# 简历

|  |
| --- |
| * 个人信息 |

姓名：喻汉生 性别：男

居住地：北京 工作经验：3年

专业：计算机科学与技术（本科） 应聘职位：大数据开发工程师

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、负载均衡等功能，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。
* 熟悉使用Datax完成关系型数据库与大数据存储框架之间的转移
* 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。
* 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。
* 能够使用Zookeeper，为分布式系统提供协调服务，高可用架构。
* 熟悉HDFS分布式文件系统、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。
* 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 内核，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）
* 熟练掌握JavaSE，深刻理解面向对象设计思想，熟练使用IO流操作和集合框架，网络编程等JavaSE主流技术。
* 能使用反射、注解、动态代理等Java高级技术，对代码的封装抽取及其性能优化有点一定的经验。
* 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linux系统，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 性质：民营

职位：大数据研发工程师 时间：2017.7 – 2019.8

* 主营业务：

海外商品类目多元丰富，包含：母婴儿童用品、美妆个护、食品生鲜、保健品、3C家电、流行鞋包、家居生活等上万款式商品，提供更丰富有趣的生活居家选择。

* 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

|  |
| --- |
| * 项目经验区 |

* 丰趣海淘数据仓库分析系统 2017.9 – 至今

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

运行环境：Flume-1.2.1+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、架构设计。
* 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
* 数仓构建分为四层，数据原始层（ods）、数据明细层（dwd）、数据服务层（dws）、数据应用层（ads），数据解耦、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。
* 数据传输
* 自定义Flume拦截器+Channel选择器对数据分流，发往Kafka不同主题。
* 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。
* Sqoop导出指标结果到Mysql时，使用staging-table解决可能出现的脏数据问题，保证数据一致性。
* 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
* 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：

* 丰趣海淘实时分析系统 2018.2 – 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。
* 配置优化关闭SparkStreaming，防止意外关闭造成数据丢失。
* 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：

* 丰趣海淘数据采集系统 2017.7 – 至今

开发环境：IntelliJ Idea+Git+Maven+Jdk1.8

运行环境：Flume-1.2.1+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：完成了数据采集系统集群的部署，离线部分传输到HDFS，由hive分析处理。实时部分传输到Kafka，由SparkStreaming分析处理。

|  |
| --- |
| * 自我评价 |