IDC | 移动云


# AI原生数据库
# 发展趋势白皮书

# 目录

CONTENT

# 核心观点摘要

AI 原生数据库（AI-Native Database）不是“数据库 +AI”的简单叠加，而是数据库内核的系统性重构。未来三年，数据库的智能能力将比存储能力更能决定企业竞争力。本白皮书提出五大关键判断：

- **从“存”向“智”的范式转移：**数据库正从被动存储向主动理解演进，语义理解、相似性推理与跨模态关联成为核心能力，向量数据库已成为 AI 时代不可或缺的关键基础设施。
- **从“外挂”到“智能内核”的架构革命：**AI 能力正深度融入数据库内核，形成“AI for DB”与“DB for AI”的双轮驱动。智能内核模式在数据流转效率、响应延迟和安全性方面全面超越传统外挂架构。
- **产品形态的三极分化与融合：**传统数据库 AI 化、专用向量数据库与 AI-Native Database 三类产品各有适用边界，企业需根据数据类型复杂度与智能需求水平进行精准选型。
- **数据库智能体（DBAgent）开启自治时代：**从辅助决策到自主决策，DBAgent 正覆盖数据库全生命周期管理，事务一致性保障、SQL 语义精确理解和任务级权限控制成为核心技术挑战。
- **国产数据库从技术跟跑到创新领跑：**中国数据库市场国产化率将超 70%，国产厂商通过生态兼容、场景深耕与 AI 原生架构的技术换道，正在全球数据库产业格局中赢得战略主动。

# 前言

人工智能大模型技术正在加速迭代，重构全球数字经济发展格局，并推动各行业从“数字化”向“智能化”深度转型。数据库作为数字基础设施的核心支柱，是激活数据要素价值、支撑 AI 技术落地的关键载体，其产业价值在智能化转型浪潮中愈发凸显。

IDC 数据显示，到 2029 年，中国数据库管理系统软件市场规模将达到 186 亿美元，从 2024 年到 2029 年的年均复合增长率将达到 20.1%，在全球所有国家和区域中，中国市场增速第一；其中，来自公有云的收入将超过 60%。

数据库不仅承担着数据存储、管理与运算的基础使命，更是衔接数据资源与 AI 应用的关键纽带。目前，全球主要数据库厂商和云服务商已纷纷布局“数据库 +AI”融合战略，推动数据基础设施向全面智能化方向演进。

目前，国内主流的云和数据库厂商都在积极推动 AI 与数据库的深度融合，引领技术变革与市场拓展新趋势。大模型技术正在被全方位融入数据库内核，构建起智能索引、故障自愈及自然语言交互查询等创新能力，大幅提升数据库自治水平。同时，多模态数据处理成为主流厂商共同关注的焦点，相关的数据库产品积极探索向量检索与全文检索的融合技术，满足 AI 应用对多样化数据的高效处理需求。在市场布局方面，头部厂商凭借全栈技术优势与生态整合能力，正努力在金融、政务、工业、能源等核心领域开展场景实践。在这个过程中，云原生体系也进一步与 AI 形成了良性协同，通过存算分离、弹性伸缩以及强大的资源调度能力，为 AI 时代的用户带来显著的降本增效效应。

本白皮书全方位聚焦 AI 大模型背景下数据库领域的核心变革与发展趋势，系统梳理业界数据库与 AI 融合发展的主要方向和最新进展，厘清行业发展痛点与机遇，帮助企业在 AI 时代全面夯实数据根基，构建出面向未来的智能化数据基础设施体系。本白皮书提出的五大趋势，围绕“AI-Native Database（AI 原生数据库）”这一创新产品体系，构成一个逐层递进、相互支撑的发展和演进逻辑，有助于用户把握相关技术、产品和市场发展的全貌：

- **趋势一，围绕架构重构，是创新的坚实基础：**向量数据库、多模融合、混合检索等技术升级构成了 AI 原生数据库的底层能力基座。没有架构层面的根本性重构，智能化升级将成为无源之水。

- **趋势二，围绕智能内核，是创新的重要核心：**在架构重构的基础上，AI 能力从外挂走向内嵌，数据库获得自主学习、自我优化、智能决策的“大脑”。这是从“有智能”到“是智能”的质变。

- **趋势三，围绕产品形态，是创新的关键体现：**架构与内核的升级最终体现在产品形态上，出现传统数据库 AI 化、专用向量库、AI 原生数据库三极分化，为不同场景提供精准适配的解决方案。

- **趋势四，围绕数据库智能体，是创新过程的载体：**数据库智能体（DB Agent）作为人与数据库系统之间的智能接口，将复杂的运维、开发、治理工作封装为自然语言交互，大幅降低使用门槛。

- **趋势五，围绕产业格局，是创新的体系化成果：**前四个趋势的技术变革，最终重塑全球数据库产业格局，为国产数据库提供“换道超车”的历史性机遇。

趋势一 01

# 由“存”向“智”数据库架构全面重构

数据库技术由“存“向“智”发展，不是简单的功能增强，而是数据库核心能力范式的根本性转变，这一转变主要体现为三个层面的能力跃迁：

- **从“被动存储”到“主动理解”：**传统数据库可视为数据的“仓库管理员”，只负责存取，不理解内容。AI 原生数据库则具备语义理解能力，能够自动识别数据内涵、发现隐含关联、预测数据趋势，成为数据的“智能分析师”。
- **从“精确匹配”到“相似性推理”：**传统检索过程基于精确的关键词匹配，无法区分同一个词在不同语境下的差异。向量检索技术使数据库能够理解语义相似性，实现“以图搜图”和“以意搜文”，突破精确匹配的局限。
- **从“单一模态”到“跨模态关联”：**传统数据库根据数据类型区分，将结构化数据、文档、图像等彼此隔离。多模态融合技术打破了数据库体系的分立状态，实现多类型数据的统一存储与关联分析，让 AI 系统获得了“全景视野”。

## 1. 向量数据库成为 AI 时代的关键基础设施

数据已经成为驱动 AI 进化的核心要素。随着 AI 应用场景的广泛拓展，文本、图像、音视频等非结构化数据呈现出爆炸式增长态势。如何高效存储、检索并利用这些数据，成为影响 AI 发展的关键性举措。在此背景下，向量数据库凭借其独特的技术特性和适配优势，正快速成为 AI 时代不可或缺的关键基础设施，为各类智能化系统的运行提供坚实支撑。

根据 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。IDC 认为进入大模型时代，向量数据库专注于存储和管理向量数据，适合语义搜索或者相

似性匹配的场景，LLM+ 向量数据库提供了非结构化数据的语义理解和检索能力，是大数据平台的补充。大模型也促进了大数据平台的智能化发展，实现了智能数据查询、数据治理等功能。

## 向量数据库是 AI 场景下非结构化数据处理的必然选择

海量非结构化数据蕴含着丰富的语义信息，是智能决策与内容生成的重要依据。在传统 IT 技术体系下，这些数据形式多样、结构复杂，难以直接被数据库系统进行处理。向量数据库通过将非结构化数据实施向量嵌入过程，将数据映射到多维向量空间，使得非结构化数据在向量空间中根据语义、语法、上下文等因素产生关联，这种映射打破了数据类型的限制，为非结构化数据的统一处理提供了基础。

- **向量数据库助力 AI 系统实现毫秒级的语义检索：**系统将用户的查询请求转换为向量，并在向量空间中快速匹配相似数据，返回最相关的结果。这种检索方式不仅速度极快，而且能够准确理解用户的语义意图，大大提高了检索的准确性和效率。例如，在智能客服场景中，向量数据库被广泛使用，实现从海量知识库中检索与用户问题最为匹配的答案，显著提升用户体验。

- **向量数据库有力解决大模型“幻觉”与实时数据调用难题：**利用向量数据库快速检索真实世界的数据，可以为大模型提供更为准确的参考信息，减少大模型自身的“幻觉”发生。在 RAG（检索增强生成）架构中，向量数据库作为底层依赖，支持实时数据更新和检索，能够及时将最新的数据反馈给大模型，确保生成内容的时效性和准确性，为大模型提供了丰富的外部知识支持，显著提升生成内容的质量和可靠性。

## 从专用向量库向混合检索引擎持续发展

向量数据库正在经历从“专用向量库”到“混合检索引擎”的重要技术突破。事实上，早期的专用向量库主要专注于向量数据的存储和检索，其功能相对单一。而随着 AI 应用场景的日益复杂化，相关的智能化系统对数据库的功能提出了更高的要求，进而推动了混合检索引擎模式的发展。

- **复杂类型信息的综合处理能力与现代 AI 系统的效能高度关联：**传统 AI 模型往往孤立处理单一模态数据，导致信息碎片化。而现代 AI 将不同类型信息映射至统一语义空间，实现跨模态关联与融合，使 AI 能模拟人类多维度推理过程，显著增强决策合理性、响应速度及场景适应性，成为推动 AI 从“单点智能”向“通用智能”跃迁的核心动力。

- **混合检索引擎大幅提升多模态环境下的数据检索效率：**将向量检索与全文检索、标量检索、图检索等多种检索方式相结合，通过整合向量检索的语义理解能力与关键词检索的精确匹配优势，在多模态数据环境中构建出高效检索范式。例如，向量检索能够解析非结构化数据（如文本、图像、音频）的深层语义关联，突破传统关键词匹配的表面限制。而关键词检索则可快速筛选结构化元数据，缩小搜索范围，标量检索适用于数值型数据的精确查询，图检索则可以挖掘数据之间的关联关系。这些检索方式的有机融合，既保证了跨模态语义匹配的全面性，又通过精确过滤提升了检索速度，显著提升多模态数据检索的效率与准确性。

例如，在 AI 赋能安全生产的场景中，系统需要同时处理视频、图像、文本等多种类型的数据。混合检索引擎可以将视频帧转换为向量进行语义检索，同时利用全文检索处理各类现场监控日志中的文本信息，通过标量检索筛选特定时间段内的数据，并借助图检索分析人员之间的关联关系。这种多模态、综合性的检索方式，能够大大提高安监、安防系统的效率和准确性，为企业安全生产提供有力保障。

## 2. 打破数据孤岛，实现多模态数据融合处理

AI 时代的数据在呈现爆炸式增长的同时，其形态也日益趋向复杂和多样化，可涵盖结构化的关系型数据、半结构化的文档数据、时序数据，以及非结构化的图像、音频、视频等多模态数据。

传统的关系型数据库、文档数据库、时序数据库、图数据库和向量数据库等多专注于单一数据模型处理，在多模态数据时代面临新的、更为严重的数据孤岛现象，进而极大地限制了数据的价值挖掘和多场景应用。因此，向量数据库的多模融合趋势，有利于企业数据体系向一体化多模引擎方向发展，旨在通过同一内核实现结构化、非结构化以及多模态数据的统一存储、查询与事务处理，打通解决数据孤岛问题的关键路径。

### 单一模型数据库在 AI 时代将产生新的数据孤岛困境

传统单一模型数据库针对特定类型数据的特点和需求进行优化。例如，关系型数据库擅长处理结构化的业务数据，通过严格的模式定义和事务处理机制保障数据的一致性和完整性；文档数据库更适合存储和查询半结构化的文本数据，具有较高的灵活性和可扩展性；时序数据库专注于处理带有时间戳的数据，在物联网、金融等领域的时间序列数据存储和分析中表现出色；图数据库适用于社交网络、知识图谱等具有复杂关系的数据处理。

然而，企业 AI 场景下的数据关系日趋复杂，不同数据库之间的数据无法直接共享和交互，导致企业难以对全量数据进行综合分析和挖掘，无法形成全面的业务洞察，从而影响 AI 技术体系下智能决策的科学性和精准性。例如，在 AI 质控领域，视频监测信息、生产日志和物料信息等多种类型的数据需要被融合处理和分析，从而以更加全面的视角评估和衡量质量态势和改进方向，单一模型数据库无法满足这种多模态数据的快速融合处理需求，严重制约相关流程的智能化转型。

## 一体化多模引擎：向量数据库多模融合的核心架构

为了避免 AI 时代新的数据孤岛现象，向量数据库正向一体化多模引擎的技术方向快速演进，旨在将多种数据模型集成在同一内核中，利用统一的存储引擎、查询引擎和事务处理机制，实现对结构化、非结构化和多模态数据的统一管理。

- **在存储层面：**一体化多模引擎融合多种数据存储结构，可根据不同类型数据的特点适配合适的存储方式。例如，对于结构化数据，可以采用传统的行式或列式存储；对于非结构化数据，如图像、音频等，可以将其元数据存储在关系型表格中，而实际数据则存储在对象存储系统中，并通过指针进行关联。

- **在查询层面：**一体化多模引擎需要支持多种查询语言，并在用户层面提供统一的查询接口，实现对不同类型数据的查询。例如在同一项 AI 任务中，通过 SQL 语句查询的结构化数据能够与通过向量相似性搜索查询的非结构化图像数据进行结合，形成多模态数据的联合查询效果。

- **在事务处理层面：**一体化多模引擎须保障不同类型数据在操作中的一致性和原子性，无论是结构化数据的更新、非结构化数据的插入，还是多模态数据的关联操作，都能够在一个事务中完成。在操作过程中以严格的机制保障不同模态数据操作的协同与可靠，在涉及文本、图像、结构化数据等多类型数据的并发修改或查询时，引擎应通过统一的事务调度，确保所有操作全部成功并持久化，或全部回滚至初始状态。

## 多模融合将在 AI 场景中产生出巨大价值

数据库的多模融合趋势具有重要的产业价值。其在打破数据孤岛的同时，通过全类型数据的统一处理，为企业提供了更全面、准确的数据视图，同时从用户层面简化了数据管理的复杂度，降低了企业的数据管理成本，提高了数据处理的效率。

随着 AI 场景的不断深入和 AI 数据类型的日趋丰富多样，各类企业 AI 技术体系对多模融合的需求将日益迫切。数据库的一体化多模引擎将持续获得完善和优化，实现更强大的数据综合处理能力和更高的协同性能，满足复杂业务场景的需求。同时，多模融合技术还有望与区块链、隐私计算等技术相结合，保障数据的安全性和隐私性，为数字产业的发展提供更稳固的基础保障。

## 3. AI 助力数据库复杂检索能力持续升级

复杂检索能力的升级是 AI 时代数据库架构重构过程中的关键环节。快速发展的 AI 应用对数据库检索效能提出了前所未有的新挑战，AI 应用普遍呈现出实时性和交互性特征，例如在智能客服、实时推荐等电子商务场景中，数据库既需要快速响应查询请求，提供低延迟的数据检索服务，还需要完成 AI 任务中对文本、图像、音频等多种类型数据的处理和理解，通过对多模态数据的综合检索与分析，从更多维度理解客户意图，响应客户的真实诉求。

- **全文检索能力的深化与推广：**为更准确、高效地处理 AI 时代的文本数据，数据库通常在全文检索方面进行了深化设计。传统的全文检索主要基于关键词匹配，而新的数据库有效引入了自然语言处理（NLP）技术，通过对文本进行分词、词性标注、命名实体识别等处理，构建语义索引，使得数据库能够根据用户的自然语言查询，准确返回相关的文本结果。例如，在智能问答系统中，用户使用自然语言提出问题，数据库体系能够进一步理解问题的语义，从海量文档中检索出最匹配的答案，全面提升查询的准确性和用户体验。

- **混合检索引擎的崛起：**混合检索引擎则是 AI 时代数据库检索能力进一步跃升的重要方向。它深度融合向量检索的语义理解能力与关键词检索、全文检索的精准定位优势，通过多模态数据统一表征与跨模态相似度计算，突破了传统检索在复杂语义和异构数据上的局限。基于 AI 模型动态优化检索策略，混合检索引擎能够实现意图理解、查询扩展与结果排序的智能化，并在大模型交互中进一步强化上下文感知与逻辑推理能力，使检索结果更贴近用户真实需求。未来，混合检索将在智能问答、内容推荐、科学决策等 AI 场景下持续发挥核心作用，成为 AI 驱动的信息服务基础设施的关键组成部分。

# 趋势二 02

# 数据库 AI 原生，数据库正从“外挂”演变成新时代“智能内核”

IDC 预测，到 2027 年，70% 的 IT 团队将开始关注数据的流通质量、数据的治理，以及打造一个 AI 就绪的数据基础设施平台。

在技术架构全面重构的趋势下，数据库领域正在经历一场深刻且具有革命性的变革。数据库相对于 IT 架构的传统“外挂”模式正在逐步演进为 AI 原生驱动的“智能内核”形态。这一转变不仅重塑了数据库的功能边界与应用场景，更成为推动各行业智能化升级的关键力量。

- **传统“外挂”模式的效率掣肘：**传统数据库聚焦于数据的存储、查询与管理，高度依赖预设规则与算法。在 AI 与数据库融合的早期实践中，通常通过在数据库外部搭建 AI 模型和工具来处理与分析数据。这种这种“外挂”模式模式虽在一定程度上实现了数据与智能算法的初步关联，但存在数据流转效率低下、模型与数据普遍割裂、开发运维复杂等固有局限，既难以充分释放 AI 在数据处理中的巨大潜能，也严重制约了数据赋能 AI 的应用路径。

- **“智能内核” 显现强大内生效应：**AI 原生数据库的重要发展趋势在于将 AI 能力深度融入数据库核心架构，使数据库具备自主学习、自我优化和智能决策的原生能力。这一根本性转变意味着数据库不再仅是数据的存储容器，而是进化为能够理解数据内涵、挖掘数据价值、为业务提供全方位智能赋能的核心中枢。通过内置机器学习算法，数据库可自动识别数据模式、预测趋势变化，并基于实时数据反馈动态调整查询与优化策略，显著提升数据处理的效率与准确性。

表 1 外挂模式 vs 智能内核模式对比

| 对比维度 | 外挂模式（传统架构） | 智能内核模式（AI-Native） |
| --- | --- | --- |
| **架构位置** | AI 模块部署在数据库外部，<br>通过 API 调用 | AI 引擎深度嵌入数据库内核，<br>与存储 / 查询引擎协同 |
| **数据流转** | 数据需在数据库与外部 AI 系统间<br>频繁搬运，产生 I/O 瓶颈 | 数据不动模型动，<br>计算下推至数据所在位置 |
| **响应延迟** | 网络传输 + 序列化开销，<br>通常为百毫秒至秒级 | 内存级计算，<br>延迟降至毫秒级甚至微秒级 |
| **开发复杂度** | 需维护多系统连接、数据同步、<br>版本兼容，运维负担重 | 统一 SQL 接口，标准化 AI 调用语法，<br>大幅降低开发门槛 |
| **安全性** | 数据出域风险高，<br>需额外加密传输和访问控制 | 数据不出域即可完成 AI 计算，<br>满足金融 / 政务合规要求 |
| **自进化能力** | 依赖人工调优，<br>模型与数据迭代不同步 | 强化学习驱动，<br>基于实时数据反馈自动优化策略 |

目前，主流云厂商已普遍关注到这一技术趋势，纷纷加大在数据库 AI 原生领域的投入与布局，并从两个关键维度推动数据库的智能化升级。

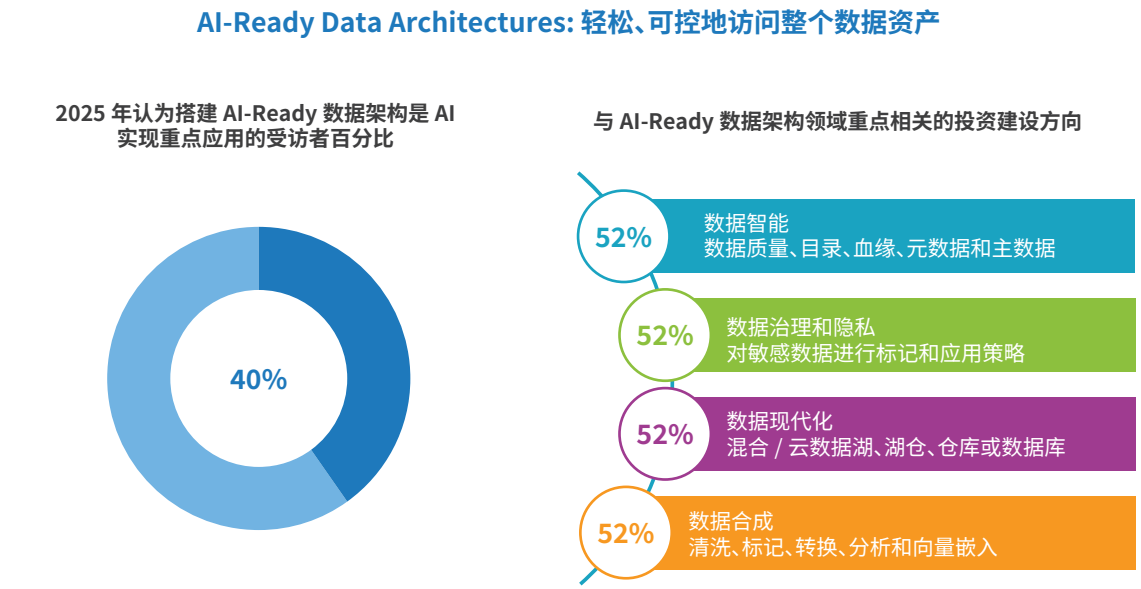
- **多维度打造数据库内生 AI 能力：**从数据存储、索引构建到查询执行等各个环节，全面引入智能算法与模型，使数据库自主适应不同的业务场景和数据特征。例如，通过内置 AI 算法优化数据存储结构，实现自动分区和索引优化，提升查询效率。利用 AI 进行数据质量监控与异常检测，保障数据可靠性。在查询处理方面集成自然语言处理技术，让用户能以自然语言直接查询数据库，降低使用门槛，并借助机器学习模型优化查询计划，加速复杂查询的执行。此外，还可引入 AI 强化数据库的自动化运维，实现智能资源调度、故障预测与自愈等，为用户提供更高效、智能的数据服务。

- **构建数据和模型的一站式应用开发范式：**将 AI 服务平台的能力深度融合至数据库的原生设计中，推动数据库与机器学习平台和深度学习框架的无缝连接，为用户提供从数据准备、模型训练到 AI 应用部署的全流程服务。用户无需在不同系统之间来回切换，即可在统一的平台上完成数据的探索、特征工程、模型构建与评估等操作，大大降低了 AI 应用开发的门槛与成本。同时，这种深度融合还促进了数据与模型的高效协同，使得模型能够实时获取最新的数据进行训练与更新，从而保证模型的准确性与时效性。

“外挂”到“智能内核”的演进，是技术发展的必然路径，也是行业数字化转型的迫切需求。云厂商基于对数据库技术和市场趋势的认知，为 AI 原生数据库的发展开辟了新的道路。随着 AI 应用场景的持续拓展，AI 原生数据库将在越来越多的领域发挥关键作用，助力企业实现数据驱动的智能化决策与业务创新，并推动整个社会向智能时代加速迈进。

IDC 的研究显示，2025 年，有 40% 的企业认为搭建 AI-Ready 数据架构是 AI 实现重点应用的前提，具体投资建设方向包括数据智能、数据治理和隐私保护、数据现代化、数据合成以及向量 /RAG 管理。以终端应用为驱动的 AI 及 Agent 技术，让用户重新审视大数据底座，希望解决动态数据管理、分布式存储、数据质量、多模态等问题。

图 1 AI-Ready 数据架构投资建设方向




来源：IDC 2026

目前，AI 原生数据库的发展进程中已经形成了“AI for DB”与“DB for AI”的双轮驱动模式，前者关注 AI 赋能所带来的功能强化和性能跃升，后者则重点体现数据库面对大模型需求时所产生的革命性变革。双轮驱动模式有利于 AI 原生数据库在快速发展的 GenAI 时代获得充足的成长动力，并成为未来数据库发展的核心形态之一。

## 1. AI for DB：AI 重塑数据库管理新范式

传统数据库的管理和运维工作高度依赖人工，例如在大型数据库实践中，普遍需要专业的数据库管理员（DBA）实施监控、故障排查和性能优化等工作。在 AI 时代，数据量的爆炸式增长和业务场景的日益复杂，使以人工为主的数据库管理工作面临越来越多的能力和效率挑战。AI for DB 为解决上述问题提供了全新的思路和方法，即通过高度智能化的自治管理和运维，改变数据库的规模化管理困境，并切实解决一系列实践问题：

- **智能运维与自治：**由 AI 驱动数据库实现自诊断、自优化、自修复等一系列高度自动化的运维与自治能力。例如通过系统日志、性能指标等数据，精准定位磁盘故障、内存泄漏、连接异常等问题，以及自动调整数据库的配置参数以实现面向应用的最优性能表现等。特别是 AI 加持的慢查询自愈能力，能够自动分析数据库应用中慢查询出现的特征和原因，适时进行自动化的调整与优化，快速消除慢查询对性能的影响。

- **智能调度与索引优化：**AI 有效预测未来业务负载的变化情况，提前调整计算与存储资源的分配策略，避免资源浪费或不足的情况发生。同时，AI 根据用户查询模式和数据分布情况，自动推荐合适的索引创建方以及调整缓存策略，帮助数据库在保证查询性能的同时，减少索引维护的开销，这对于提高数据库查询性能至关重要。

- **多模态数据融合分析：**AI 能够有效打破数据形态之间的壁垒，将文本、图片、视频等多样化的数据纳入到面向业务目标的统一分析过程中，通过跨模态语义理解，将不同的数据转化为可计算的向量形态，使不同类型的数据能够相互补充和验证。这一能力不仅使数据库支持更复杂的业务分析场景，还能挖掘出更多的数据价值反哺企业 AI 的性能和泛化能力，构建面向未来发展的竞争优势。

- **面向业务的自然语言交互：**通过自然语言处理技术，将用户输入的口语化指令精准解析为可执行的 SQL 语句，并自动完成查询以及返回查询结果。这一提升使得非技术背景的用户也能够轻松地利用自然语言进行数据查询、访问和分析，促进数据的广泛共享和应用，持续降低 AI 应用门槛，加速 AI 体系在业务中的快速落地。

## 2. DB for AI：提升数据使用效能，深度赋能 AI 应用

大模型正在自然语言处理、图像识别、多模态信息处理等领域推动巨大变革。为更好满足大模型训练和应用需要，显著提升与大模型相关的数据处理、分析与应用水平，目前已出现多种数据库与大模型深度集成的技术方案，旨在通过 DB for AI 实现数据库能力在大模型应用链路中的内置，为解决效率和成本等问题提供创新思路。

- **向量数据库融合：**通过将大模型的语义理解能力与向量数据库的高效检索能力相结合，显著提升知识检索、推理和生成的效率。例如将非结构化文本转换为高维向量以及将图像、音频等非文本数据向量化，实现跨模态检索。同时，利用向量数据实现检索增强生成（RAG），提升大模型输出的严谨性，避免幻觉现象等。

- **提升面向 AI 的数据治理成效：**聚焦 AI 训练的关键环节，利用 AI 强化数据治理、标签管理、数据补齐等过程的效能。例如自动去除数据噪声与异常值，以及自动填补缺失数据内容等，以保证数据的完整性。此外，企业还可以借助知识图谱等技术，深度剖析数据关联，消除冗余信息，纠正分布偏移，从而为 AI 模型训练提供高质量的数据。

- **实现 In-database 推理：**在数据库内部直接执行机器学习或深度学习模型，实现特定的推理过程，而无需将数据导出至外部系统处理。这是模型能力内置的重要体现，即通过将 AI 能力内嵌至数据库内核，使数据存储、查询与推理流程无缝集成，进一步提升效率、降低延迟并保障数据安全。

在未来，随着 AI 技术的持续进步和数据库在 AI 应用领域的广泛拓展，这种双轮驱动模式将获得长期的加速动力。数据库将进一步强化智能、高效和易用特性，满足行业 AI 场景对数据管理和分析的需求。同时，AI 与数据库的深度融合也将使 AI 创新能力变得更加普惠，为数字经济的发展注入全新的动力。

# 趋势三 03

# 从数据基础设施到智能中枢
# 数据库产品形态获得全面升级

面对飞速发展的 GenAI 技术，传统数据库产品体系在应对海量、多元、实时的数据处理需求时存在诸多掣肘，数据库产品形态的全面升级成为必然趋势。

事实上，当前的数据库产品发展已经显现出较为显著的形态变化。一方面，传统数据库产品通过引入 AI 技术进行改造升级，以适应新的数据处理需求；另一方面，专用向量数据库快速进化，积极适配 AI 模型训练和推理任务，成为新的焦点。其总体思想是通过插件或内核升级等过渡性策略，快速补齐向量处理和自治能力的短板，快速适应 AI 时代的新目标。

IDC 关于中国数据库管理软件技术发展的研究显示，未来的数据库产品一方面会继续强化分布式、云原生等技术特性，另一方面则会进一步将 AI、多模态融合以及共享存储集群等技术融入到体系化发展进程中。

图 2 中国数据库管理软件技术发展趋势


来源：IDC 2026

同时，新一代 AI 原生数据库也以其全新的设计理念和强大的 AI 能力，推动数据库产品的革命性变革。其重构了数据库的技术形态，将向量引擎、自治 AI 和多模能力整体融入至数据库的应用流程中，能够更好地满足海量 AI 应用对数据处理的多样化需求，为 AI 技术的深入发展提供更坚实的底层支撑。

## AI 时代数据库产品适配矩阵

面对传统数据库 AI 化、专用向量数据库、AI-Native Database 三类产品形态，企业应如何基于自身业务需求做出精准选择？以下适配矩阵以“数据类型复杂度”和“智能需求水平”为双维度分析框架，为不同应用场景实践提供科学指引：

表 2 AI 时代数据库产品选型矩阵

| 场景特征 | 推荐选择 | 选型理由 |
| --- | --- | --- |
| 以结构化数据为主，<br>智能需求低（如 ERP、财务系统） | 传统数据库 AI 化 | 充分利用现有投资，<br>通过插件快速补齐 AI 能力，迁移成本低 |
| 海量非结构化数据，极致检索性能<br>（如推荐系统、图像搜索） | 专用向量数据库 | 针对高维向量检索深度优化，<br>延迟极低、吞吐极高 |
| 多模态数据融合，高智能需求<br>（如 RAG 应用、智能客服、知识库） | AI-Native Database | 一体化架构消除数据搬运，库内训推保障<br>安全，混合检索满足复杂查询 |
| 关键业务系统，高合规要求<br>（如金融核心、政务系统） | AI-Native Database | 数据不出域即可完成 AI 计算，任务级权<br>限控制 Agent 行为，满足合规要求 |

建议：对于 AI 战略明确、数据资产丰富的企业，AI-Native Database 已经成为其重要的选项。虽然该方向的初始投入相对较高，但其“模型驱动”的架构设计能够从根本上消除数据安全风险，同时，库内训推一体化能力也能够持续降低 AI 应用的边际成本。从长期来看，该方案具有更高的投资回报率。而对于 AI 需求尚处于探索阶段的客户，可先通过传统数据库的 AI 化方案快速验证场景价值，再逐步向 AI-Native 架构迁移。

## 1. 传统数据库 AI 化：插件与内核升级补齐短板

传统数据库通过技术融合与功能升级，在保留原有优势的基础上，通过功能扩展、智能升级和生态协作，快速补齐向量计算、自治管理等关键能力，构建起支持 AI 场景的“智慧底座”，在保留高可靠性、事务一致性等核心优势的基础上，满足 AI 时代对实时性、灵活性和智能化的需求，使数据库从单纯的数据存储和管理工具升级为智能数据协作平台。

### 通过集成向量计算能力实现基于语义理解的混合查询

在 AI 应用场景中，需要突破传统数据库的精确匹配查询模式，通过图像识别、自然语言处理等过程理解语义，例如通过用户输入的图片查找相似商品等。因此，传统数据库通过插件扩展或内核升级等方式，引入向量计算能力，能够一定程度上实现基于语义理解的查询和分析：

- **强化向量存储与搜索能力：**数据库能够将图片、语音、文本等非结构化数据转换为向量，并构建专用索引，使系统能够通过向量相似度等条件检索数据，突破传统数据库精确匹配的限制，实现基于相近语义的图片或文档检索。
- **增加混合查询模式：**将向量计算与结构化查询相结合，从而实现对检索目标的多维度筛选。例如广泛使用的电商平台商品检索过程，可同时根据结构化的商品参数信息和非结构化的向 量数据（图片风格等）检索和推荐商品，提升搜索匹配度和用户体验。
- **系统深度集成：**通过一体化集成改造，使传统数据库能够为 AI 应用提供本地化的非结构化数据处理能力，避免数据在数据库与外部 AI 系统间频繁传输，在提升效率的同时降低隐私泄露风险。

### 引入自治能力，从被动维护转变为主动管理

传统数据库的手动配置参数过程高度依赖专业人员，调整周期长，响应速度慢。AI 场景下的工作负载特性更加多变，例如突发流量、访问类型变化等都会对数据库服务质量产生巨大的挑战，人工调整模式难以适应实时响应的要求。因此，数据库引入自治能力，通过机器学习实现自适应管理，使数据库从被动维护转变为主动管理，显著降低企业 AI 应用的运维门槛，是应对 AI 发展的关键举措：

- **智能资源调度：**数据库系统根据实时监测到的负载变化情况，动态调整服务资源，甚至可以通过学习历史负载特性，对未来需求进行一定程度的预测，提前规划资源状态，实现系统能效的最大化。
- **故障自愈与访问优化：**当数据库系统检测到硬件故障或性能瓶颈时，自主决策自愈策略和流程，例如切换至备用节点，修复数据错误，优化查询逻辑等，大幅缩短故障恢复时间。
- **自主安全机制：**在日常运行过程中，自主识别用户异常操作，自动触发安全防护机制。同时通过分析用户访问行为，对全局安全策略提出调整建议，降低突发安全事件造成的影响。

### 与 AI 应用形成闭环协作

为进一步体现传统数据库对于 AI 应用的价值，还需要简化其在 AI 开发、应用过程中与 AI 模型的协作模式，实现“低代码”甚至“0 代码”集成，从而缩短 AI 模型在实际场景中的落地周期。目前，传统数据库产品可以通过多种方式简化 AI 应用的流程：

- **在数据库中内置 AI 模块：**用户可以直接通过 SQL 语句实现对特定函数的调用，包括预训练模型等，实现内容分类、数据预测等功能，而无需另行编写代码，从而帮助更多非数据库专业人员发掘数据库的专业能力。
- **利用自然语言调用数据库：**即通过自然语言转 SQL 功能，使数据库系统支持用户使用日常语言进行调用，自动生成 SQL 并返回结果。
- **打通数据与大模型的工作闭环：**实现数据库与特定 AI 框架的深度对接，使大模型训练数据可直接从数据库中读取，打通“数据 - 模型 - 应用”的完整链路，避免数据搬移和转换过程中的损耗。

## 2. 专用向量数据库：以极致性能服务大规模模型训练

除传统数据库外，专用向量数据库的产品形态也正在经历从单一工具到智能数据中枢的蜕变。其在索引技术迭代、分布式架构优化以及与 AI 生态的深度融合方面，不仅满足了 AI 对高维数据管理的核心需求，更成为推动大模型训练、实时推理及多模态应用落地的关键基础设施。未来，随着 AI 向边缘计算、联邦学习等场景延伸，向量数据库将进一步向轻量化、增强隐私保护等方向演进，持续拓展人工智能的技术边界与应用深度。

## 索引性能的极致突破：从算法优化到软硬协同

向量数据库的核心价值在于快速检索高维向量数据，例如图像、文本的数字特征等，而这一能力的进化正从算法层面向系统级优化延伸，并直接决定了 AI 应用的响应速度与用户体验。早期产品受限于技术架构，面对海量数据时查询延迟高、稳定性差，难以满足实时场景需求。当前，领先产品通过软硬件协同优化，正在实现性能上的快速跃迁：

- **硬件与算法协同加速：**通过集成 TPU/GPU 等专用计算单元，利用并行计算能力加速向量计算，结合智能索引算法和内存计算优化等手段，使复杂查询的响应时间从秒级压缩至毫秒级，为电商搜推等场景的体验带来革命性提升。
- **服务资源动态优化：**根据负载请求的实时监测情况，利用自主策略自动调整计算资源的分配，在百万、千万级用户并发查询时，灵活启动弹性扩缩容机制，持续保障查询的成功率。
- **系统层面的参数平衡：**例如针对具体的访问场景，实时调整召回率等参数，平衡数据库访问的精准度与速度，并进一步让企业在成本与用户体验之间做出灵活取舍。

## 海量数据与低延时的双重挑战：分布式架构的进化

随着 AI 模型参数规模突破万亿级，训练数据量呈指数级增长，向量数据库的存储与处理能力成为关键瓶颈。专用向量数据库通过分布式架构的持续迭代，在扩展性与延迟间寻找平衡点：

- **去中心化索引架构：**即改变传统方案的中心化元数据管理模式，将数据自动分片存储于多个节点，并利用分布式哈希表（DHT）实现索引分片的自主路由，使集群规模可线性扩展至千节点以上。
- **冷热数据智能分层：**将高频访问的“热数据”存储在高速介质，低频“冷数据”迁移至低成本存储，平衡性能与成本。
- **实时数据流处理：**针对实时训练场景，将数据库任务与消息队列等技术进行集成，实现数据边读取边索引，满足工业流水线质检、自动驾驶等低时延场景下的实时检索需求。

### 从数据存储工具到加速引擎：服务规模化模型训练和应用

目前，专用向量数据库已经开始深度融入 AI 开发全流程，成为 AI 开发流水线的核心组件。其通过与训练框架、推理引擎以及 AI 应用的深度集成，直接加速大模型开发周期，承担起连接数据、模型与应用的纽带作用：

- **无缝对接训练框架：**例如提供 PyTorch、TensorFlow 等主流框架的插件，自动处理数据分片、预取与缓存，使开发者可像使用本地文件一样调用向量数据库，无需额外适配。
- **分布式训练协同优化：**例如与大模型框架以及相关的通信库配合，自动处理数据分片、梯度同步等复杂任务，动态调整数据加载策略，避免因数据倾斜导致的训练中断。
- **推理场景深度优化：**针对实时推理需求，内置模型量化、压缩工具，在保持精度的同时减少向量维度，降低计算延迟，满足高频交易、实时推送等场景的需求。

## 3.AI 原生数据库：整合多元能力，推动 AI 应用范式变革

在改造、强化、升级已有数据库技术和流程的同时，为彻底解决传统数据库在 AI 场景下的性能瓶颈与功能割裂问题，数据基础设施的技术边界与应用价值得到重新定义，AI 原生数据库应运而生，其全面突破传统架构的桎梏，通过系统性设计将内置向量引擎、自治 AI、多模能力等创新要素有机融入数据库的体系化能力中，形成从数据存储到智能决策的完整闭环。

### 整体性设计：从 AI 适配到 AI 原生

彻底改变传统数据库 AI 化改造中所采用的“外挂式”设计，以端到端的整体设计和优化思想，将 AI 能力全方位融入数据库内核，使数据库具备高水平的智能引擎和平台服务能力：

- **支撑 AI 原生的数据结构：**AI 原生数据库将直接支持向量、图、文本等非结构化数据的高效存储与索引，彻底消除数据形态转换的开销，内置的向量计算引擎可无缝处理图像、语音、文本等数据，使 AI 模型训练直接基于原始数据训练，避免过程开销带来的效率损失。

- **动态进化的查询策略：**改变传统数据库依赖静态统计信息进行优化的方式，通过强化学习动态调整执行策略。例如在复杂的 AI 分析场景，AI 原生数据库系统可自动识别数据分布特征，选择最优的索引策略与计算路径，使查询效率成倍提升。

- **资源全局智能调度：**AI 原生数据库可通过内置的 AI 调度引擎，实时分析工作负载变化，动态分配硬件计算单元、内存等资源，实现整体数据吞吐率的最大化。

### 体系化整合：从单点突破到协同增效

AI 原生数据库的核心优势是将向量引擎、自治 AI、多模态能力等创新要素整合为有机整体，通过强化的协同效应使数据库具备持续自我进化的能力，适应 AI 应用快速迭代的需求：

- **向量引擎与自治 AI 的深度融合：**AI 原生数据库的内置向量计算引擎在支持高效相似性搜索的同时，可通过自治 AI 实现索引策略的持续动态优化。例如基于查询模式分析，自动调整向量分片的粒度，并在召回率与系统延迟间取得平衡；同时，自治 AI 模块可监测索引健康度，主动触发重建或压缩操作，进一步提升查询体验。

- **多模态数据处理提升智能推理效能：**AI 原生数据库可构建统一的数据框架，在融合处理文本、图像、视频等多模数据的同时，支持跨模态检索与推理，提升推理匹配度和准确性。例如在医疗诊断场景中，将影像向量数据与病历文本关联，通过多模态分析生成诊断建议。

- **AI 数据全生命周期管理闭环：**将自治 AI 能力贯穿数据库的规划、部署、运维全流程。例如：在规划阶段基于业务负载预测自动配置集群规模；在运维阶段实现快速故障自愈等；在日常运行阶段利用强化学习手段持续开展参数调优，使系统实现自我进化。

### 重构 AI 应用：实现面向业务的价值创造

AI 原生数据库的整体性融合特性对于未来的企业级 IT 系统设计同样意义非凡，其有利于在 AI 应用的开发、测试和运营阶段，实现一系列模式创新与价值重塑：

- **激发业务人员创新意识：**其内置的大量 AI 工具链，让不熟悉数据库专业知识的开发者无需关注底层数据管理细节，即可完成模型训练与部署，通过极低的 AI 数据管理门槛使更多开发者得到技术普惠，大幅缩短创新周期。

- **强化实时智能决策：**AI 原生数据库的多模态处理与低延迟向量检索能力结合，将持续推升高并发、实时推理性能，将更多的传感器、图文、规则信息进行实时融合处理，满足智能制造、智能交通、城市管理等领域的关键性需求。

- **持续盘活存量数据资产：**通过跨模态关联分析与自治优化，AI 原生数据库还有望通过主动挖掘，发现数据中的隐藏价值，形成新的业务增长点。例如在线下消费场景中，将客户的视频动态与消费数据等进行融合分析，进一步完善用户画像并提升现场陈设的合理性。

# 趋势四 04

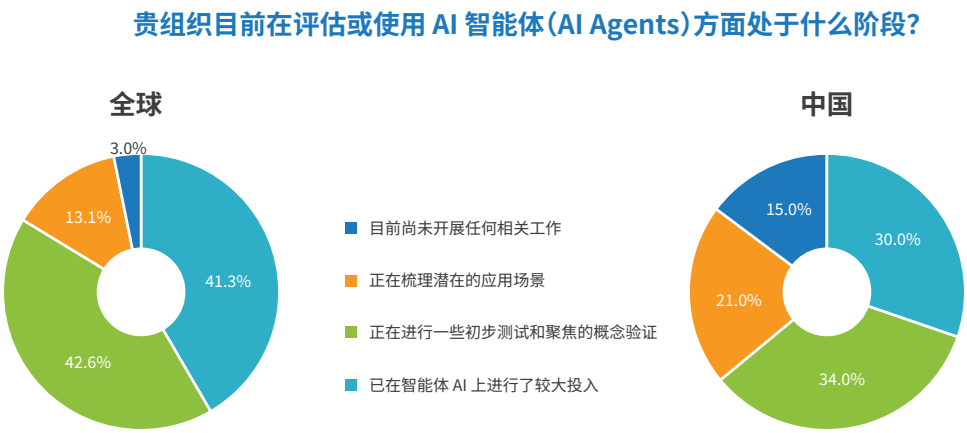
# 数据库智能体成为驱动企业智能化升级的关键力量

在 AIAgent 技术爆发的背景下，越来越多具备特定智能化能力的垂直领域智能体开始被用于数据库基础设施的开发和管理工作，以应对传统数据库管理模式中效率低下、成本高企等诸多挑战。

IDC FutureScape 2026 的“十大预测”显示，到 2027 年，80% 的中国 500 强企业将会部署代理式 AI 平台，为自动化 IT 云运营提供大规模、持续性的监控、分析、故障修复能力，最小化人工干预；此外，到 2029 年，50% 的中国企业将采用 SaaS 平台模式进行实时工作流中的预定义 APP 功能和 AI 智能体的协同，构建模块化和共享交互的解决方案，SaaS 正在向“应用 + 智能体”的平台形态演进。

IDC 报告指出，中国企业 AI 智能体（Agent）应用目前仍处于追赶全球的阶段（34% 受访企业开展测试验证，30% 进入“较大投入 + 采购培训”阶段）。

图 3 全球 vs 中国 AI 智能体采用阶段对比




来源：IDC 2026

早期，数据库领域的智能管控多以单点工具形式存在，例如简单的性能监测工具、数据备份恢复工具等。这些工具功能单一，彼此之间缺乏有效的协同与整合，难以满足数据库全生命周期管理的复杂需求。AI Agent 的引入旨在与数据库管理系统实现深度融合，一方面利用 AI 推动数据库智能开发和管理能力的跃升，同时通过构建统一的智能平台，将不同功能的垂直智能体进行持续整合，形成体系化的数据库 AI 管理解决方案。此外，利用智能体积极促进数据库产业链上下游的技术和服务整合，构建数据库与云计算、大数据、区块链等技术的生态协同格局，也能够进一步加速实现 AI 原生数据库的产业化目标。

## 1. 垂直智能体深度渗透三大核心场景，覆盖数据库开发与管理全生命周期

目前，以数据库应用场景为导向的垂直智能体正加速渗透至数据库开发与管理的核心领域，推动从开发、数据治理到运维的全流程智能化重构。在开发环节，智能体可自动完成代码生成、技术咨询、性能优化等复杂任务，显著缩短项目交付周期；在数据治理层面，智能体通过自动化分类分级以及质量、安全管控，构建强有力的基础保障能力；在运维阶段，基于实时监测、诊断、分析的预测性维护机制，能够提前识别潜在风险并自动触发修复流程，大幅提升系统稳定性。这种面向数据库全生命周期的技术范式变革，为企业构建 AI 时代的全栈数据智能提供了坚实的保障。

### 开发智能体：降低开发门槛加速业务迭代

开发智能体旨在构建低门槛、高效率的数据库智能化开发新模式，其核心特性体现在强大的自然语言转换、代码自动化生成以及实时知识服务等方面。例如，开发智能体能够基于 Text2SQL 技术实现查询需求的语义解析与结构化转换，使开发者能够以业务语言描述数据操作意图，自动生成符合语法规范的 SQL 语句，彻底改变传统开发流程对 SQL 专业知识的强依赖。在代码自动化生成方面，开发智能体普遍能够依据业务场景需求自动生成存储过程、触发器等数据库对象代码，在保证代码规范性的同时大幅降低代码调试周期。此外，开发智能体还可以在数据库性能优化、事务管理等领域构建丰富的动态知识图谱，为开发者提供即时的问题解答与最佳实践推荐，打造无所不在的开发助手。这些显著的智能体技术特性使 AI 时代的数据库开发效率实现倍增，并让更多的业务人员参与到数据库乃至整个系统的开发工作中，利用知识、技术等多重保障手段，确保整个开发过程的质量和稳定性。

### 治理智能体：保障数据安全与合规

在数据治理领域，智能体正在成为推动治理体系智能化升级的核心引擎，其通过融合自然语言处理、机器学习与知识图谱技术，构建起覆盖数据全生命周期的自动化治理能力。在金融、政务等高合规要求行业，治理智能体有望展现出独特价值，基于数据语义特征与业务上下文，自动完成敏感数据识别、分类分级及权限管控，实现数据资产的精细化、可视化管理。更为重要的是，其通过动态监测数据质量指标并触发智能修复流程，结合全链路合规审计与风险预警机制，形成自动化识别、管控、优化的数据闭环治理体系。这种智能化治理模式不仅将人工操作效率提升数倍，更通过主动防御机制显著降低数据安全事件发生率，为关键业务场景提供可信赖的数据基础保障，推动数据治理从被动响应向主动服务、从人工驱动向智能自治的根本性转变。

### 运维智能体：全链路自动化提升系统效能

运维智能体作为数据库全链路自动化管理的核心驱动力，重点聚焦故障诊断、性能优化、容量规划等领域的智能化突破，有效解决传统数据库运维的痛点，让 DBA 从重复性工作中解放出来，从而更加聚焦架构创新等战略性任务，大幅提升业务连续性、系统可用性以及响应效率。

- **在故障诊断方面：**运维智能体改变传统模式依赖人工经验、排查效率低、误判风险大等问题，重点强化实时状态感知与多维数据融合分析体系，借助机器学习算法智能识别异常模式，能快速精准定位故障根源，并自动生成修复策略，形成完整闭环。
- **在性能优化方面：**运维智能体突破静态参数配置的局限，能够基于动态负载与查询模式分析，运用自适应调优算法和执行计划对查询策略进行智能重构，推动系统综合性能获得持续提升。
- **在容量规划方面：**运维智能体能够全面整合系统的历史数据与未来业务需求趋势，通过预测分析模型提前识别资源瓶颈，生成扩容建议和流程依据，帮助企业用户有效规避资源不足的风险。

### 国产数据库厂商的三种突围路径

在 AI 重塑数据库产业格局的历史窗口期，国产数据库厂商正通过三条差异化路径实现战略突围。这三条路径并非互斥，而是构成国产数据库从“可用”到“好用”再到“领先”的完整进化阶梯。

- **路径一：**生态兼容——降低迁移成本的“捷径”：通过兼容 MySQL、PostgreSQL 等主流开源生态，国产数据库大幅降低用户迁移成本和学习曲线。这一路径的核心逻辑是“先替代、后超越”，即以生态兼容性获取市场入场券，再以差异化能力建立竞争壁垒。典型实践包括海山数据库对 PostgreSQL 生态的 100% 协议兼容，使用户可无缝迁移存量应用。

- **路径二：**场景深耕——垂直领域的“护城河”：聚焦金融、政务、能源、电信等关键行业，针对行业特有的数据特征、合规要求和业务流程，打造深度适配的场景化解决方案。这一路径的核心逻辑是“以深打宽”——在垂直领域建立不可替代的专业能力，进而向通用能力延伸。金融行业对事务一致性的极致要求，政务领域对数据主权的刚性约束，均为场景深耕提供了高价值战场。

- **路径三：**技术换道——AI 原生架构的“超车弯道”：跳过传统架构的渐进式改良，直接以 AI-Native 架构重新设计数据库内核，在智能能力上与国外厂商形成代际差异。这一路径的核心逻辑是“换道超车”——在 AI 驱动的新赛道上，传统架构优势被削弱，为国产厂商提供了前所未有的创新空间，通过库内训推一体化、任务级权限等创新，在 AI 原生数据库领域率先布局。

## 2. 数据库智能体实现生态化嵌入，融入企业通用智能体系

在当前的企业智能化转型进程中，数据库智能体更倾向于通过 SubAgent 形态深度融入企业级通用智能体框架，成为连接数据资产与业务价值的核心纽带。这种生态化的部署模式可以充分基于企业现有的服务化架构和系统解耦模式，将数据库智能体的数据解析、SQL 生成、性能优化等核心能力转化为可被调用的标准服务，进而通过统一的服务连接，与业务 Agent、决策 Agent 形成动态功能集合。

- **在业务执行层面：**当 AI 场景需求触发时，系统能够自动组合跨域服务资源，例如业务 Agent 发起数据查询请求时，数据库智能体可同步调用数据治理模块完成数据准备，并联动运维模块评估资源负载，最终生成符合业务场景的智能响应方案，实现从需求感知到价值交付的全链路赋能。

- **在数据与 AI 协作层面：**数据库智能体有助于企业构建出自下而上的价值实现模型。基础价值层通过语义理解引擎将业务语言转化为数据操作指令，加速业务系统的开发进程；数据治理层利用机器学习模型实现数据价值的深度挖掘，为决策系统提供可解释的智能分析结果；业

务价值层打造数据与业务的闭环反馈通道，在强化业务成果的同时持续优化数据资产质量与模型训练效果。这种联动模式使数据库智能体成为数据价值的创造者，其数据处理的维度和形态将获得持续的拓展与延伸，推动企业数据智能水平迈向更高阶段。

- **在技术实现层面：**企业希望实现围绕数据库智能体的生态连接，首先需要建立跨域知识融合机制，通过构建业务术语、数据模型与算法库的常态化关联，实现不同智能体间的语义互通。其次，能够根据 AI 场景对各类服务进行智能组合和动态编排，基于智能化的工作流搭建定制解决方案。此外，企业还有必要进一步引入强化学习算法，形成智能反馈机制，持续调整服务调用策略，例如根据业务 Agent 对决策结果的修正，持续优化语义理解模型的参数等。这些技术实现过程有利于企业建立一个自学习、自进化的智能应用生态，推动数据库智能体随业务发展而获得长期进步。

企业通过生态化部署所构建的通用智能应用体系，将帮助企业在 AI 时代持续提升数据利用效率、决策质量以及运维效能。更为重要的是，这种开放式的连接架构为未来的业务创新充分预留了技术发展空间，当企业引入新的业务系统或 AI 模型时，数据库智能体可通过服务重组的方式快速完成适配，从而塑造持续进化的智能业务发展态势。

## 3. 技术内核升级，从“辅助”走向“自治”

随着大模型进入规模化应用阶段，数据库智能体正从企业智能业务体系的“辅助决策”角色向“自主决策”角色演进。数据库智能体全方位重塑了企业数据库全生命周期的管理模式；依托大模型更广泛的适用性和更强的泛化能力，数据库智能体能够在不同的数据库场景中快速适应和学习，不断提升自身的性能和效率。未来，数据库智能体将具备更强大的上下文感知、多轮对话和自主学习能力，能够在无人干预的情况下，独立完成复杂的数据库管理任务，真正实现数据库的自治化运营。

### 从“辅助”到“自治”，技术跃迁驱动应用范式变革

数据库智能体引发的系统管理与数据服务范式的变革，从根本上源于大模型在语义理解、逻辑推理与任务规划层面的质变。大模型基于海量数据库管理数据的预训练，形成了对存储结构、查询逻辑、性能特征等要素的通用认知，并通过快速泛化能力使其能够快速适配多种类型的数据库环境，实现面向智能管理能力的无缝跃迁。

- **在资源规划阶段：**数据库智能体能够展现出强大的整合分析能力，自主关联存储介质选择、索引优化策略、分区设计参数等关键决策要素，并兼顾业务侧的波动特性以及历史数据特性，动态调整资源分配权重，形成兼顾性能、成本与可扩展性的最佳方案。

- **在系统建设阶段：**数据库智能体凭借多模态解析能力，可同时处理多种数据文档，提取关键信息，结合知识自动生成实用物理模型。其可具备智能化的索引生成能力，能通过分析数据访问、表关联等情况，自主构建多层级索引，评估不同索引组合的影响，输出最优方案。这种智能映射机制，让建模误差率大幅降低，建模周期也大幅缩短。

- **在持续优化阶段：**数据库智能体有望展现出显著的自我进化特性。其内置的强化学习机制可从每次管理操作中精准提取决策特征，构建动态资源优化算法与索引优化生成模型。甚至在业务查询模式发生结构性变化时，系统能够自动触发模型微调流程，在保障系统稳定性的前提下，逐步迭代出适配新业务场景的优化策略。

未来，数据库智能体将依托上下文感知、多轮对话、自主学习等多项能力的持续演进，重塑企业数据基础设施的管理逻辑，智能体的决策范围将覆盖资源调度、性能优化、安全防护等全要素管理，决策时效性从小时级、分钟级缩短至秒级。自治化运营模式使数据库系统能够根据业务发展自动调整管理策略，在确保系统稳定性的同时，最大化释放数据价值。

- **上下文感知能力推动数据库管理策略升级：**通过融合长期记忆机制与知识图谱技术，可精准捕捉跨时间维度的操作关联性，实现对复杂查询链路、存储资源分配等决策的长期影响预判，将传统被动响应式管理升级为主动优化策略执行。

- **多轮对话能力全面赋能复杂管理流程执行：**在存储规划、索引优化、复杂查询等场景中，基于强化学习的多轮对话能力可帮助用户分步明确业务需求边界，动态调整性能与成本之间的约束条件，并推动复杂查询流程趋向最优路径。

- **自主学习能力驱动企业迈向智能自治体系：**企业数据管理体系能够从每日数亿次的管理操作中提炼优化模式，自动生成场景化决策规则，显著缩短数据库管理策略的自我迭代周期，实现资源利用率、业务响应速度与系统稳定性的整体跃升。

## 数据库智能体的核心技术挑战与突破

相比于通用智能体，数据库智能体面临着一系列独特的技术挑战。这些挑战源于数据库系统对一致性、精确性和安全性的极致要求，也是衡量数据库智能体技术成熟度的关键标尺：

- **事务一致性保障：**当数据库智能体执行多步复杂操作（如跨表数据迁移、Schema 变更、批量数据修复）时，需要确保整个操作链的事务性，例如当一步失败时如何自动回滚。这是数据库智能体成为“生产工具”的第一道门槛。其突破方向在于将 Agent 的操作语义与数据库事务机制深度耦合，实现操作级别的 ACID 保障。

- **SQL 语义的精确理解：**大模型存在“幻觉”风险，在生成 SQL 时可能出现语法错误、逻辑偏差或越权操作。因此，需要让数据库智能体在理解自然语言意图的同时，确保生成 SQL 的语义精确性和执行安全性。其突破方向在于构建 SQL-specific 的验证层，通过语法检查、语义分析和影响范围评估三重校验机制，拦截高风险操作。

- **权限的细粒度控制：**传统数据库权限基于“用户 - 角色”模型，但数据库智能体的行为特征与传统用户完全不同。一个 Agent 可能同时执行多种类型的任务，不同任务对数据的访问需求差异巨大。因此，如何让 Agent 以“最小权限”运行成为关键。“任务级权限”机制成为重要突破：即针对某一具体任务赋予特定权限，任务完成后自动收回，实现最小权限原则的动态化落地。

- **可解释性与审计追溯：**数据库智能体的自主决策过程需要可追溯、可审计。当 Agent 自动执行了某项操作（如删除索引、修改参数），管理员需要了解“为什么这么做”以及“依据什么策略”。其突破方向在于构建 Agent 决策日志系统，记录每一步决策的上下文、依据和预期效果，实现“黑盒”到“白盒”的转变。

趋势五 05

# 市场格局重构
# 国内生态崛起 安全赋能创新

## 1. 厂商积极布局：从加速替代到创新引领，布局数据库 AI 技术图谱

市场规模：2026 年中国数据库市场规模预计达 106 亿美元，国产化率将超 70%，从“信创替代”转向“AI 增量支撑”。

AI 时代的数据库产业正经历从“加速替代”到“创新引领”的关键转型，在市场、技术、生态等多个维度实现全面创新突破。国产数据库厂商也在 AI 与数据库的深度融合方面进行了积极布局，推动数据库全生命周期实践向智能化、自治化方向加速演进。

### 国产替代深化，多元格局初现

在政策支持与市场需求的双重推动下，企业级数据库的国产替代进程持续深化，正在从局部领域渗透至关键行业的核心系统，在金融、政务、能源等关键行业的应用比例显著提升。目前，国内数据库市场呈现出多元化竞争格局，并覆盖了公有云与本地部署市场。其中，公有云领域以云原生数据库为主导，承载大量线上业务，并与公有云上的大模型服务共同赋能业务的智能化升级；本地部署市场则兼顾传统架构与创新架构，重点面向有较高安全性需求和自主开发能力的大中型企业。

### 技术引领，AI 赋能体系化发展

数据库领域的 AI 技术深度融合成为产品发展的关键驱动力。头部企业已将 AI 能力深度集成至数据库内核，借助深度学习模型对查询计划进行动态智能优化，使复杂数据分析场景的处理效率实现质的飞跃。各类智能压缩算法能够根据数据特征自动匹配最佳压缩策略，显著提升存储资源利用率。AI 驱动的自治运维系统将全方位帮助企业用户实时监控数据库运行状态，精准预测潜在硬件故障，大

幅增强系统稳定性。这些技术融合重塑了数据库的技术架构与商业逻辑，并有望催生更多的数据服务创新业态。

### 从单点突破到全栈协同，全面提升竞争力

AI 数据库将通过与芯片、操作系统、AI 框架的深度适配，使性能和安全性得到进一步优化，形成生态协作下的技术闭环。在产业方面，AI 数据库厂商联合高校、科研机构及行业用户，推动产学研用协同创新，并加速技术迭代与标准制定。在全球化布局方面，国内的数据库厂商通过跨地区合作与技术输出，逐步拓展海外市场，参与国际竞争，推动新技术标准的全球化应用。

## 2. 行业竞争加剧：从聚焦产品创新到“生态 + 场景”的全方位适配

新一代 AI 数据库行业正以惊人的速度演进，并持续引发产品和服务层面的激烈竞争态势。各厂商在持续强化产品创新的基础上，更加趋向于依托“生态 + 场景”的路径，全方位适配各类行业新场景。其中，生态绑定策略让 AI 数据库产品厂商与基础大模型、AI 应用等上下游企业形成联动，加速全链条的技术进步与产品迭代；场景深耕则聚焦金融、政务、工业等重点领域，打造专属解决方案，满足企业智能化转型的深层需求。

### 生态绑定：构建“数据库 + 模型 + 应用”闭环

AI 数据库与大模型深度集成，相互赋能。在产品迭代发展的进程中，AI 数据库的应用实践将为大模型输送海量且高质量的训练数据，并在一定程度上为其提供实时推理依据，确保大模型能基于最新数据做出精准判断。大模型则凭借强大的语义理解和模式识别能力持续提升数据库效能，实现智能化查询优化、自动索引生成以及异常检测等能力，让数据库更懂业务需求。这种深度集成还将进一步推动 AI 场景的普及，使大量的 AI 应用能够直接调用数据库与大模型的强大融合能力，提供智能问答、智能推荐、智能决策等高级服务，为用户带来个性化、智能化的体验。

AI 数据库产品的发展前景将紧密依赖创新生态的协同成效，其生态策略核心在于构建“数据库 + 模型 + 应用”的紧密闭环。为尽快融入主流生态体系，各数据库厂商也积极投入生态建设，与大模型厂商、应用开发商等建立广泛合作关系，通过整合各方优势资源，共同完善协作体系，以在激烈的市场竞争中能够占据有利地位，共同推动 AI 数据库行业的持续繁荣。

## 场景深耕：聚焦核心领域打造场景化解决方案

金融、政务、工业、互联网等领域成为新一代 AI 数据库竞争的主战场。各厂商针对垂类场景的特点和需求特性，积极打造场景化的解决方案，并将场景化运营作为关键目标进行持续打造，以在激烈的市场竞争中保持主动。

- **在金融领域：** AI 数据库瞄准数据实时性、准确性和安全性等关键目标，全力提升智能化分析能力，满足金融机构在风控、投研等场景中对多维分析、策略优化结果的苛刻要求，为金融服务带来效率和质量方面的显著跃升。
- **在政务领域：** 全面提升各领域的智慧治理能力成为重要诉求。AI 数据库凭借先进的多模态数据处理技术和智能 ETL，帮助用户快速整合不同部门、不同系统和不同业务场景中的数据，为政府科学决策提供精准有力的数据支撑。
- **在工业领域：** AI 数据库聚焦预测性维护等关键应用场景，结合传感器数据和 AI 算法，实时监测设备运行状态，提前预测设备故障，实现设备的智能化运维，降低生产成本，提高生产效率。
- **在泛互联网领域：** AI 数据库能够为大量使用的推荐系统和 AIGC 应用提供保障，支撑多领域的高并发处理以及各种维度的智能化生成需求，满足海量互联网用户的多样化访问请求。

## 开源与闭源融合：拓展企业级服务新路径

在持续的 AI 发展浪潮中，开源 AI 数据库具备开放架构和灵活优势，吸引大量开发者参与其中，推动产品实现快速迭代进化。借助灵活的插件机制，开源数据库产品可迅速集成各类 AI 功能，满足开发者对智能化数据库的迫切需求，极大提升了开发效率与数据库性能。

国内厂商一方面积极投身开源社区，通过参与项目、贡献代码提升影响力，吸引更多开发者与用户，构建活跃的技术生态。另一方面，很多国内厂商也在持续推出闭源的企业级数据库产品，重点聚焦安全、稳定与高性能，为企业核心业务提供坚实保障，满足其对数据安全和业务连续性的严苛要求，成为企业数字化转型的可靠基石。

开源与闭源融合是新一代 AI 数据库发展的关键路径，既可以让 AI 数据库产品借助开源力量紧跟技术潮流，实现快速创新，又能依靠闭源服务满足大型企业用户的高端需求。这种融合策略将长期助力 AI 数据库厂商负重前行，推动 AI 数据库行业向智能化、场景化、生态化加速迈进。

## 3. 安全范式变革：从产品安全防护到安全赋能行业创新

在 AI 深度重塑企业数据生态的当下，企业数据库的安全态势正遭遇前所未有的多维度结构性挑战。传统基于边界隔离与规则匹配的安全防护逻辑，在 AI 驱动的新型攻击范式面前显得力不从心，而 AI 原生数据库的发展则成为企业变革安全保障体系的关键驱动力。通过将安全能力内化，不仅显著提升防御深度与响应速度，更有望将安全保障工作转化为赋能企业用户业务创新的核心动力，从而加速重塑整个数据库市场的产品与服务格局。

### 当前 AI 原生数据库发展面临的主要风险

AI 模型对数据的规模化处理能力，使得单次攻击可获取的数据量呈指数级增长，且攻击者可通过生成式 AI 伪造合法访问请求，绕过传统安全机制，进行数据窃取或篡改。同时，数据在模型训练、推理过程中的流动轨迹难以追溯，导致泄露源定位困难。在 AI for DB 的加持下，不合规操作的场景越来越泛化，智能体的自动化特性模糊了操作边界，也引发了对其行为安全性与伦理合规性的担忧，例如可能出现的歧视性查询结果或恶意决策等。

### AI 赋能的数据安全能力跃升

AI 技术为数据库带来新的风险，同时也正在赋能数据库安全技术栈，重构数据安全格局。例如在数据保护层面，动态脱敏技术结合 AI 算法，能够根据访问者的角色特性，自动调整数据展示精度，既保障了数据使用效率，又降低了泄露风险。同时，基于深度学习的异常检测系统，通过构建用户行为基线模型，有效识别并拦截非常规访问，大幅增强主动威胁防御能力。在数据审计方面，自然语言处理与知识图谱的结合，使得审计过程更加智能化，能够快速理解复杂业务逻辑，精准定位合规风险。

### 面向未来的前沿安全发展态势

在 AI 技术加速跃进的未来，AI 原生数据库的安全边界也将持续拓展。除传统防护外，其更需直面一系列前沿风险挑战。从法律合规的刚性约束到技术迭代的深度赋能，从数据伦理的软性规范到安全架构的范式变革，每一项挑战都关乎 AI 原生数据库产品能否在创新与安全间找到平衡支点，从而在激烈的市场环境中完成规模化落地。

- **AI 安全合规：**随着全球数据法规的日益严格，AI 原生数据库将面临严峻的合规挑战。如何确保 AI 原生数据库的处理活动符合各国法律要求，成为企业必须切实解决的问题。这要求相关产品不仅具备强大的技术防护能力，还要内置合规性检查机制，能够实时监测数据处理活动，确保每一步操作合法合规，避免企业因违规而面临的法律处罚与声誉损失。

- **数据溯源与脱敏技术发展：**数据溯源与脱敏保护将迎来重要的技术升级。例如区块链技术的引入将为数据流动提供不可篡改的溯源记录，确保数据来源的透明性与可信度，有效应对数据泄露后的溯源难题。而同态加密等更先进的技术，也将进一步使数据脱敏保护的应用场景更广泛，既能够保护数据隐私，又不影响 AI 模型的训练效能，为 AI 数据应用过程提供更坚实的安全屏障。

- **AI 自主行为约束：**AI 自主行为的安全与伦理约束也将成为未来研究的重点。例如，如何确保 AI 在数据库管理过程中的每一项决策都符合人类价值观与道德规范，避免歧视性或恶意行为，这需要建立严格的伦理审查机制与行为约束框架，通过算法审计、伦理评估等手段，对 AI 行为进行全方位监管。

- **零信任架构发展：**零信任架构的推广将对 AI 原生数据库的安全产生更加深远影响，其能够对内、外部的任何访问请求都进行严格验证，并通过持续的身份验证和动态权限管理手段，构建起坚固的安全防线，在为业务提供便捷创新环境的同时，有效应对日益复杂的网络攻击威胁，为 AI 原生数据库的长远发展保驾护航。

# 总结

本文对于 AI 数据库的五大趋势预测，不仅勾勒出 AI 与数据库深度融合的发展蓝图，同时也预示着智能化转型对于各行业加速变革的重要影响力。总体来看，各行业的智能化程度将持续攀升，借助先进的模型和算法实现自动优化与决策。在 AI 的强大赋能作用下，企业数据管理与应用将迈向新高度，全面显现海量复杂数据的巨大价值。AI 原生数据库将与云原生技术紧密协同，为企业用户提供更具弹性与灵活性的数据服务能力。此外，跨领域的数据应用也会持续拓展，为各行业发展注入强劲动力。

从用户的视角看，不同的企业均希望在 AI 时代紧跟先进技术潮流，利用 AI 的影响力进一步提升创新能力。因此，企业应根据自身战略规划和基础建设情况，从本文描述的传统数据库 AI 化、专用向量数据库与 AI 原生数据库路线中进行适度选择，将智能化升级与架构创新结合，扎实迈上每一个新的台阶。

企业也需要在 AI 蓬勃发展的态势中，关注主流厂商在相关领域的发展情况。各厂商凭借深厚的技术积累与创新能力，正在紧锣密鼓地推出各具特色的 AI 原生数据库以及相关附属产品。这些产品不仅在技术和功能上贴合 AI 的发展趋势，更在性能、稳定性等方面有持续跃升的卓越表现。移动云数据库脱胎于中国移动云原生技术底座，是移动云体系内的核心基础产品。依托中国移动全域算力网络、底层云资源调度能力与多年技术沉淀，移动云完成数据库全栈自研与产品化演进，打造覆盖关系型、非关系型、分布式及信创自研海山数据库的完整产品矩阵。产品以云原生架构为根基，兼具高可用、弹性伸缩、智能运维、国密安全与全链路容灾能力，兼容主流数据库生态与国产化软硬件体系。作为企业数据存储与运算的核心底座，移动云数据库以央企安全能力为依托，简化企业数据运维成本，赋能政务、金融、互联网等千行百业数字化转型，筑牢数字业务稳定运行的数据基石。

# 关于
# 移动云数据库

## 移动云 AI 原生数据库的战略定位与核心主张

在 AI 重新定义数据基础设施的历史性时刻，移动云提出“数据不动模型动”的核心理念，致力于实现“数据在哪里，智能就在哪里诞生”。这一理念不仅是其技术架构的核心设计原则，更是对企业数据主权与 AI 普惠化价值的深度承诺。

- **独特理解：**AI-Native Database 的本质并非“给数据库加上 AI 功能”，而是围绕 AI 工作范式重新设计数据库内核。数据不应为模型所搬运，模型应向数据所在位置下沉——这是“数据不动模型动”的技术哲学内核，也是破解企业数据安全与 AI 效率矛盾的根本路径。

- **核心差异化优势：**移动云拥有全球运营商规模最大的算力网络，覆盖全国范围的云边端协同基础设施，以及服务 9 亿个人用户、数千万企业客户的丰富场景积累。这种“算力网络 + 海量场景 + 央企可信”的三位一体优势，使移动云在 AI 原生数据库竞争中具备独特战略纵深——既能提供从中心云到边缘节点的全链路算力支撑，又能将数据库产品在真实业务场景中持续打磨优化，更能为政企客户提供满足最高合规要求的可信数据底座。

- **技术愿景：**移动云致力于 “打造 AI 时代的数据基础设施新标准”——构建以内置智能为核心、以安全可信为底线、以生态开放为路径的新一代数据库范式。在这一愿景指引下，数据库将从被动的数据容器演进为主动的智能引擎，从单点技术产品升级为连接算力、数据与 AI 应用的生态枢纽。

作为中国移动旗下的云计算品牌，移动云始终秉持“云原生 + 自主可控”的技术理念，积极布局 AI 原生数据库领域。依托中国移动强大的算力网络基础设施和海量数据场景，移动云构建了覆盖数据库全生命周期的 AI 原生产品体系，为企业提供从数据存储、智能检索到 AI 应用开发的一站式解决方案。

## 移动云 AI 原生数据库：五大技术宣言

移动云在 AI 原生数据库领域的技术突破，可凝练为五个言简意赅的技术宣言。每一项宣言背后，都是对传统数据库架构的深刻变革，都是对企业 AI 应用痛点的精准回应。

- **宣言一：**“库内训推一体化”——数据不动模型动，训练推理在库内完成。传统 AI 应用需要将数据从数据库导出至外部 AI 平台，数据搬运造成安全风险和效率损耗。移动云海山数据库（He3DB）将模型训练与推理能力下沉至数据库内核，用户通过标准 SQL 即可完成数据预处理、特征工程、模型训练和推理预测的全流程操作。数据不出域，安全有保障；计算不下沉，效率无损耗。这一架构使 AI 应用的开发周期从数周缩短至数天，数据安全合规风险降低 90% 以上。

- **宣言二：**“任务级权限”——AI Agent 的最小权限粒度从“用户级”到“任务级”。传统数据库权限体系基于“用户 - 权限”的静态映射，无法满足 AI Agent 动态、多变、不确定的行为特征。移动云首创任务级权限机制：针对某一具体任务赋予特定权限，权限仅存在于任务运行的生命周期内，任务完成后自动收回。这一机制将 Agent 的权限暴露面缩减两个数量级，为“敢用 Agent”提供了安全底气。

- **宣言三：**“PGFS 共享文件系统”——让数据库成为 Agent 的共享记忆中枢。PGFS（PostgreSQL File System）将海山数据库挂载为本地目录，作为 Agent 工具的共享文件系统。多 Agent 可在同一工作空间协作，共享上下文记忆和任务状态，实现“共享大脑”的能力。当 Agent 误删除或修改文件时，可基于 PITR 技术回滚到任意时间点。PGFS 通过标准数据库连接串即可挂载，一行命令完成配置，比传统 NFS 方案更简洁高效。

- **宣言四：**“8192 个数据沙箱”——大规模 Agent 并行隔离的安全底座。每个 Agent 可从数据库实例中复制出独立的数据工作区，工作区之间完全隔离。数据沙箱支持秒级创建，最大可支持 8192 个独立工作区，满足大规模 Agent 并行运行的隔离需求。结合 Time Travel 能力，系统可恢复到过去任意时间点的数据库状态，为 Agent 操作提供“无限撤销”的安全网。

- **宣言五：**“混合检索引擎”——一条 SQL 搞定结构化查询与语义搜索。基于 PGvector 深度改造，支持 HNSW、IVFFlat 等多种向量索引算法，原生支持“向量 + 标量”混合查询。单条 SQL 即可完成语义相似度检索与业务属性过滤的联合计算，满足 RAG 场景的高效召回需求。移动云 Elasticsearch 向量数据库产品更支持高达百亿级向量的毫秒级检索，延迟控制在 2 秒以内。

# 一、移动云 AI 原生数据库技术突破、主流产品、技术图谱

## AI IN DB 能力（库内 AI、库内训推一体化、混合检索）

移动云海山数据库（He3DB）内核层面深度集成 AI 能力，实现了“AI for DB”与“DB for AI”的双轮驱动。

- **AI for Kernel：** He3DB 内置了智能优化器，能够深度分析特定工作负载的运行特征与访问模式，结合机器学习算法自动完成查询计划优化、索引推荐和参数调优。通过对历史 SQL 执行数据、资源使用趋势及业务负载规律的持续学习，系统可智能预测性能瓶颈并主动调整优化策略，实现从“人工经验驱动”到“数据智能驱动”的转变，显著降低 DBA 运维负担，提升数据库整体性能与稳定性。

- **库内训推一体化能力：** He3DB 创新性地将模型训练与推理能力下沉至数据库内核，实现了“数据不动模型动”的高效 AI 计算范式。在数据库内部集成轻量级 ML 运行的同时支持 SQL-based 模型训练语法，用户可直接使用标准 SQL 语句完成数据预处理、特征工程、模型训练和推理预测的全流程操作。该架构避免了海量数据在数据库与外部 AI 平台之间的频繁搬运，同时确保数据安全不出域，满足金融、政务等高合规场景需求。

- **混合检索引擎：** He3DB 构建了业界领先的混合检索能力矩阵，覆盖结构化、向量、全文等多种检索范式。基于 PGvector 深度改造优化，支持 HNSW、IVFFlat 等多种向量索引算法，实现向量检索性能提升。同时原生支持“向量 + 标量”混合查询，单条 SQL 即可完成语义相似度检索与业务属性过滤的联合计算，满足 RAG 场景的高效召回需求。此外，移动云 Elasticsearch 向量数据库产品同样支持混合检索能力，支持 HNSW（高召回率）与 LSH（低延迟）等多种向量索引算法，可实现高达百亿级向量的毫秒级检索，延迟控制在 2 秒以内。同时支持全文搜索、向量搜索、混合搜索以及数据分析等功能。

## AI Infra 基础设施（垂域大模型、统一数据存储、数据库工具 Hub）

移动云面向 AI 原生数据库时代，构建了完整的数据基础设施体系，为数据库智能体的快速构建提供坚实底座：

- **垂域数据库大模型：**移动云基于海量数据库运维数据和领域知识，训练了专门面向数据库场景的垂域大模型。该模型深度理解 SQL 语义、数据库架构和运维场景，在 SQL 生成、故障诊断、性能优化等任务上的表现显著优于通用大模型。通过持续微调机制，模型能够不断吸收新的运维案例和最佳实践，实现能力的持续进化。
- **统一数据存储：**移动云构建了面向 AI 应用的统一数据存储层，支持结构化、半结构化、非结构化数据的统一存储和管理。通过湖仓一体架构，实现数据湖的数据灵活性与数据仓查询性能的完美融合。统一元数据管理确保数据资产的清晰可追溯，为 AI 训练提供高质量数据供给。
- **数据库工具 Hub：**移动云打造了数据库领域的工具集市（Tool Hub），整合 Schema 设计、SQL 审核、性能分析、数据脱敏、备份恢复等丰富的数据库工具。这些工具以标准化接口对外开放，可被智能体灵活调用，快速组装成面向特定场景的解决方案，大幅降低智能体开发门槛。

表 3 移动云 AI Infra 基础设施体系

| 层级 | 核心能力 | 代表产品 / 服务 |
| --- | --- | --- |
| 垂域大模型 | 数据库领域专用模型 | 移动云数据库大模型 |
| 统一数据存储 | 多模数据统一管理 | 海山数据库、数据湖 |
| 工具 Hub | 数据库工具集市 | Schema 设计、SQL 审核、性能分析 |

来源：移动云 2026

### Agent 生态适配（运维 Agent、迁移 Agent、SQL 生成）

移动云积极拥抱 AI Agent 技术趋势，在数据库领域构建了丰富的智能体生态。

**当前实践**

- **运维 Agent：**基于垂域大模型和工具 Hub，打造运维智能体，实现故障自动诊断、性能瓶颈定位、容量智能预测，将 DBA 从重复性劳动中解放，运维效率提升 5 倍以上
- **迁移 Agent：**面向异构数据库迁移场景，打造迁移智能体，支持异构数据库的智能迁移评估、Schema 转换、数据校验，结合大模型的语义理解能力，显著降低迁移风险和人力成本

- **SQL 生成 Agent：**面向 SQL 辅助编程场景，打造 SQL 生成智能体，基于自然语言理解和垂域模型，实现 Text2SQL 智能转换，准确率达 90% 以上，大幅降低开发门槛，同时完善 SQL 生态能力，补齐 SQL 解释、SQL 自动补全等能力

### 演进方向

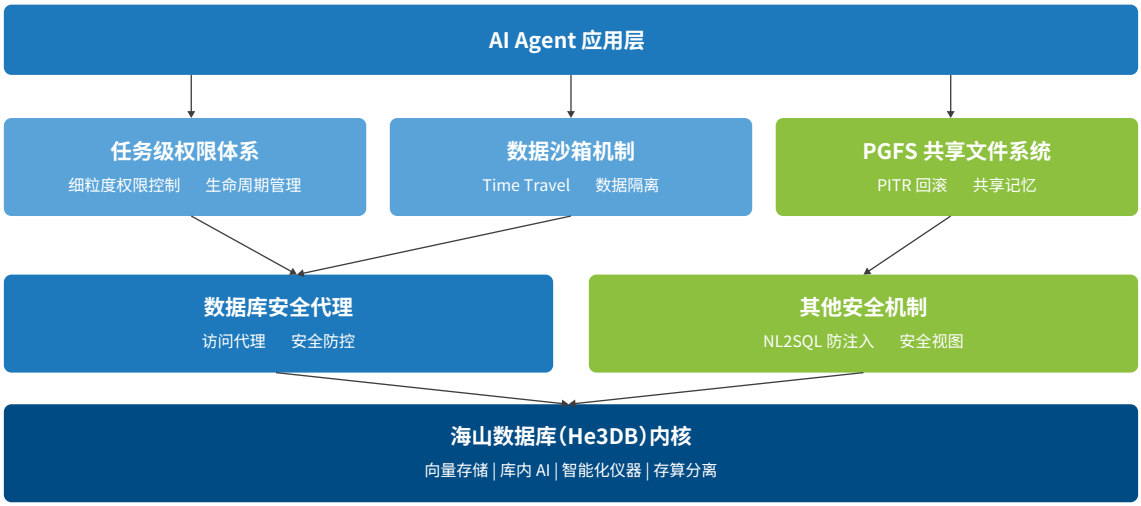
面向未来，移动云数据库智能体将向通用智能体架构演进。参考 OpenClaw 等前沿范式，打破当前垂直 Agent 的边界，构建统一的智能体基座——核心能力以“Skills（技能插件）+ MCP（工具集）”的标准化形式接入，实现“一次构建、多端复用”。

这种架构下，运维、迁移、SQL 生成等能力将沉淀为可插拔的技能模块，通过 MCP 协议与底层数据库资源无缝连接。智能体不再局限于单一场景，而是根据用户意图动态组合技能，实现真正的统一赋能：同一入口理解需求，灵活调度全栈能力，持续学习优化，最终形成自进化的数据库智能服务生态。

## AI 原生数据库的安全能力构建

随着 AI Agent 技术在企业场景的深入应用，数据库安全面临全新的挑战。与传统应用不同，AI Agent 具有行为不确定性高、任务边界模糊、多任务并行执行等特点，传统的、基于用户身份的权限体系已难以满足安全需求。移动云基于海山数据库（He3DB）内核能力，结合业界最佳实践，构建了面向 AI 原生场景的五层安全体系，为 AI Agent 的安全运行提供坚实保障。

图 4 AI 原生数据库安全体系架构



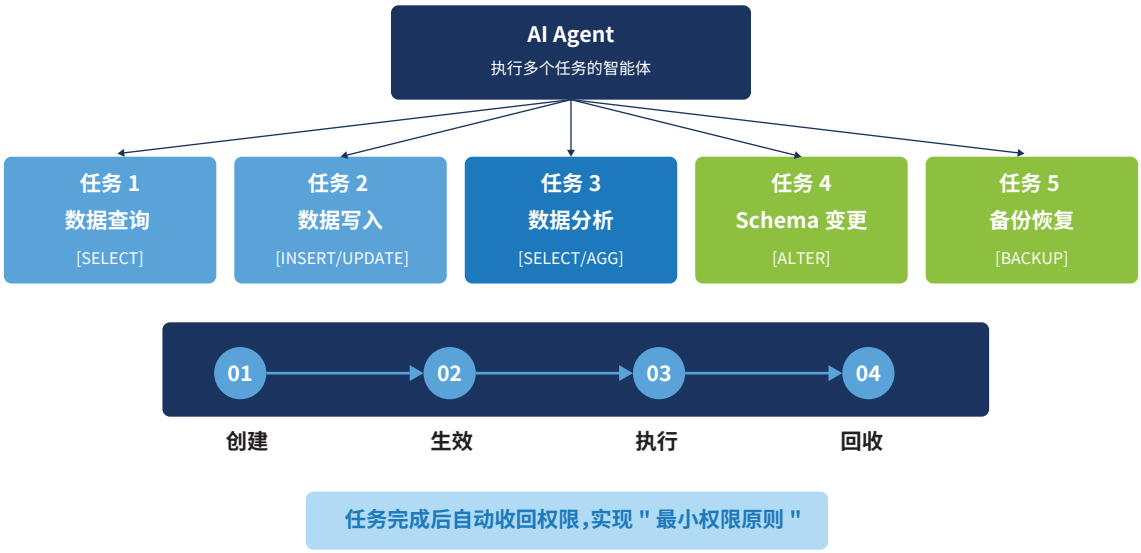

来源：移动云 2026

### 4.1 任务级权限体系

传统数据库权限体系基于“用户 - 权限”的静态映射模型，假设数据库用户是具体的人或稳定的业务应用。然而，在 AI Agent 场景下，这种模型面临两大挑战：一是 Agent 行为相比人更具不确定性，大模型的 “幻觉” 可能导致非预期操作；二是一个 Agent 执行多个任务将成为常态，不同任务对数据的访问需求差异巨大。

基于以上问题，移动云海山数据库创新性地提出任务级权限机制，针对某一具体任务（Skill）赋予特定权限，权限仅存在于任务运行的生命周期内，任务完成后自动收回。该机制实现了“最小权限原则”的动态化落地，有效降低 Agent 越权访问风险。

图 5 任务级权限体系




来源：移动云 2026

#### 数据沙箱机制

移动云海山数据库布局实现数据沙箱与 Time Travel 能力，每个 Agent 能够从数据库实例中复制出独立的数据工作区，工作区之间完全隔离，互不影响。数据沙箱支持秒级创建，最大可支持 8192 个独立工作区，满足大规模 Agent 并行运行的需求。结合 Time Travel 能力，系统可恢复到过去任意时间点的数据库状态，为 Agent 提供“无限撤销”的安全网。当 Agent 执行误操作时，可快速回滚到正确状态，大幅降低实验成本和风险。

图 6 数据沙箱与 Time Travel 机制


来源：移动云 2026

## PGFS 共享文件系统

PGFS（PostgreSQL File System）将海山数据库挂载成本地目录，作为 OpenClaw 等 Agent 工具的共享文件系统。该方案解决了两个核心痛点：

- **防止误操作：**利用海山数据库的事务机制和版本控制能力，实现文件系统的闪回能力。当 Agent 误删除或修改文件时，可基于 PITR（Point-in-Time Recovery）技术回滚到任意时间点。
- **记忆共享：**基于 PGFS 的共享文件系统，多 Agent 可在同一工作空间协作，共享上下文记忆和任务状态，实现“共享大脑”的能力。

PGFS 通过标准的数据库连接串即可挂载，一行命令完成配置，比传统 NFS、FTP 方案更加简洁高效。同时，PGFS 继承了海山数据库的高可用、高并发特性，满足企业级生产环境要求。

## 数据库安全代理

当前 PostgreSQL 生态缺乏成熟的数据库代理层，Agent 直连数据库存在安全风险。移动云海山数据库提出“Agent 只能通过代理访问数据库”的安全原则，在 Proxy 层构建系列安全防控机制：

表 4 数据库安全代理核心功能

| 功能模块 | 核心能力 |
| --- | --- |
| 连接管控 | 统一连接入口，实现连接池管理、负载均衡、故障转移 |
| SQL 审计 | 全量 SQL 日志记录，支持实时分析和离线审计 |
| 访问控制 | 基于 IP、时间、操作类型的细粒度访问策略 |
| 流量管控 | QPS 限流、并发控制、慢查询拦截 |
| 敏感操作拦截 | DDL、批量删除等高风险操作的人工审批机制 |

AI 原生数据库的安全体系建设是一个系统工程，需要从权限管理、数据隔离、访问控制、审计追溯等多个维度协同发力。移动云基于海山数据库内核能力，结合 PGFS、数据沙箱、安全代理等创新技术，构建了面向 AI Agent 场景的完整安全解决方案，为企业“敢用、能用、好用”AI Agent 提供坚实的数据安全底座。

## 主流 AI 原生数据库产品（海山 PG、ES、DMS）

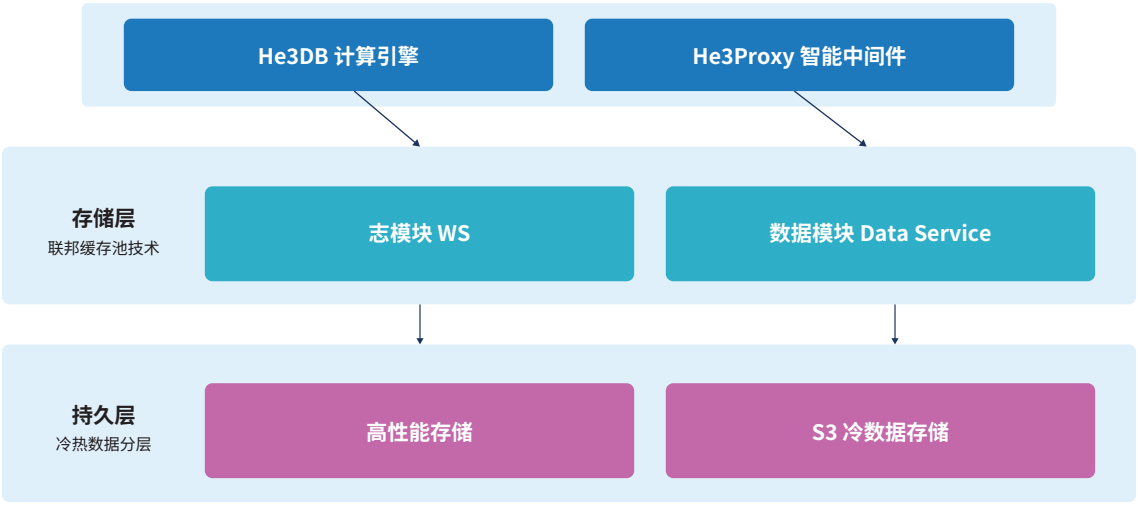
- **海山数据库（He3DB）——向 AI 原生数据库演进**

海山数据库是移动云自主研发的云原生关系型数据库产品，于 2025 年 8 月通过中国信息安全测评中心与国家保密科技测评中心的“安全可靠测评”，成为国内电信运营商首款获此认证的数据库产品。在 AI 原生演进方面，海山数据库实现了以下关键突破：

- **向量能力原生集成：**基于 Pgvector 深度优化，内置向量数据类型和索引，支持“向量 + 标量”混合查询，无需额外部署向量数据库即可满足 RAG 场景需求

- **库内 AI 计算：**内置轻量级 ML 运行时，支持 SQL-based 模型训练和推理，实现数据不出域的 AI 计算
- **智能优化器：**基于深度学习的查询优化器，自动选择最优执行计划，复杂查询性能提升 30% 以上
- **存算分离架构：**主备节点共享数据，支持一主 15 备，RTO 绝对时间小于 30 秒，满足金融级高可用要求
- **Agent 生态构建：**面向移动云海山数据库，积极拥抱 Agent 生态，打造运维、迁移和 SQL 辅助编程智能体，全面赋能数据库应用，改善用户体验

图 7 海山数据库 (He3DB) 产品架构



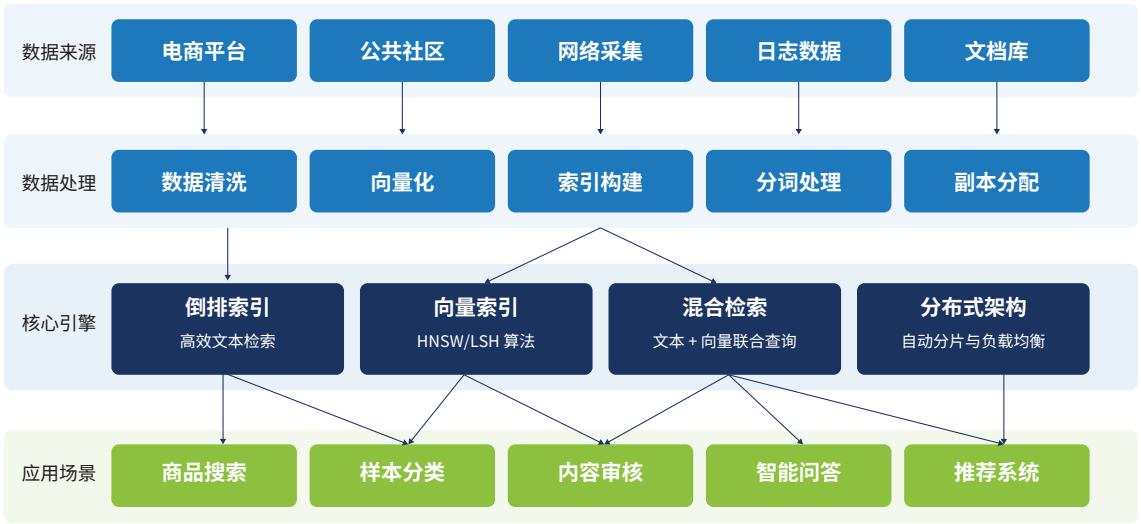

来源：移动云 2026

## Elasticsearch——向量检索服务

移动云 Elasticsearch（ES）是移动云基于开源 Elasticsearch、OpenSearch 搜索引擎打造的高可用、可伸缩的云端全托管搜索服务，支持全文检索、向量检索、混合搜索等能力。产品提供专业版与企业版两大服务体系，涵盖 Elasticsearch、OpenSearch、向量检索、RAG 平台及 AI 模型等组件，从日志分析到 AI 智能检索，满足不同业务场景需求。在 AI 原生演进与云原生架构方面，ES 产品实现了以下关键突破：

- **全场景覆盖，灵活适配：**提供专业版与企业版两大服务体系，涵盖 Elasticsearch、OpenSearch、向量检索、RAG 平台及 AI 模型等组件，从日志分析到 AI 智能检索，满足不同业务需求
- **一站集成 RAG，加速落地：**提供向量库、RAG 平台、知识库存储、AI 大模型调用等 RAG 场景全链路组件化服务，实现知识库管理、模型管理、聊天问答等全流程模块，内置 DeepSeek、Qwen 等主流大模型支持，适配 LangChain、Dify、Ragflow 等开源 AI 框架，加速 AI 应用落地
- **内核增强，性能卓越：**在开源能力基础上，具备向量检索、SQL 查询、批写入加速、物理复制、智能索引压缩等企业级增强特性；百亿向量数据召回率达 99%，吞吐量超 1 万 + 并发，毫秒级查询响应；支持超过 16000 维向量处理，兼容 LSH、HNSW 等主流索引算法，提供 L1、L2、Cosine、Jaccard、Hamming 等多种相似性计算方法，并可平滑扩展至 PB 级数据量
- **弹性扩展，稳定可靠：**提供多种节点规格与存储介质，支持按需扩缩容，单集群最大支持 99 节点，可扩展至 PB 级数据量；支持同城跨多可用区容灾部署，服务可用性不低于 99.95%
- **生态集成与国产化适配：**全链路国产化适配，兼容鲲鹏、海光、飞腾芯片；支持公有云、私有云、专属云、一体机等多形态交付，提供 SDK、Logstash 等多端数据接入方式，满足不同场景下的部署需求

图 8 移动云 Elasticsearch 产品架构




来源：移动云 2026

## DMS——数据库管理服务

移动云数据库管理服务（Database Management Terminal，简称 DMS）是移动云推出的一款用于登录并管理云上数据库的云原生 Web 服务，提供元数据可视化管理、界面化 SQL 查询、细粒度权限管控等核心能力。它实现了统一管理、跨资源池接入，支持关系型数据库（如 MySQL、PostgreSQL、SQL Server 等）与 NoSQL 数据库（如 Redis 等），为 DBA、开发者和运维人员提供一站式的数据库管理、开发与治理平台。在数据库云原生管理与智能化运维方面，DMS 实现了以下关键突破：

- **跨源统一接入与管理：**支持跨资源池统一管理，兼容关系型数据库与 NoSQL 数据库的多源接入，提供实例全生命周期管理（创建、启用、禁用、删除、状态检查）、库级 Schema 管理、表级元数据自动同步，以及视图、存储过程、触发器、函数等可编程对象的统一管理，实现异构数据库资源的一站式纳管

- **智能化 SQL 开发与分析:** 提供 SQL Console 交互式查询环境，支持 SQL 执行、执行计划分析、SQL 格式化、智能提示（TextToSQL）及 SQL 解释能力；集成常用 SQL 管理、批量导入导出、跨库操作等功能，帮助开发人员高效完成数据查询与开发工作，降低 SQL 编写门槛

- **细粒度安全权限管控：**构建“库 - 表 - 敏感列”多层级权限体系，支持库 Owner/ 表 Owner 管理、库表权限申请与审批、敏感列脱敏保护、算法配置与重置；结合 RAM 访问控制与云账号授权审计，实现操作可追溯、数据不出域的一体化安全保障，满足企业级合规要求

- **数据库自治与智能优化：**内置异常智能诊断引擎，支持 SQL 分析优化、智能索引推荐、智能参数调优（智能调参）等自治服务能力；通过运维助手与操作日志的全量采集、查询与下载，帮助 DBA 快速定位性能瓶颈与异常根因，实现数据库的自治化运维

- **DevOps 一体化工作流：**提供完整的工单驱动型数据治理流程，涵盖权限申请（库 / 表 / 敏感列 /Owner）、表结构设计、测试数据构建、批量数据导入、SQL 审核（即将上线）、无锁变更、数据恢复等能力；支持定时任务的创建、发布、有效性检查与运行历史追踪，实现数据库变更的标准化、自动化与可追溯

图 9 移动云 DMS 产品架构


来源：移动云 2026

表 5 移动云 AI 原生数据库产品矩阵

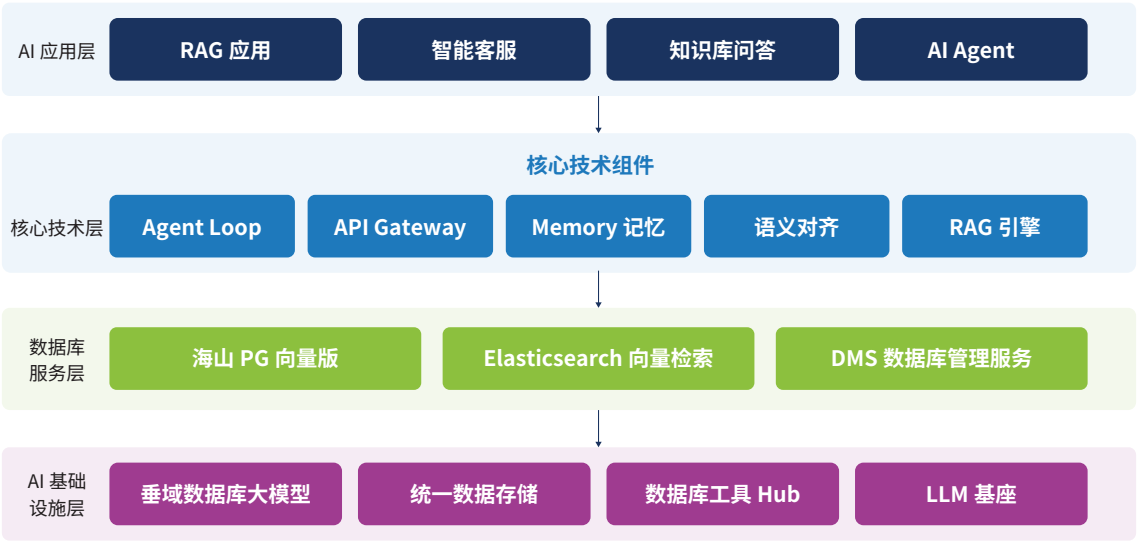
| 产品名称 | 产品定位 | 核心特性 |
| --- | --- | --- |
| 海山 PG | AI 原生关系型数据库 | 向量原生集成、库内训推一体化、智能优化器、Agent 生态构建 |
| Elasticsearch | 向量检索服务 | 百亿级向量检索、混合搜索、毫秒级响应 |
| DMS | 数据库管理服务 | 智能运维、SQL 优化、全生命周期管理 |

来源：移动云 2026

## 技术图谱

移动云 AI 原生数据库技术图谱呈现四层架构设计，从底层基础设施到上层应用形成完整闭环：

图 10 移动云 AI 原生数据库技术架构




来源：移动云 2026

如图所示，移动云 AI 原生数据库架构分为四层：AI 基础设施层提供垂域数据库大模型、统一数据存储、数据库工具 Hub 和 LLM 基座等基础能力；数据库服务层整合海山 PG 向量版、Elasticsearch 向量检索、DMS 数据库管理服务等核心产品；核心技术层提供 Agent Loop、API Gateway、Memory 记忆、语义对齐 Semantic Alignment、RAG 引擎等关键技术组件；AI 应用层支撑 RAG 应用、智能客服、知识库问答、AI Agent 等丰富场景。四层之间通过标准化接口实现数据流动和能力协同，形成面向 AI 时代的完整数据技术栈。

**核心技术层关键组件说明：**

- **Agent Loop：**智能体执行引擎，负责任务分解、工具调度、结果整合，实现复杂任务的自主完成
- **API Gateway：**统一接口网关，提供标准化的数据访问和能力调用接口，屏蔽底层复杂性

- **Memory 记忆：**上下文记忆模块，支持多轮对话和长期记忆，提升智能体的上下文理解能力
- **语义对齐 Semantic Alignment：**语义理解和对齐组件，确保用户意图与数据库操作的精准映射
- **RAG 引擎：**检索增强生成引擎，整合向量检索与大模型生成能力，提供高质量问答服务

## 二、移动云客户案例

**案例一：某大型云盘平台多模态智能检索升级**

- **项目背景：**客户原有自建开源向量数据库在支撑业务快速增长过程中暴露出多方面核心痛点：一是百亿级图片库检索耗时久、查询效率低，难以满足用户秒级响应体验要求；二是知识问答、相册识别、智能助手等不同业务场景对检索延迟和召回率的需求各异，现有单一方案无法精准适配；三是数据更新需重置整个数据集，维护成本高、响应周期长；四是自建库存在单点故障、扩容复杂、缺乏有效容灾机制等问题，运维负担沉重，已难以支撑 AI+ 业务数据的快速增长。
- **解决方案：**采用移动云 AI 原生数据库全面替换原有开源向量库，部署超 400 节点的大规模集群，整合约 480 亿基础数据与多模态数据，为用户提供更智能的查找与问答体验，重塑云盘交互模式。方案覆盖图片资产处理、文档资产处理、智能助手等核心业务场景，实现从数据存储到智能检索的全链路升级。
- **项目成效：**实现百亿级多模态数据秒级查询响应，相较原有开源 Milvus 数据库方案端到端性能提升 90%；云原生架构彻底解决了开源向量数据库的单点瓶颈与运维困难问题，显著降低运维成本；基于容错、高可用、无感知扩展等云原生能力，每年节约资源与运维成本约 460 万元

**案例二：移动云智算平台政企服务助手知识库建设**

- **项目背景：**移动云智算平台为企业提供智能体开发工具链、模型广场、智能体应用及模型运营服务等全栈能力，满足客户多样化定制诉求。平台 AI 模型层亟需高性能 AI 原生数据库作为 RAG（检索增强生成）架构的关键记忆组件，以支撑知识库问答、智能助手等场景的高效运行。
- **解决方案：**采用移动云 AI 原生数据库为智算平台提供知识库数据的高效存储与智能检索能力。以问答助手智能体搭建为例，开发人员在为智能体配置相关技能时，可基于智算平台公共向量数据库或移动云 AI 原生数据库快速创建知识库，实现专属知识问答助手的一站式构建与部署。
- **项目成效：**该方案已深度集成至移动云智算平台，成功落地苏州高新区“苏新享 ·AI+ 政企服务助手”等政务标杆场景。目前已建成覆盖 140 余项政务服务事项的标准化智能知识库，整体问答准确率突破 90%，有效精简咨询流转环节，显著提升政务咨询响应效率与便民服务体验。

# 未来展望：AI 原生数据库的下一站

站在 2026 年的时间节点展望未来，AI 原生数据库正加速迈入全新发展阶段。未来 1-3 年，围绕以下三类技术的演进有望取得突破性进展：

- **数据库与 AI Agent 的深度融合：**数据库将从被 Agent 调用的外部工具演进为 Agent 的“原生宿主”。Agent 的运行时、记忆存储、任务调度将与数据库内核深度耦合，形成“DB-Native Agent”的新型架构范式。
- **多模态推理的库内实现：**数据库将具备库内多模态推理能力，直接完成图像识别、语音理解、文本生成等复杂任务。用户仅需通过标准 SQL 即可调用多模态 AI 能力，无需关注底层模型的部署与调度细节。
- **安全技术与应用取得新突破：**在数据治理、隐私保护上持续创新，进入更加多元的应用场景。不仅全面应用同态加密等技术，还将与量子计算等新兴技术持续碰撞，开拓全新安全架构与应用场景。
- **自治等级的持续跃升：**数据库自治能力将在未来三年实现两次关键跨越：一是“条件自治”实现规模化商用部署，二是“高度自治”进入生产环境中。

从 AI 原生数据库的技术发展路线看，其首先应围绕库内训推一体化、混合检索引擎、任务级权限体系等关键能力，快速发展成熟并投入生产环境。在此基础上，由主流厂商发起构建面向数据库智能体的生态开放平台，支持第三方开发者发布自定义 Agent，数据沙箱可扩展至万级。1-2 年内，多模态推理能力将广泛内置于数据库产品中，将自治等级推至 L3 水平，一批国内厂商将参与制定全球数据语义标准。

AI 原生数据库并非遥不可及的未来蓝图，而是正在发生的现实变革。它将深度融入各行各业的发展进程中，深刻重塑数据基础设施的技术范式，并重新定义企业发展与数据能力建设的核心关系，成为未来企业发展的关键基石。

面对未来激烈竞争的全球市场环境，能率先将 AI 原生数据能力体系深度融入业务发展的企业，将获得显著的内生竞争优势；通过建立精准洞察和前瞻性预测，精准把握市场动态，极致优化运营流程，快速形成差异化竞争和体系化创新优势，从而引领行业变革潮流，全面开启“数据即智能”的新时代。

**参考文献与数据来源**

本白皮书引用的数据和观点来源于以下机构和报告：


- IDC,《Worldwide Database Management Systems Forecast, 2024-2028》
- IDC,《中国数据库市场预测，2024-2028》
- 中国移动通信集团 ,《移动云 AI 原生数据库技术白皮书》
- PostgreSQL Global Development Group,《PostgreSQL Documentation》
- VectorDBBench Project,《Vector Database Benchmark Report》


注：部分市场预测数据将根据 IDC 2025 年最新发布的研究报告进行更新。移动云保留对本文档内容的最终解释权。

# 关于 IDC

国际数据公司（IDC）是在信息技术、电信行业和消费科技领域，全球领先的专业的市场调查、咨询服务及会展活动提供商。IDC帮助IT专业人士、业务主管和投资机构制定以事实为基础的技术采购决策和业务发展战略。IDC在全球拥有超过1100名分析师，他们针对110多个国家的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。在IDC超过50年的发展历史中，众多企业客户借助IDC的战略分析实现了其关键业务目标。IDC是IDG旗下子公司，IDG是全球领先的媒体出版、会展服务及研究咨询公司。

**IDC China**

IDC中国（北京）：中国北京市东城区北三环东路36号环球贸易中心E座901室

邮编：100013

+86.10.5889.1666

Twitter: @IDC

blogs.idc.com

www.idc.com