SK네트웍스 **Family AI**과정 **3**기

모델배포 시스템 구성도

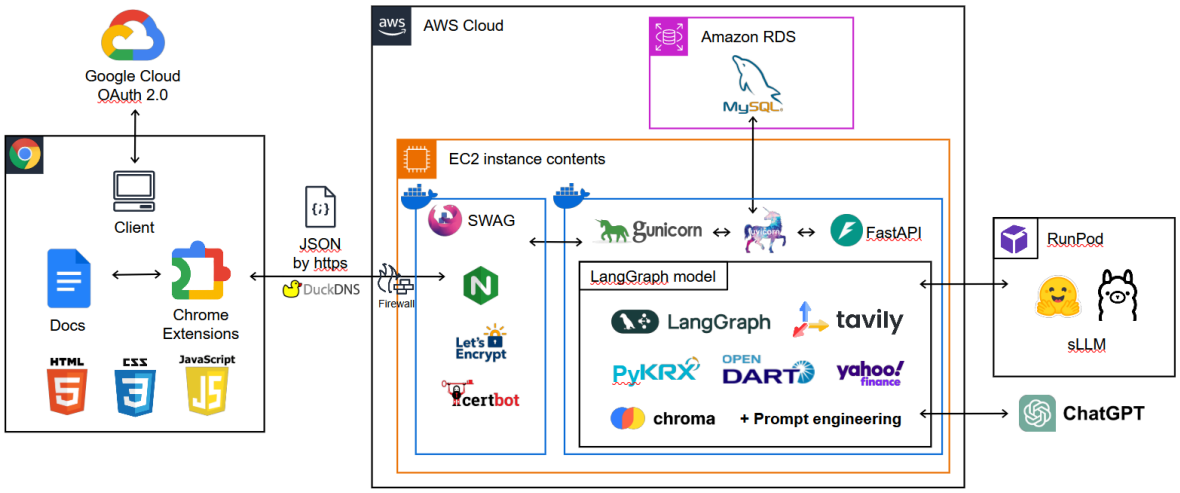
□ 개요

- 산출물 단계 : 모델배포
- 평가 산출물 : 시스템 구성도
- 제출 일자 : 2025. 02. 10
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN05-final-3Team
- 작성 팀원 : 허상호

개요	• 시스템 구성도
구성 요소	• 1. 클라이언트(Client) • 2. 데이터 전송 • 3. AWS Cloud • 4. 데이터베이스 • 5. 외부 서비스
데이터 흐름	• 데이터 흐름 및 통신 경로

개요

- 시스템 구성도



구성 요소

1. 클라이언트(Client)

- **Chrome 브라우저**: 사용자가 웹 애플리케이션에 접근하는 주요 인터페이스.
- **Docs**: Google Docs와 같은 문서 편집 애플리케이션을 통해 데이터를 처리.
- **Chrome Extensions**: 브라우저 확장 프로그램으로, 사용자 경험을 확장하거나 특정 기능(예: 데이터 전송)을 제공.

- **기술 스택**:
  - **HTML, CSS, JavaScript**: 클라이언트 측에서 사용자 인터페이스(UI)와 동작을 구현하는 기본 기술.
- **OAuth 2.0**:
  - Google Cloud를 통해 인증 및 권한 부여를 관리하며, 안전한 사용자 인증을 제공.

2. 데이터 전송

- **JSON by HTTPS**:
  - 클라이언트와 서버 간 데이터는 RESTful API 기반 JSON 형식으로 전송되며, HTTPS 프로토콜을 통해 암호화된 안전한 통신이 이루어짐.
- **DuckDNS**:
  - 동적 DNS 서비스로, 클라이언트가 서버에 안정적으로 접근할 수 있도록 도메인 이름을 관리.
- **FireWall**

- 데이터는 방화벽을 통해 허용된 포트의 요청(443, 80, 22)만 통과.

3. AWS Cloud

AWS 클라우드 환경은 애플리케이션의 서버와 데이터베이스를 호스팅하며, 주요 구성 요소는 다음과 같음:

EC2 인스턴스

- AWS의 가상 서버로, 애플리케이션 실행 환경을 제공.
- **SWAG (Secure Web Application Gateway)**:
  - Nginx와 함께 동작하며 HTTPS 요청을 처리하고 리버스 프록시 역할을 수행함.
- **Nginx**:
  - 웹 서버로 사용되며, 정적 파일 제공과 리버스 프록시 역할을 함.
- **Let's Encrypt & Certbot**:
  - 무료 SSL 인증서를 발급 및 갱신하여 HTTPS 통신을 지원함.

애플리케이션 레이어

- **Gunicorn**:
  - Python 기반 WSGI(Web Server Gateway Interface) 서버로 FastAPI 애플리케이션을 실행.
- **FastAPI**:
  - uvicorn 기반 백엔드 애플리케이션 프레임워크로, API 요청을 처리하고 데이터를 반환.

LangGraph 모델

- 애플리케이션 내에서 사용되는 언어 모델 및 관련 데이터 소스를 포함합니다:
  - **PyKRX**: KRX에서 실시간 금융 데이터를 제공하는 모듈.
  - **Open DART**: 공시 및 대한민국의 기업 데이터를 처리하는 모듈.
  - **Yahoo Finance**: 외부 금융 데이터를 가져오는 API.
  - **ChromaDB, Prompt Engineering**: LLM 기반 RAG 모델의 성능 최적화를 위한 벡터 데이터베이스 및 프롬프트 엔지니어링 도구.

4. 데이터베이스

- **Amazon RDS(MySQL)**:
  - 관계형 데이터베이스 서비스로, 애플리케이션의 데이터를 저장하고 관리.

5. 외부 서비스

RunPod

- AI 모델(sLLM)을 실행하는 외부 컴퓨팅 플랫폼으로 사용.

ChatGPT

- OpenAI의 언어 모델 API를 활용하여 자연어 처리 기능을 제공.

데이터 흐름

- 데이터 흐름 및 통신 경로

1. 사용자가 Chrome 브라우저(또는 확장 프로그램)를 통해 요청을 전송함.
2. 요청은 DuckDNS를 통해 AWS EC2 인스턴스로 전달됨.
3. Nginx가 요청을 받아 Gunicorn과 FastAPI로 전달하여 처리함.
4. 필요한 경우 Amazon RDS(MySQL)에서 데이터를 저장 및 조회하고, LangGraph 모델과 외부 API(Yahoo Finance 등)를 호출함.
5. AI 관련 요청은 RunPod(sLLM) 또는 ChatGPT로 전달됨.
6. 최종 결과가 JSON 형식으로 클라이언트에 반환됨.