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

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

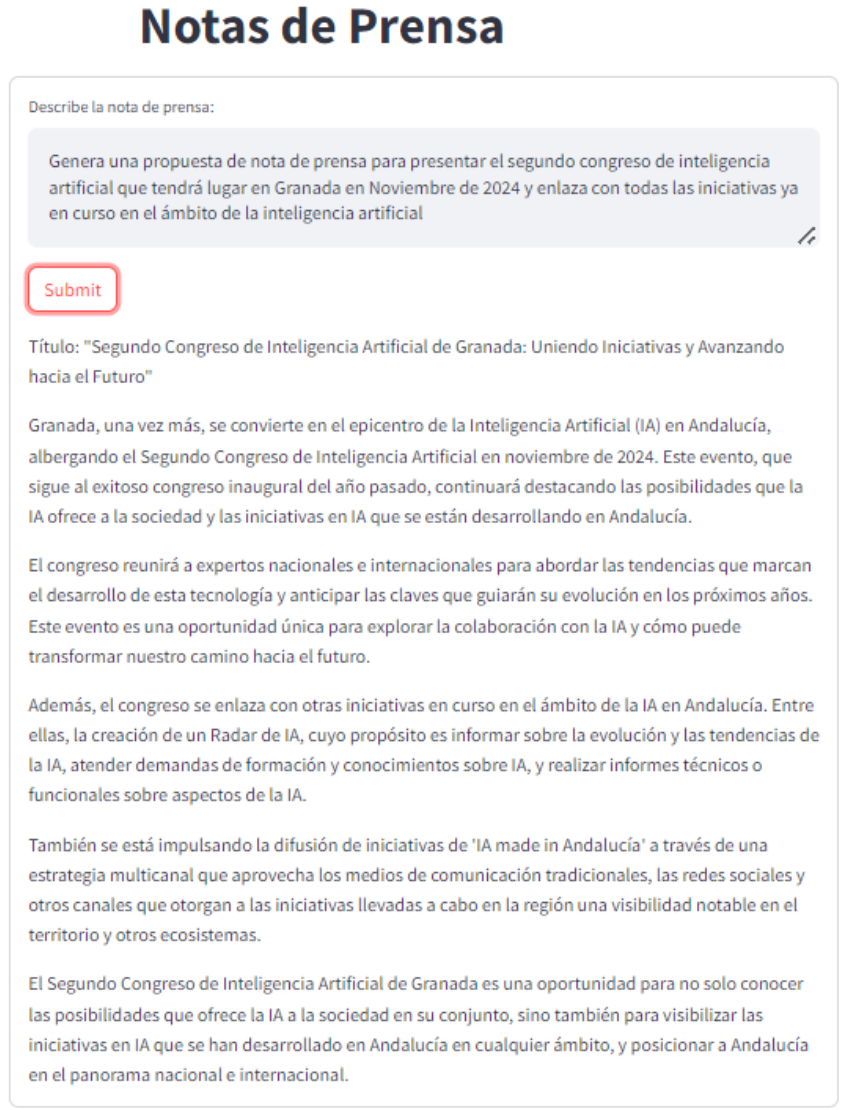
El caso de uso: generando contenidos de marketing

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.



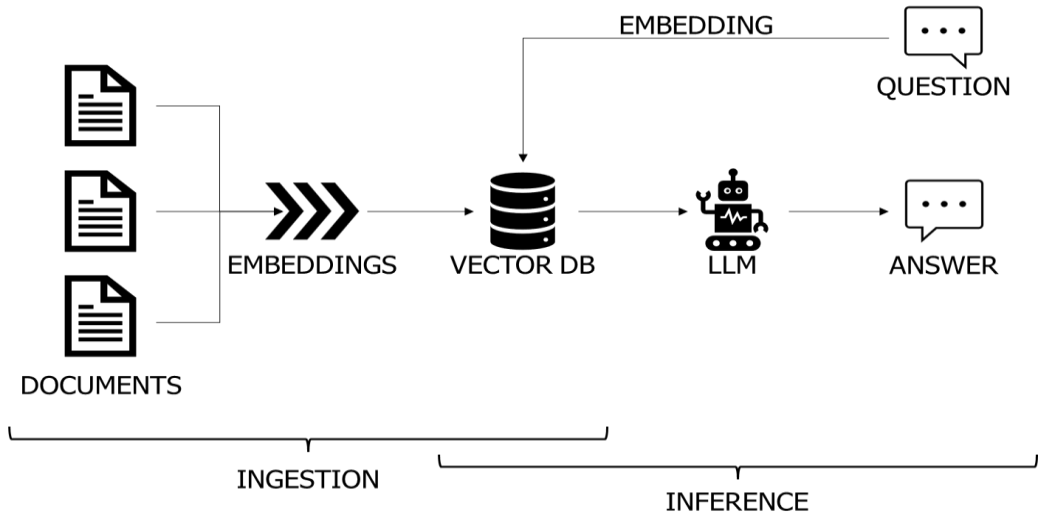
La arquitectura: RAG

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).

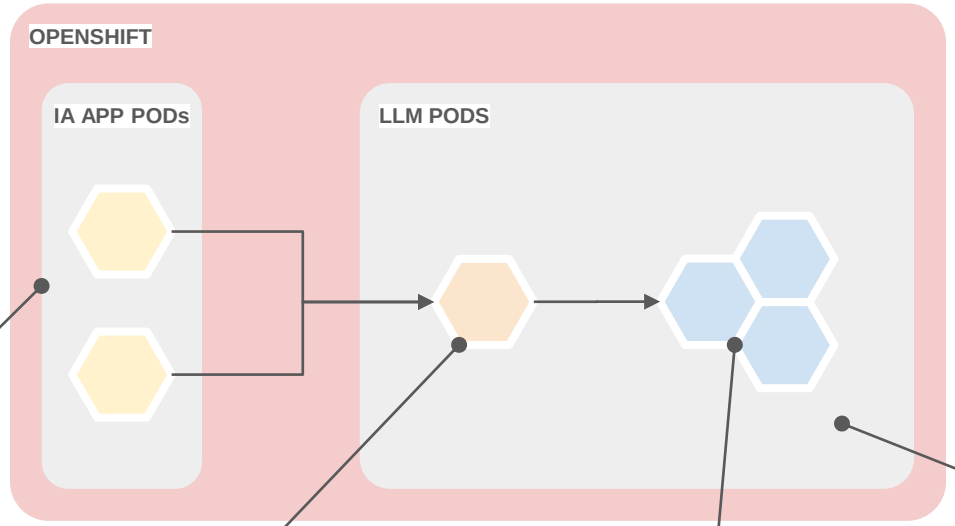
Este tipo de arquitecturas permiten indexar semánticamente conjuntos de documentos acotados en un dominio específico, guardando estos índices semánticos en una base de datos vectorial.

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



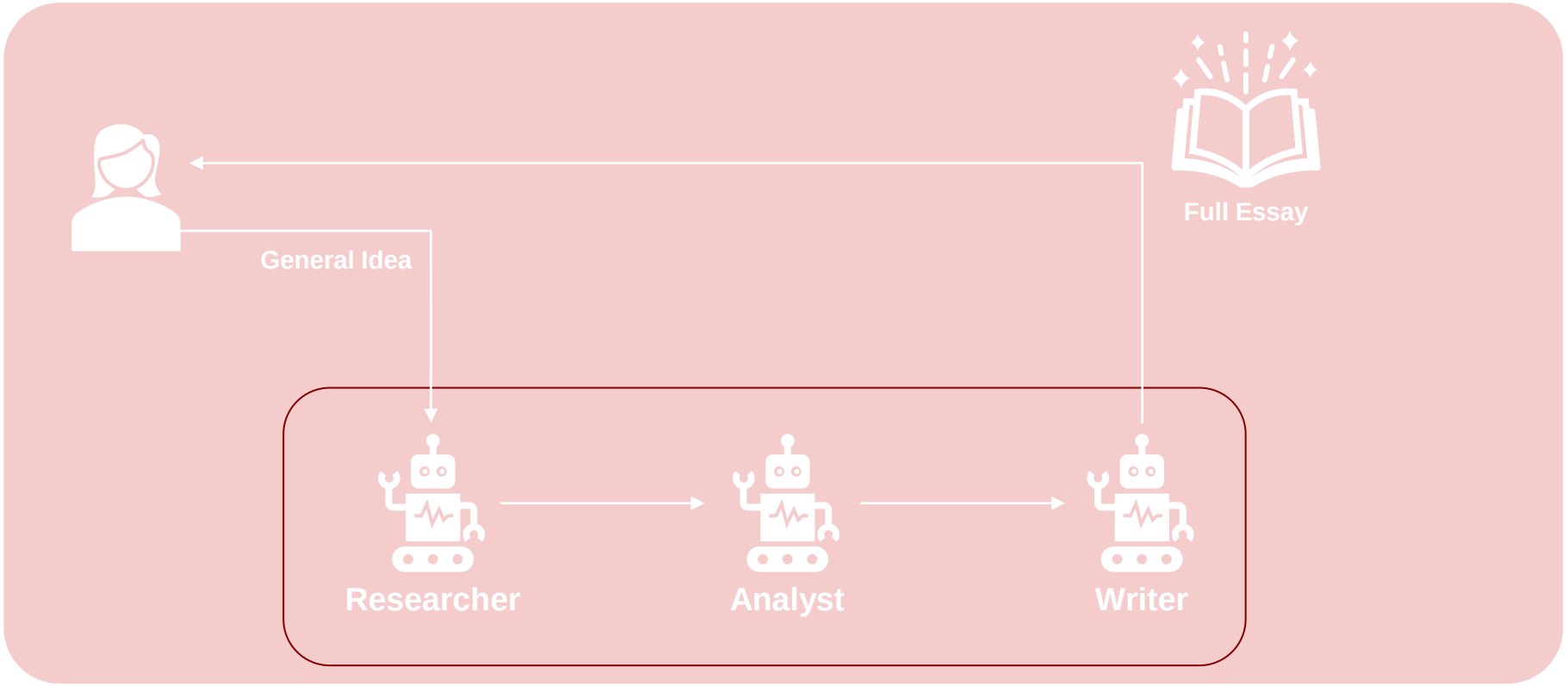
AI apps are deployed using LLMs as external services

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