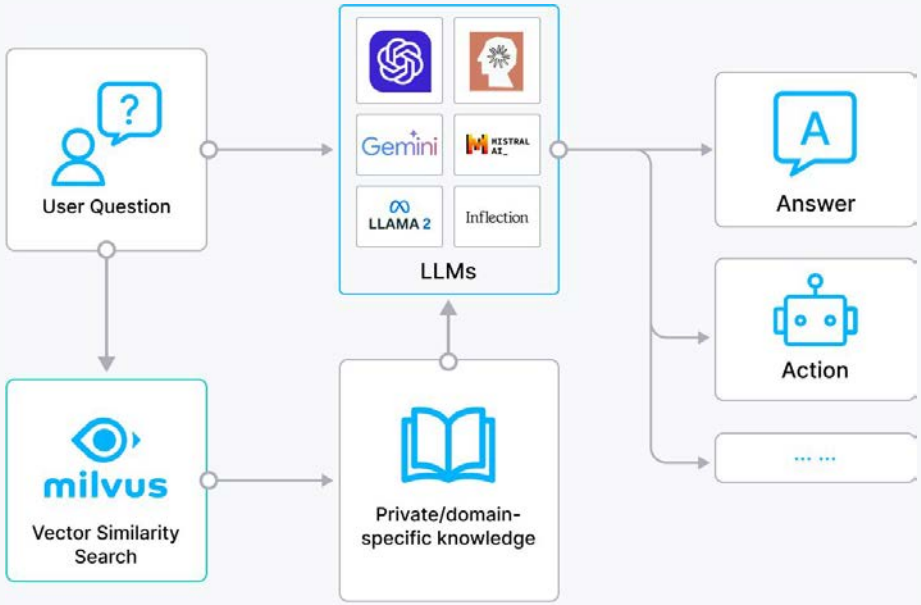
 |

zilliz

Retrieval-Augmented Generation (RAG)

A technique that combines the strength of retrieval-based and generative models:

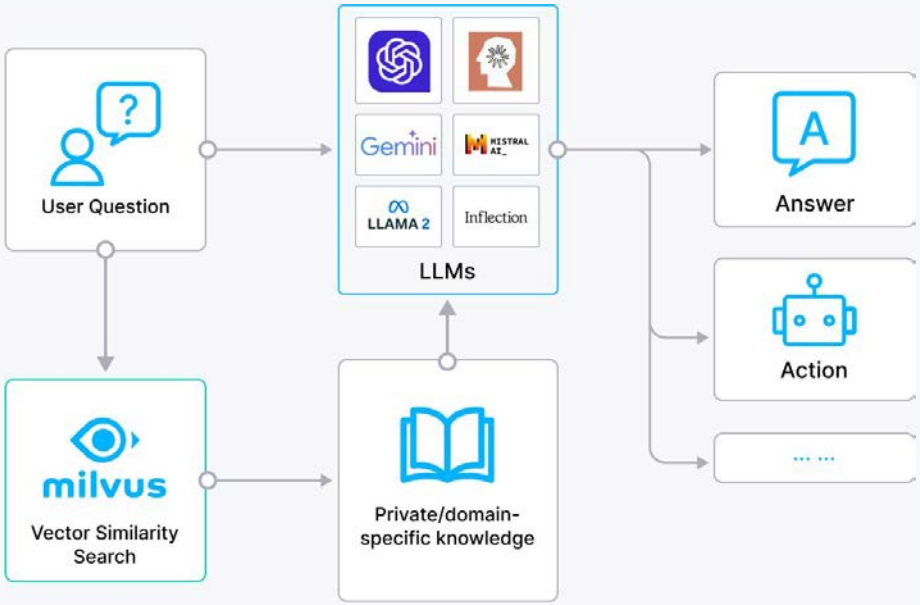
- Improve accuracy and relevance
- Eliminate hallucination
- Provide domain-specific knowledge



RAG : an economic perspective

A business model that bridges public data and private data

- Data sovereignty
- You can't and shouldn't give your private data to others



zilliz

Product Portfolio

Open Source

milvus
VECTOR DATABASE

Knowhere
VECTOR SEARCH ENGINE

GPT-Cache
SEMANTIC CACHE FOR LLM QUERIES

VDB Benchmark
VECTORDB BENCHMARK TOOL

Attu
GUI for Milvus

Commercial Offerings

Zilliz Cloud

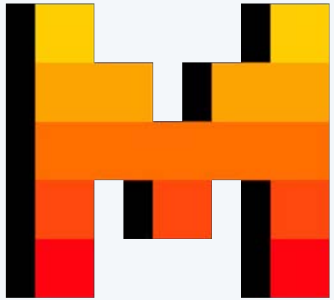
Optimized Milvus with essential data and security tools for a high-performing vector search platform

Deploy fully managed or “Bring Your Own Cloud” (BYOC)

 |

zilliz

Embeddings Models



 |

zilliz

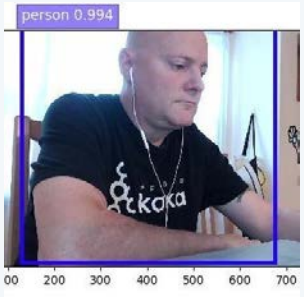
RESOURCES

 |

zilliz

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.

Generative AI Resource Hub

Tutorials, Code Examples, and Best Practices for Developing and Deploying GenAI Applications.

https://zilliz.com/learn/generative-ai

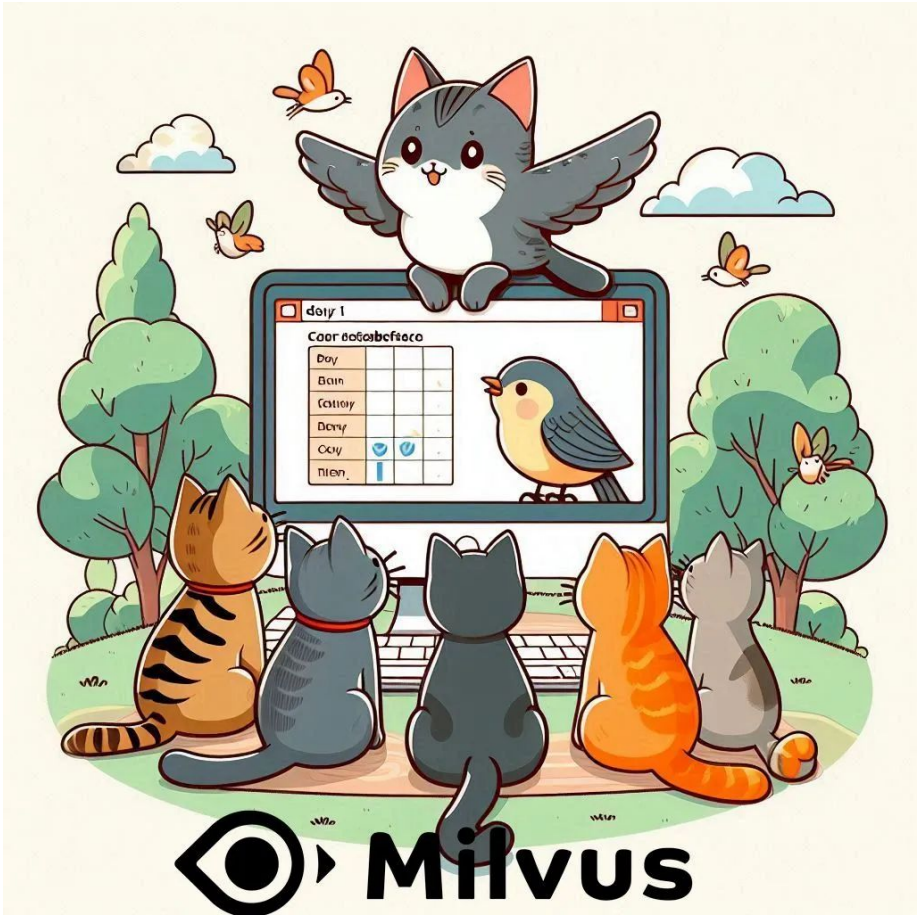
 |

zilliz

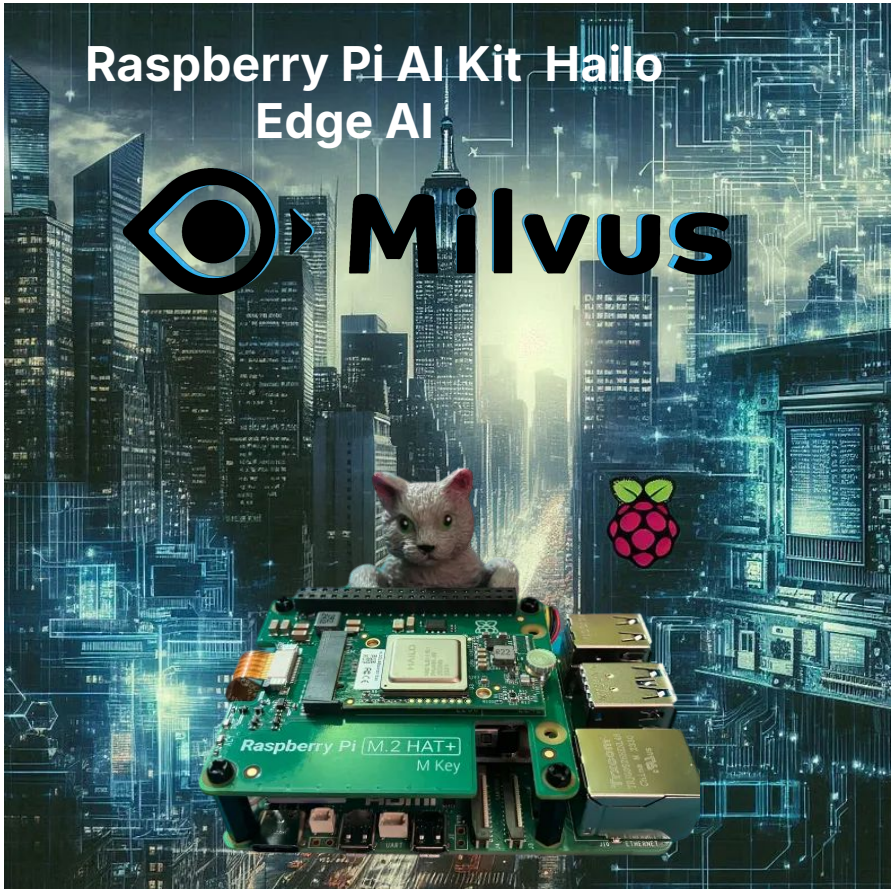

Learning Vector

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



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

AIM Weekly by Tim Spann

https://bit.ly/32dAJft
https://github.com/milvus-io/milvus

This week in Milvus, Towhee, Attu, GPT Cache, Gen AI, LLM, Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Apache Spark, Apache Iceberg, Python, Java, Vector DB and Open Source friends.

 |

zilliz

Thank you!

milvus.io

github.com/milvus-io/

@milvusio

Connect with me!

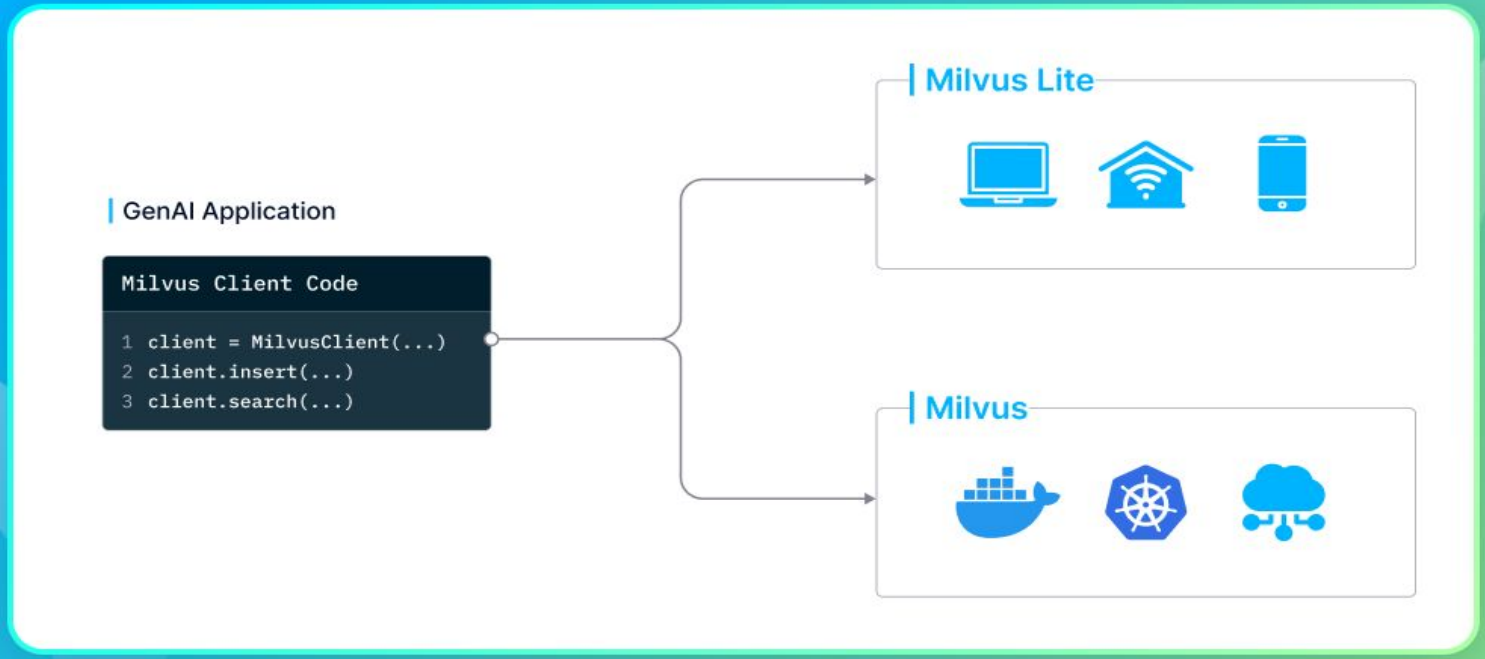
@paasDev

/in/timothyspann

 |

zilliz

Build Once
Deploy Anywhere



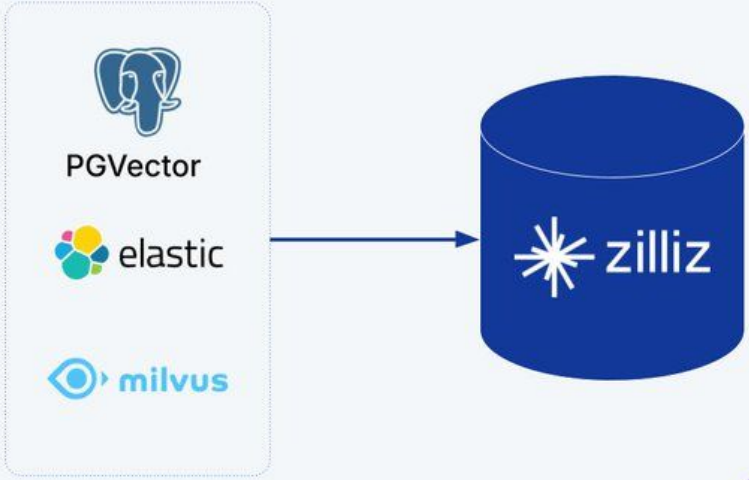
milvus

 |

zilliz

Migration Services

Zero-loss unstructured data migration across platforms, supporting both batch and streaming for AI applications.



zilliz

zilliz

BEFORE MILVUS

 |

zilliz

Join us at our next meetup!
meetup.com/unstructured-data-meetup-new-york/

THANK YOU

05

What is Similarity Search?

zilliz

Vector Search Overview

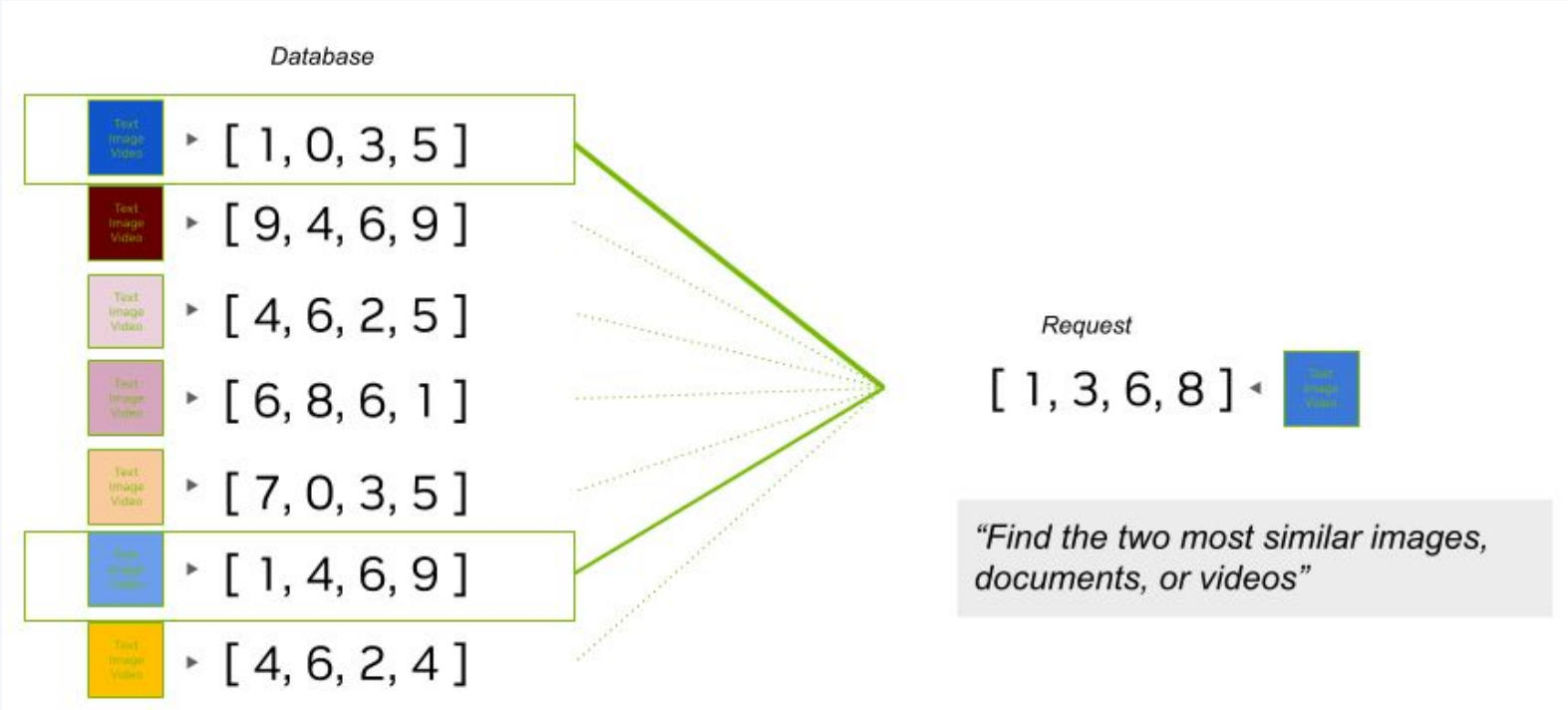


Image from Nvidia

 |

zilliz

Vector Similarity Measures: L2 (Euclidean)

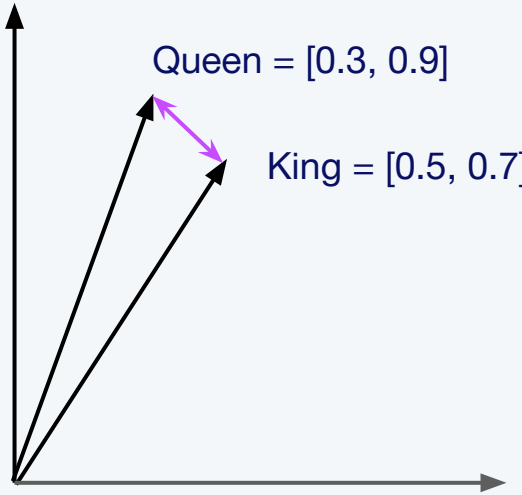
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

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

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

= $\sqrt{0.04 + 0.04}$

= $\sqrt{0.08} \cong 0.28$



 |

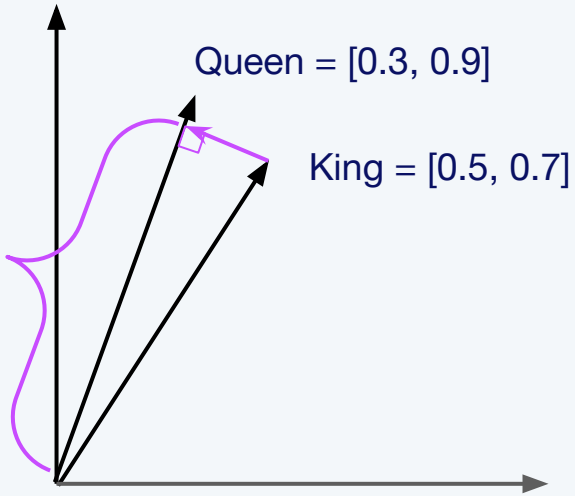
zilliz

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

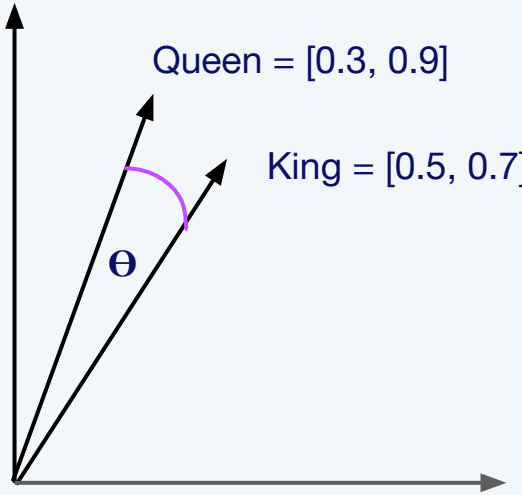


 |

zilliz

Vector Similarity Measures: Cosine

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



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

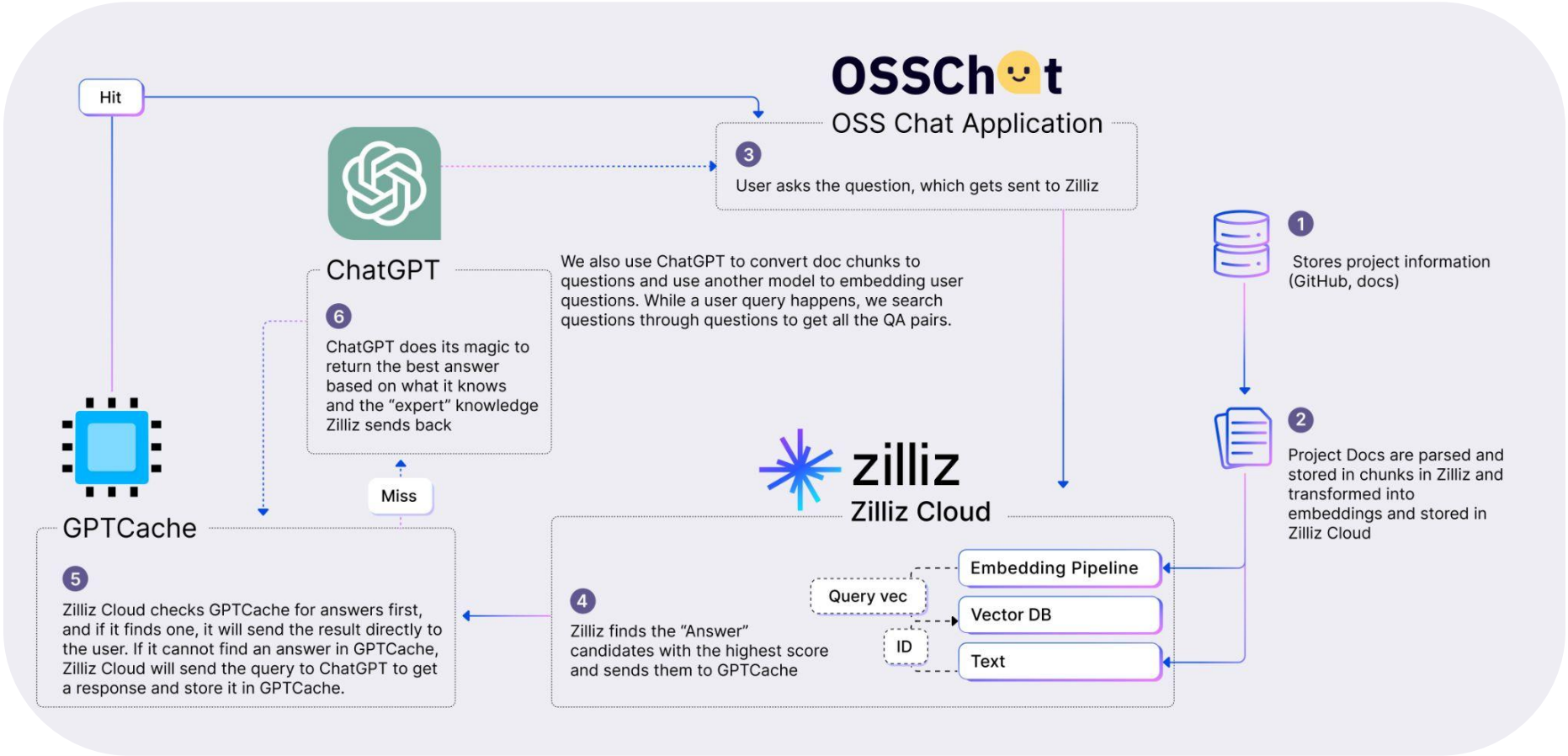
$$= \frac{0.15+0.63}{\sqrt{0.9} * \sqrt{0.74}}$$

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

$$\cong 0.03$$

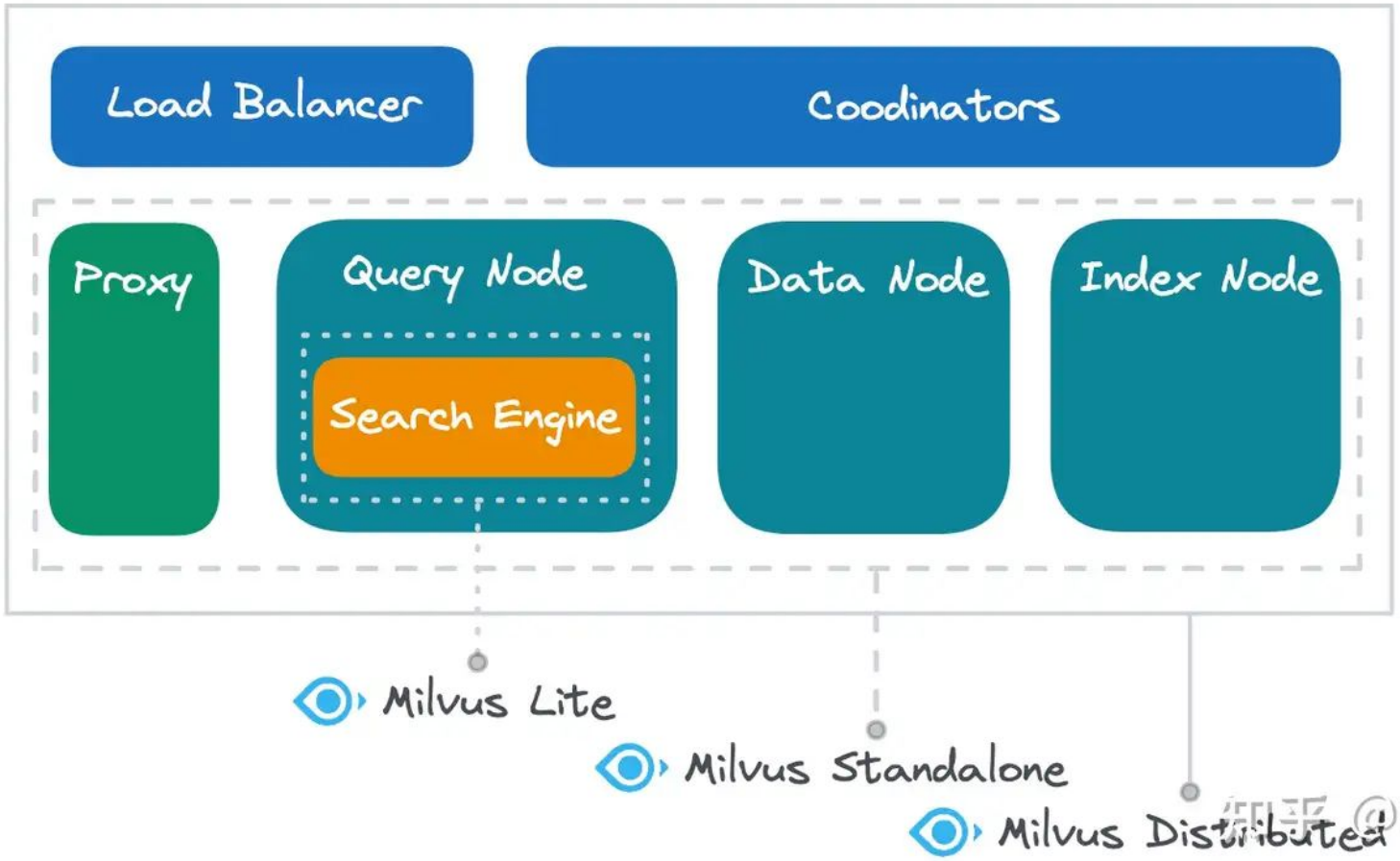
 |

zilliz



osschat.io

zilliz



zilliz