向量数据库

产品简介

腾讯云

文档目录

产品简介
腾讯云向量数据库

最近更新时间：2023-07-27 14:46:31

本页面旨在通过回答几个问题来让您大致了解腾讯云向量数据库（Tencent Cloud VectorDB）。读完本页后，您将了解腾讯云向量数据库是什么、它是如何工作的、关键概念、为什么使用腾讯云向量数据库、支持的索引和指标、架构和相关连接方式。

腾讯云向量数据库是什么?

腾讯云向量数据库是一款全托管的自研企业级分布式数据库服务，专用于存储、检索、分析多维向量数据。该数据库支持多种索引类型和相似度计算方法，单索引支持10亿级向量规模，可支持百万级 QPS 及毫秒级查询延迟。腾讯云向量数据库不仅能为大模型提供外部知识库，提高大模型回答的准确性，还可广泛应用于推荐系统、NLP 服务、计算机视觉、智能客服等 AI 领域。

关键概念

如果您不熟悉向量数据库和相似性搜索领域，请优先阅读以下基本概念，便于您对向量数据库有一个初步的了解。更多名词解释，请阅读 关键概念 。

什么是向量?

向量是指在数学和物理中用来表示大小和方向的量。它由一组有序的数值组成，这些数值代表了向量在每个坐标轴上的分量。

什么是非结构化数据?

非结构化数据，是指图像、文本、音频等数据。与结构化数据相比，非结构化数据不遵循预定义模型或组织方式，通常更难以处理和分析。

什么是 AI 中的向量表示?

当我们处理非结构化数据时，需要将其转换为计算机可以理解和处理的形式。向量表示是一种将非结构化数据转换为嵌入向量的技术，通过多维度向量数值表述某个对象或事物的属性或者特征。腾讯云向量数据库提供的模型能力，目前在开发调试中。具体上线时间，请关注 产品动态 。

什么是向量检索?

向量检索是将向量与数据库进行比较以查找与查询向量最相似的向量的过程。相似的向量通常具有相近的原始数据，通过向量检索可以挖掘出原始非结构化数据之间的联系。

为什么是腾讯云向量数据库?

腾讯云向量数据库作为一种专门存储和检索向量数据的服务提供给用户， 在高性能、高可用、大规模、低成本、简单易用、稳定可靠、智能运维等方面体现出显著优势。 具体信息，请参见 产品优势 。

腾讯云向量数据库应用示例有哪些?

腾讯云向量数据库可进行高性能向量存储和检索，主要适用于以下应用场景。

- 大规模知识库：企业的私域数据存储在向量数据库中可构建外部知识库，帮助企业更好地管理和利用自己的数据资源。
- 推荐系统：向量数据库会基于用户特征进行向量存储与检索，并返回与用户可能感兴趣的物品作为推荐结果。
- 问答系统：向量数据库会基于问题信息进行向量存储与检索，并返回最相关的问题与对应的答案。
- 文本/图像检索：向量数据库对输入的图像和文本信息进行向量存储与检索，会找到最匹配输入信息的文本或图像结果。

腾讯云向量数据库支持哪些索引类型?

索引是数据的组织单位。您必须先声明索引类型和相似性度量，然后才能搜索或查询向量数据。目前，腾讯云向量数据库支持如下类型。具体信息，请参见 Index 。

- FLAT 索引：向量会以浮点型的方式进行存储，不做任何压缩处理。搜索向量会遍历所有向量与目标向量进行比较。
- HNSW 索引：全称为 Hierarchical Navigable Small World， 是基于图的索引，适合对搜索效率要求较高的场景。
- IVF 系列：全称为 Inverted File，IVF 系列索引的核心思想是：将高维空间划分为多个聚类，并为每个聚类构建一个倒排文件。适用于高维向量数据的快速检索。（即将支持）

腾讯云向量数据库支持哪些相似度计算方法?

在 VectorDB 中，相似度度量用于衡量向量之间的相似度。选择良好的距离度量有助于显着提高分类和聚类性能。根据输入数据形式，选择特定的相似性度量以获得最佳性能。

相似性计算方法	方法说明
内积（IP）	全称为 Inner Product，是一种计算向量之间相似度的度量算法，它计算两个向量之间的点积（内积），所得值越大越与搜索值相似。
欧式距离（L2）	全称为 Euclidean distance，指欧几里得距离，它计算向量之间的直线距离，所得的值越小，越与搜索值相似。L2在低维空间中表现良好，但是在高维空间中，由于维度灾难的影响，L2的效果会逐渐变差。
余弦相似度（COSINE）	余弦相似度（Cosine Similarity）算法，是一种常用的文本相似度计算方法。它通过计算两个向量在多维空间中的夹角余弦值来衡量它们的相似程度。

腾讯云向量数据库是如何设计的?

- 部署架构：腾讯云向量数据库采用分布式部署架构，每个节点相互通信和协调，实现数据存储与检索。客户端请求通过 Load balance 分发到各节点上。具体架构图，请参见 产品架构 。
- 逻辑架构：实例是腾讯云中独立运行的数据库环境，是用户购买向量数据库服务的基本单位。腾讯云向量数据库数据存储的一个实例集群中包括 Database 、Collection 、Document 三个逻辑层级。其中，一个实例可以包含很多个 Database，一个 Database 可以包含多个 Collection，一个 Collection 可以包含多个 Document。具体信息，请参见 逻辑结构简介 。
- 数据安全：腾讯云向量数据库的多副本设计、多可用区分布节点、API 密钥认证，并运行于私有网络环境，通过安全组控制访问来源，CAM 账户授权等多方面保护向量数据的完整性和隐私。具体信息，请参见 数据安全 。
- 鉴权方式：腾讯云向量数据库使用账号（account）和 API 密钥（api_key）的组合进行鉴权，以验证用户身份并授权其访问。具体信息，请参见 鉴权方式 。
- 连接方式：腾讯云向量数据库支持通过 HTTP 协议进行数据写入和查询等操作。具体信息，请参见 连接方式 。
- 检索方法：腾讯云向量数据库支持通过标量检索、向量检索、标量向量混合检索的方法。
  - 标量检索：是基于标量字段的检索。标量是指一个单独的数值，例如文本字段、数值字段或日期字段等，区别于向量等多维数据结构。具体信息，请参见 标量检索 。
  - 向量检索：是基于向量相似度进行的检索，通过计算向量之间的相似度来找到与查询向量最相似的文档或记录。具体信息，请参见 向量检索 。
  - 混合检索：是将标量检索和向量检索结合起来的一种方式，旨在综合利用标量属性和向量特征进行更精确和全面的检索。具体信息，请参见 混合检索 。

腾讯云向量数据库如何快速体验?

腾讯云向量数据库目前是公测阶段。公测用户免费领用实例，每个地域最多申请2个，免费试用时长3个月。若1个月内未使用实例，平台将自动回收。

序号	步骤描述	具体操作
1	申请腾讯云账号并认证。	• 如需注册腾讯云账号：请单击 注册腾讯云账号 。 • 如需完成实名认证：请单击 实名认证 。
2	测试申请	请单击 产品内侧申请 ，填写用户信息。
3	了解向量数据库所支持的规格与类型	预估数据规模，选择合适的类型与规格。具体信息，请参见 产品规格 。
4	确定向量数据库所部署的地域	当前支持的地域信息，请参见 发布地域 。
5	规划数据库实例的私有网络与安全组	具体操作，请参见 创建私有网络 与 创建安全组 ，并同时设置安全组入站规则。
6	购买实例	具体操作，请参见 新建数据库实例 。购买实例中，直接选择上一步已准备的私有网络与安全组。

7	申请与腾讯云向量数据库在同一地域同一个 VPC 内的 Linux 云服务器 CVM。	具体操作，请参见 快速配置 Linux 云服务器 。
8	连接并操作向量数据库。	连接并写入数据库 ，本文使用 API 接口 从创建 DataBase 到 插入数据、检索数据到最终删除数据，均给出了具体的使用示例。您可以简单并快速体验向量数据库。
9	管理向量数据库实例	您可以体验通过控制台直接管理实例，查看实例状态或销毁实例。具体操作，请参见 管理实例 。
10	智能运维	您可以在控制台查看监控数据库实例的各项指标。具体信息，请参见 实例监控 。目前仅支持对节点信息的监控，后续还会支持更丰富的监控项目。

开发者工具

腾讯云向量数据库支持丰富的 API 接口，以促进 DevOps。具体操作，请参见 API 文档 。

产品优势

最近更新时间：2023-08-21 15:36:41

腾讯云向量数据库（Tencent Cloud VectorDB）作为一种专门存储和检索向量数据的服务提供给用户， 在高性能、高可用、大规模、低成本、简单易用、稳定可靠等方面体现出显著优势。

高性能

向量数据库单索引支持10亿级向量数据规模，可支持百万级 QPS 及毫秒级查询延迟。

高可用

向量数据库提供多副本高可用特性，其多可用区和三节点的架构可用性可达99.99%，显著提高系统的可靠性和容错性，确保数据库在面临节点故障和负载变化等挑战时仍能正常运行。

大规模

向量数据库架构支持水平扩展，单实例可支持百万级 QPS，轻松满足 AI 场景下的向量存储与检索需求。

低成本

只需在管理控制台按照指引，简单操作几个步骤，即可快速创建向量数据库实例，全流程平台托管，无需进行任何安装、部署和运维操作，有效减少机器成本、运维成本和人力成本开销。

简单易用

支持丰富的向量检索能力，用户通过 HTTP API 接口即可快速操作数据库，开发效率高。同时控制台提供了完善的数据管理和监控能力，操作简单便捷。

稳定可靠

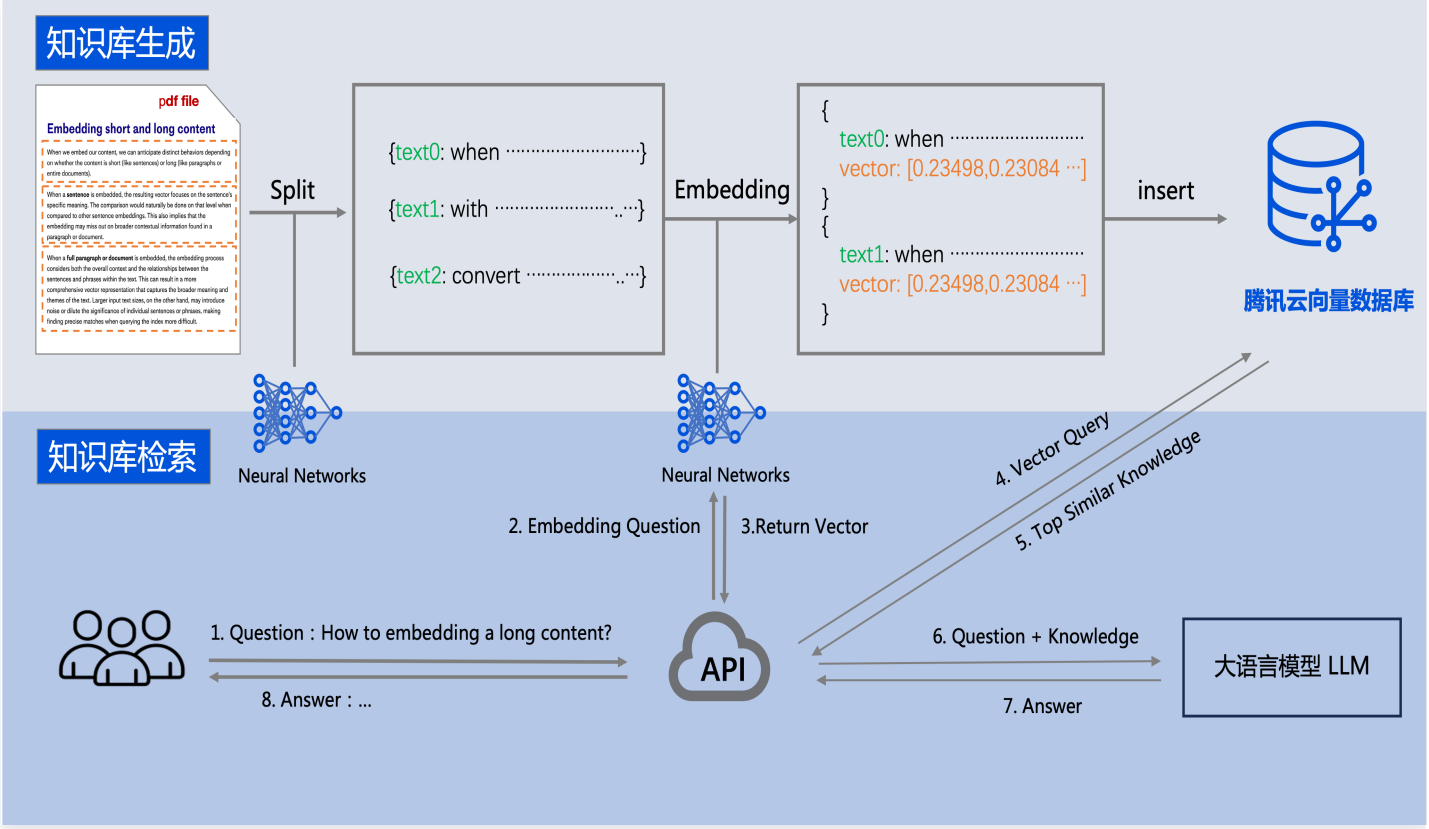
向量数据库源自腾讯集团自研的向量检索引擎 OLAMA，近40个业务线上稳定运行，日均处理的搜索请求高达千亿次，服务连续性、稳定性有保障。

应用场景

最近更新时间：2023-07-25 09:52:46

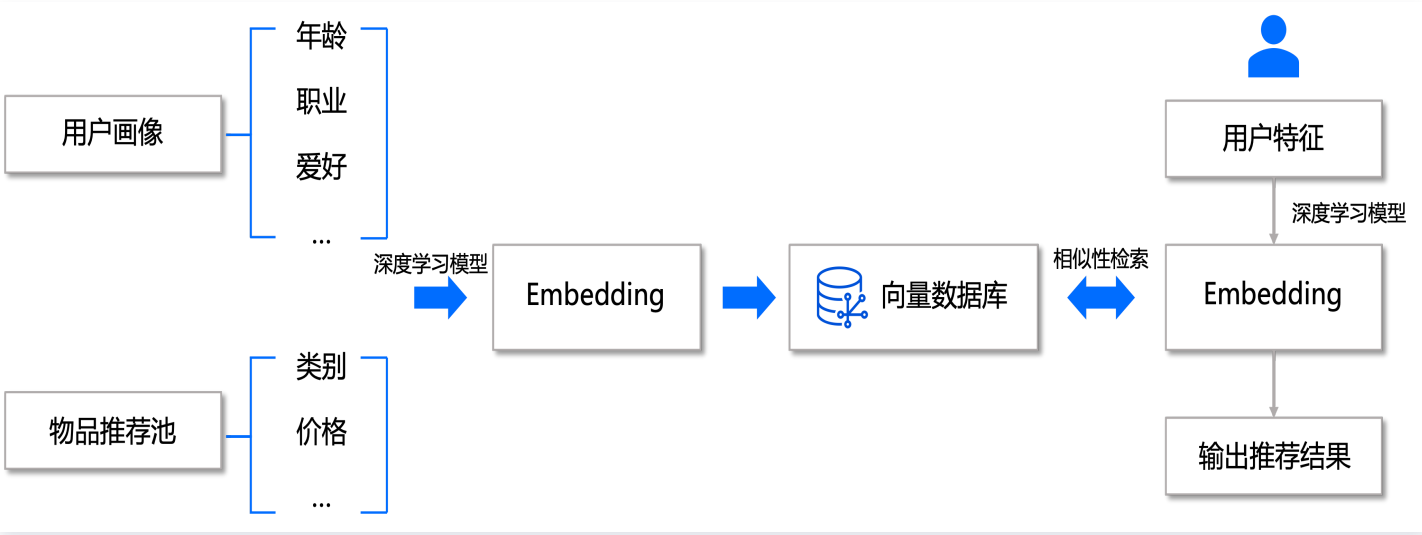
大模型知识库

腾讯云向量数据库可以和大语言模型 LLM 配合使用。企业的私域数据在经过文本分割、向量化后，可以存储在腾讯云向量数据库中，构建起企业专属的外部知识库，从而在后续的检索任务中，为大模型提供提示信息，辅助大模型生成更加准确的答案。



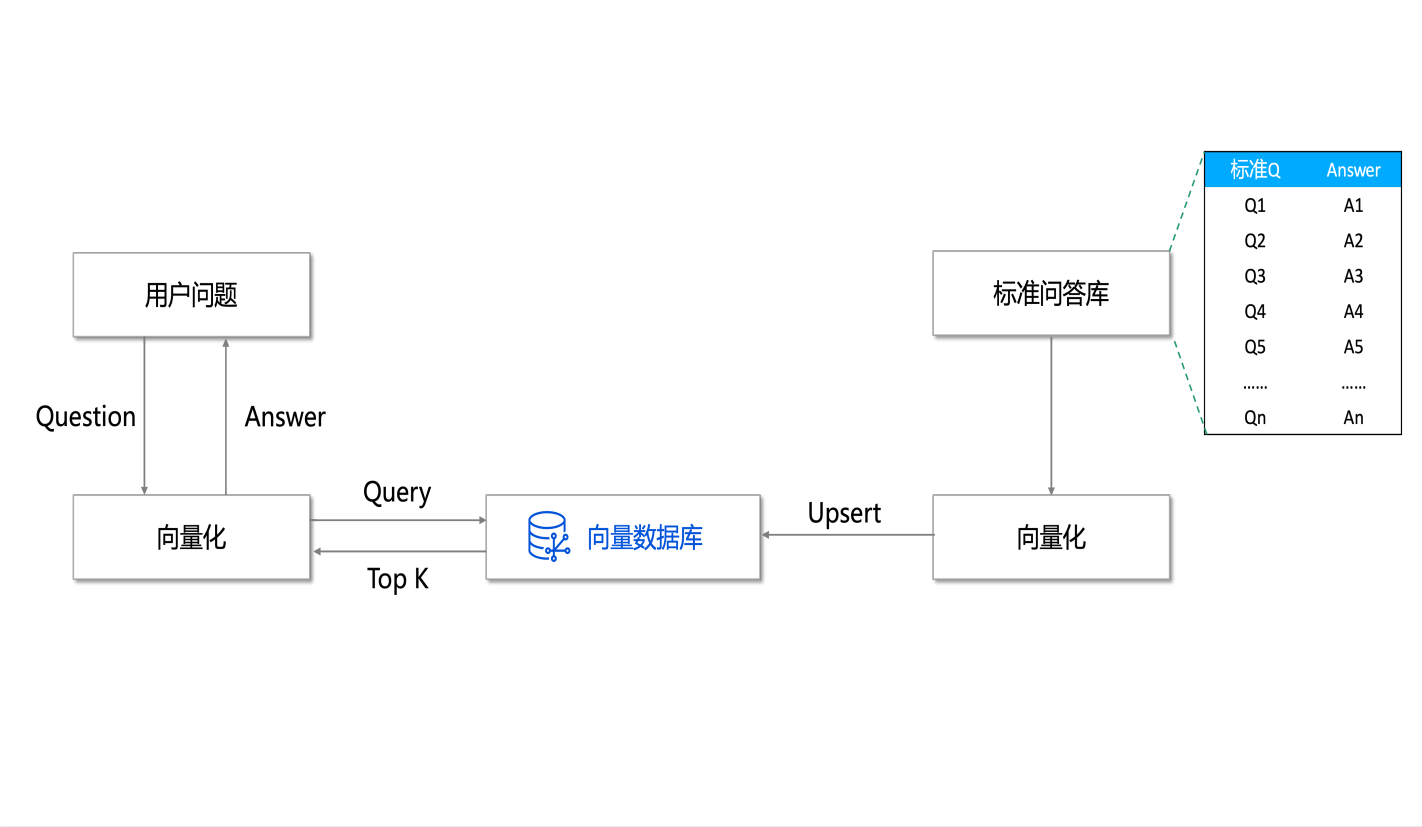
推荐系统

推荐系统的目标是根据用户的历史行为和偏好，向用户推荐可能感兴趣的物品。在这种场景下，将用户行为特征向量化存储在向量数据库。当发起推荐请求时，系统会基于用户特征进行相似度计算，然后返回与用户可能感兴趣的物品作为推荐结果。



问答系统

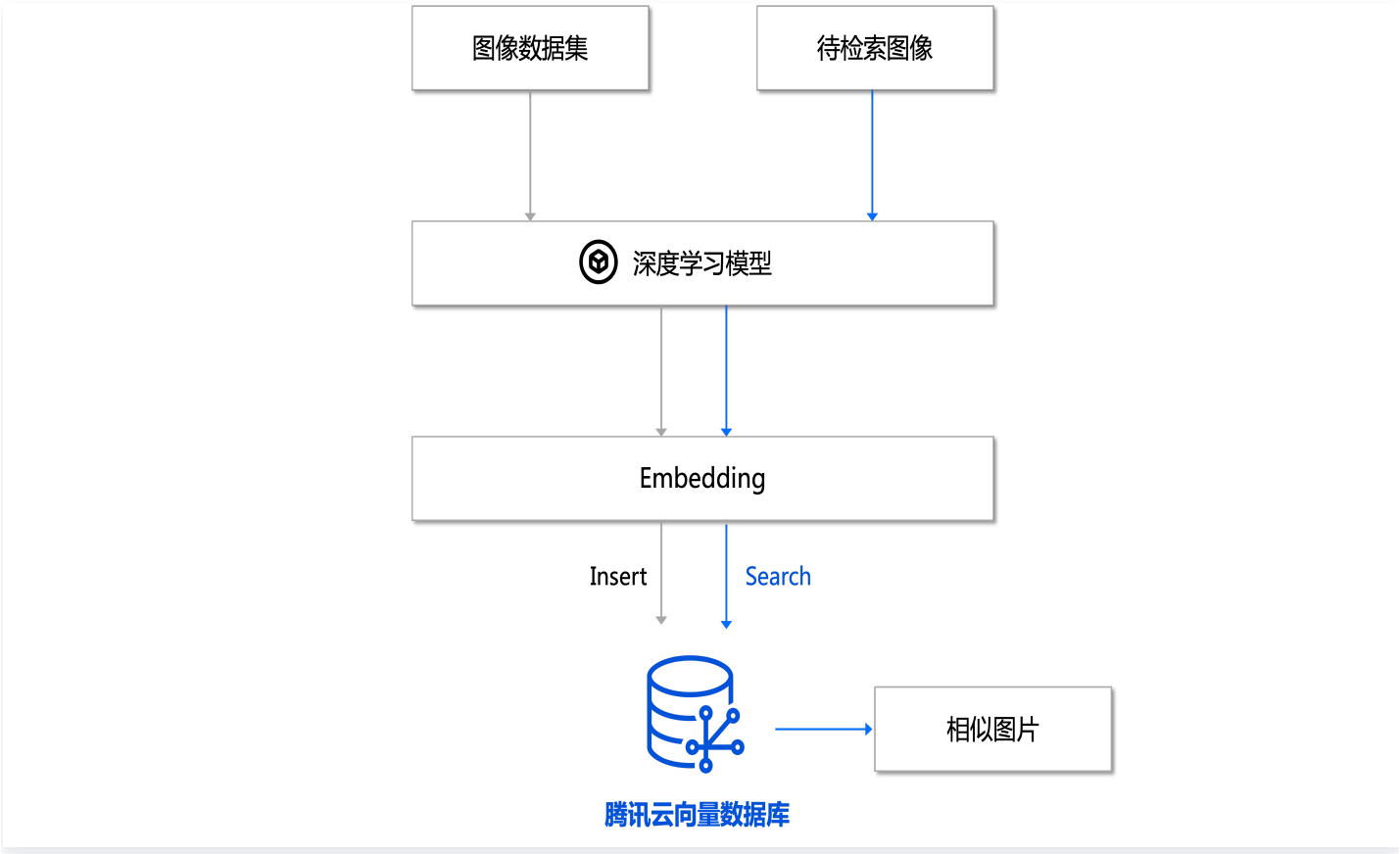
智能问答系统是一种能够回答用户提出问题的智能应用，通常使用 NLP 服务和深度学习等技术实现。在问答系统中，问题和答案通常被转换为向量表示，并存储在向量数据库中。当用户提出问题时，问答系统可以通过计算向量之间的相似度，检索最相关的问题信息并返回对应的答案信息。因此，使用向量数据库来存储和检索相关的向量数据，可以提高问答系统的检索效率和准确性。



问答系统的应用场景非常广泛，例如智能客服、智能助手、智能家居等。在这些场景中，用户可以通过自然语言提问获取相关信息，例如查询产品信息、控制家居设备等。通过使用向量数据库来存储和检索相关的向量数据，问答系统可以更快速、准确地响应用户的请求，提高用户体验。

文本/图像检索

文本/图像检索任务是指在大规模文本/图像数据库中搜索出与指定图像最相似的结果，在检索时使用到的文本/图像特征可以存储在向量数据库中，通过高性能的索引存储实现高效的相似度计算，进而返回和检索内容相匹配的文本/图像结果。下图以图像检索为例介绍任务流程。



关键概念

最近更新时间：2023-07-24 19:36:42

向量（Vector）

向量可以理解为一组数值的有序集合，通常用于表示某个对象或事物的属性或者特征。这些数值可以有不同的维度，每个维度都表示一个属性或特征。在机器学习和人工智能领域，向量常用于表示图像、文本、音频等数据，通过计算向量之间的距离或相似度来实现分类、聚类、检索等任务。

OLAMA

OLAMA 是腾讯自研的向量引擎，具有高性能、高可用、简单易用等特点。它支持单索引10亿级向量规模，适用于 AI 运算、检索场景，已稳定服务于近40个线上业务。

实例（Instance）

实例是腾讯云中独立运行的数据库环境，是用户购买向量数据库服务的基本单位，以单独的进程存在。一个数据库实例可以包含多个由用户创建的数据库。 您可以在控制台创建、修改和删除实例。实例之间相互独立、资源隔离，相互之间不存在 CPU、内存、持久内存、IO 等抢占问题。

数据库（Database）

数据库是按照数据结构来组织、存储和管理数据的仓库，一个实例可以创建多个 Database。

集合（Collection）

在向量数据库中，集合是指一组文档组，类似于关系型数据库中的表，其中可包含多条文档数据。集合没有固定的结构，可以插入不同格式和类型的数据。向量数据库支持集合维度的多分片、多副本特性，可以在创建集合时按需指定分片数和副本数。

文档（Document）

在向量数据库中，集合 可以看作是一个表格，而 Document 可以看作是表格中的一行数据。每个 Document 代表一个完整的文档对象，包含了多个 Field，每个 Field 表示文档中的一个属性或字段。向量数据库的文档是一组键值对（key：value），每个文档都有一个唯一主键（id）和一个向量字段（vector）。在插入文档时，向量数据库不需要设置相同的字段，可以在插入数据时增加或删除字段。

字段（Field）

每个 Field 是一个键值对（key：value），表示文档中的一个属性或者字段。每个 Field 都有自己的类型和取值范围，可以是字符串、数字等不同类型的数据。

节点（Node）

从向量数据库集群的资源角度来看，节点是用于存储数据的资源单位。一个运行中的向量数据库实例通常包含很多个节点，集合的多个副本和分片会分布在若干个节点上。节点是组成向量数据库集群的基本单元之一。

分片（Shard）

为了支持更大规模的数据，集合一般会按某个维度分成多个部分，每个部分就是一个分片，分布在若干个节点（Node）上。为了保证可靠性和可用性，同一个集合的多个分片会分布在不同节点（Node）上。

副本（Replica）

同一个分片（Shard）的备份数据，一个分片至少会有2个副本。副本分片作为硬件故障时保护数据不丢失的冗余备份，并为向量检索和文档查询等读操作提供服务，确保数据库在面临节点故障和负载变化等挑战时仍能正常运行。

索引（Index）

索引是一种特殊的数据结构，用于快速查找和访问数据，存储在内存中。索引本身并不存储数据，而是存储指向数据存储位置的指针或键值对。Tencent Cloud VectorDB 支持 FLAT、HNSW 等常见的向量索引。索引介绍详见向量检索 。

KNN（K-Nearest Neighbor Search）

KNN 指的是最近邻搜索（K-Nearest Neighbor Search），是一种基于暴力搜索的方法，它的原理是：计算待查询向量与数据库中所有向量之间的距离，然后按照距离从小到大排序，选择距离最近的 K 个向量作为查询结果。KNN 算法的优点是可以保证精确的结果，但是对于大规模的向量数据，计算量会非常大，效率较低。

ANN（Approximate Nearest Neighbor Search）

ANN 表示近似最近邻搜索（Approximate Nearest Neighbor Search），是一种用于高维数据空间中快速查找最近邻点的方法。与精确最近邻搜索相比，ANN 牺牲了一定的精度以换取更高的搜索速度，因此在处理大规模数据集时具有较高的效率。ANN 方法通常会对数据进行预处理，从而在查询时减少计算距离的次数。ANN 算法的优点是速度快、效率高，但是相对于 KNN 算法来说，其结果可能不够精确。

HNSW（Hierarchical Navigable Small World）

HNSW 是一种基于图的高维向量相似性搜索算法，全称为：Hierarchical Navigable Small World。它通过构建一张图来表示向量之间的相似度关系，并使用一些优化策略来加速搜索过程。

产品规格

最近更新时间：2023-08-24 14:06:41

腾讯云向量数据库（Tencent Cloud VectorDB）采用分布式部署架构，每个节点相互通信和协调，实现数据存储与检索。客户端请求通过 Load balance 分发到各节点上。具体信息，请参见 产品架构 。

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**存储型**和**计算型**两类。

- **存储型**：主要用于存储和管理大规模的向量数据，其主要优势在于：提供低查询延迟，能够高效地存储和管理向量数据，特别适用于数据量大、数据增长快、查询 QPS 相对较低的场景，例如：人脸识别、图像搜索等。
- **计算型**：主要用于快速查找和检索向量数据，支持高并发的查询请求，其主要优势在于：提供更高的查询 QPS 和更低的查询延迟，适用于流量大、延迟敏感的场景，例如：实时推荐、广告投放等。

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3~10个节点。

节点规格

节点类型不同，对应的产品规格有差异，详细信息，请参见下表。

说明：

- 存储型节点规格 S.2xLARGE、S.4xLARGE、S.8xLARGE，需 提交工单 申请。
- 计算型节点规格 P.LARGE、P.2xLARGE、P.4xLARGE、P.6xLARGE，需 提交工单 申请。

节点类型	节点规格	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）
存储型	S.MEDIUM	1	8	1,000,000	2,000,000
	S.LARGE	2	16	2,000,000	4,000,000
	S.2xLARGE	4	32	4,000,000	8,000,000
	S.4xLARGE	8	64	8,000,000	16,000,000

	S.8xLARGE	16	128	16,000,000	32,000,000
计算型	P.SMALL	2	4	500,000	1,000,000
	P.MEDIUM	4	8	1,000,000	2,000,000
	P.LARGE	8	16	2,000,000	4,000,000
	P.2xLARGE	16	32	4,000,000	8,000,000
	P.4xLARGE	32	64	8,000,000	16,000,000
	P.6xLARGE	48	96	12,000,000	24,000,000

发布地域

最近更新时间：2023-07-24 19:36:42

腾讯云数据库托管机房分布在全球多个位置，这些位置节点称为地域（Region），每个地域又由多个可用区（Zone）构成。每个地域（Region）都是一个独立的地理区域。每个地域内都有多个相互隔离的位置，称为可用区（Zone）。每个可用区都是独立的，但同一地域下的可用区通过低时延的内网链路相连。腾讯云支持用户在不同位置分配云资源，建议用户在设计系统时考虑将资源放置在不同可用区以屏蔽单点故障导致的服务不可用状态。

地域

腾讯云不同地域之间隔离，保证不同地域间最大程度的稳定性和容错性。建议您选择最靠近您用户的地域，可降低访问时延、提高下载速度。用户启动实例、查看实例等操作都是区分地域属性的。

> **注意：**
> - 同地域下（保障同一账号，且同一个 VPC 内）的云资源之间可通过内网互通，可以直接使用 内网 IP 访问。
> - 不同地域之间网络隔离，不同地域之间的云产品默认不能通过内网互通。
> - 处于不同私有网络的云产品，可以通过 云联网 进行通信，此通信方式较为高速、稳定。

可用区

可用区（Zone）是指腾讯云在同一地域内电力和网络互相独立的物理数据中心。目标是能够保证可用区间故障相互隔离（大型灾害或者大型电力故障除外），不出现故障扩散，使得用户的业务持续在线服务。通过启动独立可用区内的实例，用户可以保护应用程序不受单一位置故障的影响。

命名规则

地域、可用区名称是对机房覆盖范围最直接的体现，为便于客户理解，命名规则如下：

- 地域命名采取【覆盖范围 + 机房所在城市】的结构，前半段表示该机房的覆盖能力，后半段表示该机房所在或临近的城市。
- 可用区命名采取【城市 + 编号】的结构。

支持地域和可用区

> **说明**
> 不同地域所开放的资源可能因资源售罄而缺少，之前已售罄的资源可能又得到了重新补给。资源的开放情况会根据实际业务使用情况随时评估调整，请以控制台购买页所开放的资源为准。

中国

地域（region）		可用区（zone）
华南地区（广州）	ap-guangzhou	存储节点默认为多可用区（三可用区）分布，不支持用户自定义。
华东地区（上海）	ap-shanghai	
华北地区（北京）	ap-beijing	