What SageMaker Inference Option To Use For Model Hosting

An overview of the different SageMaker hosting options.

[Ram Vegiraju](#)
Published in

AWS in Plain English

.
3 min read
.
Oct 27, 2021
59

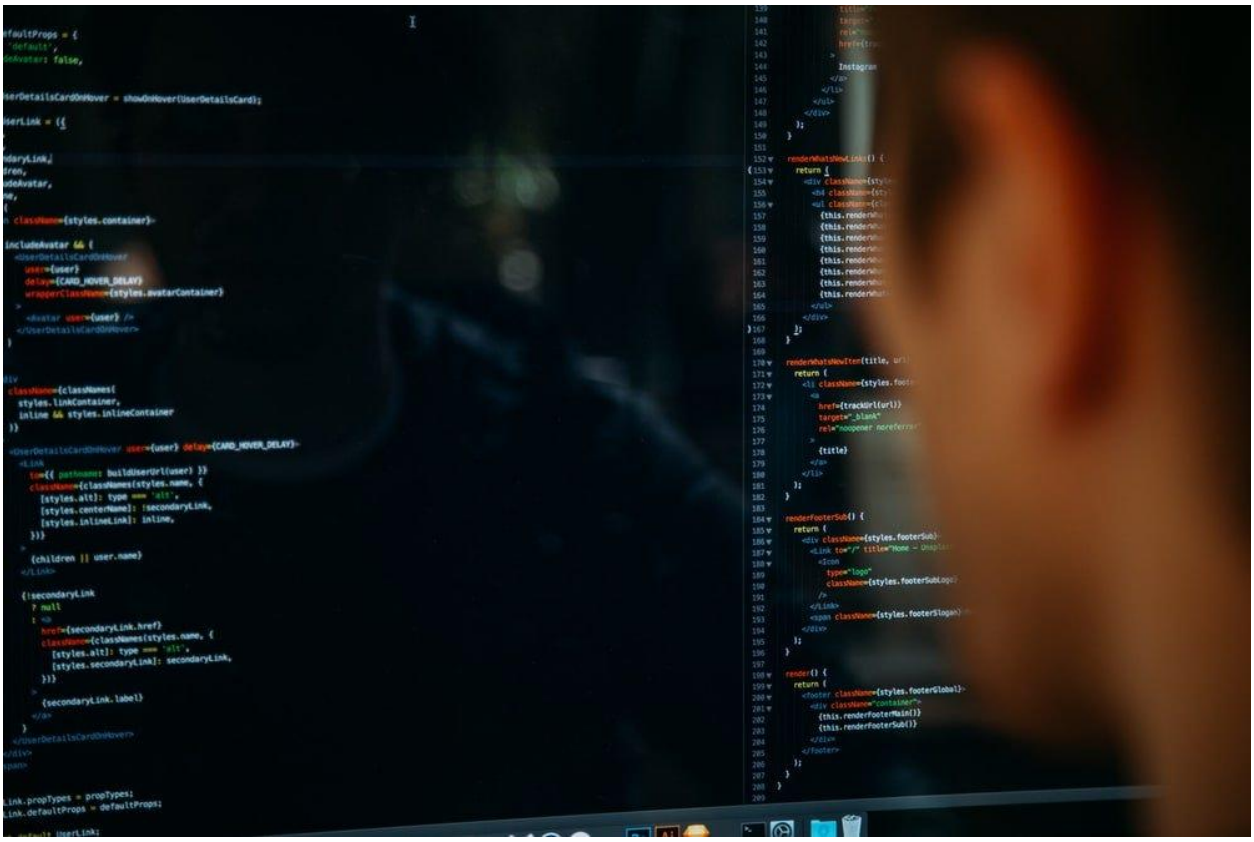

Image from Unsplash by Charles Deluvio

The end goal of any ML project is to have a point of inference for the model that you've built. With Amazon SageMaker there are currently three different inference options available: **Real-Time**, **Batch Transform**, and **Asynchronous Inference**. In this article, we'll quickly explore what each option entails and what would be the best solution for your model hosting.

Table of Contents

3. Batch Transform
4. Asynchronous Inference
5. Conclusion

Prerequisites

This article will assume an intermediate knowledge of SageMaker but also other services such as S3 and ECR. To get an introduction to SageMaker Inference follows the developer documentation for more detailed information on hosting within SageMaker. For this article, we'll be focusing purely on the Inference portion of SageMaker.

Real-Time Inference

Real-Time Inference is your go-to choice when you need a **persistent endpoint**. This option is critical for applications that need an endpoint with **certain latency and throughput requirements**. You can create a Real-Time Endpoint that is fully managed by SageMaker and comes with AutoScaling policies that you can configure based on your traffic.

There are **three main steps** in creating a real-time endpoint within SageMaker.

1. Create Model
2. Create Endpoint Configuration

3. Create Endpoint

Along with singular endpoints, SageMaker offers capabilities known as Multi-Model Endpoints (MME) and Multi-Container Endpoints (MCE) that can also be used for real-time inference. **MME** can be utilized when you have **various models** of the **same ML framework** (TensorFlow, PyTorch, etc) that you want to be invoked on the same endpoint. **MCE** can be utilized when you have **different ML frameworks** that you're using to create your models.

For examples of real-time inference check out the following repository.

Batch Transform

Batch Transform can be utilized when you **don't need a persistent endpoint** and want **inference on a large dataset**. Working with Batch Transform is a little different from creating a Real-Time Endpoint. The key in Batch Transform is **creating a Transformer object** and a **Batch Transform job** that will perform inference on a large dataset.

To explore an end to end example and explanation of Batch Transform check out this page. For examples with different ML frameworks check out the SageMaker Examples repository.

Asynchronous Inference

Asynchronous Inference is one of the newer SageMaker features but it is very similar to a real-time endpoint in both essence and creation. With Asynchronous Inference you can **queue incoming requests**. This use case is ideal when you're working with **large preprocessing times** and **near real-time workloads** in regards to latency. It can especially be handy in cases with NLP and Computer Vision where there are large payloads that require a lot of preprocessing.

The main change in creating an Asynchronous Endpoint from a Real-Time endpoint is in the Endpoint Configuration step. Here you can specify that you are dealing with an Asynchronous Endpoint Configuration. For examples of Asynchronous Configuration check out the following repository.

Conclusion & Additional Resources

The inference is crucial to bringing Machine Learning to life. It's even more essential to have the proper inference option for your use case. I'm currently **building a repository of examples for these different options in popular frameworks** that can be easily accessible and duplicated. In the meantime also follow the SageMaker Examples repository and check out the different options for your use case.

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