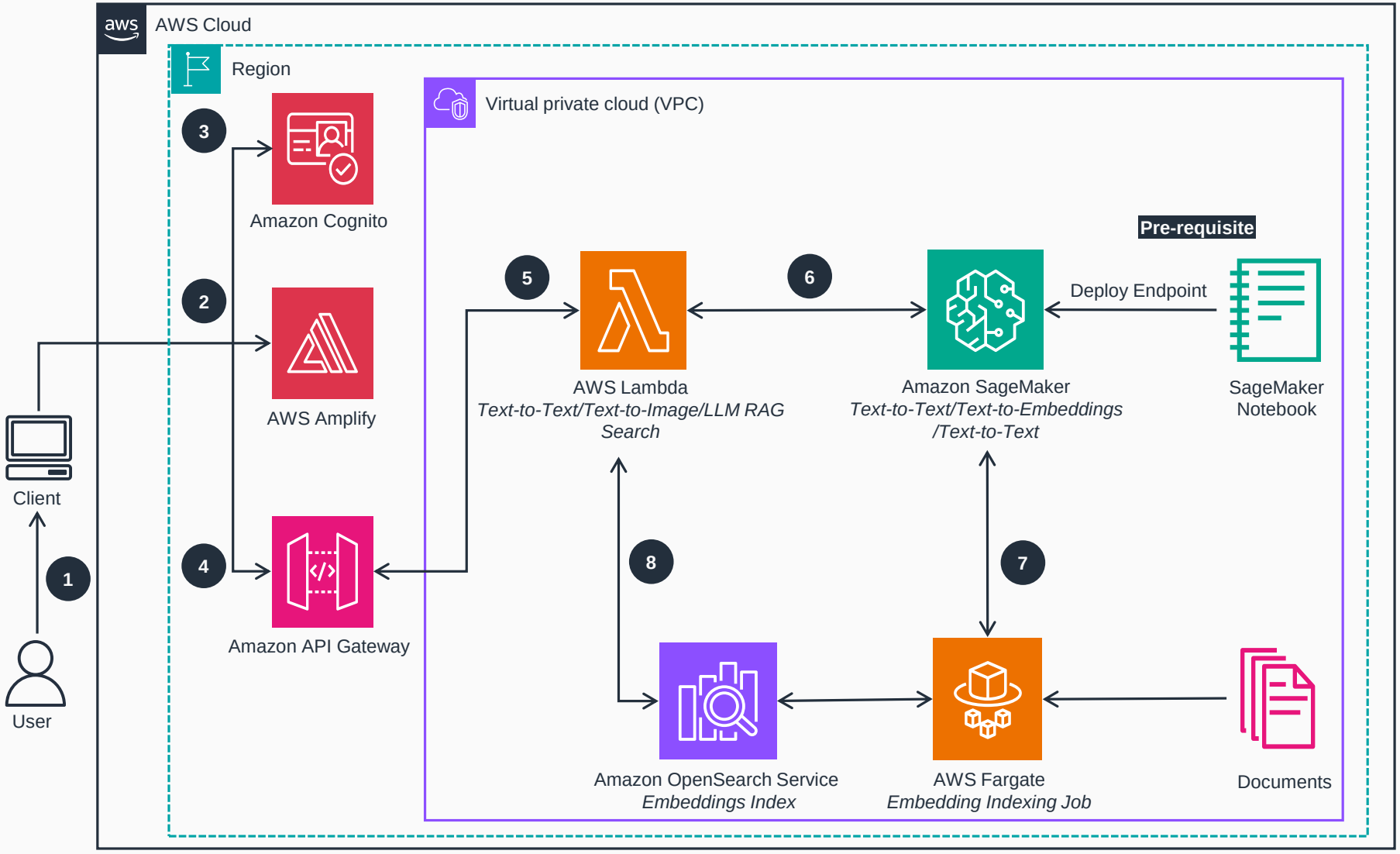
Guidance for Text Generation using Embeddings from Enterprise Data on AWS

This architecture diagram shows a secure, generative AI-based application that generates text from enterprise data.



Pre-requisite: Amazon SageMaker JumpStart is used to deploy foundational models as **Amazon SageMaker** endpoints, including Text-to-Image (Stability AI), Text-to-Text (Hugging Face Flan T5 XL), and Text-to-Embeddings (Hugging Face GPT 6B FP16) for different tasks.

1. Users access a React app with three pages: one for image prompts, one for text prompts, and one for questions that provide context-based answers from a Text-to-Text model.

2. The React app, built with **AWS Amplify libraries**, is hosted and served from an **Amplify** URL. The **Amplify** command line interface (CLI) is used to set up and deploy the app's hosting environment.

3. If a user has not been authenticated, the user will be authenticated against **Amazon Cognito** using the **Amplify** React user interface (UI) library.

4. When a user provides an input and submits the form, **Amazon API Gateway** processes the request.

5. Depending on the chosen application (Text-to-Text, Text-to-Image, or LLM RAG Search), **API Gateway** invokes the appropriate **Lambda** function. The **Lambda** function sanitizes user input and invokes the corresponding **SageMaker** endpoint, formatting prompts as needed for the language models. It also reformats the model output and returns it to the user.

6. Three distinct endpoints are deployed for Text-to-Text (Flan T5 XXL), Text-to-Embeddings (GPTJ-6B), and Text-to-Image models (Stability AI). Depending on the specific use case, these endpoints produce responses, and **Lambda** functions format the generated output.

7. **AWS Fargate** receives documents, breaks them into smaller sections, uses the Text-to-Embeddings LLM to generate embeddings, and then indexes these embeddings into **Amazon OpenSearch Service** for context-based searching.

8. The Text-to-Embeddings model creates document embeddings, which **OpenSearch Service** indexes. An index with k-Nearest Neighbor (k-NN) capability is activated, enabling efficient searching of these embeddings within **OpenSearch Service**.

Reviewed for technical accuracy October 26, 2023

AWS Reference Architecture