클라우드 네이티브 기반 재사용 데이터 파이프 라인 구축

조민준° 송인용* 김종원*

광주과학기술원 전기전자컴퓨터공학부°

광주과학기술원 AI 대학원 NetAI 연구실*

A Cloud-Native based Reusability Data Pipeline

Min-Jun Cho° In-Yong Song* Jong-Won Kim*

Dept. of Electrical Engineering and Computer Science°

Dept. of Ai Graduate School NetAI Laboratory*

GIST (Gwangju Institute of Science & Technology)

minjun_jo@gm.gist.ac.kr , {inyong, jongwon}@smartx.kr

요 약

본 논문은 복수의 AI 모델을 서비스하는 과정에서의 인프라적 효율을 위해 재사용성을 강조한 파이프라인 구조를 소개하고, 재사용성을 살려 복수의 AI 모델을 효율적으로 운영하기 위한 초기 Cloud Native 환경 구현체의 구현 방식과 결과를 제시한다.

1. 서론

새로운 인공지능 모델과 개념들이 매일 등장함에 따라, 사용자들은 다양한 AI 모델 중에서 자신의 요구에 부합하는 모델을 선택할 수 있는 시대에 살고 있다. 이로 인해 복수의 AI 모델을 사용하는 상황이 자주 발생하게 되었고 이러한 상황은 AI 모델 선택을 지속적인 과제로 만들었다. 특히 사용하는 AI 모델이 증가할 수록 맞춤형 데이터 파이프라인을 개발해야 하는 필요성이 증가할 수 밖에 없으며, 이는 개발 속도를 저해하는 중요한 요인으로 작용하게 된다. 따라서 본 논문에서는 Cloud Native 환경에서 그러한 파이프라인의 유연한 재활용 가능성을 보이고자 한다.

Cloud Native 란 Container, Microservice, Immutable Infrastructure, Declarative APIs 접근 방식으로 조직이 공개적 혹은 비공개적 또는 하이브리드 클라우드 상황에서 현대적이고 동적인 환경에서 확장가능한 애플리케이션을 개발하고 실행할 수 있도록 하는 것으로 Cloud Native 기술을 활용하여 하면 엔지니어는 영향이 큰 반경을 최소한의 노력으로 자주, 예측 가능하게 수행할 수 있다.[1] 쿠버네티스는 이러한 Cloud Native 환경을 구축할 수 있는 현대적 컨테이너 오케스트레이션 툴로 Reusability, Scalability, 그리고 Flexibility 을 만족시키는 것이 가능하다. MSA 구조[2]와 컨테이너 부하 증가에 대한 autosacler[3]는 기존 infrastructure 와 큰 차이점을 가진다. 이러한 기술들은 최종 형태의 구현에서 완전히 적용될 예정이며 본 논문에서 다룰 내용은 현재 전부 구현되어 있지 않지만 후에 적용해야 할 기술로 재사용성과, 확장성, 유연성을 단계적으로 탑재할 예정이다.

2. 구현 현황 및 활용

Kubernetes 환경에서 재사용성을 높인 파이프라인 설계를 위해 MinIO 를 데이터 레이크 포지션에 위치시키고 NATS 를 이용하여 실시간성을 보장하기 위해 다음과 같은 인프라를 설계하였다.

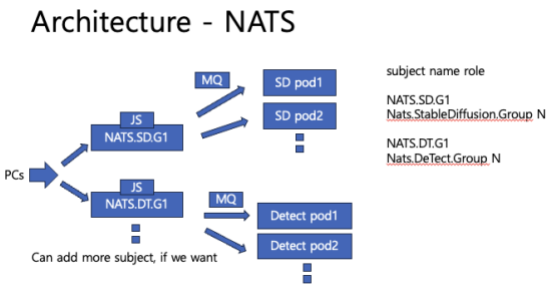

Figure 1 NATS 구조

NATS 를 통해 데이터의 저장위치와 명을 subject 로 분류하여 각 서비스는 격리된 환경에서 원하는 데이터 정보를 실시간으로 받을 수 있다. 이렇게 받은 데이터 정보는 NATS 와 병렬적으로 저장되어 Local 환경에서 MinIO 로부터 데이터를 받아 이용하게 된다. 다음은 MinIO 와 NATS 를 모두 나타낸 설계이다.

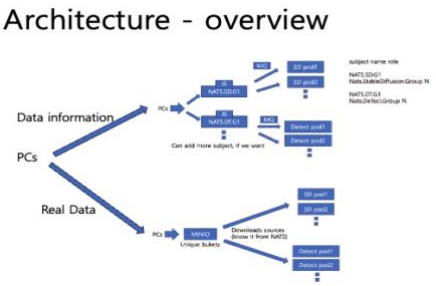

Figure 2 Overview Architecture 1

Figure 2 를 기반으로 하면 실시간 데이터가 필요한 AI 모델은 NATS 로부터 가장 최신의 데이터를 사용할 수 있고, 훈련이 필요한 AI 모델의 경우에는 그간 MinIO 를 통해 분류된 데이터를 즉각적으로 사용할 수 있다.

이렇게 구축된 파이프라인 기반하여 생성형 AI Stable Diffusion 을 활용하여 간단한 대상의 특징을 반영한 가상 캐릭터 생성 서비스 구조를 다음과 같이 설계 및 구축하였다. NetAI 의 MobileX Station – 클라우드 네이티브 기반 움직일 수 있는 PC 형태의 교육 플랫폼이다. - 10 대의 환경에서 각 스테이션들은 웹 캠의 데이터를 스테이션 번호에 따라 unique 한 subject 및 MinIO 버킷으로 발행한다. 서비스의 API 서버는 2 개 이상의 복수의 포드로 이루어져 Round-Robin 방식으로 요청을 받아 끝 단의 Stable Diffusion 포드와 상호작용을 통해 변환된 데이터를 최종적으로 사용자에게 redirecting 한다.

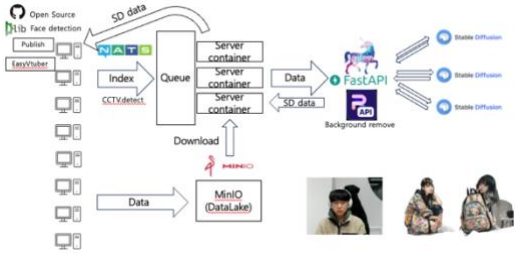

Figure 3 재사용 파이프라인을 활용한 Stable Diffusion 서비스

이 때 API 서버 혹은 최초의 edge device 에서 Stable Diffusion 에 적합한 사이즈로 이미지의 사이즈를 수정, 이미지(사용자의 흉상 모습)의 특정 구도만 필터링하여 개발자가 원하는 데이터로 가공하는 작업을 가지게 된다. 이 작업과 동시에 raw data 는 그대로 MinIO 버킷에 저장되고 수정된 데이터 역시 또 다른 버킷에 저장하는 작업을 통해 추후 어떠한 AI 모델이 사용할 수 있는 가공한 데이터를 저장함으로써 파이프라인의 재사용성을 확보할 수 있음을 확인할 수 있다.

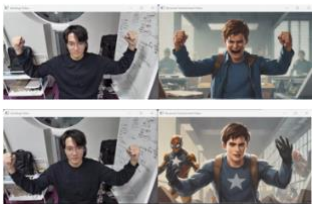

Figure 4 Hero 를 프롬포트에 넣었을 때의 실행결과

Figure 4 와 같이 기존 모습을 기반으로 그림을 생성하고, Sable Diffusion 에 기입한 프롬포트에 따라 다양한 작풍의 그림을 생성한다.

3. 결론

본 논문에서 제시한 설계를 기반으로 하여 프로토타입의 시스템을 구축하였고, 해당 시스템이 실제로 재사용성을 확보할 수 있는지 검증하기 위하여 Stable Diffusion AI 모델을 활용하여 대상의 특징을 반영한 간단한 가상 캐릭터 생성 서비스를 제작하였다.

이를 통해서 10 대의 스테이션으로부터 모이는 정보와 AI 모델이 사용하기 위해 가공한 데이터도 함께 저장하는 과정을 관찰하면서 충분히 다른 작업에서도 사용이 가능함을 관찰할 수 있었다.

해당 파이프라인을 기반으로 하여 스테이션이 각각의 움직이는 관찰자가 되어 고정된 CCTV 가 보지 못하는 사각지대를 극복함으로써 완전 관제 환경을 1:N 파이프라인 구축을 통해 효율적인 인프라의 활용이 가능할 것으로 기대한다.

4. Acknowledgment

본 논문은 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-01176, 클라우드 기반 융합형 자율주행 지능학습 데이터 생성/제공을 위한 데이터 수집가공 핵심 기술 개발; No. 2019-0-01842, 인공지능대학원지원(광주과학기술원))

5. 참고 문헌

[1] Cloud Native Computing Foundation(CNCF), about, who we are, Cloud Native Definition, (https://www.cncf.io/about/who-we-are/), Mar. 2024.

[2] Kubernetes, Documentation, Concepts, Overview, Objects in Kubernetes, (https://kubernetes.io/docs/concepts/overview/), Mar. 2024.

[3] Github, Kubernetes, autoscaler, (https://github.com/kubernetes/autoscaler/tree/master), Mar. 2024.