# Write-up for Operationalizing an AWS ML Project

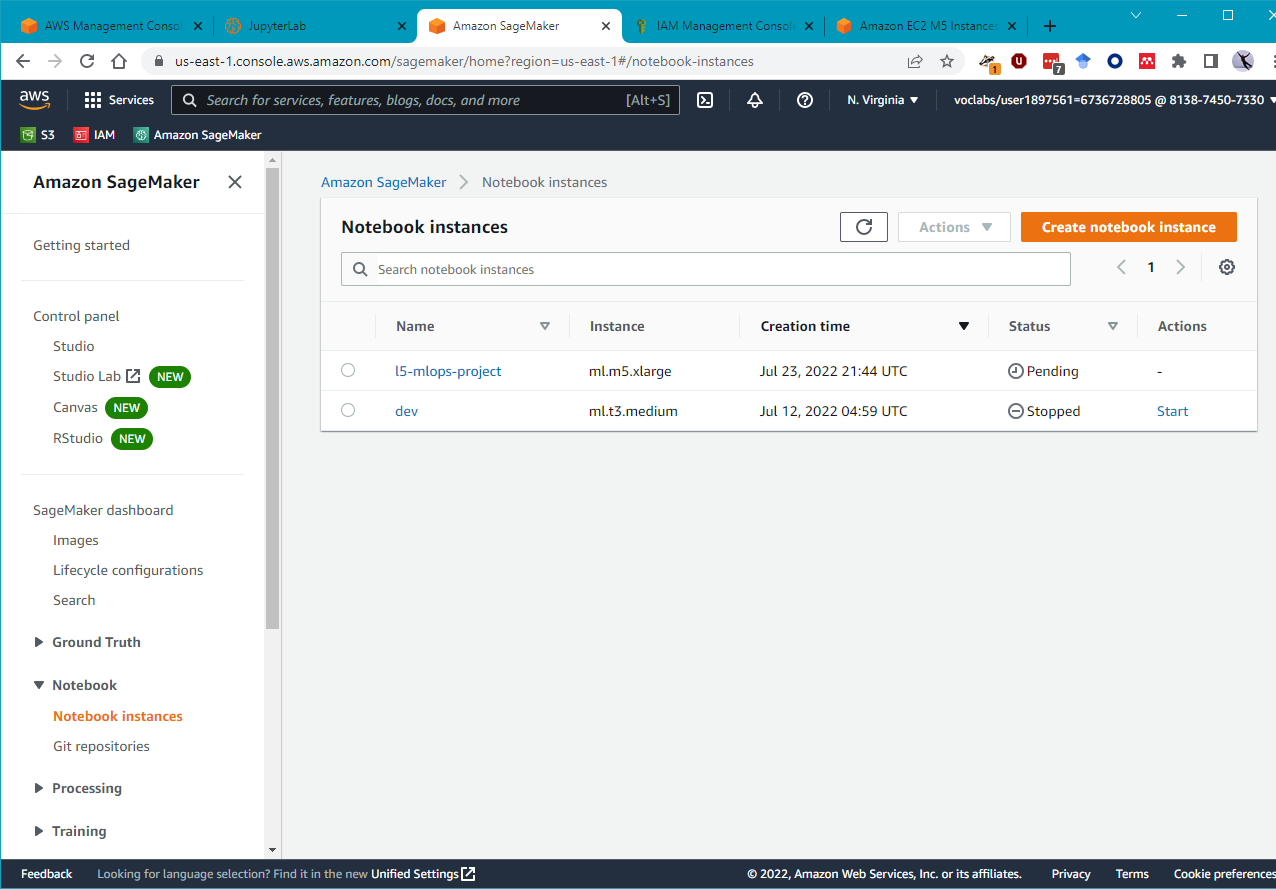
# 1. Amazon Sagemaker Notebooks

Initial setup, training and deployment: Computer vision workload and HPO (Hyperparameter optimization) requires a larger instance to run.

* Chosen Instance Type: ml.m5.xlarge with JupyterLab platform

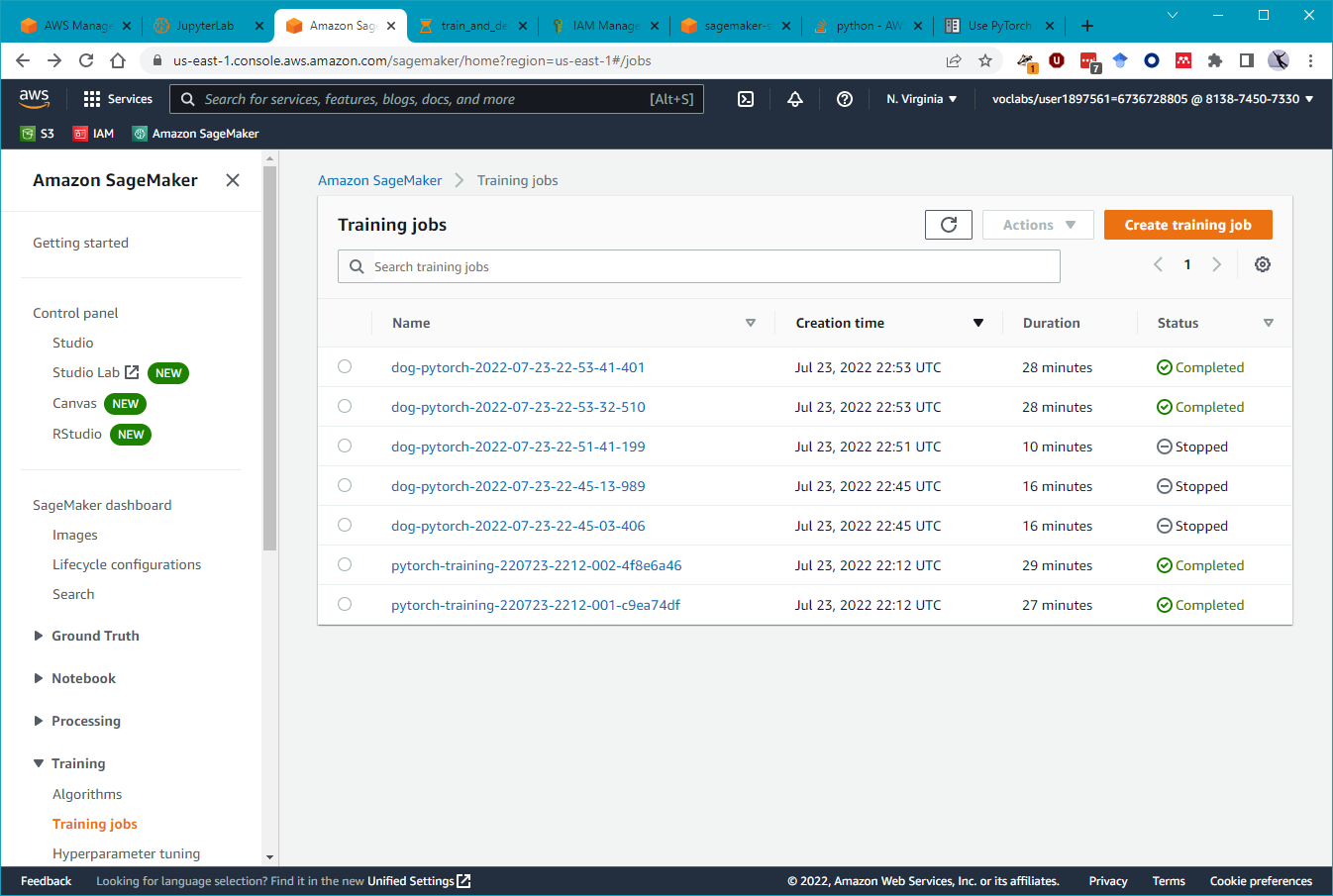
A ml.m5.large instance type is for general purpose compute workload with 4 vCPUs, 16GB RAM and high network bandwidth.

Even though there is no GPU on these instance, M5 instances provide support for the Intel Advanced Vector Extensions 512 (AVX-512) instruction set, offering up to 2x the FLOPS per core compared to the previous generation M4 instances.

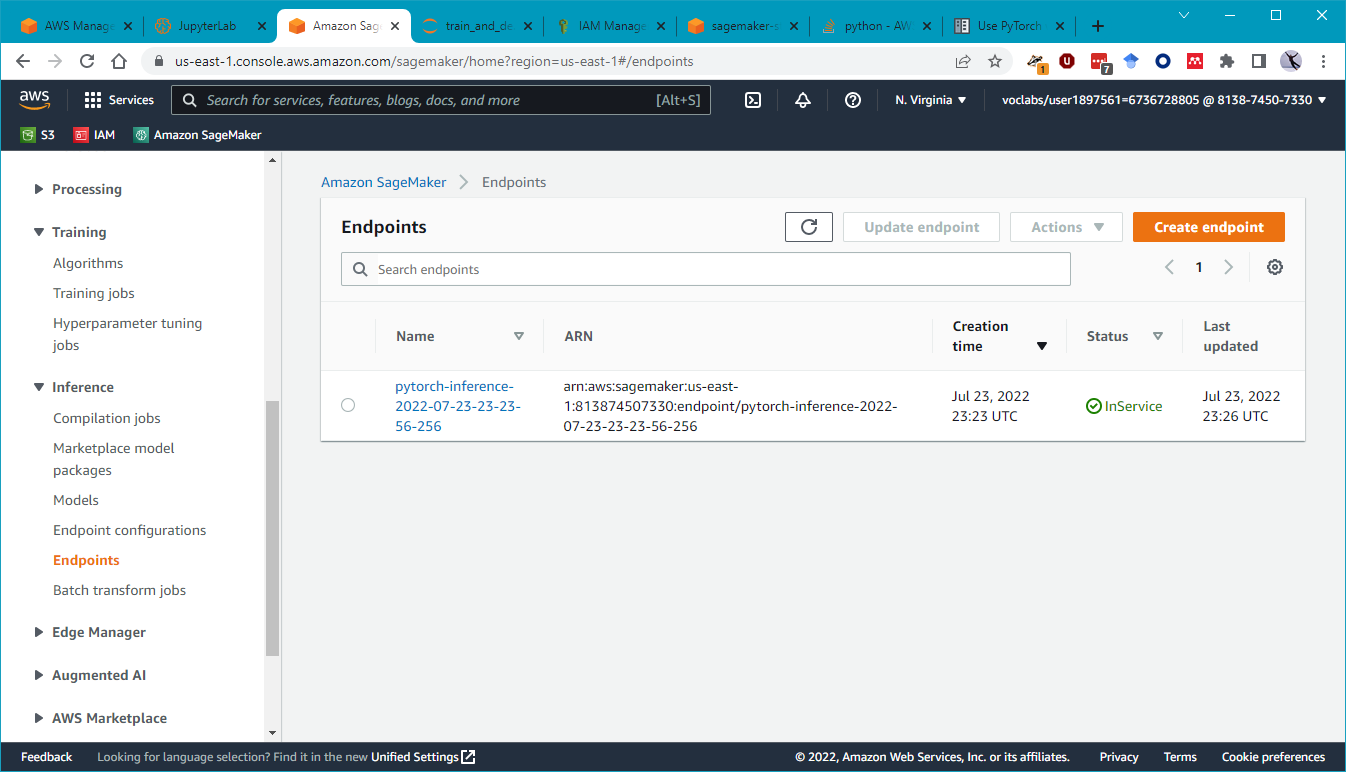


* Download data to “S3://sagemaker-studio-qm1g8vhuin/dataset”

**Training and Deployment (Takes more than 20 minutes)**



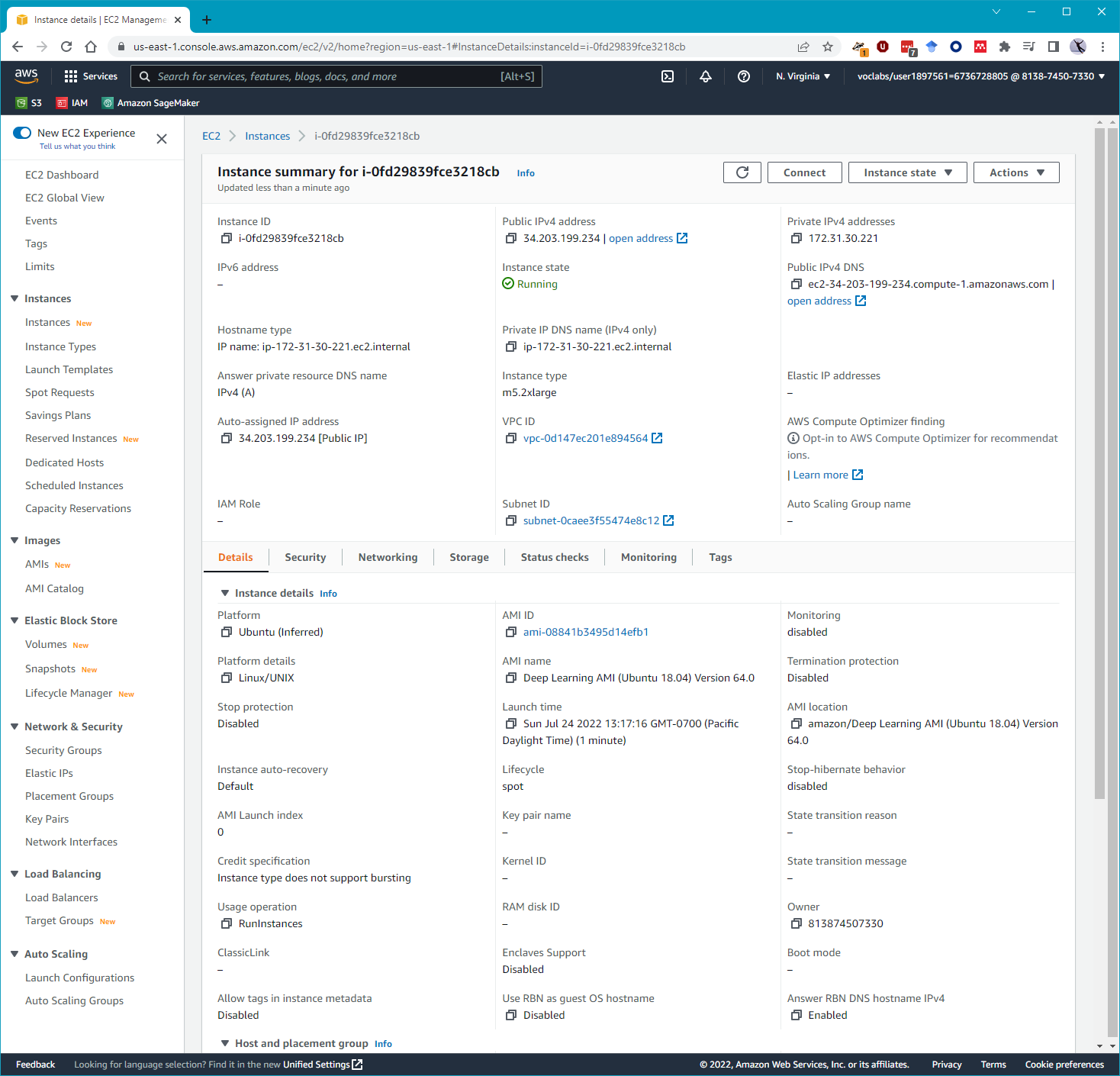
**Deployment Endpoint screenshot**



# 2. EC2 Setup

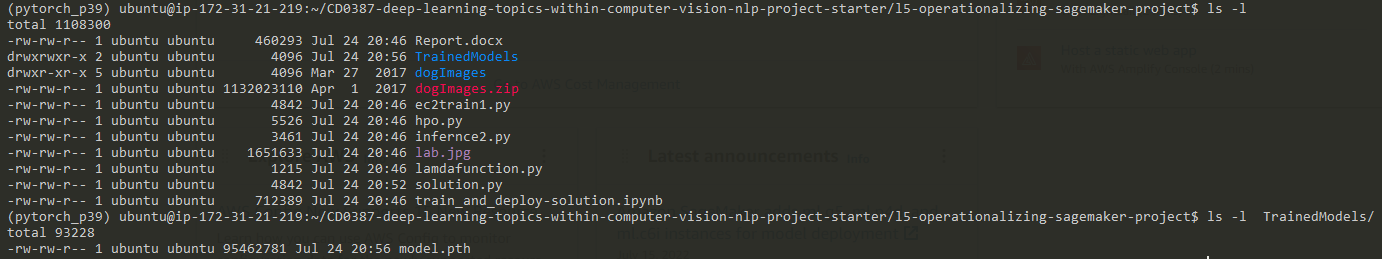
Chosen Instance Type:

* ml.m5.xlarge for the computer vision workloads. Similar computing power and pricing rational as described for notebook instance. However, we need to be careful about the EBS storage and how long we leave this instance on.

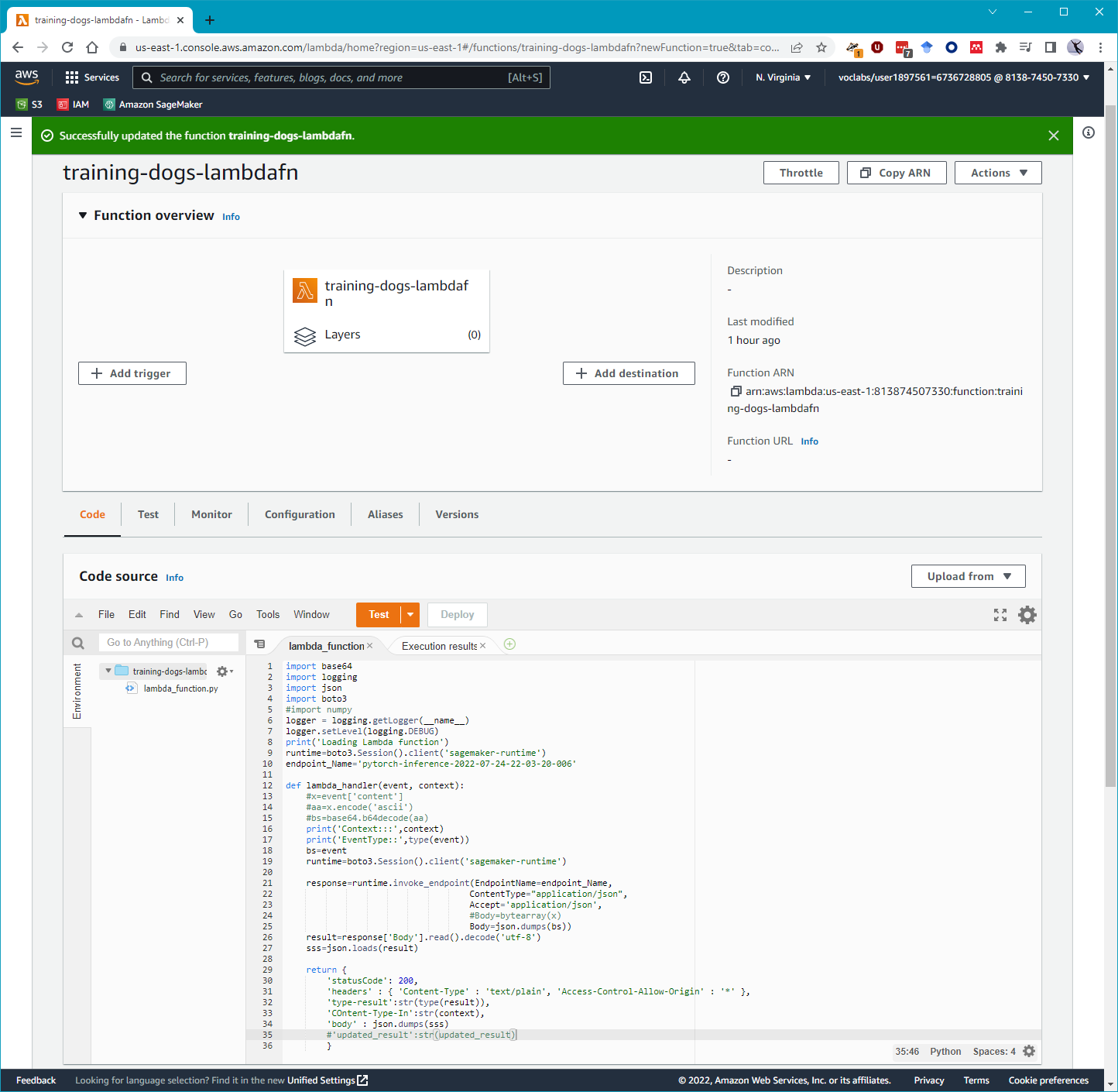


Write about the EC2 code using "Amazon Deep Learning AMI" image

* Major differences between the code is that (1) storage for data is on local path, (2) the training job is local, (3) model is also stored in a folder locally, (4) computation infrastructure used is also local and not submitted to another instance environment for execution.



# 3. Lambda Function

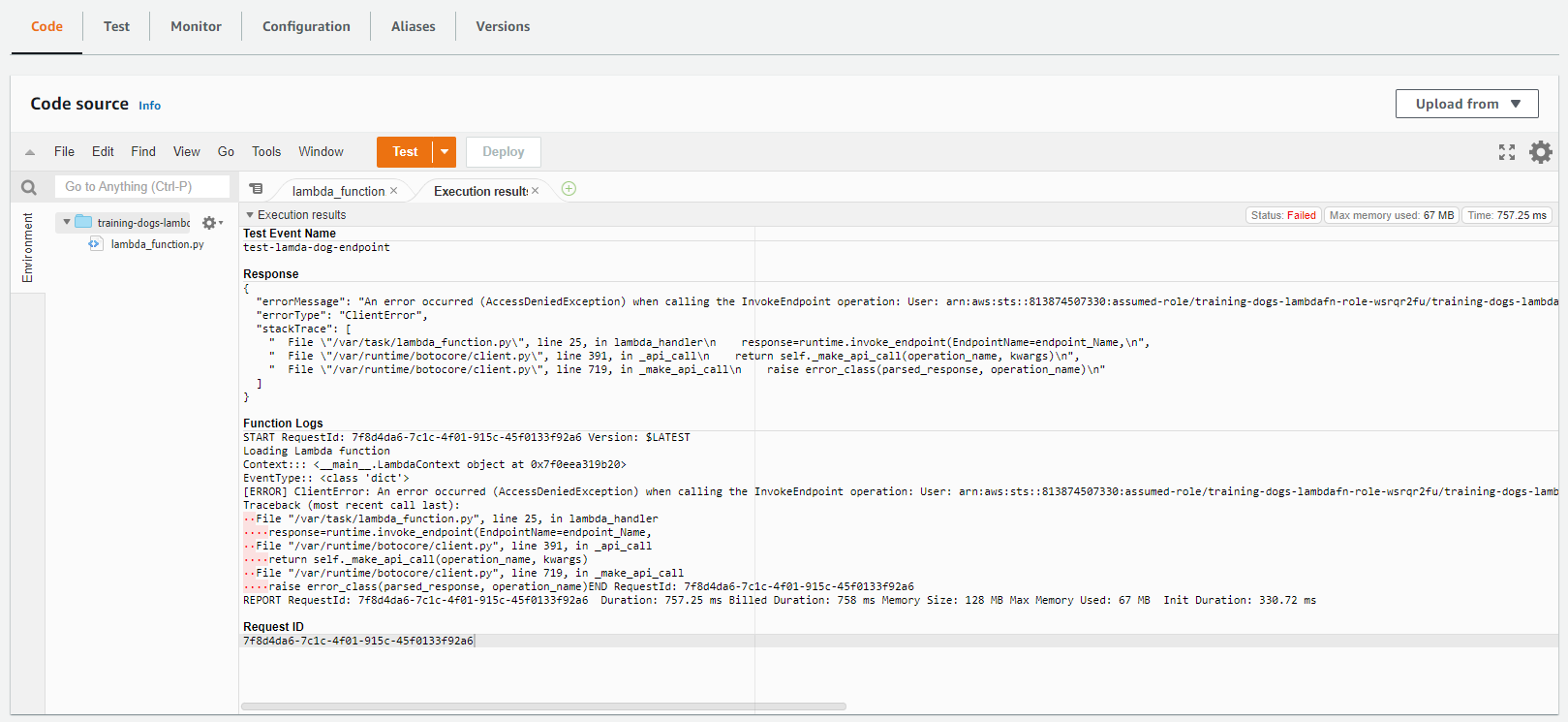


* This lambda functions invokes the deployed dog breed detector endpoint named “'pytorch-inference-2022-07-24-22-03-20-006'”
* It accepts the URL of an image and then uses endpoint inference services to execute the operations and return the results in JSON format for breed identification.

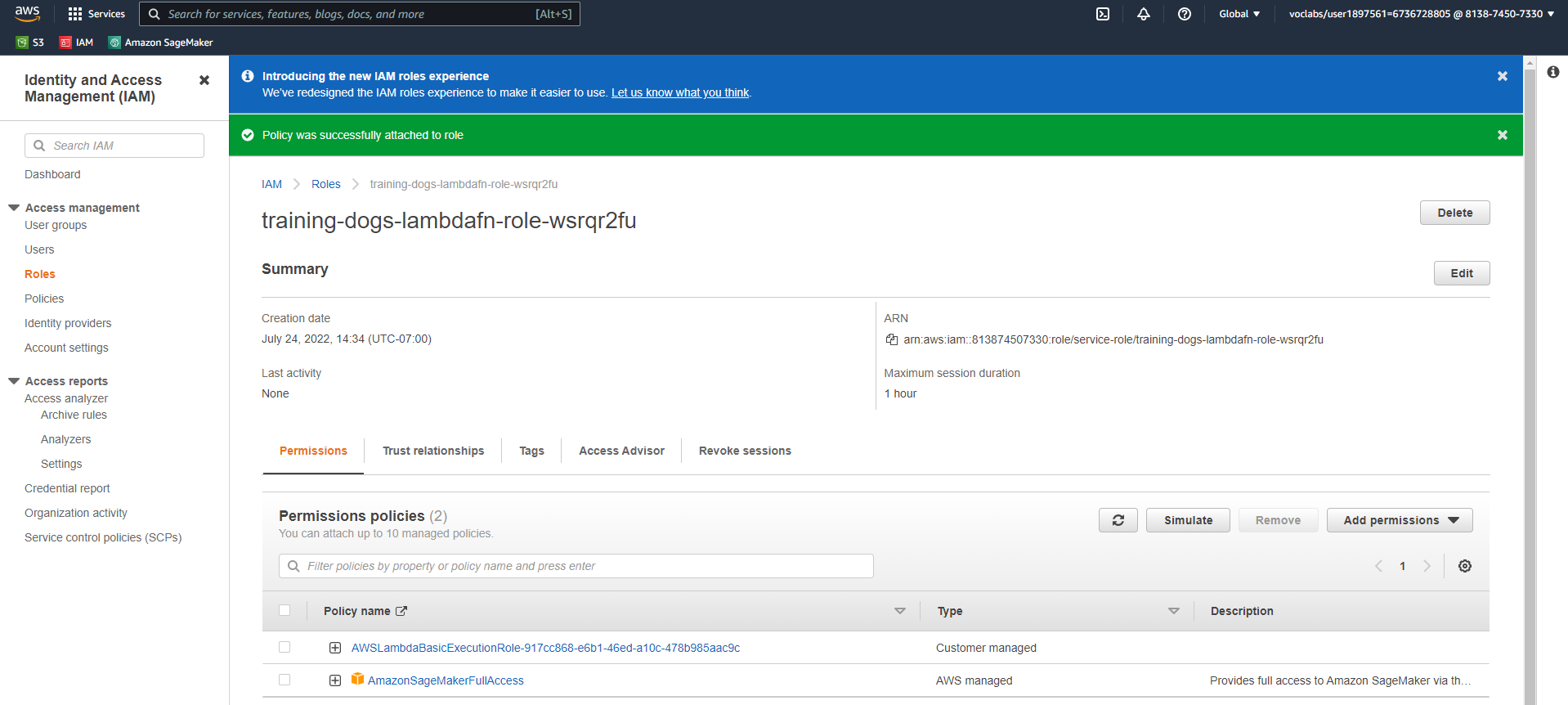
# 4. Lambda Function Security

* Execution gives an error as shown in the logs below

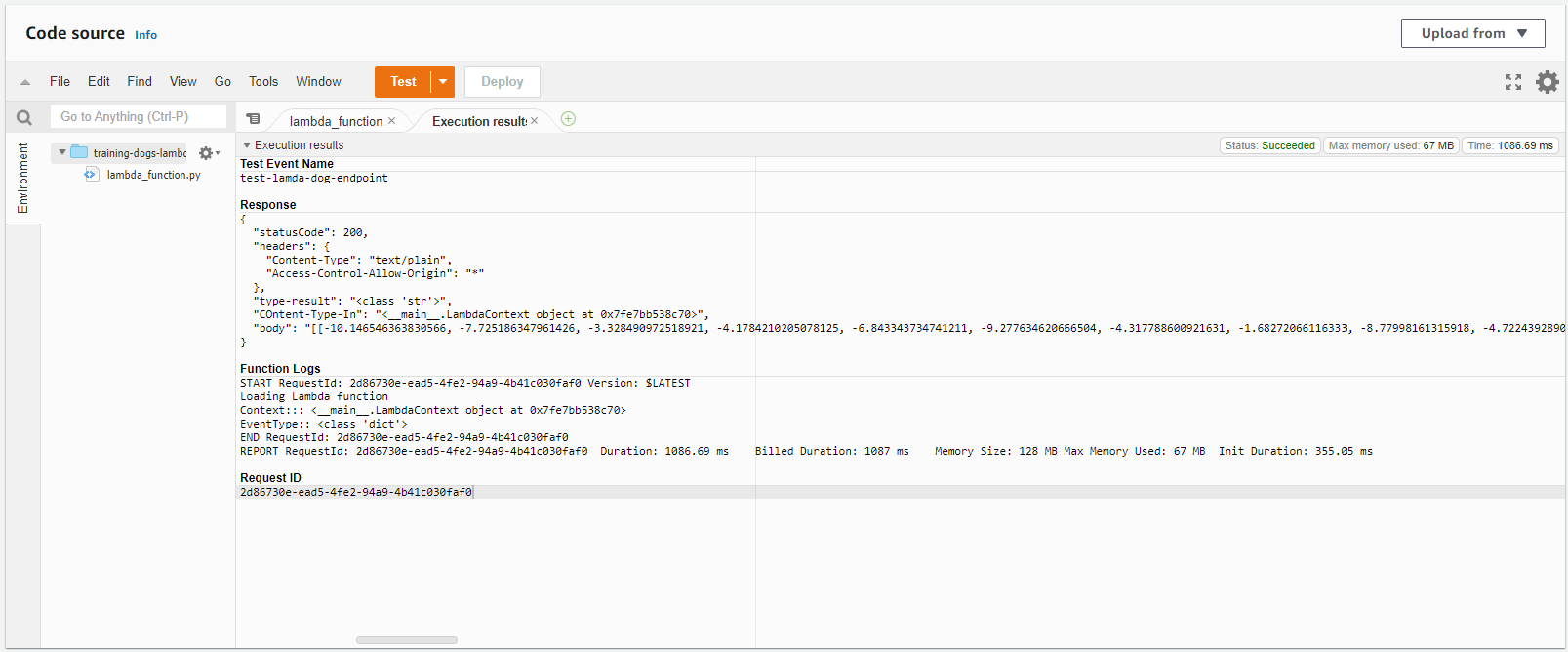
"errorMessage": "An error occurred (AccessDeniedException) when calling the InvokeEndpoint operation: User: arn:aws:sts::813874507330:assumed-role/training-dogs-lambdafn-role-wsrqr2fu/training-dogs-lambdafn is not authorized to perform: sagemaker:InvokeEndpoint on resource: arn:aws:sagemaker:us-east-1:813874507330:endpoint/pytorch-inference-2022-07-24-22-03-20-006 because no identity-based policy allows the sagemaker:InvokeEndpoint action",



* Add or attach an AmazonSagemakerFullAccess policy to the lambda roles



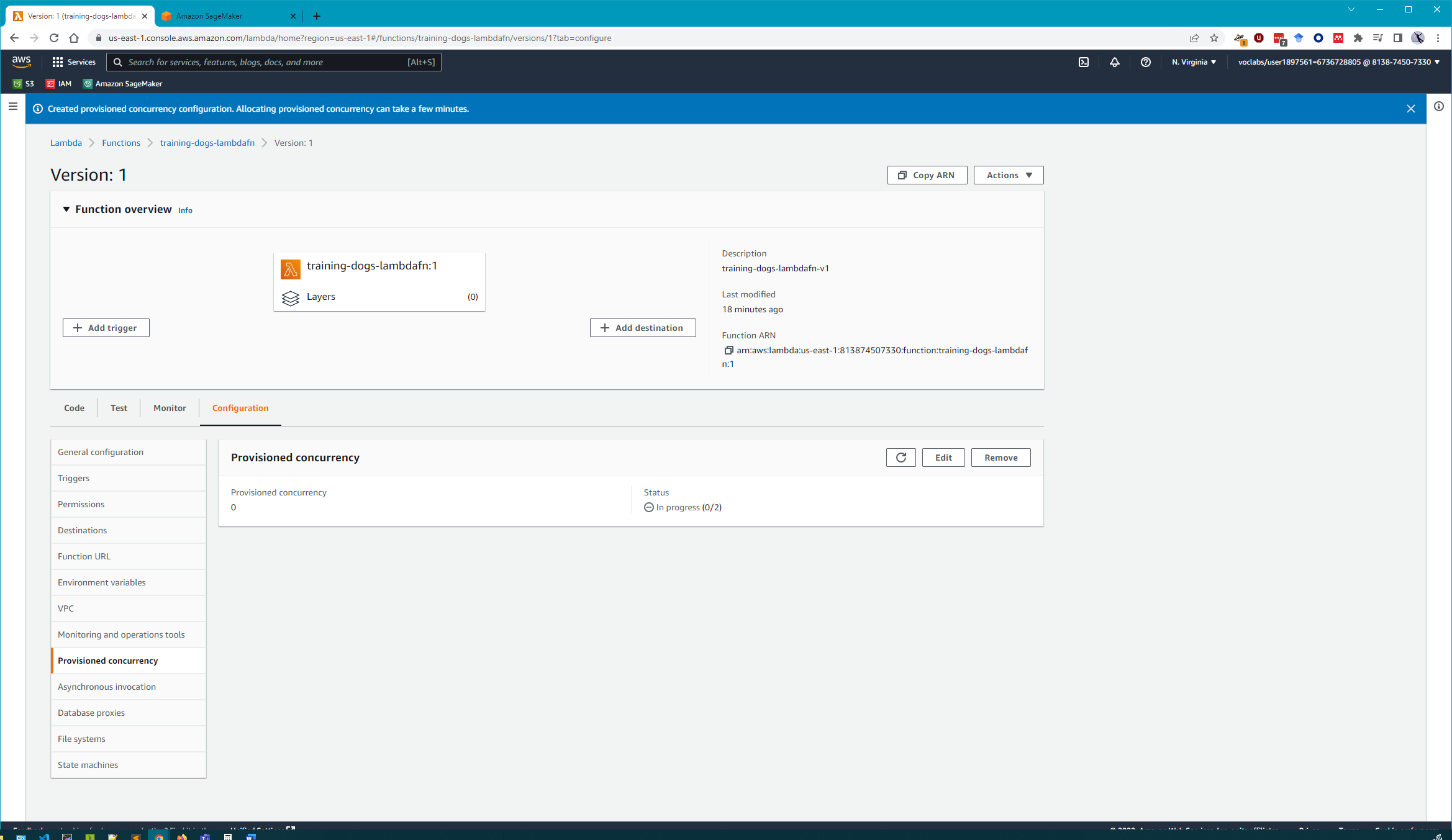
* Execute the lambda function again with success



**This can still be insecure since there are vulnerabilities that the endpoint is exposed to from external request calls. We should also use an access token or web security token to validate the service calls**.

* Concurrency – 2, to consider for at least two simultaneous request to the lambda function.

This can be decided based on assumed usage that there are only two request which may get overlapped and its okay to queue request through a proxy for others.



* Auto-scaling – initially updated weights to setup 2 instances for serving this endpoint,

Thereafter autoscaling was turned on for scaling beyond the maximum instance count of 2 which should typically start during the peak traffic for checking dogs breed. Like during evening dog park time where people see a lot of dogs outdoors.

