叶阳

广州
Tel:18520149491
Linkedin: https://www.linkedin.com/in/yang-ye/
Github: https://github.com/dsuuny?tab=repositories
Emial: D244975154@gmail.com

数据工程师（硕士）| ETL | 数据分析| Hadoop| 数据仓库| Spark | AWS | GCP|阿里云| Python |

教育

利默里克大学(一等荣誉学位) 2022年9月 -- 2023年9月
硕士 软件工程 Limerick, 爱尔兰
华北理工大学 2014年9月 - 2018年6月
本科 信息管理与信息系统 唐山,中国

技术技能

编程语言: Python(高级), Java, SQL, NOSQL, HTML5/CSS
开发工具: VS Code, MS Visual Studio, IntelliJ Idea, PyCharm, Tableau
数据开发: MySQL, PostgreSQL, Machine Learning, Hive, Spark, HDFS, Data Warehouse, Elasticsearch, MongoDB
其他: Agile, Git, MPI, CUDA, GCP/AWS

工作经历

SYMBIO（汇丰银行项目外包） 2024年2月 - 2025年7月
全栈工程师 中国，广州

数据开发自动化平台:

技术栈: Python, Google BigQuery, CI/CD, REST API, Confluence

- 设计并开发了一个基于Python的自动化平台，以简化整个数据开发生命周期。
- 设计该工具以自动生成标准化的BigQuery JSON代码，确保了代码一致性并减少了人工错误。
- 实施了自动化流水线，用于向开发（DEV）环境自动部署、执行测试，并将结果直接发布至Confluence页面，以提高项目透明度。
- 集成了多个内部API，以自动化预生产环境的发布工作流，例如工单创建和审批请求。
- **关键成果**：显著缩短了数据交付周期，将开发阶段从**1周**缩短至**1天**，生产发布准备阶段从**3天**缩短至**1天**。

ERMS（ETL需求管理服务）:

技术栈: FastAPI, React, PostgreSQL, Docker

- 作为技术负责人，领导一个4人的工程师团队，主导了项目从概念到持续开发的整个生命周期。
- 采用敏捷方法论领导开发团队，以促进迭代开发、管理任务并确保项目按时交付。
- 主导了关键技术决策，包括技术栈选型、系统架构和数据库模式设计。
- 指导新团队成员并提供技术培训，促进其个人成长，同时保证团队的代码质量。
- 实现了核心功能，包括：需求版本控制、动态审批工作流、数据血缘可视化以及AI助手。
- **关键成果**：通过为数据模型设计团队简化设计流程，并利用AI驱动的代码生成功能为模型工程师加速开发，从而优化了整个需求生命周期。

CLPS（汇丰银行项目外包） 2021年2月 - 2022年5月
数据工程师 中国，广州

Teradata - GCP, AWS & 阿里云数据迁移:
技术栈: DataStage + SQL + Python
- 设计并开发了从Teradata到GCP、AWS和阿里云的ETL流程。
- 对ETL流程进行彻底的测试和验证,以确保迁移过程中的数据完整性。
- 与内部客户密切协作,定义关键性能指标(KPI)和监控机制,以跟踪迁移过程的进展。
- 分析了 Teradata和 DataStage 的现有数据结构。
- 与技术团队协作实施优化解决方案,这可能包括调整数据传输方法、利用并行处理或优化 SQL查询。

Xcheck中间件:
技术栈: SQL + Python + JinJa
- 开发了一个数据校验工具,用于解决源端数据与目标端，因为逻辑差异导致的数据差异问题。
- 利用SQL查询从 Teradata和 BigQuery 中检索数据。

HSBC 数据迁移技术支持:
技术栈: Python
- 将原来数据库的 DDL 查询语句转换为与新平台兼容的语句。
- 协助用户将 SQL 查询和脚本从原始 Teradata 环境迁移到新平台。
- 开发了一个多线程 ETL 工具以用于高效的表数据传输。

厦门市美亚柏科信息股份有限公司 2018年11月 - 2020年2月
ETL工程师 中国，广州

数据标准化注释系统:
技术栈: SQL + Flask+ Pandas + scikit-learn
- 设计了文本分类模型(tf-idf, 逻辑回归)用于数据注释。
- 使用 Flask 开发 RESTFUL 后端API,将文本分类模型的功能开放给外部系统。

DATAX 插件:
技术栈: Java
- 根据每个数据库独特的数据结构、模式和数据集成需求,设计和开发定制的 ETL 扩展。
- 为结构化和非结构化数据库开发数据清洗功能。

广州火数银花信息科技有限公司 2017年4月 - 2018年10月
数据分析师 广州

电商消费者行为分析:
技术栈: Python + Pandas + Numpy + POWERBI + SQL
- 了解淘宝买家的行为,包括购买偏好、购物车分析、购买周期以及购买决策过程。
- 使用算法对来自 Oracle/MySQL/PostgreSQL 的客户购买记录数据进行分析,以预测潜在客户。

商品销售预测:
技术栈: Python + Pandas + Numpy + POWERBI

- 执行探索性数据分析(EDA),以发现各种因素之间的相关性和依赖性,例如产品类别、定价、促销和销售量。
- 分析商品销售收入,以深入了解历史销售模式、趋势和季节性。
- 根据分析结果,卖家优化定价、库存管理、促销活动和产品定位,以最大限度地提高客户满意度和增强销售效率。

个人项目经历

Immune cell profiling from single-cell RNA with Neo4J
Github: https://github.com/MeghanaKshirsagar/ISCBTutorial

- 构建了Seurat分析流程，用于处理GEO（Gene Expression Omnibus）数据库的单细胞数据。
- 采用降维和聚类方法对胃部数据集进行了分析。
- 利用多种细胞注释工具（如SingleR, scMRMA等）对细胞类型进行标记。
- 开发了ETL流程，将CSV文件数据导入Neo4j图数据库，并编写Cypher查询脚本以验证关键的生物标志物。

证书

- Oracle Certified Professional
- AWS Certified Developer - Associate
- IELTS 6.0

技能

- 语言: 粤语(母语), 普通话(母语), 英语(流利)
- Office软件: Word, Excel, PowerPoint
- 摄影，视频剪辑