

**SK네트웍스 Family AI 과정 10기  
 모델배포 시스템 구성도**



| **산출물 단계** | 모델배포 |
| --- | --- |
| **평가 산출물** | 시스템 구성도 |
| **제출 일자** | 2025.09.26 |
| **깃허브 경로** | https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN14-FINAL-6Team |
| **작성 팀원** | 서은선 |

| **개요** | * AWS 시스템 구성도 |
| --- | --- |

| **구성 요소** | * **AWS Cloud**   **1. CI/CD Pipeline :**  코드 빌드(Build) 및 배포(Deploy)를 자동화하는 파이프라인.  **2. Docker Image Registry (Docker Hub) :**  어플리케이션의 컨테이너 이미지를 보관하는 레지스트리. Elastic Beanstalk/EC2가 배포 시점에 이미지를 가져와 실행.  **3. Elastic Beanstalk/EC2 Instance :**  EB 환경에서 EC2가 실행되고, 인스턴스의 Docker가 Uvicorn + FastAPI + LangGraph 컨테이너를 구동.  **4. Docker Container** : 애플리케이션을 컨테이너로 실행.  **5**. **Elastic Load Balancer :** 외부 트래픽을 받아 EC2/컨테이너로 라우팅.  **5.** **Amazon RDS with MySQL :** 관계형 DB. 백엔드가 트랜젝션/조회 작업을 mysql로 읽고 씀. 백업/모니터링/패치가 자동화됨.  **6. Uvicorn :** ASGI 서버. FastAPI 앱을 고성능 비동기 방식으로 구동하며, 컨테이너 내부에서 외부로 HTTP 요청을 받아 처리.  **7. FastAPI** : 웹 프레임워크.   * **Model**  1. OpenAI: 생성/요약/분류 LLM. 2. Tavily: AI에이전트와 워크플로우를 위해 실시간 검색 엔진 제공. 3. Pinecone: FAQ/이용약관에 대한 벡터 검색(임베딩) 저장.  * **DNS 및 HTTPS 통합**   **1.**  **Amazon Route 53:** DNS 서비스. 사용자의 도메인 요청을 ELB로 라우팅해 접속 경로를 만듦. |
| --- | --- |
| **데이터 흐름** | * **개발자 작업**   1. 기능 개발·리팩터링·테스트를 로컬(VS Code)에서 수행.  2. Github에 코드 푸시 및 머지.  3. Dockerfile로 이미지를 빌드하고 Docker Hub에 푸시.  4. 배포 시 Elastic Beanstalk/EC2가 Docker Hub에서 대상 이미지를 pull하여 컨테이너를 재가동.     * **사용자 요청**   1. DNS(Route 53)를 통해 사용자 요청을 Elastic Load Balancer(ELB)로 요청 전달.  2. ELB는 타깃 그룹의 상태를 확인해 EC2 인스턴스로 트래픽 라우팅.  3. 컨테이너 내 FastAPI가 요청을 처리하여 필요 시 OpenAI/Tavily/Pinecone 및 RDS와 통신.  4. HTTPS로 안전한 통신 지원 |