

毕业设计开题报告

|  |  |
| --- | --- |
| **姓 名：** | 李保媛 |
| **学 号：** | 12422706026 |
| **学 院：** | 信息工程学院 |
| **专业班级：** | 大数据（专升本）2403 |
| **指导教师：** |  |

2025年9月3日

|  |  |
| --- | --- |
| **论文题目：** | 基于RFM模型的视频平台用户价值分群系统设计 |
| **一、选题背景与意义**  （一）国内外研究现状1、国外研究现状   * 大数据与实时分析：Netflix、YouTube 等流媒体平台广泛采用基于数据湖与实时分析的架构（如 Kafka + Flink + Druid/ClickHouse/BigQuery），支持用户行为实时埋点、千人千面推荐与 QoE 保障。国外在高并发低时延分析、A/B 测试与全链路指标可观测性方面较为成熟。 * 近实时 OLAP 技术：Apache Druid、ClickHouse、BigQuery、Snowflake 等已在海量日志与事件序列分析中验证其高吞吐与低延迟能力；Lakehouse（如 Delta Lake、Iceberg）与向量化执行、列式存储成为主流。 * 数据驱动运营与可视化：借助 Looker、Tableau、Superset 等，形成从埋点、指标体系、可视化到闭环优化的标准范式，强调指标治理与一致性计算。   2、国内研究现状   * 互联网视频平台（如爱奇艺、优酷、B站等）已形成完善的用户行为采集与增长分析体系，典型架构为 Kafka/Flink 实时计算 + OLAP 引擎（如 Doris、ClickHouse）+ 可视化（Superset、DataV、Grafana）。 * 开源 OLAP 引擎国产化：Apache Doris 在高并发、亚秒级查询、物化视图与 Rollup 聚合方面表现突出，适合行为明细与宽表分析；与 Flink、Kafka 的生态集成完善。 * 痛点与趋势：多源数据融合难、指标口径不一致、时效性与成本平衡、可视化交互从“报表式”向“探索式/自助式 BI”演进；隐私合规与数据安全（脱敏、最小化访问）成为刚需。   （二）选题目的及意义   * 目的：构建基于 Apache Doris 的视频用户行为分析与可视化系统，实现对播放、停留、跳转、搜索、互动（点赞/弹幕/评论）等核心行为的批流一体统计、近实时指标计算与主题看板展示，支撑运营增长、内容优化与产品决策。 * 意义：   1) 学术价值：验证 Doris 在视频行为明细分析中的可行性与性能边界，形成一套可复用的数据建模与指标治理方法。2) 应用价值：为中小视频平台提供低成本、高性能的实时用户行为分析方案，提升运营效率与数据驱动能力。3) 工程价值：沉淀端到端实践（埋点规范→数据接入→数据建模→物化视图→可视化→闭环运营），可复制、易运维。 | |
| 1. **主要研究内容业务域与指标体系设计：围绕拉新、留存、活跃、转化与内容消费，定义 UV、VV、完播率、播放时长、跳出率、回访率、转化漏斗、搜索转化、互动率、用户生命周期分层等指标与口径。** 2. **数据采集与接入：制定埋点规范（必填字段、时间、设备、会话、内容 ID、曝光/点击/播放/互动事件），通过 SDK→Kafka→Flink/Doris Stream Load 接入。** 3. **数据建模：采用明细事实表 + 维度表（用户、内容、设备、渠道）+ 物化聚合宽表；主题域包括播放分析、留存与回访、内容效果、渠道投放、互动与社区行为。** 4. **Doris 侧优化：键模型选择、分区分桶策略、列式压缩、预聚合、物化视图、Rollup、冷热分层、Bitmap/HLL 去重、查询路由与并发优化。** 5. **近实时与批流一体：Flink 实时计算 + Doris 明细入库与预聚合，离线补数据与口径对齐，构建数据质量校验与血缘追踪。** 6. **可视化实现：基于 Apache Superset 或自研前端，构建多主题可视化看板与自助分析能力，支持多维钻取与联动。** 7. **数据治理与安全：指标口径管理、元数据与血缘、权限与脱敏、合规审计与访问追踪。** 8. **性能与成本评测：对比不同分区/物化策略与并发下的时延与成本，给出优化策略与参数范式** | |
| * 1. **研究方法与手段** * 文献与对比研究：调研国内外流媒体数据分析架构与 OLAP 引擎对比，提炼适配视频行为的建模与查询模式。 * 原型实现与实验：构建最小可行系统（MVP），在不同数据规模（亿级明细）下评测 Doris 的导入、查询、并发性能。 * 数据建模方法：维度建模（星型/雪花）、宽表与明细并存、指标原子化与衍生分层、物化视图驱动的加速层。 * 系统工程方法：采用可观测性与数据质量治理（单调性检查、对账、采样核验）、自动化运维（冷热分层、分区生命周期管理）。 * 安全与合规方法：最小权限、字段脱敏、行列级权限控制与审计日志。 * 评估方法：以查询 P95/P99 时延、吞吐、成本（存储/计算）、可用性与易运维性为评价指标，结合真实业务画像进行 A/B 验证。 | |
| 1. **论文撰写提纲**  * 摘要与关键词 * 第1章 绪论   1.1 研究背景与问题定义1.2 国内外研究现状1.3 研究目标与贡献1.4 论文结构安排   * 第2章 相关技术与理论基础   2.1 视频用户行为分析概述与指标体系2.2 Apache Doris 架构与关键技术2.3 实时数据处理与可视化技术（Kafka、Flink、Superset 等）   * 第3章 系统需求分析与总体设计   3.1 业务场景与指标口径设计3.2 数据采集与埋点规范3.3 数据模型设计（事实/维度/宽表/物化视图）3.4 系统架构设计与模块划分   * 第4章 数据接入与存储优化实现   4.1 数据接入链路与质量校验4.2 Doris 表设计、分区分桶与导入策略4.3 物化视图与预聚合优化4.4 冷热分层与成本优化   * 第5章 实时计算与可视化系统实现   5.1 实时指标计算（Flink/Doris Stream）5.2 主题看板与自助分析实现5.3 权限、审计与数据安全   * 第6章 实验与评估   6.1 实验设计与数据集6.2 性能评测（导入、查询、并发、时延）6.3 对比分析与消融实验（分区/物化策略）6.4 案例与业务价值验证   * 第7章 总结与展望   7.1 研究结论与贡献7.2 局限性与未来工作 | |
| 五、参考文献  [1] Apache Doris 官方文档与社区实践，<https://doris.apache.org/>  [2] Carbone P., et al. Apache Flink: Stream and Batch Processing in a Single Engine.  [3] Yang X., et al. ClickHouse: The True Column-Oriented DBMS.  [4] Fang J., et al. Druid: A Real-time Analytical Data Store.  [5] Keller A., Netflix Technology Blog: Real-time Data Infrastructure for Personalization.  [6] Kimball R., Ross M. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.  [7] Superset 官方文档与最佳实践，<https://superset.apache.org/>  [8] Chen T., 国内视频平台用户行为分析实践综述，信息系统工程。  [9] Li H., 实时数仓与指标治理方法论，数据开发者大会论文集。  [10] GDPR/数据合规相关标准与行业指南。 | |
| 1. **指导教师评语**   是否同意开题： （是、否）  指导教师（签字）：  年 月 日 | |
| **七、学院审核意见**  负责人（签字） ：  年 月 日 | |