向量数据库：大模型时代的基础设施构建

刘力 Zilliz

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

演讲嘉宾

刘力

Zilliz 首席工程师

Zilliz 首席工程师，拥有多年的数据库，大数据等方向的开发经验，目前在Zilliz负责查询索引相关的研发工作。曾于Meta就任高级工程师，负责广告流式数据框架的设计和开发工作。刘力拥有卡内基梅隆大学信息技术硕士学位。

目录
CONTENTS

PART 01
什么是向量数据库

什么是向量数据库

ID	Type	Desprtion	Author
0	Image	A Short History of Nearly Everything	Bill Bryson
1	Text	Silent Spring	Rachel Carson
2	Video	Holes	Louis Sachar
…			

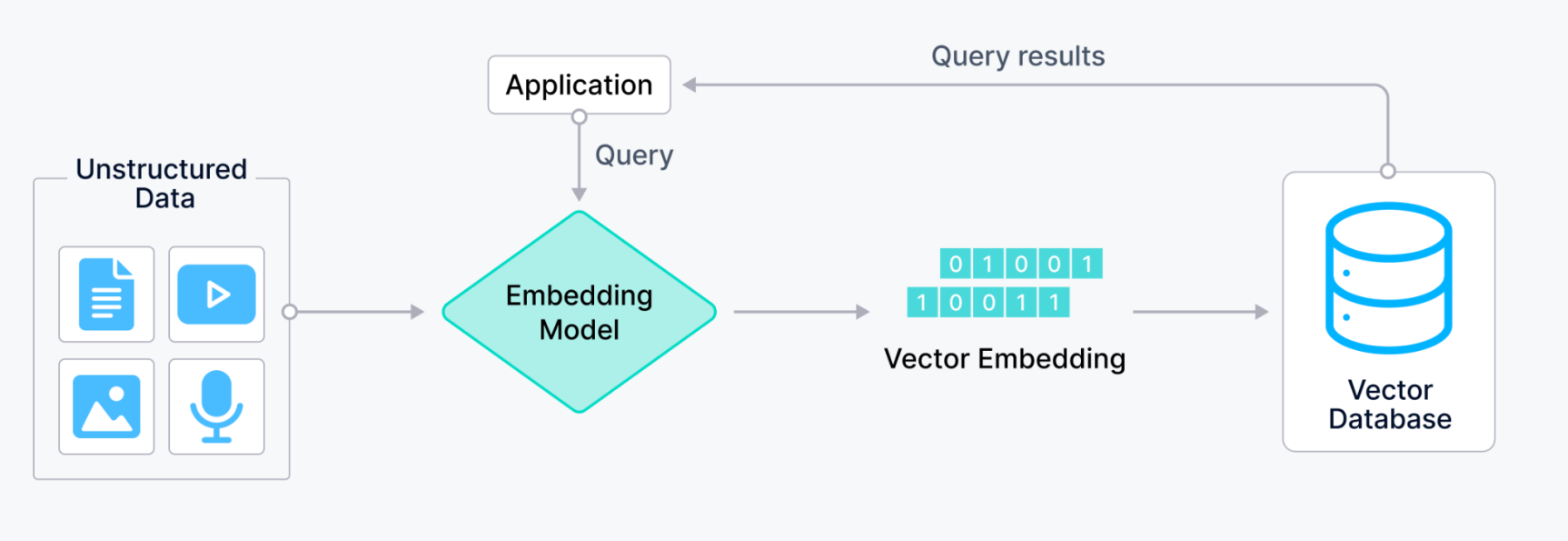
Data UID[1]	Vector representation
0	[-0.31, 0.53, -0.18, …, -0.16, -0.38]
1	[0.58, 0.25, 0.61, …, -0.03, -0.31]
2	[-0.07, -0.53, -0.02, …, -0.61, 0.59]
…	

- 过去人们通过关系型存储检索数据，这种方式无法模糊匹配，无法跨模态检索，缺少对上下文的理解
- 随着大模型的泛化能力变强，通过预训练学习数据的基本特征，利用深度学习模型提取 Embedding 用于数据检索的范式越来越常见

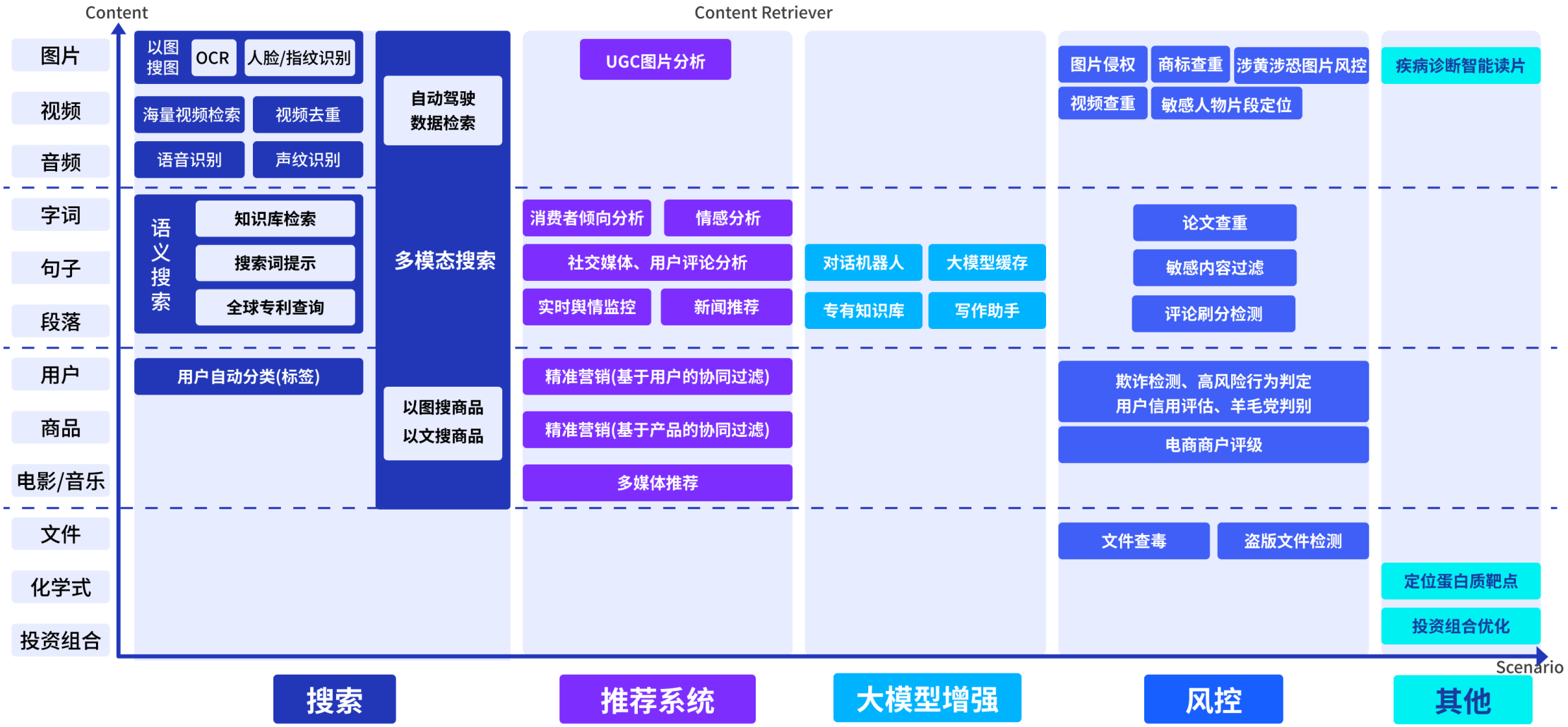
什么是向量数据库

向量检索利用向量在高维空间的距离来表征非结构化数据的相似度

向量数据库是一种专为存储和查询高维度向量数据而优化的数据库系统。



向量数据库的场景


Content
Content Retriever
图片
视频
音频
字词
句子
段落
用户
商品
电影/音乐
文件
化学式
投资组合
以图搜图
OCR
人脸/指纹识别
海量视频检索
视频去重
语音识别
声纹识别
语义搜索
知识库检索
搜索词提示
全球专利查询
用户自动分类(标签)
自动驾驶数据检索
多模态搜索
以图搜商品
以文搜商品
UGC图片分析
消费者倾向分析
情感分析
社交媒体、用户评论分析
实时舆情监控
新闻推荐
精准营销(基于用户的协同过滤)
精准营销(基于产品的协同过滤)
多媒体推荐
对话机器人
大模型缓存
专有知识库
写作助手
图片侵权
商标查重
涉黄涉恐图片风控
视频查重
敏感人物片段定位
论文查重
敏感内容过滤
评论刷分检测
欺诈检测、高风险行为判定
用户信用评估、羊毛党判别
电商商户评级
文件查毒
盗版文件检测
疾病诊断智能读片
定位蛋白质靶点
投资组合优化
Scenario
搜索
推荐系统
大模型增强
风控
其他

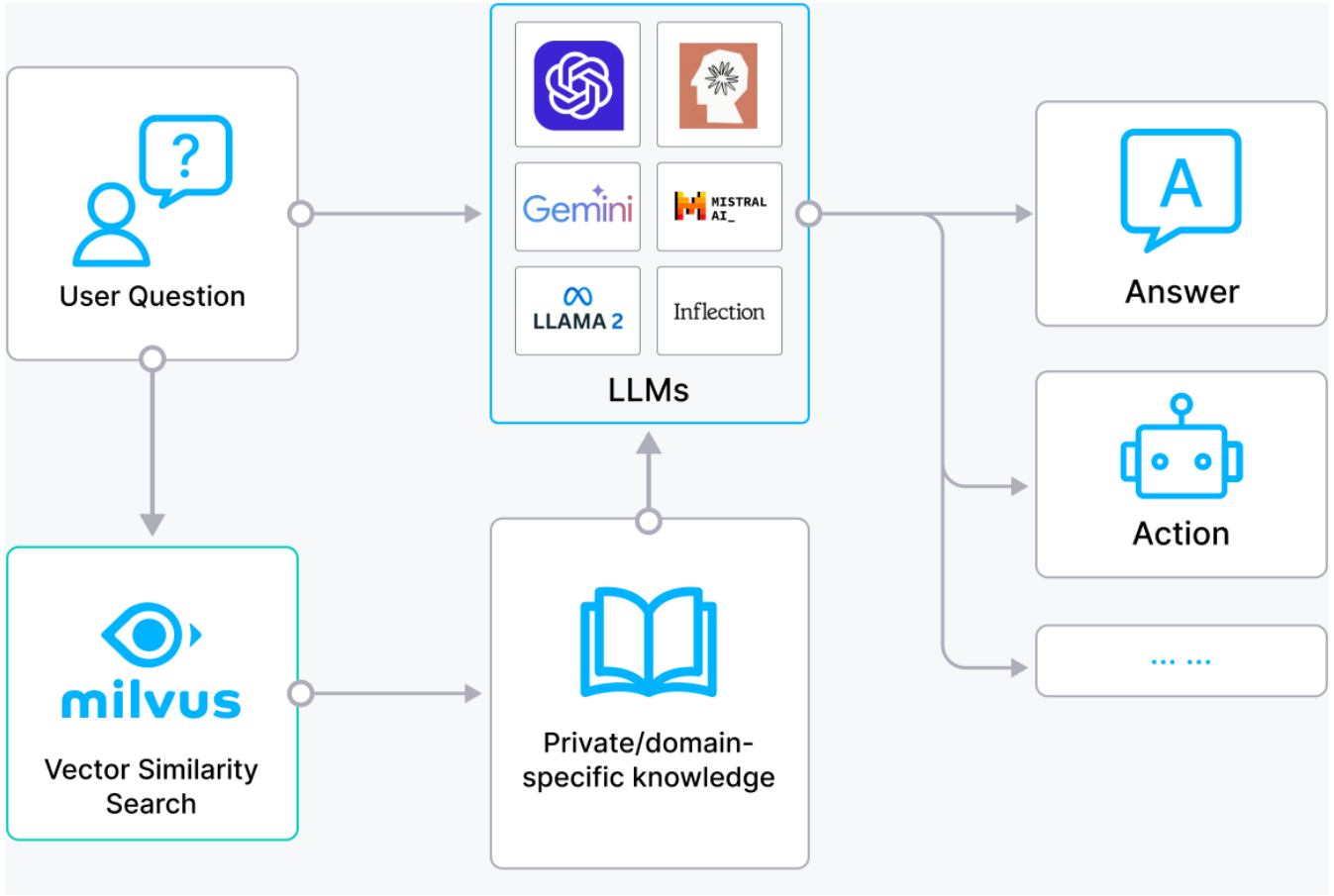
Retrieval Augmented Generation(RAG)

Prompt解决的LLM的幻觉：

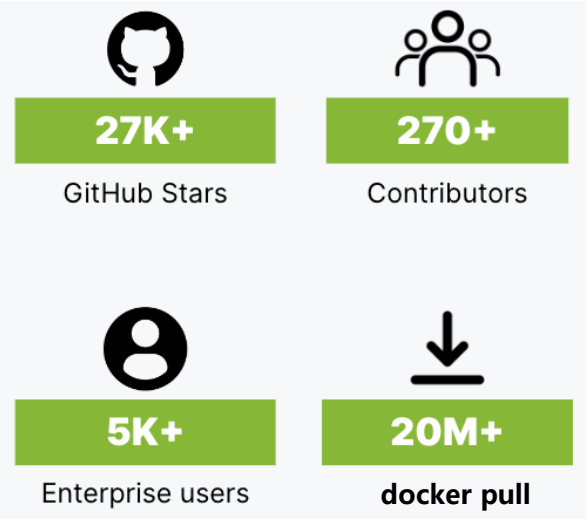
- 提高准确性和相关性
- 提升知识的实时性

通过RAG实现Prompt工程：

- 更高的效率和更低的成本
- 提供实时更新能力
- 提供私有/特定领域的知识



Milvus: 全球第一款向量数据库



中国平安 PING AN
招商银行 CHINA MERCHANTS BANK
华泰证券 HUATAI SECURITIES
中国建设银行 China Construction Bank
中国太平 CHINA TAIPING
泰康人寿
Walmart
IKEA
得物
Shopee
唯品会 品牌特卖
tokopedia
iQIYI 爱奇艺
xiaomi
Baidu 百度
快手
vivo
新浪微博 weibo.com
汽车之家 看车·买车·用车
SF TECHNOLOGY 顺丰科技
理想
哈啰
mobileye An Intel Company
WPS
MOMO
TREND MICRO
有道 youdao
INTUIT
shutterstock
COMPASS
企查查 Qcc.com
SmartNews
Hisense
Haier
dailyhunt
丁香园 WWW.DXY.CN

PART 02

从向量检索到Milvus

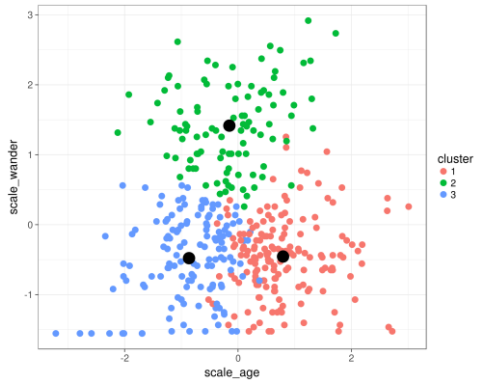
向量数据库的基石：向量索引

ANNS：Approximate Nearest Neighbor Search

利用预先插入的数据获取数据分布

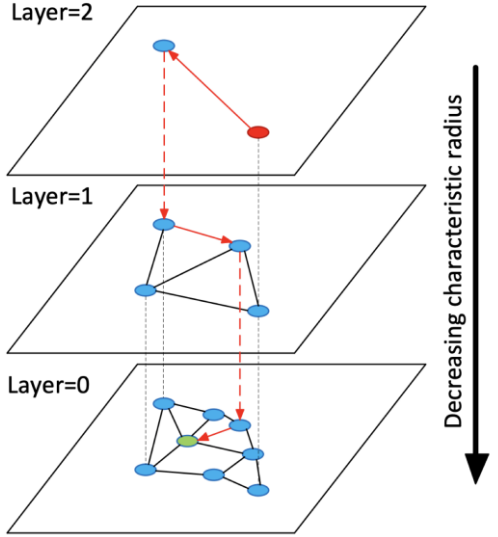
利用Graph，聚类，Hash等方式快速筛选并接近目标位置

利用SQ,PQ量化和SIMD降低单次距离的成本



IVF，SCANN，...

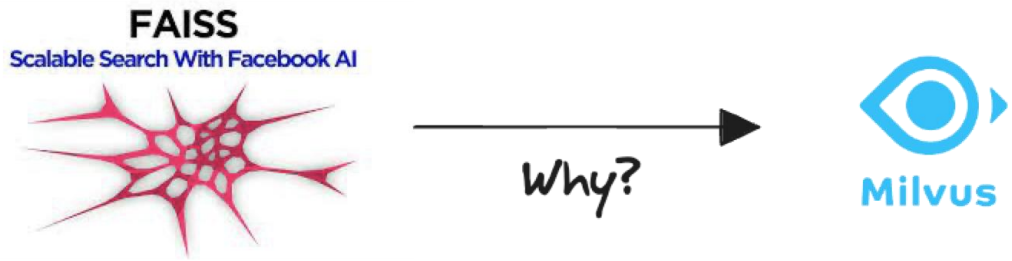
基于桶的ANN算法



HNSW，DiskANN，...

基于图的ANN算法

为什么需要向量数据库



Milvus的目的：更多，更快

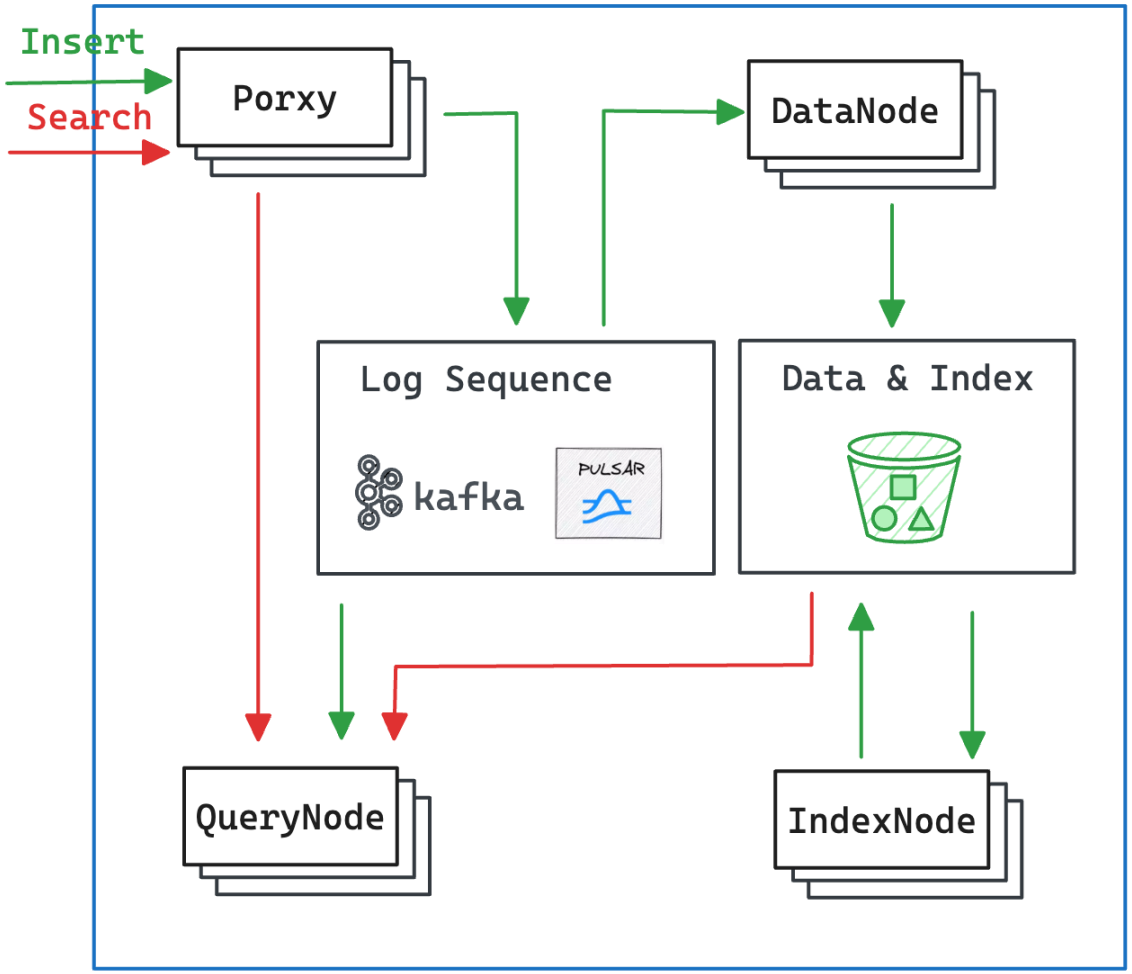
功能上：

- 增删改查
- 复杂查询：多向量、标向量混合等
- 查询一致性
- 多租户
- 监控
- RBAC

可用性上：

- 规模、性能的可扩展性
- 故障恢复能力
- 数据的备份，迁移与导入
- 微服务环境下的快速部署

更多：分布式架构支持百亿数据



设计理念：

- **存算分离**

  提供更好的可扩展性，资源管理能力和隔离性

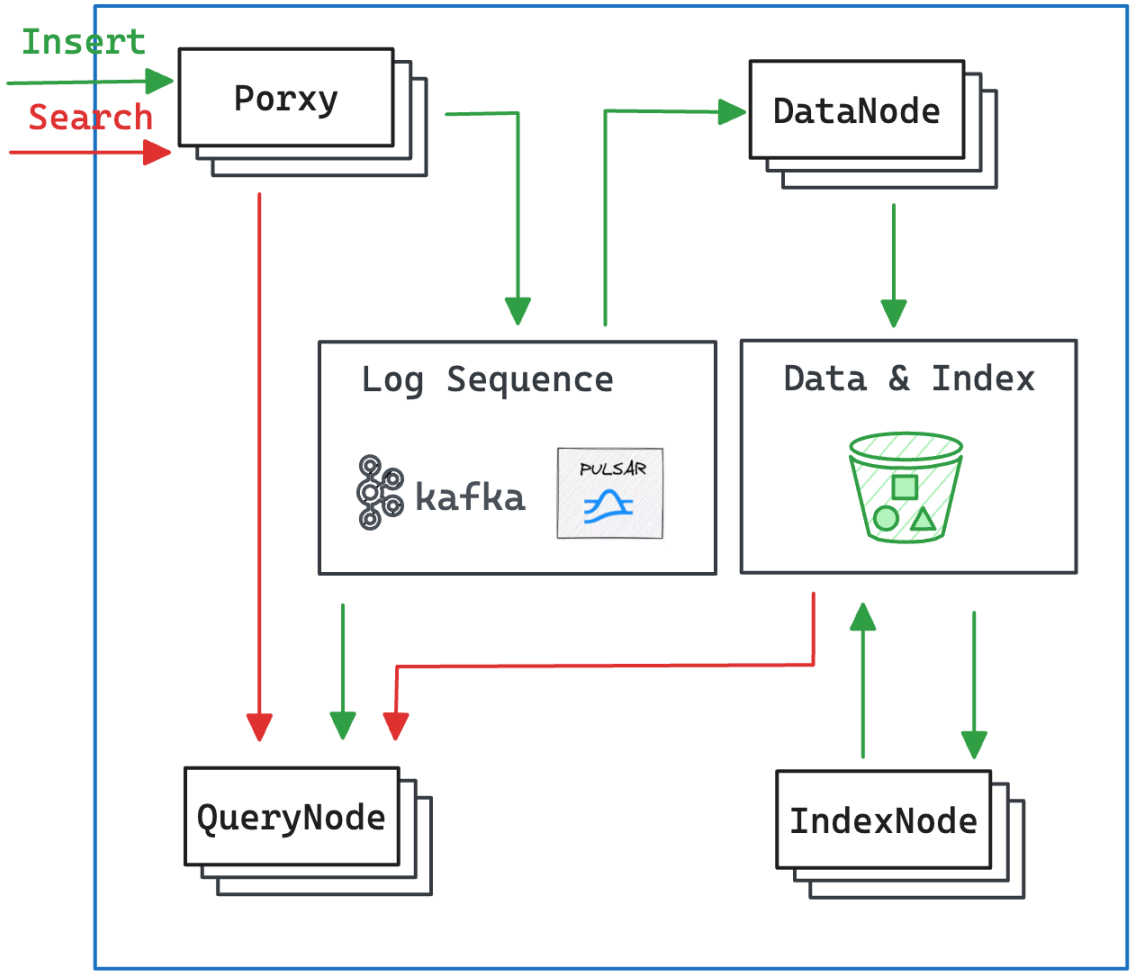
- **微服务化 + K8s**

  自动化部署，扩展

- **消息队列作为数据骨架：日志即数据**

  解耦不同组件，并且提供简单快捷的故障恢复机制

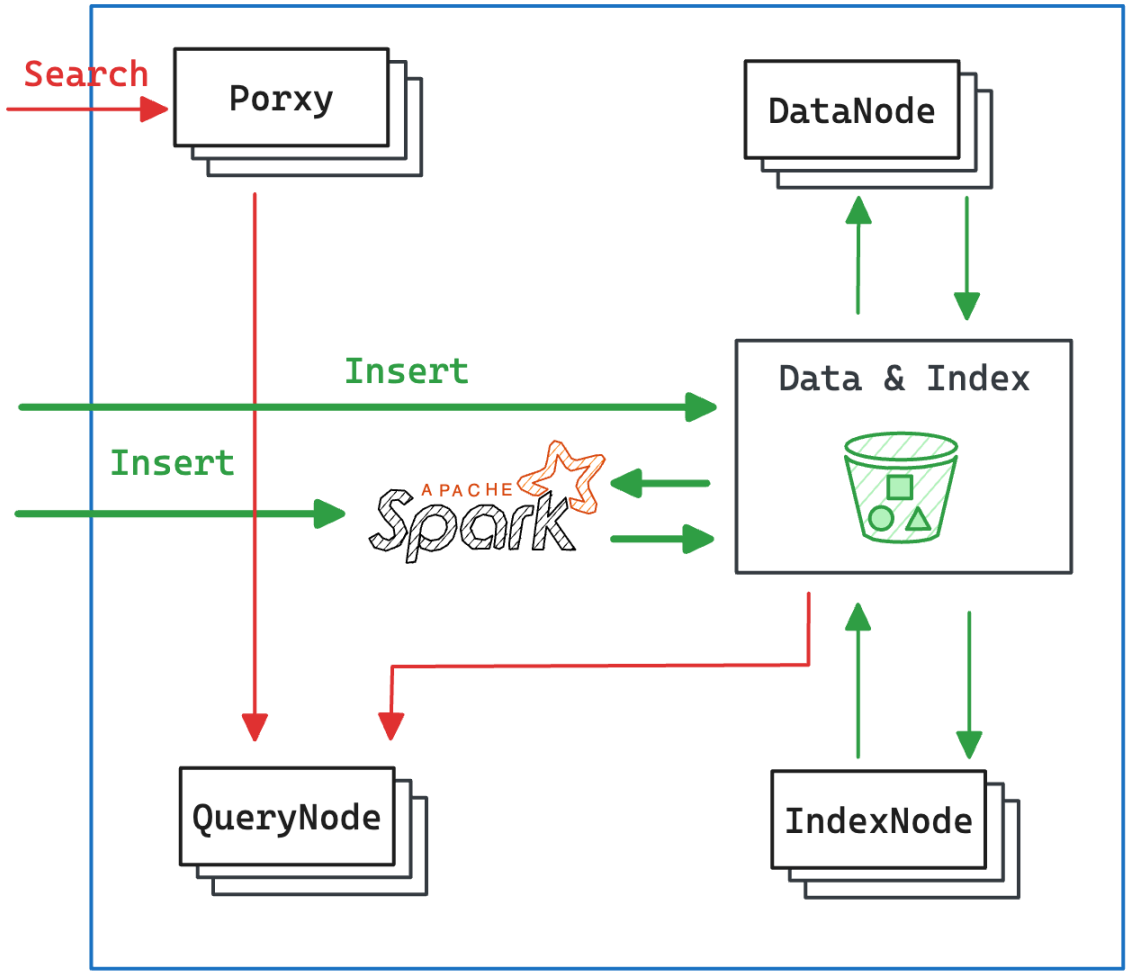
更多：分布式架构支持百亿数据



存算分离：

- **可扩展性**：储存和计算节点可以按需独立扩展
- **资源利用**：根据不同类型节点定制化的资源让系统能力更强
- **隔离性**：某个组件的升级、故障或者热点任务不会影响其他组件
- **池化**：IndexNode、和DataNode可以池化以提高资源利用率

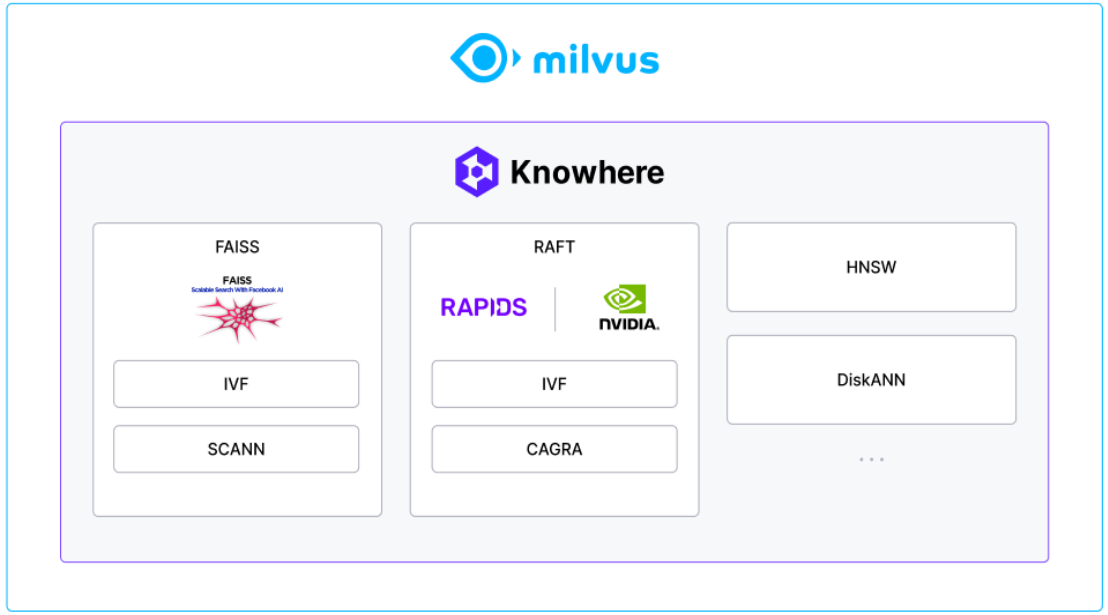
更多：分布式架构支持百亿数据



流批一体：

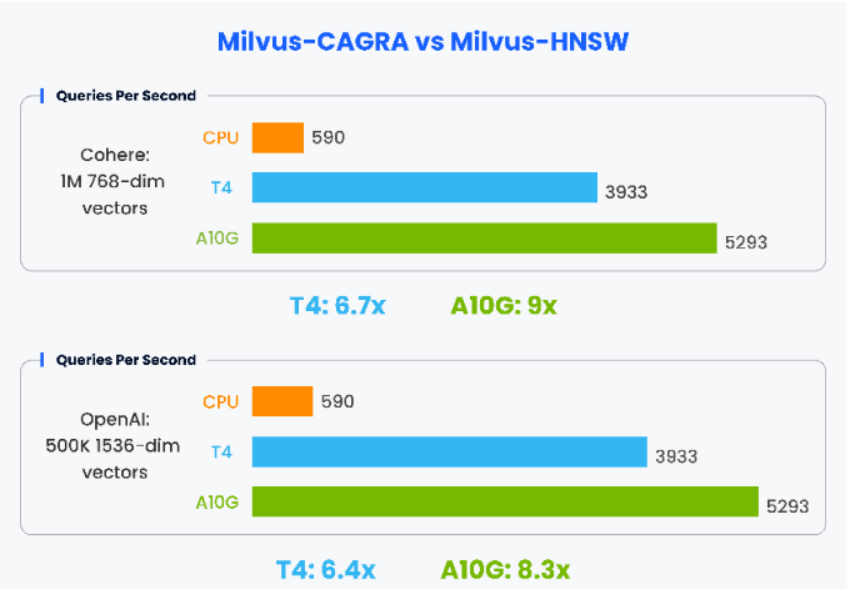
- **批量导入**：数据跳过复杂的流式系统，直接插入到对象储存
- **向量ETL**：数据可以在Spark中进行预处理（数据清洗、向量提取等）再批量导入到Milvus
- **全局优化**：数据可以从Milvus全量导出到Spark进行基于全局数据分布的优化再导回Milvus提供更高效的服务

更快：多种索引算法助推性能起飞

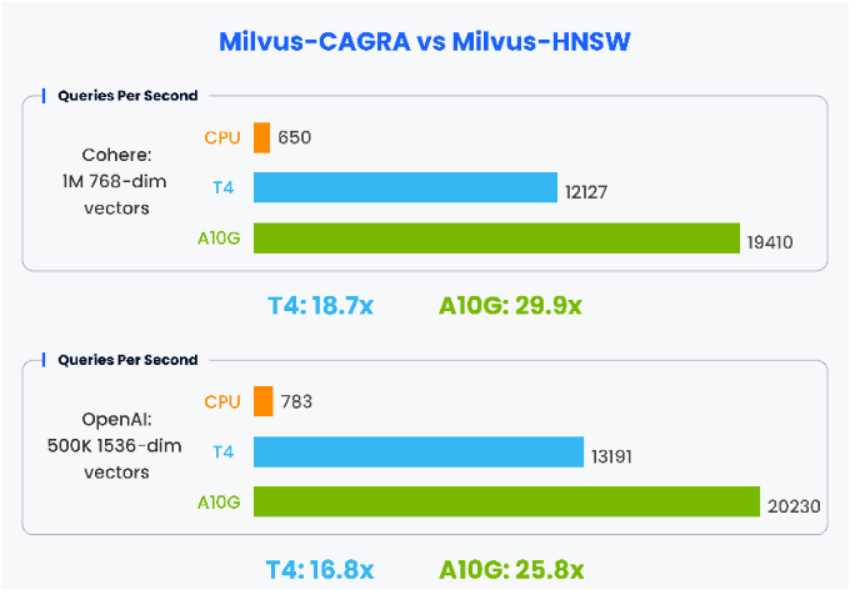


- 可插拔向量索引引擎，集成多种业界常用索引，支持不同的场景。
- 支持x86、ARM各种架构，并针对新老机型进行了SIMD调优。
- 和英伟达合作，支持使用GPU获得超高性能
- 对复杂搜索场景进行算法调优，比如过滤搜索

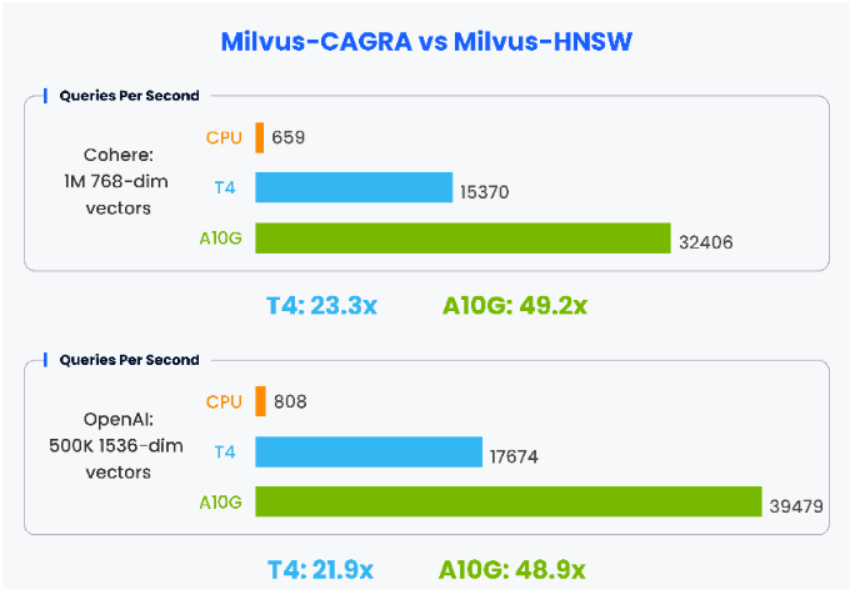
更快：GPU支持获得性能巅峰



Batch Size = 1



Batch Size = 10

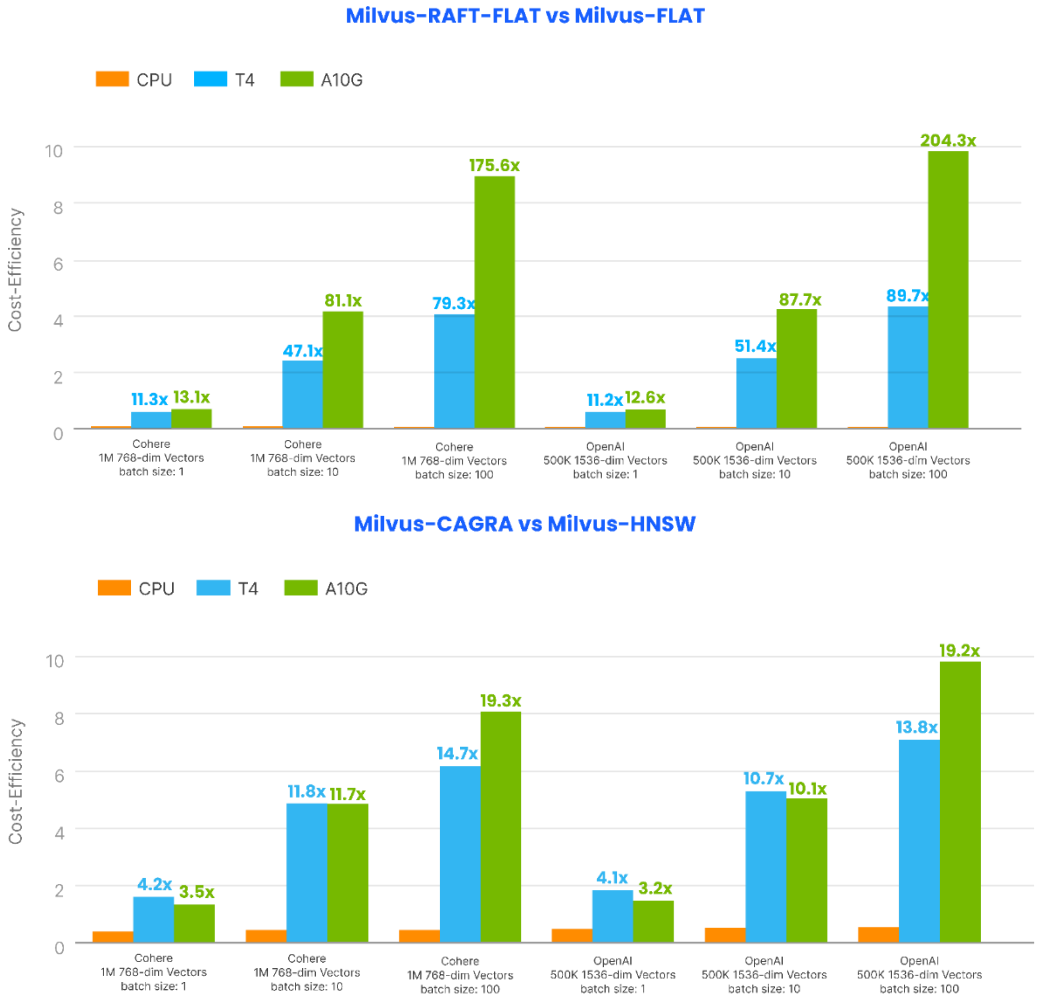


Batch Size = 100

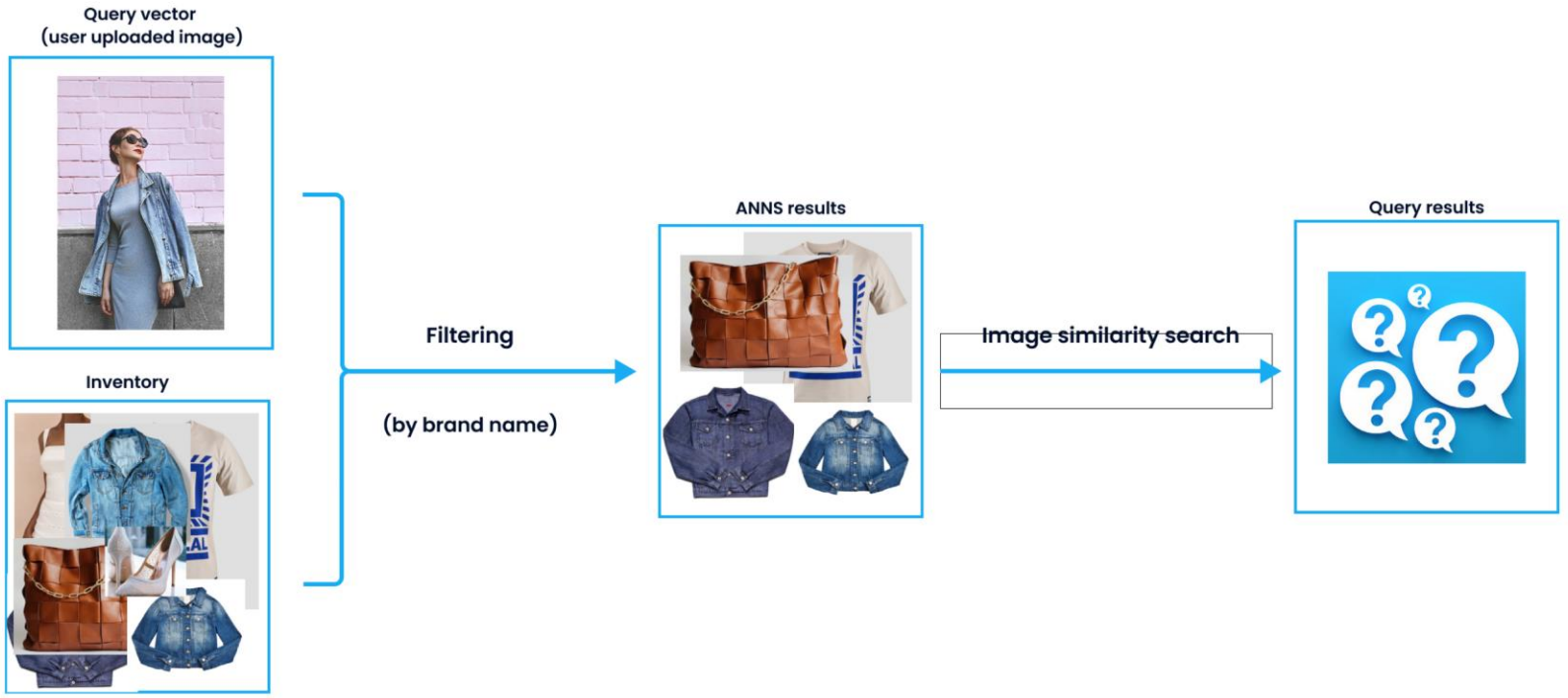
CPU: m6id.2xlarge T4: g4dn.2xlarge A10G: g5.2xlarge Top 100 Recall: 98%
Dataset:: https://github.com/zilliztech/VectorDBBench

更快：GPU支持获得性能巅峰

	Instance type	Price($/h)
T4	g4dn.2xlarge	0.752
A10G	g5.2xlarge	1.212
CPU	m6id.2xlarge	0.4746



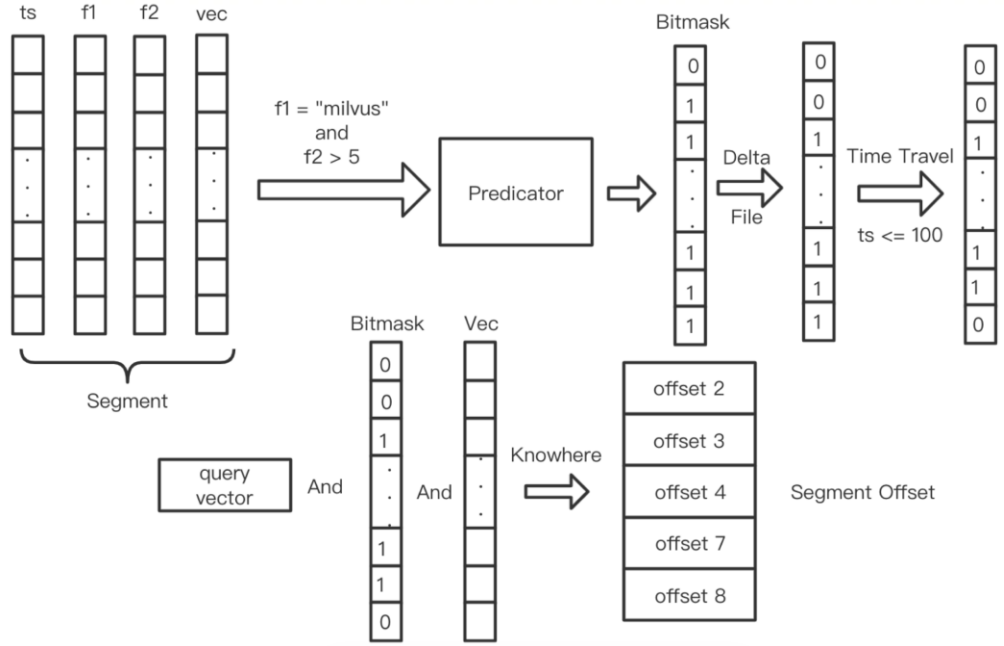
更快：过滤下的高效搜索


Query vector
(user uploaded image)
Inventory
Filtering
(by brand name)
ANNS results
Image similarity search
Query results

更快：过滤下的高效搜索

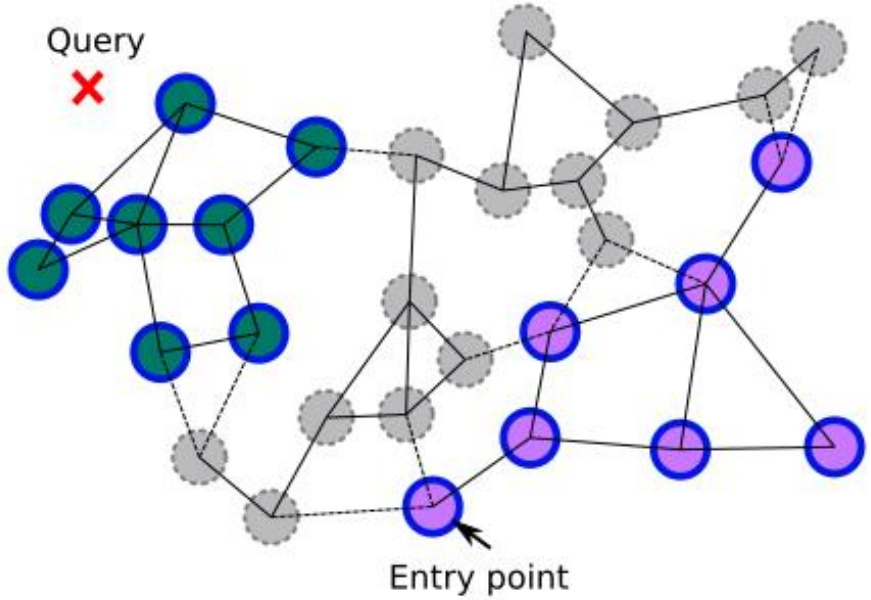
- 过滤方式
  - Pre-Filtering 边做向量检索边过滤
    适合大部分场景
  - Post-Filtering 昨晚向量检索再过滤
    只适合过滤量非常小的场景

- Validation评估方式
  - Bitset
    Overhead较大，适合过滤量较大的场景
  - Expr
    单次执行较慢，适合过滤量较小的场景



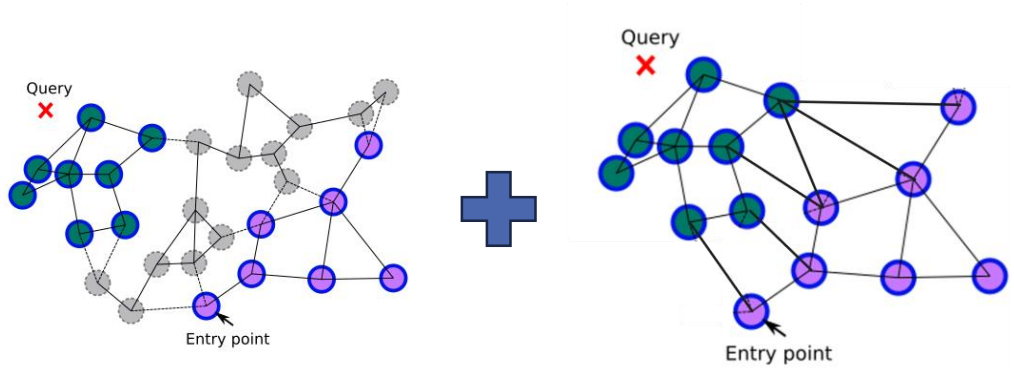
更快：过滤下的高效搜索

- 如何在图上处理过滤掉的点
  - Without candidates(绕过去)
    容易产生孤岛问题
  - With candidates(走过去):
    孤岛问题缓解，性能问题凸显

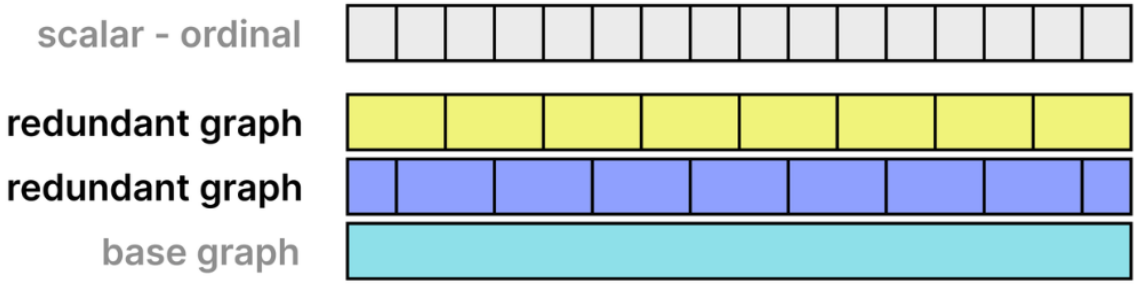


更快：过滤下的高效搜索

- 对于离散数据对不同标签及标签组合，通过分析数据分布构造小图提升连通性

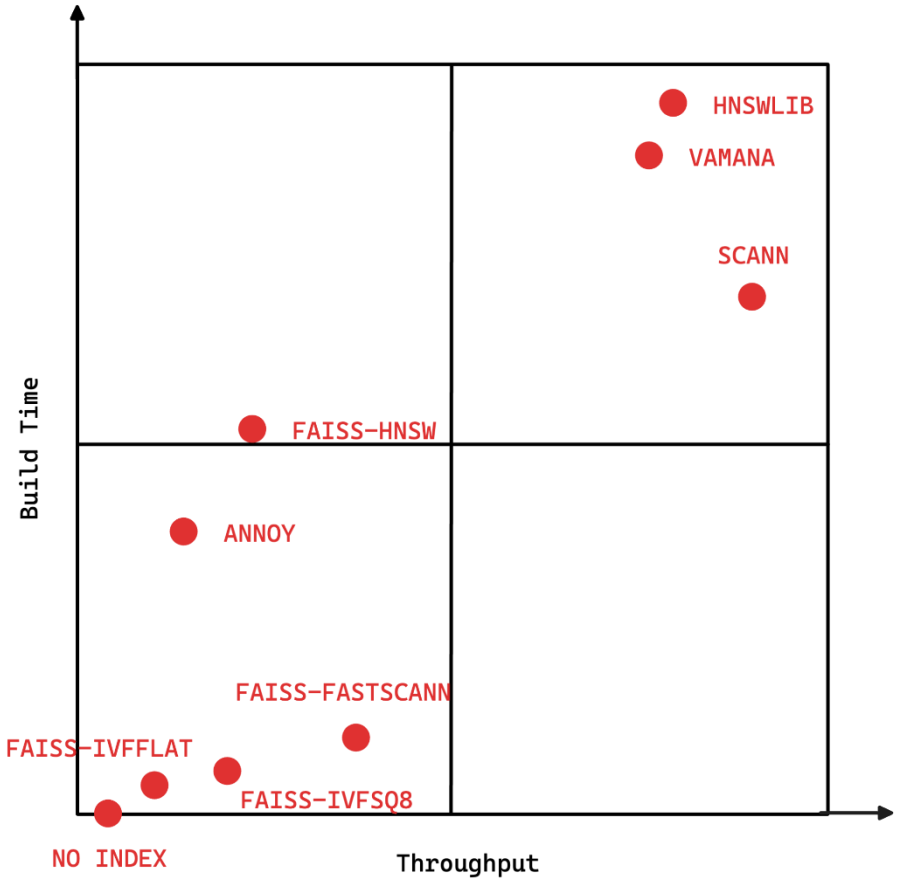


- 对于连续数据，除小图外，还需要冗余层增加不同小图之间的连通性



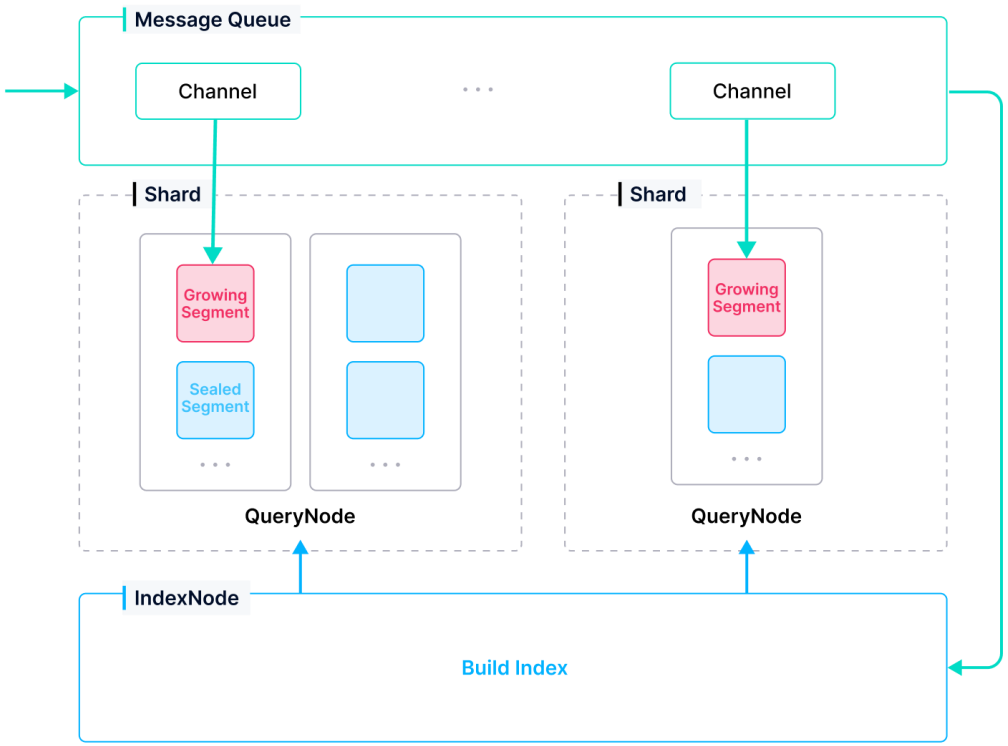
还有多列标签的And，Or结合的优化等等，性能提升2-10x

更快：实时性和性能都想要



- 实时性和性能不可兼得
- 大部分向量数据库选择直接使用HNSW，通过牺牲实时性换取更好的性能

更快：实时性和性能都想要



- Shard

  加速数据插入

- Segment

  Milvus的基本数据单元

  - Growing Segment

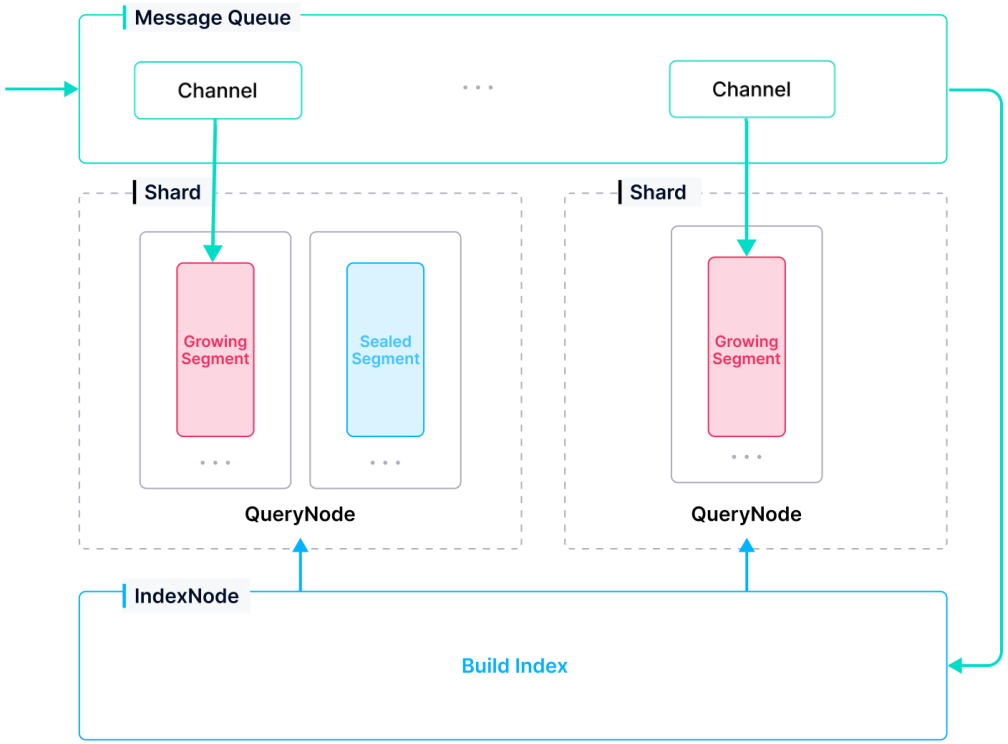
    直接从消息队列里拿取数据以提供快速检索，使用构建较快的索引

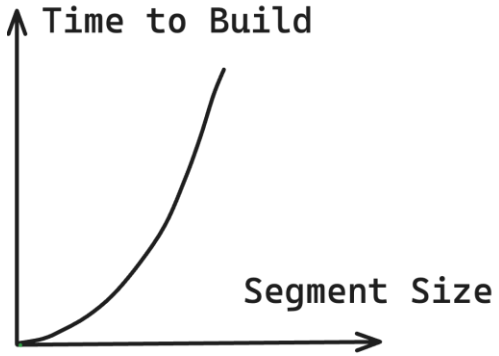
  - Sealed Segment

    使用Immutable的索引以保证查询速度

更快：实时性和性能都想要

Big Segment

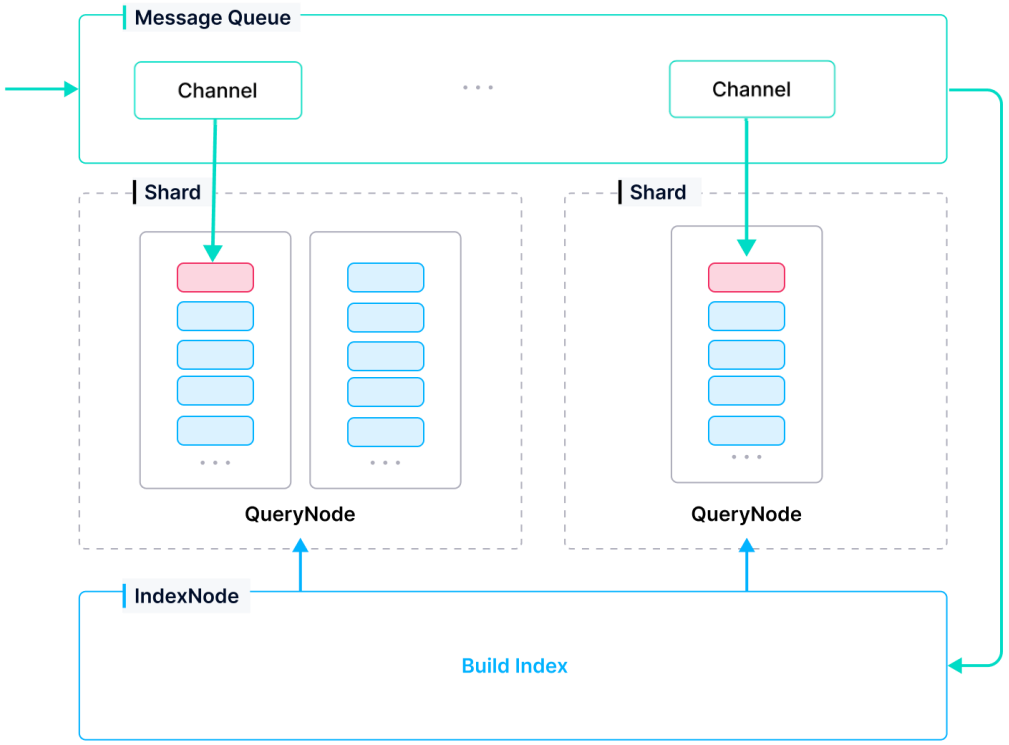


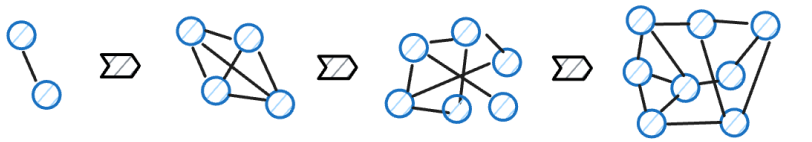


- 搜索性能降低
- 建索引速度慢
- 负载更加不均衡
- 内存overhead过大
- Failure Recovery能力变差

更快：实时性和性能都想要

Small Segment



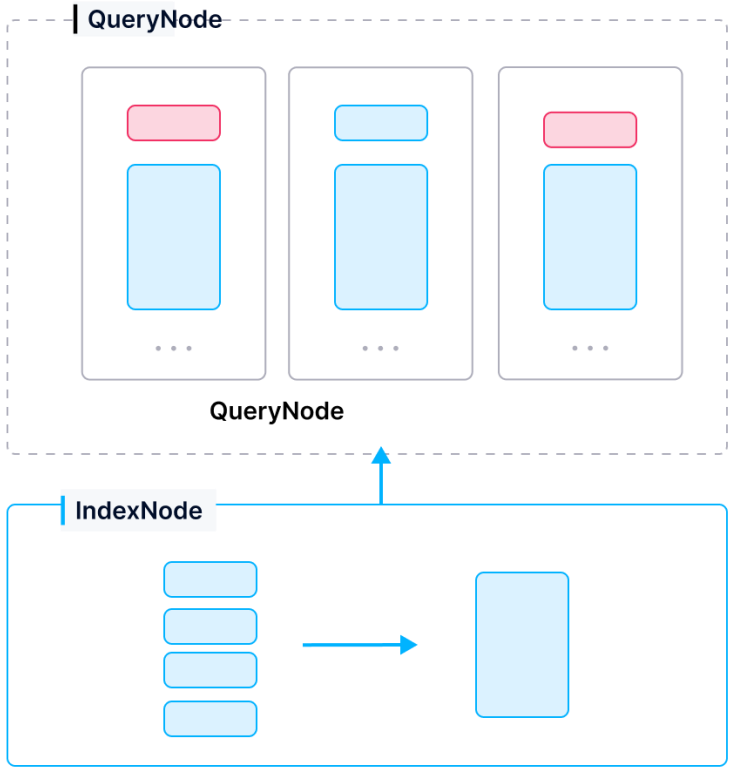


Size	0.25M	0.5M	0.75M	1M
QPS	~9.7K	~10K	~10.9K	~9.9K

- 对于图索引，大小对于性能的影响可以忽略。因此大量小索引会导致整体性能降低
- 元数据储存压力较大

更快：实时性和性能都想要

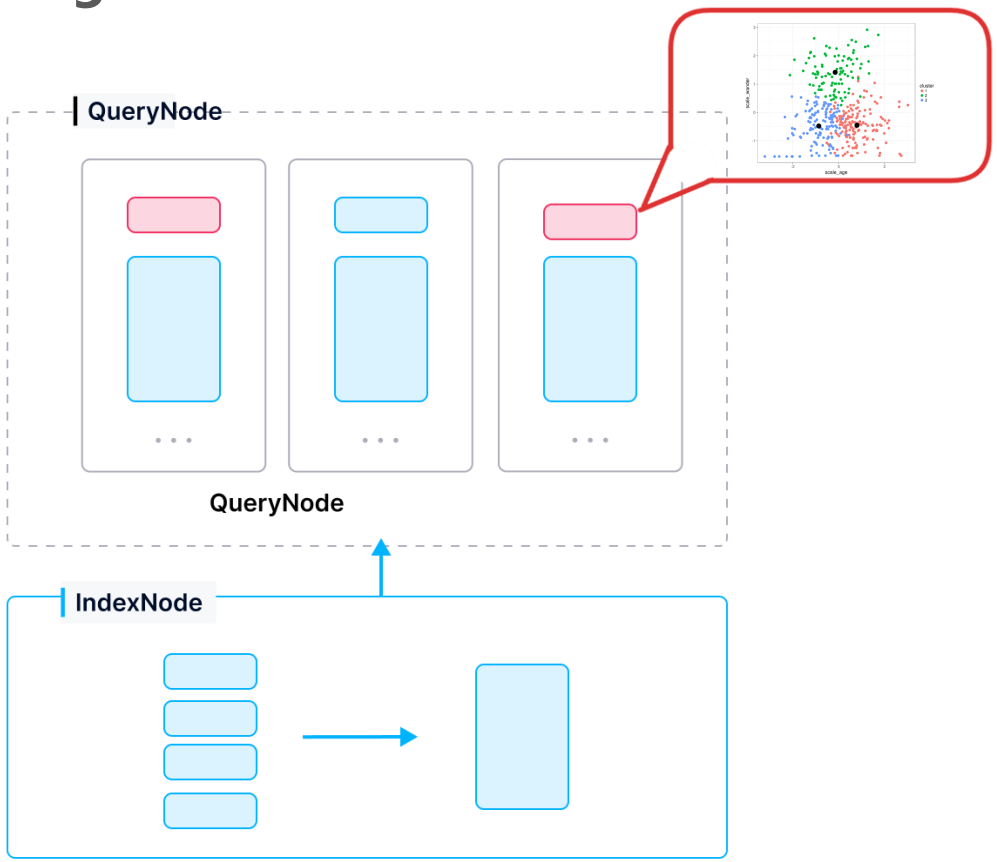
异步Compaction



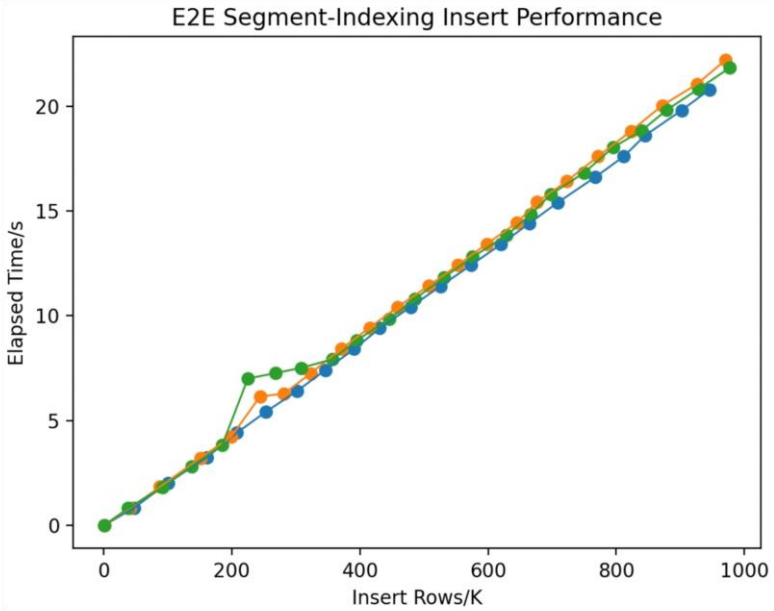
- DataNode会把小的Segment Compact成大Segment，并由IndexNode重新构建索引
- QueryNode会把大的Segment加载上来代替小的Segment
- 除此之外，Compaction还可以通过多级Compaction来删除过期数据，回收删除的数据，重新partition等

更快：实时性和性能都想要

Growing Index



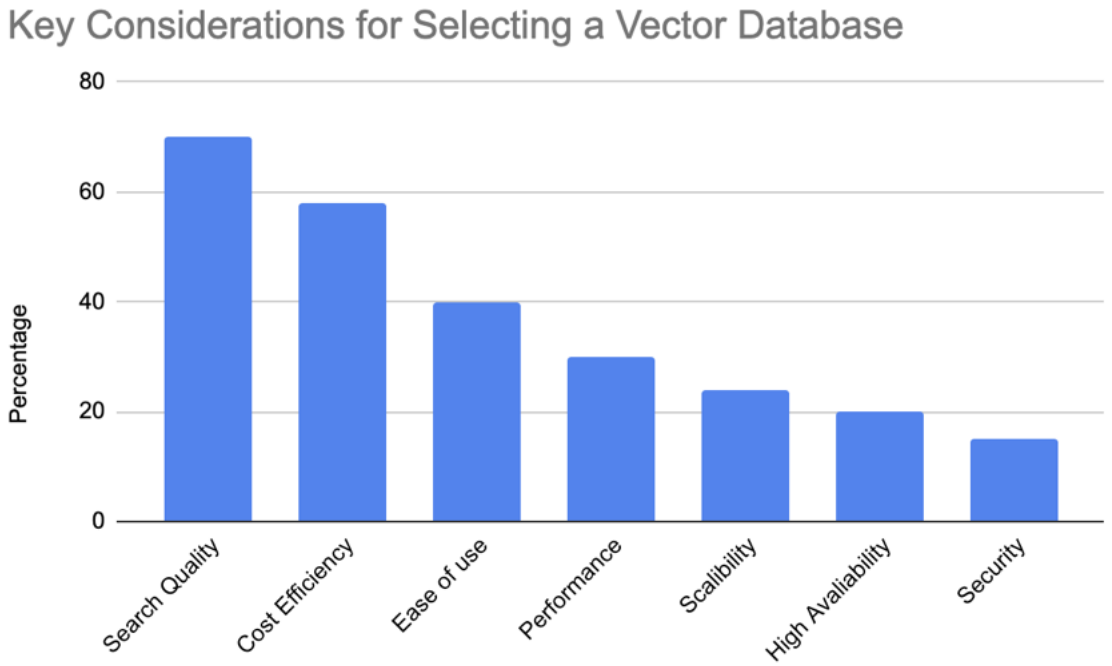
在Growing Segment中使用构建速度较快的索引而不是暴搜



除此之外，还有大K，删除，异常分布，等

用户关注什么

- 更强大的搜索质量
  - 更好的搜索效果
  - 更丰富的语义
  - 更灵活的查询方式
- 更高的性价比
  - 资源的池化使用
  - 少量查询的冷热储存



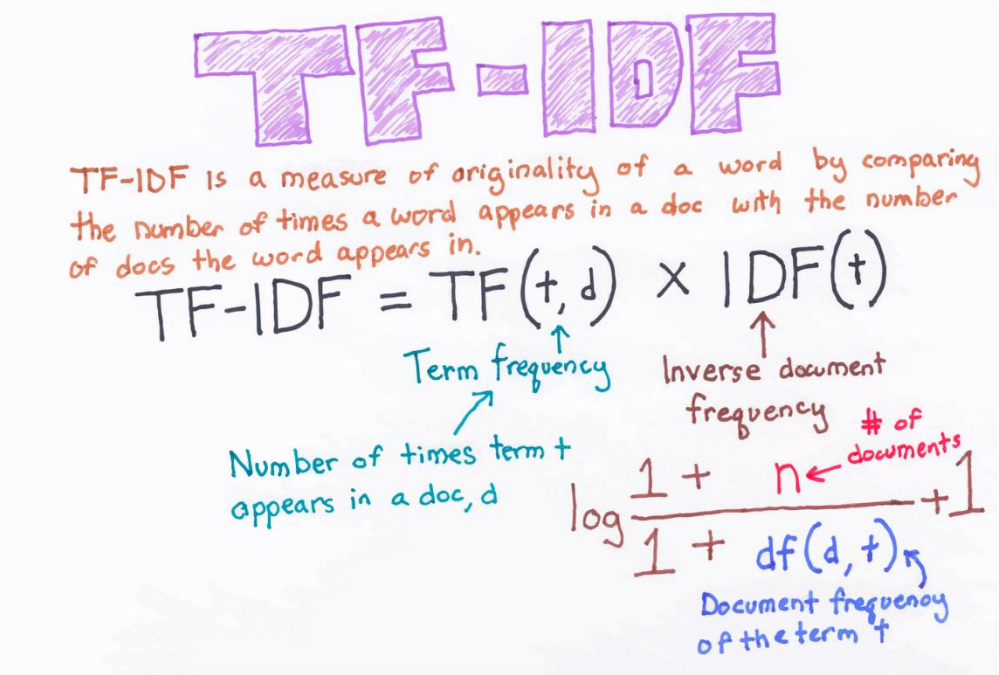
Gathering feedback from 1000 GenAI developers from Milvus community, 2024.

PART 03
不仅仅是ANN搜索

更强大的搜索质量：Sparse

TF-IDF

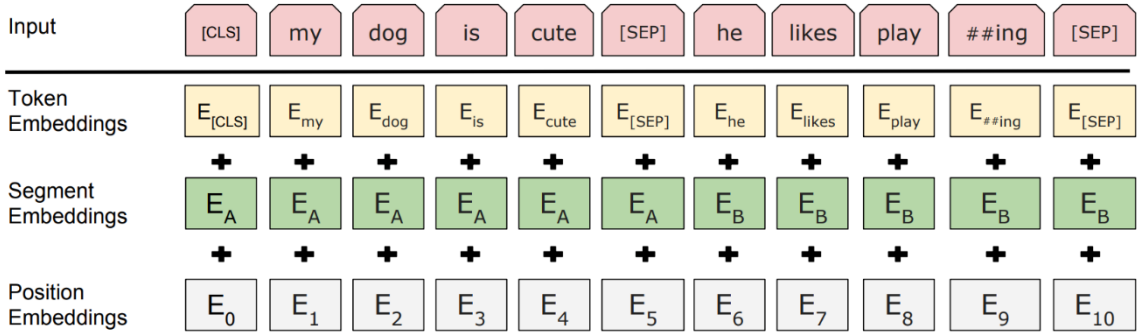
- **TF(Term Frequency, 词频)**表示词条在文本中出现的频率
- **IDF(Inverse Document Frequency, 逆文件频率)**表示关键词的普遍程度。如果包含词条的文档越少，则IDF越大，说明该词条具有很好的类别区分能力。
- TF-IDF的重要变种 - **BM25**，进一步考虑了词频的饱和度，同时会基于文本长度对词频进行归一化,以防止它偏向长的文件



更强大的搜索质量：Sparse

BERT

- BERT引入Transformer和预训练。注意力机制使得模型能够关于更的上下文，预训练使得模型对概率的理解更加充分。
- 揭开向量检索的序幕，带来了语义理解能力

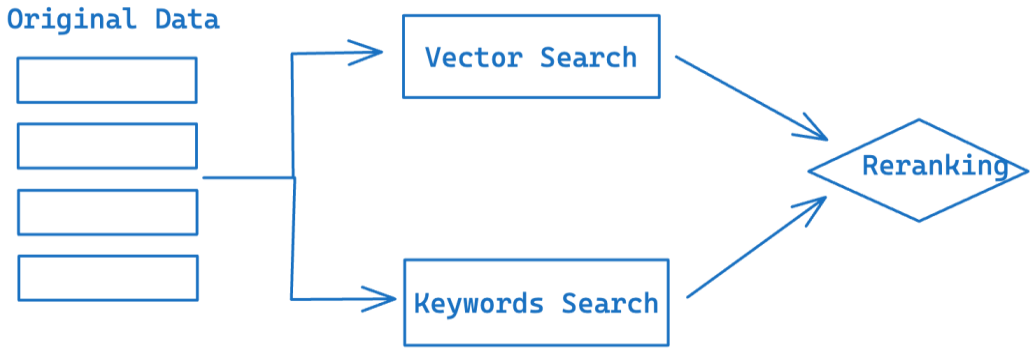


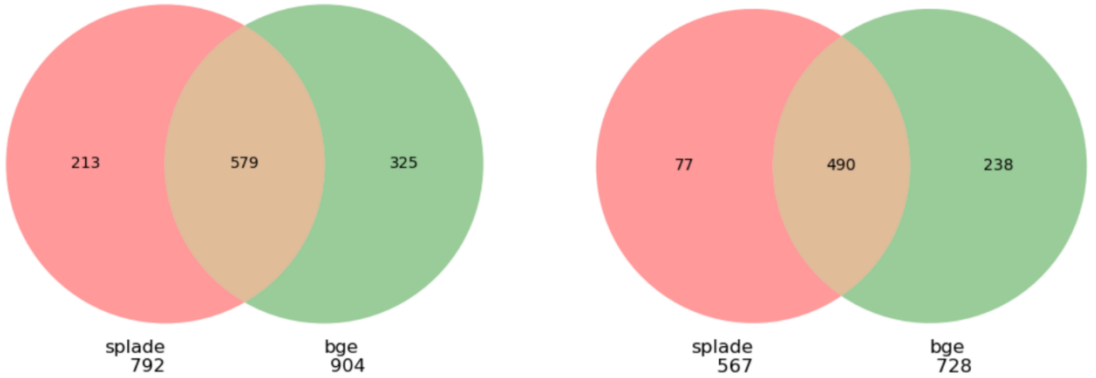
Method		MS MARCO Passage Development MRR@10	MS MARCO Passage Test MRR@10
BM25 (Microsoft Baseline)		0.167	0.165
IRNet (Deep CNN/IR Hybrid Network)	January 2nd, 2019	0.278	0.281
BERT [Nogueira and Cho, 2019]	January 7th, 2019	0.365	0.359

Table 1: The state of the leaderboard for the MS MARCO passage ranking task in January 2019, showing the introduction of BERT and the best model (IRNet) just prior to it. This large gain in effectiveness kicked off the "BERT revolution" in text ranking.

更强大的搜索质量：Sparse

向量检索不是传统关键词检索的替代，而是补充





Performing top-10 retrieval of all queries in NFCorpus/FIQA dataset based on embeddings generated by SPLADE and BGE-M3.

Red: number of ground truth documents that are only found by SPLADE

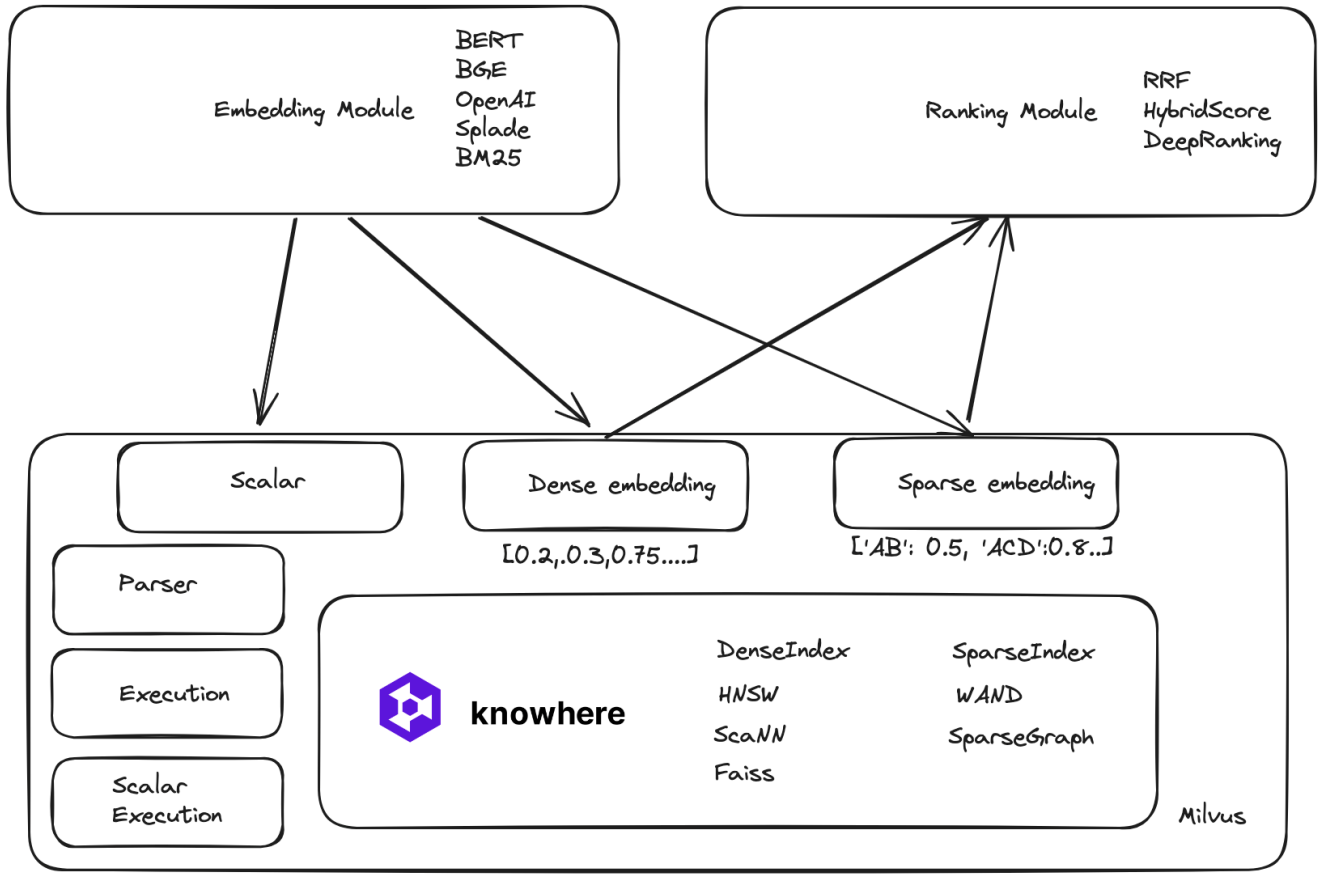
Green: number of ground truth documents that are only found by BGE-M3

Yellow: number of ground truth documents that are found by both models

传统关键词检索擅长域外信息、专有词汇的检索而

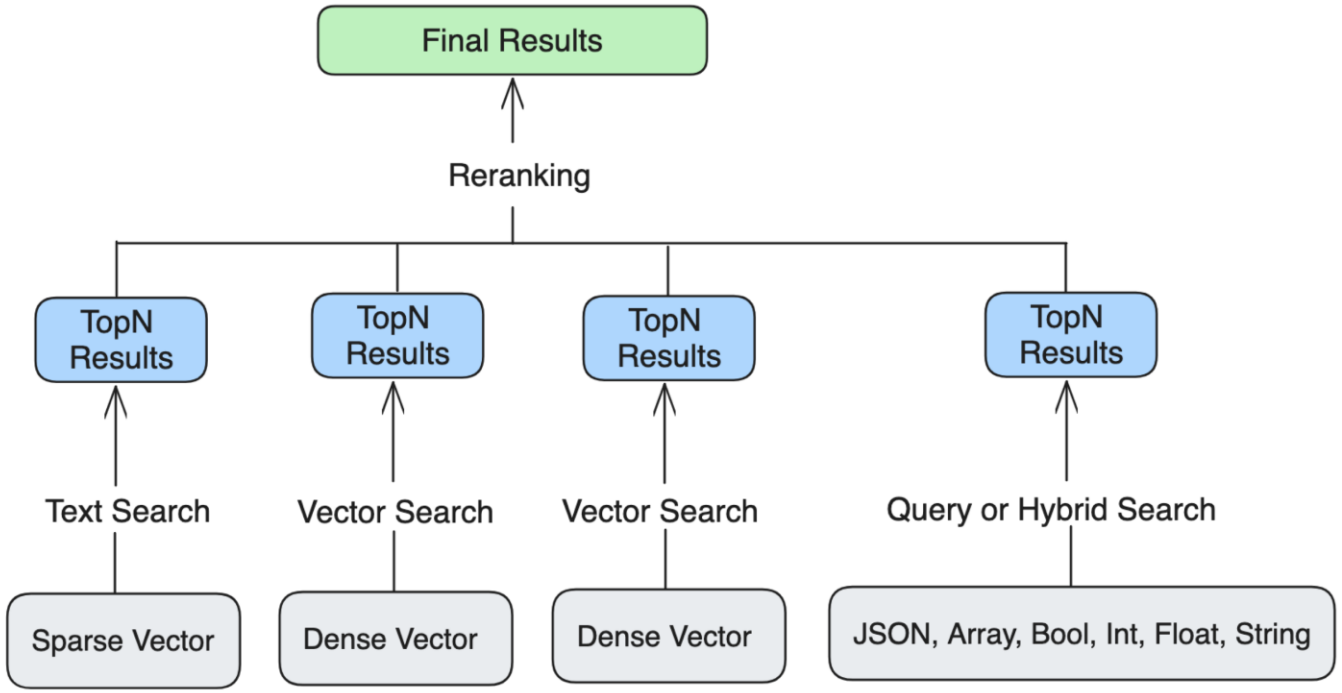
向量检索更擅长关注语义信息

更强大的搜索质量：Sparse



- All in one，解决同时维护两套系统的复杂度
- 可插拔的Embedding能力，覆盖BM25，不止BM25
- 可选择的Reranking方式

更强大的搜索质量：Hybrid Search



Dense + Dense

- 组合语义：比如分别用两人的照片搜合照
- 提升精度：比如同一个人的不同视角

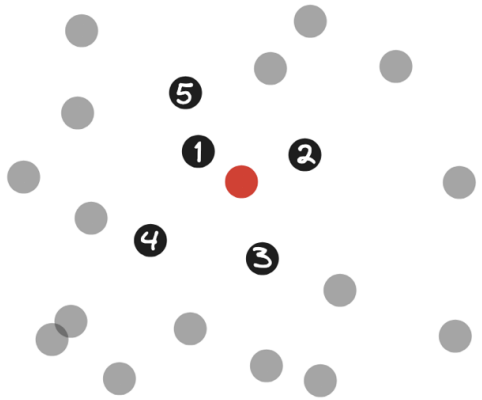
Scalar + Dense

- 更灵活的检索：比如不同距离的帖子，语义相似度占比不同
- 更全面的检索：比如通过证件号+人脸匹配信息

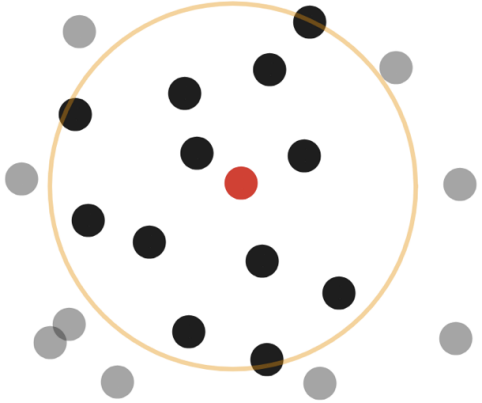
更强大的搜索质量：Range Search

丰富的搜索语义

TOP-K ANN



RANGE Search



现实场景中很多查询没有相近的匹配项，而有的查询却有很多。用户希望找到足够相近的，而不是足够多的结果

- 避免搜索出毫不相关的结果
- 快速找到未知大小的结果

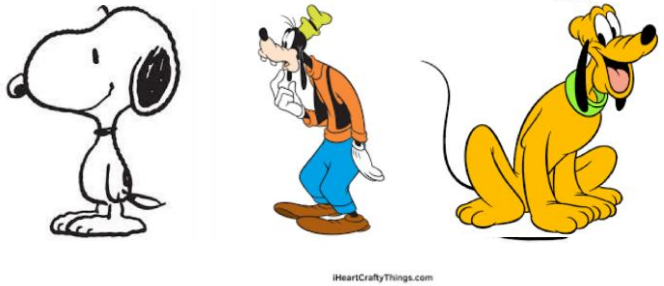
更强大的搜索质量： Groupby

向量化数据库语义的开始

如果我们想检索到三种不同的小狗怎么办

查询：

结果：



name：Snoopy Goofy Pluto

Top1 Top2 Top3

SELECT TOP1(image) GROUPBY **name** ORDER BY score LIMIT 3

如果我们没有name列怎么办？

SELECT TOP1(image) GROUPBY **image** ORDER BY score LIMIT 3

更多的Aggregator：AVG, COUNT, etc

更多的排序方法：AVG, UDF, etc.

PART 04
Zilliz Cloud 及更多

更好的性价比：Zilliz Cloud

向量数据库即服务

- Zilliz Cloud是Zilliz基于开源向量数据库Milvus打造的全托管企业级向量检索云服务
- 分为Serverless，SaaS和BYOC三个版本，面向不同需求和不同部署环境
- 提供大量企业级功能，助力用户聚焦业务逻辑
- 目前已经登陆AWS，GCP，Azure，阿里云、腾讯云等5朵云

更好的性价比：Zilliz Cloud - Saas

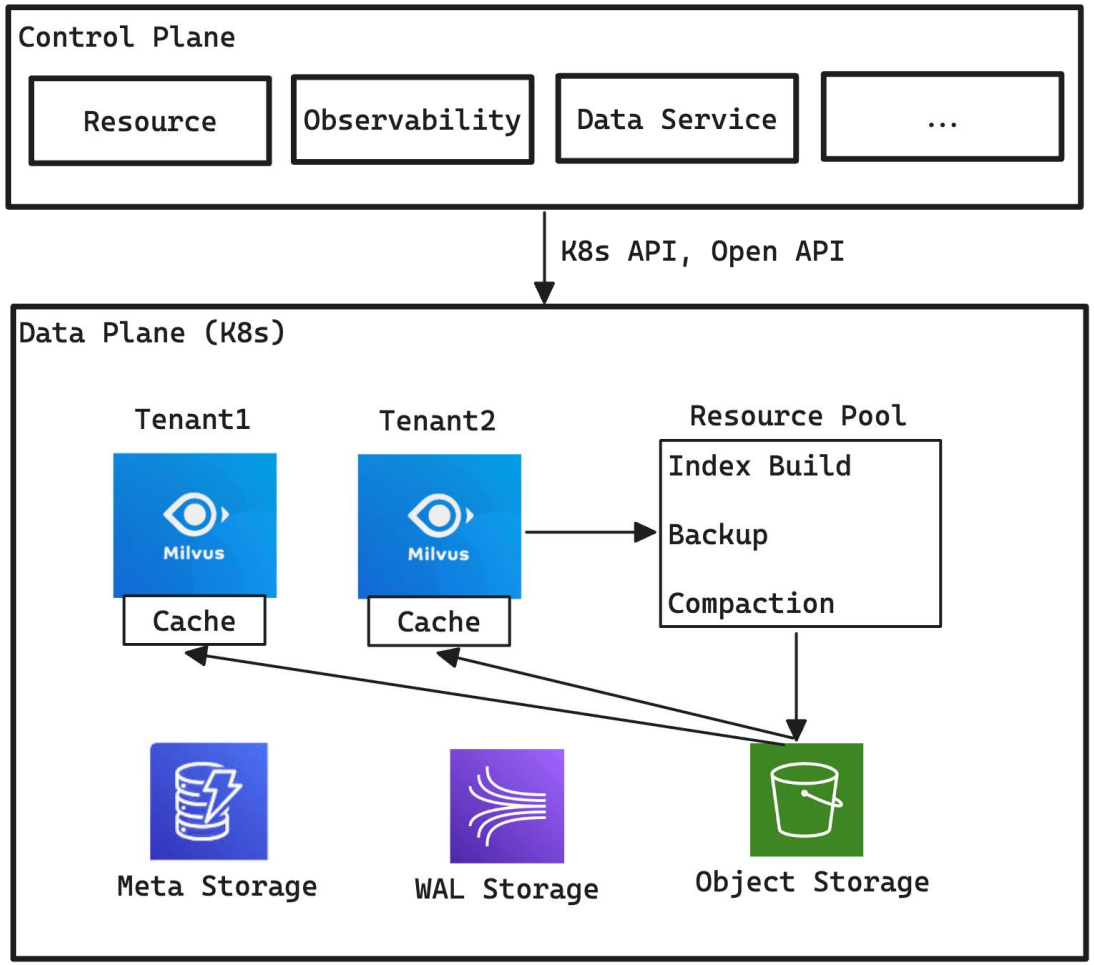
Saas架构

优点

- 弹性的的池化资源节约大量成本
- K8s namespace提供较好的隔离性
- 数据全量缓存在查询节点上提高性能

缺点

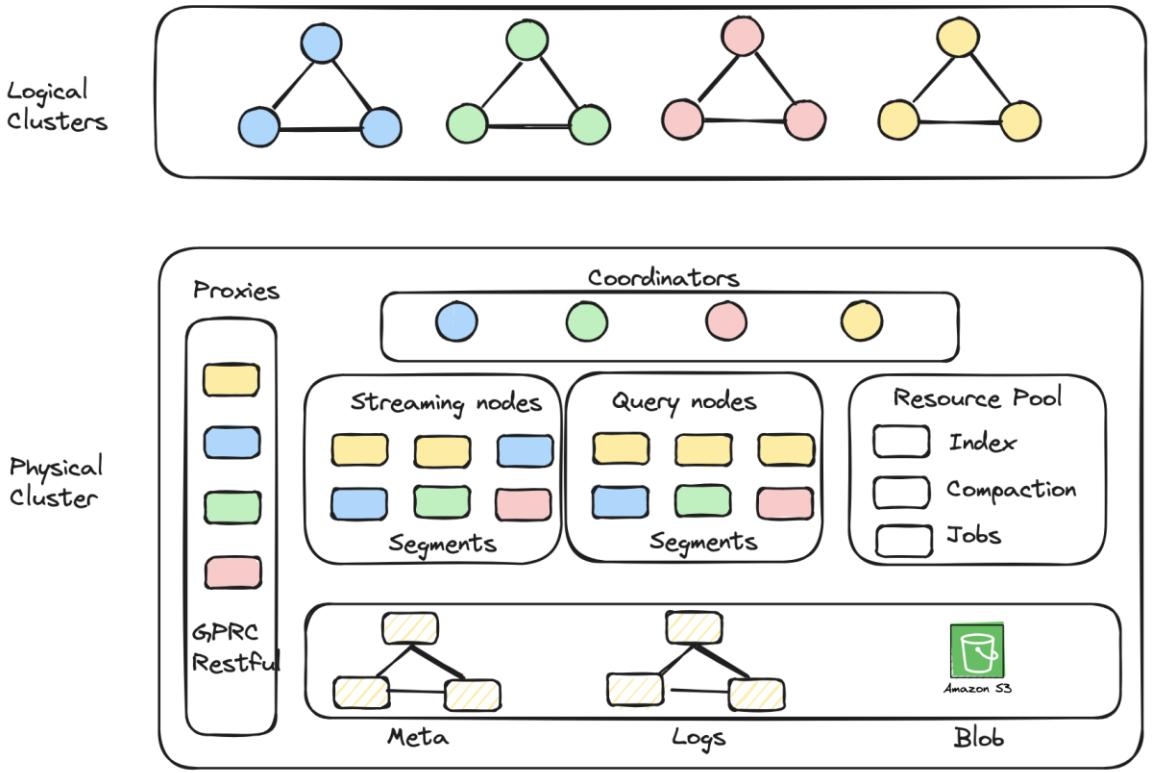
- 缓存数据消耗大量资源，不查询的时候也需要付出大量成本
- 大数据量下性能损失



更好的性价比：Zilliz Cloud - Serverless

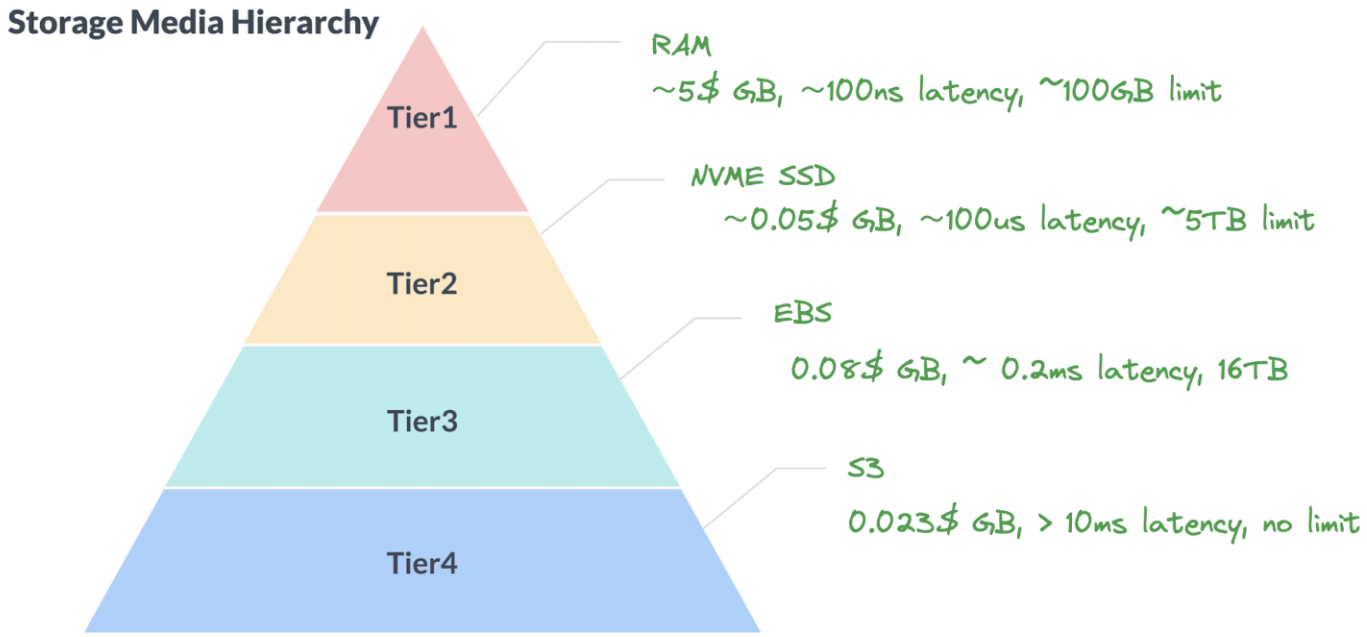
Logic Clusters

- 所有节点stateless，为任意logic cluster提供服务，用户感知不变
- 节点
  - Proxy：路由用户请求，限流
  - StreamingNode：实时数据查询
  - QueryNode：历史数据查询
  - Resource Pool：执行建索引，Compaction等batch job
  - Storage：log和meta储存
- 每种节点根据资源使用自动扩缩容



更好的性价比：Zilliz Cloud - Serverless

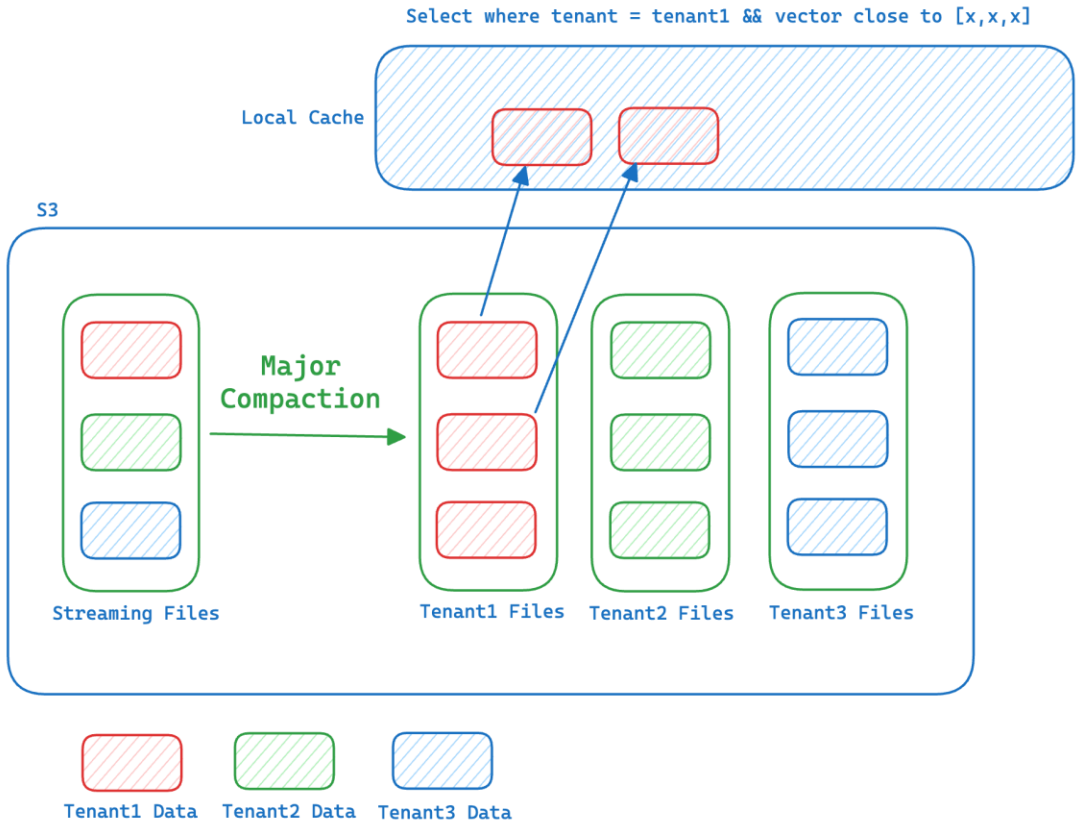
分层存储



- 从Ram到S3多级缓存
- 只加载需要的数据，多余的数据会向下层evict

更好的性价比：Zilliz Cloud - Serverless

冷热数据分离

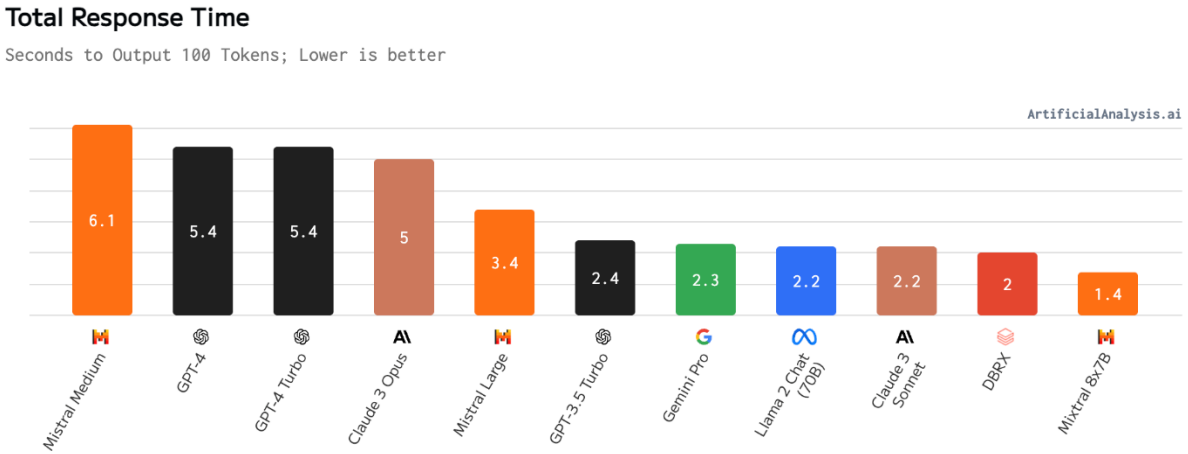


- 数据分离
  - 根据不同tenant分离数据
  - 根据过滤条件分离数据
  - 根据向量空间分布分离数据
- 多层级
- Pre-Warm

更好的性价比：Zilliz Cloud - Serverless

极致的性价比

许多场景并不性能敏感



https://artificialanalysis.ai/models

	Cold Search	Warm Search	Hot Search
1M 768Dim	2.3s	80ms	4ms
10M 768Dim	7s	150ms	7ms

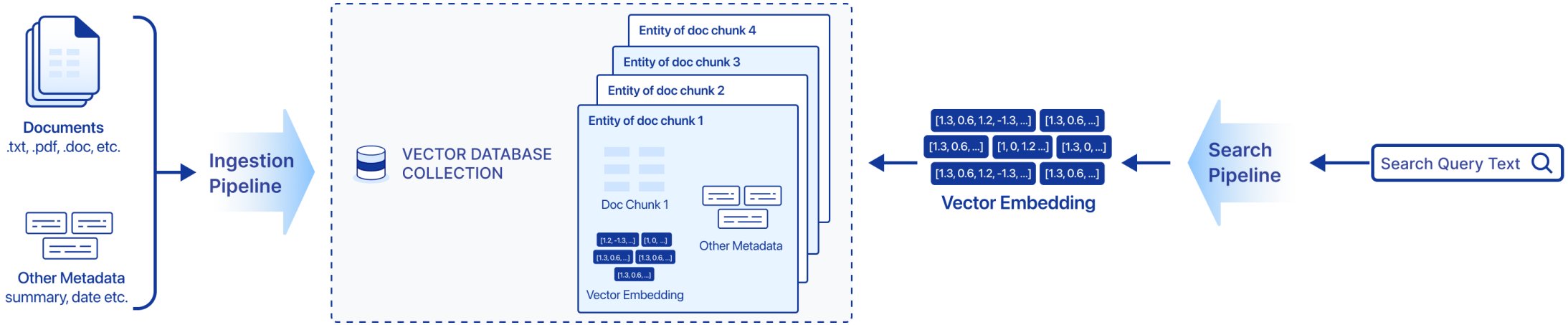
host: 1CU ZillizCloud

$915	$228	$16
Performance	Capacity	Serverless

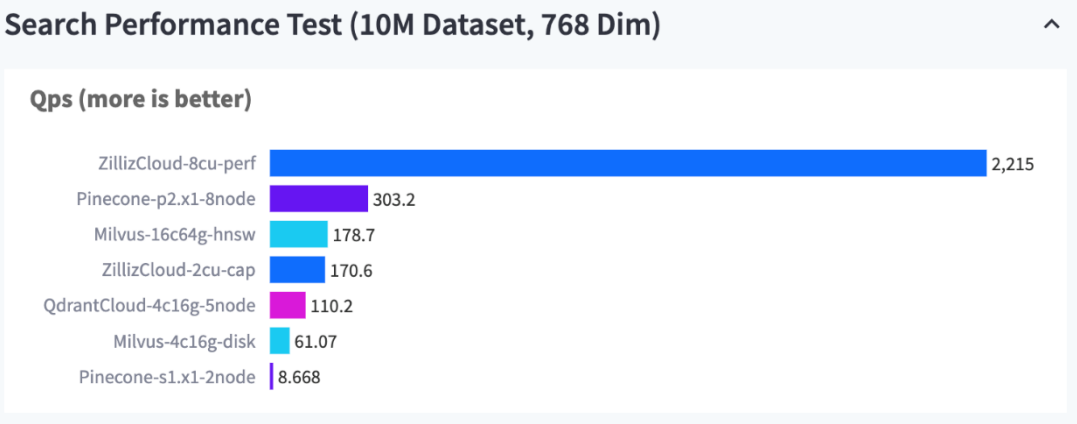
Cost per Month for 10M 768Dim Data in ZillizCloud

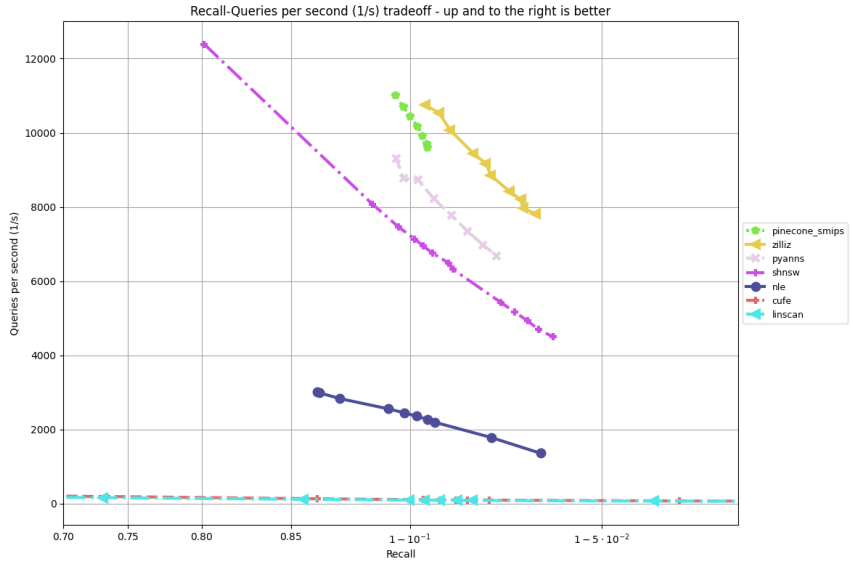
更加易用：Zilliz Cloud Pipeline

- Zilliz Cloud Pipeline是围绕Zilliz Cloud打造的一站式文本知识库系统
- 支持多种开源模型，可以任意选择



打造极致性能：Cardinal



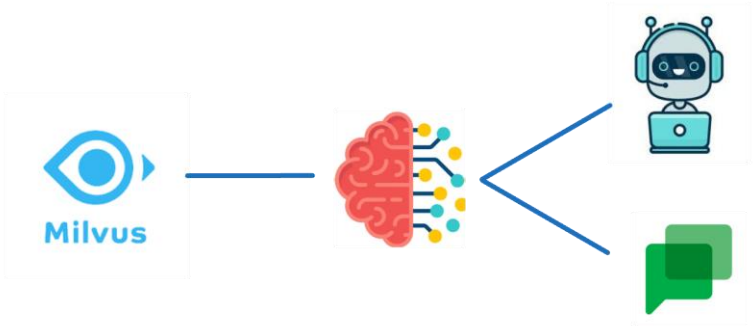


- Zilliz Cloud基于自研的Cardinal索引，性能是Milvus的5-10倍
- Cardinal的Dense Vector算法使用了创新的IVF+图的混合算法，加上极致的工程优化达到性能巅峰
- Cardinal的Sparse Vector算法在BigANN中斩获最佳
- Autoindex在方便用户使用的同时，基于数据分布的自动参数调优配置助力性能更上一层

新场景带来的新挑战

RAG/Agent

- 准确度的评价标准不止是相关性，也包含重要性，时间
- 视频场景

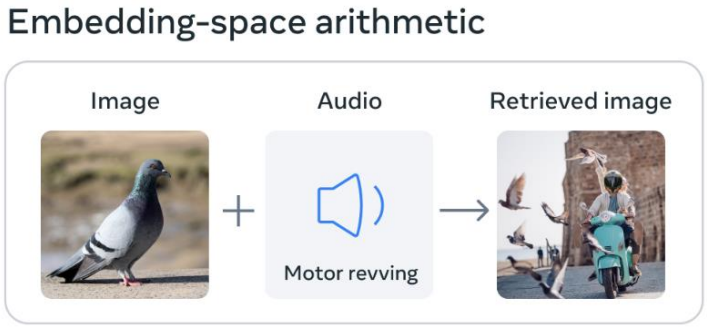


大规模离线分析场景

- 海量数据的大规模召回
- 复杂的分析语义
- 可视化

多模态

- 跨模态的查询
- 多种模态的叠加查询



AI 初创 计划

1000 万元 Zilliz Cloud 抵扣金

- 探索 AI 新应用，初创企业的加速器

0元上云

- 每个注册账户最高 10 万元的 Zilliz Cloud 抵扣金

技术支持

- Zilliz技术专家贴身服务进行产品or服务的互相集成

市场推广

- 参与 or 合作 Zilliz全球范围内的市场活动及推广

商机共享

- Zilliz提供商机共享缩短交易达成周期

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

AiDD

THANKS