JEONGMIN'S PORTFOLIO

https://www.jmseo1217.com

https://github.com/jeongmin1217

토스뱅크와 함께 성장하며 데이터 기반의 혁신적인 세계를 구축하고자 합니다.

JEONGMIN SEO

ABOUT ME

모든 규모와 형태의 데이터를 효율적으로 처리할 수 있는 환경 구축을 통한, 데이터 기반의 혁신적인 세계를 꿈꿉니다.

이의 실현을 위해서는 아래의 2가지가 가장 중요하다고 생각합니다.

1. 확장 가능하고 효율적인 데이터 인프라 및 파이프라인 구축
2. 대규모 데이터를 효율적으로 처리하고 분석할 수 있는 분석 환경 구축

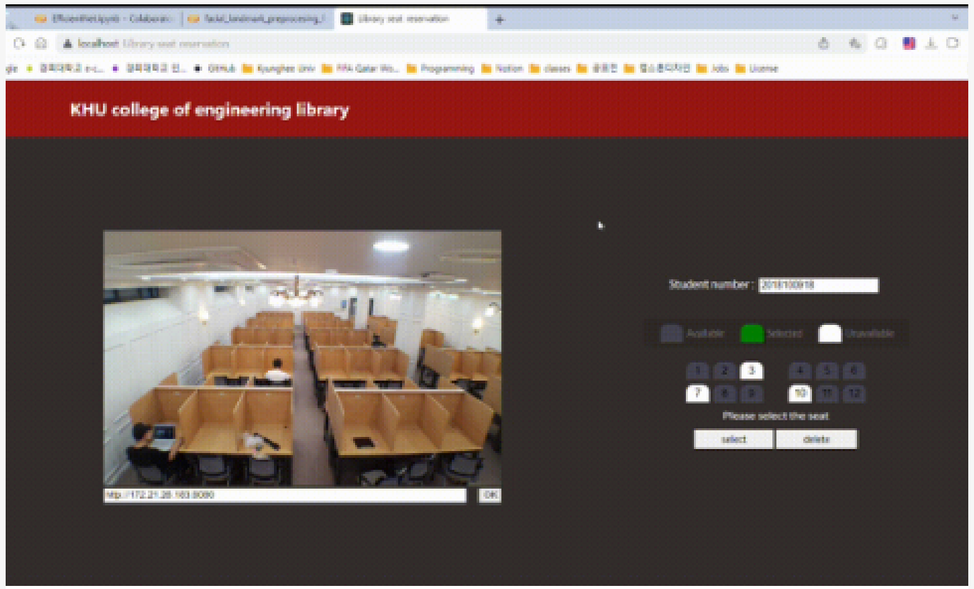
위의 2가지를 위한 지금까지 저의 여정을 프로젝트와 함께 포트폴리오로 소개드리겠습니다.

제로베이스 시절 타대학교 웹 개발 동아리에서 유일한 타학교 학생으로 활동하는 등 도전을 두려워 하지 않는 성격,
인턴 당시 사수님이 미처 고려하지 못한 데이터 정합성 오류까지 해결하는 등의 꼼꼼한 성격을 기반으로,
토스뱅크에서 핀테크 산업의 다양한 규모와 형태의 빅데이터를 다루며 함께 성장할 수 있길 바랍니다.

STEP 01

IMPROVING LIBRARY SEAT MANAGEMENT SYSTEM BASED ON COMPUTER VISION - TEAM PROJECT

컴퓨터 비전을 이용한 경희대학교 공과대학 열람실 예약시스템 개선 프로젝트



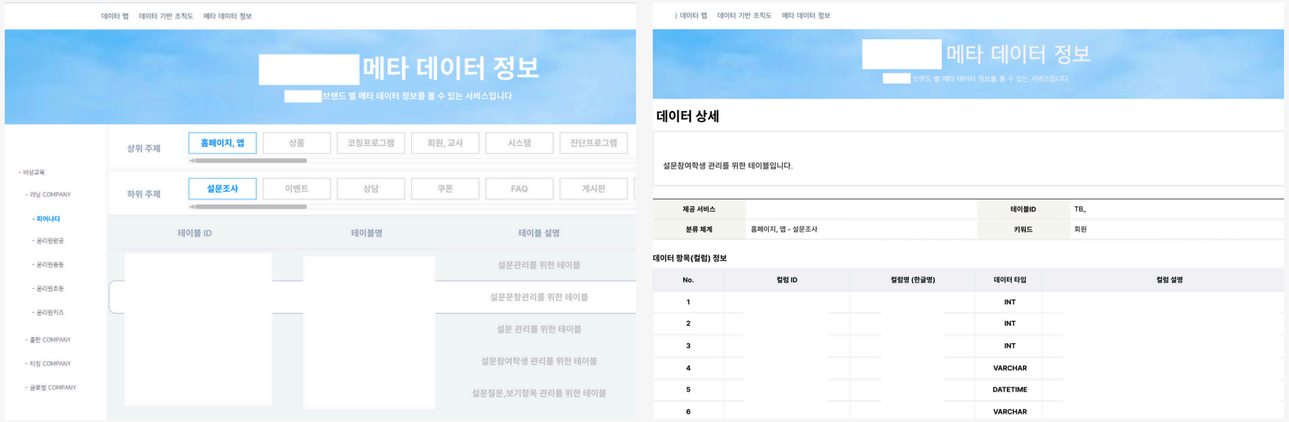
[개선된 공과대학 열람실 예약시스템 개발]

- **문제정의**
  - 빈번하게 발생하는 사유화된 좌석 혹은 예약시스템에 예약되어 있지만 공석인 좌석 문제. 그동안 이는, 학생회에 의해 수작업으로 처리되어 왔음.
- **목표**
  - 카메라를 통해 좌석을 감지하여 사유화된 좌석과 공석을 자동으로 반납 처리하는 예약 시스템 개발
- **주요 해결 과정**
  - 녹화된 영상이 아닌 실시간 카메라를 통해 좌석을 감지해야 했던 문제를 라우터를 사용한 네트워크 포트포워딩을 통해 실시간 IP 카메라 스트리밍을 클라이언트 단에 송출하여 해결
- **배운 점**
  - 프로젝트 당시, 해당 프로젝트가 엄청난 실시간성을 요구하지 않았기에, 10초 주기로 데이터를 전송하는 방식을 선택. 그러나 1초 주기로 진행을 해보았을 때, 병목현상 이슈 경험을 하였고, 이 때 처음 실시간성 데이터 처리에 궁금증을 갖게 됨.

STEP 02

VISANG EDUCATION DATA PLATFORM - 1

- **문제정의 :** 사내 브랜드 21개가 서로 다른 회사처럼 운영되어 버려지는 데이터가 많음.
- **목표 :** 전사적 Data Literacy 향상
- **주요 해결 과정 :주제영역 분류 자동화 시도**
  - 구축 초기 : 3100개 가량의 테이블 메타데이터를 직접 보고, 브랜드별 개념 데이터 모델 등을 참고해 수작업으로 데이터의 주제 영역 분류
  - ChatGPT API에 테이블/컬럼명, 브랜드 구조 입력 후 분류 및 테이블/컬럼명을 통한 테이블/컬럼 키워드 생성 후 키워드를 이용한 분류 자동화 시도
  - 유의미한 키워드는 생성하였으나, 주제 영역 분류에는 정확성이 떨어지며 실패
- **배운 점**
  - 3100개 가량의 테이블을 직접 보며, 규칙없이 운영되어 온 데이터와 이를 활용한 가치 창출의 무한함을 생각했을 때 매우 아쉬웠고, DQ의 중요성을 배움.
  - 주제영역 분류 자동화 시도 당시, 비록 성공하지 못 하였지만 키워드를 이용한 bottom-up 방식의 새로운 시도로 가능성을 느낌.



[비상교육 사내 데이터 포털]

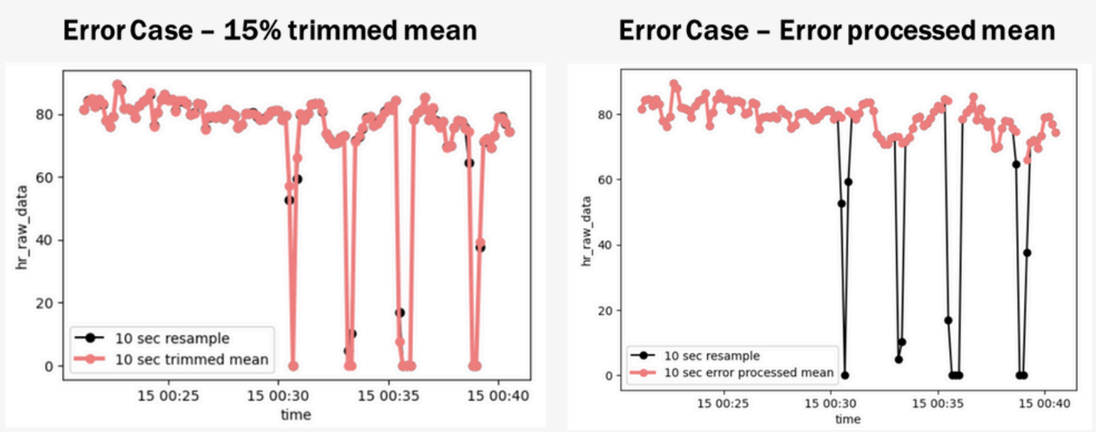
개발, 기획 분석까지 모든 단계에 대해서 아주 좋은 수행능력을 보여주었습니다. 또한 작은 것 하나도 놓치지 않고 업무적으로 디테일한 부분이 아주 좋았습니다. 업무적으로 아주 많은 도움이 되어서 정말 긍정적이였습니다.

[인턴십 당시 사수님 피드백]

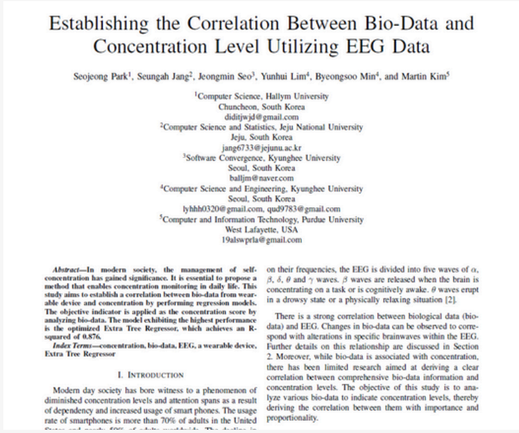
STEP 03

PURDUE UNIVERSITY DEPARTMENT OF COMPUTER & INFORMATION TECHNOLOGY STUDENT INTERN PROJECT

- **문제정의 :** 일상 생활에서 집중력의 척도를 파악하기 쉽지 않음.
- **목표 :** 복잡한 뇌파 분석 대비 간편한 웨어러블 기기를 통한 집중력 측정 및 집중력 측정 앱 개발
- **주요 해결 과정 : 데이터 전처리를 통한 모델 성능 개선 및 프로세스 자동화**
  - 앉아서 진행하는 실험 특성 상, 영향력이 없던 운동량 지표를 운동량과 집중력의 상관관계 논문을 바탕으로 재해석하여 모델 성능 R^2 값 0.031 향상
  - 두 기기의 데이터 수집 주기 병합 과정 중, 오류 데이터를 따라가던 이슈를, 센서 오류 데이터의 기준을 분석하여 어쩔 수 없게 여기던 성능 하락을 끈기 있게 최소화
- **배운 점**
  - 오류데이터를 쫓는 오류를 해결하고자 집요하게 파고들어, 오류 데이터의 기준까지 분석해보며 데이터 퀄리티의 중요성과 버려질 수도 있던 데이터를 이용한 가치 창출을 통해 도메인 기반 지식과 해석의 중요성을 배움.



[데이터 전처리 및 시각화]


Establishing the Correlation Between Bio-Data and Concentration Level Utilizing EEG Data

Seojeong Park[1], Seungah Jang[2], Jeongmin Seo[3], Yunhui Lim[4], Byeongsoo Min[4], and Martin Kim[5]

[1]Computer Science, Hallym University
Chuncheon, South Korea
didiljwjd@gmail.com
[2]Computer Science and Statistics, Jeju National University
Jeju, South Korea
jang6733@jejunu.ac.kr
[3]Software Convergence, Kyunghee University
Seoul, South Korea
balljm@naver.com
[4]Computer Science and Engineering, Kyunghee University
Seoul, South Korea
lyhhh0320@gmail.com, qus99783@gmail.com
[5]Computer and Information Technology, Purdue University
West Lafayette, USA
l9alswpta@gmail.com

***Abstract*—In modern society, the management of self-concentration has gained significance. It is essential to propose a method that enables concentration monitoring in daily life. This study aims to establish a correlation between bio-data from wearable device and concentration by performing regression models. The objective indicator is applied as the concentration score by analyzing bio-data. The model exhibiting the highest performance is the optimized Extra Tree Regressor, which achieves an R-squared of 0.876.**

***Index Terms*—concentration, bio-data, EEG, a wearable device, Extra Tree Regressor**

I. Introduction

Modern day society has bore witness to a phenomenon of diminished concentration levels and attention spans as a result of dependency and increased usage of smart phones. The usage rate of smartphones is more than 70% of adults in the United

on their frequencies, the EEG is divided into five waves of α, β, δ, θ and γ waves. β waves are released when the brain is concentrating on a task or is cognitively awake. θ waves erupt in a drowsy state or a physically relaxing situation [2].

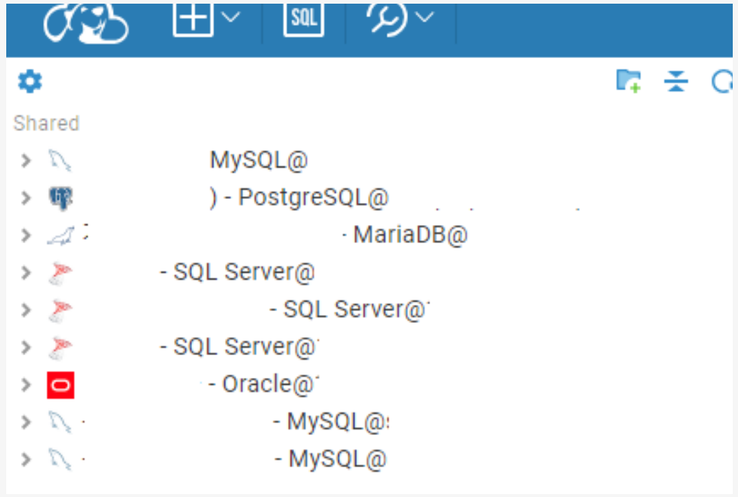
There is a strong correlation between biological data (bio-data) and EEG. Changes in bio-data can be observed to correspond with alterations in specific brainwaves within the EEG. Further details on this relationship are discussed in Section 2. Moreover, while bio-data is associated with concentration, there has been limited research aimed at deriving a clear correlation between comprehensive bio-data information and concentration levels. The objective of this study is to analyze various bio-data to indicate concentration levels, thereby deriving the correlation between them with importance and proportionality.

[최종 논문]

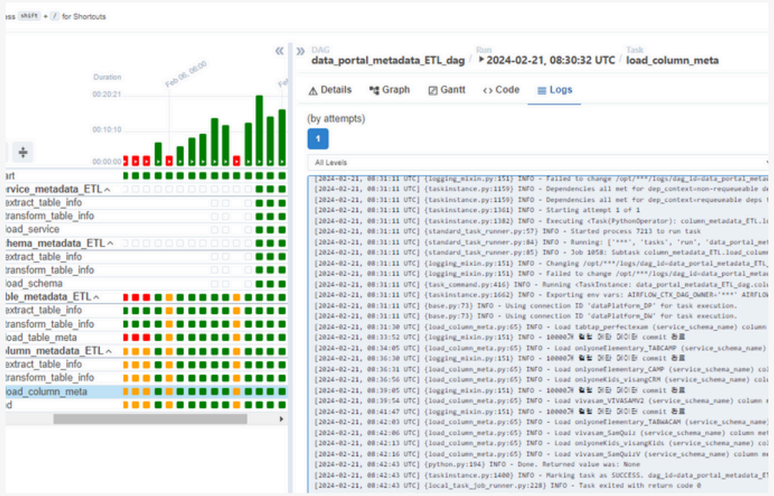
STEP 04

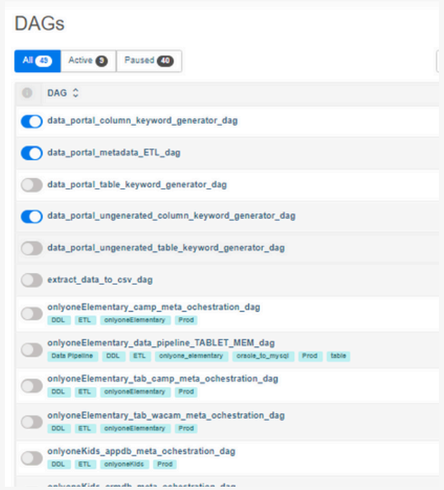
VISANG EDUCATION DATA PLATFORM - 2

- **문제정의 :** 서로 다른 회사처럼 운영되어 온 사내 브랜드 21개
- **목표 :** 데이터 플랫폼 구축 및 운영을 통한 가치 창출 목표
- **주요 해결 과정 : 전사적 데이터 파이프라인 구축**
  - 지속적으로 효율적인 코드를 고민하여 52,714개 레코드 처리 시간을 12분 9초에서 11분 32초로 약 5% 단축
    - Upsert 최적화: 테이블의 Unique Key를 활용하여 ON CONFLICT와 DO UPDATE SET 구문을 적용한 Upsert 방법을 도입하여 코드 효율을 극대화.
    - 배치 처리: execute 대신 executemany를 사용하여 배치 처리 방식을 도입, 데이터베이스 트랜잭션의 효율성을 향상.
  - 타 스키마/브랜드 내 같은 이름의 테이블로 인한 정합성 오류를 발생시켰던 사수님의 코드를 꼼꼼한 코드 리뷰와 테스트 케이스 작성을 통해 식별 및 해결
- **배운 점**
  - 외래키 제약조건, DAG간의 의존성, 코드의 재사용성을 고려한 알고리즘 작성과 테스트 케이스 작성을 통한 데이터 중복/일관성/정합성을 다양한 케이스를 통해 고려하는 것의 중요성을 배움.



[전사적 데이터 파이프라인 구축]





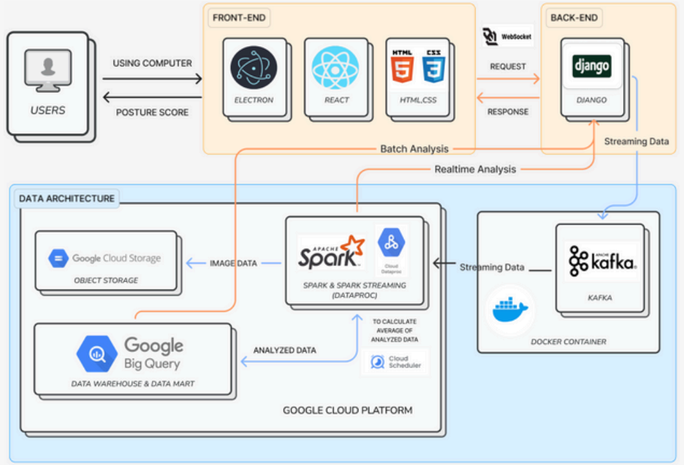
[AIRFLOW DAG]

STEP 05

POSTURE GUARD - PERSONAL PROJECT

- **문제정의 :** 현대인의 고질적인 노트북/컴퓨터 사용 중 척추 및 목 건강 문제
- **목표 :** 지속적인(실시간에 가까운) 모니터링 시스템 개발을 통한 척추 및 목 건강 개선
- **주요 해결 과정 : 실시간 데이터 처리 & 데이터 분석에 필요한 데이터 집계 마트 개발 및 구축**
  - 적절한 Spark Executor수(7개)와 Partition수(390개)로 분산 처리를 통해 성능 향상 (1만개의 평균 45KB json 파일 처리 및 분석 작업 시간 17% 단축)
  - Kafka와 Spark Streaming을 이용한 준실시간 (600ms) 데이터 처리 및 분석 과정 중, Partition수를 6개로 늘려 병목 현상 해결
    - 단일 파티션에 걸리던 과부화를 해결하기 위해 수집한 데이터를 RDD의 병렬처리적 특성을 활용
  - 주간 평균 기록 제공을 위해 일주일간의 데이터를 분석하는 것은 리소스가 소모되는 것이 많다고 판단, 수집된 준실시간 데이터를 일일 배치로 하루 평균의 이미지와 값을 분석해 BigQuery와 Google Cloud Storage에 저장하여 분석에 사용할 수 있는 데이터 집계 마트 구축
- **배운 점**
  - 명확한 문제 정의 기반의 이해하기 쉬운 데이터 아키텍처의 설계는 곧, 추후 효율적이고 간편한 분석 작업을 가능하게 하며 분석의 범주 또한 확대할 수 있는 기반을 마련한다는 점을 배움.

[POSTURE GUARD 프로젝트 앱]



[프로젝트 아키텍처]

THANK YOU.

You can see more details in

https://www.jmseo1217.com

https://github.com/jeongmin1217