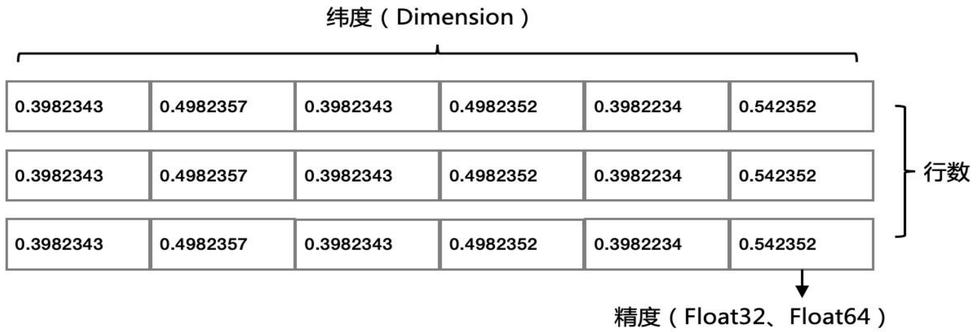
大模型时代下的向量数据库创新与挑战

腾讯云专家工程师 / 伍旭飞

什么是向量检索

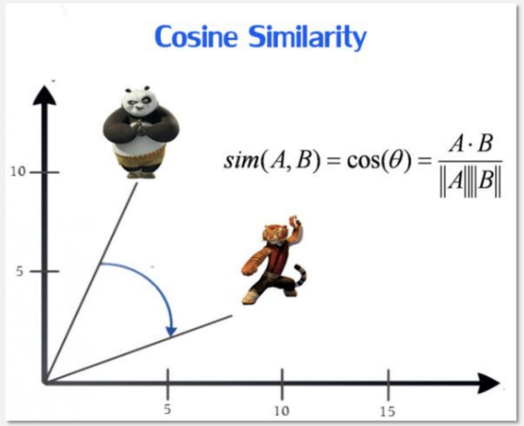
向量概念

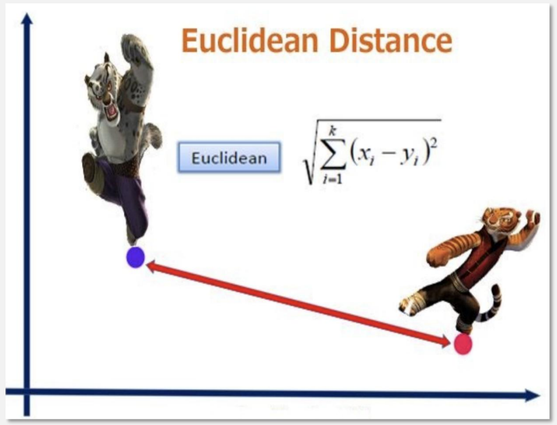


AI 中的特征向量

向量检索

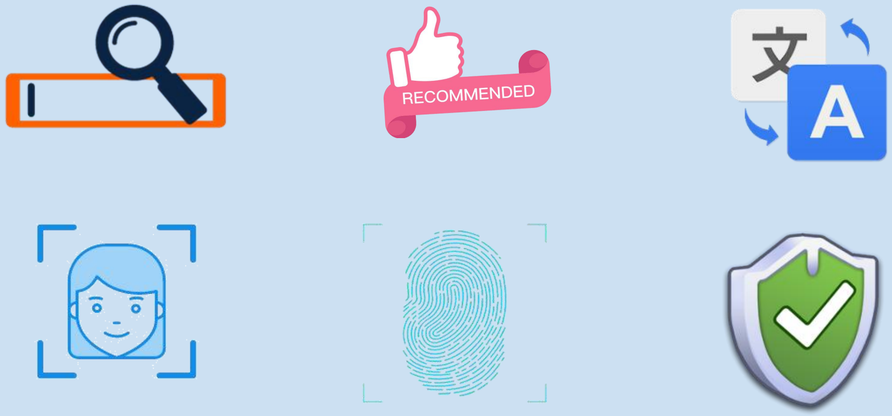
向量检索又称为近似最近邻搜索（Approximate Nearest Neighbor Search, ANNS），是一种在大规模高维向量数据中寻找与给定查询向量相似的向量的技术。向量检索在许多AI领域具有广泛的应用，如图像检索、文本检索、语音识别、推荐系统等。





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



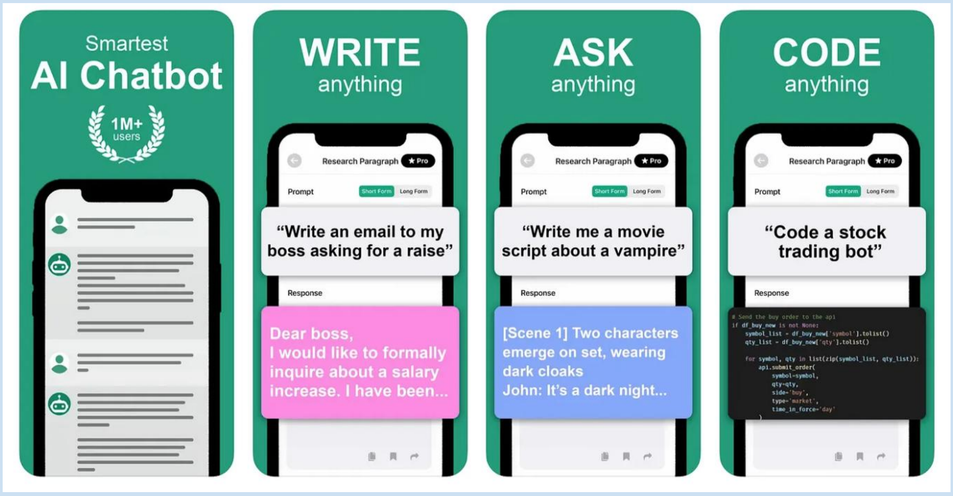
腾讯OLAMA向量检索引擎，在腾讯集团大规模应用

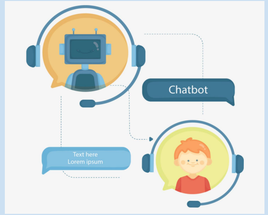
技术起源&发展历程

大语言模型引爆向量数据库

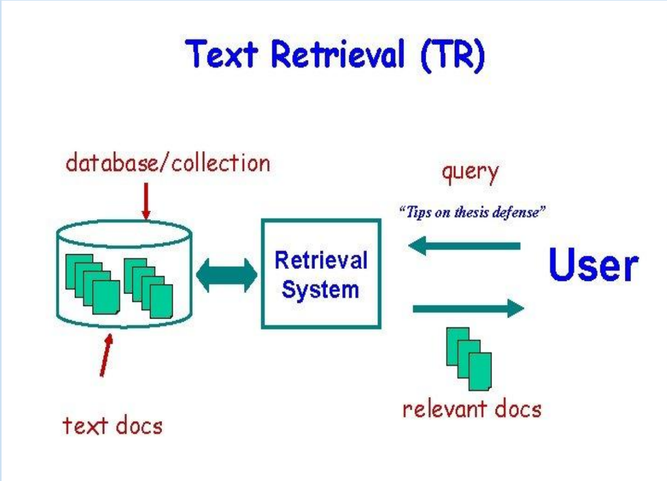
新的生产力
大语言模型

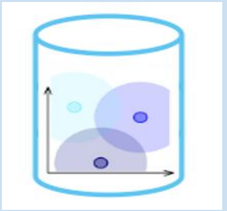
新的应用形态
对话式的交互



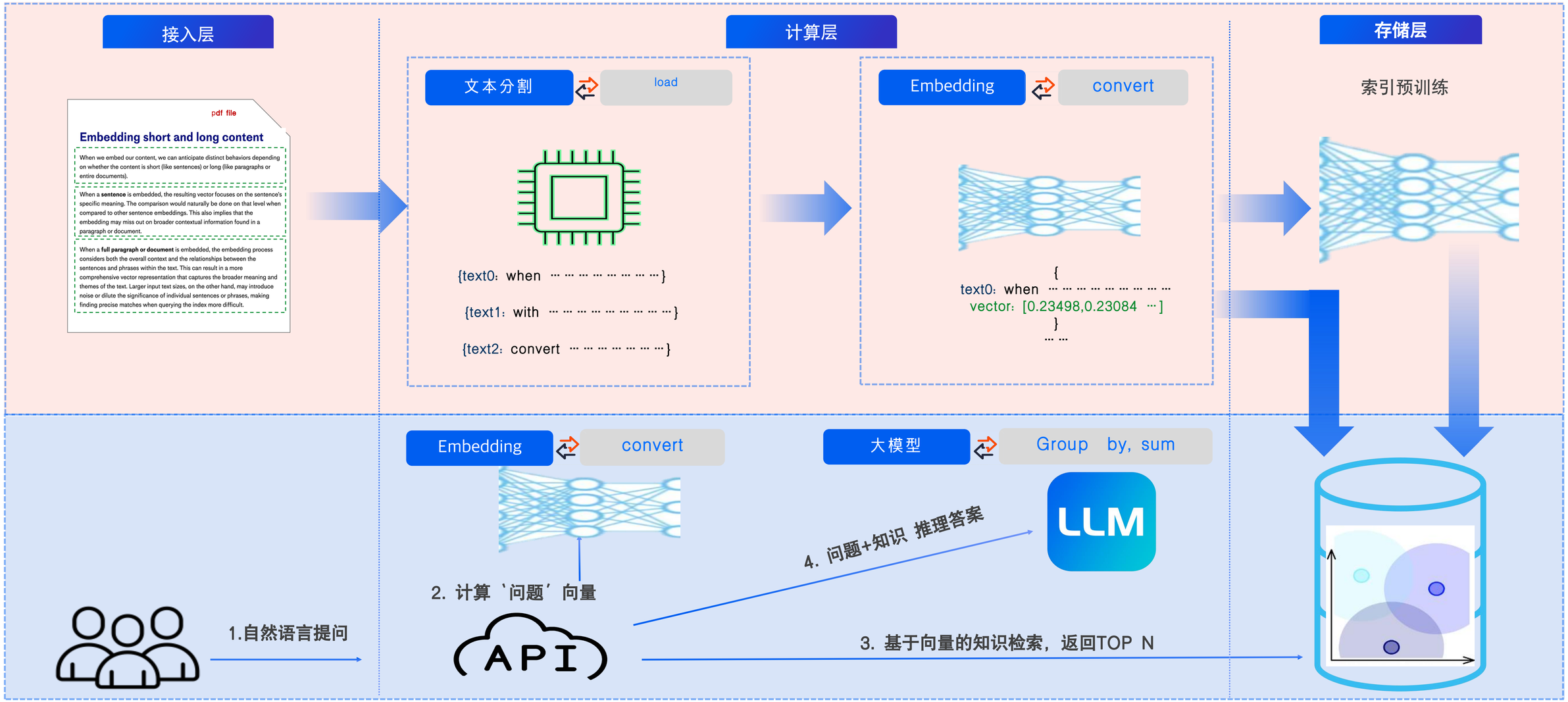


新的检索需求
向量数据库



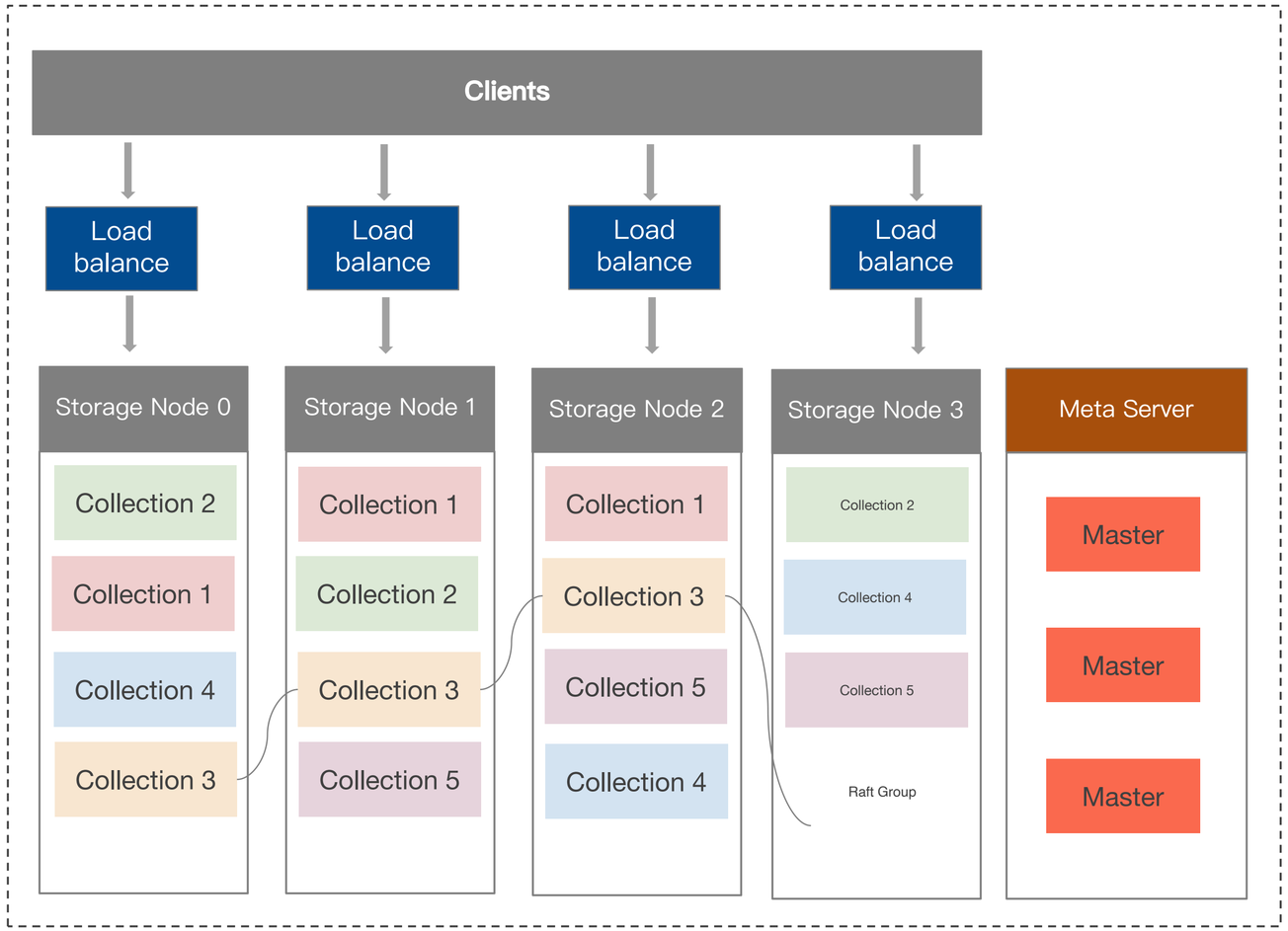


向量数据库AI Native时代来临

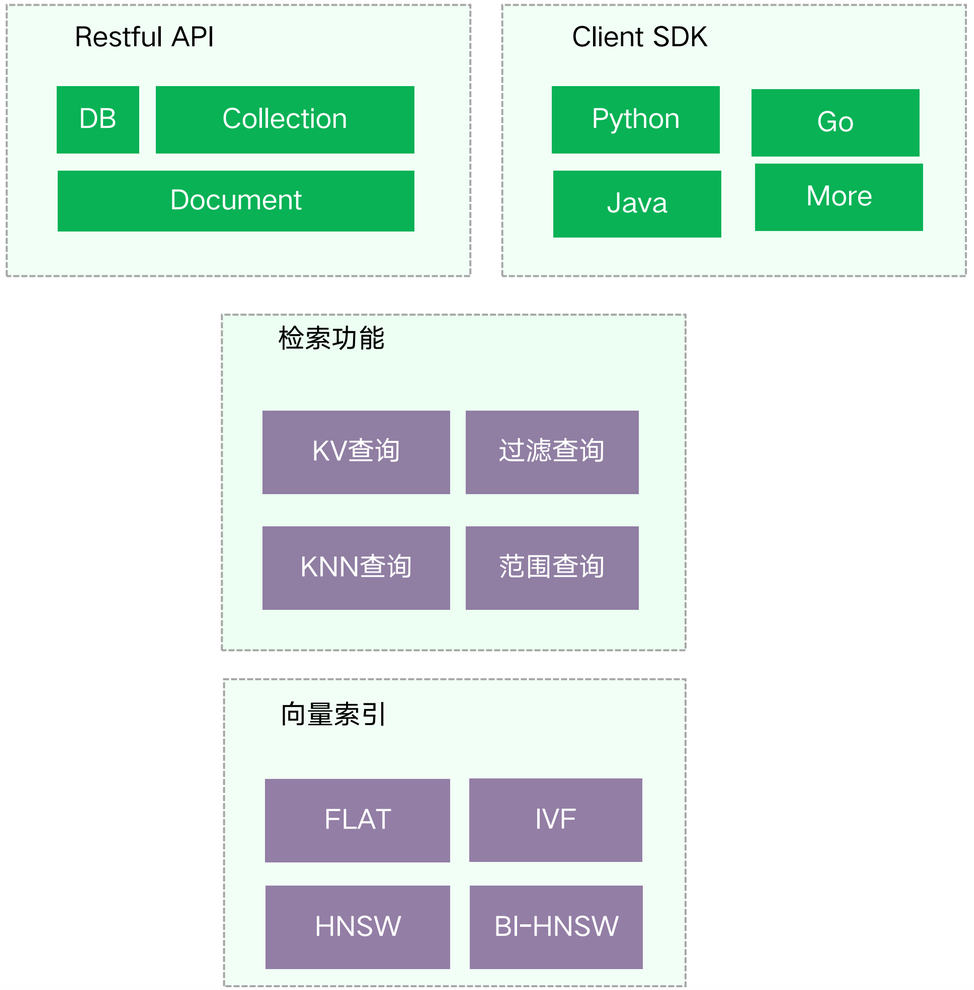


腾讯云向量数据库-极致性能-无界连接

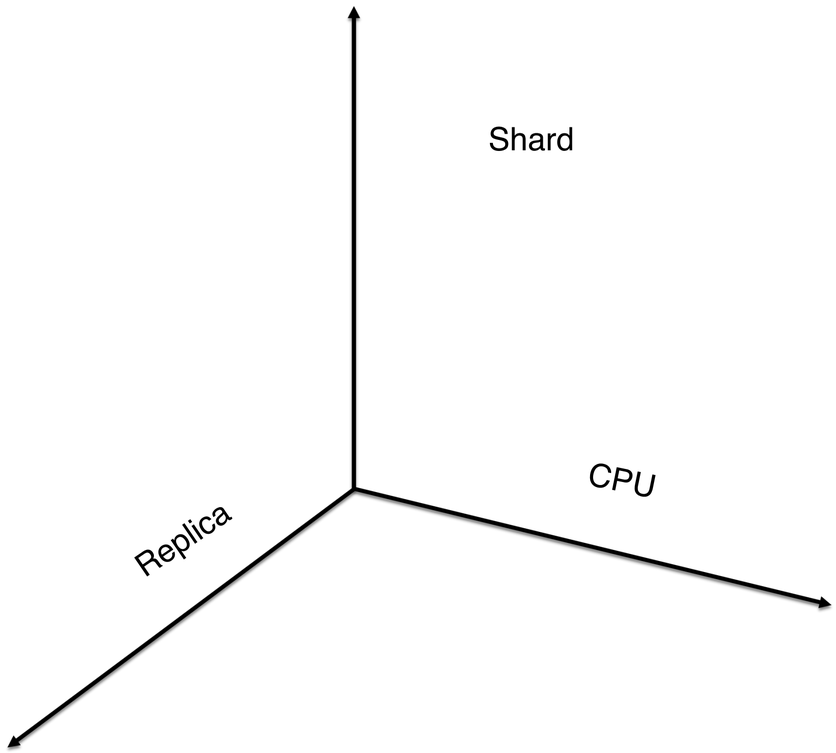
Multiple Raft-挖掘极致性能



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

无限提升查询性能

Replica

Shard

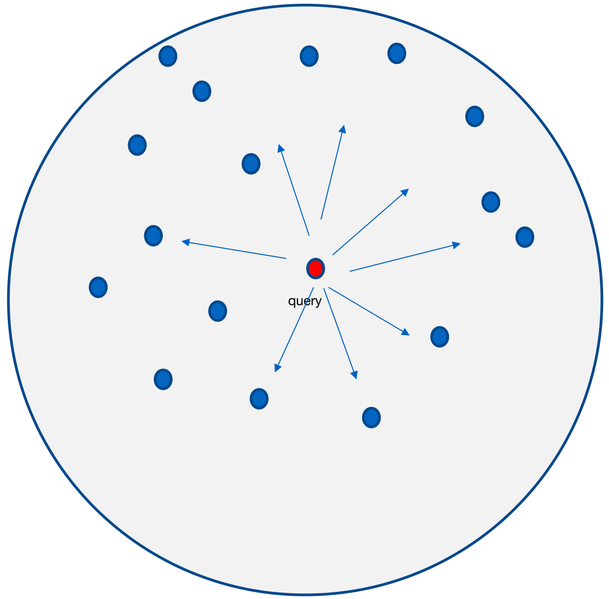
无限提升容量上限

DIsk

- 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-传统算法

Flat-暴力搜索

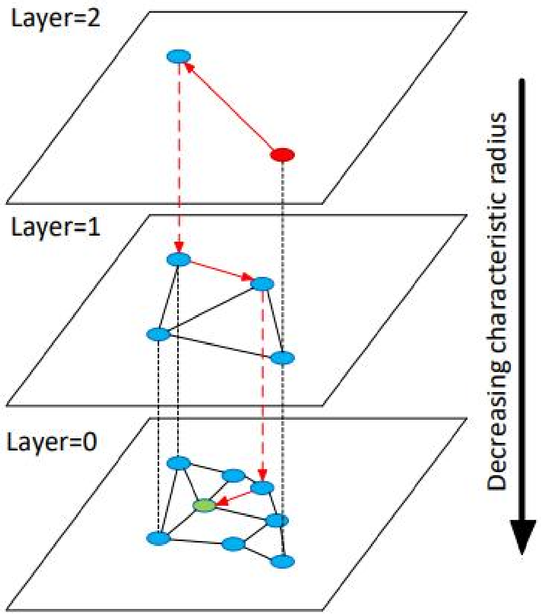


优点
- 100%召回率

缺点
- 性能低

适用
- 适用于少量数据集

hnsw

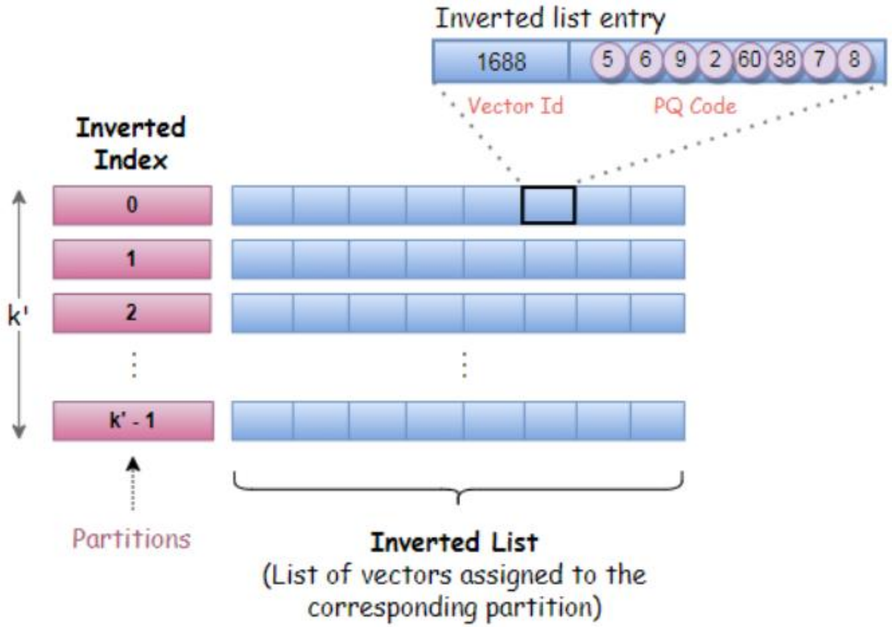


优点
- 查询性能高
- 召回率高

缺点
- 内存大

适用
- 适用百万级别数据量

IVFPQ



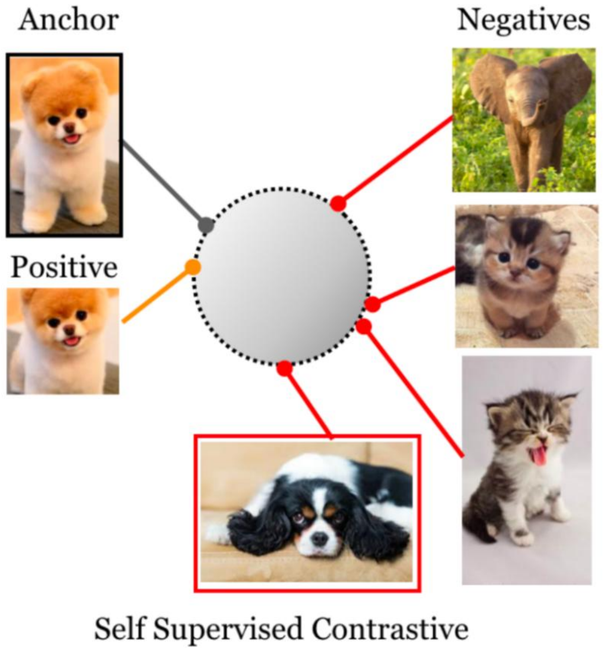
优点
- 查询性能高
- 容量高

缺点
- 召回率参数选择难

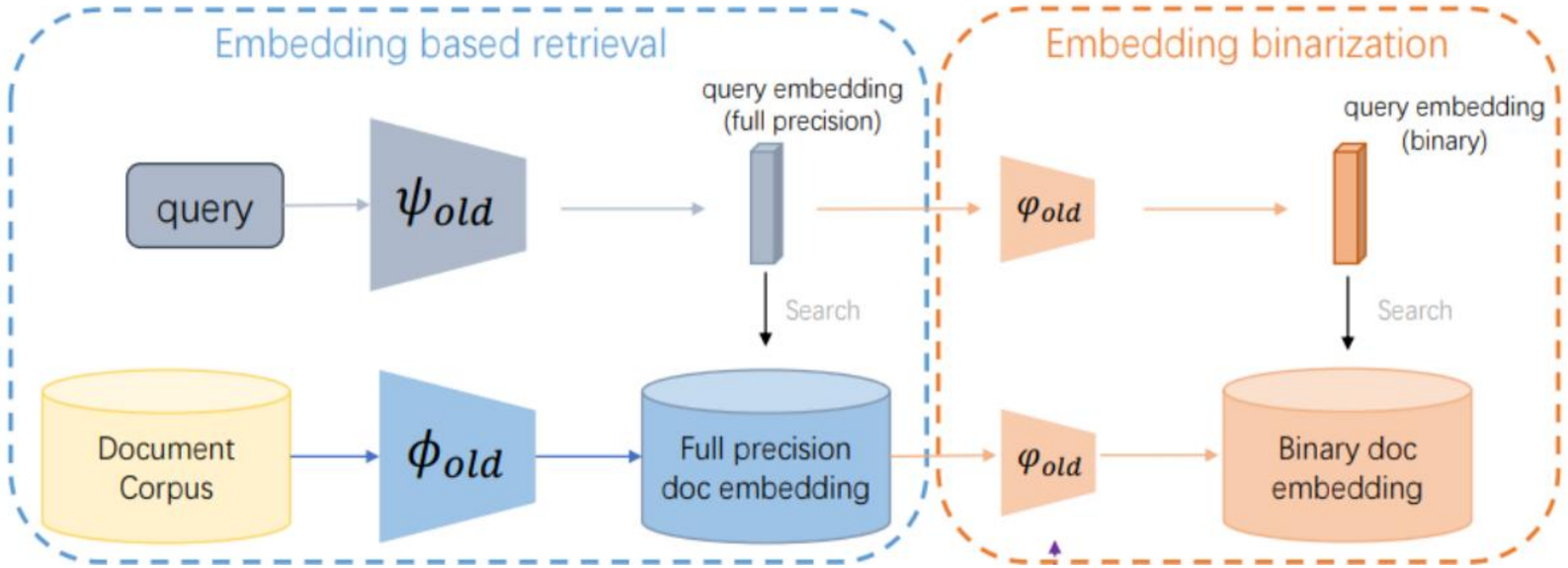
适用
- 适用亿万级别数据量

腾讯云向量数据库-AI加持-Contrastive Learning

Contrastive Learning

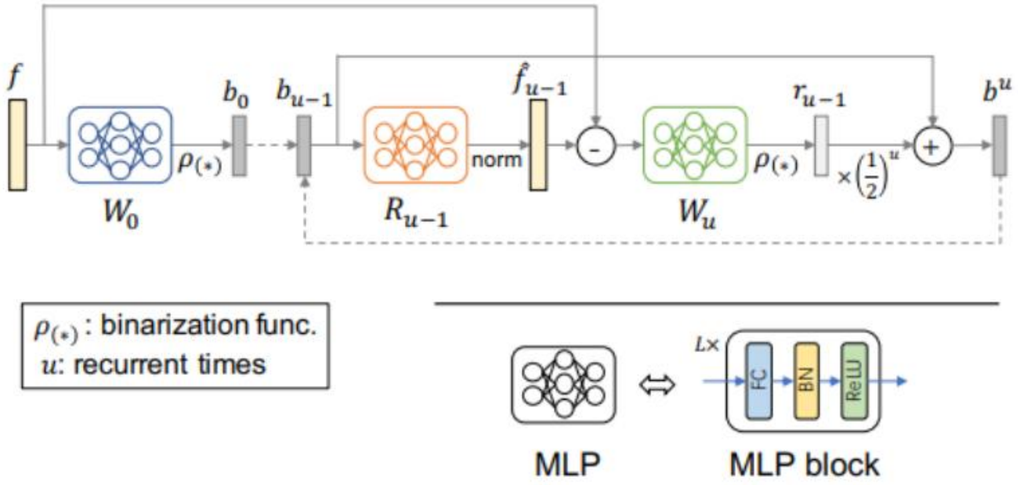


Binary vector

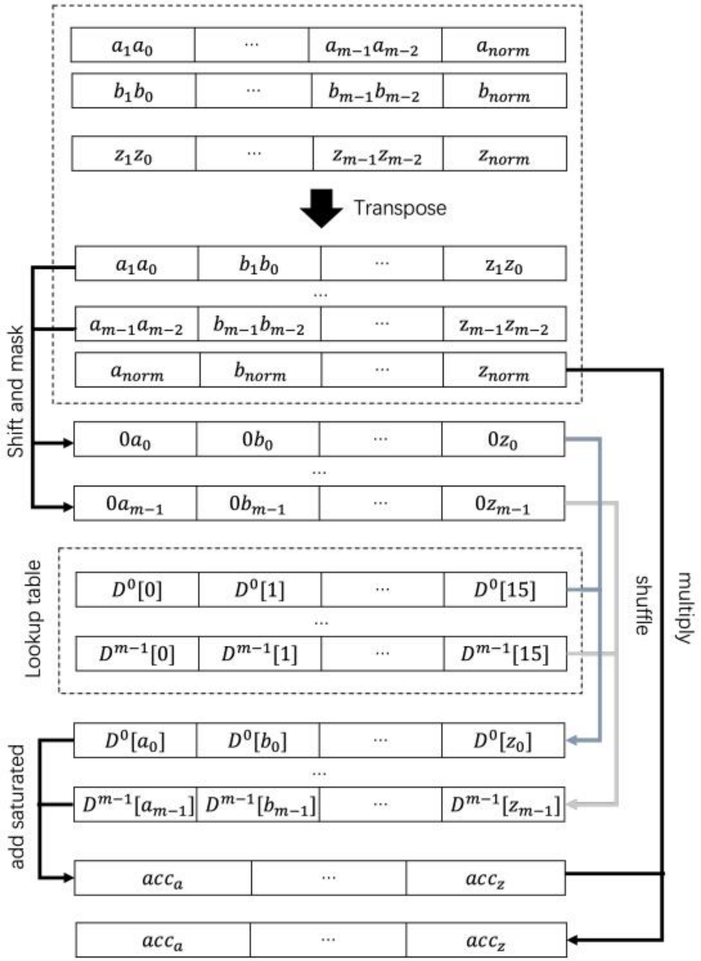


腾讯云向量数据库-二值化训练-BI-HNSW

Binary Learn Network

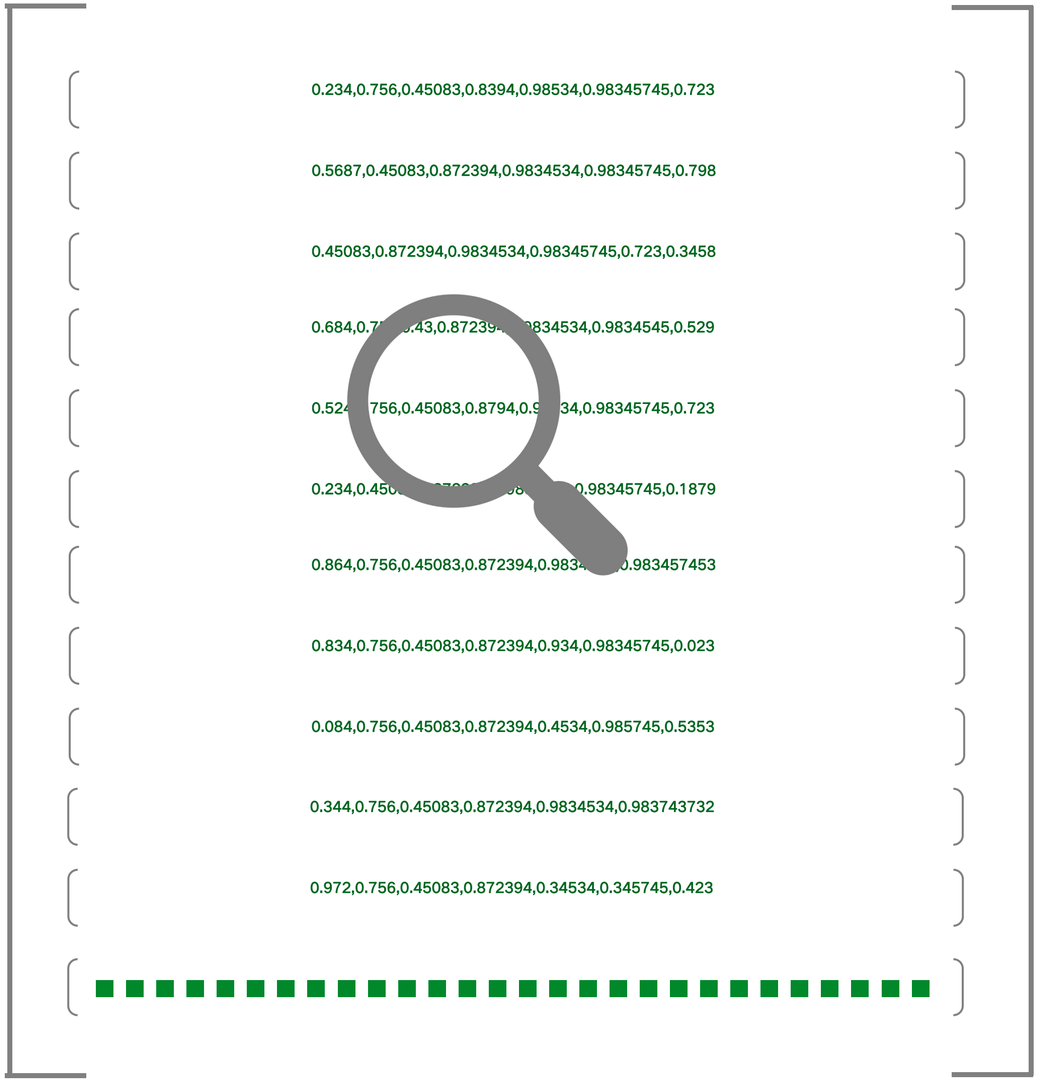


Symmetric distance calculation SIMD



- 同召回率内存节省30%~50%

腾讯云向量数据库-三维升级-性能飞跃

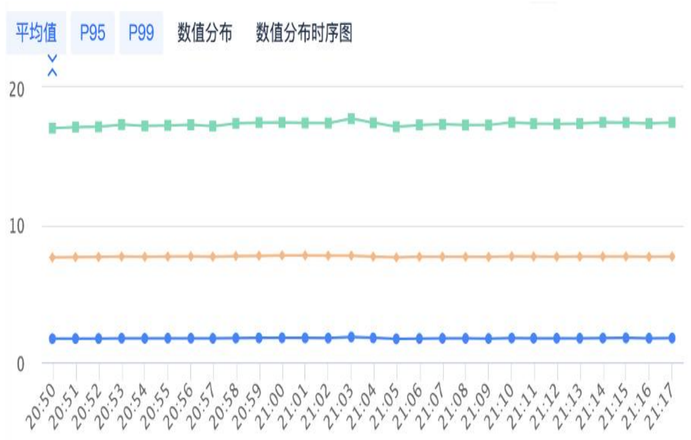


10亿

单索引行数

100万

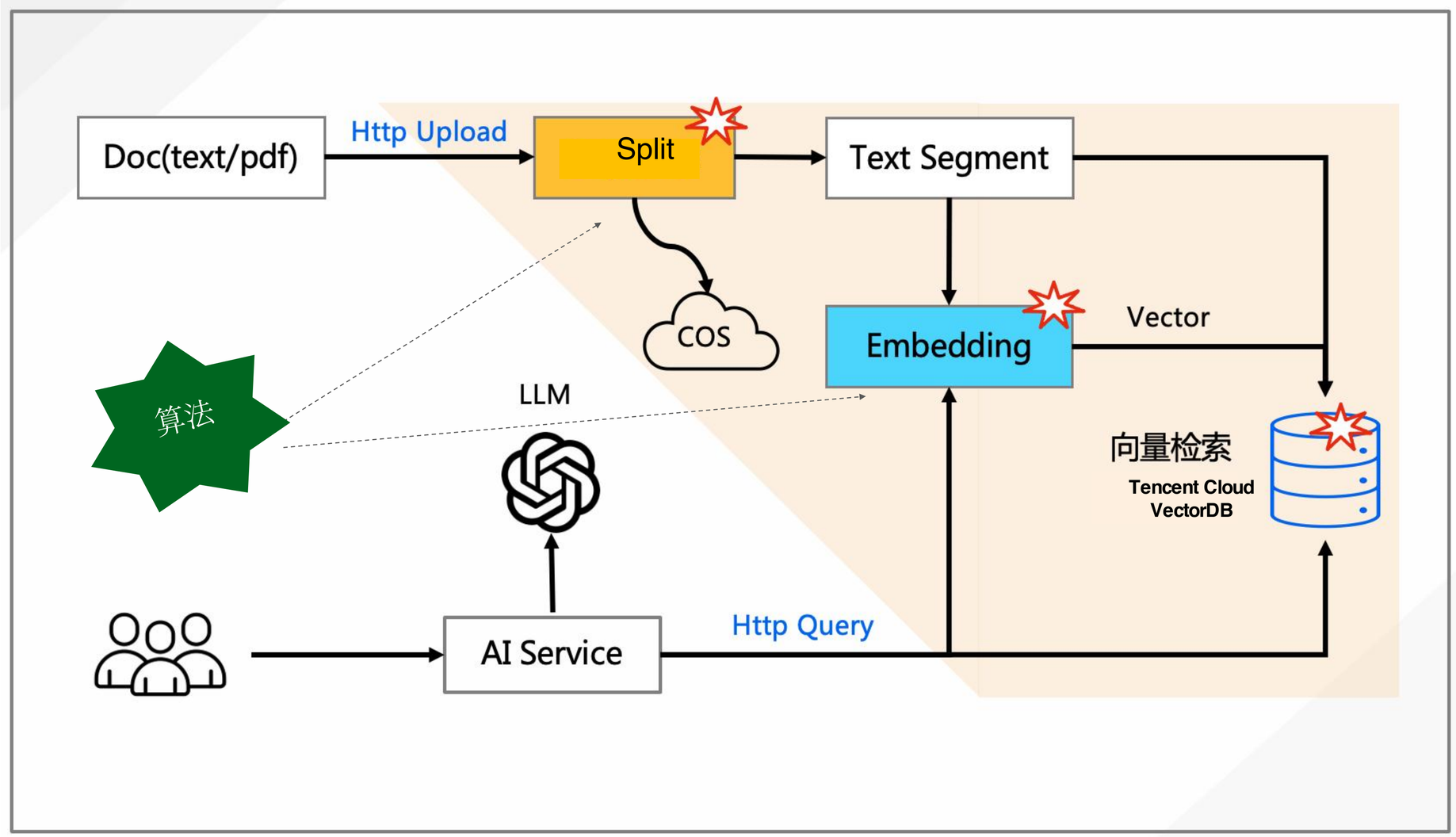
单实例QPS



20MS

P99响应延迟

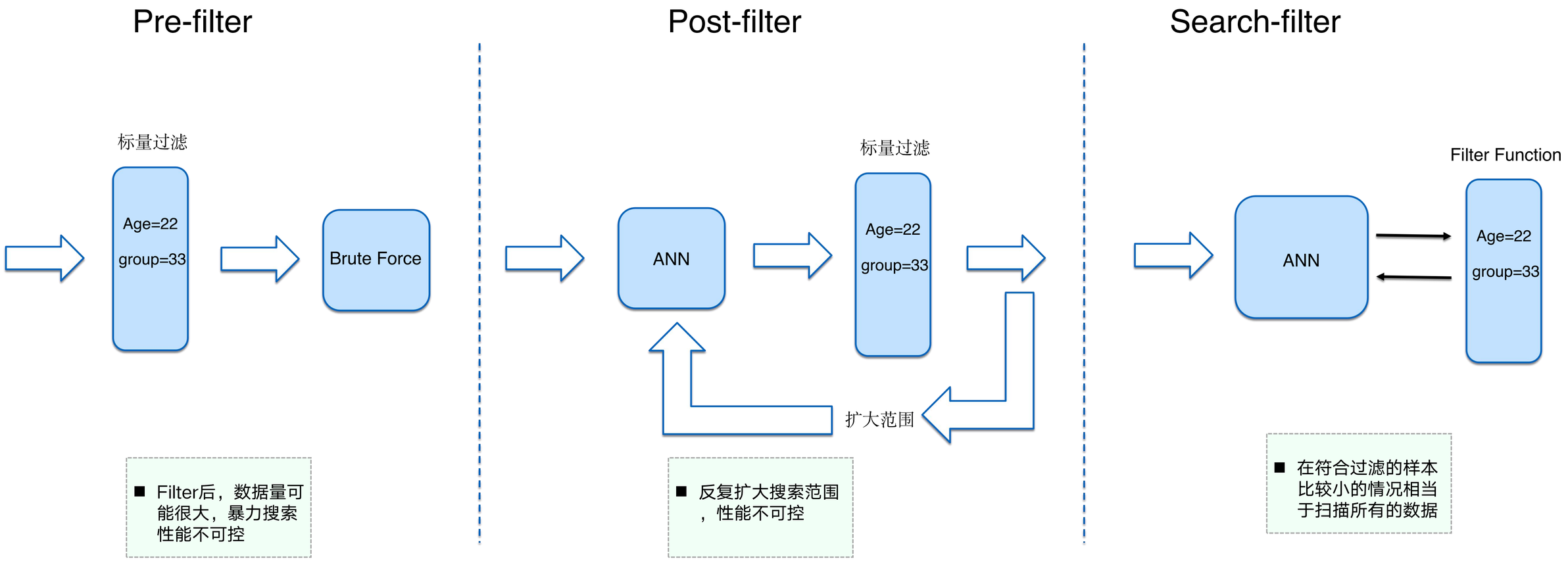
腾讯云向量数据库-一站式AI Native向量检索方案



一站式方案：

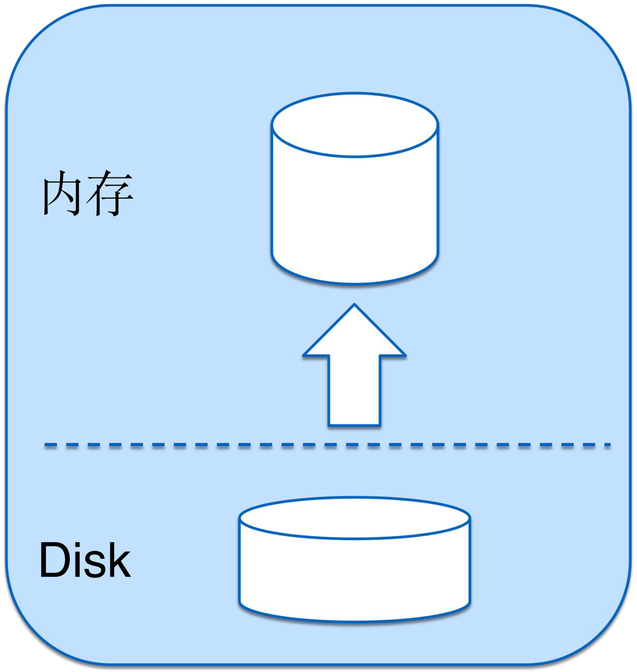
- 源自腾讯内部积累；
- 简化开发流程；
- 降低业务接入门槛；
- 提升业务接入效率；
- 降低算法工程投入；

向量数据库挑战-Filter效率



向量数据库挑战-成本/性能

索引全内存



- 高效的资源利用
- 无限的性能提升
- 无限的容量存储

算法

- Disk ANN
- BI-HNSW

硬件/架构

- RDMA
- GPU
- 基于新硬件架构

THANKS

软件正在重新定义世界

Software Is Redefining The World