西南证券
SOUTHWEST SECURITIES

人工智能专题研究

向量数据库——AI时代的技术基座

西南证券研究发展中心
通信研究团队
2023年6月

核心要点

- **受大模型热潮催化，向量数据库方兴未艾。**NVIDIA CEO 黄仁勋在3月的NVIDIA GTC Keynote 中，首次提及向量数据库，并强调其在构建专有大型语言模型的组织中的重要性。大模型作为新一代的 AI 处理器，提供了数据处理能力；而向量数据库提供了存储能力，成为大模型时代的重要基座。向量数据库是一种专门用于存储和查询向量数据的数据库系统，与传统数据库相比，向量数据库使用向量化计算，能够高速地处理大规模的复杂数据；并可以处理高维数据，例如图像、音频和视频等，解决传统关系型数据库中的痛点；同时，向量数据库支持复杂的查询操作，也可以轻松地扩展到多个节点，以处理更大规模的数据。
- **百亿蓝海市场蓄势待发，向量数据库空间广阔。**据 Statista 数据，2021 年全球数据库市场规模为 800 亿美元，同比增长约20.3%。假设增速保持20%，预计到2025年，全球数据库市场规模将达到1658.9 亿美元。据中国信通院测算，2020年中国数据库市场规模约241亿元；预计到2025年，中国数据库市场规模将达688亿元，复合增长率为23.4%。随着AI应用场景加速落地，我们预计2025年向量数据库渗透率约为30%，则全球向量数据库市场规模约为99.5亿美元，中国向量数据库市场规模约为82.56亿元。
- **海外需求逐步爆发，新兴赛道群雄并起。**目前向量数据库的赛道仍处于发展初期，随着大模型日趋成熟，越来越多玩家瞄准向量数据库的机会并选择加入赛道，呈现百花齐放的竞争格局。向量数据库的头部企业包括Zilliz、Pinecone等，目前的主要的客户还是互联网厂商随着大模型应用的不断拓宽，预计向量数据库的公司将受到更多投资者青睐，迎来投资井喷期。 Zilliz目前已与Nvidia、IBM、Mircosoft等公司展开合作，在一级市场获得1.13亿美元投资；Pinecone先后上架Google云和AWS，逐步打开市场，在一级市场获得1.38亿美元投资。
- **风险提示**：AI技术更新迭代缓慢、专业领域落地效果不及预期、市场开拓不及预期等风险。

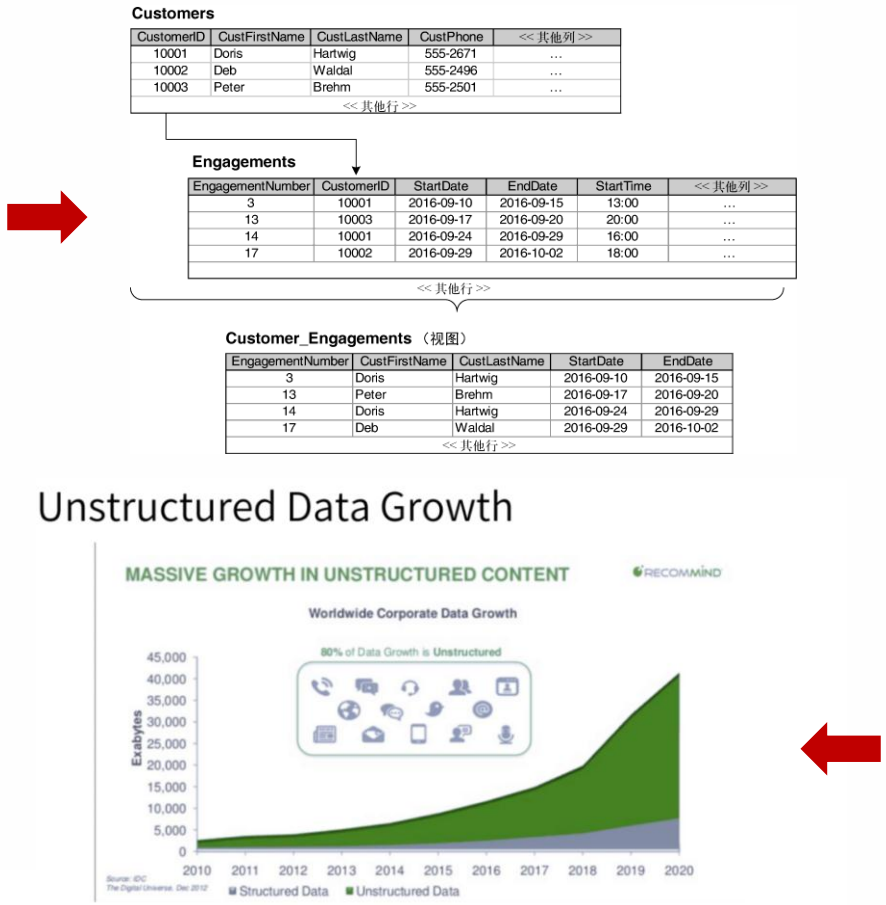
目 录

1.1 数据库分类

关系型数据库（SQL）vs. 非关系型数据库（NoSQL）

关系型数据库（SQL）

- **定义**：**依据“一对一、一对多、多对多”的关系模型创建数据库**，并将数据**以二维表格的形式储存**，各个表之间建立关系，通过这些关联的表格间分类、合并、连接或选取等运算来实现数据的管理。
- **发展情况**：1960s开始在航空领域发挥作用；因为其良好的一致性以及通用的关系型数据模型接口，使用范围广泛。
- **常见类型**：MySQL、Oracle、PostgreSQL等。
- **优点**：数据安全（磁盘）、数据一致性、二维表结构直观，易理解、使用SQL语句操作非常方便，可用于比较复杂的查询
- **缺点**：读写性能较差、不擅长处理较复杂的关系



图：关系型数据库和非关系型数据库规模对比情况

非关系型数据库（NoSQL）

- **起源**：2000年左右，互联网应用兴起，需要支持大规模的并发用户，并保持永远在线。一方面，**关系型数据库无法支持如此大规模数据和访问量**，升级CPU、内存和硬盘可以提高性能，但呈现明显的收益递减效应。另一方面，**数据库在机器间的迁移非常复杂**，需要较长的停机时间。**NoSQL因此应运而生，有效补充了SQL的适用范围，** NoSQL在Web应用领域提供了高可用性和可扩展性。
- **特点**：没有固定的表结构、**数据之间不存在表与表之间的关系、数据之间可以是独立的、NoSQL可用于分布式系统上。**
- **类型**：数据类型多样，针对不同的数据类型，出现了不同的 NoSQL，如**向量数据库**。

非关系型数据库是关系型数据库的有效补充

信息来源：《数据库分类及简介》，《数据库有哪些类型？如何根据应用场景选择？终于有人讲明白了》，IDC The Digital Universe，西南证券整理

1.1.1 数据库的分类——非关系型数据库

非关系型数据库按存储方式分为向量数据库、图形数据库、文档存储数据库、宽列数据库、键值存储数据库等，能够**实现非结构化或半结构化数据的处理和存储。**

	向量数据库	图形数据库	文档存储数据库
特点	将数据以向量形式存储，可实现向量数据的相似度搜索、聚类、降维等操作。	将数据以图的形式存储，以点、边为基础存储单元，每个节点代表一个实体，每条边代表两个实体之间的关系。	将数据以文档的形式存储，每个文档包含成对的字段和值。
优势	易处理高维度、高相似度、高并发的数据； 易与机器学习模型结合并提供智能化的服务。	易体现复杂的实体关系；支持高效的图遍历和分析。	非常灵活，可在文档中修改数据结构； 适用于处理半结构化或多变化的数据； 具有较高的性能，可快速传输、处理海量数据。
不足	技术成熟度较低，产品和相关应用较少	不适用于处理关系简单或无关系的数据；复杂性高，支持的数据规模有限。	缺乏严格的数据约束，需要小心谨慎地管理数据，避免数据出现质量问题。 通常不支持多文档操作，难以处理关联数据。

信息来源：《文档数据库介绍》，《揭秘文档数据库的优势与劣势》，《NoSQL 是什么？NoSQL 数据库详解》，西南证券整理

1.1.2 向量数据库的概述和原理

向量数据是什么？

- **“向量数据”**：向量数据是由多个数值组成的序列，可以表示一个数据量的大小和方向。通过Embedding技术，图像、声音、文本都可以被表达为一个高维的向量，比如一张图片可以转换为一个由像素值构成的向量。

向量数据库是什么？

- 向量数据库是一种专门用于**存储和查询向量数据**的数据库系统。
- 向量数据库支持对向量数据进行各种操作，例如：

向量检索：根据给定的向量，找出数据库中与之最相似的向量，例如在图像向量数据库中，用户输入一张图片进行搜索时，先将这张图片转换为一个向量，通过向量之间的近似检索，找到与输入图片最相似的图片。
向量聚类：根据给定的相似度度量，将数据库中的向量分类，例如根据图片的内容或风格，将图片分成不同的主题。
向量降维：根据给定的目标维度，将数据库中的高维向量转换成低维向量，以便于可视化或压缩存储。
向量计算：根据给定的算法或模型，对数据库中的向量进行计算或分析，例如根据神经网络模型，对图片进行分类或标注。

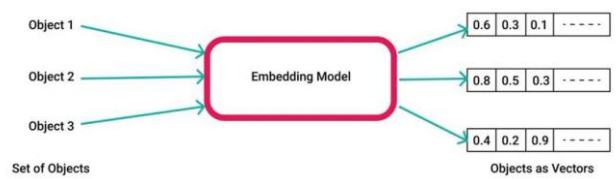
向量数据库有什么特点？

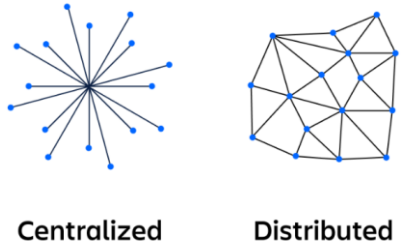
- **高维**：向量数据通常有很多元素，维度很高
- **稀疏**：向量数据中很多元素的值可能为零或接近零。
- **异构**：向量数据中的元素可能有不同的类型或含义。
- **动态**：向量数据可能随着时间或环境变化而变化。

信息来源：《和Zilliz聊聊开源商业化、向量数据库和软件全球化》，《向量数据库(Vector Database)简介》，西南证券整理

1.1.2 向量数据库的概述和原理

向量数据库的部分核心技术





Embedding 技术：
针对问题：文本、图像、音频等非结构数据存储问题。
解决方法：利用Embedding技术将高维度的数据（例如文字、图片、音频）映射到低维度空间，即把图片、声音和文字转化为向量来表示，将这些向量存储起来就构成向量数据库。实现**Embedding过程的方法包括神经网络、LSH（局部敏感哈希算法）等。**

向量索引技术：
针对问题：向量数据维度很高，直接进行全量扫描或者基于树结构的索引会导致效率低下或者内存爆炸。
解决方法：采用近似搜索算法来加速向量的检索，通常利用向量之间的距离或者相似度来检索出与查询向量相近的K个向量，距离度量包括欧式距离、余弦、内积、海明距离，向量索引技术包括 k-d tree（k-dimensional tree），PQ（乘积量化），HNSW（可导航小世界网络）等。

分布式系统架构：
针对问题：向量数据规模庞大，单机无法满足存储、计算需求。
解决方法：使用分布式系统。分布式系统是计算机程序的集合，这些程序**利用多个节点的计算资源来实现共同的目标，**节点通常代表独立的物理硬件设备，但也可代表单独的软件进程或其他递归封装的系统。

硬件加速技术：
针对问题：向量数据计算密集，单纯依靠CPU的计算能力难以满足实时性和并发性的要求。
解决方法：利用专用硬件来加速向量运算，这些硬件包括GPU，FPGA，AI芯片等，用于提供更高的浮点运算能力和并行处理能力。

信息来源：《什么是分布式系统？分布式系统和微服务架构的概述》，《向量数据库介绍》，《语义索引（向量检索）的几类经典方法》，西南证券整理

1.1.2 向量数据库的概述和原理

Embedding的步骤

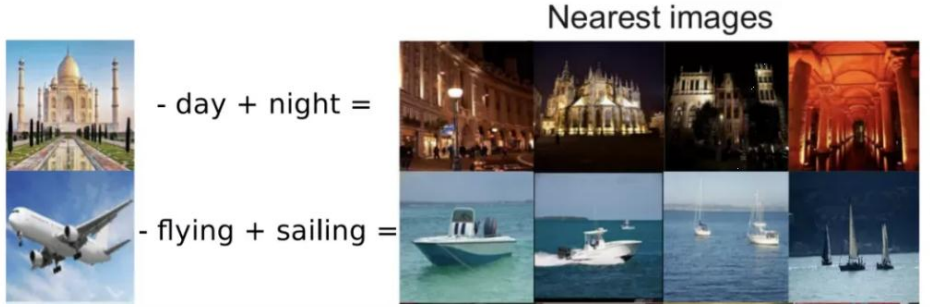
① **特征提取：将图片/音频转换成能够反映其内容或者属性的特征。**可以使用SIFT（Scale-invariant feature transform尺度不变特征转换），SURF（Speeded Up Robust Features加速稳健特征），HOG（Histogram of Oriented Gradients方向梯度直方图特征）等算法提取图片的边缘、角点、纹理等特征；可以使用MFCC（Mel频率倒谱系数）,LPC（线性预测分析）,PLP（感知线性预测）等方法提取音频的频谱、倒谱、能量等特征。

② **特征编码：将提取得到的特征进行编码，用一个固定长度的向量来表示。**方法包括BOW（词袋模型），VLAD (Aggregating local descriptors), Fisher Vector，GMM（高斯混合模型），HMM（隐马尔可夫模型），DNN（深度神经网路）等。

③ **特征压缩：将编码后的向量投影到低维度的子空间，进行向量压缩，使其能够用一个更低维度的向量来近似表示**，并保留尽可能多的信息。方法包括PCA（主成分分析算法），LDA（线性判别分析法），LSH（局部敏感哈希算法）。

Embedding的功能

- **语义搜索：**embedding向量是根据单词在上下文中的出现模式进行学习的，如果两个单词经常在上下文中一起出现，那么这两个单词**映射得到的向量在向量空间中就会有相似的位置。**用关键词进行语义搜索时，模型将关键词转化为embedding向量，然后在高维的向量空间里搜索这个embedding向量以及与其较为接近的向量，就可以得到与关键词相似的结果。
- **向量运算：**通过对embedding向量执行向量**加法和减法操作，可以推断出单词之间的语义关系。**例如，women的embedding向量可以通过下列运算得出：Embedding(woman)=Embedding(man)+[Embedding(queen)-Embedding(king)]
- **共享和迁移：**embedding向量可以**在多个自然语言处理任务中进行共享和迁移。**例如，在训练一个情感分析模型时，可以使用在句子分类任务中训练的嵌入向量，这些向量已经学习到了单词的语义和上下文信息，从而可以提高模型的准确性和泛化能力。



信息来源：《AI大模型领域的热门技术——Embedding入门介绍以及为什么Embedding在大语言模型中很重要》，《和Alex聊聊向量数据库与职业规划》，《深入解读向量数据库，Zilliz「三剑客」干货分享》，西南证券整理

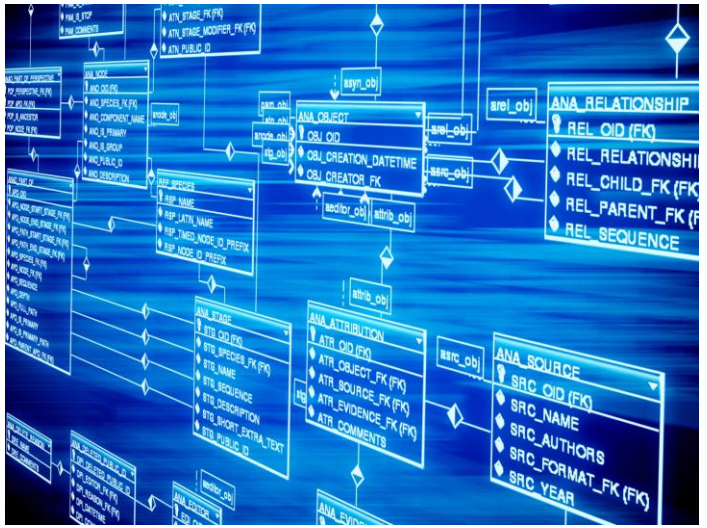
1.1.3 向量数据库的优势和不足

优点

- **处理大规模数据：**向量数据库的基本数据类型是向量，使用向量化计算能够比关系型数据库更快地处理大规模的复杂数据。
- **支持高维数据：**向量数据库可以处理关系型数据库中很难处理的高维数据，例如图像、音频和视频等。
- **支持复杂查询：**向量数据库支持复杂的查询操作，例如相似性搜索、聚类分析、降维等，并且速度快、准确度高，而关系型数据库中很难实现复杂操作。
- **易于扩展：**向量数据库可以利用分布式、云计算、边缘计算等技术轻松地扩展到多个节点，从而扩大数据处理规模，并提高向量数据的存储、管理和查询的稳定性。
- **高兼容性：**向量数据库支持多种类型和格式的向量数据，支持多种语言和平台的接口和工具。

不足

- **相对较新：**向量数据库是一种相对较新的技术，目前市场上的产品和应用还比较少。
- **学习成本高：**向量数据库需要掌握向量化计算的相关知识，学习成本较高。
- **适用场景局限：**向量数据库适用于处理大规模的复杂数据，而关系型数据库适用于处理简单数据。



信息来源：《一图文秒懂向量数据库、数据库发展过程》，《向量数据库(Vector Database)简介》，西南证券整理

1.1.4 向量数据库和传统数据库的区别

向量数据库与传统关系型数据库协同发展、相互补充。针对传统关系型数据库难以处理的大规模数据、低时延高并发检索、模糊匹配等领域，向量数据库通过数据的向量化来满足特定需求，尤其适用于人工智能领域。

	传统的关系型数据库	向量数据库
数据类型	数值、字符串、时间等传统数据类型	新的数据类型：向量数据 不存储原始数据
数据规模	小，1亿条数据对于关系型数据库来说规模很大	大，最少千亿数据是底线
数据组织方式	基于表格，按照行和列组织	基于向量，按照向量维度组织
查找方式	精确查找：点查/范围查 查询结果要么符合条件要么不符合条件	近似查找 查询结果是与输入条件最相似的，近似比较对计算能力要求非常高。
低时延，高并发	否	是
上层应用	较弱	对外提供统一的API，更适合大规模AI引用程序的部署和使用
下游应用场景	央企、国企。央企国企因工作内容要求容错率低，传统数据库能够提供更为准确的搜索结果。	互联网公司。向量数据库的结果正确率相对较低，成本低，互联网公司的场景容错率较高，能够包容这一缺陷。

信息来源：《向量数据库(Vector Database)简介》，《和Alex聊聊向量数据库与职业规划》，《向量数据库介绍》，西南证券整理