向量数据库

配置规格（选型）

腾讯云

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

文档目录

配置规格（选型）
资源估算

最近更新时间：2024-12-13 15:04:52

购买实例之前，首先需要评估自身业务数据规模，依据数据规模以及业务形态选择合适的索引类型。

步骤1：评估业务向量数据规模

- **向量数据维度**：确认实际使用的向量维度，比如768维、1024维等。如果您使用的是腾讯云向量数据库提供的 Embedding 模型，可以在 模型信息 中确认向量维度。
- **向量数据规模**：评估向量数据的条数，比如100万条、1000万条。评估时建议预留一定的增长空间。

步骤2：选择索引类型

根据预估的向量数据规模、业务写入与检索速度的需求，以及召回率的期望值，确定最适合业务需求的索引类型。

索引类型	使用场景	适用向量规模	召回率	检索速度
FLAT	• 暴力检索，召回率100%，但检索效率较低。 • 适用于数据量级不大，但对召回率要求严格的场景。	10万以内	最高，可保证100%召回率	慢
HNSW	• 基于图算法构建索引，可通过调整检索参数提升召回率，具体信息请参见 配置索引参数 。 • HNSW 索引的性能和召回率较为平衡，适用于绝大多数线上业务场景。	10万-1亿	95%+，可根据参数调整	快
IVF 系列	• 基于聚类算法构建的索引，可通过参数调整召回率。 • IVF 系列索引内存占用通常较低，同时采用指定部分聚类检索的方案，更适用于上亿规模的数据集。IVF 具体索引类型，请参见 IVF 系列索引应用指南 。	1亿以上	90%+，可根据参数调整	快

步骤3：内存估算

不同的索引类型，其节点内存容量估算方法存在差异。请根据下表中提供的计算公式来预算业务所需的**单份数据的内存容量**。

索引类型	单份数据内存估算计算公式	公共参数	其他参数
FLAT	单节点内存（GB）= 数据量 * 向量维度 * 4 * 1.2 / (1024^3)	• **数据量**：指向量数据的总条数。 • **向量维度**：指向量数据的维度。 • **4**：是一个常数，表示一个浮点数在内存中占用的字节数。 • **1.2**：是一个经验系数，内存 buffer 预留。 • **1024^3**：表示1GB的字节数。	-
HNSW	单节点内存（GB）= 数据量 *（向量维度 * 4 + M * 8）* 1.2/(1024^3)		**M**：为 HNSW 索引参数，指定每个节点在检索构图中可以连接多少个邻居节点。取值范围：[4,64]。M 值过大会影响写入效率。若无特殊要求，可直接使用默认参数**16**。
IVF_FLAT	单节点内存（GB）= 数据量 * 向量维度 * 4 * 1.2 / (1024^3)		-
IVF_PQ	单节点内存（GB）= 数据量 * 向量维度 * 4 * 1.2 * 压缩比 / (1024^3)		**压缩比 = M / (4 * 向量维度)**，其中，M 为 IVF_PQ 系列索引参数，指乘积量化中原始数据被拆分的子向量的数量。 • 取值要求：原始数据的向量的维度 D（即向量中元素的个数）必须能够被 M 整除，M 必须是一个正整数。 • 取值范围：[1,向量维度]。
IVF_SQx			**压缩比 = x / 32**，其中，x 决定了 IVF_SQ 索引的压缩比，如 IVF_SQ8 索引的 x = 8，压缩比为：8 / 32 = 0.25。

选择实例类型与副本数

最近更新时间：2024-12-03 14:58:53

选择实例类型

腾讯云向量数据库支持单机版与高可用版两类实例类型，适用于不同的高可用性要求场景。生产环境建议至少选择高可用版2AZ，如果是核心在线业务，可以选择3AZ 部署模式。

实例类型	部署模式	适用场景
单机版	单可用区	适用于对高可用性和容错性要求不高的场景，推荐个人、小型企业或测试/开发环境使用。不承诺服务可用性水平。
高可用版	多节点分布式架构，支持选择两可用区、三可用区	适合需要保证系统高可用性和容错性的大型企业或关键业务场景。腾讯云所承诺的服务等级水平，请参见 服务等级协议 。

副本（replicas）

副本，指每个主分片有多个相同的备份，用来容灾和负载均衡。搜索请求量越高的索引，建议设置越多的副本数，避免负载不均衡。并且，副本默认为可读，可以分担主节点的读压力，提升系统性能。当前支持的副本数量最大不超过实例的节点数量，两可用区最少为1个（不包含主节点），三可用区最少为2个，单机版不支持设置数据副本。具体如下：

- **单机版**：不支持设置数据副本，即副本数量 replicaNum = 0。
- **高可用版两可用区**：最少设置1个数据副本，最多10个数据副本。
- **高可用版三可用区**：最少设置2个数据副本，最多10个数据副本。

选择节点规格及数量

最近更新时间：2025-01-06 11:10:13

预估业务所需的数量规模，并确定向量数据库的实例类型、副本数量之后，便可计算业务所需的节点内存容量，选择符合业务需求的 CPU 与内存资源配比规格。

步骤1：选择节点类型

腾讯云向量数据库依据存储节点 CPU 与内存资源分配比例不同，分为**计算型、标准型、存储型**三类，如下表所示。请根据业务特点选择合适的节点类型。

节点类型	CPU 与内存分配比例	主要特点	适用场景	典型业务
计算型	1:2	• 主要用于快速查找和检索向量数据，计算性能非常高 • 价格相对较高	特别适用于高并发查询请求、流量大、延迟敏感的场景	实时推荐、广告投放等
标准型	1:4	• 具有均衡的计算和存储资源 • 价格适中	适用于绝大多数日常业务	RAG 应用
存储型	1:8	• 主要用于存储和管理大规模的向量数据，计算性能较弱 • 价格相对较低	特别适用于数据量大、数据增长快、查询 QPS 相对较低的场景	数据清洗、审核流等

步骤2：选择节点规格

请根据业务所需的节点类型以及估算的内存，选择节点所需的规格。若已预估业务有100万条数据，向量数据为1536维，QPS 要求不高，则根据公式单节点内存（GB）= 数据量 *（向量维度 * 4 + M * 8）* 1.2/(1024^3) 估算 HNSW 向量索引单节点所需内存 = 1.2 × 1000000 × (1536 × 4 + 20 × 8) / (1024^3) ≈ 7.04GB，可选择计算型4核8GB、标准型2核8GB或存储型1核8GB规格。

> **说明：**
> 选择规格时，除了考虑实际数据存储所需的内存之外，还需要评估额外的内存，例如，处理请求、缓存数据、并发处理等所需消耗的内存资源。否则，将会造成集合只读、新数据无法写入、新集合无法重建等异常而阻塞业务正常运行。
> • 向量数据库建议预留1.2 到1.3倍的内存 buffer，以确保系统在当前负载下稳定运行，也为未来的扩展和优化留出空间。

- 若业务需要存储更多标量字段，或者可能存在扩展，则需选择更大规格。
- 若所选规格，超出当前支持的规格范围，请 提交工单 联系腾讯云工程师。

说明：
在选择**磁盘规格**时，建议选择的磁盘容量是单节点内存容量的**两倍**。对于 IVF_PQ 或 IVF_SQ 索引，由于磁盘存储不会压缩向量数据，实际所需的磁盘空间需要根据内存空间除以压缩比来计算，以确保有足够的存储空间来容纳未压缩的向量数据。

节点类型	CPU	内存（GB）	建议向量数据规模（基于1536维32位 Float 存储下估算的向量规模，不包含标量数据）	建议向量数据规模（基于768维32位 Float 存储下估算的向量规模，不包含标量数据）
计算型	1	2	250,000	500,000
	2	4	500,000	1,000,000
	4	8	1,000,000	2,000,000
	8	16	2,000,000	4,000,000
	16	32	4,000,000	8,000,000
	24	48	6,000,000	12,000,000
	32	64	8,000,000	16,000,000
标准型	1	4	500,000	1,000,000
	2	8	1,000,000	2,000,000
	4	16	2,000,000	4,000,000
	8	32	4,000,000	8,000,000
	12	48	6,000,000	12,000,000
	16	64	8,000,000	16,000,000
存储型	1	8	1,000,000	2,000,000
	2	16	2,000,000	4,000,000
	4	32	4,000,000	8,000,000
	6	48	6,000,000	12,000,000

	8	64	8,000,000	16,000,000

步骤3：计算节点数量

腾讯云向量数据库采用分布式架构，支持多节点通信与协调，目前支持1~30个节点。默认选择3个节点来保证高可用。

实例类型	节点数量	说明
单机版	[1,30]	如果需30节点以上更大规格，请 提交工单 申请。
高可用版	• 两可用区：[2,30] • 三可用区：[3,30]	

如果单个节点的规格小于业务所需内存，那么就需要使用多个节点来组成一个集群，以满足业务需求。可以按照以下公式将单个节点所需的内存大小转化为节点数，计算需要的节点数量。

节点数 = 单节点内存估算大小 × （1 + 副本数） ÷ 单节点内存规格

- **单节点内存估算大小**：指一个节点所需的内存大小，包括向量索引内存和标量索引内存，即为 资源估算 的大小。
- **单节点规格**：指每个节点的内存规格，即为 步骤3：选择节点规格 对应的内存容量。
- **1 + 副本数**：副本指每个节点的副本数量。已在 选择实例类型与副本数 中确定副本数量。

示例：单份数据30GB，副本数 = 1（1主1副本），那总内存就是60GB，此时可以选3个24GB规格的节点。