程诚

电话：13409736469 丨邮箱：13409736469@163.com
个人博客：https://blog.csdn.net/TomAndersen 丨Github：
https://github.com/TomAndersen0714
27岁 丨男
当前状态：在职 丨期望职位：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，熟悉 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
- 了解 Hadoop、Hive、Spark、ClickHouse、Airflow 等组件底层基本运行原理
- 了解 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

- 计算机技术与软件专业技术资格（中级-软件设计师）
- 银行业专业人员职业资格（初级-个人理财）
- DAMA 数据治理工程师（CDGA）
- CDA 数据分析师（Level I）
- 英语（CET-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 2023年11月 - 至今

数据开发岗 武汉研发中心/开发四室 武汉

- 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
- 反洗钱数据探查、取数等数据分析类需求开发
- 反洗钱业务系统，计算资源、存储资源、数据质量等数据治理类需求开发
- 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 2021年07月 - 2023年11月

大数据开发工程师 数据平台组 成都

- 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。
- 负责组内大数据开源组件运维，保障集群运行稳定。负责数据质量治理工具开发，编写数据治理规范和操作文档。
- 负责组内大数据开发工具研发，提升组内数据开发能力和开发效率。
- 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、机器学习模型等技术方案，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、性能优化转换为 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。
  - 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，以支持业务侧使用指标创建自定义规则模型。
  - 计算任务和资源治理：迁移历史 ETL 任务统一各任务使用的脚手架工具，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 ETL 任务资源配置进行分档精细化控制，降低工作流运行时整体资源开销。
- **项目难点及解决思路**：
  - 任务性能优化问题：在针对 ETL 任务进行性能优化时，除了从代码结构和调参的角度进行优化外，还应该从原始业务需求的角度出发并结合业务数据的特征，通过重写代码逻辑来优化整体性能，而不拘泥于原逻辑。
- **项目总结与反思**：
  - 异地协作中的沟通问题：异地协作相比同地协作，沟通起来难度较大，每次启动交流前，应该做好上下文背景介绍，提升沟通效率，减少理解成本。同时，还应该利用线上文档信息共享，分批次做好信息同步，减少碎片化沟通次数，降低远程沟通成本。

数据计算存储资源和数据质量治理 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也不断增多，为了降低服务器资源开销和数据运维成本，实现大数据集群的降本增效，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，最终实现了高负载查询、无效数据表、数据质量问题报告的自动生成，提高数据运维效率，降低了服务器资源开销和运维成本。
- **项目角色**：开发负责人
- **技术栈**：Airflow、ClickHouse、CDH、Impala、Kudu、HDFS 等
- **负责内容**：
  - 数据治理工具开发：开发 Airflow 插件，支持定时任务调用插件发送消息到飞书群。修改 Airflow 源码，将 Airflow 告警信息接入飞书群，实现线上定时任务的实时监控和预警信息采集统一。
  - 数据计算资源治理：开发 Airflow 定时任务，定期查询 ClickHouse query-log 数据，同时自动生成高负载 SQL 统计报告并发送告警群，为性能调优提供数据支撑和优化方向，进而降低 ClickHouse 集群的 CPU 和内存资源开销。
  - 数据存储治理：与下游业务线数据产品负责人协商敲定数据生命周期信息，通过 Airflow 定时任务，读取表的生命周期配置，及时清理过期历史数据；通过定期查询 query-log，统计并自动报告各数据表访问次数，通过与业务方沟通对齐后，将近期内无访问记录的数据表和任务下线。
  - 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，通过 Airflow 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，降低下游开发成本。
- **项目难点及解决思路**：
  - 跨组工作事项难以推进：通过提前确定好待办事项，然后沟通预约相应部门负责人，确定该项目的参与人员与工作事项大致估时和排期，确保足够的资源分配；通过定期小型会议的形式，及时同步和对齐项目进度，遇到无法解决的卡点和难点时，需要及时向上反应，报备风险，必要时进行事态升级，保障项目整体进度。
- **项目总结与反思**：
  - 项目过程中需要做好过程指标采集，为最后项目成果汇报的量化部分提供数据支持：任何项目在最后都需要进行项目成功总结和汇报，项目开发过程中不能只关注项目进度，还需要注重项目过程指标的采集，避免某些数据因为生命周期较短（如服务器日志数据），错过采集时间后影响后续成果量化。

电商客服服务质量监控大屏 2021年09月 - 2021年11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。截至目前，已成功为 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**：Pulsar、xdvector、ClickHouse、Airflow、DataForce 等

- **负责内容**：
  - 数据集成：对接上游业务系统，通过企业自研的流数据处理平台 xdvector 开发 Pulsar 流实时消费程序，对客服实时会话日志数据进行实时 ETL 后写入 ClickHouse Buffer 表中，保证数据秒级预处理和写入，进而构建实时数仓 ODS 和 DWD 层，同时通过 Airflow 定期调度数据同步任务将各企业组织架构、员工信息等维度数据到 ClickHouse Replicated 表中，构建实时数仓的 DIM 层。
  - 实时数仓搭建：采用 ClickHouse ReplacingReplicatedMergeTree 表存储会话数据，通过数据副本保证高可用的同时，支持数据按照指定会话 ID 进行后台数据合并，搭配 Merge On Read 机制以实现数据的近实时更新，最后通过 ClickHouse Distribued 表提供数据的实时数据查询和数据分析服务。
  - 数据可视化：基于企业自研的低代码可视化平台 DataForce，以 ClickHouse 作为数据源，通过前端 EChars 组件，实现客服接待相关数据指标的实时计算和可视化。

- **项目难点及解决思路**：
  - 基于 ClickHouse 的近实时数据更新方案的设计与实现：通过技术调研，以及线下测试环境的性能对比测试，最终确定技术方案的实现路径，并将相关技术文档分享组内，实现技术沉淀。
  - 历史明细数据实时查询时会因时间范围过大而出现性能问题：在产品层面，通过和产品达成一致，及时约束数据更新操作的时间范围；在技术层面，针对历史数据创建任务进行定期聚合 DWD 层明细数据，生成统计指标写入 DWS 层，加速查询。

- **项目总结与反思**：
  - 数据需求分析阶段，理应做好数据探查 EDA 相关数据分析工作，评估数据日增量，以及后续数据查询性能，提前和业务侧确定产品方案细则，避免产品功能发布后出现变动和违约，影响客户使用体验。
  - 项目完结时，要做好相关文档沉淀，技术方案的调研和设计，需要及时同步到组内知识库，必要时进行宣传和共享，与组内成员共同交流和进步。

电商客服服务质量数据报表 2022年02月 - 2023年03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，截止目前，已成功为 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 等
- **负责内容**：
  - 数据集成：对接上游客服质检业务系统，进行数据离线 Airflow ETL 任务开发，定期同步企业组织架构、员工信息等配置数据，作为离线数仓的公共维度层 DIM ，T+1 增量同步 MongoDB 中的客服会话数据，经过反规范化后处理后与历史数据合并更新，作为搭建离线数仓的数据接入层 ODS 和明细数据层 DWD 。
  - 离线数仓搭建：基于 DWD 层中客服会话质检结果等明细事实表，以及 DIM 层相关的商家配置维度表，将不同的数据指标进行拆解细分，构建指标矩阵，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 。最后按照相同的应用主题、功能模块，将 DWS 层中的指标济进行汇聚，提升数据的易用性，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，并支持即席查询。

- **项目难点及解决思路**：
  - 业务方面，产品迭代迅速，而自身是首次加入该团队，需要快速学习业务和产品相关知识：首先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，了解产品产生的背景、目标、功能等，然后主动和产品经理进行沟通交流，查漏补缺。
  - 技术方面，数据项目历史代码参与人员较多、各模块代码和表结构参差不齐，历史包袱重，开发成本高：首先需要调研和摸清代码和数据库现状，创建文档通过图表等方式描述代码和数据现状，然后按照优先级罗列待办事项，最后按照不同优先级将待办事项排期依次解决。

- **项目总结与反思**：
  - 在解决历史包袱过程中，可以使用四象限法辅助排期，并按不同的事项主题进行汇聚，然后逐渐拆分事项，并排期到日常迭代周期中进行解决。按事项的主题进行划分，是为了避免做的事情不够聚焦，导致工作汇报时，无法体现该工作的产出和价值；而拆分排期，是为了尽量减小这些重要不紧急事项对于主线需求排期的影响，避免与主线排期冲突，导致重要项目延期。