

王皓

3 年工作经验 | 本科 | 24 岁 | 男

联系电话：13277812352

电子邮箱：2034364675@qq.com



自我描述

综合能力：责任心 | 个人能力 | 抗压能力

- 思维严谨，有较强的学习及适应能力。
- 责任心、执行力及抗压性强，乐于从事有挑战性的工作，能够积极的面对并解决工作中的问题。
- 有较强的团队合作意识，对工作尽职尽责。
- 热爱大数据工作，对新技术有着强烈的好奇心。

技能

- 1、熟悉 Java、Scala 编程语言
- 2、熟练使用 SQL 语言的基本使用，能够根据业务需求独立进行表的设计、构建和使用
- 3、熟练掌握 Linux 常用操作命令，并配合 shell 脚本进行使用
- 4、熟悉 Hadoop，深入理解 Map/Reduce，可对其进行部署和代码的编写
- 5、熟悉 hive 和 hbase 相关原理，并使其对海量数据进行处理
- 6、熟悉开源日志收集框架 flume、redis 和 sqoop 工具，能够进行基础的搭建和使用
- 7、熟悉分布式消息队列 kafka 和 zookeeper，可根据业务进行使用
- 8、熟练使用 spark core、spark streaming、spark sql 并了解其相关原理
- 9、掌握 storm 和 elasticsearch 的基本使用，可根据需求对其进行使用和开发
- 10、了解机器学习相关概念，能够对具体数据进行分析，如协同过滤等

工作经历

深圳捷誉技术有限公司

2018.10 - 2019.04

职位：大数据研发工程师

负责项目需求探讨，代码的编写，并结合 opencv 技术实现对视频流的处理。在职期间也对 Hadoop 数据存储源码部分进行研究，并尝试修改 Hadoop 的数据存储方式。

广东精点数据科技股份有限公司

2016.10 - 2018.08

职位：大数据开发工程师

负责项目前期需求调研，部分架构规划，运用 spark 对数据进行过滤筛选再封装，并运用 HDFS 和 redis 分别进行数据存储，后期参与项目部署和维护工作。

项目经历

网络视频流服务系统的研究

2018.11 - 2019.04

项目描述：

项目描述：主要针对监控摄像头数据源每秒产生的大量的数据而提出，以便使用者可以实时对画面进行管理和监控使用技术：CDH+spark+opencv+scala+kafka 项目流程：1、将摄像头数据引入到收集器中，并转换为 json 格式数据发送到 kafka 中，保证下游程序不会出现吞吐量暴增的行为

2、下游 spark Streaming 从 kafka 中获取到数据，并转换为可用数据格式

3、调用 opencv 把每帧数据绘制出图片并通过算法识别出每幅图中动态物体，并保存在 HDFS 中

4、同时将 opencv 输出的流信息可以显示在页面上

你的成就：

负责描述：主要负责 kafka 进群的搭建与配置的修改，并编写 spark 代码处理数据调用 opencv 识别动态物体，并发送给下游

珠海移动智能营销系统

2018.04 - 2018.08

项目描述：

项目描述：基于珠海移动网络商品和活动的智能推广和营销精确化的目的，成立该项目；分为渠道推荐、素材推荐和商品推荐，为了提高推广的转化率，可精准的将活动和商品推荐给用户，使其大大提高业务的转化率，提高价值度。使用技术：CDH+HDFS+Spark+shell+Oracle+hive 项目流程 客群推荐 1、源数据经过指定的业务规则由相关人员进行 hive 预处理，并存储在 hive 指定的库中供后续 spark 调度使用 2、由前端页面通过 Java 调度 shell 脚本，并使其调度 spark 3、基于内存的开源的分布式计算框架 spark 通过 ScalaAPI 调用 hive 中的处理后的数据做数据空值处理和数据的平衡问题，同时初始化处理状态写入 Oracle 4、分桶操作，结合业务做出分桶字典将数据作为矩阵实现分桶数据，并计算出其 WOE 矩阵 5、精准模式下计算出 IV 值，全量模式下可直接将 WOE 矩阵带入算法进行运算生成模型结果 6、调用模型应用于新数据，生成推荐结果 json 数据返回给前端，并将结果推荐结果表存储在 HDFS 中，并写入处理结果的状态写入 Oracle 中 渠道推荐 1、运用 spark 调用 hive 进行抓取数据，写入初始化状态到 Oracle 2、对处理后的渠道数据经过聚合，分组等操作获取到训练数据带入算法公式中 3、结果存储于 HDFS 中，并跟新结果状态到 Oracle 中

你的成就：

负责描述客群推荐：架构的规划；CDH 的搭建；前端对接接口格式和接口内容；Oracle 中状态表字段设计；分桶操作代码编写；WOE 矩阵代码编写；IV 值代码编写；HDFS 存储的规划；后期维护；渠道推荐：前端对接接口格式和接口内容；Oracle 中状态表字段设计；spark 代码的编写；HDFS 存储的规划；后期维护；

中山大数据挖掘平台项目

2018.01 - 2018.04

项目描述：

项目描述：该项目主要针对于算法的整合，使各种算法融入在一个平台中，供使用者的调度和应用使用技术：CDH + Hadoop + spark + hive + livy + mysql + kafka + impala 项目流程 1、由前端进行算法组件的托拽功能

实现算法的流程，将命令发送给程序 client 客户端 2、客户端将命令进行传递给消息中间件 Kafka3、livy 通过从 Kafka 中拿取的命令进行调度 spark，进行 spark 任务的执行并同时状态初始值写入到 MySQL 库中 4、算法包从 impala 中拿取数据，进行计算 5、结果集存储在 hive 并同步 impala 元数据，6、状态值更新

你的成就：

负责描述更替整体项目环境并修改 livy 内部文件配置，将最终结果导入 impala；对客户后期需求变更做出修改并重新部署

南航反爬虫系统

2017.08 - 2017.12

项目描述：

项目描述：该项目主要针对南航官网存在爬虫的问题提出，爬虫会消耗系统资源但是没有转化成销量，也影响正常用户的购票。该系统主要对恶意非法访问能做到实时识别、拦截和防范使用技术：

CDH+Hadoop+Spark+redis+kafka+lua+nginx+apache+mllib 项目流程 1、原系统是用 Nginx 作为 Web 服务器，Http 请求会首先经过这些服务器，通过拦截 Nginx 服务器的请求消息，将数据处理后传到 Kafka 的生产端；2、Kafka 作为数据消息缓存队列集群，将源数据传输给 Spark Streaming 进行实时处理/清洗，实时处理/清洗完成的数据返回 Kafka 队列并存入 HDFS 分别由 Spark Streaming 和 Spark 进行实时计算和离线计算，运算数据和运算结果，在 Redis 和 Mysql 中存储；3、数据应用，包括实时监控预警、报表分析和反爬策略控制等；4、整个数据传输和反馈过程，提供数据平台的监控管理和分布式协调控制服务；5、建模的结果通过反馈机制，最终作用到官网系统，将识别出的爬虫和占座用户，通过拦截处理，将后续请求阻挡在 Http 请求阶段。

你的成就：

负责描述数据处理模块：将规则抓取到内存中，从 kafka 中读取数据进行匹配，按照规定的字段进行拆分获取数据再传给 Kafka 共后续调用防爬规则模型：从 kafka 中抓取数据按照指定的规则进行统计和判断

书海小说推荐平台

2017.03 - 2017.07

项目描述：

项目描述：该项目为了解决信息过载的问题而提出，通过使用推荐算法来分析用户行为从而构建出个性化的推荐模型，来对用户做精准的推荐，提高转化率。使用技术：

nginx+mycat+MySQL+hive+Hadoop+spark+dubbo+elk 项目流程：1、从 nginx 集群中将用户日志数据加载到 MySQL 集群中 2、增量的将行为数据放到 HDFS 中，用于线下处理 3、通过 hive 对 hdfs 中的数据进行数据清洗和预处理操作 4、将处理完成的数据通过 mllib 进行模型训练，并发送给线上 5、将用户日志收集到 elk 中建立索引，供前台进行查询 6、通过用户访问数据调用前台进行查询，返回给用户推荐列表

你的成就：

责任表述：主要负责 spark 代码的编写；离线数据的清洗和特征的提取和算法模型的构建

折 800 网上商城日志分析系统

2016.10 - 2017.02

项目描述：

项目描述：通过分布式 hadoop 集群处理网站产生的大量日志，分析出网站 PV、UV、访问 IP 数量、跳出率、平均访问时长、用户地理位置分布等信息共使用者做出相应的决策。使用技术：

nginx+flume+HDFS+MapReduce+Hbase+Hive+Mysql+Sqoop 项目流程：1、通过 flume 将 nginx 集群中的日志文件导入到 HDFS 中 2、对 HDFS 中的脏数据使用 MapReduce 进行数据清洗 3、将清洗过的数据存储在 HBase 中 4、将需要的数据字段导入到 hive 中并作相应的分析，并作自动化处理 5、使用 sqoop 把 hive 分析的数据结果导入关系型数据库 Mysql 中供前台调用 6、通过 highcharts 将数据以报表的方式展现在 web 页面中

你的成就：

责任表述：通过项目方案的讨论，主要负责修改 flume 配置文件构建 flume 环境；对 MapReduce 代码做 ETL 的编写，对 hive 操作做自动化处理，和后期项目的部署

教育经历

湖南理工学院 本科 / 信息与计算科学

2013 - 2017