서정민

Data Analytics Engineer

신입 DAE를 희망하는 서정민입니다. 데이터 기반 의사 결정을 지원하기 위한 고품질 데이터의 수집, 처리, 저장 프로세스를 고민하고 개발하는 과정을 즐깁니다. **모든 규모와 형태의 데이터를 효율적으로 처리할 수 있는 환경을 구축하여, 데이터 기반의 혁신적인 세계를 꿈꾸는 저의 궁극적인 목표**를 토스와 함께 성장하며 펼쳐나가고자 합니다.

[Email] mujm1217@gmail.com
[Website] https://www.jmseo1217.com
[Github] https://github.com/jeongmin1217

Work Experience

(주)비상교육
데이터플랫폼셀

2024.01 – 2024.02
2023.07 – 2023.08

Data Engineer Intern

데이터 플랫폼 구축 및 운영

서로 다른 회사처럼 운영되어 온 21개 사내 브랜드의 데이터 통합 운영을 통한 가치 창출 목표

- **데이터 파이프라인 개발**
  - Airflow를 이용해 데이터웨어하우스로부터 사내 데이터 포털 DB에 6개 브랜드 DB **메타데이터를 ETL하는 파이프라인 구축**
    - 외래키 제약조건, DAG간의 의존성, 코드의 재사용성을 고려한 알고리즘 작성
  - 지속적으로 효율적인 코드를 고민하여 If/Else 문 대신 ON CONFLICT를 활용한 Upsert 방식 쿼리문과 executemany 배치 처리를 도입해 **52,714개 레코드** 처리 시간을 **12분 9초에서 11분 32초로 약 5% 단축**
  - 타 스키마/브랜드 내 같은 이름의 테이블로 인한 **정합성 오류**를 발생시키는 팀원의 코드를 꼼꼼한 **코드 리뷰와 테스트 케이스 작성을 통해** 식별 및 해결
  - 사내 브랜드별 DQ 통계를 관리하는 집계 테이블을 이용한 간단한 DQ 모니터링 도입
- **Data Maturity를 높이기 위한 Data Quality 관리**
  - Data Quality 통계 관리, 사내 표준 용어 사전 구축 기여를 통한 Data Maturity 향상

사내 데이터 포털 서비스 개발

21개 사내 브랜드의 데이터 통합 관리 목적의 Data Catalog 시스템

- **전사적 Data Literacy 향상을 위한 기획 및 개발**
  - **데이터 주제 영역 분류 자동화 시도**
    - 데이터 포털 구축 초기 : 커뮤니케이션 후, Non-IT 직군 데이터 리터러시 향상위해 브랜드별 개념 데이터 모델 파악 후 수작업으로 데이터 주제 영역 분류
    - 자동화 1차 시도 : ChatGPT API에 테이블/컬럼명, 브랜드 구조 입력 후 분류 시도
    - 자동화 2차 시도 : ChatGPT API로 테이블/컬럼명을 통한 테이블/컬럼 키워드 생성 후, 역으로 키워드를 이용한 주제 영역 분류 시도
    - **유의미한 키워드는 생성하였으나, 주제 영역 분류에는 정확성이 떨어지며 실패**
- 데이터 포털 서비스 Front-End 개발

Tech : Airflow, AWS(S3, RDS, EC2, VPC), Docker, PostgreSQL, MySQL, React, Tailwind CSS, Git, Figma

Purdue University
Computer & Information Technology Dept.

2023.09 – 2023.12

Student Intern (International Scholarship)

뇌파(EEG) 데이터를 활용한 바이오 데이터와 집중력 척도 간의 상관관계 분석

복잡한 뇌파 분석 대비 간편한 웨어러블 기기를 통한 집중력 측정 및 집중력 측정 앱 개발 목표

- **탄탄한 도메인 이해도를 바탕으로 분석용 데이터 제공 및 분석 결과 검증**
  - 앉아서 진행하는 실험 특성 상, **영향력이 없던 운동량 지표를** 운동량과 집중력의 상관관계 논문을 바탕으로 **재해석하여 모델 성능 R^2 값 0.031 향상**
  - 영향력이 가장 높은 지표를 도메인 바탕으로 분석하여 연구 결과의 타당성 검증
- **데이터 전처리를 통한 모델 성능 개선 및 프로세스 자동화**
  - 두 기기의 데이터 수집 주기 병합 과정 중, 오류 데이터를 따라가던 이슈를, 센서 오류 데이터의 기준을 분석하여 **어쩔 수 없게 여기던 성능 하락 이슈 최소화**
  - 전체 전처리 프로세스로 원본 데이터 실험 대비 모델 성능 R^2 값 총 0.25 증가

Tech : Python(Pandas, Numpy), Jupyter Notebook, Git

Main Project

Posture Guard

개인 프로젝트

사용자의 카메라를 활용한 실시간/일간/주간 자세 분석 및 기록 데스크탑 앱

Mediapipe를 이용한 지속적인 척추 및 목 자세 모니터링을 통한 척추 건강 관리 목표

- **대용량 데이터 처리 경험 및 지표 설계, 분석**
  - 오버헤드가 없으면서 최대 효율 성능을 내기 위한 적절한 Spark Executor 수(7개)와 Partition 수(390개)를 설정하여 분산 처리를 통한 성능 향상
    - **1만개의 평균 45KB json 파일** 처리 및 분석 작업을 42분 46초에서 35분 42초로 **약 17% 단축**
  - Kafka와 Spark Streaming을 이용한 **준실시간 (600ms)** 데이터 처리 및 분석
    - Kafka와 Spark의 partition 수를 6개로 늘려 **병목 현상 해결**
- **데이터 분석에 필요한 데이터 집계 마트 개발 및 구축**
  - **BigQuery에 데이터 집계 마트 구축**
    - 준실시간 데이터 분석을 통해 수집된 데이터를 일일 배치로 하루 평균의 이미지와 값을 분석해 BigQuery와 Google Cloud Storage에 저장
  - **데이터 집계 마트를 이용한 데이터 분석**
    - 지난 일주일간의 평균 기록을 BigQuery 데이터 집계 마트를 이용해 효율적 분석
- 명확하고 쉬운 데이터 구조 설계에 기반한 용도에 따른 데이터 처리 및 분석

Tech : Spark, Kafka, GCP(BigQuery, GCS, Dataproc, Cloud Schedular), Docker, Django, React, Electron, Mediapipe

Academic Activities

2022.02 ~ 2023.02	연합 빅데이터 동아리 BITAmin	데이터 분석/ML/DL 스터디 및 프로젝트 진행
2022.07 ~ 2022.08	SVSTIP (Silicon Valley Software Technology and Innovation Program)	실리콘밸리 SW 연수 및 사용자 리뷰 기반 OTT 추천 프로젝트 개발
2021.03 ~ 2022.02	한양대학교 멋쟁이사자차럼	풀스택 웹개발 스터디 및 위치기반 SNS 플랫폼 프로젝트 개발

Skills

Data Engineering	Airflow, Spark, Kafka, PostgreSQL, MySQL
Cloud Computing	Docker, GCP (BigQuery, GCS, Dataproc), AWS (EC2, S3, RDS)
Web, ML, DL	Django, React, Numpy, Pandas, OpenCV
Language, etc.	Python, SQL, JavaScript, Git

Education

대학교	경희대학교 산업경영공학부, 소프트웨어융합학부 (복수전공)	2018년 2월~2025년 2월(예정)
고등학교	용인한국외국어대학교부설고등학교	2014년 2월~2017년 2월

Awards

경희대학교 2023-1 캡스톤디자인 경진대회	우수상
2023년 도시형소공인 집적지구 캡스톤경진대회	장려상
SVSTIP (Silicon Valley Software Technology and Innovation Program)	Entrepreneurship Award

Extracurricular Activities

경희대학교 산업경영공학과 3학년 과대표
연합 동아리 공감강연단 꿈inDream 18기 단장
KHU Global Ambassador (교환학생 도우미)