
Middle East Technical University


Department of Computer Engineering

CENG 495

Cloud Computing
Spring 2024–2025
HW - 2

Due date: 2025-05-11 23:59

1 Introduction

For this homework, you will deploy a Kubernetes cluster, run an application within the cluster and expose the frontend of the application as a service. You will then enhance your application using an LLM agent. As the deliverable of this assignment, you will document your process and prepare a report outlining your steps and solutions.

2 Deployment

- First, select a microservices based application from the Microservices Project List, under "Demo/Toy Projects, mainly for learning or research purpose".
- Install and start minikube. This will be our local Kubernetes cluster for the homework.
- Also install and configure skaffold, we will use this for the development and deployment workflow of our application.
- Deploy the application you have chosen in the first step using `skaffold dev`. (Hint: `kubectl get deployments/pods/services` are great additions to your report at this step)
- Deploy ollama to your Kubernetes cluster. Load a *tiny* model when using ollama, like smollm2.
- Expose the frontend of your application so that it is accessible from your browser under `localhost` (Hint: `kubectl get deployments/pods/services/endpoints` are great additions to your report at this step).
- Add an LLM chatbot to your Web application powered by ollama running in your Kubernetes cluster.

You are encouraged to follow tutorials for minikube, skaffold or Kubernetes in general.

3 Submission

- Submit your report as a `.pdf` file.
- Your report should include sections documenting every step of the 2 - Deployment process: decisions you have made or how you solved the problem at hand, when applicable.
- This is an individual assignment. You can discuss your ideas with your peers but the deployment process and the report should be your own work. The violators will get no grade from this assignment and will be punished according to the department regulations.