엣지에 최적화된 학습 및 추론 환경 구성/관리 기술 발표

(GS-AI)

2020.12.10

GEdge Platform 코어 개발자
조정현(junghyuncho@sk.com)

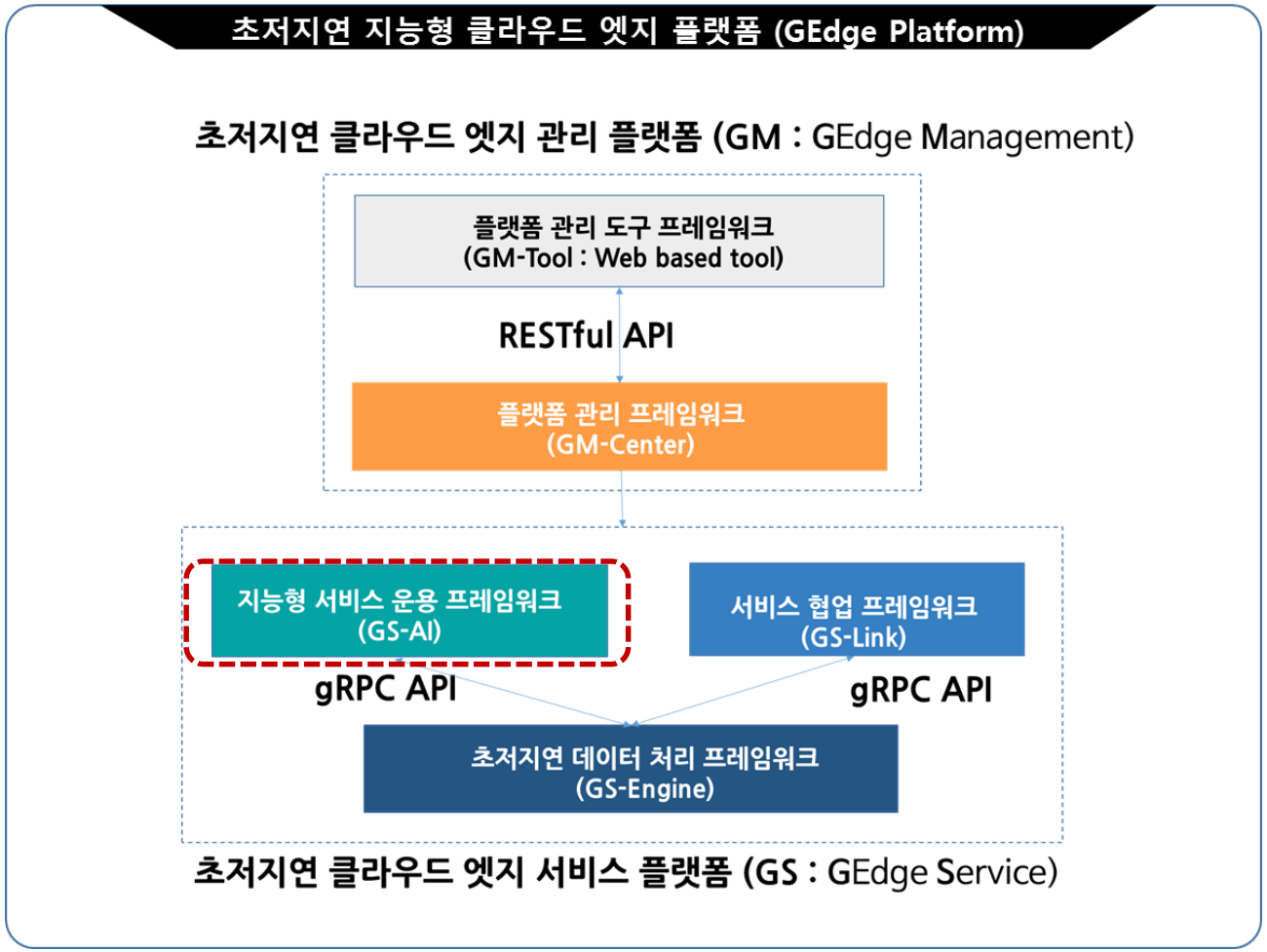
"The First talk of Edge Computing with Clouds"

- GEdge Platform 커뮤니티 멤버들의 첫번째 이야기 -

GEdge Platform Community 1st Conference

Contents

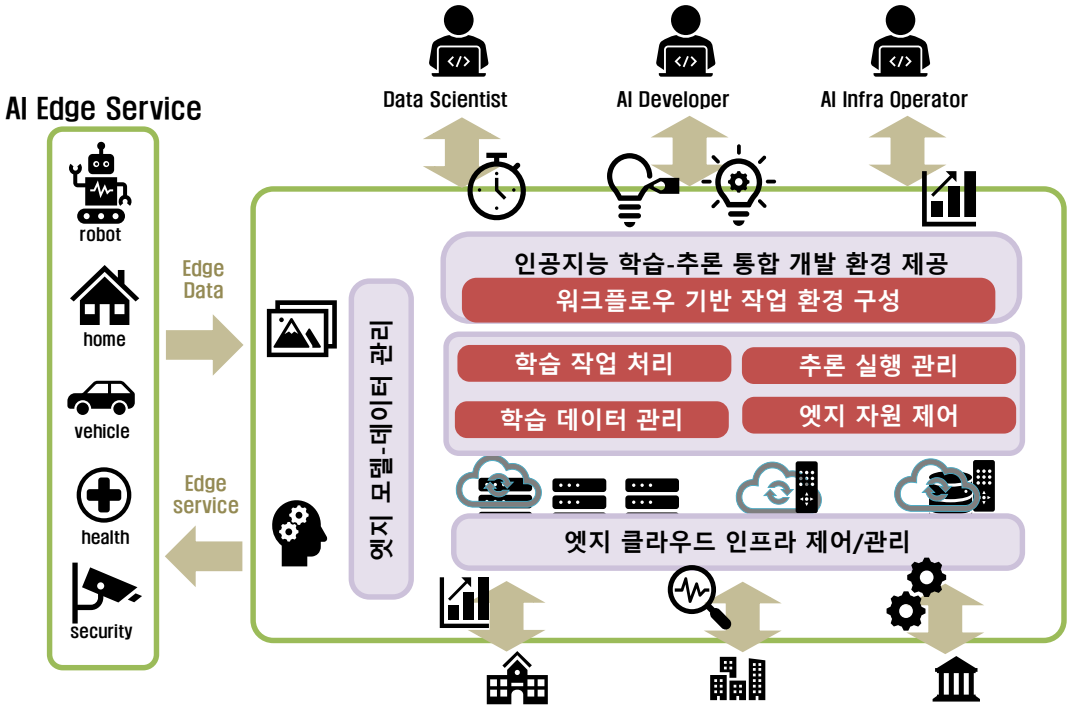
이번 세션은 …


초저지연 지능형 클라우드 엣지 플랫폼 (GEdge Platform)
초저지연 클라우드 엣지 관리 플랫폼 (GM : GEdge Management)
플랫폼 관리 도구 프레임워크
(GM-Tool : Web based tool)
RESTful API
플랫폼 관리 프레임워크
(GM-Center)
지능형 서비스 운용 프레임워크
(GS-AI)
서비스 협업 프레임워크
(GS-Link)
gRPC API
gRPC API
초저지연 데이터 처리 프레임워크
(GS-Engine)
초저지연 클라우드 엣지 서비스 플랫폼 (GS : GEdge Service)

1 기술 개요

엣지에 최적화된 학습 및 추론 환경 구성/관리 기술

- 클라우드 기반의 엣지 컴퓨팅 플랫폼에서
- 인공지능 응용을 개발하고 실행함에 있어
- 작업 특징을 분석하고 플랫폼의 특징을 반영하여 최적화된 작업 환경을 구성/제공하는 기술



[인공지능 개발자]
- 다양한 연구/개발에 대한 용이성 지원
- 시스템 구성에 대한 부담 해소

[클라우드 엣지]
- 데이터 처리, 및 전송에 대한 지연 극복
- 작업 요청에 대해 자원 성능/효율 지원

[디바이스]
- 엣지 단말 디바이스의 다양화 지원
- 계산/가속 디바이스의 다양화 지원

1 기술 개요

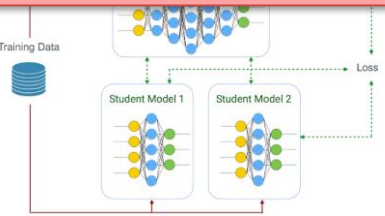
Motivation

- 저지연 처리/전송 + 대규모 데이터 처리 가능
- 인공지능 학습 형태의 다양화

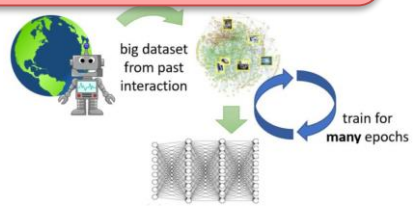
복잡도 향상

학습 형태별 맞춤형 시스템 구성 지원
엣지 플랫폼의 지능적 변화(성능/효율성) 지원

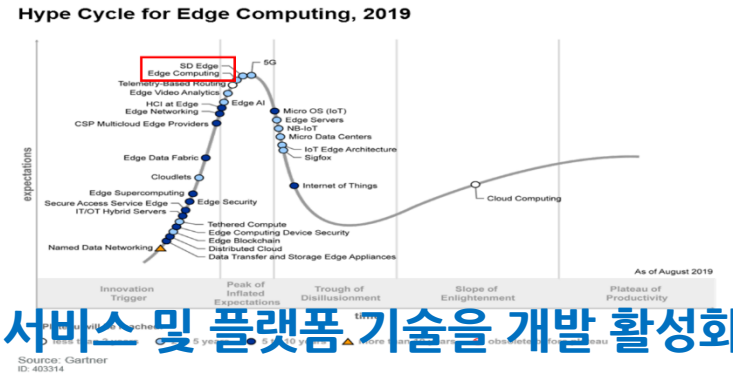
[Federated Learning]



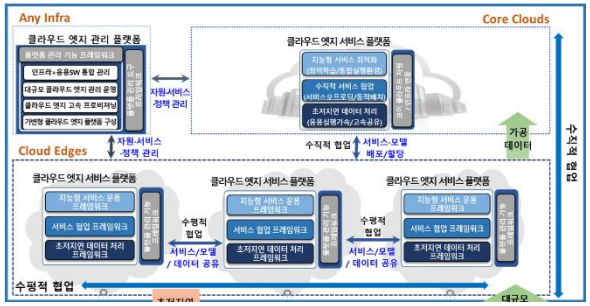
[Transfer Learning]



[Reinforce learning]



엣지 서비스 및 플랫폼 기술을 개발 활성화



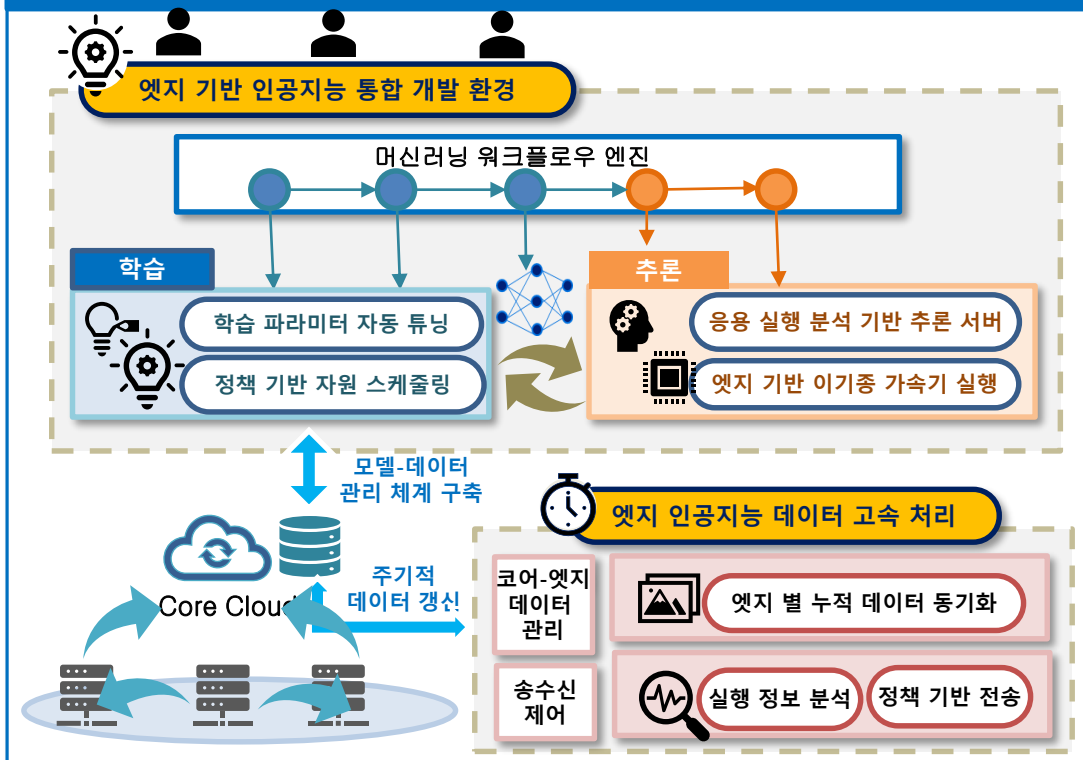
엣지 실행 응용 다양화 및 지능화

1 기술 개요

Goal

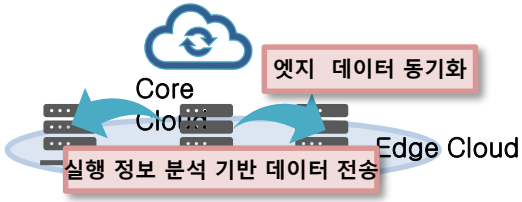
- 워크플로우 기반 인공지능 개발 환경 핵심 기술 개발 → 워크플로우 기반 학습/추론 통합 지원
- 엣지 기반 인공지능 데이터 고속 지원 기술 개발 → 기계학습 학습 타입 단계별 확장 지원

핵심 개발 내용



개선 기술 특장점

- 엔터프라이즈 급 인공지능 통합 개발 환경
- 학습 처리 속도 향상
- 용이한 재학습 환경으로 정확도 향상 기여
- 엣지 기반 이기종 가속기 실행 기능 제공

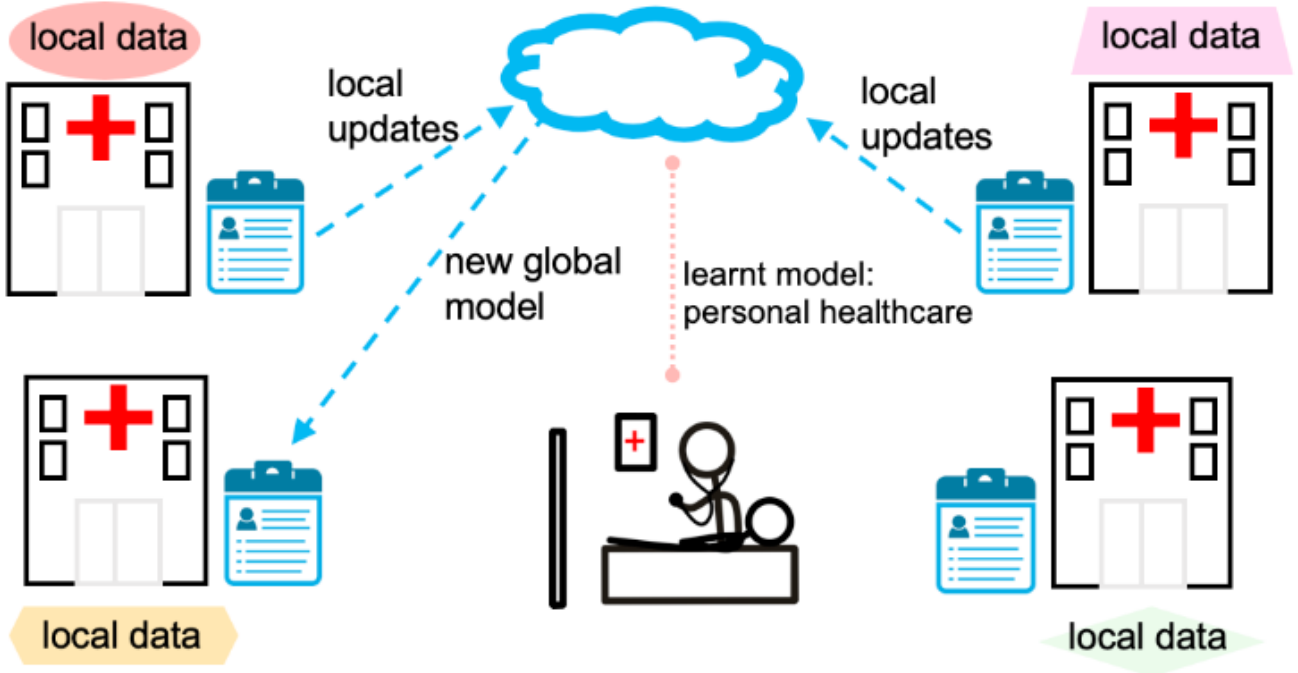


- 모델-데이터 저장 체계 구축으로 확장성 지원
- 학습 기법 별 데이터 전송 정책 관리
- 실행 정보 분석 기반 맞춤형 데이터 고속 전송
- 엣지 학습 누적 데이터 동기화 관리

1 기술 개요

1차년도 기계학습 지원 상세 : Federated Learning

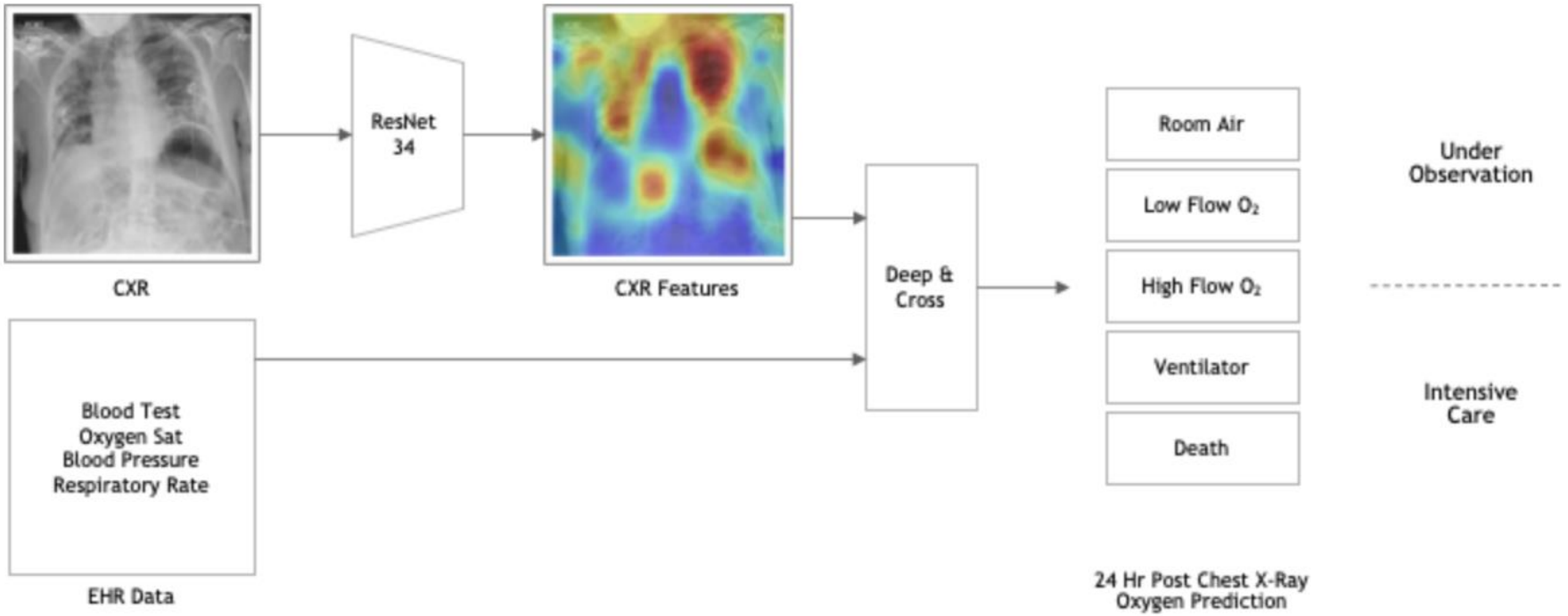
- 데이터를 직접 공유하지 않고 분산된 환경에서 사용자의 데이터로 학습하되 연합하여 중앙의 모델을 강화 시키는 방식
- 데이터 프라이버시/보안 이슈 해결 → 통신 이슈 & 로컬/글로벌 모델 관리 필요



1 기술 개요

1차년도 기계학습 지원 상세 : Federated Learning

관련 연구 : Clara FL : Federated Learning powered by NVIDIA Clara



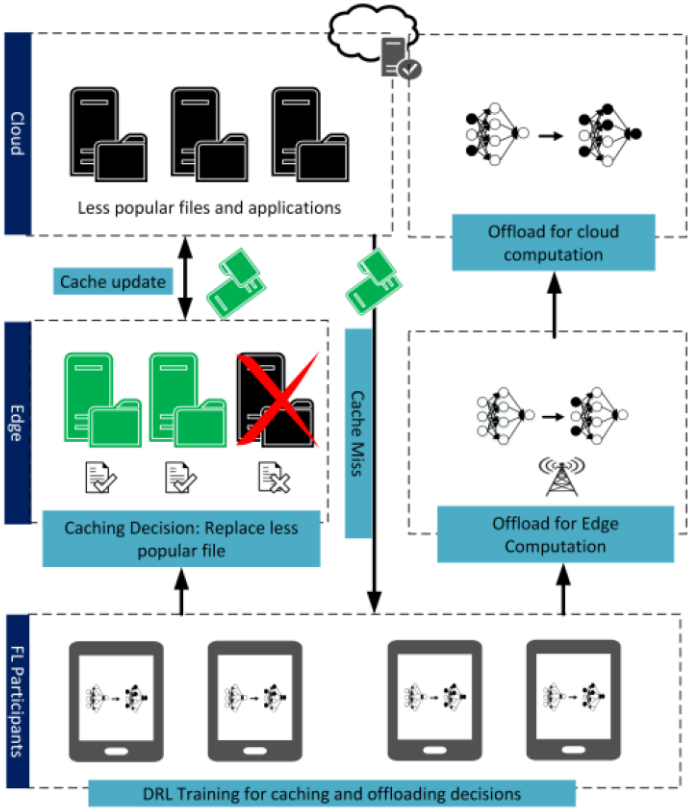
엔비디아 Clara로 코로나19 환자 산소 요구량 예측 AI 모델 구축

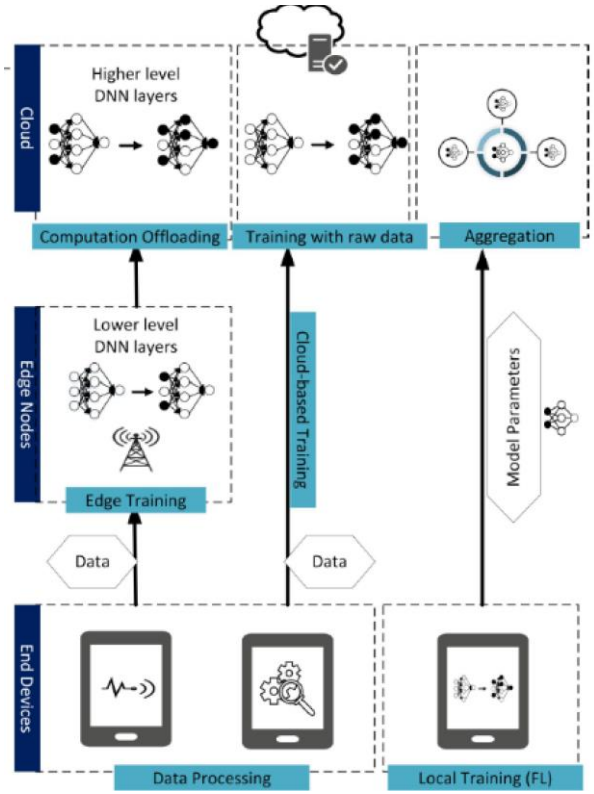
1차년도 기계학습 지원 상세 : Federated Learning

관련 연구 : Clara FL : 의료 영상 분석과 유전체학 프로파일링을 가속화하기 도구

1차년도 기계학습 지원 상세 : Federated Learning

- Federated Learning 학습/모델 데이터 전송 - 동기화
- Federated Learning 실행을 위한 최적 자원 할당 - 실행 제어 관리)



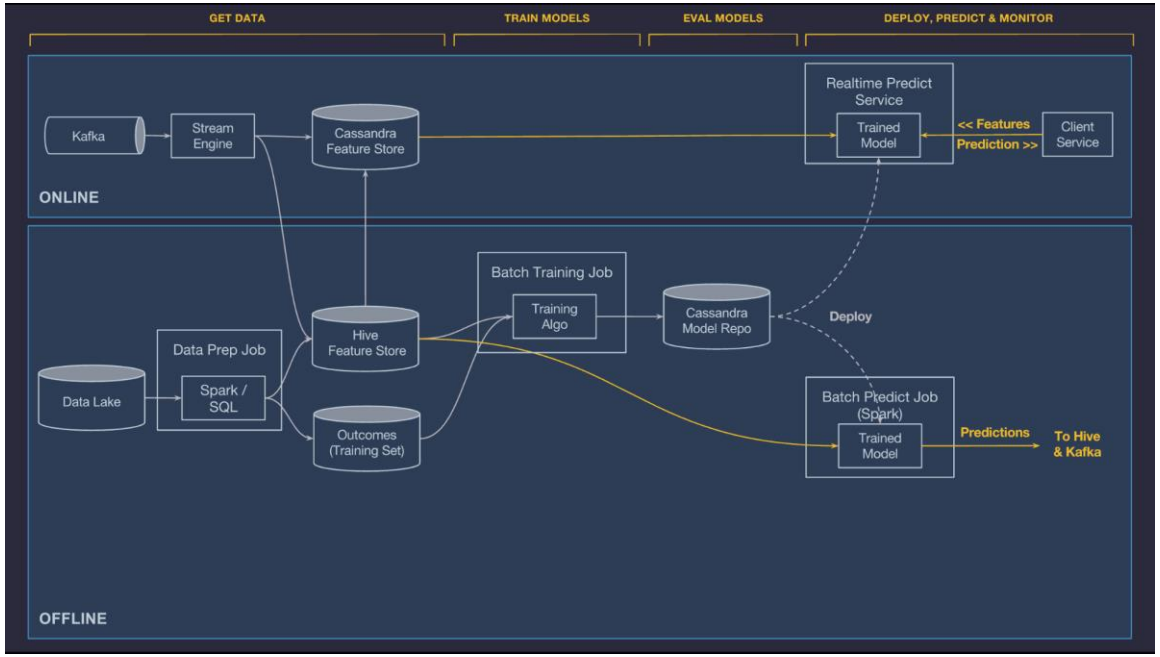


. Yang, et. al., “Federated Learning in Mobile Edge Networks: A Comprehensive Survey,” arXiv:1909.11875v1, Sept. 2019

1 기술 개요

1차년도 기계학습 지원 상세 : 워크플로우 기반 기계학습 실행

엔드-투-엔드 머신러닝 플랫폼 다양화 & 고도화



Michelangelo: Uber's Machine Learning Platform

우버 사내 엔지니어 및 데이터 엔지니어가 대규모의 기계학습 솔루션을 쉽게 구축(build)하고 배포(deploy)할 수 있는 플랫폼

전체 아키텍쳐와 EATS 플랫폼 구축에 오픈소스와 자체개발 컴포넌트를 조합하여 구축

빅데이터 기반의 학습 기반 대상으로 머신러닝 학습 지원 형식 한계
클라우드 엣지 플랫폼으로의 확장성 결여

2 기술 상세

핵심 기술 1)
클라우드 엣지 환경에서의 기계학습 작업 최적 자원 할당을 위한 정책 관리

[ML Global Policy] : 클라우드 엣지 환경에서의 기계학습 작업 최적 자원 할당

① [Training Type Classifier] 정적 학습 타입 분류를 통한 가중치 적용 - 정의된 학습 타입 분류기를 통해 타입을 분류하고, 학습 타입 별, 모델 이동과 데이터 이동에 대한 Overhead 계산 적용

② [Place Schema] 학습 실행 동안의 자원 사용량을 프로파일링하여 유사 작업 요청시, 프로파일링 결과와 가용 자원의 실시간 모니터링 정보를 취합하여 예측 배포 정책 적용

[Type-specific Local Policy]

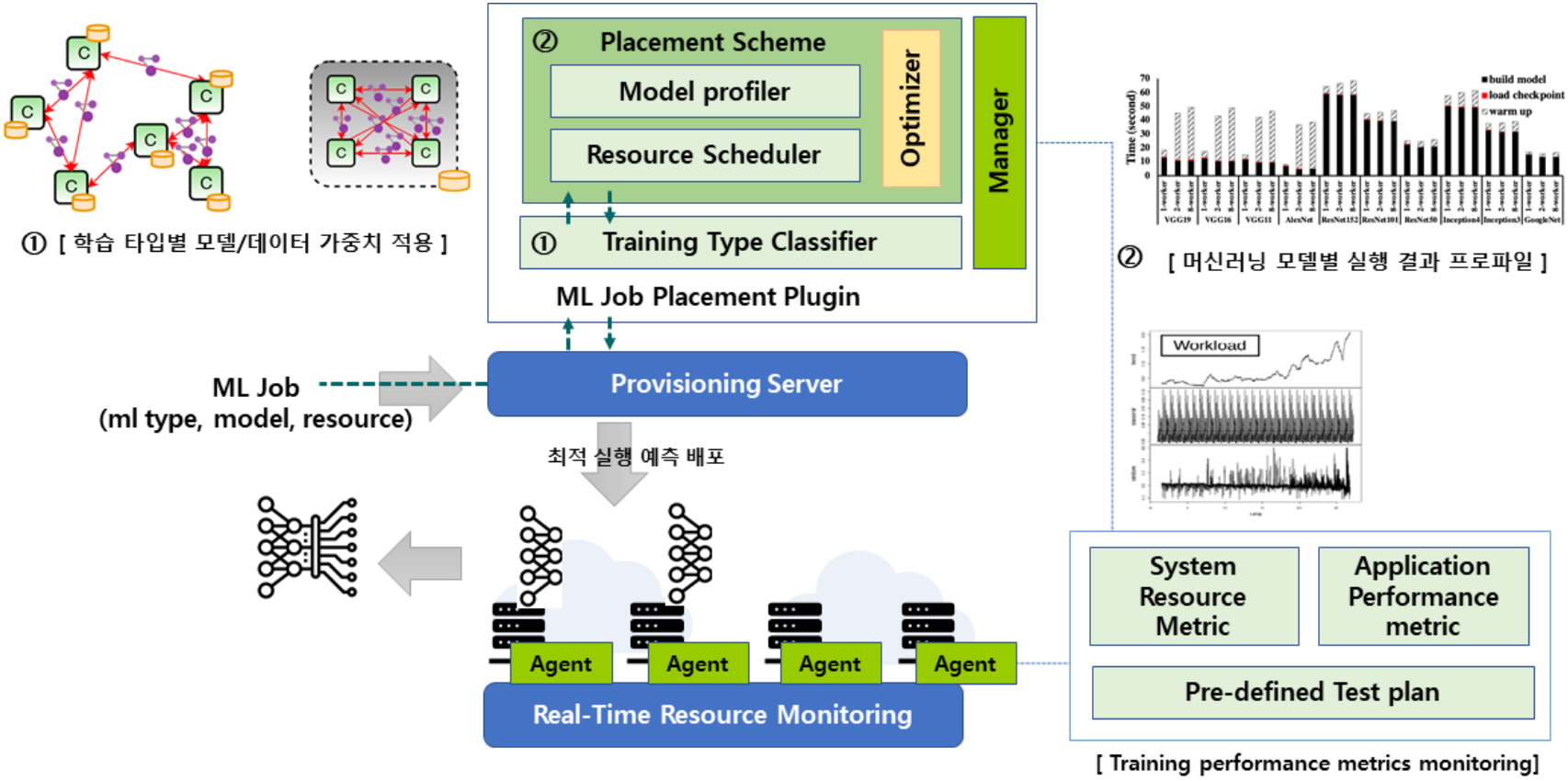
. Training Type Classifier을 통해 분류된 타입에 따라, 학습실행 특징을 반영한 계산식 도출. 연차별 학습 타입 확장 적용

Type	Type-specific Local Policy
Federated Learning (2-level optimization)	① (가용 컴퓨팅 노드의 유휴 자원 값) x (컴퓨팅 노드간 통신 값) 연산 ▶ 연합학습의 작업 지연 지표인 컴퓨팅 지연과 송수신 지연 개선 ② 모델 업데이트를 통한 협업학습 특징 고려 컴퓨팅 노드상에서의 작업 완료 시간 예측 적용. (인풋 사이즈 x 프로파일링) 연산 ▶ 분산된 환경에서의 전체 학습 완료 시간 최소화

2 기술 상세

핵심 기술 1)

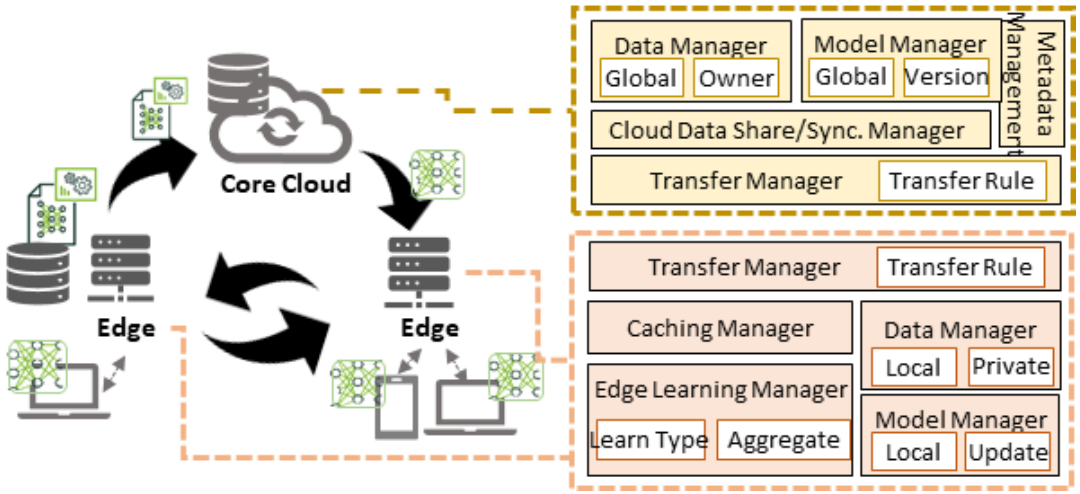
클라우드 엣지 환경에서의 기계학습 작업 최적 자원 할당을 위한 정책 관리



2 기술 상세

핵심 기술 2)

클라우드간 학습 데이터 고속 지원을 위한 데이터 분할 관리 체계 구축 및 교환 기술



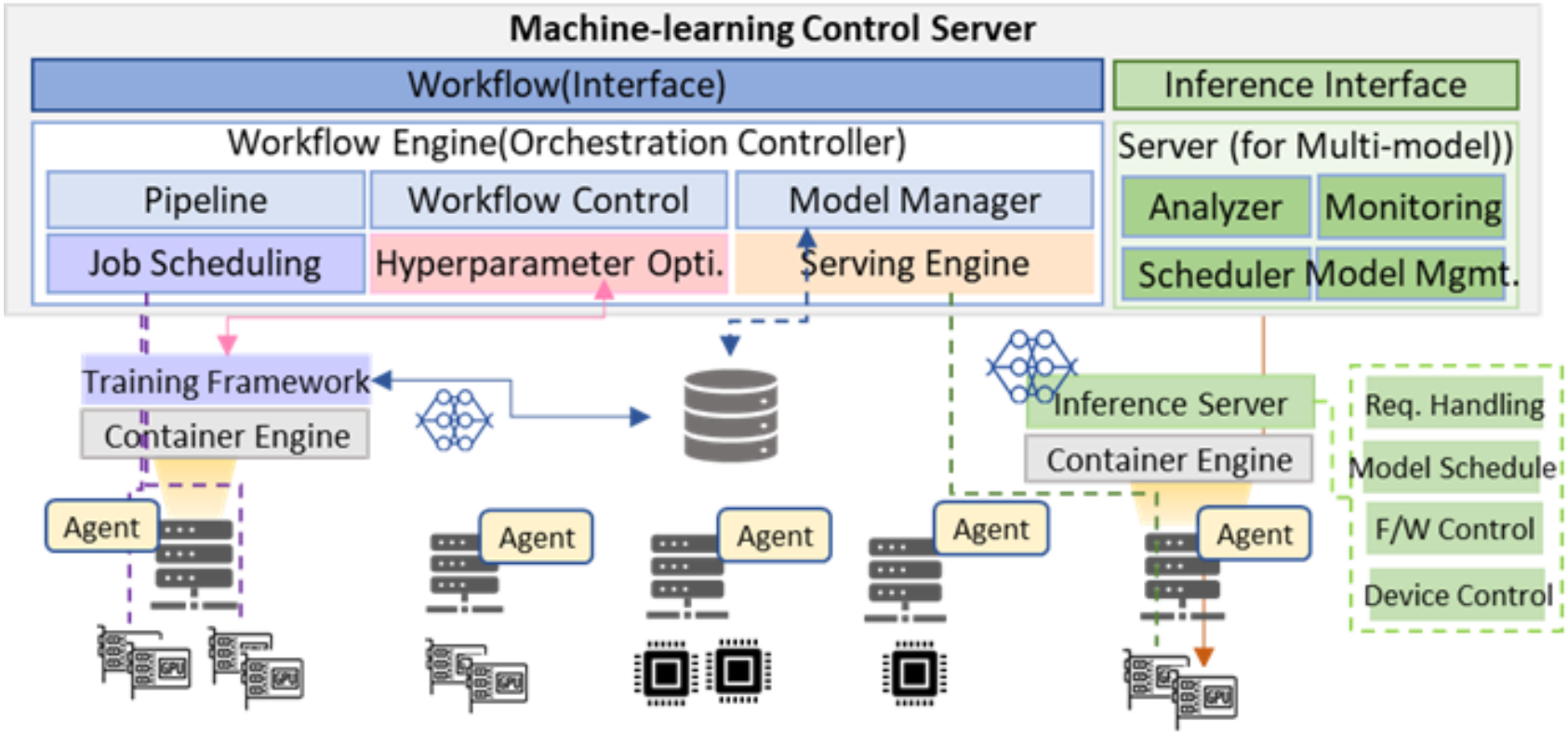
Location	Role
Core Cloud	• 학습 데이터-모델 관리 저장 • 글로벌 학습 데이터/모델 • 자원 상태 및 학습 적응형 전송 • 정책 기반 데이터 동기화
Edge Cloud	• 학습 방식 맞춤형 전송 정의 • 엣지 누적 로컬 데이터 관리 • 데이터 캐싱 적용 • 룰 기반의 클라우드 간 전송

단계별 학습 타입 확장 지원	처리 플로우 정의
1단계 : FederatedLearning	① ML Training Requirement : (Topolopy, compute plan) 입력 분석 ② 초기 모델(Core Cloud): ① 결과 x 프로비저닝 서버(학습 예측 배포 정책기)로 도출된 서버로 모델 배포(Edge Cloud) ③ (Training Data) Topology 기반의 업데이트 모델 전송. 모델 업데이트에 따라 전체 학습 지연에 영향 : 모델 업데이트 처리 이벤트 핸들링(전송상태에 따른 재시도 처리) ④ (Training Complete) 최종 학습 모델 저장(Core Cloud-Storage)

2 기술 상세

핵심 기술 3)

최적 학습 및 고속 추론을 위한 실행환경 제공 기술



핵심 기술 3)

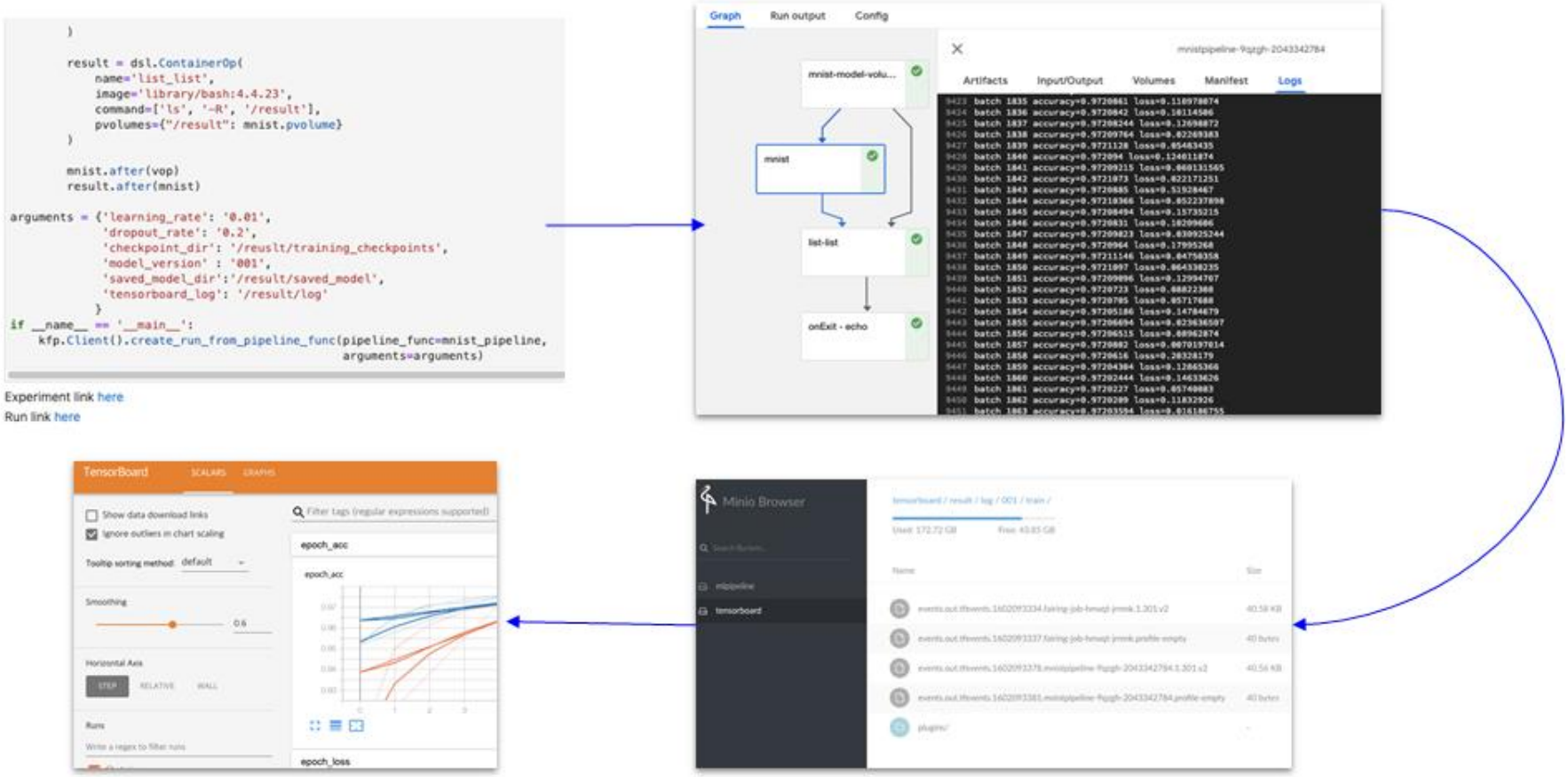
최적 학습 및 고속 추론을 위한 실행환경 제공 기술 - 관련 오픈소스

Kubeflow : 머신 러닝을 위한 클라우드 네이티브(Cloud Native) 플랫폼

컴포넌트 명	주요 기능
Pipeline	서버컨테이너 기반의 end-to-end ML 워크플로우를 만들고 배포할 수 있는 컴포넌트
Notebook Server	쿠버네티스 위에서 실행되는 쥬피터 노트북 서버
Katib	하이퍼파라미터 튜닝과 뉴럴아키텍처탐색을 수행하는 컴포넌트
Fairing	쿠베플로우가 설치된 환경에서 손쉽게 ML 모델을 학습/배포 할 수 있는 파이선 패키지
KFServing	InferenceService라는 커스텀 리소스를 가지고 인퍼런스 서버 제공

핵심 기술 3)

최적 학습 및 고속 추론을 위한 실행환경 제공 기술



JupyterNotebook 실행 및 Workflow pipeline 연동 실행 결과

2 기술 상세

핵심 기술 3)

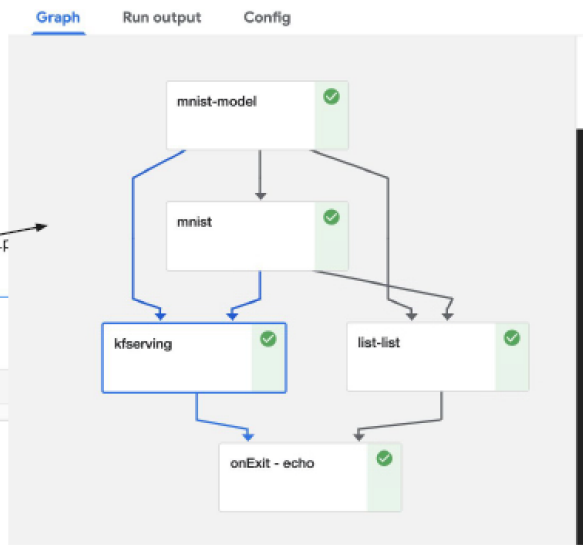
최적 학습 및 고속 추론을 위한 실행환경 제공 기술

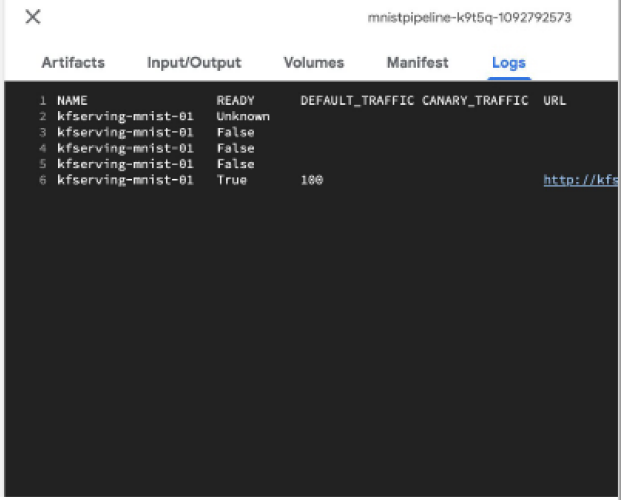
```
arguments = {'volume_size' : '5Gi',
             'learning_rate': '0.01',
             'dropout_rate': '0.2',
             'checkpoint_dir': '/result/training_checkpoints',
             'saved_model_dir': '/result/saved_model/0001',
             'tensorboard_log': '/result/log',
             'namespace': 'kubeflow',
             'storage_uri': '/saved_model',
             'name': 'kfserving-mnist-01'
             }

if __name__ == '__main__':
    kfp.Client().create_run_from_pipeline_func(pipeline_func=mnist_pip
                                               arguments=arguments)
```

Experiment link here

Run link here

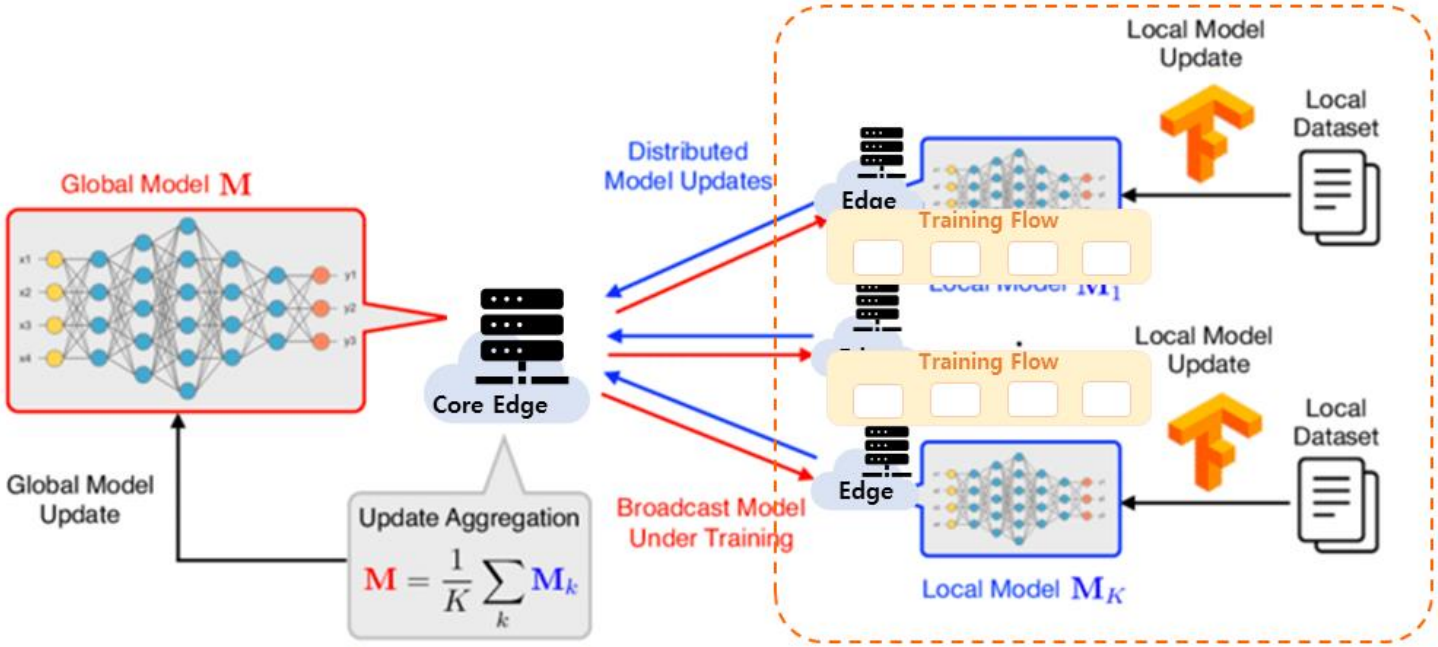




배포-실행 추론 플로우 구축 및 실행 결과

핵심 기술 3)

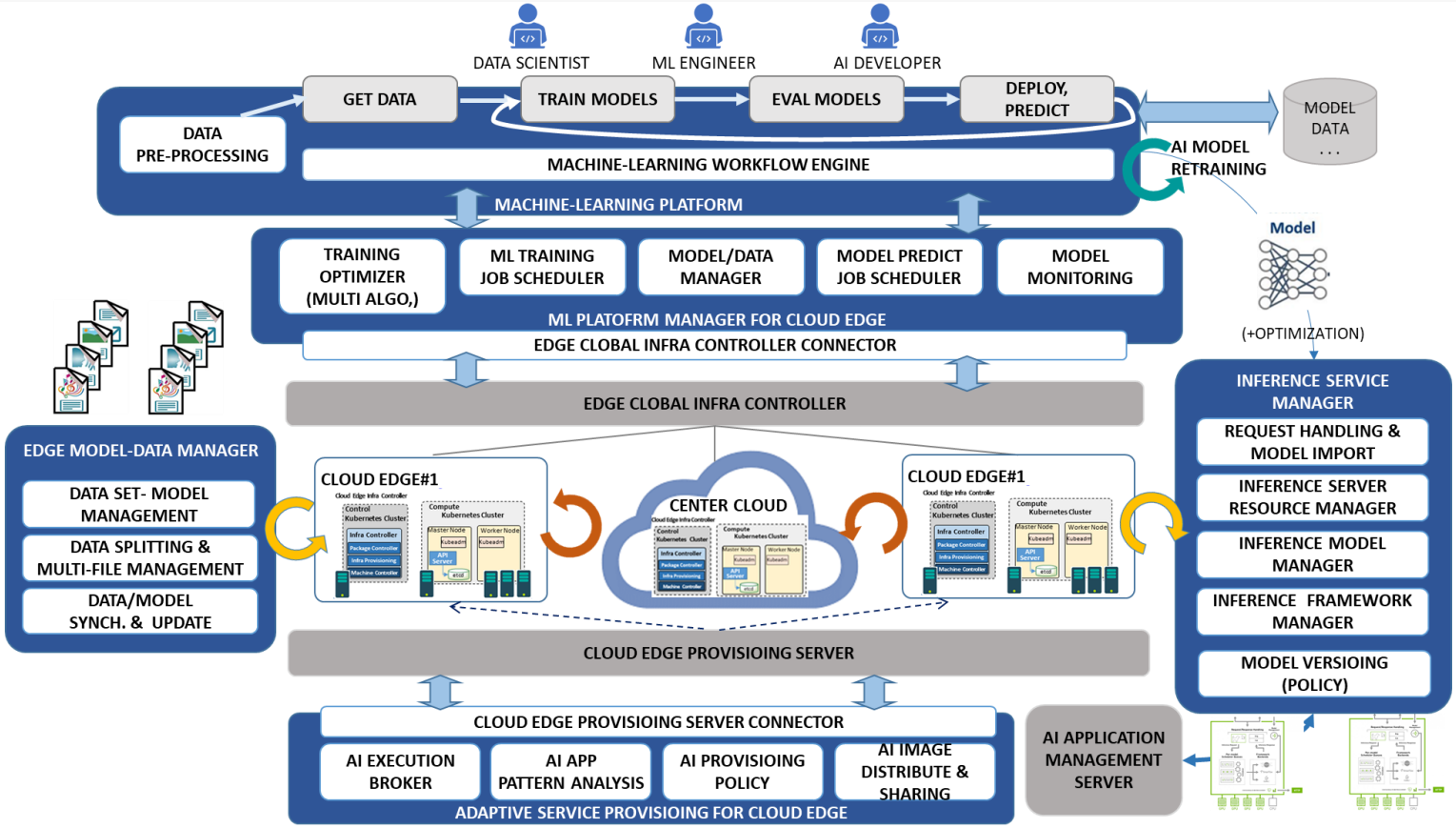
최적 학습 및 고속 추론을 위한 실행환경 제공 기술



기계학습 플로우 기반의 연합학습 실행 관리 시스템 제안

3 향후 계획

- 기계학습 모델 확장 지원
  - 1차년도 연합학습에 이어, 전이학습, 강화학습 등 학습 타입 다양화 지원
- 오픈소스 활성화 기여
  - gedge-platform 오픈소스 저장소를 통한 오픈소스 활성화 및 고도화



감사합니다.

http://gedge-platform.github.io

GEdge Platform 코어 개발자
조정현(junghyuncho@sk.com)