向量数据库选型评估

业务价值

- 1. 为 LLM 存储长期记忆
- 2. 解决 LLM 输入的 Token 限制:不会在每个 LLM 提示中都发送所有文档,而是只发送少数相关度最高的文档
- 3. 私有部署解决数据安全问题

核心功能

高效地存储、对比和检索海量的向量

Round 0:确定向量数据库的范围 √

向量数据库	LangChain 支 持 [Link]	Semantic Kernel [Link]	LlamaIndex [Link]	OpenAl 支持 [L	ink]
AnalyticDB	√-se	王汛洪T586	王汛洪7586		
Annoy	√	王	_{有效} 7586		
Atlas	$\sqrt{}$		王汛洪7586		
AwaDB	▼	毛	P洪 7586		
Azure Search	√ √ · · · ·	√玉汛港1586	王 汛洪7586	$\sqrt{}$	
BagelDB	√	毛	_{有洪} 7586		
Cassandra	$\sqrt{}$	王 汛煤T586	王汛港7586		
Chroma (1986)	▼	√	√ ·	√	
Clarifai		主 汛港 ⁷⁵⁸⁶	王汛洪7586		
ClickHouse Vector	√	王	有法 7586		
Search		王汛洪7586	王汛洪7586		
Activeloop's Deep Lake	▼	王	_{花花} T586		
Dingo	√ 1586 √	主邪港7586	主汛港1586		
DocArrayHnswSear ch	√ 王汛洪T586 孔洪T586	王州洪7586	## 7586 ● # 7586		
DocArrayInMemory Search	√ ±π∓1586	王	√ ·		
Search E		_{王汛洪} 7586	王汛洪7586		
ElasticSearch	√		pst 7586	√ 1586	
FAISS	N. 3€ 1586	王汛港 ⁷⁵⁸⁶	√		
Hologres	√ ∓₹¥7586	44.3	n拼 7586		
LanceDB	√-86	主汛洪T586	√ =用井7586		
			206		

Marqo	E 0 V 56		主 汛洪7586			王汛洪7586			
MatchingEngine	√-								
Meilisearch	₹10%√-6		主 汛洪7586			王 汛洪7586			王汛洪7586
Milvus ************************************	√	王汛洪7586	$\sqrt{}$	王汛	1505		√	王汛洪7586	
MongoDB Atlas	₹10#.7586		主 汛洪7586		$\sqrt{}$	王 汛洪7586			
MyScale MyScale	√				± 1586				
OpenSearch	王和洪7586		玉 汛洪 ⁷⁵⁸⁶		$\sqrt{}$				
pg_embedding	√							王汛洪7586	
PGVector	₹ 1580 √		√ERI# 7580				$\sqrt{}$		
Pinecone	√ 		V 25.06		X 1580		$\sqrt{}$		
Qdrant	₹ 13A 13A		V TENNAL		√ 		$\sqrt{}$		
Redis	√ 7586		V 7586		√ -		$\sqrt{}$		
Rockset	√		主州//		7586			7586	
ScaNN	√ 1586		亚亚米 7586			亚亚共 7586			
SingleStoreDB	√-								
scikit-learn	₹7586		主 汛洪7586			_{王汛洪} 7586			工汛洪 7586
StarRocks	√								
Supabase (Postgres)	王州洪 7586		宝 汛洪 ⁷⁵⁸⁶		$\sqrt{}$	主汛洪7586	$\sqrt{}$		
Tair ERA 7586	√				× 1586				
Tigris	±1/34 1580		王汛洪 ⁷⁵⁸⁰						
Typesense	√ 		7686						
USearch	₹		王州 从						
Vectara	√ 	王汛洪 ⁷⁵⁸⁶	-m:tt. 7586					王汛洪7586	
Weaviate	√		√ × × × × × × × × × × × × × × × × × × ×		√ 		$\sqrt{}$		
Xata	√ 		工用洪7586			七讯洪7586			
Zilliz	√-	= 286			7526	Specific 1	√	, 75.86	

选择在 LangChain、Semantic Kernel、LlamaIndex 和 OpenAI 中支持较多的向量数据库

Round 1: 根据类型筛选 $\sqrt{}$

向量数据库	数据库类型	开源	部署	结论
AnalyticDB	扩展向量模块×	否 ×	Managed ×	× 主汛洪 ⁷⁵⁸⁶
Azure Search	玉形洪 7586	否× 宝石洪 7586	Managed ×	×
Chroma	纯向量库	是	Memory ×	★ 玉汛洪 7586
. 4296	. 4596	. 4296	. 46	

ElasticSearch / OpenSearch	扩展向量模块 ×	是 (1787)	Managed Self-Hosted	× _王 元第7586
Milvus	纯向量库	是	Managed Self-Hosted	√ 16
PGVector	扩展向量模块 ×	是	Self-Hosted	★ E R # 1586
Pinecone	纯向量库	否 × 至刑共7589	Managed ×	×
Qdrant	纯向量库	是 _{至3} 108 1589	Managed Self-Hosted	重用港 15 556
Redis	扩展向量模块 ×	⁸⁶ 是	Self-Hosted	× 主用港 7586
Supabase (Postgres)	扩展向量模块 ×	是	Self-Hosted	×
Weaviate	纯向量库	是 _{至形第1586}	Managed Self-Hosted	▼
Zilliz (Milvus)	纯向量库	是	Managed ×	★ 至用港 T 586

以下类型的向量数据库不考虑:

- 不开源
- 不能本地部署
- 不能持久化的,只能 Memory 存储的
- 其他类型数据库(OLTP、NoSQL、OLAP)扩展出向量模块的不选择:这些数据库大多数都没有对向量的使用进行优化。另外从微服务的角度说,每个专业的服务(or工具)应该专注于解决一个的特定事情,能保持服务进行独立的扩展、部署和维护

Round 2: 文献调研

	调研项	Weaviate	Qdrant	Milvus
社区	开源许可证	BSD-3-Clause	Apache 2.0	Apache 2.0
	商业公司维护	是	是	是
	发布时间	2019	2021	2019
	Star ★	7.1k	12.2k	22k
煮 洪7586	DB- Engines Ranking	5 ERRTS86	7 2 2 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8	4 第 第 第 第 第 第 第 第 第 第 第 第 第 第 8 8 8 8 8
_{和洪7586}	网址	Link	Link	Link ERICH TSE6
功能 ASK 7586	核心特性	 支持向量与对象的存储 向量索引与倒排索引组合 集成 ML / Embedding 模型 	丰富的数据类型Payload 可以索引WAL 日志写入SIMD	极高的检索性能非结构化数据的极简管理跨平台实时检索和分析

	● 支持推荐、总结 等能力		具有很高的容灾与 障转移能力高度可拓展
数据量级	±刑淋 1586 至7	[第1586 主用序节	10 亿级
Index Types ★	倒排索引	• Payload 索引	• 标量索引
1ypes A	● 向量索引	全文索引向量索引	● 向量索引 ● FLAT
	_{三元洪} 7586	HNSW	• IVS_FLAT
	王 元并7586		IVF_SQ8IVF_PQ
	_{王用洪} 7586		• HNSW ************************************
	宝汛洪 ^{T586} 宝汛洪 ^{T586}		ANNOYBIN_FLATBIN_IVF_FLAT
Similarity	Cosine Similarity	Cosine Similarity	Euclidean Distance
Metrics ★	• Dot Product	• Dot Product	Inner Product
	Euclidean Distance	Euclidean Distance	JaccardTanimoto
	HammingManhattan		HammingSuperstructure
	Distance		Substructure
混合查询	√ Ξ∄Ħ7586 Ξ∄	▼	√
数据模型	• Class	• Collection	Collection: Table
	• Property	PointsSharding	Entity: RowField: Field (Value)
	王飛洪7586	SHATS#	or Vector)
	王汛洪T586 王汛洪T586		Primary KeySharding / Partition/ Segment
Backup & Restore ★	7585	√ Snapshots	√ Binlog、日志持久 化、日志快照
编程接口	• RESTful	RESTful	• RESTful
主 讯洪 ⁷⁵⁸⁶	GraphSQL	OpenAPI 3.0	• Java
	PythonJava	• gRPC • Python	 Go Javascript
	GoJavascript	• Go • Rust	王汛洪 ^{T586} 王汛洪 ^{T5}
	生讯洋 1586	Javascript	
RBAC	X - 17.38 75.86	★ 滋7586 _{东汛洪75}	55 ∓汛港 ⁷⁵
编程语言	Go	Rust	Go
架构图	三汛洪 ⁷⁵⁸⁶		86 王汛洪 ⁷⁵

	Manuscript of the second of th	Parameter Section 1986	A. Feyer.		Lood Rabaces - Transp.
	Nonexperit Stream Conference and Artifact and another Company of the Conference and Artifact and	Makenya		土 ///	unitationne 💼 🏣
架构复杂	简单	简单		王汛洪7586	复杂,包含组件有:
度★	王 汛洪7586				Access Layer
	≠汧洪7586	工刊洪7586		75RE	Coordinator Service
	主汛洪	王环决		王汛洪7588	• Worker Node,细分
	主 汛洪7586				为: ※1586
	_{主汛洪} 7586	王环洪7586		王汛洪7586	 data node
					 query node
	宝汛 从7586				index node
	主汛洪7586	王环洪7586		王汛洪7586	• Storage,细分为:
	王汛洪 ⁷⁵⁸⁶				。 元数据存储
	主 讯洪7586	工环洪 7586		- 4t 7588	(meta store)
	E34075 **	王孙从		王汛洪7586	。 消息存储(log
	玉 汛洪7586				broker)
	主 汛洪7586	王环洪 7586		王汛洪7586	。 对象存储 (object
					storage)
ム左式加払	√	√		- 14 7538	
分布式架构				王汛灰。"	√ 分布式 MPP 架构
存算分离	× 宝用港 ⁷⁵⁸⁶	×			★ 宝汛洪7586
Replicatio n ★	√ (Class-level replication)	√ (Co replica	llection-le ation)	vel	▼ 37.H 7586
Sharding ★	$\sqrt{}$ 静态数据 Sharding	√静态	S数据 Shar	rding	√ 动态 Segment 替换
Partition	★ 重用港7586	×			√
*	主 汛洪7586	王环洪 7586		王汛洪7586	
可扩展性	Resharding 复数且耗时	杂 • Re 耗	sharding <u>1</u>	复杂且	存算分离,节点无状态,扩缩容简单
	三环共 ⁷⁵⁸⁶	王承洪7586	- 3	王汛洪7586	动态扩缩容
					水平扩展
	王州冲	- 7586			
数据一致性	• Eventual		entual Posistopov	王汛洪 (30)	• Eventual
	Consistency Tunable		nsistency nable		ConsistencyImmediate
	consistency		nable nsistency	王汛洪7586	• Immediate Consistency
					Session Consistency
	王汛洪7586				Tunable
	王 汛洪7586	王环洪7586		王汛洪7586	Consistency ************************************
服务器成	低 宝汛洪7586	低			高 宝汛港 1586
本★	_{三汛洪} 7586	王环洪7586		王汛洪758 6	
·T· 🔨					内置组件:
组件依赖	内置组件:	内置组	11年:		内且组件.
	内置组件: • 系统进程		l件: 统进程		• Proxy

王和洪 7586 王和洪 7586	主汛洪7586	• module service (可选) 外部组件:		ernetes	王邪洪下86	Query coord Index coord Data coord
王州从"		 Kubernetes 			王汛洪7586	Query node
	主汛洪 7586	王汛洪7586			•	Index node
王汛洪7586		主 讯洪7586	王汛洪7586		王汛洪7586 ●	Data node
	_{王汛洪7586}	主 汛洪7586			外	部组件:
王荆洪 7586		至河流 7586	王开洪7586		玉汛洪 ^{TS36}	Kubernetes etcd:元数据存储
	主 汛洪 7586	主 用洪7586			•	MinIO:对象存储
王和洪7586	-05	主 流洪T586	主汛洪7586		王 汛洪T586 ●	Kafka/Pulsar: 对象 存储
	云原生	√ ±388	√		$\sqrt{}$	

关键评估指标:

- 社区
 - Star
- 功能
 - Index Types
 - Similarity Metrics
 - Backup & Restore

• 架构

- 。 架构复杂度
- Replication
- Sharding
- Partition

• 部署

- 服务器成本
- 部署组件

术语解释:

- Faiss 是由 Facebook 开发的适用于稠密向量匹配的开源库,支持 C++ 与 Python 调用。Faiss 支持多种向量检索方式,包括内积、欧氏距离等,同时支持精确检索与模糊搜索
- ANN:近似最近邻搜索,ANN算法会预先构建索引,通常,有3种索引结构:图索引、树索引、哈希索引。选择不同的索引算法会影响搜索速度、内存使用情况和准确性。各种类型的ANN索引算法主要分为2种思路:缩小搜索范围和将高维向量空间分解为低维子空间。
- HNSW: Hierarchical Navigable Small World,图索引算法,通过创建多层接近图(Proximity graph)来索引向量空间。
- ANNOH:Approximate Nearest Neighbor Oh Yeah,树索引算法,通过构建二叉树森林来索引向量空间。ANNOY 使用在向量空间中距离两个随机向量相等的超平面来将空间分成 2 个子空间。对每个子空间重复此过程,直到每个子空间中最多有 K 项。
- IVF: Inverted File Index, IVF 通过将整个 vector 空间拆分成 k 个子空间,并对每个子空间找到一个代表质心(centroid)。并将子空间的所有 vector points 都 match 到这个质心上。给定一个要查询的 vector,先通过和所有的质心比较找到最近的质心,然后在这个质心所代表的子空间里搜索最近的点

- Jaccard/Tanimoto: 是衡量两个二进制向量之间的重叠量
- Hamming: 是对两个二进制向量中不同的向量元素数量的统计
- Partition 的意义在于通过划定分区减少数据读取,而 Sharding 的意义在于多台机器上并行写入操作。
- 静态数据 Sharding:采用静态数据分片后,如果数据规模超过服务器存储上限,您需要为集群添加更多机器并重新对数据进行分片。这个过程复杂且耗时。而且,数据分片不均衡可导致性能瓶颈,降低系统效率。
- 动态 Segment 替换:动态分配节点有益于实现更轻松的扩展和更合理的资源分配,从而确保系统延迟和吞吐量。

Round 3: 试用评估

	向量数 据库	参数	Reca ll	QPS	ART (ms)	T95 (ms)	T99 (ms)	Index Build(s)
最佳召回	Weavi ate	m: 72efCon: 512ef: 512	1 王汛共7	25.8	38.5	51.5	62.5	311 ERA7586
·飛港,7586	Qdrant	• m: 72 • efCon: 512	0.99 987	89	11.5	15.2	18 ************************************	118 EXAMPSES
有满 7586	王汛共T586 王汛共T586	ef: 512quantization: Truerescore: True			主汛井7586 生汛井7586	至刑罪7586	王用并T586 至用并T586	
最佳 QPS	Weavi ate	m: 72efCon: 256ef: 16	0.97 628	5° 114	8.5	10.2	13.5 ±m#1586	101
	Qdrant EXA 7586	m: 72efCon: 256	0.99 145	231. 9	4.2	4.9	5.3	55
煮洪7586	王汛洪 ⁷⁵⁸⁶	ef: 16quantization: True			主 汛洪7586	王汛洪7586	宝汛进7586	
和洪7586		• rescore: True				王汛洪 7586		

		Weaviate	Qdrant ERATSBS
高可靠	НА	 停止1节点后,表写入√ 	 停止1节点后,表写入√
性	王 汛洪 7586	• 停止1节点后,表读取√	● 停止1节点后,表读取√
王和洪7586		• 停止 1 节点后,表删除 ×	 停止1节点后,表删除√
	王汛洪7586	停止1节点后,单副本表 创建 ×	 停止1节点后,单副本表创建 √
王荆洪7586		• 停止 1 节点后,多副本表 创建 ×	• 停止1节点后,多副本表创建 ×
	主 汛洪7586		56
王来洪7586			主用用TS86 主用用TS86

王和洪 7586	王用洪 7586 王用洪 7586	Bug 1: 停止 Following 节 点,其他节点创建单副本表 后,该 Following 节点不		
王荆洪7586	王元洪7586	能启动 Bug 2: 停止 Following 节		
_王 凯洪7586	王汛并7586 王汛拼 ⁷⁵⁸⁶	点,其他节点删除表后, 该 Following 节点不能启 动		
王机洪7586	王汛洪 ^{T586} 王汛洪 ^{T586}	 Bug 3: 停止 Master 节点, 其他节点写入表数据,启 动 Master 节点后,从 Master 节点读取报错 		
_王 汛洪 7586	容错能力	√ Replication	√ Replication	
运维成	部署复杂度	容器部署简单,裸机部署简单	容器部署简单,裸机部编译和安装多个依赖)	署复杂 (需
王祝洪 7586	扩缩容复杂度	简单	简单	
王京洪 7586	压测运行资源消 耗	CPU: 2 cores / MEM: 1.2G	CPU: 2 cores / MEM: 8	00M 3370 ⁸ 7586
	可观测性	√ (with Grafana Dashboard)	▼	

关键评估指标:

• 业务性能: 召回率

• 技术性能: QPS / 响应时间

• 高可靠性: HA、容错能力

• 运维成本: 部署复杂度、扩缩容复杂度、副本自动修复能力、数据自动均衡能力

Bug1重现步骤:

- 1. 停止 Master 节点
- 2. 停止 Follow 节点,启动 Master 节点
- 3. 用 Master Url 节点上创建单副本表
- 4. Follow 节点启动失败

Bug2重现步骤:

- 1. 停止 Master 节点
- 2. 停止 Follow 节点,启动 Master 节点
- 3. 用 Master Url 节点上删除表
- 4. Follow 节点启动失败

Bug3重现步骤:

- 1. 停止 Master 节点
- 2. 其他节点写入表数据,表为3副本
- 3. 启动 Master 节点
- 4. 从 Master 节点读取该数据报 consistency level 错误

参考文档

What is a Vector Database?

什么是向量相似性搜索?

The Best Vector Database for Stablecog's Semantic Search

Open Source Vector Database Comparison