# 简历-大数据开发工程师

|  |
| --- |
| * 个人信息 |

姓名：喻汉生 性别：男

专业：计算机科学与技术（本科） 工作经验：3年

邮箱：[18600770391@163.com](mailto:18600770391@163.com) 手机：18600770391

|  |
| --- |
| * 技术生态圈 |

* 象鼻（数据采集）
* 熟悉使用Flume完成数据采集、分流等功能，理解Flume的组成架构及工作原理。能够搭建多种拓扑结构的数据采集系统，使用Ganglia完成数据监控，并进行传输优化。
* 熟悉Kafka消息队列，理解数据的生产、存储、消费过程，能够搭建高吞吐量，易扩展，高可用架构，提高消费能力，实现Exactly-once语义。
* 熟悉使用Sqoop工具实现关系型数据库与大数据存储框架之间的数据迁移，并解决在数据传输过程中可能出现的数据不一致性问题。
* 熟悉使用Canal实时监控Mysql新增及变化的数据,为实时分析提供业务数据支持。
* 了解并使用DataX完成关系型数据库与大数据存储框架之间的转移。
* 象肚（数据存储）
* 熟悉Hadoop大数据生态体系，能够搭建高可用的Hadoop生态集群。
* 熟悉HDFS组成架构、存储原理、数据的读写流程，及优化HDFS工作性能。
* 理解HBase核心架构，数据的读写流程、存储机制，理解RowKey的设计原则以及HRegion的热点问题。
* 熟悉使用Phoenix为HBase建立二级索引，使其具有类SQL的功能，利于编程，为项目中读写场景建立本地索引和全局索引，加快系统性能。
* 理解Elasticsearch基本原理，熟悉使用Elasticsearch批量保存数据。
* 熟悉Redis内存数据库，能搭建Redis高可用集群并编程，优化缓存穿透问题。
* 熟悉Mysql关系型数据库，能搭建Mysql主从、高可用集群，建SQL索引优化性能。
* 象脑（数据分析计算）
* 熟悉数据仓库的设计、搭建、表关系的模型构建，使用Hive 实现海量结构化日志的分层管理与查询分析。
* 熟悉MapReduce工作机制与运行原理，用Tez替换计算引擎，提升Hive计算查询速度。
* 熟悉使用 Spark Core、SparkSql、SparkStreaming处理离线数据、实时数据，理解Spark 任务提交流程，优化Spark处理过程。
* 理解Flink基本架构，能够使用Flink分析处理有界和无界数据流，以事件为驱动达到最低延迟。能够处理以EventTime为时间准则的数据分析。
* 象皮（数据可视化）
* 了解JavaEE框架(SSM)并能够简单搭建Web可视化工程。
* 了解SpringBoot框架，完成数据接口的发布。
* 了解并使用了的Kibana完成数据可视化展示。
* 象脚（基础）
* 了解常见的Java设计模式和熟悉常用的排序算法，搭建程序结构，优化程序性能。
* 理解JVM模型，常见的垃圾回收器和算法，能够通过调整JVM优化程序性能。
* 熟练掌握Scala编程，能运用Scala进行Spark RDD，Spark Streaming编程。
* 熟悉常用的Linux的Shell命令，能在Linux系统下搭建开发环境，及主机性能分析。

|  |
| --- |
| * 工作经验园 |

公司名称：上海牵趣网络科技有限公司 职位：大数据研发工程师

* 参与项目：
* 丰趣海淘数据仓库分析系统
* 丰趣海淘实时分析系统
* 丰趣海淘数据采集系统
* 主要职责：
* 参与并讨论公司架构选型，技术选取，平台功能规划与设计。
* 搭建公司大数据服务平台，日常维护平台软件，优化性能。
* 负责海淘数仓数据的接入，清洗，业务主题建模。
* 结合业务需求设计数据分析场景，并形成后台实现方案。
* 独立完成业务数据建模，指标分析，并将结果转化为运营指标。
* 优化Hadoop，Kafka,Flume,Spark 参数，实现系统调优。
* 学习新技术，提高整个平台的计算能力和效率。

|  |
| --- |
| * 项目经验区 |

* 丰趣海淘数据仓库分析系统 2017.6 – 至今

开发环境：IntelliJ IDEA+Jdk1.8+Scala2.1+Maven+Git

运行环境：Flume-1.7.0+Kafka-0.11+Hadoop-2.7.2+Hive-1.2.1+Spark-2.1.1+Mysql-5.6+Sqoop-1.4.6+Azkaban-2.5.0

项目描述：主要对“丰趣海淘”App用户行为数据和业务数据库的数据进行存储和分析统计。针对各种用户，商品等多维度去统计分析，根据分析结果优化App。改变不同人群、不同地区的销售策略，帮助提升公司营业额以及市场占有率。

工作职责：

* 参与讨论数据仓库需求分析、分层设计。
* 将用户行为日志数据和业务数据采集至数仓数据原始层。
* 实现产品，运营陆续提出的需求，包括但不限于留存用户、连续活跃用户。
* 持续维护数仓健康状态，不断优化数据仓库性能。

技术要点：

* 数仓结构分析与设计
* 数仓构建为四层:数据原始层（ods）、数据明细层（dwd）、数据服务层（dws）、数据应用层（ads），保证了数据间解耦、分层处理、便于管理。
* 数据原始层（ods）：加载用户行为日志和业务数据，保持数据原貌。
* 数据明细层（dwd）：使用自定义UDF函数对用户行为日志进行解析，并过滤脏数据，对敏感数据加密，对用户重要信息脱敏，对表降维处理。
* 数据服务层（dws）：对数据轻度聚合，并将查询频度较高的表联合起来制作宽表，一表多用，解决大部分需求，提高工作效率。
* 数据应用层（ads）：对dws层的数据从不同维度聚合，结果写入Mysql。
* 数仓优化
* 初期使用Tez引擎，运行提升速度两倍，后期加入Spark-sql，提升数十倍。
* 使用Impala做即席查询，处理工作日常产品、运营提出的临时需求。
* 数仓中通过数据建模理论，对数仓中的数据采用多冗余，低范式构建星型模型，减少关联查询，提高查询效率。
* Hive元数据管理：基于Keepalived实现Mysql HA，每日凌晨后做元数据备份。
* 合理设置小表阈值，在资源允许的情况下尽量使用MapJoin以避免Shuffle。
* 避免小文件，在Map操作之前合并小文件，在任务结束时合并小文件

工作成果：通过对商品各指标、用户行为主题、下单流程各指标进行分析统计，对 app 进行优化，针对不同人群，不同区域改变销售策略，帮助公司获得了更大营收和市场占率。

自我拓展：将Sqoop工具换成了DataX，还用Kylin做了缓存查询。

* 丰趣海淘实时分析系统 2018.4 – 至今

开发环境：IntelliJ IDEA+Jdk1.8+Scala2.1+Maven+Git+Kibana-6.3.1

运行环境：Kafka-0.11+Spark-Streaming-2.1.1+Redis-3.2.5+ES-6.3.1+Spring-Boot-1.5.17+Sqoop-1.4.6+Canal-1.1.2+Mysql-5.6

项目描述：主要对“丰趣海淘”App 用户行为数据和业务数据进行实时分析统计。为实时报表提供数据支持，让产品和运营能够快速做出决策，反映本公司业绩和营收额，供PM及管理人员掌握产品销售情况。

工作职责：

* 参与讨论实时系统的需求分析，架构搭建。
* 实时监控Mysql中新增和修改的数据，为实时分析提供业务数据支持。
* 设计并实现运营及产品提出的实时需求，包括但不限于订单量、销售额。
* 维护实时系统稳定运行，优化实时性能。

技术要点：

* 数据计算
* 使用Canal实时监控MySQL的数据，获取新增以及变化业务数据，Mysql的Binlog的格式选用Row，保证Canal同步到的数据与Mysql是一致的。
* Spark-Streaming消费Kafka数据，手动维护offset，将指标结果入库和offset入库放在同一个事物中，实现精准一次消费。
* 使用updateByKey+checkpoint，实现有状态的累加操作。
* 使用window窗口函数，计算最近一小时内广告点击量Top排名数据。
* 将处理后的明细数据批量写入Elasticsearch， 提供实时检索查询。
* 将需要去重的指标结果写入Redis，利用Set数据结构去重，例如UV。
* 使用transform算子在Driver端周期性读取Redis数据， 通过广播变量发往Executor。
* 实时优化
* 使用MataData Checkpoint 做Driver的容灾恢复，提高程序的稳定性。
* 使用Kryo序列化方式，效率比Java原生序列化方式高十倍左右。
* 使用foreachPartition，将指标结果写入数据库，减少数据库连接次数。
* 配置优雅关闭SparkStreaming，代码升级方便关闭流式程序。
* 使用广播变量+map算子可以实现Map Join，避免Shuffle引起数据倾斜。

工作成果：完成不同时段，不同区域，运营及产品提出的实时需求以及报表展示，为公司管理层对销售策略和广告投放提供数据支持。

自我拓展：项目后期由于调研了Flink技术，将一些需求使用Flink实现。

* 使用Flink CEP库来防止恶意登陆。通过在流中匹配出恶意登陆事件（连续三次登陆失败），来进行报警。
* 使用Flink计算PV，UV等日常指标。
* 丰趣海淘数据采集系统 2016.12 – 至今

开发环境：IntelliJ Idea+Git+Maven+Jdk1.8

运行环境：Flume-1.7.0+Kafka-0.11+Hadoop-2.7.2+Zookeeper-3.4.10+Sqoop-1.4.6+Mysql-5.6+Canal-1.1.2

项目描述：本采集系统为数据仓库和实时分析系统提供数据来源，采集的数据来源于前端埋点的用户行为日志，以及业务数据库中的数据。

工作职责：

* 区分日志类型，根据类型对数据分流，发往Kafka中的不同主题。
* 使用Sqoop将Mysql中的业务数据导入到HDFS。
* HDFS上的数据导入到数仓的数据原始层。
* 维护数据采集系统健康运行，优化系统性能。

技术要点：

* 调整Flume JVM内存参数，增加内存的同时防止内存抖动。
* Flume选用Taildir Source，实时监控日志文件，支持多目录和断点续传。
* 自定义Flume拦截器+Channel选择器实现对数据分流。
* 使用FileChannel防止数据丢失，并配置多目录存储，提高Flume吞吐量。
* 基于时间、文件大小、Event个数三个维度，解决HdfsSink小文件问题。
* 通过调整JVM内存、网络处理线程数、IO处理线程数等优化Kafka性能。
* 使用Sqoop在业务数据库与数仓之间传输数据，并且根据数据变化特性分别采用全量、增量、拉链策略。
* 使用Canal监控 Mysql 业务库数据的实时变动，将数据处理后发送至Kafka实时主题，为实时分析系统提供业务数据支持。

工作成果：通过对商品各指标、用户行为主题、下单流程各指标进行分析统计，对APP进行优化，针对不同人群，不同区域改变销售策略，帮助公司获得了更大营收和市场占率。

自我拓展：自学了ELK日志采集系统，分析了Nginx日志，统计URL访问Top10。

|  |
| --- |
| * 自我评价 |

* 勤奋认真，做事热情积极，较强的责任心。
* 环境适应能力强，能很快适应新的工作环境。
* 具备团队合作意识，善于协调团队中的人际关系。
* 能够虚心接受他人的意见，对符合的予以采纳。
* 始终积极向上，热爱学习。