|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **刘源远** 算法开发工程师 | | |  |  |
| 性 别： | 男 | 生 日： | 1992.02.21 |
| 民 族： | 汉 | 现 住 地： | 上海长宁区 |
| 电子邮箱： | yuanyuan.liu@dsglyy.com | 电 话： | 19901721187 |
| 教育经历  **2011/09 – 2015/06 赣南师范 数学与应用数学 本科**  软件技能   * 熟练SVM、Logistic、Decision、Bayes、KNN算法的原理及编码实现； * 熟悉Spark Mlib，并能熟练使用Spark Mlib完成算法开发； * 熟悉 FPG，ALS，USER-CF，ITEM-CF等推荐算法原理，以及Spark mlib开发实现； * 熟练使用数仓，从数仓构建维度、维度清洗；精通HQL开发； * 熟练图数据库 Nebula Graph，图空间设计，数据导入、查询、以及API使用； * 熟练使用数据实时处理相关工具 Flume、Kafka、Flink、Spark Streaming、以及Mysql实时同步工具Canal; * 熟练Spark大数据内存迭代计算框架，了解Spark架构原理，调度和性能优化，熟练使用Scala实现流式和离线业务逻辑； * 精通Java1.8及相关语言特性，Java 分布式开发和高并发控制，熟悉JVM的内存结构，有多线程和高并发以及原生JAVA服务开发经验； * 精通Scala及相关语言特性及其开发，熟练Python语言及开发； * 熟练使用 Mysql，Oracle，SqlServer关系型数据库； * 熟练使用大数据生态圈Hadoop、Hbase、Kylin、Hive、Spark Core做维度清洗； * 熟练使用Linux系统，以及Shell编程，熟练使用Docker，以及复杂网络环境中各种Docker容器网络互通配置； * 熟悉计算机组成原理，计算机网络，操作系统课程的主要内容。   工作经历  **2017/09 – 现在 驴妈妈 数据挖掘工程师**   * **工作描述**  1. Spark Mlib开发，数据建模； 2. 维度清洗筛选，算发效果回收跟踪； 3. 推荐数据输出接口开发，A-B测试接口开发； 4. 参与部分数仓开发与Spark实时开发； 5. 数仓、Job、集群BUG FIX; 6. 大数据集群、实时计算集群、AI集群管控，线上数据输出服务优化管控。  * **参与项目**  1. **项目名称：用户近期下单预测**   项目描述：该项目是基于用户基本属性，近期浏览、登录、收藏行为，历史下单，历史出游数据，构建训练数据，使用多个分类模型组合训练用户近期下单预测模型。项目主要用于为用户画像提供数据支持，为客服外呼提供数据依据。  核心技术：Hive、数仓、Kettle、Anaconda3。  使用语言：Python3、HQL  项目职责：   * + 1. 项目调研，需求调研，数据支持调研；     2. 维度初步选取，训练集构建方案预演；     3. 大部分维度清洗，部分数据ETL；     4. 维度关联性分析，部分维度的计算口径定论；     5. Logistic、Decision、Bayes开发，模型训练；     6. 线上运行环境打通，任务调度配置；  1. **项目名称：用户画像**   项目描述：该项目综合各个部门历史的用户标签逻辑，构建公司统一的用户标签体系。基于用户基本信息、全平台所有订单数据、出游数据数据、页面埋点数据、用户行为记录、客服手动维护会员数据构建注册用户以及未注册用户标签体系和偏好体系。项目主要用于客服部外呼，会员部精准营销，智能数据部后续推荐提供数据支持和输出。同时提供自选时间段内用户行为、偏好聚合查询接口。项目分为两部分：数据开发部分，公司内部用户查看页面。  核心技术：Hive、数仓、Kettle、Kafka、Flume、Nebula Graph。  使用语言：Python3、HQL、Scala  项目职责：   1. 数据开发部分负责人，数据经理； 2. 前期各个部门标签逻辑对接、梳理、整合、口径确认、统一，数据支持调研； 3. 各平台埋点数据收集到数仓，埋点数据清洗，用户唯一机器标识符确定，数据中用户ID覆盖率最大化； 4. 用户登录次数、在线时长、浏览记录、收藏、下单维度标签开发，； 5. 根据各类订单数据构建用户旅游消费水平等级； 6. 用户分群开发，用户偏好模型开发，构建用出游偏好； 7. 使用Nebula Graph图数据库，构建用户行为图空间，实现根据用户行为快速生成任意时间段内用户的偏好。 8. **项目名称：智能推荐**   项目描述：该项目是基于用户近3个月浏览、收藏行为，历史下单，历史出游数据构建用户评分矩阵，同时以产品的主题、标签、价格、品类、目的地构建构建用偏好模型，实现，人推产品，产品推产品，千人千面功能，主要用于产品详情页，支付完成页，频道页产品推荐。  核心技术：Hive、数仓、Kettle、Anaconda3、Spark。  使用语言：Python3、HQL  项目职责：   1. 构建评分矩阵，以及用户偏好模型； 2. Spark milb开发 , ALS，FPG，ITEM-CF，USER-CF，基于内容推荐开发； 3. A-B测试数据输出接口开发； 4. 产品描述信息关键字提取； 5. 基于产品内容的推荐开发，产品属性、维度构建，相似度计算； 6. **项目名称：数仓建设**   项目描述：该项目为驴妈妈数据仓库从零到对其他部门提供取数需求的项目，为一般的数仓建设项目，主要为数仓平台建设，数据主题，数仓分层，数据导入，数据ETL，数据模型设计构建，数仓统一化调度……。  核心技术：Hive、Spark、Shell、Kettle、Flume、Kafka、Canal……  使用语言：Python3、HQL、Shell、Scala、Java1.8  项目职责：   1. 订单模块原始数据模型梳理与构建； 2. 使用Kettle从原BI系统数据抽取，数据ETL、宽表开发； 3. 使用Spark处理Nginx日志，入仓； 4. 使用Flume + Kafka抽取接受前段埋点，同步至数仓； 5. 使用 Mysql + Canal 实时抽取数据到数仓，Canale数据同步服务开发; 6. 使用Spark开发实时订单业务逻辑。   **2015/06 - 2017/08 溢思得瑞科技创新集团 大数据开发工程师**   * **工作描述**   + - 1. 客户大数据项目调研、需求分析、开发、维护；       2. Hadoop开发、Spark开发、Hive开发，数仓开发；       3. JRBDP平台维护，调研，服务集成开发，核心功能测试，迭代开发；       4. JRBDP 打包、平台组件RPM包制作       5. 项目部署文档，使用文档撰写。 * **参与项目**  1. **项目名称：无锡横新电站预测**   项目描述：电站使用光伏发电，为将电站的发电量接入国家电网，需要对电站的日发电量有个比较准确的预测，同时要能实时检测光伏电板的异常检测。现在需求建立电板异常检测模型，以及发电量预测模型，异常检测模型由One-Class SVM训练得到，通过辐射值构造B -Spline预测日发电量。项目和同事分工完成，我主要负责异常检测部分模型训练。  核心技术：Spark、Kafka。  使用语言：Java、Scala、Python。  项目职责   1. 调研Spark的Mlib中SVM的实现及API，以及One-ClassSVM建模； 2. 样本数据标记，以低于均发电量55%的数据为异常数据（负类样本）； 3. 使用Spark API 训练校验测试One-Class SVM； 4. 使用Python One-Class SVM 验证Spark API实现的One-Class SVM模型的准确性。 5. **项目名称：聚睿大数据平台2.0研发**   项目描述：该项目是基于Hortonworks 开源项目Ambari的二次开发，主要实现应用汉化，国内高效安装源服务器搭建、一键化安装工具开发、组件生态丰富、服务增加集成、功能扩充等。在2.0的版本中，需要增加服务Kylin和Greenplum，主要实现服务以来，配置安装，cube监控，多节点部署，服务监控等功能，我主要负责Kylin部分的开发，以及平台打包。  核心技术：Ambari、Kylin、Proxy、Hadoop、Hbase、Hive。  使用语言：Python、Shell  项目职责：   1. 调研Kylin的工作原理以及与Hadoop生态圈的协同工作机制， 2. Kylin Ambari集成化开发，包括服务控制模块、服务监控模块、页面化配置、页面化安装，挖掘Ambari和Kylin的部分隐藏Api，以及Kylin隐藏参数配置化； 3. Kylin与Ambari自带Hadoop生态圈、平台个性化后的兼容性开发。 4. 开发服务控制模块与监控模块代码，重组Kylin原生配置、以及Kylin Rpm包制作，以及Yum源搭建； 5. 使用平台现有的Haproxy对Kylin做负载均衡集群式集成； 6. 挖掘使用Kylin隐藏的Api设计Kylin的监控指标； 7. 集成功能测试，以及场景测试，迭代增加Kylin额外功能和bug修复。 8. **平台优化升级**   项目描述：该项主要实现聚睿大数据集群实时监控，数仓平台优化，机器学习平台搭建，实时计算集群搭建，数仓备份，线上对外数据服务Docker化部署，多节点多实例docke容器打通，Docker网络使用规范集成。  核心技术：Zabbix、Docker。  使用语言：Python3、Shell   1. Zabbix 服务搭建配置，问题排查，一键化机器分组监控工具开发； 2. Spark社区版本 针对Ambari适配化修改编译，安装； 3. 数仓备份工具FUSER编译装调试，数仓备份； 4. 大数据平台参数调优。 5. 针对Ooize 和 Hue 数仓调度插件开发，实现各种任务统一化调度，任务依处理，断点执行，串并行控制，超时控制，参数传递与同一，任务Log控制，实现与Ambari自带Ooize 和 Hue兼容结合，提供简易已维护任务配置入口； 6. Docker各个网络模式测试配置，包括宿主机网络桥接，Nat+静态路由表项，以及docker自带Bridge,Host模式测试。 7. Docker化项目一键部署脚本开发,Docker使用规范制定，Docker网络使用规范制定； 8. **项目名称：鹏瑞利用户行为实时跟踪系统**   项目描述：该项目是合作公司鹏瑞利全国综合体商城的每日流动顾客行为实时跟综，用终端MAC标记顾客，数据由商城提供的WIFI AP收集发送至Kafka，以报文的形式暂存于Kafka，根据甲方需求提取报文中相关长度字段实现数据还原，清洗过滤，业务逻辑，本人与同事2人协作开发。  核心技术：Spark、Kafka、Mongodb。  使用语言：Scala、Java。  项目责任：   1. SparkStreaming环境测试，独立实时计算集群安装； 2. SparkStreaming开发，数据还原，清洗过滤，格式化； 3. 编写核心业务需求代码，按分钟标记数据，计算顾客最早光临时间以及新增顾客； 4. 业务需求迭代开发更新，以及项目维护；   **2014/08 - 2015/04 Java工程师（实习）**  求职意向  **目前状况**：离职状态，一月或者年后入职  **工作性质**：全职  **期望职业**：软件/系统集成/大数据/算法  自我评价   * 注重团队合作，工作态度认真； * 性格沉稳，理性，办事谨慎，自学能力强； * 有很强的逻辑思维能力，分析能力； * 有钻研精神，对代码执行效率和性能有极致的最求； * 有很好的理解能力和团队合作能力以及沟通能力。 | | | | |