孙 亮 亮

联系电话：18810061382 电子邮箱：397585361@qq.com 出生年月：1992.01
政治面貌：中共党员 居住地址：北京回龙观昌平路 籍 贯：黑龙江齐齐哈尔

教育背景

2015.9 - 2018.6	中国石油大学（北京）	油气田开发	学术硕士	CET-6（488）
2011.9 - 2015.6	中国石油大学（北京）	石油工程	工学学士	CET-4（495）

专业技能

- 熟悉 Java、Scala、Python 语言，了解 JVM 并有基础调优经验，具有面向对象和函数式编程思维；
- 熟悉 Hadoop 集群部署和参数调优、MapReduce 计算框架，了解底层 RPC 框架及原理；
- 熟练掌握 Flume、Kafka、Zookeeper 等技术原理和应用，可进行二次开发；
- 熟悉数据仓库基础构建，数据 ETL，Azkaban 任务流调度，数据分析引擎如 Hive、Presto、Druid 等；
- 熟悉 Spark、Flink 计算框架基本原理和应用，有丰富的离线、实时任务开发和基础调优经验；
- 熟悉 Mysql 的部署模式、索引结构、锁机制、缓存策略等，Redis、Hbase 等 NoSQL 数据库；
- 了解 SSM 框架和应用、了解 ELK 日志平台。

工作经历

2018.7 - 2019.8	暴风集团股份有限公司
任职	DT 数据技术部门——数据平台开发工程师
工作简述	1.两个 hadoop 大数据平台和组件开发和维护（约 600 台服务器） 2.离线、实时任务开发（主导），基础数据接口的开发 3.各业务(ad、wireless、tv 等)120 多种 ltype 的数据 ETL 及数据仓库基础构建 4.满足其他部门如仓库、推荐、数据分析等所提需求

个人评价

胆大不妄为，保证线上正常情况下尝试新技术、优化现有任务，提高集群计算效率、工作效率；
自信不自负，工作中主动承担任务、问题，多次解决突发线上事件，有很好的 trouble-shooting 能力；
乐观而向上，热爱编程、钻研技术，热爱足球、羽毛球等运动。

项目经历一

项目名称	**用户行为实时统计**
项目周期	2019.2 - 2019.4
项目描述	通过 TV 心跳服务上报用户行为数据，实时采集消费、分析、展示，最重要的目的就是实时准确掌握用户动态，分析用户行为，为运营谋划、产品分析、相关推荐提供有力的数据支持。该项目主要包含以下需求（以下均分平台、版本）： 1.当前在线人数&当天累计在线人数趋势 2.当前用户分布&当天累计用户分布热度 3.当前用户所用第三方应用排行榜&当天第三方应用使用时间占比 4.当前用户观看影片/电视台排行榜&当天观看影片/电视台时间占比

	5.当前影片滚动信息流
承担任务	1.利用实时数据线进行埋点数据上报、采集、消费 2.利用 flink 的 SQL、Table API 和 Stream API 实现上述需求 3.将窗口计算结果写入 mysql 存储和 redis 队列供展示查询
主要技术	flume+kafka+flink+mysql+redis
主要成果	提高数据的实时性和多样性，可通过配置 sql 轻松扩展需求。
问题&解决	1.用户去重方案选择 当天累计需求去重计算时，checkpoint 时间过长导致结果延时甚至程序崩溃 方案①增加并行度，flink SQL 不分组的话是单一线程计算，但是并不起作用 方案②使用 redis 去重，弯路太多浪费计算资源且加大延时、扩展性不强 方案③使用 BloomFilter 去重，损失些许精度达到最优性能 2.TopN 实现方案选择 flink 最新版本的 SQL 还不支持 order by&limit 方案①使用 Table API，自定义实现 TopN 函数，可行但扩展性不强 方案②使用自定义聚合函数，维护最小堆

项目经历二

项目名称	**用户画像**
项目周期	2018.9 - 2018.12
项目描述	通过埋点上报、线上接口历史数据、用户信息，基于制定规则对用户进行画像，生成用户基本属性标签和会员相关衍生标签，目的是实现自动化运营及广告精准投放。 埋点数据+接口数据（用户信息，观看影片时间、类别、时长，搜索词等）； 用户信息（会员、会员类型、成为会员方式、历史消费等） 1.用户基本属性：男性、女性、老人、小孩、游戏迷、夜猫子等 2.消费属性：购买力强、中、弱 3.喜爱属性：影片类型偏好(*)
承担任务	1.利用 spark 调用 hql 过滤、计算 hive 中 org 层历史数据，同步 mysql 用户信息，时间粒度分为月级、周级、天级 2.基于制定规则的方式对每一个用户打标签 3.将结果写入 mysql&hive
主要技术	hive+mysql+spark
主要成果	推荐时长比提升 13%，广告收入提升 4%
问题&解决	1.规则的制定 由于没有训练集，所以采用规则的方式打标签，所以规则的制定会影响结果 2.标签稀疏问题(针对基本属性和喜好属性) 出现在天级计算，解决方法是 hive 中类似增量存储，mysql 中存储过去一周时间段最新的标签，极限是一周。