

# 大数据开发工程师

## 个人信息

姓 名：	xxx	性 别：	男
年 龄：	xx 岁	学 历：	本科，工科学士
籍 贯：	xxxx	联系电话：	
工作经验：	x 年	电子邮件：	

## 求职意向

工作性质：	全职	目标地点：	上海
目标职能：	大数据数仓开发工程师	目标薪资：	面议

## 教育经历

学校名称：清华大学  
层次：  
毕业时间：2017 年 6 月  
专业：生物工程

## 工作经历

单位名称：xxxxxxxxxx  
职位名称：大数据数仓开发工程师  
工作时间：2017.6-2019 年 2 月  
工作描述：1、对接各部门，根据需求进行脚本开发，对数据进行逻辑加工，源数据运维  
2、需求调研，需求分析，功能模块分析以及编码实现等。  
3、ERwin，Cognos 等工具数据仓库模型的设计优，输出逻辑数据模型  
并完成物理化工作  
4、采用 Kettle，DataStage 等工具进行 ETL、ELT 开发  
5、Tableau, PowerBI, FineBI 进行数据看板开发  
6、降范式建模，源数据 ID 调研，BD 需求调研，数据集成  
7、核心系统，个人信贷系统，对公信贷系统等 19 个数据源系统数据调研，输出调研文档，根据调研情况进行模型设计，输出 Mapping 设计文档，中科天玑大数据平台进行 HQL，SQL 开发和优化

**单位名称：** X X X X X X

**职位名称：**机器学习/数据挖掘工程师（初级）

**工作时间：**2019.2-至今

- 工作描述：**
- 1、项目的需求分析
  - 2、项目的架构设计
  - 3、项目的扩展性设计
  - 4、Atlas 元数据管理
  - 5、采用即席查询工具 Presto,Druid,Kylin，进行指标分析
  - 6、Griffin 质量监控，分析数据异常，并处理异常
  - 7、评估模型的精读，召回率，准确率
  - 8、构建并改善识别付款欺诈行为的线性模型
  - 9、处理时间数据调整数据格式，使数据适合训练机器学习模型
  - 10.使用 SageMaker 的 DeepAr 算法进行时间序列预测

## 专业技能

- 1、熟练 Spark 底层原理和运行流程，能使用 SparkCore、SparkSQL、SparkStreaming 进行开发，了解 SparkMLlib 并能使用其实现数据模型的加载，训练，系数调优以及模型的评估；
- 2、熟悉 Hadoop 工作流程以及原理，包括 HDFS 的读写流程，MapReduce 原理流程，Yarn 任务调度以及元数据的管理
- 3、熟练使用 Hive 构建数据仓库并能做一定的 Hive 调优
- 4、熟练 HBase 架构，熟悉 RowKey 设计原则，了解 HBase 优化
- 5、熟悉 Redis 内存数据库，了解 Redis 的持久化
- 6、熟悉 kafka 集群的架构，消息不丢失、负载均衡等机制，了解集群的搭建与调优
- 7、熟悉 Flink 底层原理，能够使用 Flink 开发常见的指标分析
- 8、熟练使用 Sqoop 数据迁移框架、Flume 采集工具。
- 9、精通 Java，Scala 语言，了解 JVM 调优，有前后端开发能力。
- 10、精通 Linux，了解 Shell 编程；
- 11、精通 Mysql，Oracle,有 SQL 调优能力.
- 12、了解机器学习的常见算法，如决策树，随机森林。

## 项目经历

### ● 金融产品个性化推荐平台

- **开发环境：**Windows + JDK1.8 +IDEA+Linux+Scala
- **系统架构：**Hive+Zookeeper+SparkMLlib+Flume+HDFS+Sqoop+Redis+Mysql +Kafka +Hbase
- **开发时间/开发周期：**
- **项目目标：**

公司重点开发数据挖掘、数据分析、人工智能等相关领域产品。为相关理财产品推广并获取更多用户，让用户选择更适合自己的产品。对前期用户消费行为、额度使用情况、

业务订购等不同历史行为进行维度分析；通过聚类分析建立模型，对用户业务推广提升效率。通过分析用户大量历史数据获取用户潜在 规律和价值，智能为用户推荐喜欢的业务和产品。再和基本特征权重加和排序取出

➤ **责任描述：**

1. 参与项目初期设计，需求调研、技术选型、文档编写；
2. 将 MySQL 数据导入到大数据 HDFS 文件系统，用 HQL 分析处理数据；将相关数据导入 Hive 数据仓库，进行后期数据处理，将 Hive 中数据转化成建模所需格式；
3. 通过 SparkStreaming 对数据清洗并进行分析，为算法同事提供模型调优数据，将数据模型转为推广到用户群的产品；
4. 用将清洗后的数据用来模型训练，构建特征索引，以便通过逻辑回归算法构建模型放在 redis 中；
5. 对产品特征数据进行抽取以及关联的历史购买记录维度数据进行优化，保证推荐结果准确性；
6. 使用 ALS 协同过滤算法，将推荐算法与业务系统紧密结合；
7. 对用户历史数据和负例样本作去噪处理，保证模型准确度；
8. 从 Hbase 获取产品基本特征和用户历史购买数据，从 Redis 读出模型，预测排名

● **用户画像平台**

- **开发环境：**Windows + JDK1.8 + Hadoop-2.6.0+IDEA+linux+Mysql+Scala2.11
- **系统架构：**Saprk+hadoop+Zookeeper+Flume+Hive+Hbase+scoop
- **开发时间/开发周期：**
- **项目目标：**

通过采集网站用户访问日志数据和已有业务系统的静态数据，进行数仓建模，提炼用户画像，为银行实现精准营销、推荐系统、完善产品运营等做数据支持。将业务系统的业务数据从 mysql 中的 user 表、order 表和 log 日志表导入到 Hbase 中，加载 Hbase 中的原始数据以及 MySQL 用户标签规则数据，用 spark 对匹配型、统计型和挖掘型标签进行计算,最终得到标签结果,将结果存入 HBase 中。

➤ **主要职责：**

- 1.负责前期用户画像需求分析与架构设计；
- 2.使用 Flume 采集用户信息数据、用户交易数据、用户访问日志数据到 HDFS 中；
- 3.在 Hive 中建立用户画像模型表，并将之映射到 Hbase 中
- 4.Spark 与 Hive 的整合，利用 Spark 引擎计算用户的标签属性。
- 5..负责用户画像中匹配型标签和统计型标签的计算
- 6..使用机器学习算法完成 RFE 用户活跃度模型、RFM 户价值模型、PSM 价格敏感度模型等算法挖掘型标签的开发和算法参数对的调优
- 7.建立 solr 与 Hbase 之间的映射并最终将数据展示在 web 页面上

- 信用评级风控系统

- 开发环境：Windows + JDK1.8 + Hadoop-2.6.0+IDEA+linux+Mysql+Maven3.0.5
- 软件架构：Hadoop+ Flume+Hbase+Hive+Sqoop+Mysql
- 开发时间/开发周期：
- 项目目标：

该项目通过电商部门数据结合互联网数据对商户信用评级，根据评级给商户划定贷款额度。主要分析的数据来源有公司的业务数据，通过和其他数据公司的整合数据等，通过分析商户的资产状况，生活习惯，个人信息以及交际圈得出信用等级，给公司的金融部门做决策。

- 责任描述：

- 1.使用 Flume 将业务系统产生的日志数据采集到 HDFS 中 ,为使用 Hive 做数据分析工作做准备。
- 2.使用 Sqoop 将业务数据库的数据导入到 Hive 表中，用来分析查询。
- 3.对来自第三方的数据使用 Spark 程序进行数据的预处理清洗工作，将不合法数据以及缺失字段的数据做标记逻辑删除。
- 4.将存在于 HDFS 上的数据映射到 Hive 表中 ,整合 Spark SQL 做数据 ETL 工作 ,构建数据仓库。
- 5.使用 Hive 构建数据仓库中 ODS 层数据表，DW 层维度表，事实表，使用 HQL 抽取数据转换数据到相应的数据层。
- 6.将 Hive 分析后的数据导入到 HBase 中，为业务部分查询分析做准备。
- 7.使用 Phoenix+Hbase 整合做查询优化分析，构建用户画像，分析用户风险指标。
- 8.使用 Sqoop 将数据导出到 MySQL 数据库中，为其他部门做深度数据挖掘，算法分析工作做数据支撑。
- 9.使用逻辑回归构建信用评分卡 A 卡、B 卡、C 卡

## 自我评价

- 1、参与过多个项目的开发，熟悉常见业务处理，能快速融入团队；
- 2、逻辑思维能力强，思路清楚，学习能力强，对新技术有着强烈的好奇心；
- 3、对工作尽职尽责，乐于从事有挑战性的工作；
- 4、乐于与用户以及同事和领导沟通，以便快速解决项目遇到的问题；
- 5、具有良好的英语阅读能力，能阅读英文资料、技术文档等；
- 6、高考物理满分，大学高数一 98 分，线性代数，概率论，统计学高分通过，对数字数据较为敏感；
- 7、初中开始接触计算机编程，热爱游戏和编程开发，热爱 IT 行业新技术；
- 8、扎实的 Java 基础，良好的编码风格。