令狐岩松

182-9578-8788 | 531599751@qq.com

教育经历

江苏科技大学 - 计算机科学与技术（嵌入式培养） - 本科 2018.09 - 2022.06

* 竞赛：2020全国大学生数学建模江苏省二等奖、五一建模一等奖、扬子江杯一等奖等
* 毕业设计：基于Hadoop实现的音乐推荐系统（SSM+MapReduce实现协同过滤）

实训经历

东软校企合作实训（大数据方向） 2021.09 - 2021.12

疫情封闭期间学校与东软校企合作，为难以出校实习的学生提供实训环境。期间接触Hadoop组件的部署和运维、Python数据分析以及机器学习和数据挖掘的商业应用等，我负责的内容有**基于决策树的用户分类**和**电商文本挖掘**。

项目经历

Flink+Hudi审核机制 - 个人项目

审核机制在用户违规后进行治理，需要跨出具罚单，用户申诉，奖惩结果，全局视角查看明细并且基于明细开发一些指标看板，同时数据分析师做业务分析使用。本项目主要实现审核机制业务的入湖和调度小时表，时效性为30分钟到小时级的准实时。Hudi 数据湖作为基底，通过 Flink SQL 完成从 Kafka 到 Hudi 的入湖工作，打通湖表和 Hive ，最后 Hive 进行小时表调度。

重点: 1、为了数据不延迟写入，选用 MOR 表，调整 Compaction 参数优化；定时调度**离线异步 Compaction** 脚本，避免数据高峰时同步执行 Compaction 阻塞。

2、通过 RateLimiter 构造 UDF 的方式实现**线程级限速**；利用 FlinkSQL的 TopN 原理实现去重写入。

3、针对多流时间差大、维度数据更新不及时的问题，采用 Partial Update ，将更新时的多流 join 在存储端拼接，加速下游查询。

Lambda数仓建设 - 个人项目

场景为电商业务的 Lambda 架构数据仓库，完成 ETL 、看板等功能，统计用户留存率、TopN 等多个指标，设定用户每人每天产生日志约100条，预计高峰数据量20m/s到40m/s左右。

重点:1、对 Hive on Spark 的分组聚合、数据倾斜、多表 join 等进行了调整，将宽表join耗时优化至分钟级；对 Flink的 SQL 、状态和 CP 进行了调优，为可能出现的反压和故障做了预案。

2、使用 Flink CDC 监控配置表变化；针对关联维度表的需求使用旁路缓存和异步IO缓解 jdbc 压力。

3、通过修改Flink源码配置JobListeners、Atlas源码加入Flink-bridge 来调通 Flink 与 Atlas ，**实现Kafka-任务-Kafka的表级血缘路径**。

4、准备转储和抽数两种方法，将离线和实时的宽表按进行**对数**；确保离线实时同一来源取数，各个环节过滤脏数据，保障数据质量。

职位画像与职位推荐系统 - 个人项目

利用用户的投递行为等数据，建立用户与职位之间的画像关系，通过推荐算法进行个性化职位推荐。ETL后，项目主要分为画像工程和推荐工程两部分，以文本分析为基础，pyspark为计算引擎，模拟了简易的实时推荐流程。目前实现了完整的画像工程，推荐工程还在持续探索和更新。

重点: 1、分词后通过职位描述的TFIDF和TextRank，进一步得出职位相似度。关联用户投递行为和职位关键词并去重，根据时间衰减计算用户标签权重，从而构建用户画像。

2、准备了LFM协同过滤和基于职位相似度两种召回推荐方法，再通过LR模型进行CTR预估排序。

3、使用gRPC创建服务端和客户端，模拟ABTest环境。

核心技术栈

实时:

1、熟悉**Flink**运行架构、核心编程，对窗口、watermark、状态编程、容错机制有较为深入了解，对源码中Calcite的解析过程、应用场景有实践经验，对Flink反压、数据倾斜有参数调优经验。

2、熟悉**Hudi**部署与集成、表类型原理和选择，了解Payload机制，运用过0.13版本的部分列更新。

离线:

1、掌握常见场景下的**Hive** SQL的编写，了解**MySQL**的索引机制和优化方法，对Hive on Spark环境下的多种优化方法有实践经验。了解**Hbase**的读写原理和Rowkey设计；了解clickhouse表引擎和去重原理。

2、掌握pyspark基本用法，熟悉**Spark**运行架构、RDD常用算子和执行原理，了解Spark的参数调优、算子调优、RDD优化、Shuffle优化等。

3、熟悉Hadoop生态，了解Hadoop相关组件的常用优化手段，有过MapReduce程序的编写经验。

语言:

1、熟悉 Java SE开发，了解**JVM**类加载机制、GC原理，实际遇到并解决过OOM问题。熟悉常用Linux命令和shell脚本编写，以及数据结构和算法的基本原理。

2、了解基本的python开发、numpy库和pandas库的常用函数，使用过sklearn、tf、jieba等框架进行数据分析。

个人主页

github.com/LH-YS