Amazon SageMaker

SageMaker is a fully managed service to prepare data and build, train, and deploy machine learning (ML) models for any use case with fully managed infrastructure, tools, and workflows.

AWS Regions

For a list of the supported SageMaker Regions, please visit the AWS Regional Services page. Also, for more information, see Regional endpoints in the AWS general reference guide.

Service availability of SageMaker

SageMaker is designed for high availability. There are no maintenance windows or scheduled downtimes. SageMaker APIs run in Amazon proven high-availability data centers, with service stack replication configured across three facilities in each Region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Code Security

SageMaker stores code in ML storage volumes, secured by security groups and optionally encrypted at rest.

Security measures

SageMaker ensures that ML model artifacts and other system artifacts are encrypted in transit and at rest. Requests to the SageMaker API and console are made over a secure (SSL) connection. You pass AWS Identity and Access Management roles to SageMaker to provide permissions to access resources on your behalf for training and deployment. You can use encrypted Amazon Simple Storage Service (Amazon S3) buckets for model artifacts and data, as well as pass an AWS Key Management Service (AWS KMS) key to SageMaker notebooks, training jobs, and endpoints to encrypt the attached ML storage volume. SageMaker also supports Amazon Virtual Private Cloud (Amazon VPC) and AWS PrivateLink support.

Sharing models, training data, or algorithms

SageMaker does not use or share customer models, training data, or algorithms. We know that customers care deeply about privacy and data security. That's why AWS gives you ownership and control over your content through simplified, powerful tools that allow you to determine where your content will be stored, secure your content in transit and at rest, and manage your access to AWS services and resources for your users. We also implement technical and physical controls that are designed to prevent unauthorized access

to or disclosure of your content. As a customer, you maintain ownership of your content, and you select which AWS services can process, store, and host your content. We do not access your content for any purpose without your consent.

SageMaker Charges

You pay for ML compute, storage, and data processing resources that you use for hosting the notebook, training the model, performing predictions, and logging the outputs. With SageMaker, you can select the number and type of instance used for the hosted notebook, training, and model hosting. You pay only for what you use, as you use it; there are no minimum fees and no upfront commitments. For more details, see Amazon SageMaker Pricing and the Amazon SageMaker Pricing Calculator.

Cost Optimizations

There are several best practices that you can adopt to optimize your SageMaker resource usage. Some approaches involve configuration optimizations; others involve programmatic solutions. A full guide on this concept, complete with visual tutorials and code samples, can be found in this blog post.

Development environment

SageMaker provides a full and complete workflow, but you can continue using your existing tools with SageMaker. You can easily transfer the results of each stage in and out of SageMaker as your business requirements dictate.

Language support for R

You can use R within SageMaker notebook instances, which include a preinstalled R kernel and the reticulate library. Reticulate offers an R interface for the Amazon SageMaker Python SDK, helping ML practitioners build, train, tune, and deploy R models.

Model imbalances

Amazon SageMaker Clarify helps improve model transparency by detecting statistical bias across the entire ML workflow. SageMaker Clarify checks for imbalances during data preparation, after training, and ongoing over time, and also includes tools to help explain ML models and their predictions. Findings can be shared through explainability reports.

Bias detection with SageMaker Clarify

Measuring bias in ML models is a first step to mitigating bias. Bias may be measured before training and after training, as well as for inference for a deployed model. Each measure of bias corresponds to a different

notion of fairness. Even considering simple notions of fairness leads to many different measures applicable in various contexts. You must choose bias notions and metrics that are valid for the application and the situation under investigation. SageMaker currently supports the computation of different bias metrics for training data (as part of SageMaker data preparation), for the trained model (as part of Amazon SageMaker Experiments), and for inference for a deployed model (as part of Amazon SageMaker Model Monitor). For example, before training, we provide metrics for checking whether the training data is representative (that is, whether one group is underrepresented) and whether there are differences in the label distribution across groups. After training or during deployment, metrics can be helpful to measure whether (and by how much) the performance of the model differs across groups. For example, start by comparing the error rates (how likely a model's prediction is to differ from the true label) or break further down into precision (how likely a positive prediction is to be correct) and recall (how likely the model will correctly label a positive example).

Model explainability with SageMaker Clarify

SageMaker Clarify is integrated with SageMaker Experiments to provide a feature importance graph detailing the importance of each input for your model’s overall decision-making process after the model has been trained. These details can help determine if a particular model input has more influence than it should on overall model behavior. SageMaker Clarify also makes explanations for individual predictions available through an API.

Amazon SageMaker Studio

SageMaker Studio provides a single, web-based visual interface where you can perform all ML development steps. SageMaker Studio gives you complete access, control, and visibility into each step required to prepare data and build, train, and deploy models. You can quickly upload data, create new notebooks, train and tune models, move back and forth between steps to adjust experiments, compare results, and deploy models to production all in one place, making you much more productive. All ML development activities including notebooks, experiment management, automatic model creation, debugging and profiling, and model drift detection can be performed within the unified SageMaker Studio visual interface.

RStudio on Amazon SageMaker

RStudio on SageMaker is the first fully managed RStudio Workbench in the cloud. You can quickly launch the familiar RStudio integrated development environment (IDE) and dial up and down the underlying compute resources without interrupting your work, making it easier to build ML and analytics solutions in R at scale. You can seamlessly switch between the RStudio IDE and SageMaker Studio notebooks for R and Python development. All your work, including code, datasets, repositories, and other artifacts, is

automatically synchronized between the two environments to reduce context switch and boost productivity.

SageMaker Studio pricing

There is no additional charge for using SageMaker Studio. You pay only for the underlying compute and storage charges on the services that you use within SageMaker Studio.

Regions support for SageMaker Studio

You can find the Regions where SageMaker Studio is supported in the Amazon SageMaker Developer Guide.

ML governance

ML governance tools in SageMaker

SageMaker provides purpose-built ML governance tools across the ML lifecycle. With Amazon SageMaker Role Manager, administrators can define minimum permissions in minutes. Amazon SageMaker Model Cards makes it easier to capture, retrieve, and share essential model information from conception to deployment, and Amazon SageMaker Model Dashboard keeps you informed on production model behavior, all in one place. For more information, see ML Governance with Amazon SageMaker.

SageMaker Role Manager

You can define minimum permissions in minutes with SageMaker Role Manager. It provides a baseline set of permissions for ML activities and personas with a catalog of pre-built IAM policies. You can keep the baseline permissions, or customize them further based on your specific needs. With a few self-guided prompts, you can quickly input common governance constructs such as network access boundaries and encryption keys. SageMaker Role Manager will then generate the IAM policy automatically. You can discover the generated role and associated policies through the AWS IAM console. To further tailor the permissions to your use case, attach your managed IAM policies to the IAM role that you create with SageMaker Role Manager. You can also add tags to help identify the role and organize across AWS services.

SageMaker Model Cards

SageMaker Model Cards helps you centralize and standardize model documentation throughout the ML lifecycle by creating a single source of truth for model information. SageMaker Model Cards auto-populates training details to accelerate the documentation process. You can also add details such as the purpose of the model and the performance goals. You can attach model evaluation results to your model card and

provide visualizations to gain key insights into model performance. SageMaker Model Cards can easily be shared with others by exporting to a PDF format.

SageMaker Model Dashboard

SageMaker Model Dashboard gives you a comprehensive overview of deployed models and endpoints, letting you track resources and model behavior violations through one pane. It allows you to monitor model behavior in four dimensions, including data and model quality, and bias and feature attribution drift through its integration with SageMaker Model Monitor and SageMaker Clarify. SageMaker Model Dashboard also provides an integrated experience to set up and receive alerts for missing and inactive model monitoring jobs, and deviations in model behavior for model quality, data quality, bias drift, and feature attribution drift. You can further inspect individual models and analyze factors impacting model performance over time. Then, you can follow up with ML practitioners to take corrective measures.

Foundation models

Getting started

SageMaker JumpStart helps you quickly and easily get started with ML. SageMaker JumpStart provides a set of solutions for the most common use cases that can be deployed readily in just a few steps. The solutions are fully customizable and showcase the use of AWS CloudFormation templates and reference architectures so you can accelerate your ML journey. SageMaker JumpStart also provides foundation models and supports one-step deployment and fine-tuning of more than 150 popular open-source models, such as transformer, object detection, and image classification models.