

# 众安保险 CDP 平台基于 Apache Doris 的应用实践

戴鸿文

众安保险 CDP 平台负责人

# 目录

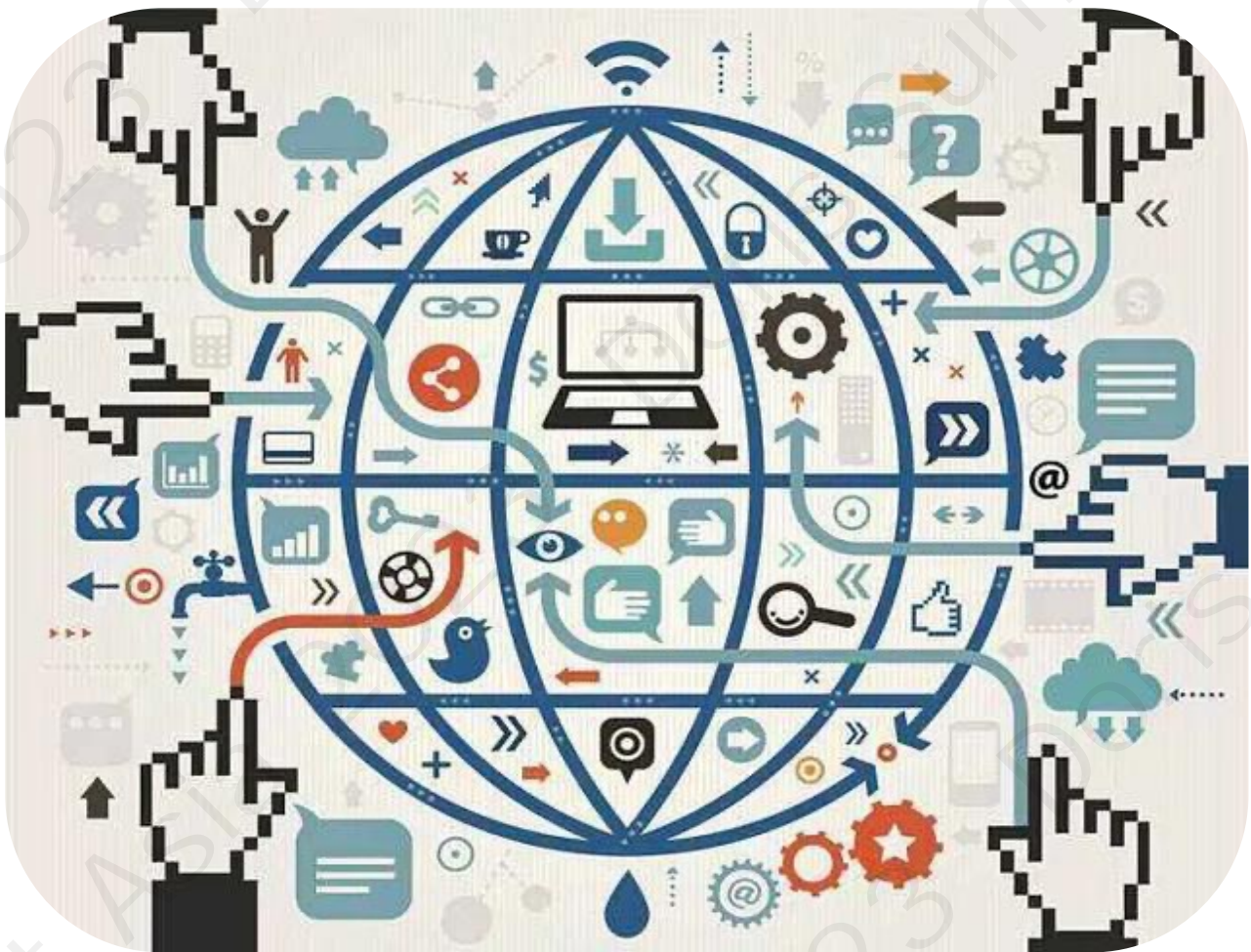
1. 众安CDP业务背景
2. 架构演进
3. 实践场景
4. 总结与展望

# 1 众安CDP业务背景



「CDP业务背景」

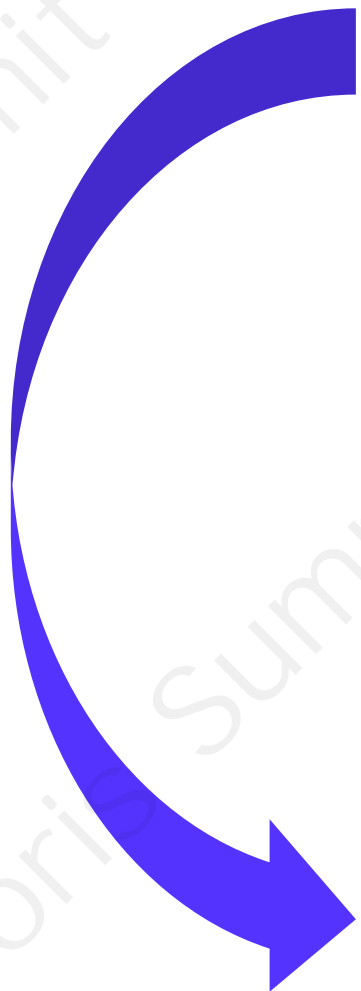
用户触点信息碎片化



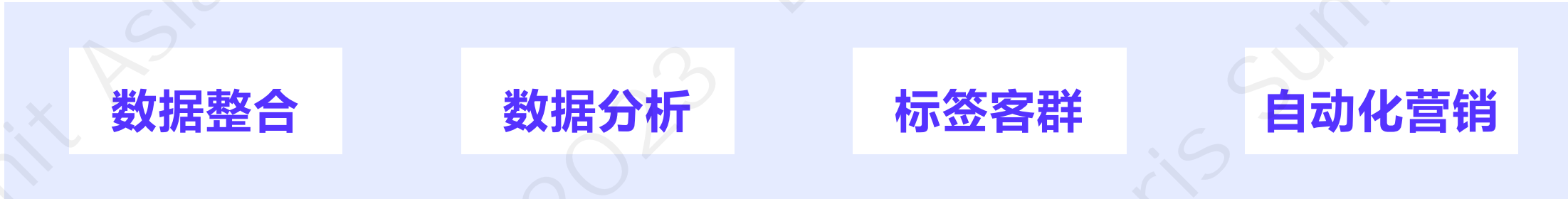
数据驱动业务增长



数据  
萃取



数据  
运营



CDP

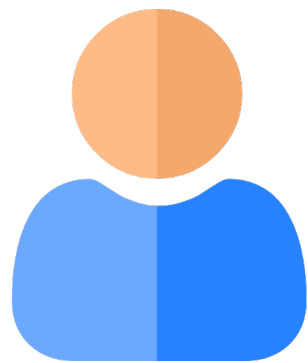


# 「CDP建设目标」



## 快速数据集成

- 打通数据孤岛，集成常见的关系型数据库 Mysql、PostgreSQL
- 集成常见的数仓如 Hive、MaxCompute
- 集成实时流数据如 Kafka



## 精准识别用户

- 在复杂的业务体系中识别多种 ID 类型
- 灵活的融合多种 ID 类型，形成统一的用户视图



## 灵活的用户标签和分群

- 统一的用户标签体系
- 强大的用户分群能力



## 多维度实时分析

- 画像分析
- 用户旅程
- 营销效果
- 赋能用户投放，提升获客ROI
- 精细化运营，提升业务转化

# 「解决方案」



✓建设离线数仓实时多渠道数据整合，使用Flink实现实时数据采集，沉淀高质量数据资产

✓通过ID-Mapping实现用户ID，用户手机，用户身份证，设备指纹,OpenID等用户身份，打通数据孤岛

✓实现用户属性，用户行为，业务交易状态，模型标签等多维度的标签的建设

✓通过规则客群的圈选能力实现客群细化

✓居于用户标签数据实现用户画像洞察

✓实时效果回收支撑营销漏斗分析

✓实现以用户标签，用户客群，用户分层，用户实时事件等多维的数据接口服务能力，赋能用户全链路智能营销

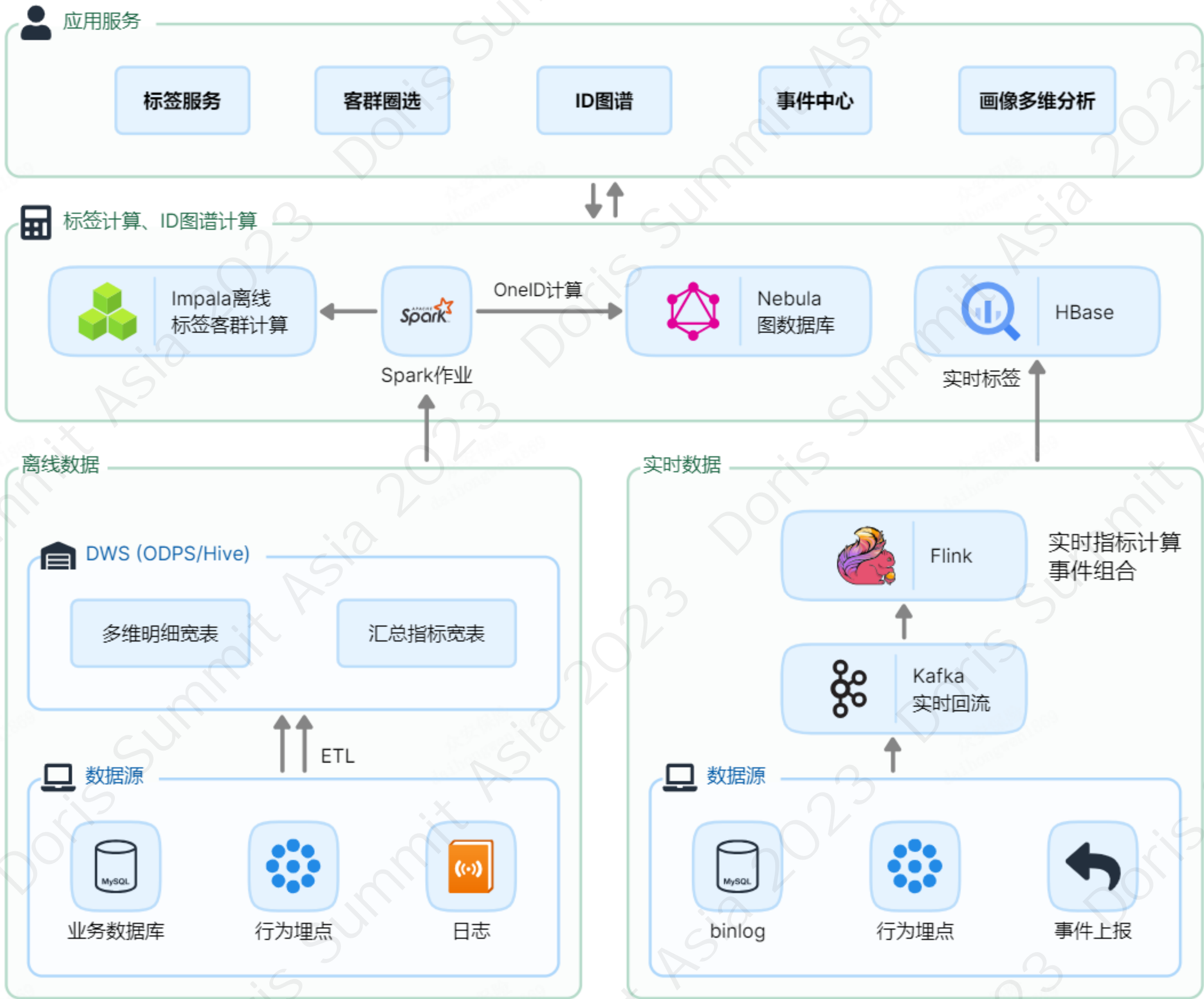
## 2 架构演进

「CDP平台架构」





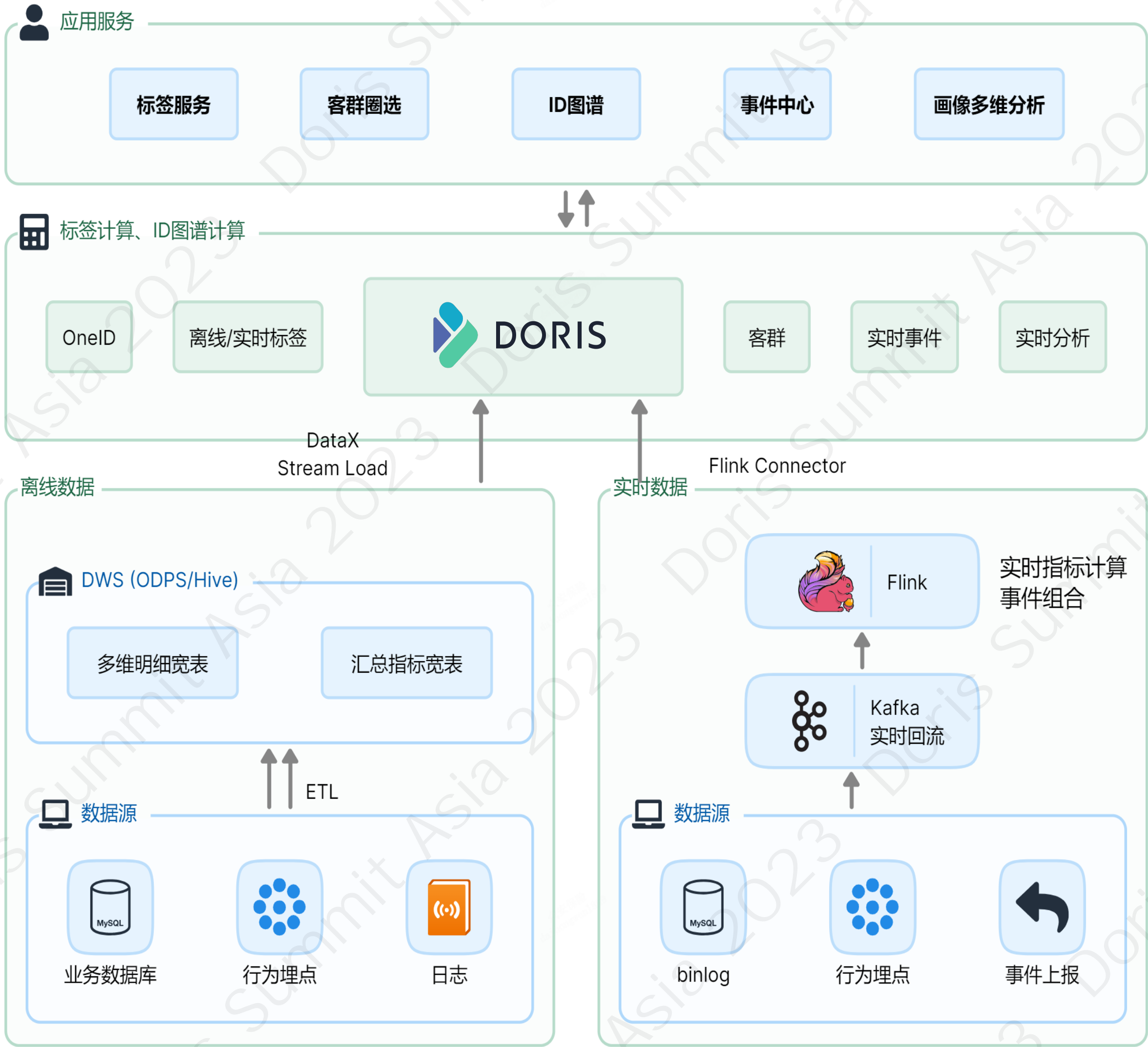
# 「CDP架构1.0」



## 场景与技术栈多样

- 数据存储不统一：离线标签与实时标签存储不统一，使用场景割裂；OneID 与标签存储不统一，数据打通需要额外存储；事件存储走离线 T+1；数据传输与存储成本高
- 技术栈复杂：离线标签客群、OneID、实时客群存储方案不一，资源与维护成本较高

# 「CDP架构2.0」



计算存储改为Doris

- 数据存储：离线标签、实时标签、OneID、事件的存储与计算统一，节约存储与计算资源，减少数据传输与耗时，提高用户体验
- 技术栈精简：从 1.0 的 Spark + Impala + Hbase + Nebula 的方案，精简为单一 Doris 的方案，极大减少维护成本



# 「CDP架构演进」

## 1.0

- 架构复杂：Spark + Impala + Hbase + Nebula
- 资源成本高：CDH + Nebula 集群
- 运维成本高：组件较多，组件间交互需要版本兼容，需要较多人力运维
- 学习成本高：涉及较多大数据组件，新人学习曲线陡峭

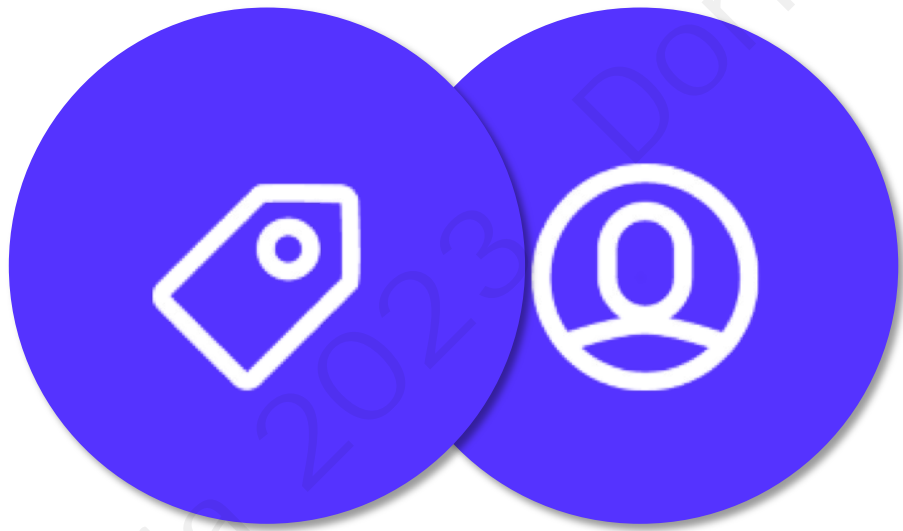
## 2.0

- 架构简单：Doris
- 资源成本降低：Doris 集群
- 运维成本低：标准 MySQL 协议、集群监控配套完善
- 学习成本低：单一组件，快速上手，数据导入导出方便

# 3 实践场景



# 「数据导入」



## DataX 离线数据集成

- 通用数据源统一接入
- 需要使用内表，计算时效要求较高场景
- 标签计算
- 客群人群预估、圈选
- 标签、客群多维分析

采用Stream Load  
多线程写入TPS 30w+



## 外表联邦查询/同步

- 实时分析报表
- 基于标签与客群，结合其他业务表数据进行实时分析或数据导入

Hive/MaxCompute Catalog  
JDBC Catalog

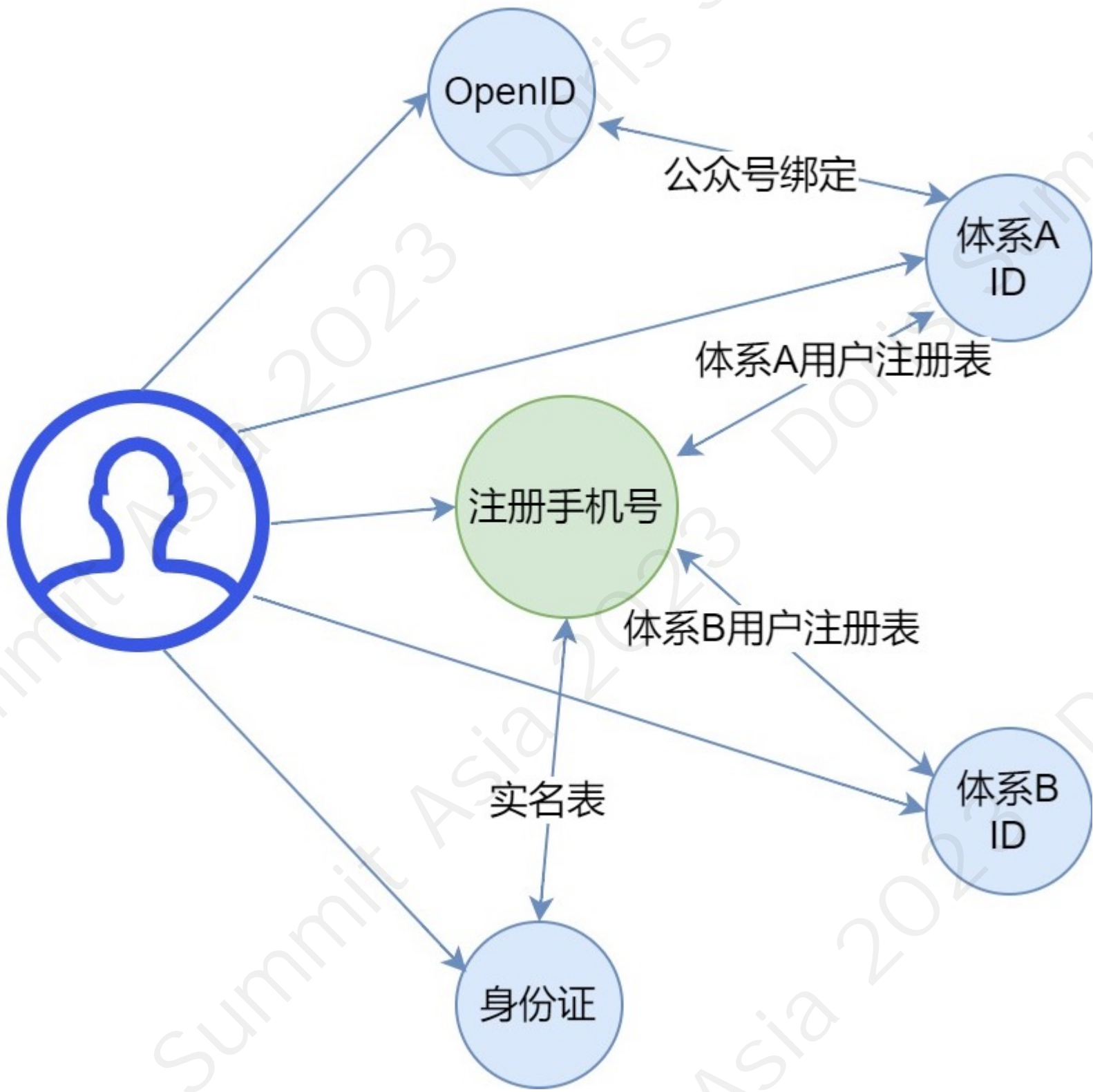


## 实时写入

- 实时事件落表
- 实时标签Flink计算后落表

Stream Load （开启部分列更新  
partial\_columns=true）

# 「数据融合-ID图谱」



- 梳理ID类型、ID关系
- 构建ID图谱

1

ID配置

+ 新增ID

身份证

OpenID

用户手机号

用户ID

2

ID关系图谱

ID信息

ID名称: 用户手机号

ID标识Code: regist\_phone

描述: 用户注册手机号

ID数据

数据来源: 离线

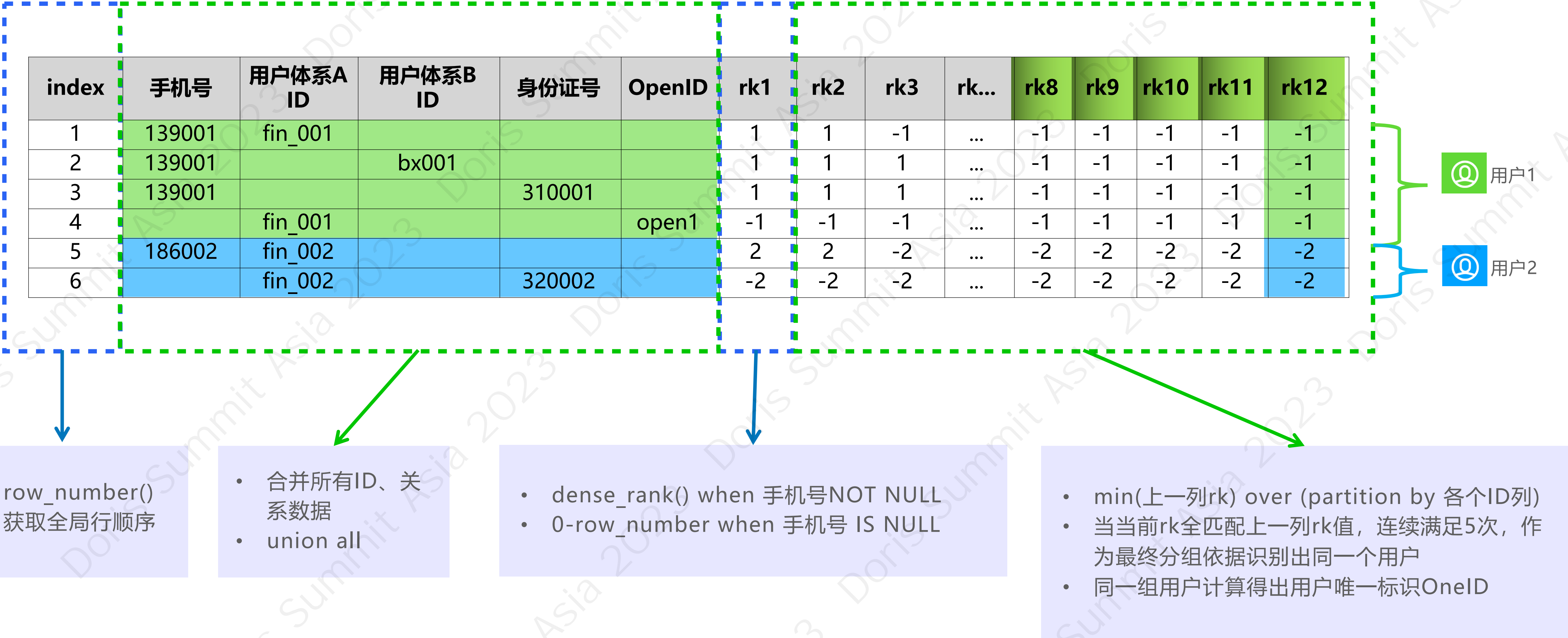
数据集: 手机号信息表

ID属性: phone\_no

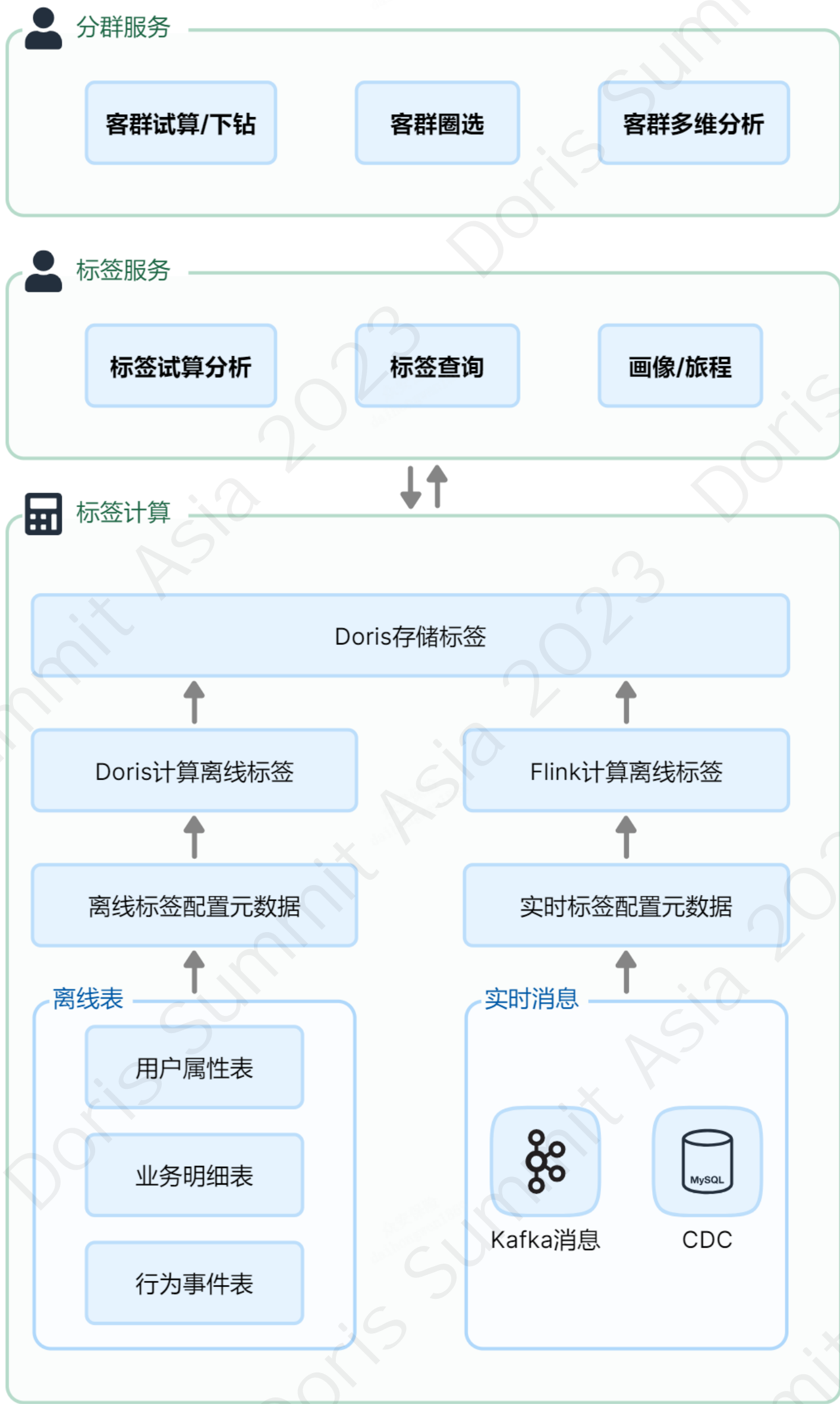
时间排序属性: N/A



# 「数据融合 - OneID 构建」



「标签体系」



▼ 标签规则 (可添加多个分类, 满足分类规则的用户标签值等于分类名称)

放款大于等于2次 ☒ 放款小于2次 放款失败 + 添加分类(1/3)

**属性**

当前是否有在贷 等于 (in) 1

最近一次支用申请时间 有值 (isNotNull)

+ 添加规则

**明细**

明细表数据 放款时间 / 总次数 \* 计算结果: 大于等于 2

放款状态 等于 (in) 1

放款时间 介于 (含两端) 前60天 至 今天

+ 筛选条件 + 添加规则

**行为**

早于 (含当... 前30天 支用事件 总次数 \* 计算结果: 大于 2

事件key 等于 (in) xxxx yyyy

+ 筛选条件 + 添加规则

- 根据标签配置元数据, 分别在Doris中计算离线标签, 在Flink中计算实时标签, 统一存储到Doris中
- 标签使用场景多样: 实时试算与分析, 标签值点查, 用户画像与旅程分析



# 「标签体系」



2000个标签



50+ 来源表



亿级用户量

## 部分列更新

```
set enable_unique_key_partial_update=true;
insert into tb_label_result(one_id, labelxx)
select one_id, label_value as labelxx
from .....
```

## 点查

使用prepareStatement  
be参数:

```
enable_unique_key_merge_on_write = true
disable_storage_row_cache = false
storage_page_cache_limit=40%
```

表参数: store\_row\_column = true  
light\_schema\_change = true

## Join优化

标签表与相关来源表统一CG（分桶列类型、数量、副本数），优先满足Colocation Join条件，本地hash join

可使用表参数"colocate\_with" = "group1" 自动创建CG中分片与副本

## Stream Load

```
curl --location-trusted -u root: -H "partial_columns:true" -H
"column_separator:;" -H "columns:id,balance,last_access_time" -T /tmp/test.csv
http://127.0.0.1:48037/api/db1/user_profile/_stream_load
```

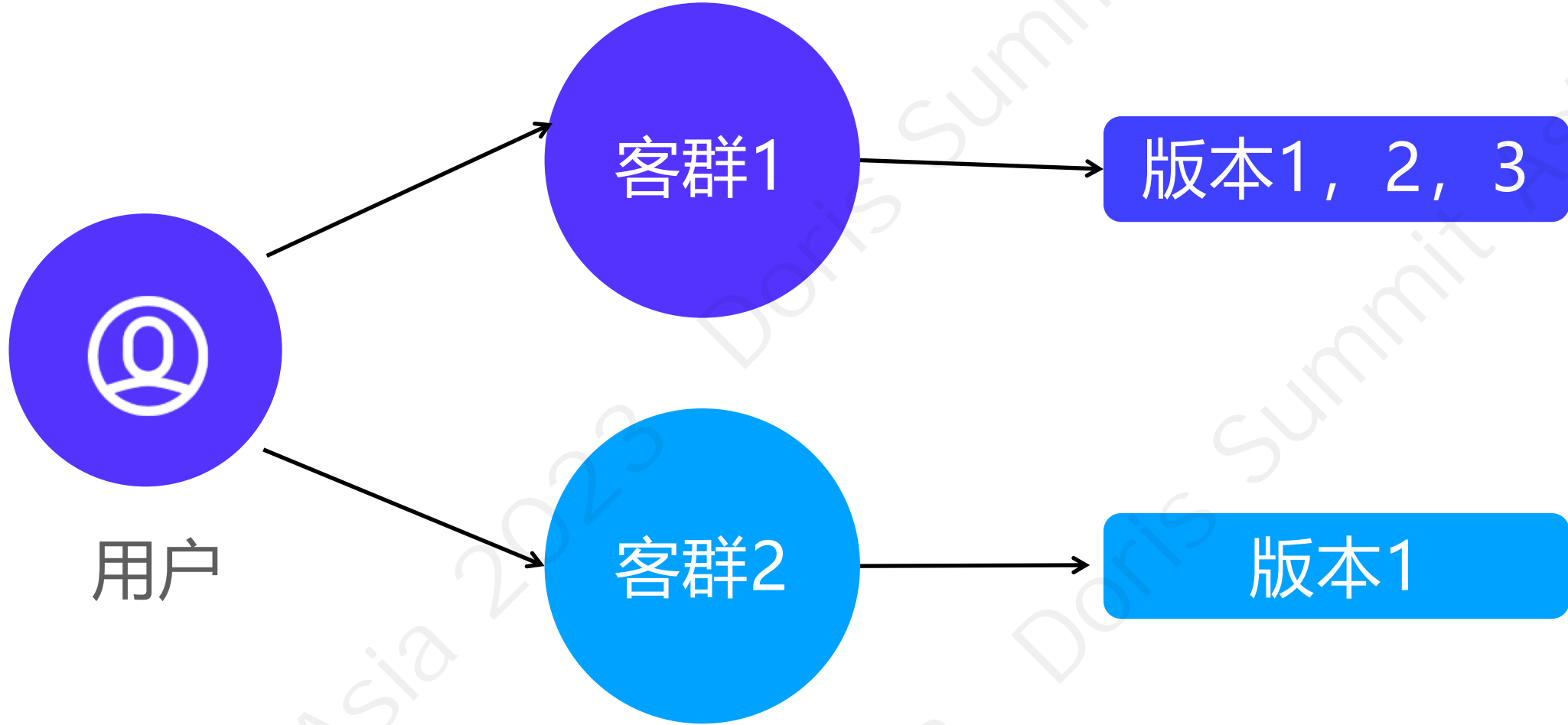
# 「客群圈选」



- 减少数据回传链路
- 由Doris将结果集直写对象存储



# 「客群归属」

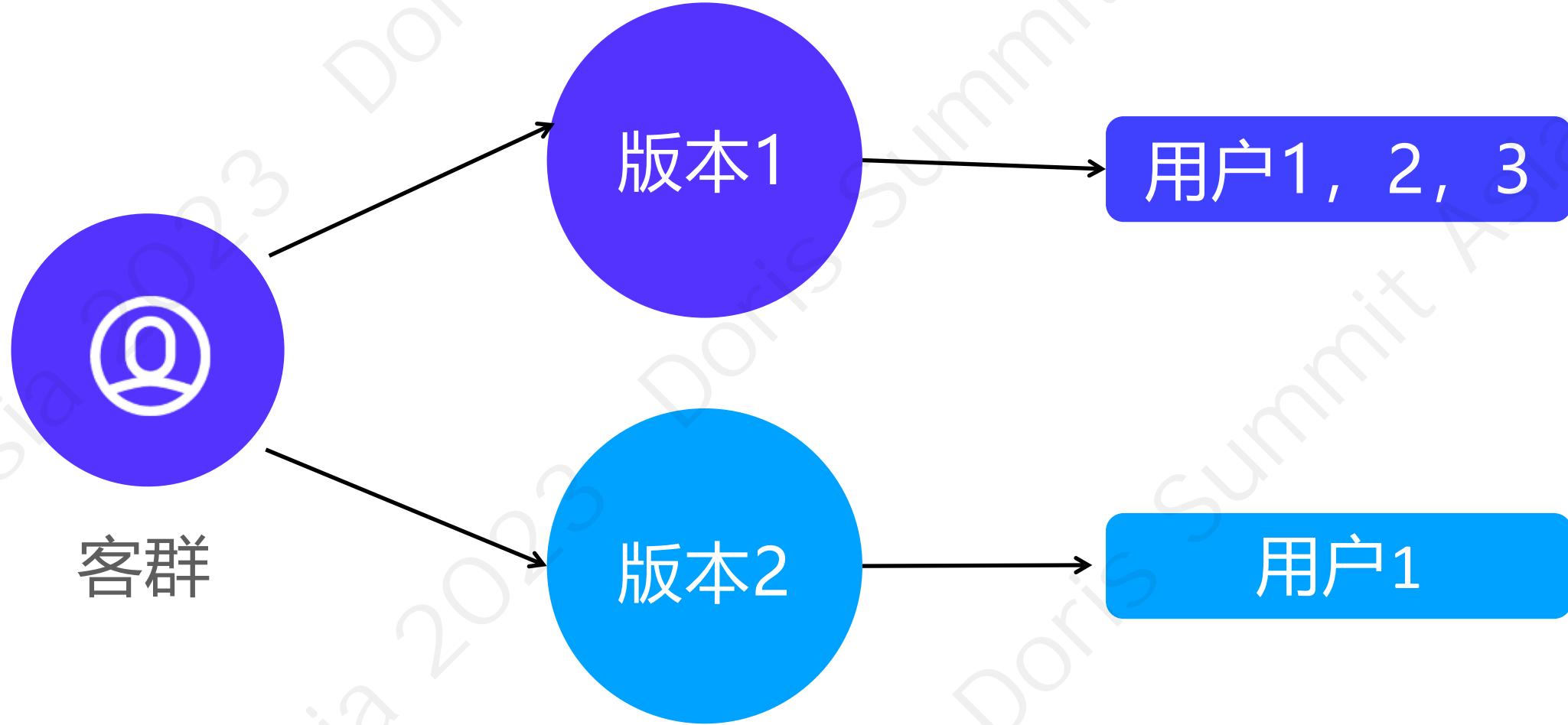


```
create table user_belong(  
  user_id varchar(100),  
  cg_version_bmp BITMAP_UNION  
) AGGREGATE key(user_id)
```

查询客群归属时，通过contains

- 查询用户所属所有客群
- 查询用户是否在某客群（版本）中

BITMAP\_CONTAINS(BITMAP bitmap, BIGINT input)



```
create table cg_belong(  
  cg_id bigint,  
  cg_version bigint,  
  user_bmp BITMAP  
) UNIQUE key(cg_id,cg_version)
```

BITMAP\_OR : 客群间并集  
BITMAP\_INTERSECT: 客群间交集  
BITMAP\_XOR: 客群间差集



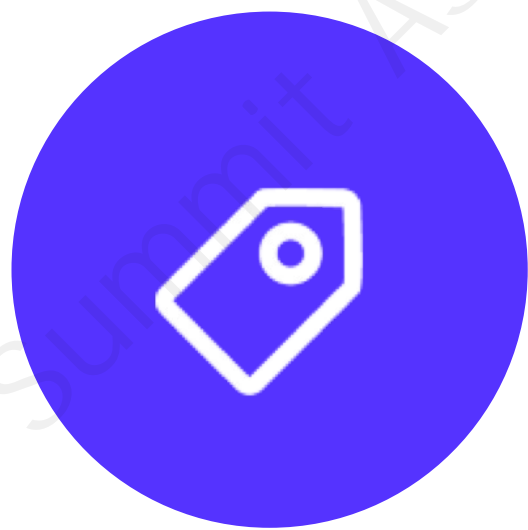
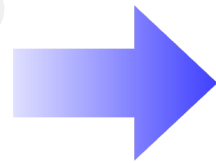
# 4 总结与展望

「总结与展望」



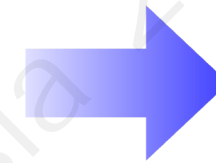
v1.0

- 架构复杂  
Spark + Impala +  
Hbase + Nebula



v2.0

- 架构改为Doris
- 标签、客群、OneID、实时  
事件存储计算使用Doris完  
成统一计算与存储



v3.0

- 基于2.0实现离线实时  
标签混合圈选
- 基于Doris计算实现实  
时OneID计算



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](https://doris.apache.org)
- Apache Doris GitHub: [github.com/apache/doris/](https://github.com/apache/doris/)

### 获取更多峰会资料:

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