

AWS Serverless e Amazon Bedrock com bases de conhecimento, usando dados privados





AWS Serverless and Amazon Bedrock with knowledge bases, using private data

Why Watch this talk?

Cloud is the **future**



Serverless is the **future**



Infrastructure as code is the **future**



Artificial Intelligence is the **future**



Your **future** is built in the **present**

About the Speaker

Weder Mariano de Sousa

Specialist Senior Java - GFT



Technician **System Development**

Graduated **Computer Science**

Post Graduate in **Midias UFG**

Post Graduate in **Information Security**



GOJava



AWS User
Group Goiânia

<https://www.linkedin.com/in/wedermarianodesousa/>

<https://github.com/weder96>

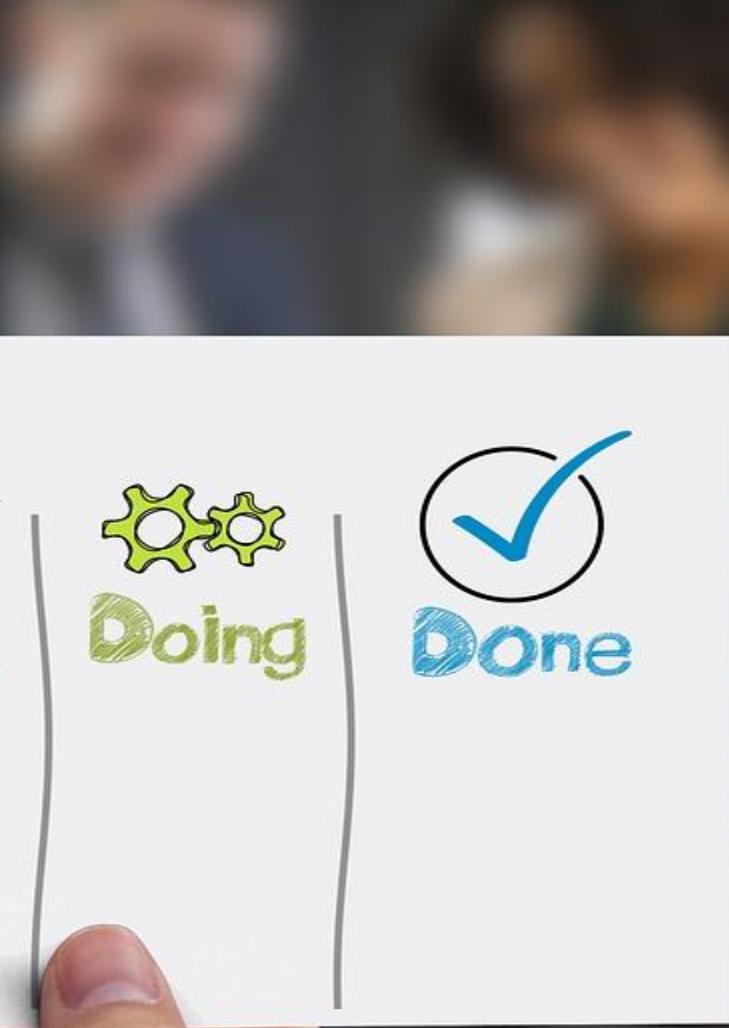
<https://twitter.com/weder96>

<https://dev.to/weder96>



Agenda

1. Generative Artificial Intelligence (Gen AI)
2. Architecture AWS Bedrock
3. What is RAG (Retrieval-Augmented Generation)?
4. What will we build?
5. Knowledge bases
6. Deploy Infra and Application with CDK
7. AWS CDK For Java and Python
8. Next Steps



Generative Artificial Intelligence (Gen AI)

Generative artificial intelligence (Gen AI) refers to a subfield of artificial intelligence that focuses on creating systems capable of generating new, original content, such as text, images, music, or videos. Unlike traditional AI systems, which analyze and process existing data, Gen AI uses algorithms to produce novel outputs. This field has various applications, including language translation, image synthesis, and content creation. Google's researchers have made significant contributions to Gen AI, exploring its potential in tasks like generating human-like conversations and creating realistic multimedia content. The technology is still in its early stages of development.

Generative artificial intelligence (Gen AI) is rapidly reshaping our world. According to McKinsey, Gen AI's influence on productivity might contribute trillions of dollars in value to the global economy annually. All industries have the opportunity to benefit from its capabilities, with new applications emerging daily. Gen AI is poised to have a profound effect on human life across a wide spectrum of sectors, including healthcare and life sciences. While its potential is undeniable, many organizations are still in the process of exploring possibilities and figuring out how to effectively integrate Gen AI to create meaningful impacts within their businesses.



INCREASE REVENUE

Gen AI is a powerful tool for business innovation. It is used to design new products, services, and solutions that satisfy customer needs and increase revenue.



DECREASE COSTS

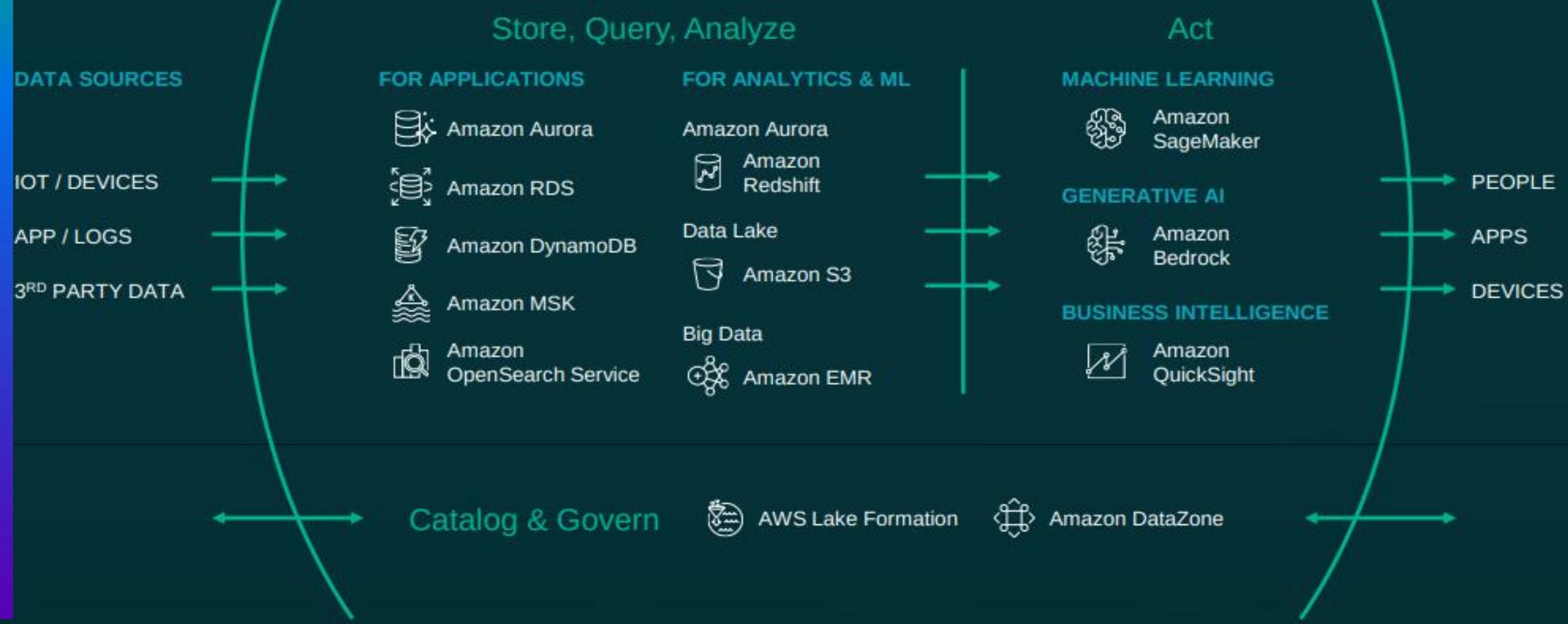
Gen AI is useful when applied to business optimization opportunities. It can help automate tasks, improve efficiency, and reduce errors that cost time and money.



ATTRACT AND RETAIN CUSTOMERS

Gen AI improves engagement through personalized experiences. This leads to attracting new customers along with an increase in brand loyalty and customer lifetime value.

A comprehensive set of services for your data foundation



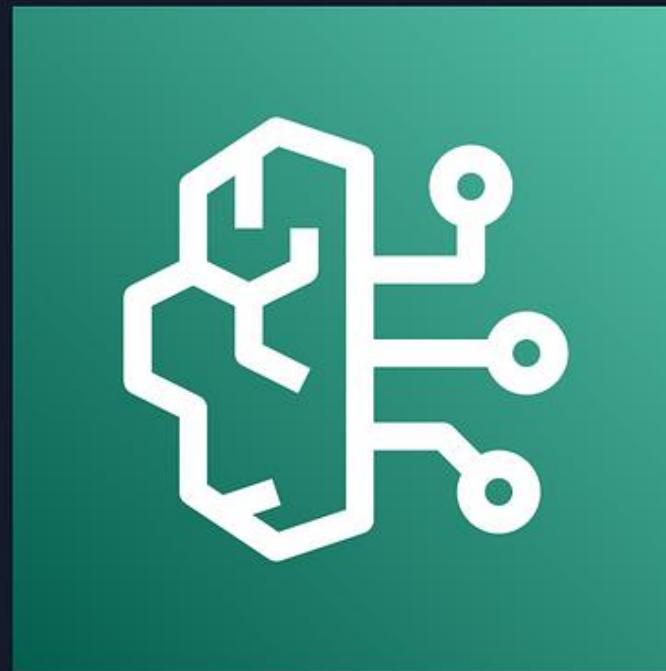
Architecture AWS Bedrock

**Topic 2: Presentation of the integration
architecture AWS Bedrock**





Amazon Bedrock



Amazon Bedrock

https://aws.amazon.com/pt/bedrock/?did=ap_card&trk=ap_card

MACHINE LEARNING



DEEP LEARNING



FOUNDATION MODELS



Foundation Models

Foundation models

Amazon Bedrock supports foundation models from industry-leading providers. Choose the model that is best suited to achieving your unique goals.





Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models

AI21labs

amazon

ANTHROPIC

cohere

Meta

MISTRAL
AI

stability.ai



Contextual answers,
summarization,
paraphrasing

Text summarization,
generation, Q&A, search,
image generation

Summarization, complex
reasoning, writing, coding

Text generation,
search, classification

Q&A and reading
comprehension

Text summarization,
Q&A, text classification,
text completion, code
generation

High-quality images
and art

Jurassic-2 Ultra

Amazon Titan Text Premier

Claude 3 Opus

Command

Llama 3 8B

Mistral Large

Stable Diffusion XL 1.0

Jurassic-2 Mid

Amazon Titan Text Lite

Claude 3 Sonnet

Command Light

Llama 3 70B

Mistral 7B

Stable Diffusion XL 0.8

Amazon Titan Text Express

Claude 3 Haiku

Embed English

Llama 2 13B

Mixtral 8x7B

Amazon Titan Text Embeddings

Claude 2.1

Embed Multilingual

Llama 2 70B

Amazon Titan Text Embeddings
V2

Claude 2

Command R+

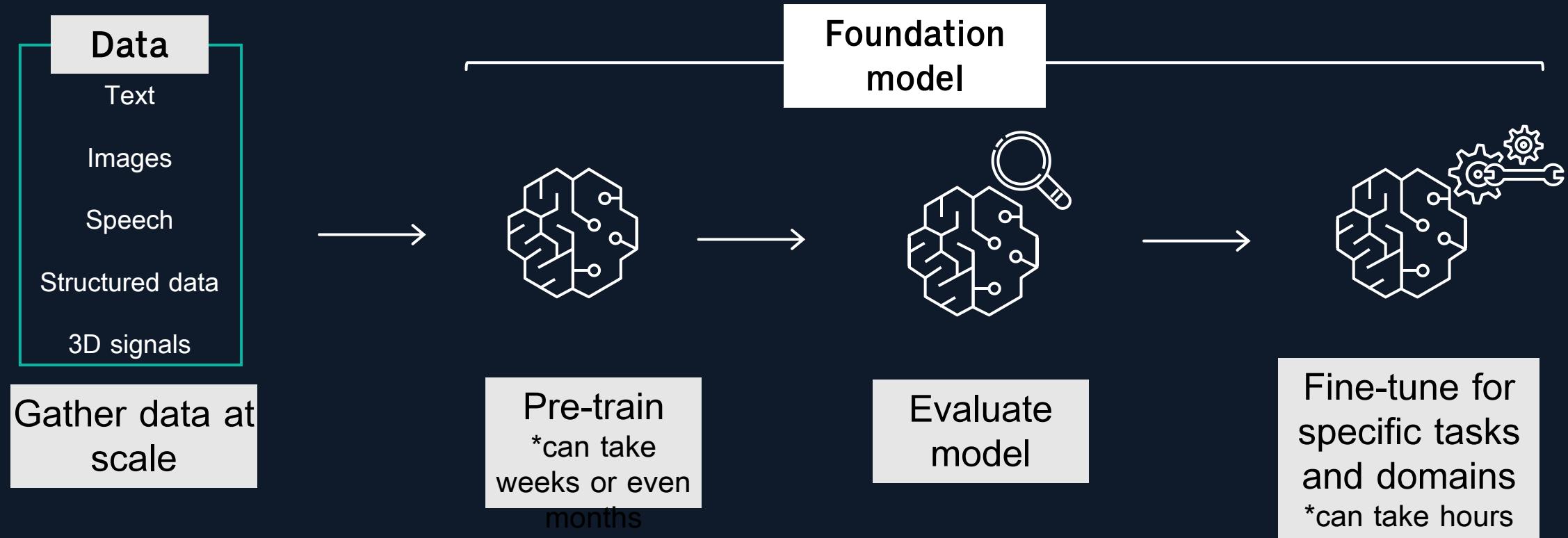
Amazon Titan Multimodal
Embeddings

Claude Instant

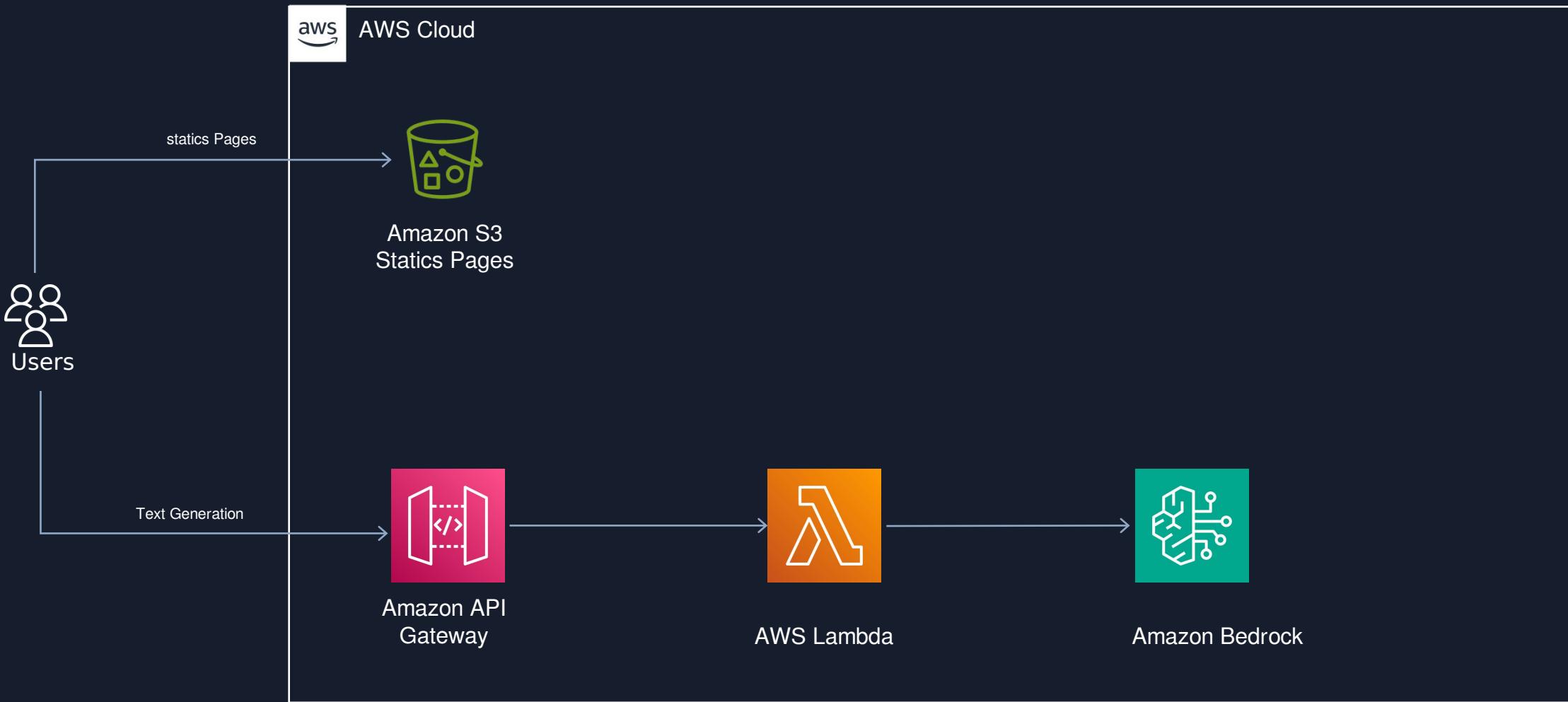
Command R

Amazon Titan Image Generator

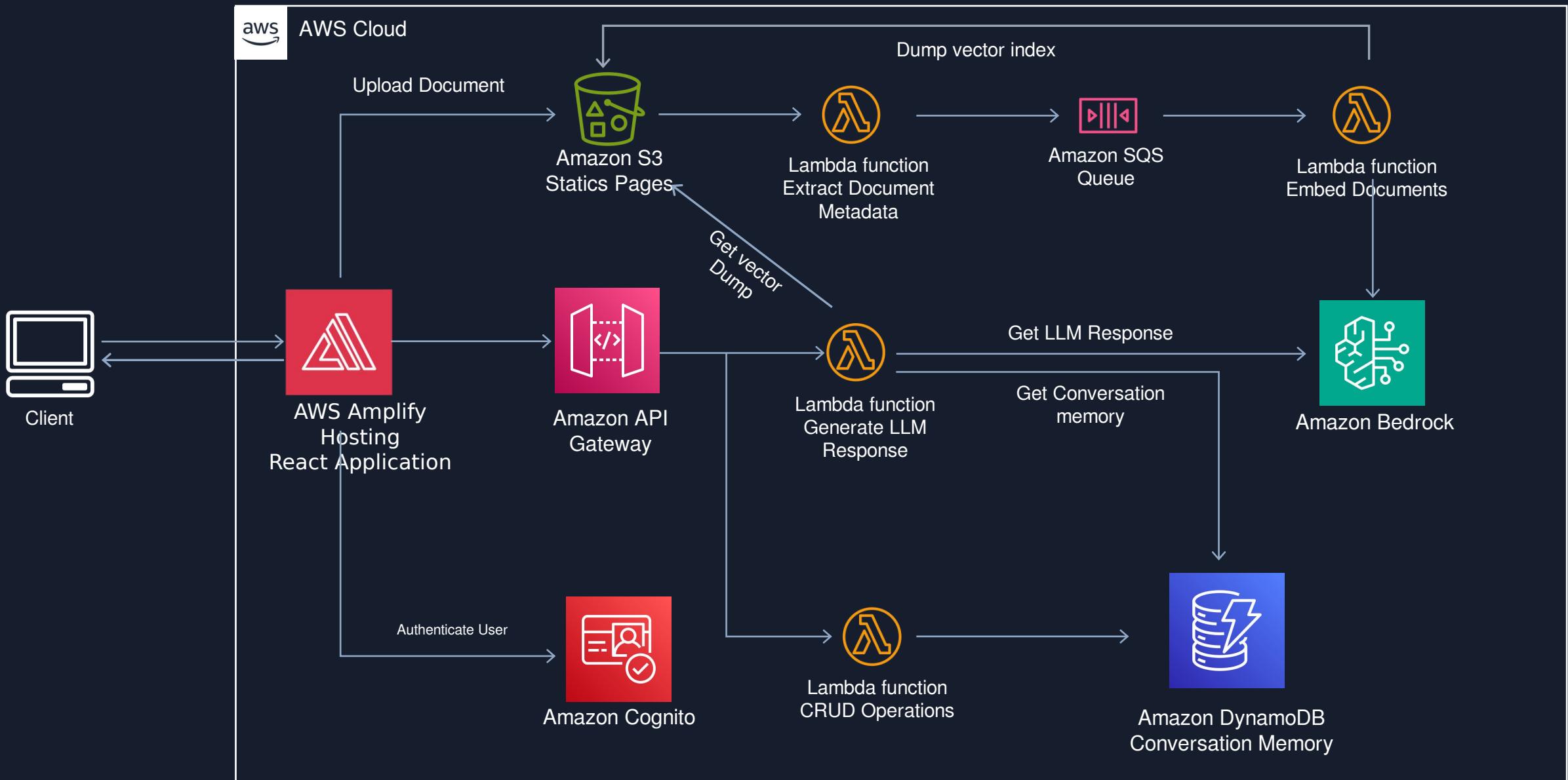
How does a foundation model work?



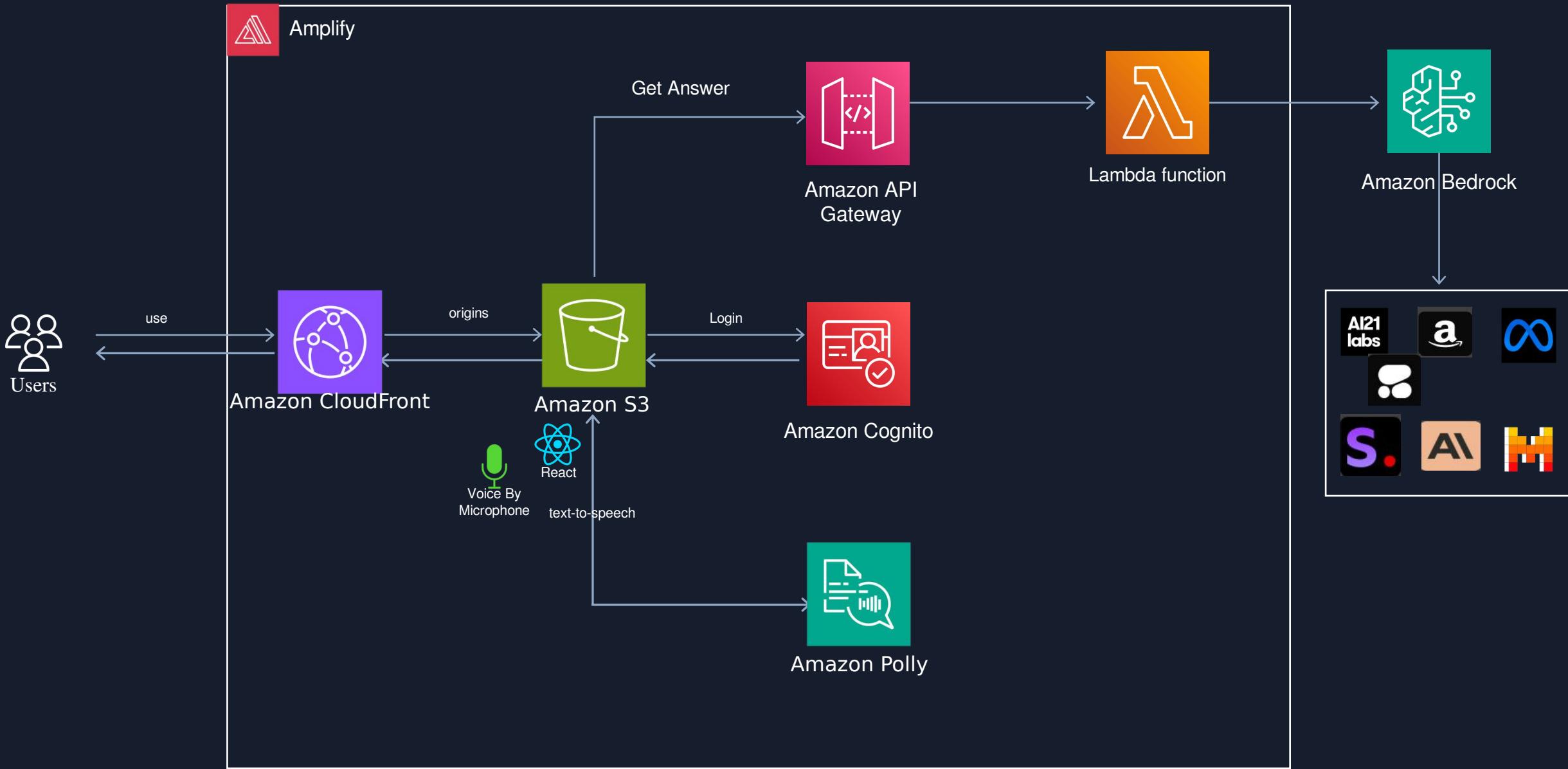
Amazon Bedrock Serveless Basic Architecture



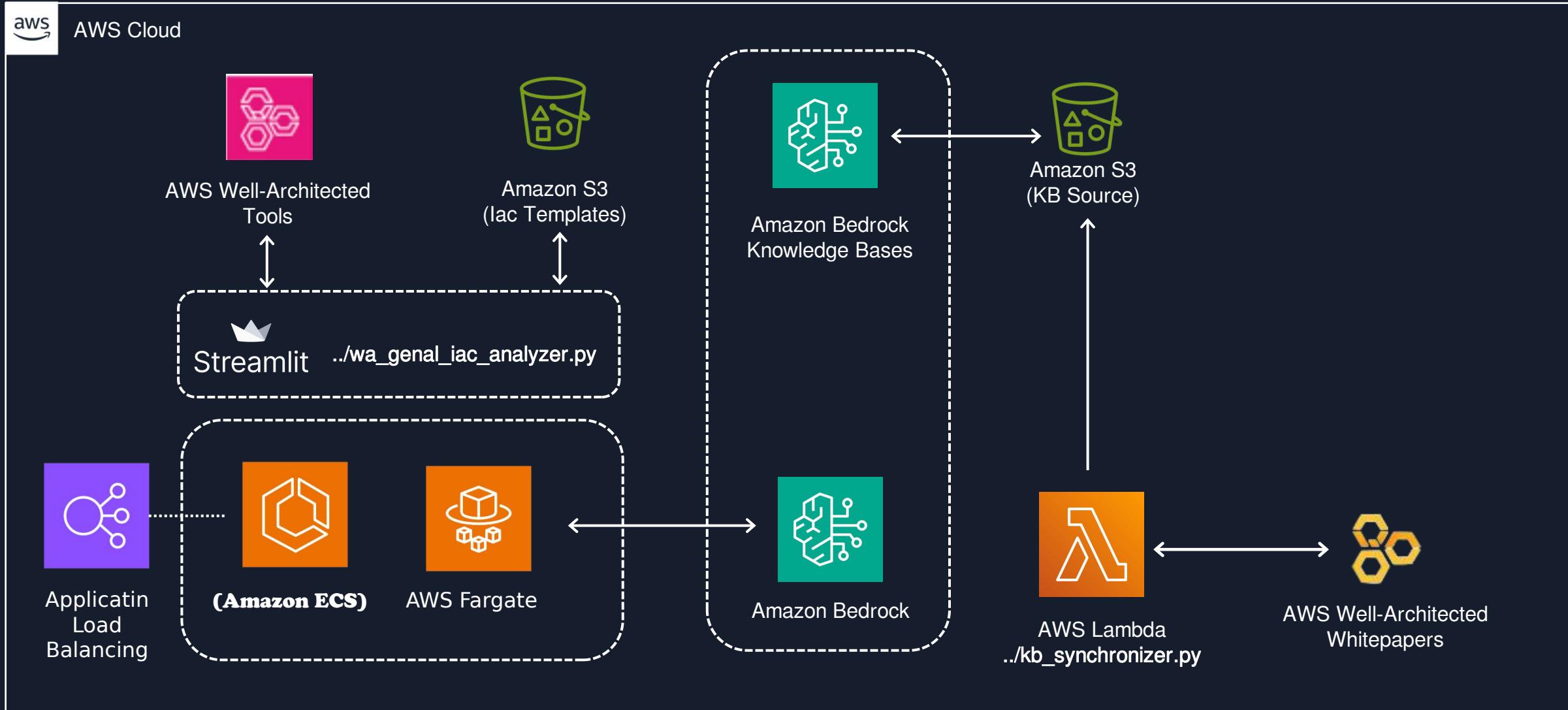
Advantages of the Serverless Approach



Advantages of the Serverless Approach



Architecture(ECS and Fargate) with Bedrock by Rossana Suarez



<https://github.com/aws-samples/well-architected-iac-analyzer>

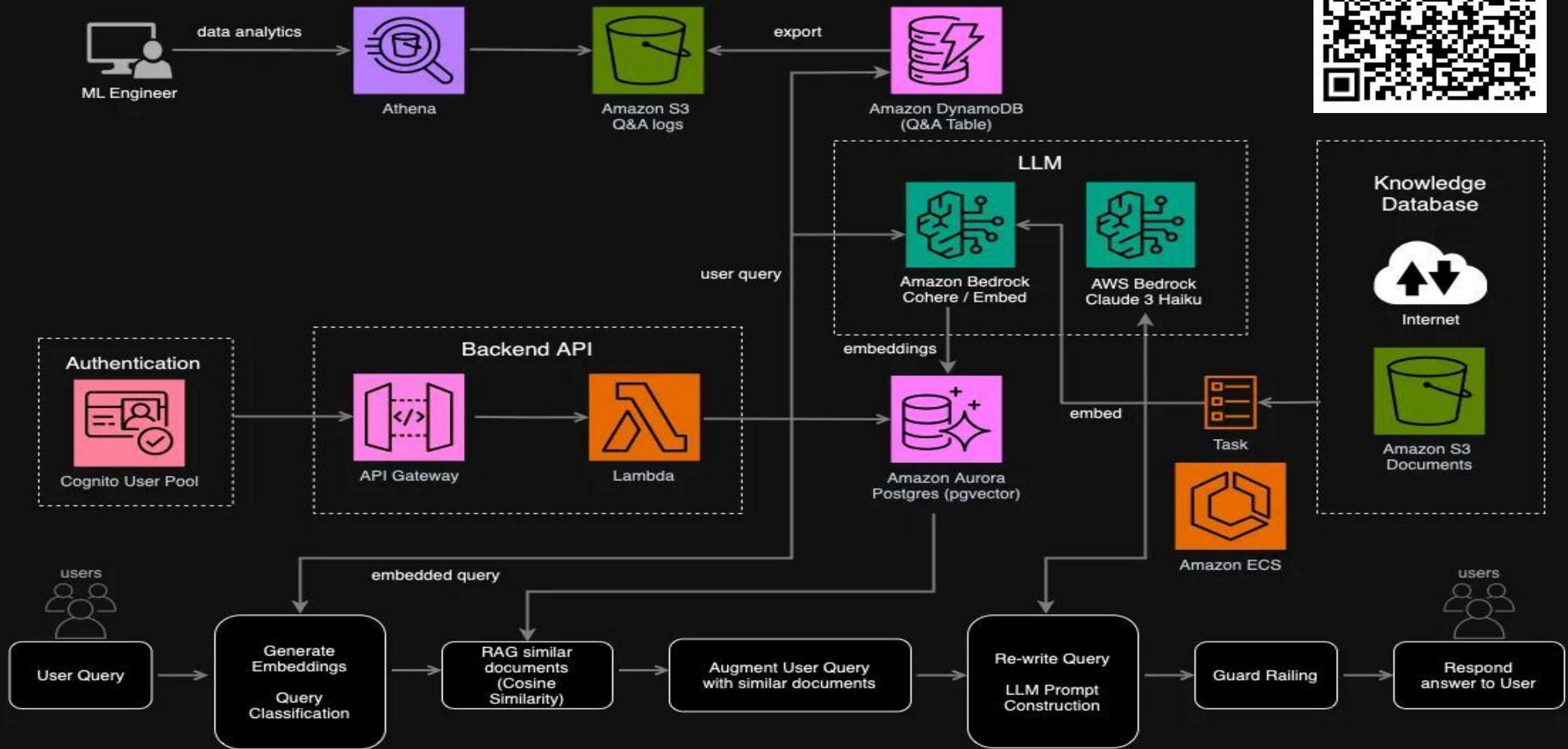
Step into the Future of AI with Amazon Bedrock



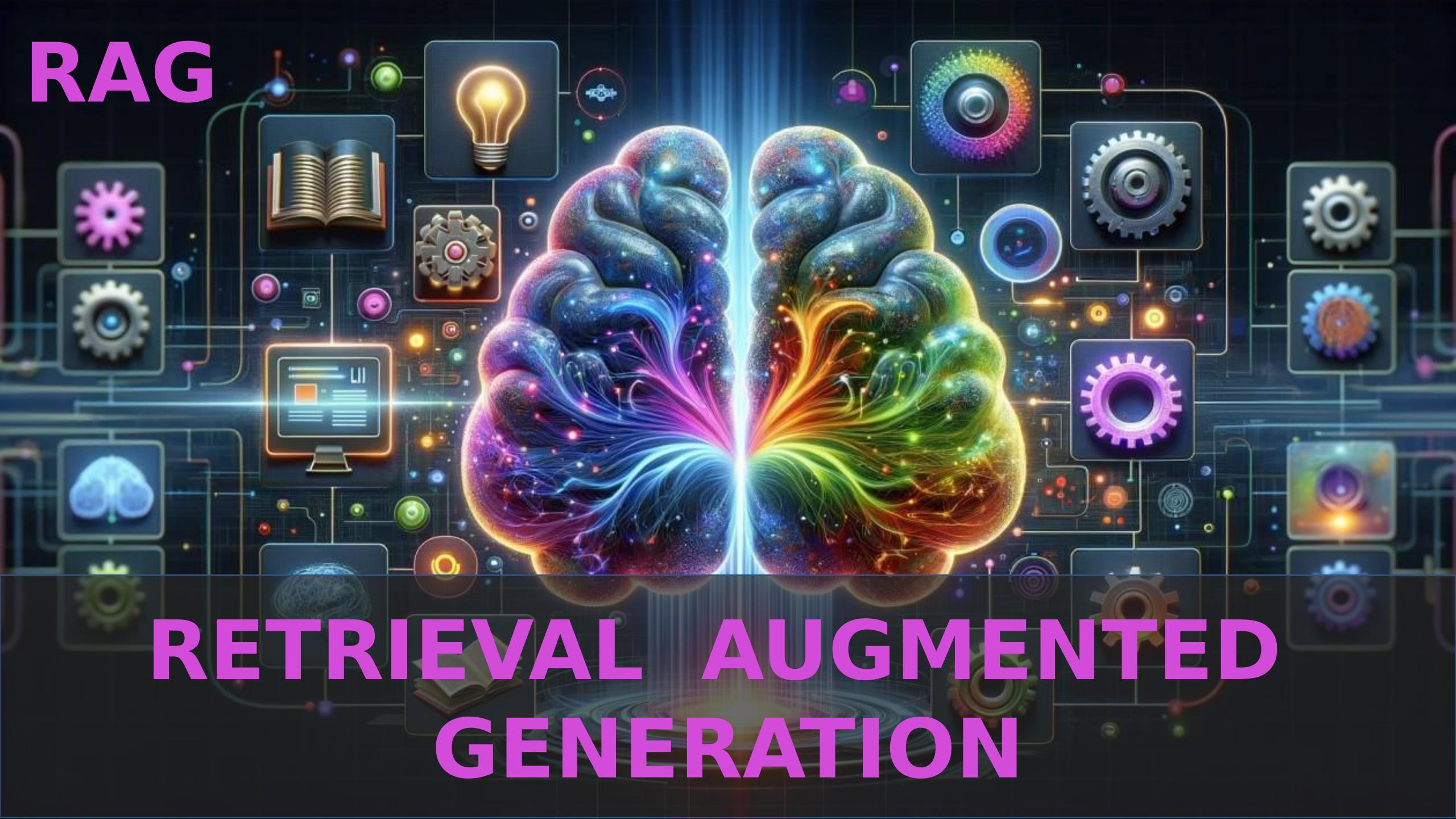
Amazon Bedrock



AWS Bedrock Generative AI Application Architecture



RAG



RETRIEVAL AUGMENTED GENERATION

What is RAG (Retrieval-Augmented Generation)?

In **Technology** there is a name that I consider a **Genius** and at the same time a **visionary**, who would this person be?

What is RAG (Retrieval-Augmented Generation)?

**He is the *creator* of the Linux
Operating System**

What is RAG (Retrieval-Augmented Generation)?

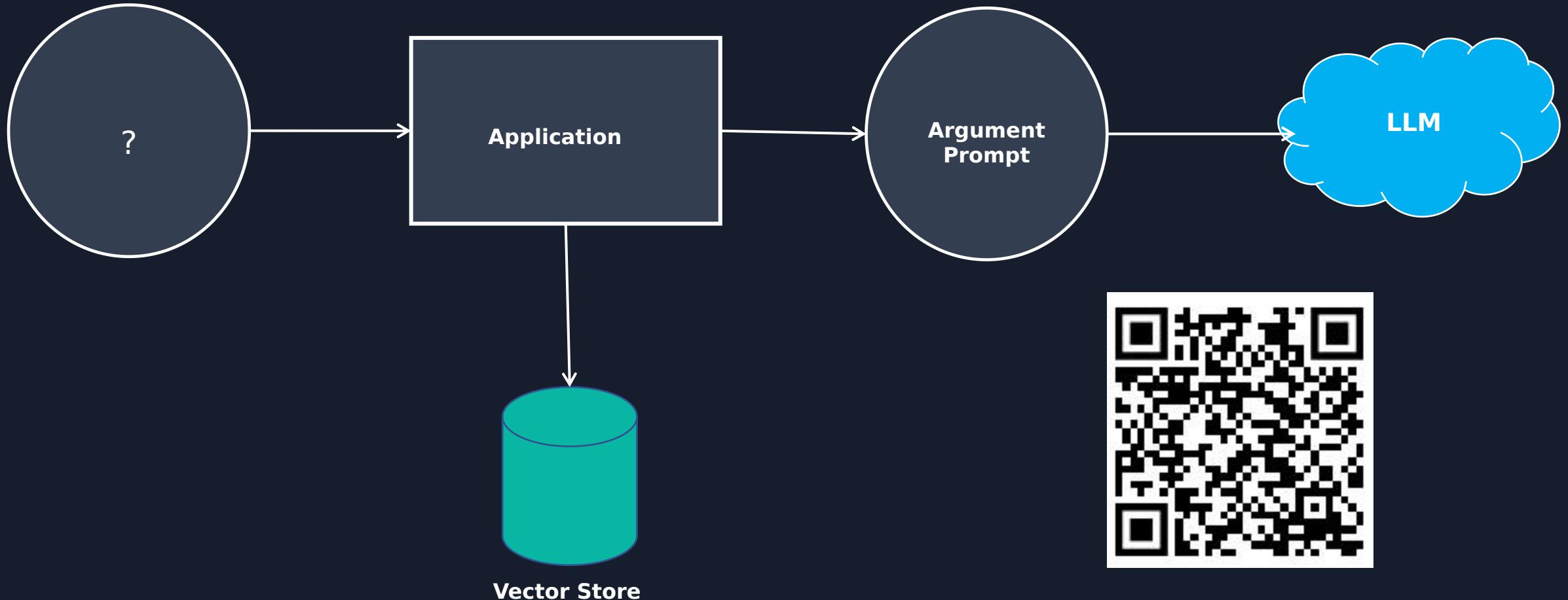


Linus Benedict Torvalds (Helsínquia, 28 de dezembro de 1969) é um engenheiro de software, nascido na Finlândia e naturalizado estado-unidense em 2010

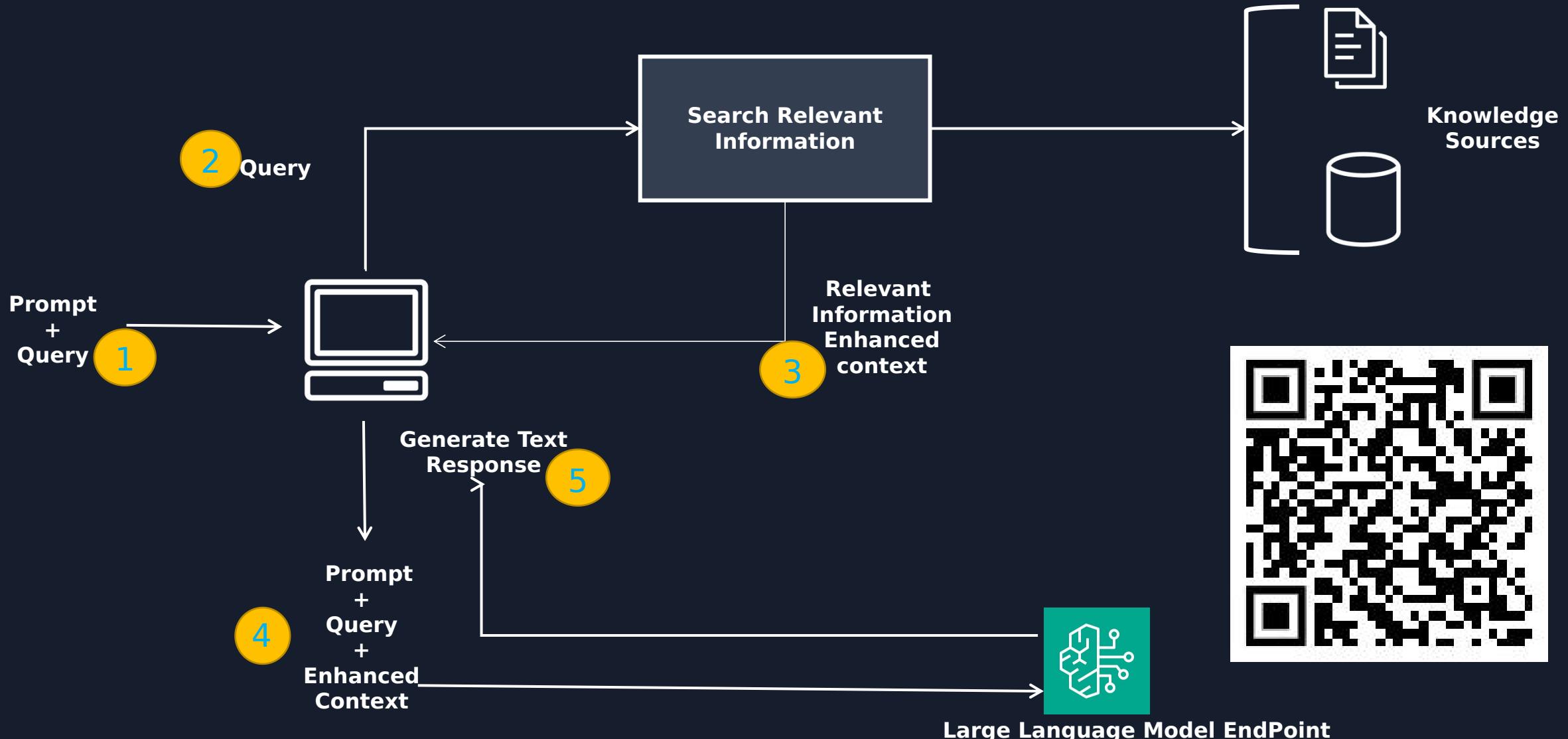
Criador, e por muito tempo o desenvolvedor mais importante do núcleo Linux, sendo utilizado em importantes sistemas Linux, Android e Chrome OS.

É também o criador do Git, sistema de controle de versão amplamente utilizado, e o aplicativo para planejamento e registro de mergulho, Subsurface.

AI Application Architecture - RAG



Conceptual Flow of using RAG with LLMs

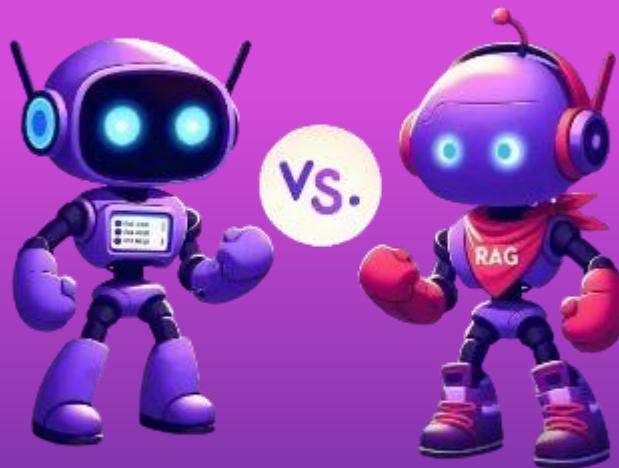


Comparing **RAG v/s Agent AI**

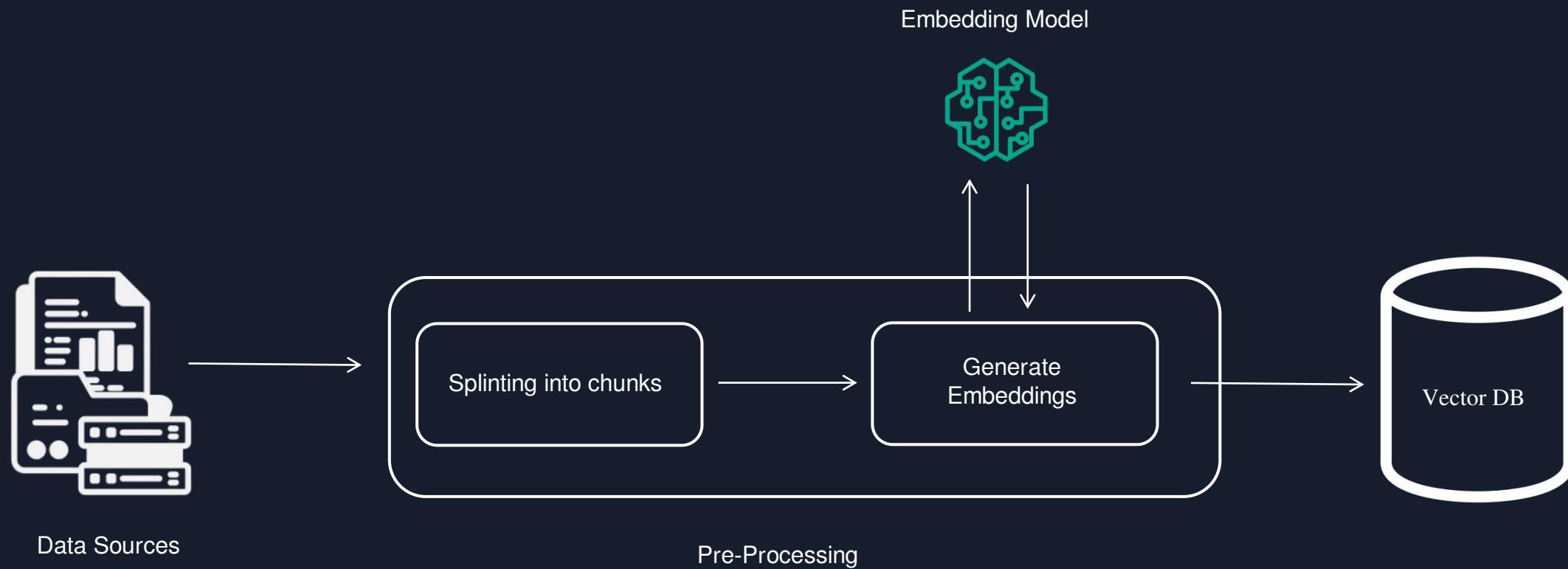
*RAG chatbots and agent AI:
knowledge-base and Automation*



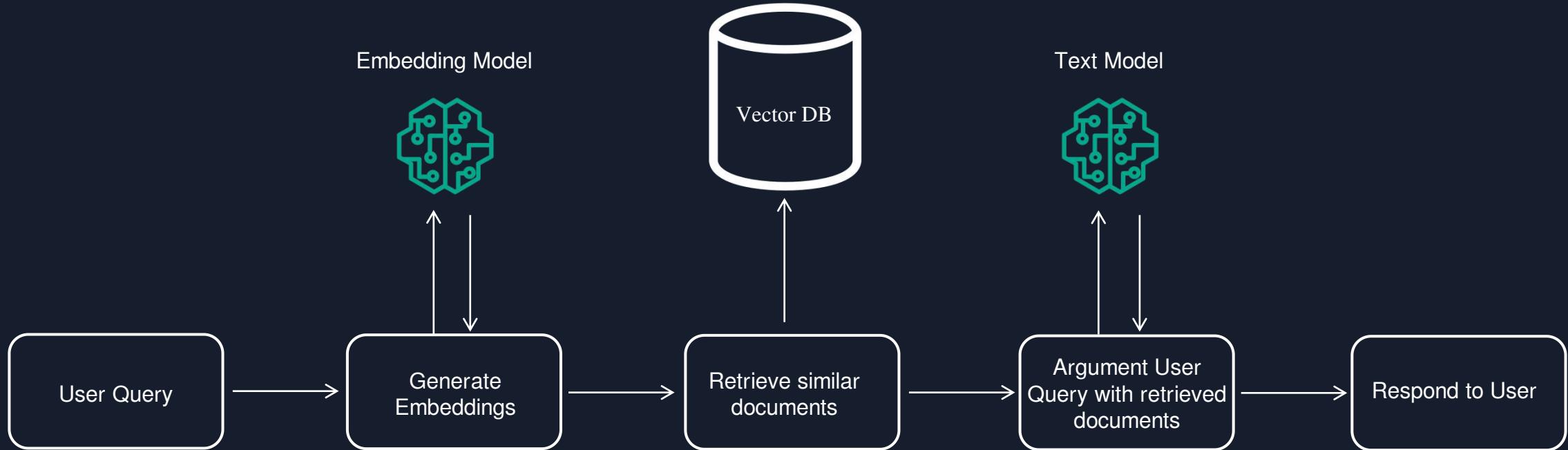
<https://yourgpt.ai/blog/general/rag-chatbot-vs-ai-agent>



Pre-processing data(Bedrock)



Runtime Execution(Bedrock)

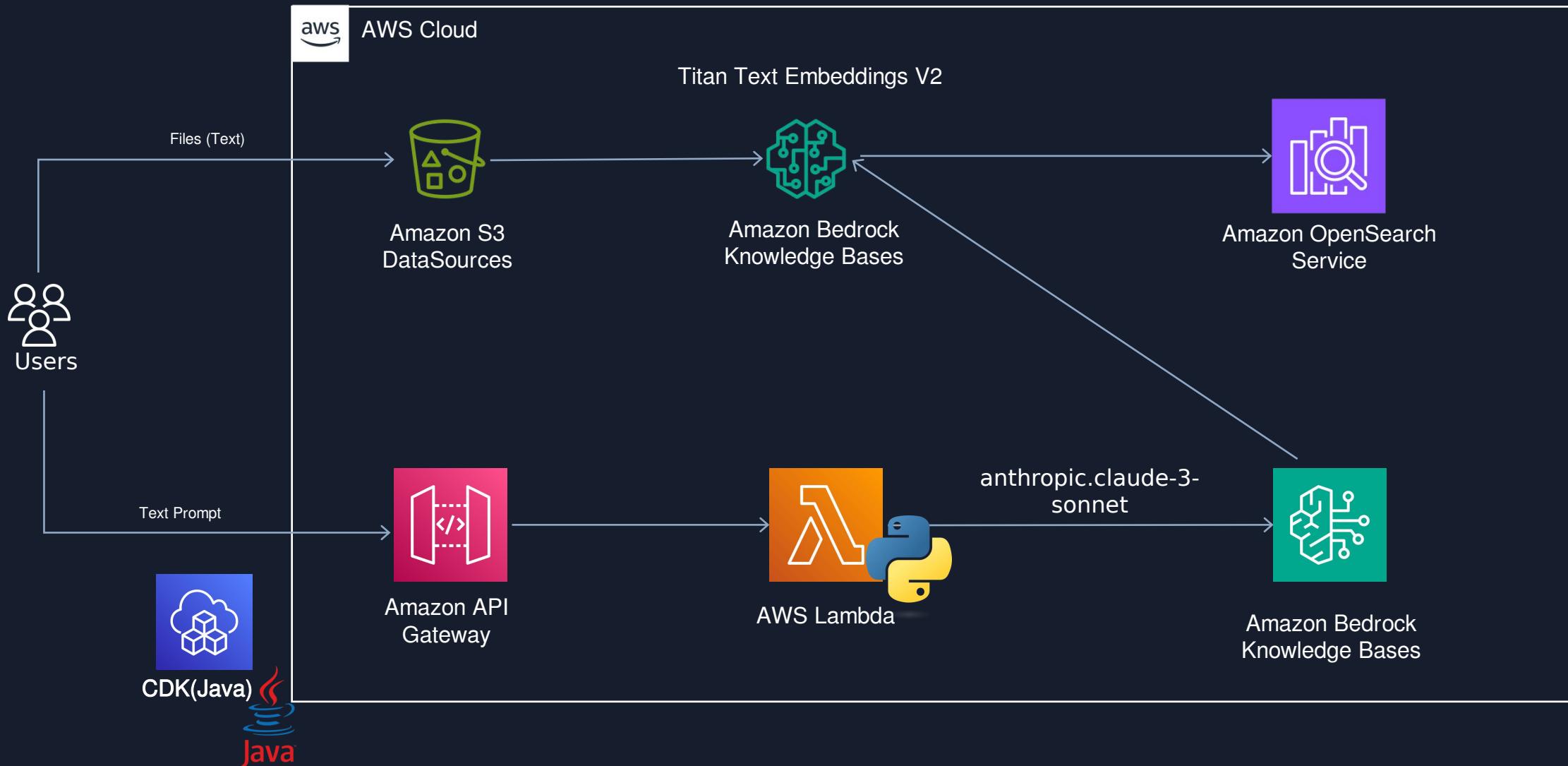




What will we build?



Amazon Bedrock Serveless Architecture



Create Data Source at S3(Storage)

Amazon S3 > Buckets > datasourcews2024

datasourcews2024 Info

Objects Properties Permissions Metrics Management Access Points

Objects (3) Info

C 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](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	marvel_min/	Folder	-	-	-
<input type="checkbox"/>	marvel/	Folder	-	-	-
<input type="checkbox"/>	ux_ui/	Folder	-	-	-

Knowledge bases Steps

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

Provide knowledge base details

Knowledge base details

Knowledge base name

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

Knowledge base description - optional

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

Choose Data Source

Choose data source

Select the data source that you want to configure in the next step. You can add up to 5 data sources in a knowledge base.



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.

Browser s3 files

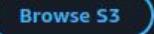
 **Amazon S3** Info
Provide details to connect Amazon Bedrock to your S3 data source.

▼ Data source: knowledge-base-quick-start-gjjuz-data-source [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

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
 X [View](#) 

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 3 more data source(s).

[Cancel](#) [Previous](#) **Next**

Choosing the model for Embeddings dataSources

Step 3
 Select embeddings model and configure vector store

Step 4
 Review and create

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

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.

Review and Create

Step 2
Configure data source

Step 3
Select embeddings model and configure vector store

Step 4
Review and create

Step 1: Provide details

Edit

Knowledge base details			
Knowledge base name	knowledge-base-campus-party-2024-ds	Knowledge base description	—
Data source type	S3	Service role	AmazonBedrockExecutionRoleForKnowledgeBase_02257
Tags (0)			
Key		Value	
No tags to display			

Step 2: Setup up data source

Edit

Data source: knowledge-base-quick-start-drksj-data-source		
Data source name	Account ID	S3 URI
knowledge-base-quick-start-drksj-data-source	(this account)	s3://datasourceaws2024/marvel/ []
Customer-managed KMS Key for S3	KMS key for transient data storage	Chunking strategy
-	-	Default
Parsing strategy	Lambda function	S3 bucket for Lambda function
Default	-	-

Sync Data Sources

Syncing data source - 'knowledge-base-quick-start-drksj-data-source'. It could take several minutes to hours to sync, depending on the size of your data.

Service Role
AmazonBedrockExecutionRoleForKnowledgeBase_02257

Created date
November 07, 2024, 13:24 (UTC-03:00)

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	Value
No tags	

[Manage tags](#)

Data source (2)
Data sources contain information returned when querying a Knowledge base.

Sync Stop sync Add ▾

Syncing data source - 'knowledge-base-quick-start-gjjuz-data-source'. It could take several minutes to hours to sync, depending on the size of your data.

Manage tags

Data source (2)
Data sources contain information returned when querying a Knowledge base.

Find data source

Sync Stop sync Add ▾

Data so... ▾	Status	Data sourc...	Account ID	Source Link	Last sync t...	Last sync ...	Chunking ...	Parsing str...	Data dele...
<input checked="" type="radio"/> knowledge-base-quick-start-drksj-data-source	Syncing	S3	71030481...	S3://datasourceaws2024/ux_ui/	November 07, 2024, 13:24 (UTC-03:00)	-	Default	Default	Delete
<input type="radio"/> knowledge-base-quick-start-gjjuz-data-source	Available	S3	71030481...	S3://datasourceaws2024/marvel/	November 07, 2024, 13:24 (UTC-03:00)	-	Default	Default	Delete

Embeddings model

Model
Titan Text Embeddings v2

Embeddings type
Float vector embeddings

Vector dimensions
1024

Sync Completed

Sync completed for data source - 'knowledge-base-quick-start-qjjuz-data-source' X

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	Value
	No tags

[Manage tags](#)

Data source (2)

Data sources contain information returned when querying a Knowledge base.

	<input type="text"/> Find data source	< 1 >							
Data so...	Status	Data sourc...	Account ID	Source Link	Last sync t...	Last sync ...	Chunking ...	Parsing str...	Data dele...
<input checked="" type="radio"/> knowledge-base-quick-start-qjjuz-data-source	<input checked="" type="checkbox"/> Available	S3	71030481...	s3://datasourceaws2024/ux_ui/ [2]	November 10, 2024	-	Default	Default	Delete
<input type="radio"/> knowledge-base-quick-start-qjjuz-data-source	<input checked="" type="checkbox"/> Available	S3	71030481...	s3://datasourceaws2024/marvel/ [2]	November 10, 2024	-	Default	Default	Delete

Embeddings model

Model Titan Text Embeddings v2	Embeddings type Float vector embeddings	Vector dimensions 1024
-----------------------------------	--	---------------------------

Selected Model And Tests

Select model

1. Categories

Model providers

 Amazon

 Anthropic

 Meta

2. Models

Models with access (1)

Titan Text G1 - Premier v1

Text model | Max 32k tokens

3. Inference

On-demand

Cancel

Apply

Not seeing a model you are interested in? Check out all supported models [here](#)

Test Knowledge Base

Amazon Bedrock > Knowledge bases > knowledge-base-campus-party-2024-ds

Test Knowledge base

Generate responses

Titan Text G1 - Premier v1 | On-demand Change

One or more data sources have not been synced. [Go to data sources](#)

What was Zola afraid of?

Zola was afraid of dying [1][2][3] [Hide source details](#)

Enter your message here [Run](#)

▼ Source chunk 1

Depois de vê-la fitar a tela por quase meio minuto, Tony não aguentou mais. – Que foi? Ela estendeu o celular para que ele pudesse ver a tela. – É do Zola – disse ela. – Enviada faz uns vinte minutos. Falo para o mundo que jamais escutou, este mundo que jaz abaixo, coberto de branco contra o maravilhoso preto do espaço. Desta vez, talvez, eu tenha sido superado. Desta vez, talvez, meus comandos não sejam ouvidos, minhas visões não sejam realizadas, meus avanços não sejam percebidos. Aírm Zola morre sabendo que sua vida foi dedicada à busca do conhecimento e do melhoramento daqueles elementos da raça humana que vale a pena melhorar. E morre sabendo que nada nunca morre, e que mesmo que esse truismo heraclitiano falhe perante um teste de gravidade ou seriedade, ele cabe à situação. Corte a cabeça de uma Hidra, sr. Stark, e dez outras crescem no lugar. Corte o corpo de um Zola... enfim, você já sabe.

► Metadata associated with this chunk

▼ Source chunk 2

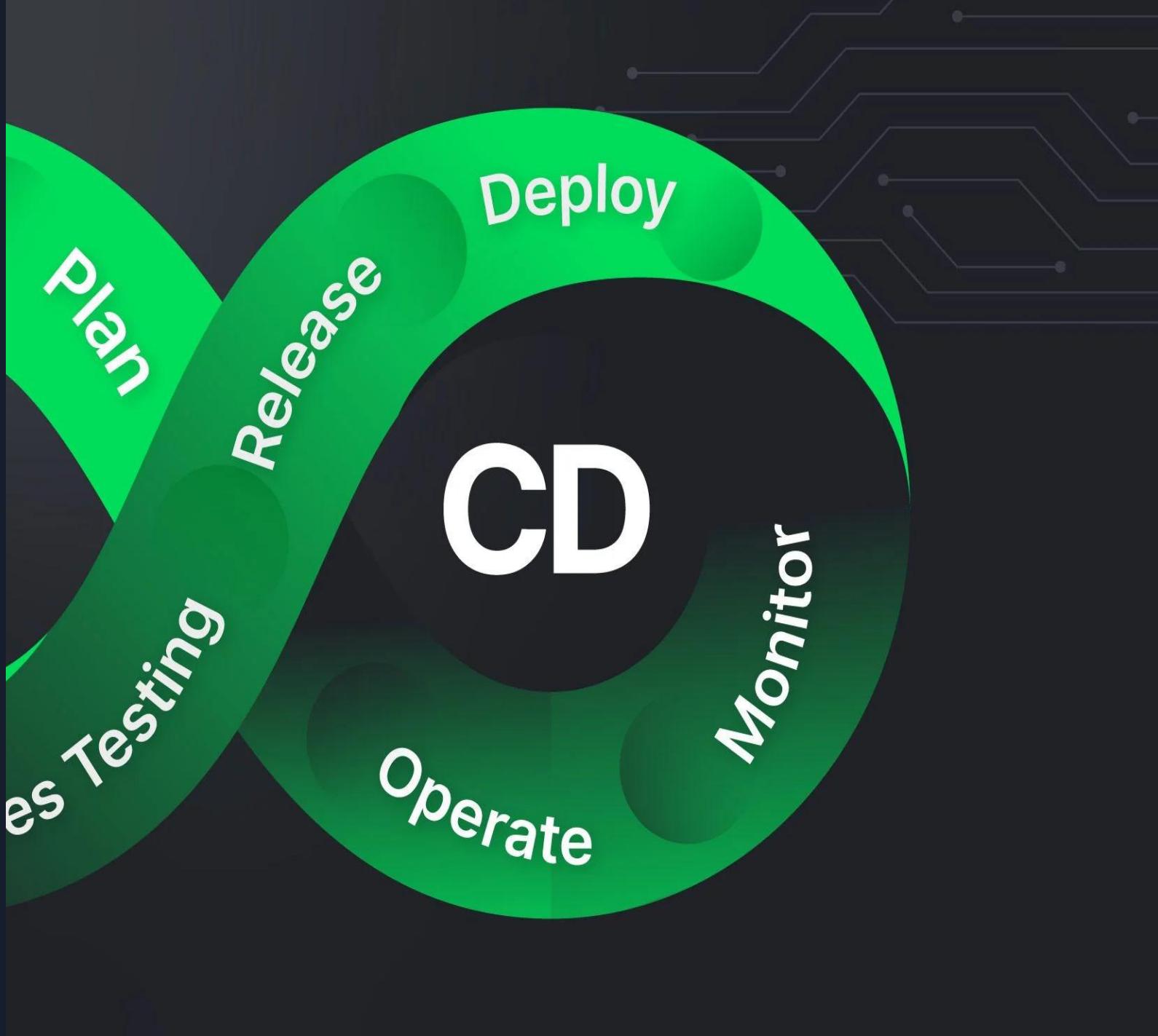
Saber que um clone estivera por aí na cidade passando-se por ela tornava tudo pior ainda. Ela não sabia que a criatura vinha fazendo, que tipo de impressão deixara. Ia morrer, e nem sabia como haviam sido suas três últimas semanas de vida, já que, enquanto o clone percorria o mundo vivendo sua vida, ela estivera aprisionada ali. Passou por sua mente que já morrera de todo modopossível. Tudo o que restava a Zola ou a uma das Oito Madames era fazer o corpo dela parar de funcionar. – Bem – disse Zola –, a qualquer momento, os membros restantes da equipe da S.H.I.E.L.D. enviada para extrair você chegarão a essa porta. Eventos desagradáveis sucederão. Espero que fique fora do caminho. – Por que me manteve viva? – Serena perguntou. No fim, não importava, mas ela quis saber. – Fui aconselhado ao contrário – disse Zola. Refletiu um pouco. – Um gosto pessoal, talvez. Não seria a primeira vez que minhas reações emocionais geraram dificuldades profissionais.

▼ Metadata associated with this chunk

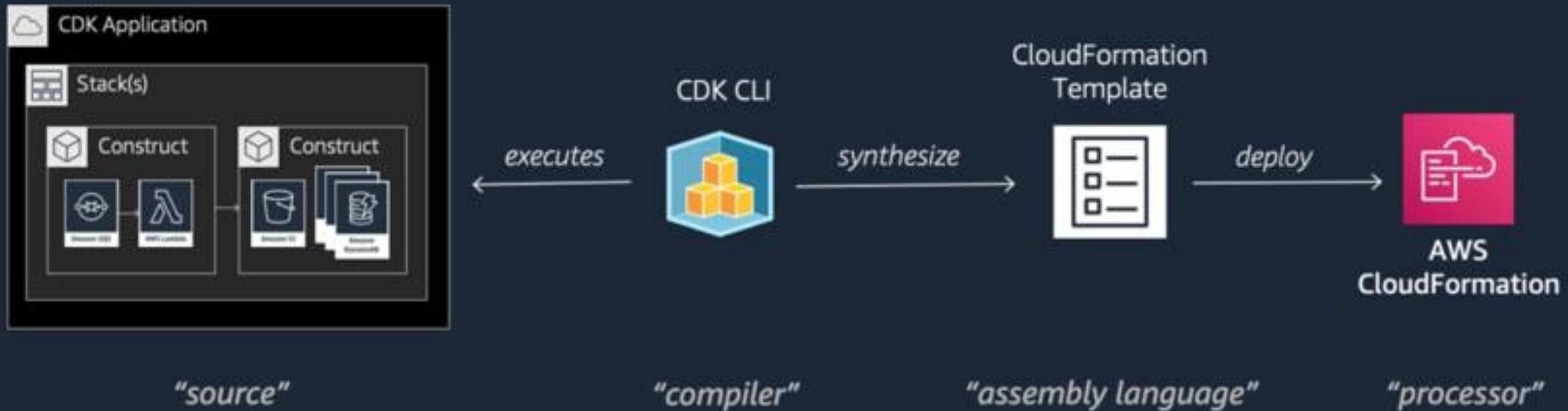
Key	Value
-----	-------



AWS CDK for Java and Python



CDK Lifecycle



AWS CDK INFRA

CampusParty2024 Private

main · 1 Branch · 0 Tags

Go to file Add file · Code

weder96 add cdk infra · b420e85 · now · 2 Commits

architectureAwsIcons · first commit · 2 days ago

infra · add cdk infra · now

lambda · add cdk infra · now

marvel · first commit · 2 days ago

marvel_min · add cdk infra · now

presentation2024 · add cdk infra · now

CampusParty2024 · first commit · 2 days ago

README.md · first commit · 2 days ago

README

CampusParty2024

Project AWS CDK PROJECT

infra-cdk-Kb-java [poc-cdk-java] ~/Documents

.idea
my-cdk
cdk.out
lambda
knowledgeBasesFunction.py
lambdaSqs
postman
src
main
java
dev.wsousa
CreateBucket
CreateDynamoDb
CreateDynamoMapDb
CreateKnowledgeBasesLamp
CreateLambdaCron
CreateSqsAndSns
InfrastructureApp
InfrastructureStack
test
target
.gitignore
cdk.json
package.json
package-lock.json
pom.xml

InfrastructureApp.java InfrastructureStack.java CreateKnowledgeBasesLambdaAPI.java knowledgeBasesFunction.py

```
15 public class CreateKnowledgeBasesLambdaAPI { 1 usage new*
16     public CreateKnowledgeBasesLambdaAPI(InfrastructureStack myStack) { 1 usage new*
17         cdkKnowledgeBasesLambdaAPI(myStack);
18     }
19
20     public void cdkKnowledgeBasesLambdaAPI(InfrastructureStack myStack){ 1 usage new*
21         Role role = createLambdaRole(myStack);
22
23         Function createItemFunction = new Function(myStack, id: "knowledgeBasesFunction", getLambdaFunctionProps(role,
24             RestApi api = new RestApi(myStack, id: "KnowbaseApi", RestApiProps.builder().restApiName("KnowledgeBases Service").build());
25             IResource items = api.getRoot().addResource( s: "Knowbase");
26             Integration createOneIntegration = new LambdaIntegration(createItemFunction);
27             items.addMethod( s: "GET", createOneIntegration);
28             addCorsOptions(items);
29     }
30
31
32     private void addCorsOptions(IResource item) { 1 usage new*
33         List<MethodResponse> methodResponses = new ArrayList<>();
34
35         Map<String, Boolean> responseParameters = new HashMap<>();
36         responseParameters.put("method.response.header.Access-Control-Allow-Headers", Boolean.TRUE);
37         responseParameters.put("method.response.header.Access-Control-Allow-Methods", Boolean.TRUE);
38         responseParameters.put("method.response.header.Access-Control-Allow-Credentials", Boolean.TRUE);
39         responseParameters.put("method.response.header.Access-Control-Allow-Origin", Boolean.TRUE);
40         methodResponses.add(MethodResponse.builder()
41             .responseParameters(responseParameters)
42             .statusCode("200")
43             .build());
44         MethodOptions methodOptions = MethodOptions.builder()
```

AWS CDK INFRA ROLES

CreateKnowledgeBasesLambdaAPI.java

```
15  public class CreateKnowledgeBasesLambdaAPI { 1usage new *
16
17      public Role createLambdaRole(InfrastructureStack myStack){ 1usage new *
18          PolicyStatement statement2 = PolicyStatement.Builder.create()
19              .effect(Effect.ALLOW)
20              .actions(Arrays.asList(new String[] {"logs:CreateLogGroup","logs:CreateLogStream","logs:PutLogEvents"}))
21              .resources(Arrays.asList(new String[] {"arn:aws:logs:*:*:*"})).build();
22
23
24
25
26
27
28      PolicyStatement statement3Lambda = PolicyStatement.Builder.create()
29          .effect(Effect.ALLOW)
30          .actions(NSArray.asList(new String[] {"bedrock:InvokeModel"}))
31          .resources(NSArray.asList(new String[] {"arn:aws:bedrock:*:*:*"})).build();
32
33
34
35
36
37      PolicyStatement statementQueryToKnowledgeBases = PolicyStatement.Builder.create()
38          .effect(Effect.ALLOW)
39          .actions(NSArray.asList(new String[] {"bedrock:Retrieve","bedrock:RetrieveAndGenerate"}))
40          .resources(NSArray.asList(new String[] {"arn:aws:bedrock:*:*:*"})).build();
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
```

AWS CDK INFRA LAMBDA

```
15  public class CreateKnowledgeBasesLambdaAPI { 1 usage new *
16
17      private void addCorsOptions(IResource item) { 1 usage new *
18          .passthroughBehavior(PassthroughBehavior.NEVER)
19          .requestTemplates(requestTemplate)
20          .build();
21
22
23          item.addMethod( s: "OPTIONS", methodIntegration, methodOptions);
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71 @private FunctionProps getLambdaFunctionProps(Role role, String nameFunction, String extension) { 1 usage new *
72     try{
73         String lambdaContent = readFileAsString("./lambda/".concat(nameFunction).concat( str: ".") .concat(extension));
74         return FunctionProps.builder()
75             .description("Lambda ".concat(nameFunction))
76             .code(Code.fromInline(lambdaContent))
77             .handler("index.lambda_handler")
78             .role(role)
79             .memorySize(512)
80             .timeout(Duration.seconds( amount: 300))
81             .runtime(Runtime.PYTHON_3_13)
82             .build();
83
84     } catch (Exception e) {
85         e.printStackTrace();
86     }
87
88     return null;
89 }
```

LAMBDA FUNCTION PYTHON

```
# https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/bedrock-agent-runtime/client/retrieve_and_generate.html
input_data = {
    'input': {
        'text': str(query)
    },
    'retrieveAndGenerateConfiguration': {
        'knowledgeBaseConfiguration': {
            'knowledgeBaseId': 'MC6K0RI4A1',
            'modelArn': 'arn:aws:bedrock:us-east-1::foundation-model/anthropic.claude-3-sonnet-20240229-v1:0'
        },
        'type': 'KNOWLEDGE_BASE'
    }
}

# SessionId
if session_id:
    input_data['sessionId'] = session_id

print(input_data)

response = {
    'statusCode': 200,
    'body': json.dumps(client.retrieve_and_generate(**input_data)),
    'headers': { 'Content-Type': 'application/json', 'Access-Control-Allow-Origin': '*' },
}
return response
```

AWS CDK INFRA COMMANDS

```
weder@weder:~/Documentos/Projects/CampusParty2024/infra/infra-cdk-Kb-java/my-cdk$ cdk synth
```

1

Resources:

 LambdaIAMRole149A597F:

 Type: AWS::IAM::Role

 Properties:

```
weder@weder:~/Documentos/Projects/CampusParty2024/infra/infra-cdk-Kb-java/my-cdk$ cdk bootstrap
```

2

 💡 Bootstrapping environment aws://710304818543/us-east-1...

Trusted accounts for deployment: (none)

Trusted accounts for lookup: (none)

Using default execution policy of 'arn:aws:iam::aws:policy/AdministratorAccess'. Pass '--cloudformation-execution-policies' to customize.

✓ Environment aws://710304818543/us-east-1 bootstrapped (no changes).

```
weder@weder:~/Documentos/Projects/CampusParty2024/infra/infra-cdk-Kb-java/my-cdk$ cdk deploy
```

3

* Synthesis time: 4.06s

dev: start: Building 2c1e37df4caf7385bd8153f4ff1f0f811f245a8f7d01308bbbc2a218e6539c8c:710304818543-us-east-1

dev: success: Built 2c1e37df4caf7385bd8153f4ff1f0f811f245a8f7d01308bbbc2a218e6539c8c:710304818543-us-east-1

dev: start: Publishing 2c1e37df4caf7385bd8153f4ff1f0f811f245a8f7d01308bbbc2a218e6539c8c:710304818543-us-east-1

dev: success: Published 2c1e37df4caf7385bd8153f4ff1f0f811f245a8f7d01308bbbc2a218e6539c8c:710304818543-us-east-1

✓ dev

* Deployment time: 53.93s

Outputs:

dev.KnowbaseApiEndpoint6FB98E75 = <https://jma4p628l8.execute-api.us-east-1.amazonaws.com/prod>

Stack ARN:

arn:aws:cloudformation:us-east-1:710304818543:stack/dev/d9c64f10-ac53-11ef-89e2-0e2da9294735

* Total time: 57.99s

4

dev-knowledgeBasesFunctionFA4B52A5-SwbFqQ3Zc90k

[Throttle](#) [Copy ARN](#)[Actions ▾](#)

▼ Function overview

[Info](#)[Export to Infrastructure Composer](#)[Download ▾](#)[Diagram](#)[Template](#)

dev-
knowledgeBasesFunctionFA4B52A5-
SwbFqQ3Zc90k



Layers

(0)



API Gateway

(2)

[+ Add destination](#)[+ Add trigger](#)

Description

Lambda knowledgeBasesFunction

Last modified

1 hour ago

Function ARN

 [arn:aws:lambda:us-east-1:XXXXXXXXXX:function:dev-knowledgeBasesFunctionFA4B52A5-SwbFqQ3Zc90k](#)

Application

dev

Function URL

[Info](#)

-

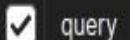
GET

[https://\[REDACTED\].execute-api.us-east-1.amazonaws.com/prod/Knowbase?query=Why+did+Tony+say+to+himself,+are+there+many+technical+hurdles+to+overcome+in+new+neurointerfaces?](https://[REDACTED].execute-api.us-east-1.amazonaws.com/prod/Knowbase?query=Why+did+Tony+say+to+himself,+are+there+many+technical+hurdles+to+overcome+in+new+neurointerfaces?)

Send ▼

Params • Authorization Headers (6) Body Scripts Tests Settings

Cookies



query

Why did Tony say to himself, are there many technical hurdles to overcome i...

Key

Value

Description

Body Cookies Headers (12) Test Results

Status: 200 OK Time: 6.63 s Size: 3.62 KB Save as example

Pretty Raw Preview Visualize JSON

□ Q

```
1 {  
2 >   "ResponseMetadata": {"  
3 >     },  
4 >   "citations": [  
5 >     ],  
6 >   "output": {  
7 >     "text": "Tony said to himself \\\"There are many technical hurdles to overcome in the new neurointerfaces\\\" because he was working on developing advanced neurointerface systems to  
8 >       control his Iron Man suit and integrate it with other systems. The neurointerfaces would allow him to control the suit directly with his thoughts and nervous system, achieving  
9 >       \\\"immediate control\\\". However, this cutting-edge technology presented many technical challenges that still needed to be overcome."  
10 >    },  
11 >   "sessionId": "8b04895f-7d9a-4812-95ea-5db9a9583a07"  
12 }
```

Citations

```
"citations": [
    {
        "generatedResponsePart": {
            "textResponsePart": {
                "span": {
                    "end": 470,
                    "start": 0
                },
                "text": "Tony said to himself \"There are many technical hurdles to overcome in the new neurointerfaces\" because he was working on developing advanced neurointerface systems to control his Iron Man suit and integrate it with other systems. The neurointerfaces would allow him to control the suit directly with his thoughts and nervous system, achieving \"immediate control\". However, this cutting-edge technology presented many technical challenges that still needed to be overcome."
            }
        }
    }
],
"retrievedReferences": [
    {
        "content": {
            "text": "E se ele pudesse acrescentar sistemas de controle de interface multipla, em tempo real, que empregassem mais aplicativos do que somente a armadura? Primeiro o mais importante, Tony disse a si mesmo.Existem muitos obstáculos técnicos para superar nas novas neurointerfaces. Ele deu meia-volta, pondo um fim proposital à sua irritação; deu um tapinha nos ombros de Rhodey e Fury e ergueu os braços feito um homem de negócios, apresentando as reluzentes novas instalações. – Não aceitei nem um centavo de isenção de impostos aqui, desde que a questão das parcerias público-privadas foi mencionada. Agora deixe-me explicar do que se trata. O prédio principal do laboratório era em formato de L, com uma área de testes exterior acoplada ao canto interno do L. A perna mais curta era ocupada por laboratórios menores e mais simples, para a confecção dos componentes individuais do sistema de neurointerface. A perna mais longa era um único piso de produção e testes, cobrindo quase 150 metros ao longo da margem de um lago que delineava o limite ao norte da propriedade."
        },
        "location": {
            "s3Location": {
                "uri": "s3://datasourcews2024/marvel_min/Marvel Vol. 05 Homem de ferro virus.pdf"
            },
            "type": "S3"
        },
        "metadata": {
            "x-amz-bedrock-kb-source-uri": "s3://datasourcews2024/marvel_min/Marvel Vol. 05 Homem de ferro virus.pdf",
            "x-amz-bedrock-kb-document-page-number": 19.0,
            "x-amz-bedrock-kb-chunk-id": "1%3A0%3AAxuiapMBhtRLgI425ycp",
            "x-amz-bedrock-kb-document-page-count": 19
        }
    }
]
```

Code Editor Function

← → 🔍 dev-knowledgeBasesFunctionFA4B52A5-SwbFqQ3Zc90k

EXPLORER

index.py

```
index.py
1 import boto3
2 import json
3
4 def lambda_handler(event, context):
5     client = boto3.client('bedrock-agent-runtime', region_name='us-east-1')
6
7     session_id = event.get('sessionId', None)
8
9     query_params = event.get('queryStringParameters', {})
10    if 'query' not in query_params:
11        return {Amazon Q Tip 1/3: Start typing to get suggestions ([ESC] to exit)
12            'statusCode': 400,
13            'body': json.dumps({'error': 'Query parameter is missing'}),
14            'headers': {'Content-Type': 'application/json', 'Access-Control-Allow-Origin': '*'}
15        }
16
17    # Extraiendo a consulta do evento
18    query = query_params['query']
```

PROBLEMS OUTPUT CODE REFERENCE LOG TERMINAL

Execution Results

{
da margem de um lago que defineava o limite ao norte da propriedade. "}, "location": {"S3LOCATION":
{"uri": "s3://datasourcews2024/marvel_min/Marvel Vol. 05 Homem de ferro vi\u0301rus.pdf"},
"type": "S3"}, "metadata": {"x-amz-bedrock-kb-source-uri": "s3://datasourcews2024/marvel_min/
Marvel Vol. 05 Homem de ferro vi\u0301rus.pdf", "x-amz-bedrock-kb-document-page-number": 19.0,
"x-amz-bedrock-kb-chunk-id": "1%3A0%3AAxuiapMBhtRLgI425ycp", "x-amz-bedrock-kb-data-source-id":
"W1BC45STUJ"}}], "output": {"text": "Tony said to himself \\\\"There are many technical hurdles
to overcome in the new neurointerfaces\\\\" because he was working on developing advanced neurointerface
systems to control his Iron Man suit and integrate it with other systems. The neurointerfaces would
allow him to control the suit directly with his thoughts and nervous system, achieving \\\\"immediate
control\\\\". However, this cutting-edge technology presented many technical challenges that still
needed to be overcome \\\"1 \"sessionId\\\" - \\\"8882bc42-0fd9-45c4-a367-657d7c246ed0\\\""}
Ln 11, Col 17 Spaces: 4 UTF-8 LF Python Lambda Layout: us

DEPLOY

Deploy (Ctrl+Shift+U)

Test (Ctrl+Shift+I)

TEST EVENTS

Private saved events

save



ENVIRONMENT VARIABLES

Next Steps



Amazon Bedrock Studio

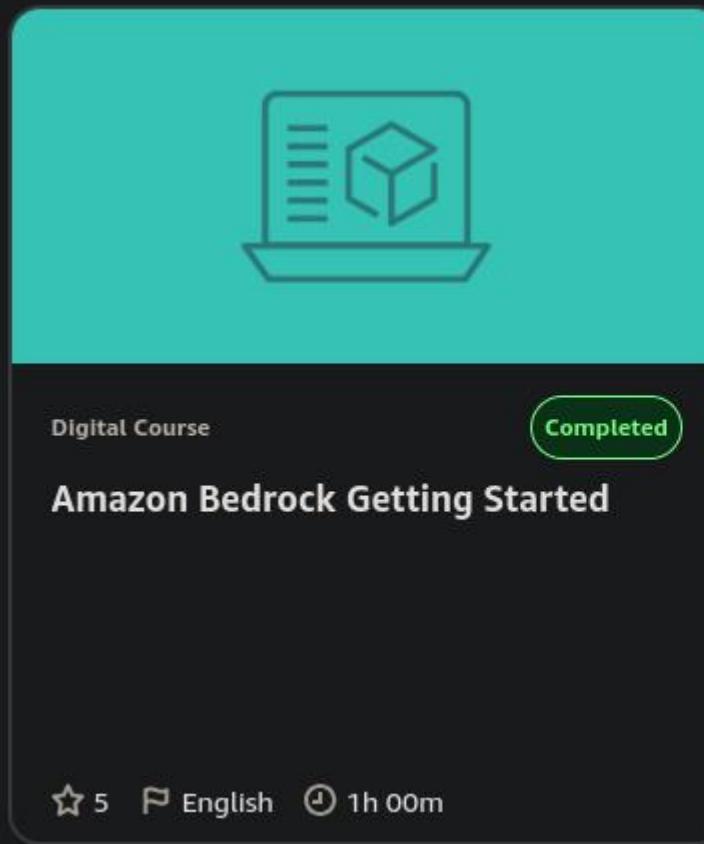


The image shows the landing page of the Amazon Bedrock Studio. It features a dark background with a teal circular icon containing a stylized AI head. Below the icon, the text "Welcome to Amazon Bedrock Studio" is displayed in large, white, sans-serif font. A descriptive paragraph follows, explaining the purpose of the studio: "Use a playground to experiment with Amazon Bedrock models. Create and collaborate on apps that integrate your own data with robust capabilities such as responsible AI guardrails, functions, and knowledge bases." At the bottom, there are two calls-to-action: a teal button labeled "Sign in with SSO" and a link labeled "Learn More" with a small icon.

<https://aws.amazon.com/bedrock/studio/>

Building Generative AI Applications

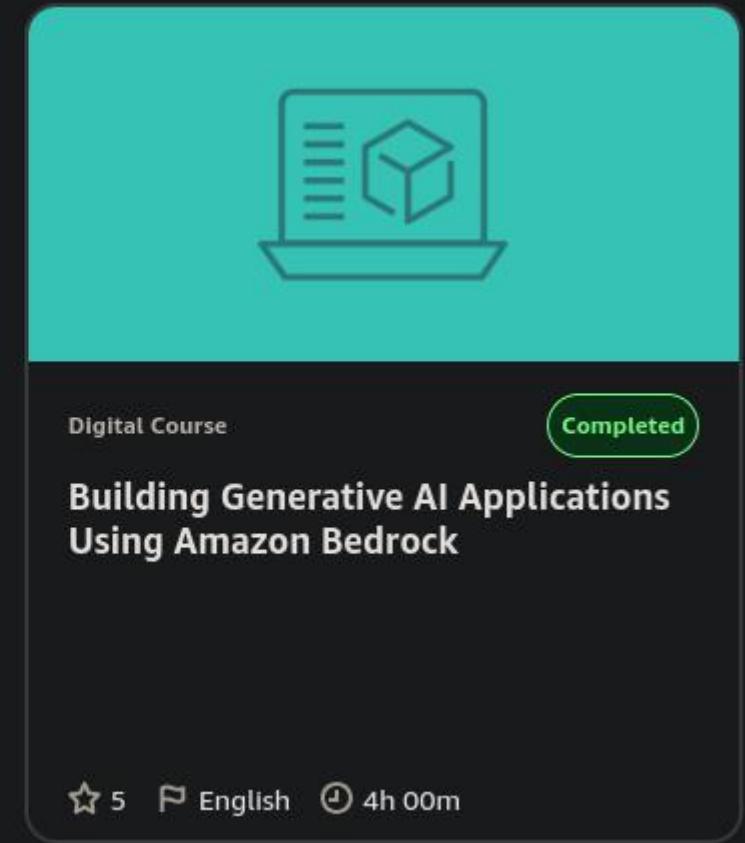
AWS Skill Builder



Digital Course Completed

Amazon Bedrock Getting Started

☆ 5 ⚒ English ⏱ 1h 00m

This block displays a course card for "Amazon Bedrock Getting Started". It features a teal header with a laptop icon containing a 3D cube. Below the header, the text "Digital Course" is followed by a green "Completed" badge. The course title "Amazon Bedrock Getting Started" is prominently displayed in bold. At the bottom, it shows a rating of 5 stars, the language as English, and a duration of 1 hour and 0 minutes.

Digital Course Completed

**Building Generative AI Applications
Using Amazon Bedrock**

☆ 5 ⚒ English ⏱ 4h 00m

This block displays a course card for "Building Generative AI Applications Using Amazon Bedrock". It has a similar teal header with a laptop icon containing a 3D cube. The text "Digital Course" and a green "Completed" badge are present. The course title is split into two lines. At the bottom, it shows a rating of 5 stars, the language as English, and a duration of 4 hours and 0 minutes.

<https://skillbuilder.aws/search?searchText=bedrock+badge&page=1&isValidSearchText=true>

Want to Try?

Chat/text playground

Amazon Bedrock > Chat/text playground

Mode Chat ▾

Compare mode Load examples :

Configurations

Titan Text G1 - ... v1 On-demand

Randomness and diversity

Temperature: 0.7

Top P: 0.9

Length

Response length: 512

Stop sequences
Enter a stop sequence
Only | allowed (maximum 20 characters)

Guardrail

Manage guardrails

Qual é o melhor modelo para gerar imagem usando bedrock

When it comes to generating images using Bedrock, there are several models that you can consider. Here are some popular models that are widely used for image generation tasks:

1. Generative Adversarial Networks (GANs): GANs are a type of neural network architecture that consists of a generator and a discriminator. The generator produces synthetic images, while the discriminator tries to distinguish between real and generated images. GANs have been successful in generating high-quality images in various domains, such as faces, landscapes, and even artistic styles.
2. Variational Autoencoders (VAEs): VAEs are another type of neural network model that can generate images. They consist of an encoder and a decoder. The encoder maps the input image to a lower-dimensional latent space, while the decoder reconstructs the image from the latent representation. VAEs can generate diverse images by sampling from the latent space.
3. PixelRNN/PixelCNN: These models are specifically designed for image generation tasks. They use recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to model the conditional probability of pixels in an image given the previous pixels.

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.

X | ↵

Choose files

The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx. Maximum 5 files, 4.5 MB each.

Run

Chat/text playground

Want to Try?

Artificial Intelligence Integration



Artificial Intelligence Integration: Using Spring AI with OpenAI (ChatGPT) and AWS Bedrock to Deliver Efficient APIs

The background features a dense grid of green binary digits (0s and 1s). Overlaid on this is a stylized blue illustration of an AI microchip with the letters "AI" in its center, connected by lines to various nodes and a circuit board below it.A green square icon containing the white OpenAI logo (two interlocking circles).

<https://github.com/weder96/JoinCommunity2024>

Want to Try? AWS Community



aws

Search for content

Ctrl K

Home Tags Featured Spaces

Amazon Q

Cost Optimization

DevOps

Generative AI

Kubernetes

Livestreams

Resilience

SaaS

Training and Certification

Community Programs

AWS Heroes

AWS Community Builders

AWS User Groups

AWS Cloud Clubs for Students

Builder ID Benefits

Site Terms, Privacy, and more.

Generative AI

Your space for generative AI tools, guides, and technology.

The Secrets of Agents

Are agents and agentic workflows taking over as the leading use case for generative AI? Let's take a look at agents...



Corneliu CROITORU

Beyond Auto-Replies: Building an AI-Powered E-commerce Support system

Implementation of an AI-driven multi-agent system for automated e-commerce customer...

Sep 12 • ai-agents



Mike Chambers

Amazon Employee

AI running its own code! : Agentic Code Interpreter

Public Preview: Code Interpreter for Agents for Amazon Bedrock - With sample code.

Jul 17 • generative-ai



Mike Chambers

Amazon Employee

Solving LLM Amnesia - Cross Session Memory

Public Preview: Long Term Memory for Agents on Amazon Bedrock

Jul 17 • generative-ai

<https://community.aws/generative-ai>

Getting started



**Get started with
Amazon Bedrock**



**Discover features with
a step-by-step tutorial**



**Dive deep with a
hands-on workshop**

GenAI Roadshow - Virtual



Amazon Web Services Latin America



- [/aws.amazon.com](http://aws.amazon.com)
- [/amazonwebservices](https://www.facebook.com/AWSLatam)
- [/AWSLatam](https://twitter.com/AWSLatam)
- [/AWSBrasil](https://www.linkedin.com/company/amazon-web-services)
- [/company/amazon-web-services](https://www.linkedin.com/company/amazon-web-services)



Want to Try? CDK - Workshop



AWS CDK
Workshop

Search

English

Prerequisites

- TypeScript Workshop
- Python Workshop
- .NET Workshop
- Java Workshop
- Go Workshop
- Construct Hub
- Congrats!

Submit a correction

CDK on GitHub

CDK on StackOverflow

CDK on Gitter

Welcome Developers!

Hey there, and thanks for joining us! Hope you can't wait to play with this new thing we call the "AWS Cloud Development Kit" or in short, the AWS CDK.

The AWS CDK is a new software development framework from AWS with the sole purpose of making it fun and easy to define cloud infrastructure in your favorite programming language and deploy it using AWS CloudFormation.

So what are we going to build? Nothing too fancy...

We'll spend some time setting up your development environment and learning a little about how to work with the CDK Toolkit to deploy your app to an AWS environment.

Then, you'll write a little "Hello, world" Lambda function and front it with an API Gateway endpoint so users can call it via an HTTP request.

Next, we'll introduce the powerful concept of **CDK constructs**. Constructs allow you to bundle up a bunch of infrastructure into reusable components which anyone can compose into their apps. We'll walk you through writing your own construct.



<https://cdkworkshop.com/>

Want to Try?

CDK Patterns



aws
CDK Patterns

[About](#)[Find A Pattern](#)

CDK Patterns is more than "just AWS CDK examples"

Check Out Our 4 Content Distribution Platforms:

[Star 2,130](#)[Follow](#)[YouTube](#)[The Practical Dev](#)

<https://cdkpatterns.com/>

Demo CDK

Clone S3 AWS with CDK

Campus Party 2022

Home / Buckets

Buckets (4) Info Buckets are containers for data stored in S3. Learn more

Name	AWS Region	Access	Owner	Creation Date
data-lake-crime-results	US East (N. Virginia) us-east-1	Não informado	wmarians	27-04-2022 09:54:41
data-lake-crimes	US East (N. Virginia) us-east-1	Não informado	wmarians	27-04-2022 09:51:08
heroes-marvel	US East (N. Virginia) us-east-1	Não informado	wmarians	10-06-2022 15:36:53
heroes-marvel-iron-men	US East (N. Virginia) us-east-1	Não informado	wmarians	16-06-2022 15:17:48

<< < 1 > >> Showing 1 to 4 of 4 buckets 10 ▾

Upload About: Seja Bem Vindo!!!

Success Info Warn Error Message Plain Success Message Plain Info Message Plain Warn Message Plain Error

aws-image-upload-wsousa / src / main / resources / application.properties



weder96

download e upload de arquivos

Code

Blame

7 lines (6 loc) · 168 Bytes

```
1 aws.access_key = ${value}
2 aws.secret_key = ${value}
3 aws.test = ${value}
4 spring.servlet.multipart.max-file-size=10MB
5
6 server.port=9090
7 springdoc.api-docs.path=/api-docs
```



<https://github.com/weder96/aws-image-upload-wsousa>

<https://github.com/weder96/presentationCampusParty2022>

Resources

#Links (BedRock)

<https://www.ranthebuilder.cloud/post/automating-api-calls-with-agents-for-amazon-bedrock-with-powertools>
<https://www.slightinsight.com/tech/developing-a-spring-boot-application-with-amazon-bedrock-api/>
<https://community.aws/content/2dhKdwY1kzhFTg9CTLbaJ9MmTN/build-generative-ai-applications-with-amazon-bedrock>
https://www.eficode.com/blog/building-ai-on-aws-bedrock-brings-brilliant-building-blocks?utm_campaign=AWS&utm_content=189323971&utm_medium=social&utm_source=twitter&hss_channel=tw-142208607
<https://cloudacademy.com/learning-paths/integrating-aws-services-with-langs-and-other-fms-14068/>
https://www.youtube.com/watch?v=CE_zrMvcuk&list=PLhr1KZpdzukfmv7jxvB0rL8SWoycA9TIM&index=5
<https://programadriano.medium.com/conhecendo-o-amazon-bedrock-c687c7e9777f>
<https://github.com/AWS-Cloud-Drops-Builders-Edition/show?tab=readme-ov-file>
<https://levelup.gitconnected.com/ai-powered-video-summarizer-with-amazon-bedrock-and-anthropic-claude-9f1832f397dc>
<https://catalog.workshops.aws/persona-based-access-genai-application/en-US/04-testing-application/01-updating-code>
<https://dev.to/aws-builders/have-fun-with-aws-partyrock-3755>
<https://dev.to/aws-builders/build-serverless-generative-ai-api-service-with-aws-lambda-and-amazon-bedrock-3abc>
<https://norahsakal.medium.com/how-to-use-aws-titans-ai-multimodal-embeddings-for-better-e-commerce-recommendations-b9f4adb60c02>

#Importante

<https://www.linkedin.com/pulse/aplica%C3%A7%C3%B5es-serverless-llm-com-amazon-bedrock-diogo-santos-yjo3f/>
<https://aws.amazon.com/pt/blogs/aws-brasil/crie-aplicativos-de-ia-generativa-usando-o-aws-step-functions-e-o-amazon-bedrock/>
<https://community.aws/generative-ai>
<https://aws.amazon.com/pt/what-is/retrieval-augmented-generation/>

Resources

#Links Spring AI OpenAI

<https://github.com/danvega/hello-gpt/blob/main/src/main/java/dev/danvega/Application.java>

https://www.youtube.com/watch?v=yyvjT0v3lpY&list=PLZV0a2jwt22uoDm3LNDFvN6i2cAVU_HTH

<https://tecnoblog.net/guias/4-sites-gratis-para-transcrever-video-do-youtube/>

#Spring AI Vector

<https://www.youtube.com/watch?v=azKntWC6d3w>

#Top

<https://github.com/kousen/openaidemo>

<https://www.youtube.com/watch?v=ZeH3bBKdqRU>

Resources

<https://cdkworkshop.com>

<https://github.com/aws-samples/aws-cdk-examples>

<https://docs.aws.amazon.com/cli/latest/userguide/getting-started-install.html>

<https://aws.amazon.com/pt/developer/language/java/>

<https://docs.aws.amazon.com/toolkit-for-jetbrains/latest/userguide/setup-toolkit.html>

<https://aws.amazon.com/pt/intellij/>

https://docs.aws.amazon.com/code-library/latest/ug/java_2_code_examples.html

https://docs.aws.amazon.com/pt_br/prescriptive-guidance/latest/patterns/deploy-a-ci-cd-pipeline-for-java-microservices-on-amazon-ecs.html

<https://docs.aws.amazon.com/lambda/latest/dg/lambda-java.html>

<https://aws.amazon.com/pt/blogs/compute/java-17-runtime-now-available-on-aws-lambda/>

<https://www.slideshare.net/AmazonWebServices/java-on-aws>

<https://www.jrebel.com/blog/aws-java-application-setup>

<https://www.slideshare.net/VadymKazulkin/adopting-java-for-the-serverless-world-at-jax-2022>

<https://towardsaws.com/deploy-spring-boot-application-to-aws-ec2-using-docker-f359e7ad2026>

<https://aws.amazon.com/pt/blogs/developer/stepfunctions-fluent-api/>

<https://aws.amazon.com/blogs/compute/java-17-runtime-now-available-on-aws-lambda/>

<https://docs.aws.amazon.com/lambda/latest/dg/snapstart.html>

About the Speaker

Weder Mariano de Sousa

Specialist Senior Java - GFT



Q & A



Technician **System Development**

Graduated **Computer Science**

Post Graduate in **Midias UFG**

Post Graduate in **Information Security**



GOJava



AWS User
Group Goiânia

- <https://www.linkedin.com/in/wedermarianodesousa/>
- <https://github.com/weder96>
- <https://twitter.com/weder96>
- <https://dev.to/weder96>



THANK YOU

Weder Sousa

 <https://www.linkedin.com/in/wedermarianodesousa/>

 <https://github.com/weder96>

 <https://twitter.com/weder96>

 <https://dev.to/weder96>