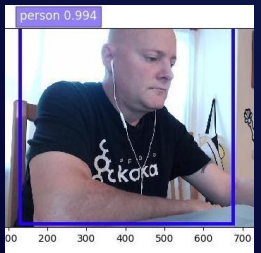
zilliz

Introduction to Milvus

June, 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

https://lu.ma/zh6ktycd

Thursday, July 25

https://www.meetup.com/unstructured-data-meetup-new-york/events/301720478/

Agenda

01

Shift Search Data Paradigm

How AI has revolutionized our search capabilities and the variety of data we can process

02

Introducing Milvus

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

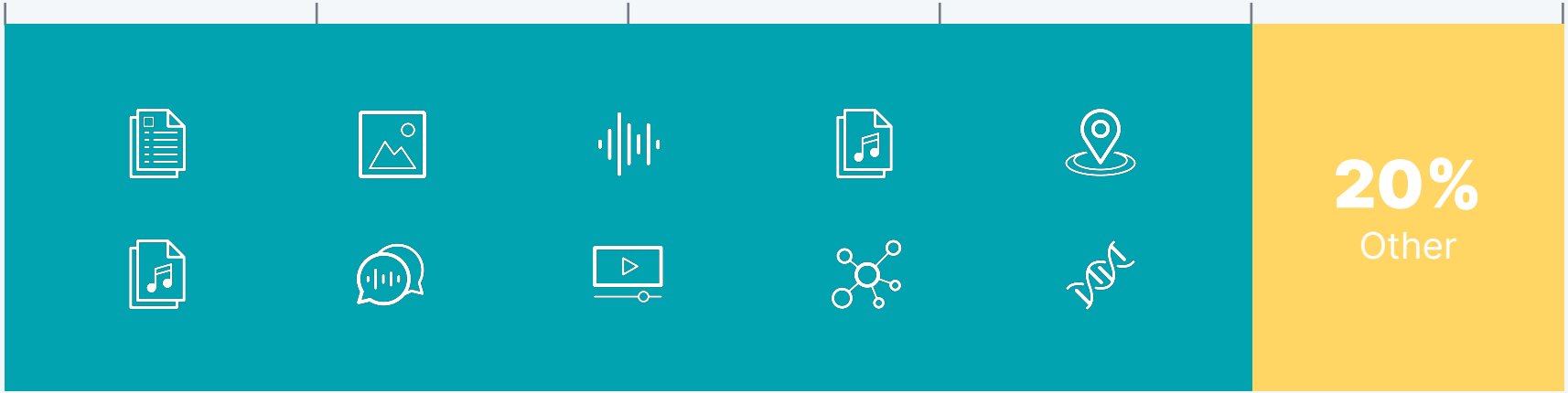
Shifted Search and Data Paradigm

 |

zilliz

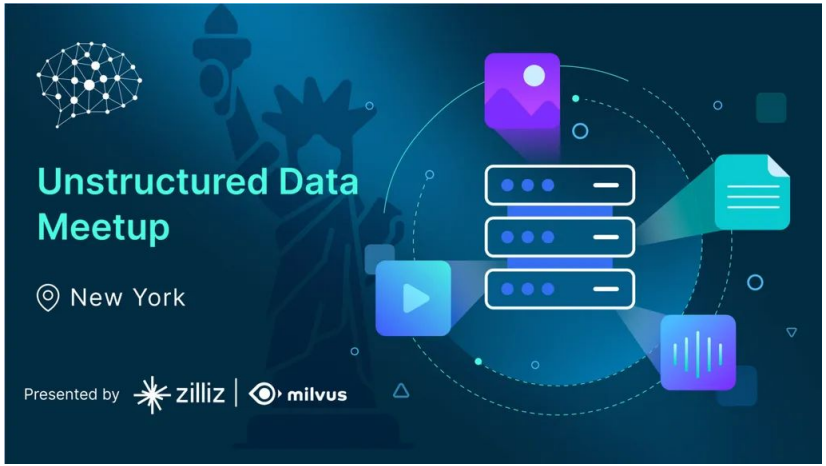
…and cannot process increasingly growing unstructured data

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



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

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.

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

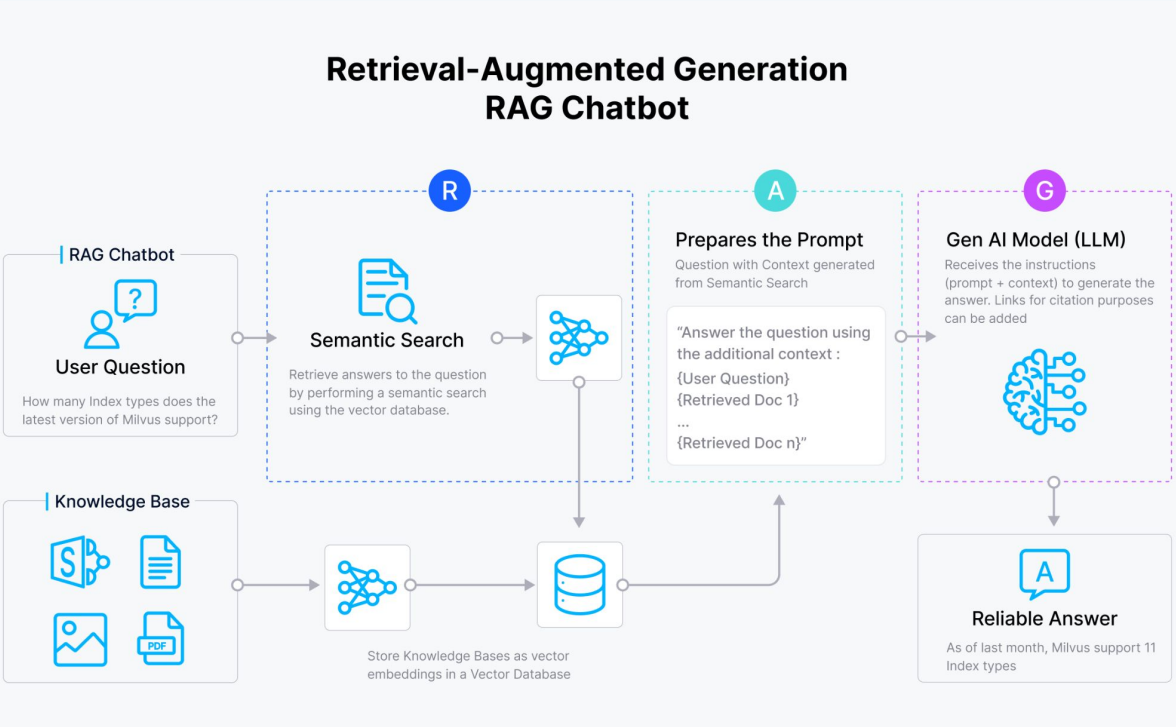
Enable robust handling of expanding data volumes and search demands

Highly performant

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

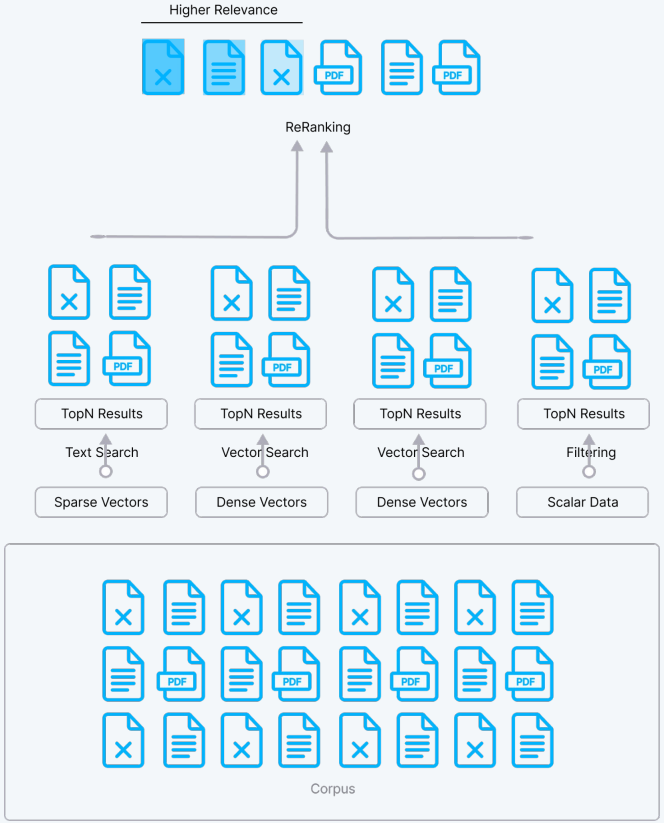
zilliz

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



…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]



Introducing Milvus

zilliz

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.

contributors stars docker pulls 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

 |

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

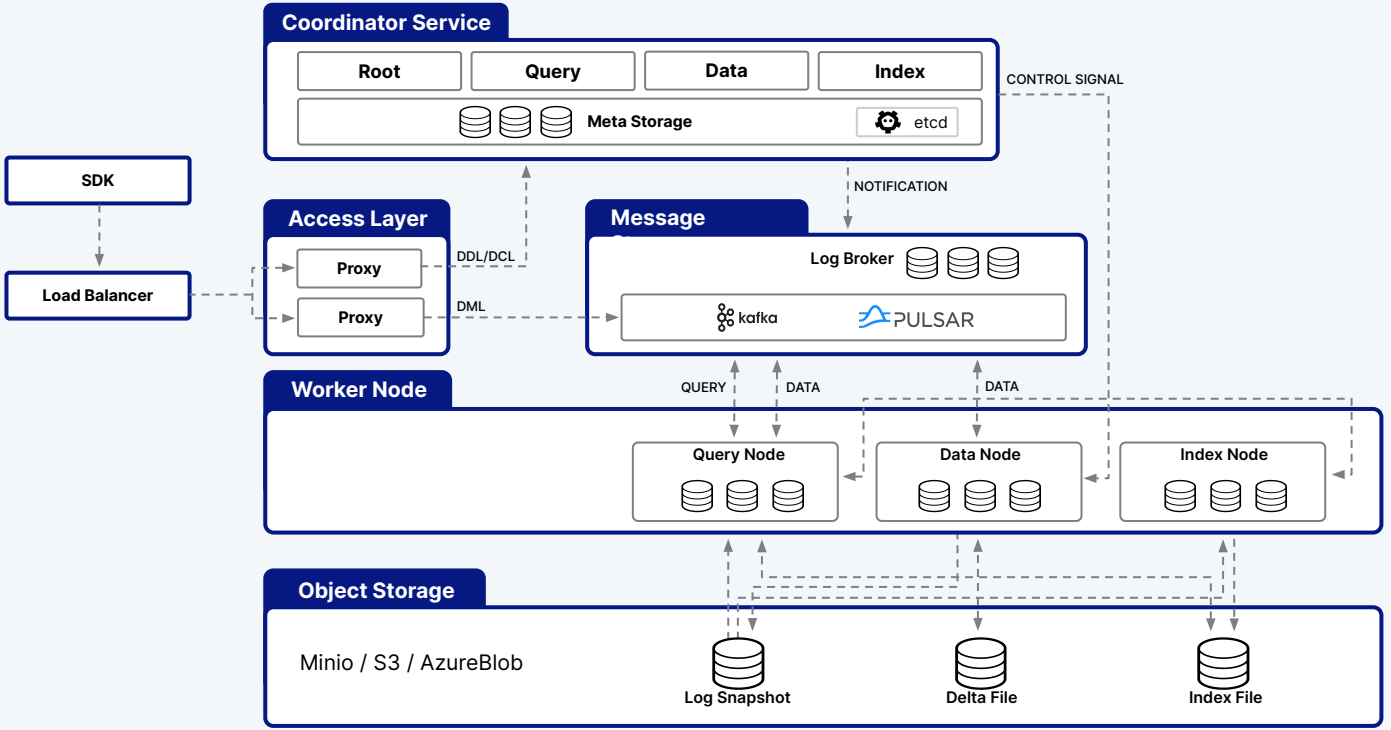
Compute Types

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

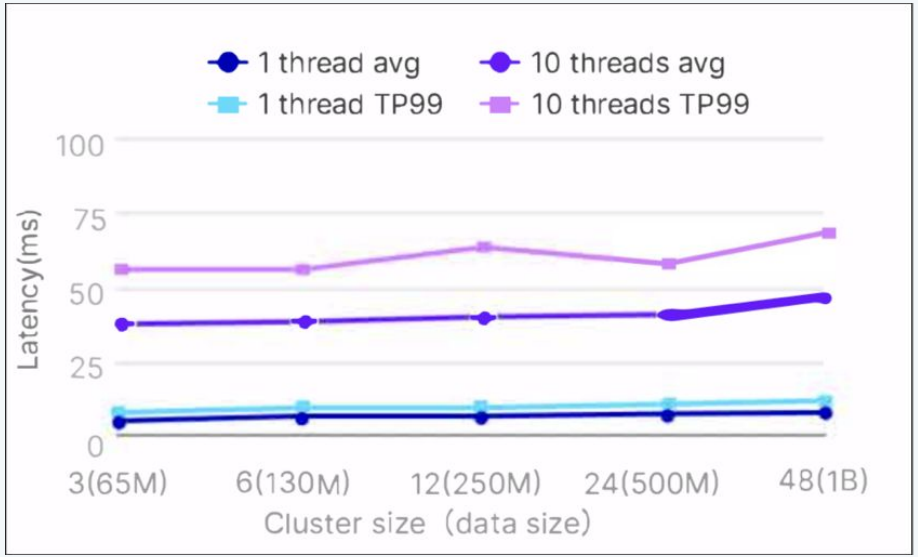
 |

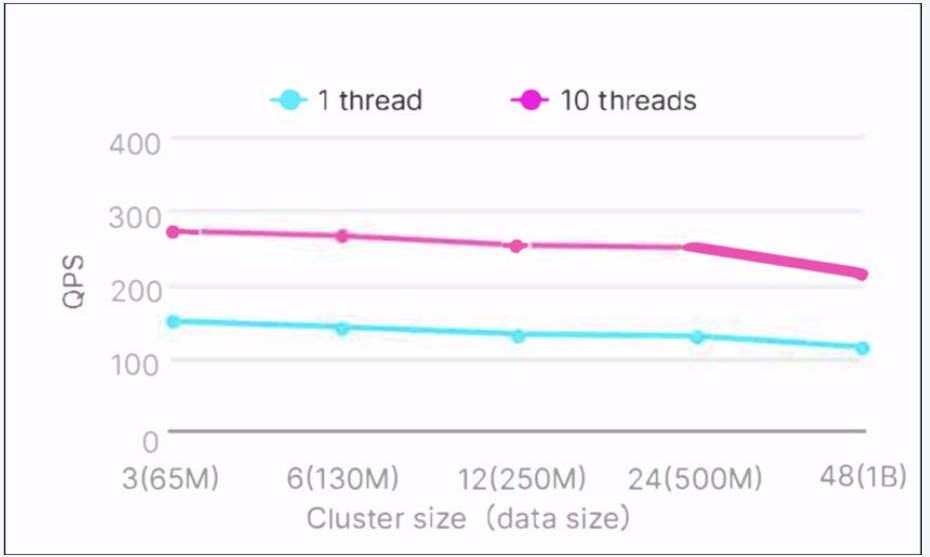
Milvus' fully distributed architecture is designed scalability and performance



 |

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

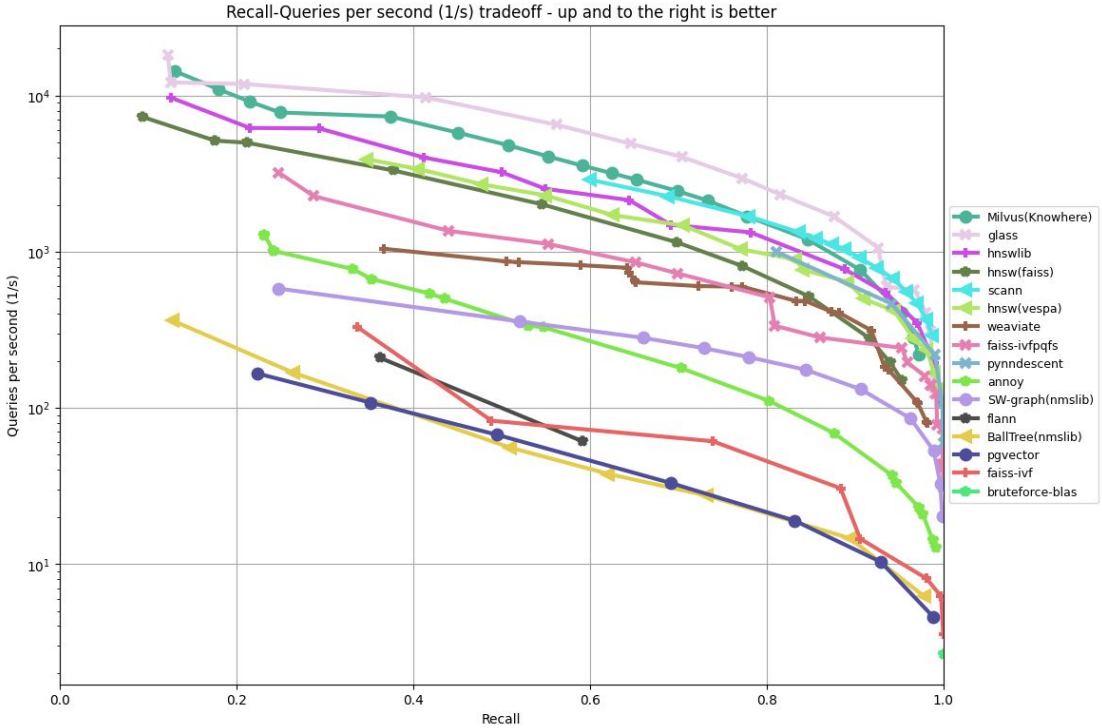




 |

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

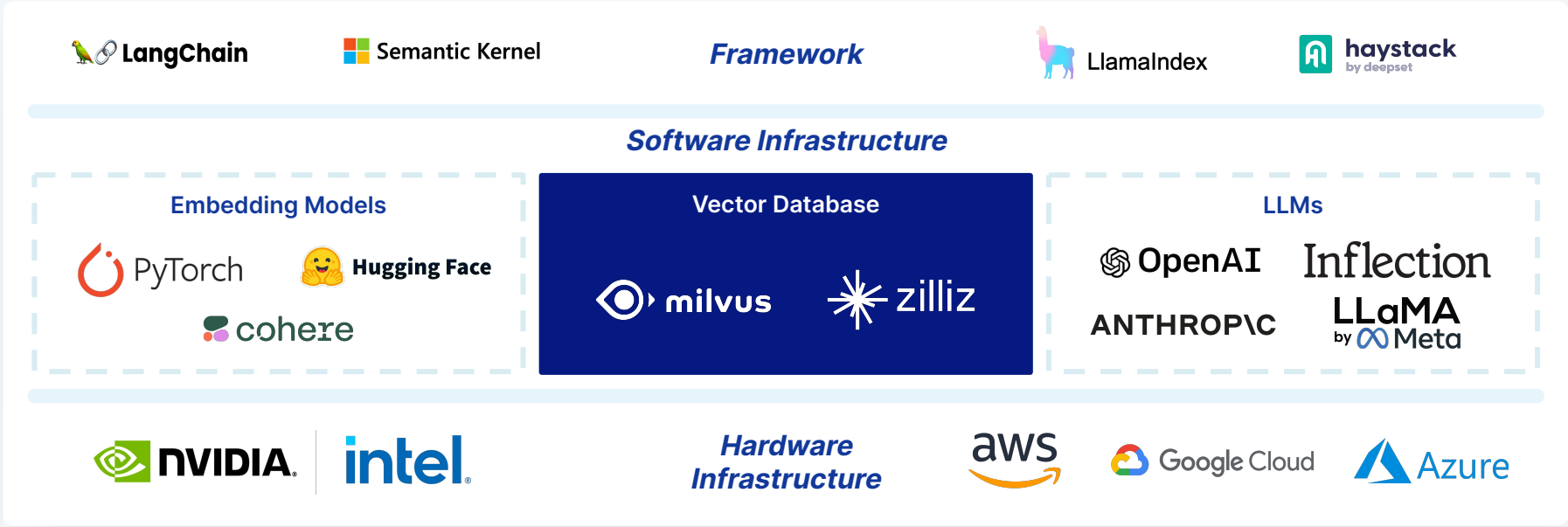
Enterprise-ready Milvus for Private VPCs

Deploy in your virtual private cloud

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

zilliz

Well-connected in LLM infrastructure to enable RAG use cases



 |

zilliz

RESOURCES

zilliz

Street Cameras

https://medium.com/cloudera-inc/streaming-street-cams-to-yolo-v8-with-python-and-nifi-to-minio-s3-3277e73723ce

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.

Join Our Slack and Interact with LLM

https://flankworkspace.slack.com/

https://join.slack.com/t/flankworkspace/shared_invite/zt-2fycjv241-~NRHZDtdfwDjlfvXK_Bz0A

Vector Database Resources

Give Milvus a Star!

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

Chat with me on Discord!

THANK YOU

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