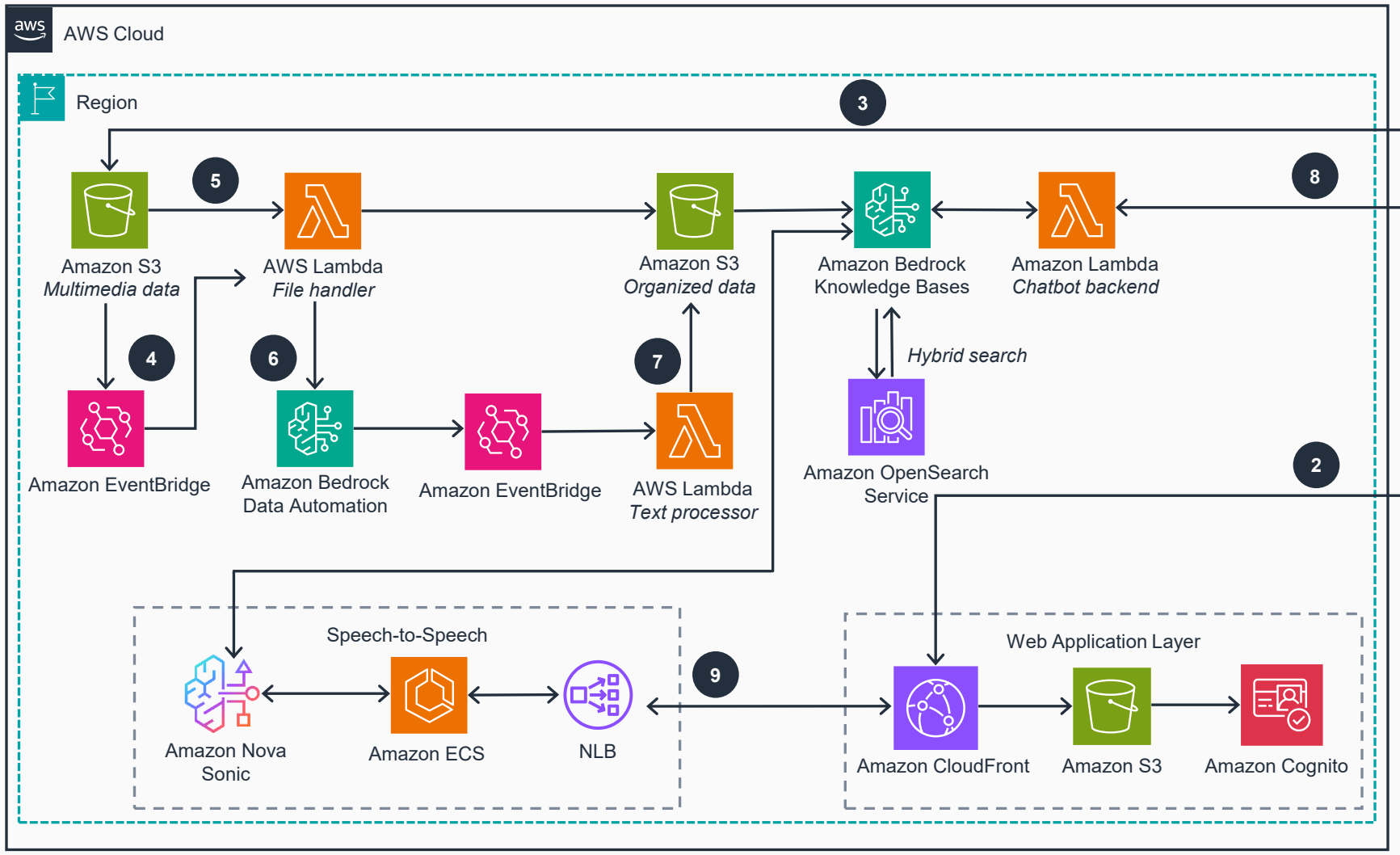
Guidance for Advanced Multimodal Chatbot with Speech-to-Speech on AWS

This architecture diagram shows how to leverage AWS AI services to process multi-format data (video, audio, PDFs, text), creating a unified knowledge base for quick insights. This slide shows Steps 1-7.



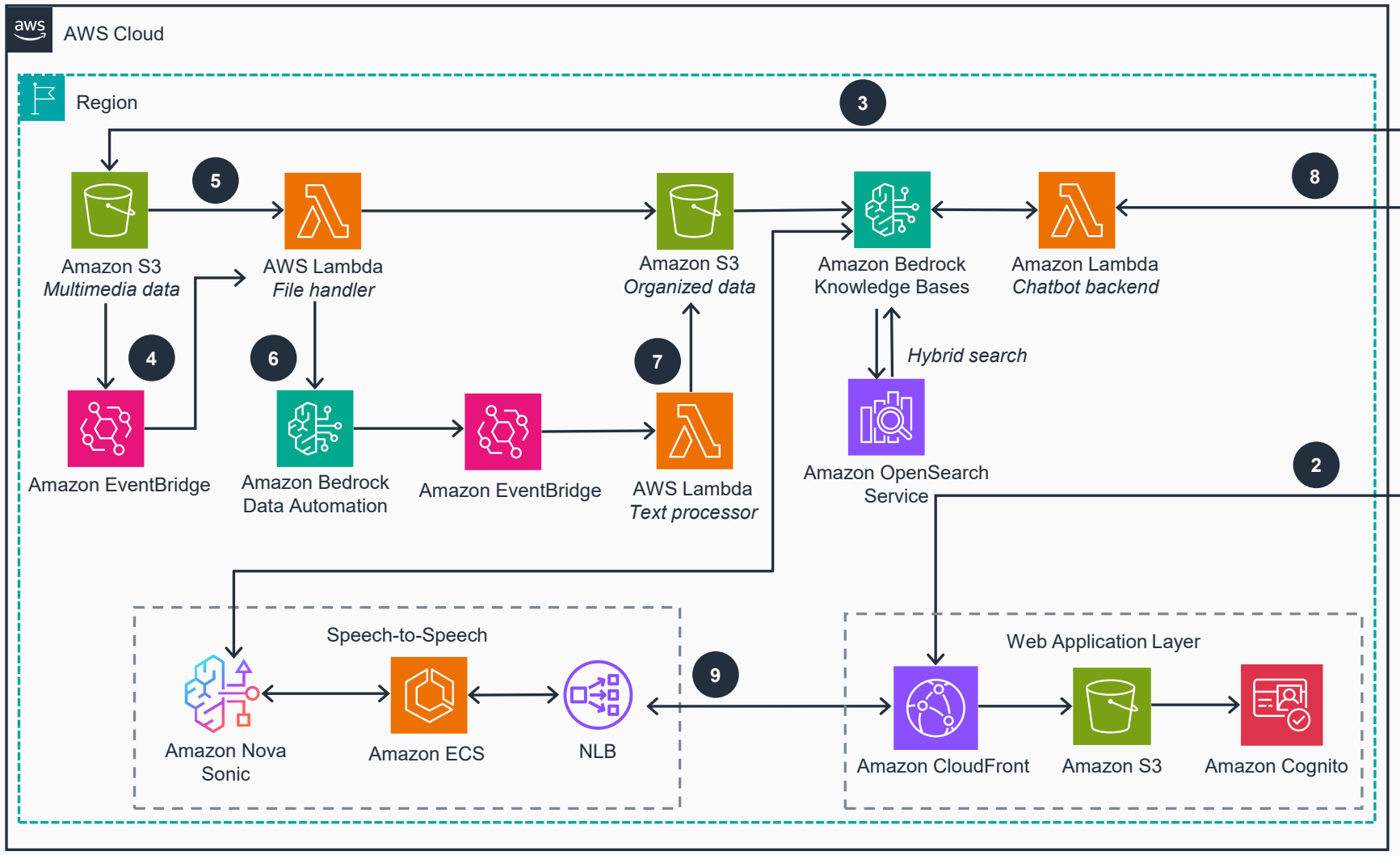
1. Users access the web application UI to authenticate, upload multimedia files, and ask questions about their data.
2. **Amazon CloudFront** delivers the web application UI, which is hosted in an **Amazon Simple Storage Service (Amazon S3)** bucket and authenticates users through **Amazon Cognito**.
3. Authenticated users upload multimedia files through the web application UI to the multimedia data **S3** bucket.
4. File uploads to the multimedia data **S3** bucket trigger an **Amazon EventBridge** rule. **EventBridge** invokes the file handler **AWS Lambda** function.
5. The **Lambda** function invokes an **Amazon Bedrock Data Automation** project with the file.
6. The **Amazon Bedrock Data Automation** project processes the file to extract information in structured JSON format. Text, image descriptions, and table data are extracted from rich documents. Image summary and categories are extracted from images. Audio segments, summary, and timestamps are extracted for both video and audio files. In addition, video frames and video summary are also extracted for video files.
7. Upon successful **Amazon Bedrock Data Automation** job completion, the **Lambda** function (text processor) is invoked by **EventBridge** and formats the JSON output, tests it, and uploads it to the documents folder in the organized data **S3** bucket.

Reviewed for technical accuracy July 14, 2025

AWS Reference Architecture

Guidance for Advanced Multimodal Chatbot with Speech-to-Speech on AWS

This architecture diagram shows how to leverage AWS AI services to process multi-format data (video, audio, PDFs, text), creating a unified knowledge base for quick insights. This slide shows Steps 8-9.



8 Users trigger the **Amazon Bedrock Knowledge Bases** retriever from the web application UI with questions about their data. The web application UI invokes the **Lambda** function (chatbot backend), which calls **Amazon Bedrock** to search the retrieved content and generate a response. The generated response is then returned to the web application UI.

9 When users initiate speech-to-speech communication, the web application UI forwards the process through **Amazon CloudFront** to the speech-to-speech processing pipeline. This pipeline consists of **Amazon Elastic Container Service (Amazon ECS)** with a running task to establish and maintain connection with **Amazon Nova Sonic**, and a network load balancer (NLB) to distribute traffic. **Amazon Nova Sonic** handles speech recognition, natural language processing, and speech synthesis workloads, enabling users to ask questions through voice and receive answers through audio, with relevant context from **Amazon Bedrock Knowledge Bases**.

Reviewed for technical accuracy July 14, 2025

AWS Reference Architecture