Samsung Cloud Platform

AI/ML

AI&MLOps Platform

Kubernetes 기반의 머신 러닝 플랫폼

AI&MLOps Platform은 머신 러닝 모델의 개발, 학습, 배포 과정 전체 파이프라인의 반복적인 작업을 자동화하는 머신 러닝 플랫폼입니다. Kubernetes 기반의 AI/MLOps[1] 환경을 제공하며, 학습 데이터와 모델, 운영 데이터의 통합적인 관리가 가능합니다.

Cloud Native MLOps 환경 제공

AI&MLOps Platform은 클라우드에 최적화 된 머신러닝 모델 개발 환경을 제공하며, Kubernetes 기반으로 다양한 오픈소스와의 연계가 편리합니다.

머신 러닝 개발 및 운영 편의성

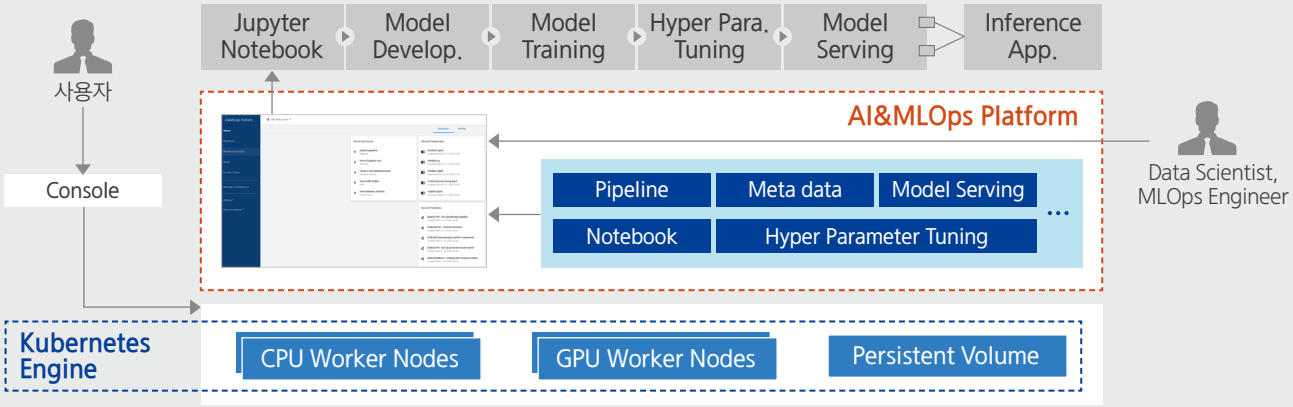
TensorFlow, PyTorch, scikit-learn, Keras 등 다양한 머신러닝 프레임워크를 지원하는 표준화된 환경을 제공합니다. 머신러닝 모델의 개발, 학습, 배포 과정의 전체 Pipeline을 자동화하여 제공함으로써 모델 구성 및 생성이 쉽고 재사용이 용이합니다.

Add-on Feature 지속 추가

분산학습 Job 실행 및 모니터링, 추론서비스 관리 및 분석, Job Queue 관리 등 MLOps 환경 구성을 위한 다양한 기능을 제공하며, 잡 스케줄러(FIFO, Bin-packing, Gang 기반), GPU Fraction, GPU 자원 모니터링 등 효율적인 GPU 자원 활용을 위한 다양한 Add-on Feature들을 추가로 제공합니다. 특히, BM 기반의 Multi Node GPU 및 GPU Direct RDMA(Remote Direct Memory Access)를 통해 LLM(Large Language Model)과 자연어처리(NLP)의 Job 속도를 획기적으로 개선할 수 있습니다.

[1] MLOps : Machine Learning Development(Dev)와 Machine Learning System Operation(Ops) 통합을 목표로 하는 ML 엔지니어링 방법론

서비스 구성도



주요 기능

- **기본 기능**
  - AI 플랫폼 생성(자동배포/구성) , 조회(플랫폼 버전, 자원 현황), 삭제
  - Jupyter Notebook 제공 : 모델개발, 학습, 추론
  - 머신 러닝 Pipeline Workflow 자동화
- **추가 기능(AI&MLOps Platform에서 가능)**
  - Advanced AI/ML 플랫폼 대시보드
  - AI/ML Notebook Server : Base 이미지, 사용자 정의 이미지
  - AI/ML Job : Job 생성, 템플릿, 아카이브, 스케줄링, 실행, 모니터링
    ※ GPU 자원 모니터링, GPU Fraction 지원
    ※ Large Language Model 학습(DeepSpeed) 지원을 위한 Job Operator 제공
  - 사용자 이미지 빌드 및 관리
  - AI JumpStarter 및 실험 추적 관리(ETM : Experiment Tracking Management)
  - Serving : 대시보드, 모델 등록/관리, Inferencing, Predictions 시각화
  - 플랫폼 자원 관리 : 프로젝트별 자원 사용량 관리, 자원 사용량 모니터링
  - 프로젝트 사용자/권한 관리, Admin 기능, 플랫폼 Configuration 조정 기능

요금 기준

- **제공 항목**
  - AI&MLOps 환경 제공을 위한 SW 패키징
- **과금**
  - 배포된 AI&MLOps Platform 규모와 사용 시간에 대해 시간 단위 과금
    ※ 사용자 환경 구성을 위한 Samsung Cloud Platform 비용 별도

※ 본 상품은 오픈소스 Machine Learning Tool인 Kubeflow를 활용합니다.

FOR MORE INFORMATION

SAMSUNG SDS

www.samsungsds.com / cloud.samsungsds.com
contact.sds@samsung.com / scp_sales@samsung.com
youtube.com/samsungsds