**Steps to Deploy Model in Vertex AI**

Assumptions/Prerequisites:

* Assuming Docker is installed on the local machine.
* Assuming the gcp cli is installed on local machine.
* Assuming gcp authentication is done on local machine
* Assuming the artifactory is Enabled on google console. to enable please refer the steps: https://cloud.google.com/artifact-registry/docs/docker/store-docker-container-images

**Generating the Docker Image and publishing it in gcp artifactory location:**

* Create a Docker file in the main code base with required packages and instructions to run the python code.
* Goto command prompt run the below docker command to create the Image.

1. docker buildx build --tag resumeapi:1.0

* Once the build is completed. we have to push the image to the gcp artifactory.
* Before pushing it to the artifactory we have to tag the local docker image with gcp cloud artifactory location i.e,

1. docker tag resumeapi:1.0 us-central1-docker.pkg.dev/wellsfargo-genai24-8049/hackathon-resume-matching/resumeapi:1.0

* Once tagging is done. we have to push it to the gcp artifactory location

1. docker push us-central1-docker.pkg.dev/wellsfargo-genai24-8049/hackathon-resume-matching/resumeapi:1.0

**Deploying it in Vertex AI**

* Goto Vertex Ai console.
* Select Deploy and Use -> Model Registry
* Click on Import new prompt will open.
* It will Open Import Model
* Please choose Import new model, then enter name and choose the region and click next
* In Model Setting, Select "Import an existing custom container". In Custom container settings section Browse the container image from artifactory, i.e choose resumeapi:1.0 and go to Environment variables section, Enter Prediction Route and Health Route i.e, /predict and /health. and specify the port if you are using other than 8080 in your code. default will be 8080 the click continue
* In Explainability section choose No Explainability and click on import.
* This Process takes 2-3 mins to import the model container in vertex ai.

**Expose Endpoint in Vertex AI :**

* Once the Model is imported, click the model and choose Deploy and Test tab.
* Click on the Deploy to Endpoint. Then it will open a prompt window.
* Define Endpoint

1. Choose create new endpoint and then name your endpoint i.e, resmue\_rest\_api and select access as "standard" then click on continue.

* Model Settings

1. Enter Traffic split as 100%, minimum node as 1 and max nodes as 2,
2. Go to Advance scaling options, Select Machine type and then select accelerator type, then select accelerator count and choose service account.
3. Go to logging section and enable access logging endpoint. and click continue.

* Model Monitoring

1. Disable Enable model monitoring for this endpoint and click on deploy.

This deployment process takes around 15-30 mins depends on machine type and accelerator once its completed we can access the endpoint.

**To Test the Endpoint we need below details**

Project Id: Model Import Id. i.e, 847355328557

Endpoint Id: Deployed Endpoint Id. i.e, 1445930358292873216

Bearer Token: To genrate the bearer token run this command gcloud auth print-access-token from google cli / we have to create a new service and deploy it in gcp with the service account which we used to deploy

Sample Request

URL : https://us-central1-aiplatform.googleapis.com/v1/projects/${PROJECT\_ID}/locations/us-central1/endpoints/${ENDPOINT\_ID}:predict

Request Method : POST

Request Body :

{

"instances": [

{ "instance": "us-central"}

],

"parameters": {

"context": "Lead engineer",

"category": "job",

"threshold": 0.7,

"noOfmatches":8,

"inputPath": "https://console.cloud.google.com/storage/browser/hackathon1415"

}

}

**Steps to Deploy Cloud Function to Expose the API without Google OAuth 2.0**

* Enable the Cloud Functions From Google Console.
* Click on Create Functions. Enter Function name i.e, resumeapi.
* Choose region as us-central1
* Goto Trigger section, select Trigger type as "HTTPS" and select Authenticaiton as "Allow unauthenticated invocations"
* Goto Runtime, build, connections and security settings section,
* select Runtime Section,

1. specify Memory 4 gb
2. CPU as "2",
3. Enter Timeout as 180

* In Autoscalling Section,

1. Specify Minimum number of instances and maximum number of instances

* In Runtime service account section,

1. Select Service account as same which we used to deploy our service in vertex ai. i.e, "App Engine Default service account"

* Choose the Langague as Python.

1. Add Requirements please refer the requirments for code/cloudfunctions/requirments.txt
2. replace main.py code with code/cloudfunctions/main.py
3. click on deploy.

Then this service is deployed in cloud

**Endpoint**

URL :https://us-central1-wellsfargo-genai24-8049.cloudfunctions.net/resumematching-api

Method: POST

Request Body

{

"context": "java dev",

"category": "job",

"threshold": 0.7,

"noOfmatches":8,

"inputPath": "https://console.cloud.google.com/storage/browser/hackathontestdata2024"

}

**Steps to Deploy React UI In Google Cloud App Engine :**

1. First create git repository project.
2. Create the GCP project and subscribe the billing.
3. Open APP Engine inside same project ID space .
4. Open APP Engine console and clone the UI project from git repository.
5. While cloning the project, login with git hub username and git token.
6. Install NPM through App Engine console.
7. Build the UI project there .
8. Delete rest of the folder and file apart from build folder.
9. Create App.yaml file touch command.It is mandatory to deploy the UI project in GCP.
10. App.yaml file will have the all the static file and handler path with runtime data.
11. Finally deploy the application with “gcloud app deploy” command.
12. While deploying the application we have to enable the build option with yes confirmation.
13. Once build is completed .Use “gcloud app brows” command to run the application.

**Steps to test Search UI:**

1. https://wellsfargo-genai24-8049.de.r.appspot.com/
2. Open the Home Screen and Select Find Jobs Button.
3. Then it redirects to the Find Jobs Screen and Enter the job title to search available jobs in the portal.
4. Click on search icon after entering the criteria then it will return the appropriate results.
5. Click on the back button to return to the home screen and select Find resumes.
6. Then it redirects to the Find resumes screen, Enter the search criteria in search box and click the search icon.
7. Then we will receive the response form backend with appropriate result set.