Operationalizing an AWS ML Project

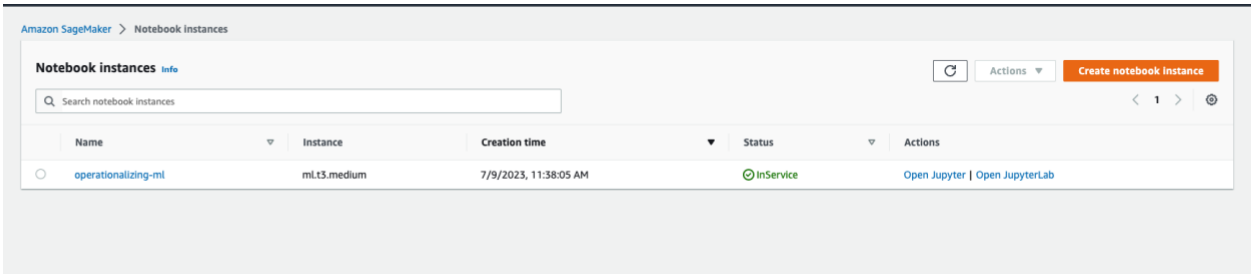
1. Initial Setup

I have chosen the "ml.t3.medium" instance type for this Notebook due to a variety of considerations, all of which cater to the specific needs of this project.

Firstly, the nature of our project doesn't necessitate the employment of a high-performance CPU or an abundance of RAM for the successful execution of our Jupyter notebooks. Rather than focusing on raw computational power, it's more prudent to consider the time duration our notebook instance will remain active.

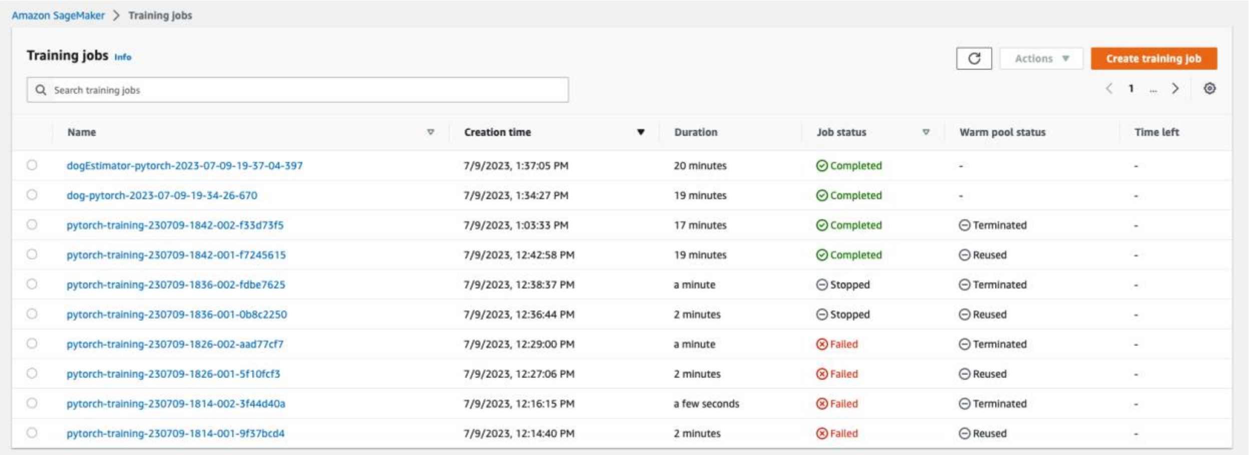
Considering the potentially substantial duration of our project, it's essential to select an instance with a cost-effective hourly rate, in addition to providing a reasonable level of CPU and RAM. This strategic decision is crucial to curbing excessive costs while ensuring the functionality and efficiency of our project are uncompromised.

In light of these considerations, the "ml.t3.medium" instance emerges as an optimal choice. It provides a balance between cost-efficiency and computational prowess that fits our project's requirements. Despite a slower startup time, this instance type is more cost-effective per hour, a trade-off that is acceptable given our project's lack of dependency on instant start-up.


Amazon SageMaker > Notebook instances
Notebook instances Info
Actions ▼
Create notebook instance
Search notebook instances
< 1 >

	Name	Instance	Creation time	Status	Actions
○	operationalizing-ml	ml.t3.medium	7/9/2023, 11:38:05 AM	✓ InService	Open Jupyter | Open JupyterLab

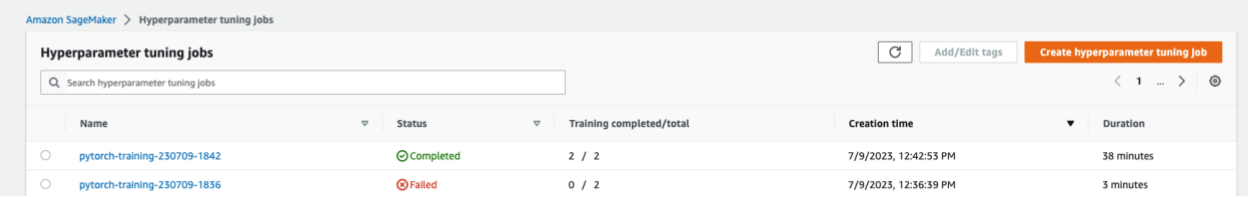
S3 Bucket



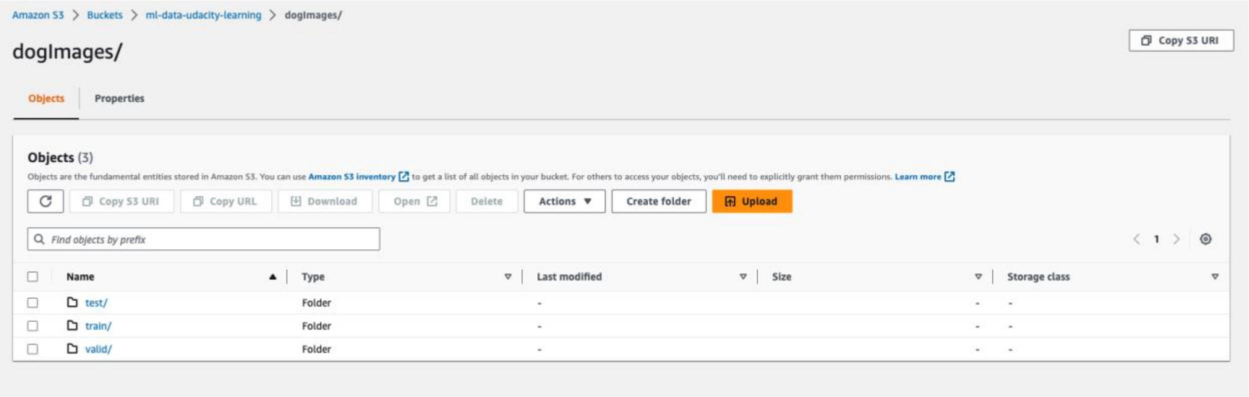
Name	Creation time	Duration	Job status	Warm pool status	Time left
dogEstimator-pytorch-2023-07-09-19-37-04-397	7/9/2023, 1:37:05 PM	20 minutes	Completed	-	-
dog-pytorch-2023-07-09-19-34-26-670	7/9/2023, 1:34:27 PM	19 minutes	Completed	-	-
pytorch-training-230709-1842-002-f33d73f5	7/9/2023, 1:03:33 PM	17 minutes	Completed	Terminated	-
pytorch-training-230709-1842-001-f7245615	7/9/2023, 12:42:58 PM	19 minutes	Completed	Reused	-
pytorch-training-230709-1836-002-fdbe7625	7/9/2023, 12:38:37 PM	a minute	Stopped	Terminated	-
pytorch-training-230709-1836-001-0b8c2250	7/9/2023, 12:36:44 PM	2 minutes	Stopped	Reused	-
pytorch-training-230709-1826-002-aad77cf7	7/9/2023, 12:29:00 PM	a minute	Failed	Terminated	-
pytorch-training-230709-1826-001-5f10fcf3	7/9/2023, 12:27:06 PM	2 minutes	Failed	Reused	-
pytorch-training-230709-1814-002-3f44d40a	7/9/2023, 12:16:15 PM	a few seconds	Failed	Terminated	-
pytorch-training-230709-1814-001-9f37bcd4	7/9/2023, 12:14:40 PM	2 minutes	Failed	Reused	-

Training and Tuning Jobs

Training and Tuning didn’t take that long because of the instance type I used.


Amazon SageMaker > Hyperparameter tuning jobs
Hyperparameter tuning jobs
Add/Edit tags
Create hyperparameter tuning job
Search hyperparameter tuning jobs
1 ...

Name	Status	Training completed/total	Creation time	Duration
pytorch-training-230709-1842	Completed	2 / 2	7/9/2023, 12:42:53 PM	38 minutes
pytorch-training-230709-1836	Failed	0 / 2	7/9/2023, 12:36:39 PM	3 minutes

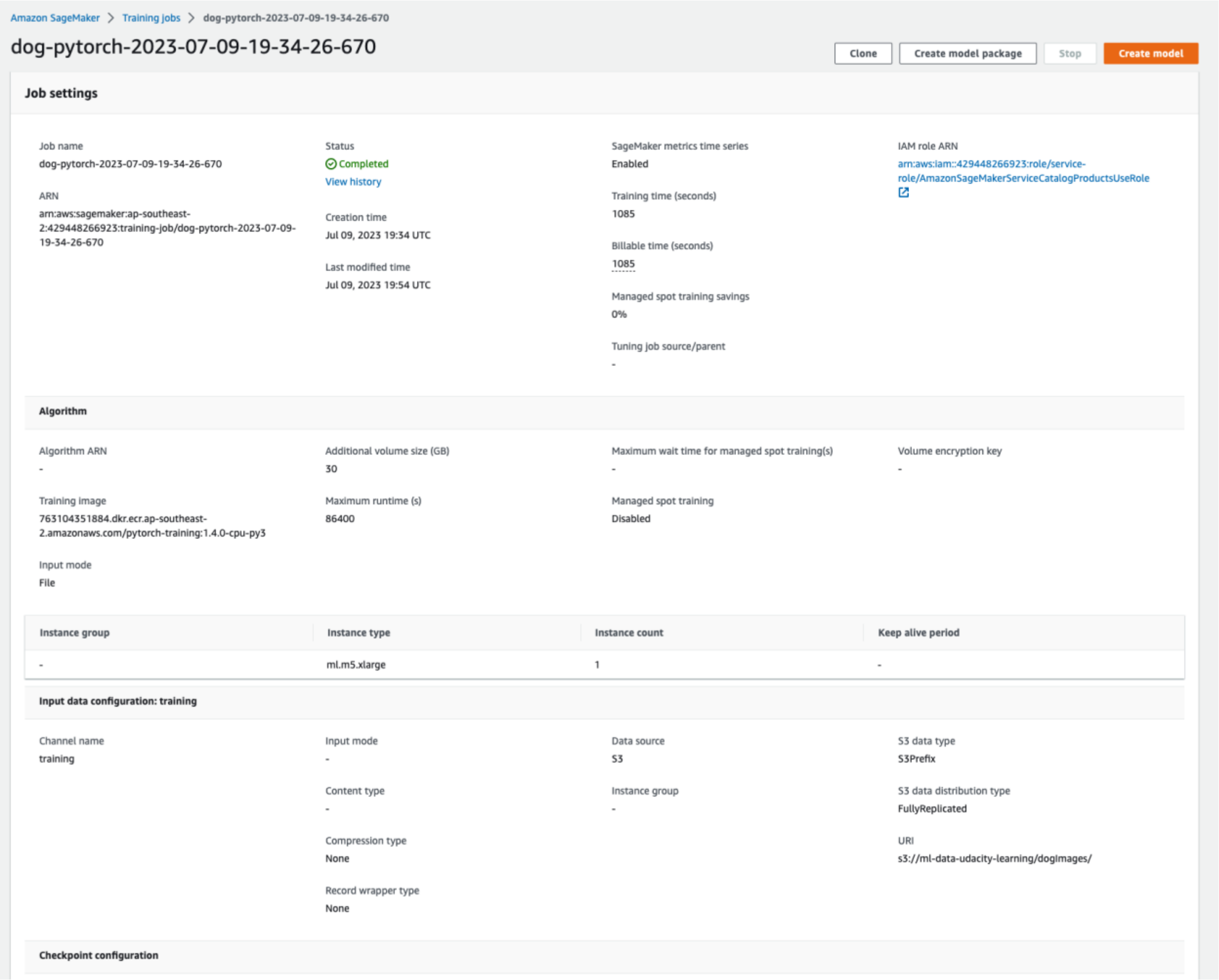


Objects (3)

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	-	-	-

Single Instance Training Job



Amazon SageMaker > Training jobs > dog-pytorch-2023-07-09-19-34-26-670

dog-pytorch-2023-07-09-19-34-26-670

Clone
Create model package
Stop
Create model

Job settings

Job name
dog-pytorch-2023-07-09-19-34-26-670

ARN
arn:aws:sagemaker:ap-southeast-2:429448266923:training-job/dog-pytorch-2023-07-09-19-34-26-670

Status
Completed
View history

Creation time
Jul 09, 2023 19:34 UTC

Last modified time
Jul 09, 2023 19:54 UTC

SageMaker metrics time series
Enabled

Training time (seconds)
1085

Billable time (seconds)
1085

Managed spot training savings
0%

Tuning job source/parent
-

IAM role ARN
arn:aws:iam::429448266923:role/service-role/AmazonSageMakerServiceCatalogProductsUseRole

Algorithm

Algorithm ARN
-

Training image
763104351884.dkr.ecr.ap-southeast-2.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	-

Input data configuration: training

Channel name
training

Input mode
-

Content type
-

Compression type
None

Record wrapper type
None

Data source
S3

Instance group
-

S3 data type
S3Prefix

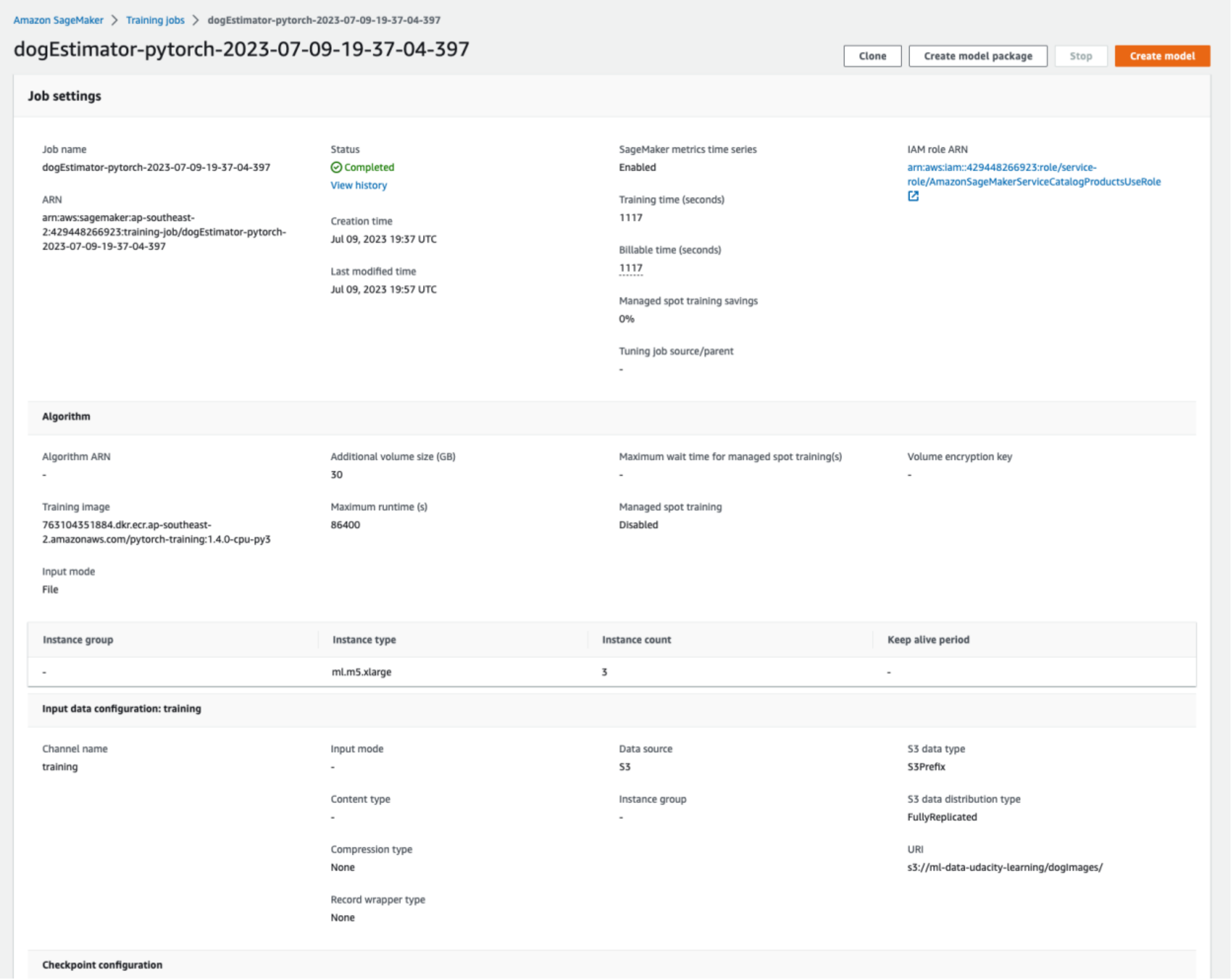
S3 data distribution type
FullyReplicated

URI
s3://ml-data-udacity-learning/dogImages/

Checkpoint configuration

Both Training Jobs took fairly equal amount of time and I am very surprised.

Multi-Instance Training Job:


Amazon SageMaker > Training jobs > dogEstimator-pytorch-2023-07-09-19-37-04-397
Clone
Create model package
Stop
Create model

dogEstimator-pytorch-2023-07-09-19-37-04-397

Job settings

Job name
dogEstimator-pytorch-2023-07-09-19-37-04-397

ARN
arn:aws:sagemaker:ap-southeast-2:429448266923:training-job/dogEstimator-pytorch-2023-07-09-19-37-04-397

Status
Completed
View history

Creation time
Jul 09, 2023 19:37 UTC

Last modified time
Jul 09, 2023 19:57 UTC

SageMaker metrics time series
Enabled

Training time (seconds)
1117

Billable time (seconds)
1117

Managed spot training savings
0%

Tuning job source/parent
-

IAM role ARN
arn:aws:iam::429448266923:role/service-role/AmazonSageMakerServiceCatalogProductsUseRole

Algorithm

Algorithm ARN
-

Training image
763104351884.dkr.ecr.ap-southeast-2.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	3	-

Input data configuration: training

Channel name
training

Input mode
-

Content type
-

Compression type
None

Record wrapper type
None

Data source
S3

Instance group
-

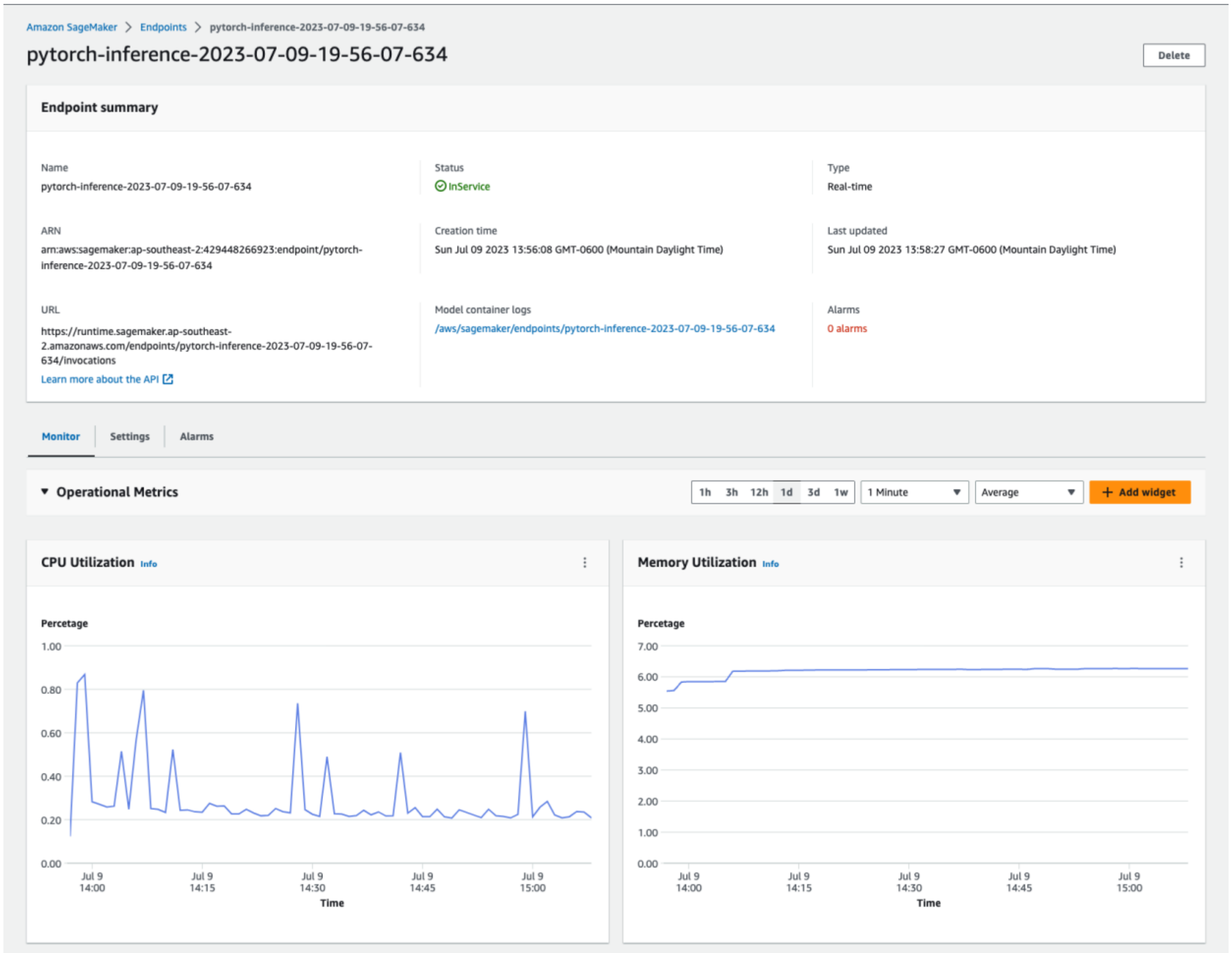
S3 data type
S3Prefix

S3 data distribution type
FullyReplicated

URI
s3://ml-data-udacity-learning/dogImages/

Checkpoint configuration

Deployment Endpoints



Amazon SageMaker > Endpoints > pytorch-inference-2023-07-09-19-56-07-634

pytorch-inference-2023-07-09-19-56-07-634

Endpoint summary

Name
pytorch-inference-2023-07-09-19-56-07-634

Status
InService

Type
Real-time

ARN
arn:aws:sagemaker:ap-southeast-2:429448266923:endpoint/pytorch-inference-2023-07-09-19-56-07-634

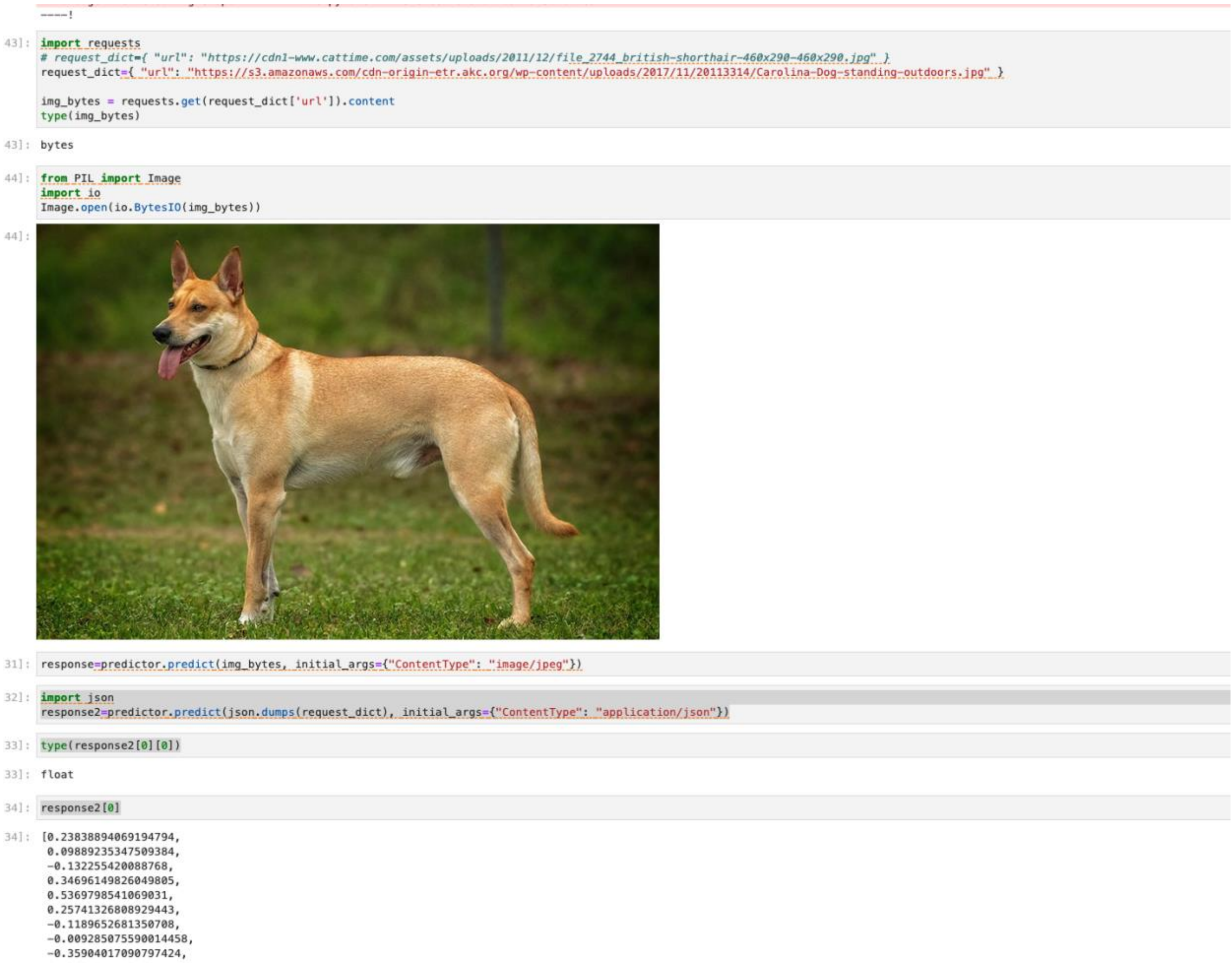
Creation time
Sun Jul 09 2023 13:56:08 GMT-0600 (Mountain Daylight Time)

Last updated
Sun Jul 09 2023 13:58:27 GMT-0600 (Mountain Daylight Time)

URL
https://runtime.sagemaker.ap-southeast-2.amazonaws.com/endpoints/pytorch-inference-2023-07-09-19-56-07-634/invocations
Learn more about the API

Model container logs
/aws/sagemaker/endpoints/pytorch-inference-2023-07-09-19-56-07-634

Alarms
0 alarms



```
----!
```

```
43]: import requests
     # request_dict={ "url": "https://cdn1-www.cattime.com/assets/uploads/2011/12/file_2744_british-shorthair-460x290-460x290.jpg" }
     request_dict={ "url": "https://s3.amazonaws.com/cdn-origin-etr.akc.org/wp-content/uploads/2017/11/20113314/Carolina-Dog-standing-outdoors.jpg" }

     img_bytes = requests.get(request_dict['url']).content
     type(img_bytes)

43]: bytes

44]: from PIL import Image
     import io
     Image.open(io.BytesIO(img_bytes))

44]:
```

```
31]: response=predictor.predict(img_bytes, initial_args={"ContentType": "image/jpeg"})

32]: import json
     response2=predictor.predict(json.dumps(request_dict), initial_args={"ContentType": "application/json"})

33]: type(response2[0][0])

33]: float

34]: response2[0]

34]: [0.23838894069194794,
      0.09889235347509384,
      -0.132255420088768,
      0.34696149826049805,
      0.5369798541069031,
      0.25741326808929443,
      -0.1189652681350708,
      -0.009285075590014458,
      -0.35904017090797424,
```

2. EC2 Training

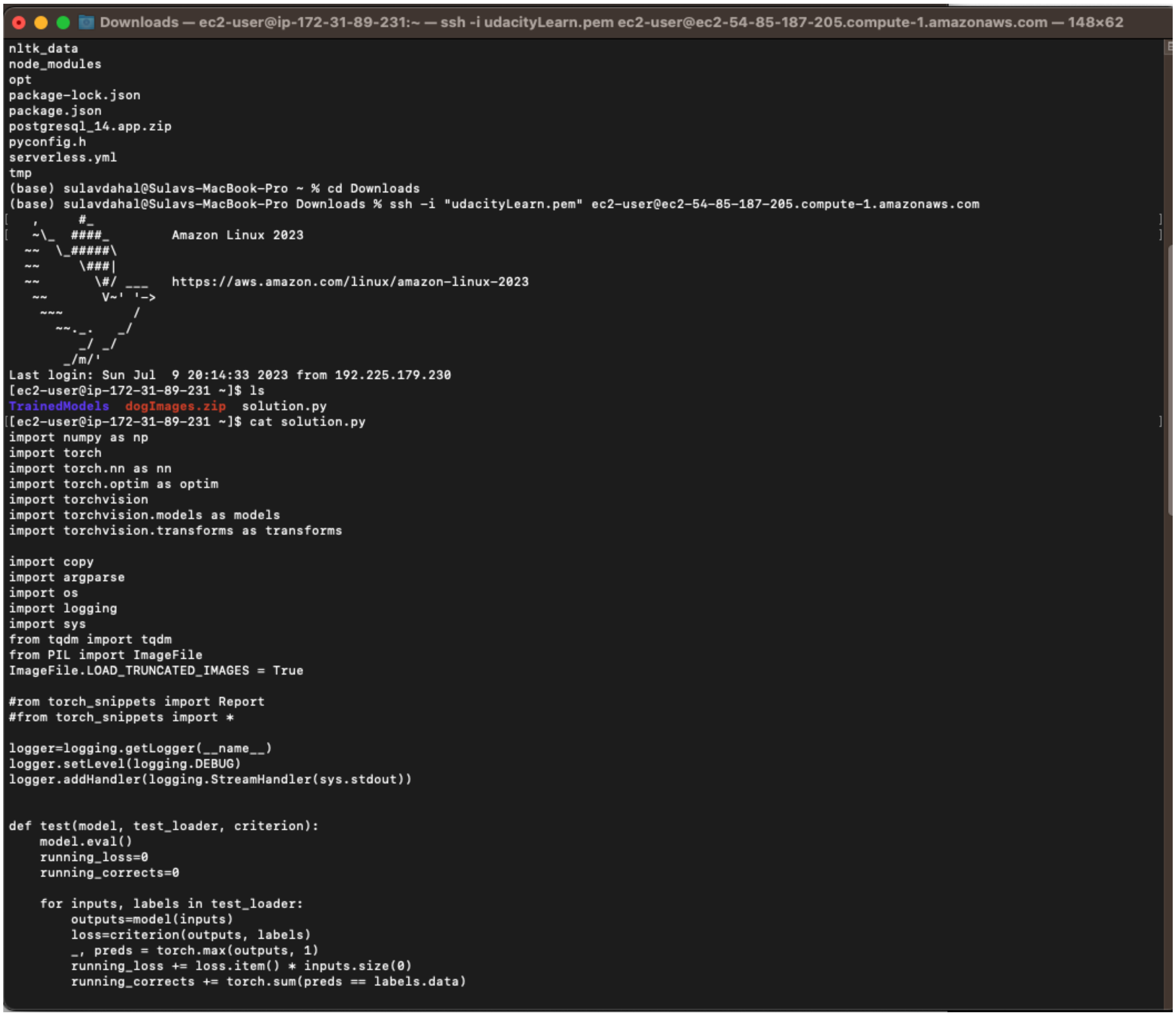
I employed the Deep Learning AMI, specifically GPU PyTorch 2.0.1 (Amazon Linux 2) 20230627, ami-051619310404cab17 (64-bit (x86)), alongside the g4dn.xlarge instance. This combination strikes an effective balance between financial considerations and performance capabilities.

To begin with, g4dn.xlarge instances emerge as the most economical option for this project, primarily due to the fact that the g4dn series generally represents the least expensive instances compatible with the Deep Learning AMI.

According to the official documentation, g4dn.xlarge instances are capable of maintaining superior CPU performance for the duration required by a task, a characteristic that adds to their suitability for this project.

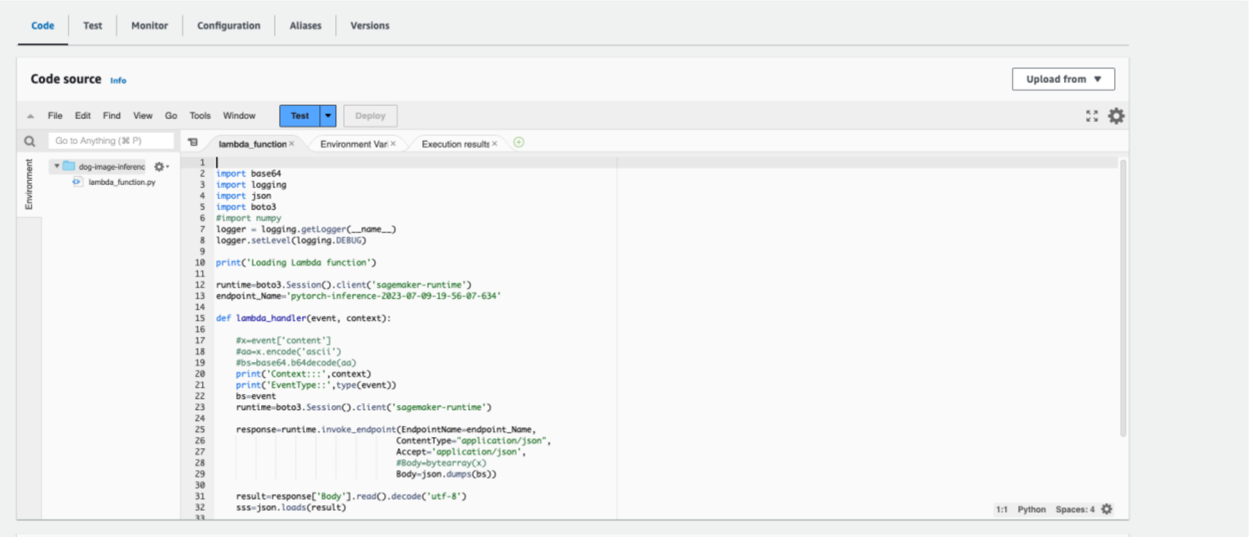
Lastly, without incurring any additional expenses, g4dn instances provide adequate performance for the majority of applications of a general nature.

This instance type is supported by several EC2 instances, namely G3, P3, P3dn, P4d, P4de, G5, and G4dn. For further details, you can refer to the release notes available at: https://docs.aws.amazon.com/dlami/latest/devguide/appendix-ami-release-notes.html.

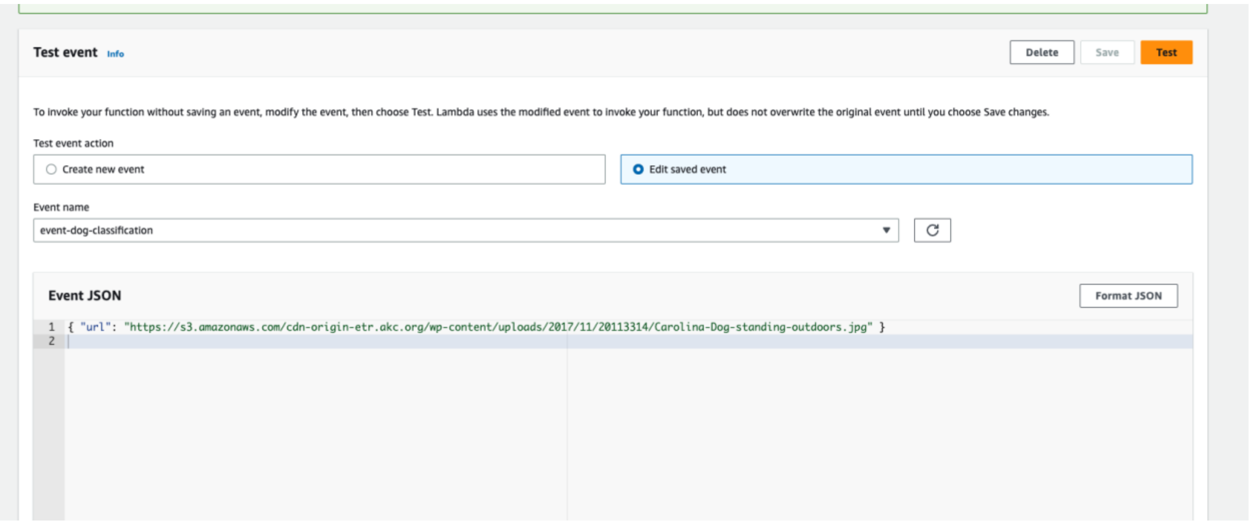

Downloads — ec2-user@ip-172-31-89-231:~ — ssh -i udacityLearn.pem ec2-user@ec2-54-85-187-205.compute-1.amazonaws.com — 148×62

```
nltk_data
node_modules
opt
package-lock.json
package.json
postgresql_14.app.zip
pyconfig.h
serverless.yml
tmp
(base) sulavdahal@Sulavs-MacBook-Pro ~ % cd Downloads
(base) sulavdahal@Sulavs-MacBook-Pro Downloads % ssh -i "udacityLearn.pem" ec2-user@ec2-54-85-187-205.compute-1.amazonaws.com
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Sun Jul  9 20:14:33 2023 from 192.225.179.230
[ec2-user@ip-172-31-89-231 ~]$ ls
TrainedModels  dogImages.zip  solution.py
[ec2-user@ip-172-31-89-231 ~]$ cat solution.py
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
import torchvision
import torchvision.models as models
import torchvision.transforms as transforms

import copy
import argparse
import os
import logging
import sys
from tqdm import tqdm
from PIL import ImageFile
ImageFile.LOAD_TRUNCATED_IMAGES = True

#rom torch_snippets import Report
#from torch_snippets import *

logger=logging.getLogger(__name__)
logger.setLevel(logging.DEBUG)
logger.addHandler(logging.StreamHandler(sys.stdout))


def test(model, test_loader, criterion):
    model.eval()
    running_loss=0
    running_corrects=0

    for inputs, labels in test_loader:
        outputs=model(inputs)
        loss=criterion(outputs, labels)
        _, preds = torch.max(outputs, 1)
        running_loss += loss.item() * inputs.size(0)
        running_corrects += torch.sum(preds == labels.data)
```

3. Lambda Function



```python

import base64
import logging
import json
import boto3
#import numpy
logger = logging.getLogger(__name__)
logger.setLevel(logging.DEBUG)

print('Loading Lambda function')

runtime=boto3.Session().client('sagemaker-runtime')
endpoint_Name='pytorch-inference-2023-07-09-19-56-07-634'

def lambda_handler(event, context):

    #x=event['content']
    #aa=x.encode('ascii')
    #bs=base64.b64decode(aa)
    print('Context:::',context)
    print('EventType::',type(event))
    bs=event
    runtime=boto3.Session().client('sagemaker-runtime')

    response=runtime.invoke_endpoint(EndpointName=endpoint_Name,
                                    ContentType="application/json",
                                    Accept='application/json',
                                    #Body=bytearray(x)
                                    Body=json.dumps(bs))

    result=response['Body'].read().decode('utf-8')
    sss=json.loads(result)
```



To invoke your function without saving an event, modify the event, then choose Test. Lambda uses the modified event to invoke your function, but does not overwrite the original event until you choose Save changes.

Test event action

○ Create new event   ● Edit saved event

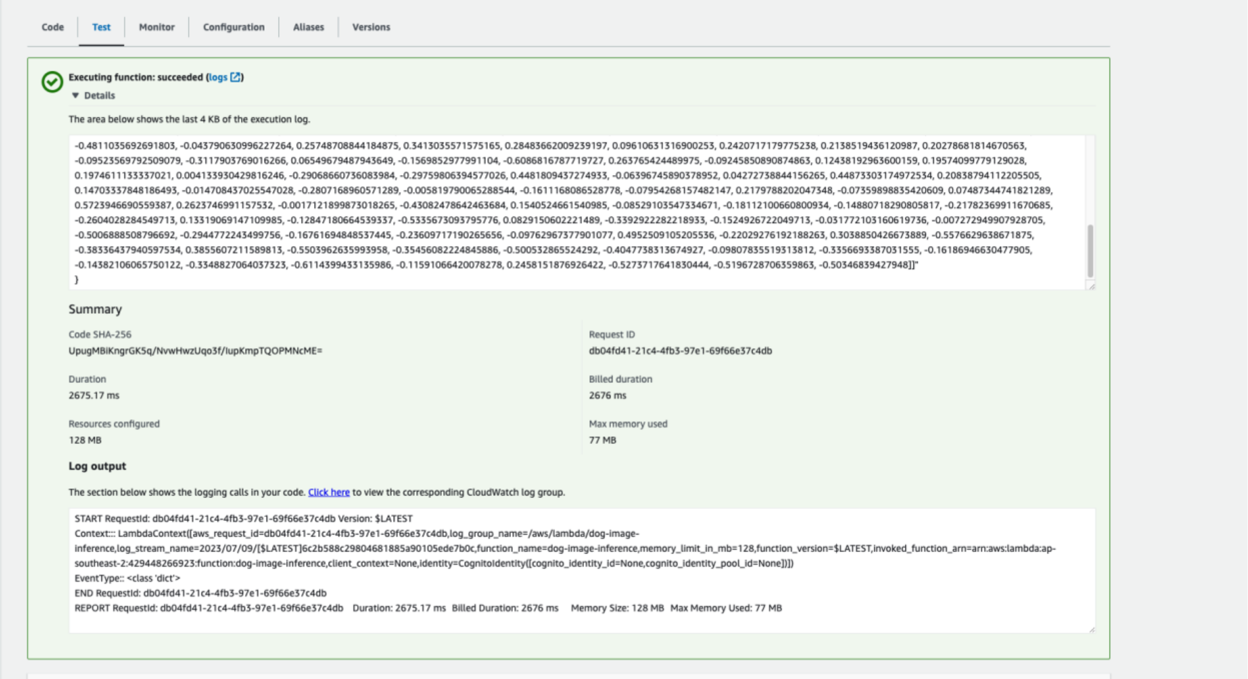
Event name

event-dog-classification

Event JSON

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

A SageMaker Policy has been attached to make the lambda code functional.



Executing function: succeeded (logs)

▼ Details

The area below shows the last 4 KB of the execution log.

```
-0.481103569269180З, -0.043790630996227264, 0.25748708844184875, 0.3413035571575165, 0.28483662009239197, 0.09610631316900253, 0.2420717179775238, 0.2138519436120987, 0.20278681814670563,
-0.09523569792509079, -0.3117903769016266, 0.06549679487943649, -0.1569852977991104, -0.6086816787719727, 0.263765424489975, -0.09245850890874863, 0.12438192963600159, 0.19574099779129028,
0.1974611133337021, 0.004133930429816246, -0.29068660736083984, -0.29759806394577026, 0.4481809437274933, -0.06396745890378952, 0.04272738844156265, 0.44873303174972534, 0.20838794112205505,
0.14703337848186493, -0.014708437025547028, -0.2807168960571289, -0.005819790065288544, -0.1611168086528778, -0.07954268157482147, 0.2179788202047348, -0.07359898835420609, 0.07487344741821289,
0.5723946690559387, 0.2623746991157532, -0.0017121899873018265, -0.43082478642463684, 0.1540524661540985, -0.08529103547334671, -0.18112100660800934, -0.14880718290805817, -0.21782369911670685,
-0.2604028284549713, 0.13319069147109985, -0.12847180664539337, -0.5335673093795776, 0.0829150602221489, -0.3392922282218933, -0.1524926722049713, -0.031772103160619736, -0.007272949907928705,
-0.5006888508796692, -0.2944772243499756, -0.16761694848537445, -0.23609717190265656, -0.09762967377901077, 0.4952509105205536, -0.22029276192188263, 0.3038850426673889, -0.5576629638671875,
-0.38336437940597534, 0.3855607211589813, -0.5503962635993958, -0.35456082224845886, -0.500532865524292, -0.4047738313674927, -0.09807835519313812, -0.3356693387031555, -0.16186946630477905,
-0.14382106065750122, -0.3348827064037323, -0.6114399433135986, -0.11591066420078278, 0.2458151876926422, -0.5273717641830444, -0.5196728706359863, -0.50346839427948]]"
}
```

Summary

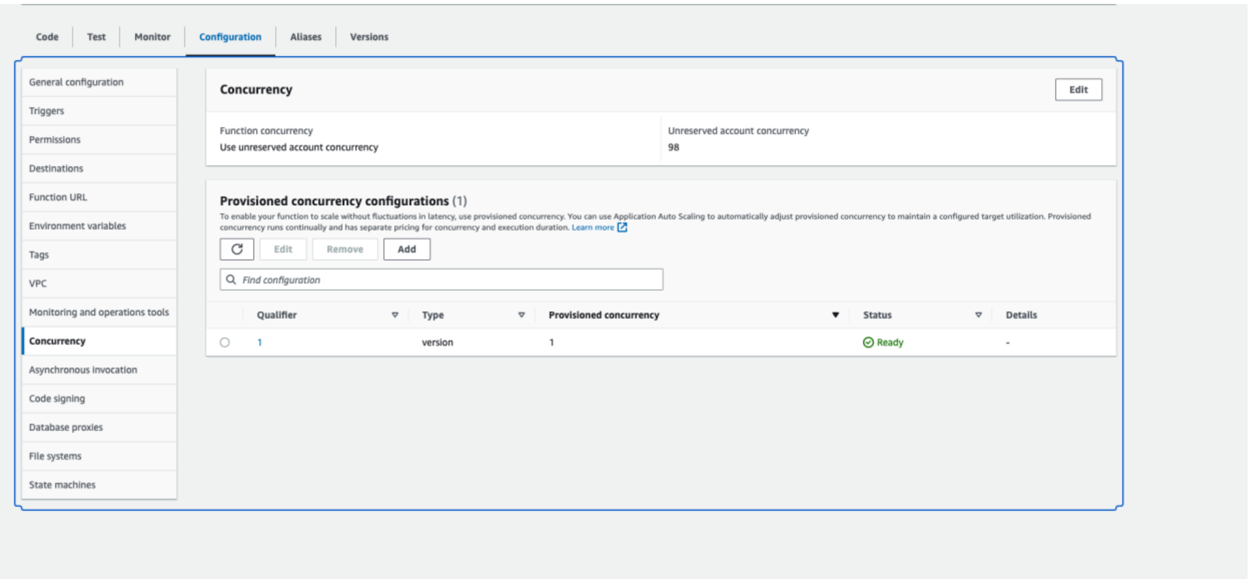
Code SHA-256 UpugMBiKngrGK5q/NvwHwzUqo3f/IupKmpTQOPMNcME=	Request ID db04fd41-21c4-4fb3-97e1-69f66e37c4db
Duration 2675.17 ms	Billed duration 2676 ms
Resources configured 128 MB	Max memory used 77 MB

Log output

The section below shows the logging calls in your code. Click here to view the corresponding CloudWatch log group.

```
START RequestId: db04fd41-21c4-4fb3-97e1-69f66e37c4db Version: $LATEST
Context::: LambdaContext([aws_request_id=db04fd41-21c4-4fb3-97e1-69f66e37c4db,log_group_name=/aws/lambda/dog-image-
inference,log_stream_name=2023/07/09/[$LATEST]6c2b588c29804681885a90105ede7b0c,function_name=dog-image-inference,memory_limit_in_mb=128,function_version=$LATEST,invoked_function_arn=arn:aws:lambda:ap-
southeast-2:429448266923:function:dog-image-inference,client_context=None,identity=CognitoIdentity([cognito_identity_id=None,cognito_identity_pool_id=None])])
EventType:: <class 'dict'>
END RequestId: db04fd41-21c4-4fb3-97e1-69f66e37c4db
REPORT RequestId: db04fd41-21c4-4fb3-97e1-69f66e37c4db	Duration: 2675.17 ms	Billed Duration: 2676 ms	Memory Size: 128 MB	Max Memory Used: 77 MB
```

Concurrency



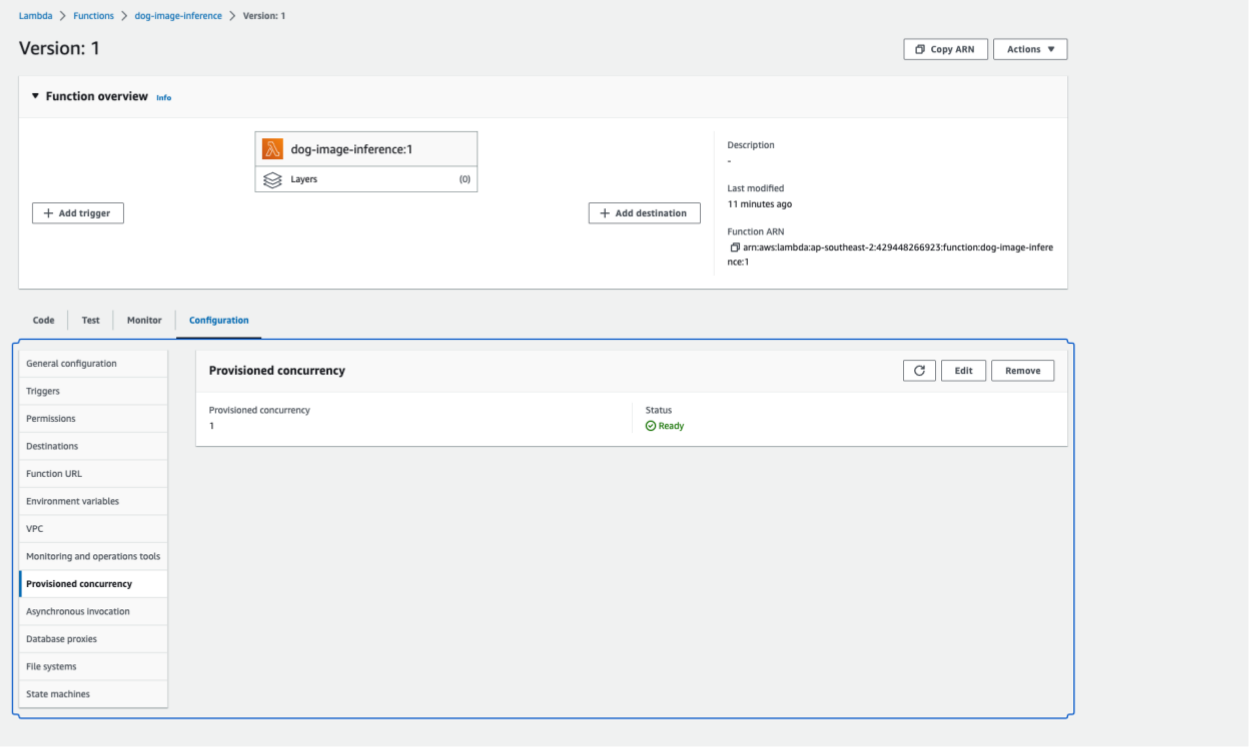
Concurrency

Function concurrency Use unreserved account concurrency	Unreserved account concurrency 98

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	-



Version: 1

Function overview Info

dog-image-inference:1

Layers (0)

Description
-

Last modified
11 minutes ago

Function ARN
arn:aws:lambda:ap-southeast-2:429448266923:function:dog-image-inference:1

Provisioned concurrency

Provisioned concurrency
1

Status
Ready