



RAG를 위한 Knowledge Base 설정

KB (Knowledge Base) for Amazon Bedrock 설정 방법 가이드입니다.
이제 Bedrock 콘솔에서 RAG 작업을 쉽게 진행하세요!

19 July 2024

김제삼 (Jesam Kim)

Solutions Architect
AWS

1. KB 생성

Amazon Bedrock 콘솔에 접속 합니다.

(1) Knowledge bases 메뉴에 접속 합니다.

(2) Create knowledge base 를 클릭하여 새로운 KB를 생성 합니다.

* KB를 만드는 것은 OpenSearch Serverless 컬렉션을 하나 만드는 것 입니다.
여기서 인덱스 구성 정보(칭킹 등) 설정이 포함 됩니다.

즉, KB를 만드는 것은 하나의 Index를 만드는 것과 같습니다.

Amazon Bedrock <

Amazon Bedrock > Knowledge bases

Knowledge bases Chat with your document

Knowledge bases

▼ **How it works**

Upload and chat

Quickly query foundation models with context provided by ad-hoc dataset.

[Chat with your document](#)

Create a knowledge base

To create a knowledge base, specify the location of your data, select an embedding model, and configure a vector store for Bedrock to store and update your embeddings.

Test the knowledge base

Query your knowledge base in the test window. You can get source text chunks, or you can use the chunks to get responses from a foundation model.

Use the knowledge base

Integrate your knowledge base into your application as is or add it to agents.

Knowledge bases (3) Edit Delete Test knowledge **2 Create knowledge base**

Find knowledge base

	Name	Status	Description	Source files	Creation time	Last sync warnings	Last sync
<input type="radio"/>	240718-test	✓ Ready	-	1	July 18, 2024, 13:14 (UTC+09...	-	July 18, 2024, 13:24 (UTC+...
<input type="radio"/>	knowledge-base-quick-star...	✓ Ready	-	1	February 26, 2024, 21:26 (UT...	-	February 26, 2024, 21:26 (...)
<input type="radio"/>	knowledge-base-quick-star...	✓ Ready	-	1	February 26, 2024, 17:27 (UT...	-	February 26, 2024, 17:28 (...)

1 Knowledge bases

Agents

Prompt management [Preview](#)

Prompt flows [Preview](#)

2. KB 세부사항

(1) Knowledge bases name 지정

(2) IAM 롤 생성

(이미 만든 롤이 있다면 기존 롤 선택 가능)

(3) Data source 는 S3 선택

(4) Next 클릭

Step 1 **Provide knowledge base details**
Step 2 Configure 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
ITB-ES_Contracts
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
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

Service role name
AmazonBedrockExecutionRoleForKnowledgeBase_240719

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.

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.

No tags associated with the resource.

[Add new tag](#)
You can add up to 50 more tags.

4
[Cancel](#) [Next](#)

3. Data source 구성 (1/2)

- (1) Data source name 지정
- (2) 문서가 업로드 된 S3 URI 선택
(Parsing Strategy를 사용하는 경우,
S3 URI 당 처리 가능 문서 **최대 100개**)
- (3) 청킹 부분에서 Custom 선택
- (4) Parsing strategy 사용으로 체크
- (5) 문서 Parsing 부분에 사용할 모델 선택
→ Claude 3 Sonnet
- (6) 선택사항 : Parsing strategy에 대한
프롬프트 수정 및 Chunking strategy는
바꿔볼 수 있음 (문서 마지막 Appendix 참조)
- (7) 선택사항 : 추가 데이터 소스가 있다면 Add
data source 선택

Amazon Bedrock > Knowledge bases > Create knowledge base

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

Configure data source

Configure for the chosen data source

Amazon S3 Info
Provide details to connect Amazon Bedrock to your S3 data source.

Data source: es_itb_1 [Delete]

1 **Data source name**
es_itb_1
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

2 **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://240719-jesamkim-bucket/ES_Contracts/1_UHP/ [X] [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.

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

4 **Parsing strategy**
Parsing analyses and extracts useful information from documents.
☒ Use foundation model for parsing [See supported formats](#)
Suitable for parsing more than standard text in supported document formats, including tables within PDFs with their structure intact. [View pricing](#)

Choose foundation model for parsing

5 AI Claude 3 Sonnet v1 [By Anthropic](#) [Refresh]

AI Claude 3 Haiku v1 [By Anthropic](#) [Radio]

6 **Instructions for the parser - optional**

Chunking strategy
Chunking breaks down the text into smaller segments before embedding. The chunking strategy can't be modified after you create the data source.
Default chunking
Automatically splits text into chunks of about 300 tokens in size, by default. If a document is less than or already 300 tokens, it's not split any further. [Dropdown]

Select Lambda function
Select an existing Lambda function to customize chunking and document metadata processing. Visit [AWS Lambda](#) to create a new function. Select the refresh button after creating your function.
[Select a Lambda function]

Function version
[Select a version] [View] [Refresh]

7 **Advanced settings - optional**
[Add data source]
You can add 4 more data source(s).

Cancel [Previous] Next



3. Data source 구성 (2/2)

(8) 선택사항 : 추가 Data source 가 있다면 앞의
과정처럼 S3 URI를 지정하고 Chunking 구성
등은 동일하게 설정 합니다.

(9) Next 버튼 클릭



8

Select Lambda function

Select an existing Lambda function to customize chunking and document metadata processing. Visit [AWS Lambda](#) to create a new function. Select the refresh button after creating your function.

Select a Lambda function

Function version

Select a version View Refresh

► Advanced settings - optional

Amazon S3

Provide details to connect Amazon Bedrock to your S3 data source.

Data source: es_itb_3

Delete

Data source name

es_itb_3

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://240719-jesamkim-bucket/ES_Contracts/UAE_F3/

X

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

Choose between default or advanced customization.

☐ Default

Uses default parsing and chunking strategy.

☒ Custom

Customize the parsing and chunking strategy, including using advanced parsing.

Parsing strategy

Parsing analyses and extracts useful information from documents.

☒ Use foundation model for parsing [See supported formats](#)

Suitable for parsing more than standard text in supported document formats, including tables within PDFs with their structure intact. [View pricing](#)

Choose foundation model for parsing

AI

Claude 3 Sonnet v1

By Anthropic

AI

Claude 3 Haiku v1

By Anthropic

► Instructions for the parser - optional

Chunking strategy

Chunking breaks down the text into smaller segments before embedding. The chunking strategy can't be modified after you create the data source.

Default chunking

Automatically splits text into chunks of about 300 tokens in size, by default. If a document is less than or already 300 tokens, it's not split any further.

Select Lambda function

Select an existing Lambda function to customize chunking and document metadata processing. Visit [AWS Lambda](#) to create a new function. Select the refresh button after creating your function.

Select a Lambda function

Function version

Select a version View Refresh

► Advanced settings - optional

Add data source

You can add 2 more data source(s).

9

Cancel

Previous

Next

4. 임베딩 모델 선택

(1) 임베딩 모델 선택

→ 여기서는 Titan Emb v2 선택

(2) Vector database 부분에서는

Quick create ~~ 를 선택 합니다.

(3) Next 클릭

Amazon Bedrock > Knowledge bases > Create knowledge base

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

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](#)

1 ☒ **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

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.

2 ☒ **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.

☐ **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.

3

5. 리뷰 & 생성

(1) 앞서 선택한 구성이 제대로 되었는지
확인하고 문제가 없으면 Create knowledge
base 버튼을 클릭 합니다.

Amazon Bedrock > Knowledge bases > Create knowledge base

Step 1: Provide knowledge base details
Step 2: Configure data source
Step 3: Select embeddings model and configure vector store
Step 4: Review and create

Review and create

Step 1: Provide details [Edit](#)

Knowledge base details		
Knowledge base name ITB-ES_Contracts	Knowledge base description —	Service role AmazonBedrockExecutionRoleForKnowledg eBase_240719

Tags (0)

Key	Value
No tags to display	

Step 2: Setup up data source [Edit](#)

Data source: es_itb_1		
Data source name es_itb_1	Account ID 376278017302 (this account)	S3 URI s3://240719-jesamkim-bucket/ ES_Contracts/1_UHP/
Customer-managed KMS Key for S3 —	KMS key for transient data storage —	Chunking strategy —

Data deletion policy Delete		
--------------------------------	--	--

Data source: es_itb_3		
Data source name es_itb_3	Account ID 376278017302 (this account)	S3 URI s3://240719-jesamkim-bucket/ ES_Contracts/UAE_F3/
Customer-managed KMS Key for S3 —	KMS key for transient data storage —	Chunking strategy Default
Parsing strategy Claude 3 Sonnet v1 (Bedrock model parsing)	Lambda function —	S3 bucket for Lambda function —
Data deletion policy Delete		

Step 3: Select embeddings model and configure vector store [Edit](#)

Embeddings model	
Model Titan Text Embeddings v2	Vector dimensions 1024

Vector store
Quick create vector store - <i>Recommended</i> We will create an Amazon OpenSearch Serverless vector store in your account on your behalf.

[Cancel](#) [Previous](#) [Create knowledge base](#)

6. Data Source Sync (인덱싱) (1/2)

(1) 앞서 추가한 data source 항목이 보입니다.

(문서 갯수가 많아서 3개로 나누어서 등록했습니다;

Data source 1개 당 문서 최대 100개)

(2) Data source를 선택하고 Sync 버튼을 누릅니다. (인덱싱 시작)

나머지 문서도 순차적으로 Sync 합니다.

The screenshot displays the Amazon Bedrock console for the 'ITB-ES_Contracts' knowledge base. The 'Data source (3)' section is highlighted with an orange box and a red circle with the number 1. The 'Sync' button is highlighted with a red circle and the number 2. The 'Test knowledge base' panel on the right shows a 'Generate responses' button and a 'Run' button.

Knowledge base overview

Knowledge base name: ITB-ES_Contracts
Knowledge base ID: JXGUVQIK6
Status: Ready
Created date: July 19, 2024, 09:42 (UTC+09:00)

Tags

No tags to display
Manage tags

Data source (3)

Find data source

Data so...	Status	Data sour...	Account ID	Source Link	Last sync ...	Last sync ...	Chunking...	Parsing st...	Data dele...
es_itb_1	Available	S3	37627801...	s3://2407...	-	-	Default	Claude 3 ...	Delete
es_itb_3	Available	S3	37627801...	s3://2407...	-	-	Default	Claude 3 ...	Delete
es_itb_2	Available	S3	37627801...	s3://2407...	-	-	Default	Claude 3 ...	Delete

Embeddings model

Model: Titan Text Embeddings v2
Vector dimensions: 1024

Vector database

Vector database: Vector engine Amazon OpenSearch Serverless
Collection ARN: arn:aws:aoss:us-east-1:376278017302:collection/aofogon7swk8wkqc5gi
Vector index name: bedrock-knowledge-base-default-index
Vector field name: bedrock-knowledge-base-default-vector
Text field name: AMAZON_BEDROCK_TEXT_CHUNK
Metadata field name: AMAZON_BEDROCK_METADATA

Test knowledge base

Generate responses
Select model
One or more data sources have not been synced.
Go to data sources
Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .
Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.
Please select a model
Run

6. Data Source Sync (인덱싱) (2/2)

Sync 가 완료되면 Last sync에 시간이 표시 됩니다.

Status도 Available 인지 확인 입니다.

Data source (3)

AddEditDeleteSync

Find data source

	Data so...	Status	Data sour...	Account ID	Source Link	Last sync ...	Last sync ...	Chunking...	Parsing st...	Data dele...
<input type="radio"/>	es_itb_1	Available	S3	37627801...	s3://2407...	July 19, 2...	3	Default	Claude 3 ...	Delete
<input type="radio"/>	es_itb_3	Available	S3	37627801...	s3://2407...	July 19, 2...	-	Default	Claude 3 ...	Delete
<input type="radio"/>	es_itb_2	Available	S3	37627801...	s3://2407...	July 19, 2...	3	Default	Claude 3 ...	Delete

7. RAG 테스트 (1/2)

KB는 챗봇 UI가 제공되어, RAG 테스트를 콘솔에서 바로 해볼 수 있습니다.

- (1) 챗봇에 사용할 모델 선택
- (2) Configurations 창 오픈
- (3) Search Type 은 하이브리드 방식으로 변경
- (4) 가져올 청크 갯수 조정
- (5) 사용자 쿼리 입력

Test Knowledge base

2 **Configurations** info Copy X

3 **Search Type**
Search Type refers to how retrieval is done on vector store.
☒ **Override default search**
Default search chooses the best option between hybrid search and semantic search for your vector store for better accuracy.
Choose your preferred option for searching your vector store:
☒ **Hybrid search (semantic & text)**
Combines relevancy scores from semantic and text search to provide greater accuracy.
☐ **Semantic search**
Uses vector embeddings to deliver relevant results.

Inference parameters info Reset V
Set values to influence the responses that the model provides when you query your knowledge base.

Randomness and diversity
Temperature 0
Top P 1

Length
Maximum length 4096
Stop sequences
Add
Observation X

Maximum number of source chunks
Specify the maximum number of retrieved results to return from the vector store.
Enter a number between 1 and 100
20

Filters info
Metadata search helps you improve response accuracy and relevancy. Add filters and then run a query to search with metadata.
Enter
Add Group

Query modifications
☒ **Break down queries**
Enabling this allows the Knowledge base to split complex queries into multiple parts to get more relevant responses. This may improve retrieval accuracy.

1 **Generate responses**
Claude 3 Sonnet
Change

감정에 대한 산식 또는 수식이 있나요?

예, 감정에 대한 수식이 있습니다. 백업 연료 요구량을 초과할 경우의 벌금(Penalty for Exceeding the Guaranteed Back-up Fuel Requirement) 계산식이 제시되어 있습니다: $FOEPm = CPbm * (FHRa / FHRg - 1)$ 여기서, $FOEPm$ = 백업 연료 요구량 초과에 대한 벌금, $CPbm$ = 기준 연료 가격, $FHRa$ = 실제 백업 연료 열량 요구량, $FHRg$ = 보장된 백업 연료 열량 요구량^[1]
[Hide source details](#)

5 Enter your message here Run

7. RAG 테스트 (2/2)

KB는 챗봇 UI에서는 사용자 쿼리에 대한 응답을 생성할때 사용한 Context를 함께 확인할 수 있습니다. (Citations)

Amazon Bedrock > Knowledge bases > ITB-ES Contracts

Test Knowledge base

Generate responses

Claude 3 Sonnet

Configure your retrieval and responses

감정에 대한 산식 또는 수식이 있나요?

예, 감정에 대한 수식이 있습니다. 백업 연료 요구량을 초과할 경우의 벌금(Penalty for Exceeding the Guaranteed Back-up Fuel Requirement) 계산식이 제시되어 있습니다: $FOEPm = CPbm * (FHRa / FHRg - 1)$ 여기서, $FOEPm$ = 백업 연료 요구량 초과에 대한 벌금 $CPbm$ = 로 열량 요구량 $FHRg$ = 보장된 백업 연료 열량 요

[1] Fujairah F3 IPP - EPC Contract - Appendix G (Final).DOCX

The Penalty for Exceeding the Guaranteed Back-up Fuel Requirement FOEP (in AED), which will be calculated if $FHRa > FHRg$ represents the portion of the actual cost of such Back-up Fuel excess require...

Enter your message here

Run

Source details (1)

Query configurations (1)

Source chunk 1

The Penalty for Exceeding the Guaranteed Back-up Fuel Requirement FOEP (in AED), which will be calculated if $FHRa > FHRg$ represents the portion of the actual cost of such Back-up Fuel excess requirement, which shall be calculated as follows: $FOEPm = CPbm * (FHRa / FHRg - 1)$ where: $CPbm$ = as previously defined $FHRa$ = actual hourly fuel heat demand under Back-up Fuel operation, at the guarantee conditions in accordance with Appendix A to the Contract, as most recently tested $FHRg$ = guaranteed hourly fuel heat demand under Back-up Fuel operation as agreed in Appendix A to the Contract The Penalty FOEP will be paid by EPC Contractor to the Owner for the period of Back-up Fuel operation, until such time that the performance for Back-up Fuel firing guaranteed in Appendix A to the Contract is proven by EPC Contractor in accordance with the procedures set forth in Appendix J of the Contract. APPENDIX G TO G-31 FUJAIRAH F3 EPC CONTRACT

Metadata associated with this chunk

Key	Value
x-amz-bedrock-kb-source-uri	s3://240719-jesamkim-bucket/ES_Contracts/UAE_F3/EPC/Fujairah F3 IPP - EPC
x-amz-bedrock-kb-chunk-id	1%3A0%3AnFR5yJAB6z4CrzVbckEb
x-amz-bedrock-kb-data-source-id	HGYNFN7RJV

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Thank you!

Appendix

Instruction for the parser 프롬프트 (default)

Transcribe the text content from an image page and output in Markdown syntax (not code blocks). Follow these steps:

1. Examine the provided page carefully.
2. Identify all elements present in the page, including headers, body text, footnotes, tables, visualizations, captions, and page numbers, etc.
3. Use markdown syntax to format your output:
 - Headings: # for main, ## for sections, ### for subsections, etc.
 - Lists: * or - for bulleted, 1. 2. 3. for numbered
 - Do not repeat yourself
4. If the element is a visualization
 - Provide a detailed description in natural language
 - Do not transcribe text in the visualization after providing the description
5. If the element is a table
 - Create a markdown table, ensuring every row has the same number of columns
 - Maintain cell alignment as closely as possible
 - Do not split a table into multiple tables
 - If a merged cell spans multiple rows or columns, place the text in the top-left cell and output ' ' for other
 - Use | for column separators, |-| for header row separators
 - If a cell has multiple items, list them in separate rows
 - If the table contains sub-headers, separate the sub-headers from the headers in another row
6. If the element is a paragraph
 - Transcribe each text element precisely as it appears
7. If the element is a header, footer, footnote, page number
 - Transcribe each text element precisely as it appears

Output Example:

A bar chart showing annual sales figures, with the y-axis labeled "Sales (\$Million)" and the x-axis labeled "Year". The chart has bars for 2018 (\$12M), 2019 (\$18M), 2020 (\$8M), and 2021 (\$22M).
Figure 3: This chart shows annual sales in millions. The year 2020 was significantly down due to the COVID-19 pandemic.

Annual Report

Financial Highlights

* Revenue: \$40M
* Profit: \$12M
* EPS: \$1.25

|| Year Ended December 31, ||
|| 2021 | 2022 |
|-|-|
| Cash provided by (used in): |||
| Operating activities | \$ 46,327 | \$ 46,752 |
| Investing activities | (58,154) | (37,601) |
| Financing activities | 6,291 | 9,718 |

Here is the image.

선택 가능한 Chunking strategy

Chunking strategy
Chunking breaks down the text into smaller segments before embedding. The chunking strategy can't be modified after you create the data source.

Default chunking
Automatically splits text into chunks of about 300 tokens in size, by default. If a document is less than or already 300 tokens, it's not split any further.

Default chunking
Automatically splits text into chunks of about 300 tokens in size, by default. If a document is less than or already 300 tokens, it's not split any further.

Fixed-size chunking
Splits text into your set approximate token size.

Hierarchical chunking
Organizes text chunks (nodes) into hierarchical structures of parent-child relationships. Each child node includes a reference to its parent node.

Semantic chunking
Organizes text chunks or groups of sentences by how semantically similar they are to each other.

No chunking
Suitable for documents that are already pre-processed or text split into separate files without any further chunking necessary.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

13