Introduction to Unstructured Data, Vector Database and Gen AI

Tim Spann @ Zilliz

Tim Spann

Principal Developer
Advocate, Zilliz

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

 |

zilliz

W

A New Data and Compute World

zilliz

The world is much more than just text and keywords

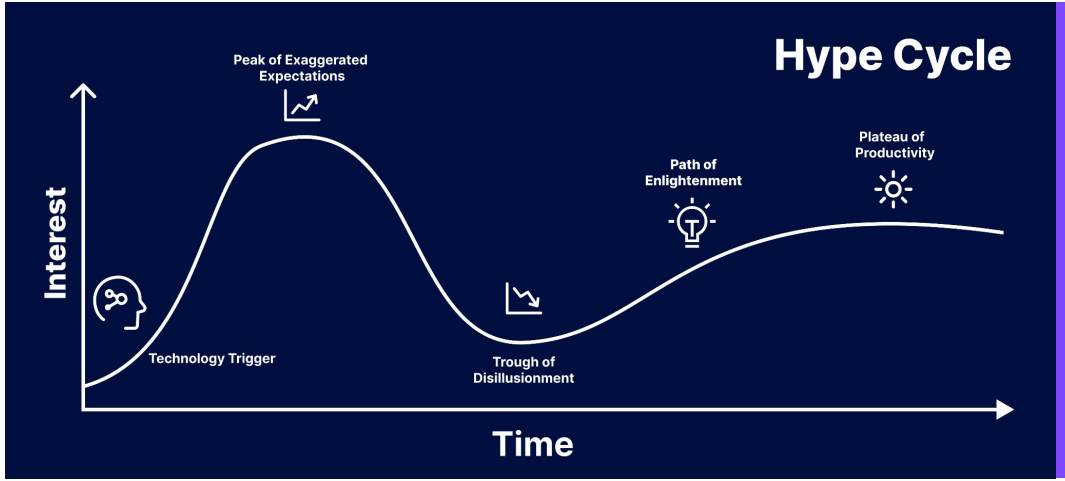
90% of newly generated data in 2025 will be unstructured data

 |

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

zilliz

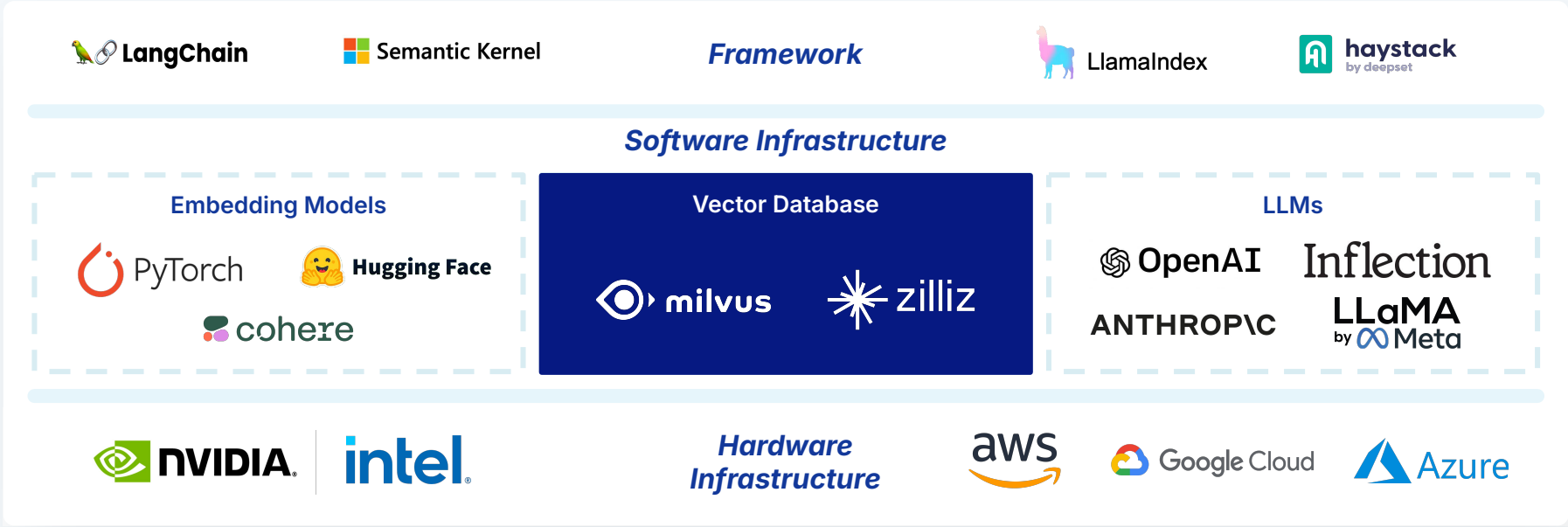
AI Hype?



What do these companies that navigated the "trough of disillusionment" have in common?

Data Volumes.

Well-connected in LLM infrastructure to enable RAG use cases



zilliz

New Hotness

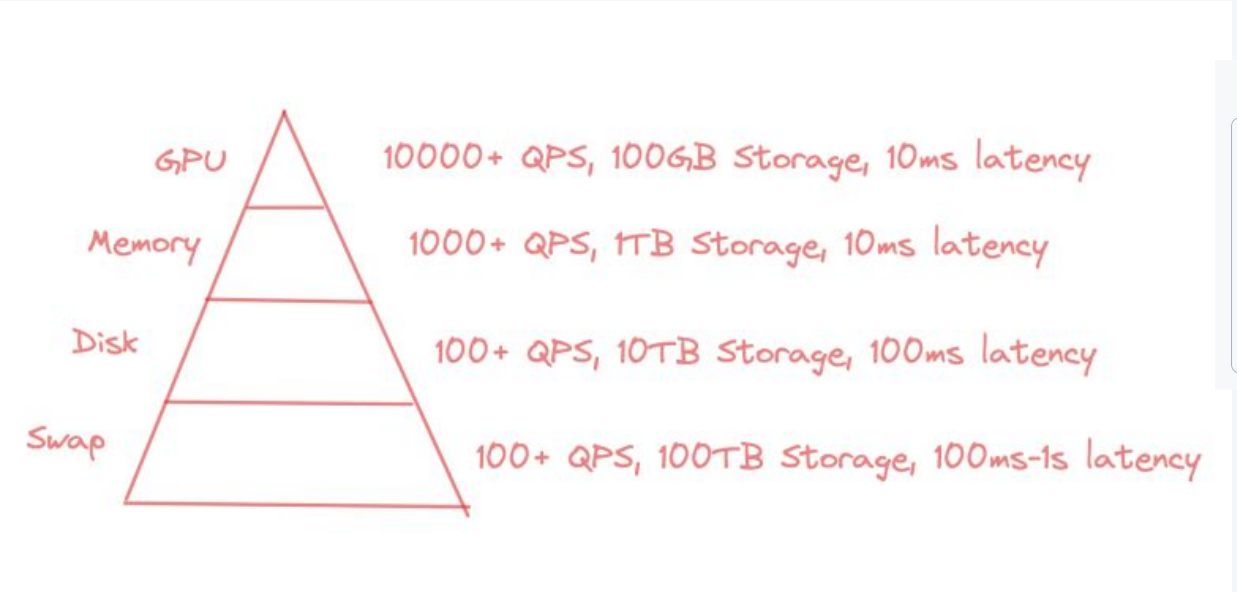
https://zilliz.com/learn/top-10-best-multimodal-ai-models-you-should-know

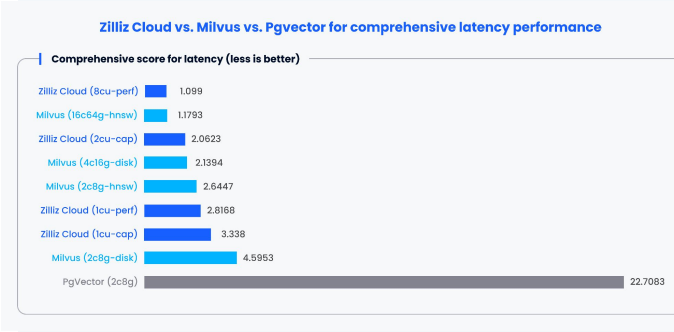
https://github.com/facebookresearch/ImageBind

zilliz

Vector vs Relational

https://zilliz.com/blog/relational-databases-vs-vector-databases





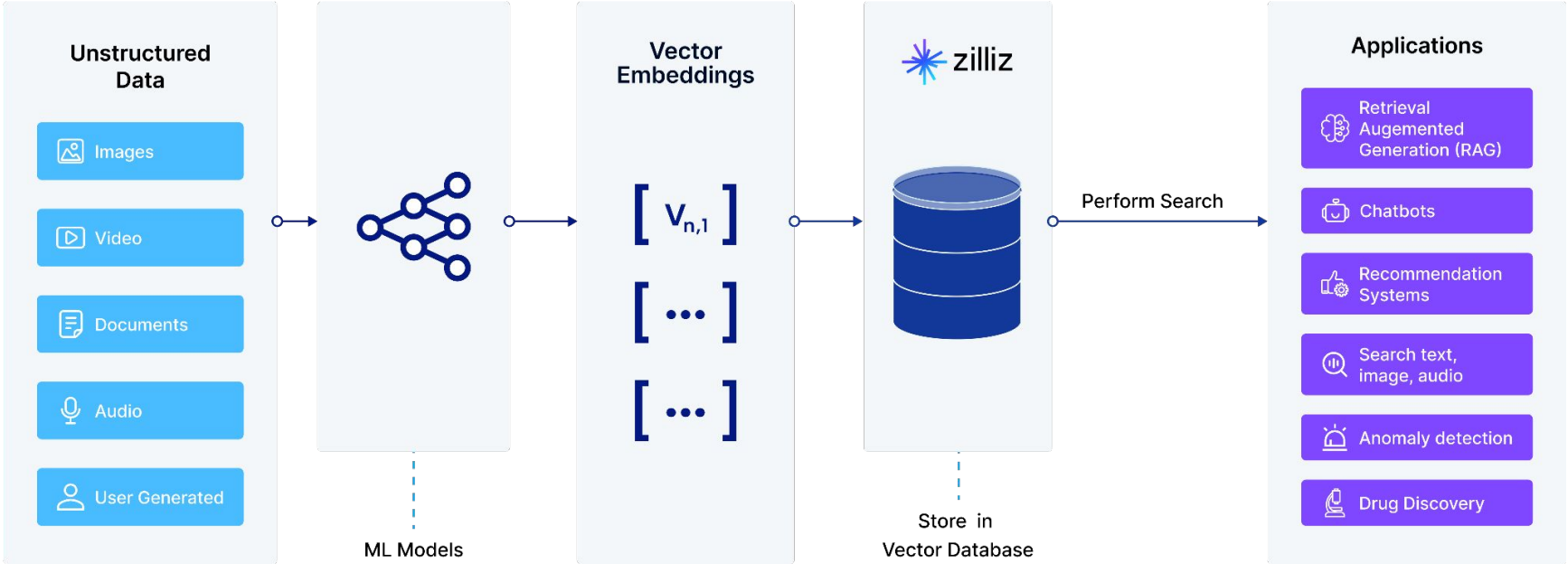
zilliz

V

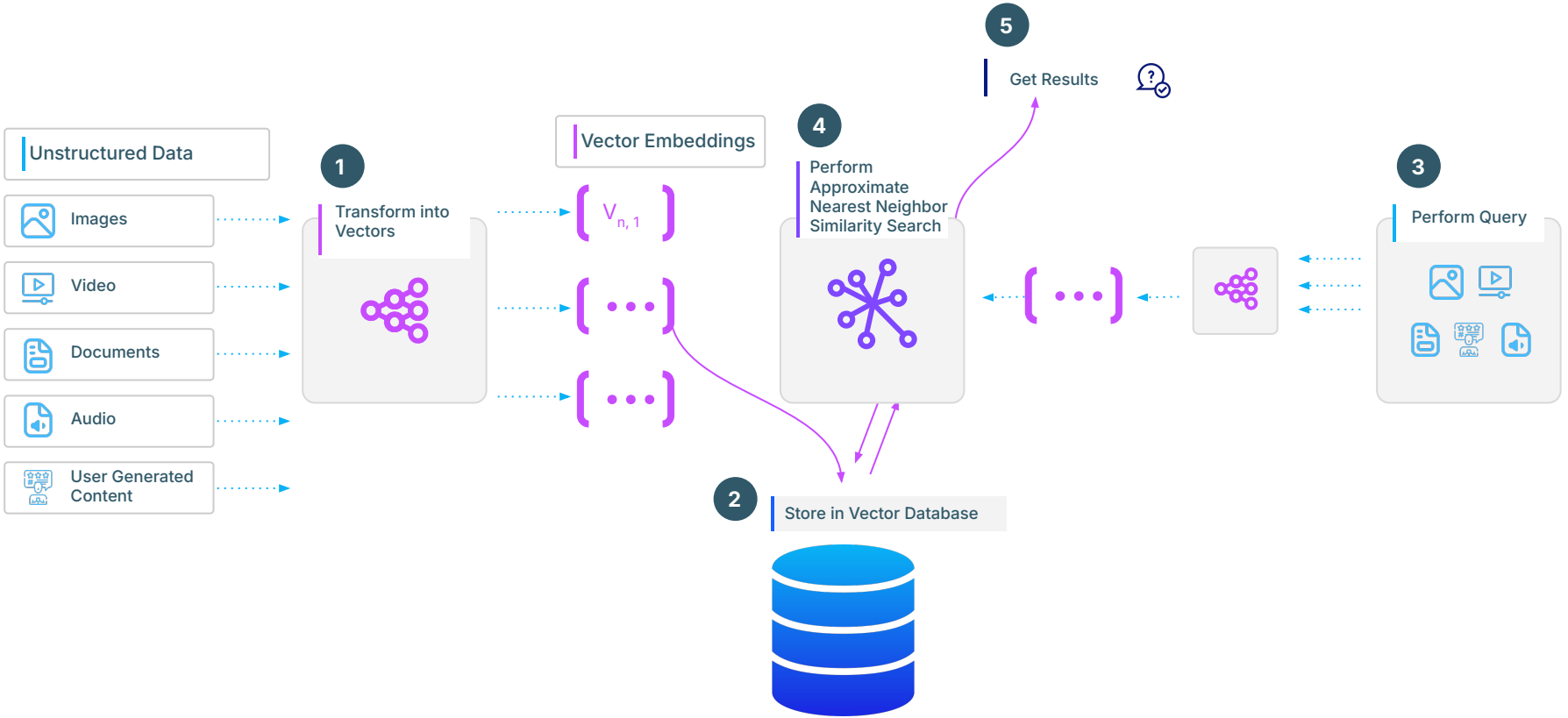
Overview of Vector Databases

zilliz

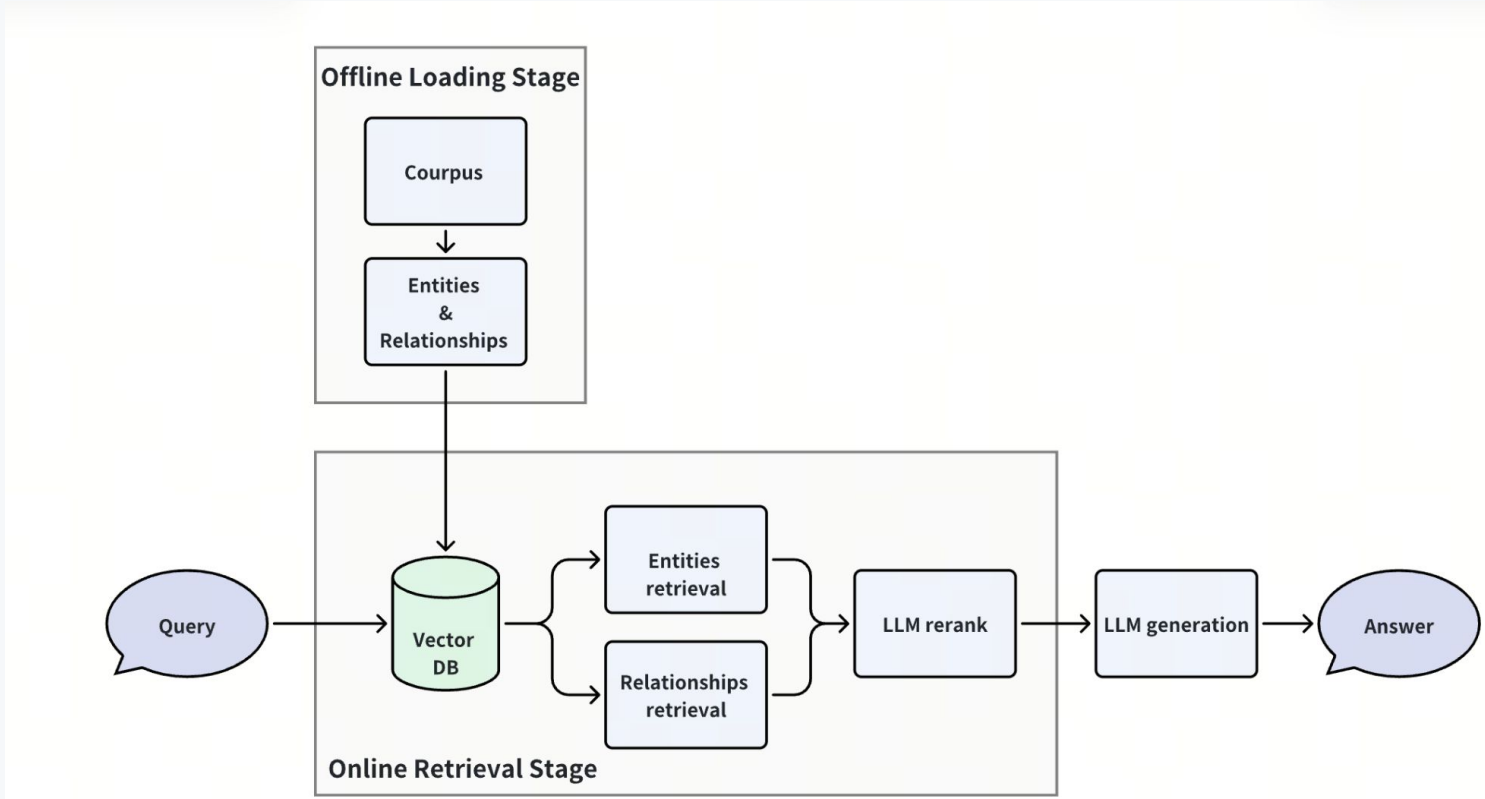
A New tool emerged. The Vector Database



How Similarity Search Works



Vector Database : making sense of unstructured data



zilliz

Vector Search
+ Indexing
+ Filtering
= 🚀

zilliz

A Quick Introduction to Milvus

zilliz

Mission:

Helping organizations make sense of unstructured data.

2017
Founded

$113M
Raised

140+
Employees

Redwood City, CA
Headquarters

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.

400+ contributors

29K+ stars

66M+ docker pulls

2.7K+ forks

 |

zilliz

Rich functionality

Dynamic Schema

Float, Binary and
Sparse Vector

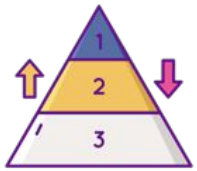
Tag + Vector
Optimized filtering

Hybrid Search
Sparse + Dense

RBAC
TLS, Encryption

Million level
Tenants support

Disk based Index



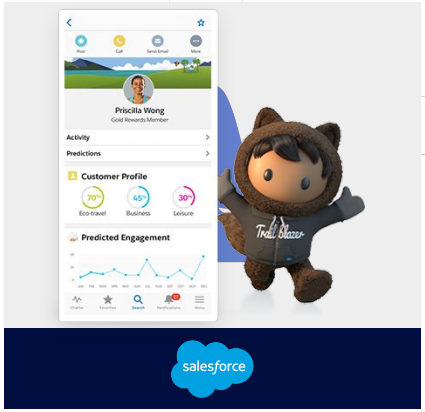
Tiered Storage

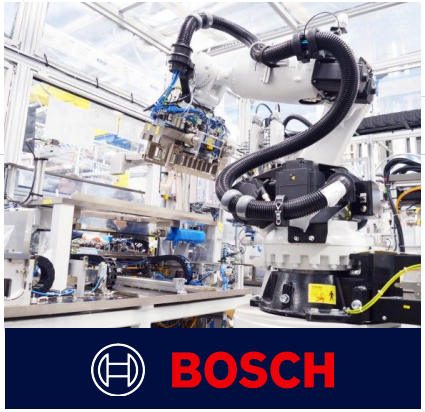
Support bulk import

GPU Support
Intel + ARM Cpu Support

Industry leaders already use vector search in their apps

Use Case: Data Search	**Use Case**: Drug Discovery	**Use Case**: Image Search	**Use Case:** Recommender System
Vectors: 2 Billion	**Vectors:** 12 Billion	**Vectors**: 20 Billion	**Vectors**: 20 Billion
Req'ts: 200 ms, Cost mgmt	**Req'ts:** High Recall	**Req'ts**: High Insertion, Cost	**Req'ts**: 5,000 QPS
Index: DiskANN for cost savings	**Index:** BIN_FLAT	**Index**: Disk Based Index	**Index**: HNSW & CAGRA





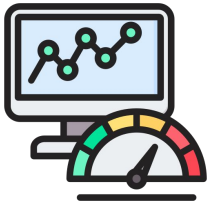
 |

zilliz

That's why we build Milvus

And it's open sourced under Apache license!

Fast & Cost effective



3X faster, 3X Cheaper

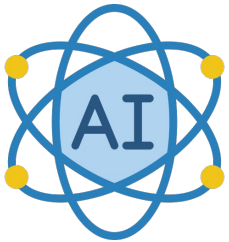
Pluggable Vector Search Lib
Tiered Storage

Scalable & Reliable

Cloud Native, K8s Native

Scale from 1 - 10B
Storage / compute disaggregation

AI Powered

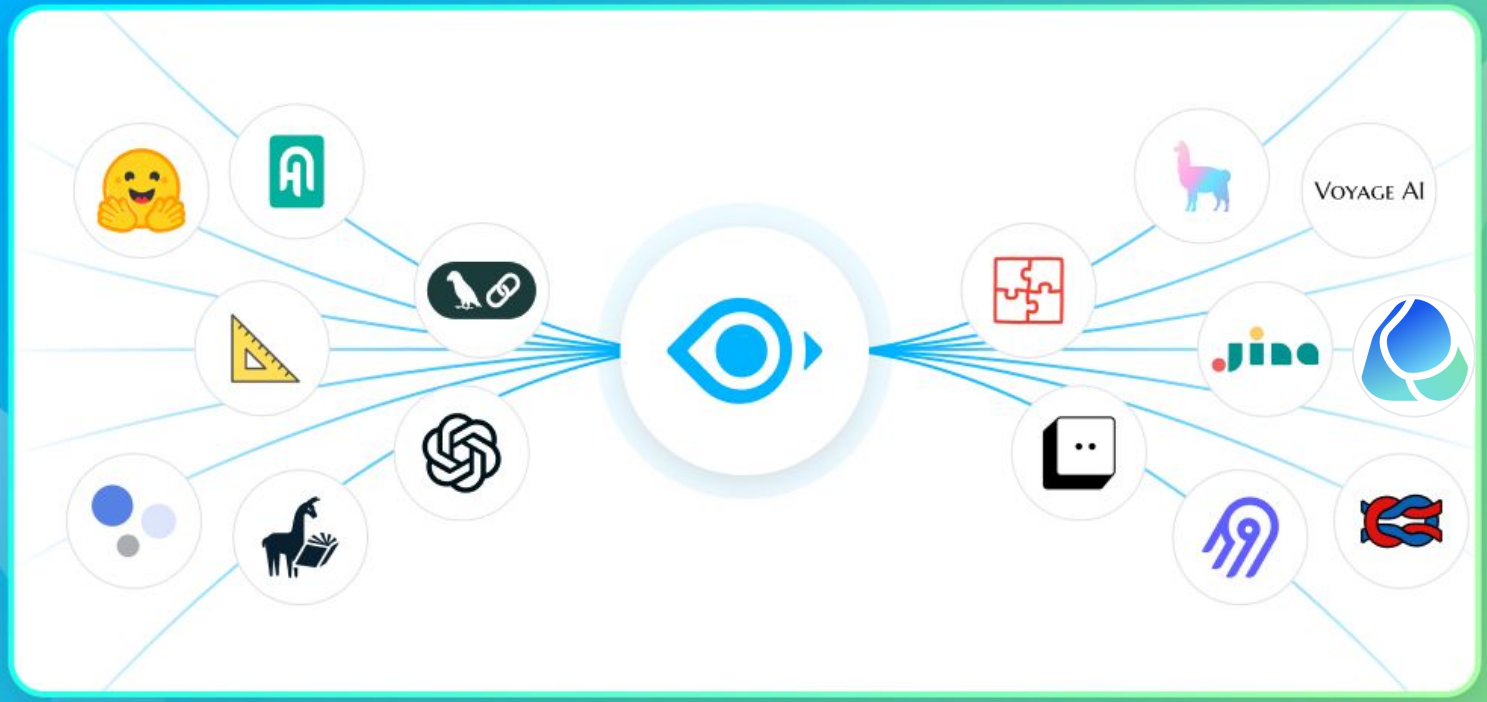


Vector Native

Rich functionality for AI
Born for vector data processing

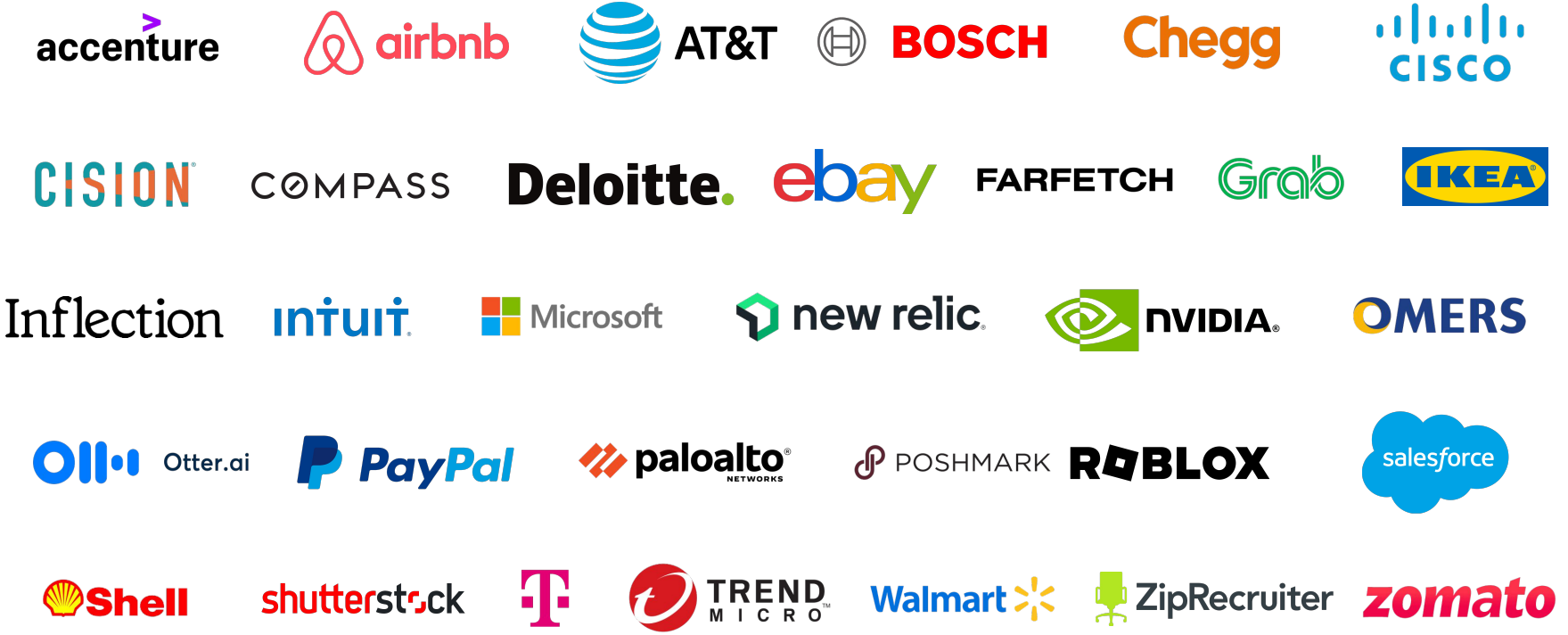
zilliz

Seamless integration with all popular AI toolkits



milvus

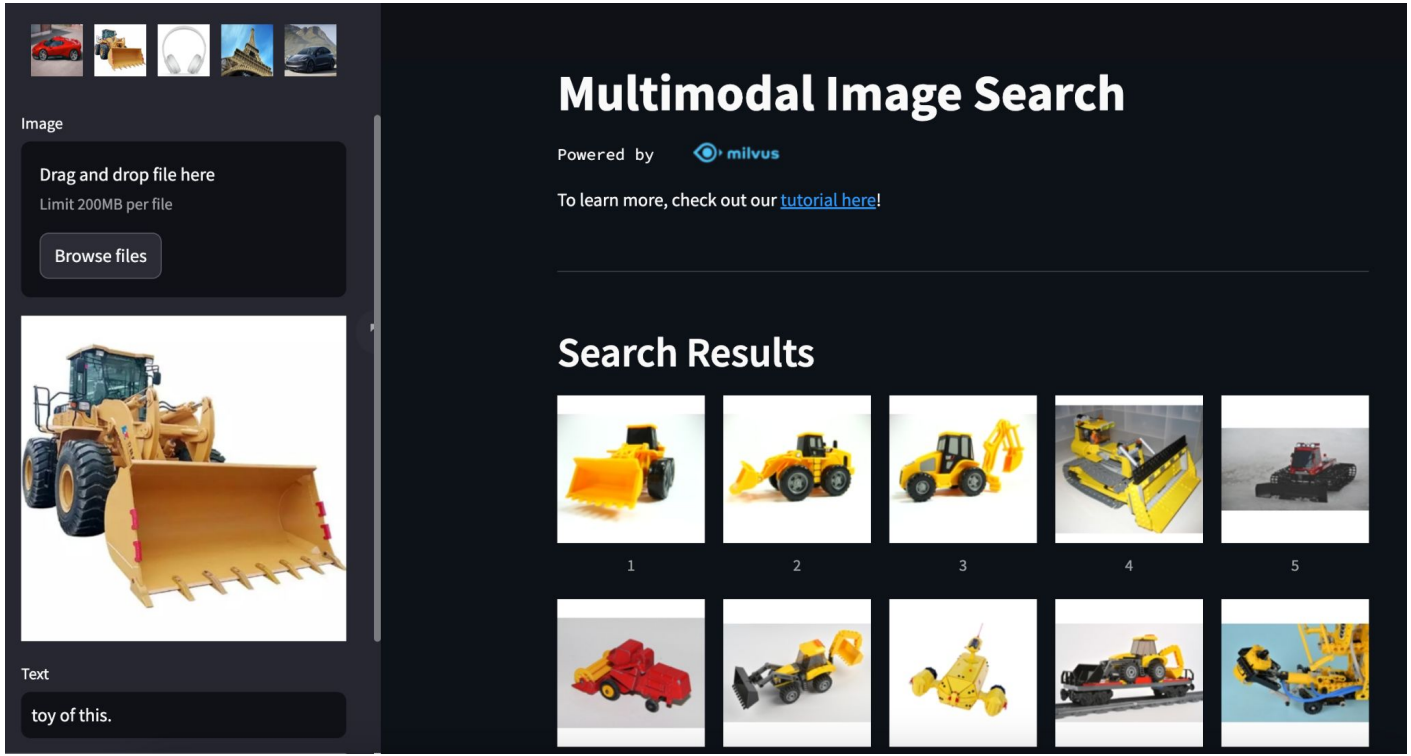
Sample of Milvus Users



 |

zilliz

Multi-modal Search



multimodal-demo.milvus.io

zilliz

zilliz

The Landscape of GenAI Ecosystem: Beyond LLMs and Vector Databases

Read Blog

Workflow Orchestration / Optimization:
LangChain, LlamaIndex, Haystack, DSPy, Semantic Kernel

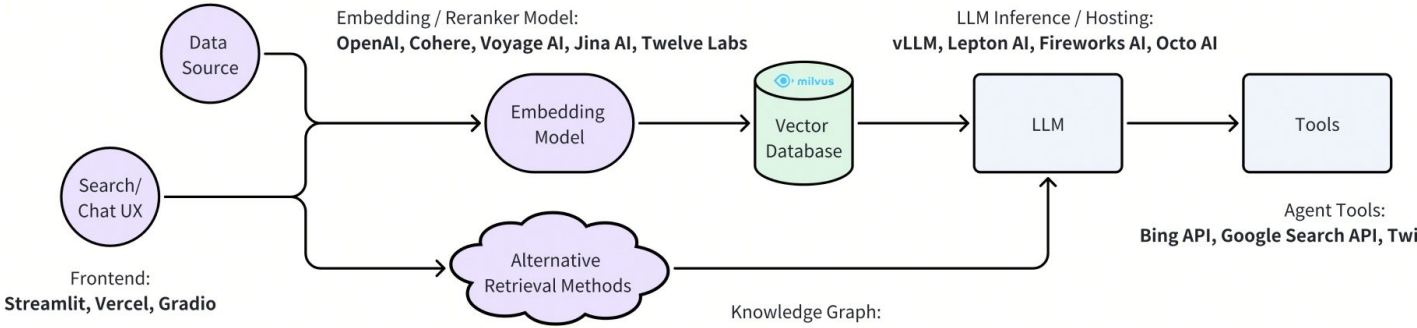
Quality Evaluation / Observability:
Ragas, Arize, Langfuse, Relari AI, Giskard, DeepEval

Agent and Memory Management:
MemGPT, Mem0, Camel, AutoGPT, CrewAI, AutoGen

Data Connector / Web Scrapping:
Airbyte, Fivetran, Apify, Zyte

Embedding / Reranker Model:
OpenAI, Cohere, Voyage AI, Jina AI, Twelve Labs

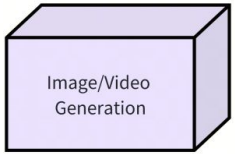
LLM Inference / Hosting:
vLLM, Lepton AI, Fireworks AI, Octo AI



Agent Tools:
Bing API, Google Search API, Twillio, OpenAPI

Frontend:
Streamlit, Vercel, Gradio

Knowledge Graph:
WhyHow, GraphRAG, HippoRAG



Out-of-the-box / Low-code RAG Service:
AnythingLLM, PrivateGPT, Dify, Shakudo, Vectara, Epsilla, FlowiseAI

Image / Video Generation, Digital Twin:
Midjourney, Stability AI, Runway, Pika, HeyGen

zilliz

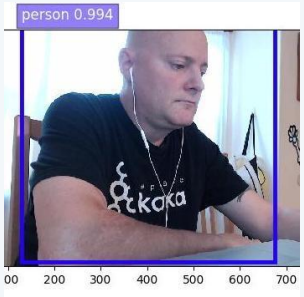
RESOURCES

 |

zilliz

Vector Database Resources

Give Milvus a Star!



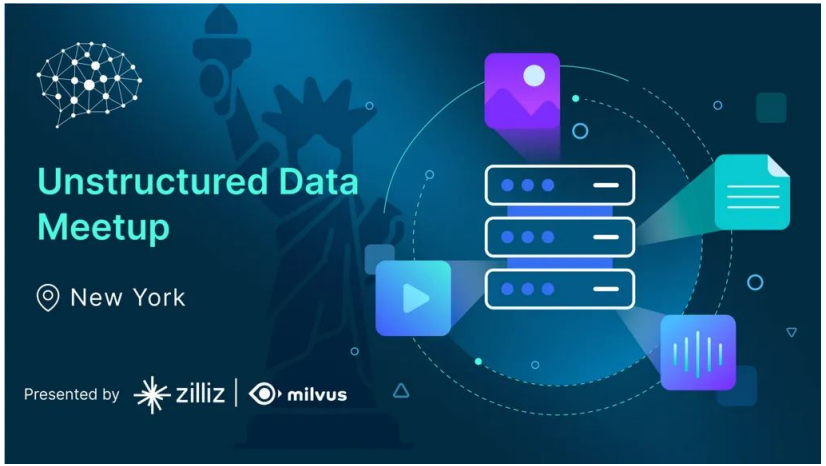
Chat with me on Discord!

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

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

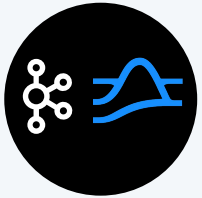
 |

zilliz

Milvus 🤝 Open-Source

MINIO

Store Vectors and Indexes
Enables Milvus' stateless architecture



Kafka/ Pulsar

Handles Data Insertion stream
Internal Component Communications
Real-time updates to Milvus

Prometheus / Grafana

Collects metrics from Milvus
Provides real-time monitoring dashboards

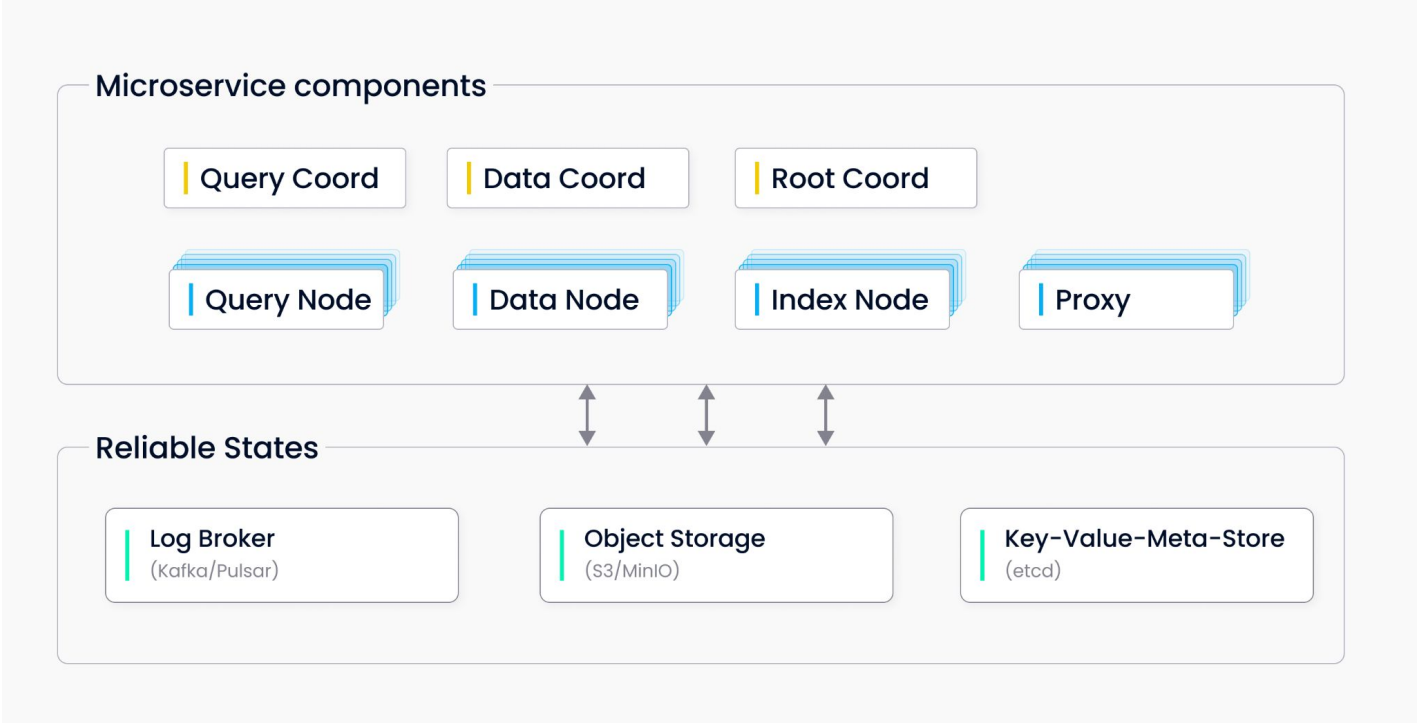
Kubernetes

Milvus Operator CRDs

 |

zilliz

Distributed Architecture



zilliz

Dynamic Scaling 🚀

Stateless components for **Easy Scaling**

- **Query**, **Index**, and **Data Nodes** can be scaled **independently**
- Allows for **optimized resource allocation** based on workload characteristics

Data sharding across multiple nodes

- **Distributes** large **datasets** across multiple **Data Nodes**
- Enables **parallel processing** for improved query performance

Horizontal Pod Autoscaler (**HPA**)

- **Automatically scales** up and down
- **Custom metrics** can be used (e.g., query latency, throughput)

 |

zilliz

Stateless Architecture

Stateless Components

All Milvus components are deployed **Stateless**.

Object Storage

Milvus relies on **Object Storage** (MinIO, S3, etc) for data **persistence**.
Vectors are stored in **Object Storage**, **Metadata** is in **etcd.**

Scaling and Failover

Scaling and failover **don't** involve traditional **data rebalancing.** When **new pods** are added or existing ones fail, they can immediately start handling requests by **accessing data** from the **shared object storage.**

zilliz

Different Consistency levels

Ensures every node or replica has the same view of data at a given time.

- **Strong**: Guaranteed up-to-date reads, highest latency
- **Bounded**: Reads may be slightly stale, but within a time bound
- **Session**: Consistent reads within a session, may be stale across sessions
- **Eventually**: Lowest latency, reads may be stale

Trade Offs

- **Strong** consistency for **critical applications** requiring accurate results
- **Eventually** consistency for **high-throughput**, latency-sensitive apps

zilliz

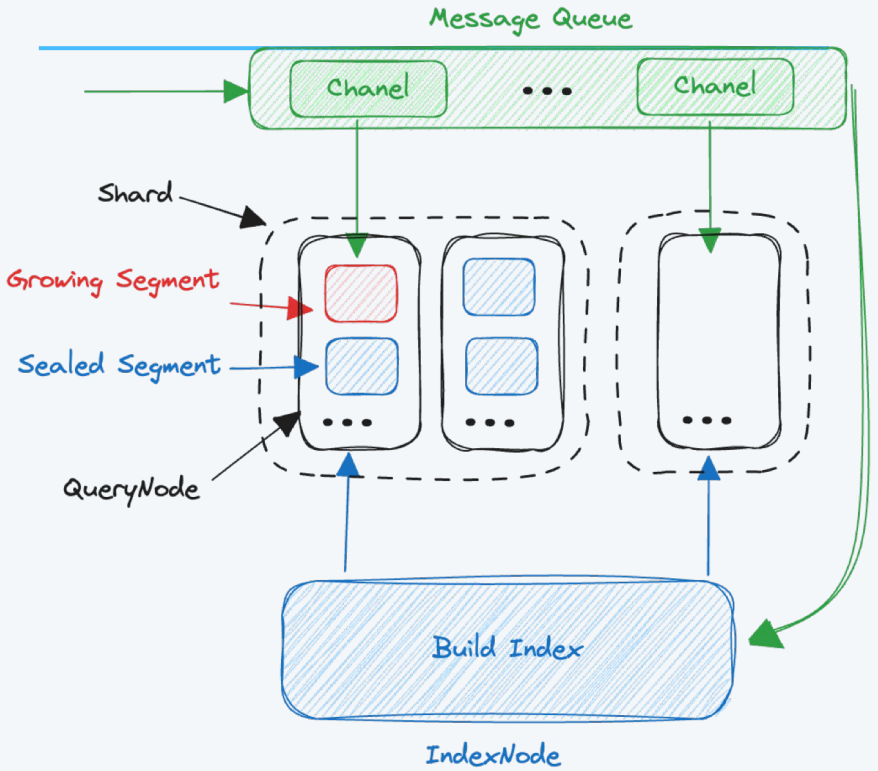
Milvus Data Layout - **Segments**

Growing Segment:

- In-memory segment replaying data from the Log Broker.
- Uses a FLAT index to ensure data is fresh and appendable.

Sealed Segment:

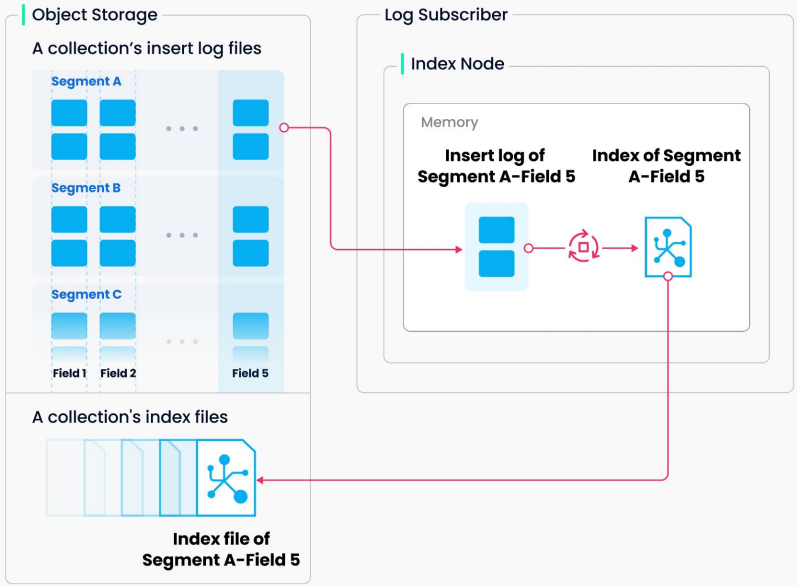
- Immutable segment using alternative indexing methods for efficiency.



 |

zilliz

Index Building



To **avoid frequent index building** for data updates.

A **collection** in Milvus is **divided** further **into segments**, each with its own index.

 |

zilliz

Picking an Index

- 100% Recall – Use `FLAT` search if you need 100% accuracy
- 10MB < `index_size` < 2GB – Standard IVF
- 2GB < `index_size` < 20GB – Consider `PQ` and `HNSW`
- 20GB < `index_size` < 200GB – Composite Index, `IVF_PQ` or `HNSW_SQ`
- Disk-based indexes

 |

zilliz

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.

New Stuff

The Zilliz Sep '24 Launch

Supercharge Your GenAI with Production-Ready Data Infrastructure

Data Sovereignty

Migration Services

Secure, complete and easy data migration to Zilliz from Milvus, PGVector, Elastic or between Zilliz clusters

Fivetran Source Connector

Enable unstructured data retrievals from 500+ system through Fivetran

Fivetran

High-performance

Auto-Scale *(Private Preview)*

Dynamically increase cluster capacity based on workload to reduce operational burden and ensure continuous service

Multi-replica *(Public Preview)*

Replicated collections for increased throughput and fault tolerance

Security & Reliability

Metrics & Alerts

18 core metrics for proactive issue identification and 39 alerts to ensure optimal performance

Auth0-based SSO

(Private Preview)

Simplify IT management and improve security w/ Google/Github/SSO/Email login

99.95% Uptime SLAs

zilliz

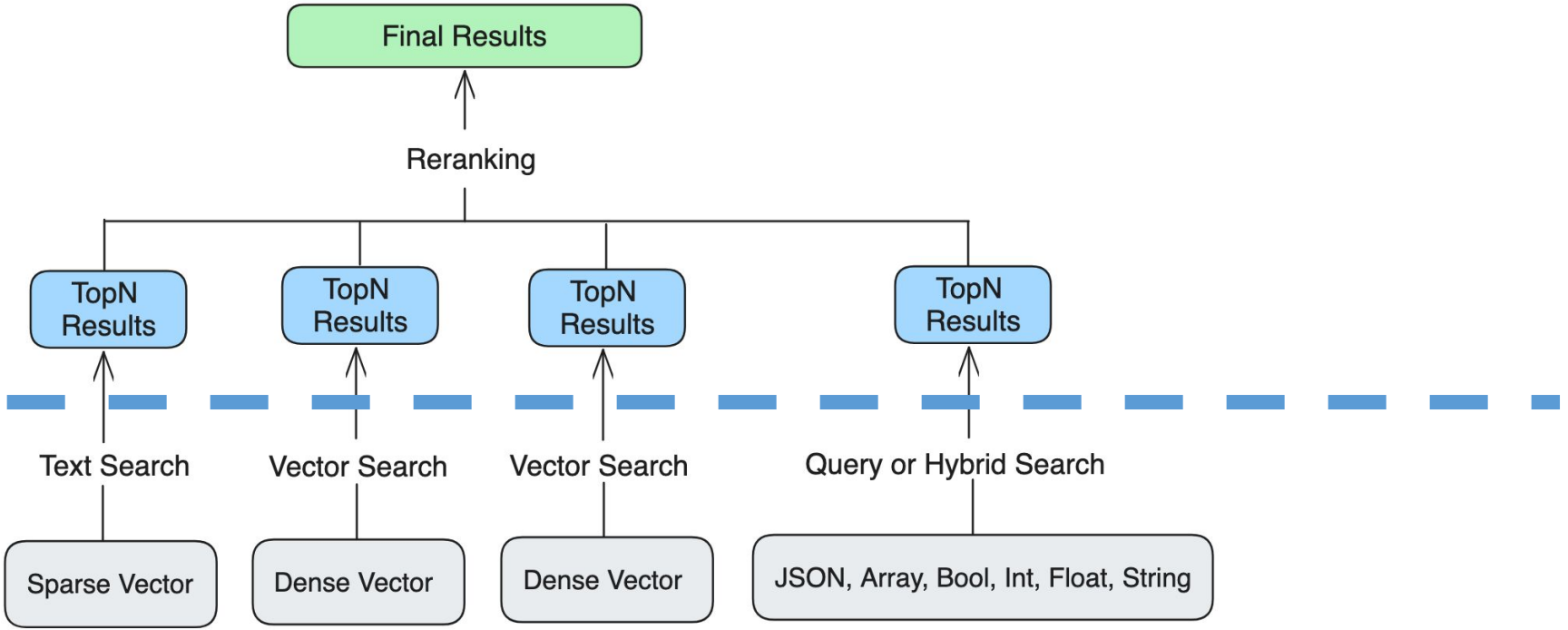
New Stuff

https://github.com/milvus-io/milvus-sdk-java/releases/tag/v2.4.4

Milvus 2.4 introduces several new features and improvements:

1. **New GPU Index - CAGRA**: This GPU-based index offers significant performance improvements, especially for batch searches
2. **Multi-vector and Hybrid Search**: This feature allows storing vector embeddings from multiple models and conducting hybrid searches.
3. **Sparse Vectors Support (Beta)**: Milvus now supports sparse vectors for processing in collections, which is particularly useful for keyword interpretation and analysis
4. **Grouping Search**: This feature enhances document-level recall for Retrieval-Augmented Generation (RAG) applications by providing categorical aggregation
5. **Inverted Index and Fuzzy Matching**: These capabilities improve keyword retrieval for scalar fields
6. **Float16 and BF16 Vector Data Type Support**: Milvus now supports these half-precision data types for vector fields, which can improve query efficiency and reduce memory usage.
7. **L0 Segment**: This new segment is designed to record deleted data, enhancing the performance of delete and upsert operations.
8. **Refactored BulkInsert**: The bulk-insert logic has been improved, allowing for importing multiple files in a single bulk-insert request.

Hybrid Search



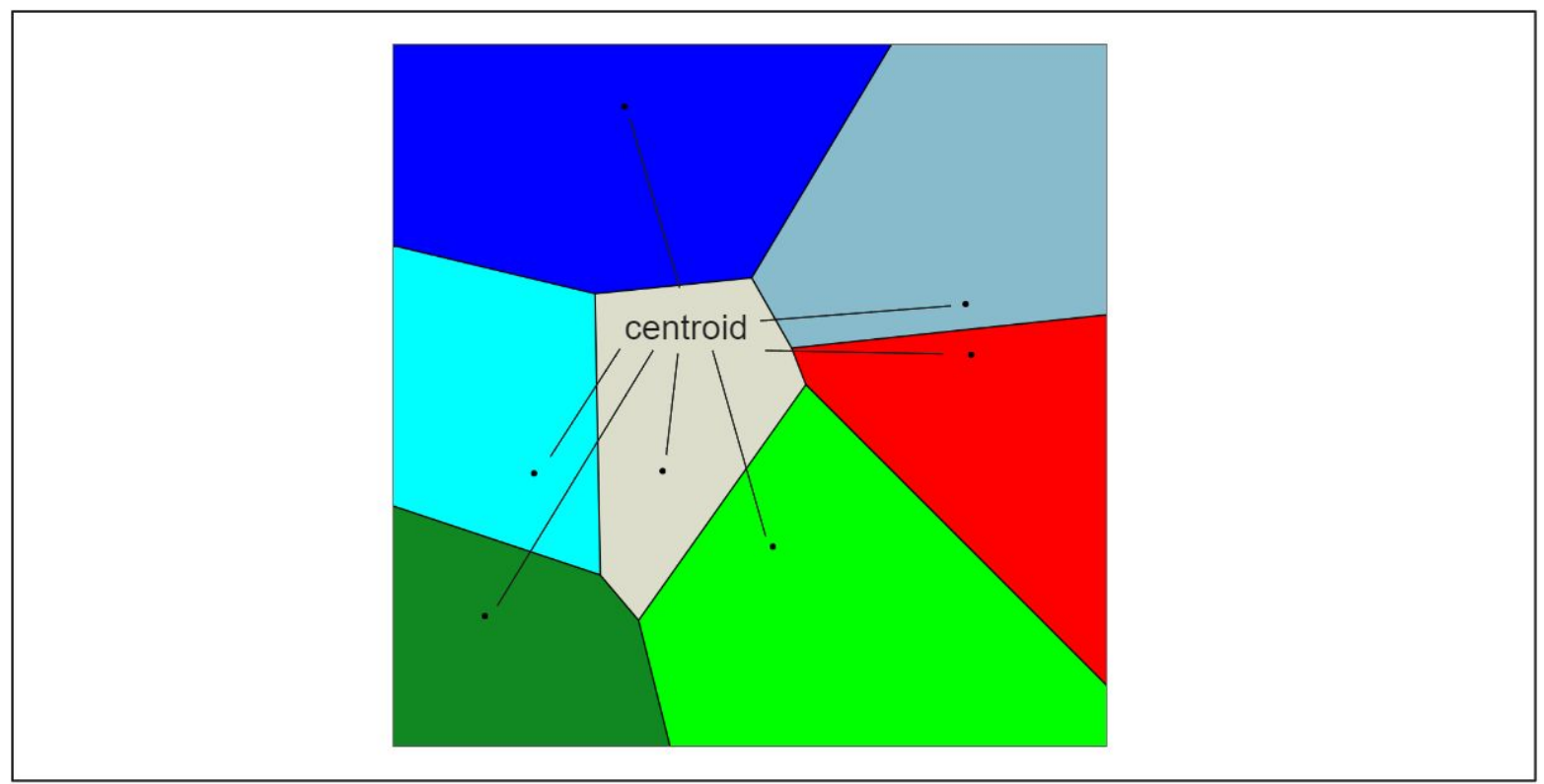
zilliz

Inverted File FLAT
(IVF-FLAT)

 |

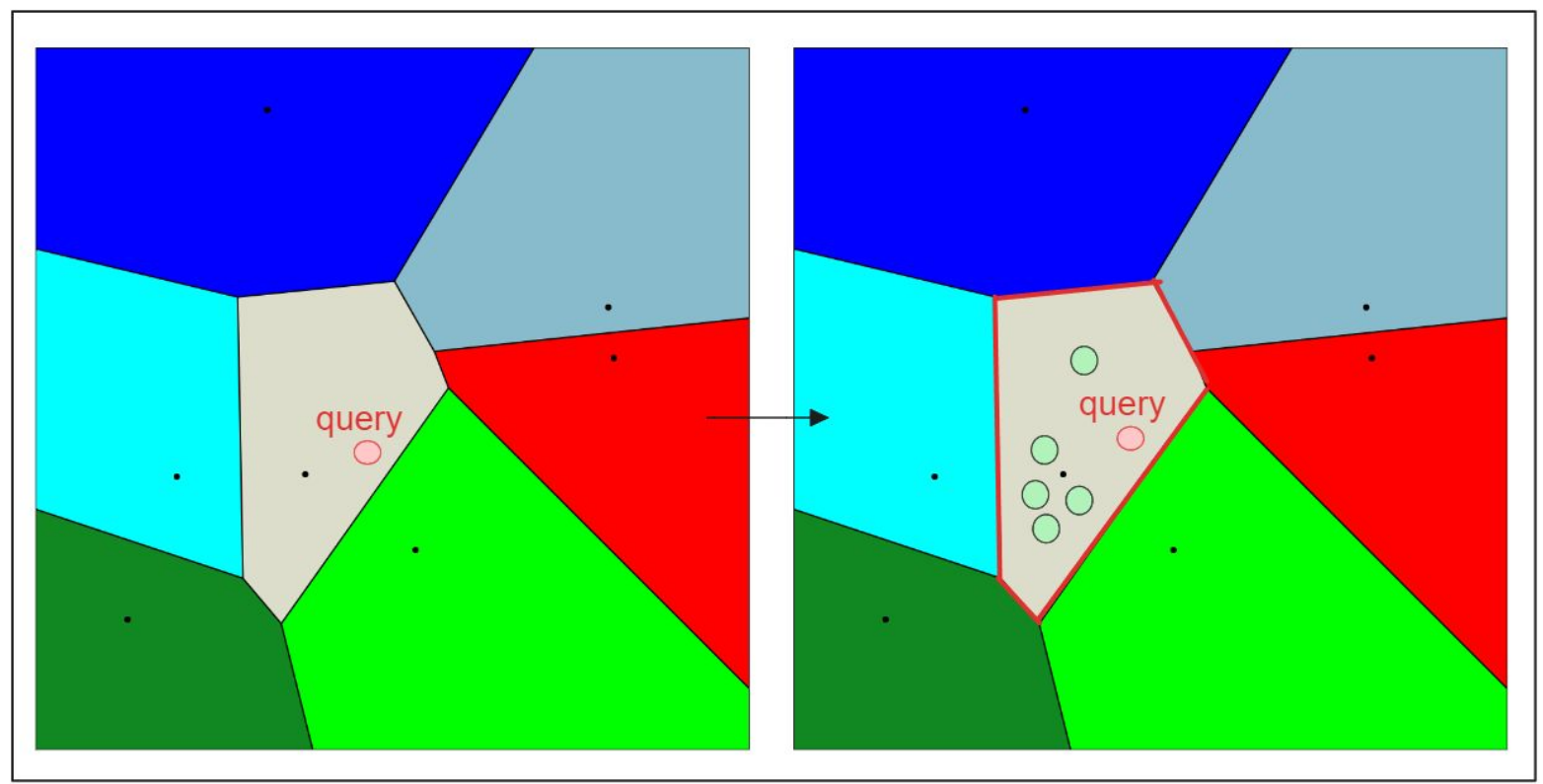
zilliz

IVF-FLAT Index



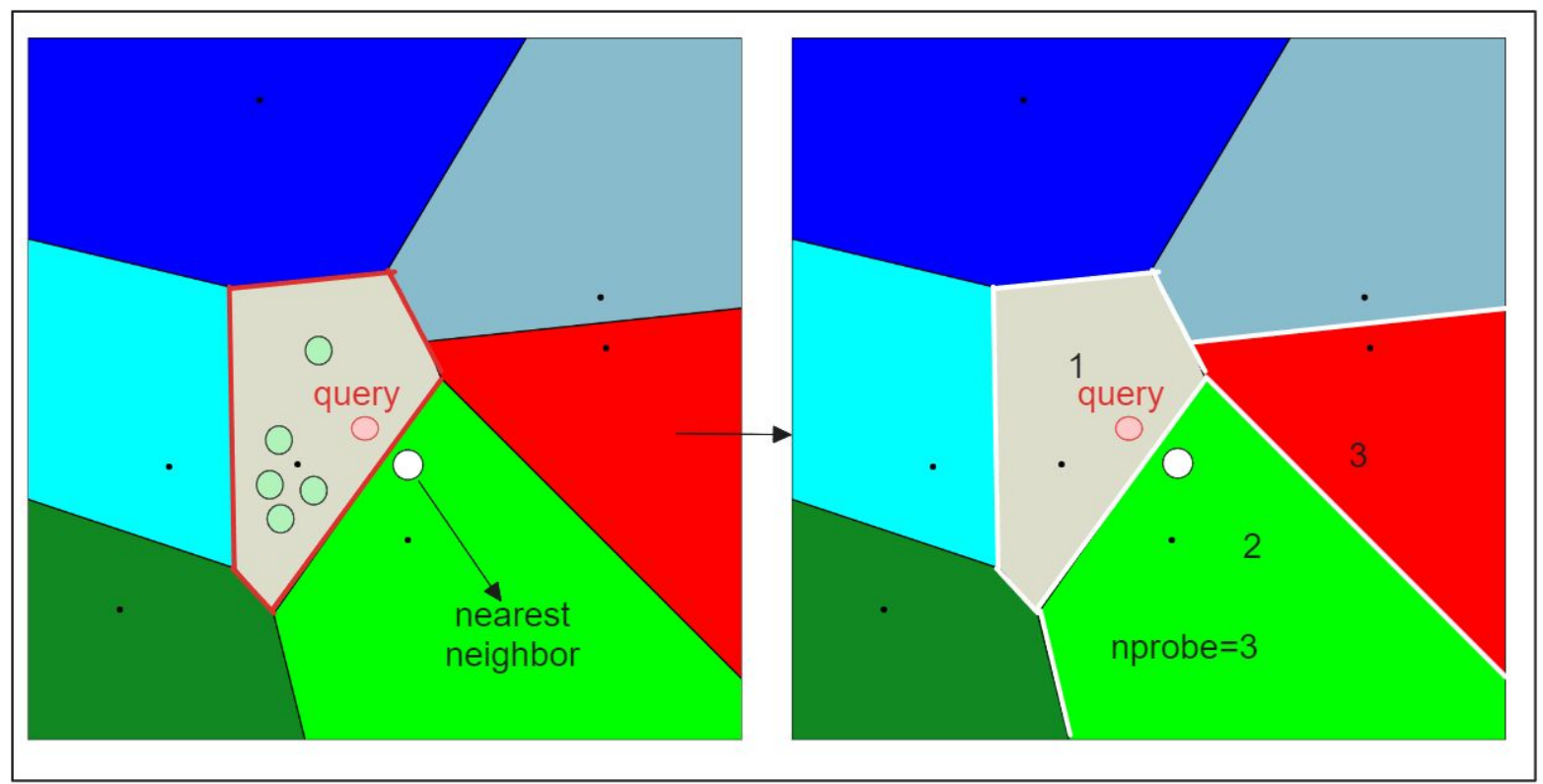
zilliz

IVF-FLAT Index



zilliz

IVF-FLAT Index

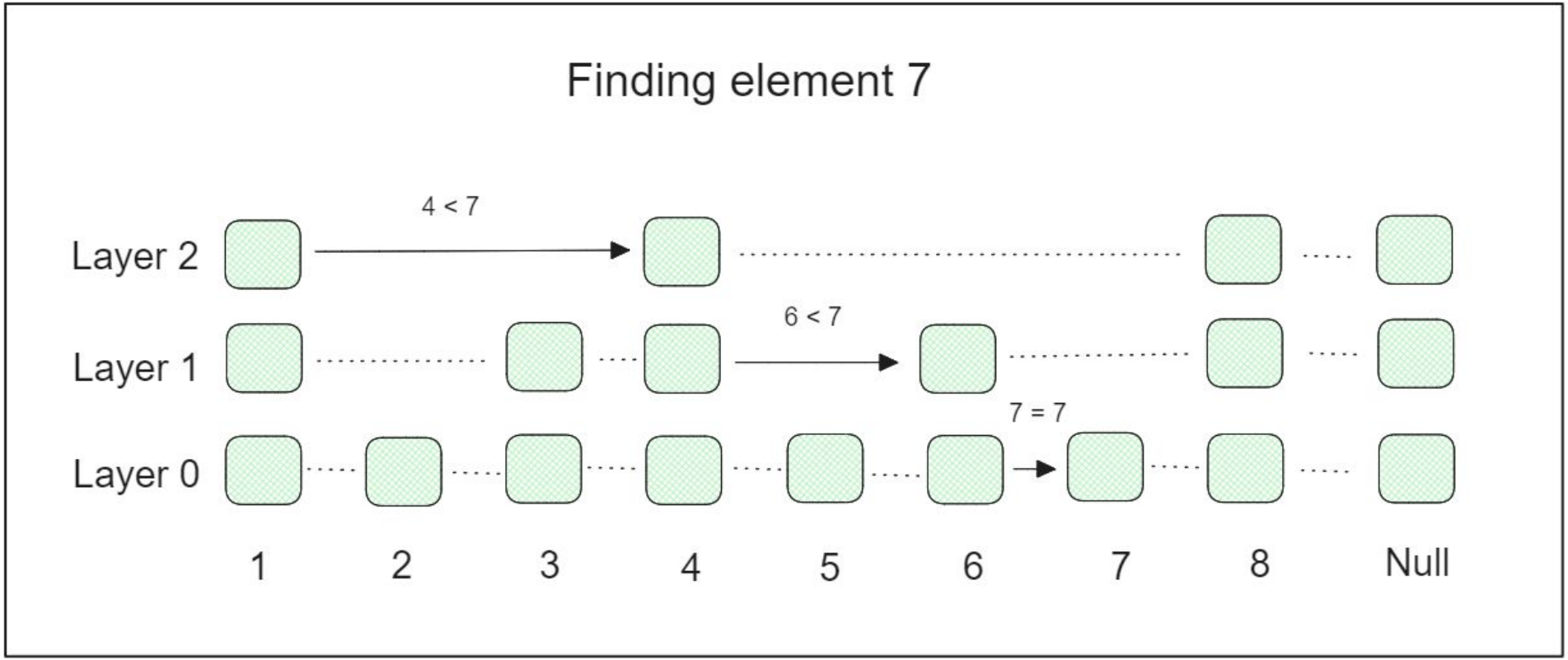


zilliz

Hierarchical Navigable Small World (HNSW)

zilliz

HNSW - Skip List



 |

zilliz

HNSW - NSW Graph

- Built by randomly shuffling data points and inserting them one by one, with each point connected to a predefined number of edges (M).

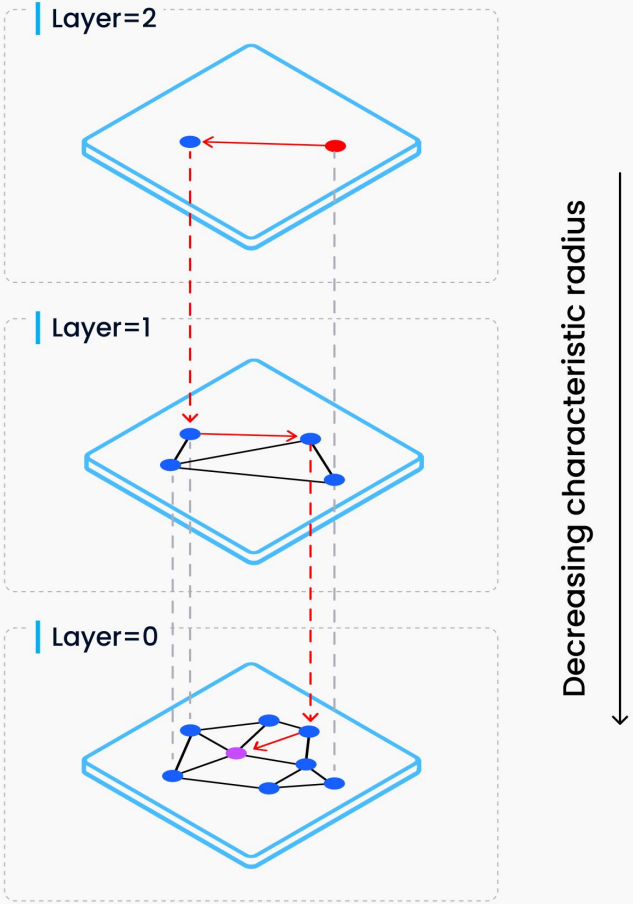
  ⇒ Creates a graph structure that exhibits the "small world".

  ⇒ Any two points are connected through a relatively short path.

 |

zilliz

HNSW



 |

zilliz

Filtering

 |

zilliz

Filtering on Metadata

- Search Space Reduction w/ **Pre-Filtering**
- Bitset Wizardry 🧙
  - Use Compact Bitsets to represent Filter Matches
  - Low-level CPU operations for speed
- Scalar Indexing
  - Bloom Filter
  - Hash
  - Tree-based

 |

zilliz