Arquitectura de Solución – RGA

YoCampo Prototipo

Arquitectura para Implementación de YoCampo en Azure

Autores: Victor Manuel Mondragon Maca

Laboratorio de Datos SNUIRA

Versión:1.0

Fecha: 19 de septiembre de 2024

Unidad de Planificación Rural Agropecuaria (UPRA)
Calle 28 Nº 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.
+57(601) 552 9820, 245 7307
www.upra.gov.co

Unidad de Planificación Rural Agropecuaria (UPRA)
Calle 28 Nº 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.
+57(601) 552 9820, 245 7307
www.upra.gov.co

Resumen

Este documento detalla la arquitectura de solución del proyecto YoCampo, una plataforma que utiliza inteligencia artificial para mejorar el acceso a información técnica en el sector agropecuario. Basado en la Arquitectura de Generación Aumentada de Recuperación (RAG), YoCampo integra la recuperación de datos relevantes mediante Azure Cognitive Search y la generación de respuestas contextualizadas usando Azure OpenAI (GPT). Los datos se almacenan y gestionan en Cosmos DB, mientras que Embeddings API permite búsquedas avanzadas. La plataforma ofrece respuestas personalizadas a través de una Web App intuitiva, adaptando la información a las necesidades de extensionistas, investigadores y productores. Esta solución optimiza la toma de decisiones y mejora la eficiencia en el sector agropecuario, facilitando el acceso a información crítica

Tabla de contenido

Índice de tablas

No se encuentran elementos de tabla de ilustraciones.

Índice de figuras

Lista de siglas y abreviaturas

Aunap	Autoridad Nacional de Acuicultura y Pesca
UPRA	Unidad de Planificación Rural Agropecuaria

Glosario

Azure Cognitive Search: Un servicio de búsqueda en la nube con capacidades de inteligencia artificial, utilizado para realizar búsquedas avanzadas y recuperación de información.

Cognitive Services: Conjunto de servicios de inteligencia artificial que incluye procesamiento de lenguaje natural, visión artificial y servicios de habla.

Virtual Machines (VM): Instancias de servidores virtuales en la nube, altamente configurables y escalables.

Azure App Service: Plataforma totalmente gestionada para construir, desplegar y escalar aplicaciones web.

Azure Storage: Servicio de almacenamiento en la nube altamente escalable y duradero.

Microsoft Defender for Cloud: Servicio de seguridad que proporciona gestión unificada de la seguridad y protección contra amenazas para las cargas de trabajo en la nube.

Azure Load Balancer: Servicio que distribuye el tráfico entrante a varios servicios en la nube, asegurando alta disponibilidad y redundancia.

Azure Cosmos DB: Base de datos NoSQL globalmente distribuida, diseñada para ofrecer escalabilidad y disponibilidad altas.

Blob Storage: Servicio de almacenamiento de objetos de Azure, optimizado para almacenar grandes cantidades de datos no estructurados.

Azure OpenAI: Servicio que proporciona acceso a modelos avanzados de inteligencia artificial de OpenAI, como GPT-4, a través de la infraestructura de Azure.

Azure AI: Conjunto de servicios y herramientas de inteligencia artificial ofrecidos por Azure, diseñados para ayudar a los desarrolladores a construir aplicaciones inteligentes.

Budget (Presupuesto): Límite financiero establecido para controlar y monitorear los gastos en Azure.

Forecast (Proyección): Estimación de costos futuros basada en el uso actual y tendencias pasadas.

Actual (Real): Costos efectivamente incurridos en un periodo específico.

Cost Management (Gestión de Costos): Herramienta de Azure que permite monitorizar, asignar y optimizar los costos asociados a los servicios utilizados.

Scoping (Alcance): Proceso de asignar y definir un presupuesto a un ámbito específico dentro de Azure, como suscripciones, grupos de recursos o servicios específicos.

Resource (Recurso): Cualquier instancia de servicio o componente de Azure utilizado en la implementación de YoCampo.

Resource Group (Grupo de Recursos): Contenedor lógico en Azure que agrupa recursos relacionados para facilitar su administración.

Subscription (Suscripción): Unidad de facturación y administración en Azure que organiza recursos y servicios.

Cost Allocation (Asignación de Costos): Proceso de distribuir los costos a diferentes proyectos, departamentos o unidades de negocio dentro de una organización.

Cost Optimization (Optimización de Costos): Estrategias y acciones para reducir los costos operativos mientras se mantiene o mejora el rendimiento y la eficiencia de los servicios en la nube.

YoCampo: Proyecto que utiliza múltiples servicios de Azure para ofrecer una solución robusta y escalable en el ámbito agropecuario.

Retrieval-Augmented Generation (RAG): Técnica que combina la recuperación de documentos relevantes con la generación de texto, mejorando la precisión y relevancia de las respuestas generadas por modelos de lenguaje.

Region (Región): Ubicación geográfica de los centros de datos de Azure donde se alojan los recursos.

Scope (Ámbito): El rango o alcance al que se aplica un presupuesto o política en Azure.

Introducción

El sector agropecuario enfrenta desafíos significativos, como la necesidad de mejorar la productividad, gestionar eficientemente los recursos naturales y adoptar nuevas tecnologías para enfrentar el cambio climático. En este contexto, el proyecto YoCampo se presenta como una solución innovadora, basada en inteligencia artificial, diseñada para ofrecer acceso rápido y preciso a información técnica crítica. La plataforma está orientada a satisfacer las necesidades de extensionistas, investigadores y productores, proporcionándoles una herramienta avanzada para la toma de decisiones.

La arquitectura de YoCampo se construye sobre el concepto de Generación Aumentada de Recuperación (RAG), que integra tecnologías de Azure como Cognitive Search, Cosmos DB, y Azure OpenAI (GPT). Esta arquitectura permite no solo la recuperación de información relevante desde diversas fuentes documentales, como la Biblioteca Agropecuaria de Colombia (BAC), sino también la generación de respuestas contextualmente enriquecidas a las consultas de los usuarios.

El proceso incluye la ingesta y almacenamiento de datos en Cosmos DB, seguido de su indexación y análisis semántico mediante Azure Cognitive Search. A través de Embeddings API, se optimiza la búsqueda semántica, permitiendo al modelo GPT generar respuestas adaptadas a cada usuario. La plataforma se accede mediante una Web App intuitiva que personaliza las respuestas en función del perfil del usuario, maximizando la utilidad de la información proporcionada.

YoCampo, por tanto, se posiciona como una herramienta estratégica para mejorar la productividad, sostenibilidad y toma de decisiones en el sector agropecuario, facilitando el acceso a datos relevantes y actualizados.

Contexto y Objetivo de la Solución

- Describir la arquitectura de solución de YoCampo, detallando el uso de tecnologías avanzadas como Azure Cognitive Search, Cosmos DB, y Azure OpenAI (GPT), y su integración dentro del concepto de Generación Aumentada de Recuperación (RAG).

  Explicar el flujo de ingesta, procesamiento y recuperación de datos, incluyendo la indexación semántica mediante Embeddings API y la generación de respuestas personalizadas a partir de consultas de usuarios en tiempo real.

Alcance

Este documento abarca la descripción detallada de la arquitectura de solución YoCampo, enfocándose en el uso de tecnologías avanzadas de Microsoft Azure para la ingesta, procesamiento, recuperación y generación de respuestas personalizadas a partir de datos agropecuarios. El contenido explora el funcionamiento de la Arquitectura de Generación Aumentada de Recuperación (RAG) y su implementación, explicando cómo herramientas como Cosmos DB, Azure Cognitive Search, y Azure OpenAI (GPT) se integran para ofrecer un acceso eficiente a la información técnica. Además, se analiza el flujo de trabajo desde la ingesta de datos hasta la presentación de respuestas a través de una Web App diseñada para diferentes perfiles de usuarios. El documento también evalúa el impacto potencial de la plataforma YoCampo en la mejora de la toma de decisiones, productividad y sostenibilidad del sector agropecuario, ofreciendo una solución robusta y escalable para enfrentar los desafíos actuales en el acceso y uso de la información técnica.

1. Arquitectura Cloud YoCampo

El proyecto YoCampo, utiliza inteligencia artificial generativa para asistir a los actores del sector agropecuario con información actualizada y personalizada. La arquitectura RAG facilita la integración de estos componentes mediante la combinación de tecnologías de recuperación de información y generación de lenguaje natural. Al implementar una solución RAG, YoCampo puede ofrecer respuestas precisas y personalizadas a consultas específicas, de esta forma mejorando el acceso a la información para el beneficio de actores del Sistema Nacional de Innovación Agropecuaria (SNIA).

A continuación, se describe la arquitectura de solución implementada en el prototipo experimental YoCampo aliñada a la estimación de costos para su replicación y proyecciones de operación.

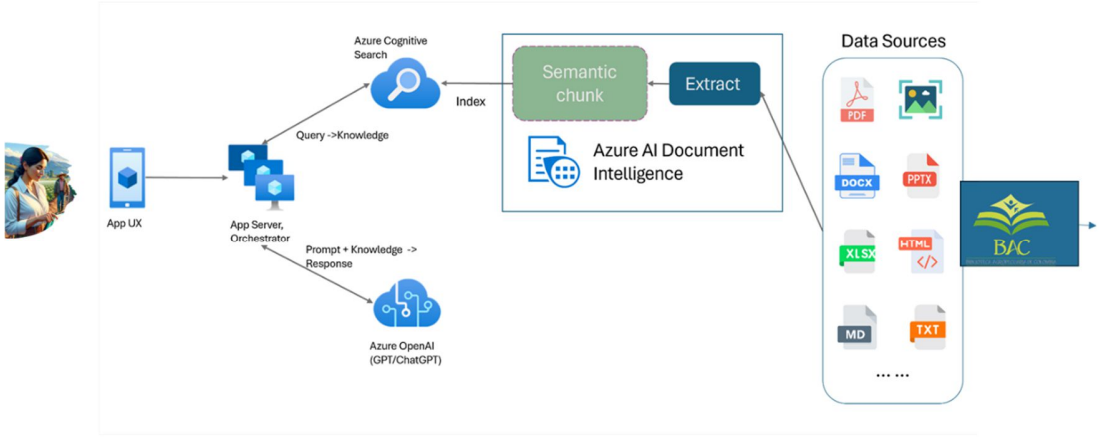
1.1. Arquitectura de Generación Aumentada de Recuperación (RAG)

La **Arquitectura de Generación Aumentada de Recuperación (RAG)** [1] es una técnica avanzada que combina la recuperación de información relevante desde bases de datos y fuentes documentales, con la capacidad de generar respuestas contextualmente enriquecidas mediante modelos de lenguaje como **Azure OpenAI (GPT)** [2]. Esta arquitectura se ha implementado en YoCampo para mejorar la toma de decisiones en el sector agropecuario, facilitando el acceso a información crítica a partir de una amplia gama de fuentes como la **Biblioteca Agropecuaria de Colombia (BAC)**.

En este tipo de arquitectura, los datos provenientes de diversas fuentes documentales son procesados y fragmentados semánticamente. Posteriormente, mediante el uso de Azure Cognitive Search, se recupera la información que luego es utilizada por Azure OpenAI para generar respuestas

personalizadas y contextualmente relevantes. El propósito principal de esta arquitectura es maximizar la eficiencia y precisión en las respuestas que reciben los usuarios.

Figura 1 Arquitectura RGA alto nivel



Fuente: Adaptada de Microsoft

La Figura 1 representa un flujo de Arquitectura de Generación Aumentada de Recuperación (RAG), en el contexto de la solución YoCampo, donde se combina la recuperación de información de diversas fuentes documentales con la generación de respuestas personalizadas mediante inteligencia artificial, específicamente usando Azure OpenAI (GPT).

La figura de igual forma representa un flujo de Arquitectura de Generación Aumentada de Recuperación (RAG), en el contexto de la solución YoCampo, donde se combina la recuperación de información de diversas fuentes documentales con la generación de respuestas personalizadas mediante inteligencia artificial, específicamente usando Azure OpenAI (GPT).

- **Fuentes de Datos**:En el extremo derecho de la imagen se muestran las diferentes fuentes de datos, las cuales incluyen documentos en diversos formatos (PDF, DOCX, PPTX, XLSX, HTML, TXT, entre otros). Estas fuentes provienen de la Biblioteca Agropecuaria de Colombia (BAC), que almacena una vasta cantidad de documentos científicos y técnicos relevantes para

el sector agropecuario. La variedad de formatos refleja la diversidad de tipos de información que el sistema puede manejar.

- **Azure AI Document Intelligence**: Una vez que los documentos son extraídos desde las fuentes de datos, son procesados a través de Azure AI Document Intelligence, una herramienta que permite la extracción del contenido relevante de los documentos. Esta extracción no solo se limita a extraer el texto, sino que también incluye el análisis y comprensión de la estructura del documento, como tablas, gráficos y otros elementos clave.
- **Chunking Semántico (Semantic Chunk):** El proceso de chunking semántico divide los documentos en fragmentos más pequeños, denominados "chunks", que están basados en la coherencia semántica del contenido. Estos chunks no se generan arbitrariamente por longitud o tamaño, sino que se agrupan de acuerdo con la lógica del contenido, asegurando que cada fragmento trate un tema o concepto específico y coherente.
- **Azure Cognitive Search**: Después de la creación de los chunks semánticos, estos se indexan en Azure Cognitive Search. Esta herramienta permite la realización de búsquedas inteligentes dentro del índice de documentos, lo que facilita que el sistema recupere rápidamente la información más relevante para una consulta específica. La búsqueda no se basa en coincidencias exactas de palabras clave, sino que utiliza un enfoque semántico, lo que aumenta la precisión de los resultados.
- **App Server y Orchestrator**: Las consultas de los usuarios se realizan a través de una App UX (User Experience), donde un servidor actúa como orquestador. Este servidor recibe las consultas, las envía a Azure Cognitive Search para recuperar la información relevante y luego organiza la respuesta que se generará para el usuario. La orquestación de este flujo es clave para asegurar que cada consulta sea manejada de manera eficiente y en tiempo real.

- **Generación de Respuestas con Azure OpenAI (GPT):** Una vez que Azure Cognitive Search recupera los fragmentos de información relevantes, estos son enviados al modelo de Azure OpenAI (GPT/ChatGPT). Este modelo no solo replica los fragmentos recuperados, sino que genera una respuesta enriquecida y coherente que responde directamente a la consulta del usuario. El GPT utiliza la información recuperada para dar contexto a la respuesta y adaptarla a la pregunta realizada.
- **Respuesta al Usuario:** Finalmente, la respuesta generada por Azure OpenAI es enviada de vuelta al usuario a través de la App UX, cerrando así el ciclo de consulta-respuesta. El resultado es una respuesta altamente personalizada, contextualizada y basada tanto en el conocimiento preentrenado del modelo GPT como en la información actualizada recuperada desde las fuentes de datos.

En general la arquitectura descrita es un proceso robusto de recuperación de información y generación de respuestas, donde cada componente juega un papel clave en garantizar que las consultas de los usuarios se respondan de manera eficiente, precisa y basada en datos actualizados. La integración de herramientas como Azure Cognitive Search y Azure OpenAI permite combinar la búsqueda avanzada con la inteligencia artificial generativa, maximizando la utilidad de las fuentes de datos almacenadas en la BAC y otras fuentes agropecuarias

1.2. Principios de la Arquitectura RAG en Azure Cognitive Search

La Arquitectura RAG en YoCampo se basa en los siguientes principios fundamentales:

- **Recuperación Inteligente de Información**: A diferencia de las arquitecturas tradicionales que únicamente generan respuestas basadas en datos previamente entrenados, la arquitectura RAG permite la recuperación de documentos y fragmentos semánticos específicos en

tiempo real desde fuentes como la BAC. Estos documentos se indexan y almacenan en una estructura eficiente utilizando Azure Cognitive Search.

- **Generación de Respuestas Enriquecidas**: El uso de un modelo de lenguaje como Azure OpenAI permite que, tras la recuperación de los documentos más relevantes, se generen respuestas textuales contextualizadas. El modelo no solo recupera información directa, sino que la combina y la presenta de manera coherente y comprensible para los usuarios.

- **Uso de Embeddings Semánticos**: La arquitectura utiliza embeddings para representar el contenido de los documentos en un formato vectorial que captura su significado semántico. Esto permite realizar búsquedas más precisas y relevantes, al identificar no solo coincidencias exactas de palabras clave, sino también el contexto general de los documentos.

- **Modularidad y Escalabilidad**: La arquitectura está diseñada para ser modular, lo que facilita su escalabilidad. Nuevas fuentes de datos y documentos pueden ser fácilmente integradas sin comprometer el rendimiento. Esto es crucial para el crecimiento de YoCampo a medida que se agregan más datos del sector agropecuario.

- **Interoperabilidad**: Un principio clave es la interoperabilidad entre diferentes fuentes de datos, como BAC, Linkata y Siembra, a través de una infraestructura de datos flexible como Azure Blob Storage y Cosmos DB, que permite manejar datos no estructurados y estructurados.

1.3. Flujo de Datos y Generación de Respuestas Contextualizadas

El flujo de datos dentro de la arquitectura RAG en YoCampo sigue una secuencia optimizada para asegurar que los usuarios reciban respuestas precisas y personalizadas. A continuación, se describe este flujo:

- **Ingesta y Almacenamiento de Datos**: Los datos son capturados desde múltiples fuentes, como archivos PDF, documentos de Word, presentaciones, y videos almacenados en plataformas como la Biblioteca Agropecuaria de Colombia (BAC). Estos archivos se ingieren y almacenan en Azure Blob Storage o Cosmos DB para su posterior procesamiento.

- **Extracción e Indexación de Documentos**: Usando Azure AI Document Intelligence, los documentos son procesados y se extrae su contenido relevante. Este contenido es indexado en Azure Cognitive Search, lo que facilita la búsqueda rápida de información específica cuando se realizan consultas.

- **Consulta del Usuario**: Cuando un usuario de la aplicación YoCampo realiza una pregunta a través de la interfaz de usuario (App UX), esa consulta es enviada al servidor donde se orquesta la búsqueda. Esta consulta es interpretada semánticamente y enviada a Azure Cognitive Search para recuperar los fragmentos de documentos más relevantes.

- **Búsqueda en el Índice**: Azure Cognitive Search busca dentro del índice creado a partir de los documentos procesados y devuelve aquellos fragmentos que mejor responden a la consulta del usuario. Estos fragmentos o chunks son piezas de información semánticamente coherentes.

- **Generación de Respuesta**: Una vez que los documentos relevantes han sido recuperados, el modelo de lenguaje Azure OpenAI (GPT) toma los

fragmentos más pertinentes y genera una respuesta coherente. Esta respuesta no es una simple repetición de los fragmentos recuperados, sino una síntesis contextual que responde directamente a la consulta del usuario.

- **Devolución de la Respuesta**: La respuesta generada se envía de vuelta al usuario a través de la interfaz de la aplicación (App UX), mostrando una combinación de la información recuperada y el conocimiento generado por el modelo GPT.

Este flujo asegura que el usuario no solo obtenga una respuesta basada en un modelo pre-entrenado, sino que esa respuesta esté enriquecida con información contextualizada y actualizada directamente desde los documentos más relevantes.

1.4. Procesamiento de Documentos y Chunking Semántico

Uno de los elementos clave de la arquitectura RAG en YoCampo es el proceso de chunking semántico, que permite dividir grandes volúmenes de texto en fragmentos más pequeños y coherentes llamados "chunks". Este proceso se lleva a cabo de la siguiente manera:

- **Extracción de Documentos**: Los documentos provenientes de diversas fuentes, como la BAC, son procesados a través de Azure AI Document Intelligence. Esta herramienta no solo permite extraer el texto de documentos en varios formatos (PDF, DOCX, PPTX, etc.), sino que también identifica las secciones más relevantes del contenido.

- **Fragmentación Semántica (Chunking):** Una vez extraído el contenido, el texto es dividido en fragmentos o chunks basados en su semántica. Esto significa que, en lugar de dividirse por tamaño o por una estructura rígida (como párrafos o páginas), los fragmentos se crean de acuerdo con temas o ideas coherentes dentro del documento. Por ejemplo, un chunk puede contener toda la información relacionada con las técnicas de cultivo de cacao, mientras que otro puede estar dedicado a las plagas comunes.

- **Indexación de los Chunks**: Cada uno de estos chunks es indexado en Azure Cognitive Search utilizando embeddings, lo que permite realizar búsquedas semánticas más precisas. Los embeddings capturan el significado de cada chunk en lugar de solo las palabras clave, lo que asegura que se puedan recuperar los fragmentos más relevantes para cualquier consulta.
- **Uso de los Chunks en la Generación de Respuestas**: Cuando un usuario realiza una consulta, los chunks más relevantes se recuperan de Azure Cognitive Search y se utilizan como base para generar una respuesta. Gracias a la granularidad de los chunks, la respuesta puede estar compuesta por información muy específica, extraída de varias partes de uno o más documentos.
- **Actualización y Adaptación Continua**: El proceso de chunking semántico también permite la actualización continua de los documentos sin necesidad de reindexar por completo los archivos. Si se agregan nuevos documentos o se actualizan los existentes, solo los nuevos chunks son creados y añadidos al índice, lo que garantiza que el sistema siempre esté actualizado con la información más reciente.

Este proceso de **chunking semántico** no solo mejora la relevancia de las respuestas generadas, sino que también reduce significativamente el tiempo de

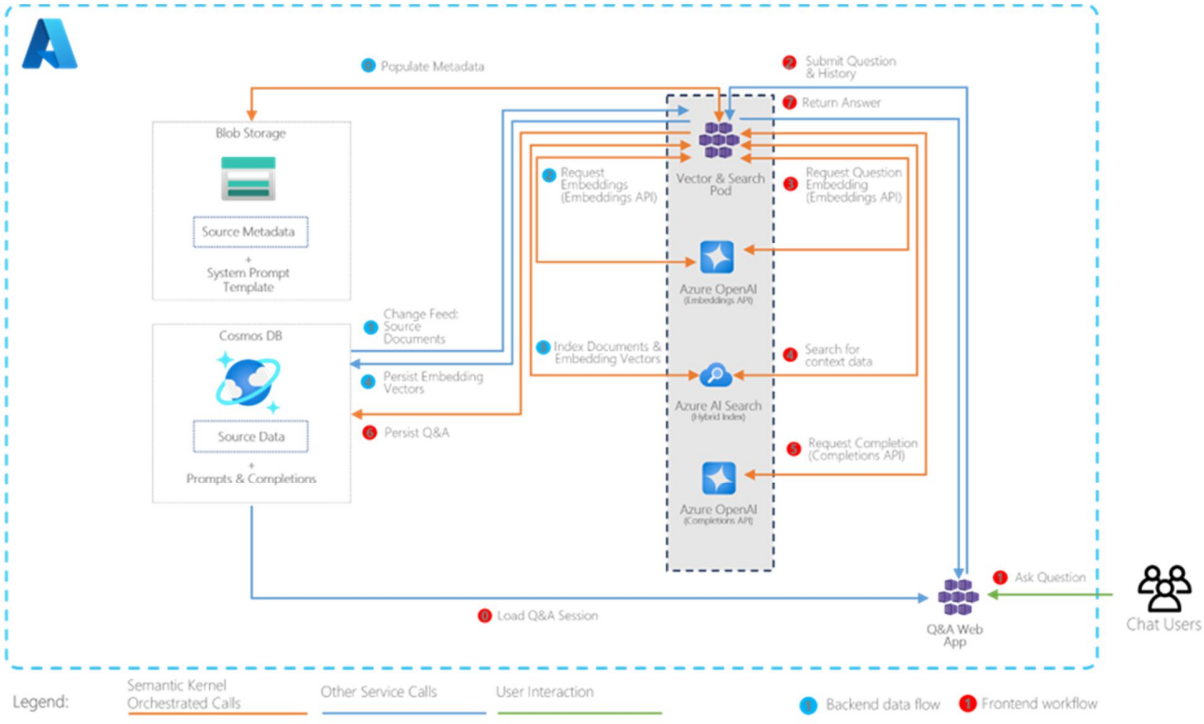
búsqueda y procesamiento, ya que solo se extraen y procesan las partes relevantes de los documentos.

2. Arquitectura Detallada de la Solución YoCampo

La implementación de la solución YoCampo se basa en una arquitectura de nube escalable y modular, utilizando tecnologías avanzadas de Azure como Cosmos DB, Azure Cognitive Search y Azure OpenAI. Esta implementación está diseñada para gestionar eficientemente la ingesta de datos, la búsqueda semántica y la generación de respuestas contextualizadas.

La arquitectura de solución representada en la Figura 2 RAG (Retrieval-Augmented Generation) se compone de los siguientes elementos principales:

Figura 2. Ejemplo de mapa de contexto general



Fuente: [2](2024).

- Blob Storage: Almacena metadatos y plantillas de prompts del sistema.
- Cosmos DB: Guarda datos fuente y vectores de incrustaciones, así como preguntas y respuestas (Q&A).
- Azure OpenAI: Proporciona API para incrustaciones y completaciones, que se utilizan para generar vectores y respuestas.
- Azure AI Search: Realiza búsquedas contextuales en los datos indexados.
- Vector & Search Pod: Maneja las solicitudes de incrustaciones y búsquedas vectoriales.

2.1. Flujo de Ingesta y Procesamiento de Datos

La ingesta y procesamiento de datos se puede detallan en los siguientes componentes de la arquitectura:

2.1.1. Uso de Cosmos DB para Almacenamiento y Gestión de Datos

En YoCampo, Cosmos DB actúa como la base de datos principal para almacenar tanto los datos originales provenientes de fuentes como BAC, Linkata y Siembra, como los resultados procesados (incluyendo embeddings, prompts y respuestas). Cosmos DB es una base de datos NoSQL de baja latencia que facilita la gestión de datos no estructurados y semi-estructurados, ideal para almacenar grandes volúmenes de documentos heterogéneos.

El flujo de ingesta se realiza mediante Azure Data Factory, que captura los datos en tiempo real (streaming) o por lotes desde las fuentes y los almacena en Cosmos DB. Esta base de datos no solo permite el almacenamiento, sino también la persistencia y rápida consulta de estos datos cuando es necesario generar respuestas a partir de consultas.

2.1.2. Azure Cognitive Search para la Indexación y Búsqueda Semántica

Una vez que los datos han sido almacenados en Cosmos DB, son indexados mediante Azure Cognitive Search. Esta herramienta permite realizar búsquedas avanzadas sobre el contenido de los documentos, utilizando técnicas de búsqueda semántica y lingüística avanzada.

El proceso de indexación consiste en analizar los documentos almacenados, extraer el contenido clave (incluyendo el texto y metadatos relevantes), y crear un índice que permite recuperar los fragmentos más relevantes para una consulta específica. Esta funcionalidad es fundamental en YoCampo, ya que permite acceder de manera eficiente a grandes volúmenes de información y recuperar contenido semánticamente adecuado para las preguntas de los usuarios.

2.1.3. Embeddings API: Creación de Vectores Semánticos

Azure Cognitive Search se complementa con la Embeddings API, que crea vectores semánticos de los documentos indexados. Los embeddings son representaciones vectoriales de alto nivel que capturan el significado contextual de las palabras y frases dentro de los documentos, lo que permite realizar búsquedas semánticas más precisas.

Estos vectores se generan a partir de los datos almacenados en Cosmos DB y se indexan en Azure Cognitive Search. Al realizar una consulta, los embeddings permiten identificar no solo los documentos que contienen las palabras clave exactas, sino aquellos que tratan sobre temas relacionados de manera contextual. Este proceso mejora significativamente la precisión de las respuestas generadas por el modelo.

2.2. Generación de Respuestas mediante GPT

Sobre la generación bajo RGA el modelo Yocampo genera el siguiente flujo de actividades:

2.2.1. Uso de Azure OpenAI (GPT) para el Procesamiento de Preguntas

El modelo de Azure OpenAI (GPT) se utiliza para procesar las preguntas de los usuarios y generar respuestas coherentes y personalizadas. Cuando un usuario realiza una consulta a través de la Web App, esta consulta se convierte en un input para el modelo GPT, que luego se encarga de generar una respuesta basada tanto en los datos preentrenados como en la información recuperada por Azure Cognitive Search.

El modelo GPT es capaz de interpretar el contexto de la pregunta, relacionarlo con los datos semánticos indexados, y generar una respuesta comprensible y relevante para el usuario. Esto le permite a YoCampo ofrecer respuestas de alto valor agregado en tiempo real.

2.2.2. Flujo Completo de Generación de Respuestas (desde la pregunta hasta la respuesta)

El flujo completo de generación de respuestas en YoCampo sigue los siguientes pasos:

- Pregunta del Usuario: El usuario ingresa una pregunta a través de la interfaz de la Web App.
- Consulta en Azure Cognitive Search: La consulta es enviada a Azure Cognitive Search, que recupera los documentos y fragmentos relevantes del índice basado en embeddings semánticos.

- Generación de Respuesta: Los fragmentos recuperados son enviados al modelo Azure OpenAI (GPT), que genera una respuesta basada en la información contextual y el conocimiento almacenado en el modelo.
- Devolución de la Respuesta: La respuesta generada se envía de vuelta al usuario, proporcionando una solución precisa, basada tanto en el contexto del modelo como en los datos más recientes y relevantes de las fuentes documentales.

Este flujo asegura que las **respuestas no solo sean precisas, sino que estén basadas en el conocimiento más actualizado posible.**

2.2.3. Persistencia del Historial y Optimización de las Respuestas

YoCampo permite la persistencia del historial de consultas y respuestas, lo que facilita la mejora continua de las respuestas proporcionadas por el sistema. A medida que los usuarios realizan más preguntas, el sistema puede aprender de las interacciones anteriores y optimizar las futuras respuestas mediante técnicas de machine learning.

El historial de consultas se almacena en Cosmos DB y se utiliza para analizar patrones de búsqueda y mejorar la precisión y relevancia de las respuestas generadas por el modelo. Esto permite una personalización más detallada y una optimización continua de la plataforma.

2.3. Interacción con el Usuario

Interfaz de Usuario Q&A (Pregunta y Respuesta). La interacción del usuario con YoCampo se realiza a través de una Web App de Pregunta y Respuesta (Q&A), que actúa como el principal punto de acceso para realizar consultas. La interfaz es intuitiva y permite a los usuarios formular preguntas en lenguaje natural, las cuales son procesadas por el backend de la solución.

La aplicación está diseñada para proporcionar respuestas rápidas y contextuales, basadas en los documentos relevantes extraídos de las bases de datos del sistema.

2.3.1. Uso de la Web App para Consultas y Visualización de Respuestas

La Web App permite que los usuarios realicen consultas y reciban respuestas de manera eficiente. Una vez que la pregunta es ingresada, el sistema se encarga de procesarla en segundo plano utilizando las tecnologías de Azure mencionadas anteriormente.

La visualización de las respuestas se realiza en tiempo real y el usuario puede interactuar con los resultados, solicitar aclaraciones adicionales o realizar nuevas preguntas relacionadas. Esto crea una experiencia de usuario dinámica y adaptable.

Frontend: Interfaz de Usuario

La Web App presenta una interfaz gráfica amigable y sencilla para los usuarios, desarrollada principalmente con React. La estructura del frontend incluye los siguientes elementos clave:

- Caja de texto de consulta: Aquí el usuario ingresa su pregunta en lenguaje natural. La aplicación está diseñada para soportar una variedad de consultas relacionadas con el agro.
- Botón de envío: Envía la consulta al backend, iniciando el proceso de generación de respuesta.
- Área de visualización de respuesta: Una vez que el backend ha procesado la consulta, la respuesta generada por Azure OpenAI se muestra aquí de forma clara y concisa.

El diseño de la interfaz permite que los usuarios sin experiencia técnica puedan interactuar fácilmente con la plataforma, haciendo preguntas complejas sobre temas agropecuarios y recibiendo respuestas de alta calidad.

Backend: Orquestación de Consultas

El backend, implementado principalmente en Node.js (basado en el repositorio de GitHub), se encarga de orquestar el flujo de consultas y respuestas:

- Recepción de la consulta del usuario: Una vez que el usuario envía la pregunta, esta es enviada al backend a través de una API que maneja la comunicación entre el frontend y las herramientas de Azure.
- Envío a Azure Cognitive Search: La pregunta es analizada y enviada a Azure Cognitive Search, que recupera los fragmentos relevantes de la base de datos y documentos indexados en Cosmos DB.
- Procesamiento con Azure OpenAI (GPT): Los fragmentos recuperados por Azure Cognitive Search son utilizados por Azure OpenAI (GPT) para generar una respuesta. GPT se encarga de sintetizar la información y entregar una respuesta contextualizada y coherente.

Este flujo asegura que las respuestas no solo sean precisas, sino también relevantes, ya que se basan en el contenido más actualizado disponible en las fuentes de datos indexadas.

Respuesta Personalizada Basada en el Perfil del Usuario

El sistema también está diseñado para personalizar las respuestas según el perfil del usuario. Dado que diferentes usuarios (extensionistas, investigadores, productores) pueden necesitar respuestas de diferente nivel de profundidad, la Web App ajusta el tipo de respuesta generada por GPT:

- Respuestas técnicas para investigadores: Si el perfil del usuario es de investigador, la Web App puede generar respuestas más detalladas y técnicas.

- Respuestas prácticas para productores: Si el usuario es un productor, el sistema puede generar respuestas más prácticas y fáciles de implementar en el campo.

Este nivel de personalización permite que la plataforma sea útil para una amplia gama de usuarios dentro del sector agropecuario.

Integración con Azure OpenAI y Cognitive Search

El nivel de acceso a OPEN AI esta dado por:

Azure Cognitive Search: La Web App está integrada con Azure Cognitive Search, lo que le permite realizar búsquedas inteligentes en una gran cantidad de documentos y fuentes de información. Azure Cognitive Search se encarga de:

- Indexar documentos: Procesar documentos y convertirlos en índices semánticos que se puedan consultar.
- Recuperar fragmentos relevantes: Identificar los fragmentos más pertinentes de los documentos basados en las consultas de los usuarios.

Azure OpenAI (GPT)

El modelo de lenguaje GPT de Azure OpenAI se encarga de la generación de respuestas. Este modelo utiliza los fragmentos recuperados por Cognitive Search para generar respuestas coherentes, utilizando tanto el contenido de los documentos como su conocimiento preentrenado.

El flujo se resume en los siguientes pasos:

- El usuario ingresa la consulta en la Web App.
- La consulta es enviada a Cognitive Search, que encuentra los fragmentos de información más relevantes.
- Azure OpenAI (GPT) toma los fragmentos y genera una respuesta personalizada y coherente.
- La respuesta es devuelta al usuario en la Web App.

2.3.2. Personalización de Respuestas en Función del Perfil del Usuario

Una característica clave de la solución YoCampo es la capacidad de personalizar las respuestas en función del perfil del usuario. Esto se logra mediante la integración de datos de los usuarios (extensionistas, investigadores, productores) y la personalización de las respuestas en función de las necesidades específicas de cada uno.

El sistema puede ajustar las respuestas para ser más técnicas o prácticas, dependiendo del perfil del usuario, lo que maximiza la utilidad de la información proporcionada. A medida que el sistema aprende más sobre los patrones de búsqueda de cada usuario, las respuestas pueden volverse aún más personalizadas y relevantes.

En resumen, la arquitectura de solución RAG [3] proporciona la infraestructura necesaria para que YoCampo pueda gestionar y procesar grandes volúmenes de datos, generar respuestas en tiempo real, y ofrecer un servicio eficiente y efectivo a los investigadores, extensionistas y productores agropecuarios en Colombia.

3. Conclusiones

- Optimización en el acceso a la información agropecuaria: YoCampo, mediante su arquitectura basada en Generación Aumentada de Recuperación (RAG), proporciona un acceso rápido y preciso a información técnica crítica, optimizando la capacidad de respuesta y toma de decisiones para extensionistas, investigadores y productores del sector agropecuario.

- Integración eficaz de tecnologías avanzadas: La combinación de herramientas como Azure Cognitive Search, Cosmos DB, y Azure OpenAI (GPT) demuestra cómo las tecnologías de inteligencia artificial y búsqueda semántica pueden integrarse para crear una solución escalable y personalizada que mejora la eficiencia y precisión en la recuperación de datos y generación de respuestas.

4. Referencias

[1] Microsoft AI, «Generación aumentada de recuperación (RAG) en Azure AI Search,» 2024. [En línea]. Available: https://learn.microsoft.com/es-es/azure/search/retrieval-augmented-generation-overview.

[2] Microsoft Azure AI, «github - Vector-Search-AI-Assistant,» github , Marzo 2024. [En línea]. Available: https://github.com/Azure/Vector-Search-AI-Assistant/tree/cognitive-search-vector. [Último acceso: mayo 2024].

[3] Mondragon, V.M y García-Díaz, V, «Adaptive contents for interactive TV guided by machine learning based on predictive sentiment analysis of data,» *Springer.*

[4] Microsoft, «Azure Machine Learning prompt flow,» 01 mayo 2024. [En línea]. Available: https://learn.microsoft.com/en-us/azure/machine-learning/prompt-flow/overview-what-is-prompt-flow?view=azureml-api-2.

[5] Microsoft-architecture, «Architecting applications on Azure,» 20 06 2023. [En línea]. Available: https://learn.microsoft.com/en-us/azure/architecture/.

[6] microsoft, «Arquitectura de análisis moderno con Azure Databricks,» 23 Sep 2023. [En línea]. Available: https://learn.microsoft.com/es-es/azure/architecture/solution-ideas/articles/azure-databricks-modern-analytics-architecture.

[7] ScienceDirect IBM, «A Method for Implementation of Machine Learning Solutions for Predictive Maintenance in Small and Medium Sized Enterprises,» *ELSEVIER,* pp.

https://pdf.sciencedirectassets.com/282173/1-s2.0-S2212827120X00102/1-s2.0-S2212827120306223/, 2020.

[8] B. Peng y C. Li, «Instruction Tuning with GPT-4,» *arXiv,* 2023.

Firmado digitalmente por Victor Manuel Mondragon Maca