

# Cloud기반 MLOps 서비스

## 개요

Machine Learning (이하 ML)은 분석 및 모델 개발 단계를 넘어 운영 단계로 확장하는 단계에 있습니다. 최근 ML모델을 만들어 서비스에 적용하는 업무가 증가하고 있지만, 현실에서는 모델을 개발하는 시간보다 데이터 수집/분석/모델 튜닝에 더 많은 시간이 필요합니다. 그리고 Production 레벨의 ML System에서는 모델 개발/학습 - 모델 튜닝 - 모델 빌드/배포 - 모델 관리에 이르기까지 MLOps 라이프사이클 상에 복잡한 ML Workflow를 자동화하고 관리하는 기능이 필요합니다.

**Kubeflow**는 이러한 ML Workflow를 지원하는 Kubernetes 기반 오픈 소스 Machine Learning 플랫폼입니다. Google, Cisco, IBM, Red hat 등이 참여하여 '18년 오픈 소스 프로젝트로 공개하였고, '20년 3월 1.0 Version이 Release 되었습니다. **Kubeflow**는 ML Workflow 각 영역을 Kubernetes기반의 여러 오픈 소스들(istio, knative, argo 등)을 적절히 조합하여 확장성 있게 제공합니다.

이 문서는 MLOps를 지원하는 오픈 소스 **Kubeflow**를 SDS Cloud에 구성하고 활용하는 방법을 설명합니다.

## 아키텍처 다이어그램

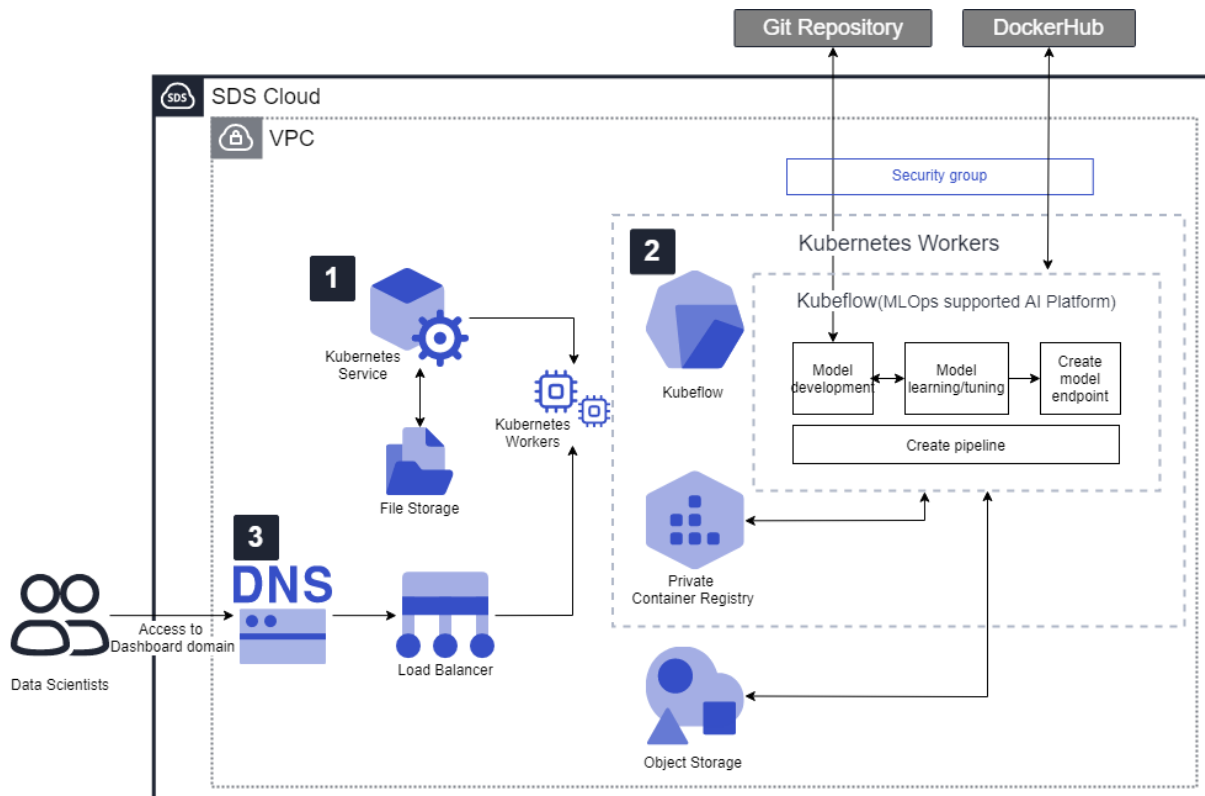


Figure 1. Kubernetes 기반 MLOps를 지원하는 Kubeflow 구성

1. **Kubeflow** 상품 설치를 위해서는 **Kubernetes Cluster**가 필요하다. 사용자는 **Kubernetes Engine** 상품을 먼저 생성한다. 이 때 **Kubernetes Cluster**의 **Persistent Volume(PV)**는 **File Storage** 상품에서 생성한다.
  2. 사용자의 **VPC**의 **Kubernetes Cluster**를 선택하여 **Kubeflow**를 배포할 수 있다. **Kubeflow** 설치가 완료 되면 사용자는 **Kubeflow** Dashboard URL에 접속하여 **Kubeflow**가 제공하는 ML Workflow 기능을 사용할 수 있다.
    - A. 분석 데이터 셋 저장 및 모델 파일 저장 용도로 **Object Storage**를 사용하여 Jupyter Notebook에서 S3 SDK를 이용하여 연계할 수 있다.
    - B. **Kubernetes Cluster**에 컨테이너 이미지 저장소인 **Private Container Registry**를 배포 (Docker Registry 등) 또는 외부 **DockerHub**를 이용할 수 있다.
- ※ **DockerHub**, **GitHub** 등 사용자 **VPC** 외부의 서비스를 이용할 경우 **Security Group**에 방화벽 룰 추가가 필요하다.
3. **Kubeflow** Dashboard를 Domain을 활용하여 사용하려는 경우, **DNS** 상품에서 Domain을 생성하고, **Load Balancer**에서 해당 Domain, **Kubernetes Worker**와 연결하여 통해 Domain 기반의 접근이 가능하다.

## 사용 사례

### A. Kubeflow 기반 생산 공정 불량 판정 시스템 구축

Kubeflow는 Kubernetes 기반으로 뛰어난 재사용성, 확장성, 안정성을 제공합니다.

Kubeflow는 Jupyter Notebook 기반 모델 개발 환경을 제공하고, 하이퍼파라미터 조정 및 결과 검증 기능을 제공하여 여러 모델 간 성능 비교를 할 수 있습니다.

또한 ML Framework(Tensorflow, Pytorch 등)의 GPU를 활용한 분산 학습을 지원하여 모델 학습 시간을 단축할 수 있으며, 모델 배포 시 어플리케이션에서 호출하고 서비스 확장이 가능한 Endpoint API를 제공합니다.

GPU 기반 고성능 인프라(GPU Direct RDMA) 사용 시, 분산 학습 성능을 평균적으로 1.5 배 향상 시킬 수 있습니다.

### B. ML을 위한 CI/CD (모델 개발 및 운영 배포 체계 개발)

Kubeflow Pipeline를 통해 재사용 가능한 Workflow를 생성하여 모델 개발부터 배포까지 CI/CD 구성이 가능하며, 추가 데이터로 모델 재 학습 시 동일 pipeline으로 자동화할 수 있습니다.

## 선결 사항

Kubeflow 설치를 위해 최소 사양 이상의 Kubernetes Cluster, File Storage가 필요합니다.

## 제약 사항

Kubeflow 모델 배포는 현재 구성된 Kubernetes Cluster로만 가능 합니다.

모델 개발 환경과 추론 서비스를 위한 운영 환경을 분리하지 않고 단일 클러스터에서 제공합니다.

## 고려 사항

데이터의 저장 및 모델 저장을 위해 **Object Storage**의 사용을 고려할 수 있습니다.

또한 컨테이너 이미지저장을 위한 Container Registry 구성을 Kubernetes Cluster에 Private Container Registry 구성 또는 외부의 DockerHub 사용을 고려할 수 있습니다.

추가적으로 Kale, Elyra, Feast 등 Pipeline 구성 오픈소스 툴을 배포하여 **Kubeflow**의 기능을 확장 할 수 있습니다.

## 관련 상품

- Kubernetes Engine
- File Storage
- Object Storage
- Load Balancer
- DNS