

云原生数据库向 数芯智能体的演进

分享嘉宾：唐成

KING大咖面对面-杭州站

数据库演讲路线



什么是云原生数据库?

01

两种定义

1. 能跑在各种公有云、私有云上的数据库
2. 完全为云环境开发的数据库

02

云从来都“云里雾里”，无明确的边界。
云原生是一种广告语



存算分离架构

计算和存储解耦、分布式存储、资源灵活调配：可根据业务需求单独扩展计算或存储资源



弹性伸缩能力

Serverless高弹性、按需付费，应对业务峰值：如大促期间业务量激增时可快速扩容，结束后自动缩容



云原生技术栈

支持Docker/Kubernetes容器化部署、微服务架构、多租户隔离



极致资源利用率

一台物理机上运行了数百个数据库实例，实例可以在不同的物理机之间进行漂移

理解这些特点需要有云里雾里的觉悟，不能较真



高可用与可靠性

单机、单AZ、多AZ、多Region等多级容灾，主流云原生数据库可提供99.95%-99.995%的服务可用性，海量分布式智能存储，保障数据安全



生态兼容性

100%兼容MySQL、PostgreSQL等开源数据库，部分产品兼容Oracle语法，可无缝迁移现有应用到云原生数据库



自动化与智能化

利用云的自动化能力，降低运维成本，提供完善的监控、日志、追踪能力，实现智能优化与自治运维



全球化支持

跨地域部署，适应多云环境，满足不同地区的数据合规要求。

AI增强数据库 (AI-Enhanced Database) 是指将人工智能（特别是机器学习和深度学习）技术深度集成到数据库内核、管理流程或交互界面中的新一代数据库系统

01

AI增强数据库（三个层面的革新）

- 1.内核级智能（性能与优化）：智能索引与查询优化、自适应资源管理、异常检测与自愈
- 2.运维与管理自动化（DBA的助手）：自动调优、安全增强、容量规划
- 3.交互与应用赋能（开发者的新能力）：原生向量搜索、自然语言交互（Text-to-SQL）、库内推理、多模型支持

02

典型代表产品（2025-2026视角）

PostgreSQL+pgvector：通过插件pgvector补齐了多模查询的短板

Oracle AI Database (26ai)：强调 “AI for Data”，将向量搜索、图神经网络和机器学习直接融入关系型引擎，支持构建无代码AI Agent

Microsoft SQL Server 2025：引入了原生的AI模型管理功能，支持在T-SQL中直接调用外部大模型，并增强了向量搜索能力

AI原生数据库 (AI-Native Database) 是“AI数据库”的另一种形态。如果说“AI增强数据库”是在传统架构上加装“AI引擎”，那么AI原生数据库则是**号称**在设计之初就为了适应人工智能（特别是大语言模型和向量计算），但在国内是小众

AI时代是结构化数据、文本、向量数据并存的时代

定义与架构
差异

AI增强数据库

AI增强数据库：以关系型模型（行/列）和向量搜索、机器学习推理为双核心，既可以服务于传统的报表和事务处理（OLTP/OLAP），也可以平滑服务AI业务场景

AI原生数据库

以非结构化数据和向量为第一公民（First-class Citizen）。号称是为AI而生。对结构化数据处理可能存在功能缺失。

都在索引结构如HNSW、IVF-PQ做深层优化

四大关键特征

多模态数据统一存储 (Unified Multimodal Storage)

同时处理文本、图像、音频、视频等多种格式。直接存储原始数据及其对应的向量嵌入 (Embeddings)，并在同一引擎内实现结构化数据与非结构化向量的混合查询 (Hybrid Search)。

场景：你可以同时查询“价格低于100元”（结构化）且“看起来像复古风格”（非结构化向量相似度）的商品。

语义理解优先

传统数据库基于关键词匹配 (Keyword Match)，而AI原生数据库基于语义相似度 (Semantic Similarity)。

它理解“苹果”在水果语境和科技语境下的不同含义。

查询不再依赖精确的 `WHERE name = '...'`，而是基于意图的 `SIMILAR_TO('...')`。

为大模型记忆而生 (Memory for LLMs)

它是构建RAG（检索增强生成）系统的最佳后端。

长上下文管理：能够高效存储和检索海量的对话历史、文档片段，作为大模型的“长期记忆”。

低延迟高并发：针对向量检索进行了极致的性能优化，确保在大模型生成回答前，能在毫秒级时间内从亿级数据中召回最相关的上下文。

动态演化与自适应 (Dynamic Evolution)

随着大模型版本的迭代或业务数据的变化，数据的向量表示可能需要重新计算 (Re-embedding)。

AI原生数据库通常内置了更灵活的数据流水线，支持在线重索引和模型版本的热切换，而无需停机迁移数据。

AI增强 vs. AI原生：如何选择？

传统认知：

特性	AI增强数据库 (AI-Enhanced)	AI原生数据库 (AI-Native)
核心架构	关系型/文档型为主，向量是附加功能	向量/多模态为核心，可能兼容SQL
主要优势	事务一致性(ACID)、生态成熟、存量系统兼容	极致的向量检索性能、语义理解、大模型集成度
适用场景	传统企业核心系统升级、混合负载、复杂事务处理	大模型应用(RAG)、推荐系统、图像/视频搜索、知识库问答
数据范式	表格 (Rows/Columns)	向量 (Vectors) + 元数据
典型代表	Oracle 26ai, SQL Server 2025, PostgreSQL (pgvector)	Milvus, Pinecone, Weaviate, Qdrant

实际情况:AI原生数据库能做的事情,AI增强数据库都能做,反之则不行

很多时候考虑是其他方面的事情，在轻量级的场景下用AI原生数据库？

核心定义：数据即芯片

Data-Chip Agent：颠覆传统“硬件承载软件”的层级，提出未来视角：**数据本身就是逻辑、记忆和能力的固化形式。**

传统计算：芯片(物理载体) + 数据(流动信息)

智能体时代：数据结构决定“算力”与“智商”

没有数据，智能体就是空壳

制造类比：数据加工即造芯

数据不仅是血液，更是智能体的**骨骼、神经与硅片。**

输入数据 = 晶圆制造

训练微调 = 光刻制程

推理应用 = 芯片运行

反馈学习 = 工艺迭代

AI时代就是在造数字芯片



未来AI竞争：制造数字芯能力的比拼

表面是模型的竞争，实则是“**数据加工与治理**”能力的竞争。谁能制造出更纯净、更致密、更逻辑自治的“数据芯片”，谁就能打造出最强大的智能体。

为什么数据就是“芯片”？



逻辑的固化 (Logic Solidification)

传统芯片通过光刻刻电路，智能体通过数据将规律“刻”在权重和向量库中。数据结构的复杂度，等同于晶体管的密度。



能力的边界 (Capability Boundary)

芯片性能由架构决定，智能体能力由数据决定。数据越丰富精准，“核心频率”越高，处理能力越强。数据是构建大脑的砖瓦。



不可篡改的“固件” (Immutable Firmware)

对于某些专用智能体，高质量的领域数据（如法律条文、医疗图谱）一旦被内化，就构成了它的底层操作系统，就像烧录在芯片里的固件一样，定义了它“是谁”以及“能做什么”。

特性对比：物理芯片 vs 数芯

特性	物理芯片 (硅基)	数芯 (数据基)
制造	光刻/蚀刻 (耗时昂贵)	清洗/训练 (迭代快)
升级	更换硬件	热更新 (注入数据)
并行	受限于核心数	无限分割处理
磨损	老化、发热	越用越“聪明”
定制	开模成本极高	千人千面 (专属生成)

制造晶圆 (数据清洗、结构化)

旧视角

数据入库(ETL)

01

02

Data-Chip 视角

Data-Chip 视角:晶圆制备。原始数据是沙硅,必须经过极致的清洗、对齐和结构化,才能成为制造“逻辑芯片”的高纯度基材。杂质(脏数据)会导致芯片短路(幻觉/错误决策)

01

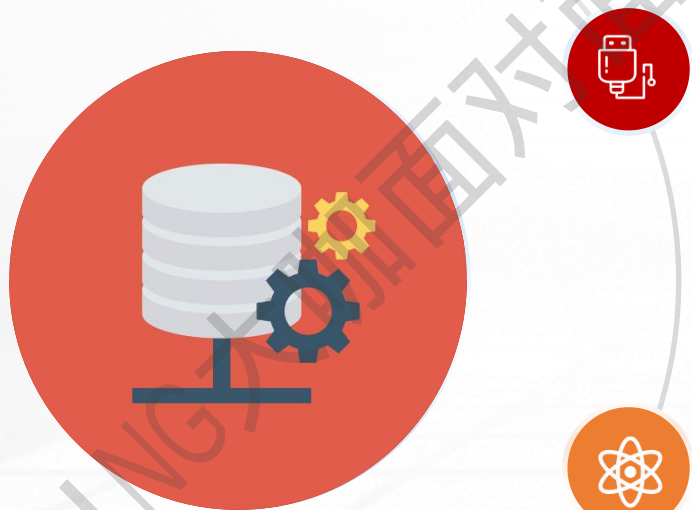
旧视角

建立索引和关联

02

Data-Chip 视角:电路光刻

- ★ 向量嵌入:将语义信息“刻”入高维空间,形成神经连接电路
- ★ 知识图谱:将业务逻辑“刻”为实体关系网,形成逻辑门电路
- ★ 这一步决定了这枚“数芯”的晶体管密度(知识密度)和运算效率



旧视角

模型训练与接口开发

Data-Chip 视角:模型微调与RAG组装

- 1.将特定的领域数据(如医疗法规、金融风控规则)封装成独立的 Function Chip(功能芯片)
- 2.通过 RAG 技术,将这些“数据芯片”动态挂载到主模型上,实现“热插拔”升级。需要法律能力?插上“法律数芯”;需要代码能力?换上“编程数芯”

- 一条8GT/s的QPI的单向带宽: $8\text{GT/s} \times 16\text{bit}/8 = 16\text{GB/s}$
- 一条9.6GT/s的QPI的单向带宽: $9.6\text{GT/s} \times 16\text{bit}/8 = 19.2\text{GB/s}$
- 一条10.4GT/s的UPI的单向带宽: $10.4\text{GT/s} \times 16\text{bit}/8 = 20.8\text{GB/s}$



01

旧视角:微服务调用

1. 毫秒级别、最多万兆带宽

Data-Chip 视角:SoC (System on Chip) 架构

- 1.Data-Chip Agent 不再是一个单一的大模型,而是一个由无数“数据芯片”组成的片上系统。极高带宽、微秒延迟
- 2.感知芯片处理输入,记忆芯片提供上下文,逻辑芯片执行推理,行动芯片输出结果。它们在同一极高速低延迟数据总线上高速协同

02

- NVLink 1.0有4条Link, 每个Link有8个Lan, 单向带宽是: $20\text{GT/s} \times 4\text{Link} \times 8\text{Lan}/8 = 80\text{GB/s}$
- NVLink 2.0有6条Link, 每个Link有8个Lan, 单向带宽是: $25\text{GT/s} \times 6\text{Link} \times 8\text{Lan}/8 = 150\text{GB/s}$

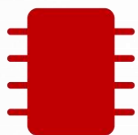
GPU比CPU厉害的点就是 互联带宽远远超过CPU。CXL、NVLink、InfiniBand流行的原因

旧视角:版本更新



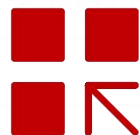
Data-Chip 视角:制程工艺迭代

随着新数据的流入,原有的“电路”被优化,新的“晶体管”被生成
Data-Chip Agent 能够自我重绘电路图,实现“越用制程越先进”的进化效果



理念变革：存算一体

- 打破传统架构中数据存储与计算引擎的分离。
- 数据包含计算逻辑，不再是被动对象，而是驱动智能体运行的核心。



动态重构能力

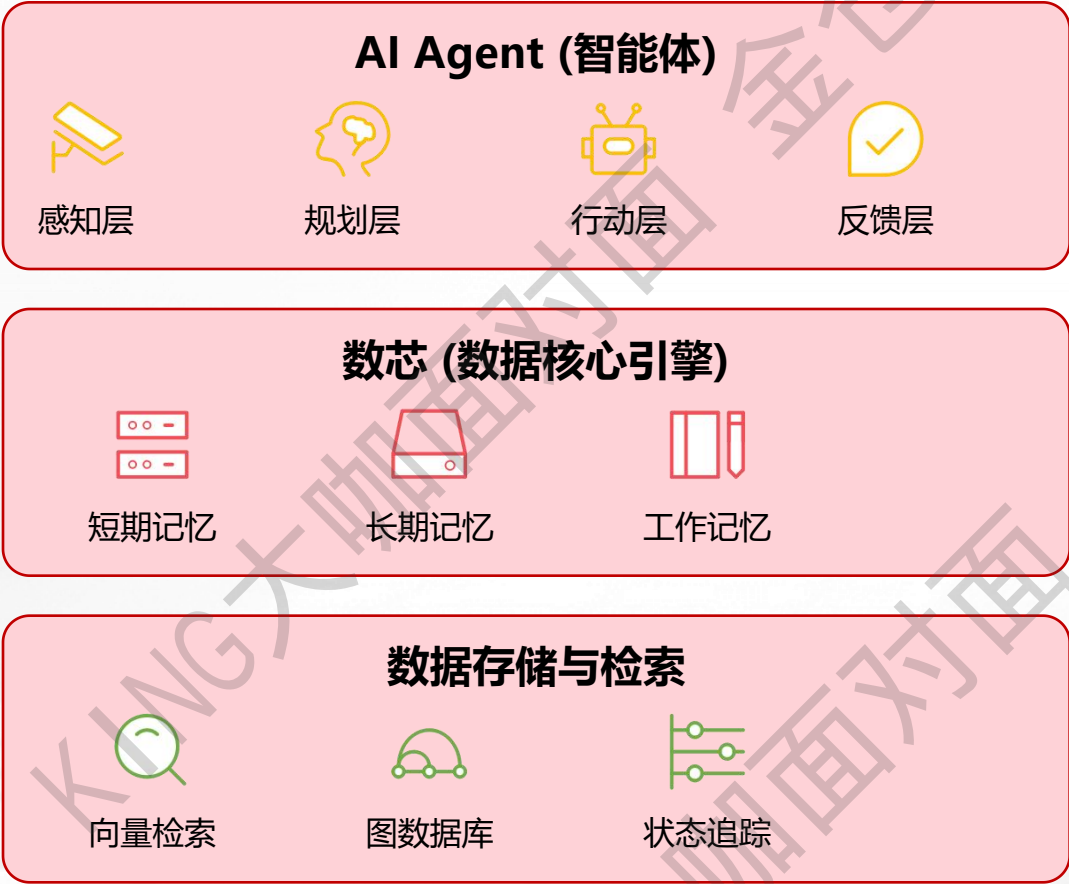
- 区别于物理芯片的固定电路，智能体可实时重组。
- 根据任务场景（如数学、编程）即时切换形态，成为流动的芯片工厂。



知识的“光刻”

- 将大模型能力压缩或海量数据提炼为高质量知识库。
- 微观层面的精密“光刻”过程，制造更高效的数芯，持续提升性能。

总结：数芯智能体通过“存算一体”、“动态重构”与“知识光刻”，实现了从静态硬件到动态智能系统的进化。



记忆类型特性对比

类型	存储内容	延迟	方案
短期记忆	当前对话上下文	< 10ms	内存/Redis
长期记忆	知识库, 经验	< 50ms	向量数据库
工作记忆	任务执行状态	< 20ms	图/时序库

数芯智能体：以数据为核心引擎，驱动智能决策与执行

3大强调

01

强调“逻辑即数据”：芯片的本质是电路逻辑。叫“Chip”暗示了数据本身已经包含了执行逻辑,而不仅仅是原材料。

02

强调“模块化与组合”：芯片是可以组合的。这意味着我们可以像搭积木一样,通过组合不同的“数据芯片”来快速构建垂直领域的超级智能体。

03

强调“高性能与低延迟”：芯片追求极致效率。这暗示了未来的智能体将把计算尽可能下沉到数据层(In-Database AI),减少数据搬运,实现芯片级的响应速度。

核心技术栈

**GraphRAG**
知识关联推理引擎

**A2A 协议**
智能体通信协作标准

**MCP 协议**
工具调用标准化接口

**ContextExp**
长时记忆优化算法

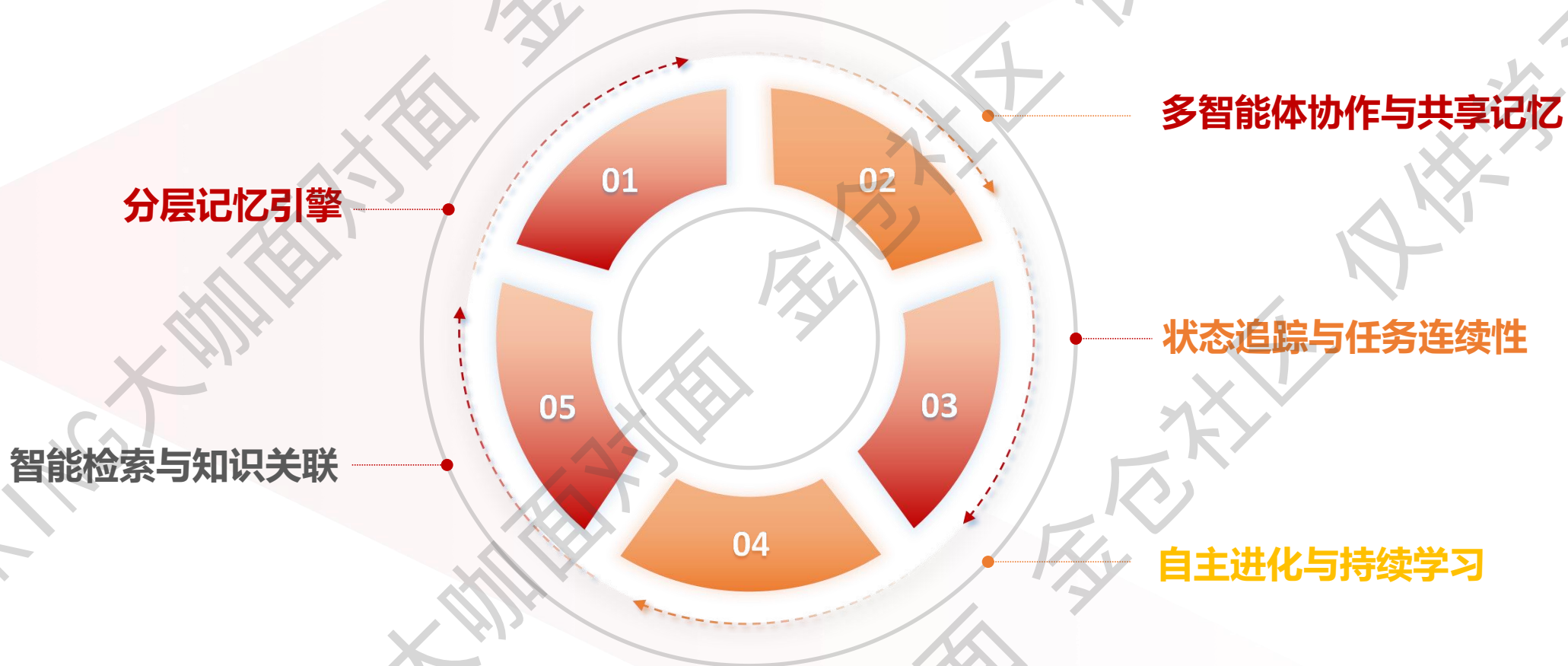
关键技术突破演进

领域	云原生时代	AI赋能时代	数芯智能体
记忆管理	无	向量检索	三叠记忆体系
状态追踪	事务日志	单态状态	GraphRAG图
知识检索	SQL查询	语义搜索	RAG+向量+图
自主进化	无	有限学习	持续反馈优化

方案优势对比

维度	传统 AI Agent	数芯智能体方案
速度	慢 (响应延迟高)	快 (引擎级优化)
记忆架构	外挂多个数据库	统一数芯引擎
状态管理	应用各自维护	引擎统一主责
知识检索	向量库+KG分离	多模统一检索
协作协议	复杂定制通信	内置A2A标准协议
架构复杂度	高 (4-5个组件)	低 (1个统一引擎)
运维成本	高 (多组件维护)	降低 60%
任务连贯性	单次对话记忆	数周级长时记忆

数芯智能体的5大核心功能



挑战与争议

类型	具体问题	解决思路
技术成熟度	2026年仍处早期阶段	优先选择成熟技术
标准化缺失	A2A/MCP协议尚未统一	持续关注开源社区进展
数据安全	记忆包含敏感信息风险	数据加密 + 严格访问控制
人才缺口	AI+数据复合人才少	加强培训 + 工具模板化

场景类型推荐方案

场景类型	推荐方案	理由
简单问答机器人	传统方案	成本效益最优，无需复杂记忆
企业知识库	数芯智能体	深度知识关联与长时记忆检索刚需
多智能体协作系统	数芯智能体	多Agent间共享记忆与状态同步刚需
长期任务(数周)	数芯智能体	超长时记忆保持与任务断点续传刚需

未来展望：演进路线图

- 2026年：概念普及期**
头部厂商发布数芯智能体产品，市场认知度显著提升，早期尝鲜项目启动。
- 2027年：标准统一期**
A2A/MCP通信协议逐步统一，记忆模块实现标准化，多厂商生态开始互通。
- 2028年：协作成熟期**
预计70%数字化企业部署多智能体协作系统，数芯智能体成为企业级标配基础设施。
- 2030年：自主进化期**
具备持续学习能力的自主进化智能体出现，数芯支持智能体自我迭代与优化。

THANKS

成为世界卓越的数据库产品与服务提供商

