南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

DBGroup

向量数据库

南方科技大学
唐　博
tangb3@sustech.edu.cn

专栏 中国计算机学会通讯 第19卷 第11期 2023年11月

学术观点

向量数据库：关键技术、系统架构与未来挑战

唐 博[1] 秦建斌[2] 毛 睿[2]
[1]南方科技大学
[2]深圳大学

关键词：向量数据库 大语言模型 AI基座

受邀在中国计算机学会通讯上综述向量数据库系统热点技术

课程提纲

- ✓ 研究历史: 千呼万唤始出来，犹抱琵琶半遮面
  - 追本溯源、推波助澜、登峰造极
- ❑ 系统特性: 不畏浮云遮望眼，只缘身在此山中
  - 核心功能、系统组件、关键技术
- ❑ 研究积累: 衣带渐宽终不悔，为伊消得人憔悴
  - 十年耕耘、两个冠军、多篇顶会
- ❑ 未来挑战: 问渠那得清如许，为有源头活水来
  - 百亿千维、端侧服务、软硬协同

追本溯源：Top-K查询

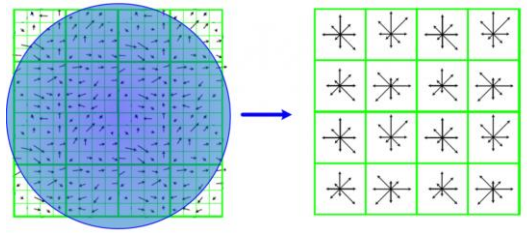
- 时空数据库的Top-K查询
  - Top-*k* query: shortlist **top options** (i.e., rank by score) from a set of alternatives
  - Skyline query, Reverse Top-k query, Uncertain Top-k Query, k Shorted-list Query, Top-Rank Region Query, ….

Hotel	Location	Service
香格里拉	3	8
希尔顿	9	4
洲际酒店	8	3
四季酒店	4	3
博林天瑞	5	5

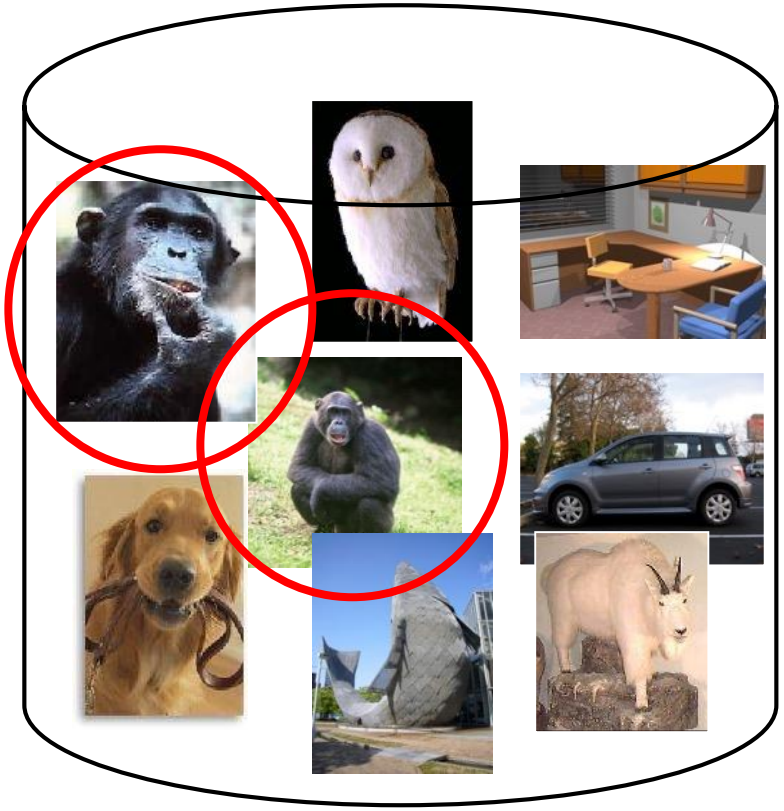
- 查询权重范围[0,1]
- Score = Loc*q_1 + Service*q_2
- $\boldsymbol{q}$ = <0.7, 0.3>,
  博林天瑞 属于 top-**3**
- $\boldsymbol{q}$ = <0.1, 0.9>
  博林天瑞 属于 top-**2**

数据库领域的Top-K查询具备物理意义的数字表征

追本溯源：图片检索

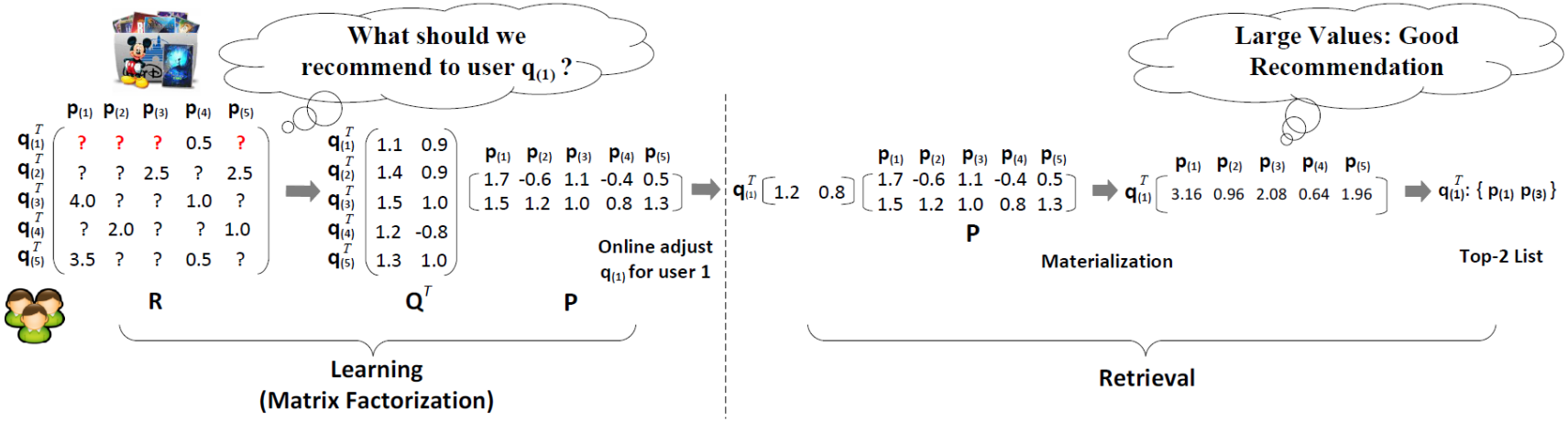


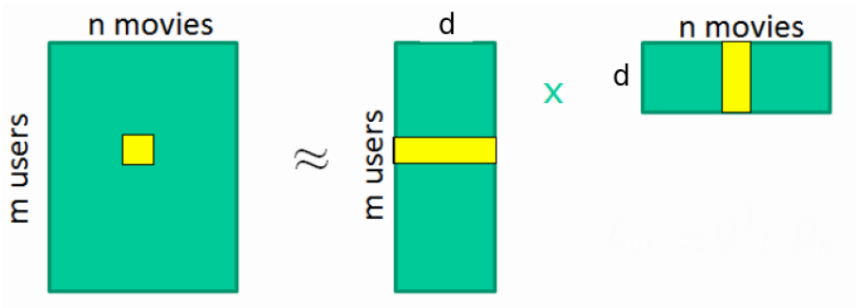
SIFT算子



计算机视觉领域的SIFT是最早的Embedding模型，尺度不变特征变换

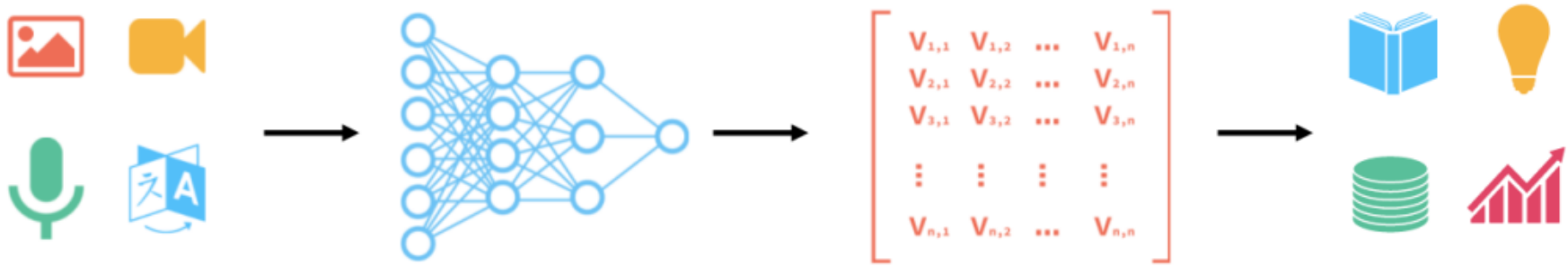
推波助澜：推荐系统





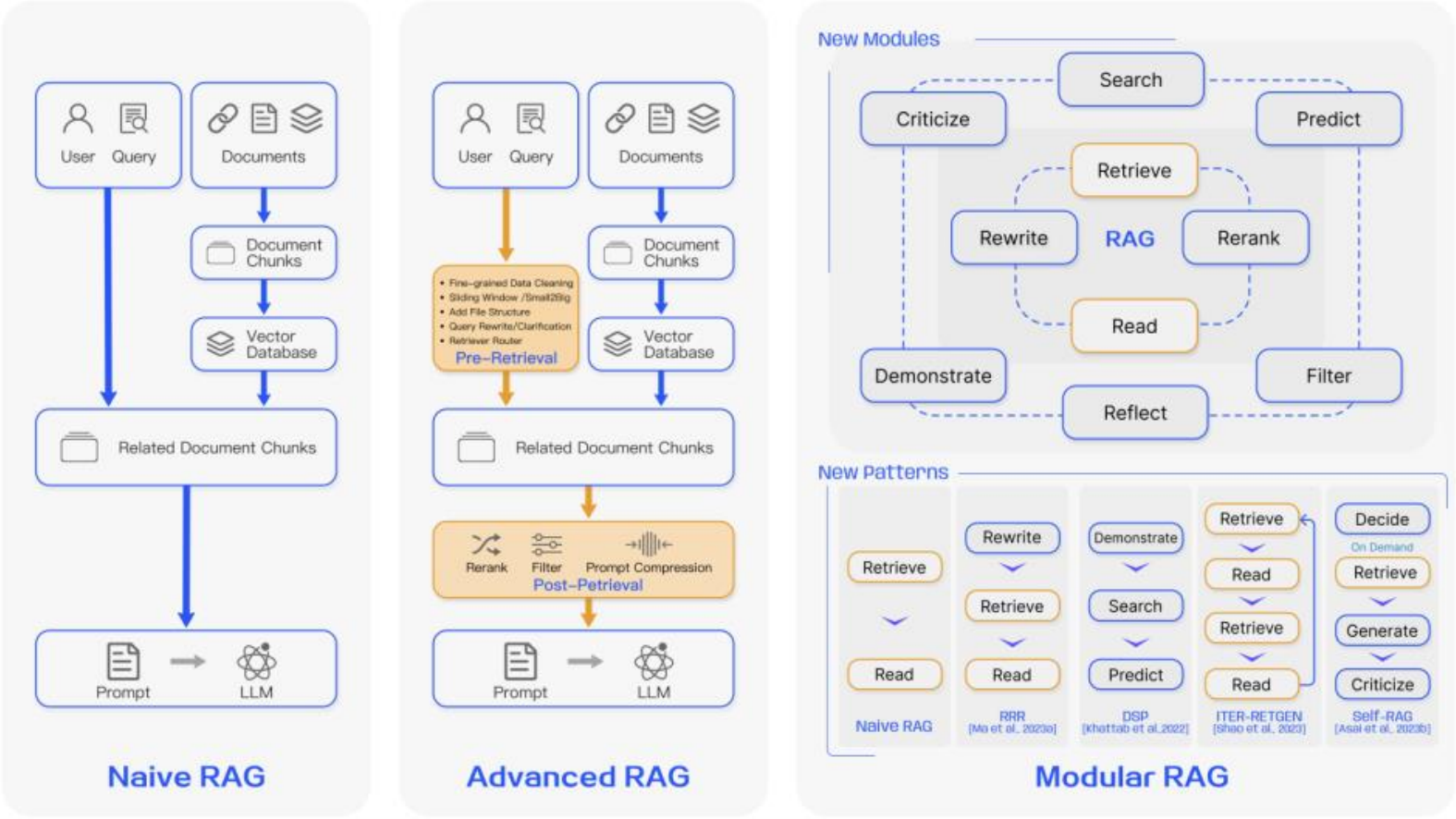
推荐系统领域的矩阵分解技术带来了高维向量相似度高效计算的硬需求

推波助澜：深度学习



火热的向量数据库源于深度学习模型

登峰造极：大语言模型



向量数据库被广泛应用于大语言模型检索增强生成(RAG)

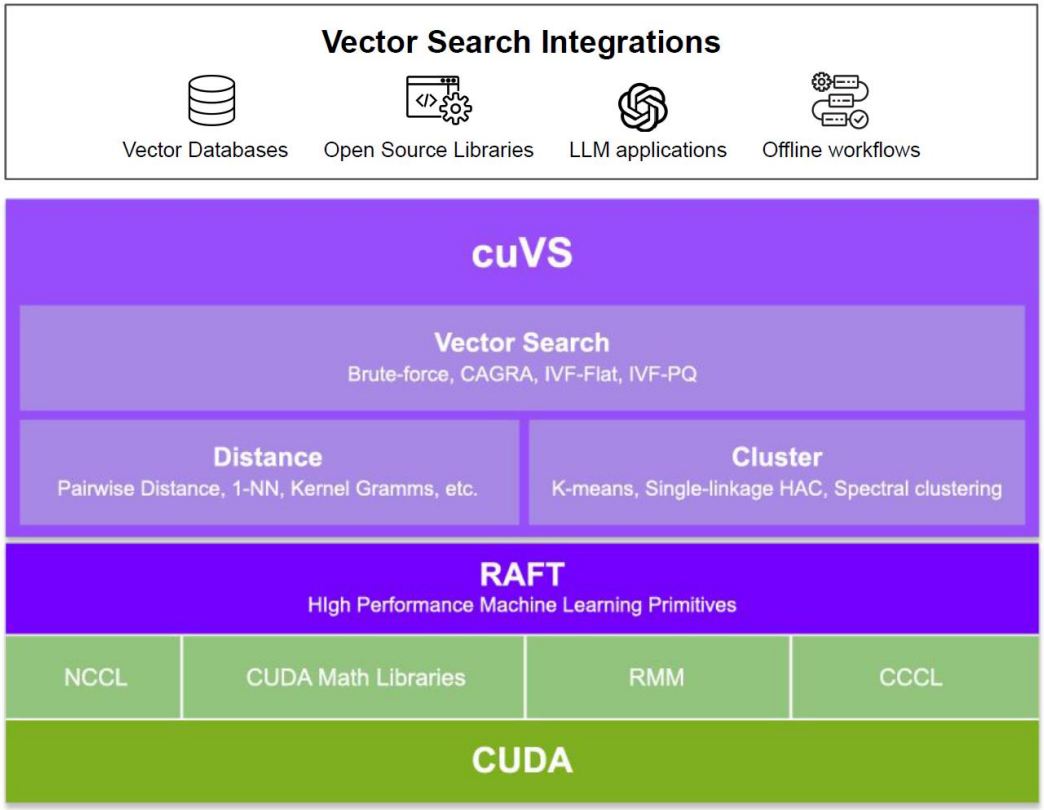
登峰造极：硬件加速

NVIDIA
https://blogs.nvidia.cn › 2020/03/17 · Translate this page

ZILLIZ借助NVIDIA GPU加速AI全流程，打造新一代海量非结构信息

...

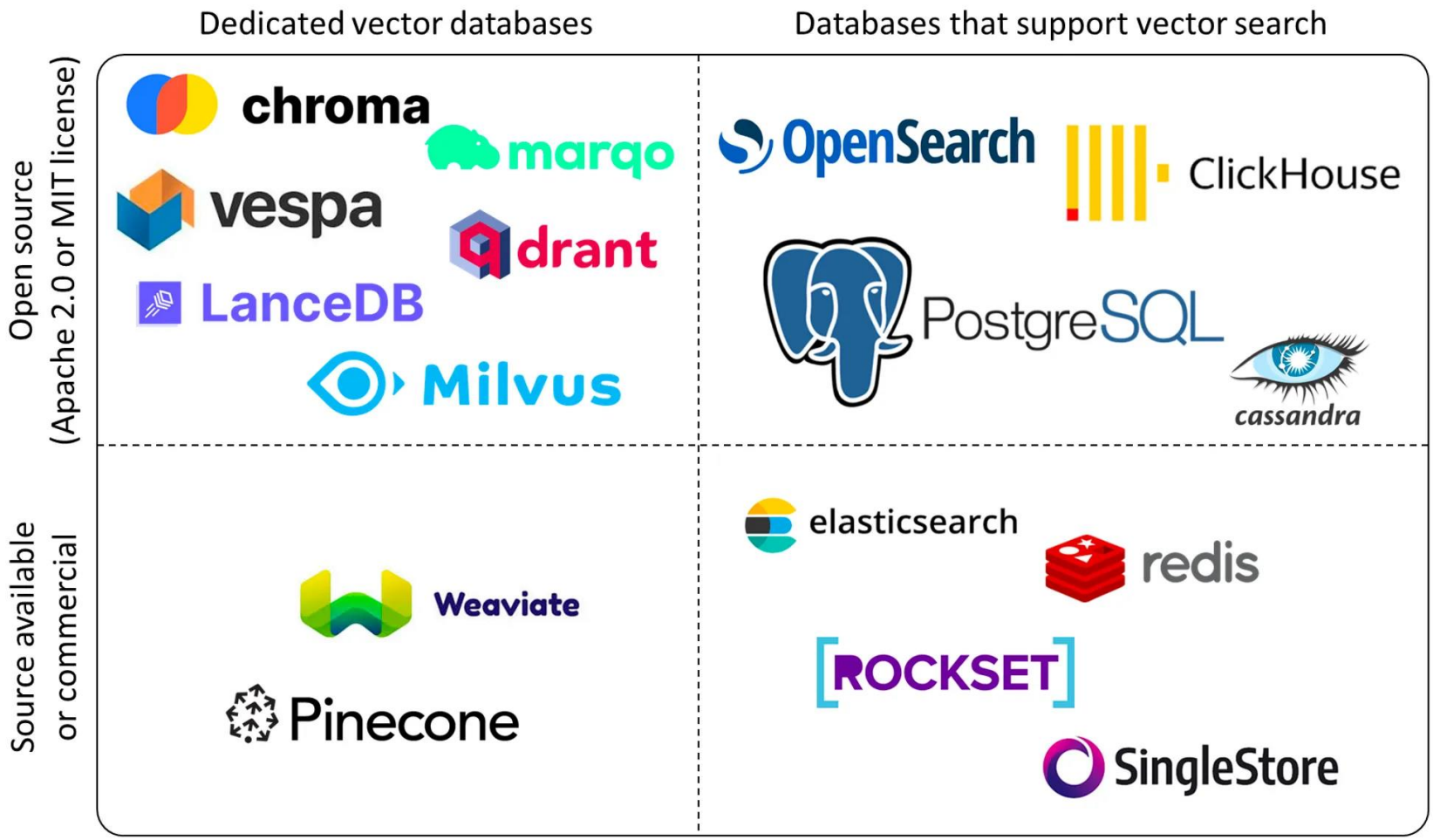
ZILLIZ Milvus是一款针对AI应用大规模落地而研发的，面向海量特征向量搜索比对的数据引擎。 • 在本案例中，借助于NVIDIA GPU的强大算 ...



向量数据库系统汇报提纲

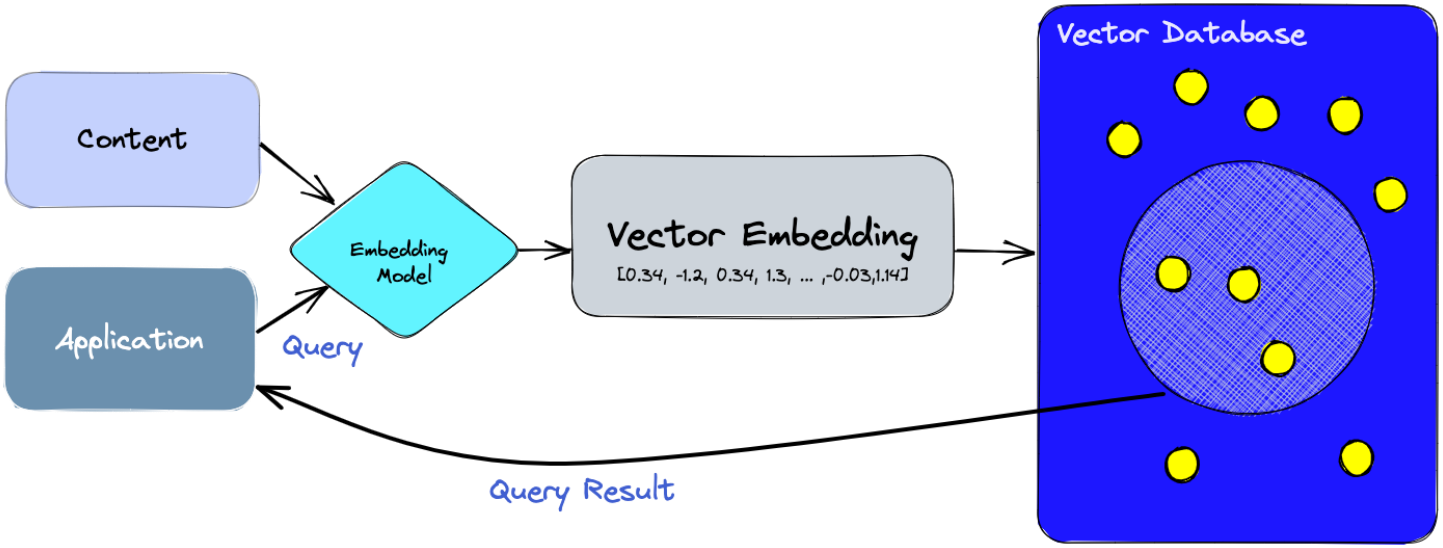
- ✓ 研究历史: 千呼万唤始出来，犹抱琵琶半遮面
  - 追本溯源、推波助澜、登峰造极
- ✓ 系统特性: 不畏浮云遮望眼，只缘身在此山中
  - 核心功能、系统组件、关键技术
- ❑ 研究积累: 衣带渐宽终不悔，为伊消得人憔悴
  - 十年耕耘、两个冠军、多篇顶会
- ❑ 未来挑战: 问渠那得清如许，为有源头活水来
  - 百亿千维、端侧服务、软硬协同

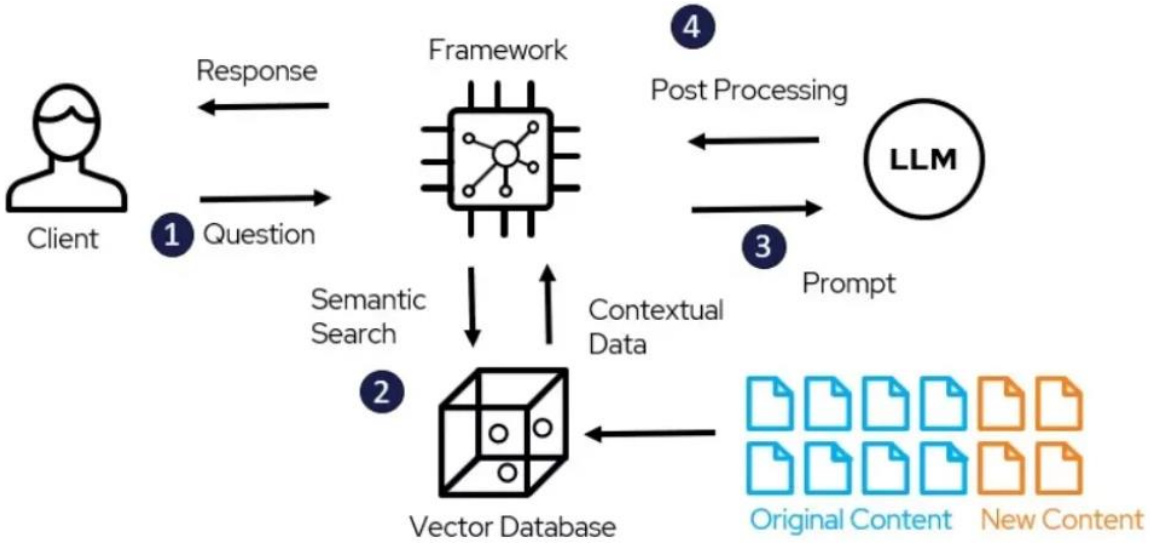
全球向量数据库产品



路线之争: 开源 vs 商业？ 从零打造 vs 现有魔改？

核心功能





向量即知识，向量即洞察力，向量即智能！

相似查询方法论

- Tree-based Solution (基于树形数据结构)
  - R-tree, KD-tree, VP-tree, iDistance
  - The curse of dimensionality

维灾难

- Locality Sensitive Hashing (局部敏感度哈希)
  - LSH, LSB-tree, C2LSH, DSH, LazyLSH, …
  - It provides approximate bound of results

理论保证

- Proximity Graph (邻近图)
  - kNN graph, HNSW, …
  - The core idea is the neighbor's neighbor is neighbor

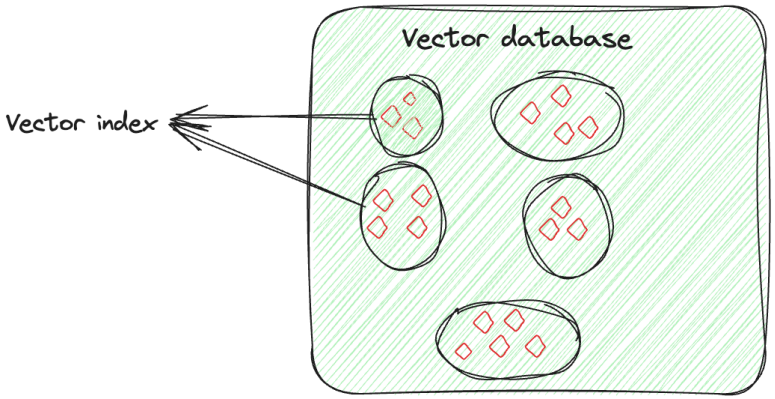
准确率高

- Product Quantization(向量量化)
  - Compress high-dimensional vectors
  - FAISS provides efficient product quantization implementation

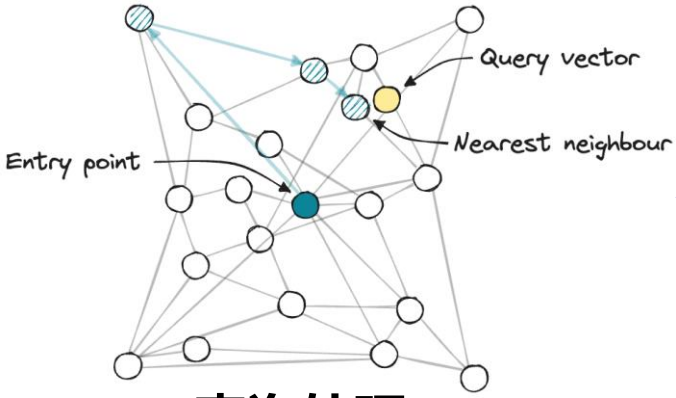
内存小

阿喀琉斯之踵: 准确率recall和处理速度latency不可兼得

核心组件



索引构建



查询处理

Compute Unit

CPU, FPGA, GPU, DSA

New Storage

DRAM, SSD, PMEM

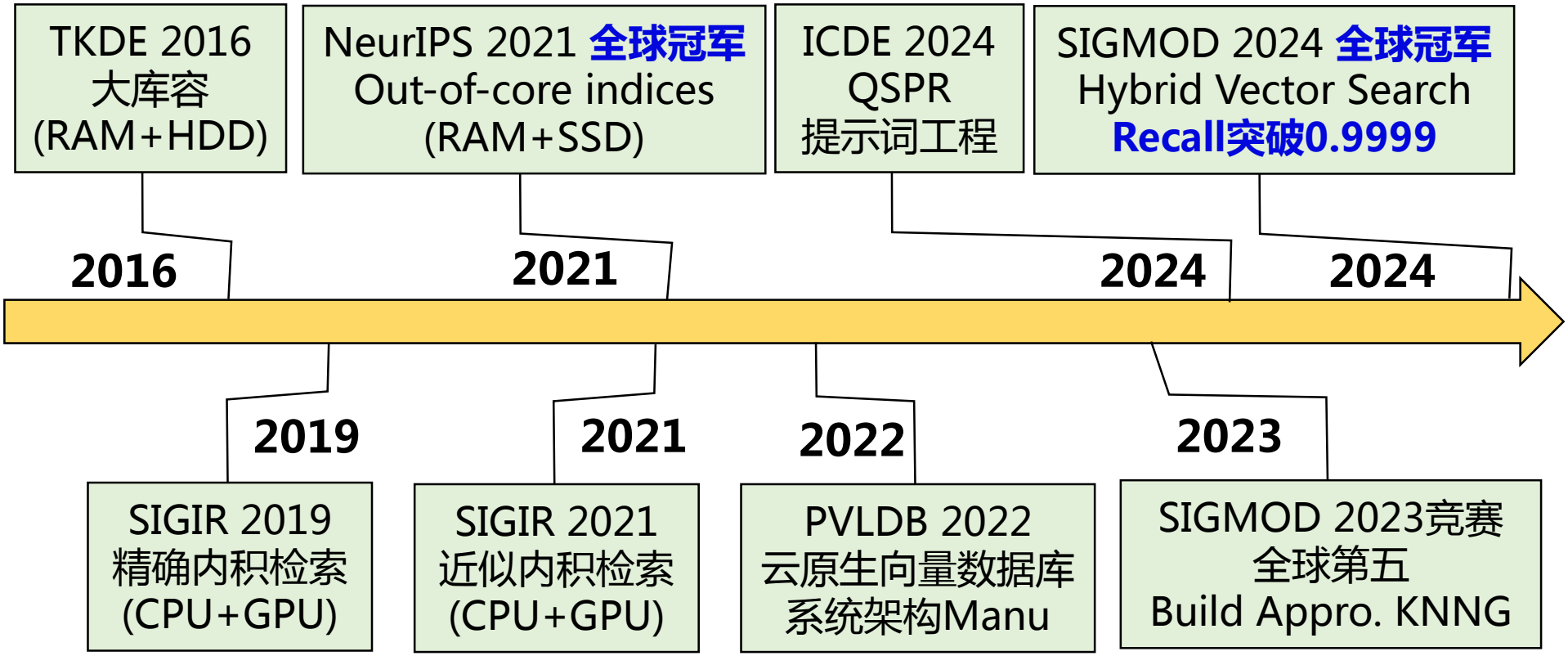
Network

RDMA, CXL, Ethernet

向量数据库系统汇报提纲

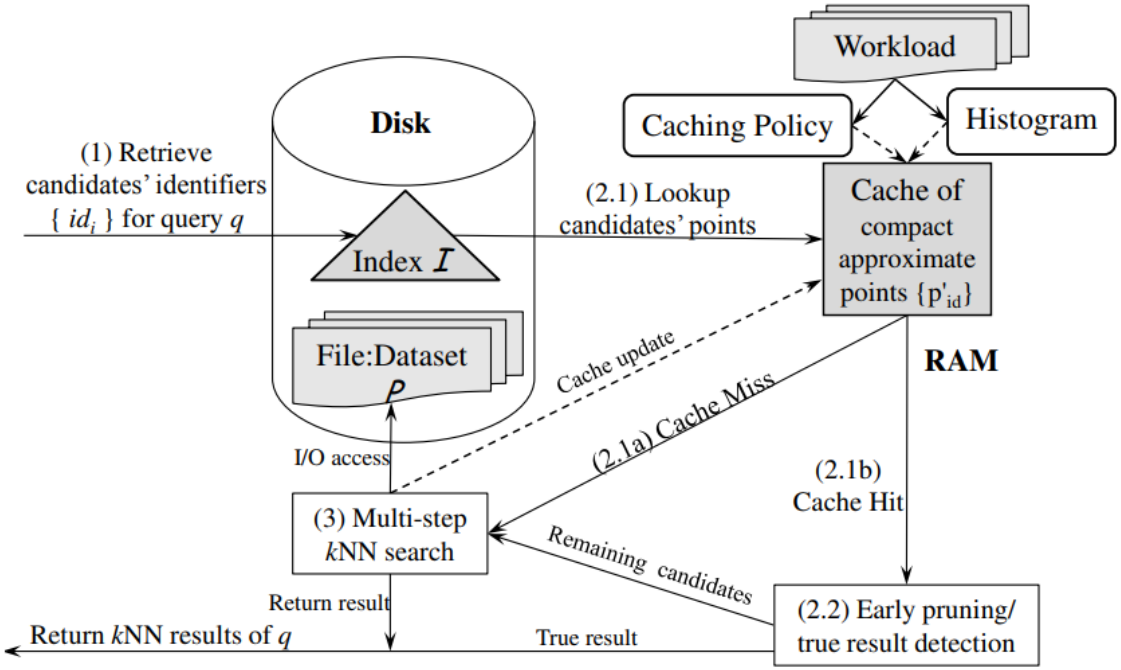
- ✓ 研究历史: 千呼万唤始出来，犹抱琵琶半遮面
  - 追本溯源、推波助澜、登峰造极
- ✓ 系统特性: 不畏浮云遮望眼，只缘身在此山中
  - 核心功能、系统组件、关键技术
- ✓ 研究积累: 衣带渐宽终不悔，为伊消得人憔悴
  - 十年耕耘、两个冠军、多篇顶会
- ❑ 未来挑战: 问渠那得清如许，为有源头活水来
  - 百亿千维、端侧服务、软硬协同

研究成果总结

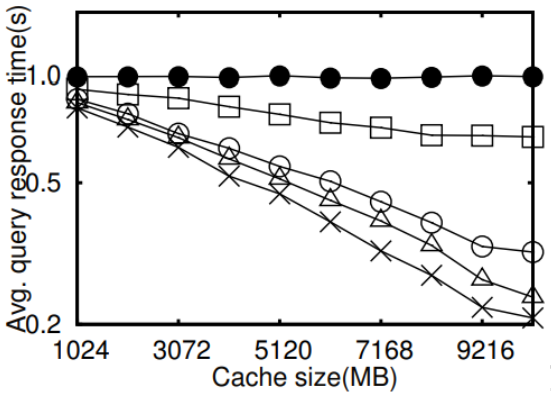


两个全球冠军，多篇CCF-A类论文，技术涵盖大库容到异构计算

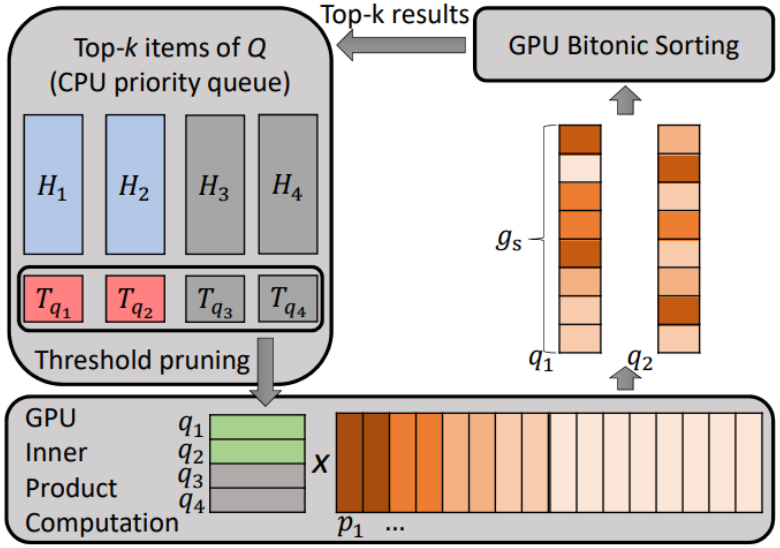
大库容向量检索TKDE 2016

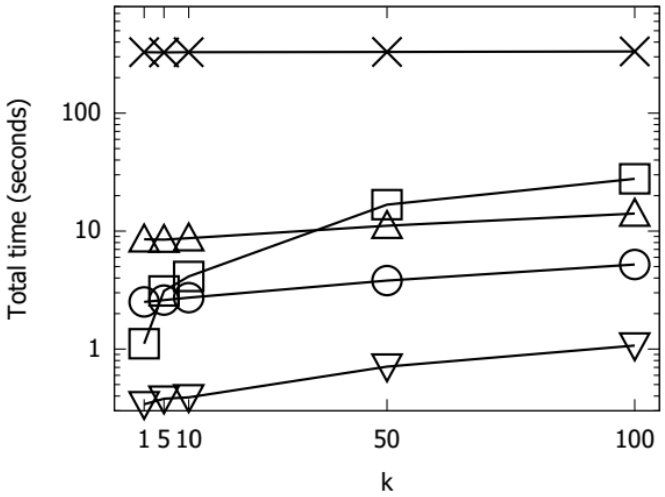


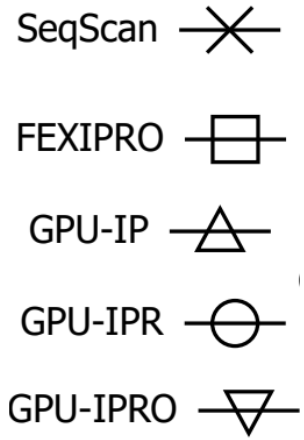
- **大库容高维向量检索时间开销主要来自于IO**
  - 充分利用内存加速准确距离计算 → exploit every bit!
  - 设计简短紧凑的高维向量近似表达方式 → **向量量化**
  - 提出基于近似表达提出距离上下界计算公式



精确距离检索SIGIR 2019

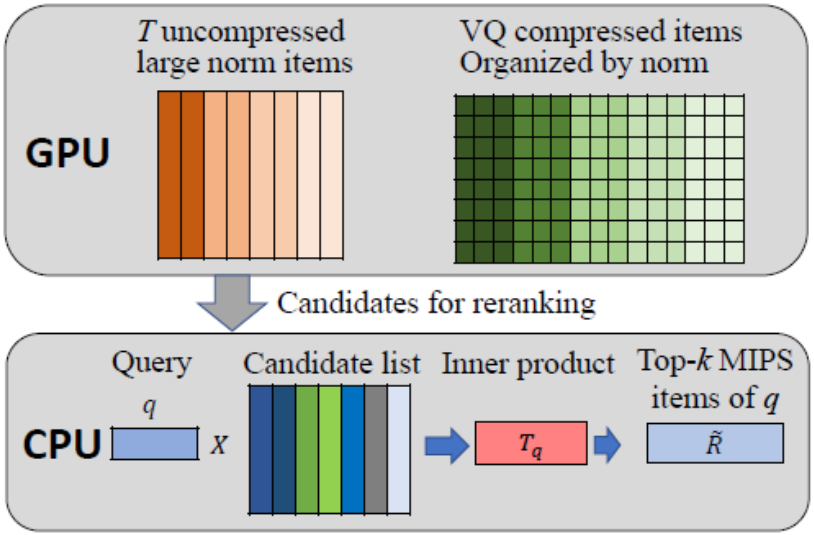


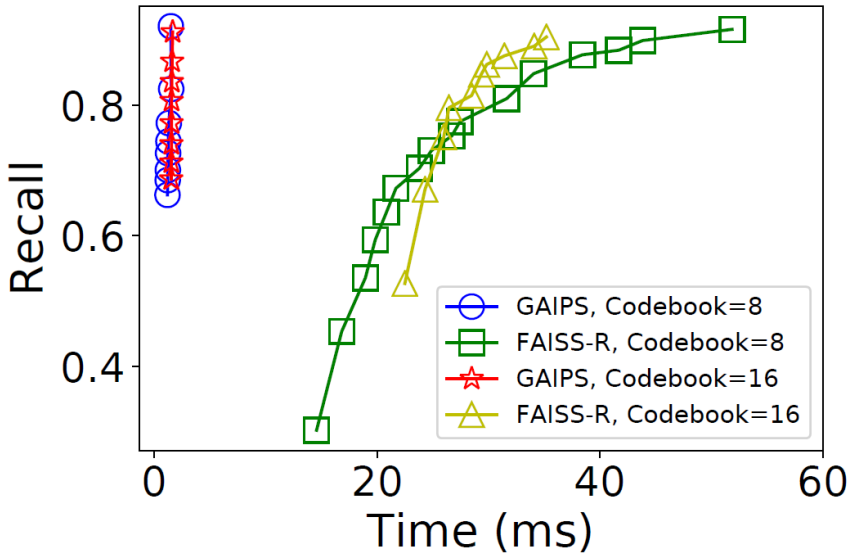




- **向量数据集规模大，精确距离计算是计算密集型任务**
  - 数据切分方案充分利用GPU计算性能
  - 发挥GPU计算效率，提前剪枝
  - 数据预处理提升剪枝效率

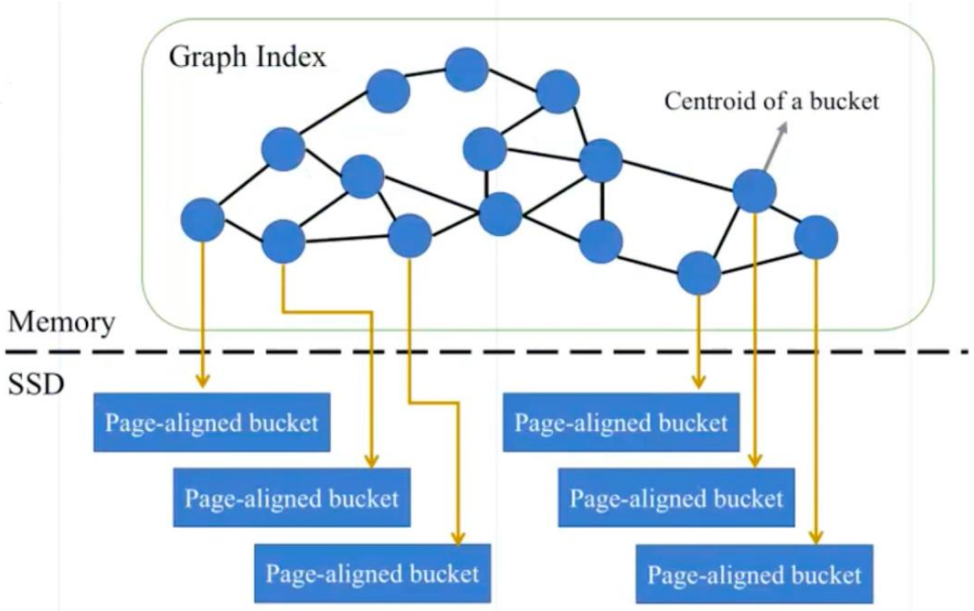
近似距离检索SIGIR 2021





- **Approximate Inner Product Retrieval**
  - 合适的索引设计，充分使用GPU并行度；大数据集对GPU的适配
  - 在线内积检索GAIPS性能优于FAISS
  - 向量量化节约GPU存储空间、有理论保证的量化误差剪枝

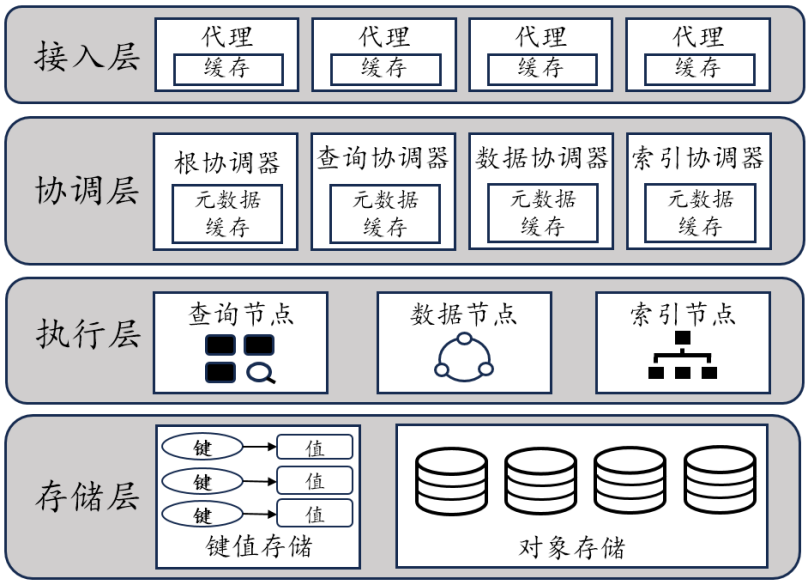
NeurIPS 2021竞赛冠军

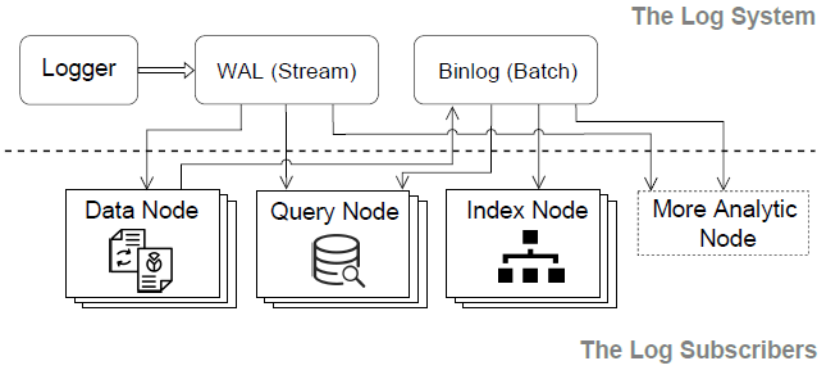


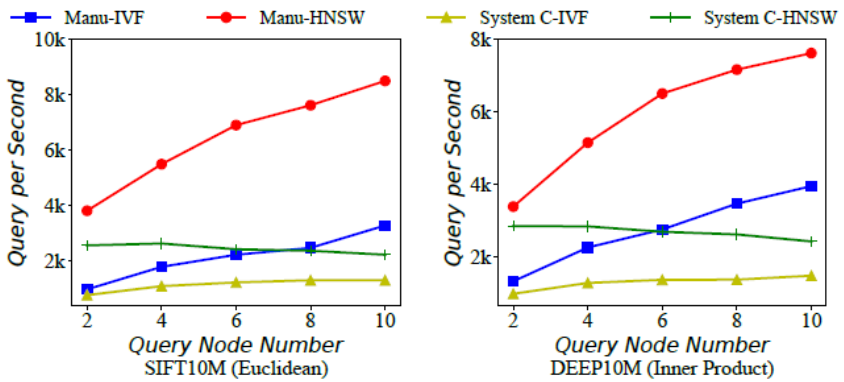
- **Billion-Scale Approximate Nearest Neighbor Search (亿级近似检索)**
  - SSD+RAM: 关注利用小内存和廉价SDD实现大规模数据集上的高效检索
  - 设计合适的索引结构，在减少磁盘读取的同时，保证高检索质量
  - Recall@1500QPS: 88.5% 远高于16.3% (SimSearchNet++ 1Billion), 12 hours for index
  - Recall@1500QPS: 49.5% 远高于48.8% (Text2image 1Billion), 28 hours for index

Manu系统架构 VLDB 2022

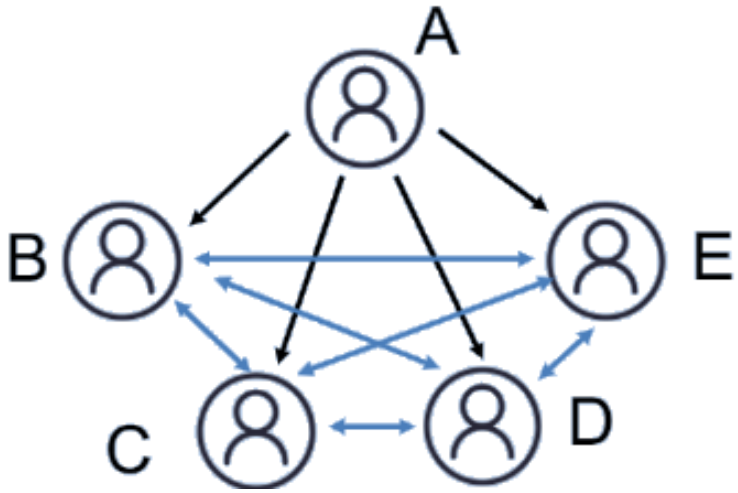
- **云原生向量数据库系统架构**
  - **关键设计：**基于微服务的架构，实现模块间松耦合和高弹性；log is the database
  - 业界首款云原生向量数据库Manu（VLDB'22），获OpenAI官方推荐

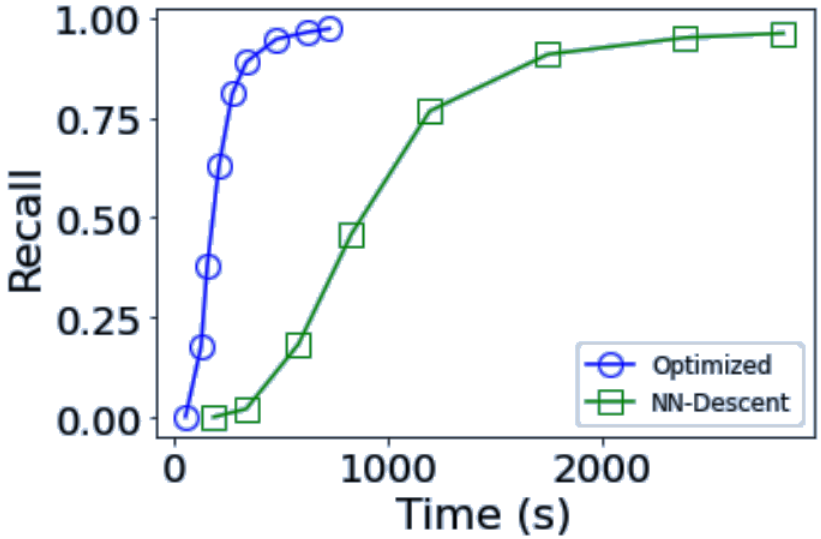






SIGMOD 2023竞赛全球第五





- **Build Approximate KNN Graph (检索构建), all pairs NN search**
  - 数据集大小: 10 million, d=100
  - 评测指标: Recall of 10K query points, 总共时间开销小于 30min
  - 核心思想: 邻居的邻居也是邻居，如何快速计算二跳邻居的距离
  - 关键技术: SIMD距离计算加速，Sorted Array与Heap的性能增益

反向前K查询 ICDE 2024

- **Reverse k-Ranks Query 反向前K查询处理**
  - 可以根据用户喜好构建其独有的提示工程，实现用户与大语言模型的高效协作
  - 现有Reverse k-Ranks Query 算法性能处理严重不同
  - QSRP实现数量级的性能提升
    - 基于采用的索引构建、基于回归的计算剪枝、查询感知的采样技术

Product set			
Inception($\mathbf{p}_1$)	(0.6,0.9)	Léon($\mathbf{p}_5$)	(0.3,0.6)
Devotion($\mathbf{p}_2$)	(0.2,2.7)	Iron Man($\mathbf{p}_6$)	(2.4,0.9)
Troll($\mathbf{p}_3$)	(2.1,1.8)	Titanic($\mathbf{p}_7$)	(0.9,0.0)
Smile($\mathbf{p}_4$)	(1.8,2.7)	Fight Club($\mathbf{q}$)	(2.7,0.6)

User set	
Bob($\mathbf{u}_1$)	(1.5,0.9)
Mary($\mathbf{u}_2$)	(2.7,1.2)
Alan($\mathbf{u}_3$)	(0.3,2.7)
Tom($\mathbf{u}_4$)	(1.2,0.0)
Jerry($\mathbf{u}_5$)	(0.9,1.2)

(a) User and product embeddings

Score		Sorted list
$f(\mathbf{u}_1, \mathbf{q})$	4.59	$\langle \mathbf{p}_4, 5.13\rangle, \langle \mathbf{p}_3, 4.77\rangle, \langle \mathbf{q}, 4.59\rangle, \langle \mathbf{p}_6, 4.41\rangle, \langle \mathbf{p}_2, 2.73\rangle, \langle \mathbf{p}_1, 1.71\rangle, \langle \mathbf{p}_7, 1.35\rangle, \langle \mathbf{p}_5, 0.99\rangle$
$f(\mathbf{u}_2, \mathbf{q})$	8.01	$\langle \mathbf{p}_4, 8.10\rangle, \langle \mathbf{q}, 8.01\rangle, \langle \mathbf{p}_3, 7.83\rangle, \langle \mathbf{p}_6, 7.56\rangle, \langle \mathbf{p}_2, 3.78\rangle, \langle \mathbf{p}_1, 2.70\rangle, \langle \mathbf{p}_7, 2.43\rangle, \langle \mathbf{p}_5, 1.53\rangle$
$f(\mathbf{u}_3, \mathbf{q})$	2.43	$\langle \mathbf{p}_4, 7.83\rangle, \langle \mathbf{p}_2, 7.35\rangle, \langle \mathbf{p}_3, 5.49\rangle, \langle \mathbf{p}_6, 3.15\rangle, \langle \mathbf{p}_1, 2.61\rangle, \langle \mathbf{q}, 2.43\rangle, \langle \mathbf{p}_5, 1.71\rangle, \langle \mathbf{p}_7, 0.27\rangle$
$f(\mathbf{u}_4, \mathbf{q})$	3.24	$\langle \mathbf{q}, 3.24\rangle, \langle \mathbf{p}_6, 2.88\rangle, \langle \mathbf{p}_3, 2.52\rangle, \langle \mathbf{p}_4, 2.16\rangle, \langle \mathbf{p}_7, 1.08\rangle, \langle \mathbf{p}_1, 0.72\rangle, \langle \mathbf{p}_5, 0.36\rangle, \langle \mathbf{p}_2, 0.24\rangle$
$f(\mathbf{u}_5, \mathbf{q})$	3.15	$\langle \mathbf{p}_4, 4.86\rangle, \langle \mathbf{p}_3, 4.05\rangle, \langle \mathbf{p}_2, 3.42\rangle, \langle \mathbf{p}_6, 3.24\rangle, \langle \mathbf{q}, 3.15\rangle, \langle \mathbf{p}_1, 1.62\rangle, \langle \mathbf{p}_5, 0.99\rangle, \langle \mathbf{p}_7, 0.81\rangle$

(b) User ranks for products

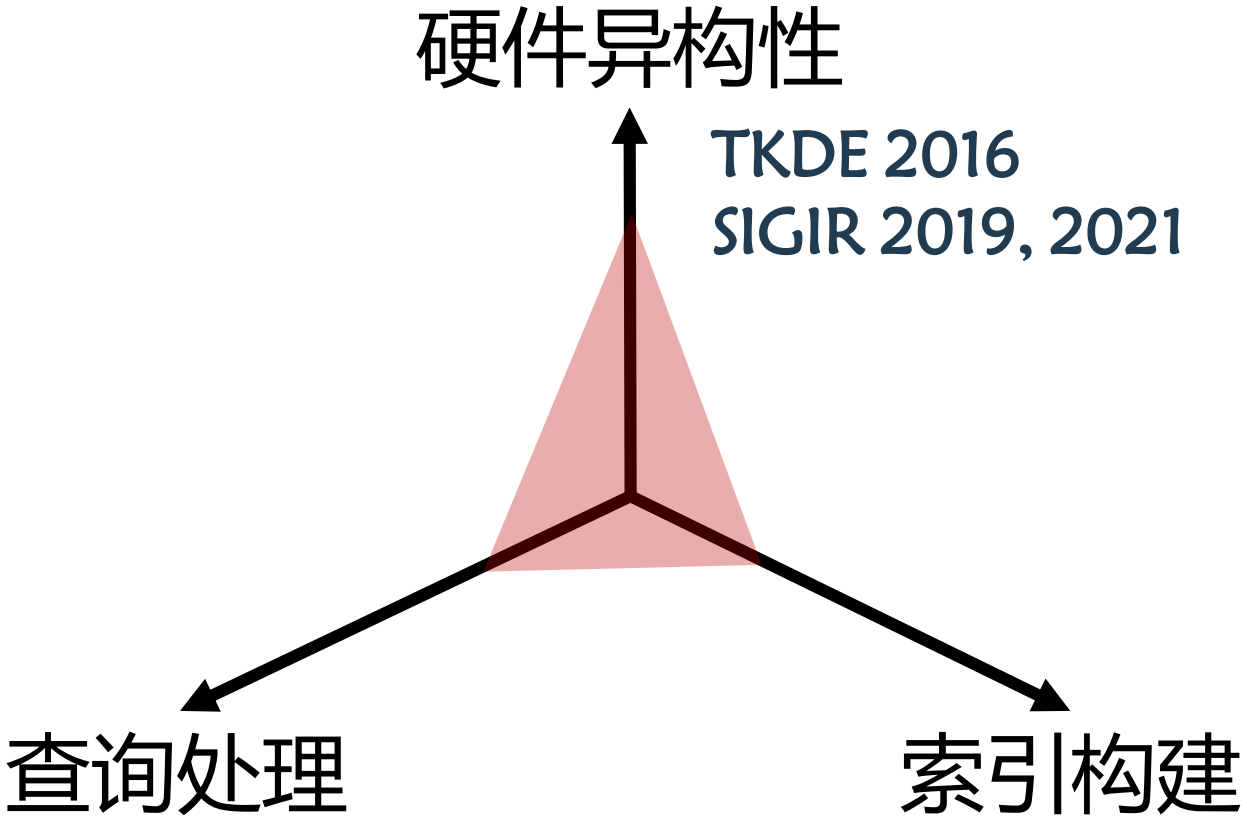
SIGMOD 2024竞赛全球冠军

Rank	Team	Best Recall	Runtime(s)
#1	Alaya (SUSTech and ZJU)	1.0000	1160
#2	wahaha (Xiamen University)	0.9998	1066
#3	BitBigBang (Xidian University and The Chinese University of Hong Kong)	0.9998	1172
#4	biejuanle (SUSTech and ZJU)	0.9998	1177
#5	Daily_Lab	0.9995	1188
#6	nju_lands (Nanjing University)	0.9994	1127

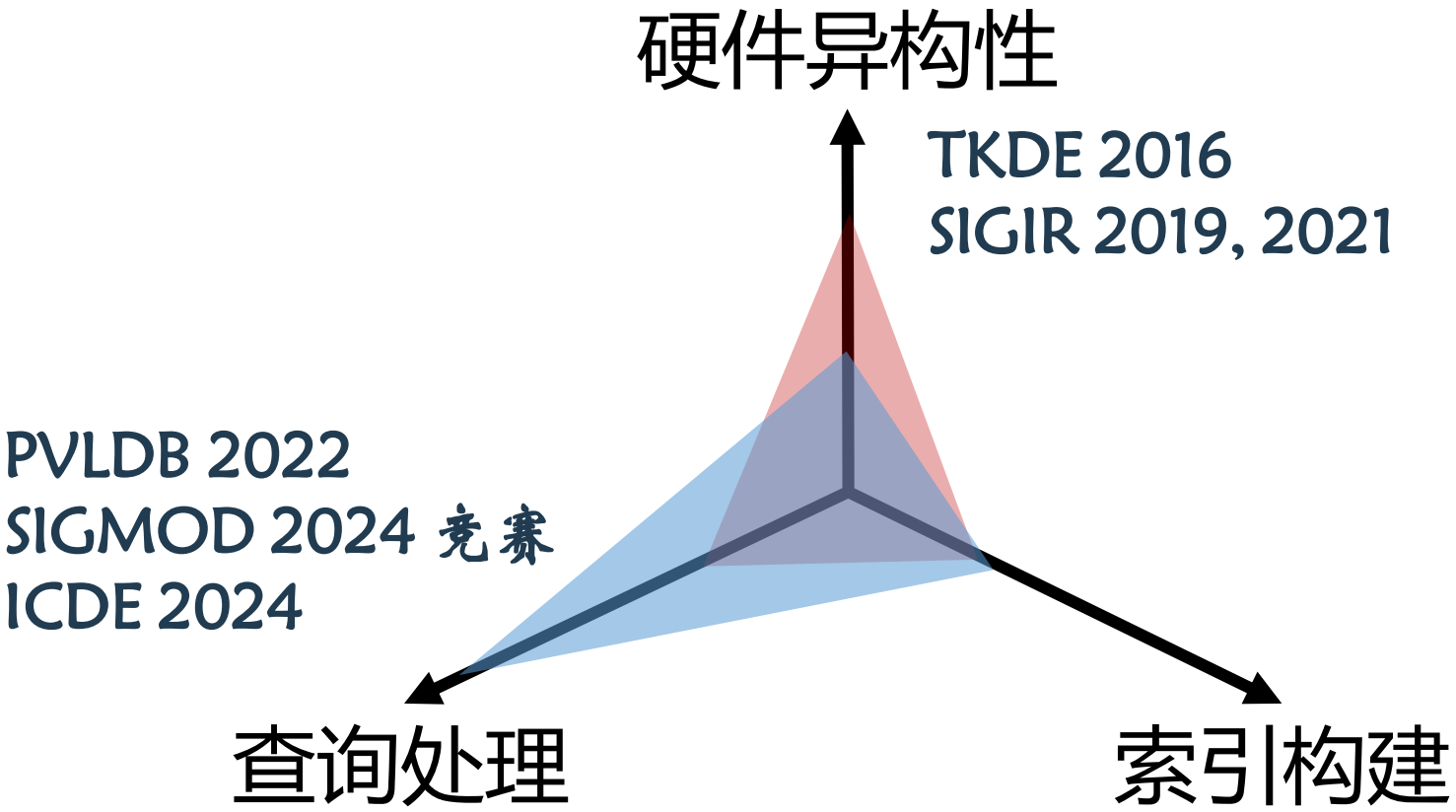
Hybrid Vector Search (混合向量检索)

- 纯向量检索，属性过滤向量检索，向量与时间范围检索，向量+属性+时间范围
- 评测指标: Recall 和 Runtime
- 核心思想: 基于属性划分建立多临近图索引
- 突出指标: 检索精度突破4个9，即0.99995以上

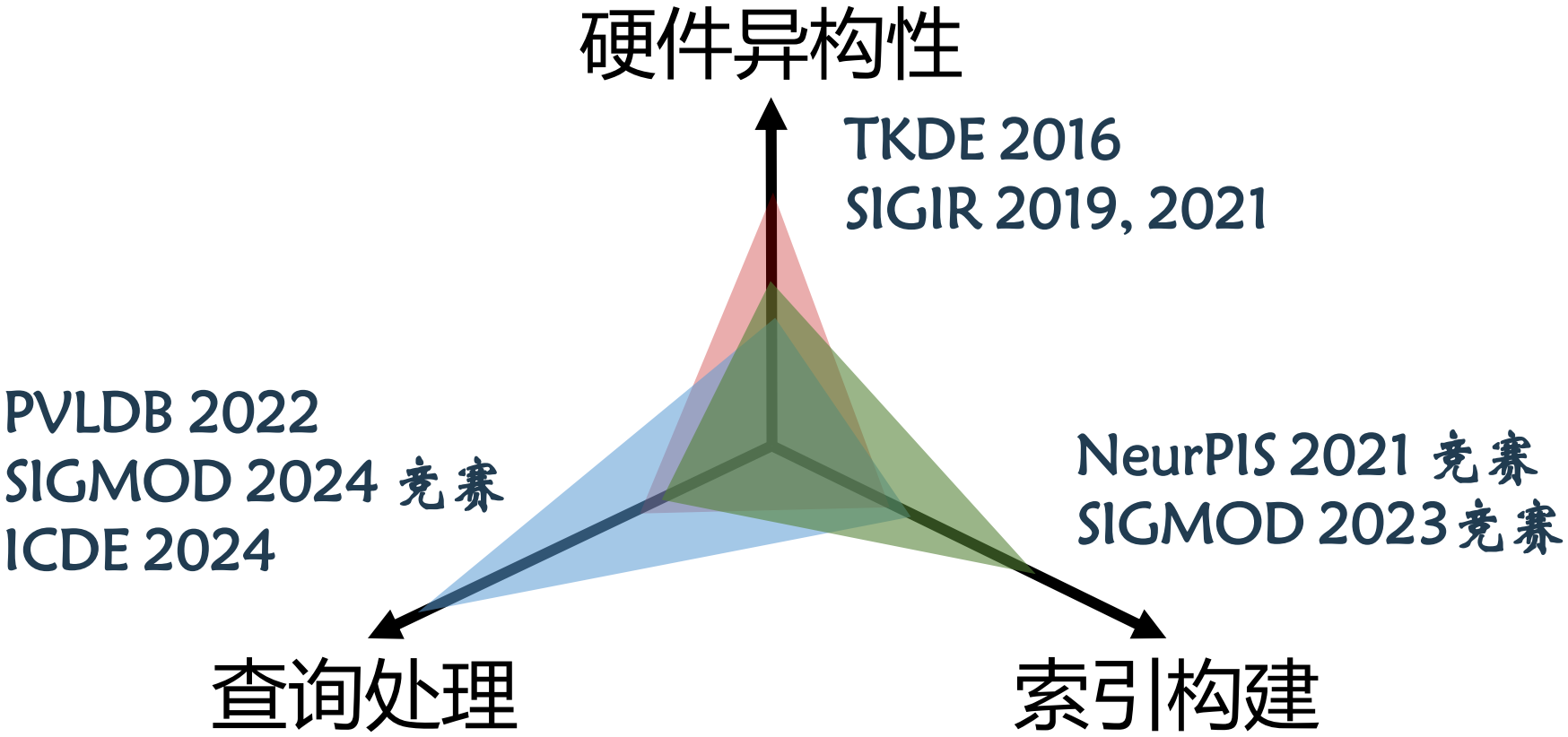
研究工作总结



研究工作总结


硬件异构性
TKDE 2016
SIGIR 2019, 2021
PVLDB 2022
SIGMOD 2024 竞赛
ICDE 2024
查询处理
索引构建

研究工作总结

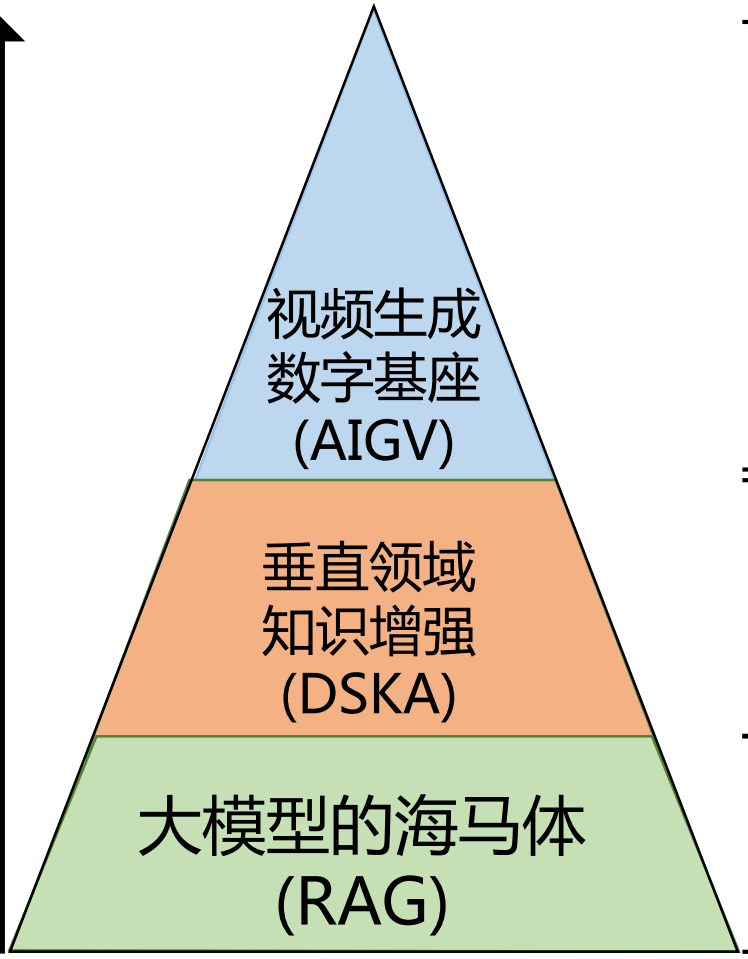

硬件异构性
TKDE 2016
SIGIR 2019, 2021
PVLDB 2022
SIGMOD 2024 竞赛
ICDE 2024
NeurPIS 2021 竞赛
SIGMOD 2023 竞赛
查询处理
索引构建

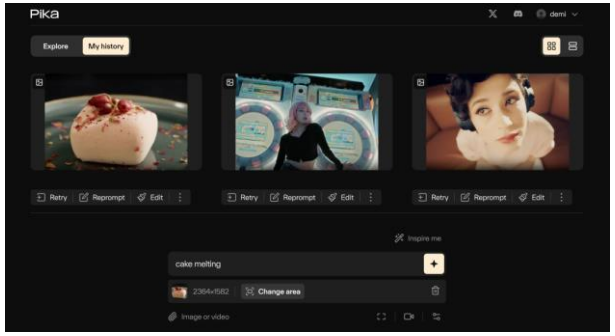
向量数据库系统汇报提纲

- ✓ 研究历史: 千呼万唤始出来，犹抱琵琶半遮面
  - 追本溯源、推波助澜、登峰造极
- ✓ 系统特性: 不畏浮云遮望眼，只缘身在此山中
  - 核心功能、系统组件、关键技术
- ✓ 研究积累: 衣带渐宽终不悔，为伊消得人憔悴
  - 十年耕耘、两个冠军、多篇顶会
- ✓ 未来挑战: 问渠那得清如许，为有源头活水来
  - 百亿千维、端侧服务、软硬协同

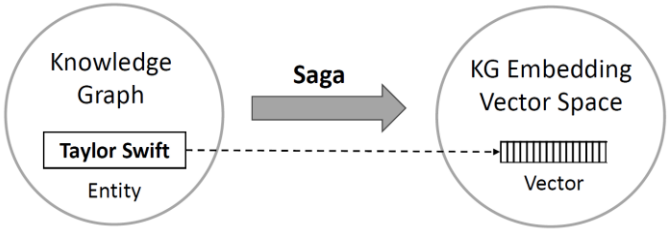
应用驱动技术创新

向量数据库系统应用研究热度

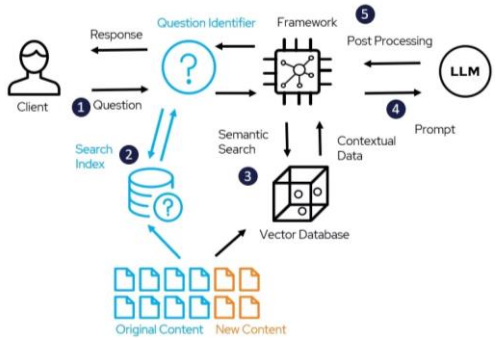




视频生成数字基座，鲜有讨论



基于向量的知识图谱构建,方兴未艾

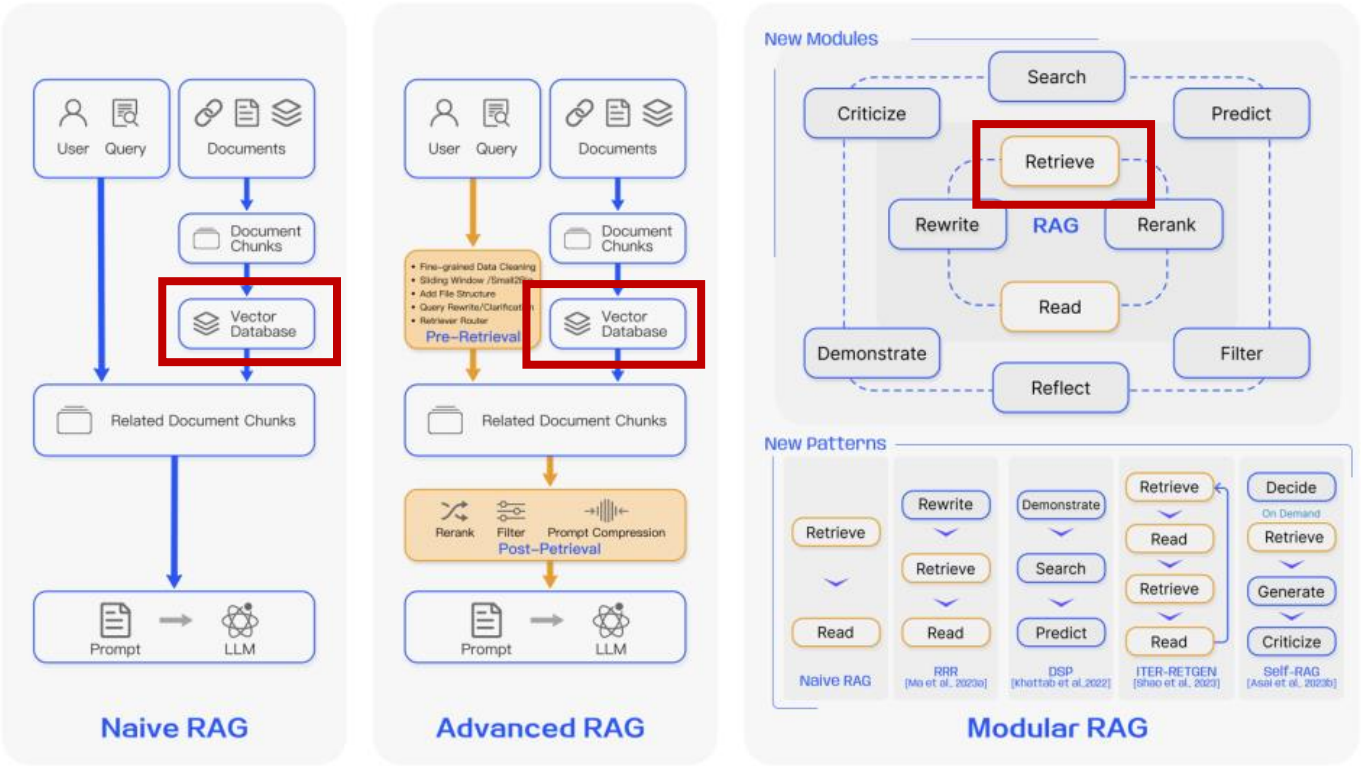


向量数据库最火应用,没有之一

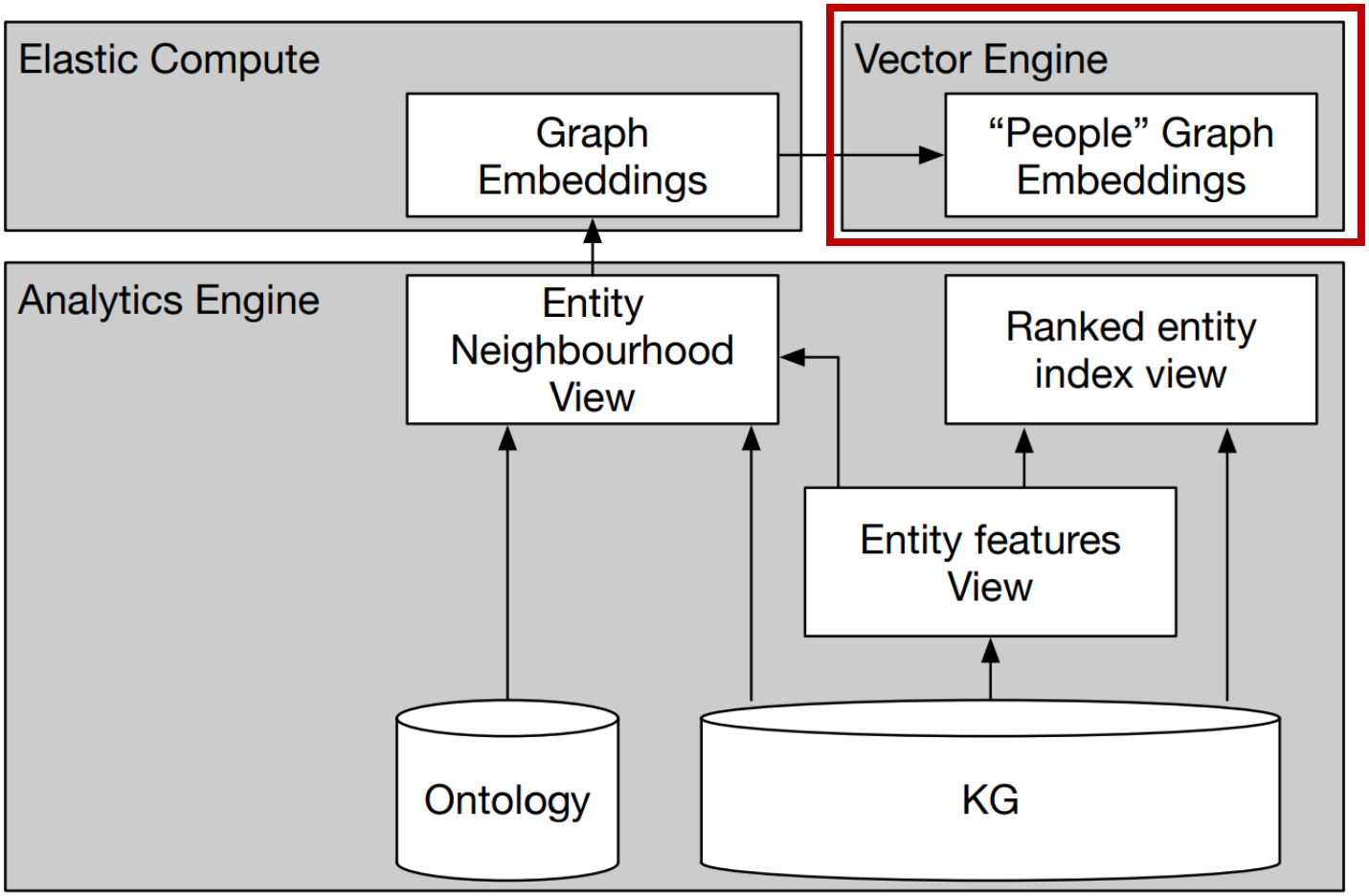
技术挑战：RAG（百亿千维）

5000万tokens大海捞针创纪录，百川智能192K超长上文+搜索增强
破解商用难题！解决99%企业定制需求

Original 新智元 新智元 2023-12-21 12:59 Posted on 北京



技术挑战：DSKA（高准确率）



Apple Research: Saga (SIGMOD 2022工业轨):

基于向量数据库的知识图谱持续构建和大规模知识服务

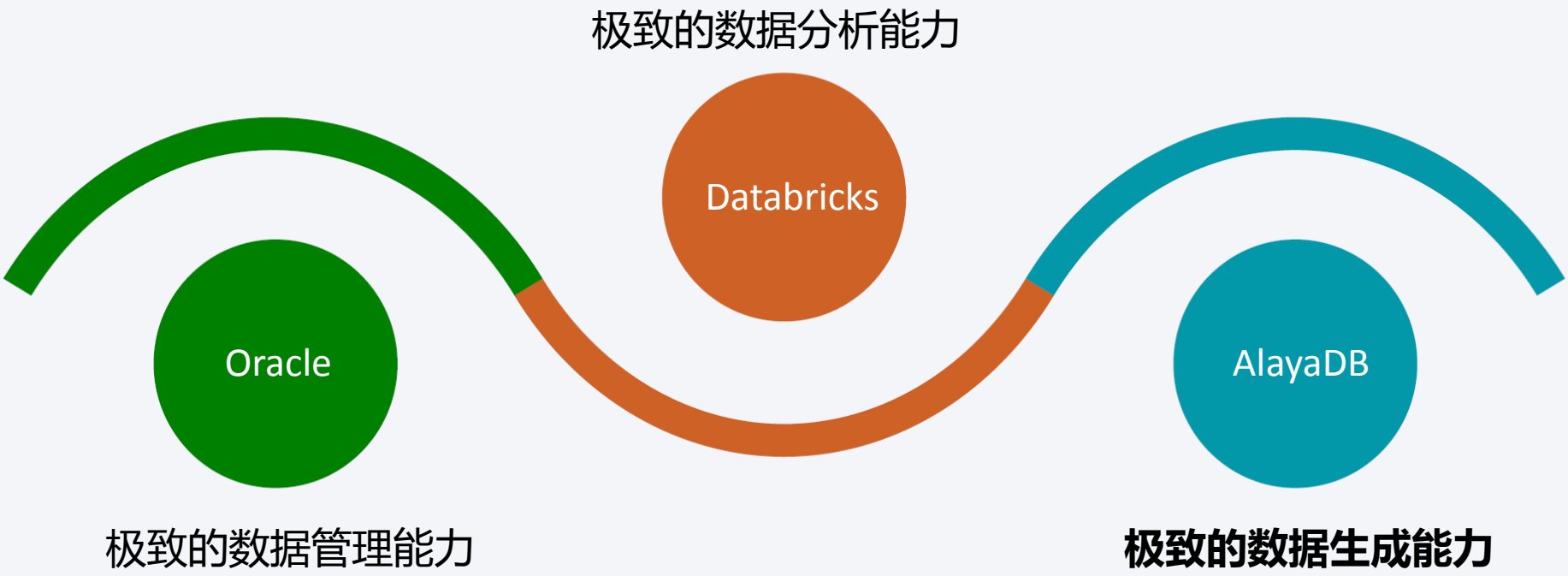
技术挑战：AIGV（端侧服务）

技术引领未来

人机协作创新力量

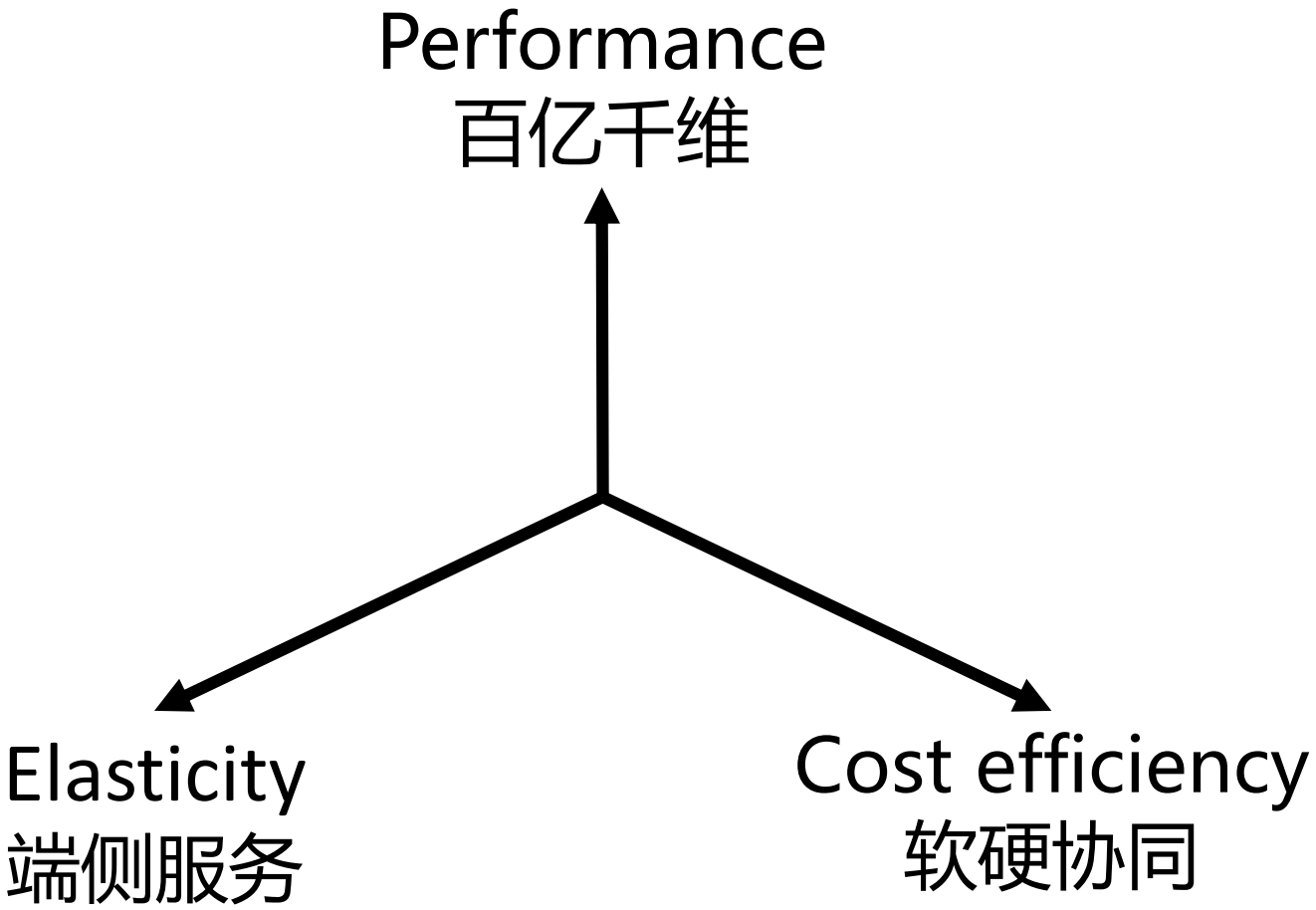
基于信仰克服挑战

下一个时代



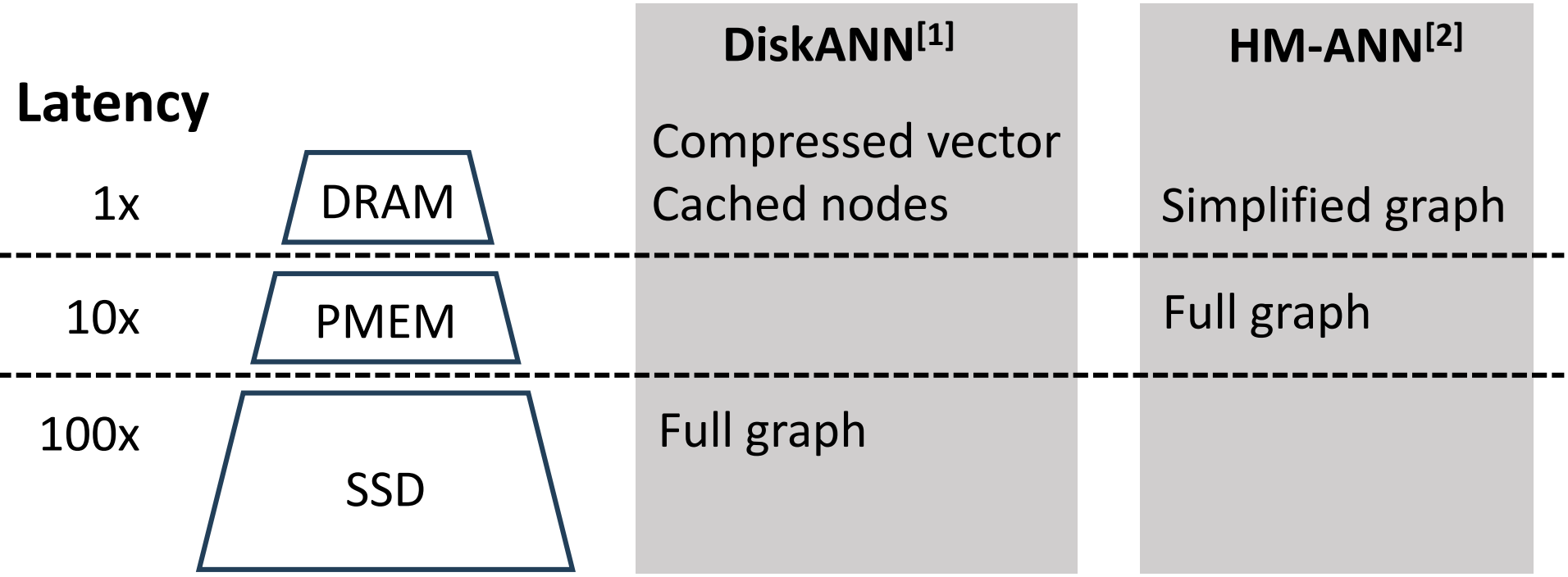
大模型时代亟需全新的数据库系统!

未来挑战



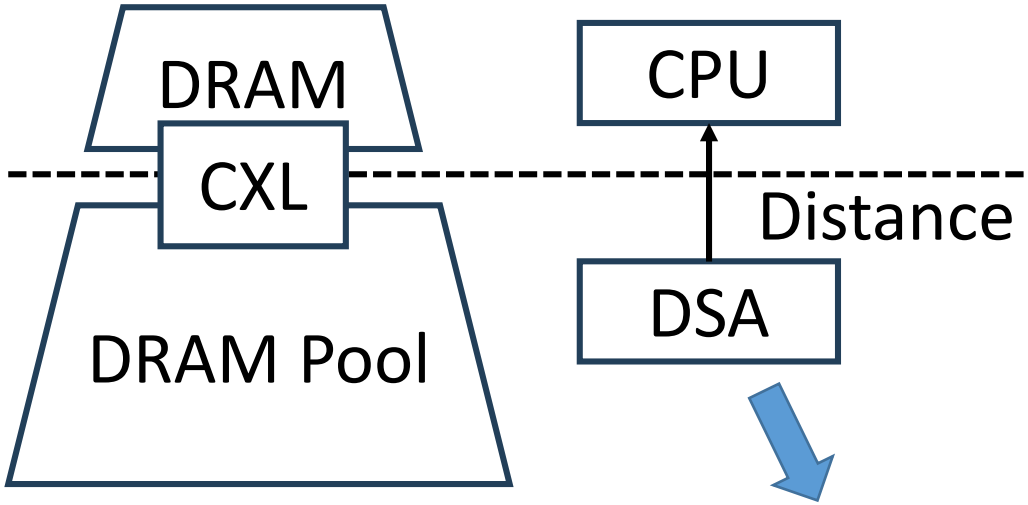
百亿千维：纯软方案

Hierarchical Approach



[1] DiskANN: fast accurate billion-point nearest neighbor search on a single node. NIPS'19.
[2] HM-ANN: efficient billion-point nearest neighbor search on heterogeneous memory. NIPS'20.

百亿千维：软硬结合

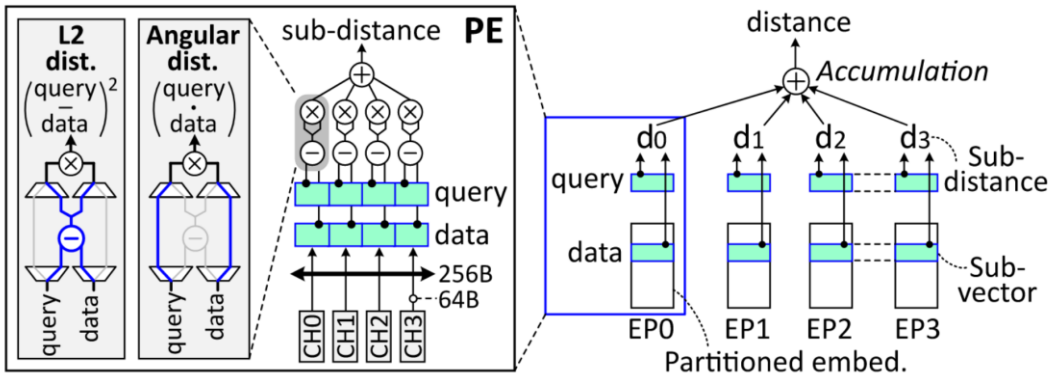


Software Design

Cached Nodes

Graph & Vectors

Hardware Design



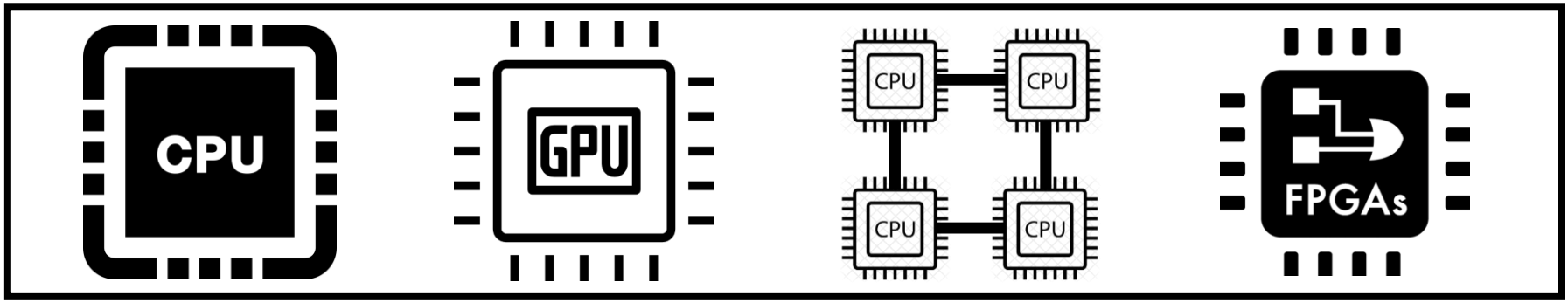
[1] CXL-ANNS: Software-Hardware Collaborative Memory Disaggregation and Computation for Billion-Scale Approximate Nearest Neighbor Search, ATC'23

系统架构重思考

- **中央处理器:** CPU is the first-class citizen, it is not true in vector database
- **近数据计算范式:** Data is the first-class citizen, distance computation should offload

摩尔定律已经失效!

--- 英伟达CEO黄仁勋, CES2019



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

DBGroup

谢谢！

DBGroup @ SUSTech

Dr. Bo Tang (唐博)

tangb3@sustech.edu.cn