王贺松

18810901677 1172700121@qq.com https://wanghesong2019.github.io/

意向岗位：大数据高级开发工程师

教育经历

中国科学院大学 2018.09 – 2021.06
应用数学 博士 数学与系统科学研究院 北京

中国科学院大学 2016.09 – 2018.06
应用数学 硕士 数学与系统科学研究院 北京

南阳师范学院 2008.09 – 2012.06
数学与应用数学 学士 数学与统计学院 南阳

- 国家励志奖学金

个人技能

- **高可用平台设计**：主导设计高可用、可扩展的分层数仓架构（ODS-ADS），精通基于 K8s/Docker 的分布式计算框架容器化部署，并具备集群资源调度优化经验。
- **全栈引擎优化**：核心掌握 Flink 实时流处理技术，擅长高吞吐量（TPS）日志处理。对 Spark/Flink 计算引擎执行计划进行调优，具备支撑数百万 QPS 级别应用的性能优化实践。
- **特征工程与治理**：具备大规模多源数据（包括用户行为、媒体数据等）的清洗治理、存储优化及数据偏差修正能力。设计并落地异常检测、预警模型，支撑 AI x 安全等场景的特征工程与业务创新。
- **系统优化与可靠性**：具备丰富的系统稳定性实践经验，包括内存管理、规则动态管理机制设计，以及对 HDFS/YARN 等组件的稳定性调优，保障核心数据服务的连续性与质量。

工作经历

浙江大华技术股份有限公司 2021.07 – 2023.09
大数据开发工程师 杭州

- 主导参与部门离线数仓（基于 Hadoop/Spark 生态）的产品化建设与架构设计。重点解决复杂实体关系建模（如人像、车辆聚档）及多源异构数据的清洗治理，确保数据基座的高可用性与高准确性。
- 作为核心开发人员，将垂直行业（原公安安防）的复杂业务需求抽象化、产品化为通用的实时预警、风控模型等大数据解决方案，推动技术在 AI x 安全等场景的创新应用与快速落地。
- 负责基于客户安防与缉私业务需求，进行高性能定制化项目的开发与交付。具备在严格的技术标准和时间约束下，快速实现业务逻辑建模和系统部署的实践能力。

杭州高新区（滨江）区块链与数据安全研究院 2023.10 – current
资深工程师 杭州

- 主导设计并实现基于 K8s/Docker 的高可用实时数仓环境（Kafka+Flink）。通过容器化部署实现了计算资源的弹性伸缩和快速迭代，为大规模分布式应用提供了高可移植性和稳定性的数据支撑。
- 核心负责网络数据安全运营平台（MSS）的异常流量实时告警模块。基于 Flink/Kafka 实时流架构，设计并优化了高吞吐量特征提取和规则引擎，为 AI x 安全场景提供了亚秒级的异常检测能力。
- 负责基于复杂数据安全需求，主导推进现场定制化风控项目（如“内鬼检测”）的全栈研发与业务落地。成功将垂直领域的业务逻辑抽象为可复用的大数据建模方案。

项目经历

异常账号行为实时检测系统 2024.10 – 2025.03

- **项目背景**：针对数据安全和防泄露需求，构建基于 Flink 的高并发实时检测平台。目标是秒级发现僵尸账户登录、敏感文件频繁下载等关键风险行为。该系统直接赋能 AI x 安全场景。
- **个人职责**：主导高性能、可扩展的实时检测系统架构设计与落地。专注于 Flink SQL 时态表关联性能瓶化解决和数据治理，实现100%准确率的高风险账号识别，保障核心数据资产安全。
- **方案流程**：
  - **技术架构**：采用 Kafka + Flink (SQL/API) + Hive + HDFS 的 Kappa架构。

  - **核心设计**：僵尸账户检测通过 Flink SQL 时态表（Temporal Table Join）实现实时登录流与 Hive 历史/弱口令维表的低延迟关联；频繁操作识别基于 Flink API 滑动窗口聚合，实现按账号 ID 的高吞吐量实时统计与阈值告警。
  - **性能优化**：对 Hive 天表攒批配置进行深度调优（提升批次规模），使小文件数量减少 90%，大幅提升 HDFS 存储效率和后续批处理性能。
- **工作成果**：
  - **系统性能**：系统连续稳定运行 2 个月，告警端到端延迟不大于 1 秒，成功满足实时风控要求。
  - **风险识别**：上线 2 个月内检测异常账号 400+，前 5% 高风险账号经验证准确率 100%。
  - **资源效率**：通过数据治理与优化，实现 HDFS 资源消耗降低 60%，计算成本下降 40%，有效优化了客户硬件资源利用率。

网络风险流量实时告警项目 2024.02 – 2024.08

- **项目背景**：为增强网络边界安全能力，主导构建高并发、低延迟的实时告警系统。目标是对设备服务请求中的 20+种高级网络攻击（如 SQL 注入、文件上传漏洞）进行亚秒级精准检测，直接赋能 AI x 安全风控业务。
- **个人职责**：主导基于 K8s+Flink 的实时流式规则引擎的架构设计与核心开发。专注于解决万级规则的动态加载与高性能匹配难题，通过工程优化提供支撑数十万 QPS 级别实时检测应用的稳定能力。
- **方案流程**：
  - **技术架构**：采用 Kubernetes + Flink + Kafka 容器化栈，实现计算资源的弹性伸缩。
  - **规则动态管理**：创新结合 Debezium CDC（监听规则配置更新）与 Flink 广播状态，实现万级规则的秒级热更新。规则逻辑表达式使用 ANTLR 进行高性能解析。
  - **匹配优化**：为突破计算瓶颈，引入 ForkJoin 框架实现规则逻辑的并行处理，并设计递归剪枝策略，将匹配耗时降低 39%。
  - **鲁棒性保证**：精细配置 Flink Checkpoint 和 Kafka Offset 持久化，保障故障恢复时的状态完整性。
- **工作成果**：
  - **性能突破**：通过并行化和剪枝优化，单条告警匹配耗时从 36ms 降低 39% 至 22ms，日均处理量提升 1.6 倍。
  - **系统可靠性**：规则热更新实现秒级生效，无需重启服务。系统成功实现端到端精准一次性语义，规则恢复完整性 100%，告警重复率 < 0.1%，确保了高风险业务场景的数据准确性。

XX省公安厅视频大数据建模平台项目 2023.03 – 2023.08

- **项目背景**：针对公安缉私总队提出的油气缉私团伙挖掘、窝点发现等复杂业务需求，主导平台建设，旨在整合日均近 4 亿条多模态感知数据（车辆、人像），解决数据孤岛和线索发现滞后（72小时）的痛点，为 AI x 安全场景提供高效、准确的特征工程基座。
- **个人职责**：主导基于 Kafka + Spark + Hive + Hadoop 的高性能混合大数据架构设计与落地。专注于多源异构数据的高鲁棒性清洗与治理，并构建千亿级实体关系和团伙挖掘特征工程，实现业务线索的模型化、产品化输出。
- **方案流程**：
  - **架构设计**：采用数仓主题建模设计规范，设计 ODS/DWD/DWS 三层分层数仓模型，划分 7 大主题域。
  - **数据治理**：开发高鲁棒性 ETL 流程，针对多厂商、多地市接入数据中的 12 类脏数据进行深度清洗，使全链路数据合格率达 99.8%。
  - **特征与模型**：基于 Spark 平台构建人车关联算法，实现了团伙成员关系图谱的构建（平均关联深度达 4 层）；设计动态特征工程模块，自动生成 12 类业务标签。
  - **集群优化**：通过对 HDFS/YARN 参数进行深度调优（如调整最大文件大小），彻底消除了计算节点崩溃等稳定性隐患。
- **工作成果**：
  - **效率提升**：支撑 20+ 缉私行动，线索发现效率提升 80%。团伙挖掘响应时间从 72 小时缩短至 8 小时（效率提升 89%），满足业务快速响应要求。
  - **模型沉淀**：成功构建 3 套标准化技战法模型，形成了可复用的模型资产。
  - **降本增效**：通过冷热分层存储管理等优化手段，数据处理成本降低 60

大华人像聚档算法效果验证系统开发 2022.07 – 2022.10

- **项目背景**：作为人像聚档算法效果评估的核心系统，目标是解决客户对聚档与跨天合档质量的实时巡检痛点。初期方案因跨天合档逻辑缺失，导致搜档率仅 0.5%，且计算耗时高达 1.5小时，无法满足客户实时化（≤30分钟）的评估要求，急需在数据完整性与计算效率间实现平衡。

- **个人职责**：主导搜档逻辑重构与性能优化，专注于解决 Spark 计算中复杂关联的效率瓶颈和数据倾斜问题。设计并落地创新算法（倒序合档），将算法效果评估的耗时大幅压缩，并将其产品化，赋能算法的规模化推广。
- **方案流程**：
  - **架构设计**：基于 Hive 分区表和 Spark SQL/Core 实现核心计算逻辑。
  - **创新算法**：通过问题诊断，确定正序合档的复杂度为平方级，因此创新设计了倒序合档算法，将时间关联次数从 435 次降至 29 次（线性级）。
  - **性能优化**：针对合档流程中的严重数据倾斜（差异超 100 倍），应用哈希分区 + 动态负载均衡策略，将任务平均执行时间缩短 40%；通过优化 Spark 内存配置，彻底消除 OOM及节点崩溃问题。
  - **业务集成**：构建全链路验证流程，将优化后的数据集成至大华云库，支撑客户对算法的实时巡检。
- **工作成果**：
  - **业务价值**：通过逻辑重构，搜档量从 2 万提升至 20 万，搜档率从 0.5% 提升至 5%，精准反映了聚档算法的真实效果。
  - **效率突破**：计算效率从 1.5 小时大幅降至 20 分钟，提升效率 77.8%，成功满足了客户对实时巡检的苛刻要求。
  - **商业落地**：为客户提供可信赖的评估工具，助力大华聚档算法在 20+ 地市实现规模化落地推广。

驾乘数据分析离线数仓项目 2021.12 – 2022.05

- **项目背景**：作为核心数仓开发者，主导公安安防驾乘数据离线数仓的搭建。旨在整合抓拍数据，构建精准数据模型，解决同乘分析、营运车监管、疑似代驾识别等复杂业务场景的数据支撑需求。
- **个人职责**：主导数仓 ODS-DWD-DWS-ADS 全链路分层架构设计与落地。专注于数据质量和流程鲁棒性优化，通过高级 ETL 策略（如重分区、结果反刷）大幅提升数据准确性，并实现业务价值的模型化输出。
- **方案流程**：
  - **架构设计**：基于 Hive on Spark 实现高性能 ETL，采用 Hive 分区表构建数仓。
  - **建模设计**：完成人车关系域、人人关系域等数据域划分，构建总线矩阵。设计 ODS-ADS 四层分层模型。
  - **数据治理**：在 ETL 流程中创新引入抓拍时间重分区机制，有效解决延迟数据漏采问题，使数据延迟统计偏差从 15% 降至 <5%。设计合档关系维表更新流程及结果反刷机制，使跨天合档错误率从 5% 修正至 1%，确保核心指标的准确性与可靠性。
- **工作成果**：
  - **业务价值**：通过逻辑重构，搜档量从 2 万提升至 20 万，搜档率从 0.5% 提升至 5%，精准反映了聚档算法的真实效果。
  - **效率突破**：计算效率从 1.5 小时大幅降至 20 分钟，提升效率 77.8%，成功满足了客户对实时巡检的苛刻要求。
  - **商业落地**：为客户提供可信赖的评估工具，助力大华聚档算法在 20+ 地市实现规模化落地推广。

论文专利

- Dingkang Wang, Hesong Wang, and Fanghui Xiao. An Extended GCD Algorithm for Parametric Univariate Polynomials and Application to Parametric Smith Normal Form, Proceedings of ISSAC 2020, 442-449.
- Dingkang Wang, Hesong Wang, and Fanghui Xiao. An extended GCRD algorithm for parametric univariate polynomial matrices and application to parametric Smith form. Journal of Symbolic Computation Volume 115, March–April 2023, Pages 248-265.
- 基于逻辑语法树的嵌套条件匹配、系统及存储介质. 发明人:王贺松
- 日志数据获取方法、装置、设备及存储介质. 发明人:王贺松
- 告警数据处理方法、装置、设备及存储介质. 发明人:王贺松
- 一种基于实时计算引擎的网络流量日志抓取系统,.发明人:王贺松