Instruction Manual for Cloud Platform-Agnostic Deployment Framework

Overview

This guide provides instructions for using a cloud platform-agnostic deployment framework. It facilitates the deployment of Machine Learning models using GitLab CI/CD pipelines across various cloud providers such as Azure, GCP, and OpenStack.

Prerequisites

Before starting, ensure you have the following:
GitLab Account: Must have access to GitLab Runner.
Machine Learning Models: One or two models prepared for deployment.
Cloud Provider Account: Active accounts with Azure, GCP, or OpenStack.

Repositories Overview

There are two GitLab repositories involved in this setup:

1. Provisioning with Terraform

This repository contains the configuration files for provisioning virtual machines and Kubernetes clusters on the cloud based on user inputs.
Repository URL: https://gitlab.com/thesis3640311/Provisioning_with_Terraform

Key Activities:

Modify the config.yaml file to select the cloud provider and enter credentials.
Commit changes to trigger the deployment pipeline which provisions cloud resources.

Required CI/CD Variables:

APP_REPO_URL: SSH URL of the Machine Learning Models repository.
CURRENT_REPO_URL: SSH URL of the current provisioning repository.
GITLAB_USER_EMAIL: Email associated with your GitLab account.
GITLAB_USER_NAME: Username for GitLab.
RUNNER_TAG: Your GitLab Runner tag (must have SSH access to projects).
SSH_PRIVATE_KEY: Your private SSH key (set to Protected, Masked, Expanded).

2. Machine Learning Models

This repository should contain your machine learning models, ready to be deployed to the provisioned infrastructure.
Repository URL: https://gitlab.com/thesis3640311/IA-Deployement

Key Activities:

A CI/CD pipeline builds a Docker image of the models and pushes it to Docker Hub.
The application is then deployed on the Kubernetes cluster.

Required CI/CD Variables:

APP_NAME1: Deployment name for model 1.
APP_NAME2: Deployment name for model 2 (if applicable).
REGISTRY_PASS: Docker Hub registry password (set to Protected, Masked, Expanded).
REGISTRY_USER: Docker Hub registry username.

RUNNER_TAG: Your GitLab Runner tag.
SSH_PRIVATE_KEY: Your private SSH key (set to Protected, Masked, Expanded).

Deployment Instructions

Step 1: Clone the Repositories

Clone both the Provisioning with Terraform and Machine Learning Models repositories to your local machine.

Step 2: Configure the Repositories

2.1. Machine Learning Models Repository:
- Replace the model1.py and model2.py files with your machine learning models and any associated data sets or folders.
- Commit and push these changes.

2.2. Provisioning Repository:
- Navigate to the config.yaml file.
- Edit the file to specify your chosen cloud provider and credentials.
- Commit and push the changes to trigger the infrastructure provisioning.

Step 3: Monitor Deployment

Once the changes are committed in the provisioning repository, the pipeline will automatically provision the VMs and set up a Kubernetes cluster.
After the cloud infrastructure is ready, the model's repository will trigger its own pipeline to deploy the models onto the newly created VMs.

Step 4: Verify Deployment

You can verify the deployment by SSH-ing into the VM using the provided IP address and credentials.

By following these steps, you can deploy machine learning models seamlessly across multiple cloud platforms using a unified framework.