

**Star Milvus
for good
energy!**



The world is much more than just text and keywords

◀ **90%** of newly generated data in 2025 will be unstructured data ▶



10%
Other

Mission:

Helping organizations make sense of unstructured data.



2017
Founded



\$113M
Raised



140+
Employees



Redwood City, CA
Headquarters

Building a Research Agent to Extract Insight from Internal Data



Stefan Webb

Developer Advocate, Zilliz

stefan.webb@zilliz.com

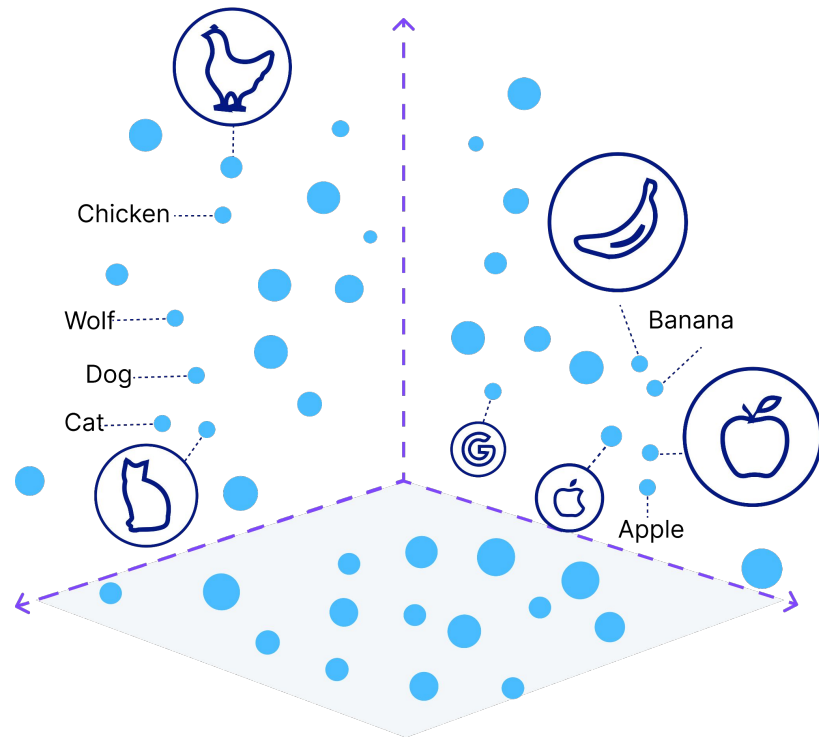
<https://www.linkedin.com/in/stefan-webb>



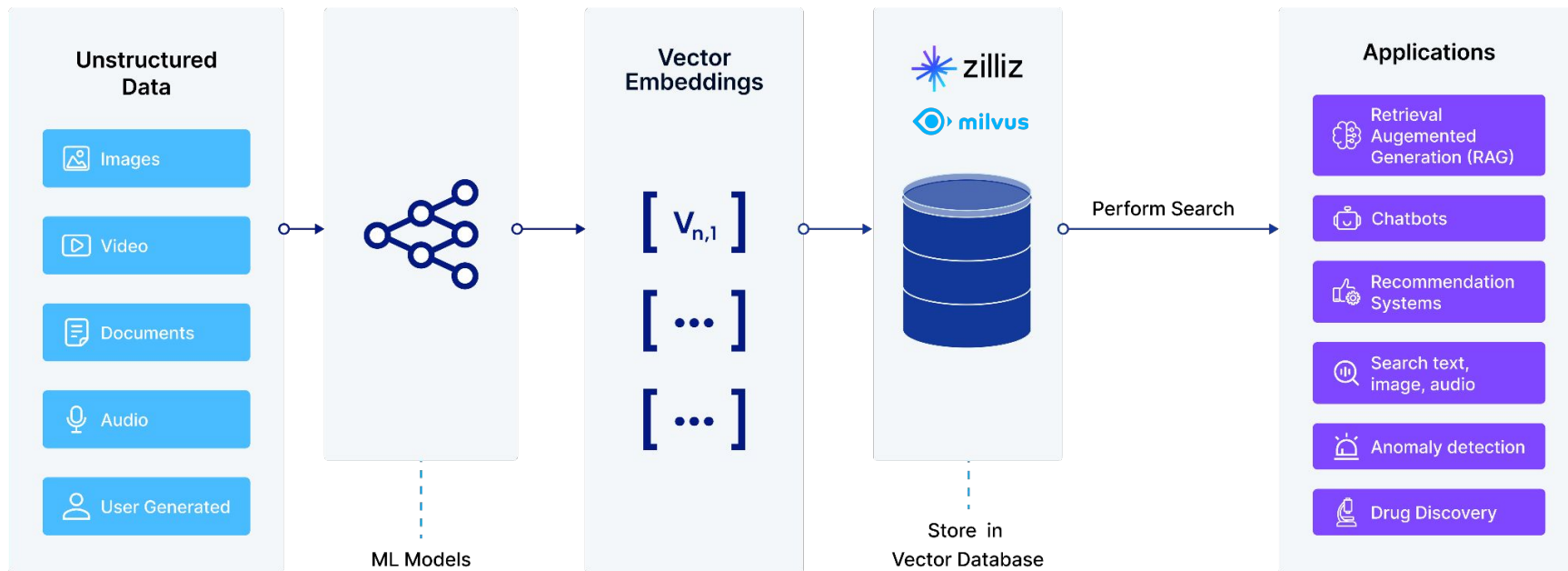
01

Why Milvus?

Vector search is the new standard

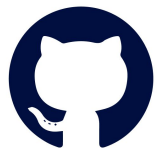


A New tool emerged. The Vector Database



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.



430+

contributors



33K+

stars



67M+

docker pulls



2.8K
+

forks

Built by database & AI experts

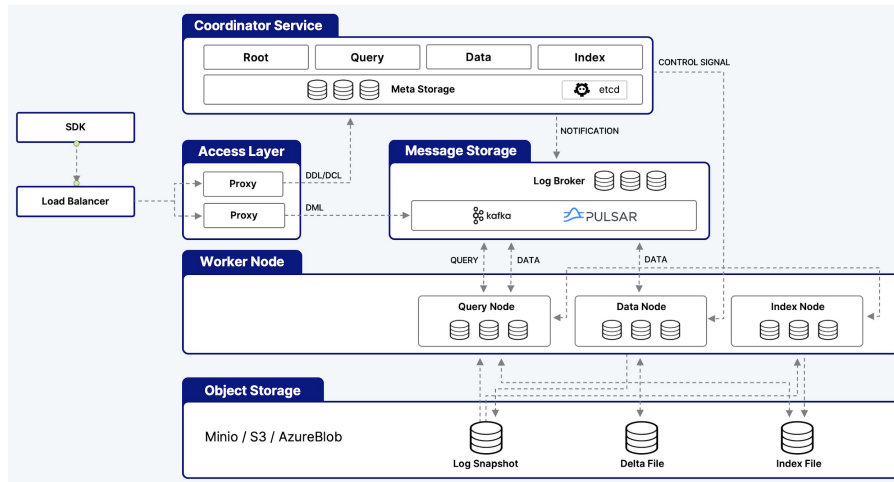
Zilliz was built by a top-tier team
of **algorithm and database
engineers** with a strong
pedigree in developing
high-performance, scalable,
and **highly available** distributed
systems, uniquely tailored for
vector search.

Milvus Architecture

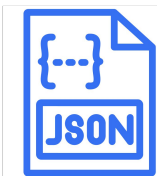
Design Principles

- 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

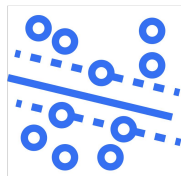
Fully distributed, designed for scalability



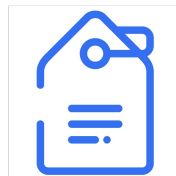
Rich functionality



Dynamic Schema



Float, Binary, &
Sparse Vector



Tag+Vector
Optimized Filtering



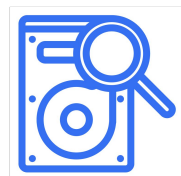
Hybrid Search
Dense & Sparse



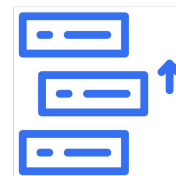
RBAC, TLS,
Encryption



Million+ level
tenant support



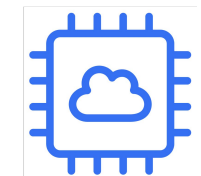
Disk Based
Index



Tiered Storage



Bulk Import



GPU, Intel & ARM
CPU support

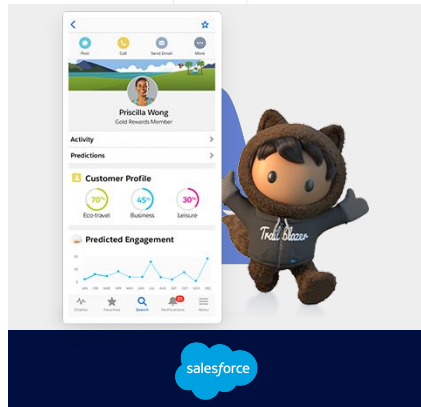
Industry leaders already use vector search in their apps

Use Case: Data Search

Vectors: 2 Billion

Req'ts: 200 ms, Cost mgmt

Index: DiskANN for cost savings



Use Case: Drug Discovery

Vectors: 12 Billion

Req'ts: High Recall

Index: BIN_FLAT

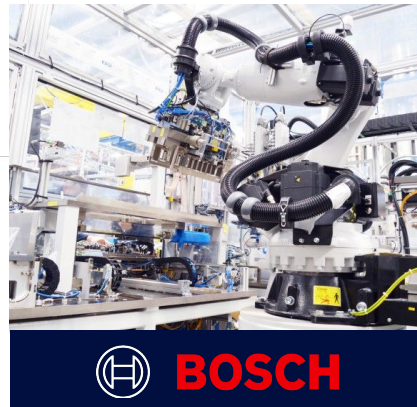


Use Case: Image Search

Vectors: 20 Billion

Req'ts: High Insertion, Cost

Index: Disk Based Index



Use Case: Recommender System

Vectors: 20 Billion

Req'ts: 5,000 QPS

Index: HNSW & CAGRA



Milvus Users

accenture

airbnb

AT&T



BOSCH

Chegg

CISCO

CISION

COMPASS

Deloitte.

ebay

FARFETCH

Grab

IKEA

Inflection

intuit.

Microsoft

new relic.

nVIDIA.

OMERS

Otter.ai

PayPal

paloalto
NETWORKS

POSHMARK ROBLOX

salesforce

Shell

shutterstock

T

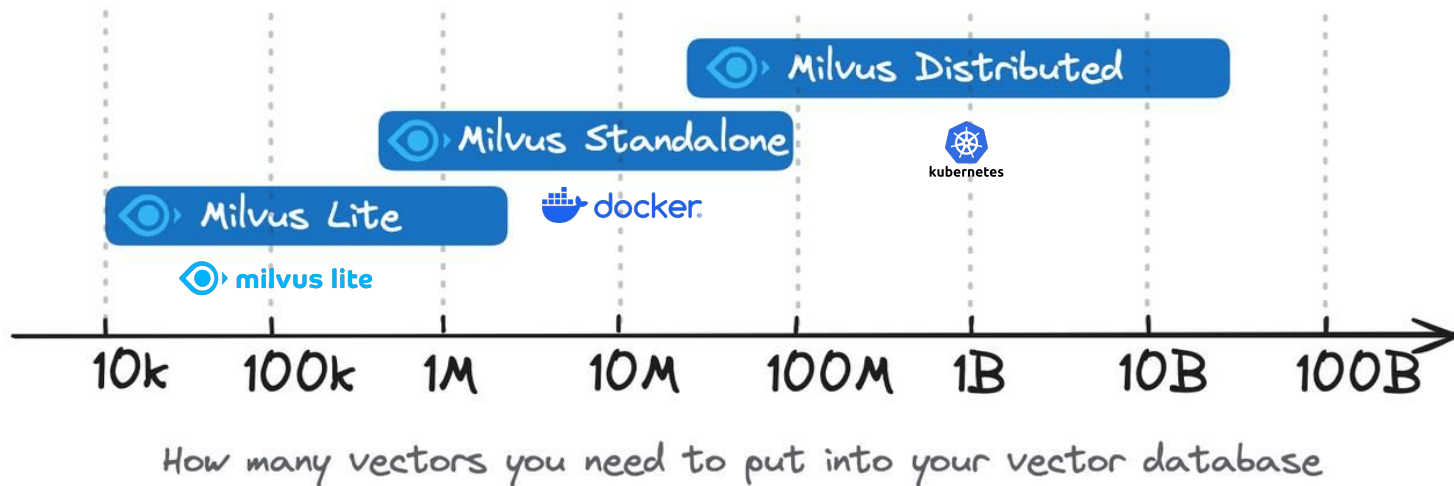
TREND
MICRO

Walmart

ZipRecruiter

zomato

Flexible Deployment Options for Milvus



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

AI Stack



Software Infrastructure

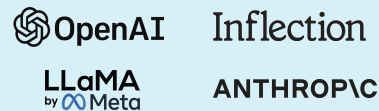
Embedding Models



Vector Database



LLMs



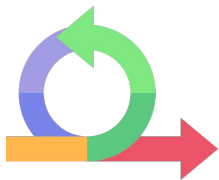
Hardware Infrastructure



02

DeepSearcher demo

Research Agents



Iteration

“...learned to plan and execute a multi-step trajectory...”

“...backtracking and reacting to real-time information...”

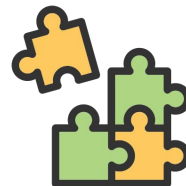
“...pivoting as needed in reaction to information it encounters...”



Search

“...trained using end-to-end reinforcement learning on hard browsing and reasoning tasks across a range of domains...”

“...optimized for web browsing and data analysis...”

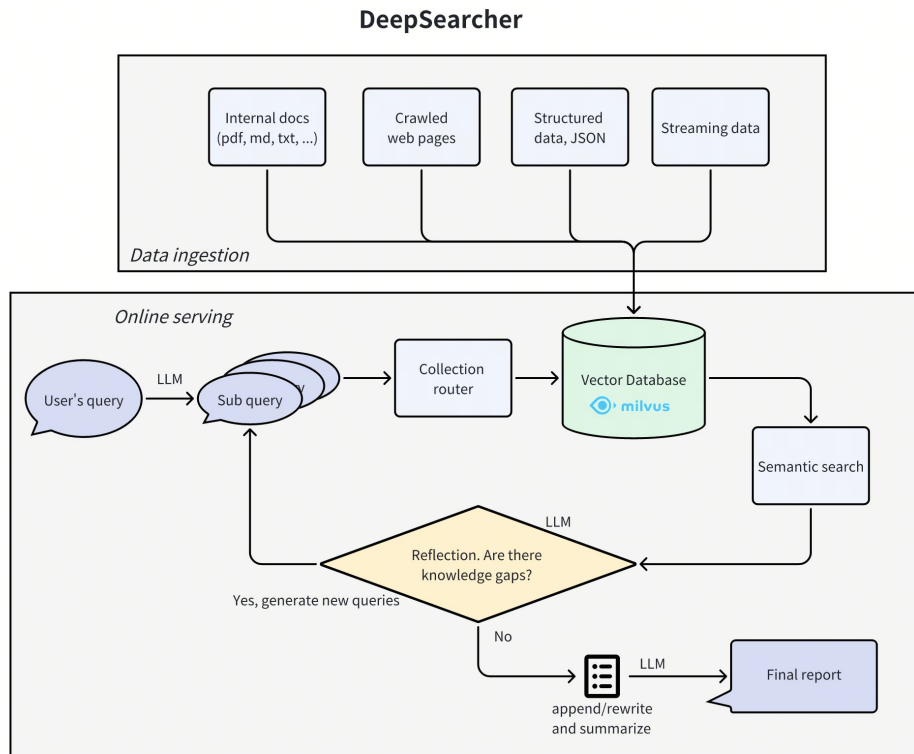


Reasoning

“...fine-tuned on the upcoming OpenAI o3 reasoning model...”

““...leverages reasoning to search, interpret, and analyze massive amounts of text...””

DeepSearcher from Zilliz



```
(py310) zilliz@zillizdeMacBook-Pro-6 deep-rag-agent % deepsearcher --query "Write a report comparing Milvus with other vector databases"
```

```
(py310) zilliz@zillizdeMacBook-Pro-6 deep-rag-agent %deepsearcher --query "Write a report comparing Milvus with other vector databases"
<query> Write a report comparing Milvus with other vector databases </query>
```

our query/task


```
(py310) zilliz@zillizdeMacBook-Pro-6 deep-rag-agent %deepsearcher —query "Write a report comparing Milvus with other vector databases"
```

```
<query> Write a report comparing Milvus with other vector databases </query>
```

```
<think> Break down the original query into new sub queries: ['What is Milvus?', 'What are the key features of Milvus?', 'What are some other popular vector databases?', 'How does Milvus compare with these other vector databases in terms of performance?', 'What are the use cases for Milvus and other vector databases?', 'What are the advantages and disadvantages of using Milvus compared to other vector databases?', 'How is the community support and documentation for Milvus versus other vector databases?', 'What are the scalability and deployment options for Milvus compared to other vector databases?']</think>
```

```
>> Iteration: 1
```

```
|
```

initial subqueries

```
(py310) zilliz@zillizdeMacBook-Pro-6 deep-rag-agent %deepsearcher —query "Write a report comparing Milvus with other vector databases"
<query> Write a report comparing Milvus with other vector databases </query>

<think> Break down the original query into new sub queries: ['What is Milvus?', 'What are the key features of Milvus?', 'What are some other popular vector databases?', 'How does Milvus compare with these other vector databases in terms of performance?', 'What are the use cases for Milvus and other vector databases?', 'What are the advantages and disadvantages of using Milvus compared to other vector databases?', 'How is the community support and documentation for Milvus versus other vector databases?', 'What are the scalability and deployment options for Milvus compared to other vector databases?']</think>

>> Iteration: 1

<think> Perform search [What is Milvus?] on the vector DB collections: ['milvus_docs', 'what_is_milvus'] </think>
>

<search> Search [What is Milvus?] in [milvus_docs]... </search>
```

**think what action to perform
search 1st subquery, 1st source**

```
(py310) zilliz@zillizdeMacBook-Pro-6 deep-rag-agent %deepsearcher —query "Write a report comparing Milvus with other vector databases"
<query> Write a report comparing Milvus with other vector databases </query>

<think> Break down the original query into new sub queries: ['What is Milvus?', 'What are the key features of Milvus?', 'What are some other popular vector databases?', 'How does Milvus compare with these other vector databases in terms of performance?', 'What are the use cases for Milvus and other vector databases?', 'What are the advantages and disadvantages of using Milvus compared to other vector databases?', 'How is the community support and documentation for Milvus versus other vector databases?', 'What are the scalability and deployment options for Milvus compared to other vector databases?']</think>

>> Iteration: 1

<think> Perform search [What is Milvus?] on the vector DB collections: ['milvus_docs', 'what_is_milvus'] </think>
>

<search> Search [What is Milvus?] in [milvus_docs]... </search>

<search> Accept 5 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/integrations/apify_milvus_rag.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/roadmap.md'] </search>

<search> Search [What is Milvus?] in [what_is_milvus]... </search>
```

retrieve chunks with Milvus

```
(py310) zilliz@zillizdeMacBook-Pro-6 deep-rag-agent %deepsearcher —query "Write a report comparing Milvus with
other vector databases"
<query> Write a report comparing Milvus with other vector databases </query>

<think> Break down the original query into new sub queries: ['What is Milvus?', 'What are the key features of Mi
lvus?', 'What are some other popular vector databases?', 'How does Milvus compare with these other vector databa
ses in terms of performance?', 'What are the use cases for Milvus and other vector databases?', 'What are the ad
vantages and disadvantages of using Milvus compared to other vector databases?', 'How is the community support a
nd documentation for Milvus versus other vector databases?', 'What are the scalability and deployment options fo
r Milvus compared to other vector databases?']</think>

>> Iteration: 1

<think> Perform search [What is Milvus?] on the vector DB collection: ['milvus-docs', 'what_is_milvus'] </think>
>

<search> Search [What is Milvus?] in milvus-docs... </search>

<search> Accept 5 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/o
verview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/integrations/apify_milvus_rag.md', '/Users/zilli
z/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.
x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/roadmap.md'] </search>

<search> Search [What is Milvus?] in [what_is_milvus]... </search>

<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatisMilvus.pdf', './examples/data/Whati
sMilvus.pdf', './examples/data/WhatisMilvus.pdf'] </search>
```

search 1st subquery, 2nd source
retrieve chunks with Milvus


```
<think> Break down the original query into new sub queries: ['What is Milvus?', 'What are the key features of Milvus?', 'What are some other popular vector databases?', 'How does Milvus compare with these other vector databases in terms of performance?', 'What are the use cases for Milvus and other vector databases?', 'What are the advantages and disadvantages of using Milvus compared to other vector databases?', 'How is the community support and documentation for Milvus versus other vector databases?', 'What are the scalability and deployment options for Milvus compared to other vector databases?']</think>
```

```
>> Iteration: 1
```

```
<think> Perform search [What is Milvus?] on the vector DB collections: ['milvus_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What is Milvus?] in [milvus_docs]... </search>
```

```
<search> Accept 5 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/roadmap.md'] </search>
```

```
<search> Search [What is Milvus?] in [what_is_milvus]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>
```

```
<think> Perform search [What are the key features of Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What are the key features of Milvus?] in [milvus_docs]... </search>
```

think what action to perform
search 2nd subquery, 1st source

```
>> Iteration: 1
```

```
<think> Perform search [What is Milvus?] on the vector DB collections: ['milvus_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What is Milvus?] in [milvus_docs]... </search>
```

```
<search> Accept 5 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/integrations/apify_milvus_rag.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/roadmap.md'] </search>
```

```
<search> Search [What is Milvus?] in [what_is_milvus]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>
```

```
<think> Perform search [What are the key features of Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

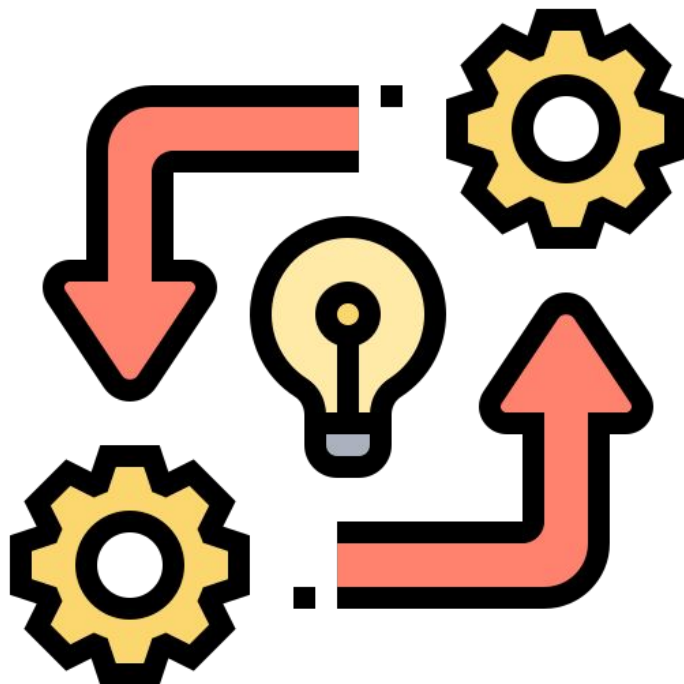
```
<search> Search [What are the key features of Milvus?] in [milvus_docs]... </search>
```

```
<search> Accept 4 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/architecture/architecture.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/integrations/apify_milvus_rag.md'] </search>
```

```
<search> Search [What are the key features of Milvus?] in [milvus_embedding_docs]... </search>
```

Retrieve chunks with Milvus





```
<search> Search [What are the scalability and deployment options for Milvus compared to other vector databases?] in [what_is_milvus]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>
```

```
<search> Search [What are the scalability and deployment options for Milvus compared to other vector databases?] in [milvus_docs]... </search>
```

```
<search> Accept 5 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/integrations/apify_milvus_rag.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/architecture/architecture_overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md'] </search>
```

```
<think> Reflecting on the search results... </think>
```

```
<think> New search queries for next iteration: ['What are the performance benchmarks of Milvus compared to its competitors?', 'What specific use cases are best suited for other vector databases compared to Milvus?', 'What are the recent developments and future directions in vector databases beyond Milvus?'] </think>
```

```
>> Iteration: 2
```

```
<think> Perform search [What are the performance benchmarks of Milvus compared to its competitors?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_docs]... </search>
```

Accelerated Playback

reflect on iter 1, new subqueries


```
<search> Search [What are the scalability and deployment options for Milvus compared to other vector databases?] in [milvus_docs]... </search>
```

```
<search> Accept 5 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/integrations/apify_milvus_rag.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/accelerated-playback.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/architecture/architecture_overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md'] </search>
```

```
<think> Reflecting on the search results... </think>
```

```
<think> New search queries for next iteration: ['What are the performance benchmarks of Milvus compared to its competitors?', 'What specific use cases are best suited for other vector databases compared to Milvus?', 'What are the recent developments and future directions in vector databases beyond Milvus?'] </think>
```

```
>> Iteration: 2
```

```
<think> Perform search [What are the performance benchmarks of Milvus compared to its competitors?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_docs]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/comparison.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/benchmark.md'] </search>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_embedding_docs]... </search>
```

Accelerated Playback

search 1st subquery



```
<search> Accept 5 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/integrations/apify_milvus_rag.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/architecture/architecture_overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md'] </search>
```

Accelerated Playback

```
<think> Reflecting on the search results... </think>
```

```
<think> New search queries for next iteration: ['What are the performance benchmarks of Milvus compared to its competitors?', 'What specific use cases are best suited for other vector databases compared to Milvus?', 'What are the recent developments and future directions in vector databases beyond Milvus?'] </think>
```

```
>> Iteration: 2
```

```
<think> Perform search [What are the performance benchmarks of Milvus compared to its competitors?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_docs]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/comparison.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/benchmark.md'] </search>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_embedding_docs]... </search>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [what_is_milvus]... </search>
```

```
x/site/en/reference/architecture/architecture_overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/getstarted/install-overview.md'] </search>
```

```
<think> Reflecting on the search results... </think>
```

Accelerated Playback

```
<think> New search queries for next iteration: ['What are the performance benchmarks of Milvus compared to its competitors?', 'What specific use cases are best suited for other vector databases compared to Milvus?', 'What are the recent developments and future directions in vector databases beyond Milvus?'] </think>
```

```
>> Iteration: 2
```

```
<think> Perform search [What are the performance benchmarks of Milvus compared to its competitors?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_docs]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/comparison.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/benchmark.md'] </search>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_embedding_docs]... </search>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [what_is_milvus]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>
```



```
ompetitors?', 'What specific use cases are best suited for other vector databases compared to Milvus?', 'What are the recent developments and future directions in vector databases beyond Milvus?'] </think>
```

```
>> Iteration: 2
```

```
<think> Perform search [What are the performance benchmarks of Milvus compared to its competitors?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

Accelerated Playback

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_docs]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/comparison.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/benchmark.md'] </search>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_embedding_docs]... </search>
```

search 2nd subquery

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [what_is_milvus]... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>
```

```
<think> Perform search [What specific use cases are best suited for other vector databases compared to Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_docs]... </search>
```

```
r DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_docs]...  
</search>
```

Accelerated Playback

```
<search> Accept 3 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/c  
omparison.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/m  
ilvus-docs-2.5.x/site/en/reference/benchmark.md'] </search>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_embeddin  
g_docs]... </search>
```

```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [what_is_milvus]  
... </search>
```

```
<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/Whati  
sMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>
```

```
<think> Perform search [What specific use cases are best suited for other vector databases compared to Milvus?]   
on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [mil  
vus_docs]... </search>
```

```
<search> Accept 1 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/c  
omparison.md'] </search>
```

```
<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [mil  
vus_embedding_docs]... </search>
```

</search>

<search> Accept 3 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/comparison.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/overview.md', '/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/reference/benchmark.md'] </search>

Accelerated Playback

<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_embedding_docs]... </search>

<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [what_is_milvus]... </search>

<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>

<think> Perform search [What specific use cases are best suited for other vector databases compared to Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_docs]... </search>

<search> Accept 1 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/comparison.md'] </search>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_embedding_docs]... </search>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [what_is_milvus]... </search>


```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [milvus_embedding_docs]... </search>

<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [what_is_milvus]
... </search>

<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>

<think> Perform search [What specific use cases are best suited for other vector databases compared to Milvus?]
on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_docs]... </search>

<search> Accept 1 document chunk(s) from references: ['./Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/comparison.md'] </search>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_embedding_docs]... </search>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [what_is_milvus]... </search>

<think> Perform search [What are the recent developments and future directions in vector databases beyond Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>

<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [milvus_docs]... </search>
```

Accelerated Playback

search 3rd subquery



```
<search> Search [What are the performance benchmarks of Milvus compared to its competitors?] in [what_is_milvus] ... </search>

<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>

<think> Perform search [What specific use cases are best suited for other vector databases compared to Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_docs]... </search>

<search> Accept 1 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/compare.md'] </search>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_embedding_docs]... </search>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [what_is_milvus]... </search>

<think> Perform search [What are the recent developments and future directions in vector databases beyond Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>

<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [milvus_docs]... </search>

<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [milvus_embedding_docs]... </search>
```

Accelerated Playback


```
<search> Accept 3 document chunk(s) from references: ['./examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf', './examples/data/WhatIsMilvus.pdf'] </search>

<think> Perform search [What specific use cases are best suited for other vector databases compared to Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_docs]... </search>

<search> Accept 1 document chunk(s) from references: ['/Users/zilliz/Downloads/milvus-docs-2.5.x/site/en/about/comparison.md'] </search>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_embedding_docs]... </search>

<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [what_is_milvus]... </search>

<think> Perform search [What are the recent developments and future directions in vector databases beyond Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>

<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [milvus_docs]... </search>

<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [milvus_embedding_docs]... </search>

<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [what_is_milvus]... </search>
```

Accelerated Playback

```
omparison.md'] </search>
```

```
<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [milvus_embedding_docs]... </search>
```

Accelerated Playback

```
<search> Search [What specific use cases are best suited for other vector databases compared to Milvus?] in [what_is_milvus]... </search>
```

```
<think> Perform search [What are the recent developments and future directions in vector databases beyond Milvus?] on the vector DB collections: ['milvus_docs', 'milvus_embedding_docs', 'what_is_milvus'] </think>
```

```
<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [milvus_docs]... </search>
```

```
<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [milvus_embedding_docs]... </search>
```

```
<search> Search [What are the recent developments and future directions in vector databases beyond Milvus?] in [what_is_milvus]... </search>
```

```
<search> Accept 1 document chunk($1) from references: ['./examples/data/WhatIsMilvus.pdf'] </search>
```

```
<think> Reflecting on the search results... </think>
```

```
<think> No new search queries were generated. Exiting. </think>
```

```
<think> Generating final answer... </think>
```

```
<think> Summarize answer from all 21 retrieved chunks... </think>
```

reflect on iter 2,
no new subqueries

```
<think> No new search queries were generated. Exiting. </think>
```

```
<think> Generating final answer... </think>
```

```
<think> Summarize answer from all 21 retrieved chunks... </think>
```

← generate report

```
=== FINAL ANSWER===
```

```
### Report: Comparison of Milvus with Other Vector Databases
```

```
**Introduction to Milvus:**
```

Milvus is a high-performance, scalable vector database designed for managing unstructured data like images, text, and audio by converting them into numerical vectors. It is open-source and available as a cloud service, supporting various deployment options: Milvus Lite, Standalone, and Distributed. These variants cater to different scales, from local testing environments to large Kubernetes deployments capable of handling tens of billions of vectors.

```
**Key Features of Milvus:**
```

1. ****High Performance****: Milvus is optimized for speed and efficiency. It employs advanced data access patterns and hardware-aware optimizations, achieving a performance boost of 30%-70% over popular implementations like FAISS and HNSWLib.
2. ****Flexible Deployment****: It supports various deployment options—Milvus Lite for lightweight applications, Standalone for single-host setups, and Distributed architecture for high-scale scenarios.
3. ****Robust Indexing Algorithms****: Milvus supports diverse indexing/search algorithms (IVF, HNSW, DiskANN) enhancing query speed through real-time processing.
4. ****Advanced Querying Capabilities****: The database supports multiple types of searches, including ANN, hybrid, and filtered searches, alongside maintaining support for advanced data types like sparse and binary vectors.

```
**Comparative Analysis:**
```


2. ****Flexible Deployment****: It supports various deployment options—Milvus Lite for lightweight applications, Standalone for single-host setups, and Distributed architecture for high-scale scenarios.
3. ****Robust Indexing Algorithms****: Milvus supports diverse indexing/search algorithms (IVF, HNSW, DiskANN) enhancing query speed through real-time processing.
4. ****Advanced Querying Capabilities****: The database supports multiple types of searches, including ANN, hybrid, and filtered searches, alongside maintaining support for advanced data types like sparse and binary vectors.

****Comparative Analysis****

1. ****Performance****: Milvus often outperforms its competitors (like Pinecone and others) by a factor of 2-5x due to its efficient engine implemented in C++, optimized for low-level resource management and multi-threading.
2. ****Scalability****: With its cloud-native architecture, Milvus scales linearly to handle significant loads and supports concurrent queries efficiently. It managed billion-scale vectors with minimal performance degradation during extensive testing.
3. ****Community Support and Documentation****: Milvus has robust community engagement and extensive documentation compared to many other vector databases. Its integrations and user contributions, especially from large corporations, enhance its reliability and development support.
4. ****Use Cases****: Milvus excels in applications needing intense similarity searches, such as image retrieval, AI-driven applications, and machine learning frameworks. However, some scenarios may better suit specialized databases like Pinecone for specific niche applications.

****Advantages and Disadvantages****

- **Advantages**

- Superior performance and speed.
- Flexible regarding deployment options.
- Comprehensive SDK support and community-driven documentation.

- **Disadvantages**

- Complexity in setup for new users.
- Requires careful optimization for specific workloads.

****Recent Developments and Future Directions****

Milvus is continuously evolving with planned updates that focus on enhancing its speed, adopting more advanced

****Comparative Analysis:****

1. ****Performance****: Milvus often outperforms its competitors (like Pinecone and others) by a factor of 2-5x due to its efficient engine implemented in C++, optimized for low-level resource management and multi-threading.
2. ****Scalability****: With its cloud-native architecture, Milvus scales linearly to handle significant loads and supports concurrent queries efficiently. It managed billion-scale vectors with minimal performance degradation during extensive testing.
3. ****Community Support and Documentation****: Milvus has robust community engagement and extensive documentation compared to many other vector databases. Its integrations and user contributions, especially from large corporations, enhance its reliability and development support.
4. ****Use Cases****: Milvus excels in applications needing intense similarity searches, such as image retrieval, AI-driven applications, and machine learning frameworks. However, some scenarios may better suit specialized databases like Pinecone for specific niche applications.

****Advantages and Disadvantages:****

- ****Advantages****:
 - Superior performance and speed.
 - Flexible regarding deployment options.
 - Comprehensive SDK support and community-driven documentation.
- ****Disadvantages****:
 - Complexity in setup for new users.
 - Requires careful optimization for specific workloads.

****Recent Developments and Future Directions:****

Milvus is continuously evolving with planned updates that focus on enhancing its speed, adopting more advanced indexing mechanisms, and exploring serverless architectures for even greater scalability and cost-effectiveness.

****Conclusion:****

Milvus is a compelling choice among vector databases, noteworthy for its high performance, flexible deployment options, and strong community support. While it shines in many applications, specific use cases might warrant looking into tailored alternatives depending on organizational needs and operational scale.

(py310) zilliz@zillizdeMacBook-Pro-6 deep-rag-agent %

03

Let's build it ourselves!



It's Coding Time!

<https://bit.ly/41HjZbJ>



04

Where to from here?

Where to from here?



Data

- Open-source
- High-quality
- Reasoning traces



Cost

- Specialized hardware
- Algorithmic improvements



Scaling

- Per-se

Zilliz Offerings

<https://cloud.zilliz.com/signup>

SELF MANAGED SOFTWARE



Milvus

Most widely-adopted open source vector database



FULLY MANAGED SERVICE



Zilliz Cloud

AI Powered Search that is performant and scales



Google Cloud



Azure

BRING YOUR OWN CLOUD



Zilliz BYOC

For Private VPCs



Google Cloud



Azure
Coming Soon!

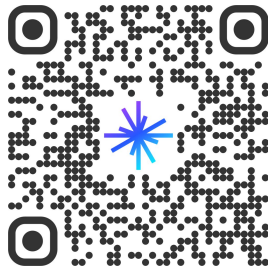


Set up Once: Common API across all products regardless of architecture

LET'S STAY CONNECTED!

Stefan Webb

Developer Advocate, Zilliz



<https://milvus.io/discord>

Book a Free 1:1 Session

For Support Productionizing Milvus



<http://bit.ly/43YkEYW>