Tutorial

# AI&MLOps Platform에서 Model Serve 생성하기

## (Creating a New model Serving Instance in AI&MLOps Platform)

October 2024



SAMSUNG SDS

# Contents

# 1. Tutorial Learning Objectives

튜토리얼의 목표는 삼성 클라우드 플랫폼에서 AI&MLOps Platform을 사용하여 Model Server를 생성하고 추론기능을 검증하는 것입니다. 튜토리얼의 각 모듈은 AI&MLOps Platform의 기능에 대한 간결한 접근 방법을 제공합니다.

# 2. Prerequisites

튜토리얼 진행에 필요한 사전요건은 아래와 같습니다.

- AI&MLOps Platform에 대한 이해
- Kubernetes와 KServe에 대한 이해
- YAML 파일 형식 및 사용법에 대한 이해
- Iris 데이터셋에 대한 이해
- “AI&MLOps Platform에서 Jupyter 노트북 생성하기” 튜토리얼 참조

## 2.1 Consult the User Guide

✔ AI&MLOps Platform의 전체적인 이해를 위해 사용자 가이드를 참고합니다.

➜ 아래 단계를 따라 Samsung Cloud Platform의 Service Portal로 이동합니다

① https://cloud.samsungsds.com/serviceportal/index.html

② Service Portal 메인 페이지 우측 상단의 "Console ↗" 버튼을 클릭합니다.

③ SCP Console의 회원가입 또는 로그인을 진행합니다.

➜ AI&MLOps Platform 온라인 사용자 가이드로 이동합니다.

① Console 페이지의 우측 상단에 위치한 "서포트" 아이콘을 클릭합니다.

② "서포트" 아이콘을 클릭하면 드롭다운 메뉴가 펼쳐집니다.

③ 드롭다운 메뉴에서 "사용자 가이드"를 선택합니다.

④ 처음에는 "Samsung Cloud Platform 사용자 가이드" 메인 페이지로 연결됩니다.

⑤ "AI&MLOps Platform 사용자 가이드"는 메인 페이지 상단 메뉴에서 이동합니다.

ⓘ 튜토리얼은 "AI&MLOps Platform 사용자 가이드 (AMP 1.6)"을 참조합니다.

➜ 사용자 가이드에서 원하는 섹션으로 이동합니다.

① “Part VI. Deployment”로 이동합니다.

② “1. Model Servers” 또는 “2. Inference Templates”을 선택합니다.

### 2.2 Secure Necessary Permissions

✔ 튜토리얼 진행을 위한 AI&MLOps Platform 사용자 권한을 확인합니다.

✔ 필요시 AI&MLOps Platform 관리자를 통해 권한을 확보합니다.

## 3. Quick Start

AI&MLOps Platform에서 Model server 인스턴스를 생성하여 추론하는 과정을 학습합니다.

### 3.1 Preliminary Checks

➜ Model server 인스턴스 생성에 충분한 리소스를 확보합니다.

• AI&MLOps Platform 대시보드의 사이드바를 통해 "Project>Resource Summary"로 이동합니다.

• AI&MLOps Platform은 프로젝트 단위로 가용 리소스의 총량을 정의합니다. 프로젝트에 할당된 리소스가 부족하면, 프로젝트 내 리소스 할당을 재조정합니다.

• 기존 프로젝트의 리소스 부족 시 추가 프로젝트를 생성하여 리소스를 확보합니다.

## 3.2 Creating a New Model Server Instance

➜ AI&MLOps Platform 대시보드의 사이드바에서 "Deployment > Inference Templates"을 선택합니다.

➢ "Inference Templates"은 KServe 컴포넌트를 이용하여 Model Server 인스턴스를 생성하는 템플릿입니다. YAML 파일로 정의된 Model Server를 생성하고, 생성된 Model Server는 REST API를 통해 JSON 형식의 추론 데이터를 입력 받습니다.

➜ “Inference Templates” 화면 우측 상단의 “+ NEW TEMPLATE” 버튼을 클릭하면 “New Template” 팝업 페이지가 나타납니다.

➜ “New Template” 팝업 페이지에 아래와 같이 세부 사항을 입력합니다.

① Name: 템플릿 이름을 “MyFirst-Inference-Template”로 입력합니다.

② Type: 템플릿 유형을 “Inference”로 지정합니다.

③ Description (Optional): 생성목적을 “Iris-Inferencing”로 입력합니다.

④ 템플릿 등록 방법은 “Contents”를 선택합니다.

⑤ 아래 YAML 코드 블록을 복사하여 텍스트 상자에 붙여넣기를 합니다.

```
apiVersion: "serving.kserve.io/v1beta1"
kind: "InferenceService"
metadata:
  name: "sklearn-iris"
spec:
  predictor:
    model:
      modelFormat:
        name: sklearn

    storageUri: "gs://kfserving-examples/models/sklearn/1.0/model"
```

⑥ “Add”를 클릭합니다.

➜ AI&MLOps Platform 대시보드의 사이드바에서 “Deployment > Model Servers" 을 선택합니다.

"Model Servers" 화면 우측 상단의 "+ NEW MODEL SERVER" 버튼을 클릭하면 "NEW MODEL SERVER" 팝업 페이지가 나타납니다.

➜ Model Server 생성에 필요한 세부 사항을 아래와 같이 입력합니다.

① Model Server에 활용될 템플릿의 Type은 "My Templates"를 선택합니다.

② 템플릿 드롭다운 박스의 항목에서 "Iris-Inferencing"를 선택합니다.

③ "START"를 클릭합니다.

➜ 생성된 Model Server의 추론기능 검증을 위해 아래와 같이 진행합니다.

① Model Server의 Status가 "InferenceService is Ready"로 변경되어야 합니다.

② "Status" 필드의 아이콘에 마우스 포인터를 가져가면 상태를 확인할 수 있습니다.

③ Model Server 준비가 완료되면 "Status" 필드의 아이콘이 초록색으로 바뀝니다.

④ 추론기능 검증을 진행하기 위해 "Predict" 버튼을 클릭합니다.

⑤ "Predict" 팝업 페이지 상단의 "Body"에 아래의 JSON 코드 블록을 복사하여 붙여넣기를 합니다.

```
{
  "instances": [
    [6.8, 2.8, 4.8, 1.4],
    [6.0, 3.4, 4.5, 1.6]
  ]
}
```

⑥ "Predict" 팝업 페이지 우측 상단에 있는 "EXECUTE"를 클릭합니다.

⑥ "Predict" 팝업 페이지 하단 "Body"에 표시되는 예측 결과를 다음과 비교합니다.

{"predictions": [1, 1]}.

# 4. Exploring Advanced Features

✔ KServe는 A/B 배포(A/B testing)와 다양한 딥러닝/머신러닝 프레임워크를 위한 표준 API 패키지 등의 편리한 기능들을 제공합니다.

✔ 튜토리얼처럼 YAML 파일을 업로드하여 Model Server를 생성할 수 있고, Python SDK를 활용하여 Model Server를 생성할 수 있습니다.

✔ Jupyter Notebook과 Kubeflow Pipelines (kfp) Python SDK를 사용하면, Kubeflow 기존 버전에서 제공된 Fairing 기능을 대신할 수 있습니다.

ⓘ Kubeflow 기존 버전에서 제공된 Fairing기능은 Pipeline기능으로 대체되어 개발이 더 진행되지 않습니다.

# 5. Troubleshooting

AI&MLOps Platform 사용오류를 최소화하는 방안은,

✔ 다양한 오픈소스 환경에 따른 컴포넌트 간 버전관리를 철저히 하는 것

✔ 사용자와 자원의 권한/인증에 대한 정확한 정보를 유지하는 것

✔ 사용자 가이드 등 공식문서를 사전에 참조하여 적용하는 것입니다.

# 6. Summary

✔ AI&MLOps Platform에서의 모델 서빙은 KServe 컴포넌트가 담당하고 있으며, YAML 파일을 이용한 직관적인 생성과 검증방식을 제공하고, TensorFlow 등 다양한 딥러닝 프레임워크 기반 Model Server의 배치와 추론도 효율적으로 처리합니다.

# 7. Legend

[✔] Tip

[➜] 따라하기

[🚀] Note

[ⓘ] Notice

[➢] Definition or Terminology

[⚠] Warning

[★] See User Guide or Documentation

[•] 순서가 없는 리스트

[①~⑮] 순서가 있는 리스트