

# Apache Doris x AI 应用的全场景解析

张彬华@SelectDB资深解决方案架构师



# 打造全球最知名开源软件基金会 Apache 的顶级项目



Open-Source Real-Time Data Warehouse

2013 Project Creation

2017 Open Source

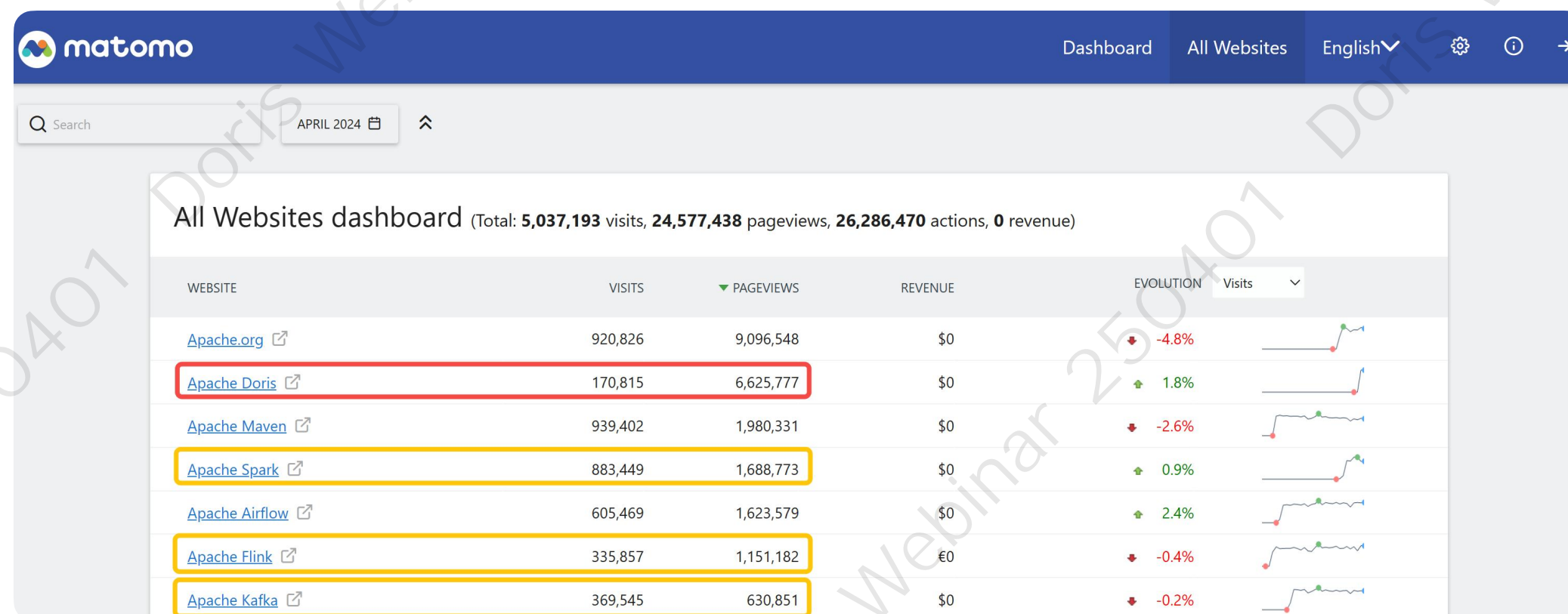
2022 ASF Top Project

13k+ GitHub Stars

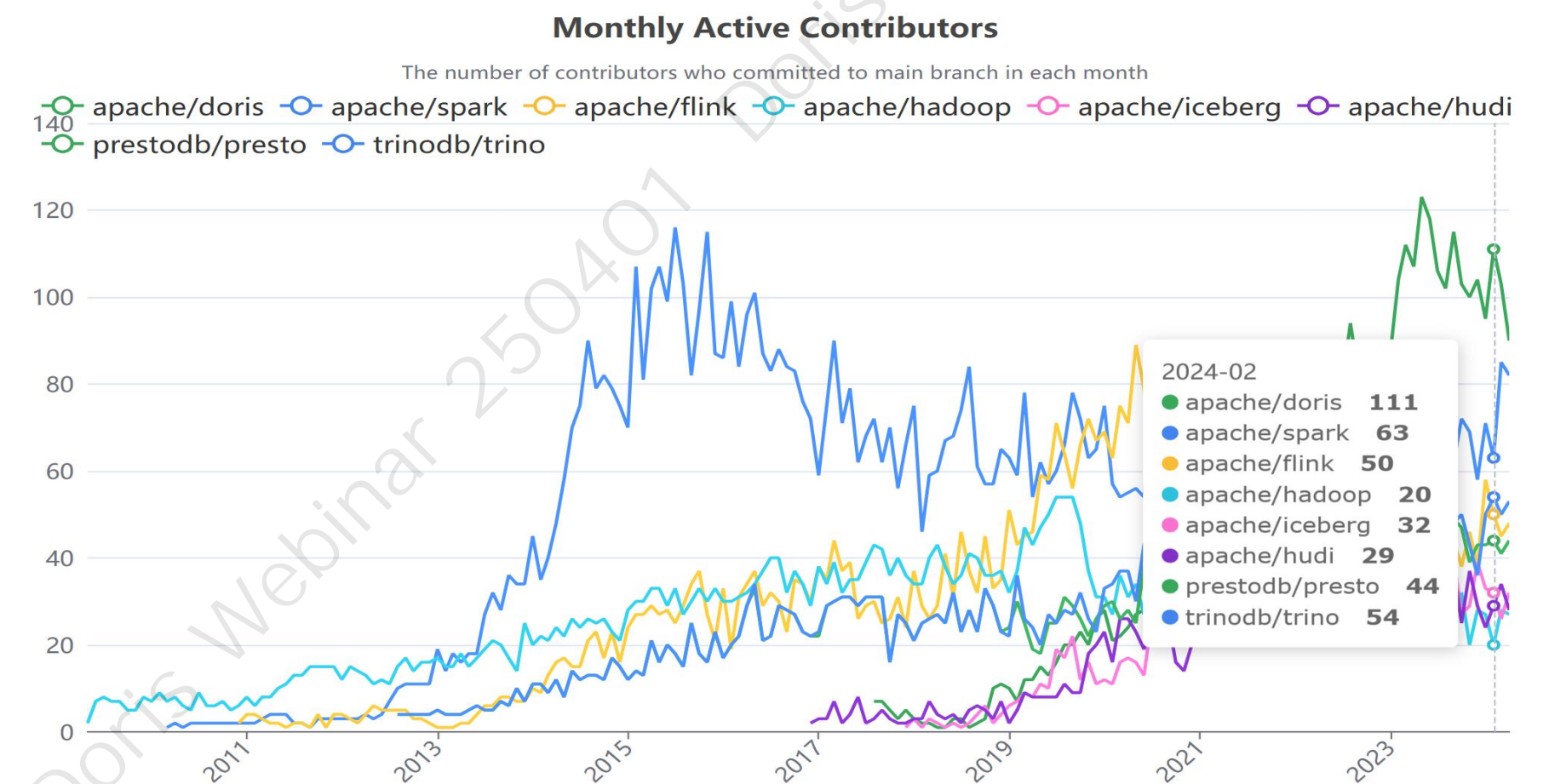
650+ Contributors

5000+ Enterprises

Apache 所有项目中 官网每日浏览量 No.1



开源大数据、数据库项目中 月活开发者 No.1



# 目录

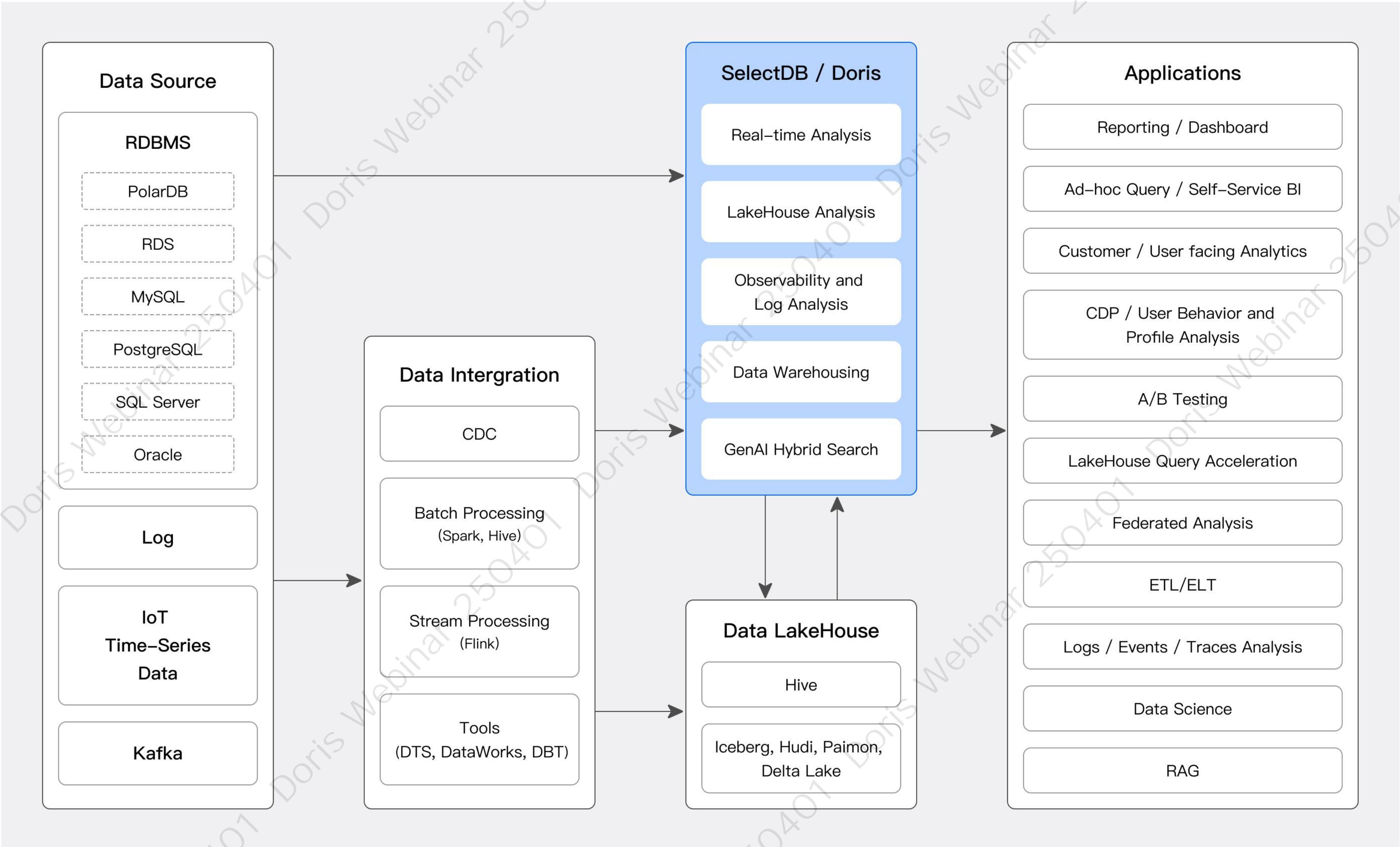
Apache Doris x AI 关键特性

Apache Doris x AI 应用场景

Apache Doris x AI 案例演示

Apache Doris x AI 未来规划

数据源经过各种数据集成和加工处理后，通常会进入实时数据仓库 Doris 和离线湖仓（如 Hive、Iceberg、Hudi、Paimon），广泛应用于 OLAP 分析场景，同时也可以作为 LLM 上下游的数据底座。



# 兼容 MySQL 的语法和协议

兼容 MySQL 协议，学习成本低，和各类 LLM 项目及 BI 工具兼容性好

使用 MySQL Client 连接

```
> mysql -h 59.110.10.92 -P 11827 -u admin

Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 119952
Server version: 5.7.37 SelectDB Core version: 2.0.0

Copyright (c) 2000, 2022, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

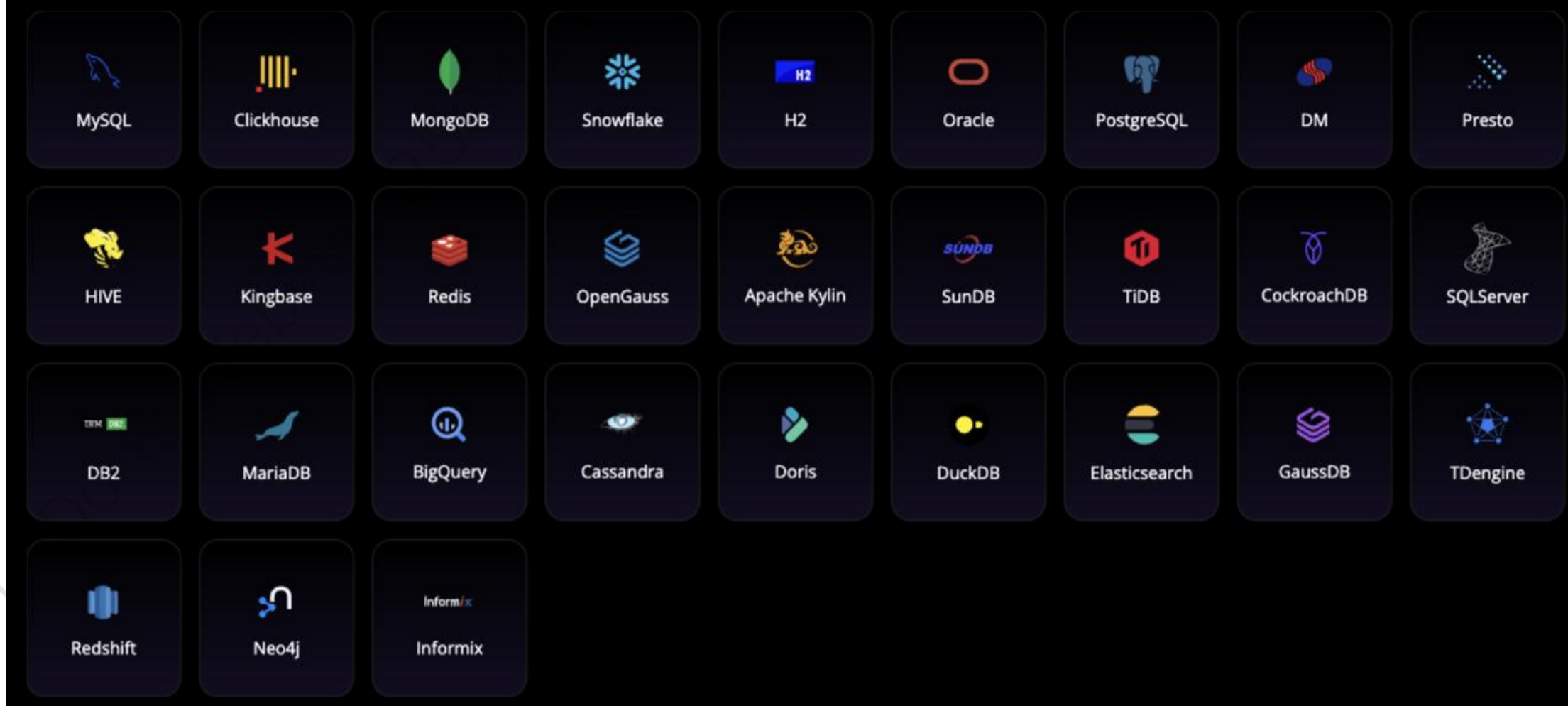
SQL Chat is built by [Next.js](#), it supports the following databases and will add more over time:

- MySQL
- PostgreSQL
- MSSQL
- TiDB Cloud

Text2SQL 项目

## 支持多种数据库

告别频繁切换工具的烦恼！现在你可以通过Chat2DB轻松连接和管理任意数据库，无论什么样的数据库类型，我们都提供了无缝集成，让你能够专注于深入分析数据并挖掘有价值的洞察。



使用 DBeaver 连接

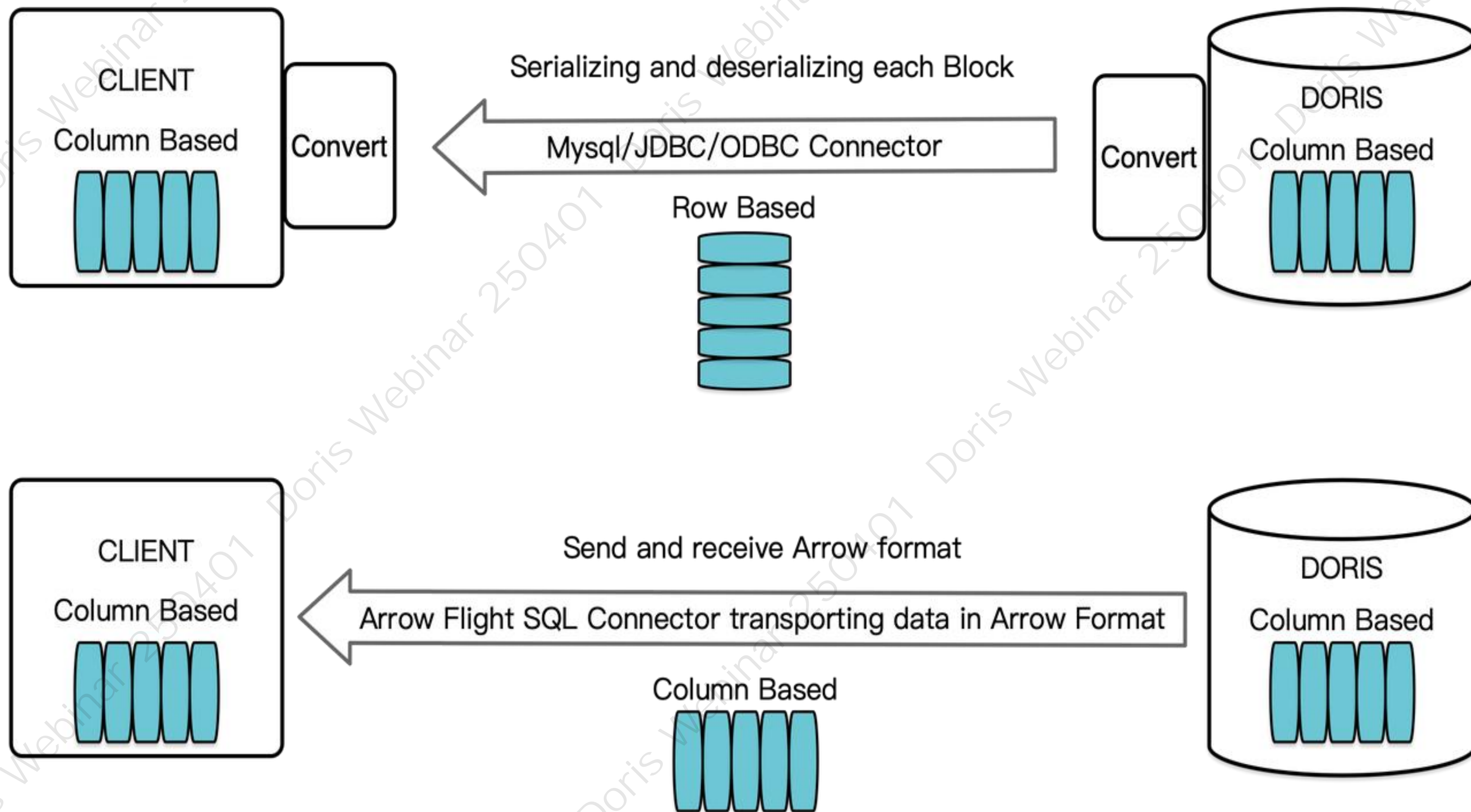


# 支持 Arrow Flight 高速读取传输服务数据科学、机器学习等AI 场景

MySQL 具有良好兼容性和广泛的工具支持，但在数据科学、大规模数据导出场景，FE 容易成为瓶颈、文本协议效率差。

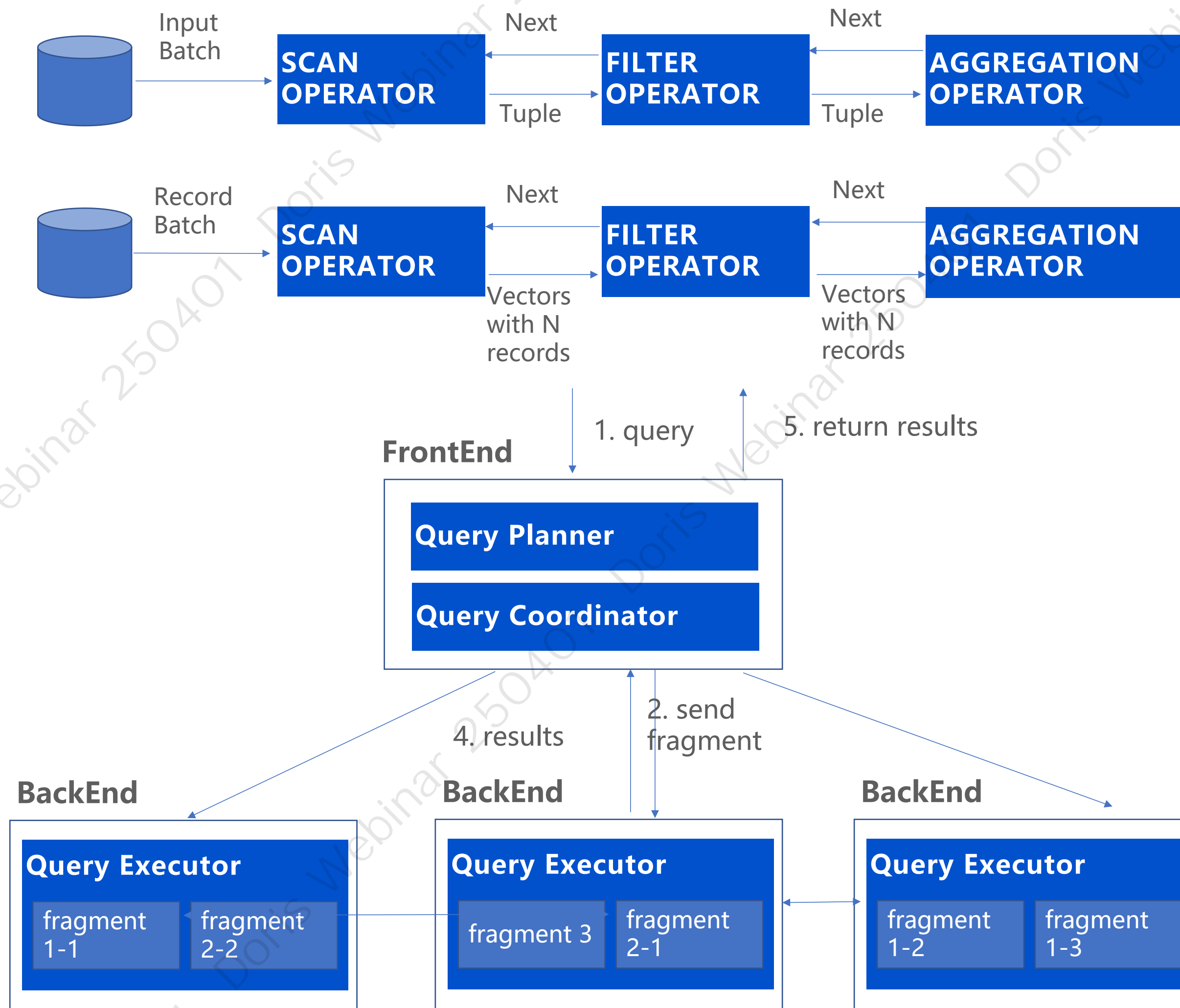
引入 Arrow Flight 实现高速数据读取、数据直接通过 BE 传递到 Pandas 客户端，列式数据传输。

Pandas ( NLP预处理/模型训练 )  
测试数据吞吐提升 100 倍！



# 极致的分析性能

|                 |  |
|-----------------|--|
| <b>执行引擎</b>     | <ul style="list-style-type: none"><li>• MPP 执行框架</li><li>• Pipeline 数据处理模型</li><li>• 向量化执行算子</li></ul> |
| <b>查询优化器</b>    | <ul style="list-style-type: none"><li>• 结合RBO和CBO，在处理多表 Join 等复杂查询时，能够游刃有余，轻松应对</li></ul>              |
| <b>缓存加速</b>     | <ul style="list-style-type: none"><li>• Data Cache</li><li>• SQL Cache</li></ul>                       |
| <b>物化视图透明加速</b> | <ul style="list-style-type: none"><li>• 同步物化视图</li><li>• 多表异步物化视图</li></ul>                            |
| <b>智能索引</b>     | <ul style="list-style-type: none"><li>• 倒排索引，全文检索（LLM日志分析）</li><li>• 自动创建前缀索引和 ZoneMap 索引</li></ul>    |



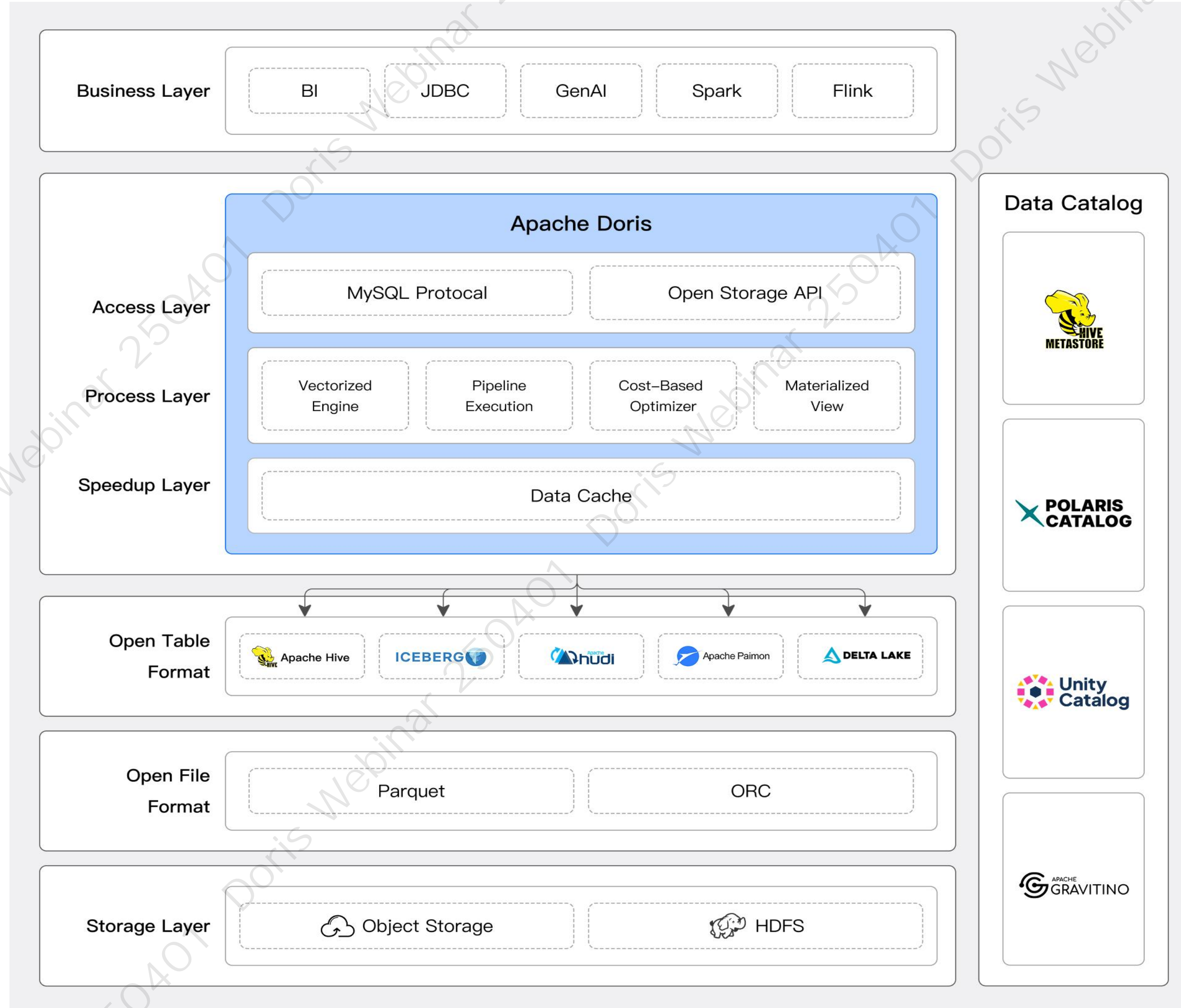
# 开放的湖仓一体

Doris 通过可扩展的连接框架、存算分离架构、数据生态开放性和极致的分析性能，为用户提供了优秀的湖仓一体解决方案。

1) 可扩展的连接框架：Doris 定义了标准的数据目录 (Catalog)、库(Database)、表(Table)三个层级，帮助开发人员快速对接企业内部特有的数据源(S3/HDFS：溯源分析/跨域模型训练)，实现数据快速互通。

2) 存算分离架构：不同时间点使用不同规模的计算资源服务业务请求，按需使用计算资源，节约成本。

3) 数据生态开放性：2.1 版本起，Doris 支持多种 SQL 方言，如 Presto、Trino、Hive、PostgreSQL、Spark、Clickhouse 等等。





# 目录

Apache Doris x AI 关键特性

Apache Doris x AI 应用场景

Apache Doris x AI 案例演示

Apache Doris x AI 未来规划

# Doris x DataAgent

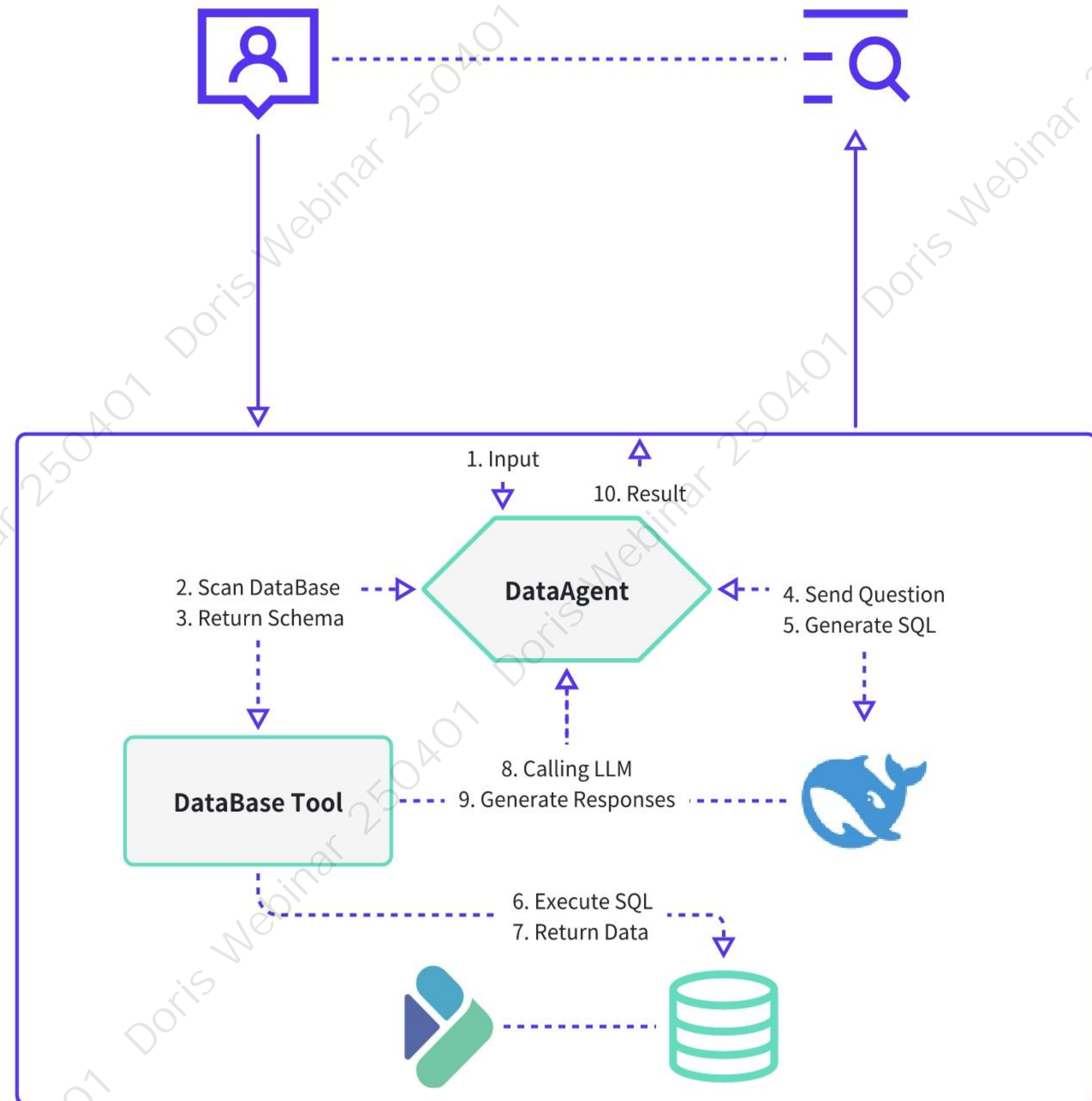
Data Agent 是一种数据智能助手，它能像数据专家一样理解你的问题，自动连接到数据库，结合 LLM 独立完成数据查询、分析和洞察。

## 应用方案

- 大模型（DeepSeek）根据输入的提示词，以及数据库相关的 Schema 信息，生成查询 SQL
- 基于 Doris，执行上一步生成的 SQL，返回得到的结果数据集
- 大模型（DeepSeek）根据用户的问题，以及第二步得到的结果数据集，生成自然语言回答

## 应用场景

- **AI 舆情分析**：Doris 存储整合社交、新闻、论坛及客户交互等基础数据，通过 DataAgent 调用大模型进行多维分析，自动进行舆情分类、紧急定义及处理建议输出，并预警潜在风险
- **实时反欺诈决策**：Doris 毫秒级处理交易流水（支持每秒百万级写入），结合 DataAgent 通过图神经网络识别异常交易模式，自动拦截高风险交易并生成风险报告



# Doris x RAG

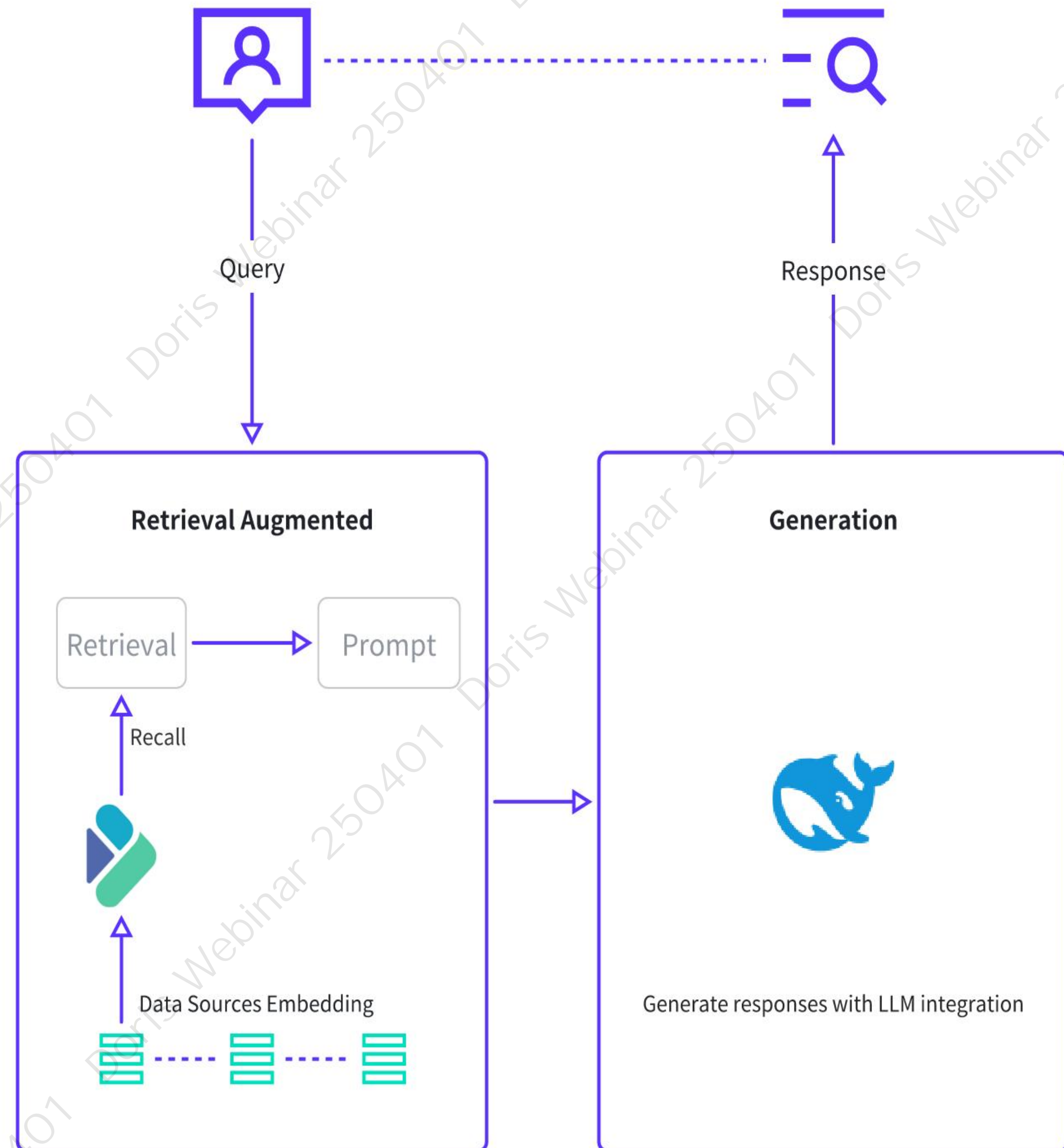
RAG (检索增强生成) 是给大模型添加一个知识库, 让它遇到问题时先查资料再回答, 好比人类解决问题前先查找相关信息一样, 也相当于开卷考试。

## 应用方案

- 本地数据文件进行向量化 Embedding 后存储至 Doris 中
- 结合 Doris (高性能查询) 与 **distance** 向量函数 (相似检索), 实现召回
- 大模型 (DeepSeek) 选择最合适的数据, 生成回答

## 应用场景

- **问答系统**: RAG 在问答系统中用于处理复杂查询与长尾知识。例如法律咨询场景, 系统可检索法律条文、案例库及历史咨询记录, 生成带引用来源的答案。
- **企业知识库**: RAG 技术通过结合检索与生成能力, 解决传统知识库 FAQ 维护成本高、结构化/非结构化数据难以混合、知识更新滞后等问题。



# Doris x ChatBI

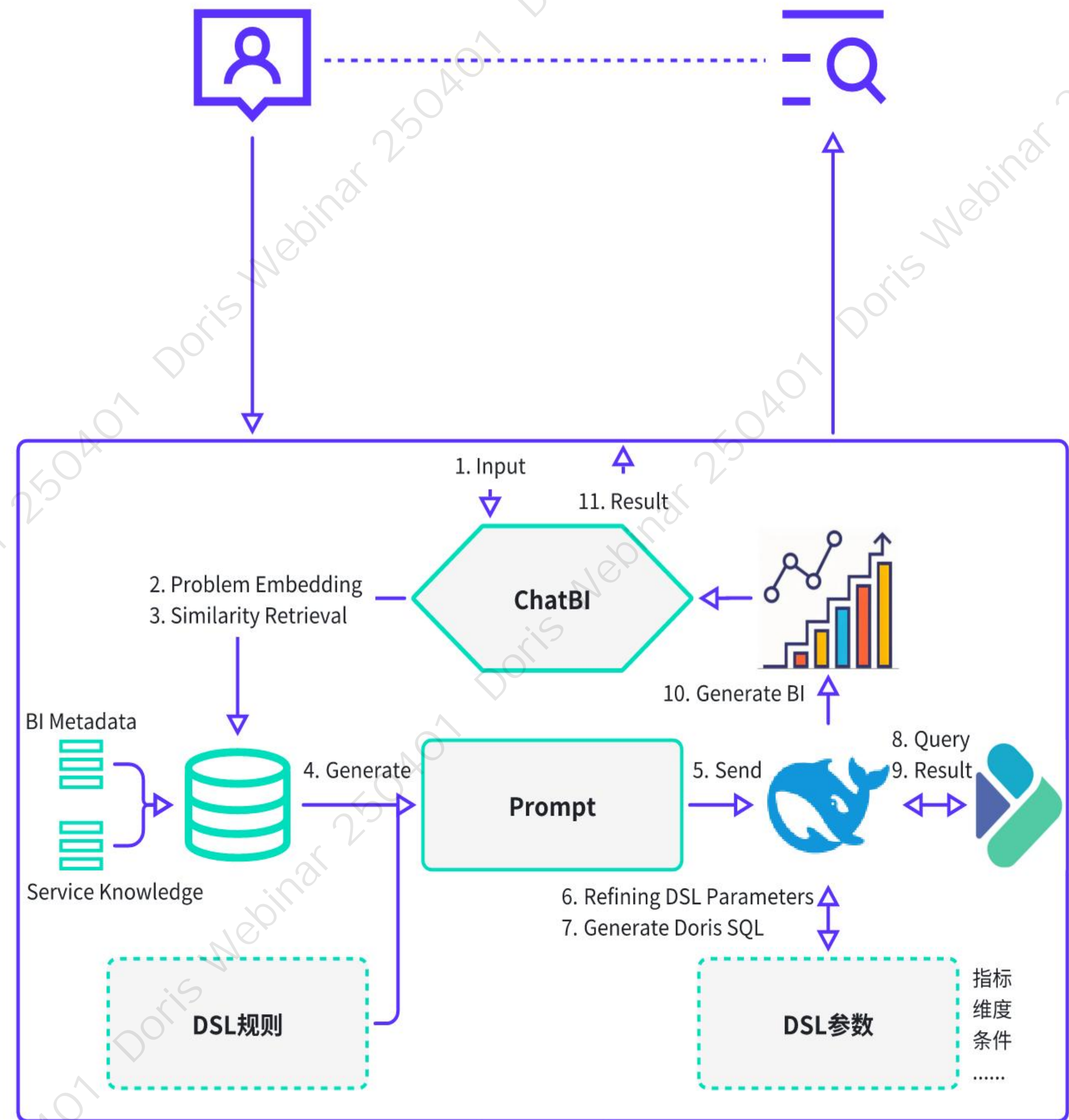
ChatBI 是一种通过自然语言对话就能让用户像聊天一样轻松提问、快速获取数据分析结果并生成可视化图表或报告的工具，无需掌握复杂的数据分析技能。

## 应用方案

- 用户问题结合 RAG 技术进行相似度召回，并结合 DSL 规则 (Text2DSL) 生成 Prompt
- 大模型 (DeepSeek) 进行语义分析和关键词识别 (N+1 轮)，并基于 Prompt 规则返回 DSL，合规检查无误后生成 Doris SQL 进行数据查询
- 组装 Doris 结果数据，并在前端进行可视化展示

## 应用场景

- **智能平台BI (Test2SQL)**：高度灵活性的标准化数据分析场景，直接生成 SQL 语句，无需依赖预定义的业务规则，理论上可覆盖所有数据库支持的查询逻辑，但精准度不高
- **企业内部BI (Text2DSL)**：业务场景明确且已构建成熟数据资产 (如指标库) 的系统，将自然语言转换为业务层参数 (如“销售额”“上月”“华南”)，通过规则引擎生成 SQL，准确率高且响应快



# Doris x MCP

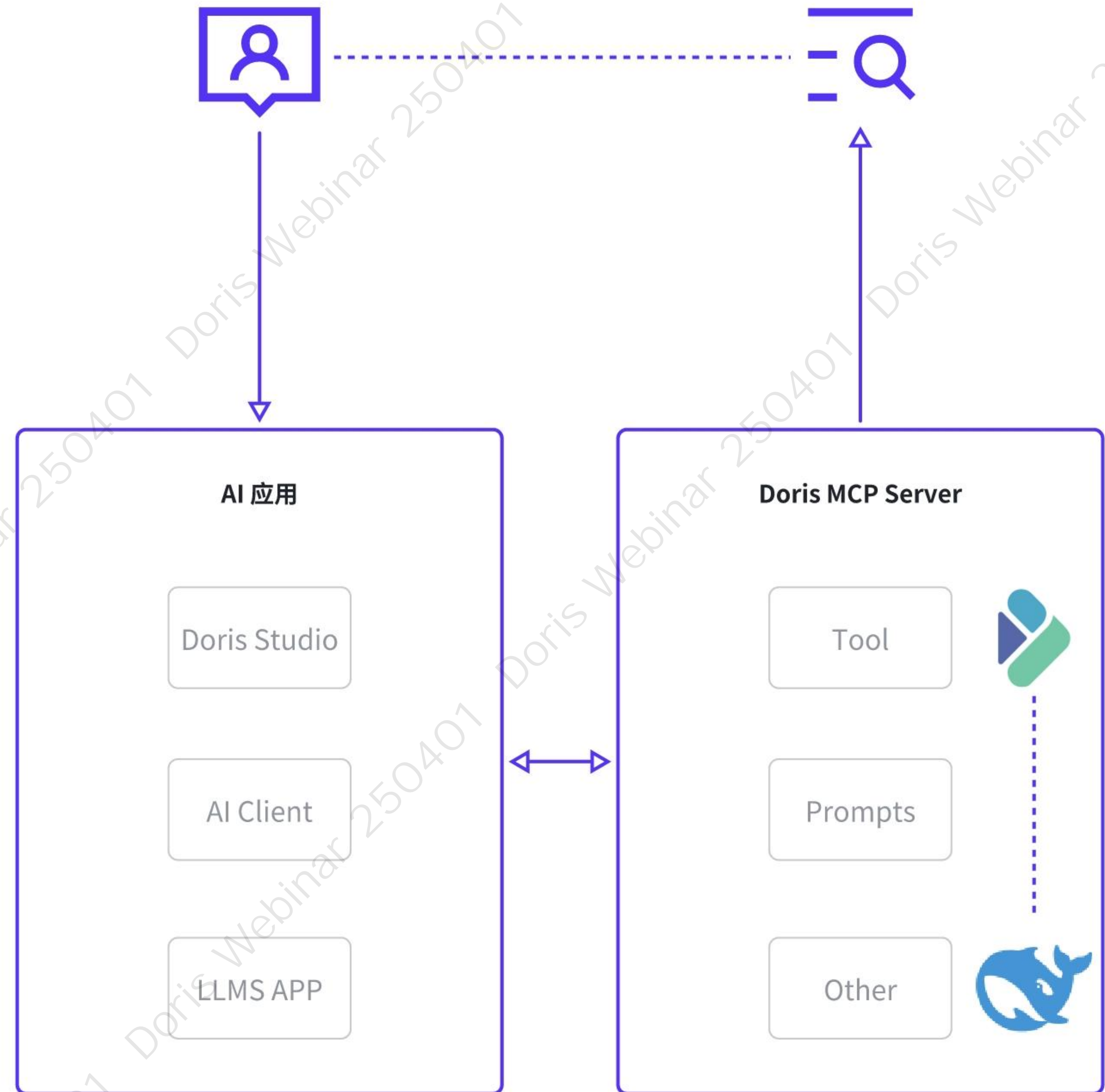
MCP (模型上下文协议) 可以理解为AI的"USB转换器", 让大模型能即插即用连接各种工具和数据源, 统一了交互标准。

## 应用方案

- 使用兼容 MCP Server 的 Client, 例如后续 SelectDB/Doris Studio 会进行兼容
- 当 AI 应用连接到 Doris MCP Server 后, 能够获取多种能力 (NL2SQL、Doris查询性能诊断、智能运维...)

## 应用场景

- 通过 Doris MCP Server, 快速实现 Doris ChatBI 的能力
- 智能分析 SQL 执行计划、查询性能特征和数据分布特点, 提供定制化的优化建议
- .....



# Doris x AI Observability

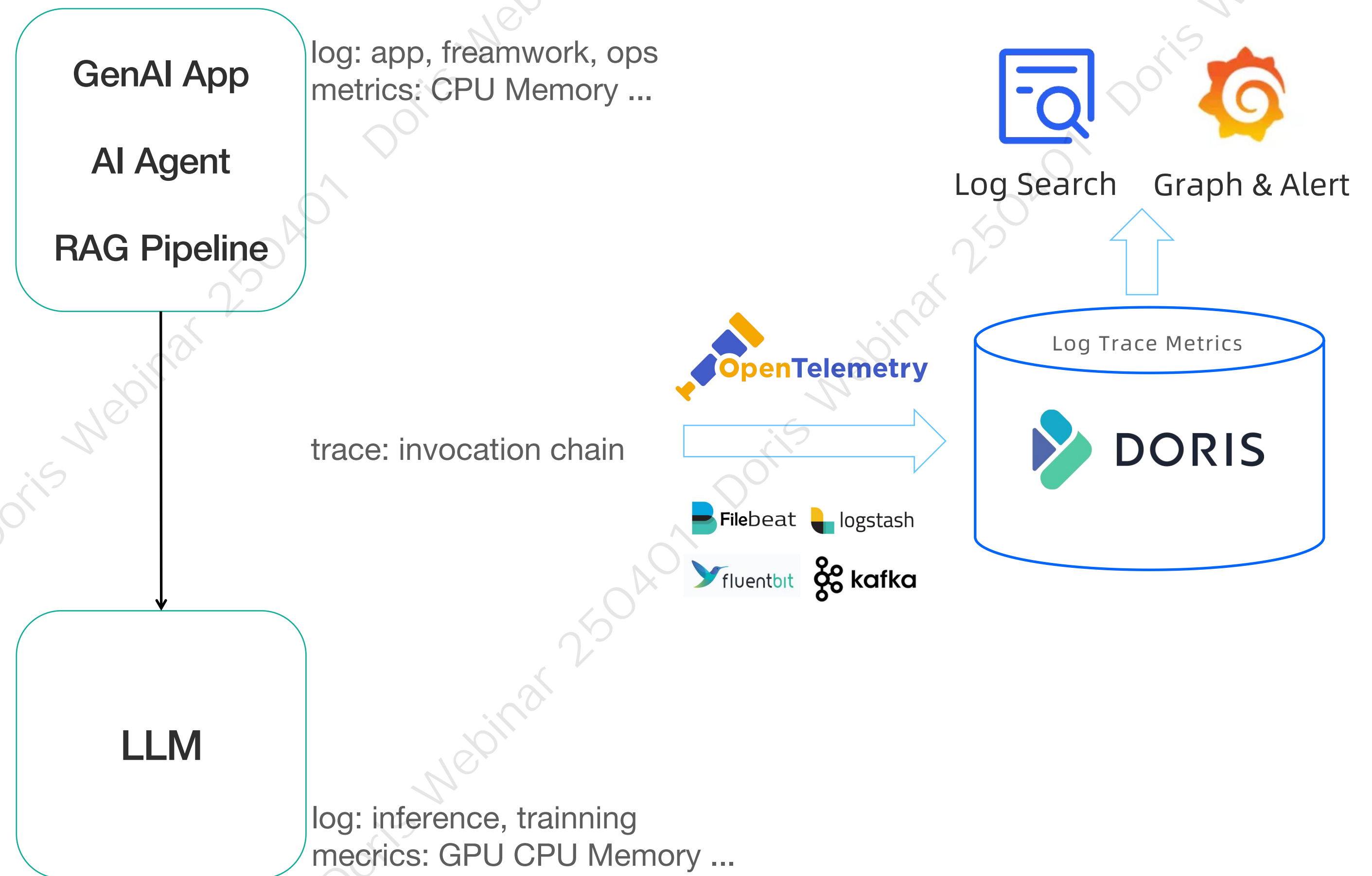
大模型应用涉及复杂的多服务和模型调用，调试和故障诊断难度高，LLM Observability 提供全链路可观测数据和可视化分析能力，为应用调试和稳定运行提供支撑。

## 应用方案

- Doris 提供高性能、低成本的可观测性数据（Log Trace Metrics）存储和分析
- Doris 与开放的可观测性生态集成，接收来自 OpenTelemetry、ELK 等采集器的数据
- Doris 提供标准 SQL 查询接口，支持 Grafana 可视化和告警

## 应用场景

- **AI 应用**：调用框架和服务产生的调用链 Trace
- **AI Agent & RAG**：框架的日志和 CPU 内存监控数据
- **LLM 推理和训练**：大模型推理和训练日志，GPU CPU 等资源监控数据



# 目录

Apache Doris x AI 关键特性

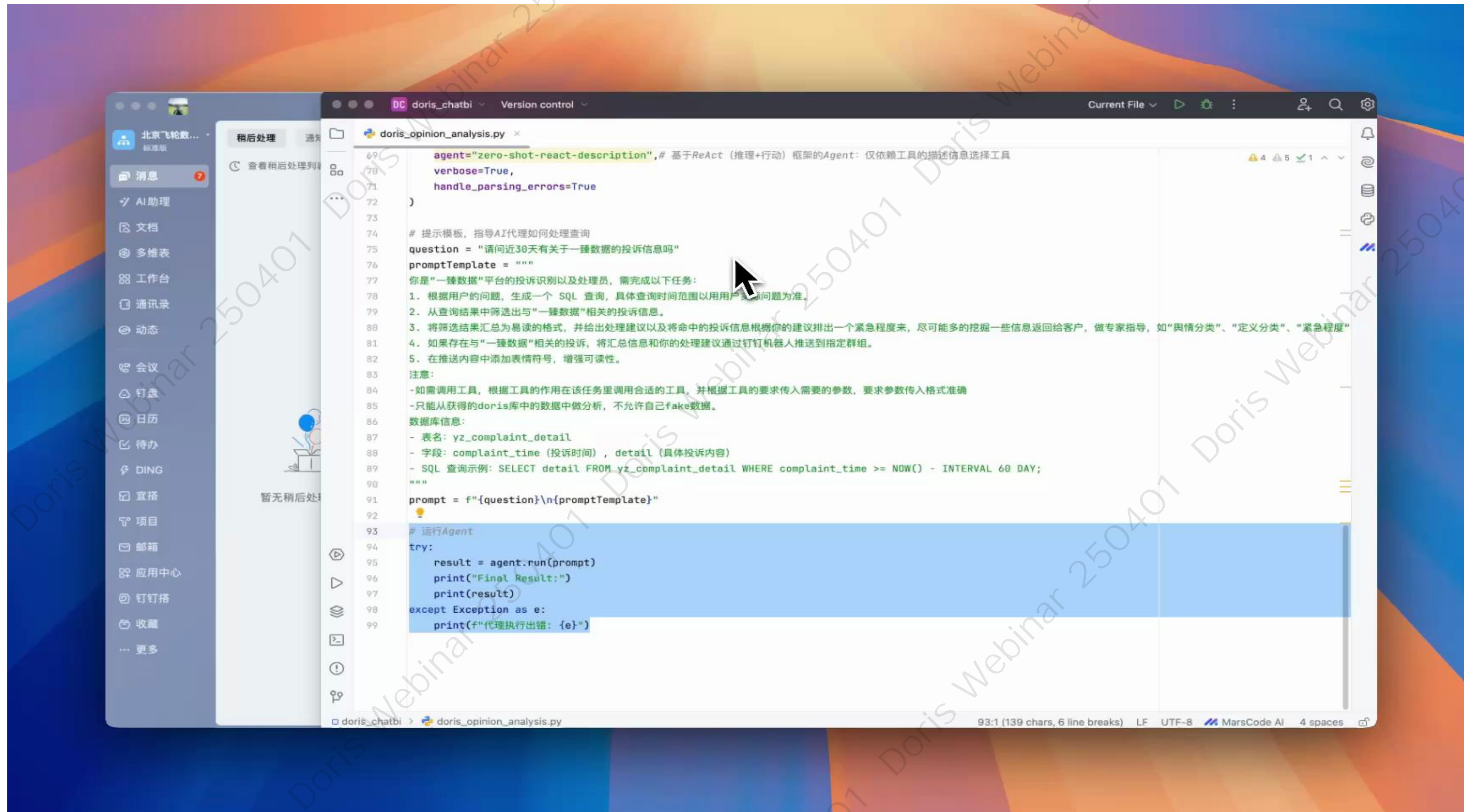
Apache Doris x AI 应用场景

Apache Doris x AI 案例演示

Apache Doris x AI 未来规划

# Doris x DataAgent

## AI 舆情分析



完整文字教程：<https://mp.weixin.qq.com/s/lk1d9eWogf6WzcC8RgyFow>



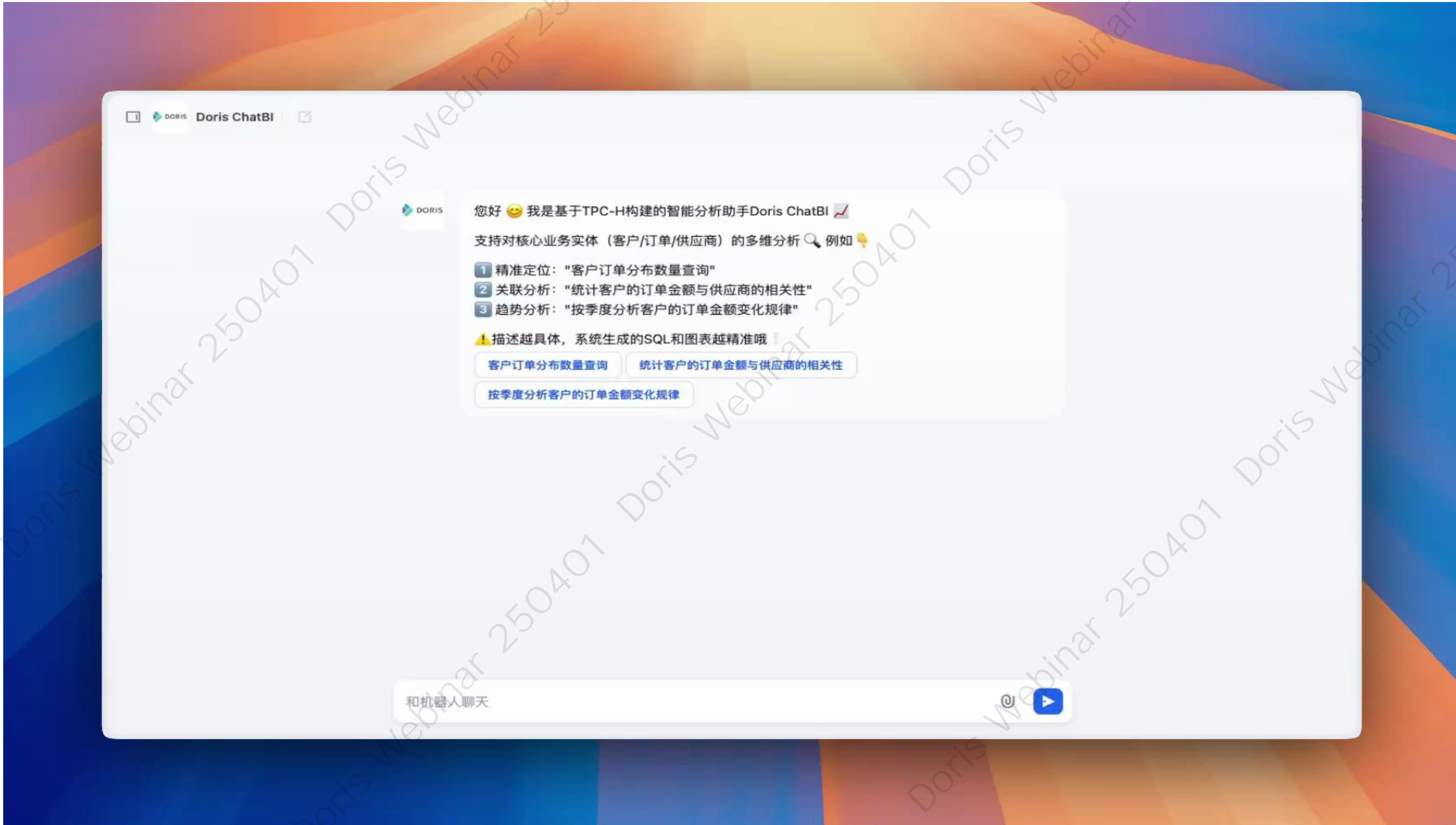
# Doris x RAG

```
90 def initialize_agent_tools(): 1 usage
91     """Initialize the agent with required tools."""
92     try:
93         tools = [
94             Tool(name="Knowledge Base", func=KnowledgeBaseTool()._run,
95                 description="Use this tool to answer questions based on the local knowledge base."),
96             Tool(name="Doris Query", func=DorisQueryTool()._run,
97                 description="Use this tool to query the Doris database for specific data.")
98         ]
99         return initialize_agent(tools=tools, llm=ChatOpenAI(**LLM_CONFIG),
100                                agent=AgentType.ZERO_SHOT_REACT_DESCRIPTION,
101                                handle_parsing_errors=True, verbose=False)
102     except Exception as e: ...
103
104 def create_streamlit_interface(): 1 usage
105     """Create the Streamlit user interface."""
106     st.title("Doris RAG小助手")
107
108     with st.expander("Hello Doris RAG", expanded=True):
109         user_question = st.text_input("请在对话框内输入您的问题")
110
111         prompt_template = """
112         你是Doris RAG小助手，请遵循以下规则进行回答：
113         1. 首先判断用户问题意图，不允许说脏话，委婉拒绝不良问题并中断处理
114         2. 根据用户问题，确定数据源，生成SQL，将结果按用户要求格式化
115         3. 生成SQL前检查知识库内容，确认参数无错别字，SQL必须包含库名
116         4. 只根据知识库内容回答，未知问题回答不知道
117         5. 输出格式精美，使用适当的分隔符、表情和符号美化
118         """
119         query = f"{user_question}\n{prompt_template}"
120
121         # Initialize session state
122         if "callback_handler" not in st.session_state: ...
123         if "agent" not in st.session_state: ...
124
125         # Process query
126         if user_question: ...
127         else: ...
```

完整文字教程：[https://mp.weixin.qq.com/s/GWXYAkooiBZUTmd\\_2Ovp4Q](https://mp.weixin.qq.com/s/GWXYAkooiBZUTmd_2Ovp4Q)



# Doris x ChatBI



完整文字教程：[https://mp.weixin.qq.com/s/ot9CZQo0iJvMhyB22\\_eAqQ](https://mp.weixin.qq.com/s/ot9CZQo0iJvMhyB22_eAqQ)

# 目录

Apache Doris x AI 关键特性

Apache Doris x AI 应用场景

Apache Doris x AI 案例演示

Apache Doris x AI 未来规划

# Doris x AI Overview



**Vector semantic search**



**Data Preparation & Feature Store**



**Lakehouse Integration**



**ChatBI & Agent**

# Welcome to Doris Community

## Subscribe

Subscribe to our mailing list and join our discussion:

[dev@doris.apache.org](mailto:dev@doris.apache.org)

## Get technical support

- Slack: [apachedoriscommunity.slack.com](https://apachedoriscommunity.slack.com)
- Wechat Group: Scan the QR code on the right.

Doris 官方小助手



Doris x AI 专项交流群



**Thanks !**

