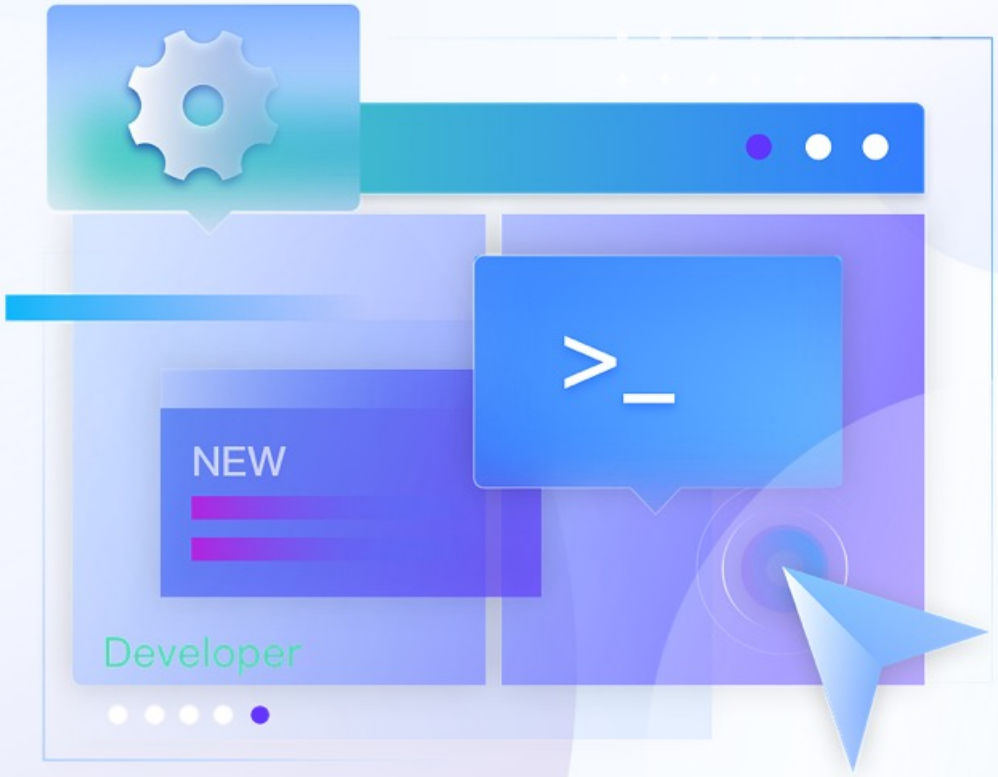
DB-GPT 在京东零售大数据平台的落地实践

从传统数据应用到智能数据应用



程方银

DB-GPT 核心开发者兼架构师
京东零售大数据平台智能化技术负责人

2024/07/06

Contents
目录

01 大模型能给大数据领域带来什么？

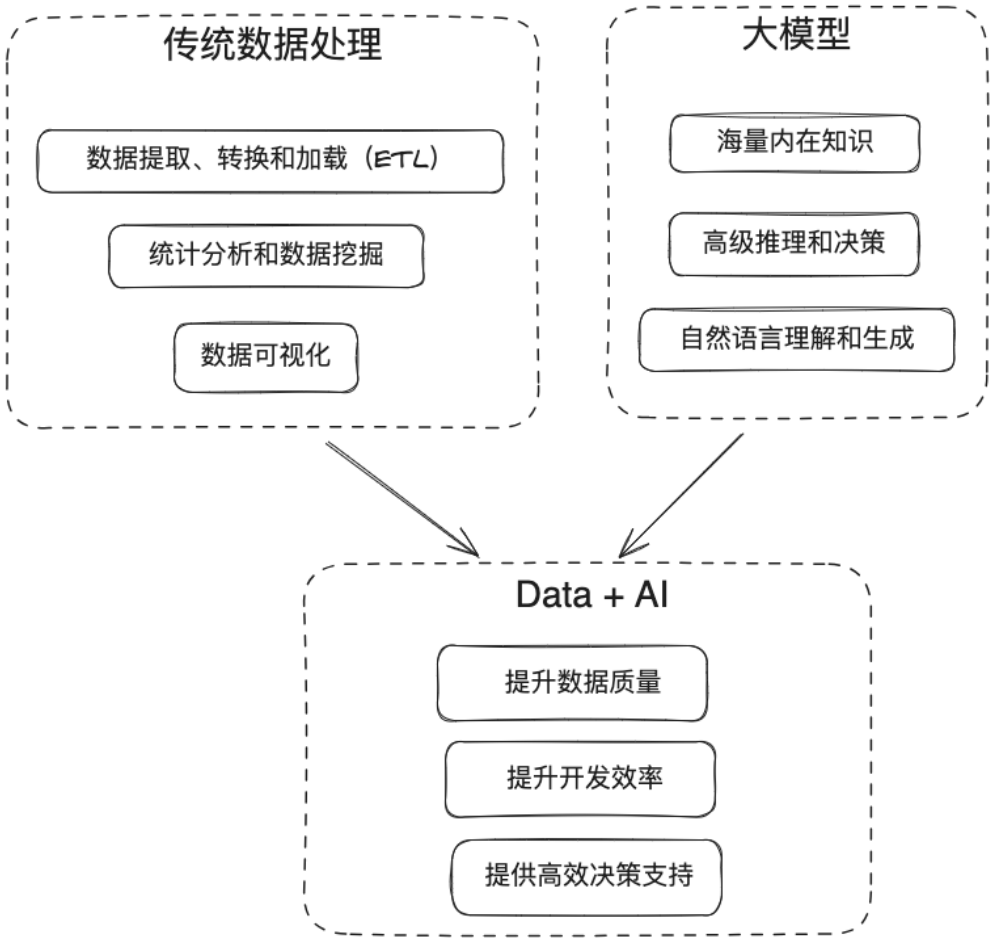
02 大数据平台智能化落地案例

03 生产部署实践

04 总结

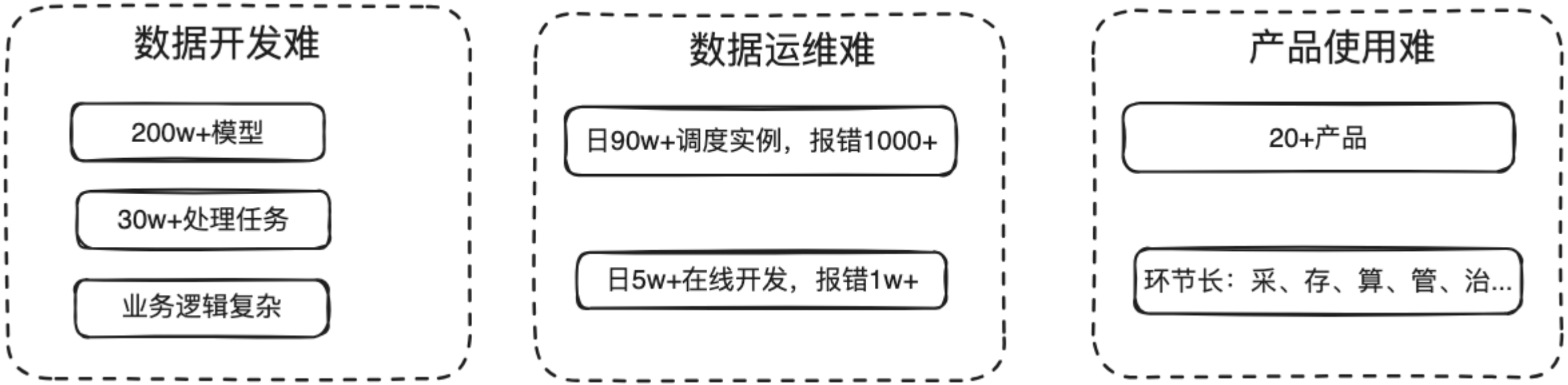
大模型给大数据领域带来什么

从传统数据应用走向智能数据应用



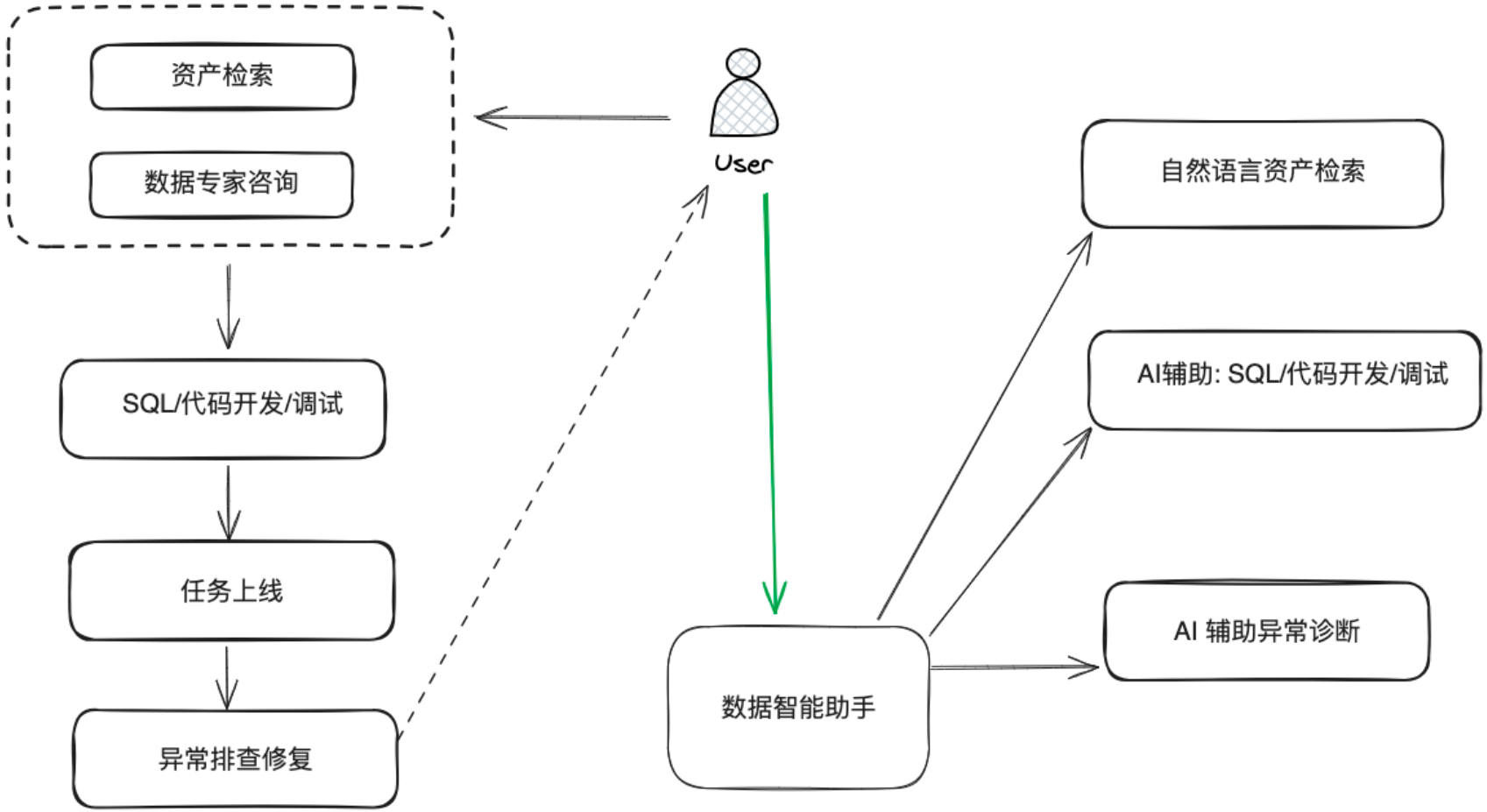
大数据平台面临哪些问题

京东零售大数据平台的“三难”



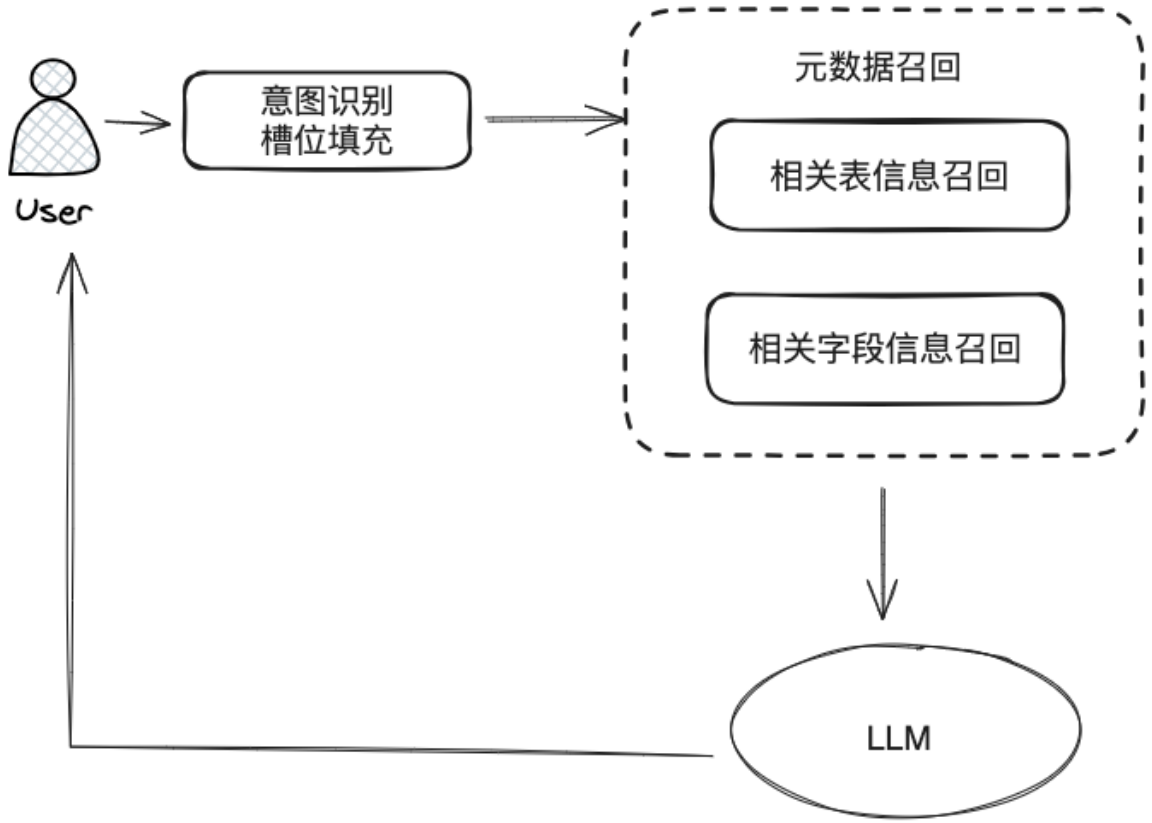
案例-智能数据开发

大模型赋能数据开发链路



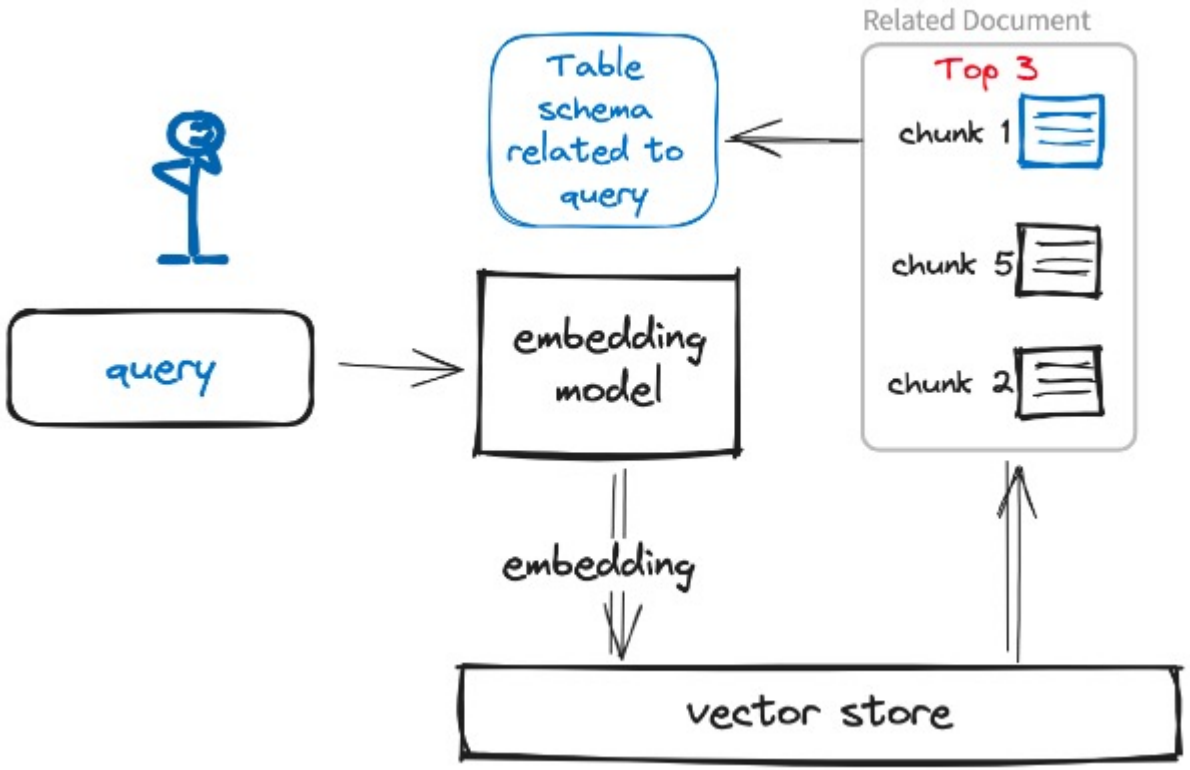
辅助数据开发

SQL开发基础流程及难点



1. 表多, 200w+表
2. 字段多, 平均30+字段, 超过1w字段30+张
3. 质量层次不齐， 一半表没有注释

表基础信息召回-朴素召回

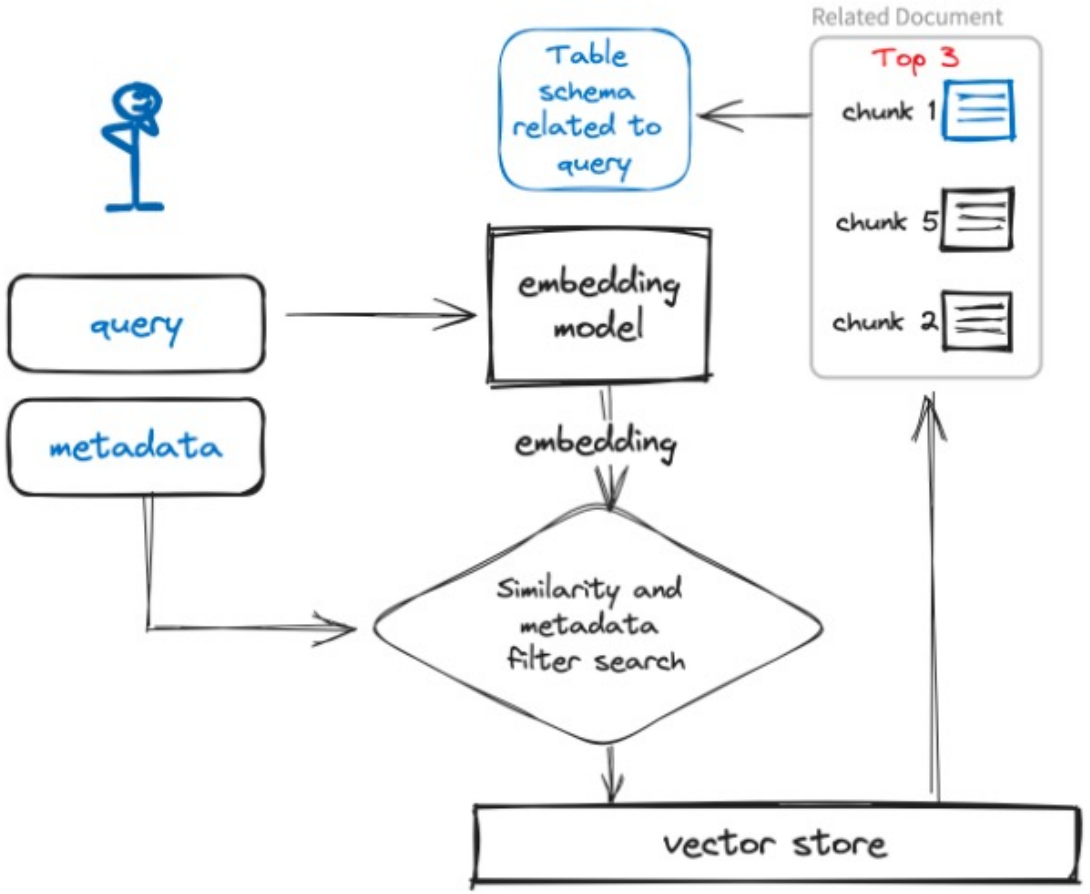


$$HitRate = \frac{\text{命中的表总数量}}{\text{测试集表总数}}$$

目前只处理 1w 张高频表

命中率：0.16

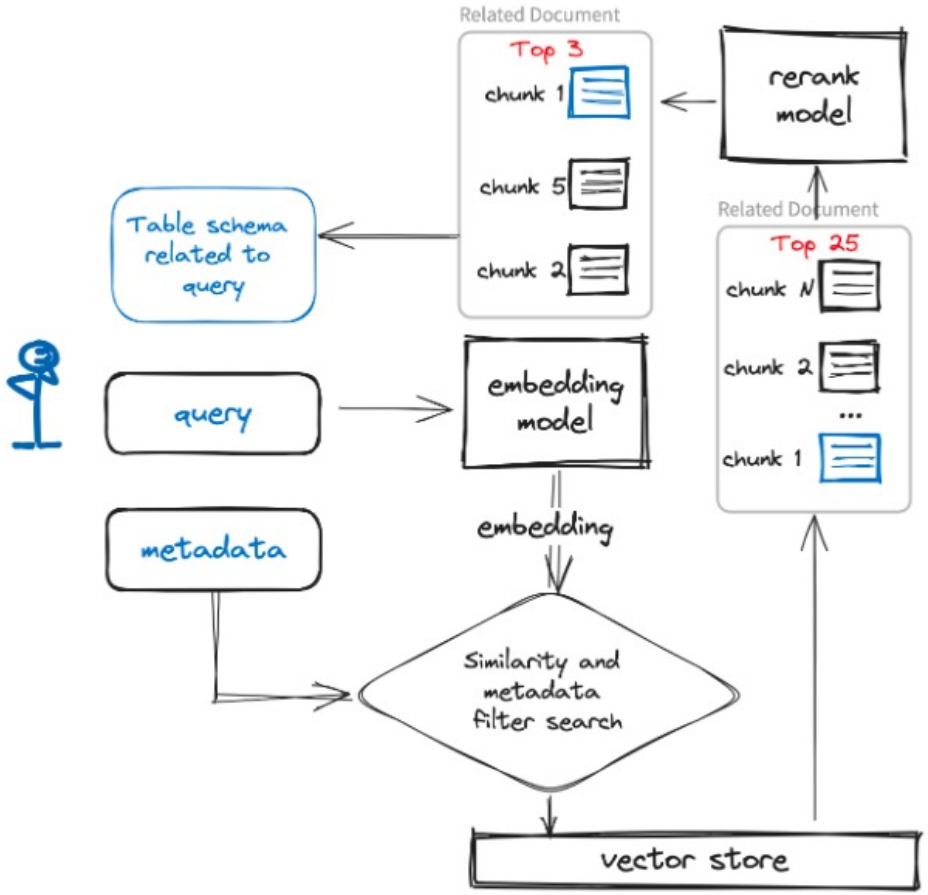
表基础信息召回-元数据索引



通过预处理构建索引
例如：
业务域：3C数码
业务主题：订单域
业务架构：京东零售

命中率：0.31

表基础信息召回-元数据索引+重排优化



通过预处理构建索引
例如：
业务域：3C数码
业务主题：订单域
业务架构：京东零售

命中率：0.51

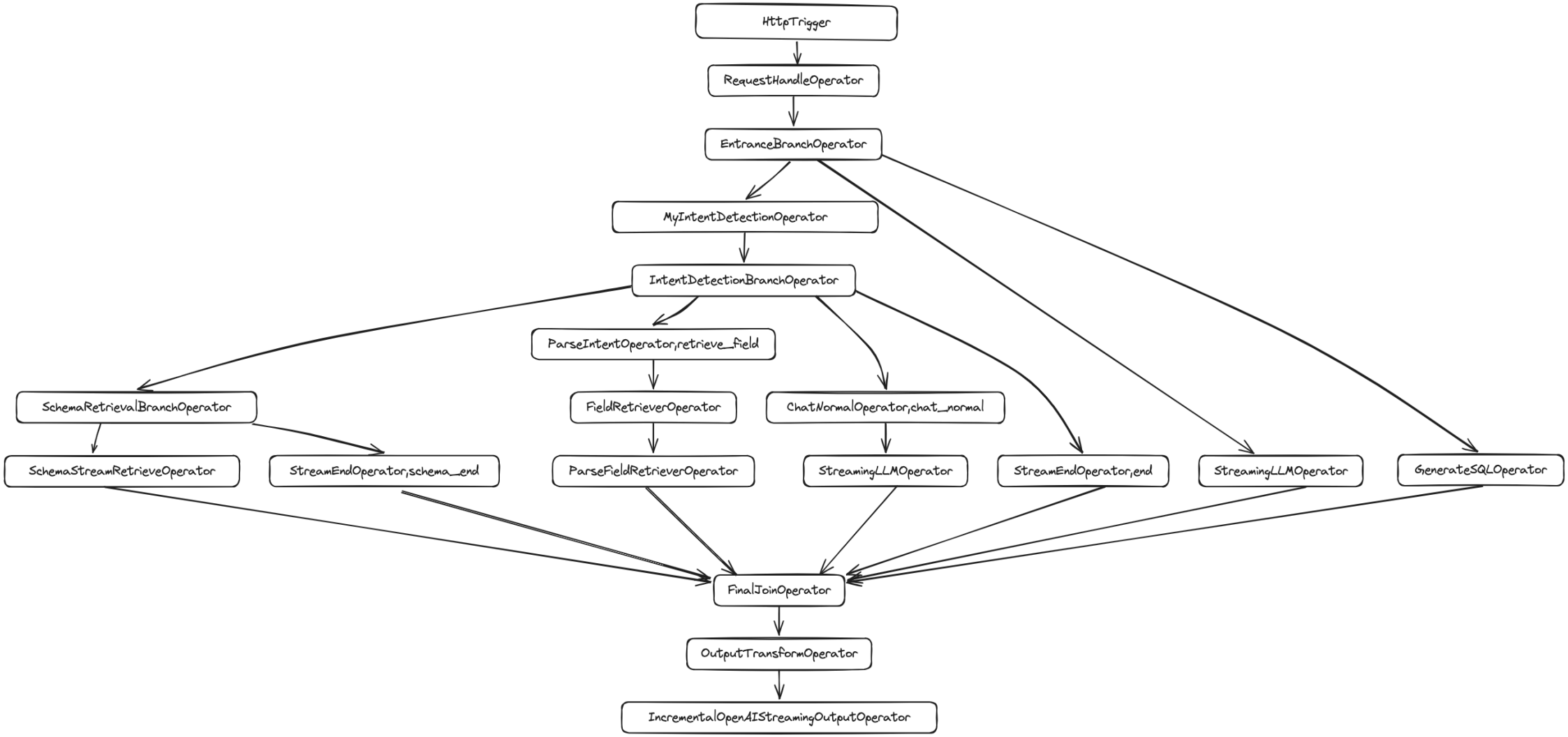
表字段信息召回

总体上与表基础信息召回相似：将字段信息拆分为 chunk，通过索引+重排优化检索。

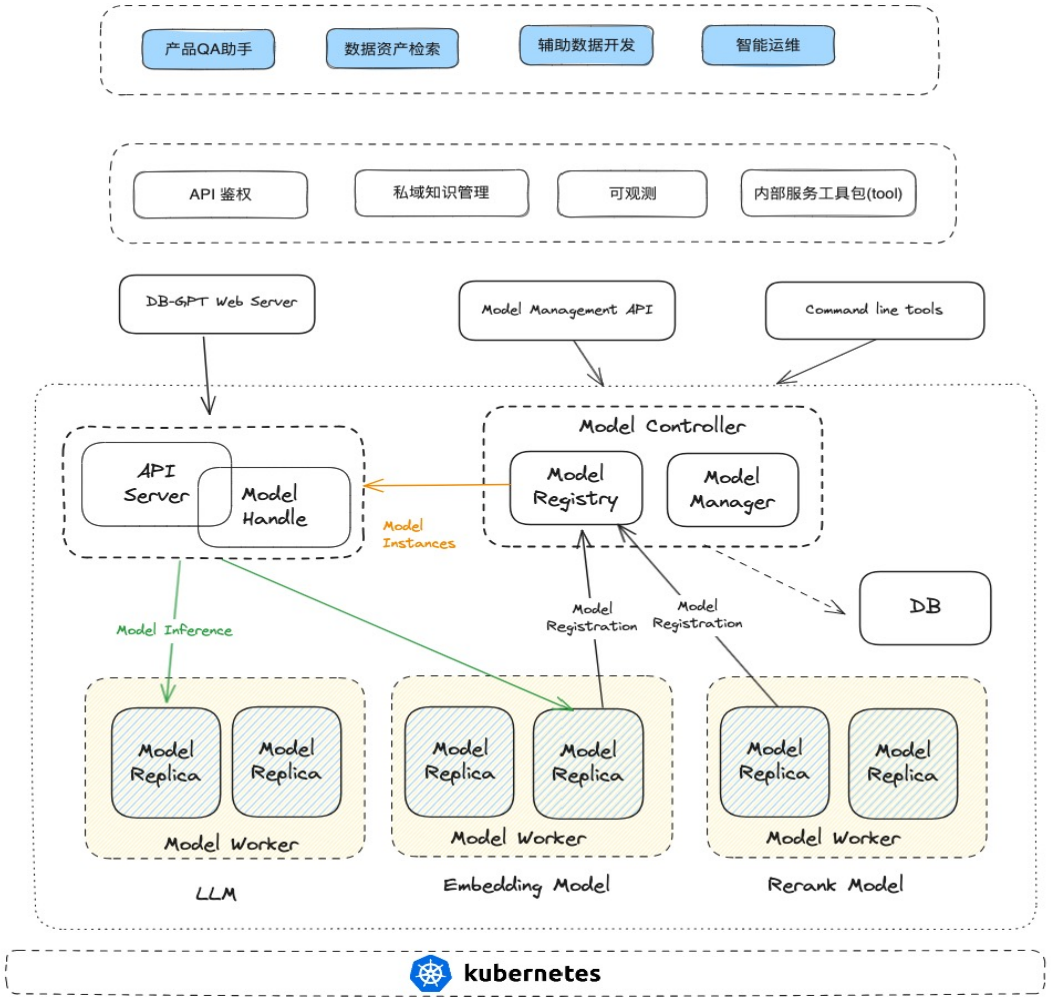
注意点：
1. chunk 大小：一般不超过 embedding 模型上下文长度
2. chunk 打标：附加表名、总长度和业务信息等
3. 字段使用频率分组排序：解析加工任务SQL，提取字段热度

命中率：0.72

辅助数据开发- AWEL workflow 概览



生产部署架构



分层建设：

DB-GPT基础设施、内部公共业务层、产品对接层

DB-GPT 集群部署：

1. 数据库（MySQL）作为模型元数据注册中心
2. Model Controller/API Server 各部署3个节点
3. Webserver 独立部署、至少3节点
4. 内部工具基础设施多节点部署，至少3节点
5. 基于 Kubernetes 部署
6. 基于 OTLP + Jaeger 实现分布式可观测能力

总结

Data + AI 三个方向：

1. 数据开发/运维/治理的智能化
2. 数据挖掘、洞察智能化
3. 数据种类多样化，实现语音/视频等数据的理解和管理

智能体工作流（Agentic Workflow）是落地最佳助手

1. 将不确定的结果流程化、确定化
2. 复杂问题分而治之
3. 抽象很重要，通过工具化/算子化抽象对世界的理解

Thank you!