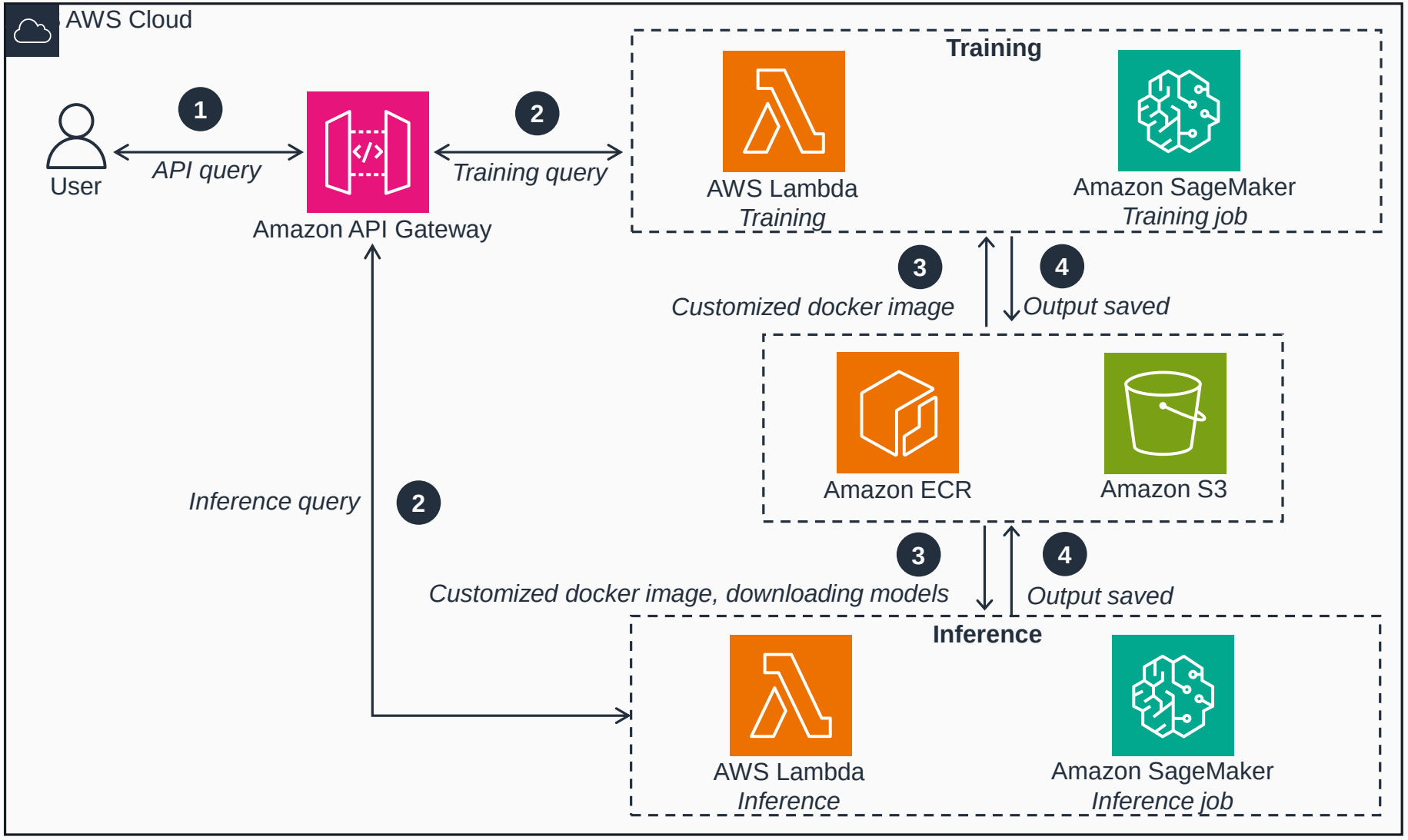
Guidance for AI-Generated Images with Stable Diffusion on AWS

This architecture diagram shows how to use Stable Diffusion APIs to decouple applications into training and inference components that are hosted on Amazon SageMaker.



1. The user submits the training or inference API query to **Amazon API Gateway**. The resource authorizers in **API Gateway** help limit the resource accessed by IPs.

2. **API Gateway** passes the training or inference query to separate components. Both components contain **AWS Lambda** and **Amazon SageMaker**. **Lambda** gets the parameters in the query, then uses the parameters to initiate **SageMaker** training or the inference job. These parameters are used to train the Stable Diffusion model or used for the Stable Diffusion model to do image generation.

3. **SageMaker** uses your customized Stable Diffusion docker image stored in **Amazon Elastic Container Registry (Amazon ECR)**. Or, it can download the pre-trained Stable Diffusion models from **Amazon Simple Storage Service (Amazon S3)** to train the Stable Diffusion model, or generate images. **SageMaker** automatically scales based on request volume.

4. The outputs of the training job are models or images. The outputs of the inference job are images. You can customize the outputs of the training jobs and inference jobs, and save the outputs into an **Amazon S3** bucket. Then, **API Gateway** returns the generated image link to the user.

Reviewed for technical accuracy February 2, 2024

AWS Reference Architecture