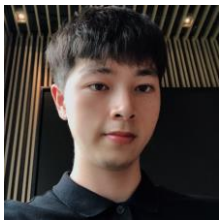

黄海峰 Pruce

湖南大学·11 级本科·软件工程

手机：138-0749-4071

邮箱：1756983926@qq.com

工作经历

深圳腾讯科技有限公司 数据分析工程师 2017.11-至今

为心悦俱乐部创建一套业务指标分析经分系统，搭建腾讯精品游戏用户画像，主导开发线上活动分析系统，为心悦线上活动提供埋点数据生成、采集、校验、处理、个性化统计分析、展示等一套完整的数据闭环分析系统。

深圳市英威诺科技有限公司 大数据工程师 2016.04-2017.11

负责数据收集（Flume、kafka 跨机房传输）、数据清洗（reformat）、数据存储（HDFS、S3）、离线数据计算（Spark、Hive）、实时数据计算（Spark Streaming）、数据统计落地（Mysql、Redis、Druid，Dashbord）等数据流工作。为推荐系统提供资讯、用户的累积、衰减(GMP)实时反馈数据，为算法模型提供训练数据，为报表展示提供格式化的聚合数据，并设计流量探索模型。

北京猎豹移动有限公司 Java 工程师 2014.07-2016.04

猎豹移动从工具转型内容，有幸成为第一批内容研发团队成员，研发猎豹移动手机 APP 内容推荐系统，负责 CMS 重构、初选规则、Hadoop Streaming 数据反馈统计。

项目经验

活动分析系统 深圳市腾讯科技有限公司 2018.08-2019.04

项目职责：系统负责人

项目描述：活动分析系统以线上活动数据为主线，对线上活动的各方面效果进行分析，日更新数亿条记录。整个系统包括活动配置管理、数据埋点上报、用户数据画像、Hermes 实时数据查询系统、心悦 UO 平台实时交互、用户自定义漏斗、元数据管理、任务监控等几个模块。活动分析系统打通用户画像系统，参与活动的用户也可直接进行用户画像分析。

活动的效果数据可分为：

- 活动基本数据分析，包括活动的 PV、UV、活动参与用户个性化漏斗数据分析；
- APP 效果，包括 APP 拉新、拉回流用户数据分析；
- 游戏效果，包括游戏拉活人数，拉充值人数、金额，预约人数、下载人数等；
- 分发效果，包括游戏渠道包染色用户数、染色用户游戏活跃数等。

用户画像系统 深圳市腾讯科技有限公司 2017.12-2019.04

项目职责：系统主要参与者

项目描述：腾讯精品游戏用户画像，每周约更新十几亿条记录，上万个标签，包含有用户自然属性标签、游戏大盘标签、单游戏活跃充值标签、个性化兴趣标签等等。

用户画像系统可分为

- 业务功能层，人群分析、产品分析、自助提包、营销自助化
- 实时存储层，实时数据查询、实时数据存储
- 画像层，画像总表、基础模块、游戏模块、心悦模块、偏好模块、数据字典配置模块

✧ 算法层，聚类生成新用户标签

✧ 数据源层，腾讯众多的底层数据表

个性化推荐系统 深圳市英威诺科技有限公司 2016.04-2017.11

项目职责：大数据团队核心成员

项目描述：1、原始 log 数据采集收集（kafka，MirrorMaker）

经历了一个从 log kafka 到直接写入 Kafka 的迁移过程，克服 kafka 跨机房传输数据吞吐量不高、高峰期数据延迟大、网络抖动导致传输失败的问题，完美替换 MirrorMaker 传输工具

2、数据清洗（reformat）

线上所有接口的原始日志都将汇总到 reformat，由 reformat 格式化后产生统一规范化的数据，输入和输出都是 Kafka，针对每个 partition 都起一个单独的线程处理

3、格式化的数据落地（HDFS、S3）

国内用 HDFS、海外用 S3，一个单独的程序会每隔十分钟将 Kafka 内的所有数据打包一份存入集群，数据上传的时候我们可以选择性加上 md5 数据校验

4、离线数据计算（Spark、Hive、AWK）

少量数据查询分析用 hive，每天规定的统计报表任务用 spark 来做，甚至再再小一点的十分钟级别任务用 AWK 来统计

5、实时数据计算（Spark Streaming）

实时反馈是实时个性化推荐系统必不可少的关键节点，目前的主要任务分为用户、文章十分钟、一小时数据统计，文章 impression 统计，用户居于 impression 的已读消重，文章的 GMP（实时衰减 CTR）等等。保证服务的高可用、高容灾、高时效是最关键点

6、统计数据落地（Kafka、Mysql、Redis、Druid，Dashbord，ES）

数据产出的多样化在个性化推荐系统中表现的淋漓尽致，不同类型的任务将会用到各种个样的输出存储方式

日处理数据约数 T，服务日活几千万

7、流量探索

为推荐系统设计并实现流量探索模型(已申请专利)，为系统设计出三大资讯池（探索池、进行池、等待池）来筛选优质的内容资讯，海外每天探索文章、视频几万篇。此外还配套设计实现了流量探索实时监控和天级流量探索效率反馈报表。有利于准确把握流量探索的各项关键指标

8、已读服务，确保不能让用户重复看到相同的资讯。

自我评价

多年的大数据、数据分析、商业分析经验，熟练掌握流式大数据处理方案，熟悉 Hadoop、Hive、Spark SQL、Kafka、Redis。一直负责数据工作，将数据收集采集、清洗、聚合、落地、报表展示各个环节打通，形成一套完整的数据闭环。

乐于分享，近半年写了 4 篇 KM 文章，收藏率 10%，且 4 次收入 KM 精选好文，3 次入选 KM 头条，2 次入选腾讯知识奖，且被腾讯技术工程公众号外部转载。

善于总结提炼，将项目中的核心技术点总结归纳深化，撰写了 3 篇专利。