



**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

**Diseño e implementación de una solución de datos
para el análisis del uso de los recursos públicos del
Sistema General de Participaciones a través de las
Cuentas Maestras en Colombia**

CRISTIAN CAMILO CHAPARRO MUÑOZ

Dirigido por

Dr. ALBERTO PARTIDA RODRÍGUEZ

CURSO 2025-2026

TÍTULO: Diseño e implementación de una solución de datos para el análisis del uso de los recursos públicos del Sistema General de Participaciones a través de las Cuentas Maestras en Colombia.

AUTOR: CRISTIAN CAMILO CHAPARRO MUÑOZ

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

DIRECTOR DEL PROYECTO: Dr. ALBERTO PARTIDA RODRÍGUEZ

FECHA: abril de 2026

RESUMEN

El presente Trabajo Fin de Máster aborda el problema de la fragmentación técnica y normativa de la información reportada por las Cuentas Maestras y las Cuentas Maestras Pagadoras del Sistema General de Participaciones (SGP) en Colombia, situación que dificulta su integración, trazabilidad y explotación analítica para fines de seguimiento y control del uso de los recursos públicos. Como respuesta, se diseñó y gestionó una base de datos relacional unificada, implementada en SQL Server, orientada a consolidar la información de cuentas, titulares, movimientos, beneficiarios y tablas de referencia bajo una estructura consistente y reutilizable. De manera complementaria, se desarrollaron procesos ETL en Python para la lectura, homologación, validación y cargue de archivos normativos en formato “.txt”, permitiendo integrar información histórica del periodo 2018-2025. Sobre esta base consolidada se implementaron dos herramientas analíticas: una orientada a la detección de posibles operaciones no permitidas y patrones anómalos en los movimientos financieros, y otra enfocada en el análisis de concentración de mercado sobre beneficiarios receptores de recursos. Los resultados obtenidos muestran que la solución desarrollada permite transformar información dispersa en un insumo estructurado, trazable y analíticamente explotable, fortaleciendo la identificación de riesgos y la generación de evidencia para procesos de monitoreo, seguimiento y control. En conjunto, el trabajo aporta una base técnica y metodológica aplicable al análisis del uso adecuado de los recursos públicos del SGP.

Palabras clave: Sistema General de Participaciones; Cuentas Maestras; base de datos relacional; ETL; detección de anomalías; concentración de mercado.

ABSTRACT

This Master's Thesis addresses the problem of the technical and regulatory fragmentation of the information reported through the Master Accounts and Master Payment Accounts of Colombia's General System of Participations (SGP), a situation that hinders data integration, traceability, and analytical use for monitoring and control of public resource management. In response, a unified relational database was designed and managed, implemented in SQL Server, to consolidate information on accounts, holders, transactions, beneficiaries, and reference tables within a consistent and reusable structure.

Additionally, ETL processes were developed in Python for reading, standardizing, validating, and loading regulatory ".txt" files, enabling the integration of historical information for the 2018-2025 period. On top of this consolidated database, two analytical tools were implemented: one aimed at detecting potentially unauthorized operations and anomalous financial transaction patterns, and another focused on market concentration analysis of resource-receiving beneficiaries.

The results show that the proposed solution makes it possible to transform dispersed information into a structured, traceable, and analytically exploitable resource, thereby strengthening risk identification and the generation of evidence for monitoring, oversight, and control processes. Overall, the thesis provides a technical and methodological foundation applicable to the analysis of the proper use of SGP public resources.

Keywords: General System of Participations; Master Accounts; relational database; ETL; anomaly detection; market concentration.

AGRADECIMIENTOS

Agradezco a mi director, el Dr. Alberto Partida Rodríguez, por su orientación, sus observaciones y su acompañamiento académico durante el desarrollo de este Trabajo Fin de Máster. Expreso también mi agradecimiento al Dr. Fernando Olivera, asesor de la Dirección General de Apoyo Fiscal del Ministerio de Hacienda y Crédito Público, por su confianza y por brindarme la oportunidad de trabajar en el tema de las Cuentas Maestras del Sistema General de Participaciones.

Dedicatoria

A mi familia, pareja y amigos, por su apoyo constante, su paciencia y su confianza durante este proceso académico y profesional.

Cita célebre

“Lo que no se define no se puede medir. Lo que no se mide no se puede mejorar. Lo que no se mejora, se degrada siempre.”

— Lord Kelvin

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Cristian Camilo Chaparro Muñoz
Título del proyecto:	Diseño e implementación de una solución de datos para el análisis del uso de los recursos públicos del Sistema General de Participaciones a través de las Cuentas Maestras en Colombia.
Directores del proyecto:	Alberto Partida Rodríguez
El proyecto se ha realizado en colaboración con una empresa o a petición de una empresa:	SI
El proyecto ha implementado un producto:	NO
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SI
Objetivo general del proyecto:	Diseñar y gestionar una base de datos relacional unificada para las Cuentas Maestras y Cuentas Maestras Pagadoras del SGP, orientada a mejorar la trazabilidad, el seguimiento y el control del uso de recursos públicos en Colombia.

Índice

RESUMEN	3
ABSTRACT	4
TABLA RESUMEN	7
Capítulo 1. RESUMEN DEL PROYECTO	13
1.1 Contexto y justificación.....	13
1.2 Planteamiento del problema	13
1.3 Objetivos del proyecto.....	14
1.4 Resultados obtenidos	14
1.5 Estructura de la memoria	15
Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE	16
2.1 Estado del arte	16
2.2 Contexto y justificación.....	19
2.3 Planteamiento del problema	24
Capítulo 3. OBJETIVOS.....	26
3.1 Objetivos generales	26
3.2 Objetivos específicos	26
3.3 Beneficios del proyecto	26
Capítulo 4. DESARROLLO DEL PROYECTO	28
4.1 Planificación del proyecto.....	28
4.2 Descripción de la solución, metodologías y herramientas empleadas.....	29
4.3 Recursos requeridos	46
4.4 Presupuesto	47
4.5 Viabilidad	48
4.6 Resultados del proyecto	50
Capítulo 5. DISCUSIÓN.....	77
5.1 Discusión sobre la utilidad de la metodología adoptada.....	77
5.2 Adaptaciones realizadas durante el desarrollo del proyecto	77
5.3 Limitaciones del estudio y de la solución desarrollada	78
5.4 Implicaciones metodológicas de los falsos positivos.....	79

5.5	Alcance e impacto del resultado obtenido	80
Capítulo 6.	CONCLUSIONES	83
6.1	Conclusiones del trabajo.....	83
6.2	Conclusiones personales.....	84
Capítulo 7.	FUTURAS LÍNEAS DE TRABAJO	86
7.1	Integración de la estructura unificada prevista en la Resolución 2394 de 2023.....	86
7.2	Fortalecimiento de la calidad del dato y ampliación de validaciones	86
7.3	Evolución de las herramientas de detección de riesgo	86
7.4	Profundización del análisis de concentración de mercado	87
7.5	Desarrollo de herramientas de visualización y uso operativo	88
7.6	Desarrollo de herramientas de consulta y exploración a diferentes niveles de granularidad	88
Capítulo 8.	UTILIZACIÓN DE IA.	89
8.1	Apoyo en la planificación del proyecto.....	89
8.2	Apoyo en la búsqueda y orientación bibliográfica	90
8.3	Apoyo en la revisión de redacción académica.....	90
8.4	Apoyo en el procesamiento de la información de contraste	91
8.5	Apoyo en la revisión técnica de scripts en Python	91
8.6	Consideraciones finales sobre el uso de IA.....	92
Capítulo 9.	OBJETIVOS DE DESARROLLO SOSTENIBLE DE LAS NACIONES UNIDAS.	93
9.1	Objetivo 2. Hambre Cero	93
9.2	Objetivo 9. Industria, innovación e infraestructura.....	93
9.3	Objetivo 10. Reducción de las desigualdades.....	94
9.4	Objetivo 11. Ciudades y comunidades sostenibles.....	94
Capítulo 10.	REFERENCIAS.....	95
Capítulo 11.	ANEXOS	97
11.1	Anexo A. Repositorio técnico del proyecto	97
11.2	Anexo B. Diccionario de datos	97
11.3	Anexo C. Catálogo y glosario de datos.....	97

Índice de Figuras

Ilustración 1 Esquema general del alcance, la solución y los resultados esperados del TFM	14
Ilustración 2 Estructura general de la memoria del TFM.....	15
Ilustración 3 Evolución Normativa del SGP y las Cuentas Maestras en Colombia.....	22
Ilustración 4 Estructura de Reporte de información de Cuentas Maestras.....	23
Ilustración 5 Estructura de Reporte de información de Cuentas Maestras Pagadoras	24
Ilustración 6 Enfoque general y metodología de la solución planteada en el TFM	30
Ilustración 7 Arquitectura de solución para integración y análisis	30
Ilustración 8 Modelo relacional simplificado de la solución	31
Ilustración 9 Homogeneización de estructuras normativas heterogéneas	33
Ilustración 10 Flujo general del proceso ETL para tablas FACT.....	38
Ilustración 11 Proceso de construcción de reglas de detección de riesgo.	39
Ilustración 12 Ejemplo aplicado de la regla A3: concentración en un único beneficiario	41
Ilustración 13 Flujo general de alertas de riesgo en el uso de los recursos.....	42
Ilustración 14 Flujo general Clústeres de concentración de mercado	46
Ilustración 15 Demostración de la Implementación del Modelo en SQL Server.	51
Ilustración 16 DER completo del modelo físico actual.....	52
Ilustración 17 Distribución de Alertas por Regla de Negocio.....	56
Ilustración 18 Evolución Temporal de Alertas de Riesgo	56
Ilustración 19 Top 15 Cuentas con Mayor Número de Alertas.....	57
Ilustración 20 Mapa de Calor: Alertas por Regla y Periodo	58
Ilustración 21 Matriz de Confusión - Reglas de Negocio.....	59
Ilustración 22 Métricas de Eficacia - Reglas de Negocio	60
Ilustración 23 Composición de Resultados - Reglas de Negocio.....	60
Ilustración 24 Método del Codo - Justificación de Clústeres de Riesgo.....	61
Ilustración 25 Distribución de Scores de Riesgo (IA).....	62
Ilustración 26 Perfiles Transaccionales vs Score de Riesgo.....	63
Ilustración 27 Matriz de Confusión IA Híbrida	64
Ilustración 28 Métricas de Eficacia - IA Híbrida.....	65
Ilustración 29 Composición de Resultados - IA Híbrida	66

Ilustración 30 Concentración de Capital	69
Ilustración 31 Top 20 Beneficiarios por Monto Total Recibido.....	70
Ilustración 32 Top 20 Beneficiarios por Cantidad de Territorios de origen	71
Ilustración 33 Probabilidad Acumulada de Montos Recibidos	72
Ilustración 34 Dispersión de Territorios de Origen por Cuartil de Monto	73
Ilustración 35 Método del Codo - Selección Óptima de Clústeres.....	74
Ilustración 36 Clústeres K-Means: Patrones Transaccionales y Riesgo.....	75

Índice de Tablas

Tabla 1 Planificación del Proyecto	28
Tabla 2 Presupuesto estimado del proyecto.....	47
Tabla 3 Tablas del modelo SQL con su total de filas	53
Tabla 4 Resultados de los procesos de ETL en las tablas FACT	54
Tabla 5 Comparación de resultados entre el modelo basado en reglas y el modelo híbrido de aprendizaje automático	66
Tabla 6 Resumen de resultados de la herramienta de concentración de mercado	75

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

El presente Trabajo Fin de Máster se desarrolla en el contexto del seguimiento al uso de los recursos públicos del Sistema General de Participaciones (SGP) en Colombia, especialmente de aquellos administrados a través de las Cuentas Maestras y las Cuentas Maestras Pagadoras. Estos instrumentos financieros surgieron como mecanismos para garantizar la destinación específica, la trazabilidad y el control de los recursos transferidos a las entidades territoriales, en el marco de la normativa aplicable.

Sin embargo, la evolución progresiva de dicha regulación dio lugar a múltiples estructuras de reporte, anexos técnicos y reglas operativas que dificultan la consolidación y el análisis integral de la información. En este escenario, el proyecto se justifica por la necesidad de transformar un conjunto de archivos dispersos y heterogéneos en una base de datos estructurada, consistente y reutilizable, que facilite tanto la explotación analítica como la identificación de riesgos en el uso de los recursos del SGP.

1.2 Planteamiento del problema

El análisis del uso adecuado de los recursos públicos administrados a través de las Cuentas Maestras y las Cuentas Maestras Pagadoras del Sistema General de Participaciones enfrenta una limitación estructural: la información reportada por las entidades financieras no se encuentra organizada bajo un modelo de datos homogéneo, integrado y directamente explotable para fines de seguimiento, control y análisis.

Aunque la normativa vigente define de manera precisa las obligaciones de reporte, en la práctica dicha información se materializa en archivos planos derivados de anexos técnicos construidos en momentos normativos distintos. Esta evolución regulatoria ha producido estructuras heterogéneas de datos, diferencias en los tipos de registro, variaciones en la organización de campos y dificultades de compatibilidad entre reportes. En consecuencia, el aprovechamiento analítico de la información depende de procesos previos de limpieza, homologación e integración que, en muchos casos, se desarrollan de forma manual o mediante soluciones parciales.

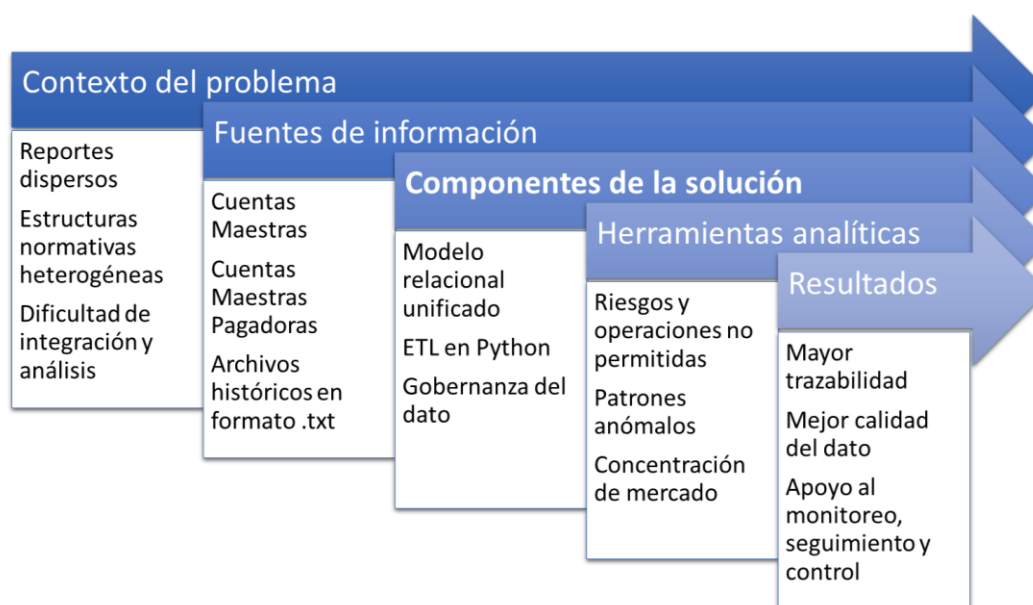
Esta fragmentación técnica reduce la capacidad institucional para consolidar información histórica, dificulta la trazabilidad de los movimientos financieros y limita la identificación temprana de posibles situaciones de riesgo, tales como operaciones no permitidas o patrones de concentración en la recepción de recursos. Por ello, el problema central de este trabajo no se restringe a una dificultad de almacenamiento, sino a la necesidad de construir una solución de datos que permita integrar, depurar, estructurar y explotar de manera consistente la información reportada. Desde esta perspectiva, la pregunta que orienta el proyecto es la siguiente: ¿cómo diseñar e implementar un modelo de datos y un conjunto de validaciones automáticas que, a partir de la información reportada por las Cuentas Maestras y las Cuentas Maestras Pagadoras del SGP, permitan consolidar y mejorar la calidad de los datos disponibles

para el análisis del uso adecuado de los recursos públicos transferidos a las entidades territoriales?

1.3 Objetivos del proyecto

El objetivo general del proyecto consiste en diseñar y gestionar una base de datos relacional unificada para las Cuentas Maestras y Cuentas Maestras Pagadoras del SGP, orientada a mejorar la trazabilidad, el seguimiento y el control del uso de recursos públicos en Colombia. Para ello, el trabajo plantea como objetivos resumidos la normalización del modelo de datos, la construcción de instrumentos de gobernanza como el diccionario y el glosario de datos, el desarrollo de procesos ETL en Python para la integración de archivos normativos y la implementación de herramientas analíticas que permitan identificar concentración de mercado y operaciones no permitidas o atípicas que puedan brindar un apoyo importante en el seguimiento, monitoreo y control al uso de los recursos públicos.

Ilustración 1 Esquema general del alcance, la solución y los resultados esperados del TFM



Elaboración Propia

1.4 Resultados obtenidos

Como resultado del proyecto, se obtuvo una solución integral orientada a transformar información dispersa y heterogénea en un insumo estructurado y analíticamente explotable. En primer lugar, se diseñó e implementó un modelo relacional unificado en SQL Server para consolidar la información de Cuentas Maestras y Cuentas Maestras Pagadoras. En segundo lugar, se desarrollaron procesos ETL reproducibles en Python para la lectura, homologación, validación y cargue de archivos normativos históricos. En tercer lugar, se incorporaron instrumentos de gobernanza del dato, como el diccionario, el catálogo y el glosario, para fortalecer la consistencia y trazabilidad de la información. Finalmente, se implementaron

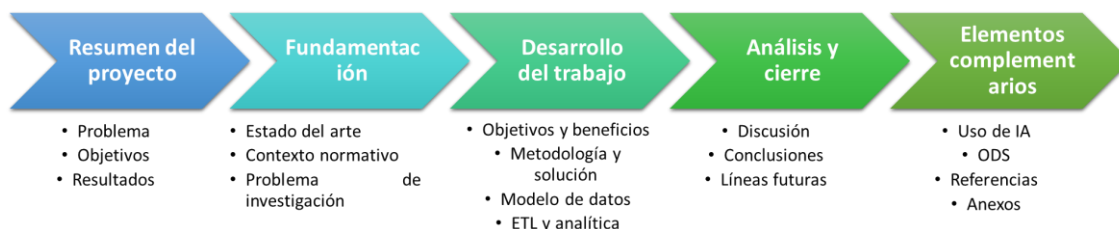
herramientas analíticas orientadas a la identificación de operaciones no permitidas, patrones anómalos y escenarios de concentración de mercado. En conjunto, estos resultados permiten mejorar la trazabilidad, la calidad del dato y el apoyo al monitoreo, seguimiento y control del uso de los recursos del SGP.

1.5 Estructura de la memoria

La memoria se organiza en once capítulos. Tras este capítulo introductorio, el segundo presenta los antecedentes, el estado del arte y el contexto normativo del problema. El tercero expone los objetivos generales, específicos y los beneficios esperados del trabajo. El cuarto desarrolla la solución propuesta, incluyendo la planificación, la metodología, las herramientas empleadas, los recursos, el presupuesto, la viabilidad y los resultados del proyecto.

El quinto recoge la discusión crítica sobre la metodología adoptada, las adaptaciones realizadas, las limitaciones y el alcance del resultado. El sexto presenta las conclusiones del trabajo y las conclusiones personales. El séptimo plantea futuras líneas de trabajo. El octavo documenta el uso de inteligencia artificial durante el desarrollo del TFM. El noveno vincula el proyecto con los Objetivos de Desarrollo Sostenible. Finalmente, los capítulos décimo y undécimo contienen las referencias bibliográficas y los anexos, respectivamente.

Ilustración 2 Estructura general de la memoria del TFM



Elaboración Propia

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

El problema por tratar en este Trabajo Fin de Máster está relacionado con diseñar e implementar un modelo de datos y un conjunto de validaciones automáticas que, a partir de la información reportada por las Cuentas Maestras y Cuentas Maestras Pagadoras del SGP, permitan consolidar y mejorar la calidad de los datos disponibles para el análisis del uso adecuado de los recursos públicos transferidos a las entidades territoriales. Es por lo anterior, que es importante manejar tres (3) enfoques de investigación para el estado del arte de este trabajo para que de esta forma se encuentren todos alineados con solucionar el problema planteado.

Estos enfoques son:

- **GovTech:** por la importancia de entender el contexto de cómo se han implementado las tecnologías del manejo de la información para el caso de información de dominio público y/o el gobierno digital.
- **Diseño y gestión de bases de datos:** para poder consolidar toda la información reportada por las Cuentas Maestras y Cuentas Maestras Pagadoras del SGP en un modelo óptimo para poder ser utilizado y consultado.
- **Detección de concentración de mercado o anomalías:** con el objetivo de identificar modelos e instrumentos utilizados en la literatura que puedan servir de apoyo en el desarrollo de herramientas que puedan aportar en el análisis del uso adecuado de los recursos del SGP.

2.1.1 GovTech

El seguimiento y la evaluación del uso de recursos públicos se dificulta cuando la información se reporta a través de múltiples estructuras, formatos y estándares, con niveles heterogéneos de calidad y consistencia, lo que limita la validación, la transparencia y la rendición de cuentas. En el marco del gobierno digital, se ha señalado que las TIC pueden contribuir a construir una cultura de transparencia cuando se orientan a ampliar el acceso a la información y fortalecer el control social sobre decisiones públicas (Bertot, Jaeger, & Grimes, 2010). De forma complementaria, se ha propuesto que estrategias de e-government apoyadas en apertura y uso de canales digitales pueden contribuir a reducir riesgos de corrupción al aumentar visibilidad y seguimiento ciudadano (Bansal, 2012).

En paralelo, el enfoque de datos abiertos advierte que la publicación de información no garantiza valor público por sí sola, ya que requiere condiciones de gobernanza, estandarización y contexto para que los usuarios puedan interpretar y reutilizar los datos adecuadamente (Janssen, Charalabidis, & Zuidervijk, 2012). Esta necesidad es especialmente relevante en sistemas de financiación intergubernamental, donde los flujos de recursos demandan criterios consistentes para asegurar trazabilidad e interpretación homogénea. En el caso colombiano, el análisis histórico del SGP muestra reformas y ajustes institucionales y operativos que inciden en la forma en que se distribuyen y gestionan los recursos, reforzando la necesidad de mecanismos

de información confiables para seguimiento y análisis (Bonet, Pérez, & Ayala, 2014). En términos operativos, los lineamientos de distribución constituyen un insumo para formalizar reglas de negocio y variables analíticas asociadas a recursos, territorios y asignaciones, facilitando su traducción a estructuras de datos consistentes para evaluación y monitoreo (Departamento Nacional de Planeación, 2023).

Adicionalmente, la literatura sobre datos abiertos señala que los beneficios esperados, como la transparencia, la innovación y la mejora de los servicios, dependen de la superación de barreras recurrentes, entre ellas la insuficiente estandarización, los problemas de calidad y la ausencia de contexto para interpretar adecuadamente los datos (Janssen, Charalabidis, & Zuidervijk, 2012). En este sentido, un enfoque GovTech aplicado al control del SGP puede entenderse como una intervención basada en datos cuya prioridad no se limita a estandarizar reportes, sino que también busca fortalecer la trazabilidad y la calidad de la información para hacerla reutilizable por analistas y órganos de control (Bertot, Jaeger, & Grimes, 2010).

2.1.2 Diseño y gestión de bases de datos

Las diferentes estructuras de reporte en datos financieros públicos incrementan el riesgo de inconsistencias y dificultan la consolidación de un repositorio único para análisis, control y trazabilidad. En este contexto, el modelo relacional proporciona una base sólida para organizar la información con integridad, mediante el uso de claves primarias y foráneas, así como de reglas de consistencia que reducen ambigüedades en la representación de entidades y relaciones (Date, 2001). La normalización, particularmente hasta tercera forma normal, se emplea como práctica para disminuir redundancias y evitar anomalías de actualización; en esta forma, cada atributo no clave depende únicamente de la clave primaria y no de otros atributos no clave, lo que resulta especialmente pertinente cuando se integran reportes periódicos provenientes de distintos actores (Elmasri & Navathe, 2007).

Para implementar prototipos y esquemas de integración de manera local, SQLite constituye una alternativa práctica como motor relacional ligero, útil para construir y validar modelos entidad-relación y ejecutar consultas SQL sin requerir infraestructura compleja (Owens, 2006). Sobre este tipo de esquema, los procesos ETL con Python se apoyan en bibliotecas como pandas para lectura de archivos, estandarización de formatos, limpieza y transformación de datos, habilitando cargas consistentes hacia tablas relacionales (McKinney, 2022). De este modo, un diseño relacional normalizado combinado con un ETL reproducible permite sentar las bases para explotar la información posteriormente en herramientas analíticas como Power BI, manteniendo coherencia entre reglas de negocio, estructura de datos y consultas.

Además, cuando la información se origina en archivos planos o reportes repetitivos, la combinación de SQL y pipelines en Python facilita la estandarización de tipos de datos, el control de duplicados y la aplicación de reglas de consistencia antes de la explotación analítica (McKinney, 2022). Desde el punto de vista del diseño relacional, estas etapas previas al cargue refuerzan el objetivo de que las tablas resultantes representen entidades estables y relaciones claras, lo cual se alinea con el propósito del modelo relacional de soportar consistencia, integridad y consultas confiables (Date, 2001). En otras palabras, el valor del esquema

normalizado no depende solo del diagrama o del SQL, sino de su integración con un proceso ETL que garantice que los datos que llegan al modelo respetan las reglas que este impone (Elmasri & Navathe, 2007).

2.1.3 Detección de concentración de mercado o anomalías

Cuando los recursos públicos se canalizan a través de entidades financieras y generan movimientos recurrentes, se vuelve necesario identificar tanto patrones de concentración de mercado como comportamientos transaccionales atípicos que puedan indicar riesgos operativos o de control en el uso de los recursos. Para el primer caso, el índice HHI se emplea como medida cuantitativa del nivel de concentración de mercado y se ha analizado su sensibilidad para interpretar cambios ante variaciones en participaciones o entrada/salida de actores, lo que facilita caracterizar escenarios de concentración en términos operativos (Brezina, Pekár, Čičková, & Reiff, 2016). Este enfoque es aplicable para evaluar concentración de mercado por beneficiarios, siempre que las participaciones se definan de forma consistente.

Además de métricas agregadas, el análisis de redes permite modelar flujos como grafos (por ejemplo, relaciones cuenta–beneficiario) y calcular propiedades estructurales para identificar nodos dominantes o patrones de conectividad relevantes para control (Hagberg, Schult, & Swart, 2008). En cuanto a detección de fraude y comportamientos anómalos, la literatura en minería de datos describe el uso de técnicas de clasificación, agrupamiento y detección de patrones para identificar transacciones sospechosas, especialmente en dominios financieros (Bhowmik, 2008). Los trabajos de revisión evidencian que las investigaciones sobre fraude financiero han utilizado de manera amplia técnicas de minería de datos, consolidando el campo como un área activa de investigación aplicada (Al-Hashedi & Magalingam, 2021) y estudios más recientes destacan la necesidad de reforzar la detección en la medida en que aumentan transacciones digitales y riesgos asociados (Lenka & Ratha, 2024).

Para escenarios con datos masivos y sin etiquetas confiables de fraude, se destacan técnicas no supervisadas como Isolation Forest, que detecta anomalías mediante el aislamiento eficiente de observaciones raras con particiones aleatorias, ofreciendo ventajas prácticas en escalabilidad (Liu, Ting, & Zhou, 2008). En un contexto como el de las CM/CMP, el soporte de estas técnicas es relevante porque permite construir “alertas” basadas en desviaciones del comportamiento considerado natural en aspectos transaccionales (por ejemplo, montos inusuales, frecuencias atípicas o patrones temporales anómalos), lo cual se alinea con la lógica de minería de datos aplicada a fraude y a detección de patrones raros en transacciones (Bhowmik, 2008).

Adicionalmente, en los problemas de detección de fraude y anomalías financieras, la literatura ha señalado que los eventos de interés suelen representar una fracción reducida del universo total de transacciones, por lo que la evaluación de los modelos no puede limitarse a métricas globales de acierto, sino que debe considerar de forma explícita la capacidad de recuperar casos potencialmente riesgosos. En este tipo de escenarios, las técnicas de minería de datos se utilizan como mecanismos de priorización para auditoría y revisión posterior, más que como sistemas de decisión automática, lo que otorga especial relevancia a métricas como la sensibilidad y a la interpretación del equilibrio entre recuperación de casos sospechosos y generación de falsas

alertas (Al-Hashedi & Magalingam, 2021). En consecuencia, para dominios financieros con fuerte desbalance entre transacciones ordinarias y eventos anómalos, resulta metodológicamente razonable valorar los modelos no solo por su precisión, sino también por su capacidad para minimizar la omisión de riesgos relevantes.

A nivel analítico, la combinación de un indicador agregado de concentración para identificar beneficiarios con alta dominancia en el mercado (Brezina, Pekár, Čičková, & Reiff, 2016) y el análisis de grafos para ubicar actores estructuralmente centrales en los flujos (Hagberg, Schult, & Swart, 2008) puede funcionar como una estrategia complementaria a los modelos de anomalías, al aportar contexto sobre “dónde se concentra” y “cómo circula” el recurso antes de priorizar casos atípicos para revisión (Liu, Ting, & Zhou, 2008).

Bajo esta lógica de priorización, la literatura sobre fraude financiero y detección de anomalías sugiere que, en contextos con fuerte desbalance de clases, la evaluación del desempeño no debe centrarse exclusivamente en la precisión, sino también en la capacidad del modelo para recuperar la mayor proporción posible de casos potencialmente riesgosos. En términos aplicados, ello otorga una relevancia particular al *recall*, especialmente cuando los modelos se utilizan como herramientas de apoyo para auditoría y revisión posterior, y no como sistemas de decisión automática. En el contexto de los recursos públicos, la omisión de un evento de riesgo puede tener consecuencias significativas sobre la trazabilidad, el control y la protección del erario, por lo que resulta metodológicamente razonable valorar modelos que prioricen la recuperación de anomalías, aun cuando ello implique una carga adicional de validación para los equipos de auditoría (Al-Hashedi & Magalingam, 2021) (Lenka & Ratha, 2024).

2.2 Contexto y justificación

2.2.1 Contexto histórico y normativo del SGP y las Cuentas Maestras

2.2.1.1 Origen constitucional y descentralización territorial

La descentralización en Colombia ha sido un proceso gradual que se formalizó y sentó sus bases con la promulgación de la Constitución Política de 1991, en la cual se definió a Colombia como República unitaria, descentralizada y con autonomía de sus entidades territoriales. Dicha autonomía se representa en una descentralización de las competencias a un nivel político, fiscal y administrativo. De esta forma, las entidades territoriales percibirían transferencias por parte del Estado para atender sus nuevas competencias y el cumplimiento de sus funciones, las cuales se establecieron en los artículos 356 y 357 de la Constitución Política, en donde se exponían dos tipos de transferencias: el Situado Fiscal, dirigido a los departamentos y distritos, y la Participación de los Municipios en los Ingresos Corrientes de la Nación – ICN, respectivamente.

2.2.1.2 Creación y evolución del Sistema General de Participaciones

Con la coyuntura de las finanzas públicas del país a finales del siglo XX y comienzos del presente, cuando se vio seriamente amenazada la sostenibilidad fiscal del Estado colombiano, se modificaron las transferencias intergubernamentales establecidas en la Constitución Política mediante la introducción del Sistema General de Participaciones – SGP (Ministerio de Hacienda

y Crédito Público, 2015). El Acto Legislativo 01 de 2001 dio origen al SGP, que derogó el sistema de transferencias establecido en la Ley 60 de 1993 y agrupó en una bolsa única lo que comprendía el Situado Fiscal y las transferencias municipales.

Como resultado de esta modificación constitucional se expidió la Ley 715 de 2001, mediante la cual se estableció la composición y distribución del SGP. Al respecto, se entiende el SGP como aquella bolsa de recursos que el Gobierno Nacional Central transfiere a las entidades territoriales, ya sean departamentos, distritos o municipios, destinados a salud, educación, agua potable y saneamiento básico, y propósito general (Bonet, Pérez, & Ayala, 2014). Posteriormente, el Acto Legislativo 04 de 2007 y la Ley 1176 de 2007 ajustaron nuevamente la estructura del sistema, separando los recursos de agua potable y saneamiento básico de la participación de propósito general y reforzando la necesidad de una estrategia de monitoreo, seguimiento y control, la cual se materializó mediante el Decreto Ley 028 de 2008.

2.2.1.3 Trazabilidad y administración separada de los recursos

Frente a la normatividad expuesta y la importancia de una adecuada administración de los recursos del SGP, cabe resaltar que el artículo 91 de la Ley 715 de 2001 estipula que estos recursos no harán unidad de caja con los demás recursos del presupuesto y que su administración deberá realizarse en cuentas separadas de los recursos de la entidad y por sectores. Esta característica refuerza la destinación específica de los recursos del SGP y la necesidad de contar con instrumentos que permitan garantizar su trazabilidad y el cumplimiento de los fines para los cuales fueron creados.

2.2.1.4 Antecedente sectorial: cuentas maestras en salud

En esta misma línea, el artículo 13 de la Ley 1122 de 2007 dispuso que los recursos del sector salud serían administrados a través de cuentas maestras. En desarrollo de esta disposición, el Ministerio de Salud y Protección Social expidió la Resolución 3042 de 2007, que reguló la organización de los Fondos de Salud de las entidades territoriales y definió las Cuentas Maestras de Salud como las cuentas destinadas a recibir la totalidad de los recursos del Sistema General de Participaciones en Salud y efectuar pagos exclusivamente mediante transferencias electrónicas, de acuerdo con los conceptos de gasto autorizados. Posteriormente, las Resoluciones 4204 de 2008, 1021 de 2009, 1127 de 2013 y 1128 de 2013 precisaron las reglas de operación de estas cuentas y la obligación de las entidades financieras de reportar periódicamente los movimientos y beneficiarios de los Fondos Locales de Salud. En conjunto, estas normas consolidaron entre 2007 y 2015 el modelo de Cuentas Maestras en el sector salud, que constituyó un antecedente y referente técnico para el diseño posterior de las Cuentas Maestras del Sistema General de Participaciones, foco principal de este trabajo.

2.2.1.5 Reglamentación de las Cuentas Maestras del SGP

Con el objetivo de contar con información de tesorería que reforzara el seguimiento y control al buen manejo de los recursos públicos trasladados por concepto del SGP, el Gobierno Nacional estableció, por medio de la Ley 1753 de 2015, en su artículo 140, que los recursos del Sistema General de Participaciones se manejarían a través de cuentas bancarias debidamente

registradas, las cuales solo aceptarían operaciones de débito por transferencia electrónica a cuentas pertenecientes a beneficiarios identificados formalmente como receptores de estos recursos.

En consecuencia, los Ministerios de Hacienda y Crédito Público, Educación Nacional y Vivienda, Ciudad y Territorio reglamentaron el producto de las Cuentas Maestras mediante diferentes resoluciones, entendiendo estas cuentas como aquellas aperturadas en entidades bancarias vigiladas por la Superintendencia Financiera de Colombia, que aceptan únicamente operaciones crédito y débito por transferencias electrónicas a beneficiarios previamente inscritos. En este marco, el Ministerio de Hacienda y Crédito Público expidió inicialmente la Resolución 3841 de 2015 para la administración de los recursos de la Asignación Especial del SGP para Resguardos Indígenas Certificados, Territorios Indígenas y asociaciones que estos conformen. Posteriormente, expidió la Resolución 4835 de 2015 para las Cuentas Maestras de propósito general, asignaciones especiales y atención integral a la primera infancia. A su vez, el Ministerio de Vivienda, Ciudad y Territorio expidió la Resolución 0857 de 2016 para agua potable y saneamiento básico, y el Ministerio de Educación Nacional la Resolución 12829 de 2017 para educación. Finalmente, la Resolución 2248 de 2018 reglamentó las Cuentas Maestras para la administración del Programa de Alimentación Escolar.

2.2.1.6 Origen de las Cuentas Maestras Pagadoras

Con la implementación de estos productos se presentaron dificultades para realizar pagos electrónicos a través del botón de pago electrónico seguro en línea – PSE en conceptos autorizados por la ley, dado que las Cuentas Maestras solo aceptaban como operaciones débito las transferencias electrónicas a beneficiarios previamente inscritos. En este contexto, se hizo necesaria la regulación de una cuenta subsidiaria y complementaria que permitiera el uso del PSE.

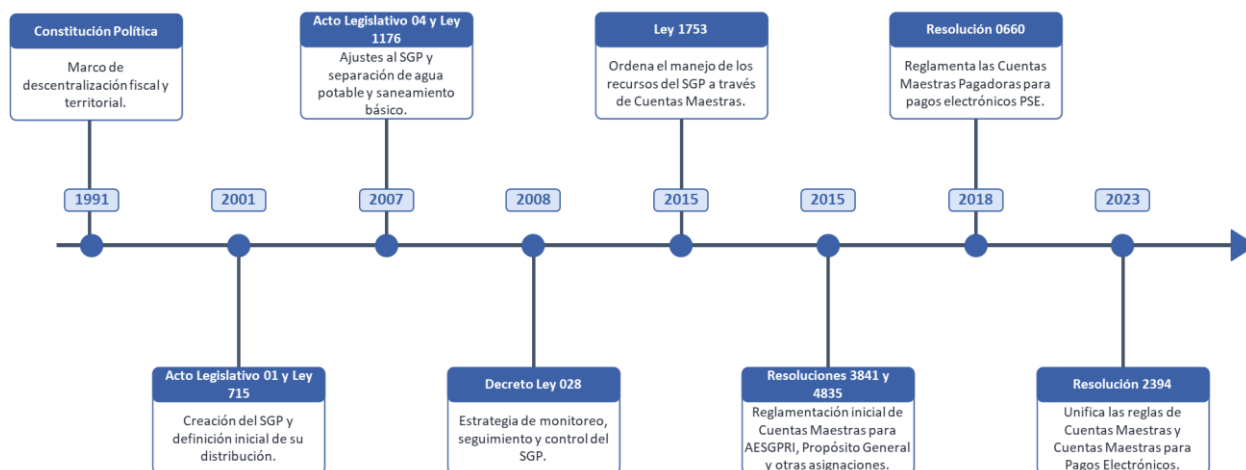
Por ello, de manera conjunta entre los Ministerios de Hacienda y Crédito Público, Educación Nacional y Vivienda, Ciudad y Territorio se expidió la Resolución 0660 del 9 de marzo de 2018, por la cual se reglamentaron las Cuentas Maestras Pagadoras. Estas surgieron como cuentas de ahorros subsidiarias y complementarias de las Cuentas Maestras para la administración de los recursos del SGP. En consecuencia, solo aceptan como operaciones crédito las transferencias provenientes de su respectiva Cuenta Maestra y como operaciones débito exclusivamente pagos electrónicos mediante PSE en los conceptos autorizados.

2.2.1.7 Unificación normativa reciente

Por último, con el propósito de ofrecer una normatividad más compacta y fácil de consultar para los usuarios y titulares de Cuentas Maestras, y aprovechando el trabajo conjunto entre los Ministerios de Hacienda y Crédito Público, Educación Nacional y Vivienda, Ciudad y Territorio, junto con la Unidad Administrativa Especial de Alimentación Escolar – Alimentos para Aprender, se expidió la Resolución 2394 de 2023. Esta resolución unifica en un solo marco regulatorio las reglas de negocio de las Cuentas Maestras y de las Cuentas Maestras para Pagos Electrónicos – PSE, deroga la regulación previa y, además, busca unificar los reportes de información y poner

en marcha el aplicativo web SISFEVAL del Ministerio de Hacienda y Crédito Público para la recepción de dichos reportes.

Ilustración 3 Evolución Normativa del SGP y las Cuentas Maestras en Colombia



Elaboración Propia

2.2.2 Reporte de información de Cuentas Maestras

Las resoluciones que reglamentan las Cuentas Maestras no solo definen las condiciones de apertura, operación, sustitución y cierre de estos productos financieros, sino que también establecen la forma en que las entidades bancarias deben reportar la información de todos los movimientos y beneficiarios de estas cuentas al Gobierno Nacional. Para ello, cada resolución incorpora un anexo técnico con la estructura detallada de los archivos planos (archivos de texto separados por punto y coma) que deben enviarse periódicamente a través de las plataformas dispuestas para tal fin, como PISIS del Ministerio de Salud y Protección Social y el Aplicativo Web SISFEVAL del Ministerio de Hacienda y Crédito Público.

En la Resolución 3841 de 2015 se adopta un anexo técnico que define el nombre de los archivos, los tipos de registro y los campos que deben reportarse para las Cuentas Maestras. Al respecto, los archivos contienen cuatro tipos de registro, cada uno con sus campos propios y su finalidad, teniendo un registro de control seguido de registros de detalle de cada Cuenta Maestra, un registro de los beneficiarios de cada cuenta y un registro para el detalle de todos los movimientos. De manera análoga, la Resolución 4835 de 2015 establece un anexo técnico para las Cuentas Maestras de Propósito General, las Asignaciones Especiales y la Asignación para la Atención Integral a la Primera Infancia, manteniendo la lógica de un registro de control seguido de registros de detalle para Cuentas Maestras, beneficiarios y movimientos, estructura que también es replicada por las Resoluciones 0857 de 2016, 12829 de 2017 y 2248 de 2018.

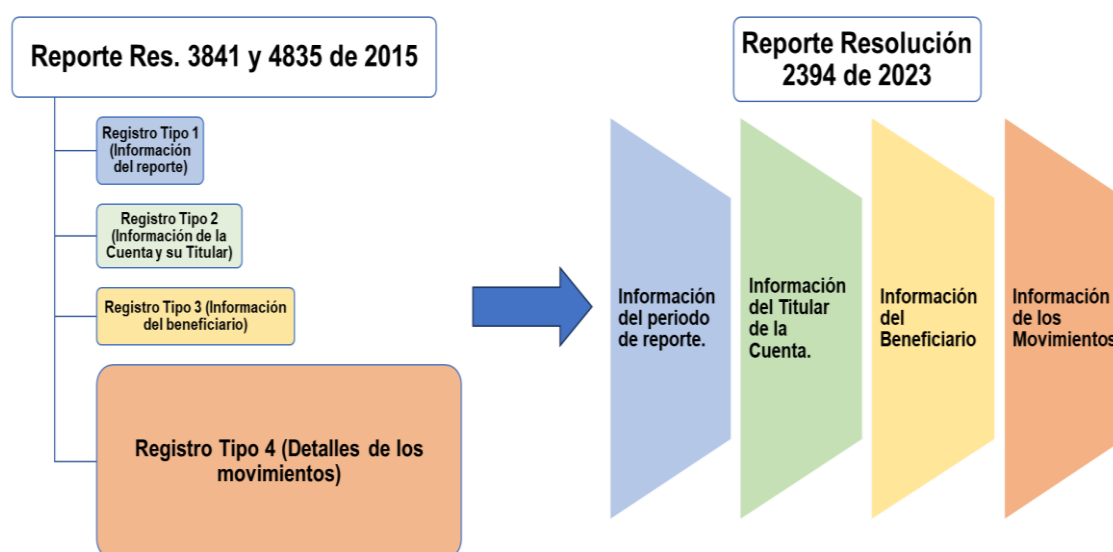
Por su parte, la Resolución 0660 de 2018 extiende este esquema al reglamentar las Cuentas Maestras Pagadoras y adopta un anexo técnico específico para el reporte de los pagos realizados mediante el botón de pago electrónico seguro en línea (PSE), pasando de tener cuatro tipos de

registro a contar con solo tres: un registro de control seguido de registros de detalle de cada Cuenta Maestra Pagadora y un registro para el detalle de todos los movimientos.

Por último, con la consolidación de toda esta normatividad en la Resolución 2394 de 2023, se planteó un anexo técnico que, de igual forma, compilara todos los reportes en uno solo y estableciera un único tipo de registro que contenga la información que en otros reportes se manejaba en diferentes tipos de registro.

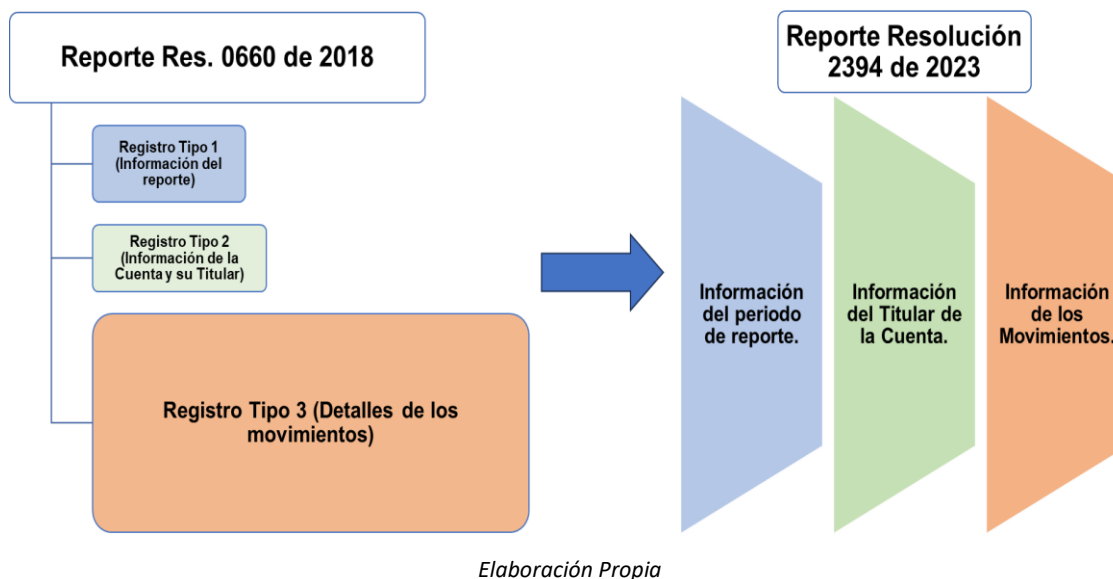
En conjunto, estos anexos técnicos, con sus diferentes estructuras, constituyen el insumo normativo directo para el diseño del modelo de base de datos y de los procesos de identificación de riesgos desarrollados en este trabajo.

Ilustración 4 Estructura de Reporte de información de Cuentas Maestras



Elaboración Propia

Ilustración 5 Estructura de Reporte de información de Cuentas Maestras Pagadoras



2.3 Planteamiento del problema

El Sistema General de Participaciones (SGP) canaliza una parte relevante de los recursos públicos de inversión en Colombia hacia las entidades territoriales, en los términos de los artículos 356 y 357 de la Constitución Política de 1991 (Asamblea Nacional Constituyente, 1991). Estos recursos financian sectores como agua potable y saneamiento básico, educación, propósito general y otras asignaciones especiales, y se administran a través de productos financieros denominados Cuentas Maestras y Cuentas Maestras Pagadoras.

La regulación vigente define las condiciones de apertura, operación y cierre de estos productos, así como la estructura de los archivos de reporte que las entidades financieras deben remitir periódicamente al Gobierno nacional. Sin embargo, esta normativa se ha desarrollado de manera progresiva mediante resoluciones sectoriales y anexos técnicos independientes, dando lugar a múltiples estructuras de datos, nomenclaturas y tipos de registro que no siempre resultan plenamente compatibles entre sí.

En la práctica, esta fragmentación normativa y técnica implica que la información reportada por los establecimientos bancarios llegue dispersa en diferentes archivos de texto y con esquemas heterogéneos. Como consecuencia, las entidades responsables del seguimiento y control del SGP deben invertir esfuerzos significativos en procesos manuales o semimanuales de limpieza, homologación e integración antes de poder realizar análisis sustantivos sobre el uso de los recursos. Esto limita la detección oportuna de anomalías y dificulta la evaluación sistemática del uso adecuado de los recursos a gran escala.

Aunque la normativa establece qué información debe reportarse y con qué periodicidad, no existe un modelo de datos consolidado ni un conjunto de herramientas estándar para integrar, depurar y explotar esa información de forma uniforme. En su lugar, coexisten soluciones ad hoc, como hojas de cálculo, consultas aisladas y scripts dispersos, que resuelven necesidades

puntuales, pero no garantizan trazabilidad, comparabilidad entre sectores ni reutilización de reglas de negocio a lo largo del tiempo.

De ello surge la necesidad de diseñar un modelo de base de datos que consolide las diferentes estructuras de reporte previstas en la normativa vigente y permita almacenar de manera integrada la información de cuentas, titulares, movimientos y beneficiarios, así como desarrollar herramientas que apoyen la identificación de posibles situaciones de riesgo en el uso de los recursos. Dicho modelo y dichas herramientas constituyen un primer paso para fortalecer el análisis del uso adecuado de los recursos del SGP, al facilitar la detección de inconsistencias y la generación de insumos confiables para la toma de decisiones.

En síntesis, el problema que aborda este Trabajo Fin de Máster puede formularse así: **¿cómo diseñar e implementar un modelo de datos y un conjunto de validaciones automáticas que, a partir de la información reportada por las Cuentas Maestras y Cuentas Maestras Pagadoras del SGP, permitan consolidar y mejorar la calidad de los datos disponibles para el análisis del uso adecuado de los recursos públicos transferidos a las entidades territoriales?**

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

El objetivo general del presente trabajo es diseñar y gestionar una base de datos relacional unificada para las Cuentas Maestras y Cuentas Maestras Pagadoras del SGP, orientada a mejorar la trazabilidad, el seguimiento y el control del uso de recursos públicos en Colombia.

3.2 Objetivos específicos

- Diseñar y normalizar el modelo entidad–relación para CM y CMP (cuentas, titulares, beneficiarios, movimientos y tablas de referencia), definiendo claves primarias y foráneas y su implementación en SQL Server.
- Construir y documentar el diccionario de datos y el glosario de términos del modelo unificado, incluyendo dominios, definiciones de negocio y reglas base de validación para asegurar interpretación consistente.
- Implementar y documentar un proceso ETL en Python que ingeste archivos “.txt” separados por “;”, realice homologación, limpieza y validaciones, y cargue la información en la base de datos relacional, generando trazabilidad de ejecución (registros procesados, rechazados y motivos de error).
- Diseñar, aplicar y evaluar una herramienta analítica en Python para identificar escenarios de concentración (por beneficiario y/o territorio) sobre los recursos administrados, mediante métricas e indicadores que permitan priorizar casos relevantes y sustentar su interpretación con evidencia.
- Diseñar e implementar una herramienta en Python para la detección explicable de operaciones no permitidas o patrones fuera de norma, basada en reglas verificables sobre los movimientos, generando un reporte auditable que incluya el criterio activado (regla) y la evidencia asociada (campos del registro y contexto).

3.3 Beneficios del proyecto

El desarrollo de este Trabajo Fin de Máster guarda una relación directa con las labores que se realizan en el marco de la Estrategia de Monitoreo, Seguimiento y Control a los recursos del Sistema General de Participaciones (SGP), reglamentada por el Decreto Ley 028 de 2008. En este contexto, la solución propuesta aporta beneficios relevantes desde el punto de vista técnico, metodológico y operativo, al fortalecer la organización, trazabilidad y explotación analítica de la información reportada por las Cuentas Maestras y las Cuentas Maestras Pagadoras.

En primer lugar, el proyecto permite la homogeneización de diferentes estructuras normativas de reporte en dos estructuras consistentes, correspondientes a Cuentas Maestras y Cuentas Maestras Pagadoras. Esta estandarización facilita la integración de información históricamente dispersa y reduce la complejidad derivada de la coexistencia de múltiples anexos técnicos y esquemas de datos.

En segundo lugar, la consolidación de la información en una base de datos relacional unificada favorece una mayor trazabilidad y transparencia en el seguimiento del uso de los recursos del SGP. Al contar con una estructura clara, organizada y reutilizable, se facilita la consulta de la información, su validación y su explotación posterior con fines de análisis y control.

Del mismo modo, el desarrollo del diccionario de datos y del glosario de términos aporta una base de gobernanza escalable, que facilita el mantenimiento del modelo y su adaptación frente a cambios normativos futuros. Estos instrumentos contribuyen a preservar la consistencia conceptual y técnica del sistema, y fortalecen la interpretación homogénea de la información por parte de distintos usuarios.

Por su parte, la solución mejora la eficiencia operativa al automatizar el cargue, la homologación y la validación de los archivos de reporte de información. Esto reduce la dependencia de tareas manuales repetitivas y permite disponer de una base más confiable para los procesos de seguimiento y análisis.

Adicionalmente, el diseño modular del modelo y de los procesos desarrollados favorece la interoperabilidad futura, en la medida en que facilita la incorporación de nuevas fuentes de información o la ampliación de las estructuras existentes sin necesidad de rediseñar completamente la solución.

Por último, el proyecto permite generar alertas tempranas sobre posibles situaciones de riesgo en el uso de los recursos públicos de destinación específica, particularmente en aspectos como concentración de mercado y operaciones no permitidas o atípicas. Este componente analítico amplía el valor de la solución, al no limitarse a la organización de datos, sino convertirlos en insumos útiles para la priorización de revisiones y el fortalecimiento del control.

En conjunto, estos beneficios pueden contribuir a sustentar hallazgos dentro del marco de la Estrategia de Monitoreo, Seguimiento y Control a los recursos del SGP, al facilitar la identificación de eventos de riesgo en su uso o administración (Presidencia de la República de Colombia, 2008). En este sentido, la solución desarrollada constituye un apoyo técnico que puede fortalecer la adopción de Medidas Preventivas¹ o Correctivas² por parte de las entidades competentes, y aportar al mejoramiento de la administración pública a nivel territorial.

¹ Medidas Preventivas: Plan de desempeño en el cual se obliga a desarrollar las actividades orientadas a mitigar o eliminar los eventos de riesgo en los términos y plazos que allí se fijen. (Presidencia de la República de Colombia, 2008)

² Medidas Correctivas: Suspensión de giros a la entidad territorial, es la medida por medio de la cual se suspende el giro sectorial o general de recursos a la entidad territorial, sin que se afecte el derecho jurídico de la misma a participar en los recursos del Sistema General de Participaciones, ni la continuidad en la prestación del servicio, conforme lo determine el reglamento. (Presidencia de la República de Colombia, 2008)

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

Tabla 1 Planificación del Proyecto

Semana	Fase / Actividad	Entregable	Esfuerzo (horas)
1	Estudio del arte: normativa SGP (CM/CMP), anexos técnicos, Diseño de bases de datos y modelos de detección de anomalías	Capítulo 2: Antecedentes y Estado del arte (borrador)	20
2	Homogeneizar las estructuras en 2 modelos (CM y CMP)	Anexo: Matriz de homologación (campo_origen → campo_destino + reglas)	14
3	Diseño del sistema (modelo conceptual y lógico)	Capítulo 4: Metodología y diseño – modelo conceptual y lógico (borrador)	10
4	Diseño del modelo de datos (SQL Server): DDL v1, PK/FK/índices/checks	Capítulo 4: Metodología y diseño – modelo de datos (borrador) + Script DDL v1	14
5	Diccionario de datos v1 (técnico)	Anexo A2: Diccionario de datos (v1)	10
6	Glosario de datos v1 (negocio) + catálogo de reglas	Anexo A3: Glosario de datos (v1) + catálogos de reglas	10
7	Análisis técnico de interfaces (archivos .txt normativos)	Capítulo 4: Metodología – especificaciones de ingesta y validación (ETL)	12
8	ETL en Python – lectura y limpieza	Archivo .py con el código del proceso ETL para CM y CMP.	16
9	Validaciones y carga a modelo operativo	Código de validaciones + MERGE a productivo + LOG_CARGA/LOG_ERROR	14

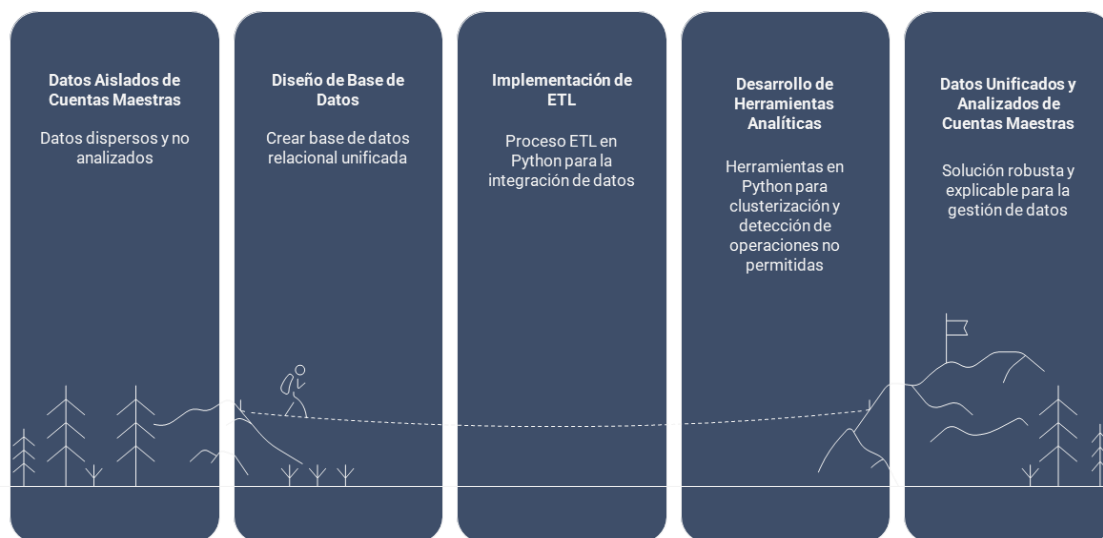
10	Modelo de identificación de Clústeres de concentración de mercado – prototipo	Archivo .py con el código del modelo de identificación de clústeres de concentración de mercado (prototipo) + Anexo A4: Metodología de clustering (borrador)	16
11	Clústeres – versión final	CSV/Excel de resultados + Capítulo 5: Resultados – Modelo de Clústeres (borrador)	12
12	Operaciones no permitidas – prototipo explicable	Archivo .py con el código del modelo de identificación de Operaciones no permitidas.	16
13	Modificaciones y correcciones (con director)	Cap. 2, Cap. 4, Cap. 5 – versión final + Anexos consolidados + Resumen/Abstract	8
14	Modificaciones y correcciones (cierre)	Memoria TFM lista para entrega y diapositivas de presentación del TFM.	8
TOTAL			180

4.2 Descripción de la solución, metodologías y herramientas empleadas

4.2.1 Enfoque general de la solución

El proyecto adopta un enfoque aplicado con soporte metodológico científico. De forma general, la solución se concibe como un proceso que parte de información dispersa y culmina en una base estructurada y analíticamente explotable.

Ilustración 6 Enfoque general y metodología de la solución planteada en el TFM



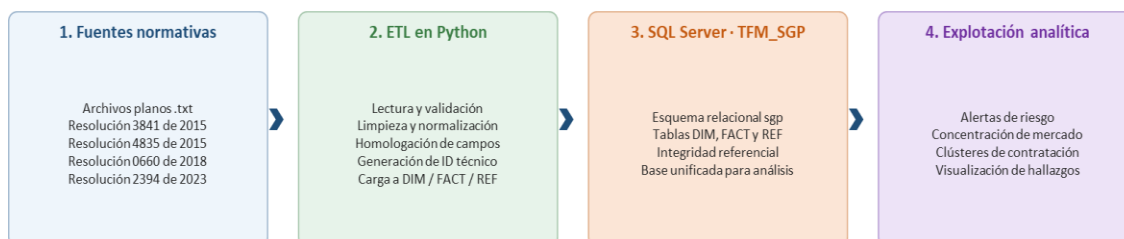
Elaboración Propia

La solución se estructura en tres componentes principales:

1. **Diseñar y gestionar** un modelo de datos relacional unificado para Cuentas Maestras (CM) y Cuentas Maestras Pagadoras (CMP), normalizado y documentado mediante diccionario y catálogo de datos.
2. **Implementar** un proceso ETL en Python orientado a la ingesta, homologación, validación y carga de los archivos normativos de reporte.
3. **Desarrollar** un conjunto de herramientas analíticas en Python destinado a apoyar la identificación de posibles situaciones de riesgo y de concentración en el uso de los recursos.

A partir de este enfoque general, la arquitectura funcional de la solución se organiza en cuatro bloques: fuentes normativas, proceso ETL, base de datos relacional y explotación analítica

Ilustración 7 Arquitectura de solución para integración y análisis



Elaboración Propia

4.2.2 Modelo de datos relacional propuesto

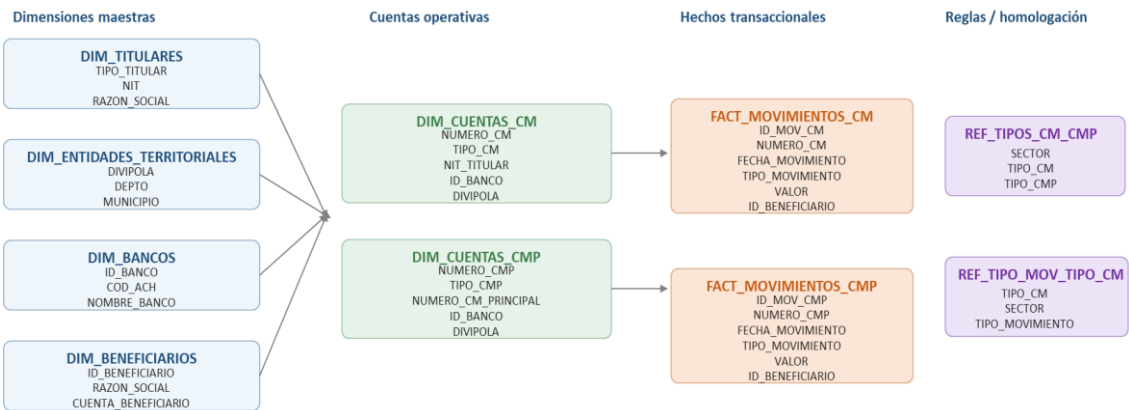
El modelo de datos se construyó bajo un enfoque relacional que separa la información en tres grupos: dimensiones (maestros), tablas de hechos (movimientos) y tablas de referencia

(correspondencias). Esta organización permite consolidar la información de Cuentas Maestras (CM) y Cuentas Maestras Pagadoras (CMP) con integridad, trazabilidad y consistencia para análisis posteriores.

Desde el punto de vista metodológico, el diseño se apoya en dos principios:

1. Normalización mediante formas normales: el modelo busca reducir redundancia y anomalías de actualización aplicando criterios de 2FN y 3FN. En particular, se separan entidades repetibles (por ejemplo, bancos, tipos y entidades territoriales) en tablas maestras, y se evita que atributos dependientes de un tercero o de una clasificación se almacenen de forma duplicada en múltiples tablas operativas. Con ello se mejora la calidad del dato, se facilita el mantenimiento y se refuerza la integridad referencial mediante PK/FK.
2. Diseño guiado por normativa y anexos técnicos: la selección de tablas y campos se realizó a partir del análisis del marco normativo del SGP y de la normativa específica de CM/CMP, utilizando los anexos técnicos como fuente para comprender la estructura real de los reportes “.txt”, sus codificaciones y sus reglas de operación. Este enfoque permite que el modelo propuesto sea consistente con la información efectivamente reportada y que sirva como base para validaciones y controles alineados con la regulación.

Ilustración 8 Modelo relacional simplificado de la solución



Elaboración Propia

a) Dimensiones principales (maestros)³

³ Nota metodológica: la separación de estos catálogos en tablas maestras permite que las clasificaciones se mantengan consistentes y actualizables sin modificar registros históricos, reduciendo redundancia y mejorando la gobernanza del dato.

- **DIM_TITULARES:** Su objetivo es enlistar quién es titular de Cuentas Maestras en las que se administran los recursos del SGP. Incluye campos como TIPO_TITULAR, NIT, DV, RAZON_SOCIAL, DIVIPOLA, COD_DEPARTAMENTO, COD_MUNICIPIO, entre otros.
- **DIM_ENTIDADES_TERRITORIALES:** Se usa para análisis territoriales y cruces geográficos, complementa la perspectiva territorial incorporando variables de localización como LATITUD, LONGITUD y UBICACIÓN, además de los identificadores DIVIPOLA, COD_DEPARTAMENTO, COD_MUNICIPIO.
- **DIM_CUENTAS_CM:** Esta tabla representa el inventario unificado de CM y permite relacionarlas con sus movimientos, describe las cuentas maestras con sus atributos de identificación, clasificación y vinculación con titular y banco. Incluye, entre otros, DIVIPOLA, NIT_TITULAR, TIPO_TITULAR, SECTOR, RUBRO, TIPO_CM, NUMERO_CM, TIPO_CUENTA, NIT_BANCO, CODIGO_ACH_BANCO.
- **DIM_CUENTAS_CMP:** Describe las cuentas pagadoras (CMP) y su relación con la cuenta principal. Incluye TIPO_CMP, NUMERO_CMP, NUMERO_CM_PRINCIPAL, además de campos territoriales, del titular y del banco como DIVIPOLA, NIT_TITULAR, TIPO_TITULAR, SECTOR, RUBRO, TIPO_CM, TIPO_CUENTA, NIT_BANCO, CODIGO_ACH_BANCO.
- **DIM_BENEFICIARIOS:** Permite estandarizar terceros y habilitar análisis de concentración y riesgos por receptor. Adicionalmente, registra los beneficiarios inscritos y la información bancaria asociada para pagos/transferencias. Incluye TIPO_ID_BENEFICIARIO, ID_BENEFICIARIO, DV, RAZON_SOCIAL, CODIGO_BANCO, CUENTA_BENEFICIARIO, TIPO_CUENTA_BENEFICIARIO.
- **DIM_BANCOS:** Catálogo de entidades financieras, con campos como COD_SUPERFINANCIERA, COD_ACH, NOMBRE_BANCO, NIT_BANCO, NIT_DV, ESTADO_CM.
- **DIM_TIPO_CM:** Catálogo de todos los tipos de Cuenta Maestra habilitados en la normatividad, cuyo objetivo es identificar los recursos que se administran en la cuenta.
- **DIM_TIPO_CMP:** Catálogo de todos los tipos de Cuenta Maestra Pagadora habilitados en la normatividad, cuyo objetivo es identificar los recursos que se administran en la cuenta.

b) Tablas de hechos (movimientos)

- **FACT_MOVIMIENTOS_CM:** registra movimientos asociados a las cuentas maestras, con campos como TIPO_CM, NUMERO_CM, FECHA_MOVIMIENTO, TIPO_MOVIMIENTO, DESCRIPCION_MOVIMIENTO, VALOR y, cuando aplique, datos del beneficiario (TIPO_ID_BENEFICIARIO, ID_BENEFICIARIO, DV, RAZON_SOCIAL) y de la cuenta destino (CODIGO_BANCO, CUENTA_BENEFICIARIO, TIPO_CUENTA_BENEFICIARIO). Es la base transaccional para trazabilidad y análisis de comportamiento.
- **FACT_MOVIMIENTOS_CMP:** registra movimientos de cuentas pagadoras, con campos como TIPO_CMP, NUMERO_CMP, FECHA_MOVIMIENTO, TIPO_MOVIMIENTO, DESCRIPCION_MOVIMIENTO, VALOR, identificadores del beneficiario y campos

operativos adicionales como CODIGO_SERVICIO y CUS. Es la base para el control de operaciones autorizadas y detección de posibles situaciones de riesgo.

c) Catálogos y referencias (estandarización y reglas)

- **REF_TIPOS_CM_CMP:** Matriz de correspondencia entre SECTOR, TIPO_CM, ASIGNACION y su TIPO_CMP, incluyendo NOMBRE_SECTOR.
- **REF_TIPO_MOV_TIPO_CM:** Relaciona TIPO_CM + SECTOR con TIPO_MOVIMIENTO, lo que habilita reglas de movimientos permitidos/no permitidos por tipo de CM.

El desarrollo de este modelo se basó en una estrategia de **homologación progresiva** que permite convertir múltiples estructuras de reporte, que es la situación que se presenta actualmente, en un esquema relacional consistente:

1. **Inventario de estructuras y campos:** se identificó la totalidad de campos disponibles por tipo de reporte y se clasificaron según su función.
2. **Homologación hacia el modelo objetivo:** se definió una correspondencia para consolidar la información en las tablas finales, principalmente en **DIM_CUENTAS_CM**, **DIM_CUENTAS_CMP**, **DIM_TITULARES**, **DIM_BENEFICIARIOS** y los hechos **FACT_MOVIMIENTOS_CM/CMP**.
3. **Normalización y catálogos:** se separaron entidades repetibles (bancos, tipos, territorios) en dimensiones y catálogos, reduciendo redundancia y facilitando validaciones.
4. **Trazabilidad:** se incorporaron tablas de referencia (REF_*) para soportar reglas de consistencia y clasificación, necesarias para las herramientas analíticas posteriores.

Ilustración 9 Homogeneización de estructuras normativas heterogéneas



Elaboración Propia

En conjunto, el modelo refleja la información reportada en los anexos técnicos y la traduce a una estructura normalizada, preparada para controles automáticos y análisis reproducibles sobre el uso de recursos del SGP.

Por último, aunque se utiliza la nomenclatura “DIM/FACT/REF” por claridad y por su utilidad analítica, el modelo propuesto mantiene un enfoque relacional normalizado. En particular, las entidades maestras se separan para reducir redundancia y dependencias transitivas (2FN/3FN),

mientras que las tablas de hechos concentran eventos transaccionales vinculados por claves a dichas entidades. De este modo, se preserva integridad y consistencia sin depender de un modelo puramente dimensional.

4.2.3 Diccionario de datos y catálogo de datos

Con el fin de garantizar consistencia, trazabilidad y reutilización de la información consolidada de Cuentas Maestras (CM) y Cuentas Maestras Pagadoras (CMP), se desarrollaron dos artefactos de gobierno del dato: el Diccionario de Datos y el Catálogo de Datos. Ambos complementan el modelo relacional propuesto, ya que formalizan definiciones, dominios y reglas que soportan la ingesta (ETL), la validación y el análisis posterior. Este enfoque se alinea con las prácticas de gestión de metadatos y gobierno de datos, donde la documentación de estructuras, definiciones y reglas operativas se considera esencial para asegurar calidad, control y uso consistente a lo largo del ciclo de vida del dato (DAMA International, 2017).

4.2.3.1 Diccionario de datos

El Diccionario de Datos se construyó como una especificación técnica detallada a nivel tabla–campo, orientada a estandarizar la interpretación de la información y a servir como referencia para el diseño e implementación de la base de datos. En este documento se consolidaron los nombres finales de las tablas y sus campos, garantizando consistencia con el modelo y con los procesos de carga. Para cada campo se documentaron, entre otros, los siguientes elementos: descripción funcional, tipo de dato, longitud y/o escala (cuando aplica), obligatoriedad (permitir o no valores nulos), participación en clave primaria (PK) o clave foránea (FK), y observaciones relevantes para su uso.

Con el propósito de hacer operativa esta documentación dentro del ciclo de vida del dato, se incorporaron metadatos complementarios que conectan las definiciones con el control de calidad y la implementación:

- **Dominios (valores permitidos):** catálogos de valores permitidos, con su fuente y vigencia cuando aplica, para estandarizar campos codificados y reducir ambigüedad. Para dominios masivos (p. ej., “DIVIPOLA” o entidades financieras), la estandarización se gestiona mediante tablas de referencia (“DIM”) y su fuente oficial, evitando duplicar listados extensos en el documento.
- **Glosario de términos:** términos clave del contexto CM/CMP (definiciones, sinónimos, campos asociados y ejemplos) para armonizar el lenguaje normativo, operacional y técnico.
- **Reglas de validación:** conjunto inicial de reglas con identificador, tipo (unicidad, integridad referencial, consistencia), severidad (error/alerta), mecanismo de implementación y mensaje estandarizado, como insumo directo para el módulo de validaciones.

En términos prácticos, el diccionario funciona como una fuente de referencia para que el ETL y la validación automaticen controles (por ejemplo, tipos de dato, obligatoriedad, unicidad o consistencia de códigos) y para que el consumo analítico mediante consultas SQL y procesos

posteriores mantenga un significado estable de cada variable, reduciendo interpretaciones divergentes.

Adicionalmente, en las tablas de hechos de movimientos se definió un identificador técnico: el campo “ID_MOV_*”, el cual se genera durante el proceso ETL de forma determinística a partir de la combinación “NOMBRE_ARCHIVO” + “CONSECUTIVO_FILA”, garantizando unicidad por registro reportado y evitando duplicidades derivadas de recargas del mismo archivo.

4.2.3.2 Catálogo de datos

El Catálogo de Datos se elaboró como un documento de nivel superior (a nivel de tabla) enfocado en describir cómo se administra y utiliza cada conjunto de datos dentro de la solución. Para cada tabla del modelo se consignaron aspectos que facilitan su operación y consumo controlado: descripción del propósito, origen del dato, frecuencia de actualización, granularidad, criterio de identificación/unicidad, responsables sugeridos (data owner), sensibilidad y dependencias/lineage (flujo desde la fuente → transformación → tabla destino).

Este catálogo cumple dos propósitos: primero, soporta la operación del dato al clarificar periodicidad y responsabilidades, lo cual es relevante en el manejo de información asociada a cuentas, entidades y movimientos. Segundo, habilita el análisis al formalizar la granularidad y el rol de cada tabla dentro del modelo, evitando que los indicadores se construyan sobre interpretaciones inconsistentes. En línea con DAMA-DMBOK, el catálogo actúa como un mecanismo práctico de gestión de metadatos que facilita el gobierno, el control y la reutilización de la información por diferentes perfiles (técnicos y de negocio) (DAMA International, 2017).

Finalmente, el diccionario y el catálogo se gestionan como artefactos versionados, de manera que puedan actualizarse cuando cambien estructuras, dominios o reglas, garantizando coherencia entre el modelo físico, el proceso ETL y las validaciones.

4.2.4 Implementación del modelo físico en SQL Server

La solución contempla la implementación del modelo físico en SQL Server mediante scripts DDL destinados a la creación de las tablas de dimensiones (DIM), hechos (FACT) y referencias (REF) definidas en el modelo relacional. Estos scripts incorporan la definición de tipos de dato, claves primarias y foráneas, restricciones de nulabilidad y reglas básicas de consistencia, en coherencia con el Diccionario de Datos y el Catálogo de Datos elaborados. De forma complementaria, el diseño considera índices mínimos asociados a claves primarias y foráneas para favorecer la integridad y el desempeño de las consultas.

Para preservar códigos con ceros a la izquierda, como los identificadores territoriales (DIVIPOLA) y algunos códigos bancarios, el modelo prioriza el uso de tipos de dato de texto cuando ello resulta necesario, evitando pérdidas de formato derivadas de conversiones numéricas.

En cuanto a la incorporación de información, la solución diferencia dos estrategias. Por una parte, las tablas dimensionales paramétricas y las tablas de referencia se gestionan como datos semilla, dado que su actualización depende principalmente de cambios normativos o ajustes puntuales del contexto regulatorio. Por otra, las tablas asociadas a información operacional

periódica, especialmente las tablas de hechos se alimentan mediante procesos ETL desarrollados en Python. Para determinadas dimensiones maestras con cambios controlados, se contempla además un mini-ETL que permite realizar cargas iniciales y actualizaciones puntuales de manera mantenible.

4.2.5 Proceso ETL en Python

La carga de información hacia el modelo relacional se implementa mediante procesos ETL desarrollados en Python. El diseño contempla dos niveles: (i) un **mini-ETL** para tablas maestras y de referencia (DIM/REF), orientado a cargas iniciales y/o actualizaciones; y (ii) un ETL mensual para las tablas de hechos (FACT), orientado a ingerir los movimientos reportados en archivos “.txt” y transformarlos a una estructura unificada y validable.

- **4.2.5.1 Mini-ETL para tablas DIM/REF (cargas iniciales y actualizaciones puntuales)**

El mini-ETL se orienta al mantenimiento controlado de las dimensiones y tablas de referencia que soportan la integridad del modelo. Su lógica funcional comprende la extracción desde fuentes oficiales o archivos de entrada, la estandarización de tipos y formatos, la validación de unicidad y consistencia, y la carga o actualización controlada mediante mecanismos de inserción y actualización. Con ello se busca que las tablas maestras actúen como una base estable para los procesos periódicos de ingesta y validación.

De forma general, el mini-ETL sigue cuatro reglas operativas:

- **“Extraer”** desde la fuente oficial o archivo de entrada correspondiente (listado de Entidades Territoriales o registros de cuentas/titulares).
- **“Estandarizar”** tipos y formatos (tipos de dato, limpieza de espacios, normalización de claves y consistencia de codificaciones).
- **“Validar”** unicidad y consistencia (por ejemplo, coherencia “DIVIPOLA – departamento – municipio” cuando aplique, control de duplicados en códigos y verificación de dominios).
- **“Cargar/actualizar”** mediante inserción o actualización controlada (upsert), registrando trazabilidad de ejecución (archivo, fecha, registros insertados/actualizados y registros rechazados).

El resultado de este mini-ETL es un conjunto de tablas DIM/REF estable y versionado, que reduce errores de carga y facilita que las validaciones del ETL mensual se basen en “fuentes de verdad” consistentes (dominios y catálogos).

- **4.2.5.2 ETL mensual para tablas FACT (movimientos CM y CMP)**

El ETL periódico se orienta a integrar los movimientos reportados por las entidades financieras y a transformarlos en registros comparables dentro de las tablas de hechos del modelo. Para ello, el diseño considera la coexistencia de estructuras normativas históricas y estructuras más recientes, de manera que el proceso pueda identificar el tipo de archivo, extraer la información transaccional relevante, homologar campos, validar consistencias mínimas y conservar trazabilidad suficiente entre el dato cargado y su archivo de origen.

Para cubrir las variaciones normativas, el ETL soporta dos esquemas de reporte:

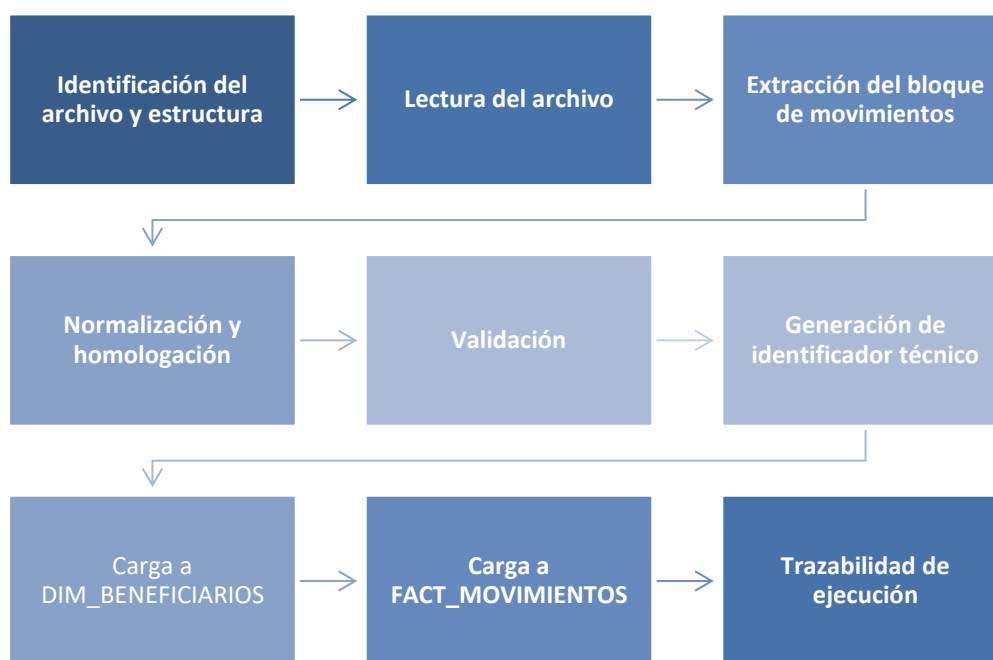
- **Estructuras históricas por tipos de registro:** para CM (estructura 2005/2015) y CMP (estructura 2018), donde la información se organiza por “tipos” (control/periodo, cuenta/titular, beneficiario y movimientos).
- **Estructura 2023 por bloques de columnas:** donde la misma información se encuentra organizada en secciones de columnas dentro de un mismo registro, pero conservando bloques equivalentes a los anteriores (periodo, cuenta/titular, beneficiario si aplica y movimientos).

En ambos casos, el ETL se concentra en extraer y consolidar la información correspondiente a **movimientos**, dado que esta es la que alimenta las tablas FACT. El flujo general se resume así:

1. **Identificación del archivo y estructura:** se determina si el archivo corresponde a CM o CMP y si su organización es por “tipos de registro” o por “bloques de columnas”.
2. **Lectura del archivo:** se lee el archivo “.txt” con separador “;”, preservando los códigos como texto cuando aplica para evitar pérdida de información.
3. **Extracción del bloque de movimientos:**
 - En estructura por tipos: se recorre el archivo y se captura la información de movimientos del tipo correspondiente (CM: tipo de movimiento; CMP: tipo de movimiento según estructura).
 - En estructura 2023: se toma el bloque de columnas equivalente a movimientos y se construyen registros transaccionales.
4. **Normalización y homologación:** se estandarizan formatos (fechas y valores), se preservan códigos, se mapean campos hacia el modelo unificado (según la matriz de homologación) y por último se genera un dataframe con la información de los beneficiarios identificados en los movimientos.
5. **Validación:** se aplican reglas mínimas de calidad y consistencia (obligatoriedad, dominios, consistencias básicas), registrando errores y alertas con trazabilidad.
6. **Generación de identificador técnico:** para asegurar unicidad por registro reportado, se define el campo “ID_MOV_*”, generado de forma determinística a partir de la combinación “NOMBRE_ARCHIVO + CONSECUTIVO_FILA”.
7. **Carga a DIM_BENEFICIARIOS:** Se carga el dataframe de los beneficiarios identificados en la tabla de dimensión correspondiente.
8. **Carga a FACT:** se insertan los movimientos válidos en la tabla de hechos correspondiente, conservando evidencia suficiente para auditoría.
9. **Trazabilidad de ejecución:** se registra un resumen del proceso (archivo, periodo, total de movimientos detectados, cargados y rechazados) y se genera un reporte de errores para revisión.

Esta estrategia permite integrar reportes de diferentes años y estructuras sin perder consistencia, manteniendo un proceso reproducible donde cada movimiento cargado puede rastrearse hasta su origen (archivo y registro), y donde las reglas de validación aportan evidencia para el análisis posterior y para auditorías más profundas.

Ilustración 10 Flujo general del proceso ETL para tablas FACT



Elaboración Propia

4.2.6 Herramientas analíticas en Python

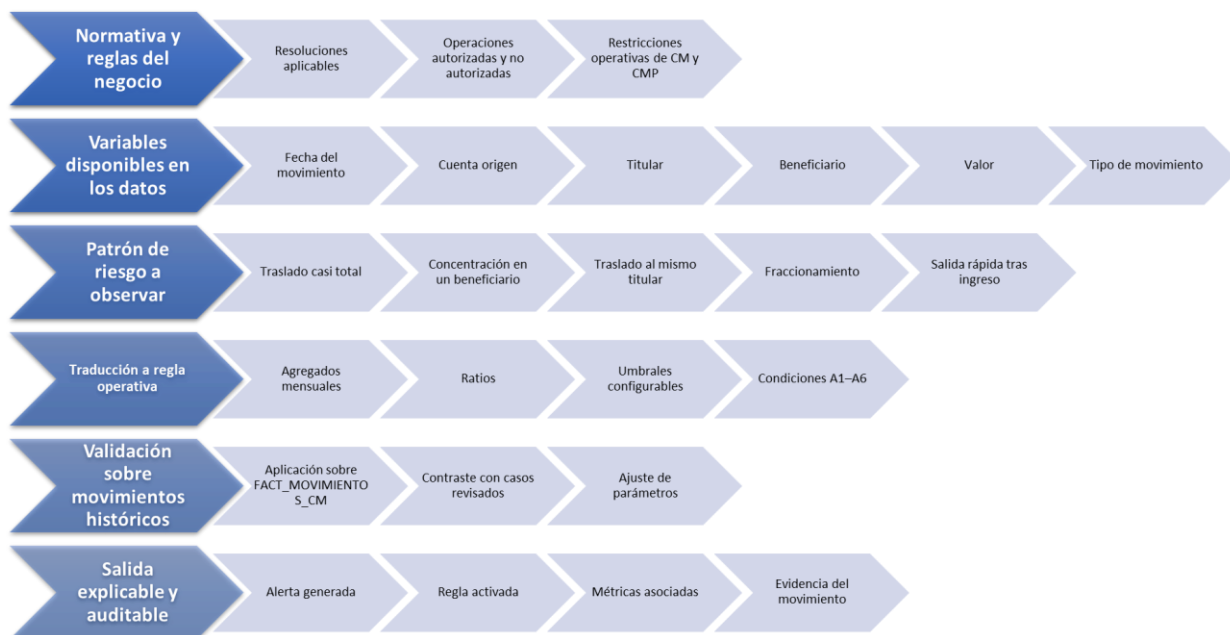
Una vez consolidada la información en el modelo relacional (especialmente en las tablas de hechos de movimientos), este trabajo contempla el desarrollo de dos herramientas analíticas en Python orientadas a apoyar el seguimiento del uso de recursos del SGP administrados en Cuentas Maestras. Ambas herramientas se diseñan bajo un enfoque **explicable y auditable**, es decir, cada resultado debe estar acompañado por el criterio aplicado y la evidencia que lo sustenta (campos, agregados y registros relevantes). Esto permite que los hallazgos se interpreten como **alertas para auditoría**, no como conclusiones definitivas.

4.2.6.1 Herramienta 1: Identificación de alertas de riesgo por traslado de recursos

El objetivo de esta herramienta es identificar situaciones en las que los recursos podrían estar siendo trasladados desde la cuenta maestra hacia cuentas no autorizadas o con patrones que justifican una revisión posterior. El diseño se basa en reglas de comportamiento mensual por cuenta y en agregaciones por beneficiario, utilizando como variables clave la fecha del movimiento, el número de cuenta, el NIT del titular, el NIT o cuenta del beneficiario y el valor del movimiento.

En este modelo, la detección no surge de una inferencia opaca, sino de la transformación de criterios normativos y patrones operativos en reglas explícitas aplicadas sobre variables observables. Esto permite que la herramienta funcione como un sistema de priorización analítica explicable, en el que cada alerta puede rastrearse hasta su fundamento lógico y su soporte en los datos.

Ilustración 11 Proceso de construcción de reglas de detección de riesgo.



Elaboración Propia

Variables para construir el análisis de alertas de riesgo (por cuenta y mes) **Identificadores base (por movimiento)**

- “FECHA_MOVIMIENTO”.
- “NUMERO_CUENTA”.
- “NIT_TITULAR”.
- “NIT_BENEFICIARIO”.
- “CUENTA_BENEFICIARIO”.
- “VALOR”.
- “TIPO_MOVIMIENTO” o “CATEGORIA_MOV” (INGRESO/EGRESO), según la referencia disponible.

Variables agregadas por “cuenta–mes”

- “INGRESO_MES”: suma de “VALOR” para movimientos clasificados como ingreso en el mes.

- “EGRESO_MES”: suma de “VALOR” para movimientos clasificados como egreso en el mes.
- “RATIO_EGRESO_INGRESO”: “EGRESO_MES / INGRESO_MES” (cuando “INGRESO_MES > 0”).
- “FECHA_PRIMER_INGRESO”: fecha mínima de ingreso en el mes (si existe).
- “EGRESOS_EN_VENTANA”: suma de egresos en la ventana “FECHA_PRIMER_INGRESO + VENTANA_DIAS_RAPIDA”.
- “RATIO_SALIDA_RAPIDA”: “EGRESOS_EN_VENTANA / INGRESO_MES” (si “INGRESO_MES > 0”).
- “EGRESO_MISMO_TITULAR”: suma de egresos donde “NIT_BENEFICIARIO = NIT_TITULAR”.
- “RATIO_MISMO_TITULAR”: “EGRESO_MISMO_TITULAR / INGRESO_MES” (si “INGRESO_MES > 0”).

Variables agregadas por “cuenta-mes-beneficiario” (beneficiario definido como “NIT_BENEFICIARIO + CUENTA_BENEFICIARIO”)

- “EGRESO_BENEF”: suma de egresos hacia un beneficiario específico en el mes.
- “N_TX_BENEF”: número de transferencias hacia ese mismo beneficiario en el mes.
- “RATIO_BENEF”: “EGRESO_BENEF / INGRESO_MES” (si “INGRESO_MES > 0”).
- “TOP_BENEF_RATIO”: mayor valor de “RATIO_BENEF” entre beneficiarios de la cuenta en el mes.

Parámetros (umbrales ajustables)

- “UMBRAL_TRASLADO_TOTAL”.
- “UMBRAL_SIGNIFICATIVO”.
- “UMBRAL_SIGNIFICATIVO_MISMO_TITULAR”.
- “N_MIN_TRANSACCIONES” (para fraccionamiento al mismo beneficiario).
- “VENTANA_DIAS_RAPIDA”.

A continuación, se presenta el pseudocódigo del proceso (reglas A1–A6) que genera alertas explicables con métricas y evidencia:

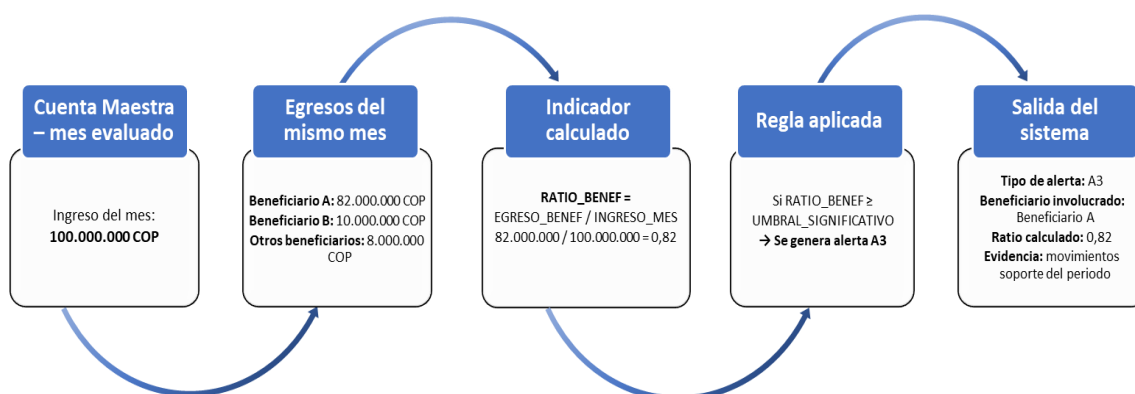
Pseudocódigo – motor de alertas de riesgo

1. Cargar movimientos del mes o periodo a evaluar desde ‘FACT_MOVIMIENTOS_CM’.
2. Clasificar movimientos en ingresos y egresos según el tipo/categoría disponible.
3. Para cada cuenta-mes, calcular agregados: ‘INGRESO_MES’, ‘EGRESO_MES’ y ‘FECHA_PRIMER_INGRESO’ (si existe ingreso).

4. Para cada cuenta-mes y beneficiario (NIT_BENEFICIARIO + CUENTA_BENEFICIARIO), calcular: 'EGRESO_BENEF' y 'N_TX_BENEF'.
5. Evaluar reglas de alerta:
 - a. A1 (mismo titular, traslado casi total): si " $\text{EGRESO_MISMO_TITULAR} / \text{INGRESO_MES} \geq \text{UMBRAL_TRASLADO_TOTAL}$ " → alerta.
 - b. A2 (cualquier beneficiario, traslado casi total): si " $\text{EGRESO_MES} / \text{INGRESO_MES} \geq \text{UMBRAL_TRASLADO_TOTAL}$ " → alerta.
 - c. A3 (concentración en un único destino): si el mayor " $\text{EGRESO_BENEF} / \text{INGRESO_MES} \geq \text{UMBRAL_SIGNIFICATIVO}$ " → alerta (una transferencia o varias).
 - d. A4 (mismo titular, proporcional significativo): si " $\text{EGRESO_MISMO_TITULAR} / \text{INGRESO_MES} \geq \text{UMBRAL_SIGNIFICATIVO_MISMO_TITULAR}$ " → alerta.
 - e. A5 (fraccionamiento al mismo beneficiario): si " $\text{N_TX_BENEF} \geq \text{N_MIN_TRANSACCIONES}$ " y " $\text{EGRESO_BENEF} / \text{INGRESO_MES} \geq \text{UMBRAL_SIGNIFICATIVO}$ " → alerta.
 - f. A6 (salida rápida tras ingreso): si " $\text{EGRESOS_EN_VENTANA} / \text{INGRESO_MES} \geq \text{UMBRAL_SIGNIFICATIVO}$ ", donde la ventana es " $\text{FECHA_PRIMER_INGRESO} + \text{VENTANA_DIAS_RAPIDA}$ " → alerta.
6. Para cada alerta: registrar tipo, métricas (ratios), beneficiario(s) involucrado(s) y evidencia (movimientos soporte: fecha, valor, NIT beneficiario y número de cuenta de beneficiario).
7. Exportar reporte auditable de alertas (CSV/Excel) y un resumen por tipo de alerta y periodo.

A continuación, se presenta un ejemplo simplificado de aplicación de la regla A3, orientada a detectar concentraciones significativas de recursos en un único beneficiario dentro de un mismo mes.

Ilustración 12 Ejemplo aplicado de la regla A3: concentración en un único beneficiario

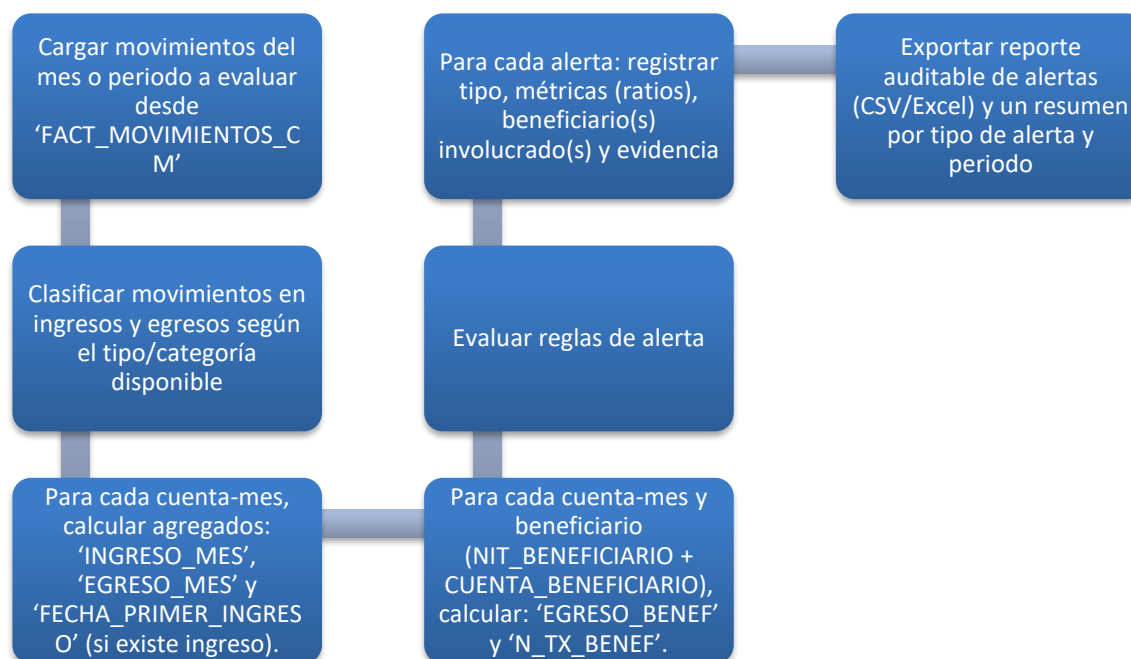


Elaboración Propia

En este caso, la alerta no implica por sí misma la existencia de una irregularidad, sino que señala un patrón de concentración suficientemente relevante como para justificar una revisión posterior desde la perspectiva operativa y normativa.

Este motor de alertas se mantiene parametrizable (umbrales y condiciones) para facilitar ajustes por sector, por contexto territorial o por recomendaciones del usuario, conservando trazabilidad y reproducibilidad de resultados.

Ilustración 13 Flujo general de alertas de riesgo en el uso de los recursos



Elaboración Propia

4.2.6.2 Herramienta complementaria de detección de anomalías mediante aprendizaje automático

Como complemento a la herramienta basada en reglas de negocio, se plantea una segunda herramienta analítica orientada a la identificación de movimientos atípicos mediante técnicas de aprendizaje automático. Su propósito es ampliar la capacidad de detección del sistema frente a situaciones que no siempre pueden anticiparse mediante reglas fijas, umbrales definidos manualmente o combinaciones previamente conocidas de condiciones de riesgo. De esta manera, ambas herramientas se conciben como enfoques complementarios para el análisis de posibles movimientos que requieran revisión.

Mientras la herramienta basada en reglas se fundamenta en conocimiento experto y criterios explícitos de control, esta segunda herramienta se orienta a detectar desviaciones respecto al comportamiento esperado de los movimientos financieros registrados en las Cuentas Maestras. Su lógica general parte de que una operación no debe evaluarse únicamente por su valor absoluto, sino también en función de su contexto transaccional. En otras palabras, un

movimiento puede resultar normal en una cuenta con altos volúmenes de operación, pero representar una señal atípica en otra con menor actividad o con un patrón histórico distinto.

Bajo este criterio, se propone un enfoque híbrido de aprendizaje automático no supervisado, en el que primero se identifiquen perfiles transaccionales relativamente homogéneos y, posteriormente, se evalúe la presencia de anomalías dentro de dichos perfiles. Esta aproximación busca incorporar contexto al proceso de detección, evitando que todos los registros sean analizados como si respondieran a una misma escala o dinámica operativa. Para ello, se plantea utilizar una combinación de técnicas de agrupamiento y detección de outliers, específicamente K-Means para la generación de perfiles transaccionales e Isolation Forest para la identificación de observaciones anómalas dentro del espacio de análisis definido.

Desde el punto de vista metodológico, la herramienta se concibe como un mecanismo de priorización analítica. Su función no será determinar de manera concluyente la existencia de una irregularidad normativa, sino señalar movimientos con comportamiento atípico que ameriten revisión posterior. En este sentido, la anomalía detectada por el modelo debe entenderse como una alerta técnica basada en desviaciones estadísticas o transaccionales, cuya interpretación final deberá realizarse a la luz del contexto operativo, las reglas de negocio y la normativa aplicable.

La incorporación de esta herramienta dentro del TFM responde, por tanto, a una lógica de complementariedad metodológica. Por una parte, se dispone de una herramienta explicable y auditable, sustentada en reglas de negocio. Por otra, se incorpora un enfoque de aprendizaje automático capaz de explorar patrones menos evidentes en los datos. La comparación entre ambos enfoques permitirá valorar posteriormente sus ventajas, limitaciones y nivel de aporte al proceso de identificación de posibles movimientos de riesgo.

4.2.6.3 Herramienta 2: Clústeres de concentración de mercado (beneficiarios recurrentes)

La segunda herramienta tiene como objetivo identificar escenarios de concentración de mercado y patrones de contratación recurrente hacia un mismo beneficiario (persona natural o jurídica), en los que un beneficiario reciba recursos de múltiples cuentas/titulares y por valores significativamente altos. Este análisis es relevante para priorizar auditorías, dado que puede sugerir dependencia de proveedores, baja diversificación o concentración sostenida a nivel territorial (por ejemplo, varios municipios contratando repetidamente al mismo beneficiario).

El enfoque se basa en dos componentes complementarios:

1. **Concentración por beneficiario:** mediciones tipo ranking y participación (por ejemplo, proporción del total recibido, top N y concentración acumulada).
2. **Clusterización (agrupamiento):** agrupación de beneficiarios según su perfil de recepción (monto total, número de cuentas origen, frecuencia y recurrencia), con el fin de identificar “tipologías” de concentración.
3. **Variables para construir el análisis de concentración y clústeres (por beneficiario y periodo)**

Identificadores base (por movimiento)

- “FECHA_MOVIMIENTO”.
- “NUMERO_CUENTA” (cuenta origen).
- “NIT_TITULAR” (titular de la cuenta origen).
- “DIVIPOLA” (si está disponible vía cuenta/titular; se usa para análisis territorial).
- “NIT_BENEFICIARIO”.
- “CUENTA_BENEFICIARIO”.
- “VALOR”.
- “TIPO_MOVIMIENTO” o “CATEGORIA_MOV” (INGRESO/EGRESO), según la referencia disponible.

Unidad de análisis (beneficiario)

- “BENEFICIARIO_CLAVE”: combinación “NIT_BENEFICIARIO + CUENTA_BENEFICIARIO”.

Variables agregadas por beneficiario (por mes o ventana definida)

- “TOTAL_RECIBIDO”: suma de “VALOR” donde el beneficiario es receptor.
- “N_CUENTAS_ORIGEN”: número de cuentas distintas (“NUMERO_CUENTA”) que transfieren al beneficiario.
- “N_TITULARES_ORIGEN”: número de titulares distintos (“NIT_TITULAR”) que transfieren al beneficiario.
- “N_TERRITORIOS_ORIGEN”: número de territorios distintos (“DIVIPOLA”) desde los cuales recibe recursos (cuando está disponible).
- “N_TRANSACCIONES”: cantidad de movimientos recibidos (conteo de registros).
- “PROMEDIO_TRANSACCION” y “MAX_TRANSACCION”: promedio y máximo de “VALOR” recibido.
- “RECURRENCIA”: número de meses en los que el beneficiario aparece recibiendo recursos (si se usa una ventana multimensual).

Indicadores derivados para concentración (interpretables)

- “PARTICIPACION”: “TOTAL_RECIBIDO / TOTAL_GENERAL” (en el periodo).
- “RANKING_RECIBIDO”: posición del beneficiario según “TOTAL_RECIBIDO”.
- “TOP_N_ACUMULADO”: participación acumulada de los N beneficiarios principales.

A continuación, se presenta el pseudocódigo general de la herramienta.

Pseudocódigo – clústeres de concentración por beneficiario

1. Cargar movimientos del periodo de análisis desde ‘FACT_MOVIMIENTOS_CM’.

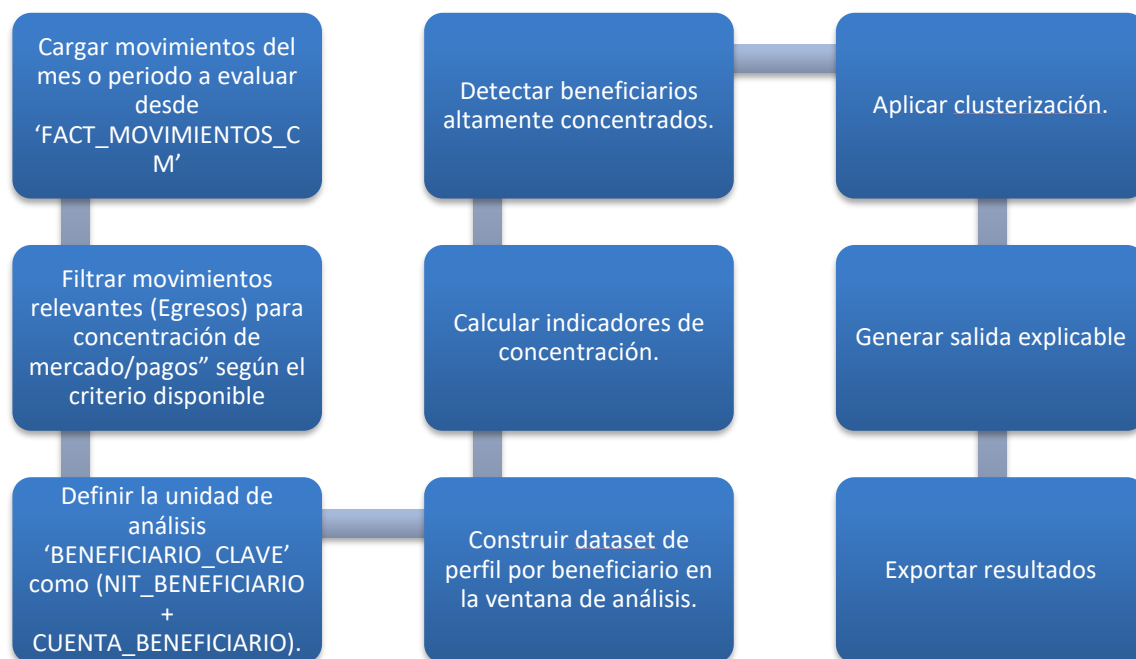
2. Filtrar movimientos relevantes (Egresos) para contratación/pagos según el criterio disponible.
3. Definir la unidad de análisis 'BENEFICIARIO_CLAVE' como (NIT_BENEFICIARIO + CUENTA_BENEFICIARIO).
4. Construir dataset de perfil por beneficiario en la ventana de análisis:
 - "TOTAL_RECIBIDO = suma (VALOR)"
 - "N_CUENTAS_ORIGEN = conteo de NUMERO_CUENTA distinto"
 - "N_TITULARES_ORIGEN = conteo de NIT_TITULAR distinto"
 - "N_TERRITORIOS_ORIGEN = conteo de DIVIPOLA distinto" (si está disponible vía cuenta/titular)
 - "N_TRANSACCIONES = conteo de movimientos"
 - "PROMEDIO_TRANSACCION, MAX_TRANSACCION"
 - "RECURRENCIA = número de meses con recepción" (si se usa ventana multimensual)
5. Calcular indicadores de concentración:
 - "Ranking de beneficiarios por TOTAL_RECIBIDO".
 - "Participación del beneficiario = TOTAL_RECIBIDO / TOTAL_GENERAL".
 - "Top N beneficiarios y participación acumulada".
6. Detectar beneficiarios altamente concentrados:
 - Si "TOTAL_RECIBIDO" o "Participación" supera un umbral, o si "N_TITULARES_ORIGEN" y "N_TERRITORIOS_ORIGEN" son altos, marcar como caso prioritario.
7. Aplicar clusterización:
 - "Seleccionar variables del perfil (TOTAL_RECIBIDO, N_TITULARES_ORIGEN, N_TERRITORIOS_ORIGEN, N_TRANSACCIONES, RECURRENCIA...)".
 - "Estandarizar variables para comparabilidad".
 - "Ejecutar algoritmo de agrupamiento (por ejemplo, k-means o clustering jerárquico)".
 - "Asignar etiqueta de clúster a cada beneficiario".
8. Generar salida explicable:
 - Tabla de beneficiarios con: métricas del perfil, ranking, clúster (si aplica) y motivo de priorización.
 - Resumen por clúster: tamaño, rangos típicos y principales beneficiarios.

9. Exportar resultados:

- Reporte en CSV/Excel con ranking, indicadores y clusterización (si aplica), listo para revisión.

El objetivo no es concluir irregularidades, sino detectar patrones de concentración y recurrencia que justifican análisis posterior. En particular, un beneficiario que recibe montos altos desde múltiples cuentas y territorios puede representar un caso de concentración relevante para auditoría, especialmente si el patrón se repite en el tiempo.

Ilustración 14 Flujo general Clústeres de concentración de mercado



Elaboración Propia

4.3 Recursos requeridos

Para el desarrollo del proyecto se requieren recursos tecnológicos, herramientas de software y material normativo/técnico asociado a CM y CMP.

Recursos tecnológicos (hardware)

- Computador con procesador de cuatro núcleos o superior.
- Memoria RAM mínima de 8 GB.
- Almacenamiento mínimo de 256 GB en unidad SSD; se recomienda 512 GB o superior para mejorar el desempeño en los procesos de carga y análisis.
- Sistema operativo Windows 10/11 o Linux.

Software y herramientas

- Motor de base de datos relacional: SQL Server.
- Herramienta de administración de base de datos, como SQL Server Management Studio (SSMS).
- Python 3.9 o superior, con librerías orientadas al procesamiento de datos, conexión a bases de datos, análisis y generación de visualizaciones.
- Entorno de desarrollo para scripting y pruebas, como Visual Studio Code o Jupyter Notebook.
- Controladores de conexión a SQL Server, como ODBC Driver compatible.
- Herramientas ofimáticas y de apoyo a la documentación, revisión y gestión de artefactos, como Word y Excel.

Insumos

- Normativa aplicable al reporte de CM y CMP, junto con sus anexos técnicos.
- Archivos históricos de reporte en formato “.txt” para pruebas, homologación y validación del proceso ETL.
- Archivos maestros en formato Excel utilizados como base para la carga de dimensiones y tablas de referencia.
- Archivos de contraste o soportes de validación utilizados para evaluar el desempeño de las herramientas analíticas.

4.4 Presupuesto

El presupuesto del proyecto se estimó a partir de los recursos requeridos para su desarrollo, incluyendo tanto el valor del tiempo invertido como el costo de referencia del equipo técnico y del software utilizado. Dado que se trata de un trabajo académico desarrollado principalmente con herramientas disponibles o de libre uso, el costo principal del proyecto corresponde al tiempo de dedicación requerido para el diseño, implementación, validación y documentación de la solución. Para efectos del presupuesto se adoptó una estimación revisada de **180 horas** de trabajo, valorada a una tarifa de referencia de **75.000 COP por hora**.

De igual forma, los costos de hardware y software se presentan como valores aproximados de mercado, incluso cuando no implicaron una compra específica para el desarrollo del trabajo, con el fin de reflejar el valor económico equivalente de los recursos empleados.

Tabla 2 Presupuesto estimado del proyecto

Tipo de coste	Valor (COP)	Comentarios
Horas de trabajo en el proyecto	13.500.000	Estimado con base en 180 horas de trabajo y una tarifa de referencia de 75.000 COP por hora.

Equipo técnico utilizado	8.000.000	Valor aproximado de un equipo de cómputo de gama media-alta, con procesador de 4 núcleos o superior, memoria RAM suficiente para procesos ETL y analíticos, y unidad SSD de 512 GB o superior.
Software utilizado	500.000	Valor estimado de referencia asociado a herramientas ofimáticas y de apoyo a la documentación, revisión y gestión de artefactos utilizadas durante el proyecto. Las demás herramientas técnicas empleadas, como “SQL Server”, “SSMS”, “Python”, “Visual Studio Code” y “Jupyter Notebook”, no implicaron costo de licencia.
Estudios e informes	0	No se adquirieron informes pagos ni estudios de consultoría para el desarrollo del proyecto.
Materiales empleados	0	No se requirieron materiales físicos especializados para la ejecución del trabajo.
Total estimado del proyecto	22.000.000	Incluye tiempo de dedicación, equipo técnico y software de apoyo utilizados en el desarrollo del trabajo.

Nota. El presupuesto total estimado del proyecto asciende a **22.000.000 COP**, equivalentes aproximadamente a **5.164,56 EUR**, según una tasa de cambio de referencia de **1 EUR = 4.259,80 COP**, consultada el **13 de abril de 2026**. Esta conversión se presenta únicamente con fines comparativos e ilustrativos para facilitar la interpretación internacional del presupuesto.

A partir de esta estimación económica, resulta pertinente valorar no solo el costo de referencia del proyecto, sino también su relación con los beneficios esperados y con la sostenibilidad futura de la solución propuesta.

4.5 Viabilidad

El proyecto desarrollado presenta viabilidad tanto desde la perspectiva económica como desde la sostenibilidad técnica y operativa de sus resultados. La solución propuesta se apoya en un conjunto de herramientas tecnológicas accesibles y ampliamente utilizadas en contextos de análisis de datos, como SQL Server, Python y utilitarios ofimáticos de apoyo, lo que permite

reducir barreras de entrada para su implementación. En este sentido, aunque el presupuesto estimado del proyecto asciende a 22.000.000 COP, la mayor parte de este valor corresponde al tiempo de dedicación invertido en el diseño, desarrollo, validación y documentación de la solución, más que a costos elevados de licenciamiento o infraestructura especializada.

Desde el punto de vista de la relación coste-beneficio, la propuesta resulta favorable en la medida en que transforma un proceso tradicionalmente disperso, manual y de difícil explotación en una solución estructurada de integración, trazabilidad y análisis. El modelo relacional permite consolidar la información histórica de Cuentas Maestras y Cuentas Maestras Pagadoras en una base organizada y consultable; el proceso ETL automatiza tareas de lectura, homologación, validación y carga de archivos; y las herramientas analíticas permiten priorizar movimientos potencialmente riesgosos y patrones de concentración que, sin este tipo de apoyo, requerirían revisiones manuales mucho más extensas.

En términos operativos, uno de los principales factores de viabilidad del proyecto radica en su capacidad de automatizar actividades de identificación de riesgos que, de otra forma, demandarían tiempo significativo de análisis humano. Esta automatización puede traducirse en ahorros potenciales de tiempo y de personal, especialmente en contextos donde las labores de revisión y seguimiento son desarrolladas por perfiles cuyo costo horario puede situarse aproximadamente entre 35.000 y 40.000 COP. En consecuencia, la solución no solo aporta una mejora metodológica y técnica, sino que también puede contribuir a una utilización más eficiente de los recursos humanos dedicados al monitoreo, seguimiento y control. Bajo esta lógica, incluso reducciones parciales en la carga manual de revisión pueden representar beneficios acumulados relevantes en términos de eficiencia operativa, especialmente cuando el análisis debe repetirse de forma periódica sobre grandes volúmenes de información.

Desde la perspectiva de sostenibilidad, el proyecto presenta condiciones favorables para su continuidad futura. El modelo de datos fue diseñado bajo una estructura relacional normalizada, lo que facilita su mantenimiento, actualización y ampliación. De forma complementaria, los procesos ETL y las herramientas analíticas fueron desarrollados con una lógica modular, lo que permite ajustar reglas, incorporar nuevas estructuras de entrada o fortalecer los modelos sin necesidad de rediseñar completamente la solución. Esta característica resulta especialmente importante en un contexto normativo cambiante, como el asociado a la entrada en operación de la estructura unificada prevista en la Resolución 2394 de 2023.

De igual forma, la sostenibilidad futura se ve favorecida por el uso de tecnologías ampliamente conocidas y disponibles en entornos institucionales y académicos. Esto facilita tanto la transferencia del conocimiento como la posibilidad de mantenimiento por parte de otros perfiles técnicos. No obstante, la sostenibilidad efectiva de la solución dependerá también de factores externos, especialmente de la calidad del dato reportado por las entidades bancarias, de la disponibilidad operativa de las nuevas estructuras normativas y de la continuidad en los procesos de actualización y validación de la información.

En síntesis, la solución propuesta puede considerarse viable porque combina un costo de implementación razonable con beneficios concretos en organización, trazabilidad, automatización y apoyo analítico. A ello se suma su potencial de reutilización y evolución futura,

lo que permite afirmar que el proyecto no solo es factible en el marco académico en el que fue desarrollado, sino también sostenible como base para desarrollos posteriores orientados al seguimiento y análisis del uso de recursos públicos.

4.6 Resultados del proyecto

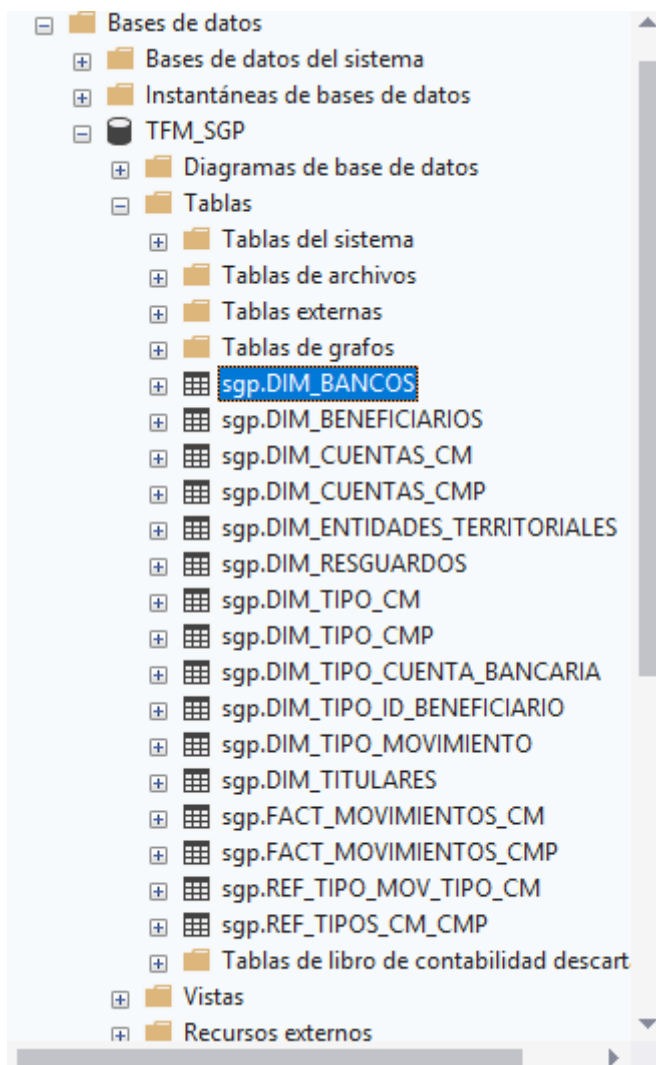
En este apartado se presentan los principales resultados obtenidos a partir del desarrollo e implementación de la solución propuesta. La exposición se organiza de acuerdo con los componentes definidos en la metodología del proyecto: la implementación del modelo físico en SQL Server, el proceso ETL de integración y validación de datos, y las herramientas analíticas desarrolladas para la identificación de posibles situaciones de riesgo y de concentración en el uso de los recursos. Esta estructura permite relacionar los resultados alcanzados con los objetivos específicos planteados y mostrar de forma progresiva cómo el proyecto avanzó desde el diseño del modelo hasta su aplicación analítica sobre los datos consolidados.

4.6.1 Resultados de la implementación del modelo físico en SQL Server.

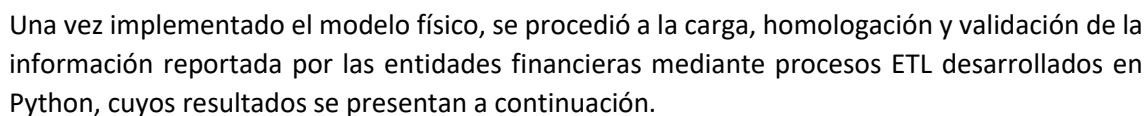
La implementación del modelo físico se materializó en SQL Server mediante la ejecución del script maestro “99_run_all.sql”, diseñado para orquestar en un único flujo la creación del esquema, las tablas y la carga inicial de datos semilla correspondientes a aquellas dimensiones paramétricas y tablas de referencia con valores previamente definidos. Esta estrategia permitió estandarizar el proceso de despliegue del modelo, reducir errores asociados al orden de creación de objetos con dependencias entre claves primarias y foráneas y favorecer la reproducibilidad del entorno de base de datos. La orquestación se apoyó en el modo SQLCMD y en la directiva “:r”, lo que permitió incluir y ejecutar scripts en secuencia de manera controlada y automatizada (Microsoft).

Como resultado de la ejecución satisfactoria del script maestro, se generaron en la base de datos “TFM_SGP”, bajo el esquema “sgp”, las tablas requeridas para soportar el modelo relacional propuesto. Esto permitió materializar la estructura física definida en la etapa de diseño, incluyendo tablas de dimensiones, tablas de hechos y tablas de referencia, con sus respectivas relaciones y restricciones básicas de integridad.

Ilustración 15 Demostración de la Implementación del Modelo en SQL Server.



La evidencia de este despliegue permite confirmar que la solución propuesta trascendió el plano conceptual y lógico, y fue implementada en una estructura técnica funcional, preparada para soportar los procesos de carga, validación y análisis desarrollados en las fases posteriores del proyecto. La captura del entorno de SQL Server y el diagrama entidad-relación final permiten observar tanto la materialización de los objetos creados como la estructura completa del modelo físico resultante.



El proceso ETL permitió integrar y cargar en el modelo relacional la información reportada por las entidades bancarias en archivos planos “.txt” de Cuentas Maestras (CM) y Cuentas Maestras Pagadoras (CMP), correspondientes al periodo comprendido entre enero de 2018 y diciembre de 2025. Para el caso de CM se procesaron 1.614 archivos en total, incluyendo 408 archivos asociados a Resguardos Indígenas Certificados, mientras que para CMP se procesaron 1.116 archivos. Este alcance permitió trabajar con un volumen amplio de información histórica y validar la capacidad del modelo para consolidar estructuras heterogéneas en un esquema relacional unificado.

52

de 2023 aún no se encontraba disponible para su explotación operativa, debido a que el nuevo aplicativo web SISFEVAL se encuentra en fase final de implementación, los procesos ETL y las tablas desarrolladas fueron diseñados para permitir la homogeneización de ambas estructuras. En consecuencia, la solución no solo permitió integrar los archivos históricos efectivamente disponibles, sino que quedó preparada para incorporar la nueva estructura sin necesidad de rediseñar el modelo de datos.

Como resultado de la ejecución de los procesos ETL, se logró poblar satisfactoriamente la totalidad de las tablas principales del modelo. En particular, se cargaron las tablas de hechos “FACT_MOVIMIENTOS_CM” y “FACT_MOVIMIENTOS_CMP”, junto con las dimensiones y tablas de apoyo necesarias para soportar trazabilidad, validación y análisis posterior, entre ellas “DIM_BENEFICIARIOS”, “DIM_CUENTAS_CM”, “DIM_CUENTAS_CMP”, “DIM_TITULARES”, “DIM_ENTIDADES_TERRITORIALES”, “DIM_RESGUARDOS”, “DIM_BANCOS”, “DIM_TIPO_MOVIMIENTO”, “REF_TIPO_MOV_TIPO_CM” y las demás tablas paramétricas del modelo. Los volúmenes finales cargados evidencian la materialización efectiva del proceso de integración y la capacidad de la solución para almacenar información transaccional y maestra dentro de una misma estructura consistente.

Tabla 3 Tablas del modelo SQL con su total de filas

Tabla	Total Filas
FACT_MOVIMIENTOS_CM	14340242
FACT_MOVIMIENTOS_CMP	6910269
DIM_BENEFICIARIOS	1027996
DIM_CUENTAS_CM	17194
DIM_TITULARES	8627
DIM_CUENTAS_CMP	6234
DIM_ENTIDADES_TERRITORIALES	1135
DIM_RESGUARDOS	1070
REF_TIPO_MOV_TIPO_CM	141
DIM_BANCOS	112
DIM_TIPO_MOVIMIENTO	32
REF_TIPOS_CM_CMP	13
DIM_TIPO_CMP	13
DIM_TIPO_CM	12
DIM_TIPO_ID_BENEFICIARIO	4
DIM_TIPO_CUENTA_BANCARIA	2

Un resultado metodológicamente relevante del proceso fue la necesidad de incorporar un mecanismo adicional para identificar cuentas maestras válidas que no se encontraban registradas en la fuente maestra inicial de cuentas. Para ello se desarrolló el script “00_CM_NO_REGISTRADAS”, orientado a detectar cuentas presentes en los archivos históricos que, aunque no figuraban en el inventario vigente, sí habían estado registradas en algún

momento del periodo analizado y posteriormente fueron objeto de sustitución. A partir de este procedimiento se realizó una validación de dichas cuentas frente a los criterios normativos aplicables y, una vez confirmada su validez, estas fueron incorporadas mediante el proceso ETL de la dimensión “DIM_CUENTAS_CM”. Este ajuste permitió ampliar la cobertura histórica de la dimensión y reducir pérdidas de información derivadas de sustituciones o cambios en el inventario de cuentas.

En términos de calidad de datos, el proceso mostró un nivel de error controlado. Estas incidencias estuvieron asociadas principalmente a errores atribuibles a los establecimientos bancarios responsables del reporte, especialmente por la inclusión de registros correspondientes a cuentas que no cumplían las condiciones normativas para ser consideradas cuentas maestras válidas.

Tabla 4 Resultados de los procesos de ETL en las tablas FACT

Componente	Archivos procesados	Registros válidos cargados	Registros con error	Porcentaje de error
Cuentas Maestras (CM)	1.614	14.532.579 movimientos	207.541 movimientos	1,41%
Cuentas Maestras Pagadoras (CMP)	1.116	6.910.269 movimientos	17.398 movimientos	0,25%

Otra incidencia identificada durante la carga estuvo relacionada con la aparición de beneficiarios vacíos en determinados registros. Este comportamiento se observó principalmente en movimientos que no correspondían a egresos, por lo que se decidió construir la dimensión “DIM_BENEFICIARIOS” únicamente a partir de movimientos de egreso, en los cuales la información del tercero receptor resultaba relevante y consistente para los fines analíticos del proyecto. Como resultado de este criterio, se consolidaron 5.791.310 registros válidos en dicha dimensión, evitando errores de poblamiento asociados a registros sin contenido útil para el análisis de riesgo y concentración.

Desde el punto de vista técnico, el ETL conservó como texto los campos de identificación y codificación, con excepción del valor monetario, con el fin de evitar pérdidas de formato en elementos sensibles como números de cuenta, identificadores territoriales o códigos bancarios. Esta decisión fue especialmente importante para preservar la integridad de la información original reportada en los archivos planos y asegurar su correcta homologación dentro del modelo relacional.

Adicionalmente, el proceso generó archivos de control y trazabilidad para el seguimiento de la calidad de los datos, incluyendo listados de errores por vigencia anual. Esta salida permitió identificar incidencias específicas para cada periodo de procesamiento y facilitó la revisión de inconsistencias persistentes en los reportes de las entidades financieras.

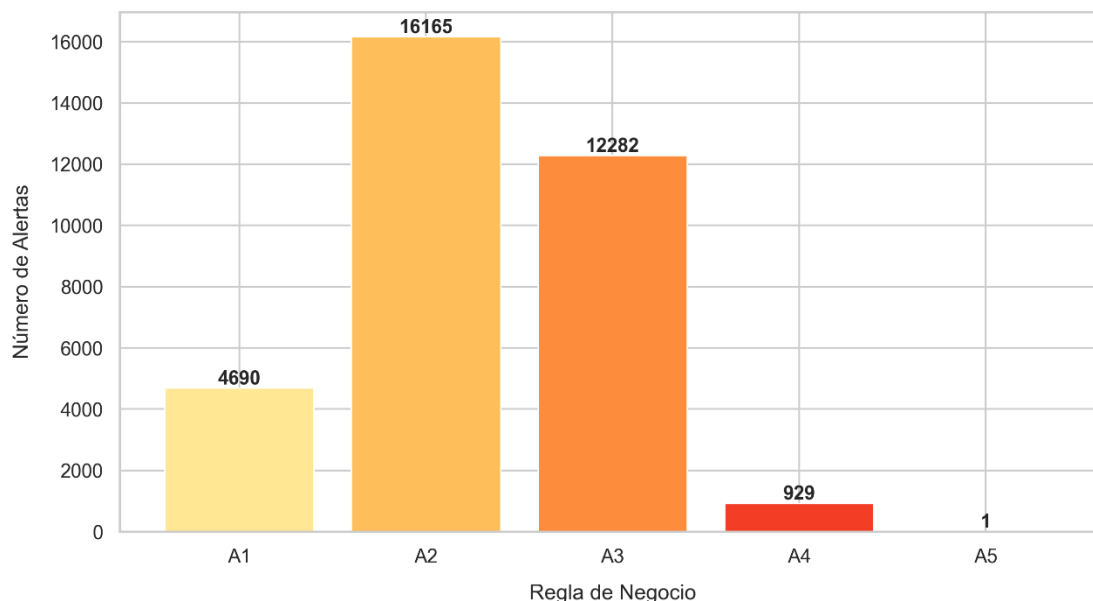
En términos funcionales, el principal resultado del ETL consistió en consolidar información originalmente dispersa en múltiples archivos “.txt”, con estructuras heterogéneas y difícil explotación mediante métodos convencionales, en una base de datos relacional organizada y consultable. Esta transformación permitió facilitar el acceso a la información en distintos niveles de granularidad, fortalecer la trazabilidad de los movimientos y dejar una base sólida para la aplicación de las herramientas analíticas desarrolladas en el proyecto.

4.6.3 Resultados de la herramienta de alertas de riesgo basada en reglas de negocio

La primera herramienta analítica desarrollada en el proyecto correspondió a un modelo explicable de detección de alertas de riesgo basado en reglas de negocio, aplicado sobre la tabla “*FACT_MOVIMIENTOS_CM*”. Su objetivo fue identificar patrones transaccionales que, por su comportamiento financiero y operativo, ameritaran revisión posterior en el marco del seguimiento al uso de los recursos del SGP. Para ello, el modelo operó sobre movimientos de Cuentas Maestras, excluyó los tipos de movimientos “500” y “600” por no representar transacciones reales, e incorporó umbrales específicos para las reglas “A1” a “A5”, priorizando la detección de traslados casi totales, concentraciones significativas en beneficiarios, traslados al mismo titular y fraccionamiento de operaciones.

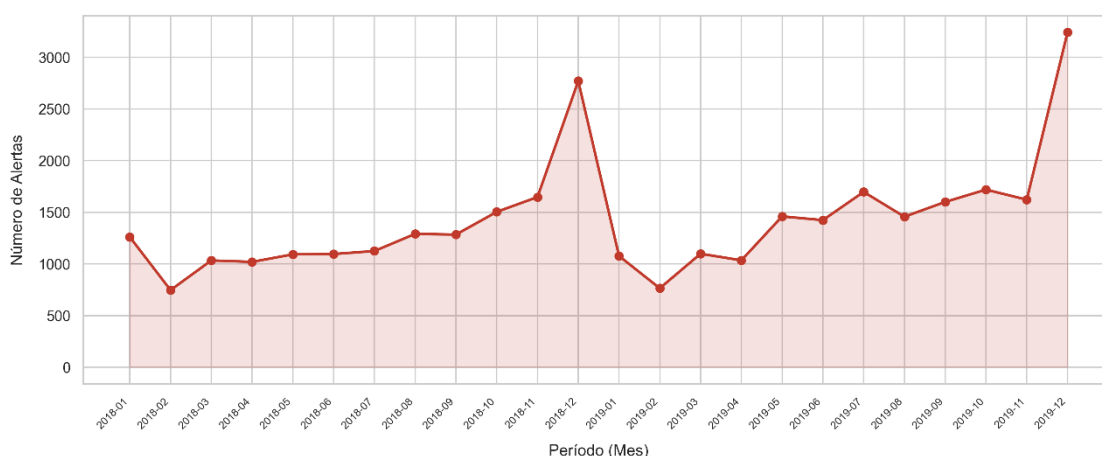
Como resultado de su ejecución, el modelo generó 34.067 alertas a nivel de regla. La distribución obtenida mostró que la mayor parte de los casos correspondió a la regla “A2”, asociada a salidas casi totales de recursos, con 16.165 alertas, seguida por la regla “A3”, relacionada con concentración en un único beneficiario, con 12.282 alertas. En una proporción menor se identificaron 4.690 alertas bajo la regla “A1”, 929 alertas bajo la regla “A4” y un solo caso bajo la regla “A5”. Esta distribución sugiere que, dentro del universo analizado, los patrones más frecuentes de riesgo estuvieron vinculados a salidas amplias de recursos y a comportamientos de concentración, mientras que el fraccionamiento, bajo los umbrales definidos, fue poco frecuente en el contraste realizado.

Ilustración 17 Distribución de Alertas por Regla de Negocio



Desde una perspectiva temporal, la herramienta permitió observar que las alertas no se distribuyeron de manera uniforme entre periodos, sino que presentaron variaciones mensuales asociadas al comportamiento operativo de las cuentas y a la intensidad del uso de los recursos. Esta lectura temporal resulta útil porque no solo cuantifica las alertas generadas, sino que permite identificar meses con mayor densidad de eventos y, por tanto, periodos susceptibles de priorización en revisiones posteriores. Se observa que los picos se presentan al final de cada vigencia.

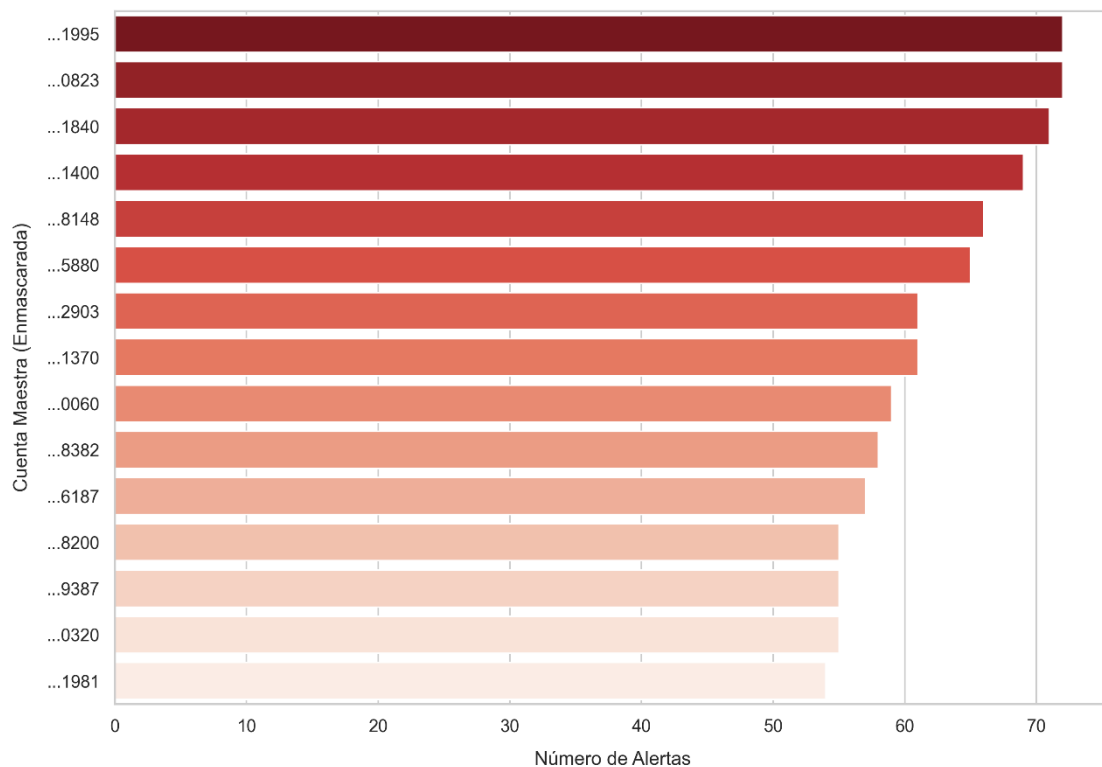
Ilustración 18 Evolución Temporal de Alertas de Riesgo



A nivel de cuentas, el modelo también permitió identificar cuáles concentraron un mayor número de alertas. Este resultado es importante porque desplaza el análisis desde la lógica puramente transaccional hacia una lógica de focalización operativa: no solo interesa saber

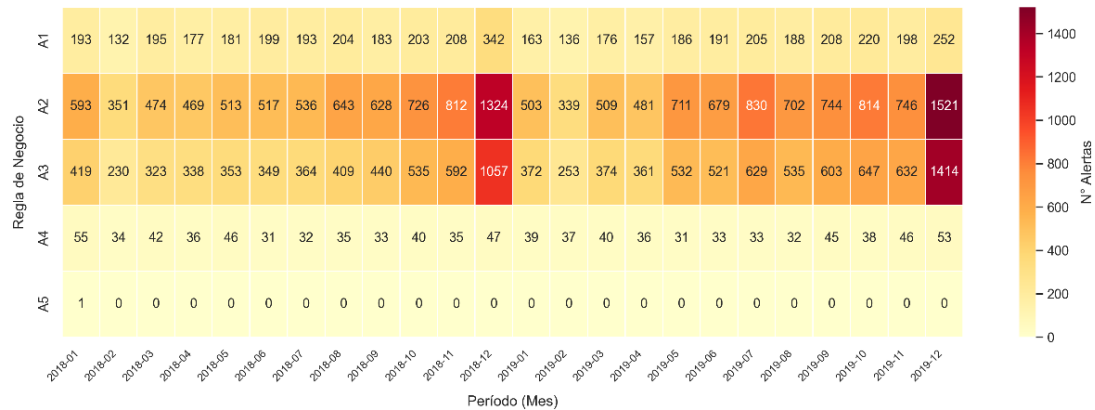
cuántas alertas se generaron, sino en qué cuentas se acumularon de manera recurrente. La visualización de las cuentas con mayor frecuencia de alertamiento aporta así un criterio adicional para priorizar análisis más profundos o cruces con otra información administrativa y financiera.

Ilustración 19 Top 15 Cuentas con Mayor Número de Alertas



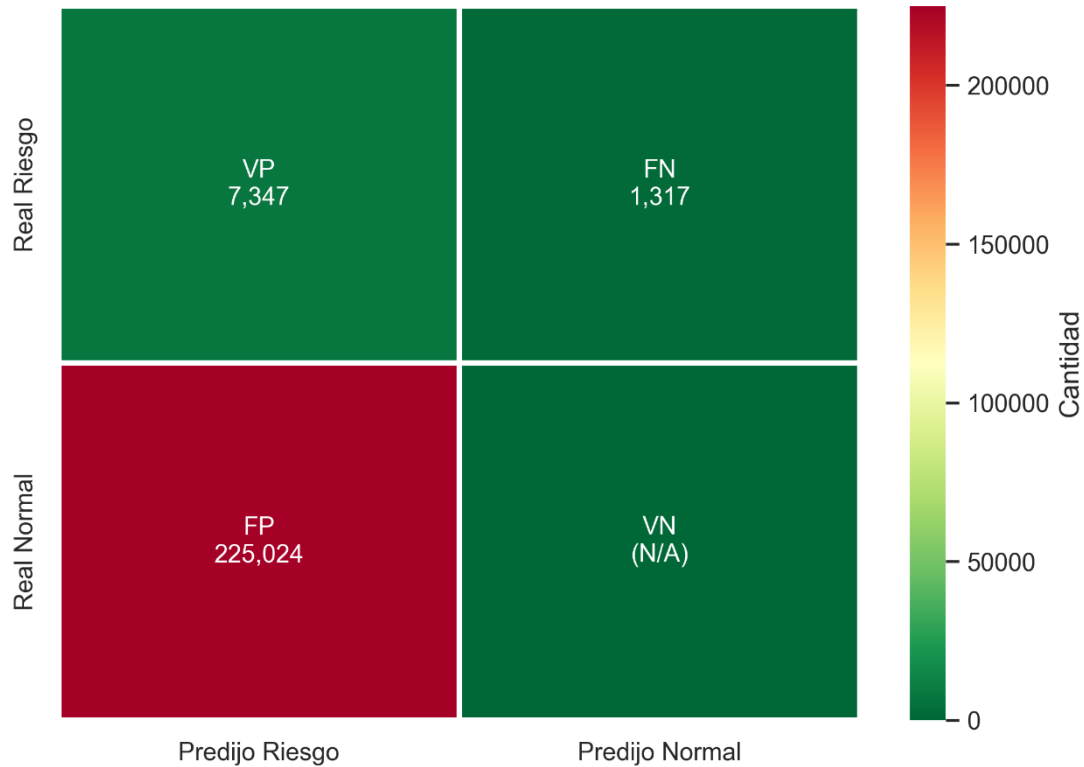
De manera complementaria, la combinación entre regla aplicada y periodo permitió observar que ciertos patrones se activaron con mayor intensidad en momentos específicos. Este cruce ofrece una visión más rica del comportamiento del modelo, dado que no todas las reglas capturan el mismo tipo de riesgo ni se manifiestan con la misma frecuencia en el tiempo. En este sentido, la lectura conjunta entre categoría de alerta y periodo contribuye a una interpretación más contextualizada de los resultados.

Ilustración 20 Mapa de Calor: Alertas por Regla y Periodo



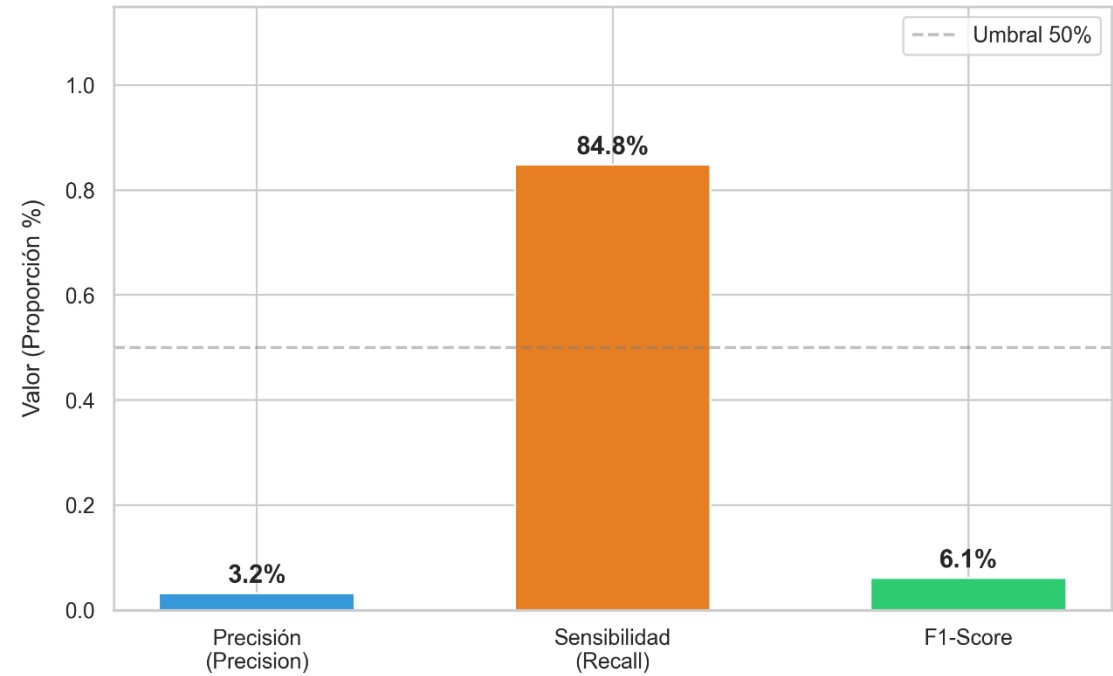
Con el fin de evaluar la eficacia del modelo, sus resultados fueron contrastados frente a un conjunto de movimientos previamente identificados por revisión humana para los años 2018 y 2019. En este ejercicio, el modelo basado en reglas detectó 232.371 movimientos como riesgosos, frente a un universo de 8.664 riesgos reales contenidos en el archivo de contraste. El cruce permitió identificar 7.347 verdaderos positivos, 225.024 falsos positivos y 1.317 falsos negativos. Aunque la precisión obtenida fue baja (3,16%), el resultado más relevante es que el modelo logró recuperar el **84,80%** de los casos reales identificados, dejando sin señalar **1.317** casos frente a más de **233.000** registros evaluados como potencialmente riesgosos. Este comportamiento es consistente con una herramienta orientada a la priorización y apoyo al monitoreo, donde resulta más crítico minimizar la omisión de riesgos relevantes que reducir al máximo la generación de alertas adicionales. En conjunto, el modelo alcanzó un *F1-score* de 6,10%.

Ilustración 21 Matriz de Confusión - Reglas de Negocio



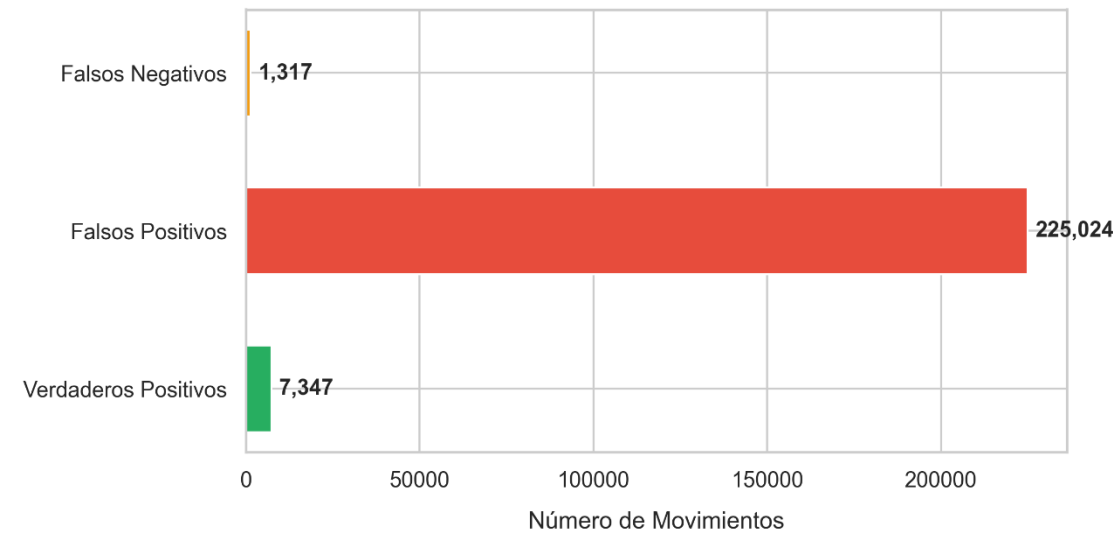
Aunque la precisión obtenida fue baja, este resultado debe interpretarse a la luz del contexto del problema. En escenarios de auditoría y control fiscal, el costo de omitir un movimiento riesgoso es mayor que el costo de revisar una alerta adicional que finalmente no se confirme. En consecuencia, el modelo fue calibrado para privilegiar la sensibilidad del sistema, es decir, su capacidad para recuperar la mayor cantidad posible de riesgos potenciales. En ese marco, el *recall* superior al 84% constituye un resultado relevante, ya que evidencia que la herramienta logró capturar una proporción alta de los casos previamente identificados como problemáticos, aun al costo de generar un volumen considerable de alertas adicionales. Este resultado es consistente con la literatura sobre fraude y anomalías financieras en escenarios de fuerte desbalance de clases, donde la evaluación del modelo no se limita a la precisión, sino que también considera su capacidad para recuperar casos potencialmente riesgosos (Al-Hashedi & Magalingam, 2021).

Ilustración 22 Métricas de Eficacia - Reglas de Negocio



La composición de resultados confirma este comportamiento. El modelo logró una cobertura amplia de los riesgos reales, aunque acompañada de una cantidad elevada de falsas alarmas. Desde el punto de vista operativo, esto refuerza la idea de que la herramienta no debe entenderse como un mecanismo de decisión automática, sino como un sistema de priorización explicable para auditoría, donde la trazabilidad de la regla activada y la evidencia asociada constituyen su principal fortaleza.

Ilustración 23 Composición de Resultados - Reglas de Negocio



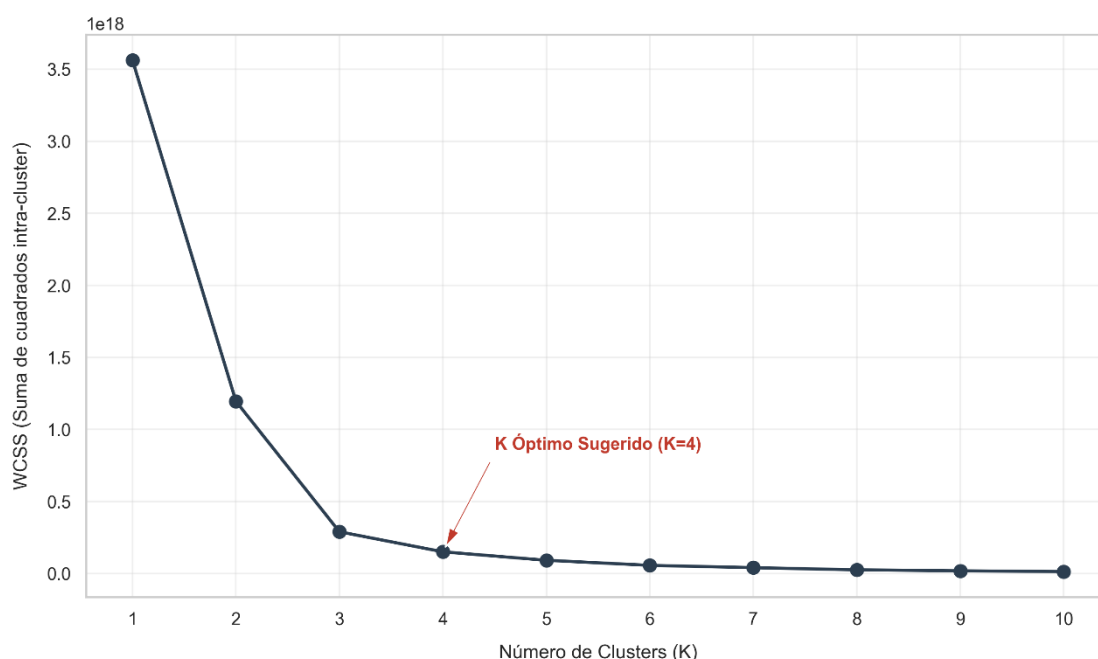
Si bien el modelo por reglas mostró una capacidad alta para recuperar riesgos previamente identificados, sus resultados también evidenciaron la existencia de un volumen importante de alertas que no coincidían con el archivo de contraste. A partir de esta observación, resultó pertinente evaluar un segundo enfoque complementario, basado en aprendizaje automático, orientado a detectar anomalías transaccionales desde una lógica no supervisada y contextualizada.

4.6.4 Resultados de la herramienta complementaria de detección de anomalías mediante aprendizaje automático

Como complemento al modelo basado en reglas, se implementó una herramienta híbrida de aprendizaje automático que combinó técnicas de agrupamiento y detección de anomalías. En particular, se utilizó “*KMeans*” para construir perfiles transaccionales relativamente homogéneos e “*Isolation Forest*” para identificar movimientos atípicos dentro de dichos perfiles.

Esta combinación respondió a la necesidad de incorporar contexto al análisis, bajo el supuesto de que un mismo valor monetario no representa necesariamente el mismo nivel de riesgo en cuentas con comportamientos operativos distintos. Al respecto, a través del método del codo se determinó que la cantidad óptima de clústeres sería de 4 para el conjunto de datos que se tiene.

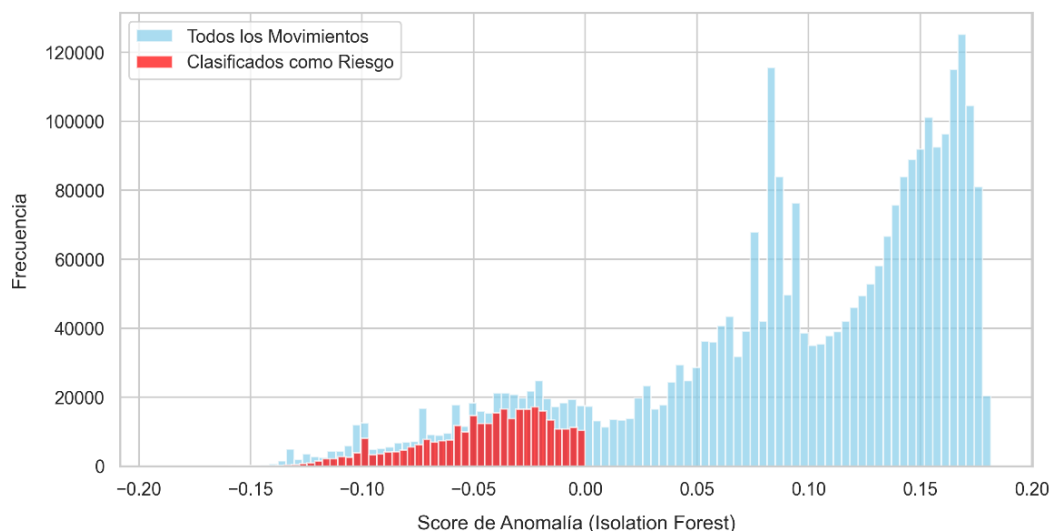
Ilustración 24 Método del Codo - Justificación de Clústeres de Riesgo



La ejecución del modelo permitió identificar 242.067 movimientos como anómalos y potencialmente riesgosos dentro del periodo de contraste. A diferencia del enfoque por reglas, este modelo no dependió de umbrales de negocio explícitos para cada patrón, sino de la capacidad del algoritmo para detectar desviaciones respecto al comportamiento esperado según la interacción entre variables como el valor del movimiento, el peso relativo de los

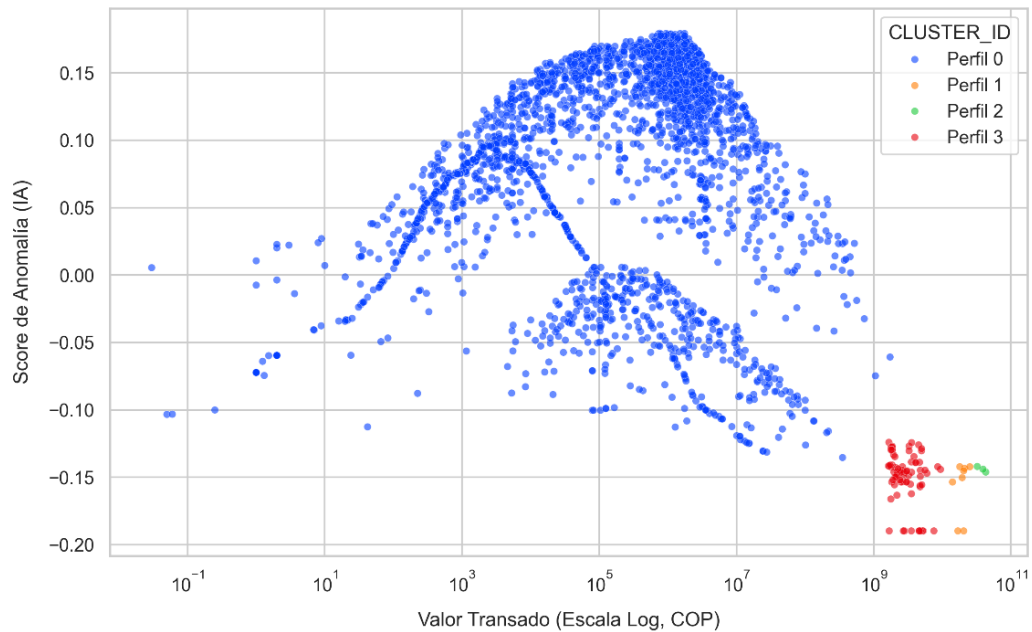
egresos, la relación con el titular, la frecuencia de egresos y el perfil transaccional asociado a cada cuenta.

Ilustración 25 Distribución de Scores de Riesgo (IA)



La distribución de puntajes de anomalía mostró una diferenciación clara entre el universo general de movimientos y el subconjunto de casos priorizados como riesgo. De manera complementaria, la visualización conjunta de perfiles y puntajes permitió observar que el modelo no solo identificó valores extremos aislados, sino patrones inusuales dentro de contextos transaccionales específicos. Este resultado es relevante porque respalda metodológicamente el uso del enfoque híbrido: la anomalía no se interpreta únicamente como un dato extremo, sino como una desviación respecto a un perfil de comportamiento.

Ilustración 26 Perfiles Transaccionales vs Score de Riesgo



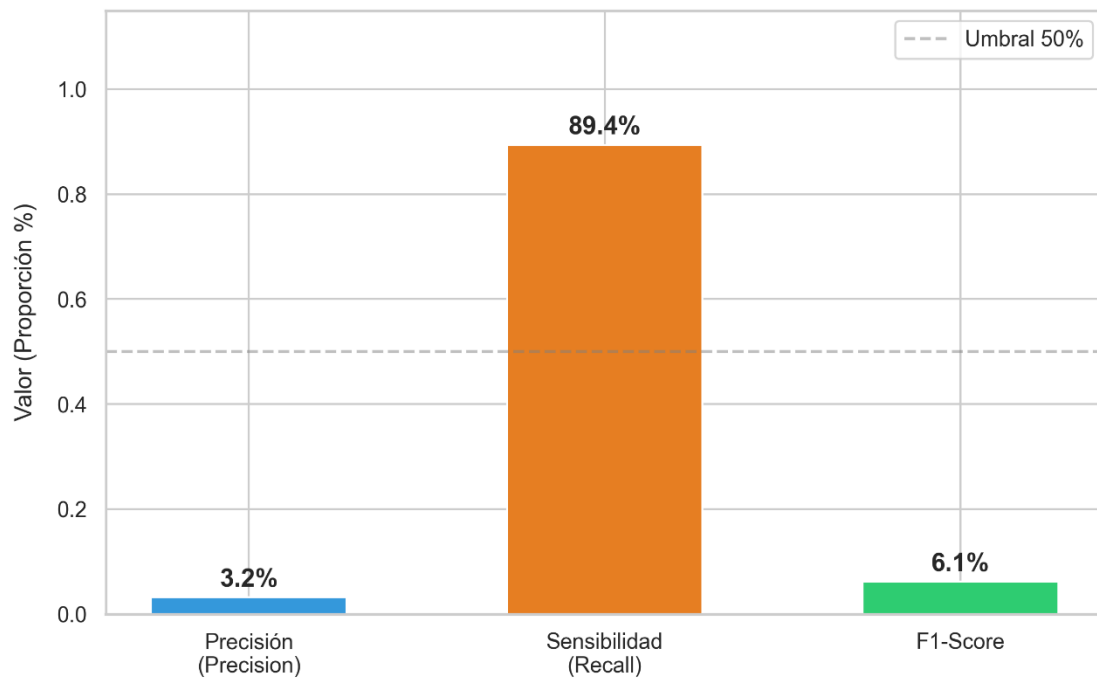
Al contrastar sus resultados frente a la base de movimientos previamente identificados por revisión humana para 2018 y 2019, el modelo híbrido detectó 242.067 movimientos, frente a 8.656 riesgos reales en el archivo de contraste. El cruce permitió identificar 7.737 verdaderos positivos, 234.330 falsos positivos y 919 falsos negativos. A partir de ello, el modelo alcanzó una *precisión* de 3,20%, un *recall* de 89,38% y un *F1-score* de 6,17%.

Ilustración 27 Matriz de Confusión IA Híbrida



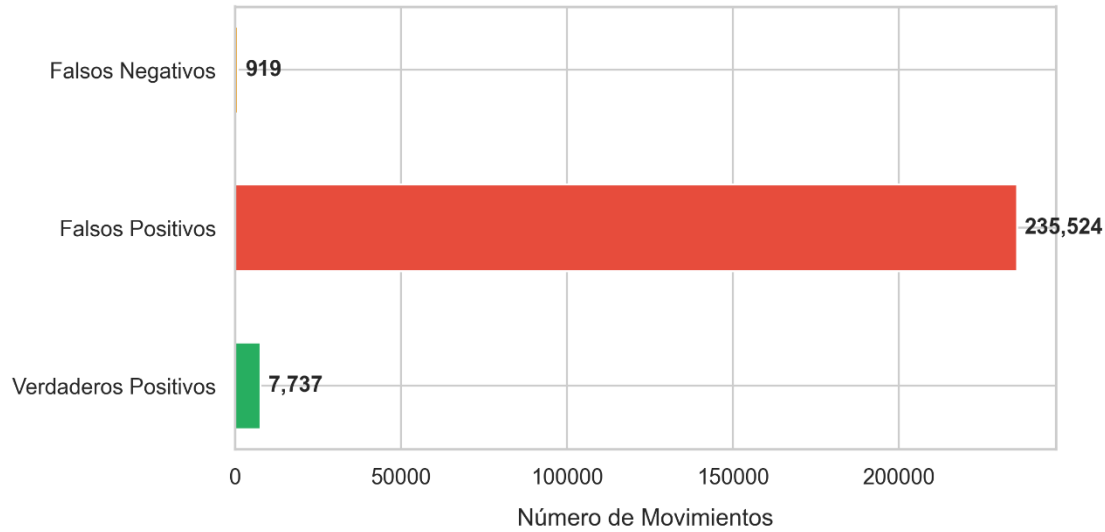
El principal resultado de este contraste fue la mejora en la sensibilidad frente al modelo basado en reglas. El *recall* del enfoque híbrido superó el 89%, lo que indica una mayor capacidad para recuperar riesgos reales y, en particular, una menor cantidad de falsos negativos. Este hallazgo es coherente con la justificación técnica del uso de “*Isolation Forest*”, cuya lógica de aislamiento resulta especialmente adecuada para contextos multivariados, ruidosos y de gran volumen, como los movimientos financieros del SGP. A ello se suma la ventaja de haber incorporado previamente una fase de agrupamiento, lo que permitió detectar anomalías dentro de perfiles transaccionales y no solo sobre el universo completo de movimientos.

Ilustración 28 Métricas de Eficacia - IA Híbrida



No obstante, al igual que en el modelo por reglas, la precisión se mantuvo en niveles bajos. Este comportamiento no invalida el uso del modelo, sino que refleja la naturaleza del problema analizado: un universo masivo de transacciones donde los riesgos reales constituyen una fracción muy pequeña del total. En este tipo de dominios, el desempeño debe interpretarse considerando el desbalance extremo entre clases y el hecho de que el archivo de contraste no necesariamente agota todos los riesgos existentes, sino únicamente aquellos ya identificados por revisión humana. En consecuencia, parte de los falsos positivos del modelo podrían corresponder, en realidad, a movimientos no detectados previamente por el proceso manual de contraste.

Ilustración 29 Composición de Resultados - IA Híbrida



En conjunto, los resultados muestran que el modelo híbrido ofreció una mejora moderada frente al modelo por reglas en términos de recuperación de riesgos, con una reducción de falsos negativos y un *F1-score* ligeramente superior. Esto permite concluir que ambos enfoques son complementarios: el modelo basado en reglas aporta explicabilidad directa y trazabilidad normativa, mientras que el modelo híbrido amplía la cobertura de detección al identificar patrones menos evidentes desde una lógica estadística y contextual.

Tabla 5 Comparación de resultados entre el modelo basado en reglas y el modelo híbrido de aprendizaje automático

Modelo	Movimi entos detecta dos	Riesgos reales	Verdad eros positiv os	Falsos positiv os	Falsos negativ os	<i>Precisió n</i>	<i>Recall</i>	<i>F1- score</i>
Reglas de negocio “A1- A5”	232.371	8.664	7.347	225.024	1.317	3,16%	84,80%	6,10%

IA híbrida "Isolation Forest + KMeans"	242.067	8.656	7.737	234.330	919	3,20%	89,38%	6,17%
--	---------	-------	-------	---------	-----	-------	--------	-------

En conjunto, los resultados obtenidos muestran que ambos modelos cumplieron una función útil como mecanismos de priorización analítica sobre movimientos potencialmente riesgosos. Mientras el enfoque basado en reglas ofreció una lectura directa y auditable de los criterios activados, el enfoque híbrido de aprendizaje automático mostró una mayor capacidad de recuperación de riesgos en el contraste disponible. Esta complementariedad refuerza la utilidad del sistema propuesto como apoyo al seguimiento y control del uso de los recursos del SGP.

4.6.5 Estrategia de tratamiento de falsos positivos

En el contexto de este trabajo, la generación de falsos positivos no se interpreta como un fallo aislado del modelo, sino como una consecuencia esperable de un enfoque orientado a la priorización de riesgos dentro de un universo masivo de transacciones. Tanto el modelo basado en reglas como el modelo híbrido fueron diseñados para maximizar la recuperación de casos potencialmente relevantes, incluso a costa de producir alertas que posteriormente puedan descartarse en revisión. Esta decisión responde a la lógica del control fiscal y del seguimiento preventivo, en la que resulta más costoso omitir un evento verdaderamente riesgoso que revisar una alerta adicional sin hallazgo concluyente.

En consecuencia, lo que se propone hacer con los falsos positivos no es eliminarlos completamente en esta fase, sino gestionarlos de manera estructurada. Para ello, las alertas generadas por la solución deben entenderse como un mecanismo de tamizado inicial que alimenta un proceso posterior de validación analítica. En términos operativos, esto implica agrupar, clasificar y ordenar las alertas según criterios de priorización, tales como el monto involucrado, la recurrencia del patrón, el tipo de regla activada, la concentración por beneficiario, la repetición temporal del comportamiento y el sector o entidad territorial afectada. De esta forma, el volumen de alertas deja de ser un listado plano y se convierte en una bandeja de casos jerarquizada para revisión.

Bajo este esquema, la revisión humana posterior cumple una función central. La solución desarrollada no reemplaza el juicio técnico de los equipos de monitoreo, seguimiento y control, sino que les entrega evidencia preliminar auditable para focalizar esfuerzos. Cada alerta debe poder ser contrastada con el contexto del movimiento, la normativa aplicable, la trazabilidad del beneficiario y, cuando sea necesario, con soportes adicionales o cruces complementarios. Así, el falso positivo deja de verse únicamente como "error" y pasa a ser parte del costo operativo razonable de una estrategia de vigilancia temprana.

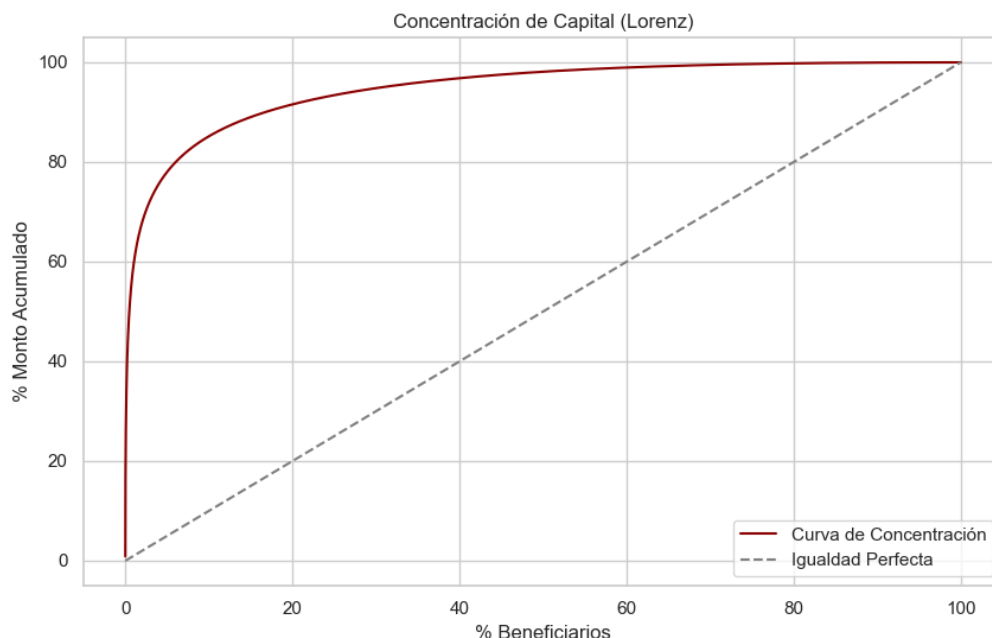
Finalmente, se propone que los resultados de esa revisión posterior alimenten un ciclo de mejora continua. Las alertas descartadas de forma recurrente pueden utilizarse para ajustar umbrales, redefinir reglas, incorporar nuevas variables o recalibrar los criterios del modelo híbrido. Con ello, el tratamiento de falsos positivos no solo reduce carga operativa en iteraciones futuras, sino que fortalece progresivamente la precisión práctica de la solución sin sacrificar su capacidad de detección.

4.6.6 Resultados de la herramienta de concentración de mercado

La herramienta de concentración de mercado se aplicó sobre los movimientos de egreso registrados en *"FACT_MOVIMIENTOS_CM"*, con el propósito de identificar beneficiarios que, por su nivel de recepción de recursos, frecuencia transaccional y diversidad territorial de origen de los recursos, pudieran representar escenarios de concentración relevantes para análisis posterior. Para ello, se construyó un perfil agregado por beneficiario a partir de variables como *"TOTAL_RECIBIDO"*, *"N_TRANSACCIONES"*, *"N_CUENTAS_ORIGEN"*, *"N_TERRITORIOS_ORIGEN"* y *"RECURRENCIA"*, complementando el análisis descriptivo con una fase de agrupamiento mediante *"KMeans"* para segmentar patrones transaccionales.

Como resultado del proceso, se consolidó un universo de **711.540 beneficiarios únicos**, con un monto total recibido cercano a **78,88 billones de pesos colombianos**. Este volumen evidencia que la herramienta no se limitó a identificar casos extremos aislados, sino que permitió construir una visión amplia del mercado de receptores de recursos asociados a las Cuentas Maestras. Al mismo tiempo, los resultados muestran que la distribución del capital no fue homogénea entre beneficiarios, sino claramente asimétrica, lo que justificó la aplicación de métricas y visualizaciones de concentración.

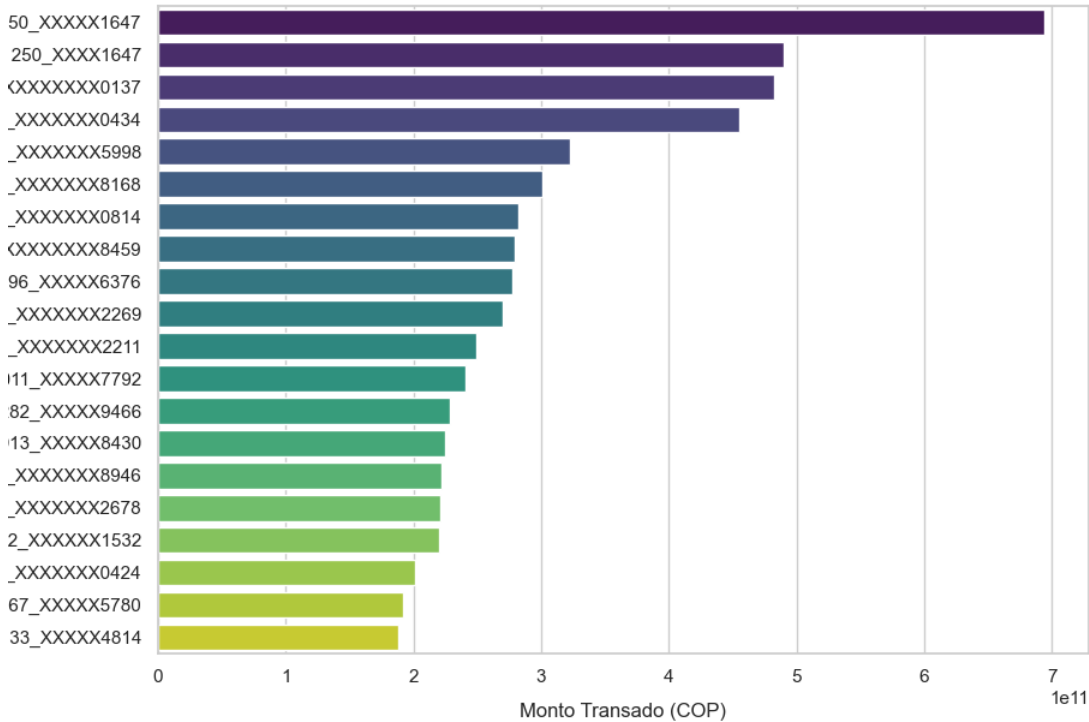
Ilustración 30 Concentración de Capital



La curva de concentración muestra una desviación pronunciada respecto a la línea de igualdad perfecta, lo que permite concluir que existe una concentración materialmente alta en la distribución de recursos. Esta lectura se refuerza con los resultados acumulados: el **1% de los beneficiarios** concentró aproximadamente **59,57%** del monto total recibido, el **5%** acumuló cerca de **78,02%**, y el **10%** alcanzó **85,17%** del total. En otras palabras, una fracción relativamente reducida de beneficiarios absorbió la mayor parte de los recursos analizados, mientras que el resto participó en proporciones significativamente menores. Desde el punto de vista analítico, este comportamiento es consistente con mercados altamente concentrados, donde un número pequeño de actores domina una porción sustancial del flujo financiero.

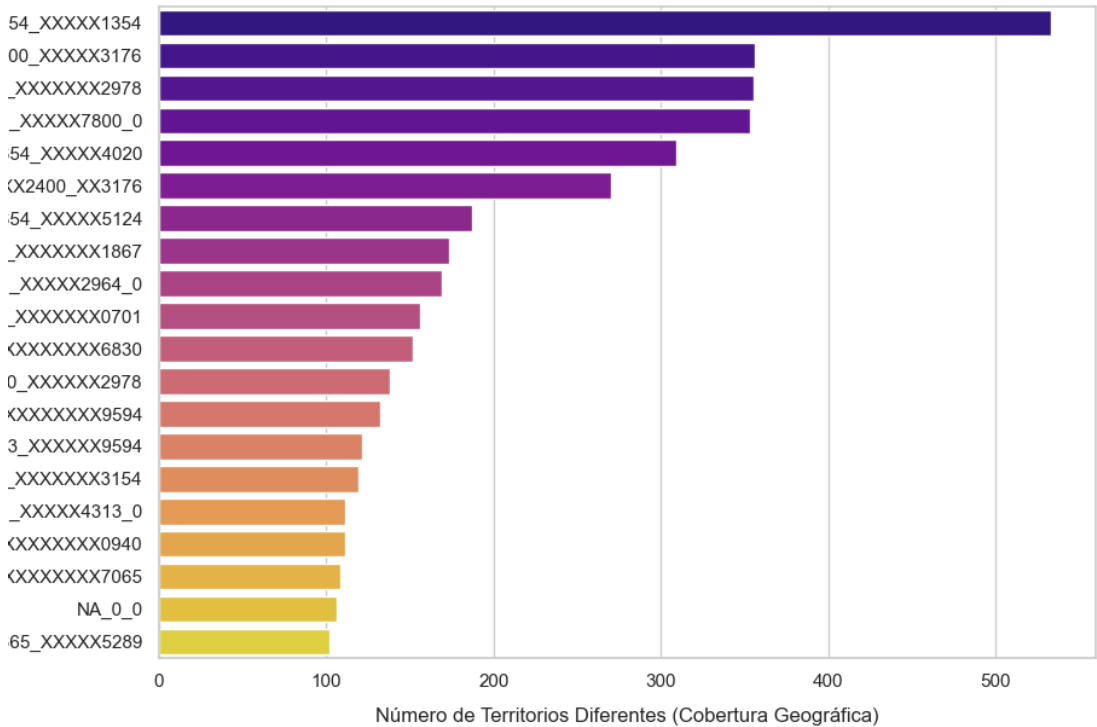
Al observar los beneficiarios con mayor monto total recibido, se evidencia que los primeros lugares del ranking concentraron valores de magnitud muy superior al resto de casos. El beneficiario ubicado en la primera posición acumuló alrededor de **694,3 mil millones de pesos**, mientras que los veinte primeros beneficiarios concentraron en conjunto aproximadamente **7,66%** del total general. Este resultado refuerza la existencia de actores dominantes y sugiere la necesidad de revisar, desde el punto de vista contractual y operativo, las razones por las cuales ciertos beneficiarios concentran valores tan elevados frente al resto del universo analizado.

Ilustración 31 Top 20 Beneficiarios por Monto Total Recibido



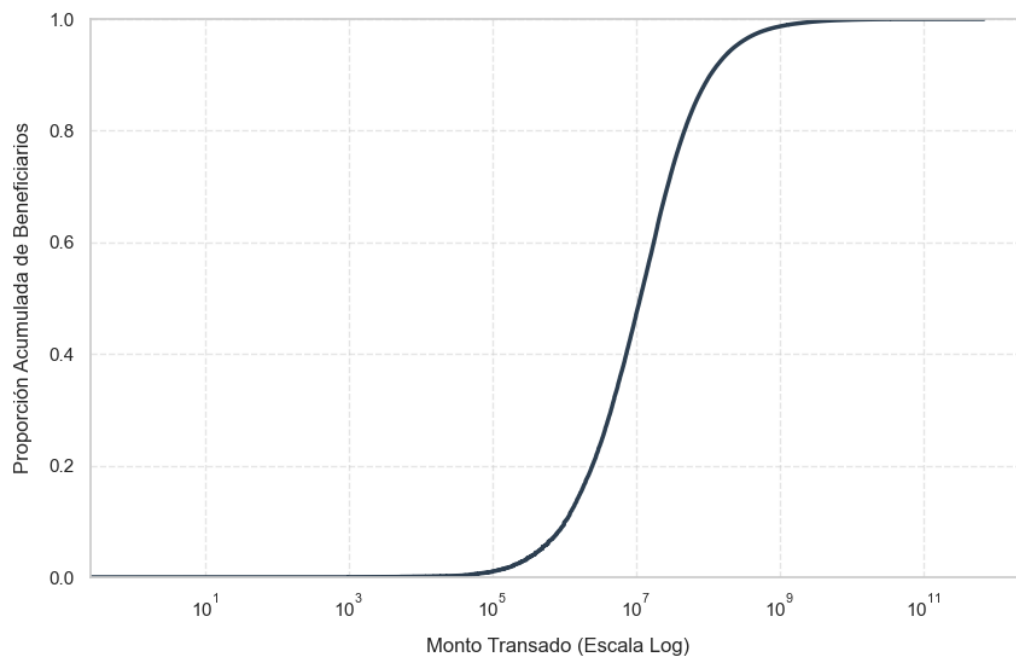
No obstante, la concentración no se expresa únicamente en términos de monto. Cuando se observa la cantidad de territorios de origen desde los cuales recibe recursos cada beneficiario, aparece una segunda dimensión del fenómeno: algunos actores no solo concentran montos altos, sino que también reciben recursos desde una cobertura territorial amplia. El caso más extremo alcanzó **532 territorios de origen**, mientras que otros beneficiarios del grupo superior superaron los **300 territorios**. Este resultado es especialmente relevante porque sugiere patrones de recepción extendidos territorialmente, lo que puede asociarse a beneficiarios con alta capilaridad operativa o, alternatively, a casos que ameritan revisión por su presencia recurrente en múltiples territorios.

Ilustración 32 Top 20 Beneficiarios por Cantidad de Territorios de origen



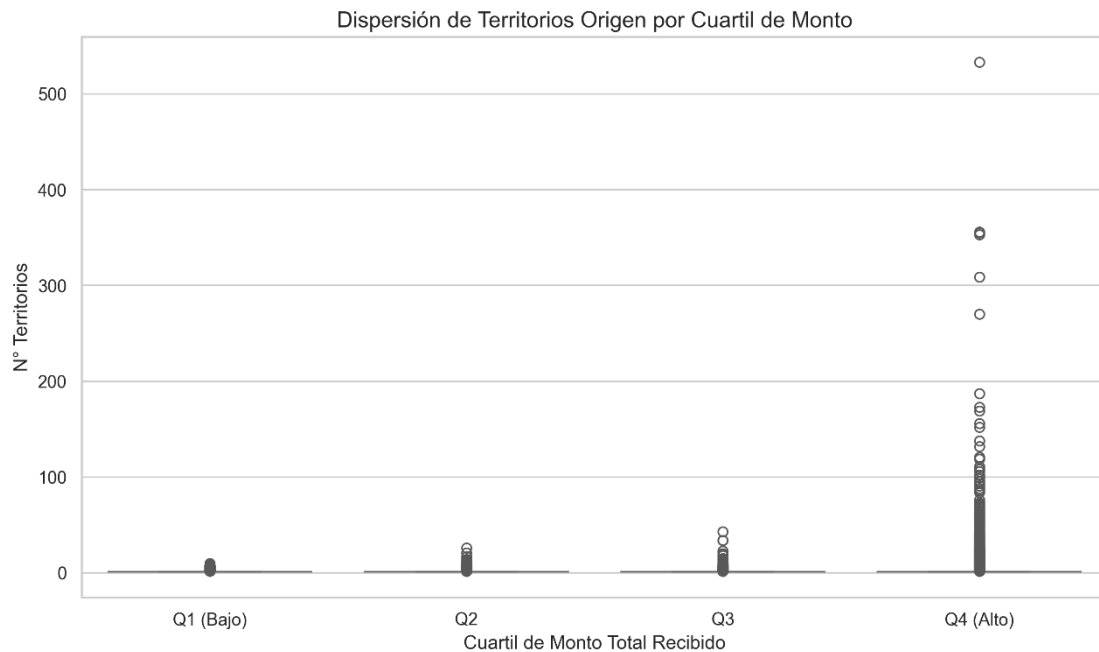
La distribución acumulada de montos recibidos confirma además que la mayoría de los beneficiarios se concentra en niveles relativamente bajos de recepción, mientras que una proporción muy reducida alcanza montos extraordinariamente altos. Las medidas descriptivas refuerzan esta lectura: la mediana del monto total recibido fue cercana a **11,5 millones de pesos**, el percentil 90 se ubicó en torno a **108,9 millones**, el percentil 99 en **1,37 mil millones** y el percentil 99,9 en **12,29 mil millones**. Esta distancia entre la mediana y los percentiles superiores evidencia una distribución fuertemente sesgada hacia la derecha, propia de fenómenos de concentración económica.

Ilustración 33 Probabilidad Acumulada de Montos Recibidos



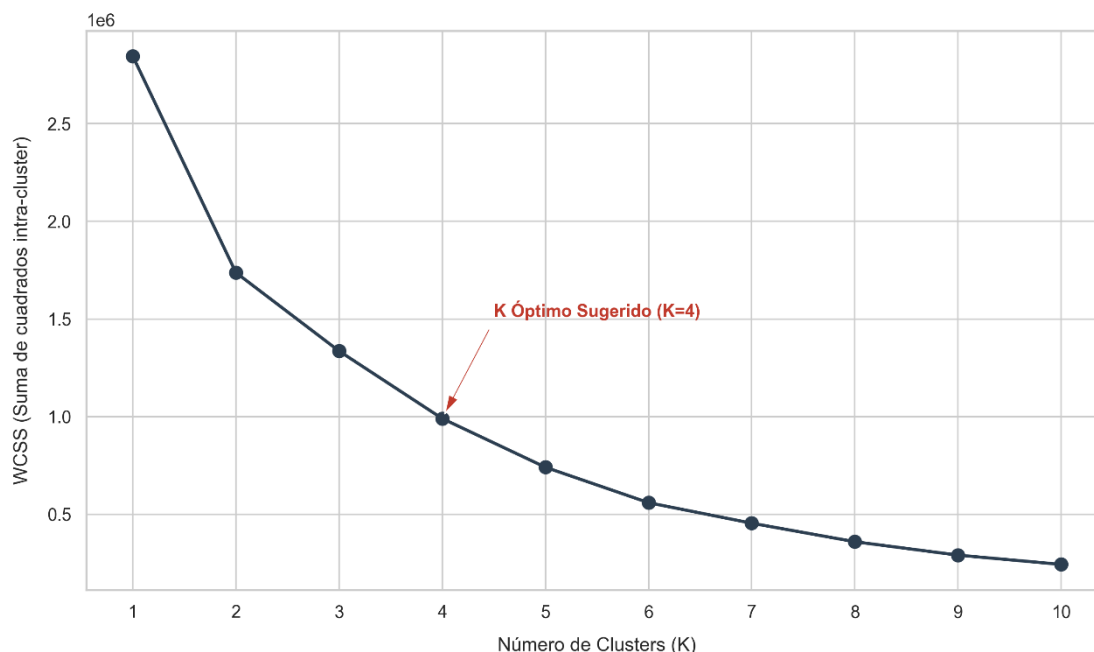
La relación entre monto recibido y diversidad territorial también mostró un comportamiento diferencial por rangos. Al agrupar los beneficiarios por cuartiles de monto, se observó que el cuartil superior presentó una dispersión territorial mucho más amplia que los grupos inferiores, incluyendo numerosos valores atípicos y casos extremos. Esto sugiere que los beneficiarios con mayores montos no solo concentran más recursos, sino que además tienden a participar en redes de recepción más amplias, lo que incrementa su relevancia analítica dentro del estudio de concentración de mercado.

Ilustración 34 Dispersión de Territorios de Origen por Cuartil de Monto



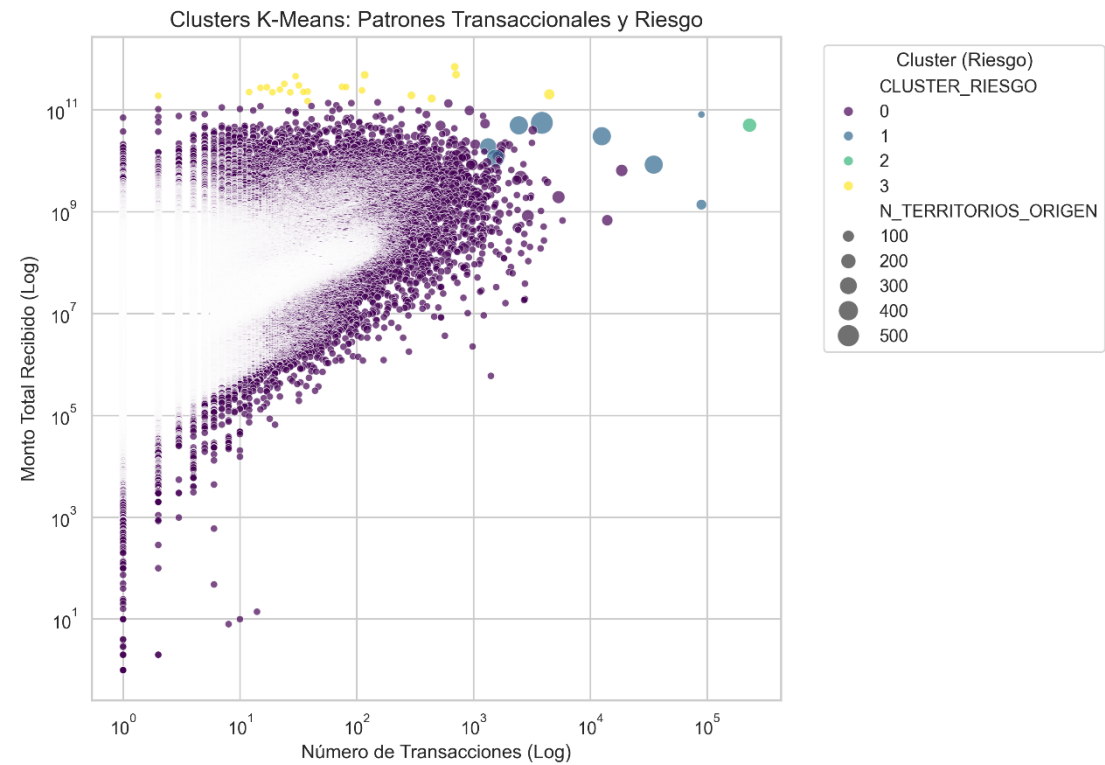
Con el fin de complementar el análisis descriptivo, se aplicó una segmentación mediante “KMeans” sobre las variables “TOTAL_RECIBIDO”, “N_TRANSACCIONES”, “N_TERRITORIOS_ORIGEN” y “RECURRENCIA”, utilizando **cuatro clústeres**, de acuerdo con el punto de inflexión sugerido por el método del codo. Esta fase permitió identificar perfiles diferenciados de beneficiarios y no solo ordenar casos por monto. El resultado más importante fue que la mayor parte del universo quedó concentrada en el clúster de comportamiento ordinario, mientras que un grupo muy reducido agrupó los casos de mayor monto y mayor intensidad analítica.

Ilustración 35 Método del Codo - Selección Óptima de Clústeres



La segmentación final mostró que el clúster “0” agrupó **711.509 beneficiarios**, equivalentes a prácticamente la totalidad del universo, y concentró cerca de **91,55%** del monto total. Aunque a primera vista esta participación puede parecer contradictoria, lo que refleja en realidad es que el clúster dominante reúne tanto una gran masa de beneficiarios de comportamiento ordinario como una parte importante del flujo general. En contraste, el clúster “3” reunió solo **22 beneficiarios**, pero concentró aproximadamente **8,06%** del monto total, con valores medios muy superiores al resto. A ello se suman el clúster “1”, con **8 beneficiarios** y alta dispersión territorial, y el clúster “2”, con **1 beneficiario** de comportamiento singular. Esta distribución confirma que los casos más relevantes de concentración pueden identificarse como grupos muy pequeños, pero altamente diferenciados en términos de monto, frecuencia o cobertura territorial.

Ilustración 36 Clústeres K-Means: Patrones Transaccionales y Riesgo



En conjunto, los resultados de esta herramienta permiten afirmar que el mercado de beneficiarios receptores de recursos del SGP, observado a través de las Cuentas Maestras, presenta una estructura marcadamente concentrada. La utilidad principal del modelo no radica en afirmar automáticamente la existencia de una irregularidad, sino en priorizar beneficiarios y patrones de recepción que merecen revisión adicional, especialmente cuando combinan montos elevados, alta frecuencia de transacciones y diversidad territorial de origen. Desde esta perspectiva, la herramienta aporta un insumo analítico complementario a los modelos de riesgo, al ampliar la lectura desde la anomalía transaccional hacia la estructura de concentración en la circulación de los recursos.

Tabla 6 Resumen de resultados de la herramienta de concentración de mercado

Indicador	Resultado
Beneficiarios únicos analizados	711.540
Monto total recibido	78,88 billones COP
Participación del beneficiario con mayor monto	0,88%
Participación acumulada del top 20 beneficiarios	7,66%

Participación acumulada del 1% superior de beneficiarios	59,57%
Participación acumulada del 10% superior de beneficiarios	85,17%
Mediana de monto recibido	11,5 millones COP
Percentil 99 del monto recibido	1,37 mil millones COP
Máximo número de territorios de origen observado	532
Número de clústeres utilizados en “KMeans”	4

En conjunto, los resultados obtenidos evidencian que la solución desarrollada logró avanzar desde la consolidación técnica de un modelo relacional y un proceso ETL reproducible hasta la construcción de herramientas analíticas aplicables al seguimiento del uso de los recursos del SGP. Mientras los modelos de riesgo permitieron priorizar movimientos potencialmente problemáticos desde enfoques complementarios —reglas de negocio y aprendizaje automático—, la herramienta de concentración de mercado aportó una lectura estructural sobre la distribución de los recursos entre beneficiarios. Esta articulación entre integración de datos y análisis constituye uno de los principales aportes del proyecto.

Capítulo 5. DISCUSIÓN

La metodología adoptada permitió integrar de forma coherente el modelado relacional, el proceso ETL y las herramientas analíticas en una misma solución. No obstante, su desarrollo requirió ajustes técnicos y metodológicos frente a la realidad de los datos, la disponibilidad normativa y las condiciones de validación. En este contexto, los resultados deben interpretarse como apoyo al monitoreo y a la priorización analítica, más que como mecanismos de decisión automática.

5.1 Discusión sobre la utilidad de la metodología adoptada

La metodología adoptada resultó adecuada para abordar un problema caracterizado por complejidad normativa, heterogeneidad estructural de los datos y necesidad de análisis posterior sobre grandes volúmenes de movimientos financieros. La decisión de estructurar la solución en tres componentes articulados (modelo relacional, proceso ETL y herramientas analíticas) permitió avanzar de manera progresiva desde la organización de la información hasta su explotación con fines de seguimiento y control.

En primer lugar, el modelo relacional permitió transformar estructuras normativas dispersas en una base de datos unificada, consistente y apta para consultas reproducibles. Su utilidad no fue solo técnica, sino también funcional respecto del problema público abordado, ya que la trazabilidad de los recursos del SGP exige una organización de datos que preserve integridad, facilite cruces y soporte reglas de validación.

En segundo lugar, el proceso ETL demostró ser un componente central de la solución. Su aporte no consistió únicamente en cargar información, sino en homologar archivos históricos con estructuras distintas, preservar la integridad de códigos y cuentas, y dejar evidencia suficiente para trazabilidad y revisión. Esto confirma que, en este tipo de proyectos, el valor analítico posterior depende en gran medida de la calidad y consistencia logradas durante la integración de los datos.

En tercer lugar, las herramientas analíticas desarrolladas mostraron que la combinación entre conocimiento de negocio y técnicas computacionales constituye una estrategia útil para apoyar el seguimiento del uso de recursos públicos. El modelo basado en reglas aportó explicabilidad y trazabilidad, el modelo híbrido amplió la cobertura de detección sobre patrones menos evidentes y la herramienta de concentración de mercado permitió complementar la lectura del problema desde una dimensión estructural.

En conjunto, la metodología fue útil porque permitió abordar el problema desde una lógica integral: primero organizar la información, luego validarla y finalmente analizarla. Esta secuencia evitó construir herramientas analíticas sobre datos no consolidados y permitió que cada componente reforzara al siguiente dentro de una arquitectura coherente.

5.2 Adaptaciones realizadas durante el desarrollo del proyecto

Aunque la metodología general se mantuvo, durante el desarrollo del proyecto fue necesario realizar varias adaptaciones respecto al planteamiento inicial. Estas modificaciones no alteraron el objetivo general, pero sí permitieron ajustar la solución a las condiciones reales de los datos, a las restricciones normativas disponibles y a los hallazgos encontrados durante la implementación.

Uno de los primeros ajustes se produjo en el diseño del modelo de datos. Aunque el objetivo inicial era construir una estructura unificada para Cuentas Maestras y Cuentas Maestras Pagadoras, el trabajo práctico mostró que algunas relaciones previstas desde el diseño conceptual no podían mantenerse de forma estricta en el plano físico sin generar inconsistencias o rigidez excesiva. Esto obligó a revisar ciertas claves foráneas, redefinir criterios de unicidad y adoptar decisiones pragmáticas para preservar la funcionalidad del modelo sin perder coherencia relacional.

Una segunda adaptación relevante surgió durante el desarrollo del ETL. Inicialmente se contemplaba una carga relativamente directa desde las estructuras normativas hacia las tablas del modelo; sin embargo, en la práctica fue necesario incorporar procesos complementarios de saneamiento y backfill, como ocurrió con la identificación de cuentas no registradas que habían sido válidas en algún momento del periodo analizado. Esto mostró que el problema no era solo de transformación técnica, sino también de reconstrucción histórica del inventario de cuentas.

También fue necesario adaptar el proyecto frente a la disponibilidad efectiva de información. Aunque la solución fue diseñada para soportar tanto las estructuras históricas como la estructura unificada prevista en la Resolución 2394 de 2023, en la práctica el análisis se realizó únicamente con estructuras históricas, dado que la nueva estructura aún no se encontraba operativa dentro del aplicativo SISFEVAL. Esto confirma que se optó por una arquitectura con capacidad de adaptación futura, aunque los resultados empíricos correspondan al universo de datos efectivamente disponible durante su ejecución.

En el caso de las herramientas analíticas, las adaptaciones más importantes estuvieron relacionadas con la calibración de umbrales, filtros y criterios de evaluación. El modelo basado en reglas requirió ajustes para reducir ruido y concentrar la detección en movimientos con mayor relevancia analítica. De forma similar, el modelo híbrido fue orientado a privilegiar la recuperación de riesgos potenciales, asumiendo que en un entorno de auditoría resulta preferible revisar alertas adicionales a omitir eventos relevantes. En ese sentido, el proyecto se desarrolló como un proceso iterativo de ajuste entre diseño, datos y propósito analítico.

5.3 Limitaciones del estudio y de la solución desarrollada

A pesar de los resultados obtenidos, el trabajo presenta limitaciones que deben reconocerse explícitamente. La primera se relaciona con la disponibilidad y cobertura de la información. Aunque se procesó un volumen amplio de archivos entre 2018 y 2025, el proyecto trabajó con estructuras históricas de reporte y no con la nueva estructura unificada prevista por la Resolución 2394 de 2023, ya que esta aún no se encontraba implementada operativamente al momento del análisis.

La segunda limitación corresponde a la calidad de los datos fuente. Parte de los errores detectados en el ETL provinieron de inconsistencias atribuibles a los establecimientos bancarios que reportan la información, incluyendo registros asociados a cuentas no válidas normativamente o campos incompletos en ciertos movimientos. Aunque el proceso ETL permitió controlar, filtrar y documentar estas incidencias, la solución sigue dependiendo de la calidad del dato reportado en origen.

Una tercera limitación corresponde a la construcción del contraste utilizado para evaluar los modelos de riesgo. Tanto el modelo basado en reglas como el modelo híbrido fueron comparados contra una base de movimientos previamente identificados por revisión humana para 2018 y 2019. Esta estrategia fue útil para contar con una referencia empírica, pero no debe interpretarse como una representación exhaustiva del universo de riesgos existente. En consecuencia, parte de los falsos positivos identificados por los modelos puede corresponder a movimientos potencialmente relevantes que no habían sido incorporados en la muestra de contraste.

De igual forma, las herramientas analíticas desarrolladas no constituyen mecanismos de decisión automática. El modelo de reglas identifica alertas explicables con base en criterios definidos, mientras que el modelo híbrido detecta anomalías estadísticas o contextuales. Ninguno de los dos, por sí mismo, prueba de manera concluyente la existencia de una irregularidad administrativa, fiscal o disciplinaria. Su función debe entenderse como apoyo a la priorización analítica y a la focalización de revisiones posteriores.

Finalmente, la herramienta de concentración de mercado debe interpretarse dentro de su propósito exploratorio. La identificación de beneficiarios con alta participación o amplia cobertura territorial no implica, por sí sola, una conducta irregular, sino una lectura estructural sobre la distribución de los recursos útil para orientar preguntas de auditoría.

5.3.1 Límite de interpretación

En este sentido, los resultados del proyecto deben entenderse como apoyo analítico para orientar revisiones y priorizar riesgos, pero no como evidencia concluyente ni como sustituto del análisis institucional posterior. Esta limitación no invalida la solución, pero sí exige definir explícitamente cómo deben interpretarse y gestionarse los falsos positivos generados por los modelos.

5.4 Implicaciones metodológicas de los falsos positivos

Uno de los aspectos metodológicos más relevantes del presente trabajo es la elevada proporción de falsos positivos obtenida en los ejercicios de contraste. Lejos de ocultar esta situación, conviene interpretarla a la luz del problema abordado. El dominio analizado presenta un fuerte desbalance entre el volumen total de transacciones y la cantidad de casos previamente identificados como riesgos reales, lo que hace improbable alcanzar niveles altos de precisión sin sacrificar de manera significativa la sensibilidad del sistema. En este marco, una lectura aislada de la precisión puede conducir a conclusiones incompletas sobre la utilidad de la solución.

Desde una perspectiva aplicada, el sentido de la herramienta no es sancionatorio ni decisorio, sino preventivo y analítico. Su función principal consiste en reducir el espacio de búsqueda dentro de millones de registros y señalar subconjuntos que merecen revisión posterior. Por ello, el falso positivo debe analizarse como parte de una estrategia de priorización y no como una clasificación definitiva errónea en términos absolutos. Incluso, dado que el archivo de contraste proviene de una identificación humana previa y no agota necesariamente el universo real de riesgos, es posible que una parte de los falsos positivos corresponda en realidad a eventos que no habían sido detectados en el proceso manual de referencia.

Esta situación tiene dos implicaciones. La primera es que la utilidad de la solución depende de su integración con un proceso institucional de validación posterior, en el que las alertas sean revisadas por analistas con criterios normativos y conocimiento del contexto. La segunda es que la mejora del desempeño no debe plantearse únicamente como una reducción mecánica del número de falsos positivos, sino como una optimización del equilibrio entre cobertura, explicabilidad y carga operativa. En otras palabras, un sistema de alertas útil no es necesariamente el que menos alertas genera, sino el que permite identificar con oportunidad los casos más relevantes para revisión.

En este sentido, los falsos positivos también cumplen una función informativa dentro del proyecto. Su análisis permite detectar reglas demasiado amplias, perfiles transaccionales ruidosos, variables insuficientemente discriminantes o comportamientos fronterizos entre lo permitido y lo atípico. Por tanto, más que un residuo indeseable, constituyen una fuente de aprendizaje para el refinamiento del modelo y para la definición de futuras líneas de trabajo orientadas a segmentación de alertas, aprendizaje supervisado y enriquecimiento de variables contextuales.

5.5 Alcance e impacto del resultado obtenido

El principal aporte del trabajo consiste en haber demostrado que es posible diseñar e implementar una solución integral para organizar, validar y analizar la información reportada por las Cuentas Maestras y Cuentas Maestras Pagadoras del SGP. El resultado obtenido no se limita a la construcción de una base de datos, sino que articula una arquitectura funcional que convierte archivos aislados y heterogéneos en insumos trazables para el análisis técnico y la priorización de riesgos.

Desde el punto de vista práctico, el impacto más importante de la solución es la reducción de la dependencia de procesos manuales o dispersos para la consulta y análisis de la información. La consolidación del modelo relacional y la automatización del ETL permiten contar con una estructura más estable para consulta histórica, cruces entre entidades, análisis de movimientos y desarrollo de nuevas reglas o modelos analíticos.

En el plano analítico, el proyecto mostró que la combinación entre reglas de negocio, aprendizaje automático y análisis de concentración puede aportar valor real al seguimiento del uso de los recursos públicos. El enfoque por reglas ofreció trazabilidad y explicación, el modelo

híbrido amplió la recuperación de riesgos y la herramienta de concentración permitió visibilizar patrones estructurales de distribución entre beneficiarios. Esta complementariedad constituye uno de los aportes más relevantes del trabajo, ya que evita depender de un único enfoque para abordar un problema que es, por naturaleza, multidimensional.

En términos institucionales, el trabajo también tiene impacto potencial como base para desarrollos futuros dentro de estrategias de monitoreo, seguimiento y control. La solución puede servir como punto de partida para fortalecer procesos de validación, generar tableros analíticos, incorporar nuevas fuentes de información o ampliar el análisis hacia otras dimensiones del uso de los recursos del SGP.

En síntesis, el impacto del proyecto radica en haber convertido un problema de dispersión y heterogeneidad de datos en una solución estructurada de integración y análisis. Aun con las limitaciones señaladas, los resultados muestran que el trabajo aporta una base técnica y metodológica sólida para mejorar la trazabilidad, la consulta y la identificación de riesgos en el uso de recursos públicos administrados a través de Cuentas Maestras en Colombia.

Impacto principal del proyecto

El valor más relevante de la solución desarrollada radica en haber convertido información históricamente dispersa y heterogénea en una plataforma estructurada, trazable y analíticamente explotable para fortalecer el seguimiento del uso de los recursos del SGP.

5.5.1 Impacto económico de la solución y ahorro potencial asociado a la detección temprana

Además de su utilidad técnica y metodológica, el impacto de la solución también puede evaluarse desde una perspectiva económica y operativa. En el escenario actual, una revisión manual exhaustiva de la información mensual resulta materialmente inviable. Para un solo mes, el analista debe primero identificar las cuentas incluidas en los archivos reportados por las entidades bancarias, lo que puede tomar entre 10 y 15 minutos, y posteriormente realizar cruces adicionales para establecer el monto de recursos asignados a cada cuenta, proceso que puede requerir entre 5 y 10 minutos adicionales. A ello se suma la revisión de movimientos por cada cuenta, que puede tardar entre 2 y 5 minutos por cuenta. Considerando que el universo mensual se sitúa aproximadamente entre 8.000 y 10.000 cuentas, una revisión exhaustiva implicaría entre 267 y 834 horas de trabajo por mes, sin contar tiempos adicionales de documentación, contraste y escalamiento de hallazgos.

Tomando como referencia un perfil técnico con formación en economía, contaduría, administración pública o áreas afines, y un costo estimado de **35.000 COP por hora**, equivalente aproximadamente a entre **5.600.000 y 6.200.000 COP mensuales** según una jornada laboral de 40 a 44 horas semanales, la revisión manual exhaustiva de **un solo mes** podría representar un costo aproximado de entre **9.300.000 y 29.200.000 COP**. En contraste, la solución desarrollada no pretende reemplazar la revisión humana, sino reducir el espacio de búsqueda y transformar un universo masivo de transacciones en una bandeja priorizada de alertas para análisis posterior. Bajo esta lógica, incluso cuando persistan falsos positivos, el costo total del proceso sigue siendo sensiblemente menor al de una revisión manual global, y su atención resulta operativamente asumible con los recursos existentes.

Desde esta perspectiva, el valor económico de la solución no depende únicamente del ahorro de tiempo analítico, sino también de su capacidad para anticipar casos cuyo monto justifica una revisión inmediata. Esto puede ilustrarse con ejemplos anonimizados provenientes de la base de contraste utilizada en el proyecto, la cual contiene movimientos previamente señalados como posibles riesgos. En un primer caso, correspondiente a un municipio que para febrero de 2019 se identificaron tres movimientos marcados como posibles riesgos por valores de 192.908.923 COP, 9.067.253 COP y 151.183.066 COP, para un total de 353.159.242 COP en un solo corte mensual. En un segundo caso, en diciembre de 2018 se registraron movimientos señalados como posibles riesgos por 282.887.330 COP, 141.443.665 COP, 13.157.550 COP y 29.604.489 COP, para un total acumulado de 467.093.034 COP en el mismo periodo.

Estos ejemplos no deben interpretarse automáticamente como fraudes probados, sino como situaciones de riesgo que, de haber sido detectadas de forma temprana, habrían podido activar revisiones oportunas, medidas preventivas o correctivas y decisiones orientadas a proteger recursos públicos por cuantías materialmente relevantes. Además, debe tenerse en cuenta que estos ejemplos corresponden únicamente a un mes de información y a dos municipios dentro de un universo aproximado de 300 identificados con alertas de este tipo. En consecuencia, incluso si una parte de las alertas generadas por el sistema termina siendo descartada tras la validación humana, la solución mantiene rentabilidad práctica e institucional si permite focalizar el esfuerzo analítico y acelerar la identificación de casos que involucran cientos de millones de pesos. Bajo esta lógica, el costo de revisar falsos positivos constituye una carga razonable frente al beneficio potencial de reducir tiempos de análisis y fortalecer la protección preventiva del erario.

Capítulo 6. CONCLUSIONES

Conclusiones clave del TFM

- Se diseñó e implementó una base de datos relacional unificada para Cuentas Maestras y Cuentas Maestras Pagadoras.
- Se desarrollaron procesos ETL en Python para integrar y validar información histórica del periodo 2018–2025.
- Se construyeron herramientas analíticas para detección de riesgos y análisis de concentración de mercado.
- La solución mejora la trazabilidad, la calidad del dato y el apoyo al monitoreo del uso de los recursos del SGP.

6.1 Conclusiones del trabajo

El presente trabajo permitió cumplir el objetivo general de diseñar y gestionar una base de datos relacional unificada para las Cuentas Maestras y Cuentas Maestras Pagadoras del Sistema General de Participaciones, orientada a mejorar la trazabilidad, el seguimiento y el análisis del uso adecuado de los recursos públicos en Colombia. A partir del desarrollo realizado, fue posible transformar un conjunto de estructuras normativas dispersas en una solución integrada, compuesta por un modelo relacional, procesos ETL y herramientas analíticas aplicadas sobre los datos consolidados.

En relación con los objetivos específicos, se diseñó e implementó un modelo de datos relacional en SQL Server que permitió representar de forma consistente titulares, cuentas, beneficiarios, movimientos y tablas de referencia, preservando integridad, trazabilidad y capacidad de consulta. Asimismo, se elaboraron el diccionario de datos y el catálogo de datos como artefactos de apoyo para la gobernanza y la interpretación homogénea de la información.

De igual forma, se implementó un proceso ETL en Python capaz de integrar archivos históricos de Cuentas Maestras y Cuentas Maestras Pagadoras en formato “.txt”, homologando estructuras heterogéneas, validando datos y cargando la información al modelo relacional. Este proceso permitió poblar las tablas del modelo con información histórica del periodo 2018–2025 y consolidar una base organizada para su explotación posterior.

Sobre esta base, se desarrollaron y aplicaron herramientas analíticas orientadas a dos propósitos complementarios. Por una parte, se construyeron modelos para la identificación de posibles movimientos de riesgo, tanto mediante reglas de negocio explicables como mediante un enfoque híbrido de aprendizaje automático. Por otra, se desarrolló una herramienta para analizar la concentración de mercado sobre beneficiarios receptores de recursos. En conjunto, estos resultados muestran que la solución no solo organiza la información, sino que la convierte en un insumo útil para la priorización analítica y el apoyo al seguimiento del uso de los recursos públicos.

En síntesis, el trabajo demuestra que es posible articular diseño de bases de datos, procesos ETL y analítica de datos en una solución aplicada al contexto del SGP. Aunque persisten limitaciones derivadas de la calidad del dato fuente, la disponibilidad normativa efectiva y la naturaleza exploratoria de algunas herramientas, el proyecto aporta una base técnica y metodológica sólida

para fortalecer la consulta, la trazabilidad y la identificación de riesgos asociados al uso de los recursos administrados a través de Cuentas Maestras.

Aporte principal del trabajo

El principal aporte del TFM consiste en demostrar que la integración de estructuras normativas heterogéneas en una base relacional unificada puede convertirse en una plataforma útil no solo para almacenar información, sino también para habilitar trazabilidad, validación y analítica aplicada al seguimiento de recursos públicos.

6.1.1 Conclusión sobre el tratamiento de falsos positivos

Como conclusión específica, este trabajo propone asumir los falsos positivos como un elemento gestionable dentro de una arquitectura de alertas orientada al control preventivo, y no como una razón para descartar la utilidad de la solución desarrollada. En problemas de seguimiento al uso de recursos públicos, donde el costo de no detectar una operación riesgosa puede ser alto en términos de trazabilidad, oportunidad de intervención y protección del erario, resulta metodológicamente defendible privilegiar la sensibilidad del sistema frente a una reducción estricta del número de alertas.

Por esta razón, la solución presentada debe entenderse como un instrumento de apoyo a la priorización de casos y no como un mecanismo automático de determinación de irregularidades. El tratamiento propuesto para los falsos positivos consiste en su clasificación, jerarquización y revisión posterior por parte de equipos técnicos, acompañadas de un proceso de retroalimentación que permita ajustar progresivamente reglas, umbrales y variables analíticas. Bajo esta lógica, el valor de la herramienta no reside únicamente en “acertar” de forma cerrada, sino en hacer visible un conjunto de movimientos que, de otro modo, permanecerían dispersos dentro de grandes volúmenes de información.

En consecuencia, la principal aportación del trabajo no es la eliminación total del error de clasificación, sino la construcción de una base de datos y de una capa analítica que permiten transformar la revisión de riesgos en un proceso más estructurado, trazable y escalable. Así, los falsos positivos dejan de ser un problema meramente estadístico y pasan a formar parte de una estrategia institucional de vigilancia, aprendizaje y mejora continua.

6.2 Conclusiones personales

El desarrollo de este trabajo tuvo una importancia especial para mí, no solo por su valor académico, sino por su relación directa con un problema real de gestión y seguimiento de recursos públicos en Colombia. Trabajar sobre información asociada al Sistema General de Participaciones me permitió vincular mi experiencia profesional con una propuesta técnica de mayor alcance, orientada no solo a organizar datos, sino también a convertirlos en una herramienta útil para el análisis y el control.

A nivel personal, este proyecto representó un proceso de aprendizaje integral. Me permitió profundizar en aspectos de diseño de bases de datos relacionales, normalización, procesos ETL y aplicación de herramientas analíticas en Python, así como comprender mejor la importancia de articular adecuadamente la dimensión técnica con la dimensión metodológica en un trabajo de investigación aplicada. También me mostró que, en proyectos de datos reales, la mayor dificultad no siempre está en el modelo analítico, sino en la calidad, estructura y contexto de la información que se debe integrar.

Otro aprendizaje importante fue confirmar que el valor de una solución de datos no depende únicamente de la tecnología empleada, sino de su capacidad para responder a una necesidad

concreta y generar utilidad práctica. En este caso, el proyecto me permitió evidenciar que una buena estructura de datos puede convertirse en el punto de partida para desarrollar mecanismos de validación, trazabilidad y priorización analítica que aporten a procesos institucionales de seguimiento y control.

Finalmente, considero que este trabajo tuvo una trascendencia significativa en mi formación, ya que consolidó conocimientos técnicos que forman parte de mi perfil profesional y, al mismo tiempo, me permitió abordar un tema con impacto potencial en la administración pública. Más allá del resultado obtenido, el proceso confirmó mi interés por seguir desarrollando soluciones de datos aplicadas a problemas complejos, especialmente en contextos donde la calidad de la información y la analítica pueden contribuir a una mejor toma de decisiones.

Aprendizaje personal principal

Este trabajo confirmó que el valor de la analítica no depende únicamente del modelo o de la tecnología utilizada, sino de la capacidad para integrar datos complejos, contextualizarlos adecuadamente y convertirlos en herramientas útiles para la toma de decisiones.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

Las futuras líneas de trabajo del presente proyecto se orientan a ampliar su madurez técnica, su robustez metodológica y su utilidad operativa. Aunque la solución desarrollada permitió integrar información históricamente fragmentada y construir herramientas analíticas funcionales, todavía existen oportunidades relevantes para fortalecer su alcance, mejorar su precisión y facilitar su adopción en escenarios reales de monitoreo, seguimiento y control.

7.1 Integración de la estructura unificada prevista en la Resolución 2394 de 2023

Una primera línea de trabajo consiste en incorporar de manera operativa la estructura unificada de reporte prevista en la Resolución 2394 de 2023, una vez esta se encuentre plenamente habilitada en el aplicativo SISFEVAL. Aunque en el presente trabajo la validación empírica se realizó sobre estructuras históricas, tanto el modelo de datos como los procesos ETL fueron diseñados con la flexibilidad suficiente para soportar también dicho esquema.

En consecuencia, un desarrollo futuro natural será ejecutar pruebas reales de carga sobre esta nueva estructura, validar los ajustes técnicos requeridos y documentar su comportamiento frente a la arquitectura ya implementada. Esta evolución permitiría, además, contrastar el desempeño de la solución en un entorno normativo más homogéneo, reduciendo la dependencia de procesos de homologación y facilitando la consolidación institucional de la información bajo un único estándar de reporte.

7.2 Fortalecimiento de la calidad del dato y ampliación de validaciones

Aunque la solución desarrollada ya incorpora controles de estructura, consistencia y trazabilidad, a futuro sería valioso profundizar en reglas de validación más sofisticadas, incluyendo cruces temporales entre periodos, controles de completitud histórica, alertas por cambios atípicos en patrones de reporte y validaciones específicas según sector, rubro o tipo de cuenta.

De igual forma, sería pertinente incorporar mecanismos de retroalimentación sistemática a partir de los errores identificados en los reportes de las entidades financieras. Esto permitiría construir una capa de control de calidad más robusta, orientada no solo a cargar datos correctamente, sino también a mejorar progresivamente la confiabilidad de la información desde su origen.

7.3 Evolución de las herramientas de detección de riesgo

En cuanto a las herramientas analíticas, una línea de trabajo futura consiste en profundizar el desarrollo de los modelos de detección de riesgo. Sin embargo, la mejora efectiva del sistema no dependería únicamente de incorporar más reglas o algoritmos, sino de que se cumplan varias condiciones metodológicas y operativas.

En primer lugar, sería necesario contar con un conjunto de contraste más amplio, consistente y actualizado, construido a partir de revisiones humanas sistemáticas sobre periodos adicionales y distintos sectores del SGP. Esto permitiría refinar umbrales, validar reglas de negocio con mayor robustez y reducir el sesgo derivado de evaluar el sistema sobre una muestra histórica limitada.

En segundo lugar, la herramienta mejoraría si incorporara mayor contexto analítico en cada movimiento evaluado. Para ello, sería conveniente añadir variables relacionadas con comportamiento histórico de la cuenta, frecuencia transaccional, sector, fuente del recurso, tipo de beneficiario, recurrencia territorial y estacionalidad de los pagos. Cuanto más contextualizada esté la observación, mayor sería la capacidad del sistema para distinguir entre comportamientos operativos normales y eventos realmente atípicos.

En tercer lugar, resultaría clave avanzar hacia una lógica de segmentación más especializada. En lugar de aplicar los mismos criterios a universos heterogéneos, el sistema podría calibrarse de manera diferenciada por sector, tipo de cuenta, rubro o territorio, dado que los patrones esperados de operación no son iguales en educación, agua potable, propósito general o asignaciones especiales. Esta segmentación permitiría mejorar la precisión sin sacrificar de manera significativa la capacidad de recuperación de riesgos.

En cuarto lugar, una mejora sustancial requeriría establecer un ciclo de retroalimentación institucional. Si las alertas revisadas por analistas o equipos de control se clasifican posteriormente como confirmadas, descartadas o dudosas, dicha información podría utilizarse para ajustar reglas, recalibrar umbrales y fortalecer progresivamente el modelo híbrido. En ese sentido, el sistema mejoraría en la medida en que dejara de ser una herramienta estática y evolucionara hacia un esquema de aprendizaje continuo, apoyado en evidencia real de uso.

Finalmente, también sería relevante evaluar nuevamente el sistema una vez se encuentre operativa la estructura unificada de reporte prevista en la Resolución 2394 de 2023, ya que un entorno normativo y técnico más homogéneo podría reducir parte del ruido introducido actualmente por la heterogeneidad de las estructuras históricas. En conjunto, estas condiciones permitirían avanzar hacia una herramienta con mejor equilibrio entre recuperación de riesgos, interpretabilidad y reducción de falsas alertas.

7.4 Profundización del análisis de concentración de mercado

Otra línea de continuidad consiste en ampliar la herramienta de concentración de mercado. En su estado actual, esta permite identificar beneficiarios con alta dominancia en la recepción de recursos, alta recurrencia o amplia cobertura territorial. Sin embargo, a futuro sería valioso complementar esta lectura con indicadores adicionales, como medidas de concentración por sector, por fuente de recursos o por periodo, así como incorporar análisis longitudinales que permitan observar la evolución de la concentración en el tiempo.

Del mismo modo, podría explorarse una integración más profunda con enfoques de análisis de redes, representando explícitamente las relaciones entre cuentas, titulares, beneficiarios y territorios. Esto permitiría pasar de una lectura centrada en concentración agregada a una visión

más estructural sobre la circulación de los recursos y la posición relativa de ciertos actores dentro de la red de transferencias.

7.5 Desarrollo de herramientas de visualización y uso operativo

Desde una perspectiva aplicada, una línea de trabajo especialmente valiosa consiste en desarrollar mecanismos de explotación operativa sobre la base construida. En este sentido, sería pertinente implementar tableros de control en herramientas como Power BI, que permitan consultar de manera dinámica la información consolidada, explorar alertas, identificar beneficiarios prioritarios y realizar análisis territoriales o temporales sin depender exclusivamente de consultas técnicas sobre la base de datos.

Este tipo de desarrollo incrementaría de forma importante el valor práctico del proyecto, al acercar los resultados a perfiles de usuario no técnicos y facilitar su incorporación dentro de procesos institucionales de monitoreo, seguimiento y control.

7.6 Desarrollo de herramientas de consulta y exploración a diferentes niveles de granularidad

Una línea de trabajo futura especialmente relevante consiste en desarrollar herramientas orientadas a la consulta, exploración y análisis de la información a diferentes niveles de granularidad, aprovechando la estructura consolidada de la base de datos relacional construida en este proyecto. Una vez organizada la información en dimensiones, hechos y tablas de referencia, se abre la posibilidad de diseñar mecanismos que permitan navegar desde una visión agregada del comportamiento de los recursos hasta el detalle específico de cuentas, movimientos, beneficiarios, periodos y territorios.

Este tipo de desarrollo permitiría que la solución no se limite al almacenamiento y procesamiento técnico de la información, sino que evolucione hacia un entorno más accesible para la consulta analítica y operativa.

Capítulo 8. UTILIZACIÓN DE IA.

El uso de herramientas de inteligencia artificial en este Trabajo Fin de Máster se orientó como apoyo complementario en tareas de planificación, búsqueda bibliográfica, revisión de redacción, procesamiento de información y revisión técnica de código. En todos los casos, la IA se empleó como una herramienta de asistencia y no como un mecanismo de sustitución del análisis, la redacción académica o la toma de decisiones metodológicas del autor. Las salidas generadas fueron revisadas, contrastadas y ajustadas manualmente antes de su incorporación al trabajo.

8.1 Apoyo en la planificación del proyecto

La inteligencia artificial se utilizó para apoyar la construcción inicial del cronograma de trabajo del TFM, a partir del objetivo general del proyecto y de las fases previstas para su desarrollo. En particular, se solicitó apoyo para organizar temporalmente actividades relacionadas con la revisión de literatura, la homogeneización de estructuras, la construcción del diccionario y glosario de datos, el proceso de cargue de la nueva estructura y el desarrollo de herramientas analíticas para concentración de mercado e identificación de operaciones no permitidas.

El prompt utilizado fue el siguiente:

“Ayúdame a crear un cronograma de trabajo para mi TFM en Análisis de Datos Masivos (Big Data), el cual el objetivo principal es Diseñar y gestionar una base de datos relacional unificada para las Cuentas Maestras y Cuentas Maestras Pagadoras del SGP, orientada a mejorar la trazabilidad y el seguimiento del uso de recursos públicos en Colombia. Para este cronograma necesito incluir las fases de revisión de literatura, desarrollo del modelo que consiste en las siguientes fases:

Homogeneizar información de distintas estructuras

1.1. Construir un diccionario de datos con la nueva estructura

1.2. Construir un glosario de datos con la nueva estructura

1.3. Proceso de llenado de la nueva estructura

Sacarle más provecho esta información

2.1. Identificar Clústeres de concentración de mercado

2.2. Identificación de operaciones no permitidas

Adicionalmente, que este cronograma tiempo en semanas para la ejecución de las tareas, la idea es que la duración sea de aproximadamente 12-14 semanas.”

A partir de la propuesta inicial generada por la IA, se realizaron ajustes manuales sobre la estructura del cronograma, los entregables definidos y la asignación de horas por actividad. Esto permitió adaptar la planificación a las necesidades reales del proyecto y alinearla con la estructura final de la memoria.

8.2 Apoyo en la búsqueda y orientación bibliográfica

La IA también se utilizó como apoyo inicial para identificar líneas bibliográficas pertinentes para la construcción del estado del arte. En este caso, su función consistió en sugerir referencias y temas de búsqueda asociados al diseño y gestión de bases de datos relacionales, prácticas de GovTech y modelos para la identificación de concentración de mercado y movimientos anómalos.

El prompt utilizado fue el siguiente:

“Podrías ayudarme con la consulta de fuentes científicas o artículos que puedan ser consultadas en internet para completar la sección del estado del arte de mi TFM. Al respecto, lo importante es que la bibliografía se enfoque en creación y gestión de bases de datos relacionales, si se puede enfocada a casos de recursos públicos o prácticas correctas de GovTech y la otra bibliografía que este enfocada en modelos para la identificación de concentración de mercado y para identificación de movimientos o transferencias anómalos”.

Como resultado, la IA propuso varios bloques de referencias potenciales. No obstante, la selección final de la bibliografía utilizada en el trabajo no fue automática. Las referencias sugeridas fueron revisadas de manera crítica y contrastadas mediante búsquedas adicionales en Google Scholar, con el fin de seleccionar únicamente aquellas que resultaban pertinentes para el problema, el enfoque metodológico y el nivel académico del TFM.

8.3 Apoyo en la revisión de redacción académica

Otra aplicación de la IA consistió en apoyar la revisión de párrafos y secciones de la memoria desde una perspectiva de redacción académica. En este uso, la herramienta se empleó para identificar oportunidades de mejora en claridad, cohesión, precisión y estilo formal, manteniendo como criterio que la redacción final debía ser revisada y ajustada manualmente.

El prompt base utilizado fue el siguiente:

*“Eres experto en redacción académica y en metodología de la investigación.
Ayúdame a revisar la redacción del siguiente párrafo de mi Trabajo de Fin de Máster.
El nivel académico corresponde a un Máster.
Debes seguir el formato de normas APA.
Revisa que la redacción cuente con un estilo académico, fluido, neutro y preciso.
El tono debe mantenerse académico, coherente y sin redundancias.
Mantén precisión, claridad y rigor académico en todo el texto.
X Negative prompt:
X No uses o permitas tono personal ("creo", "en mi opinión").
X No permitas o repitas ideas.*

X No permitas que sea vago, comercial ni narrativo.

X No generes contenido fuera de tema.

El párrafo es el siguiente: (Párrafo)”.

Las salidas generadas por la IA fueron tratadas como propuestas de mejora y no como versiones definitivas. En cada caso se realizó una revisión manual posterior para ajustar la puntuación, la consistencia terminológica y la fidelidad de las ideas frente al contenido técnico y metodológico del trabajo.

8.4 Apoyo en el procesamiento de la información de contraste

La IA también se utilizó como apoyo técnico para preparar la base de contraste empleada en la evaluación de los modelos de riesgo. En este caso, se partió de un conjunto de archivos de Excel organizados por departamento y municipio, en los cuales ciertos movimientos estaban resaltados como posibles riesgos. La tarea consistió en consolidar dicha información en un único archivo utilizable como insumo para el contraste de resultados.

El prompt utilizado fue el siguiente:

“Eres experto en procesamiento de datos.

Para mi herramienta de alerta de riesgos, que estoy elaborando para mi TFM, en la que busco identificar movimientos anómalos o no permitidos que está en los dos archivos de Python que te comparto, quiero hacer una prueba de contraste frente a unos movimientos anómalos que están totalmente identificados. Los tengo en una carpeta zip que contiene diferentes carpetas por departamento y por municipio en este departamento, cada carpeta de municipio contiene un archivo de Excel con los movimientos identificados como posibles riesgos con relleno de color, podrías ayudarme a consolidar todos estos archivos en un solo txt y que estos que tienen este relleno de color le coloquemos un nuevo campo que se llame posible riesgo. te adjunto el zip soportes suspensión masiva que es el que contiene la información que te indico”.

El archivo resultante fue utilizado como insumo para el contraste frente a los resultados obtenidos por los modelos desarrollados, previa verificación manual de su estructura y utilidad para el ejercicio comparativo.

8.5 Apoyo en la revisión técnica de scripts en Python

Finalmente, la IA se empleó como apoyo en la revisión técnica de los scripts de Python desarrollados en el proyecto, particularmente en aspectos relacionados con coherencia del código, eficiencia, modularidad y prevención de errores de ejecución.

El prompt base utilizado fue el siguiente:

“Eres experto en programación en Python con más de 10 años de experiencia.

Ayúdame a revisar los scripts de Python elaborados para mi Trabajo de Fin de Máster.

Revisa la coherencia del código y su eficiencia.

Que no existan loops que dejen el proceso congelado o errores de ejecución.

Mantén precisión, claridad y rigor en todo el código.

X Negative prompt:

X No cambies el código sin sustento técnico.

X No hagas cambios sin previa autorización y explicación de la razón.”

En este caso, la revisión se realizó mediante un agente de IA integrado en Visual Studio Code, lo que permitió mantener control directo sobre las sugerencias propuestas antes de su aceptación. De esta forma, cualquier cambio aplicado al código fue previamente revisado y validado por el autor, garantizando supervisión humana sobre la evolución de los scripts y evitando modificaciones automáticas sin justificación técnica.

8.6 Consideraciones finales sobre el uso de IA

En conjunto, la IA se utilizó en este trabajo como una herramienta de apoyo para tareas auxiliares de organización, revisión, orientación bibliográfica y asistencia técnica. Su uso no reemplazó la toma de decisiones metodológicas, la selección final de bibliografía, la interpretación de resultados ni la redacción definitiva del documento. En todos los casos, el criterio final, la validación de contenidos y la responsabilidad sobre el resultado presentado recayeron en el autor.

Capítulo 9. OBJETIVOS DE DESARROLLO SOSTENIBLE DE LAS NACIONES UNIDAS.

Con los resultados obtenidos en este trabajo se identifican situaciones de riesgo en la ejecución de los recursos del Sistema General de Participaciones (SGP), tales como posibles concentraciones de mercado en el uso de las Cuentas Maestras y eventuales transferencias no permitidas. Dado que estos recursos están destinados a la inversión pública en Colombia, el fortalecimiento de su trazabilidad, control y seguimiento se articula de forma directa con varios Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, entre ellos los ODS 2, 9, 10 y 11, como se expone a continuación.

9.1 Objetivo 2. Hambre Cero

La bolsa de la Asignación Especial de Alimentación Escolar del SGP tiene como principal objetivo financiar el Programa de Alimentación Escolar (PAE), política pública de permanencia escolar que busca que niños, niñas y adolescentes reciban alimentos durante su jornada académica, reduciendo así la deserción escolar en hogares de menores ingresos. En la medida en que la adecuada ejecución de estos recursos contribuye tanto a la seguridad alimentaria como a la continuidad educativa, su gestión se relaciona directamente con el ODS 2, “Hambre cero”.

En este contexto, los mecanismos desarrollados en el presente trabajo (orientados a detectar riesgos en la utilización de las Cuentas Maestras, tales como transferencias no permitidas o patrones atípicos en la ejecución de los recursos) aportan a disminuir la probabilidad de malversación o uso ineficiente de los recursos destinados al PAE, especialmente en las zonas más pobres del país. Adicionalmente, la capacidad de identificar operaciones inusuales puede contribuir a la identificación temprana de patrones asociados a usos ineficientes o potencialmente irregulares de los recursos. De esta forma, una mejor calidad de la información y mejores herramientas en el control sobre las operaciones realizadas a través de las Cuentas Maestras fortalece la garantía efectiva del derecho a la alimentación escolar y, por extensión, la contribución del SGP al ODS 2.

9.2 Objetivo 9. Industria, innovación e infraestructura

Dentro de los sectores financiados por la Participación de Propósito General del SGP se encuentran la promoción del desarrollo, el sector vivienda y el desarrollo comunitario (Congreso de la República de Colombia, 2001), los cuales están estrechamente vinculados con la infraestructura social y productiva del país. La correcta focalización y ejecución de estos recursos permite financiar proyectos que mejoran las condiciones de vida, dinamizan las economías locales y reducen brechas en infraestructura entre territorios.

En este sentido, el diseño de una base de datos relacional que consolida la información de las Cuentas Maestras y Cuentas Maestras Pagadoras, junto con los mecanismos de detección de riesgos propuestos, aporta a la modernización de la gestión financiera pública y al uso de la analítica de datos como herramienta de innovación en el seguimiento de la inversión. La

posibilidad de identificar patrones de concentración de mercado, flujos de recursos atípicos y posibles desvíos incrementa la transparencia y la eficiencia en el uso de los recursos del SGP, creando condiciones más favorables para que la inversión pública se traduzca efectivamente en infraestructura y servicios, en línea con el ODS 9, “Industria, innovación e infraestructura”.

9.3 Objetivo 10. Reducción de las desigualdades

El SGP es, por diseño, una bolsa de recursos de destinación específica para inversión pública, cuyo objetivo central es reducir las desigualdades territoriales en el acceso a servicios y oportunidades. Los criterios de distribución de estos recursos consideran variables como población, equidad, eficiencia, ruralidad y densidad poblacional de cada entidad territorial (Departamento Nacional de Planeación, 2023), lo que permite orientar más recursos hacia los territorios con mayores vulnerabilidades.

Bajo este enfoque, cualquier herramienta que contribuya a la ejecución eficiente, oportuna y transparente de los recursos del SGP —como los mecanismos de consolidación de información y detección de riesgos desarrollados en este trabajo— se alinea con el ODS 10, “Reducción de las desigualdades”. En particular, la identificación de transferencias no permitidas y de posibles desviaciones contribuye a que los recursos lleguen a los territorios y sectores para los cuales fueron concebidos, evitando pérdidas o reasignaciones injustificadas que puedan profundizar brechas. Asimismo, el análisis de concentración de mercado en el uso de las Cuentas Maestras ofrece insumos para reflexionar sobre las condiciones de acceso a servicios financieros de las entidades territoriales y, potencialmente, ofrecer insumos para revisar posibles patrones de concentración en el uso de servicios financieros por parte de las entidades territoriales.

9.4 Objetivo 11. Ciudades y comunidades sostenibles

El propósito principal del SGP es garantizar la prestación de servicios básicos y financiar inversiones que contribuyan a reducir las brechas sociales y de desarrollo entre regiones, lo que incluye componentes clave para el ODS 11, “Ciudades y comunidades sostenibles”, tales como agua potable y saneamiento básico, educación, infraestructura urbana y servicios comunitarios. El uso adecuado de estos recursos es determinante para que ciudades y municipios avancen hacia entornos más inclusivos, seguros, resilientes y sostenibles.

Los mecanismos propuestos en este trabajo para identificar transferencias no permitidas y flujos de recursos inconsistentes entre Cuentas Maestras fortalecen el control sobre el destino de la inversión pública a nivel territorial. Al mejorar la trazabilidad de los movimientos financieros y facilitar la detección temprana de posibles irregularidades, se contribuye a que los recursos del SGP se orienten efectivamente a proyectos que impactan de forma positiva la calidad de vida en las ciudades y comunidades. En consecuencia, el desarrollo de esta solución de integración, seguimiento y gestión se convierte en un soporte técnico para la planificación territorial y la sostenibilidad urbana y rural, en consonancia con las metas del ODS 11.

Capítulo 10. REFERENCIAS

- Al-Hashedi, K., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40. doi:10.1016/j.cosrev.2021.100402
- Asamblea Nacional Constituyente. (1991). Constitución Política de Colombia. Bogotá D. C., Colombia. Obtenido de http://www.secretariassenado.gov.co/senado/basedoc/constitucion_politica_1991.html
- Bansal, A. (Agosto de 2012). E-Government and Social Media as Openness and Anti-Corruption Strategy. *Research Journal of Management Sciences*, 1(1), 48–52.
- Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 27, 264–271.
- Bhowmik, R. (2008). Data Mining Techniques in Fraud Detection. *Journal of Digital Forensics, Security and Law*, 3(2). doi:10.15394/jdfsl.2008.1040
- Bonet, J., Pérez, G. J., & Ayala, J. (julio de 2014). Contexto histórico y evolución del SGP en Colombia. Bogotá D.C., Colombia. Obtenido de https://www.banrep.gov.co/sites/default/files/publicaciones/archivos/dtser_205.pdf
- Brezina, I., Pekár, J., Čičková, Z., & Reiff, M. (2016). Herfindahl–Hirschman index level of concentration values modification and analysis of their change. *Central European Journal of Operations Research*(24), 49–72.
- Congreso de la República de Colombia. (21 de diciembre de 2001). Ley 715 de 2001. Bogotá D.C., Colombia. Obtenido de http://www.secretariassenado.gov.co/senado/basedoc/ley_0715_2001.html
- DAMA International. (2017). *The DAMA Guide to the Data Management Body of Knowledge* (Second ed.). Basking Ridge, NJ: Technics Publications.
- Date, C. J. (2001). *Introducción a los sistemas de bases de datos*. México: Pearson Educación.
- Departamento Nacional de Planeación. (2023). Manual para la distribución del Sistema General de Participaciones (SGP). Bogotá D.C., Colombia. Obtenido de <https://colaboracion.dnp.gov.co/CDT/DNP/SIG/M-CA-02%20Manual%20para%20la%20distribuci%C3%B3n%20del%20SGP.Pu.pdf>
- Elmasri, R., & Navathe, S. B. (2007). *Fundamentos de Sistemas de Bases de Datos*. Madrid.
- Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, 11–16.

- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4), 258–268.
- Lenka, S., & Ratha, B. (2024). Financial Fraud Detection using Data Mining: A Survey. *IJCSNS International Journal of Computer Science and Network Security*, 24(9), 169–185. doi:10.22937/IJCSNS.2024.24.9.20
- Liu, F., Ting, K., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, (págs. 413–422). Pisa. doi:10.1109/ICDM.2008.17
- McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter* (Third edition ed.). Sebastopol, CA: O'Reilly Media, Inc.
- Microsoft. (s.f.). *sqlcmd utility*. Obtenido de Microsoft Learn: <https://learn.microsoft.com/en-us/sql/tools/sqlcmd/sqlcmd-utility?view=sql-server-ver17&tabs=go%2Cwindows-support&pivots=cs1-bash>
- Ministerio de Hacienda y Crédito Público. (2015). *Responsabilidad Fiscal Subnacional y Descentralización en Colombia: Quince años de Consolidación*. Bogotá : Ministerio de Hacienda y Crédito Público.
- Owens, M. (2006). *The Definitive Guide to SQLite*. Berkeley: Apress.
- Presidencia de la República de Colombia. (10 de enero de 2008). Decreto 028 de 2008. Bogotá D. C., Colombia. Recuperado el 12 de 2025, de <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=28232>

Capítulo 11. ANEXOS

En este apartado se presentan los materiales complementarios que respaldan técnica y metodológicamente el desarrollo del proyecto. Su finalidad es ampliar la trazabilidad del trabajo realizado y ofrecer evidencia adicional sobre la implementación de la solución propuesta, sin sobrecargar el cuerpo principal de la memoria.

Los anexos incluyen el repositorio técnico del proyecto, el diccionario y catálogo de datos.

11.1 Anexo A. Repositorio técnico del proyecto

El desarrollo técnico del presente Trabajo Fin de Máster se encuentra documentado y respaldado en un repositorio de GitHub que reúne los principales artefactos construidos durante el proyecto. En este repositorio se incluyen los scripts SQL asociados a la creación del modelo relacional, los procesos ETL desarrollados en Python, las herramientas analíticas para identificación de riesgos y concentración de mercado, así como documentación técnica complementaria.

Repositorio del proyecto:

<https://github.com/ChapaCamilo-collab/CuentasMaestrasSGP>

De forma general, el repositorio se organiza en componentes orientados a documentación de contexto, scripts SQL, procesos ETL en Python, herramientas analíticas y artefactos de soporte técnico, lo que permite dar trazabilidad a la implementación integral de la solución.

11.2 Anexo B. Diccionario de datos

Se anexa el diccionario de datos del modelo relacional propuesto, en el cual se documentan las tablas, campos, tipos de dato, longitudes, nulabilidad, claves primarias y foráneas, así como observaciones relevantes para la interpretación e implementación de la base de datos.



Diccionario de Datos
CM_CMP.xlsx

11.3 Anexo C. Catálogo y glosario de datos

Se anexa el catálogo de datos y el glosario de términos del proyecto, orientados a documentar el propósito de cada tabla, su granularidad, periodicidad, dependencias, definiciones de negocio y términos clave asociados al contexto de Cuentas Maestras y Cuentas Maestras Pagadoras.



Catalogo_de_Datos_C
M_CMP.xlsx

[PÁGINA INTENCIONADAMENTE EN BLANCO]