

# Operationalizing an AWS ML Project

The goal of this project will be to use several important tools and features of AWS to adjust improve, configure, and prepare a machine-learning model for production-grade deployment.

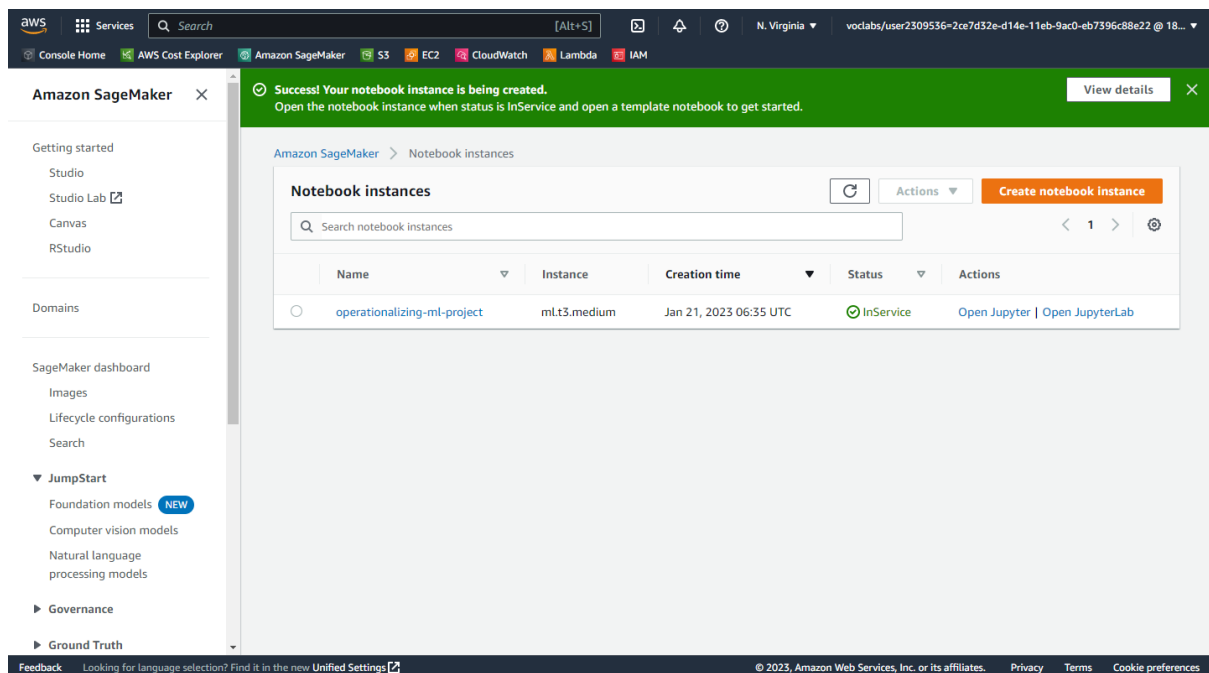
## Training and Deployment on Sagemaker

### Initial Setup

1. Create and open a sagemaker notebook instance
2. Install unzip command

```
sudo yum install unzip -y
```

I choose `ml.t3.medium` as the cheapest compute instance just for running the jupyter notebook.



### Download and Upload Data to an S3 Bucket

I create an s3 bucket with the name `operationalizing-ml`.

```
wget -nc https://s3-us-west-1.amazonaws.com/udacity-aind/dog-project/dogImages.zip
unzip -q dogImages.zip
aws s3 cp ./dogImages s3://operationalizing-ml/dataset
```

The screenshot shows the Amazon S3 console interface. On the left, there's a navigation menu with options like Buckets, Access Points, and Storage Lens. The main area displays an 'Account snapshot' and a list of buckets. The bucket 'operationalizing-ml-project' is highlighted with a green box. Below it, other buckets like 'sagemaker-us-east-1-188932755969' and 'udacitysolutions' are listed.

Name	AWS Region	Access	Creation date
operationalizing-ml-project	US East (N. Virginia) us-east-1	Bucket and objects not public	January 21, 2023, 14:00:07 (UTC+07:00)
sagemaker-us-east-1-188932755969	US East (N. Virginia) us-east-1	Objects can be public	January 19, 2023, 12:40:25 (UTC+07:00)
udacitysolutions	US East (N. Virginia) us-east-1	Bucket and objects not public	January 20, 2023, 14:20:33 (UTC+07:00)

## Training and Deployment

### Single Instance

In my opinion because this training uses a small epoch, I don't need to spend more for a GPU instance, and a CPU-optimised instance like `m1.c5.2xlarge` is enough. It charges **\$0.408/hour** and took **11 minutes and 3 seconds** to finish training.

### Multi Instance

For the multi-instance training, I used spot instances. And the cheapest spot instance available in this account is `m1.c5.2xlarge`. It provides an affordable price of **\$0.174/hour**. I used 2 instances and took **22 minutes and 33 seconds** to finish training in total.

### Endpoint

I created an endpoint from multi-instance estimator

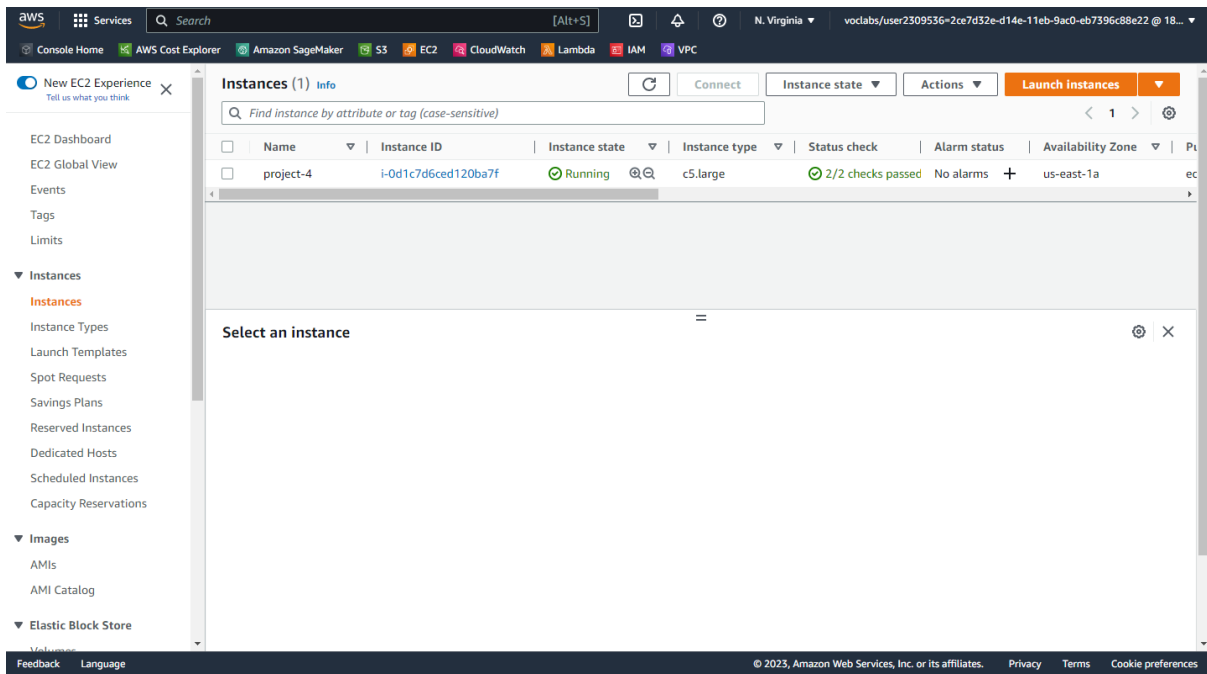
The screenshot shows the Amazon SageMaker console interface. On the left, there's a navigation menu with options like Ground Truth, Notebook, Processing, Training, and Inference. The main area displays the 'Endpoints' page with a table listing the endpoints.

Name	ARN	Creation time	Status	Last updated
pytorch-inference-2023-01-24-07-44-31-770	arn:aws:sagemaker:us-east-1:188932755969:endpoint/pytorch-inference-2023-01-24-07-44-31-770	Jan 24, 2023 07:44 UTC	InService	Jan 24, 2023 07:46 UTC

# EC2 Training

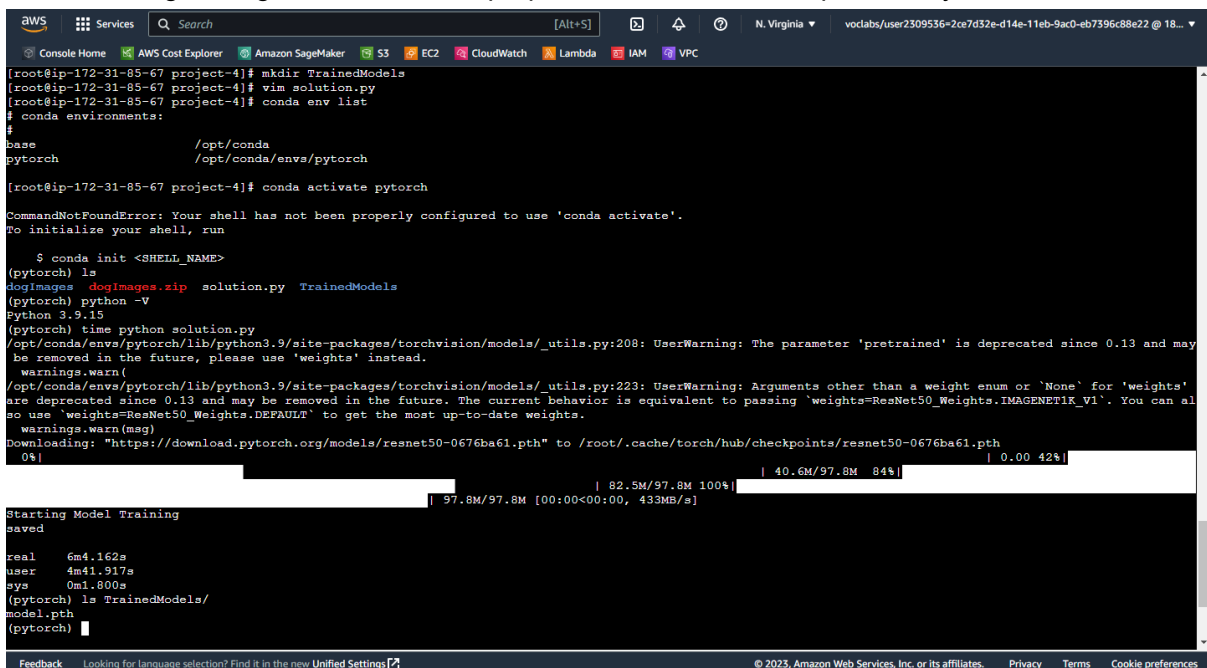
## EC2 Setup

From previous training on Sagemaker, I used `ml.c5.2xlarge`. When I take a look at CloudWatch data. It shows me that the training doesn't use that much storage. So I choose `ml.c5.large` with **Amazon Deep Learning AMI PyTorch 1.13.1 (Amazon Linux 2)**.



## EC2 Model training

Unlike training on sagemaker, the data preparation has to be set up manually.



# Sagemaker Training vs EC2 Training

sagemaker:

- invoke a training job
- input data from s3
- save model data to s3

ec2:

- training on local instance
- load data from local instance directory
- save model data to local directory

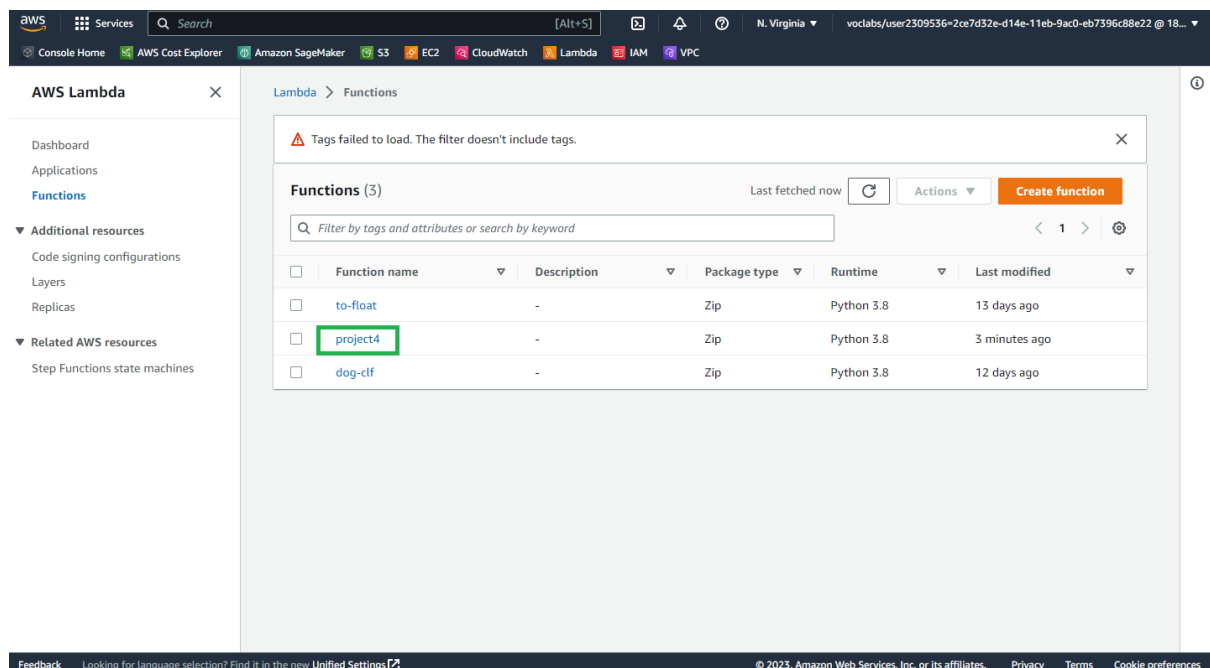
## Lambda Function

This lambda function will invoke an endpoint

pytorch-inference-2023-01-24-07-44-31-770. It only accepts a request with content type `application/json` and returns data with the following format:

```
{
  'statusCode': 200,
  'headers': { 'Content-Type': 'text/plain', 'Access-Control-Allow-Origin':
  '*' },
  'type-result':str(type(result)),
  'Content-Type-In':str(context),
  'body': json.dumps(sss)
}
```

The detail about this lambda function can be found at `src/code/lambdafunction.py`.



The screenshot shows the AWS Lambda console interface. At the top, there's a navigation bar with the AWS logo, a search bar, and various service icons. Below this, the 'AWS Lambda' console is open, displaying a list of functions. A sidebar on the left contains navigation links like 'Dashboard', 'Applications', 'Functions', and 'Additional resources'. The main content area shows a table of functions with columns for 'Function name', 'Description', 'Package type', 'Runtime', and 'Last modified'. Three functions are listed: 'to-float', 'project4', and 'dog-clf'. The 'project4' function is highlighted with a green box. Above the table, there's a message 'Tags failed to load. The filter doesn't include tags.' and a 'Create function' button.

Function name	Description	Package type	Runtime	Last modified
to-float	-	Zip	Python 3.8	13 days ago
project4	-	Zip	Python 3.8	3 minutes ago
dog-clf	-	Zip	Python 3.8	12 days ago

## Security and Testing

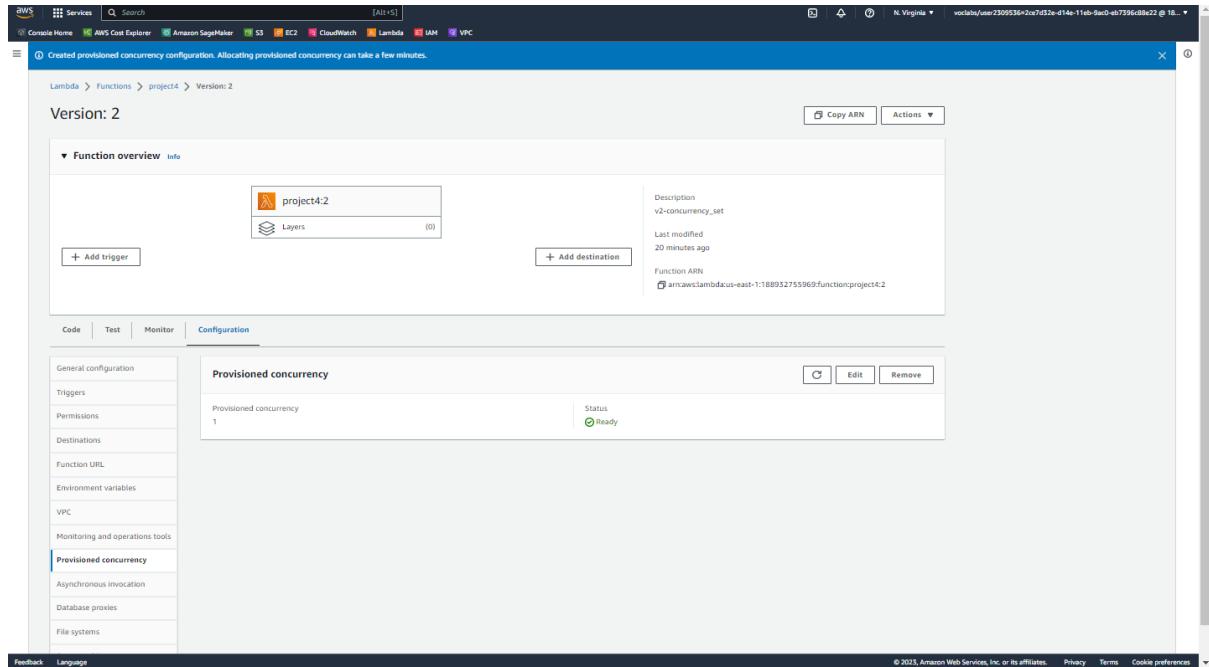
By default for security purposes, AWS Lambda does not have access to AWS Sagemaker. Certain authentication methods are required. One of the methods is by using the IAM role with the corresponding policies. In this case `AmazonSageMakerFullAccess` policy.

The first screenshot shows the AWS IAM console with the role `project4-role-ktvzixj1` selected. The `Permissions` tab is active, showing two policies: `AWSLambdaBasicExecutionRole-f2dddfcb-7d0b-4e15-bb5c-921...` (Customer managed) and `AmazonSageMakerFullAccess` (AWS managed). A green notification bar at the top states "Policy was successfully attached to role."

The second screenshot shows the AWS Lambda console with the function `project4` selected. The `Test` button is clicked, and the `Execution result` is displayed. The status is `Succeeded`. The response is a JSON object with `statusCode: 200` and `headers: { "Content-Type": "text/plain", "Access-Control-Allow-Origin": "*" }`. The function logs show the execution details, including the request ID `8417169e-7013-4e2d-8395-84a8df6874cd` and the duration `1119.77 ms`.

## Concurrency and Auto-scaling

The main purpose of concurrency and auto-scaling are to reduce latency during high-traffic scenarios. In this project, I just worked with low-traffic situations. So a small amount of concurrency is acceptable. I configured concurrency after creating a version of the lambda function. The only concurrency setting available is provisioned concurrency.



For the same reason as concurrency, a small amount of instances is acceptable for auto-scaling. I choose **3** for the maximum instances. It will be triggered if **10** requests come simultaneously.