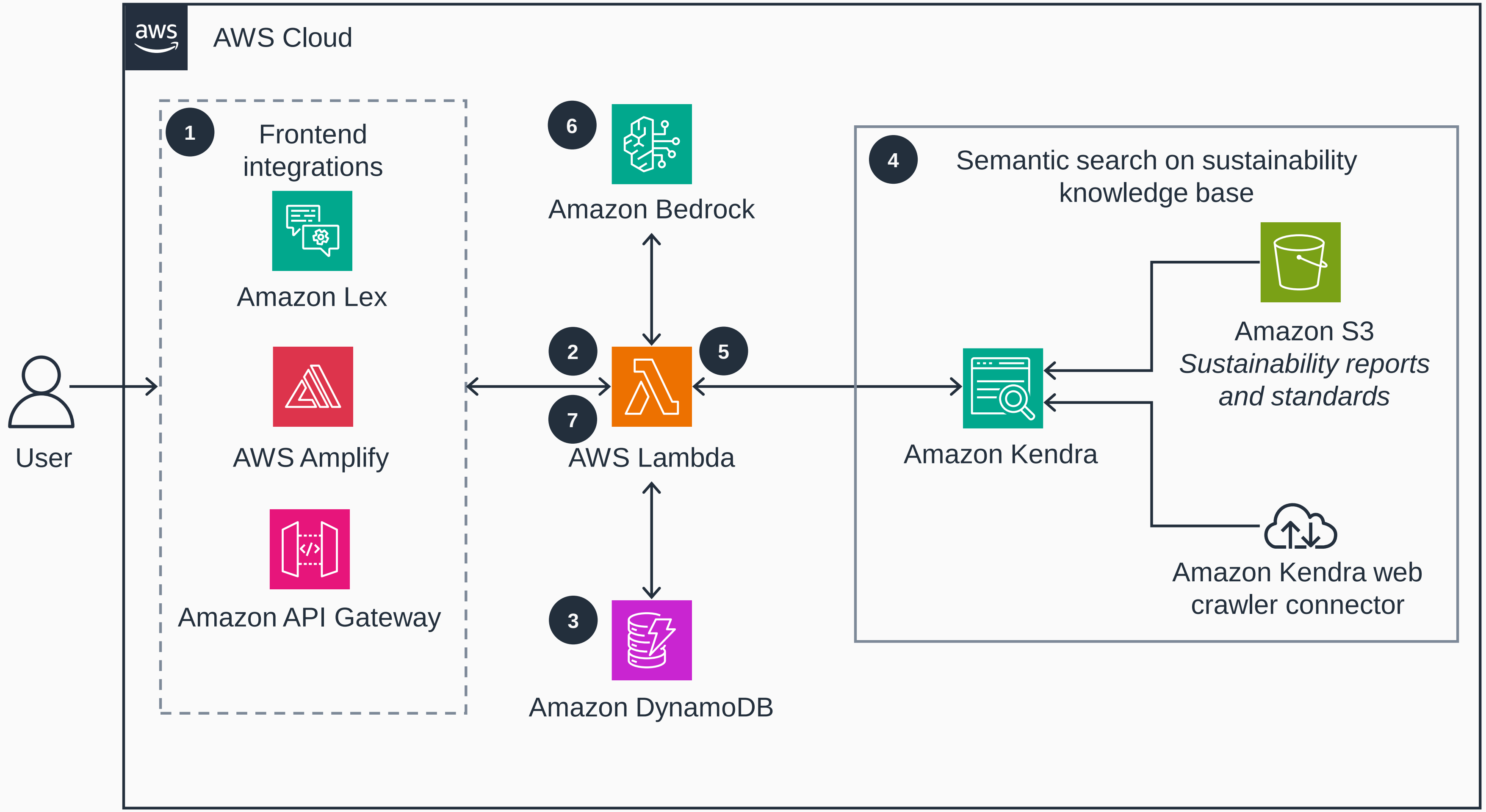
Guidance for Querying Sustainability Documents Using Generative AI for ESG Reporting on AWS

This architecture diagram demonstrates how to implement a Retrieval-Augmented Generation (RAG) process into your sustainability workflow by combining the capabilities of Amazon Kendra with a large language model (LLM) on Amazon Bedrock.



1 A user asks questions and receives generated responses through various frontend integration options. For example, **Amazon Lex** for conversational chatbots, **AWS Amplify** for custom frontend web applications, and **Amazon API Gateway** for processing user requests with backend services.

2 **AWS Lambda** acts as a backend response orchestrator.

3 **Lambda** stores all inputted questions and generated responses into **Amazon DynamoDB** as conversational memory to facilitate future user requests.

4 **Amazon Kendra** performs semantic searches on your sustainability knowledge base. This consists of objects related to sustainability frameworks, such as the Corporate Sustainability Reporting Directive (CSRD) and the International Sustainability Standards Board (ISSB). It also consists of corporate reports, such as the Carbon Disclosure Project (CDP) questionnaires and the Form 10-K.

The knowledge base can be stored on **Amazon Simple Storage Service (Amazon S3)** or third-party repositories like Dropbox and Confluence. It can also be accessed with public or internal websites over HTTPS using an **Amazon Kendra** Web Crawler.

5 The **Lambda** function uses the **Amazon Kendra** Retrieve API, which is optimized for Retrieval-Augmented Generation (RAG), to identify and extract relevant passages. Document metadata filters are specified in the query API to help narrow the results, including only documents relevant to the user's question.

6 The user's question and the relevant context are passed by the **Lambda** function to a large language model (LLM) hosted on **Amazon Bedrock**, a fully managed service that offers a choice of high-performing foundation models (FMs). LLMs such as Anthropic's Claude or Meta Llama can compare, analyze, and summarize large volumes of text from the sustainability knowledge base.

7 The generated response from the LLM on **Amazon Bedrock** is returned to the **Lambda** function, which updates the conversational memory in **DynamoDB** and presents the response back to the user through your implementation of frontend integrations.

Reviewed for technical accuracy May 10, 2024

AWS Reference Architecture