**[ 基本信息 ]**

姓 名：徐凯 性 别：男

年 龄：28岁 联系电话：13231423298

电子邮箱：xukai5265@126.com 毕业院校：中国地质大学

专 业：计算机科学与技术 学 历：大专

工作年限：2年Java服务端开发经验;2年+大数据开发经验

**[ 求职意向 ]**

SPARK 大数据开发方向

**[ 个人技能 ]**

1. 掌握JAVA、多线程等JAVA高级特性的使用，并深入理解面向对象编程;
2. 掌握SCALA函数式编程，并深入理解隐式转换与类型类等scala高级特性;
3. 掌握爬虫经验，熟练使用Jsoup、HtmlUnit 爬虫工具;
4. 掌握flume、sqoop、hive、kafka、hadoop生态系统的使用, 理解Spark工作机制，掌握Spark core、Spark sql、spark streaming的开发;
5. 熟练常用设计模式的使用、贝叶斯算法、决策树算法的原理;
6. 熟练使用redis缓存，kafka消息队列的使用;
7. 熟悉ElasticSearch 开发，对ES生态系统有一定的认识；有基于ELK系统监控、日志收集的处理方案相关经验;
8. 能够熟练的在Linux环境下的开发，部署，配置;
9. 了解spring cloud 微服务

**[ 工作经历 ]**

2017/05 --- 至今 信和财富 | Hadoop 开发工程师

工作内容：

1. 提供大数据解决方案
2. 征信档案

2016/05 --- 2017/05 灵伴科技有限公司 | 大数据开发

工作内容：

1. 质检系统研发

2015/07 --- 2016/05 百分点信息科技有限公司 | 大数据平台开发

工作内容：

1. BDOS 大数据操作系统平台研发
2. 告警系统研发
3. 调研 ambari 的架构，实现对ambari 进行二次开发

2013/09 --- 2015/05 北京视点科技有限公司 | 后台研发工程师

工作内容：

1. 负责进销存系统开发
2. 技术指导及培训

**[ 教育背景 ]**

时间：2010年09月~2013年07月

学校：中国地质大学

专业：计算机科学与技术专业

主修课程：数字逻辑电路、计算机组成原理、微机系统结构、数据库概论、数据通信与计算机网、计算机网络工程、C语言基础、JAVA基础

**[ 项目经验 ]**

1. **舆情分析:**

**项目简介：**第一时间知晓公司新闻动态信息，了解有哪些是正面的新闻，哪些是负面的新闻。

**项目架构：**Jsoup、HtmlUnit + Spring Cloud + Kafka + Spark2.2

**算 法：**贝叶斯

**实现原理：**

1. 数据爬取阶段：

使用爬虫爬取互联网（百度，360新闻，搜孤，知乎等网站）关于本公司的数据，持久化到关系型数据库中。

1. 数据标记阶段：

人工标记爬取到的网页数据，标记结果： 1 正面 0 中立 -1 负面

1. 训练数据模型

获取标记结果数据，使用中文分词工具ansj，对文本进行分词处理。

计算分词数据词频信息，根据词频数据计算TF-IDF 值

1. 使用spark ml NaiveBayes算法训练分类模型
2. 将 步骤1 中的数据写入到kafka中
3. 使用spark streaming 实时获取爬虫数据，进行分词处理后，使用分类模型预测数据，将结果写入到关系行数据库中。

**总 结：**

1. 在分词阶段使用不同的分词算法会对分类结果有一定的影响
2. 在分词中需要去掉标点符号，需要使用停词库
3. 由于贴吧类网页，其内容大多是回复类的，比如提问帖子，每个人的回复既存在正面的也存在负面的，所以贴吧类的帖子暂时无法分类，
4. **征信档案：**

**项目简介：**贷款系统用户越来越多，导致系统响应性能下降，为了提升系统的响应性能，提升客户的贷款效率，通过对数据的预处理，提升系统响应能力。

**项目架构：**Hive 数据仓库、Tez计算引擎、sqoop、azikaban调度系统

**实现原理：**

1. 编写sqoop 脚本，同步关系型数据库数据到HDFS中
2. 创建外部表映射hdfs目录中的数据表
3. 数据模型

数据分为三个逻辑层 ODS层、IDS层、ADS层

ODS层：为贴源层，与关系型数据库数据基本保持一致，多为外部表，分为HDFS目录和hive数据库表映射。

IDS层:为中间层，hive 内部表，对ODS层数据进行ETL后的中间结果存储位置。

ADS层:为应用层，hive 内部表，最终应用数据的存储位置。

1. 根据应用系统处理数据表的关联关系
2. 将最终结果转为json 格式的数据表存储在ADS层
3. 将脚本打包到azikaban中，定时执行。

总 结：

1. 在开发sqoop 脚本阶段，需要注意将表中的字段一一写在脚本中，最好不用select \* 方式获取原表数据，目的防止对方的数据结构改变，导致获取到的数据错误
2. 在开发hive 脚本过程中，遇到数据倾斜的问题。首先分析原因是多表关联过程中，产生一对多的关系，导致出现大量的重复数据，导致在reduce阶段计算量大。
3. 原系统虽然查询一条记录会慢一些，但是可以对当天的数据实时查询，而采用该解决方案虽然查询性能上有提升，但是由于sqoop 获取数据不及时，导致离线计算不能够保证数据的即时性 。