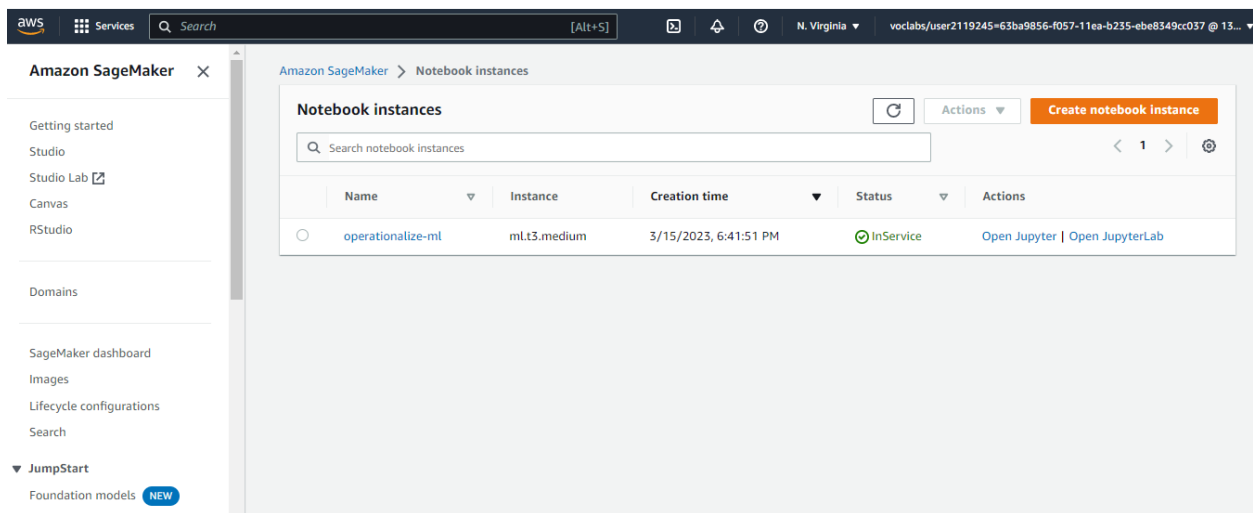


Operationalize AWS ML Project

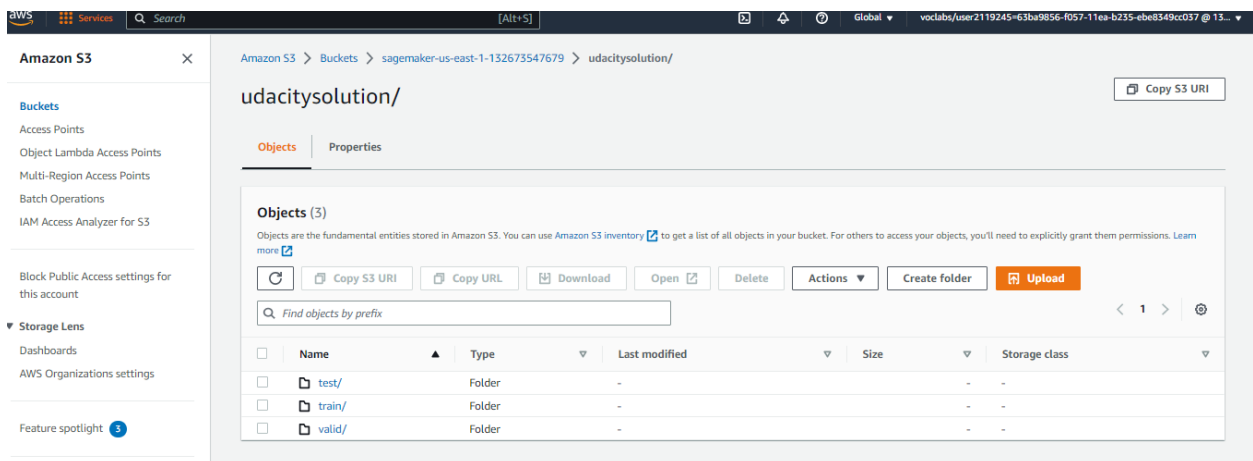
SageMaker Notebook instance:

I have chosen an ml.t3.medium instance because it has sufficient computing power (vCPU =2, Memory = 4 GiB) and it's cheap (Price per Hour = \$0.05). So, it does the job while maintaining a low cost.

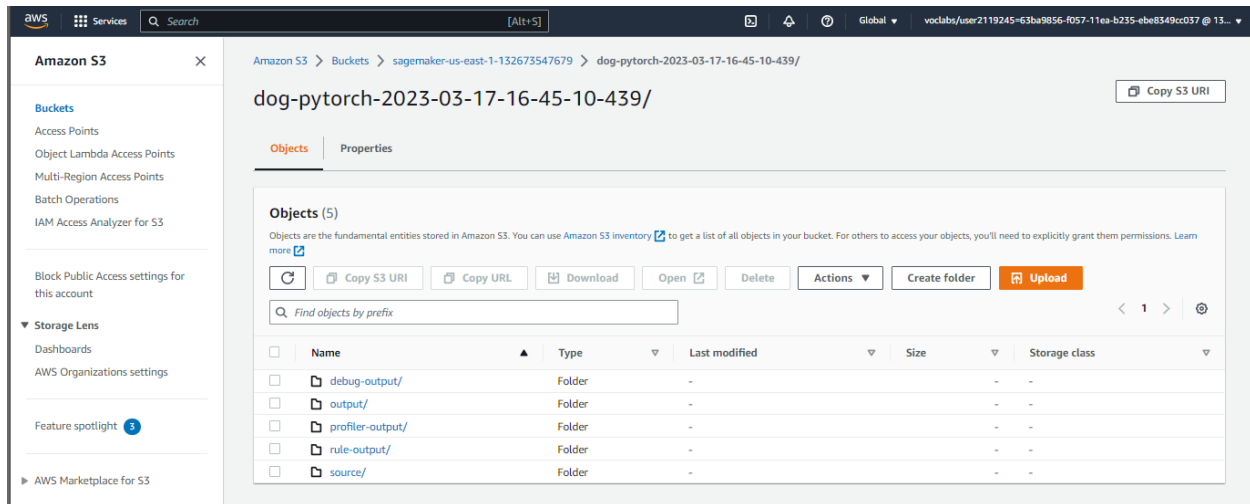


S3 Bucket:

The Training Dataset



The Output of Training



Amazon S3 > Buckets > sagemaker-us-east-1-132673547679 > dog-pytorch-2023-03-17-16-45-10-439/

dog-pytorch-2023-03-17-16-45-10-439/ Copy S3 URI

Objects | Properties

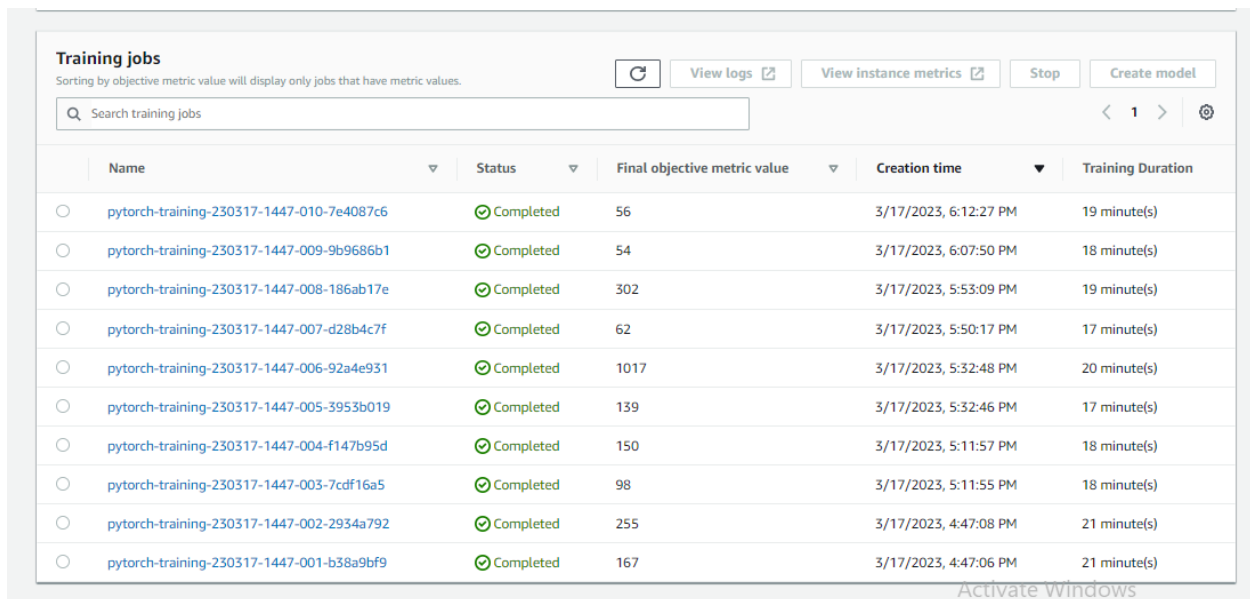
Objects (5)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Refresh Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	debug-output/	Folder	-	-	-
<input type="checkbox"/>	output/	Folder	-	-	-
<input type="checkbox"/>	profiler-output/	Folder	-	-	-
<input type="checkbox"/>	rule-output/	Folder	-	-	-
<input type="checkbox"/>	source/	Folder	-	-	-

Training and Deployment: Hyperparameter Training Jobs



Training jobs

Sorting by objective metric value will display only jobs that have metric values. Refresh View logs View instance metrics Stop Create model

<input type="radio"/>	Name	Status	Final objective metric value	Creation time	Training Duration
<input type="radio"/>	pytorch-training-230317-1447-010-7e4087c6	Completed	56	3/17/2023, 6:12:27 PM	19 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-009-9b9686b1	Completed	54	3/17/2023, 6:07:50 PM	18 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-008-186ab17e	Completed	302	3/17/2023, 5:53:09 PM	19 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-007-d28b4c7f	Completed	62	3/17/2023, 5:50:17 PM	17 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-006-92a4e931	Completed	1017	3/17/2023, 5:32:48 PM	20 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-005-3953b019	Completed	139	3/17/2023, 5:32:46 PM	17 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-004-f147b95d	Completed	150	3/17/2023, 5:11:57 PM	18 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-003-7cdf16a5	Completed	98	3/17/2023, 5:11:55 PM	18 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-002-2934a792	Completed	255	3/17/2023, 4:47:08 PM	21 minute(s)
<input type="radio"/>	pytorch-training-230317-1447-001-b38a9bf9	Completed	167	3/17/2023, 4:47:06 PM	21 minute(s)

For the training I started training using one instance and another time using multi-instance.

- One instance

The screenshot displays the Amazon SageMaker console interface. The left sidebar shows the navigation menu with categories like 'Getting started', 'Domains', 'SageMaker dashboard', and 'Training'. The main content area shows the details for a specific training job.

Job settings

Job name dog-pytorch-2023-03-17-16-45-10-439	Status ✔ Completed View history	SageMaker metrics time series Enabled	IAM role ARN arn:aws:iam::132673547679:role/service-role/AmazonSageMaker-ExecutionRole-20230309T200693
ARN arn:aws:sagemaker:us-east-1:132673547679:training-job/dog-pytorch-2023-03-17-16-45-10-439	Creation time Mar 17, 2023 16:45 UTC	Training time (seconds) 1111	
	Last modified time Mar 17, 2023 17:05 UTC	Billable time (seconds) 1111	
		Managed spot training savings 0%	
		Tuning job source/parent -	

Algorithm

Algorithm ARN -	Additional volume size (GB) 30	Maximum wait time for managed spot training(s) -	Volume encryption key -
Training image 763104351884.dkr.ecr.us-east-1.amazonaws.com/pytorch-training:1.4.0-cpu-py3	Maximum runtime (s) 86400	Managed spot training Disabled	
Input mode File			

Instance group	Instance type	Instance count	Keep alive period
-	ml.m5.xlarge	1	-

- Multi instances

The screenshot shows the Amazon SageMaker console interface. On the left is a navigation sidebar with options like 'Getting started', 'Studio', 'Domains', 'SageMaker dashboard', 'Images', 'Lifecycle configurations', 'Search', 'JumpStart', 'Foundation models', 'Computer vision models', 'Natural language processing models', 'Governance', 'Ground Truth', 'Notebook', 'Processing', 'Training', 'Algorithms', 'Training jobs', 'Hyperparameter tuning jobs', and 'Inference'. The main panel displays the details for a specific training job: 'dog-pytorch-2023-03-17-17-06-20-335'. At the top right of the job details are buttons for 'Clone', 'Create model package', 'Stop', and 'Create model'. Below this is the 'Job settings' section, which includes fields for Job name, Status (Completed), SageMaker metrics time series (Enabled), IAM role ARN, ARN, Creation time (Mar 17, 2023 17:06 UTC), Last modified time (Mar 17, 2023 17:26 UTC), Training time (seconds) (1132), Billable time (seconds) (1132), Managed spot training savings (0%), and Tuning job source/parent (-). Below the job settings is the 'Algorithm' section, which includes fields for Algorithm ARN, Additional volume size (GB) (30), Maximum wait time for managed spot training(s) (-), Volume encryption key (-), Training image (763104351884.dkr.ecr.us-east-1.amazonaws.com/pytorch-training:1.4.0-cpu-py3), Maximum runtime (s) (86400), Managed spot training (Disabled), and Input mode (File). At the bottom is a table for 'Instance group' with columns: Instance group, Instance type, Instance count, and Keep alive period. The table shows one instance group with type 'ml.m5.xlarge', count '5', and keep alive period '-'.

Instance group	Instance type	Instance count	Keep alive period
-	ml.m5.xlarge	5	-

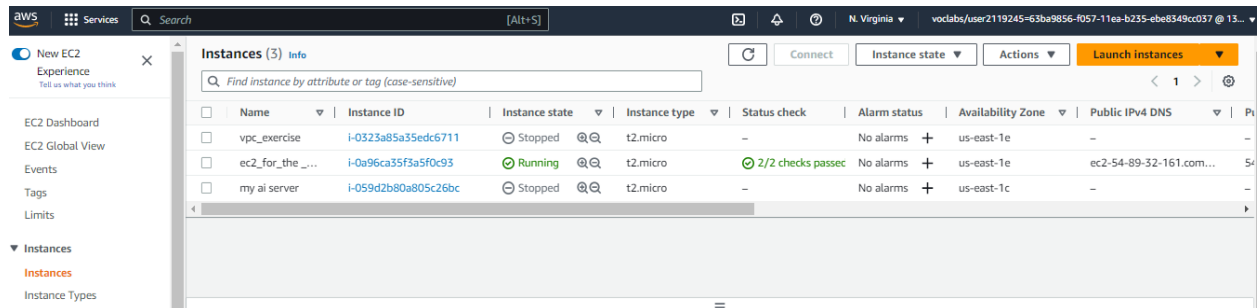
And here is the deployed end points (one using training with one instance and the other with multi-instance)

The screenshot shows the Amazon SageMaker console interface for the 'Endpoints' section. On the left is a navigation sidebar with options like 'Domains', 'SageMaker dashboard', 'Images', 'Lifecycle configurations', 'Search', 'JumpStart', 'Foundation models', 'Computer vision models', and 'Inference'. The main panel displays the 'Endpoints' section with buttons for 'Update endpoint', 'Actions', and 'Create endpoint'. Below this is a search bar and a table of endpoints. The table has columns: Name, ARN, Creation time, Status, and Last updated. There are two endpoints listed, both with status 'InService'.

Name	ARN	Creation time	Status	Last updated
pytorch-inference-2023-03-17-18-02-16-600	arn:aws:sagemaker:us-east-1:132673547679:endpoint/pytorch-inference-2023-03-17-18-02-16-600	3/17/2023, 8:02:17 PM	InService	3/17/2023, 8:04:32 PM
pytorch-inference-2023-03-17-17-57-12-629	arn:aws:sagemaker:us-east-1:132673547679:endpoint/pytorch-inference-2023-03-17-17-57-12-629	3/17/2023, 7:57:13 PM	InService	3/17/2023, 7:59:28 PM

EC2:

I have used a t2.micro instance because it's eligible for the free tier and provide the needed computation power.



Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
vpc_exercise	i-0323a85a35edc6711	Stopped	t2.micro	-	No alarms	us-east-1e	-
ec2_for_the_...	i-0a96ca35f3a5f0c93	Running	t2.micro	2/2 checks passed	No alarms	us-east-1e	ec2-54-89-32-161.com...
my ai server	i-059d2b80a805c26bc	Stopped	t2.micro	-	No alarms	us-east-1c	-

Difference between EC2 training code and SageMaker's:

- In EC2 code there is no calling for any Estimator or Tuner functions. The code in the EC2 script is responsible for saving the model to the local path. While in the SageMaker scripts this was handled internally by SageMaker where the model data was stored to a S3 location.
- In the EC2 code, all the variables already mentioned in the code itself.
- In the EC2 the training happens on the same while in the SageMaker the training job runs on a separate container than the one on which the SageMaker notebook is running.

Lambda function:

Lambda functions are used for invoking our deployed endpoints. And the endpoint used is the one of the mult-instance training (pytorch-inference-2023-03-17-18-02-16-600). Also we attached an Amazon SageMaker full Accesses policy to be able to interact with SageMaker successfully with no errors.

In the following images there are the role and result of the testing

The screenshot shows the AWS IAM console interface. On the left is a navigation menu with sections like 'Identity and Access Management (IAM)', 'Access management', 'Access reports', and 'Service control policies (SCPs)'. The main area displays the details for the role 'lambdafunction-role-j9rt5bvx'. The 'Summary' tab is active, showing the role's creation date (March 17, 2023, 21:02 UTC+02:00), its ARN, and its maximum session duration (1 hour). Below the summary, there are tabs for 'Permissions', 'Trust relationships', 'Tags', 'Access Advisor', and 'Revoke sessions'. The 'Permissions' tab is selected, showing a list of attached policies. Two policies are listed: 'AWSLambdaBasicExecutionRole-f79e3cd6-a3a6-49f0-a9e3-9d0749b71509' (Customer managed) and 'AmazonSageMakerFullAccess' (AWS managed). At the bottom right, there is a watermark for 'Activate Windows'.

The screenshot shows the AWS Lambda console interface. At the top, a green banner states 'The test event test was successfully saved.' Below this, the 'lambda_function' is selected, and the 'Execution results' tab is active. The 'Test Event Name' is 'test'. The 'Response' section shows a JSON object with a status code of 200, headers, and a body. The 'Function Logs' section shows the execution details, including the request ID, request time, and duration. The 'Request ID' is 'd08fec49-957d-4eba-b3cb7c05716b'.

It's not a good idea to give full access permission because it may result in a security breach, but always choose the right policies and permission and remove them when the task is done. Below is the policies attached to my SageMaker execution role.

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

Analzers

Settings

Credential report

Organization activity

Service control policies (SCPs)

Related consoles

IAM Identity Center

AWS Organizations

IAM > Roles > AmazonSageMaker-ExecutionRole-20230309T200693

AmazonSageMaker-ExecutionRole-20230309T200693

SageMaker execution role created from the SageMaker AWS Management Console.

Summary

Creation date

March 09, 2023, 20:06 (UTC+02:00)

ARN

arn:aws:iam::132673547679:role/service-role/AmazonSageMaker-ExecutionRole-20230309T200693

Last activity

30 minutes ago

Maximum session duration

1 hour

Permissions

Trust relationships

Tags

Access Advisor

Revoke sessions

Permissions policies (4)

You can attach up to 10 managed policies.

Filter policies by property or policy name and press enter.

1

	Policy name	Type	Description
<input type="checkbox"/>	AmazonSageMaker-ExecutionPolicy-20230309T200693	Customer managed	
<input type="checkbox"/>	AmazonS3FullAccess	AWS managed	Provides full access to all
<input type="checkbox"/>	AmazonSageMakerFullAccess	AWS managed	Provides full access to An
<input type="checkbox"/>	AmazonSageMakerCanvasFullAccess	AWS managed	Provides full access to An

Concurrency and Auto Scaling:

I have configured a provisioned concurrency after publishing a version for lambda. Also I have configured Auto Scaling to cope with the traffic requests.

The screenshot displays the AWS Lambda console interface for a function named 'lambdafunction'. The top navigation bar includes the AWS logo, 'Services', a search bar, and the user's account information. The main content area is divided into two sections: 'Function overview' and 'Configuration'.

Function overview: This section provides a high-level view of the function. It includes a 'Layers' section with '(0)' layers listed. On the right, there is a 'Description' section with the following details:

- Description: -
- Last modified: 24 minutes ago
- Function ARN: `arn:aws:lambda:us-east-1:132673547679:function:lambdafunction`
- Function URL: [Info](#)

Configuration: This section is further divided into several tabs: 'Code', 'Test', 'Monitor', 'Configuration' (selected), 'Aliases', and 'Versions'. The 'Configuration' tab is active, showing the following settings:

- General configuration:** Includes options for Triggers, Permissions, Destinations, Function URL, Environment variables, Tags, VPC, and Monitoring and operations tools.
- Concurrency:** This section allows you to manage the function's concurrency. It includes an 'Edit' button and the following settings:
 - Function concurrency: Use reserved concurrency
 - Reserved concurrency: 5
- Provisioned concurrency configurations (1):** This section shows the current provisioned concurrency configuration. It includes a search bar and a table of configurations.

Provisioned concurrency configurations table:

Qualifier	Type	Provisioned concurrency	Status	Details
1	version	1	Ready	-

aws

Services

Q lambda

X

N. Virginia

voclabs/user2119245-63ba9856-f057-11ea-b235-ebe8549cc037 @ 13...

▼ JumpStart

Foundation models NEW

Computer vision models

Natural language processing models

► Governance

► Ground Truth

▼ Notebook

Notebook instances

Git repositories

► Processing

▼ Training

Algorithms

Training jobs

Hyperparameter tuning jobs

▼ Inference

Compilation jobs

Marketplace model packages

Models

Endpoint configurations

Endpoints

Batch transform jobs

Shadow tests


► Edge Manager

► Augmented AI

► AWS Marketplace

Tutorials

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	-	mLm5.large	1	1	1 - 5	Yes

Endpoint configuration settings

Change

Clone

Endpoint configuration

Name	ARN	Encryption key	Creation time
pytorch-inference-2023-03-17-18-02-16-600	arn:aws:sagemaker:us-east-1:132673547679:endpoint-config/pytorch-inference-2023-03-17-18-02-16-600	-	3/17/2023, 8:02:16 PM

Data capture

Enable data capture	Data capture options	S3 location to store data collected	Capture content type
...			