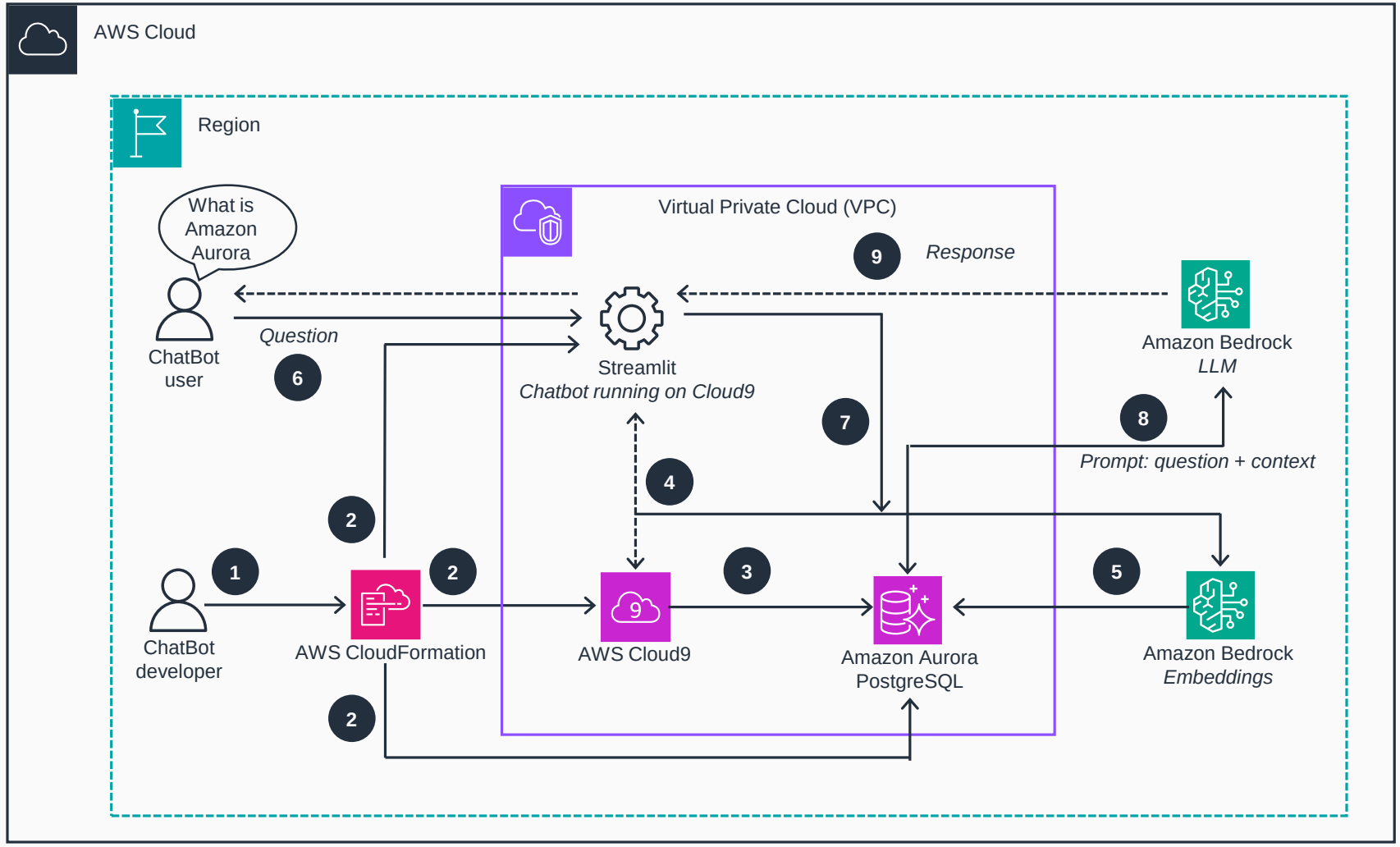
Guidance for High-Speed RAG Chatbots on AWS

This architecture diagram shows how to build an artificial intelligence (AI)-powered chatbot that lets you ask questions based on content in your PDF files in natural language. Once you upload PDF files to the application, type in questions in simple English. The AI-powered application will process the questions and use the Retrieval-Augmented Generation (RAG) technique to generate a response based on the relevant content from the PDFs.



1. Download the **AWS CloudFormation** template from the GitHub repository and deploy the **CloudFormation** stack.
2. The **CloudFormation** stack deploys an **AWS Cloud9** instance, an **Amazon Aurora PostgreSQL** cluster, a Streamlit custom chatbot application, and other pre-requisites required for this Guidance.
3. Set up the environment variables to connect to the **Aurora PostgreSQL** instance, create the pgvector extension, and start the Streamlit application.
4. Once the Streamlit application starts, upload the PDF document for processing. This will segment the document into chunks and convert them into vectors using an **Amazon Titan** model from **Amazon Bedrock. Amazon Bedrock** is a fully managed service that offers a choice of high-performing foundation models (FMs).
5. Load the vector embeddings into an **Aurora PostgreSQL** cluster.
6. The user asks a question in natural language in the chatbot application.
7. The question from the Streamlit application is converted into embeddings using the **Amazon Titan** model. The vectors are then compared with the **Aurora PostgreSQL** vector store to identify the most semantically similar vectors.
8. Pass the user question and the context from the vector database to the large language model (LLM). In this example, the Claude 3 model from Anthropic on **Amazon Bedrock** is used.
9. The LLM generates a response based on the relevant content and displays the response in the chatbot application.

Reviewed for technical accuracy May 30, 2024

AWS Reference Architecture