zilliz

Beyond RAG: Vector Databases

Yujian Tang | Zilliz

Speaker

Yujian Tang

Senior Developer Advocate, Zilliz

yujian@zilliz.com
https://www.linkedin.com/in/yujiantang
https://www.twitter.com/yujian_tang

CONTENTS

01

Why Vector Databases?

Compare data that you couldn’t compare before

Unstructured Data is Everywhere

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

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

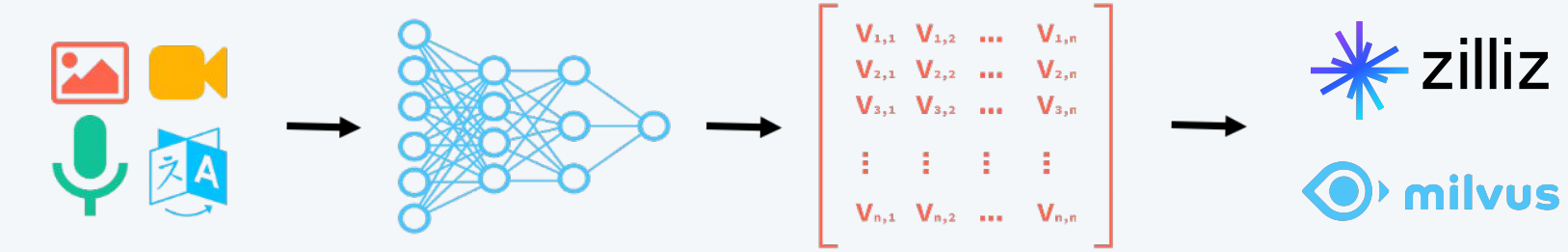
Text

Images

Video

and more!

Unstructured Data + ML = Vector Magic



Find Semantically Similar Data

Apple made profits of $97 Billion in 2023

I like to eat apple pie for profit in 2023

Apple’s bottom line increased by record numbers in 2023

But wait! There’s more!

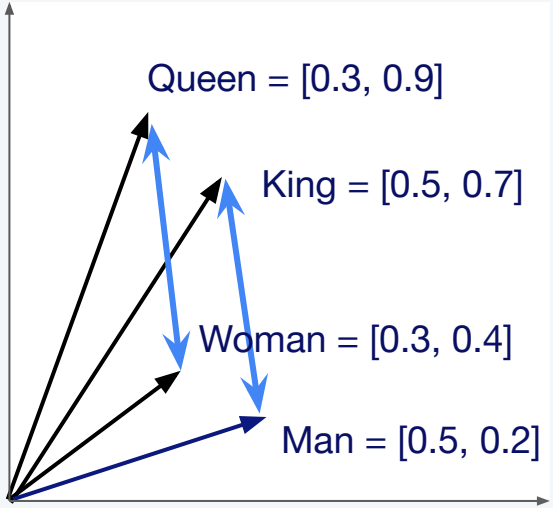
Use math to quantify relationships between entities

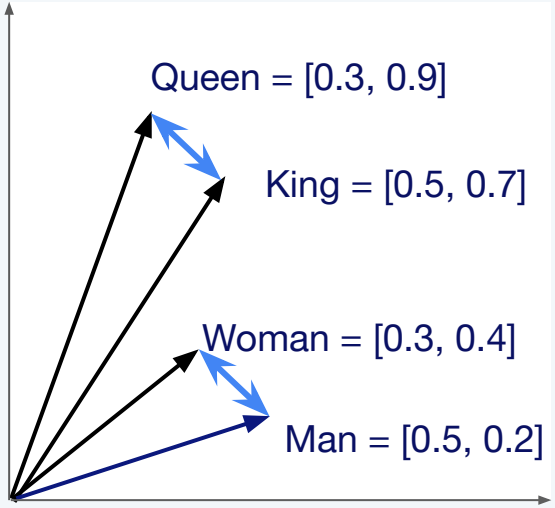
02

How Do Vector Databases Work?

Vector similarity is a mathematical measure of how close two vectors are

Semantic Similarity





Queen - Woman + Man = King

Queen = [0.3, 0.9]
\- Woman = [0.3, 0.4]

[0.0, 0.5]
\+ Man = [0.5, 0.2]

King = [0.5, 0.7]

Image from Sutor et al

Similarity metrics are ways to measure distance in vector space

Vector Similarity Metric: L2 (Euclidean)

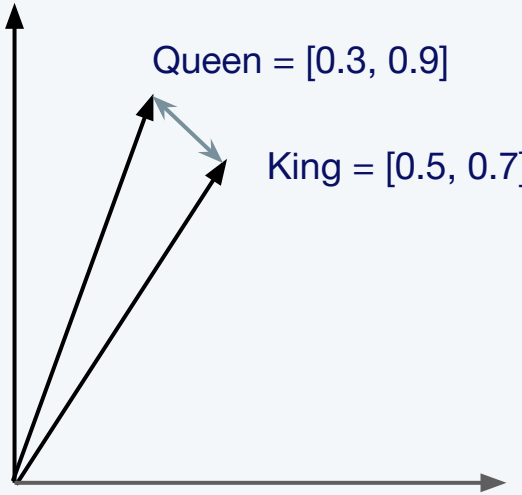
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

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

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

= $\sqrt{0.04 + 0.04}$

= $\sqrt{0.08} \cong 0.28$

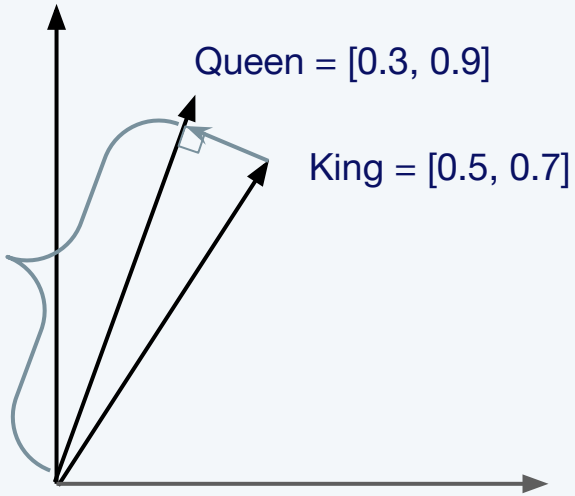


Vector Similarity Metric: Inner Product (IP)

$$a \cdot b = \sum_{i=1}^{n} a_i b_i$$

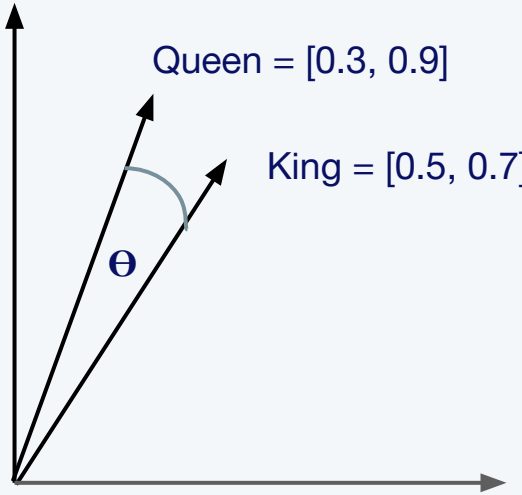
Queen · King = (0.3*0.5) + (0.9*0.7)

= 0.15 + 0.63 = 0.78



Vector Similarity Metric: Cosine

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



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

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

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

≅ 0.03

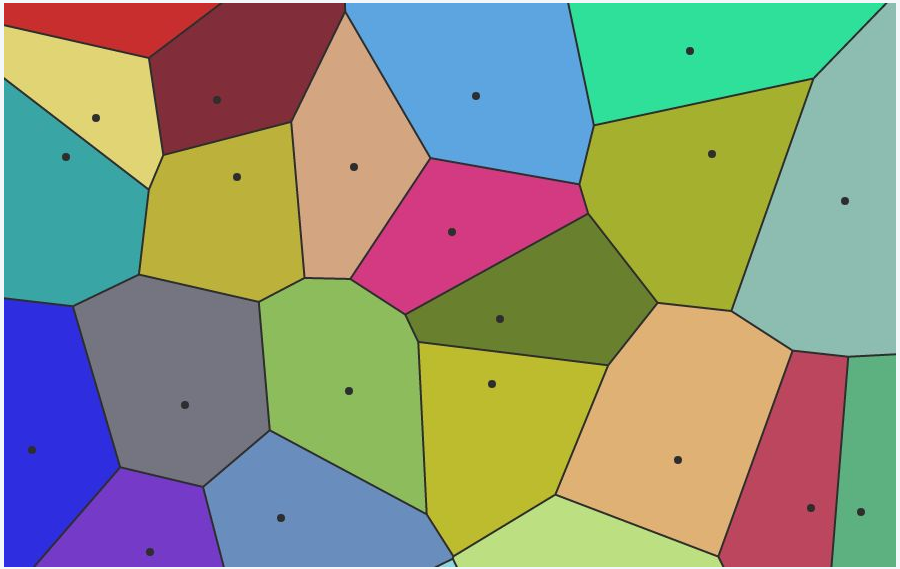
Vector Similarity Metrics

Euclidean - Spatial distance
Cosine - Orientational distance
Inner Product - Both

With normalized vectors, IP = Cosine

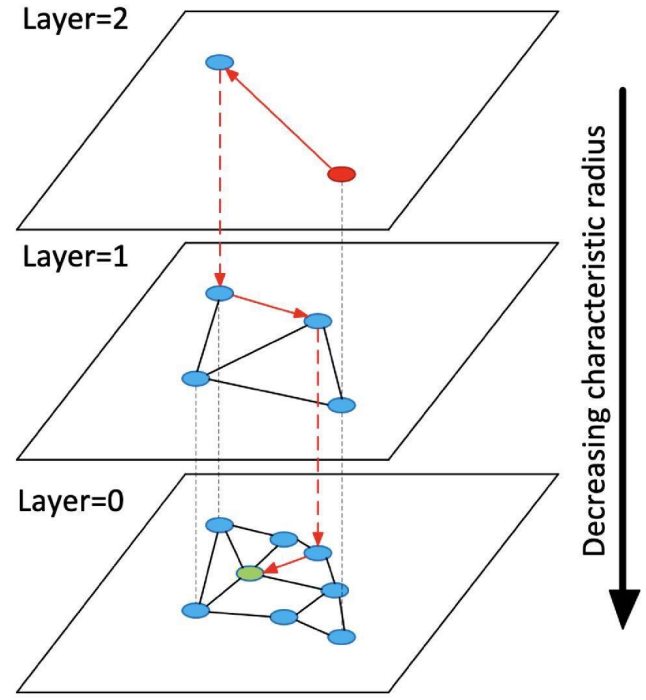
Indexes organize the way we access our data

Inverted File Index



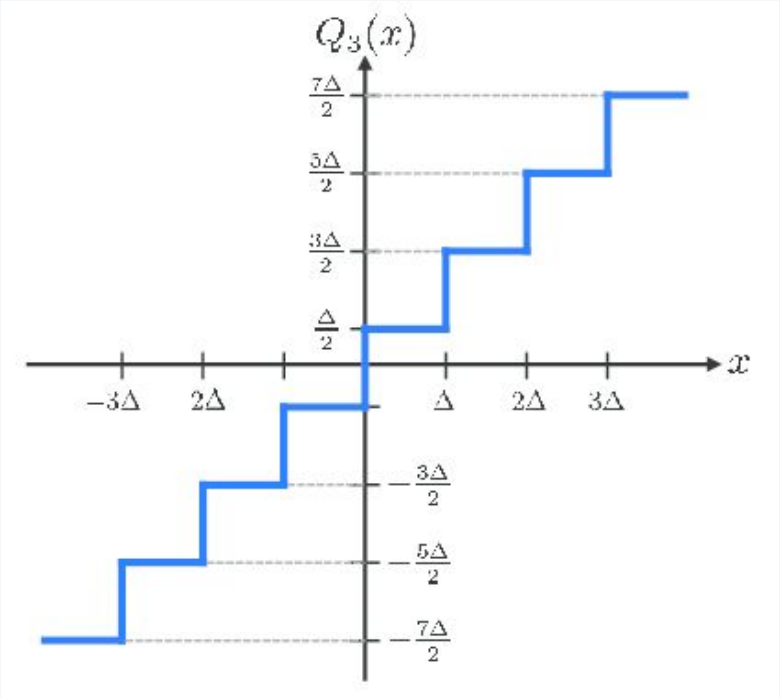
Source:
https://towardsdatascience.com/similarity-search-with-ivfpq-9c6348fd4db3

Hierarchical Navigable Small Worlds (HNSW)

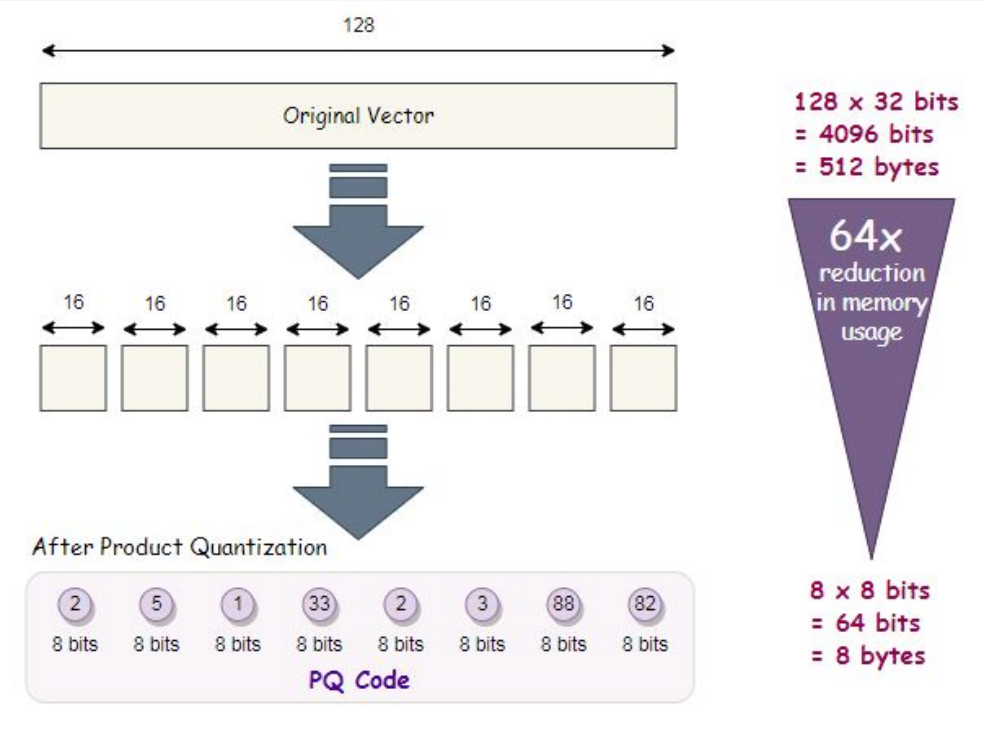


Source:
https://arxiv.org/ftp/arxiv/papers/1603/1603.09320.pdf

Scalar Quantization (SQ)



Product Quantization



Source:
https://towardsdatascience.com/product-quantization-for-similarity-search-2f1f67c5fddd

Indexes Overview

- IVF = Intuitive, medium memory, performant
- HNSW = Graph based, high memory, highly performant
- Flat = brute force
- SQ = bucketize across one dimension, accuracy x memory tradeoff
- PQ = bucketize across two dimensions, more accuracy x memory tradeoff

Vector databases efficiently store, index, and relate entities by a quantitative value

03

Use Cases

What Does Vector Data Look Like?

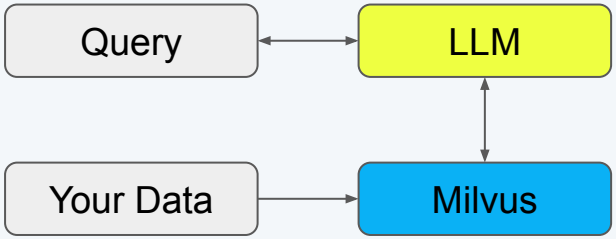
"id": "https://towardsdatascience.com/detection-of-credit-card-fraud-with-an-autoencoder-9275854
"embedding": [-0.042092223,-0.0154002765,-0.014588429,-0.031147376,0.03801204,0.013369046,0
"date": "2023-06-01"
"paragraph": "We define an anomaly as follows:"
"reading_time": "11"
"subtitle": "A guide for the implementation of an anomaly..."
"publication": "Towards Data Science"
"responses": "1"
"article_url": "https://towardsdatascience.com/detection-of-credit-card-fraud-with-an-autoencoder-
"title": "Detection of Credit Card Fraud with an Autoencoder"
"claps": "229"

↑ Hide 6 fields Vector search

RAG

RAG

Inject your data via a vector database like Milvus/Zilliz



Primary Use Case

- Factual Recall
- Forced Data Injection
- Cost Optimization

Common AI Use Cases

LLM Augmented Retrieval

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 behaviors or features 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 from massive amounts of audio data to perform tasks such as genre classification, or recognize speech.

Molecular Similarity Search

Search for similar substructures, superstructures, and other structures for a specific molecule.

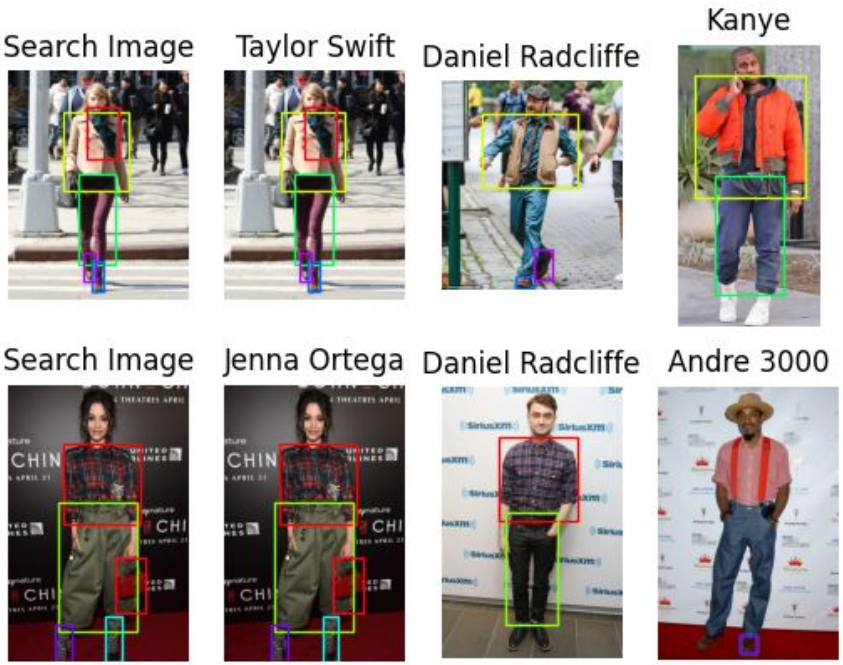
Question Answering System

Interactive QA chatbot that automatically answers user questions

Multimodal Similarity Search

Search over multiple types of data simultaneously, e.g. text and images

Example Use Case



Example Use Case

Example Use Case

```
pprint(res_512_plot.response)
```

```
('The plot of "The Nightmare Before Christmas" revolves around Jack '
 'Skellington, the Pumpkin King of Halloween Town, who becomes tired of the '
 'same routine of Halloween and discovers Christmas Town. Intrigued by the '
 'concept of Christmas, Jack decides that Halloween Town will take over '
 'Christmas this year. He assigns the residents various Christmas-themed '
 "tasks, but his efforts lead to disastrous consequences. Jack's love "
 'interest, Sally, warns him about the potential disaster, but he dismisses '
 'her warnings. Eventually, Jack realizes his mistake and sets out to fix the '
 'chaos he has caused. With the help of Santa Claus, Jack saves Christmas and '
 'learns the true meaning of the holiday. The film ends with Jack and Sally '
 'declaring their love for each other [6][7][8][9].')
```

04

Vector Database Architecture

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

Why Not Use a Vector Search Library?

- Have to manually implement filtering
- Not optimized to take advantage of the latest hardware
- Unable to handle large scale data
- Lack of lifecycle management
- Inefficient indexing capabilities
- No built in safety mechanisms

TL;DR: Vector search libraries **lack the infrastructure** to help you **scale**, **deploy**, and **manage** your apps in production.

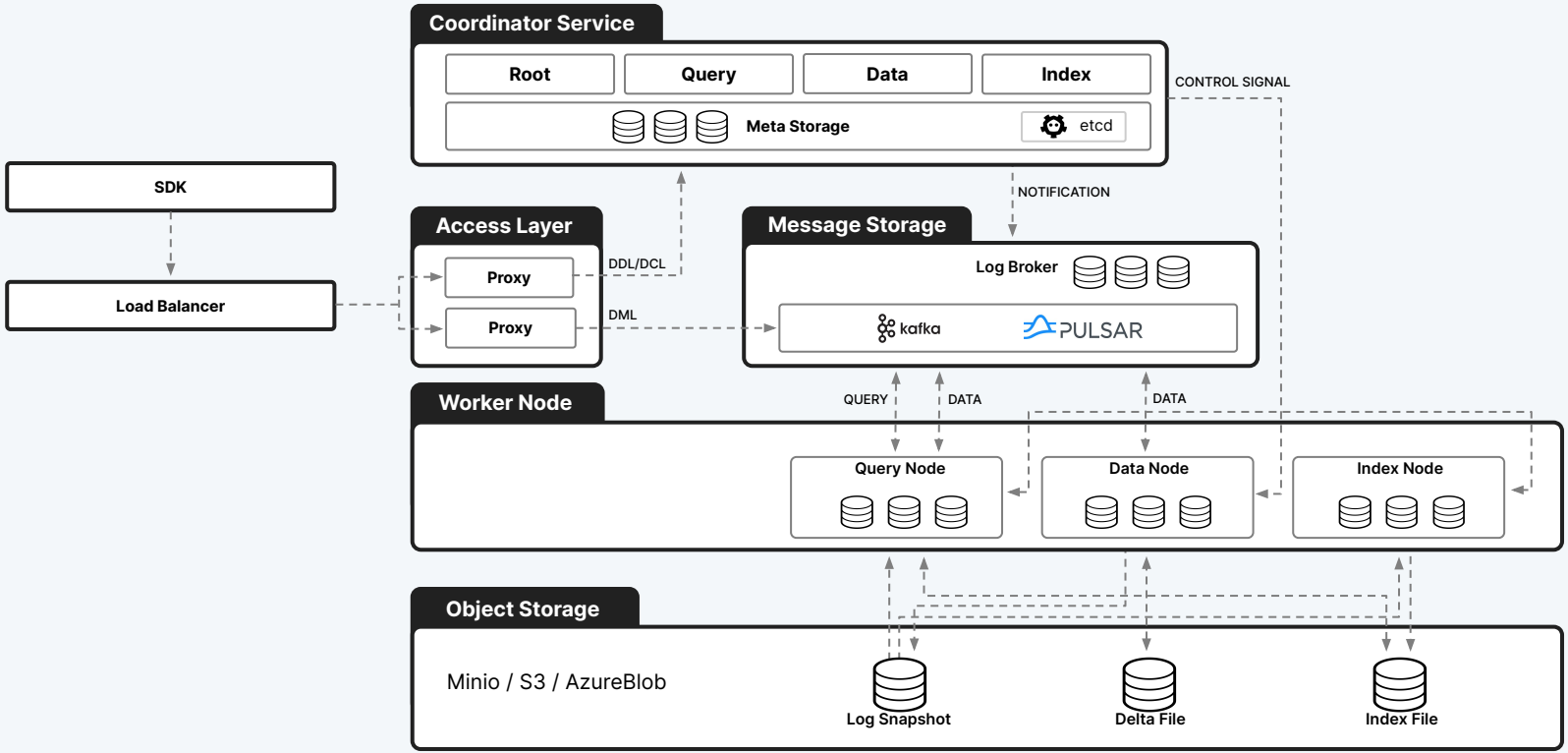
What is Milvus/Zilliz ideal for?

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

- Advanced filtering
- Hybrid 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

High-level overview of Milvus' Architecture



Start building
with Zilliz Cloud today!
zilliz.com/cloud

Appendix

 |

zilliz

Important Notes

- Cosine, IP, and L2 are all the SAME rank order.
- They differ in **use case**
  - L2 for when you need magnitude
  - Cosine for orientation
  - IP for magnitude and orientation

- OR
  - Cosine = IP for normalized vectors

Embeddings Models

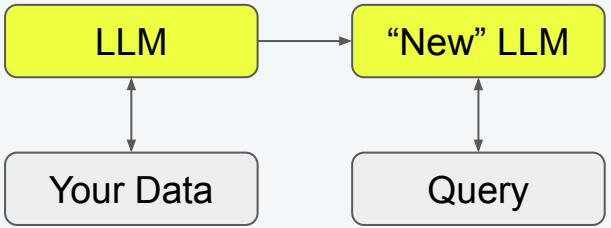
Basic Idea

You want to use your data with a large language model

RAG vs Fine Tuning

Fine Tuning

Augment an LLM by training it on your data



Primary Use Case

- Style transfer

Takeaway

Use RAG to force the LLM to work with your data by injecting it via a vector database like Milvus or Zilliz

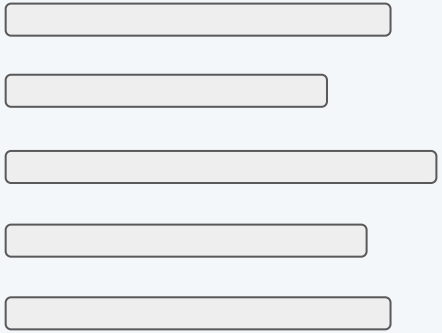
Chunking Considerations

Chunk Size

Chunk Overlap

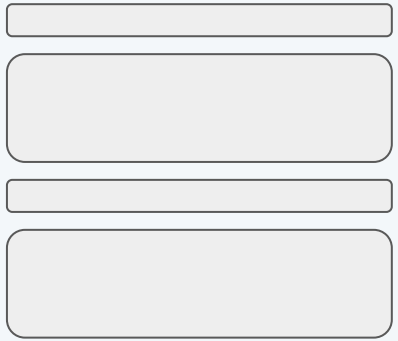
Character Splitters

How Does Your Data Look?



Conversation Data

Documentation Data



Lecture or Q/A Data

Examples

```
(exps) yujiantang@Yujians-MacBook-Pro experimentation_scripts % python test_langchain_chunking.py
Responses from chunking strategy:
         32, 4
page_content='is a Distinguished Engineer.'
page_content='engineer.'
page_content='as an engineer.'
page_content='engineers) junior engineer.'
Responses from chunking strategy:
         64, 8
page_content='engineer.'
page_content='accomplishments while working as an engineer.'
page_content='- Werner Vogels, CTO of Amazon, is a Distinguished Engineer.'
page_content='of Amazon/Microsoft/Google engineers) junior engineer.'
Responses from chunking strategy:
         128, 16
page_content='- Has achieved noteworthy technical, professional accomplishments while working as an engineer.'
page_content='senior software engineer.'
page_content='- Provides solid technical leadership\xa0*beyond the company*.\n- Werner Vogels, CTO of Amazon, is a Distinguished Engineer.'
page_content='- Everything on the\xa0*Senior Staff Engineer*\xa0list, plus:\n- Generally 15-20 years of experience.'
Responses from chunking strategy:
         256, 32
page_content='- This will ordinarily require significant amounts of time over the lifetime of the individual given this distinction.\n- Provides solid
technical leadership\xa0*beyond the company*.\n- Werner Vogels, CTO of Amazon, is a Distinguished Engineer.'
page_content='- More and better than the last. Less day-to-day coding, and more thinking, planning, and directing.\n- Has achieved noteworthy technical
, professional accomplishments while working as an engineer.'
page_content='- Everything on the\xa0*Senior Staff Engineer*\xa0list, plus:\n- Generally 15-20 years of experience.\n- Has led or spearheaded several n
on-trivial or important projects.\n- Provides solid technical leadership\xa0*across the company*.'
page_content='- Everything on the\xa0*Staff Engineer*\xa0list, plus:\n- Generally 10-15 years of experience.\n- Has led or spearheaded multiple non-tri
vial or important projects.\n- Provides solid technical leadership\xa0*beyond*\xa0their team.'
Responses from chunking strategy:
         512, 64
page_content='- More and better than the last. Less day-to-day coding, and more thinking, planning, and directing.\n- Has achieved noteworthy technical
, professional accomplishments while working as an engineer.\n- This will ordinarily require significant amounts of time over the lifetime of the indiv
idual given this distinction.\n- Provides solid technical leadership\xa0*beyond the company*.\n- Werner Vogels, CTO of Amazon, is a Distinguished Engin
eer.'
page_content='- Everything on the\xa0*Senior Staff Engineer*\xa0list, plus:\n- Generally 15-20 years of experience.\n- Has led or spearheaded several n
on-trivial or important projects.\n- Provides solid technical leadership\xa0*across the company*.\n- Typically has one or more patents.'
page_content='- Everything on the\xa0*Staff Engineer*\xa0list, plus:\n- Generally 10-15 years of experience.\n- Has led or spearheaded multiple non-tri
vial or important projects.\n- Provides solid technical leadership\xa0*beyond*\xa0their team.'
page_content='- More and better than the last.\n- Provides solid technical leadership\xa0*across the industry*.\n- James Gosling, creator of the Java p
rogramming language, is a Fellow.'
```

Examples

```
Responses from chunking strategy:
          32, 4
page_content='is a Distinguished Engineer.'
page_content='engineer.'
page_content='as an engineer.'
page_content='engineers) junior engineer.'
```

Examples

```
Responses from chunking strategy:
          64, 8
page_content='engineer.'
page_content='accomplishments while working as an engineer.'
page_content='- Werner Vogels, CTO of Amazon, is a Distinguished Engineer.'
page_content='of Amazon/Microsoft/Google engineers) junior engineer.'
```

Examples

```
Responses from chunking strategy:
          128, 16
page_content='- Has achieved noteworthy technical, professional accomplishments while working a
s an engineer.'
page_content='senior software engineer.'
page_content='- Provides solid technical leadership\xa0*beyond the company*.\n- Werner Vogels,
CTO of Amazon, is a Distinguished Engineer.'
page_content='- Everything on the\xa0*Senior Staff Engineer*\xa0list, plus:\n- Generally 15-20
years of experience.'
```

Examples

```
Responses from chunking strategy:
          256, 32
page_content='- This will ordinarily require significant amounts of time
over the lifetime of the individual given this distinction.\n- Provides s
olid technical leadership\xa0*beyond the company*.\n- Werner Vogels, CTO
of Amazon, is a Distinguished Engineer.'
page_content='- More and better than the last. Less day-to-day coding, an
d more thinking, planning, and directing.\n- Has achieved noteworthy tech
nical, professional accomplishments while working as an engineer.'
page_content='- Everything on the\xa0*Senior Staff Engineer*\xa0list, plu
s:\n- Generally 15-20 years of experience.\n- Has led or spearheaded seve
ral non-trivial or important projects.\n- Provides solid technical leader
ship\xa0*across the company*.'
page_content='- Everything on the\xa0*Staff Engineer*\xa0list, plus:\n- G
enerally 10-15 years of experience.\n- Has led or spearheaded multiple no
n-trivial or important projects.\n- Provides solid technical leadership\x
a0*beyond*\xa0their team.'
```

Examples

```
Responses from chunking strategy:
          512, 64
page_content='- More and better than the last. Less day-to-day coding, an
d more thinking, planning, and directing.\n- Has achieved noteworthy tech
nical, professional accomplishments while working as an engineer.\n- This
 will ordinarily require significant amounts of time over the lifetime of
 the individual given this distinction.\n- Provides solid technical leade
rship\xa0*beyond the company*.\n- Werner Vogels, CTO of Amazon, is a Dist
inguished Engineer.'
page_content='- Everything on the\xa0*Senior Staff Engineer*\xa0list, plu
s:\n- Generally 15-20 years of experience.\n- Has led or spearheaded seve
ral non-trivial or important projects.\n- Provides solid technical leader
ship\xa0*across the company*.\n- Typically has one or more patents.'
page_content='- Everything on the\xa0*Staff Engineer*\xa0list, plus:\n- G
enerally 10-15 years of experience.\n- Has led or spearheaded multiple no
n-trivial or important projects.\n- Provides solid technical leadership\x
a0*beyond*\xa0their team.'
page_content='- More and better than the last.\n- Provides solid technica
l leadership\xa0*across the industry*.\n- James Gosling, creator of the J
ava programming language, is a Fellow.'
```

Takeaway:

Your chunking strategy depends on **what your data looks like** and **what you need from it.**

Examining Embeddings

Picking a model

What to embed

Metadata

Embeddings Strategies

Level 1: Embedding Chunks Directly

Level 2: Embedding Sub and Super Chunks

Level 3: Incorporating Chunking and Non-Chunking Metadata

Metadata Examples

Chunking

- Paragraph position
- Section header
- Larger paragraph
- Sentence Number
- …

Non-Chunking

- Author
- Publisher
- Organization
- Role Based Access Control
- …

Takeaway:

Your embeddings strategy depends on your **accuracy**,

cost, and **use case** needs

Basic Idea

Vector Databases provide the ability to inject your data via semantic similarity

Considerations include: scale, performance, and flexibility

Milvus Architecture: Differentiation

1. Cloud Native, Distributed System Architecture
2. True Separation of Concerns
3. Scalable Index Creation Strategy with 512 MB Segments

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.

Vector Database Resources

Give Milvus a Star!

Chat with me on Discord!

Get Started Free

Milvus	Zilliz Cloud
Open Source Self-Managed	SaaS Fully-Managed
github.com/milvus-io/milvus	zilliz.com/cloud

Got questions? Stop by our booth!