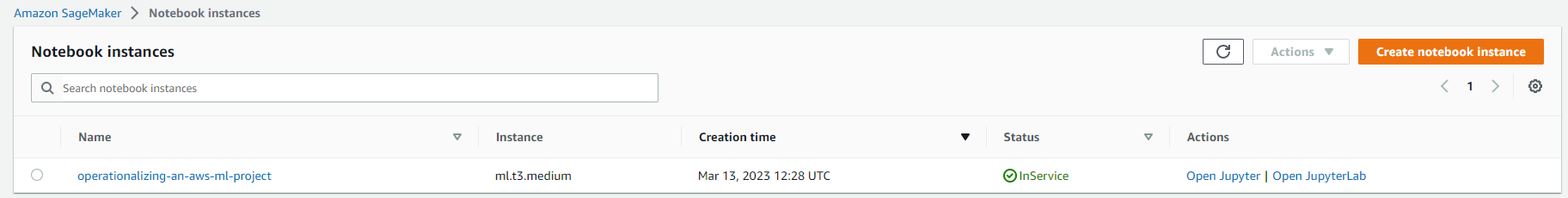
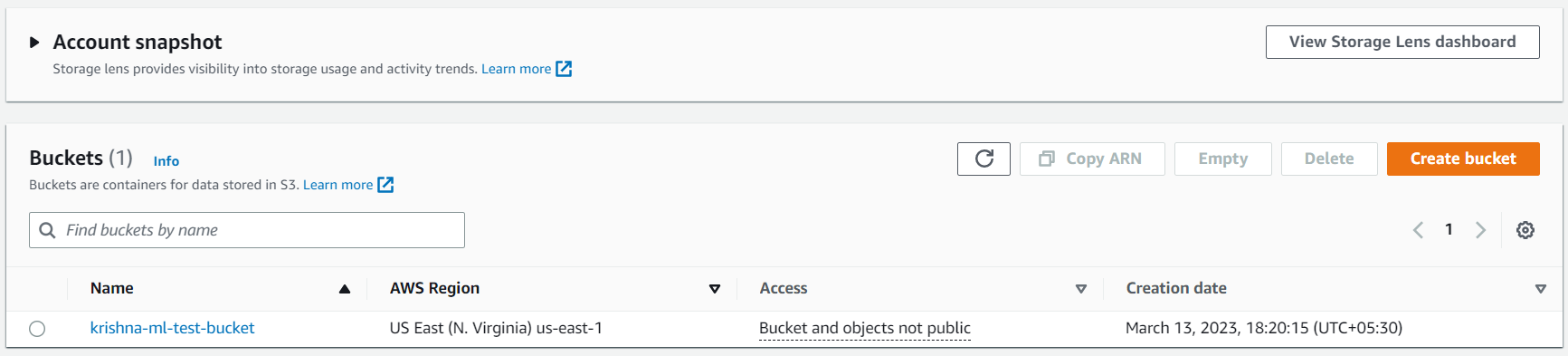
Step 1:Training and deployment on Sagemaker

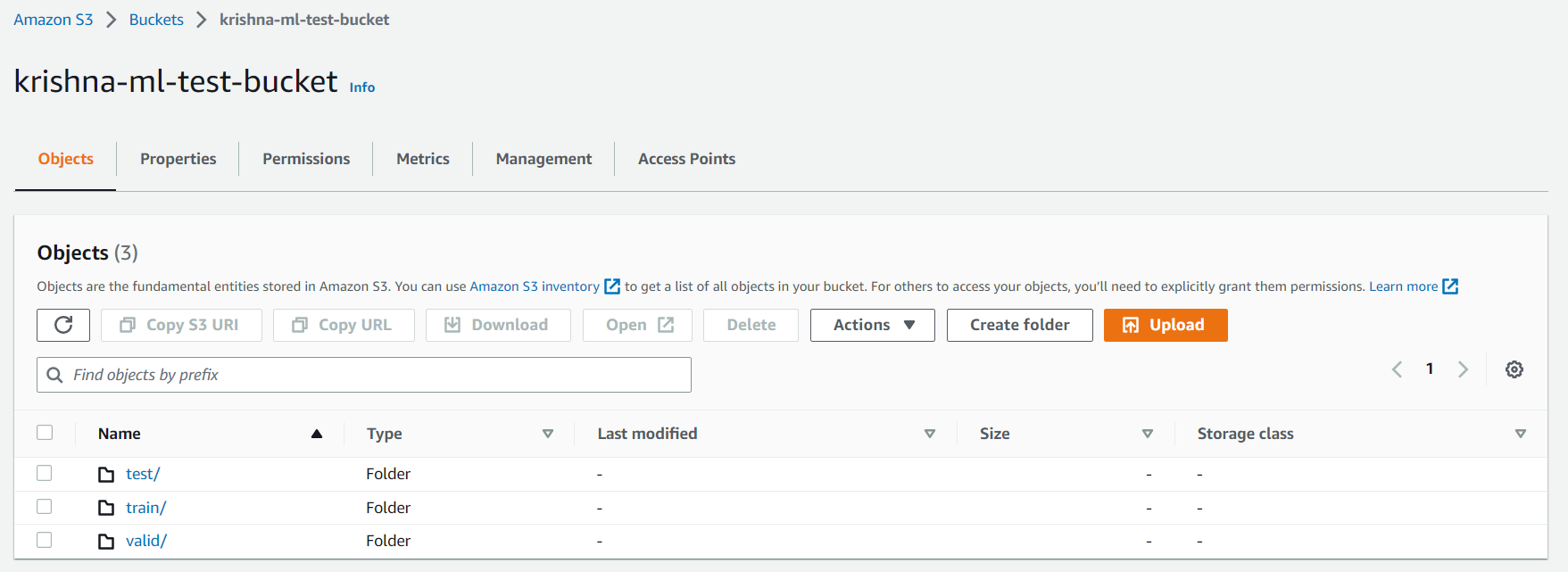
1. Screenshot of created notebook instance(ml.t3.medium) this type was chosen as optimal option because of as an offer of a monthly free tier of 250 hours in aws.

Training and deployment: I have tried m5.xlarge(but traceback occured), so then was changed to ml.p3.2xlarge, which was used and more effective and used also for deploying ML models for inference

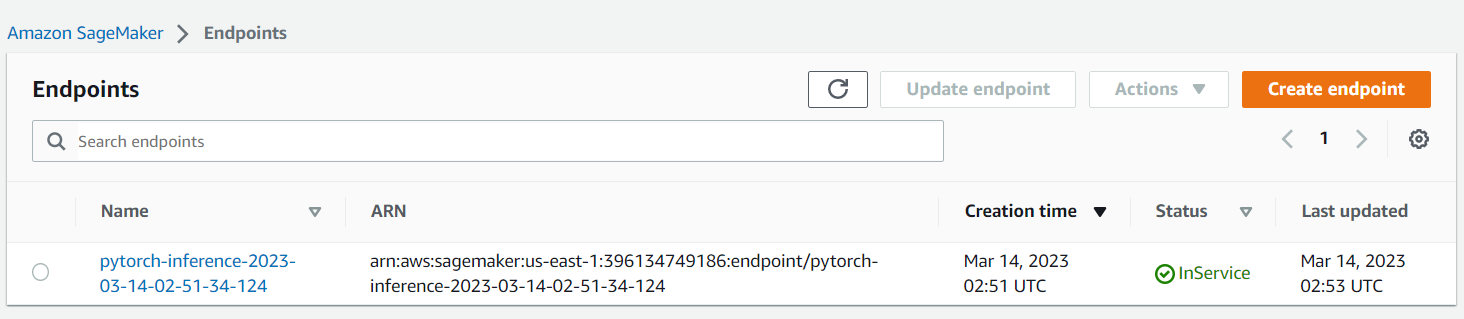


1. “krishna-ml-test-bucket” is created



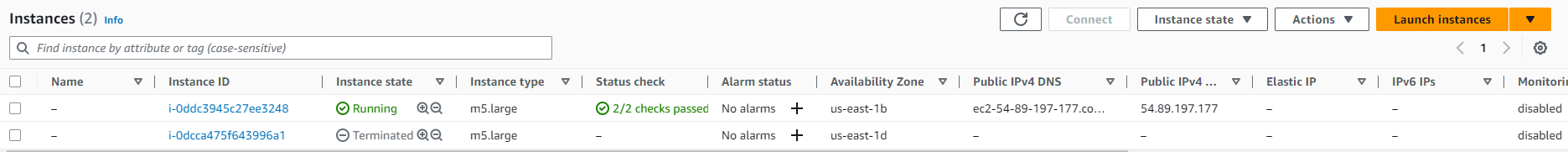


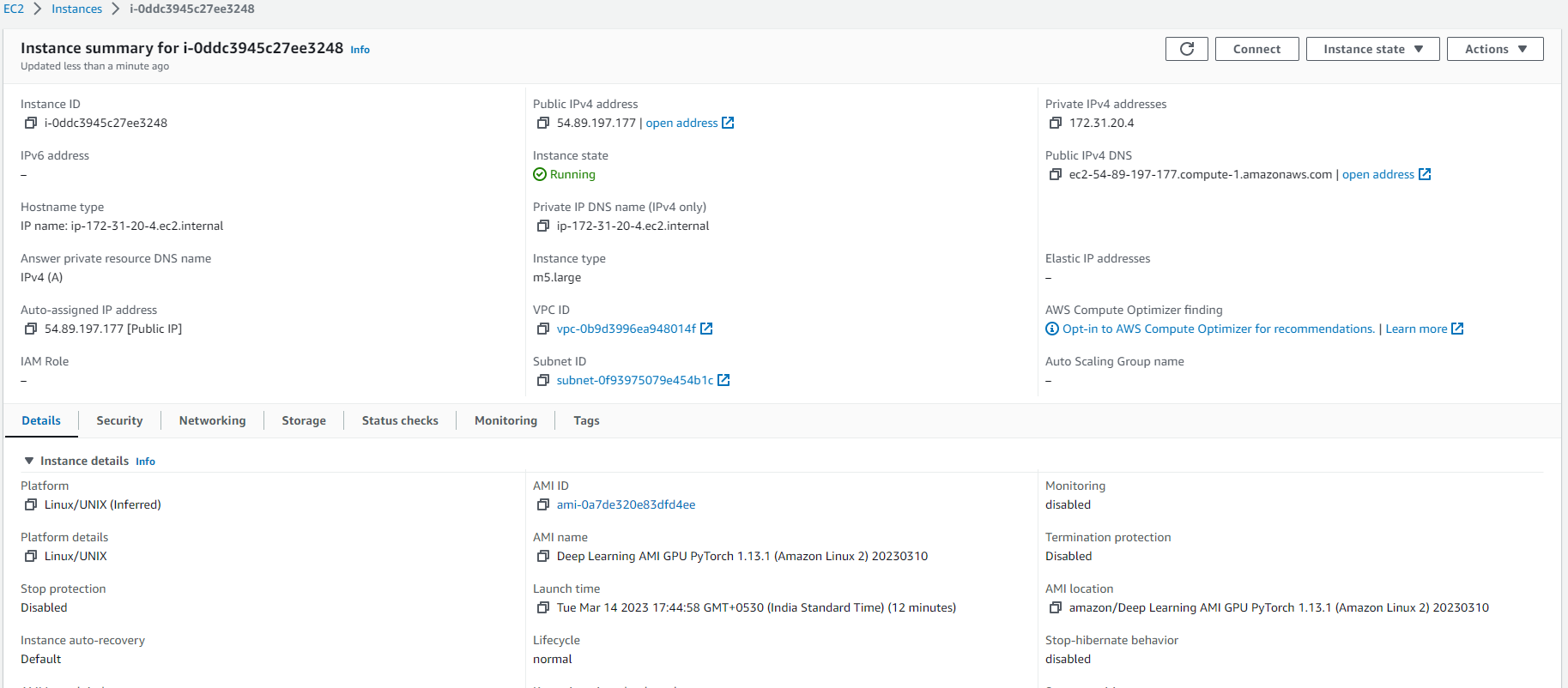
1. After training model was deployed as Inference endpoint



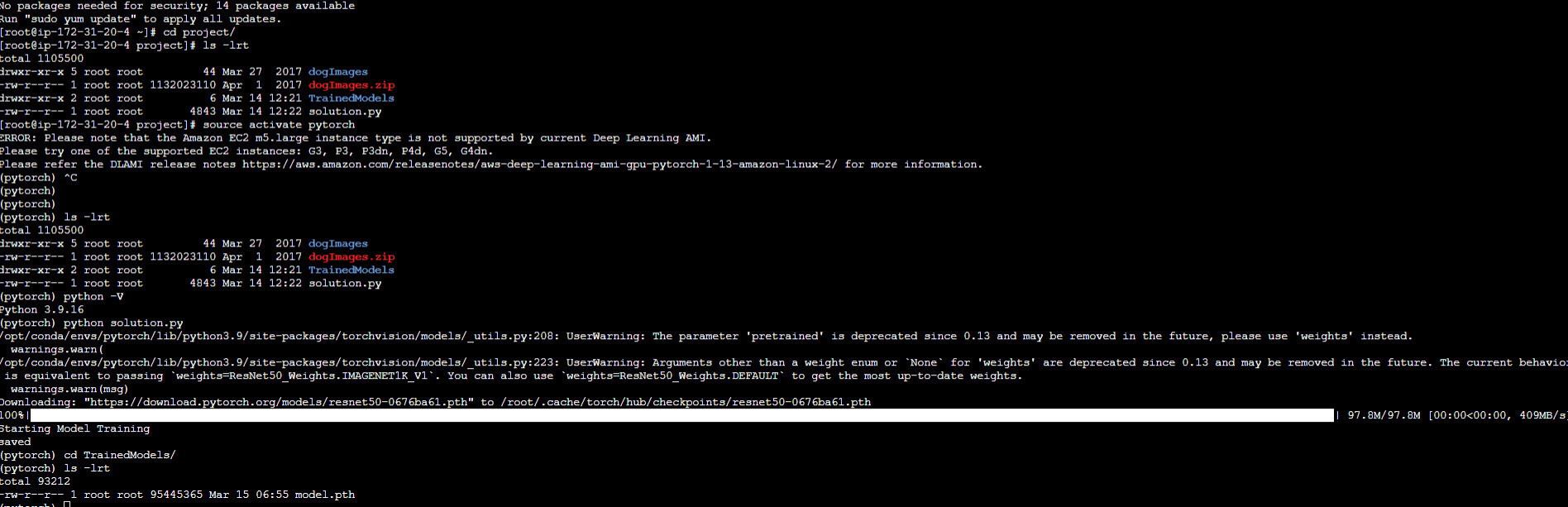
Step 2 : EC2 Training

I have used m5.large instance type in EC2 with spot instance for making cheapier when configured instance type in Advanced details





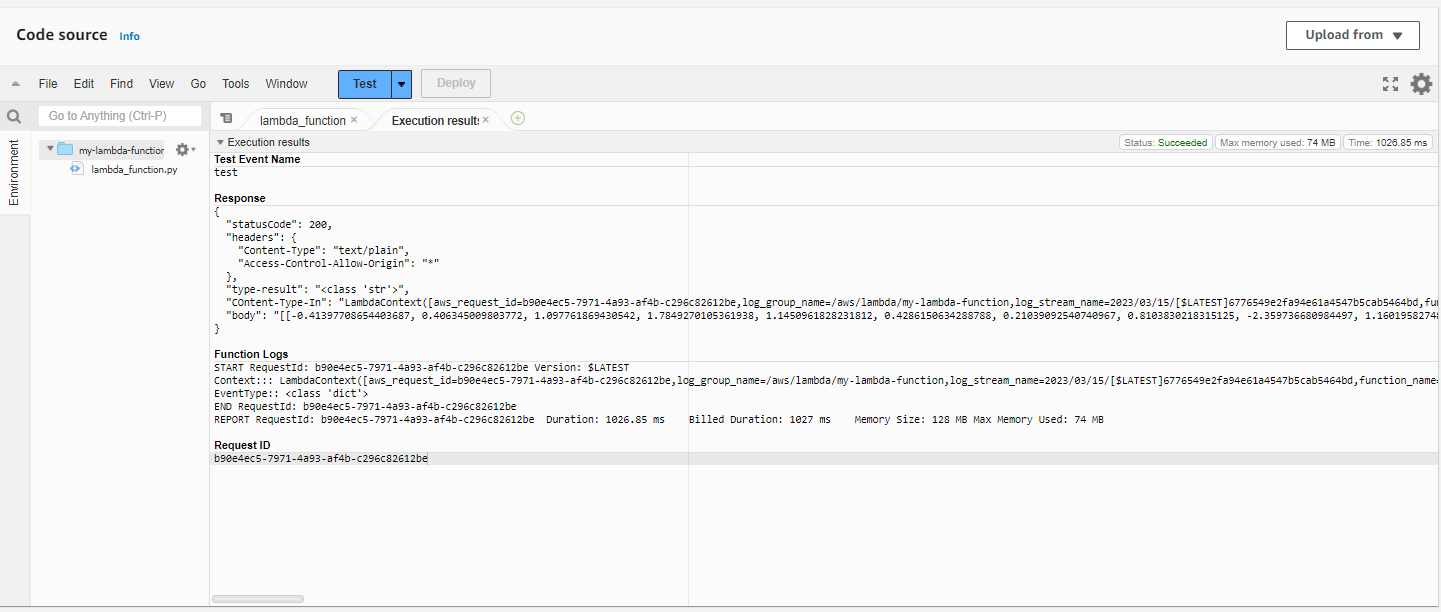
EC2 uses python files with predefined hyperparameters and contains environment(pytorch) that need to activate(in Sagemaker need to install all required dependencies).Training in EC2 takes less time (appr.5 minuts) with comparison in Sagemaker(20minuts), but working in vim/nano editor with pasting of code(ec2instance.py) was little tricky.



Step 3, 4: Lambda function setup, security and testing

Lambda function was created with replacing of endpoint name and configurated test event by replacing url address of test image data. The response shows in output attribute “body” with values of predictions.

Security in using of role in different services is fine, I haven’t issues with attaching policies throughout. But why it’s not allowed by default, it’s the question that following me.



Step 5: Concurrency and auto-scaling

In Lambda configuration in concurrency was changed to reserved with number of 5 instances, chosen because it satisfied our needs. Range of instances between 1 and 5, target metric value 20, scale-in/out cool down is 30-so if I calculate right maximum requests count is 100.This option of configuration was used are optional for cost saving, customizing of them depends on number of triggered innovations->users, budget of project (option to change concurrency to provisioned concurrency>100)

