Propuesta de Arquitectura de Expertos: Sistema Híbrido RAG (Retrieval-Augmented Generation)

Para: Pedro (InfoSubvenciones) De: [Tu Nombre/Nombre de la Consultoría]
Fecha: 24 de Noviembre de 2025
Asunto: Implementación de un sistema de Inteligencia Artificial para la extracción precisa y la consulta avanzada de datos de convocatorias de subvenciones.

Resumen Ejecutivo

La presente propuesta detalla la implementación de una arquitectura de Búsqueda Aumentada por Recuperación (RAG) de última generación. A diferencia de las soluciones estándar que encarecen el procesamiento de grandes volúmenes, nuestra estrategia se centra en la **eficiencia económica**. Al aplicar la **"Fase de Ingesta Inteligente"** (sección 1.5), transformamos los 90GB de texto bruto en metadatos ligeros y vectorizados, logrando una **reducción de costes de procesamiento superior al 99%** en la fase inicial, que es tradicionalmente la más costosa.

Utilizaremos la potencia de los Modelos de Lenguaje Grande (LLMs) de Google (Gemini) y una base de datos híbrida (Supabase) para transformar los documentos en **datos estructurados y consultables**. Este enfoque asegura un **"Experto GPT"** capaz de responder con alta precisión, citando fuentes específicas y filtrando información con la fiabilidad de una base de datos tradicional.

1. La Arquitectura del "Experto Inteligente"

El sistema se divide en tres fases críticas para garantizar la velocidad, la precisión y la rentabilidad del servicio.

1.1. Fase de Ingesta Inteligente (Smart Ingestion)

Esta fase sustituye el simple "troceado" de documentos por un proceso de análisis impulsado por IA.

Componente	Objetivo	Tecnología
Extractor LLM	Analizar el PDF completo para extraer campos clave y generar un resumen semántico de alta calidad.	**Gemini 2.5 Flash-Lite**

API de InfoSubvenciones	Fuente de datos (metadatos, estado, fechas, región).	Python Script
Deduplicación	Identificar y evitar el procesamiento repetido de "Bases Reguladoras" idénticas.	Checksum (Hash MD5)

El Valor Añadido: Utilizamos Gemini para generar un **Resumen de Experto** (aprox. 200 palabras) que destaca el *propósito*, los *beneficiarios específicos* y las *exclusiones*. Este resumen es la clave para la búsqueda semántica, ya que es mucho más denso y preciso que un fragmento de texto aleatorio.

1.2. Fase del Cerebro: Base de Datos Híbrida (Supabase)

Almacenaremos los datos de forma dual en una única tabla, combinando la precisión de SQL con la comprensión semántica de los vectores.

Tipo de Dato	Columna de la Base de Datos	Función en la Consulta
Datos Estructurados (SQL)	region, deadline, beneficiary_type, numConv	Filtro **Obligatorio** (Ejemplo: WHERE region = 'Madrid' AND deadline > NOW()).
Datos Vectoriales (Embeddings)	summary_embedding	Búsqueda por **Significado** (Ejemplo: "ayudas para equipamiento de artes escénicas").
El Documento Fuente	full_text_reference (Link al PDF/texto completo)	La **Única Fuente de la Verdad** (se usa solo para la respuesta final).

1.3. Fase del Experto: El Módulo de Respuesta (Parent Document Retrieval)

Este es el proceso de cuatro pasos que garantiza la precisión de la respuesta final:

1. **Filtro SQL:** El sistema aplica filtros duros (región, tipo de beneficiario) usando la tabla de metadatos SQL.

2. **Búsqueda Semántica:** Se realiza una búsqueda vectorial sobre los **resúmenes de alta calidad** de los documentos filtrados. Se seleccionan los 3 mejores candidatos.
3. **Recuperación del Documento Padre:** El sistema toma el link (full_text_reference) de los 3 candidatos y recupera la sección relevante del **texto original completo** del PDF.
4. **Generación de la Respuesta:** El texto original (la "verdad") se inyecta como contexto a un LLM de alta gama (**Gemini 2.5 Flash**) con la instrucción de responder **únicamente** basándose en dicho texto y citando el artículo o la base reguladora.

Ejemplo de Consulta Avanzada (Solo posible con el sistema híbrido):

Pregunta del Usuario	**Proceso Híbrido**	**Respuesta del Experto (Garantía de Cita)**
*"Quiero ayudas para mi **agencia de publicidad en Cataluña**, solo las que cierran en **2026**."*	1. **SQL Filter:** region='Cataluña' AND deadline > 2026. 2. **Vector Search:** Busca publicidad en los resúmenes.	*"La convocatoria 'Impulso Digital Plus' (ID 4567) de la Agencia Catalana de Comercio está abierta hasta el 15/03/2026. Según la Base 3.1, su agencia es elegible."*
*"¿La subvención para teatro de Madrid cubre **gastos de viaje internacional**?"*	1. **SQL Filter:** region='Madrid'. 2. **Vector Search:** Busca teatro y viaje internacional. 3. **Retrieval:** Carga la sección COMPLETA de "Gastos Subvencionables" del PDF.	*"Sí. El Artículo 14 de las Bases establece que los gastos de viaje para representaciones internacionales están cubiertos al 100%, hasta un máximo de 4.000€ por beneficiario."*

1.4. ¿Qué son los Embeddings (o Búsqueda Vectorial)?

El concepto de *Embedding* es clave para que el sistema entienda el significado, y no solo las palabras.

Imagina un texto ("ayuda para contratar personal técnico") y a ese texto se le asignan miles de coordenadas en un espacio matemático. Este punto es el *vector* o *embedding*.

- **Búsqueda tradicional:** Requiere que uses las palabras exactas (personal, técnico). Si buscas staff, no encontraría el documento.
- **Búsqueda Vectorial:** El sistema traduce tu pregunta (ayuda para contratar staff) a un vector y busca otros vectores que estén **cerca** en ese espacio. Dos vectores cerca significan que los textos tienen un significado similar, aunque usen palabras diferentes.

Esto permite que el sistema responda a preguntas conceptuales y no solo a búsquedas de palabras clave.

1.5. Estrategia de Vectorización Económica: El Reto de los 90GB

La arquitectura está diseñada específicamente para manejar grandes volúmenes de datos (90GB+) con un coste mínimo en la fase de ingesta, resolviendo la preocupación sobre la escalabilidad inicial.

1. **Evitar Vectorizar el Volumen Bruto:** Vectorizar 90GB de texto completo (que se traduce en miles de millones de tokens y miles de dólares) sería prohibitivo.
2. **Estrategia de Resumen Inteligente (La Clave):** En lugar de vectorizar el PDF entero, solo vectorizamos el **Resumen de Experto** (aprox. 200 palabras) que el LLM genera.
3. **Reducción Masiva de Tokens y Costes:** Si un PDF de 30 páginas se convierte en un resumen de 200 palabras, reducimos la cantidad de texto a vectorizar en más del 90%. Este ahorro se traduce directamente en una **reducción drástica de costes de procesamiento inicial** y un menor requerimiento de RAM en la base de datos de vectores.
4. **Coste Fijo y Único:** Este proceso de vectorización solo ocurre una vez por documento. Usando modelos ligeros y optimizados para esta tarea (como Gemini 2.5 Flash-Lite), garantizamos un coste muy bajo y predecible para incorporar miles de documentos.

En resumen: Al vectorizar solo la esencia (el resumen), convertimos un reto de volumen costoso en una tarea de *procesamiento de metadatos* barata, manteniendo la precisión.

2. Presupuesto y Costes Operativos

El principal objetivo de esta arquitectura es mantener los costes variables al mínimo, migrando la mayor parte del procesamiento a la fase inicial de ingesta, que es poco frecuente y mucho más económica que la generación de respuestas.

2.1. Costes de Infraestructura (Supabase)

Utilizaremos Supabase como motor principal de la base de datos híbrida, combinando PostgreSQL (SQL) con el motor de vectores pgvector.

Concepto	Coste Estimado	Notas
Plan Pro de Supabase	**$25.00 USD/mes**	Incluye 8GB de base de datos y 100GB de almacenamiento de archivos. La estrategia de vectorización ligera evita la necesidad de grandes

		cantidades de RAM para el índice de vectores.
Almacenamiento Adicional	$0.125 USD/GB	Coste escalable solo si la base de datos supera los 8GB.
TOTAL INFRAESTRUCTURA	**A partir de $25 USD/mes**	Coste fijo, bajo y predecible para un entorno de producción.

2.2. Costes de Inteligencia Artificial (Gemini API)

Utilizaremos dos modelos para optimizar el coste: Flash-Lite para tareas de volumen (Ingesta) y Flash para respuestas de calidad (Consulta).

Concepto	Modelo Usado	Coste por 1M de Tokens	Estimación Mensual (1.000 consultas/mes)
Ingesta de Documentos (Input)	Gemini 2.5 Flash-Lite	$0.10 USD/M tokens	~$0.25 USD/mes
Ingesta de Documentos (Output)	Gemini 2.5 Flash-Lite	$0.40 USD/M tokens	~$1.00 USD/mes
Consulta del Usuario (Input)	Gemini 2.5 Flash	$0.30 USD/M tokens	~$3.00 USD/mes
Consulta del Usuario (Output)	Gemini 2.5 Flash	$2.50 USD/M tokens	~$1.25 USD/mes
TOTAL ESTIMADO AI			**Aprox. $5.50 USD/mes**

PRESUPUESTO OPERATIVO TOTAL ESTIMADO (MENSUAL): $30 – $50 USD

3. Honorarios de Desarrollo e Implementación

Este coste cubre el diseño, desarrollo, implementación, pruebas y documentación completa

de la arquitectura RAG Híbrida Inteligente.

Servicio	Descripción	Honorarios (Tarifa Fija)
Fase 1: Configuración de Infraestructura	Creación y optimización de la instancia de Supabase (tablas, pgvector, autenticación). Configuración del entorno de desarrollo.	€
Fase 2: Script de Ingesta Inteligente	Desarrollo del script Python para la API de InfoSubvenciones. Lógica de deduplicación. Implementación del *prompt* de extracción JSON con Gemini 2.5 Flash-Lite.	€
Fase 3: Módulo de Expertos y Query	Desarrollo del *endpoint* de consulta (API). Implementación de la lógica de **Filtro SQL + Búsqueda Vectorial**. Implementación de la **Recuperación del Documento Padre** y la llamada final a Gemini 2.5 Flash para la generación de la respuesta citada.	€
Fase 4: Pruebas y Documentación	Pruebas de precisión con casos críticos. Documentación completa del código y del proceso de despliegue.	€
TOTAL HONORARIOS DE IMPLEMENTACIÓN		**€ - € (Tarifa Fija)**

Este modelo garantiza una inversión inicial fija para la construcción del sistema, seguida de unos costes operativos mensuales mínimos y altamente predecibles.

Agradezco su tiempo y estoy a su disposición para discutir cualquier detalle técnico o ajustar el cronograma y las tarifas.