

# Operationalizing an AWS ML Project

## Step 1: AWS Sagemaker

Write about the Sagemaker instance you created, including a justification of why you chose the instance type you did.

- Chosen **[ml.t2.medium]**:
  - Cost saving.
  - The size of my project doesn't exceed 5 GB EBS.
  - It comes with specs 2 vCPUs, 4GB main memory, no GPU.
  - Amazon Linux 2, Jupyter Lab 3(notebook-al2-v2)
  - Estimated training time was about 20 mins which is reasonable.
  - Fast launch.

The screenshot displays the AWS Management Console for Amazon SageMaker. The top navigation bar shows the AWS logo, Services menu, search bar, and user information (N. Virginia, voclabs/user2074122=5828682428 @ 9338-4504-5900). The main content area is titled 'Amazon SageMaker > Notebook instances'. It features a 'Notebook instances' section with a search bar and a table of instances. One instance is listed: 'MLOps-Operationalizing-AWS-ML-Project' with type 'ml.t2.medium', creation time 'Feb 06, 2023 08:00 UTC', and status 'InService'. Below the table, the 'Notebook instance settings' for the selected instance are shown, including Name, Status, Notebook instance type, Platform identifier, ARN, Creation time, Elastic Inference, Volume Size, Last updated, and Lifecycle configuration.

Name	Instance	Creation time	Status	Actions
MLOps-Operationalizing-AWS-ML-Project	ml.t2.medium	Feb 06, 2023 08:00 UTC	InService	Open Jupyter   Open JupyterLab

Notebook instance settings			
Name	Status	Notebook instance type	Platform identifier
MLOps-Operationalizing-AWS-ML-Project	InService	ml.t2.medium	Amazon Linux 2, Jupyter Lab 3 (notebook-al2-v2)
ARN	Creation time	Elastic Inference	Minimum IMDS Version
arn:aws:sagemaker:us-east-1:933845045900:notebook-instance/mlops-operationalizing-aws-ml-project	Feb 06, 2023 08:00 UTC	-	2
Lifecycle configuration	Last updated	Volume Size	
-	Feb 06, 2023 08:05 UTC	5GB EBS	

# Operationalizing an AWS ML Project

## S3 Bucket:

The screenshot displays the AWS S3 console configuration page for the bucket 'mlops-operationalizing-aws-ml-project-bucket'. The page is organized into several sections, each with a title and a set of configuration options. The 'Bucket overview' section shows the bucket's region (US East (N. Virginia) us-east-1), its Amazon Resource Name (ARN), and its creation date (February 6, 2025). The 'Bucket Versioning' section indicates that versioning is disabled. The 'Tags' section shows no tags are associated with the bucket. The 'Default encryption' section shows that server-side encryption is enabled using Amazon S3-managed keys. The 'Intelligent-Tiering Archive configurations' section shows no configurations. The 'Server access logging' section shows that server access logging is disabled. The 'AWS CloudTrail data events' section shows a permission error. The 'Event notifications' section shows no event notifications. The 'Amazon EventBridge' section shows that notifications are disabled. The 'Transfer acceleration' section shows that transfer acceleration is disabled. The 'Object Lock' section shows that object lock is disabled. The 'Requester pays' section shows that requester pays is disabled. The 'Static website hosting' section shows that static website hosting is disabled.

**Bucket overview**

AWS Region: US East (N. Virginia) us-east-1

Amazon Resource Name (ARN): arn:aws:s3::mlops-operationalizing-aws-ml-project-bucket

Creation date: February 6, 2025, 11:22:12 (UTC+02:00)

**Bucket Versioning**

Versioning is a means of keeping multiple variants of an object in the same bucket. You can use versioning to retrieve previous versions, and restore every version of every object stored in your Amazon S3 bucket. With versioning, you can easily recover from both unintended user actions and application failures. [Learn more](#)

Bucket Versioning: Disabled

Multifactor authentication (MFA) delete: Disabled

An additional layer of security that requires multi-factor authentication for changing Bucket Versioning settings and permanently deleting object versions. To modify MFA delete settings, use the AWS CLI, AWS SDK, or the Amazon S3 REST API. [Learn more](#)

**Tags (0)**

You can use bucket tags to track storage costs and organize buckets. [Learn more](#)

Key: Value

No tags associated with this resource.

**Default encryption**

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption key type: [info](#)

Amazon S3-managed keys (SSE-S3)

Bucket Key: Disabled

When SSE encryption is used to encrypt new objects in this bucket, the bucket key reduces encryption costs by lowering calls to AWS KMS. [Learn more](#)

**Intelligent-Tiering Archive configurations (0)**

Objects stored in the Intelligent-Tiering storage class can be moved to the Archive Access tier or the Deep Archive Access tier which are optimized for objects that will be rarely accessed for long periods of time. [Learn more](#)

Name	Status	Scope	Days until transition to Archive Access tier	Days until transition to Deep Archive Access tier
No archive configurations				
No configurations to display.				
<input type="button" value="Create configuration"/>				

**Server access logging**

Log requests for access to your bucket. [Learn more](#)

Server access logging: Disabled

**AWS CloudTrail data events**

Configure CloudTrail data events to log Amazon S3 object-level API operations in the CloudTrail console. [Learn more](#)

Name: Access

**Event notifications (0)**

Send a notification when specific events occur in your bucket. [Learn more](#)

Name	Event types	Filters	Destination type	Destination
No event notifications				
Choose Create event notification to be notified when a specific event occurs.				
<input type="button" value="Create event notification"/>				

**Amazon EventBridge**

For additional capabilities, use Amazon EventBridge to build event-driven applications at scale using S3 event notifications. [Learn more](#) or see EventBridge pricing.

Send notifications to Amazon EventBridge for all events in this bucket: Off

**Transfer acceleration**

Use an accelerated endpoint for faster data transfers. [Learn more](#)

Transfer acceleration: Disabled

**Object Lock**

Store objects using a write-once-read-many (WORM) model to help you prevent objects from being deleted or overwritten for a fixed amount of time or indefinitely. [Learn more](#)

Object Lock: Disabled

Amazon S3 currently does not support enabling Object Lock after a bucket has been created. To enable Object Lock for this bucket, contact [Customer Support](#).

**Requester pays**

When enabled, the requester pays for requests and data transfer costs, and anonymous access to this bucket is disabled. [Learn more](#)

Requester pays: Disabled

**Static website hosting**

Use this bucket to host a website or redirect requests. [Learn more](#)

Static website hosting: Disabled

## Operationalizing an AWS ML Project

### Training Job (in Progress):

The screenshot shows the Amazon SageMaker Training jobs console. The top navigation bar includes the AWS logo, Services menu, a search bar, and user information for 'voclabs/user2074122=5828682428 @ 9338-4504-5900' in the N. Virginia region. Below the navigation bar, the 'Amazon SageMaker' section is active, and the 'Training jobs' page is displayed. The page features a 'Training jobs' header with an 'Info' link, a refresh button, an 'Actions' dropdown, and a 'Create training job' button. A search bar for training jobs is also present. The main table lists training jobs with columns: Name, Creation time, Duration, Job status, Warm pool status, and Time left. The first job, 'pytorch-training-230206-1010-001-9d2eb29a', is in the 'InProgress' state. The other three jobs are 'Completed'. The footer includes links for Feedback, Language, Privacy, Terms, and Cookie preferences, along with the copyright notice '© 2023, Amazon Web Services, Inc. or its affiliates.'

Name	Creation time	Duration	Job status	Warm pool status	Time left
pytorch-training-230206-1010-001-9d2eb29a	Feb 06, 2023 10:10 UTC	-	InProgress	-	-
pytorch-training-2023-02-02-12-55-20-920	Feb 02, 2023 12:55 UTC	11 minutes	Completed	-	-
pytorch-training-230202-1014-003-c9793356	Feb 02, 2023 10:56 UTC	28 minutes	Completed	Terminated	-
pytorch-training-230202-1014-002-e082004c	Feb 02, 2023 10:14 UTC	39 minutes	Completed	Reused	-

### Training Job (Completed):

The screenshot shows the Amazon SageMaker Training jobs console with a list of completed training jobs. The interface is similar to the previous screenshot, but the 'Job status' column now shows 'Completed' for all jobs. The table lists six training jobs, including the one that was previously in progress. The footer remains the same, with links for Feedback, Language, Privacy, Terms, and Cookie preferences, and the copyright notice '© 2023, Amazon Web Services, Inc. or its affiliates.'

Name	Last updated	Creation time	Duration	Job status	Warm pool status	Time left
pytorch-training-230202-1014-002-e082004c	Feb 02, 2023 10:53 UTC	Feb 02, 2023 10:14 UTC	39 minutes	Completed	Reused	-
pytorch-training-230202-1014-001-6fdc85f6	Feb 02, 2023 11:01 UTC	Feb 02, 2023 10:14 UTC	an hour	Completed	Terminated	-
pytorch-training-230202-1014-003-c9793356	Feb 02, 2023 11:24 UTC	Feb 02, 2023 10:56 UTC	28 minutes	Completed	Terminated	-
pytorch-training-2023-02-02-12-55-20-920	Feb 02, 2023 13:06 UTC	Feb 02, 2023 12:55 UTC	11 minutes	Completed	-	-
pytorch-training-230206-1010-001-9d2eb29a	Feb 06, 2023 10:30 UTC	Feb 06, 2023 10:10 UTC	20 minutes	Completed	Reused	-
pytorch-training-230206-1010-003-27090636	Feb 06, 2023 10:53 UTC	Feb 06, 2023 10:34 UTC	19 minutes	Completed	Terminated	-

## Operationalizing an AWS ML Project

### Training Job (Multi-Instance):

The screenshot shows the Amazon SageMaker console's 'Training jobs' page. At the top, there's a search bar and a 'Create training job' button. Below is a table listing training jobs. Two jobs are visible, both in a 'Completed' state. The first job is 'dog-pytorch-2023-02-06-11-30-12-708' and the second is 'pytorch-training-230206-1010-003-27090636'. Both were created on Feb 06, 2023, and took 19 minutes to complete. The second job's status is 'Completed' with a 'Terminated' warm pool status.

Name	Last updated	Creation time	Duration	Job status	Warm pool status	Time left
dog-pytorch-2023-02-06-11-30-12-708	Feb 06, 2023 11:49 UTC	Feb 06, 2023 11:30 UTC	19 minutes	Completed	-	-
pytorch-training-230206-1010-003-27090636	Feb 06, 2023 10:53 UTC	Feb 06, 2023 10:34 UTC	19 minutes	Completed	Terminated	-

### Endpoint deployment:

The screenshot shows a Jupyter Notebook cell with a code snippet for deploying a model to an endpoint. The code is: `predictor = pytorch_model.deploy(initial_instance_count=1, instance_type='ml.m5.large')`. Below the code, the execution logs are visible, showing the creation of the endpoint configuration and the endpoint itself. A red arrow points to the log line: `INFO:sagemaker:Creating endpoint with name pytorch-inference-2023-02-06-11-55-06-147`.

```
[24]: predictor = pytorch_model.deploy(initial_instance_count=1, instance_type='ml.m5.large')
```

INFO:botocore.credentials:Found credentials from IAM Role: BaseNotebookInstanceEc2InstanceRole  
INFO:sagemaker:Creating model with name: pytorch-inference-2023-02-06-11-55-05-593  
INFO:sagemaker:Creating endpoint-config with name pytorch-inference-2023-02-06-11-55-06-147  
INFO:sagemaker:Creating endpoint with name pytorch-inference-2023-02-06-11-55-06-147  
-----

The screenshot shows the Amazon SageMaker console's 'Endpoints' page. At the top, there's a search bar and buttons for 'Update endpoint' and 'Create endpoint'. Below is a table listing endpoints. One endpoint is visible, 'pytorch-inference-2023-02-06-11-55-06-147', which is in an 'InService' state. It was created on Feb 06, 2023, at 11:55 UTC.

Name	ARN	Creation time	Status	Last updated
pytorch-inference-2023-02-06-11-55-06-147	arn:aws:sagemaker:us-east-1:933845045900:endpoint/pytorch-inference-2023-02-06-11-55-06-147	Feb 06, 2023 11:55 UTC	InService	Feb 06, 2023 11:57 UTC

# Operationalizing an AWS ML Project

## Step 2: AWS EC2 Workspace

**Write about the EC2 you created, write a justification of why you chose the instance type.**

- Storage Instance [m5.xlarge] with 4 vCPU and 16 Memory (GiB).
- EBS-Only Instance Storage(GB) with Up to 10 Network Bandwidth (Gbps) and Up to 4,750 EBS Bandwidth (Mbps).
- Suited for computer vision classification workloads, but must be careful as the EBS storage cost increase with time instance is on.
- With [Deep Learning AMI GPU PyTorch 1.13.1 (Amazon Linux 2) 20230201]

The screenshot shows the AWS Management Console 'Instances' page. The instance 'Operationalizing-AWS-ML-EC2' (i-05414b70d1d850c0d) is in a 'Running' state. The instance type is 'm5.xlarge'. The console shows various details including IP addresses, DNS names, and VPC ID.

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone
Operationalizing-AWS-ML-EC2	i-05414b70d1d850c0d	Running	m5.xlarge	Initializing	No alarms	us-east-1b

**Instance: i-05414b70d1d850c0d (Operationalizing-AWS-ML-EC2)**

**Instance summary**

Instance ID	Public IPv4 address	Private IPv4 addresses
i-05414b70d1d850c0d (Operationalizing-AWS-ML-EC2)	34.229.129.241   <a href="#">open address</a>	172.31.16.4

**Instance state**

Running

**Private IP DNS name (IPv4 only)**

ip-172-31-16-4.ec2.internal

**Instance type**

m5.xlarge

**VPC ID**

vpc-003a6539f44a92cd2

**Subnet ID**

subnet-0c4a3d6b8f4c40f70

The screenshot shows the AWS Management Console 'Instance details' page for 'Operationalizing-AWS-ML-EC2'. The page displays various configuration details including IAM Role, Subnet ID, and Auto Scaling Group name.

Instance ID	Public IPv4 address	Private IPv4 addresses
i-05414b70d1d850c0d (Operationalizing-AWS-ML-EC2)	34.229.129.241   <a href="#">open address</a>	172.31.16.4

**Instance state**

Running

**Private IP DNS name (IPv4 only)**

ip-172-31-16-4.ec2.internal

**Instance type**

m5.xlarge

**VPC ID**

vpc-003a6539f44a92cd2

**Subnet ID**

subnet-0c4a3d6b8f4c40f70

**IAM Role**

-

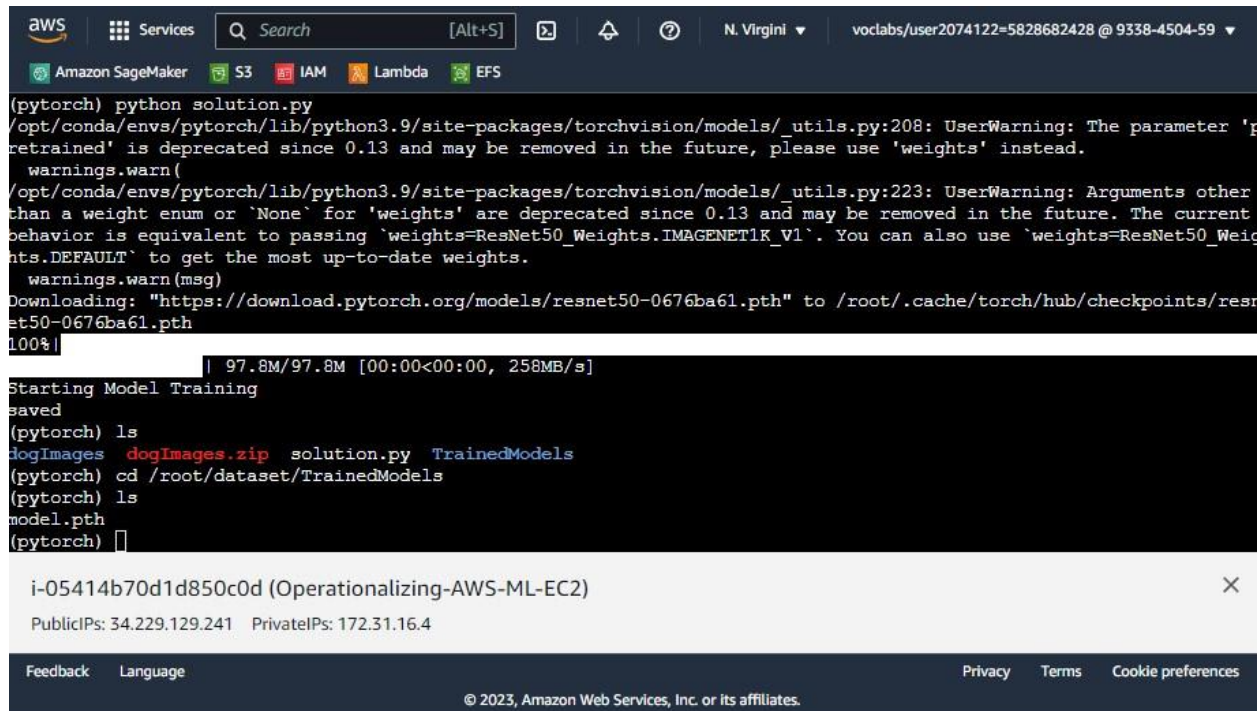
**Auto Scaling Group name**

-



Write at least one paragraph about the differences between the code in *ec2train1.py* and the code you used in Step 1.

- it is self-containing script which trains the model on any machine using local file system to access the data, while Sagemaker scripts required methods to access S3 dataset. Using EC2 instances is similar to usage of local machine.
  - storage for data is on local path is different (the training job is local)
  - model is also stored in a folder locally.
  - computation infrastructure used is local with not submission to another instance environment for execution.



```
aws
Services
Search [Alt+S]
N. Virgini
voclabs/user2074122-5828682428 @ 9338-4504-59
Amazon SageMaker S3 IAM Lambda EFS

(pytorch) python solution.py
/opt/conda/envs/pytorch/lib/python3.9/site-packages/torchvision/models/_utils.py:208: UserWarning: The parameter 'pretrained' is deprecated since 0.13 and may be removed in the future, please use 'weights' instead.
  warnings.warn(
/opt/conda/envs/pytorch/lib/python3.9/site-packages/torchvision/models/_utils.py:223: UserWarning: Arguments other than a weight enum or 'None' for 'weights' are deprecated since 0.13 and may be removed in the future. The current behavior is equivalent to passing 'weights=ResNet50_Weights.IMAGENET1K_V1'. You can also use 'weights=ResNet50_Weights.DEFAULT' to get the most up-to-date weights.
  warnings.warn(msg)
Downloading: "https://download.pytorch.org/models/resnet50-0676ba61.pth" to /root/.cache/torch/hub/checkpoints/resnet50-0676ba61.pth
100% |
| 97.8M/97.8M [00:00<00:00, 258MB/s]
Starting Model Training
saved
(pytorch) ls
dogImages dogImages.zip solution.py TrainedModels
(pytorch) cd /root/dataset/TrainedModels
(pytorch) ls
model.pth
(pytorch) []

i-05414b70d1d850c0d (Operationalizing-AWS-ML-EC2)
PublicIPs: 34.229.129.241 PrivateIPs: 172.31.16.4

Feedback Language
© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences
```

## Step 3: Lambda Function

### Describe how the function is written and how it works

- This lambda functions invokes the deployed endpoint created from the first step. 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.

The screenshot displays the AWS Lambda console interface. At the top, there's a navigation bar with 'Services', 'Search', and a list of services including Amazon SageMaker, S3, IAM, Lambda, and EFS. Below this, a blue banner indicates 'Created provisioned concurrency configuration. Allocating provisioned concurrency can take a few minutes.'

The main content area shows the 'MLOps-Operationalizing-AWS-ML-Project:2' function. It includes a 'Function overview' section with a description: 'increased concurrency with additional price', 'Last modified 49 seconds ago', and the 'Function ARN: arn:aws:lambda:us-east-1:933845045900:function:MLOps-Operationalizing-AWS-ML-Project:2'. There are buttons for 'Add trigger' and 'Add destination'.

Below the overview, the 'Execution result: succeeded' is shown. It includes a 'Details' section with a log snippet:
 

```
{
  "statusCode": 200,
  "headers": {
    "Content-Type": "text/plain",
    "Access-Control-Allow-Origin": "*"
  },
  "type-result": "<class 'str'>",
  "Content-Type-In": "<__main__.LambdaContext object at 0x7f77236a7ac0>",
  "body": "[[-7.698862552642822, -5.347639083862305, -3.352297782897949, -1.4173848628997803, -3.677257537841797, -4.9470038414001465, -4.244888782501221, -1.391355276107788, -5.949262619018555, -1.240530252456665, -0.42824578285217285, -3.1313931941986084, -1.9111168384552002, 0.8243446946144104, -2.878235101699829, ...]]"
```

 A 'Summary' section provides metadata: Code SHA-256, Request ID, Duration (897.56 ms), Billed duration (898 ms), Resources configured (128 MB), and Max memory used (71 MB).

The 'Log output' section shows the start and end of the request, including the event name 'SampleTest' and the event JSON:
 

```
{
  "url": "https://s3.amazonaws.com/cdn-origin-etr.akc.org/wp-content/uploads/2017/11/20113314/Carolina-Dog-standing-outdoors.jpg"
}
```

At the bottom, the 'Test event' section allows for testing the function. It includes a 'Test event action' dropdown (set to 'Edit saved event'), an 'Event name' field (set to 'SampleTest'), and a 'Format JSON' button. The event JSON is displayed in a text area.

# Operationalizing an AWS ML Project

Attach a security policy to your Lambda function so it can access your Sagemaker endpoint.

The screenshot displays the AWS IAM console interface. The left-hand navigation pane shows the 'Identity and Access Management (IAM)' section, with 'Roles' selected. The main content area shows the details for the role 'MLOps-Operationalizing-AWS-ML-Project-role-tu0jvt4'. The 'Summary' tab is active, displaying the role's creation date (February 07, 2023) and its ARN. Below this, the 'Permissions' tab is selected, showing a list of attached permissions policies. The list includes 'AWSLambdaBasicExecutionRole-5415aa71-d66a-4353-a676-4f39bbfc...', 'AmazonEC2FullAccess', 'AmazonS3FullAccess', and 'AmazonSageMakerFullAccess'. The 'Permissions boundary' section indicates that no boundary is currently set. At the bottom, there is a 'Generate policy based on CloudTrail events' section with a 'Generate policy' button.

**Identity and Access Management (IAM)**

Unable to load search  
Dashboard

▼ Access management  
User groups  
Users  
**Roles**  
Policies  
Identity providers  
Account settings

▼ Access reports  
Access analyzer  
Archive rules  
Analyzers  
Settings  
Credential report  
Organization activity  
Service control policies (SCPs)

**MLOps-Operationalizing-AWS-ML-Project-role-tu0jvt4** Delete

**Summary** Edit

Creation date  
February 07, 2023, 11:22 (UTC+02:00)

ARN  
arn:aws:iam::933845045900:role/service-role/MLOps-Operationalizing-AWS-ML-Project-role-tu0jvt4

Last activity  
25 minutes ago

Maximum session duration  
1 hour

**Permissions** | Trust relationships | Tags | Access Advisor | Revoke sessions

**Permissions policies (5)** Info Refresh Simulate Remove Add permissions

You can attach up to 10 managed policies.

Filter policies by property or policy name and press enter.

<input type="checkbox"/>	Policy name	Type	Description
<input type="checkbox"/>	<a href="#">AWSLambdaBasicExecutionRole-5415aa71-d66a-4353-a676-4f39bbfc...</a>	Customer managed	
<input type="checkbox"/>	<a href="#">AmazonEC2FullAccess</a>	AWS managed	Provides full access to Amazon EC2 vi...
<input type="checkbox"/>	<a href="#">AmazonS3FullAccess</a>	AWS managed	Provides full access to all buckets via t...
<input type="checkbox"/>	<a href="#">AmazonSageMakerFullAccess</a>	AWS managed	Provides full access to Amazon SageM...

**Permissions boundary - (not set)** Info

Set a permissions boundary to control the maximum permissions this role can have. This is not a common setting but can be used to delegate permission management to others.

Set permissions boundary

▼ Generate policy based on CloudTrail events

You can generate a new policy based on the access activity for this role, then customize, create, and attach it to this role. AWS uses your CloudTrail events to identify the services and actions used and generate a policy. [Learn more](#)

Generate policy

No requests to generate a policy in the past 7 days.

Feedback Language

© 2023, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)



Test your Lambda function, and add the result of your test to your writeup

Test Response Result:

7.698862552642822, -5.347639083862305, -3.352297782897949, -1.4173848628997803, -  
3.677257537841797, -4.9470038414001465, -4.244888782501221, -1.391355276107788, -  
5.949262619018555, -1.240530252456665, -0.42824578285217285, -3.1313931941986084, -  
1.9111168384552002, 0.8243446946144104, -2.878235101699829, -4.649499416351318, -  
6.185570240020752, -4.9580864906311035, -4.424695014953613, 0.41801193356513977, -  
2.9353291988372803, -2.7232508659362793, -7.775493621826172, -7.063922882080078, -  
7.322050094604492, -7.940590858459473, -4.056085109710693, -7.363180160522461, -  
5.64165735244751, -3.111562728881836, -4.576196670532227, -5.910926818847656, -  
6.551362037658691, -5.1530656814575195, -7.405170917510986, -7.548816204071045, -  
6.143508434295654, -5.465755939483643, -1.5719480514526367, -5.441058158874512, -  
6.1667304039001465, -3.9466495513916016, 0.4702502191066742, -4.881587028503418, -  
1.9118139743804932, -12.494391441345215, -3.2319061756134033, -1.2295335531234741, -  
4.9419426918029785, -2.914275884628296, -4.3234686851501465, -7.038097381591797, -  
5.984295845031738, -2.81803035736084, -6.6579909324646, -3.4884591102600098, -  
8.707923889160156, -7.949629783630371, -3.3004395961761475, -3.240016460418701, -  
5.740108013153076, -7.424511432647705, -7.692864894866943, -7.383284568786621, -  
5.518505096435547, -6.879883289337158, 0.0034765787422657013, -7.727499485015869, -  
4.429338455200195, -2.8595452308654785, -1.6434866189956665, -4.072175025939941, -  
4.924370288848877, -5.589024066925049, -6.112706661224365, -2.314589023590088, -  
7.842854022979736, -3.007693290710449, -6.61555290222168, -6.093110084533691, -  
2.6826772689819336, -7.160601615905762, 0.6826741695404053, -1.168666958808899, -  
7.320164203643799, -5.28354549407959, -2.049327850341797, -7.931869983673096, -  
3.4631261825561523, -1.7312875986099243, -7.897009372711182, -6.111726760864258, -  
5.259680271148682, -8.255675315856934, -6.626457214355469, -3.025027275085449, -  
4.766608715057373, -3.9160499572753906, -6.909079074859619, -5.580389022827148, -  
8.45504379272461, -3.9216482639312744, -4.523221492767334, -4.135554313659668, -  
5.4708075523376465, -8.630441665649414, -3.0037379264831543, -1.521710991859436, -  
2.8514323234558105, -1.0108898878097534, -2.2446963787078857, -2.4315717220306396, -  
4.832242488861084, -6.192027568817139, -7.673048496246338, -2.0159521102905273, -  
8.957904815673828, -0.31093719601631165, -4.665895462036133, 0.1316596418619156, -  
0.8865324854850769, -4.975542068481445, -3.7051947116851807, -4.13137149810791, -  
8.761499404907227, -7.0732550621032715, -1.6587605476379395, -0.0707918107509613, -  
5.985174655914307, -5.818277359008789, -5.7232465744018555, -1.8361643552780151, -  
6.063624858856201]]

## Operationalizing an AWS ML Project

Take a screenshot of your Lambda setup and add it to your solution archive.

The screenshot displays the AWS Lambda console interface. At the top, there's a navigation bar with the AWS logo, a search bar, and a user profile. The main header shows the path 'Lambda > Functions > MLOps-Operationalizing-AWS-ML-Project'. Below this, the function name 'MLOps-Operationalizing-AWS-ML-Project' is prominently displayed. To the right of the name are buttons for 'Throttle', 'Copy ARN', and 'Actions'. The 'Function overview' section shows a card for the function with a description, last modified time (2 hours ago), and function ARN. Below this, there are tabs for 'Code', 'Test', 'Monitor', 'Configuration', 'Aliases', and 'Versions'. The 'Code source' tab is active, showing a code editor with a file named 'lambda\_function.py'. The code is a Python script that uses boto3 to interact with the SageMaker runtime. It defines a 'lambda\_handler' function that takes an event and context as input, encodes the event content, and then calls the SageMaker runtime endpoint to invoke the model. The code also includes logging and error handling. At the bottom, the 'Code properties' section shows the package size (700.0 byte), SHA256 hash, and last modified date (February 7, 2023 at 12:25 PM GMT+2).

**Write about whether you think your AWS workspace is secure, and whether you think there are any vulnerabilities.**

- insecure since there are vulnerabilities that the endpoint is exposed to from external request calls. Granting full access to lambda function can result unexpected errors. So it should be monitored or limited.

# Operationalizing an AWS ML Project

## Step 5: Concurrency and auto-scaling

### Set up concurrency for your Lambda function

#### Version:

The screenshot shows the AWS Lambda console interface. At the top, there's a navigation bar with the AWS logo, a search bar, and a list of services (Amazon SageMaker, S3, IAM, Lambda, EFS). Below the navigation bar, the breadcrumb trail reads 'Lambda > Functions > MLOps-Operationalizing-AWS-ML-Project'. The main heading is 'MLOps-Operationalizing-AWS-ML-Project'. To the right of the heading are buttons for 'Throttle', 'Copy ARN', and 'Actions'. Below the heading is a 'Function overview' section with a card for 'MLOps-Operationalizing-AWS-ML-Project' showing 'Layers (0)' and buttons for '+ Add trigger' and '+ Add destination'. To the right of the card, there's a 'Description' section with fields for 'Last modified' (1 minute ago), 'Function ARN' (arn:aws:lambda:us-east-1:933845045900:function:MLOps-Operationalizing-AWS-ML-Project), and 'Function URL' (Info). Below the overview is a 'Versions' tab with a 'Delete' button and a 'Publish new version' button. A search bar 'Find versions' is present. The table below shows two versions:

Version	Aliases	Description	Last modified	Architecture
2	-	increased concurrency with additional price	1 minute ago	x86_64
1	-	this version is the start point version with basic concurrency	1 hour ago	x86_64

#### Concurrency:

The screenshot shows the AWS Lambda console interface, specifically the 'Configuration' tab for the function 'MLOps-Operationalizing-AWS-ML-Project'. The left sidebar contains a list of configuration options: General configuration, Triggers, Permissions, Destinations, Function URL, Environment variables, VPC, Monitoring and operations tools, Provisioned concurrency (selected), Asynchronous invocation, Database proxies, File systems, and State machines. The main content area is titled 'Provisioned concurrency' and has buttons for 'Edit' and 'Remove'. Below the title, there's a section for 'Provisioned concurrency' showing a value of '0' and a 'Status' section showing 'In progress (0/2)'.

# Operationalizing an AWS ML Project

## Set up auto-scaling for your deployed endpoint

Automatic scaling was configured for variant AllTraffic

pytorch-inference-2023-02-06-11-55-06-147 [Delete](#)

### Endpoint settings

Name pytorch-inference-2023-02-06-11-55-06-147	Status InService	Type Real-time	URL <a href="https://runtime.sagemaker.us-east-1.amazonaws.com/endpoint/pytorch-inference-2023-02-06-11-55-06-147/invocations">https://runtime.sagemaker.us-east-1.amazonaws.com/endpoint/pytorch-inference-2023-02-06-11-55-06-147/invocations</a> <a href="#">Learn more about the API</a>
ARN arn:aws:sagemaker:us-east-1:933845045900:endpoint/pytorch-inference-2023-02-06-11-55-06-147	Creation time Mon Feb 06 2023 13:55:06 GMT+0200 (Eastern European Standard Time)	Last updated Mon Feb 06 2023 13:57:21 GMT+0200 (Eastern European Standard Time)	

### Data capture settings

Enable data capture No	Current sampling percentage (%) -	S3 location to store data collected -
Data capture status -		

### Monitor

Access CloudWatch logs to view your Jupyter notebook's debugging and progress reporting. [Learn more](#)

[View invocation metrics](#) [View instance metrics](#) [View logs](#)

1h 3h 12h 1d 3d 1w [Add to dashboard](#)

No widget on this dashboard.

### Endpoint runtime settings

[Update weights](#) [Update instance count](#) [Configure auto scaling](#)

	Variant name	Current weight	Desired weight	Elastic inference	Instance type	Current instance count	Desired instance count	Instance min - max	Automatic scaling
<input type="radio"/>	AllTraffic	1	1	-	ml.m5.large	1	1	1 - 2	Yes

### Endpoint configuration settings

[Change](#) [Clone](#)

#### Endpoint configuration

Name pytorch-inference-2023-02-06-11-55-06-147	ARN arn:aws:sagemaker:us-east-1:933845045900:endpoint-config/pytorch-inference-2023-02-06-11-55-06-147	Encryption key -	Creation time Feb 06, 2023 11:55 UTC
---	---	---------------------	---

#### Data capture

Enable data capture No	Data capture options -	S3 location to store data collected -	Capture content type -
Sampling percentage (%) -			

#### Variants

Identifies a model that you want to host and the resources chosen to deploy for hosting it.

#### Production

Model name	Training job	Variant name	Instance type	Elastic inference	Initial instance count	Initial weight
pytorch-inference-2023-02-06-11-55-05-593	-	AllTraffic	ml.m5.large	-	1	1

#### Shadow

Model name	Training job	Variant name	Instance type	Elastic inference	Initial instance count	Initial weight
There are currently no resources.						

**Write about your configuration of concurrency and auto-scaling. Make sure to mention what kind of concurrency you set up, how you set up auto-scaling, and why you made your decisions.**

- **Concurrency:** decided based on assumed usage that there are only two request which may get overlapped and extra request can be queried.
- **Autoscaling:** was turned on for scaling to the maximum of 2 instance which should typically start during the peak traffic (30 simultaneous requests) for min of 30 seconds and cool down in 30 seconds.