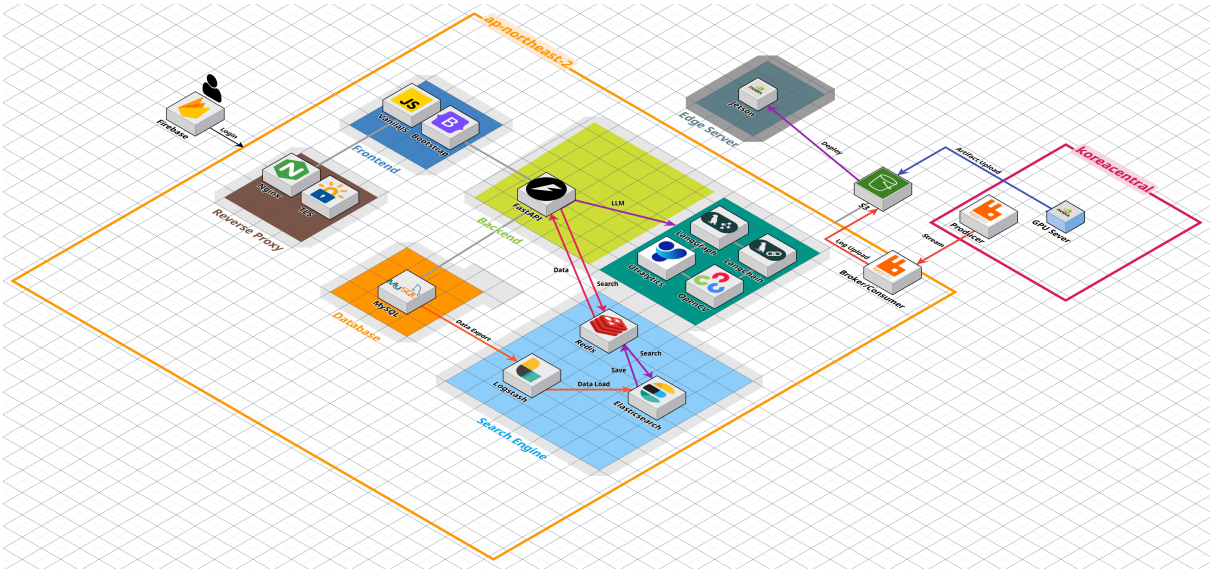
서비스 아키텍처



VisionInApp - AI 기반 객체 탐지 모델 전주기 통합 플랫폼

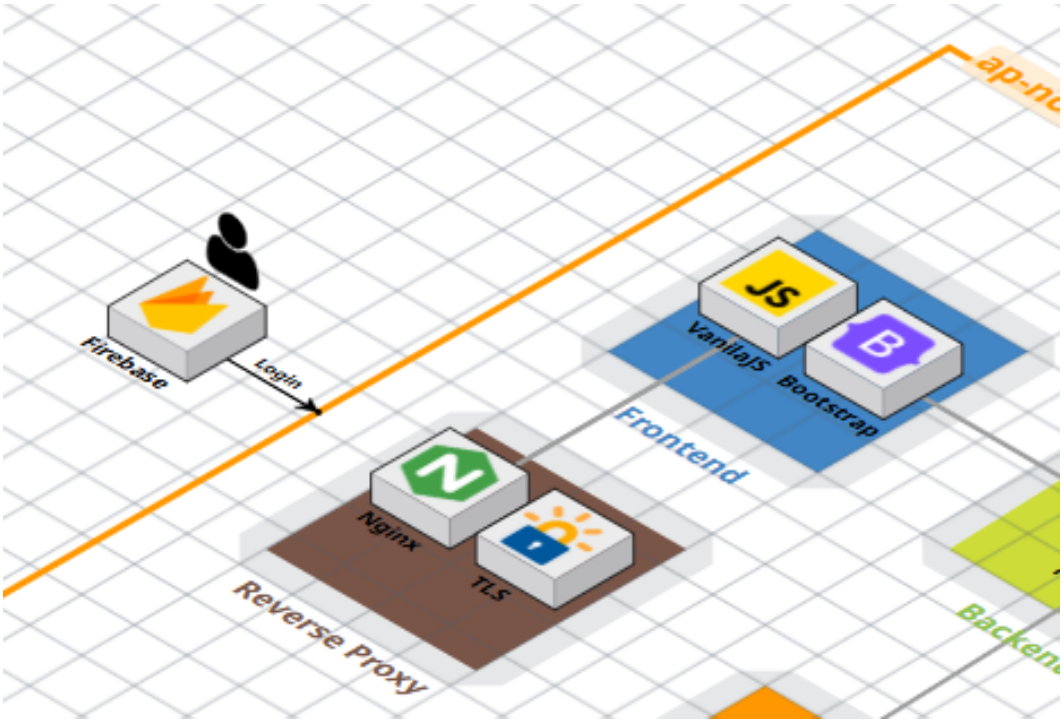
전체 개요

본 아키텍처는 **AWS(ap-northeast-2)** 지역과 **Azure(korea-central)** 을 혼합 사용하는 하이브리드 구조로,

모델 학습–배포–피드백 루프가 자동으로 순환하는 형태입니다.

클라우드 의존도를 최소화하면서도 **온디바이스/엣지 디바이스(Jetson)** 에 모델을 배포할 수 있는 구조로 설계되었습니다.

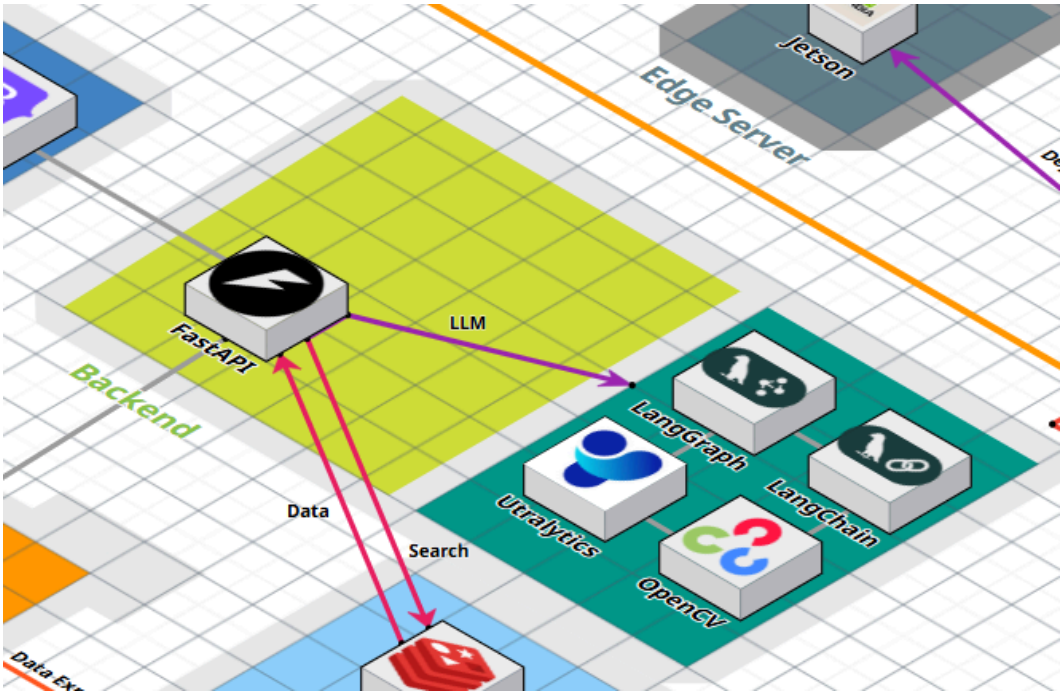
주요 구성 요소



1️⃣ Frontend (Vanilla JS + Bootstrap) / Reverse Proxy (Nginx + TLS)

- 사용자는 웹 브라우저를 통해 모델 관리, 학습, 배포를 직관적으로 조작할 수 있습니다.
- Firebase Auth 기반으로 로그인 인증을 수행합니다.
- FastAPI 백엔드와 REST API로 통신합니다.
- HTTPS 인증서를 통한 TLS 암호화를 제공합니다.
- Frontend, Backend, API 엔드포인트로의 요청을 라우팅합니다.
- SSL 인증서 관리 및 트래픽 부하 분산 담당.
- **인증 및 사용자 관리**

  : Firebase Authentication (OAuth 2.0)

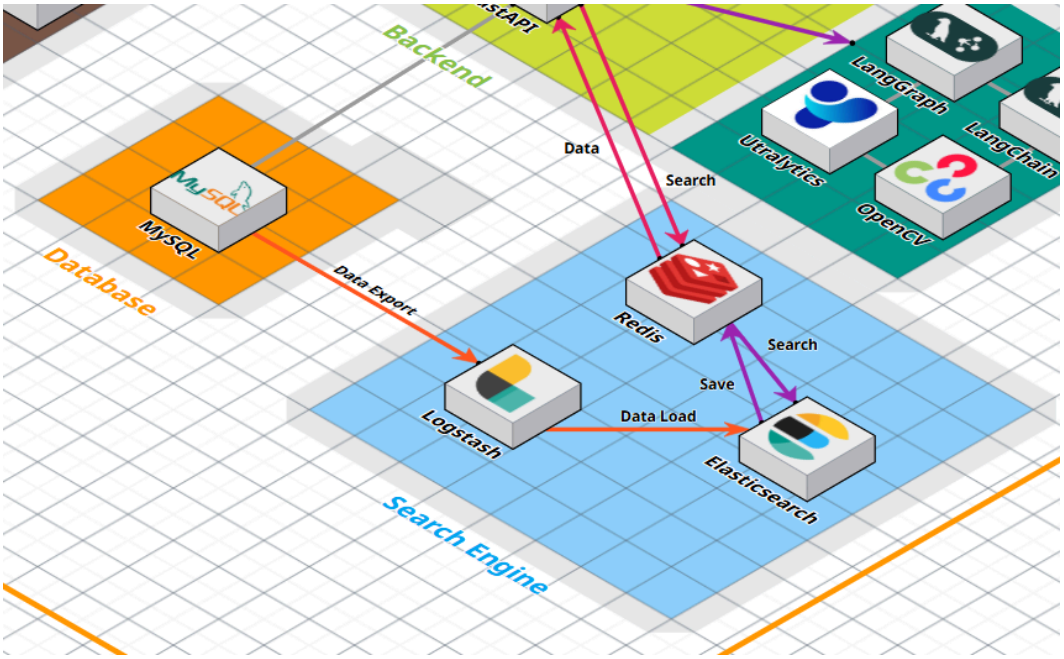


2️⃣ Backend (FastAPI + Python)

- 플랫폼의 핵심 로직이 구현된 영역으로, 다음 기능들을 담당합니다:
  - **데이터 처리 및 모델 학습 요청**
  - **LLM 및 LangGraph 기반 워크플로우 관리**
  - **Redis 및 Elasticsearch 연동을 통한 고속 검색 및 캐싱**
  - **RabbitMQ Stream을 통해 GPU 서버와 실시간 통신**

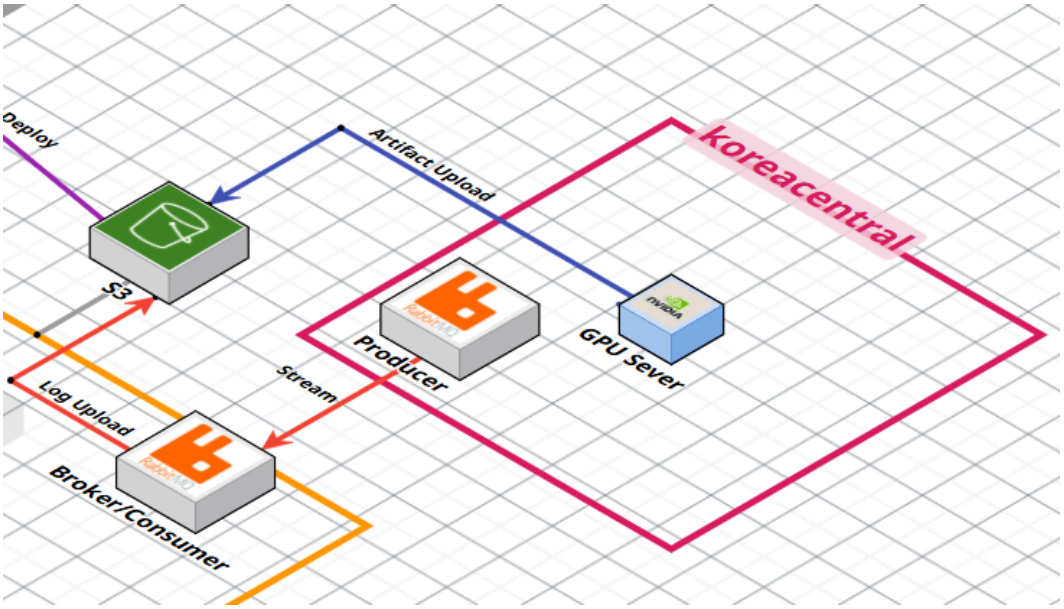
3️⃣ AI / LLM 모듈 (LangGraph, LangChain, OpenCV)

- FastAPI 내부에서 호출되어:
  - 데이터 전처리 (OpenCV)
  - 모델 학습/튜닝 (LangGraph, LangChain)
  - LLM 기반 모델 설명/결과 요약 기능 제공



4 Database Layer

- **MySQL:** 프로젝트, 데이터셋, 주석 이력 등 구조화된 데이터 관리
- **Redis:** 학습/검색 결과 캐싱 및 임시 데이터 저장
- **Elasticsearch:** 고속 검색엔진 (Logstash 파이프라인으로 MySQL → Elasticsearch 데이터 전송)



5 Message Broker (RabbitMQ)

- **Producer (GPU Server):** 학습 완료 후 결과 로그 및 모델 파일을 S3에 업로드

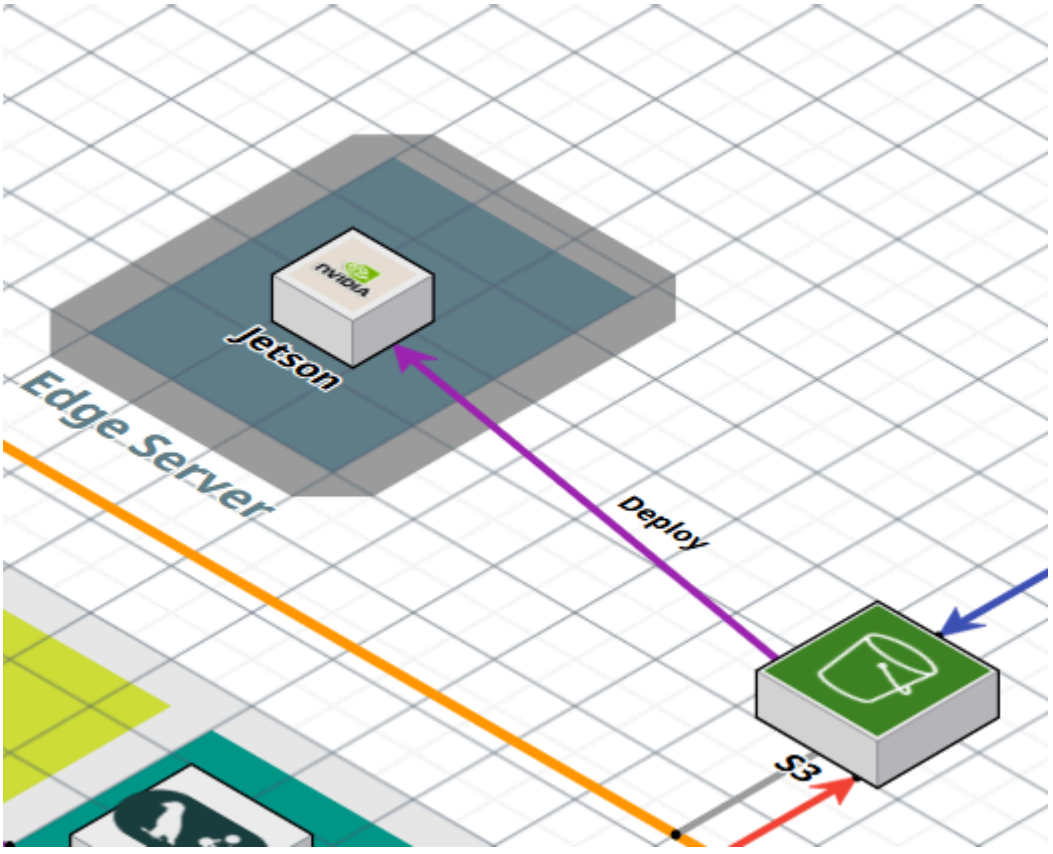
- **Consumer (Backend 데몬)**: S3 업로드 이벤트를 감지하고 MySQL/Redis/Elasticsearch로 메타데이터를 반영
- **Streams Plugin**을 사용하여 대용량 로그와 모델 메타데이터를 안정적으로 전송

6️⃣ GPU Server (korea-central)

- YOLOv12/YoloX 기반 객체 탐지 모델을 학습하는 전용 환경
- 학습 완료 시 결과(가중치, 로그 등)를 **S3에 업로드**
- FastAPI는 S3에 업로드된 artifact를 받아 모델 배포 및 추론에 활용

7️⃣ S3 Storage

- 학습 결과(모델 가중치, 로그 등)와 업로드된 데이터셋 저장소
- Jetson Edge Device로 배포될 모델 아티팩트를 관리



8️⃣ Edge Server (Jetson Orin Nano)

- 클라우드에서 학습된 모델을 **온디바이스로 배포 및 추론 수행**

기술 스택 요약

구분	기술 스택
Frontend	Vanilla JS, Bootstrap
Backend	FastAPI (Python)
AI/ML	PyTorch, LangChain, LangGraph, OpenCV
Database	MySQL, Redis
Search Engine	Elasticsearch + Logstash
Broker	RabbitMQ (Streams)
Cloud	AWS(ap-northeast-2), Azure(korea-central)
Storage	S3
Edge	NVIDIA Jetson Orin Nano
Auth	Firebase OAuth2.0
Proxy	Nginx + TLS

아키텍처 요약

✅ **End-to-End 통합성:** 데이터 수집부터 배포까지 하나의 플랫폼에서 자동화

✅ **Hybrid Cloud 구조:** GPU 서버와 Edge 디바이스를 연동한 유연한 배포

✅ **실시간 피드백 루프:** 학습→배포→운영→재학습 자동 순환

✅ **보안성:** Firebase 인증 + Private 네트워크 환경

✅ **확장성:** LLM 및 다양한 모델 포맷(ONNX/TensorRT) 대응 가능