

OceanBase

向量能力与AI应用

案例集 (2025)



目录

一、引言	01
1.1 为什么现在必须重新思考“数据库与AI的边界”	
1.2 OceanBase Vector 部分代表用户	
二、Data x AI 革命	03
2.1 Data x AI 发展趋势：从非结构化数据爆发到技术架构革新	
2.2 向量嵌入可以帮助理解语义	
2.3 向量数据库的作用及场景	
三、OceanBase Vector 构建一体化向量数据库	07
3.1 OceanBase 向量特性介绍	
3.2 OceanBase 向量的五大优势	
3.3 OceanBase：全球技术领先的分布式数据库	
3.4 Data x AI Landscape：开源开放，深度融入全球AI生态圈	
四、向量数据库选型和未来趋势	17
4.1 选择向量数据库的考虑因素	
4.2 性能基准	
4.3 迁移服务	
4.4 未来趋势	
五、案例	21
6.1 货拉拉：资损代码识别与数仓 AI 答疑	
6.2 携程：基于 OceanBase 向量搭建酒店图片检索	
6.3 中国联通软研院：ChatDBA 构建运维 RAG	
6.4 九讯云：电商导购 + 商品快速搜索 + 智能客服	
6.5 慧视通：智能登签	
6.6 三维家：基于 OceanBase 向量的家具智能推荐	
6.7 视源股份：学习机绘本题库	
6.8 当贝投影：观影记录推荐	
6.9 伯俊科技：知识库	
6.10 in银泰商业：零售业智能问数平台	
6.11 某跨境电商：电商网站推荐系统	
6.12 百宝箱：AI Agent	

一、引言

IDC DataSphere 数据显示，到 2027 年，全球非结构化数据将占到数据总量的 86.8%，达到 246.9ZB。全球数据总量从 103.67ZB 增长至 284.30ZB，复合增长率为 22.4%，呈现稳定增长态势。随着互联网内容化的飞速发展，以音视频等多媒体内容为代表的非结构化数据呈现出高速增长的趋势。文本、图片、音频和视频等非结构化数据的存储和检索需求也变得越来越。

当前我们大量的基础设施几乎都是面向结构化数据而设计实现的，可以预见数据处理技术需要从结构化向非结构化转变，以更好的应对海量非结构化数据的管理需求。

在生成式 AI 时代，非结构数据的存储、计算和应用将得到急剧发展。非结构化数据的爆发不仅是技术挑战，更是商业机遇。向量数据库正从单一检索工具演变为 AI 原生数据基础设施，其与云计算、边缘计算、具身智能等融合将催生更复杂的生态系统。未来具备混合查询、一体化架构与企业级数据安全保护能力的向量数据库，将成为企业智能化转型的核心引擎。

1.1 为什么现在必须重新思考“数据库与 AI 的边界”

人工智能技术的快速发展，尤其是生成式 AI（如 ChatGPT）、检索增强生成（RAG）和多模态检索的兴起，正深刻改变企业对数据的需求。如今，数据不仅仅是简单的存储和查询对象，而成为智能应用和实时决策的核心资产。与此同时，海量的非结构化数据——文本、图像、音频、视频等多模态信息，正以前所未有的速度爆发增长，这对数据基础设施提出了新的挑战。

传统数据库主要面向结构化数据和事务处理设计，虽然在稳定性、一致性和高效事务处理方面表现优异，但面对非结构化数据的存储、向量检索以及实时智能计算时，功能和性能明显不足。特别是在 AI 驱动的实时应用场景中，数据库需要支持从高并发的在线事务处理（OLTP）到复杂的实时分析（HTAP），并且要无缝集成 AI 能力，才能满足业务快速变化的需求。因此，必须重新定义数据库与 AI 的边界，打造一个既支持事务一致性，又具备实时分析和 AI 推理能力的统一数据平台。这样的平台不仅能够处理结构化和非结构化数据，还能融合语义搜索、向量计算、多模态数据处理等 AI 功能，实现业务数据与智能计算的深度融合。

重新思考“数据库与 AI 的边界”，意味着推动数据库向实时智能化转型，构建面向未来的统一数据底座。这样的数据平台不仅是企业数字化的基石，更是实现智能经济和 AI 赋能业务的关键引擎。

1.2 OceanBase Vector 部分代表用户

零售先行者 in银泰商业 基于向量能力打造“智能问数平台”	跨境电商先行者 某跨境电商品牌 基于向量能力搭建“商品推荐系统”	智能设备先行者 当贝投影 基于向量能力搭建“观影推荐”
旅游先行者 携程 基于向量能力打造“智能图片系统”	物流先行者 货拉拉 基于向量能力搭建“资损代码识别与数仓 AI 答疑”	运营商先行者 中国联通软研院 基于向量能力构建“智能运维RAG”
智能交通先行者 慧视通 基于向量能力构建“实时核身系统”	零售服务先行者 伯俊科技 基于向量能力打造“伯俊 AI 通识小助手”	家居服务先行者 三维家 基于向量能力搭建“家具智能推荐系统”

二、Data x AI 革命

2.1 Data x AI 发展趋势： 从非结构化数据爆发到技术架构革新

在人工智能技术的推动下，数据的角色正经历深刻变革。过去，数据主要由结构化信息组成，便于存储、管理和处理。然而，随着 AI 的普及，数据的种类和形式变得更加多样，尤其是非结构化数据的爆发，推动了从传统数据存储到复杂数据处理的转型。这一转变不仅改变了数据的存储与查询方式，还带来了更为复杂的数据分析与处理需求。

随着 AI 的深入应用，数据不仅仅是“被存储”的对象，更是“智能化处理”的核心要素。尤其是在多模态数据处理和知识图谱构建等场景中，数据的多维度特性要求新的技术架构来支持更高效的处理与推理能力。因此，未来的数据处理不仅需要应对结构化数据的管理，还要支持非结构化数据、混合数据、以及实时数据的智能分析和多模态检索等多重需求。

在这一背景下，Data × AI 的深度融合正在成为技术发展的关键趋势。非结构化数据的爆发、技术架构的革新，以及 AI 技术的融合，这一趋势不仅重塑了数据处理范式，更催生出全新的技术架构与应用场景。在 Data x AI 的深度融合下，以下是预测的五个关键趋势方向：

非结构化数据爆发与多模态检索需求激增

非结构化数据的复杂性（高维度、语义关联性）迫使传统数据库向向量化转型。例如，企业需对视频流中的行为特征、医疗影像的病理特征进行实时分析。向量数据库通过将多模态数据（人脸、声纹、指纹等）统一转化为向量，支持跨模态联合检索。这一能力在认证场景中尤为重要，例如结合人脸与声纹的综合打分机制，准确率远超单一模态检索。预计到 2028 年，全球向量数据库市场规模将达 43 亿美元，年复合增长率 23.3%，其中多模态场景占比将超过 40%。

技术架构革新：混合查询与分层存储成为标配

为应对非结构化数据的高维特性，向量数据库正从单一稠密向量检索向混合查询演进：

稀疏与稠密向量融合：
结合 BM25（稀疏向量）与深度学习模型（稠密向量）的混合检索方案，可提升语义搜索准确率（如电商场景中关键词与商品特征的联合匹配）。

多级存储优化成本：
基于内存、本地磁盘、对象存储的分层架构，将冷热数据分离存储，降低内存依赖成本。例如，热数据（高频检索的实时视频特征）存于内存，历史数据转存至低成本对象存储。

AI 融合：从 RAG 到大模型的“数据中枢”

生成式 AI 的爆发使向量数据库成为大模型的“长期记忆体”与知识增强工具：

RAG（检索增强生成）：
通过向量数据库存储外部知识库（如企业私有文档），结合用户提问生成上下文增强的 Prompt，减少大模型“幻觉”问题。

训练与推理优化：
在模型训练中，向量数据库统一多源数据格式（文本、图像等），加速特征提取与检索；在推理阶段，其作为隐私屏障，隔离用户敏感数据与通用模型，防止信息泄露。

云上的向量数据库的技术栈加速迭代

未来会出现新的存储形态，即 VaaS（Vector as-a-Service），即用户将使用公有云服务，利用更方便的向量处理接口，支持存储和梳理向量数据。同时可以结合 GPU 等硬件加速索引构建与相似度计算，提升大规模数据吞吐量。另外可以结合云的弹性和分布式结构特性，满足 AIGC 业务的突发流量。

非结构化数据爆发与多模态检索需求激增

在 AGI 时代，数据的质量、可获取性和组织方式正逐渐成为制约智能化能力上限的核心因素。相比单点技术的突破，真正支撑 Data x AI 落地的，是底层数据基础设施的统一与演进。
当前，大模型能力的提升已经从“算法为王”过渡到“数据为本”。生成式 AI 的幻觉问题、本地知识补全、多模态理解能力、实时交互性能等核心挑战，本质上都指向一个共同根源——对高质量数据的高效组织、融合与使用能力的缺失。这不仅需要更强的模型，还需要一个能够支撑全链路数据流转、智能处理和安全管理的数据底座。

统一的数据底座不仅提供结构化与非结构化数据的融合能力，还具备以下关键能力：

低成本获取与高质量治理：
应对数据获取难、数据质量评估难的问题。

跨模态融合与统一建模：
满足文本、图像、行为等多模态数据的统一表达与查询需求。

实时性与可扩展性结合：
支持海量长尾场景下的在线智能应用。

流动性与可控性并存：
保障金融、医疗等高严谨行业数据的安全可控流动。

2.2 向量嵌入可以帮助理解语义

非结构化数据是指没有明确定义的数据格式和组织结构的数据。非结构化数据通常包括文本、图像、音频、视频等形式的数据，以及社交媒体内容、电子邮件、日志文件等。由于非结构化数据的复杂性和多样性，处理这些数据需要采用特定的工具和技术，例如自然语言处理、图像识别、机器学习等。

传统搜索技术允许我们使用关键词来从非结构化数据中获得信息。这些方法通常在理解内容的深层含义方面存在不足：丢失了上下文、语义和用户意图。因此需要使用高级技术如自然语言处理来识别文本中的关键实体、情感和主题。

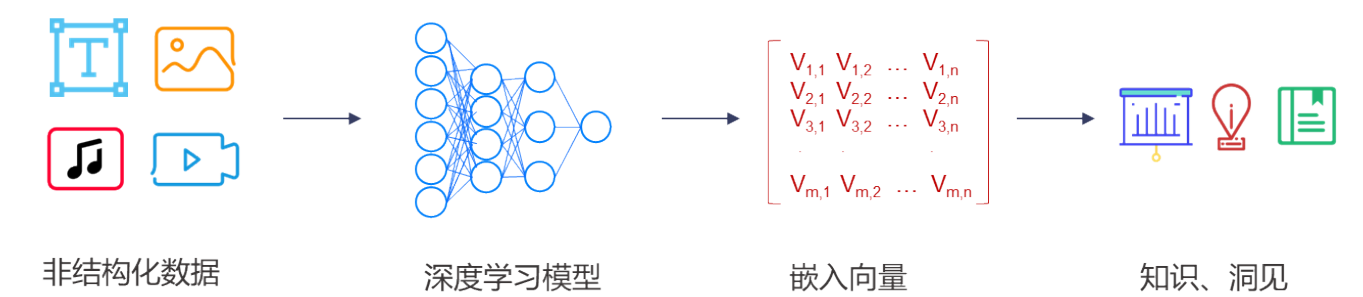
处理非结构化数据最强大的方法之一是使用向量 Embedding（向量嵌入）：通过深度学习神经网络提取非结构化数据里的内容和语义，将非结构化数据（文图音视）变成特征向量的过程。Embedding 技术将原始数据从高维度空间映射到低维度空间，将具有丰富特征的多模态数据转换为多维向量数据。这些 Embedding 由许多维度组成，有时被称为特征，而传统的基于关键词的搜索不同，向量 Embedding 采用语义相似性搜索，其中在高维空间中接近的向量代表具有相似含义或内容的项目。

2.3 向量数据库的作用及场景

人工智能技术的快速发展，尤其是生成式 AI（如 ChatGPT）、检索增强生成（RAG）和多模态检索的兴起，正深刻改变企业对数据的需求。如今，数据不仅仅是简单的存储和查询对象，而成为智能应用和实时决策的核心资产。与此同时，海量的非结构化数据——文本、图像、音频、视频等多模态信息，正以前所未有的速度爆发增长，这对数据基础设施提出了新的挑战。

传统数据库主要面向结构化数据和事务处理设计，虽然在稳定性、一致性和高效事务处理方面表现优异，但面对非结构化数据的存储、向量检索以及实时智能计算时，功能和性能明显不足。特别是在 AI 驱动的实时应用场景中，数据库需要支持从高并发的在线事务处理（OLTP）到复杂的实时分析（HTAP），并且要无缝集成 AI 能力，才能满足业务快速变化的需求。因此，必须重新定义数据库与 AI 的边界，打造一个既支持事务一致性，又具备实时分析和 AI 推理能力的统一数据平台。这样的平台不仅能够处理结构化和非结构化数据，还能融合语义搜索、向量计算、多模态数据处理等 AI 功能，实现业务数据与智能计算的深度融合。

重新思考“数据库与 AI 的边界”，意味着推动数据库向实时智能化转型，构建面向未来的统一数据底座。这样的数据平台不仅是企业数字化的基石，更是实现智能经济和 AI 赋能业务的关键引擎。



现代向量数据库能够处理不同规模的向量数据，从千万到亿到百亿，并且支持多租户，提供全面的 API 支持，涵盖多种的编程语言，并提供全面和易用的管理界面。向量数据库还具有混合搜索能力，结合传统关系数据库中的标量过滤和向量相似性搜索以提供更准确的结果。向量数据库是语义搜索、推荐系统、图像识别和自然语言处理等应用的基础，并赋能 GenAI 应用。

检索增强生成（RAG）：

RAG 是一个人工智能框架，用于从外部知识库中检索事实，以便为大型语言模型（LLM）提供最准确、最新的信息，并让用户深入了解 LLM 的生成过程，常应用于智能问答、知识库等。

高级搜索（图像/文字）：

图像/文本检索任务是指在大规模图像/文本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，通过高性能的索引组织实现高效的相似度计算，进而返回和检索内容相匹配的图像/文本结果，典型场景如人脸识别等。

个性化推荐：

推荐系统可以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。当发起推荐请求时，系统会基于用户特征进行相似度计算，然后返回用户可能感兴趣的物品作为推荐结果，如饭店推荐、景点推荐等。

AI 智能体：

企业级 Agent 支持利用向量搜索实时访问和检索高维数据，实现多模态数据处理、RAG 驱动的多步决策等。Agent 对数据库提出综合型需求，包括：混合查询能力、高并发与低延迟、多模态交互、数据安全与合规等。

三、OceanBase Vector 构建一体化向量数据库

3.1 OceanBase 向量特性介绍

OceanBase 向量数据库提供了完备的向量检索功能，支持基于图的 HNSW 以及基于倒排的 IVF 的索引组织算法；支持通用的量化算法，Flat、SQ8、PQ、BQ 等，量化算法和索引组织算法按需结合，以适用各种复杂的向量检索场景。提供丰富的距离算法，支持曼哈顿距离、欧式距离、内积、余弦距离等多种类型向量距离的计算，以支持不同应用场景对距离算法的诉求。同时支持稠密向量、稀疏向量，也支持多种类型向量的混合检索，以提升检索效果。

OceanBase 向量检索具备带有标量过滤的混合检索能力，标量过滤可以辅助快速而且精确地缩小查询空间。基于 OceanBase 长期在 OLTP 领域的打磨，OceanBase 标量和向量的混合检索提供更强的查询性能。

在开发者生态中，OceanBase 的向量提供灵活的访问接口，支持通过 MySQL 协议及各种语言的客户端访问，也可以使用 Python 和 Java 的 SDK 访问。同时 OceanBase 数据库也完成了对 AI 应用开发框架 LlamaIndex、DB-GPT 及 AI 应用开发平台 Dify 等的适配，更好的服务于 AI 应用开发。

OceanBase 向量整体架构

用户访问层	SQL 多语言客户端			Python SDK		
多模数据层	基础类型	JSON	GIS	Vector	HBase	Redis
检索层	融合查询			并行执行		
索引层	向量索引	全文索引	GIS 空间索引	多值索引	二级索引	
内核	事务	高可用	多分区水平扩展	高效存储	弹性扩缩容	读写分离
工具体系	ODC	OMS	OAS	OCP	OMA	备份恢复

3.2 OceanBase 向量的五大优势

OceanBase 向量具备以下的五大优势：

海量向量

简化开发运维

融合查询

不止于向量

创新&安全可靠

海量向量，超群性能：原生分布式架构，支撑从千万到千亿+数据量的向量场景

分布式架构：高增速的非结构数据，需要分布式架构处理海量向量；

海量向量：OceanBase 一体化架构，通过混合向量+磁盘的向量算法，以及叠加 OceanBase 的分布式技术，支持千万、亿级、千亿+不同场景的向量处理；

- 千万**

针对中大型用户企业，提供从数十万到千万级向量支撑。
- 亿**

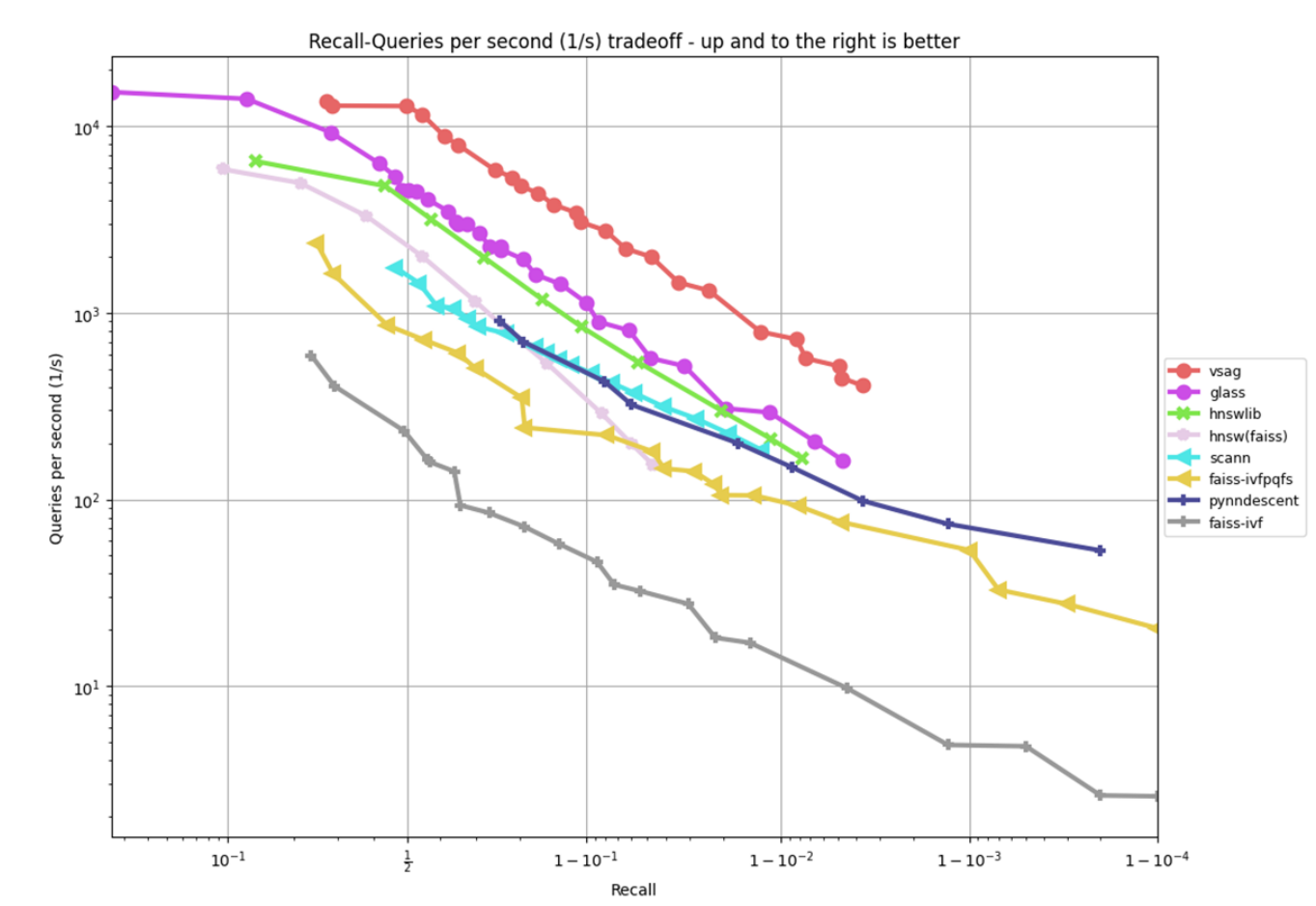
对超大型用户企业，提供从千万到亿向量支撑。
- 千亿+**

对泛互联网或者面向 C 端的场景，支持十亿到千亿的向量场景支持。

超高性能：OceanBase 和蚂蚁自研了高性能的向量计算库，该算法库（Vsag，已经开源）

在 ANN-Benchmarks 评测，在 gist-960 数据集上达到了 SOTA 的水平。

和目前最好的 glass 比，VSAG 在 96.7% 召回率的情况下 QPS 提升了 90% 。

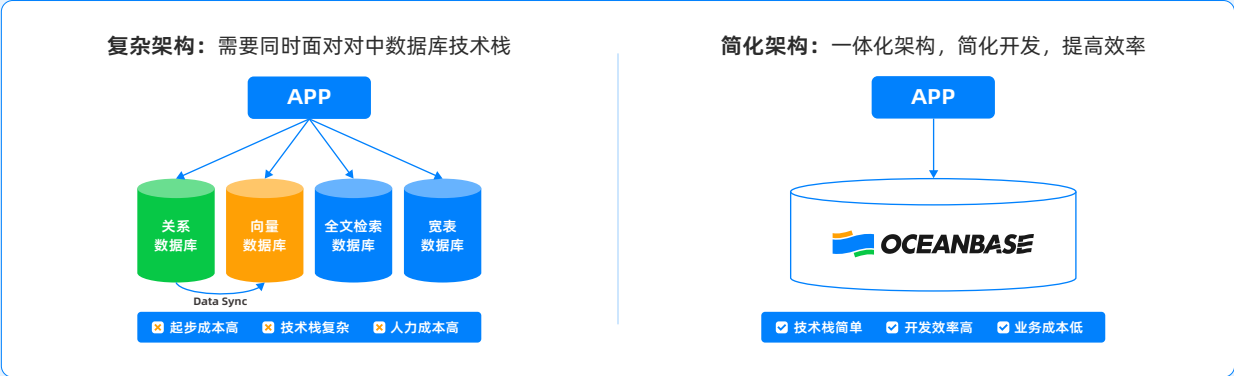


极简架构，敏捷运维：对开发、运维、架构友好

对开发而言：SQL Native，支持多种开发语言和生态

- | 兼容 MySQL 协议
- | 接口简单, 尤其是 SQL 接口
- | SQL 原生 AI: SQL+AI, 同时支持 Python 和 Java 原生调用
- | 生态丰富,主流框架已经对接

对架构师和开发人员而言，多库合一



对运维而言：对称架构方便运维、通过多租户解决数据权限

多模融合，混合检索：提升检索效率和效果

融合查询：同时支持关系、NoSQL、KV、GIS 等多种数据存储，可以同时支持多种数据的融合查询

先计算向量还是标量？

Hybrid Search Optimizer

后过滤 暴力搜索 前过滤

更快

丰富的执行策略
基于代价的自动选择性能最好计划

先计算向量数据会不会少？

标量查询 (数据库) 后过滤 前过滤

向量查询 (向量库) 标量迭代器 标量 Bitmap

更准

自研向量算法库
深度和数据库集成，兼顾查询效率&准确性

RAG 等场景能否一条 SQL 完成多路查询？

一条 SQL

Hybrid Search

标量索引 向量索引 全文索引 空间索引 JSON 多值索引

标量 文本 空间 JSON

一行数据

OceanBase 数据库

更易用

内核级多模混合查询
更丰富的“模”做更有效率的融合查询

在向量数据库中，标量（传统结构化数据）与向量（高维嵌入表示）的融合设计，是提升 AI 应用性能与功能完整性的核心优势。这种融合不仅解决了多模态数据处理难题，还显著优化了复杂查询场景的效率与精准度。

以下是具体优势分析说明：

提升查询精度与语义理解能力

- 技术原理：标量数据（如用户 ID、时间戳、类别标签）提供上下文和业务约束条件，向量数据则捕捉语义相似性。融合查询通过标量过滤缩小候选集，再通过向量相似度排序返回最相关结果，避免纯向量检索的语义漂移问题。
- 场景：金融风控系统--检测异常交易行为
- 融合查询：标量过滤：筛选“近 24 小时内、金额 > 10 万元”的交易记录。
 - 向量检索：在候选集中，通过用户行为向量（如操作序列 Embedding）匹配异常模式。
 - 效果：降低误报率，提升召回率。

优化响应效率与资源消耗

- 技术原理：纯向量检索需全库扫描，计算开销大。标量条件（如时间范围、状态标签）可快速定位分区数据，减少需计算的向量量级，显著降低延迟和 CPU / 内存压力。
- 场景：电商推荐系统--为用户推荐“夏季新款连衣裙”。
- 融合查询：标量过滤：限定“品类=女装、上架时间 > 2025-05-01”。
- 向量检索：基于用户浏览向量匹配相似商品。
- 效果：大幅提升 QPS，显著降低响应延迟。

实现多模态数据联合分析

- 技术原理：非结构化数据（如图片、文本）转为向量后，需与结构化元数据（如来源路径、作者信息）关联。融合存储支持跨模态联合查询，突破传统数据库的模态壁垒。
- 场景：医疗知识库检索--查找“肺炎 CT 影像及治疗方案”。
- 融合查询：标量过滤：限定“患者年龄 < 12 岁、疾病类型=肺炎”。
- 向量检索：匹配 CT 影像向量相似的病例，并关联对应文本诊疗方案。
- 效果：提升诊疗决策效率。

增强动态数据实时性

- 技术原理：标量数据常需实时更新（如库存状态、用户权限），向量数据更新成本较高。融合设计允许独立更新标量字段，无需重建向量索引，保障业务敏捷性。
- 场景：实时库存管理--快速响应“限量款球鞋”的抢购请求。
- 融合操作：标量实时更新：库存数量随交易动态调整。向量索引不变：商品特征向量无需重新训练。
- 效果：减少数据更新延迟，更好支持高并发秒杀场景。

支持复杂业务逻辑的灵活扩展

- 技术原理：通过 MySQL 兼容接口，开发者可直接编写含标量条件与向量函数的混合查询，降低 AI 应用开发门槛。
- 场景：智能编程助手--IDE 插件中搜索“Python 初始化日志文件”的代码片段。
- 效果：提升代码定位准确率，提高问题解决效率。

技术降本，工具降本：高压缩比、完整的数据库工具降低降本

超强压缩，轻量成本：

内存降本：

支持二进制定量化算法，同等向量性能内存成本降低 95%；

磁盘降本：

支持半结构化（JSON）数据压缩，助力 AI 应用中半结构化数据存储成本降低超 40%。

业界领先的压缩算法，内存与磁盘多层面显著降低存储成本

不止于向量：

提供完整的数据库能力：

分布式	基于 OceanBase 的水平扩展能力及多分区能力，支持海量向量数据储存	低成本	基于透明压缩能力，支持超过万维向量存
高可用	支持主备/跨机房/跨地域容灾，数据备份恢复能力即使是内存 HNSW 索引，容灾切换后也能实时访问	融合查询	支持向量索引、标量索引以及半结构化索引（多值索引、空间索引）全文索引支持同时访问多路索引
事务	分布式事务能力保证了向量数据的一致性和完整性	多租户	完善的资源共享和资源隔离能力
企业级安全特效	身份鉴别和认证、访问控制、数据加密、监控告警、安全审计	完善的工具体系	ODC/OCP/OMA/OMS

提供完整的 OceanBase 的工具体系：



包括如下企业级特性：

- | 完善的资源共享和资源隔离能力
- | 完整的企业级安全特性
- | 身份鉴别和认证、访问控制、数据加密、监控告警、安全审计能力；
- | OCP 监控无死角, 报警多通道
- | 限流降级, 快速止损
- | 稳定可靠的逃生方案, 完整容灾恢复
- | 复用数据库高效故障诊断

企业技术创新、国家信息技术应用创新

代码根自研：



根自研代码

内核代码级掌控，深耕复杂场景
为关键业务系统兜底



自建研发环境和流程

自研编码优化技术
自建研发流程



没有开源协议风险

不受外部社区制约和断供
拥有完整自主知识产权



根自研开源

自主掌握技术路线
主导产品发展和社区方向



共建根自研生态圈

构建自研技术生态圈
打造高可靠性的产品生态体系

100%根自研，确保自主可控

引领数据库行业多项创新

性能创新

全球唯一刷新过 TPC-C 和 TPC-H 测试世界纪录的分布式数据库

2019 年 10 月 19 日，OceanBase 首次参加被誉为数据库领域“世界杯”的事务型基准测试 TPC-C，以 **6088 万 tpmC** 打破 Oracle 保持 9 年的世界纪录。2020 年 5 月 20 日 OceanBase 在被誉为数据库领域“世界杯”的事务型基准测试 TPC-C 中，打破自己保持的世界纪录，获得 **7.07 亿 tpmC** 较上次测试提升 11 倍，创下新的世界纪录。2021 年 5 月 19 日，OceanBase 基于 64 台阿里云 ECS 服务器（共计 5120 核）的计算集群，在分析型基准测试 TPC-H 中，以 **1526 万 QphH@30000GB** 的性能总分，创下新的世界纪录。

容灾创新

首创“三地五中心”城市级故障自动无损容灾解决方案

OceanBase 首创“三地五中心”城市级故障自动无损容灾解决方案，**满足国标金融 6 级容灾标准**。依托原生分布式内核技术整个过程完全自动切换，无需人工干预，实现机柜级、机房级、城市级无损容灾，保障企业数据安全、业务稳定。

4.0 版本再造“RPO = 0, RTO < 8s”新标准

2014 年，OceanBase 0.5 版本发布在故障自愈上做到

**RPO = 0
RTO < 30s**

成为行业标准

2022 年，OceanBase 4.0 版本发布再次取得创新突破做到

**RPO = 0
RTO < 8s**

成为业内首个将 RTO 做到 8 秒的数据库

架构创新

业内首个单机分布式一体化数据库

OceanBase 4.0（小鱼）是业内首个单机分布式一体化数据库，打破分布式技术的不可能，有效兼顾**分布式架构的扩展性与集中式架构的性能优势**，斩获 2022 年世界互联网大会“领先科技成果奖”。

一体化创新

TP & AP、多模一体化融合

2024 年，OceanBase **推出列式存储**，进一步强化 TP & AP 一体化，**提供 SQL + NoSQL 多模能力**。可在一个数据库、一份数据，同时实现联机事务处理与秒级实时分析，统一管理多种数据类型，简化技术栈。

多云原生创新

从单机、分布式 2F1A、2F1L、3F 等**多样化部署形态**，到同城多活、异地冷备、异地热备、异地多活、跨洲际多活、三地五中心等灵活的**多云容灾能力**，全面保障数据高可用性和业务连续性。

根自研：2024 年 9 月通过首批通过分布式数据库“安全可靠测评”，可基于一个数据库产品同时提供交易、分析、原生 AI 工作负载支持。



 **OceanBase 首批分布式数据库通过“安全可靠测评”** 

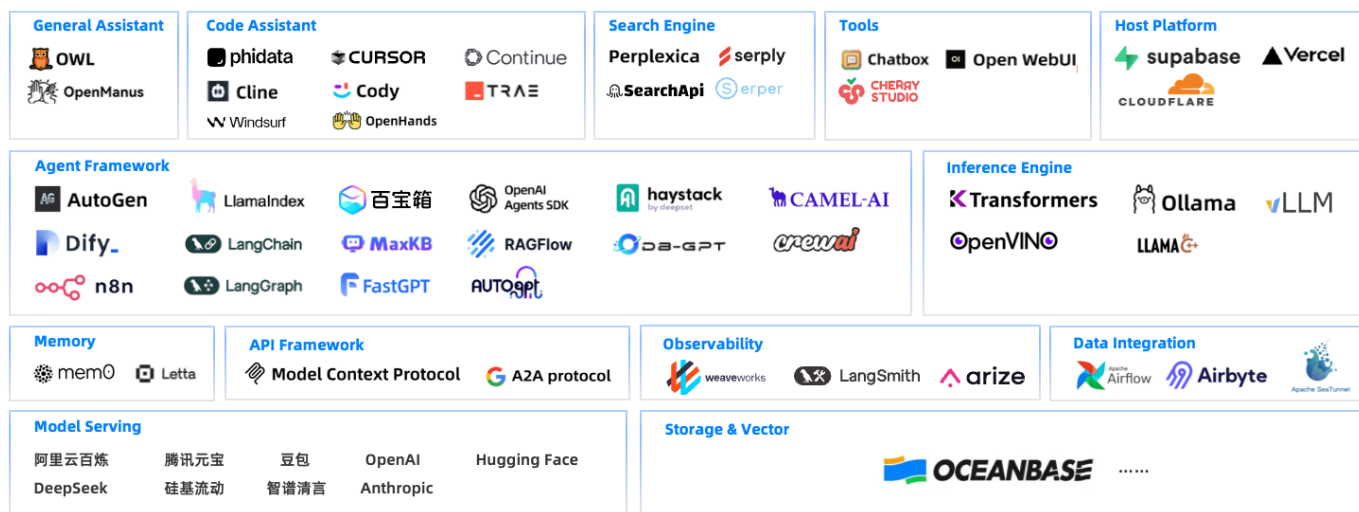
3.3 OceanBase：全球技术领先的分布式数据库

OceanBase 是 100% 根自研的原生分布式数据库，基于单机分布式一体化架构，高度兼容 Oracle 和 MySQL，支持事务处理（TP）和实时分析处理（AP）等多工作负载，原生支持向量检索与多模数据混合检索，助力客户构建面向 AI 的一体化数据底座。已广泛应用于金融、运营商、零售、互联网及公共服务等行业，助力 2000+ 客户实现关键业务系统升级。



3.4 Data x AI Landscape: 开源开放，深度融入全球AI生态圈

OceanBase 正在构建面向 AI 应用的开放技术生态，通过与主流开发框架和协议的集成，降低数据库在智能应用中的接入门槛。当前，OceanBase 已与 LlamaIndex、LangChain、Dify 等 60 余个 AI 生态工具打通，支持语义检索、上下文构建、嵌入式问答等主流能力，适配 RAG、搜索、推荐及 Agent 等典型 AI 应用场景，为企业打造更高效、易用且具备高度扩展性的一体化数据底座。



在大模型标准协议层面，OceanBase 也已原生支持 MCP，通过 OceanBase MCP Server 提供对数据库的结构化查询、实时监控、智能诊断等能力的封装。目前该能力已接入阿里云魔搭社区、anserPACK 等平台，开发者可以在大模型界面中通过自然语言提出问题，由 MCP Client 调用数据库工具链自动完成执行与结果返回，从而降低 AI 应用的数据访问门槛，简化上下游集成路径。这种从接口协议到生态对接的整合，不仅让 OceanBase 更容易被纳入现有的 AI workflows 中，也让数据库本身在生成式 AI 场景下具备了更强的主动能力。

全面支持MCP，构建智能体开发新范式



四、向量数据库选型和未来趋势

4.1 选择向量数据库的考虑因素

支持多种访问方式：
包括 SQL、Python 和 Java API 的支持。

支持多模数据和混合检索：
在一个数据库同时支持多种数据类型包括标量、向量、空间和 NoSQL 等。同时提供向量和标量，向量和全文本的混合检索。

高性能和可扩展性：
支持分布式架构和数据分片，处理能力需远超单机数据库，特别是在处理数百亿向量或数 TB 数据时。

完整的数据持久性管理：
数据库应具备强大的备份、复制和灾难恢复功能。

企业级服务：
企业可以提供企业级服务，特别是对于关键的应用程序。

4.2 性能基准

VectorDBBench 是一款为主流向量数据库和云服务提供基准测试结果的工具，在 VectorDBBench 中集成了主流向量数据库的测试客户端，基于 VectorDBBench，可以轻松对比各个数据库在典型测试负载下的性能，详细的性能测试流程，可以参考 OceanBase 官网，向量数据库 VectorDBBench 测试章节。

如下介绍 2 个典型性能测试数据集下的性能对比：

1.

第一个数据集有 100 万条向量数据，每条向量 768 维。

2.

第二个数据集有 50 万条向量数据，每条向量 1536 维。

在每个数据集中，测试了三种场景：

1.

纯粹的向量检索场景，查询条件中只有向量。

2.

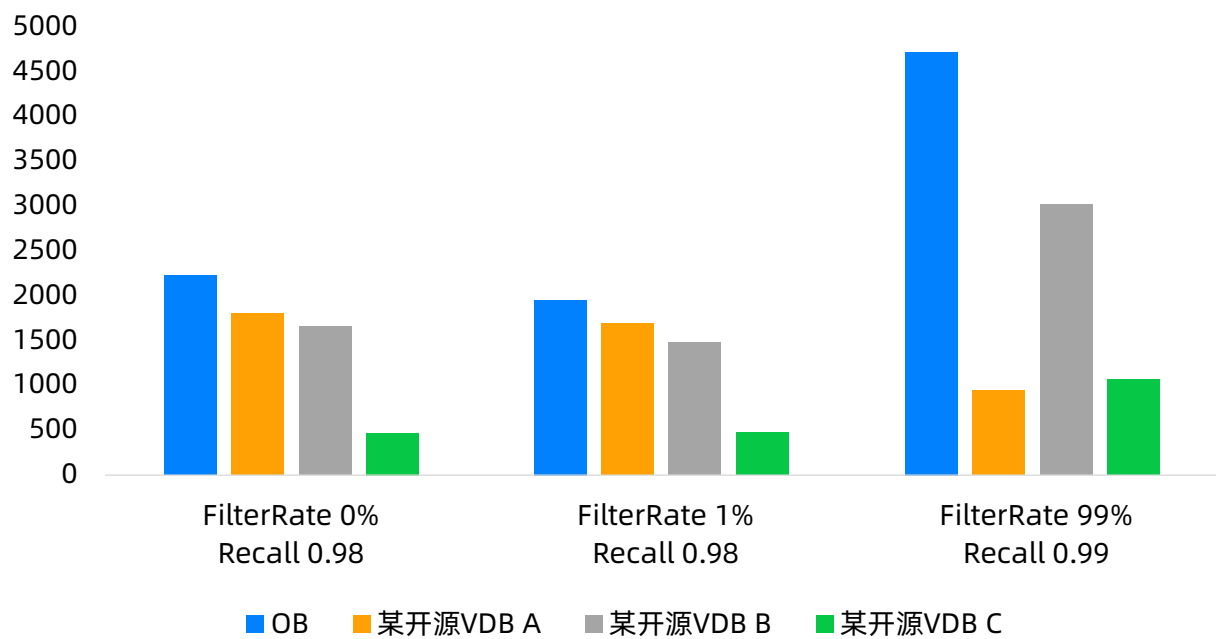
标量和向量混合检索的场景，同时通过标量，只有 1% 的数据被过滤。

3.

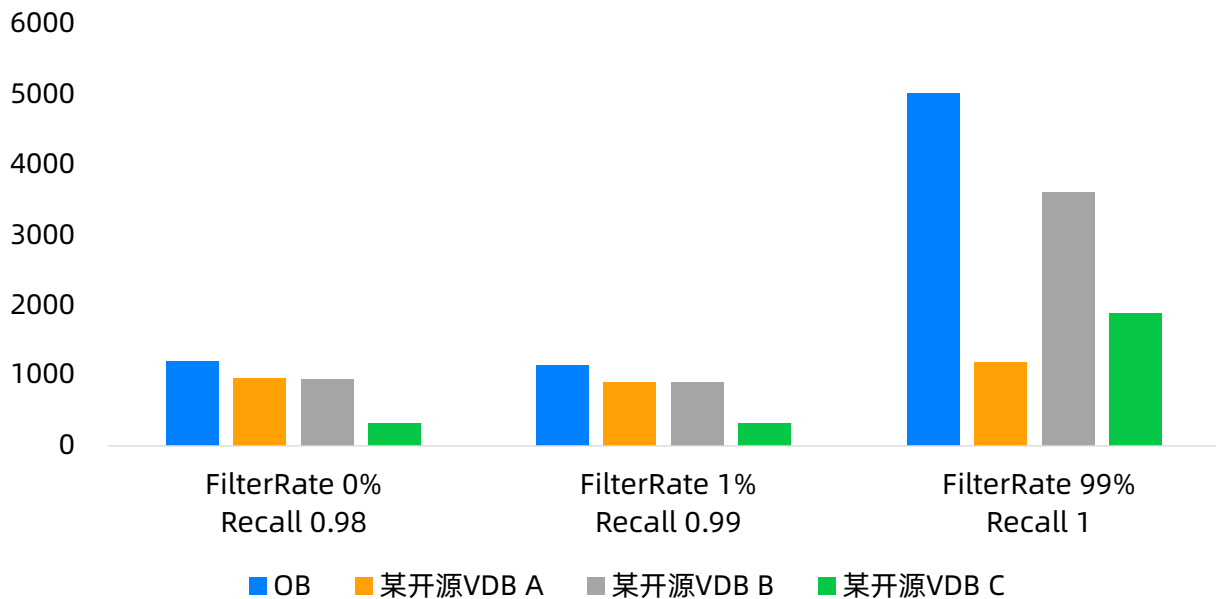
标量和向量混合检索的场景，同时通过标量，有 99% 的数据被过滤。

向量检索是个近似检索，相比于传统关系数据库的各种 Benchmark，引入了一个新的指标 Recall。暴力搜索（查询向量和数据集中所有向量做距离检索）的 Recall 被定义为1（100%），其他更高效的向量算法召回的结果集和暴力搜索的结果集之比，是该向量算法的召回率。召回率在在一定程度上是可以通过调整检索参数来提升，但是一般情况下，会导致性能的下降。在本次 VectorDB Bench 的对比中，每一组测试场景，都会将各个数据库的 Recall保持在一个固定值，最终对比在当前召回率下各个数据查询的 QPS。

Performance768D1M



Performance1536D500K



4.3 迁移服务

OceanBase 提供了完整的向量数据迁移方案，帮助您轻松地将现有向量数据库中的数据迁移至 OceanBase。

目前支持以下主流向量数据库的数据迁移：

Milvus 数据迁移

支持将 Milvus 数据库中的向量集合(Collection)、向量数据及其元数据无缝迁移至 OceanBase。迁移过程保证数据的完整性和一致性，让您的业务平滑过渡。

Elastic Search 数据迁移

支持全量迁移 Elasticsearch 数据库的数据至 OceanBase，迁移步骤请参考官网。

PGVector 数据迁移

OceanBase 支持迁移 PostgreSQL 数据库的数据至 OceanBase，Pgvector 作为 PostgreSQL 生态的一个插件，可以复用 PostgreSQL 到 OceanBase 的迁移流程，迁移步骤请参考官网。

Qdrant 数据迁移

支持将 Qdrant 数据库中的数据迁移至 OceanBase，迁移工具会自动处理数据格式转换，确保迁移后的数据可以直接使用，请参见迁移 Qdrant 数据库的数据至 OceanBase。

通过 OceanBase 的向量数据迁移能力,您可以：

统一数据管理,减少多个数据库带来的运维成本。

充分利用 OceanBase 的高性能向量检索能力。

实现数据库功能的无缝切换和升级。

4.4 未来趋势

正如数据库教授 Andy Pavlo 所言：“向量检索本质是索引技术的演进，而非数据库范式革命”。

技术趋势：分布式架构为骨、多模查询为脉

企业中非结构化数据的比重远大于结构化数据，而且前者的增长速度远远高于后者，未来企业应面对海量的向量存储需求。

数据规模驱动：

大模型需处理千亿级向量（如 OpenAI 训练数据），传统单机库仅支持百万级，而分布式架构通过分片与弹性扩缩容实现横向扩展，单索引支持 10 亿级向量。

数据规模驱动：

分布式并行计算将查询延迟降至毫秒级，同时支持增量索引更新，避免全量重建。

混合查询能力：

混合查询能力：企业 90% 场景需同时处理向量（语义）与标量（业务过滤），如金融风控需先筛选“金额 > 10 万”再匹配行为向量。

跨模态统一引擎：

多模态大模型需统一文本/图像/音频的语义空间（如医疗中 CT 影像关联诊断文本）。

产品趋势：通用数据库是向量数据库主流

技术同质化：

向量数据库核心是索引算法（如 HNSW、IVF），而通用数据库通过补充相关索引算法即可实现向量检索，开发周期很短。

运维成本高：

独立向量库需单独管理数据安全、备份、权限，企业需维护多套系统（如 Elasticsearch+ 向量库），增加复杂度。

资本降温佐证：

2024 年开始资本回归理性，纯向量数据库厂商转向垂直场景。

生态整合需求：

企业数据大量存储结构化数据（订单、用户画像），向量数据需与业务库联动。纯向量数据库在企业中的适用场景非常窄。

成本与易用性：

OceanBase 引入向量引擎后，企业做 AI 应用的准备周期从周缩到天。如果使用 OB Cloud，更可以缩短到小时。

五、案例

6.1 货拉拉：资损代码识别与数仓 AI 答疑

公司简介

货拉拉创立于2013年，是从事同城/跨城货运、企业版物流服务、搬家、零担、跑腿、冷运、汽车租赁及车后市场服务的互联网物流商城。2025年，月活司机达100万，月活用户达1350万。

客户收益

- 业务稳定运行：避免之前 Weaviate 在高并发下经常宕机的问题；
- 支持混合检索：同时支持标量和向量的混合检索，简化技术栈；
- 高可用：避免之前发生故障后不能平滑自动迁移；另外提升了备份/恢复能力；
- 未来规划：货拉拉 AI 平台底座将基于OceanBase，将目前其他向量库，包括京东 vsearch和Milvus 迁移到 OceanBase。

业务挑战和需求

- 资损代码识别：是OceanBase 向量检索在货拉拉的重要应用场景，结合大模型能力与 OceanBase 向量检索，开发了自动化代码风险识别系统，该系统通过向量化历史案例数据并检索相似代码，利用大模型进行分析判断资损风险，从而提高代码审查的效率和准确性，控制开发过程中的风险。
- 数仓 AI 答疑项目：是 OceanBase 向量检索在货拉拉的另一个重要落地场景，将向量检索的能力应用于数仓 AI 答疑助手，提高数据查询效率，减轻了数仓开发人员的工作压力。

解决方案和场景描述

- 向量方案：压缩算法 HNSW：从而可以单分区支撑业务，无须分库分表；未来使用 IVF-SQ8：数据量更大后，升级到 IVF-SQ8 算法，节省内存；快速索引重建：重建索引仅 25 分钟。

The diagram illustrates the interaction between users and administrators with OceanBase vector search and Large Language Models (LLM). It is divided into two main sections: '用户' (User) and '管理员' (Administrator).

用户 (User) Section:

- Answer (答案):** The user provides a question, which is processed by the '大语言模型 (LLM)' to generate an answer.
- Question (问题):** The user provides a question, which is processed by the '向量嵌入模型 (Embedding Mode)' to generate a vector representation.
- Question + Related Information (问题+关联信息):** The user provides a question, which is processed by the '向量嵌入模型 (Embedding Mode)' to generate a vector representation, and then the 'OceanBase 向量检索' (OceanBase Vector Search) to retrieve related information.

管理员 (Administrator) Section:

- Document Knowledge Base (文档知识库):** The administrator manages the document knowledge base, including 'OceanBase 向量检索使用.md' and '文档切片' (Document Chunks).
- Vector Embedding Model (向量嵌入模型):** The administrator uses the '向量嵌入模型 (Embedding Mode)' to process document chunks into vectors.
- OceanBase Vector Search (OceanBase 向量检索):** The administrator uses the 'OceanBase 向量检索' to perform vector search on the document chunks, returning results like '例如: insert ... values ([1-3,1415926, 2.7182818, ...]);'.

6.2 携程：基于OceanBase向量搭建酒店图片检索

公司简介

携程是一个在线票务服务公司，创立于 1999 年，总部设在中国上海。携程旅行网拥有国内外六十余万家会员酒店可供预订，是中国领先的酒店预订服务中心。

客户收益

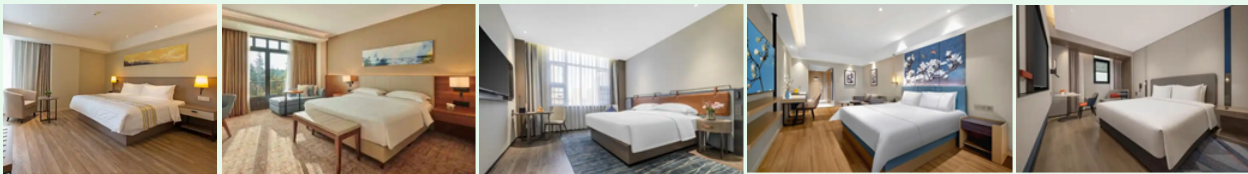
- 提升高可用，简化运维：解决之前 Elasticsearch + Faiss 的问题。
- 简化开发：支持标量和向量混合检索。
- 更好的支持数据增量：每天数百万数据的刷新。
- 稳定运行：上线近一年，0 故障。

业务挑战和需求

- 相似酒店匹配：寻找价格最便宜的相似酒店；根据酒店的照片，找到与之最相似的酒店。
- 识别不同渠道的同一酒店：一家酒店，识别可能有多家供应商。
- 大数据量和高可用性：数据量大, 数据库规模是 10 亿级；而且每天有增量数据刷新，同时要求极高的可用性。

解决方案和场景描述

- 高可用：使用 OceanBase 标准 1-1-1 3 节点保证高可用；
- 混合检索：支持标量和向量混合检索，同时使用标准 SQL 开发。
- 向量设计加速构建：采用 IVF 向量索引算法，支持增量自动刷新；向量查询使用 cosine 算法，对 location 进行分区，分区数 3000，向量维度 384；构建索引 4 倍速 加速构建。



6.3 中国联通软研院：ChatDBA 构建运维 RAG

公司简介

中国联通软件研究院成立于 2014 年，是中国联通集团总部提升自主研发能力的重要战略规划，公司在哈尔滨、济南、广州、西安、南京设五个分院，主要承担中国联通集团的内部业务支撑系统、管理支撑系统、大数据、互联网软件研发任务，以及前瞻性技术研究与应用。

客户收益

- 资源成本优化：硬件资源使用量减少约 30%，通过一体化数据库架构替代原有 MySQL+ ES 双库模式，降低硬件投入。
- 运维效率提升：技术栈简化，运维复杂度降低，无需同时维护关系型与向量数据库，通过 OceanBase OCP 平台实现统一监控、备份和扩缩容。
- 系统稳定性增强：OceanBase 分布式架构支持高可用、容灾与弹性扩展，减少单点故障风险，保障生产环境稳定性。
- 响应速度与准确性提升：RAG 架构结合 OceanBase 向量检索能力，实现基于企业数据的实时精准问答（如“ChatDBA”），提升数据库问题解决效率。
- 开发敏捷性优化：支持标准 SQL 和 SDK 接口，降低开发门槛；迁移工具 OMS 简化数据迁移，适配改造周期仅 1-2 周。

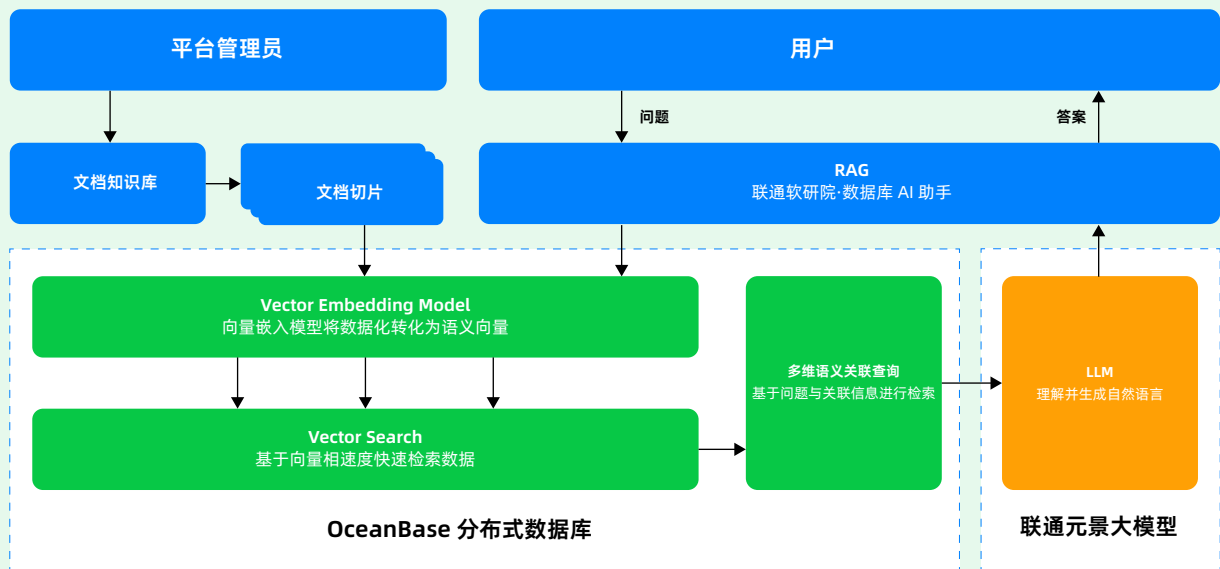
业务挑战和需求

- 复杂技术栈：针对内外部的数据库通识类和特性类文档进行系统性梳理，形成文档知识库，技术栈复杂。
- 可扩展性：使用 MySQL + ES 架构，不能水平扩展。可用性。

解决方案和场景描述

- 技术选型升级：从 MySQL+ ES 双库架构升级为 OceanBase 一体化数据库，支持关系型数据与向量数据的混合查询及分布式扩展。
- 向量检索能力应用：利用 OceanBase 的向量索引（HNSW）、多维度距离计算（余弦/欧式等）及 Filter 混合过滤功能，构建 RAG 知识库，提升问答召回率。
- 分布式架构优势：通过 Paxos 协议实现高可用，支持横向扩展，满足海量数据处理需求，并具备多租户资源隔离能力。

- **统一数据底座构建：**整合文档知识库，通过向量化模型嵌入和 OceanBase 存储，实现通识与私有数据的统一管理，支撑智能问答系统“ChatDBA”。
- **迁移与运维工具链：**使用 OMS 工具完成 Milvus 向量数据迁移，复用 OCP 平台实现部署、监控和备份，降低运维复杂度。



6.4 九讯云：电商导购 + 商品快速搜索 + 智能客服

公司简介

九讯云隶属于云南九机信息技术有限公司旗下，是基于九机零售模式，为手机零售行业定制开发的系统服务平台，包含 O2O 电商平台、全功能 OA 管理系统、系统高度融合 CRM、旧机回收转售、智慧物流、财务、营销、运营管控、店面管理、人力资源、培训管理等功能模块。

客户收益

- **电商导购提速：**下单流程从 20 分钟，下降到 1 分钟。
- **稳定运行：**电商导购之前用 Milvus 内存经常被打爆，而且运维复杂。
- **统一商品总体管理：**各门店系统迁移至九讯云平台时，需要将原有系统的商品库存等信息导入九讯云商品库存，但是每家系统的不同商品名称实际是同一个，使用语义快速匹配，能够快速匹配不同租户的同一个商品信息，快速实现商品总体管理。
- **销售知识库提效：**客服、呼叫中心等同事每天面对用户各种问题，对于商品参数信息、活动等无法做到准确记忆，为了帮助客户及呼叫中心同事快速检索商品参数，提升客服响应速度，将商城手机参数信息等构造成 QA 问答，并存入向量数据库，方便同事快速检索。

业务挑战和需求

- **电商导购：**需要用图片拍照的方法，提升导购效率；
- **商品快速搜索：**需要通过向量检索，识别语义匹配；
- **智能客服：**需要客服和呼叫中心，提供不同商品参数的快速检索。

解决方案和场景描述

- **统一数据库方案：**原来 AP 分析业务，从 PG 迁移到了 OceanBase；原来 TP 业务从 TiDB 迁移到 OceanBase；
- **多租户：**原来大量多租户的需求迁移到了 OceanBase；
- **商品搜索向量匹配：**解决各门店系统迁移至九讯云平台时，能够快速匹配不同租户的同一个商品信息，快速实现商品总体管理。
- **RAG 方案：**提供给客服和呼叫中心，商品知识库。

```
graph TD
    A[所有产品图片集] --> B[图片特征提取]
    B --> C[特征向量]
    C -- "录入数据" --> D[(OceanBase 向量数据库 商品 SKU 信息)]
    E[客户下单商品] --> F[店员拍照]
    F -- "APP 等上传图片" --> B
    F --> G[店员录入]
    G -- "获取商品信息" --> D
    D --> H[订单系统]
    D --> I[库存系统]
```

6.5 慧视通：智能登签

公司简介

深圳市慧视通科技有限公司经营范围包括全球定位系统、动态管理系统、电子软件及硬件产品等。

客户收益

- 简化运营：当前需要由出租车公司通知后台监控中心客服人员，将司机信息在后台进行绑定后下发到终端，流程繁琐且耗费人力。
- 自研：需要使用符合自研要求的向量数据量。

业务挑战和需求

- 人脸登签业务：巡游出租车业务，人、证、车绑定，司机运营前需人脸识别，人、证一致后才能运营。需要通过面部识别，做人和司机的登签，简化流程。

解决方案和场景描述

- 向量方案：用人脸识别，自动匹配司机照片，基于 OceanBase 向量数据库的 TOPN 排序相似度计算和标量点查，快速帮助机动司机绑定设备终端，快速上线运营，大幅提升机动司机运营效率。

Before

```
graph LR
    Driver -- Signin --> Device
    Device -- Notify --> CS[Customer Service]
    CS -- Bind --> OceanBase
    OceanBase -- Photo --> OS[Object Storage]
    OS -- Photo --> Cloud
    Cloud -- Photo --> Device
```

After

```
graph LR
    Driver -- Signin --> Device
    Device -- Embedding Search --> Cloud
    Cloud <--> OBV[OceanBase Vector]
```

6.6 三维家：基于 OceanBase 向量的家具智能推荐

公司简介

广东三维家信息科技有限公司是一家专注于大家居产业的领先企业，致力于提供基于云计算的工业软件解决方案服务的各个环节，包括营销、设计、生产、制造管理以及生产装备。公司成立于 2013 年，构建了强大的产品矩阵，包括 3D 云设计、3D 云制造和数控系统等产品，覆盖了家居产业级，推动了家居产业和装备制造业的数字化升级。

客户收益

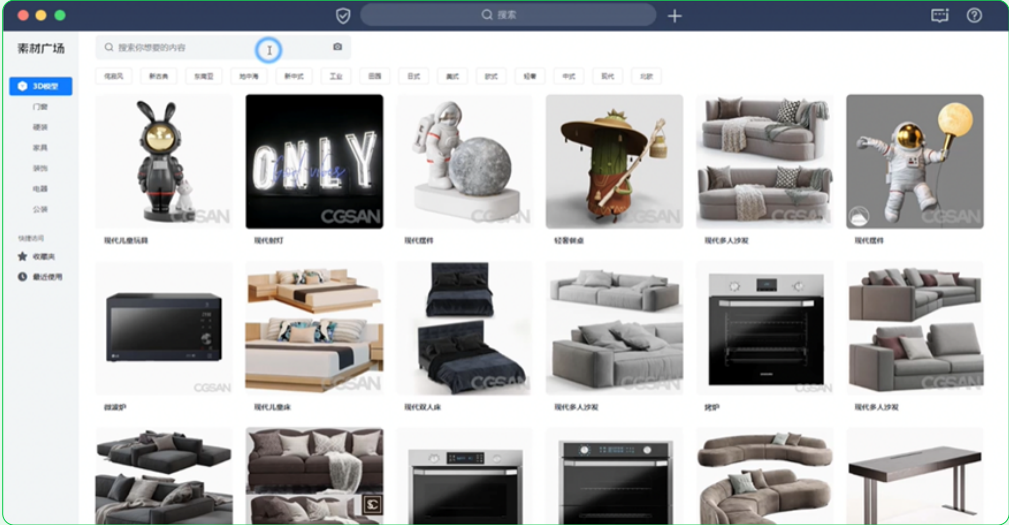
- 提升高可用：解决之前 Milvus 频繁内存爆的问题。
- 提升稳定性：上线后，稳定运行。
- 统一技术栈：Milvus 和 TiDB 基本都迁移到 OceanBase。

业务挑战和需求

- 以图搜图：AI 智能推荐家具，快速查找相似家具模型，轻松替换应用。基于智能搜索技术，具备语义识别、模糊、精准联想用户需求帮助用户全网搜索软硬装素材，快速应用至方案素材广场集成在线编辑功能，用户可轻量建模，修改材质。

解决方案和场景描述

- 向量方案：使用压缩算法 HNSW-SQ，降低内存量到原来 1/4，从而可以单分区支撑业务，无须分库分表。
- 扩展性设计：未来数据量进一步增大时，切换到 IVF-SQ8。



6.7 视源股份：学习机绘本题库

公司简介

教育数字化工具及服务提供商希沃（seewo）、智慧协同平台 MAXHUB 的母公司。视源股份的主营业务为液晶显示主控板卡和交互智能平板等显控教育信息化产品的设计、研发与销售，产品已广泛应用于家电领域、领域、企业服务领域等。

客户收益

- **简化技术栈：**使用融合查询，一个数据库解决标量和向量的查询。
- **提升稳定性：**上线后，稳定运行。

业务挑战和需求

- **智能教育：**搜索知识库场景，每个绘本约 4W 左右问句，每个问句场景小于 20 个字。业务逻辑是将绘本的问题及答案存储于数据库中，基于用户的提问转化为向量的相似性搜索，进而返回找到标准问题及结果。如果没有相似的问题，则返回固定文案。

解决方案和场景描述

- **向量方案：**数据量大概 4000 万，纬度 1024，1000 多绘本，每个绘本 4 万个问题。
- **融合查询：**标量和向量，根据绘本做标量过滤，根据语义做向量查询。未来会增加全文检索。



6.8 当贝投影：观影记录推荐

公司简介

杭州当贝网络科技有限公司是中国知名的智能大屏增值服务提供商之一。公司专注于智能大屏领域，提供软件、硬件和操作系统全生态服务，旗下拥有当贝市场、当贝影视、当贝酷狗音乐等多款优质大屏应用，以及当贝投影、当贝盒子等智能硬件产品。当贝自主研发的当贝 oS 大屏操作系统，已被国内外大屏硬件企业广泛采用。

客户收益

- 统一技术栈降本：解决 RDS+PG-vector 多技术栈问题。
- 简化运维：20+ 实例整合至两个 OceanBase 实例。

业务挑战和需求

- 视频推荐：历史观影记录存储+观影记录推荐。
- 技术栈复杂：数据库种类多、实例多，管理复杂

解决方案和场景描述

- 向量方案：阿里云 PAI + OceanBase 方案，数据按天采集入库并进行加工，生成向量数据。按天新建分区存入OceanBase 向量库。OceanBase-Vector 给推荐业务提供数据召回。

6.9 伯俊科技：知识库

公司简介

伯俊成立于 1999 年，深耕零售行业 26 年。其公司的核心分销系统，OMS，全渠道一盘货等广泛使用在 5000+ 中大型行业客户，属于零售行业的头部 ISV。

客户收益

- **简化技术栈：**伯俊原已经使用腾讯的 VectorDB，采用 Vectordb+mongodb 复杂传统架构。
- **融合检索：**将记忆体，向量数据，文本数据混合存放，提升了大模型的开发效率。另外通过减少多副本的方式极大削减了在库存预测以及专业知识库的搭建的投入。

业务挑战和需求

- **知识库：**随着零售行业的深化应用，伯俊从知识库开始探索建立 RAG 应用，逐步延伸到门店的AI导购，AI配货等。

解决方案和场景描述

- **统一数据库方案：**TP 场景和向量场景统一使用 OceanBase ；
- **融合查询：**对向量、文本数据库的融合查询。

The diagram illustrates a data architecture for a retail company. At the top, six business centers (业务中心1-6) are shown. Centers 1-4 are grouped together, and centers 5-6 are grouped together. Data flows from these centers to a central data processing layer. This layer includes OSS (Object Storage Service) for HTTP read/write, DRDS (Distributed Relational Database Service) for sharding and sub-table design, and MySQL databases. Data is written to DRDS and then to MySQL. A BinLog (Binary Log) is used for data replication, which is then processed by Dump, Canal/DTS, and Kafka. The DataSync tool is used for data synchronization. The final data is stored in Elasticsearch (ES) for search. A reporting center (报表中心) is also shown, which reads data from the ES and writes it to a reporting tool. The diagram also shows a Redis cache layer for query results.

6.10 in银泰商业：零售业智能问数平台

公司简介

银泰集团（全称“中国银泰投资有限公司”）由沈国军先生于 1997 年创立，立足实业发展与产业投资，业务涵盖商业零售、商业地产运营与开发、股权投资等领域，在境内外拥有多家控股、参股公司，已发展成为一个主业突出、多元发展的现代企业集团。

客户收益

- 高可用性：Faiss 只提供基于内存的能力，缺乏高可用性能力。
- 统一技术栈：目前企业使用 Oracle，MySQL，未来有计划统一选择 OceanBase 。解决C端积分场景 TP 数据库的高性能；HTAP 特性加工实时报表等能力。

业务挑战和需求

- 智能问数平台：提供一套通用问数的产品，实现管理层的问题：询问业务数据的真实值；问诊：对店长提供数据分析和原因（内外部数据+报表）。提供门店日常数据查询与经营分析、门店业绩对标分析与经营问诊、集团预算管理与业绩分析等。
- 知识库查询，客户问答：同时对员工提供制度查询，对客户相服务查询。

解决方案和场景描述

- 向量方案：RAG+LLM，实时数据快速分析，AI 智能辅助科学决策；NL2SQL，自然语言交互，秒级结果输出，问数更简单。
- RAG：基于 OceanBase 的知识库方案。

```
graph TD
    subgraph User_Path [用户链路]
        direction TB
        UR[语音识别]
        SA[方案助手]
        AI[AI 客服]
        ET[企业内训]
        PC[潜在客户]
        IC[意向客户]
        SC[签约客户]
    end

    subgraph Business_Path [业务链路]
        direction TB
        IDI[智能问数/智能问诊]
        CPG[客户画像与分级]
        NI[需求整合]
        SAO[方案吸引力]
        CF[关注特征]
        BS[标化说明]
        US[升级方案]
        NPA[需求精准分析]
        KLO[客单及利润优化]
        ML[材料升级]
        LA[关联卖点]
        RP[复购及后]
    end

    subgraph Tech_Support [技术支持]
        direction TB
        DCM[数据清洗/模型建设]
        CC[内容采集]
        DA[数据对接]
        MD[模型设计]
        RA[推荐算法]
        CO[策略优化]
        DB[数据看板]
        OM[运营管理]
        FM[流程管理]
        ND[需求挖掘]
        SAH[售后助手]
        MT[营销工具]
    end

    UR --> ET
    ET --> PC
    PC --> IC
    IC --> SC
    SC --> IDI
    IDI --> CPG
    CPG --> NI
    NI --> SAO
    SAO --> CF
    CF --> BS
    BS --> US
    US --> NPA
    NPA --> KLO
    KLO --> ML
    ML --> LA
    LA --> RP
    RP --> DCM
    DCM --> CC
    CC --> DA
    DA --> MD
    MD --> RA
    RA --> CO
    CO --> DB
    DB --> OM
    OM --> FM
    FM --> ND
    ND --> SAH
    SAH --> MT
```

6.11 某跨境电商品牌：电商网站推荐系统

背景简介

在跨境电商竞争日益激烈的今天，面对海量用户咨询与个性化推荐的需求挑战，消费者体验和运营效率升级已成为跨境电商破局的关键。某全球知名的时尚品牌跨境电商，正通过结合OceanBase向量数据库与前沿AI技术，为其智能客服系统与商品推荐业务注入全新动力，实现服务智能化升级与业务创新。

客户收益

- 基于 OceanBase 知识库开源代码，进行二次开发，一周完成智能商品推荐售前业务，一周完成对接订单，物流等售后客服对接。
- 智能商品推荐可以满足客户问题的快速检索以及同款商品的搜索与导购。

业务挑战和需求

- **智能客服推荐：**之前主要是使用传统客服，遇到相应效率以及客服主观经验问题导致体验较差。
- 使用传统的 Milvus+Mongo 方案，需要学习成本以及资源成本

解决方案和场景描述

- OceanBase 具备文档以及向量多模一体化能力，一套数据库解决数据库底座问题。
- OceanBase 与 Milvus 一样拥有多种的向量索引，HNSW/HNSW_SQ/HNSW_BQ/IVF 等，可以做到性能，准确率，成本的 tradeoff。
- OceanBase 与 Dify, MCP 等 AI 生态结合，更方便接入。

6.12 百宝箱：AI Agent

公司简介

百宝箱通过领先的大模型技术，极大地缩短了应用的开发链路。在充分发挥大语言模型的优势的同时，百宝箱还支持开发者通过引用插件、卡片、知识库、数据库、图片炼丹炉等资源，进一步增强应用的能力，以保证应用在处理用户查询等任务时的回复更加符合预期。

客户收益

- 应用敏捷：无需额外开发数据隔离逻辑，借助原生表隔离功能即可实现用户级安全隔离。
- 结构灵活：支持基于 JSON 灵活扩展和变更字段结构，满足智能体应用的多样化数据需求。
- 简化运维：原先部署于多台实例的上万张小表统一管理，大幅降低运维复杂度与资源消耗。

业务挑战和需求

- 表数量巨大，难以管理：平台需运维超过 10 万张结构高度相似的小表，传统方案带来高运维成本和实例资源浪费。
- 多租户安全隔离要求高：每张表需根据用户维度进行严格的数据隔离，确保多租户间数据安全。
- 结构频繁变更，开发效率受限：平台用户经常需要快速调整表结构，以适配不同的智能体构建需求。

解决方案和场景描述

- 百宝箱采用 OceanBase 提供的 JSON 虚拟表能力，将原本分散在多实例的小表整合为统一结构，替代物理表，实现更高效的数据存储和访问方式。同时，结合原生的表级别用户隔离机制，保障多用户数据安全。通过 JSON 虚拟表 + Text2SQL 技术，开发者能够在智能体中实现结构化数据的快速读写和语义查询。

用户提问

回答

17:24

17:27

请推荐距离 3000 米以内评分 4.5 分以上人均消费 200 元以下的潮汕牛肉火锅店

请推荐距离 3000 米以内评分 4.5 分以上人均消费 200 元以下的潮汕牛肉火锅店

3000 米以内

评分 4.5 分以上

人均消费 200 以内

潮汕牛肉火锅店

空间搜索和计算

标量过滤

向量检索