Guía Práctica: AWS Bedrock con RAG

Introducción

Esta guía te llevará paso a paso en la implementación de **AWS Bedrock** con **Retrieval-Augmented Generation (RAG)**, un enfoque que combina modelos de lenguaje con bases de conocimiento personalizadas para generar respuestas más precisas y contextuales.

¿Qué es RAG?

RAG (Retrieval-Augmented Generation) es una técnica que:

- **Recupera** información relevante de una base de conocimiento
- **Aumenta** el prompt del modelo con esta información
- **Genera** respuestas más precisas y fundamentadas

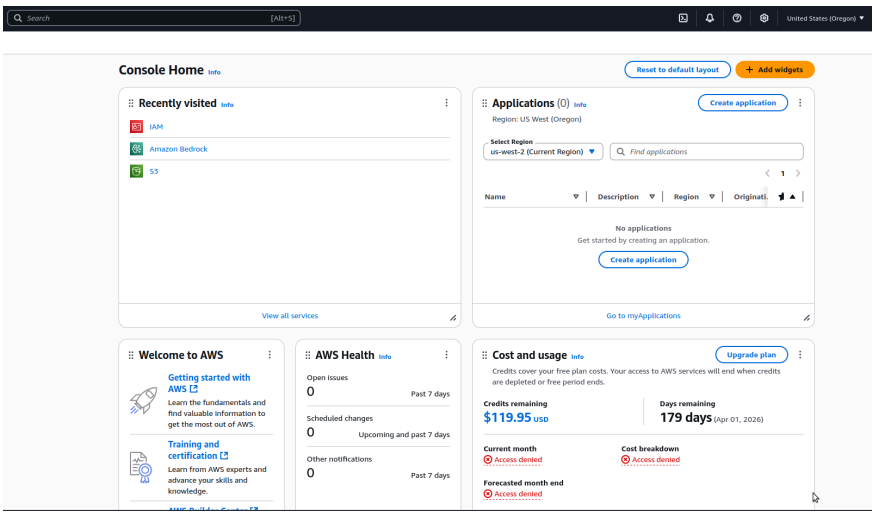
Paso 1: Configuración Inicial de AWS

1.1 Crear Cuenta AWS

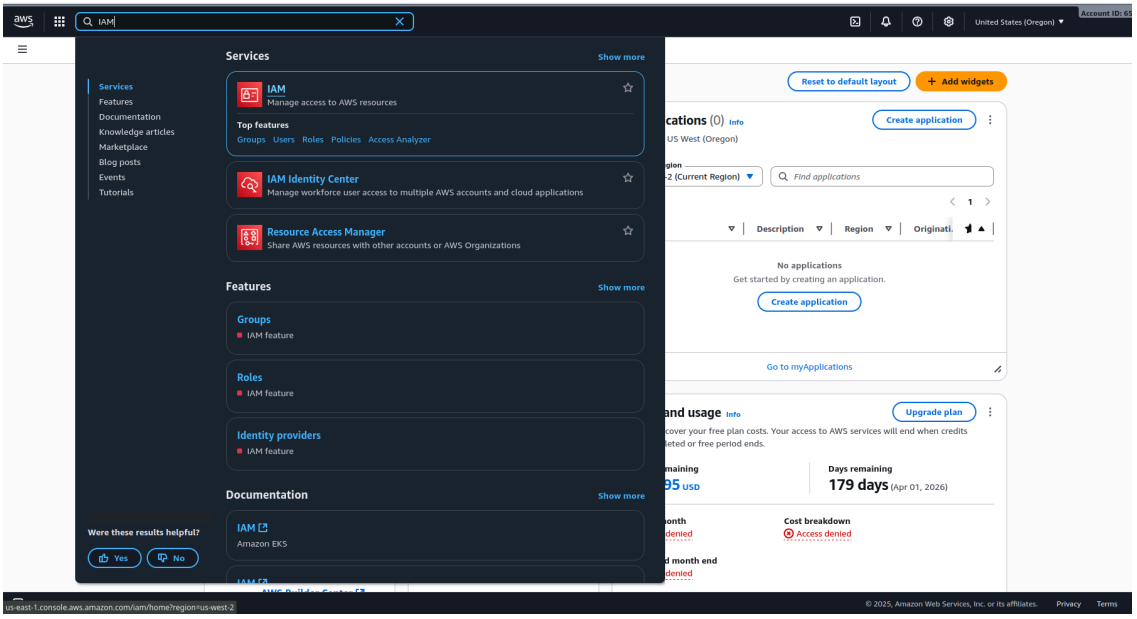
- Ve a https://aws.amazon.com
- Haz clic en "Create an AWS Account"
- Completa el proceso de registro (requiere tarjeta de crédito, pero hay tier gratuito)

1.2 Configurar IAM User

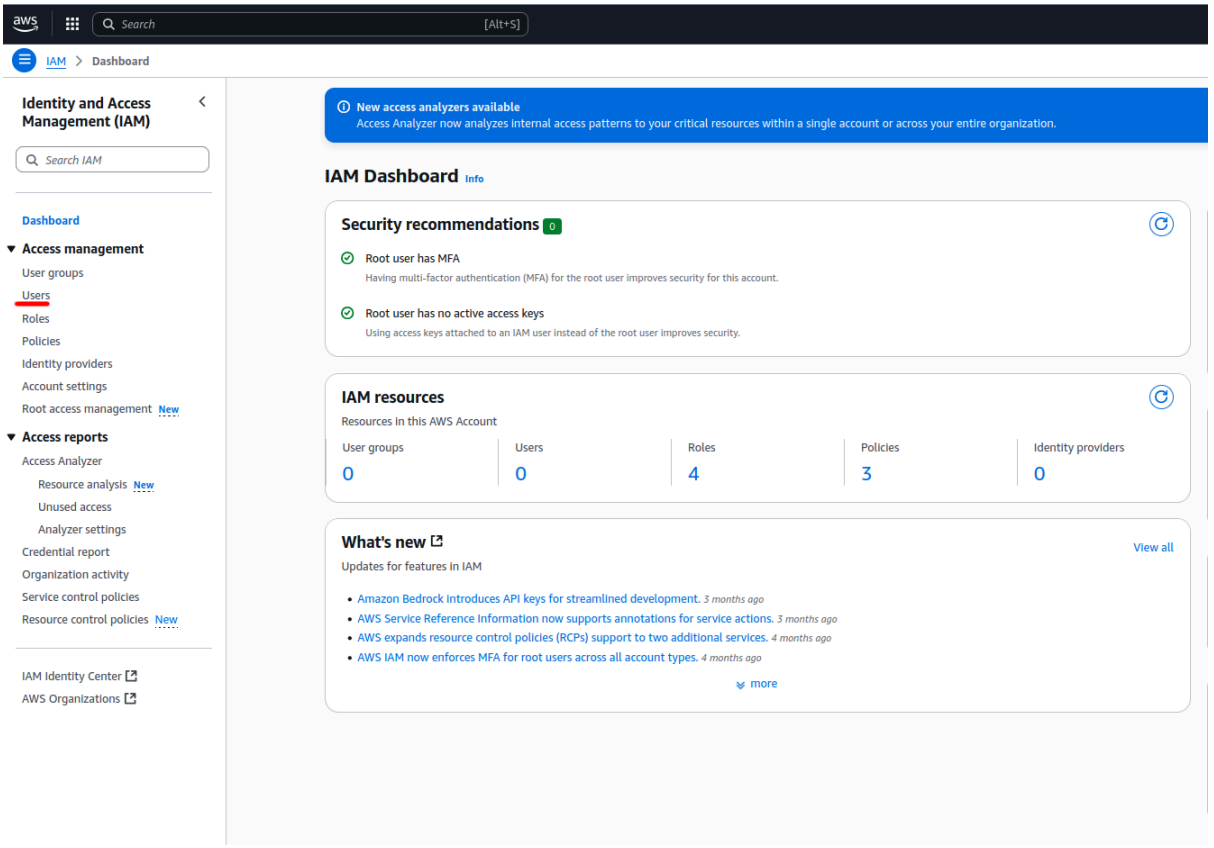
Una vez completado el anterior paso, seremos redirigidos aquí:



1. **Buscar IAM** en la barra de navegación superior



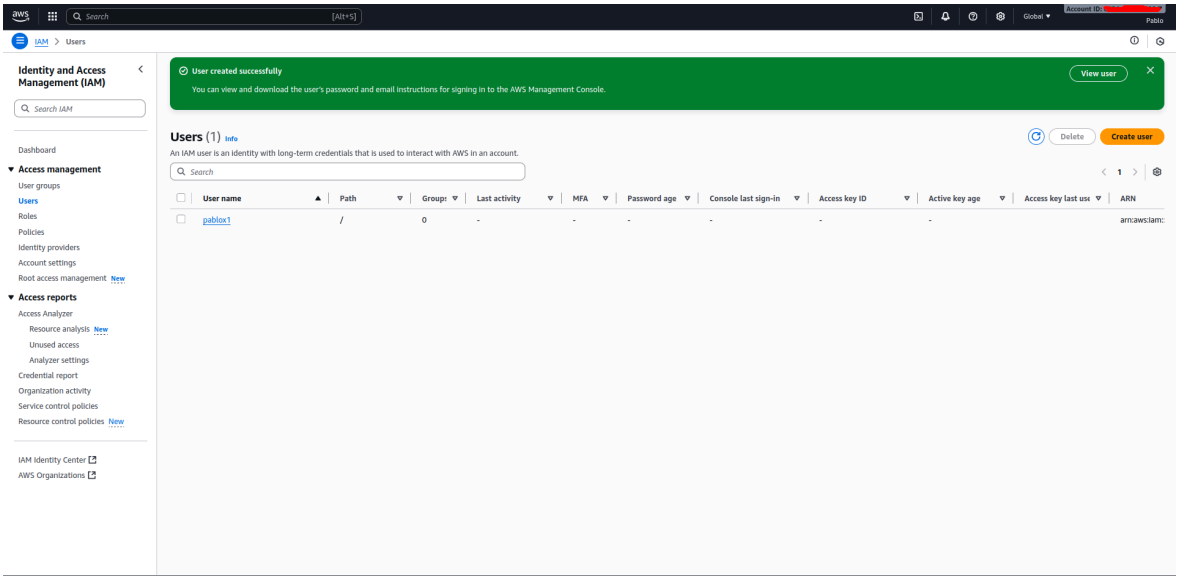
una vez hecho click serás redirigido a:



Ir a "users" en el menú lateral

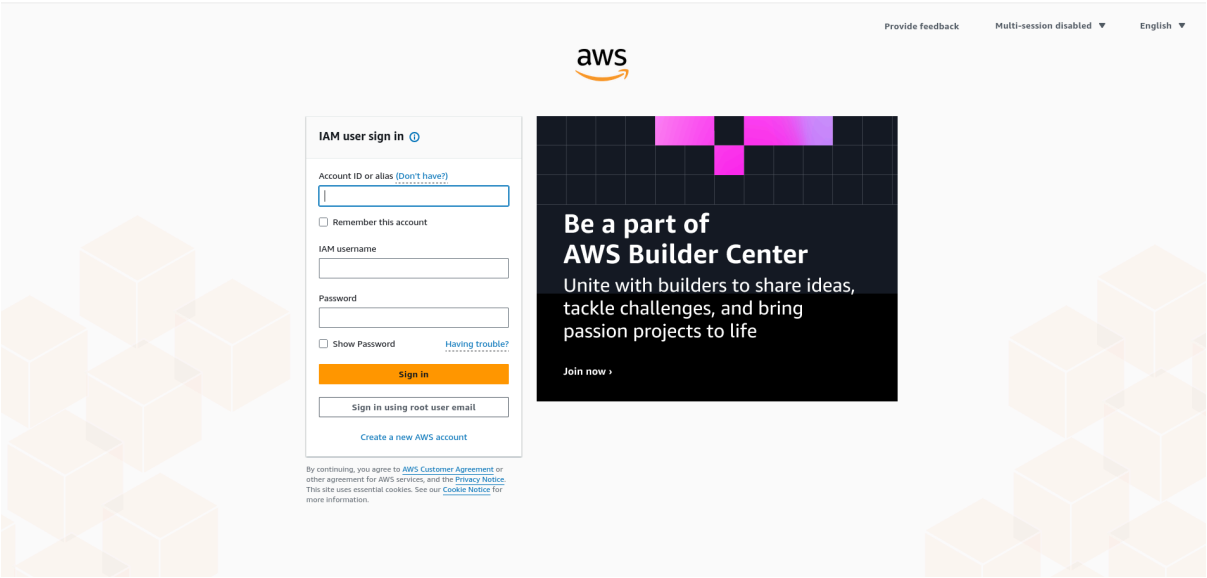
2. **Crear usuario:**
    - User name

3. **Permisos:**
    - Attach existing policies directly
    - Seleccionar: por simpleza le asignaremos el `AdministratorAccess` y ya
4. **Crear usuario** y **guardar las credenciales**
    - Hacer click en crear usuario y ser redirigido a:



5. Con el usuario creado, copiamos nuestro account-id que sale en la parte superior derecha que se puede observar en la anterior imagen, para posteriormente hacer sign out.

6. **Iniciar sesión con el usuario creado.** Se realiza el inicio de sesión



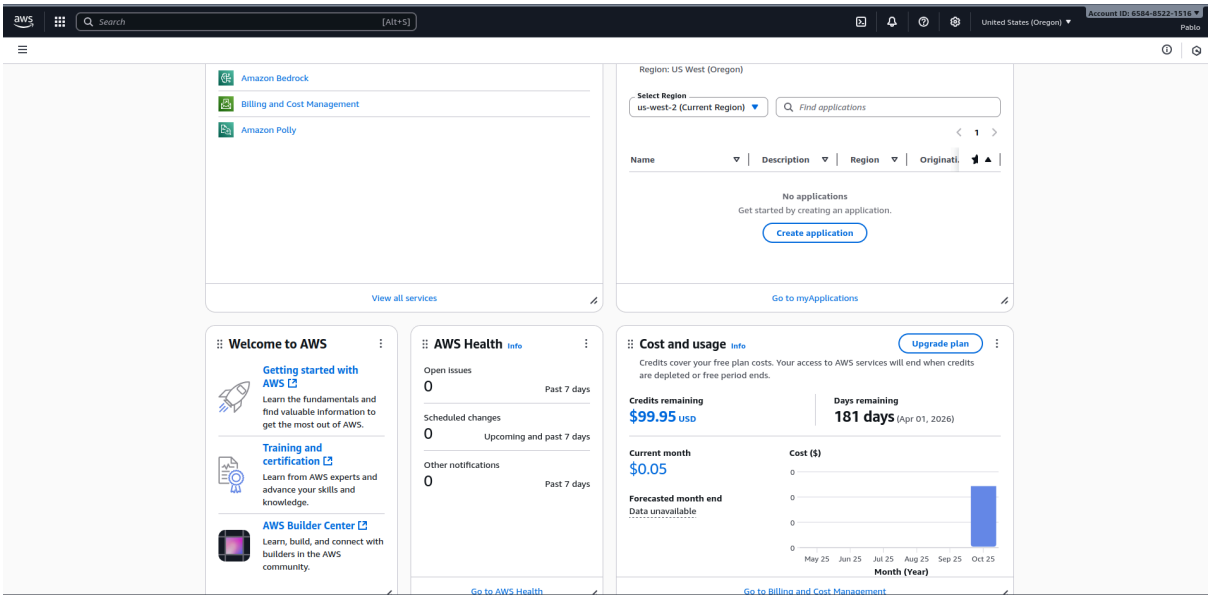
Nota: Inicialmente la password es la misma que la cuenta principal, pero cuando entres con el usuario que acabas de crear te pedirá cambiarla para ese usuario en específico.

Paso 2: Preparar Datos en Amazon S3

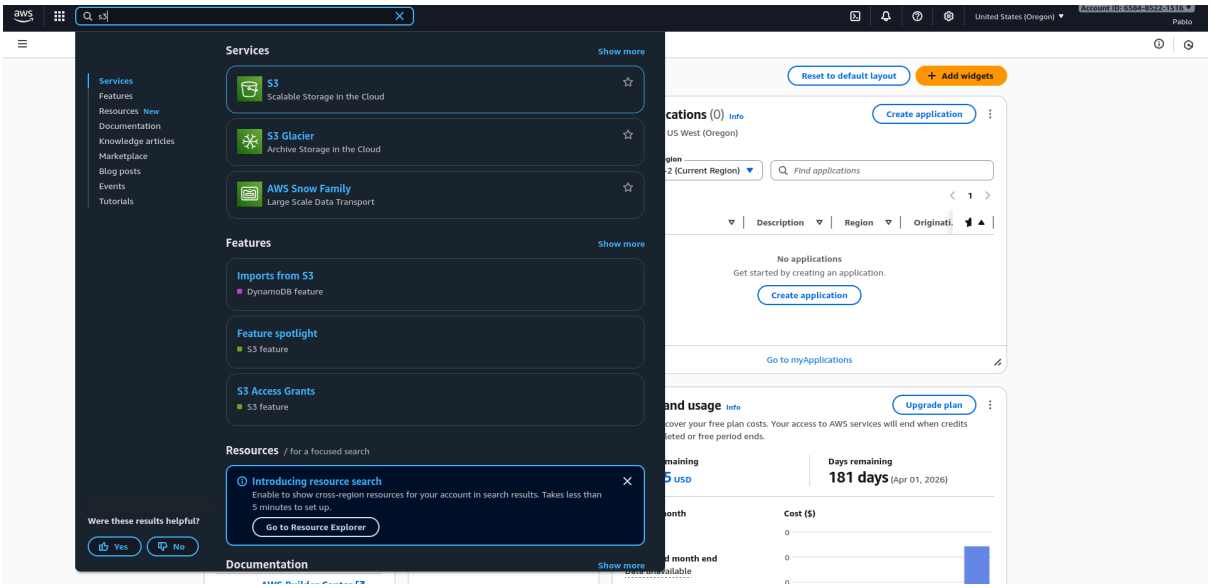
2.1 ¿Por qué usar S3?

Amazon S3 nos permite almacenar los documentos que nuestra base de conocimiento utilizará para el RAG.

1. Ubicarse en el home

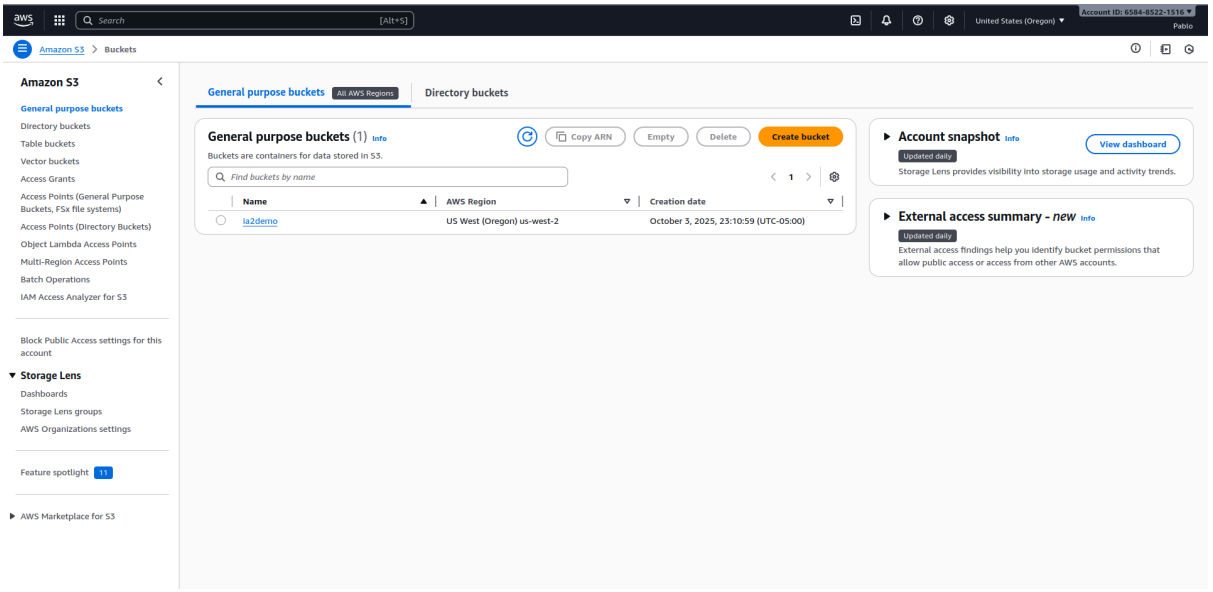


2. **Buscar "S3"** en la barra de navegación superior



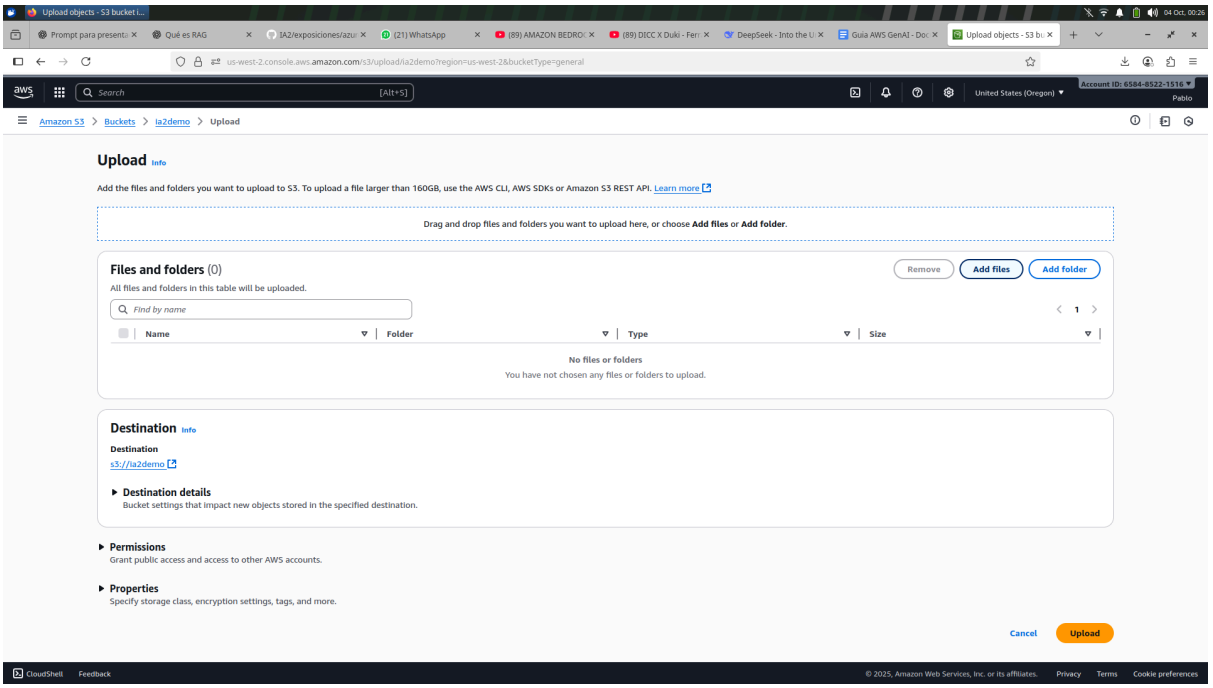
3. **Crear bucket:**

a. dejamos todos los campos por defecto, en este caso lo único que se cambió fue el nombre, se le colocó “ia2demo”
b. Después de crear el bucket, serás redireccionado a la lista de buckets.



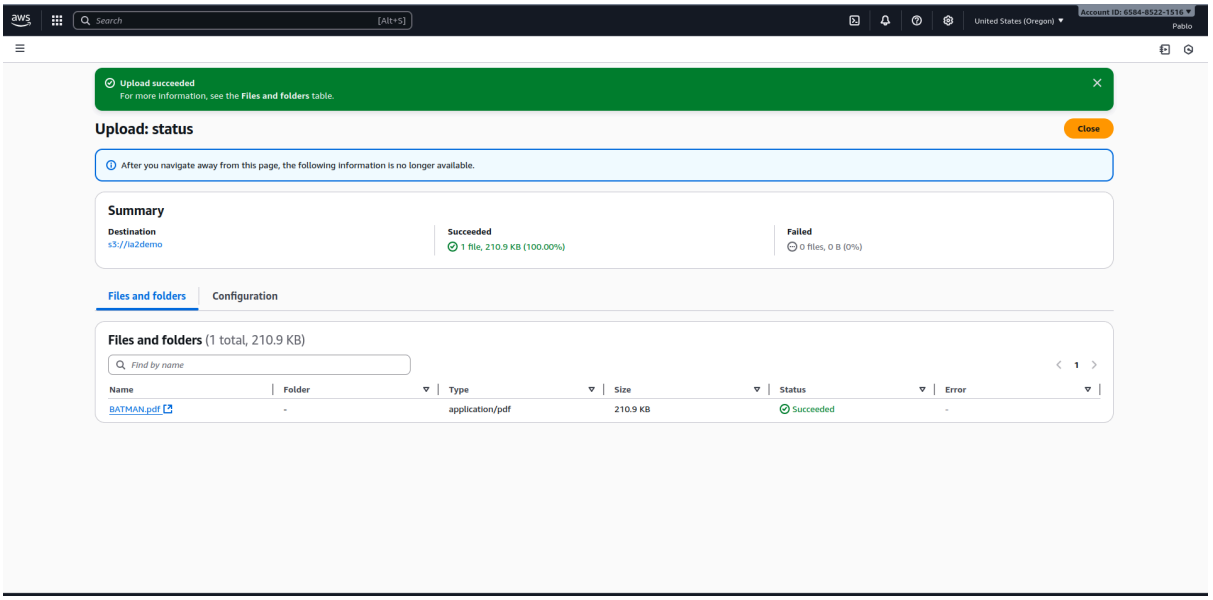
2.2 Subir Documentos

1. **Seleccionar el bucket** recién creado
2. **Hacer clic en "Upload" para ser redireccionado aquí:**



3. **Añadir archivos:** Seleccionar documentos PDF, TXT o DOC, en este caso añadiremos el pdf que se encuentra en el repositorio llamado “BATMAN”:

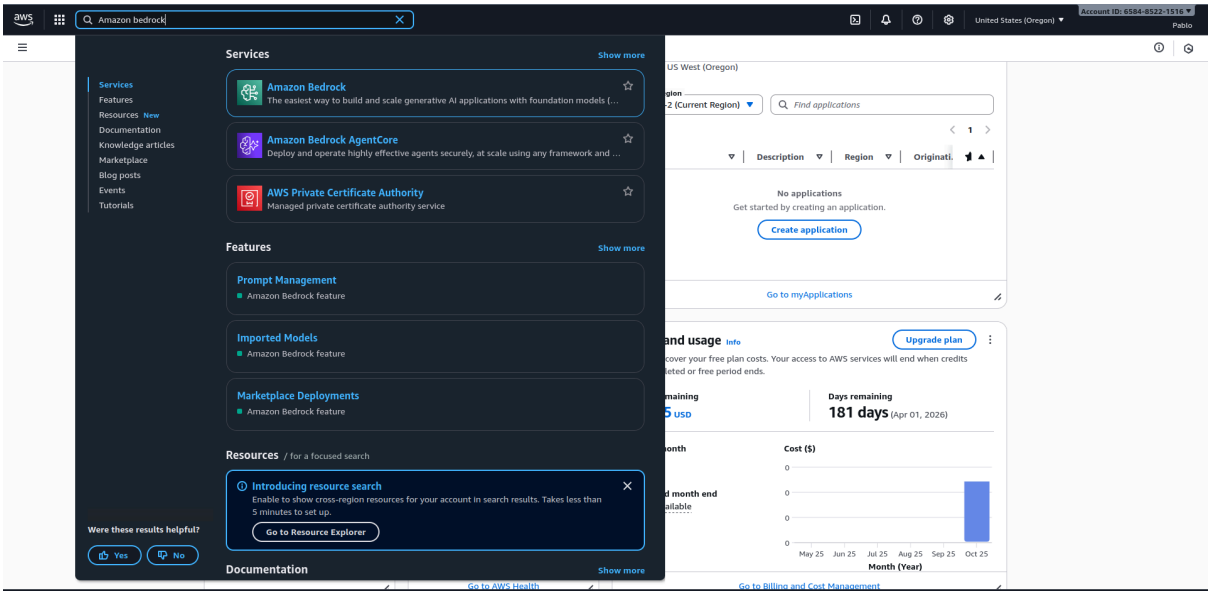
4. **Hacer clic en "Upload"** y debería de mostrarse esto:



Paso 3: Configurar AWS Bedrock

3.1 Acceder a Bedrock

- Volver al home
- Buscar **"Amazon Bedrock"** en la barra de navegación

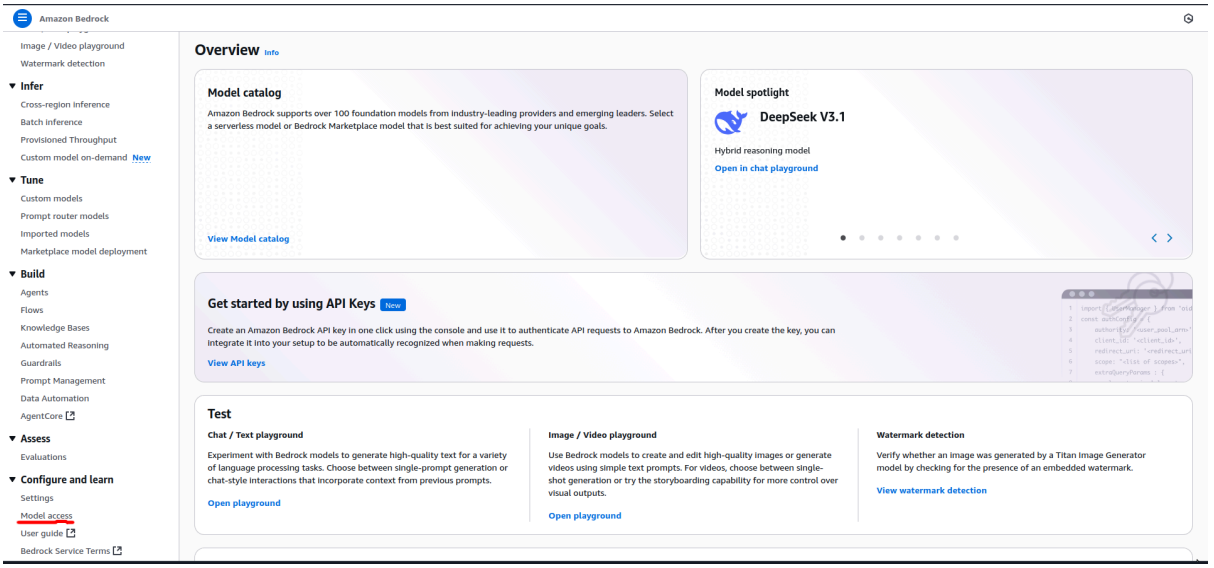


- Seleccionar el servicio

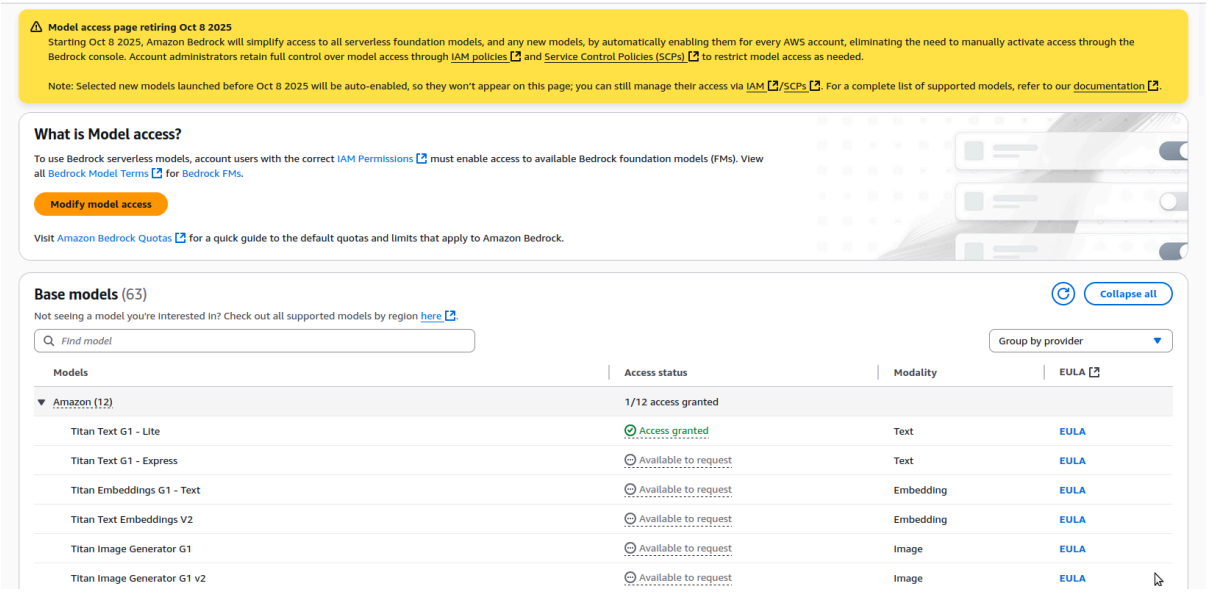
3.2 Habilitar Model Access

¿Por qué? AWS requiere que habilites explícitamente los modelos que quieres usar.

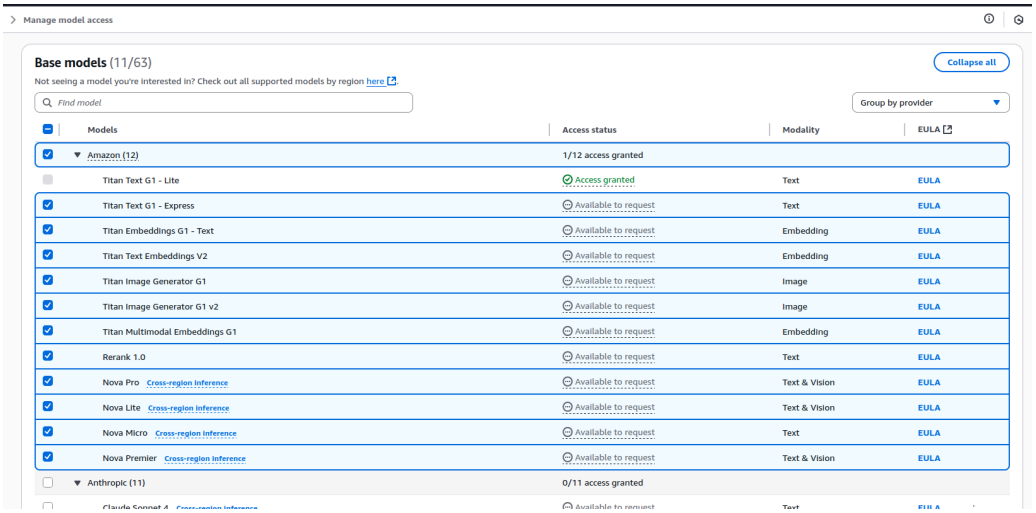
1. **Ir a "Model access"** en el menú lateral



2. **Hacer clic en "Modify model access"**



3. **Seleccionar modelos:**
    - Seleccionamos toda la familia de modelos de amazon:



4. **Hacer clic en “Next” y posteriormente en "Save changes"**

Nota: Hay que esperar unos minutos para que nos habiliten el acceso a estos modelos.

Paso 4: Crear Base de Conocimiento (Knowledge Base)

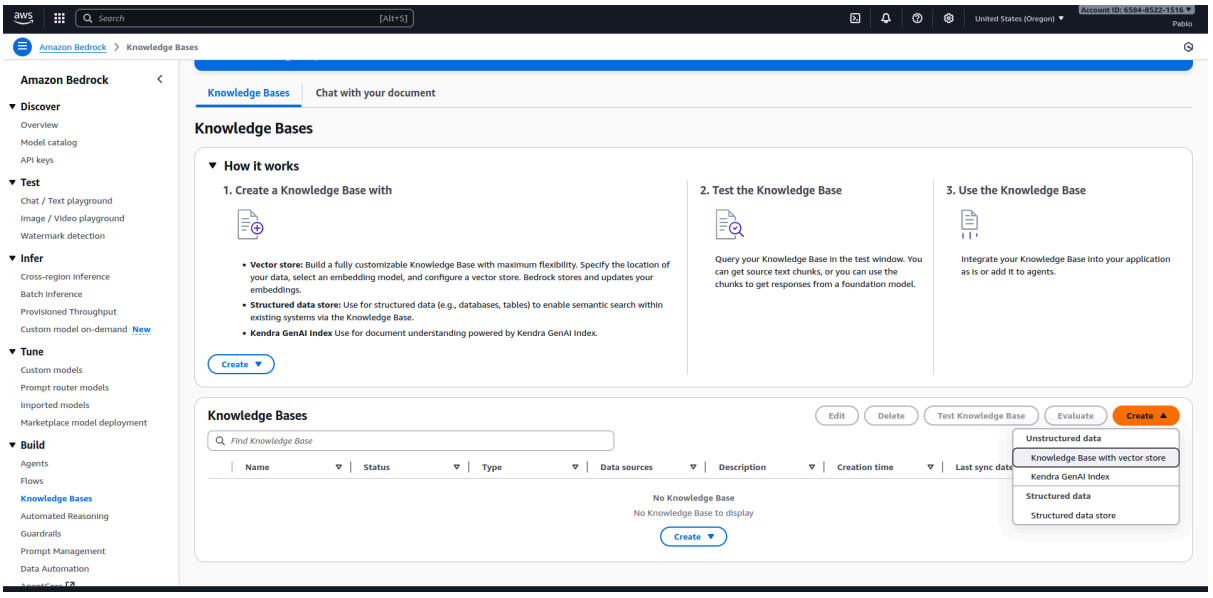
4.1 ¿Qué es una Knowledge Base?

Es el componente central del RAG que:

- **Almacena** tus documentos procesados
- **Crea embeddings** (representaciones vectoriales del texto)
- **Permite búsquedas semánticas** para encontrar información relevante

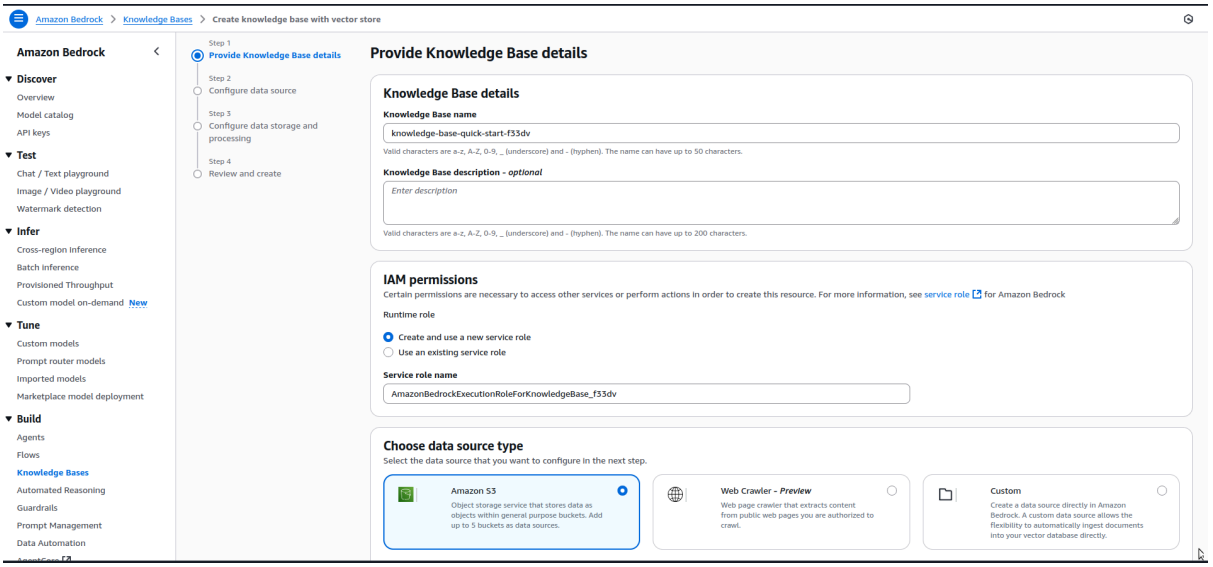
4.2 Proceso de Creación

1. **Ir a "Knowledge bases"** en el menú lateral
2. **Seleccionar "Create knowledge base" → "With vector store"**

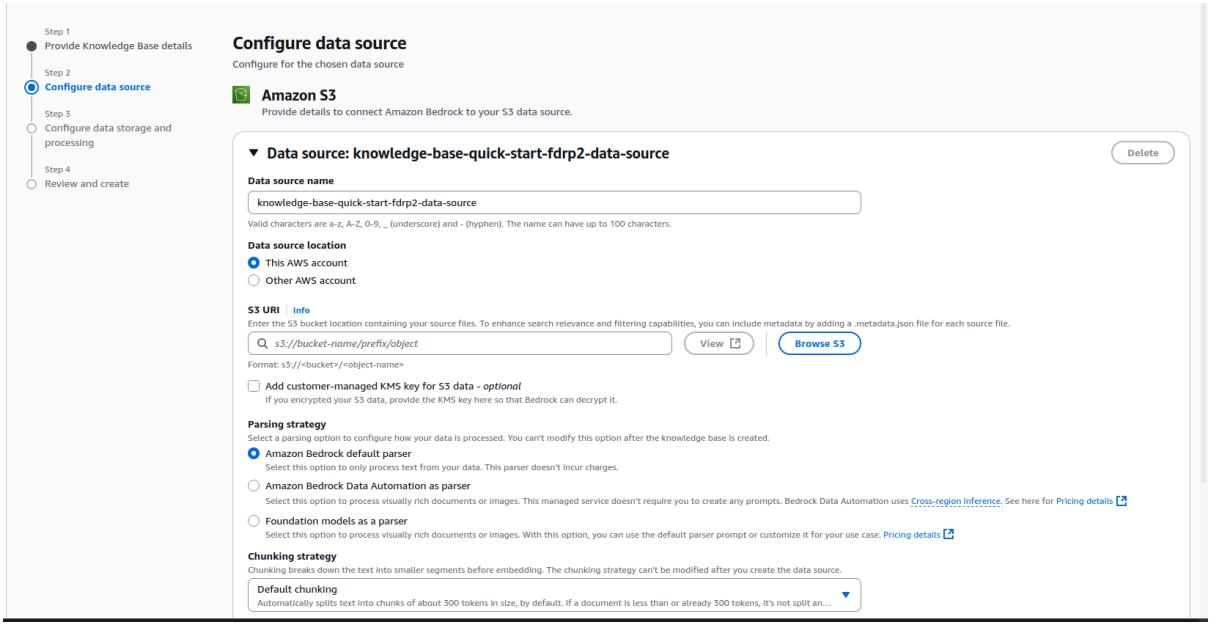


Paso 4.2.1: Configure knowledge base details

- dejaremos todos los campos por defecto y hacemos click en “Next”:

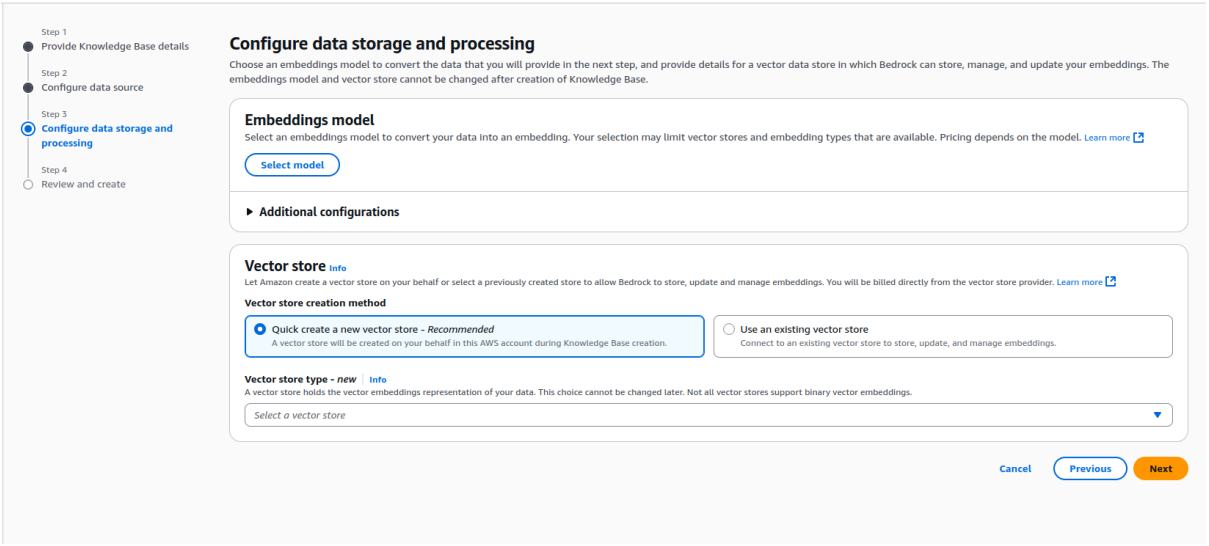


Paso 4.2.2: Configure data source



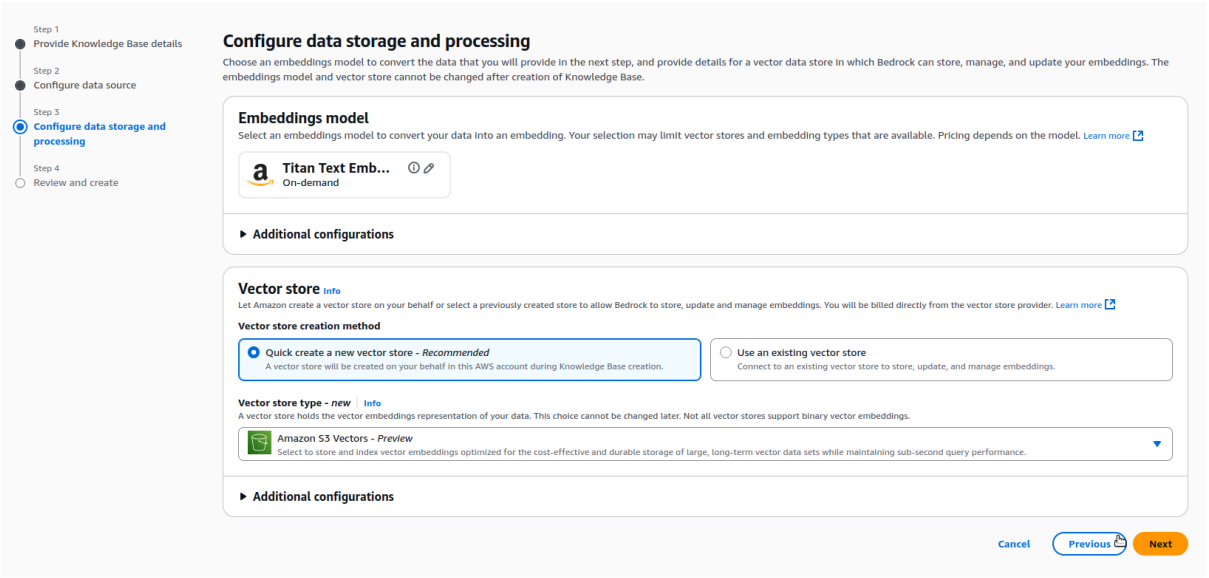
Los campos los dejamos por defecto a excepción del S3 URI, ahí le haremos click en “Browse S3” y seleccionaremos el bucket que creamos anteriormente para posteriormente clickear en “Next”.

Paso 4.2.3: Configure embeddings and vector store

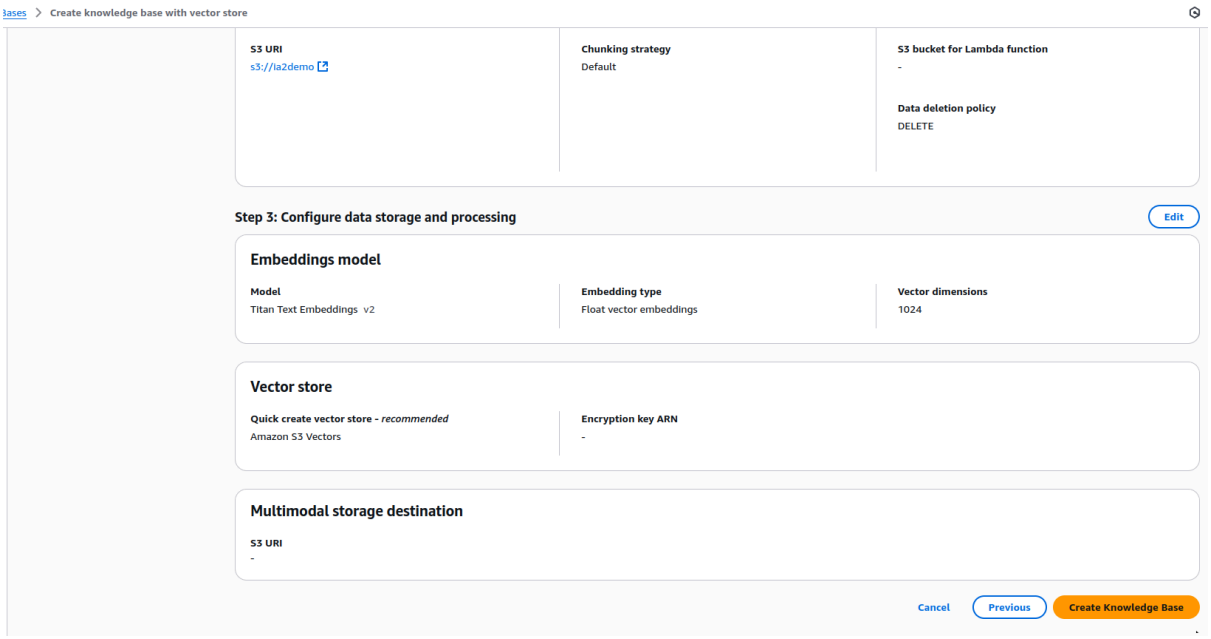


- **Embeddings model:** seleccionamos el modelo "Amazon Titan Embeddings V2"
    - **¿Qué son embeddings?** Son representaciones numéricas del texto que permiten comparar similitudes semánticas
- **Vector store:** "Quick create a new vector store"
- **Vector store type:** Amazon S3 vectores preview

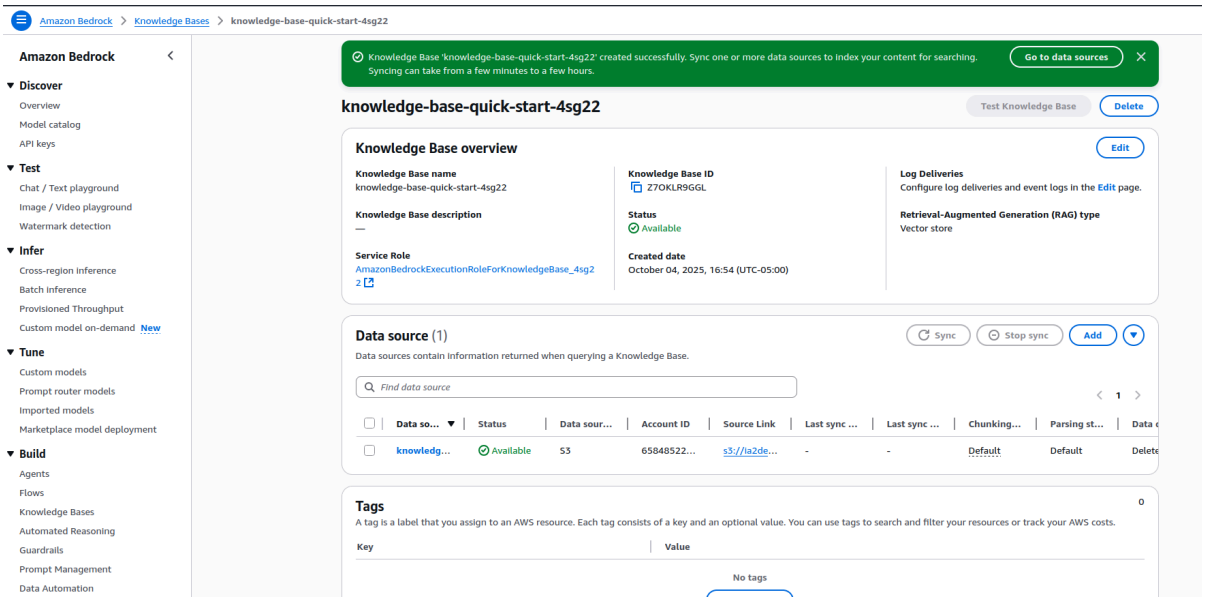
Así quedaría la configuración completa para posteriormente clickear en “Next”:



Paso 4.2.4: Review and create



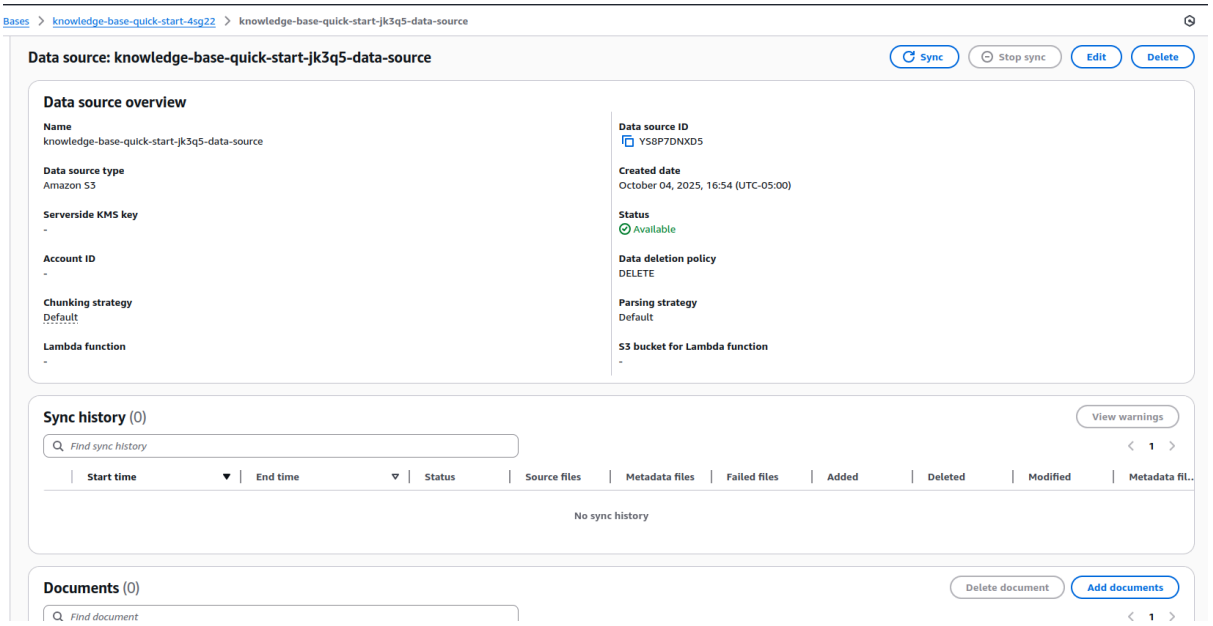
- Revisar el resumen de la configuración
- Hacer clic en "Create knowledge base"
- Si se ha creado correctamente la base de conocimiento serás redireccionado aquí:



4.3 Sincronizar Data Source

¿Por qué? Para procesar los documentos y crear los embeddings.

1. **Seleccionar la knowledge base** creada
2. **Ir a la pestaña "Data sources"**
3. **Seleccionar el data source**



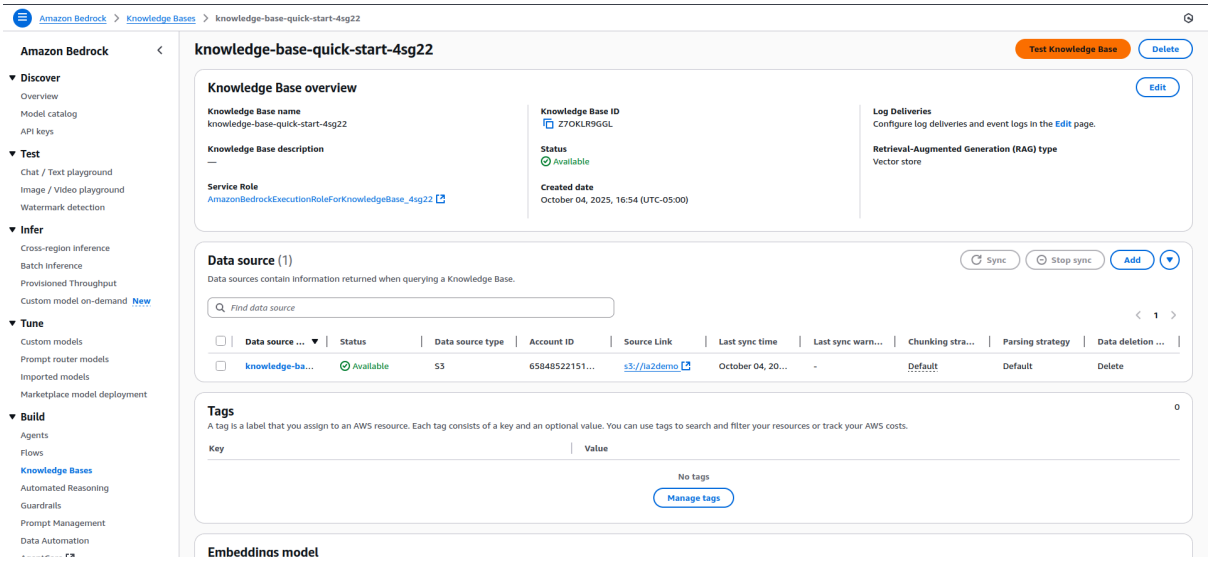
4. **Hacer clic en "Sync"**

Nota: Si aparece error de permisos, verificar que el modelo de embedding que colocaste esté habilitado en "Model access".

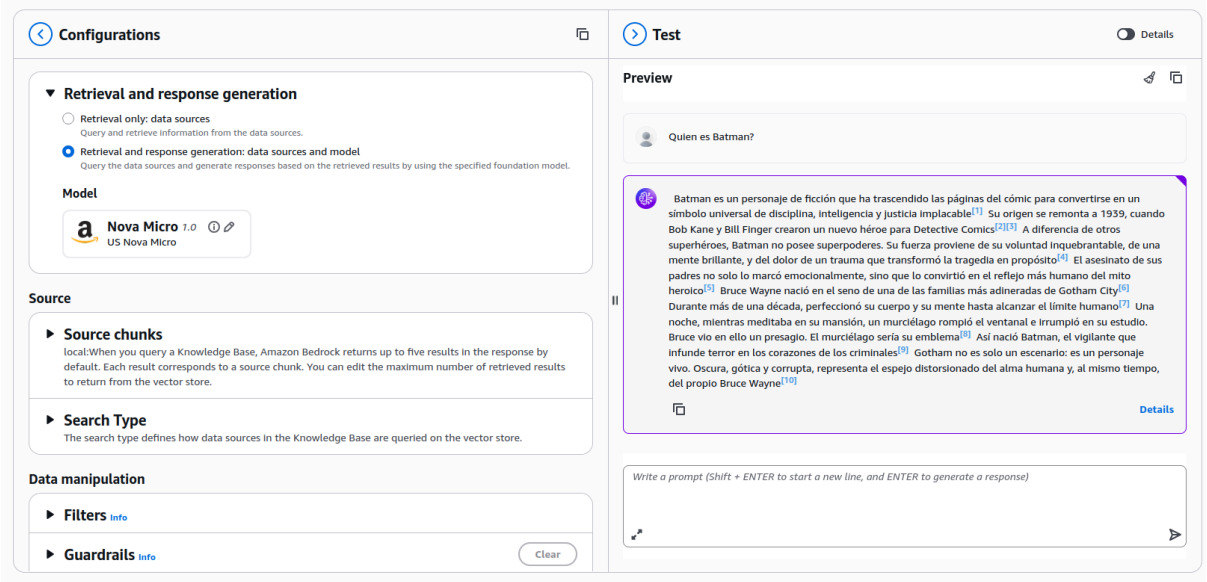
Paso 5: Probar la Knowledge Base

5.1 Test en Consola

1. **Ir a "Knowledge bases"** en el menú lateral
2. **Ir a tu knowledge base creada**
3. **Hacer clic en "Test knowledge base"**



4. **Seleccionar modelo:** Puedes usar cualquiera, sin embargo para esta guía se usará el "Nova Micro 1.0" ya que habilitamos el uso de todos estos modelos de amazon anteriormente.
5. **Hacer preguntas** sobre el contenido de tus documentos



5.2 ¿Qué observar?

- **Respuestas contextuales** basadas en tus documentos
- **Citas numeradas** que referencian las fuentes
- **Hacer clic en las citas** para ver el texto original

Paso 6: Demo Práctica con Python

6.1 Código de la Demo

descarga el archivo llamado “`demo_bedrock.py`” del repositorio

6.2 Instalación de Librerías Requeridas

Antes de ejecutar el código, instala la librería necesaria:

```
pip install boto3
```

¿Qué es boto3?

- Es el **SDK oficial de AWS para Python**
- Permite interactuar con servicios de AWS como Bedrock, S3, IAM, etc.

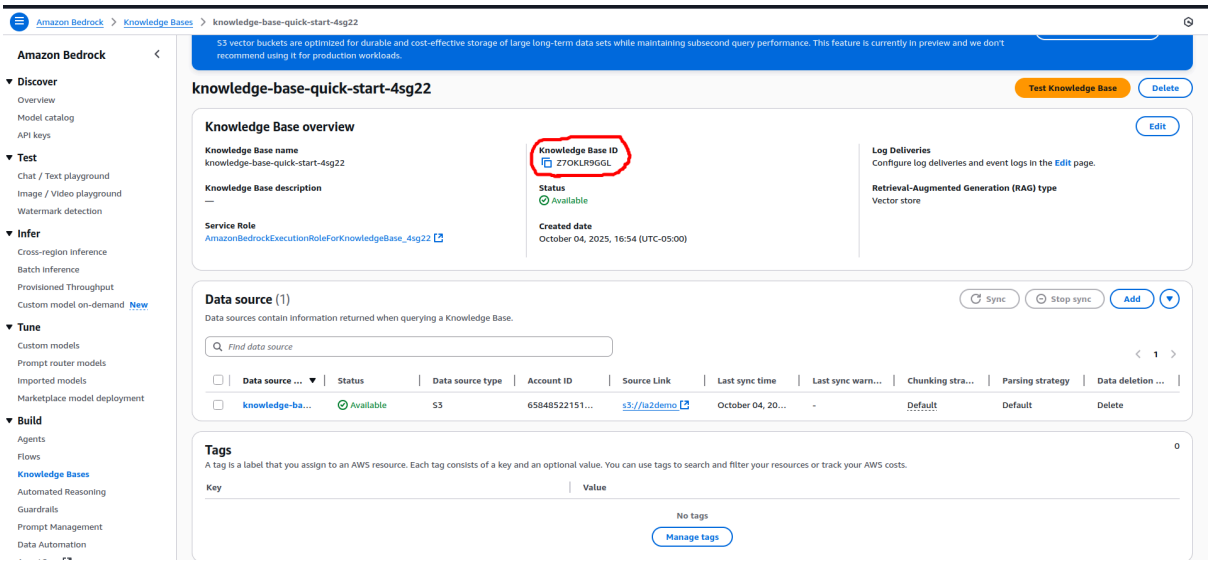
6.3 Configuración de Valores

Reemplaza estos 4 valores en el código:

1. **`KNOWLEDGE_BASE_ID`**: Tu Knowledge Base ID de Bedrock
2. **`AWS_ACCESS_KEY`**: Tu AWS Access Key ID
3. **`AWS_SECRET_KEY`**: Tu AWS Secret Access Key
4. **`AWS_REGION`**: `us-west-2` (ya configurada, pero debes cambiarla según tu región)
5. **`MODEL_NAME`**: `nova-micro-1-0` (ya configurado)

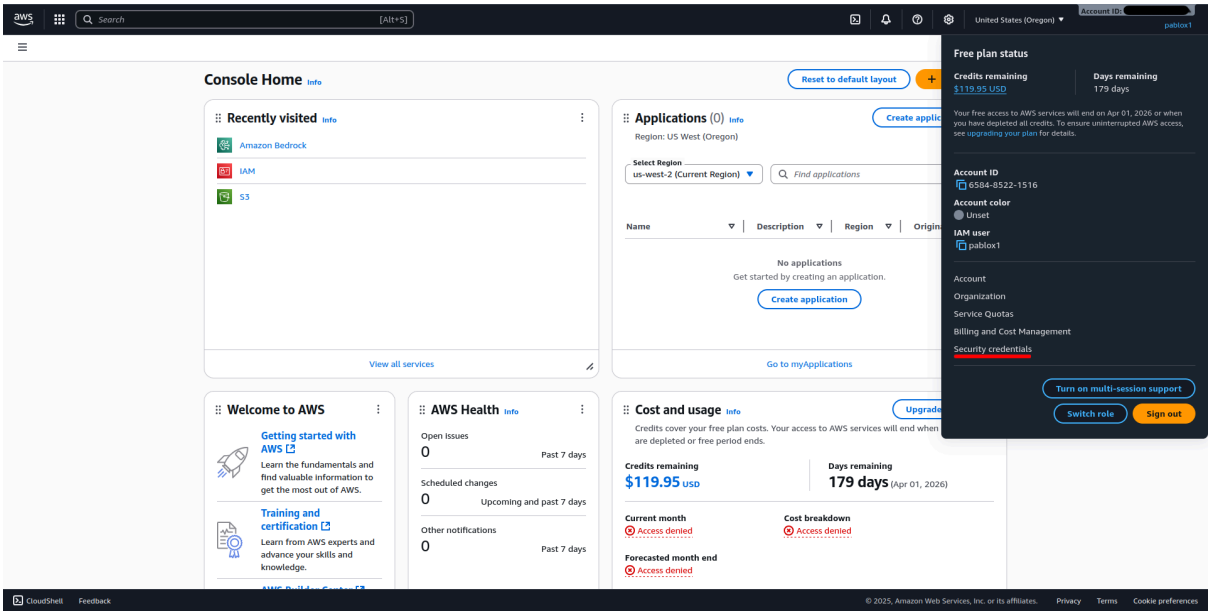
¿Dónde encontrar el **`KNOWLEDGE_BASE_ID`**?

se encuentra dirigiéndose a la base de conocimiento que creamos anteriormente:

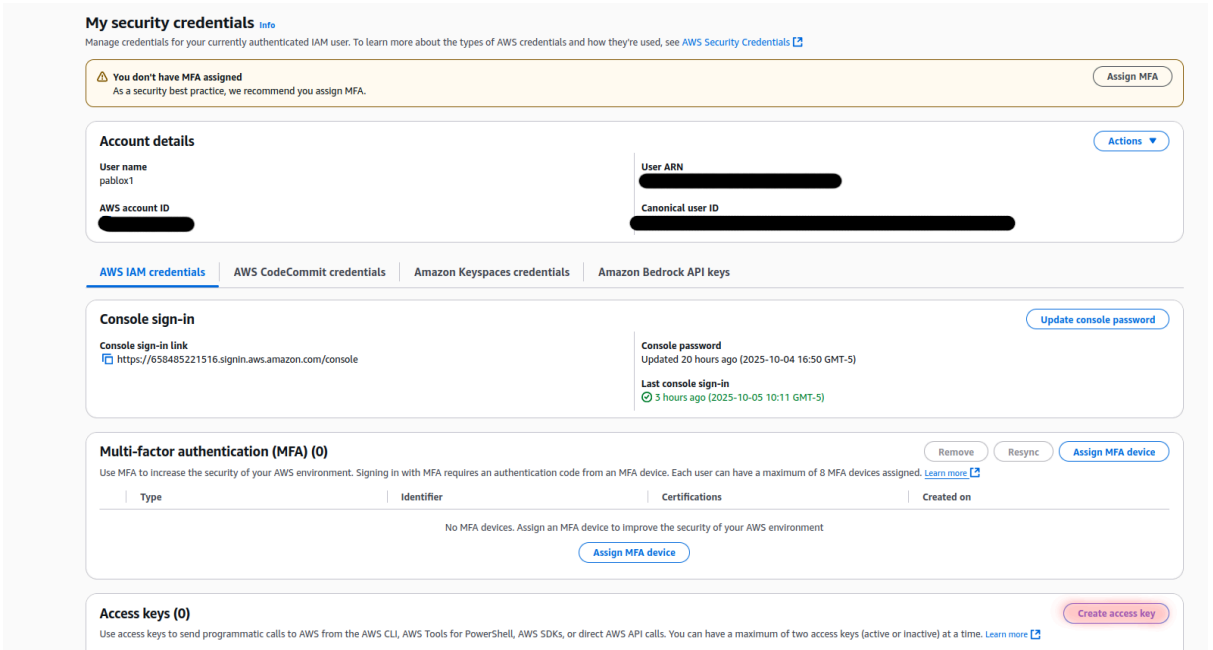


¿Dónde encontrar el **`AWS_ACCESS_KEY`** y **`AWS_SECRET_KEY`**?

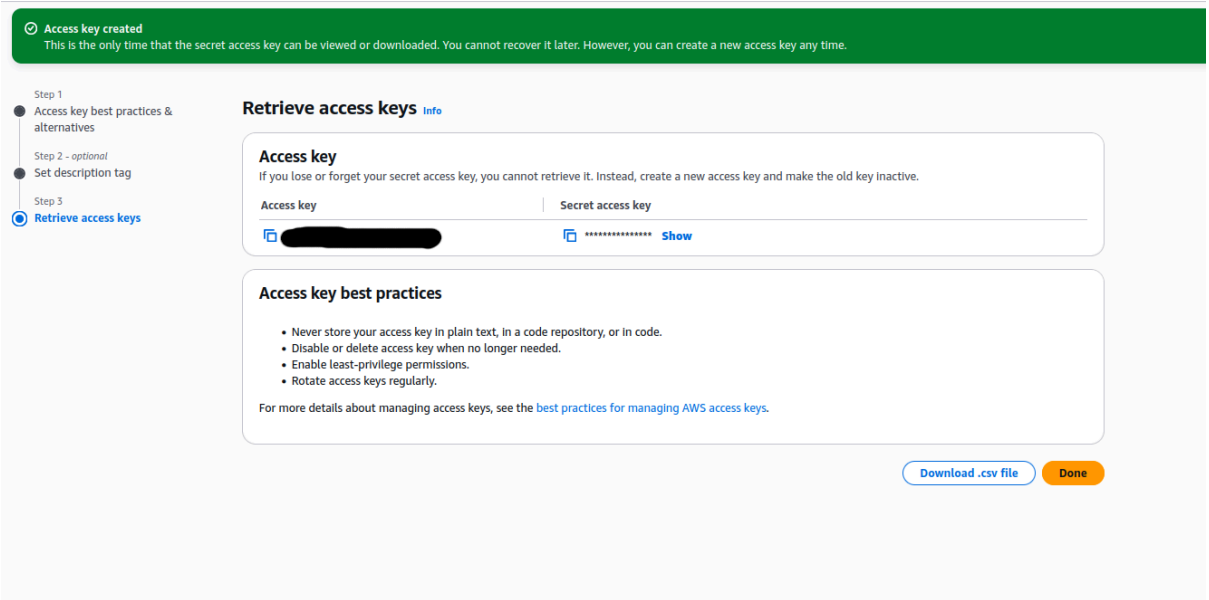
Haremos click al nuestro nombre de usuario que aparece en la parte superior derecha de la pantalla y le haremos click a security credentials



Serás redirigido a la siguiente página, donde deberás hacer click en “create access key para después seleccionar la opción de CLI”:



Después de los pasos anteriores nos arrojará las credenciales las cuales deberemos reemplazar en el código:



Nota: cuando tenga las credenciales inmediatamente pasalas al código, una vez hecho click en “Done” ya no puedes verlas.

6.3 Ejecutar la Demo

```
(env) pablo@pablo-VivoBook-ASUSLaptop-X415DA-M415DA:~/Desktop/IA2/practica aws$ python demo_bedrock.py
🤖 Pregunta lo que quieras sobre tus documentos:

Tú: quien es batman?
Bot: Batman es un personaje ficticio que ha trascendido las páginas del cómic para convertirse en un símbolo universal de disciplina, inteligencia y justicia implacab
le. Su origen se remonta a 1939, cuando Bob Kane y Bill Finger crearon un nuevo héroe para Detective Comics. A diferencia de otros superhéroes, Batman no posee superp
oderes, sino que su fuerza proviene de su voluntad inquebrantable, de una mente brillante, y del dolor de un trauma que transformó la tragedia en propósito. El asesin
ato de sus padres no solo lo marcó emocionalmente, sino que lo convirtió en el reflejo más humano del mito heroico. Desde su creación, Batman ha sido reinventado y re
interpretado por escritores, artistas y cineastas. Cada generación ha visto una nueva versión de este arquetipo moderno, una representación de la capacidad humana par
a enfrentarse a la oscuridad interior y exterior sin perder la cordura ni la ética.
```

Paso 7: Limpieza de Recursos (Opcional pero Recomendado si ya no harás uso de estas herramientas para que no consuma creditos)

7.1 Eliminar Knowledge Base

1. **Ve a AWS Bedrock → Knowledge bases**
2. **Selecciona tu knowledge base**
3. **Haz clic en "Delete"**
4. **Confirma la eliminación**

7.2 Eliminar Bucket S3

1. **Ve a Amazon S3 → Buckets**
2. **Selecciona tu bucket**
3. **Haz clic en "Empty"** (vaciar bucket primero)

4. **Luego haz clic en "Delete"**

Conclusiones

Esta práctica demostró la implementación exitosa de un sistema de IA generativa empresarial usando AWS Bedrock con RAG, donde se creó una base de conocimiento que combina modelos fundacionales como Nova Micro con documentos personalizados almacenados en S3, permitiendo respuestas precisas basadas en información específica. El proceso incluyó la configuración completa desde la creación de usuarios IAM con políticas adecuadas, implementación de buckets S3 para almacenamiento documental, habilitación de modelos en Bedrock, construcción de knowledge bases con vector stores para búsqueda semántica, y desarrollo de una aplicación Python funcional que consume las APIs de Bedrock, validando así la capacidad de AWS para democratizar el acceso a IA avanzada y permitir la creación de asistentes inteligentes contextuales que responden basándose en documentación corporativa específica, todo ello implementable en cuestión de horas sin requerir expertise especializado en machine learning.