



**Instituto
Europeo
de Posgrado**

AGENTE DE IA ONPREMISE Y DATAWAREHOUSE INTEGRADO: ECOSISTEMA DE INTELIGENCIA
INSTITUCIONAL PARA LA AUDITORÍA GENERAL DE LA REPÚBLICA DE COLOMBIA (AGR)

Trabajo fin de estudio presentado por:	Miguel Leonardo Aragón Garzón Nuria Leal Quintana
Tipo de trabajo:	Investigación-Acción
Director/a:	Julián Camilo Ortega Gómez
Fecha:	Marzo 27, 2026

Índice

1	Resumen.....	3
2	Abstract	3
3	Introducción.....	5
3.1	Objetivo General.....	5
3.2	Objetivos Específicos.....	5
3.3	Metodología.....	5
3.3.1	Identificación y Análisis del Problema.....	6
3.3.2	Planificación de la Solución	6
4	Identificación y descripción un problema al interior de la organización	8
4.1	Descripción de la empresa a intervenir.....	9
4.1.1	Misión	9
4.1.2	Visión	9
4.1.3	Mapa de procesos	9
4.1.4	Descripción del proceso a intervenir.....	9
4.1.5	Descripción del problema que presenta el proceso	10
5	Establecimiento de planes de acción para la solución del problema.....	11
5.1	Establecimiento de objetivos.....	11
5.2	Descripción de la solución.....	11
5.2.1	Componentes Open Source por Capas	11
5.3	Secuencia de actividades para ejecución de la solución.....	12
5.4	Descripción de recursos necesarios para la ejecución de la solución.....	13
6	Definición del alcance de la solución, partes interesadas, cronograma y presupuestos.....	15
6.1	Alcance final de la solución	15
6.2	Identificación de responsables de las actividades para ejecución	15
6.3	Cronograma de ejecución.....	16
6.4	Presupuesto.....	16
7	Referencias bibliográficas	18

1 Resumen

El presente proyecto de grado desarrolla el diseño e implementación de una arquitectura de Inteligencia Artificial Generativa local (On-premise) para la Auditoría General de la República (AGR), orientada a optimizar la vigilancia y control fiscal en Colombia. La solución propone un Agente de IA basado en la técnica de Generación Aumentada por Recuperación (RAG), diseñado para operar en un entorno de red privada sin salida a Internet, garantizando la privacidad y gobernanza del dato y seguridad de la información sensible.

El sistema integra datos de cuatro aplicativos misionales estructuralmente diversos: SIA Misional y SIA Contralorías (arquitectura de microservicios en .NET 6/8 con PostgreSQL) y SIA Observa y SIA Fiscal (arquitectura monolítica en .NET Framework 4.8 con SQL Server). Para centralizar esta información sin afectar el rendimiento operativo, se diseña un Datawarehouse (DWH) alimentado por réplicas de bases de datos y repositorios de archivos (PDF, Word, Excel, Audio). La infraestructura técnica se fundamenta en servidores de alto rendimiento con aceleración de cómputo mediante GPUs externas vía Thunderbolt 4¹ y almacenamiento NVMe (opcional). Finalmente, el proyecto se alinea con la norma ISO 42001:2023², estableciendo un marco de gobernanza, ética y gestión de riesgos para sistemas de IA en el sector público.

Palabras Clave: AGR, Inteligencia Artificial On-premise, RAG (Retrieval-Augmented Generation), Control Fiscal, Microservicios .NET, ISO 42001:2023, Datawarehouse, Agente de IA, GGUF (GPT-Generated Unified Format).

2 Abstract

This master's thesis presents the design and implementation of a Local Generative Artificial Intelligence (On-premise) architecture for EntidadXX (AGR), aimed at optimizing fiscal oversight and control in Colombia. The solution proposes an AI Agent based on the Retrieval-Augmented Generation (RAG) technique, specifically designed to operate within a private network environment without Internet access, thereby ensuring data sovereignty and the security of sensitive information.

The system integrates data from four structurally diverse mission-critical applications: SIA Misional and SIA Contralorías (.NET 6/8 microservices architecture with PostgreSQL) and SIA Observa and SIA Fiscal (.NET Framework 4.8 monolithic architecture with SQL Server). To centralize this information without compromising operational performance, a Datawarehouse (DWH) is designed, fed by database replicas and file repositories (PDF, Word, Excel, Audio). The technical infrastructure is based on high-performance servers

¹ Términos técnicos como Thunderbolt 4 y RAG son términos estándar de la industria por tal motivo se mencionan sin traducción.

² Norma técnica internacional para guiar a las empresas a gestionar los riesgos inherentes al uso de la IA en las organizaciones (ISO Tools, Consultado en marzo 2026).

with computational acceleration through external GPUs via Thunderbolt 4 and NVMe storage (optional). Finally, the project is aligned with the ISO 42001 standard, establishing a governance, ethics, and risk management framework for AI systems within the public sector.

Keywords: AGR, On-premise Artificial Intelligence, RAG (Retrieval-Augmented Generation), Fiscal Control, .NET Microservices, ISO 42001:2023, Datawarehouse, AI Agent, GGUF (GPT-Generated Unified Format).

Instituto Europeo de Posgrado ©

3 Introducción

Este proyecto de aplicación en investigación-acción se centró en la identificación y resolución de **la complejidad en las consultas transversales de información existente en varios aplicativos misionales y repositorios documentales en la Auditoría General de la República (AGR)**, permitiendo integrar teoría y práctica en un entorno real. Mediante un enfoque de investigación-acción, se desarrollaron habilidades de diagnóstico, análisis y planificación orientadas a mejorar la eficiencia y efectividad de un proceso crítico en la organización. La estructura del proyecto incluyó una revisión exhaustiva del contexto organizacional y la implementación de una solución factible que responde a las necesidades de la organización y optimiza su desempeño.

3.1 Objetivo General

El objetivo principal del proyecto fue desarrollar una solución operativa al problema **la dificultad para realizar análisis transversales (datos insumos para construcción de indicadores de gestión por dominio de datos) y consultas en lenguaje natural sobre la información distribuida en los diversos aplicativos misionales y repositorios documentales de la AGR**. La intervención se orientó no solo a corregir el problema identificado, sino también a proporcionar una propuesta sostenible que pudiera integrarse a la estructura de la organización, elevando los niveles de eficiencia y alcanzando los estándares de desempeño requeridos en el proceso intervenido.

3.2 Objetivos Específicos

Para cumplir con el objetivo general, se establecieron tres objetivos específicos:

1. Realizar una identificación detallada del problema en su contexto organizacional, determinando causas raíz y evaluando su impacto en el desempeño general del proceso.
2. Diseñar un plan de acción estructurado con objetivos claros, actividades específicas y recursos definidos para la ejecución de la solución.
3. Establecer el alcance de la intervención, definiendo responsables, cronograma y presupuesto, para asegurar que la implementación sea organizada y efectiva.

3.3 Metodología

El proyecto se desarrolló mediante la metodología de investigación-acción, la cual facilitó la colaboración continua entre el equipo de estudiantes y docente asesor. Esta metodología permitió un enfoque sistemático para identificar problemas, formular hipótesis y aplicar soluciones en un contexto real, ajustando las estrategias conforme se avanzaba en el proyecto. La investigación-acción proporcionó un marco flexible y

participativo para ejecutar el proyecto de manera estructurada, generando una intervención bien fundamentada y con alto potencial de impacto.

El desarrollo del proyecto se estructuró en tres etapas secuenciales, cada una de las cuales abordó un aspecto específico de la intervención. En la primera etapa, se identificó y analizó el problema organizacional, incluyendo una descripción detallada de la organización, su misión, visión y el proceso específico afectado. La segunda etapa se centró en la planificación de la solución, estableciendo objetivos y actividades concretas para abordar el problema. En la tercera fase, se definieron aspectos operativos como el cronograma, el presupuesto y la asignación de roles.

3.3.1 Identificación y Análisis del Problema

En la primera etapa, se realizó un análisis exhaustivo de la organización y del contexto en el cual se desarrolla el proceso afectado. Se documentó el problema identificando sus causas raíz, efectos y puntos críticos que afectaban el desempeño del proceso. Este diagnóstico inicial fue esencial para estructurar la solución y asegurar que la intervención respondiera de manera precisa a las necesidades de la organización.

Para el diagnóstico de la problemática descrita como objetivo general, se utilizará el método de Diagnóstico interdisciplinario y multidimensional de Activos de Información que posee la entidad, donde se estructurará en tres técnicas fundamentales:

- Mediante entrevistas diseñadas como formularios, donde se realice el levantamiento de datos sobre los Activos de Información, con el objetivo de mapear el flujo de datos según el proceso y procedimientos que lleve la dependencia entrevistada.
- Auditoría sobre las Fuentes de Datos, posterior a la recolección y mapeo de los flujos de datos, se ejecuta el análisis técnico sobre los esquemas de las bases de datos de los aplicativos misionales y sobre los repositorios documentales. Esto permitirá identificar documentos según el **dominio** del dato y sus características principales.
- Diagrama de Ishikawa, herramienta fundamental para detectar los efectos a partir de las causas raíz sobre tres ejes: sobre Tecnología (burbujas de datos), Procesos y sus Procedimientos (tiempos de búsqueda manual), Personal (curva de aprendizaje para manejo de los sistemas de información) y el Entorno (basado en el marco normativo institucional y de cumplimiento que se rige en la Entidad, justificando por qué la solución se implementará On Premise (por temas de seguridad y privacidad de la información) y no en Nube o en Nube Híbrida).

3.3.2 Planificación de la Solución

En la segunda etapa, se desarrolló un plan de acción detallado que incluyó la definición de objetivos específicos y una secuencia de actividades que permitiría mitigar o resolver el problema identificado. Esta fase incluyó el diseño de una estrategia de ejecución organizada y estructurada que facilitara el logro de los objetivos propuestos y maximizara

la eficiencia del proceso intervenido. La planificación detallada de actividades y recursos fue clave para asegurar una implementación precisa y efectiva.

3.3.2.1 Definición de Alcance y Recursos

La tercera etapa se enfocó en delimitar el alcance de la intervención y los recursos necesarios. Se establecieron responsables para cada actividad, y se elaboró un cronograma de ejecución para controlar el avance de la implementación. Asimismo, se desarrolló un presupuesto que detalló los costos asociados a cada fase, permitiendo optimizar la asignación de recursos. Esta fase aseguró que todas las variables críticas estuvieran alineadas para una ejecución sin interrupciones.

4 Identificación y descripción un problema al interior de la organización

La AGR, en su misión de vigilancia y control fiscal, ejerce esta labor mediante una planeación de auditoría programadas sobre el año fiscal inmediatamente anterior. Sobre este proceso fundamental, la Entidad recurre a la ejecución de rendiciones de cuentas sobre sus sujetos vigilados, los cuales lo conforman 67 entidades principales y 11 sujetos subordinados.

Para este proyecto, el diagnóstico se ha centrado en cuatro dependencias misionales: la Oficina de Estudios Especiales, la Oficina Delegada para la Vigilancia, la Oficina de Planeación junto con el Grupo T.I.C. y la Dirección de Responsabilidad Fiscal.

Cada una de estas áreas lidera procesos críticos que dependen directamente de la calidad del dato y la oportuna disposición de la información acopiada. A pesar de ello, la entidad tiene segmentada en "islas de datos" y catalogados en dominios de información que no interactúan entre sí.

El problema central radica en que la AGR gestiona un gran volumen de información catalogada en dominios de datos a través de sus aplicativos misionales (SIA³ Misional, Observa, Contralorías y Fiscal)

- Diversidad Tecnológica: Coexistencia de arquitecturas modernas en .NET 8 (Microservicios/PostgreSQL) con sistemas legados en .NET Framework 4.8 (Monolíticos/SQL Server). Esto impide una consulta unificada y ágil.
- Saturación de la Capacidad Analítica: Los funcionarios de las oficinas delegadas deben realizar búsquedas manuales en múltiples plataformas y bases de datos para consolidar un solo informe de responsabilidad fiscal o contratación, lo que genera cuellos de botella operativos.
- Información No Estructurada: Existe una vasta cantidad de conocimiento en repositorios de archivos (PDF, Word, Audios de audiencias) que no es indexable ni se puede consultar como si fueran "queries" SQL tradicionales, perdiendo trazabilidad crítica en los procesos de misionales.
- Restricción de Seguridad y Privacidad: Al manejar datos sensibles de sujetos vigilados y procesos de jurisdicción coactiva, la entidad no puede utilizar soluciones de IA en la nube (Cloud Public) por riesgos de fuga de datos y cumplimiento normativo, quedando rezagada en la adopción de modelos de lenguaje (LLMs).

³ SIA acrónimo de Sistema Integral de Auditoría

4.1 Descripción de la empresa a intervenir

4.1.1 Misión

"La Auditoría General de la República coadyuva a la transformación, depuración y modernización de los órganos instituidos para el control de la gestión fiscal, mediante la promoción de los principios, finalidades y cometidos de la función administrativa consagrados en la Constitución Política, el fomento de la cultura del autocontrol y el estímulo de la participación ciudadana en la lucha para erradicar la corrupción." (AGR, Principios Corporativos, Consultado en marzo 2026)

4.1.2 Visión

"En 2027, la Entidad será reconocida en el ámbito nacional e internacional, por el impacto que los resultados de su operación tienen sobre la lucha contra la corrupción y el fortalecimiento del Sistema Nacional de Control Fiscal." (AGR, Principios Corporativos, Consultado en marzo 2026)

4.1.3 Mapa de procesos



Ilustración 1. Mapa de Procesos de la AGR, Imagen mejorada con Gemini AI Nano banana 2

4.1.4 Descripción del proceso a intervenir

Dentro del mapa de procesos que se rige por el Sistema de Gestión de la Calidad en la AGR, se destacan 2 procesos misionales fundamentales para el ejercicio de la Vigilancia y el Control Fiscal a los entes vigilados.

1. Gestión del Proceso Auditor (PA)
2. Gestión de Procesos Fiscales (PF)

Las áreas que gestionan estos procesos críticos y que dependen directamente de la calidad del dato y la oportuna disposición de la información acopiada. A pesar de ello, la

entidad tiene segmentada en "islas de datos" y catalogados en dominios de información que no interactúan entre sí.

4.1.5 Descripción del problema que presenta el proceso

Los procesos misionales, aunque están bien definidos y estructurados según el Sistema de Gestión de la Calidad SGS de la AGR, los sistemas de información SIA no tienen Interoperabilidad interna entre ellos, esto imposibilita el intercambio de datos de manera automática y con disponibilidad inmediata.

Por lo anterior, se obtienen los siguientes puntos sobre el problema, las causas y los efectos:

Efectos (Consecuencias)	Problema Central	Causas (Raíz)
Retraso en la toma de decisiones fiscales y baja eficiencia en tiempos de consolidación de información en procesos sancionatorios.	Dificultad para consolidar, analizar y consultar información misional transversal, realizando consulta en lenguaje natural.	Dispersión de datos bajo múltiples arquitecturas (.NET 4.8 vs .NET 8).
Alto riesgo de errores humanos en la consolidación de reportes para la Alta Gerencia.		Inexistencia de un repositorio centralizado de conocimiento (DWH/Vector Store).
Desaprovechamiento de evidencia documental en formatos no estructurados.		Restricciones de seguridad que impiden el uso de IA externa (Internet).

5 Establecimiento de planes de acción para la solución del problema

El proyecto se ejecutará en un ciclo de vida incremental, garantizando que el Agente de IA aprenda de los dominios de datos de forma progresiva. De igual forma, este plan de acción define la hoja de ruta técnica y administrativa para transformar la gestión del control fiscal mediante inteligencia artificial soberana.

5.1 Establecimiento de objetivos

Se definen los siguientes objetivos del Plan de Acción:

1. Centralizar el Conocimiento: Integrar en un plazo de 15 semanas los 13 dominios de datos de los sistemas SIA en un repositorio único de consulta.
2. Eliminar la Fragmentación: Reducir en un 80% el tiempo de búsqueda manual de información transversal entre aplicativos SIAs.
3. Garantizar la Soberanía de Datos: Desplegar una solución 100% on-premise bajo la norma ISO 42001, sin dependencia de nubes públicas.
4. Habilitar Consultas de Lenguaje Natural: Permitir que personal no técnico genere informes complejos mediante una interfaz de chat.

5.2 Descripción de la solución

Se propone un **Agente de IA OnPremise** con arquitectura de microservicios y que utiliza una arquitectura Multi-Step RAG (M., Consultado en marzo 17 de 2026). La solución aborda la fragmentación de la AGR mediante un **Datawarehouse (DWH)** intermedio que actúa como fuente de conocimiento propio de la AGR. El uso de **Sprints** cortos (15 días por sprint) garantiza que se entreguen productos al final de cada uno, construcción de los adaptadores para los servicios web API RESTful de cada SIA, donde se prueben de manera aislada y segura antes de su integración final en el orquestador de IA. Toda la solución se implementará bajo el concepto de OpenSource.

5.2.1 Componentes Open Source por Capas

1. Capa de Inteligencia Artificial (LLM y Orquestación)

- **ollama / IA Local⁴**: Herramientas para servir modelos de lenguaje (LLMs) localmente. Permiten ejecutar modelos como **Llama 3** o **Mistral** sin salir de la red de la AGR.

⁴ Se puede obtener una estadística sobre la compatibilidad y el uso de modelos tales como Llama 3 o Llama 3.1 y mediante dispositivos eGPU como Tesla P40 o Tesla T4, se puede consultar desde el sitio web <https://www.canirun.ai/>

- **LangChain / Semantic Kernel:** Frameworks para orquestar la lógica del agente, las cadenas de pensamiento y la conexión con las herramientas (bases de datos y archivos).
- **Hugging Face Transformers:** Librerías para gestionar los modelos de **Embeddings** (como BGE-M3) que convierten el texto en vectores.

2. Capa de Datos y Almacenamiento (Vector Store & DWH)

- **Qdrant / ChromaDB:** Bases de datos vectoriales de código abierto optimizadas para búsquedas semánticas de alta velocidad.
- **PostgreSQL:** El motor para el **Datawarehouse (DWH)** intermedio. Es robusto, altamente escalable y gratuito.
- **Apache Hop / Meltano:** Herramientas de **ETL** (Extracción, Transformación y Carga) para mover los datos desde las réplicas de los SIA hacia el DWH. Mirar probabilidad de que sea Pentaho.

3. Capa de Conectividad y Backend

- **Python (FastAPI):** Para el desarrollo de los adaptadores y conectores. Es el estándar de la industria para microservicios de IA por su velocidad y ecosistema de librerías.
- **Docker / Kubernetes (K3s):** Para la orquestación de los contenedores donde correrán los conectores y el agente, facilitando el despliegue en un servidor local.

5.3 Secuencia de actividades para ejecución de la solución

❖ Fase I: Infraestructura y Datos Maestros (Meses 1-2), Gobierno de Datos.

- ✓ **Sprints 1 y 2:** Adquisición, montaje de hardware y configuración de la eGPU. Definición de la Política de IA (ISO 42001).
- ✓ **Sprints 3 y 4:** Despliegue de réplicas de bases de datos y diseño del esquema del DWH para los primeros dominios (Contratación).
- ☑ **Recursos:** Arquitecto de Infraestructura, DBA, Servidor Host, Almacenamiento NVMe.

❖ Fase II: Conectividad Misional y ETL (Meses 3-4)

- ✓ **Sprints 5 y 6:** Desarrollo de adaptadores para SIA Misional y Contralorías (.NET 8). Extracción de Cuentas Contables.
- ✓ **Sprints 7 y 8:** Desarrollo de adaptadores para SIA Observa y Fiscal (.NET 4.8). Extracción de Hallazgos y Procesos.
- ☑ **Recursos:** Desarrolladores .NET (Senior y Legacy), Python Developer, APIs RESTful.

❖ **Fase III: Cerebro de IA y Vectorización (Meses 5-6)**

- ✓ **Sprints 9 y 10:** Implementación del Vector Store (Qdrant) y pipeline de embeddings para documentos (PDF/Word/Excel).
- ✓ **Sprints 11 y 12:** Configuración del Orquestador LangChain y carga masiva de los 13 dominios al banco de conocimiento.
- ☑ **Recursos:** Ingeniero de IA, Especialista en NLP, GPU con alta vRAM.

❖ **Fase IV: Refinamiento y Pruebas de Calidad (Meses 7-8)**

- ✓ **Sprints 13 y 14:** Ciclos de "Fine-tuning" y ajuste de prompts para eliminar engaños en temas fiscales.
- ✓ **Sprints 15 y 16:** Pruebas de usuario (UAT) con funcionarios delegados y auditoría de seguridad perimetral (On-premise).
- ☑ **Recursos:** Analista de QA, Líderes de la AGR, Herramientas de evaluación (Ragas (IBM, Consultado en marzo 19 de 2026)).

❖ **Fase V: Despliegue y Mejora Sostenible (Mes 9)**

- ✓ **Sprint 17:** Despliegue controlado en la Intranet y capacitación técnica.
- ✓ **Sprint 18:** Cierre de proyecto, entrega de documentación de esta maestría y plan de mantenimiento.
- ☑ **Recursos:** Project Manager, UX Designer, Servidor de aplicaciones.

5.4 Descripción de recursos necesarios para la ejecución de la solución

En cuanto a los recursos tecnológicos, talento humano y de materiales, tenemos:

❖ **Recursos Tecnológicos (Infraestructura Virtual y GPUs Legacy)**

- **Entorno de Virtualización (Hyper-V / VMware):** En lugar de un servidor físico dedicado de última generación (por costos altos), la solución se desplegará sobre una **Máquina Virtual (VM)** dentro del clúster existente de la AGR. Se deben asignar recursos garantizados (vCPU y RAM reservada) para evitar latencia. Se recomienda una configuración de 8 a 12 vCPUs y 64GB de RAM para soportar el flujo de datos de los 13 dominios y el Datawarehouse.
- **Aceleradoras de Cómputo (GPUs de Clase Data Center Económicas):**
 - **Opción 1, NVIDIA Tesla T4:** Ideal para la fase de **inferencia** (respuesta del chat). Aunque tiene solo 16GB de VRAM, su bajo consumo energético y arquitectura Turing permiten ejecutar modelos cuantizados de forma muy eficiente en entornos virtuales. (Nvidia Tesla T4, Consultado en 3 de marzo de 2026)

- **Opción 2, NVIDIA Tesla P40:** La mejor opción para el **almacenamiento y procesamiento masivo**. Con **24GB de VRAM**, permite cargar modelos de lenguaje más grandes (Llama 3.1 8B o Mistral) con ventanas de contexto amplias, lo cual es vital para leer PDFs largos. **Nota:** Al ser tarjetas pasivas, requieren que el servidor virtualizado cuente con un chasis que soporte flujo de aire forzado o un adaptador de ventilación*. (Nvidia Tesla P40, Consultado en 3 de marzo de 2026)
- **Almacenamiento SSD / NVMe Virtualizado:** El Vector Store (Qdrant) se alojará en discos de estado sólido de alta velocidad. Si el almacenamiento central de la entidad es híbrido, se debe priorizar el uso de un *Tier* de SSDs para las bases de datos vectoriales para asegurar que la búsqueda semántica no se degrade.
- **Stack de Contenedores Open Source:** Uso de **Docker** sobre la VM para aislar el Agente, el DWH en PostgreSQL y las herramientas de ETL (Apache Hop), facilitando la migración o escalabilidad futura.

6 Definición del alcance de la solución, partes interesadas, cronograma y presupuestos.

6.1 Alcance final de la solución

Este proyecto engloba el ciclo de el diseño, desarrollo, pruebas, despliegue y observabilidad, de un Agente de IA Local (On-Premise) capaz de interactuar con el ecosistema de datos de la AGR.

- ❖ **Límites:** El sistema operará exclusivamente en la red interna (LAN). Se excluyen integración con nubes públicas y servicios externos. El alcance se limita a los 13 dominios de datos definidos de los sistemas SIA.
- ❖ **Objetivos Específicos:**
 - Implementar un motor de búsqueda semántica sobre documentos (PDF, Word, Excel, Audio).
 - Habilitar consultas SQL en lenguaje natural sobre el Datawarehouse.
 - Asegurar un tiempo de respuesta inferior a 5 segundos para consultas complejas.
- ❖ **Resultados Esperados:** Un prototipo funcional integrado a los 4 aplicativos misionales de la AGR, un repositorio de conocimiento vectorial y un manual de gobernanza bajo ISO 42001:2023.

6.2 Identificación de responsables de las actividades para ejecución

La siguiente tabla, muestra la matriz de responsabilidades que tiene cada parte interesada (Stakeholder) en la solución:

Parte Interesada	Rol en el Proyecto	Responsabilidad Principal
Oficina de Tecnología (Sistemas)	Propietario de Infraestructura	Proveer la VM, el almacenamiento y el <i>passthrough</i> (Scale Computing, Consultado en marzo 18 de 2026) de las GPUs Tesla.
Líderes de las áreas involucradas de la AGR	Usuarios Líderes	Validar la precisión de las respuestas y definir los casos de uso críticos.
Comité de Ética y Cumplimiento en IA	Garante Normativo	Supervisar el cumplimiento de la ISO 42001 y privacidad de datos.
Equipo de Desarrollo (3 Full-Stack)	Codificación de conectores, adaptadores, interfaces, etc.	Desarrollo de conectores .NET/Python y arquitectura RAG.

Tabla 2. Presupuesto General Estimado.

Rubro	Descripción	Valor (COP)
Hardware GPU NVIDIA Tesla P40	2x NVIDIA Tesla P40 (24GB), \$2.500.000 c/u	\$5.000.000
Hardware GPU NVIDIA Tesla T4	1x NVIDIA Tesla T4 (16GB), \$3.200.000	\$3.200.000
Infraestructura Almacenamiento NVMe Gen4 2TB	2x NVMe Gen4 2TB \$9.50.000 c/u	\$1.900.000
Kit de Enfriamiento Activo GPUs	3x Kits enfriamiento \$350.000 c/u	\$1.050.000
Cableado y Periféricos	Conectividad 10GB/40GB \$1.200.000	\$1.200.000
Consumo Eléctrico Estimado	Incremento por cómputo de GPUs 24/7. (9 meses, \$250.000 c/u)	\$2.250.000
Total, Hardware y adecuación		\$14.600.000
Talento Humano Lider de Proyecto / Arq. IA	Diseño de RAG, ISO 42001 y Gerencia. (9 meses, \$12.000.000 c/u)	\$108.000.000
Talento Humano Dev. Full-Stack (.NET/Python)	Construcción de conectores SIA y API. (7 meses, \$8.500.000 c/u)	\$59.500.000
Talento Humano Ingeniero de Datos (DBA)	ETL, Réplicas SQL/Postgres y DWH. (5 meses, \$7.000.000 c/u)	\$35.000.000
Total, Talento Humano		\$202.500.000
Seguridad	Auditoria ISO 42001:2023 Costo de certificación \$1.200.000	\$1.200.000
TOTAL.		\$218.300.000

7 Referencias bibliográficas

- AGR, Principios Corporativos. (Consultado en marzo 2026). *Misión y Visión*. Principios Corporativos: <https://www.auditoria.gov.co/web/guest/auditoria/la-auditoria/principios-corporativos>
- IBM. (Consultado en marzo 19 de 2026). *Evalúe un pipeline de RAG usando Ragas en Python con watsonx*. IBM Think: <https://www.ibm.com/es-es/think/tutorials/evaluate-rag-pipeline-using-ragas-in-python-with-watsonx>
- ISO Tools. (Consultado en marzo 2026). *Qué es ISO 42001:2023*. <https://acortar.link/eINqXQ>
- M., B. (Consultado en marzo 17 de 2026). *¿Qué es Multi-Step RAG (Guía completa)?* F22LABS.com: <https://www.f22labs.com/blogs/what-is-multi-step-rag-a-complete-guide/>
- Nvidia Tesla P40. (Consultado en 3 de marzo de 2026). *Nvidia Dispositivos para datacenter*. Nvidia.com: <https://images.nvidia.com/content/pdf/tesla/184427-Tesla-P40-Datasheet-NV-Final-Letter-Web.pdf>
- Nvidia Tesla T4. (Consultado en 3 de marzo de 2026). *Nvidia Dispositivos para datacenter*. Nvidia.com: <https://www.nvidia.com/es-la/data-center/tesla-t4/>
- Scale Computing. (Consultado en marzo 18 de 2026). *GPU virtual frente a GPU Passthrough*. [scalecomputing.com: https://www.scalecomputing.com/resources/virtual-gpu-vs-gpu-passthrough](https://www.scalecomputing.com/resources/virtual-gpu-vs-gpu-passthrough)