An Introduction to Amazon SageMaker for Beginners

A comprehensive guide on how to get started with Amazon SageMaker.

[Ram Vegiraju](#)

Published in

AWS in Plain English

.

4 min read

.

Dec 15, 2021

58

Image from Unsplash by Glenn Carstens-Peters

As Reinvent recently took place, Amazon SageMaker continues to expand its offerings. Amazon SageMaker is a managed Machine Learning Service that helps users scale up different parts of the ML lifecycle with a variety of different features. The power it provides is incredible, but it can be a little difficult to navigate such a large set of offerings at first. With this article, I'll cover some of the main features that SageMaker provides for ML at scale. We can navigate what features can be used for what, and I'll provide resources and examples to get started.

The **three** main **ML domains** I'll be covering are **MLOps**, **Training**, and **Inference**. All of these intersect with each other at one point or another, so a lot of real-life solutions will be a mix of the SageMaker features we dissect here.

Table of Contents

1. MLOps

2. Training

3. Inference

1. MLOps

MLOps is a must for deploying ML solutions at scale and in production. The main MLOps offering SageMaker provides is SageMaker Pipelines. You can integrate different features such as SageMaker Model Registry and SageMaker Experiments to build full-fledged ML solutions.

SageMaker Pipelines

With SageMaker Pipelines you can create ML Workflows using the SageMaker Python SDK. You can work with the SDK in SageMaker Studio to visualize your workflows in a DAG (visual) that tracks all the steps of your workflow. All SageMaker Pipelines are composed of steps, you can define specific steps such as Processing, Training, CreateModel, and RegisterModel. Through these steps, you create a Pipeline that is accompanied by a visual DAG displaying your workflow.

Pipelines Resources

Pipeline Documentation

End to End Guide

Pipeline Examples

SageMaker Model Registry

ML is an iterative process, you're rarely going to have just one model. To be able to track your models, SageMaker provides Model Registry. With Model Registry you can create a Model Group. You can create a SageMaker Pipeline which contains a RegisterModel step. Here you interface with Model Registry and for every execution of the Pipeline, you can register a model version in the Model Group that you've created.

Model Registry Resources

Model Registry Documentation

RegisterModel Pipeline Step

SageMaker Experiments

SageMaker Experiments lets you track your various ML iterations as Experiments. Each iteration's inputs, parameters, and configurations are captured in entities known as Trials. Trials will then contain Trial Components, which can be training, preprocessing, and different types of jobs that you can compare. You can view these components and order them in the UI through metrics such as RMSE based on the type of model that you are working with. For Experiments, there is a specific SDK.

Experiments Resources

Example and Documentation

End to End Example and Guide

2. Training

With SageMaker Training, we need to understand how SageMaker operates at the core. The simplest method to train SageMaker is known as Script Mode. Everything in SageMaker is containerized. To make life easier, SageMaker provides managed containers and estimators for popular ML frameworks such as TensorFlow, PyTorch, SkLearn, and HuggingFace. You can retrieve these images using the SageMaker Python SDK and simply provide a script (Script Mode) with your model building and training code. In the case that the framework is not supported by SageMaker, we can go with a Bring Your Own Container approach where you provide a Dockerfile with whatever dependencies are necessary. Both of these paradigms can be used for both training and inference.

The other key portion to understand with SageMaker Training is Distributed Training. There are two main different options with Distributed Training: Data and Model Parallelism.

Training Resources

SageMaker Training Compiler Introduction

Script Mode Examples

Distributed Training Examples

3. Inference

SageMaker currently supports four different inference options to pick from based on your use case.

Real-Time Inference

Real-Time Inference is an ideal option when you are dealing with **stringent latency requirements**. With Real-Time Inference you can create a **persistent endpoint**, that you can AutoScale and optimize for performance as necessary. Within Real-Time Inference there are a plethora of sub-options: Multi-Model Endpoints, Multi-Container Endpoints, and Serial Inference Pipelines.

Batch Transform

Batch Transform is meant for large **offline predictions** with no need for a persistent endpoint. No endpoint is created, rather a Transformer object is used to take in and output large datasets.

Asynchronous Inference

Asynchronous Inference is the inference option to use when you have long preprocessing times, large payload sizes, and near real-time latency requirements. This is especially ideal for NLP and Computer

Vision workloads where you're dealing with large text and visual datasets that will take longer to process.

Serverless Inference

SageMaker's latest inference option is available in Preview at the moment. Serverless Inference is ideal for **intermittent workloads** that can tolerate a **cold-start** and have **idle periods between traffic spurts**. The need to manage infrastructure is removed as the scaling and provisioning of instances is taken care of underneath the hood.

Inference Resources

SageMaker Inference Examples

SageMaker Inference Documentation

Conclusion

These were just a subset of the vast amount of offerings SageMaker has. Through utilizing a combination of the features we've discussed there's a variety of ML solutions that can be built. I've attached a few additional resources below to help get started.

Additional Resources

All SageMaker Offerings

SageMaker Examples

If you enjoyed this article feel free to connect with me on LinkedIn and subscribe to my Medium Newsletter. If you're new to Medium, sign up using my Membership Referral.

More content at plainenglish.io. Sign up for our free weekly newsletter here.