

KBase多模数据库管理系统 技术白皮书



一、产品背景

1. 强化数据库核心支撑

数字经济时代，全球数据量激增且类型日益复杂（结构化、半结构化、非结构化），驱动各行业对数据库的高效、智能、安全处理能力需求迫切。作为与操作系统、中间件并列的三大基础核心软件，数据库是所有应用运行和数据交互的基石，战略地位至关重要。

2. 响应国家战略要求

政策导向明确： 国家发展改革委等六部门《关于促进数据产业高质量发展的指导意见》（2024 年 12 月 28 日）明确提出“加速向量数据库、多模态数据分析等实时检索分析技术突破”，并支持大模型等 AI 技术创新。

信创国产化刚性要求： 国资委《79 号文》（2022 年 9 月）等政策文件，强力要求央国企等关键领域落实信息化系统信创国产化改造，覆盖芯片、基础软件（含数据库）、操作系统、中间件等核心环节，破解“卡脖子”困局。

3. 把握 AI 融合机遇

国外技术垄断带来双重风险： Elasticsearch 在非结构化数据领域的长期垄断，导致核心技术受制于人与开源协议风险 (SSPL/Elastic License) 双重安全威胁。

多模数据库支撑智能应用： AI 与大模型的快速发展，对底层数据基础设施提出了更高要求。融合多模处理能力（尤其是向量检索）的数据库，正成为支撑智能应用的关键。这类数据库的核心价值在于：

- **打破数据孤岛**：多模态融合，统一处理多种数据类型，避免多库并存带来的架构复杂与数据割裂。
- **支撑实时向量检索**：提供高性能向量处理能力（支持千万级 QPS 与毫秒级响应），满足大模型高并发、低延迟需求。
- **分角色高效协同**：为开发者提供灵活 API，为数据分析师提供可视化工具，为系统管理员提供全链路监控，满足 AI 应用全角色协同需求。
- **分级资源管控**：实现计算/存储/内存资源分级管理，保障核心业务稳定性。
- **赋能大模型**：为大模型的“智慧外脑”，解决大模型“幻觉”问题，弥补了大模型在知识实时性、长期记忆、事实准确性、隐私保护、成本效率和多模态处理等方面的关键短板。

4. 确立国产化多模数据库战略定位

在国家强力推进信创国产化与数据产业智能化升级的背景下，直面 ES 等国外技术垄断风险，抓住向量技术发展浪潮，自主研发安全可控、性能领先、深度融合搜索与向量能力的国产多模数据库，并全面适配国产化环境，已成为保障国家信息安全、突破技术封锁、支撑产业智能化转型的战略选择。

二、方案概述

KBase 是以搜索型数据库为主要模型，集成向量、NXD、RDF 的多模型数据库管理系统。KBase 发展的主要驱动力，来自于知网数

据管理与应用开发的需求。为了提高基于 KBase 的应用开发效率，KBase 借鉴 SQL 标准，研发了 KSQL 查询语言。同时提供 C/C++、Java、Python 等 6 种开发语言接口，以及一系列开发管理及运维工具。支持并行计算和集群部署，确保了系统的高可用性和高扩展性。KBase 集成了智能文本挖掘引擎，支持文本的分类、聚类、自动摘要等智能文本处理能力。全新推出的 AIKBase 向量数据库专注于高效处理大规模向量数据，支持标量、全文与向量数据的混合查询，可在单机上实现十亿级高维向量数据的毫秒级检索。

三、产品功能



1. 提供高性能向量数据库引擎

专为 AI 场景设计的新一代向量数据库引擎，支持跨模态相似性搜索与复杂向量计算，高效赋能大模型应用开发。其核心能力包括：

- **十亿级向量毫秒响应**：支持单机十亿级高维向量数据的毫秒级检索。
- **智能混合查询**：兼容标量数据、全文数据与向量数据的统一查询，提供灵活数据管理。
- **多模态融合检索**：开箱即用的文本、图像存储与检索，支持文本、图像等跨模态相似性搜索与复杂向量计算。
- **大模型深度集成**：提供便捷接口，无缝对接各类 AI 大模型应用。
- **自适应智能向量索引**：面向多样化计算与检索场景，支持多种高效索引算法。
- **海量数据处理**：确保大规模向量数据环境下仍保持高性能检索与分析。

2. 搭载领先全文索引

支持近 10 种全文索引类型和模式，具备高度灵活性以适应多样化的功能与性能需求。支持多值字段、虚字段与字段映射等特色功能，满足复杂文本检索与相关度排序需求。

3. 创新多模数据架构

为满足多源不同类型数据的存储和检索需求，KBase 多模数据库架构具有创新性和广泛的应用领域。支持四种数据库模型：搜索型数据库模型、向量数据库模型、NXD 数据库模型、RDF 数据库模型。

4. 应用创新指纹索引

创新性提出并应用指纹索引技术，在非向量场景下也能实现稳定可靠的相似性检索。该技术作为底层核心，可应用于海量文本、学术图像检索等产品，具备在海量数据（如学术论文、图像库）中快速匹配并输出相似结果的高效能力。

5. 集成智能中文信息处理

深度集成知网先进的文本挖掘引擎、自然语言处理引擎及概念关系词典，构建高度智能化的中文信息处理环境。支持自动分类、自动聚类、关键词自动标引、内容相似性分析、自动文摘、信息过滤、特征词分布规律分析、通用关联规则分析。

6. 独创 KSQL 语言

支持文档检索、关系数据检索，融合 XPath 和 XQuery，支持 SPARQL 语言，提出排重合表等特色功能，具备强大的检索功能。

四、产品优势

1. 确保信创适配自主可控

内核自主可控、共获得国产化平台互认证证书 35 个，可以满足国内绝大部分国产化机型和操作系统的适配需要。针对鲲鹏等国产化机型专项性能优化。

2. 实现全文检索毫秒级响应

经国家工业信息安全发展研究中心赛昇实验室检测，500 万期刊数据全文检索，300 秒并发事务响应时间为 3.9ms，百万级的文献量

相似检索速度达到毫秒级。支持实时信息排重，拥有 400 万数量级的概念关系词典，采用全切分算法，切词准确率高达 98%，使得 KBase 的查准率、查全率在同类产品中领先。相对而言，传统的关系数据库如 SQL Server 全文检索速度仅为 500MB/S，其速度不到 KBase 的千分之一。

3. 赋能统一数据平台

多模融合架构消除数据孤岛与架构冗余，实现标量、全文、向量、图数据等异构数据的统一存储与查询，大幅简化应用开发与运维。

4. 承载超大规模并发

基于分布式集群架构，实现卓越水平扩展能力：

- **弹性集群部署**：支持上千节点大规模集群灵活扩展；
- **秒级复杂查询**：日均处理亿级请求，复杂场景响应保持秒级；
- **持续高可用服务**：保障 7x24 小时稳定运行，作为知网核心引擎服务 3.4 万家机构；
- **动态负载扩容**：日更新文献超万篇，业务增长时无缝扩展计算存储资源；

5. 优化实时数据同步

支持读写分离架构，通过统一版本管理+增量同步机制，将数百台机器的更新操作简化为数台源端更新，大幅降低运维成本，保障 24 小时不间断数据同步与强一致性。

6. 提供智能语义检索

突破传统关键词检索局限，通过“两路召回、综合排序”技术解

决“词不达意”与“文档不匹配”问题，显著提升查全/查准率。已应用于 AI Pat+ 产品，TOP50 召回率达 36.4%，赋能 AI 专利分析等场景。

7. 高精度相似检测

满足海量文本、图像的高精度相似查找需要，如论文查重、图像检索以及段落、全文、篇名、摘要等信息的高性能实时的文本语义检索能力；支撑 AI 领域中 RAG 检索增强相关应用的需要，如 AI 生成式问答、AI 学术研究助手、AI 写作、智能医疗等。

8. 高安全性

强调安全性，支持帐号、权限管理和数据授权，用户需验证身份以保护数据安全。同时支持分布式集群、读写分离系统的安全加密帐号，采用多层次安全措施，提供可靠的数据保护。

五、应用场景

KBase 支持超大规模分布式集群计算、跨平台私有化部署，广泛应用于教育、科研、党政、军队、企业、医卫等各个专业领域的数字化信息化建设与知识管理。

行业	关键需求	应用场景
----	------	------

党政机关	• 安全保密 • 权威准确 • 版式规范 • 长期归档保存 • 国产化自主可控	▶ 公文管理与流转 ▸ 全流程加密公文流转 ▶ 政策法规库建设 ▸ 政策库建设与实时更新 ▸ 高效、精准语义检索 ▶ 调查研究数据采集与管理 ▸ 多源数据采集管理+分析 ▶ 政策研究 ▸ 基于 RAG 的智能政策解读 ▸ 资源整合，规模数据分析 ▶ 智能政府助手 ▸ 大模型+AIKBase 向量数据库
军队	• 高安全性 • 高可靠性 • 高性能 • 可扩展性 • 多源数据统一管理 • 数据完整性	▶ 内部/人事档案管理 ▸ 文本、图像、音频等向量化，跨模态检索 ▶ 情报数据采集分析 ▸ 自主研发+多级安全策略 ▸ 数据加密，保障数据安全 ▶ 智能军队教育 ▸ 知识库整合，混合检索

企业	• 低成本高效益 • 知识资产化 • 风险可控 • 多源数据统一管理	▶ 企业知识库建设 ▸ 知识资源整合、检索、更新、挖掘，知识图谱 ▶ 舆情采集分析 ▸ 舆情整合+趋势分析 ▶ 内部档案管理 ▸ 合同、人事、财务、法务等 ▶ 企业知识社区建设 ▸ RAG 增强的企业问答
科研	• 海量数据 • 数据多样性 • 高性能查询与分析 • 数据质量与完整性 • 元数据管理 • 数据安全和隐私 • 长期保存与归档 • 多源数据统一管理 • 可扩展性	▶ 科学数据管理 ▸ 原始存储、实时更新、检索、分析 ▶ 知识资源整合 ▸ 文献数据库、知识图谱 ▶ 科研助手 ▸ 大模型+AIKBase 向量数据库

医院	• 高可靠性 • 高安全性 • 海量异构数据 • 数据完整性与一致性 • 长期归档保存 • 可扩展	▶ 电子病历档案库 ▸ 结构化/半结构化统一管理 ▶ 辅助诊疗：CDSS ▸ 综合分析医学知识和患者信息 ▶ 科研与培训知识库 ▸ 文本、图像、音频等医疗数据向量化，跨模态检索
----	--	---

六、最佳实践

1. 某大学智能制造联合实验室 智能科研平台

客户痛点：多源异构数据（实验流/文献/视频）日均增量 50TB，MySQL 延迟 > 8 秒；材料热实验分析延迟 1 分钟致 23%实验重做，跨库分析失败率 > 60%；高并发频发系统崩溃，年损失超 300 万元。

解决方案：部署 KBase 替代 MySQL，分布式架构支撑 10 万+并发毫秒响应，实验流实时分析延迟降至 200ms。AIKBase 向量化关联《Nature》方案动态优化参数，实验重做率降至 6%，论文撰写效率提升 70%，设备故障风险下降 80%。

产品搭配：AIKBase 向量数据库+自建数据库+知网数据库+第三方数据库+AI 学术研究助手

2. XX 省高院法律 法律知识库

客户痛点：年处理 5000+案件，卷宗 90%为非结构化文本；检察官日均 3 小时查案例，历史案件复用率仅 8%；敏感数据禁用公有云，本地工具落后。

解决方案：私有部署 AIKBase 向量数据库作为法知大模型外挂知识库，构建本地法律知识大脑，秒级生成类案报告。

产品搭配：AIKBase 向量数据库+自建数据库+法知大模型

3. XX 重装集团 科技管理平台

客户痛点：管理超 10 万份科技数据（项目文档、标准、专利、三维图、视频等），分散在 20 余家二级单位，需实时汇总支撑 8 大装备领域研发决策。但奖项、专利、标准、论文分属独立系统，成果关联分析耗时超 2 周/次；近 27%专利未应用，创新资源匹配失误率高（>40%），成果转化滞后；二级单位数据报送延迟 3-5 天，影响国家级项目申报。

解决方案：KBase 搜索型数据库毫秒级检索专利、论文、设计图，构建 XXX 数字化科技管理平台；自主可控系统，开放 API 对接二单位系统，实时同步试验数据。

产品方案：KBase 搜索型数据库 + 国产化系统平台

4. 某军工单位 科技创新平台项目

客户痛点：管理超 500 万份高密级异构数据（三维设计图、工艺手册、试验视频等），错误率 > 35%；数据分离，技术复用率 < 15%；科技研发需实时关联内部技术沉淀（如某军事装备设计库）与外购知识资源（知网等多个文献数据库产品相关领域专辑专题）；亟需自主可控

方案管理涉密数据。

解决方案：构建科技创新平台，无缝整合知网等外购文献（200 万篇）与自建技术库（文档、专利、图纸、案例）。基于 KBase 搜索型数据库实现“设计参数→工艺标准→内外部文献”统一检索，一站式快速获取资料。采用自主可控国产数据库保障涉密资产安全。

产品搭配：KBase 搜索型数据库+自建数据库+知网数据库+第三方数据库

5. 某档案管历史资源信息化项目

客户痛点：藏有 10 万卷民国户籍胶片，数字化进程缓慢，策展高度依赖专家经验。胶片依赖手工检索，错误率超 50%；年度策展仅 2 场，市民参与度低，亟需提升馆藏资源利用效率和公众服务能力。

解决方案：采用 AIKBase 向量数据库对胶片元数据进行图像识别与理解（向量化），并整合采购的知网民国历史专题库资源。基于 KBase 搜索型数据库构建“文献-胶片-音视频-历史主题”关系图谱，实现多源异构历史资源的统一智能管理与辅助策展。

产品搭配：AIKBase 向量数据库+自建数据库+知网数据库

6. 某三甲医院 智能辅助诊疗系统

客户痛点：病历/检验/影像数据分散，分析延迟 > 5 分钟；VTE 人工筛查漏诊率 34%，预防实施率 < 40%；年发生 VTE 事件超 500 例（含致死肺栓塞）

解决方案：AIKBase 实时关联患者病史、检验值、用药记录，动态计算血栓风险。结合 CDSS 系统自动推送防治方案，高危医嘱实时拦截。

VTE 筛查效率提升 15 倍，漏诊率下降至 9%，年致死性肺栓塞减少 82%。

产品搭配：AIKBase 向量数据库+自建数据库+同方知网临床决策支持系统（CDSS）

七、团队介绍

知网团队深耕学术资源数字化三十载，从光盘出版到 AI 知识服务，持续推动技术跨越：

1. 核心引擎突破

- 自主研发 KBase 多模数据库，率先实现搜索/向量/NXD/RDF 四模融合，支持十亿级向量毫秒检索
- 创新指纹索引与智能向量索引技术，中文处理精度行业领先

2. 工程能力验证

- 分布式集群管理超千节点，日均处理亿级查询
- 复杂请求秒级响应，日更新文献超万篇，服务 3.4 万+机构

3. AI 前沿布局

- 推出 AIKBase 向量引擎，支撑 RAG 增强生成、跨模态检索等智能服务
- 构建全球最大中文学术文献库，驱动知识服务智能化转型

4. 国产化根基

- 全自主知识产权数据库，完成 35 项信创认证
- 深度适配鲲鹏/飞腾等国产芯片，满足党政军安全需求

团队使命：以自主核心技术筑牢知识基础设施，助力中国学术生态全球跃升。

八、技术认证与奖项

权威评测

- 信通院 搜索型数据库+向量数据库 双认证
- 2022-2023 信创飞越产品奖（信息技术应用创新榜）

专利壁垒

- 20 项核心技术专利（涵盖数据处理、索引管理、全文检索、数据恢复等）
- 2 项软件著作权

创新赛事荣誉

- 2024 年，同方知网 KBase 数据库团队作品“同方知网 KBase 数据库解决方案”荣获二等奖！
- 2024 年 12 月，“同方知网 KBase 数据库”项目荣获企业赛大企业赛道银奖！
- 2024 年，同方知网 KBase 数据库团队荣获龙头企业专项赛·华为鲲鹏昇腾创新专项赛特别奖！