SageMaker Serverless Inference

Amazon SageMaker Serverless Inference is a purpose-built inference option that makes it easy to deploy and scale ML models.

Serverless Inference is ideal for workloads which have **idle periods** between traffic peaks and can tolerate **cold starts**. Serverless endpoints also automatically launch compute resources and scale them in and out depending on traffic, eliminating the need to choose instance types or manage scaling policies. Tiis is ideal for scenarios that we have constrains in cost and at the same time we have no idea about the type of instance we need.

For this notebook we'll be working with the **SageMaker XGBoost Algorithm** to train a model and then deploy a serverless endpoint. We will be using the public S3 Abalone regression dataset for this example.

In [1]:
```
! pip install sagemaker botocore boto3 awscli --upgrade
```

```
Requirement already satisfied: sagemaker in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages
(2.188.0)
Collecting sagemaker
  Downloading sagemaker-2.196.0.tar.gz (916 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 916.9/916.9 kB 36.6 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: botocore in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages
(1.31.57)
Collecting botocore
  Obtaining dependency information for botocore from https://files.pythonhosted.org/packages/64/93/d472917cc800
0157d1220f18edf175e06b7f1ac92970e7cdb719d4405b5a/botocore-1.31.78-py3-none-any.whl.metadata
  Downloading botocore-1.31.78-py3-none-any.whl.metadata (6.1 kB)
Requirement already satisfied: boto3 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages (1.2
8.57)
Collecting boto3
  Obtaining dependency information for boto3 from https://files.pythonhosted.org/packages/17/a4/e161e777d445a5b
1029eff4c4272896410d7dac55e17f2889bd9a3517eed/boto3-1.28.78-py3-none-any.whl.metadata
  Downloading boto3-1.28.78-py3-none-any.whl.metadata (6.7 kB)
Requirement already satisfied: awscli in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages (1.
29.57)
Collecting awscli
  Obtaining dependency information for awscli from https://files.pythonhosted.org/packages/4a/3a/2bc87e42c30f79
423c8ded2efd9e94a0d50cba07e109d08ceb76861272d2/awscli-1.29.78-py3-none-any.whl.metadata
  Downloading awscli-1.29.78-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: attrs<24,>=23.1.0 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-p
ackages (from sagemaker) (23.1.0)
Requirement already satisfied: cloudpickle==2.2.1 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-
packages (from sagemaker) (2.2.1)
Requirement already satisfied: google-pasta in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packag
es (from sagemaker) (0.2.0)
Requirement already satisfied: numpy<2.0,>=1.9.0 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-p
ackages (from sagemaker) (1.22.3)
Requirement already satisfied: protobuf<5.0,>=3.12 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site
-packages (from sagemaker) (4.24.3)
Requirement already satisfied: smdebug_rulesconfig==1.0.1 in /home/ec2-user/anaconda3/envs/python3/lib/python3.
10/site-packages (from sagemaker) (1.0.1)
Requirement already satisfied: importlib-metadata<7.0,>=1.4.0 in /home/ec2-user/anaconda3/envs/python3/lib/pyth
on3.10/site-packages (from sagemaker) (6.8.0)
Requirement already satisfied: packaging>=20.0 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-pac
kages (from sagemaker) (21.3)
Requirement already satisfied: pandas in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages (fr
om sagemaker) (2.0.3)
Requirement already satisfied: pathos in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages (fr
om sagemaker) (0.3.1)
Requirement already satisfied: schema in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages (fr
om sagemaker) (0.7.5)
Requirement already satisfied: PyYAML~=6.0 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-package
s (from sagemaker) (6.0)
Requirement already satisfied: jsonschema in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages
(from sagemaker) (4.18.4)
Requirement already satisfied: platformdirs in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packag
es (from sagemaker) (3.9.1)
Requirement already satisfied: tblib==1.7.0 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packag
es (from sagemaker) (1.7.0)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/s
ite-packages (from botocore) (1.0.1)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /home/ec2-user/anaconda3/envs/python3/lib/python
3.10/site-packages (from botocore) (2.8.2)
Requirement already satisfied: urllib3<2.1,>=1.25.4 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/sit
e-packages (from botocore) (1.26.14)
Requirement already satisfied: s3transfer<0.8.0,>=0.7.0 in /home/ec2-user/anaconda3/envs/python3/lib/python3.1
0/site-packages (from boto3) (0.7.0)
Requirement already satisfied: docutils<0.17,>=0.10 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/sit
e-packages (from awscli) (0.16)
Requirement already satisfied: colorama<0.4.5,>=0.2.5 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/s
ite-packages (from awscli) (0.4.4)
Requirement already satisfied: rsa<4.8,>=3.1.2 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-pac
kages (from awscli) (4.7.2)
Requirement already satisfied: zipp>=0.5 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages
(from importlib-metadata<7.0,>=1.4.0->sagemaker) (3.16.2)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /home/ec2-user/anaconda3/envs/python3/lib/python3.1
0/site-packages (from packaging>=20.0->sagemaker) (3.0.9)
```

```
Requirement already satisfied: six>=1.5 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages
(from python-dateutil<3.0.0,>=2.1->botocore) (1.16.0)
Requirement already satisfied: pyasn1>=0.1.3 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packa
ges (from rsa<4.8,>=3.1.2->awscli) (0.5.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /home/ec2-user/anaconda3/envs/python3/li
b/python3.10/site-packages (from jsonschema->sagemaker) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site
-packages (from jsonschema->sagemaker) (0.30.0)
Requirement already satisfied: rpds-py>=0.7.1 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-pack
ages (from jsonschema->sagemaker) (0.9.2)
Requirement already satisfied: pytz>=2020.1 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packag
es (from pandas->sagemaker) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-pack
ages (from pandas->sagemaker) (2023.3)
Requirement already satisfied: ppft>=1.7.6.7 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packa
ges (from pathos->sagemaker) (1.7.6.7)
Requirement already satisfied: dill>=0.3.7 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-package
s (from pathos->sagemaker) (0.3.7)
Requirement already satisfied: pox>=0.3.3 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages
(from pathos->sagemaker) (0.3.3)
Requirement already satisfied: multiprocess>=0.70.15 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/si
te-packages (from pathos->sagemaker) (0.70.15)
Requirement already satisfied: contextlib2>=0.5.5 in /home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-
packages (from schema->sagemaker) (21.6.0)
Downloading botocore-1.31.78-py3-none-any.whl (11.3 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 11.3/11.3 MB 72.0 MB/s eta 0:00:00:00:010:01
Downloading boto3-1.28.78-py3-none-any.whl (135 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 135.8/135.8 kB 21.6 MB/s eta 0:00:00
Downloading awscli-1.29.78-py3-none-any.whl (4.3 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 4.3/4.3 MB 88.8 MB/s eta 0:00:00:00:01
Building wheels for collected packages: sagemaker
  Building wheel for sagemaker (setup.py) ... done
  Created wheel for sagemaker: filename=sagemaker-2.196.0-py2.py3-none-any.whl size=1223208 sha256=7631e4223651
abbcb64e3f291bc7dbf07ee2a98bb8e6e5f5d34703463d67789f
  Stored in directory: /home/ec2-user/.cache/pip/wheels/22/86/1b/11b1150764a78929af99b11b7789b8f3ed340d2c31d425
cfe2
Successfully built sagemaker
Installing collected packages: botocore, boto3, awscli, sagemaker
  Attempting uninstall: botocore
    Found existing installation: botocore 1.31.57
    Uninstalling botocore-1.31.57:
      Successfully uninstalled botocore-1.31.57
  Attempting uninstall: boto3
    Found existing installation: boto3 1.28.57
    Uninstalling boto3-1.28.57:
      Successfully uninstalled boto3-1.28.57
  Attempting uninstall: awscli
    Found existing installation: awscli 1.29.57
    Uninstalling awscli-1.29.57:
      Successfully uninstalled awscli-1.29.57
  Attempting uninstall: sagemaker
    Found existing installation: sagemaker 2.188.0
    Uninstalling sagemaker-2.188.0:
      Successfully uninstalled sagemaker-2.188.0
Successfully installed awscli-1.29.78 boto3-1.28.78 botocore-1.31.78 sagemaker-2.196.0
```

In [2]:

```
# Setup clients
import boto3

client = boto3.client(service_name="sagemaker")
runtime = boto3.client(service_name="sagemaker-runtime")
```

To begin, we import the AWS SDK for Python (Boto3) and set up our environment, including an IAM role and an S3 bucket to store our data.

In [3]:

```
import boto3
import sagemaker
from sagemaker.estimator import Estimator

boto_session = boto3.session.Session()
region = boto_session.region_name
```

```
print(region)

sagemaker_session = sagemaker.Session()
base_job_prefix = "xgboost-example"
role = sagemaker.get_execution_role()
print(role)

default_bucket = 'day07-mk1'
s3_prefix = base_job_prefix

training_instance_type = "ml.m5.xlarge"
```

```
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
us-east-1
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
arn:aws:iam::239630988601:role/fast-ai-academic-1-Student-Azure
```

Retrieve the Abalone dataset from a publicly hosted S3 bucket.

In [4]:
```
# retrieve data
! curl https://sagemaker-sample-files.s3.amazonaws.com/datasets/tabular/uci_abalone/train_csv/abalone_dataset1_t
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  131k  100  131k    0     0  1004k      0 --:--:-- --:--:-- --:--:-- 1008k
```

Upload the Abalone dataset to the default S3 bucket.

In [5]:
```
# upload data to S3
!aws s3 cp abalone_dataset1_train.csv s3://{default_bucket}/xgboost-regression/train.csv
```

```
upload: ./abalone_dataset1_train.csv to s3://day07-mk1/xgboost-regression/train.csv
```

Model Training

Now, we train an ML model using the XGBoost Algorithm. In this example, we use a SageMaker-provided XGBoost container image and configure an estimator to train our model.

In [6]:
```
from sagemaker.inputs import TrainingInput

training_path = f"s3://{default_bucket}/xgboost-regression/train.csv"
train_input = TrainingInput(training_path, content_type="text/csv")
```

In [7]:
```
model_path = f"s3://{default_bucket}/{s3_prefix}/xgb_model"

# retrieve xgboost image
image_uri = sagemaker.image_uris.retrieve(
    framework="xgboost",
    region=region,
    version="1.0-1",
    py_version="py3",
    instance_type=training_instance_type,
)

# Configure Training Estimator
xgb_train = Estimator(
    image_uri=image_uri,
    instance_type=training_instance_type,
    instance_count=1,
    output_path=model_path,
    sagemaker_session=sagemaker_session,
    role=role,
)

# Set Hyperparameters
xgb_train.set_hyperparameters(
    objective="reg:linear",
```

```
    num_round=50,
    max_depth=5,
    eta=0.2,
    gamma=4,
    min_child_weight=6,
    subsample=0.7,
    silent=0,
)
```

Train the model on the Abalone dataset.

In [8]:
```
# Fit model
xgb_train.fit({"train": train_input})
```

```
INFO:sagemaker:Creating training-job with name: sagemaker-xgboost-2023-11-05-16-42-21-650
```

```
2023-11-05 16:42:21 Starting - Starting the training job...
2023-11-05 16:42:37 Starting - Preparing the instances for training......
2023-11-05 16:43:32 Downloading - Downloading input data...
2023-11-05 16:44:08 Training - Downloading the training image..[2023-11-05 16:44:34.410 ip-10-0-118-156.ec2.int
ernal:7 INFO utils.py:27] RULE_JOB_STOP_SIGNAL_FILENAME: None
INFO:sagemaker-containers:Imported framework sagemaker_xgboost_container.training
INFO:sagemaker-containers:Failed to parse hyperparameter objective value reg:linear to Json.
Returning the value itself
INFO:sagemaker-containers:No GPUs detected (normal if no gpus installed)
INFO:sagemaker_xgboost_container.training:Running XGBoost Sagemaker in algorithm mode
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Single node training.
[16:44:34] 2923x8 matrix with 23384 entries loaded from /opt/ml/input/data/train?format=csv&label_column=0&deli
miter=,
[2023-11-05 16:44:34.479 ip-10-0-118-156.ec2.internal:7 INFO json_config.py:91] Creating hook from json_config
at /opt/ml/input/config/debughookconfig.json.
[2023-11-05 16:44:34.480 ip-10-0-118-156.ec2.internal:7 INFO hook.py:201] tensorboard_dir has not been set for
the hook. SMDebug will not be exporting tensorboard summaries.
[2023-11-05 16:44:34.480 ip-10-0-118-156.ec2.internal:7 INFO profiler_config_parser.py:102] User has disabled p
rofiler.
[2023-11-05 16:44:34.481 ip-10-0-118-156.ec2.internal:7 INFO hook.py:255] Saving to /opt/ml/output/tensors
[2023-11-05 16:44:34.481 ip-10-0-118-156.ec2.internal:7 INFO state_store.py:77] The checkpoint config file /op
t/ml/input/config/checkpointconfig.json does not exist.
INFO:root:Debug hook created from config
INFO:root:Train matrix has 2923 rows
[16:44:34] WARNING: /workspace/src/objective/regression_obj.cu:167: reg:linear is now deprecated in favor of re
g:squarederror.
[16:44:34] WARNING: /workspace/src/learner.cc:328:
Parameters: { num_round, silent } might not be used.
  This may not be accurate due to some parameters are only used in language bindings but
  passed down to XGBoost core.  Or some parameters are not used but slip through this
  verification. Please open an issue if you find above cases.
[0]#011train-rmse:8.09123
[2023-11-05 16:44:34.497 ip-10-0-118-156.ec2.internal:7 INFO hook.py:423] Monitoring the collections: metrics
[2023-11-05 16:44:34.500 ip-10-0-118-156.ec2.internal:7 INFO hook.py:486] Hook is writing from the hook with pi
d: 7
[1]#011train-rmse:6.61298
[2]#011train-rmse:5.45157
[3]#011train-rmse:4.54038
[4]#011train-rmse:3.84707
[5]#011train-rmse:3.31465
[6]#011train-rmse:2.91636
[7]#011train-rmse:2.62415
[8]#011train-rmse:2.40885
[9]#011train-rmse:2.24929
[10]#011train-rmse:2.13106
[11]#011train-rmse:2.04974
[12]#011train-rmse:1.98240
[13]#011train-rmse:1.93888
[14]#011train-rmse:1.89701
[15]#011train-rmse:1.87329
[16]#011train-rmse:1.85216
[17]#011train-rmse:1.82408
[18]#011train-rmse:1.81372
[19]#011train-rmse:1.80362
[20]#011train-rmse:1.78164
[21]#011train-rmse:1.77341
[22]#011train-rmse:1.76766
[23]#011train-rmse:1.75940
[24]#011train-rmse:1.74632
[25]#011train-rmse:1.74385
[26]#011train-rmse:1.73876
[27]#011train-rmse:1.73410
[28]#011train-rmse:1.72847
[29]#011train-rmse:1.72384
[30]#011train-rmse:1.71492
[31]#011train-rmse:1.69789
[32]#011train-rmse:1.69073
[33]#011train-rmse:1.68621
[34]#011train-rmse:1.67960
[35]#011train-rmse:1.67194
```

```
[36]#011train-rmse:1.65883
[37]#011train-rmse:1.65463
[38]#011train-rmse:1.65199
[39]#011train-rmse:1.63903
[40]#011train-rmse:1.63353
[41]#011train-rmse:1.62607
[42]#011train-rmse:1.61662
[43]#011train-rmse:1.60241
[44]#011train-rmse:1.59173
[45]#011train-rmse:1.58875
[46]#011train-rmse:1.57816
[47]#011train-rmse:1.56941
[48]#011train-rmse:1.56063
[49]#011train-rmse:1.55822

2023-11-05 16:44:51 Uploading - Uploading generated training model
2023-11-05 16:44:51 Completed - Training job completed
Training seconds: 77
Billable seconds: 77
```

Deployment

After training the model, retrieve the model artifacts so that we can deploy the model to an endpoint.

In [9]:
```
# Retrieve model data from training job
model_artifacts = xgb_train.model_data
model_artifacts
```

Out[9]:
```
's3://day07-mk1/xgboost-example/xgb_model/sagemaker-xgboost-2023-11-05-16-42-21-650/output/model.tar.gz'
```

Model Creation

Create a model by providing your model artifacts, the container image URI, environment variables for the container (if applicable), a model name, and the SageMaker IAM role.

In [10]:
```
from time import gmtime, strftime

model_name = "xgboost-serverless" + strftime("%Y-%m-%d-%H-%M-%S", gmtime())
print("Model name: " + model_name)

# dummy environment variables
byo_container_env_vars = {"SAGEMAKER_CONTAINER_LOG_LEVEL": "20", "SOME_ENV_VAR": "myEnvVar"}

create_model_response = client.create_model(
    ModelName=model_name,
    Containers=[
        {
            "Image": image_uri,
            "Mode": "SingleModel",
            "ModelDataUrl": model_artifacts,
            "Environment": byo_container_env_vars,
        }
    ],
    ExecutionRoleArn=role,
)

print("Model Arn: " + create_model_response["ModelArn"])
```

```
Model name: xgboost-serverless2023-11-05-16-46-38
Model Arn: arn:aws:sagemaker:us-east-1:239630988601:model/xgboost-serverless2023-11-05-16-46-38
```

Endpoint Configuration Creation

This is where you can adjust the **Serverless Configuration** for your endpoint. The current max concurrent invocations for a single endpoint, known as MaxConcurrency, can be any value from **1 to 200**, and MemorySize can be any of the following: **1024 MB, 2048 MB, 3072 MB, 4096 MB, 5120 MB, or 6144 MB.**

In [13]:
```
xgboost_epc_name = "xgboost-serverless-epc" + strftime("%Y-%m-%d-%H-%M-%S", gmtime())

endpoint_config_response = client.create_endpoint_config(
    EndpointConfigName=xgboost_epc_name,
    ProductionVariants=[
        {
            "VariantName": "byoVariant",
            "ModelName": model_name,
            "ServerlessConfig": {
                "MemorySizeInMB": 3072,
                "MaxConcurrency": 1,
            },
        },
    ],
)

print("Endpoint Configuration Arn: " + endpoint_config_response["EndpointConfigArn"])
```

```
Endpoint Configuration Arn: arn:aws:sagemaker:us-east-1:239630988601:endpoint-config/xgboost-serverless-epc2023
-11-05-16-47-46
```

Serverless Endpoint Creation

Now that we have an endpoint configuration, we can create a **serverless endpoint** and deploy our model to it. When creating the endpoint, provide the name of your endpoint configuration and a name for the new endpoint.

In [14]:
```
endpoint_name = "xgboost-serverless-ep" + strftime("%Y-%m-%d-%H-%M-%S", gmtime())

create_endpoint_response = client.create_endpoint(
    EndpointName=endpoint_name,
    EndpointConfigName=xgboost_epc_name,
)

print("Endpoint Arn: " + create_endpoint_response["EndpointArn"])
```

```
Endpoint Arn: arn:aws:sagemaker:us-east-1:239630988601:endpoint/xgboost-serverless-ep2023-11-05-16-47-47
```

Wait until the endpoint status is InService before invoking the endpoint.

In [15]:
```
# wait for endpoint to reach a terminal state (InService) using describe endpoint
import time

describe_endpoint_response = client.describe_endpoint(EndpointName=endpoint_name)

while describe_endpoint_response["EndpointStatus"] == "Creating":
    describe_endpoint_response = client.describe_endpoint(EndpointName=endpoint_name)
    print(describe_endpoint_response["EndpointStatus"])
    time.sleep(15)

describe_endpoint_response
```

```
Creating
Creating
Creating
Creating
Creating
Creating
Creating
InService
```

Out[15]:

```
{'EndpointName': 'xgboost-serverless-ep2023-11-05-16-47-47',
 'EndpointArn': 'arn:aws:sagemaker:us-east-1:239630988601:endpoint/xgboost-serverless-ep2023-11-05-16-47-47',
 'EndpointConfigName': 'xgboost-serverless-epc2023-11-05-16-47-46',
 'ProductionVariants': [{'VariantName': 'byoVariant',
   'DeployedImages': [{'SpecifiedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.0-1-cpu-py3',
     'ResolvedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost@sha256:1b38de0dbca9bfe90990430d0c4639fe5660c5dbc3964e8af33621c3b07a5a97',
     'ResolutionTime': datetime.datetime(2023, 11, 5, 16, 47, 49, 242000, tzinfo=tzlocal())}],
   'CurrentWeight': 1.0,
   'DesiredWeight': 1.0,
   'CurrentInstanceCount': 0,
   'CurrentServerlessConfig': {'MemorySizeInMB': 3072, 'MaxConcurrency': 1}}],
 'EndpointStatus': 'InService',
 'CreationTime': datetime.datetime(2023, 11, 5, 16, 47, 48, 433000, tzinfo=tzlocal()),
 'LastModifiedTime': datetime.datetime(2023, 11, 5, 16, 50, 28, 710000, tzinfo=tzlocal()),
 'ResponseMetadata': {'RequestId': '6ea76a18-2957-4780-b21a-bfdb01a061fa',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amzn-requestid': '6ea76a18-2957-4780-b21a-bfdb01a061fa',
   'content-type': 'application/x-amz-json-1.1',
   'content-length': '817',
   'date': 'Sun, 05 Nov 2023 16:50:35 GMT'},
  'RetryAttempts': 0}}
```


Endpoint Invocation

Invoke the endpoint by sending a request to it. The following is a sample data point grabbed from the CSV file downloaded from the public Abalone dataset.

In [16]:
```
%%time
response = runtime.invoke_endpoint(
    EndpointName=endpoint_name,
    Body=b".345,0.224414,.131102,0.042329,.279923,-0.110329,-0.099358,0.0",
    ContentType="text/csv",
)

print(response["Body"].read())
# container is not running, it will be created after invoking
```

```
b'4.566554546356201'
CPU times: user 26.2 ms, sys: 0 ns, total: 26.2 ms
Wall time: 179 ms
```

Let's review the above output : CPU times: user 36.8 ms, sys: 0 ns, total: 36.8 ms

user 36.8 ms ---> that shows the amount of time spent executing user-level code, which is code that is written in a higher-level language (in this case Python).

sys 0 ns --> that shows the amount of time spent executing system-level code, which is code that interacts directly with the operating system. In this case is very small value that is considered as 0.

total 36.8 --> The total amount of CPU time spent executing the code (i.e., the sum of user and sys values).

Wall time: 326 ms --> the wall-clock time is the elapsed time between the start and end of the code block, as measured by a clock on the wall (i.e., a real-world clock). This includes time spent waiting for input/output operations, sleep calls, or other types ofactions or blocking calls that may not consume any CPU time but still takes time.

In this example, the wall time was 326 ms, which is significantly longer than the CPU time of 36.8 ms. This indicates that the code spent a significant amount of time waiting for other things that as you are aware thar is becuase no container was there before the first request come.

Assignment:

1- Now that you learn how to get the time spent to get the reponse, go back to that real-time notebook and compare the CPU time and Wall time of real time with serverless and explain why they are different.

- Realtime Notebook - CPU times: user 28.2 ms, sys: 0 ns, total: 28.2 ms, Wall time: 178 ms
- Serverless Notebook (The above cell) - CPU times: user 26.2 ms, sys: 0 ns, total: 26.2 ms, Wall time: 179 ms

Realtime(user=28.2ms) > Serverless(user=26.2ms) - it means that the amount of time spent executing user-level code was greater in realtime notebook as compared to serverless.

Realtime(sys=0ns) = Serverless(user=0ns) - it means that the amount of time spent executing system-level code, which is code that interacts directly with the operating system, was very small that is considered as 0 in both notebooks.

Realtime(walltime=178ms) < Serverless(user=179ms) - it means that the cell in realtime notebook includes time spent waiting for input/output operations, sleep calls, or other types of actions or blocking calls that may not consume any CPU time but still takes time was approximately same as serverless notebook.

2- Use the above cell to re-send an inferencing request to the serverless endpoint immidiately after the first request. Do you see an improvmen in the wall time. If yes why?

- Serverless Notebook (Before)- CPU times: user 26.2 ms, sys: 0 ns, total: 26.2 ms, Wall time: 180 ms
- Serverless Notebook (After)- CPU times: user 0 ns, sys: 4.2 ms, total: 4.2 ms, Wall time: 124 ms

Why? Because When we run the cell for the first time, it establishes a connection to the serverless endpoint, sends the request, and receives the response. Subsequent runs may benefit from potential optimizations like connection reuse or endpoint warm-up, which results in a potentially faster execution time. If the serverless environment caches certain resources or computations, it may provide faster responses for subsequent, similar requests.

Please delete the endpoint in the console before you close this session

In [ ]:

In [ ]:
```
! pip install sagemaker botocore boto3 awscli --upgrade
```

In [17]:
```
# Setup clients
import boto3

client = boto3.client(service_name="sagemaker")
runtime = boto3.client(service_name="sagemaker-runtime")
```

In [18]:
```
import sagemaker
from sagemaker.estimator import Estimator

boto_session = boto3.session.Session()
region = boto_session.region_name
print(region)

sagemaker_session = sagemaker.Session()
base_job_prefix = "xgboost-example"
role = sagemaker.get_execution_role()
print(role)

default_bucket = 'day07-mk1-realtime-serverless'
s3_prefix = base_job_prefix

training_instance_type = "ml.m5.xlarge"
```

```
us-east-1
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
arn:aws:iam::585522057818:role/fast-ai-academic-11-Student-Azure
```

In [19]:
```
# retrieve data
! curl https://sagemaker-sample-files.s3.amazonaws.com/datasets/tabular/uci_abalone/train_csv/abalone_dataset1_t
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  131k  100  131k    0     0  1088k      0 --:--:-- --:--:-- --:--:-- 1092k
```

In [20]:
```
# upload data to S3
!aws s3 cp abalone_dataset1_train.csv s3://{default_bucket}/xgboost-regression/train.csv
```

```
upload: ./abalone_dataset1_train.csv to s3://day07-mk1-realtime-serverless/xgboost-regression/train.csv
```

Modle Training

In [21]:
```
from sagemaker.inputs import TrainingInput

training_path = f"s3://{default_bucket}/xgboost-regression/train.csv"
train_input = TrainingInput(training_path, content_type="text/csv")
```

In [22]:
```
model_path = f"s3://{default_bucket}/{s3_prefix}/xgb_model"

# retrieve xgboost image
image_uri = sagemaker.image_uris.retrieve(
    framework="xgboost",
    region=region,
    version="1.0-1",
    py_version="py3",
    instance_type=training_instance_type,
)

# Configure Training Estimator
xgb_train = Estimator(
    image_uri=image_uri,
    instance_type=training_instance_type,
    instance_count=1,
    output_path=model_path,
    sagemaker_session=sagemaker_session,
    role=role,
)

# Set Hyperparameters
xgb_train.set_hyperparameters(
    objective="reg:linear",
    num_round=50,
    max_depth=5,
    eta=0.2,
    gamma=4,
    min_child_weight=6,
    subsample=0.7,
    silent=0,
)
```

In [23]:
```
# Fit model
xgb_train.fit({"train": train_input})
```

```
INFO:sagemaker:Creating training-job with name: sagemaker-xgboost-2023-11-14-06-11-34-110
```

```
2023-11-14 06:11:34 Starting - Starting the training job...
2023-11-14 06:11:48 Starting - Preparing the instances for training......
2023-11-14 06:12:54 Downloading - Downloading input data...
2023-11-14 06:13:34 Training - Downloading the training image...
2023-11-14 06:13:59 Uploading - Uploading generated training model[2023-11-14 06:13:55.349 ip-10-2-246-118.ec2.
internal:7 INFO utils.py:27] RULE_JOB_STOP_SIGNAL_FILENAME: None
INFO:sagemaker-containers:Imported framework sagemaker_xgboost_container.training
INFO:sagemaker-containers:Failed to parse hyperparameter objective value reg:linear to Json.
Returning the value itself
INFO:sagemaker-containers:No GPUs detected (normal if no gpus installed)
INFO:sagemaker_xgboost_container.training:Running XGBoost Sagemaker in algorithm mode
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Single node training.
[06:13:55] 2923x8 matrix with 23384 entries loaded from /opt/ml/input/data/train?format=csv&label_column=0&deli
miter=,
[2023-11-14 06:13:55.426 ip-10-2-246-118.ec2.internal:7 INFO json_config.py:91] Creating hook from json_config
at /opt/ml/input/config/debughookconfig.json.
[2023-11-14 06:13:55.427 ip-10-2-246-118.ec2.internal:7 INFO hook.py:201] tensorboard_dir has not been set for
the hook. SMDebug will not be exporting tensorboard summaries.
[2023-11-14 06:13:55.427 ip-10-2-246-118.ec2.internal:7 INFO profiler_config_parser.py:102] User has disabled p
rofiler.
[2023-11-14 06:13:55.428 ip-10-2-246-118.ec2.internal:7 INFO hook.py:255] Saving to /opt/ml/output/tensors
[2023-11-14 06:13:55.428 ip-10-2-246-118.ec2.internal:7 INFO state_store.py:77] The checkpoint config file /op
t/ml/input/config/checkpointconfig.json does not exist.
INFO:root:Debug hook created from config
INFO:root:Train matrix has 2923 rows
[06:13:55] WARNING: /workspace/src/objective/regression_obj.cu:167: reg:linear is now deprecated in favor of re
g:squarederror.
[06:13:55] WARNING: /workspace/src/learner.cc:328:
Parameters: { num_round, silent } might not be used.
  This may not be accurate due to some parameters are only used in language bindings but
  passed down to XGBoost core.  Or some parameters are not used but slip through this
  verification. Please open an issue if you find above cases.
[0]#011train-rmse:8.09123
[2023-11-14 06:13:55.433 ip-10-2-246-118.ec2.internal:7 INFO hook.py:423] Monitoring the collections: metrics
[2023-11-14 06:13:55.435 ip-10-2-246-118.ec2.internal:7 INFO hook.py:486] Hook is writing from the hook with pi
d: 7
[1]#011train-rmse:6.61298
[2]#011train-rmse:5.45157
[3]#011train-rmse:4.54038
[4]#011train-rmse:3.84707
[5]#011train-rmse:3.31465
[6]#011train-rmse:2.91636
[7]#011train-rmse:2.62415
[8]#011train-rmse:2.40885
[9]#011train-rmse:2.24929
[10]#011train-rmse:2.13106
[11]#011train-rmse:2.04974
[12]#011train-rmse:1.98240
[13]#011train-rmse:1.93888
[14]#011train-rmse:1.89701
[15]#011train-rmse:1.87329
[16]#011train-rmse:1.85216
[17]#011train-rmse:1.82408
[18]#011train-rmse:1.81372
[19]#011train-rmse:1.80362
[20]#011train-rmse:1.78164
[21]#011train-rmse:1.77341
[22]#011train-rmse:1.76766
[23]#011train-rmse:1.75940
[24]#011train-rmse:1.74632
[25]#011train-rmse:1.74385
[26]#011train-rmse:1.73876
[27]#011train-rmse:1.73410
[28]#011train-rmse:1.72847
[29]#011train-rmse:1.72384
[30]#011train-rmse:1.71492
[31]#011train-rmse:1.69789
[32]#011train-rmse:1.69073
[33]#011train-rmse:1.68621
[34]#011train-rmse:1.67960
```

```
[35]#011train-rmse:1.67194
[36]#011train-rmse:1.65883
[37]#011train-rmse:1.65463
[38]#011train-rmse:1.65199
[39]#011train-rmse:1.63903
[40]#011train-rmse:1.63353
[41]#011train-rmse:1.62607
[42]#011train-rmse:1.61662
[43]#011train-rmse:1.60241
[44]#011train-rmse:1.59173
[45]#011train-rmse:1.58875
[46]#011train-rmse:1.57816
[47]#011train-rmse:1.56941
[48]#011train-rmse:1.56063
[49]#011train-rmse:1.55822

2023-11-14 06:14:16 Completed - Training job completed
Training seconds: 82
Billable seconds: 82
```

In [24]:
```
# Retrieve model data from training job
model_artifacts = xgb_train.model_data
model_artifacts
```

Out[24]:
```
's3://day07-mk1-realtime-serverless/xgboost-example/xgb_model/sagemaker-xgboost-2023-11-14-06-11-34-110/output/model.tar.gz'
```

In [25]:
```
from time import gmtime, strftime

model_name = "xgboost-serverless" + strftime("%Y-%m-%d-%H-%M-%S", gmtime())
print("Model name: " + model_name)

# dummy environment variables
byo_container_env_vars = {"SAGEMAKER_CONTAINER_LOG_LEVEL": "20", "SOME_ENV_VAR": "myEnvVar"}

create_model_response = client.create_model(
    ModelName=model_name,
    Containers=[
        {
            "Image": image_uri,
            "Mode": "SingleModel",
            "ModelDataUrl": model_artifacts,
            "Environment": byo_container_env_vars,
        }
    ],
    ExecutionRoleArn=role,
)

print("Model Arn: " + create_model_response["ModelArn"])
```

```
Model name: xgboost-serverless2023-11-14-06-17-35
Model Arn: arn:aws:sagemaker:us-east-1:585522057818:model/xgboost-serverless2023-11-14-06-17-35
```

Endpoint

In [26]:
```
xgboost_epc_name = "xgboost-serverless-epc" + strftime("%Y-%m-%d-%H-%M-%S", gmtime())

endpoint_config_response = client.create_endpoint_config(
    EndpointConfigName=xgboost_epc_name,
    ProductionVariants=[
        {
            "VariantName": "byoVariant",
            "ModelName": model_name,
            "ServerlessConfig": {
                "MemorySizeInMB": 3072,
                "MaxConcurrency": 1,
            },
        },
    ],
)
```

```
print("Endpoint Configuration Arn: " + endpoint_config_response["EndpointConfigArn"])
```

```
Endpoint Configuration Arn: arn:aws:sagemaker:us-east-1:585522057818:endpoint-config/xgboost-serverless-epc2023
-11-14-06-17-38
```

In [27]:
```
endpoint_name = "xgboost-serverless-ep" + strftime("%Y-%m-%d-%H-%M-%S", gmtime())

create_endpoint_response = client.create_endpoint(
    EndpointName=endpoint_name,
    EndpointConfigName=xgboost_epc_name,
)

print("Endpoint Arn: " + create_endpoint_response["EndpointArn"])
```

```
Endpoint Arn: arn:aws:sagemaker:us-east-1:585522057818:endpoint/xgboost-serverless-ep2023-11-14-06-17-39
```

In [28]:
```
# wait for endpoint to reach a terminal state (InService) using describe endpoint
import time

describe_endpoint_response = client.describe_endpoint(EndpointName=endpoint_name)

while describe_endpoint_response["EndpointStatus"] == "Creating":
    describe_endpoint_response = client.describe_endpoint(EndpointName=endpoint_name)
    print(describe_endpoint_response["EndpointStatus"])
    time.sleep(15)

describe_endpoint_response
```

```
Creating
Creating
Creating
Creating
Creating
Creating
Creating
Creating
Creating
Creating
Creating
InService
```

Out[28]:
```
{'EndpointName': 'xgboost-serverless-ep2023-11-14-06-17-39',
 'EndpointArn': 'arn:aws:sagemaker:us-east-1:585522057818:endpoint/xgboost-serverless-ep2023-11-14-06-17-39',
 'EndpointConfigName': 'xgboost-serverless-epc2023-11-14-06-17-38',
 'ProductionVariants': [{'VariantName': 'byoVariant',
   'DeployedImages': [{'SpecifiedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.0-1-
cpu-py3',
     'ResolvedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost@sha256:be62b677e1dcac42c0
333b10e0eb0b0c2aecd1f6a2ac08db68ae0235494f0de3',
     'ResolutionTime': datetime.datetime(2023, 11, 14, 6, 17, 40, 870000, tzinfo=tzlocal())}],
   'CurrentWeight': 1.0,
   'DesiredWeight': 1.0,
   'CurrentInstanceCount': 0,
   'CurrentServerlessConfig': {'MemorySizeInMB': 3072, 'MaxConcurrency': 1}}],
 'EndpointStatus': 'InService',
 'CreationTime': datetime.datetime(2023, 11, 14, 6, 17, 40, 27000, tzinfo=tzlocal()),
 'LastModifiedTime': datetime.datetime(2023, 11, 14, 6, 20, 25, 837000, tzinfo=tzlocal()),
 'ResponseMetadata': {'RequestId': '074c9cc2-4fd1-464d-9863-58f369c6ab77',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amzn-requestid': '074c9cc2-4fd1-464d-9863-58f369c6ab77',
   'content-type': 'application/x-amz-json-1.1',
   'content-length': '817',
   'date': 'Tue, 14 Nov 2023 06:20:29 GMT'},
  'RetryAttempts': 0}}
```

Enpoint invocation

In [30]:
```
%%time
response = runtime.invoke_endpoint(
    EndpointName=endpoint_name,
    Body=b".345,0.224414,.131102,0.042329,.279923,-0.110329,-0.099358,0.0",
    ContentType="text/csv",
```

```
)

print(response["Body"].read())
```

```
b'4.566554546356201'
CPU times: user 0 ns, sys: 4.2 ms, total: 4.2 ms
Wall time: 124 ms
```

In [31]:
```
client.delete_endpoint(EndpointName=endpoint_name)
```

Out[31]:
```
{'ResponseMetadata': {'RequestId': 'fd5eb57c-d9f1-49d2-92aa-07df78a16037',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amzn-requestid': 'fd5eb57c-d9f1-49d2-92aa-07df78a16037',
   'content-type': 'application/x-amz-json-1.1',
   'content-length': '0',
   'date': 'Tue, 14 Nov 2023 06:21:53 GMT'},
  'RetryAttempts': 0}}
```

In [ ]:

In [ ]: