Dohee Kim

Data Engineer

Contact

010-5857-6510

kheedogg@naver.com

Gyeonggi-do, South Korea

linkedin.com/in/kheedogg/

GitHub kheedogg - Overview

Education

통계학(주전공) & 컴퓨터공학(복수전공) 학사 학위

부산대학교

2017.03-2022.02

Language

Language

New

English	Korean	+ New page
Opic-IH	Native	

Awards

UKU 학부 연구 포스터 장려상 수상

About me

숫자로 세상을 해석하는 것을 좋아하는 저는 자연과학대학 통계학을 전공하였고, 보다 넓은 세상을 이해하기 위해 컴퓨터 공학을 복수전공하여 이중 학사학위를 취득했습니다.

졸업 전, 글로벌 인턴십 프로그램을 통해 버클리 소재 스타트업에서 데이터 분석 업무를 처음으로 경험했습니다. 정규직 전환 후, 다양한 데이터 관련 업무를 수행하며 스키마 설계, 데이터 파이프라인 구축, AI 모델 개발 등 엔지니어링 전반의 프로세스를 익히는 소중한 시간이었습니다.

5개 주요 앱의 약 1,300만명의 사용자가 생성하는 매일 500만건 로그 데이터와 최대 10만명의 사용자를 커버하는 AIDT 디지털 교과서 데이터 파이프라인 작업을 주로 진행하며 데이터 파이프라인 구조 최적화와 데이터 처리 기술에 깊은 관심을 갖게 되었습니다. 다양한 환경에서 효율적으로 대규모 데이터를 관리해 나갈 수 있도록 데이터 엔지니어로서의 역량을 한층 더 발전시키고자 합니다.

Work Experience

에누마코리아 정규직, 2022.1 ~ 재직중

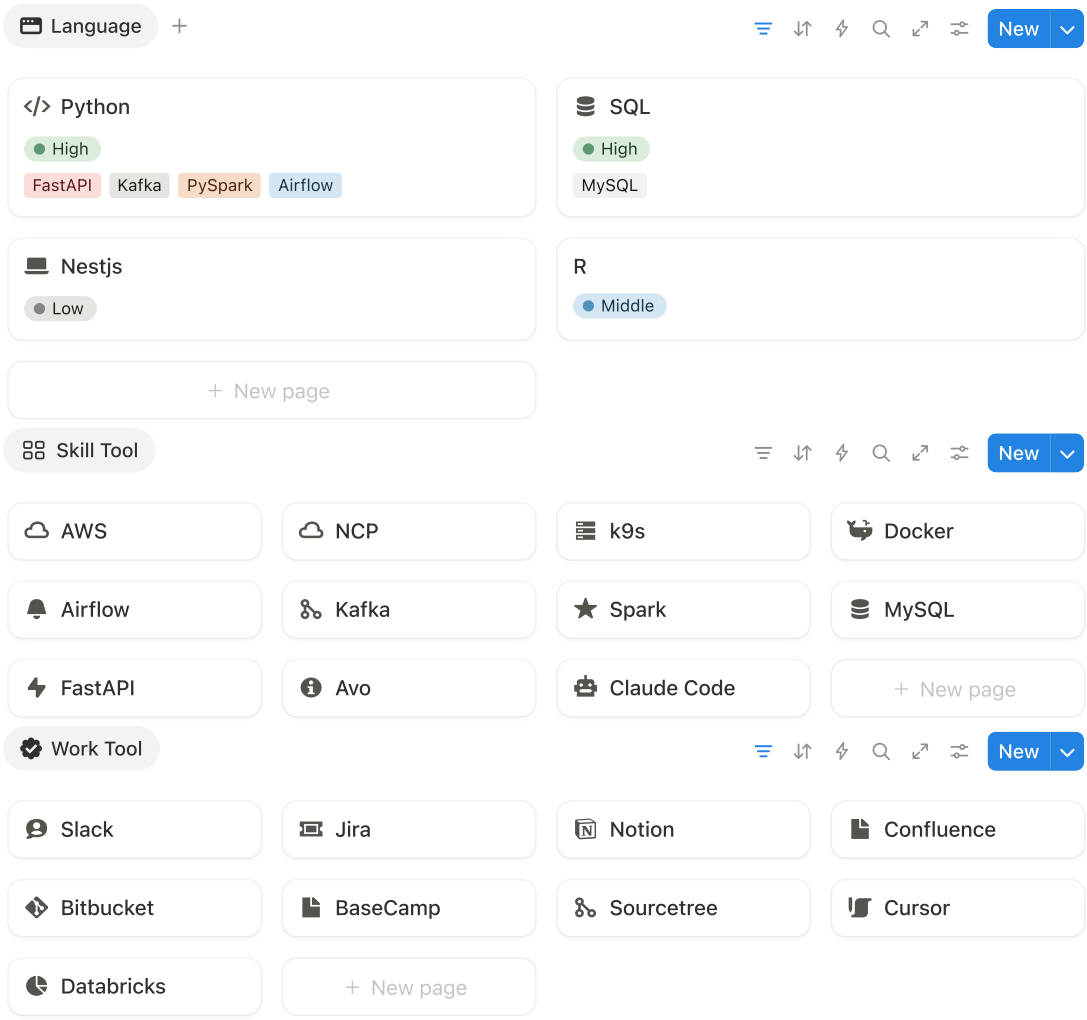
- AIDT 디지털 교과서 LRS(Learning Record Store) DB 관리 및 데이터 서빙 작업 진행 ~2025
- 개발팀&데이터팀 소속 토도시리즈 앱 데이터 파이프라인 이전 작업 진행 ~2022

- AITFT팀 일환으로 모델 구축 작업 진행 (Text Recognition Model, Knowledge Tracing Model) ~2022

에누마 인턴, 2021.07 ~ 2021.12

- 토도앱 사용자 로그 데이터 가공/시각화/분석 진행 ~ 2021

Skills



Publications

- Kim, D., et al. (2025). ES-KT-24: A Multimodal Knowledge Tracing Benchmark Dataset with Educational Game Playing Video and Synthetic Text Generation. *Intelligent Tutoring Systems*. arXiv:2409.10244. **[First Author]**
- Lee, U., Bae, J., Kim, D. (2024). Language Model Can Do Knowledge Tracing: Simple but Effective Method to Integrate Language Model and Knowledge Tracing Task. arXiv:2406.02893. **[Co-author]**
- Lee, U., et al. (2024). From Prediction to Application: Language Model-based Code Knowledge Tracing with Domain Adaptive Pre-Training and Automatic Feedback System with Pedagogical Prompting for Comprehensive Programming Education. arXiv:2409.00323. **[Contributing Author]**

Work History

Jan 2022-
Present

Data Engineer

Full Time

- **AI 디지털 교과서 공동 개발:** 초등 및 중고등 수학, 영어, 사회 교과목의 데이터 실시간 및 배치 파이프라인 구축
  - 학습 로그 데이터 스키마 설계
  - MySQL Service DB 실시간 데이터 파이프라인 구성
    - 인덱스 구조 최적화를 통해 API 응답 시간을 4초에서 1초 미만으로 75% 개선
  - Hadoop 내에서 Zeppelin 및 Airflow를 사용하여 Spark 기반의 Learning Record Log ETL 배치 파이프라인 구성
    - 청크 단위 비동기 처리 도입으로 사용자당 배치 처리 시간을 1.2초에서 0.07초로 단축
  - Learning Record 및 모델(Paper 2.) 추론 결과 데이터 전송을 위한 FastAPI 백엔드 작업 진행
- 토도 앱 DB 구조 개선
  - 매일 약 500만건의 로그 데이터를 처리하는 배치 파이프라인을 Airflow에서 Databricks로 이전 작업 진행
  - Medallion Architecture 구조 및 snapshot을 활용하여 로그 데이터 무결성 보장 강화
- 토도 앱 파이프라인 최적화: Databricks에 Mysql 연결 및 DynamoDB 복제 테이블 구성
  - 테이블 스캔 대신 Secondary Index 및 Query를 사용하여 배치 작업 속도를 4배 이상 개선
  - DynamoDB의 Capacity Mode를 On-demand에서 Provisioned 모드로 변경하여 비용을 35% 절감

AI Engineer

Full Time

- 개인 맞춤 이해도 모델 개발: 사용자 게임(학습) 데이터를 기반으로 LKT 모델(Paper 1 & Paper 2) 구축 (Knowledge Tracing Model)
  - 사내 게임 로그 데이터를 모델 추론 형식에 맞춰 변형 (정오답 산정 기준 수립)
  - 기존 DKT 모델과 AUC 및 ACC 성능 비교
- 수학, 한자, 한글, 영어 등 과목별 손글씨 이미지에 대한 CNN 기반 모델 구축 (Text Recognition Model)
  - 성능 유지 + 레이어 단축으로 8MB 크기에서 3MB로 모델 크기 축소
  - 모델 테스트 페이지 직접 개설하여 모델 버저닝 관리, 오인식 이미지 수집, 모델 추가학습 진행

July 2021-
Dec 2021

Data Analyst

Intern

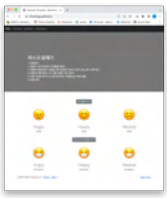
- Adhoc 이벤트 처리: 데이터 가공 및 추출 작업 수행
- 게임 데이터 분석: 음성 데이터 및 인지 능력 관련 게임 데이터 분석
  - 게임(학습) 이해도 분석 후, 통계적 결과를 기반으로 커리큘럼 및 게임 난이도 수정
- 대시보드 이전 및 보완 작업: 게임(학습) 데이터 관련 대시보드를 Apache Superset에서 Databricks로 이전 및 보완
- QA 로그 검증: 신규 및 기존 로그 데이터 무결성 보장 파이프라인 추가

Projects

Featured | All Projects

New

산학협력 with LOCS



졸업과제

(URO-학부연구생) COVID-19 Analysis

KBL ALL STAR전 농구 선수 선발

태블로 신병 훈련소 _ 19기

+ New page