

# 程诚

电话：13409736469 | 邮箱：13409736469@163.com  
个人博客：<https://blog.csdn.net/TomAndersen> | Github：  
<https://github.com/TomAndersen0714>  
27岁 | 男  
当前状态：在职 | 意向岗位：大数据开发工程师

## 专业技能

- 熟悉 Python 开发，了解 Pandas、NumPy、Argparse 等常用开发模块，以及大数据组件 API
- 熟悉 Java、Scala 开发，了解 JVM 基本原理，了解 Collection、Stream 等常用开发框架
- 熟悉 Airflow 工作流开发，熟悉 Airflow ETL 任务及其 Operator、Hook 等自定义插件开发
- 熟悉 Azkaban 工作流开发，了解 SeaTunnel ETL 任务及其 Source、Transform、Sink 等自定义插件开发
- 熟悉 Hive SQL、Spark SQL 开发及其性能调优，熟悉 ClickHouse SQL 开发及其性能调优
- 了解 Pulsar、Kafka 消息队列，具备相关应用开发经验
- 熟悉 AWS Deequ 数据质量校验工具，具有相关源码开发经验
- 熟悉 Hadoop、Hive、Spark、ClickHouse、Airflow 等组件底层基本运行原理
- 熟悉 CDH、ClickHouse、Airflow 等大数据组件的部署和运维
- 熟悉数仓搭建方法论，熟悉数仓优化和开发规范
- 了解数据治理理论，具有数据治理应用相关开发经验

## 职业证书

1. 计算机技术与软件专业技术资格（中级-软件设计师）
2. 银行业专业人员职业资格（初级-个人理财）
3. DAMA 数据治理工程师（CDGA）
4. CDA 数据分析师（Level I）
5. 英语（CET-6）
6. 普通话（二级乙等）

## 教育经历

华中科技大学 985	2019年09月 - 2021年06月
计算机技术 硕士 计算机科学与技术学院	
中国地质大学（武汉） 211	2015年09月 - 2019年06月
空间信息与数字技术 本科 计算机科学与技术学院	

## 工作经历

深圳前海微众银行股份有限公司（微众银行）	2023年11月 - 至今
数据开发岗 武汉研发中心/开发四室	武汉
1. 反洗钱业务系统的数据 ETL、机器学习模型部署等数据处理类需求开发	
2. 反洗钱数据探查、取数等数据分析类需求开发	
3. 反洗钱业务系统中计算资源、存储资源、数据质量等数据治理类需求开发	
4. 反洗钱业务子系统线上生产问题排查和处理	
成都晓多科技有限公司（晓多科技）	2021年07月 - 2023年11月
大数据开发工程师 数据平台组	成都
1. 负责特定业务线的所有数据需求开发，其中包括 ETL 任务开发、数仓搭建、数据报表搭建、数据大屏搭建等。	
2. 负责组内大数据开源组件和 CDH 集群运维，保障集群运行稳定。负责数据质量治理及监控工具开发，编写数据开发规范和数据治理操作文档。	
3. 负责组内大数据任务开发工具研发，提升组内研发能力和效率。	
4. 获奖情况：2022 年度奖项-突飞猛进奖-CEM 产品线团队（核心成员）、2021 年度奖项-最佳协作奖（成员）、2021 Q3-创新探索项目奖（成员）	

项目经历

反洗钱大额和可疑交易报告报送

2023年11月 - 2024年08月

- **项目介绍**：根据中国人民银行令《金融机构大额交易和可疑交易报告管理办法》等相关规定，境内依法设立的指定金融机构需要依照规章要求，向人民银行上报大额交易和可疑交易。为支持反洗钱业务侧进行大额交易、可疑交易的识别、审批和上报，本项目通过规则模型、机器学习模型等方案，实现了大额交易、可疑交易的自动化识别，同时支持业务侧根据业务场景自定义调整模型规则。
- **项目角色**：数据开发
- **技术栈**：Azkaban (WTSS)、SeaTunnel (Blanca)、Hive、Spark 等
- **负责内容**：
  - 可疑交易客户机器学习模型部署：与中台组对接各可疑交易机器学习模块，将模型的数据处理流程通过规范化、数据模型优化、性能优化后转换为大数据 ETL 工作流实现工程化部署，接入到现有的反洗钱业务系统中，用于圈定可疑交易客户案例。
  - 可疑交易客户规则模型基础指标计算：按业务需求开发对应的基础指标并注册到规则模型中，以支持业务侧根据不同的业务场景使用基础指标创建自定义规则模型，圈定可疑交易客户案例。
  - 计算资源和存储资源治理：历史任务重构迁移，迁移重构线上历史任务适配新版本大数据任务客户端工具，解决历史包袱减少后续开发和维护成本，同时，针对工作流中的所有任务资源进行分档精细化配置，降低工作流运行时整体资源开销。计算资源、存储资源治理，针对高资源开销任务、慢查询任务进行性能优化；按照数据表和数据处理任务的资源占用指标确定治理优先级，并通过设置数据表生命周期、定期归档冷备数据、性能优化等方式进行资源治理。
- **项目难点及解决思路**：
  - TB 级数据处理任务性能优化：通过 WebUI、日志等工具定位具体病灶代码段，采取任务拆分、谓词下推、重复查询合并下沉、数据缓存、调整并行度等方式进行性能优化，必要时还可以尝试从原始业务需求的角度出发并结合业务数据的特征。
  - 异地协作沟通难：异地协作中，提前准备介绍文档并做好会议纪要，提升效率，减少沟通成本。会中做好相关会议纪要，对齐约定内容，尽量减少会后碎片化沟通，避免出现约定不一致。

大数据集群计算和存储资源治理

2023年03月 - 2023年06月

- **项目介绍**：随着内部的数据相关业务和数据量的不断增长，大数据集群的数据量以及负载的数据查询请求数量不断增多，为了降低服务器资源开销和数据运维成本，本项目通过开发数据治理基础工具，以及针对计算和存储资源开销、数据质量进行治理，通过下线无效数据表、合并下沉数据处理分支、优化高负载查询、元数据指标的自动监控和告警等手段，提高了数据运维效率，降低了服务器资源开销和运维成本，节约了集群磁盘空间约 32%，CPU 和内存平均占用各下降约 13%，生产环境集群资源告警从约 20 次/天下降至约 3 次/天。
- **项目角色**：开发负责人
- **技术栈**：Airflow、ClickHouse、Impala、Kudu、CDH、Grafana 等
- **负责内容**：
  - 数据治理基础工具开发：扩展 Airflow 配置和功能，支持调度任务消息同步至飞书群，实现了线上定时任务执行过程的实时监控和告警。
  - 计算资源治理：开发 Airflow 定时调度任务，从 ClickHouse 系统表中获取 SQL 查询日志，采集扫描数据量、CPU 时间、内存开销等元数据，针对高资源开销任务，进行实时告警和性能优化。
  - 存储资源治理：与各数据下游业务线数据产品负责人协商敲定数据生命周期配置，开发调度任务，读取生命周期配置清理过期历史数据；合并下沉数据处理分支，实现数据复用，减少冗余存储；基于 ClickHouse 系统表元数据统计和监控各数据表近期访问次数，并告警无用数据表，同对应责任人敲定下线时间。
  - 数据质量治理：基于业务线指标统计规则，开发相应的数据测试用例旁路工作流，通过定时执行数据测试用例，实现数据质量的在线监控，尽早发现和处理数据质量问题，减小下游数据质量修复成本和影响范围。
- **项目难点及解决思路**：
  - 跨部门协作难推进：通过排定整改事项及价值点，协调人力和资源，定期会议对齐进度，识别并暴露风险点，必要时进行升级，确保项目整体进度。
  - 性能优化方案实施时需保证数据质量：在针对数据处理任务进行性能优化时，为保证优化前后数据一致性，需要做好数据质量校验，通过采取数据双写和旁路校验的方式，做好生产环境预发布验证，避免正式上线切换后产生数据质量问题。

电商客服服务质量可视化大屏

2021年12月 - 2022年06月

- **项目介绍**：为助力电商企业，帮助提升客服团队的服务质量，本项目基于客户的实时会话消息流，开发搭建了客服接待实时监控数据大屏，实现了客服接待过程中各项数据指标在约 10w qps 下的秒级实时计算、更新和可视化，提供了针对客服接待质检内容的统一在线实时监控、实时告警等能力，成功赋能客户帮助企业大幅提高了各部门客服团队的服务质量。截止项目交付，已为 100+ 企业，1500+ 店铺提供了相关服务，满足了客户针对在线接待过程的实时监控需求。
- **项目角色**：数据仓库开发工程师
- **技术栈**：Airflow、Pulsar、xdvector、ClickHouse、DataForce (BI)、EChars 等

- **负责内容：**

- 数据集成：对接上游业务系统，通过内部流数据处理平台 xdvector 开发 Pulsar 实时消费应用，对客服实时聊天会话、会话打标等明细数据进行实时预处理后写入 ClickHouse Buffer 表后缓冲入库，搭建实时数仓的 ODS 层；开发 Airflow 任务，定期同步各企业组织架构、员工信息等维度数据到 ClickHouse Replicated 表中搭建实时数仓的 DIM 层。
- 实时数仓搭建：采用 ClickHouse Replacing 表存储会话、消息记录等明细数据，通过后台数据合并、Final 机制以实现明细数据近实时更新，同时定时聚合统计会话量、消息告警等会话质量指标，分别搭建 ODS/DWD/DWS 数据层。
- 数据可视化：基于企业自研平台 DataForce，以 ClickHouse 作为数据源搭建数据大屏，实现客服接待相关数据指标的近实时更新和大屏数据可视化。

- **项目难点及解决思路：**

- 近实时数据更新技术方案设计：通过数据探查 EDA 摸清数据画像，评估好数据日增量和数据特征；基于现有平台和工具的特性，通过开源社区方案调研、内部评审，以及测试环境可行性、压力测试，确定技术方案的实现路径，并将相关技术文档分享给组内，实现技术沉淀。
- 历史数据查询出现性能问题：在技术层面，针对历史数据开发预聚合任务进行定期合并 DWD 层明细数据，预聚合写入 DWS 层，近期数据查明细表，历史数据查统计表，两者结合实现大时间范围的查询加速；在产品层面，通过和业务侧敲定功能模块的数据更新、查询操作的时间范围。

## 电商客服服务质量数据报表

2021年11月 - 2022年02月

- **项目介绍：**为提高电商企业的客服服务质量，优化电商企业数据分析师的工作效率，本项目基于上游客服接待质检平台的质检结果，通过构建离线数仓 T+1 离线统计 100+ 客服服务质量关键数据指标，实现了不同模块的质检数据报表自动化产出。截止项目交付，已为 400+ 企业、3000+ 店铺客服团队的服务质量评估、绩效考核、客服能力评估和培训、客服数据分析报告等工作内容提供了数据支撑和流程提效。

- **项目角色：**数据仓库开发工程师

- **技术栈：**MongoDB、xdvector、ClickHouse、Airflow、DataForce (BI) 等

- **负责内容：**

- 数据集成：对接上游客服质检业务系统，开发 Airflow ETL 任务，定期同步企业组织架构、客服员工、客户子账号等维度数据，搭建 DIM 维度数据层，增量同步 MongoDB 中的客服会话、会话质检标签等数据，搭建 ODS 原始数据层。
- 离线数仓搭建：基于 ODS 层中当天客服会话质检结果等明细事实表，并与历史数据拉链合并更新后写入明细数据层 DWD。拆解需求指标构建总线矩阵，将相同的数据域和指标粒度进行汇聚，实现指标聚合统计以构建数据汇总层 DWS，然后按照不同功能模块、应用，将 DWS 层中的指标进行汇聚和个性化处理，形成数据应用层 ADS。
- 数据可视化：基于内部自研的低代码可视化平台 Dataforce，搭建数据报表的可视化看板按不同维度展示对应的数据指标，并展示至上游业务系统中。

- **项目难点及解决思路：**

- 新人需尽快熟悉相关业务和技术：业务线快速扩展时，需新人尽快熟悉业务与研发流程，通过查阅业务线对应产品知识库中的文档、视频等相关资料，以及邀约相关人员咨询，先摸清整体框架快速启动，具体细节在后续具体开发过程中逐步填充。阅读源码时，通过编写技术方案文档、使用 UML 图表工具等方式描述代码结构和执行流程方便记忆和共享交流。
- 历史包袱重需要代码重构：针对日常开发过程中发现的历史包袱，需要做好记录，并周期性按照不同的主题汇聚并上报如实体现其价值和重要性，并在后续排期中按照事项逐步优化和迭代，避免自己的任务过于零碎不够聚焦，上报时无法体现自己的产出和价值。