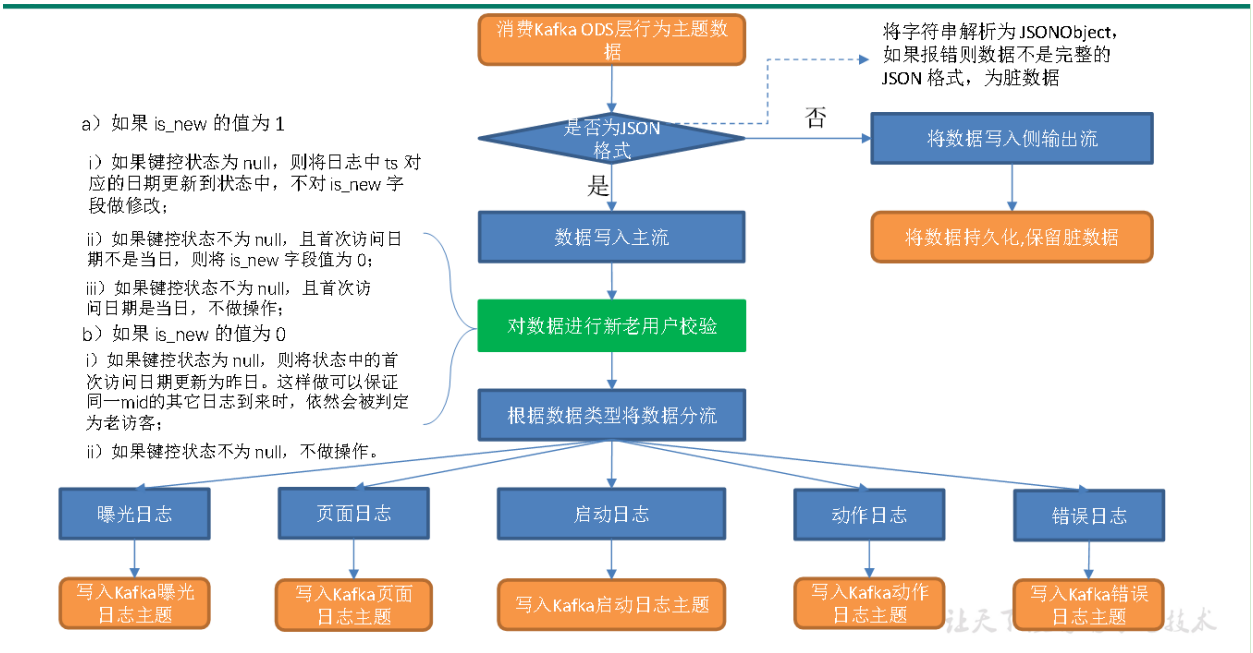
用户行为分析流程拆解

一、ODS-DWD处理流程

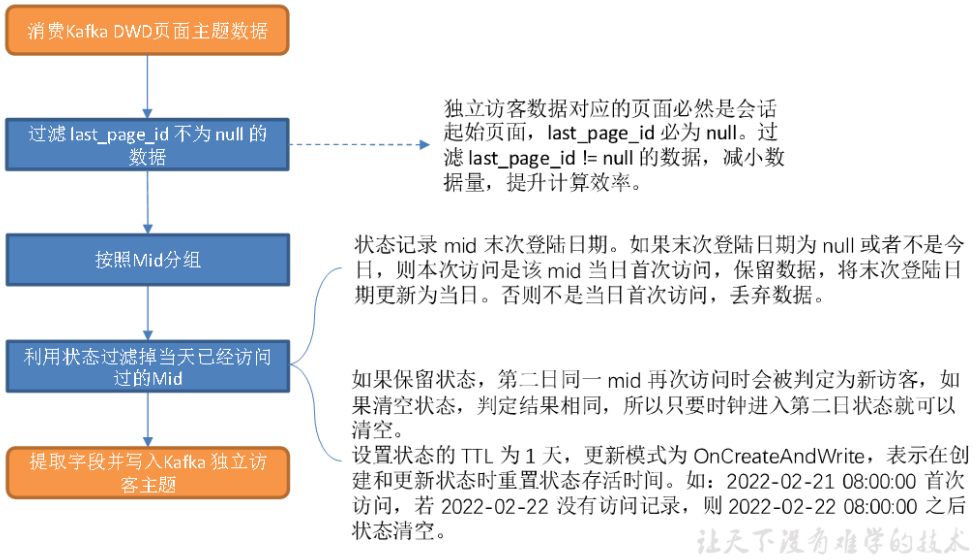
日志数据common 字段下的is_new 字段是用来标记新老访客状态的，1 表示新访客，0 表示老访客。前端埋点采集到的数据可靠性无法保证，可能会出现老访客被标记为新访客的问题，因此需要对该标记进行修复。



二、DWD-DWS处理流程

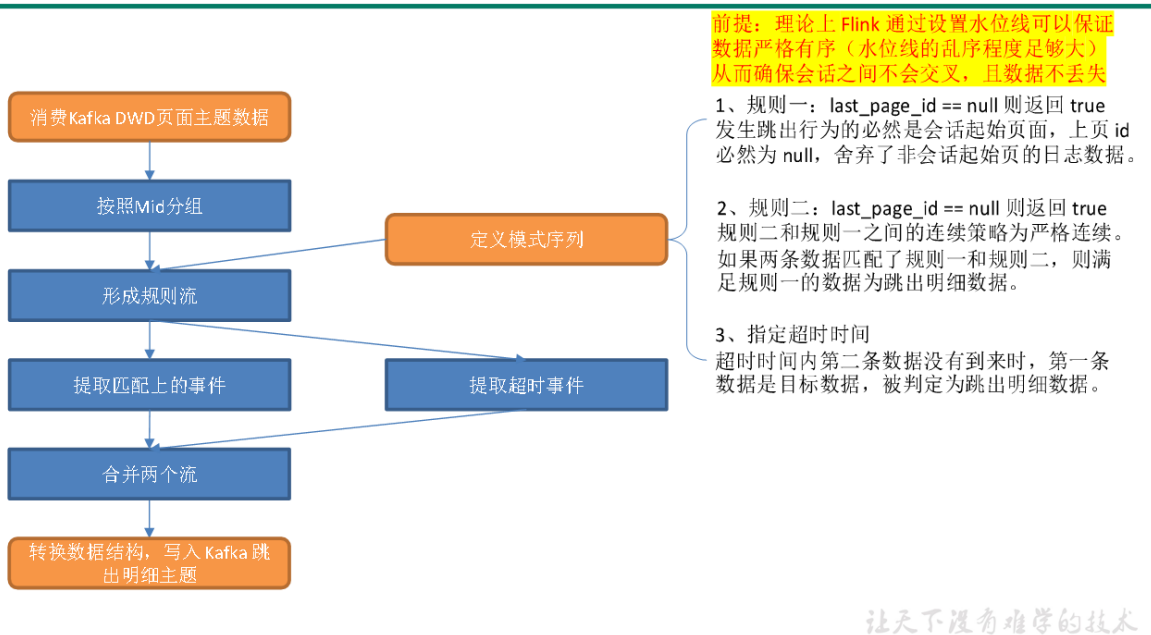
流量域独立访客事务事实表

也叫独立访客明细表，做一个UV，为了未来的UV需求做准备。



流量域用户跳出事务事实表

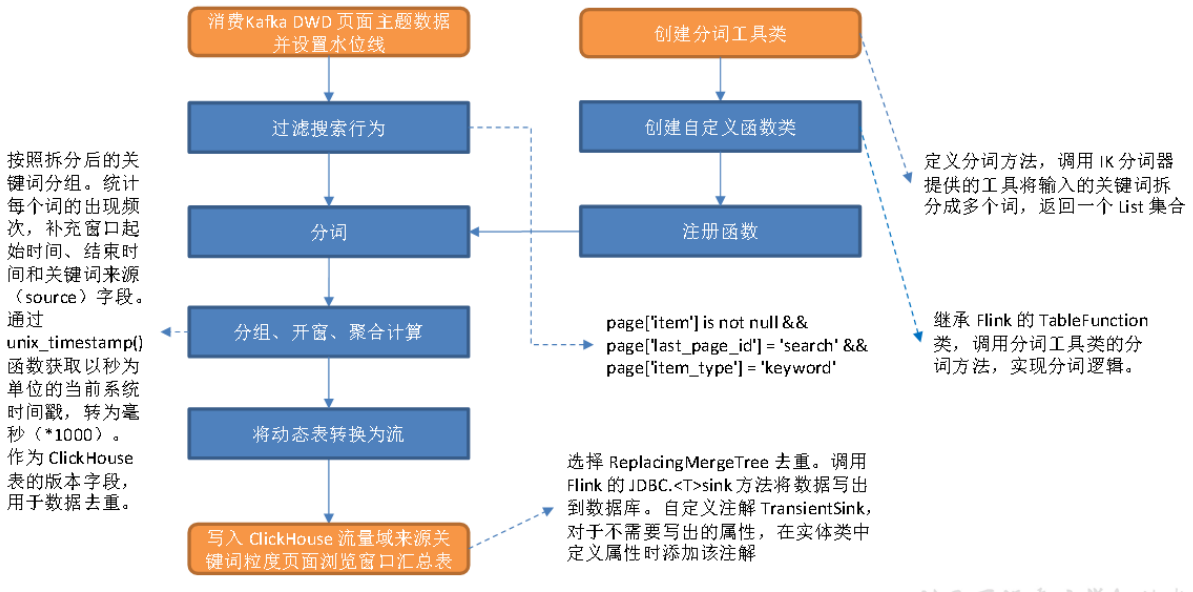
过滤用户跳出明细数据。跳出是指会话中只有一个页面的访问行为，如果能获取会话的所有页面，只要筛选页面数为1 的会话即可获取跳出明细数据。



三、DWS-ADS处理流程

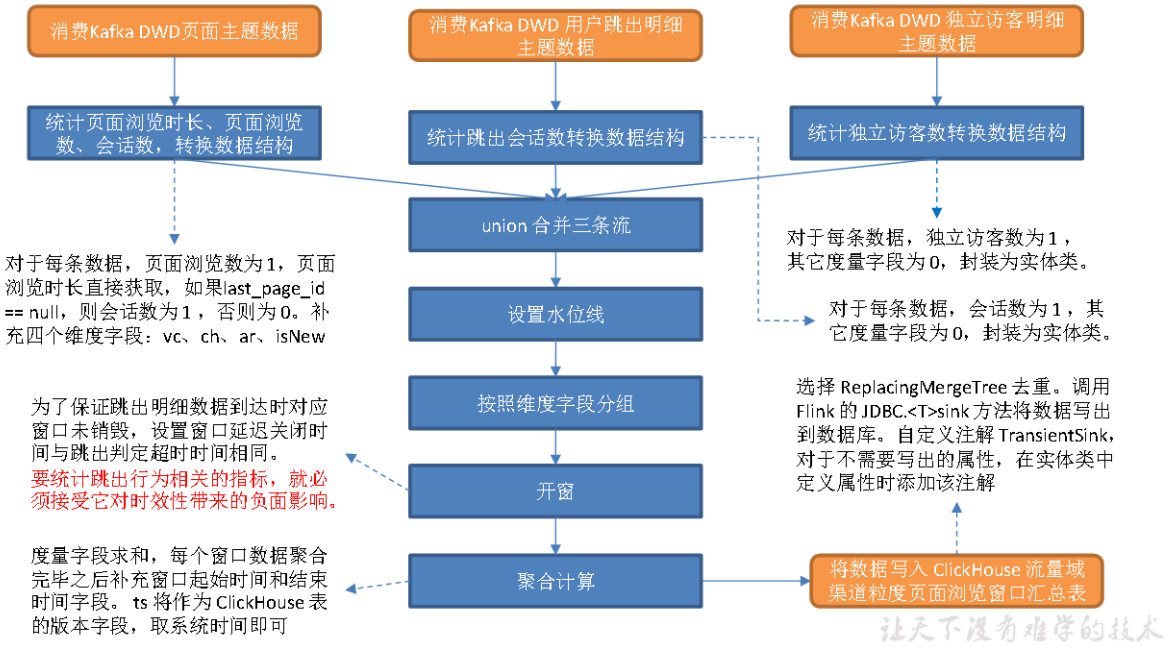
流量域来源关键词粒度页面浏览各窗口汇总表

页面浏览明细主题读取数据，过滤搜索行为，使用自定义UDTF（一进多出）函数对搜索内容分词。统计各窗口各关键词出现频次，（查看用户搜索的频次哪个最高）写入ClickHouse。



渠道-地区-访客类别粒度页面浏览各窗口汇总表

汇总表中需要有会话数、页面浏览数、浏览总时长、独立访客数、跳出会话数五个度量字段。本节的任务是统计这五个指标，并将维度和度量数据写入ClickHouse 汇总表。



流量域页面浏览各窗口汇总表

从Kafka 页面日志主题读取数据，统计当日的首页和商品详情页独立访客数。