江润冬_大数据开发工程师_3 年

个人信息

姓　　名：江润冬
性　　别：男
毕业学校：河南科技学院
学　　历：本科
专　　业：计算机科学与技术
年　　龄：27
现住城市：北京
手 机 号：18303620306
邮　　箱：18303620306@163.com
就职状态：离职

求职意向

工作性质：全职
求职地点：北京
意向岗位：大数据开发工程师
薪资要求：面议

工作经验

公司名称：北京果敢时代科技有限公司
部　　门：研发部
职位名称：大数据开发工程师
时　　间：2020.9- 2023.9
汇报上级：项目经理
工作职责：1. 参与框架的选型，集群的搭建，以及数据仓库的建模、分层；
2. 根据业务需求，参与大数据平台开发；
3. 使用 MapReduce，Hive，HQL 来进行离线数据分析；
4. 使用 Flink、Flink SQL 进行实时数据开发。

专业技能

1. 熟悉 Hadoop 生态核心组件，能独立搭建 Hadoop 集群环境，对 HDFS 架构、读写机制、MapReduce 任务切片机制有较好的理解，了解 MapReduce 参数优化方法，能够处理数据倾斜和小文件问题；
2. 熟悉 Linux 开发环境和常用的高级命令，熟悉常用 Shell 命令，可以编写简单的 Shell 脚本（集群启停分发）；
3. 熟悉 Java 开发语言，有一定的面向对象编程和函数式编程功底；
4. 熟练使用 Flume 采集用户日志，能够自定义拦截器实现脏数据的过滤、轻度清洗以及零点漂移问题解决，能够使用监控器监控 Flume 的数据传输情况，熟悉 Taildir Source 的断点续传原理以及 Channel 和 Sink 的场景选型；
5. 熟悉 Kafka 生产者、Broker 及消费者的工作流程，分区分配策略以及 ISR 副本同步策略，集群部署以及参数调优，能够解决 Kafka 宕机，数据丢失，数据重复，数据积压等问题，能够通过相关参数配置提升 Kafka 消费能力；
6. 熟悉 Hive 架构和 HQL 编译为 MR 的原理，熟悉 Hive 中常见系统函数、自定义 UDF、UDTF 函数、窗口函数及 Hive 参数优化，能够处理 hive 数据倾斜问题，能熟练使用 HQL 对数据进行分析；

7. 掌握 DataX 和 Maxwell 同步工具的部署，能够处理 DataX 传输时空值问题，熟悉 Maxwell 底层主从复制原理，熟练使用 Maxwell 实时监控 MySQL 的 Binlog 文件同步增量数据，使用 BootStrap 同步历史全量数据；
8. 熟悉 Hbase 的基本架构、读写流程、以及运用 RowKey 预分区和写入反转 ID 的方式处理数据热点问题。
9. 熟悉 Zookeeper 客户端命令行操作、选举机制、状态监听器原理以及写数据流程；
10. 熟悉 Redis 的五大数据类型及 API、RDB 与 AOF 的持久化机制，能够熟练使用 Redis 进行数据存储；
11. 熟悉 ClickHouse 数据库，以及表引擎 MergeTree，可以使用 Bitmap 表进行存储和查询，能够利用 SpringBoot 实现接口读取 Clickhouse 中的数据提供给 Sugar 可视化平台展示；
12. 熟悉使用 Flink 的常用 API 使用、对乱序和迟到数据的处理方式，能够使用 Flink 对流式数据进行处理、理解 Watermark 机制，时间语义，状态特点，了解检查点算法原理，双流 join 编程。

项目经验

项目名称： “MAMA+”社群电商实时数仓

开发环境：IDEA、Linux、jdk1.8、Maven、Git

项目架构：Nginx、Flume、Mysql、Maxwell、Kafka、Zookeeper、Flink、FlinkCDC、Redis、Hbase、Phoenix、ClickHouse、Sugar、SpringBoot

项目描述：离线数仓的“T+1”隔天反馈结果，无法应对时效性较高的场景，比如：想及时查看广告投放以及 6.18、11.11 活动推广的效果等等，这种场景就需要能够实时的获取统计结果，实时的监控当天的数据，从而为公司运营决策提供数据支撑。基于 Flink 技术搭建实时数仓，统计结果的查询仅有毫秒级的延迟，数据存储在 Kafka、Hbase、ClickHouse 中，将实时数据通过大屏监控，动态精准营销。

项目职责：

1. 对 Flink 框架进行调研学习，根据实际情况对项目流程进行设计，进行技术选型和平台搭建；
2. 将业务数据动态分流，使用 FlinkCDC 监测捕获配置表数据的变化并广播然后与主流相连接；
3. 使用旁路缓存和异步 IO 的方式对读取 hbase 数据进行优化；
4. 使用 SpringBoot 技术搭建数据接口服务，将 Clickhouse 中的数据通过 Sugar 实现实时大屏展示；
5. 特殊场景需求的解决：使用滑动窗口实时展示热门 TopN 商品/配合布隆过滤器优化 UV 计算过程中出现的内存压力过大的问题/使用 Flink CEP 监测用户连续登录异常/使用状态编程实现订单的超时检测。
6. 负责的 DWS 层指标：
   (1) 版本、渠道、地区、访客类别粒度页面浏览各窗口轻度聚合
   (2) 统计用户域 7 日回流用户数以及当日独立用户数；
   (3) 配合智能分词器通过注册自定义 UDTF 函数实现流量域的来源关键词统计；
   (4) 通过 Flink 状态编程结合定时器实现去重，配合旁路缓存和异步 IO 优化实现交易域 sku 粒度下单聚合统计。

技术要点：

1. 实时数仓数据分层为：原始数据层（ODS）、公共维度层（DIM）、明细数据层（DWD）、汇总数据层（DWS）；
2. 利用前端埋点的日志数据通过 Nginx 的反向代理实现负载均衡将数据发送到指定的日志服务器中，Flume 将数据采集到 Kafka 中的 ODS 层中；
3. MySQL 数据库中的业务数据通过 Maxwell 实时监控数据的变化，采集到 Kafka 的 ODS 层中；
4. 通过 Flink 主动消费 Kafka 中的 ODS 层的日志数据，过滤空值，利用 ListState 状态保存首次访问时间；

5. 在 MySQL 中维护一张配置表，利用 Flink CDC 实时监控 MySQL 配置表中的变化，以此实现动态分流，将事实数据写到 Kafka 的 DWD 层，将维度数据通过 Phoenix 写到 HBase 中；
6. 采用双流 join 将订单表与订单详情表进行 join，并进行相关维度的关联；
7. 维度关联利用 Phoenix 的类 SQL 语言查询 HBase 中维度表数据，采用 Redis 旁路缓存及 Flink 的异步 IO 查询进行优化，将处理后的数据写到 Kafka 的 DWS 层；
8. 从 Kafka 的 DWS 层数据读取数据，利用 union 算子将不同来源的流的数据进行合并，指定 Watermark 及提取时间字段，然后对流中数据进行分组、开窗、聚合，将数据写到 ClickHouse；
9. 从 ClickHouse 中读取数据并发布数据接口，利用 Sugar 对数据进行可视化展示。

项目名称： “MAMA+”实时数仓性能监控

项目架构：Prometheus、Pushgateway、Node Exporter、Grafana、睿象云

项目描述：实时流计算与离线批处理计算不同，它需要 7*24 小时不间断运行，实时任务运行的好坏直接影响用户体验，这对实时任务运行情况的监控就显得尤为重要。我们需要对任务运行时的集群资源使用情况以及任务的各类指标进行监控，以便通过大屏及预警通道实时掌握 flink 任务运行情况并能够在必要时及时处理突发预警。

项目职责：

1. 对 Zabbix、Nagios、Prometheus 等监控工具进行调研并作最后的选型；
2. 设计实时监控架构并安装搭建监控平台；
3. 设计集群监控指标，使用 Prometheus 拉取相关指标，并保存到 TSDB 时序数据库；
4. 添加 Prometheus 数据源集成 Grafana，设计看板模型、各指标展示模型；
5. 监控以 Jvm 为主体的各项性能指标，包括：CPU、内存、GC、线程。监控以实时任务为主体的各项性能指标，包括：运行状况、运行时间、重启次数、子任务数量、checkpoint、延迟、缓存命中、Hbase 查询耗时；
6. 在 Grafana 配置 Webhook URL 以集成睿象云，配置告警任务分派策略以及通知策略；
7. 使用时序图在 Grafana 配置任务故障预警规则，指定告警触发阈值，触发延迟以及检测周期。

技术要点：

1. 配置 Grafana 二级联动菜单，实现通过条件筛选查看各任务监控图表。
2. 配置 Status 状态图，添加值映射、阈值变色以及单位配置等功能来监控任务运行情况。配置 Table 表格图，groupby 维度列，对事实列进行聚合计算，来监控任务算子的收发数据量以及背压情况。
3. 对线程不安全的 Counter 指标对象加重入锁，来解决自定义缓存命中率指标时出现的多线程并发问题。
4. 利用 push_time_seconds 指标以及推送网关提供的删除接口配合 OkHttp 网络请求库自制检查清理工具，解决 pushgateway 历史数据沉积的问题。

项目名称： “MAMA+”社群电商离线数仓

开发环境：IDEA、DataGrip、Linux、JDK、Maven、Git

项目架构：MySQL、Flume、DataX 、Maxwell、Kafka、Hadoop、Hive、Spark、MySQL、Zookeeper、Superset、DolphinScheduler

项目描述：搭建数据采集通道，使用 Flume 采集用户行为日志数据到 Kafka 然后通过 Flume 采集到 HDFS 上，使用 Maxwell 和 DataX 将业务数据传输到 HDFS 上面，两条数据线数据上传到集群后用于之后的数据分析，接着通过对数据进行清洗，过滤，脱敏，转换，处理实现数据价值最大化，为公司运营决策提供数据支撑。

项目职责：

1. 参与数据采集平台的搭建，部署配置 Flume、Kafka 组件；
2. 负责制定数仓分层策略、数据域的划分、维度建模方案选择；
3. 负责拉链表思路优化、主题宽表的设计工作以及调度脚本的编写和测试；
4. 参与框架调优，解决 Hive 数据倾斜问题；
5. 参与一些难点指标的编写：“MAMA+”课堂同时在线最多人数统计、使用窗口跨行函数对页面的浏览会话进行划分统计、对用户间隔登录连续天数的统计、对妈妈商学院各类课程的优惠天数的统计；
6. 负责的主题指标：
   交易主题：最近 1/7/30 天各省份的订单交易统计；
   商品主题：最近 7/30 天各品牌复购率统计、三级分类的商品购物车存量 Top10 统计；
   流量主题：最近 1/7/30 天各个渠道的流量统计、最近 1/7/30 天用户访问路径跳转次数统计；
   用户主题：用户变动（流失数/回流数）统计、新增用户留存率分析、新增用户数及活跃用户数统计、用户行为漏斗分析；

技术要点：

1. 根据维度建模理论对数仓采取了如下分层：原始数据层（ODS），公共维度层（DIM），明细数据层（DWD），汇总数据层（DWS），数据应用层（ADS）；
2. 部署 Flume，使用 Taildir source 实现断点续传，自定义拦截器进行数据清洗以及解决零点漂移问题。对 Flume 进行配置，解决 HDFS Sink 对接 HDFS 时产生大量小文件问题；
3. 通过使用 DataX 进行 MySQL 数据库与 HDFS 的数据导入导出，并对空值进行处理；
4. 使用 Maxwell 的断点续传功能对业务数据库 Binlog 文件进行监控，实时进行数据同步；
5. 通过使用 MySQL 替换 Hive 默认元数据库，允许多客户端访问。使用 Spark 引擎替换 MR 引擎，提升查询速度。
6. 合理采用分区技术，防止全表扫面，采用外部表，避免误删元数据；
7. 采用星型模型进行建模，减少 Join 次数，提高查询效率；
8. ODS 层采用 Gzip 压缩，其它层采用 Snappy 压缩、列式存储，减少磁盘空间，提高查询效率；
9. 根据业务情况划分数据域，便于数据的管理和应用：用户域、流量域、交易域、工具域、互动域；
10. 通过 DataX 将 ADS 层的指标分析结果导入到 MySQL，对接 Superset 进行可视化展示；
11. 通过 DolphinScheduler 进行定时任务调度，设置短信，电话报警对失败任务报警通知。

自我评价

1. 性格积极开朗，团队意识强，能换位思考，有团队精神，与团队成员融洽相处，共同进步；
2. 适应能力以及抗压能力强，不谦虚的说学习能力也强，有较强的解决问题的能力，能够很好的完成上级布置的任务；
3. 喜欢钻研各种新技术，私下里喜欢逛博客园、Csdn、GitHub，手动搭建目前已经运行两年多的博客，对于擅长的技术有强烈的探讨和分享的欲望。