# Enhancement & Customizations for

## Workshop: RAG using Amazon Bedrock Agents and Knowledge Base

## Description:

The original "**RAG using Amazon Bedrock Agents and Knowledge base**" workshop guides you how to create a GenAI bedrock agent that can answer questions about Amazon SageMaker.

This document will guide you how to enhance this workshop by:

1) Use your own data as the data source for the Bedrock agent.
2) Customize your own Agent & Knowledge Base instructions (Prompts)
3) Creating a webpage that will allow users from outside of this workshop environment to experiment with your deliverables and send prompts to the agent that you created.

# Use your own data

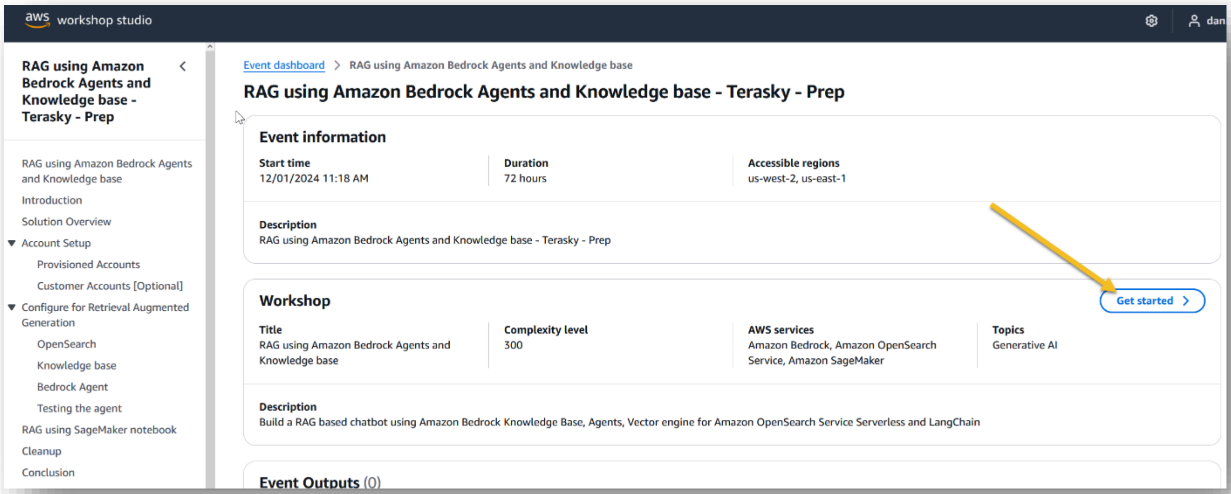
The workshop session starts with the deployment of the “**rag-w-bedrock-kb**” CloudFormation stack.
(In AWS hosted events this stack is pre-deployed before you login to the workshop account)

The **rag-w-bedrock-kb** stack is used to define resources like… IAM roles, Access Policies, S3-Bucket and more… that you get to use in the workshop tasks.

The **rag-w-bedrock-kb** stack also import SageMaker-documentation HTML files into the S3-Bucket, and these files are used as the data source for creating a vector database that the bedrock agent will use.

In order to use your own data in this workshop, we need to delete these SageMaker-documentation files and upload your files to the same S3-Bucket.

1) Once you have uploaded the relevant files to the S3-Bucket. Go to the Event dashboard page and click - **Get started >**



RAG using Amazon Bedrock Agents and Knowledge base - Terasky - Prep

RAG using Amazon Bedrock Agents and Knowledge base
Introduction
Solution Overview
Account Setup
Provisioned Accounts
Customer Accounts [Optional]
Configure for Retrieval Augmented Generation
OpenSearch
Knowledge base
Bedrock Agent
Testing the agent
RAG using SageMaker notebook
Cleanup
Conclusion

Event dashboard > RAG using Amazon Bedrock Agents and Knowledge base

RAG using Amazon Bedrock Agents and Knowledge base - Terasky - Prep

Event information

| Start time | Duration | Accessible regions |
|---|---|---|
| 12/01/2024 11:18 AM | 72 hours | us-west-2, us-east-1 |

Description
RAG using Amazon Bedrock Agents and Knowledge base - Terasky - Prep

Workshop

Get started >

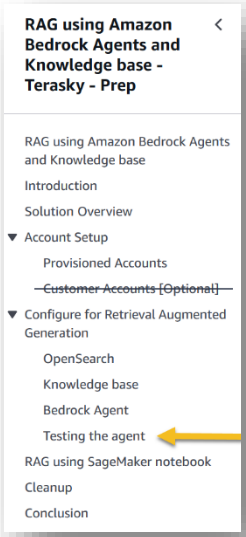
| Title | Complexity level | AWS services | Topics |
|---|---|---|---|
| RAG using Amazon Bedrock Agents and Knowledge base | 300 | Amazon Bedrock, Amazon OpenSearch Service, Amazon SageMaker | Generative AI |

Description
Build a RAG based chatbot using Amazon Bedrock Knowledge Base, Agents, Vector engine for Amazon OpenSearch Service Serverless and LangChain

Event Outputs (0)

2) Follow the workshop lab instructions till you complete the **Testing the agent** phase.

**Customer Accounts [Optional]** phase is not required.



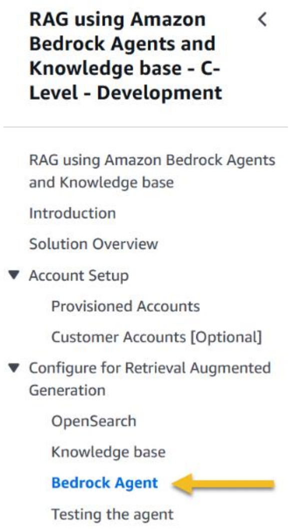
RAG using Amazon Bedrock Agents and Knowledge base - Terasky - Prep

RAG using Amazon Bedrock Agents and Knowledge base
Introduction
Solution Overview
Account Setup
Provisioned Accounts
~~Customer Accounts [Optional]~~
Configure for Retrieval Augmented Generation
OpenSearch
Knowledge base
Bedrock Agent
Testing the agent
RAG using SageMaker notebook
Cleanup
Conclusion

**Important:**
In these workshop steps you will be asked to define resources with specific names like "**sagemaker-qa**" (Bedrock agent) or "**sagemaker-docs**" (Bedrock Knowledge Base).

Please make sure that you use the exact names because the initial permissions rules that were defined in this environment are using these names.

# Define your own agent & knowledge Base instructions


RAG using Amazon Bedrock Agents and Knowledge base - C-Level - Development
RAG using Amazon Bedrock Agents and Knowledge base
Introduction
Solution Overview
Account Setup
Provisioned Accounts
Customer Accounts [Optional]
Configure for Retrieval Augmented Generation
OpenSearch
Knowledge base
Bedrock Agent
Testing the agent

On step **Bedrock Agent** tasks 2 & 4 – you are asked to define
**Instructions for the Agent** & **Knowledge base instructions for agent**

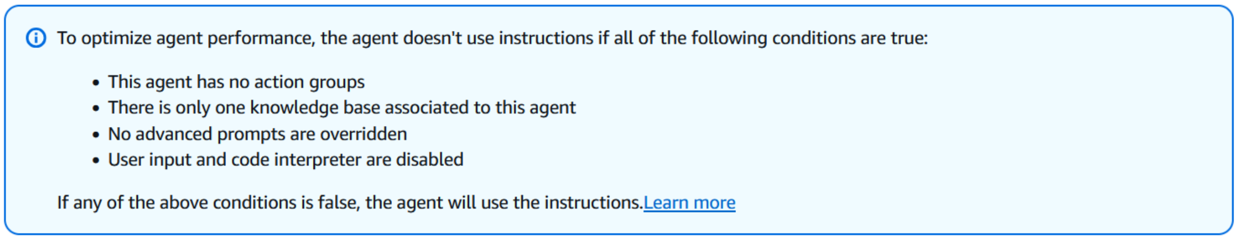
The suggested prompts are for a Q&A agent that answer to SageMaker related questions.

This is the time to get creative –

- What prompt will you use?
- What LLM works best for you?
- How can you make your prompt helpful for users?
- Will the prompt hold any output examples?

**Troubleshooting: If your prompt is not working like it should…**

- Are you getting this info message?



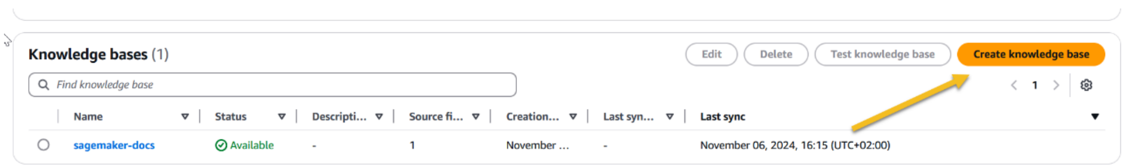
To optimize agent performance, the agent doesn't use instructions if all of the following conditions are true:

- This agent has no action groups
- There is only one knowledge base associated to this agent
- No advanced prompts are overridden
- User input and code interpreter are disabled

If any of the above conditions is false, the agent will use the instructions.Learn more

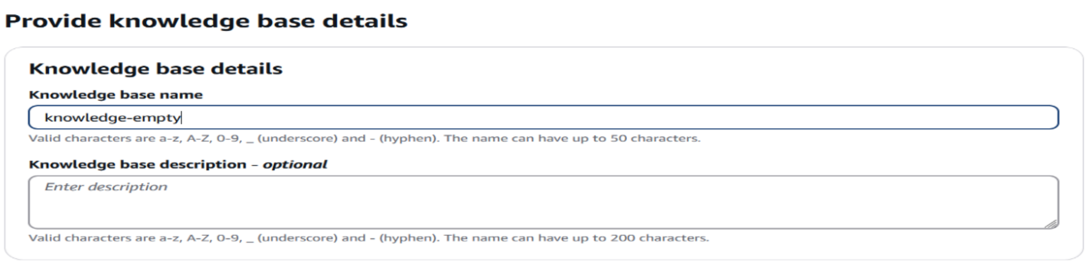
In order to overcome this limitation let's add a second knowledge base to our agent.

- On the bedrock page, open the Knowledge base pane
- Click **Create Knowledge base**



| Name | Status | Descripti... | Source fi... | Creation... | Last syn... | Last sync |
|---|---|---|---|---|---|---|
| sagemaker-docs | Available | - | 1 | November ... | - | November 06, 2024, 16:15 (UTC+02:00) |

- Name the knowledge base "knowledge-empty" and click **Next**



**Provide knowledge base details**

**Knowledge base details**

**Knowledge base name**

knowledge-empty

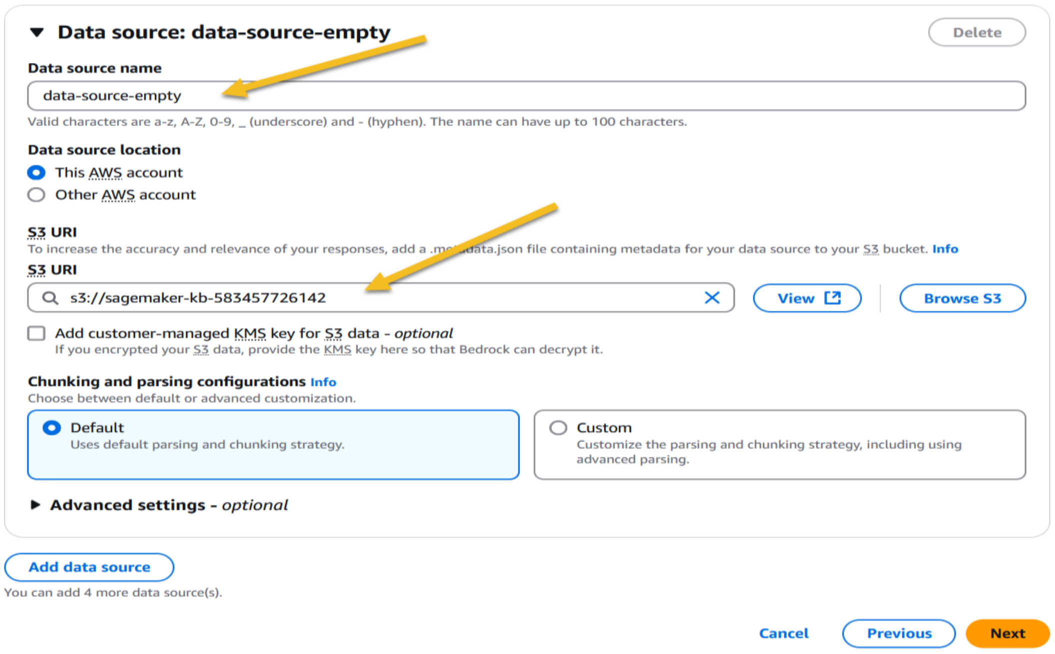
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 50 characters.

**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.

- Name the Data source “data-source-empty”, Select the S3-Bucket and click **Next**



**Data source: data-source-empty**

Delete

**Data source name**

data-source-empty

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

**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

**S3 URI**

s3://sagemaker-kb-583457726142

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.

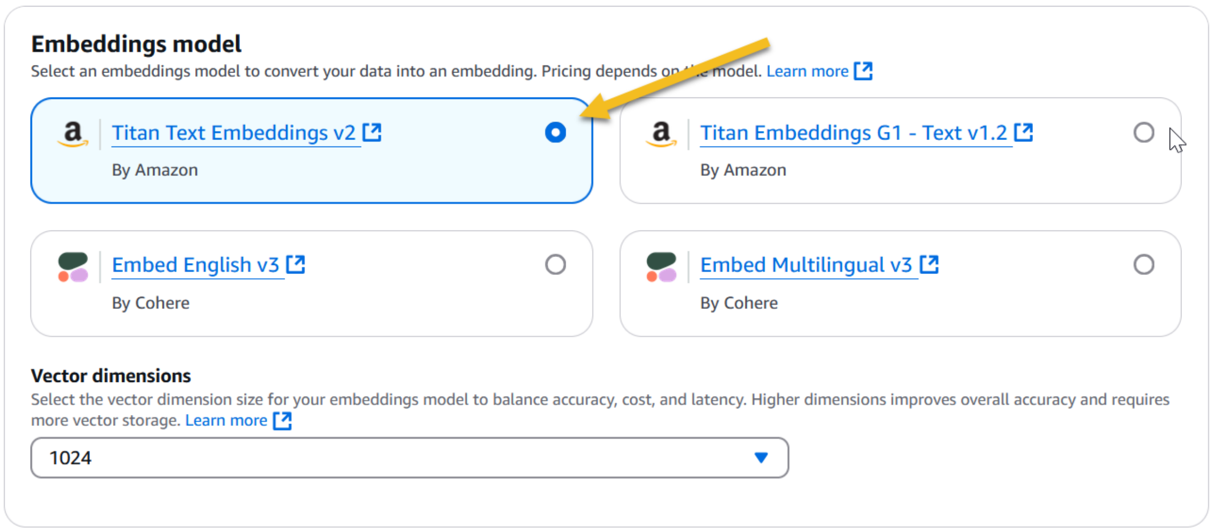
[✓ Default] Uses default parsing and chunking strategy.
Custom — Customize the parsing and chunking strategy, including using advanced parsing.

**Advanced settings** - *optional*

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

Cancel | Previous | Next

- Select the **Titan Text Embeddings V2** model and click **Next**.
  (Without setting the vector DB)



**Embeddings model**

Select an embeddings model to convert your data into an embedding. Pricing depends on the model. Learn more

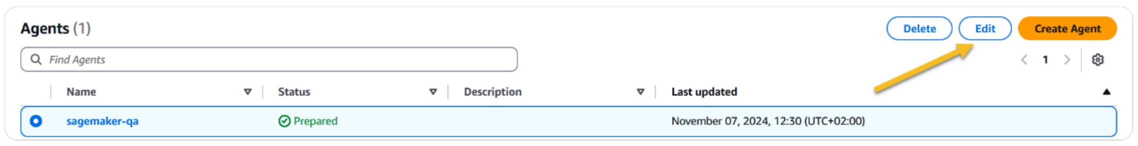
[✓ Titan Text Embeddings v2 — By Amazon]
Titan Embeddings G1 - Text v1.2 — By Amazon
Embed English v3 — By Cohere
Embed Multilingual v3 — By Cohere

**Vector dimensions**

Select the vector dimension size for your embeddings model to balance accuracy, cost, and latency. Higher dimensions improves overall accuracy and requires more vector storage. Learn more

1024

- Click **Create Knowledge base** and wait for the creation to be completed
- Navigate back to the **Bedrock** → **Agents** pane
- Select the **sagemaker-qa** agent and click **Edit**



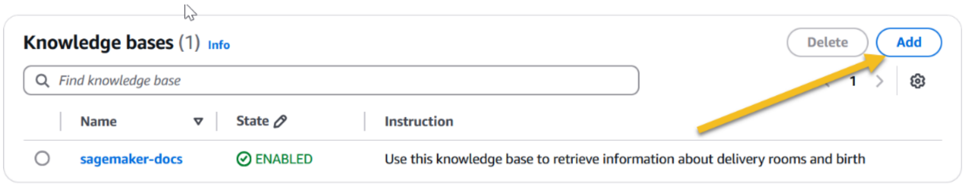
**Agents** (1)

Delete | Edit | Create Agent

Find Agents

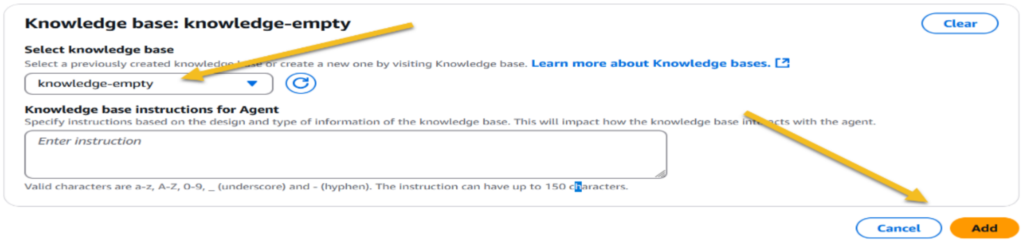
| | Name | Status | Description | Last updated |
|---|---|---|---|---|
| ● | sagemaker-qa | ✓ Prepared | | November 07, 2024, 12:30 (UTC+02:00) |

- Under the **Knowledge base** pane click **add**



| Name | State | Instruction |
|---|---|---|
| sagemaker-docs | ENABLED | Use this knowledge base to retrieve information about delivery rooms and birth |

- Select the empty Knowledge base that you created and click **Add**



Knowledge base: knowledge-empty

Select knowledge base

Select a previously created knowledge base or create a new one by visiting Knowledge base. Learn more about Knowledge bases.

knowledge-empty

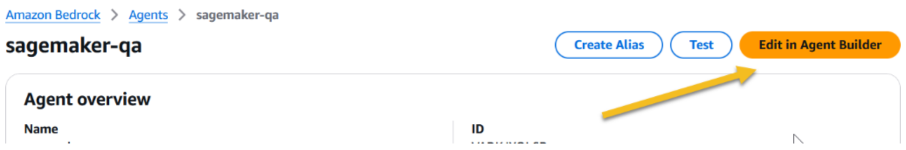
Knowledge base instructions for Agent

Specify instructions based on the design and type of information of the knowledge base. This will impact how the knowledge base interacts with the agent.

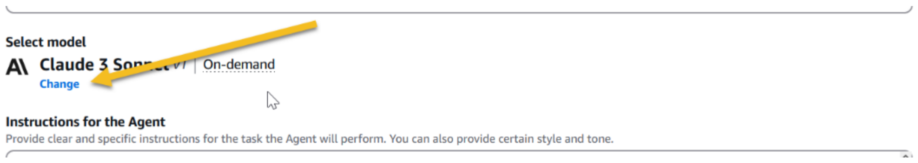
Enter instruction

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The instruction can have up to 150 characters.

- Click **Save and Exit** in the Agent Builder page
- Click **Prepare**
- Test your agent…
- Still not getting the reply that you are expecting…
  - Click **Edit in Agent Builder**



  - Click **change** and select a different model



Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

  - **Save** the agent, **Prepare** the agent and test it
  - Experiment with different models till you get the required results

## Prompt example:

- **Instructions for the Agent:**
  You are a medical assistant by the name Daniela.
  Your task is to answer patient's questions based on the data in the knowledge base only.

  The answer should be written in simple language in an E-mail format with new lines between sentences to format a well-structured email.

  The context of the questions is always around birth and delivery rooms.

  Example email reply:
  Dear patient,

  It is recommended that the expectant mother eat snacks that she enjoys during childbirth,
  as the labor process can be long and the mother may feel hungry.
  However, heavy meals are not recommended, especially if the mother is receiving an epidural,
  as she can only drink water during that time.

  I recommend packing light snacks in the delivery room bag to help sustain the mother's energy during labor.

  Thank you and see you at the delivery room,
  Daniela

- **Knowledge base instructions for agent:**
  Your task is to generate email based on the patient's query. Answers questions that are only based on the knowledge base and rephrase the answer in simple English.

## Congratulations!

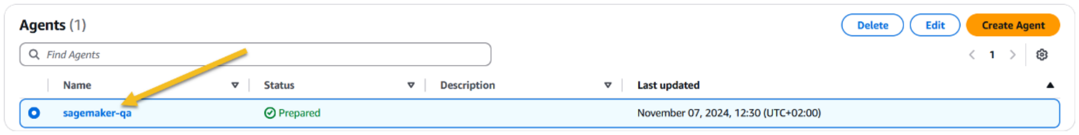
You have created a GenAI bedrock agent with an OpenSearch vector Database that hosts your embedded vectorized data as a RAG.
Now…
Let's expose your agent to the world

# Expose your Agent to the world

This section will instruct you on how to deploy a CloudFormation template that will create a webpage that will allow external users to send questions to & get-replies from your Bedrock Agent from outside the workshop environment.

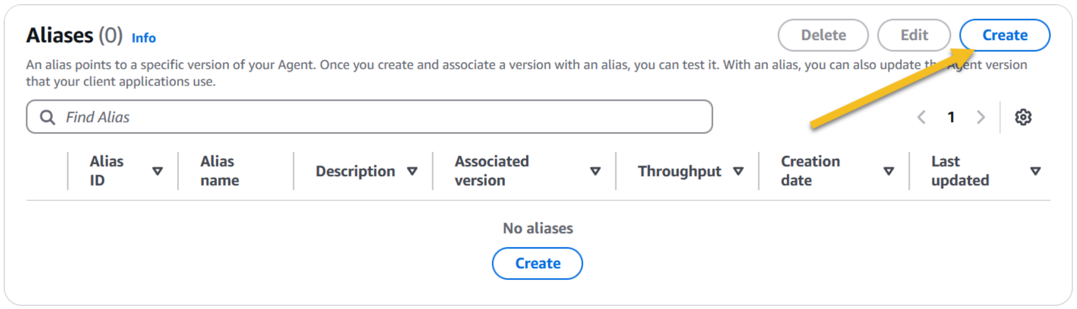
In order to access our Bedrock agent with a lambda function, we have to define an Agent Alias.

1) Navigate to the **Bedrock → Agents** pane
2) Click on the **sagemaker-qa** agent



| Name | Status | Description | Last updated |
|---|---|---|---|
| sagemaker-qa | Prepared | | November 07, 2024, 12:30 (UTC+02:00) |

3) Scroll down to the Aliases pane and click **Create**



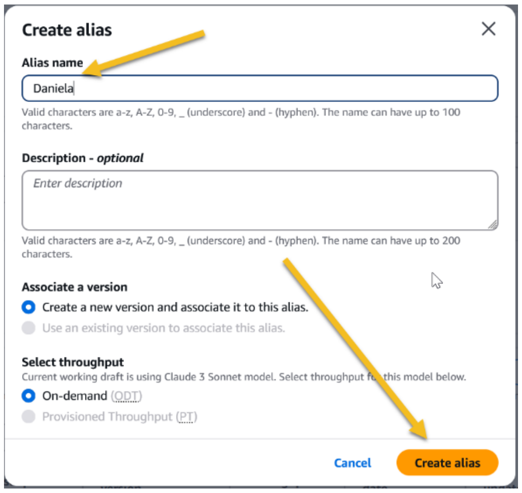
An alias points to a specific version of your Agent. Once you create and associate a version with an alias, you can test it. With an alias, you can also update the Agent version that your client applications use.

| Alias ID | Alias name | Description | Associated version | Throughput | Creation date | Last updated |
|---|---|---|---|---|---|---|

No aliases

Create

4) Type an **Alias name** and click **Create Alias**



Create alias

Alias name

Daniela

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

Description - *optional*

Enter description

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

Associate a version

[✓ Create a new version and associate it to this alias.]

Use an existing version to associate this alias.

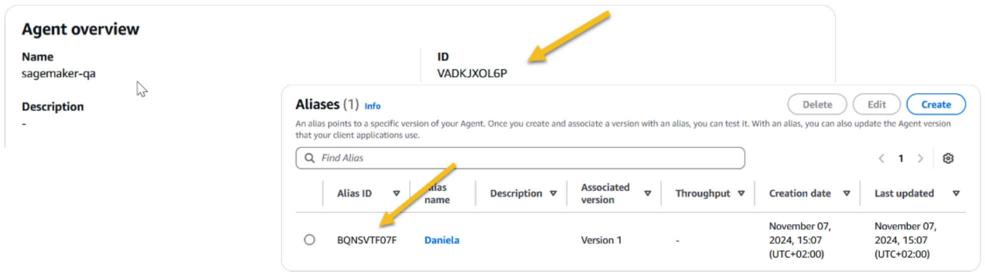
Select throughput

Current working draft is using Claude 3 Sonnet model. Select throughput for this model below.

[✓ On-demand (ODT)]

Provisioned Throughput (PT)

5) From the agent page copy the **Alias ID** & the **Agent ID**



Agent overview

Name
sagemaker-qa

Description
-

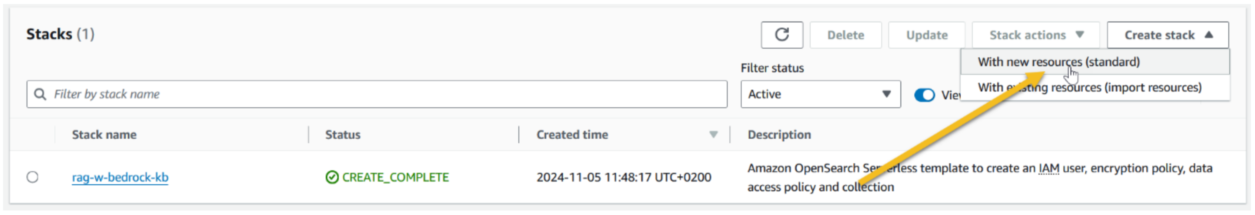
ID
VADKJXOL6P

An alias points to a specific version of your Agent. Once you create and associate a version with an alias, you can test it. With an alias, you can also update the Agent version that your client applications use.

| Alias ID | Alias name | Description | Associated version | Throughput | Creation date | Last updated |
|---|---|---|---|---|---|---|
| BQNSVTF07F | Daniela | | Version 1 | - | November 07, 2024, 15:07 (UTC+02:00) | November 07, 2024, 15:07 (UTC+02:00) |

## Deploy the Agent-Frontend CloudFormation

1) Navigate to the CloudFormation page and click **Create Stack → With new resources**



Stacks (1)

Delete | Update | Stack actions | Create stack

With new resources (standard)

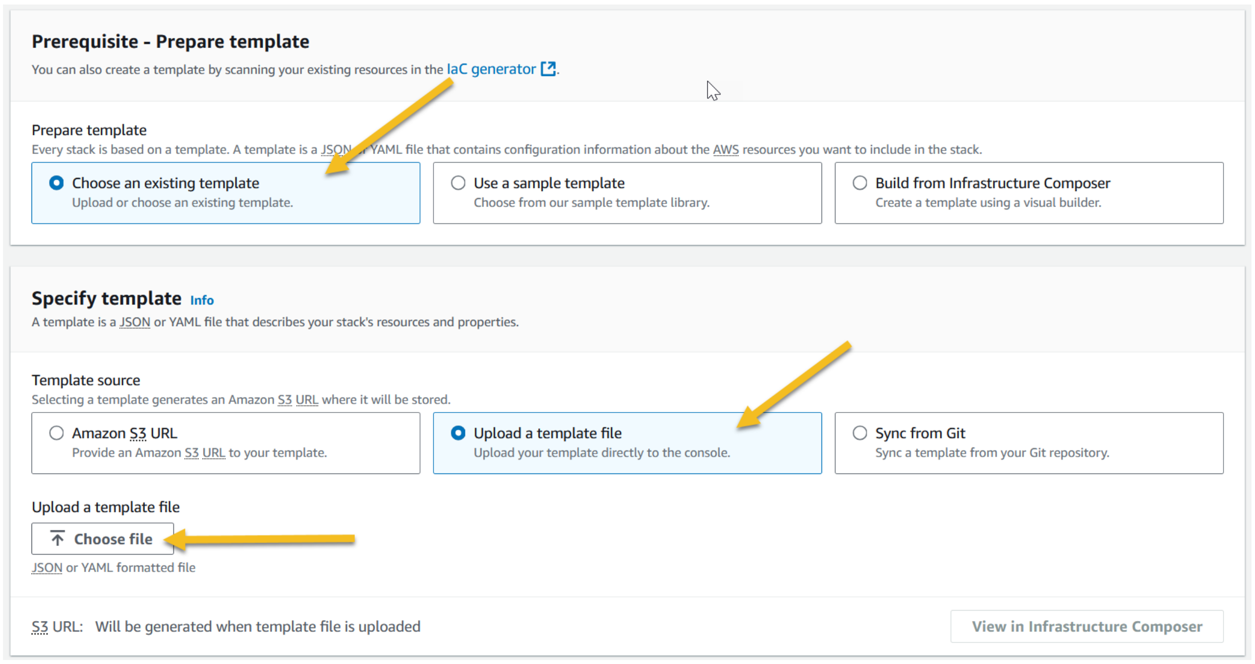
With existing resources (import resources)

Filter by stack name

Filter status: Active

| Stack name | Status | Created time | Description |
|---|---|---|---|
| rag-w-bedrock-kb | CREATE_COMPLETE | 2024-11-05 11:48:17 UTC+0200 | Amazon OpenSearch Serverless template to create an IAM user, encryption policy, data access policy and collection |

2) Select **Choose an existing template** & **upload a template file**, and click Choose file



**Prerequisite - Prepare template**

You can also create a template by scanning your existing resources in the IaC generator.

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 Infrastructure Composer — Create a template using a visual builder.

**Specify template** Info

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 — Sync a template from your Git repository.

Upload a template file

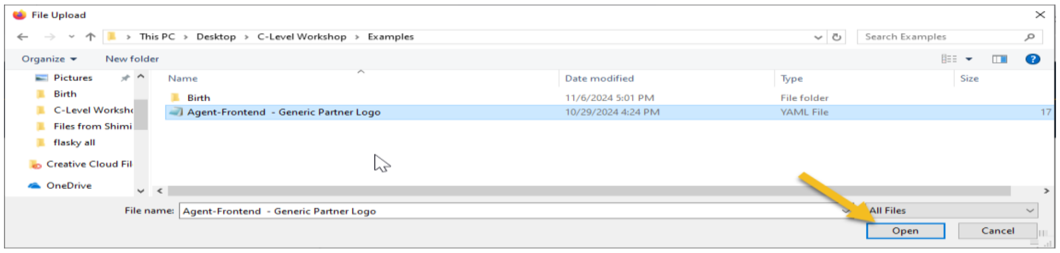
Choose file

JSON or YAML formatted file

S3 URL: Will be generated when template file is uploaded

View in Infrastructure Composer

3) Select the stack file that you got from the workshop facilitator and click open



File Upload

This PC › Desktop › C-Level Workshop › Examples

| Name | Date modified | Type | Size |
|---|---|---|---|
| Birth | 11/6/2024 5:01 PM | File folder | |
| Agent-Frontend - Generic Partner Logo | 10/29/2024 4:24 PM | YAML File | 17 |

File name: Agent-Frontend - Generic Partner Logo

All Files

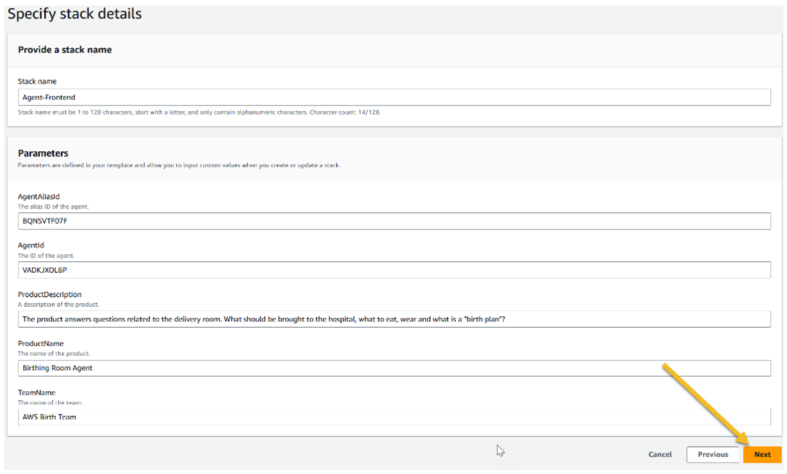
Open | Cancel

4) Click **Next**

5) Fill the required information

|  |  |  |
|---|---|---|
| a. | **Stack name**: | Agent-Frontend |
| b. | **AgentAliasId**: | <The bedrock agent alias ID that we saved before> |
| c. | **AgentId**: | <The bedrock agent ID that we saved before> |
| d. | **ProductDescription**: | <A description of the product> |
| e. | **ProductName**: | <The name of the product> |
| f. | **TeamName**: | <Your Name/s> |

And click **Next**



Specify stack details

Provide a stack name

Stack name: Agent-Frontend

Stack name must be 1 to 128 characters, start with a letter, and only contain alphanumeric characters. Character count: 14/128.

Parameters

Parameters are defined in your template and allow you to input custom values when you create or update a stack.

AgentAliasId
The alias ID of the agent.
BQNSVTF07F

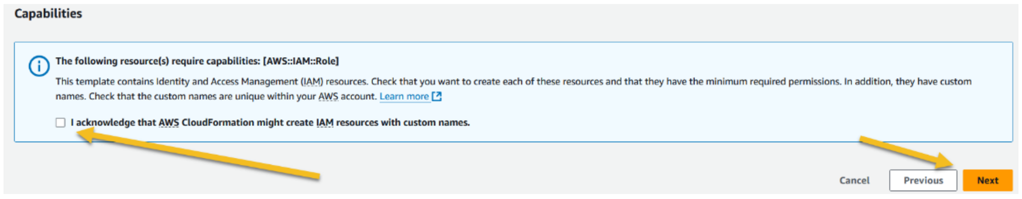
AgentId
The ID of the agent.
VADKJXOL6P

ProductDescription
A description of the product.
The product answers questions related to the delivery room. What should be brought to the hospital, what to eat, wear and what is a "birth plan"?

ProductName
The name of the product.
Birthing Room Agent

TeamName
The name of the team.
AWS Birth Team

6) Acknowledge and click **Next**



Capabilities

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

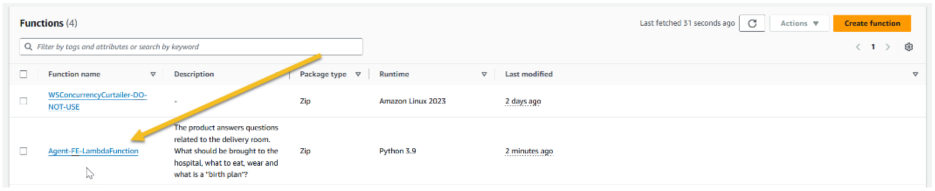
This template contains Identity and Access Management (IAM) resources. Check that you want to create each of these resources and that they have the minimum required permissions. In addition, they have custom names. Check that the custom names are unique within your AWS account. Learn more

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

7) Review and click **Submit**

8) Wait 2-3 minutes till all the resources get created and navigate to the Lambda page

9) In the Lambda page, click on the **Agent-FE-LambdaFunction** function

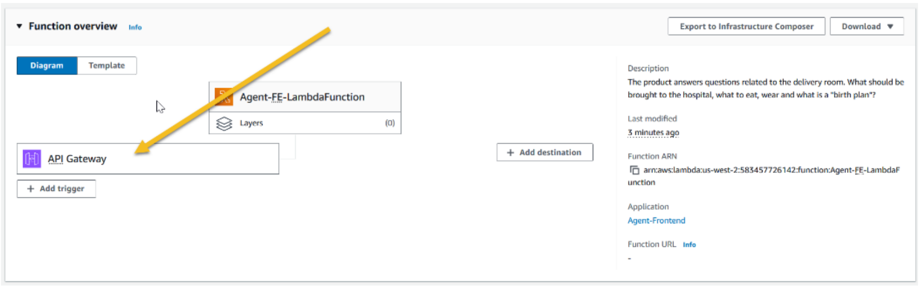


Functions (4)

Last fetched 31 seconds ago

| Function name | Description | Package type | Runtime | Last modified |
|---|---|---|---|---|
| WSConcurrencyCurtailer-DO-NOT-USE | - | Zip | Amazon Linux 2023 | 2 days ago |
| Agent-FE-LambdaFunction | The product answers questions related to the delivery room. What should be brought to the hospital, what to eat, wear and what is a "birth plan"? | Zip | Python 3.9 | 2 minutes ago |

10) Click on the API Gateway component



Description
The product answers questions related to the delivery room. What should be brought to the hospital, what to eat, wear and what is a "birth plan"?

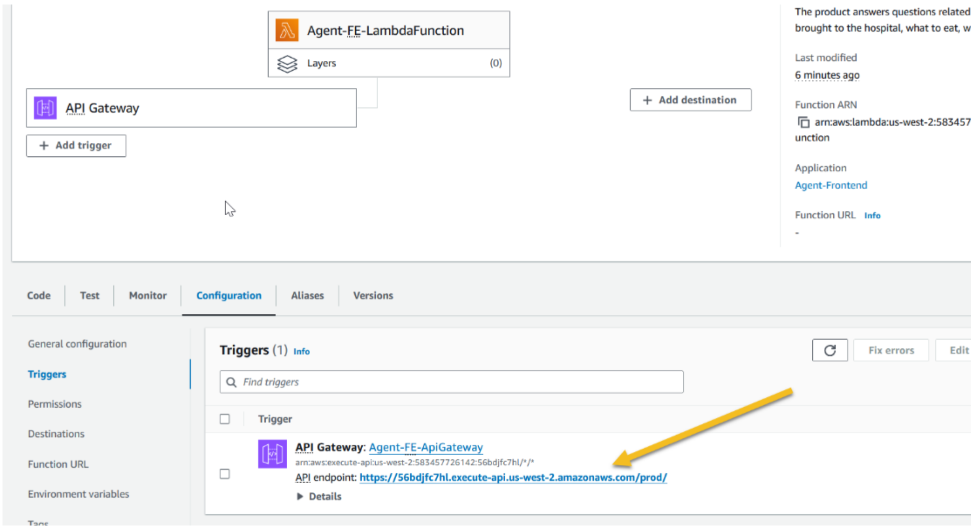
Last modified
3 minutes ago

Function ARN
arn:aws:lambda:us-west-2:583457726142:function:Agent-FE-LambdaFunction

Application
Agent-Frontend

Function URL Info
-

11) Under configuration click on the API endpoint link and access your agent via the internet.



The product answers questions related
brought to the hospital, what to eat, w

Last modified
6 minutes ago

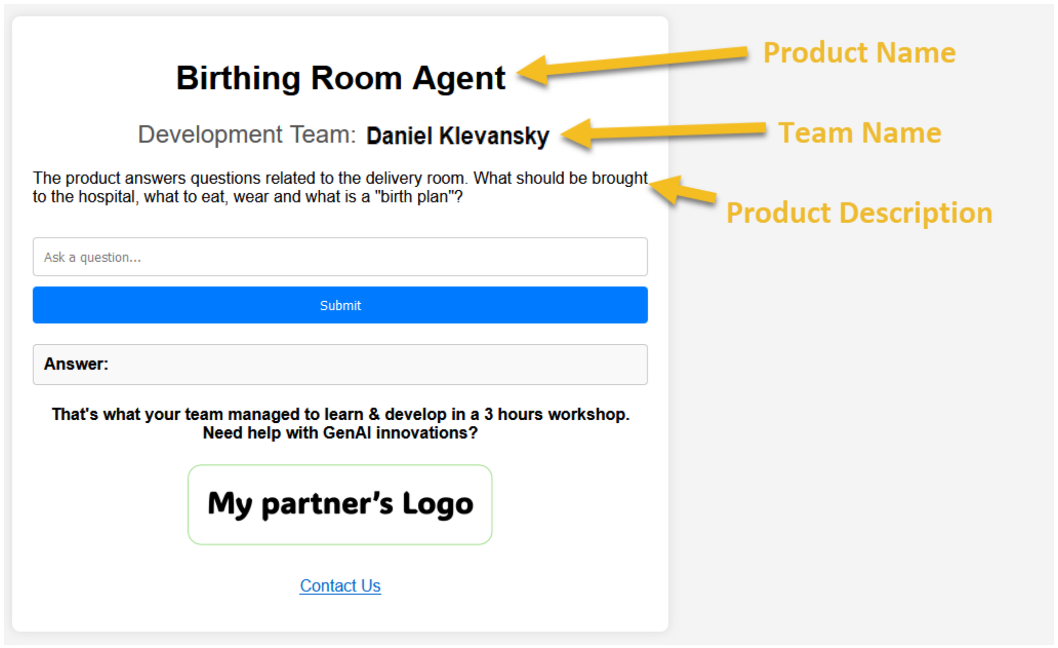
Function ARN
arn:aws:lambda:us-west-2:583457
unction

Application
Agent-Frontend

Function URL Info
-

API Gateway: Agent-FE-ApiGateway
arn:aws:execute-api:us-west-2:583457726142:56bdjfc7hl/*/*
API endpoint: https://56bdjfc7hl.execute-api.us-west-2.amazonaws.com/prod/

12) You should see a web page that looks like this…



**Birthing Room Agent**

Development Team: **Daniel Klevansky**

The product answers questions related to the delivery room. What should be brought to the hospital, what to eat, wear and what is a "birth plan"?

**Answer:**

**That's what your team managed to learn & develop in a 3 hours workshop. Need help with GenAI innovations?**

13) Access this link from your phone & enjoy your first Bedrock Agent