Cloudturing

기술 아키텍처 백서

꿈많은청년들 CTO
cto@cloudturing.com

2025년 12월 16일

Abstract

본 문서는 Cloudturing 플랫폼의 핵심 기술과 아키텍처를 설명합니다. Google의 최신 Gemini AI 모델을 활용한 인텐트(Intent) 자동 생성 기술과 검색 증강 생성(RAG) 기반의 답변 처리 시스템을 소개합니다. 또한, Google Kubernetes Engine(GKE) 기반의 클라우드 네이티브 아키텍처를 통해 어떻게 대규모 트래픽을 안정적으로 처리하고 높은 가용성을 보장하는지 기술합니다.

Contents

1 개요

1.1 기술 비전

Cloudturing는 "누구나 쉽게 만드는 강력한 AI"를 목표로 합니다. 복잡한 AI 모델 튜닝이나 인프라 관리 없이, 비즈니스 로직에만 집중할 수 있는 환경을 제공합니다. 이를 위해 우리는 다음과 같은 기술적 가치를 추구합니다:

- **자동화 (Automation)**: 인텐트 분류부터 답변 생성까지 AI가 자동으로 수행하여 관리 소요를 최소화합니다.
- **확장성 (Scalability)**: 사용자 및 트래픽 증가에 유연하게 대응하는 클라우드 네이티브 아키텍처를 지향합니다.
- **최적화 (Optimization)**: 최신 LLM 기술을 경량화하고 캐싱 전략을 최적화하여 빠른 응답 속도를 보장합니다.

2 AI 핵심 기술 (Core Engine)

2.1 Gemini 기반 인텐트 자동 생성

Cloudturing는 Google의 멀티모달 AI 모델인 **Gemini**를 핵심 엔진으로 사용합니다. 단순한 키워드 매칭을 넘어, 사용자의 발화 의도(Intent)를 심층적으로 이해하고 자동으로 분류 체계를 생성합니다.

> **자동 학습 파이프라인**
> 사용자가 업로드한 문서(PDF, TXT 등)나 입력한 예시 문장을 분석하여, AI가 스스로 예상 질문과 답변 쌍을 생성합니다. 이는 기존의 수동 시나리오 설계 방식을 혁신적으로 단축시킵니다.

2.2 검색 증강 생성 (RAG)

할루시네이션(Hallucination, 환각) 현상을 최소화하고 정확한 정보만을 제공하기 위해 **RAG (Retrieval-Augmented Generation)** 기술을 적용했습니다.

1. **지식베이스 인덱싱**: 고객이 제공한 데이터를 벡터화하여 고성능 벡터 데이터베이스에 저장합니다.
2. **의미론적 검색 (Semantic Search)**: 사용자 질문이 들어오면 키워드가 일치하지 않더라도 의미적으로 가장 유사한 문서를 검색합니다.
3. **컨텍스트 주입**: 검색된 정확한 지식을 LLM의 프롬프트에 컨텍스트로 주입하여, 그에 기반한 사실적인 답변을 생성하도록 유도합니다.

3 시스템 아키텍처

3.1 클라우드 네이티브 디자인

본 서비스는 마이크로서비스 아키텍처(MSA)를 기반으로 설계되었으며, Google Kubernetes Engine(GKE) 상에서 운영됩니다.

계층	구성 요소
오케스트레이션	GKE Autopilot (Private Cluster, Workload Identity)
트래픽 관리	Google Cloud Load Balancer (IPv4/IPv6 Dual Stack)
네트워크	Cloud NAT (프라이빗 노드 아웃바운드 통신)
보안 및 가속	Cloud Armor (DDoS 방어), Cloud CDN
데이터 스토리지	Cloud SQL (PostgreSQL), BigQuery

Table 1: 시스템 아키텍처 계층

3.2 고가용성 및 오토스케일링

트래픽 변동에 능동적으로 대응하기 위해 수평적 자동 확장(HPA, Horizontal Pod Autoscaling)이 적용되어 있습니다.

- **Pod 오토스케일링**: CPU 및 메모리 사용량에 따라 서비스 인스턴스(Pod) 수가 자동으로 증감합니다.
- **무중단 배포**: 롤링 업데이트 방식을 채택하여 서비스 중단 없이 새로운 기능을 배포할 수 있습니다.
- **자가 치유 (Self-healing)**: 장애가 발생한 컨테이너를 자동으로 감지하고 재시작하여 서비스 연속성을 유지합니다.

4 성능 최적화 전략

4.1 글로벌 CDN 및 캐싱

전 세계 어디서든 빠른 챗봇 위젯 로딩을 위해 엣지 컴퓨팅 기술을 활용합니다.

- **정적 자산 가속**: 챗봇 아이콘, 스크립트 파일 등 정적 리소스는 Google Cloud CDN을 통해 사용자에게 가장 가까운 엣지 서버에서 전송됩니다.
- **인메모리 캐싱**: 빈번하게 요청되는 챗봇 설정 및 세션 정보는 Valkey(Redis 호환) 기반의 인메모리 저장소에서 초고속으로 처리됩니다.

4.2 데이터베이스 최적화

대규모 대화 로그 처리를 위해 읽기와 쓰기 경로를 분리하고, 분석용 데이터는 BigQuery에서 다음과 같은 최적화 전략을 적용하여 비용과 성능을 관리합니다:

- **파티셔닝 (Partitioning)**: 데이터 생성 시간(`createdAt`) 기준 월별(MONTH) 파티셔닝을 적용하여 쿼리 스캔 범위를 최소화합니다.
- **클러스터링 (Clustering)**: 자주 조회되는 필드를 기준으로 데이터를 정렬 저장하여 검색 속도를 극대화합니다.

5 결론

Cloudturing의 기술 아키텍처는 최신 AI 기술의 혁신성과 엔터프라이즈급 인프라의 안정성을 동시에 달성하도록 설계되었습니다. Gemini AI를 통한 지능형 처리와 GKE 기반의 탄탄한 인프라는 기업이 안심하고 AI를 비즈니스에 도입할 수 있는 강력한 기반을 제공합니다.

6 부록

6.1 용어 정의

용어	설명
LLM	Large Language Model, 대규모 언어 모델
RAG	Retrieval-Augmented Generation, 검색 증강 생성
MSA	Microservices Architecture, 마이크로서비스 아키텍처
GKE	Google Kubernetes Engine, 구글의 관리형 쿠버네티스
HPA	Horizontal Pod Autoscaling, 수평적 파드 자동 확장
CDN	Content Delivery Network, 콘텐츠 전송 네트워크
CQRS	Command and Query Responsibility Segregation, 명령과 조회의 책임 분리

6.2 문서 이력

버전	날짜	변경 내용
1.0	2025년 12월 16일	최초 작성

Table 3: 문서 이력

Cloudturing
본 문서에 대한 문의: cto@cloudturing.com