

薛鹏瑞-大数据开发工程师-3年

个人信息

姓名：薛鹏瑞

年龄：24

电话：189-0993-7248

学校：辽宁工程技术大学

专业：电子信息工程

邮箱：xuepengrui1@163.com

求职意向

期望职位：大数据开发工程师

期望薪资：面议

期望地点：北京

到岗时间：即时到岗

技术能力

1. 熟悉Hadoop分布式系统, HDFS读写流程以及小文件的处理, 能够搭建Hadoop运行环境, 进行相关参数调优。
2. 熟悉MapReduce运行原理、Yarn的工作机制及调度器, 熟悉shuffle过程及其优化,
3. 熟悉Linux系统, 掌握常用操作命令, 能够进行简单shell脚本编写。
4. 熟悉Flume采集系统的组成架构, 熟悉三大组件(Source、Channel、Sink)的作用, 能够自定义拦截器和使用Channel选择器, 利用Ganglia进行监控。
5. 熟悉Zookeeper的半数选举机制以及常用命令, 使用可视化工具对znode进行操作。
6. 熟悉Kafka基本架构、ISR副本同步队列和相关参数优化, 能够处理Kafka碰到的问题(数据积压、重复等)。
7. 熟悉Hive基本架构及使用, 熟悉常用系统函数、窗口函数, 自定义UDF、UDTF函数步骤及Hive优化, 运用HQL进行相关指标的分析。
8. 熟悉Spark计算引擎架构, 运行模式, 常用的RDD算子及用法, 运用SparkCore、SparkSQL处理离线数据。
9. 熟悉Hbase的存储结构, RowKey的设计原则, 利用Phoenix为Hbase表创建二级索引, 实现SQL查询。
10. 熟悉Redis存储系统的使用, 持久化机制, 哨兵模式以及集群的配置, 利用Redis对数据进行缓存。
11. 熟悉Flink框架的使用, 熟悉Watermark机制, 乱序和迟到数据的处理方式、三种时间语义、状态编程, CEP机制。
12. 熟悉使用Sqoop工具, 完成HDFS与传统关系型数据库(MySQL)之间进行数据的传递。
13. 熟悉Azkaban批量任务调度器的使用, 具体任务流的编写, 以及指标挂掉后的解决办法。
14. 了解Maxwell的工作原理, 以及其在实际工作场景中的应用。
15. 了解Kylin的基本架构, 并在开发中简单的使用查询SQL。
16. 了解Atlas元数据管理, Griffin数据质量监控和Ranger权限管理。

工作经历

易纳购科技(北京)有限公司

大数据开发工程师

2018年11月 - 2021年10月

项目经验

项目一：豌豆公主离线数据管理系统

2019年6月 - 2020年10月

软件架构：Hadoop+Flume+Kafka+Zookeeper+Hive+Sqoop+Mysql+Spark+Hbase+Azkaban+Kylin+Superset

项目描述：本项目通过Hadoop集群来处理豌豆公主跨境电商网站和APP产生的大量日志数据, 为了对这些数据进行集中的处理, 搭建了一个离线数据仓库, 将公司中的业务数据和采集到的行为数据进行汇总, 进行整合统计。主要对数据进行分层、备份、清洗、聚合、统计各项指标提取有价值的信息, 通过这个管理平台来对公司的业务进行改进、提高以及下一步的决策。

责任描述：

1. 负责跨境电商大数据管理平台的分层搭建。
2. 对数据进行ETL, 参与拉链表与主题宽表的设计。
3. 调度脚本的编写和测试。
4. 用户行为模块：分析周回流用户、流失用户、连续3周登录等指标。

5. 商品模块：热门商品按全部商品、类别、区域TopN的分析。
6. 订单分析模块：分析GMV、转化率、品牌复购率、支付区域TopN等指标。

技术描述：

1. 日志数据通过Flume-Kafka-Flume导入到HDFS, 业务数据通过Sqoop导入到HDFS中, 然后将HDFS中的数load到Hive中的ODS层当中。
2. 在双11、618等活动, Flume采集通道抗不住这么大的数据量, 我们通过临时购买阿里云的日志服务器, 然后将Flume部署到日志服务器上。
3. ODS层保持数据原貌, 起到备份的作用, 创建分区表, 防止后续全盘扫描, 采用snappy压缩进行存储。
4. DWD层
 - (1) 基于ODS层数据使用hql进行清洗, 过滤重复和空值数据, 使用UDF和UDTF函数对日志数据进行解析。
 - (2) DWD层对业务数据进行维度退化和降维, 保存维度数据, 主要是对业务事实的描述, 分为访客、商品、优惠券、活动、地区、时间、用户维度, 其中用户表制作成拉链表。
 - (3) 采用snappy压缩和列式存储, 并且对身份证、电话进行脱敏操作。
5. DWS层
 - (1) 基于DWD层数据, 按天进行轻度聚合成宽表, 一行信息代表一个主题对象一天的汇总行为, 例如每个地区当天的下单次数, 下单金额等,
 - (2) 分为用户行为宽表、商品宽表、访客宽表、活动宽表、优惠券、地区表, 统计当天的指标, 导入到DWS层。
6. DWT层以DWS层数据为基础, 对数据进行累计汇总, 一行信息代表一个主题对象的累积行为。例如每个地区最近7天的下单次数、下单金额等。
7. ADS层为各种统计报表提供数据, 统计过日活、月活、周活、留存、留存率、新增、转化率、流失、回流、GMV、复购率、点赞、评论、收藏、领优惠券人数、使用优惠券人数、退款人数、TopN等指标。
8. 使用Sqoop工具将数仓中表导入到MySQL中, 并使用可视化工具Superset进行大屏展示。
9. 搭建Kylin环境, 获取DWD层相应事实表和维度表数据, 构建星型模型, 进行OLAP分析。
10. 使用Azkaban作为全流程调度, 配置邮件、电话报警, 集成了第三方工具, 睿像云来打电话。
11. 业务优化: 创建分区表, 避免全局扫描; 创建分桶表避免数据倾斜; 合理设置reduce的个数; 小文件问题进行相应处理。

项目二：豌豆公主实时数据分析系统

2020年11月 - 2021年10月

软件架构：

Nginx+Flume+Kafka+Zookeeper+Spark+Flink+Mysql+Hadoop+Hbase+Clickhouse+Redis+Maxwell+Phoenix+Sugar

项目描述：仅靠离线计算, 数据时效性无法保证, 通过实时数据分析系统计算当日分钟级、秒级的数据更加利于公司对业务进行快速反应与调整。通过实时计算结果与离线数据结合展示在数据大屏中, 更加直观的看见公司数据变化。

责任描述：

1. 根据业务需求及规划, 进行数据选型,
2. 参与环境搭建, 框架调优
3. 指标分析:
 - (1) 每日日活实时统计
 - (2) 每日订单量实时统计

技术描述：

1. 搭建好的Nginx将前端埋点的日志数据发到日志服务器, 保存30天, 通过Flume发往Kafka的topic中。
2. 实时业务数据采用Maxwell同步历史数据, 获取binlog中的数据信息发往Kafka。
3. ODS层保存原始业务数据和日志数据, 起到备份的作用。
4. DWD层主要功能使用状态编程修复新老访客统计、使用侧输出流完成分流操作, 把分好的流保存到对应表、主题当中。

- (1) 动态分流功能添加一张配置表, 用mysql数据库存储, 使用FlinkCDC读取。
5. DWM层主要功能对部分数据对象进行进一步加工过滤独立访客, 跳出明细计算, 订单宽表, 支付宽表。
 - (1) 跳出明细计算使用FlinkCEP技术, 通过多条数据组合来完成跳出判断。
 - (2) 双流join采用Flink中的基于状态缓存的intervaljoin。
6. DWS层对DWD层和DWM层数据进行轻度聚合统计, 主要对访客主题宽表、商品主题宽表、地区主题宽表、关键词主题宽表的计算, 将聚合结果写到Clickhouse中。
7. ADS层把Clickhouse中的数据根据可视化需要进行筛选聚合, 没有持久化, 前端调用Sugar进行可视化展示。
8. 处理维度关联, 并通过旁路缓存和异步查询对其进行性能优化。

项目三：豌豆公主数据采集平台

2018年11月 - 2019年5月

软件架构：Sqoop+Flume+Ganglia+Kafka+Kafka eagle+Hadoop+Zookeeper+Nginx+MySQL

项目描述：本项目搭建数据采集平台, 将豌豆公主APP和网站所产生的用户行为数据和业务数据进行采集, 方便离线平台使用。使用Flume采集日志服务器中的数据发送至Kafka, 再通过Flume将用户行为数据导入到HDFS, 业务数据使用Sqoop导入到HDFS。

责任描述：

1. 框架版本的选型
2. 服务器选型 (机器成本到运维成本)
3. 估算数据量, 确定集群规模
4. 用户行为数据采集平台搭建
5. 业务数据采集平台搭建

技术描述：

1. 搭建Nginx实现负载均衡, 将前端埋点日志发到不同的日志服务器, 存储为文件, 保存30天。
2. 通过Flume对用户行为数据进行采集, 并自定义拦截器对数据进行简单的ETL清洗, 使用KafkaChannel对接Kafka。
3. 搭建Kafka集群, 配置常用参数, 对Kafka集群拉取数据的速率进行压力测试, 确定Kafka集群的上限。
4. 搭建第二层Flume消费Kafka日志数据, 并发送到HDFS, 优化Flume配置防止HDFS上出现大量小文件。
5. 使用Sqoop工具将业务数据直接导入到HDFS, 并对NULL值存储、数据一致性问题进行存储。

自我评价

1. 独立思考能力强, 具备良好自学能力, 喜欢分析数据, 对计算机编程有浓厚兴趣。
2. 有较强的抗压能力、环境适应能力, 能够快速融入到新的团队当中。
3. 擅长沟通交流, 积极与同事、领导进行交流, 解决工作中的问题。