SKN13-FINAL-4Team

3. 모델링 및 평가

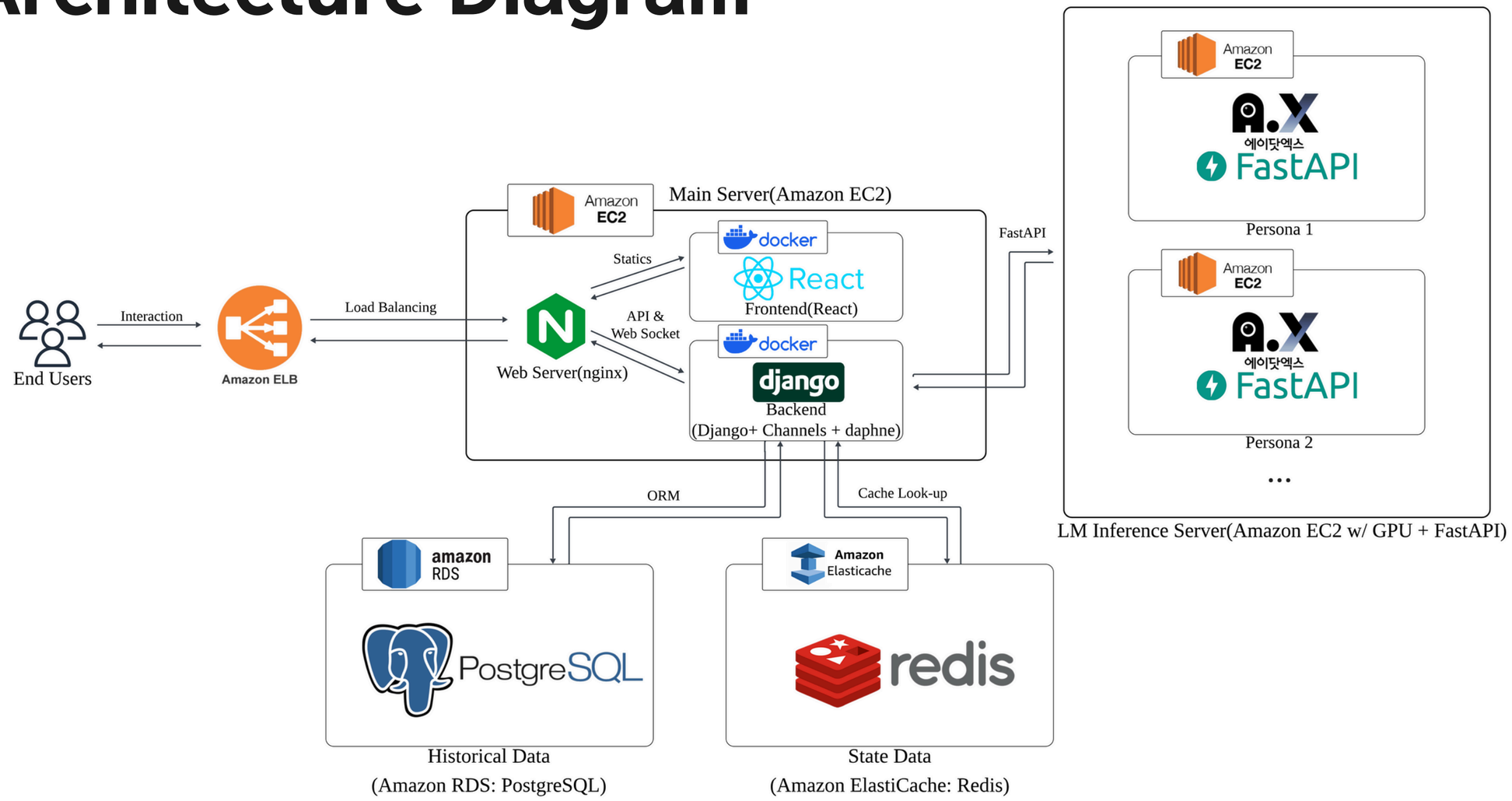
System Architecture

2025-08-22 (금)

작성자: 이재범

https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-4Team

System Architecture Diagram



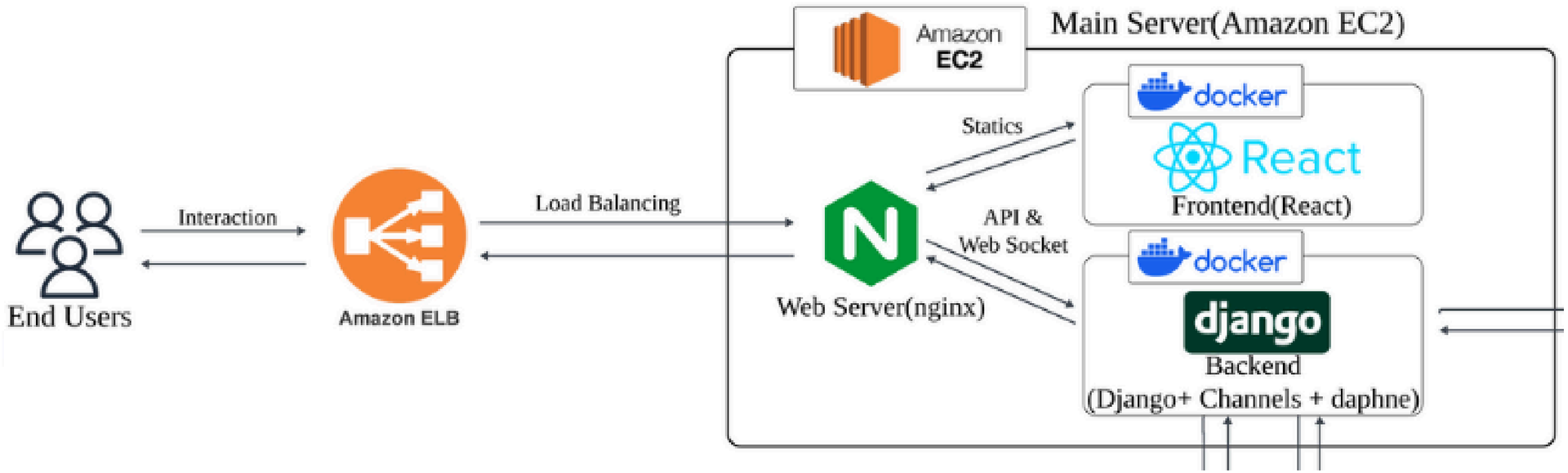
System Architecture Diagram

주요 고려 사항

- 스트리밍 서비스 고유의 특징
  - 특정 시간대에 집중되는 부하
  - 실시간 1 대 다 의사소통 시스템 구현

- LM(Language Model) 추론 과정과 결과 처리의 신속성
  - 메인 서버 – LM 추론 서버 분산 구조 상에서의 속도 최적화

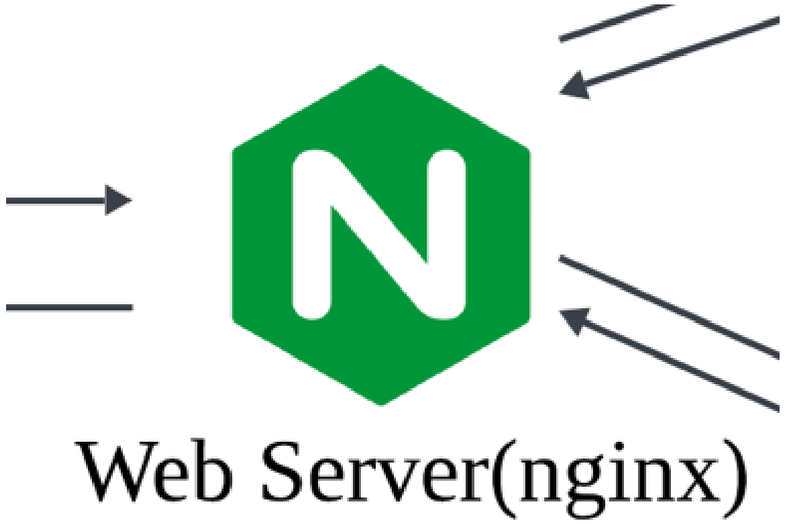
Details: Main Server



Main Server: Amazon EC2

- 높은 설정 자유도, Amazon 생태계 내 타 서비스와의 유연한 연계
- 1개 서버 내 수용 가능한 부하 추정치: 500
- Amazon ELB를 함께 운영하여 적절한 부하 분산 시스템 구축
- 각 하위 시스템은 docker 컨테이너로 패키징하여 실행 환경의 일관성 보장

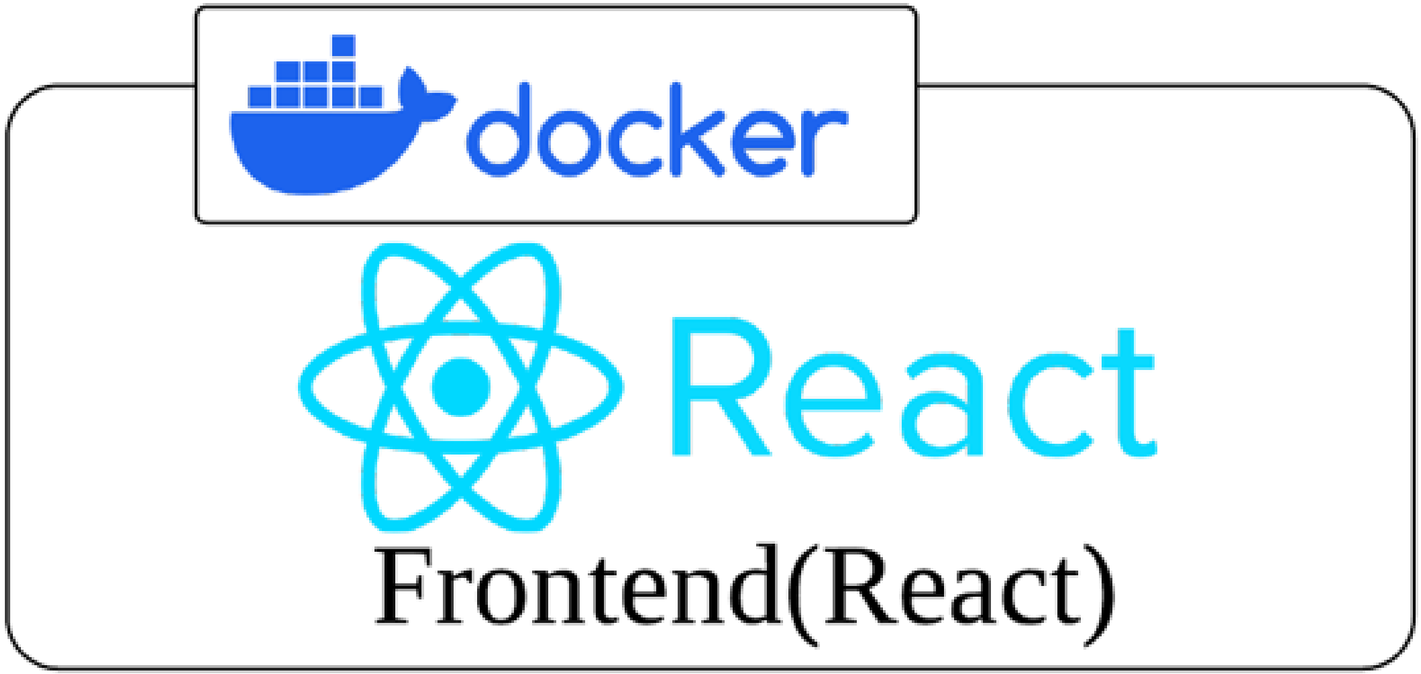
Details: Main Server



Web Server: nginx

- End Users의 요청에 응답하기 위해 필요한 자원을 적절한 하위 시스템에 분산 요청
- 비동기 이벤트 기반의 효율적인 동시 처리
- Reverse Proxy 고유의 강력한 WAS/App 보안

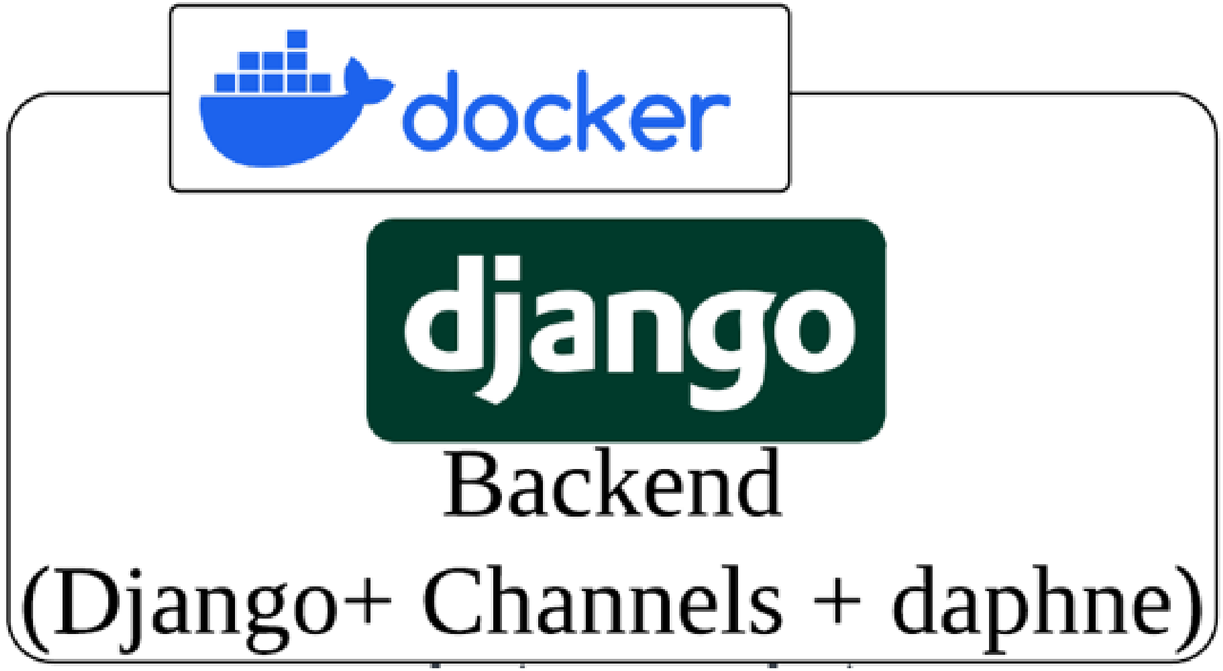
Details: Main Server



Frontend: React

- 화면, 정적 파일 처리 후 웹 서버에 전달
- Virtual DOM을 통한 압도적 반응 속도
- Component 기반 구조의 UI 일관성

Details: Main Server



Backend: Django + Channels + daphne

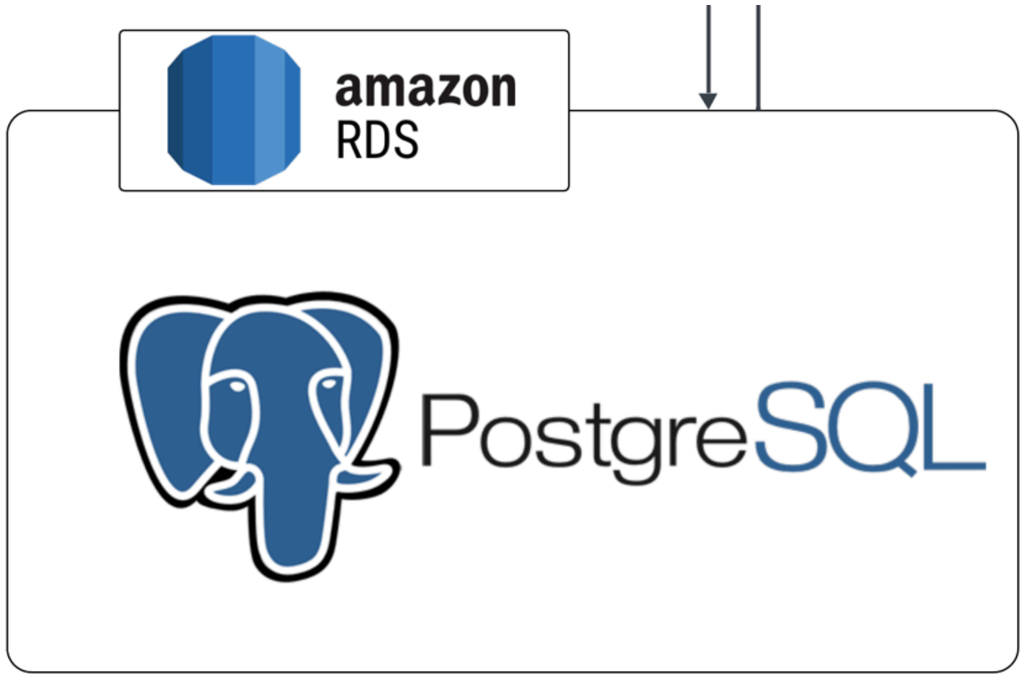
- Django: API(HTTP) 통신 담당. PostgreSQL에 영구적으로 저장될 데이터 생성 및 조회.
- Channels: WebSocket 기반 지속적 양방향 통신 담당. Redis를 통해 실시간 메시지를 모든 유저에게 전달.
- daphne: 웹 서버로부터의 모든 요청을 Django 또는 Channels로 적절히 분산하는 웹 어플리케이션 서버

Details: Main Server

Backend: Django Apps

- config: 프로젝트 총괄
  - Django 프로젝트 기본 설정 및 WebSocket 통신용 ASGI 서버 설정
- chat:실시간 채팅 기능과 스트리밍 기능 담당
  - 실시간 채팅 로직, 스트리머 음성 구현(TTS API)
- users: 사용자 인증 및 프로필 관리
  - 회원 정보 데이터 정형화(JSON) 및 API 제공

Details: Databases



Historical Data: Amazon RDS(PostgreSQL)

- PostgreSQL: 오픈소스 RDB, 복잡한 쿼리 수행 능력 우수, 엄격한 데이터 무결성&안정성
- 사용자 기본 정보, 결제 정보, 채팅 정보, 방송 정보 등 영속적 데이터 저장소
- Django ORM으로 데이터 생성 및 조회

Details: Databases

State Data: Amazon ElastiCache(Redis)

- 실시간 WebSocket 통신을 위한 메시지 브로커 및 고속 캐시 저장소
  - 별도의 메시지 브로커 전용 솔루션이 필요 없는 통합 서비스
- Channels를 통해 전달되는 실시간 채팅 메시지를 모든 클라이언트에 Broadcast
- 자주 조회되는 데이터의 복사본을 cache로 저장하여 메인 DB 부하 완화 및 데이터 처리 속도 향상

Details: LM Inference Servers

LM Inference Server: Amazon EC2

- 각 페르소나 LM(A.X-4.0-Light)을 담당하는 Amazon EC2 운영
- g5.xlarge(24GB VRAM) 인스턴스 사용하여 LM 추론 속도 보장
- FastAPI 통신으로 추론 결과를 신속하게 메인 서버로 전달
  - Python 웹 프레임워크 중 최고 성능
  - Python type hint기반 간단한 데이터 검증