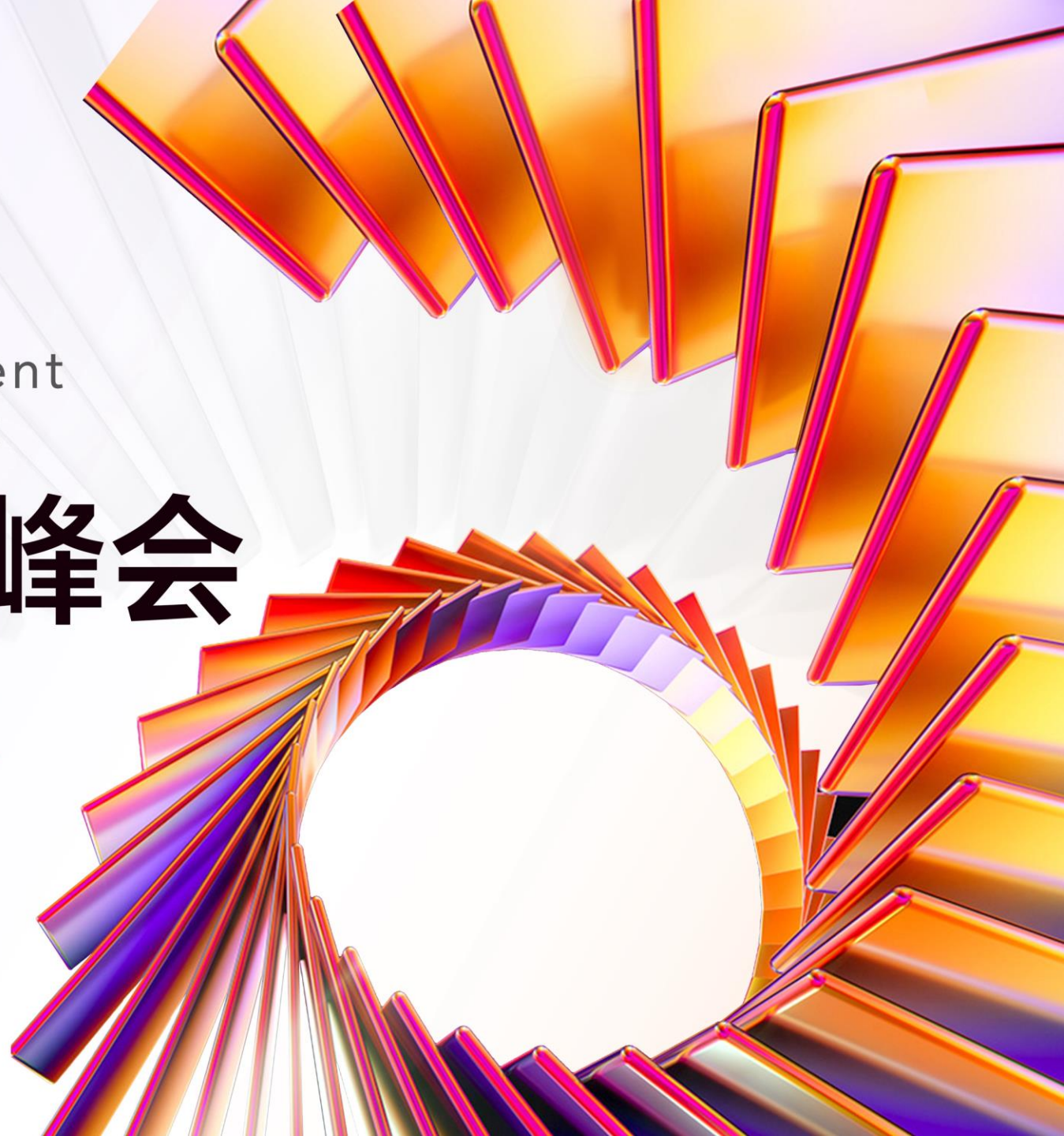
AiDD 7th 2025

第7届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

8月8-9日 | 北京站



AiDD 8th 2025

第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

查看会议详情

下一站预告 11/14-15 | 深圳站 12/19-20 | 上海站

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

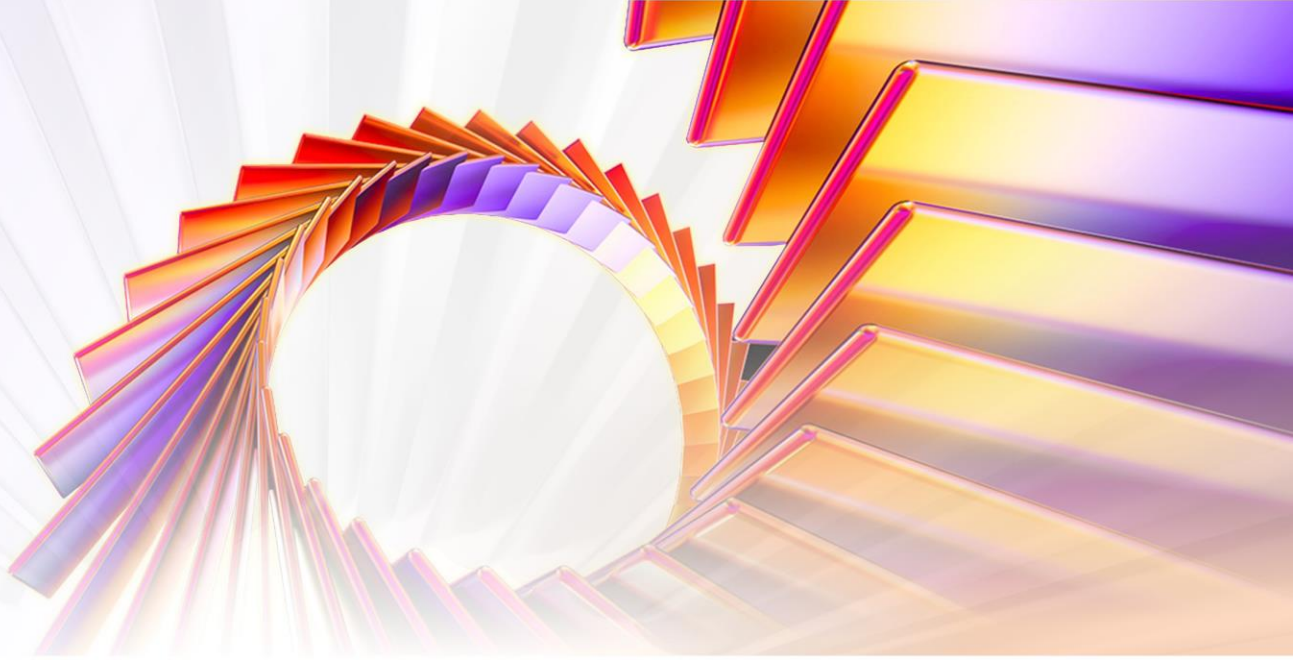
上下文工程

AiDD 7th 2025 | 8月8-9日 | 北京站

第7届 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



蚂蚁开源向量检索库 VSAG 与业务实践

王翔宇 | 蚂蚁集团

王翔宇

蚂蚁集团全模态检索部技术专家

- 开源向量数据库 Milvus 核心开发者，BigANN 21 Track 2 第一名团队成员。曾在 Zilliz 负责存储和 GPU 算法相关开发工作。
- 2023 年加入蚂蚁集团，主要负责蚂蚁向量检索算法研发以及千亿规模向量数据库在蚂蚁业务场景落地。对向量检索算法与系统有丰富经验。

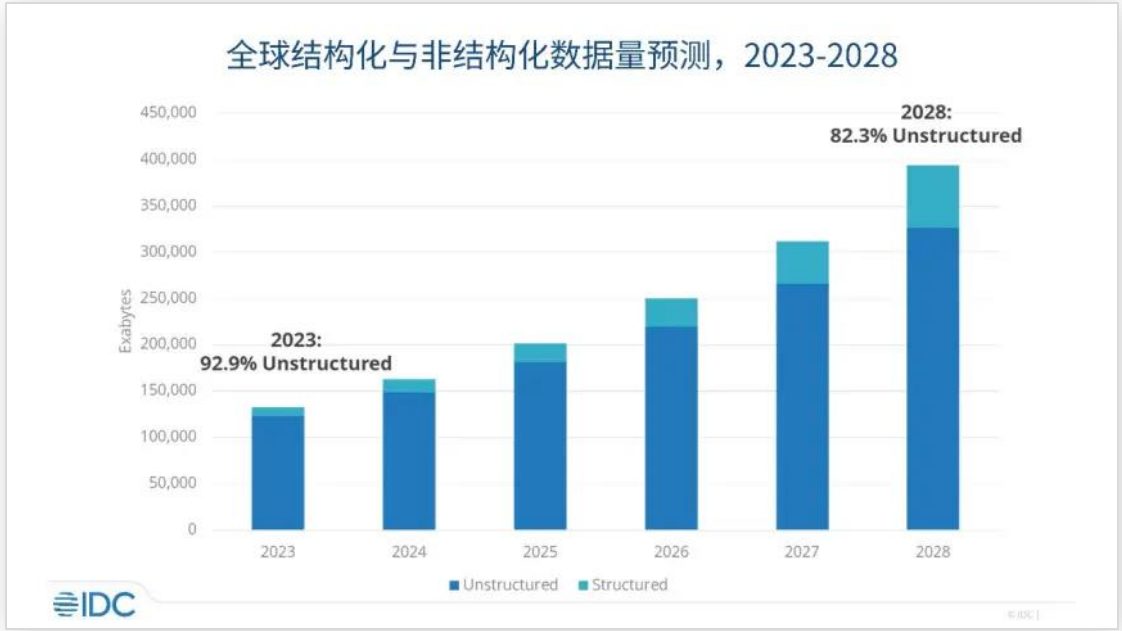
目录
CONTENTS

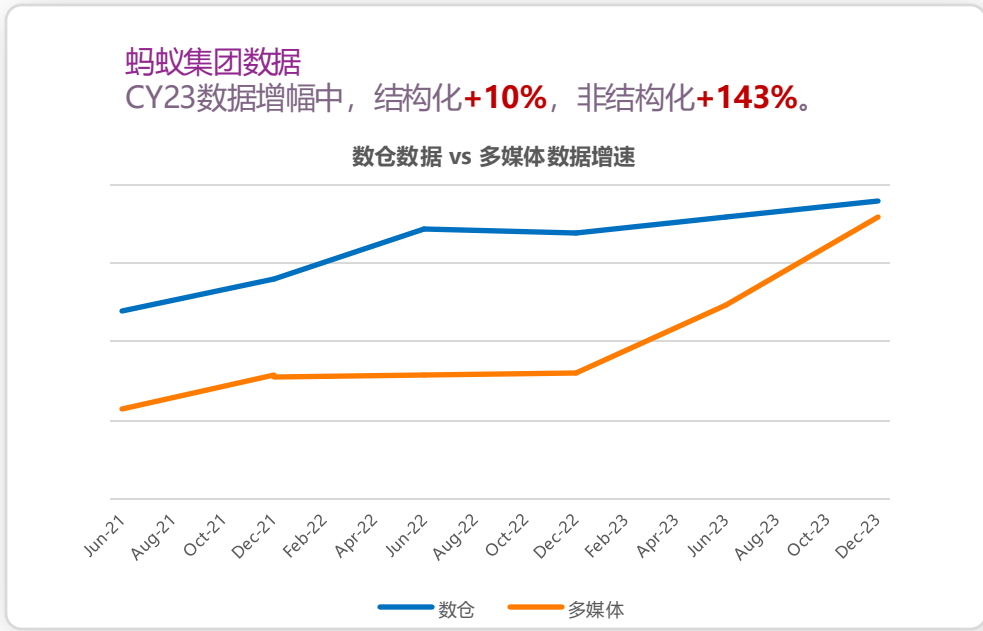
AiDD 7th 2025

PART 01

向量检索技术介绍

背景 – 持续增长的数据

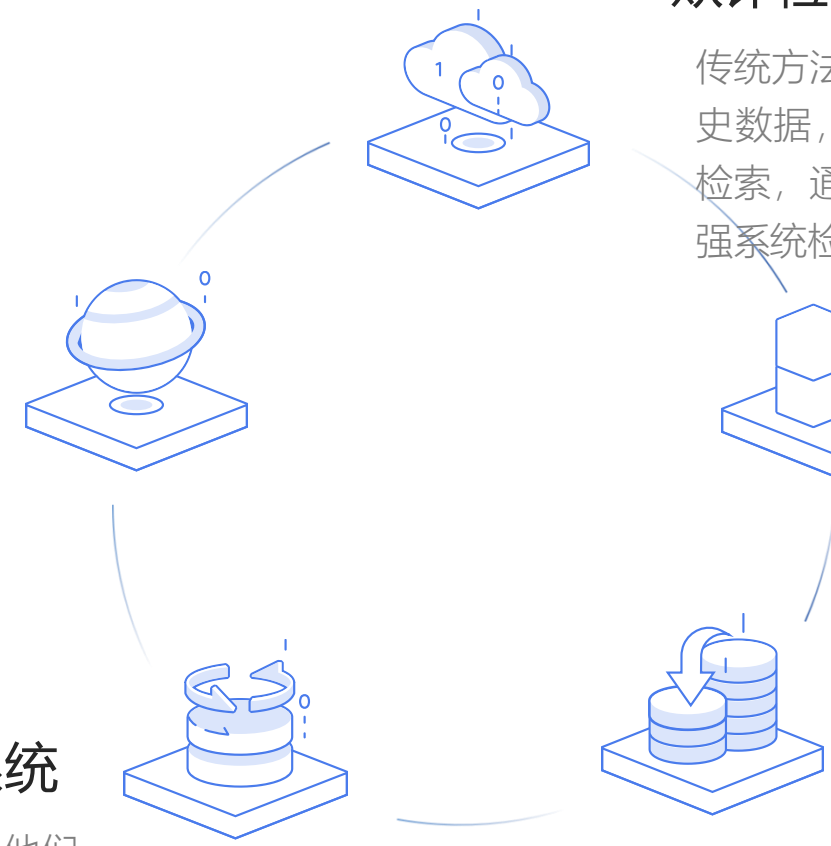




趋势：

- 企业中非结构话数据增长迅速，占比超过 80%；
- 蚂蚁集团 2023 年非结构化数据增速超过结构化数据

背景 – 越来越多的 RAG 应用



欺诈检测和风险评估

传统方法严重依赖于预定义的规则和历史数据，RAG 支持动态、上下文数据检索，通过使用最新的外部信息能够增强系统检测异常的能力。

客户支持机器人

当用户提出问题时，聊天机器人会从知识库、FAQ或客户记录等来源检索相关信息，并使用模型生成个性化响应。

电商产品推荐

RAG 通过实施了解客户需求来动态生成结果，相较于传统的方法，能产生更相关和准确的推荐，提高销售额。

个性化学习和辅导系统

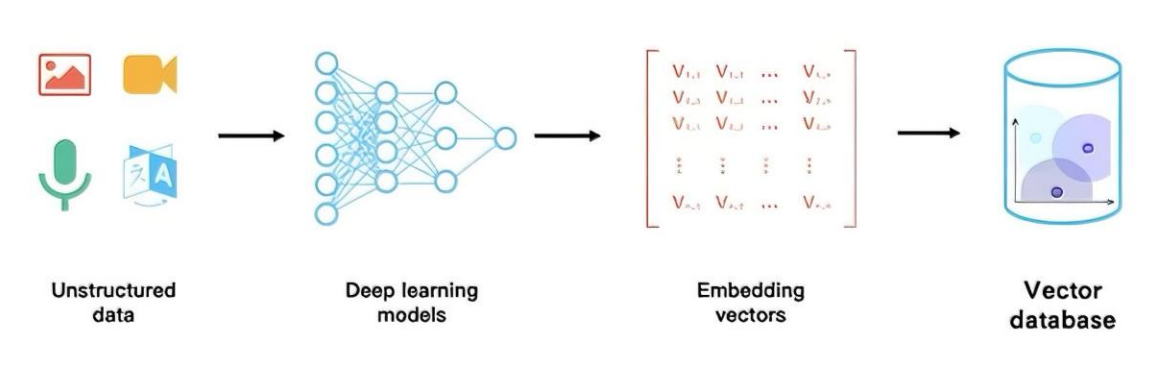
通过对话，LLM 帮助学习者了解他们的偏好和职业目标，提供更多相关的课程选择和个性化的学习计划，提升学习体验。

企业知识管理

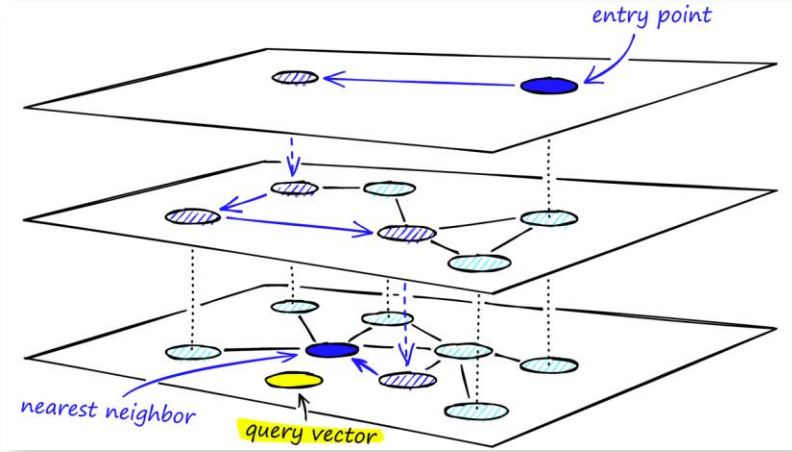
使用 RAG，企业可以为员工和客户提供即使、量身定制的查询答案，减少手动搜索的需要并提高效率。

背景 – 非结构化数据与向量检索

- 非结构化数据
  - 音频、视频、图片、文本
  - 数据规模大，信息密度高，处理成本高
- 向量化表示
  - 通过神经网络提取非结构化数据特征，形成向量化表示
  - 向量具备语义表达能力，能用于相似性检索
- 向量检索
  - 向量索引以树/图/倒排/哈希等方式组织数据，加速检索过程
  - 通过向量间的距离计算，找出最相近向量
  - 检索过程包含着大量浮点数计算



Vector Search



AiDD 7th 2025

PART 02

蚂蚁开源检索库 VSAG

VSAG – 简介

开源向量检索库 VSAG

VSAG 是一个用于向量相似性检索的索引库。其中的索引算法允许用户在大小各种规模的数据集上进行高效检索，特别是那些无法放进内存的数据集。索引库 VSAG 使用 C++ 编写，在蚂蚁集团内部已服务百亿级别非结构化数据存储和检索。

VSAG
Discover the unseen similarities

高召回率 & 低内存使用

创新的混合索引融合了主成分分析、量化、误差估计、共轭图、重排序和磁盘存储等先进方法，实现了内存占用的大幅优化。即便在低内存环境下，也能保持较高的向量召回精度。

高性能

检索方面融合了多种近似距离计算方法和剪枝技术，大幅提升检索效率。结合 extrainfo 技术避免回表操作，实现卓越的端到端性能。

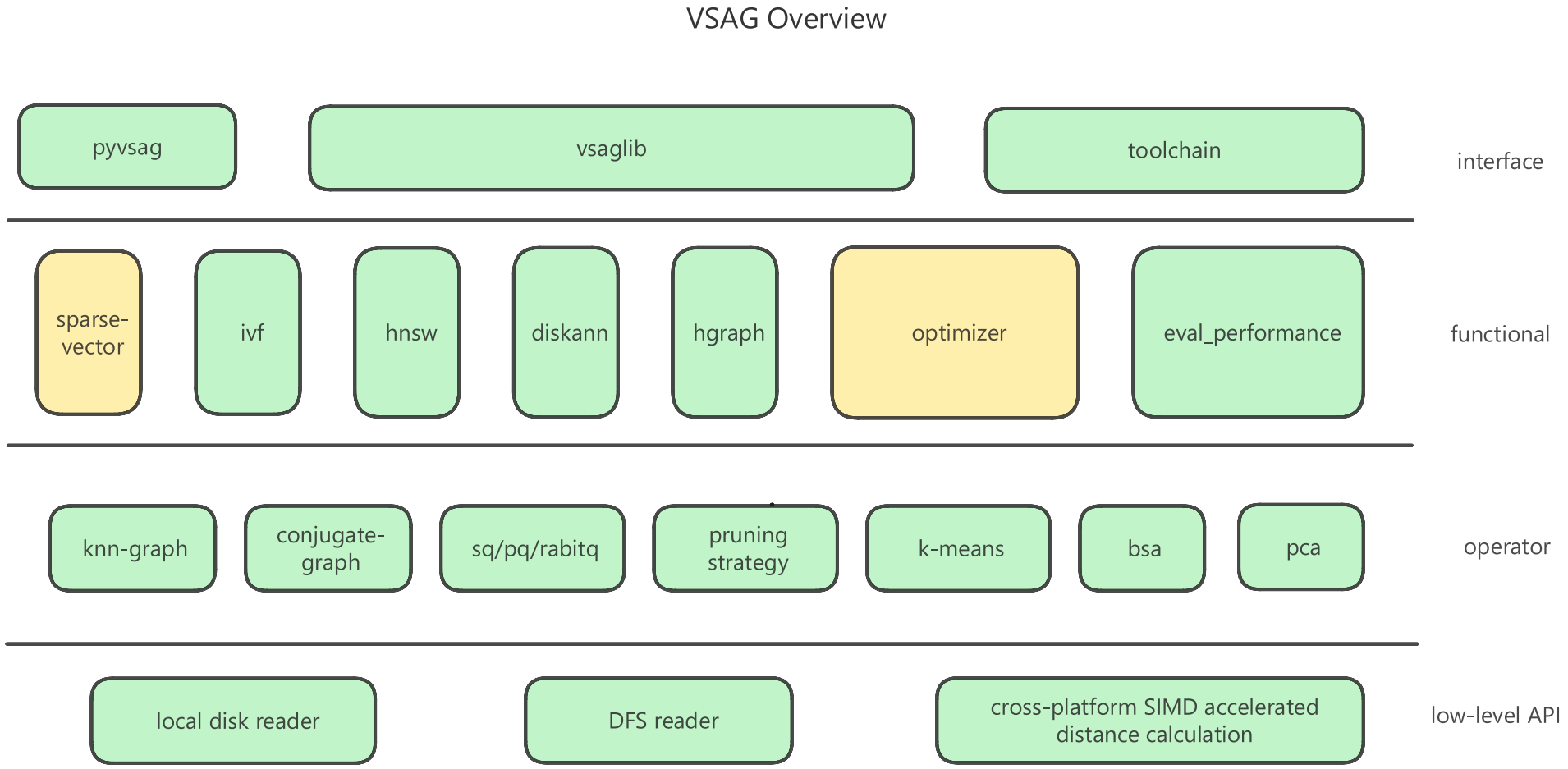
混合搜索支持

VSAG 除标准的 TopK 搜索以及 Range 搜索外，还支持基于 Bitmap 的前置过滤和基于 Callback 的后置过滤两种混合搜索方法。

易于使用

使用 C++ 编写，支持以 CMake 方式快速集成到存储产品中。同时提供 Python 封装，无缝接入 AI 生态应用。支持无参数构建索引以及运行搜索。

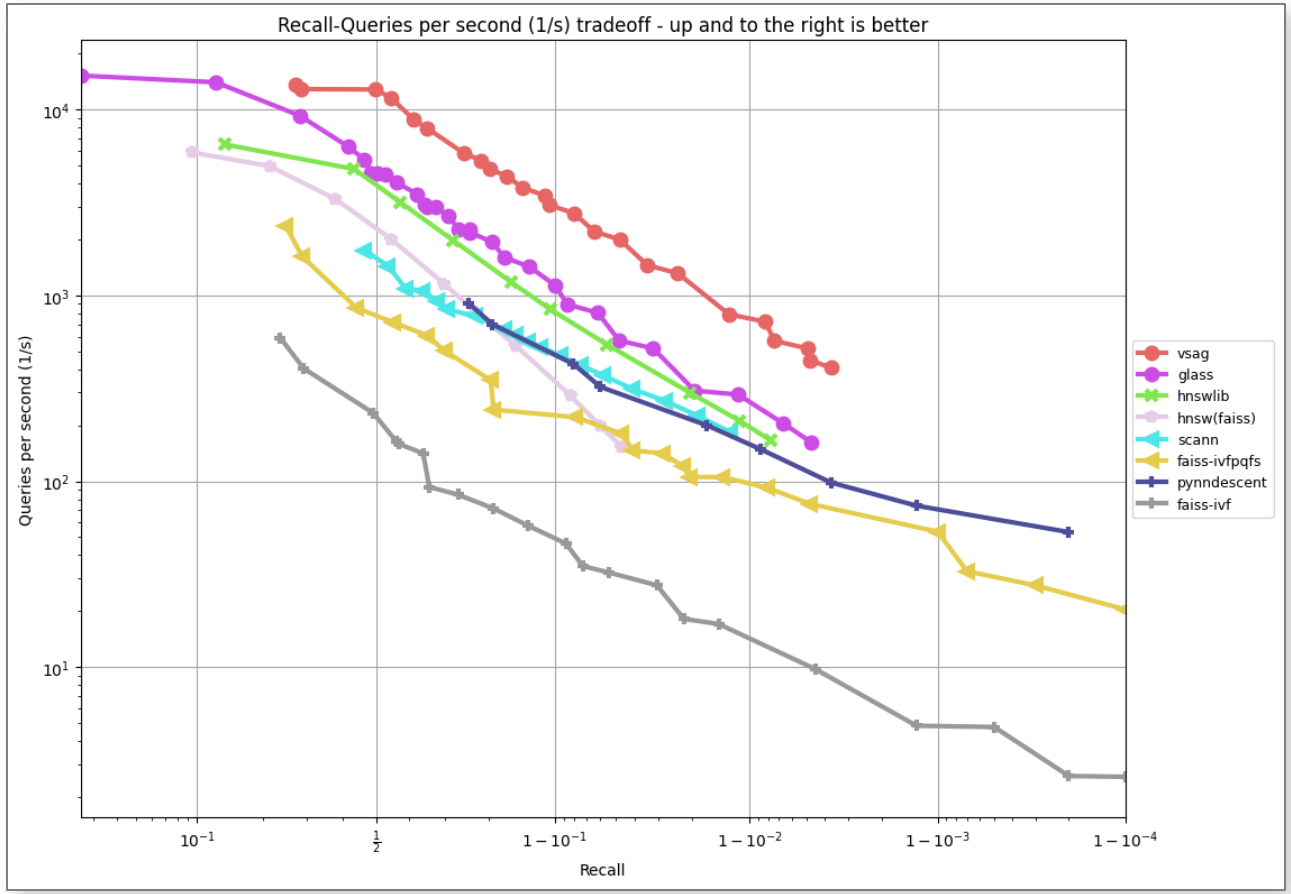
VSAG – 架构一览


VSAG Overview
pyvsag
vsaglib
toolchain
interface
sparse-vector
ivf
hnsw
diskann
hgraph
optimizer
eval_performance
functional
knn-graph
conjugate-graph
sq/pq/rabitq
pruning strategy
k-means
bsa
pca
operator
local disk reader
DFS reader
cross-platform SIMD accelerated distance calculation
low-level API

VSAG – 性能

优化策略

- BSA剪枝：使用学习分类器来对量化和完整距离进行分类，以过滤不必要的距离计算；+60%
- 8-bit量化：大幅降低向量距离计算量；+50%
- Prefetch：使用预取指令提升数据cache命中率；+20%
- 指令重排：将预取指令和计算指令混合排布提高预取有效率；+20%
- 量化重排：使用4-bit量化进行搜索，再用高精度向量重排；+25%
- CPI优化：减少无效的预取指令，避免指令堆积；+10%
- re-rank剪枝：根据距离过滤无效re-rank过程；+5%
- 指令依赖优化：减少访存指令之间的依赖，提高指令吞吐；+2.5%
- 内存排布优化：在图结构上冗余存储部分向量，提升数据locality；+25%
- Core Bound优化：进一步减少无效的访存指令，降低同时间指令发射数；+10%



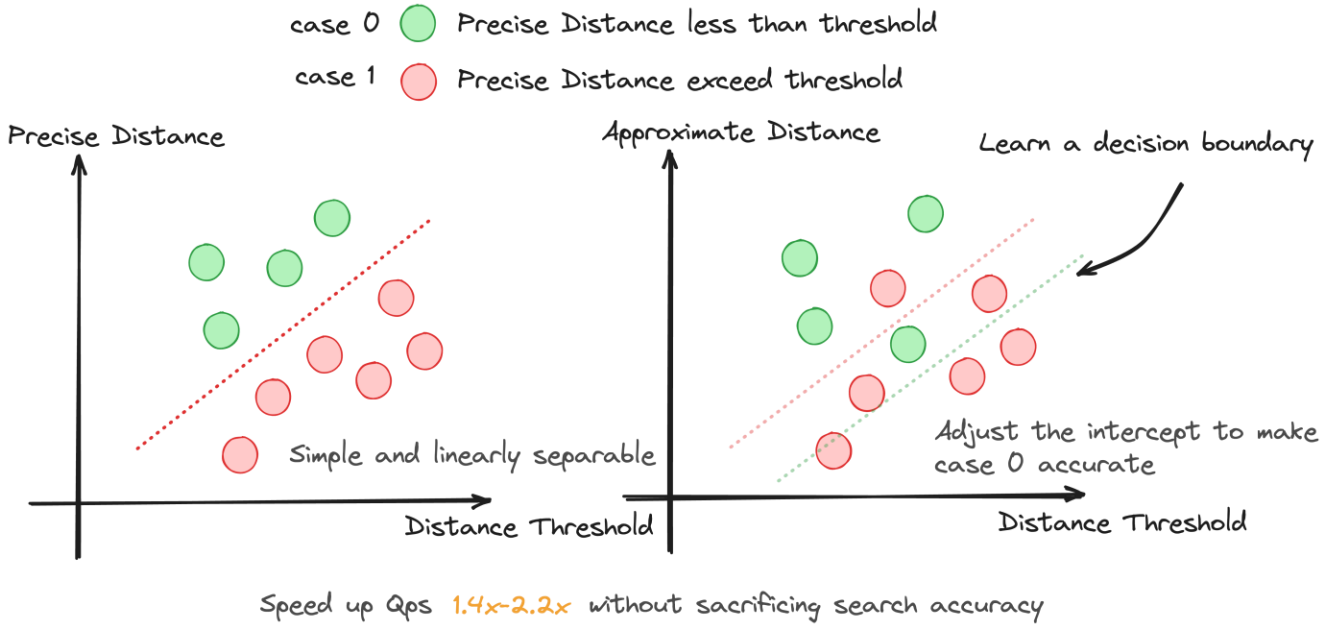
PART 03

前沿向量检索算法和优化

算法和优化 – BSA 剪枝框架

BSA 剪枝框架 – 基于近似距离和线性分类器的距离计算加速

- 引入近似距离计算机制
- 第一阶段使用压缩向量和近似距离搜索
- 第二阶段使用原始向量和精确向量进行二次重排
- 向量查找速度有 1.4x ~ 2.2x 提升（实验数据）
- 论文已被 ICDE 2024 接收
  （https://arxiv.org/abs/2404.16322）

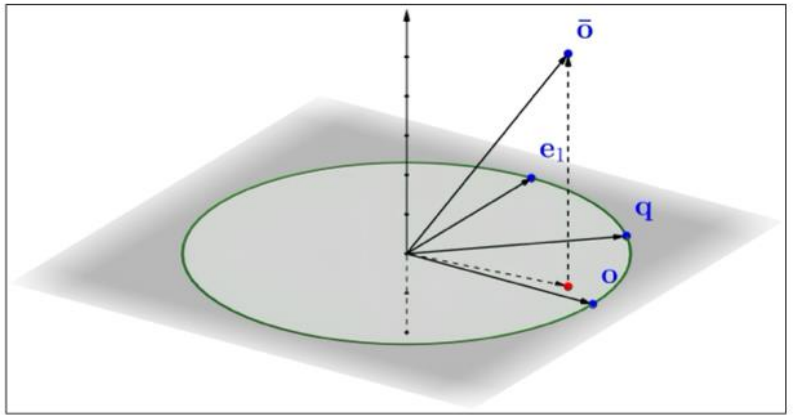


学术合作项目：港科大 x 蚂蚁集团，ICDE 2024

算法和优化 – Binary 量化

Binary 量化（RabitQ）

- 量化技术能将向量数据用更少的比特位代替
- 常见的量化方法有 标量量化、乘积量化
- 新加坡南洋理工大学在 SIGMOD 2025 提出 RabitQ 二值量化方法，在 32x 压缩比的情况下做到很少的精度损失
- 基于 RabitQ 论文进行算法实现，提供生产可用版本



$$\langle \bar{o}, q \rangle = \langle \bar{o}, o \rangle \cdot \langle o, q \rangle + \sqrt{1 - \langle o, q \rangle^2} \cdot \langle \bar{o}, e_1 \rangle$$

$$\frac{\langle \bar{o}, q \rangle}{\langle \bar{o}, o \rangle} = \langle o, q \rangle + \sqrt{1 - \langle o, q \rangle^2} \cdot \frac{\langle \bar{o}, e_1 \rangle}{\langle \bar{o}, o \rangle}$$

Estimator　Target　Error

新加坡南洋理工大学 SIGMOD 2025 RabitQ 论文复现

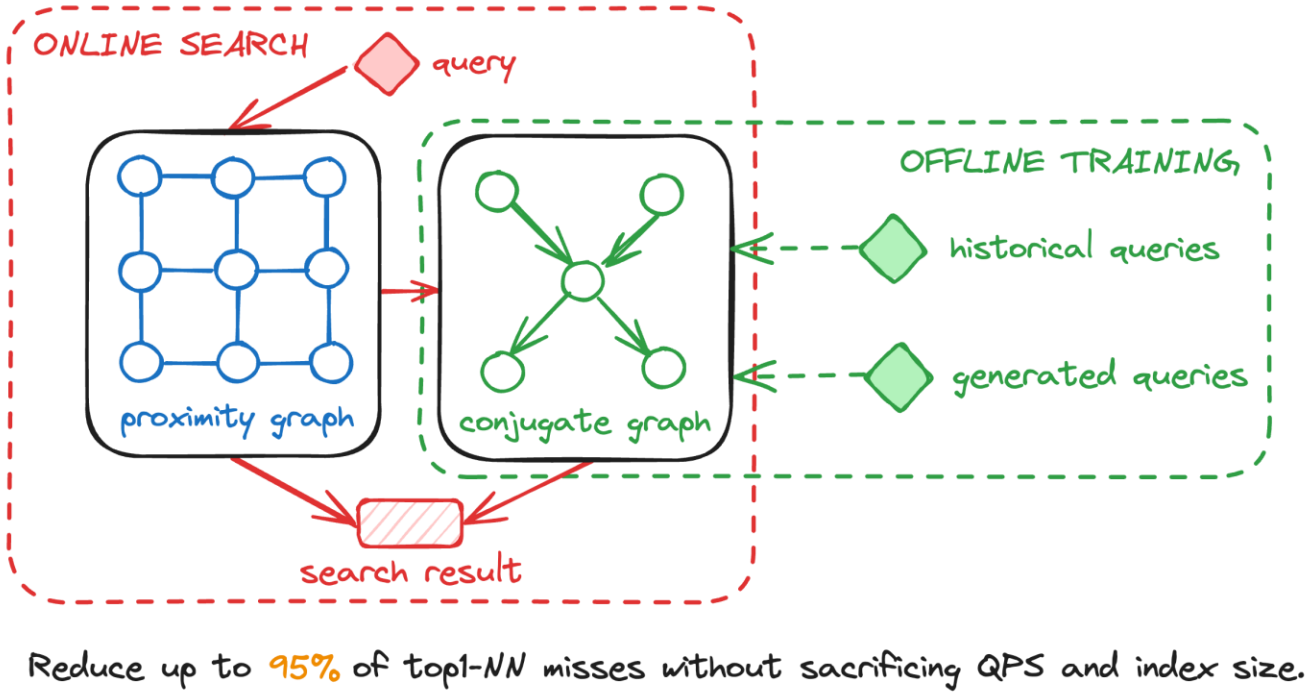
算法和优化 – 基于反馈的召回优化

更高召回率突破（EnhanceGraph）

- 引入创新性的共轭图（conjugate graph）
- 根据用户查询反馈来改进图结构
- 不断优化近邻图的连通性

100 万数据集优化效果（实验数据）

- 基于生成：召回率从 99.8% 到 99.96%
- 基于反馈：召回率从 99.8% 到 99.97%
- 召回失败的人群有最高 95% 以上概率不再失败
- 成本：内存消耗增长约 3%，吞吐不受影响

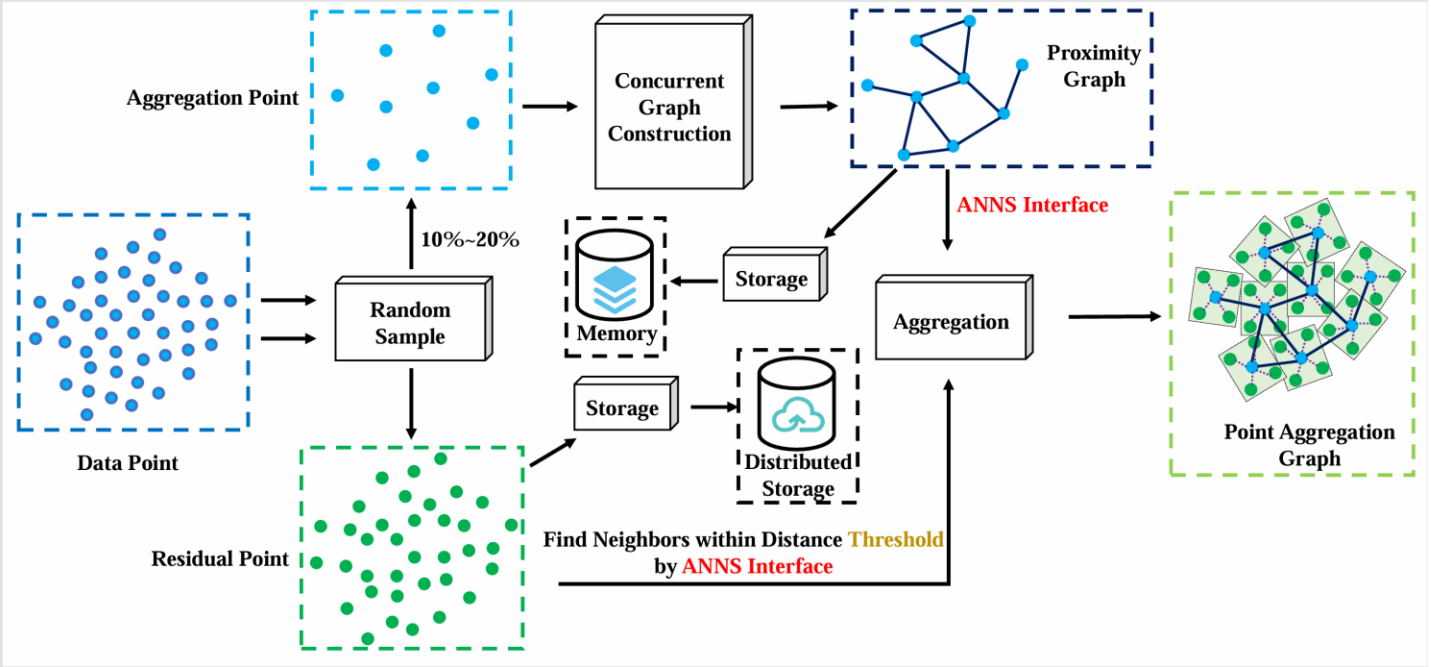


学术合作项目：华师大 x 蚂蚁集团

算法和优化 – 磁盘索引改进

内存 + 磁盘索引上的改进（PAG）

- PAG（Point Aggregation Graph）是一种新的图-聚类混合索引
- 为分布式存储架构优化
- 将 Graph 中距离接近的点压缩成一个节点，以降低存储成本，减少 IO 次数

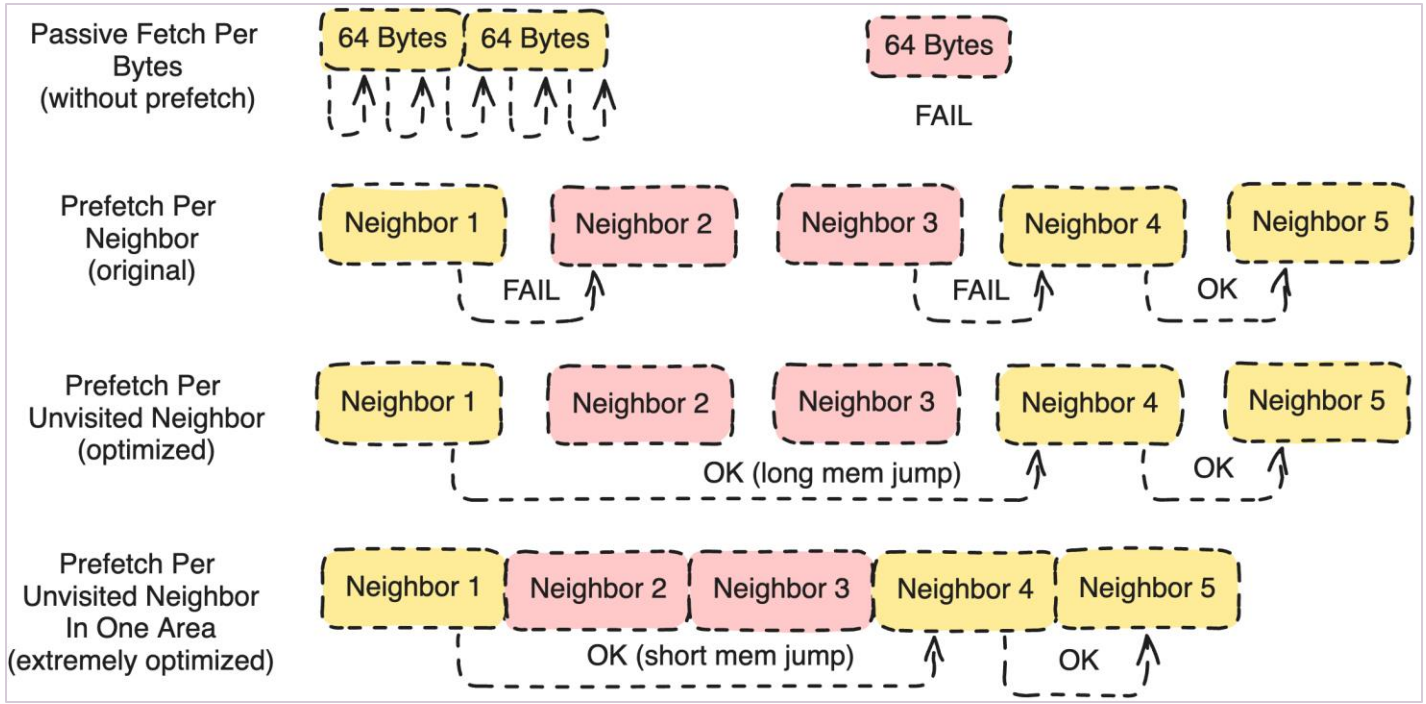


学术合作项目：华师大 x 蚂蚁集团

算法和优化 – 数据预取

数据排布和预取

- 近邻图在性能和延迟上都有更好表现
- 但近邻图检索过程有大量的随机内存访问
- 内存排布和数据预取优化非常重要
- 收益：内存排布+25%，数据预取+20%
- 论文已被 VLDB 2025 接收
  (https://arxiv.org/abs/2503.17911)

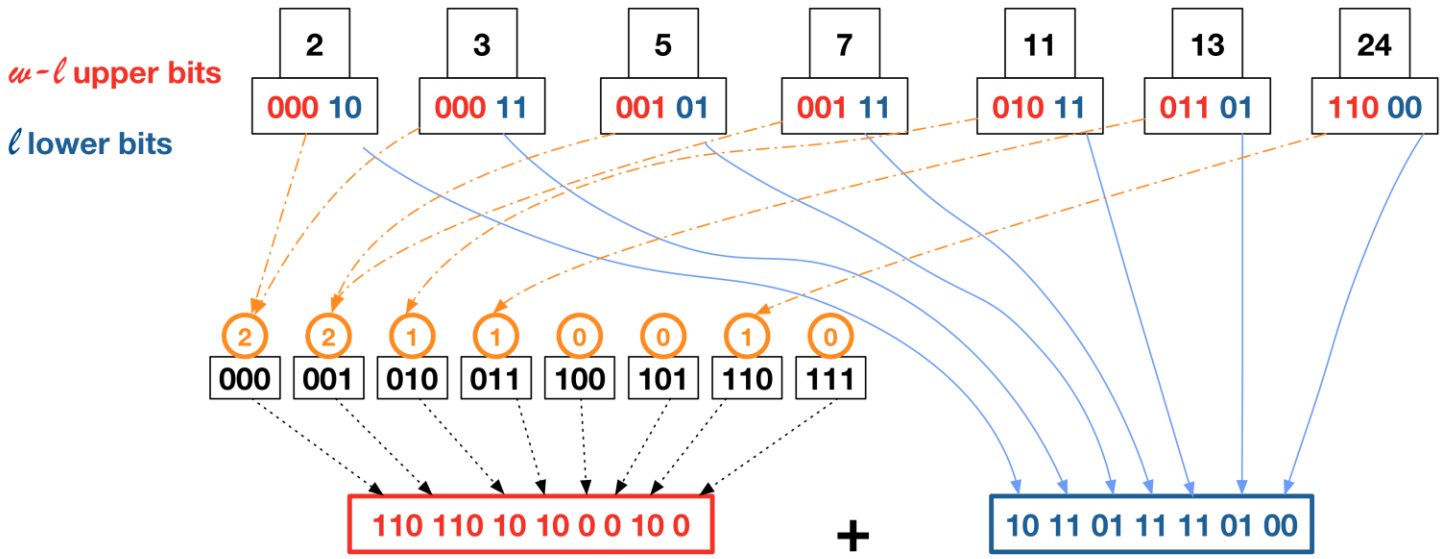


蚂蚁集团，VLDB 2025

算法和优化 – 图结构压缩

图结构压缩

- 在使用高倍率量化向量后（例如 PQ、RabitQ），图结构可能占到向量索引的 70% 以上
- 通过使用 elias-fano-encoding 方法对邻居 ID 进行编码，可以大幅减少临接表占用内存空间，可达原始临接表的 1/3 左右
- 相较于原始算法，引入图结构压缩算法后，索引大小能降低约 50%

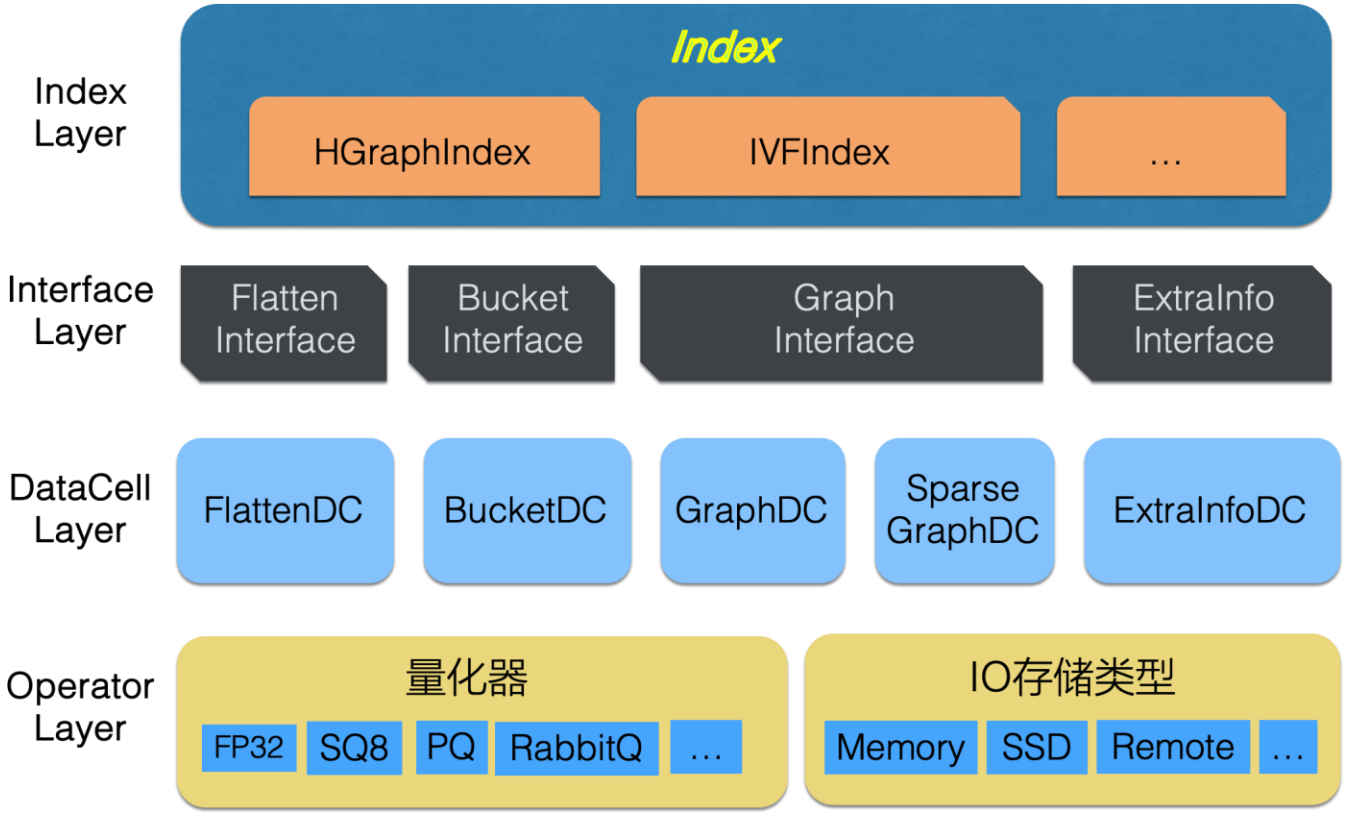


插图来自：https://www.antoniomallia.it/sorted-integers-compression-with-elias-fano-encoding.html

算法和优化 – HGraph 层次图索引

HGraph 层次图索引

- HGraph 是层次图的一种新实现
- 支持构建多级量化结构，能根据运行环境和数据规模配置索引
- 通过将图/倒排索引拆分、数据访问接口的抽象，并结合多种量化能力，HGraph 提供了机遇量化和多阶段重排的索引框架
- 通过该框架，能快速构建出高度自定义的全新索引



PART 04

业务落地中的实践案例

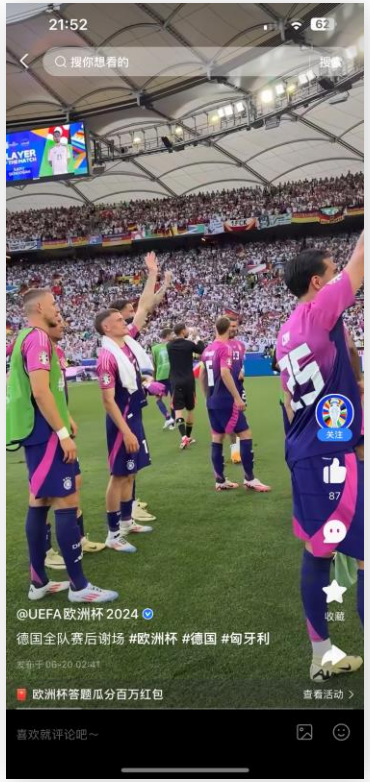
业务 – 多媒体内容检索成本降低

向量存储成本挑战

- 内容爆发式增长
- 向量存储/检索成本高
- 大规模数据查询耗时高

场景

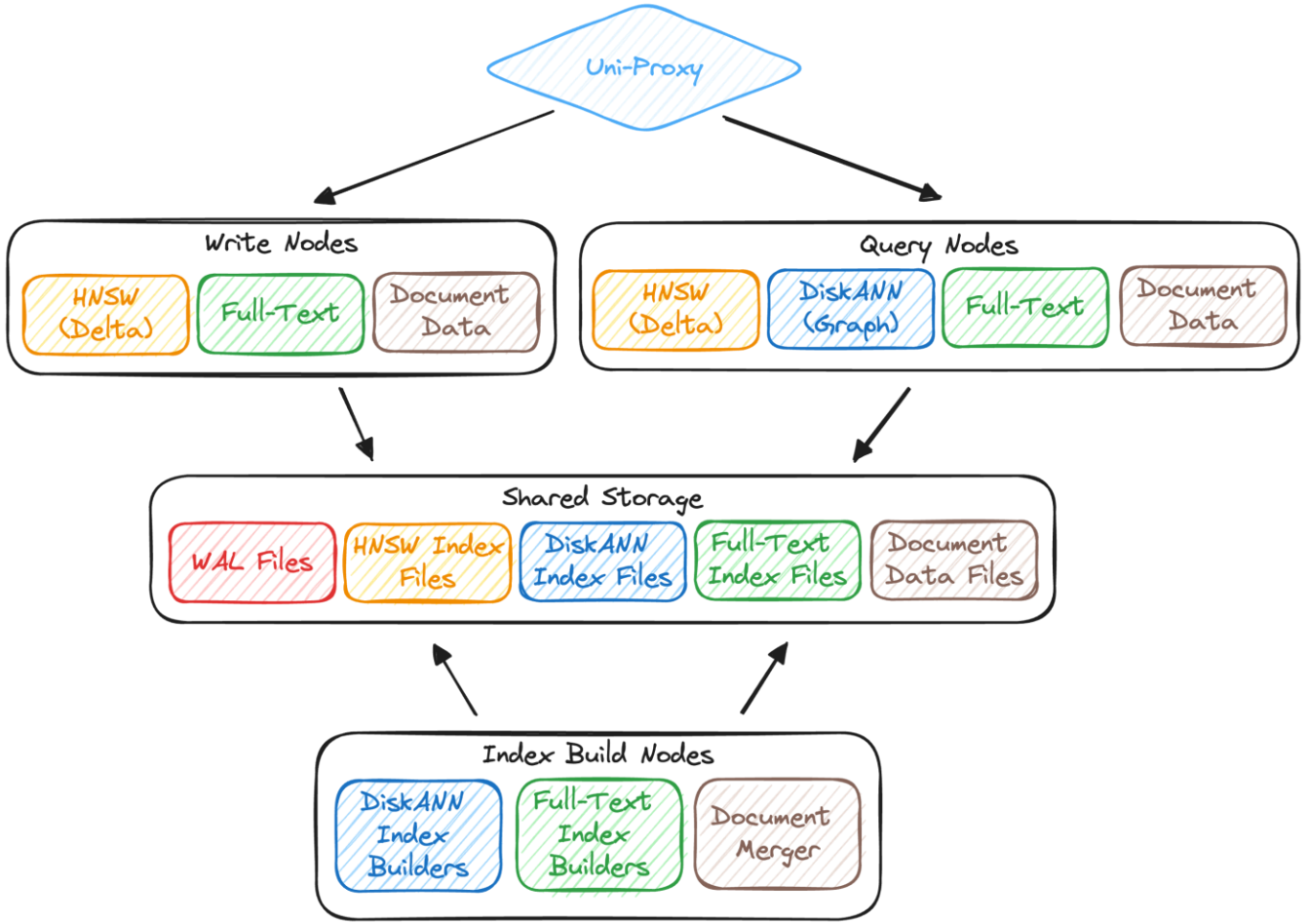
- 内容安全防控
- 多媒体内容管理



业务 – 多媒体内容检索成本降低

采用 HNSW + DiskANN 混合索引方案
（应对成本挑战）

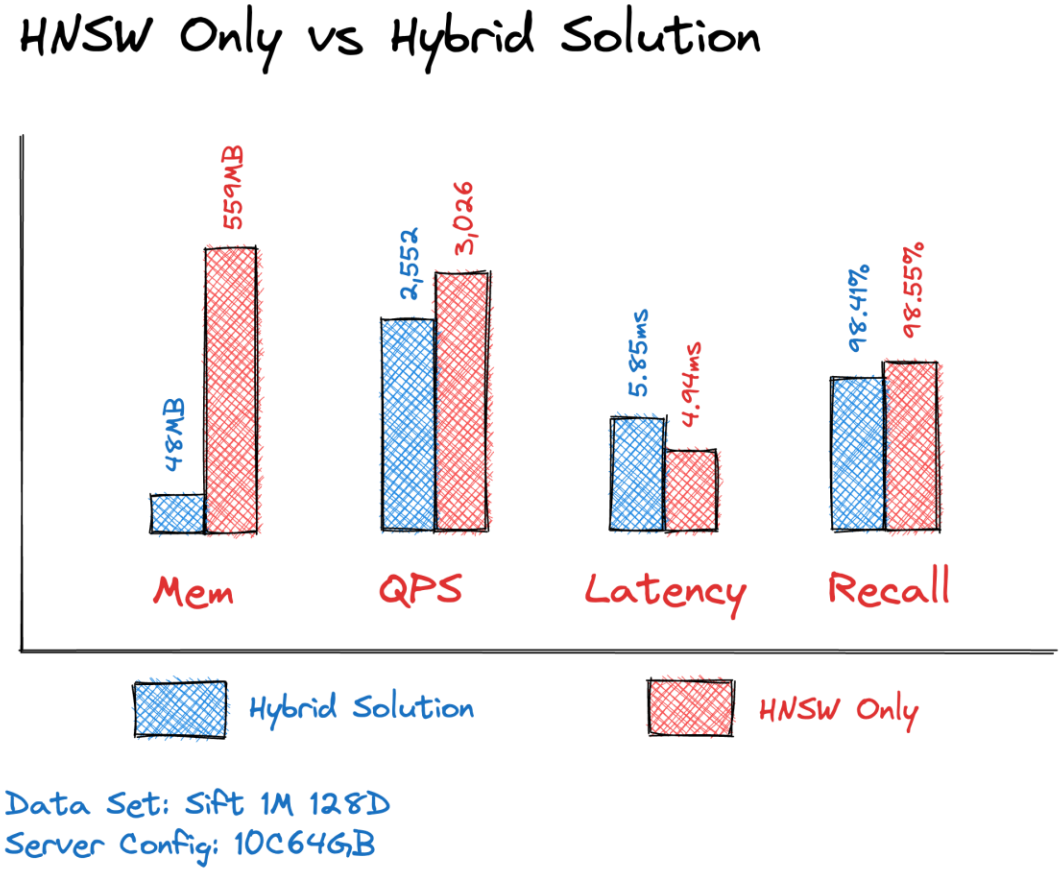
- 融合使用 HNSW 和 DiskANN 索引
- 充分利用两个索引的优点
- 通过工程优化巧妙规避缺点



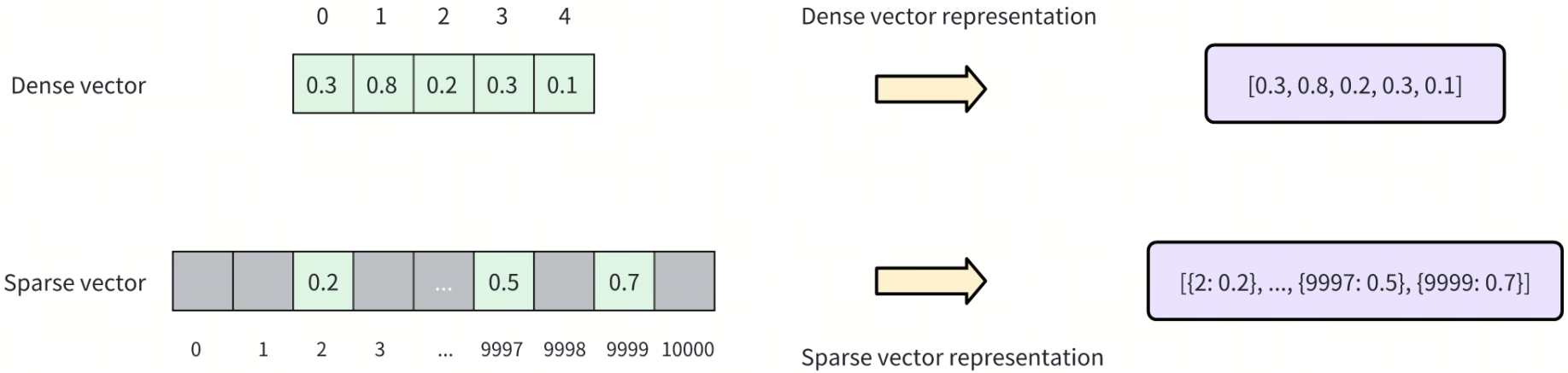
业务 – 多媒体内容检索成本降低

HNSW + DiskANN 混合索引方案

- 内存需求是纯 HNSW 方案的 1/10
- QPS 和延迟与 HNSW 相当
- 召回率与 HNSW 相当



业务 – RAG 文本检索召回率提升



稀疏向量特点

- 一种高维向量的特殊表示方式，每个维度对应一个单词，其中大部分元素为零
- 相比于稠密向量，计算成本更低，精确匹配关键词或者短语时效果更好

业务 – RAG 文本检索召回率提升

稠密向量/稀疏向量/BM25的召回效果

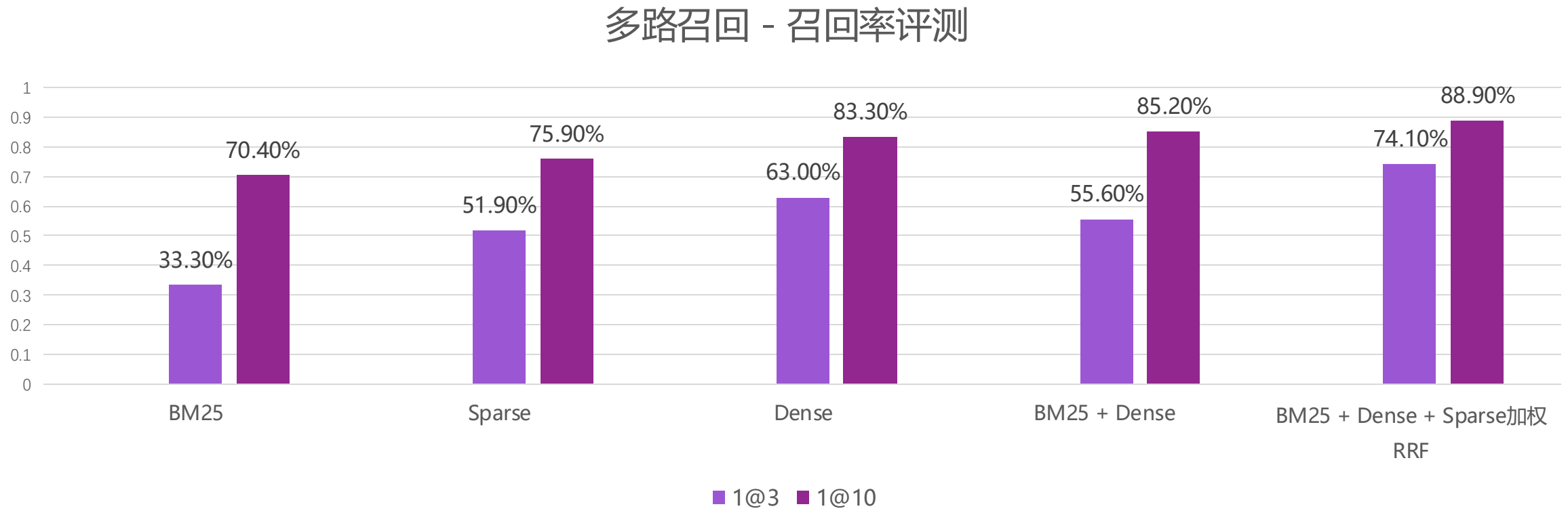
	Dense	Sparse	BM25
1@1	63.89%	68.97%	68.46%
1@2	70.77%	77.18%	76.52%
1@3	74.18%	80.40%	80.07%
1@5	78.38%	84.08%	83.77%
1@10	82.78%	88.16%	87.35%
1@20	86.69%	90.80%	90.02%
1@100	93.14%	95.25%	94.15%

A 或 B：算法在 Top-10 中找到正确结果的查询集合

A	B	A & B	A - B	B - A	A U B
Sparse	Dense	80.03%	8.14%	2.75%	91.92%
Sparse	BM25	83.28%	4.88%	4.07%	92.23%
Dense	BM25	77.62%	5.15%	9.72%	92.49%

总结：不同索引能够实现互补，提升召回率

业务 – RAG 文本检索召回率提升



增加稀疏向量的收益：

- 1@10 召回率提升 3.7%，1@3 召回率提升 18%

PART 05

开源社区与展望

VSAG v1.0 Roadmap

功能方面

- 支持常见的数据类型，满足不同场景的非结构化数据检索需求
  - FP32 向量：满足主流向量检索场景使用
  - INT8、BF16、FP16 向量：适配量化的 embedding 模型，避免额外的存储开销
  - 稀疏向量：扩展文本检索方式
- 提供全面优化的核心索引类型，覆盖绝大部分检索场景
  - 图索引 HGraph：满足对高精度和低延迟的要求
  - 倒排索引 IVF：满足大 K 和批量查询的需求
- 提供丰富的量化方式，满足内存/召回率的平衡
  - RabitQ（BQ）：超高倍率的压缩，极少的内存使用
  - PQ：灵活的压缩倍率，适合低精度要求的场景
  - SQ4、SQ8：常规压缩方式，少量牺牲召回率获得内存和性能收益

平台与资源方面

- 多平台指令集适配，减少系统集成分发工作量
  - x86_64 平台：SSE，AVX，AVX2，AVX512
  - ARM 平台：Neon，SVE
  - 可选的矩阵乘法加速库：intel-mkl，openblas
- 支持资源隔离，提供细粒度的运行资源可配置
  - 内存资源：支持以索引为单位设置内存分配器，以实现类似租户级内存管理
  - CPU 资源：支持注入线程池，从而提升写入吞吐和搜索吞吐

开源社区

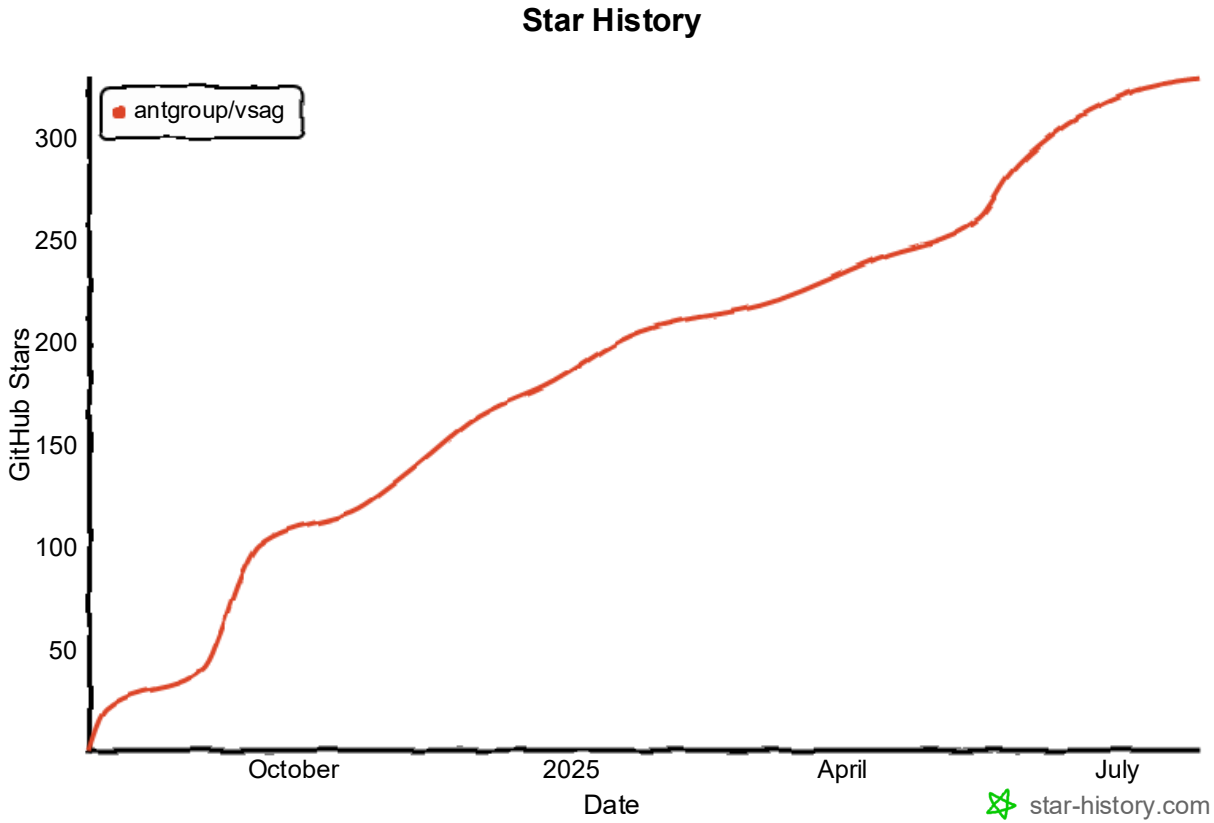
项目地址：https://github.com/antgroup/vsag

邮箱地址：the.vsag.project@gmail.com

微信公众号：StorageScale

公众号

VSAG 开源交流



AiDD 8th 2025

第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

查看会议详情

下一站预告 11/14-15 | 深圳站 12/19-20 | 上海站

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

上下文工程

AiDD

「深行·浅智」

Walk Deep, Think Light.

2025.11.16

AiDD首届麦理浩径徒步

KEYLINK ing

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

AiDD峰会详情

上海站

AI+研发数字峰会

AiDD 峰会

时间：

2025.05.23-24

北京站

AI+研发数字峰会

AiDD 峰会

时间：

2025.08.08-09

深圳站

AI+研发数字峰会

AiDD 峰会

时间：

2025.11.14-15

香港站

「深行·浅智」

AiDD 首届麦理浩径徒步

时间：

2025.11.16

上海站

AI+产品创新峰会

AiDD × 联合峰会

时间：

2025.12.19-20

第7届AI+研发数字峰会
AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料