

平安人寿基于 Apache Doris 统一 OLAP 技术栈实践

孙顺

平安人寿 大数据架构师

目录

- 1 大数据平台建设沿革与总览
- 2 早期架构与应用痛点
- 3 基于 Apache Doris 统一 OLAP 技术栈的演进之路
- 4 总结与未来规划

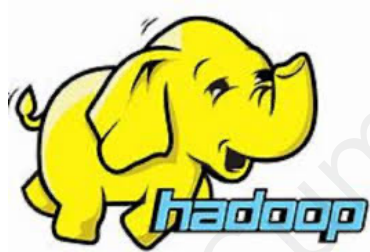
1 大数据平台建设沿革与总览

大数据平台建设历程

ORACLE

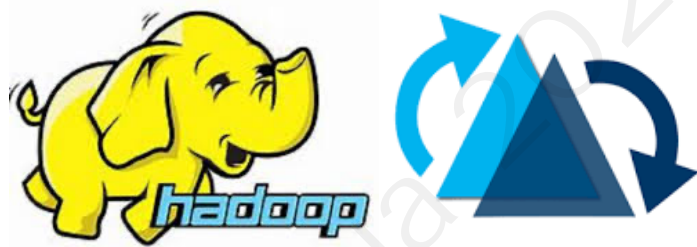
- 数据仓库与数据集市
- 1. 集中业务系统数据
 - 2. 按需开发数据报表

2005



- 大数据 Hadoop 平台
- 1. 存储力和算力大
 - 2. 提升报表刷新时效

2009



- 设计数据中台
- 1. 数据治理
 - 2. 数据底座
 - 3. 数据产品

2016



大数据产品体系

2022

至今

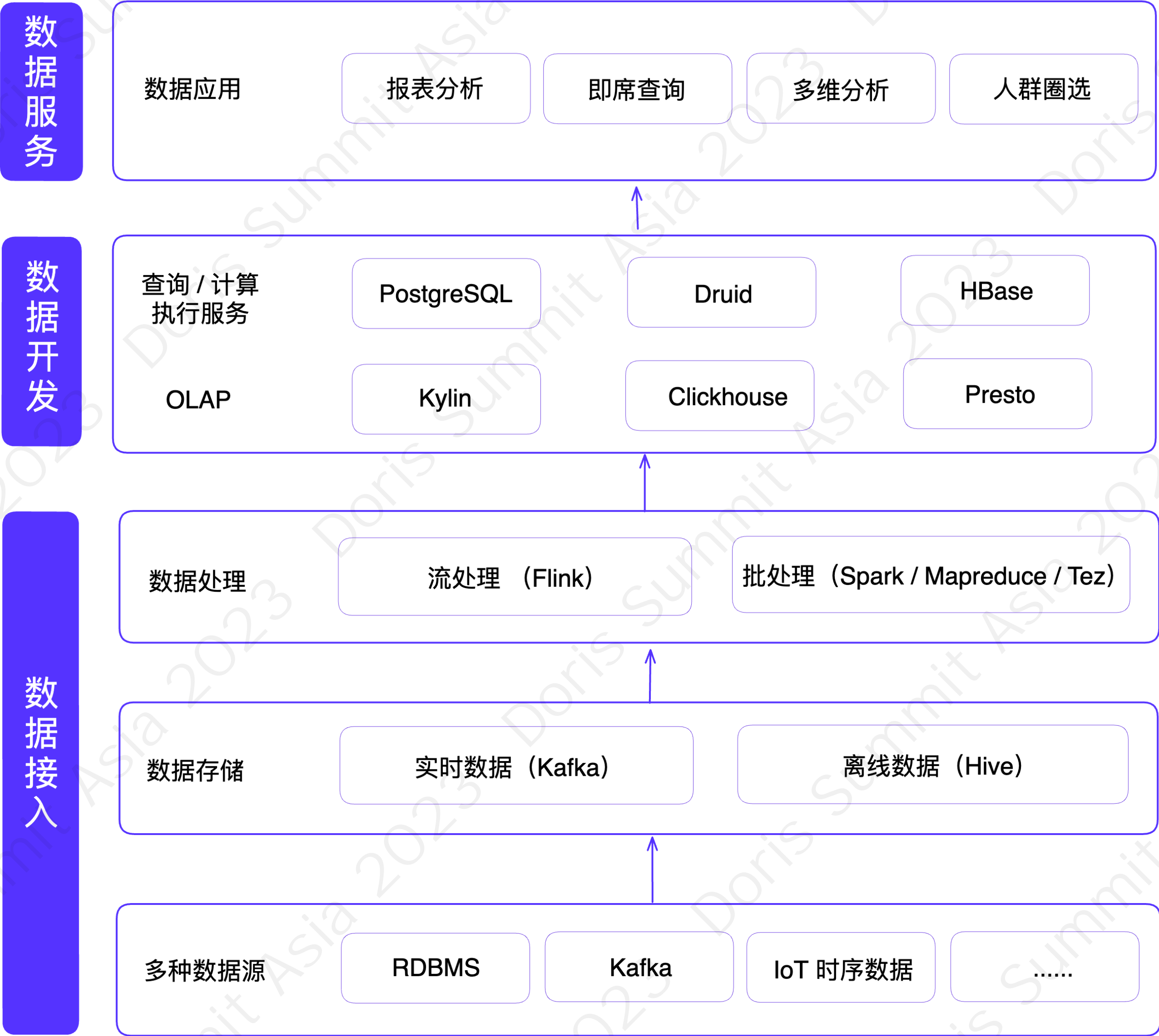
统一 OLAP 技术栈、统一指标与标签管建、统一数据服务

大数据产品体系总览



2 早期架构与应用痛点

早期架构总览



早期应用痛点

<div>报表分析</div> <div>Hive PostgreSQL</div>	<div>管理层经营数据探查</div> <div><div>✓ 业务洞察</div><div>✓ 问题定位</div><div>✓ 趋势预测</div><div>✓ 经营全貌概览</div></div>	<div>痛点</div> <div><div>✓ 多轮清洗计算，数据链路长</div><div>✓ PG 扩容成本高</div><div>✓ 导数冗余、资源浪费</div><div>✓ 应用层需配合改造</div></div>	<div>升级需求</div> <div><div>✓ 大规模数据处理能力</div><div>✓ 报表产出时效性</div><div>✓ 报表查询响应在百毫秒内</div><div>✓ 资源成本控制</div></div>
<div>即席查询</div> <div>Hive Presto PostgreSQL</div>	<div>总部业务运营人员</div> <div><div>✓ 客户经营</div><div>✓ 产品、理赔、核保等数据 可视化分析</div><div>✓ 指标数据集订阅</div></div>	<div>痛点</div> <div><div>✓ 查询性能不及预期</div><div>✓ Hive 权限管理不清晰</div><div>✓ 缺少资源隔离限制</div></div>	<div>升级需求</div> <div><div>✓ 支持高并发查询</div><div>✓ 支持低延迟响应</div><div>✓ 指标复用率</div></div>

早期应用痛点

<div>多维分析</div> <div>Druid</div>	<div>一线业务运营人员</div> <div><div>✓ 业务运营决策</div><div>✓ 业绩报表</div><div>✓ 实时数据监控和分析</div></div>	<div>痛点</div> <div><div>✓ 节点重启时间超 24 小时</div><div>✓ 查询灵活度低</div><div>✓ 非精确去重</div><div>✓ 运维成本高</div></div>	<div>升级需求</div> <div><div>✓ 高并发查询访问</div><div>✓ 报表查询响应时间</div><div>✓ 指标开发扩展性</div><div>✓ 精确去重与关联查询</div></div>
<div>人群圈选</div> <div>Clickhouse Kylin</div>	<div>总部与分公司营销人员</div> <div><div>✓ 人群圈选</div><div>✓ 画像分析</div><div>✓ 人群检测与追踪</div></div>	<div>痛点</div> <div><div>✓ CK 难以支撑 200 人并发查询</div><div>✓ 组件语法差别，适配成本高</div><div>✓ 依赖人工判断、灵活度不足</div><div>✓ 运维成本高</div></div>	<div>升级需求</div> <div><div>✓ 支持高并发查询</div><div>✓ 多表关联查询能力</div><div>✓ 数值间圈选</div><div>✓ Array 字段</div></div>

早期数据开发系统痛点

标签数据开发		指标数据开发	
组件	痛点	组件	痛点
<ul style="list-style-type: none">• HBase	<ul style="list-style-type: none">• 不支持二级索引• 日期、数字需额外转化• 不支持标签跨对象	<ul style="list-style-type: none">• PG• Presto• Druid	<ul style="list-style-type: none">• 依赖预计算• 不同组件支持指标存储与查询• 不支持复合指标开发

-
1. 平台组件不同，数据未打通
 2. 一个需求配置一个接口，无法动态配置与灵活调用

架构升级目标

持续构建一站式数据门户

开发、存储、运维成本控制

有效简化数据链路、架构技术栈，降低对技术人员开发的依赖

数据治理体系化、数据一致性

打通平台数据读取，将指标与标签数据统一存储，实现复合指标加工



综合性强、灵活度高

覆盖不同应用系统的日常分析需求，面向更多样化的业务场景

高效写入与极速分析

强化查询性能
丰富查询与写入的优化功能

3 基于 Apache Doris 统一 OLAP 技术栈的演进之路

为什么选择 Apache Doris

简单易用的架构

- 采用 MySQL 协议，支持标准 SQL
- MPP 架构，仅 FE 与 BE 两类进程
- 节点间负载均衡
- 列式存储，极高压缩比

极速的查询分析

- 支持多种存储模型
- 支持物化视图与 Rollup 预聚合
- 支持高并发与低延迟查询
- 支持多表关联查询
- 提供灵活的数值区间圈选 支持 Bitmap 精确去重，HyerLogLog 近似去重

高效的写入性能

- 支持主键模型写时合并
- 支持 Schema 在线变更
- 支持 Upsert 所有导入方式，条件更新、条件删除、部分列更新以及分区覆盖

基于 Apache Doris 构建全新大数据产品体系



统一 OLAP 技术栈

- 统一的数据查询与计算引擎
- 可插拔引擎
- 减少重复开发、冗余存储

统一的指标标签设计平台

- Doris 集群之上构建统一的指标标签设计平台
- 形成“上下经营一张表”
- 完善经营指标追踪体系

统一的数据服务平台

- API 接口服务化
- 打通各类应用系统的数据调取

统一 OLAP 技术栈，加速业务需求交付

交付效率提升

14 倍

原架构

- 链路长、导数重复、定制开发
- 14 个 PG 数据库成本压力
- 业务交付耗时 2 周

新架构

- 架构统一、运维成本降低
- 丰富的数据模型简化链路
- 业务交付仅需 1- 2 天

查询响应提速

10 倍

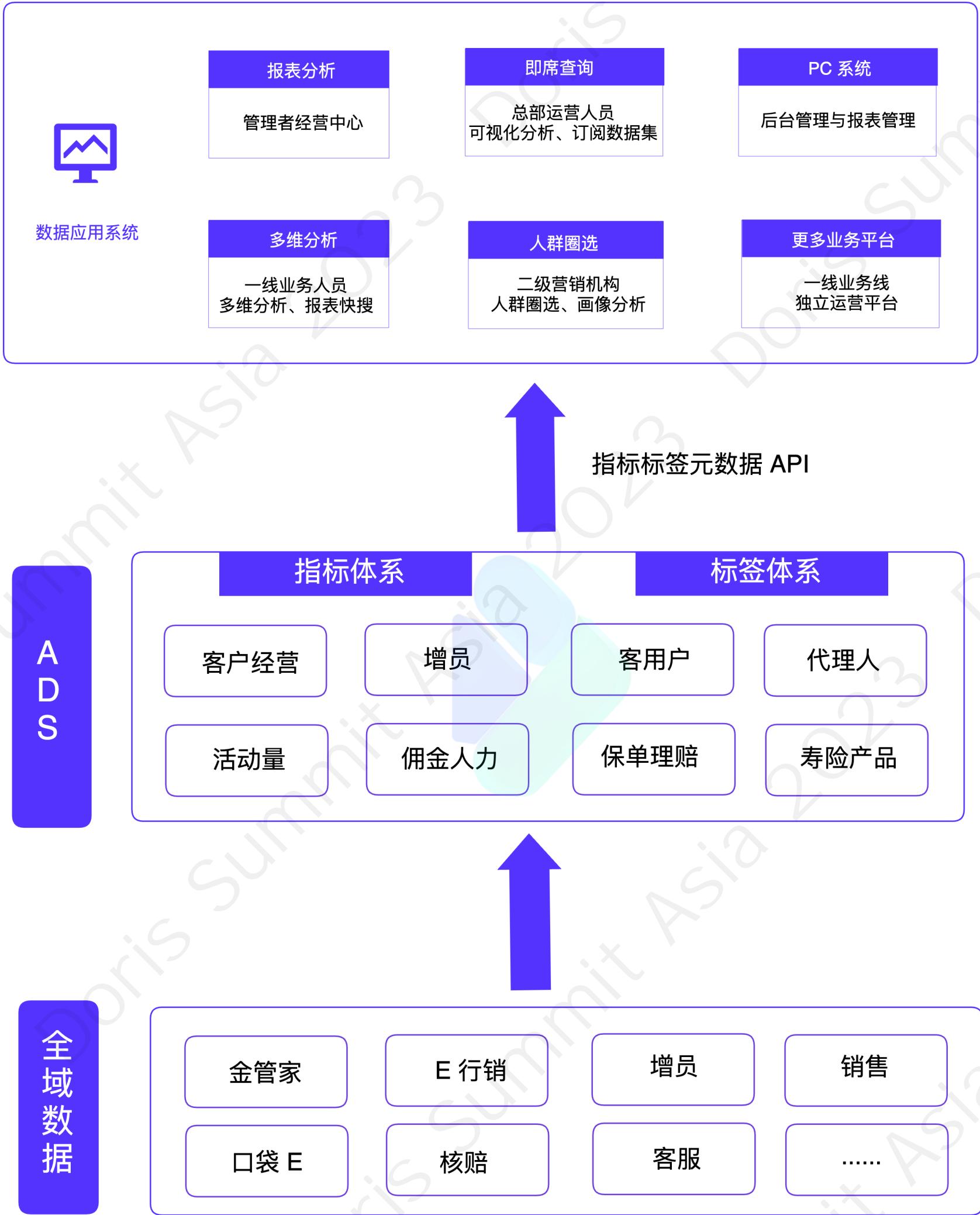
原架构

- 缺少权限、表交叉使用
- 查询分钟级耗时

新架构

- 业务人员只可以访问 ADS 层数据
- 提升数据指标复用率
- 查询秒级毫秒级响应

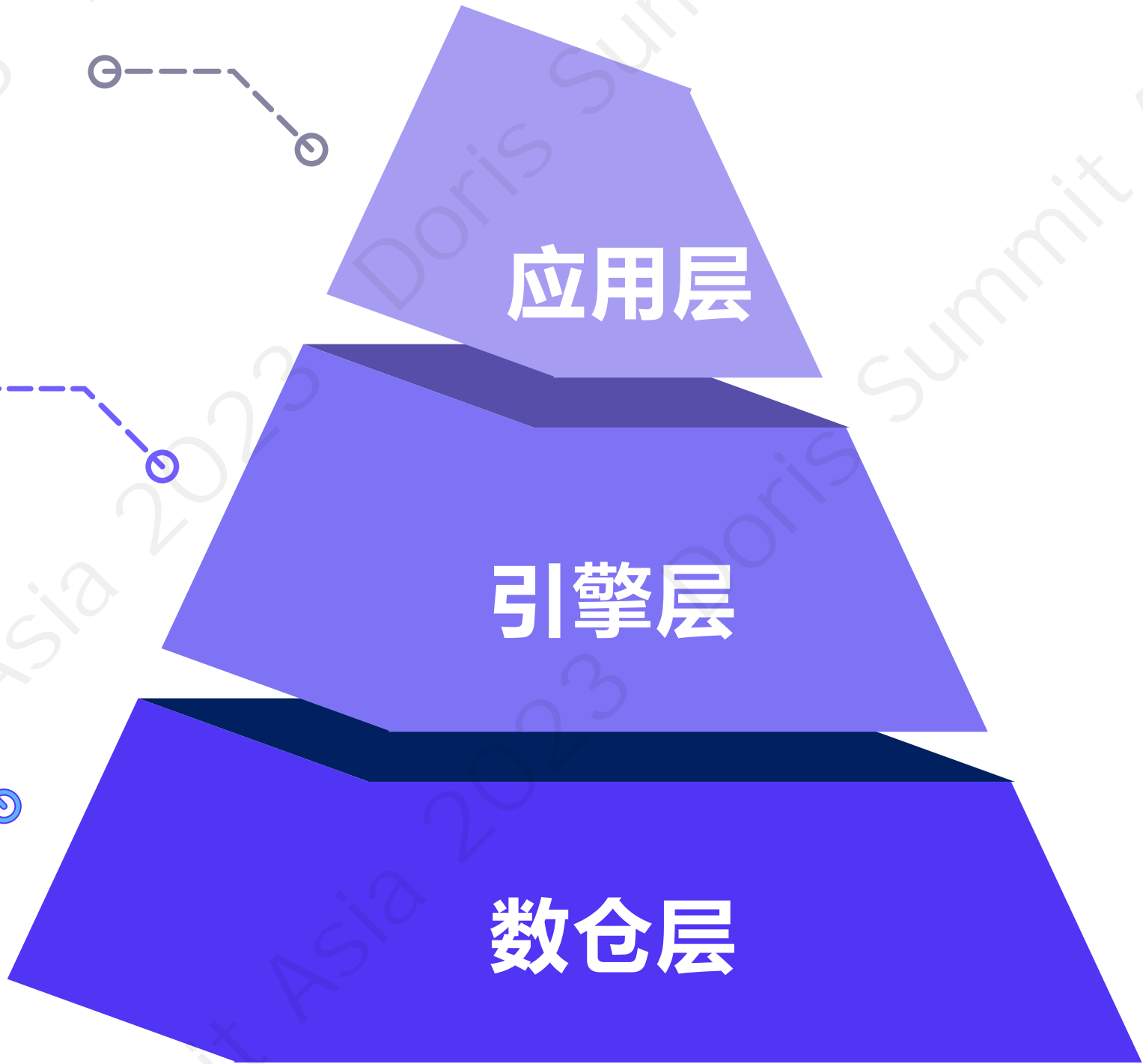
统一指标和标签管理，直通应用场景



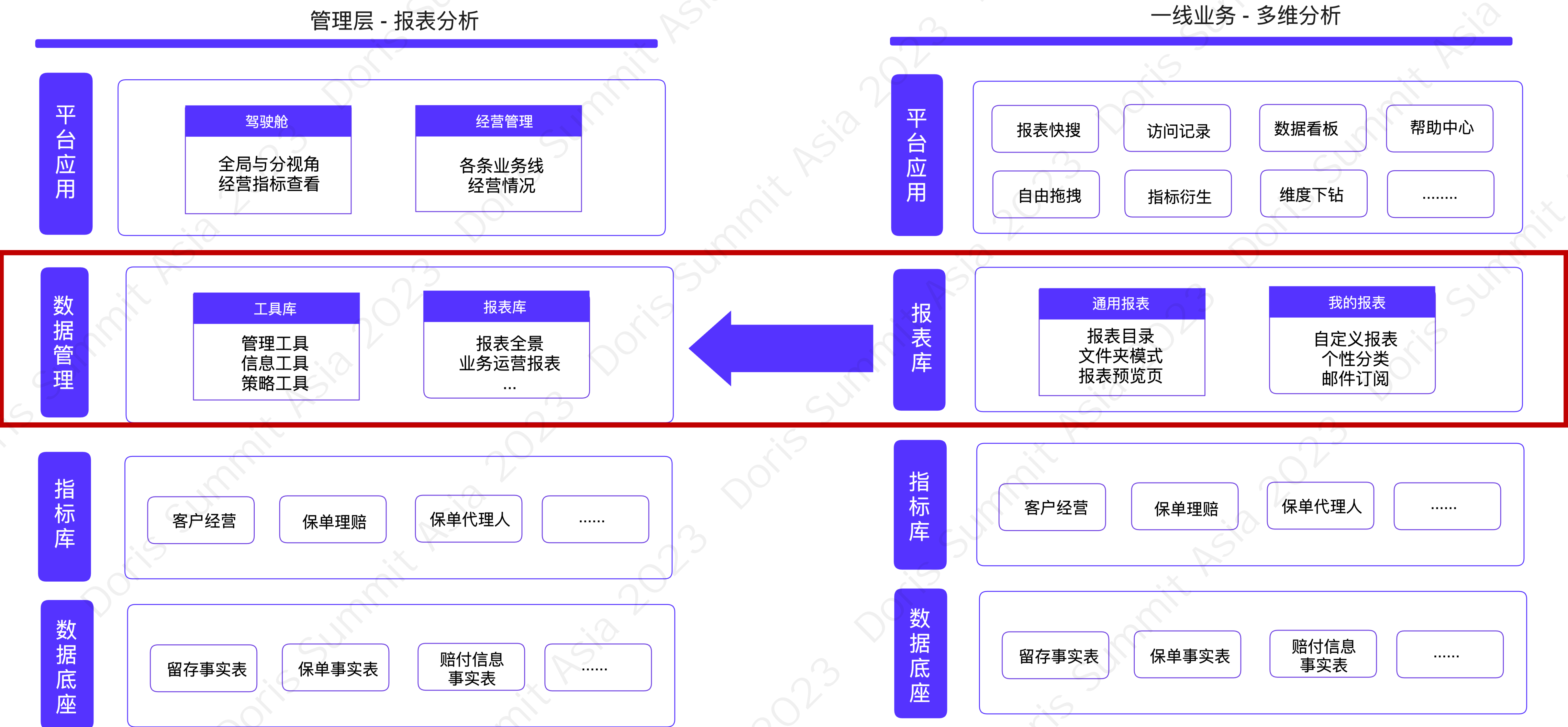
指标和标签服务 API

指标与标签统一计算

指标与标签定义



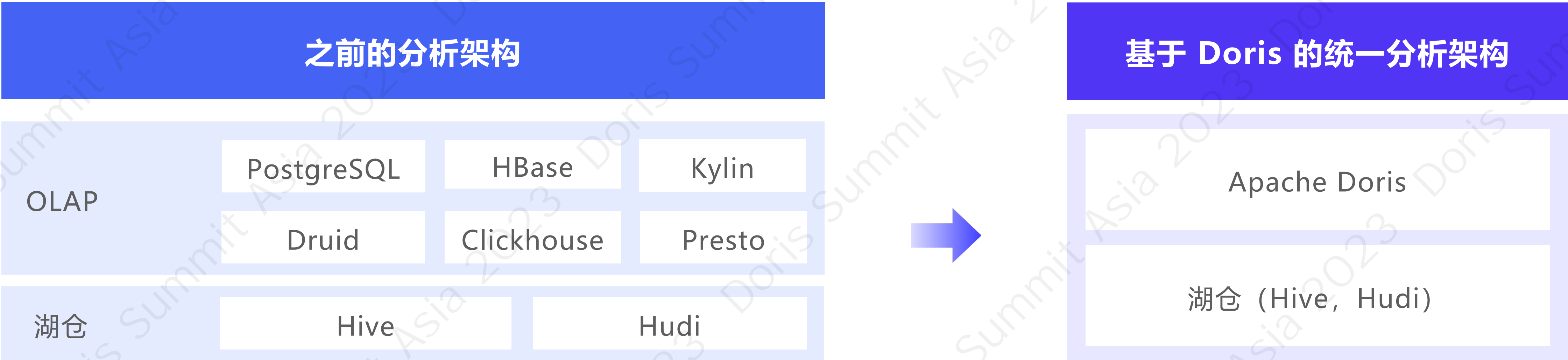
统一数据服务平台，打破数据孤岛



4 总结与未来规划

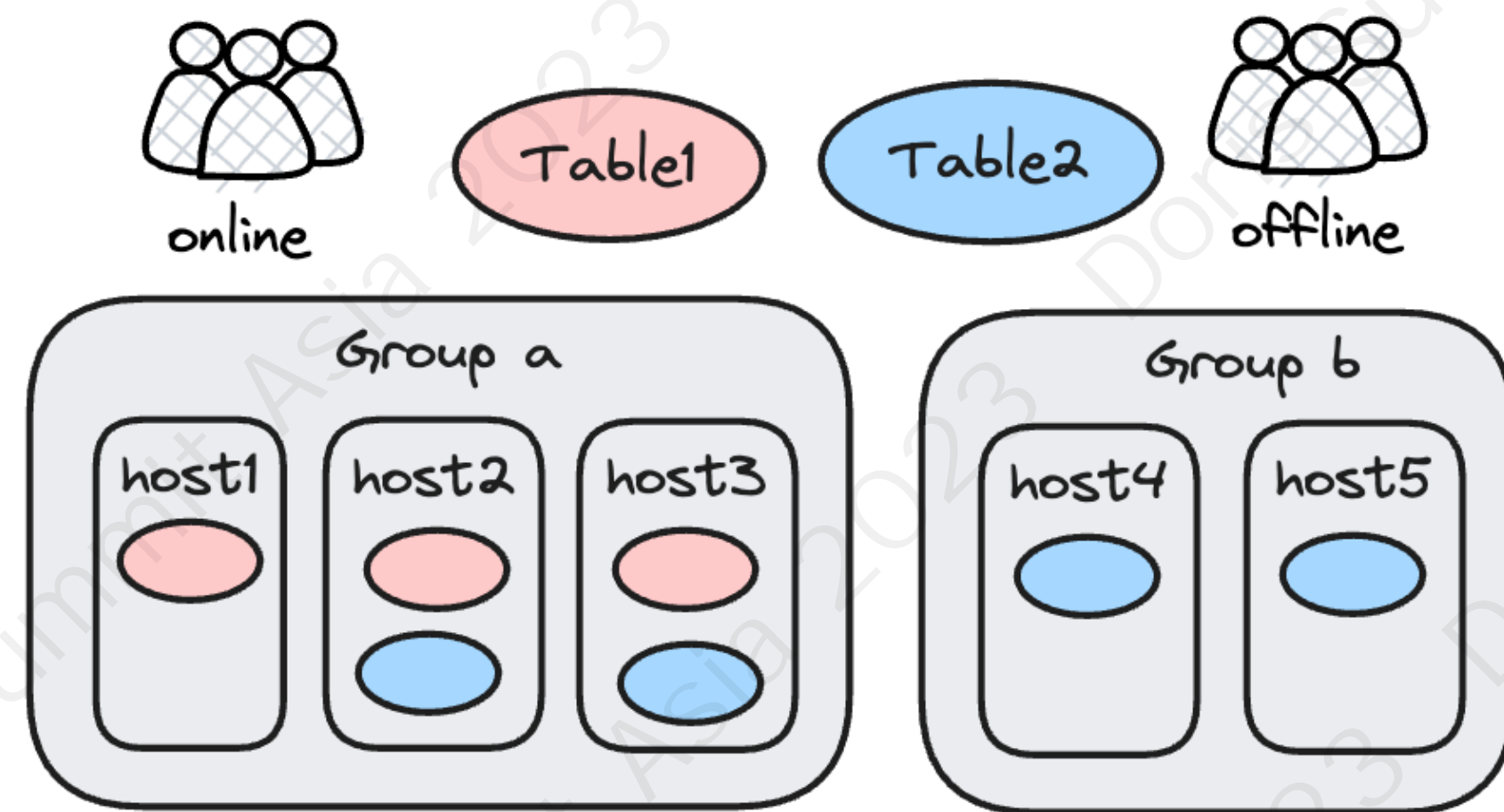
湖仓一体化

更开放、灵活、可扩展的企业级管理与分析大数据产品体系



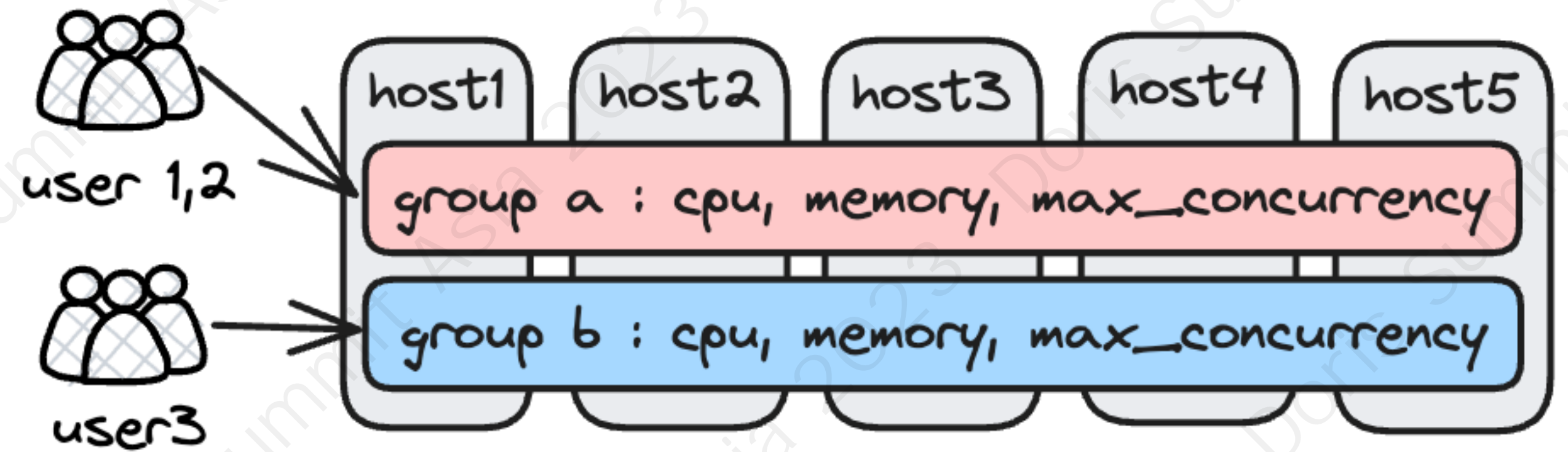
多租户资源隔离

完善应用系统之间负载均衡性能，使集群资源更合理分配



Resource Group

- 机器分组
- 副本放置到分组
- User 绑定分组



WorkLoad Group

- 对特定查询模式、分区/分片数大的查询拦截
- 单查询内存限制
- 多 Cluster 机制和 Resource Group 整合



获取更多社区动态与最佳实践

Apache Doris 官方平台:

- Apache Doris 官网: doris.apache.org
- Apache Doris GitHub: github.com/apache/doris/

获取更多峰会资料:

- Doris Summit 峰会官网: doris-summit.org.cn
- Doris Summit 峰会回放: <https://space.bilibili.com/1196172099/channel/collectiondetail?sid=1824324>