Accelerate AWS Well-Architected reviews with Generative AI

Shoeb Bustani

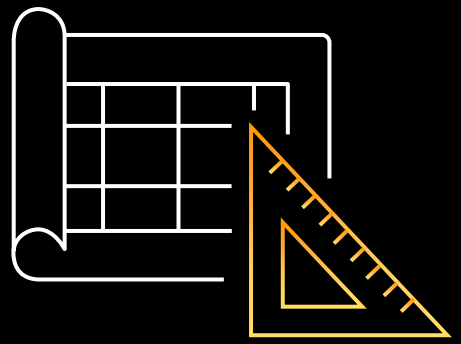
Senior Solutions Architect
AWS

When you look at the workloads your team is building, can you answer the question:

"Are you Well-Architected?"

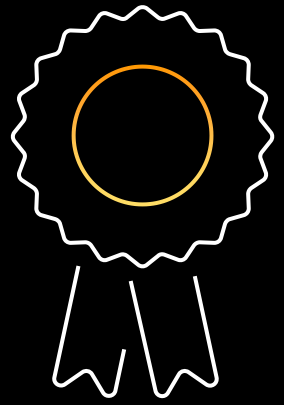
What is the AWS Well-Architected Framework?

Pillars & Lenses



Design principles

Questions



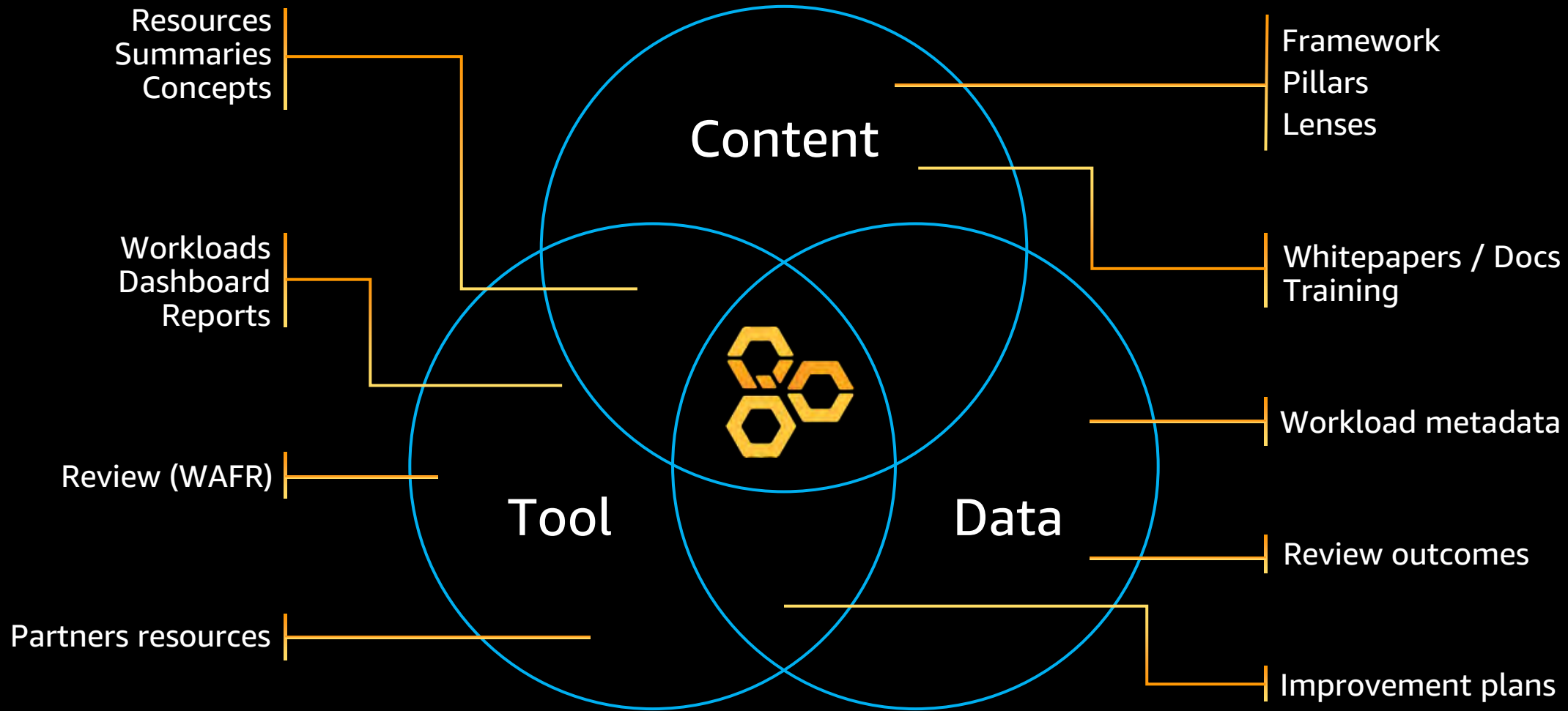
Best Practices

https://aws.amazon.com/architecture/well-architected

Pillars of the AWS Well-Architected Framework

Pillars & Lenses Design principles Questions Best Practices

What is available?



Scaling architecture reviews

Time & resource utilization

Manual reviews are time consuming and resource intensive

Consistency

Inconsistent application of Well-Architected principles across different teams

Playing catch-up

Difficulty in keeping pace with the latest best practices

Review volume and cycles

Challenges in scaling reviews for large or numerous architectures

Scaling architecture reviews

Time & resource utilization

Manual reviews are time consuming and resource intensive

Consistency

Inconsistent application of Well-Architected principles across different teams

Playing catch-up

Difficulty in keeping pace with the latest best practices

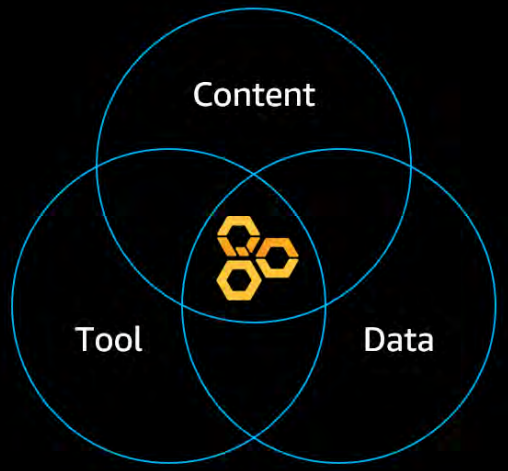
Review volume and cycles

Challenges in scaling reviews for large or numerous architectures

WAFR Acceleration Using Generative AI

Accelerate AWS Well-Architected Framework Review (WAFR) velocity and enterprise adoption by leveraging the power of Generative AI (GenAI) and provide organizations with automated comprehensive analysis and recommendations for optimizing their AWS architectures.

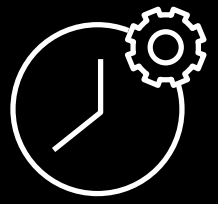
WAFR acceleration approach



Generative AI

AWS Well-Architected Assessment

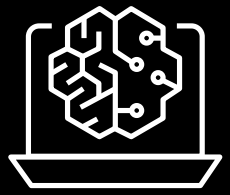
Benefits



Rapid Analysis - Time Efficiency

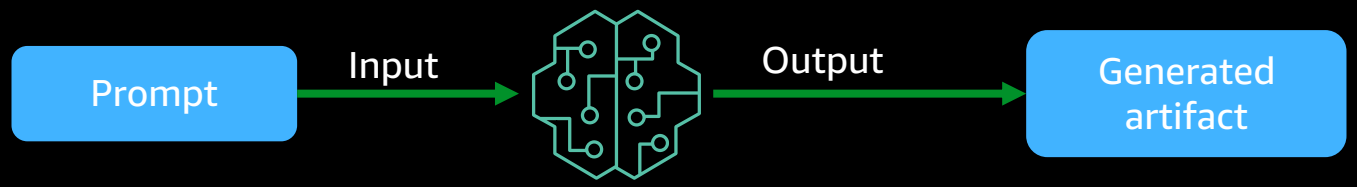
Cost Savings

Consistency



Continuous Improvement

GenAI Prompt engineering



Similar to the decision process of human beings by learning from analogy

Tweak the input or "prompt", get desired output or ”completion”.

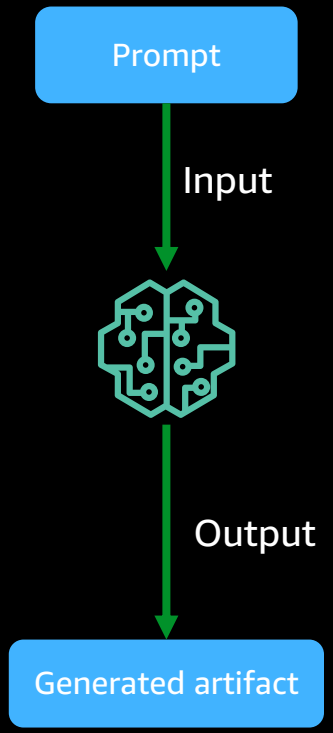
Slight changes to the prompt can have significant impact

Effectiveness also depends on how the model was trained.

GenAI - Prompt engineering types

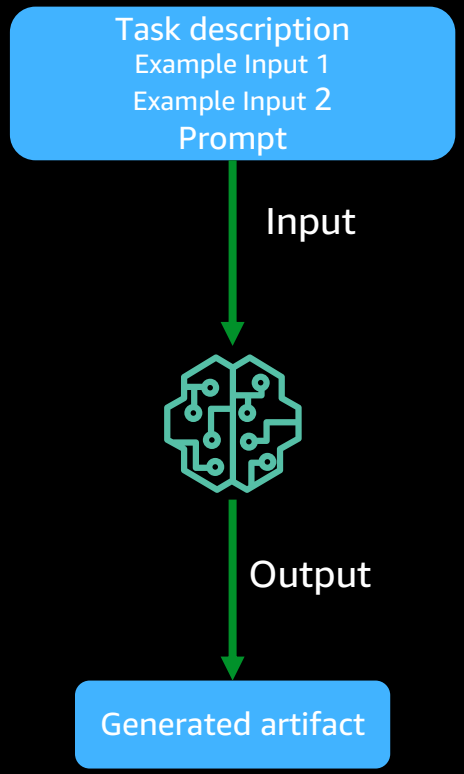
Zero shot prompts

- Direct request with sufficient context



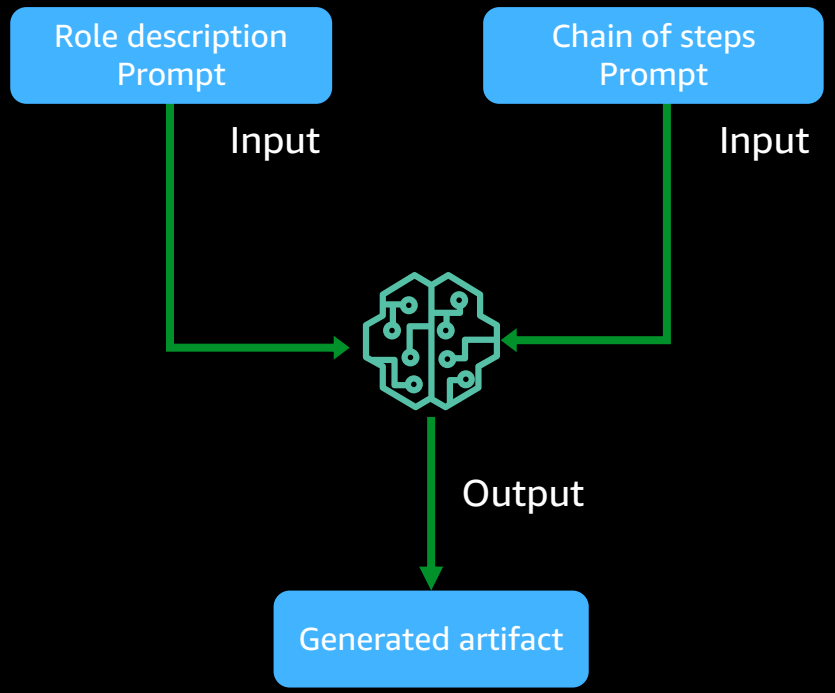
One shot or few shot prompts

- Provide one or more examples with a request



Role or Chain of Thought prompts

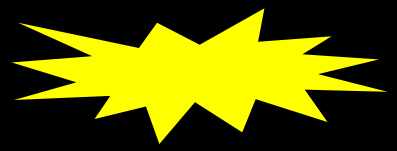
- Provide the model with a role or persona for the task
- Provide a chain of steps for the model to follow



Prompting limitations

Prompting limitations:

- Tweaking the input or "*prompt*“ to get desired output or "*completion”*
- Poor memory
- Limited context
- Accessing external knowledge sources to complete tasks.

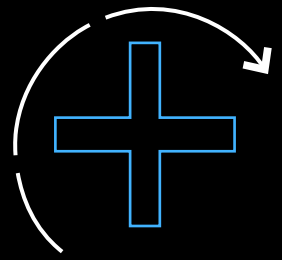


=> Retrieval Augmented Generation (RAG) method

Retrieval Augmented Generation?

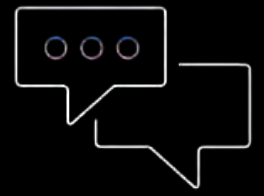
Retrieval

Fetches the relevant content from the external knowledge base or data sources based on a user query



Augmentation

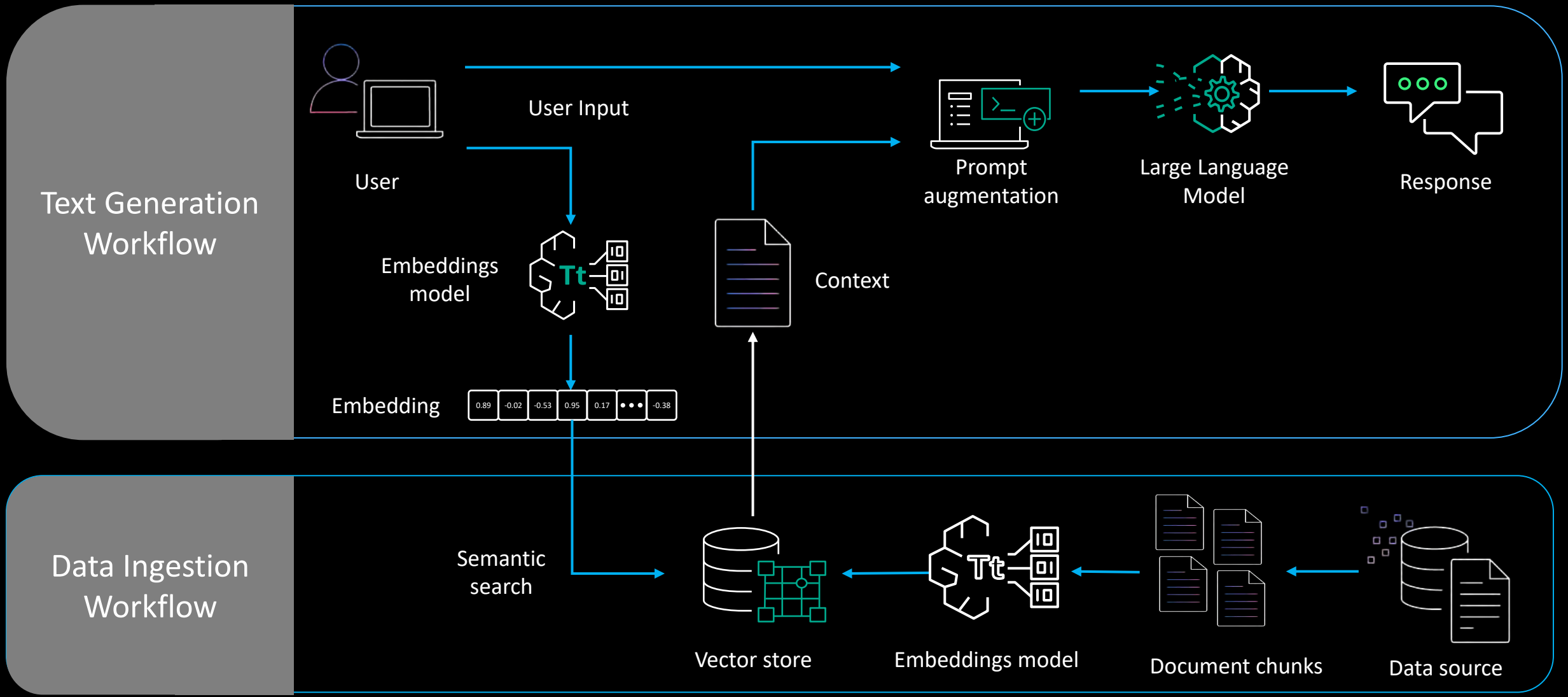
Adding the retrieved relevant context to the user prompt, which goes as an input to the foundation model



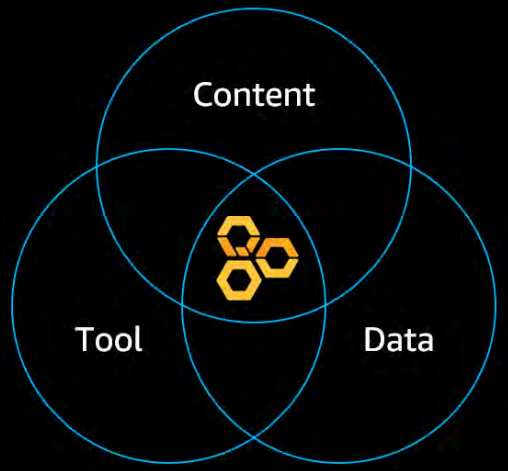
Generation

Response from the foundation model based on the augmented prompt.

RAG flow



WAFR acceleration approach



Amazon Bedrock
+
Knowledge Bases for Amazon Bedrock
+
LLM

AWS Well-Architected Assessment Report

**LLM – Large Language Model*

WAFR Accelerator

WAFR Accelerator is a **one-click** comprehensive sample designed to facilitate and expedite the AWS Well-Architected Framework Review (WAFR) process.

Use workload documentation

Ability to upload workload technical documentation like Solution Architecture Document, Technical Design, Infrastructure Design etc.

Mainstream File format

Ability to upload PDF documents

AWS WAFR knowledge data sets

Drive Amazon Bedrock knowledge bases from formal Well Architected documentation and best practices

Customer knowledge data sets (roadmap)

Organization’s enterprise and solution architecture standards. Design templates and developer standards

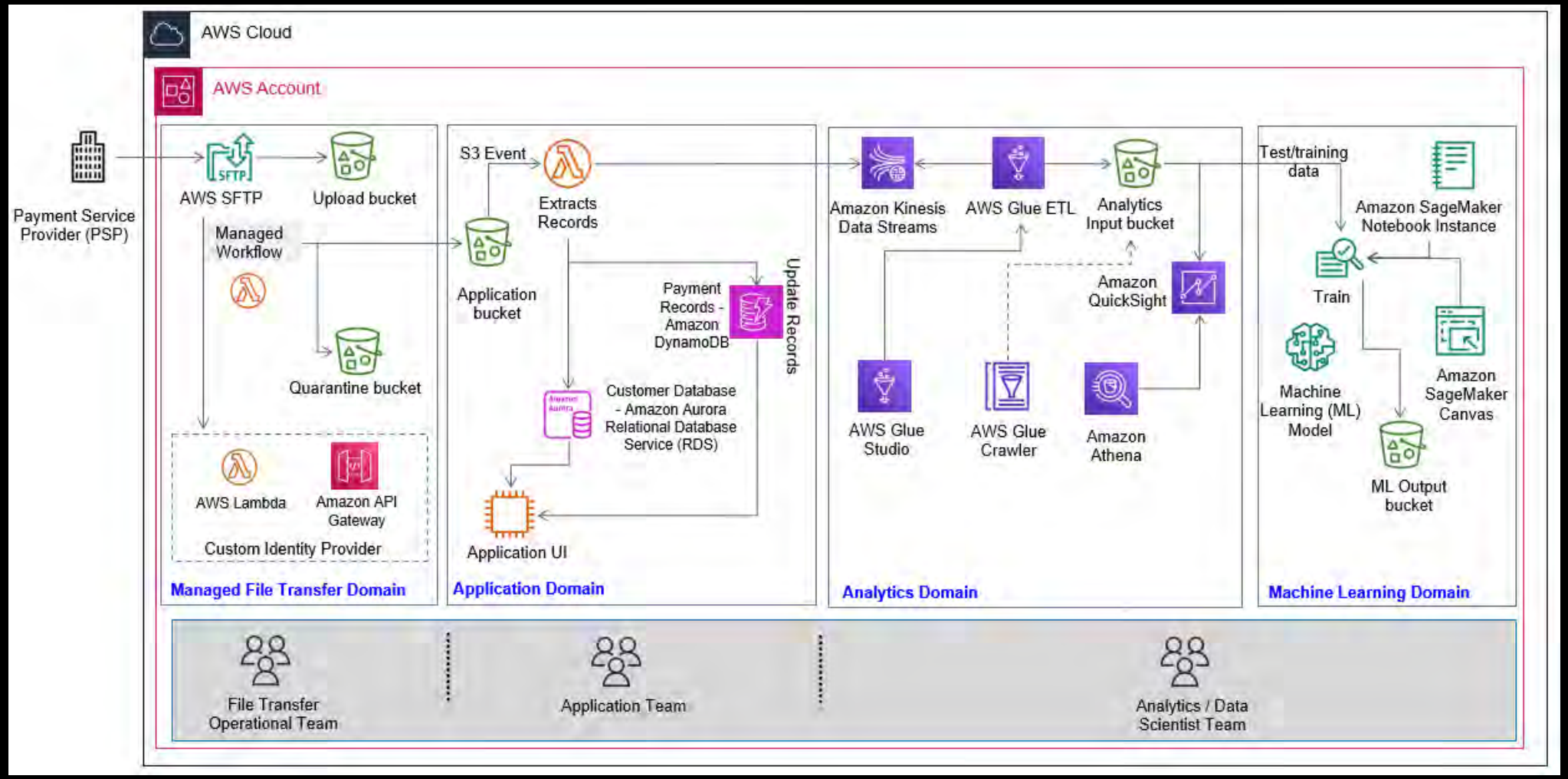
One-click deployment

All resources are created using **AWS Cloud Development Kit (CDK),** including:

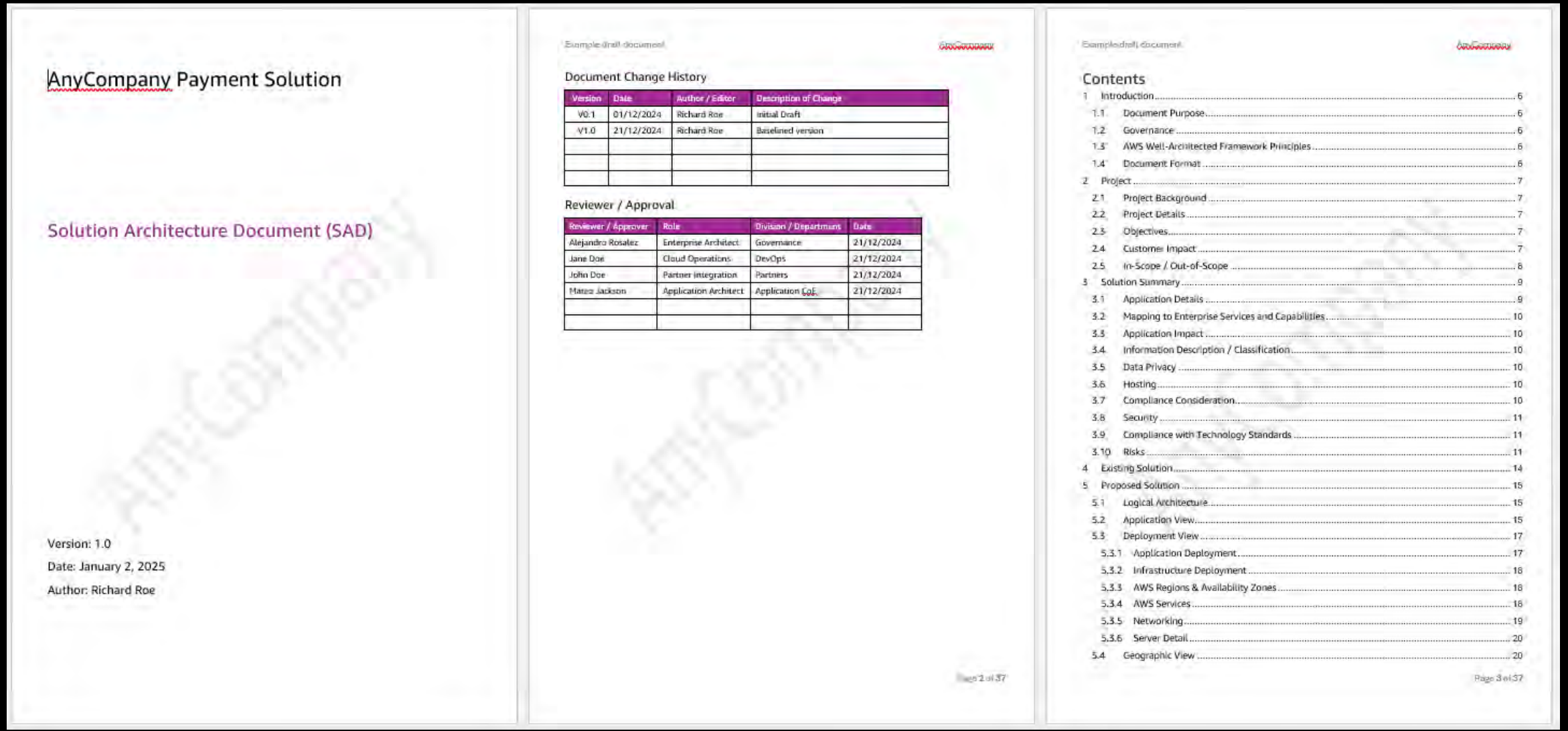
- Amazon Bedrock with Anthropic Claude Large Language Models (LLMs)
- Amazon Opensearch (serverless) used as Amazon Bedrock RAG Knowledge base
- Amazon DynamoDB table which is our core database to store information about review runs
- Amazon Simple Queue Service (SQS) queue
- EC2 instance for hosting Streamlit front end application
- Amazon S3 buckets
- AWS Lambda and Step functions for managing WAFR analysis runs
- Amazon Cognito for user management
- Amazon CloudFront distribution with Application Load Balancer (ALB) and initial AWS Web Application Firewall (WAF) rule

Demo

Demo – AnyCompany Solution Architecture



Demo - Accelerated WAFR Review



AnyCompany Payment Solution

Solution Architecture Document (SAD)

Version: 1.0

Date: January 2, 2025

Author: Richard Roe

Example draft document AnyCompany

Document Change History

Version	Date	Author / Editor	Description of Change
V0.1	01/12/2024	Richard Roe	Initial Draft
V1.0	21/12/2024	Richard Roe	Baselined version

Reviewer / Approval

Reviewer / Approver	Role	Division / Department	Date
Alejandro Rosalez	Enterprise Architect	Governance	21/12/2024
Jane Doe	Cloud Operations	DevOps	21/12/2024
John Doe	Partner Integration	Partners	21/12/2024
Mateo Jackson	Application Architect	Application CoE	21/12/2024

Page 2 of 37

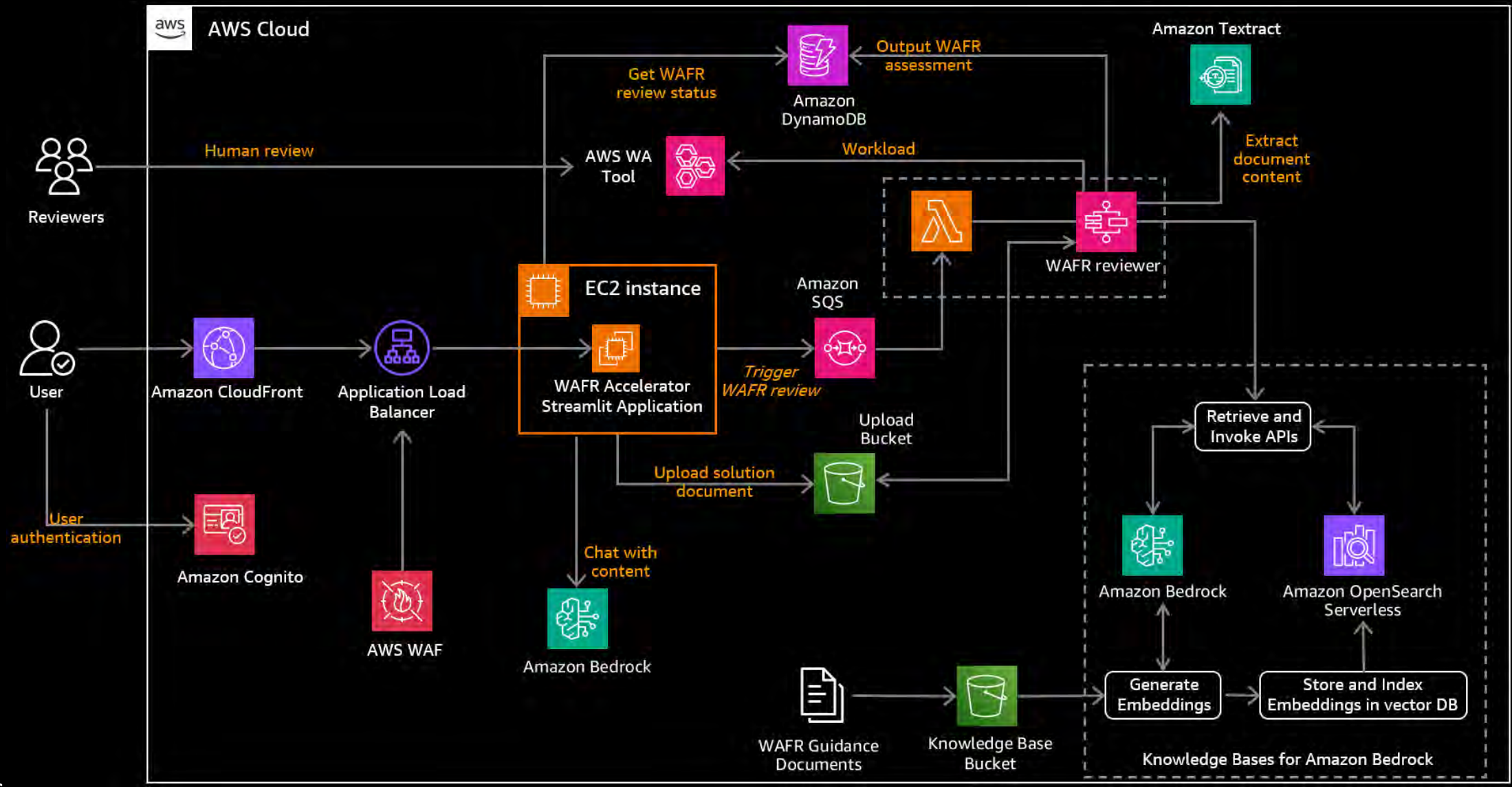
Example draft document AnyCompany

Contents

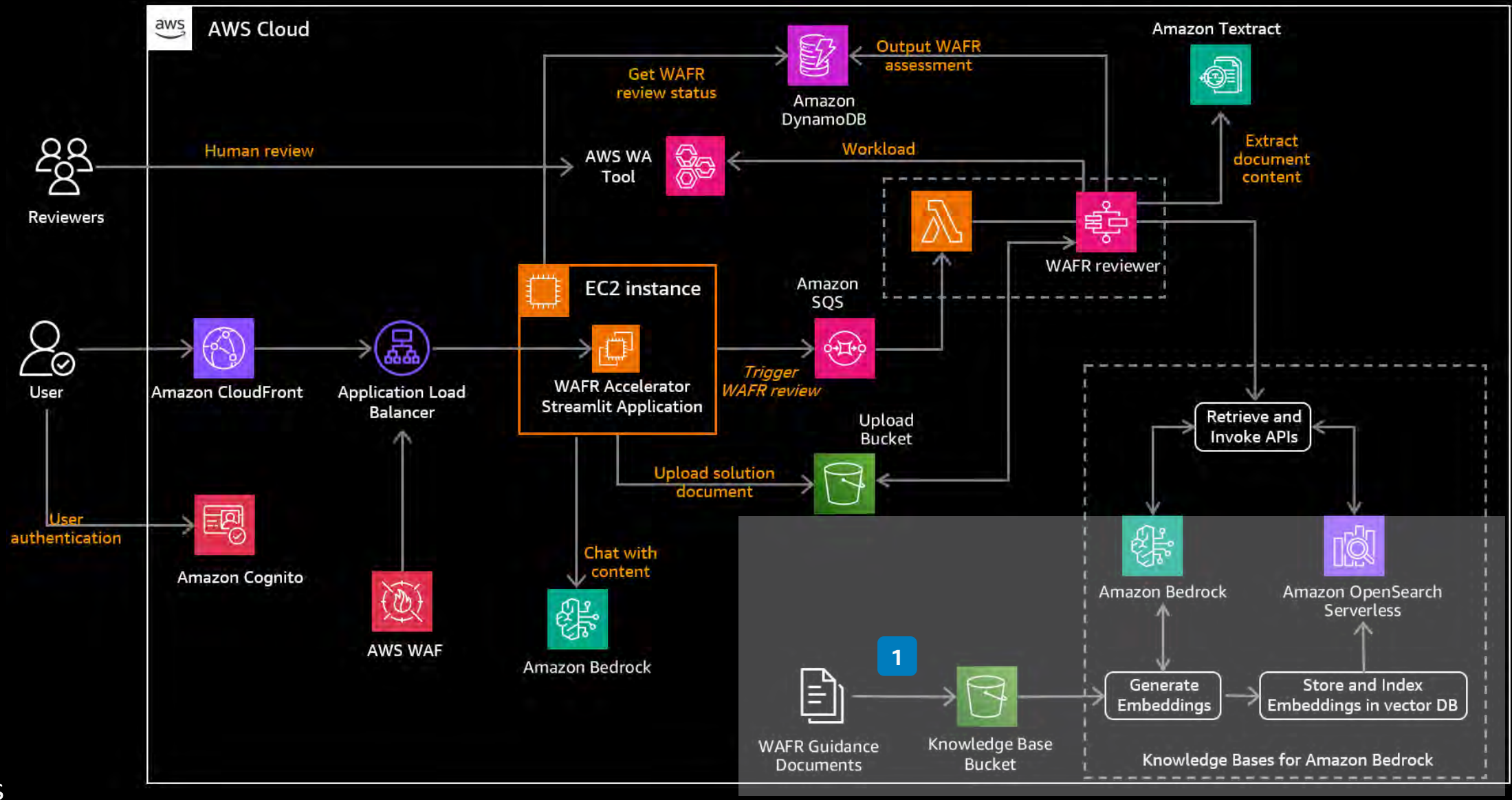
Page 3 of 37

WAFR Accelerator Architecture

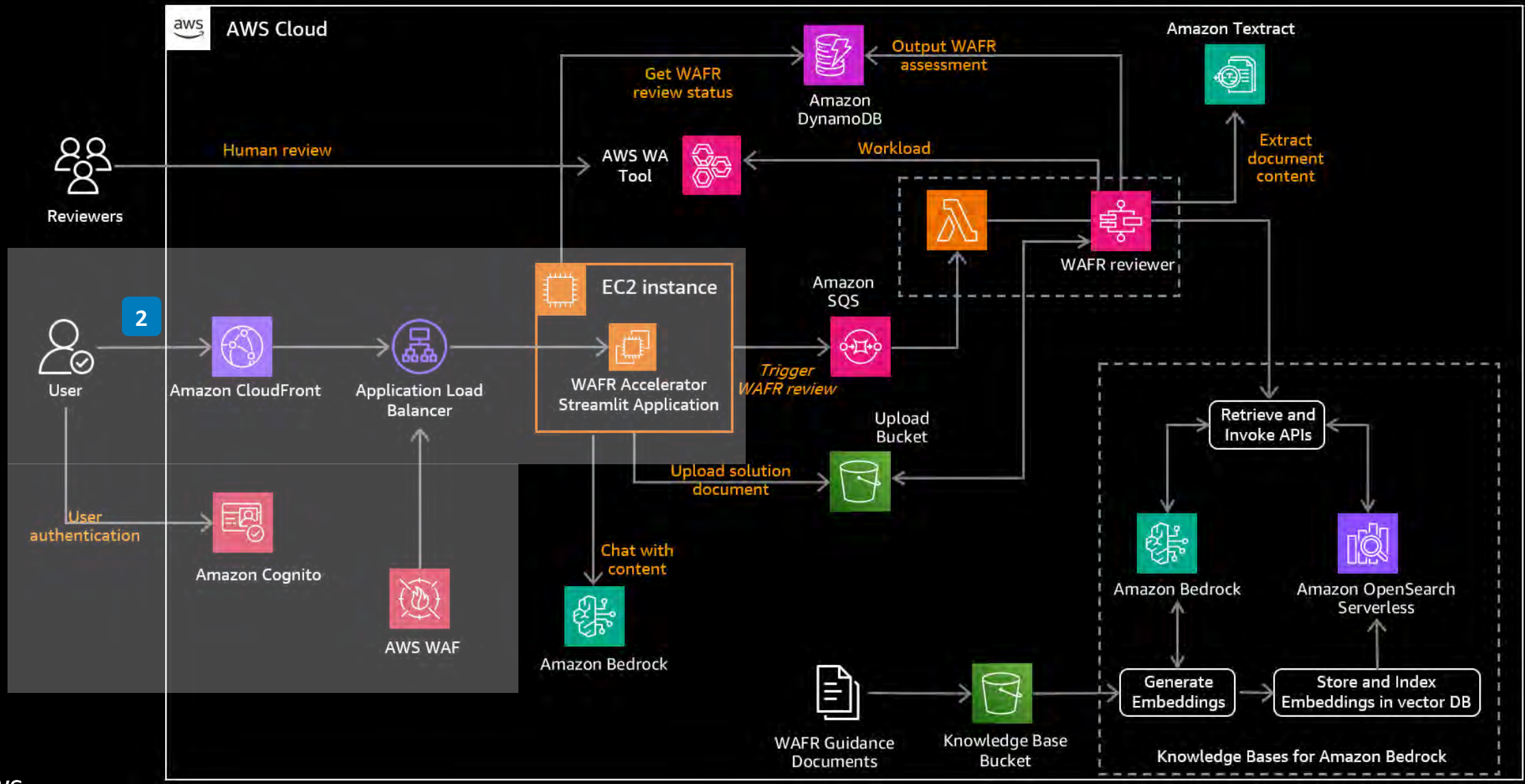
WAFR Accelerator - Architecture



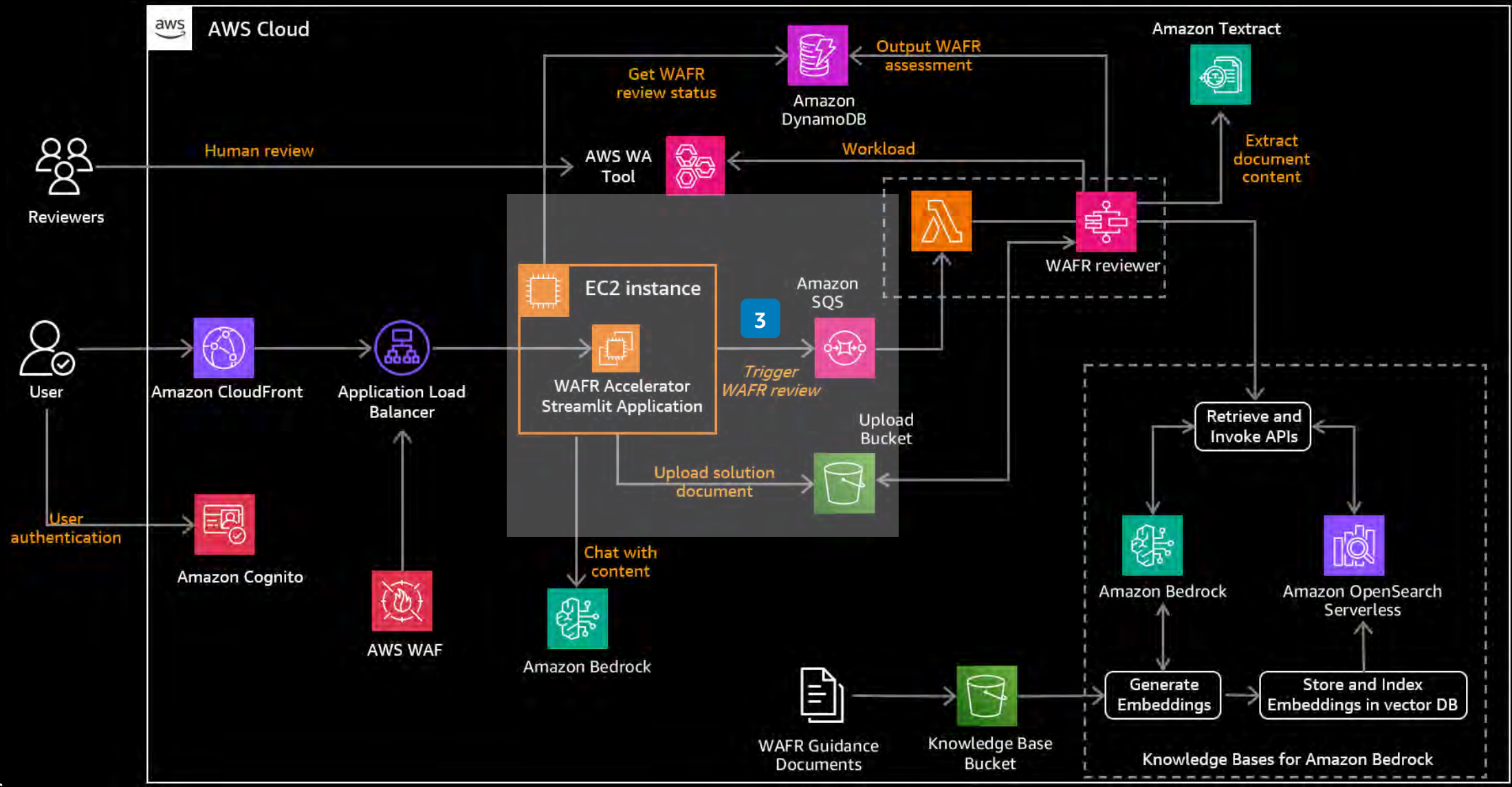
WAFR Accelerator - Architecture



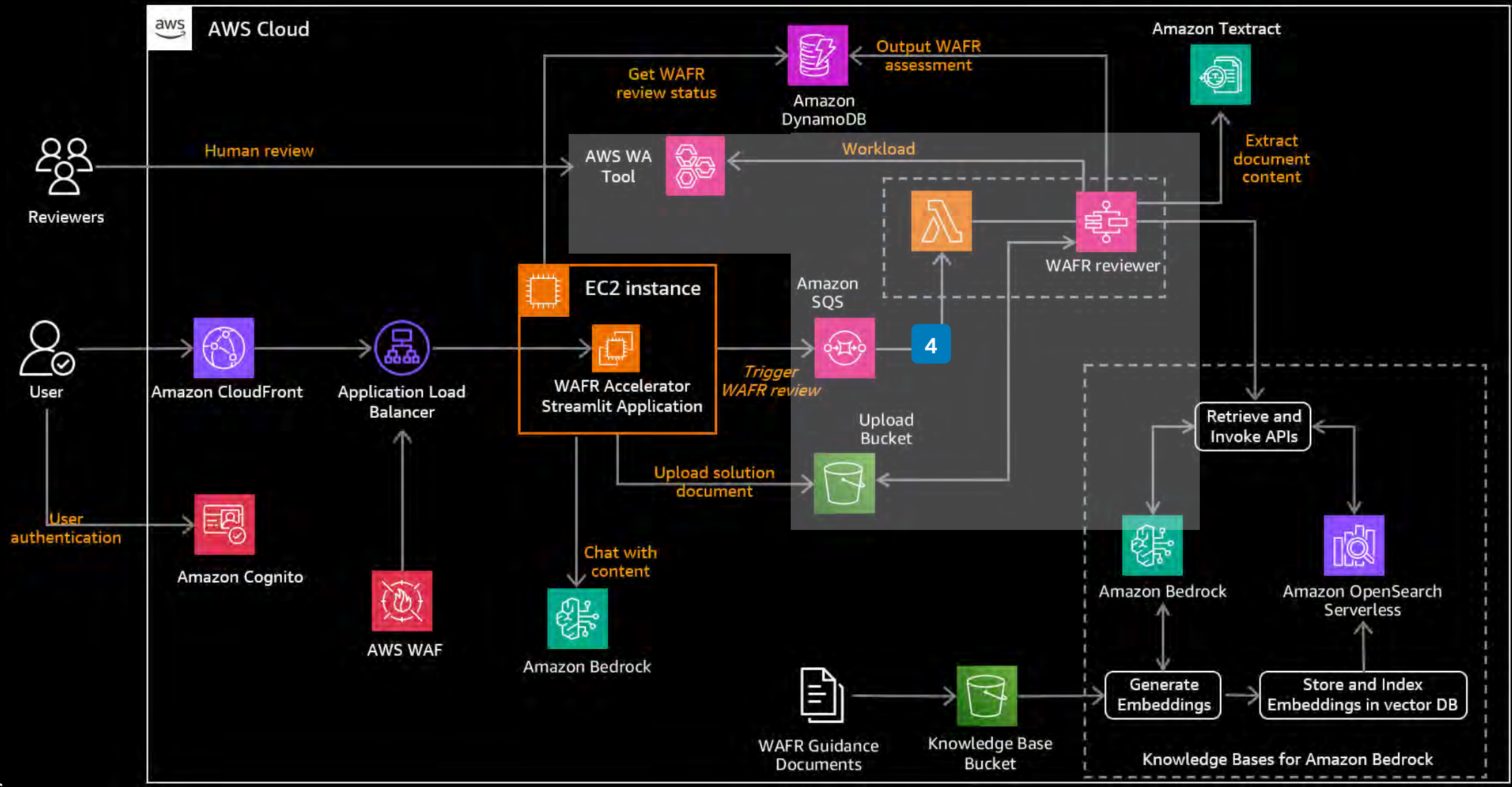
WAFR Accelerator - Architecture



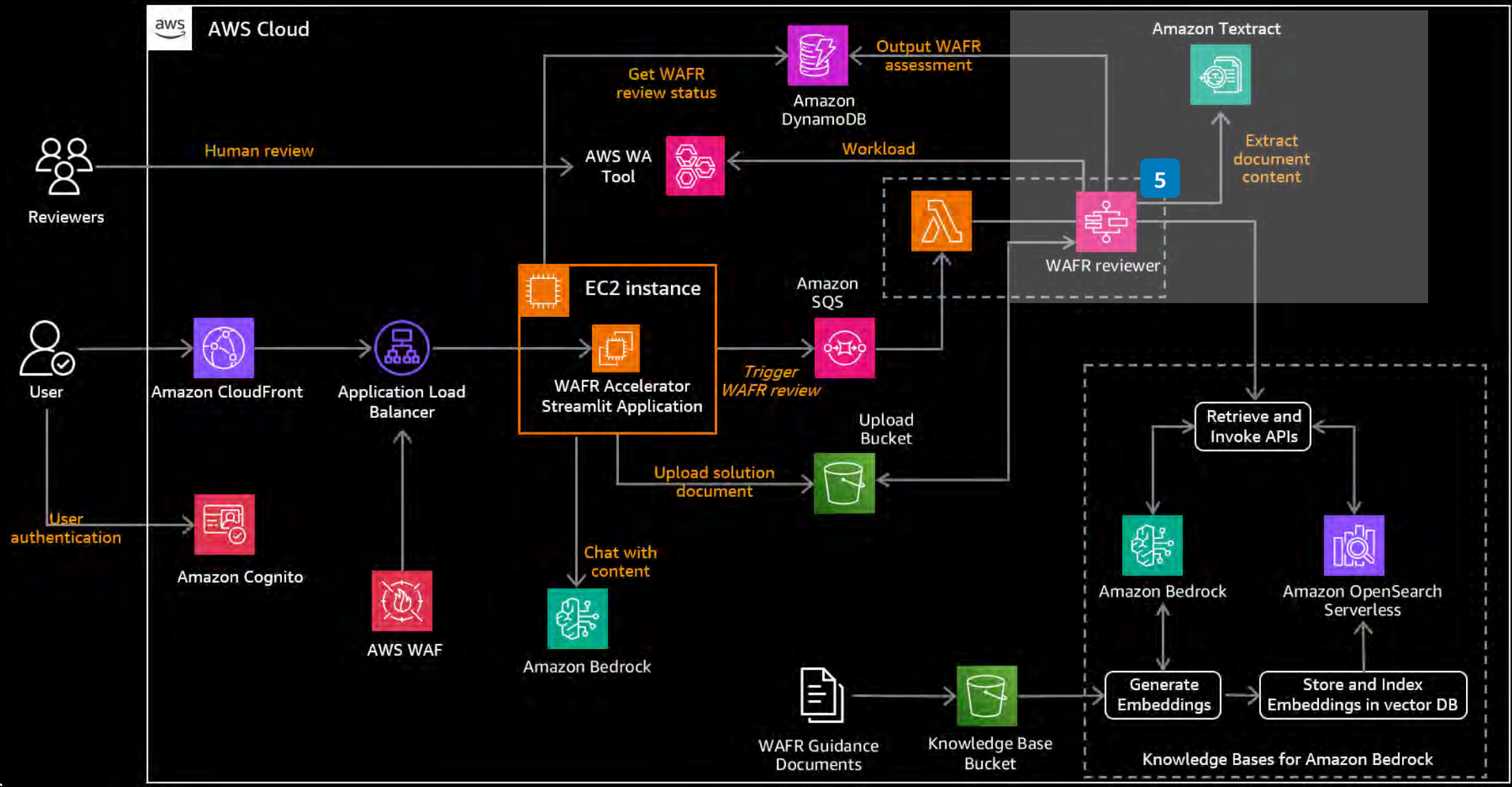
WAFR Accelerator - Architecture



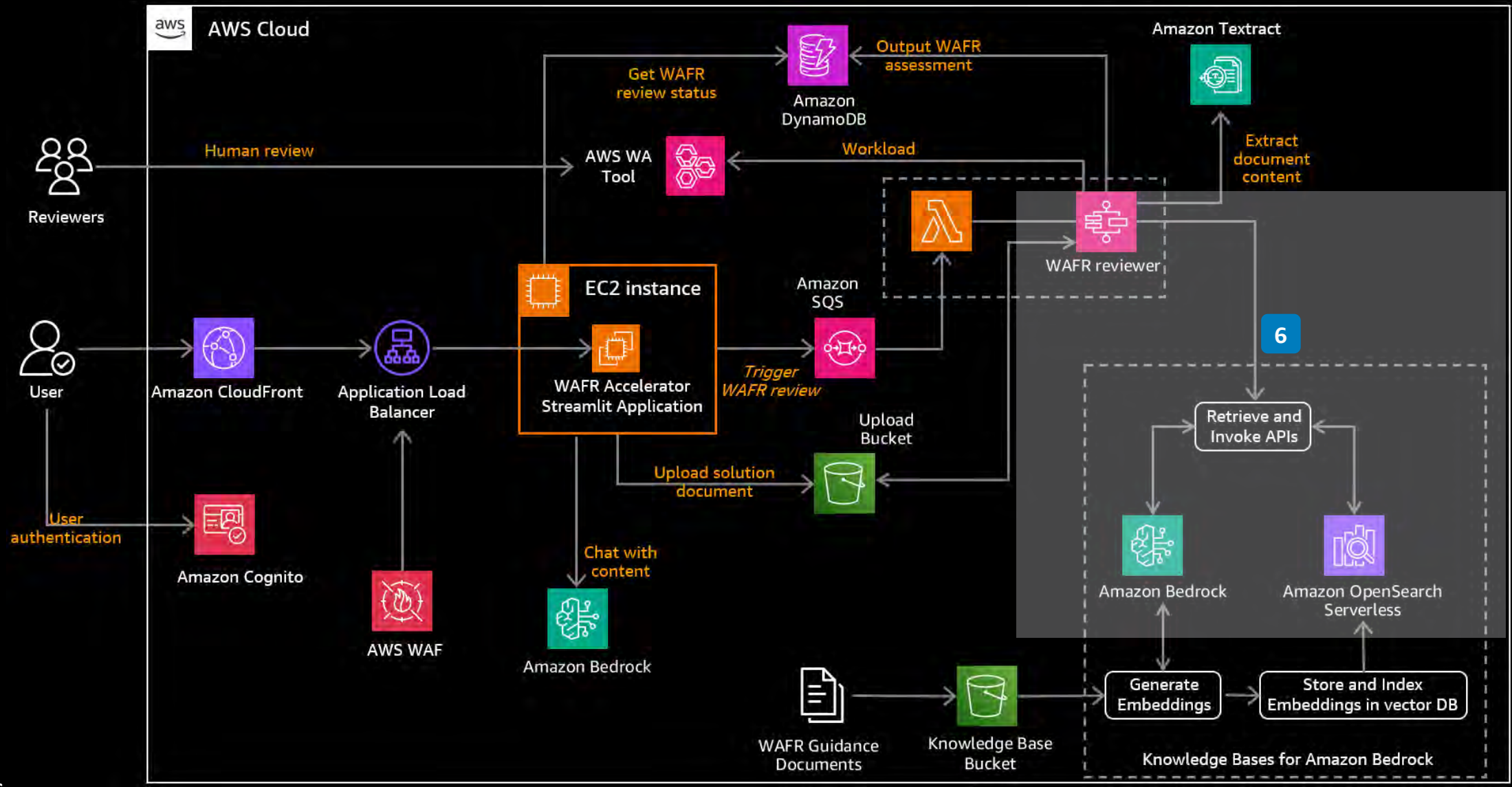
WAFR Accelerator - Architecture



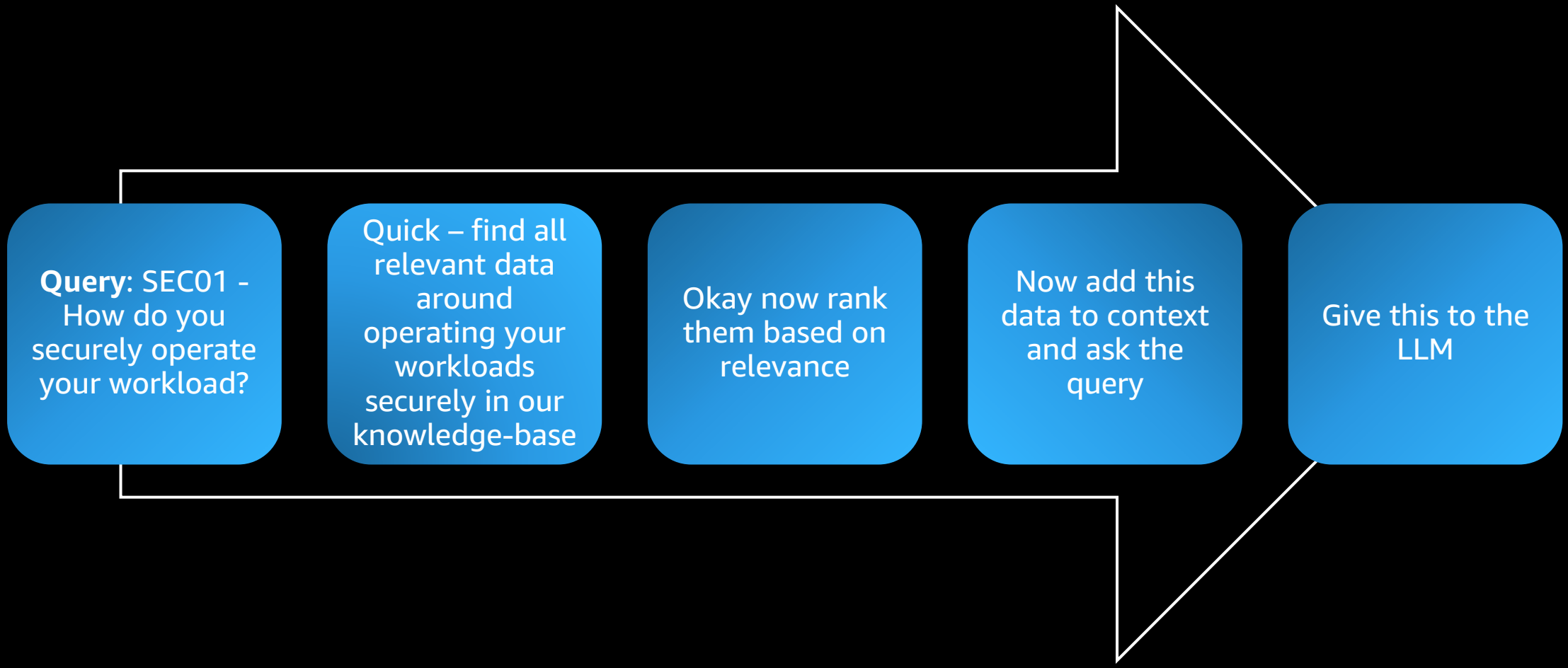
WAFR Accelerator - Architecture



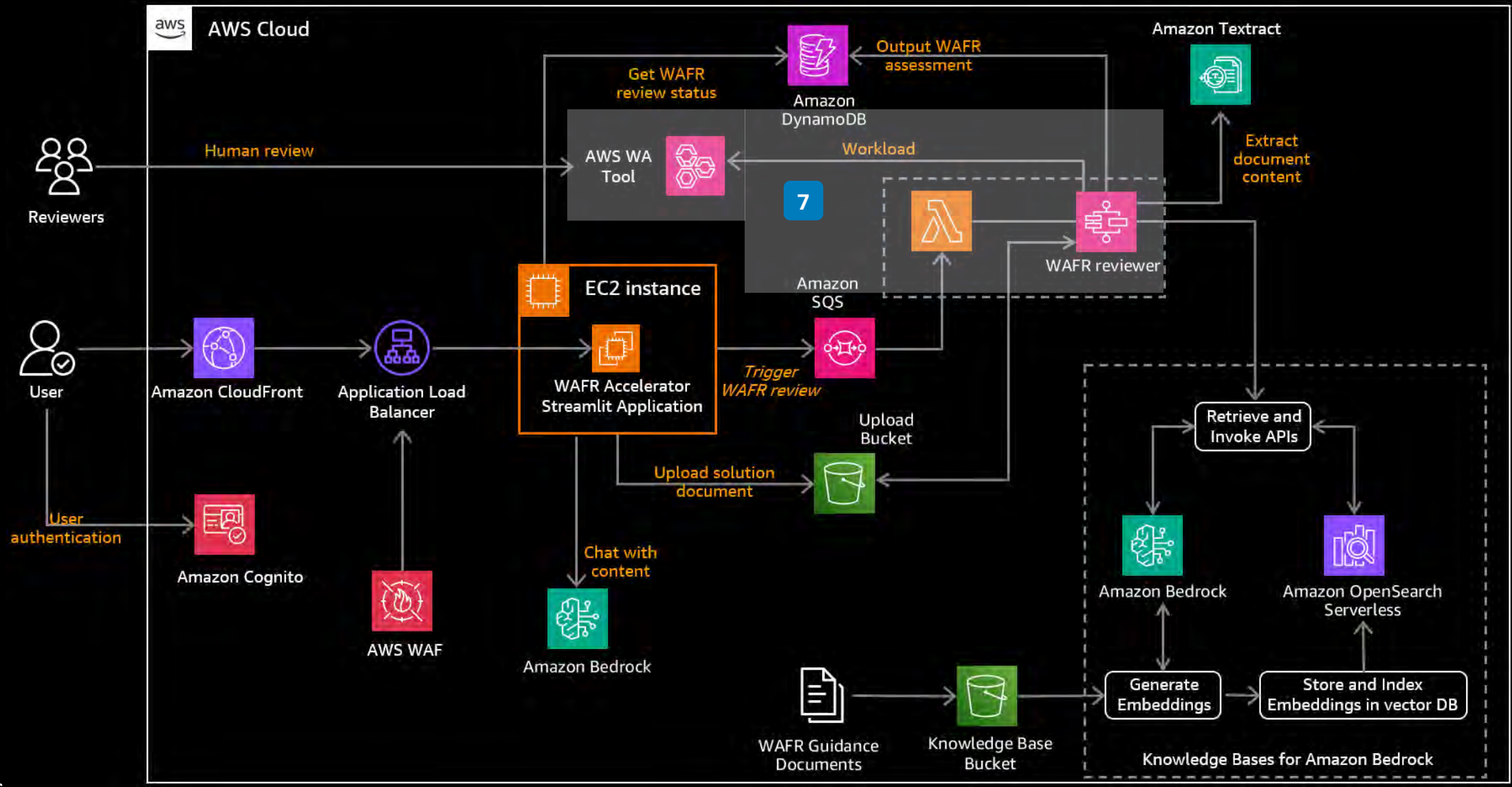
WAFR Accelerator - Architecture



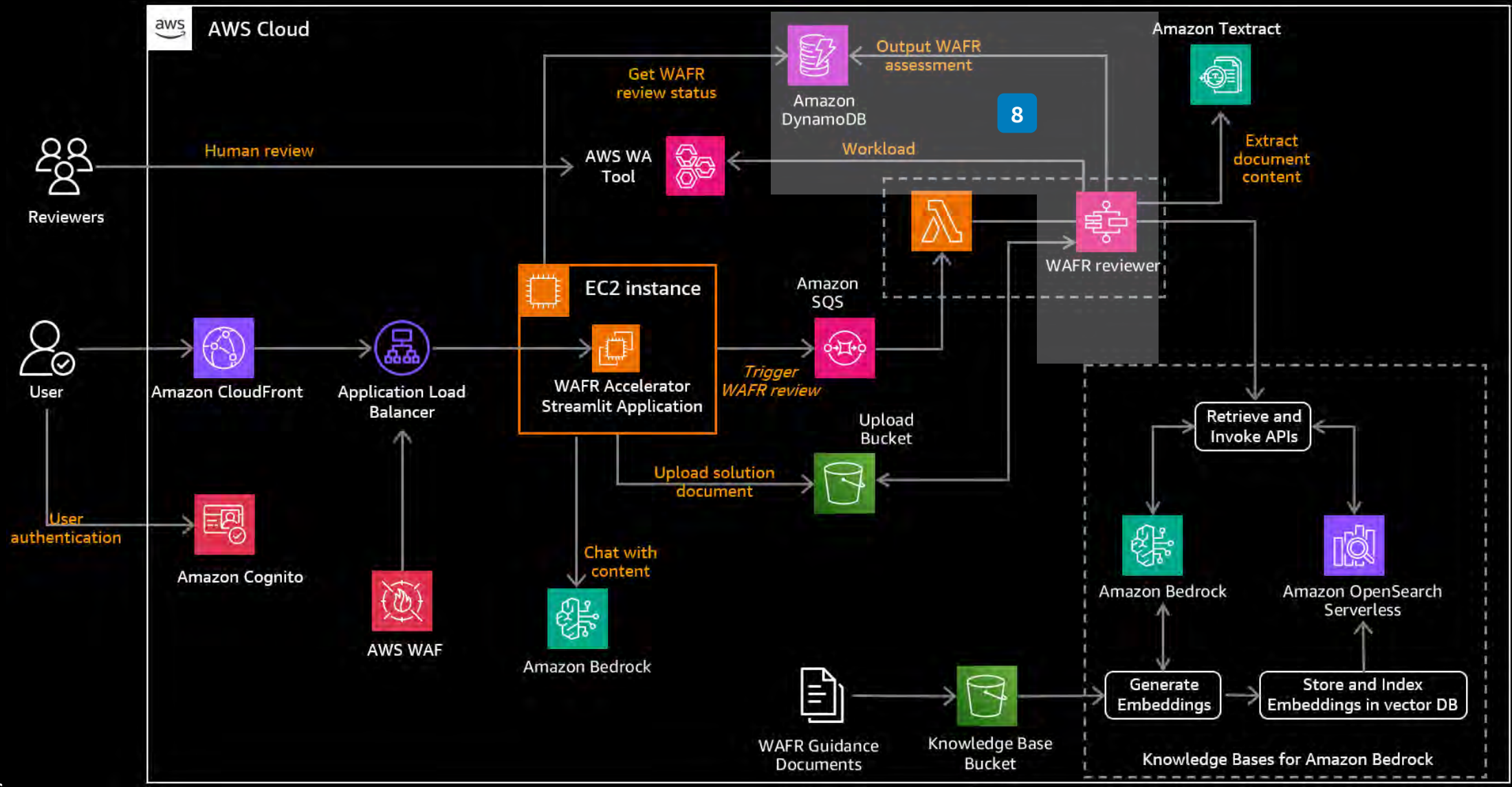
Retrieve and invoke working



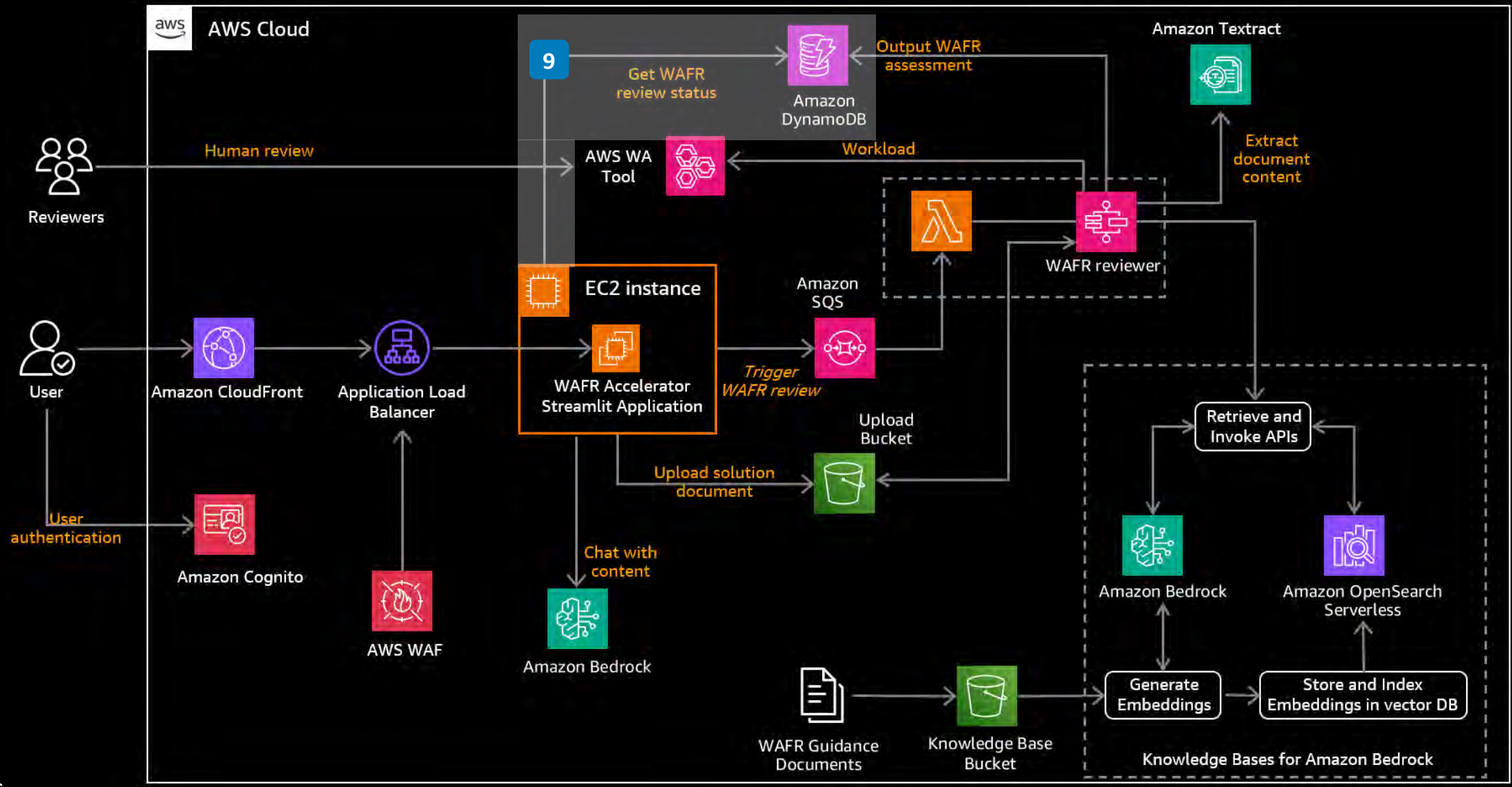
WAFR Accelerator - Architecture



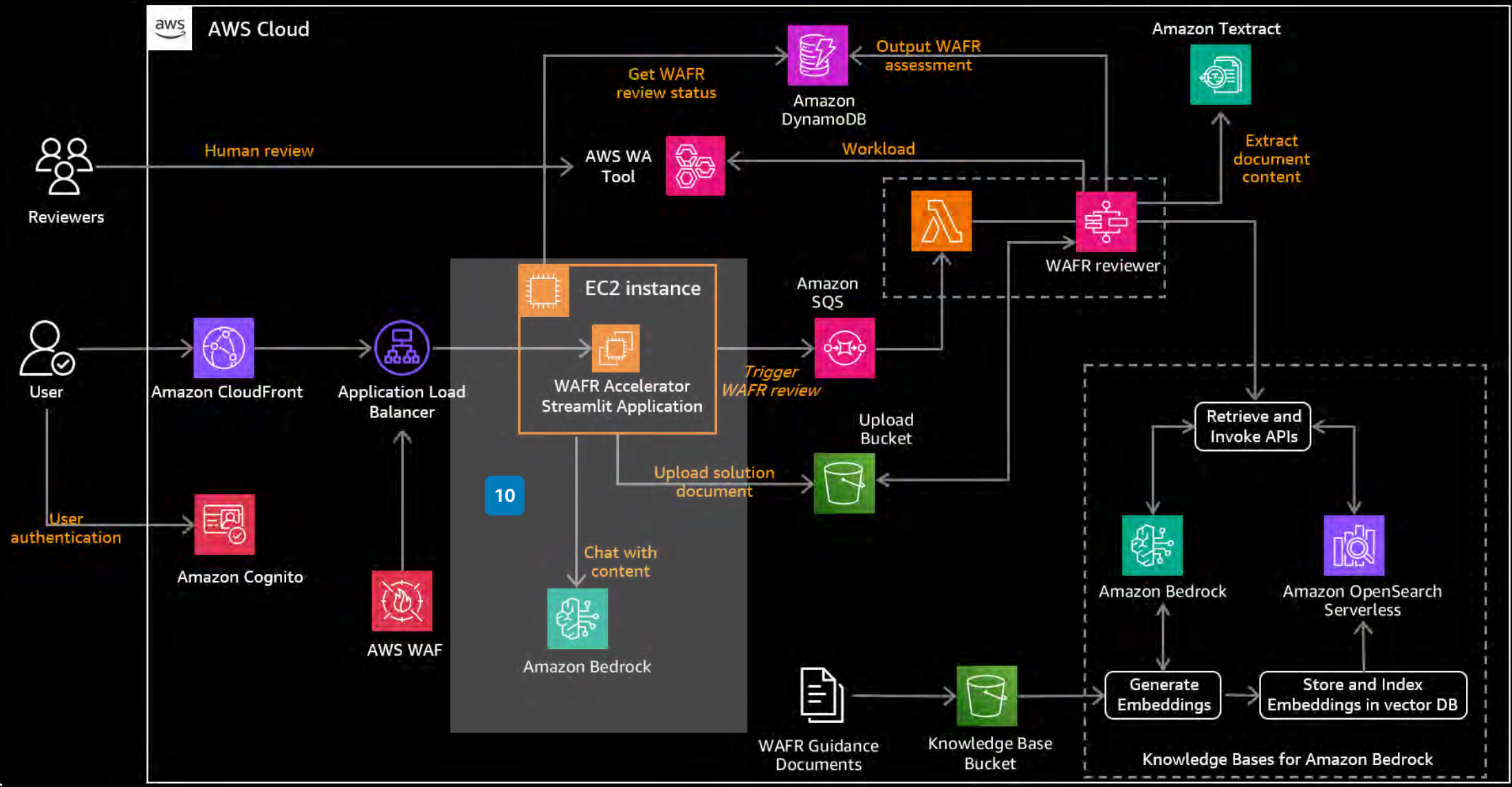
WAFR Accelerator - Architecture



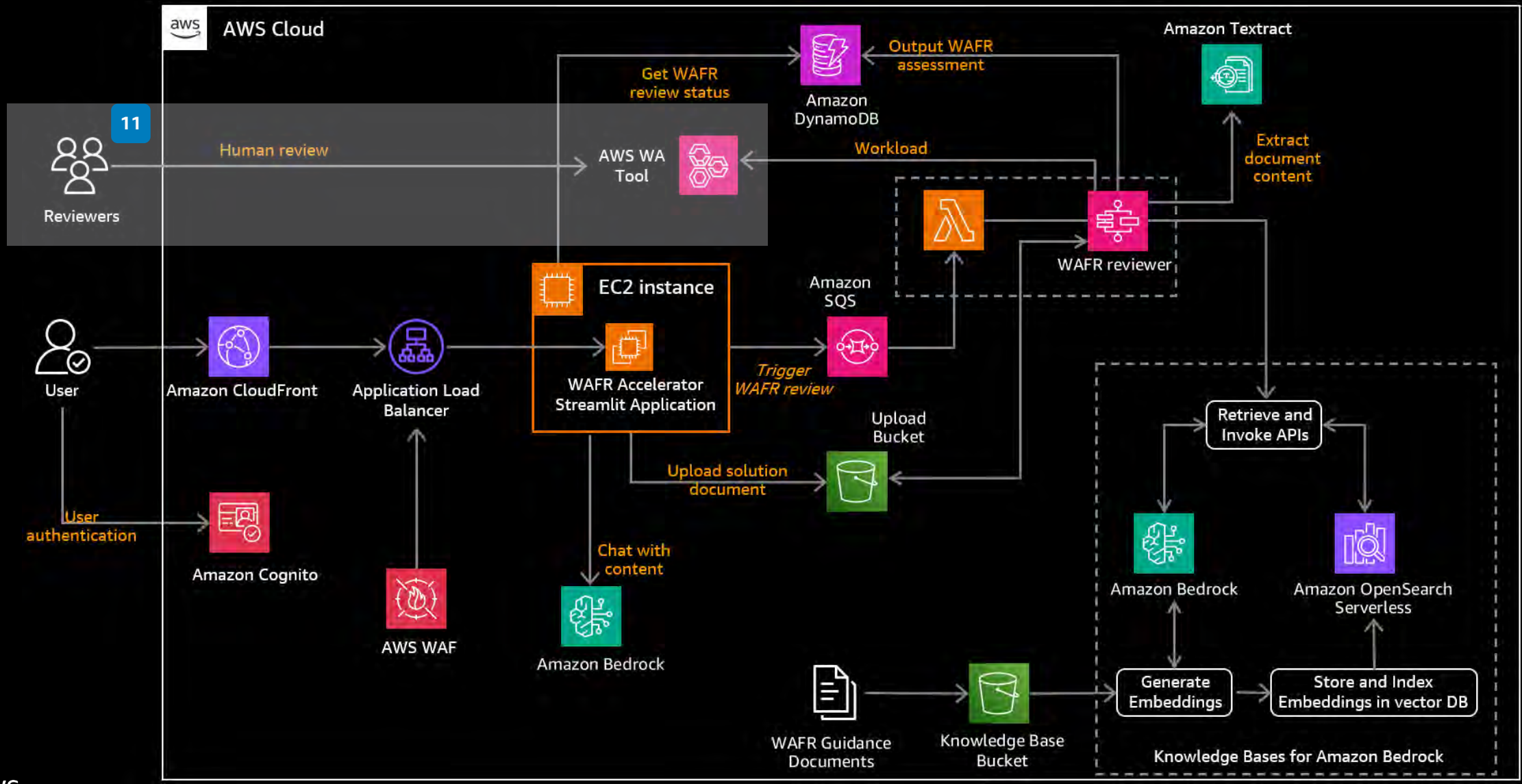
WAFR Accelerator - Architecture



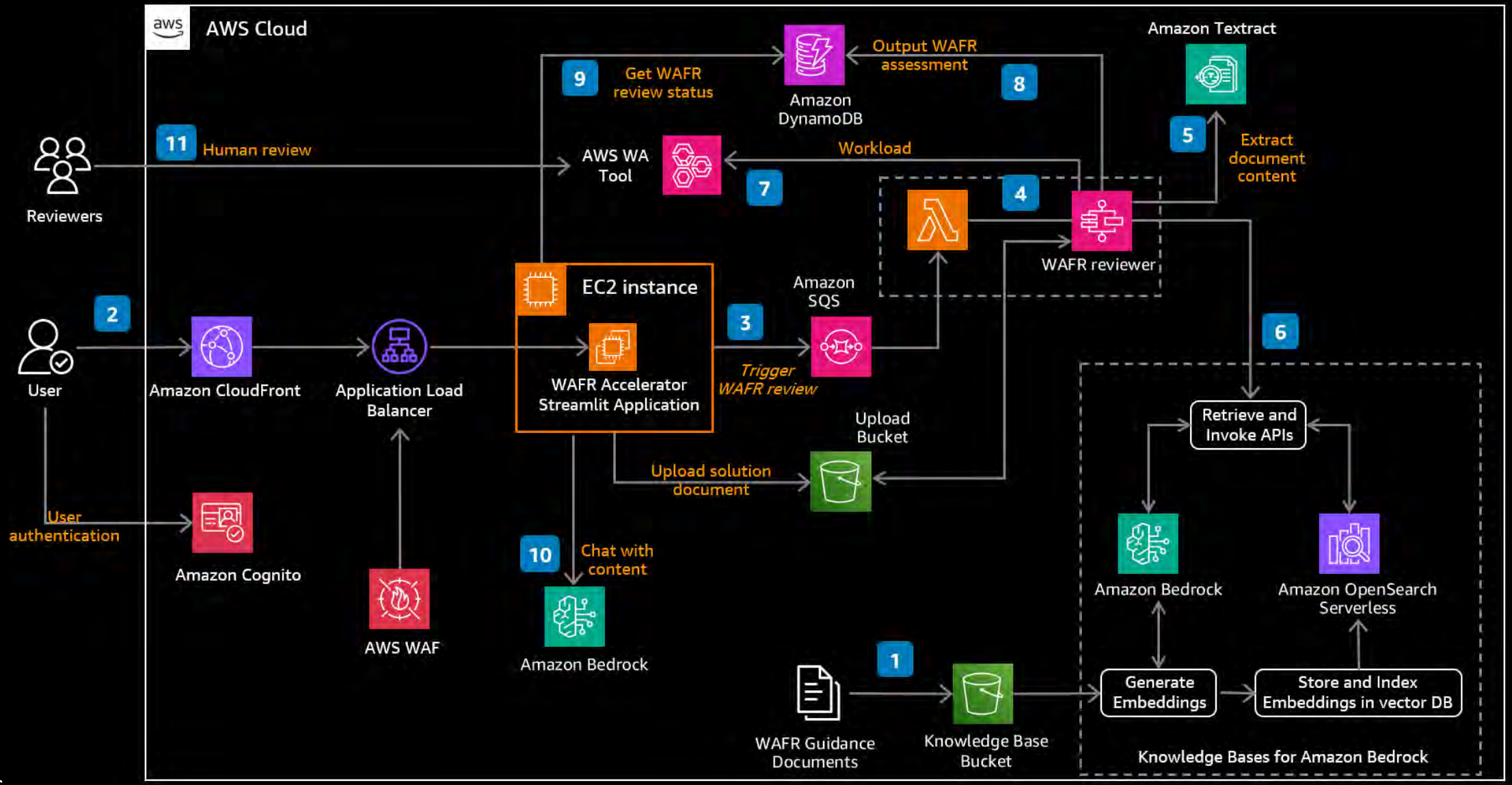
WAFR Accelerator - Architecture



WAFR Accelerator - Architecture

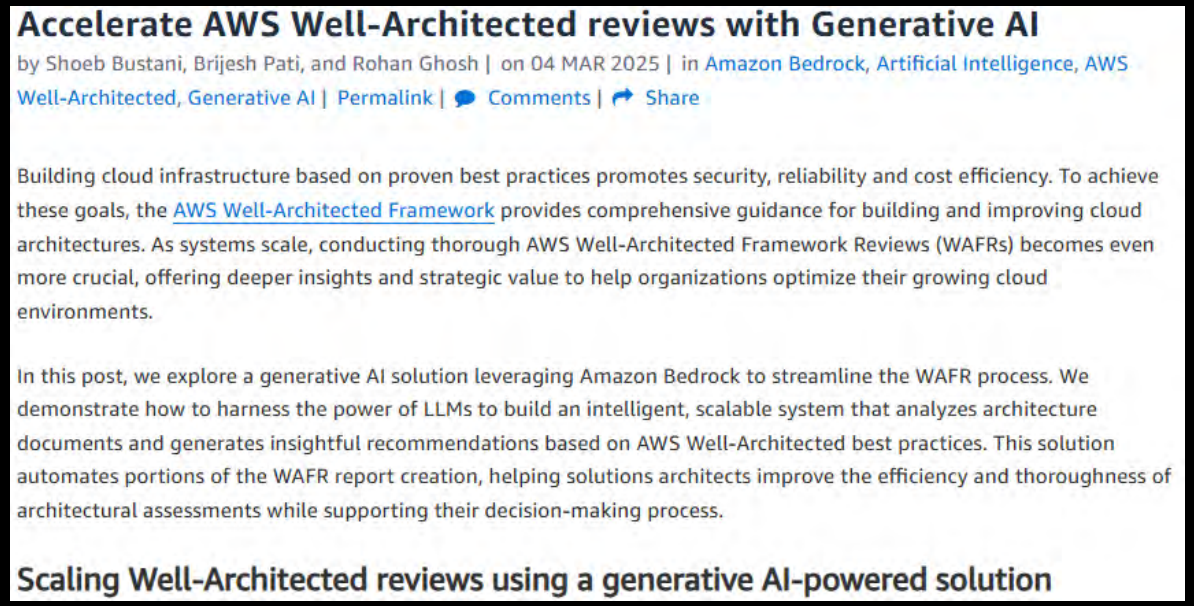


WAFR Accelerator - Architecture



Resources

AWS Blog:



Accelerate AWS Well-Architected reviews with Generative AI

by Shoeb Bustani, Brijesh Pati, and Rohan Ghosh | on 04 MAR 2025 | in Amazon Bedrock, Artificial Intelligence, AWS Well-Architected, Generative AI | Permalink | Comments | Share

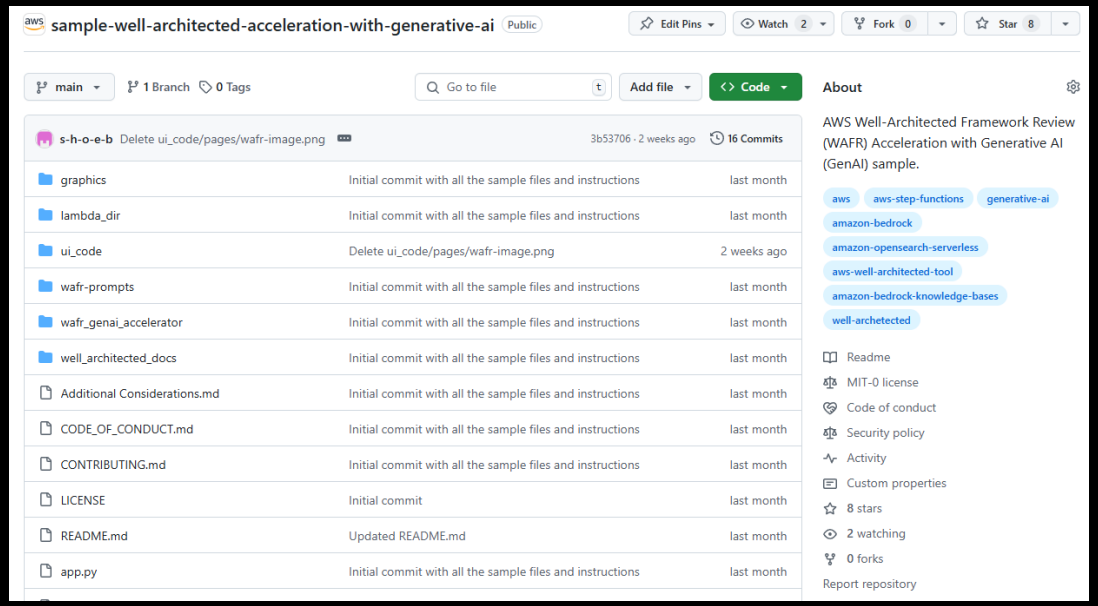
Building cloud infrastructure based on proven best practices promotes security, reliability and cost efficiency. To achieve these goals, the AWS Well-Architected Framework provides comprehensive guidance for building and improving cloud architectures. As systems scale, conducting thorough AWS Well-Architected Framework Reviews (WAFRs) becomes even more crucial, offering deeper insights and strategic value to help organizations optimize their growing cloud environments.

In this post, we explore a generative AI solution leveraging Amazon Bedrock to streamline the WAFR process. We demonstrate how to harness the power of LLMs to build an intelligent, scalable system that analyzes architecture documents and generates insightful recommendations based on AWS Well-Architected best practices. This solution automates portions of the WAFR report creation, helping solutions architects improve the efficiency and thoroughness of architectural assessments while supporting their decision-making process.

Scaling Well-Architected reviews using a generative AI-powered solution

URL: https://aws.amazon.com/blogs/machine-learning/accelerate-aws-well-architected-reviews-with-generative-ai/

AWS Sample:



sample-well-architected-acceleration-with-generative-ai Public

Edit Pins | Watch 2 | Fork 0 | Star 8

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

s-h-o-e-b Delete ui_code/pages/wafr-image.png 3b53706 · 2 weeks ago 16 Commits

graphics	Initial commit with all the sample files and instructions	last month
lambda_dir	Initial commit with all the sample files and instructions	last month
ui_code	Delete ui_code/pages/wafr-image.png	2 weeks ago
wafr-prompts	Initial commit with all the sample files and instructions	last month
wafr_genai_accelerator	Initial commit with all the sample files and instructions	last month
well_architected_docs	Initial commit with all the sample files and instructions	last month
Additional Considerations.md	Initial commit with all the sample files and instructions	last month
CODE_OF_CONDUCT.md	Initial commit with all the sample files and instructions	last month
CONTRIBUTING.md	Initial commit with all the sample files and instructions	last month
LICENSE	Initial commit	last month
README.md	Updated README.md	last month
app.py	Initial commit with all the sample files and instructions	last month

About

AWS Well-Architected Framework Review (WAFR) Acceleration with Generative AI (GenAI) sample.

aws | aws-step-functions | generative-ai | amazon-bedrock | amazon-opensearch-serverless | aws-well-architected-tool | amazon-bedrock-knowledge-bases | well-archetected

Readme | MIT-0 license | Code of conduct | Security policy | Activity | Custom properties | 8 stars | 2 watching | 0 forks | Report repository

Git Hub: https://github.com/aws-samples/sample-well-architected-acceleration-with-generative-ai

Recap

AWS Well-Architected Framework

Scaling opportunities

Generative AI

WAFR Accelerator

Thank you!

Shoeb Bustani
Senior Solutions Architect