

Westcon AWS

GenAI Deployment Guide

Table of contents

1. Background	1
2. Architecture	1
2.1. High level apps integration architecture	1
2.2. How it works	2
3. Solutions component	2
3.1. Amazon Bedrock	2
3.2. Amazon Aurora PostgreSQL	2
3.3. AWS Lambda	3
3.4. Amazon API Gateway	3
3.5. Amazon S3	3
4. AWS Account Configuration	3
4.1. AWS workload account	3
5. AWS workload configuration	4
5.1. Foundation model	4
5.2. Bedrock foundation model availability	4
5.2.1. Activate Foundation model	5
5.3. CloudFormation	7
5.3.1. CloudFormation parameters	7
5.3.2. CloudFormation execution	8
5.4. Database configuration	10
5.4.1. Query editor setup	10
5.4.2. SQL script extension on PostgreSQL	11
5.5. Knowledge base configuration	12
5.5.1. Create knowledge base	12
5.5.2. Setup knowledge base integration	18
5.6. Lambda configuration	20
5.6.1. Lambda Layer Boto3	21
5.6.2. Lambda Layer JWT	23
5.6.3. Lambda token configuration	24
5.6.4. Lambda prompt configuration	25
6. GenAI test API	29
6.1. Get token	29
6.2. Prompt test	30

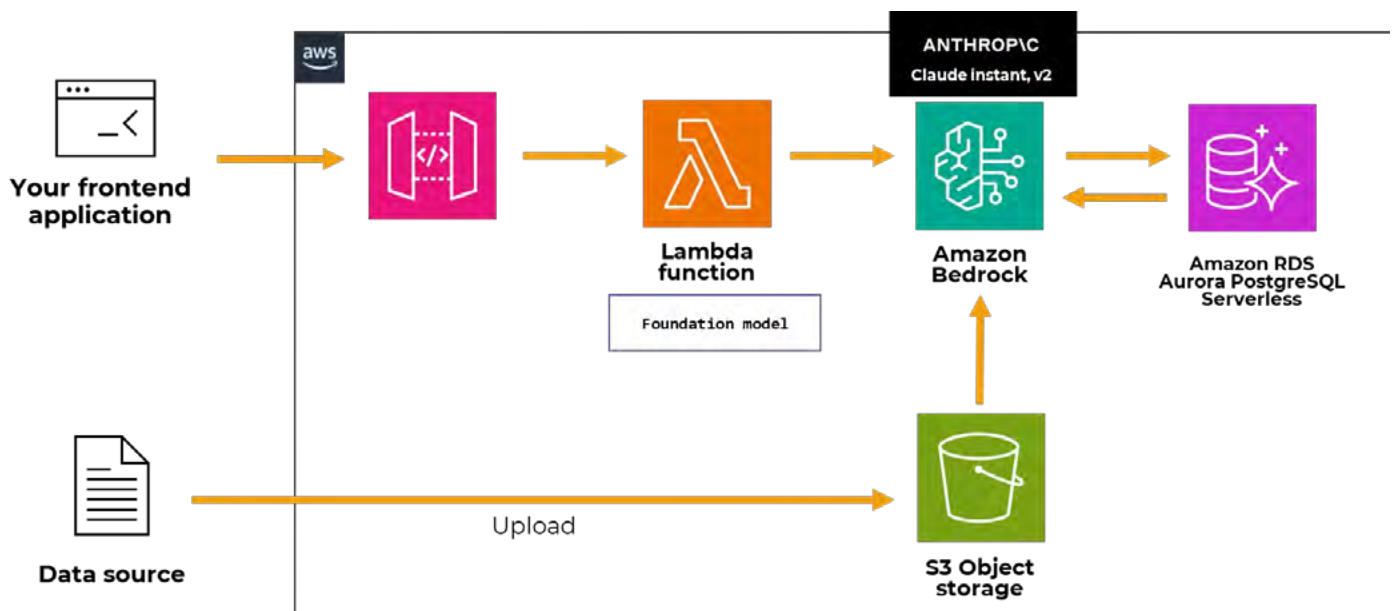
1. Background

AWS Bedrock offers a transformative approach to utilizing generative AI (GenAI) in business and technology, providing a robust, scalable, and cost-effective platform for innovation. The power of AWS Bedrock lies in its ability to streamline and democratize access to cutting-edge GenAI models, making sophisticated AI capabilities available to a broader range of industries and developers. This accessibility fosters innovation, as organizations can leverage advanced AI without the need for extensive infrastructure or expertise. By integrating GenAI into AWS Bedrock, businesses can automate complex processes, enhance customer experiences, and drive decision-making with data-driven insights, thereby staying competitive in an increasingly AI-driven market.

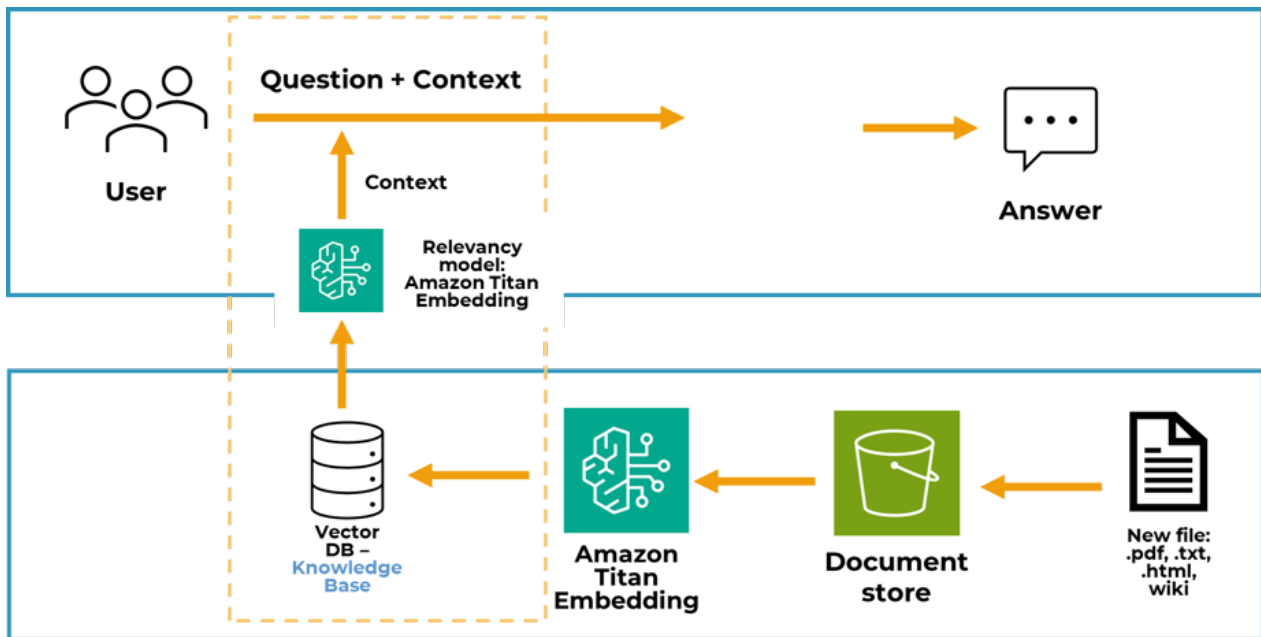
Moreover, AWS Bedrock ensures reliability and security, critical factors when handling sensitive data and deploying AI at scale. With AWS's proven track record in cloud services, Bedrock provides a seamless and secure environment for developing, training, and deploying AI models. This integration allows businesses to focus on innovation rather than infrastructure management. Additionally, AWS Bedrock's flexibility supports various AI use cases, from natural language processing to image generation, enabling organizations to tailor AI solutions to their specific needs. Embracing AWS Bedrock for GenAI thus not only enhances operational efficiency but also drives strategic growth by unlocking the full potential of artificial intelligence.

2. Architecture

2.1. High level apps integration architecture



2.2. How it works



- A document is broken up into chunks of text. The chunks are passed to Titan Embeddings to be converted to vectors. The vectors are then saved to the vector database.
- The user submits a question.
- The question is converted to a vector using Amazon Titan Embeddings, then matched to the closest vectors in the vector database.
- The combined content from the matching vectors + the original question is then passed to the large language model to get the best answer.

3. Solutions component

3.1. Amazon Bedrock

AWS Bedrock is a fully managed service from Amazon Web Services (AWS) that enables users to build and scale generative AI applications with foundation models (FMs) from AI21 Labs, Anthropic, and Stability AI. It simplifies access to these advanced models, offering APIs to integrate them into various applications. AWS Bedrock is designed to support tasks such as text generation, image creation, and data summarization, and it allows customization and fine-tuning of these models according to specific needs.

3.2. Amazon Aurora PostgreSQL

AWS Aurora PostgreSQL is a fully managed relational database service by Amazon Web Services (AWS) that is compatible with PostgreSQL. It combines the performance and availability of high-end commercial databases with the simplicity and cost-effectiveness of open-source databases. Aurora PostgreSQL provides features such as automated backups, replication, and seamless scaling of storage and compute resources. It is designed for enterprise-grade performance and reliability, supporting complex queries and large-scale database applications while ensuring high availability with multi-AZ (Availability Zone) deployments. Aurora also integrates with other AWS services, offering a robust solution for applications requiring a powerful and scalable PostgreSQL-compatible database.

3.3. AWS Lambda

AWS Lambda is a serverless compute service provided by Amazon Web Services (AWS) that allows users to run code without provisioning or managing servers. It automatically scales applications by running code in response to events, such as changes in data, shifts in system state, or user actions. With AWS Lambda, users pay only for the compute time consumed, with billing based on the number of requests and the duration of code execution. It supports various programming languages, including Node.js, Python, Java, and Go, enabling developers to build and deploy applications quickly and efficiently while focusing solely on writing code, without the overhead of managing infrastructure.

3.4. Amazon API Gateway

AWS API Gateway is a fully managed service that allows developers to create, publish, maintain, monitor, and secure APIs at any scale. It acts as a “front door” for applications to access data, business logic, or functionality from backend services such as workloads running on Amazon EC2, AWS Lambda, or any web application. API Gateway handles all the tasks associated with accepting and processing up to hundreds of thousands of concurrent API calls, including traffic management, authorization and access control, monitoring, and API version management. It supports RESTful APIs and WebSocket APIs, enabling real-time two-way communication applications. By using API Gateway, developers can create robust, secure APIs that are easy to maintain and scalable to meet the demands of any application.

3.5. Amazon S3

Amazon Simple Storage Service (AWS S3) is a scalable, high-speed, web-based cloud storage service designed for online data and file storage. Offered by Amazon Web Services (AWS), S3 provides secure, durable, and highly available object storage, capable of storing and retrieving any amount of data from anywhere on the web. It is designed for 99.999999999% (11 9's) durability and offers comprehensive security and compliance capabilities. AWS S3 is used for a wide range of applications, including backup and restore, data archiving, big data analytics, content distribution, and hosting static websites. Users pay only for the storage they use, making it a cost-effective solution for businesses of all sizes.

4. AWS Account Configuration

4.1. AWS workload account

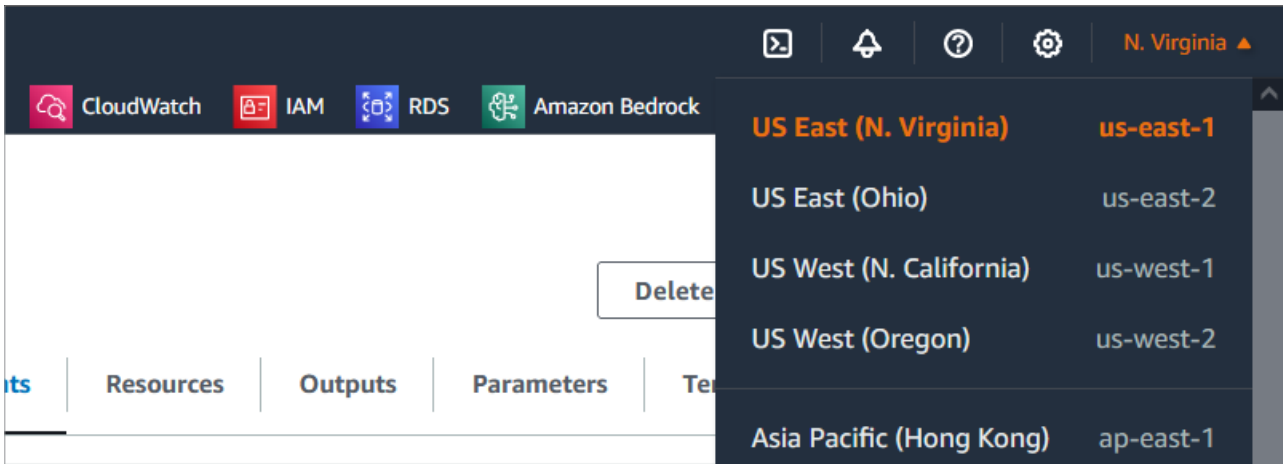
1. **Visit the AWS Website** - Go to the [AWS website](#) and click on the “Create an AWS Account” button.
2. **Sign Up** - Provide your email address, choose a password, and enter an AWS account name. Click on “Continue.”
3. **Contact Information** - Enter your contact information, including your name, address, and phone number. You will also need to choose an account type (Personal or Professional).
4. **Payment Information** - Enter your credit or debit card information. AWS requires this to verify your identity and for billing purposes. Your card will not be charged unless you use paid services.
5. **Identity Verification** - AWS will verify your identity via a phone call or text message. Enter the verification code you receive.
6. **Select a Support Plan** - Choose a support plan that fits your needs. The Basic Support plan is free and sufficient for getting started.
7. **Complete the Setup** - Review the details and complete the setup process.

After completing these steps, you will receive a confirmation email. You can then log in to the AWS Management Console and start using AWS services.

5. AWS workload configuration

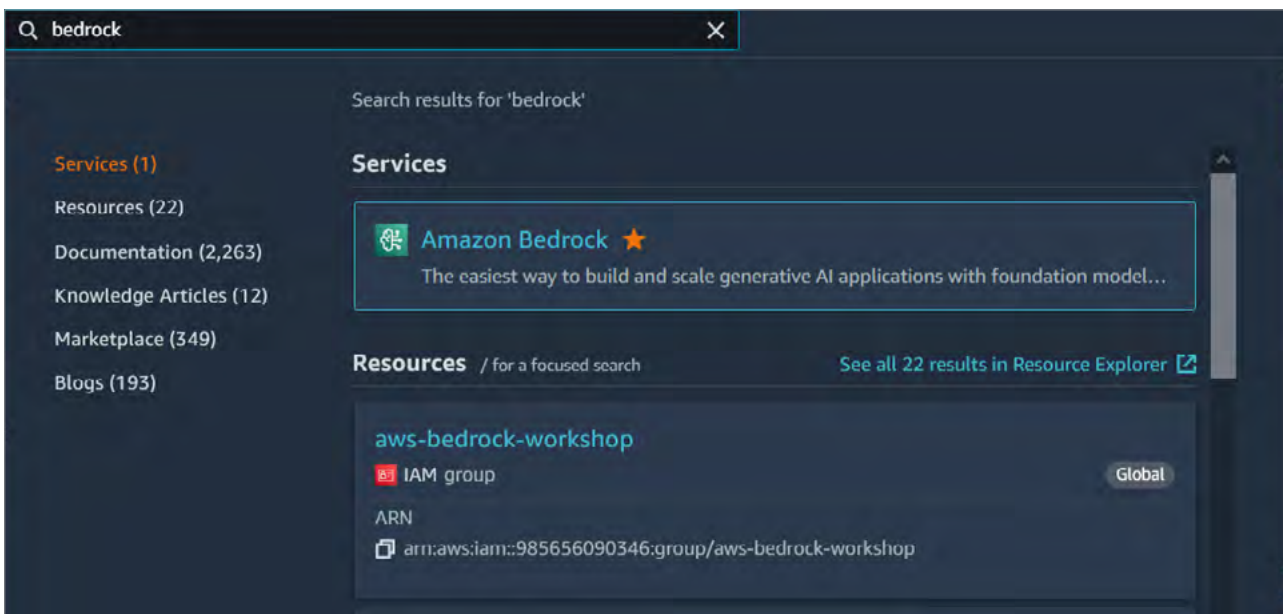
5.1. Foundation model

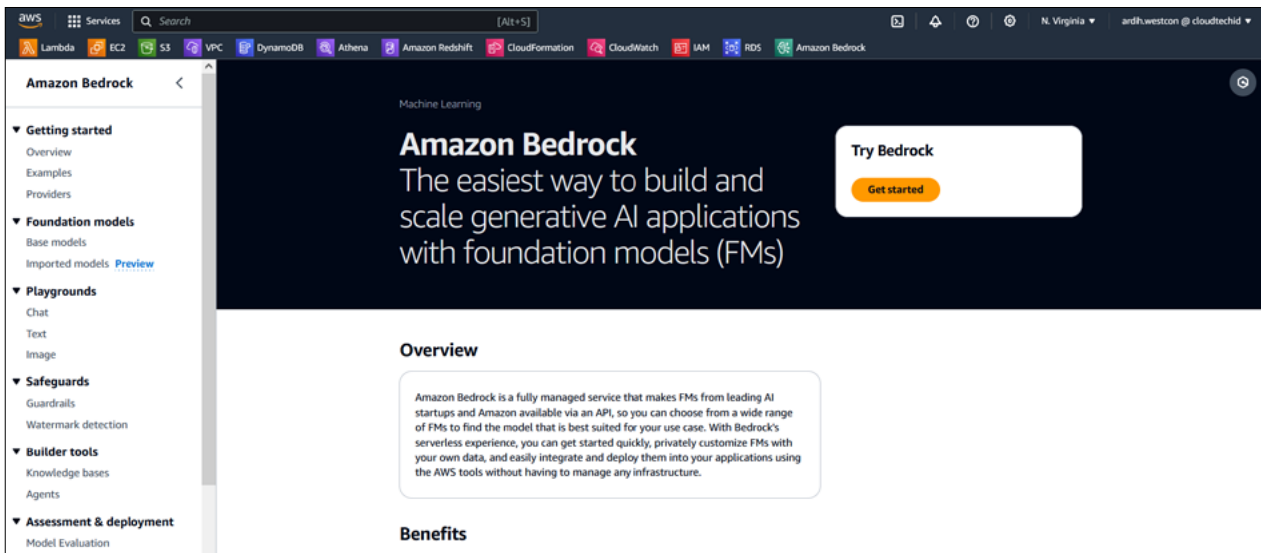
Enabling an AWS Bedrock foundation model involves leveraging AWS's managed service to access and deploy powerful generative AI models. To get started, you must sign in to your AWS Management Console and navigate to the AWS Bedrock service. From there, you can select and activate a foundation model from providers like AI21 Labs, Anthropic, or Stability AI, tailored to your specific use case. AWS Bedrock offers APIs to seamlessly integrate these models into your applications, allowing for tasks such as natural language processing, image generation, and more.



5.2. Bedrock foundation model availability

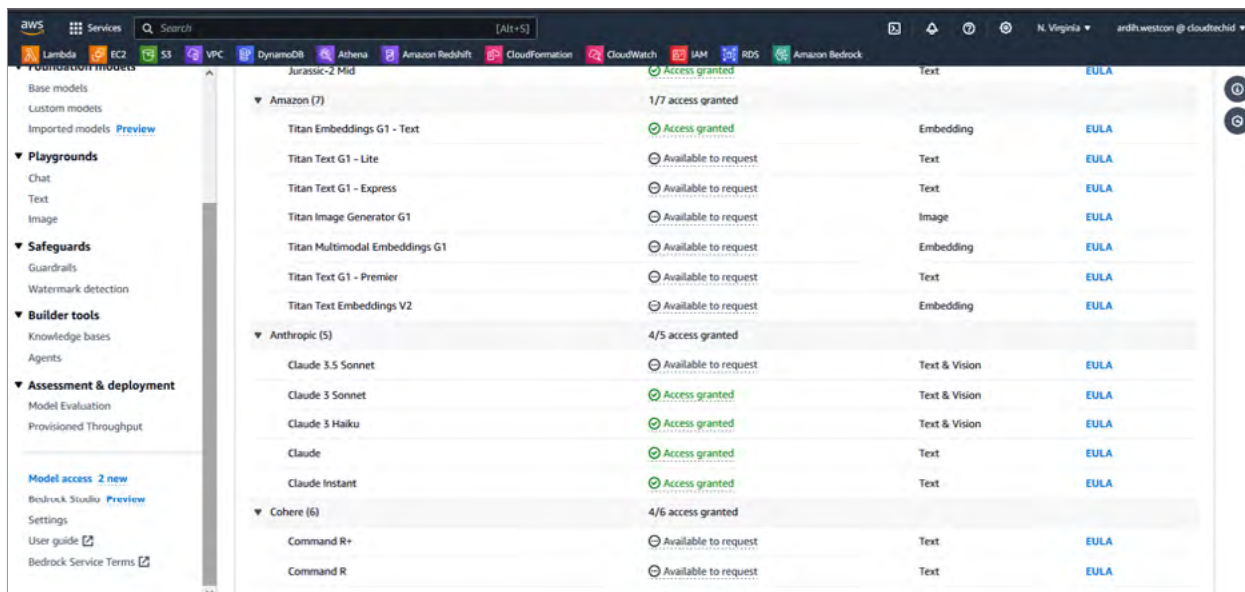
1. Use your AWS credentials to log in to the console. Make sure to check your region. The region should be N. Virginia
2. In the AWS Management Console, type "Bedrock" in the search bar and select it from the services list. Check Amazon Bedrock Foundation Model





5.2.1. Activate Foundation model

In the AWS Bedrock dashboard, you will find options to view and manage available foundation models. These models come from providers like AI21 Labs, Anthropic, and Stability AI.



Provider	Model Name	Access Status	Model Type	EULA Link
Amazon (7)	Titan Embeddings G1 - Text	Access granted	Embedding	EULA
	Titan Text G1 - Lite	Available to request	Text	EULA
	Titan Text G1 - Express	Available to request	Text	EULA
	Titan Image Generator G1	Available to request	Image	EULA
	Titan Multimodal Embeddings G1	Available to request	Embedding	EULA
	Titan Text G1 - Premier	Available to request	Text	EULA
	Titan Text Embeddings V2	Available to request	Embedding	EULA
Anthropic (5)	Claude 3.5 Sonnet	Available to request	Text & Vision	EULA
	Claude 3 Sonnet	Access granted	Text & Vision	EULA
	Claude 3 Haiku	Access granted	Text & Vision	EULA
	Claude	Access granted	Text	EULA
	Claude Instant	Access granted	Text	EULA
Cohere (6)	Command R+	Available to request	Text	EULA
	Command R	Available to request	Text	EULA

3. In the AWS Bedrock dashboard, you will find options to view and manage available foundation models. These models come from providers like AI21 Labs, Anthropic, and Stability AI.

The screenshot displays the AWS Bedrock dashboard interface. It shows a list of foundation models categorized by provider. A modal dialog is open, providing information about models that are 'Available to request'.

Provider	Model Name	Access Status	Model Type	EULA
Amazon (7)	Titan Embeddings G1 - Text	Access granted	Embedding	EULA
	Titan Text G1 - Lite	Available to request	Text	EULA
	Titan Text G1 - Express	Available to request	Text	EULA
	Titan Image Generator G1	Available to request	Image	EULA
	Titan Multimodal Embeddings G1	Available to request	Embedding	EULA
	Titan Text G1 - Premier	Available to request	Text	EULA
	Titan Text Embeddings V2	Available to request		EULA
Anthropic (5)	Claude 3.5 Sonnet	Available to request		EULA
	Claude 3 Sonnet	Access granted	Text & Vision	EULA
	Claude 3 Haiku	Access granted	Text & Vision	EULA
	Claude	Access granted	Text	EULA
	Claude Instant	Access granted	Text	EULA
	Mixtral 8x7B Instruct	Available to request	Text	EULA
Stability AI (1)	SDXL 1.0	Available to request	Image	EULA

Available to request modal text:

You can request access to this model. Billing will start after you are granted access and start using the model in Bedrock. [Request model access](#)

Buttons: Cancel, Next

5.3. CloudFormation

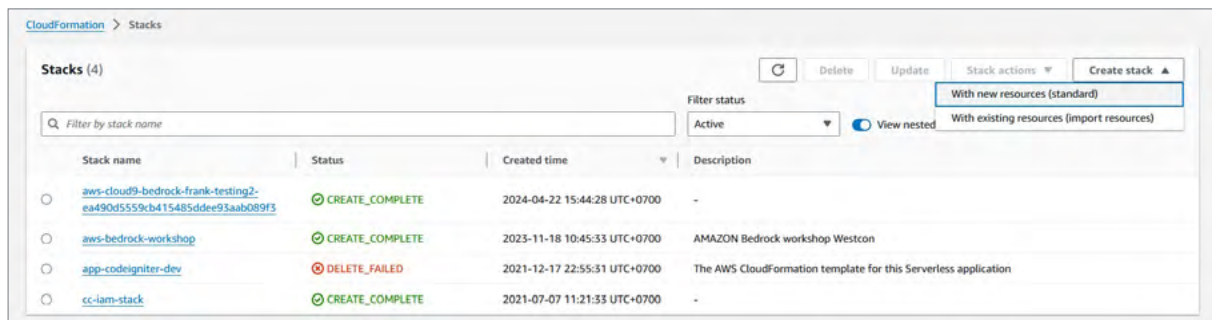
Deploying an AWS CloudFormation stack involves several steps to define and launch AWS resources using a CloudFormation template.

5.3.1. CloudFormation parameters

- 1) **Stack name**
CloudFormation stack name.
- 2) **VpcCidrBlock**
VPC CIDR block, default value is 10.1.0.0/16. Equivalent to 65,536 IP.
- 3) **PublicSubnetCidrBlock**
Public subnet that interacts with internet gateway. Provided 3 subnets for public subnet. Default value:
 - a. **10.1.10.0/24** → PublicSubnetACidrBlock
 - b. **10.1.11.0/24** → PublicSubnetBCidrBlock
 - c. **10.1.12.0/24** → PublicSubnetCCidrBlock
- 4) **PrivateSubnetCidrBlock**
Private subnet that interacts without internet gateway. Provided 3 subnets for private subnet. Default value:
 - a. **10.1.0.0/24** → PrivateSubnetACidrBlock
 - b. **10.1.1.0/24** → PrivateSubnetBCidrBlock
 - c. **10.1.2.0/24** → PrivateSubnetCCidrBlock
- 5) **VPCName**
VPC name for your project.
- 6) **ProjectName**
Project name that affect your service name. For example, your s3 bucket name format would be **<ProjectName>-lambda-layer**
- 7) **AppCode**
Application code to get token from your Lambda function
- 8) **AppSecret**
Application secret to interact with Lambda application JWT. Information about JWT can be found in this link [Introduction to JSON Web Token](#).
- 9) **DBMasterUsername**
User name for your vector database in Aurora PostgreSQL
- 10) **DBmasterUserPassword**
Password for your vector database in Aurora PostgreSQL

5.3.2. CloudFormation execution

- 1) Log in to the AWS Management Console, then navigate to the CloudFormation service.



- 2) Click on “Create stack” and choose either “With new resources (standard)” or “With existing resources (import resources).”

Create stack

Prerequisite - Prepare template

Prepare template

Every stack is based on a template. A template is a JSON or YAML file that contains configuration information about the AWS resources you want to include in the stack.

Choose an existing template

Upload or choose an existing template.

Use a sample template

Choose from our sample template library.

Build from Application Composer

Create a template using a visual builder.

Specify template

A template is a JSON or YAML file that describes your stack's resources and properties.

Template source

Selecting a template generates an Amazon S3 URL where it will be stored.

Amazon S3 URL

Provide an Amazon S3 URL to your template.

Upload a template file

Upload your template directly to the console.

Sync from Git - new

Sync a template from your Git repository.

Upload a template file

Choose file

JSON or YAML formatted file

S3 URL: Will be generated when template file is uploaded

View in Application Composer

- 3) Choose the source of your template

Name	Status	Date modified	Type	Size
genai-starterkit-v.1.0.yaml		6/28/2024 6:02 PM	Yaml Source File	16 KB

- 4) Configure stack details
 - a) **Stack Name:** Provide a unique name for your stack.
 - b) **Parameters:** If your template includes parameters, enter the required values.

Specify stack details

Provide a stack name

Stack name

westcon-genai

Stack name must be 1 to 128 characters, start with a letter, and only contain alphanumeric characters. Character count: 13/128.

Westcon Starter Kit GenAI Implementation Guide

8 | 31

Naming convention

VPCName
VPC name. For example: project-vpc.

ProjectName
Project name without space. For example: project, project-prod, project-dev

Database settings

DBMasterUsername
The database admin account username

DBMasterUserPassword
The database admin account password

Cancel

Previous

Next

Capabilities

 The following resource(s) require capabilities: [AWS::IAM::Role]

This template contains Identity and Access Management (IAM) resources. Check that you want to create each of these resources and that they have the minimum required permissions. In addition, they have custom names. Check that the custom names are unique within your AWS account. [Learn more](#)

☒ I acknowledge that AWS CloudFormation might create IAM resources with custom names.

Create change set

Cancel

Previous

Submit

- 5) Review all the configurations and template details. Acknowledge that AWS CloudFormation might create IAM resources if your template includes IAM roles or policies. Then, click “Create stack”
- 6) CloudFormation will start creating the resources defined in your template. You can monitor the progress in the “Events” tab of your stack in the CloudFormation console
- 7) Once the stack creation is complete, the status will change to “CREATE_COMPLETE.” You can now see the created resources in the AWS Management Console.

westcon-genai

Delete

Update

Stack actions ▼

Create stack ▼

Stack info

Events

Resources

Outputs

Parameters

Template

Change sets

Git sync - new

Events (98)

Detect root cause



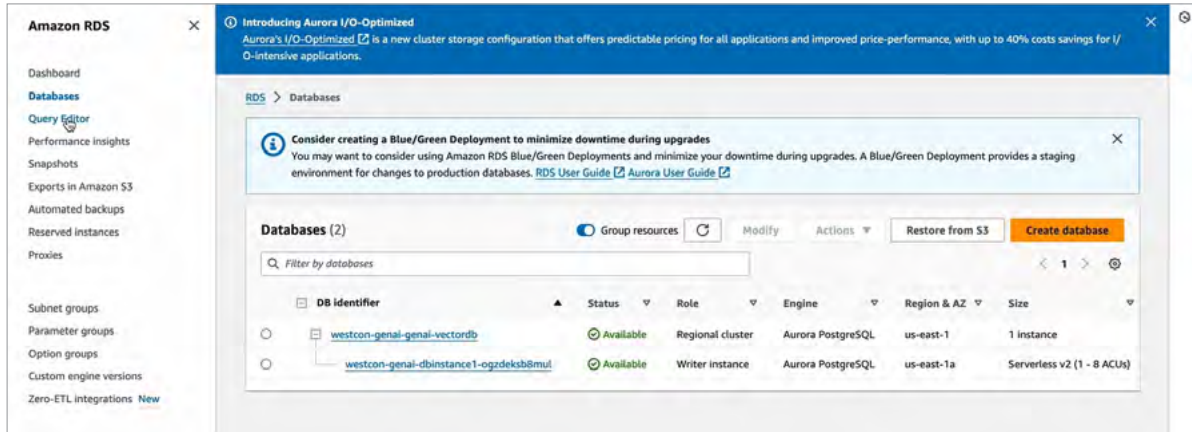


Timestamp	Logical ID	Status	Detailed status	Status reason
2024-06-28 20:44:10 UTC+0700	westcon-genai	✔ CREATE_COMPLETE	-	-
2024-06-28 20:44:09 UTC+0700	DBInstance1	✔ CREATE_COMPLETE	-	-

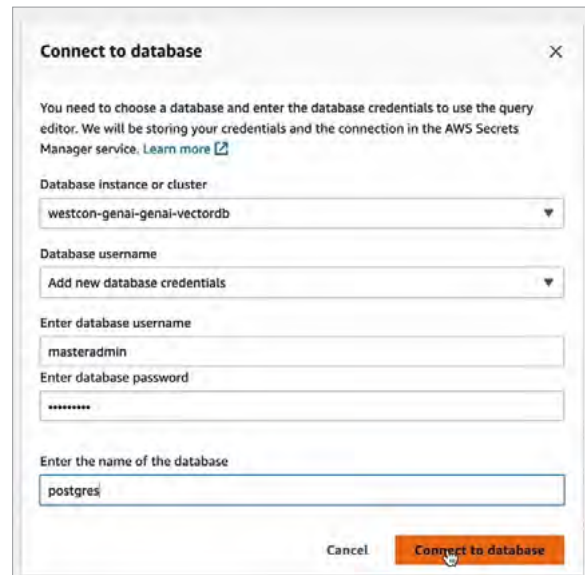
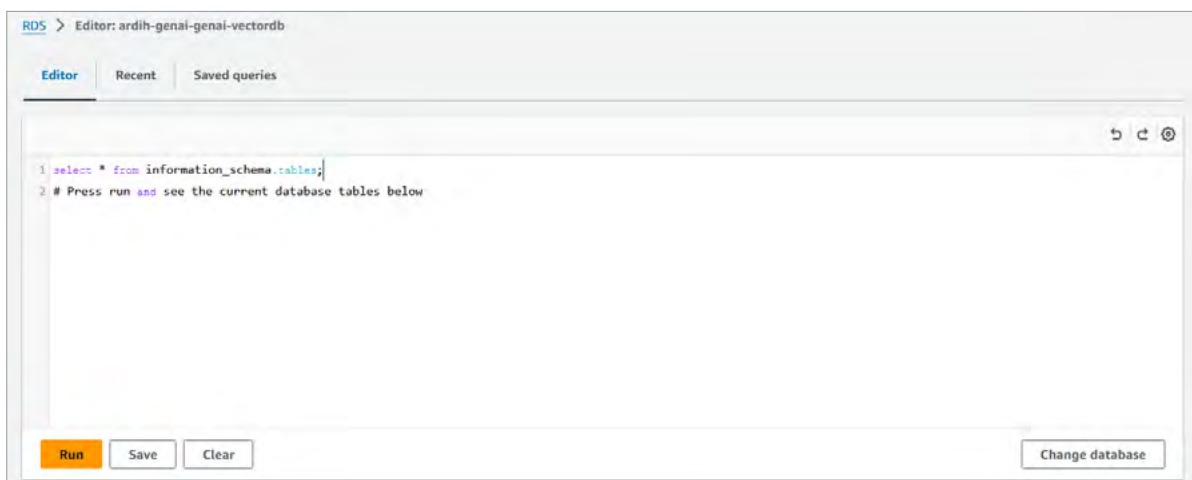
5.4. Database configuration

5.4.1. Query editor setup

1. In the AWS Management Console, type “RDS” in the search bar and select it from the services list.



2. Setup database credentials. Ensure you have the database username and password for your RDS instance, as you will need these credentials to connect to the Query Editor.

Output **Result set 1 (197)**

Rows returned (197)

Q Search rows

Export to csv

table_catalog	table_schema	table_name	table_type	self_referencing_column_name	reference_generation	user_defined_type_catalog
postgres	pg_catalog	pg_type	BASE TABLE	NULL	NULL	NULL
postgres	pg_catalog	pg_foreign_table	BASE TABLE	NULL	NULL	NULL
postgres	pg_catalog	pg_roles	VIEW	NULL	NULL	NULL

5.4.2. SQL script extension on PostgreSQL

- Here are the query command to create vector extension.

```
CREATE EXTENSION IF NOT EXISTS vector;

SELECT extversion FROM pg_extension WHERE extname='vector';

CREATE SCHEMA bedrock_integration;

CREATE ROLE bedrock_user WITH PASSWORD <yourPassword> LOGIN;

GRANT ALL ON SCHEMA bedrock_integration to bedrock_user;

CREATE TABLE bedrock_integration.bedrock_kb (id uuid PRIMARY KEY,
embedding vector(1536), chunks text, metadata json);

CREATE INDEX on bedrock_integration.bedrock_kb USING hnsu (embedding
vector_cosine_ops);
```

- Block each line of the code and hit Run.

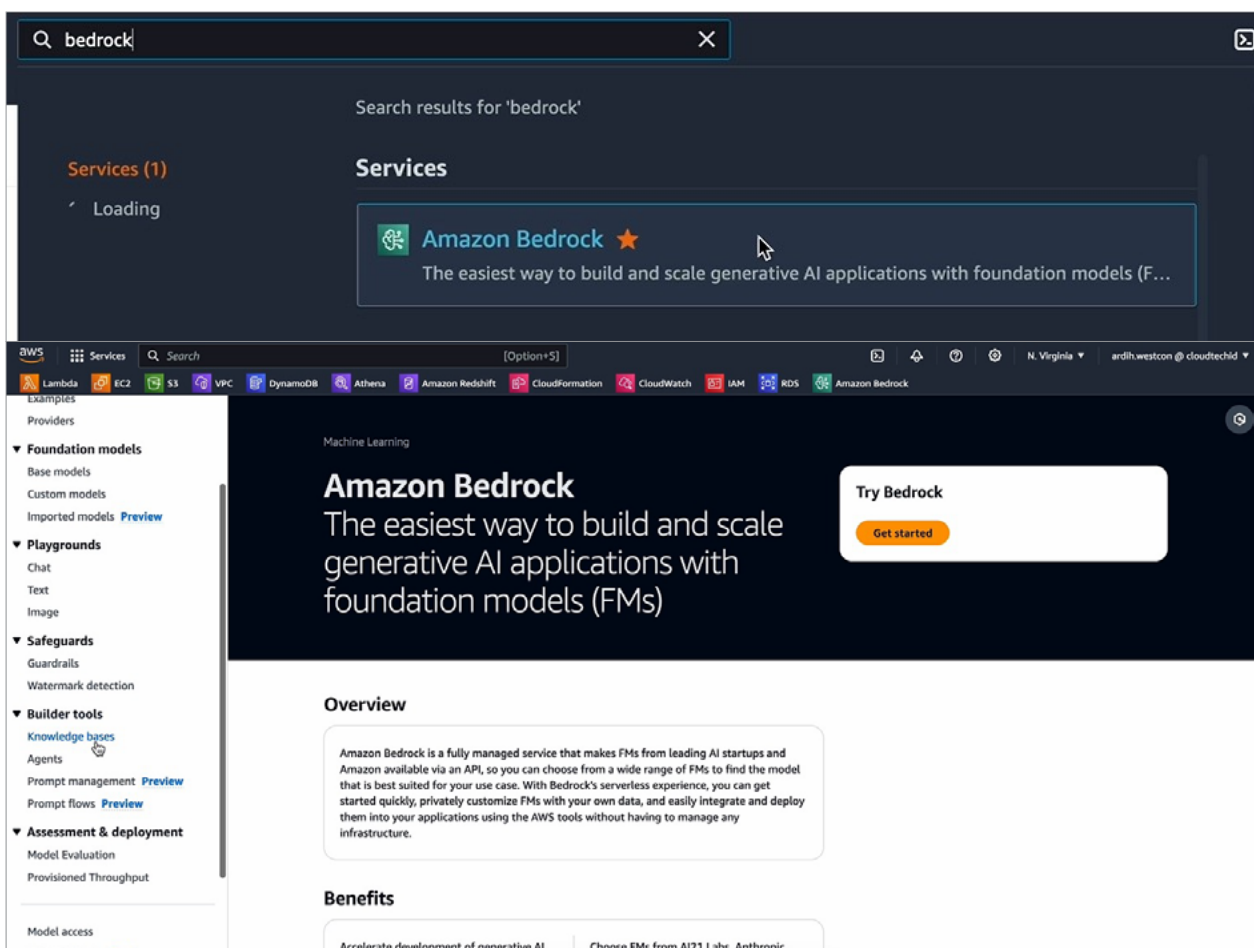


5.5. Knowledge base configuration

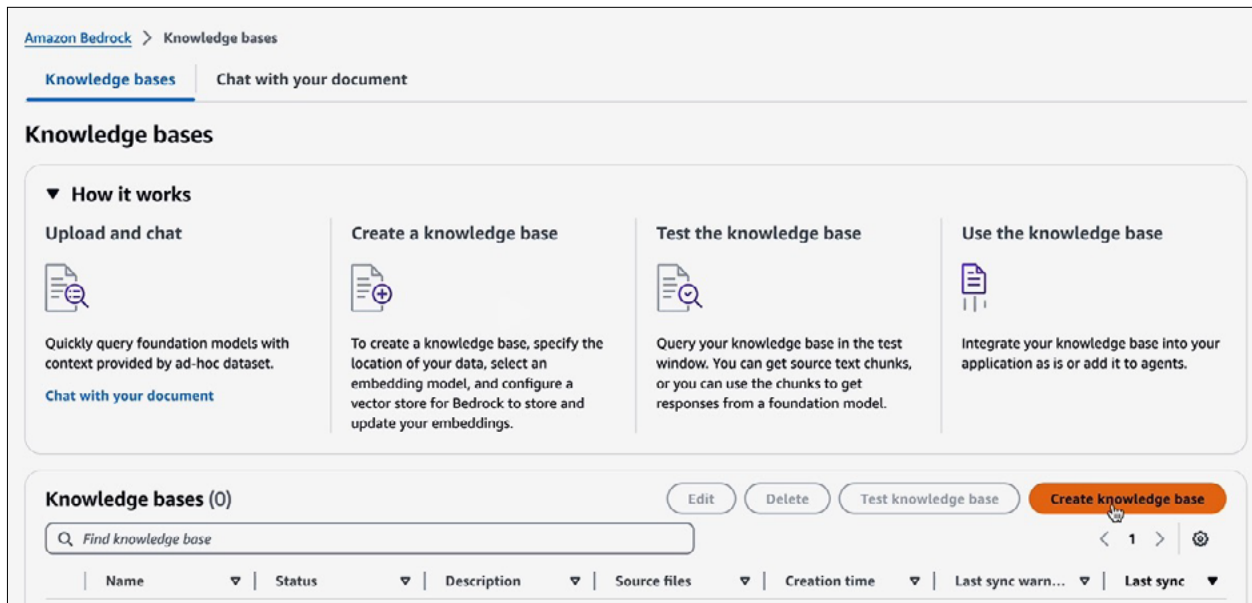
In AWS Bedrock, a knowledge base refers to a centralized repository of information that can be utilized by generative AI models to enhance their performance and accuracy. This repository stores structured and unstructured data, such as documents, FAQs, manuals, and other relevant content, which the AI models can access and learn from. By leveraging a knowledge base, AWS Bedrock can provide more accurate and contextually relevant responses, improve decision-making processes, and streamline information retrieval. This capability is particularly useful for applications requiring detailed and specific knowledge, such as customer support, technical documentation, and content creation, enabling more intelligent and informed AI-driven interactions.

5.5.1. Create knowledge base

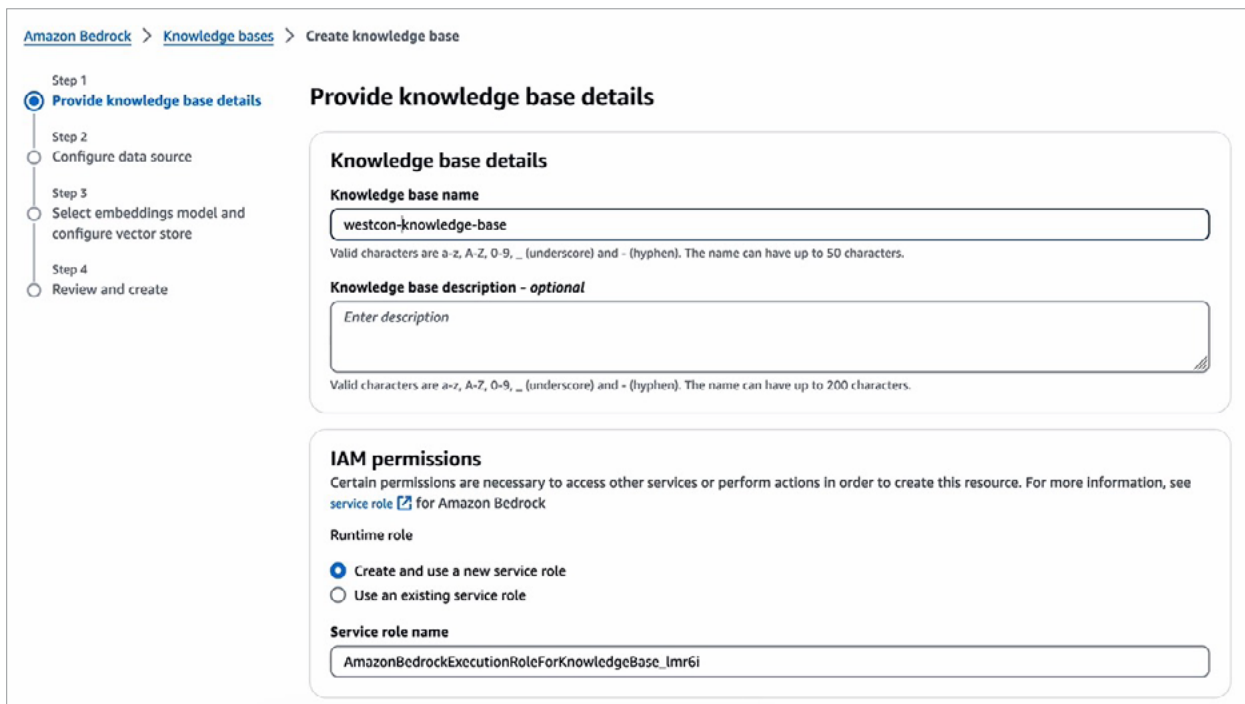
1. In the AWS Management Console, search for “Bedrock” and select it from the services list.



2. Create a New Knowledge Base



3. Provide a name and description for your Knowledge Base.



4. Create IAM permission or use existing role.

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

Certain permissions are necessary to access other services or perform actions in order to create this resource. For more information, see [service role](#) for Amazon Bedrock

Runtime role

☐ Create and use a new service role

☒ Use an existing service role

Choose an existing role or create a new one
To learn more, visit the section *Create a service role and configure permissions* in [our documentation](#).

ArdihAmazonBedrockExecutionRoleForKnowledgeBase_vvf94

If you choose an existing role, check that it is configured with the appropriate permissions.

5. Choose Amazon S3 for data source.

Choose data source

Select the data source that you want to configure in the next step.

Amazon S3 ☒

Object storage service that stores data as objects within buckets.

Web Crawler - Preview ☐

Web page crawler that extracts content from public web pages you are authorized to crawl.

Third party data sources

Confluence - Preview ☐

Collaborative work-management tool designed for project planning, software development and product management.

Salesforce - Preview ☐

Customer relationship management (CRM) tool for managing support, sales, and marketing data.

Sharepoint - Preview ☐

Collaborative web-based service for working on documents, web pages, web sites, lists, and more.

Westcon Starter Kit GenAI Implementation Guide

14 | 31

6. Navigate to S3 bucket from CloudFormation template.

Amazon S3 [Info](#)
Provide details to connect Amazon Bedrock to your S3 data source.

▼ Data source: westcon-knowledge-base Delete

Data source name

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

Data source location
☒ This AWS account
☐ Other AWS account

Choose an archive in S3 ×

S3 buckets
Buckets (1/24)
 2 No matches 1

Name	Creation date
☒ westcon-genai-kb-us-east-1-985656090346	2024-07-18T16:02:37.000Z
☐ westcon-genai-lambda-layer-us-east-1-985656090346	2024-07-18T16:02:37.000Z

Cancel Choose

S3 URI
To increase the accuracy and relevance of your responses, add a .metadata.json file containing metadata for your data source to your S3 bucket. [Info](#)
 × View Browse S3

☐ **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.

Chunking and parsing configurations [Info](#)
Choose between default or advanced customization.

☒ **Default**
Uses default parsing and chunking strategy.

☐ **Custom**
Customize the parsing and chunking strategy, including using advanced parsing.

► Advanced settings - optional

Add data source
You can add 4 more data source(s).

Cancel Previous Next

7. Use Titan Embeddings V1.2 for Embeddings model

Select embeddings model and configure vector store
Choose an embeddings model to convert the data that you will provide in the next step, and provide details for a vector data store in which Bedrock can store, manage, and update your embeddings. The embeddings model and vector store cannot be changed after creation of knowledge base.

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

 [Titan Text Embeddings v2](#) ×
By Amazon

☐

 [Titan Embeddings G1 - Text v1.2](#) ×
By Amazon

☒

 [Embed English v3](#) ×
By Cohere

☐

 [Embed Multilingual v3](#) ×
By Cohere

☐

8. Choose vector store that we have created, and select for Amazon Aurora.

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, MongoDB Atlas, Pinecone or Redis Enterprise Cloud and provide field mappings.

Select your vector store.

☐ **Vector engine for Amazon OpenSearch Serverless**
 If you are a first time user, create a vector database by visiting [OpenSearch Service](#)

☒ **Amazon Aurora**
 If you are a first time user, create a vector database by visiting [RDS Console](#)

☐ **MongoDB Atlas**
 If you are a first-time user, create a MongoDB Atlas Cluster and Vector Search Index by visiting [MongoDB Atlas](#)

☐ **Pinecone**
 If you are a first time user, create a vector database by visiting [Pinecone](#)

☐ **Redis Enterprise Cloud**
 If you are a first time user, create a vector database by visiting [Redis Enterprise Cloud](#)

9. Fill Aurora DB ARN. Database name would be **postgres** and table name be seen from query editor in this [link](#)

Amazon Aurora DB Cluster ARN

Specify the ARN of your Amazon Aurora DB cluster.

Database name

Specify the name of your Database.

Table name

Specify the Table Name of your Amazon Aurora DB cluster.

10. To find Secret ARN, you can go to AWS Secret Manager page.



11. Copy and paste the ARN to secret ARN field.



12. Index field mapping would like this according to query editor step. You can find the field name in this [link](#).

Index field mapping

Specify names for the following fields in the vector database. Bedrock uses these names to store and index your data. [Learn more](#)

Vector field name

Provide the name of the vector field that you configured in Amazon Aurora. Bedrock will store the vector embeddings in this field.

Metadata field mapping

Specify the names of the following fields in your vector database. Bedrock uses these fields to store your data. If you haven't configured these fields in your vector database before this step, your knowledge base will fail to be created.

Text field name

Provide the name for the text field that you configured in Amazon Aurora. Bedrock will store raw text from your data in chunks in this field.

Bedrock-managed metadata field

Provide a name for the field where Bedrock will store metadata that is required to carry out source attribution and enable data ingestion and query

Primary key

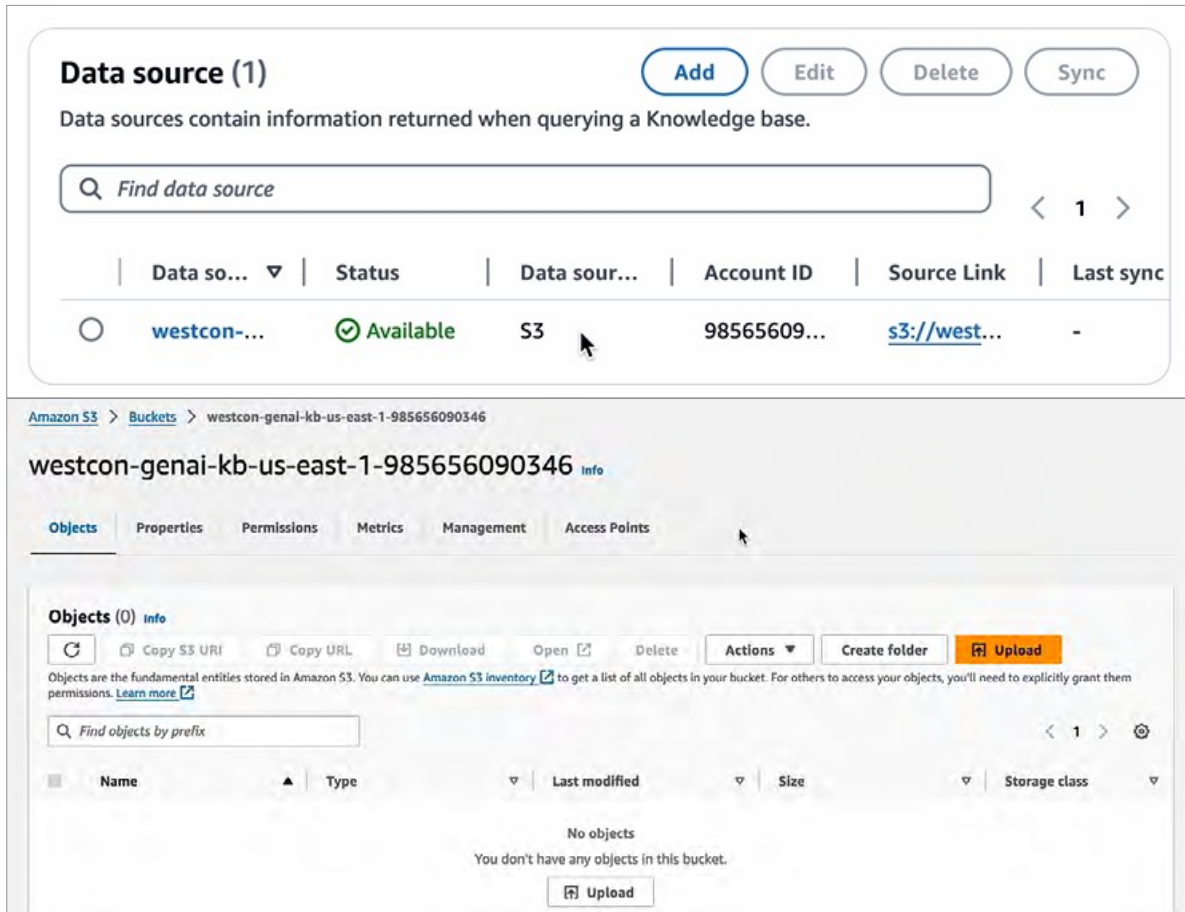
Provide the primary key that you configured in Amazon Aurora.

Cancel
Previous
Next

5.5.2. Setup knowledge base integration

After knowledge base creation, navigate to data source and you should find data source that we have setup.

1. Select data source, and hit the source link. This will navigate you to S3 page.



Data source (1) Add Edit Delete Sync

Data sources contain information returned when querying a Knowledge base.

	Data so...	Status	Data sour...	Account ID	Source Link	Last sync
☐	westcon-...	Available	S3	98565609...	s3://west...	-

[Amazon S3](#) > [Buckets](#) > westcon-genai-kb-us-east-1-985656090346

westcon-genai-kb-us-east-1-985656090346 [Info](#)

[Objects](#) [Properties](#) [Permissions](#) [Metrics](#) [Management](#) [Access Points](#)

Objects (0) [Info](#)

Refresh Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

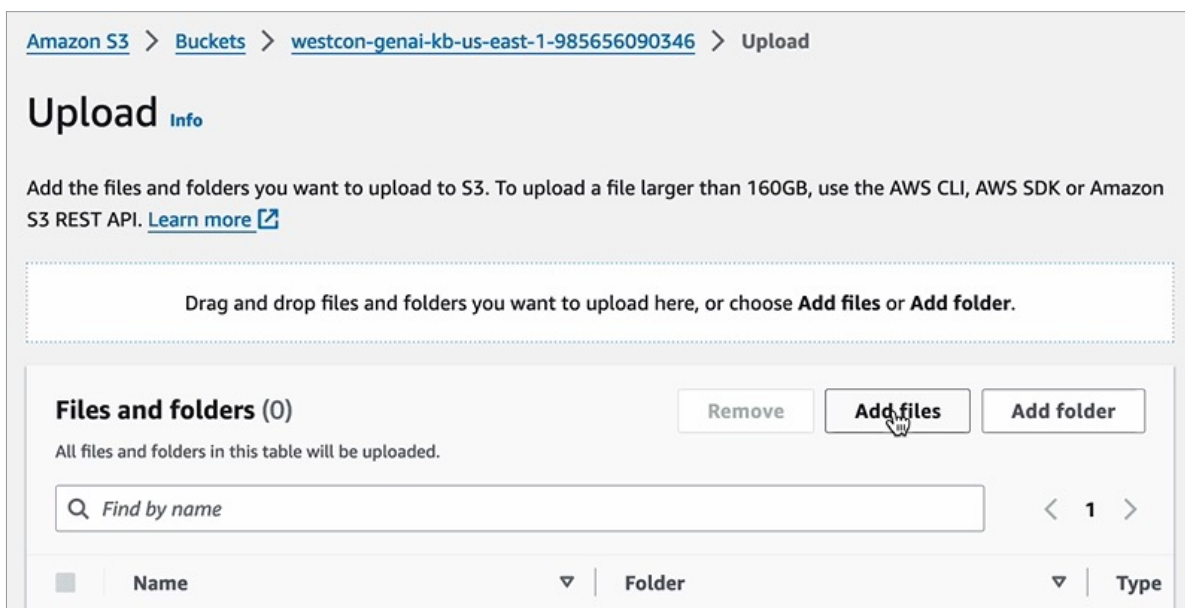
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Name	Type	Last modified	Size	Storage class
No objects				

You don't have any objects in this bucket.

Upload

2. Try to upload one or more documents to test.



[Amazon S3](#) > [Buckets](#) > [westcon-genai-kb-us-east-1-985656090346](#) > Upload

Upload [Info](#)

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

Files and folders (0) Remove Add files Add folder

All files and folders in this table will be uploaded.

Name	Folder	Type
------	--------	------

Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

Files and folders (1 Total, 627.2 KB)

Remove
Add files
Add folder

All files and folders in this table will be uploaded.

< 1 >

☐	Name	Folder	Type
☐	UULaluLintasAngkutanNomor22Tahun2009....	-	applic

Amazon S3 > Buckets > westcon-genai-kb-us-east-1-985656090346

westcon-genai-kb-us-east-1-985656090346 Info

Objects Properties Permissions Metrics Management Access Points

Objects (1) Info

Refresh
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

< 1 >

☐	Name	Type	Last modified	Size	Storage class
☐	UULaluLintasAngkutanNomor22Tahun2009.pdf	pdf	July 19, 2024, 00:01:40 (UTC+07:00)	627.2 KB	Standard

3. Select data source, and hit sync.

Data source (1)

Add
Edit
Delete
Sync

Data sources contain information returned when querying a Knowledge base.

< 1 >

	Data source name	Status	Data sour...	Account ID	Source Link
	westcon-knowledge-base	Available	S3	98565609...	s3://west...

Data source (1)

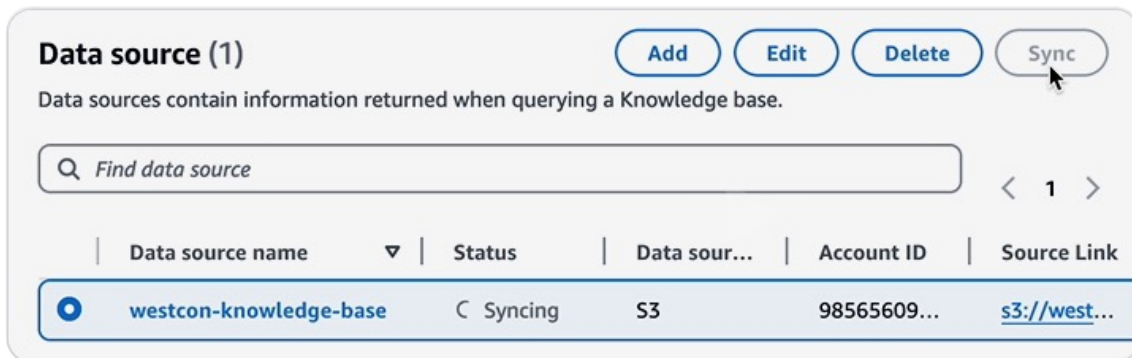
Add
Edit
Delete
Sync

Data sources contain information returned when querying a Knowledge base.

< 1 >

	Data source name	Status	Data sour...	Account ID	Source Link
	westcon-knowledge-base	Available	S3	98565609...	s3://west...

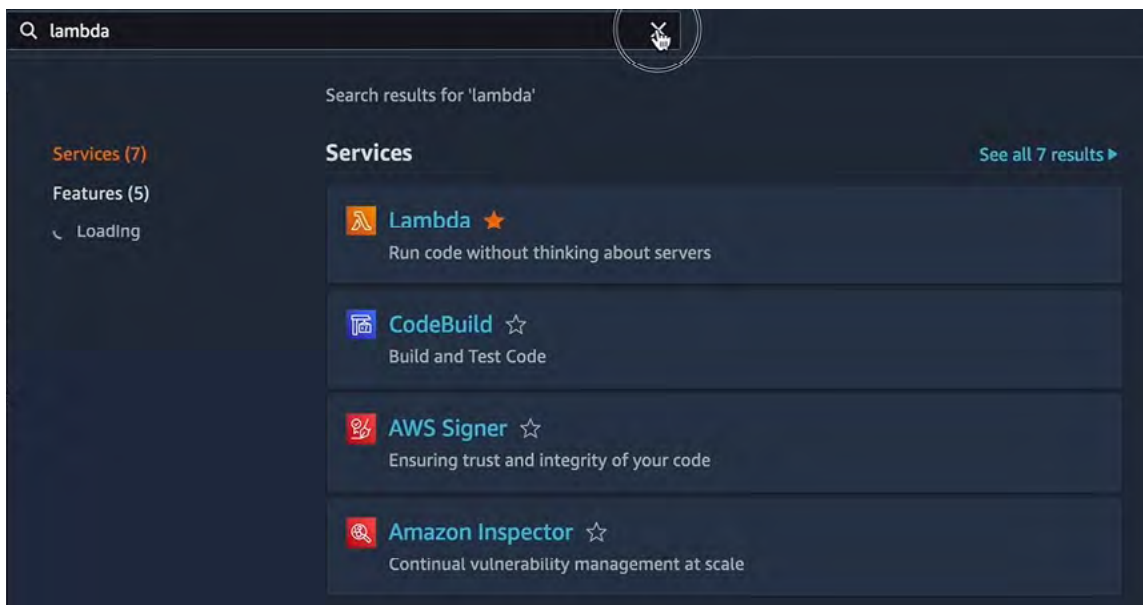
- 4) Wait until syncing process is completed.



5.6. Lambda configuration

Before we setup AWS Lambda Function, we need to add 2 Lambda Layer to work, **Boto3** and **PyJWT**. AWS Lambda layers are used to simplify and streamline the management of dependencies and shared code across multiple Lambda functions. By creating a layer, you can package libraries, custom runtime environments, and other dependencies separately from your function code, which allows for cleaner and more modular code management. This separation enhances reusability and efficiency, as the same layer can be shared across different functions without duplicating code. Layers also make it easier to update dependencies independently of the function code, reducing deployment time and minimizing the risk of errors. This approach not only optimizes the development process but also ensures better version control and consistency across serverless applications.

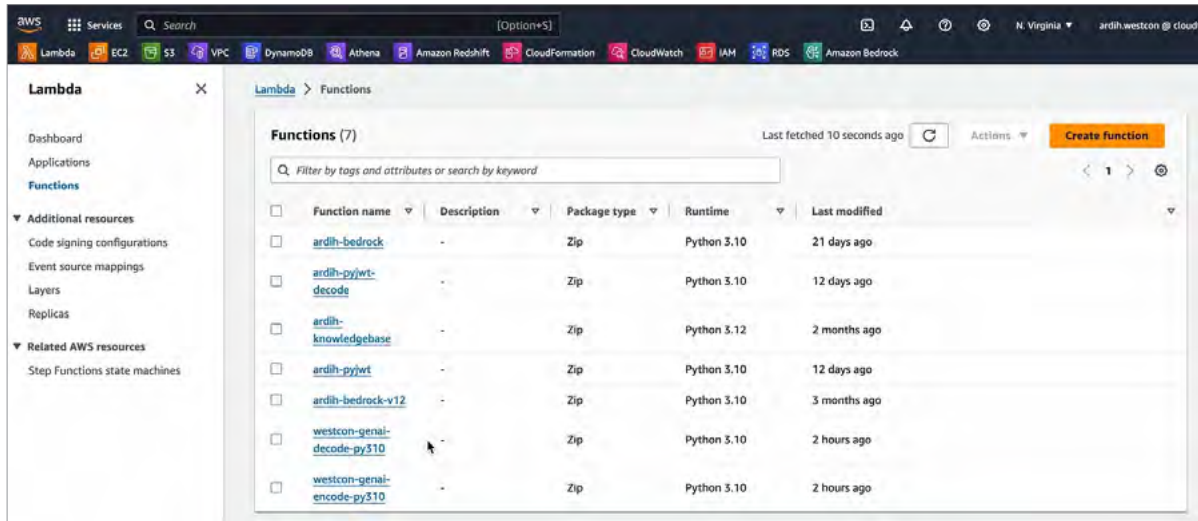
1. Navigate to Lambda page in the AWS console.



5.6.1. Lambda Layer Boto3

Boto3 is the Amazon Web Services (AWS) Software Development Kit (SDK) for Python, enabling developers to interact with AWS services programmatically. It simplifies the process of integrating Python applications with AWS by providing an easy-to-use, consistent interface for managing and utilizing AWS resources like S3, EC2, DynamoDB, Lambda, and more.

1. Expand the left panel and search for Layers menu. Click Create layer.



2. Upload the zip file you created or specify an S3 URL if you've uploaded the zip to an S3 bucket. Fill the field with your preferred name that indicate library name. For this instance, **Westcon-boto3** is the name of layer.

Name

westcon-boto3

Description - optional

boto3

☒ Upload a .zip file

☐ Upload a file from Amazon S3

 Upload

For files larger than 10 MB, consider uploading using Amazon S3.

Compatible architectures - optional [Info](#)

Choose the compatible instruction set architectures for your layer.

☐ x86_64
 ☐ arm64

Compatible runtimes - optional [Info](#)

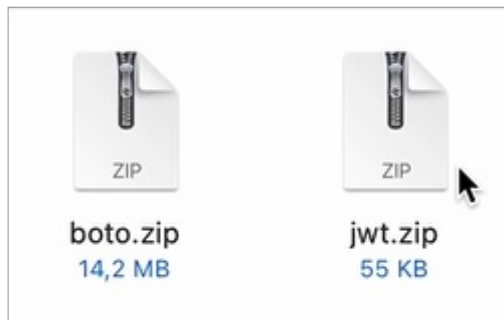
Choose up to 15 runtimes.

Runtimes



License - optional [Info](#)

3. Choose boto.zip



4. Choose x86_64 for compatible architectures, and select all compatible runtime from Python 3.8 to Python 3.12

☒ Upload a .zip file
☐ Upload a file from Amazon S3

[Upload](#)

boto.zip
14.18 MB

For files larger than 10 MB, consider uploading using Amazon S3.

Compatible architectures - optional [Info](#)
Choose the compatible instruction set architectures for your layer.

☒ x86_64
☐ arm64

Compatible runtimes - optional [Info](#)
Choose up to 15 runtimes.

Runtimes ▼ [Refresh](#)

Python 3.12 ✕ Python 3.10 ✕ Python 3.11 ✕ Python 3.8 ✕
Python 3.9 ✕

License - optional [Info](#)

5.6.2. Lambda Layer JWT

PyJWT is a Python library used for encoding and decoding JSON Web Tokens (JWT). JWTs are a compact, URL-safe means of representing claims to be transferred between two parties, commonly used for authentication and information exchange.

1. Upload the zip file you created or specify an S3 URL if you've uploaded the zip to an S3 bucket. Fill the field with your preferred name that indicate library name. For this instance, **Westcon-jwt** is the name of layer.

Name

westcon-jwt

Description - *optional*

jwt

☒ Upload a .zip file

☐ Upload a file from Amazon S3

Upload

For files larger than 50 MB, consider uploading using Amazon S3.

Compatible architectures - *optional* [Info](#)

Choose the compatible instruction set architectures for your layer.

☐ x86_64

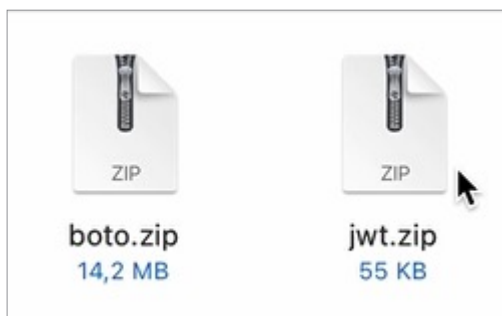
☐ arm64

Compatible runtimes - *optional* [Info](#)

Choose up to 15 runtimes.

Runtimes

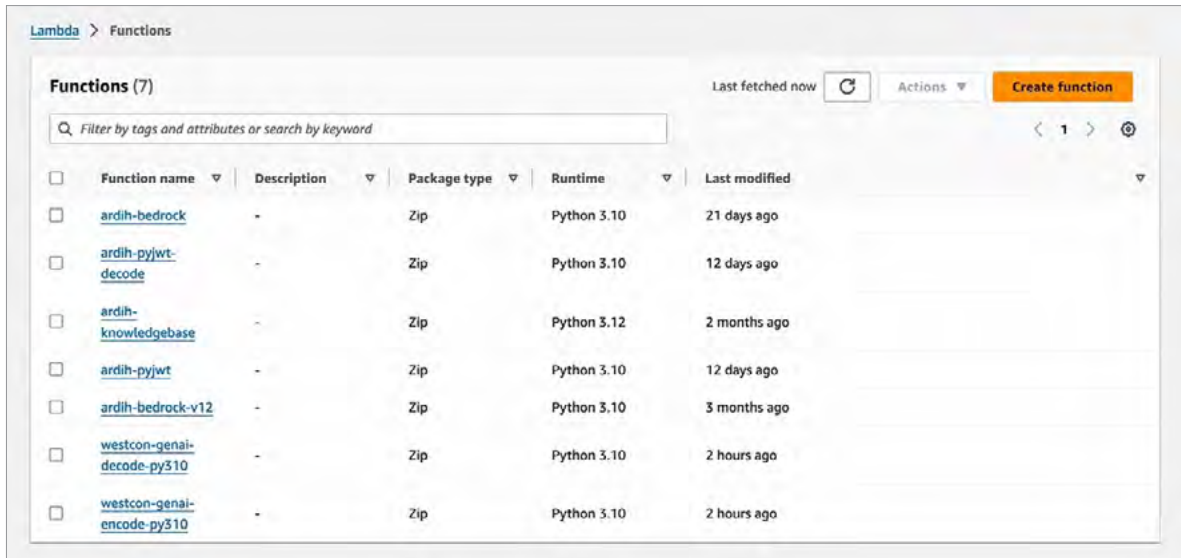
2. Choose jwt.zip



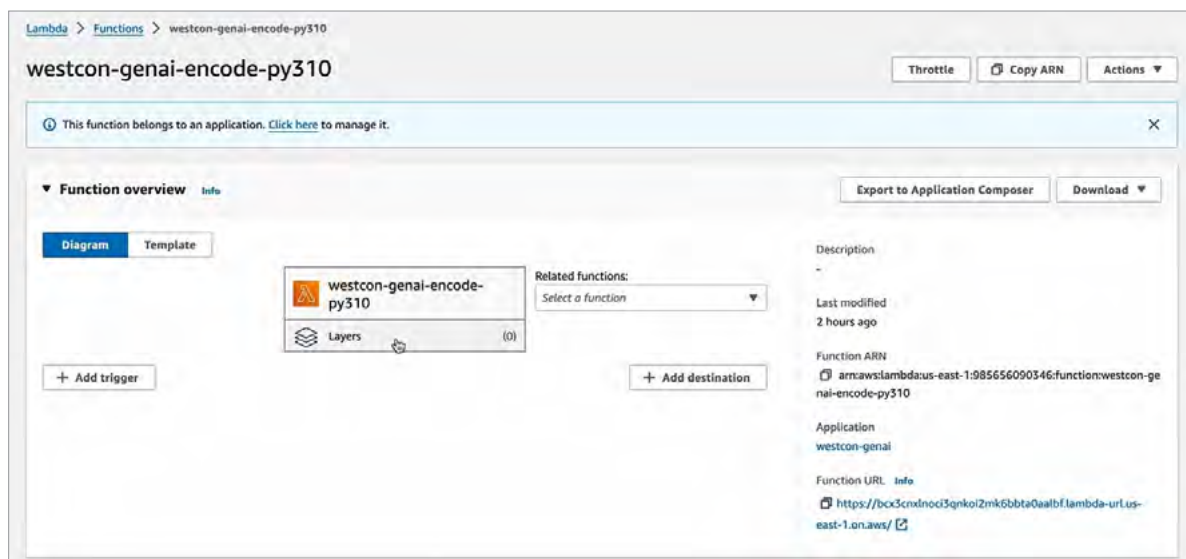
3. Choose [x86_64](#) for compatible architectures, and select all compatible runtime from Python 3.8 to Python 3.12

5.6.3. Lambda token configuration

1. Log in to the AWS Management Console and navigate to the AWS Lambda service. In the Lambda console, select “Functions” from the left navigation pane.



2. Select layers.



3. Add new layer



4. Choose custom layer. Select custom layer named **Westcon-jwt** and version 1.

Choose a layer

Layer source [Info](#)

Choose from layers with a compatible runtime and instruction set architecture or specify the Amazon Resource Name (ARN) of a layer version. You can also [create a new layer](#).

☐ **AWS layers**
Choose a layer from a list of layers provided by AWS.

☒ **Custom layers**
Choose a layer from a list of layers created by your AWS account or organization.

☐ **Specify an ARN**
Specify a layer by providing the ARN.

Custom layers

Layers created by your AWS account or organization that are compatible with your function's runtime.

westcon-jwt

▼

Version

1

▼

Cancel

Add

5.6.4. Lambda prompt configuration

1. Log in to the AWS Management Console and navigate to the AWS Lambda service. In the Lambda console, select “Functions” from the left navigation pane.

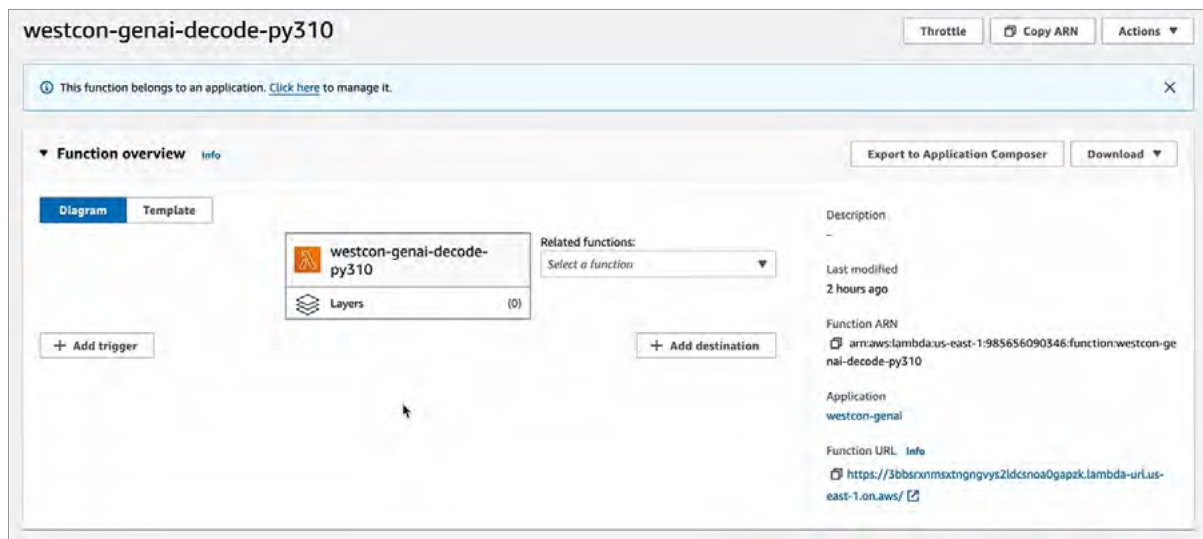
Lambda > Functions

Functions (7) Last fetched now Actions Create function

Filter by tags and attributes or search by keyword

	Function name	Description	Package type	Runtime	Last modified
☐	ardih-bedrock	-	Zip	Python 3.10	21 days ago
☐	ardih-pyjwt-decode	-	Zip	Python 3.10	12 days ago
☐	ardih-knowledgebase	-	Zip	Python 3.12	2 months ago
☐	ardih-pyjwt	-	Zip	Python 3.10	12 days ago
☐	ardih-bedrock-v12	-	Zip	Python 3.10	3 months ago
☐	westcon-genai-decode-py310	-	Zip	Python 3.10	2 hours ago
☐	westcon-genai-encode-py310	-	Zip	Python 3.10	2 hours ago

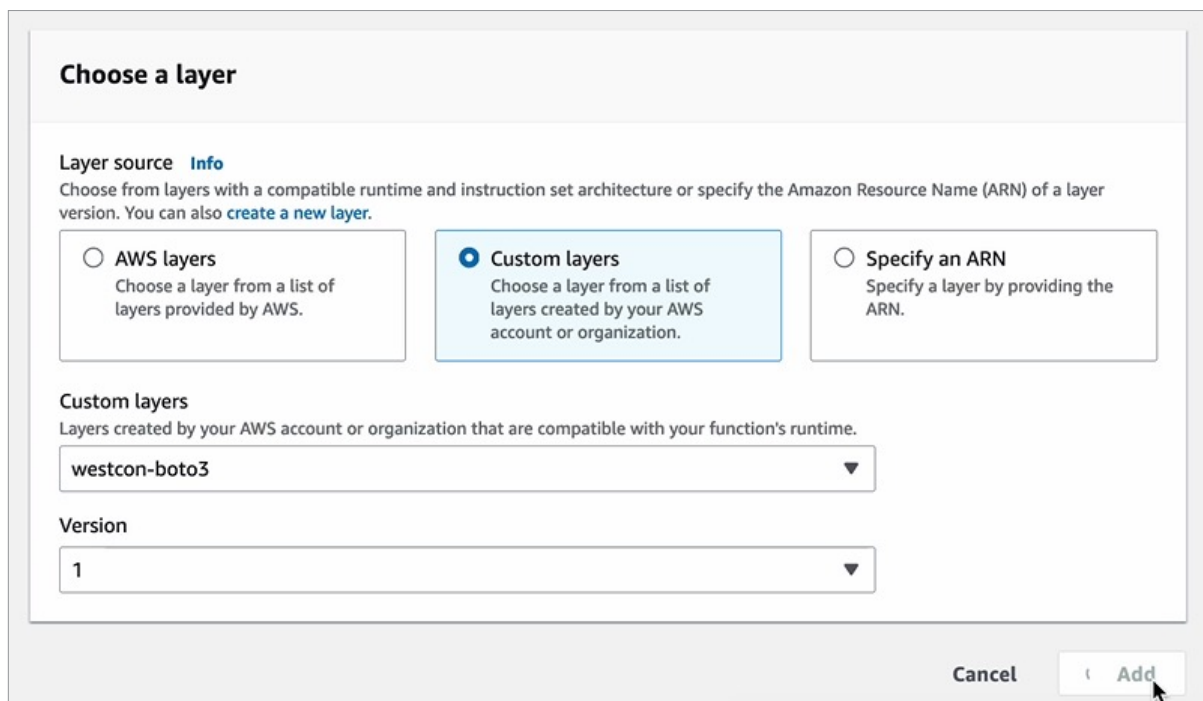
2. Select layers.



3. Add new layer



4. Add custom layer named **Westcon-boto3** version 1.



- Repeat the process for **Westcon-jwt**.

Choose a layer

Layer source [Info](#)

Choose from layers with a compatible runtime and instruction set architecture or specify the Amazon Resource Name (ARN) of a layer version. You can also [create a new layer](#).

☐ **AWS layers**
Choose a layer from a list of layers provided by AWS.

☒ **Custom layers**
Choose a layer from a list of layers created by your AWS account or organization.

☐ **Specify an ARN**
Specify a layer by providing the ARN.

Custom layers

Layers created by your AWS account or organization that are compatible with your function's runtime.

Version

Cancel
Add

- Go to Amazon Bedrock page, select the knowledge base and copy the knowledge base ID.

[Amazon Bedrock](#) > [Knowledge bases](#) > westcon-knowledge-base

Test
Delete

westcon-knowledge-base

Edit

Knowledge base overview

Knowledge base name westcon-knowledge-base	Knowledge base ID H36HFT9STO	Log Deliveries Configure log deliveries and event logs in the Edit page.
Knowledge base description —	Status Ready	
Service Role ArdihAmazonBedrockExecutionRoleForKnowledgeBase_vvf94	Created date July 19, 2024, 00:00 (UTC+07:00)	

- Back to Lambda page, click on **Configuration** tab.

Code
Test
Monitor
Configuration
Aliases
Versions

Code source [Info](#)
Upload from

File
Edit
Find
View
Go
Tools
Window
Test
Deploy

westcon-genai-dec

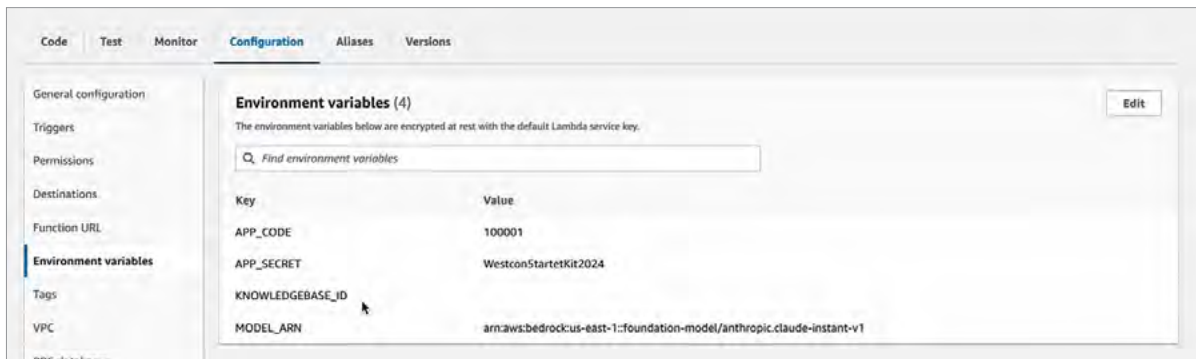
index.py

Environment Var

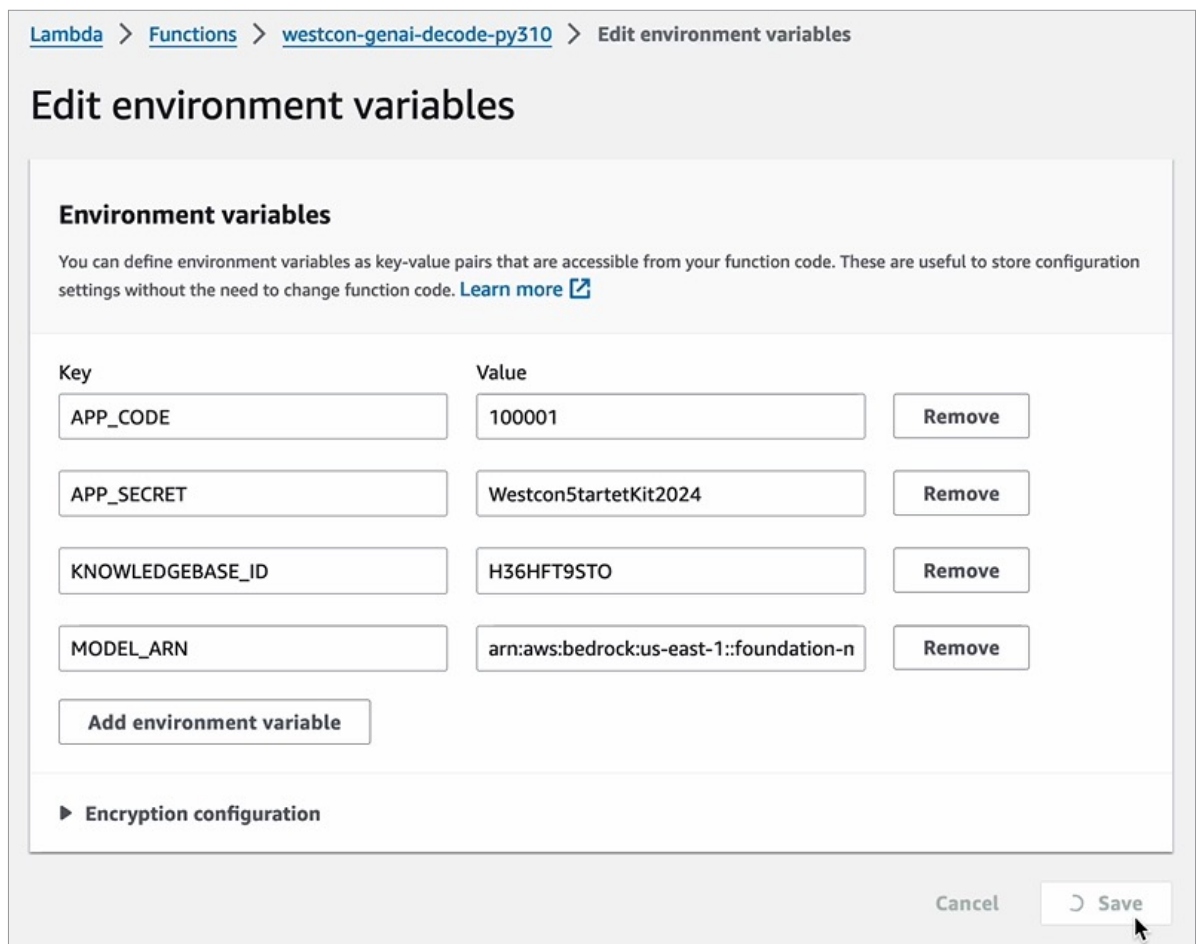
```

1 import json
2 import boto3
3 import os
4 import jwt
5 from botocore.config import Config
6
7 secret = os.environ['APP_SECRET']
8 my_config = Config(region_name = 'us-east-1')
9 client = boto3.client('bedrock-agent-runtime', config=my_config)
10
11 def jwtdecode(token):
12     try:
13         decode = jwt.decode(token, secret, algorithms=['HS256'])
14         return decode
            
```

8. Scroll down to **Environment variables**, and edit.



9. Paste the **Knowledge base ID** and hit **Save**.



6.2. Prompt test

1. Open CloudFormation page and open **Outputs** tab.

The screenshot shows the AWS CloudFormation console. On the left, a list of stacks is visible, including 'westcon-genai', 'ardih-jwt-decode', 'ardih-lambda-pyjwt', 'aws-cloud9-bedrock-frank-testing2-ea490d5559cb415485ddee93aab089f3', and 'aws-bedrock-workshop'. The 'westcon-genai' stack is selected, and the 'Outputs' tab is active. The 'Outputs' table lists the following outputs:

Key	Value	Description	Export name
LoginURL	https://1h2pxt7on8.execute-api.us-east-1.amazonaws.com/v1/login	Login API Gateway endpoint URL	-
PromptURL	https://1h2pxt7on8.execute-api.us-east-1.amazonaws.com/v1/prompt	Prompt API Gateway endpoint URL	-
PublicSubnetId2	subnet-060d154ee9b6b04b	Public Subnet ID 2	-
PublicSubnetId3	subnet-081df3ffa32e7a233	Public Subnet ID 3	-
PrivateSubnet2Id	subnet-095c54d50d1522678	Private Subnet 2 ID	-
PrivateSubnet3Id	subnet-0a37e1ecadfc32bac	Private Subnet 3 ID	-
PublicSubnetId1	subnet-0a7d3f581634851e2	Public Subnet ID 1	-
PrivateSubnet1Id	subnet-0fbc8cf1597c7a4f7	Private Subnet 1 ID	-

2. Copy **PromptURL** value
3. Copy **PromptURL** and token from previous operation

The screenshot shows a REST client interface with a POST request to the URL `https://3bbsrxnmsxtngngvys2ldcsnoa0gapzk.lambda-url.us-east-1.amazonaws.com/v1/prompt`. The request body is a JSON object:

```

{
  "headers": {
    "Authorization": "eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJ1bm90VTENPTIj9.LQ097NkMaPQ-jTBH9WQSDamDqjS7yI0BtDScP2F7pEY"
  },
  "input": "apa saja perlengkapan kendaraan bermotor?"
}

```

The response is a 200 OK status with a size of 2.29 KB and a time of 18.04 s. The response body is a JSON object:

```

{
  "message": "Token is valid",
  "payload": "100001",
  "input": "apa saja perlengkapan kendaraan bermotor?",
  "result": {
    "ResponseMetadata": {
      "RequestId": "22b995b0-3639-4c39-abee-33fd9387fac5",
      "HTTPStatusCode": 200,
      "HTTPHeaders": {
        "date": "Thu, 18 Jul 2024 17:45:43 GMT",
        "content-type": "application/json",
        "content-length": "1881",
        "connection": "keep-alive",
        "x-amzn-requestid": "22b995b0-3639-4c39-abee-33fd9387fac5"
      },
      "RetryAttempts": 0
    },
    "citations": [
      {
        "generatedResponsePart": {
          "textResponsePart": {
            "span": {
              "end": 260,
              "start": 0
            },
            "text": "Berdasarkan hasil pencarian, perlengkapan kendaraan bermotor terdiri atas susunan, perlengkapan, ukuran, karoseri, rancangan teknis kendaraan sesuai dengan peruntukannya, pemuatan, penggunaan, pengandangan kendaraan bermotor, dan penempelan kendaraan bermotor."
          }
        }
      }
    ]
  }
}

```

4. Follow this JSON structure to POST data

JSON POST structure

```

{
  "headers": {
    "Authorization": "<token>"
  },
  "input": "<your-question>"
}

```

JSON response structure

```
{
  "message": "<token-message>",
  "payload": "<YourAppCode>",
  "input": "<question>",
  "result": {
    "ResponseMetadata": {
      "RequestId": "08b8e586-9940-4e4c-88b9-b5cce47fc797",
      "HTTPStatusCode": 200,
      "HTTPHeaders": {
        "date": "Fri, 19 Jul 2024 08:17:19 GMT",
        "content-type": "application/json",
        "content-length": "3689",
        "connection": "keep-alive",
        "x-amzn-requestid": "08b8e586-9940-4e4c-88b9-b5cce47fc797"
      },
      "RetryAttempts": 0
    },
    "citations": [
      {
        "generatedResponsePart": {
          "textResponsePart": {
            "span": {
              "end": 381,
              "start": 0
            },
            "text": "<text>"
          }
        },
        "retrievedReferences": [
          {
            "content": {
              "text": "<text>"
            },
            "location": {
              "s3Location": {
                "uri": "s3://<s3-object-path>"
              },
              "type": "S3"
            }
          },
          {
            "content": {
              "text": "<text>"
            },
            "location": {
              "s3Location": {
                "uri": "s3://<s3-object-path>"
              },
              "type": "S3"
            }
          }
        ]
      }
    ],
    "output": {
      "text": "<answer>"
    },
    "sessionId": "e7bba22d-0749-4402-9b10-3b96cee90655"
  }
}
```



Have a question?

Contact us

NZ Cloud Sales: +64 9 477 7211
cloudsales.nz@westcon.com

AU Cloud Sales: +61 2 8412 1212
cloudsales.au@westcon.com

SG Cloud Sales: +65 6424 0570
cloudsales.sg@westcon.com

ID Cloud Sales: +62 21 8062 1470
cloudsales.id@westcon.com

VN Cloud Sales: +84 2836203338
cloudsales.vn@westcon.com