Milvus的概述

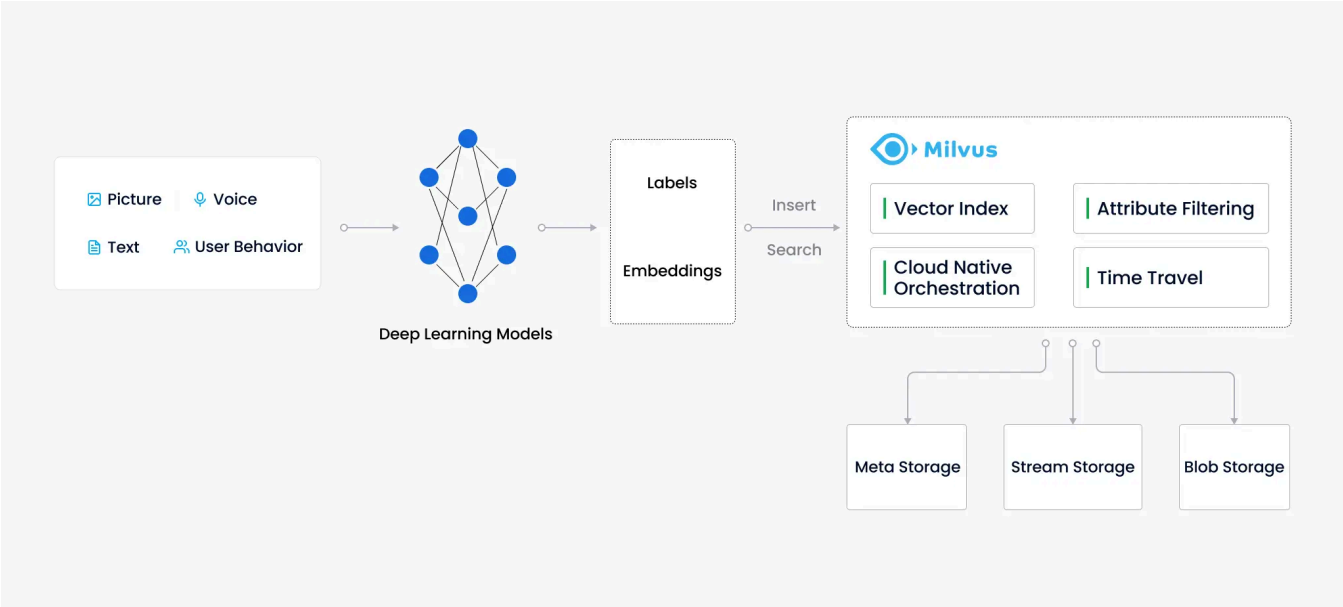
介绍

本页旨在通过回答几个问题，给你提供 Milvus 的概述。阅读本页后，你将了解到 Milvus 是什么，以及它是如何工作的，以及关键概念、为什么使用 Milvus、支持的索引和度量、示例应用程序、体系结构和相关工具。

什么是 Milvus 向量数据库?

Milvus 是在 2019 年创建的，其目标是存储、索引和管理深度神经网络和其他机器学习（ML）模型生成的大规模 嵌入向量。作为一种专门设计用于处理对输入向量的查询的数据库，它能够处理万亿级规模的向量索引。与现有的主要处理按照预定义模式遵循结构化数据的关系型数据库不同，Milvus 从底层开始设计，主要处理从 非结构化数据 转换而来的嵌入向量。

随着互联网的发展和演变，非结构化数据变得越来越常见，其中包括电子邮件、论文、物联网传感器数据、Facebook 照片、蛋白质结构等等。为了使计算机能够理解和处理非结构化数据，这些数据会使用嵌入技术转换为向量。Milvus 存储并索引这些向量。Milvus 能够通过计算它们的相似性距离分析两个向量之间的相关性。如果两个嵌入向量非常相似，那么意味着原始数据源也是相似的。



关键概念

如果你对向量数据库和相似性搜索的世界还不太了解，可以阅读以下关键概念的解释，以更好地理解。

了解更多有关 Milvus 的术语表。

非结构化数据

非结构化数据，包括图像、视频、音频和自然语言等，是指不遵循预定义模型或组织方式的信息。这种数据类型占据了全球数据的约 80%，可以使用各种人工智能（AI）和机器学习（ML）模型将其转换为向量。

嵌入向量

嵌入向量是非结构化数据的特征抽象，例如电子邮件、物联网传感器数据、Instagram 照片、蛋白质结构等。从数学上讲，嵌入向量是浮点数或二进制数的数组。现代嵌入技术用于将非结构化数据转换为嵌入向量。

向量相似性搜索

向量相似性搜索是将向量与数据库进行比较，以找到与查询向量最相似的向量的过程。使用近似最近邻搜索（ANNS）算法来加速搜索过程。如果两个嵌入向量非常相似，那么意味着原始数据源也是相似的。

为什么选择 Milvus?

- 在大规模数据集上进行向量搜索时具有高性能。
- 提供多语言支持和工具链的开发者优先社区。
- 即使在发生中断的情况下，也具有云扩展性和高可靠性。
- 通过将标量过滤与向量相似性搜索相结合，实现混合搜索。

支持的索引和度量有哪些?

索引是数据的组织单元。在可以搜索或查询插入的实体之前，必须声明索引类型和相似性度量。**如果未指定索引类型，Milvus 将默认使用暴力搜索。**

索引类型

Milvus 支持的大多数向量索引类型使用近似最近邻搜索（ANNS），包括：

- **FLAT**：FLAT 最适合于在小型百万级数据集上寻求完全准确和精确的搜索结果的场景。
- **IVF_FLAT**：IVF_FLAT 是基于量化的索引，最适合于在准确性和查询速度之间寻求理想平衡的场景。还有一个 GPU 版本 **GPU_IVF_FLAT**。
- **IVF_SQ8**：IVF_SQ8 是一种基于量化的索引，最适合于在磁盘、CPU 和 GPU 内存消耗非常有限的场景。
- **IVF_PQ**：IVF_PQ 是一种基于量化的索引，最适合于在牺牲准确性的情况下追求高查询速度的场景。还有一个 GPU 版本 **GPU_IVF_PQ**。
- **HNSW**：HNSW 是一种基于图的索引，最适合于对搜索效率有很高要求的场景。

有关更多详细信息，请参见 向量索引。

相似性度量

在 Milvus 中，相似性度量用于衡量向量之间的相似性。选择一个好的距离度量有助于显著提高分类和聚类性能。根据输入数据的形式，选择特定的相似性度量以实现最佳性能。

浮点嵌入中广泛使用的度量包括：

- **欧氏距离（L2）**：该度量通常在计算机视觉（CV）领域中使用。
- **内积（IP）**：该度量通常在自然语言处理（NLP）领域中使用。

二进制嵌入中广泛使用的度量包括：

- **汉明距离（Hamming）**：该度量通常在自然语言处理（NLP）领域中使用。
- **杰卡德相似系数（Jaccard）**：该度量通常在分子相似性搜索领域中使用。

有关更多信息，请参见 相似性度量。

示例应用程序

Milvus 使向你的应用程序添加相似性搜索变得很容易。Milvus 的示例应用包括：

- 图像相似性搜索：使图像可搜索，并立即从庞大的数据库中返回最相似的图像。

- 视频相似性搜索：通过将关键帧转换为向量，然后将结果输入到 Milvus 中，可以在几十亿个视频中进行搜索和推荐。
- 音频相似性搜索：快速查询大量音频数据，如语音、音乐、音效和表面相似声音。
- 推荐系统：根据用户行为和需求推荐信息或产品。
- 问答系统：自动回答用户问题的交互式数字 QA 聊天机器人。
- DNA 序列分类：通过比较相似的 DNA 序列，在毫秒级别准确地对基因进行分类。
- 文本搜索引擎：通过将关键字与文本数据库进行比较，帮助用户找到他们正在寻找的信息。

有关更多 Milvus 应用场景，请参见 Milvus 教程 和 Milvus 采用者。

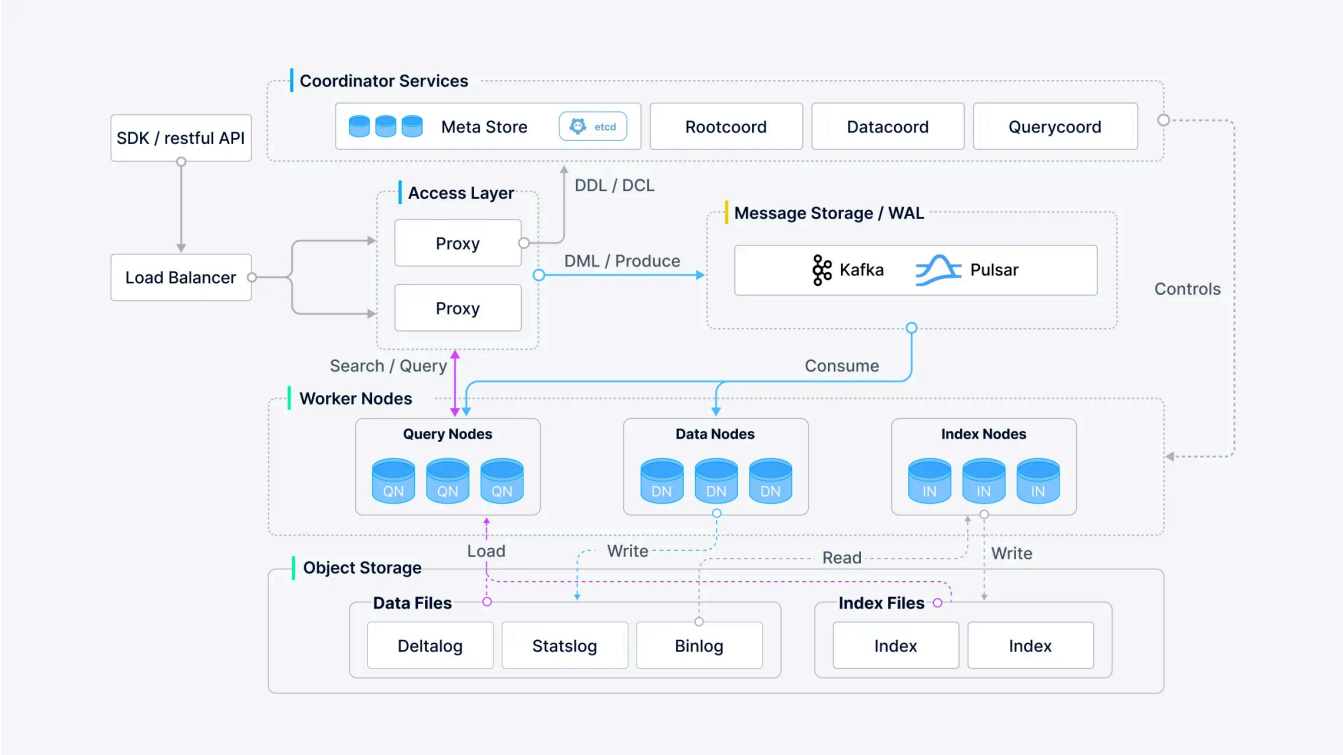
Milvus 的设计理念如何?

作为一个云原生的向量数据库，Milvus 在设计上将存储和计算分离。为了增强弹性和灵活性，Milvus 中的所有组件都是无状态的。

系统分为四个层次：

- 接入层：接入层由一组无状态代理组成，作为系统的前端层和用户的终端点。
- 协调服务：协调服务将任务分配给工作节点，并作为系统的大脑。
- 工作节点：工作节点充当手脚，是彻底的执行者，遵循协调服务的指示并执行用户触发的 DML/DDL 命令。
- 存储：存储是系统的骨架，负责数据持久性。它包括元数据存储、日志代理和对象存储。

更多信息，请参见 架构概览。



开发工具

Milvus 得到了丰富的 API 和工具的支持，以促进 DevOps。

API 接入

Milvus 拥有在 Milvus API 之上封装的客户端库，可以从应用程序代码中以编程方式插入、删除和查询数据：

- PyMilvus
- Node.js SDK
- Go SDK
- Java SDK

我们正在努力支持更多的新客户端库。如果你想贡献，请访问 Milvus 项目 的相应仓库。

Milvus 生态系统工具

Milvus 生态系统提供了有用的工具，包括：

- Milvus CLI

- Attu，Milvus 的图形管理系统。
- MilvusDM（Milvus 数据迁移），一个专门为 Milvus 设计的数据导入导出的开源工具。
- Milvus sizing 工具，用于估算不同索引类型所需的指定向量数量的原始文件大小、内存大小和稳定磁盘大小。

下一步

- 快速入门：3 分钟教程：
  - Hello Milvus
- 为你的测试或生产环境安装 Milvus：
  - 安装前提条件
  - 安装独立的 Milvus
- 如果你对 Milvus 的设计细节感兴趣：
  - 阅读有关 Milvus 架构概述

Last updated on July 4, 2024