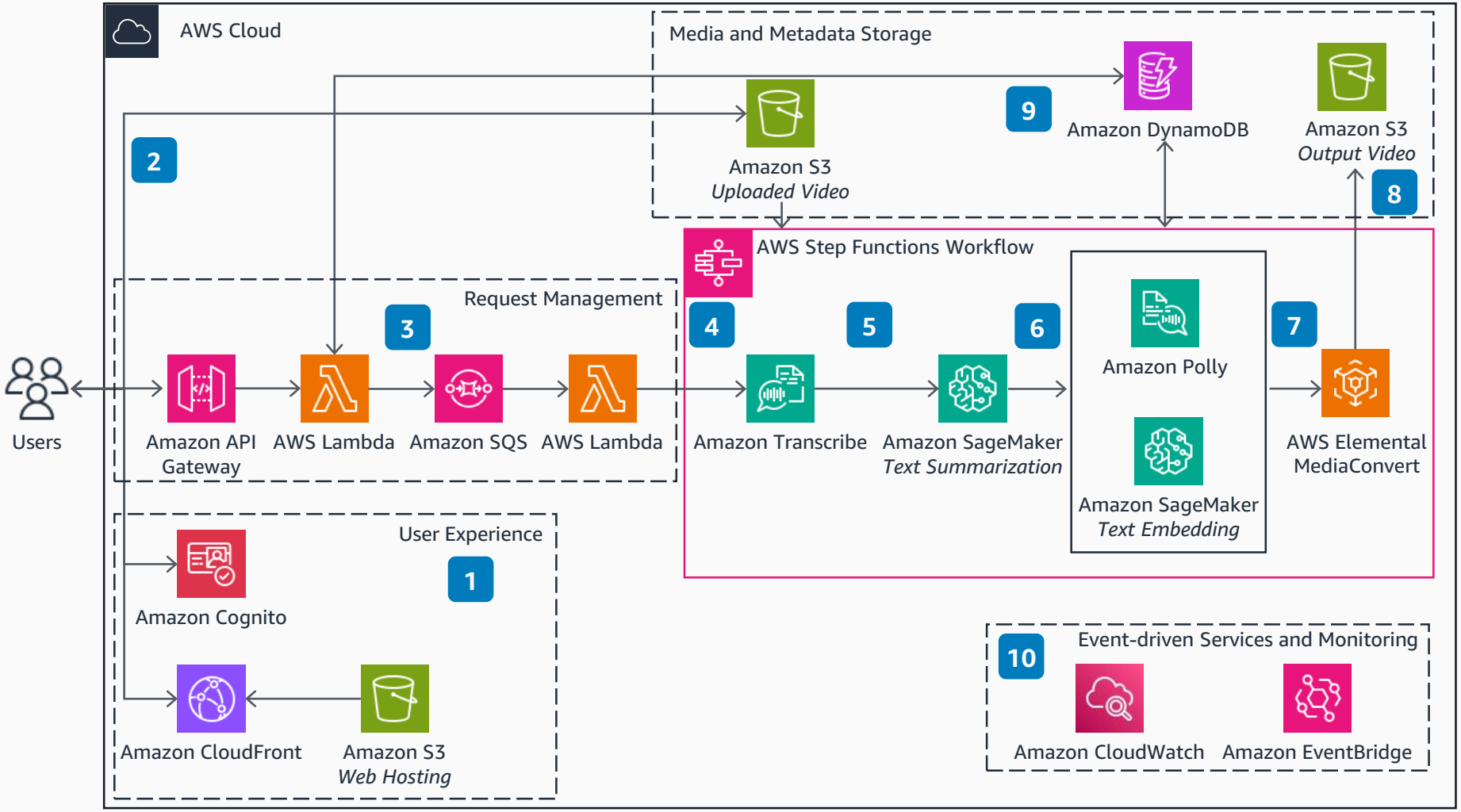
Guidance for Video Summarization using Amazon SageMaker and AI Services

This architecture diagram shows how to produce a short-form summary video from a longer video file by extracting and summarizing speech in the video, identifying the most relevant video segments, and compiling a final video output.



1. **Amazon Simple Storage Service (Amazon S3)** hosts a static website for the video summarization workload, served by an **Amazon CloudFront** distribution. **Amazon Cognito** provides customer identity and sign-in functionality to the web application.

2. **Amazon S3** stores the source videos, which are uploaded through pre-signed URLs.

3. To perform a video summarization, make an API call to **Amazon API Gateway** that invokes an **AWS Lambda** function to put the request into an **Amazon Simple Queue Service (Amazon SQS)** queue. New messages added to the queue are processed by a **Lambda** function that processes a new **AWS Step Functions** workflow.

4. **Amazon Transcribe** converts the speech in the source video into text, generating an output subtitle file containing the transcript and timestamps.

5. **Amazon SageMaker** foundation model endpoint summarizes the text, retaining the story from the original video but in shorter form.

6. **Amazon Polly** generates a voice narration. The **SageMaker** text embedding model endpoint pairs each sentence in the summarized text with its corresponding sentences in the original subtitle file. The output is the most relevant video segments and their timestamps.

7. **AWS Elemental MediaConvert** transcodes the final video output using the original video input clipping timestamps. It inserts the voice narration, generated from **Amazon Polly,** with optional background music of your preference.

8. **Amazon S3** stores the output video that offers durable, highly available, and scalable data storage at low cost.

9. **Amazon DynamoDB** tables store profiling and task metadata. This helps you keep track of the tasks' status and other relevant information.

10. **Amazon CloudWatch** and **Amazon EventBridge** monitor in near real-time every component, and can be used to integrate this workflow into other systems.

Reviewed for technical accuracy September 5, 2023

AWS Reference Architecture