

刘刚键-大数据开发简历

个人信息

28 岁 | 男 | 本科 | 5 年经验 | 18786708783 | arlenliugj@sina.cn

个人优势

编程语言：Java、Scala、OC 等

iOS 技术：响应式、组件、音视频开发

后端技术：Spring、SpringBoot、MyBatis、Activiti、MySQL、Redis、Docker、K8S、神通、达梦

大数据技术（熟练）：Linux、HDFS、Yarn、MapReduce、HBase、Hive、Kettle、Spark（SQL/Streaming）

大数据技术（了解）：Kafka、Flink、Livy、Zookeeper、Flume、Sqoop、ClouderaManager、数仓维度建模（Kimball）

个人博客：<https://liugangjian.com>

教育经历

井冈山大学 | 计算机科学与技术 | 本科 | 2012-2016

工作经历

北京计算机技术及应用研究所 | 大数据研发工程师 | 2019.04—2021.2

1、开源大数据产品研究、适配 arm64 处理器及二次开发，Docker 部署、K8S 管理。

2、Java/Scala 项目，从原型图设计、数据库设计、代码编写、测试上线等全流程参与，主要实现功能：工作流审核、ETL 系统开发、字段建模供文本关键字要素池抽取、数据处理算法应用、JVM 参数调优等。

3、承担大数据平台的实时维护优化、日常监控、故障处理、日志分析、数据备份、性能瓶颈分析等工作，提高系统的可用率及可维护性。

贵州痛客梦工场科技有限公司 | 大数据运维开发工程师 | 2018.03—2019.03

1、基于大数据解决企业服务痛点的平台，解决工商财税、品牌营销、IT 研发、创新方案等问题。

2、职责：CDH 大数据集群的运维监控、日志管理（集群日志、网站和 APP 埋点日志）、Java 接口开发等。

贵州宏思锐达科技有限公司 | iOS 工程师 | 2017.03—2018.03

居家游旅游有限公司 | iOS 工程师 | 2016.04—2017.03

项目经历

JJW 纪检监察数据融合应用 | 大数据开发工程师 | 2020.04—2021.2

描述：存储计算用阿里 EMR 离线部署版本，NLP 提取用阿里自学习平台离线部署版本，数据来源为历史系统数据、纸质文件扫描、其他单位引接数据、业务员实时录入等，涉及七大业务类型。所负责功能：数据引接口、批处理和实时处理、文本提取及识别、数据检验查重等；其他工作：参与服务器部署规划、网络拓扑设计、各系统数据流程及数据库表设计、集群运维调优等。

1、数据引接口：实现与阿里 NLP 服务器的 HTTP 交互、HBase 增删查改、HDFS 上传下载及文件系统管理、MyBatis 读写 Hive、Livy 提交 Spark 代码片段接口、递归扫描目录、NLP 结果通过 Java Runtime 转换成具体业务模型等工具封装。

2、离线场景：HBase 存储某任务相关的半结构化及纯文本数据，Hive 存储各阶段中间处理结果，Spark 常用算子解析 OCR 结果、NLP 模型迭代、NLP 预测、解析 NLP 识别结果（分区、过滤、topN 排序等）；部分同一业务不同关键字涉及多次循环扫描，需对任务 ID 进行哈希分区，并广播依赖的原文本；部分业务类型使用结巴分词的精确模式或基于正则规则提取更优；提取结果通过 JDBC 写入神通数据库。

3、实时场景：各下属单位、外部单位引接数据（如业务员实时语音）进入 Kafka，本模块消费使用 SparkStreaming 每 60 秒（批次间隔 10 秒、窗口大小为 6、滑动步长为 5）汇总文本数据，走 NLP 实体提取后入 Hive，给其他功能模块使用，如再走离线场景。

4、模型迭代：收集的用户修改 NLP 识别结果的日志，Spark 过滤后处理后转储为 NLP 语料数据集放入 HDFS，后台进程扫描目录上传阿里 NLP 迭代模型；制造数据丰富语料库，对 NLP 提取效果差的人工标注。

5、作为技术负责人之一与需求侧对接，收集需求字段建表提供给 NLP 提取，维护 Gitlab 管理代码等。

GXB 数据资产管理子系统 大数据开发工程师 2019.10—2020.06

描述：内网部署项目，负责数据清洗转换功能模块的原型图设计、数据库设计及接口开发。

1、数据来源为下属单位填报数据、互联网端爬取动态建表数据，提供增量和全量抽取方式、整表抽取和部分字段抽取，应对维度表和事实表主题；

2、Kettle 二次开发：复杂作业转换流程配置、数据维度建模，从数据二义性、重复、不完整、映射、维度转换、合并拆分等角度对数据清洗转换后，Upsert 加载入库为其他子系统提供数据支撑，调度任务、转换、算子任意纬度组合，任务执行异常、延时、重试、报警、生产结果异常需时自主判断写入日志，供前端及日志管理模块读取；关系数据库为达梦，项目开发使用 SpringBoot+MyBatis，部署为 Docker 与 K8S。

GXB 信息窗管理子系统 大数据研发工程师 2019.10—2020.06

描述：互联网、内网部署项目，负责 workflow 审核功能模块的原型图设计、数据库设计及接口开发。

1、中台提供基于 Activiti 6.0 搭建的工作流引擎、BPMN2.0 流程设计，基于此开发业务功能。

2、因走网关 workflow 无法检索，使用网络数据、本地用户信息数据与 Redis 结合方式，对 Set 做交集并集实现多维度筛选数据；使用 Pipeline 批量操作大幅提升性能。

3、数据存储达梦数据库，达梦提供单向网闸做外网、内网数据同步，代码定时器监控数据状态。

4、不同网络下对审核通过的数据推送到不同的用户、单位；开发环境部署环境同上一个项目。

军队某数据服务平台 大数据研发工程师 2019.04—2019.10

描述：1、流计算，建立 Kettle 任务抽取数据、跑 Spark 任务测试，任务监控为星环 ETL 工作流调度平台。

2、数据治理平台外协部署跟踪，编写测试用例，包括元数据管理、数据共享、血缘关系、数据目录等。

3、写 MapReduce 程序打包运行，清洗固定共享目数据结果输出到 HDFS；Neo4j 命令脚本实现导入导出，数据在 Neo4j 多维度展示。

4、其他，如 Docker、K8S 操作使用；RAID 卡安装做阵列；系统硬盘拷盘机备份。

痛客网 大数据运维开发工程师 2018.03—2019.03

描述：部分具体功能业务如下：服务商详情页咨询、交易详情咨询、一键拨号等咨询日志；发布需求的三种方式、优选服务商广告、banner 广告等需求及广告日志；用户网站/APP 交互、登入登出注册、第三方分享、分类面包屑筛选及多类型搜索等日志。

CDH 大数据集群安装、日常运维监控，核心数据埋点数据服务功能设计及后台开发、APP 开发参与。数仓建模共约 50 个字段，支持增量、全量做 ETL，数据处理（清洗脏数据、关联聚合、数据倾斜等）后提供给分析侧做报表，供不同分析需求（趋势、搜索、受访、转化、来源、访客等）。项目在数博会上展示。

技术架构：Hadoop + Hive + HBase + Flume + Sqoop + Nginx + Shell + Linux。服务器配置（测试/生产）：8/12 个，CPU 8 核，内存 64G。

星厨餐帮 APP iOS 工程师 2017.10—2018.03

描述：星厨课堂、星厨资讯、星厨动态、星厨上门、星厨俱乐部、星厨发布、厨友圈、消息、星餐厅等。

全智易联 APP（用户端与商户端） iOS 工程师 2017.04—2017.10

描述：用户端：基于地图搜索附近店家，导航到店消费；商户端：分销系统与店铺管理。

居家游 APP iOS 工程师 2016.04—2017.03

描述：房客查看目的地民宿房源，选择满意的房间下单订房；房东可以登录房东端进行房源和订单管理。