向量数据库

关于向量数据库

腾讯云

文档目录

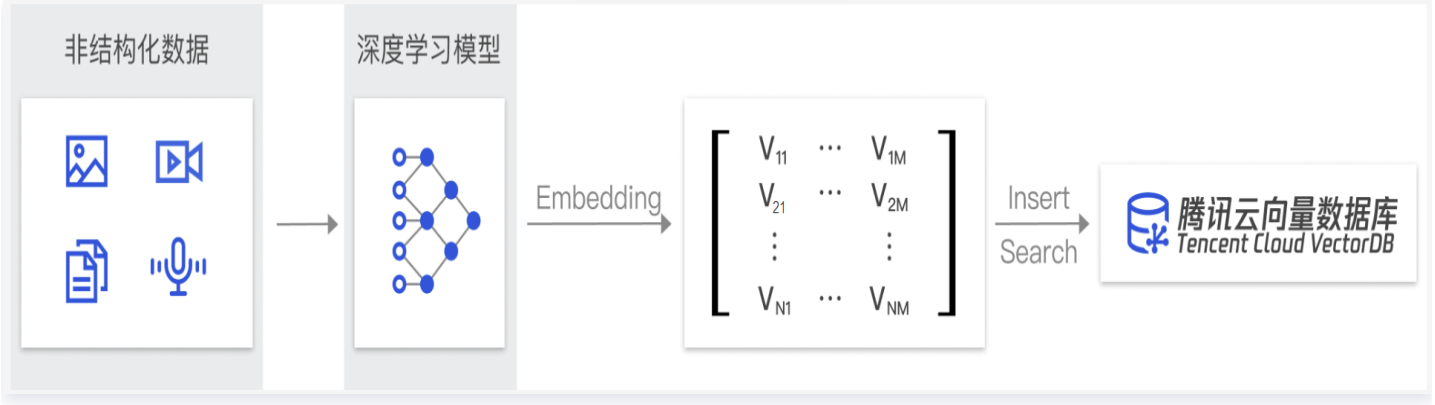
关于向量数据库
什么是腾讯云向量数据库

最近更新时间：2025-07-09 14:49:52

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?

随着互联网的普及，越来越多的非结构化数据如电子邮件、图像、音视频和文本等变得普遍。为了让计算机能够理解和处理这些非结构化数据，使用嵌入技术将这些数据转换为向量形式。
腾讯云向量数据库（Tencent Cloud VectorDB）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、索引、检索、管理由深度神经网络或其他机器学习模型生成的大量多维嵌入向量。作为专门为处理输入向量查询而设计的数据库，它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，高达百万级 QPS 及毫秒级查询延迟。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 名词解释 。

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。当前支持的模型能力，请参见 Embedding 。

什么是向量相似性检索?

向量检索是一种基于向量空间模型的信息检索方法。向量数据库通过相似度计算方法计算两个向量之间的相似距离来分析它们之间的相关性。如果两个嵌入向量非常相似，则意味着原始数据源也相似。

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 。

支持哪些索引和指标?

在建表时，需指定向量的 索引类型（如 HNSW 等）与 相似度计算方法。数据库存储的向量将会按照指定的索引类型进行索引。那么，在向量检索时，便会依据索引并使用已选择的相似性计算方法进行匹配，快速高效地获取目标向量。如果不指定索引类型，向量数据库将默认进行暴力搜索。

索引类型

向量数据库支持的向量索引类型大部分采用近似最近邻搜索（ANNS）。目前，支持如下类型。

- 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。
- BIN_FLAT：用于处理二进制向量数据的索引类型。二进制向量通常表示数据点，每个向量由一系列二进制值（0或1）组成，BIN_FLAT 索引并不对二进制向量进行压缩，适用于图像二值数据或者其他二进制特征数据的处理场景。

相似度计算方法

选择良好的距离度量有助于显著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方法，获得数据库最佳性能。目前支持的相似度计算方法如下表所示。

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。所得值越大越与搜索值相似。
汉明距离（Hamming Distance）	汉明距离是一种简单而有效的相似度计算方法，尤其适用于处理等长的二进制数据。通过计算两个字符串对应位置上不同字符的数量来定义，如果字符不同，那么它们之间的汉明距离就会加一。对于长度为 n 的二进制向量，如果两个向量的对应位有 d 个不同，则它们的汉明距离为 d。

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，最终筛选用户可能感兴趣的物品推荐给用户。
- 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。客户端请求通过 Load Balancer 分发到各节点上，每个节点相互通信和协调，实现数据存储与检索。
- 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。更多信息，请参见 设计架构 。

开发者工具

开发者工具	API
HTTP API	API 接口
Python SDK	Python SDK Demo
Java SDK	Java SDK Demo

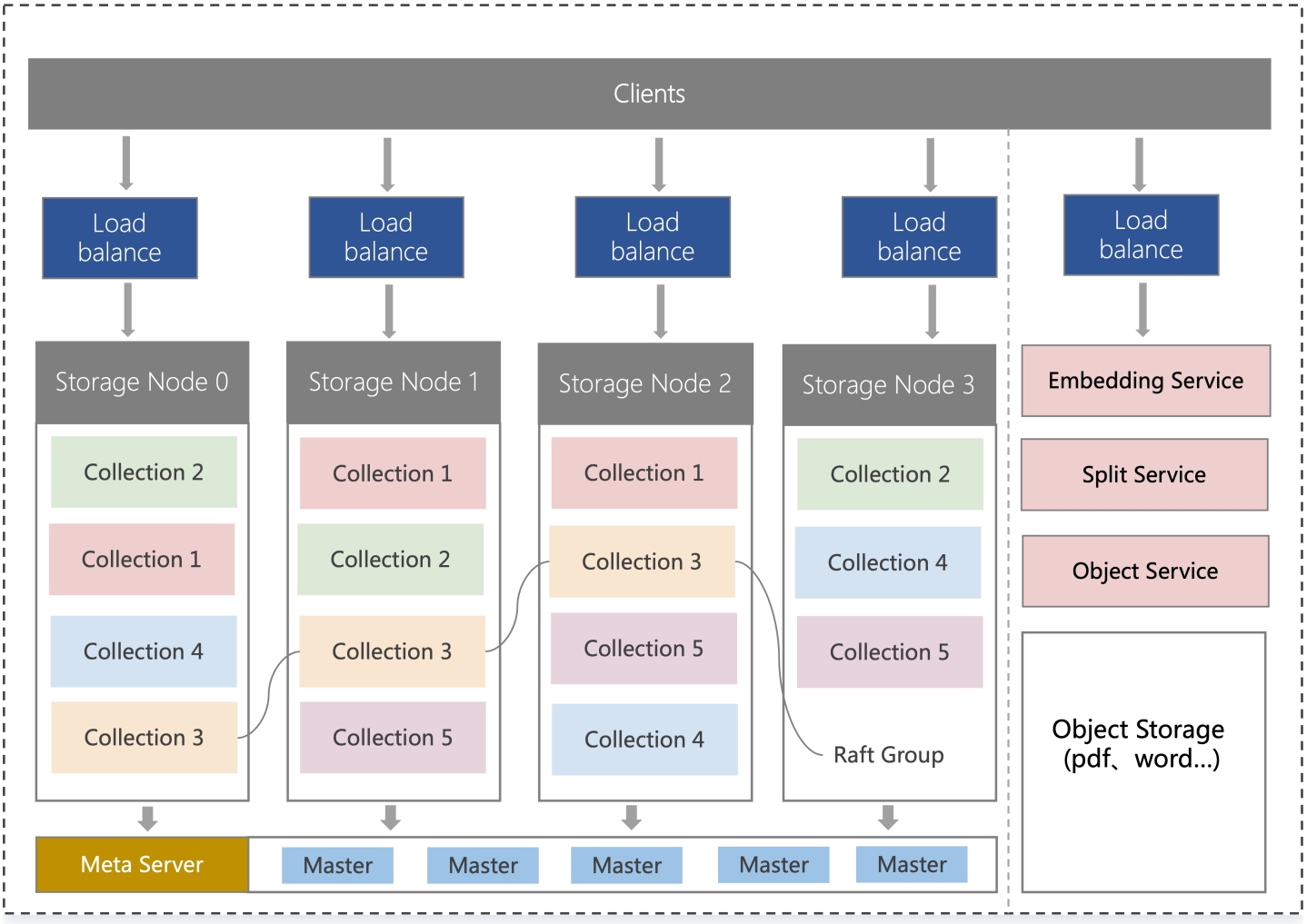
Go SDK	Go SDK Demo
C++ SDK	C++ SDK Demo

设计架构

最近更新时间：2024-12-02 19:01:12

部署架构

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

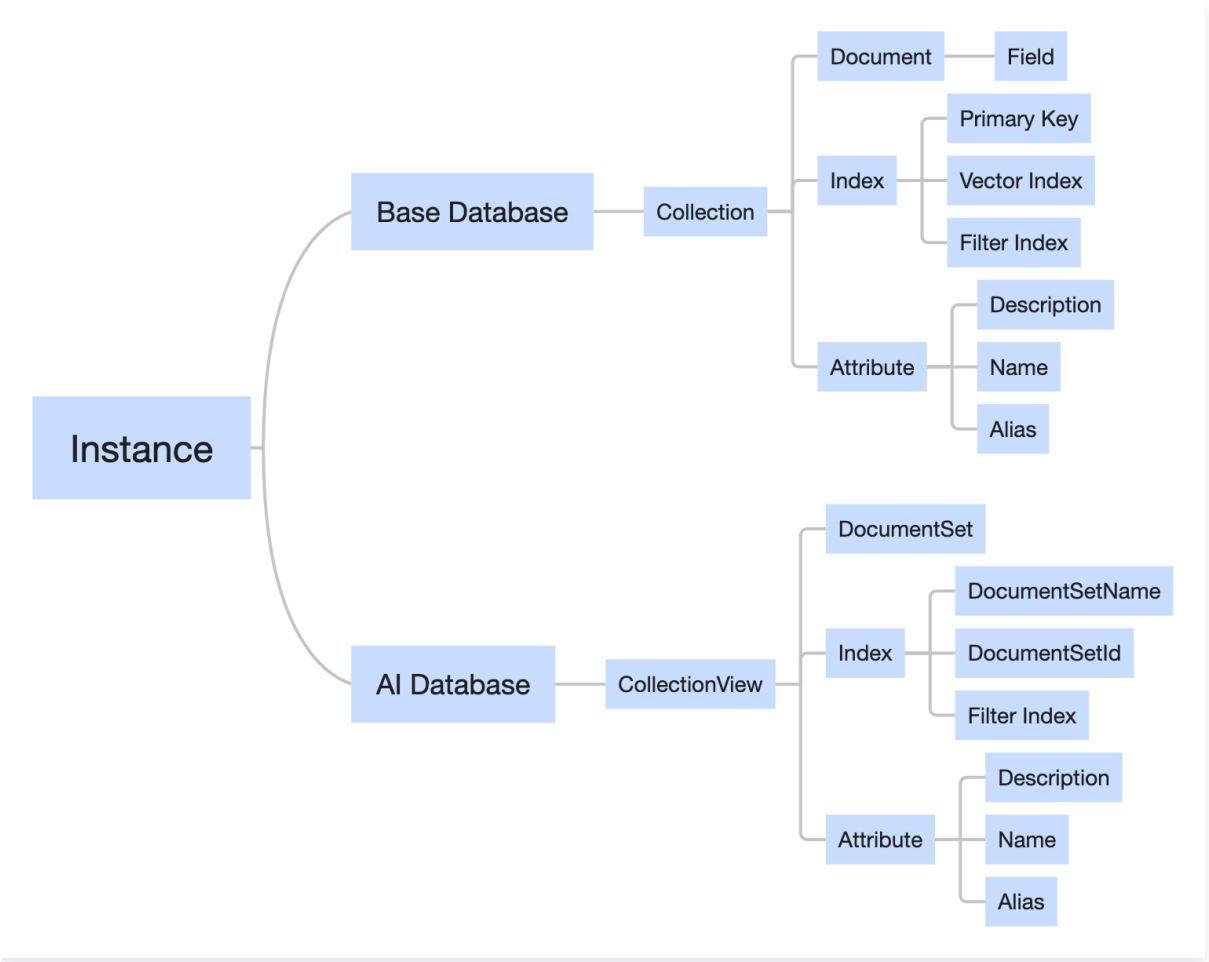


- **负载均衡（Load Balancer，LB）**：是对多台后端服务器进行流量分发的服务。向量数据库集群架构节点数量 >= 3，自动通过 Load balance 来均衡访问。
- **分布式 Storage Node**：向量数据库集群由多个节点构成，每个节点均可直接进行读/写操作，负责数据的计算及存储。Collection 是向量数据的基本组织形式，将向量集合拆分成多个分片，并分配到不同的节点上进行存储和处理；每个分片还会在其他节点上同步产生多个副本，以保证数据库服务的可扩展性与高可用性。
- **Meta Server**：集群管理模块，由一组 Master 节点组成，负责存储集群的节点信息、数据分片信息等元数据信息。
- **Embedding Service**：是一种将非结构化数据（如文本、图像、音频等）转换为向量表示的服务，从而方便进行分析、聚类等操作。具体信息，请参见 Embedding 介绍 。

- **Split Service**：是一种将文本拆分成短语或句子等的服务。
- **Object Service**：负责将数据批量导入到指定集合，支持多种数据导入格式。
- **Object Storage**：用于存储和管理数据导入服务中上传的数据文件。

逻辑架构

在 VectorDB 中，数据库实例是最高层级的容器，用于存储多个 Database，而一个 Database 包含多个集合。Collection 与 CollectionView 是具有相同数据结构和语义的 Document 容器。其逻辑结构，如下图所示。



基本概念	含义
实例（Instance）	用户购买向量数据库服务的基本单位。
Base Database	可直接操作向量数据的数据库，包含一系列相关 Collection 的集合。 • 该类数据库可直接写入向量数据，可对存储的向量数据进行修改、清理等。 • 该类数据库可写入文本信息，通过 Embedding 自动向量化，存储原始文本与向量数据，可修改原始文本信息，进而修改数据。

AI Database	指专门用于 AI 套件上传并存储文件的向量数据库， 包含一系列相关 CollectionView 集合视图。该类数据库对已上传的文件不支持更新文件内容，无法直接对向量数据进行操作。
Collection	Base 类数据库具有相似属性文档组的集合，相当于关系型数据库中的表。一个 Database 中可以包含多个 Collection。 • **副本与分片：**创建 Collection 时需根据预估数据规模和业务需求，指定集合所需的副本数和分片数。分片是将数据集合拆分为多个更小的部分，每个部分放在不同的节点，以提高系统性能；而副本则是对每一个数据分片的备份，保障系统的可靠性。 • **动态 Schema：**创建 Collection 时支持根据数据属性按需自定义字段，设计灵活、易用。
CollectionView	AI 类数据库中一组具有相似属性的文档组集合，相当于关系型数据库中的表。只是 CollectionView 是针对 AI 类数据库文档组的集合视图，由多个 DocumentSet 组成。每个 DocumentSet 存储一组数据，对应一个文件。多个 DocumentSet 构成一个CollectionView。 • **副本与分片：**创建 CollectionView 无需指定实例副本与分片，系统根据文件内容的大小自动调整分配资源。 • **动态 Schema：**创建 CollectionView 支持灵活扩展文件 Metadata 信息的标量字段。
Document	Base 类数据库中一个 Collection 中可以包含多个 Document。每个 Document 由多个**字段（Field）**组成形成一条向量数据。每个 Field 以键值对（key：value）的形式存储，相当于关系型数据库中的数据字段。 **说明：** **动态 Schema：**腾讯云向量数据库（Tencent Cloud VectorDB）支持在创建 Collection 时灵活扩展属性字段，同时支持在更新数据时新增字段。这意味着用户无需预先定义所有的字段，可以根据需要在插入数据时自动识别并调整模式。这种设计使得 VectorDB 更加灵活和易于使用，同时也避免了预定义所有字段带来的限制。
DocumentSet	DocumentSet 是 AI 类数据库中 CollectionView 中的一个概念，用于存储文件的单元。在 CollectionView 中，一个文件会被拆分成多个 Document，这些 Document 组成了一个完整的文件数据，也就是一个DocumentSet。DocumentSet 可看作是一组相关的 Document 的集合，它们对应着同一个文件的数据。 **说明：**

	动态 Schema：腾讯云向量数据库（Tencent Cloud VectorDB）支持在创建 CollectionView 时灵活扩展文件 Meta 信息的标量字段，同时支持更新或新增标量字段。
Index	一种用于加速检索的数据结构，分为主键检索、向量索引和 Filter 索引。 • **主键索引（Primary Key Index）**：是一种用于快速查找特定行的数据库索引类型。在主键索引中，每个行都有一个唯一的标识符，称为主键。主键索引使用这些主键来快速查找特定行，而不需要扫描整个表。腾讯云向量数据库默认将 Document id 或 DocumentSet ID 作为主键来构建主键索引。 • **向量索引（Vector Index）**：是一种用于快速查找相似向量的数据库索引类型。在向量索引中，每个向量都被索引，并且可以通过计算它们之间的相似度来快速查找最相似的向量。向量索引通常用于处理大规模的高维数据，当前所支持的相似度算法，请参见 相似性计算 。 • **Filter 索引（Filter Index）**：是建立在标量字段的索引。标量字段被建立 Filter 索引之后，向量检索时，将依据 Filter 指定的标量字段的条件表达式来查找相似向量，请参见 Filter 条件表达式 。
属性（Attribute）	标识 Collection 或 CollectionView 的其他属性，基本属性如下所示，可以根据具体需求进行自定义和扩展，以适应不同的业务场景和数据类型。 • **Name**：集合名称。 • **Description**：集合的备注信息，用于描述集合的业务属性或其他信息。 • **Alias**：集合别名。别名可以是一个简短的字符串，方便标识和访问对应的集合。当使用别名访问时，用户不感知真实集合名的变化，适用于数据迁移到新集合后的一键切换场景。

安全设计

- **多副本设计**：腾讯云向量数据库（Tencent Cloud VectorDB）的多副本设计是指将数据库的数据复制到多个副本中，以提高系统的可用性和容错性。在多副本设计中，用户可以在 Collection 级别指定副本数，每个副本分布在不同的节点。当某个节点发生故障时，系统可以自动从其他节点的副本中获取数据，从而保证系统的连续性和稳定性。同时，多副本设计还可以提高系统的性能和可扩展性。
- **多可用区**：腾讯云向量数据库默认为多可用区分布式集群化部署，这意味着将数据的副本分布在多个可用区域，以确保系统的高可用性和容错性。每个可用区都是一个独立的数据中心，拥有自己的计算、存储和网络资源，避免单点故障和保证系统的稳定性。当前默认为多可用区部署，暂不支持用户自定义可用区。
- **登录鉴权**：腾讯云向量数据库（Tencent Cloud VectorDB）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。客户端向服务端发起 HTTP 请求时，需要在请求的 Header 中携带 account 和 api_key。服务端在接收到请求后，会对 api_key 的合法性进行判断。如果 api_key 合法，则服务端会根据请求的内容做出正确的响应。鉴权时 Token 的格式为 `account=root&api_key=xxx` 。
- **私有网络**：腾讯云向量数据库运行于**私有网络**环境中。私有网络（Virtual Private Cloud，VPC）是一块在腾讯云上自定义的逻辑隔离网络空间，基于隧道技术，在物理网络上构造虚拟网络，使用虚拟化技术，实现不同

私有网络之间内网完全隔离。这种网络环境为用户提供了独立、隔离的安全云网络，可以有效地保护用户的数据安全和隐私，确保数据库的安全性和可靠性。

- **安全组**：一种虚拟防火墙，具备有状态的数据包过滤功能，用于设置云服务器、负载均衡、云数据库等实例的网络访问控制，控制实例级别的出入流量，是重要的网络安全隔离手段。腾讯云向量数据库支持配置安全组，以控制云数据库实例的网络访问，从而保护云资源的安全性。安全组支持基于 IP 地址、端口号、协议等多种条件进行访问控制，可以根据不同的业务需求，设置不同的访问规则。
- **访问管理（Cloud Access Management，CAM）**：腾讯云提供的一套 Web 服务，用于帮助客户安全地管理腾讯云账户的访问权限，资源管理和使用权限。腾讯云向量数据库通过 CAM 可以创建、管理和销毁用户（组），并通过身份管理和策略管理控制哪些人可以使用哪些数据库资源，资源细粒度控制，提供企业级的安全防护。

通信协议

腾讯云向量数据库（Tencent Cloud VectorDB）支持通过 **HTTP** 协议进行数据写入和查询等操作。您可以将不同类型的请求消息以 JSON 格式放入 HTTP 请求消息 Body 中，将请求发送到 VectorDB 的HTTP API 地址即可。VectorDB 将自动解析请求消息 Body 中的 JSON 数据，并将其存储到数据库中。

- VectorDB 使用 HTTP 协议中的 GET、POST 等标准 HTTP 方法来执行不同的操作。同时，VectorDB 的 HTTP API 还使用 HTTP 状态码和响应头来表示请求的结果和状态，以及使用 URL 来标识资源。这种设计风格使得 VectorDB 的 HTTP API 易于使用和理解，并且可以方便地与其他系统进行集成。
- VectorDB 通过提供 VPC 网络隔离和发送请求时在 HTTP 头部信息中携带实例账号和 API 密钥进行身份认证来保证数据的安全性，通过 JSON 格式的结构体进行数据交换，每个请求都会返回一个标准的 HTTP 响应状态码和响应内容。若操作失败，用户可以根据响应内容获取到具体错误信息。
- 在发送请求时，客户端在 HTTP 头部信息中携带账号和 API 密钥，这些信息会以 HTTP 参数的形式进行传递（HTTP 参数格式为account=root&api_key=xxx）。服务端会对 API 密钥的合法性进行判断，并根据判断结果做出相应的响应。

您可以使用 `curl` 命令发送 HTTP 请求消息，如下为创建数据库的请求示例。其中，请求头部信息中携带了账号和 API 密钥，分别为“root”和“abcdefg”。

```
curl -i -X POST \
  -H 'Content-Type: application/json' \
  -H 'Authorization: Bearer account=root&api_key=abcdefg' \
  http://10.0.X.X:80/database/create \
  -d '{
    "database": "db-test"
  }'
```

参数	参数解释
curl	执行命令

-X	• 指定 HTTP 请求方法，常用的请求方法如下所示：
	  ○ GET：从数据库集群获取资源。
	  ○ POST：向数据库提交数据，用于创建资源或处理数据。

-H	指定 HTTP 请求头，该示例设置了两个请求头，含义如下所示：
	• 'Content-Type: application/json' ，表示请求的消息 Body 是 JSON 格式的。
	• 'Authorization: Bearer account=root&api_key=abcdefg' \http://10.0.X.X:80/database/create，表示请求消息的安全认证信息。
	  ○ account=root：向量数据库账号信息。数据库当前仅支持 root 账号。
	  ○ api_key=abcdefg：API 请求密钥信息。
	  ○ http://10.0.X.X:80/database/create：请求访问的 URL 地址。由以下参数拼接而成。
	  ○ 10.0.X.X：指向量数据库的内网 IP 地址。
	  ○ 80：向量数据库默认的访问端口。
	  ○ /database/create：指创建向量数据库的 API 接口。

-d	指定 HTTP 请求消息 Body。这里使用 JSON 格式传递了一个参数为"database"，值为"db-test"的数据。

产品优势

最近更新时间：2024-09-09 15:16:31

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 或者 SDK 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。

功能全面

功能	描述
动态数据管理	您可以随时对数据进行插入、删除、搜索、更新等操作而无需受到静态数据带来的困扰。
近实时搜索	在插入或更新数据之后，您可以几乎立刻对插入或更新过的数据进行搜索。向量数据库负责保证搜索结果的准确率和数据一致性。

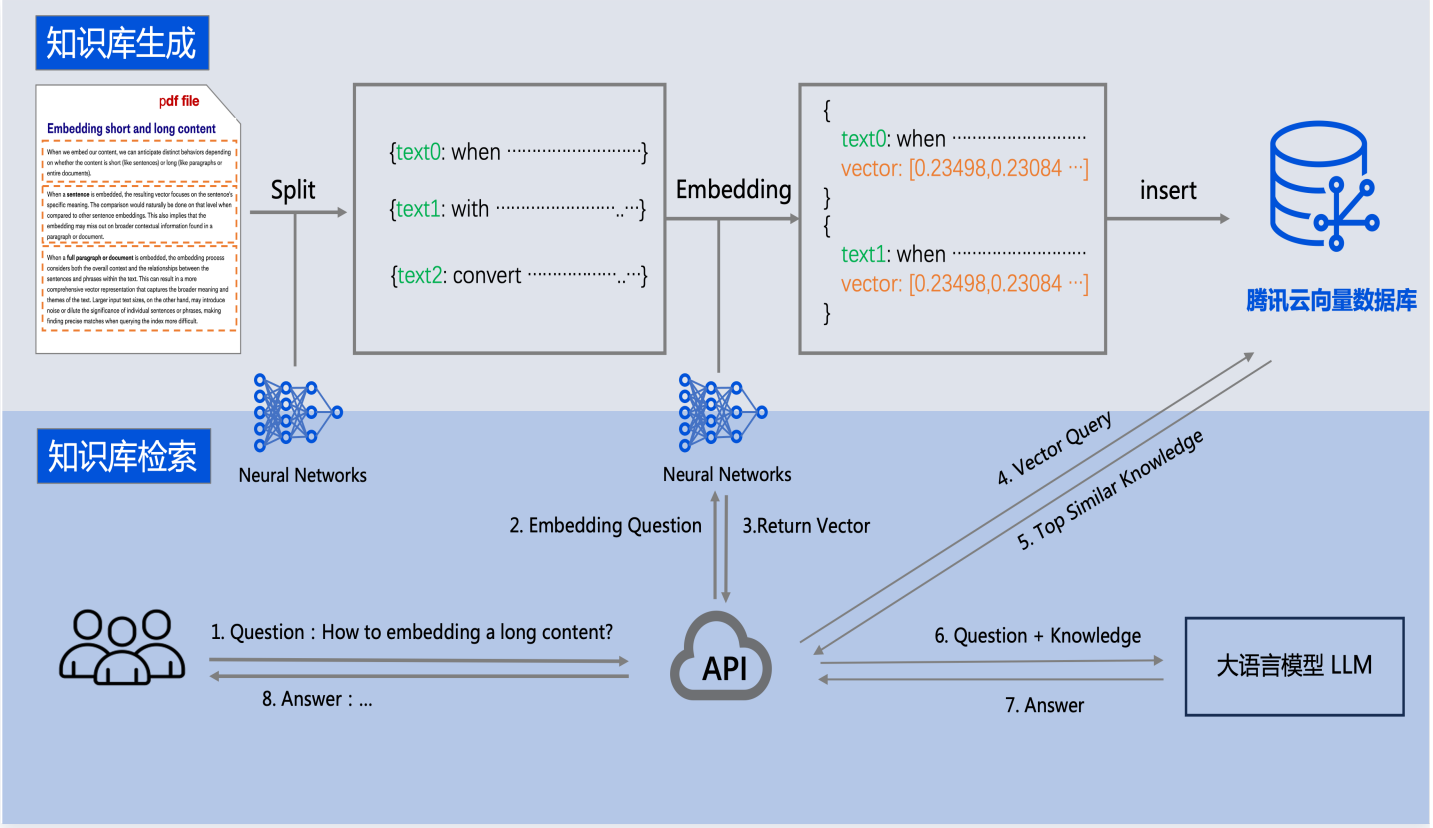
全面的相似度指标	支持各种常用的相似度计算指标，包括欧氏距离、内积、余弦相似度等。您可以根据应用需求来选择最有效的向量相似度计算方式。
原始文本自动向量化	向量数据库的 Embedding 功能会自动将原始文本进行转换，生成对应的向量数据并插入数据库或进行相似性检索，实现了文本到向量数据的一体化转换，减少了用户的操作步骤，极大降低了使用门槛。
可视化操作数据库	腾讯云向量数据库支持通过 DMC（Database Management Center，DMC）可视化管理数据库，在线执行数据插入、精确查询、相似度查询等操作，直观地查看执行结果。
一站式文档检索	用户只需上传 Markdown 格式的文档文件。腾讯云向量数据库将自动进行文本切分（Split）、信息补充、向量化(Embedding)和索引构建等一系列操作，完成知识库的建立。

应用场景

最近更新时间：2024-09-09 15:16:31

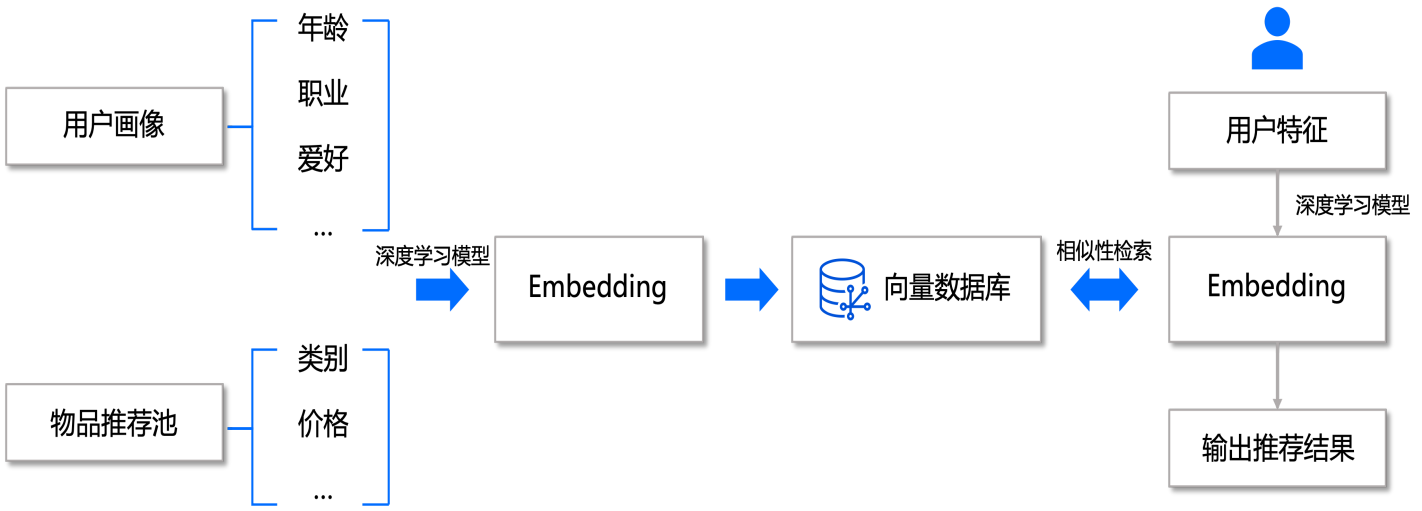
大模型知识库

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。



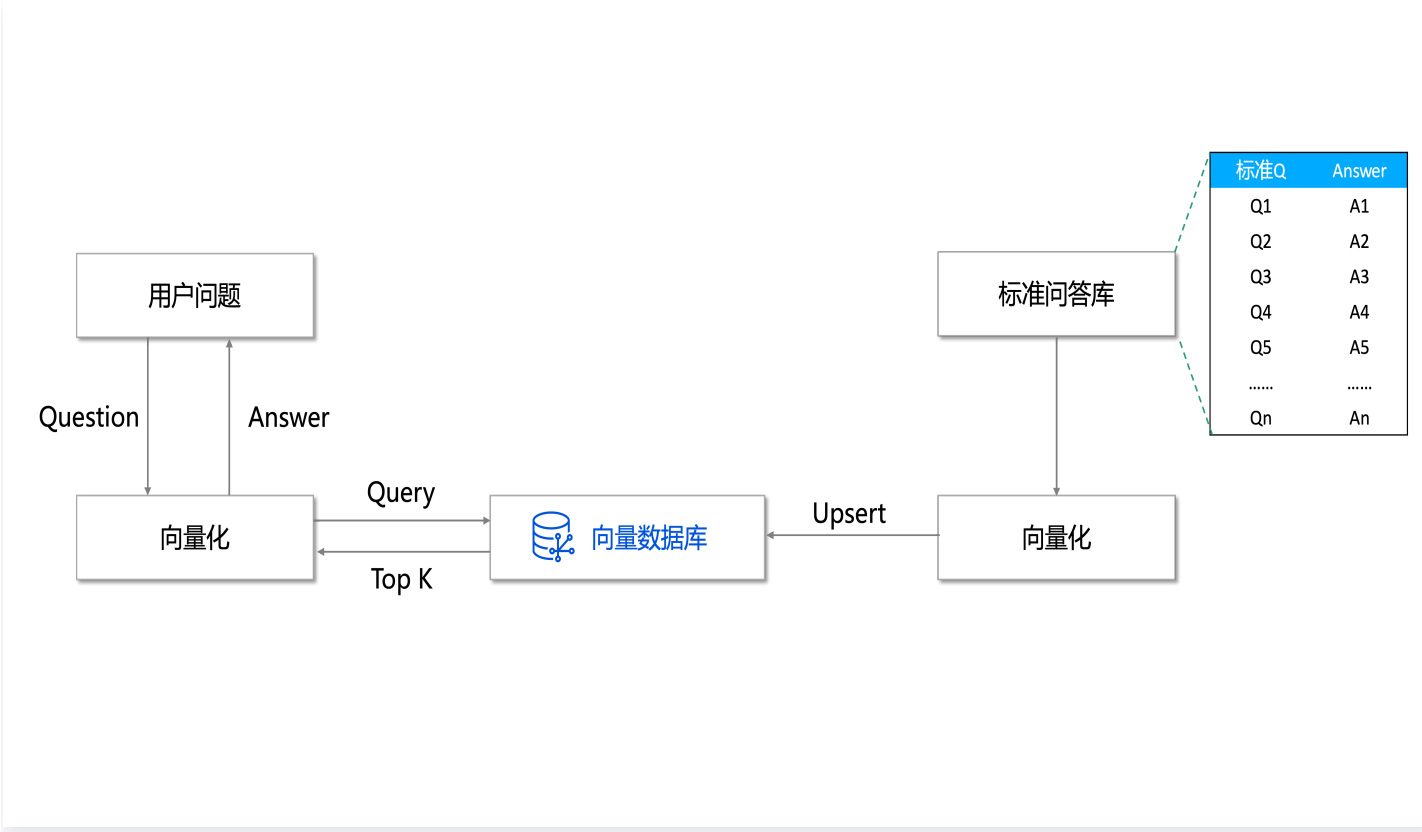
推荐系统

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，最终筛选用户可能感兴趣的物品推荐给用户。



问答系统

智能问答系统是一种能够回答用户提问的智能应用，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。

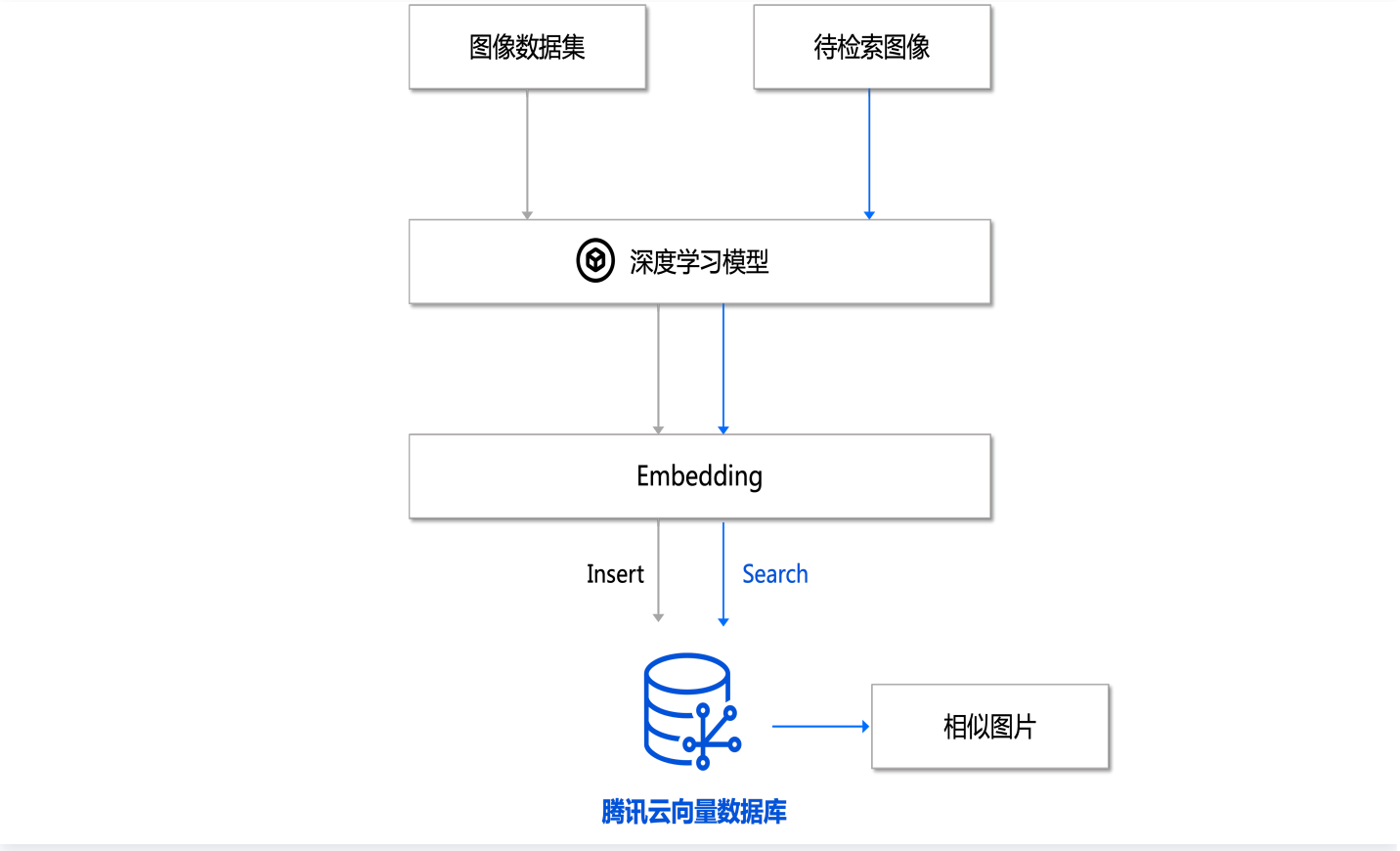


问答系统的应用场景非常广泛，例如智能客服、智能助手、智能家居等。在这些场景中，用户可以通过自然语言提问获取相关信息，例如查询产品信息、控制家居设备等。通过使用向量数据库来存储和检索相关的向量数据，问答系统

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。



名词解释

最近更新时间：2024-12-02 11:17:01

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。它支持单索引千亿级向量规模，适用于 AI 运算、检索场景，已稳定服务于近60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。

发布地域

最近更新时间：2025-04-24 11:24:52

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。

可用区

存储节点默认为三可用区部署，可根据不同的实例类型选择可用区的部署模式，不支持用户自定义具体可用区。

- 高可用版：支持选择三可用区与两可用区。
- 单机版：仅支持单可用区部署。

中国

地区	地域（region）		说明
华南地区	广州	ap-guangzhou	-
	深圳金融	ap-shenzhen-fsi	仅限金融机构和企业通过 在线咨询 申请开通
华东地区	上海	ap-shanghai	-
	上海自动驾驶云	ap-shanghai-adc	-
	南京	ap-nanjing	-
华北地区	北京	ap-beijing	-
西南地区	成都	ap-chengdu	-
港澳台地区	中国香港	ap-hongkong	中国香港节点可用于覆盖港澳台地区

其他国家和地区

地区	地域（region）		说明
亚太东南	新加坡	ap-singapore	新加坡节点可用于覆盖亚太东南地区
	雅加达	ap-jakarta	雅加达节点可用于覆盖亚太东南地区
美国西部	硅谷	na-siliconvalley	硅谷节点可用于覆盖美国西部
美国东部	弗吉尼亚	na-ashburn	弗吉尼亚节点可用于覆盖美国东部地区
欧洲地区	法兰克福	eu-frankfurt	法兰克福节点可用于覆盖欧洲地区

如何选择地域和可用区

购买云服务时建议选择最靠近您的地域，可降低访问时延、提高下载速。