

DEPLOY HUGGINGFACE MODELS IN SAGEMAKER

**8 STEPS TO
GET
INFERENCE
ENDPOINT**



aws



WHAT CHALLENGE SAGEMAKER SOLVES & HOW

LABEL DATA

SAGEMAKER STUDIO: LETS YOU BUILD, TRAIN, DEBUG, DEPLOY, AND MONITOR YOUR MACHINE LEARNING MODELS.

BUILD

SAGEMAKER **NOTEBOOK** INSTANCES: PREPARE, PROCESS DATA, TRAIN & DEPLOY MACHINE LEARNING MODELS FROM A COMPUTE INSTANCE RUNNING THE JUPYTER NOTEBOOK APPLICATION. (VERY SIMILAR TO COLAB ENVIRONMENT)

TRAIN

SAGEMAKER **STUDIO LAB:** STUDIO LAB **IS A FREE SERVICE THAT GIVES YOU ACCESS TO AWS COMPUTE RESOURCES**, IN AN ENVIRONMENT BASED ON OPEN-SOURCE JUPYTERLAB, WITHOUT REQUIRING AN AWS ACCOUNT.

TUNE

SAGEMAKER CANVAS: GIVES YOU THE ABILITY TO USE MACHINE LEARNING TO GENERATE PREDICTIONS WITHOUT NEEDING TO CODE.

DEPLOY

SAGEMAKER **GEOSPATIAL:** GIVES YOU THE ABILITY TO BUILD, TRAIN, AND DEPLOY GEOSPATIAL MODELS.

DISCOVER

RSTUDIO: RSTUDIO IS AN IDE FOR R, WITH A CONSOLE, SYNTAX-HIGHLIGHTING EDITOR THAT SUPPORTS DIRECT CODE EXECUTION, AND TOOLS FOR PLOTTING, HISTORY, DEBUGGING AND WORKSPACE MANAGEMENT.

STEPS TO DEPLOY THE MODELS

8 STEPS:

- 1.CREATE ROLE
- 2.CREATE DOMAIN
- 3.CREATE USER
- 4.CREATE STUDIO INSTANCE
- 5.UNDERSTAND SAGEMAKER CLASSES
- 6.PULL THE MODEL & STORE IN S3
- 7.CREATE INFERENCE END POINT
- 8.PREDICT

CONNECTED WITH:

S3 BUCKETS,
HUGGING FACE HUB,
GIT REPOSITORIES
LINUX USERS

ENVIRONMENT:

DOMAIN,
USERPROFILE
SHARED SPACE
APP

MODELS ACTIVITIES:

SAGEMAKER STUDIO,
SAGEMAKER STUDIO NOTEBOOKS,
RSTUDIO

CHOICES TO BE MADE

Amazon SageMaker capability	Free Tier usage per month for the first 2 months
Studio notebooks, and notebook instances	250 hours of ml.t3.medium instance on Studio notebooks OR 250 hours of ml.t2 medium instance or ml.t3.medium instance on notebook instances
RStudio on SageMaker	250 hours of ml.t3.medium instance on RSession app AND free ml.t3.medium instance for RStudioServerPro app
Data Wrangler	25 hours of ml.m5.4xlarge instance
Feature Store	10 million write units, 10 million read units, 25 GB storage
Training	50 hours of m4.xlarge or m5.xlarge instances
Real-Time Inference	125 hours of m4.xlarge or m5.xlarge instances
Serverless Inference	150,000 seconds of inference duration
Canvas	750 hours/month for session time, and up to 10 model creation requests/month, each with up to 1 million cells/model creation request
Free Tier usage per month for the first 6 months	
Experiments	100,000 metric records ingested per month, 1 million metric records retrieved per month, and 100,000 metric records stored per month

ml.m5.xlarge	4	16 GiB	\$0.23
ml.g4dn.xlarge	4	16 GiB	\$0.94

AMAZON SAGEMAKER HOSTING: PROVIDES REAL-TIME INFERENCE FOR YOUR USE CASES NEEDING REAL-TIME PREDICTIONS. YOU ARE CHARGED FOR USAGE OF THE **INSTANCE TYPE** YOU CHOOSE. **BUILT-IN RULES**, YOU GET UP TO 30 HOURS OF MONITORING AT NO CHARGE. CHARGES WILL BE BASED ON DURATION OF USAGE. YOU ARE CHARGED SEPARATELY WHEN YOU USE YOUR OWN CUSTOM RULES.

Amazon SageMaker

Build, train, and deploy machine learning models at scale

The quickest and easiest way to get ML models from idea to production.

New to SageMaker?

Get started with Amazon SageMaker by completing the quick start guide.

Get Started

Configure role

Establish roles and quickly configure permission policies for your users.

Create a role

Configure SageMaker Domain

Quick setup (1 min)

Standard setup (10 min)



A quick setup for a single user that allows full use of SageMaker in less than one minute.

Set up SageMaker Domain

User profile

Name

insight01

The name can have up to 63 characters. Valid characters: A-Z, a-z, 0-9, and - (hyphen)

Execution role

The default execution role for both users and spaces in the domain. The execution role must have the [AmazonSageMakerServiceCatalogProductsExecutionRole](#) attached.

AmazonSageMakerServiceCatalogProductsExecutionRole

Create role using the role creation wizard

☒ Enable SageMaker Canvas permissions [Info](#)

Enable SageMaker Canvas permissions to attach the [AmazonSageMakerCanvasFullAccess](#) policy to the default execution role. This policy enables the following features in Canvas: 1) Ready-to-use models - The [AmazonSageMakerCanvasAIServiceCatalogProductsExecutionRole](#) attached to the default execution role. 2) Time series forecasting - The [AmazonSagemakerCanvasForecastRolePolicy](#) attached to the default execution role called AmazonSageMakerCanvasForecastRole.

Step 1

Enter role information

Step 2
Configure ML activities

Step 3
Add additional policies & tags

Step 4
Review role

Enter role information

A role is an IAM identity that has permissions to perform actions with AWS services.

Custom role settings

No preselected settings or default options. Complete customization starting with empty settings.

Data Scientist

A persona that performs machine learning activities from within a sagemaker environment. Permitted to process S3 data, perform experiments and produce models.

MLOps

A persona that deals with operational activities from within a sagemaker environment. Permitted to manage models, endpoints and pipelines, and audit resources.

SageMaker Compute Role

A persona used by SageMaker compute resources such as jobs and endpoints. Permitted to access S3 resources, ECR repositories, Cloudwatch, and other services for ML computation.

Select a persona

<input type="checkbox"/>	Name	Description
<input type="checkbox"/>	Access Required AWS Services	Permissions to access S3, ECR, Cloudwatch and EC2. Required for execution roles for jobs and endpoints.
<input checked="" type="checkbox"/>	Run Studio Applications	Permissions to operate within a Studio environment. Required for domain and user-profile execution roles.
<input type="checkbox"/>	Manage ML Jobs	Permissions to manage SageMaker Jobs across their lifecycles.
<input type="checkbox"/>	Manage Models	Permissions to manage SageMaker models and Model Registry.
<input checked="" type="checkbox"/>	Manage Endpoints	Permissions to manage SageMaker Endpoint deployments and updates.
<input type="checkbox"/>	Manage Pipelines	Permissions to manage SageMaker Pipelines and pipeline executions.
<input type="checkbox"/>	Manage Experiments	Permissions to manage experiments and trials.
<input checked="" type="checkbox"/>	Search and visualize experiments	Permissions to audit, query lineage and visualize experiments.
<input type="checkbox"/>	Manage Model Monitoring	Permissions to manage monitoring schedules for SageMaker Model Monitor.
<input checked="" type="checkbox"/>	S3 Full Access	Permissions to perform all S3 operations
<input type="checkbox"/>	S3 Bucket Access	Permissions to perform operations on specified buckets.
<input type="checkbox"/>	Query Athena Workgroups	Permissions to execute and manage Amazon Athena queries.
<input type="checkbox"/>	Manage Glue Tables	Permissions to create and manage Glue tables for SageMaker Feature Store and Data Wrangler.

CODE TO EXECUTE IN SAGEMAKER

HUB MODEL CONFIGURATION.

```
HTTPS://HUGGINGFACE.CO/MODELS
```

```
HUB = {  
    'HF_MODEL_ID': 'DISTILBERT-BASE-UNCASED-FINETUNED-SST-2-ENGLISH',  
    'HF_TASK': 'TEXT-CLASSIFICATION'  
}
```

CREATE HUGGING FACE MODEL CLASS

```
HUGGINGFACE_MODEL = HUGGINGFACEMODEL(  
    TRANSFORMERS_VERSION='4.17.0',  
    PYTORCH_VERSION='1.10.2',  
    PY_VERSION='PY38',  
    ENV=HUB,  
    ROLE=ROLE,  
)
```

DEPLOY MODEL TO SAGEMAKER INFERENCE

```
PREDICTOR = HUGGINGFACE_MODEL.DEPLOY(  
    INITIAL_INSTANCE_COUNT=1, # NUMBER OF INSTANCES  
    INSTANCE_TYPE='ML.M5.XLARGE' # EC2 INSTANCE TYPE  
)
```

PUBLIC S3 URI TO GPT-J ARTIFACT

```
MODEL_URI="S3://HUGGINGFACE-SAGEMAKER-MODELS/TRANSFORMERS/4.12.3/PYTORCH/1.9.1/GPT-J/MODEL.TAR.GZ"
```

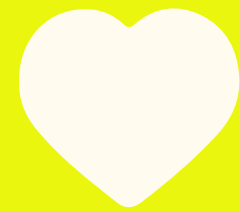
FROM TRANSFORMERS IMPORT GPTJFORCAUSALLM IMPORT TORCH

```
MODEL = GPTJFORCAUSALLM.FROM_PRETRAINED(  
    "ELEUTHERAI/GPT-J-6B",  
    REVISION="FLOAT16",  
    TORCH_DTYPE=TORCH.FLOAT16,  
    LOW_CPU_MEM_USAGE=TRUE  
)
```

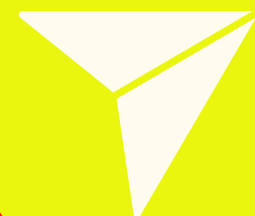
DEPLOY MODEL TO SAGEMAKER INFERENCE

```
PREDICTOR = HUGGINGFACE_MODEL.DEPLOY(  
    INITIAL_INSTANCE_COUNT=1, # NUMBER OF INSTANCES  
    INSTANCE_TYPE='ML.G4DN.XLARGE'  
    #'ML.P3.2XLARGE' # EC2 INSTANCE TYPE  
)
```

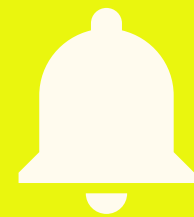
THANKS FOR WATCHING



LIKE



SHARE



SUBSCRIBE