

AI/ML

AI&MLOps Platform



Kubernetes-based Machine Learning Platform

AI&MLOps automates repetitive work in the overall pipeline for development, learning, and deployment processes of the machine learning model.

AI/MLOps¹ environments based on the machine learning platform are provided, offering integrated management of learning data and models as well as operational data.



Cloud Native MLOps Environments

AI&MLOps Platform offers ML model development environments optimized for cloud, enabling Kubernetes-based linking with various open source software.



Usage Convenience for Big Data

It provides the standardized environments to support a range of machine learning frameworks from TensorFlow, PyTorch, scikit-learn, and Keras. Also, it allows users to configure the ML pipeline automation from development to training and development of ML models. Thus, the automated process makes it easy to configure, create and reuse models as needed.

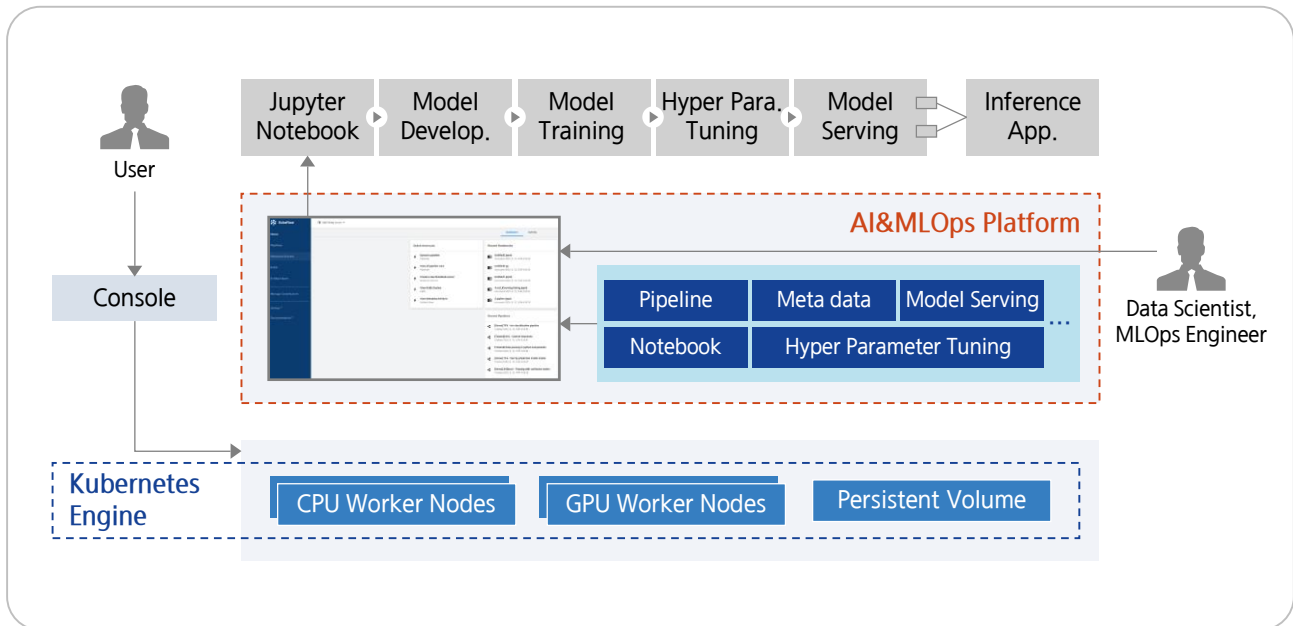


Add-on Features

AI&MLOps Platform provides various features for configuring MLOps environments, including distributed learning job execution and monitoring, inference service management and analysis, and job queue management. Users can also enjoy job schedulers(FIFO, Bin-packing, and Gang-based), GPU fraction, GPU resource monitoring and more add-on features for efficient GPU resource utilization. In particular, a Bare Metal Server-based multi-node GPU and GPU Direct RDMA(Remote Direct Memory Access) help achieve faster processing for large language model(LLM) and natural language processing(NLP).

¹ MLOps : A set of engineering practices aiming to streamline the processes of ML development(Dev) and operations(Ops).

Service Architecture



Key Features

- **Basic features**
 - Create AI platform(auto-deployment/configuration), view(platform version, resource status), and delete
 - Provide Jupyter Notebook : Model development, learning, inference
 - Automate machine learning pipeline workflow
- **Additional features(Available on AI&MLOps Platform)**
 - Advanced AI/ML platform dashboard
 - AI/ML notebook server : Base image, user-defined image
 - AI/ML job : Job creation, template, archive, scheduling, execution, monitoring
 - Support GPU resource monitoring, GPU fraction
 - Providing job operator for Large Language Model training(DeepSpeed)
 - Build and manage user image
 - AI JumpStarter and ETM(Experiment Tracking Management)
 - Serving : Dashboard, register/manage model, inference, predictions visualization
 - Managing platform resource : Manage resource usage by account, monitor resource usage
 - Manage account user/permissions, admin feature, adjust platform configuration

※ This service utilizes Kubeflow, an open source machine learning platform

SAMSUNG SDS

www.samsungsds.com / cloud.samsungsds.com
contact.sds@samsung.com / scp_sales@samsung.com
[youtube.com/samsungsds](https://www.youtube.com/samsungsds)

