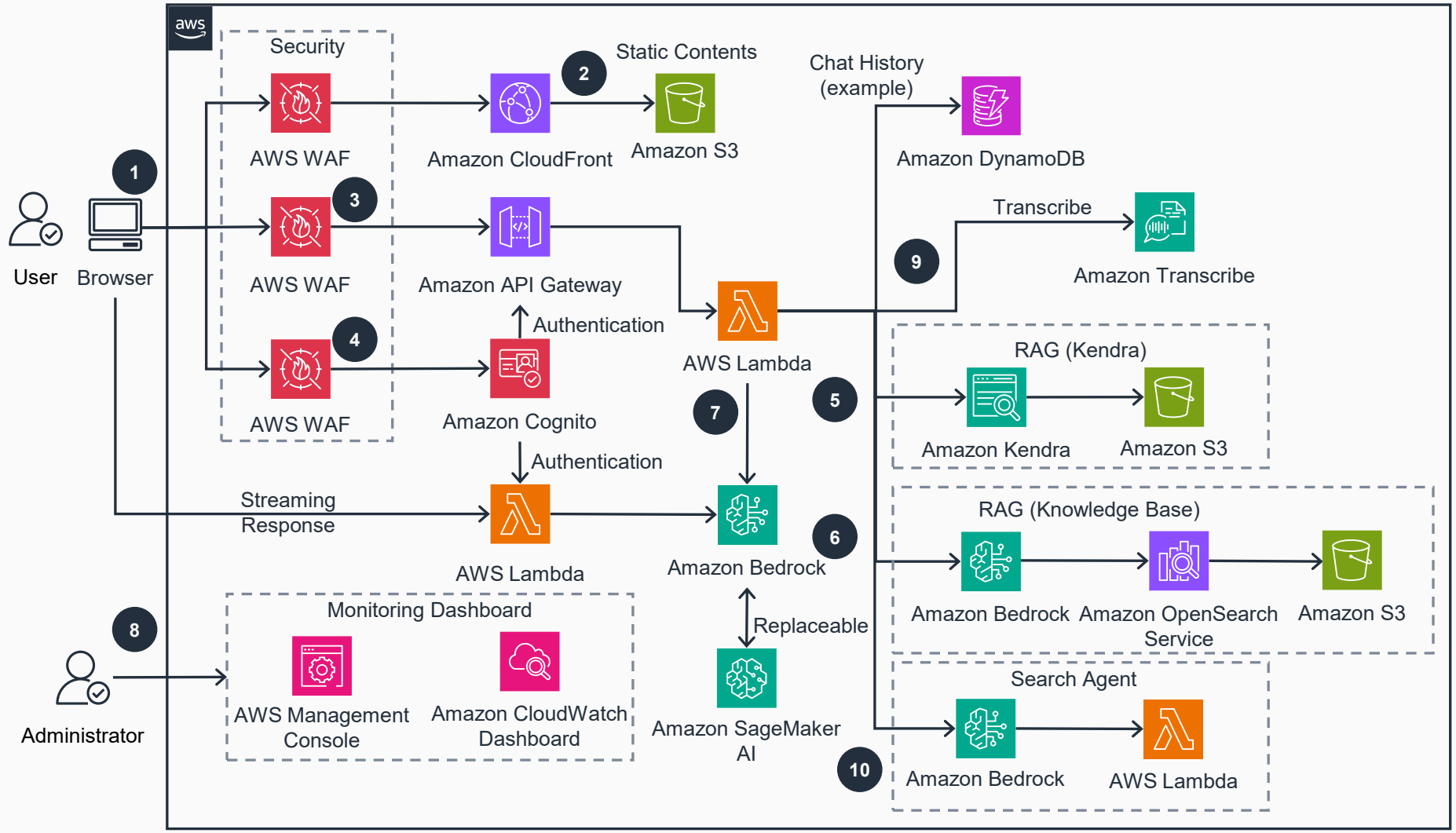
Guidance for Rapidly Deploying Generative AI for Business Use Cases on AWS

This architecture diagram illustrates how to effectively support GenU on AWS. It shows the key components and their interactions, providing an overview of the architecture's structure and functionality.



1. Users access the application through a browser, where **AWS Web Application Firewall (AWS WAF)** provides security at multiple entry points.
2. Serve static content through **Amazon CloudFront** and store in **Amazon Simple Storage Service (Amazon S3)**, while maintaining chat history in **Amazon DynamoDB.**
3. **Amazon API Gateway** handles API requests and integrates with **AWS Lambda** functions for backend processing.
4. Manage user authentication through **Amazon Cognito**, providing secure access control.
5. [Optional] The RAG (Retrieval Augmented Generation) system can use **Amazon Kendra** for intelligent search, with documents stored in **Amazon S3**.
6. [Optional] **Amazon Bedrock** can provide fully managed RAG functionality to power the knowledge base component, working with **Amazon OpenSearch Service** for efficient data retrieval.
7. **Amazon SageMaker AI** and **Amazon Bedrock** process queries, with results handled by **AWS Lambda** functions.
8. [Optional] Administrators can monitor the system through the **AWS Management Console** and **Amazon CloudWatch** Dashboard.
9. **Amazon Transcribe** processes audio content when required.
10. [Optional] Enable the Search Agent using **Amazon Bedrock** and **AWS Lambda**, providing enhanced search capabilities when specified in the deployment parameters.

Reviewed for technical accuracy December 17, 2025

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Reference Architecture