Chapter 7. MLOps for AWS

By Noah Gift

> *Everybody was scared of him [Dr. Abbott (because he yelled at everyone)]. When I was attending, there was a new resident named Harris. Harris was still afraid of him [Dr. Abbott] even though he was the chief resident and had been there for 5 years. Later [Dr. Abbott] has a heart attack and then his heart stops. A nurse yells, "Quick, he just had an arrest, come in!" So Harris went in there…leaning on the sternum and breaking ribs and stuff. So Harris starts pumping on Abbott and he woke up. He woke up! And he looked up at Harris and he said, "You! STOP THAT!" So Harris stopped. And that was the last story about Abbott.*
>
> —Dr. Joseph Bogen

One of the most common questions I get from students is, "which cloud do I pick?" I tell them the safe choice is Amazon. It has the widest selection of technology and the largest market share. Once you master the AWS cloud, it is easier to master other cloud offerings since they also assume you might know AWS. This chapter covers the foundations of AWS for MLOps and explores practical MLOps patterns.

I have a long, rich history of working with AWS. At a sports social network I ran as the CTO and General Manager, AWS was a secret ingredient that allowed us to scale to millions of users worldwide. I also worked as an SME (subject matter expert) on the AWS Machine Learning Certification from scratch in the last several years. I have been recognized as an AWS ML Hero, I am also a part of the AWS Faculty Cloud Ambassador program, and have taught thousands of students cloud certifications at UC Davis, Northwestern, Duke, and the University of Tennessee. So you can say I am a fan of AWS!

Due to the sheer size of the AWS platform, it is impossible to cover every single aspect of MLOps. If you want a more exhaustive coverage of all of the available AWS platform options, you can also check out *Data Science*

on AWS by Chris Fregly and Antje Barth (O'Reilly), for which I was one of the technical editors.

Instead, this chapter focuses on higher-level services like AWS Lambda and AWS App Runner. More complicated systems like AWS SageMaker are covered in other chapters in this and the previously mentioned book. Next, let's get started building AWS MLOps solutions.

Introduction to AWS

AWS is the leader in cloud computing for several reasons, including its early start and corporate culture. In 2005 and 2006, Amazon launched Amazon Web Services, including MTurk (Mechanical Turk), Amazon S3, Amazon EC2, and Amazon SQS.

To this day, these core offerings are not only still around, but they get better by the quarter and year. The reason AWS continues to improve its offerings is due to its culture. They say it is always "Day 1" at Amazon, meaning the same energy and enthusiasm that happens on Day 1 should happen every day. They also place the customer at the "heart of everything" they do. I have used these building blocks to build systems scaled to thousands of nodes to do machine learning, computer vision, and AI tasks. Figure 7-1 is an actual whiteboard drawing of a working AWS Cloud Technical Architecture shown in a coworking spot in downtown San Francisco.

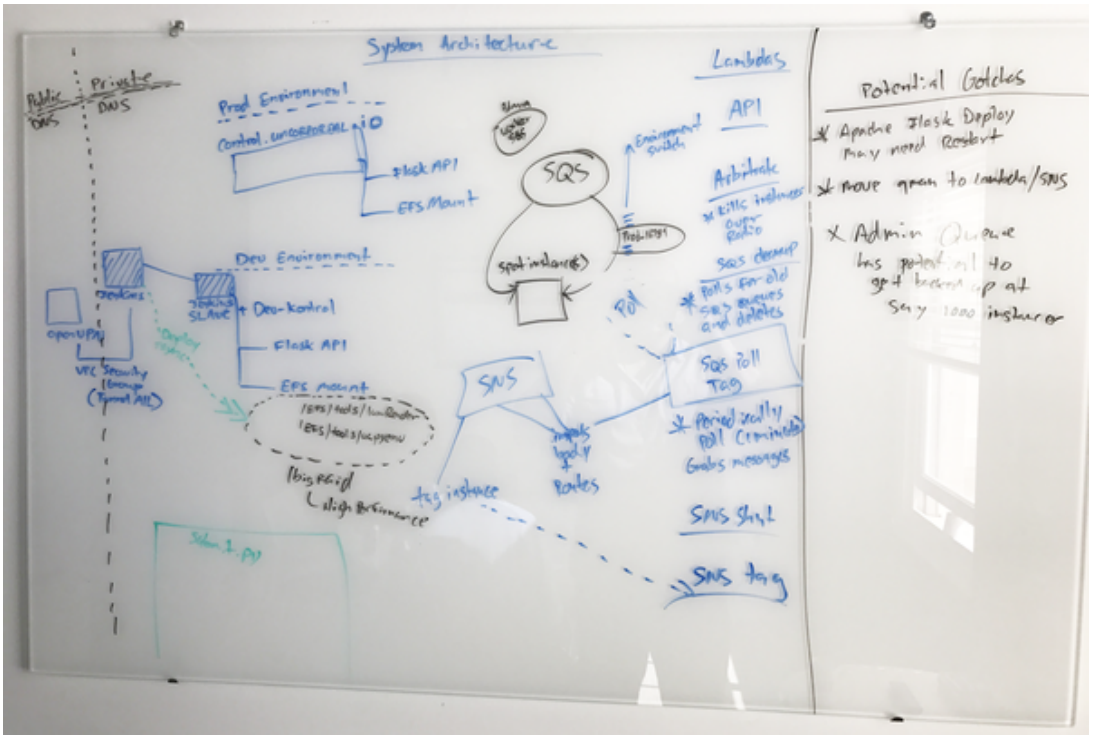

Figure 7-1. AWS Cloud Architecture on whiteboard

This type of culture is very different than Google, which has struggled in cloud computing. The Google culture is research-oriented with heavy academic hiring. The benefits of this culture are open source projects like Kubernetes and TensorFlow. Both are complex engineering marvels. The downside is that the customer is not first in the culture, which has hurt Google in the cloud computing market share. Many organizations balk at not buying professional services and supporting something as critical as cloud computing.

Next, let's take a look at getting started with AWS Services.

Getting Started with AWS Services

To get started with AWS initially requires only a Free Tier account. If you are at university, you can also use both AWS Academy and AWS Educate. AWS Academy provides hands-on certification material and labs. AWS Educate delivers sandboxes for classrooms.

Once you have an account, the next step is to experiment with the various offerings. One way to think about AWS is to compare it to a bulk wholesale store. For example, according to statista, Costco has 795 worldwide locations in multiple countries and approximately 1 in 5 Americans shop there. Similarly, according to the 2020 Amazon Web Services whitepaper,

AWS provides an infrastructure that “powers hundreds of thousands of businesses in 190 countries around the world.”

At Costco, there are three approaches considered relevant to an AWS analogy:

- A customer could walk in and order a very inexpensive but reasonable quality pizza in bulk in the first scenario.
- In a second scenario, a customer new to Costco needs to figure out all of the bulk items available to investigate the best way to use Costco. They will need to walk through the store and look at the bulk items like sugar to see what they could build from these raw ingredients.
- In a third scenario, when a customer of Costco knows about the premade meals (such as rotisserie chicken, pizza, and more), and they know about all of the raw ingredients they could buy, they could, in theory, start a catering company, a local deli, or a restaurant utilizing their knowledge of Costco and the services it provides.

A big box retailer the size of Costco charges more money for prepared food, and less money for the raw ingredients. The Costco customer who owns a restaurant may choose a different food preparation level depending on the maturity of their organization and the problem they aim to solve. Figure 7-2 illustrates how these Costco options compare to AWS options.

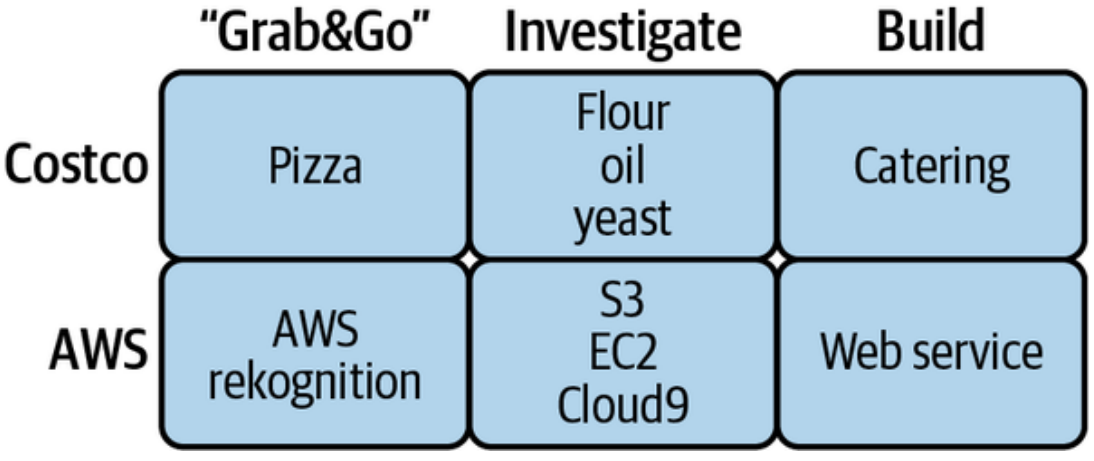

Figure 7-2. Costco versus AWS

Let’s take the case of a local Hawaii Poke Bowl stand near a famous Hawaii beach. The owner could buy Costco’s premade Poke in bulk and sell it at a price that is approximately twice their cost. But, on the other

hand, a more mature BBQ restaurant in Hawaii, with employees who can cook and prepare food, Costco sells this uncooked food at a much lower price than the fully prepared Poke.

Much like Costco, AWS provides different levels of product and it's up to the customer to decide how much they're going to utilize. Let's dig into those options a bit.

Using the "No Code/Low Code" AWS Comprehend solution

The last example showed how using Costco can benefit different restaurant businesses, from one with little to no staff—like a pop-up stand—to a more extensive sit-down facility. The more work Costco does in preparing the food, the higher the benefit to the customer purchasing it, and also the higher the cost for the food. The same concept applies with AWS; the more work AWS does for you, the more you pay and the fewer people you need to maintain the service.

NOTE

In economics, the theory of comparative advantage says that you shouldn't compare the cost of something directly. Instead, it would help if you compared the opportunity cost of doing it yourself. All cloud providers have this assumption baked in, since running a data center and building services on top of that data center is their specialization. An organization doing MLOps should focus on creating a product for its customers that generates revenue, not re-creating what cloud providers do poorly.

With AWS, an excellent place to start is to act like the Costco customer who orders Costco pizza in bulk. Likewise, a Fortune 500 company may have essential requirements to add natural language processing (NLP) to its customer service products. It could spend nine months to a year hiring a team and then building out these capabilities, or it could start using valuable high-level services like AWS Comprehend for Natural Language Processing. AWS Comprehend also enables users to leverage the Amazon API to perform many NLP operations, including the following:

- Entity detection
- Key phrase detection

- PII
- Language detection
- Sentiment

For example, you can cut and paste text into the Amazon Comprehend Console, and AWS Comprehend will find all of the text's entities. In the example in Figure 7-3, I grabbed the first paragraph of LeBron James's Wikipedia bio, pasted it into the console, clicked Analyze, and it highlighted the entities for me.

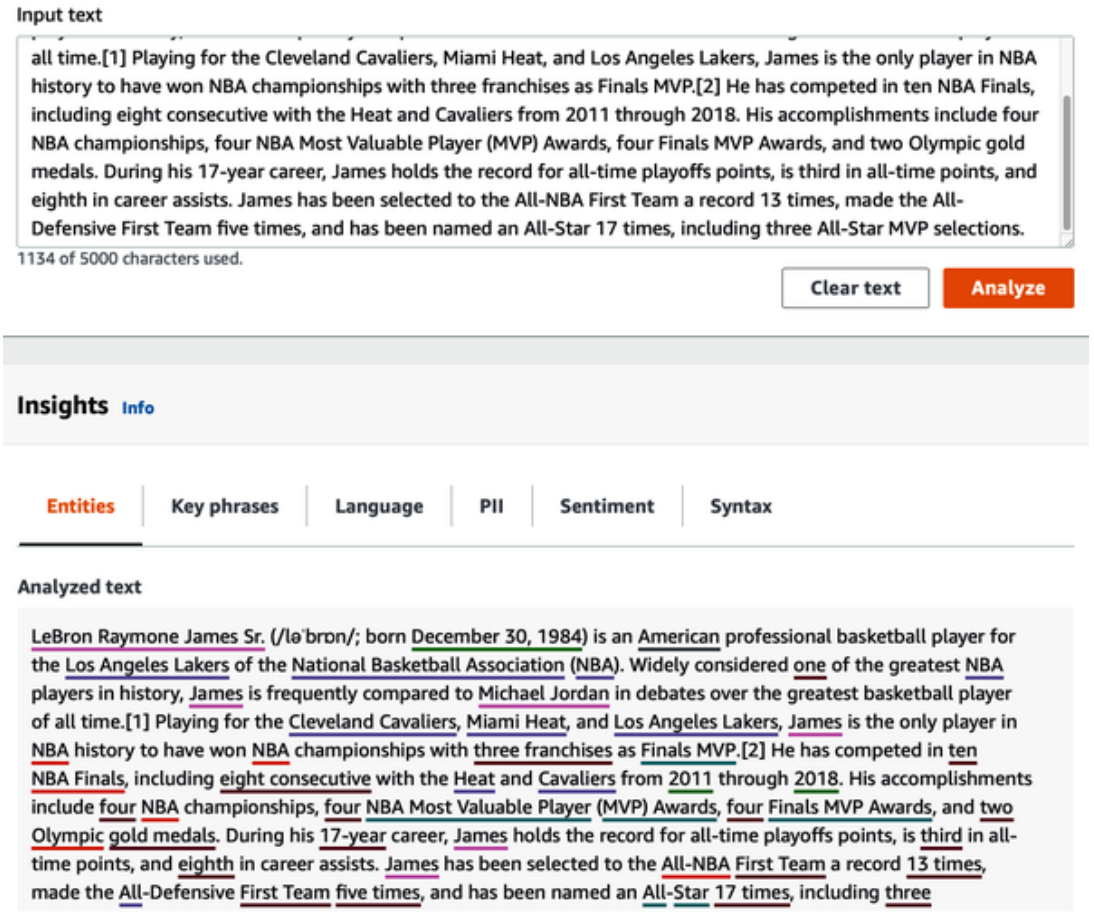

Figure 7-3. AWS Comprehend

Other use cases like reviewing medical records or determining the sentiment of a customer service response are equally straightforward with AWS Comprehend and the boto3 Python SDK. Next, let's cover a "hello world" project for DevOps on AWS, deploying a static website using Amazon S3.

Using Hugo static S3 websites

In the following scenario, an excellent way to explore AWS is to "walk around" the console, just as you would walk around Costco in awe the first time you visit. You do this by first looking at the foundational components of AWS, i.e., IaaS (infrastructure as code). These core services include AWS S3 object storage and AWS EC2 virtual machines.

Here, I'll walk you through deploying a Hugo website on AWS S3 static website hosting using "hello world" as an example. The reason for doing a hello world with Hugo is that it is relatively simple to set up and will give you a good understanding of host services using core infrastructure. These skills will come in handy later when you learn about deploying machine learning applications using continuous delivery.

NOTE

It is worth noting that Amazon S3 is low cost yet highly reliable. The pricing of S3 approaches a penny per GB. This low-cost yet highly reliable infrastructure is one of the reasons cloud computing is so compelling.

You can view the whole project in the GitHub repo. Notice how GitHub is the source of truth for the website because the entire project consists of text files: markdown files, the Hugo template, and the build commands for the AWS Code build server. Additionally, you can walk through a screencast of the continuous deployment there as well. Figure 7-4 shows the high-level architecture of this project.

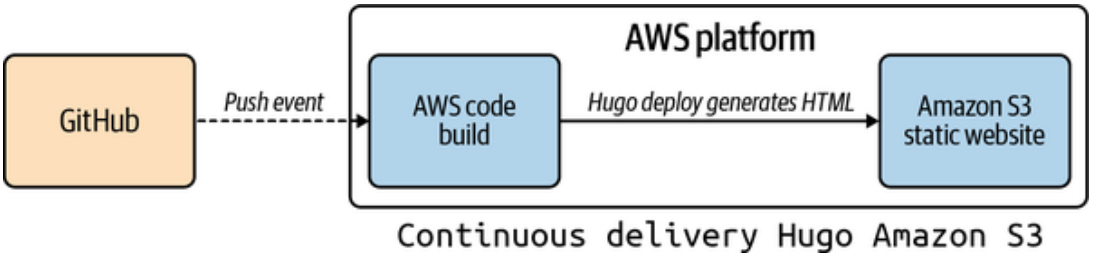

Figure 7-4. Hugo

The short version of how this project works is through the magic of the *buildspec.yml* file. Let's take a look at how this works in the following example. First, note that the `hugo` binary installs, then the `hugo` command runs to generate HTML files from the checked-out repo. Finally, the `aws`

command `aws s3 sync --delete public s3://dukefeb1` is the entire deployment process due to the power of S3 bucket hosting:

```
version: 0.1

environment_variables:
  plaintext:
    HUGO_VERSION: "0.79.1"

phases:
  install:
    commands:
      - cd /tmp
      - wget https://github.com/gohugoio/hugo/releases/download/v0.80.0/\
 hugo_extended_0.80.0_Linux-64bit.tar.gz
      - tar -xzf hugo_extended_0.80.0_Linux-64bit.tar.gz
      - mv hugo /usr/bin/hugo
      - cd
      - rm -rf /tmp/*
  build:
    commands:
      - rm -rf public
      - hugo
  post_build:
    commands:
      - aws s3 sync --delete public s3://dukefeb1
      - echo Build completed on `date`
```

Another way of describing a build system file is that it is a recipe. The information in the build configuration files is a "how-to" to perform the same actions in an AWS Cloud9 development environment.

As discussed in Chapter 2, AWS Cloud9 holds a special place in my heart because it solves a particular problem. Cloud-based development environments let you develop in the exact location where all of the action takes place. The example shows how powerful this concept is. Check out the code, test it in the cloud, and verify the same tool's deployment. In Figure 7-5 the AWS Cloud9 environment invokes a Python microservice.

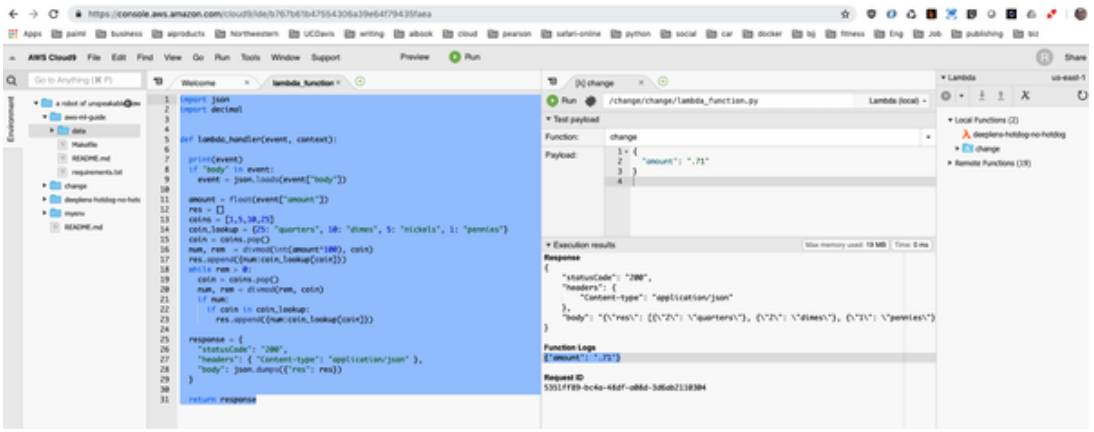

Figure 7-5. Cloud9

NOTE

You can watch a walkthrough of Hugo deployment on AWS on the O'Reilly platform, as well follow an additional, more detailed guide on the Pragmatic AI Labs website.

With the foundations of continuous delivery behind us, let's get into serverless on the AWS Platform.

Serverless Cookbook

Serverless is a crucial methodology for MLOps. In Chapter 2, I brought up the importance of Python functions. A Python function is a unit of work that can both take an input and optionally return an output. If a Python function is like a toaster, where you put in some bread, it heats the bread, and ejects toast, then serverless is the source of electricity.

A Python function needs to run somewhere, just like a toaster needs to plug into something to work. This concept is what serverless does; it enables code to run in the cloud. The most generic definition of serverless is code that runs without servers. The servers themselves are abstracted away to allow a developer to focus on writing functions. These functions do specific tasks, and these tasks could be chained together to build more complex systems, like servers that respond to events.

The function is the center of the universe with cloud computing. In practice, this means anything that is a function could map into a technology that solves a problem: containers, Kubernetes, GPUs, or AWS Lambda. As

you can see in Figure 7-6, there is a rich ecosystem of solutions in Python that map directly to a function.

The lowest level service that performs serverless on the AWS platform is AWS Lambda. Let's take a look at a few examples from this repo.

First, one of the simpler Lambda functions to write on AWS is a Marco Polo function. A Marco Polo function takes in an event with a name in it. So, for example, if the event name is "Marco," it returns "Polo." If the event name is something else, it returns "No!"

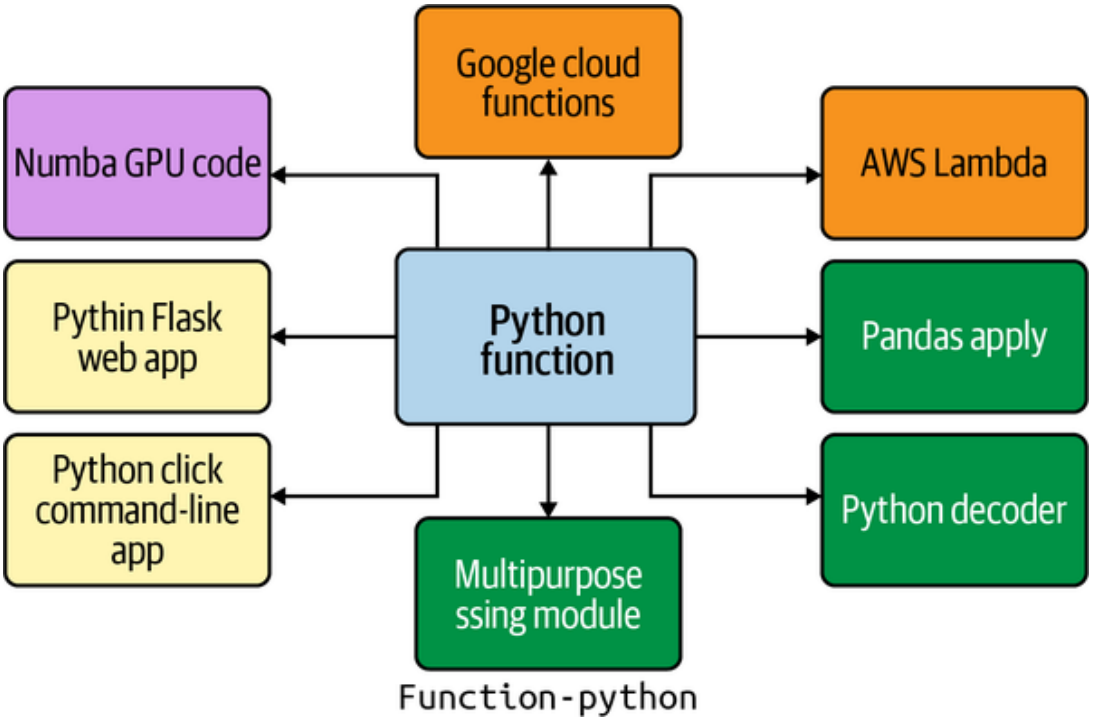

Figure 7-6. Python functions

NOTE

Growing up as a teenager in the 1980s and 1990s, Marco Polo was a typical game to play in the summer swimming pool. When I worked as a camp counselor at a pool near my house it was a favorite of the kids I supervised. The game works by everyone getting into a pool and one person closing their eyes and yelling "Marco." Next, the other players in the pool must say, "Polo." The person with their eyes closed uses sound to locate someone to tag. Once someone is tagged, then they are "It."

Here is the AWS Lambda Marco Polo code; note that an `event` passes into the `lambda_handler`:

```
def lambda_handler(event, context):
    print(f"This was the raw event: {event}")
    if event["name"] == "Marco":
        print(f"This event was 'Marco'")
        return "Polo"
    print(f"This event was not 'Marco'")
    return "No!"
```

With serverless cloud computing, think about a lightbulb in your garage. A lightbulb can be turned on many ways, such as manually via the light switch or automatically via the garage door open event. Likewise, an AWS Lambda responds to many signals as well.

Let's enumerate the ways both lightbulbs and lambdas can trigger:

- Lightbulbs
  - Manually flip on the switch
  - Via the garage door opener
  - Nightly security timer that turns on the light at midnight until 6 a.m.
- AWS Lambda
  - Manually invoke via console, AWS command line, or AWS Boto3 SDK
  - Respond to S3 events like uploading a file to a bucket
  - Timer invokes nightly to download data

What about a more complex example? With AWS Lambda, it is straight-forward to integrate an S3 Trigger with computer vision labeling on all new images dropped in a folder, with a trivial amount of code:

```
import boto3
from urllib.parse import unquote_plus

def label_function(bucket, name):
    """This takes an S3 bucket and a image name!"""
    print(f"This is the bucketname {bucket} !")
    print(f"This is the imagename {name} !")
    rekognition = boto3.client("rekognition")
    response = rekognition.detect_labels(
        Image={"S3Object": {"Bucket": bucket, "Name": name,}},
```

```
        )
        labels = response["Labels"]
        print(f"I found these labels {labels}")
        return labels


    def lambda_handler(event, context):
        """This is a computer vision lambda handler"""

        print(f"This is my S3 event {event}")
        for record in event['Records']:
            bucket = record['s3']['bucket']['name']
            print(f"This is my bucket {bucket}")
            key = unquote_plus(record['s3']['object']['key'])
            print(f"This is my key {key}")

        my_labels = label_function(bucket=bucket,
            name=key)
        return my_labels
```

Finally, you can chain multiple AWS Lambda functions together via AWS Step Functions:

```
{
  "Comment": "This is Marco Polo",
  "StartAt": "Marco",
  "States": {
    "Marco": {
      "Type": "Task",
      "Resource": "arn:aws:lambda:us-east-1:561744971673:function:marco20",
      "Next": "Polo"
    },
    "Polo": {
      "Type": "Task",
      "Resource": "arn:aws:lambda:us-east-1:561744971673:function:polo",
      "Next": "Finish"
    },
    "Finish": {
      "Type": "Pass",
      "Result": "Finished",
      "End": true
    }
```

```
  }
}
```

You can see this workflow in action in Figure 7-7.

Figure 7-7. Step Functions

Even more fun is the fact that you can call AWS Lambda functions via a CLI. Here is an example:

```
aws lambda invoke \
   --cli-binary-format raw-in-base64-out \
  --function-name marcopython \
  --payload '{"name": "Marco"}' \
  response.json
```

NOTE

It is important to always refer to the latest AWS documentation for the CLI as it is an actively moving target. As of the writing of this book, the current CLI version is V2, but you may need to adjust the command line examples as things change in the future. You can find the latest documentation at the AWS CLI Command Reference site.

The response of the payload is as follows:

```
{
    "StatusCode": 200,
    "ExecutedVersion": "$LATEST"
}
(.venv) [cloudshell-user@ip-10-1-14-160 ~]$ cat response.json
"Polo"(.venv) [cloudshell-user@ip-10-1-14-160 ~]$
```

NOTE

For a more advanced walkthrough of AWS Lambda using Cloud9 and the AWS SAM (Serverless Application Model), you can view a walkthrough of a small Wikipedia microservice on the Pragmatic AI Labs YouTube Channel or the O'Reilly Learning Platform.

AWS Lambda is perhaps the most valuable and flexible type of computing you can use to serve out predictions for machine learning pipelines or wrangle events in the service of an MLOps process. The reason for this is the speed of development and testing. Next, let's talk about a couple of CaaS, or container as a service, offerings.

AWS CaaS

Fargate is a container as a service offering from AWS that allows developers to focus on building containerized microservices. For example, in Figure 7-8, when this microservice works in the container, the entire runtime, including the packages necessary for deployment, will work in a new environment. The cloud platform handles the rest of the deployment.

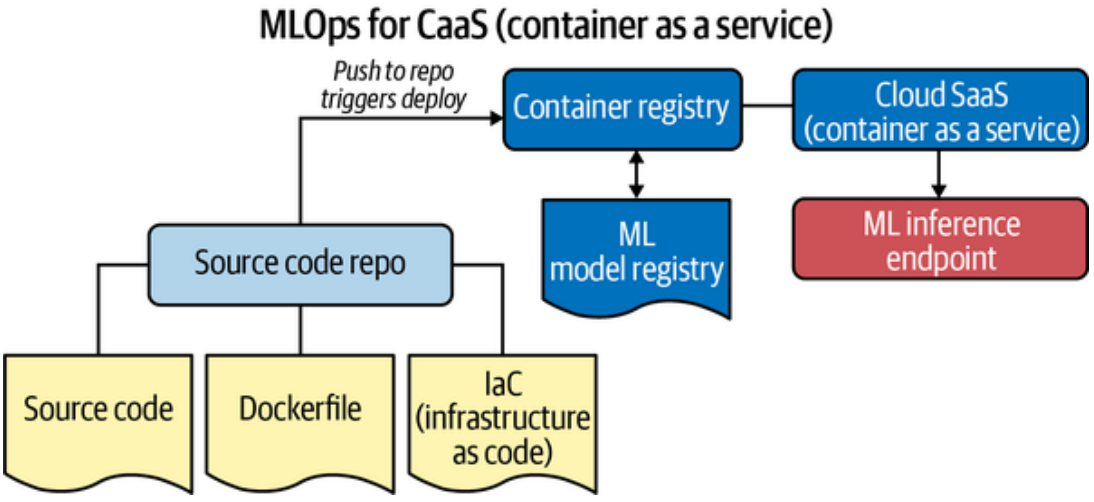

Figure 7-8. MLOps for CaaS

NOTE

Containers solve many problems that have plagued the software industry. So as a general rule, it is a good idea to use them for MLOps projects. Here is a partial list of the advantages of containers in projects:

- Allows the developer to mimic the production service locally on their desktop
- Allows easy software runtime distribution to customers through public container registries like Docker Hub, GitHub Container Registry, and Amazon Elastic Container Registry
- Allows for GitHub or a source code repo to be "source of truth" and contain all aspects of microservice: model, code, IaC, and runtime
- Allows for easy production deployment via CaaS services

Let's look at how you can build a microservice that returns the correct change using Flask. Figure 7-9 shows a development workflow on AWS Cloud9. Cloud9 is the development environment; a container gets built and pushed to ECR. Later that container runs in ECS.

Figure 7-9. ECS workflow

The following is the Python code for *app.py*:

```
from flask import Flask
from flask import jsonify
app = Flask(__name__)

def change(amount):
    # calculate the resultant change and store the result (res)
    res = []
    coins = [1,5,10,25] # value of pennies, nickels, dimes, quarters
    coin_lookup = {25: "quarters", 10: "dimes", 5: "nickels", 1: "pennies"}

    # divide the amount*100 (the amount in cents) by a coin value
    # record the number of coins that evenly divide and the remainder
    coin = coins.pop()
    num, rem  = divmod(int(amount*100), coin)
    # append the coin type and number of coins that had no remainder
    res.append({num:coin_lookup[coin]})

    # while there is still some remainder, continue adding coins to the result
```

```
        while rem > 0:
            coin = coins.pop()
            num, rem = divmod(rem, coin)
            if num:
                if coin in coin_lookup:
                    res.append({num:coin_lookup[coin]})
        return res

    @app.route('/')
    def hello():
        """Return a friendly HTTP greeting."""
        print("I am inside hello world")
        return 'Hello World! I can make change at route: /change'

    @app.route('/change/<dollar>/<cents>')
    def changeroute(dollar, cents):
        print(f"Make Change for {dollar}.{cents}")
        amount = f"{dollar}.{cents}"
        result = change(float(amount))
        return jsonify(result)


    if __name__ == '__main__':
        app.run(host='0.0.0.0', port=8080, debug=True)
```

Notice that Flask web microservice responds to change requests via web requests to the URL pattern `/change/<dollar>/<cents>` . You can view the complete source code for this Fargate example in the following GitHub repo. The steps are as follows:

1. Setup app: virtualenv + `make all`
2. Test app local: `python app.py`
3. Curl it to test: `curl localhost:8080/change/1/34`
4. Create ECR (Amazon Container Registry)

   In Figure 7-10, an ECR repository enables later Fargate deployment.

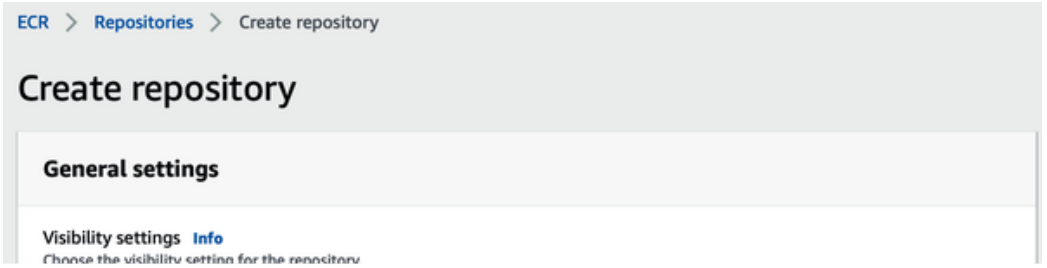

Figure 7-10. ECR

5. Build container
6. Push container
7. Run docker local: `docker run -p 8080:8080 changemachine`
8. Deploy to Fargate
9. Test the public service

NOTE

Any cloud service will have rapid, constant changes in functionality, so it is best to read the current documentation. The current Fargate documentation is a great place to read more about the latest ways to deploy to the service.

You can also, optionally, watch a complete walkthrough of a Fargate deployment on the O'Reilly platform.

Another option for CaaS is AWS App Runner, which further simplifies things. For example, you can deploy straight from source code or point to a container. In Figure 7-11, AWS App Runner creates a streamlined workflow that connects a source repo, deploy environment, and a resulting secure URL.

Figure 7-11. AWS App Runner

This repository can easily be converted to an AWS App Runner method in the AWS wizard with the following steps:

1. To build the project, use the command: `pip install -r requirements.txt`.
2. To run the project, use: `python app.py`.
3. Finally, configure the port to use: `8080`.

A key innovation, shown in Figure 7-12, is the ability to connect many different services on AWS, i.e., core infrastructure, like AWS CloudWatch, Load Balancers, Container Services, and API Gateways into one complete offering.

Figure 7-12. AWS App Runner service created

The final deployed service in Figure 7-13 shows a secure URL and the ability to invoke the endpoint and return correct change.

Figure 7-13. AWS App Runner deployed

Why is this helpful magic? In a nutshell, all of the logic, including perhaps a machine learning model, are in one repo. Thus, this is a compelling style for building MLOps-friendly products. In addition, one of the more complex aspects of delivering a machine learning application to production is microservice deployment. AWS App Runner makes most of the complexity disappear, capturing time for other parts of the MLOps problem. Next, let's discuss how AWS deals with computer vision.

Computer vision

I teach an applied computer vision course at Northwestern's graduate data science program. This class is delightful to teach because we do the following:

- Weekly video demos

- Focus on problem-solving versus coding or modeling
- Use high-level tools like AWS DeepLens, a deep learning–enabled video camera

In practice, this allows a rapid feedback loop that focuses on the problem versus the technology to solve the problem. One of the technologies in use is the AWS Deep Lens device as shown in Figure 7-14. The device is a complete computer vision hardware developer kit in that it contains a 1080p camera, operating system, and wireless capabilities. In particular, this solves the prototyping problem of computer vision.

Figure 7-14. DeepLens

Once AWS DeepLens starts capturing video, it splits the video into two streams. The Project Stream shown in Figure 7-15 adds real-time annotation and sends the packets to the MQ Telemetry Transport (MQTT) service, which is a publish-subscribe network protocol.

Figure 7-15. Detect

In Figure 7-16, the MQTT packets arrive in real time as the objects get detected in the stream.

The DeepLens is a “plug and play” technology since it tackles perhaps the most challenging part of the problem of building a real-time computer vision prototyping system—capturing the data, and sending it somewhere. The “fun” factor of this is how easy it is to go from zero to one building solutions with AWS DeepLens. Next, let’s get more specific and move beyond just building microservices and into building microservices that deploy machine learning code.

Figure 7-16. MQTT

MLOps on AWS

One way to get started with MLOps on AWS is to consider the following question. When given a machine learning problem with three constraints

—prediction accuracy, explainability, and operations, which two would you focus on to achieve success, and in what order?

Many academic-focused data scientists immediately jump to prediction accuracy. Building ever-better prediction models is a fun challenge, like playing Tetris. Also, the modeling is glamorous and a coveted aspect of the job. Data scientists like to show how accurate they can train an academic model using increasingly sophisticated techniques.

The entire Kaggle platform works on increasing prediction accuracy, and there are monetary rewards for the precise model. A different approach is to focus on operationalizing the model. This approach's advantage is that later, model accuracy can improve alongside improvements in the software system. Just as the Japanese automobile industry focused on Kaizen or continuous improvement, an ML system can focus on reasonable initial prediction accuracy and improve quickly.

The culture of AWS supports this concept in the leadership principles of "Bias for Action" and "Deliver Results." "Bias for Action" refers to defaulting to speed and delivery results, focusing on the critical inputs to a business, and delivering results quickly. As a result, the AWS products around machine learning, like AWS SageMaker, show the spirit of this culture of action and results.

Continuous delivery (CD) is a core component in MLOps. Before you can automate delivery for machine learning, the Microservice itself needs automation. The specifics change depending on the type of AWS service involved. Let's start with an end-to-end example.

In the following example, an Elastic Beanstalk Flask app continuously deploys using all AWS technology from AWS Code Build to AWS Elastic Beanstalk. This "stack" is also ideal for deploying ML models. Elastic Beanstalk is a platform as a service technology offered by AWS that streamlines much of the work of deploying an application.

In Figure 7-17, notice that AWS Cloud9 is a recommended starting point for development. Next, a GitHub repository holds the source code for the project, and as change events occur, it triggers the cloud native build server, AWS CodeBuild. Finally, the AWS Code Build process runs continu-

ous integration, tests for the code, and provides continuous delivery to AWS Elastic Beanstalk.

Figure 7-17. Elastic Beanstalk

NOTE

The source code and a walkthrough of this example are at the following links:

- Source Code
- O'Reilly Platform Video Walkthrough

To replicate this exact project, you can do the following steps:

1. Check out the repository in AWS Cloud9 or AWS CloudShell if you have strong command line skills.
2. Create a Python virtualenv and source it and run `make all`:

   ```
   python3 -m venv ~/.eb
   source ~/.eb/bin/activate
   make all
   ```

   Note that `awsebcli` installs via requirements, and this tool controls Elastic Beanstalk from the CLI.
3. Initialize new `eb` app:

   ```
   eb init -p python-3.7 flask-continuous-delivery --region us-east-1
   ```

   Optionally, you use `eb init` to create SSH keys to shell into the running instances.
4. Create remote eb instance:

   ```
   eb create flask-continuous-delivery-env
   ```

5. Setup AWS Code Build Project. Note your Makefile needs to reflect your project names:

```
version: 0.2

phases:
  install:
    runtime-versions:
      python: 3.7
  pre_build:
    commands:
      - python3.7 -m venv ~/.venv
      - source ~/.venv/bin/activate
      - make install
      - make lint

  build:
    commands:
      - make deploy
```

After you get the project working with continuous deployment, you are ready to move to the next step, deploying an ML model. I highly recommend getting a "hello world" type project working with continuous deployment like this one before you proceed directly into a complex ML project when you are learning new technology. Next, let's look at an intentionally simple MLOps Cookbook that is the foundation for many new AWS Services deployments.

MLOps Cookbook on AWS

With the foundational components out of the way, let's look at a basic machine learning recipe and apply it to several scenarios. Notice that this core recipe is deployable to many services on AWS and many other cloud environments. This following MLOps Cookbook project is intentionally spartan, so the focus is on deploying machine learning. For example, this project predicts height from a weight input for Major League Baseball players.

In Figure 7-18, GitHub is the source of truth and contains the project scaffolding. Next, the build service is GitHub Actions, and the container service is GitHub Container Registry. Both of these services can easily replace any similar offering in the cloud. In particular, on the AWS Cloud,

you can use AWS CodeBuild for CI/CD and AWS ECR (Elastic Container Registry). Finally, once a project has been "containerized," it opens up the project to many deployment targets. On AWS, these include AWS Lambda, AWS Elastic Beanstalk, and AWS App Runner.

Figure 7-18. MLOps Cookbook

The following files are all useful to build solutions in many different recipes:

Makefile

The Makefile is both a list of recipes and a way to invoke those recipes. View the Makefile in the example GitHub project.

requirements.txt

The requirements file contains the list of Python packages for the project. Typically these packages are "pinned" to a version number, which limits unexpected package dependencies. View the requirements.txt in the example GitHub project.

cli.py

This command line shows how an ML library can also be invoked from the CLI, not just via a web application. View the cli.py in the example GitHub project.

utilscli.py

The utilscli.py is a utility that allows the user to invoke different endpoints, i.e., AWS, GCP, Azure, or any production environment. Most machine learning algorithms require data to be scaled. This tool simplifies scaling the input and scaling back out the output.View the utilscli.py in the example GitHub project.

app.py

The application file is the Flask web microservice that accepts and returns a JSON prediction result via the `/predict` URL endpoint. View the app.py in the example GitHub project.

mlib.py

The model handling library does much of the heavy lifting in a centralized location. This library is intentionally very basic and doesn't solve more complicated issues like caching loading of the model or other production issues unique to production deployment. View the mlib.py in the example GitHub project.

htwtmlb.csv

A CSV file is helpful for input scaling. View the htwtmlb.csv in the example GitHub project.

model.joblib

This model is exported from sklearn but could easily be in another format such as ONNX or TensorFlow. Other real-world production considerations could be keeping this model in a different location like Amazon S3, in a container, or even hosted by AWS SageMaker. View the model.joblib in the example GitHub project.

Dockerfile

This file enables project containerization and, as a result, opens up many new deployment options, both on the AWS platform as well as other clouds. View the Dockerfile in the example GitHub project.

Baseball_Predictions_Export_Model.ipynb

The Jupyter notebook is a crucial artifact to include in a machine learning project. It shows another developer the thinking behind creating the model and provides valuable context for maintaining the project in production. View the Baseball_Predictions_Export_Model.ipynb in the example GitHub project.

These project artifacts are helpful as an educational tool in explaining MLOps but may be different or more complex in a unique production scenario. Next, let's discuss how CLI (command line interface) tools help operationalize a machine learning project.

CLI Tools

In this project, there are two CLI tools. First, the main *cli.py* is the endpoint that serves out predictions. For example, to predict the height of an MLB player, you use the following command to create a forecast: `./cli.py --weight 180`. Notice in Figure 7-19 that the command line option of `--weight` allows the user to test out many new prediction inputs quickly.

Figure 7-19. CLI predict

So how does this work? Most of the "magic" is via a library that does the heavy lifting of scaling the data, making the prediction, then doing an inverse transform back:

```
"""MLOps Library"""

import numpy as np
import pandas as pd
from sklearn.linear_model import Ridge
import joblib
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
import logging

logging.basicConfig(level=logging.INFO)

import warnings

warnings.filterwarnings("ignore", category=UserWarning)


def load_model(model="model.joblib"):
    """Grabs model from disk"""

    clf = joblib.load(model)
    return clf


def data():
    df = pd.read_csv("htwtmlb.csv")
    return df
```

```
def retrain(tsize=0.1, model_name="model.joblib"):
    """Retrains the model

    See this notebook: Baseball_Predictions_Export_Model.ipynb
    """
    df = data()
    y = df["Height"].values  # Target
    y = y.reshape(-1, 1)
    X = df["Weight"].values  # Feature(s)
    X = X.reshape(-1, 1)
    scaler = StandardScaler()
    X_scaler = scaler.fit(X)
    X = X_scaler.transform(X)
    y_scaler = scaler.fit(y)
    y = y_scaler.transform(y)
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=tsize, random_state=3
    )
    clf = Ridge()
    model = clf.fit(X_train, y_train)
    accuracy = model.score(X_test, y_test)
    logging.debug(f"Model Accuracy: {accuracy}")
    joblib.dump(model, model_name)
    return accuracy, model_name


def format_input(x):
    """Takes int and converts to numpy array"""

    val = np.array(x)
    feature = val.reshape(-1, 1)
    return feature


def scale_input(val):
    """Scales input to training feature values"""

    df = data()
    features = df["Weight"].values
    features = features.reshape(-1, 1)
    input_scaler = StandardScaler().fit(features)
    scaled_input = input_scaler.transform(val)
    return scaled_input
```

```
def scale_target(target):
    """Scales Target 'y' Value"""

    df = data()
    y = df["Height"].values  # Target
    y = y.reshape(-1, 1)  # Reshape
    scaler = StandardScaler()
    y_scaler = scaler.fit(y)
    scaled_target = y_scaler.inverse_transform(target)
    return scaled_target


def height_human(float_inches):
    """Takes float inches and converts to human height in ft/inches"""

    feet = int(round(float_inches / 12, 2))  # round down
    inches_left = round(float_inches - feet * 12)
    result = f"{feet} foot, {inches_left} inches"
    return result


def human_readable_payload(predict_value):
    """Takes numpy array and returns back human readable dictionary"""

    height_inches = float(np.round(predict_value, 2))
    result = {
        "height_inches": height_inches,
        "height_human_readable": height_human(height_inches),
    }
    return result


def predict(weight):
    """Takes weight and predicts height"""

    clf = load_model()  # loadmodel
    np_array_weight = format_input(weight)
    scaled_input_result = scale_input(np_array_weight)
    scaled_height_prediction = clf.predict(scaled_input_result)
    height_predict = scale_target(scaled_height_prediction)
    payload = human_readable_payload(height_predict)
    predict_log_data = {
        "weight": weight,
        "scaled_input_result": scaled_input_result,
        "scaled_height_prediction": scaled_height_prediction,
```

```
            "height_predict": height_predict,
            "human_readable_payload": payload,
        }
        logging.debug(f"Prediction: {predict_log_data}")
        return payload
```

Next, the Click framework wraps the library calls to *mlib.py* and makes a clean interface to serve out predictions. There are many advantages to using command line tools as the primary interface for interacting with machine learning models. The speed to develop and deploy a command line machine learning tool is perhaps the most important:

```
#!/usr/bin/env python
import click
from mlib import predict


@click.command()
@click.option(
    "--weight",
    prompt="MLB Player Weight",
    help="Pass in the weight of a MLB player to predict the height",
)
def predictcli(weight):
    """Predicts Height of an MLB player based on weight"""

    result = predict(weight)
    inches = result["height_inches"]
    human_readable = result["height_human_readable"]
    if int(inches) > 72:
        click.echo(click.style(human_readable, bg="green", fg="white"))
    else:
        click.echo(click.style(human_readable, bg="red", fg="white"))


if __name__ == "__main__":
    # pylint: disable=no-value-for-parameter
    predictcli()
```

The second CLI tool is *utilscli.py*, which performs model retraining and could serve as the entry point to do more tasks. For example, this version

doesn't change the default `model_name` , but you could add that as an option by forking this repo:

```
./utilscli.py retrain --tsize 0.4
```

Notice that the *mlib.py* again does the heavy lifting, but the CLI provides a convenient way to do rapid prototyping of an ML model:

```
#!/usr/bin/env python
import click
import mlib
import requests

@click.group()
@click.version_option("1.0")
def cli():
    """Machine Learning Utility Belt"""


@cli.command("retrain")
@click.option("--tsize", default=0.1, help="Test Size")
def retrain(tsize):
    """Retrain Model
    You may want to extend this with more options, such as setting model_name
    """

    click.echo(click.style("Retraining Model", bg="green", fg="white"))
    accuracy, model_name = mlib.retrain(tsize=tsize)
    click.echo(
        click.style(f"Retrained Model Accuracy: {accuracy}", bg="blue",
                    fg="white")
    )
    click.echo(click.style(f"Retrained Model Name: {model_name}", bg="red",
                           fg="white"))


@cli.command("predict")
@click.option("--weight", default=225, help="Weight to Pass In")
@click.option("--host", default="http://localhost:8080/predict",
              help="Host to query")
def mkrequest(weight, host):
    """Sends prediction to ML Endpoint"""
```

```
        click.echo(click.style(f"Querying host {host} with weight: {weight}",
            bg="green", fg="white"))
        payload = {"Weight":weight}
        result = requests.post(url=host, json=payload)
        click.echo(click.style(f"result: {result.text}", bg="red", fg="white"))


if __name__ == "__main__":
    cli()
```

Figure 7-20 is an example of retraining the model.

Figure 7-20. Model retrain

You can also query a deployed API, which will soon tackle the CLI, allowing you to change both the host and the value passed into the API. This step uses the `requests` library. It can help build a "pure" Python example of a prediction tool versus only predicting with a `curl` command. You can see an example of the output in Figure 7-21:

```
./utilscli.py predict --weight 400
```

Figure 7-21. Predict requests

Perhaps you are sold on CLI tools as the ideal way to rapidly deploy ML models, thus being a true MLOps-oriented organization. What else could you do? Here are two more ideas.

First, you could build a more sophisticated client that makes async HTTP requests a deployed web service. This functionality is one of the advantages of building utility tools in pure Python. One library to consider for async HTTPS is Fast API.

Second, you can continuously deploy the CLI itself. For many SaaS companies, university labs, and many more scenarios, this could be the ideal workflow to adapt speed and agility as a primary goal. In this example GitHub project there is a simple example of how to containerize a com-

mand-line tool. The three critical files are the Makefile, the *cli.py*, and the Dockerfile.

Notice that the Makefile makes it easy to "lint" the syntax of the Dockerfile using `hadolint`:

```
install:
  pip install --upgrade pip &&\
    pip install -r requirements.txt

lint:
  docker run --rm -i hadolint/hadolint < Dockerfile
```

The CLI itself is pretty small, one of the valuable aspects of the Click framework:

```
#!/usr/bin/env python
import click

@click.command()
@click.option("--name")
def hello(name):
    click.echo(f'Hello {name}!')

if __name__ == '__main__':
    #pylint: disable=no-value-for-parameter
    hello()
```

Finally, the Dockerfile builds the container:

```
FROM python:3.7.3-stretch

# Working Directory
WORKDIR /app

# Copy source code to working directory
COPY . app.py /app/

# Install packages from requirements.txt
# hadolint ignore=DL3013
```

```
RUN pip install --no-cache-dir --upgrade pip &&\
    pip install --no-cache-dir --trusted-host pypi.python.org -r requirements.t
```

To run this exact container, you can do the following:

```
docker run -it noahgift/cloudapp python app.py --name "Big John"
```

The output is the following:

```
Hello Big John!
```

This workflow is ideal for ML-based CLI tools! For example, to build this container yourself and push it, you could do the following workflow:

```
docker build --tag=<tagname> .
docker push <repo>/<name>:<tagname>
```

This section covered ideas on how to scaffold out Machine Learning projects. In particular, three main ideas are worth considering: using containers, building a web microservice, and using command line tools. Next, let's cover Flask microservices in more detail.

Flask Microservice

When dealing with MLOps workflows, it is essential to note that a Flask ML microservice can operate in many ways. This section covers many of these examples.

Let's first take a look at the core of the Flask machine learning microservice application in the following example. Note, again, that most of the heavy lifting is via the *mlib.py* library. The only "real" code is the Flask route that does the following `@app.route("/predict", methods=['POST'])` post request. It accepts a JSON payload looking similar to `{"Weight": 200}`, then returns a JSON result:

```
from flask import Flask, request, jsonify
from flask.logging import create_logger
```

```
import logging

from flask import Flask, request, jsonify
from flask.logging import create_logger
import logging

import mlib

app = Flask(__name__)
LOG = create_logger(app)
LOG.setLevel(logging.INFO)

@app.route("/")
def home():
    html = f"<h3>Predict the Height From Weight of MLB Players</h3>"
    return html.format(format)

@app.route("/predict", methods=['POST'])
def predict():
    """Predicts the Height of MLB Players"""

    json_payload = request.json
    LOG.info(f"JSON payload: {json_payload}")
    prediction = mlib.predict(json_payload['Weight'])
    return jsonify({'prediction': prediction})

if __name__ == "__main__":
    app.run(host='0.0.0.0', port=8080, debug=True)
```

This Flask web service runs in a straightforward manner using `python app.py`. For example, you run the Flask microservice as follows with the command `python app.py`:

```
(.venv) ec2-user:~/environment/Python-MLOps-Cookbook (main) $ python app.py
* Serving Flask app "app" (lazy loading)
* Environment: production
WARNING: This is a development server. Do not use it in a production...
Use a production WSGI server instead.
* Debug mode: on
INFO:werkzeug: * Running on http://127.0.0.1:8080/ (Press CTRL+C to quit)
INFO:werkzeug: * Restarting with stat
WARNING:werkzeug: * Debugger is active!
INFO:werkzeug: * Debugger PIN: 251-481-511
```

To serve a prediction against the application, run the `predict.sh`. Notice, a small `bash` script can help debug your application without having to type out all of the jargon of `curl`, which can cause you to make syntax mistakes:

```
#!/usr/bin/env bash

PORT=8080
echo "Port: $PORT"

# POST method predict
curl -d '{
   "Weight":200
}'\
     -H "Content-Type: application/json" \
     -X POST http://localhost:$PORT/predict
```

The results of the prediction show that the Flask endpoint returns a JSON payload:

```
(.venv) ec2-user:~/environment/Python-MLOps-Cookbook (main) $ ./predict.sh
Port: 8080
{
  "prediction": {
    "height_human_readable": "6 foot, 2 inches",
    "height_inches": 73.61
  }
}
```

Note that the earlier *utilscli.py* tool could also make web requests to this endpoint. You could also use httpie or the postman tool. Next, let's discuss how a containerization strategy works for this microservice.

Containerized Flask microservice

The following is an example of how to build the container and run it locally. (You can find the contents of *predict.sh* on GitHub.)

```
#!/usr/bin/env bash
```

```
# Build image
#change tag for new container registery, gcr.io/bob
docker build --tag=noahgift/mlops-cookbook .

# List docker images
docker image ls

# Run flask app
docker run -p 127.0.0.1:8080:8080 noahgift/mlops-cookbook
```

Adding a container workflow is straightforward, and it enables an easier development method since you can share a container with another person on your team. It also opens up the option to deploy your machine learning application to many more platforms. Next, let's talk about how to build and deploy containers automatically.

Automatically build container via GitHub Actions and push to GitHub Container Registry

As covered earlier in the book, the container workflow of GitHub Actions is a valuable ingredient for many recipes. It may make sense to build a container programmatically with GitHub Actions and push it to the GitHub Container Registry. This step could serve both as a test of the container build process and deploy target—say you are deploying the CLI tool discussed earlier. This example is what that code looks like in practice. Note you would need to change the `tags` to your Container Registry (shown in Figure 7-22):

```
build-container:
  runs-on: ubuntu-latest
  steps:
  - uses: actions/checkout@v2
  - name: Loging to GitHub registry
    uses: docker/login-action@v1
    with:
      registry: ghcr.io
      username: ${{ github.repository_owner }}
      password: ${{ secrets.BUILDCONTAINERS }}
  - name: build flask app
    uses: docker/build-push-action@v2
    with:
```

```
context: ./
#tags: alfredodeza/flask-roberta:latest
tags: ghcr.io/noahgift/python-mlops-cookbook:latest
push: true
```

Figure 7-22. GitHub Container Registry

SaaS (software as a service) build systems and SaaS container registries are helpful outside of just the core cloud environment. They verify to the developers both inside and outside your company that the container workflow is valid. Next, let's tie together many of the concepts in this chapter and use a high-level PaaS offering.

AWS App Runner Flask microservice

AWS App Runner is a high-level service that dramatically simplifies MLOps. For example, the previous Python MLOps cookbook recipes are straightforward to integrate with just a few AWS App Runner service clicks. In addition, as Figure 7-23 shows, AWS App Runner points to a source code repo, and it will auto-deploy on each change to GitHub.

Once it's deployed, you can open up either AWS Cloud9 or AWS CloudShell, clone the Python MLOps Cookbook repo, and then use *utilscli.py* to query the endpoint given to you from the App Runner service. The successful query in AWS CloudShell is displayed in Figure 7-24.

Figure 7-23. AWS App Runner

Figure 7-24. AWS App Runner prediction result

In a nutshell, high-level AWS services allow you to do MLOps more efficiently since less effort goes toward DevOps build processes. So next, let's move onto another computer option for AWS, AWS Lambda.

AWS Lambda Recipes

Install SAM (AWS Serverless Application Model) as AWS documentation instructs. AWS Cloud9 has it installed already. You can find the recipes on GitHub. AWS Lambda is essential because of how deeply integral it is to AWS. Let's explore how to use modern best practices to deploy a serverless ML model next.

An efficient and recommended method to deploy software to production is through SAM. This approach's innovation combines Lambda functions, event sources, and other resources as a deployment process and development toolkit.

In particular, the key benefits of SAM, according to AWS, include single-deployment configuration, an extension of AWS CloudFormation, built-in best practices, local debugging and testing, and deep integration with development tools, including my favorite, Cloud9.

To get started, first, you should install the AWS SAM CLI. After that, you can refer to the official guide for the best results.

AWS Lambda-SAM Local

To get started with SAM Local, you can try the following workflow for a new project:

- Install SAM (as shown previously)
- `sam init`
- `sam local invoke`

NOTE

If building on Cloud9, it's probably a good idea to resize using utils/resize.sh:

```
utils/resize.sh 30
```

This trick gives you more disk size to build multiple containers with SAM local or any other AWS Container workflow.

Here is a typical SAM init layout, which is only slightly different for an ML project:

```
├── sam-app/
│   ├── README.md
│   ├── app.py
│   ├── requirements.txt
│   ├── template.yaml
│   └── tests
│       └── unit
│           ├── __init__.py
│           └── test_handler.py
```

With this foundational knowledge in place, let's move on to doing more with AWS Lambda and SAM.

AWS Lambda-SAM Containerized Deploy

Now let's dive into a containerized workflow for SAM since it supports registering a container that AWS Lambda uses. You can see the containerized SAM-Lambda deploy the project in the following repo. First, let's cover the key components. The critical steps to deploy to SAM include the following files:

- *App.py* (the AWS Lambda entry point)
- The Dockerfile (what gets built and sent to Amazon ECR)
- *Template.yaml* (used by SAM to deploy the app)

The lambda handler does very little because the hard work is still happening in the *mlib.py* library. One "gotcha" to be aware of with AWS Lambda is that you need to use different logic handling in lambda functions depending on how they are invoked. For example, if the Lambda invokes via the console or Python, then there is no web request body, but in the case of integration with API Gateway, the payload needs to be extracted from the `body` of the event:

```
import json
import mlib
```

```
def lambda_handler(event, context):
    """Sample pure Lambda function"""

    #Toggle Between Lambda function calls and API Gateway Requests
    print(f"RAW LAMBDA EVENT BODY: {event}")
    if 'body' in event:
        event = json.loads(event["body"])
        print("API Gateway Request event")
    else:
        print("Function Request")

    #If the payload is correct predict it
    if event and "Weight" in event:
        weight = event["Weight"]
        prediction = mlib.predict(weight)
        print(f"Prediction: {prediction}")
        return {
            "statusCode": 200,
            "body": json.dumps(prediction),
        }
    else:
        payload = {"Message": "Incorrect or Empty Payload"}
        return {
            "statusCode": 200,
            "body": json.dumps(payload),
        }
```

The project contains a Dockerfile that builds the lambda for the ECR location. Notice that it can pack into the Docker container in the case of a smaller ML model and even be programmatically retrained and packed via AWS CodeBuild:

```
FROM public.ecr.aws/lambda/python:3.8

COPY model.joblib mlib.py htwtmlb.csv app.py requirements.txt ./

RUN python3.8 -m pip install -r requirements.txt -t .

# Command can be overwritten by providing a different command
CMD ["app.lambda_handler"]
```

The SAM template controls the IaC (Infrastructure as Code) layer, allowing for an easy deployment process:

```
AWSTemplateFormatVersion: '2010-09-09'
Transform: AWS::Serverless-2016-10-31
Description: >
  python3.8

  Sample SAM Template for sam-ml-predict

Globals:
  Function:
    Timeout: 3

Resources:
  HelloWorldMLFunction:
    Type: AWS::Serverless::Function
    Properties:
      PackageType: Image
      Events:
        HelloWorld:
          Type: Api
          Properties:
            Path: /predict
            Method: post
    Metadata:
      Dockerfile: Dockerfile
      DockerContext: ./ml_hello_world
      DockerTag: python3.8-v1

Outputs:
  HelloWorldMLApi:
    Description: "API Gateway endpoint URL for Prod stage for Hello World ML
    function"
    Value: !Sub "https://${ServerlessRestApi}.execute-api.${AWS::Region}.\
      amazonaws.com/Prod/predict/"
  HelloWorldMLFunction:
    Description: "Hello World ML Predict Lambda Function ARN"
    Value: !GetAtt HelloWorldMLFunction.Arn
  HelloWorldMLFunctionIamRole:
    Description: "Implicit IAM Role created for Hello World ML function"
    Value: !GetAtt HelloWorldMLFunctionRole.Arn
```

The only steps left to do are to run two commands: `sam build` and `sam deploy --guided`, which allows you to walk through the deployment process. For example, in Figure 7-25, the `sam build` notice builds the container and then prompts you to either test locally via `sam local invoke` or do the guided deploy.

Figure 7-25. SAM build

You can invoke a `sam local invoke -e payload.json` with the following body to test locally:

```
{
    "Weight": 200
}
```

Notice that the container spins up, and a payload is sent and returned. This testing process is invaluable before your deploy the application to ECR:

```
(.venv) ec2-user:~/environment/Python-MLOps-Cookbook/recipes/aws-lambda-sam/
sam-ml-predict
Building image.................
Skip pulling image and use local one: helloworldmlfunction:rapid-1.20.0.
START RequestId: d104cf8a-ce6b-4f50-9f2b-c82e99b8016f Version: $LATEST
RAW LAMBDA EVENT BODY: {'Weight': 200}
Function Request
Prediction: {'height_inches': 73.61, 'height_human_readable': '6 foot, 2 inches
END RequestId: d104cf8a-ce6b-4f50-9f2b-c82e99b8016f
REPORT RequestId: d104cf8a-ce6b-4f50-9f2b-c82e99b8016f Init
Duration: 0.08 ms Duration: 2187.82
{"statusCode": 200, "body": "{\"height_inches\": 73.61,\
\"height_human_readable\": \"6 foot, 2 inches\"}"}
```

You can do a guided deployment with testing out of the way, as shown in Figure 7-26. Take special note that the prompts guide each step of the deployment process if you select this option.

Figure 7-26. SAM guided deploy

Once you have deployed your lambda, there are multiple ways to both use it and test it. Perhaps one of the easiest is to verify the image is via the AWS Lambda Console (see Figure 7-27).

Figure 7-27. Test AWS Lambda in Console

Two ways to test the actual API itself include the AWS Cloud9 Console and also the Postman tool. Figures 7-29 and 7-30 show examples of both.

Figure 7-28. Invoke Lambda Cloud9

Figure 7-29. Test AWS Lambda with Postman

Other deployment targets to consider deploying this base machine learning recipe include Flask Elastic Beanstalk and AWS Fargate. Different variations on the recipe include other HTTP services, such as fastapi, or the use of a pretrained model available via the AWS Boto3 API versus training your model.

AWS Lambda is one of the most exciting technologies to build distributed systems that incorporate data engineering and machine learning engineering. Let's talk about a real-world case study next.

Applying AWS Machine Learning to the Real World

Let's dive into examples of how to use AWS machine learning resources in the real world. In this section, several MLOps take on what is involved in actual companies doing machine learning.

NOTE

There are many ways to use AWS for MLOps and an almost infinite number of real-world combinations of services. Here is a partial list of recommended machine learning engineering patterns on AWS:

Container as a service (CaaS)

For organizations struggling to get something going, CaaS is a great place to start the MLOps journey. A recommended service is AWS App Runner.

SageMaker

For larger organizations with different teams and big data, SageMaker is an excellent platform to focus on because it allows for fine-grained security and enterprise-level deployment and training.

Serverless with AI APIs

For small startups that need to move very quickly, an excellent initial approach to doing MLOps is to use pretrained models via an API along with a serverless technology like AWS Lambda.

CASE STUDY: SPORTS SOCIAL NETWORK

This section is a case study on a Sports Social Network built on top of AWS machine learning, with the former Director of Product Management, Rob Loranger. Rob's entry into the world of data and its importance for decision-making began in 2005 when he became a sales engineer specializing in software tools used to design, develop, and manage relational databases, and eventually become a product manager for an enterprise data architecture platform. Since then, Rob has remained a product manager, gaining experience across machine learning–driven software platforms ranging from social networking to data center and telecom network analysis and control.

Rob earned an MBA from the University of California, Davis, where he studied data analytics and machine learning. He gained experience in all ML pipeline activities, from formulating the business problem to choosing, evaluating, and deploying a machine learning model. The tools he has used along the way include Amazon SageMaker, Amazon Forecast, Jupyter Notebooks, Minitab, Stata, Weka, Python, and R.

Q: What are 3–5 of the most important things to be aware of deploying and maintaining machine learning systems at scale?

A: Machine learning is not a one-and-done process. You must be committed to continually evaluating the data and the model's accuracy as these change over time. As with all impactful projects, involve the SME (subject matter expert) early and often so that you can build the best understanding of the business problem and the data to be used for testing and validating your machine learning model.

Be sure to make the business problem that you are attempting to solve with machine learning and the impact of solving that problem as straightforward as possible to garner maximum support from critical stakeholders over the project's lifetime.

Q: Explain how a product manager for AI/ML thinks about building ML products.

A: As a product manager for the Sqor Sports social platform, the KPI that mattered most to me, as it was a strong indicator of the platform's stickiness, was monthly active users (MAU). Moreover, to determine that users truly valued the platform, it was essential to see this metric grow significantly over time. Even if the metric maintained a seemingly good steady-state, the platform could still be failing. The steady-state could attribute to situations such as an influx of new users month-over-month that were active only for their first month before leaving or a solid core of niche users who found the platform valuable and continued to use it over time. Either of these would fall well short of the mission of Sqor Sports to be the number one social platform for athletes, teams, and their fans to meaningfully engage with one another.

It was essential to build a rich platform that attracted new users and provide an exciting and unique experience not found anywhere else. Therefore, our strategy improved fan acquisition, fan retention, and viral growth through fans (i.e., fans helping to acquire new fans). Improvements to the platform in these areas would reflect on our core KPI, and MAU would not grow satisfactorily until we got all of these areas right.

By leveraging existing relationships with athletes and teams held by our executive team, board, and business team, we quickly grew the number of athletes and groups on the platform. In turn, by leveraging the athlete and team social content created on Sqor and syndicated across other social networks, we rapidly acquired fans of these athletes and teams. Moreover, viral growth is a large topic that we will not fully explore here. Still, it was also a core part of our strategy reflected in the product roadmap as features that provided fans with unique content and experiences, and improved fans' ability to create and foster relationships and share content on- and off-platform.

However, early on, fan retention proved challenging as, even though we had a lot of content created by our athletes and teams, fans found it hard to discover and follow new athletes/teams. Therefore, fans typically had a small network of athletes and teams and did not receive enough content from the platform. As a result, the majority of fans became inactive after their first month.

One key aspect of solving this retention problem was rooted in our machine learning–based recommendation engine. Unfortunately, the engine's first several iterations suffered from low data entry leading to poor fan feeds recommendations. However, as we improved the recommendation engine data and algorithm, we began to see the improvements to fan retention we were after, as seen through cohort analysis.

Finally, only with a sizable and growing MAU could we ultimately support our revenue strategy that, early on, consisted solely of a revenue share with athlete merchandise sold on the platform, but with the ultimate goal of providing an enterprise platform to manage social presence (e.g., social media activity, top fans, merchandise sales, etc.).

Q: How can people get ahold of you, and what would you like to share that you are working on with the readers?

A: You can get ahold of me via email (*rtloranger@gmail.com*) or Linkedin.

Julien Simon is a prolific content creator and evangelist for the AWS machine learning platform. I was lucky to grab him for some insight into machine learning on the AWS platform.

Q: What is your background, and how did you get involved with operationalizing machine learning?

A: I'm a software engineer who ended up spending 10 years leading large software and infrastructure teams in several startups. Collecting, storing, and processing data was always a central activity in my teams. Over time, it evolved from relational databases to business intelligence to real-time analytics to machine learning. The latter was actually at the core of one of the companies I worked for. Given our scale at the time (about 500K hits per second), training and deploying models created all kinds of interesting challenges, from storing data to deploying (and sometimes rolling back) models on hundreds of servers.

Q: What are 3–5 of the most important things to be aware of deploying and maintaining machine learning systems at scale?

A: First, can we please agree on the fact that machine learning is software engineering? If we do, then let's also agree that best practices like traceability, versioning, testing, automation, and documentation are not optional. I'm still surprised how deviant machine learning workflows can be, because "things are different." I don't think they are at all. Of course, data science is a very specific discipline, but at the end of the day, it produces code and data artifacts, so let's not reinvent the wheel. A lot of things that have worked well for decades still work with machine learning.

Code and datasets should be versioned and traced. Code and models should be tested. Models should be safely deployed in production, with quality gates and well-understood strategies (canary, blue-green, etc.). Monitoring is also critical, both from an infrastructure perspective (throughput, latency) and from a machine learning perspective (prediction quality, data drift). Finally, your models should also scale up and down according to incoming traffic. And yes, all these processes should be as automated as possible.

Again, nothing really new, but let's start doing it for machine learning too!

Q: What are you most excited about right now with machine learning and why?

A: The lazy guy in me loves how we can download and deploy models with just a couple of lines of code. Initiatives like Hugging Face make it extremely easy to add state-of-the-art models to your application, even if you're not a machine learning expert. In general, anything that makes machine learning more accessible is a great step forward: transfer learning, AutoML, and high-level AI services that completely hide all complexity. I often say that we should make machine learning as simple as Amazon S3, so any step in that direction is welcome.

From a technology perspective, I keep a close eye on new hardware that could massively accelerate training and inference. GPUs are nice, yet they feel like a brute force option most of the time. I'm definitely looking forward to more elegant, more cost-effective, and less power-hungry alternatives!

Q: What are the top 3–5 things a person reading this book can do to succeed in their career doing MLOps?

A: My number one advice is to learn as much as you can about the business problem that you're trying to solve. Unless you have a deep understanding of it, you won't be able to call the right shots. What do users really care about? What are the key metrics you should be watching? What's the best cost versus performance versus time to market trade-off? All these are critical in achieving success in any machine learning project. Technology is just a tool. Mastering it for its own sake or to beef up your resume is pointless.

In addition, you should also obsess about automation, without which you can't scale your operations. Your data science teams should be able to work completely on their own, training, deploying, and testing their models in their own account. Ops team should only be involved in managing production infrastructure, not dev and test infrastructure. Of course, you need to define the appropri-

ate quality gates to prevent bad models from finding their way into production. If manual approval works best for you, that's fine, but everything else should be automated.

Last but not least, please make sure that you've got your security covered. Here too, best practices should apply (principle of least privilege, encryption, auditing, etc.). AWS has lots of security-related services (IAM, KMS, CloudTrail, S3 bucket policies, etc.), and you should definitely look into them to build a secure data science and machine learning platform.

Q: How can people get ahold of you and what would you like to share that you are working on?

A: I'm always happy to connect on LinkedIn, on Twitter @julsimon, or YouTube, where I build and share as much content as I can, in order to help developers be successful with machine learning on AWS!

Conclusion

This chapter covers both well-traveled roads and unique corners of AWS. A key takeaway is that AWS is the largest cloud platform. There are many different ways to approach a problem using AWS technology, from building an entire company that does machine learning, as the case studies discuss, to using high-level Computer Vision APIs.

In particular, for business and technical leaders, I would recommend the following best practices to bootstrap MLOps capabilities as quickly as possible:

- Engage with AWS Enterprise Support.
- Get your team certified on AWS starting with the AWS Solutions Architect or Cloud Practitioner exam and AWS Certified Machine Learning Specialty.
- Obtain quick wins by using AI APIs like AWS Comprehend, AWS Rekognition, and high-level PaaS offerings like AWS App Runner or

AWS Lambda.
- Focus on automation and ensure that you automate everything you can, from the data ingestion and the feature store to the modeling and the deployment of the ML model.
- Start using SageMaker as an MLOps long-term investment and use it for longer term or more complex projects alongside more accessible solutions.

Finally, if you are serious about using AWS as an individual, including starting a career as an AWS machine learning expert, it can be beneficial and lucrative to get certified. Appendix B gives you a quick primer on how to prepare for the AWS Certification exams. A final recommended task would be to go through some exercises and critical thinking questions to practice AWS further.

Our next chapter moves away from AWS and digs into Azure.

Exercises

- Build a machine learning continuous delivery pipeline for a Flask web service using Elastic Beanstalk. You can refer to the GitHub repository as a starting point.
- Start an Amazon SageMaker instance and build and deploy the US census data for the population segmentation example.
- Build a CaaS machine learning prediction service using AWS Fargate. You can use this GitHub repo as a starting point.
- Build a serverless data engineering prototype using this GitHub repo as a starting point.
- Build a computer vision trigger that detects labels; use this GitHub repo as a starting point.
- Use the MLOps Cookbook base project and deploy to as many different targets as you can: Containerized CLI, EKS, Elastic Beanstalk, Spot Instances, and anything else you can think of.

Critical Thinking Discussion Questions

- Why do organizations doing machine learning use a data lake? What is the core problem they solve?
- What is a use case for using prebuilt models like AWS Comprehend versus training your sentiment analysis model?
- Why would an organization use AWS SageMaker versus Pandas, sklearn, Flask, and Elastic Beanstalk? What are the use cases for both?
- What are the advantages of a containerized machine learning model deployment process?
- A colleague says they are confused about where to start with machine learning on AWS due to the variety of offerings. How would you recommend they approach their search?