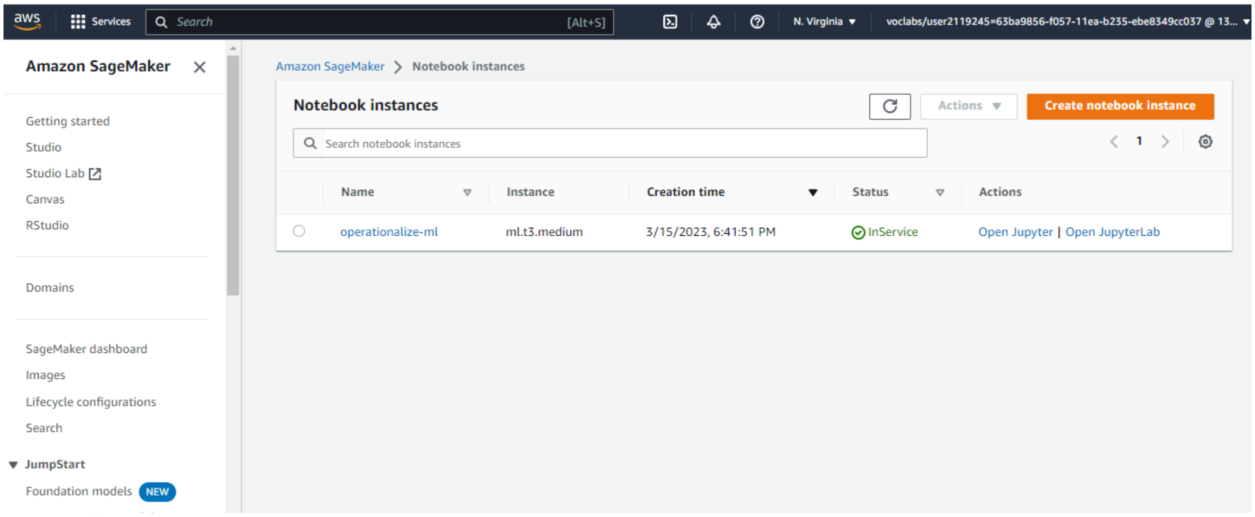
Operationalize AWS ML Project

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.

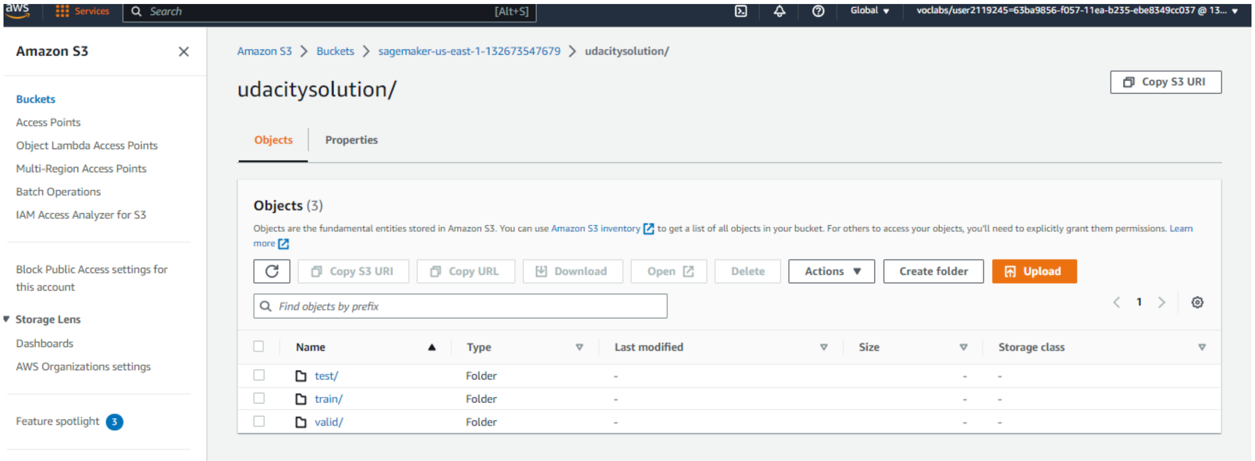


Notebook instances

	Name	Instance	Creation time	Status	Actions
○	operationalize-ml	ml.t3.medium	3/15/2023, 6:41:51 PM	InService	Open Jupyter | Open JupyterLab

S3 Bucket:

The Training Dataset


aws
Services
Search [Alt+S]
Global
voclabs/user2119245=63ba9856-f057-11ea-b235-ebe8349cc037 @ 13...
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
Amazon S3 > Buckets > sagemaker-us-east-1-132673547679 > udacitysolution/
Copy S3 URI
Objects
Properties
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1

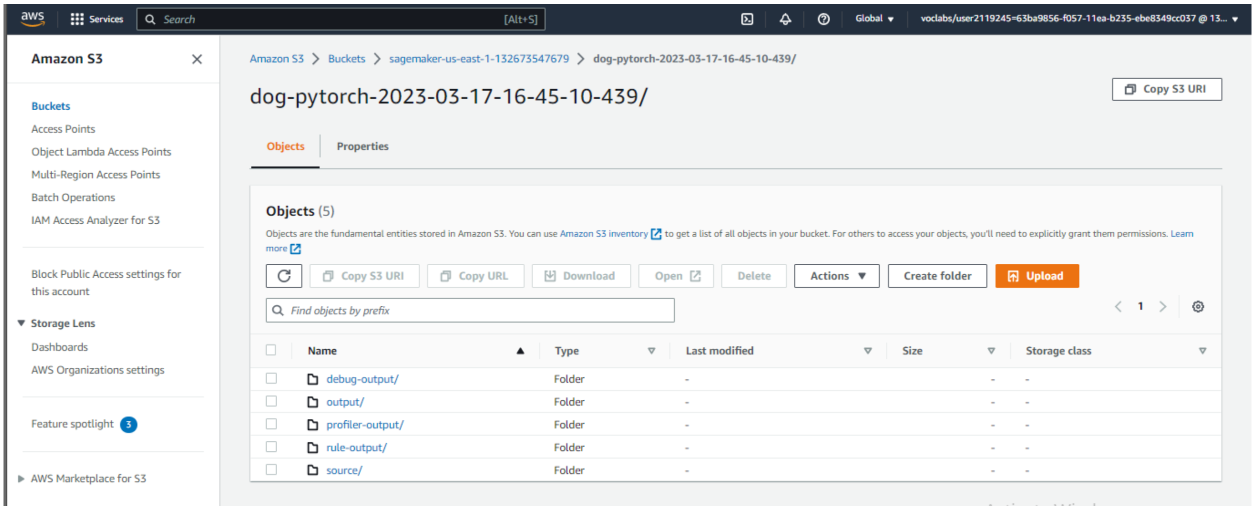
udacitysolution/

Objects (3)

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

□	Name	Type	Last modified	Size	Storage class
□	test/	Folder	-	-	-
□	train/	Folder	-	-	-
□	valid/	Folder	-	-	-

The Output of Training


aws
Services
Search
[Alt+S]
Global
voclabs/user2119245=63ba9856-f057-11ea-b235-ebe8349cc037 @ 13...
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
AWS Marketplace for S3
Amazon S3 > Buckets > sagemaker-us-east-1-132673547679 > dog-pytorch-2023-03-17-16-45-10-439/
Copy S3 URI
Objects
Properties
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1

dog-pytorch-2023-03-17-16-45-10-439/

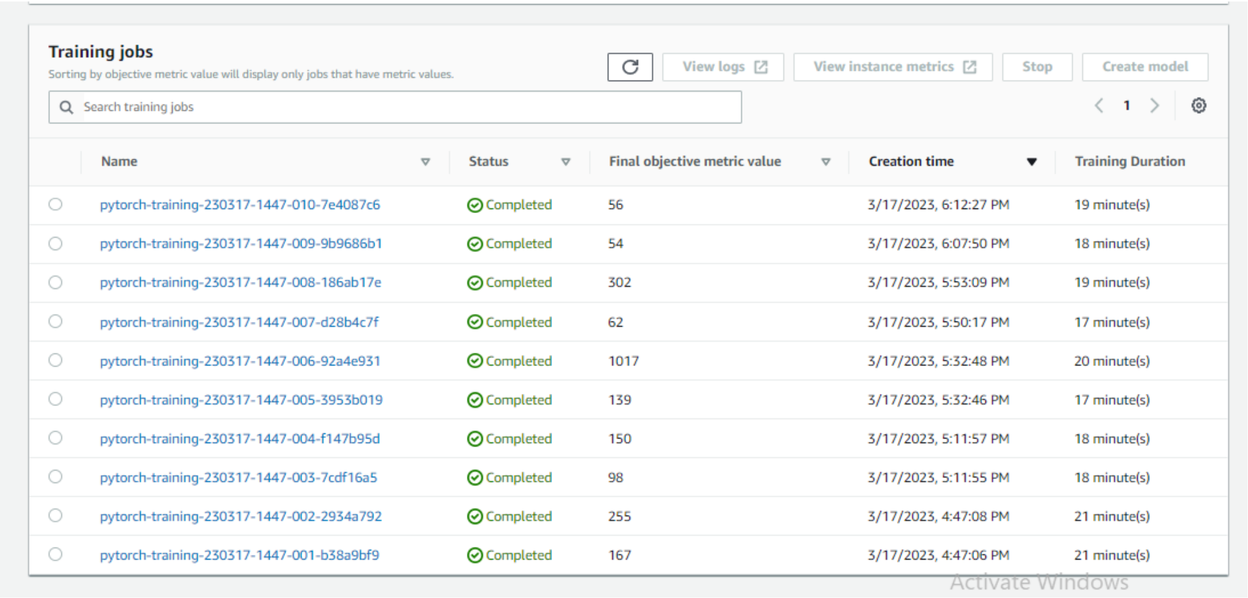
Objects (5)

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

	Name	Type	Last modified	Size	Storage class
□	debug-output/	Folder	-	-	-
□	output/	Folder	-	-	-
□	profiler-output/	Folder	-	-	-
□	rule-output/	Folder	-	-	-
□	source/	Folder	-	-	-

Training and Deployment:

Hyperparameter Training Jobs



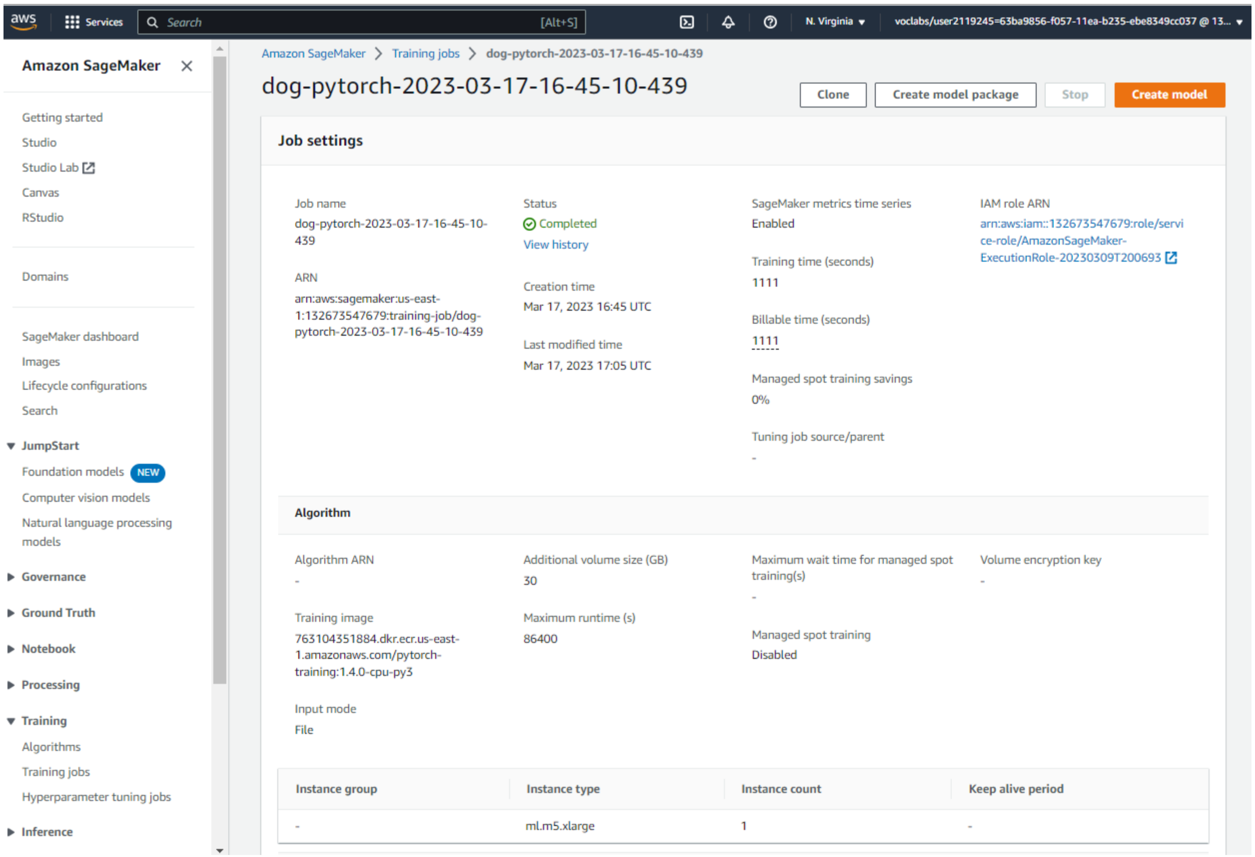
Training jobs

Sorting by objective metric value will display only jobs that have metric values.

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



dog-pytorch-2023-03-17-16-45-10-439

Job settings

Job name
dog-pytorch-2023-03-17-16-45-10-439

ARN
arn:aws:sagemaker:us-east-1:132673547679:training-job/dog-pytorch-2023-03-17-16-45-10-439

Status
Completed
View history

Creation time
Mar 17, 2023 16:45 UTC

Last modified time
Mar 17, 2023 17:05 UTC

SageMaker metrics time series
Enabled

Training time (seconds)
1111

Billable time (seconds)
1111

Managed spot training savings
0%

Tuning job source/parent
-

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

Algorithm

Algorithm ARN
-

Training image
763104351884.dkr.ecr.us-east-1.amazonaws.com/pytorch-training:1.4.0-cpu-py3

Input mode
File

Additional volume size (GB)
30

Maximum runtime (s)
86400

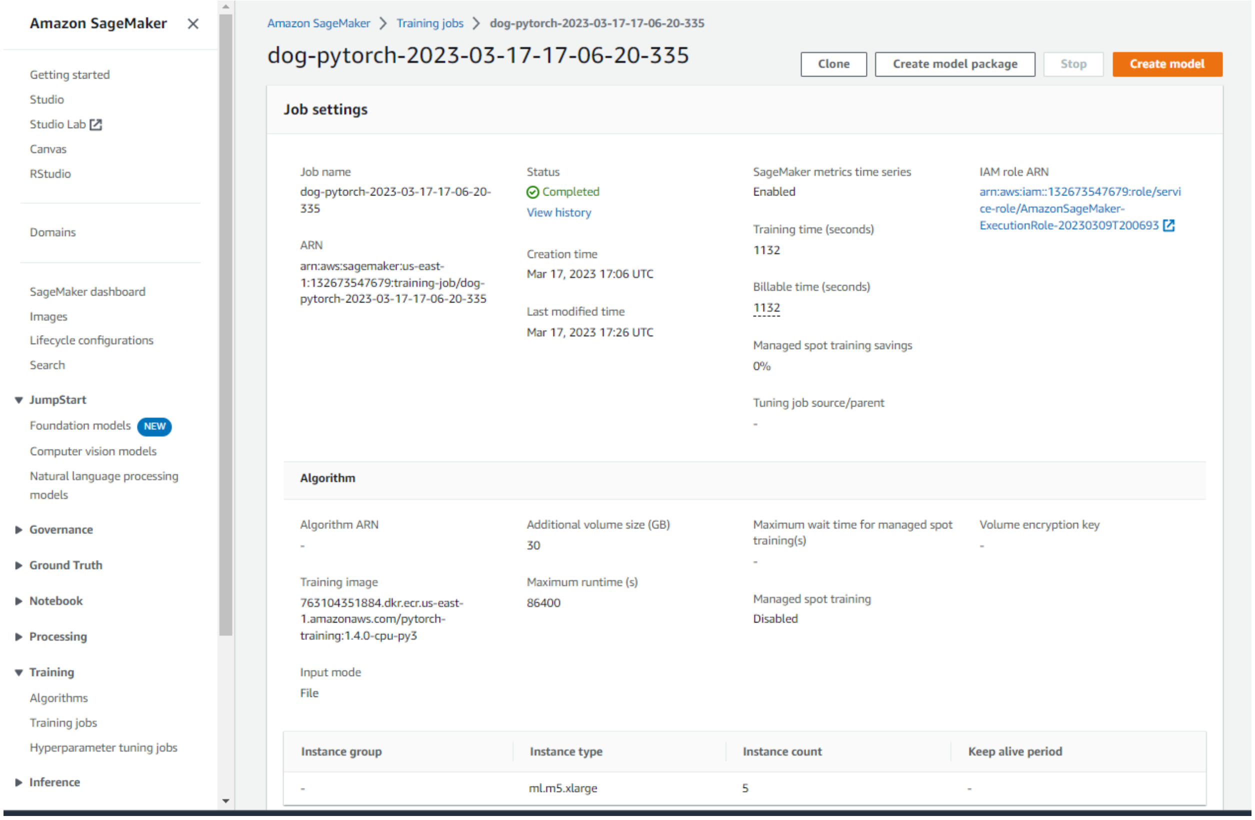
Maximum wait time for managed spot training(s)
-

Managed spot training
Disabled

Volume encryption key
-

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

- Multi instances



dog-pytorch-2023-03-17-17-06-20-335

Job settings

Job name
dog-pytorch-2023-03-17-17-06-20-335

ARN
arn:aws:sagemaker:us-east-1:132673547679:training-job/dog-pytorch-2023-03-17-17-06-20-335

Status
Completed
View history

Creation time
Mar 17, 2023 17:06 UTC

Last modified time
Mar 17, 2023 17:26 UTC

SageMaker metrics time series
Enabled

Training time (seconds)
1132

Billable time (seconds)
1132

Managed spot training savings
0%

Tuning job source/parent
-

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

Algorithm

Algorithm ARN
-

Training image
763104351884.dkr.ecr.us-east-1.amazonaws.com/pytorch-training:1.4.0-cpu-py3

Input mode
File

Additional volume size (GB)
30

Maximum runtime (s)
86400

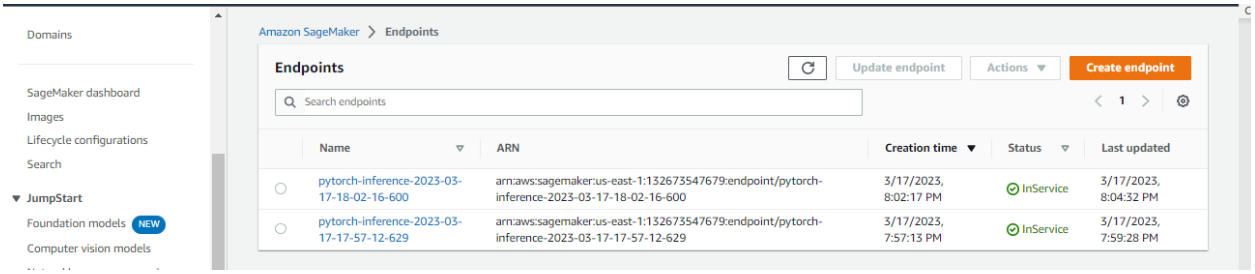
Maximum wait time for managed spot training(s)
-

Managed spot training
Disabled

Volume encryption key
-

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)

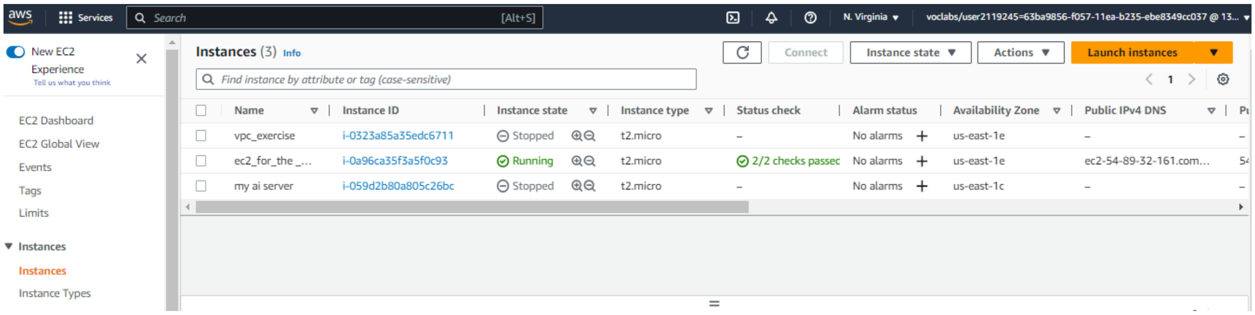


Endpoints

	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	Pı
☐	vpc_exercise	i-0323a85a35edc6711	⊖ Stopped	t2.micro	–	No alarms +	us-east-1e	–	–
☐	ec2_for_the _...	i-0a96ca35f3a5f0c93	✓ Running	t2.micro	✓ 2/2 checks passec	No alarms +	us-east-1e	ec2-54-89-32-161.com...	54
☐	my ai server	i-059d2b80a805c26bc	⊖ Stopped	t2.micro	–	No alarms +	us-east-1c	–	–

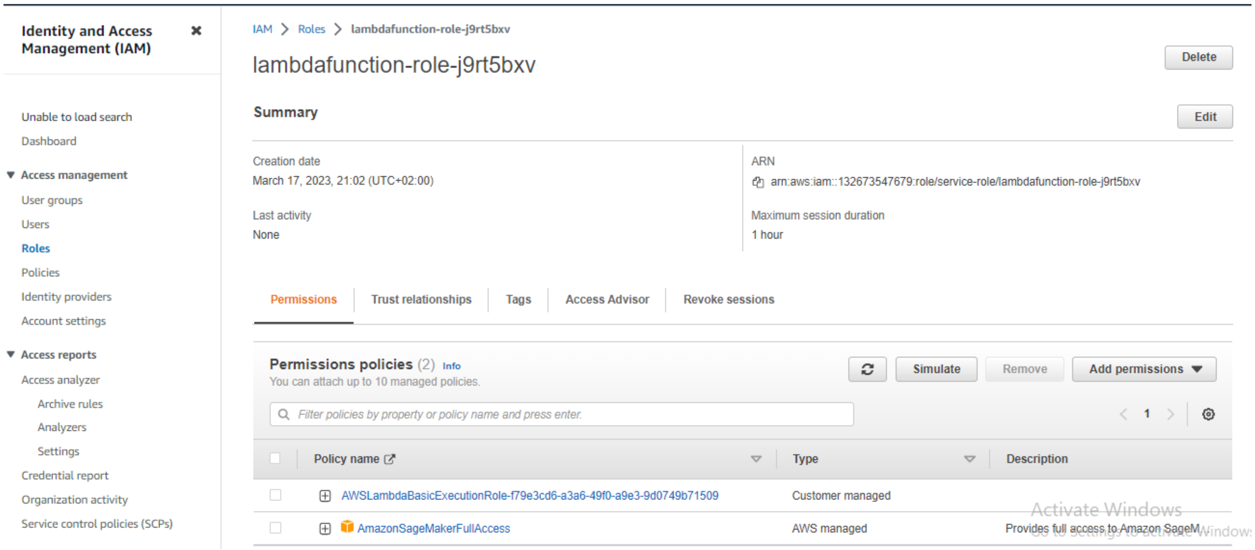
Difference between EC2 training code and SageMaker’s:

- **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 end point 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



lambdafunction-role-j9rt5bxv

Summary

Creation date
March 17, 2023, 21:02 (UTC+02:00)

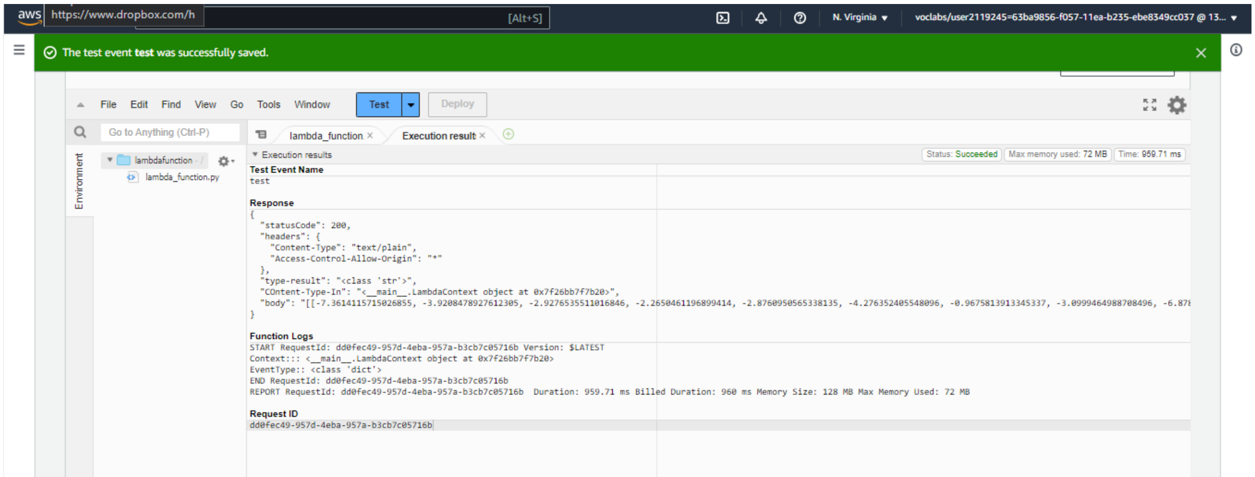
ARN
arn:aws:iam::132673547679:role/service-role/lambdafunction-role-j9rt5bxv

Last activity
None

Maximum session duration
1 hour

Permissions policies (2) Info
You can attach up to 10 managed policies.

Policy name	Type	Description
AWSLambdaBasicExecutionRole-f79e3cd6-a3a6-49f0-a9e3-9d0749b71509	Customer managed	
AmazonSageMakerFullAccess	AWS managed	Provides full access to Amazon SageM



The test event **test** was successfully saved.

Execution results

Status: Succeeded | Max memory used: 72 MB | Time: 959.71 ms

Test Event Name

```
test
```

Response

```
{
  "statusCode": 200,
  "headers": {
    "Content-Type": "text/plain",
    "Access-Control-Allow-Origin": "*"
  },
  "type-result": "<class 'str'>",
  "COntent-Type-In": "<__main__.LambdaContext object at 0x7f26bb7f7b20>",
  "body": "[[-7.3614115715026855, -3.9208478927612305, -2.9276535511016846, -2.2650461196899414, -2.8760950565338135, -4.276352405548096, -0.9675813913345337, -3.0999464988708496, -6.878
}
```

Function Logs

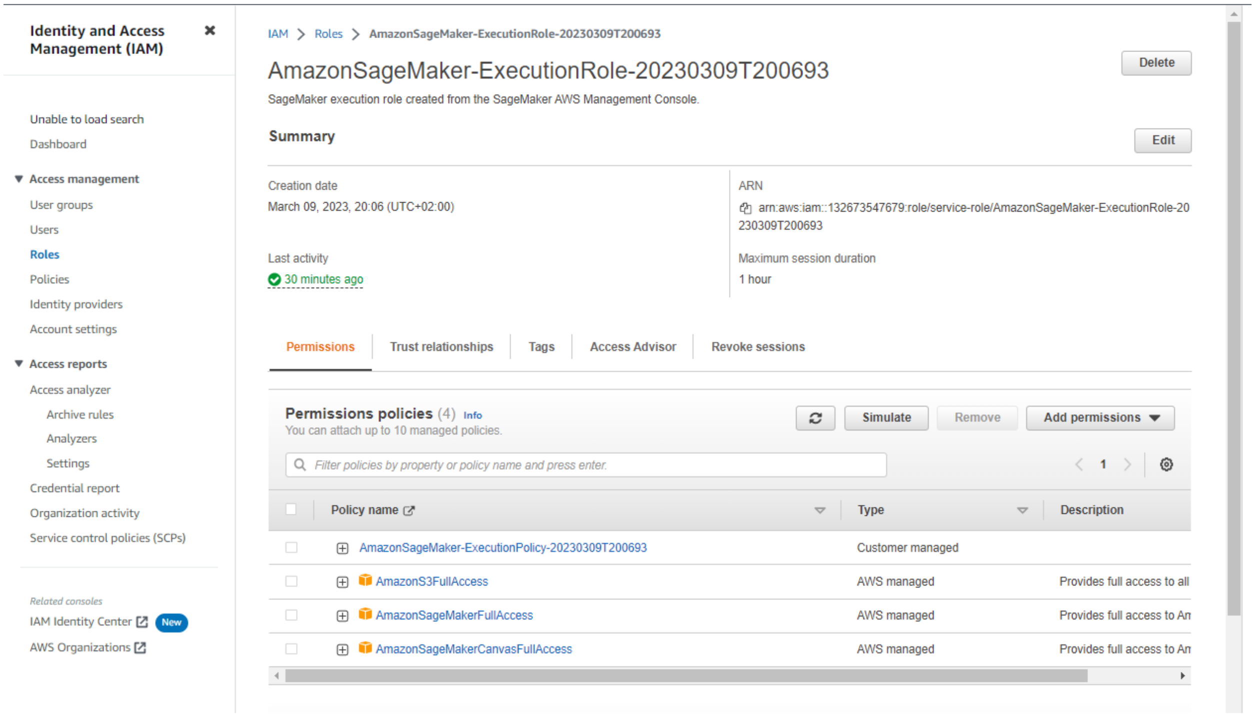
```
START RequestId: dd0fec49-957d-4eba-957a-b3cb7c05716b Version: $LATEST
Context::: <__main__.LambdaContext object at 0x7f26bb7f7b20>
EventType:: <class 'dict'>
END RequestId: dd0fec49-957d-4eba-957a-b3cb7c05716b
REPORT RequestId: dd0fec49-957d-4eba-957a-b3cb7c05716b  Duration: 959.71 ms Billed Duration: 960 ms Memory Size: 128 MB Max Memory Used: 72 MB
```

Request ID

```
dd0fec49-957d-4eba-957a-b3cb7c05716b
```

Security:

It’s not a good idea to give full access permission because it may result in a security preach, 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.



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

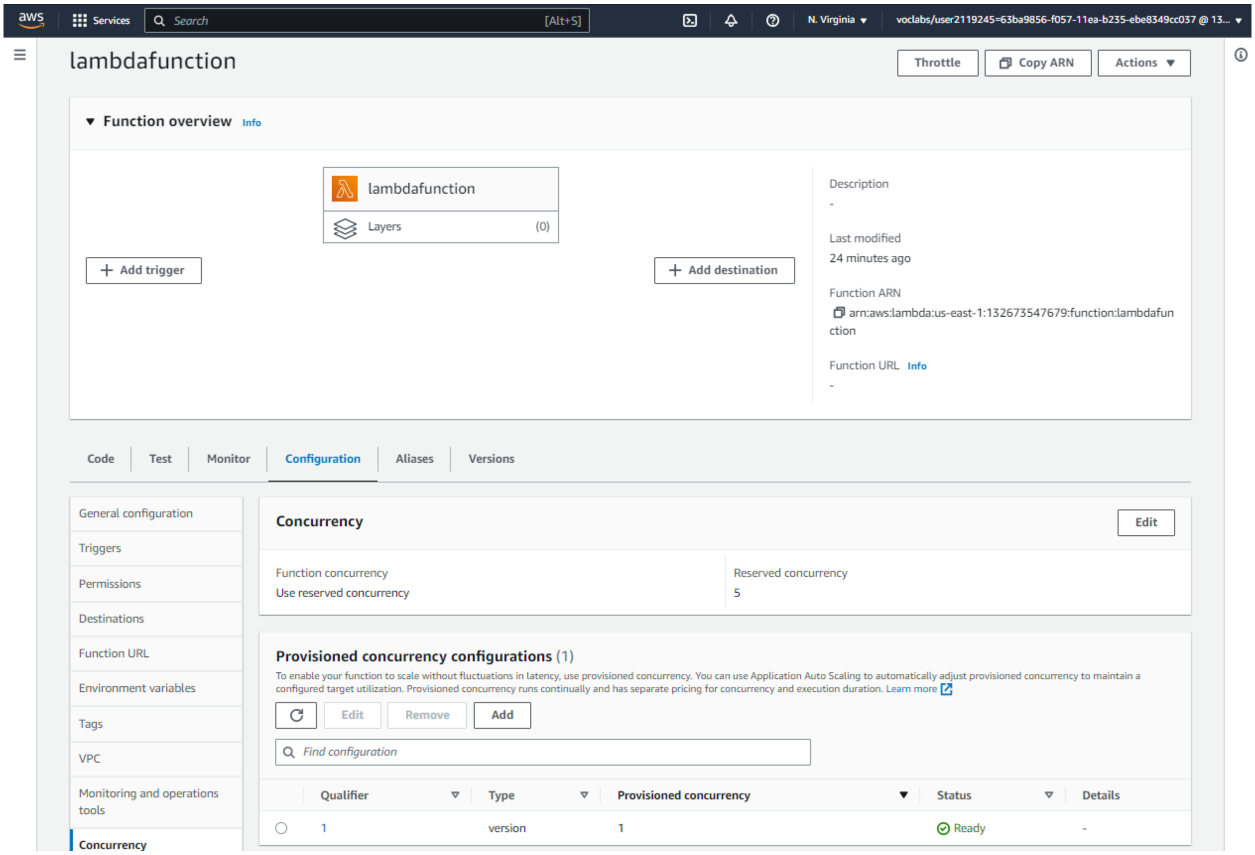
Permissions policies (4) Info

You can attach up to 10 managed policies.

Policy name	Type	Description
AmazonSageMaker-ExecutionPolicy-20230309T200693	Customer managed	
AmazonS3FullAccess	AWS managed	Provides full access to all
AmazonSageMakerFullAccess	AWS managed	Provides full access to An
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.



lambdafunction

Description
-

Last modified
24 minutes ago

Function ARN
arn:aws:lambda:us-east-1:132673547679:function:lambdafunction

Function URL Info
-

Concurrency

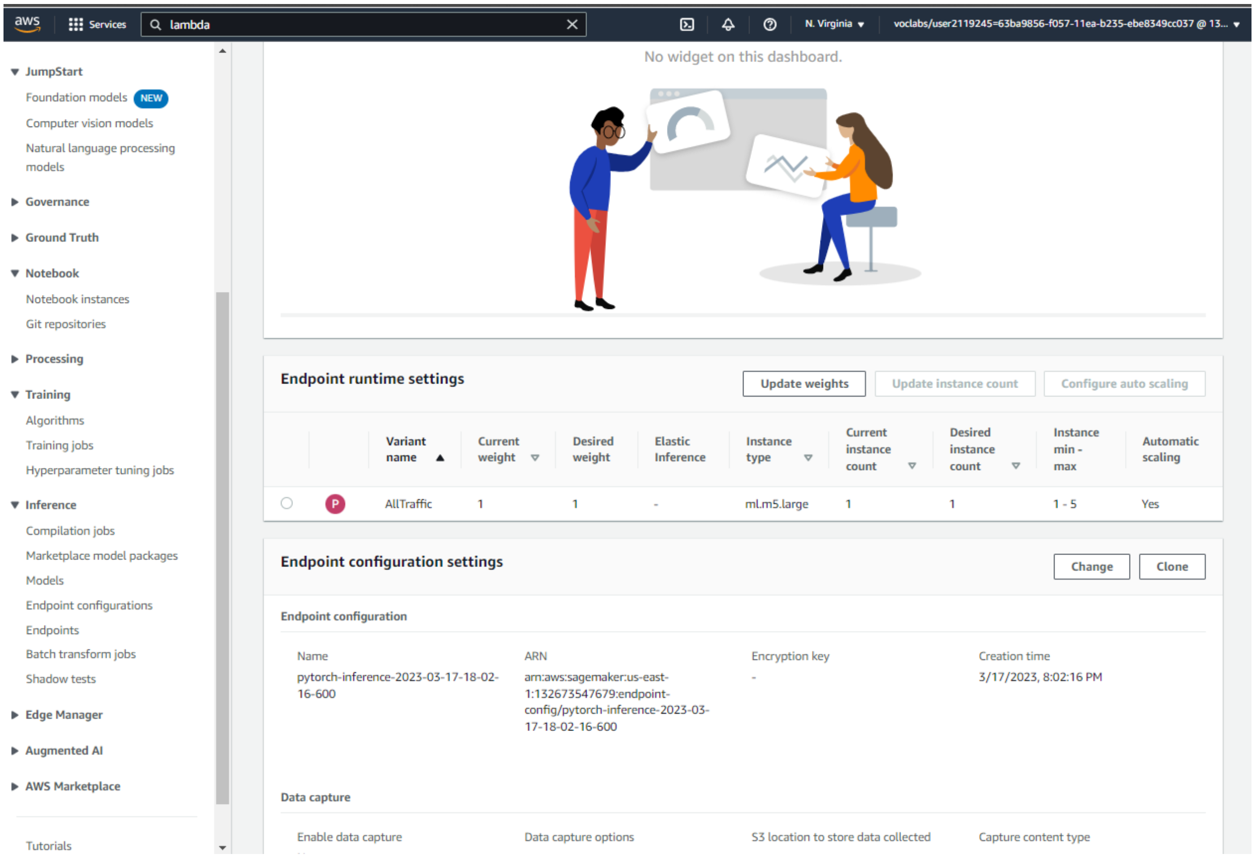
Function concurrency
Use reserved concurrency

Reserved concurrency
5

Provisioned concurrency configurations (1)

To enable your function to scale without fluctuations in latency, use provisioned concurrency. You can use Application Auto Scaling to automatically adjust provisioned concurrency to maintain a configured target utilization. Provisioned concurrency runs continually and has separate pricing for concurrency and execution duration. Learn more

	Qualifier	Type	Provisioned concurrency	Status	Details
○	1	version	1	✓ Ready	-



No widget on this dashboard.

Endpoint runtime settings

		Variant name ▲	Current weight ▽	Desired weight	Elastic Inference	Instance type ▽	Current instance count ▽	Desired instance count ▽	Instance min - max	Automatic scaling
○	P	AllTraffic	1	1	-	ml.m5.large	1	1	1 - 5	Yes

Endpoint configuration settings

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