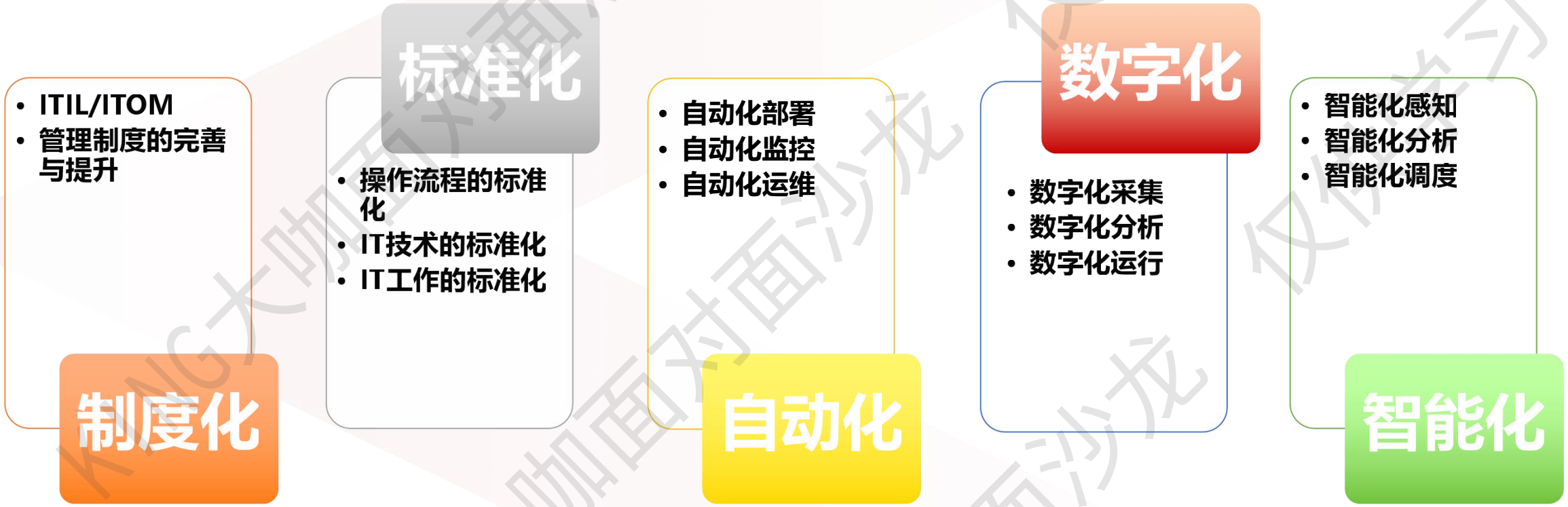
KINGBASE | 金仓社区

数据库可观测性与智能化运维

分享嘉宾：白鳝
佰晟智算(深圳)技术有限公司CEO

KING大咖面对面-厦门站

从制度化到智能化的演进路线



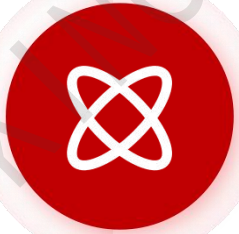
国产数据库AIOPS的技术难点

模型本地化部署的算力成本问题

30B中等规模模型可以实现准确分析，纯国产ARM服务器无GPU可实现诊断推理

RAG召回率不足的隐患

知识图谱替代向量嵌入，同样问题多次问答数据一致。通过数据预处理替大模型矫正容易出问题的数据



MCP工具太多，128K上下文不够用

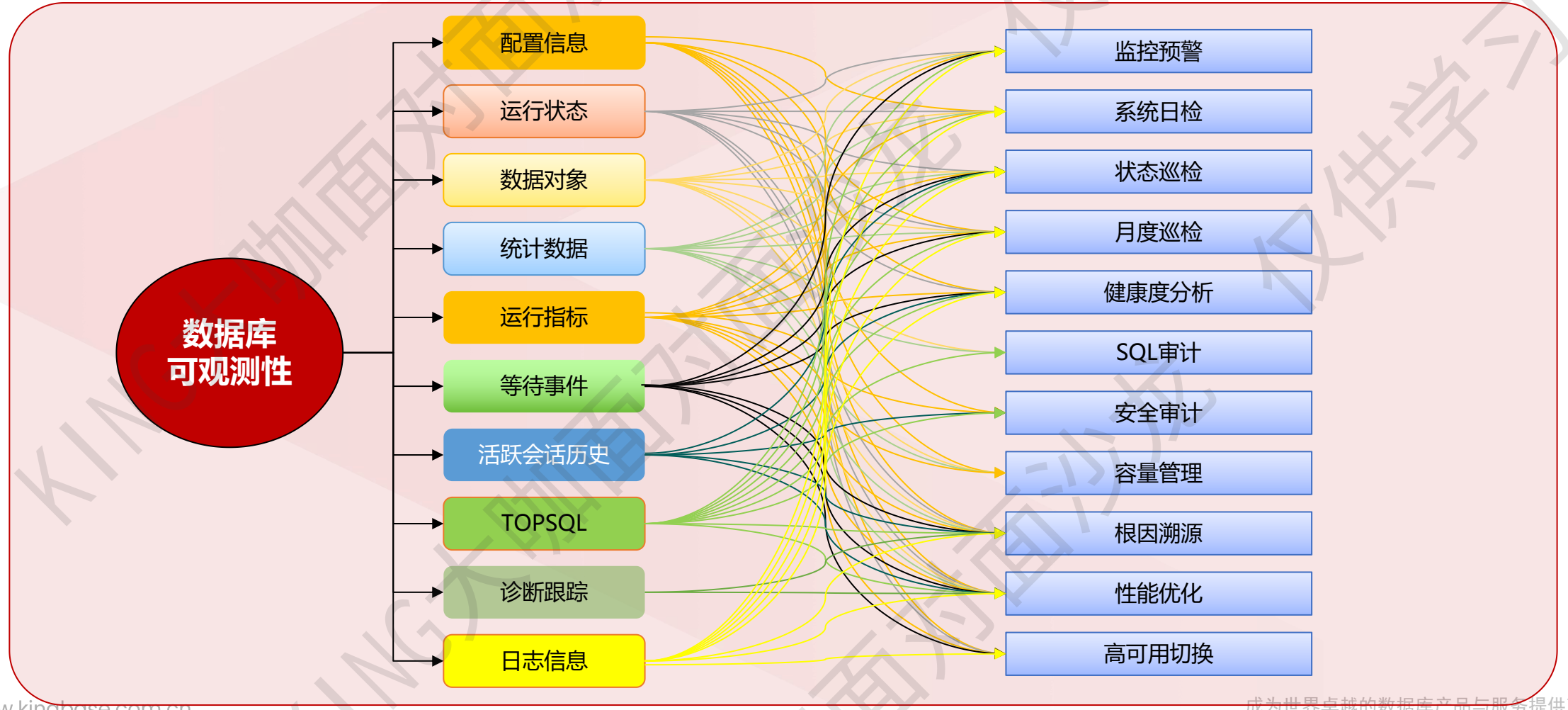
独有的多意图识别算法，自动多轮筛选单元化工具功能，通过编排实现少量工具组合覆盖多场景

未命中的CASE幻觉严重

强大的数字化底座作为支撑，没有数字化就没有智能化。上下文学习与运维知识图谱是天然好搭档

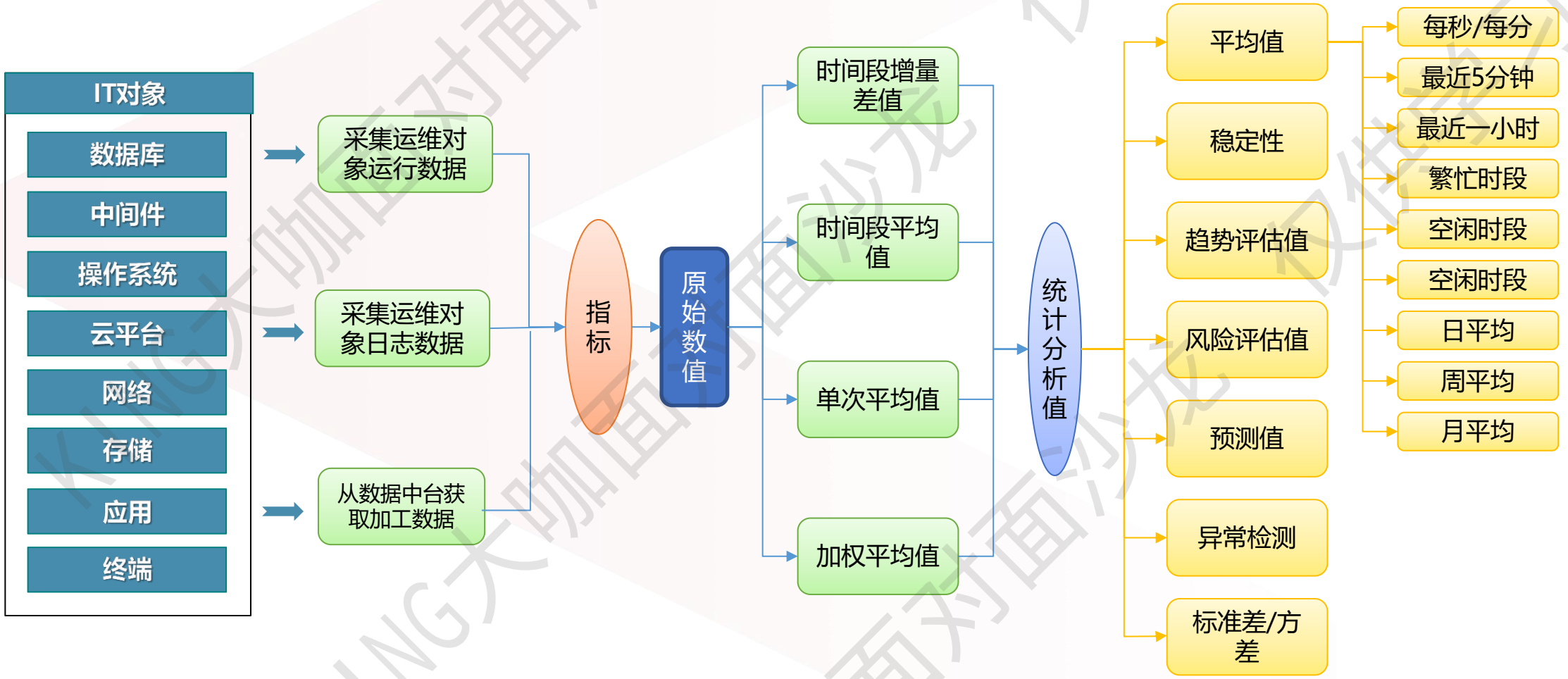
数字化底座-数据来源

AIOps的核心能力来自于数据库的可观测性能力。利用数据库、操作系统等提供的可观测性接口，AIOps使用低开销、低风险的采集方法采集大量的数据库指标，并通过AIOps内置的算法引擎进行深度加工，形成指标、模型等。为自动化运维功能提供基础能力。



数字化底座-数据深度加工

未经指标化的状态数据是无法被用于预警、诊断、巡检等工作的，获得原始状态数据仅仅是指标采集与加工的第一步，AIOps的指标大多数并不是从数据库中采集回来的原始值，而是经过加工计算后的。一切数据库的状态、容量、性能、风险等最终都会被指标化，指标化有利于今后数字化分析的开展。



数字化底座-运维模型

运行模型描述运维对象的运行状态，包括健康模型、性能模型、负载模型、容量模型、故障模型，使用运维对象的指标数据来构建状态模型生成运行状态评价指标。模型分专家模型和智能模型，实现对运维对象的数字化评价，并可实现超越专家思维的泛化分析。

健康模型

原始数据
- 原始数据包括：原始指标、日志分析结果、日检结果等数据

指标集
- 将原始数据指标化，建立运维对象指标集

关键指标
- 分析总结出用于模型的关键指标

指标维度分解
- 对关键指标进行分类

设置评价标准
- 对指标按维度设置评价规则

健康模型
- 汇总各个维度评分形成健康模型

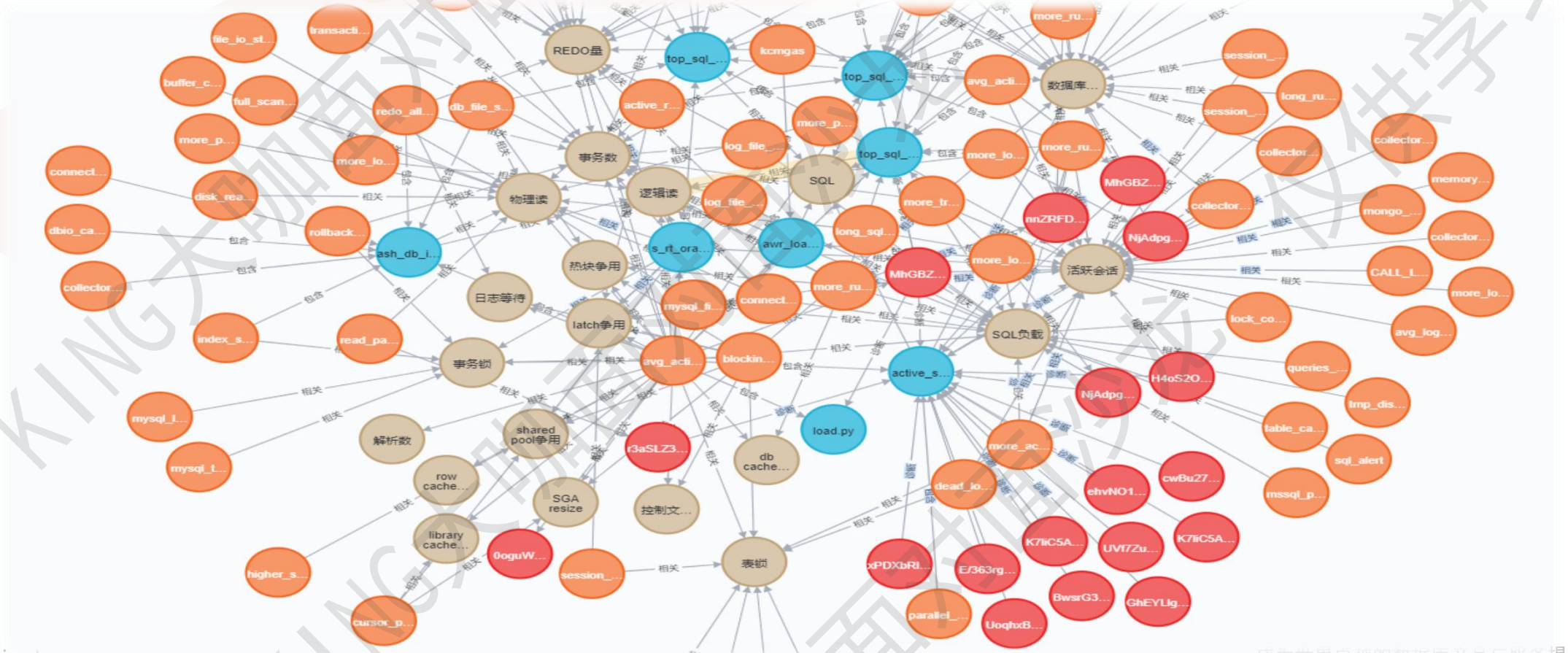
泛化分析模型（机器学习、深度学习算法模型）

异常检测	动态基线	波动检测	波动态势
	异常检测		
关联性分析	影响分析	临近发现	诊断路径泛化
	根因推理		
问题归纳	告警收敛		
	归纳分析		

数字化底座-运维知识图谱

KINGBASE | 金仓社区

运维知识图谱建设是数字化能力的基础。通过知识梳理形成了初始化的运维知识图谱，并根据实际应用案例不断提炼和丰富知识图谱，使其分析能力不断提升。基于运维知识图谱可以实现智能化推理，获得超越专家的能力



核心技术-基于上下文学习的推理框架

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam
Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss
Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh
Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

Abstract

We demonstrate that scaling up language models greatly improves task-agnostic,few-shot performance, sometimes even becoming competitive with prior state-of_x0002_the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non_x0002_sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model.GPT-3 achieves strong performance on many NLP datasets, including translation,question-answering, and cloze tasks. We also identify some datasets where GPT-3’ s few-shot learning still struggles.

核心技术- DBAIOps：一种基于知识图谱的推理LLM增强型数据库操作与维护系统

DBAIOps: A Reasoning LLM-Enhanced Database Operation and Maintenance System using Knowledge Graphs

Wei Zhou
Shanghai Jiao Tong University
weizhoudb@sjtu.edu.cn

Peng Sun
Baisheng (Shenzhen) Technology Co., Ltd.
sunpeng@dbaiops.com

Xuanhe Zhou
Shanghai Jiao Tong University
zhouxh@cs.sjtu.edu.cn

Qianglei Zang
Baisheng (Shenzhen) Technology Co., Ltd.
zangqianglei@dbaiops.com

Ji Xu
Baisheng (Shenzhen) Technology Co., Ltd.
xuji@dbaiops.com

Tieying Zhang
Bytedance
tieying.zhang@bytedance.com

Guoliang Li
Tsinghua University
liguoliang@tsinghua.edu.cn

Fan Wu
Shanghai Jiao Tong University
fwu@cs.sjtu.edu.cn

ABSTRACT

The operation and maintenance (O&M) of database systems is critical to ensuring system availability and performance, typically requiring expert experience (e.g., identifying metric-to-anomaly relations) for effective diagnosis and recovery. However, existing automatic database O&M methods, including commercial products, cannot effectively utilize expert experience. On the one hand, rule-based methods only support basic O&M tasks (e.g., metric-based anomaly detection), which are mostly numerical equations and cannot effectively incorporate literal O&M experience (e.g., troubleshooting guidance in manuals). On the other hand, LLM-based methods, which retrieve fragmented information (e.g., standard documents + RAG), often generate inaccurate or generic results.

To address these limitations, we present DBAIOps, a novel hybrid database O&M system that *combines reasoning LLMs with knowledge graphs* to achieve DBA-style diagnosis. First, DBAIOps introduces a heterogeneous graph model for representing the diagnosis experience, and proposes a semi-automatic graph construction algorithm to build that graph from thousands of documents. Second, DBAIOps develops a collection of (800+) reusable anomaly models that identify both directly alerted metrics and implicitly correlated experience and metrics. Third, for any given anomaly, DBAIOps employs an automatic graph evolution mechanism that explores the relevant paths over the graph and dynamically explores potential gaps (missing paths) without human intervention. Based on the explored diagnosis paths, DBAIOps leverages reasoning LLM (e.g., DeepSeek-R1) that inputs the relevant pathways, identifies root causes, and generates clear diagnosis reports for both DBAs and common users. Our evaluation over four mainstream database

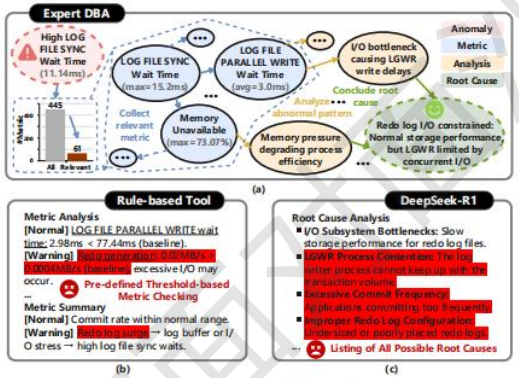


Figure 1: Automatic database O&M is challenging - (a) Expert DBA needs to analyze diverse information from triggered anomalies. (b) Empirical O&M may apply misleading rules (caused by incorrect thresholds). (c) LLMs may lack O&M experience and fail to diagnose *even with necessary abnormal information like relevant metrics.*

(e.g., achieving 99.99% four nines availability with less than 52.6 minutes of downtime per year for critical services such as financial and e-commerce systems [28]) and performance (e.g., service-level agreements (SLAs) enforced by cloud service providers [11, 22, 45]). For instance, the NOTAM database outage (an honest mistake that cost the country millions) resulted in over 10,000 flight delays and more than 1,300 cancellations [9, 14].

- 佰晟联合上海交大、清华大学、字节跳动等基于佰晟智算的BIC-IA系统实践发表的一篇论文
- 包含了BIC-IA实现国产数据库AIOPS的核心技术路线
- 提出了一种以运维知识图谱+上下文学习的AI推理新算法

核心技术-基于上下文学习的智能分析

- DeepSeek-R1
- Qwen3
- Kimi
- …

基础大模型的推理能力

- 官方文档
- 专家知识
- 管理规范
- 历史案例

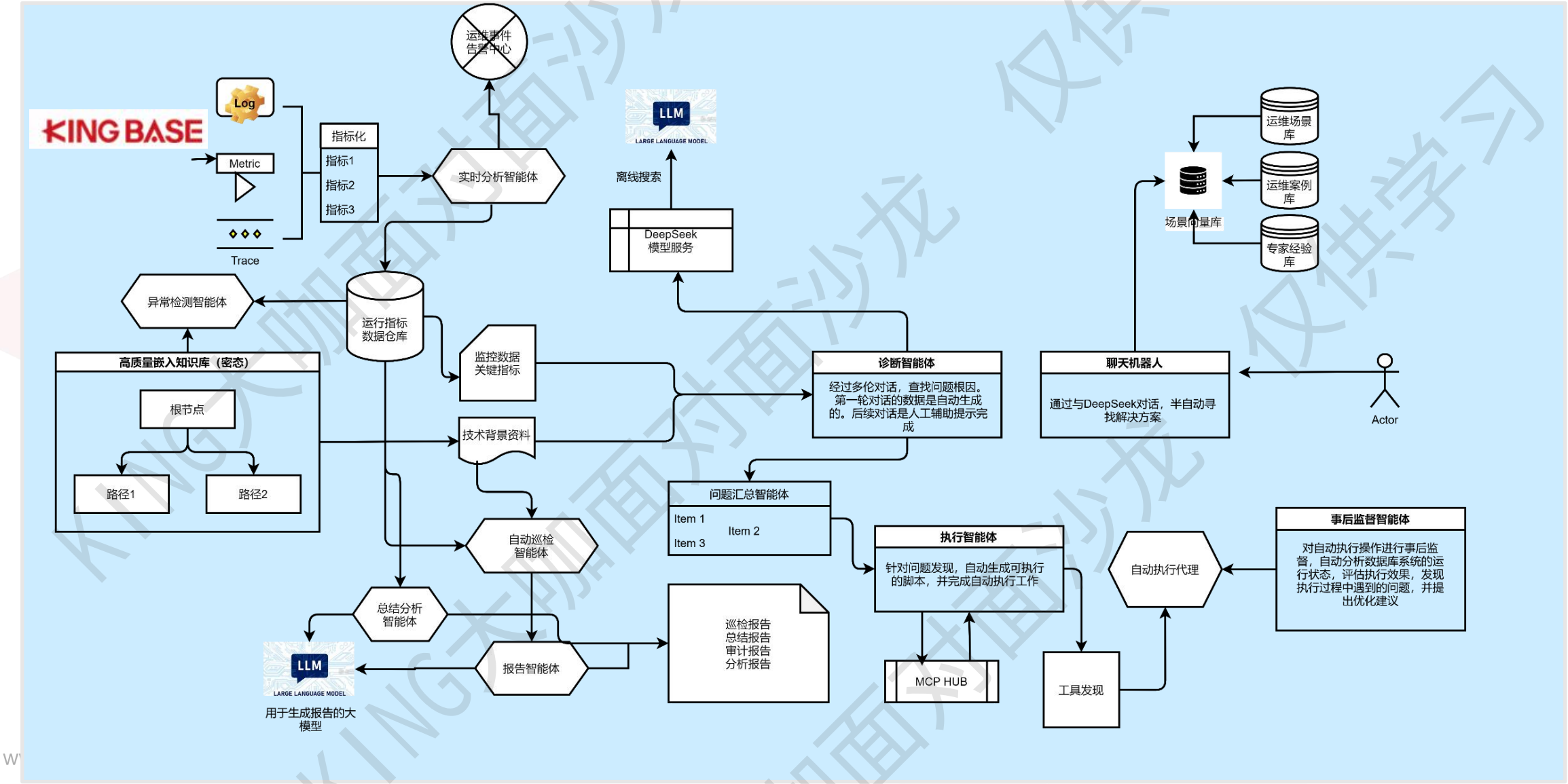
覆盖推理的基础语料

- 指标数据
- 日志提取
- 跟踪采样
- 应用探针

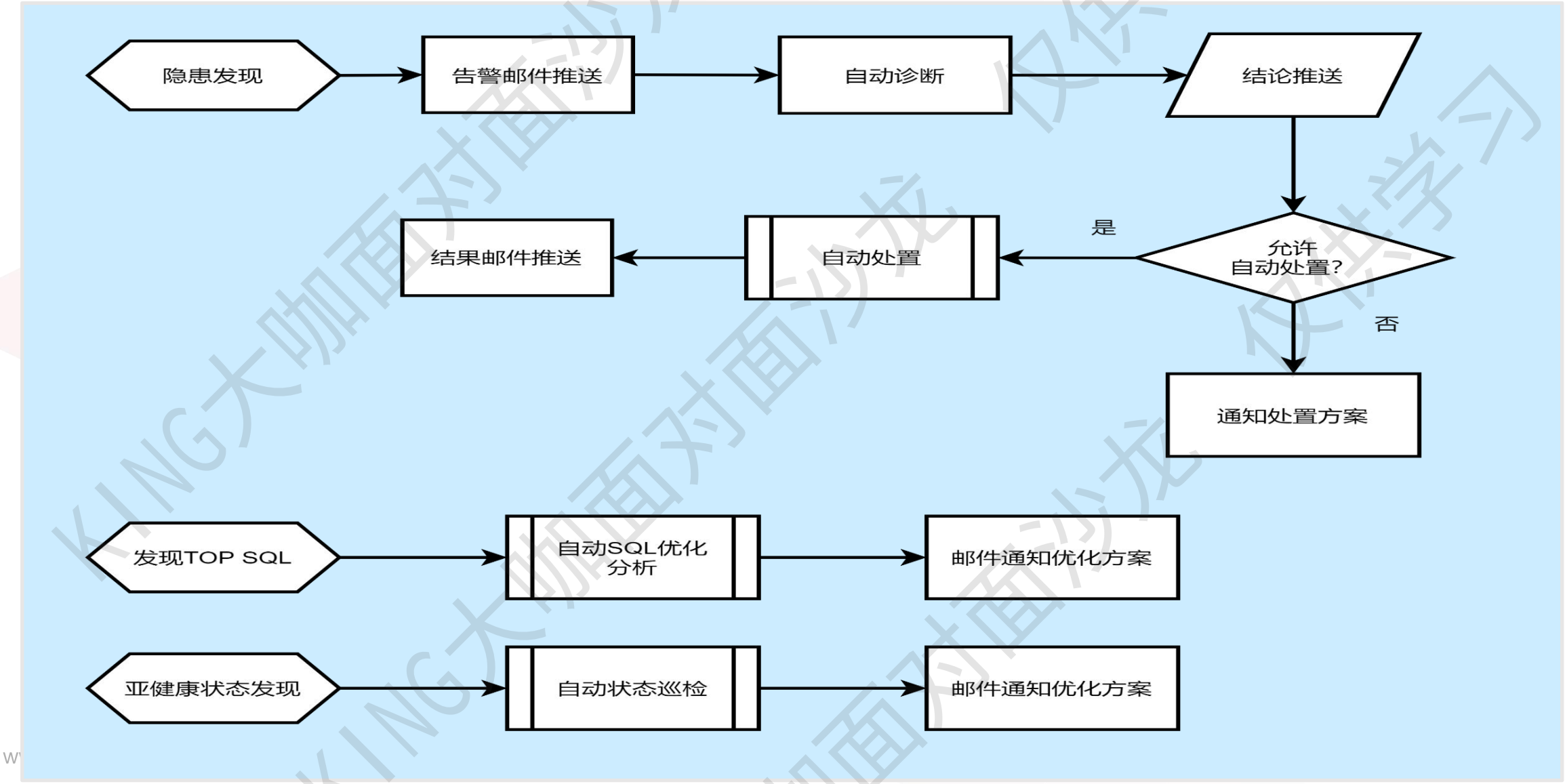
构成推理逻辑证据链的数据

高质量的推理

核心技术-多智能体协同


运维事件
告警中心
Log
KINGBASE
Metric
Trace
指标化
指标1
指标2
指标3
实时分析智能体
LLM
LARGE LANGUAGE MODEL
离线搜索
DeepSeek
模型服务
运维场景
库
运维案例
库
专家经验
库
场景向量库
异常检测智能体
运行指标
数据仓库
监控数据
关键指标
高质量嵌入知识库（密态）
根节点
路径1
路径2
技术背景资料
诊断智能体
经过多伦对话，查找问题根因。第一轮对话的数据是自动生成的。后续对话是人工辅助提示完成
聊天机器人
通过与DeepSeek对话，半自动寻找解决方案
Actor
问题汇总智能体
Item 1
Item 2
Item 3
执行智能体
针对问题发现，自动生成可执行的脚本，并完成自动执行工作
自动巡检
智能体
总结分析
智能体
报告智能体
巡检报告
总结报告
审计报告
分析报告
MCP HUB
工具发现
自动执行代理
事后监督智能体
对自动执行操作进行事后监督，自动分析数据库系统的运行状态，评估执行效果，发现执行过程中遇到的问题，并提出优化建议
LLM
LARGE LANGUAGE MODEL
用于生成报告的大模型

颠覆性理念-无屏运维



实践效果：超越专家的能力

针对一些典型场景的测试中，基于大模型的诊断分析的准确性已经达到甚至超过了专业运维人员在使用专家提供的诊断工具的帮助下的结果。说明基于大模型的诊断推理水平已经具备实战能力。

数据库	故障模型	QWQ:32B	DeepSeek-r1:671b	DeepSeek-r1:32b	运维专家+传统工具
Oracle	日志同步延时异常	100%	100%	90%	100%
	活跃日志组过多	100%	100%	70%	70%
	热块冲突	100%	80%	80%	70%
	逻辑读异常增长	90%	100%	100%	95%
	活跃会话数过多	90%	80%	80%	60%
MYSQL	活跃会话过多	100%	100%	100%	90%
	CPU异常增长	100%	100%	90%	30%
KingBaseES	backend进程大量刷脏	100%	100%	100%	80%
	大量全表扫描	100%	100%	100%	90%

BIC-QA：覆盖绝大多数关键信创数据库

BIC-QA公共知识库版本已经于8月18日上线，目前共计155万条Q/A和65万条文本条目，支持17种数据库，包括Oracle、SQL SERVER等商用数据库，MySQL/PostgreSQL等开源数据库，以及达梦、金仓、Oceanbase、GaussDB等信创数据库。

BIC-QA：公共知识库与商用版

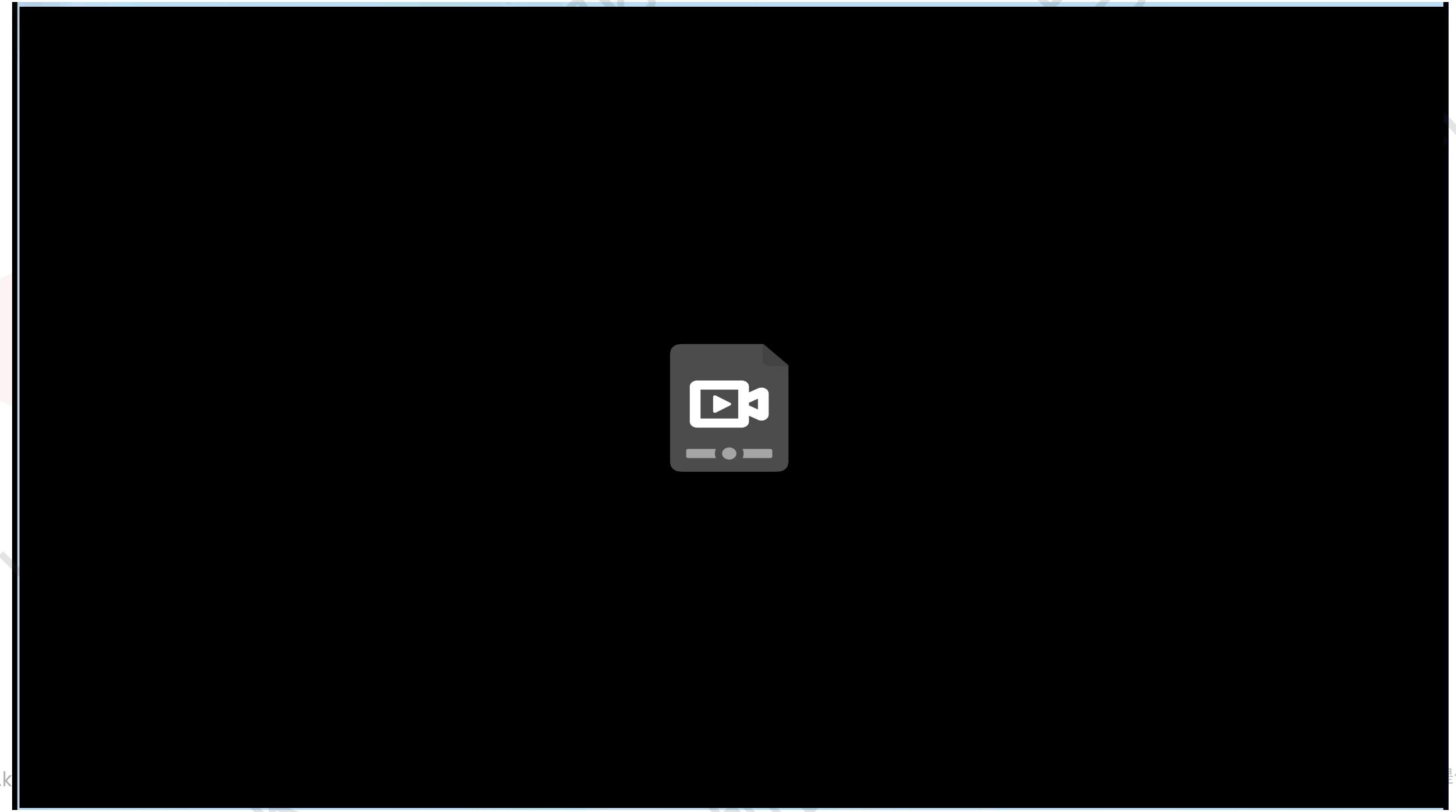
公共知识库

- 向所有DBA免费开放的公共知识库
- 通过开源的浏览器插件提供服务（https://gitee.com/BIC-QA/bic-qa）
- 与各大数据库厂商密切合作，随时更新最新的运维知识与文档

商用版

- 在企业内部离线部署
- 分为公共知识库和私有知识库两部分，企业可自建私有知识库（不限于数据库知识）
- 公共知识库可定期增量更新到商用版

小睿智能运维助手-展示



佰晟智算，做AI时代的践行者！

KINGBASE | 金仓社区

THANKS

成为世界卓越的数据库产品与服务提供商