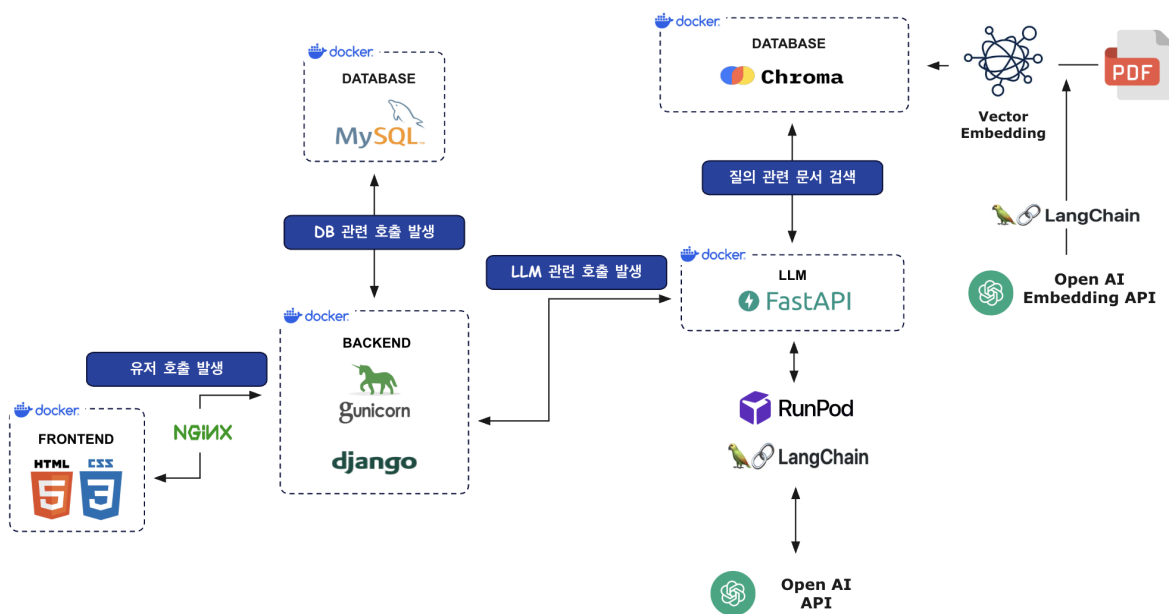


AI 푸드닥터(푸드) 시스템 구성도

AI 푸드닥터(푸드)은 AI 기반 기업 맞춤형 식단 도시락 서비스 및 영양 관리 플랫폼으로, LLM(대규모 언어 모델)을 활용하여 개인화된 식단 추천 및 구독, 영양 분석, 실시간 질의응답(Q&A) 기능을 제공한다. 문서 검색 기능을 도입함으로써 AI가 식단 및 영양 정보를 효율적으로 검색하여 사용자 맞춤형 정보를 제공할 수 있도록 설계하였다.



1. 데이터 전송

- 기능
 - 클라이언트(프론트엔드)와 백엔드 간의 데이터 요청 및 응답을 처리
 - 백엔드에서 데이터베이스(MySQL, ChromaDB) 및 LLM 서버(FastAPI)와 통신
 - API 요청 및 응답은 HTTPS를 통해 암호화하여 보안성 유지
 - OAuth2 및 JWT 기반 사용자 인증 및 권한 관리
- 기술스택
 - RESTful API

- 클라이언트와 서버 간 효율적인 데이터 교환 가능
- JSON 포맷을 활용한 빠른 데이터 전송
- TLS/SSL 암호화
 - HTTPS를 적용하여 네트워크 통신 중 데이터 암호화 및 보안 강화
 - TLS/SSL을 통해 데이터 도청 및 위변조 방지
- OAuth2 & JWT
 - OAuth2를 활용한 사용자 인증 및 API 접근 권한 부여
 - JWT(Json Web Token)를 사용하여 인증 상태를 유지하며 보안 강화

2. 프론트엔드 - 사용자 인터페이스 및 요청 처리

- 기능
 - 웹 브라우저에서 서비스 접속 및 사용자 인터랙션 수행
 - AI 추천 식단 조회, 식사 구독 예약, 영양 정보 확인, 챗봇 서비스 제공
 - 백엔드 API와 통신하여 사용자 요청을 처리
- 기술스택
 - HTML, CSS, JavaScript (React.js)
 - HTML: 웹 페이지의 구조 정의
 - CSS: 디자인 및 스타일링 적용
 - React.js: 컴포넌트 기반 UI 개발 및 상태 관리, 가상 DOM 활용으로 렌더링 성능 최적화
 - Nginx
 - 정적 파일 제공 및 리버스 프록시 기능 수행
 - 웹 서버와 백엔드 서버 간 부하 분산을 통한 트래픽 관리 최적

3. 백엔드 - 사용자 요청 처리 및 데이터 관리

- 기능
 - 사용자 요청을 수신하고, 데이터베이스(MySQL, ChromaDB) 및 LLM과 연동하여 응답 생성
 - AI 기반 식단 추천 로직을 저장 및 관리
 - 사용자 인증 및 접근 제어 수행
- 기술스택
 - 프레임워크: Django + FastAPI
 - Django: 데이터 모델링 및 ORM(Object-Relational Mapping) 활용, REST API 엔드포인트 제공
 - FastAPI: 비동기 방식의 API 처리로 빠른 응답 속도 보장, AI 모델과의 통신 최적화
 - 웹 서버: Gunicorn + Nginx
 - Gunicorn: Django 애플리케이션을 실행하는 WSGI 서버
 - Nginx: 리버스 프록시 및 정적 파일 제공, Gunicorn과의 연동을 통해 트래픽 관리

4. 데이터베이스 - 사용자 및 식단 데이터 저장

- 기능
 - 사용자 프로필 데이터 저장
 - 사용자 선호도를 반영한 AI 식단 추천 내역 및 구독 신청 내역 저장
 - 벡터 검색을 위한 문서 임베딩 데이터 저장
 - 카테고리별 음식 영양성분 데이터 저장

- 기술스택
 - 관계형 DB: Amazon RDS (MySQL)
 - 정형 데이터를 저장하고 ACID(Atomicity, Consistency, Isolation, Durability) 트랜잭션 지원
 - Amazon RDS를 활용하여 자동 백업 및 가용성 보장
 - 벡터 DB: ChromaDB
 - 고속 벡터 검색 및 AI 기반 정보 검색 최적화
 - LLM이 활용할 문서 데이터(PDF)를 벡터화하여 저장 및 검색 수행

5. LLM 서비스 - AI 기반 맞춤형 식단 추천 및 Q&A 처리

- 기능
 - OpenAI API 및 자체 sLLM을 활용하여 사용자 맞춤형 식단 추천
 - RAG (Retrieval-Augmented Generation) 방식으로 ChromaDB를 활용한 식단 조합 생성
 - LangChain을 활용한 챗봇 질의응답 및 대화 흐름 관리
- 기술스택
 - LLM 운영: RunPod
 - GPU 기반 클라우드 환경에서 AI 모델을 실행하여 높은 연산 성능 제공
 - 모델 학습 및 추론을 위한 최적의 환경 제공
 - OpenAI API
 - GPT 기반의 자연어 처리 기능을 제공
 - 대화형 인터페이스에서 사용자의 요구를 해석하고 응답 생성
 - 문서 검색: Chroma DB + Langchain
 - ChromaDB: 벡터 기반 문서 검색 최적화
 - LangChain: 대화 흐름 관리 및 사용자 질의응답 처리

6. 데이터 임베딩 및 문서 검색 - 사내 문서 기반 Q&A 시스템

- 기능
 - PDF 파일 내 식단 관련 정보를 추출하고 벡터화하여 ChromaDB에 저장
 - 사용자의 질의에 맞는 문서를 검색하고, LLM을 활용한 동적 응답 생성
- 기술스택
 - 임베딩 모델: OpenAI Embedding API
 - 텍스트 데이터를 벡터화하여 빠른 검색 및 AI 모델 학습 지원
 - 고유한 의미를 파악하여 문서 검색 성능 향상
 - 프레임워크: Langchain
 - PDF 문서를 작은 단위(chunk)로 나누고 벡터화하여 검색 최적화
 - 자연어 질의에 적절한 문서를 검색 후 AI가 답변을 생성하도록 지원
 - ChromaDB
 - 대규모 벡터 데이터를 저장하고 검색하는 NoSQL 기반 데이터베이스
 - AI 모델이 활용할 수 있도록 문서 데이터의 고속 검색 지원