大数据开发工程师

基本信息：

姓名：郑钤(zheng qian)　　**性别：**男

学历：硕士　　**工作经验：**3 年

联系电话：17190412671　　**邮箱：**whzhengqian@163.com

教育背景：

2015.9-2018.6	**湖北民族大学**	信息安全	硕士
2009.9-2013.6	**郑州大学**	计算机科学与技术	本科

求职意向：

目标职位：大数据开发工程师　　**期望薪资：**面议

工作地点：深圳　　**到岗时间：**随时

专业技能：

1. 熟悉 Hadoop 生态体系，MapReduce 编程模型原理，HDFS 的读写流程，Yarn 的工作机制及调度策略。
2. 了解 Spark 工作机制，熟练掌握 SparkStreaming 在Yarn 集群模式下的任务提交流程。
3. 熟悉 Scala 编程，能运用 Scala 进行 SparkRDD，SparkStreaming 编程。
4. 熟悉 Flume 、Kafka 等日志收集，分发框架的使用，能够将它们与 Spark，Flink 进行整合进行实时数据的处理。
5. 熟悉Hive的组成架构，Hive 的优化，内部表与外部表，UDF，UDTF。
6. 熟悉 Hbase，理解 Hbase 的存储原理和存储架构，Hbase 的读写流程，Rowkey 的设计原则。
7. 熟练使用 MySQL 数据库，擅长 SQL 语句的编写。
8. 掌握 Sqoop 数据传输工具的使用。
9. 熟悉 zookeeper 的选举机制、常用命令以及集群规模的配备。
10. 了解 Azkaban 的调度流程，极端情况的处理，任务挂了的解决方法：自动重跑和手动重跑。
11. 熟悉 Redis 的基本操作，熟悉 key 操作和五大数据类型的使用。
12. 熟练掌握 Linux 中的Shell 命令以及启动，分发 Shell 脚本的编写。

13. 熟练掌握 JavaSE，了解 JVM、多线程，熟悉基本的数据结构与算法。

工作经历：

2018.3-2021.2　　武汉氪细胞网络技术有限公司　　大数据开发工程师

工作描述：1. 参与数据仓库系统的规划，技术的调研与选型

2. 参与数据采集系统平台的搭建

3. 对实时、离线指标计算和数据服务的性能进行优化建设

项目经验：

项目一：大数据实时分析项目

技术选型：Nginx+MySQL+Canal+Kafka+SparkStreaming+Redis+HBase+ElasticSearch

项目描述：

对 Timing 在线教育 APP 的实时数据产生和流向做分析，后台实时收集数据并统计，对 APP 注册人数，学员播放视频各时长，学员做题正确率与知识点掌握度，付费课程商品页到订单页、订单页到支付页转换率进行实时统计，并将统计结果做成可视化展示出来，实时了解业务指标。

责任描述：

1. 参与数据采集平台的搭建及框架选型；
2. 调研 SparkStreaming 框架，实现 SparkStreaming 对接 Kafka 接受数据以及实时指标计算；
3. 对 SparkStreaming 进行优化，以及解决生产过程中出现的数据倾斜问题；

技术描述：

1. 使用 nginx 做负载均衡器，SpringBoot 做日志服务器，日志数据经 nginx 转发给日志服务器，再发送到 Kafka；

2. 使用 Canal 实时采集 MySQL 中业务数据到 Kafka；

3. 统计日活时，从 Kafka 读取启动日志，借助 Redis 对每台设备启动记录去重，只保留当日首次启动记录，然后把第一次启动记录保存在 ES 以供其他应用查询；

4. 在 SparkStreaming 消费 Kafka 数据时，为实现精准一次消费，采用手动保存偏移量+幂等处理的方案，即将偏移量保存在 Redis，日活明细数据保存在支持幂等的 ES 中；

5. dwd 层是数据明细层，存储事实表的明细数据和维护维度表，并把事实表和维度表做 join，把维度冗余到事实表中，将结果写入 ES；

6. 事实表连接维度表时，一般需查找全部维度表才能完成关联，维度表放 Kafka 不合适，考虑到有些维度表数据量比较大，所以采集 Kafka 中维度表数据放入 hbase 中；

7. 对于事实表与事实表的双流 join 处理中，考虑到会有一方延迟的问题，采用缓存的方案，如果在对方流的同批次中找不到数据，则可以去对方的缓存中查找；

项目二：APP 用户行为分析系统

技术选型： Hive+HDFS+Presto+Sqoop+Azkaban+MySQL

项目描述：

主要基于 Hive 数据仓库之上的数据分析，其中用户基础信息主要包括年龄、地域、性别、职业等信息，网站流量趋势分析，其中包括：访问趋势，新增访客，活跃访客，访问量(每日，每周，每月)。访问信息分析：地域，客户端环境，设备属性，移动终端，网络连接(运营商)。

责任描述：

1. 参与离线数据系统的规划、设计、数仓的搭建；
2. 对系统的调优，对 Hive 进行优化操作；
3. 参与指标分析，实现日活，周活，月活，每日新增，留存用户，留存率， 沉默用户，本周回流，流失用户，连续四周活跃，半月连续 5 天活跃，漏斗分析，重复报课率，报课率排行，GMV 等。

技术描述：

1. 系统采用维度建模的方式，数仓共分为四层：ods 层、dwd 层、dws 层、ads 层；
2. ods 层数据采用 LZO 压缩，减少磁盘存储空间，采用分区表；
3. dwd 层对业务数据采用维度模型重新建模；
4. 通过过滤关键字段将日志数据解析到启动、页面、动作、曝光、错误五张表中；
5. 对于 dwd 层业务数据中极少变化的表采用全量导入方式，变化周期频繁的表采用增量导入，缓慢变化的表则采用拉链表进行处理；
6. dws 层主要对数据做预聚合，形成一些大的宽表如时间维度宽表、地区维度宽表、商品维度宽表、用户维度宽表；
7. ads 层主要是用来存放我们计算的各种指标数据。
8. 利用 Sqoop 将 ads 层数据导入 MySQL。
9. 利用 Azkaban 调度各层脚本；
10. 利用 Presto 即席查询工具，查询 hive 表中的数据。

项目三：用户行为信息采集项目

技术选型：MySQL+Nginx+Flume+Kafka+HDFS

项目描述：

APP 用户在产品的使用中进行的各种操作，包括产品功能的使用、页面的浏览、使用路径等，会产生大量数据，该项目主要是对这些数据进行采集，使用 Flume+Kafka 将日志数据从落盘文件采

集到HDFS，通过 Sqoop 将 MySQL 中的业务数据同步到HDFS。

责任描述：

1. 参与数据采集框架的调研，版本的选型；
2. 参与数据采集系统架构设计和平台搭建；
3. 对各框架的参数进行调优；
4. 处理采集过程中遇到问题框架不稳定挂掉；

技术描述：

1. 使用 Nginx 负载均衡数据，将日志数据发送至日志服务器；
2. 使用 Flume 采集日志服务器中的数据，并发送到 Kafka 集群；
3. 使用 Kafka 对数据进行削峰填谷；
4. Kafka 设置 ack 级别和幂等性对数据实现不丢；
5. 使用 Flume 将 Kafka 中的数据导入到 HDFS；
6. 通过设置参数减少 Flume 消费时出现的小文件；
7. HDFS 本身设置参数对小文件的处理。

自我评价：

1. 对待学习工作认真负责，可以承受工作中的压力；
2. 为人随和，具备团队合作精神；
3. 学习能力强，对新技术有强烈的好奇心；
4. 具有良好的英文阅读能力，能阅读英文资料、技术文档等；