



DORIS

Webinar

Release

Apache Doris 4.0 企业级 AI 应用解读 (三)

观看直播

🕒 12月18日 周四 19:30-20:30

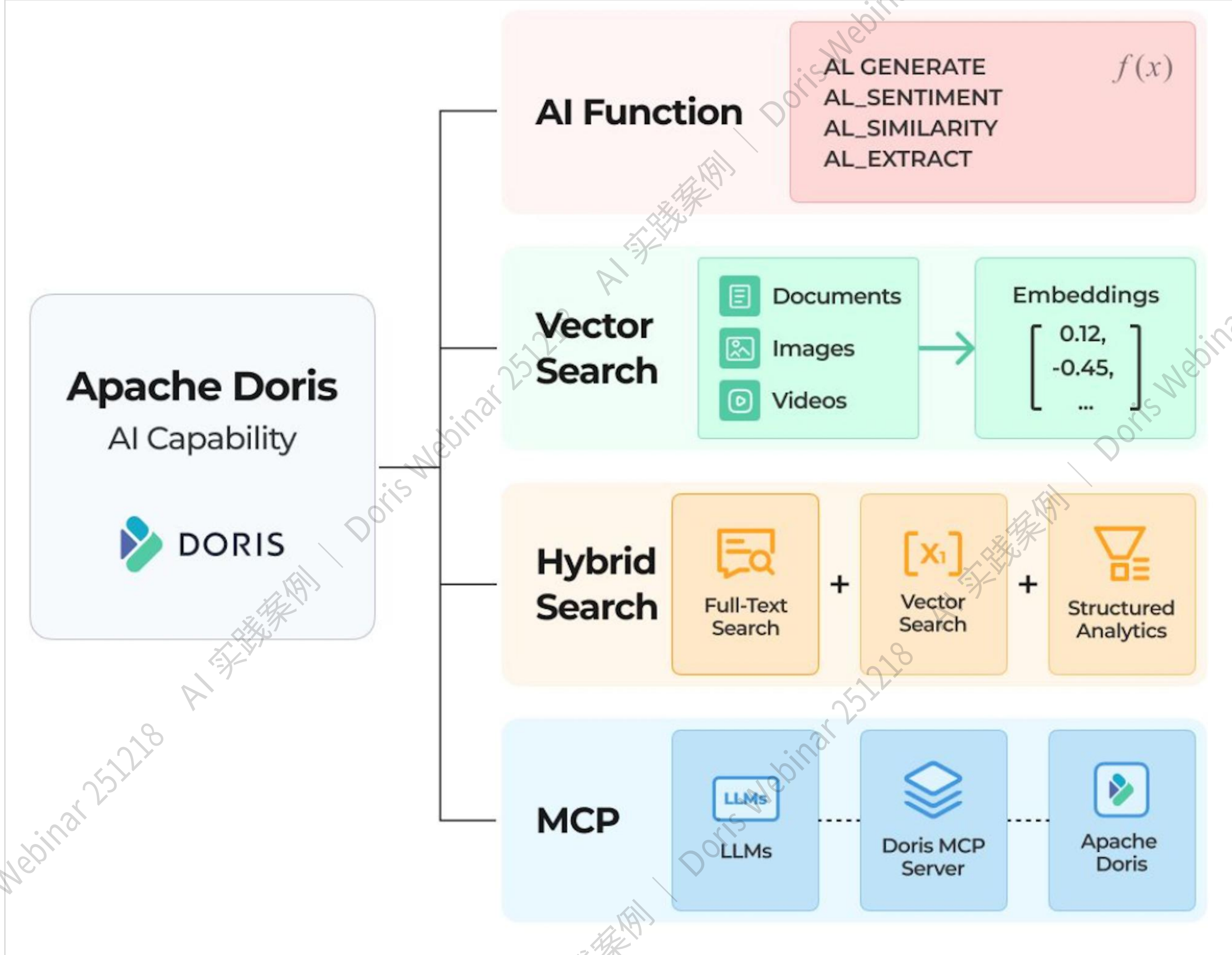
Apache Doris 4.0

从 AI 能力到业务价值

朱伟 SelectDB 内核产品负责人



Apache Doris 4.0 for AI



- Agent Facing Analytics
- Hybrid Search and Analytics Processing
- Lakehouse for AI
- RAG (Retrieval-Augmented Generation)
- Semantic Search
- AI Observability

业务价值?



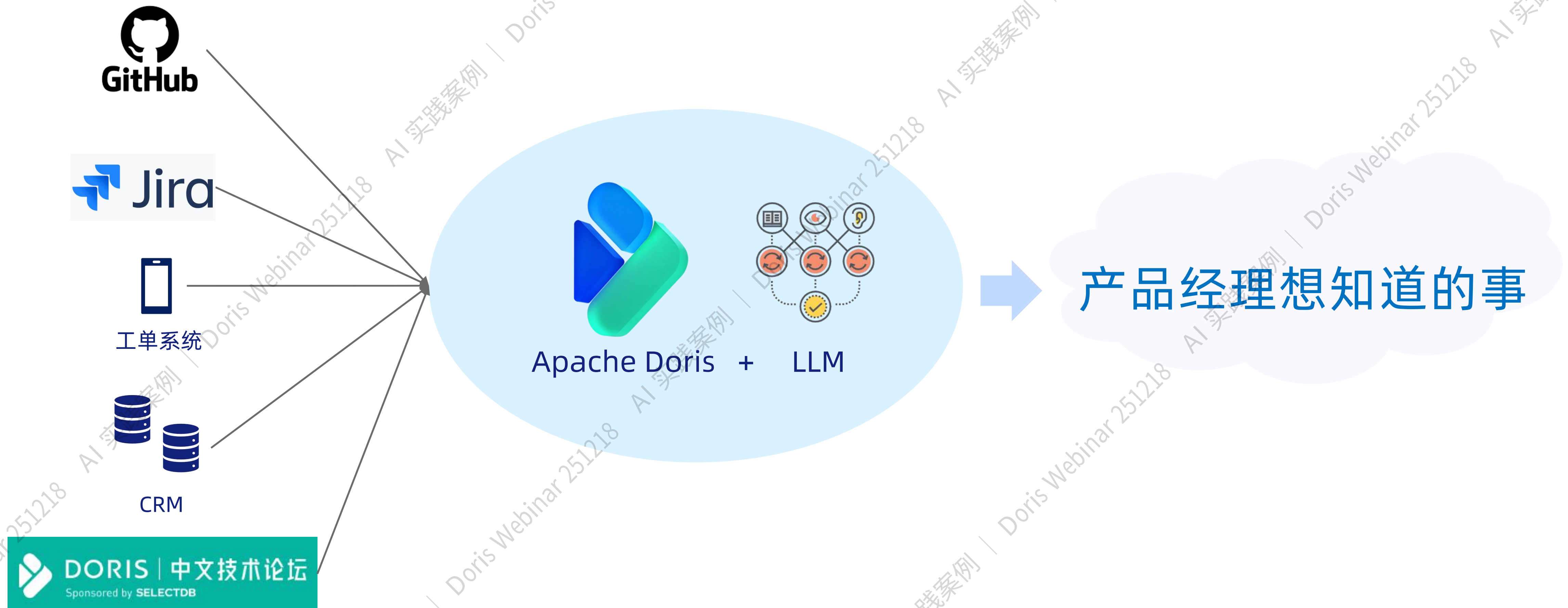
一个实际用例



产品经理想知道的事？

- 物化视图功能中，用户最常使用哪些外部数据源，分别有多少用户反馈？
- 有多少用户提到了数据时效性的痛点或需求，他们都是什么数据源？
- ...

对非结构化用户反馈的搜索与分析



整体步骤

1.初始化

- Apache Doris 建表与原始数据同步
将论坛、GitHub、Jira、工单系统等非结构化数据接入
- AI资源初始化

2.数据处理

- Chunking & Embedding
对非结构化数据进行切片和向量化处理，以向量形式存储。
- 用户反馈中的关键信息提取
利用AI Function提取用户反馈中的关键信息，以 Variant 类型存储

3.混合检索与分析

- 问题一的检索与分析：物化视图功能中，用户最常使用哪些外部数据源，分别有多少用户反馈？
- 问题二的检索与分析：有多少用户提到了数据时效性的痛点或需求，他们都是什么数据源？

Doris中文技术论坛的明细数据表

```
CREATE TABLE forum_questions (  
  question_id BIGINT,      --问题ID  
  user_id BIGINT,          --提问用户ID  
  title STRING,            --标题  
  content STRING,          --用户提问和回复的内容  
  INDEX idx_content (`content`) USING INVERTED PROPERTIES("lower_case" = "true",  
"parser" = "chinese", "support_phrase" = "true")  
)  
UNIQUE KEY(question_id)
```

question_id	user_id	title	content
100100000000000035	680	【已解决】doris 什么时候支...	QUESTION: doris 什么时候...
100100000000000040	1699	【已解决】flink-doris-conn...	QUESTION: 我是用 flink-dor...
100100000000000053	1720	【已解决】关于doris tablet...	QUESTION: 新接触doris, 看...
100100000000000059	1729	【已解决】SHOW ALTER TA...	QUESTION: 1.SHOW ALTER... ![image.png](https://ask.sel...
100100000000000063	315	【已解决】通过nginx 代理跳...	QUESTION: ![](https://ask.s...
100100000000000064	1733	【已解决】doris 高版本两阶...	QUESTION: 现实场景中有一...
100100000000000065	1732	【已解决】stream load erro...	QUESTION: 1, 请问有没有... 2, 使用 Flink 一次checkpoi...
100100000000000068	1724	sink.batch.size的大小问题	QUESTION: > spakr 写入 si...
100100000000000074	1734	【已解决】x2Doris 创建任务...	QUESTION: x2Doris 创建任... ![85803390-ee35-4807-9...
100100000000000097	1413	【已解决】Doris支持提交的S...	QUESTION: SQL最大长度是...

论坛问答向量表

```
CREATE TABLE forum_question_embeddings (  
  user_id      BIGINT NOT NULL,      -- 原问题所属用户  
  question_id  BIGINT NOT NULL,      -- 原问题ID  
  chunk_id     BIGINT NOT NULL AUTO_INCREMENT, -- 文本切片 ID (每片对应一条向量)  
  content_chunk STRING NOT NULL,      -- 文本切片内容  
  embedding    array<float> NOT NULL, -- 1024维向量  
  INDEX ann_index (embedding) USING ANN PROPERTIES(  
    "index_type"="hnsw",  
    "metric_type"="inner_product",  
    "dim"="1024",  
    "quantizer"="flat"  
  )  
);  
DUPLICATE KEY(user_id, question_id)  
DISTRIBUTED BY HASH(user_id)  
PROPERTIES (  
  "replication_num" = 3  
);
```

user_id	question_id	chunk_id	content_chunk	embedding
71	10010000000002481	3805	QUESTION: 海豚调度定时调... ``java	[0.02958452, 0.01621151]
71	10010000000002481	4248	xecuteTask(DefaultWorkerD... at org.apache.dolphinsch...	[-0.008080235, 0.02962]
71	10010000000002481	4560	131) at com.google.common.u...	[-0.02299667, 0.0515031]
71	10010000000002481	4755	ava:748) Caused by: java.sql.SQLExc...	[0.001762386, -0.001981]
71	10010000000002481	4896	ysql.cj.jdbc.ClientPrepared... at com.mysql.cj.jdbc.Clie...	[-0.03416767, 0.0124982]
71	10010000000002481	4998	pache.dolphinscheduler.plu... at org.apache.dolphinsch...	[0.05752784, 0.01667112]
71	10010000000002481	5081	: `meta_delay_toleration_s...	[0.02766938, 0.0428481]
122	1001000000000396	305	QUESTION: 版本为2.0.4, 按... 但是将域名指向ipv6后, 启动...	[0.02647048, 0.03317994]
122	1001000000000396	366	24-03-08 15:54:23,240 W... > 2024-03-08 15:54:23,30...	[-0.03165316, 0.0457278]
122	1001000000000396	408	2024-03-08 15:54:23,407 I...	[0.004303984, 0.042514]

用户反馈解析表

```
CREATE TABLE parsed_results (  
  id BIGINT NOT NULL AUTO_INCREMENT,           -- 主键ID  
  feedback_id BIGINT NOT NULL,                  -- 用户反馈ID (外键关联 forum_questions 或其他表)  
  parsed_data VARIANT NOT NULL                  -- 解析结果 (任意 JSON 数据)  
)  
UNIQUE KEY(id)  
DISTRIBUTED BY HASH(id) BUCKETS 2  
PROPERTIES (  
  "replication_num" = "1"  
);
```

id	feedback_id	parsed_data
10	100100000000000454	{"problems_or_gaps":"总执行时间2...
11	100100000000000191	{"other_insights":"建物化视图调整...
42	100100000000000598	{"other_insights":"分区物化视图增...
43	100100000000000657	{"other_insights":"分区字段在物化...


```
CREATE RESOURCE "zhipu-embedding-2"
PROPERTIES (
  'type' = 'ai',
  'ai.provider_type' = 'Zhipu',
  'ai.endpoint' = 'https://open.bigmodel.cn/api/paas/v4/embeddings',
  'ai.model_name' = 'embedding-2',
  'ai.api_key' = 'xxx',
  'ai.temperature' = '0.7',
  'ai.max_token' = '1024',
  'ai.max_retries' = '3',
  'ai.retry_delay_second' = '1',
  'ai.dimensions' = '1024'
);

CREATE RESOURCE "glm-4-flash-250414"
PROPERTIES (
  'type' = 'ai',
  'ai.provider_type' = 'Zhipu',
  'ai.endpoint' = 'https://open.bigmodel.cn/api/paas/v4/chat/completions',
  'ai.model_name' = 'glm-4.5-flash',
  'ai.api_key' = 'xxx',
  'ai.temperature' = '0.7',
  'ai.max_token' = '-1',
  'ai.max_retries' = '3',
  'ai.retry_delay_second' = '1'
);
```

AI资源初始化


```

WITH raw AS (
    SELECT
        question_id,
        user_id,
        content,
        LENGTH(content) AS content_len
    FROM forum_questions_auto_inc
),
t AS (
    SELECT 0 AS d UNION ALL SELECT 1 UNION ALL SELECT 2 UNION ALL SELECT 3 UNION ALL
    SELECT 4
    UNION ALL SELECT 5 UNION ALL SELECT 6 UNION ALL SELECT 7 UNION ALL SELECT 8 UNION
    ALL SELECT 9
),
num AS (
    SELECT
        (a.d * 100 + b.d * 10 + c.d + 1) AS n
    FROM t a
    CROSS JOIN t b
    CROSS JOIN t c
),
chunks AS (
    SELECT
        r.user_id,
        r.question_id,
        SUBSTRING(r.content, (n - 1) * 400 + 1, 400) AS content_chunk
    FROM raw r
    JOIN num ON (n - 1) * 400 + 1 <= r.content_len
)
INSERT INTO forum_question_embeddings (user_id, question_id, content_chunk, embedding)
SELECT
    user_id,
    question_id,
    content_chunk,
    EMBED("zhipu-embedding-2", content_chunk) AS embedding
FROM chunks;

```

Chunking & Embedding

SET @extract_prompt = "你是一名专业的产品反馈分析助手。现提供一段来自用户的反馈文本（feedback_text），请从中提取与“物化视图（Materialized View, MV）”相关的关键信息，并以严格的JSON 格式输出。

如果反馈文本与 MV 无关，或无法识别任何相关信息，则直接返回字符串 NULL。

请从 feedback_text 中提取以下信息并生成 JSON：

```
{
  \"used_external_source\": \"用户是否提到使用了哪些外部数据源，例如 MySQL、Oracle、Hive、Iceberg 等。如果未提及则为 null。\",
  \"transparent_rewrite\": \"是否使用或提到透明改写（true / false / null）。\",
  \"usage_scenario\": \"用户描述的使用场景。若无则为 null。\",
  \"problems_or_gaps\": \"用户认为 MV 不足或存在的问题。若无则为 null。\",
  \"other_insights\": \"任何与 MV 相关、对产品改进有价值的额外信息。若无则为 null。\"
}
```

要求：

1. 输出内容必须是纯 JSON，对象必须以 { 开头、以 } 结尾。
2. 禁止输出任何 Markdown 代码块标记，如 json、`json`、``` 等。
3. 禁止输出任何额外文字、描述、解释、前缀或后缀。
4. 不得在 JSON 外输出换行、空格或注释。";

```
INSERT INTO parsed_results (feedback_id, parsed_data)
SELECT
  question_id,
  AI_GENERATE(
    'glm-4-flash-250414',
    concat(@extract_prompt ,'\nfeedback_text: ',content, '```)
  ) AS extracted_info
FROM forum_questions where search('content:物化视图');
```

用户反馈关键信息提取

混合检索与分析

- 物化视图功能中，用户最常使用哪些外部数据源，分别有多少用户反馈？

```
SELECT
  lower(CAST(c.used_external_source AS string)) AS used_external_source,
  COUNT(*)
FROM (
  SELECT
    a.parsed_data.usage_scenario,
    a.parsed_data.transparent_rewrite,
    a.parsed_data.used_external_source AS used_external_source,
    a.parsed_data.problems_or_gaps,
    a.parsed_data.other_insights,
    b.content
  FROM parsed_results a
  JOIN forum_questions b ON a.feedback_id = b.question_id
  WHERE a.parsed_data.used_external_source != ""
  AND a.parsed_data.used_external_source != "null"
) c
GROUP BY lower(CAST(c.used_external_source AS string));
```

分析结果

37ms 🔍

used_external_source	feedback_cnt
mysql	10
hive	3
oracle	2
sqlserver	2
kafka	2
app_span_info	1
dim dim_calendar_month	1


```
SET @search_text="数据时效性差、 ETL 延迟、物化视图刷新不及时、数据不实时、实时查询不满足业务、报表延迟、同步滞后、时延高、更新慢、不能满足实时监控需求、数据落地慢、数据延时影响决策";
```

```
CREATE TEMPORARY TABLE retrieved_question_ids (  
    question_id BIGINT  
)
```

```
UNIQUE KEY(question_id)  
DISTRIBUTED BY HASH(question_id) BUCKETS 2  
PROPERTIES (  
    "replication_num" = "1"  
);
```

```
insert into retrieved_question_ids(question_id)
```

```
SELECT question_id  
FROM forum_question_embeddings  
where inner_product_approximate(  
    embedding,  
    EMBED("zhipu-embedding-2", @search_text)  
) > 0.5
```

```
union ALL
```

```
select a.question_id from
```

```
(  
    SELECT question_id,score() as relevance  
    FROM forum_questions_auto_inc  
    where content MATCH_ANY '时效性 实时 延迟 延时 更新不及时 刷新慢 滞后'  
    ORDER BY relevance DESC  
    LIMIT 100  
)a;
```

混合检索与分析

- 有多少用户提到了数据时效性的痛点或需求，他们都是什么数据源？

混合检索与分析

- 有多少用户提到了数据时效性的痛点或需求，他们都是什么数据源？

```
SELECT
  lower(CAST(c.used_external_source AS STRING)) AS used_external_source,
  COUNT(DISTINCT c.question_id) AS feedback_cnt
FROM (
  SELECT
    a.feedback_id AS question_id,
    a.parsed_data.used_external_source AS used_external_source
  FROM parsed_results a
  JOIN retrieved_question_ids r
    ON a.feedback_id = r.question_id
  WHERE a.parsed_data.used_external_source IS NOT NULL
    AND a.parsed_data.used_external_source != ""
    AND lower(a.parsed_data.used_external_source) != "null"
) c
GROUP BY lower(CAST(c.used_external_source AS STRING))
ORDER BY feedback_cnt DESC;
```

分析结果

ert into retr...		SELECT lower...
35ms		Q
used_external_source	feedback_cnt	
mysql	2	
hive	1	
iceberg, paimon	1	
sqlserver	1	
s3	1	

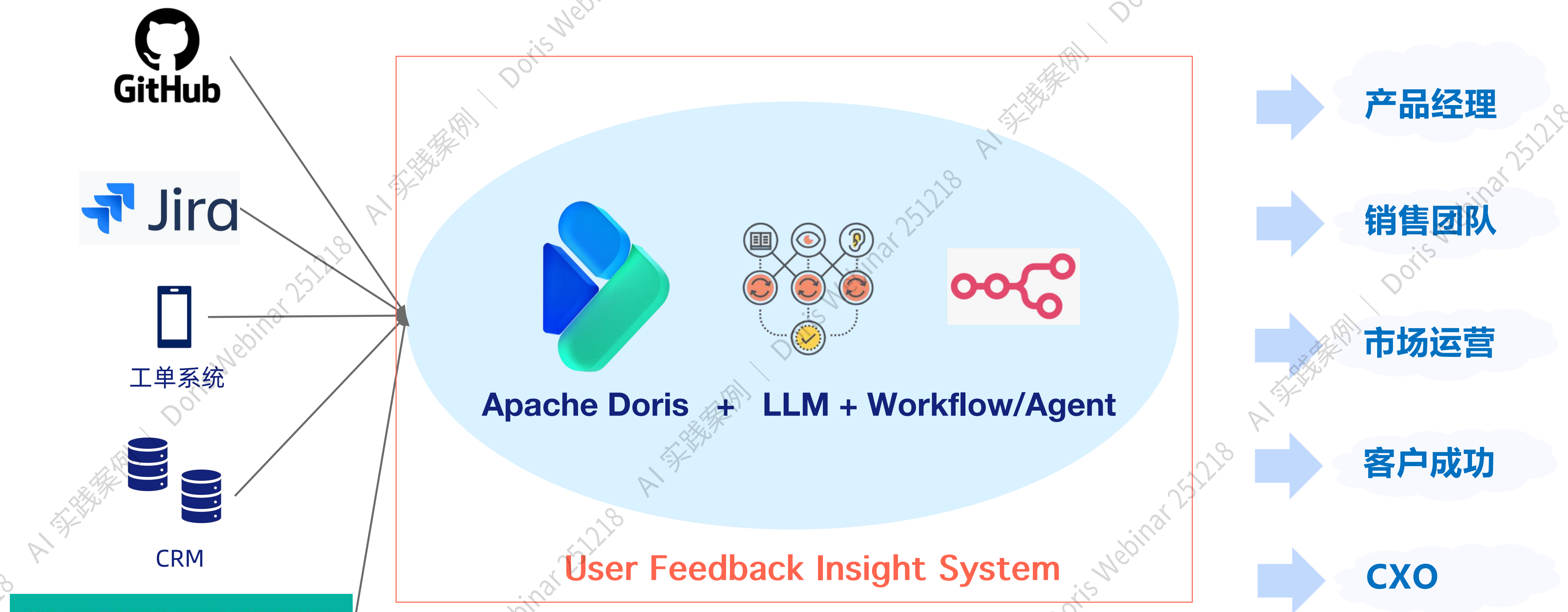


DORIS



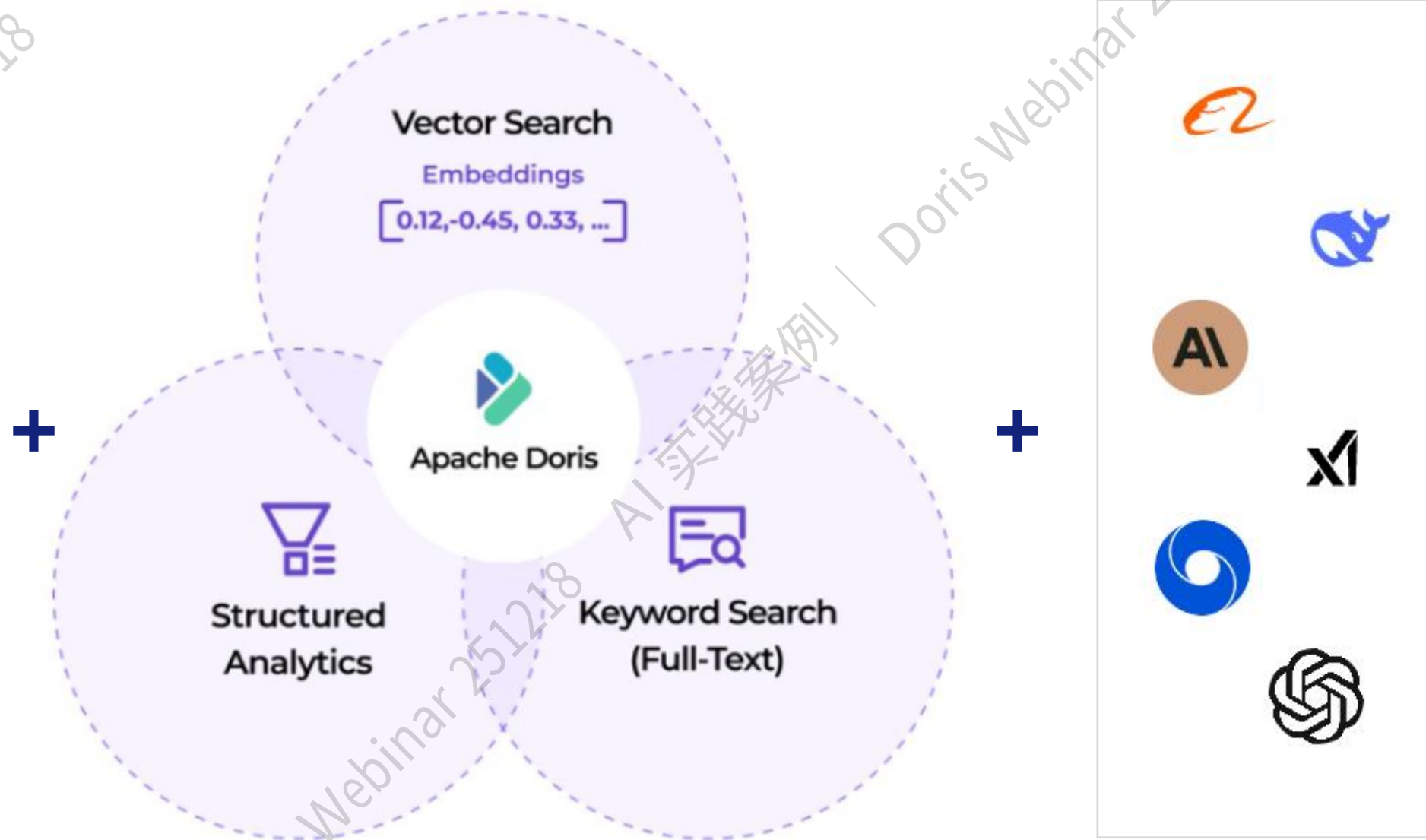
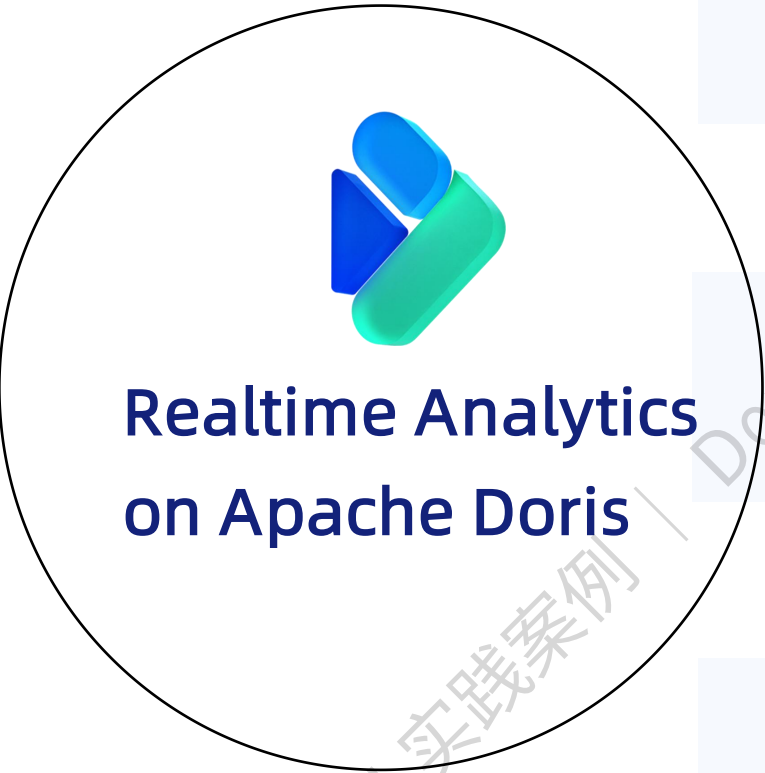
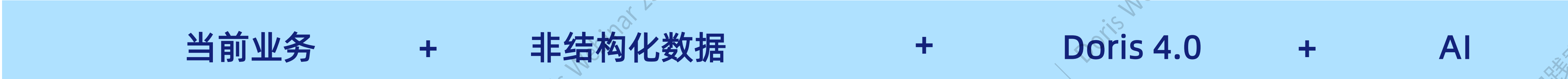
SELECTDB

构建一个用户反馈洞察系统



1. 用户可创建感兴趣的Topic, 例如物化视图的使用情况、客户的流失风险
2. 由所有非结构化用户反馈, 自动生成并定期更新Topic报告
3. 在Topic中, 用户可选择添加数据模型, 展开进一步对话

Doris + AI 能力到业务价值



Hybrid Search and Analytics Processing

Thanks !





DORIS

Webinar

Release

回放与演讲资料获取

请关注 SelectDB 公众号发送

20251218

联系我们

 www.selectdb.com

 400-092-6099



微信公众号



免费试用



在线咨询



加入社区