快手基于指标中台的实验数据链路建设

陈硕-快手-指标平台技术负责人

DataFunSummit # 2023

快手 DataFun.

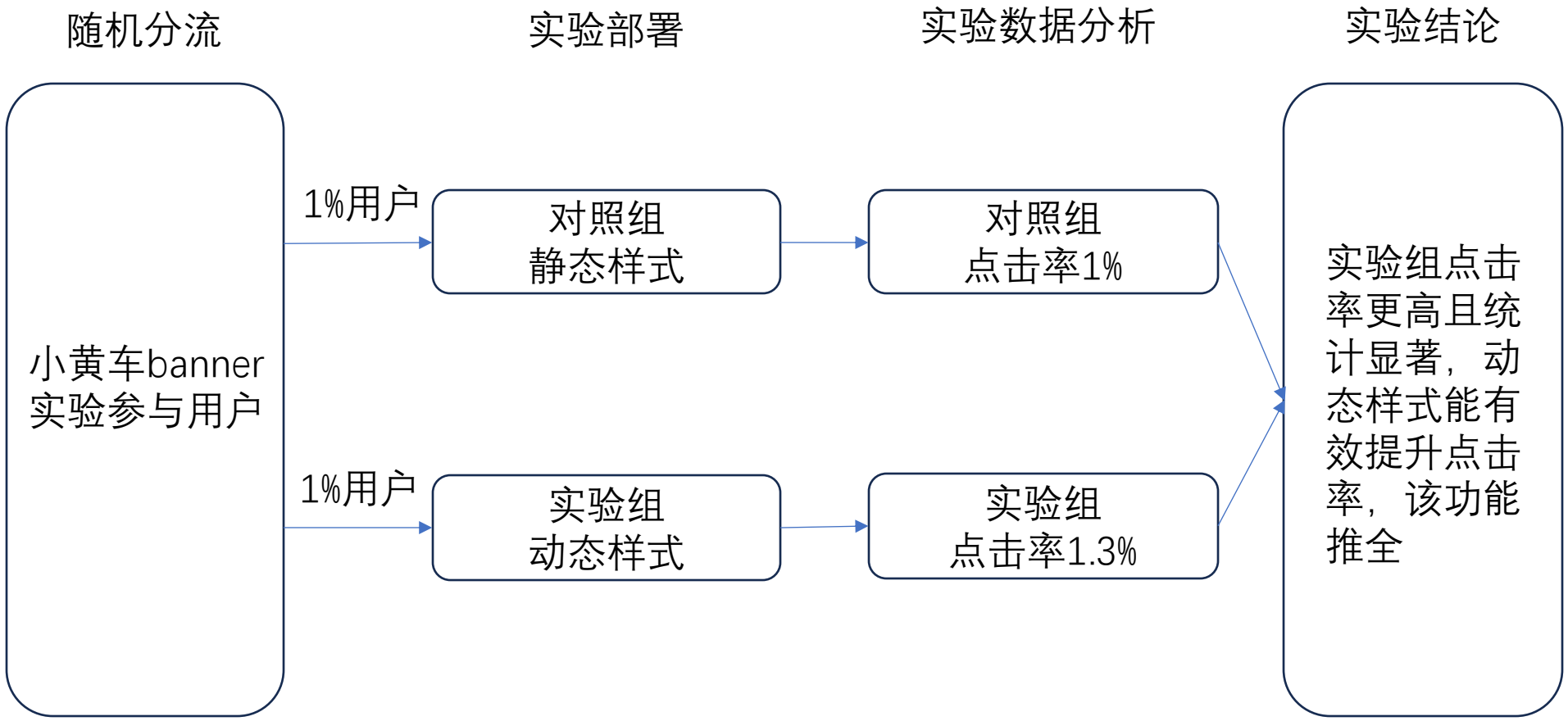
目录
CONTENT

01

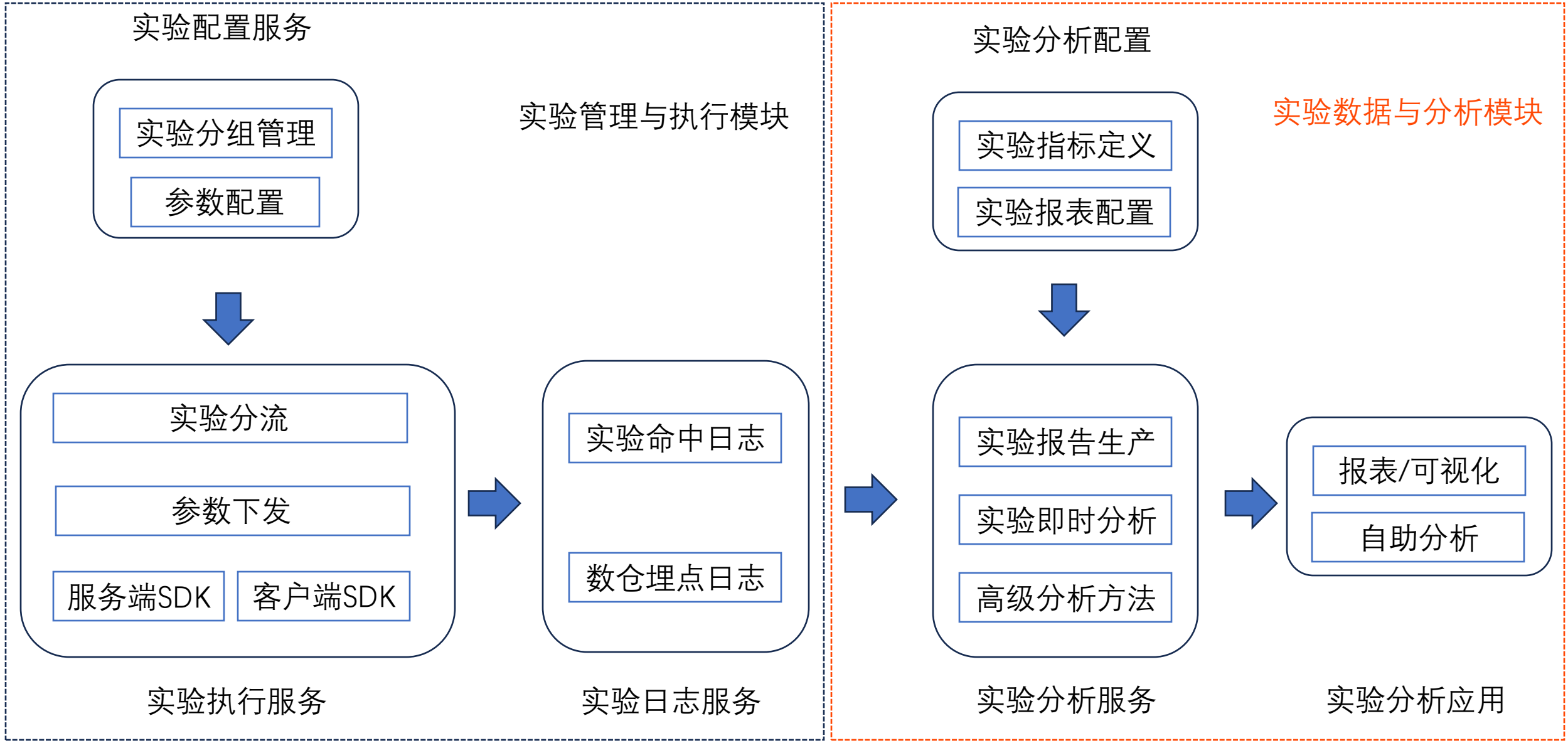
实验领域介绍

什么是AB实验？

购物车图标动态话，是否能够提升点击率？



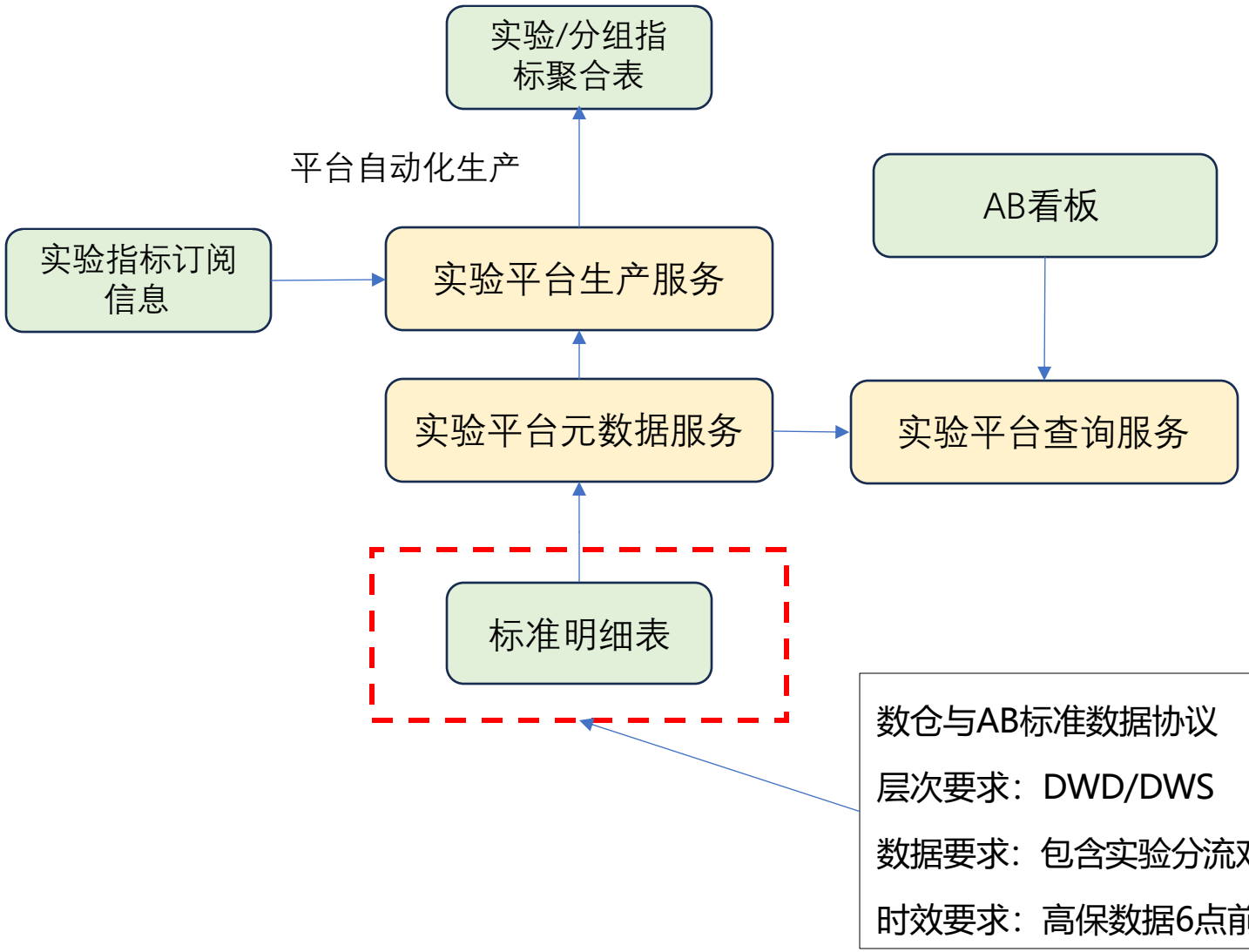
实验平台基本架构


实验配置服务
实验分组管理
参数配置
实验管理与执行模块
实验分流
参数下发
服务端SDK
客户端SDK
实验命中日志
数仓埋点日志
实验执行服务
实验日志服务
实验分析配置
实验指标定义
实验报表配置
实验数据与分析模块
实验报告生产
实验即时分析
高级分析方法
报表/可视化
自助分析
实验分析服务
实验分析应用

02
问题与解法

问题：AB实验指标的质效问题

基于指标中台重构前的架构



AB实验指标/生产/查询能力闭环建设

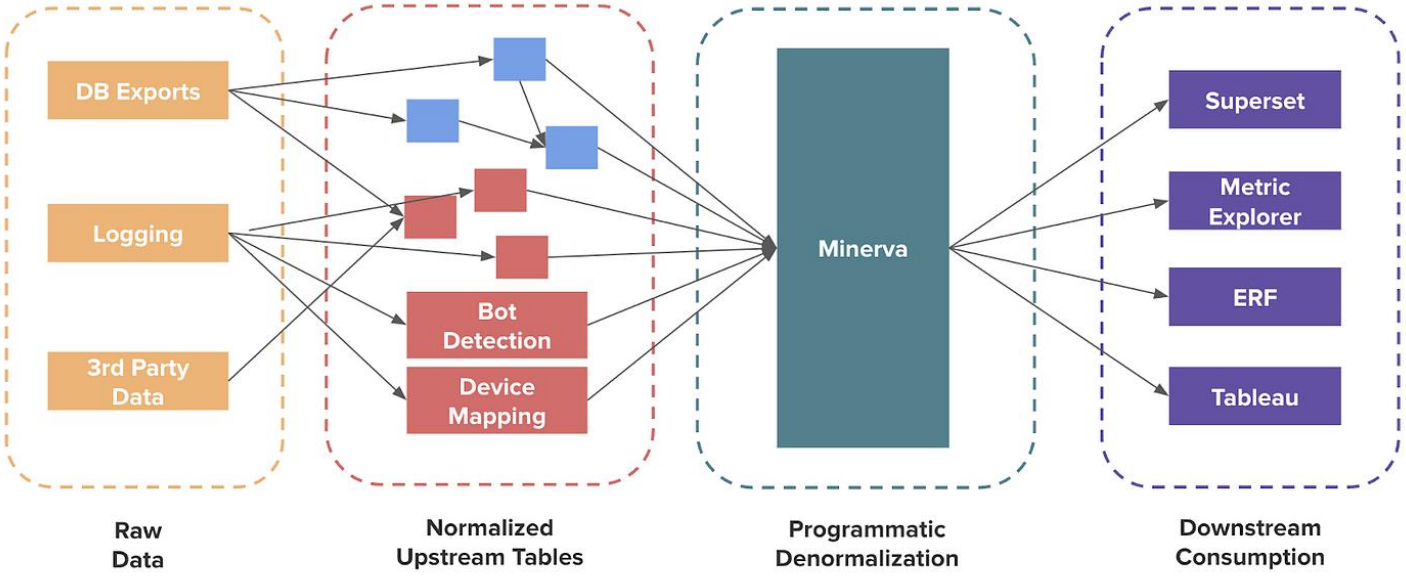
问题：

- 质量：分析师经常要将AB看板指标和BI看板指标交叉分析，但是很难保障相同指标在两个平台数据是一致的
- 效率：数据工程师需要重复在实验平台和BI平台重复开发与定义指标，浪费人力

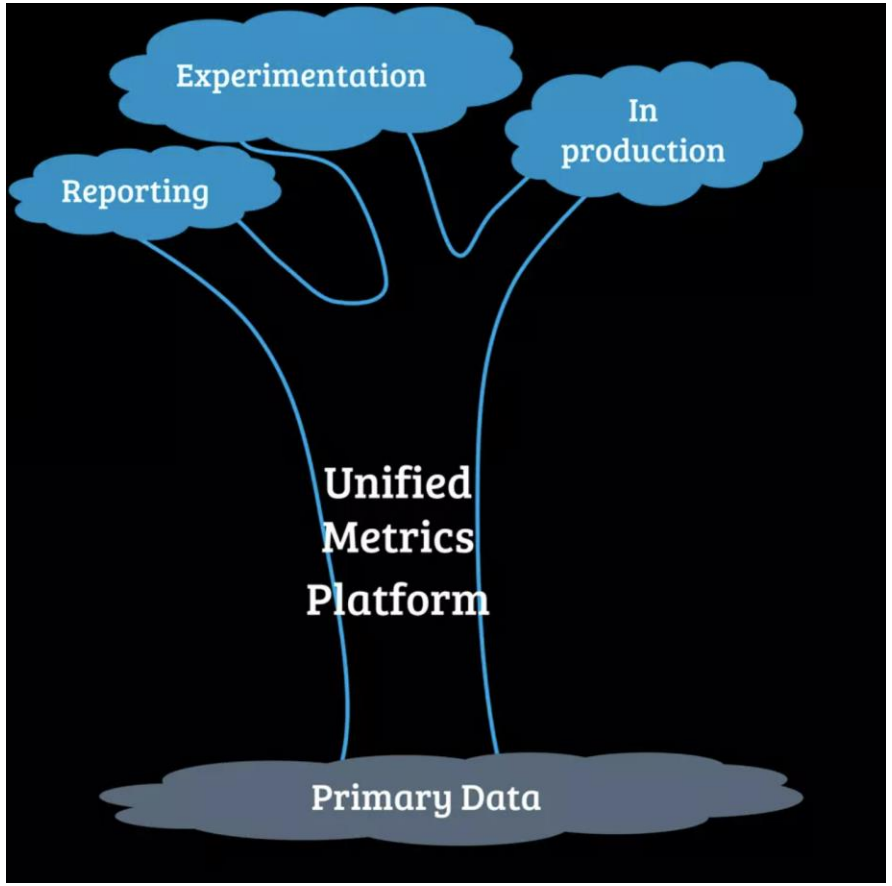
解法：基于指标中台建设实验数据服务

行业实践

Airbnb Minerva

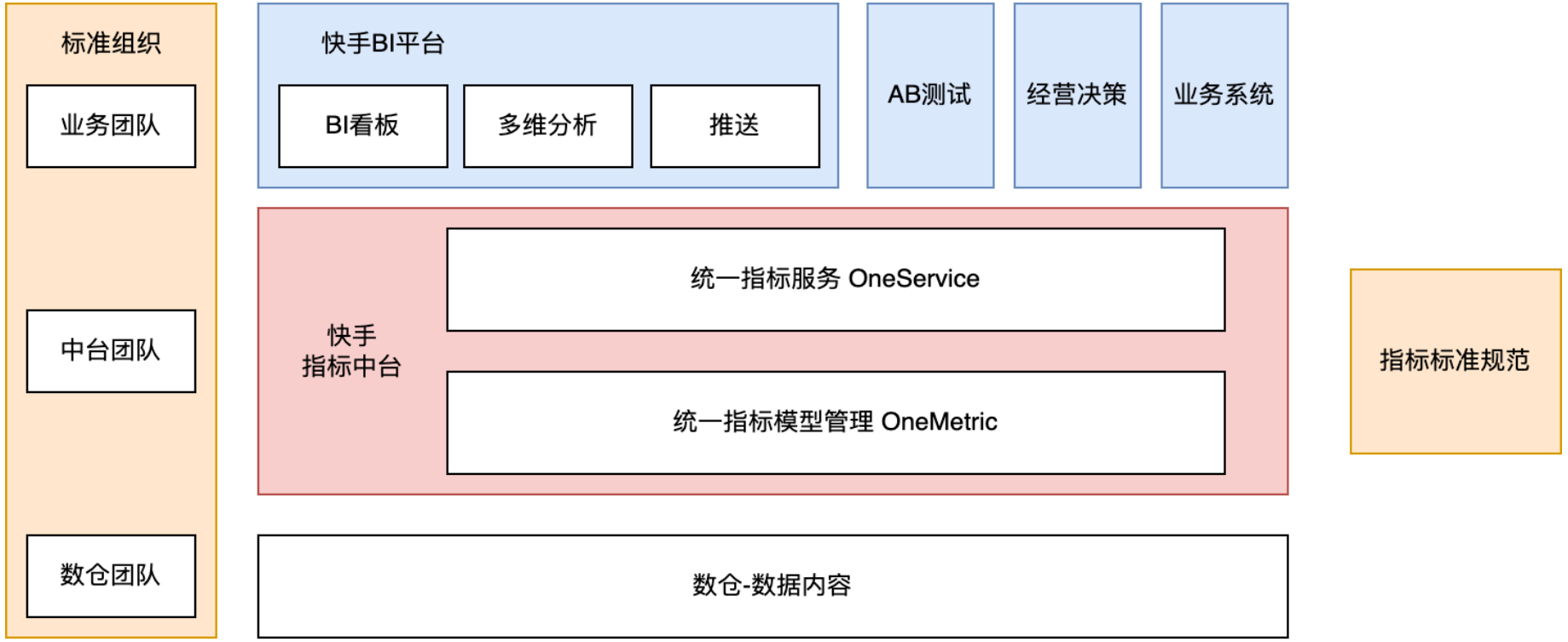


Linkedin UMP



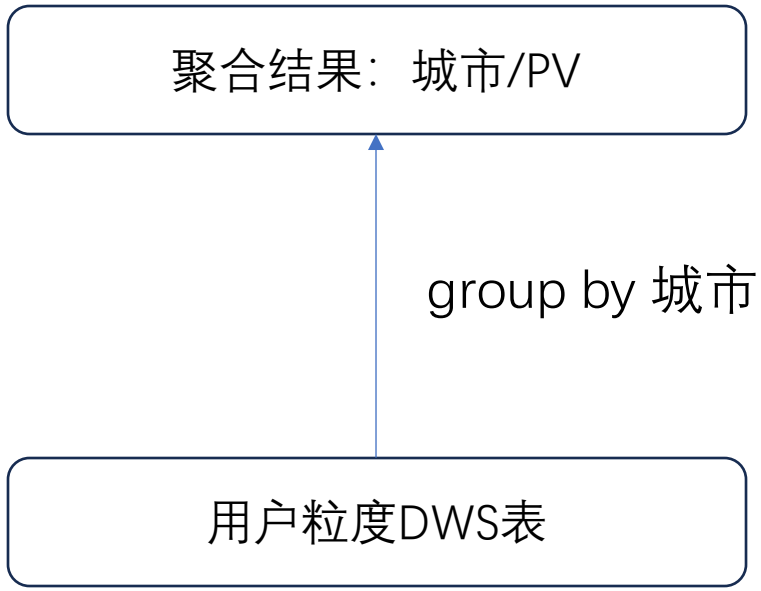
快手指标中台介绍

快手指标中台，其核心设计思想是以指标管理驱动数据服务，实现了从技术语言到管理语言的抽象，对外提供统一的指标管理以及统一的指标服务能力，进而达到“**一处定义、多处使用**”。

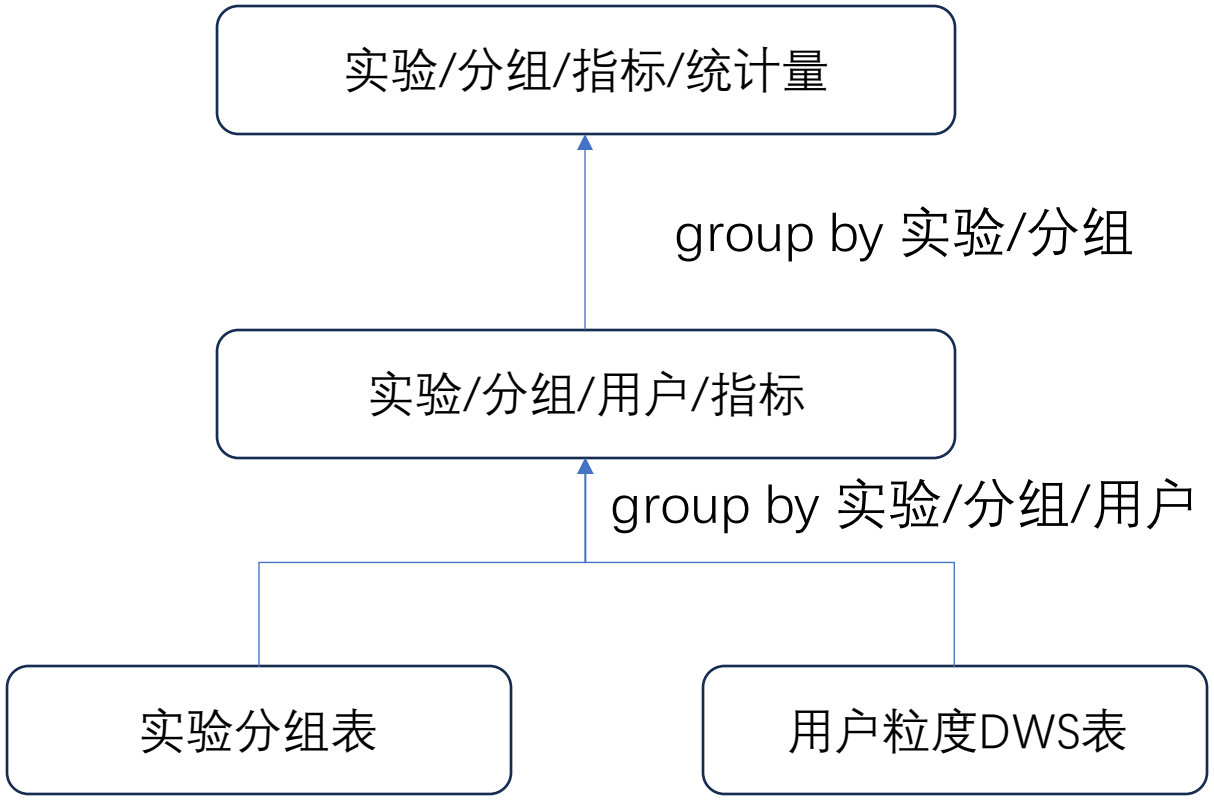


问题：指标中台在实验领域的复杂性问题

BI场景：分析不同城市的PV



AB场景：T检验分析某实验实验组与对照组PV差异



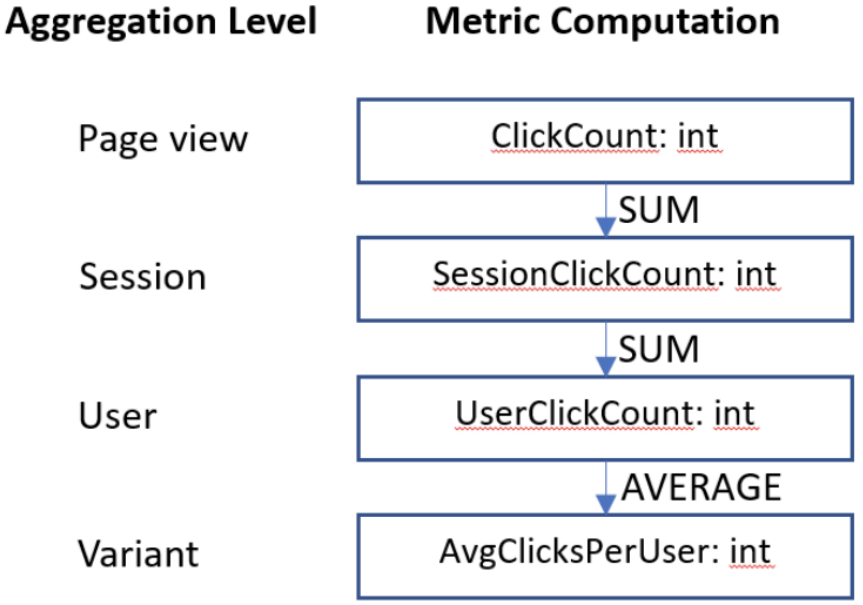
问题：

- 实验分析领域性：指标中台查询/生产服务需要能够支持实验领域的复杂逻辑

解法：支持实验领域复杂逻辑的DSL

行业实践：微软实验平台MDL语言

MDL Definition: `AVG(SUM<User>(ClickCount))`



快手实践：快手开放分析语言

(OAX, Open Analysis eXpression)

```
SELECT select_expr [, …]                                  --指标、维度
FROM [ dataset_urn | sub_query ]                          --数据集
[ WHERE where_condition ]                                 --纬度值过滤
[ CONTEXT_WHERE context_where_condition ]                 --上下文过滤
[ HAVING having_condition]                                --结果值过滤
[ GROUP BY grouping_element [, …] ]
[ ORDER BY expression [ ASC | DESC ][, …] ]
[ LIMIT offset,size ]
```

03
详细介绍

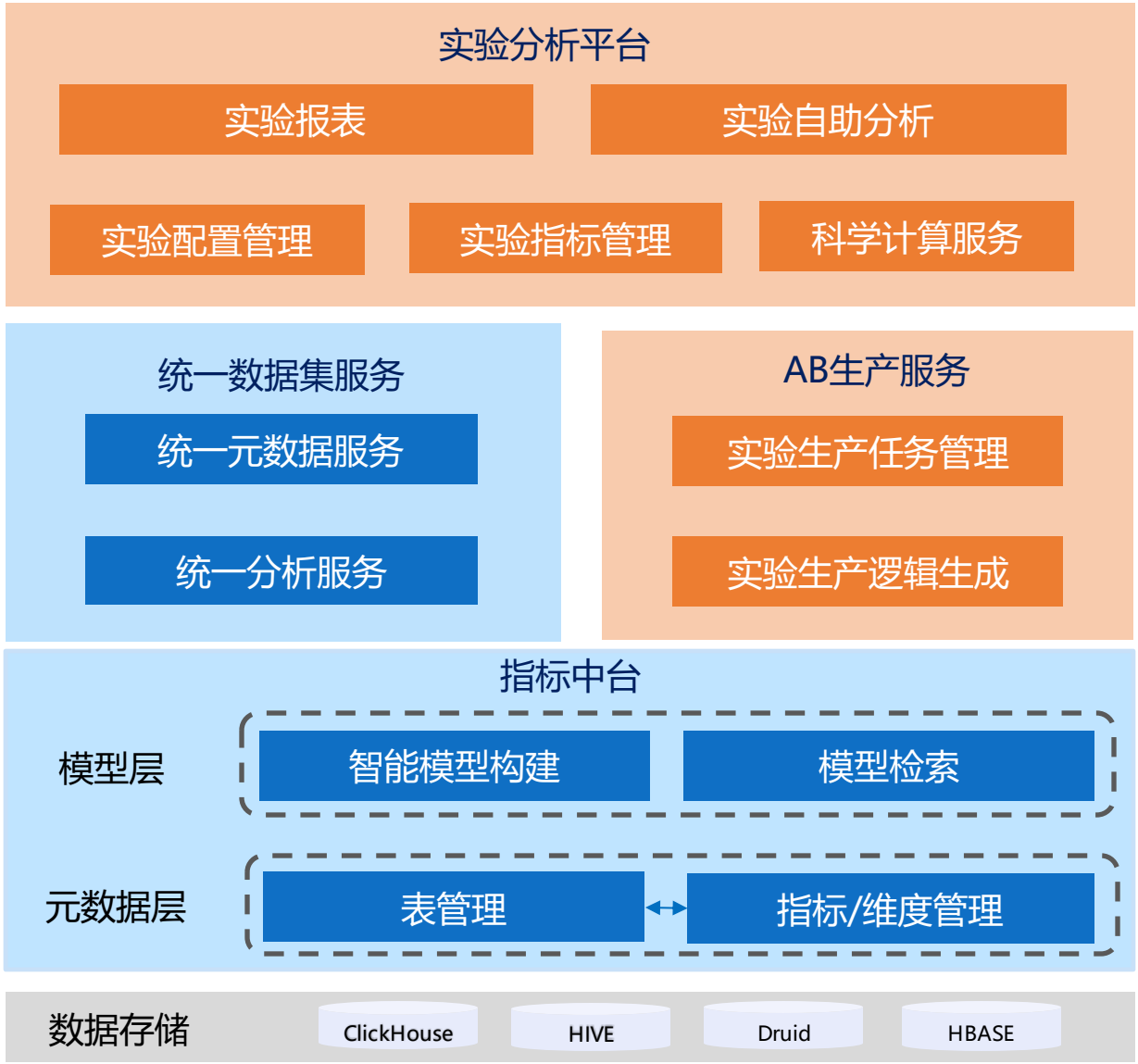
DataFunSummit # 2023

快手 DataFun.

大纲

- 总体架构
- 指标定义语言与代码生成架构
- 实验数据生产优化

基于指标中台的实验数据系统架构



2023年，重构基于指标中台的实验数据平台，并扩展实验分析能力

基于统一的指标模型元数据构建生产/分析服务

- 统一数据集服务：统一的BI与AB的指标查询能力
- AB生产服务：领域化建设的高效AB数据生产服务

基于指标中台的实验数据服务场景

分类	生产链路	分析链路
分析场景	• 固化分析思路	• 灵活分析探索
数据模式	• 数据预生产	• ADHOC即时查询
计算效率	• 效率优先，高度优化 • 保障时效 • 批量实验生产	• 尽量高效 • 即时分析

AB生产链路由于需要面向大规模实验进行高效生产，需要面向实验领域高度优化，所以比较适合领域化建设

指标定义语言 – OAX计算模型抽象

A 基本计算

允许用户在原始值或计算结果值粒度进行计算，分为：数字函数、字符串函数、日期函数、类型转换、逻辑函数、聚合函数、指标函数、高级计算

例如：SUM([消耗])，CONCAT([姓], [名]), DATE_ADD(‘day’ , NOW(), 1),
计算周同比：YOY(SUM([消耗]), 1, 'week', 'value')

B 动态粒度计算

允许计算过程中改变数据计算粒度，可以在较高粒度(EXCLUDE)、较低粒（INCLUDE)、独立粒度(FIXED)进行计算

例如计算各省的消耗占比：SUM([消耗]) / {EXCLUDE [省份]: SUM([消耗])}

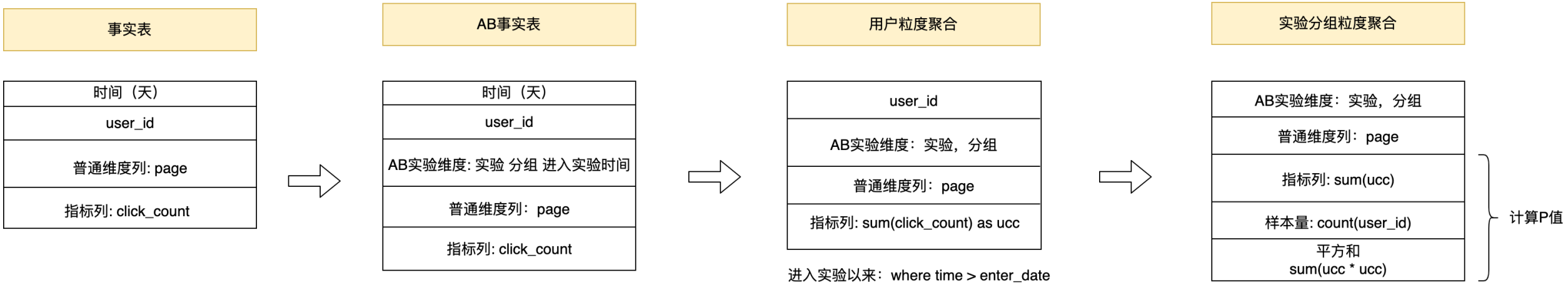
C 表计算

允许用户对结果值再进行计算，例如：滚动类函数、窗口类函数、偏移类函数等

例如要计算MTD消耗 (本月1号至本月当前日期的消耗累计)：RUNNING_SUM(SUM([消耗])) ALONG([日期])

指标定义语言在实验领域应用

t检验逻辑计算过程



引入动态粒度计算解决不同粒度聚合问题

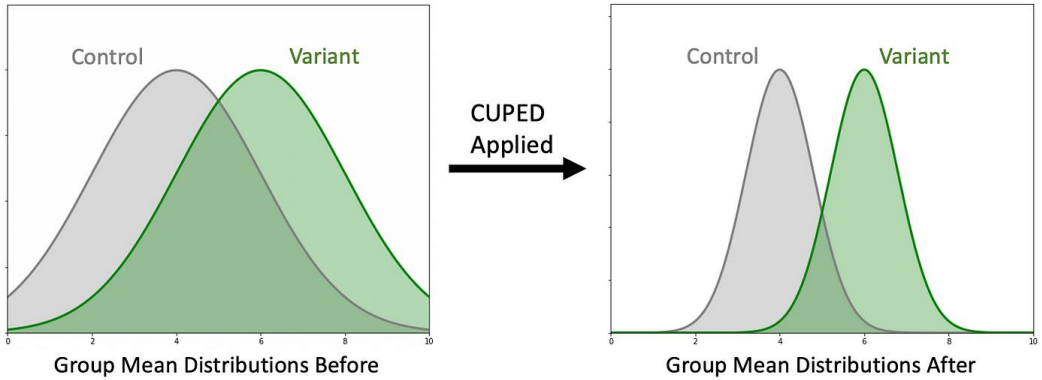
T检验在计算方差的过程中，需要计算样本指标值的平方和

先聚合计算用户粒度的指标值，再计算实验分组粒度指标值的平方和

平方和：SUM(POWER({INCLUDE [user_id]: SUM(IF time > enter_date THEN [click_count] END)}, 2))

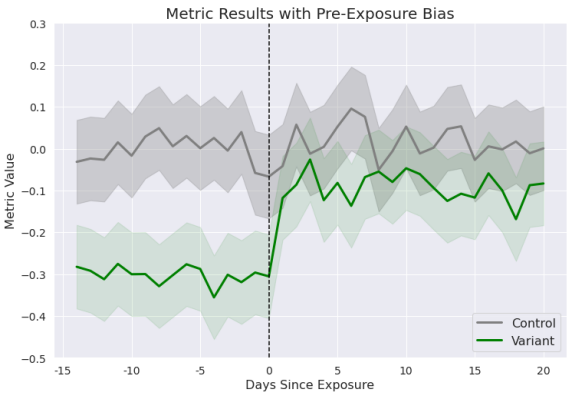
指标定义语言在实验领域应用

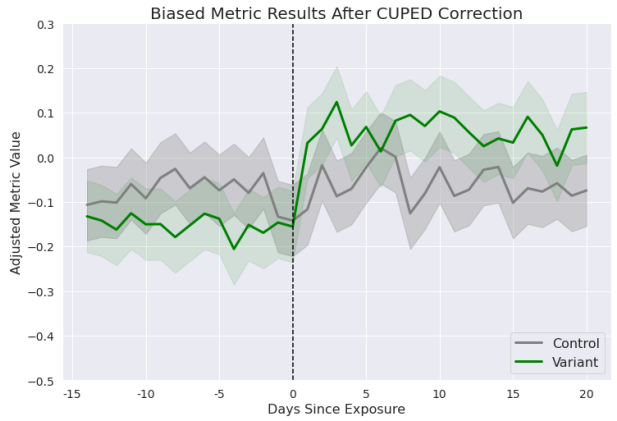
Cuped方法：一种有效的提升实验统计功效的方法



Cuped方法的核心是利用用户进入实验前的行为降低噪声，提升统计功效。

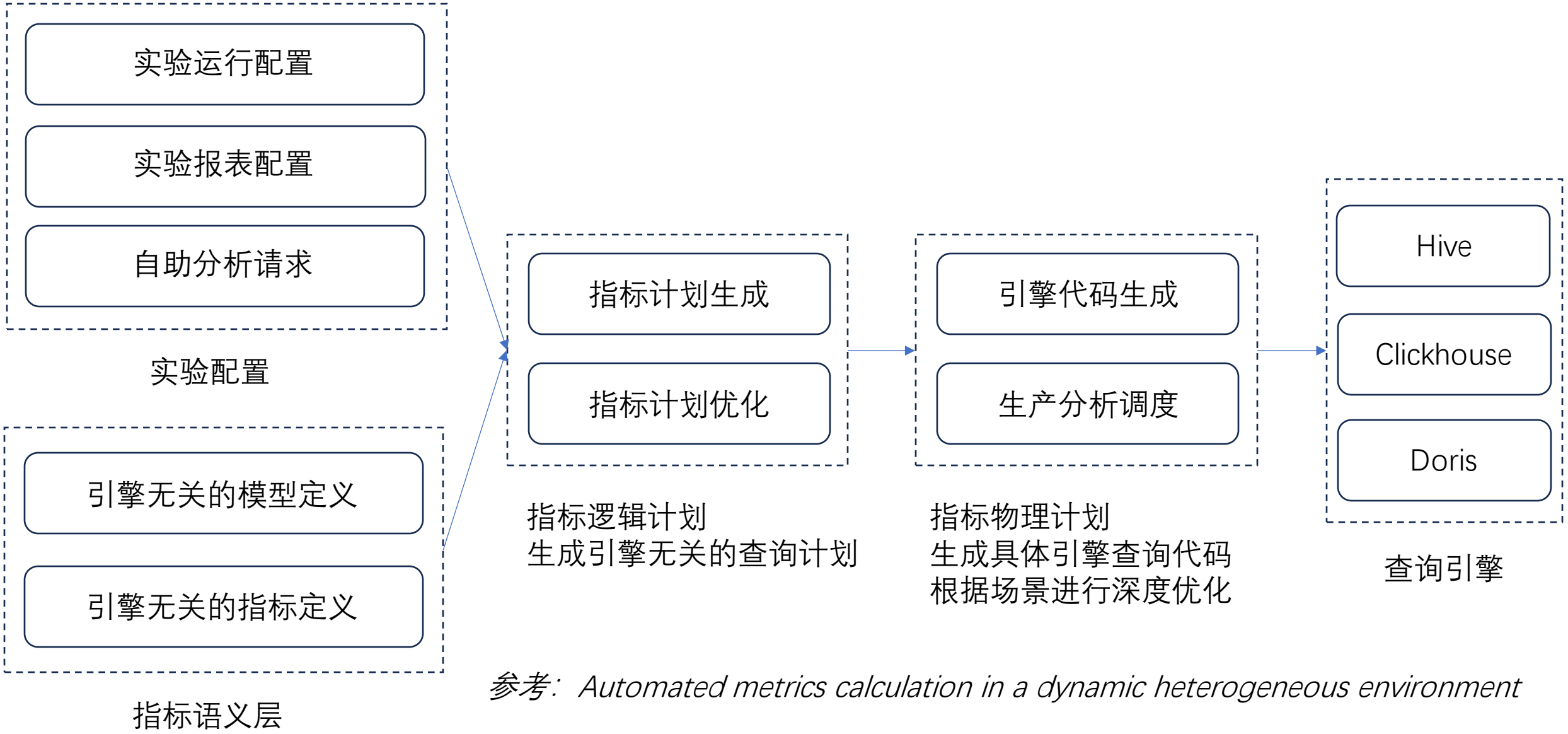
引入窗口计算与粒度计算解决
实验前样本在一段时间的指标值问题





{INCLUDE[user_id]: WINDOW_SUM(click_count, enter_date - N, enter_date)}

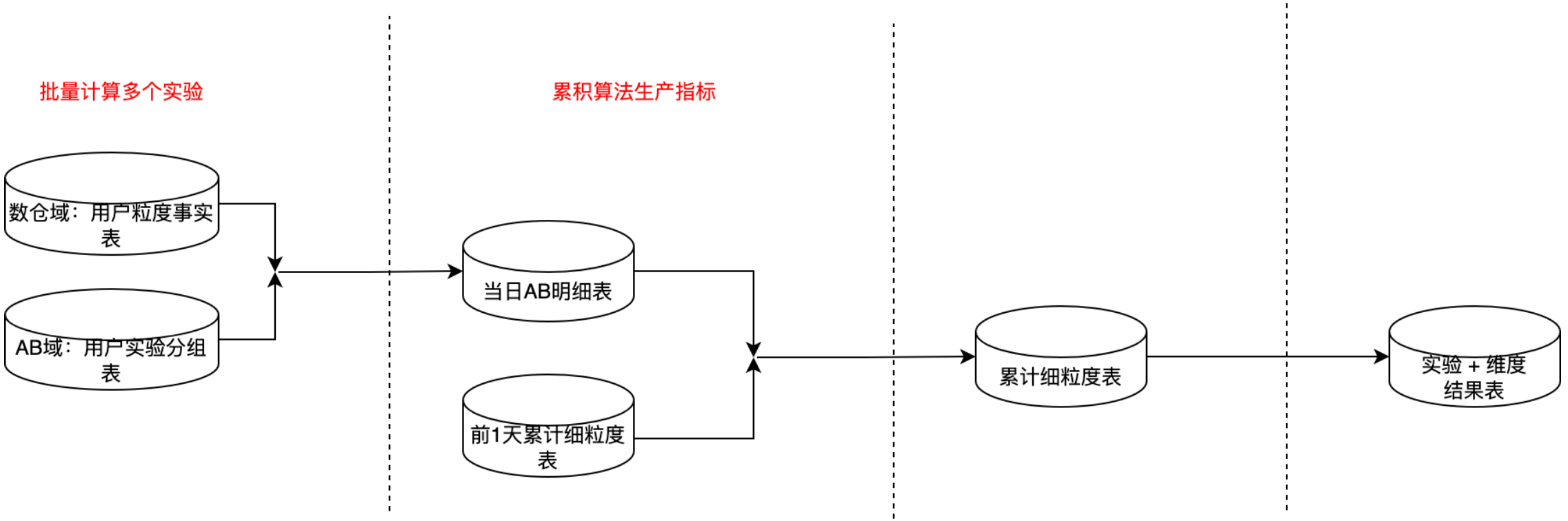
基于指标定义语言的代码生成架构



参考：Automated metrics calculation in a dynamic heterogeneous environment

实验数据生产优化 – T检验指标生产优化

面向大规模高效生产的领域化建设：T检验生产过程累积算法



第一阶段：AB明细表
计算当日各个用户在
各个实验分组明细维度和指标

第二阶段：细粒度聚合表
计算各个用户在各个实验分组
进入实验以来累计的指标

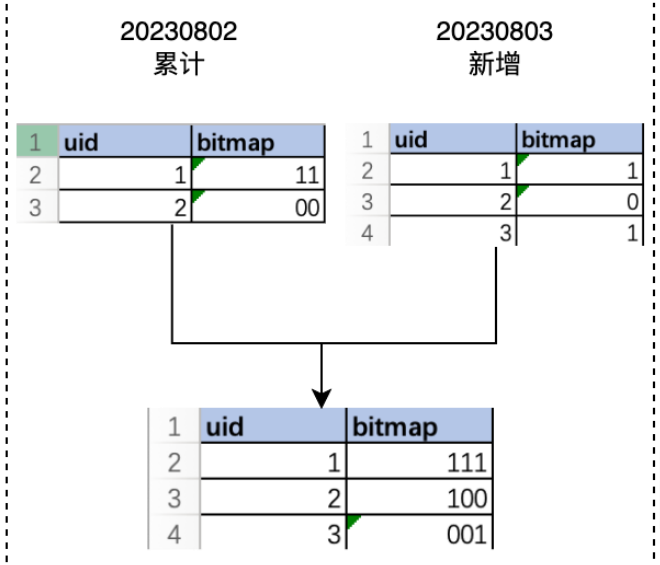
第三阶段：粗粒度聚合表
聚合各个实验分组在不同维度
下指标的聚合值

实验数据生产优化 – 去重指标生产链路优化

面向大规模高效生产的领域化建设：活跃用户数与活跃用户天数Bitmap优化

	date	uid
2	20230801	1
3	20230801	2
4	20230802	1
5	20230802	3
6	20230803	1

第一阶段：AB明细表
记录用户在不同日期的活跃情况



	uid	bitmap
2	1	11
3	2	00

	uid	bitmap
2	1	1
3	2	0
4	3	1

	uid	bitmap
2	1	111
3	2	100
4	3	001

第二阶段：细粒度聚合表
计算各个用户在各个实验分组
进入实验以来累计的指标

	uid	bitmap	活跃用户数	活跃用户天	活跃用户数平方	活跃用户天数平方
2	1	111	1	3	1	9
3	2	100	1	1	1	1
4	3	001	1	1	1	1

	活跃用户数	总活跃用户天数	活跃用户数平方和	活跃用户天数平方和
2	3	5	3	11

第三阶段：粗粒度聚合表
聚合各个实验分组在不同维度
下指标的聚合值

04
未来规划

DataFunSummit # 2023

快手 DataFun.

未来规划

指标与大模型的结合提升分析效率

高质量的指标元数据与大模型的结合

- 利用自然语言描述数据分析诉求，降低实验分析门槛
- 智能化总结实验分析结论，提升实验分析效率

基于HUDI进行生产效率提效

upsert能力带来的优势

- 优化指标累积过程样本部分更新效率
- 提前样本实验分组表的产出时效
- POC验证已完成


感谢观看
快手
DataFun.

快手实验平台和数据分析平台持续招聘
感兴趣请联系我
邮箱：chenshuo@kuaishou.com