DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

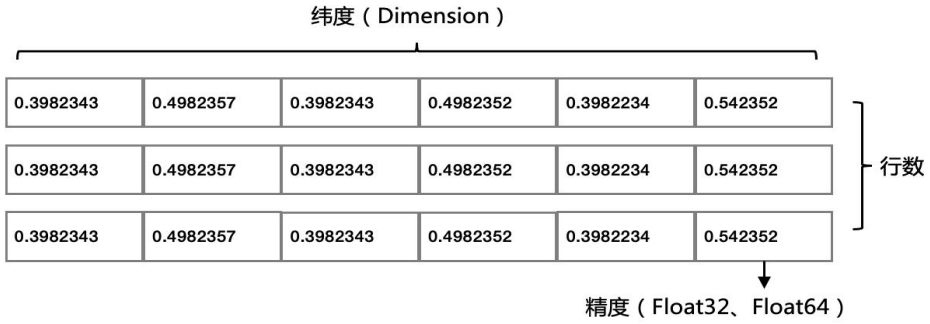
北京国际会议中心 | 2023/8/16-18

大模型时代下的向量数据库创新与挑战

腾讯云数据库专家工程师
伍旭飞

什么是向量检索

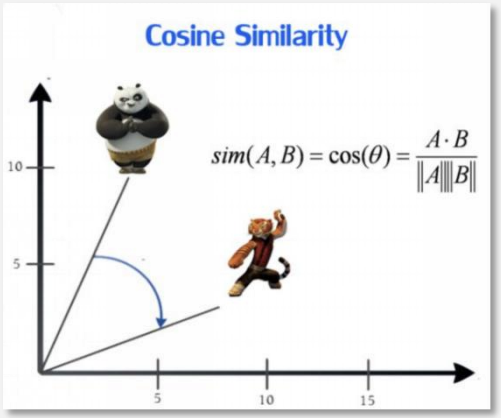
向量概念

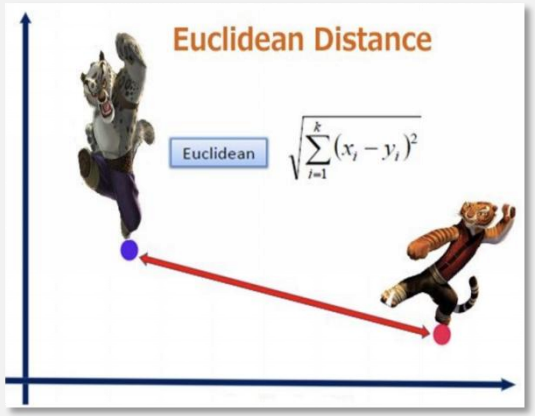


AI 中的特征向量

向量检索

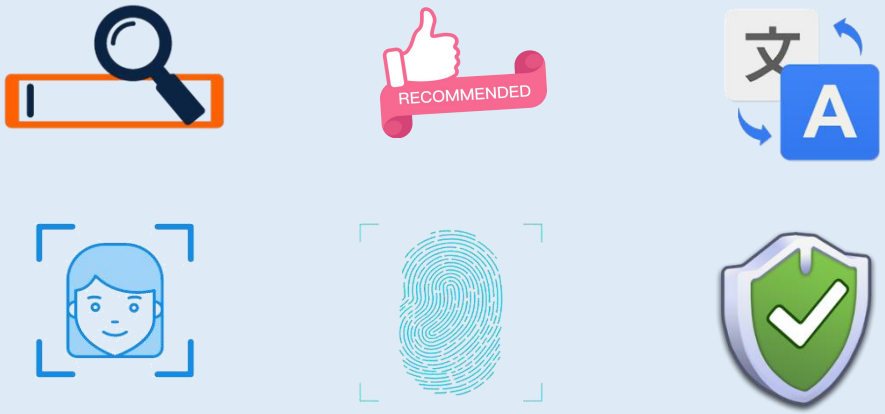
向量检索又称为近似最近邻搜索（Approximate Nearest Neighbor Search, ANNS），是一种在大规模高维向量数据中寻找与给定查询向量相似的向量的技术。向量检索在许多AI领域具有广泛的应用，如图像检索、文本检索、语音识别、推荐系统等。





腾讯在向量检索的积累

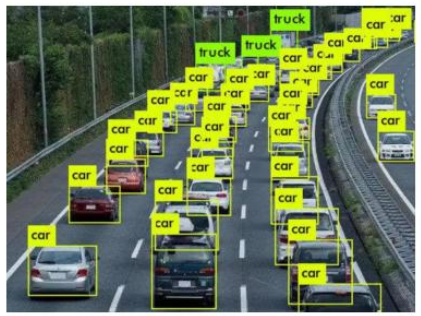
狭义人工智能时代，向量检索技术已经广泛应用



腾讯OLAMA向量检索引擎，在腾讯集团大规模应用

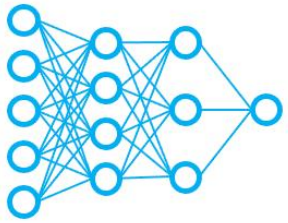
技术起源&发展历程

大模型下新场景-模型训练



Object Detection

1. Train



Neural Networks

2. Test

Bus or Traffic Signal?

3. Corner Case

4. Search for Similar Cases

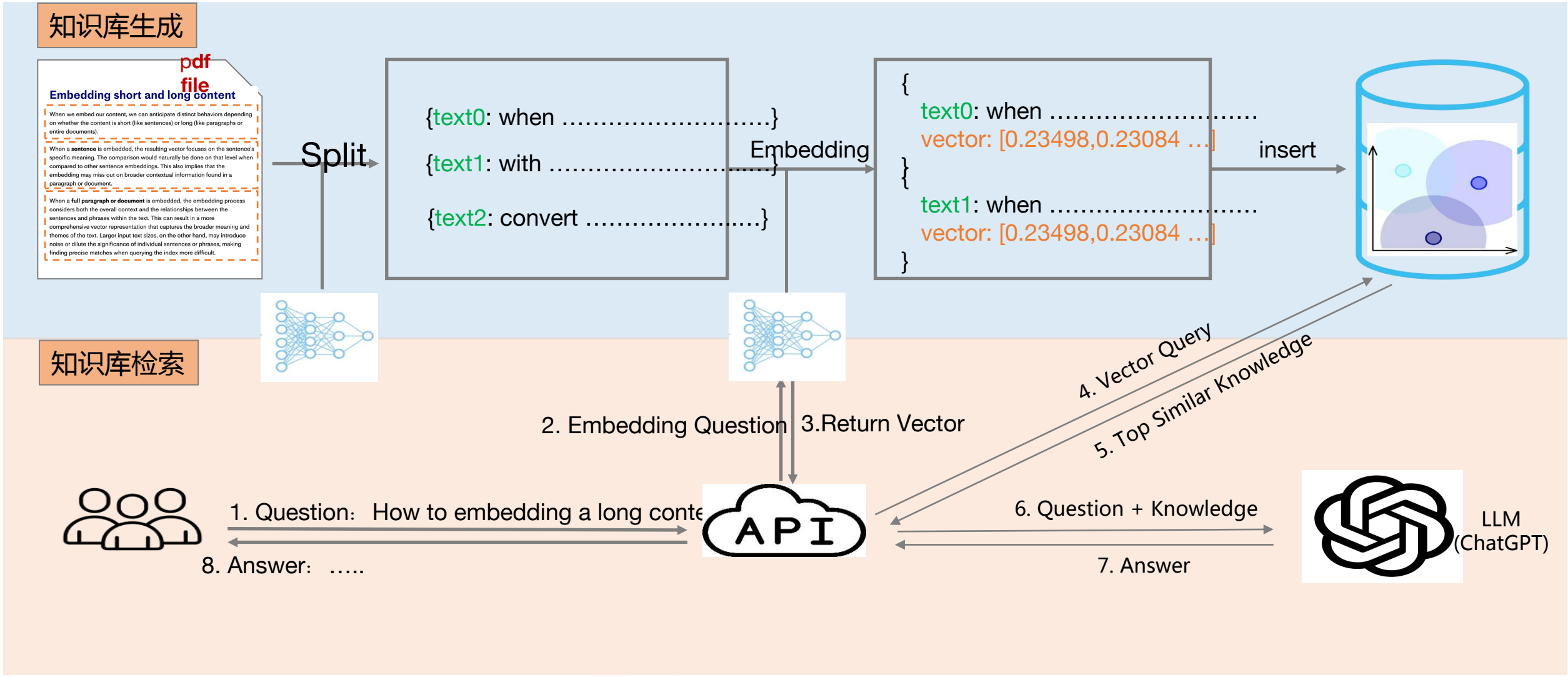
Upsert

More Dataset

Tencent Cloud
VectorDB

5. Fine-tune

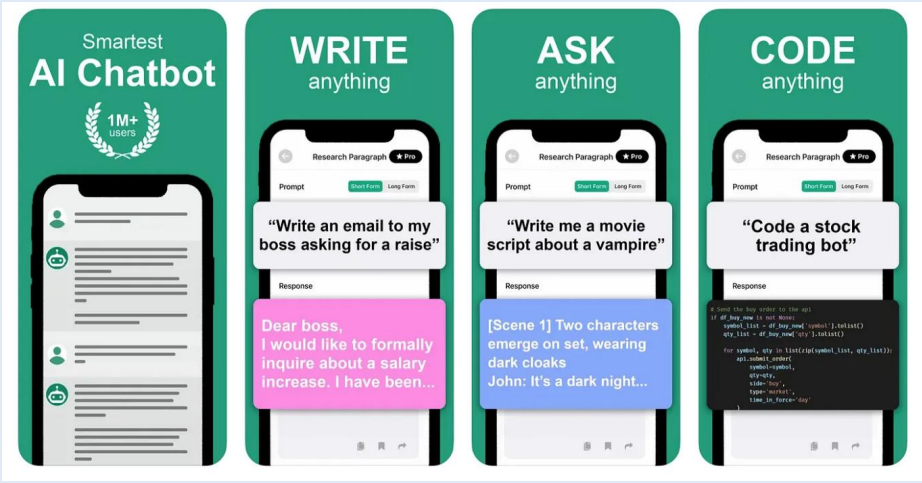
大模型下新场景-私有知识库

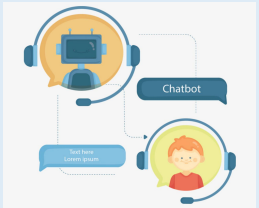


大语言模型引爆向量数据库

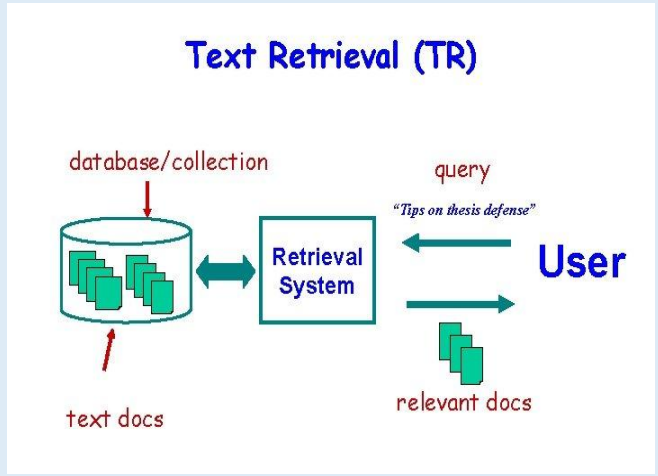
新的生产力
大语言模型

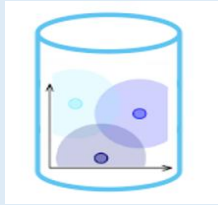
新的应用形态
对话式的交互



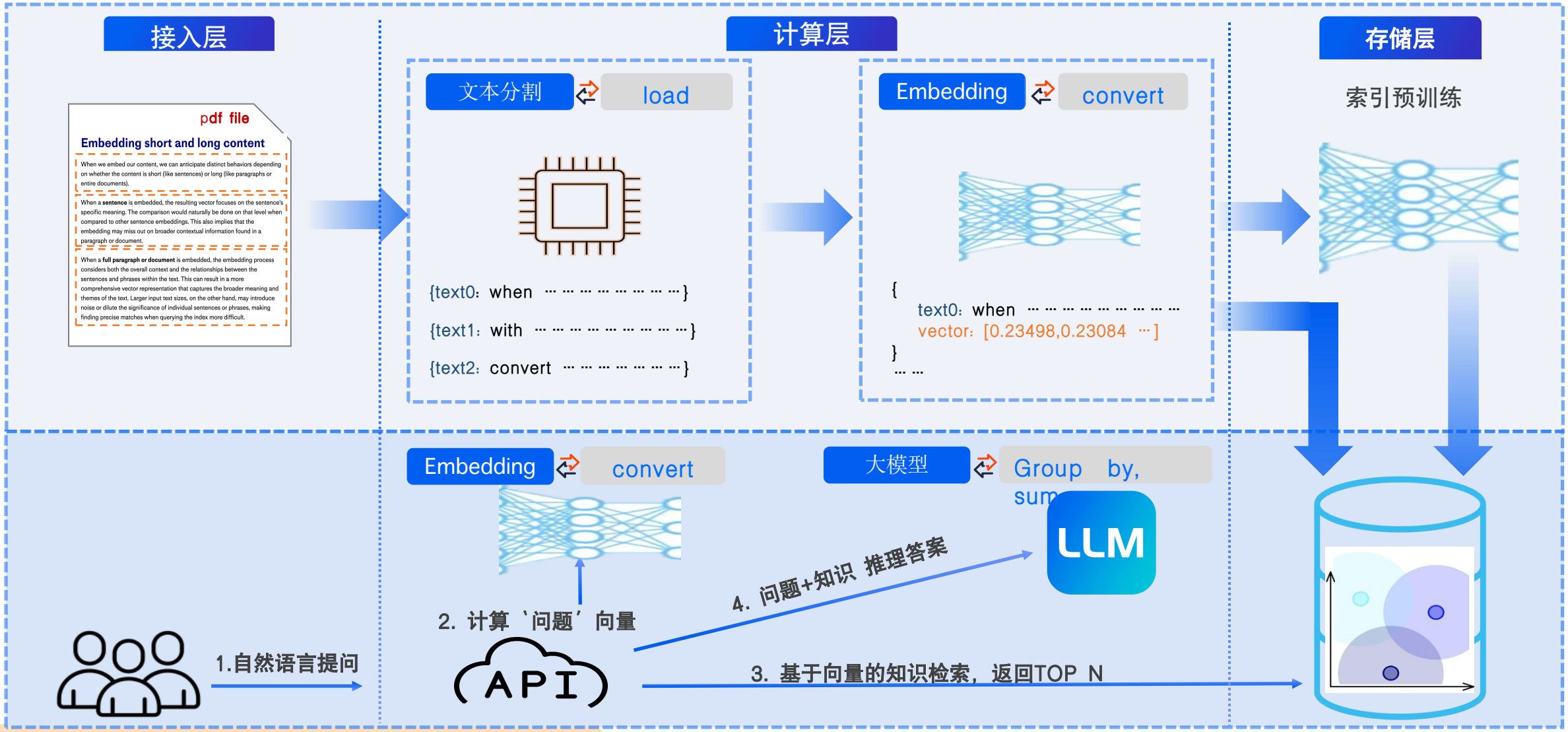


新的检索需求
向量数据库



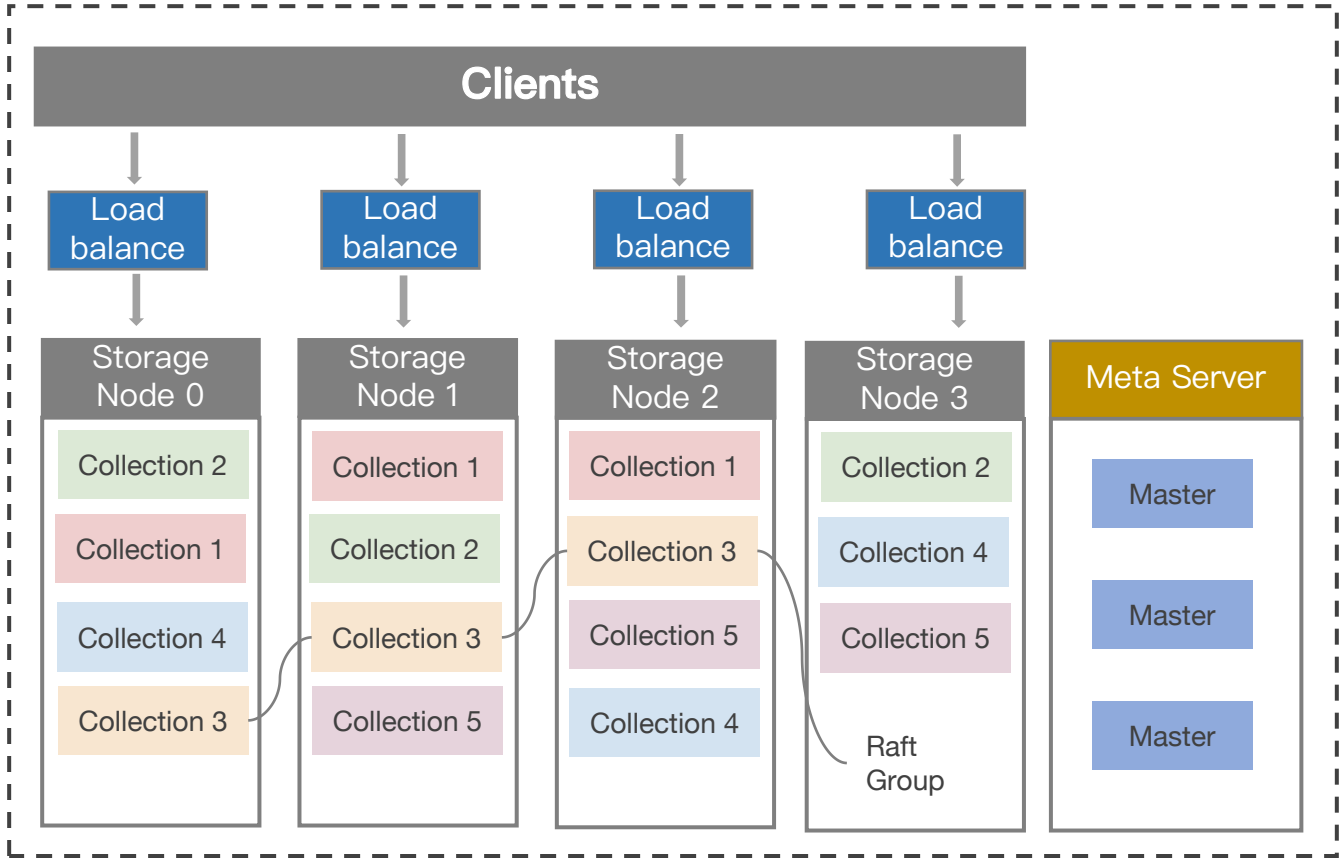


向量数据库AI Native时代来临

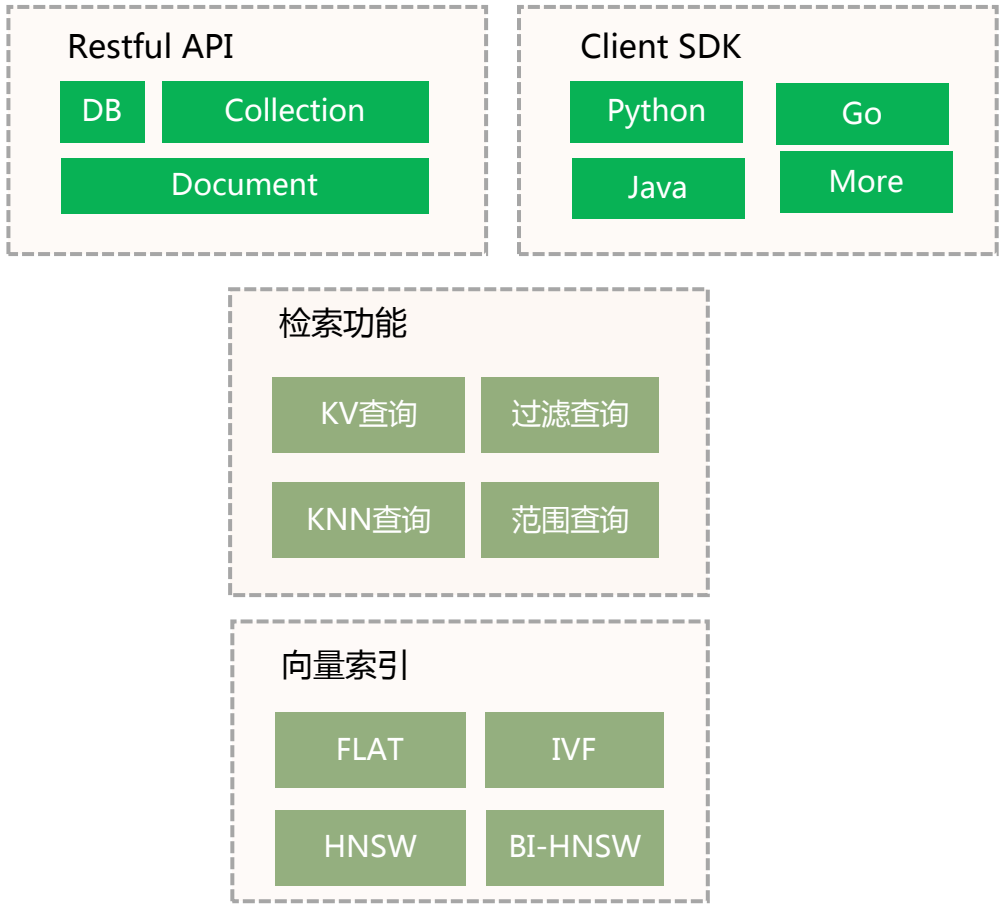


腾讯云向量数据库-极致性能-无界连接

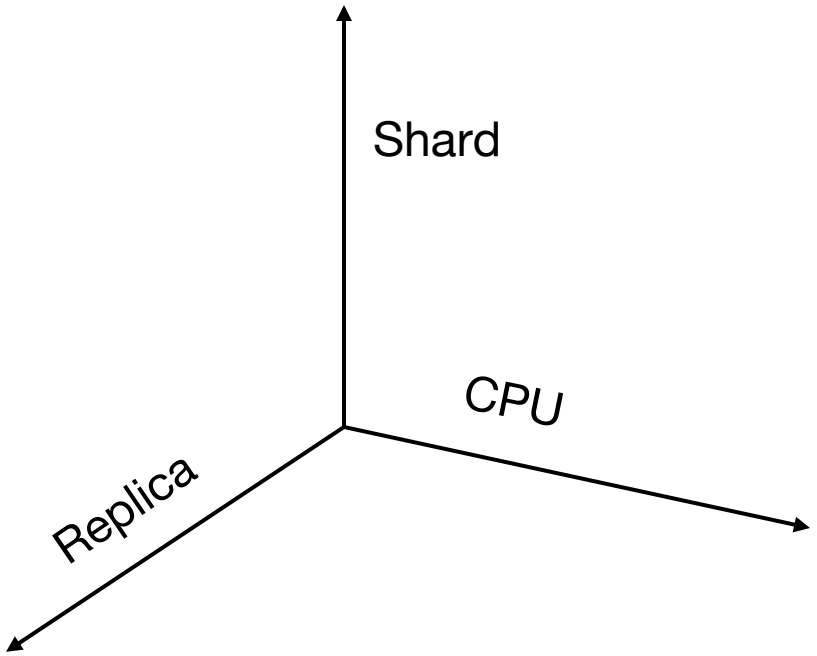
Multiple Raft-挖掘极致性能



无界连接，畅通无阻



腾讯云向量数据库-三维升级-性能飞跃



Scale UP

CPU

Memory

DIsk

提升单节点性能和容量，提升内存和磁盘利用率

Scale Out

Replica

无限提升查询性能

Shard

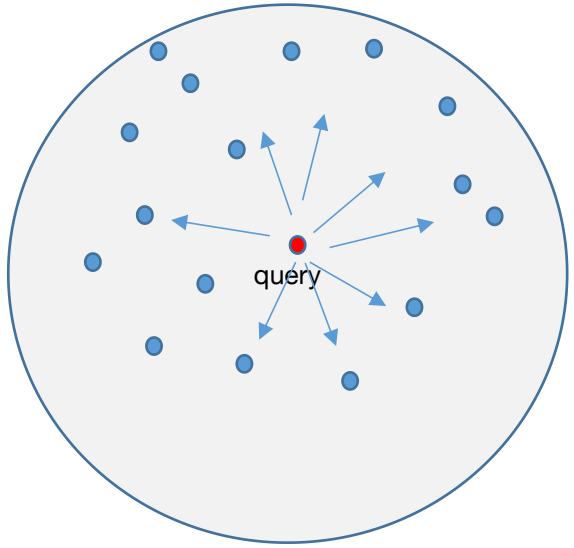
DIsk

无限提升容量上限

- 高效的资源利用
- 无限的性能提升
- 无限的容量存储

腾讯云向量数据库-算法世界,无限可能

Flat-暴力搜索

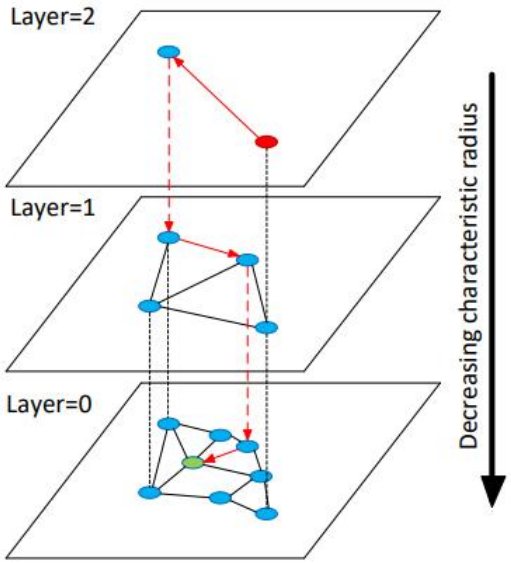


优点
- 100%召回率

缺点
- 性能低

适用
- 适用于少量数据集

hnsw

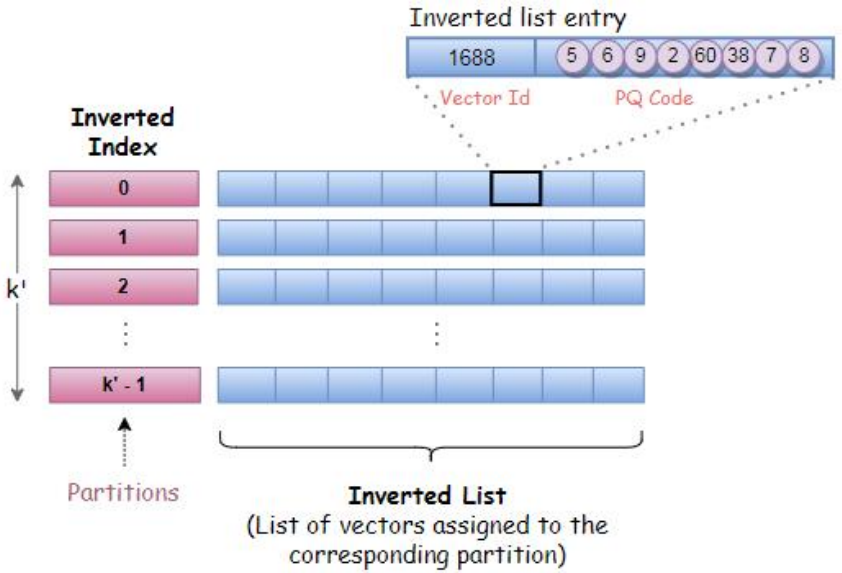


优点
- 查询性能高
- 召回率高

缺点
- 内存大

适用
- 适用百万级别数据量

IVFPQ



优点
- 查询性能高
- 容量高

缺点
- 召回率参数选择难

适用
- 适用亿万级别数据量

腾讯云向量数据库-大规模、高吞吐、低延迟

0.234,0.756,0.45083,0.8394,0.98534,0.98345745,0.723

0.5687,0.45083,0.872394,0.9834534,0.98345745,0.798

0.45083,0.872394,0.9834534,0.98345745,0.723,0.3458

0.684,0.756,0.43,0.872394,0.9834534,0.9834545,0.529

0.524,0.756,0.45083,0.8794,0.94534,0.98345745,0.723

0.234,0.45083,0.872394,0.9834534,0.98345745,0.1879

0.864,0.756,0.45083,0.872394,0.9834534,0.983457453

0.834,0.756,0.45083,0.872394,0.934,0.98345745,0.023

0.084,0.756,0.45083,0.872394,0.4534,0.985745,0.5353

0.344,0.756,0.45083,0.872394,0.9834534,0.983743732

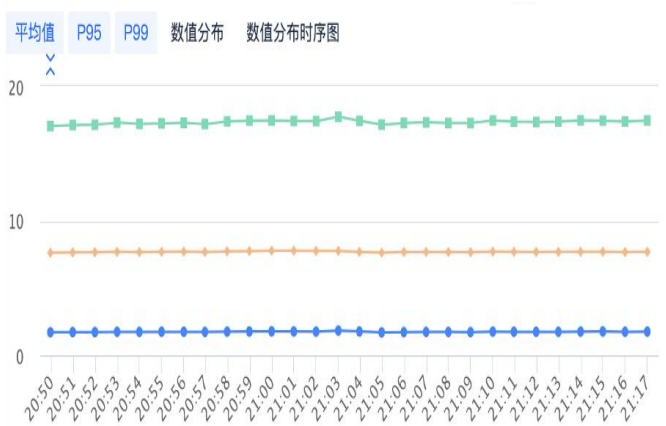
0.972,0.756,0.45083,0.872394,0.34534,0.345745,0.423

10亿

单索引行数

100万

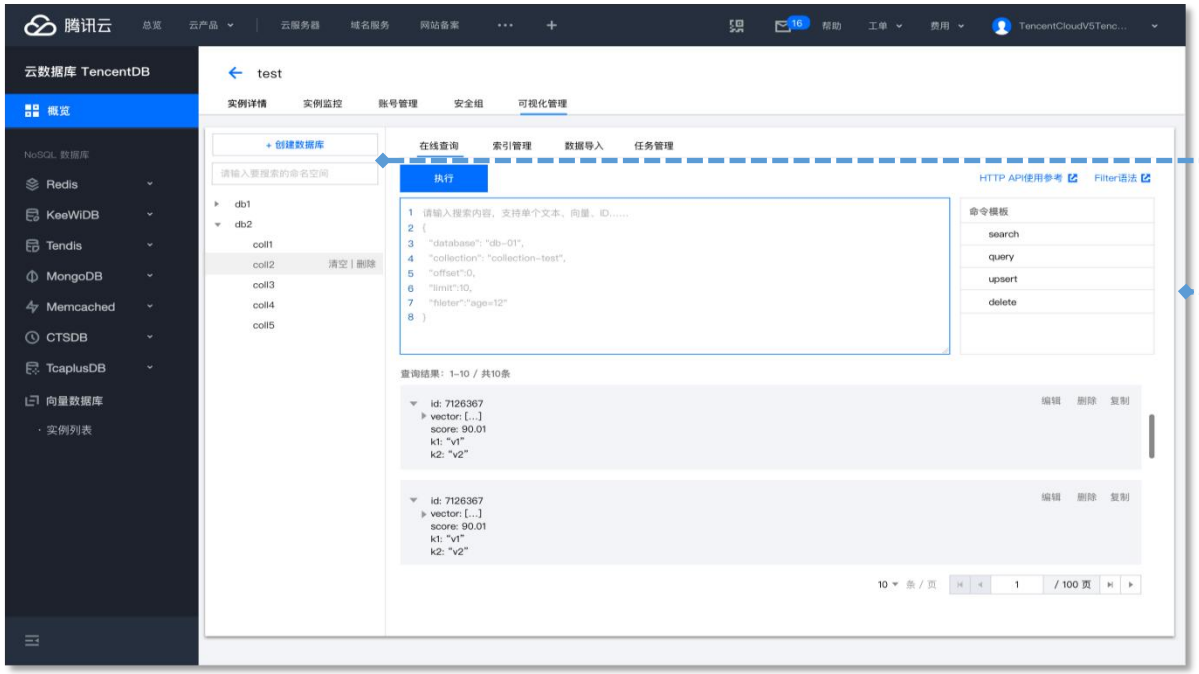
单实例QPS



20MS

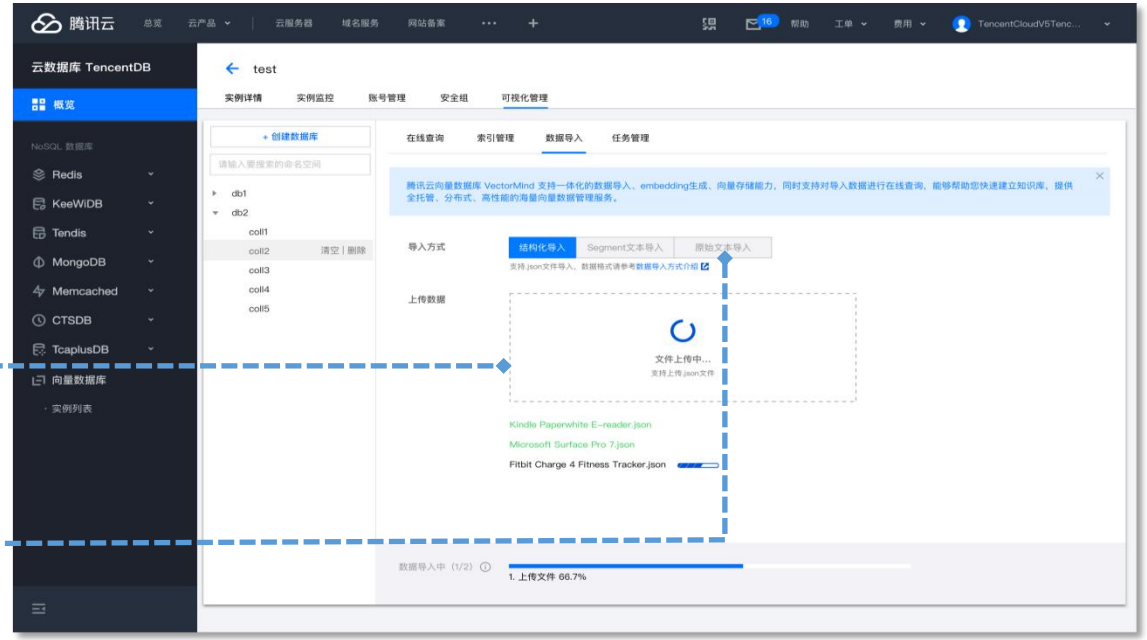
P99响应延迟

腾讯云向量数据库-简单易用的可视化数据管理



DB/Collection**在线管理**

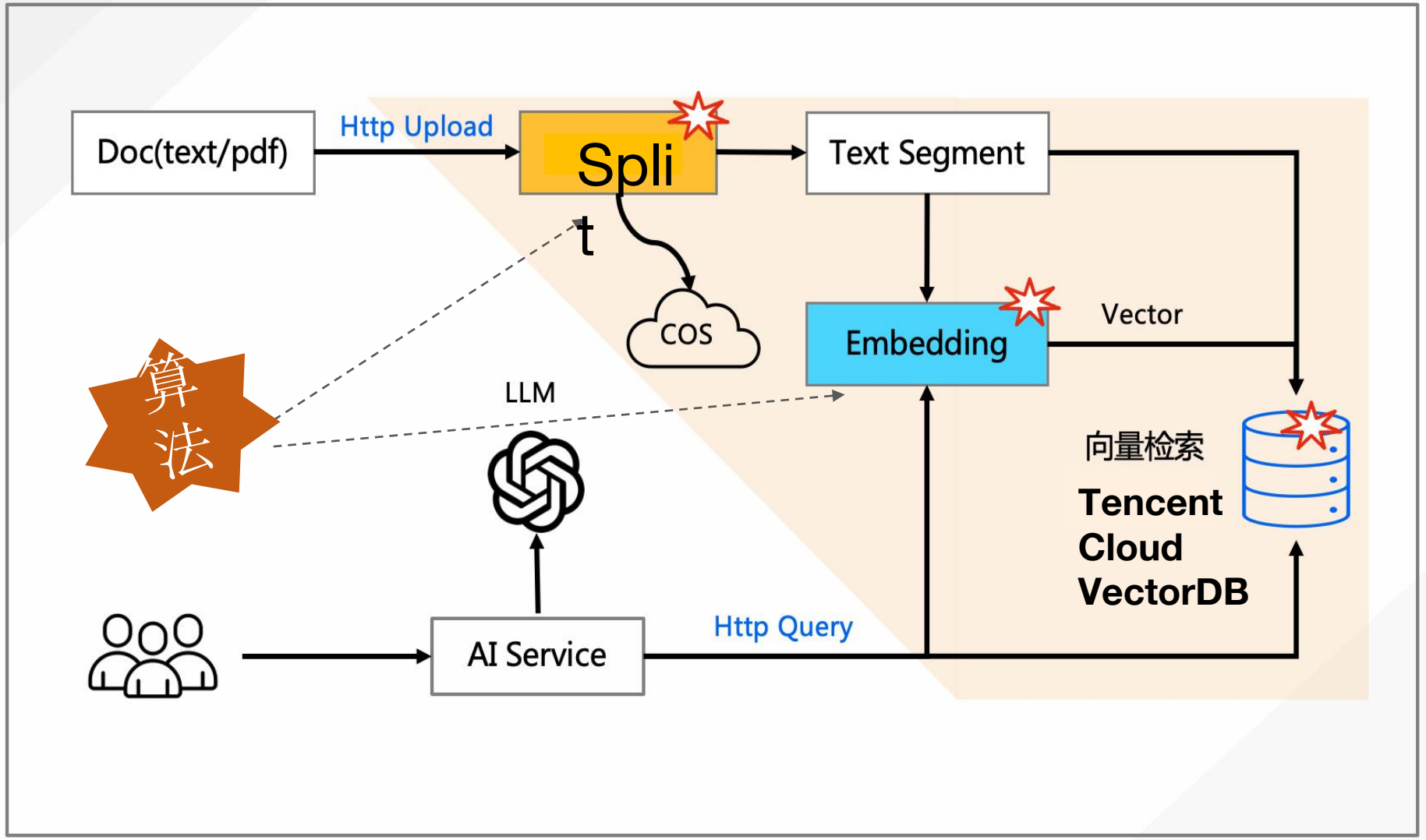
常用命令模板化
快速执行向量检索操作



一键上传数据
自动构建索引

支持文本导入与查询
实现**端到端检索能力**

腾讯云向量数据库-一站式AI Native向量检索方案



一站式方案：

- 源自腾讯内部积累；
- 简化开发流程；
- 降低业务接入门槛；
- 提升业务接入效率；
- 降低算法工程投入；

向量数据库面临的挑战


挑战
成本
为了性能，向量索引基本都在内存
高维向量存储空间大，压缩比低
Embedding依赖GPU，成本高
性能
向量检索性能相对传统索引性能偏低
Embedding性能不高
数据实时处理能力不高
高性能存储设备+DiskANN
高性能的Embedding模型
低成本AI硬件
AI加持Ann算法
More?
算法
算法选择门槛高
算法边界突破难

THANKS