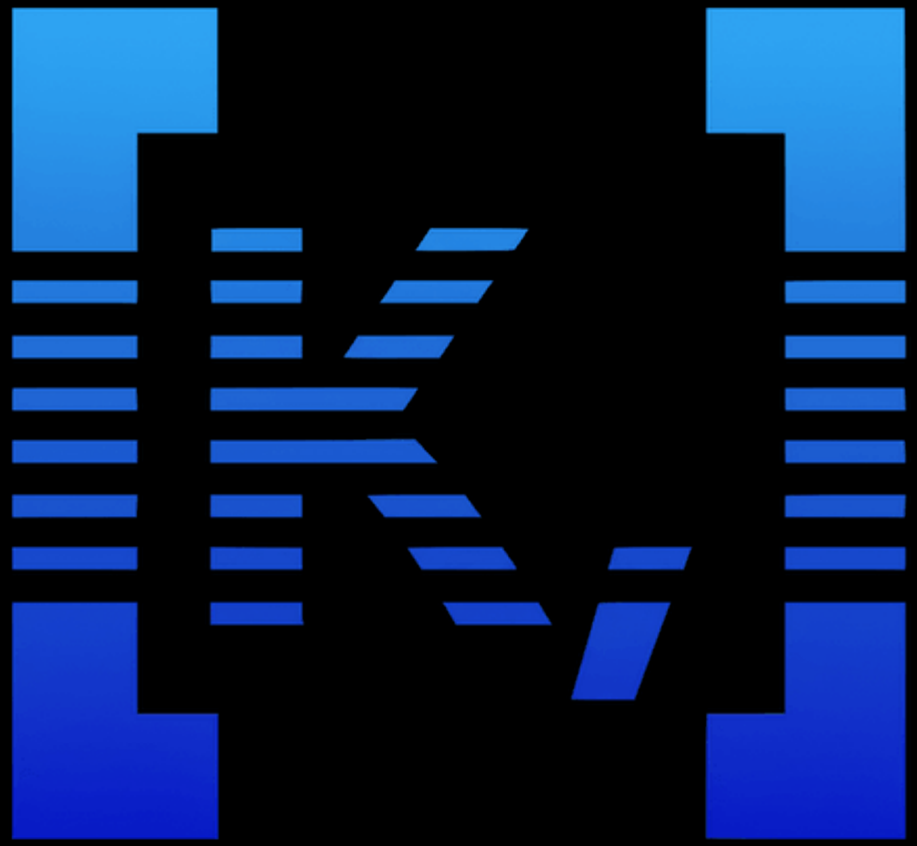


KVectors
向量数据库

王福弟

KeYou

背景与趋势

KEJOU

AI 是泡沫还是趋势？

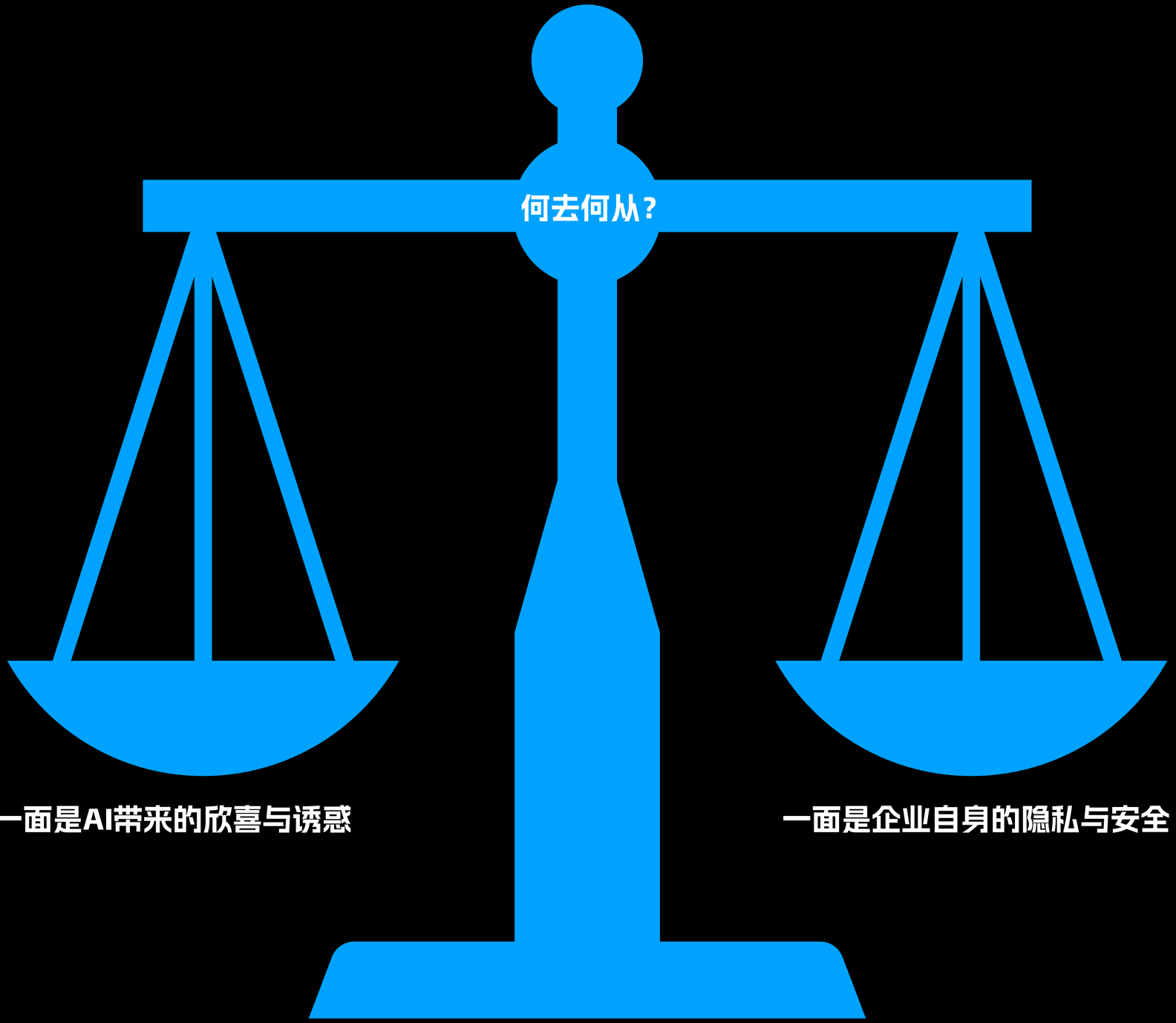
- 自 ChatGPT 在2022年底（11.30）发布之后，引领了新一轮的AI热潮
- 从电力（兆瓦）、算力（万卡、10万卡集群）到大模型（万亿参数），新一轮AI大基建正在如火如荼地极速发展当中…
- 大模型为企业带来了新的降本增效的能力，也带来了更多探索未知领域的可能性
- 从营销到研发，企业从前到后整个链路上的场景和流程都将被AI重塑

场景与痛点

- 企业对数据和大模型能力的渴求极度热切，却又往往不知道在什么场景进行投入，大多数时候都是在使用大模型厂商推出的标准化服务和API

- 企业垂直领域的知识是企业的隐形资产，如果通过公有云访问GPT能力，势必造成企业隐形资产的泄漏，所以企业在硬件上大多选择一体机或者自建

- 但依然无法解决数据相关的问题，比如如何有效利用大模型的上下文窗口并规避其限制，如何有效将大模型与企业数据结合完成业务场景的需求，这些都是企业当下关切却还没有探索出有效路径的事情。



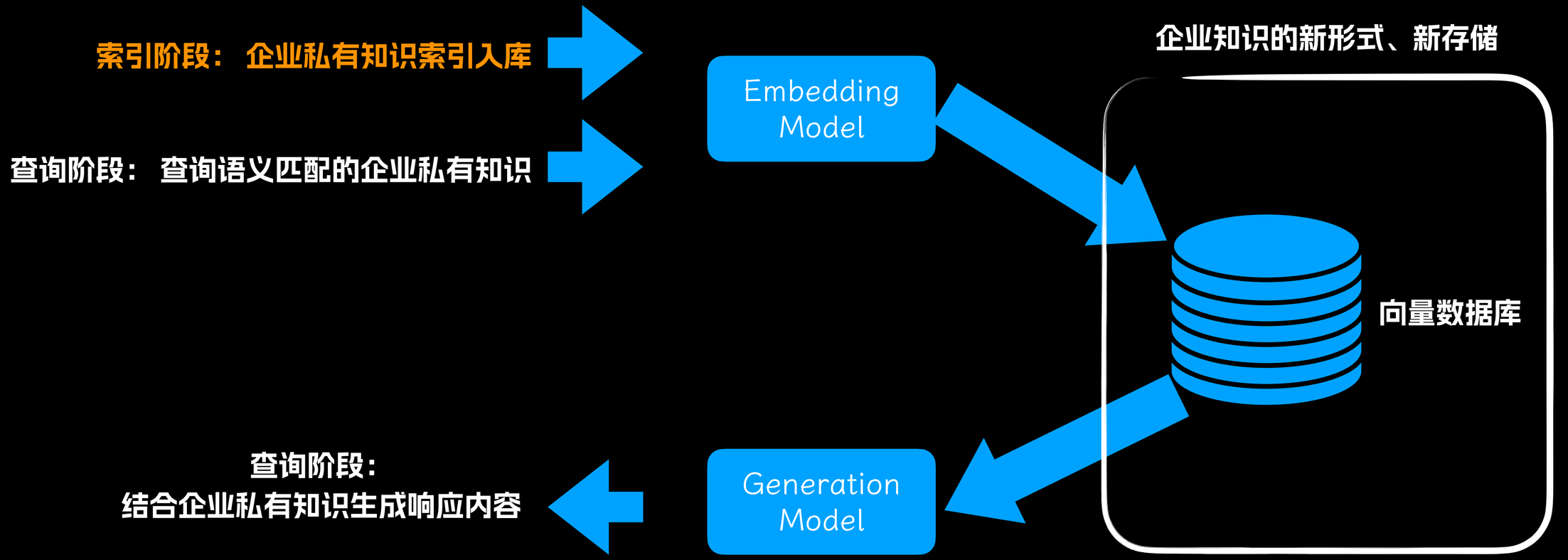
一面是AI带来的欣喜与诱惑

一面是企业自身的隐私与安全

如何有效结合
大模型与企业私有数据和知识？

RAG

典型 RAG 架构



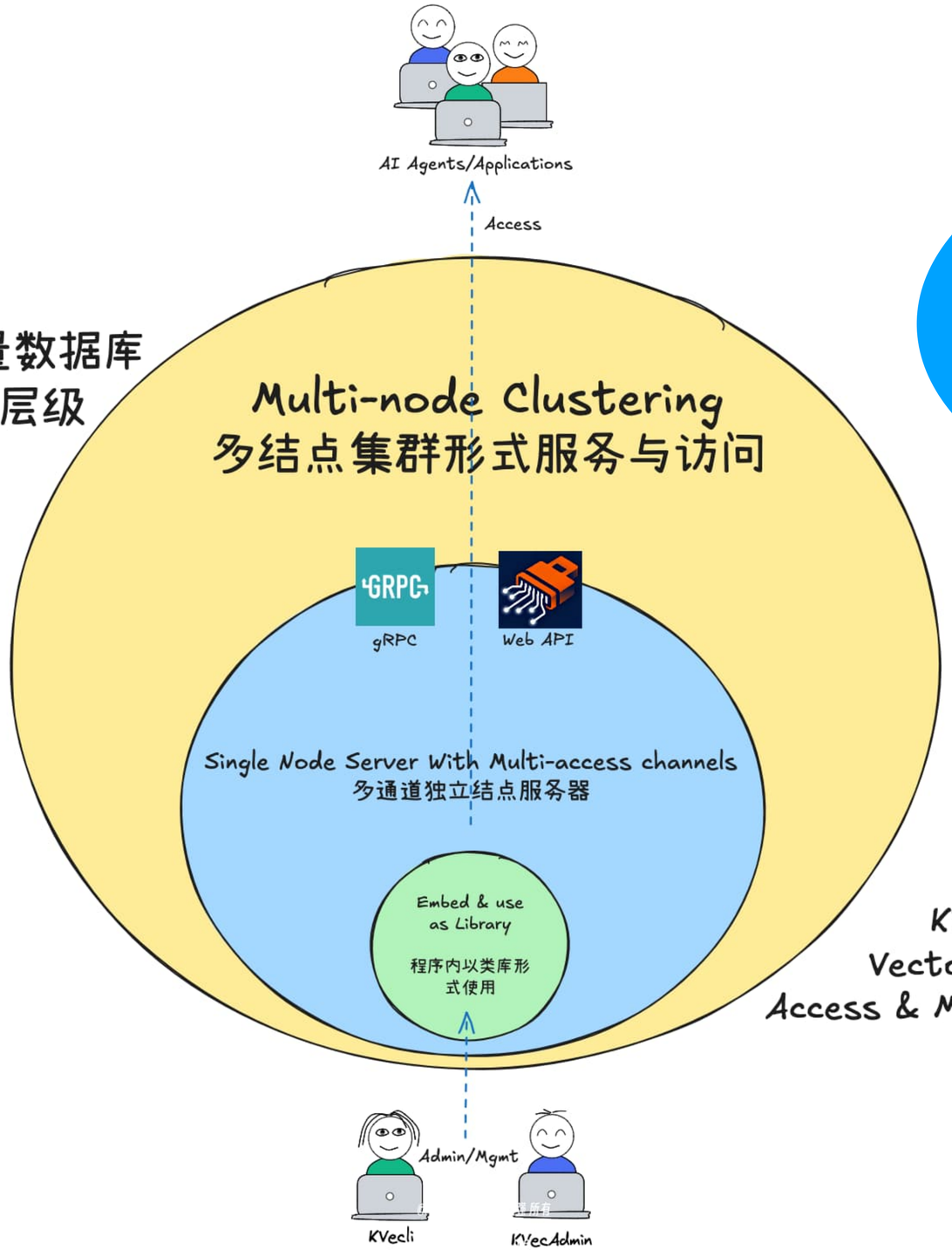
KVectors 应运而生

KVectors是什么？

- 一款 向量数据库
- 一款从一开始就追求极致性能的向量数据库
  - *消费级机器上查询时延小于10毫秒，个别情况下可以达到1.1毫秒*
- 一款始终业界追求前沿的向量数据库
  - *HNSW、IVF、FLAT甚至LSH等经历业界考验是向量索引全都支持*
  - *同时也支持PQ、RaBitQ等业界前沿的向量压缩技术*

KVectors 的设计与实现

KVectors 向量数据库
访问与管理层级

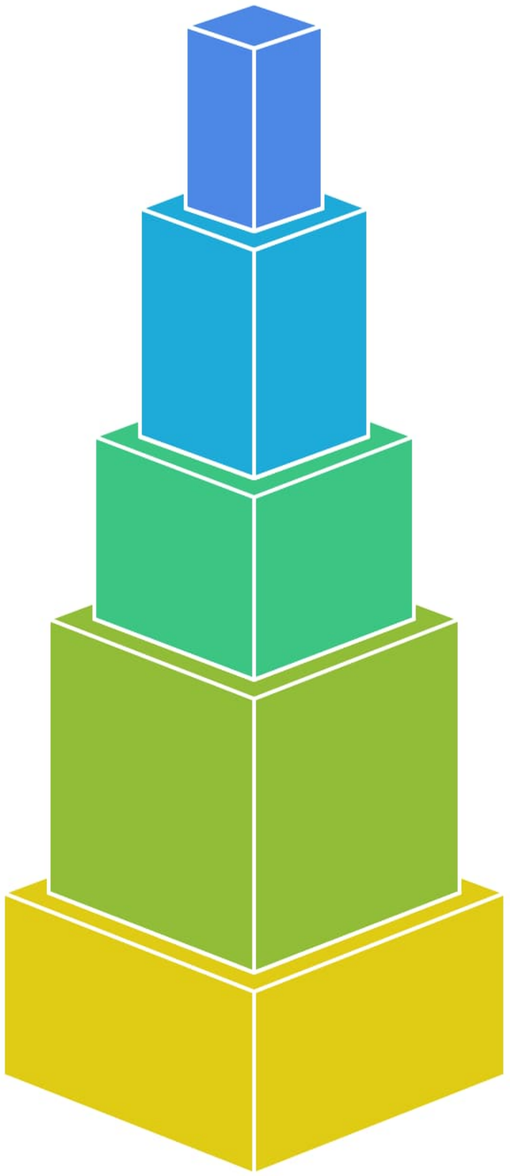


KVectors
Vector Database
Access & Management Layers

从访问
到管理
再到部署与运维
我们提供全方位的
呵护与支持

KEYOU

KVectors
Storage Layer Hierarchy



Transient Memory

Data stored temporarily in memory

Memory & Filesystem

Data stored in memory and filesystem

ANN Index

Index stored in memory with filesystem backup

Mmap & Filesystem

Memory-mapped data with filesystem persistence

Cloud Persistence

Data stored in cloud with high latency

五级存储设计

瞬时内存向量集合
Transient Vector Collections

内存+文件系统相结合的向量集合
Vector Database Storage

近似最近邻搜索向量集合
ANN indexed Vector Collection Storage

内存映射+文件系统相结合的向量集合
Mmapped Vector Collection Storage

云上向量集合
Cloud Vector Collection Storage

KVEC

KVectors 核心指标汇总参考

数据量	从0到万、十万、百万、千万、亿以及10亿以上
时延	从1.1毫秒到10毫秒（消费级硬件测试结果）
索引类型	FLAT, IVF, LSH, HNSW, DiskANN, etc.
向量压缩算法	FLAT, PQ, RaBitQ(南洋理工2024年提出的新算法）
存储类型	Memory，HDD，SDD，OSS over cloud

客户成功案例

KEAOU

智能客服的 AI 升级

橙子科技依托KVectors向量数据库，首先升级了自身的智能客服系统，将日常客服过程中发现和沉淀的FAQ，通过大模型Embedding，然后存入KVectors向量数据库，在后续的客服服务流程中，先通过AI介入，AI没有合适策略的情况下再转人工介入。 一个小小的改变，为公司客服减少了80%以上的人工客服成本。

KEYOU

千人千面的个性化推荐

在智能客服升级完成后，橙子科技又在自身电商平台上尝试基于大模型的个性化推荐系统升级，将过去固定的商品推荐逻辑升级为个性化的千人千面推荐，依托大模型和KVectors向量数据库代为管理的企业自身商品库语义数据，橙子科技完成了新一轮的流量转化和GMV增长。

KEEVO

以图搜图

传统的以图搜图基于纹理等老的AI算法，效果不佳。

而且搜索引擎提供的以图搜图，数据都是公开数据，企业内部如果牵扯到数据隐私和安全，通常并不适合走公开的搜索引擎做以图搜图。

这时候，企业就可以考虑基于大模型和KVectors向量数据库打造一套自己的以图搜图系统，从而规避数据安全和隐私泄露等问题。

尤其对于电商平台来说，用户通常习惯在某处看到一张商品图的时候上传查找类似的商品，这时候，以图搜图就可以针对电商平台自身的商品库进行搜索，同时又不用泄漏自身平台的SKU。在提升用户体验的同时，也增加了电商平台的转化率和GMV。

Agent Memory

有人说 2025年是Agent元年，而谈到Agent，就不得不谈到它的Memory系统，受限于大模型的上下文窗口限制， Agent智能体要完成大规模的任务的规划和实施，就必须依托外部Memory才能完成复杂的任务。

KVectors向量数据库为Agent智能体提供了完美的Memory存储。

为什么选择我们？

团队介绍

KEAOU

王福强

aka. 扶墙老师

杭州福强科技有限公司创始人
25+年互联网与金融技术研发与管理经验
《Spring揭秘》、《SpringBoot揭秘》作者
腾讯云最具价值专家(TVP)

- ✓ 历任多家公司顾问/CTO
- ✓ 原挖财技术VP及首席架构师(Chief Architect)
- ✓ 原天猫产品技术部资深架构师
- ✓ 原阿里巴巴平台技术部海量数据部门高级技术专家
  原阿里巴巴大数据中间件canal的产品与技术奠基人

杭州福强科技有限公司专注于技术、人才与组织成长

主要为企业用户提供技术战略、组织管理、数字/数智化转型
与互联网产品和技术顾问/咨询服务。
行业上，互联网和金融领域涉猎居多，
主导过包括信贷，理财，债券，外汇(保证金)交易等多种金融系统的设计与研发。
除了金融行业客户，亦服务过房地产、人力资源、医疗服务、平台SaaS等多个行业的企业客户。

大连理工大学
DALIAN UNIVERSITY OF TECHNOLOGY

联系洽谈

KEAOU

请掏出手机开始扫描下方二维码

KEYOU

Any Questions?

KEVOL

福强科技做生意
尽心尽力尽本分

KEEVOL

KEEP EVOLUTION...

福强

ワンマン会社