

Unleashing the Power of RAG with AWS Bedrock

In today's fast-paced business world, organizations are constantly seeking ways to leverage cutting-edge technologies to streamline their operations and enhance customer experiences. Retrieval-Augmented Generation (RAG) is a game-changing approach that combines the power of large language models with the ability to retrieve relevant information from vast knowledge bases. However, implementing RAG can be a daunting task, requiring complex orchestration frameworks and the management of vector databases, embeddings, and storage.

Enter AWS Bedrock, a game-changer in the realm of RAG deployment. This innovative service from Amazon Web Services (AWS) offers a fully managed and cost-effective solution that simplifies the process of building, deploying, and maintaining RAG applications.

One of the standout features of AWS Bedrock is its seamless integration with Amazon S3 and the Bedrock console. With just a single API call from your client application, you can leverage the RAG implementation managed by Bedrock. This streamlined approach eliminates the need for complex orchestration frameworks, allowing you to focus on building compelling applications rather than grappling with the underlying infrastructure.

At the heart of AWS Bedrock lies the powerful Bedrock orchestrator, which handles the heavy lifting of querying knowledge bases stored in Amazon S3, passing context to large language models like Claude, and returning the generated response to your client application. This robust orchestration process ensures that your RAG applications can leverage the latest advancements in natural language processing while abstracting away the complexities of managing vector databases, embeddings, and storage.

A Real-World Example

To showcase the simplicity of RAG deployment with AWS Bedrock, let's consider a real-world example. Imagine you run a training company, AI Elevate, and you want to provide your users with a seamless experience when searching for course descriptions.

1. Create a Knowledge Base

To use AWS Bedrock, you start by creating a knowledge base in the Bedrock console:

- Step 1 **Provide knowledge base details**
- Step 2 Set up data source
- Step 3 Select embeddings model and configure vector store
- Step 4 Review and create

Provide knowledge base details

Knowledge base details

Knowledge base name

RAGonBedrockDemo

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 50 characters.

Knowledge base description - optional

Enter description

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 200 characters.

IAM permissions

IAM roles are used to access other services on your behalf.

Runtime role

- Create and use a new service role
- Use an existing service role

Service role name

AmazonBedrockExecutionRoleForKnowledgeBase_AI-Elevate

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key

Name

Value - optional

BedrockDemo

Remove

Add new tag

You can add up to 49 more tags.

Cancel

Next

2. Setup your Vector Database

But that's not all – AWS Bedrock also simplifies the management of your vector databases and embeddings.

With just a few clicks in the Bedrock console, you can select an Amazon S3 bucket as a source to store your documents in:

- Step 1
● Provide knowledge base details
- Step 2
● **Set up data source**
- Step 3
○ Select embeddings model and configure vector store
- Step 4
○ Review and create

Set up data source

Set up your data source by specifying the S3 location of your data.

▼ Data source: knowledge-base-quick-start-906mv-data-source

Data source name

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

S3 URI

Add customer-managed KMS key for S3 data - optional
If you encrypted your S3 data, provide the KMS key here so that Bedrock can decrypt it.

▶ Advanced settings - optional

Choose an archive in S3

S3 buckets

Buckets (1/11)

< **1** 2 >

Name	Creation date
<input type="radio"/> ai-elevate-aiops-demo	2024-03-09T15:22:35.000Z
<input type="radio"/> aiops-demo-ct	2023-10-30T13:51:30.000Z
<input type="radio"/> aiops-demo-source	2024-03-10T09:57:00.000Z
<input type="radio"/> app-comp-demp-ct	2024-01-21T16:56:18.000Z
<input type="radio"/> aws-cloudtrail-logs-033886330600-dc129d84	2023-03-13T15:47:17.000Z
<input type="radio"/> aws-sam-cli-managed-default-samclisourcebucket-xduwmlcolaws	2024-01-12T17:42:21.000Z
<input checked="" type="radio"/> bedrock-demo-ct	2024-03-21T11:55:06.000Z
<input type="radio"/> cf-templates-19k4xrvy7t25s-us-east-1	2024-02-27T18:15:57.000Z
<input type="radio"/> ct-storage-gateway-demo	2024-02-13T16:09:08.000Z
<input type="radio"/> sagemaker-studio-033886330600-um6oe0f5pqqo	2024-02-07T15:59:51.000Z

3. Choose a Vector Embedding Model

Then you choose a vector embedding model, and sync your data with the vector database manually or through automation. This level of ease and flexibility empowers you to keep your knowledge bases up-to-date effortlessly, ensuring that your RAG applications always have access to the most recent and relevant information.

The screenshot shows the AWS Bedrock console configuration for a knowledge base. On the left, a progress indicator shows three steps: Step 3 (selected), Step 4, and Review and create. The main content area is titled 'Embeddings model' and 'Vector database'.

Embeddings model
Select an embeddings model to convert your data into an embedding. Pricing depends on the model. [Learn more](#)

- Titan Embeddings G1 - Text v1.2
By Amazon | Vector dimensions: 1536
- Embed English v3
By Cohere | Vector dimensions: 1024
- Embed Multilingual v3
By Cohere | Vector dimensions: 1024

Vector database
Let Amazon create a vector store on your behalf or select a previously created store to allow Bedrock to store, update and manage embeddings. You will be billed directly from the vector store provider. [Learn more](#)

Select how you want to create your vector store.

- Quick create a new vector store - *Recommended*
We will create an Amazon OpenSearch Serverless vector store on your behalf. This cost-efficient option is intended only for development and can't be migrated to production workload later. [Learn more](#)
- Choose a vector store you have created
Select Amazon OpenSearch Serverless, Amazon Aurora, Pinecone or Redis Enterprise Cloud and provide field mappings.

Enable redundancy (active replicas) - *optional*
The default configuration has active replicas disabled, which is optimal for development workloads. Enable this option if you want to enable redundant active replicas, which may increase storage costs.

Add customer-managed KMS key for Amazon OpenSearch Serverless vector - *optional*
If you encrypted your OpenSearch data, provide the KMS key here so that Bedrock can decrypt it.

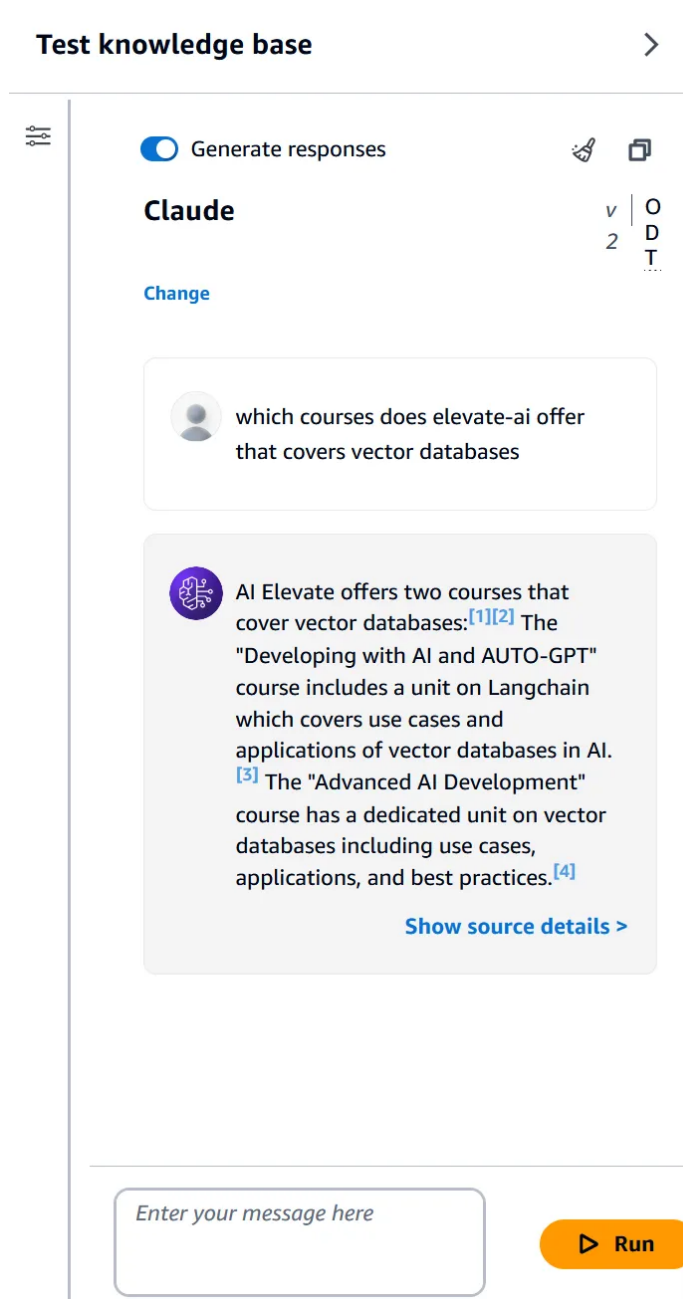
Navigation buttons: Cancel, Previous, Next

With just a few test documents detailing the course offerings loaded into the S3 bucket, the RAG system can instantly retrieve and generate relevant information based on user prompts.

If new course offerings are added by AI Elevate, you simply add the text description to the S3 bucket and Sync the knowledge base, either via the console or via an AWS CLI command.

Testing

But this is not all – the Bedrock console also has a test option built in, so you can test your model:



Whether a user is searching for a specific course topic or exploring related offerings, the RAG application powered by AWS Bedrock will provide accurate and contextual responses, enhancing the overall user experience.

Conclusion

In conclusion, AWS Bedrock is revolutionizing the way organizations approach RAG deployment. By eliminating the complexities of orchestration frameworks and vector embedding models, Bedrock empowers developers to focus on building innovative applications that leverage the full potential of natural language processing. With its cost-effective pricing model, seamless integration with AWS services, and user-friendly console, Bedrock is poised to become one of the go-to solution for businesses seeking to unlock the power of RAG in a streamlined and efficient manner.