zilliz

Zilliz Cloud

$whoami

- Solutions Architect with Zilliz
- Previously
  - SA with Databricks, Confluent
  - Data scientist @ GIC
  - Data and AI practitioner for ~10 years

zilliz

Mission:

Helping organizations make sense of unstructured data.

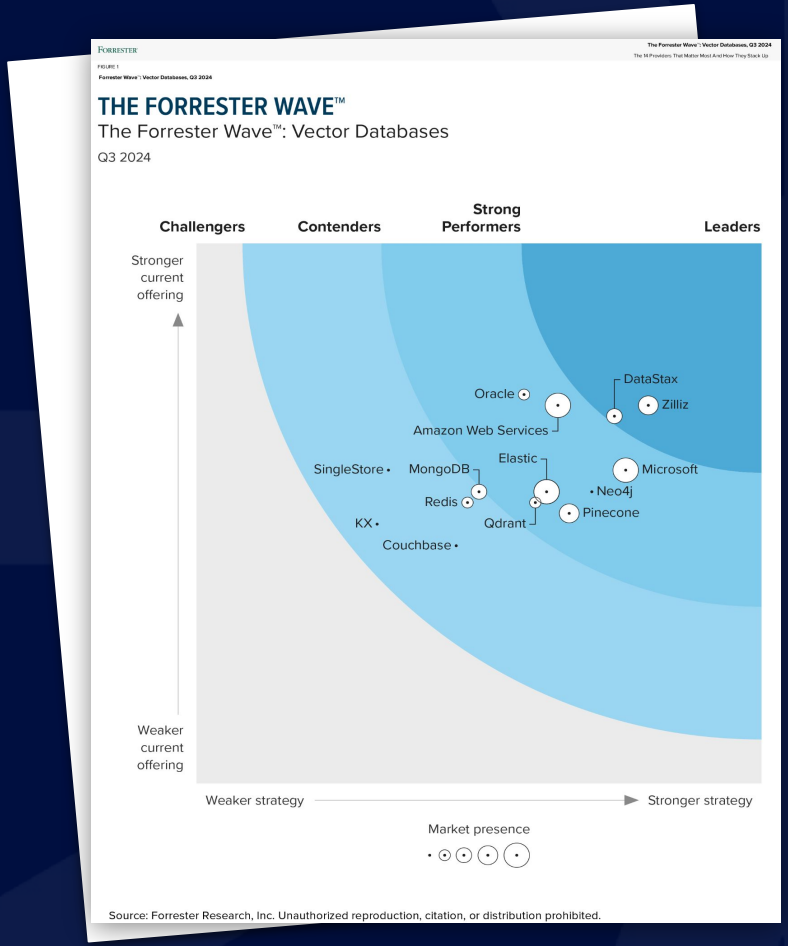
2017
Founded

$113M
Raised

140+
Employees

Redwood City, CA
Headquarters

zilliz



The top industry analysts endorse our leading position

Specifically, we perform top at the following core vector database competency categories

- **Vector Index** - *the only vendor in the market leverage multiple types of index and benefit from the unique strengths of each*
- **Performance** - *the only vendor provides advanced performance acceleration through various hardware optimization, query tuning*
- **Scalability** - *the only vendor offer compute storage separation and auto tiered storage*

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.

milvus
LF AI &DATA

283+
contributors

30K+
stars

67M+
docker pulls

2.8K+
forks

zilliz

Built by database & AI experts


Milvus: A Purpose-Built Vector Data Management System

Jianguo Wang*, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, Ruiyi Jiang, Yi Wei, Charles Xie

ABSTRACT

KEYWORDS

1 INTRODUCTION

CCS CONCEPTS


Manu: A Cloud Native Vector Database Management System

arXiv:2206.13843v1 [cs.DB] 28 Jun 2022

ABSTRACT

1 INTRODUCTION

Zilliz was built by a top-tier team of **algorithm and database engineers** with a strong pedigree in developing **high-performance, scalable**, and **highly available** distributed systems, uniquely tailored for **vector search.**

zilliz

Our Customers

zilliz

What is a vector database?

zilliz

What is a vector database?

What are vector embeddings?

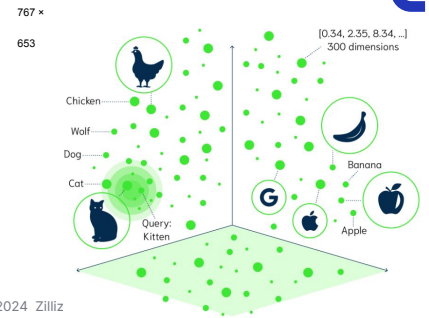
Numerical representations of unstructured data (text, images, audio, video, etc) that captures their semantics and relationships.

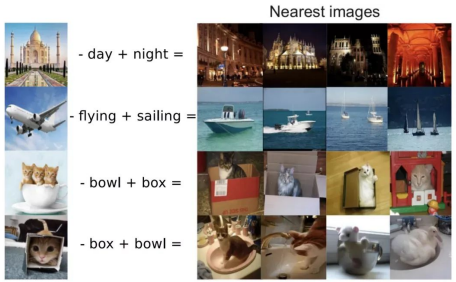
What is a vector database?

Vector database supports efficient storage, search and management of vector embeddings

How to use vector db

extract → insert → index → search / operations

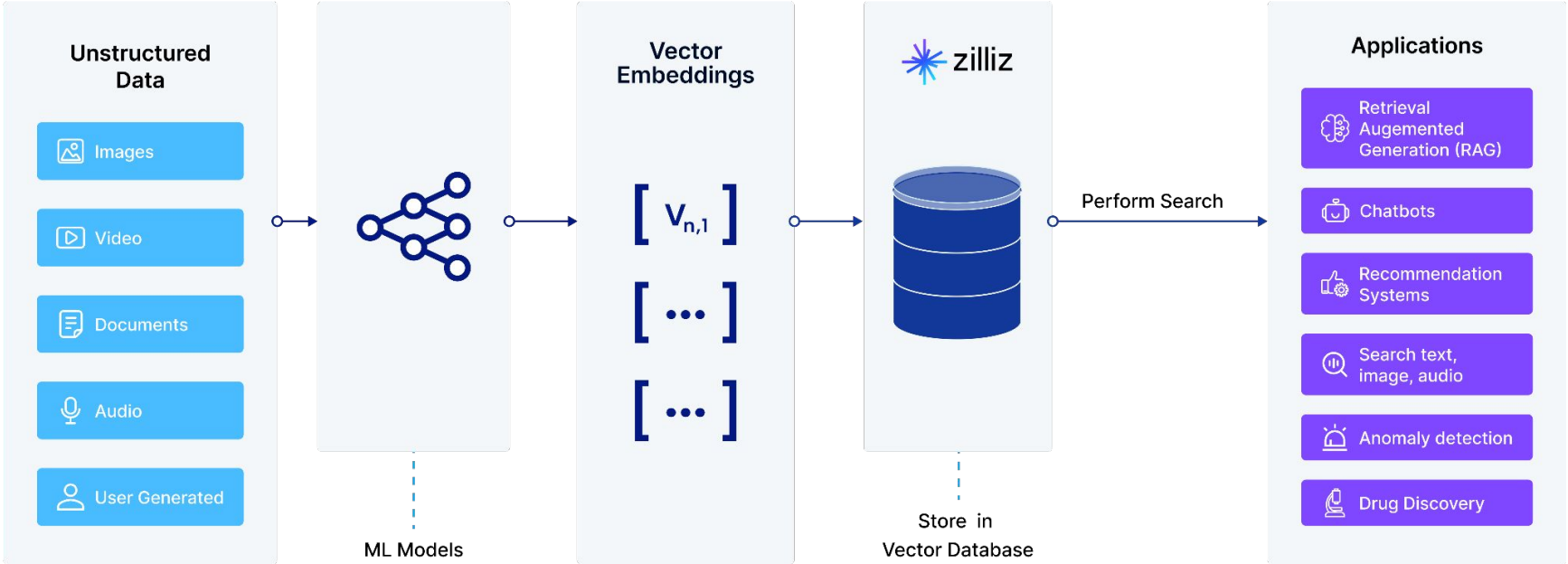




 |

zilliz

A New tool emerged. The Vector Database



 |

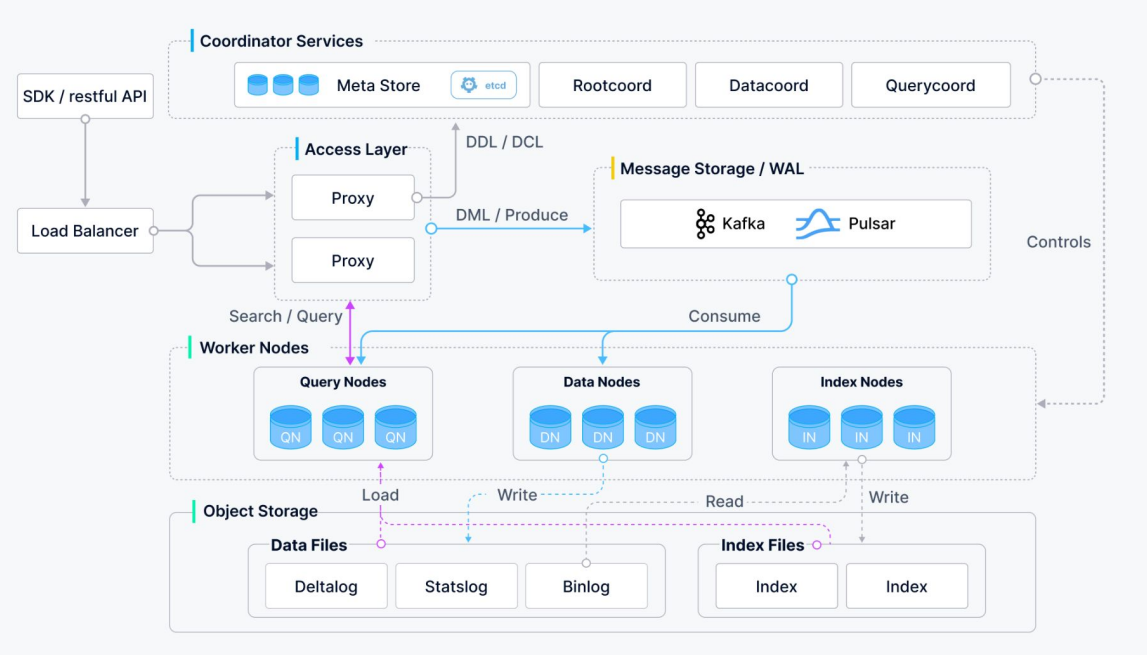
zilliz

Milvus Architecture

Design Principles

- Separation of storage and compute
- Fully depend on mature storage systems
- Microservice - scale by functionality
- Separate streaming and historical data
- Pluggable engine, storage and index
- Log as data

Fully distributed, designed for scalability

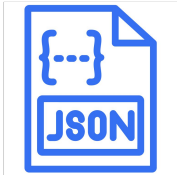


Milvus v2.4.x architecture overview

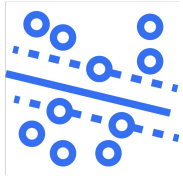
 |

zilliz

Rich functionality



Dynamic Schema



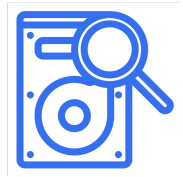
Float, Binary, & Sparse Vector

Tag+Vector Optimized Filtering

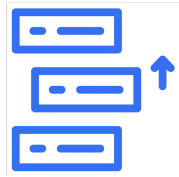
Hybrid Search Dense & Sparse

RBAC, TLS, Encryption

Million+ level tenant support



Disk Based Index

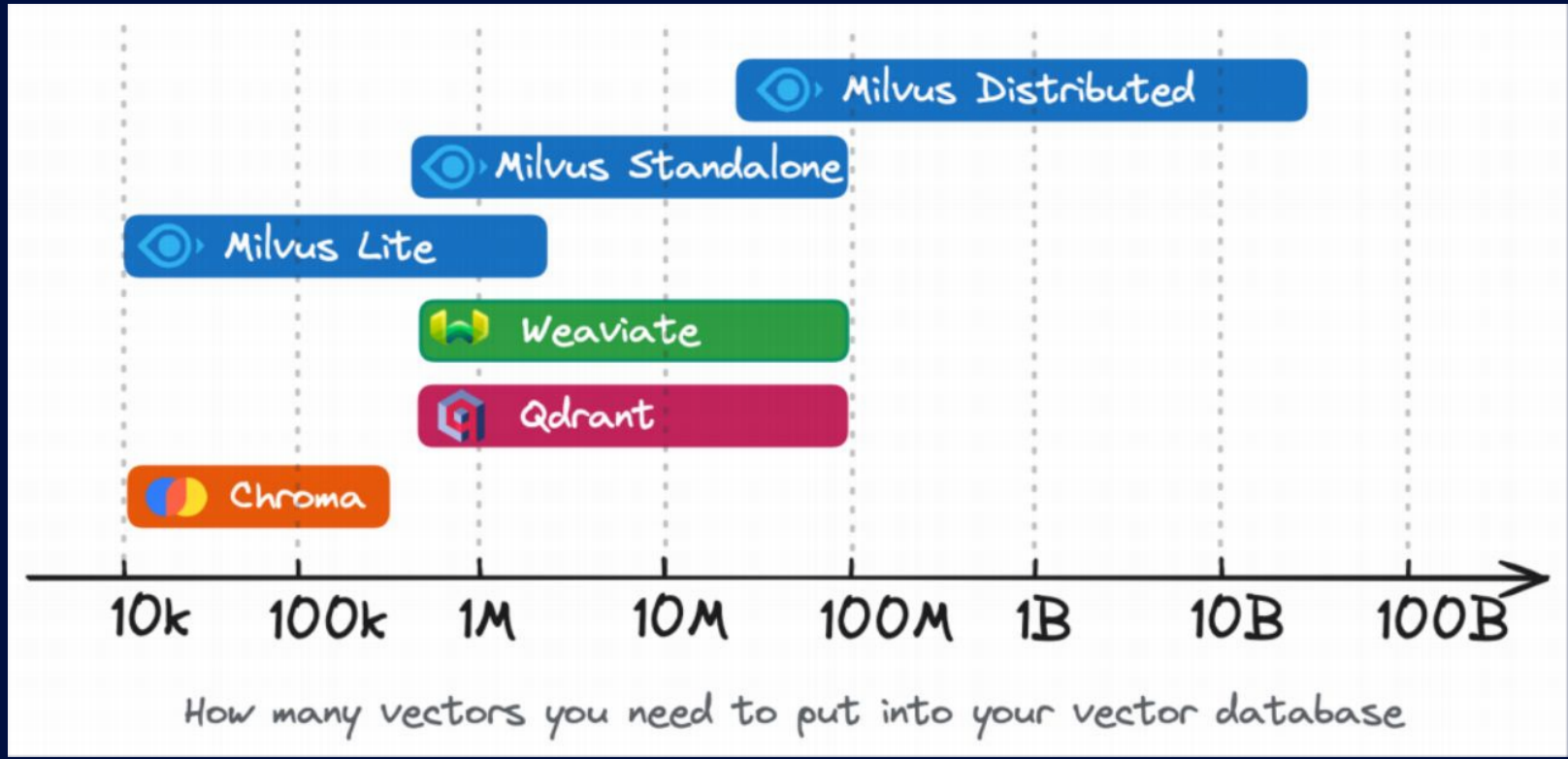


Tiered Storage

Bulk Import

GPU, Intel & ARM CPU support

Milvus vs other open source vector databases



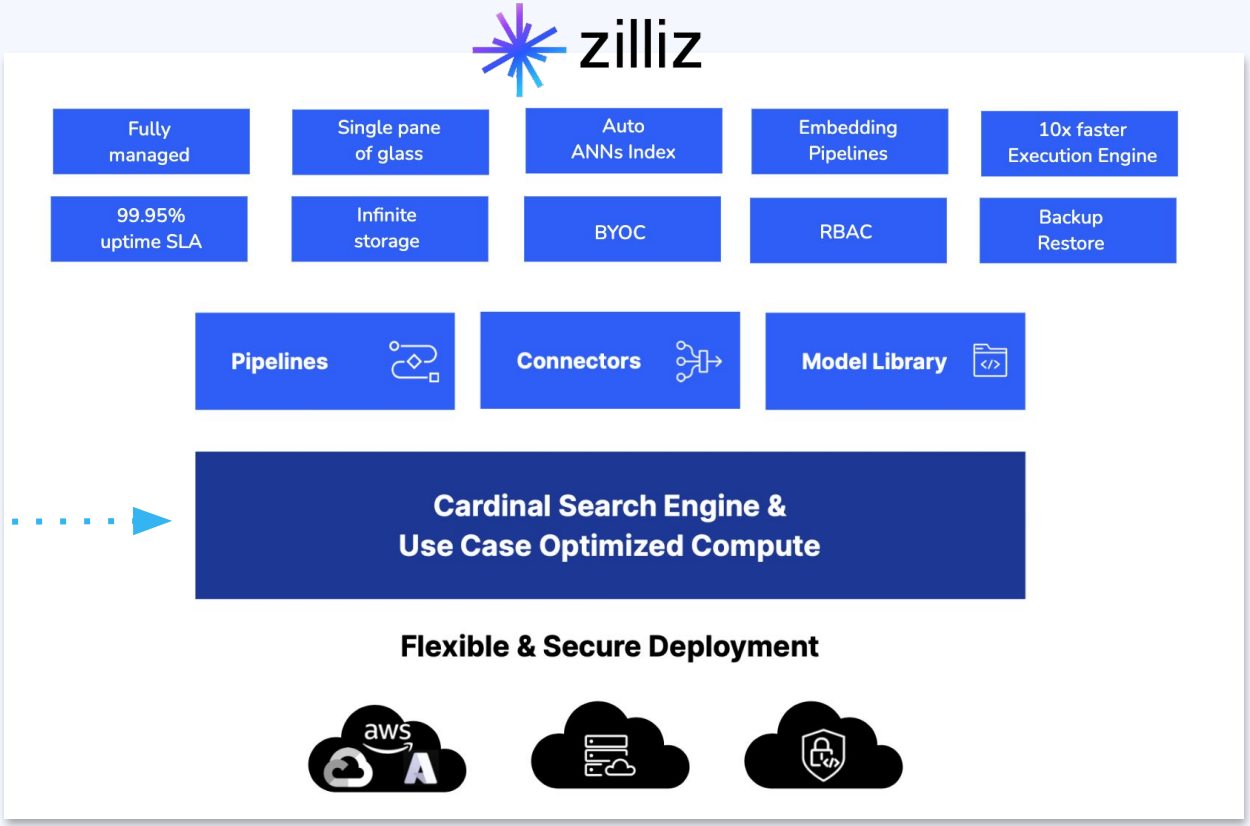
 |

zilliz

..but Zilliz Cloud is much more than OSS Milvus

Milvus

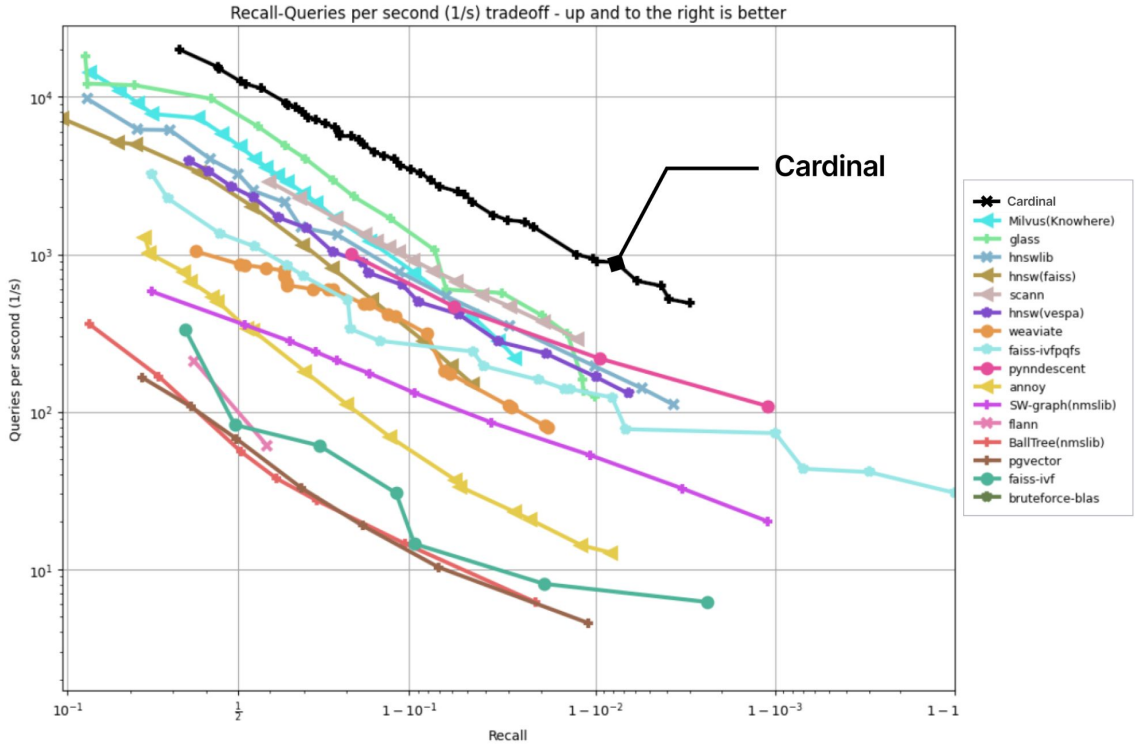
Stable Milvus versions are continuously deployed to Zilliz Cloud



 |

zilliz

ANN Benchmarks



Dataset: gist-960-euclidean

 |

zilliz

Zilliz Offerings

SELF MANAGED SOFTWARE

Milvus

Most widely-adopted open source vector database

milvus lite · docker · kubernetes

FULLY MANAGED SERVICE

Zilliz Cloud

AI Powered Search that is performant and scales

aws · Google Cloud · Azure

BRING YOUR OWN CLOUD

Zilliz BYOC

For Private VPCs

aws · Google Cloud · Azure *Coming Soon!*

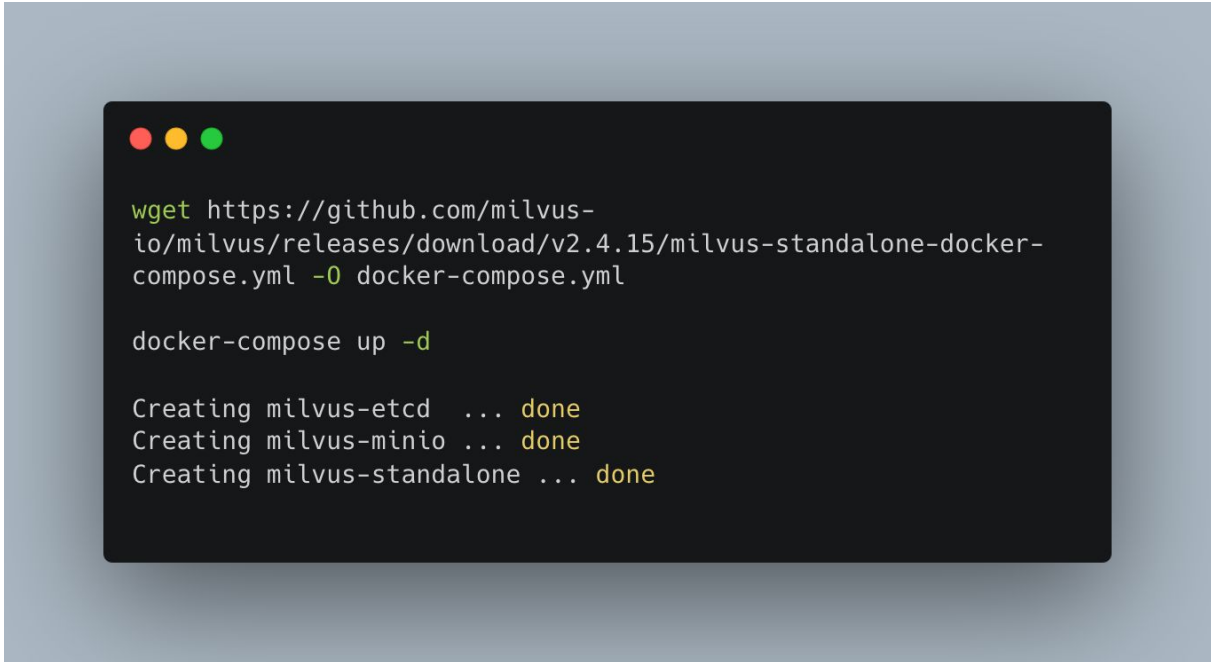
milvus

Set up Once: Common API across all products regardless of architecture

 |

zilliz

Getting started with Milvus Standalone



 |

zilliz

Integrations

The Vector Database Integration Hub

Amplify your unstructured data potential and access a variety of vector database integration options for Zilliz Cloud.

Search

All Integrations | AI Models | Client Libraries | Data Sources | Observability | AI Frameworks | Orchestration | Other

CLIENT LIBRARIES

.NET

Use the .NET SDK with Milvus or Zilliz Cloud

CLIENT LIBRARIES

Go

Use the Go SDK with Milvus or Zilliz Cloud

CLIENT LIBRARIES

JavaScript

Use the JavaScript SDK with Milvus or Zilliz Cloud

CLIENT LIBRARIES

Ruby SDK

Use the Ruby SDK with Milvus or Zilliz Cloud

CLIENT LIBRARIES

Python

Use the Python SDK to build Semantic Similarity Search with Zilliz Cloud

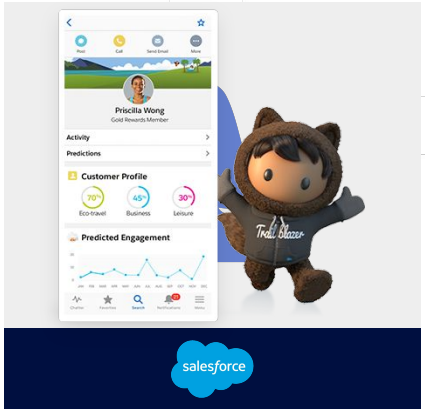
 |

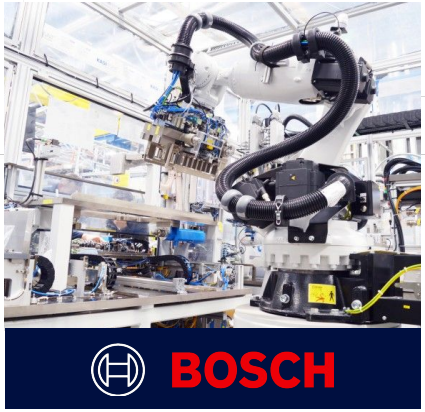
zilliz

Zilliz Use Cases

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

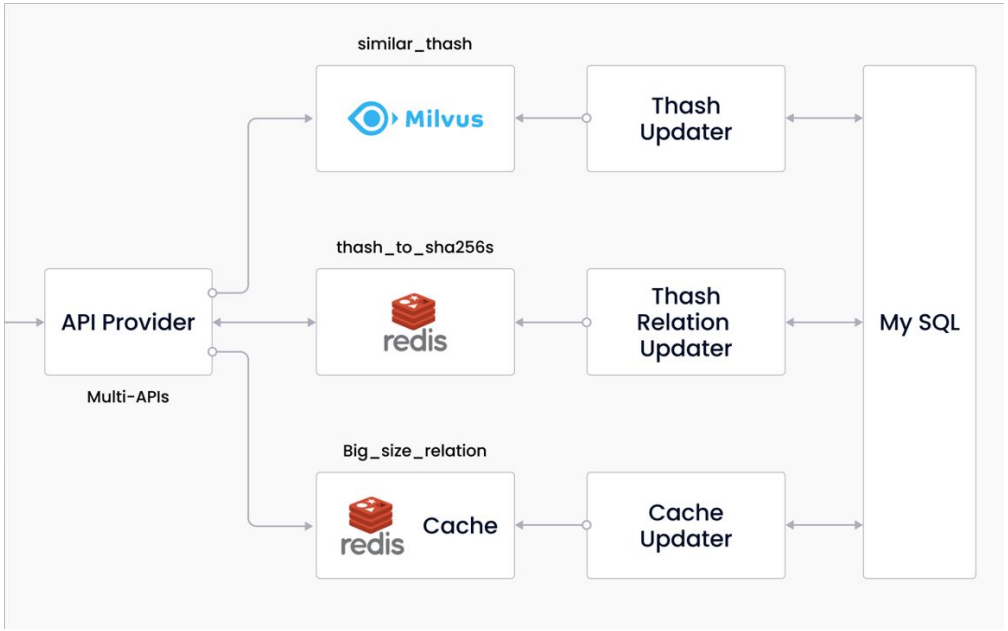




 |

zilliz

Virus scan for Android app packages



 |

e-Commerce: Similarity Search Engine

tokopedia

10x Smarter
search experience

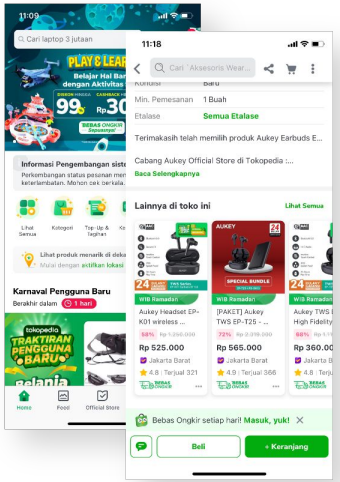
Improved
user experience

Enhanced
scalability and reliability

Moving from Elasticsearch to Milvus

Compared FAISS, Vearch and Milvus

- Milvus proved remarkably user-friendly. They found that you only need to pull its Docker image and adjust the parameters to suit your specific use cases.
- Milvus offers a broader range of supported indexes. Besides FAISS, HSNW, DISK_ANN, and ScaNN, there are 11 indexes to choose from.
- Milvus provides comprehensive documentation to aid users in their implementation.



https://zilliz.com/customers/tokopedia

 |

zilliz

e-Commerce: Similarity Search Engine

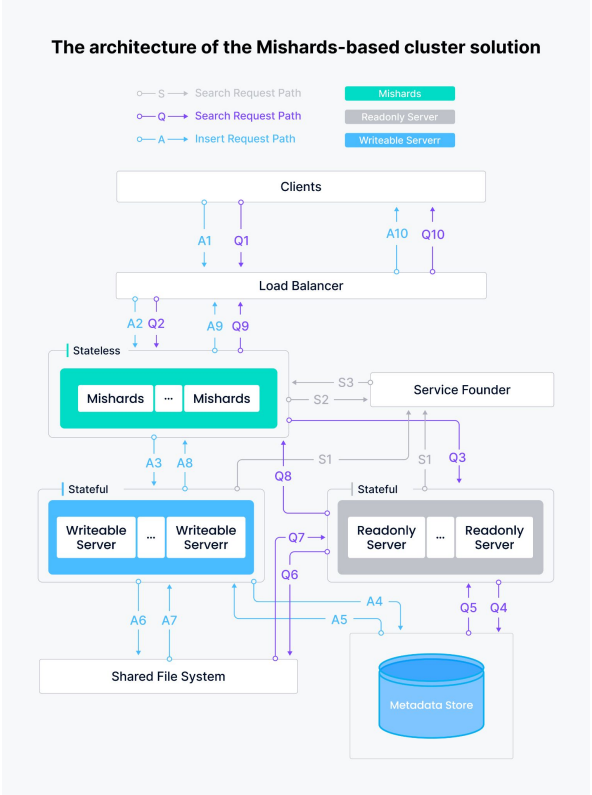
tokopedia

Ads service to match low-fill rate keywords with high-fill rate keywords

Compared Elasticsearch (Incumbent solution)

- 10x higher click-through-rate (CTR) and Conversion rate (CVR)
- Required HA implementation

Milvus offers two tools: Mishards, a cluster sharding middleware, and Milvus-Helm for streamlined configuration. At Tokopedia, they use Ansible playbooks for infrastructure setup, prompting them to create a playbook to orchestrate the infrastructure.



 |

https://zilliz.com/customers/tokopedia

zilliz

Useful resources

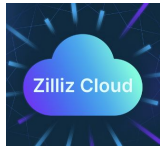
- Forrester Wave Report: Vector Databases, Q3 2024
- Interactive Vector Database comparisons
- Zilliz Integration Hub
- Milvus demo notebooks

@MILVUSAP

Discord

GitHub

Telegram
Milvus Asia Pacific



 |

zilliz

Backup slides

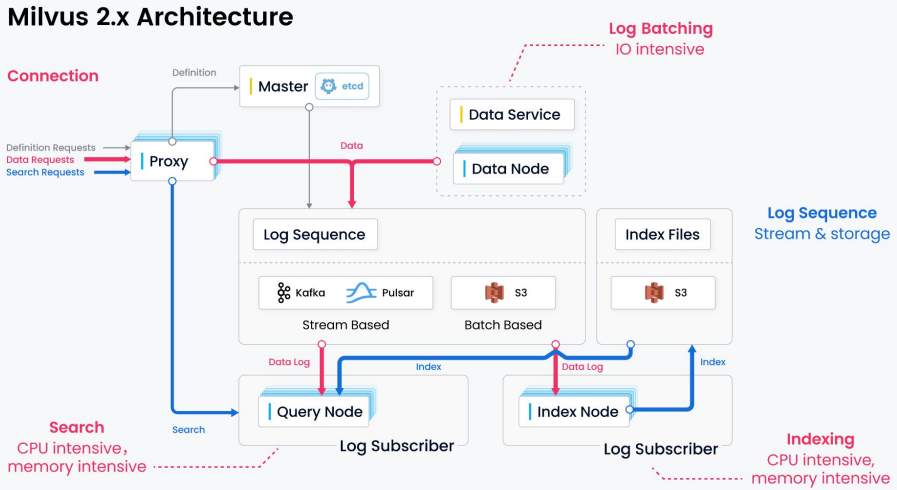
Multimedia Understanding

Shopee

> 100M
embedding vector storage and searching

Seamlessly integration
with various internal systems and tech stacks

Enhanced real-time data retrieval
with reduced latency and increased system availability



The Solution: Building a Vector Search Engine Using Milvus

After extensive research, Milvus emerged as the perfect fit. Milvus can handle billions of vectors and quickly scale out as data volume rises. Milvus' cloud-native architecture seamlessly integrated with Shopee's internal ecosystem, enabling the rapid setup of vector retrieval systems from scratch. Its feature-rich offerings, including distributed processing, GPU support, incremental updates, and scalar support, comprehensively addressed Shopee's multifaceted requirements. After careful consideration, the team selected Milvus as the foundation for their vector search engine to construct their vector search systems from scratch.

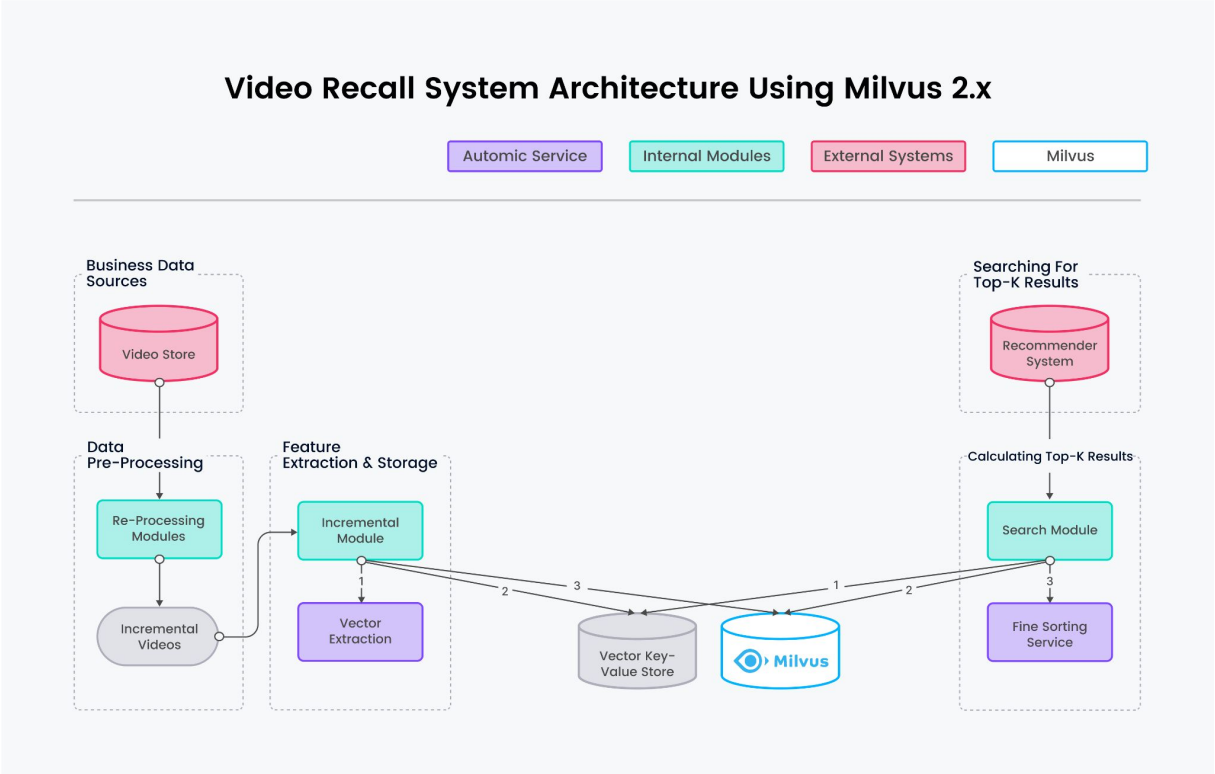
- Video Recall System
- Copyright Match System
- Video Deduplication System

https://zilliz.com/customers/shopee

zilliz

 |

Video Recall System: Improving Video Recommendation



Shopee

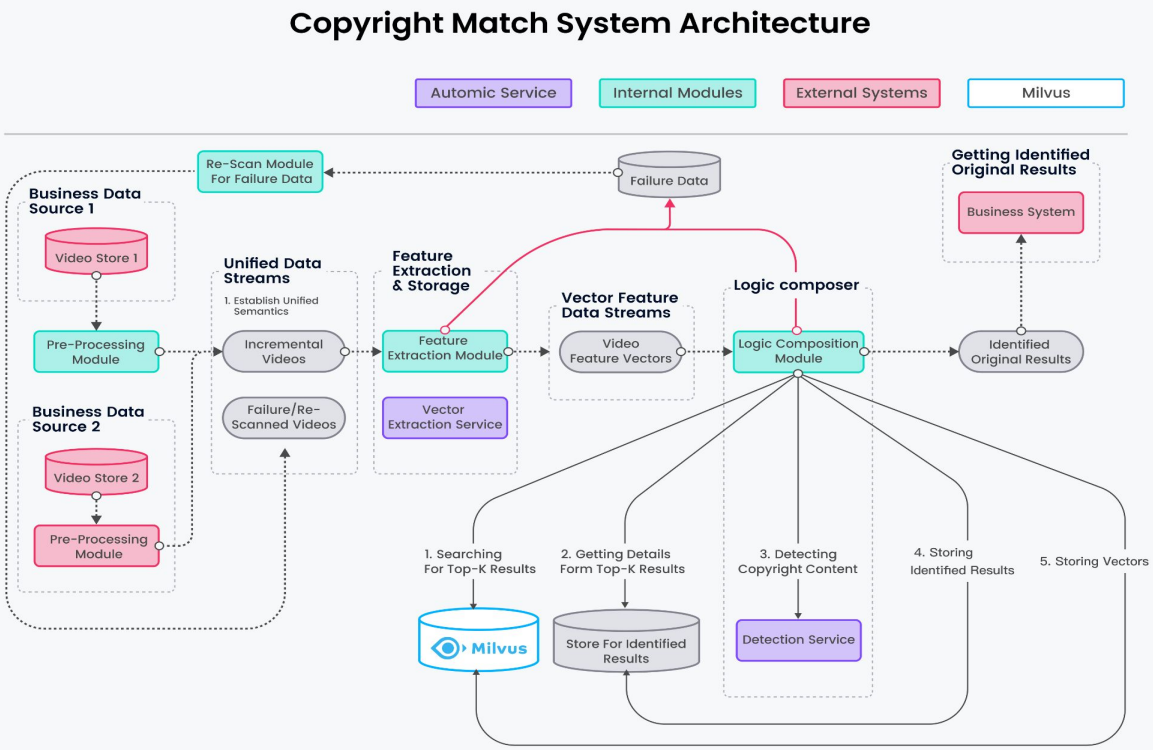
Shopee's video recall system uses Milvus as a cornerstone in the process of recommending videos. When a user searches for a video, the business requests access to Milvus to retrieve the most similar Top-K candidates. These results undergo refinement through post-ranking algorithms before being returned to the user.

https://zilliz.com/customers/shopee

 |

zilliz

Copyright Match System



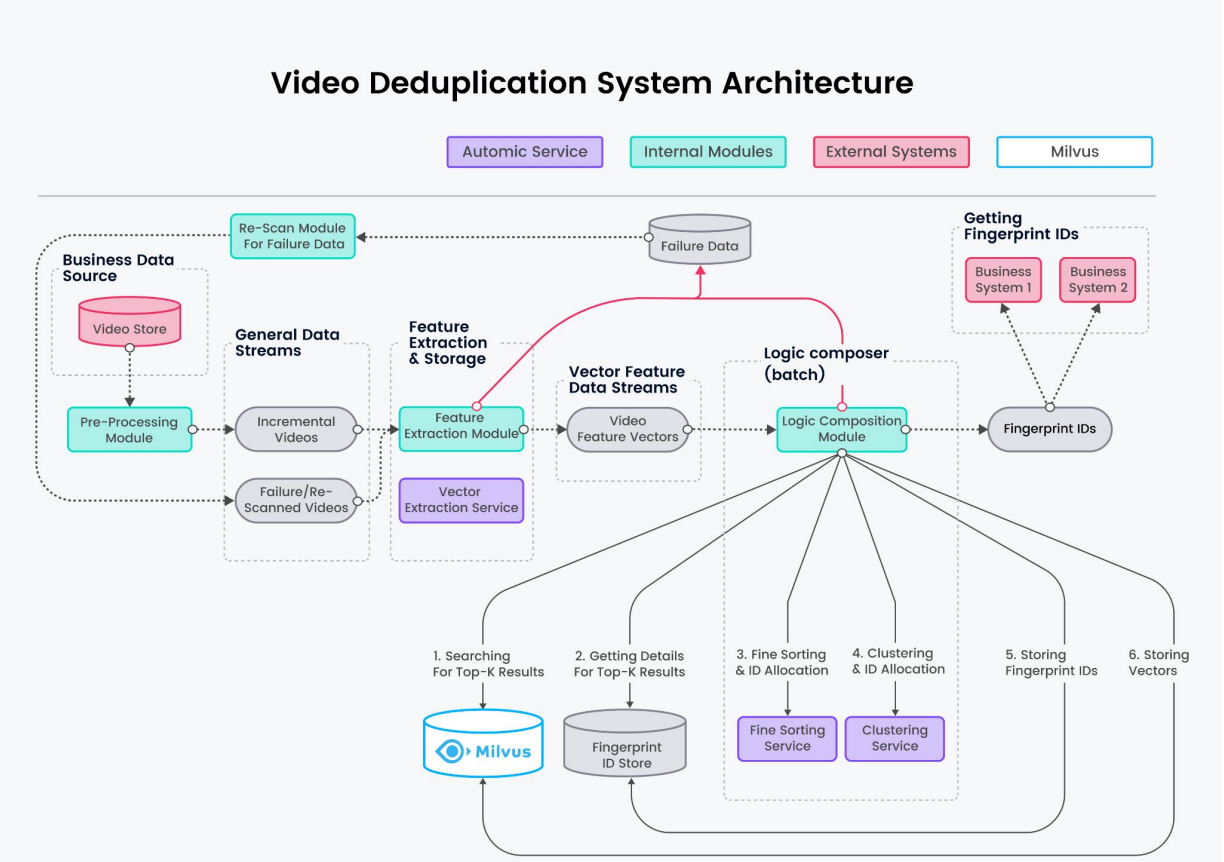
Shopee

To maintain an excellent user experience and protect the copyrights of video creators, Shopee has implemented a copyright match system using Milvus. All released video features are transformed into vectors and stored in Milvus, and every newly uploaded video is matched with those held in Milvus by using similarity searches

https://zilliz.com/customers/shopee

zilliz

Video Deduplication System



Shopee

The video deduplication system is designed to eliminate redundant content from Shopee's video platform. Like Shopee's copyright match system, the deduplication system uses Milvus to store embedding vectors transformed from video features. The system efficiently identifies and eliminates duplicate videos by searching for Top-K results in Milvus that are most similar to a specific part. Apart from the Top-K similarity search, the system involves other processing techniques such as batch data searching, post-ranking, clustering, and fingerprint assignment. In the end, Milvus stores all these results, providing valuable insights to various business units.

https://zilliz.com/customers/shopee

zilliz

Compliance and Privacy

SOC 2 Type II

Zilliz Cloud's SOC2 Type II report offers important, third-party validation of our security practices, upheld consistently throughout the reporting period. This report provides a robust, evidence-based evaluation of our commitment to maintaining the highest security standards. By eliminating discrepancies, we aim to offer you enhanced confidence in the security posture of Zilliz Cloud.

ISO/ICE 27001

The ISO/IEC 27001 certification is an international benchmark for Information Security Management Systems (ISMS). Zilliz Cloud's adherence to this standard underscores a systematic approach to managing sensitive data, aligning with global best practices. By meeting this standard, we provide a stronger assurance that your information assets are well-protected.

GDPR Readiness

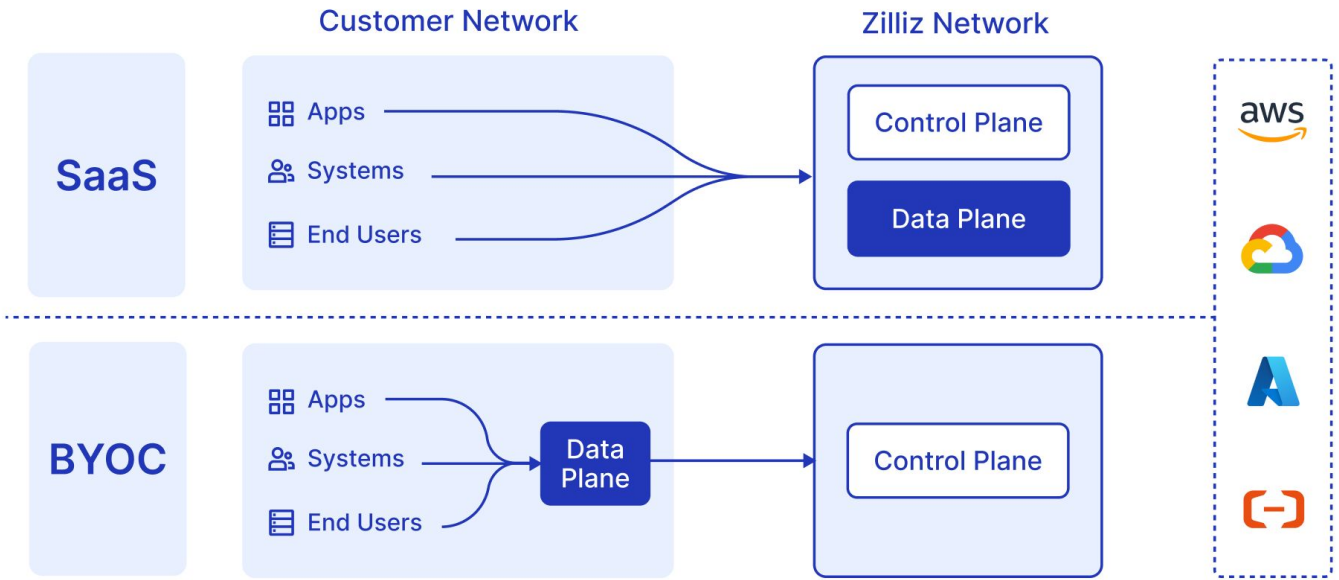
The General Data Protection Regulation (GDPR) sets standards for handling and protecting personal data from the European Economic Area (EEA) and ensures individuals' data rights. Zilliz is GDPR-ready and committed to supporting our customers' compliance efforts.

HIPAA Readiness

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) regulates the protection of the privacy and security of health information. Zilliz is HIPAA-ready, enabling covered entities and their associates to use our secure cloud database to process, maintain, and store protected health information (PHI).

zilliz

BYOC Versus SaaS



 |

zilliz

Cardinal Search Engine

AutoIndex

Algorithm

Index Builder

Index Searcher

Index Strueture

Calculation

Calculator

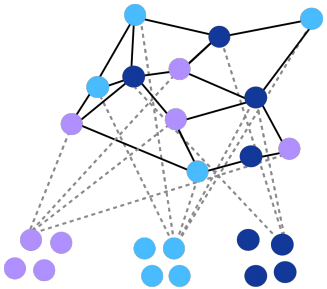
Quantizer

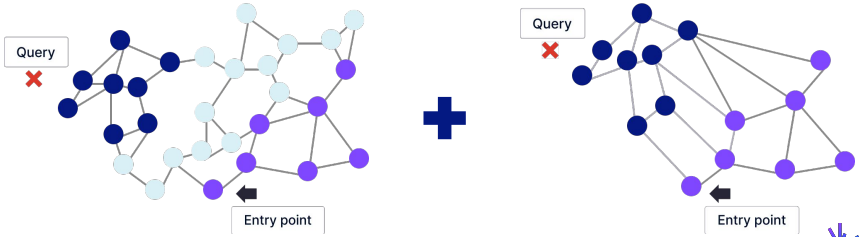
Refiner

Storage

Graph-IVF hybrid solution helps:

Configurable storage media for all modules brings flexible performance & capacity tradeoff



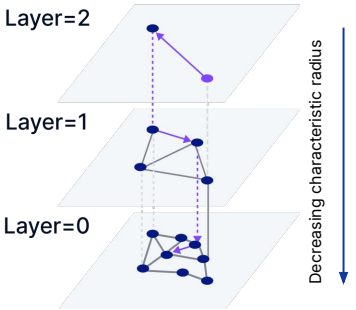


 |

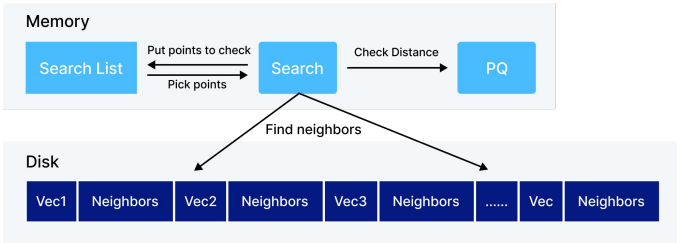
zilliz

Intelligent Tiered Storage

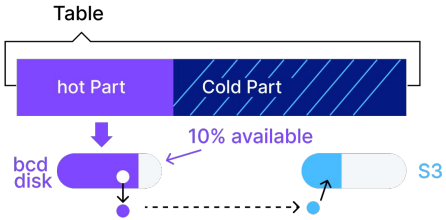
Memory Based

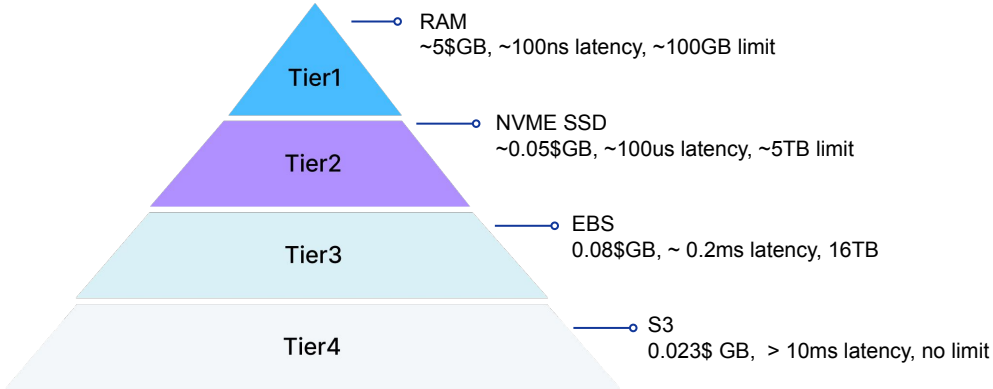


Disk Based



Smart Tiers





 |

zilliz