GeneXus by Globant

Chatear con documentos – RAG Assistant

Alejandra Caggiano

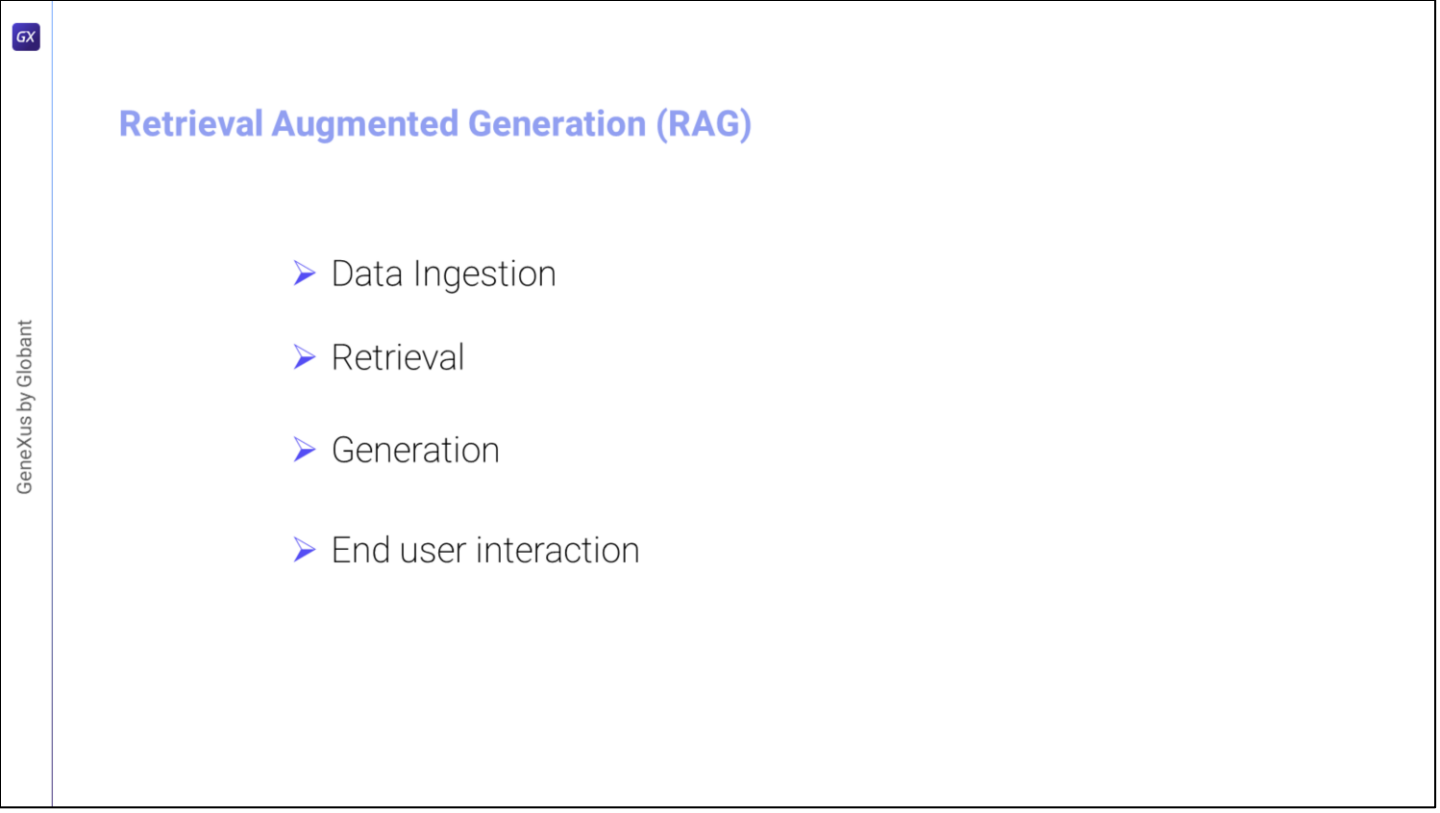
Hemos visto que GeneXus Enterprise AI permite crear distintos tipos de asistentes de inteligencia artificial. En particular, ya hemos logrado crear una conversación interactiva.

Queremos ahora poder definir un asistente que nos permita chatear con documentos. Para eso, vamos a trabajar con asistentes RAG.

GeneXus by Globant

Retrieval Augmented Generation (RAG)

La Recuperación de Generación Aumentada (RAG) es un enfoque que combina la recuperación de información a partir de datos no estructurados y la generación de texto para mejorar el rendimiento en tareas como ser la respuesta a preguntas.



Este proceso se compone de las siguientes cuatro fases:

- La primera fase es la **ingesta de datos:** Implica cargar varios tipos de documentos, en diferentes formatos y desde múltiples fuentes.

- Luego sigue la fase de **Recuperación**: En esta etapa se inicia el proceso de recuperación de datos, aprovechando la información previamente cargada y organizada. Se realiza una búsqueda selectiva sobre un conjunto de documentos, identificando la información relacionada y reduciendo eficientemente el espacio de búsqueda. Este enfoque garantiza que la atención se centre en la información más relevante y significativa.

- La siguiente fase es la de **Generación**: El foco aquí está en generar respuestas relevantes y contextualmente consistentes. En este proceso, el sistema utiliza la configuración del asistente RAG para saber a qué modelo acceder y con qué parámetros. Este asistente incorpora los elementos necesarios para definir la estrategia de búsqueda, y lograr coherencia y relevancia en el contenido generado.

- La fase final es **la Interacción con el usuario final:** GeneXus Enterprise AI facilita una comunicación fluida y eficiente entre los usuarios finales y los asistentes RAG, completando el ciclo y brindando respuestas a las consultas de manera eficiente.

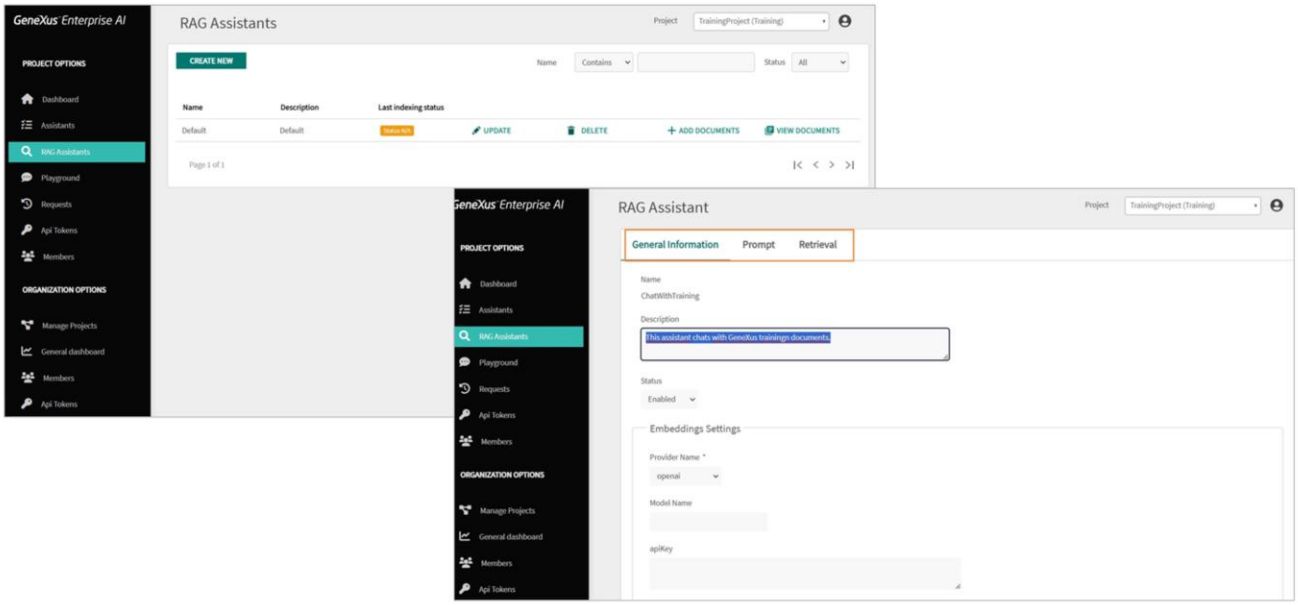
Retrieval Augmented Generation (RAG)

Ejemplo:

Chatear con documentos de GeneXus Training

Bien. Como plantemos al comienzo, nuestro objetivo ahora es crear un asistente que nos permita chatear con un conjunto de documentos, y lo haremos con documentos de GeneXus Training.

RAG Assistants

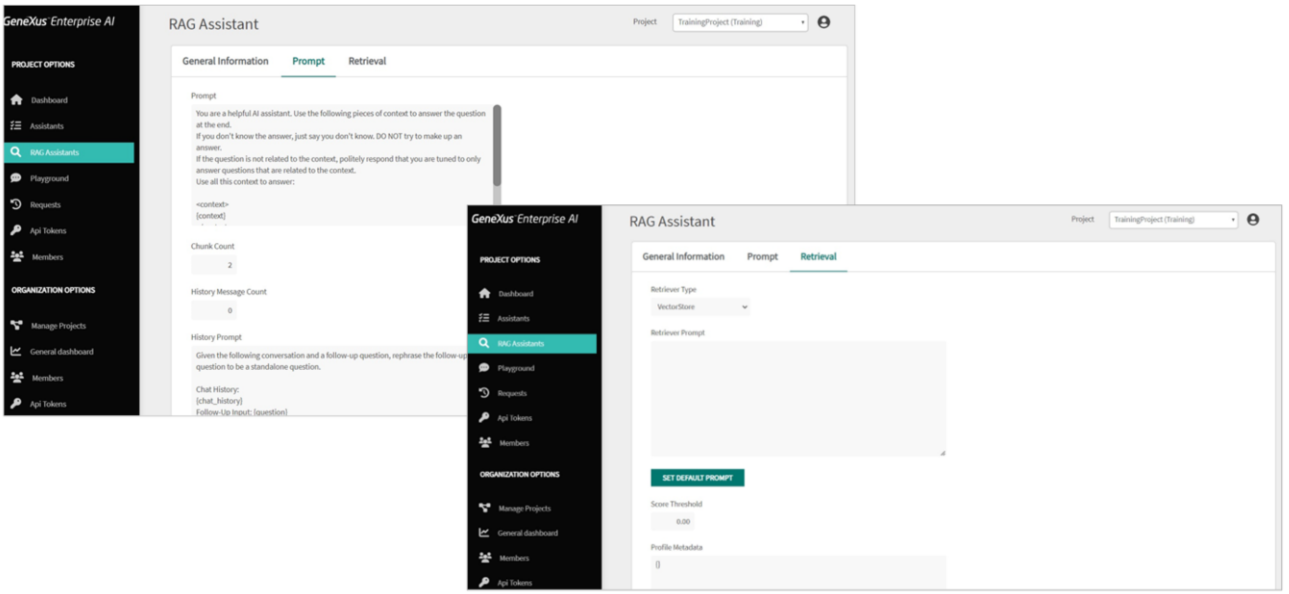


Así que ingresamos a la plataforma, seleccionamos el proyecto sobe el cual vamos a trabajar, y en el menú elegimos RAG Assistant. Por defecto, al abrir esta sección vemos un RAG Assistant de nombre Default, que puede ser personalizado, o bien podemos crear nuevos.

Presionamos Create new. Ponemos como nombre ChatWithGXTraining e ingresamos una breve descripción. Presionamos Confirm.

Desde la opción Update podemos personalizar la definición del asistente, según sea necesario. El conjunto de settings se organiza en estas solapas:

RAG Assistants



La Información general, el Prompt, que contiene instrucciones que orienten al asistente sobre cómo abordar y responder preguntas. Estas instrucciones establecen pautas claras para que el asistente brinde respuestas relevantes y útiles basadas en el contexto proporcionado.

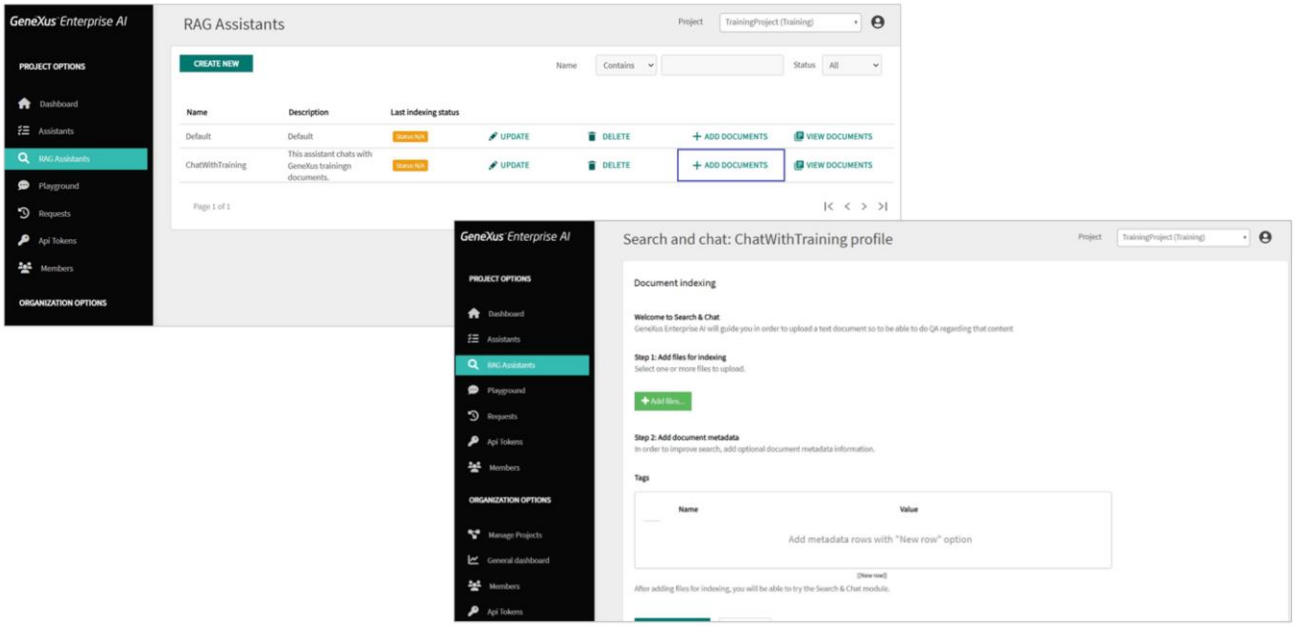
Esta opción indica la cantidad de fragmentos que se recuperan para aumentar el contexto.

Luego esta opción de Historial de mensajes establece la cantidad de mensajes históricos que se tienen en cuenta en la conversación. Esto es útil para rastrear el historial de interacción y comprender el contexto recopilado en la conversación.

Si este valor está establecido en 4, significa que interesa considerar las últimas 4 interacciones:

El valor mínimo que puede tomar es 0, lo que indica que el historial de conversaciones no es de interés. Cuando el valor es mayor que 0, se utiliza junto con el mensaje indicado en la siguiente opción History Prompt.

RAG Assistants



Luego vemos las opciones para establecer la configuración del modelo utilizado por el asistente para generar la respuesta. Esto incluye el proveedor de servicios, el nombre del modelo, la temperatura, el límite máximo de tokens y otros parámetros que afectan la forma en que se generan las respuestas.

Finalmente, la solapa de Recuperación, que especifica cómo se recupera la información.

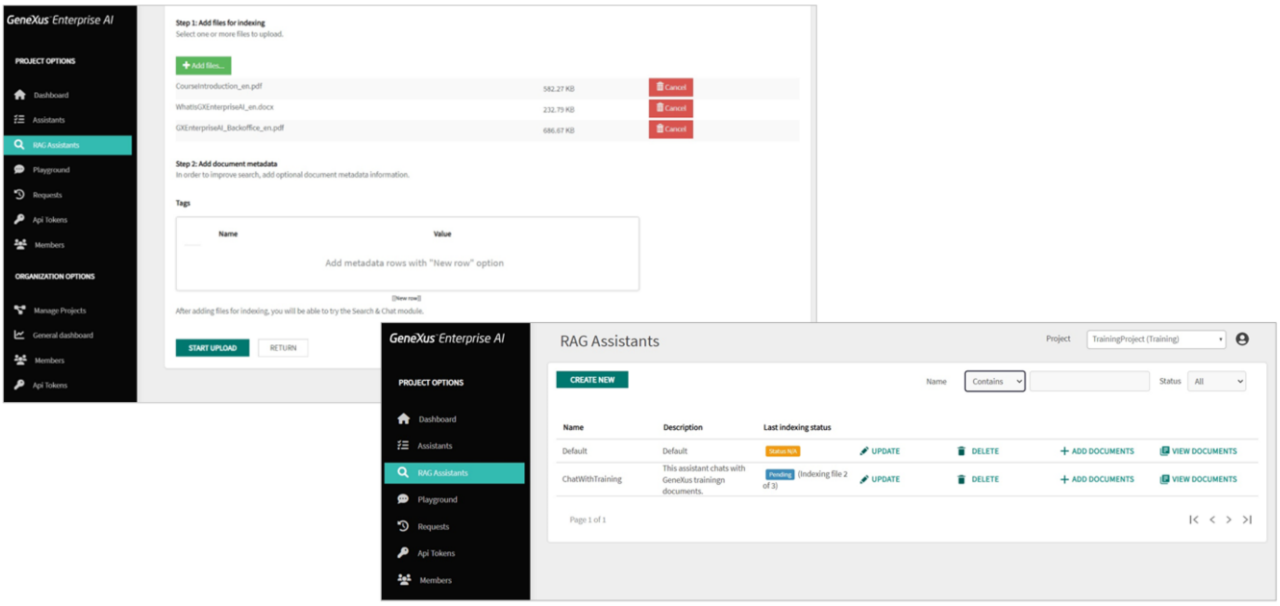
Dejamos los valores por defecto.

Bien. Vamos ahora a cargar los documentos. Presionamos Add Documents. El botón Add Files permite realizar la carga de archivos de diversos formatos:

En nuestro ejemplo vamos a cargar un pequeño conjunto de pdfs y docs que corresponden a material de GeneXus training.

Presionamos Add Files

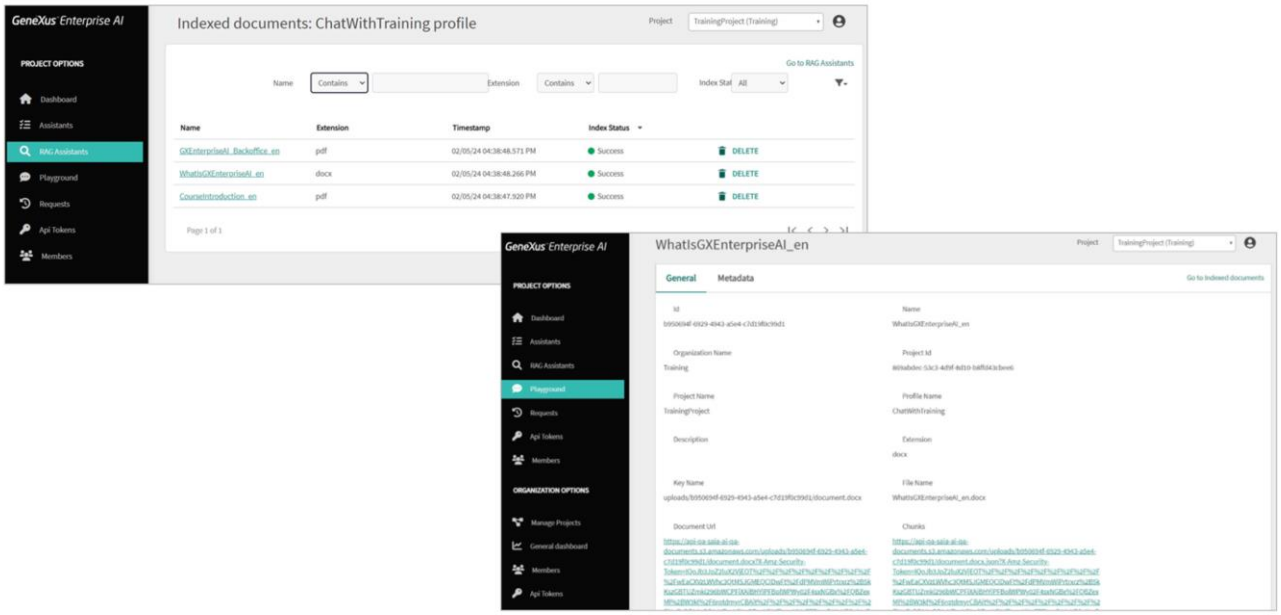
RAG Assistants



Una vez cargados los documentos presionamos Start upload

Para ver los documentos cargados, cada uno con su detalle, presionamos View Documents

RAG Assistants



Seleccionando el nombre del archivo vemos toda la información asociada, pudiendo ver y descargar el archivo desde la URL.

Bien. Ya tenemos creado nuestro RAG Assistant y lo hemos cargado con los archivos correspondientes. Estamos entonteces en condiciones de testearlo.

Lo haremos a continuación desde la opción Playground del menú.

GeneXus by Globant

GeneXus™
by Globant

training.genexus.com