**AWS 교육 서비스를 진행하기 위해 관련 업체의 데이터 수집 프로젝트 착수보고서**

**천재교육 K-Digital Training 빅데이터7기**

**이주복, 김재경, 전소현**

**Ⅰ. 일반현황**

1. 역할 분배 (R&R, Role and Responsibility)

* 이주복: 전체 프로젝트 관리 및 조정, 보고서 작성, 데이터 수집
* 김재경: 데이터 분석 및 시각화, 데이터 요약
* 전소현: 데이터 전처리 및 기초 통계분석

1. 업무 분류 체계(WBS, Work Breakdown Structure)



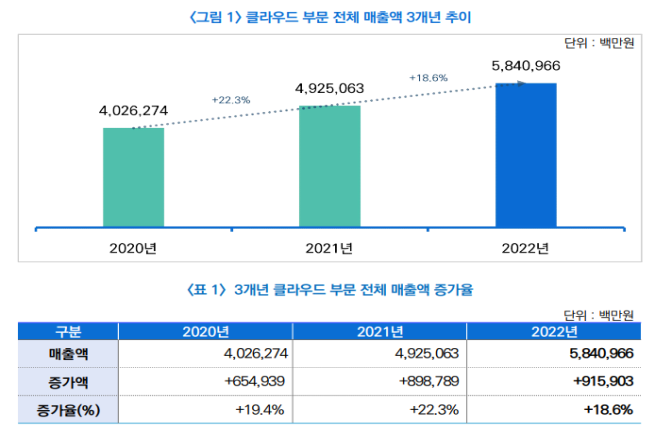
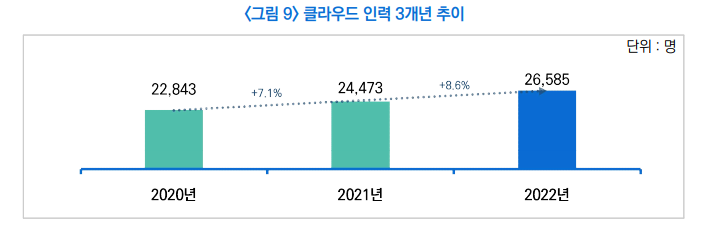
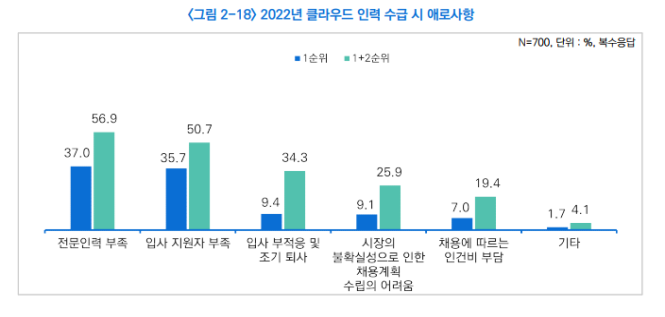
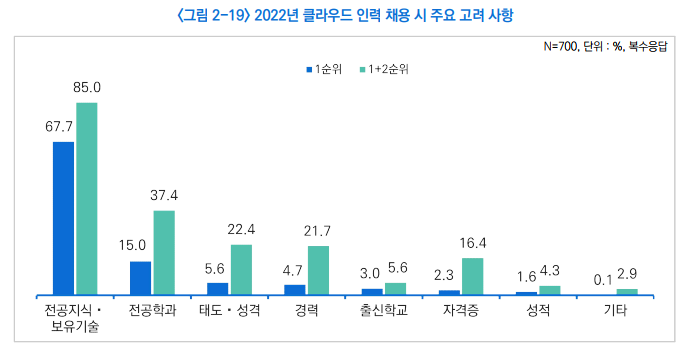
**Ⅱ. 전략 및 방법론**

1. 제안요약서

* 목적 및 배경: AWS 교육 서비스를 진행하기 위해 관련 업체의 채용 정보 수집
* 기대 결과: 채용정보 데이터를 통해 AWS 수요를 확인하여 교육 진행여부 결정
* 프로젝트 주요 목표 및 범위: AWS 역량을 필요로 하는 기업의 수요조사 및 교육 서비스 전략 수립, 개발자들이 자주 이용하는 구인구직 사이트(원티드, 랠릿, 프로그래머스)

1. 참고자료 분석 내용

* 2023 클라우드 산업 실태조사 보고서를 분석 후 클라우드 인력 교육 필요성 확인하여 AWS 수요조사 프로젝트를 진행 결정



* 2020년부터 2022년까지의 클라우드 인력 및 클라우드 부문 전체 매출액이 증가. (그림1, 그림9 참고)
* 클라우드 인력 채용 시 전공지식 및 보유기술을 갖춘 전문인력이 부족. (그림2-19, 2-20 참고)

1. 콘텐츠 정의서

* 수집할 데이터의 종류와 형식
  + 데이터 종류: 기업명, 업종분류, 사원수, 설립연도, 기업형태, 매출액, 주소, 직무 설명(JD), 사용 기술스택, 자격요건, 직급, 우대사항, 해당 페이지 URL, 기업 홈페이지 URL 등.
  + 데이터 형식: 수집된 데이터는 구조화된 형태로, 데이터프레임 형식에 맞추어 CSV 파일로 저장. 이는 분석 및 처리에 용이하여 일반적인 데이터 분석 도구에서 쉽게 읽을 수 있는 포맷임.
* 시각화 자료의 형식
  + 시각화 자료 형식: 분석 결과의 시각화는 PNG 파일 형식으로 저장. PNG 형식은 고해상도 이미지를 제공하며, 대부분의 플랫폼 및 장치에서 널리 지원되므로 시각화 자료 공유 및 표시에 적합함.
* 데이터 사용 목적 및 관리 방법
  + 사용 목적: 수집된 데이터와 시각화 자료는 AWS 관련 수요 및 요구조건을 분석하여 교육 서비스 전략 수립에 사용. 이를 통해 효과적인 교육 서비스 진행에 기여.
  + 관리 방법: 데이터 보안을 위해 크롤링 데이터의 공유는 제안자의 slack 또는 e-mail로 전달함.

1. 진행 계획

* 문제 정의: 클라우드 시장이 확대됨에 따라 클라우드 점유율이 높은 AWS(지난해 4분기 기준 31%) 관련 산업의 매출이 증가할 것으로 예상되어 AWS 수요를 확인하여 전략적 마케팅 수립
* 데이터 수집: 분석 로드맵 작성, 데이터 크롤링
* 데이터 전처리: 이상치 및 결측치 처리, 탐색적 데이터 분석(Exploratory Data Analysis), 데이터 형식 및 구조 조정
* 데이터 분석
  + 업종 분류별 AWS 수요의 차이를 확인하고 그 수요의 차이와 독립변수들 간의 상관관계를 알아보기 위해 중앙값을 기준으로 그룹화하여 로지스틱회귀 기획
* 결론 도출: 추출된 데이터 요약, 데이터를 통한 가정 및 한계점, 추가 분석 제안, 보고서 및 발표자료 준비

1. 기술 및 기능 정의서 요약
   * 크롤링 도구: Python을 활용한 웹 크롤링을 수행하기 위해 Selenium, BeautifulSoup, Requests 패키지 활용
   * 분석 소프트웨어: 데이터 분석 및 시각화를 위해 Python의 Numpy, Pandas, Sklearn, Matplotlib, Seaborn 라이브러리 사용
2. 개발 방법론

* Notion을 활용한 프로젝트 관리
* 프로젝트 계획 및 일정 관리

Notion을 사용하여 프로젝트의 전체 계획, 단계별 목표, 일정 관리. 이를 통해 팀원들이 프로젝트의 전반적인 진행 상황 파악 및 업무 조정

* 업무 할당 및 추적

각 팀원에게 할당된 업무와 그 진행 상태 관리. 이는 팀원들이 서로의 업무 진행 상황 파악

* 문서 공유 및 협업

프로젝트 관련 문서, 회의록, 리서치 자료 등을 Notion에 저장 및 공유. 이는 정보의 중앙화를 통해 자료 접근성을 높이고, 팀원 간 정보 공유 촉진

* + 피드백 및 의사소통

Notion의 댓글 기능을 활용하여 문서나 업무에 대한 피드백. 이는 실시간 의사소통 및 팀원들 간의 상호작용 강화

* Git을 활용한 협업
  + 소스 코드 관리

프로젝트의 모든 소스 코드는 Git을 통해 버전 관리함. 이를 통해 코드의 변경사항을 추적하고, 필요시 이전 버전으로 롤백 가능

* + 브랜치 전략

기능 개발, 버그 수정, 실험적 시도 등 각각의 목적에 따라 별도의 브랜치를 생성. 이는 코드의 안정성을 유지하면서 다양한 작업을 병행할 수 있도록 도움

* + 코드 리뷰 및 통합

Git의 머지(Merge) 기능을 사용하여 팀원들은 서로의 코드를 검토하고 피드백을 제공하여 코드의 품질 향상

**Ⅲ. 기술 및 기능**

1. 시스템 요구사항 확인 내용

* 하드웨어 요구사항: 적절한 처리 속도 및 메모리 용량을 갖춘 서버 또는 컴퓨터
* 소프트웨어 요구사항: Python 3.8이상, 필요한 라이브러리(Selenium, BeautifulSoup, Pandas 등)

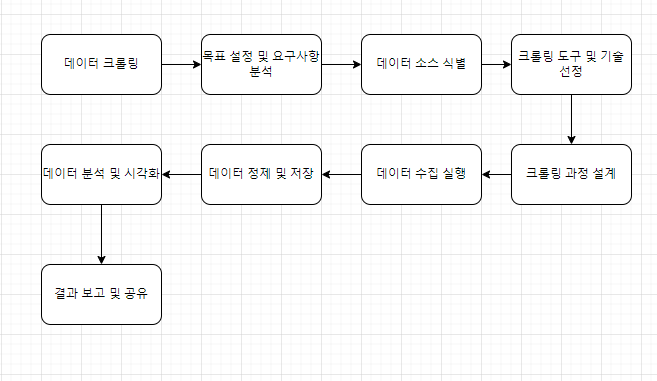
1. 기능 및 보안 요구사항 확인 내용

* 기능 요구사항: 크롤링 자동화 및 데이터 추출, 데이터 정제 및 분석, 결과 시각화
* 보안 요구사항: 수집된 데이터 암호화 저장

1. DB 구성도

* DataBase는 따로 구성하지 않음

1. 스토리보드



1. 제약사항

* 프로젝트 완료 기한: 2024년 2월 27일 화요일부터 2024년 3월 5일 화요일까지 영업일 기준 총 5일, 40시간
* 법적 및 윤리적 제약: 데이터 수집 시 개인정보 보호법 준수. 작성 코드는 공개 불가능한 내용을 제외하고 공유
* 기술적 제약: 천재교육 빅데이터 7기 강의 과정에서 배운 프로그래밍 언어 및 라이브러리 활용. 강의에 없는 기술 활용시 별도 명시함.

**Ⅳ. 성능 및 품질**

1. 성능요구사항 확인 내용

* 사용자 입력 없이 동작할 수 있도록 크롤링 프로그램 설계
* 시스템 안정성: 프로그램 실행 중 발생할 수 있는 예상치 못한 오류 코드 예외 처리.

2. 품질요구사항 확인 내용

* 데이터 정확성: 수집된 데이터의 오류가 없도록 함.
* 데이터 완전성: 중요한 데이터 필드는 누락 없이 수집되도록 함.
* 데이터 일관성: 수집된 데이터는 일관된 형식과 구조를 유지하며, 변환 과정에서 손실이 발생하지 않도록 함.

3. 인터페이스 제안

* 사용자 인터페이스: 사용자가 데이터 수집 및 분석 결과를 확인할 수 있도록 시각화
  + 범주형 데이터: Stacked barplot
  + 연속형 데이터: histogram
  + 업종 분류별 수요 차이 통계자료: barplot
* 소스코드는 코드의 가독성 및 유지보수를 위해 주석 작성

**Ⅴ. 프로젝트 관리**

1. 관리방법론

* 변경사항 있을 시 기존 진행 내역과 변경사항을 대조할 수 있도록 변경요청서를 작성하여 협의 후 진행.
* 프로젝트 문의사항이 있을 시 담당자에게 문의.

2. 일정계획

* 프로젝트 시작일: 2024년 2월 27일
* 데이터 수집 및 처리: 1일 + 4시간
* 데이터 분석 및 시각화: 1일 + 2시간
* 최종 보고서 작성 및 제출: 마지막 날에 완료
* 프로젝트 종료일: 2024년 3월 5일

**Ⅶ. 기타사항**

1. 업무분류체계(WBS) 엑셀파일 별도 첨부

2. 참고자료로 활용할 “2023 클라우드 실태조사 보고서” PDF파일 별도 첨부

3. 클라우드 시장 점유율 출처URL: “https://www.ddaily.co.kr/pag

e/view/2024021822092340107”