**Serverless deployment using IBM cloud code engine**

**Step1**: Install the IBM Cloud CLI

* 1. For linux OS we will have to run the following command :

**curl -fsSL https://clis.cloud.ibm.com/install/linux | sh**

* 1. For Windows™, run the following command in PowerShell as an administrator:

**iex (New-Object Net.WebClient).DownloadString('https://clis.cloud.ibm.com/install/powershell')**

**Step2**: Install the Code Engine CLI by running the following command:

**ibmcloud plugin install code-engine**

**Step3**: Travel to the specified directory and activate the env where these above installations have been made.

**Step4**: login to ibm cloud using this command

-> **ibmcloud.exe login –sso (For Windows OS)**

**ibmcloud login –sso (For Linux OS)**

**Step5**: check the resource groups available to you using this command -> **ibmcloud resource groups**

**Step6**: select the resource group using this command -> **ibmcloud target -g RESOURCE\_GROUP\_NAME**

**Step7**: next we have to create a project using this command ->

**ibmcloud ce project create --name** **my-translator-project (my-translator-project is the name of the project used as example)**

**Step8**: To pull an already created image from the dockerhub and deploy it using the ibm cloud code engine use the following command ->

**ibmcloud ce app create --name my-opt-app5 --image example99/opt-125:v1 --ephemeral-storage 4G --memory 24G --cpu 6 --port 8026 --min-scale 0 --scale-down-delay 300**

where , my-opt-app5 : app\_name

example99/opt-125:v1 – REGISTRY/NAMESPACE/REPOSITORY:TAG

memory – specify the amount of RAM to allocate

CPU – specify the number of CPU’s to allocate

Port – specify the port where the endpoint is exposed (as specified in the Dockerfile or the main python script file)

scale-down-delay – the time (in seconds) duration in which the server will be up after the initial request is sent. It will die down after that if it doesn’t request any request in that time frame.

**Step9**: Send a curl command to the ip/endpoint which will include the prompt for inferencing ->

**curl -X POST https://my-opt-app5.1abcde2345d.eu-de.codeengine.appdomain.cloud/generate -H "Content-Type: application/json" -d '{"prompt": "Write a creative story."}'**