

Red Hat
Summit

Connect

Generación de contenidos de marketing
utilizando tecnologías de IA Generativa



Red Hat

Eduardo Sánchez

Head of Data & AI Architecture
Sopra Steria Next



Red Hat

Agenda

El caso de uso: generando contenidos de marketing

La arquitectura: RAG

La infraestructura: desplegando LLMs en OpenShift

El futuro: desplegando sistemas de agentes en arquitecturas elásticas

El caso de uso: generando contenidos de marketing

Para la generación de contenidos de marketing se ha utilizado la arquitectura RAG desarrollada en SopraSteria como acelerador al proyecto.

Para ello, ha sido necesario indexar las últimas notas de prensa, así como la información de contexto adicional como pueden ser iniciativas, discursos o planes estratégicos.

Esto genera una nota de prensa en base a un prompt conciso que proporcione un usuario. La visualización permite generar notas de prensa y tener en cuenta el histórico de conversación con el usuario.

Tanto para los embeddings como para la generación se han utilizado modelos open-source en castellano y se ha desplegado en hardware generalista sin necesitar de GPU ni de servicios en la nube.

Notas de Prensa

Describe la nota de prensa:

Genera una propuesta de nota de prensa para presentar el segundo congreso de inteligencia artificial que tendrá lugar en Granada en Noviembre de 2024 y enlaza con todas las iniciativas ya en curso en el ámbito de la inteligencia artificial

Submit

Título: "Segundo Congreso de Inteligencia Artificial de Granada: Uniendo Iniciativas y Avanzando hacia el Futuro"

Granada, una vez más, se convierte en el epicentro de la Inteligencia Artificial (IA) en Andalucía, albergando el Segundo Congreso de Inteligencia Artificial en noviembre de 2024. Este evento, que sigue al exitoso congreso inaugural del año pasado, continuará destacando las posibilidades que la IA ofrece a la sociedad y las iniciativas en IA que se están desarrollando en Andalucía.

El congreso reunirá a expertos nacionales e internacionales para abordar las tendencias que marcan el desarrollo de esta tecnología y anticipar las claves que guiarán su evolución en los próximos años. Este evento es una oportunidad única para explorar la colaboración con la IA y cómo puede transformar nuestro camino hacia el futuro.

Además, el congreso se enlaza con otras iniciativas en curso en el ámbito de la IA en Andalucía. Entre ellas, la creación de un Radar de IA, cuyo propósito es informar sobre la evolución y las tendencias de la IA, atender demandas de formación y conocimientos sobre IA, y realizar informes técnicos o funcionales sobre aspectos de la IA.

También se está impulsando la difusión de iniciativas de 'IA made in Andalucía' a través de una estrategia multicanal que aprovecha los medios de comunicación tradicionales, las redes sociales y otros canales que otorgan a las iniciativas llevadas a cabo en la región una visibilidad notable en el territorio y otros ecosistemas.

El Segundo Congreso de Inteligencia Artificial de Granada es una oportunidad para no solo conocer las posibilidades que ofrece la IA a la sociedad en su conjunto, sino también para visibilizar las iniciativas en IA que se han desarrollado en Andalucía en cualquier ámbito, y posicionar a Andalucía en el panorama nacional e internacional.

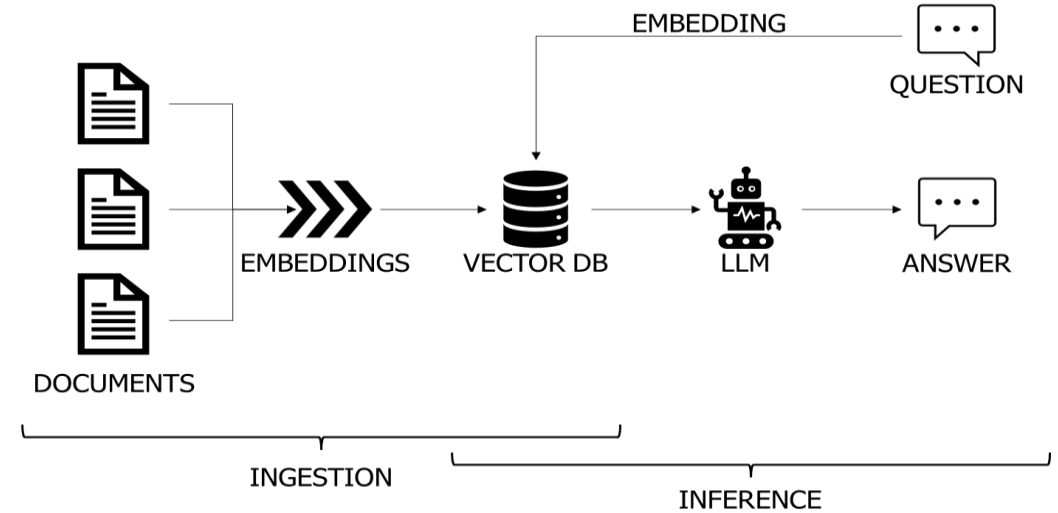
La arquitectura: RAG

La forma óptima para resolver sistemas de búsqueda inteligente y generación de contenido textual dentro de un contexto acotado es mediante una arquitectura RAG (Retrieval-Augmented Generation).

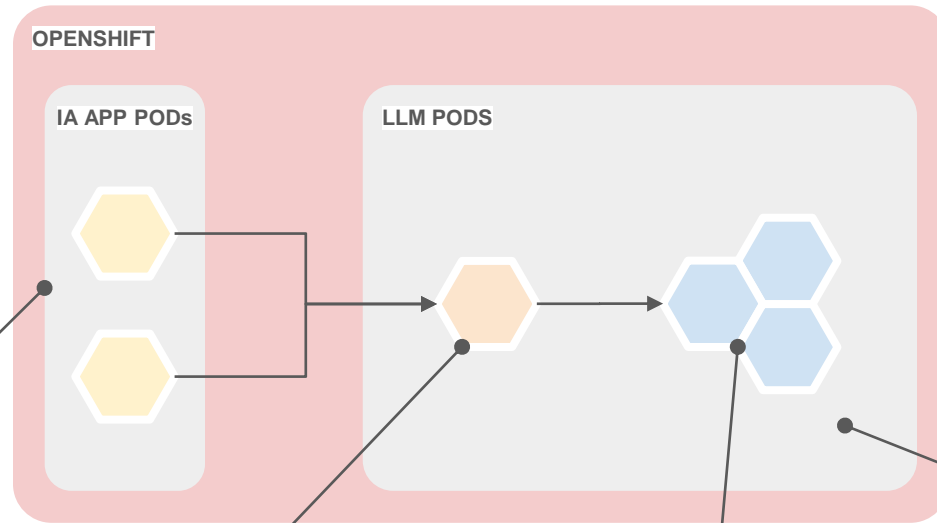
Este tipo de arquitecturas permiten indexar semánticamente conjuntos de documentos acotados en un dominio específico, guardando estos índices semánticos en una base de datos vectorial.

Cuando se produce una consulta, ya sea una búsqueda o una petición de generación de contenido, esta solicitud también se analiza semánticamente y, de la base de datos vectorial, se extraen los fragmentos de documentos similares para que el modelo de lenguaje (LLM) lo utilice como contexto.

De esta forma, el modelo de lenguaje es capaz de producir una respuesta acotada en contexto a la documentación específica y, habitualmente, sin alucinaciones.



La infraestructura: desplegando LLMs en OpenShift



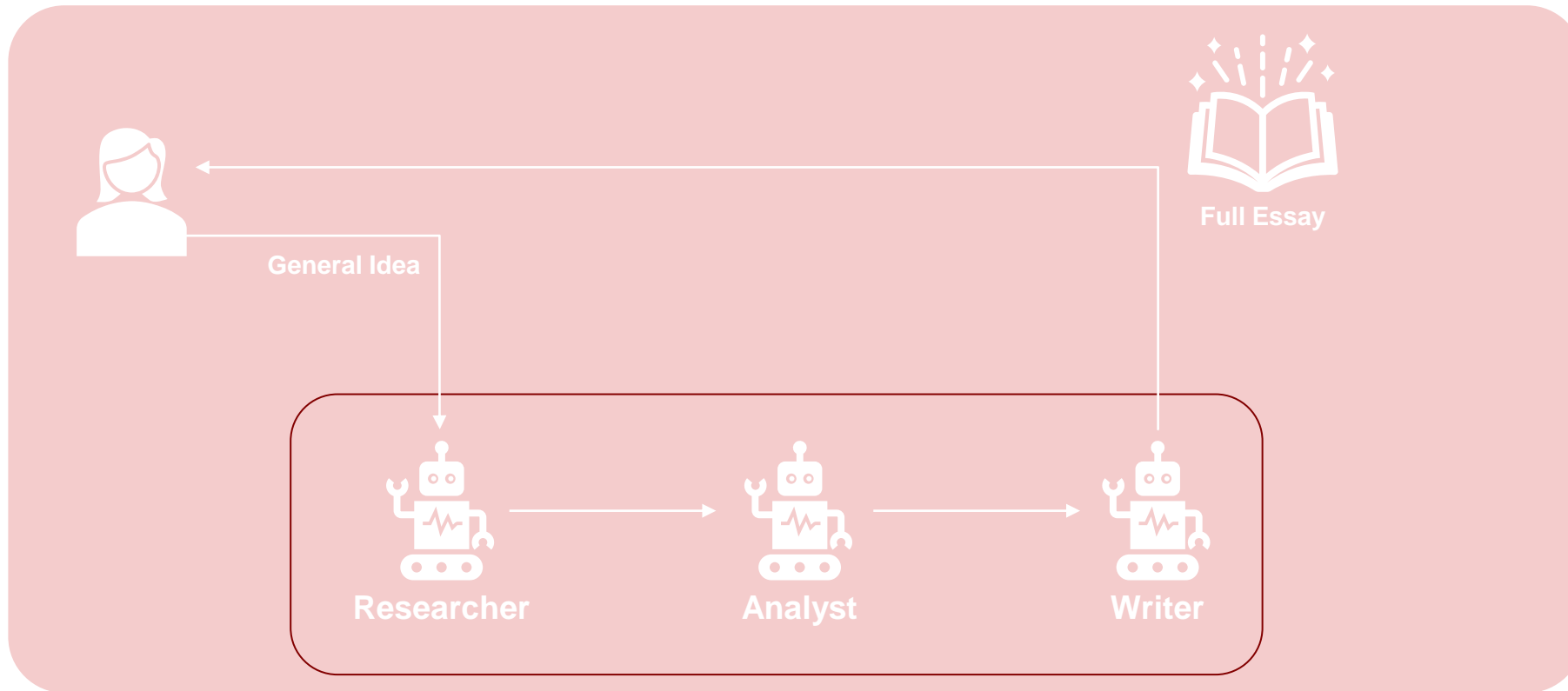
AI apps are deployed using LLMs as external services

LLMs are deployed as REST services behind a load balancer

LLMs pods can scale horizontally, as they are stateless

Current LLMs are meant to be deployed with or without GPUs

El futuro: desplegando sistemas de agentes en arquitecturas elásticas



Q&A

