首页推荐

黄东旭："向量数据库"还是"向量搜索插件 + SQL 数据库"？| 我对 2024 年数据库发展趋势的思考

PingCAP 2024-02-15 246

本文由 PingCAP 黄东旭撰写，讨论了数据库技术在 2023 年的快速变革，并对 2024 年的数据库发展趋势进行了预测。文章重点关注了 GenAI 时代对数据库的影响，提出了在数据库选择上的两种路径："向量数据库"和"向量搜索插件 + SQL 数据库"。文章强调了个性化数据服务的重要性，以及数据库在实时交互和弹性方面所起到的关键作用。

如果我们用一个词来总结 2023 年的数据技术领域，那个词无疑是"急速变革"。我们见证了数据库内核技术与云原生架构的融合演进，AI+Data 的浪潮涌现，以及用户工作负载的深刻转变。GenAI 时代的到来，就像一股不可抗拒的潮流，推动着数据技术的每一朵浪花，朝着更智能化、更灵活化的巨浪之海奔流。

2023 年，我们的眼前充满了夺目的 AI Demo 与炫技，你追我赶。转眼间，当我 们步入 2024 年，这个年份将因为 "AI 在从 Demo 到真实场景落地"的急剧转变而被人们记住。随着开源大模型成本的加速下降，企业和开发者对数据的关注也急剧上升，对数据的关注度将很快取代对模型的关注度。有预测认为，在 2023 年，用户愿意在 AI 模型上投入 80% 的预算，然而在未来这一两年里，随着模型成本的降低，这一比重可能会逆转，用户将更多的投资（甚至大于 80%）倾向于数据，数据处理和分析能力变得更加重要。

毫无疑问，AI 将会对数据处理提出非常多新的诉求，数据技术领域也会面临着多重挑战与机遇，AI 正在重塑数据技术的全新生态。我们不禁要问：在 GenAI 的大潮中，选择 "向量数据库"还是"以 SQL 数据库作为核心，添加向量搜索插件"？数据库如何应对 Gen AI 对数据库扩展性和实时交互的诉求？浪涌般海量数据的实时查询会不会带来巨大的成本压力？AI 带来的自然交互方式催生怎样的开发者体验 ？这些问题将在本文中一一解答

/ 预测一 /

"向量数据库"还是"向量搜索插件 + SQL 数据库"？这是一个答案很明确的问题。

如果说过去 CRUD 应用是对数据库访问的静态封装，那么随着 GenAI 的普及，尤其是 Chatbot 或 Agent 的产品形态，对数据的使用会是更加灵活和动态的。过去，集中的数据存储和应用是因为技术的局限，很难为个人提供个性化的服务，尽管现代的 SaaS 其实很希望往这个方向发展，但是为每个用户都提供个性化的体验对算力和开发的挑战太高，而 GenAI 和 LLM 将提供个性化服务的成本降得很低（可能就是几段 Prompt），以至于对于数据库而言，带来几个变化：

- 个人（或一个组织）产生的数据价值会变得越来越高，但这类数据通常不会很大
- GenAI 会使用更加动态和灵活的方式直接访问数据，这样效率最高
- 对数据的访问从边缘发起（从 Agent 或者 GenAI 直接发起）

一个很好的例子是 GPTs， GPTs 支持通过的自定义的 Prompt 和用户提供的 RESTful API 来创建自己的 ChatGPT，基础的 ChatGPT 会在它认为需要的时候以灵活的方式调用你给定的 Action。这个调用发生方式和参数是后端的 Action 提供者无法预料的。而且可以预料的是很快 GPTs 将会提供标记个人身份信息的机制，这样对于 Action 的提供者来说，相当于后端的数据库有了最重要的索引：UserID，剩下的就很好理解了。

这里你可能会提出质疑，RAG 不是标准的做法吗？但现有的 RAG 构建的方式几乎都是静态的，而知识应该是可以实时被更新的，这里不得不提到向量数据库。

对向量的支持，在去年是数据库迭代的一个热门方向，产生了很多专门的向量数据库， **但是我认为，更丰富的数据访问接口，使得向量搜索成为标配，然而 SQL 仍然是基石。向量搜索并不值得专门使用一个独立的数据库来支持，更应该是现有的数据库中的一个功能，就像** ：

```
Plaintext
Rust   INSERT INTO tbl (user_id, vec, ...) VALUES (xxx, [f32, f32,
f32 ...], ...);   SELECT * FROM tbl WHERE user_id = xxx and
vector_search([f32,f32,f32,f32 ...])
```

类似的访问可能是更符合开发者直觉的。

而 **关系型数据库天然支持插入和更新** ，另外配合向量索引的搜索能力，便可以将 RAG 变成一个可以实时更新实时查找的正反馈循环（利用 LLM 引入进行二次的 Summary ，然后将更新的 Index 储存在 DB 中）。更重要的是， **关系型数据库的引入消除了向量数据库带来的数据孤岛的问题** ，当你可以将向量索引筛出来的数据关联（JOIN）到同一个 DB 中其他的数据的时候，灵活性带来的价值就得以显现。

另一个好处是，Serverless 的产品形态，同样也将数据的所有权归还给用户本人，大家思考一下，在我们熟知的 Web2 时代，我们的数据是隐藏在一个个互联网公司的服务背后的黑箱，我们没有办法直接访问；而在 GenAI 的应用场景下，数据的交互变成一个三角的关系，用户 - 数据 (RAG) - GenAI。很有意思的是，这个正是 Web3 的理想之一，GenAI 的普及很可能顺手也将 Web3 想实现的将数据的所有权交还给用户的理想，这在 Web2 时代是不可能实现的，这其实是一种技术理想的回归。

当然，我相信在未来 RAG 会成为数据库的很重要的一种新应用场景，在这种场景中 Serverless 形态提供的云数据库服务会变成标准化的。

/ 预测二 /

由高价值数据驱动的应用成为 GenAI 应用的主流，弹性与实时交互成为数据库能力的基石。

在预测一里我们提到， GenAI 时代的应用要求知识和数据是可以被实时更新的，这对数据库的弹性以及实时交互提出了非常直接的需求。

数据库的可扩展性一直是过去十年间，业界关注的重点之一。根据我们的观察，大多数单一在线业务，100TB 已经是很大规模，而这个规模下的一般 OLTP 业务，已经可以被市场上很多系统自信的解决。

但这些数据库大多是 Shared Nothing 的系统，Shared nothing 的系统通常会有一个假设：在集群中的节点是对等的，只有这样数据和 Workload 才能均匀的分散在各个节点上。这个假设对于海量数据 + 访问模式均匀的场景没有问题，但是仍然 **有很多的业务具有明显的冷热特征， 尤其是在 GenAI 带来的数据访问方式越来越动态和灵活的 2024 年及以后** 。

我们最经常处理的数据库问题之一就是局部热点。如果数据访问倾斜是一个业务的天然属性的话，对等的假设就不再是合理的，更合理的方式是将更好的硬件资源倾斜给热点的数据，而冷数据库使用更廉价的存储，例如，TiDB 从一开始将存储节点（TiKV）/ 计算节点（TiDB）/ 元信息（PD）分离，以及在后来 TiDB 5.0 中引入自定义 Placement Rule 让用户能够尽可能决定数据摆放策略，就是为了尽可能弱化节点对等假设。

但是更终极的解决办法在云端，在基本的扩展性问题得到解决后，人们开始追求更高的资源利用效率，在这个阶段，对于 OLTP 业务来说，我想可能更好的评价标准是 Cost Per Request。因为在云端，计算和存储的成本差别是巨大的，对于冷数据来说，如果没有 Traffic，你甚至可以认为成本几乎为 0，但是计算却是昂贵的，而在线服务不可避免的需要计算（CPU 资源），所以 **高效利用计算资源，云提供弹性将成为关键** 。

另外，请不要误解 ，弹性并不意味着便宜，on-demand（ 随需提供的 ）的资源在云上通常比 provisioned（预分配）的资源更贵，持续的 burst 一定是不划算的，这种时候使用预留资源更合适，burst 那部分的成本是用户为不确定性支付的费用。仔细思考这个过程，这可能会是未来云上数据库的一种盈利模式，

与弹性同样重要的需求就是实时交互 。GenAI 时代的应用需要数据库不仅要有强大的数据处理能力，还需要有高效的实时数据广播和同步机制。这不只是让数据能够实时更新，而是确保数据流能够实时流动，让数据库能即时捕捉到每一次交互，每一个查询，确保每一个决策都是基于最新、最准确的信息。(就是用户愿意为更高价值的实时交互付钱，想想股票实时交易和直播电商的场景就知道了)

于是整个系统——从数据的产生到处理、再到存储和检索——都必须要在实时的框架下工作，能够在毫秒级别做出实时响应，这也需要数据库能实时在事务处理（OLTP）和分析处理（OLAP）之间无缝同步。这样的实时交互能力，将会是现代数据库区别于传统数据库的决定性因素之一。

/ 预测三 /

成本分析已经成为所有人关心的问题，在云数据库的可观测性中成为独立新视角。

今天我还想谈的一点是云数据库的可观测性，尤其是它是否能让我的云消费更透明。对于数据库云服务来说，可观测性的要求会更高，因为对于开发者来说，服务商提供的 Dashboard 几乎是唯一的诊断手段。介绍可观测性的文章也很多，相似的部分因为篇幅关系我也不打算说太多。

与传统的可观测性不一样的是： **在云上，一切 Workload 都会成为客户的帐单的一部分** 。对于用户来说一个新的问题便是：为什么我的帐单看起来是这样？我需要做什么才能让我的帐单更便宜？账单的可解释性做得越好，用户体验也就越好。

但是如果计费测量的粒度过细，也会影响产品本身的性能以及增加实现的成本。这里面需要平衡。但可以确定的是，在思考可观测性产品的方向上，成本分析可以作为一个独立的新视角。

成本分析可以帮助用户发现系统运行中的潜在问题，并采取措施予以优化。例如，如果用户观测到某个数据库实例的 CPU 使用率较低，但成本却很高，就可以考虑将该实例的规格调整为更低的级别。

AWS 今年发布的 Cost and Usage Dashboard 和 Reinvent 上 Amazon CTO Dr. Werner 的演讲专注于成本的架构艺术也同样可以看到这个趋势。他提出了“俭约架构”七大法则来在云的环境中打造更加高效、可持续的系统，为我们提供了一个系统性的指导框架。

/ 预测四 /

当 GenAI 时代的各种应用和工具变得越来越轻巧，开发者体验将成为现代数据库设计的核心目标之一。

数据库平台化不仅仅是漂亮的 Web 管控界面以及一些花哨的功能堆砌。我很喜欢 PlanetScale 的 CEO Sam Lambert 在他的个人 Blog 里面关 Develop Experience 的描述他引用了乔布斯的一句话“Great art stretches taste, it doesn’t follow tastes（ 伟大的艺术拓展审美边界，而不是刻意迎合。)”。

好用的工具之所以好用，是因为其中是饱含了设计者的巧思和品味，而且这个设计者也必须是重度的使用者，这样人们才能体会到那些细微的快乐与痛苦，但是又不至于沉浸其中使其盲目 ，其实这对负责开发者体验的产品经理来说是极高的要求。

数据库管理工具作为一种频率不算高频、但每次使用都很严肃的工具，在 AI 和云的时代，我认为有一些与体验紧密相关的设计原则是需要遵守的：

API First, 数据库平台应该提供稳定的 / 前向兼容的 API，一切在管控平台里能干的事情，API 都要能做到，最好你的管控平台是基于你的 API 构造的。这为你提供一个功能齐备的好用的 CLI Tool 也是关键的

必要条件。

使用统一的认证体系，在设计阶段将管控的认证和用户体系与数据库内部的认证体系打通，传统的数据库基于用户名和密码的权限体系在云的时代是不够的。这为了后续与云的 IAM 和 Secret 管理体系对接打下基础。

对不同的功能构建不同的 / 稳定的小工具 (Do one thing, do things well)，但是通过一个统一的 CLI 入口和语义系统进行调用。比较好的例子是 rustup, 甚至 git 也是个很好的例子。

稍微总结一下，2024 年，数据和数据库技术仍然处于巨大的变革期，谁也没办法预测未来，因为我们就身处这么一个不确定性巨大的时代。但好的一面是，创新仍然层出不穷。我今天预测的，很可能过几个月就会被我自己全部推翻，也是很正常的事情，如果能给当下的你有所启发，那就够了。

tidb

最后修改时间：2024-02-19 10:36:18

评论

分享你的看法，一起交流吧～

Oracle6 LV.5

叨咕来叨咕去，其实就是一个事：在AI的裹挟下，DB是跟还是不跟？不跟，必死；跟，拖死，你说咋办？

1年前 点赞 评论

雪狼sunny LV.6

东旭：“向量数据库”还是“向量搜索插件 + SQL 数据库”？| 我对 2024 年数据库发展趋势的思考

1年前 点赞 评论

相关阅读

【大盘点】2024年国产数据库行业有哪些大事发生？
墨天轮编辑部 789次阅读 2025-01-20 12:30:33

TiDB x DeepSeek 打造更好用的国产知识库问答系统解决方案
严少安 287次阅读 2025-02-07 00:51:17

重磅！近4800万国产数据库大单落锤！达梦、海量数据库、金仓、虚谷、TiDB 五大厂商中标！网思科技双包中标~
天下观查 196次阅读 2025-02-05 07:00:45

3.1 TiDB 社区活动深圳站 | 大规模 TiDB 国产化替代与成本优化在金融、跨境电商等行业的最新实践
PingCAP 95次阅读 2025-02-18 09:38:22

一行代码不用写，用 Autoflow + Gitee AI 搭建本地知识库问答机器人
PingCAP 55次阅读 2025-01-24 10:03:56

黄东旭：2025 数据库技术展望
数据库应用创新实验室 52次阅读 2025-02-06 10:07:08

TiDB 的高可用实践：一文了解代理组件 TiProxy 的原理与应用
PingCAP 48次阅读 2025-01-20 09:47:54

百亿大表的实时分析：华安基金 HTAP 数据库的选型历程与 TiDB 使用体验
PingCAP 46次阅读 2025-01-20 09:47:53

TiDB 分布式数据库多业务资源隔离应用实践
PingCAP 45次阅读 2025-01-24 10:03:57

攻克多版本运维难题：爱奇艺百套 TiDB 集群升级至 v7.1.5 实战宝典来袭！
PingCAP 38次阅读 2025-01-22 09:32:13