

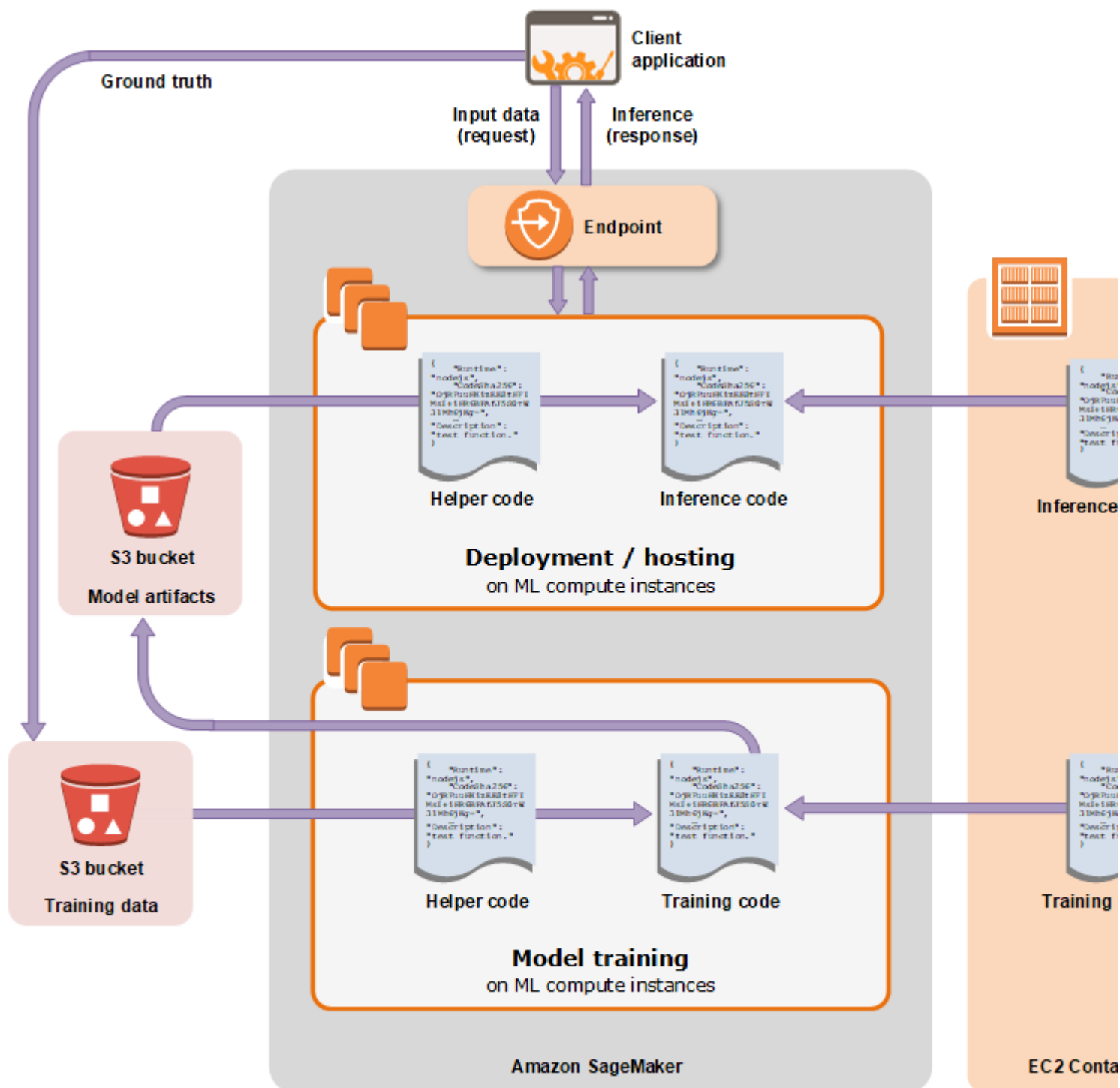
Deploy a Model on Amazon SageMaker Hosting Services

PDF ([sagemaker-dg.pdf#how-it-works-hosting](#))

Kindle (<https://www.amazon.com/dp/B07JVSB59J>)

RSS ([amazon-sagemaker-release-notes.rss](#))

Amazon SageMaker also provides model hosting services for model deployment, as shown in the following diagram. Amazon SageMaker provides an HTTPS endpoint where your machine learning model is available to provide inferences.



Deploying a model using Amazon SageMaker hosting services is a three-step process:

1. **Create a model in Amazon SageMaker**—By creating a model, you tell Amazon SageMaker where it can find the model components. This includes the S3 path where the model artifacts are stored and the Docker registry path for the image that contains the inference code. In subsequent deployment steps, you specify the model by name. For more information, see the [CreateModel \(./API_CreateModel.html\)](#) API.


2. **Create an endpoint configuration for an HTTPS endpoint**—You specify the name of one or more models in production variants and the ML compute instances that you want Amazon SageMaker to launch to host each production variant.

When hosting models in production, you can configure the endpoint to elastically scale the deployed ML compute instances. For each production variant, you specify the number of ML compute instances that you want to deploy. When you specify two or more instances, Amazon SageMaker launches them in multiple Availability Zones. This ensures continuous availability. Amazon SageMaker manages deploying the instances. For more information, see the [CreateEndpointConfig \(./API_CreateEndpointConfig.html\)](#) API.

3. **Create an HTTPS endpoint**—Provide the endpoint configuration to Amazon SageMaker. The service launches the ML compute instances and deploys the model or models as specified in the configuration. For more information, see the [CreateEndpoint \(./API_CreateEndpoint.html\)](#) API. To get inferences from the model, client applications send requests to the Amazon SageMaker Runtime HTTPS endpoint. For more information about the API, see the [InvokeEndpoint \(./API_runtime_InvokeEndpoint.html\)](#) API.

Note

Endpoints are scoped to an individual AWS account, and are not public. The URL does not contain the account ID, but Amazon SageMaker determines the account ID from the authentication token that is supplied by the caller.

For an example of how to use Amazon API Gateway and AWS Lambda to set up and deploy a web service that you can call from a client application that is not within the scope of your account, see [Call an Amazon SageMaker model endpoint using Amazon API Gateway and AWS Lambda](#)  (<https://aws.amazon.com/blogs/machine-learning/call-an-amazon-sagemaker-model-endpoint-using-amazon-api-gateway-and-aws-lambda/>) in the *AWS Machine Learning Blog*.

Note

When you create an endpoint, Amazon SageMaker attaches an Amazon EBS storage volume to each ML compute instance that hosts the endpoint. The size of the storage volume depends on the instance type. For a list of instance types that Amazon SageMaker hosting service supports, see [AWS Service Limits](#) (https://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html#limits_sagemaker) . For a list of the sizes of the storage volumes that Amazon SageMaker attaches to each instance, see [Host Instance Storage Volumes \(./host-instance-storage.html\)](#) .

To increase a model's accuracy, you might choose to save the user's input data and ground truth, if available, as part of the training data. You can then retrain the model periodically with a larger, improved training dataset.

Best Practices for Deploying Models on Amazon SageMaker Hosting Services

When hosting models using Amazon SageMaker hosting services, consider the following:

- Typically, a client application sends requests to the Amazon SageMaker HTTPS endpoint to obtain inferences from a deployed model. You can also send requests to this endpoint from your Jupyter notebook during testing.
- You can deploy a model trained with Amazon SageMaker to your own deployment target. To do that, you need to know the algorithm-specific format of the model artifacts that were generated by model training. For more

information about output formats, see the section corresponding to the algorithm you are using in [Training Data Formats \(./cdf-training.html#td-serialization\)](#) .

- You can deploy multiple variants of a model to the same Amazon SageMaker HTTPS endpoint. This is useful for testing variations of a model in production. For example, suppose that you've deployed a model into production. You want to test a variation of the model by directing a small amount of traffic, say 5%, to the new model. To do this, create an endpoint configuration that describes both variants of the model. You specify the `ProductionVariant` in your request to the `CreateEndpointConfig`. For more information, see [ProductionVariant \(./API_ProductionVariant.html\)](#) .
- You can configure a `ProductionVariant` to use Application Auto Scaling. For information about configuring automatic scaling, see [Automatically Scale Amazon SageMaker Models \(./endpoint-auto-scaling.html\)](#) .
- You can modify an endpoint without taking models that are already deployed into production out of service. For example, you can add new model variants, update the ML Compute instance configurations of existing model variants, or change the distribution of traffic among model variants. To modify an endpoint, you provide a new endpoint configuration. Amazon SageMaker implements the changes without any downtime. For more information see, [UpdateEndpoint \(./API_UpdateEndpoint.html\)](#) and [UpdateEndpointWeightsAndCapacities \(./API_UpdateEndpointWeightsAndCapacities.html\)](#) .
- Changing or deleting model artifacts or changing inference code after deploying a model produces unpredictable results. If you need to change or delete model artifacts or change inference code, modify the endpoint by providing a new endpoint configuration. Once you provide the new endpoint configuration, you can change or delete the model artifacts corresponding to the old endpoint configuration.
- If you want to get inferences on entire datasets, consider using batch transform as an alternative to hosting services. For information, see [Get Inferences for an Entire Dataset with Batch Transform \(./how-it-works-batch.html\)](#)

How It Works: Next Topic

[Validate a Machine Learning Model \(./how-it-works-model-validation.html\)](#)