Project-4-Report

Masinde

2023-02-24

Contents

Notebook Instance

The note book instance where the starter code for the project as well as the project can be seen in figure 1. The instance name project4

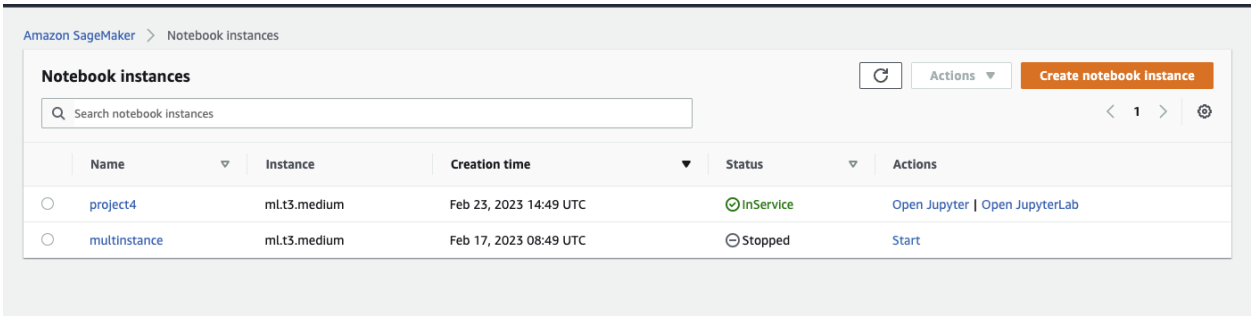

Figure 1: Note Book Instance

Data Buckets (s3)

The data was stored in the s3 bucket as per project instructions. The name of the s3 bucket is project-n4 can be seen in figure 2.

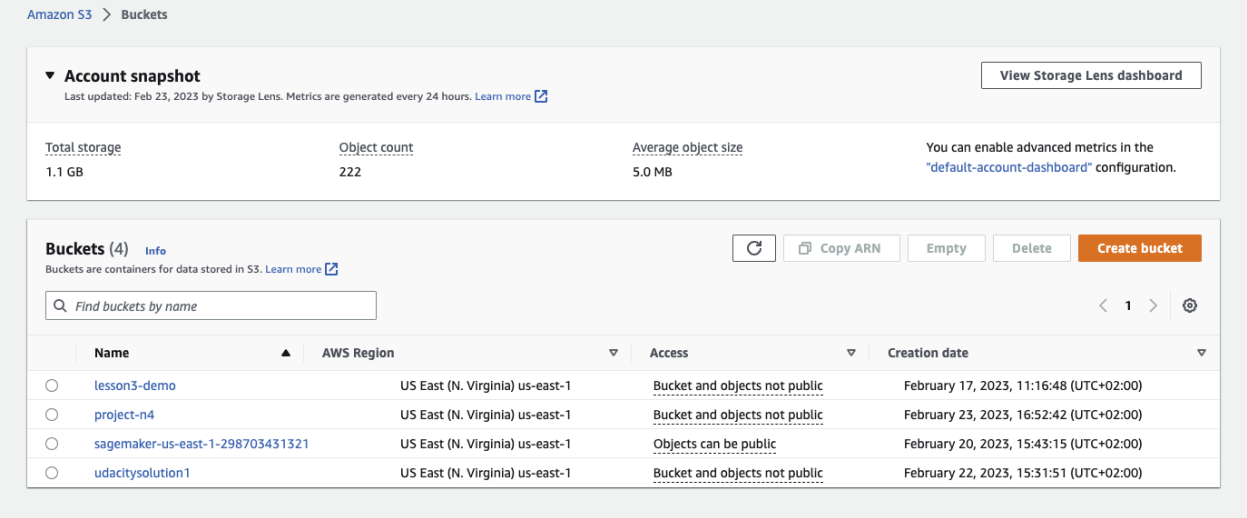

Figure 2: buckets

EC2 Instance

The instance which was used in this project was Deep Learning AMI GPU PyTorch 1.13.1 (Amazon Linux 2) 20230221 figure 2 shows instance configurations.

The reason for choosing this instance

- pre-configuration of Pytorch environment hence no need to install libraries.
- GPU capability of the instance which enhance faster training

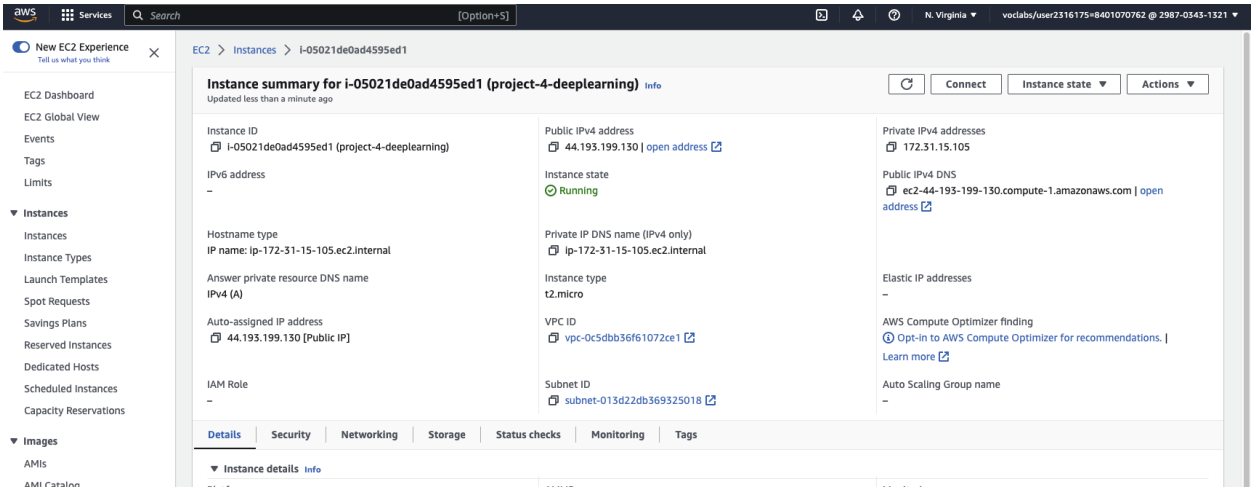

Figure 3: instance for project for 4

Model training

The model was trained as per project instruction. Figure 3 shows model training in the EC2 instance while figure 4 shows completed training.

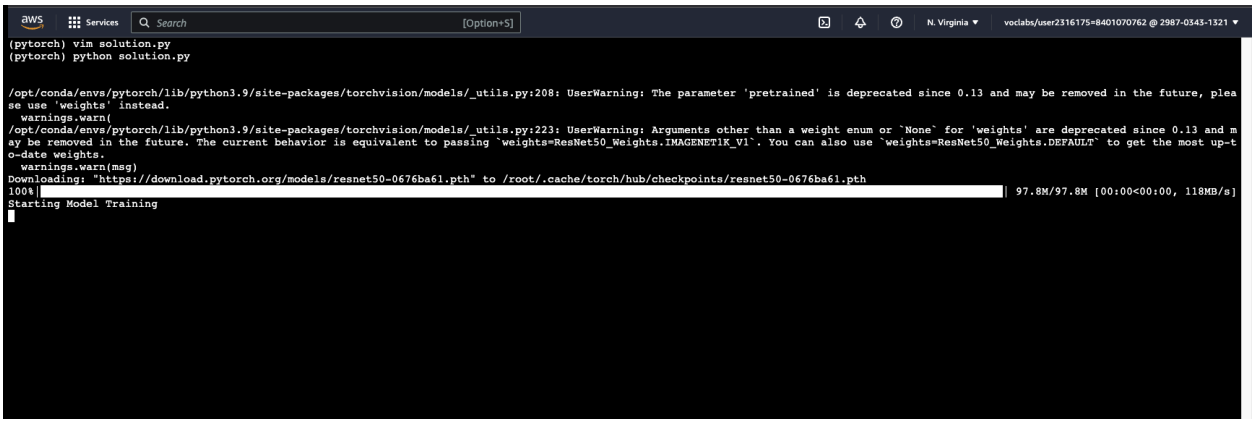

Figure 4: instance training

Lambda Functions

The lambda function which was used in this project to envoke the end point can seen in figure 6 envoking the end point.

Lambda security

The lambda security can be seen in the following figure 6. In order for the lambda function to envoke sagemaker end point lambda function was given full access to sagemaker.

Endpoints

The deployed end points can seen the figure 7 where it shows the single instance training as well as multi-nstance training. The multi-instance training enhances faster training. The number of instances was 8 for multi-instance training and one for the singe instance training.

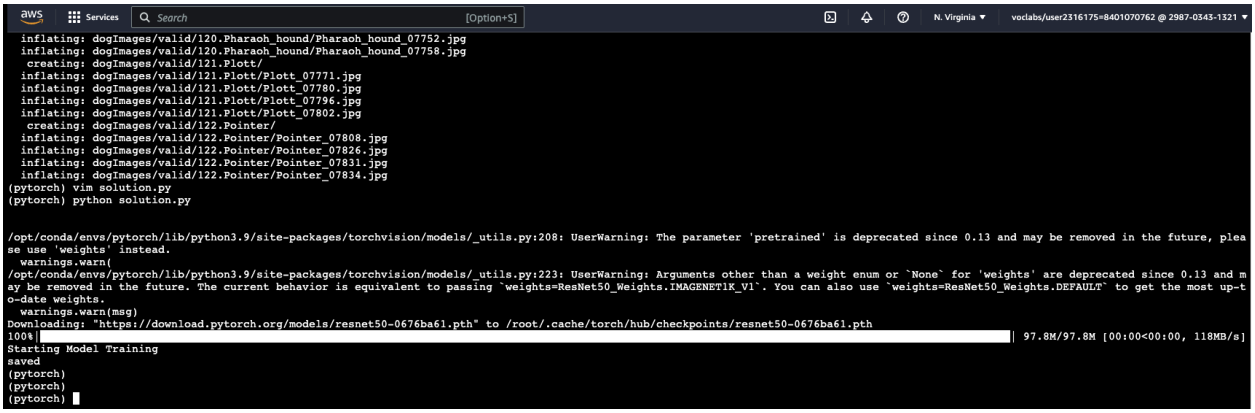

Figure 5: Finished training

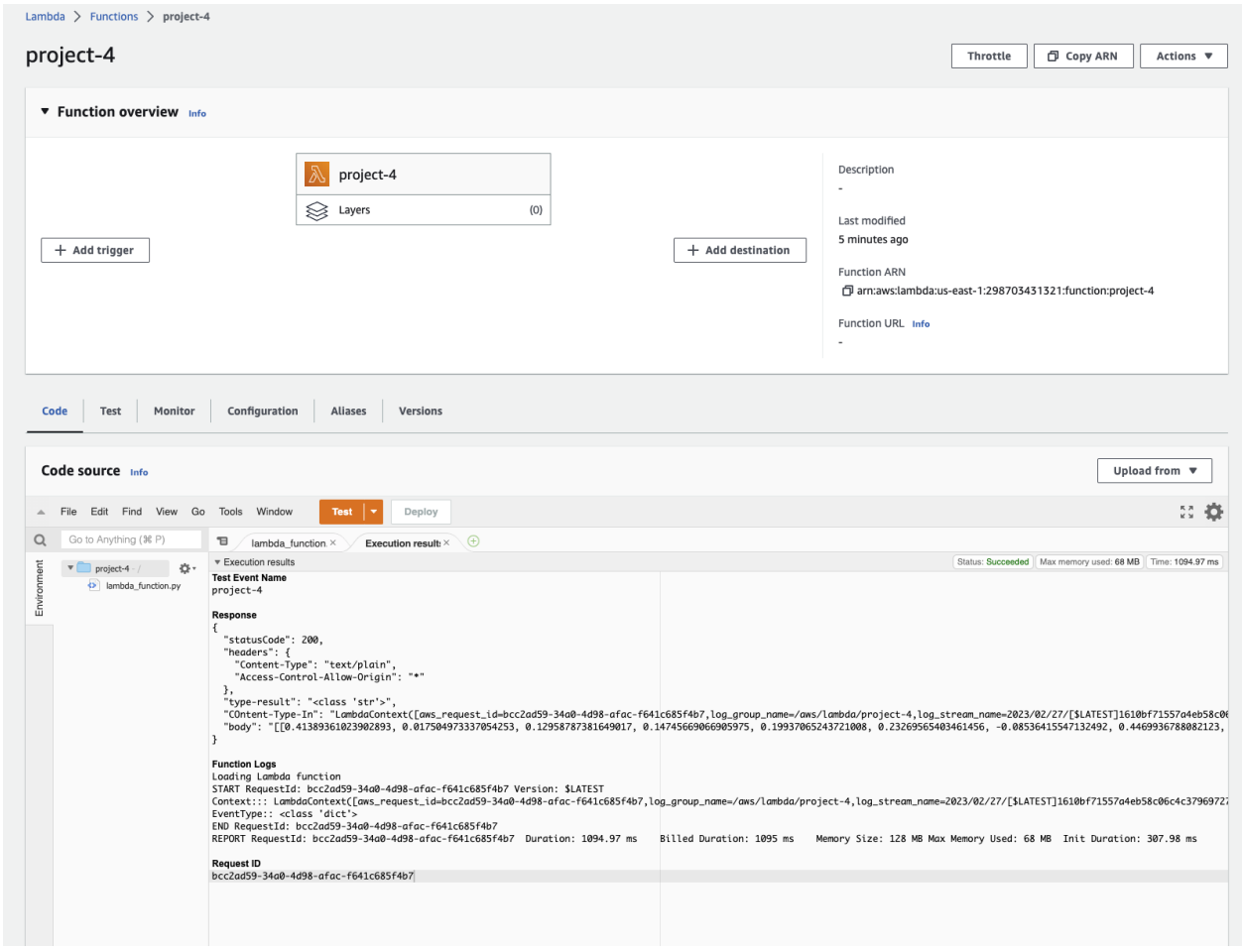

Figure 6: Lambda Function

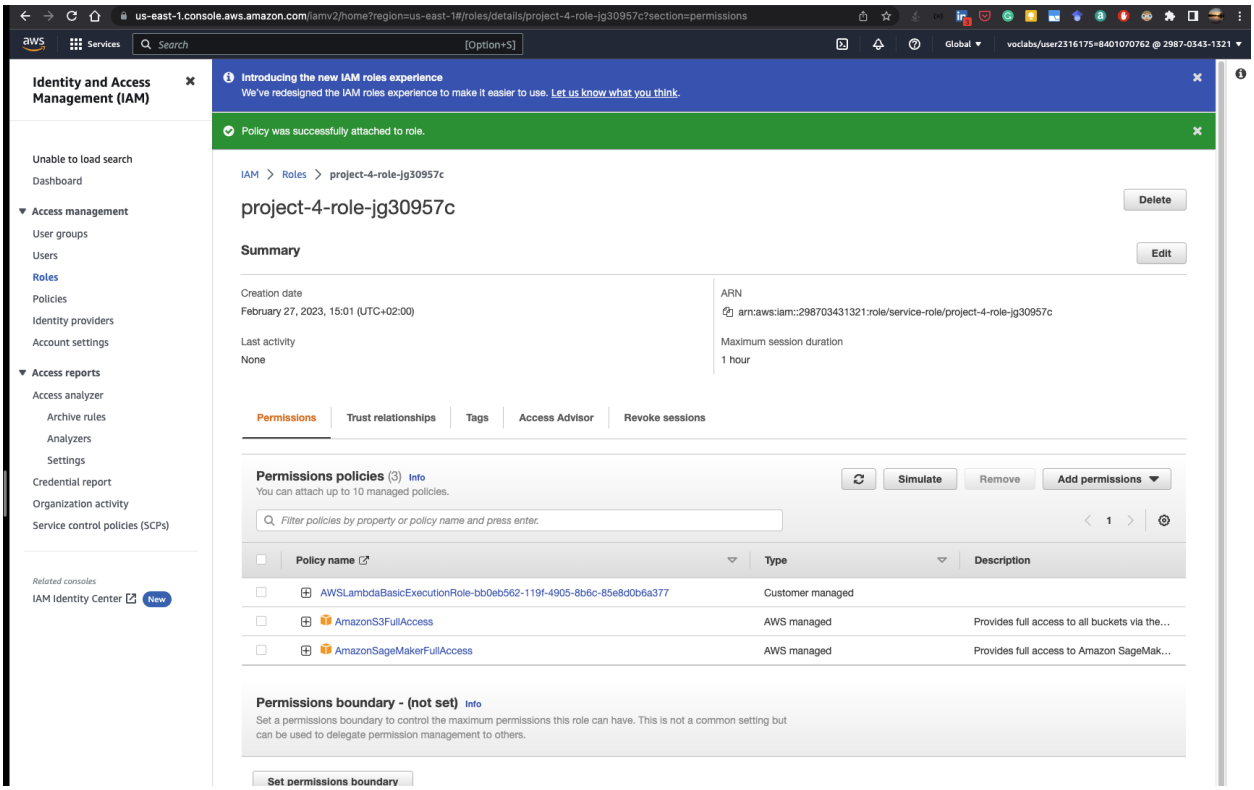

Figure 7: lambda security setup

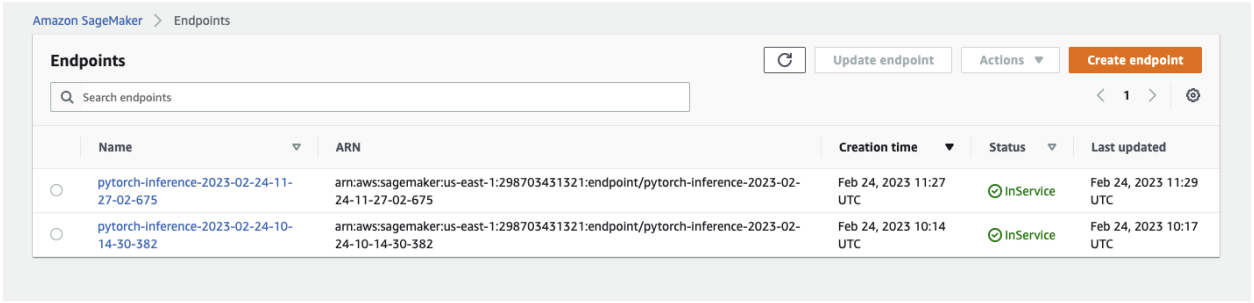

Figure 8: EndPoints

End point auto scaling

Auto scaling was setup to enable traffic when needed.The scaling up time was 30 seconds when the traffic is high hence another instance is needed, as well as cooling down when the traffic is low. Figure 8 shows the number of instance for high traffic which is 5 max instance.

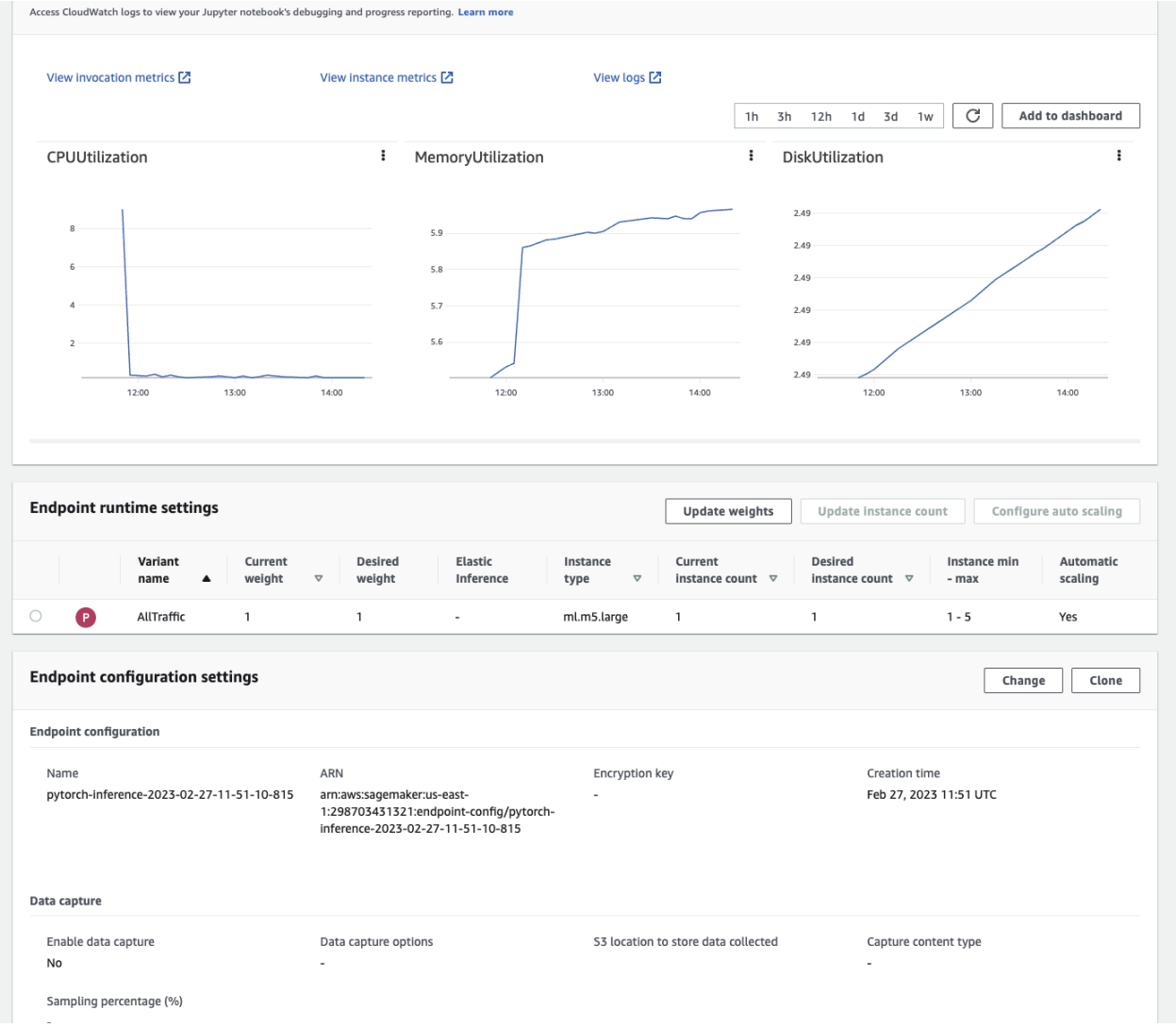

Figure 9: end point auto scaling