

DORIS
SUMMIT



大模型时代指标平台建设实践

刘豹 数势科技 大数据技术负责人

目录

01 LLM Agent vs 数据普惠

02 数据普惠实现路径

03 数势科技基于 Apache Doris 的数据语义平台建设实践

04 案例分享

数势科技：行业领先的数据智能产品提供商

深耕大金融、高科技制造和泛零售等领域，为企业提供基于大模型增强的智能分析助手 SwiftAgent、指标平台 SwiftMetrics、智能标签平台 SwiftXDP 等系列产品，提升企业的数字化决策能力，推动企业数字化升级。

部分代表客户	<div>       </div> <div>        </div> <div>        </div> <div>       </div>
投资方	<div>     </div>
生态合作	<div>       </div> <div>      </div>

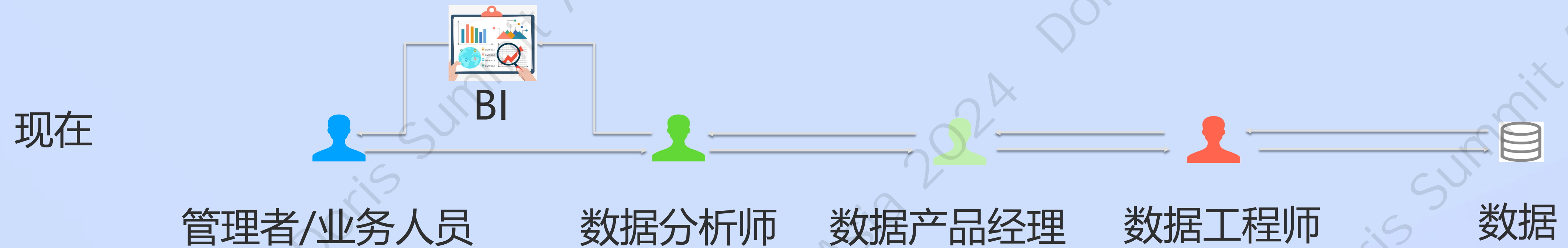
01

LLM Agent & 数据普惠

LLM Agent 在 ToB 行业常见落地场景

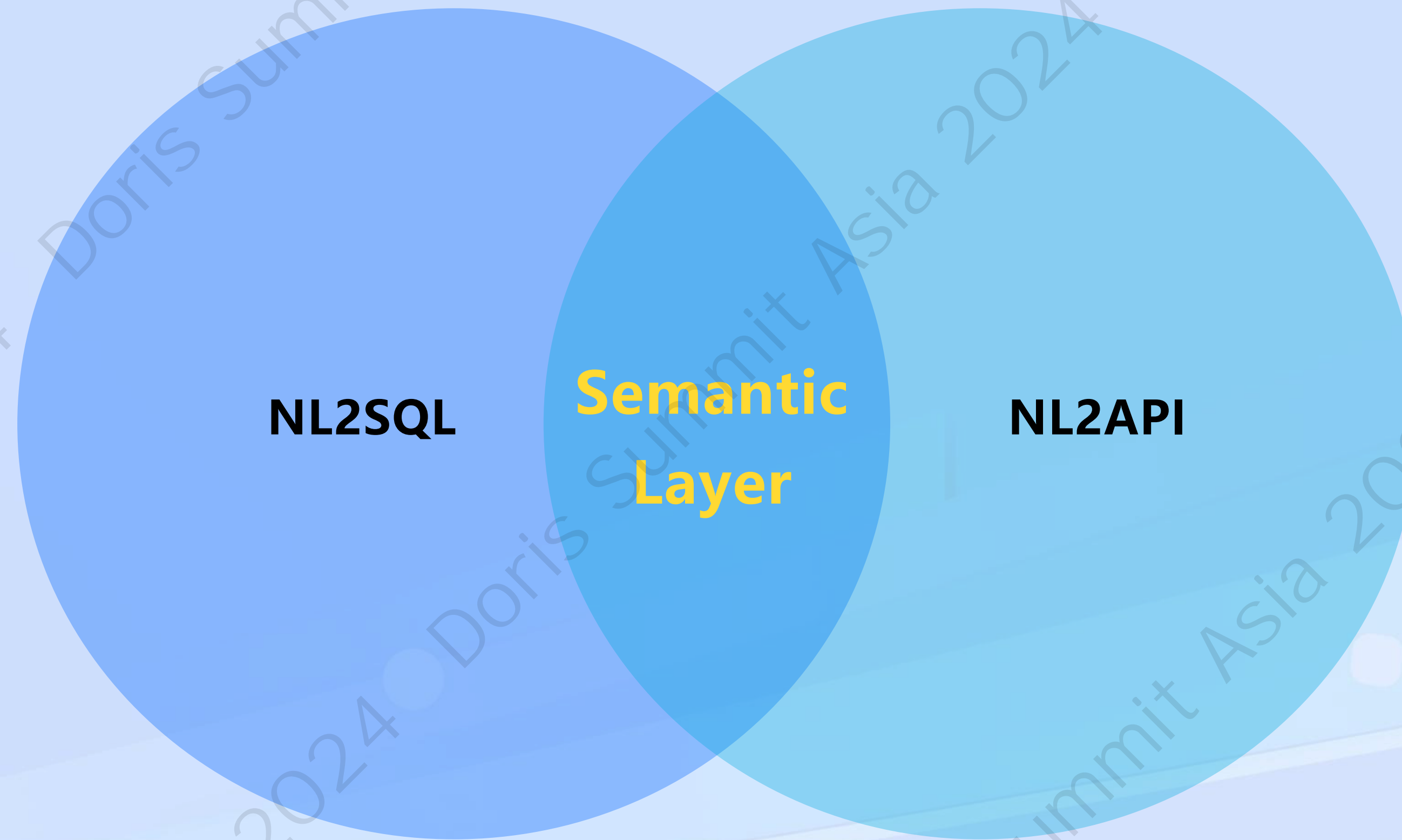


智能分析 LLM Agent 与数据普惠



管理者/一线业务员直达数据，用数门槛降低，提升企业经营决策与日常业务流中数据参与度
(普惠化)

智能分析 LLM Agent 常见方案对比

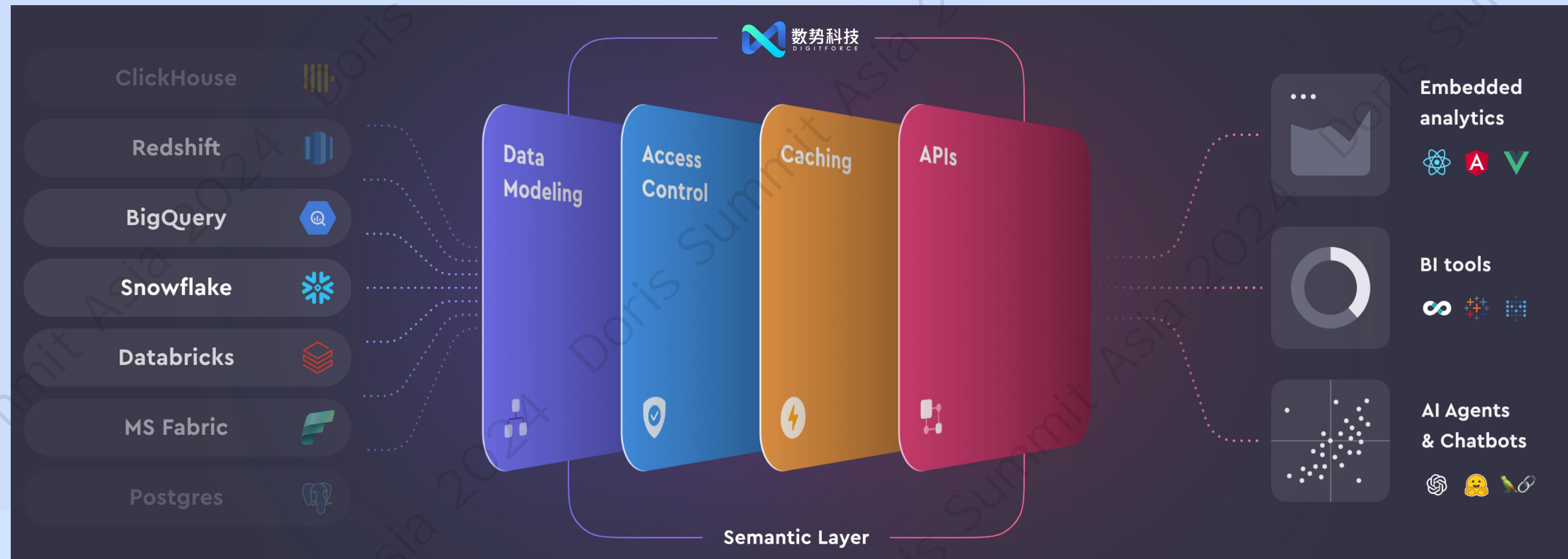


02

数据普惠实现路径

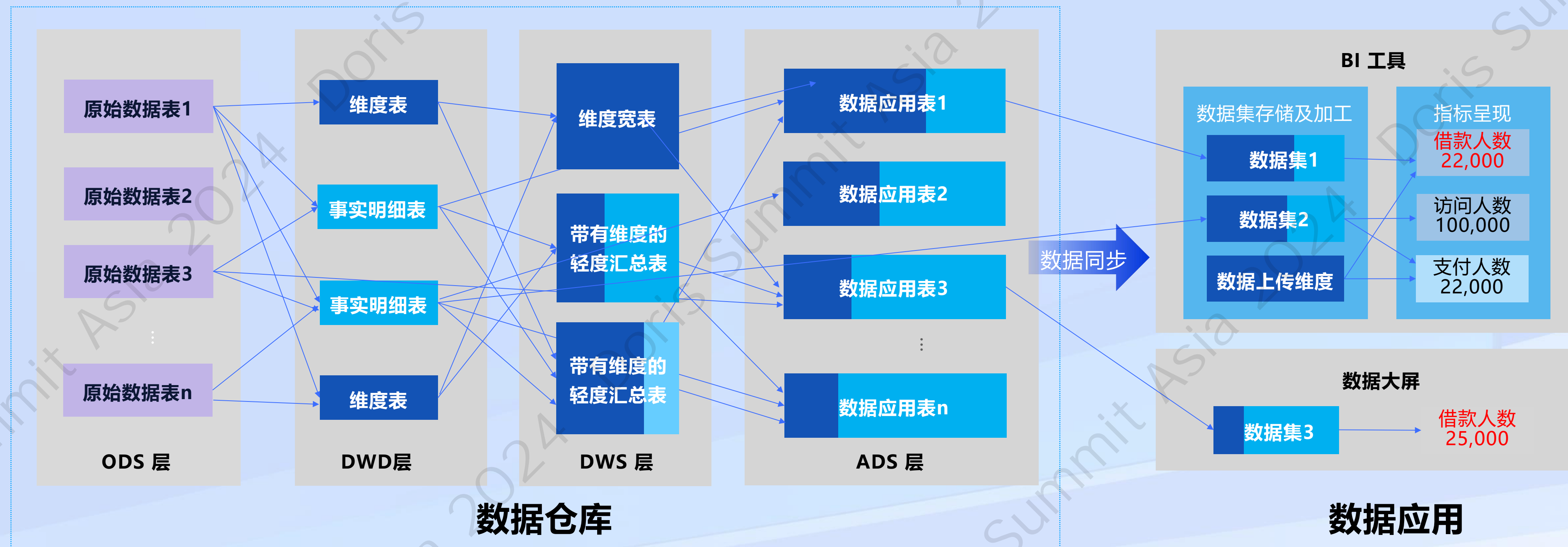
统一语义层：提高准确率、降低成本

统一语义层是现代数据栈中的一个独立且可互操作的部分，它位于数据源与数据使用者之间。统一语义层使得所有的数据端点，无论是 BI（商业智能）工具、嵌入式分析，还是 AI Agent 和聊天机器人，**都能使用相同的语义和底层数据**，从而得到一致且可信赖的洞察。

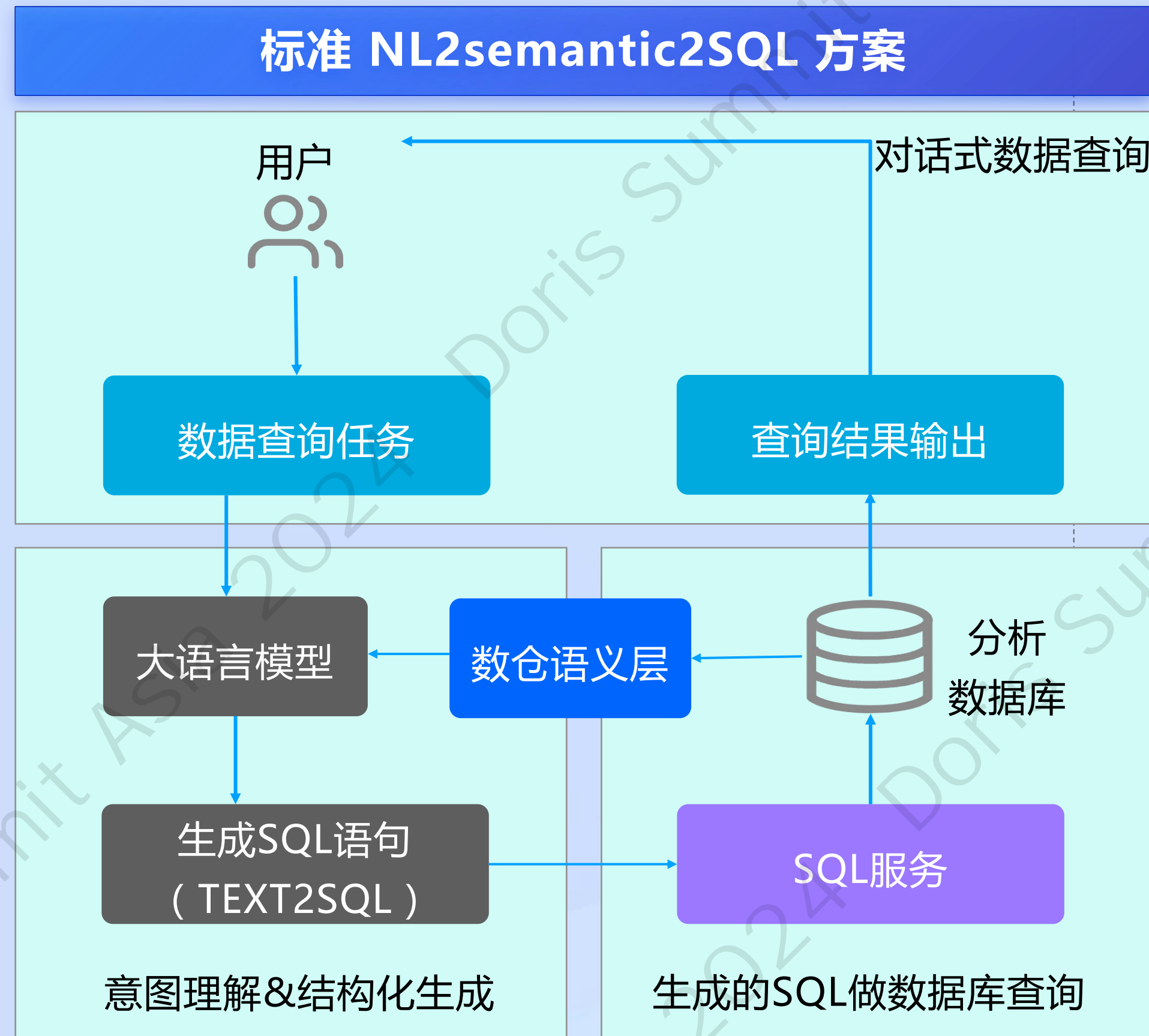


仓内语义：繁琐复杂、业务指向性差

- ODS -> DWD -> DWS -> ADS，语义建在哪一层？
- 数据产品经理、数据开发、终端数据使用者，谁来建数据语义？



仓内语义 LLM Agent 方案：NL2SQL



1、准确率低

- 数仓分层语义差异小，LLM 意图识别准确率低；
- LLM world knowledge 包含的企业级知识非常少；

2、性能不稳定

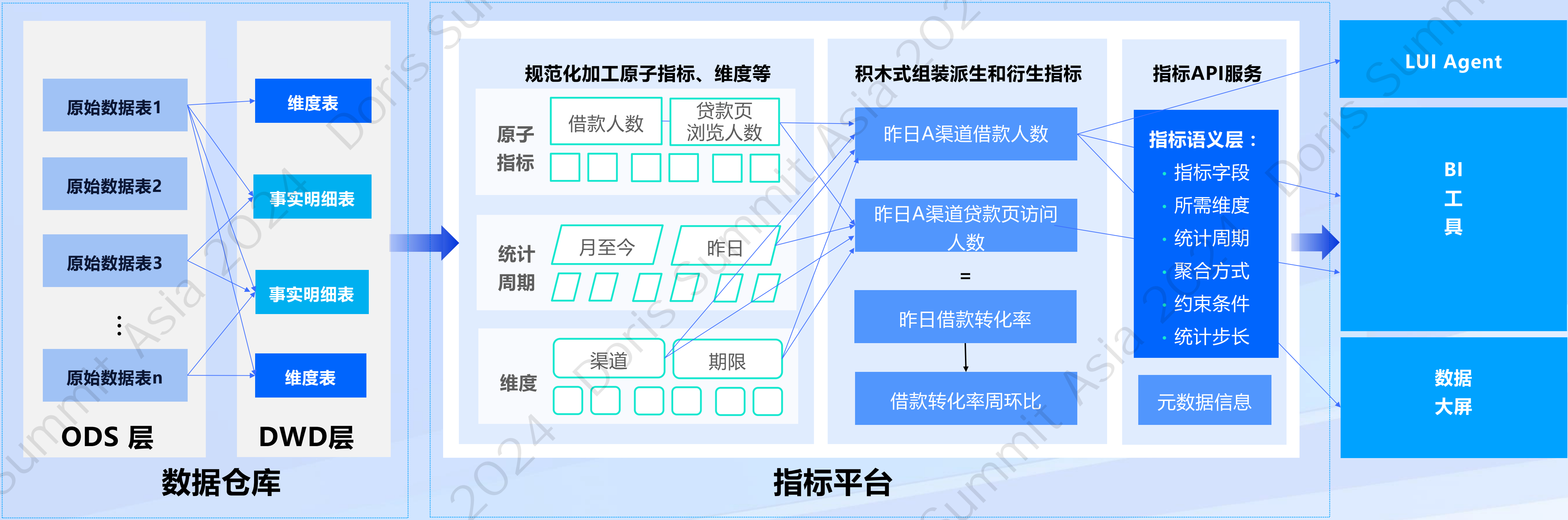
- 大模型写出的 SQL 未经优化，实际上仅能做参考；
- 大查询有将数据库拖垮的风险，影响全局系统稳定；

3、数据安全风险

- LLM 无法确定提问人对企业数据的权限范围，直接 toSQL 容易出现数据安全风险

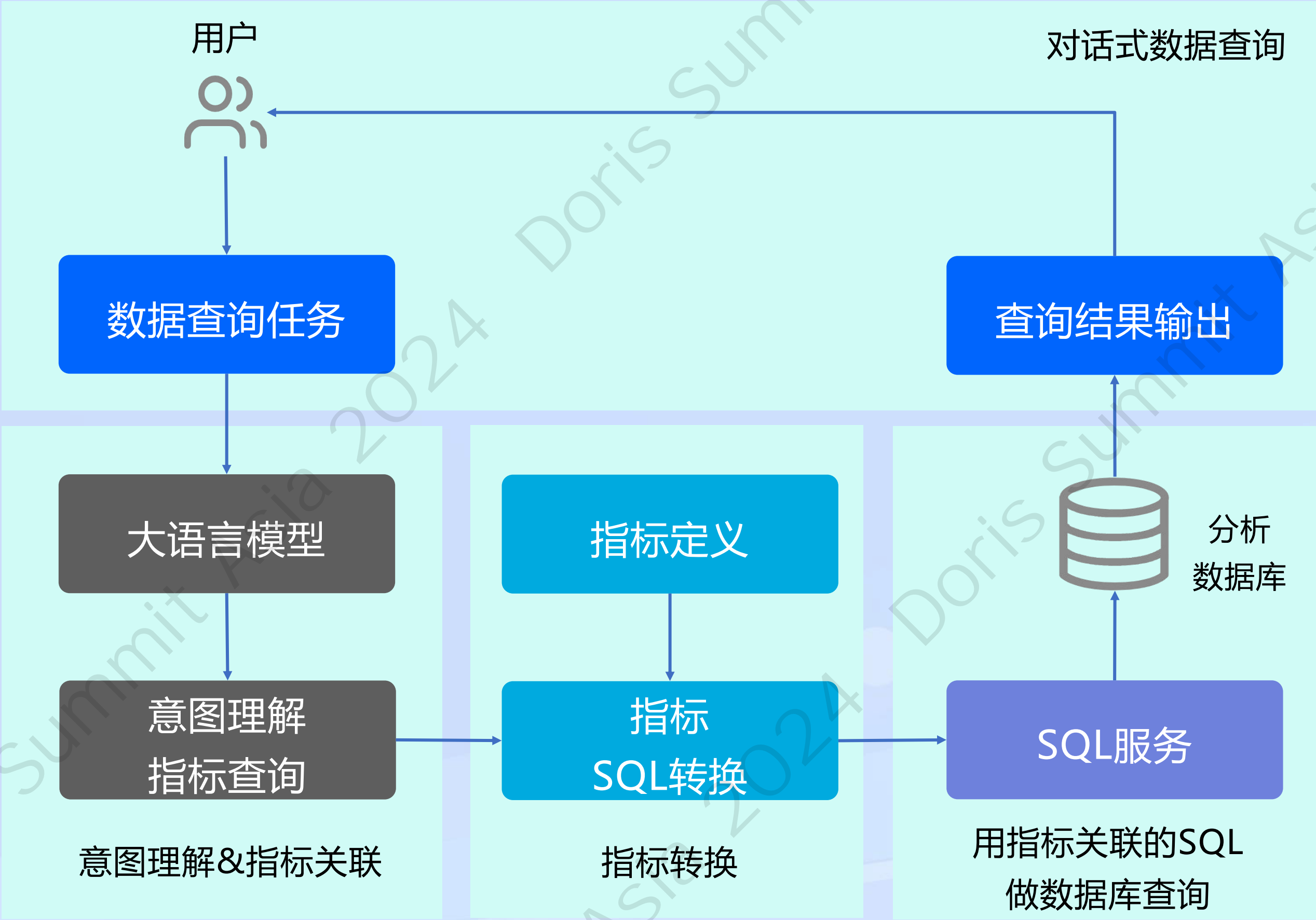
仓外语义：灵活便捷、贴近业务

- 数据建模右移，更贴近数据消费 endpoint，更便于 LLM Agent 规划推理；
- 基于虚拟层做数据编织，口径管理更灵活、便捷。



仓外语义 LLM Agent 方案：NL2API

结合大模型 + 指标分析平台，NL2semantic，提供更优方案



① 数据可信，准确率高

- 预设数据指标的定义与管理，避免业务理解对不齐
- 借助思维链分析与歧义反问，提升泛化性，避免直接从文本到SQL

② 学习成本低

- 无需对大模型做预训练，仅需依据指标语义和知识库做用户意图理解，增强prompt语义
- 全流程白盒，企业客户用业务语言描述查询过程，方便快速排查

③ 性能提升且稳定

- 基于自研的数据查询加速引擎，智能优化查询语句
- P95 可实现 从检索到回答的 **秒级出数**

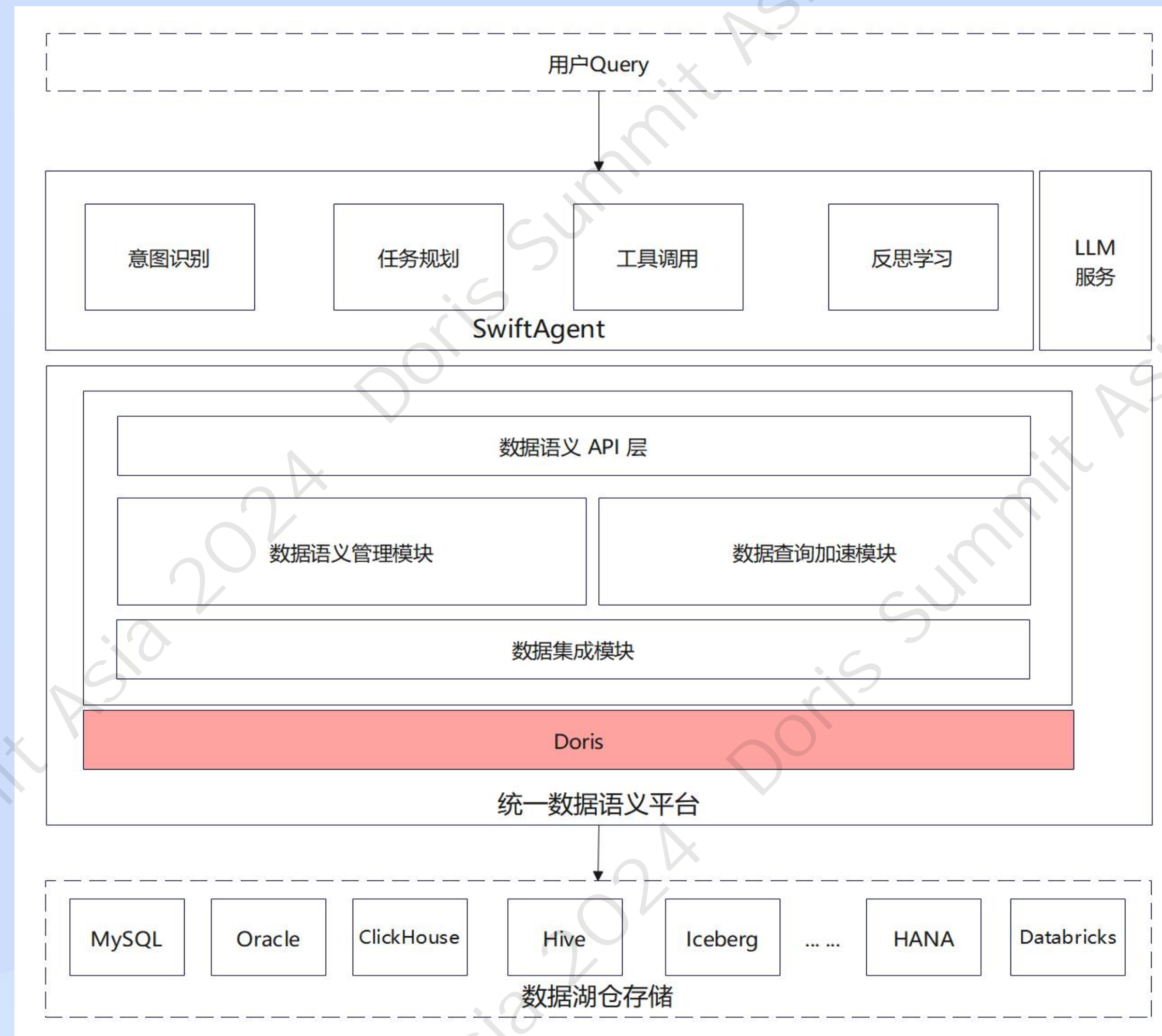
④ 数据安全可靠保障

- 利用指标分析平台的**权限管理能力**，结合RBAC基础，对数据与指标进行精细化的权限管控，**实现数据查询的安全可控**

⑤ 能力覆盖更全

- 高级数据分析问题，可通过精准的指标进行关联与展示，实现单项数据可查、报表可展示、总结报告可生成

数势 SwiftAgent -- NL2MetricsAPI



1. 统一数据语义管理
2. 数据计算加速引擎
3. 多源异构数据接入
4. 精确用户意图理解
5. 持续反思学习，自主进化

03

数势科技基于 Apache Doris 数据语义平台建设实践

数势指标平台产品概况

新建指标

指标信息

填写“指标名称”及“指标英文名”后即可暂存;

*指标名称:

可疑贷款数量

6/200

*指标英文名:

doubtful_load_cnt

17/200

*业务含义:

有可能成为不良贷款的数目

12/200

趋势属性:

反向指标

指标别名:

可疑贷款数量

6/200

数值格式:

示例: 123456789.123 @自定义配置

SwiftAgent问答:

计算逻辑

配置模式

原子指标 ①

派生指标 ①

衍生指标 ①

数据来源

贷款申请模型表

指标计算

计算字段:

字段

贷款业务是否有民事判决记录(is_civil)

聚合方式:

sum

过滤条件

贷款业务申请状态(at)

=

常量

授信失败

×

时点属性

时点属性:

时点计算方式:

周期开始

可用维度:

贷款业务客户性质 ×

贷款业务产品类型 ×

贷款业务产品名称 ×

贷款业务贷款目的 ×

贷款业务申请方式 ×

贷款业务还款方式 ×

贷款业务期限 ×

贷款业务申请状态 ×

所属总行 ×

所属分行 ×

所属员工姓名 ×

贷款业务是否多头借贷 ×

贷款业务利率 ×

贷款业务是否征信白名单 ×

贷款业务是否有行政处罚记录 ×

贷款业务是否有民事判决记录 ×

统计粒度:

日 (来源字段: 统计时间-日[dt])

数据预览

编辑维度

维度信息 填写“维度名称”及“维度英文名”“维度类型”后即可暂存；

* 维度名称：

信用卡客户是否透支交易

11/32

* 维度英文名：

dim_isoverdrawninatransaction

29/40

* 维度类型：

枚举类型

维度分类：

信用卡

业务含义：

信用卡客户是否透支交易

11/200

SwiftAgent问答：

计算逻辑

计算规则 数据来源更新后，会自动触发维度的计算

* 数据来源：

bank_demo_hm_user_credit_trade_di

* 维度名取值：

字段

is_overdraft

级联项：

过滤条件：

输入#号可插入字段

</>插入字段

初始化

☐ 固定范围 ☒ 全量

更新规则 根据选择的时间范围和更新方式，定期刷新维度值

☐ 覆盖 ☒ 追加 ☐ 无

根据每次计算的结果，增加新的数据，更新已有的数据。

* 更新取值范围：

dt

最近

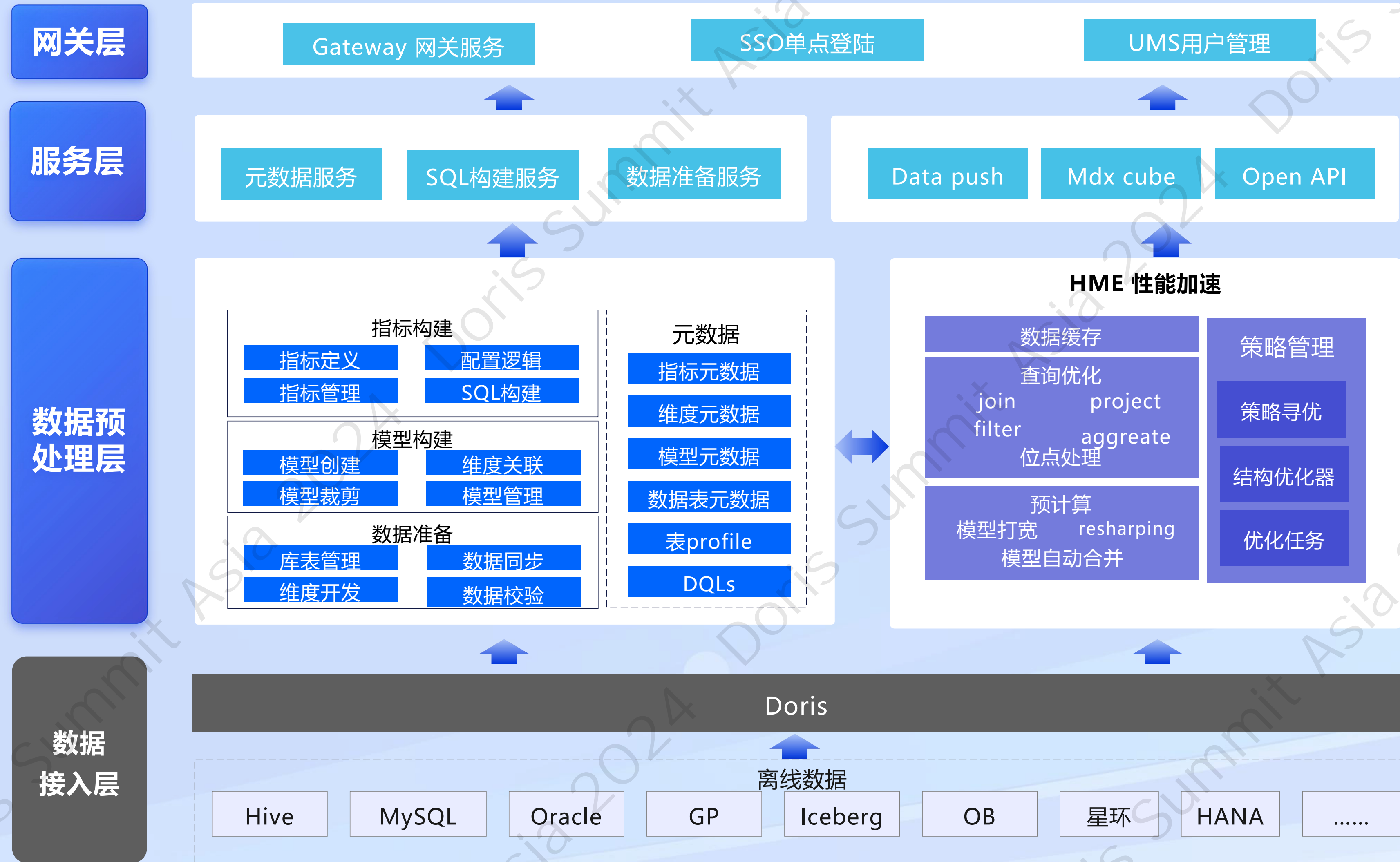
30

天

数据预览

[illegible]

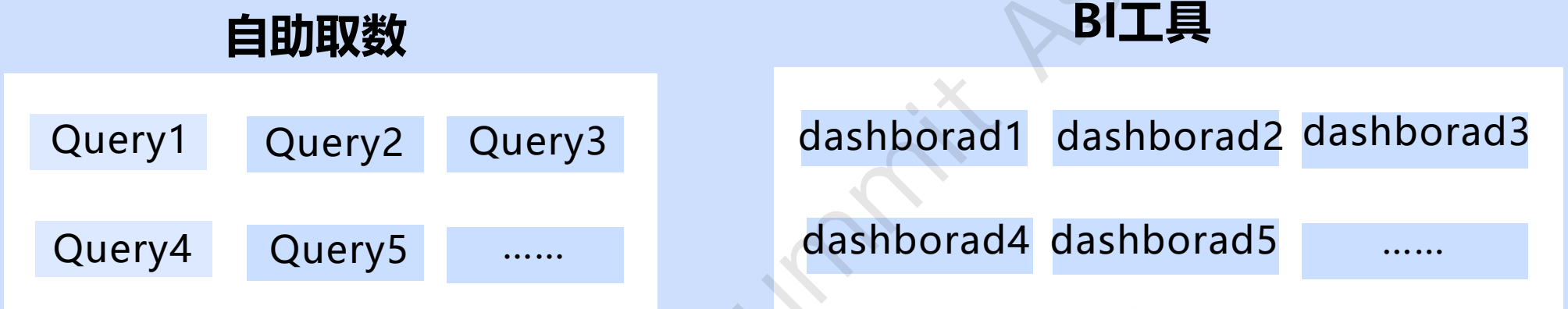
数势科技基于 APache Doris 的指标平台架构



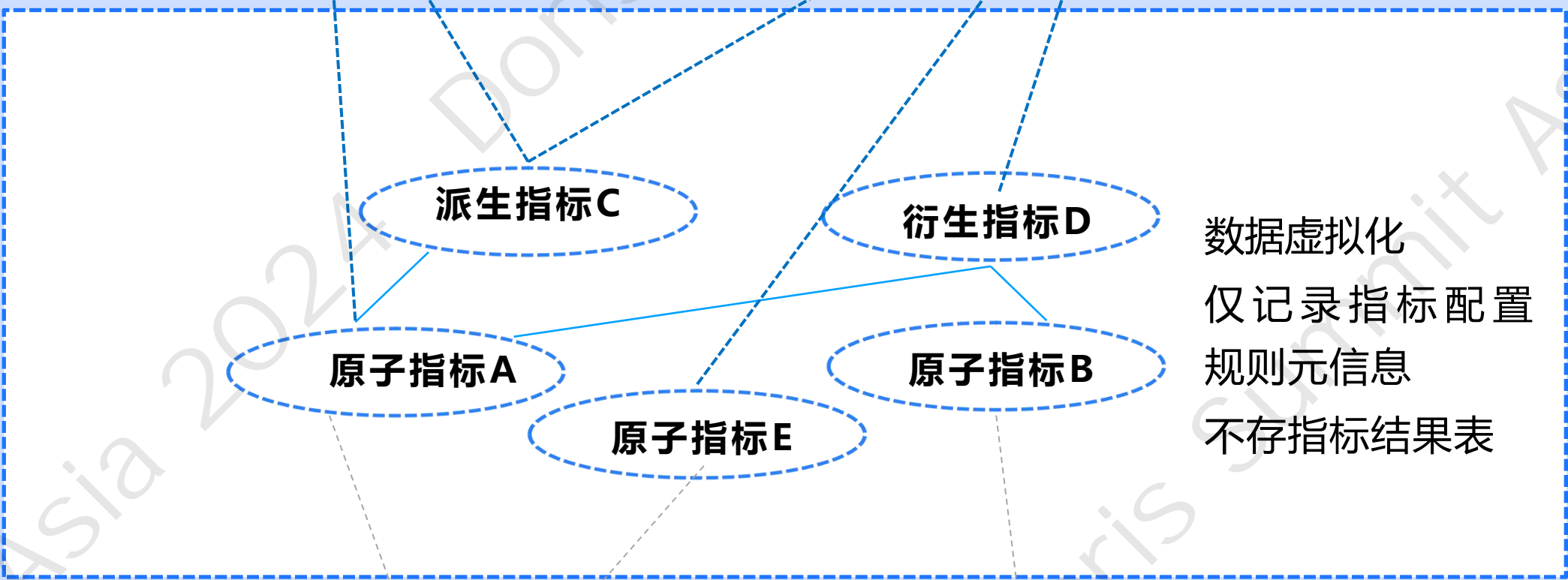
1. 指标管理高效、便捷；
2. 指标查询快速、灵活
3. 数据安全精细、可靠

DataFabric 理念：指标管理高效、便捷

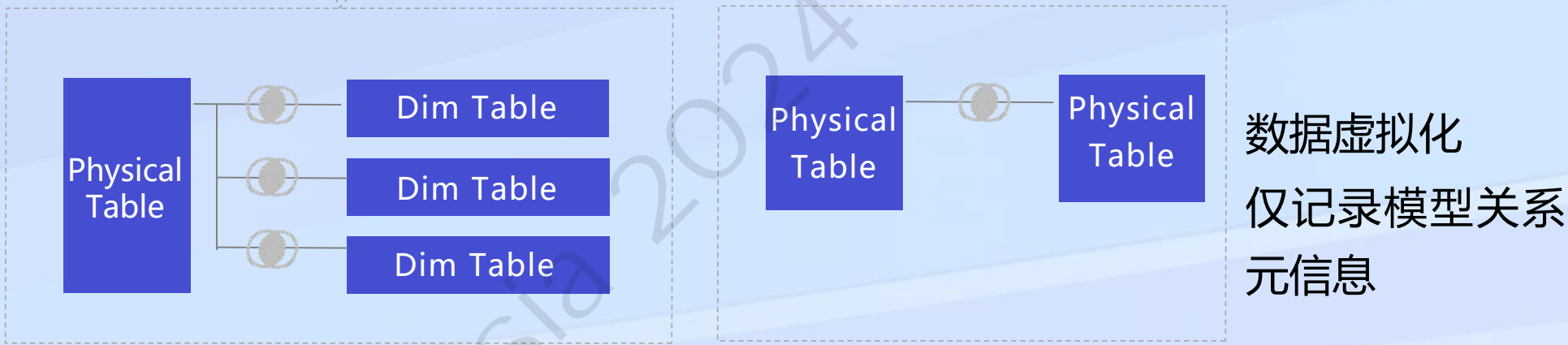
指标应用层



指标语义层



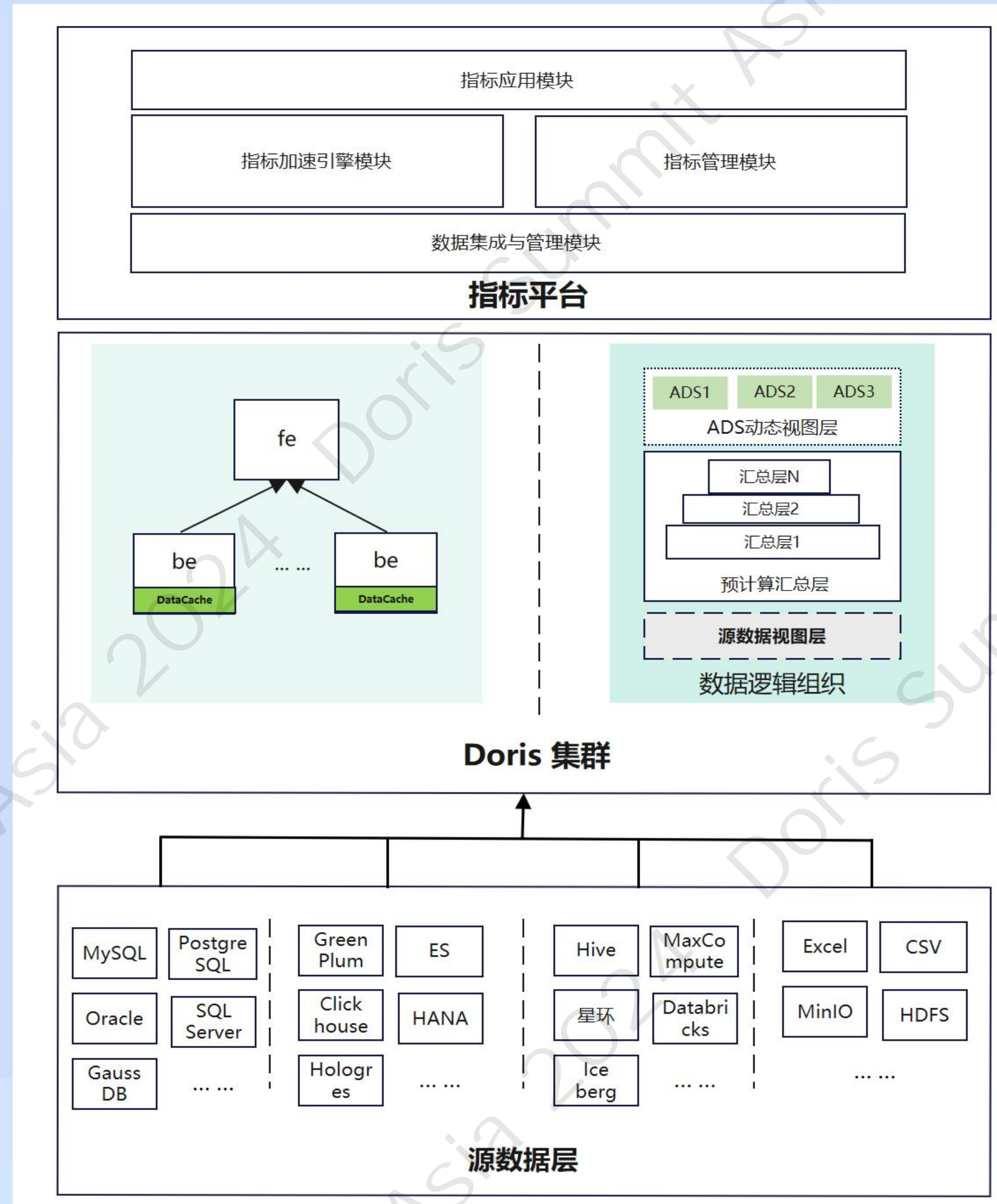
模型构建层



数据虚拟化：指标定义与数据解耦，灵活性

- 业务人员前台定义指标更灵活，所见即所得的获取结果，无需等待
- 指标加工函数和二次计算的算子都可以被独立封装成产品功能，技术实现方案可以分步增加

多源融合、灵活、极速

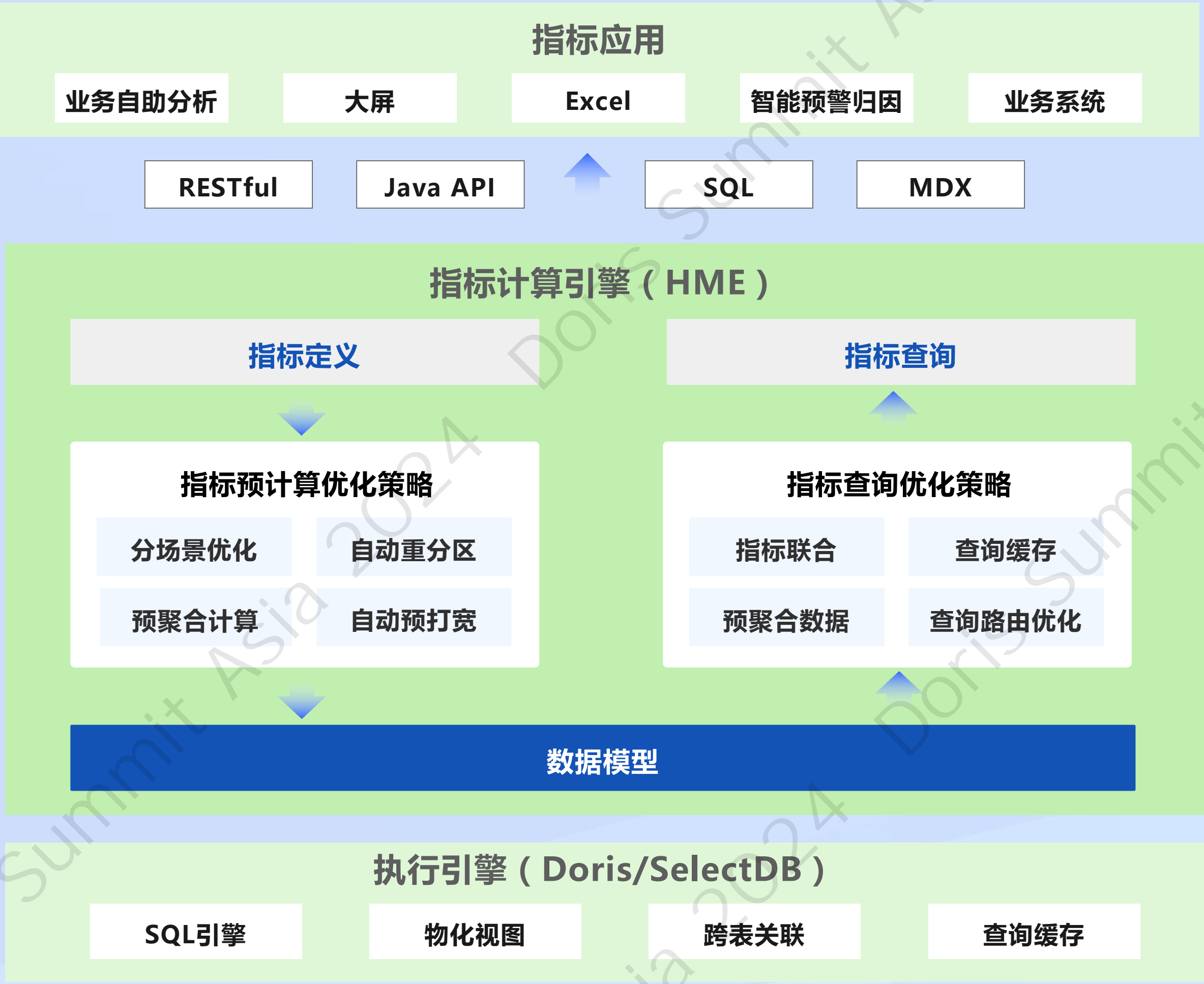


- Doris 存算分离、湖仓一体分析
- 跨源联邦数据查询分析、不搬运、更灵活
- Doris DataCache 机制 & 自研预计算加速引擎，兼顾性能

OLAP 引擎选型：指标查询快速、灵活

引擎特性		Doris	ClickHouse	Impala + Kudu + HDFS	Kylin
基础功能	列存支持	支持	支持	支持	支持
	动态分区	支持	不支持	不支持	不支持
	智能物化视图	支持很好	一般	一般	不支持
	事务支持	支持 ACID	100w以内支持原子性，DDL无事务保证	暂不支持多行事务	不支持
	Bitmap特性	支持	支持	不支持	不支持
查询能力	标准SQL	兼容标准 SQL	兼容性稍差	兼容标准SQL	兼容标准SQL
	数据查询Join	Join 方式最多	Join方式少	Join方式一般	依赖预先定义
	联邦查询	Hive/MySQL/ES/Hudi/IceBerg...	Hive/MySQL	不支持	不支持
引擎性能	查询性能	多表性能高，单表也不差	单表性能高	性能一般	定义范围内性能高
	向量化执行	支持	支持	不支持	不支持
	存储副本粒度	表级别	集群级别	表级别	表级别
集群能力	元数据管理	支持	没有，节点自己管理	支持	支持
	分布式能力	强大	较差	存算分离	存算分离、读写分离
	支持数据量级	PB 级	PB 级	PB 级	PB 级
	集群扩缩容	非常灵活	复杂且繁琐	灵活但繁琐	灵活但繁琐
场景支持	应用场景	圈群/实时&高并发查询/实时更新	普通分析场景	实时更新场景	多维分析

旁路智能预计算加速引擎：HME



核心加速策略：

- 1、自动预打宽「Join」：根据模型定义将常用的维度与明细数据进行打宽关联；
- 2、自动重分区「resharding」：根据指标口径的业务时间对数据进行重分布，提升数据扫描效率；
- 3、自动预聚合「rollup」：根据指标聚合粒度和聚合维度对明细数据进行多粒度/多维度的聚合；
- 4、自动去重「merge」：根据指标业务含义，对一定范围内的重复更新数据进行去重；
- 5、自动缓存「cache」：对常用/热度较高的指标计算结果进行缓存；
- 6、多预计算选取：自动选择执行代价最小的「预计算模型」；
- 7、支持复杂指标（衍生/派生）的查询优化；
- 8、ROI：预计算策略的调优；

... ..

贴合业务场景需要的加速引擎

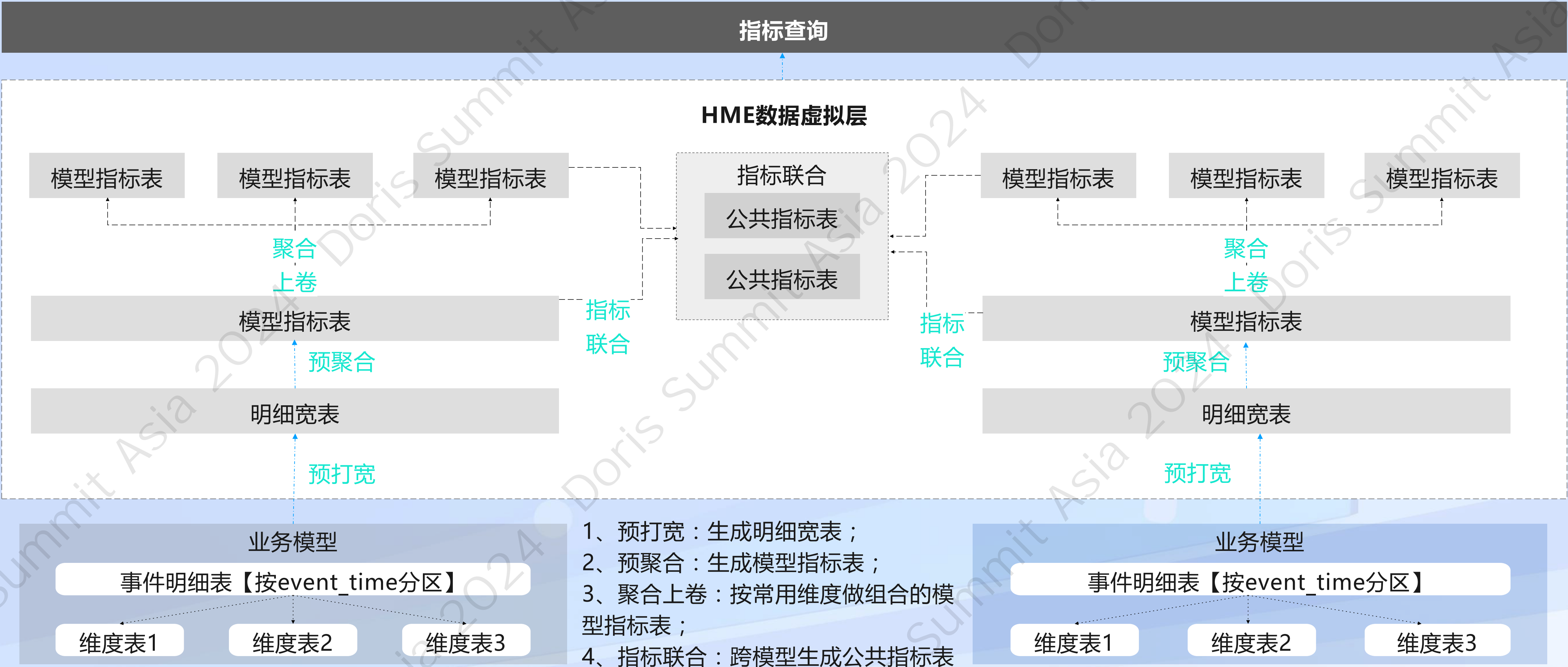
指标加工与查询方式，推导出HME需要解决的典型场景和功能点，进而抽象出通用的HME优化策略

指标加工过程抽象	查询方式	查询指标	指标类型	查询模型	模型类型
	单指标	A1	原子	模型1	单模型
		A3		模型3	
		B1	派生	模型1	单模型
		A2 (or B2/B3)	带CD(特殊的原子和派生)	模型2	单模型
		C1	衍生	模型1	单模型
		C2	衍生	模型2、模型3	跨模型
		B1	同环比(特殊的衍生)	模型1	单模型
	多指标	A1、B1、C1	不区分，同单指标查询	模型1	单模型
		A2、B2、B3		模型2	
		A1、A2、B1、B2、B3、C1	不区分，视为因子指标	模型1、模型2	跨模型
		A1、A3、B1、C1		模型1、模型3	
		A2、A3、B2、B3、C2		模型2、模型3	
		A1、A2、A3、B1、B2、B3、C1、C2		模型1、模型2、模型3	

通用抽象

模型类型	典型场景/功能	HME策略
单模型	大数据量基础指标	跨模型指标联合
	复杂指标-衍生	
	复杂指标-带CD	
	复杂指标-同环比	聚合上卷
	大数据量复杂指标	
	单模型多指标	单模型指标预聚合
跨模型	跨模型多指标	数据模型预打宽
	跨模型衍生指标	
模型变更		
计算任务管理		

HME 核心优化策略



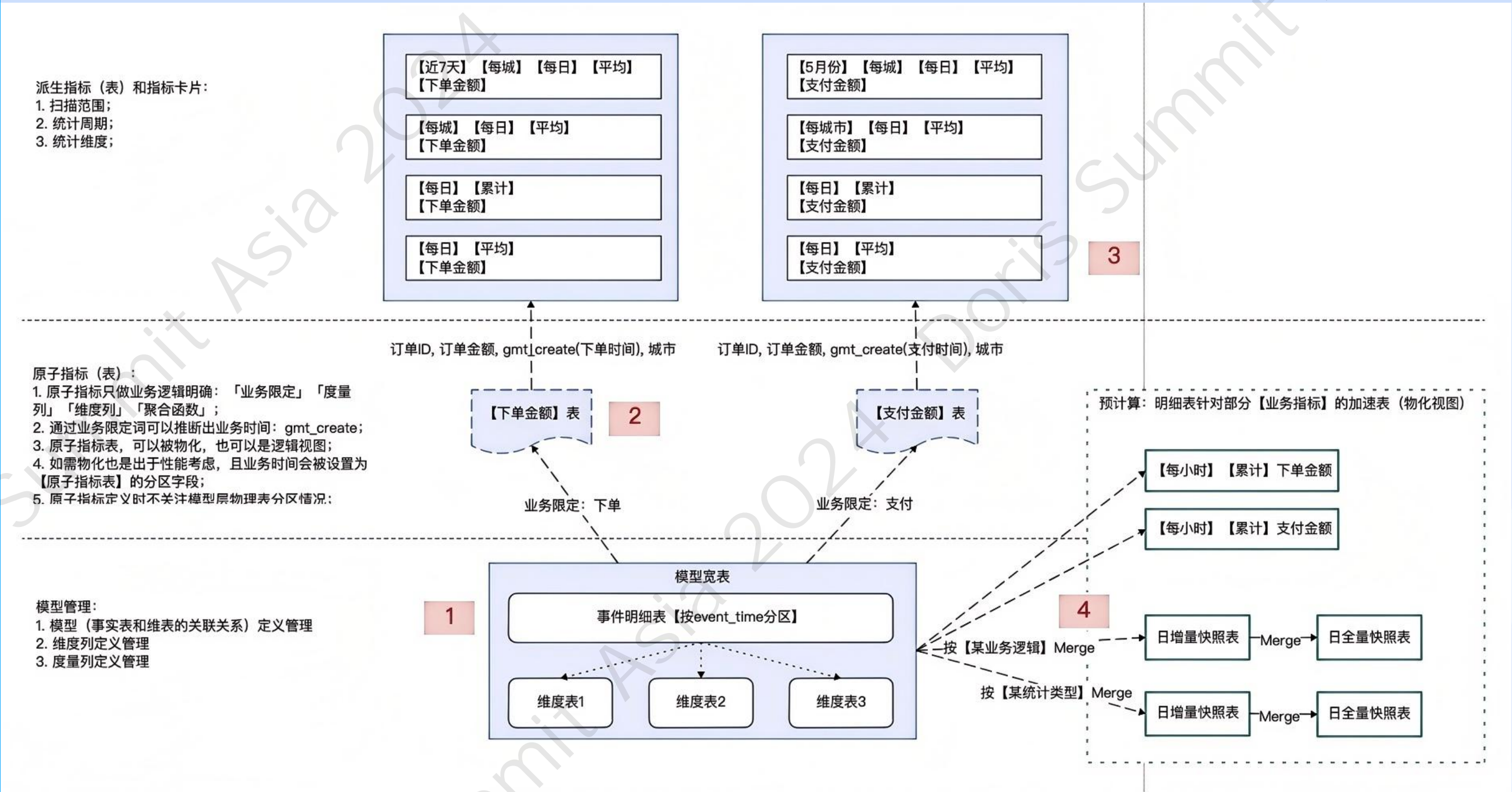
HME 加速示例

HME 优化 3：命中聚合结果

```
select
  star_time as group_time,
  comeon1 as val
from
(
  select
    star_time as star_time,
    SUM(comeon1) as comeon1
  from
    (
      select
        date_format(hme_dwd_order_info_hm0_202.star_time, '%Y-%m-%d') as star_time,
        hme_dwd_order_info_hm0_202.comeon1 as comeon1,
        hme_dwd_order_info_hm0_202.city_type_id as city_type_id,
        hme_dwd_order_info_hm0_202.city_type_name as city_type_name,
        hme_dwd_order_info_hm0_202.cate_type_id as cate_type_id,
        hme_dwd_order_info_hm0_202.cate_type_name as cate_type_name
      from
        `hme_dwd_order_info_hm0_202` as `hme_dwd_order_info_hm0_202`
      where star_time= '2023-03-06'
    ) dim_model

  group by
    star_time
) tbl
where
  (star_time = '2023-03-06')
```

执行耗时：359ms



客户真实环境性能压测报告

集群规模：

Doris fe：3台*<8C, 32G, 500G>;

Doris be: 5台*<32C, 128G, 10T>

数据准备：

事实表名	行数	Disk 占用
tbl_fact_1	136亿	19T
tbl_fact_2	136亿	3.2T
tbl_fact_3	136亿	2.9T
tbl_fact_4	136亿	1.6T
tbl_fact_5	136亿	1.4T

压测指标：

- 单表(明细表)模型数据量：136亿
- 查询维度基数组合：<500W
- 指标函数类型：sum、avg、max、min、count、count_distinct ...
- 响应时长：P95 < 2 s , P50 < 500 ms
- 可用性：>= 99.9%
- QPS: 50 / 100
- 网络：50Mbps
- CPU：大于90%的时间不超过5秒，大于60%时间不超过3分钟
- 内存：超过80%时间不大于5分钟

5.2.1 场景一、日粒度 3 指标，维度 2 个，查询周、月

【20+用户并发，持续 300s】

场景说明：所有自然日单指标一起取数，涉及的指标类型有：原子、派生、衍生。时间粒度：日，时间偏移：季至今
指标取数结果集在 1000+条数据。

指标类型（自然日）	并发用户	请求总数	错误率	最小耗时(ms)	最大耗时(ms)	50%耗时(ms)	90%耗时(ms)	95%耗时(ms)	每秒事务数
场景 1-日粒度 3 指标，维度 2 个，查询周	20	4444	0.00%	85	516	146	182	206	132.6
场景 1-日粒度 3 指标，维度 2 个，查询月	20	4340	0.00%	86	580	145	179	200	135.83

小结：

- 1、3 指标，2 维度，周查询 P50 146ms <500ms,P95 206ms< 2s TPS 132 大于 100，整体性能非常满足本次压测标准
- 2、3 指标，2 维度，月查询 P50 145ms <500ms,P95 200ms< 2s TPS 135 大于 100，整体性能非常满足本次压测标准
- 3、数据表明系统在该场景下处理 20 并发请求时，整体稳定性较好，最大耗时和平均耗时分布比较均匀
- 4、该场景下 HM 服务的 cpu 接近 90%，BE、FE、Mysql cpu 和内存均低于 50%

5.2.3 场景三、月粒度 10 个指标，带 3 个枚举维度，查询月、季、年

【35 用户并发，持续 300s】

场景说明：所有自然日单指标一起取数，涉及的指标类型有：原子、派生、衍生。时间粒度：月，时间偏移：最近 7 日、月至今、季至今、年至今，时间跨度通过月粒度转化

指标取数结果集在 20w 到 50w 条数据。

指标类型（自然日）	并发用户	请求总数	错误率	最小耗时(ms)	最大耗时(ms)	50%耗时(ms)	90%耗时(ms)	95%耗时(ms)	每秒事务数
场景 3-月粒度 10 个指标，带 3 个枚举维度，查询月	35	16699	0.00%	154	1103	361	547	596	92.61
场景 3-月粒度 10 个指标，带 3 个枚举维度，查询季	35	16917	0.00%	182	1084	362	489	531	93.83
场景 3-月粒度 10 个指标，带 3 个枚举维度，查询年	35	12724	0.00%	222	1028	489	578	612	70.52

小结：

- 1、10 指标，3 维度，查询 2 月 P50 361ms <500ms,P95 596ms< 2s TPS 92 大于 50，整体性能满足本次压测标准
- 2、10 指标，3 维度，查询 1 季 P50 362ms <500ms,P95 531ms< 2s TPS 93 大于 50，整体性能满足本次压测标准
- 3、10 指标，3 维度，查询 1 年 P50 489ms <500ms,P95 612ms< 2s TPS 70 大于 50，整体性能满足本次压测标准
- 4、数据表明系统在该场景下处理 35 并发请求时，整体稳定性较好，对月、季请求的处理能力基本相同，年请求处理能力稍弱
- 5、该场景下 BE、HM 服务的 cpu 接近 70%，FE、Mysql cpu 和内存均低于 50%

5.2.2 场景二、月粒度 6 个指标，带 5 个维度，查询月、季、年

【35 用户并发，持续 300s】

场景说明：所有自然日单指标一起取数，涉及的指标类型有：原子、派生、衍生。时间粒度：月，时间偏移：最近 7 日、月至今、季至今、年至今，时间跨度按实际月、季、年单位查询

指标取数结果集在 40w 到 50w 条数据。

指标类型（自然日）	并发用户	请求总数	错误率	最小耗时(ms)	最大耗时(ms)	50%耗时(ms)	90%耗时(ms)	95%耗时(ms)	每秒事务数
场景 2-月粒度 6 个指标，带 5 个维度，查询月	35	10214	0.00%	222	1371	604	738	781	56.52
场景 2-月粒度 6 个指标，带 5 个维度，查询季	35	11017	0.00%	198	1410	565	700	760	61.03
场景 2-月粒度 6 个指标，带 5 个维度，查询年	35	9590	0.00%	233	1605	645	786	832	53.12

小结：

- 1、6 指标，5 维度，以月为单位查询 P50 604ms >500ms,P95 781ms< 2s TPS 56 大于 50，整体性能满足本次压测标准
- 2、6 指标，5 维度，以季为单位查询 P50 565ms >500ms,P95 760ms< 2s TPS 61 大于 50，整体性能满足本次压测标准
- 3、6 指标，5 维度，以年为单位查询 P50 645ms >500ms,P95 832ms< 2s TPS 53 大于 50，整体性能满足本次压测标准
- 4、数据表明系统在该场景下处理 35 并发请求时，整体稳定性较好，对月、季、年请求的处理能力基本相同
- 5、该场景下 BE 服务的 cpu 接近 80%，HM、FE、Mysql cpu 和内存均低于 50%

5.2.4 场景四、月粒度 3 个指标，带 3 个维度，带 3 个指标季、年同环比

【35 用户并发，持续 300s】

场景说明：所有自然日单指标一起取数，涉及的指标类型有：原子、派生、衍生。时间粒度：月，时间偏移：最近 7 日、月至今、季至今、年至今

指标取数结果集在 40w 到 50w 条数据。

指标类型（自然日）	并发用户	请求总数	错误率	最小耗时(ms)	最大耗时(ms)	50%耗时(ms)	90%耗时(ms)	95%耗时(ms)	每秒事务数
场景 4-月粒度 3 个指标，带 3 个维度，带 3 个指标季环比	35	16200	0.00%	176	1412	364	572	633	89.76
场景 4-月粒度 3 个指标，带 3 个维度，带 3 个指标年环比	35	17918	0.00%	109	783	334	520	558	99.4
场景 4-月粒度 3 个指标，带 3 个维度，带 3 个指标同环比	35	8582	0.00%	165	2038	740	978	1050.85	47.57

小结：

- 1、3 指标，3 维度，带季环比+同比，P50 364ms <500ms,P95 633ms<2s TPS 89 大于 50，整体性能满足本次压测标准
- 2、3 指标，3 维度，带年环比+同比，P50 334ms <500ms,P95 558ms<2s TPS 99 大于 50，整体性能满足本次压测标准
- 3、3 指标，3 维度，带 3 指标同环比，P50 740ms >500ms,P95 1050ms<2s TPS 47 接近 50，整体性能不满足本次压测标准
- 4、该场景下 HM 服务的 cpu 接近 80%，BE、FE、Mysql cpu 和内存均低于 50%

.....

数据安全精细、可靠 -- 指标平台权限控制

RBAC 权限模型，行列级精细化控制，安全可靠

配置查看权限

配置使用权限

<input type="checkbox"/>	ID	指标名称	指标类型	状态	查看权限	操作
<input type="checkbox"/>	1440	原子1	原子指标	已上线	指定授权	查看权限 使用权限
<input type="checkbox"/>	1439	zxeexsr5	派生指标	已上线	公开	查看权限 使用权限
<input type="checkbox"/>	1438	zxeexsr4	派生指标	已上线	公开	查看权限 使用权限
<input type="checkbox"/>	1437	zxeexsr3	派生指标	已上线	公开	查看权限 使用权限
<input type="checkbox"/>	1436	zxeexsr	原子指标	已上线	公开	查看权限 使用权限
<input type="checkbox"/>	1374	逾期销售额	派生指标	已上线	指定授权	查看权限 使用权限

配置查看权限

原子指标 ID: 1440

原子1

权限类型: 指定授权

授权用户

授权角色

搜索用户

批量操作

+ 添加用户

<input type="checkbox"/>	用户名称	可查看信息	配置时间	配置人	操作
<input type="checkbox"/>	毕先生(wx_...	全部	2023-09-18	admin10005	编辑 删除

添加用户

全部用户 (116)

搜索用户

☐ 李里(1661615423150014464)

☐ 高楠(1665532399267852288)

☐ 谢明昊(1674323767564353536)

☒ 客户试用账号(CS-LVMH)

☐ Cindy(Cindy)

☐ TB_POC_管理员(TB_POC_Administrator)

☐ TB_POC_业务用户(TB_POC_user)

已选用户 (1)

搜索用户

客户试用账号(CS-LVMH)

☒ 业务会议

☒ 业务负责人

☒ 英文名称

☒ 统计粒度

☒ 可用维度

☒ 聚合方式

添加角色

全部角色 (6)

搜索角色

☐ 管理员

☒ 全部客户数据运营

☒ 北京地区客户数据运营

☐ 业务人员

☐ 上海地区客户数据运营

☐ 技术人员

已选角色 (2)

搜索角色

全部客户数据运营

北京地区客户数据运营

可用数据范围: 指定数据范围

供应商 = 固定值 鸭肉供应商

品牌 = 固定值 jw-4

门店 = 固定值 东莞市4店

东莞市4店

广州市9店

东莞市8店

广州市5店

福州市1店

福州市5店

福州市9店

04

案例分享

案例分享（一）某零售客户

基于数势SwiftAgent，某头部茶饮连锁品牌近万家门店店长实现基于数据的运营变革



简化决策，放大成效

准确性

- 用户意图识别率 **>98%**
- 复杂任务规划准确率 **>95%**

效率提升

- 分析工作处理时长 **减少80%**
- 每人每周**减少10小时**+数据处理工作

用户满意度

- 使用者满意度 **9.3分+**

方案覆盖度

- 分析解决方案覆盖度 **>90%**

交互友好度

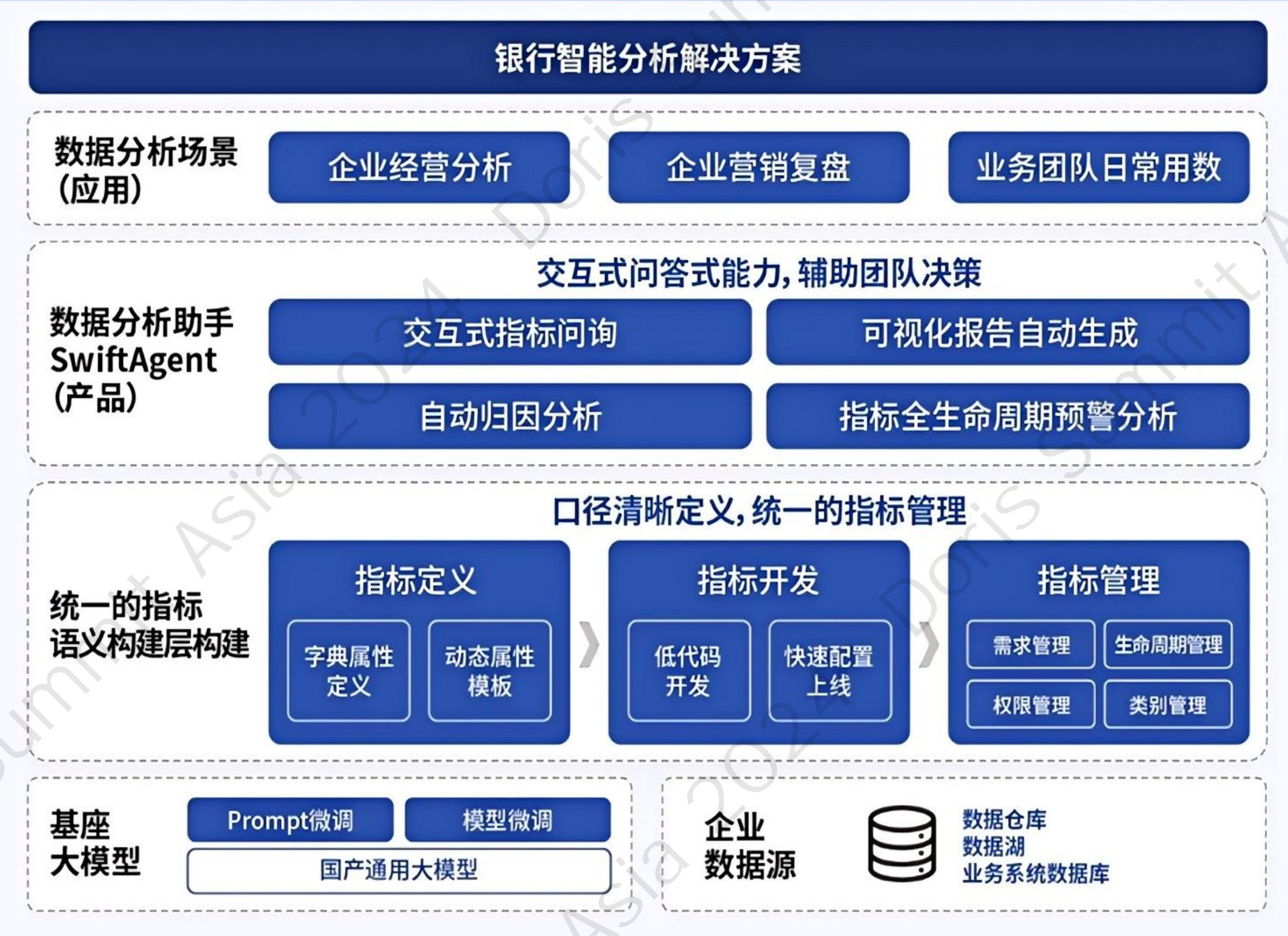
- 用户界面友好度 **9.5分**

强化学习

- 具备**强化学习正反馈与追问能力**

案例分享（二）某银行客户

数势科技以 SwiftAgent 产品为核心，利用行业知识和数据分析模型，在理解策略目标基础上，对某城商行经营矩阵实现了从数据到价值的快速转化。



- 1. 降低人工成本：平均取数工单每天减少约 50%；
- 2. 提高决策效率：平均取数周期由 3 天降为 1 分钟；
- 3. 提高数据利用率：数仓已有数据资产表利用率提高 20%。
- 4. 提升员工满意度：系统满意度 9.5 分，Top 3。

DORIS
SUMMIT



Thanks for Watching !



扫码申请产品试用
领取演讲资料