RAG不只是检索：Milvus在Context Engineering中的湖仓一体实践

刘力

Zilliz / 研发总监兼Milvus负责人

AiCon

全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

目录

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

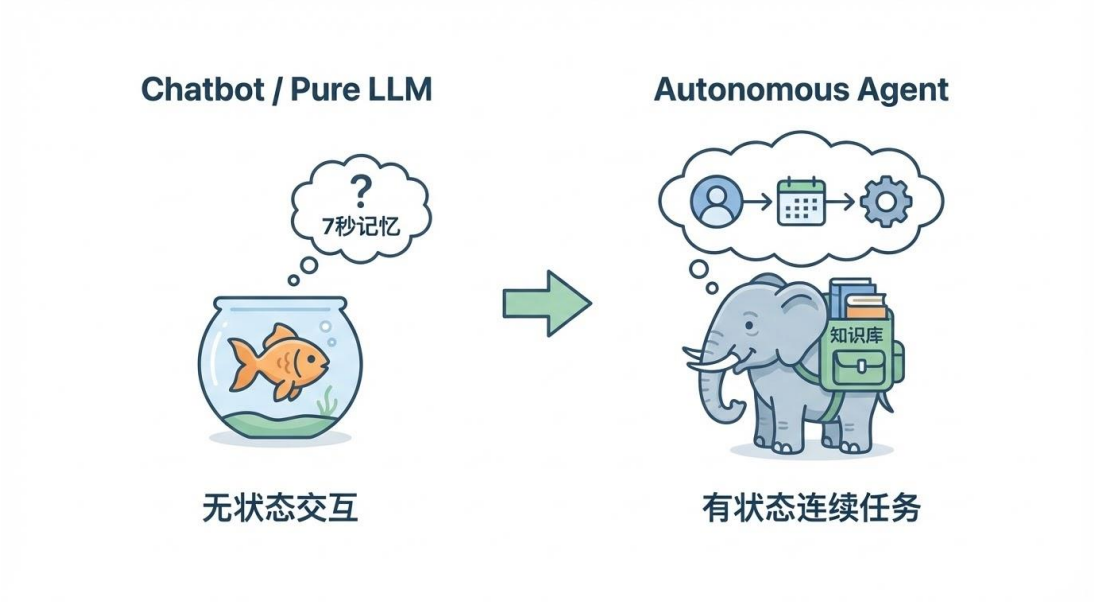
- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

01 从 RAG 到 Agent

Context Engineering 的演进

范式转变：从无状态 Chatbot 到有状态 Agent



- 从Chatbot (Stateless): 单次交互，依赖预训练知识。
  - 无法承担可规模化的复杂任务
- Agent (Stateful): 连续决策，依赖动态上下文。
  - 核心特征 1 - 长期记忆 (Long-term Memory): 记住跨会话的历史交互。
  - 核心特征 2 - 环境感知 (State Awareness): 实时感知 API 返回、工具执行结果。

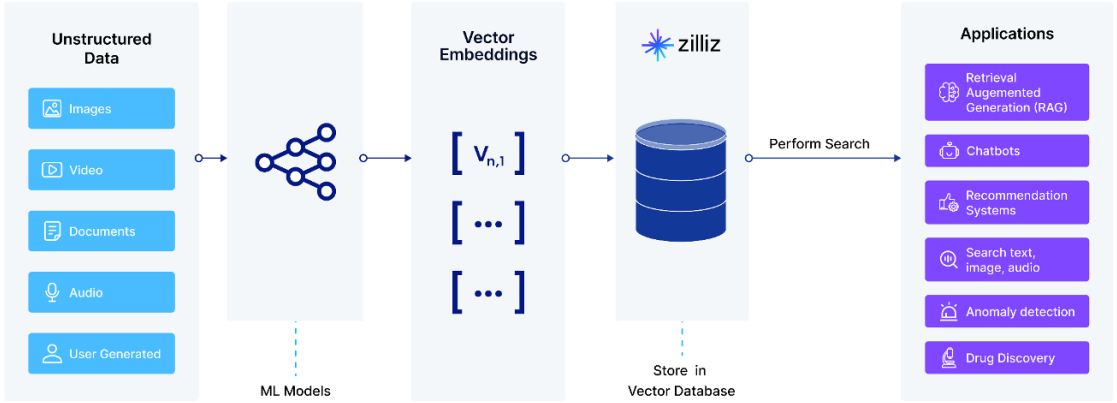
Context Engineering 的全景技术栈

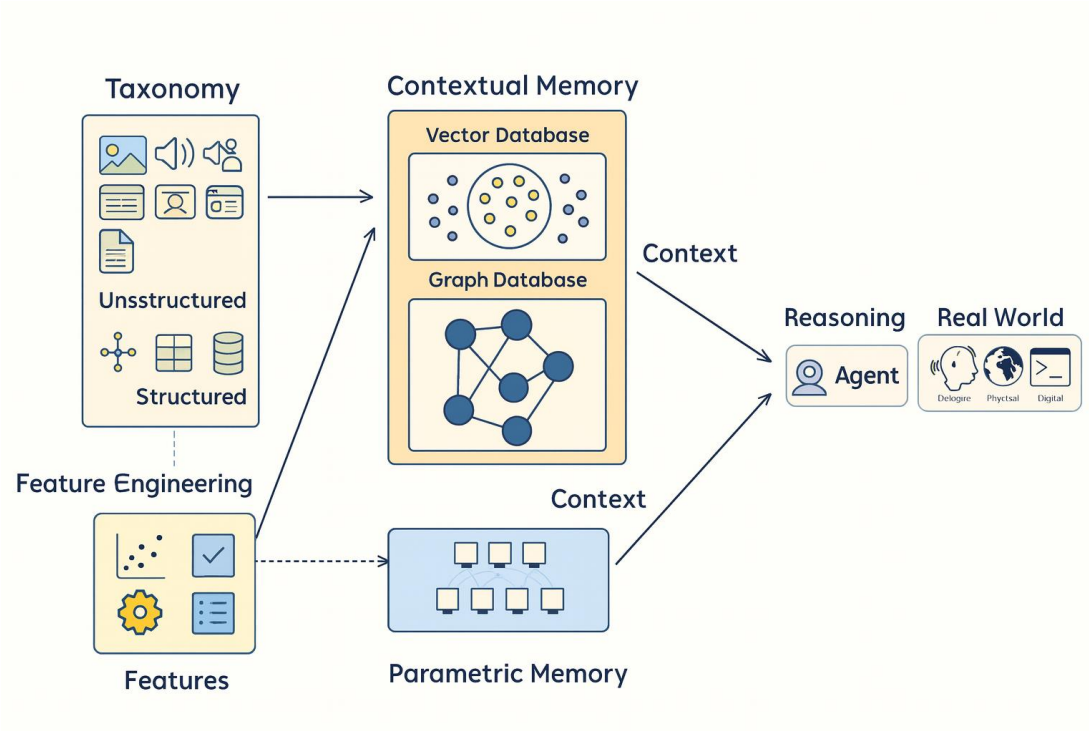
从“输入数据”到“认知循环”，Context 有自己的生命线

- **上下文的搜索**
  丰富的检索方式提高搜索准确性，解决模型回答准确度的问题
- **上下文的处理**
  强大的不同类型数据的处理能力，帮助数据的准备和迭代
- **上下文管理（储存和扩展）**
  稳定、灵活、高性价的数据管理基础， 提供稳定的上下文环境

向量数据库在 Context Memory 中的角色

- 向量是Context中流转的语义信息
- 向量搜索利用向量在高维空间的距离来表征非结构化数据的相似度
- 向量数据库是一种专为存储和查询高维度向量数据而优化的数据库系统。

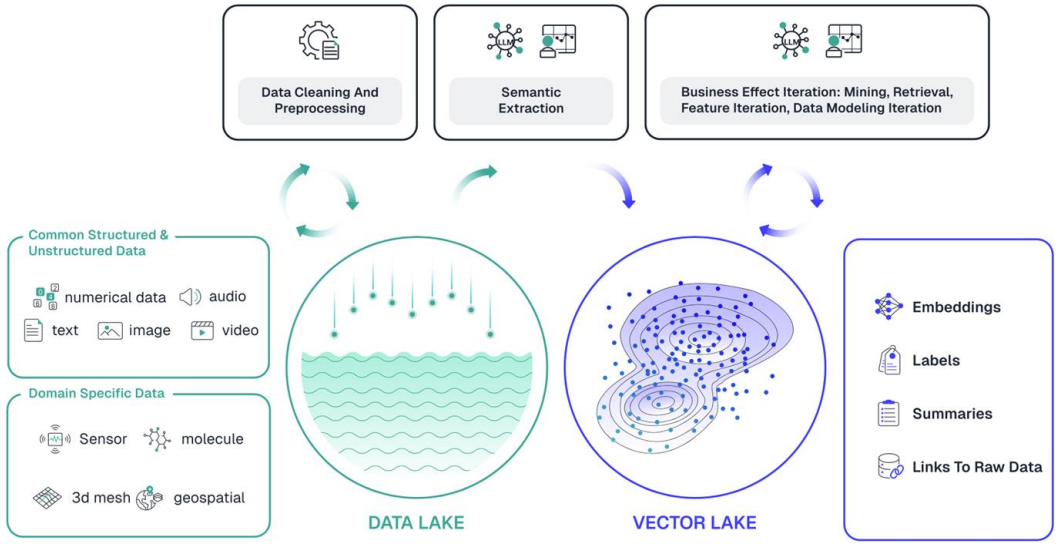




向量库 vs 向量湖：多模态数据的闭环

为什么需要湖?

- **数据规模**：海量数据对存储扩展性、成本控制和访问灵活性提出更高要求。
- **数据多样性**：多模态数据为基础，多元的分析需求催生出多样的数据结构，并要求 schema 能够灵活演进。
- **数据处理**：海量异构数据需要对接多种计算引擎，以支撑丰富的业务场景。



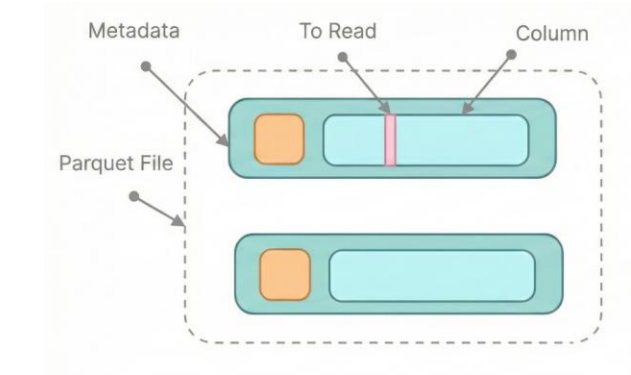
02 Context的管理

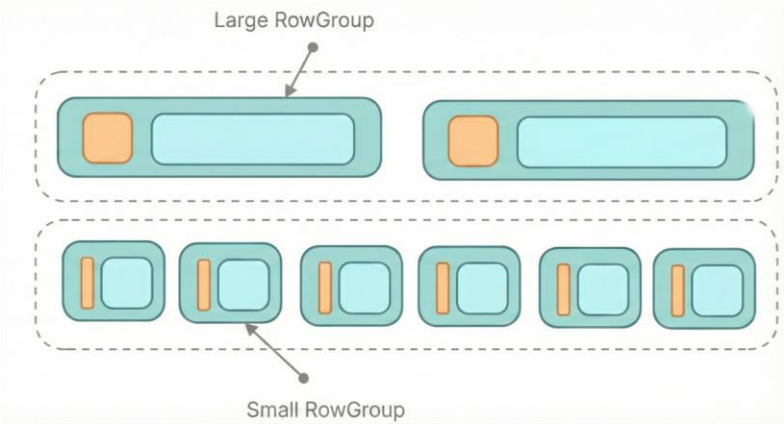
向量数据湖作为非结构化数据统一底座

Format：Parquet 在向量场景下的痛点

Parquet-OLAP时代的利器

- **RowGroup-一切的原罪**
  - 点查性能：每次获取某列中的一条数据就需要加载整个page甚至RowGroup
  - RowGroup作为压缩粒度，增加点查难度
  - 宽窄列：向量数据和其他多模态数据带来大宽列
    - 大 RowGroup 导致随机读写效率低
    - 小 RowGroup 导致无法利用S3带宽，无法充分压缩，同时会导致footer膨胀
- **糟糕的多模态数据支持**
  - 不支持外接Binary文件，内置造成无意义开销
  - 一切RowGroup带来问题更加严重

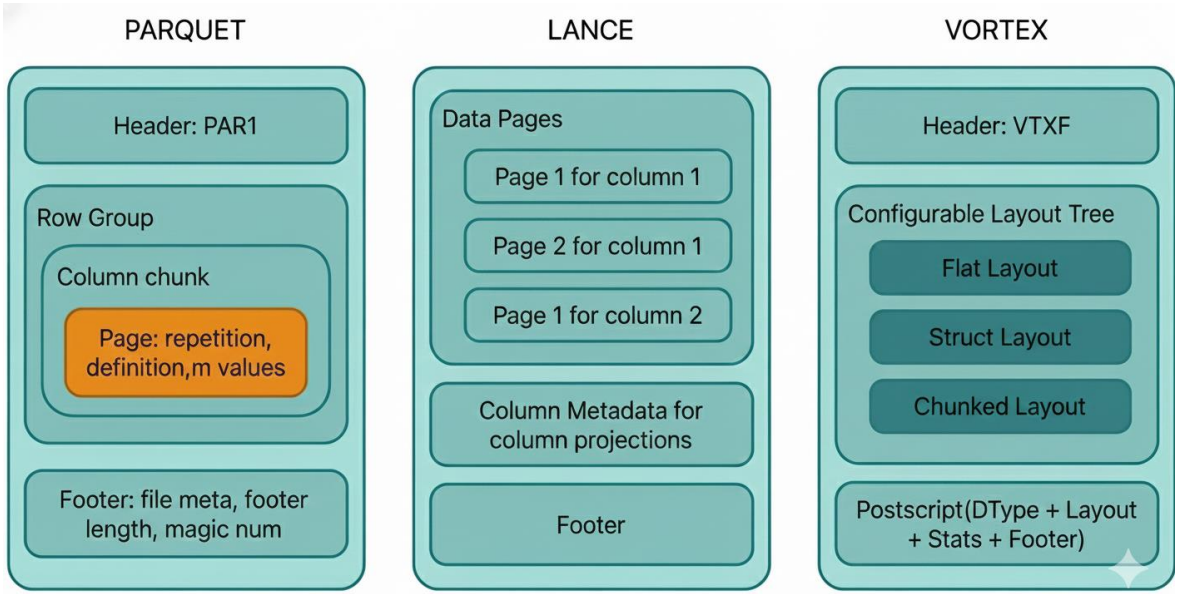




Format：新老格式解决问题

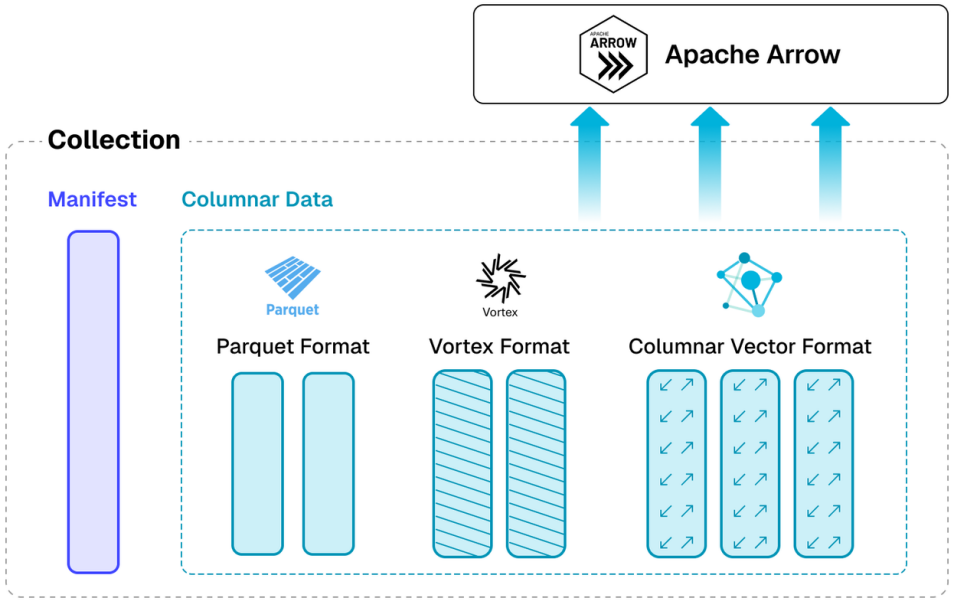
- **解决方案**
  - 动态列划分，窄列合并，宽列分离，RowGroup大小可调整
    - 宽窄列分离解耦，解决读写效率和压缩问题
    - 针对对象存储的访问优化，配合列分离解决点查问题
  - RowGroup大小可调整，适配云对象储存粒度
  - 支持外部储存Blob、Text等大型文件

- **其他Format**
  - Lance: 去掉RowGroup和压缩以支持点查
  - Vortex：灵活配置，特殊压缩格式支持点查



Format：各取所长，统一格式拥抱生态

- **兼容多种储存格式**
  不同数据类型适合不同储存格式，不同数据源使用不同格式。
- **兼容丰富生态**
  通过Apache Arrow兼容不同的执行引擎
- **统一远近端数据格式**
  格式最好直接使用远端格式作为近端使用，在非对象储存场景下可通过Posix通信

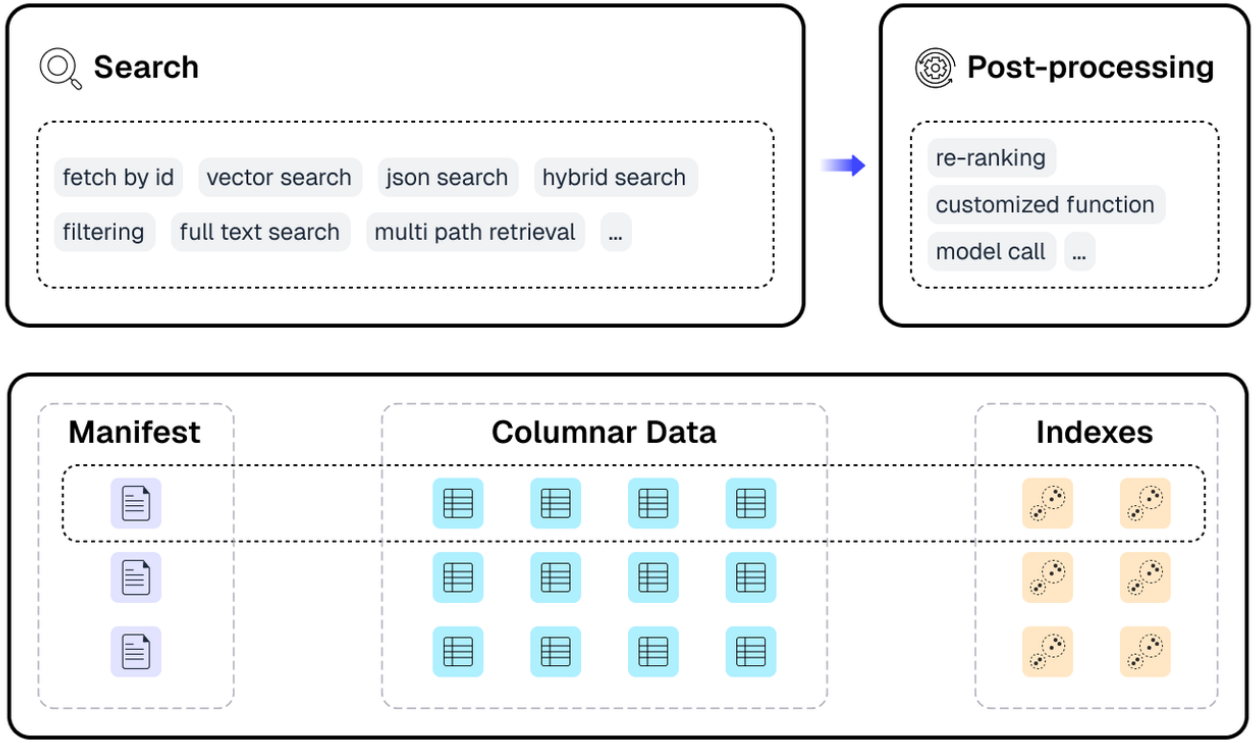


索引是向量数据湖的一等公民

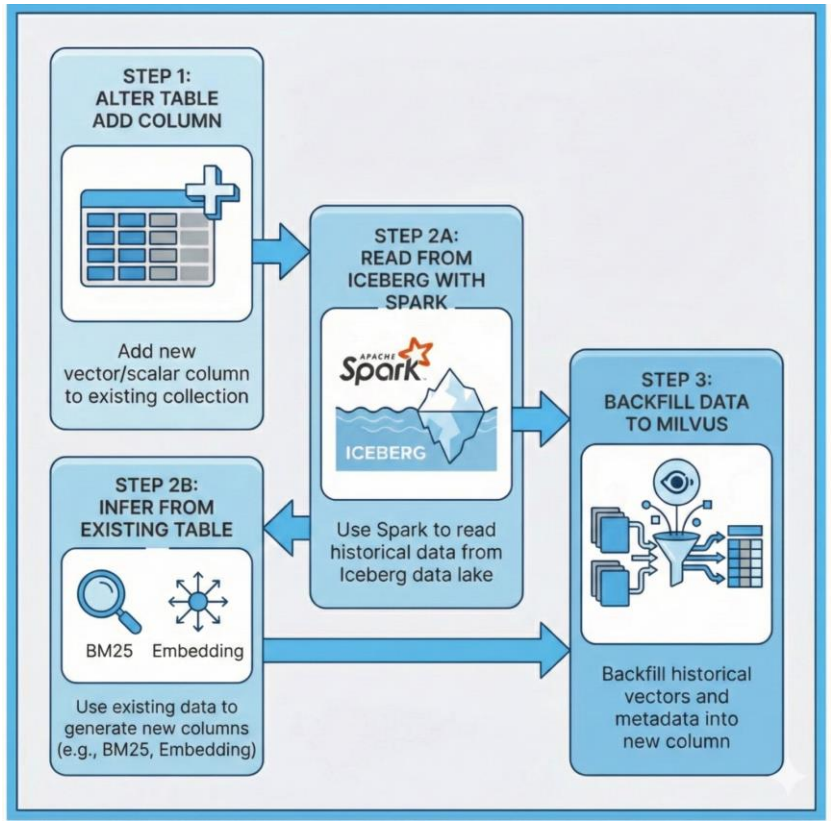
- **索引也是数据**
  相较于传统的数据湖，向量数据湖的不仅服务于**查询**，还要直接服务于**搜索**

- **每一种数据类型都提供 SOTA 水平的索引支持：**
  在语义宽表中，复杂的搜索场景要求每一个字段都支持构建高性能索引，保证上下文信息中任意部分都可以高效检索。

- **大规模搜索能力：**
  面向“性能不敏感但成本敏感”的批量搜索场景，单次任务可稳定取出十万至百万级数据并输送至下游垂直处理流程，兼顾吞吐与成本。



Schema Evolution：在生产中迭代



- **即时即用：**
  数据湖场景中，一张表的数据在持续的数据挖掘和迭代中需要动态的扩缩。需要支持在线瞬间向现有集合添加新的标量或向量字段或者删除，实现零重建。

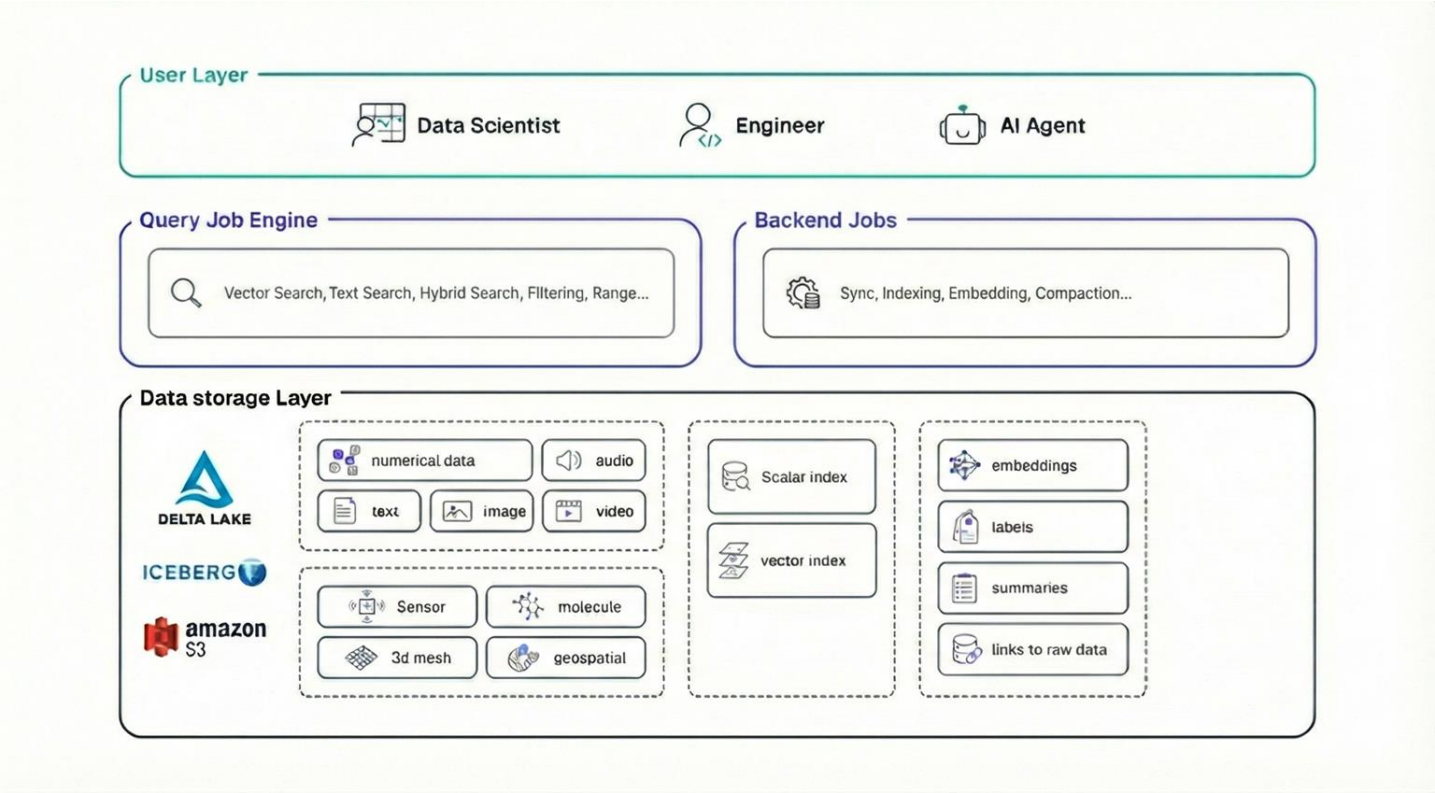
- **无缝回填：**
  - 提供 Iceberg 连接器，利用 Spark 直接从数据湖摄入并回填历史数据到新字段。
  - 利用表内现有数据生成新数据回填，比如BM25，Embedding提取

- **权限管理：**
  列级别的权限管理助力企业多团队分工协作

03 Context的处理 & 搜索

构建下一代向量数据湖

湖仓一体：外表架构支撑全局一份数据



- **移动计算比移动数据更划算**

  对于已有湖场景，向量的处理能力可作为外置能力，拒绝数据冗余

- **数据的增量同步**

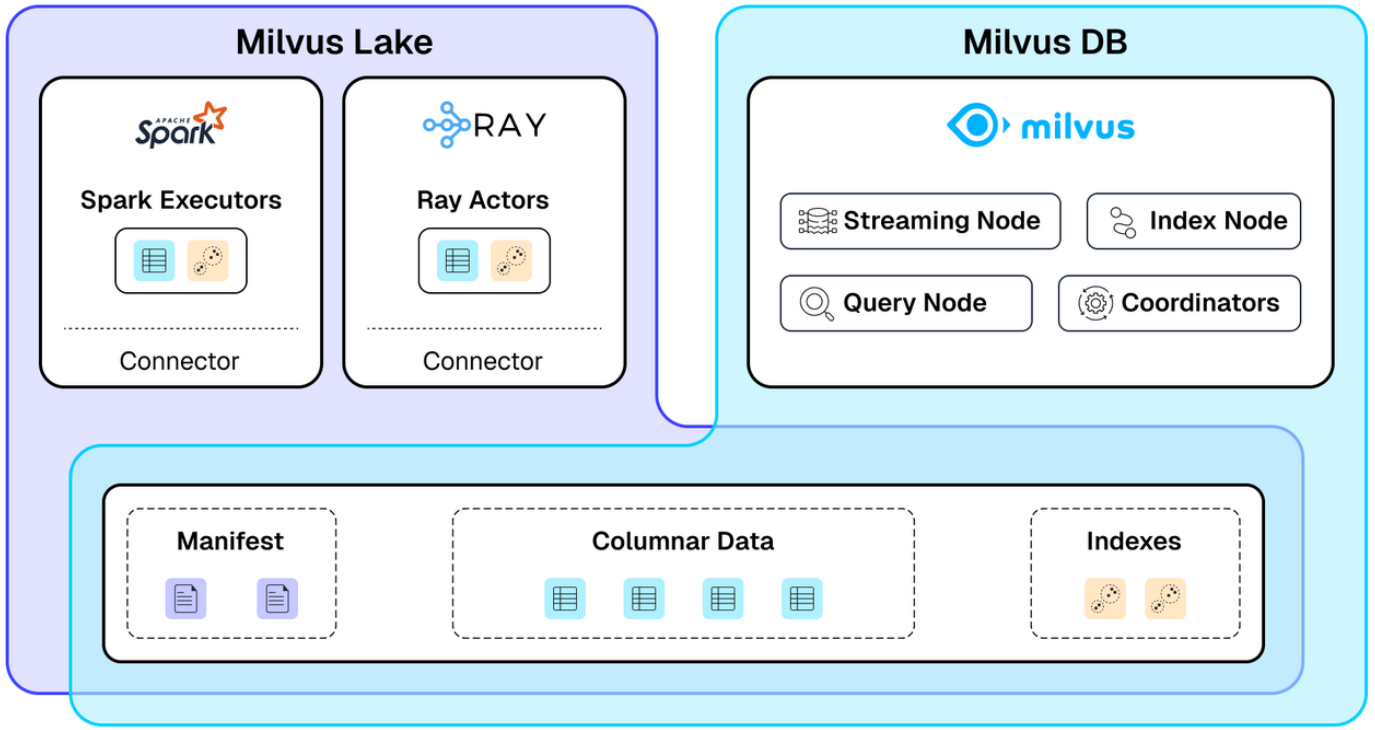
  感知湖底座的增量并同步至外置仓侧解决日常数据写入问题

- **混合检索能力**

  支持与原始湖仓进行联合查询与检索，兼顾性能与灵活性

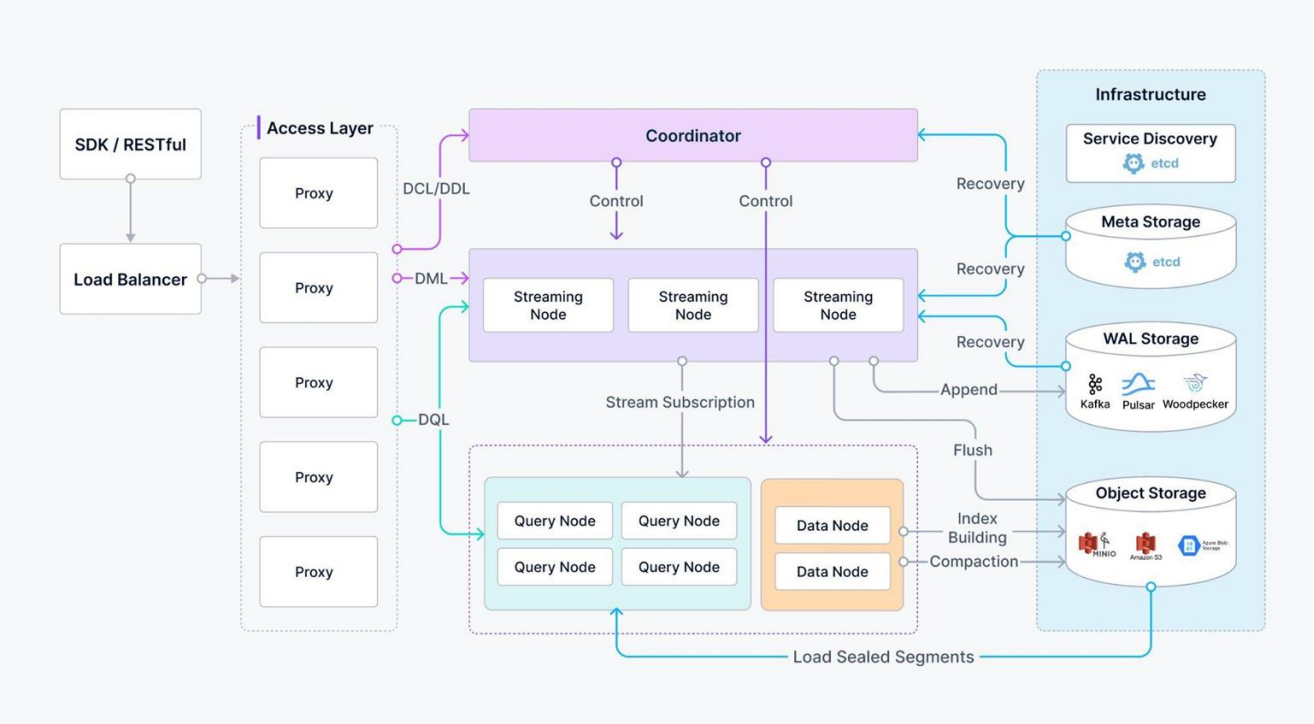
湖仓一体：存算分离与多引擎协同

- **统一在线/离线查询：** 同一套存储上同时支持大规模批处理分析和低延迟在线查询，消除数据孤岛。
- **管道可扩展性：** 支持用户在数据生命周期各阶段插入自定义逻辑，如特征提取、rerank等。
- **多模态统一索引：** 对向量、BM25、Sparse、标量(geo, JSON等)提供统一的高性能索引能力。
- **内置数据治理原语：** 架构层原生支持大规模Kmeans、数据去重等常见治理任务。



湖仓一体：数据库架构的云原生与完全分布式

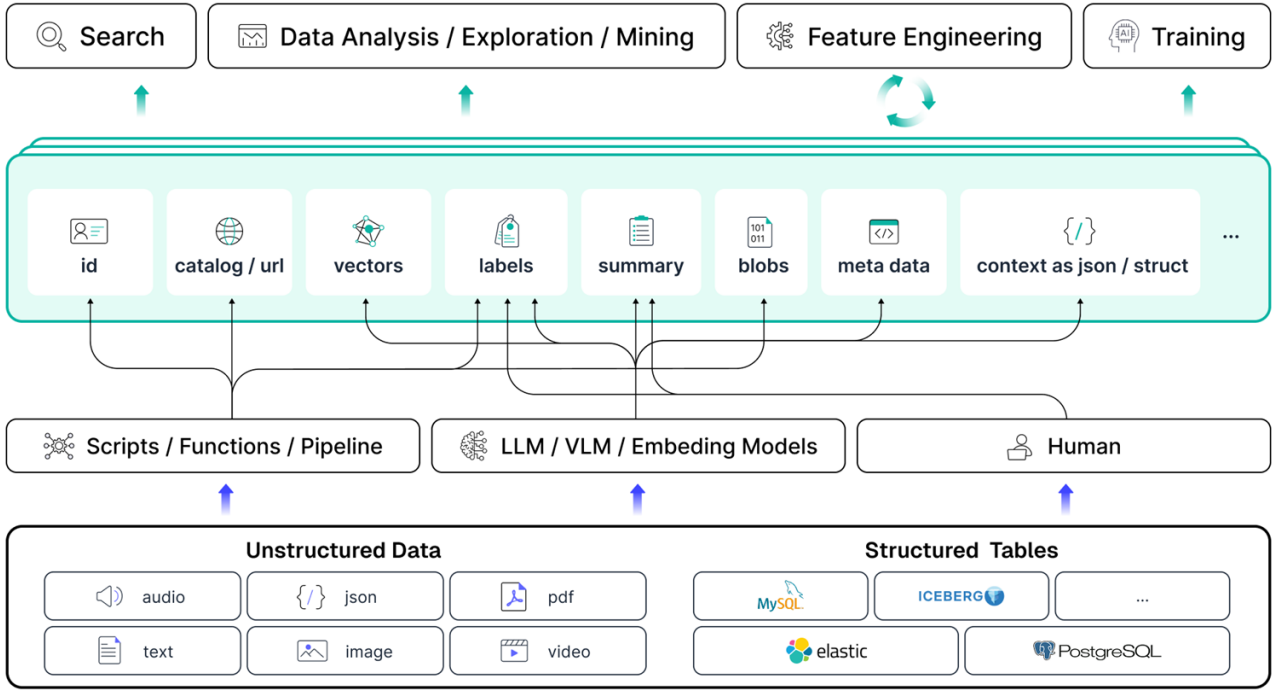
- **云原生基座**：基于 S3、etcd 与 Kubernetes 构建，可扩展性作为一等设计原则
- **微服务架构**：查询服务、索引构建与Compaction 独立部署，隔离性更优
- **存算完全分离**：存储与计算彻底解耦，实现真正的弹性伸缩
- **实时数据和历史数据分离**：兼顾实时性和性能，支持不同阶段使用不同执行引擎



数据类型演进：多模态数据的结构化归纳

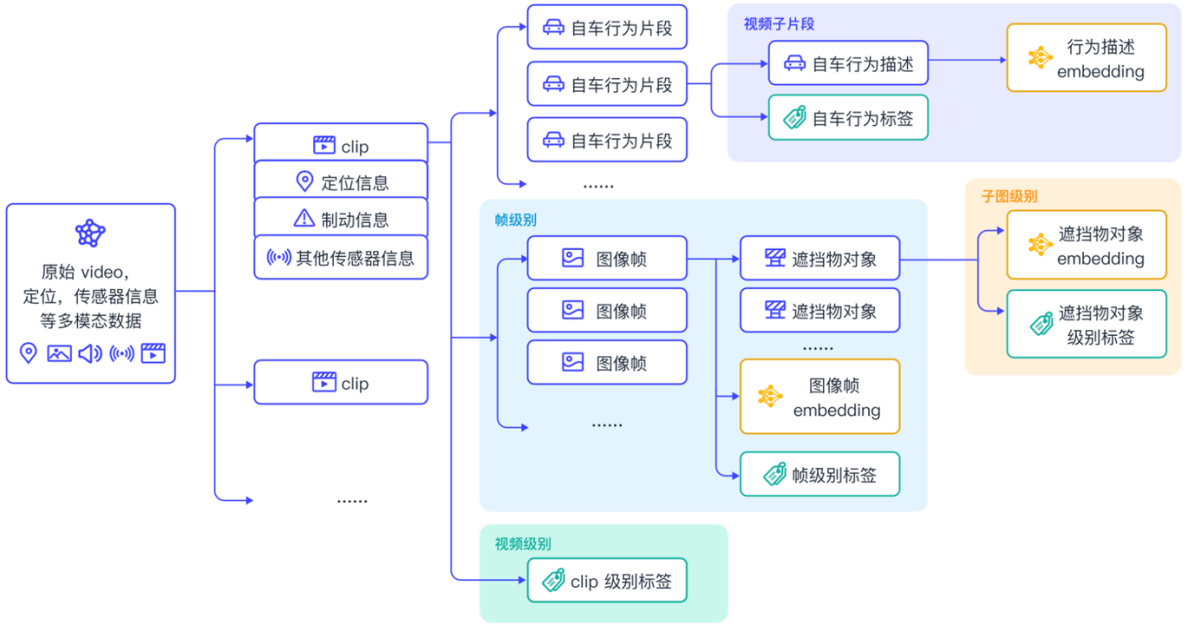
- **语义宽表建模**：单表承载完整业务实体，实体与 Row 一一映射，避免 JOIN/聚合的性能损耗与多表治理负担

- **复合类型原生支持**：对 Struct、JSON、Array 等嵌套结构提供一等公民级别支持，统一建模多模态上下文



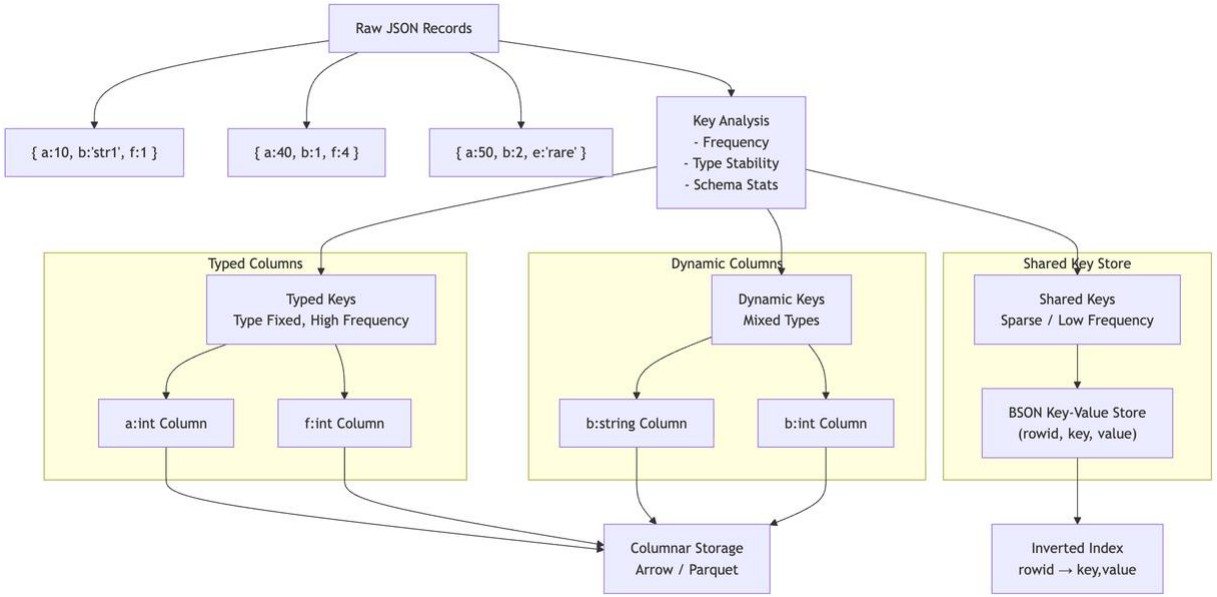
数据类型演进：处理复杂的多模态数据

- **向量类型**：FloatVector、BinaryVector、Float16/BFloat16、SparseFloatVector，覆盖稠密/稀疏/二值场景
- **标量类型**：Boolean、Integer、Float/Double、VarChar，支持标签、数值、URL等
- **地理空间类型 (GeoSpatial)**：原生WKT/WKB格式，R-Tree索引，支持ST_WITHIN、ST_CONTAINS等空间算子
- **时间类型 (TIMESTAMPTZ)**：时区感知时间戳，支持时间范围过滤、排序、Time Decay Reranker
- **文本类型 (Text)**：长文本、关键字检索、全文检索、短语匹配

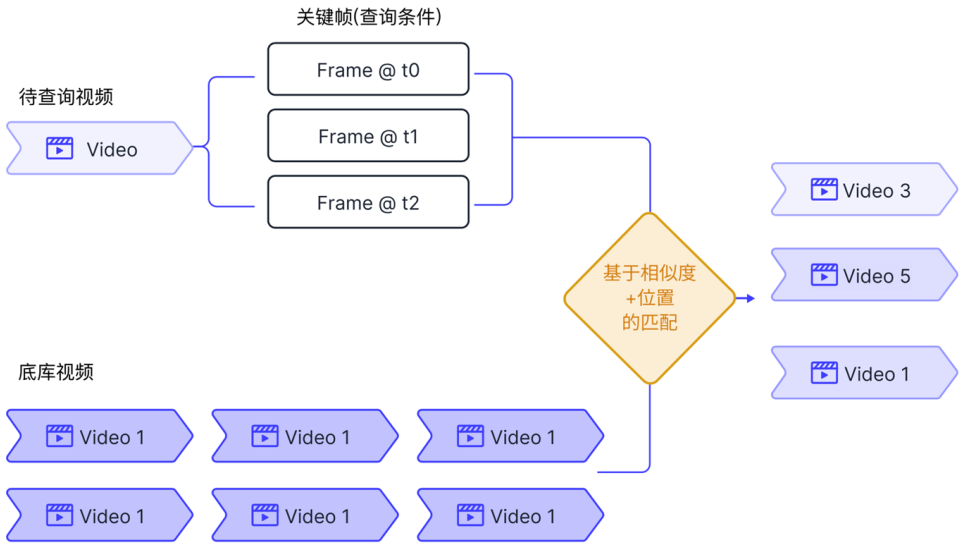


数据类型演进：JSON支持Schema-less范式

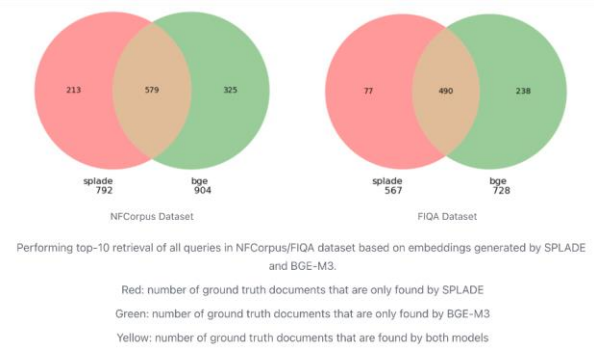
- **JSON Shredding**：自动分析字段频率，高频字段提取为独立列，低频字段构建倒排索引，无需用户配置，查询性能提升14~89×
- **JSON PATH索引**：为固定热点路径创建强类型索引，支持布尔/数值/字符串/数组类型，查询性能最优
- **JSON FLAT索引**：索引整个JSON子树，查询时动态指定路径，适合Schema动态变化场景
- **灵活组合**：Shredding兜底大部分查询，PATH/FLAT索引针对特定场景进一步优化

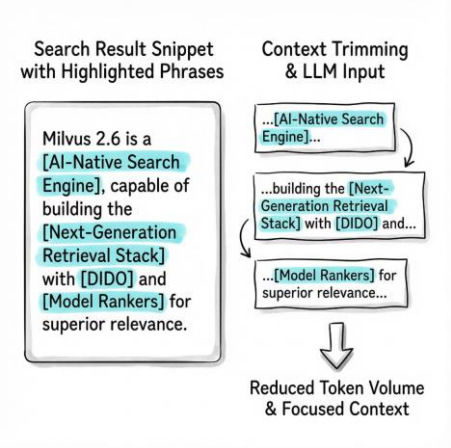


数据类型演进： Struct完整表征数据Entry



混合搜索：高质量Context的钥匙





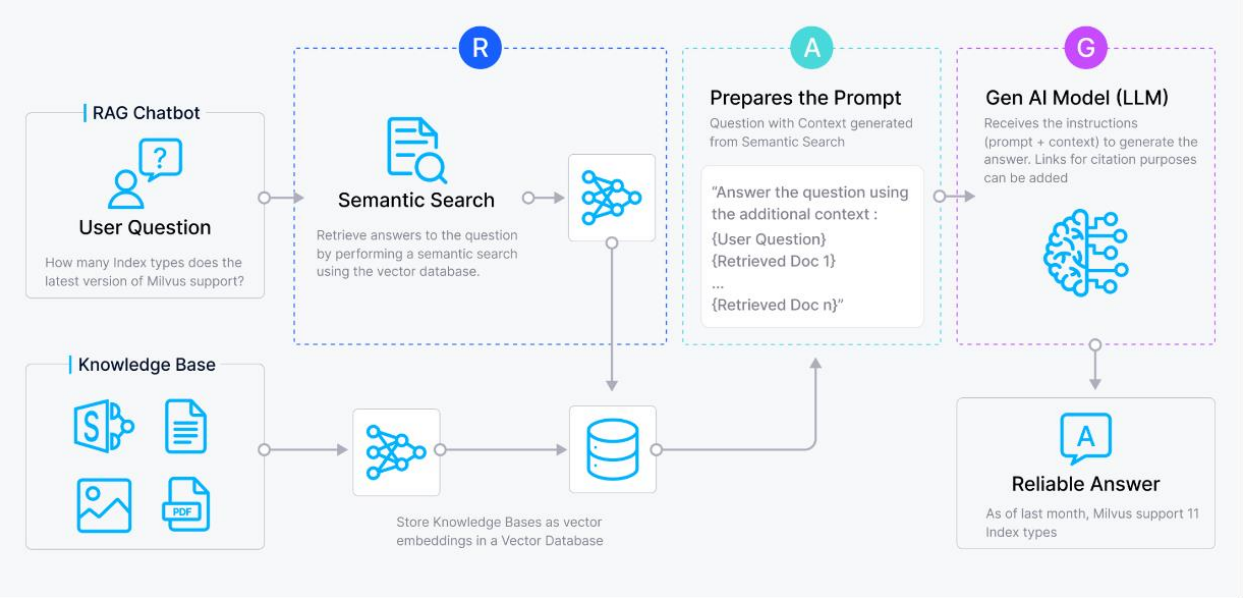
- **Context质量是关键因素**
  - **Context Poisoning：** 错误信息反复引用 → 幻觉放大。
  - **Context Distraction：** 太长导致模型失焦。
  - **Context Confusion：** 噪声干扰推理。
  - **Context Clash：** 上下文矛盾引发崩溃。
- 解决方案
  - **Sparse/BM25 + Dense：**关键词、语义双管齐下
  - **GraphRAG：**关注相似的同时关注相关
  - **Decay：**回归记忆本质，提升正确率
  - **Highlight：**降低噪声的关键
  - **Rerank：**搜索的最后一道防线

04 海量数据的处理

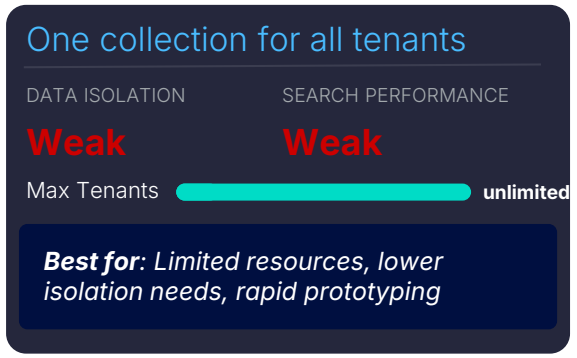
多租户架构与冷热分层

场景定义：从 RAG Chatbot 到多租户隔离

- **多租户**
  - ToB场景：用户的业务面向公司，数据需要粗粒度隔离
  - ToC场景：用户业务面向个人，数据需要细粒度隔离。
  - 基于租户的智能冷热分离：基于时序、容量的替换策略
  - 灵活控制能力
- **单租户**
  - 数据挖掘，数据准备场景
  - 基于数据分布的冷热分离



多租户策略：隔离性与性能的权衡

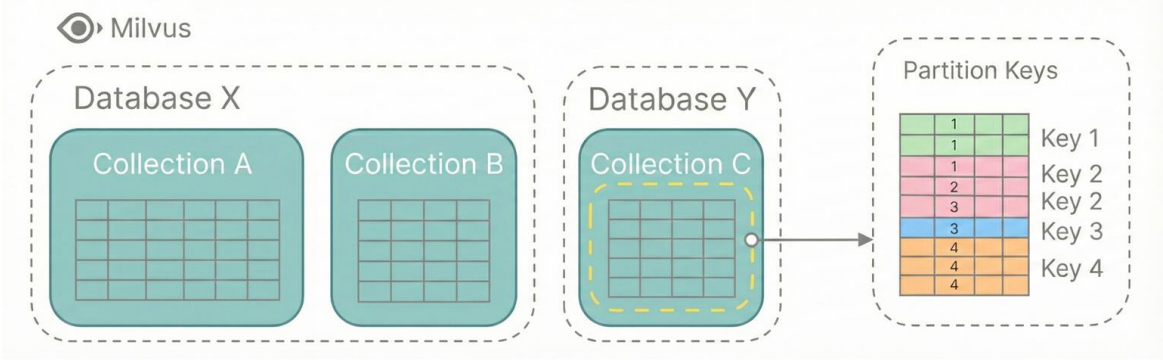


方案

- 通过内置过滤条件来达到租户分离的效果
- 通过基于PartitonKey的数据分片来降低计算量

遗留问题

- 过滤：常用的图索引在高过滤下性能较差
- 扇出和容量



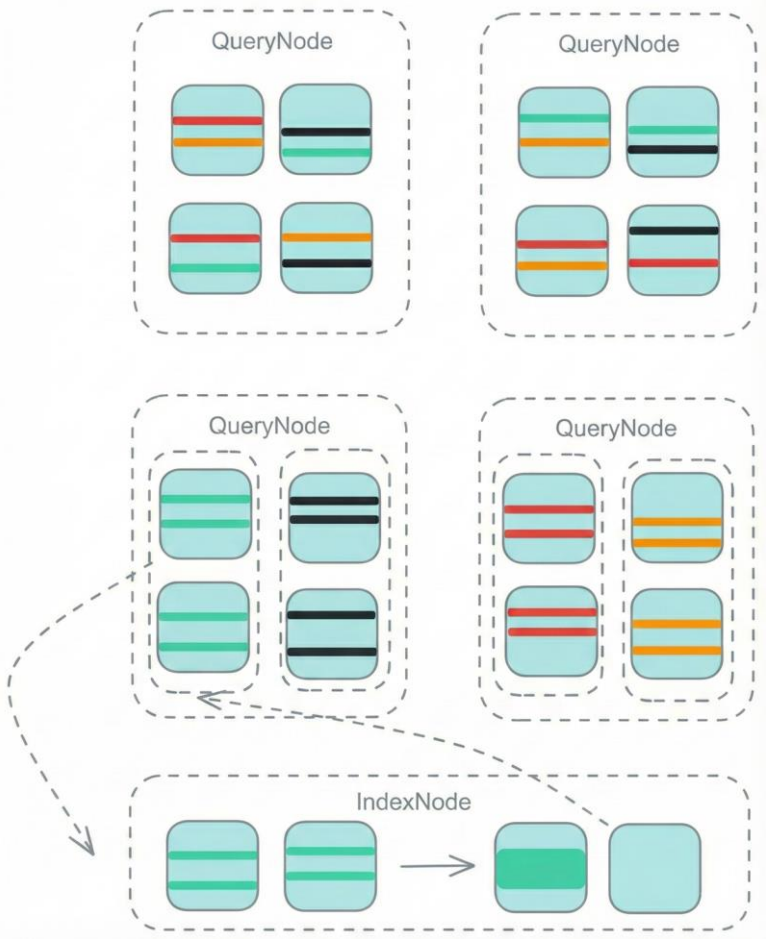
性能优化：通过数据聚集减少计算扇出

多租户数据的挑战

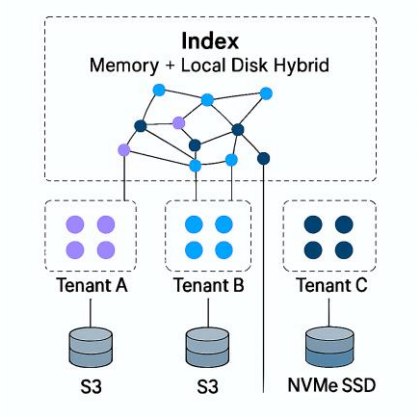
- 租户数据通常离散写入，会呈现较为严重的分散
  - 每个数据块都需要执行过滤，Overhead严重
  - 对于每个块来说，Brute-force (暴搜) 反而是最优解，无法充分利用ANN索引带来的性能提升
- 租户数量巨大，受限于元数据，和存算分离架构，无法给每个租户分配一个数据段

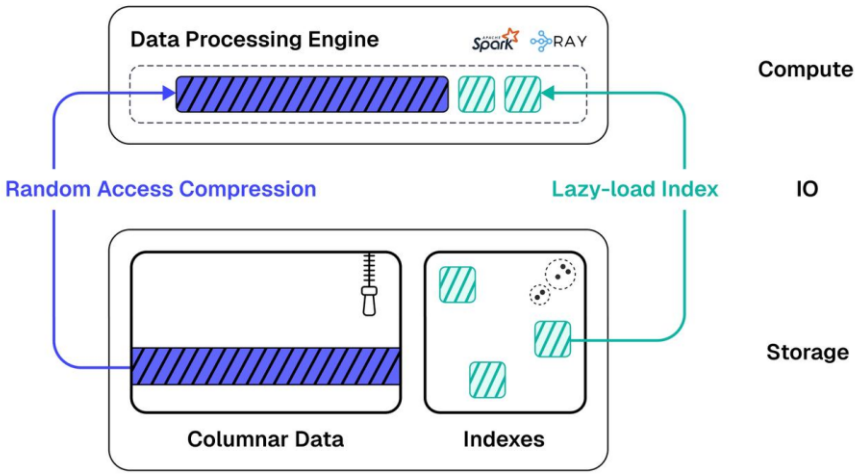
解决方案

- 根据租户分桶，减少扇出压力
- 后台周期性重构数据块，富集租户，提高效率
- 如果桶过大，使用LSM模式分摊富集成本



成本优化：智能冷热分层存储





- **分层存储：**
  - 数据可以在RAM，NVMe和S3中自由流动
  - 按需加载，无感故障恢复
- **动态驱逐：**
  - 支持基于时间，基于容量的驱逐策略
- **数据粒度优化：**
  - 对于小租户支持租户级别的数据粒度
  - 对于大租户
    - 支持按数据聚类为粒度
    - 同时后台加载图后能切换图算法，实现超高性能
  - 实现百毫秒到毫秒级延迟，数十到数千QPS的冷热查询

One More Thing: Zilliz Cloud

The Forrester Wave™ Vector Database Providers, Q3 2024

Zilliz 在向量数据库领域深耕八年，
陪伴全球客户深度迭代解决方案。

41K+ GitHub Stars

400+ Contributors

5K+ Enterprise users

100M+ Downloads

xiaomi · 得物 · 智慧树 · 贝壳 · WPS · 滴滴 · 智联招聘
理想 · Shopee · Likee · 新浪微博 weibo.com · 哈啰 · 智慧芽 · C12.ai
快手 · 唯品会 vip.com · tokopedia · 汽车之家 · Ctrip 携程 · SUPIEDT
vivo · iQIYI · 点石科技 · SmartNews · 企查查 Qcc.com · notta · 合邦电力 HEBANG ELECTRIC POWER
oppo · BIGO · SF TECHNOLOGY 顺丰科技 · 丁香园 · SHULEX · 豆神教育
快看! · 有道 youdao · mobileye · moj · 粉笔 · EviMed · 全诊通 All-In-One Clinic · Clobotics · 转转 · 焦点科技
accenture · airbnb · AT&T · BOSCH · Chegg · CISCO
CISION · COMPASS · Deloitte. · ebay · FARFETCH · Grab · IKEA
Inflection · intuit · Microsoft · new relic · NVIDIA · OMERS
Otter.ai · PayPal · paloalto NETWORKS · POSHMARK · ROBLOX · salesforce
Shell · shutterstock · T · TREND MICRO · Walmart · ZipRecruiter · zomato

One More Thing: Zilliz Cloud

SELF MANAGED SOFTWARE

Milvus

Most widely-adopted open source vector database

milvus lite docker kubernetes

FULLY MANAGED SERVICE

Zilliz Cloud

AI Powered Search that is performant and scales

aws Google Cloud Azure

BRING YOUR OWN CLOUD

Zilliz Cloud BYOC

For Private VPCs

aws Google Cloud Azure *Coming Soon!*

milvus

Set up Once: Common API across all products regardless of architecture

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

THANKS

探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会