向量数据库？不要投资！不要投资！不要投资！

作者：吴英骏 策划：Tina，赵钰莹

2023-06-04 北京

本文字数：3400 字 阅读完需约11分钟

> 我对生成式 AI 大模型的未来充满了希望，同样，我对向量数据库行业也非常看好。只不过如果有人想新入局向量数据库赛道，我只能表示劝退。与其投资新的向量数据库项目，还不如关注现有数据库中哪些加上向量引擎可以变得更加强大。

推特上关于向量数据库的调侃

由于疫情、通货膨胀、美联储加息、国际局势等诸多因素，尤其科技领域的风险投资市场其活跃度在 2022 年降至冰点。相信很多投资人都是抱着"躺平"的观点度过了过去的一年。庆幸的是，ChatGPT 的诞生点燃了全世界对科技领域的热情，投资活动如雨后春笋般蓬勃兴起，重新焕发了活力。很显然，生成式 AI 大模型的底层系统以及基于生成式 AI 大模型的应用都是投资热点。除了 OpenAI 获得微软 100 亿美金投资之外，AI 创业公司例如 Hugging Face、Jasper、 Stability AI、Midjourney、MiniMax 等，都在资本市场上颇受追喷，公司的估值也是水涨船高。

推荐阅读

连代码都没写就敢要融资：被 C
火的向量数据库，带来了一大波
AI，数据库，大数据，云原生，架

20 | Embedding：深入挖掘用
2023-05-31

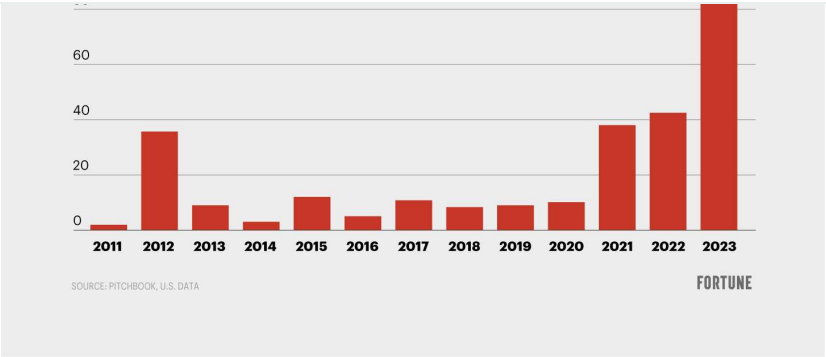
为什么 AIGC 和大模型创业者
量数据库?
2023-06-01

GGV 对话 Zilliz 星爵：向量数
AI 原生数据基础软件时代
2023-05-19

09 | 语义检索，利用 Embedd
的搜索功能
2023-04-03

33 | AI 前沿：ChatGPT 资料精
2023-05-01

数据库内核杂谈（三十三）-
（1）
数据库，运维，数据处理



生成式 AI 大模型初创公司的投前估值已经接近 1 亿美金。图片来源：https://fortune.com/2023/04/06/how-much-are-generative-ai-startups-worth-venture-capital/。

作为数据基础设施领域的创业者，我一直专注在数据库与实时流计算赛道，似乎这次的 AI 大爆发应该与我无缘。然而有意思的是，向量数据库，这一数据库领域中的细分赛道，却在短期内成为了万众瞩目的焦点，让本该相对沉寂的数据库市场再次热闹了起来。最近有不少投资人联系我，询问我对向量数据库的看法。毕竟，对于过去一整年出手甚少的投资者来说，数据库系统这一技术壁垒较高的领域出现了一个热点，自然不应该错过这个良机。然而，我的回答却是十分干脆："不要投资"。更准确的说，如果你已经投资了一些向量数据库，那么恭喜你，可以期待在这个新的时代一飞冲天；如果你在之前没有入场向量数据库的话，那现在入场可能并非明智的选择。为什么呢？我们可以从技术、应用、与市场三个方面来探讨。

向量数据库的技术

在传统的关系型数据库中，数据通常以表格为形式来存储。然而，随着 AI 时代的到来，我们面临着图像、音频和文本等海量的非结构化数据。这些数据无法简单地以表格形式存储，而是需要通过机器学习算法从这些数据中提取出以向量为表示形式的"特征"。向量数据库的兴起便是为了解决对这些向量进行存储与计算的问题。

向量数据库的核心在于对数据的索引。使用倒排索引等技术，向量数据库可以通过将向量的特征进行分组和索引，以实现高效的相似性搜索。同时，向量量化技术可以帮助向量数据库将高维向量映射到低维空间，从而减少存储和计算成本。基于索引技术，向量数据库通过自身的各类向量操作，如向量相加、相似度计算和聚类分析等，使得用户能够对向量进行高效搜索。

至于向量数据库的底层存储，实际上相比于索引技术来说，显得不那么重要。事实上，很多数据库都可以直接添加索引模块来实现高效向量搜索。而现有数据库，尤其是基于列式存储的实时分析数据库，本身便具有卓越的数据压缩率。对于向量数据而言，由于每个向量都是由大量的维度组成，通过列存储可以将相同维度的数据连续存储，从而提高存储效率和查询性能。此外，列存数据库还能够针对列级别的操作进行优化，如向量相似性计算和聚合操作。这也是为什么网络上纷纷流传新晋向量数据库 Chroma"仅仅"是在著名实时分析数据库 ClickHouse 上封装了一层而已。当然，Chroma 的联合创始人也出来澄清，表示他们会很快去除对 ClickHouse 的依赖。

Chroma 联合创始人 Jeff Huber 澄清说，“本周末 Chroma 便将不再使用 ClickHouse，并会转变成一个云原生数据库。”

不论 Chroma 的未来如何，我们都不得不承认，想要使现有数据库支持向量搜索功能并非很难实现，而大量现有数据库很有可能在不久的未来便会推出自己的向量搜索功能。

向量数据库与生成式 AI 大模型

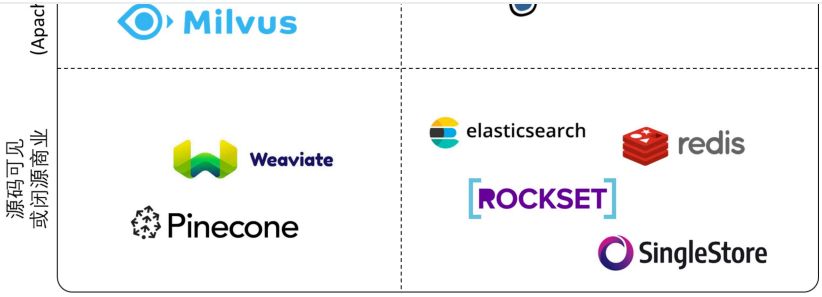
我们再来说说为什么向量数据库在最近火了起来。向量数据库并非在这两年兴起的新兴物种，而现有的向量数据库公司例如 Zilliz（2017 年）、Pinecone（2019 年）、Weaviate（2019 年）等都已经有了 4-6 年的历史。

那为什么最近的生成式 AI 大模型能促进向量数据库的火爆？这有几方面原因。其一，生成式 AI 大模型需要大量的数据进行训练，以获取丰富的语义和上下文信息。这导致了数据量的爆发式增长。向量数据库作为数据的管理者，能够高效的帮助管理数据。其二，生成式 AI 大模型生成的文本往往需要进行相似性搜索和匹配，以提供准确的回复、推荐或匹配结果。传统基于关键词的搜索方法可能无法满足复杂的语义和上下文要求，而这也使得向量数据库有了用武之地。其三，生成式 AI 大模型不仅限于处理文本数据，还可以处理图像、语音等多模态数据。向量数据库作为一种能够存储和处理多种数据类型的系统，能够有效地支持多模态数据的存储、索引和查询。

以上几点原因都能推导出一个观点，便是向量数据库的发展与生成式 AI 大模型高度绑定。只要生成式 AI 大模型在未来的几年内继续高速发展，向量数据库也一定能够获得足够多的需求。

向量数据库的市场需求与格局

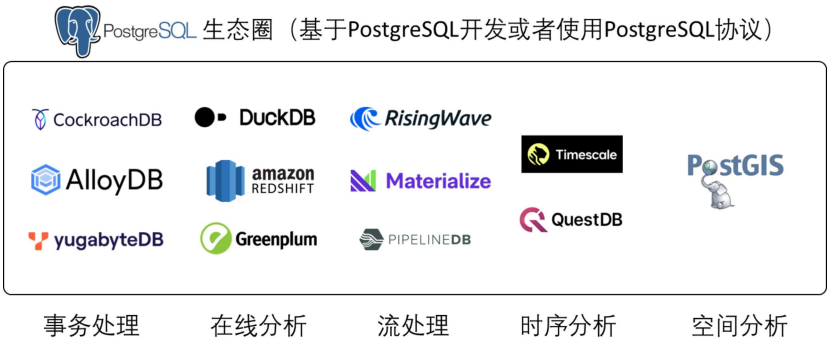
在谈了向量数据库的技术与应用之后，我们来谈谈市场。任何投资行为都是要追求收益。想要预估收益，必定需要评估现有市场需求与供给情况，再来判断投资是否能够获得有吸引力的回报。为什么我不推荐现在入场投资向量数据库呢？这是因为向量数据库已经拥有了足够多的产品，而向量数据库的用户几乎总是能够在现有的市场中找到合适的产品，这使得新入场的玩家变得机会渺茫。



市场上主流的特化向量数据库与支持向量检索的数据库。

当一家公司拥有强大的技术基础和需要先进的向量搜索功能的大量工作负载时，他们真正需要的是一款特化的向量数据库。在这个领域中，领先的选择包括 Chroma（2000 万美金融资）、Milvus（1.13 亿美金融资）、Pinecone（1.38 亿美金融资）、Qdrant（980 万美金融资）、Weaviate（6770 万美金融资）等等。这些玩家在最近的几年内都收获了大量的融资，有望占据重要的市场份额。这些向量数据库提供了高效的向量存储、索引和相似性搜索功能。它们通常具有针对向量数据的特定优化，如基于倒排索引的相似性搜索和高效的向量计算。这使得它们能够满足公司在推荐系统、图像搜索和自然语言处理等领域的需求。

而如果一家公司已经购买了 Elastic、Redis、SingleStore 或 Rockset 等商业数据库，并且不需要特别先进的向量搜索功能，他们可以充分利用这些数据库现有的功能。这些商业数据库在非向量数据处理方面表现出色，适用于各种用例和场景，而在向量数据处理方面只要能做到及格，便能够满足一般用户的需求。此外，数据库技术正在不断发展，许多数据库正在考虑引入向量搜索功能以满足自身现有用户需求。对于目前缺乏向量搜索功能的数据库，它们实现这些功能只是时间问题。



基于 PostgreSQL 开发或者使用 PostgreSQL 协议的数据库已经覆盖了各个细分领域。

事实上，即使没有这些商业数据库，用户可以很轻易的安装 PostgreSQL，并使用 PostgreSQL 内置的 pgvector 功能进行向量搜索。PostgreSQL 可以被认为是开源数据库领域的黄金标准，在数据库的各个赛道，包括事务处理、在线分析、流处理、时序分析、空间分析等方面，都有着相当完整的支持。对于那些仅仅想尝试使用向量数据库的非专业用户来讲，它们完全可以自己下载开源的 PostgreSQL，或者使用例如 Supabase 和 Neon 这样的托管服务，便能够搭建出自己的简易 AI 应用。

加强人。向量数据库，不要投资！不要投资！不要投资！

作者简介：

吴英骏，流数据库公司 RisingWave（risingwave.dev） 创始人 &CEO。博士毕业于新加坡国立大学计算机系，为前 Amazon Redshift 工程师和前 IBM Research Almaden 研究员。常年担任数据库三大顶会 SIGMOD/VLDB/ICDE 的评审委员会成员。技术交流可以扫码关注如下公众号“RisingWave 中文开源社区”或者添加微信“risingwave_assistant”。

RisingWave 中文开源社区公众号

RisingWave 中文开源社区技术交流群

本文内容仅为提供更多信息以供参考或交流学习，不代表平台立场，如有不同意见，欢迎大家投稿！

相关阅读：

连代码都没写就敢要融资：被 ChatGPT 带火的向量数据库，带来了一大波造富神话

发布于：2023-06-04 20:13

阅读数：2902

语言 & 开发 开源 数据库 架构 数据处理 技术选型 Oracle MySQL

评论 2 条评论

写下你的想法，一起交流 发布

superman

AI试用体验，懒人专项：https://monica.im/?c=IRHLTZYV 免费使用：GPT3.5、GPT4.0、AI作画、文章摘要、视频摘要、智能搜索等

13 小时前 · 北京 0 回复

Geek_a4f58c

背仔角灯塔?

2023-06-05 10:36 · 广东 0 回复

没有更多了

更多内容推荐

2022 年数据库发展总结：中国和海外数据库差距还有多远?

中国数据库行业随着 2021 年 7 月 PingCAP 完成 3.4 亿美元融资，估值达到 30 亿美金，把中国数据库行业引爆了。

数据库，QCon，文化 & 方法，方法论，InfoQ，Oracle

LLM 快人一步的秘籍 —— Zilliz Cloud，热门功能详解来啦!

最近，我们发布了可处理十亿级向量数据的 Zilliz Cloud GA 版本，为用户提供开箱即用的向量数据库服务，大大降低了数据库的...

2023-04-11

2022 年 10 月《中国数据库行业分析报告》重磅发布！精彩抢先看

墨天轮10月《中国数据库行业分析报告》新鲜出炉！本月重点介绍了向量数据库与向量化执行引擎的核心能力与技术价值，并发布...

2022-10-21

NoSQL 案例：Doris 分析案例（一）

2022-09-10

如何用一款数据库解决企业的核心问题？ | 专访矩阵起源创始人兼研发负责人张颖峰

入行近20年，张颖峰先后负责搜索引擎内核、大数据系统、分布式高可用基础架构，一路见证了互联网科技的发展与变迁。

数据库，数字化转型，数据处理，Serverless

破解加密的 LastPass 数据库

最近，LastPass泄露了电子邮件地址、家庭住址、姓名和加密的用户数据库。在这篇文章中，我将演示攻击者如何利用Hashcat等...

2022-12-27

支持向量机 - 线性 SVM 决策过程的可视化

我们可以使用sklearn中的式子来为可视化我们的决策边界，支持向量，以及决策边界平行的两个超平面。

2022-11-23

SQL 碎碎念, 你可能用不到但不能不知道的数据库技巧 (2)

古语有云,牙疼不是病,疼起来真要命.平时可能看起来不是很重要的内容,等到真正用到时候才是心急如焚.本期讲解你可能不知道但是...

2022-11-05

开心档之 C++ STL 教程

在前面的章节中，我们已经学习了 C++ 模板的概念。C++ STL（标准模板库）是一套功能强大的 C++ 模板类，提供了通用的模...

2023-04-27

yield 语句

2022-09-08

亚信科技 AntDB 数据库荣膺第十二届数据技术嘉年华（DTC 2023）“最具潜力数据库”大奖

AntDB荣获墨天轮“2022年度最具潜力数据库”

2023-04-13

被 ChatGPT 点燃的向量数据库们

在 AIGC 革命大爆发的日子，一个特别的挑战是大规模存储和查询非结构化数据（比如图像、视频、文本）的能力。

2023-05-09

KaiwuDB CTO 魏可伟：1.0 时序数据库技术解读

KaiwuDB 1.0 作为国内数据库新生品牌力量，是浪潮集团控股的数据库企业，我们聚焦在工业物联网、数字能源、交通车联网、智...

2023-01-19

数据库的基本原理

2022-09-10

Gartner 数据库魔力象限解读 | DBTalk 技术公开课第 6 期

随着数字化进程的加快，数据越来越成为数字时代的基础性战略资源和革命性关键要素。作为数据存储与计算的基础软件，数据库...

数据库，腾讯云，最佳实践

「非结构化数据峰会」精彩速递 | Zilliz Cloud 首发、Milvus 技术演进、生态实践全揭秘

2022 年 9 月 24-25 日，首届非结构化数据峰会（2022 Unstructured Data Summit）在线上举行。围绕人工智能在非结构化搜...

语言 & 开发，数据库，数据处理，企业动态