KINGBASE 金仓数据库 | 金仓社区

杭州站

KING大咖面对面

——大模型时代的数据库平替新范式——

KINGBASE | 金仓社区

基于数据库的 LLM体系化落地

分享嘉宾：张昊

King大咖面对面-杭州站

引子 – 印巴空战的体系化启示

战机?

导弹?

飞行员?

体系化

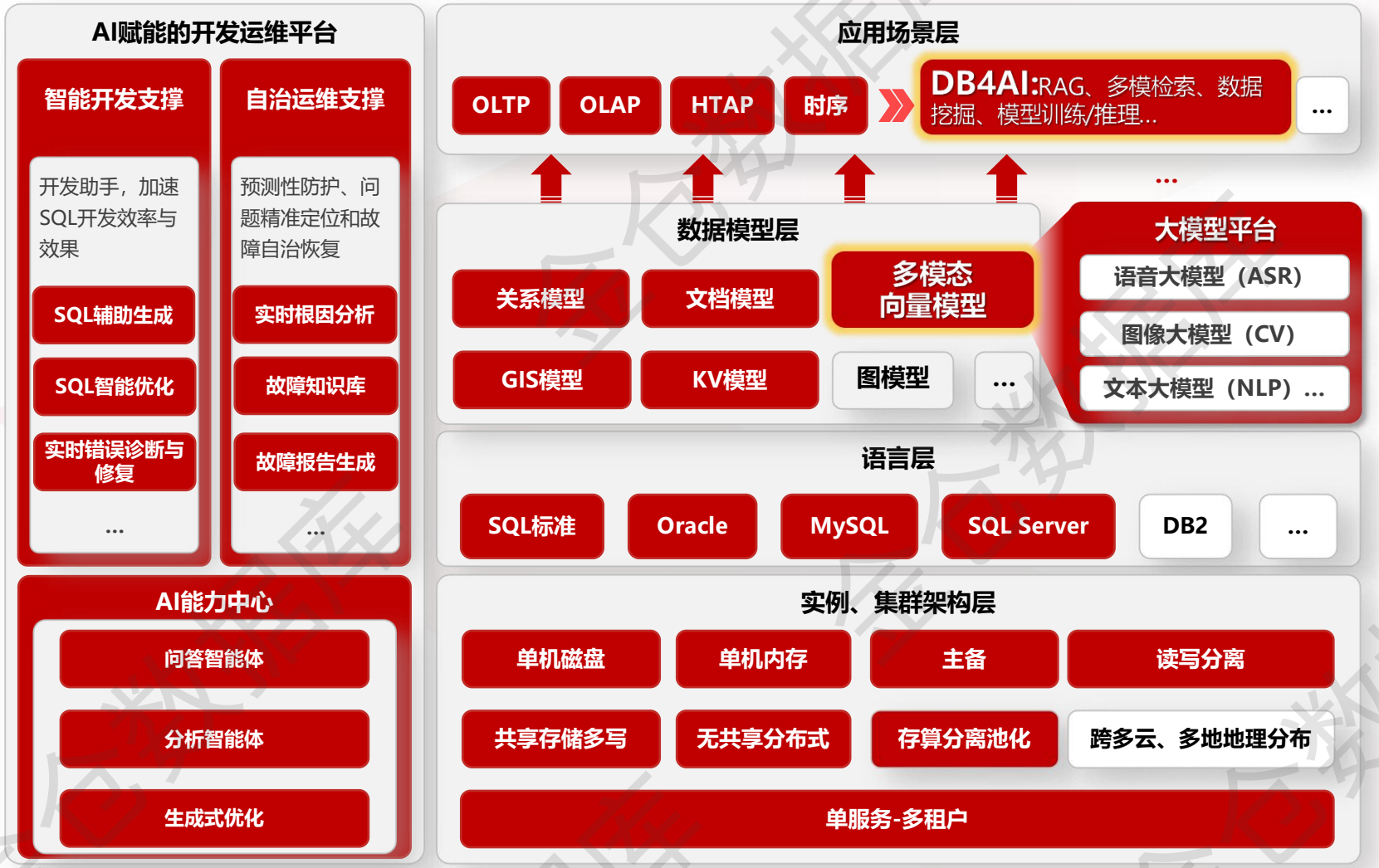
数据链

LLM落地的体系化需求

算力 + 算法 + 数据

挑战一	模型训练与推理
挑战二	数据获取与利用
挑战三	隐私与数据安全

KES – AI时代的融合数据库架构



一体化层次	一体化目标
多语法体系一体化兼容	**平替：** “0”代码修改完成应用迁移
集中分布一体化架构	多集群架构**满足不同级别的**可用/**业务连续性、性能扩展性、成本需求**，最大化投资价值
多应用场景一体化处理	满足**企业级应用与AI创新等不同场景**的数据存算需求 • 结构化数据的存算问题，在关系库框架内解决 • 库内计算，减少数据传递
多模数据一体化存储	**收敛技术栈**，降低应用复杂度和成本 减少库间数据同步，**降低同步开销**
开发运维一体化管理	**AI赋能开发运维：**基于大模型的开发与运维解决方案，显著提升工作效率

打造支撑构建人工智能系统的数据底座

AI时代的融合数据库 -支持人工智能的数据管理技术

业务领域	DB发挥的作用	
计算	内置AI算法	数据挖掘算法
		机器学习算法
	内嵌大模型	模型训练
		模型推理
存储	向量数据库	向量数据、向量索引
		相似度计算、基于相似度聚类、向量化模型比较

多模态向量数据库 – LLM体系化落地的基石

KINGBASE | 金仓社区

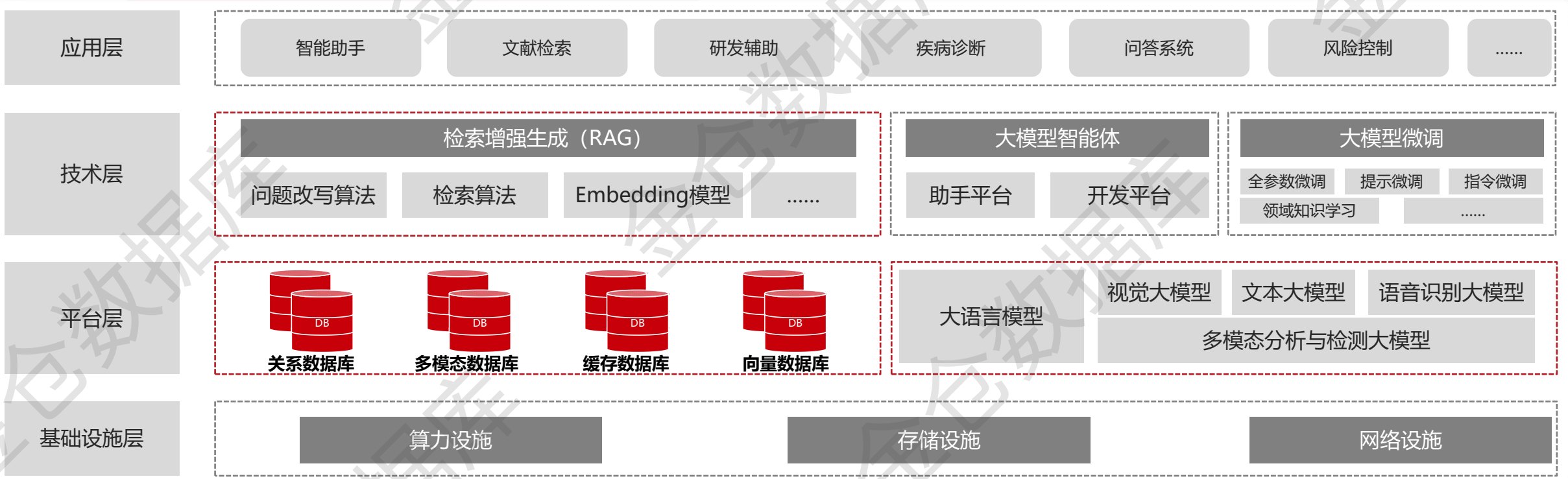
金仓向量数据库可解决大模型容易产生幻觉、知识更新不及时、数据隐私风险等问题，突破大模型在时间和空间上的限制，加速行业场景大模型落地

Deepseek从爆火到应用面临诸多挑战

- **AI“幻觉”**：LLM生成的内容与输入、上下文或已知世界不符，产生“幻觉”；
- **知识匮乏**：训练数据的范围局限、时效性等问题，无法应对时效性知识要求场景；
- **数据安全**：私域数据、问答数据、用户信息暴露在公开模型，面临严重的安全风险。

向量数据库助力AI应用实践落地

- **数据赋能**：数据赋能AI，解决大模型“幻觉”问题，让AI更靠谱；
- **实时导入**：实时更新的私域数据，确保私域知识库完备性与时效性；
- **数据安全**：纵深防御的数据库安全能力，确保AI应用不会出现数据泄露风险。



金仓向量数据库 – 关键特性

金仓向量数据库是在关系型数据库KES之上的向量功能增强，除向量数据外也提供完整的关系、文档、GIS以及时序数据的一体化处理能力，天然继承了KES的高性能、高可用、高安全、易使用、易管理等特点，**提供一站式的大模型应用开发与落地解决方案，大幅提高AI技术在企业应用落地效果，降低使用成本。**

高性能

- 支持高效存储和检索高维向量数据，实现语义搜索和相似性匹配，支持亿级向量数据毫秒级召回，索引性能领先友商20%

高安全

- 从事前的身份鉴别、访问控制，到访问过程的传输加密，到数据存储的加密、脱敏，至事后的安全审计，全栈纵深防御。

高可用

- 从实例、集群到多中心的高可用保障，确保数据零丢失、服务持续在线

高兼容

- 支持多种异构数据库的应用API级兼容，以及开源数据库的原生协议兼容，最大程度降低迁移难度

全场景支持

- 支持向量数据与关系数据、文档数据、GIS以及时序数据的混合检索，一站式全场景支持

全周期自治

- 提供数据库部署、迁移、开发、测试、调优、运维等全生命周期智能自治管理

高性能 – 数据算法算力全引擎优化

执行算法优化

- **并行执行：**Scan / Join / Append /...，多个进程并行处理一条SQL，充分利用CPU资源
- **共享执行计划缓存**，全局复用，生成更快，内存更高效
- **编译执行**，将**SQL表达式**，将**PLSQL函数及存储过程**，并编译成动态库

数据存储优化

- **基于逻辑时钟的快照优化**，有效降低MVCC版本判断时间，并支持更高并发响应。
- **热点页面固定缓冲池技术**，降低IO和封锁开销
- **WAL并行回放技术**

多核优化

- **进程绑核**，避免进程在核间飘移
- **NUMA化数据结构改造**，减少跨核内存访问
- **数据分区**，减少进程访问冲突
- **原子指令**， 减少计算开销

量化压缩优化

- **标量量化：**高精度浮点向量映射到低比特整数表示，降低空间占用
- **乘积化向量：**高维向量分解为多个低维子空间
- **量化索引结构：**在HNSW中融入量化思想，提高检索效率、降低内存占用

内外协同检索优化

- **内存与磁盘数据协同设计**，将压缩后的索引信息固化到内存，结合磁盘索引将查询目标快速定位到磁盘中的局部搜索空间，从而降低整体的内存占用需求、提高检索效率

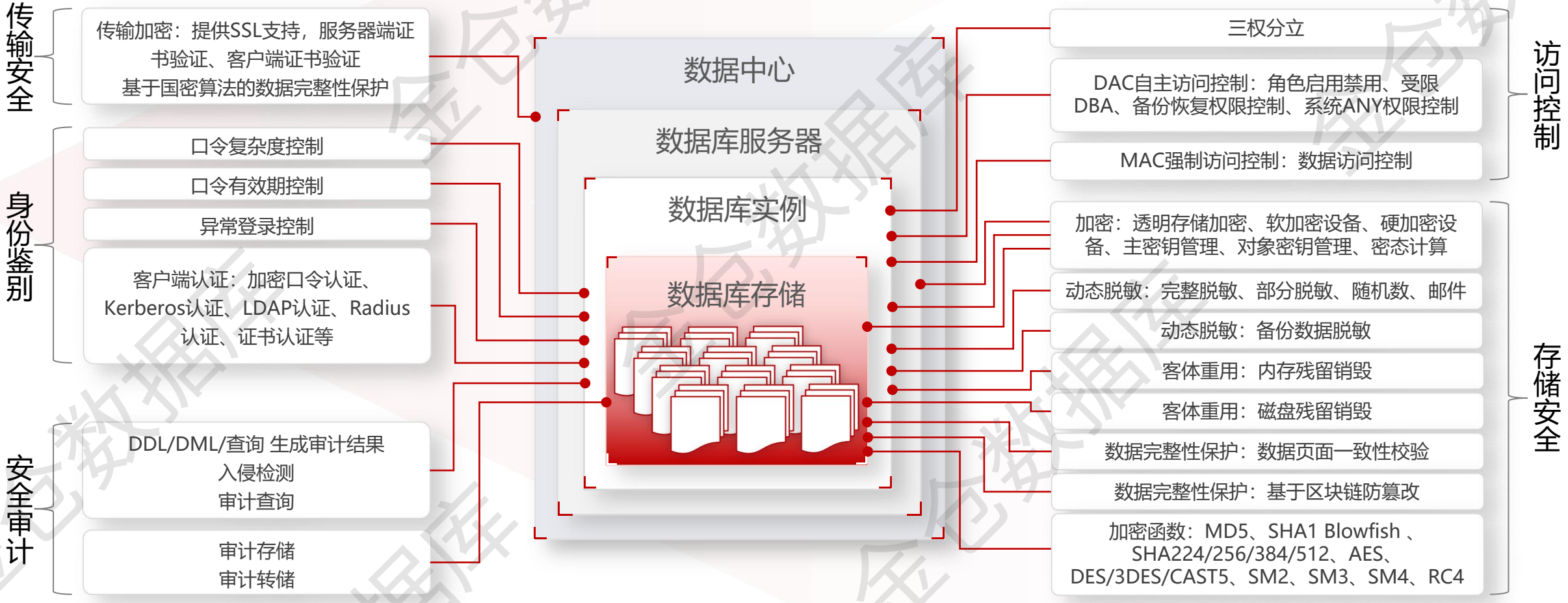
高安全 – 纵深防御，资质完备

KINGBASE | 金仓社区

信息技术产品安全分级评估证书
评估保障级4增强级（EAL4+）
国家网络与信息系统安全产品质量监督检验中心

中国国家信息安全产品认证
网络关键设备和网络安全专用产品安全认证
中国网络安全审查技术与认证中心

商用密码产品型号证书（二级）
数据库管理密码模块
国家密码管理局



高可用 - 从实例、集群到多中心的高可用保障

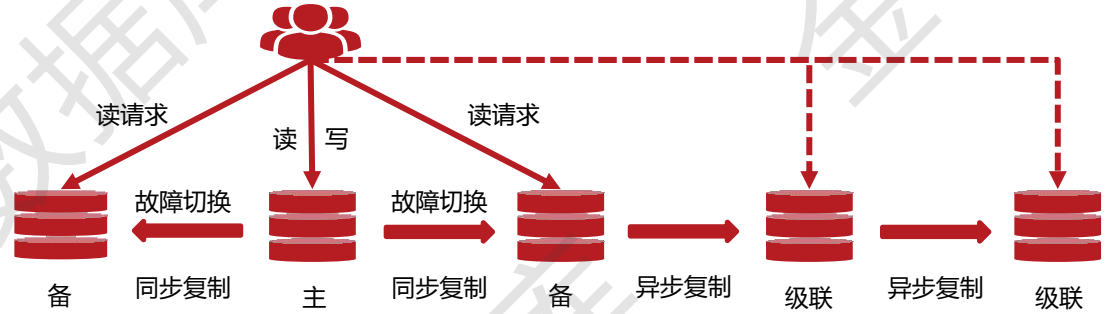
KINGBASE | 金仓社区

实例级

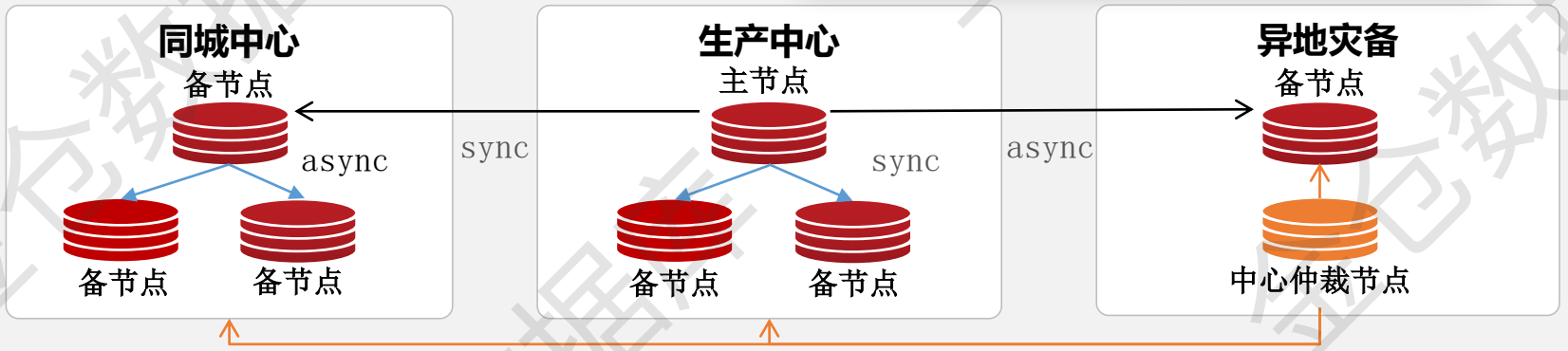
全量/增量/差异备份	全库/用户/对象备份	文本/二进制/加密输出	远程备份	备份管理
全库/事务/时刻恢复	全库/用户/对象还原	增量检查点	指定对象/时刻闪回	第三方备份平台集成
日志管理	数据一致性校验	控制文件多副本	进程级运行监控	故障进程自动重启

集群级

- 基于物理日志的全同步复制，数据不丢失
- 备节点支持异步复制，不阻塞事务，业务持续可用
- 备机日志回放并行加速，故障快速切换
- 增强型一致性协议选主，避免脑裂
- 节点故障恢复后自动回归，保持集群规模稳定



中心级



- 同城双中心物理全同步，确保数据“零”丢失，故障“秒” 切换
- 跨城市物理日志高速复制，最优可实现秒级RPO
- 中心级多数派一致性切换，避免中心级双主

高兼容 – 多种异构数据库原生协议兼容

KINGBASE | 金仓社区

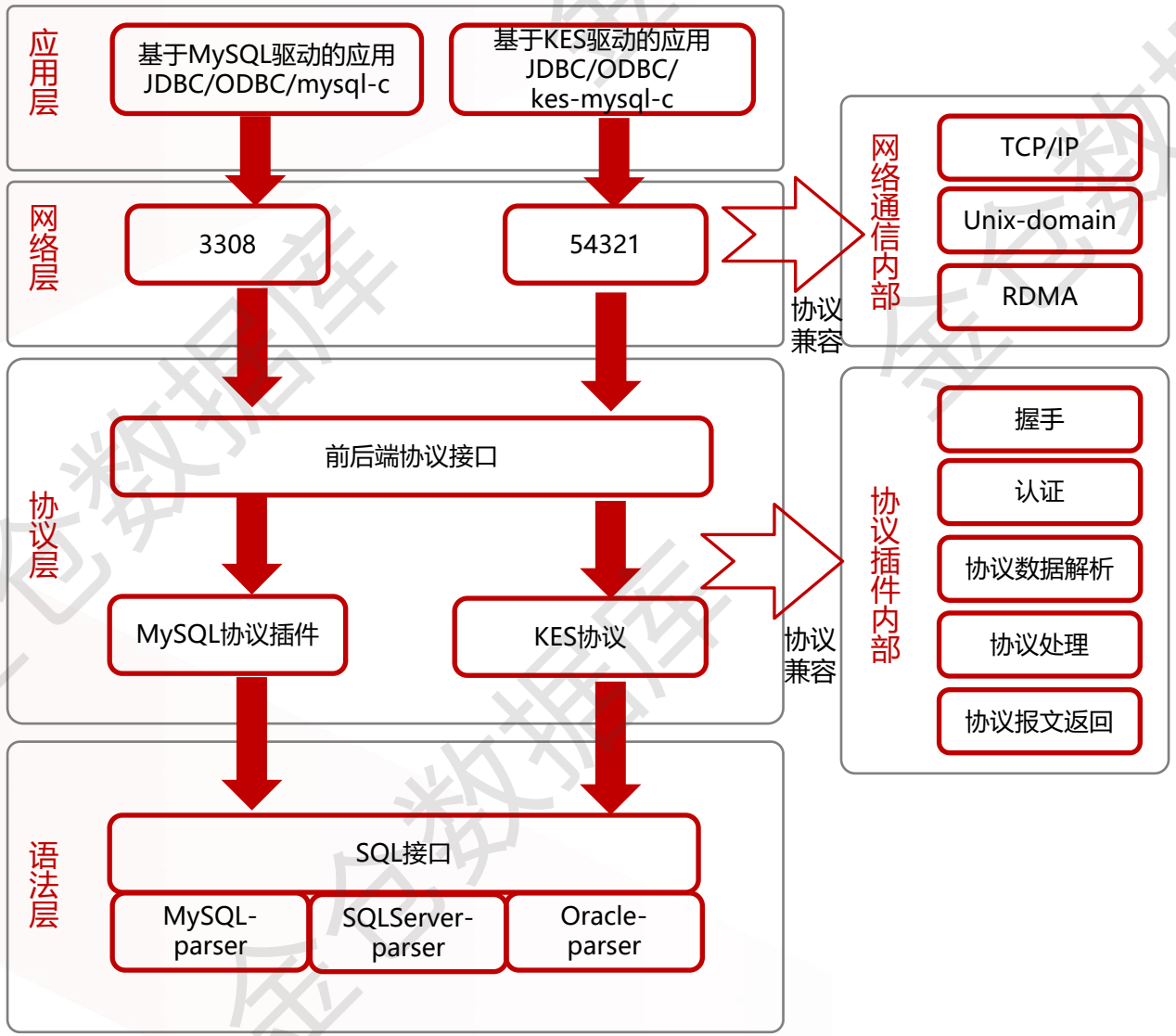
应用API级兼容

- 优势
  - ✓ 方案成熟
  - ✓ 适合协议不开源的产品，如兼容Oracle 应用API
- 缺点
  - ✓ 不同语言框架的API种类众多，尤其各个数据库都有非标准API
  - ✓ 需要应用替换驱动库和连接串等。

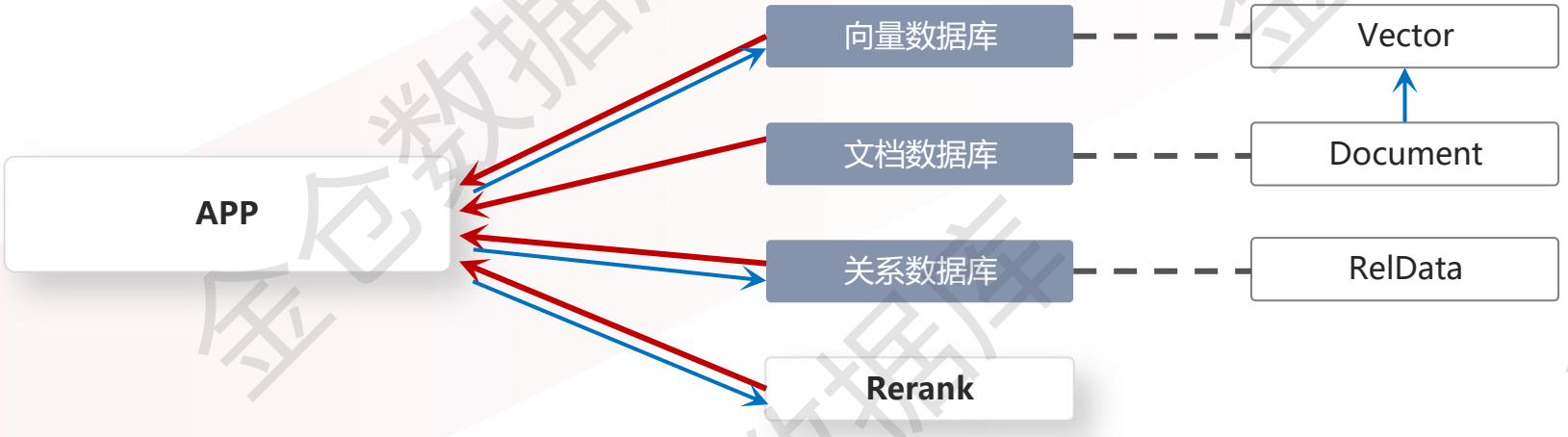
原生协议兼容

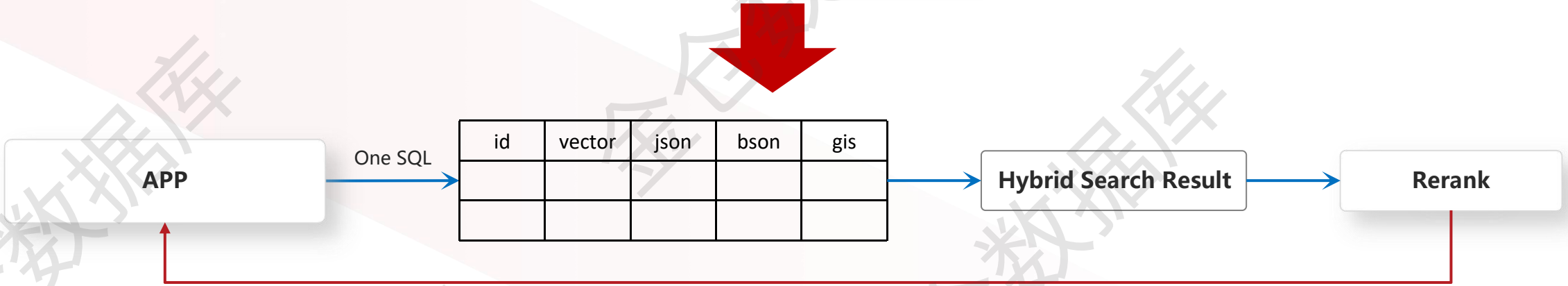
- 上层使用统一的协议接口，内部协议插件提供不同兼容协议的转换，达到一个库多种协议的效果。
- MySQL协议插件：
  - ✓ 使用原生MySQL客户端驱动直接连接KES
  - ✓ 使用原生MySQL客户端程序直接连接KES

保留所有MySQL生态



全场景 – 强大的混合检索能力





全周期自治 – 运维管控一体化平台

KINGBASE | 金仓社区

展示层

统一页面访问

统一服务接口

统一服务层

安装部署	集中管理	备份与恢复	监控与告警	健康巡检	性能诊断分析
服务器/安装介质	集群/实例管理	全量/增量/差异备份	全局概览监控	健康巡检任务管理	性能指标采集
安装环境检测	集群扩/缩容	自动备份策略	集群/实例/性能	健康/自定义巡检	数据库时间分析
集群/实例部署	主备节点切换	备份集校验/还原	慢/高频SQL监控	健康评分	TOP SQL分析
灾备部署	数据库会话管理	库/实例/表级备份	数据/资源监控	健康画像	锁与等待分析
批量安装/纳管	库/实例级参数设置	按时间点还原	运行/审计日志	健康巡检报告	压力负载分析
参数智能优化	补丁/license更新	还原至新实例	告警提醒/管理	健康档案管理	性能优化建议

标准RestFul接口
监控数据
告警数据
邮件通知接口
即时通讯接口

兼容监控平台接口
监控数据
告警数据

安全审计			
审计规则配置	审计日志	异常行为分析	运维操作审计

系统管理		
用户角色管理	功能权限管理	数据库权限管理

安全保障体系

服务支撑层

基础组件	
任务调度	资源管理
消息通知	指标管理
资源分发	告警阈值
模型管理	……

监控/巡检指标体系
数据库指标
集群指标
性能指标
……

基于数据库时间的决策树模型

健康评估模型

性能诊断模型

日志分析

规则引擎

……

采集层

节点资源数据采集	实例运行数据采集	集群运行数据采集	数据库性能指标采样	数据库/审计日志采集	……

数据库服务器 智能代理 →

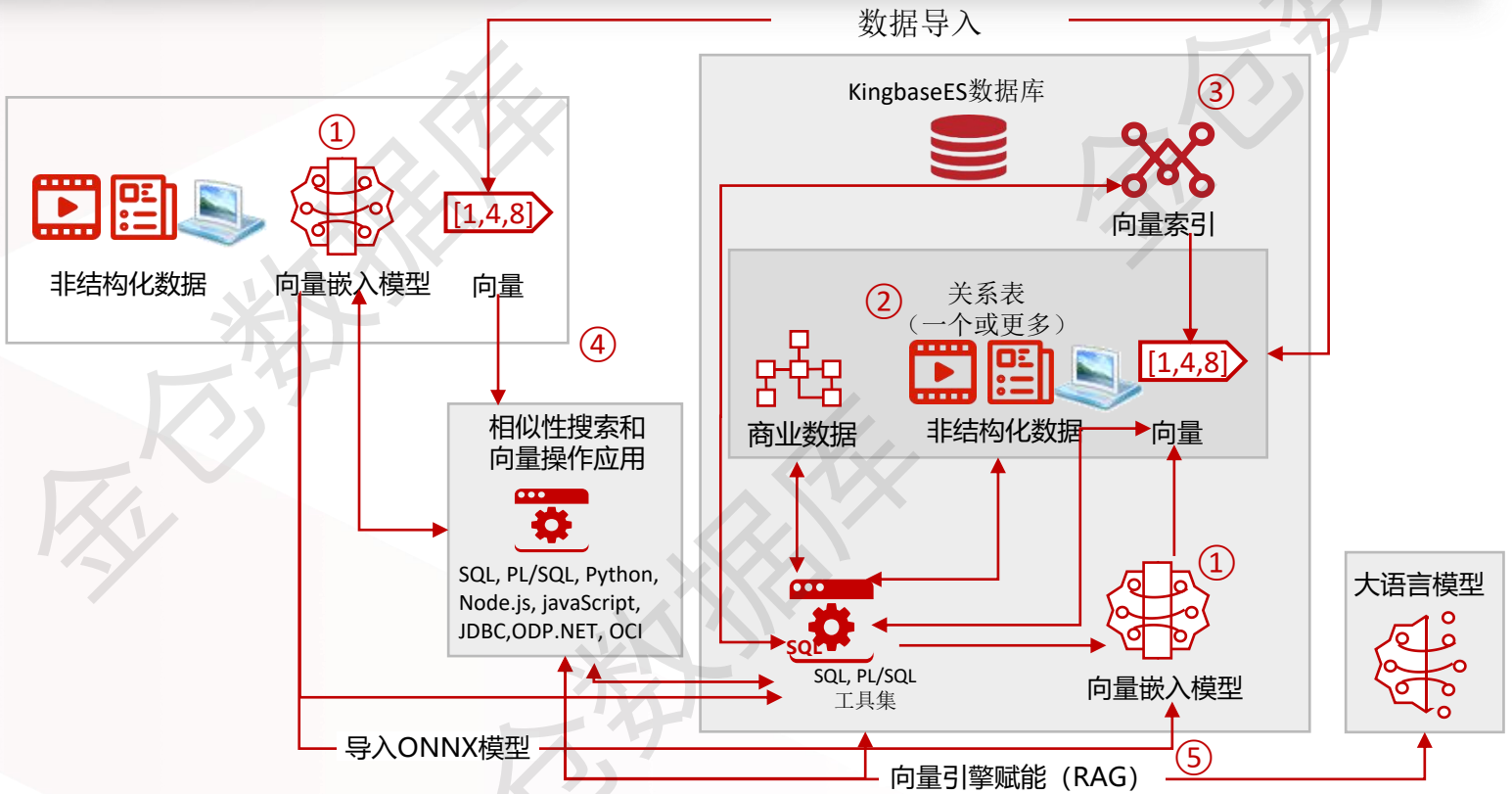
数据库服务器 智能代理 →

数据库服务器 智能代理 →

RAG赋能AI应用实践

RAG将传统信息检索系统的优势与生成式大语言模型 (LLM) 的功能相结合。通过将已知的数据和知识与LLM语言技能相结合，使能AI应用的输出更准确、更及时。

步骤	完成的任务		
检索	从知识库中检索到的问答对	**问题向量化**	将用户输入查询问题转换成向量
		相似度检索	在向量数据库中检索与问题向量最相似的知识库片段
		结果排序	按相似度得分对检索到结果排序，选择最相关片段作为后续生成的输入
增强	增强LLM的提示词（prompt）		
生成	LLM拿着增强后的Prompt生成问题答案		

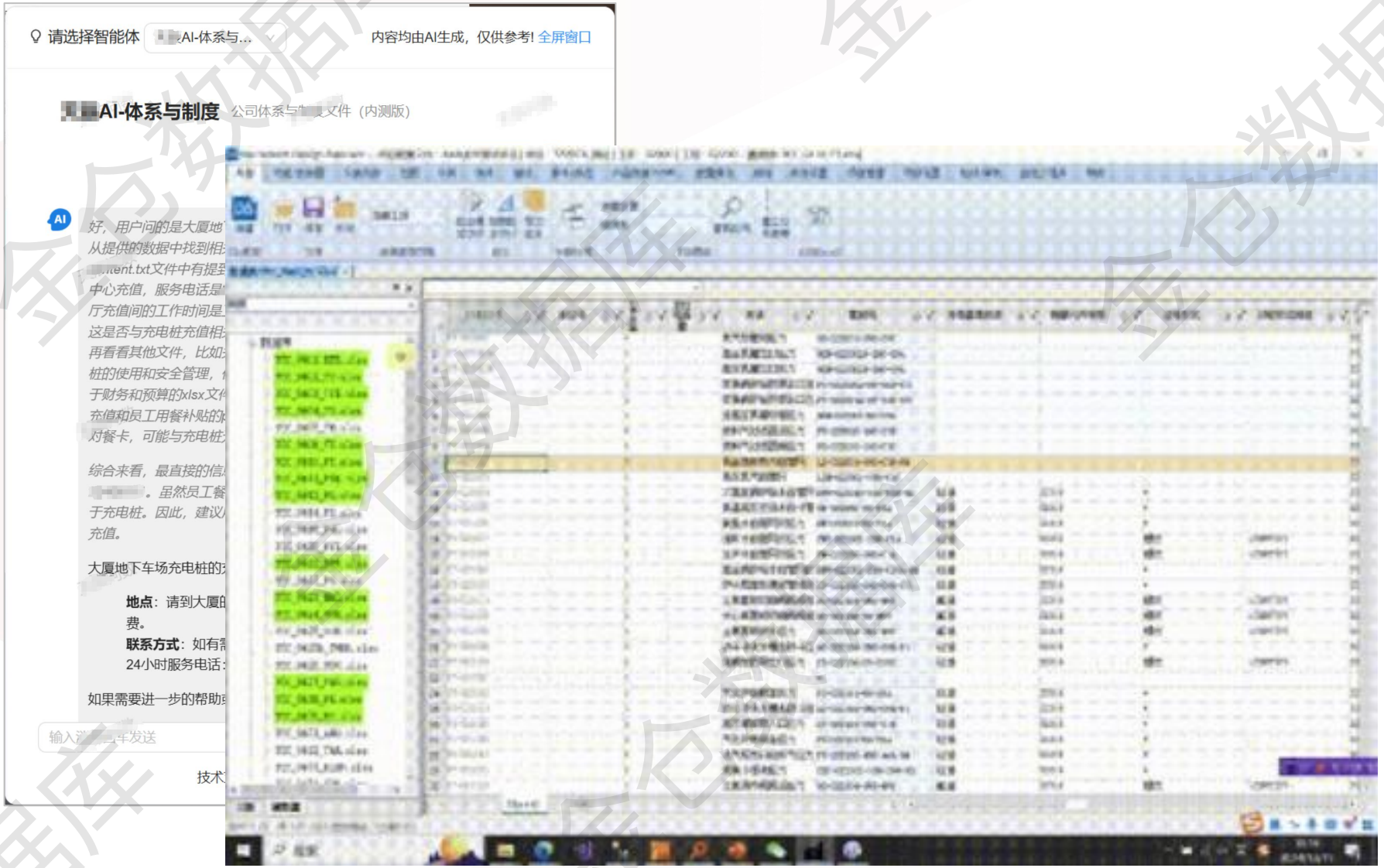


RAG场景 – 支撑某央企智能体全面落地

KINGBASE | 金仓社区

落地实践

- 某央企通过KES向量数据库构建知识库，支持多种格式文档快速导入，助力OA智能问答、AI开发平台等智能体上线，为员工提供高效、准确的信息检索、知识推荐和设计开发辅助。
- AI智能体的上线显著提高该企业的数据资产利用率，践行AI赋能企业高质量发展。



创新应用 - 多模态融合检索支撑城市管理新探索

KINGBASE | 金仓社区

典型场景需求

查询某市过去30天内与雨天+分支道路+追尾形式类似的交通事故频发（至少3次）的5个热点区域

返回结果包含

- 事故数量
- 事故详情
- 最早和最后发生时间
- 与预期特征的平均相似度

关键数据模型

- **某市热点区域**：固定地理范围，以及细化区块/网格 -> GIS数据模型
- **雨天+分支道路+追尾形式**：相似度检索 -> 向量数据模型
- **区域内各事故详情**：文档聚合 -> 文档数据模型

```
WITH city_boundary AS (
  SELECT geom AS city_geom
  FROM areas
  WHERE name = '某市'
  LIMIT 1
),
filtered_accidents AS (
  SELECT
    a.accident_id,
    a.accident_time,
    a.location,
    a.details,
    a.embedding,
    ar.area_id,
    ar.name AS area_name,
    a.embedding <=> '[1,2,3]'::vector AS vector_distance
  FROM accidents a
    CROSS JOIN city_boundary
    JOIN areas ar ON ST_Contains(ar.geom, a.location)
  WHERE ST_Within(a.location, city_boundary.city_geom)
    AND a.accident_time >= NOW() - INTERVAL '30 days'
    AND a.embedding <=> '[1,2,3]'::vector < 0.5
)
SELECT
  fa.area_id,
  fa.area_name,
  COUNT(*) AS accident_count,
  json_agg(fa.details) AS aggregated_details,
  MIN(fa.accident_time) AS first_accident,
  MAX(fa.accident_time) AS last_accident,
  AVG(fa.vector_distance) AS avg_vector_distance
FROM filtered_accidents fa
GROUP BY fa.area_id, fa.area_name
HAVING COUNT(*) > 3
ORDER BY accident_count DESC
LIMIT 5;
```

向量类型为 vector，3 维表示

- 事故类型：1（追尾）2（侧撞）3（翻车）4（...）
- 天气情况：1（晴天）2（雨天）3（雪天）4（...）
- 道路类型：1（高速）2（城市干道）3（分支道路）4（...）

GIS计算
明确事故所在区域

向量计算
事故特征与[1,2,3] 的欧氏距离

文档处理
分区域聚合事故详情清单

分组排序
分组计算热点区域排序

KINGBASE | 金仓社区

THANKS

成为世界卓越的数据库产品与服务提供商