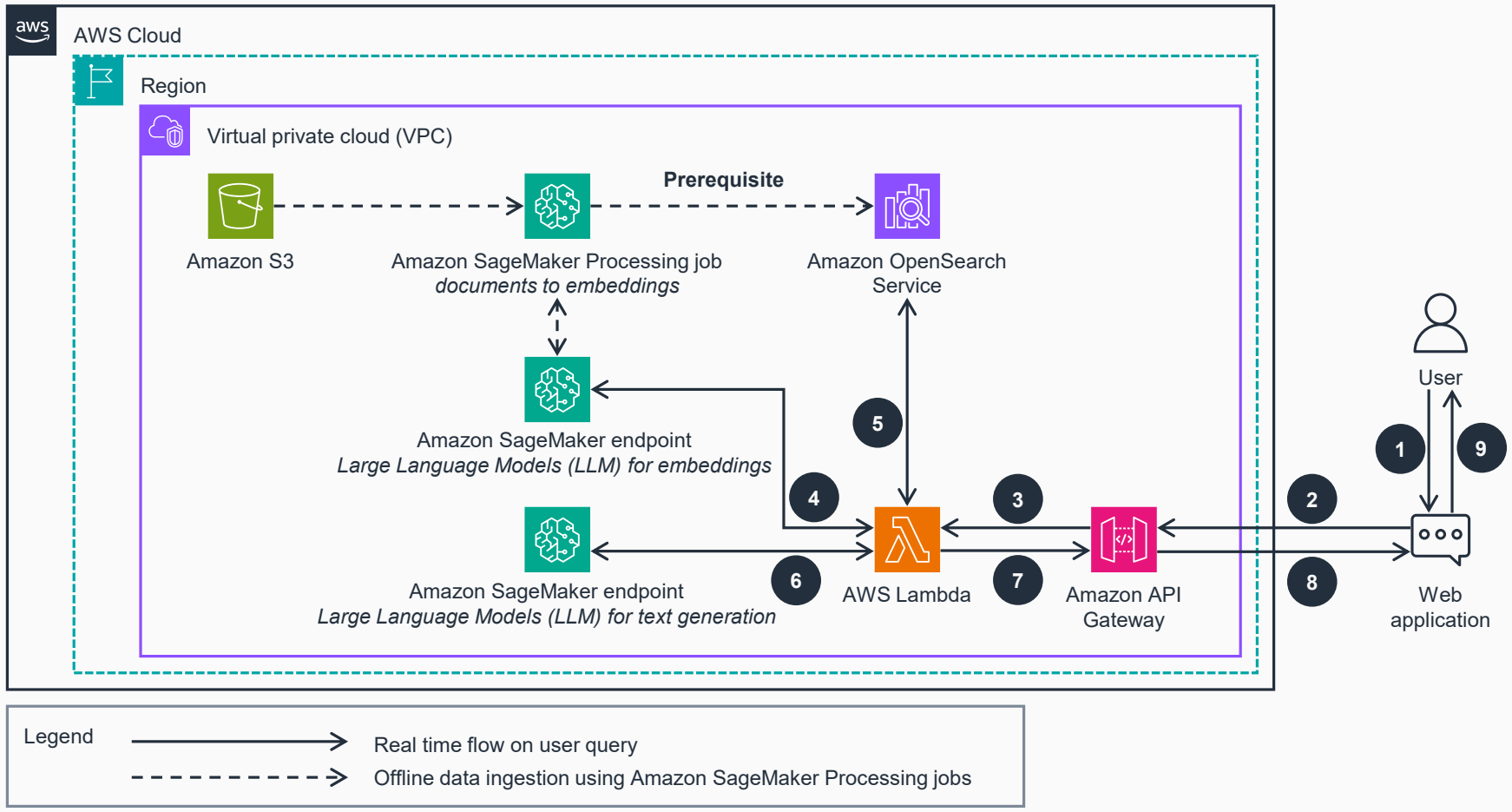
Guidance for Chatbots with Vector Databases on AWS

This architecture diagram shows how you can create a retrieval-augmented generation (RAG) application, such as a question-answering bot, using Amazon OpenSearch Service.



Prerequisite: Amazon SageMaker Processing jobs are used for large scale data ingestion into **Amazon OpenSearch Service**. In this offline data ingestion step, download the dataset locally into the **SageMaker** notebook, then ingest it into the **OpenSearch Service** index. Split the documents into segments, which can then be converted into embeddings to be ingested into **OpenSearch Service**.

1. The user provides a question using a Streamlit web application.
2. The web application invokes the **Amazon API Gateway** endpoint's representational state transfer API.
3. **API Gateway** invokes an **AWS Lambda** function.
4. The function invokes the **SageMaker** endpoint to convert the user's question into embeddings.
5. The function invokes an **OpenSearch Service** API to find documents similar to the user's question.
6. The function creates a prompt, with the user's query and the similar documents as context. It then asks the **SageMaker** endpoint to generate a response.
7. The **Lambda** function provides the response to **API Gateway**.
8. **API Gateway** provides the response to the Streamlit application.
9. The user can view the response on the Streamlit application.

Reviewed for technical accuracy November 2, 2023

AWS Reference Architecture