

박동진

서버/백엔드 개발자 | 01041674419 | straycat405@gmail.com

 Github

 Blog

 포트폴리오



MSA 설계와 인증 시스템 구현을 통한 확장성과 안정성을 중시하는 백엔드 개발자입니다. 빅팀INC(주)에서 LLM 챗봇 플랫폼 백엔드 & 프론트엔드를 개발했고, 내일배움캠프 JAVA 단기심화 과정을 통해 Spring Cloud 기반 MSA 프로젝트를 경험했습니다. 실무와 사이드 프로젝트에서 배운 설계 고민과 트러블슈팅을 기록하고 있습니다. 최근에는 AI와 AX에 관심을 갖고 학습중에 있습니다.

프로젝트

Auto Translator

2026.07~ 2026.07 (1개월)

 Github

- (주)딥라이트 인턴십 기간 단독 설계,개발
- LLM 번역 결과물 초안을 검수·검증 후 PostgreSQL에 반영하는 4개 언어 번역 운영 콘솔 (1인 개발, 백엔드·인프라 담당)
- 핵심 과제는 운영 DB의 영향, 부하를 최소화하는 것
- Python, PostgreSQL, SQLite, psycopg3, OpenAI API, Gemini API, JavaScript, HTML, Docker

운영 DB 공유 배치의 락 점유·Worker 중복 실행 차단

- 번역 Worker가 운영 서비스와 동일 PostgreSQL 공유, 배치 UPSERT 락 장기 점유 시 서비스 트래픽 지연·중복 실행 위험 존재
- 세션 단위 lock_timeout 1.5초·statement_timeout 5초 적용으로 배치 쿼리 fail-fast 처리, session advisory lock 도입으로 인스턴스 하나만 Worker 실행·나머지 안전 취소
- connection pool min 1·max 2 고정으로 LLM 응답 대기 중 저장 연결 미점유, pool 대기 1초 초과 시 TranslationWriteError로 수동 처리 이관, __exit__ pg_advisory_unlock_all로 락 누수 차단
- 배치 유발 서비스 쿼리 최대 대기 무제한 → 1.5초 상한, 인스턴스 간 Worker 중복 실행 상시 위험 → 0건

LLM 요청 병렬화와 저장 경로 직렬화로 pool 고갈 방지

- 자동 배치가 테이블·chunk를 순차 LLM 호출, 응답 대기 누적으로 실행 시간 과다 발생, 저장 연결 2개뿐이라 무분별 병렬화 시 pool 고갈 위험
- ThreadPoolExecutor로 LLM 요청 동시성 3 제한 rolling window 실행, FIRST_COMPLETED wait로 완료 즉시 다음 chunk 제출
- save_lock 기반 단일 writer로 저장 직렬화, chunk별 immutable source_items·독립 API error context 분리로 동시 실패 상태 혼선 차단, preserve_existing 적용으로 선행 저장 행 덮어쓰기 skip
- LLM 요청 순차 처리 → 동시성 3 병렬로 실행 시간 약 60% 단축, 저장 경로 pool 연결 경합 → 단일 writer 직렬화로 0건

first ticket

2026.04~ 2026.05 (2개월)



공연 티켓팅 플랫폼 MSA 백엔드

- 6인 팀
- Gateway · User Service · Eureka-Keycloak 인증 전체 담당
- Java 21 / Spring Boot 3.5.13 / Spring Cloud 2025.0.2 / Keycloak 26 / Redis / Kafka KRaft / AWS ECS Fargate / nGrinder

Redis 스파이크 95.8% 감소 / 로그인 TPS +796% 개선 / ECS Fargate 운영 배포

- Caffeine L1(TTL 5s)+Redis L2 블랙리스트 캐시 → MTT 스파이크 24회→1회 (-95.8%)
- Redis AT Cache → Keycloak ROPC 병목 우회, TPS 175→1,568(+796%), 응답 135ms→64ms
- Resilience4j Bulkhead 기본값 25 발견 및 수정 → 안정 vuser 25→125 이상

로컬 RAG 챗봇

2026.02~ 2026.03 (2개월)



- 8GB VRAM 개인 PC에서 문서 외부 유출 없이 동작하는 로컬 RAG (1인 개발)
- 백엔드·프론트엔드·인프라 전 영역 담당
- Java 17, Spring Boot 3.5, Spring AI 1.1.6, Spring Security, Ollama(qwen3:8b), pgVector, PostgreSQL 17, bge-m3, React 18, TypeScript, Vite, Docker Compose, GitHub Actions

단일 GPU 동시성 제어로 콜드스타트 37.2s→1.2s 개선

- 8GB VRAM 단일 GPU에 다중 채팅 요청 동시 유입 시 동시 실행 제어 부재로 GPU 큐 무제한 적재, k6 부하 시 에러율 100% 발생
- OllamaGenerationQueue 도입으로 동시 실행 1건 제한 + 대기열 상한 (BTLLM_GPU_QUEUE_CAPACITY=8) 초과 요청 즉시 거부(fail-fast) 적용
- WebSocketHandler에 세션별 in-flight cancellation 적용해 진행 중 생성 취소 지원, k6 부하 테스트로 before/after 검증
- 콜드스타트 37.2s→1.2s(97%), 에러율 100%→0%, 평균 지연 88.9s→33.0s(63%), p95 188.6s→74.5s(60%) 개선

청크 크기 재설계로 골든셋 통과율 0%→50% 개선

- 한국어 공고문에서 1500토큰 청크가 평균 5,425자로 생성, topK=3 근거 약 4,521토큰이 num_ctx 4096 초과로 컨텍스트 잘림 발생
- trackAttributionDiagnostic 도구로 근거 존재 여부·topK별 토큰량을 분리 관측해 '근거 부재'와 '근거 받고도 실패' 구분
- TokenTextSplitter 청크 크기 1500→800토큰 조정(문장 단위 분할 유지) + 문서 개요 요약 청크 추가로 개요형 질의 커버
- goldenSet 자동 평가 하네스로 통과율 0.0%→50.0%, 정답 핵심내용 포함률 38.1%→76.2%(gpt-4o-mini 동일 수준) 정량 검증

하이브리드 재정렬 구현으로 컨텍스트 오염 억제

- 벡터 유사도 단독 검색 시 임베딩만으로는 키워드 정확 일치 문서를 놓쳐 무관한 청크가 상위 노출-컨텍스트 오염 발생
- HybridReranker 구현으로 벡터 유사도 + 키워드 매칭 하이브리드 스코어 재정렬, RagSearchSettings로 파라미터 외부화
- 하위 검색 노이즈 억제 로직으로 무관 청크 유입 차단, HybridRerankerTest 단위 테스트(DB 불필요)로 재정렬 로직 검증
- 정답 핵심내용 포함률 38.1%→76.2%(100%) 개선

별자리 운세 사이트

2026.02~ 2026.02 (1개월)

 서비스 페이지  Github

개인 프로젝트 - 바이브 코딩 기반 일일 별자리 운세 사이트입니다.

Cloudflare를 통해 도메인 구입 후 배포했습니다.

일일 별자리 운세를 자동 크롤링해서 업데이트합니다.

대외활동 및 수상

내일배움캠프 JAVA 단기심화 6기

2026.02~ 2026.05 (4개월)

- 프로젝트 1 : 배달 주문 플랫폼 API
 - 모놀리식 아키텍처
 - 담당 : User 도메인, 인증 / 인가 도메인 / Github 관리
- 프로젝트 2 : B2B 물류 관리 플랫폼
 - MSA
 - 담당 : 상품 정보, 기업 정보 관리, Eureka server
- 프로젝트 3 : 콘서트 티켓팅 플랫폼 (최종 프로젝트)
 - MSA
 - 담당 : user-service, gateway-server, eureka-server 등

2024년 6차

전자정부표준프레임워크 온라인교육

2024.12~ 2024.12 (1개월)

- 한국지능정보사회진흥원 주최
 - 전자정부 표준프레임워크 v4.2
 - 표준프레임워크 및 공통컴포넌트 활성화와 활용도 제고 과정 수료
-

KH정보교육원 - 공공데이터 융합
자바개발자 양성과정
2023.11~ 2024.04 (6개월)

Java, JQuery, HTML, CSS, JSP, MyBatis, Spring, React, Node.js, Oracle, Git
등 기초 교육 과정 수료

팀프로젝트 - 웹 가게부 서비스 (팀장)

• 주요역할 : 사용자, 관리자 서비스 및 페이지 구축

경력 사항

딥라이트

인턴십
2026.07~ 2026.07 (1개월)

- QA업무 및 자동배치 번역 프로그램 개발
- AI 캐릭터 채팅 플랫폼 구축 (CCUC) 백엔드 및 프론트엔드 신규 기능 개발에 대한
검수 및 QA업무 담당
 - DB 다국어화 번역 작업을 위한 번역 프로그램 제작 및 개선작업

(주)빅팀아이앤씨

개발팀 연구원
2024.09~ 2025.06 (10개월)

- 공공기관 납품용 LLM 챗봇 플랫폼 개발
- 회사 자체 LLM 챗봇 개발 , 입찰설명회 발표 및 시연
 - Websocket 기반으로 LLM 응답 스트리밍 구현
 - 파일 업로드, 웹 크롤링, Excel 내보내기 등

- 현충시설정보서비스 웹사이트 유지보수
- 국가보훈부 담당 현충시설정보서비스 웹사이트 유지보수 업무
 - 주요 Frontend / Backend 버그 및 오류내용 확인 및 수정
 - DB 데이터 대조 및 최신화 반영 작업

- 주요 사용 기술스택
- Java Spring

- PostgreSQL
- Mybatis
- React
- JSP
- TailwindCSS
- Wildfly
- SVN

학력 사항

건국대학교

건축학과 학사

2020.03~ 2022.02(졸업)

평균 학점 : 4.16

건축설계, 건축디자인

졸업작품

- PERSONA - 파라메트릭 디자인을 이용한 노들섬 문화공간제안

외국어 및 자격증

1종보통운전면허

최종합격 2014.04

실내건축산업기사

최종합격 2019.08