Monetize model with JumpStart

For model provider

*Contents are not final and may change prior to GA

June Won
Product Manager for SageMaker JumpStart

Amazon SageMaker JumpStart

ML hub with foundation models, built-in algorithms, and prebuilt ML solutions that you can deploy with just a few clicks
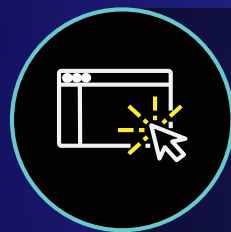
Machine learning hub

Browse through 400+ built-in algorithms with pretrained models, pretrained foundation models, solutions, and example notebooks

Pre-built training and inference scripts

Compatible with SageMaker and configurable with custom dataset

UI as well as API-based

Use the user interface for single click model deployment or API for the Python SDK-based workflow

Notebooks with examples

Jump into notebooks to use selected model with examples to guide you through the entire ML workflow

Share and collaborate within your organization

Share models and notebooks with others within your organization, and allow them to train with their own data or deploy as-is for inferencing

aws

Foundation models in JumpStart

- Model developers can easily monetize ML models by listing on Marketplace so SageMaker customers can discover and deploy them through SageMaker JumpStart.

- AWS does not expose model weights to customers with its escrow infrastructure. AWS does not expose inference data from customers to model providers.

- Model developers can choose from available pricing options such as per token and per hours of usage, and allow either specific AWS account or all users to discover and purchase.

- Once listed, models will be discoverable in JumpStart and in the AWS Marketplace. Model developer as a seller can centrally manage listed models from the seller portal in AWS Marketplace.

- SageMaker customers can deploy models into their own end-point to integrate into their SageMaker based infrastructure.

- Fine-tuning will be supported in Q3 2023 as well.

Security and intellectual property

- Seller's model weights and script codes are protected with highest levels of AWS security standards.

- When seller creates a product, the code is packaged in Docker container images. When uploading a container image, the image and artifacts are encrypted in transit and at rest. The images are also scanned for vulnerabilities before being published.

- To help safeguard your intellectual property, SageMaker allows buyers to access your product through AWS service endpoints. Buyers cannot directly access, or pull container images or model artifacts, nor can they access the underlying infrastructure.

- When buyers launch your product from AWS Marketplace, the models are deployed without network access. SageMaker deploys images in an environment with no access to the network. For example, a container image can't make outbound API calls to services on the internet, VPC endpoints, or any other AWS services.

- For buyers, your product runs in SageMaker within the buyer's SageMaker instance accessible from buyer account. So, when a buyer uses seller's product to perform data inference, you as the seller can't access their data.

Product Pricing

- Buyers pay for the *infrastructure costs* that AWS sets and *software price* that sellers set.

- **Infrastructure cost** is based on SageMaker while using your product. This is set by AWS and are available on the Amazon SageMaker pricing page.

Software pricing

- You determine the software price that AWS Marketplace charges the buyer for using your product.

- You can offer:
  - Free Pricing
  - Hourly Pricing
  - Token based / Inference based pricing (e.g., price per 1000 tokens)

- You can offer special pricing to a specific customer through private offer

- Marketplace will take 20% from software revenue

Token Based Inference Pricing

- Default Metering: Buyer is charged one unit of pricing per each invocation

- Custom Metering: Customers are charged based on volume of output generated, number of units your LLM model returns in the response

ML model-package offer

All fields required unless marked - *optional*

Pricing Info

Pricing model

- ○ Free
- ○ Hourly pricing
- ● Inference pricing (real-time)

Tax and banking information must be verified before publishing a paid product

Offer a free trial?

- ○ Yes
- ● No

Real-time inference pricing

Inference pricing metering options

Inference pricing metering option cannot be modified once the product has been created

- ○ Default metering - customers are billed one unit per invocation.
- ● Custom metering - customers are billed the number of units you return in your response header.

☑ I confirm that my model package supports the response header for custom metering.
Example response header: X-Amzn-Inference-Metering:
{"Dimension": "inference.count", "ConsumedUnits": 3}
I understand that in absence of this header, default metering will be used instead.

Inference price (real-time)

Price per inference for real-time inferences

| $ | 0 |

Terms and conditions

- Model providers sign up for Marketplace seller agreement terms
  - [Terms](#)

- Model providers upload their own end user license agreement (EULA) or use standard agreement terms provided by Marketplace
  - [Documentation](#)

Steps and process

Onboard to Marketplace as Seller	Fully self-serviced process that takes few hours [Documentation](#)
↓	
Onboard as JumpStart Provider	Give us account ID of seller account. Takes ~1 week to onboard.
↓	
List product to Marketplace	Follow our step by step process and example notebook. This can take ~2 weeks.
↓	
Surface to JumpStart	We will create model card based on product info from Marketplace. This can take ~2 weeks.

Post onboarding for premium providers (draft)

Service team support during onboarding

Quality check and feedback based on common use cases within 30 days of onboarding

Weekly check-in with key customer feedback for 60 days

Internal enablement sessions to account team and SA community

Connect with key customers that can benefit from your model

*eligibility and benefits are not final, but will need field enablement kit from provider (battlecard, use case, sample prompts, and demo)

Support model (draft)

- **For sellers**
  - Contact JumpStart team to onboard as JumpStart foundation model provider
  - Leverage AWS Marketplace seller support for Marketplace onboarding: [Link](#)
  - Work with AWS account team such as SA for more hands-on help
  - (Escalation)

- **For buyers**
  - Buyers will contact AWS support as initial contact: [Link](#)
    - AWS support will direct customers to seller, AWS Marketplace, or SageMaker depending on topic
  - Buyers will also work with AWS account team such as SA for more hands-on help
  - Buyers will contact sellers for discount and customized license terms
  - (Technical support for buyers)
  - (Escalation)

  *AWS does not provide pricing strategy or information about other sellers

Preparing your product for Marketplace listing
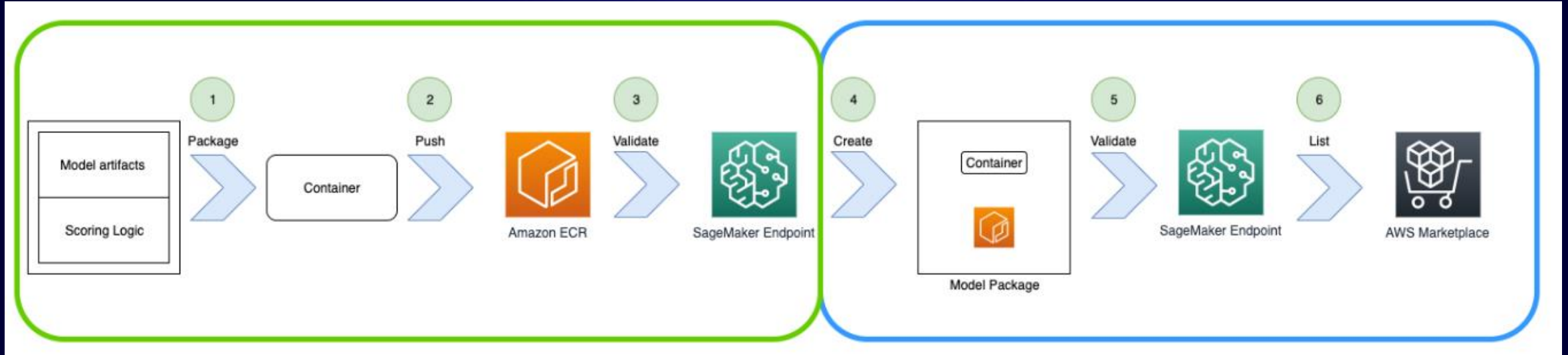
*You will be responsible for cost of holding your product on AWS (e.g., ECR cost), but there is no cost to list to Marketplace

1) A detailed step by step guide is also being sent alongside this presentation (Model Partner Jupyter Notebook)
2) Key steps are
   - Containerize your model
   - Update your container to run on SageMaker Inference
   - Create a Model Package and list it on the AWS Marketplace
   - Create a Buyer facing notebook & SDK that demonstrates how to use your model in the AWS marketplace
   - Follow AWS General guidelines to sell at the Marketplace

General Best Practices

For product descriptions, include the following:

- What your model does

- Who the target customer is

- What the most important use case is

- How your model was trained or the amount of data that was used

- What the performance metrics are and the validation data used

By default, machine learning products are configured to have public visibility. However, you can create a product with private visibility. For more information, see [Creating your product listing](#).

(Optional) For paid products, offer a free trial of 14–30 days for customers to try your product.

(Optional) For model package products, if you want to enable a real-time product demo on your product listing page, contact the AWS Marketplace Seller Operations team. The product demo allows a prospective buyer to try your model directly on the listing page without subscribing to or deploying the model themselves.

Seller roadmap (tentative)

- Preview of proprietary model in JumpStart (Current)

- General availability of proprietary model with inference only in JumpStart 6/9/2023

- Fine-tuning of proprietary model in JumpStart Q3-2023

- Self-serviced and automated onboarding in Q4-2023