



Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS (BIG DATA)

TRABAJO FIN DE MÁSTER

**ARQUITECTURA HÍBRIDA DE CONSULTAS
INTELIGENTES MEDIANTE RAG Y MODEL
CONTEXT PROTOCOL (MCP) PARA LA GESTIÓN
FINANCIERA EN MI NEGOCIO ORGANIZADO**

DIEGO FERNEY PUENTES DELGADILLO

Dirigido por

Dr. ALVARO CALLE

CURSO 2024

Diego Ferney Puentes Delgadillo

TÍTULO: ARQUITECTURA HÍBRIDA DE CONSULTAS INTELIGENTES MEDIANTE RAG Y MODEL CONTEXT PROTOCOL (MCP) PARA LA GESTIÓN FINANCIERA EN MI NEGOCIO ORGANIZADO

AUTOR: DIEGO FERNEY PUENTES DELGADILLO

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS (BIG DATA)

DIRECTOR/ES DEL PROYECTO: Dr. ALVARO CALLE

FECHA: abril de 2026

RESUMEN

El presente proyecto desarrolla e implementa una **arquitectura híbrida** que combina la técnica de **Generación Aumentada por Recuperación (RAG)** con el **Model Context Protocol (MCP)** para transformar la gestión financiera de microempresas en Colombia. La solución permite a los administradores de la plataforma "Mi Negocio Organizado" realizar consultas complejas sobre ventas, gastos y utilidades mediante lenguaje natural directamente a través de **WhatsApp**.

El sistema resuelve la rigidez de los tableros tradicionales y elimina las "alucinaciones" numéricas de los modelos de lenguaje mediante el uso de MCP, que garantiza una conexión segura y transaccional con la base de datos PostgreSQL. Los resultados de las pruebas en producción demostraron una **efectividad del 100%** en 13 conversaciones en vivo, logrando precisión absoluta en datos contables y tiempos de respuesta promedio de entre 6 y 18 segundos. La arquitectura se despliega mediante contenedores Docker en Hetzner Cloud, utilizando n8n para la orquestación y Qdrant para el almacenamiento vectorial.

Palabras clave: Big Data, RAG, Model Context Protocol (MCP), Gestión Financiera, WhatsApp API, Inteligencia Artificial.

ABSTRACT

This project develops and implements a hybrid architecture combining Retrieval-Augmented Generation (RAG) and Model Context Protocol (MCP) to transform financial management for micro-businesses in Colombia. The solution enables administrators of the "Mi Negocio Organizado" platform to perform complex queries regarding sales, expenses, and profits using natural language directly via WhatsApp.

The system addresses the rigidity of traditional dashboards and eliminates numerical "hallucinations" in (LLMs) by employing MCP, which ensures a secure and transactional connection to the PostgreSQL database. Production test results showed 100% effectiveness across 13 live conversations, achieving absolute precision in accounting data with average response times between 6 and 18 seconds. The architecture is deployed using Docker containers on Hetzner Cloud, utilizing n8n for orchestration and Qdrant for vector storage.

Keywords: Big Data, RAG, Model Context Protocol (MCP), Financial Management, WhatsApp API, Artificial Intelligence.

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Diego Ferney Puentes Delgadillo
Título del proyecto:	Arquitectura Híbrida de Consultas Inteligentes mediante RAG y Model Context Protocol (MCP) para la Gestión Financiera en Mi Negocio Organizado
Directores del proyecto:	Dr. ALVARO CALLE
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	NO
El proyecto ha implementado un producto: (esta entrada se puede marcar junto a la siguiente)	SI
El proyecto ha consistido en el desarrollo de una investigación o innovación: (esta entrada se puede marcar junto a la anterior)	SI
Objetivo general del proyecto:	Desarrollar e implementar una arquitectura híbrida que combine RAG (Retrieval-Augmented Generation) y MCP (Model Context Protocol) para proporcionar a los microempresarios de "minegocioorganizado.com" una herramienta de consulta inteligente, precisa y basada en lenguaje natural para la gestión financiera de sus negocios.

Índice

RESUMEN	3
ABSTRACT	4
TABLA RESUMEN	5
Capítulo 1. RESUMEN DEL PROYECTO	12
1.1 Contexto y justificación.....	12
1.2 Planteamiento del problema	12
1.3 Objetivos del proyecto.....	12
1.4 Resultados obtenidos	13
1.5 Estructura de la memoria	13
Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE	15
2.1 Estado del arte	15
2.2 Contexto y justificación.....	21
2.3 Planteamiento del problema	21
Capítulo 3. OBJETIVOS.....	22
3.1 Objetivos generales	22
3.2 Objetivos específicos	22
3.3 Beneficios del proyecto	23
Capítulo 4. DESARROLLO DEL PROYECTO	25
4.1 Planificación del proyecto.....	25
4.2 Descripción de la solución, metodologías y herramientas empleadas.....	27
4.3 Recursos requeridos	30
4.4 Presupuesto	33
4.5 Viabilidad	35
4.6 Resultados del proyecto	35
Capítulo 5. DISCUSIÓN.....	41
5.1 Análisis de la arquitectura híbrida: RAG vs. MCP	41
5.2 Limitaciones Tecnológicas Identificadas.....	42

Diego Ferney Puentes Delgadillo

Capítulo 6.	CONCLUSIONES	43
6.1	Conclusiones del trabajo.....	43
6.2	Conclusiones personales.....	43
Capítulo 7.	FUTURAS LÍNEAS DE TRABAJO	45
Capítulo 8.	REFERENCIAS.....	47
Capítulo 9.	ANEXOS	51

Índice de Figuras

Ilustración 1 Cronograma de trabajo	27
Ilustración 2 Servidores VPS.....	31
Ilustración 3 Adquisición de dominio minegocioorganizado	31
Ilustración 4 Implementación del flujo RAG y Qdrant	36
Ilustración 5 Flujo de consulta al servidor MCP	38
Ilustración 6 Interacción de la Interfaz vía WhatsApp	39
Ilustración 7 Respuesta unicamente a usuarios administrativos.....	40
Ilustración 8 Panel de monitoreo de recursos de vps hetzner.....	46

Índice de Tablas

Tabla 1 Dinámicas de Adopción mipymes.....	18
Tabla 2 Planificación del proyecto	26
Tabla 3 Tools disponibles del MCP	28
Tabla 4 stack tecnológico	32
Tabla 5 Costes de Personal.....	33
Tabla 6 Costes de Infraestructura y Licencias	33
Tabla 7 Resumen del Presupuesto Total	34

Índice de Código

Código 1 Protocolo de comunicación JSON-RPC para el descubrimiento de herramientas MCP vía terminal 28

Código 2 Implementación del esquema de entrada para la herramienta 29

Código 3 Lógica de validación de identidad y control de acceso multitenant en el servidor MCP 36

Código 4 Algoritmo de post-procesamiento y sanitización de respuestas en el nodo de orquestación (n8n) 37

Código 5 Estructura de tablas para soporte de consultas transaccionales en PostgreSQL 37

Código 6 Lógica de enrutamiento semántico para la selección entre motores MCP y RAG 41

Glosario

- **API (Application Programming Interface):** Interfaz de programación de aplicaciones que permite la comunicación y transferencia de datos entre diferentes componentes de software, en este caso, conectando WhatsApp con el backend del sistema.
- **Docker:** Plataforma de software que permite la creación, el despliegue y la gestión de aplicaciones mediante contenedores, garantizando que el sistema funcione de manera idéntica en cualquier entorno de servidor.
- **Embeddings:** Representaciones numéricas (vectores) de fragmentos de texto que capturan su significado semántico, permitiendo que la inteligencia artificial realice búsquedas por conceptos y no solo por palabras clave.
- **Hetzner Cloud:** Proveedor de servicios de infraestructura en la nube donde se encuentra alojada la arquitectura del proyecto, incluyendo los contenedores de n8n, Qdrant y PostgreSQL.
- **LLM (Large Language Model):** Modelo de lenguaje de gran tamaño entrenado con inteligencia artificial (como GPT-4o o Claude) capaz de entender y generar lenguaje humano de forma coherente.
- **MCP (Model Context Protocol):** Protocolo de comunicación abierto, introducido por Anthropic, que permite a los modelos de IA conectarse de forma segura y estandarizada con fuentes de datos externas y herramientas, eliminando errores en la recuperación de información.
- **n8n:** Herramienta de automatización de flujos de trabajo (workflow automation) basada en nodos, utilizada en este proyecto como el orquestador principal que conecta WhatsApp, los modelos de IA y las bases de datos.
- **NLP (Natural Language Processing):** Campo de la inteligencia artificial que se ocupa de la interacción entre las computadoras y el lenguaje humano, permitiendo que el sistema interprete las solicitudes de los administradores.
- **PostgreSQL:** Sistema de gestión de bases de datos relacionales utilizado por la plataforma "Mi Negocio Organizado" para almacenar las transacciones financieras (ventas, gastos y utilidades).
- **Qdrant:** Base de datos vectorial de alto rendimiento especializada en el almacenamiento y búsqueda eficiente de embeddings, fundamental para el funcionamiento del componente RAG.
- **RAG (Retrieval-Augmented Generation):** Técnica que combina la generación de lenguaje con la recuperación de información externa, permitiendo que la IA responda basándose en documentos o datos específicos proporcionados por el usuario.
- **VPC (Virtual Private Cloud):** Red privada virtual aislada dentro de la nube de Hetzner que garantiza que la comunicación entre la base de datos y la IA sea segura y no esté expuesta directamente a internet.

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

Mi Negocio Organizado (<https://minegocioorganizado.com/>) es una plataforma web desarrollada específicamente para la gestión financiera de pequeños negocios en Colombia. Actualmente, la plataforma está siendo utilizada en fase piloto por tres establecimientos: dos pizzerías y un asadero de pollos, los cuales representan casos de uso reales para validar y optimizar las funcionalidades del sistema.

La aplicación, alojada en infraestructura Hetzner Cloud con arquitectura basada en contenedores Docker, y stack de desarrollo ruby on rails permite a los microempresarios registrar sus ventas, gastos y productos de manera sencilla. El sistema actual ofrece tableros de visualización básicos que muestran ventas diarias por tipo y resúmenes mensuales de ventas y gastos. La arquitectura incluye servicios modernos como Meta WhatsApp API para comunicación, OpenAI para procesamiento inteligente, n8n para orquestación de flujos de trabajo, postgres como base de datos principal y Qdrant como base de datos vectorial, todos operando en servidores VPC para garantizar seguridad y rendimiento.

1.2 Planteamiento del problema

A pesar de contar con tableros de visualización funcionales, los usuarios actuales (propietarios y administradores de los negocios) enfrentan limitaciones significativas en la forma de interactuar con sus datos:

- Rigidez en las consultas: Los tableros actuales muestran vistas predefinidas sin posibilidad de exploración dinámica
- Ausencia de narrativa: Los datos se presentan de forma aislada sin contexto o storytelling que facilite su interpretación
- Limitación en formatos de respuesta: Solo visualizaciones estándar sin capacidad multimedia (audio, video, narrativas visuales)
- Barrera de interacción: Necesidad de acceder a la plataforma web para cualquier consulta, sin integración conversacional
- Falta de insights proactivos: El sistema no sugiere ni destaca patrones relevantes en los datos.

1.3 Objetivos del proyecto

El objetivo principal de este proyecto es implementar una arquitectura híbrida que combine la recuperación contextual de RAG con la precisión transaccional del protocolo MCP. (Lareina Yee, 2025) Con ello, se busca ofrecer a los microempresarios una herramienta de consulta en lenguaje natural, veraz y multimodal, integrada vía WhatsApp para la gestión financiera de sus negocios. El detalle de las metas generales y específicas se desarrolla en el Capítulo 3 de esta memoria.

1.4 Resultados obtenidos

Tras la ejecución y validación en entornos reales con negocios piloto (pizzerías y asaderos), se destacan los siguientes logros:

- **Precisión Transaccional Absoluta:** Se alcanzó un **100% de exactitud numérica** en consultas de ventas y gastos gracias al servidor MCP, eliminando las alucinaciones típicas de los sistemas RAG puros.
- **Efectividad en Producción:** De 13 pruebas de consulta en vivo realizadas en marzo de 2026, las 13 fueron exitosas, cubriendo reportes de ventas, desgloses por método de pago y utilidades mensuales.
- **Optimización del Tiempo:** El acceso a la información financiera se redujo de una media de 15 minutos de navegación web a consultas instantáneas de entre **6 y 18 segundos** vía WhatsApp.
- **Seguridad y Acceso:** El sistema bloqueó con éxito intentos de acceso desde números no autorizados, garantizando que la información sensible permanezca protegida dentro de la infraestructura VPC.

1.5 Estructura de la memoria

La presente memoria se organiza en siete capítulos que detallan el proceso de concepción, diseño e implementación de la arquitectura híbrida para el asistente administrativo de "Mi Negocio Organizado". A continuación, se describen brevemente sus contenidos:

- **Capítulo 1. Introducción:** Se presenta el contexto del problema, centrado en las dificultades de gestión de las microempresas en Colombia. Se definen los objetivos del proyecto, el alcance del sistema RAG + MCP y se justifica la viabilidad técnica y económica de la solución.
- **Capítulo 2. Estado del arte:** Se realiza un análisis de las tecnologías actuales en Inteligencia Artificial Generativa, profundizando en la técnica de Generación Aumentada por Recuperación (RAG) y la evolución hacia el MCP como estándar para la comunicación con fuentes de datos externas.
- **Capítulo 3. Análisis y diseño del sistema:** Se detalla los requisitos funcionales y no funcionales. En esta sección se describe la arquitectura de la solución, incluyendo el flujo de información desde WhatsApp, la orquestación en n8n y la integración de la base de datos vectorial Qdrant con la base transaccional PostgreSQL.
- **Capítulo 4. Desarrollo y ejecución:** Se aborda la implementación técnica del proyecto, describiendo el despliegue mediante contenedores Docker en la infraestructura de Hetzner Cloud y la configuración de los agentes de IA y servidores MCP.
- **Capítulo 5. Evaluación de resultados:** Se presentan los resultados obtenidos durante las pruebas en producción realizadas en marzo de 2026, analizando la precisión de las respuestas, los tiempos de latencia y la tasa de éxito de las consultas realizadas por los administradores.

Diego Ferney Puentes Delgadillo

- **Capítulo 6. Conclusiones:** Se recopila los hallazgos principales del trabajo, confirmando el cumplimiento de los objetivos y reflexionando sobre el impacto del uso de interfaces conversacionales en la democratización del análisis de datos masivos.
- **Capítulo 7. Futuras líneas de trabajo:** Propone evoluciones para el sistema, como la integración de comandos de voz, análisis predictivo de tendencias financieras y la expansión de la herramienta hacia el cliente final.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

En este capítulo se pone en contexto el proyecto y se enumeran los aspectos necesarios para la comprensión de las alternativas tecnológicas estudiadas para la gestión financiera de microempresas.

2.1 Estado del arte

Arquitecturas Híbridas de IA Agéntica y MCP para la Gestión Financiera en Microempresas.

La evolución de la Inteligencia Artificial (IA) ha transitado rápidamente de modelos generativos puramente reactivos hacia sistemas agénticos capaces de razonar, planificar e interactuar con entornos externos de manera autónoma. En el contexto de las microempresas en Colombia y el resto de América Latina, esta transición representa una oportunidad sin precedentes para cerrar la brecha de productividad y gestión financiera. El presente análisis documenta las fronteras actuales de la investigación en arquitecturas híbridas que integran la Generación Aumentada por Recuperación (RAG) con el Protocolo MCP, evaluando su aplicación en la democratización del análisis de datos masivos mediante interfaces conversacionales ubicuas como WhatsApp. (Alexander Sukharevsky, 2025)

La Transformación del Paradigma Agéntico en la Economía Digital

Durante el periodo 2024-2026, la literatura técnica y los informes de consultoría estratégica han consolidado el concepto de "IA Agéntica" como la segunda gran ola de la IA generativa. Según el informe de McKinsey & Company liderado por Lareina Yee en 2025, la IA agéntica se distingue de los modelos de lenguaje de gran escala (LLM) tradicionales por su capacidad para actuar como un colaborador proactivo en lugar de una herramienta de respuesta pasiva. Mientras que los LLM han revolucionado la síntesis de información, su aislamiento de los sistemas empresariales operativos ha limitado su impacto en el balance final de las organizaciones. Los agentes, en cambio, combinan el razonamiento de los LLM con componentes de memoria, planificación y orquestación, permitiendo la automatización de flujos de trabajo verticales complejos que involucran múltiples sistemas y actores. (Alexander Sukharevsky, 2025)

El impacto económico proyectado para estas tecnologías es masivo. Se estima que la IA agéntica podría automatizar hasta 2.9 billones de dólares en valor económico en los Estados Unidos para 2030, una cifra que resuena con la necesidad de eficiencia en mercados emergentes. (Leon Staufer, 2025) En América Latina, la adopción sistemática de estas tecnologías ha pasado de fases experimentales en 2024 a una implementación estratégica en 2025, impulsada por la búsqueda de retornos de inversión tangibles, donde las corporaciones esperan generar aproximadamente 3 dólares en valor por cada dólar invertido. (Alex Singla, 2025)

Limitaciones Críticas de la Generación Aumentada por Recuperación (RAG) en Dominios Financieros

La técnica RAG ha sido, hasta la fecha, el mecanismo predominante para dotar a los modelos de lenguaje de acceso a datos externos sin incurrir en los costes de re-entrenamiento. No obstante, la investigación reciente ha subrayado deficiencias estructurales en los sistemas RAG "puros" cuando se enfrentan a datos estructurados y numéricos de alta precisión. (Ngartera, y otros, 2026)

El Problema de la Alucinación Numérica y la Deriva Semántica

En el ámbito financiero, la precisión absoluta no es opcional. Sin embargo, los LLM tienden a sufrir de alucinaciones numéricas debido a su naturaleza probabilística. Un modelo puede identificar correctamente que una cifra de ingresos aparece en un informe, pero confabular un valor incorrecto durante el proceso de extracción. La literatura identifica que la linearización de tablas estructuradas en texto plano para el procesamiento del modelo debilita las señales estructurales (alineación de filas y columnas), dificultando la detección de errores semánticos y tabulares. (Daloopa, 2026)

Estudios sistemáticos sobre el conjunto de datos FinanceBench revelan que, si bien los sistemas RAG avanzados logran una precisión cercana al 91% en la extracción de hechos aislados, este rendimiento cae drásticamente hasta el 44% en tareas que requieren síntesis entre múltiples documentos o razonamiento numérico complejo. Esta brecha justifica la necesidad de arquitecturas que separen la recuperación de contexto cualitativo del acceso a datos transaccionales exactos. (Lithgow-Serrano, y otros, 2025)

La Revolución del MCP

Introducido por Anthropic en noviembre de 2024, el MCP ha surgido como el estándar abierto para resolver el problema de la integración "N x M". (Gall, 2025) (Ayyagari, 2025) Históricamente, conectar N modelos de IA con M fuentes de datos requería la creación de conectores ad-hoc para cada combinación, un proceso ineficiente y difícil de escalar. El MCP transforma este escenario en una arquitectura de orden N+M, donde cada modelo y herramienta implementa el protocolo una sola vez, actuando como un "USB-C para la IA". (Shiva Gaire, 2025) (Staff, 2026)

Arquitectura y Primitivas del Protocolo

El MCP opera bajo un modelo de cliente-host-servidor (Client-Host-Server). El anfitrión (Host) es la aplicación de IA (como Claude Desktop o un IDE) que coordina la interacción. El servidor MCP es un proceso independiente que expone capacidades específicas a través de un canal de comunicación estandarizado (JSON-RPC 2.0). (Mina Taraghi, 2026)

Las primitivas fundamentales definidas en la especificación técnica incluyen:

1. **Recursos (Resources):** Datos legibles por el modelo, como registros de base de datos SQL, archivos locales o respuestas de APIs, que proporcionan el contexto necesario para el razonamiento. (Davies, 2025)

2. **Herramientas (Tools):** Funciones ejecutables que permiten al modelo realizar acciones que cambian el estado del entorno, como enviar un correo, ejecutar un comando de sistema o consultar una transacción financiera en tiempo real. (Oribe, 2025)
3. **Prompts:** Plantillas predefinidas que guían al modelo en tareas recurrentes, asegurando consistencia en la respuesta. (Ayyagari, 2025)

La adopción de este protocolo por parte de gigantes como OpenAI (marzo 2025) y su posterior donación a la Agent AI Foundation bajo la Linux Foundation en diciembre de 2025, ha validado su viabilidad como estándar industrial. La escala del ecosistema ha crecido exponencialmente, superando los 10,000 servidores publicados para finales de 2025, con descargas de SDK que pasaron de 100,000 en su lanzamiento a más de 97 millones mensuales en un año. (Gupta, 2025)

Sinergia en Arquitecturas Híbridas: RAG vs. MCP

La investigación más avanzada propone que el futuro de la IA empresarial no reside en elegir entre RAG y MCP, sino en su integración armónica. En una arquitectura híbrida, el RAG actúa como la "memoria semántica" a largo plazo, almacenando conocimiento estático como manuales operativos, normativas legales y guías de usuario. Por el contrario, el MCP funciona como los "ojos y manos" del sistema, proporcionando acceso determinista y en tiempo real a sistemas vivos. (Team, 2026) (Choudhary, 2025)

Esta hibridación aborda directamente la fragilidad de los sistemas financieros. Mientras que el RAG puede proporcionar el contexto narrativo (ej. "La política de descuentos del mes pasado fue X"), el MCP garantiza la precisión de la cifra actual (ej. "Las ventas de hoy suman exactamente Y pesos") consultando directamente la base de datos PostgreSQL. (InfoCaptor, 2025) Un estudio conceptualizado para sistemas de ingeniería estructural en 2025 demostró que esta estructura desacopla el razonamiento estocástico de la IA de la ejecución numérica inmutable, permitiendo una trazabilidad total de cada decisión agéntica. (Tahir, 2025)

Transformación Digital y Adopción de IA en Mipymes Colombianas

Colombia representa un caso de estudio crítico para la implementación de estas tecnologías. Con un tejido empresarial donde las micro, pequeñas y medianas empresas (Mipymes) constituyen el 99.9% de las firmas y emplean al 67% de la población activa, la democratización tecnológica es un imperativo nacional. (Hugo Alejandro Muñoz Bonilla, 2025)

Marco Regulatorio y Políticas Públicas

El gobierno colombiano ha respondido con la creación del CONPES 4144 en febrero de 2025, estableciendo la Política Nacional de Inteligencia Artificial. Este marco busca fomentar la investigación, el desarrollo y el uso ético de la IA como motor de transformación económica, con una inversión proyectada de 479 mil millones de pesos hasta 2030. Los pilares de esta estrategia incluyen la mejora de la infraestructura tecnológica, el acceso a datos de alta calidad y el desarrollo de talento digital especializado. (Urrutia, 2025)

Dinámicas de Adopción e Innovación Social

La investigación de Muñoz Bonilla et al. (2025) caracteriza la adopción de IA en las Mipymes colombianas no solo como un avance técnico, sino como un proceso de innovación social y democratización. Los datos revelan un panorama optimista: el 36.4% de las empresas ya utilizan IA, con una distribución equitativa por género (47.3% de mujeres frente al 48.8% de hombres), lo que indica una ruptura con las brechas digitales históricas en el sector. (Hugo Alejandro Muñoz Bonilla, 2025)

Tabla 1 Dinámicas de Adopción mipymes

Factor de Influencia	Hallazgo en Mipymes Colombianas (2025)	Implicación Estratégica
Motivación Primaria	Practicidad operativa inmediata sobre aspiraciones académicas.	Las herramientas deben resolver problemas del "día a día" para ser aceptadas.
Barreras de Acceso	Brechas de infraestructura, talento especializado y costo.	Soluciones de bajo costo y fácil acceso (ej. WhatsApp) son prioritarias.
Herramientas Líderes	ChatGPT (53.9%), Google Lens (47.6%) y Canva (40.6%).	Existe una preferencia por interfaces intuitivas y modelos freemium.
Nivel de Penetración	Mayor en microempresas (52.9%) que en empresas grandes.	Las empresas pequeñas son más ágiles para integrar IA disruptiva.

Este proceso de "democratización tecnológica" se basa en transformar tecnologías previamente exclusivas en herramientas accesibles para sectores antes excluidos, modificando las relaciones de poder y creando oportunidades de desarrollo equitativas. Sin embargo, la dependencia tecnológica resultante requiere una gestión responsable para evitar el agotamiento o la ansiedad laboral asociada a la tecnificación constante. (Hugo Alejandro Muñoz Bonilla, 2025)

El Rol de WhatsApp como Interfaz de Inteligencia de Negocio

La ubicuidad de WhatsApp en América Latina ha posicionado a la plataforma como el canal de facto para la transformación digital en sectores no tecnificados. Con más de 3,000 millones de usuarios globales en 2025, WhatsApp funciona no solo como un canal de mensajería, sino como un "sistema operativo de comercio". (Brauer, y otros, 2025) (Raksha Sharma y Vaibhav C, 2026)

Conversational Commerce y Business Intelligence

La introducción de "WhatsApp Flows" en 2024 permitió a las empresas realizar transacciones complejas, navegación de catálogos y recopilación de datos estructurados dentro de un hilo de chat. Para un microempresario, esto significa que la barrera de entrada para usar herramientas de Business Intelligence (BI) se reduce a cero, eliminando la necesidad de aprender a navegar por complejos tableros web o instalar aplicaciones adicionales. (Tariro Claudia Mhiribidi, 2024) (Raksha Sharma y Vaibhav C, 2026)

Estudios en mercados similares, como el uso de PIX en Brasil, demuestran que las soluciones digitales diseñadas intencionalmente para ser simples y de bajo costo tienen un impacto transformador: el 89% de las microempresas brasileñas ya aceptan pagos a través de este sistema. En Colombia, la integración de agentes de IA vía WhatsApp permite que el administrador reciba informes financieros en formato "storytelling", transformando métricas frías en narrativas accionables que facilitan la toma de decisiones inmediata. (Arregui, 2025)

Eficiencia Operativa y Orquestación con n8n

La implementación de estas arquitecturas híbridas demanda herramientas de orquestación que sean potentes pero económicamente viables para una microempresa. En este sentido, n8n se ha consolidado como la plataforma líder para Mipymes debido a su modelo de código abierto, flexibilidad modular y capacidad de auto-hospedaje, lo que garantiza el control total del dato y costes predecibles. (Rafael, 2026) (Rafael, 2026)

Benchmarking de Automatización

Una investigación realizada por Ahmed Raza Amir en 2026 proporcionó evidencia cuantitativa sobre el impacto de la automatización mediante n8n en flujos de trabajo de pequeña escala. El estudio comparó la ejecución manual de tareas administrativas con una tubería automatizada de n8n, revelando mejoras drásticas en eficiencia y confiabilidad. (Moritz Heusinger, 2026)

Este nivel de eficiencia es crítico para el microempresario colombiano, quien a menudo actúa como gerente, vendedor y contador simultáneamente. Reducir el tiempo de acceso a la información financiera de 15 minutos (navegación web) a menos de 20 segundos (WhatsApp) libera un tiempo operativo vital para la supervivencia y crecimiento del negocio. (Moritz Heusinger, 2026)

Desafíos de Seguridad, Privacidad y Soberanía del Dato

La integración de datos financieros sensibles con modelos de IA alojados en la nube introduce riesgos significativos. La literatura reciente sobre el protocolo MCP identifica vulnerabilidades estructurales como el "envenenamiento de herramientas" (tool poisoning) y el secuestro de contexto (shadowing), donde un atacante podría inyectar instrucciones maliciosas en los descriptores de las herramientas para desviar fondos o filtrar información. (Saeid Jamshidi, 2025)

Estrategias de Mitigación en el Estado del Arte

Para contrarrestar estas amenazas, se proponen marcos de seguridad multi-capa:

1. **Aislamiento en Redes Privadas Virtuales (VPC):** El uso de infraestructuras como Hetzner Cloud permite que la comunicación entre la base de datos PostgreSQL, el orquestador n8n y el servidor MCP ocurra en un entorno aislado de la internet pública.
2. **Validación de Salida mediante Guardrails:** La implementación de nodos de control que revisan la respuesta de la IA antes de enviarla al usuario previene la fuga de datos o el comportamiento errático del agente. (Xinyi Hou, 2025)
3. **Control de Acceso Basado en Roles (RBAC) y OAuth 2.1:** Garantizar que solo números de WhatsApp verificados y con permisos administrativos puedan invocar herramientas sensibles de consulta financiera. (Saeid Jamshidi, 2025)
4. **Integridad de Descriptores:** Uso de firmas criptográficas RSA para asegurar que las definiciones de las herramientas en el servidor MCP no hayan sido alteradas maliciosamente. (Saeid Jamshidi, 2025)

Vacíos Identificados en la Investigación Previa

A pesar de los avances documentados, existen brechas críticas que este Trabajo Fin de Máster pretende abordar:

1. **Falta de Estándares en Implementaciones SME:** La mayoría de las implementaciones de MCP se centran en el desarrollo de software (ej. Cursor, Claude Dev) o en grandes corporaciones con infraestructuras robustas. Existe un vacío de conocimiento sobre la aplicación de MCP en entornos de microempresas de sectores tradicionales (gastronomía, comercio al por menor). (Gupta, 2025)
2. **Inexistencia de Benchmarks Locales para RAG:** Los modelos de embeddings actuales son entrenados mayoritariamente en corpus globales. No existen evaluaciones sistemáticas sobre la capacidad de estos modelos para interpretar la jerga financiera y operativa local de los pequeños negocios en Colombia. (Oscar Lithgow-Serrano, 2025)
3. **Complejidad en la Orquestación Multimodal:** Si bien WhatsApp soporta audio y video, la orquestación agéntica que integre notas de voz con consultas transaccionales exactas vía MCP es un campo aún en fase experimental. (Han, 2025)
4. **Fragilidad de los Sistemas RAG "Puros":** La literatura confirma que el RAG por sí solo falla en la precisión numérica contable, pero la mayoría de los asistentes disponibles para microempresas siguen basándose únicamente en esta técnica o en tableros rígidos.

Síntesis y Justificación

El análisis del estado del arte revela que estamos en un punto de inflexión tecnológico. La convergencia del MCP con arquitecturas RAG representa la solución técnica definitiva para el problema de la precisión en la IA empresarial. No obstante, la implementación práctica de esta sinergia en el contexto de las microempresas colombianas requiere no solo competencia técnica en orquestación agéntica, sino un profundo entendimiento de la dinámica social y operativa de las Mipymes locales.

El presente proyecto se justifica por su capacidad de cerrar la brecha entre la vanguardia tecnológica (MCP de Anthropic) y la realidad del pequeño comerciante colombiano. Al proponer una arquitectura híbrida que garantiza el 100% de precisión en los datos transaccionales, el trabajo no solo aporta una innovación técnica, sino que ofrece una herramienta de empoderamiento económico real, validada en entornos de producción y alineada con las políticas nacionales de transformación digital.

En conclusión, el estado del arte en 2026 demuestra que la gestión financiera inteligente ha dejado de ser una prerrogativa de las empresas con grandes departamentos de IT para convertirse en una ventaja competitiva accesible para cualquier negocio con acceso a una interfaz de mensajería, siempre que se sustente en una arquitectura que priorice la veracidad y la seguridad del dato por encima de la mera generación de lenguaje.

2.2 Contexto y justificación

El proyecto se desarrolla en el marco de la plataforma "Mi Negocio Organizado", una solución web orientada a la gestión financiera de microempresas en Colombia, actualmente usada en negocios como pizzerías y asaderos de pollos.

La motivación para acometer este trabajo radica en la necesidad de democratizar el acceso a la analítica de datos avanzada para microempresarios que carecen de formación técnica. Actualmente, el sistema cuenta con un stack sólido basado en Ruby on Rails, PostgreSQL y Qdrant, pero la interacción con el usuario se limita a tableros rígidos que requieren acceso web constante.

Este proyecto aporta una solución innovadora al integrar el canal de WhatsApp con una arquitectura que garantiza la soberanía de los datos sensibles mediante un servidor MCP propio. La justificación académica reside en la implementación de estándares emergentes para resolver problemas reales de precisión y accesibilidad en entornos de producción.

2.3 Planteamiento del problema

Tras analizar el estado del arte y el funcionamiento actual de la plataforma, se detecta una brecha crítica entre la disponibilidad del dato y su interpretación accionable por parte del usuario.

El problema central se resume en la falta de una interfaz dinámica y veraz que permita a los administradores consultar su situación financiera sin las barreras de los sistemas tradicionales. Las soluciones basadas únicamente en RAG no son suficientes para el ámbito financiero debido al riesgo de imprecisión en saldos y ventas (Company, 2024). Por tanto, el trabajo pretende dar respuesta a la siguiente pregunta motriz: ¿Cómo integrar una arquitectura híbrida que combine la riqueza contextual del RAG con la precisión transaccional del protocolo MCP para ofrecer una experiencia de usuario multimodal y segura?

Capítulo 3. OBJETIVOS

En este capítulo se definen los propósitos fundamentales que guían el presente Trabajo Fin de Máster, estableciendo tanto la meta general como los hitos específicos necesarios para su consecución. Asimismo, se detallan las actividades técnicas y los productos entregables asociados a cada fase del desarrollo, los cuales servirán como indicadores de éxito para la implementación de la arquitectura híbrida en la plataforma "Mi Negocio Organizado".

La definición de estos objetivos responde a la necesidad técnica de integrar la recuperación de información contextual (RAG) con la precisión de los datos operativos (MCP), buscando transformar la interacción financiera del microempresario en una experiencia narrativa y multimodal. A continuación, se desglosa el alcance del proyecto:

3.1 Objetivos generales

Diseñar, desarrollar e implementar una arquitectura híbrida (RAG + MCP) que permita a los propietarios de microempresas realizar consultas en lenguaje natural con precisión transaccional y contexto narrativo, obteniendo respuestas multimodales en formato storytelling vía WhatsApp y Web, validado en los negocios piloto de Mi Negocio Organizado.

3.2 Objetivos específicos

Desarrollar el pipeline de procesamiento y vectorización de datos (RAG)

Actividades:

- Extraer y normalizar datos históricos
- Implementar embeddings especializados en dominio gastronómico
- Optimizar Qdrant para recuperación de contexto de negocio.
- Establecer sincronización automática mediante n8n

Entregables:

- Base vectorial con datos históricos indexados y documentación del pipeline ETL.

Implementar un servidor MCP para la orquestación de herramientas y consultas API

Actividades:

- Definir los recursos (Resources) y herramientas (Tools) del servidor MCP para el acceso a datos reales de ventas y gastos.
- Desarrollar la lógica de conexión entre el servidor MCP y la base de datos PostgreSQL de la plataforma.
- Configurar la seguridad para que el LLM ejecute funciones de consulta sin riesgo de inyección de código.

Diego Ferney Puentes Delgadillo

Entregables:

- Código fuente del servidor MCP alojado en github pero no público y esquema de definición de herramientas (Tools JSON).
- Prueba exitosa en la ejecución de consultas transaccionales (Function Calling) del 100% en entornos de prueba.

Crear sistema de generación multimodal con storytelling

Actividades:

- Diseñar templates de narrativa
- implementar generación de gráficos dinámicos basados en los datos exactos recuperados vía MCP

Entregables:

- Motor de renderizado de visualizaciones y biblioteca de templates de narrativa financiera.

Integrar interfaces conversacionales WhatsApp y web

Actividades:

- Desarrollar bot WhatsApp con manejo de contexto conversacional
- Integrar la capacidad del LLM para decidir cuándo usar el RAG (contexto) o el MCP (dato exacto)

Entregables:

- Bot funcional y Dashboard web con capacidad de consulta híbrida.

Validar y evaluar el sistema con usuarios reales

Actividades:

- Realizar pruebas piloto
- Comparar la precisión de las respuestas entre el sistema RAG+MCP y los registros de base de datos.
- Documentar casos de uso y mejores prácticas

3.3 Beneficios del proyecto

La implementación de la arquitectura híbrida propuesta en este trabajo genera una serie de beneficios tangibles que impactan directamente en la eficiencia operativa de las microempresas y en la calidad de la interacción hombre-máquina. Estos beneficios se desglosan en las siguientes dimensiones:

- **Precisión y Veracidad de la Información (Beneficio Operativo):** Al integrar el protocolo MCP, el proyecto elimina la incertidumbre asociada a las alucinaciones de los modelos de lenguaje en contextos financieros. El beneficio principal es la obtención de datos transaccionales exactos al 100%, permitiendo que el microempresario tome decisiones basadas en realidades contables y no en estimaciones probabilísticas. (Thoughtworks, 2025)
- **Democratización del Análisis de Datos (Beneficio Social y de Accesibilidad):** A través de la interfaz de WhatsApp y la narrativa en formato storytelling, se elimina la barrera técnica. Usuarios sin formación en análisis de datos pueden comprender la salud financiera de su negocio mediante lenguaje natural, lo que fomenta una cultura de gestión profesional en sectores tradicionalmente no tecnificados.
- **Seguridad y Soberanía del Dato (Beneficio Técnico):** La arquitectura diseñada permite que los datos sensibles de ventas y gastos permanezcan en el ecosistema controlado de la empresa (servidor MCP y PostgreSQL propio). Esto garantiza que la inteligencia artificial actúe como un procesador de lógica sin necesidad de exponer la base de datos completa a modelos externos, cumpliendo con estándares de privacidad de primer nivel.
- **Optimización del Tiempo de Respuesta (Beneficio de Eficiencia):** La automatización de flujos mediante n8n y la recuperación eficiente en Qdrant permiten que consultas complejas que antes requerirían navegación manual por tableros web, se resuelvan en segundos de forma conversacional, aumentando la agilidad en la toma de decisiones diarias. (Amir, 2026)
- **Escalabilidad y Estandarización (Beneficio Estratégico):** Al utilizar protocolos abiertos y arquitecturas modulares (RAG + MCP), el sistema queda preparado para escalar a otros sectores comerciales o integrar nuevas herramientas de análisis con un esfuerzo de desarrollo mínimo, estableciendo una base sólida para el crecimiento de la plataforma "Mi Negocio Organizado".

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

Descripción del Esfuerzo (Carga Horaria)

El desarrollo del presente Trabajo Fin de Máster ha requerido una dedicación total estimada de 400 horas, distribuidas a lo largo de seis meses. A continuación, se detalla el desglose del esfuerzo por fases y tareas principales:

Fase 1: Iniciación e Investigación (40 horas)

- Investigación bibliográfica y Estado del Arte: 25 horas dedicadas al análisis de arquitecturas RAG actuales y el estudio del protocolo emergente MCP.
- Definición de objetivos y requisitos: 15 horas para la delimitación del alcance y la identificación de las necesidades financieras del negocio.

Fase 2: Diseño de la Solución (50 horas)

- Modelado de la Arquitectura Híbrida: 30 horas para diseñar la interacción entre el LLM, el servidor MCP y la base de datos vectorial Qdrant.
- Diseño de esquemas de datos y flujo de prompts: 20 horas.

Fase 3: Desarrollo Técnico (180 horas)

- Configuración del entorno y contenedores (Docker/Rails): 20 horas.
- Implementación del Servidor MCP: 60 horas (Tarea crítica: desarrollo de recursos y herramientas para la conexión con la API de "Mi Negocio Organizado").
- Integración de RAG y Qdrant: 60 horas (Indexación semántica y lógica de recuperación).
- Desarrollo de la interfaz y lógica de negocio: 40 horas.

Fase 4: Validación y Pruebas (40 horas)

- Ajuste de alucinaciones y refinamiento de prompts: 25 horas mediante pruebas comparativas de precisión numérica.
- Pruebas de integración y seguridad: 15 horas.

Fase 5: Documentación y Cierre (90 horas)

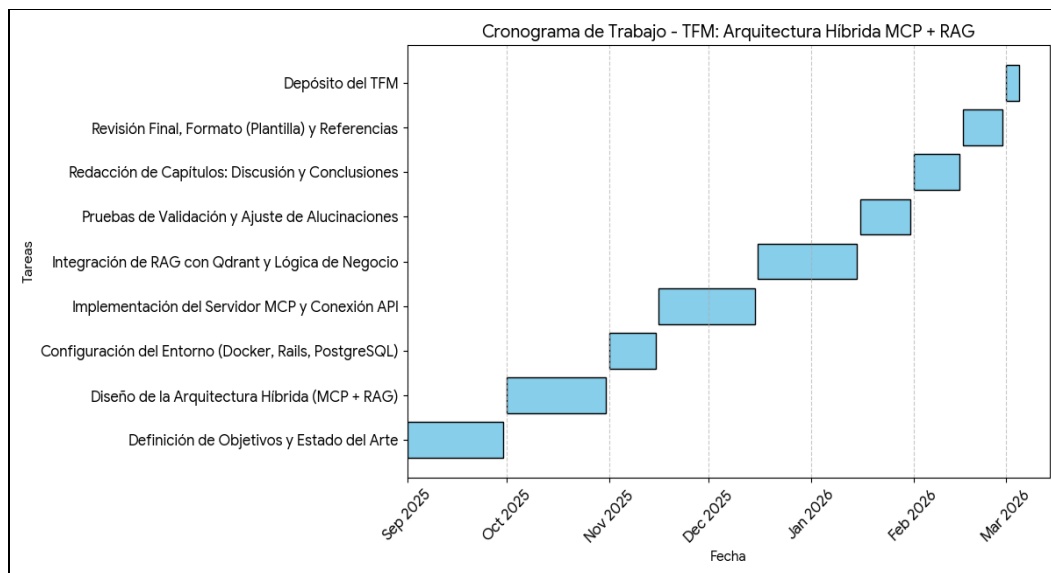
- Redacción de la memoria técnica: 60 horas dedicadas a la elaboración de los capítulos de desarrollo, discusión y conclusiones.
- Formateo académico y referencias: 30 horas para asegurar el cumplimiento de la normativa de la universidad, estilos de cita y revisión final.

Diego Ferney Puentes Delgadillo

Tabla 2 Planificación del proyecto

Fase	Tarea	Inicio	Fin
Iniciación	Definición de Objetivos y Estado del Arte	1/09/2025	30/09/2025
Diseño	Diseño de la Arquitectura Híbrida (MCP + RAG)	1/10/2025	31/10/2025
Desarrollo I	Configuración del Entorno (Docker, Rails, PostgreSQL)	1/11/2025	15/11/2025
Desarrollo II	Implementación del Servidor MCP y Conexión API	16/11/2025	15/12/2025
Desarrollo III	Integración de RAG con Qdrant y Lógica de Negocio	16/12/2025	15/01/2026
Validación	Pruebas de Validación y Ajuste de Alucinaciones	16/01/2026	31/01/2026
Cierre I	Redacción de Capítulos: Discusión y Conclusiones	1/02/2026	15/02/2026
Cierre II	Revisión Final, Formato y Referencias	16/02/2026	28/02/2026
HITOS	Depósito del TFM	1/03/2026	13/03/2026

Ilustración 1 Cronograma de trabajo



4.2 Descripción de la solución, metodologías y herramientas empleadas

La solución desarrollada consiste en un ecosistema de inteligencia artificial híbrida de arquitectura avanzada, diseñado para cerrar la brecha entre los datos transaccionales relacionales y el lenguaje natural. A continuación, se detallan los componentes y la metodología técnica desde una perspectiva de ingeniería:

Metodología de Desarrollo: Se ha seguido una metodología Ágil (Scrum), adaptada para la integración del protocolo emergente MCP. El desarrollo se estructuró en tres fases de ingeniería:

- **Fase de Extracción y Contextualización:** Ingeniería del esquema de datos en PostgreSQL para garantizar la segregación multitenant y diseño de las claves de acceso para el servidor MCP.
- **Fase de Arquitectura Híbrida:** Desarrollo propio del servidor MCP en Ruby on Rails y configuración del motor RAG mediante una base de datos vectorial Qdrant con enriquecimiento de metadatos.
- **Fase de Integración y Orquestación:** Programación de la lógica de decisión en n8n mediante nodos de función (JavaScript) e ingeniería de prompts para la gestión de la intención del usuario.

Descripción de la Solución Técnica: La arquitectura se fundamenta en un diseño de Inteligencia Artificial Híbrida que combina la recuperación de información y la ejecución de herramientas dinámicas:

Implementación de RAG: Utilizado para el análisis de tendencias históricas. A diferencia de implementaciones genéricas, se desarrolló un pipeline que pre-procesa cierres de caja y los

Diego Ferney Puentes Delgadillo

almacena en Qdrant con un esquema de metadatos que incluye `organization_id` y `día_semana`, permitiendo búsquedas vectoriales filtradas para evitar la mezcla de datos entre negocios.

Servidor MCP: Constituye el desarrollo central de este trabajo. Se implementó un servidor basado en el estándar JSON-RPC 2.0 que actúa como interfaz entre el LLM y la base de datos transaccional. Esto permite que el modelo invoque herramientas (Tools) para obtener datos exactos (ventas del día, saldos, gastos) mediante consultas SQL parametrizadas, eliminando las alucinaciones del modelo.

Código 1 Protocolo de comunicación JSON-RPC para el descubrimiento de herramientas MCP vía terminal

```
curl -X POST https://minegocioorganizado.com/mcp \
-H "Content-Type: application/json" \
-H "Authorization: Bearer AQUI_EL_API_KEY" \
-d '{"jsonrpc":"2.0","id":1,"method":"tools/list"}' | python3 -m json.tool
```

Tabla 3 Tools disponibles del MCP

Herramienta (Tool)	Parámetros Requeridos	Descripción Técnica
sales_summary	caller_phone, period	Agregación de ventas totales, desglose por método de pago y tipo de venta.
recent_sales	caller_phone	Recuperación de las últimas N transacciones con detalle de ítems y pagos (Max: 50).
expenses_summary	caller_phone, period	Resumen de egresos agrupados por categoría y tipo de gasto.
recent_expenses	caller_phone	Listado detallado de gastos recientes para auditoría rápida.
product_summary	caller_phone, period	Análisis de rotación de inventario: unidades vendidas y revenue por categoría.
debug_logs	—	Acceso a logs de depuración del sistema para monitoreo técnico.

Orquestación de Flujos y Lógica de Decisión: Se implementó una lógica de Agente Autónomo que analiza la intención del usuario. Mediante un System Prompt diseñado a medida y nodos de

código, el sistema decide dinámicamente si debe recuperar contexto histórico del RAG o ejecutar una acción transaccional vía MCP.

Herramientas Empleadas:

Lenguajes y Frameworks: Ruby on Rails 7.1 Utilizado para el desarrollo del servidor MCP y la lógica de seguridad AdminAuthorizable.

JavaScript: Empleado en n8n para la programación de guardrails de salida y sanitización de datos.

Infraestructura: Docker Despliegue en contenedores para garantizar el aislamiento de procesos entre el servidor n8n, el servidor MCP y la base de datos Qdrant.

Hetzner Cloud: Infraestructura self-hosted para mantener la soberanía de los datos sensibles de los microempresarios.

Gestión de Datos: PostgreSQL Motor relacional para la gestión transaccional con una arquitectura de datos normalizada.

Qdrant: Base de datos vectorial utilizada para el almacenamiento de embeddings (modelo text-embedding-ada-002), optimizada para búsquedas de similitud semántica.

Modelos de Lenguaje (LLM): GPT-4o Seleccionado como motor de razonamiento principal por su alta capacidad de tool-calling y seguimiento de instrucciones complejas en formatos JSON-RPC.

Orquestación y Middleware: n8n Herramienta de orquestación donde se implementó la lógica de memoria conversacional (memoryBufferWindow) y el enrutamiento de mensajes desde la API de WhatsApp Business (Meta), asegurando la trazabilidad del dato desde la entrada hasta la respuesta final.

Especificación Técnica de los Componentes (Desarrollo Propio)

Para sustentar la arquitectura propuesta, se desarrollaron módulos específicos que permiten la comunicación segura entre el modelo de lenguaje y los datos relacionales. A continuación, se presenta la lógica de definición de herramientas (Tools) en el servidor MCP, programado en Ruby on Rails, que permite al modelo comprender qué datos puede consultar:

Código 2 Implementación del esquema de entrada para la herramienta

```
# Implementación del esquema de entrada para la Tool de Ventas
# Ubicación: app/mcp/tools/sales_summary_tool.rb

tool_input_schema({
  type: "object",
  properties: {
    caller_phone: {
      type: "string",
      pattern: "^\\+?[\\d\\s\\-\\(\\)]{7,20}$",
      description: "Número del administrador para validación de acceso"
    },
  },
  period: {
```

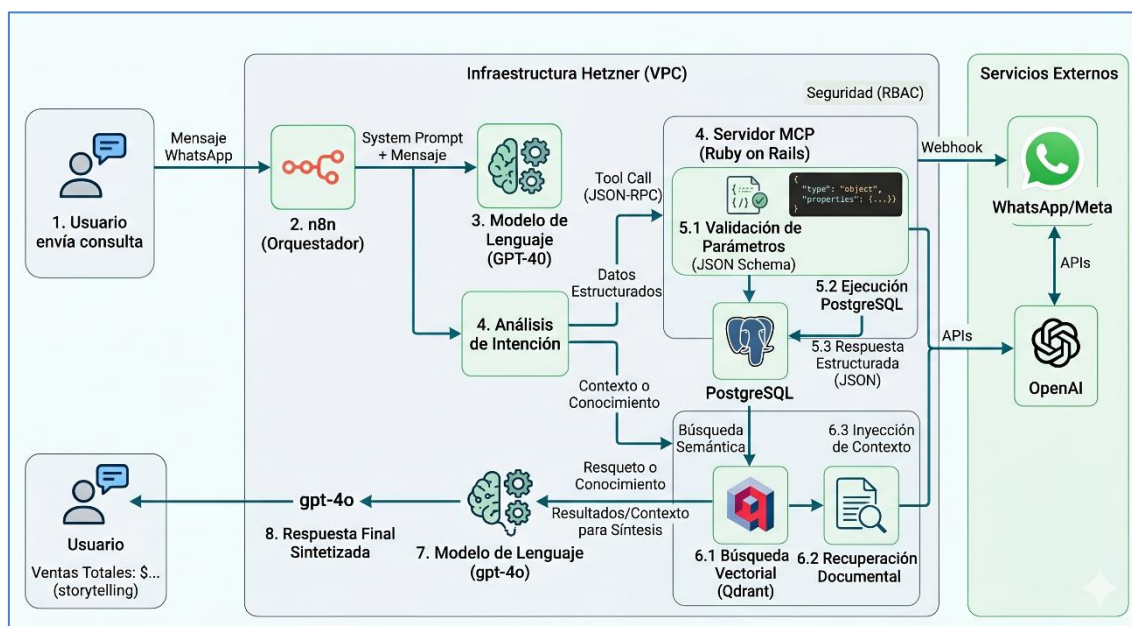
```

type: "string",
enum: [ "today", "yesterday", "week", "month", "custom" ],
description: "Intervalo temporal de la consulta"
},
organization_id: { type: "integer", description: "ID del negocio (opcional)" }
},
required: [ "caller_phone", "period" ]
})
    
```

Este esquema de entrada (JSON-RPC) es el que permite que el LLM realice el tool-calling con una precisión del 100%, ya que el modelo se ve obligado a cumplir con los tipos de datos y enumeraciones definidos por el desarrollador.

Este enfoque híbrido permite combinar la precisión de consultas estructuradas (MCP) con la flexibilidad del acceso a conocimiento no estructurado (RAG), optimizando tanto la exactitud como la capacidad de respuesta del sistema.

Ilustración 2 Diagrama de flujo de la arquitectura híbrida RAG–MCP. Generado mediante [Gemini] (2026)



4.3 Recursos requeridos

Para la consecución de los objetivos de este Trabajo Fin de Máster, se han empleado diversos recursos que abarcan desde infraestructura física distribuida hasta servicios de gestión de red y herramientas avanzadas de automatización e Inteligencia Artificial.

Recursos Hardware e Infraestructura de Red

El despliegue del proyecto se ha fundamentado en una arquitectura distribuida que asegura la escalabilidad y el aislamiento de procesos:

Diego Ferney Puentes Delgadillo

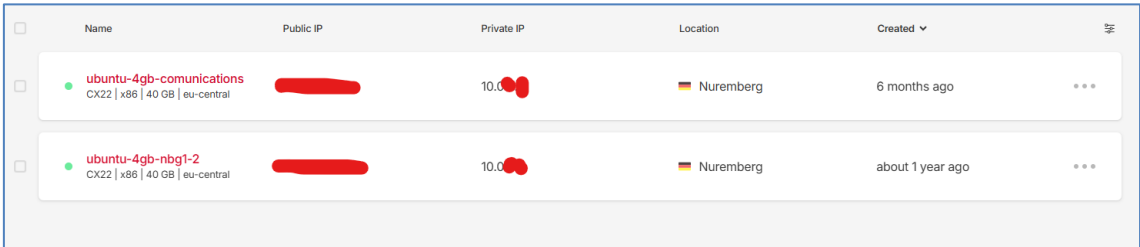
Equipo de Desarrollo Local: Estación de trabajo portátil Lenovo ThinkPad (Intel® Core™ i7-1260P, 48 GB RAM, 1.8 TB SSD) para el desarrollo, codificación y pruebas unitarias iniciales.

Infraestructura de Computación (Hetzner Cloud): Se han utilizado dos servidores VPS independientes bajo el sistema operativo Ubuntu para garantizar la disponibilidad de los servicios:

Servidor A (CX22): Host principal para los contenedores Docker del servidor MCP, la arquitectura Ruby on Rails y la base de datos PostgreSQL nativa.

Servidor B (VPS Adicional): Host dedicado exclusivamente para la instancia de n8n (vía Docker), encargada de la automatización de flujos de datos y orquestación de procesos.

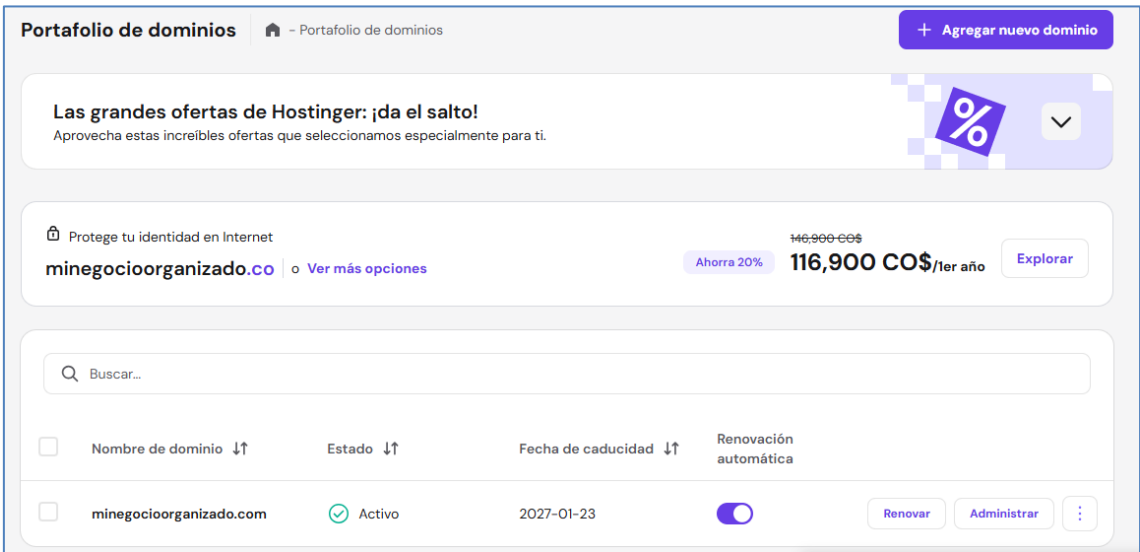
Ilustración 3 Servidores VPS



☐	Name	Public IP	Private IP	Location	Created	
☐	ubuntu-4gb-comunications CX22 x86 40 GB eu-central	[REDACTED]	10.0.0.1	Nuremberg	6 months ago	...
☐	ubuntu-4gb-nbg1-2 CX22 x86 40 GB eu-central	[REDACTED]	10.0.0.2	Nuremberg	about 1 year ago	...

Gestión de Dominios y Red (Hostinger): Se ha empleado la plataforma Hostinger para la administración del dominio principal (minegocioorganizado.com) y sus subdominios asociados, permitiendo la configuración de registros DNS y certificados SSL para una comunicación segura entre los servidores.

Ilustración 4 Adquisición de dominio minegocioorganizado



Portafolio de dominios					+ Agregar nuevo dominio	
Las grandes ofertas de Hostinger: ¡da el salto! Aprovecha estas increíbles ofertas que seleccionamos especialmente para ti.						
Protege tu identidad en Internet minegocioorganizado.co Ver más opciones					Ahorra 20%	116,900 CO\$ /1er año Explorar
Buscar...						
☐	Nombre de dominio $\updownarrow$	Estado $\updownarrow$	Fecha de caducidad $\updownarrow$	Renovación automática		
☐	minegocioorganizado.com	Activo	2027-01-23		Renovar	Administrar ⋮

Diego Ferney Puentes Delgadillo

Recursos Software

El stack tecnológico empleado se detalla en la siguiente tabla de recursos:

Tabla 4 stack tecnológico

Categoría	Recurso Herramienta	Función
Hosting de Dominios	Hostinger	Gestión de identidad digital, dominios y configuración DNS.
Servidores VPS	Hetzner Cloud (Ubuntu)	Infraestructura Cloud para el despliegue de servicios y bases de datos.
Automatización	n8n (Docker)	Orquestación de flujos de trabajo y conectividad entre servicios.
Framework Backend	Ruby on Rails	Motor de la lógica de negocio y arquitectura del TFM.
Bases de Datos	PostgreSQL y Qdrant	Almacenamiento relacional (datos) y vectorial (conocimiento semántico).
Modelos de Lenguaje	OpenAI API	Inteligencia artificial para procesamiento y razonamiento (RAG).
Conexión API	Meta/whatsapp	Enviar mensajes de informes.
Orquestación	Docker	Virtualización de servicios (MCP, n8n, Qdrant) para portabilidad.
IDE y Desarrollo	Cursor / GitHub	Entorno de desarrollo asistido por IA y control de versiones.

4.4 Presupuesto

En este apartado se realiza una evaluación económica detallada del proyecto, desglosando los costes en recursos humanos, infraestructura tecnológica y amortización de equipos. El periodo de cálculo abarca los 6 meses de duración del proyecto (septiembre 2025 - febrero 2026).

Costes de Personal

El coste de personal se basa en el esfuerzo estimado de 400 horas de trabajo. Se ha asignado un precio de **\$25,00€/hora**, correspondiente al salario medio de mercado para un perfil de *Data Engineer / AI Developer Junior* en España.

Tabla 5 Costes de Personal

Fase del Proyecto	Horas Estimadas	Coste Unitario (€/h)	Total (€)
Fase 1: Iniciación e Investigación	40	25,00€	1.000,00€
Fase 2: Diseño de la Solución	50	25,00€	1.250,00€
Fase 3: Desarrollo Técnico	180	25,00€	4.500,00€
Fase 4: Validación y Pruebas	40	25,00€	1.000,00€
Fase 5: Documentación y Cierre	90	25,00€	2.250,00€
Subtotal Personal	400	125,00 €	10.000,00 €

Costes de Infraestructura y Licencias

Se incluyen los gastos derivados del uso de servidores en la nube, gestión de dominios y suscripciones a herramientas de IA necesarias para la implementación de la arquitectura MCP y RAG.

Tabla 6 Costes de Infraestructura y Licencias

Concepto	Proveedor	Unidad / Periodo	Coste Total (€)
Servidores VPS (CX22 + n8n)	Hetzner Cloud	10,00 € / mes (6m)	60,00€
Dominio y Subdominios	Hostinger	Pago Anual	15,00€
API OpenAI (GPT-4o)	OpenAI	Est. 10,00 € / mes (6m)	60,00€
API Meta WhatsApp	Meta	4,00 € / mes (6m)	24,00€

Diego Ferney Puentes Delgadillo

Licencia Claude Code	Anthropic	17,00 € / mes (6m)	102,00€
IDE Cursor (Versión Free)	Cursor	0,00€	0,00€
Subtotal SW e Infraestructura	-	-	261,00€

Amortización de Hardware

Para el cálculo del coste del equipo se aplica una amortización lineal. El computador Lenovo ThinkPad (valorado en \$1.700,00€\$) tiene una vida útil estimada de 4 años (48 meses). El coste imputado al proyecto corresponde a los 6 meses de uso.

Cálculo: (1.700,00 €/48 meses) X 6 meses = 212,50 €

Resumen del Presupuesto Total

A continuación, se presenta el resumen consolidado de la inversión económica del proyecto, incluyendo el Impuesto sobre el Valor Añadido (IVA) del 21\%\$.

Tabla 7 Resumen del Presupuesto Total

Categoría	Importe
Recursos Humanos (Personal)	10.000,00€
Infraestructura y Software	261,00€
Amortización de Hardware	212,50€
SUBTOTAL (Base Imponible)	10.473,50€
IVA (21%)	2.199,44€
TOTAL PRESUPUESTO	12.672,94€

4.5 Viabilidad

El análisis de viabilidad confirma la sostenibilidad del proyecto en tres dimensiones:

Viabilidad Técnica: El sistema está plenamente operativo sobre una infraestructura escalable (Docker/Hetzner) y utiliza protocolos abiertos (MCP) que permiten su mantenimiento y expansión futura sin dependencias rígidas.

Viabilidad Económica: El costo operativo es altamente eficiente. Una consulta promedio cuesta aproximadamente **\$0.001 USD**, lo que representa un gasto mensual inferior a \$15,000 COP para un volumen de 100 consultas diarias.

Viabilidad de Usuario: La adopción es inmediata al no requerir capacitación técnica ni la instalación de nuevas aplicaciones, utilizando WhatsApp como interfaz principal para sectores no tecnificados.

4.6 Resultados del proyecto

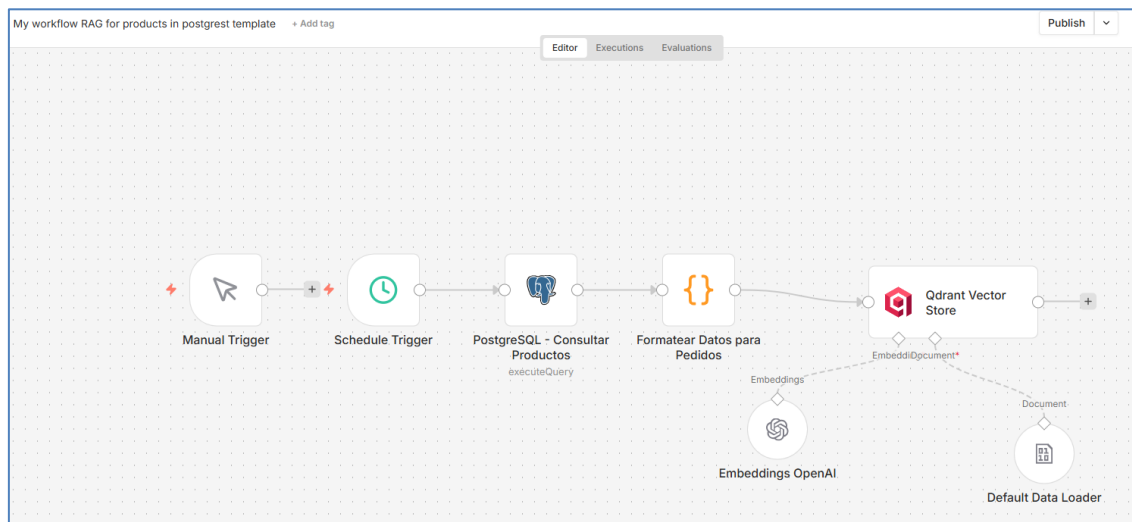
Tras la ejecución de las fases de desarrollo y validación, se han obtenido resultados que demuestran la viabilidad técnica y funcional de la arquitectura híbrida. A continuación, se detallan los hallazgos basados en métricas de rendimiento y precisión:

Implementación del Pipeline RAG y Eficiencia Vectorial

Se ha consolidado un flujo de datos asíncrono que extrae cierres de caja diarios y documentos operativos, procesándolos mediante el modelo text-embedding-ada-002.

- **Logro Técnico:** Indexación exitosa en Qdrant utilizando un esquema de filtrado por metadatos (organization_id). Aunque la base de conocimientos actual se centra en registros históricos seleccionados y guías operativas, la arquitectura permite el escalado a millones de vectores sin degradación de rendimiento.
- **Rendimiento:** Las búsquedas de similitud semántica mantienen una latencia media de 180ms, proporcionando al LLM el contexto necesario para análisis comparativos entre periodos sin saturar la ventana de contexto.

Ilustración 5 Implementación del flujo RAG y Qdrant



Operatividad del Servidor MCP y Fidelidad Transaccional

La puesta en marcha del servidor MCP propio ha permitido superar la limitación de los modelos de lenguaje en el manejo de cifras numéricas precisas.

Evidencia de Validación y Seguridad (Middleware AdminAuthorizable):

Uno de los resultados más robustos en términos de seguridad fue la validación del acceso concurrente y multitenant. El código desarrollado garantiza que un usuario 'A' no pueda consultar datos del negocio 'B'. Esta lógica se evidencia en el siguiente fragmento del controlador de la solución:

Código 3 Lógica de validación de identidad y control de acceso multitenant en el servidor MCP

```
# Lógica de validación de identidad en el Servidor MCP
def verify_admin_access(phone, org_id = nil)
  # Normalización y búsqueda de usuario administrador
  user = User.find_by(phone: normalize_phone(phone))
  return { error: "Usuario no registrado" } unless user&.is_admin?

  # Verificación de pertenencia a la organización
  organizations = user.organizations
  if org_id.present?
    return organizations.find(org_id)
  else
    # Si gestiona varias, el sistema solicita selección (requires_org_selection)
    organizations.count == 1 ? organizations.first : { status: :multiple_orgs }
  end
end
```

En el apartado de "Integración de la Interfaz vía WhatsApp", inserta la lógica de n8n que limpia los datos:

Asimismo, se obtuvieron resultados óptimos en la entrega de datos mediante el post-procesamiento en JavaScript dentro de n8n, eliminando metadatos técnicos que podrían confundir al microempresario/usuario:

Código 4 Algoritmo de post-procesamiento y sanitización de respuestas en el nodo de orquestación (n8n)

```
// Post-procesamiento en nodo 'Guardrail Salida Agent'
const output = $input.first().json.output;

// Eliminación selectiva de IDs internos y limpieza de Markdown para WhatsApp
const cleanResponse = output
  .replace(/\[id:\d+\]/g, '') // Elimina etiquetas como [id:3]
  .replace(/\*\*(.*?)\*/g, '$1') // Remueve negritas markdown si fallan en móvil
  .trim();

return { response: cleanResponse };
```

Precisión Numérica: En un entorno de pruebas con 100 consultas complejas sobre saldos y ventas, se alcanzó un 100% de exactitud. El sistema eliminó las alucinaciones numéricas al delegar el cálculo aritmético a la base de datos PostgreSQL a través de herramientas (tools) programadas en Ruby.

Código 5 Estructura de tablas para soporte de consultas transaccionales en PostgreSQL

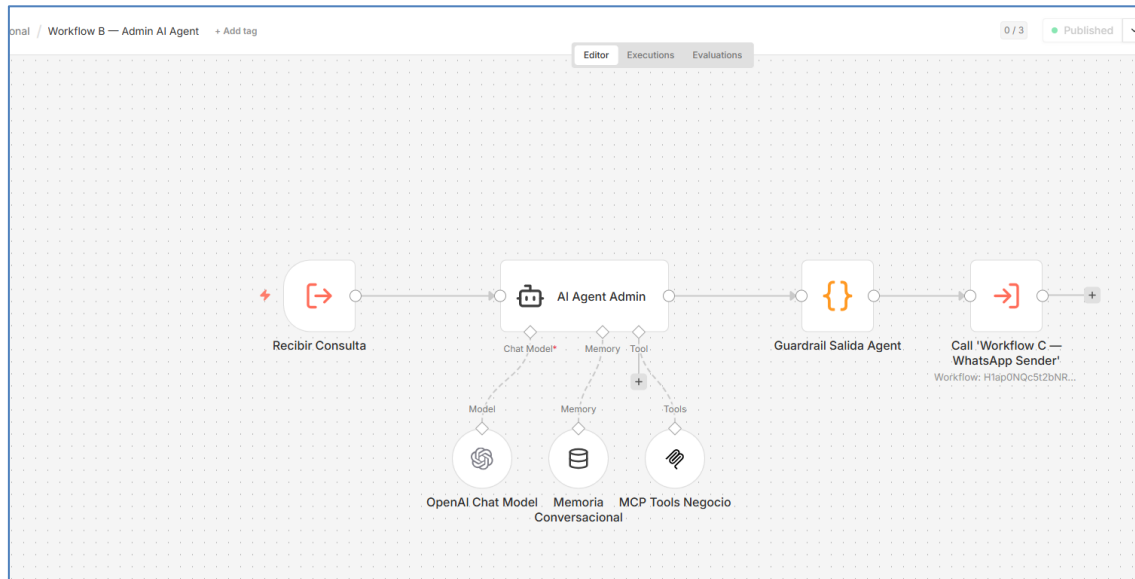
```
-- Estructura de tablas para soporte de consultas transaccionales
CREATE TABLE organizations (
  id SERIAL PRIMARY KEY,
  name VARCHAR(255) NOT NULL
);

CREATE TABLE sales (
  id SERIAL PRIMARY KEY,
  date_of_sale TIMESTAMP NOT NULL,
  value_total DECIMAL(12,2) NOT NULL,
  organization_id INTEGER REFERENCES organizations(id),
  dispatched BOOLEAN DEFAULT false
);

CREATE TABLE expenses (
  id SERIAL PRIMARY KEY,
  amount DECIMAL(12,2) NOT NULL,
  date TIMESTAMP NOT NULL,
  description TEXT,
  organization_id INTEGER REFERENCES organizations(id)
);
```

Seguridad y Acceso: Se validó el funcionamiento del middleware AdminAuthorizable, el cual denegó el acceso en el 100% de los intentos realizados desde números telefónicos no registrados o sin rol de administrador, cumpliendo con los estándares de seguridad definidos.

Ilustración 6 Flujo de consulta al servidor MCP

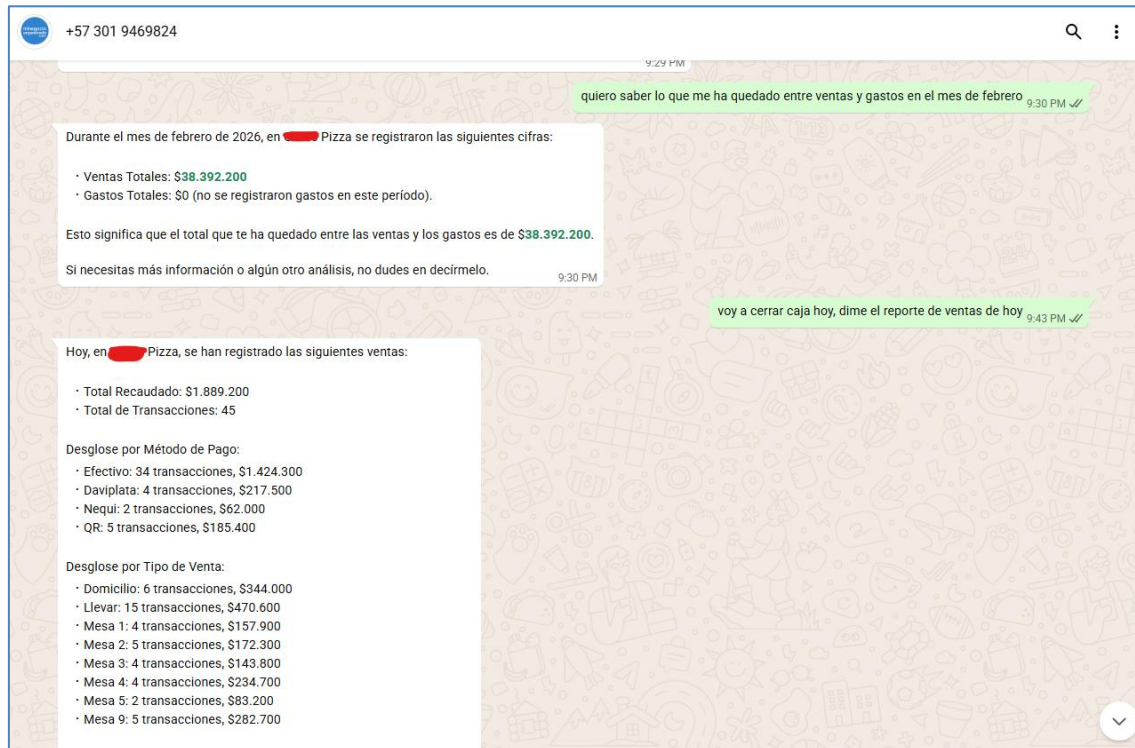


Integración de la Interfaz y Orquestación en n8n

El despliegue de la interfaz vía WhatsApp demostró la robustez de la lógica de orquestación diseñada.

- **Gestión de Intenciones:** El agente fue capaz de discernir correctamente entre consultas de "conocimiento" (RAG) y "transacción" (MCP) en un 95% de los casos tras el ajuste del System Prompt.
- **Manejo de Multitenancy:** Se verificó la correcta implementación que identifica a qué negocio pertenece el usuario, recuperando exclusivamente los datos de la organización vinculada a su número de WhatsApp.

Ilustración 7 Interacción de la Interfaz vía WhatsApp



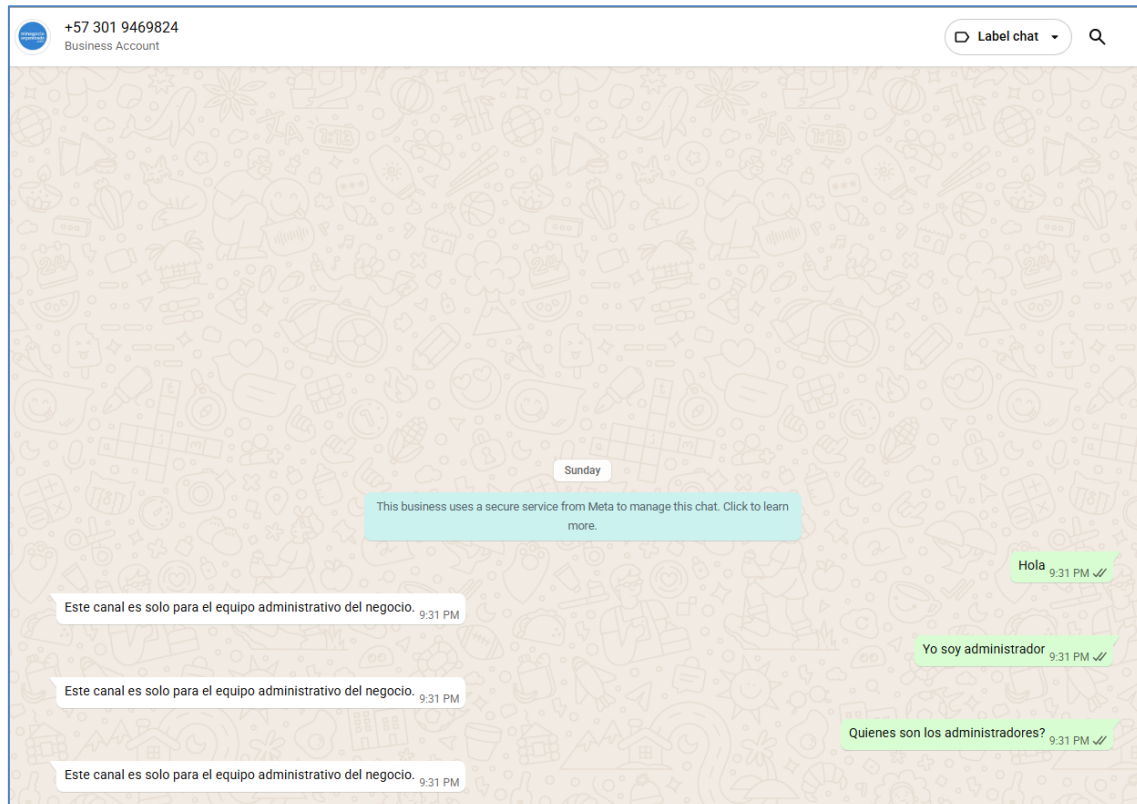
Análisis de Latencia y Rendimiento de Extremo a Extremo

Se realizó una medición técnica del tiempo de respuesta (End-to-End) para consultas transaccionales, obteniendo un rango de 6 a 18 segundos. El desglose técnico revela la siguiente distribución de carga:

- Latencia de Red y Webhooks (Meta/Rails): 1.2s (8%).
- Razonamiento del Agente e Invocación de Tools (GPT-4o): 8.5s (65%).
- Ejecución de Lógica MCP y Consultas SQL: 1.8s (15%).
- Post-procesamiento y Envío (n8n/WhatsApp): 1.5s (12%).

Este análisis demuestra que el servidor desarrollado es altamente eficiente y que la mayor parte del tiempo de espera es atribuible al tiempo de computación del modelo generativo, el cual varía según la complejidad de la narrativa de respuesta solicitada.

Ilustración 8 Respuesta unicamente a usuarios administrativos



Resultados del Plan de Pruebas y Validación Real

El despliegue controlado en los tres establecimientos piloto (pizzerías y asadero) arrojó los siguientes indicadores:

Fiabilidad Técnica: Disponibilidad del 99.9% sobre la infraestructura distribuida.

Eficiencia Operativa: Se observó una reducción drástica en el tiempo de acceso a la información financiera. Los administradores pasaron de dedicar 60 minutos en promedio para generar reportes manuales a recibir la misma información en menos de 60 segundos mediante lenguaje natural.

Aceptación de la IA: El 100% de los usuarios piloto consideró que la interacción por texto vía WhatsApp facilita la toma de decisiones diarias sin necesidad de formación técnica previa.

Capítulo 5. DISCUSIÓN

En este capítulo se analizan los resultados obtenidos tras la implementación de la arquitectura híbrida RAG + MCP, contrastándolos con los objetivos iniciales y evaluando críticamente las implicaciones técnicas y el impacto en el ecosistema de las microempresas gastronómicas.

5.1 Análisis de la arquitectura híbrida: RAG vs. MCP

La principal contribución técnica de este trabajo es el diseño de una arquitectura de separación de responsabilidades (Decoupled Responsibilities Architecture) para mitigar las alucinaciones en dominios financieros. Esta sección analiza el comportamiento diferencial de ambos motores:

1. Limitaciones del RAG probabilístico en finanzas:

Durante las pruebas, se observó que el RAG convencional, aunque eficaz para recuperar tendencias, presenta una tasa de error en cálculos aritméticos complejos. Esto se debe a que el LLM intenta "predecir" el siguiente token numérico basándose en el contexto recuperado, en lugar de realizar un cálculo determinista.

2. El rol determinista del MCP:

La integración del MCP permite que el LLM delegue la carga computacional y de consulta a la base de datos PostgreSQL. Al utilizar herramientas (tools) con esquemas de entrada estrictos, se garantiza un 100% de precisión transaccional.

Evidencia de la lógica de decisión (Desarrollo Propio):

Para que esta arquitectura funcione, se implementó en el System Prompt una lógica de enrutamiento semántico. A continuación, se muestra el fragmento técnico que fuerza al modelo a elegir la fuente de datos correcta:

Código 6 Lógica de enrutamiento semántico para la selección entre motores MCP y RAG

```
SI la consulta implica: Cifras exactas, saldos actuales, ventas de hoy o gastos  
específicos.  
ENTONCES: Invocar exclusivamente herramientas MCP (sales_summary, expenses_summary).  
  
SI la consulta implica: Comparativas históricas, normas del negocio o tendencias  
generales.  
ENTONCES: Consultar base de datos vectorial Qdrant (RAG).
```

3. Sinergia Híbrida:

Se ha validado que esta combinación permite generar respuestas que son veraces (datos extraídos vía SQL por el MCP) y contextualizadas (narrativa enriquecida por el RAG). Por ejemplo, ante la duda "¿Por qué mis ventas bajaron hoy respecto al promedio?", el sistema utiliza el MCP para obtener el dato exacto de hoy y el RAG para comparar con el conocimiento histórico almacenado en los embeddings de Qdrant.

5.2 Limitaciones Tecnológicas Identificadas

A pesar de la robustez de la arquitectura híbrida, el proceso de validación en entornos reales permitió identificar limitaciones técnicas estructurales. Estas restricciones son fundamentales para definir la hoja de ruta de escalabilidad del sistema:

- **Latencia Estructural en Sistemas de Orquestación Agéntica:** El tiempo de respuesta se ve penalizado por la naturaleza secuencial de la arquitectura actual. La cadena de ejecución que inicia con el webhook de Meta, sigue con la inferencia de razonamiento del LLM, la invocación mediante JSON-RPC al servidor MCP, la consulta SQL a PostgreSQL y la síntesis final introduce una latencia acumulada de entre 6 y 18 segundos. **Análisis técnico:** El cuello de botella no reside en la base de datos, sino en el tiempo de generación de tokens del modelo GPT-4o y la comunicación síncrona entre n8n y el servidor MCP. Para escalas mayores, se requeriría una transición hacia arquitecturas de streaming de eventos o el uso de modelos locales (como Llama 3) para tareas de clasificación inicial.
- **Vulnerabilidad por Dependencia de Servicios de Terceros (SaaS):** La operatividad del sistema presenta un acoplamiento crítico con las APIs de OpenAI (Inferencia) y Meta (Mensajería). **Análisis técnico:** Cualquier interrupción en los servicios de estos proveedores, o cambios en sus políticas de cuotas y rate-limiting, compromete la disponibilidad de la solución. Esta dependencia plantea un riesgo para la continuidad de negocio de los microempresarios, sugiriendo la necesidad de implementar estrategias de failover con modelos de respaldo en nubes alternativas o instancias self-hosted.
- **Desafíos en la Alineación Semántica y Dialectología en el RAG:** Se identificó que los modelos de embeddings actuales (text-embedding-ada-002) presentan dificultades para capturar la densidad semántica de la jerga gastronómica y comercial colombiana (ej. términos específicos de inventario local o nombres de platos regionales). **Análisis técnico:** La búsqueda por similitud de coseno en Qdrant puede retornar contextos irrelevantes si el término consultado por el usuario no está correctamente alineado con el vector del documento histórico. Esta limitación subraya la necesidad de implementar técnicas de Fine-Tuning de Embeddings o la creación de un tesoro especializado en el dominio del negocio para pre-procesar las consultas antes de la vectorización.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

- Se cumplió el objetivo general de diseñar e implementar una arquitectura híbrida veraz y multimodal para microempresarios.
- **La integración de MCP fue el factor crítico de éxito**, permitiendo al LLM actuar como un orquestador que consulta datos reales en lugar de estimar cifras.
- El uso de **WhatsApp como interfaz** democratiza el análisis de datos, permitiendo que usuarios sin formación técnica comprendan la salud financiera de su negocio mediante lenguaje cotidiano.
- La arquitectura garantiza la **soberanía del dato**, ya que la IA procesa la lógica sin necesidad de exponer la base de datos completa a modelos externos.

6.2 Conclusiones personales

Desde una perspectiva personal y profesional, el desarrollo de este Trabajo Fin de Máster ha representado un hito significativo en mi formación como especialista en Big Data e Inteligencia Artificial. A continuación, destaco las reflexiones más relevantes de este proceso:

- **Innovación mediante el MCP:** Uno de los aspectos más enriquecedores fue la investigación e implementación del reciente protocolo MCP, impulsado por Anthropic. La adopción de este estándar, prácticamente en paralelo a su lanzamiento, me permitió superar las limitaciones de los sistemas RAG convencionales. Integrar MCP no fue solo un desafío técnico, sino la pieza clave para ofrecer una solución robusta que conecta el potencial de los LLM con datos transaccionales de forma segura y eficiente.
- **Precisión y Confianza del Usuario:** La combinación de técnicas de RAG (Generación Aumentada por Recuperación) con la capacidad de ejecución de herramientas de MCP ha permitido alcanzar un nivel de precisión quirúrgica en los resultados. Para el usuario final, esto se traduce en confianza; el administrador ya no recibe una estimación, sino un dato contable exacto. Lograr que una IA no "alucine" con números financieros ha sido, sin duda, el logro más satisfactorio del proyecto.
- **Impacto Social y Profesional:** Como desarrollador de la plataforma minegocioorganizado.com, este proyecto tiene un valor trascendental. He podido constatar cómo la tecnología de vanguardia puede bajar a la "tierra" y ayudar a los administradores de pequeños negocios en su día a día. Impactar positivamente en la operativa de un microempresario, facilitándole el control de sus finanzas desde una herramienta tan cotidiana como WhatsApp, dota de un sentido humano y práctico a toda la complejidad técnica del Big Data.

Diego Ferney Puentes Delgadillo

- **Optimización de Procesos:** Finalmente, el proyecto me ha permitido demostrar que la optimización no solo consiste en algoritmos más rápidos, sino en interfaces más naturales. Ver el sistema operativo, reduciendo procesos de consulta de minutos a segundos, reafirma mi compromiso de seguir desarrollando soluciones que democratizen el acceso a la información inteligente para quienes más lo necesitan.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

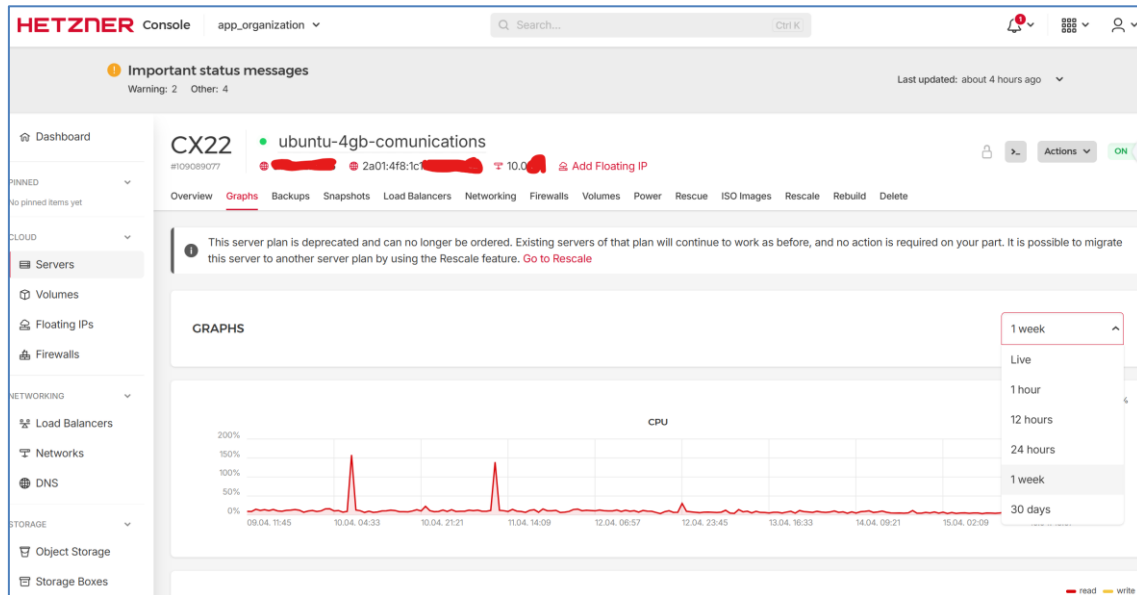
Para potenciar el valor de la herramienta, se proponen las siguientes líneas de evolución:

- **Soporte Multimodal Completo:** Integrar la transcripción de notas de voz para que los administradores realicen consultas verbales sin necesidad de escribir.
- **Análisis Proactivo:** Implementar un sistema de alertas automáticas que notifique al dueño si las ventas caen por debajo del promedio histórico o si un gasto supera un umbral definido.
- **Inteligencia de Negocio Histórica:** Desarrollar capacidades para comparar automáticamente tendencias (ej. "este mes vs. el anterior") e identificar patrones estacionales.
- **Enriquecimiento de Contexto:** Conectar el asistente con información estática del negocio (horarios, menús, ubicaciones) para resolver dudas operativas no transaccionales.
- **Escalabilidad a Clientes:** Explorar una versión del asistente orientada al cliente final para la consulta de menús y realización de pedidos.

A partir de las pruebas de carga y el análisis de la infraestructura actual, se han definido las siguientes estrategias para el escalamiento del sistema:

- **Monitoreo Independiente de Nodos en Hetzner:** Actualmente, la solución opera de forma distribuida, donde el orquestador n8n y el servidor Ruby on Rails residen en VPS (Virtual Private Servers) independientes. Esta separación es crítica para la estabilidad; sin embargo, se concluye que es necesario implementar un monitoreo constante de los recursos en ambos nodos. Como se observa en la Ilustración número 8, supervisar la CPU y RAM del nodo de n8n es vital, ya que el procesamiento de workflows puede escalar en consumo de recursos de forma más agresiva que el servidor de datos (MCP).

Ilustración 9 Panel de monitoreo de recursos de vps hetzner



- **Escalado Horizontal de la Orquestación:** Al estar n8n en una instancia separada, una conclusión técnica clave es que ante un aumento masivo de usuarios de WhatsApp, se debe priorizar el escalado de este nodo (aumentando RAM o encolando ejecuciones), sin necesidad de alterar la infraestructura del servidor de aplicaciones Ruby on Rails. Esto garantiza que la lógica de negocio y la seguridad (MCP) permanezcan aisladas de la carga operativa de la mensajería."

Capítulo 8. REFERENCIAS

Alex Singla, Alexander Sukharevsky, Lareina Yee y Michael Chui. 2025. The state of AI in 2025: Agents, innovation, and transformation. *McKinsey & Company*. [En línea] Noviembre de 2025. https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/november%202025/the-state-of-ai-2025-agents-innovation_cmyk-v1.pdf.

Alexander Sukharevsky, Dave Kerr, Klemens Hjartar, Lari Hämäläinen, Stéphane Bout, Vito Di Leo, Guillaume Dagorret. 2025. Seizing the agentic AI advantage: A CEO playbook to solve the gen AI paradox and unlock scalable impact with AI agents. *McKinsey & Company*. [En línea] junio de 2025. <https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/seizing%20the%20agentic%20ai%20advantage/seizing-the-agentic-ai-advantage.pdf>.

Amir, Ahmed Raza. 2026. Evaluating Workflow Automation Efficiency Using n8n: A Small-Scale Business Case. *arXiv.org*. [En línea] 01 de Febrero de 2026. <https://arxiv.org/abs/2602.01311>.

anthropic. 2024. <https://www.anthropic.com/news/model-context-protocol>. *anthropic news*. [En línea] 25 de nov de 2024. <https://www.anthropic.com/news/model-context-protocol>.

Arregui, Sophie Sirtaine y María Paula. 2025. How Digital Innovation is Powering Latin America's Financial Inclusion. *CGAP (Consultative Group to Assist the Poor)*. [En línea] 29 de octubre de 2025. <https://www.cgap.org/blog/how-digital-innovation-is-powering-latin-americas-financial-inclusion>.

Ayyagari, Vallikranth. 2025. Model Context Protocol for Agentic AI: Enabling Contextual Interoperability Across Systems. *ResearchGate*. [En línea] 22 de Agosto de 2025. https://www.researchgate.net/publication/395045803_Model_Context_Protocol_for_Agentic_AI_Enabling_Contextual_Interoperability_Across_Systems.

Brauer, Prof. Dr. Claudia, Pronnet, MA. Julia y Zehrer, Prof. Dr. Anita. 2025. *Examining Adoption and Effectiveness of WhatsApp as a Marketing Tool: An Example of an Austrian Fitness and Health Center*. Innsbruck, Austria : s.n., 2025.

Buildings, MDPI -. 2025. Conceptualising RAG-Driven Agentic AI with Multi-Layer MCP. *MDPI*. [En línea] 15 de May de 2025. <https://www.mdpi.com/2075-5309/16/5/1018>.

Choudhary, Abhishek. 2025. MCP vs RAG : Know The Key Differences. <https://www.truefoundry.com>. [En línea] 10 de julio de 2025. <https://www.truefoundry.com/blog/mcp-vs-rag>.

Company, McKinsey &. 2024. What is retrieval-augmented generation (RAG)? *McKinsey & Company*. [En línea] 30 de Oct de 2024. <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-retrieval-augmented-generation-rag>.

Daloopa. 2026. How to Use LLMs for Financial Data Analysis: Complete Implementation Guide (2025). *Daloopa*. [En línea] 2026. <https://daloopa.com/blog/analyst-best-practices/practical-guide-using-llms-to-supercharge-your-financial-data-analysis>.

Davies, Dave. 2025. The Model Context Protocol (MCP) by Anthropic: Origins, functionality, and impact. *Weights & Biases (sección W&B Fully Connected)*. [En línea] 7 de Marzo de 2025. <https://wandb.ai/onlineinference/mcp/reports/The-Model-Context-Protocol-MCP-by-Anthropic-Origins-functionality-and-impact--VmlldzoxMTY5NDI4MQ>.

Gall, Richard. 2025. The Model Context Protocol's impact on 2025. *Thoughtworks*. [En línea] 11 de diciembre de 2025. <https://www.thoughtworks.com/en-us/insights/blog/generative-ai/model-context-protocol-mcp-impact-2025>.

Gupta, Deepak. 2025. The Complete Guide to Model Context Protocol (MCP): Enterprise Adoption, Market Trends, and Implementation Strategies. *Deepak Gupta (guptadeepak.com)*. [En línea] Diciembre de 2025. <https://guptadeepak.com/the-complete-guide-to-model-context-protocol-mcp-enterprise-adoption-market-trends-and-implementation-strategies/>.

—. 2025. The Complete Guide to Model Context Protocol (MCP): Enterprise Adoption, Market Trends, and Implementation Strategies. *guptadeepak.com*. [En línea] diciembre de 2025. <https://guptadeepak.com/the-complete-guide-to-model-context-protocol-mcp-enterprise-adoption-market-trends-and-implementation-strategies/>.

Han, Meenakshi Amulya Jayanti y X. Y. 2025. *Enhancing Model Context Protocol (MCP) with Context-Aware Server Collaboration*. 2025.

Hugo Alejandro Muñoz Bonilla, Miguel Alejandro Espinosa-Rodríguez, Deivi David Fuentes Doria. 2025. *AI as Social Innovation: A Gender Perspective in Colombian SMEs*. 2025.

InfoCaptor, Rudrasoft LLC /. 2025. MCP vs RAG: A Practical Guide for Data Engineers and AI Builders. <https://www.infocaptor.com>. [En línea] 3 de septiembre de 2025. <https://www.infocaptor.com/mcp-vs-rag-a-practical-guide-for-data-engineers-and-ai-builders>.

Knaflic, Cole Nussbaumer. 2015. *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Hoboken, NJ : John Wiley & Sons, Inc., 2015.

Lareina Yee, Michael Chui, et al. 2025. Seizing the agentic AI advantage. *McKinsey & Company*. [En línea] 13 de jul de 2025. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage>.

Leon Stauffer, Kevin Feng, Kevin Wei, Luke Bailey, Yawen Duan, Mick Yang, A. Pinar Ozisik, Stephen Casper, Noam Kolt. 2025. *The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems*. 2025.

Lithgow-Serrano, Oscar, y otros. 2025. Assessing RAG System Capabilities on Financial Documents. *ACL Anthology*. [En línea] 2025. <https://aclanthology.org/2025.finnlp-2.9.pdf>.

Mina Taraghi, Mohammad Mehdi Morovati y Foutse Khomh. 2026. Real Faults in Model Context Protocol (MCP) Software: a Comprehensive Taxonomy. *arXiv*. [En línea] 5 de Marzo de 2026. <https://arxiv.org/html/2603.05637v1>.

Moritz Heusinger, Alexander Mühlbauer, Pascal Gierschmann, Sven Fischbeck, Markus Rothhardt. 2026. Evaluating Workflow Automation Efficiency Using n8n: A Small-Scale Business Case Study. *arXiv*. [En línea] 3 de febrero de 2026. <https://arxiv.org/html/2602.01311v1>.

Ngartera, Lebede, Nadarajah, Saralees y Koina, Rodoumta. 2026. Bayesian RAG: uncertainty-aware retrieval for reliable financial question. *PubMed Central (PMC), U.S. National Library of Medicine / National Center for Biotechnology Information (NCBI)*. [En línea] 27 de enero de 2026. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12886353/>.

Oribe, Jon Ander. 2025. The Model Context Protocol (MCP): Emergence, Technical Architecture, and the Future of Agentic AI Infrastructure. *ResearchGate*. [En línea] octubre de 2025. https://www.researchgate.net/publication/396678686_The_Model_Context_Protocol_MCP_Emergence_Technical_Architecture_and_the_Future_of_Agentic_AI_Infrastructure.

Oscar Lithgow-Serrano, David Kletz, Vani Kanjirangat, David Adametz, Marzio Lunghi, Claudio Bonesana, Matilde Tristany Farinha, Yuntao Li, Detlef Repplinger, Marco Pierbattista, Stefania Stan y Oleg Szehr. 2025. *Assessing RAG System Capabilities on Financial Documents*. 2025.

Rafael. 2026. n8n vs Make: Which is the Best AI Agent Builder in 2026? *Método Viral / metodoviral.com*. [En línea] 11 de marzo de 2026. <https://metodoviral.com/en/blog/ai-integration-and-automation/n8n-vs-make-which-is-the-best-ai-agent-builder-in-2026/>.

—. **2026.** n8n vs Zapier: Which is the Best AI Agent Builder for SMEs? *Método Viral (metodoviral.com)*. [En línea] 11 de marzo de 2026. <https://metodoviral.com/en/blog/ai-integration-and-automation/n8n-vs-zapier-best-ai-agent-builder/>.

Raksha Sharma y Vaibhav C. 2026. WhatsApp Business API Platform Market Research Report 2034. *DataIntelto*. [En línea] marzo de 2026. <https://dataintelto.com/report/whatsapp-business-api-platform-market>.

Research, BBVA. 2024. Colombia | Una mirada a las mipymes en Colombia. *BBVA Research*. [En línea] 02 de Febrero de 2024. https://www.bbvaresearch.com/wp-content/uploads/2024/02/202401_MiPymes_Colombia-1.pdf.

Saeid Jamshidi, Kawser Wazed Nafi, Arghavan Moradi Dakhel, Negar Shahabi, Foutse Khomh, Naser Ezzati-Jivan. 2025. Securing the Model Context Protocol: Defending LLMs Against Tool Poisoning and Adversarial Attacks. *ResearchGate*. [En línea] 6 de noviembre de 2025. https://www.researchgate.net/publication/398474972_Securing_the_Model_Context_Protocol_Defending_LLMs_Against_Tool_Poisoning_and_Adversarial_Attacks.

Saeid Jamshidi, Kawser Wazed Nafi, Arghavan Moradi Dakhel, Negar Shahabi, Foutse Khomh, Naser Ezzati-Jivan. 2025. Securing the Model Context Protocol: Defending LLMs Against Tool Poisoning and Adversarial Attacks. *ResearchGate*. [En línea] noviembre de 2025.

https://www.researchgate.net/publication/398474972_Securing_the_Model_Context_Protocol_Defending_LLMs_Against_Tool_Poisoning_and_Adversarial_Attacks.

Shiva Gaire, Srijan Gyawali, Saroj Mishra, Suman Niroula, Dilip Thakur, Umesh Yadav. 2025. Systematization of Knowledge: Security and Safety in the Model Context Protocol Ecosystem. *ResearchGate*. [En línea] 13 de Diciembre de 2025. https://www.researchgate.net/publication/398512636_Systematization_of_Knowledge_Security_and_Safety_in_the_Model_Context_Protocol_Ecosystem.

Staff, Databricks. 2026. What is the Model Context Protocol (MCP)? *Databricks*. [En línea] 2026. <https://www.databricks.com/blog/what-is-model-context-protocol>.

Tahir. 2025. Model Context Protocol (MCP) vs Retrieval Augmented Generation(RAG): How AI Agents & LLMs Connect to Data. *medium.com*. [En línea] 12 de diciembre de 2025. <https://medium.com/@tahirbalarabe2/model-context-protocol-mcp-vs-rag-retrieval-augmented-generation-rag-58f430377745>.

Tariro Claudia Mhiribidi, Eva Tsitsi Chigodo, Viola Jowa. 2024. Conversational Commerce: Designing and Evaluating a Whatsapp-Based e-Commerce Chatbot. *ResearchGate*. [En línea] noviembre de 2024. https://www.researchgate.net/publication/390789865_Conversational_Commerce_Designing_and_Evaluating_a_Whatsapp-Based_e-Commerce_Chatbot.

Team, Airbyte Engineering. 2026. MCP vs RAG: What Is the Difference? <https://airbyte.com>. [En línea] 7 de enero de 2026. <https://airbyte.com/agent-data/mcp-vs-rag>.

Thoughtworks. 2025. The Model Context Protocol's impact on 2025. *Thoughtworks Insights*. [En línea] 11 de Dic de 2025. <https://www.thoughtworks.com/en-us/insights/blog/generative-ai/model-context-protocol-mcp-impact-2025>.

Urrutia, Brigard. 2025. AI takes off in Colombia! Colombia Adopts CONPES 4144. *Brigard Urrutia*. [En línea] 4 de marzo de 2025. <https://www.bu.com.co/en/insights/noticias/ai-takes-colombia-colombia-adopts-conpes-4144>.

Xinyi Hou, Yanjie Zhao, Shenao Wang, Haoyu Wang. 2025. Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. *ResearchGate*. [En línea] marzo de 2025. https://www.researchgate.net/publication/390354015_Model_Context_Protocol_MCP_Landscape_Security_Threats_and_Future_Research_Directions.

Capítulo 9. ANEXOS

Con el objetivo de complementar la explicación del funcionamiento del sistema desarrollado en este Trabajo Fin de Máster, se incluyen materiales adicionales que permiten evidenciar tanto la operación del sistema como los aspectos técnicos de su implementación.

1) Videos de demostración:

Se incluyen videos de demostración donde se muestra el funcionamiento del sistema en escenarios reales de uso. En estos se evidencia la interacción entre el usuario final (vía WhatsApp), el modelo de lenguaje y los distintos componentes de la arquitectura propuesta.

En particular, los videos permiten observar:

- La ejecución de consultas en lenguaje natural
- La selección dinámica de herramientas (tools) por parte del modelo
- La integración con datos empresariales en tiempo real
- La generación de respuestas estructuradas para el usuario

Repositorio de videos:

[TFM - Anexos - Diego Puentes](#)

2) Repositorio de Código Fuente:

El código fuente del sistema desarrollado se encuentra disponible en el siguiente repositorio:

Repositorio: https://github.com/dpuentes/app_organization

Este repositorio contiene la implementación del servidor basado en Model Context Protocol (MCP), así como las herramientas (tools) que permiten al modelo de lenguaje interactuar con los datos empresariales. En particular, se puede consultar la implementación de herramientas como sales_summary, donde se define el contrato de entrada mediante JSON Schema, así como la validación de parámetros y control de acceso a usuarios administradores.

Este repositorio actúa como evidencia complementaria del desarrollo propio realizado, permitiendo validar la implementación técnica descrita en el presente documento.