

# B站基于LLM的大数据智能体实践

郭跃鹏 / 大数据架构师

张巍 / 计算平台Leader

DataFunSummit # 2024

- 背景介绍
- 架构介绍
- 技术落地
- 挑战展望

# 01

## 背景介绍

B站大数据架构介绍

# 平台架构

1. 计算平台
2. 湖仓一体
3. 存储集成
4. 资源调度
5. 平台工程



# 计算集群



## 任务量

- 27万/日 离线任务
- 2万/日 Kyuubi 任务
- 7千/日 实时任务



## 咨询量

- 上千条/周
- 3人天/周/方向

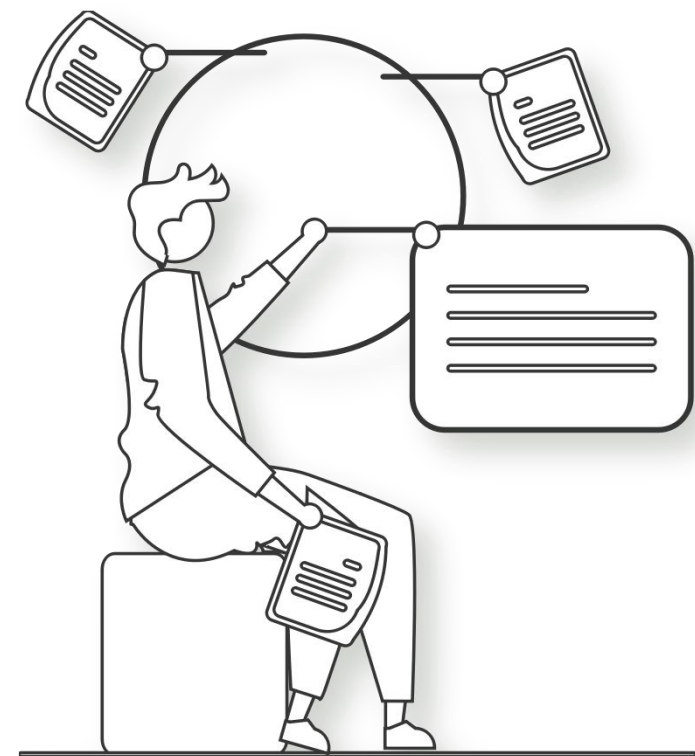
# 离线计算引擎



- 为什么任务失败
- 为什么任务变慢

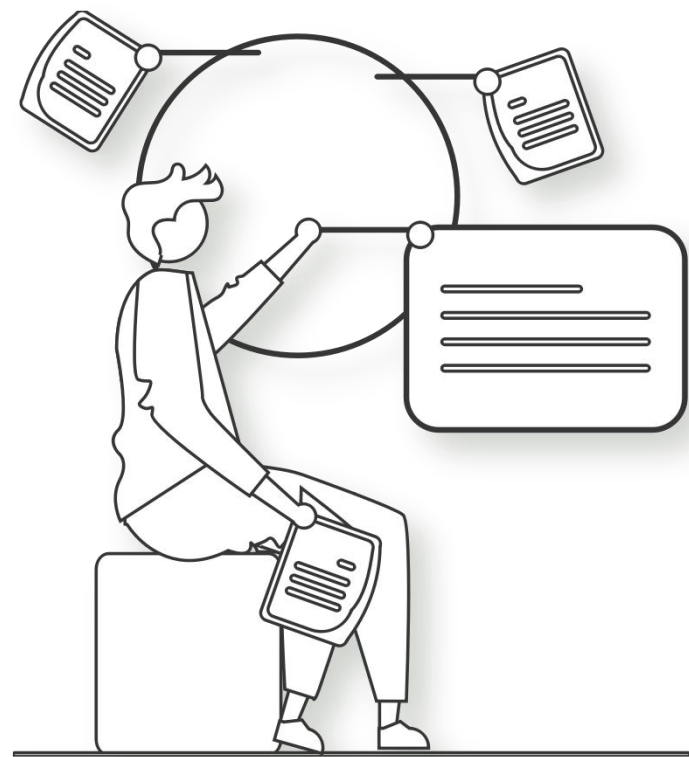
# 任务失败

- 系统内核缺陷
- 依赖组件问题
- 数据质量问题



# 任务变慢

- 硬件老化引起
- 资源调度相关
- 数据分布问题





# 用户提问的习惯

- 1. 抛链接+问题



- 2. 截图和链接+问题



# 智能小助手

- 私域咨询问题
- 诊断问题
- 时光机
- 其他Agent



# 02

## 架构介绍

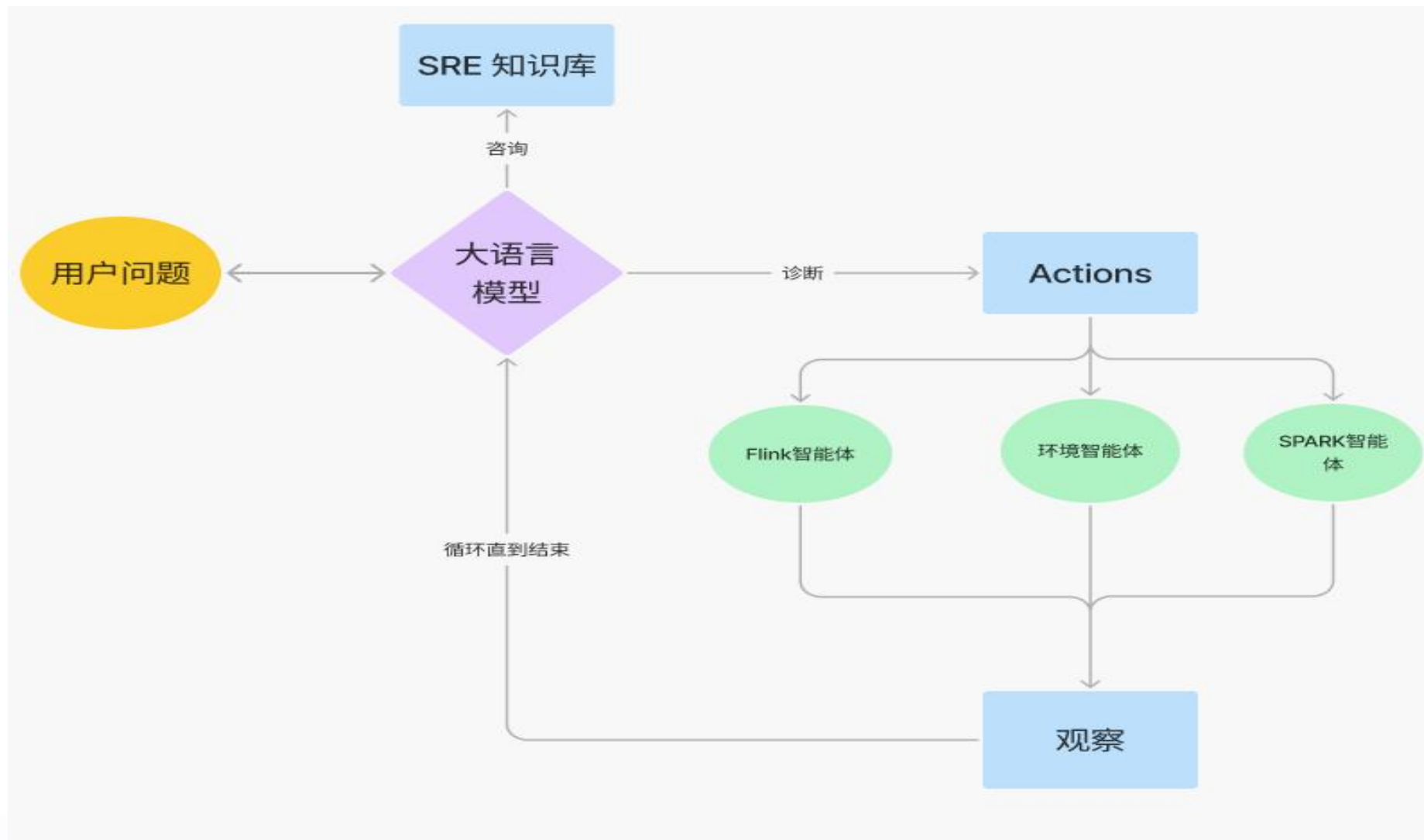
# 架构与挑战

## 架构介绍

—

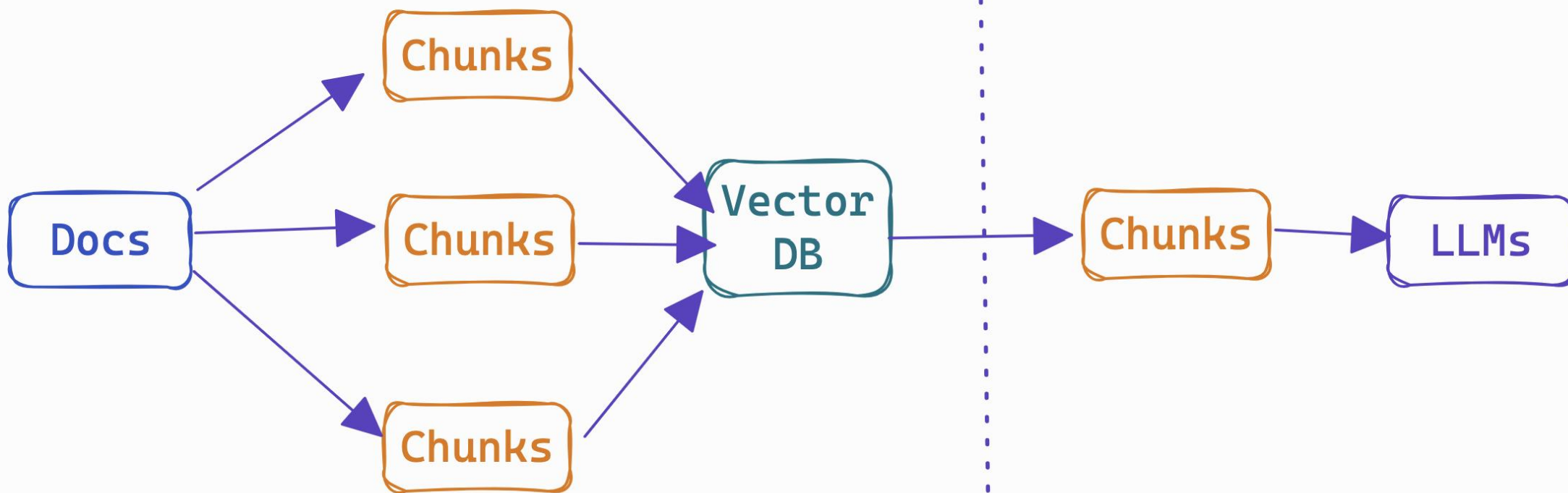
## 挑战

- precision问题
- recall问题

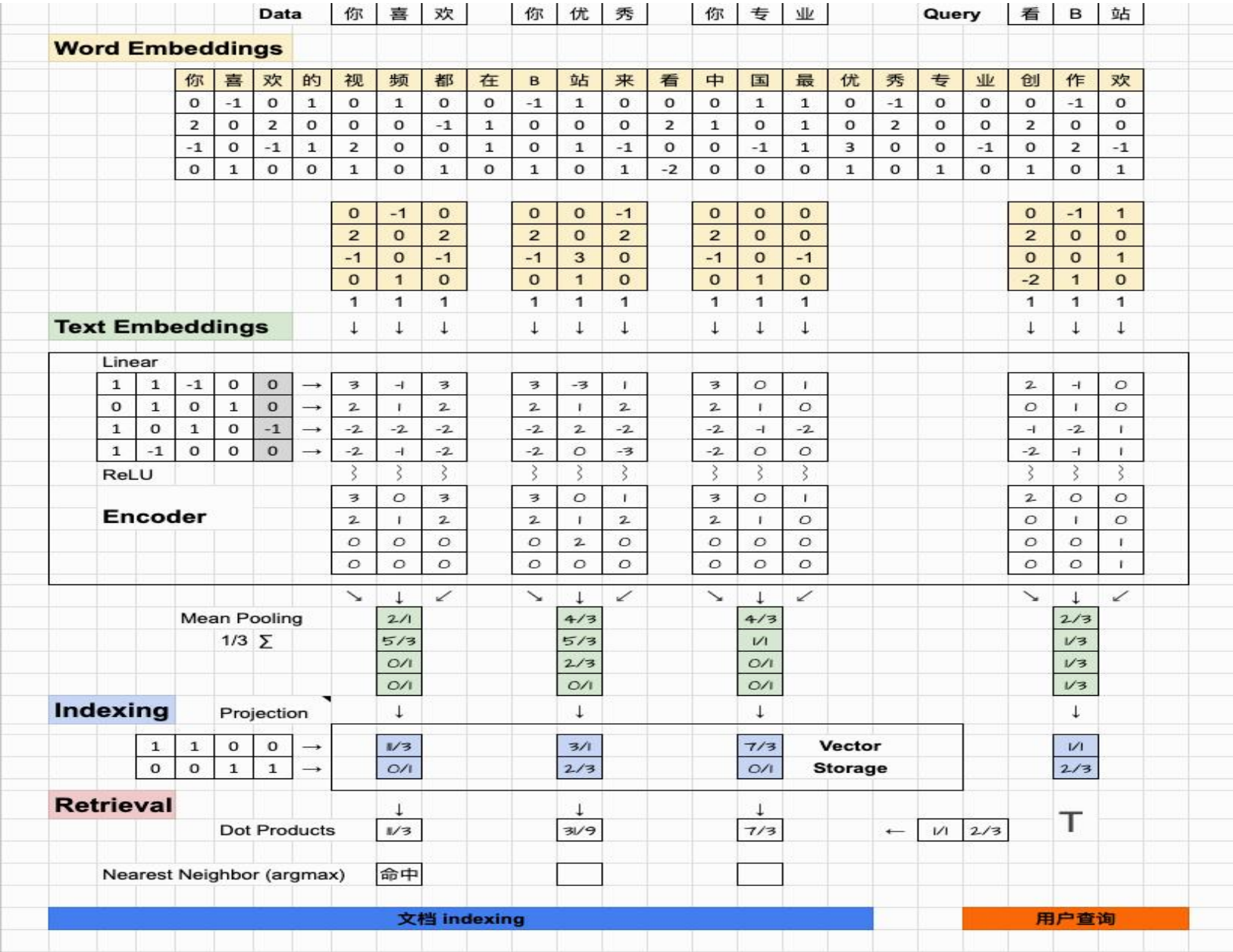


Data Indexing

Query

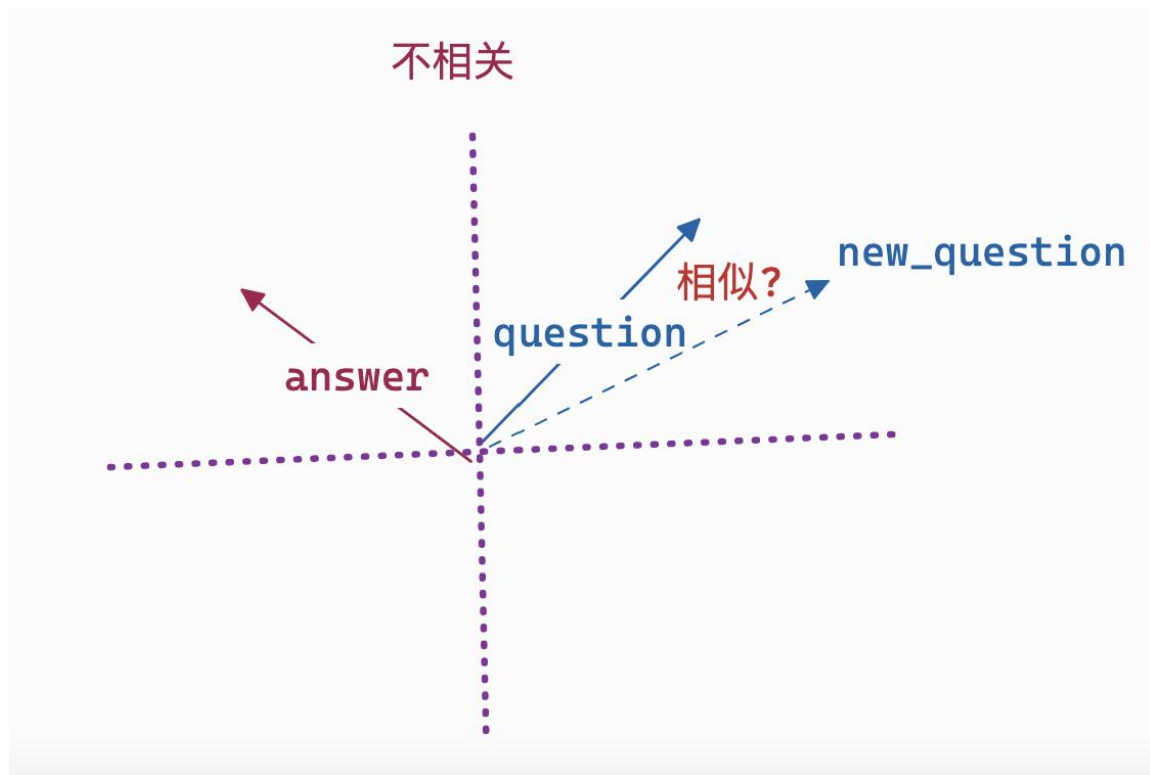


# Retrieval基础



Precision问题： 建议只embedding 问题

不要将问题和答案一起embedding





# Precision问题: 建议基于语义做chunk

- 字符级别的 splitting
- 递归字符级别的 splitting
- 句子级别 splitting
- 语义级别 splitting
- Agentic splitting

Splitter: Recursive Character Text Splitter

Chunk Size: 25

Chunk Overlap: 0

Total Characters: 604

Number of chunks: 31

Average chunk size: 19.5

在《三国：谋定天下》的沙盘里，出生州为9个：并州、凉州、益州、荆州、扬州、徐州、兖州、青州。资源州为3个：冀州、雍州、豫州。开服第三天立国玩法就会开放了，那么什么是立国玩法呢？简单理解的话，只需要占领一个出生州的州府，然后点击州府就能选择立国了。出生州的州府该怎么看呢？首先需要我们把地图逐步缩小，可以看到我们所在的出生州，每个出生州都会有一个州府，在地图上显示10级城标识。这里也为大家做了个简单的梳理，

每个出生州对应的州府如下：并州：晋阳 幽州：蓟县 凉州：下邳 益州：成都 荆州：襄阳 扬州：建业 徐州：下邳 兖州：濮阳 青州：临淄

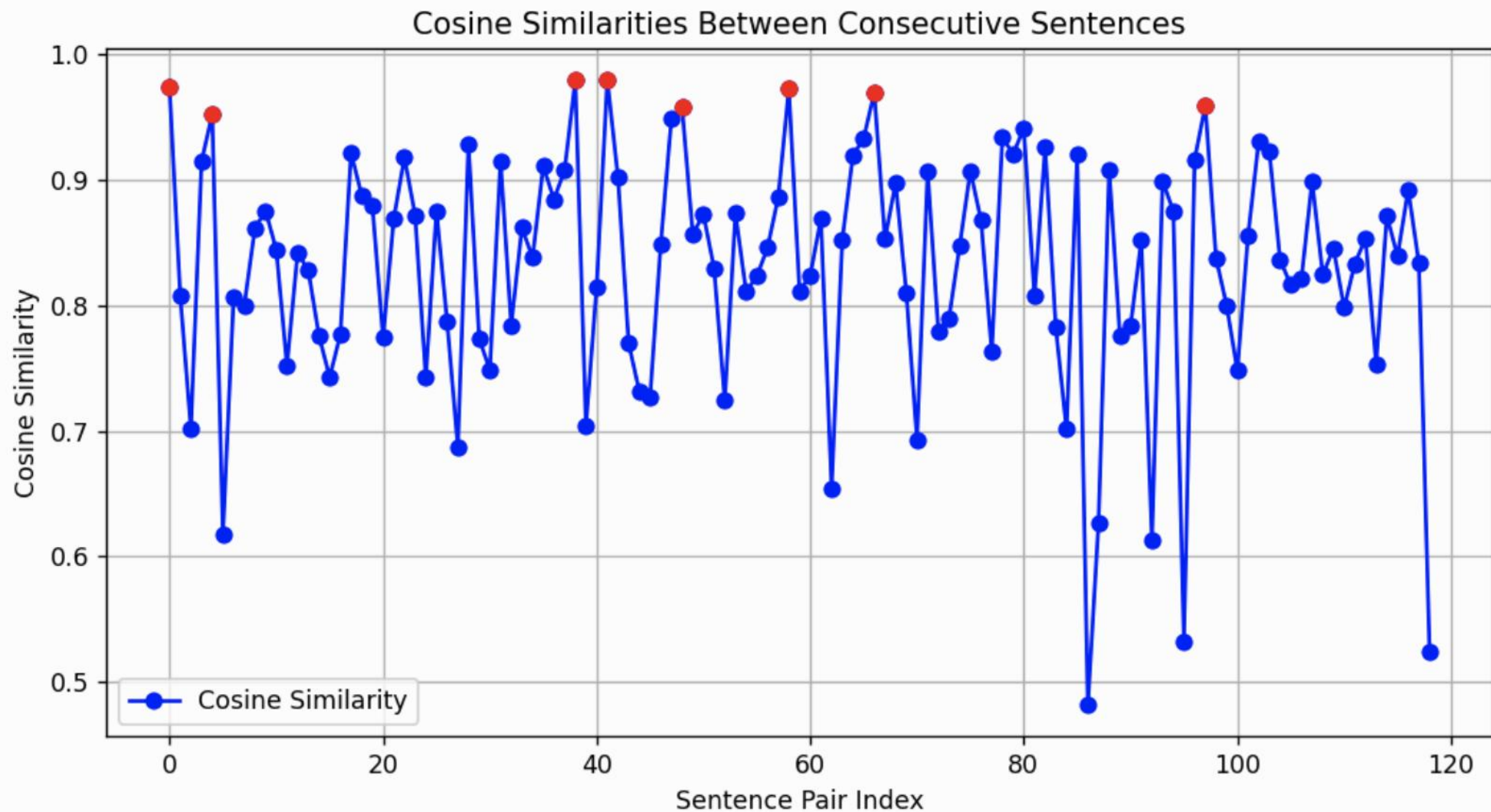
当我们把地图缩放后就可以看到每个州所对应的10级州府了。为了能够稳开州府，我们需要完成一些必要的准备工作：

1. 靠近州府的前几排低级地全部放弃，给天工打地起器械
2. 天工职业组每个人都需要造器械投石车+冲车，青囊职业组每个人都要点亮救治技能，如果没有点亮的可以选择重置点亮。
3. 每人准备3个满兵满技能的队伍，一队主力两队拆迁。投石车由天工的二队三队控制，冲车由其他职业的二队控制。
4. 打城的时候需要提前五分钟上到对应的攻城器械，防止上下车浪费时间。同时青囊在关键占位标记给到的救治也非常重要，只有齐心协力步调一致才能顺利拿下州府。

同时我们立国完毕后可以领取虎视中原的奖励，已立国的同盟成员可获得最少1000黄金，而未立国的其他同盟，成员也能获得800黄金保底。



# Precision问题：建议基于语义做chunk



# Recall问题: 建议对知识库做元数据过滤

## Metadata Filtering

```
{  
  "id": 1,  
  "engine": "SPARK",  
  "component": "HMS",  
  "sub_component": "CONSTRAINT",  
  "oncall": "离线小助手",  
  "question": " Internal error processing add_partitions_req",  
  "unique_keywords": "add_partitions_req",  
  "answer": "hive metastore 请求异常, 请联系离线小助手",  
  "explanation": ""  
},
```

Recall问题： 建议top-k之后rerank

## Rerank & Filter

- LLMRerank
- CohereRerank
- bge-reranker-v2-m3

```
def _postprocess_nodes(  
    self,  
    nodes: List[NodeWithScore],  
    query_bundle: Optional[QueryBundle] = None,  
) -> List[NodeWithScore]:  
    """Postprocess nodes."""  
  
    new_nodes = [node for node in nodes if node.score > 0]  
  
    return new_nodes
```

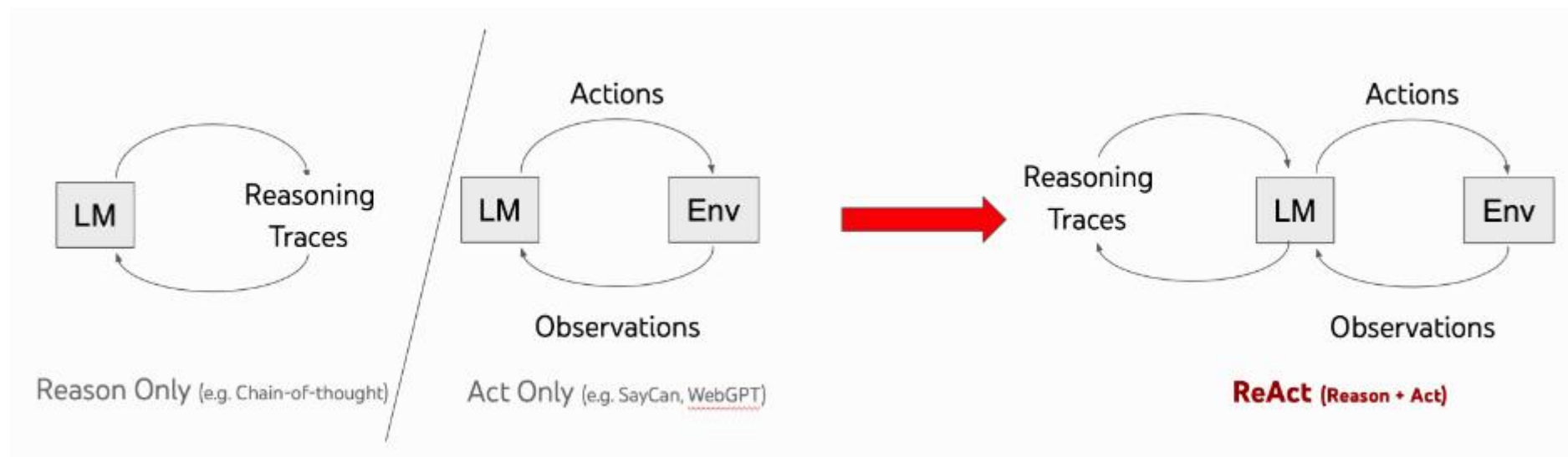


# Agent

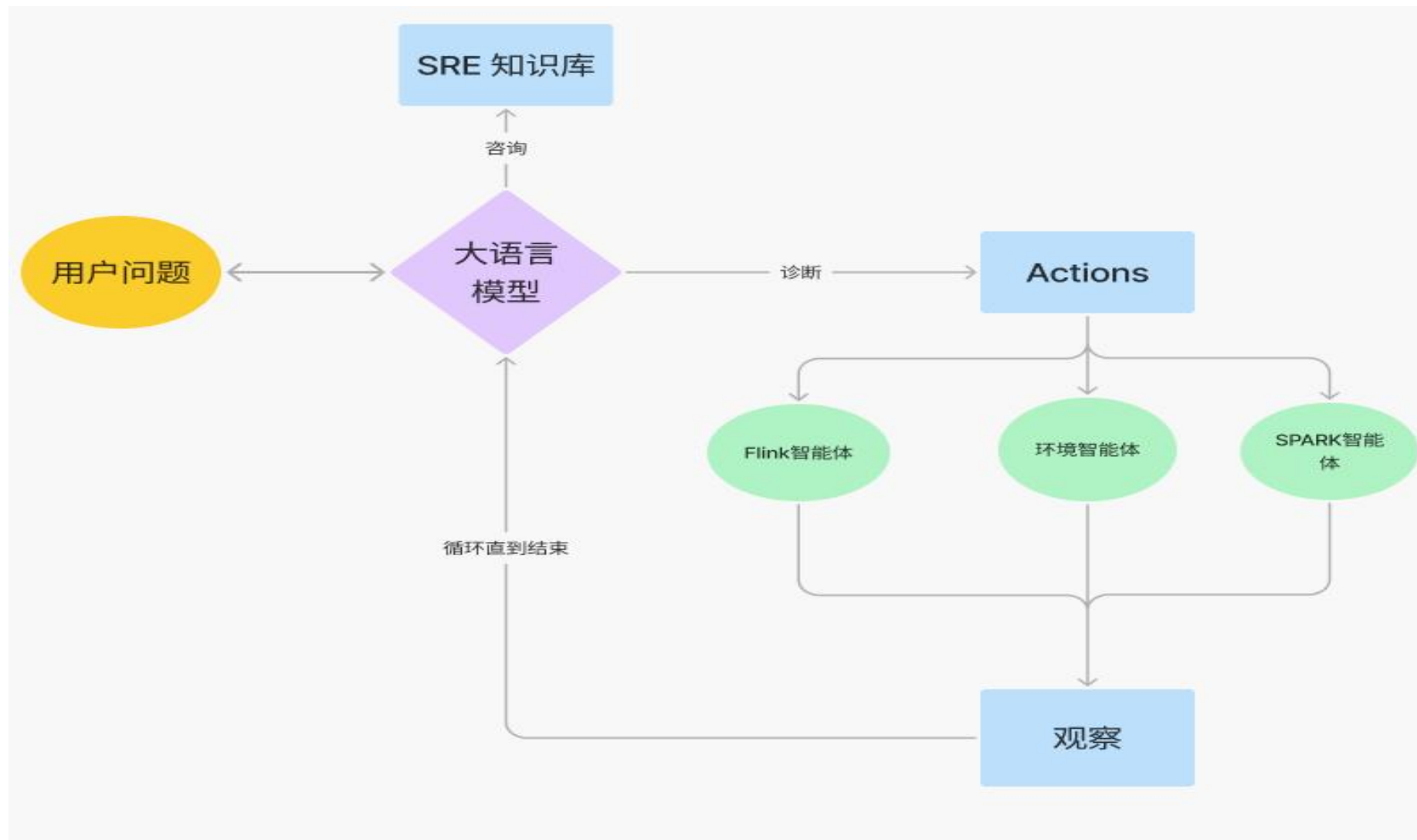
## 2024 Agent 原生应用之元年

- 自动化企业的workflow
- agent模式
  - + llm as router
  - + state management
  - + agents

# ReAct



# ReAct



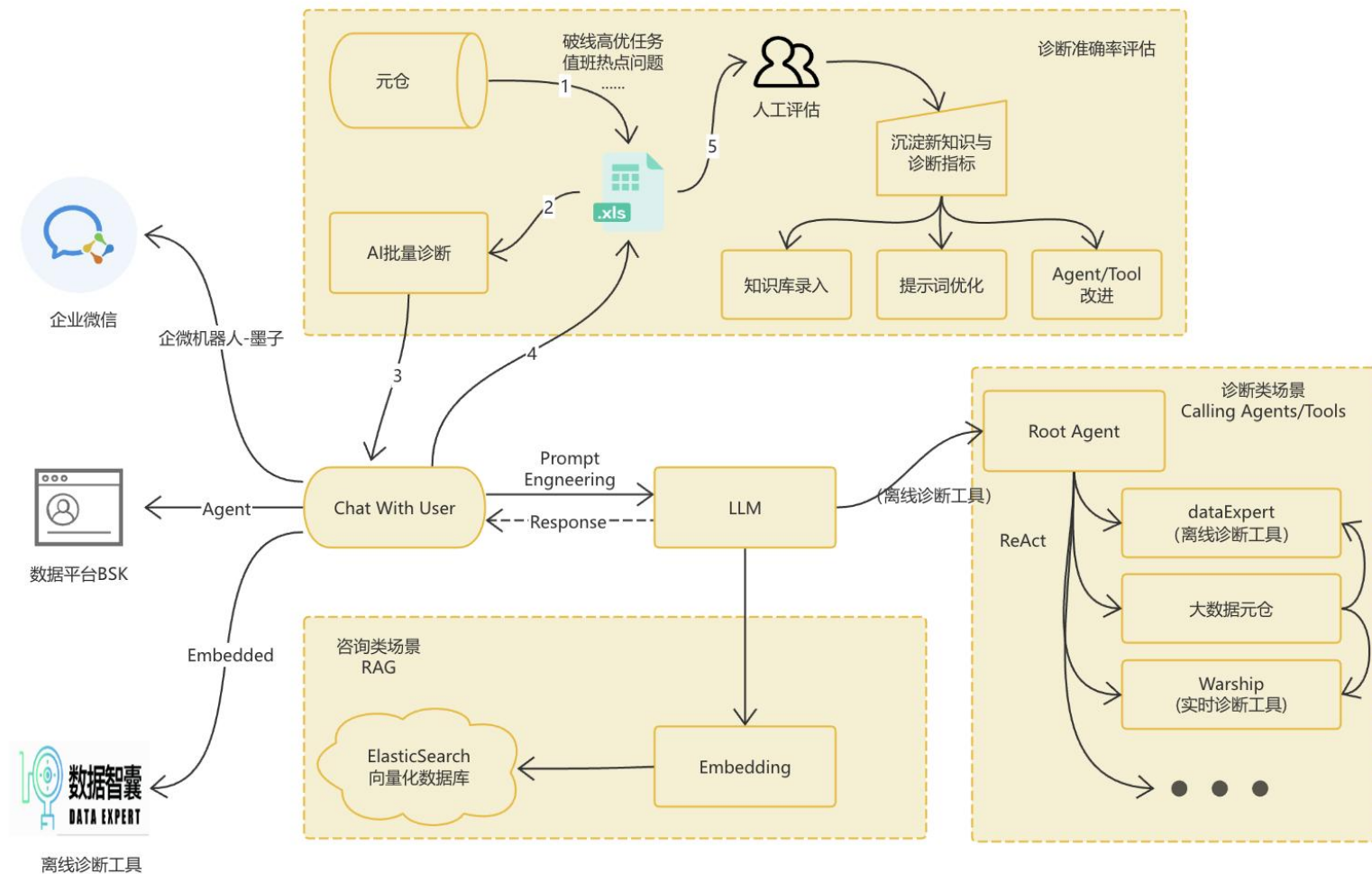
# 03

## 技术落地



# 架构设计

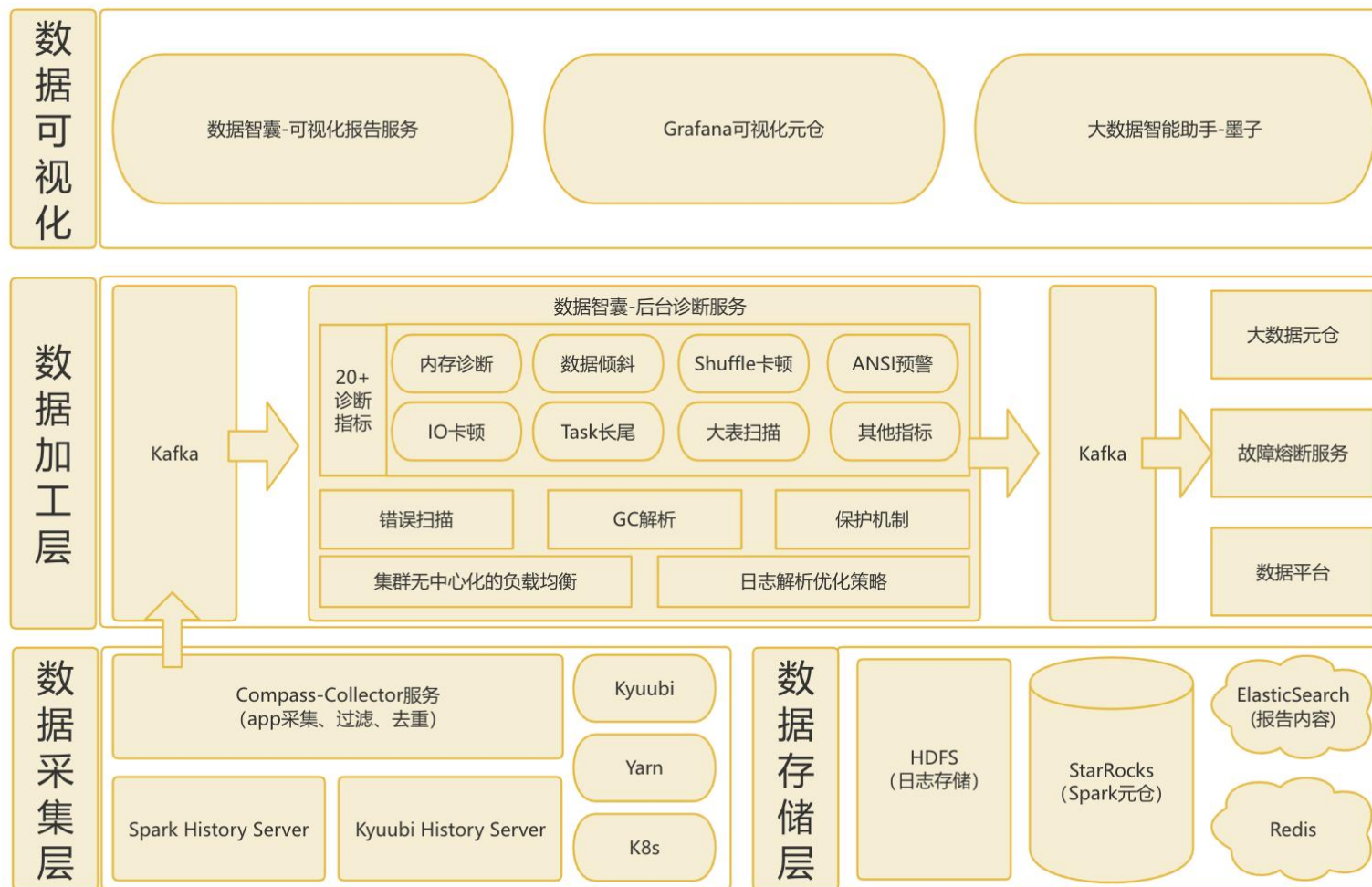
- 元仓
- 诊断
- Rag
- 评估



# 离线诊断的架构设计

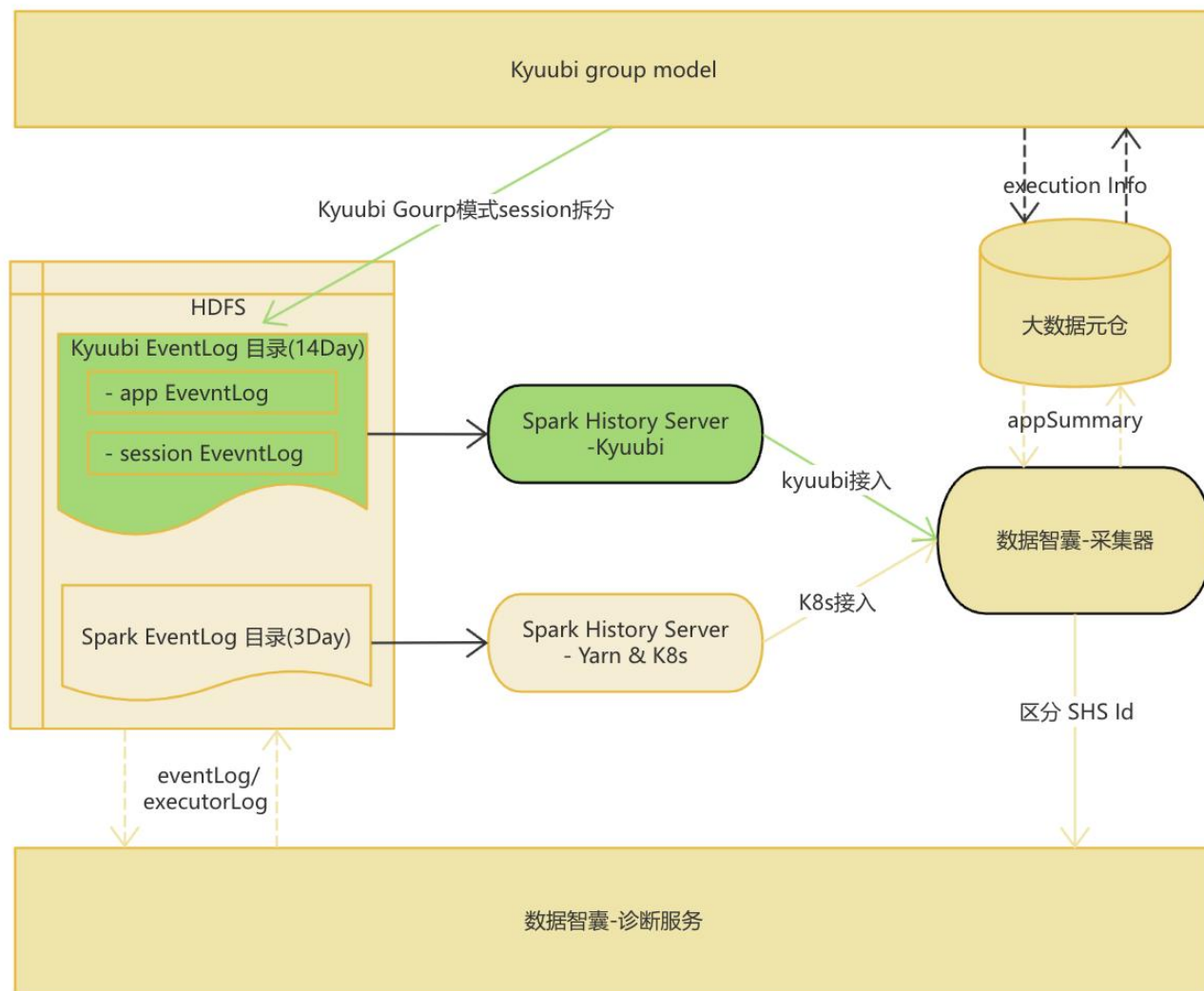


- 27万/天诊断任务
- 3千万日志/天
- Kyuubi/K8s/Yarn



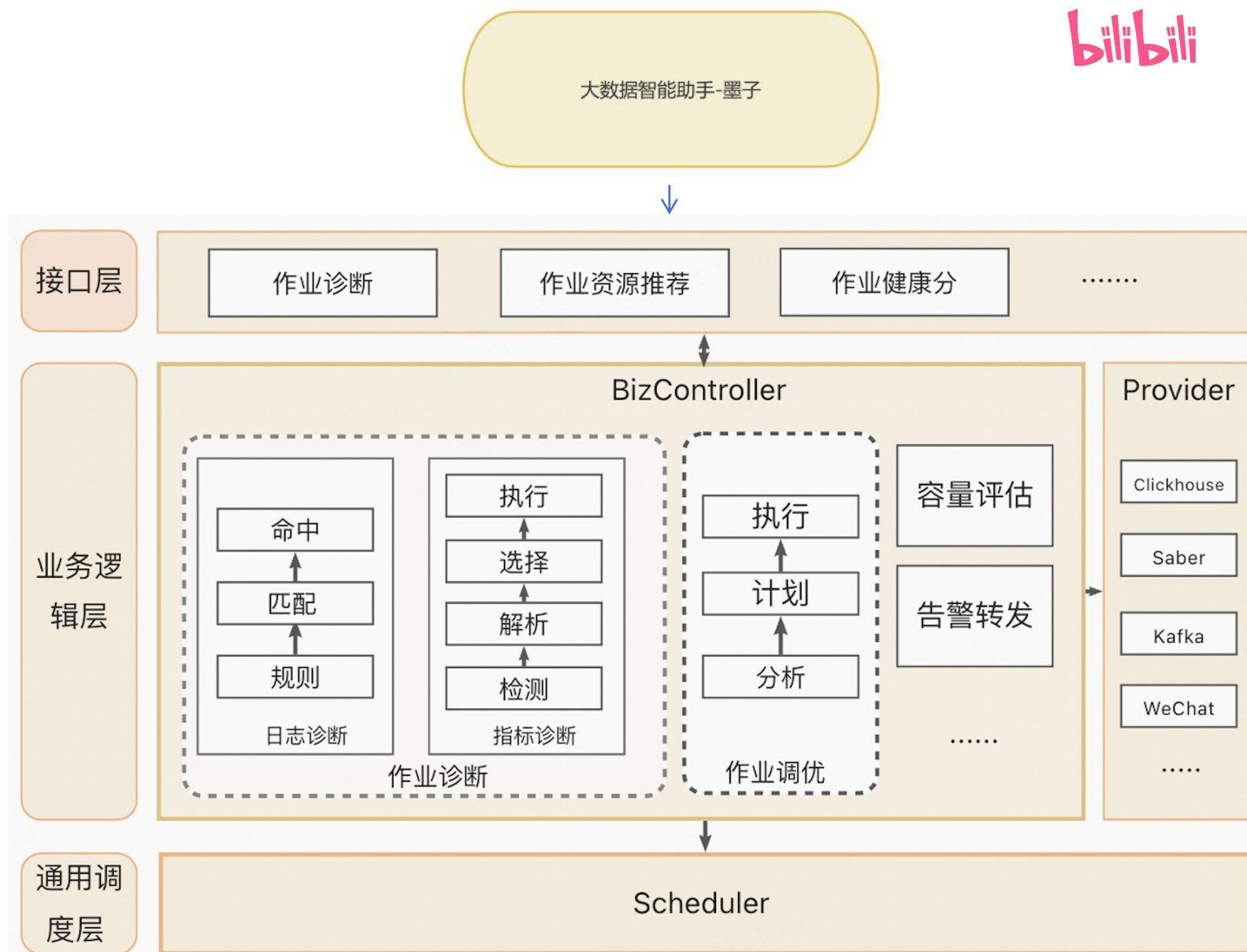
# Kyuubi 诊断

- 历史画像
- Session 拆分



# 实时诊断的架构设计

- 实时诊断
- 自动扩缩



# 智能小助手服务人群



- SRE工程师
- 组件运维人员
- 数仓专业用户
- 非相关领域用户

# SRE 用户

- SRE 工程师从大量的指标中判断中脱离出来，通过询问大模型直接给出机器潜在的问题。

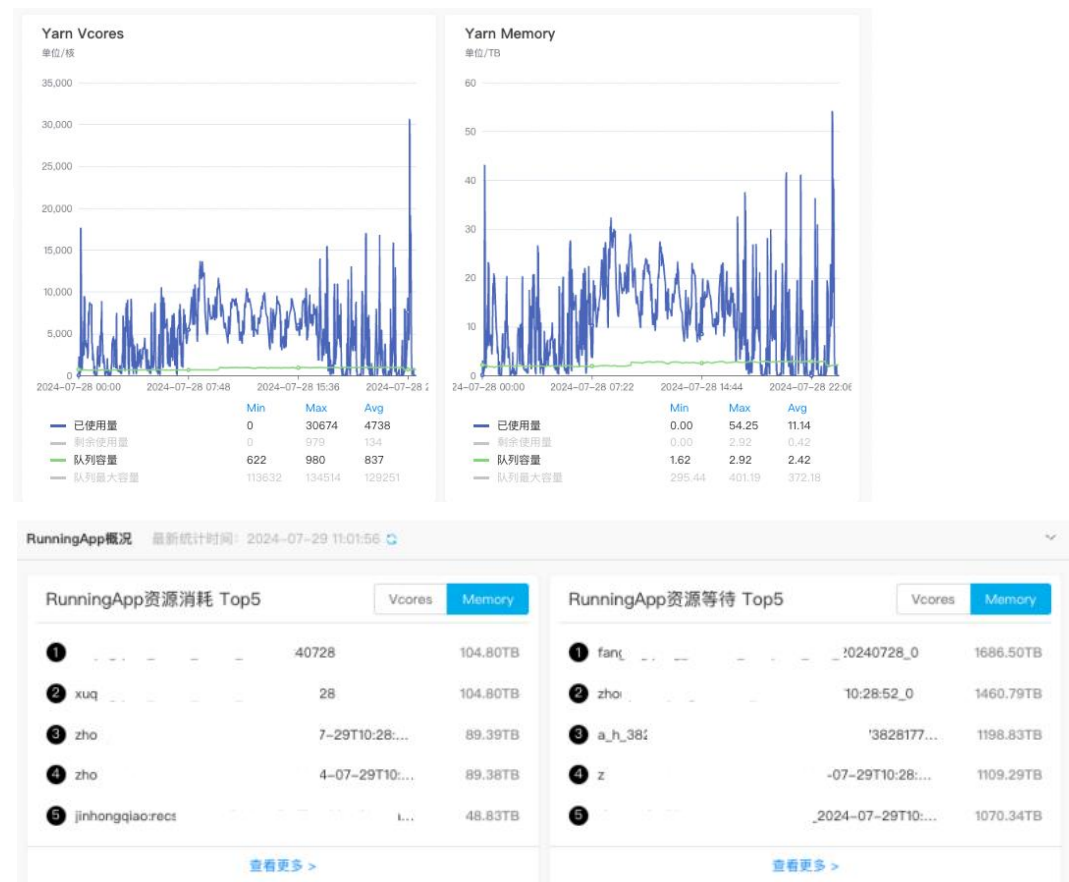




# 组件运维人员



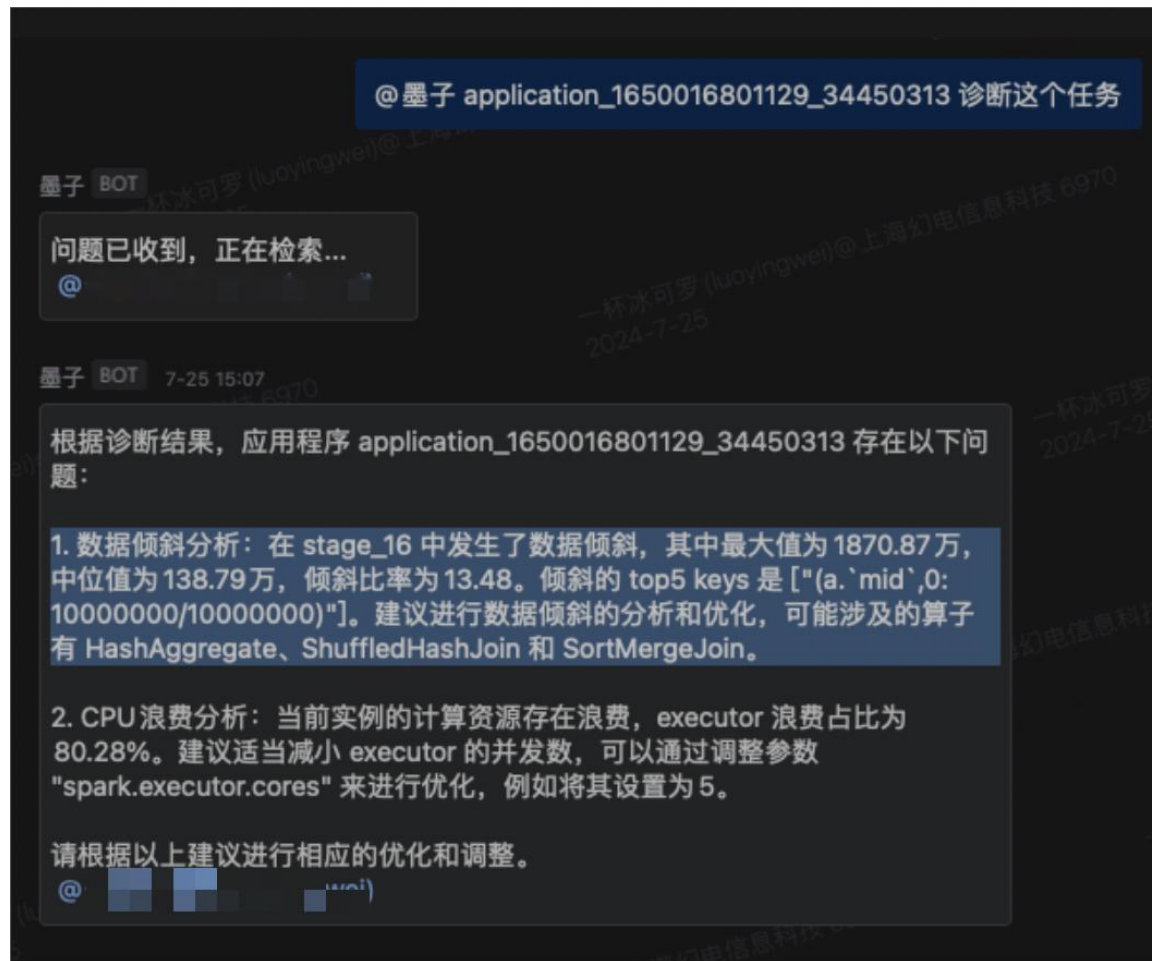
- 当发现任务慢，组件运维人员可以通过询问墨子，诊断出作业运行慢是因为资源调度出现拥塞，并能否通过可视化的看到资源拥塞时序变化



# 数仓专业用户



- 当任务变慢的时候用户可以询问智能体，智能体发现作业存在数据倾斜，并且将倾斜的key，和key对应的值展示给用户，帮助用户快速定位数据质量问题。





# 非相关领域用户



- 有些刚涉及相关领域的用户对于底层相关的专业知识也可以提供问答

大数据问答助手 BOT

回复您的这条消息：编写一个消费 kafka 数据写入到 hiv...

```
CREATE TABLE source (  
  ts AS TO_TIMESTAMP(show_timestamp 1000),  
  WATERMARK FOR ts as ts - INTERVAL '1' HOUR -- 按时间落分区 一定要在 source 端定义 Watermark  
) WITH ('offsetReset' = 'latest','blacklist.enable' = 'true') like Kafka_1_r_ai. r_ai.ai_tpc_feature_merge_wctr_new ;  
CREATE TABLE sink (  
  label_raw varchar,  
  mid bigint,  
  itemid bigint,  
  show_timestamp bigint,  
  instance varchar,  
  reserved string  
) WITH (  
  'connector' = 'bsql-hive', -- 连接器  
  'tableName' = 'test.sink_hive_catalog_test_log_date', -- 表名  
  'partitionKey' = 'log_date,log_hour', -- 分区键  
  'timeField' = 'show_timestamp', -- 指定分区字段  
  'format' = 'orc' -- 指定表格式  
) ;  
INSERT INTO  
sink  
SELECT  
label_raw,  
mid,  
itemid,  
show_timestamp 1000 AS show_timestamp,  
instance,  
reserved  
FROM  
source;
```

👍有用 | 🗑️无用

# 04

## 挑战展望

# 挑战

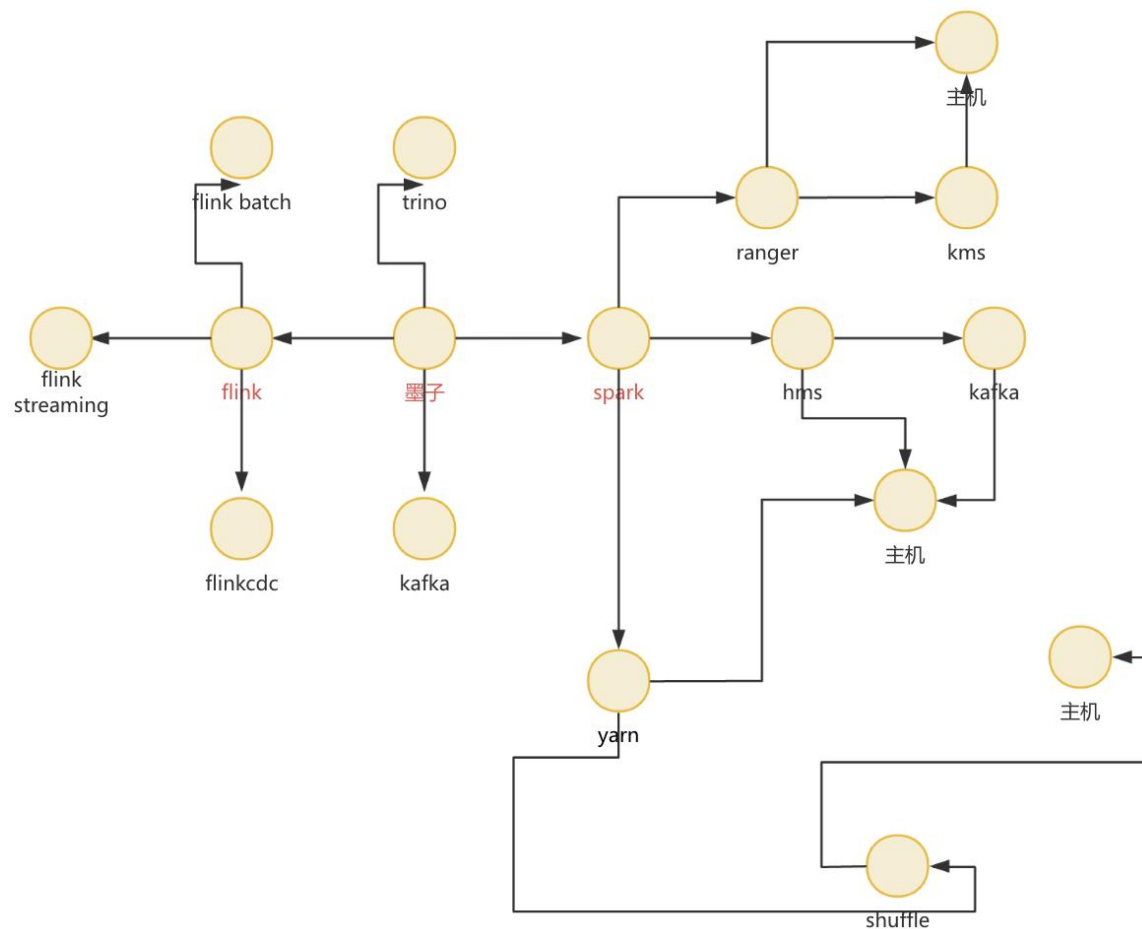
- 问答过程中准确度
- 数据质量
- 用户问题复杂多样

# 展望

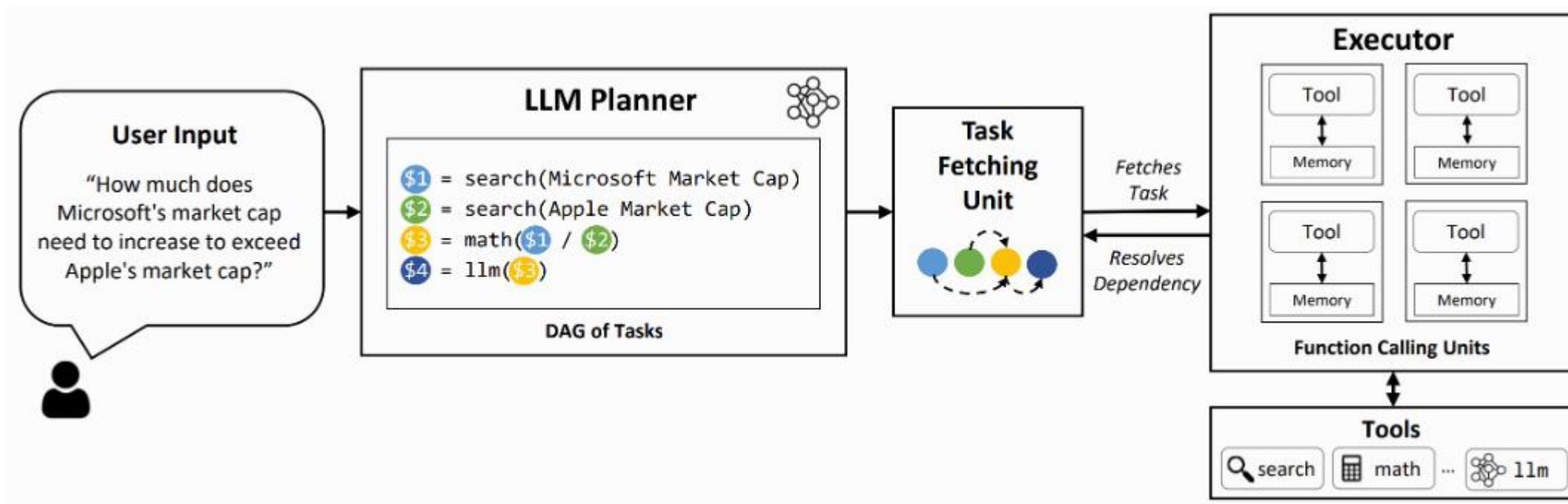
- 多专家系统
- 减少推理latency
- 提升产品体验

# MutiAgent

- 更多的组件
- 构建思维链
- 故障半径侦测



# LLMCompiler



# 欢迎关注哔哩哔哩技术



Scan to Follow

# THANKS