
GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
数据与数据库技术
为大模型构建的AI原生数据库Milvus
栾小凡 2023年05月25日

栾小凡

Zilliz 技术合伙人，工程总监

毕业于康奈尔大学计算机工程系

前阿里巴巴高级技术专家、Oracle 数据库加速工程师

热爱开源，Milvus 社区 Chair、HBase/Cassandra 贡献者

前 DBer, 目前主要兴趣在AIGC,多模态大模型，非结构数据处理

zilliz

Zilliz — 构建开源+云的大模型加强方案

Milvus

Milvus is an open source vector database used to store, index, and manage massive embedding vectors generated by deep neural networks and other machine learning (ML) models.

大模型知识库

Towhee

Towhee makes it easy to build neural data processing pipelines for AI applications. With hundreds of models, algorithms, and transformations, Towhee helps you encode your unstructured data into embeddings.

大模型编排

GPTCache

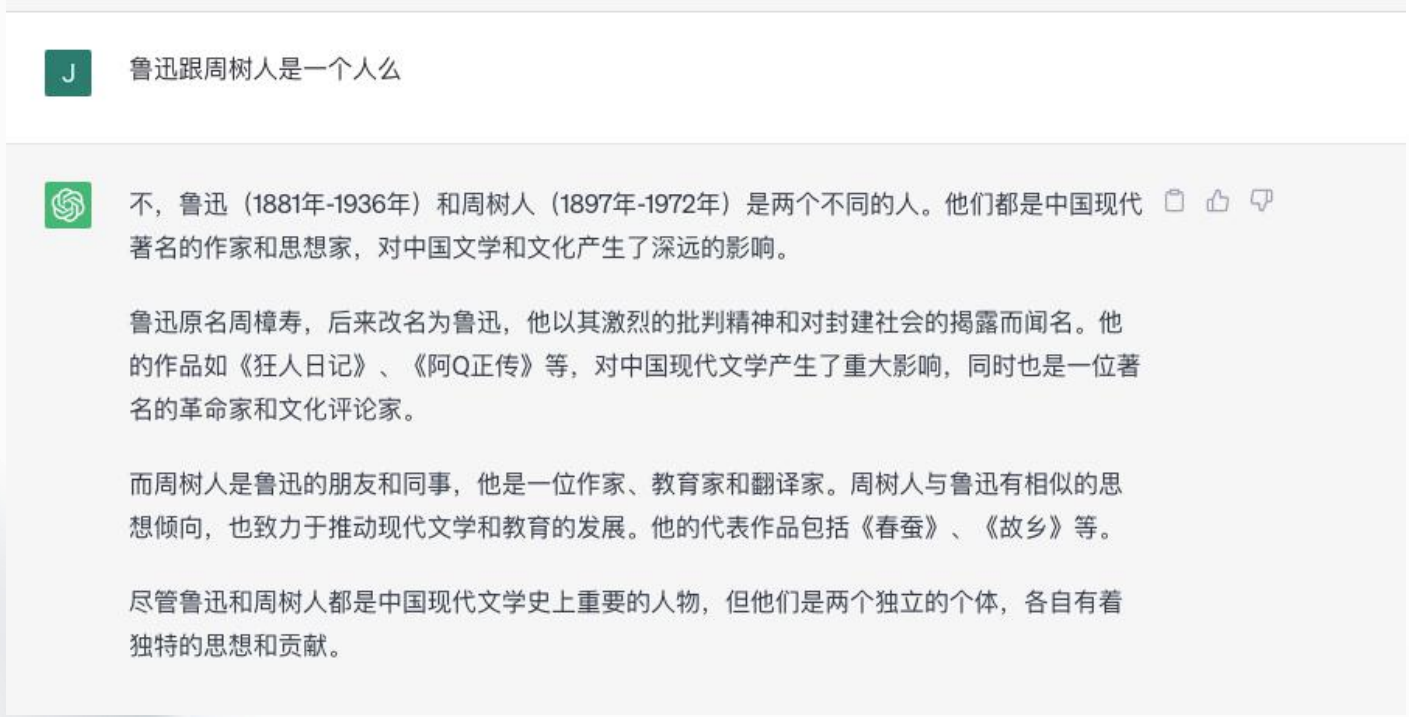
GPTCache is an open-source tool designed to improve the efficiency and speed of GPT-based applications by implementing a cache to store the responses generated by language models.

大模型缓存

Zilliz Cloud —
开箱即用的全托管
向量检索服务

Content

大模型的幻觉



在自然语言处理中，幻觉通常被定义为“生成的内容与提供的源内容无意义或不可信”。

Meta 将其定义为“自信的假话”。

维基百科：幻觉（人工智能）

大模型的幻觉

幻觉产生的原因 – 信息密度不足

大模型的幻觉

如何解决“幻觉”？

微调模型

- 数据准备
- 训练资源
- 训练时间

提示工程

- 内容选择
- 长度限制

知识库

- 内容检索
- 语义理解
- 数据存储

停止幻觉 — AIGC时代的CVP架构

ChatGPT：

- 利用大语言模型（LLM）实现以 ChatGPT 为代表的智能问答

Vector database：

- 通过向量数据库为 ChatGPT 提供大规模的、可靠的知识库

Prompt-as-code：

- 使用提示匹配用户问题与来自知识库的参考内容

ChatGPT

- 智能对话机器人
- 预训练大语言模型 GPT 模型（GPT-3.5, GPT-4，...）
- OpenAI

更多大语言模型（LLMs）

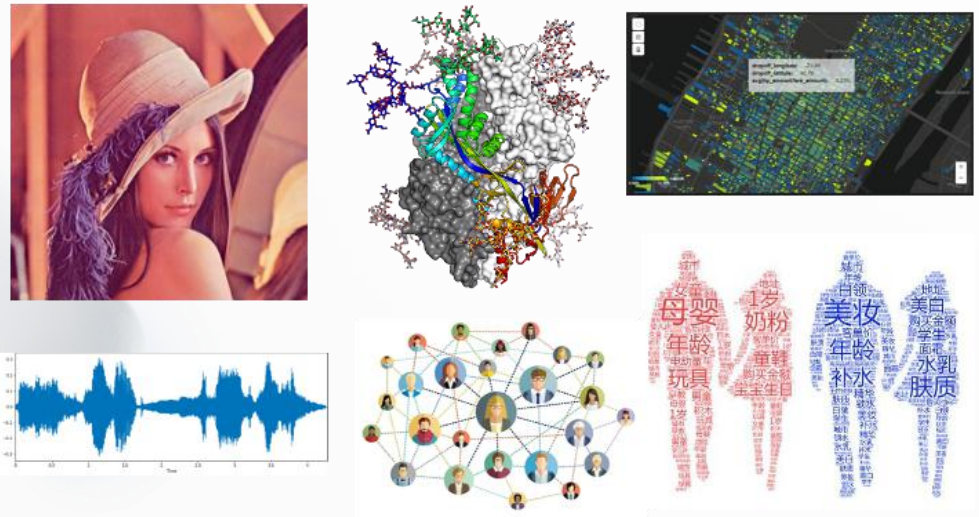
- PaLM （谷歌）
- LLaMA（Meta AI）
- Alpaca（斯坦福）
- 文心一言（百度）
- ...

如何对接比较多个大语言模型

1. Langchain
2. GptCache

向量：来自预训练模型的特征向量可以表示非结构化数据的深层语义，向量相似度用于语义搜索

向量数据库：存储向量数据，并能进行高效率的检索和编辑的程序



Data UID[1]	Vector representation
0	[-0.31, 0.53, -0.18, …, -0.16, -0.38]
1	[0.58, 0.25, 0.61, …, -0.03, -0.31]
2	[-0.07, -0.53, -0.02, …, -0.61, 0.59]
…	

CVP — Prompt Engineering

Prompts from Langchain:

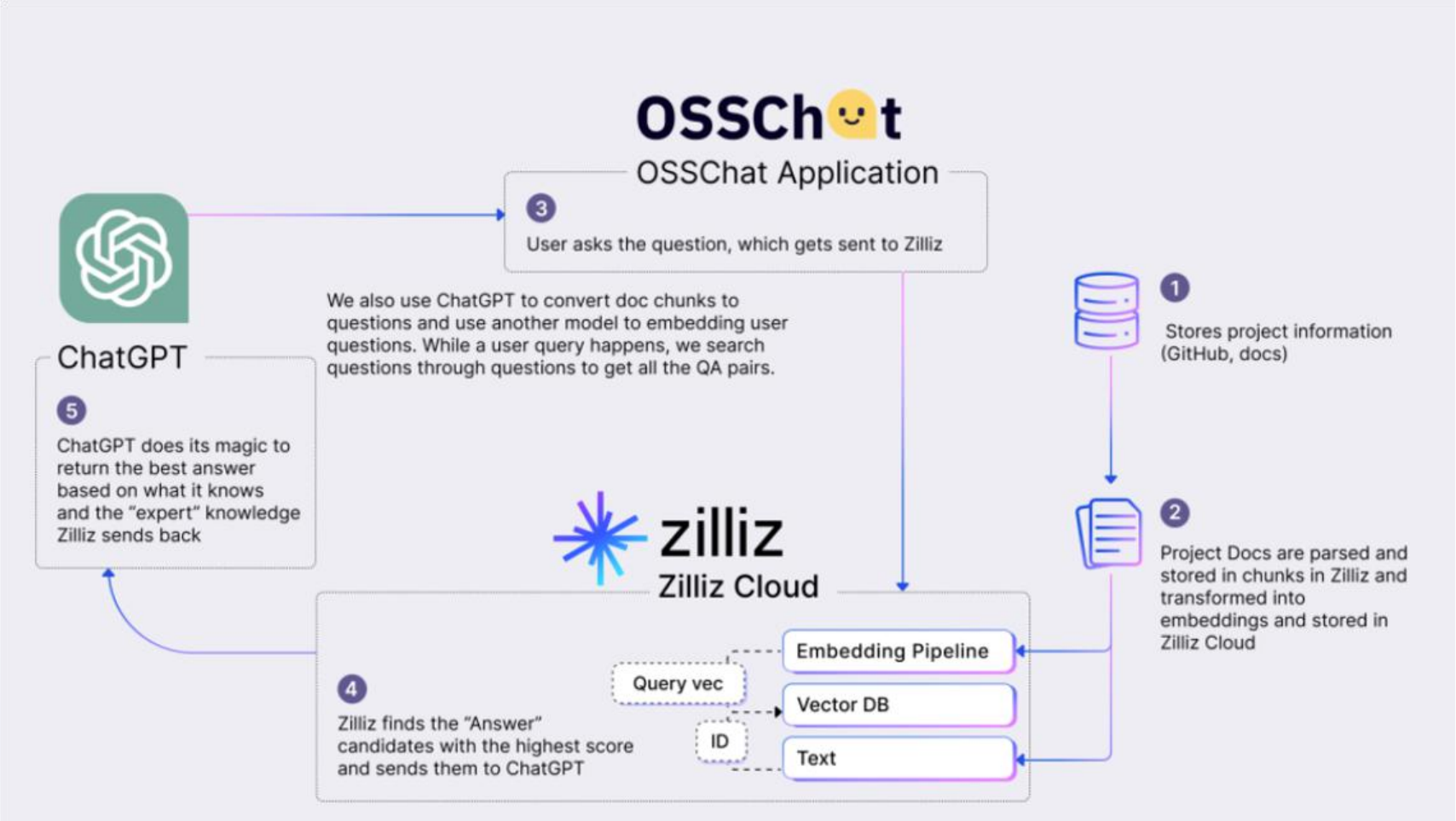
```
Use the following pieces of context to an
swer the question at the end. If you do
n't know the answer, just say that you do
n't know, don't try to make up an answer.
{context}
Question: {question}
Helpful Answer:
```

```
Context information is below.
---------------------
{context_str}
---------------------
Given the context information and not pri
or knowledge,
answer the question: {question}
```

提示工程（Prompt engineering）是人工智能中的一个概念，特别是自然语言处理（NLP）。 在提示工程中，任务的描述会被嵌入到输入中。例如，不是隐含地给予模型一定的参数，而是以问题的形式直接输入。

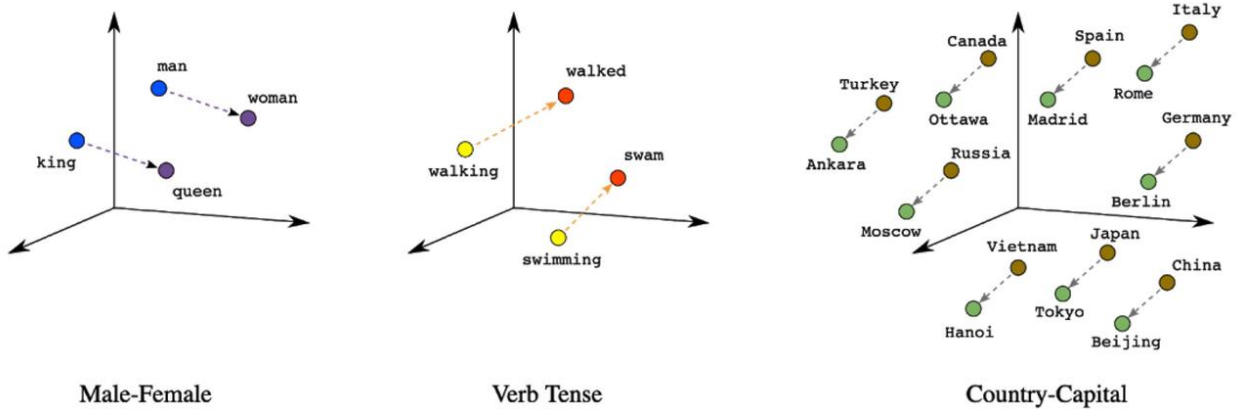
—— 维基百科：提示工程

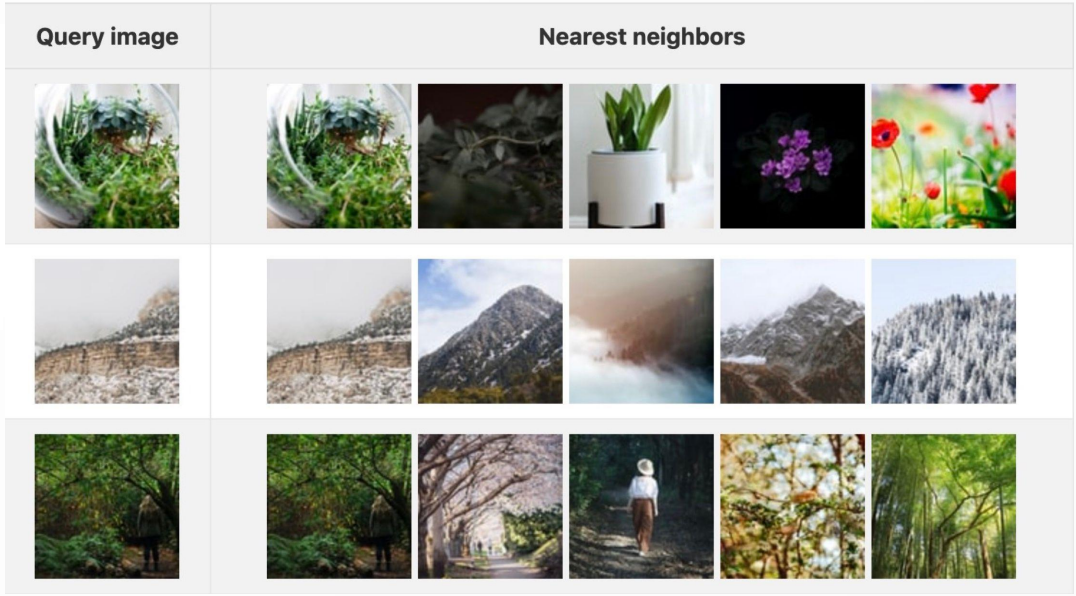
CVP应用

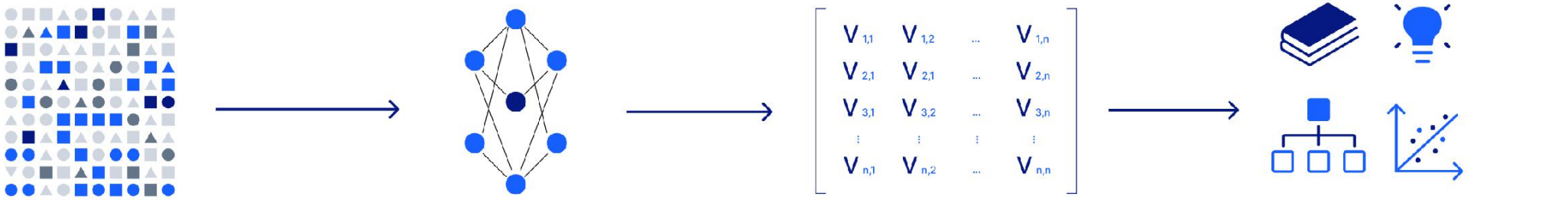


osschat.io 面向开源用户和开源社区运维人员的问答机器人

面向AI的数据库 — 向量数据库







构建一个向量数据库

你需要关注

- 低成本的向量存储和持久化
- 高性能的向量ANN检索
- 数据的增删改查实现
- 标量向量混合查询
- 可扩展性
- 易用的查询能力 - Restful，Python，SQL..
- 高可
- 异构硬件的支持
- 数据备份、Snapshot、迁移、导入
- 监控，报警，流控.....

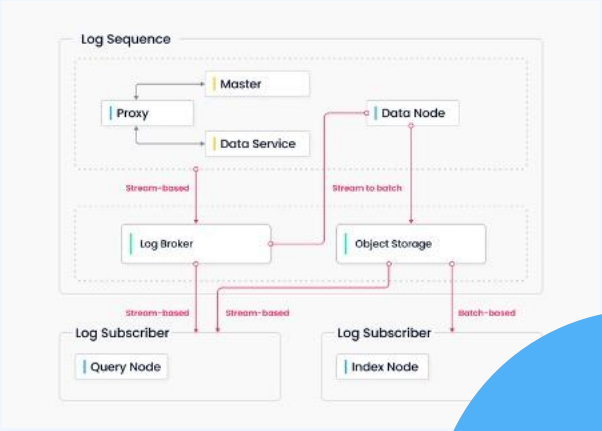
Milvus — 全球第一款向量数据库

Milvus 2.0

云原生分布式

Cloud Native

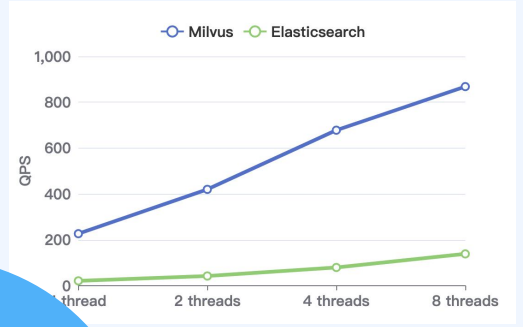
- 百亿规模向量扩展性
- 存储计算分离
- 离在线一体化
- 基于K8s实现高可用容灾



超高性能

Blazing Fast

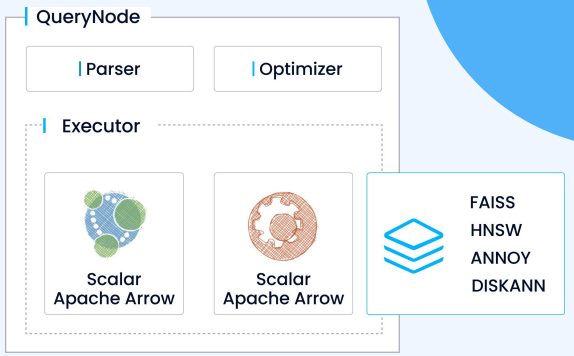
- 查询速度高于ES 10倍，高于主流竞品2倍
- 毫秒级延迟响应
- 查询性能根据物理资源线性扩展



可插拔引擎

Pluggable Engine

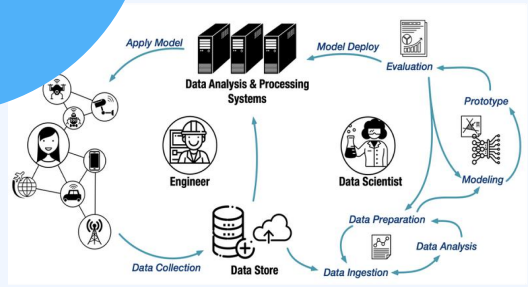
- 向量与标量混合查询
- 提供标量倒排索引支持
- 集成了 FAISS、HNSW、DISKANN 等SOTA 向量索引



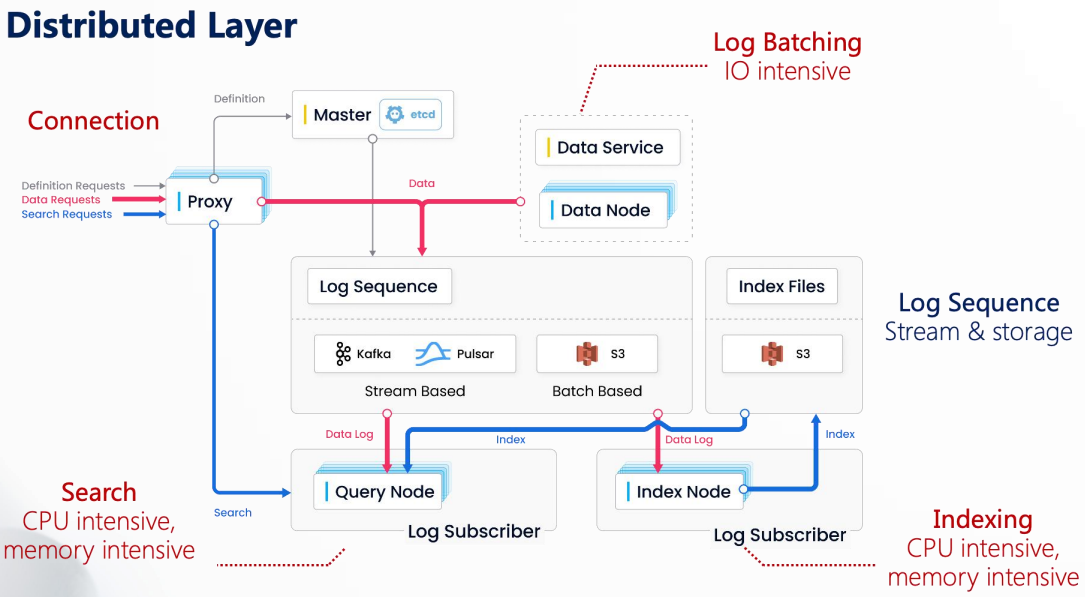
云端一体

Unify Cloud and Desktop

- 提供从笔记本，到线下机房到云完全一致的使用体验
- 丰富的部署方式，可观测性支持



为云而生的向量数据库

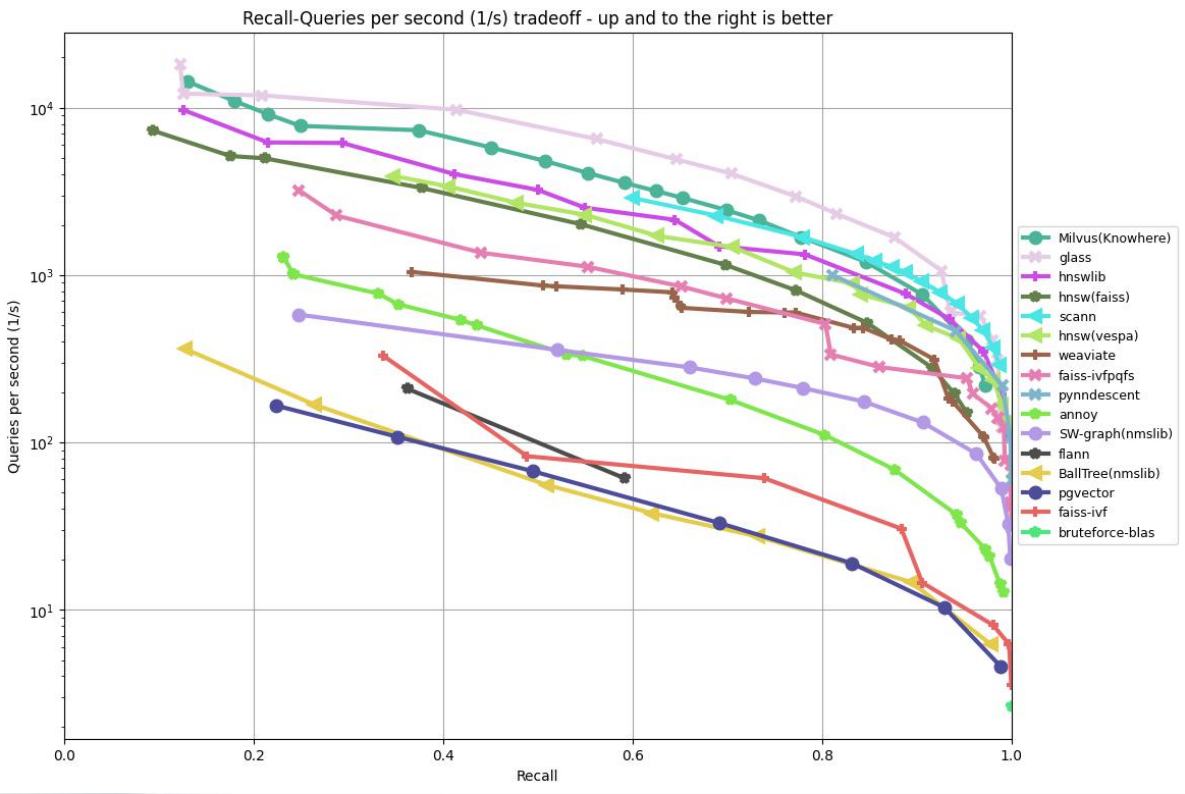


- 分布式云原生，基于K8s进行微服务化设计
- 存储计算分离，弹性扩缩容
- 高可用，故障分钟级恢复
- 百亿级向量的扩展能力
- 基于消息队列实现数据的实时增删
- 强大的生态工具 – GUI，CLI，监控，备份

向量数据库 + AIGC

- **动态 Schema/ Json 支持**
- **面向构建 SaaS 用户，通过 Partition 能力支持百万级租户**
- 支持磁盘索引 - 存储成本降低10倍
- 离线导入海量文档
- 多向量支持
- 与 OpenAI， Langchain， Llama-Index， AutoGPT，Towhee， Hugging face深度集成
- Python， Js， Golang， Java， C#， Restful **等丰富的客户端支持**

向量数据库性能 — 皇冠上的明珠



- Glass 是 Zilliz 自研的向量检索库实验室版本，未开源
- Knowhere 是目前开源的 Milvus 依赖的内核
- 通过 SIMD 优化，并发优化，索引数据结构调整，量化策略和缓存策略调整，实现了开源方案 ScaNN/HNSW 三倍以上的性能提升

GIST960数据集 https://ann-benchmarks.com/

Open Vector Search Benchmark - 即将开源敬请期待

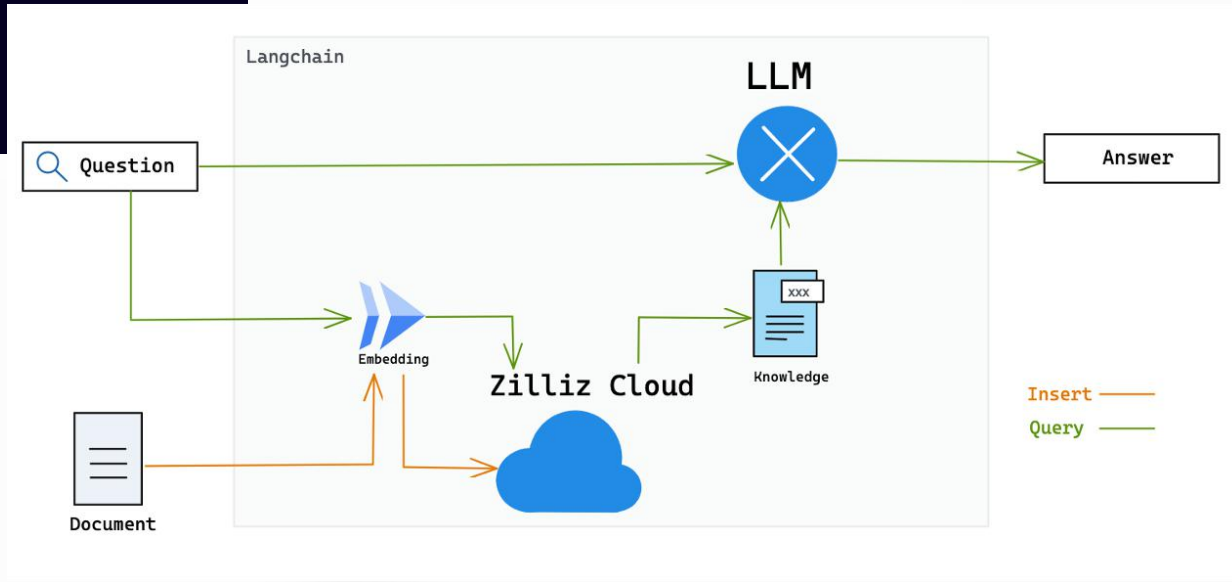
应用场景

Milvus 被全球超过1000家企业用户所信赖，超过350万次下载和安装

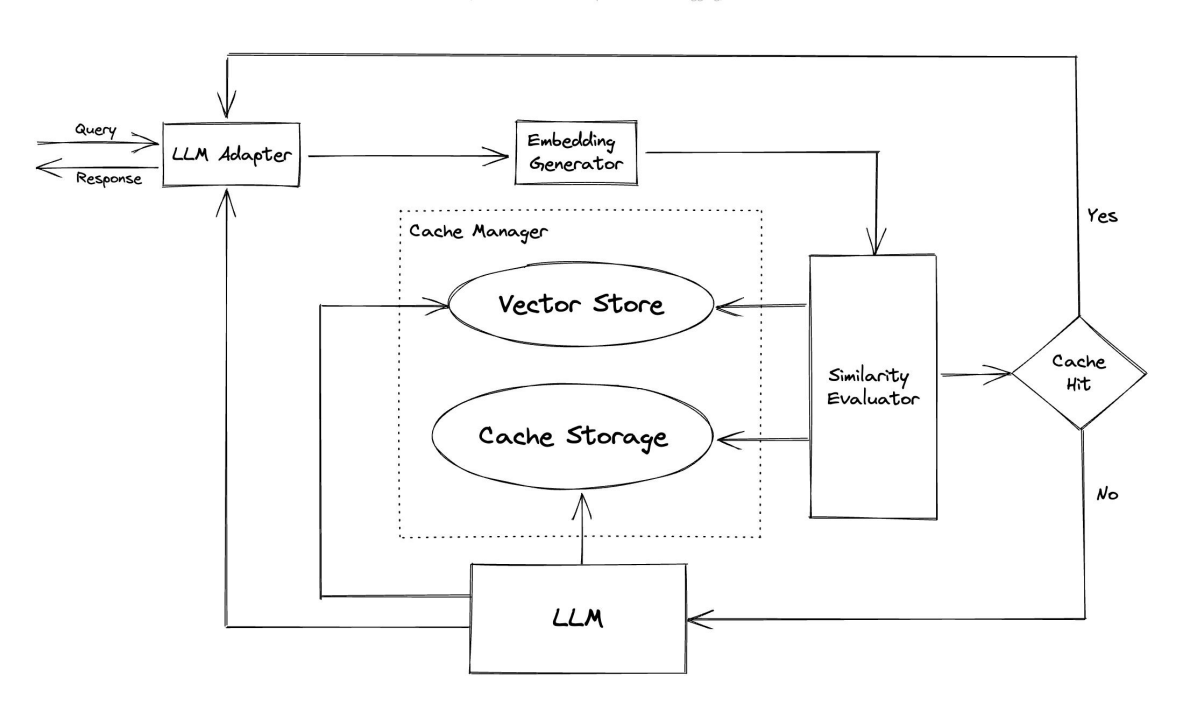
Milvus Github Star数目超过1.8万，贡献者人数超过300

Milvus DB-Engine 引擎排名 223，并且在SIGMOD和VLDB等数据库顶会上发表了论文

应用场景 – 基于大模型和向量数据库的问答系统

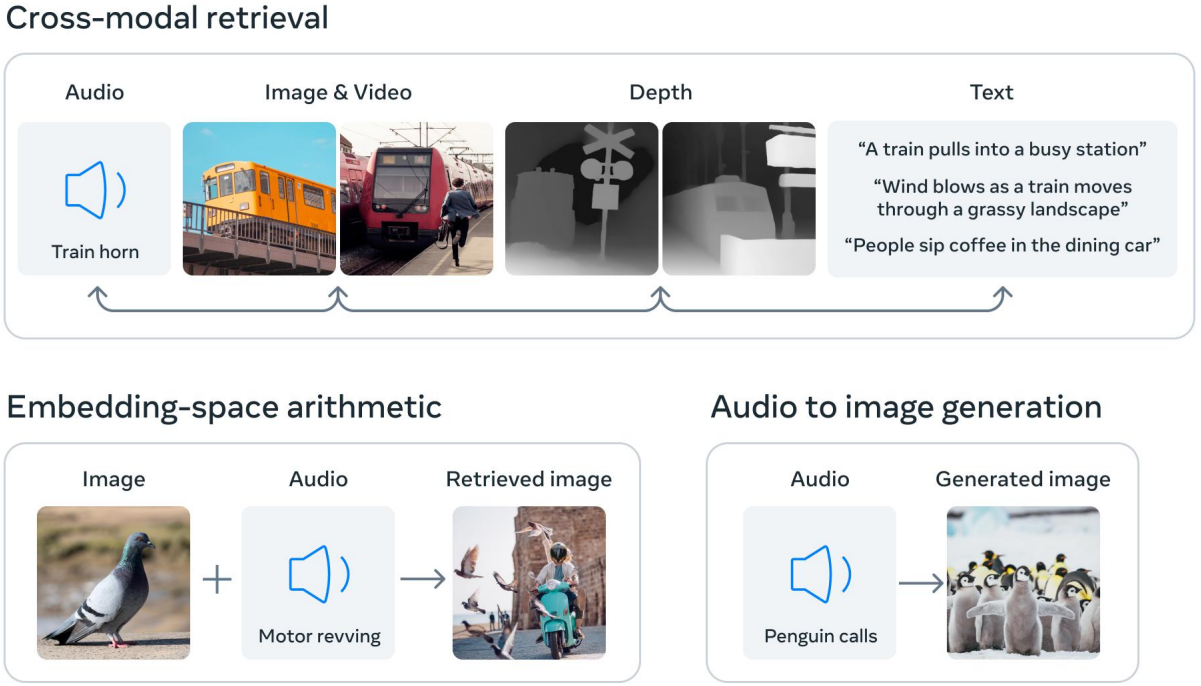


应用场景 — GPTCache



- 缓存大模型生成的结果和Context
- 支持多种向量/标量存储 – 如Milvus, FAISS， PGVector
- 支持丰富的Embedding模型
- 支持OpenAI以及本地部署LLamaCpp和Dolly
- 多模态支持

应用场景 – 多模态搜索



- Meta AI Imagebind
- 支持图像，文本，音频，深度，热，惯性测量传感器等六个模态

多模态搜索，Embedding语意叠加，多模态生成

https://imagebind.metademolab.com/

向量数据库的未来

部署运维

一键部署
滚动升级
弹性扩缩容

查询能力

混合查询
高性能过滤倒排
更丰富的数据类型

丰富的功能

数据备份
可视化
数据管理

易用的接口

SQL查询
GraphQL

性能成本

异构硬件加速
基于磁盘的索引

智能化

智能索引选择
智能过滤

向量数据库的未来 – 全托管SaaS服务 Zilliz Cloud

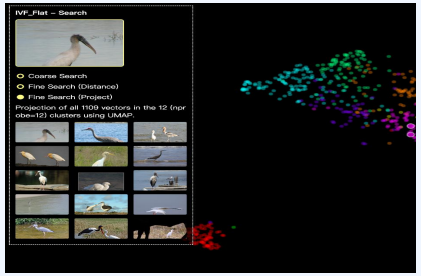
维护成本低
low maintenance cost

- 一键创建实例资源
- 动态扩缩容
- 完善的监控报警
- 多云支持

Google Cloud
阿里云
amazon

使用门槛低
Low threshold for use

- 免费实例
- 可视化界面
- 多语言SDK
- 丰富的生态支持
- 数据迁移



丰富的企业级特性
Enterprise-level features

- 7 * 24服务支持
- 99.9 SLA保障
- 数据备份，订阅
- 组织架构管理
- Dedicated Cloud

安全放心
Data Security

- RBAC权限管理
- TLS，白名单
- PrivateLink
- 审计日志
- SOC2合规认证

AICPA SOC
auth0

即将登陆阿里云，尽请期待

- Milvus: 开源向量数据库，免费使用，社区支持
- Towhee: 开箱即用的非结构化数据处理 ETL 工具
- GPTCache：为LLMs做记忆存储，省时又省钱

THANKS

Github

公众号

扫码并回复“技术交流”
加入用户交流群