

楚佳科

☎ 15544033052 ✉ 15544033052@163.com

👤 27岁 👦 男 📅 5年经验 🎯 大数据开发工程师 📍 河南

教育背景

大连东软信息学院

数据科学与大数据技术 (本科)

2017.09 - 2021.07

专业技能

• 开发基础与运维能力

熟练掌握 **Java**、**Scala**、**Python** 开发语言，熟悉 Linux 常用命令与**Shell**脚本编写；可独立完成任务部署、日志分析、故障排查，具备大数据自动化开发运维能力。

• 数据采集与监控体系

熟练运用 **Kafka**、**Flume** 搭建实时数据链路；依托 **DataX**、**MaxWell**、**Sqoop** 实现多源数据全量、增量同步；熟练使用 **Prometheus+Grafana** 搭建集群监控大盘，监控任务背压、消费 Lag，快速排查链路异常，保障数据稳定传输。

• 大数据集群与计算框架

熟悉 **Hadoop(HDFS/YARN)** 生态架构，具备参数调优、故障排查能力；熟练使用 **Flink** 实时计算，掌握窗口算子、Checkpoint/Savepoint、状态后端、Exactly-Once、反压调优，可完成实时清洗、维表关联、告警指标统计；熟练 **SparkSQL** 离线计算，精通 SQL 调优，能够处理数据倾斜、热点 key 问题，优化离线任务执行性能。

• OLAP 引擎与数仓建设

熟练使用 **Doris**、**ClickHouse** 主流 OLAP 引擎，能根据业务场景选型适配，支撑海量数据多维极速分析；深入理解 **ODS-DWD-DWS-ADS** 数仓分层设计，可独立主导数仓建模，具备统一指标口径、提升数据复用性的落地经验。

• 数据库与存储

熟练使用 **MySQL**、**HBase**、**Redis**；掌握慢 SQL 定位与索引优化；熟悉 HBase RowKey 设计，利用 HBase 实现海量历史数据归档、高并发时序数据读写。

• ETL 治理与任务调度

熟练基于 SQL 完成 ETL 清洗脱敏、脏数据治理、数据质量校验，统一业务指标口径；熟练使用 **DolphinScheduler** 实现任务编排、依赖调度、失败告警，完成大数据任务全流程自动化。

• 可视化开发

熟练使用 **FineBI**、**FineReport** 制作可视化大屏与业务报表，支撑政务、车联网等业务可视化决策分析。

• 风控特征工程

了解 **WOE/IV**、**PSI** 特征检验、缺失值与异常值处理；熟悉 **LightGBM** 等算法，掌握 AUC、KS、召回率、精确率等模型评估指标，适配信贷风控场景数据开发。

• AI 辅助开发

熟练使用 **Claude Code**、**Trae** 等大模型工具辅助代码编写、报错排查、方案梳理；了解向量库、**RAG**架构，可完成大模型相关数据侧应用落地。

工作经历

东华智慧城市科技
成都优卡数财信息科技

大数据开发
大数据开发

2025.09 - 至今
2021.07 - 2025.08

项目经验

智慧城市数据中台项目

2025.09 - 至今

项目描述：

区级政务信息化重点项目，汇聚公安、城管、民政、市监等多委办局**异构数据源**，搭建**政务离线数据中台**。通过数仓建设与**ETL 数据治理**破除**数据孤岛**，统一指标口径，为城市治理、民生统计、大屏可视化提供离线数据底座。

技术架构：

Hadoop、Hive、Spark SQL、DataX、MySQL、DolphinScheduler、Prometheus、Grafana、Fine BI

个人职责：

1. 搭建**ODS/DWD/DWS/ADS 四层离线数仓**，使用**DataX**实现业务库 T+1 全量、增量同步，完成海量业务明细数据入库。
2. 负责政务**ETL 治理**，完成清洗、脱敏、标准化，规整**千万级存量脏数据**，统一业务口径。
3. 业务域分层建模，产出**20 + 主题宽表**，通过**Spark SQL**批量加工政务统计指标。
4. 基于**DolphinScheduler+Shell**编排调度工作流，维护**50 + 离线任务**，管控依赖与重跑，保障 T+1 准时产出。
5. 落地**数据质量校验规则**，巡检修复脏数据，提升数据完整度与准确率。
6. 对接**FineBI**输出指标，支撑报表与大屏展示，完成口径核对及项目验收。
- 7.使用**Prometheus+Grafana**搭建监控大盘，配置任务失败、延迟告警，快速定位集群调度故障。

项目成果：

- 1、接入 12 个委办局，汇聚**80 + 数据表**，构建**20 + 主题宽表**，完成**千万级政务数据治理**。
- 2、实现**T+1 自动化更新**，替代人工统计，打通跨部门数据壁垒。
- 3、稳定支撑政务日常报表统计、**城市治理态势分析**及大屏展示；完善任务监控告警手段，保障离线任务稳定运行，项目顺利通过官方验收，长期稳定落地运行。

新能源车辆车联网大数据监管平台

2024.09 - 2025.08

项目描述：

平台面向物流车队、网约车运营企业，对接新能源车车辆车载Tbox终端，持续采集车辆**定位、电池状态、行驶轨迹、故障告警、能耗、驾驶行为**等海量时序数据，搭建**实时 + 离线一体化大数据平台**。依托实时链路实现车辆状态监控、**危险驾驶秒级预警**，基于离线数仓完成动力电池健康分析、车队运维数据统计分析，支撑多家车队数字化管理系统稳定落地，日均处理车载实时上报数据**9000万条**。

集群架构：

Hadoop、Hive、Spark SQL、Flink、Kafka、DataX、DolphinScheduler、HBase、Redis、MySQL

个人职责：

1.实时数据流开发

Flink消费**Kafka**车载消息，过滤**GPS 漂移**、清洗报文，识别**危险驾驶**；**历史轨迹**存入**HBase**，**热点车辆状态**写入**Redis**；做简易**SpringBoot**查询接口封装，对外提供轨迹、车辆状态查询服务，供业务与大屏调用。

2.四层离线数仓搭建

使用**DataX**同步 MySQL 中车辆、车队档案业务主数据；基于**Hadoop+Hive**构建**ODS/DWD/DWS/ADS 四层数仓**，借助**DolphinScheduler**调度，使用**Spark SQL**完成离线计算。

3.车队运营指标开发

设计日均里程、电池衰减系数等 30 + 车队运营指标，分区分桶优化解决数据倾斜、小文件问题。

4.数据治理与质量校验

制定时序数据校验规则，监控缺失、越界、重复数据，统一指标统计口径并推送异常报表。

5.任务调度与平台运维

依托DolphinScheduler调度离线任务，同步ADS层数据至MySQL；在Linux环境排查实时 / 离线任务故障，保障告警与T+1报表稳定输出。

项目成果：

- 1.搭建车联网实时 + 离线混合链路，整合时序与业务数据；
2. 优化存储策略解决小文件痛点，提升查询响应效率；
3. 危险驾驶秒级预警，降低合作车队事故率；标准化指标支撑20余家车队平台上线；
4. 沉淀可复用的数据清洗、ETL脚本模板，减少重复开发工作，提升新项目交付效率；

银行信贷客户风险识别数据中台

2023.03 - 2024.08

项目描述：

面向商业银行零售信贷业务，整合客户信息、贷款流水、征信、司法等多源异构数据，搭建信贷风险识别数据中台，构建客户准入、贷中监控、贷后预警指标体系，支撑消费贷、经营贷风控业务，日均处理数据 5000 万条。

集群架构：

Hadoop、Hive、Spark SQL、DataX、Kafka、Maxwell、ClickHouse、MySQL、DolphinScheduler

个人职责：

1.多源信贷数据采集开发

Maxwell同步业务库 binlog 增量数据，消费Kafka信贷申请日志落 ODS 层；DataX同步征信、司法离线数据，Shell脚本清洗异构数据，统一客户、账户维度口径。

2.分层信贷数仓搭建

搭建ODS/DWD/DWS/ADS四层风控主题数仓，采用Spark SQL优化计算逻辑，任务运行效率提升 43%。

3.信贷风险指标开发

开发负债、多头借贷、司法风险等 45 + 风控指标，利用窗口函数解决数据倾斜，加工客户风险标签，支撑信贷审批与贷后预警。

4.数据治理与质量监控

配置字段校验规则，梳理信贷指标字典，统一统计口径，监控异常数据并推动整改，保障风控数据准确。

5.报表接口与调度运维

ADS指标同步至ClickHouse、MySQL，支撑风控看板；通过DolphinScheduler调度离线任务，排查故障延迟，保障T+1风险报表按时产出。

项目成果：

- 1.打通内外部信贷数据源，搭建完整客户风险数据体系；
2. 离线任务效率提升 43%，报表查询耗时下降 38%；
3. 标准化指标体系，减少重复开发，提升开发效率；
- 4.建立数据质量校验机制，辅助识别高风险客户，降低信贷逾期风险。

多平台电商经营数据中台

2021.07 - 2023.03

项目描述：

公司面向淘系、拼多多、抖音电商商家提供数字化运营服务，整合多平台店铺订单、流量、广告投放、会员、仓储ERP异构数据，搭建分层离线数仓，统一电商数据口径，构建商品、流量、转化、广告投放、库存相关指标体系，支撑商家经营后台、运营报表、商品销售分析场景，日均处理数据 7000 万条。

集群架构：

Hadoop、Hive、Spark SQL、DataX、Kafka、MySQL、DolphinScheduler

个人职责：

1.多源数据采集开发

用DataX同步电商业务库全量与增量数据，消费Kafka实时流量日志落地ODS层；编写Shell脚本规整多平台异构数据，统一渠道字段格式。

2.分层电商数仓搭建

搭建ODS/DWD/DWS/ADS四层电商主题数仓；以Spark SQL替代原生Hive执行计算，运行效率提升45%。

3.电商核心指标开发

设计ROI、复购率、库存周转天数等40+运营指标；通过分区分桶、多表关联与窗口函数SQL解决数据倾斜。

4.数据治理与质量监控

配置数据校验，核验订单金额、访客唯一性等维度；梳理指标字典，统一多渠道统计口径并推送异常数据整改。

5.报表接口与调度运维

把ADS层指标同步至MySQL支撑可视化报表；依托DolphinScheduler调度离线任务，排查任务故障与数据延迟，保障T+1报表准时产出。

项目成果：

- 1.打通淘系、拼多多、抖音多渠道数据，破除商家数据孤岛；
2. 标准化指标体系，支撑上百家商家经营分析与平台运营决策；
3. 沉淀通用脚本与调度模板，提高新项目开发效率，降低重复开发成本；
4. 搭建质量监控与校验机制，增强数据稳定性，提升数据开发与运维效率。

荣誉证书

- FCA的FineRepor认证
- FCA的FineBI认证

自我评价

- 良好的沟通能力和协调能力，能够与多方合作解决问题。
- 有较强的自学能力和良好的语言表达能力，具备良好的团队协作精神。
- 擅长任务故障排查，重视数据质量，习惯整理文档资料，可对接业务、算法人员统一指标口径。
- 抗压与沟通协作能力良好，善于借助工具提升工作效率，主动学习大数据相关技术。