MÁSTER EN BIG DATA

TRABAJO FIN DE MÁSTER

ANÁLISIS DE RIESGO EN LA CONTRATACIÓN PÚBLICA DE LA ADMINISTRACIÓN GENERAL DEL ESTADO MEDIANTE UN ÍNDICE COMPUESTO Y TECNOLOGÍAS BIG DATA

Autor: Carlos García Revillas

Director: Carlos Miguel Vallez Fernández

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

ANÁLISIS DE RIESGO EN LA CONTRATACIÓN PÚBLICA DE LA ADMINISTRACIÓN GENERAL DEL ESTADO MEDIANTE UN ÍNDICE COMPUESTO Y TECNOLOGÍAS BIG DATA

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2025/26 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.
El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Carlos García Revillas Fecha: 22 / 06 / 2026

Autorizada la entrega del proyecto
EL DIRECTOR DEL PROYECTO

Carlos Miguel Vallez Fernández

Firmado digitalmente por Carlos Miguel Vallez Fernández
Fecha: 2026.06.23 11:50:06 +02'00'

Fdo.: Carlos Miguel Vallez Fernández Fecha: ……/ ……/ ……

Vº Bº del Coordinador de Proyectos

Fdo.: Carlos Morrás Ruiz-Falcó Fecha: ……/ ……/ ……

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. Carlos García Revillas_____

DECLARA ser el titular de los derechos de propiedad intelectual de la obra: análisis de riesgo en la contratación pública de la administración general del estado mediante un índice compuesto y tecnologías big data, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.
b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
e) Asignar por defecto a estos trabajos una licencia Creative Commons.
f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente)*.

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

a) Que la Universidad identifique claramente su nombre como autor de la misma
b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
c) Solicitar la retirada de la obra del repositorio por causa justificada.
d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

- El autor se compromete a:

a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que

pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a ……22…….. de ……junio……... de ….2026

ACEPTA

Fdo……

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

MÁSTER EN BIG DATA

TRABAJO FIN DE MÁSTER

ANÁLISIS DE RIESGO EN LA CONTRATACIÓN PÚBLICA DE LA ADMINISTRACIÓN GENERAL DEL ESTADO MEDIANTE UN ÍNDICE COMPUESTO Y TECNOLOGÍAS BIG DATA

Autor: Carlos García Revillas

Director: Carlos Miguel Vallez Fernández

Madrid

ANÁLISIS DE RIESGO EN LA CONTRATACIÓN PÚBLICA DE LA ADMINISTRACIÓN GENERAL DEL ESTADO MEDIANTE UN ÍNDICE COMPUESTO Y TECNOLOGÍAS BIG DATA

Autor: García Revillas, Carlos.
Director: Miguel Vallez Fernández, Carlos.
Entidad Colaboradora: ICAI - Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Se ha diseñado e implementado un sistema Big Data que, a partir de los datos abiertos de la Plataforma de Contratación del Sector Público (PLACSP), calcula cinco indicadores estructurales de riesgo y los combina en un índice compuesto para ordenar los órganos de la Administración General del Estado (AGE) según su nivel de riesgo relativo, como herramienta de apoyo a la priorización de auditorías. El ranking resultante muestra una elevada robustez frente a la elección de pesos, con correlaciones de Spearman entre esquemas de ponderación comprendidas entre 0,93 y 0,99, y presenta una correspondencia cualitativa razonable con los casos señalados por fuentes públicas de control.

Palabras clave: contratación pública, red flags, índice compuesto de riesgo, Big Data, PLACSP, Apache Spark

1. Introducción

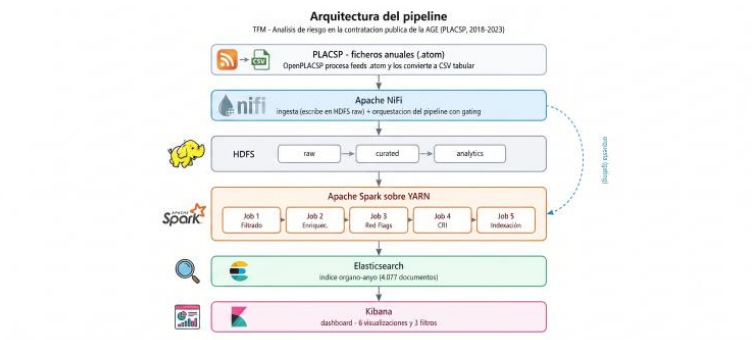
La contratación pública supone alrededor del 11 % del PIB en España y es un ámbito clave por su impacto económico y su relación con la integridad institucional. La Ley 9/2017 reforzó la publicación de las licitaciones en PLACSP, pero disponer de datos abiertos no equivale a poder explotarlos: la literatura muestra que patrones como la licitación única, los procedimientos poco competitivos o la ausencia de descuento se asocian a un mayor riesgo, y en España no existía un sistema reproducible que los calculase sobre PLACSP y los integrase en un índice por órgano. Este trabajo aborda ese vacío.

2. Definición del proyecto

El objetivo es un sistema de análisis de riesgo en la contratación de la AGE mediante Big Data, capaz de generar rankings de órganos según su riesgo relativo. Se concreta en cinco objetivos: un pipeline distribuido y reproducible (OE1), el cálculo de las red flags por licitación (OE2), su agregación por órgano-año y la construcción de un Composite Risk Index con tres esquemas de ponderación (OE3), la evaluación de la estabilidad de los rankings (OE4) y su visualización en un dashboard (OE5). El alcance son las licitaciones de la AGE con resultado conocido para Obras, Servicios y Suministros en 2018-2023. El sistema no detecta corrupción: identifica patrones asociados a un mayor riesgo y produce un ranking relativo para priorizar revisiones.

3. Descripción del modelo/sistema/herramienta

El sistema es un pipeline por capas orquestado de extremo a extremo (Ilustración 1): ingesta y orquestación con Apache NiFi, almacenamiento en HDFS, procesamiento en Apache Spark sobre YARN, e indexación y visualización en Elasticsearch y Kibana. NiFi deposita los ficheros en la capa raw con enrutado dinámico y encadena el procesamiento con un gating que solo avanza si cada etapa termina correctamente. Se reparte en cinco jobs (filtrado, enriquecimiento, red flags y agregación, índice e indexación), y los cinco indicadores son la licitación única, los procedimientos no competitivos, el precio relativo, la concentración de proveedores y el plazo de presentación.

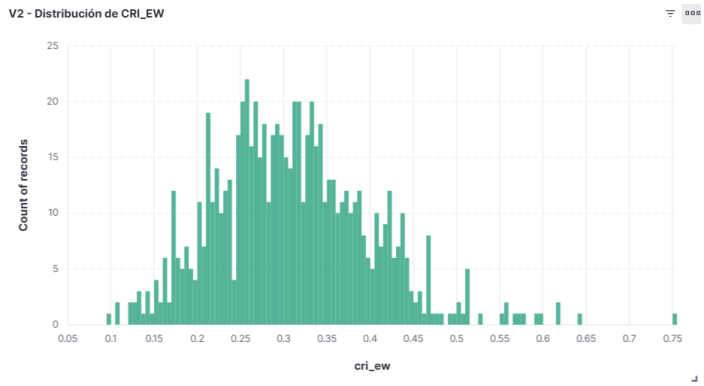


Arquitectura del sistema: pipeline por capas con orquestación NiFi.

4. Resultados

El pipeline procesa 124.761 registros de entrada que, tras la limpieza y la deduplicación de los snapshots acumulativos de PLACSP, quedan en 104.909 contratos únicos. La agregación produce 4.077 grupos órgano-año, de los cuales 743 superan el umbral de calidad y concentran cerca del ochenta por ciento del volumen contratado. La distribución del índice (Ilustración 2) muestra que el riesgo no se reparte de forma difusa, sino que se concentra en una cola estrecha de órganos, lo que hace el resultado directamente accionable.

El resultado más relevante es la robustez del índice. Las correlaciones de Spearman entre los tres esquemas de ponderación oscilan entre 0,93 y 0,99, y la coincidencia de órganos en las primeras posiciones del ranking alcanza el 68% en los diez primeros y el 84% en los cincuenta primeros, lo que indica que el orden apenas depende de la elección de pesos. El análisis de casos confirma además la utilidad interpretativa del sistema: órganos como la Subdirección General de Adquisiciones de Armamento y Material, INGESA o la AECID aparecen de forma recurrente, con perfiles de señal que admiten tanto lecturas de riesgo como explicaciones legítimas asociadas a su contexto. Por último, el contraste cualitativo de estos casos con informes del Tribunal de Cuentas y resoluciones de la CNMC aporta un indicio razonable de validez externa: el sistema, construido sin conocimiento previo de esos informes, vuelve a señalar órganos ya sometidos al escrutinio del control institucional.



Distribución del índice de riesgo sobre la población de calidad alta

5. Conclusiones

El proyecto cumple los cinco objetivos planteados y demuestra la viabilidad de construir un sistema de análisis de riesgo reproducible sobre datos abiertos de PLACSP. Sus principales aportaciones son tres: empírica, al constituir el primer análisis sistemático centrado en la AGE a partir de datos primarios de PLACSP; técnica, por el diseño de un pipeline auditable con tecnologías Big Data de código abierto; y metodológica, al medir explícitamente la robustez del índice mediante tres esquemas de ponderación en lugar de darla por supuesta.

Como líneas de trabajo futuro destacan la incorporación de un análisis de tipologías mediante clustering, la red flag de fragmentación de contratos basada en similitud textual y la extensión del sistema a comunidades autónomas y entidades locales.

6. Referencias

[5] M. Fazekas y G. Kocsis, «Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data,» *British Journal of Political Science*, vol. 50, nº 1, pp. 155-164, 2020.

[6] Open Contracting Partnership, «Red Flags in Public Procurement: A Guide to Using Data to Detect and Mitigate Risks,» 2024. [En línea]. Available: https://www.open-contracting.org/resources/red-flags-in-public-procurement-a-guide-to-using-data-to-detect-and-mitigate-risks/.

[10] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker y I. Stoica, «Apache Spark: A Unified Engine for Big Data Processing,» *Communications of the ACM*, vol. 59, nº 11, pp. 56-65, 2016.

[15] M. Fazekas, I. J. Tóth y L. P. King, «An Objective Corruption Risk Index Using Public Procurement Data,» *European Journal on Criminal Policy and Research*, vol. 22, nº 3, pp. 369-397, 2016.

RISK ANALYSIS IN PUBLIC PROCUREMENT OF THE SPANISH GENERAL STATE ADMINISTRATION (AGE) THROUGH A COMPOSITE INDEX AND BIG DATA TECHNOLOGIES

Author: García Revillas, Carlos.
Supervisor: Miguel Vallez Fernández, Carlos.
Collaborating Entity: ICAI - Universidad Pontificia Comillas

ABSTRACT

A Big Data system has been designed and implemented that, using the open data of the Public Sector Procurement Platform (PLACSP), computes five structural risk indicators and combines them into a composite index to rank the bodies of the General State Administration (AGE) according to their relative risk level, as a tool to support the prioritization of audits. The resulting ranking shows high robustness against the choice of weights, with Spearman correlations between weighting schemes ranging from 0.93 to 0.99, and a reasonable qualitative correspondence with the cases flagged by public oversight bodies.

Keywords: public procurement, red flags, composite risk index, Big Data, PLACSP, Apache Spark.

1. Introduction

Public procurement accounts for around 11% of GDP in Spain and is a key area owing to its economic impact and its relationship with institutional integrity. Law 9/2017 strengthened the publication of tenders in PLACSP, but having open data does not equate to being able to exploit it: the literature shows that patterns such as single bidding, non-competitive procedures or the absence of a discount are associated with higher risk, and in Spain there was no reproducible system that computed them from PLACSP and integrated them into an index by contracting body. This work addresses that gap.

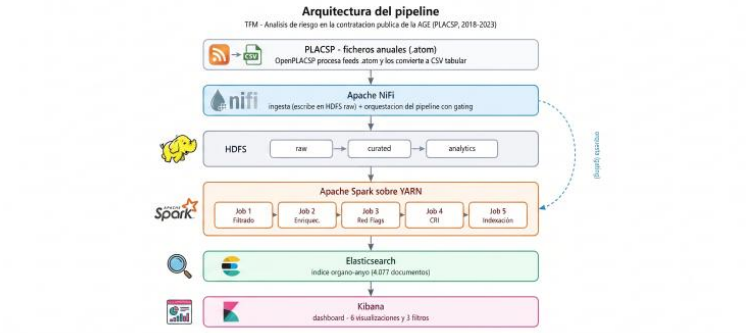
2. Project definition

The aim is a risk-analysis system for AGE procurement using Big Data, capable of generating rankings of bodies according to their relative risk. It is broken down into five specific objectives: a distributed and reproducible pipeline (SO1), the computation of red flags at tender level (SO2), their aggregation by body-year and the construction of a Composite Risk Index under three weighting schemes (SO3), the evaluation of ranking stability (SO4) and their visualization in an interactive dashboard (SO5). The scope is limited to AGE tenders with a known outcome for Works, Services and Supplies over the period 2018-2023. The system does not detect corruption: it identifies patterns associated with higher risk and produces a relative ranking to prioritize reviews.

3. System description

The system is a layered pipeline orchestrated end to end (Figure 1): ingestion and orchestration with Apache NiFi, storage in HDFS, processing in Apache Spark on YARN, and indexing and visualization in Elasticsearch and Kibana. NiFi places the files in the raw layer with dynamic routing and chains the processing through a gating mechanism that only advances if each stage finishes correctly. It is split into five Spark

jobs (filtering, enrichment, red flags and aggregation, index construction, and indexing), and the five indicators are single bidding, non-competitive procedures, relative award price, supplier concentration and submission period.

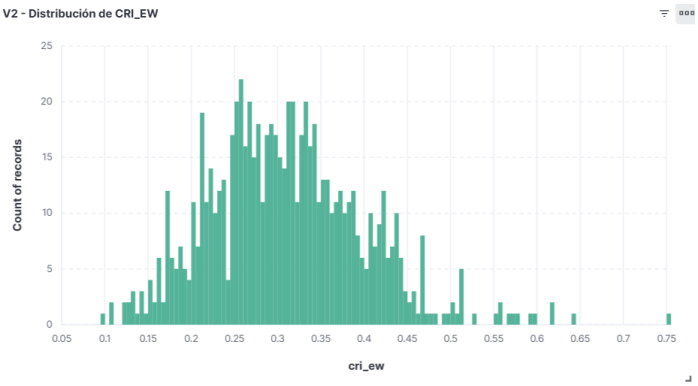


System architecture: layered pipeline with NiFi orchestration.

4. Results

The pipeline processes 124,761 input records which, after cleaning and deduplication of PLACSP's cumulative snapshots, are reduced to 104,909 unique contracts. Aggregation produces 4,077 body-year groups, of which 743 exceed the quality threshold and account for nearly eighty per cent of the contracted volume. The distribution of the index (Figure 2) shows that risk is not spread diffusely but concentrated in a narrow tail of bodies, which makes the result directly actionable.

The most relevant result is the robustness of the index. The Spearman correlations between the three weighting schemes range from 0.93 to 0.99, and the overlap of bodies in the top positions of the ranking reaches 68% in the top ten and 84% in the top fifty, indicating that the ordering barely depends on the choice of weights. The case analysis further confirms the interpretive usefulness of the system: bodies such as the Subdirectorate-General for Armament and Materiel Procurement, INGESA or AECID appear recurrently, with signal profiles that admit both risk readings and legitimate explanations associated with their context. Finally, the qualitative contrast of these cases with reports from the Court of Audit (Tribunal de Cuentas) and resolutions from the CNMC provides reasonable evidence of external validity: the system, built without prior knowledge of those reports, again flags bodies already subject to institutional oversight.



Distribution of the risk index over the high-quality population.

5. Conclusions

The project meets the five objectives set out and demonstrates the feasibility of building a reproducible risk-analysis system on PLACSP open data. Its main contributions are threefold: empirical, as the first systematic analysis focused on the AGE built from PLACSP primary data; technical, through the design of an auditable pipeline with open-source Big Data technologies; and methodological, by explicitly measuring the robustness of the index through three weighting schemes rather than assuming it. Future lines of work include a typology analysis through clustering, a contract-fragmentation red flag based on textual similarity, and the extension of the system to regional and local administrations.

6. References

[5] M. Fazekas y G. Kocsis, «Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data,» *British Journal of Political Science*, vol. 50, nº 1, pp. 155-164, 2020.

[6] Open Contracting Partnership, «Red Flags in Public Procurement: A Guide to Using Data to Detect and Mitigate Risks,» 2024. [En línea]. Available: https://www.open-contracting.org/resources/red-flags-in-public-procurement-a-guide-to-using-data-to-detect-and-mitigate-risks/.

[10] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker y I. Stoica, «Apache Spark: A Unified Engine for Big Data Processing,» *Communications of the ACM*, vol. 59, nº 11, pp. 56-65, 2016.

[15] M. Fazekas, I. J. Tóth y L. P. King, «An Objective Corruption Risk Index Using Public Procurement Data,» *European Journal on Criminal Policy and Research*, vol. 22, nº 3, pp. 369-397, 2016.

Índice de la Memoria

Índice de Figuras

Índice de Tablas

Capítulo 1. INTRODUCCIÓN

1.1MOTIVACIÓN DEL PROYECTO

La contratación pública representa una parte importante del gasto en cualquier economía desarrollada. En España, los contratos públicos adjudicados suponen habitualmente alrededor del once por ciento del Producto Interior Bruto, según los datos de contabilidad nacional publicados por Eurostat [1]. Esta dimensión hace que la contratación pública sea un ámbito especialmente relevante, tanto por su impacto en el uso de los recursos públicos como por su relación con la transparencia y el buen funcionamiento de las instituciones.

En este contexto, la Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público (LCSP), incorporó al marco jurídico español las Directivas europeas 2014/23/UE y 2014/24/UE. Su objetivo fue avanzar hacia un sistema de contratación más eficiente, transparente y con mayores garantías de integridad [2]. Entre los principales cambios introducidos por esta norma destacan la ampliación del procedimiento abierto simplificado, una regulación más estricta del contrato menor, la obligación de publicar los contratos en la Plataforma de Contratación del Sector Público (PLACSP) y el refuerzo de principios como la igualdad de trato, la no discriminación, la transparencia y la libre competencia [3].

Además, la entrada en vigor efectiva de la LCSP en marzo de 2018 marca el inicio de un periodo de mayor estabilidad regulatoria en el núcleo procedimental de la contratación pública, sin perjuicio de modificaciones puntuales introducidas durante la pandemia de COVID-19. Por ello, este intervalo se toma como marco temporal de referencia para el desarrollo del trabajo.

De forma paralela a los cambios normativos, la Dirección General del Patrimonio del Estado ha puesto a disposición pública, a través del portal del Ministerio de Hacienda, los datos estructurados de las licitaciones registradas en PLACSP desde 2012. Estos datos incluyen información como el órgano de contratación, el tipo y el objeto del contrato, el presupuesto

base, el procedimiento utilizado, el número de ofertas recibidas y el resultado de la adjudicación [4]. Esta apertura se apoya en la herramienta oficial OpenPLACSP, que permite transformar los ficheros ATOM publicados cada año en conjuntos de datos estructurados. Gracias a ello, tanto investigadores como instituciones pueden acceder a una fuente de información especialmente útil para analizar de forma sistemática el comportamiento de los órganos de contratación.

Sin embargo, disponer de datos abiertos no significa, por sí solo, disponer de herramientas que permitan analizarlos de forma útil. La literatura internacional ha mostrado que ciertos patrones en los procesos de contratación pública, como una proporción elevada de licitaciones con un único licitador, el uso frecuente de procedimientos poco competitivos o reducciones muy bajas del precio final respecto al presupuesto base, suelen estar asociados a un mayor riesgo de irregularidad [5] [6]. En el caso español, y hasta donde alcanza el conocimiento del autor, no existe un sistema de análisis basado en datos de PLACSP que permita calcular estos indicadores de forma reproducible, combinarlos en un índice compuesto y generar rankings de riesgo relativo por órgano de contratación.

Esta diferencia entre la disponibilidad de los datos y la falta de herramientas para explotarlos de forma sistemática es la principal motivación de este trabajo. Publicar la información es un paso necesario, pero no suficiente. Para que la transparencia tenga un efecto real sobre la integridad de los procesos, hace falta no solo que la información sea pública, sino que sea accesible de forma temprana y procesable. La literatura empírica demuestra que es precisamente la transparencia ex ante, la que ofrece información útil antes del cierre del plazo, la que reduce los riesgos de restricción de la competencia [7]. Trasladar ese principio al análisis sistemático del comportamiento de los órganos requiere herramientas técnicas que permitan estructurar y explotar los datos públicos disponibles.

Desde el inicio, conviene dejar claro que el sistema no detecta corrupción ni clasifica contratos como irregulares. Su función es identificar patrones estadísticos asociados a un mayor riesgo de irregularidad o a una menor competencia efectiva, y generar un ranking de

riesgo relativo que sirva para orientar la priorización de revisiones, no para extraer conclusiones sobre casos concretos.

1.2 OBJETIVOS

La pregunta de investigación que guía este Trabajo Fin de Máster es la siguiente: ¿es posible construir, a partir de los datos abiertos de la Plataforma de Contratación del Sector Público, un índice compuesto de riesgo en contratación pública para los órganos de la Administración General del Estado, basado en indicadores estructurales objetivos? Y, además, ¿son estables los rankings obtenidos cuando se aplican distintos esquemas de agregación?

Esta pregunta puede abordarse desde tres dimensiones complementarias. La primera es técnica: si es posible desarrollar un pipeline de datos reproducible que procese los ficheros anuales de PLACSP, calcule los indicadores de riesgo y genere resultados auditables. La segunda es metodológica: si los indicadores seleccionados representan dimensiones del riesgo lo bastante diferenciadas como para justificar la construcción de un índice compuesto. La tercera se centra en la validación: dado que no existen etiquetas directas de corrupción confirmada, cómo puede evaluarse la robustez y la utilidad del índice obtenido.

A partir de esta cuestión, el objetivo general del proyecto es diseñar e implementar un sistema de análisis de riesgo en la contratación pública española utilizando tecnologías Big Data, capaz de generar rankings de órganos de contratación de la Administración General del Estado según su nivel de riesgo relativo. El sistema se plantea como una herramienta de apoyo para priorizar auditorías y tareas de supervisión.

Para alcanzar este objetivo general, se establecen los siguientes objetivos específicos:

- OE1. Diseñar e implementar un pipeline de datos distribuido, reproducible y escalable, basado en Apache NiFi, HDFS y Apache Spark, que procese los datos de

PLACSP para el periodo 2018-2023 y los almacene en un formato estructurado optimizado para su análisis.

- OE2. Definir y calcular un conjunto de indicadores estructurales de riesgo, o red flags, a nivel de licitación, utilizando las variables disponibles en PLACSP y siguiendo metodologías ya consolidadas en la literatura internacional.

- OE3. Agregar estos indicadores a nivel de órgano-año y construir un Composite Risk Index mediante tres enfoques de ponderación: pesos iguales, pesos basados en la jerarquía empírica del marco de Fazekas y pesos derivados de la varianza observada en los datos.

- OE4. Evaluar la estabilidad de los rankings obtenidos mediante correlaciones de Spearman entre esquemas de ponderación, análisis de estabilidad en los percentiles superiores y pruebas de sensibilidad frente a variaciones en los pesos, todo ello en ausencia de etiquetas de validación directa.

- OE5. Desarrollar un dashboard interactivo en Kibana que permita explorar los resultados por órgano de contratación, tipo de contrato y año, facilitando su interpretación y su uso práctico.

1.3ESTRUCTURA DE LA MEMORIA

La memoria se estructura en nueve capítulos, que recorren el proyecto desde el contexto inicial hasta los resultados obtenidos.

El Capítulo 2 presenta el marco tecnológico y describe las principales herramientas del ecosistema Big Data utilizadas en el desarrollo del trabajo. El Capítulo 3 se centra en el marco teórico y en el estado del arte sobre indicadores de riesgo en contratación pública, con especial atención a las aportaciones de Fazekas y a las guías publicadas por organismos

internacionales como OCP y el Banco Mundial. El Capítulo 4 delimita formalmente el proyecto, incluyendo sus objetivos, alcance, limitaciones y planificación.

Los Capítulos 5 y 6 recogen el núcleo técnico del trabajo. En ellos se describen las fuentes de datos y la arquitectura del sistema, el diseño del pipeline de procesamiento, la definición de los indicadores de riesgo y, por último, la construcción y evaluación del Composite Risk Index.

El Capítulo 7 aborda el análisis de los resultados, desde la exploración inicial de los datos hasta la obtención de rankings y la validación cualitativa de casos extremos. Finalmente, el Capítulo 8 presenta las conclusiones principales, y algunas líneas de trabajo futuro y el capítulo 9 la bibliografía del proyecto.

Capítulo 2. MARCO TECNOLÓGICO

La arquitectura técnica del proyecto sigue un pipeline de datos por capas. La ingesta y automatización se apoyan en Apache NiFi, el almacenamiento distribuido en HDFS, el procesamiento en Apache Spark y la indexación y visualización en Elasticsearch y Kibana.

La elección de este stack responde a tres criterios. El primero es la escalabilidad, ya que permite ampliar el sistema a nuevos años o a otros ámbitos de la contratación pública sin rediseñar la arquitectura. El segundo es la reproducibilidad, porque cada fase del pipeline puede ejecutarse de forma independiente a partir de los datos originales. El tercero es la alineación con herramientas ampliamente utilizadas en entornos reales, lo que facilita trasladar el trabajo a contextos de uso más allá del TFM.

2.1 HADOOP Y HDFS

HDFS es el sistema de almacenamiento distribuido sobre el que se apoya el pipeline. Está diseñado para trabajar con grandes volúmenes de datos en clústeres de hardware convencional y ofrece tolerancia a fallos mediante replicación de bloques [8].

En este proyecto, HDFS actúa como capa de almacenamiento persistente y vertebra el pipeline. Los datos se organizan en tres niveles: una capa raw con los ficheros originales, una capa curated con los datos filtrados y enriquecidos en formato Parquet, y una capa analytics con las tablas analíticas y el Composite Risk Index. Esta organización facilita la trazabilidad del sistema y permite reejecutar cada fase del pipeline de forma independiente. Una descripción detallada de la arquitectura de HDFS y de su mecanismo de replicación puede consultarse en [8].

2.2APACHE NIFI

Apache NiFi es una plataforma orientada a la automatización de flujos de datos entre sistemas heterogéneos, gestionada a través de una interfaz gráfica [9]. Sus capacidades de trazabilidad, gestión de errores y control del flujo la hacen adecuada para pipelines de ingesta.

En este proyecto, NiFi cumple dos funciones. Por un lado, automatiza la ingesta en HDFS de los ficheros ya convertidos a formato tabular, depositándolos en la capa raw mediante un flujo que enruta cada fichero a su destino según su contenido. Por otro, orquesta la ejecución del pipeline de procesamiento: encadena la verificación de integridad y los sucesivos jobs de Spark de modo que cada etapa se ejecuta únicamente si la anterior finaliza correctamente. Aunque durante el desarrollo inicial estas fases se ejecutaron de forma manual, su automatización con NiFi reduce la incorporación de nuevos datos y el reprocesamiento del pipeline a una intervención mínima. El detalle del flujo se desarrolla en el Capítulo 5.

2.3APACHE SPARK Y PYSPARK

Apache Spark es el motor de procesamiento distribuido utilizado en el proyecto. A través de la API de DataFrames de PySpark y de Spark SQL, permite expresar de forma declarativa operaciones como filtros, joins, agregaciones, funciones de ventana o cálculo de percentiles [10].

El pipeline se organiza en cinco jobs principales: filtrado y limpieza del subconjunto de trabajo, enriquecimiento con información de órganos de contratación, cálculo de los indicadores de riesgo y su agregación, construcción del Composite Risk Index e indexación de los resultados en Elasticsearch.

La justificación de esta elección frente a alternativas de nodo único se desarrolla, por su carácter metodológico, en la sección 4.4.

2.4ELASTICSEARCH Y KIBANA

Elasticsearch es un motor distribuido de búsqueda y análisis orientado a la indexación eficiente de datos en formato JSON [11].Kibana es la herramienta de visualización del ecosistema Elastic y permite construir dashboards interactivos sobre los datos indexados.

En este trabajo, Elasticsearch actúa como capa de servicio de los resultados y Kibana como interfaz de exploración. El dashboard desarrollado permite consultar los rankings de riesgo por órgano de contratación, filtrar por año, ministerio y calidad de la estimación, y explorar la distribución del índice, su evolución temporal y el desglose de los indicadores que lo componen para cada órgano-año.

El objetivo es hacer que esta información sea accesible para perfiles no técnicos, como analistas de integridad, y facilitar su uso en tareas de supervisión y priorización de auditorías.

2.5OPENPLACSP Y LA PLATAFORMA DE CONTRATACIÓN DEL SECTOR PÚBLICO

La Plataforma de Contratación del Sector Público (PLACSP) es la fuente de datos del proyecto. A través de ella, la Dirección General del Patrimonio del Estado publica conjuntos de datos en formato ATOM con información sobre las licitaciones registradas desde 2012 [4].

Dado que estos ficheros no están preparados para su análisis directo, se utiliza OpenPLACSP para transformarlos en tablas estructuradas. Durante la fase inicial de ingesta se procesaron los seis ficheros correspondientes al periodo 2018-2023. Como resultado, se obtuvieron 802.617 licitaciones en la tabla de Licitaciones y 1.060.218 registros en la tabla de Resultados. La validación de integridad mostró una cobertura completa en variables clave como Estado, Tipo de contrato, Tipo de procedimiento, Tipo de Administración, Presupuesto base y Objeto del contrato, y superior al 85 por ciento en variables como Fecha de

presentación de ofertas e Importe de adjudicación. En conjunto, estos niveles de calidad se consideran suficientes para calcular los indicadores de riesgo definidos en el proyecto.

Capítulo 3. MARCO TEÓRICO Y ESTADO DEL ARTE

3.1 CONTRATACIÓN PÚBLICA EN ESPAÑA: CONTEXTO REGULATORIO

La contratación pública en España se articula en torno a la Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público (LCSP), que incorpora al ordenamiento jurídico español las Directivas 2014/23/UE y 2014/24/UE, sustituyendo el régimen anterior basado en las Directivas 2004/17/CE y 2004/18/CE [2].Su entrada en vigor efectiva en marzo de 2018 marca el inicio de un periodo regulatorio más homogéneo, que constituye el marco temporal de este trabajo.

Más allá de la transposición, la LCSP introdujo cambios con impacto directo en los indicadores de riesgo utilizados en este proyecto. Por un lado, reforzó las obligaciones de publicidad y transparencia, estableciendo la publicación obligatoria en la Plataforma de Contratación del Sector Público y avanzando hacia la tramitación electrónica del expediente. Por otro, limitó el uso de procedimientos discrecionales: el contrato menor quedó sometido a límites más estrictos, desapareció el negociado sin publicidad por razón de la cuantía y el artículo 64 incorporó de forma expresa la lucha contra la corrupción y la prevención de conflictos de intereses como principios del proceso [2].

El periodo 2018-2023 resulta por ello especialmente adecuado para el análisis comparativo: el núcleo procedimental se mantiene estable, lo que permite atribuir las diferencias en los indicadores al comportamiento de los órganos de contratación. Las modificaciones temporales introducidas por la legislación de emergencia durante la pandemia constituyen una discontinuidad puntual documentada en el Capítulo 4. La incorporación de las directivas europeas garantiza además la comparabilidad con estudios en otros Estados miembros, ya que procedimientos, umbrales y categorías contractuales responden a una base común, lo que permite aplicar al caso español los indicadores desarrollados en la literatura internacional sin adaptaciones de fondo [12].

A pesar de esta disponibilidad técnica, los estudios empíricos sistemáticos sobre contratación pública en España siguen siendo escasos. Algunos trabajos recientes han recurrido al Boletín Oficial del Estado como fuente alternativa [13],pero el uso de los datos primarios de PLACSP para calcular indicadores de riesgo de forma reproducible sigue siendo un ámbito poco explorado, lo que constituye una de las motivaciones específicas de este trabajo.

3.2 INDICADORES DE RIESGO EN CONTRATACIÓN PÚBLICA

3.2.1 DEFINICIÓN Y FUNDAMENTO TEÓRICO

Un indicador de riesgo, o red flag, es una variable estadística que señala la presencia de un patrón asociado a una mayor probabilidad de irregularidad o de restricción artificial de la competencia [6]. La lógica de fondo es que la corrupción en contratación pública, para sostenerse en el tiempo, requiere algún tipo de manipulación sistemática de los procedimientos: reducir plazos para limitar la concurrencia, recurrir a procedimientos sin publicidad o adjudicar de forma reiterada al mismo proveedor dejan huella en los datos administrativos aunque el comportamiento irregular no sea directamente observable [5].

Este enfoque se basa en una medición indirecta. Una red flag no prueba que exista una irregularidad: identifica un patrón anómalo que, sin una explicación legítima, eleva la probabilidad de que existan problemas de integridad. El Banco Mundial ha señalado las limitaciones de interpretar estos indicadores de forma aislada: Kenny y Musatova encontraron que prácticamente todos los contratos revisados presentaban al menos una de las trece red flags analizadas, y que los contratos potencialmente problemáticos no acumulaban un número significativamente mayor de señales que los contratos de control [14]. Esto confirma que la utilidad de las red flags aumenta cuando se analizan en conjunto y de forma agregada, lo que justifica la construcción de índices compuestos.

3.2.2 EL MARCO DE REFERENCIA DE FAZEKAS

El trabajo de referencia en la aplicación sistemática de red flags a datos de contratación pública es el de Fazekas, Tóth y King, que formalizan el Corruption Risk Index como un índice compuesto calculado a nivel de contrato a partir de indicadores binarios que reflejan la restricción artificial de la competencia [15]. Su premisa central es que la corrupción sostenida requiere adjudicaciones recurrentes a una misma red de empresas, lo que implica eludir sistemáticamente las normas de publicidad y concurrencia y deja una huella estadística detectable.

Fazekas y Kocsis aplicaron esta metodología a 2,8 millones de contratos de 28 países europeos para el periodo 2009-2014, construyendo indicadores comparables gracias al marco regulatorio común de las directivas europeas. Los índices obtenidos mostraron correlaciones significativas con medidas de corrupción consolidadas a nivel país, aportando evidencia de validez externa [5]. Este trabajo adopta directamente esta metodología, adaptando los indicadores a las variables disponibles en PLACSP y al contexto de la LCSP. Trabajos posteriores han extendido el enfoque: Basdevant y colaboradores muestran su correlación con el sobreprecio en contratos de países en desarrollo [16], y Decarolis y Giorgiantonio demuestran que ampliar el conjunto de indicadores con datos de los pliegos mejora la capacidad predictiva del modelo [17].

3.2.3 EL MARCO OPERATIVO DE OPEN CONTRACTING PARTNERSHIP

La Open Contracting Partnership ha desarrollado una guía operativa con definiciones, fórmulas de cálculo y criterios de interpretación para un catálogo de 73 indicadores de riesgo, organizados siguiendo el ciclo completo de la contratación pública [6]. Frente al enfoque de Fazekas, orientado a la comparación entre países, la guía de OCP tiene una vocación más práctica: proporcionar herramientas concretas para la monitorización a organismos de auditoría y sociedad civil. Sus indicadores están mapeados al estándar Open Contracting Data Standard, lo que facilita su aplicación en distintos contextos nacionales.

En este proyecto, la guía de OCP sirve como referencia operativa para la definición de las cinco red flags implementadas, complementando el marco teórico de Fazekas con definiciones más detalladas y fórmulas directamente aplicables sobre los campos disponibles en PLACSP.

3.3ÍNDICES COMPUESTOS DE RIESGO: TRABAJOS PREVIOS

3.3.1 AGREGACIÓN SIN ETIQUETAS DE CORRUPCIÓN

La construcción de un índice compuesto de riesgo se enfrenta a una dificultad metodológica central: la ausencia de un ground truth etiquetado. Las condenas judiciales son escasas, tardías y potencialmente sesgadas hacia los casos que han salido a la luz, no hacia la distribución real de la irregularidad. Una revisión sistemática reciente sobre detección de fraude en contratación pública confirma que la falta de datos etiquetados es la limitación más extendida del campo, y que la mayoría de los trabajos recurren por ello a enfoques no supervisados o semi-supervisados [18].

La solución adoptada por Fazekas, y replicada en este trabajo, es construir un índice no supervisado basado en la agregación ponderada de indicadores, cuya validez se evalúa mediante criterios internos de robustez y criterios externos de coherencia [15]. Este enfoque es equivalente al utilizado en otros índices compuestos en ciencias sociales, como el Índice de Desarrollo Humano o los Worldwide Governance Indicators, donde tampoco existe una medida directa del fenómeno y es necesario trabajar con proxies y evaluar la consistencia de los resultados bajo distintos supuestos.

3.3.2 ESQUEMAS DE AGREGACIÓN Y ANÁLISIS DE SENSIBILIDAD

El esquema más sencillo asigna pesos iguales a todos los indicadores, con la ventaja de la transparencia y la comparabilidad. Es la línea base utilizada por Fazekas y Kocsis y por la

mayoría de los trabajos posteriores [5]. Las alternativas incluyen pesos derivados del análisis de componentes principales, esquemas basados en entropía o varianza, y esquemas teóricos que reflejan la jerarquía empírica de la literatura. Fazekas y Kocsis identifican el single bidding como el indicador con mayor respaldo empírico, mientras que Decarolis y Giorgiantonio señalan los procedimientos discrecionales y el incumplimiento de plazos mínimos entre las señales más eficaces [17], lo que justifica asignarles un mayor peso frente a indicadores de resultado.

El análisis de sensibilidad de los rankings ante distintos esquemas es la principal herramienta de validación interna cuando no se dispone de etiquetas. Si los mismos órganos aparecen en las posiciones extremas con independencia del esquema utilizado, hay evidencia de que el índice captura una señal estructural. Las métricas estándar son la correlación de Spearman entre rankings y el análisis de estabilidad en los percentiles superiores, recogidas en el manual del OECD y el Joint Research Centre como marco de referencia para este tipo de índices [19]. Este trabajo aplica tres esquemas concretos: pesos iguales como línea base, pesos teóricos basados en la jerarquía empírica de Fazekas y pesos derivados de la varianza observada en los datos. La justificación detallada y los resultados se desarrollan en el Capítulo 6.

3.4 LITERATURA SOBRE INDICADORES ESPECÍFICOS

3.4.1 SINGLE BIDDING

La presencia de un único licitador en un procedimiento competitivo es el indicador más estudiado en la literatura y el que cuenta con mayor respaldo empírico como proxy de restricción artificial de la competencia [5]. Bauhr y colaboradores, a partir de más de tres millones y medio de contratos europeos, muestran que la transparencia ex ante del proceso es uno de los principales factores asociados a su reducción, lo que refuerza la relación entre opacidad procedimental y menor competencia [7]. En el subconjunto español analizado, las tasas de single bidding en la AGE se sitúan en torno al cuarenta por ciento para el periodo

2018-2023, con una desviación típica entre órganos de unos veinte puntos porcentuales, lo que indica una capacidad discriminante elevada para este indicador.

3.4.2 PROCEDIMIENTOS NO COMPETITIVOS

El recurso a procedimientos que evitan la convocatoria pública abierta, como el negociado sin publicidad, el contrato menor o la adjudicación directa, es uno de los mecanismos de restricción de la competencia más señalados en la literatura [6]. Decarolis y Giorgiantonio encuentran una asociación sistemática entre su uso y una mayor probabilidad de irregularidad en contratos de obras en Italia [17]. En el contexto español, la LCSP establece umbrales específicos que permiten definir este indicador con precisión. Conviene matizar que la categoría "Derivado de acuerdo marco" no se incluye como procedimiento no competitivo en este trabajo, ya que la competencia se produce en la fase previa de licitación del propio acuerdo.

3.4.3 PRECIO RELATIVO DE ADJUDICACIÓN

El ratio entre el importe de adjudicación y el presupuesto base aproxima el nivel de competencia efectiva en precio. Con competencia real, los licitadores tienen incentivos para presentar ofertas por debajo del máximo, sin ella, el contrato puede adjudicarse cerca del presupuesto base. Fazekas y Kocsis señalan la ausencia de descuento como una red flag coherente con patrones de reparto de mercado [5], y Basdevant y colaboradores cuantifican que los órganos con mayor riesgo compuesto pagan precios significativamente superiores a los de órganos de bajo riesgo para contratos comparables [16].

3.4.4 CONCENTRACIÓN DE PROVEEDORES

La concentración de contratos en un número reducido de proveedores, medida mediante el índice de Herfindahl-Hirschman o la cuota del proveedor principal, es un indicador a nivel de comprador que captura la falta de competencia sostenida a lo largo del tiempo. Kenny y Musatova lo identifican como uno de los de mayor capacidad discriminante en su análisis de proyectos del Banco Mundial, especialmente cuando las red flags individuales resultan poco informativas por separado [14]. Su relevancia en este proyecto es especialmente clara

porque se calcula directamente sobre la unidad de análisis principal, el órgano-año, lo que reduce el ruido inherente a los indicadores binarios a nivel de licitación.

3.4.5 PLAZO DE PRESENTACIÓN DE OFERTAS

Un plazo anormalmente breve entre la publicación de la convocatoria y el cierre de la presentación de ofertas limita la capacidad de nuevos licitadores para preparar propuestas competitivas y favorece a quienes ya conocen el contrato de antemano [6]. Fazekas y Kocsis lo incluyen entre los indicadores del Corruption Risk Index [5], y Bauhr y colaboradores muestran que disponer de tiempo suficiente para acceder a la información es uno de los factores clave para reducir el single bidding [7]. En este trabajo, el indicador se operacionaliza comparando el plazo efectivo con la distribución observada en cada tipo de contrato y marcando como red flag las licitaciones por debajo del percentil 10 de la población de referencia.

3.4.6 FRAGMENTACIÓN DE CONTRATOS

Un sexto indicador documentado en la literatura, no implementado en este trabajo, es la fragmentación de contratos o contract splitting. La división artificial de un contrato de mayor importe en varios menores para eludir los umbrales que obligan a procedimientos competitivos, es un mecanismo documentado en la literatura. Caires aporta evidencia en el contexto portugués: tras una reforma que rebajó los umbrales discrecionales, muestra que la fragmentación fue el principal mecanismo de adaptación observado y que su motivación parece más vinculada al favoritismo que a razones de eficiencia [20]. Su detección automatizada requiere combinar proximidad temporal, importes cercanos al umbral regulatorio y similitud semántica del objeto del contrato, lo que obliga a incorporar técnicas de procesamiento de lenguaje natural. Por su complejidad técnica y su dependencia de la calidad del campo textual en PLACSP, este indicador se reserva como extensión futura del pipeline.

Conviene precisar el alcance estadístico de todos los indicadores descritos: ninguno mide corrupción de forma directa, sino que funcionan como proxies asociados a contextos de

mayor riesgo. Su interpretación es probabilística, no determinista, y solo resultan informativos en agregado y sobre muestras suficientes. La coincidencia de varios indicadores elevados refuerza la señal, pero no confirma la irregularidad. El resultado del sistema es una herramienta de priorización para revisiones posteriores, no un veredicto.

3.5 HUECO QUE CUBRE ESTE PROYECTO

La revisión de la literatura realizada en este capítulo permite identificar tres aportaciones concretas de este trabajo con respecto al estado del arte.

La primera aportación es de alcance geográfico. La metodología de Fazekas se ha aplicado de forma sistemática en países de Europa central y oriental [5], en contextos latinoamericanos a través de guías como la de la Red Interamericana de Compras Gubernamentales [21], y en Italia en trabajos más recientes [17]. Los enfoques revisados presentan ventajas y limitaciones complementarias. El marco de Fazekas, Tóth y King destaca por su objetividad y escalabilidad, pero fue pensado para comparaciones agregadas entre países y apenas explora la sensibilidad del índice a distintos esquemas de ponderación [15]. La guía de Open Contracting Partnership ofrece un catálogo de indicadores muy completo y operativo, aunque su planteamiento es esencialmente descriptivo y no propone ni un método de agregación ni una estrategia de validación [6]. Por su parte, los enfoques basados en aprendizaje automático, como el de Decarolis y Giorgiantonio, mejoran la capacidad predictiva, pero dependen de información más rica y menos accesible, lo que dificulta su reproducibilidad [17].

Sin embargo, España, pese a contar desde 2012 con una plataforma pública de datos abiertos de contratación de cobertura nacional como PLACSP, no ha sido analizada de forma sistemática con esta metodología a partir de datos de PLACSP y con foco en la Administración General del Estado.

La segunda aportación es técnica. Buena parte de los trabajos previos sobre red flags en contratación pública se apoyan en datasets de tamaño reducido, construidos específicamente para cada investigación, o en bases de datos propietarias mantenidas por organismos internacionales. La revisión sistemática más reciente del área, que analiza 93 trabajos sobre detección de fraude en contratación pública mediante métodos basados en datos, evidencia la heterogeneidad de los conjuntos de datos utilizados y la dependencia frecuente de bases propietarias o construidas ad hoc para cada investigación [18]. Frente a ello, este trabajo diseña y documenta un pipeline de datos reproducible, implementado con tecnologías Big Data de código abierto, que puede ser reejecutado por cualquier investigador o institución con acceso a los datos públicos de PLACSP. En este sentido, el propio diseño del pipeline, con separación en capas raw, curated y analytics, jobs de Spark independientes y auditables, y persistencia en formato columnar, constituye también una aportación metodológica.

La tercera aportación se centra en la evaluación de la robustez del índice compuesto bajo distintos esquemas de ponderación. Mientras que muchos trabajos anteriores presentan resultados a partir de un único esquema, normalmente basado en pesos iguales, este proyecto aplica tres esquemas diferentes sobre el mismo conjunto de datos y analiza la estabilidad de los rankings obtenidos mediante correlaciones de Spearman entre años consecutivos. En ausencia de un ground truth de corrupción, este tipo de validación interna es una de las estrategias metodológicas más sólidas disponibles para evaluar si la señal capturada por el índice responde a un patrón estructural o si depende en exceso del método de agregación utilizado. Por ello, esta parte del trabajo también aporta una contribución directa al debate sobre la validez de los índices compuestos de riesgo en contratación pública.

Frente a ello, este trabajo combina elementos de esas tres líneas. Utiliza indicadores objetivos en la tradición de Fazekas y OCP, los calcula exclusivamente a partir de datos abiertos de PLACSP mediante un pipeline reproducible, y añade un análisis que no se ha abordado de forma conjunta para el caso español: la construcción del índice con tres esquemas de ponderación alternativos y la evaluación explícita de su estabilidad. Así, la robustez del ranking no se presupone, sino que se mide.

Conviene delimitar con precisión qué parte del trabajo constituye una aportación propia y cuál responde a la adopción de marcos ya consolidados en la literatura. La definición de los cinco indicadores procede directamente del marco de Fazekas y de la guía de la Open Contracting Partnership, por lo que no se presenta aquí como una novedad. La contribución metodológica del trabajo se sitúa en otro nivel: en la evaluación explícita de la robustez del índice mediante tres esquemas de ponderación, en lugar de asumir uno solo; en el tratamiento de la heterogeneidad de tamaño a través de un umbral muestral y de una etiqueta de calidad de la estimación; y en la implementación de todo el proceso como un pipeline reproducible con tecnologías de código abierto. La aportación no reside, por tanto, en los indicadores que se miden, sino en la forma en que esa medición se construye, se valida y se hace reproducible.

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 JUSTIFICACIÓN

La justificación de este proyecto se apoya en tres ideas complementarias que apuntan en la misma dirección: actualmente existen unas condiciones especialmente favorables para desarrollar un sistema de análisis de riesgo en contratación pública española a partir de datos abiertos.

La primera tiene que ver con el marco normativo. La entrada en vigor de la LCSP en marzo de 2018 marcó el inicio de un periodo de estabilidad regulatoria en el núcleo procedimental de la contratación pública española. El régimen de procedimientos, los principios de publicidad y la estructura general de obligaciones se mantienen sin alteraciones sustanciales hasta la actualidad, sin perjuicio de modificaciones específicas introducidas por la legislación de emergencia durante la pandemia de COVID-19, cuyo impacto se documenta como limitación en la sección 4.3.2. Esto hace posible construir indicadores comparables entre órganos y entre años sin tener que introducir ajustes continuos por cambios normativos. Por ello, el intervalo 2018-2023 puede considerarse una ventana especialmente adecuada para este tipo de análisis.

La segunda razón es la disponibilidad de los datos. La Dirección General del Patrimonio del Estado publica desde 2012 los datos estructurados de las licitaciones registradas en la Plataforma de Contratación del Sector Público, y la herramienta OpenPLACSP permite transformarlos a formatos tabulares sin coste ni restricciones de uso. La validación realizada en este trabajo muestra que el dataset correspondiente al periodo 2018-2023 contiene más de 800.000 licitaciones. De ellas, el subconjunto filtrado a la Administración General del Estado y a los tres tipos principales de contrato alcanza 123.966 registros antes del proceso de deduplicación. Además, las columnas críticas para el cómputo de los cinco indicadores de riesgo presentan coberturas elevadas, superiores al 95 por ciento para las variables de

procedimiento y resultado y por debajo de ese umbral para las variables temporales y de adjudicatario, tal y como se documenta en la sección 4.3.2.

La tercera razón es la brecha existente a nivel tecnológico e institucional. La literatura internacional ha mostrado que el análisis sistemático de red flags sobre datos de contratación pública es viable desde el punto de vista técnico y útil para los organismos de control. Así lo demuestran, por ejemplo, las aplicaciones del Corruption Risk Index en veintiocho países europeos y distintas implementaciones institucionales en otros contextos [5]. Sin embargo, en España no existe, un sistema equivalente basado en datos de PLACSP que sea abierto, reproducible y esté documentado de forma completa. Esa es precisamente la brecha que este trabajo busca reducir.

En conjunto, la combinación de datos públicos accesibles, herramientas de procesamiento distribuido de código abierto y evidencia empírica internacional sobre la utilidad de este enfoque hace que el proyecto resulte técnicamente viable, metodológicamente sólido y relevante desde un punto de vista institucional.

4.2 OBJETIVOS

El objetivo general del proyecto, ya planteado en la sección 1.2, es diseñar e implementar un sistema de análisis de riesgo en contratación pública española mediante técnicas de Big Data que permita generar rankings de órganos de contratación de la Administración General del Estado según su nivel de riesgo relativo. Este sistema se concibe como una herramienta de apoyo para la priorización de auditorías.

Los cinco objetivos específicos se plantean de forma progresiva, de manera que cada uno se apoya en el anterior y amplía el alcance del sistema. El OE1, centrado en el pipeline de datos, constituye la base técnica del proyecto, ya que sin datos limpios, estructurados y accesibles en HDFS no es posible calcular los indicadores de riesgo. El OE2, dedicado al cálculo de los

cinco indicadores definidos, representa el núcleo metodológico del trabajo y el punto de conexión más directo con la literatura de referencia.

A partir de ahí, el OE3 aborda la construcción del Composite Risk Index mediante tres esquemas de ponderación: pesos iguales, pesos basados en la jerarquía empírica propuesta en el marco de Fazekas y pesos derivados de la varianza observada en los datos. Este objetivo constituye una de las principales aportaciones metodológicas del proyecto frente a otros enfoques que trabajan con un único esquema de agregación. El OE4 se centra en la evaluación de la estabilidad de los rankings, utilizando correlaciones de Spearman entre años consecutivos y entre distintos esquemas de ponderación. Con ello se busca aportar una validación interna del índice en ausencia de un ground truth directo. Por último, el OE5 completa el ciclo mediante la visualización de los resultados en Kibana, de forma que puedan ser consultados e interpretados también por usuarios no técnicos.

4.3 ALCANCE Y LIMITACIONES

4.3.1 ALCANCE DEL PROYECTO

El ámbito de aplicación del sistema se limita, en su alcance principal, a las licitaciones de la Administración General del Estado con resultado conocido, es decir, aquellas cuyo estado aparece como Resuelta o Adjudicada, para los contratos de Obras, Servicios y Suministros durante el periodo 2018-2023. Esta delimitación responde principalmente a criterios de homogeneidad. Por un lado, la Administración General del Estado es el único nivel administrativo para el que PLACSP ofrece una cobertura completa y consistente a lo largo de todo el periodo analizado. Por otro, los tres tipos de contrato mencionados concentran la gran mayoría del subconjunto resultante, ya que las categorías excluidas, como concesiones, contratos administrativos especiales y otros tipos residuales, representan una fracción reducida del total.

A efectos de este trabajo, conviene precisar la terminología empleada. Se entiende por licitación cada procedimiento de contratación publicado en PLACSP. Cuando una licitación se divide en lotes, cada uno de ellos puede adjudicarse de manera independiente, por lo que la unidad de cálculo más desagregada es la licitación-lote. Una vez deduplicados los snapshots acumulativos, cada licitación-lote única pasa a denominarse contrato. La unidad de análisis y agregación del sistema es, sin embargo, el órgano-año, esto es, cada órgano de contratación considerado en cada uno de los años del periodo analizado. En esta distinción, el término órgano de contratación se refiere a la entidad, mientras que órgano-año designa la unidad sobre la que se calculan las tasas y se construye el ranking.

Tras el filtrado realizado en el Job 1, el subconjunto curated está formado por 123.966 registros de licitaciones-lote. Después del proceso de deduplicación, descrito en el Capítulo 6, esa cifra se reduce a 104.909 contratos únicos. La posterior agregación a nivel de órgano-año genera 4.077 grupos, de los cuales 743 superan el umbral mínimo de 30 contratos por año exigido para formar parte del ranking principal y cubren aproximadamente el 80 por ciento del número total de contratos analizados.

El sistema se plantea en dos niveles de alcance. El primero incluye el pipeline completo de datos, formado por los Jobs 1, 2 y 3 de Spark, las red flags RF1 a RF5, calculadas tanto a nivel de licitación como de órgano-año, el Composite Risk Index obtenido mediante el Job 4 con tres esquemas de agregación, la indexación de los resultados en Elasticsearch mediante el Job 5, el análisis de estabilidad de rankings y el dashboard de visualización en Kibana.

El segundo nivel, condicionado por la disponibilidad de tiempo y por la calidad real de los datos textuales, contemplaba dos extensiones: la red flag RF6 de fragmentación de contratos a partir de la similitud semántica del campo de objeto del contrato, y la orquestación completa del pipeline mediante Apache NiFi. De ambas, la orquestación con NiFi se completó dentro del proyecto y se describe en el Capítulo 5, mientras que RF6 se mantiene como línea de trabajo futuro.

4.3.2 LIMITACIONES

La primera limitación del sistema es la ausencia de un ground truth que permita confirmar o descartar la existencia de irregularidades en los contratos analizados. En España, las condenas judiciales firmes relacionadas con contratación pública son escasas, tienen un acceso limitado y, además, pueden estar sesgadas hacia los casos que han sido detectados, no hacia la distribución real de la irregularidad. Por este motivo, la validación del índice se apoya únicamente en criterios internos de robustez, como la estabilidad de los rankings bajo distintos esquemas de ponderación y entre años consecutivos, así como en una revisión cualitativa de los casos extremos contrastada con fuentes públicas, entre ellas informes del Tribunal de Cuentas y resoluciones de la CNMC. En consecuencia, el resultado del proyecto debe interpretarse como un ranking de riesgo relativo y no como una clasificación de contratos corruptos.

Otra limitación importante es la heterogeneidad en el tamaño de los órganos de contratación. El número de licitaciones varía mucho entre unos órganos y otros: algunos adjudican varios cientos de contratos al año, mientras que otros apenas tramitan unos pocos. Esta diferencia afecta directamente a la fiabilidad de los indicadores, ya que se trata de proporciones o índices de concentración cuya variabilidad aumenta cuando el número de contratos es reducido. Así, una tasa de licitación única del cien por cien en un órgano con tres contratos puede deberse al azar, mientras que ese mismo valor en un órgano con trescientos contratos constituye una señal mucho más sólida.

Para reducir este problema, el proyecto establece un umbral mínimo de 30 contratos por órgano-año para la inclusión en el ranking principal. Este criterio se justifica por el error estándar admisible en una proporción de Bernoulli. Además, el sistema incorpora un campo de calidad de la estimación que clasifica cada grupo según su tamaño muestral. De este modo, las puntuaciones calculadas a partir de muestras pequeñas se conservan en el sistema, pero quedan fuera del ranking interpretable. El funcionamiento detallado de este mecanismo se explica en el capítulo de implementación.

Una tercera limitación viene dada por la estructura acumulativa de los snapshots anuales de PLACSP. Los ficheros publicados por el Ministerio de Hacienda son acumulativos, de modo que un contrato adjudicado en un año puede volver a aparecer en el fichero del año siguiente con información actualizada. Si este fenómeno no se corrige, los indicadores agregados quedan distorsionados por duplicados estructurales. Para resolverlo, el proyecto aplica un proceso de deduplicación en dos pasos: primero se eliminan las filas idénticas y después se conserva únicamente el snapshot más reciente para cada par identificador-lote. Como resultado, el dataset pasa de 123.966 a 104.909 registros. Esta solución introduce, no obstante, una asimetría temporal, ya que algunos contratos atribuibles a un año pueden quedar asociados al snapshot del año siguiente. Esta circunstancia se documenta como una característica metodológica conocida del análisis.

Otra limitación importante afecta a la identificación de los órganos de contratación entre años. El campo identificador del órgano en PLACSP, que permite distinguir de forma unívoca cada órgano, no está disponible en los datos de 2018 y 2019 y solo presenta cobertura parcial en 2020 y 2021. En esos casos se utiliza el NIF del órgano de contratación como identificador alternativo. Esta solución permite mantener el análisis, pero puede agrupar bajo un mismo NIF a varios órganos dependientes de una misma entidad, lo que reduce la granularidad en los años más antiguos. Además, la cobertura del nombre del órgano tras el cruce con el maestro de PLACSP se sitúa en torno al 80 por ciento del subconjunto deduplicado. Los registros sin nombre asociado se conservan utilizando su identificador, aunque su interpretación requiere trabajo manual adicional.

También existen limitaciones relacionadas con la cobertura de la fecha de presentación de ofertas. El cálculo del indicador RF5 requiere combinar la fecha de publicación del anuncio con la fecha límite de presentación. Esta segunda variable presenta una cobertura del 75,8 por ciento sobre el dataset deduplicado, y el plazo en días solo puede calcularse finalmente para el 67,4 por ciento de los contratos, una vez descartados los casos con información incompleta o valores inconsistentes. Los registros sin plazo computable se excluyen tanto del numerador como del denominador en la agregación de RF5. Esta decisión es

metodológicamente la más adecuada para evitar sesgos, aunque implica que, en algunos órganos, la tasa agregada se calcula sobre un número reducido de observaciones.

El periodo de emergencia sanitaria constituye otra fuente de distorsión que conviene tener en cuenta. La pandemia de COVID-19 dio lugar a cambios regulatorios temporales que ampliaron el uso de procedimientos de emergencia y redujeron los plazos de presentación en determinadas categorías de contratos durante 2020 y 2021. Gnaldi y Del Sarto muestran que, en contextos de crisis, el uso de red flags convencionales puede sobreestimar el riesgo de corrupción, ya que algunos valores anómalos pueden responder a adaptaciones legítimas al marco regulatorio excepcional y no necesariamente a comportamientos irregulares. Por esta razón, en este trabajo los resultados se presentan desagregados por año, para facilitar una interpretación más contextualizada [22]. Además, se documenta la tasa anómalamente elevada de procedimientos no competitivos observada en 2018, que podría estar relacionada con el periodo de transición a la LCSP, dado que la norma entró en vigor en marzo de ese año y convivió durante varios meses con expedientes iniciados bajo la legislación anterior, aunque esa atribución causal no se evalúa de forma cuantitativa en este trabajo.

Por último, la posible incorporación de la red flag RF6, basada en similitud textual para detectar fragmentación de contratos, depende en gran medida de la calidad del campo objeto del contrato en PLACSP. Cuando las descripciones son demasiado genéricas o repetitivas, como ocurre con frecuencia en ciertas categorías de suministros, la similitud textual pierde capacidad discriminante. Por ello, RF6 se plantea como una extensión condicionada a que los datos ofrezcan un nivel de detalle suficiente. En caso contrario, se documenta como línea de trabajo futuro y queda fuera del índice final.

4.4METODOLOGÍA

La metodología del proyecto se basa en un marco teórico ya consolidado en la literatura internacional sobre integridad en contratación pública y se implementa mediante un flujo de

trabajo distribuido en cuatro fases secuenciales. El marco teórico define qué se mide y cómo deben interpretarse los resultados, mientras que el flujo de trabajo concreta cómo se lleva todo eso a la práctica.

Este marco combina tres referencias principales. La primera es el Corruption Risk Index desarrollado por Fazekas, Tóth y King, y posteriormente validado a escala europea por Fazekas y Kocsis. De este enfoque se toma la idea central de medir el riesgo de irregularidad a partir de patrones estadísticos anómalos observables en los contratos y agregarlos después por órgano de contratación en un índice compuesto [15] [5]. La segunda referencia es la guía operativa de Open Contracting Partnership, que aporta definiciones técnicas y fórmulas de cálculo para los indicadores a partir de datos estructurados de contratación [6]. La tercera es el manual de la OECD y el Joint Research Centre de la Comisión Europea sobre construcción de índices compuestos, que sirve de base para la normalización de indicadores, la definición de esquemas de ponderación y el análisis de sensibilidad [19]. La descripción detallada de cada indicador, así como la construcción y validación del Composite Risk Index, se desarrolla en el Capítulo 6.

A partir de este marco, el trabajo se organiza en cuatro fases que cubren todo el proceso, desde la ingesta de los datos brutos publicados por el Ministerio de Hacienda hasta la visualización interactiva de los resultados.

La primera fase corresponde a la ingesta y al almacenamiento. Los ficheros anuales publicados en formato ATOM se procesan con OpenPLACSP para generar los conjuntos de datos tabulares, que posteriormente se almacenan en la capa raw de HDFS. Durante el desarrollo inicial esta fase se ejecutó de forma manual. Una vez consolidadas las fases analíticas se automatizó mediante Apache NiFi, que no solo deposita los ficheros en la capa raw de HDFS, sino que orquesta la ejecución completa del pipeline con control de errores entre etapas, tal como se detalla en el Capítulo 5.

La segunda fase es la de limpieza y enriquecimiento. En ella, el Job 1 de Spark parte del subconjunto de trabajo ya delimitado, licitaciones de la AGE con resultado conocido para los tres tipos de contrato y aplica la limpieza necesaria: depuración temporal, exclusión de

los presupuestos no positivos y normalización de tipos, además de la winsorización al percentil 99 del importe de adjudicación por año y tipo de contrato para limitar el efecto de valores extremos. A continuación, el Job 2 enriquece el dataset mediante un join en dos pasos con el maestro de órganos de contratación de PLACSP. Para ello se utiliza como clave principal el identificador de plataforma y, cuando este no está disponible, el NIF como clave alternativa. El resultado es un dataset curated y enriquecido en formato Parquet, particionado por año.

La tercera fase corresponde al feature engineering. En esta etapa, el Job 3 aplica primero varias correcciones estructurales, como la limpieza del identificador del órgano, la deduplicación de los snapshots acumulativos y la winsorización del presupuesto base. Después calcula los cinco indicadores de riesgo a nivel de licitación-lote y los agrega a nivel de órgano-año mediante tasas, índices de concentración y promedios condicionales. Los percentiles utilizados para definir algunos umbrales operativos, en particular el percentil 10 del plazo de presentación por tipo de contrato en RF5, se calculan sobre el dataset deduplicado completo y se interpretan como estadísticos descriptivos de la población observada, no como parámetros aprendidos por un modelo predictivo. Dado que el proyecto no plantea un enfoque supervisado, no resulta necesario dividir los datos en conjuntos de entrenamiento, validación y test. En su lugar, la fiabilidad estadística se refuerza estableciendo un umbral mínimo de 30 contratos por órgano-año para entrar en el ranking principal, criterio que se justifica por el error estándar admisible en una proporción de Bernoulli y por su coherencia con prácticas utilizadas en sistemas de monitorización similares.

La cuarta y última fase es la de construcción del índice y evaluación de resultados. En ella, el Job 4 normaliza los indicadores mediante min-max scaling sobre la población de referencia y los agrega aplicando los tres esquemas de ponderación definidos en el OE3: pesos iguales como línea base, pesos centrados en los indicadores de competencia, siguiendo la jerarquía empírica propuesta por Fazekas, y pesos proporcionales a la varianza observada en cada indicador normalizado. La estabilidad de los rankings se evalúa mediante correlaciones de Spearman entre años consecutivos para cada esquema, mientras que la

robustez frente al método de agregación se analiza comparando el grado de concordancia entre los rankings generados por los tres enfoques. Finalmente, el Job 5 indexa los resultados en Elasticsearch y estos se visualizan en un dashboard de Kibana que permite explorar el ranking de órganos, la distribución del índice, su evolución temporal y el desglose de los indicadores que lo componen para cada órgano-año.

Por último, la metodología incorpora una decisión de naturaleza infraestructural: la elección del motor de procesamiento. La elección de Spark requiere una justificación explícita, porque el volumen analizado, en torno a 124.000 licitaciones, podría procesarse con mayor rapidez y menor sobrecarga mediante herramientas de nodo único como DuckDB o Polars. La decisión no responde, por tanto, a la búsqueda de la opción más eficiente para el tamaño actual de los datos, sino a su coherencia con la arquitectura distribuida del sistema. Los datos residen en HDFS y la computación se planifica mediante YARN, dos entornos con los que Spark se integra de forma nativa, y además el mismo código puede escalar sin reescritura si el sistema se amplía al conjunto completo de PLACSP o a otros niveles de la administración. En consecuencia, se ha priorizado la integración con el clúster y la escalabilidad futura sobre la rapidez en el escenario presente. No se trata de una contradicción, sino de un compromiso deliberado: la herramienta se elige por la arquitectura en la que debe operar, no por el volumen inicial de datos.

Capítulo 5. SISTEMA/MODELO DESARROLLADO

5.1 VISIÓN GENERAL DE LA ARQUITECTURA

La arquitectura del sistema se organiza en torno a dos responsabilidades claramente diferenciadas, como muestra la ilustración del diseño. La primera es la ingesta, que toma los ficheros ya convertidos a formato tabular y los deposita en el almacenamiento distribuido. La segunda es la orquestación del procesamiento, que encadena la verificación de integridad y los distintos jobs de Spark hasta generar el índice de riesgo y publicarlo en el motor de búsqueda.

Ambas responsabilidades recaen sobre Apache NiFi, pero se implementan como flujos independientes. La ingesta se activa cuando llegan nuevos ficheros, mientras que el procesamiento se lanza de forma controlada una vez que los datos están disponibles en HDFS.

El eje de todo el sistema es la organización por capas en HDFS. Los datos entran en la capa raw, se filtran y enriquecen en la capa curated, y se transforman en variables analíticas e índice de riesgo en la capa analytics. A partir de esta última capa, los resultados se indexan en Elasticsearch y se ponen a disposición del usuario final mediante un dashboard de Kibana.

Esta separación por capas no responde solo a una cuestión organizativa. También permite reejecutar cualquier fase del pipeline a partir del resultado persistido por la fase anterior, sin necesidad de repetir el proceso completo.

5.2 INFRAESTRUCTURA DE LA EJECUCIÓN

El sistema se ejecuta sobre el clúster Hadoop del programa de máster, administrado con Cloudera. El nodo Edge concentra el entorno de desarrollo y aloja los contenedores Docker

de NiFi y Kibana. Por su parte, el almacenamiento en HDFS opera en alta disponibilidad sobre dos nodos maestros, mientras que la planificación de los trabajos distribuidos se delega en YARN, que asigna los recursos del clúster a cada job de Spark.

Una decisión de infraestructura especialmente relevante afecta a la versión de Spark utilizada. El clúster incorpora una distribución antigua, incompatible con las versiones de Python necesarias para el proyecto. Por este motivo, el sistema utiliza una instalación de Spark 3.5.6 disponible en un directorio compartido por NFS y accesible desde todos los nodos.

Esta elección obliga a configurar de forma explícita el entorno de ejecución de cada job. En concreto, es necesario definir el intérprete de Python del driver y de los executors, la ruta de instalación de Spark y la ubicación de la configuración de Hadoop, de forma que los trabajos se lancen contra la instalación correcta y se ejecuten en modo YARN sobre el clúster, y no en local.

Los contenedores de NiFi y Kibana se ejecutan en el nodo Edge. En el caso de NiFi, el contenedor se configura con la identidad del usuario propietario del espacio HDFS del proyecto, para disponer de permisos de escritura sobre las rutas de destino. Elasticsearch, por su parte, reside en un nodo independiente del clúster y se accede mediante autenticación básica.

5.3 INGESTA DE DATOS CON NIFI

El flujo de ingesta resuelve el primer tramo del pipeline: trasladar los ficheros tabulares a la capa raw de HDFS. Está formado por dos procesadores. El primero monitoriza de forma continua un directorio del contenedor y captura los ficheros que se depositan en él, filtrando por extensión para procesar únicamente los que resultan relevantes. El segundo escribe cada fichero en HDFS.

El elemento más característico de este flujo es el enrutado dinámico del destino. En lugar de utilizar una única ruta fija, el procesador de escritura evalúa el nombre del fichero de entrada

y decide su ubicación. Los ficheros de licitaciones se dirigen al subdirectorio correspondiente dentro de la capa raw, mientras que los ficheros maestros de órganos de contratación se almacenan en su ruta específica.

La estrategia de resolución de conflictos se configura para reemplazar los ficheros existentes. Esto hace que la ingesta pueda reejecutarse sobre un mismo fichero sin generar duplicados. De este modo, la incorporación de los datos de un nuevo año se reduce a depositar el fichero correspondiente en el directorio monitorizado.

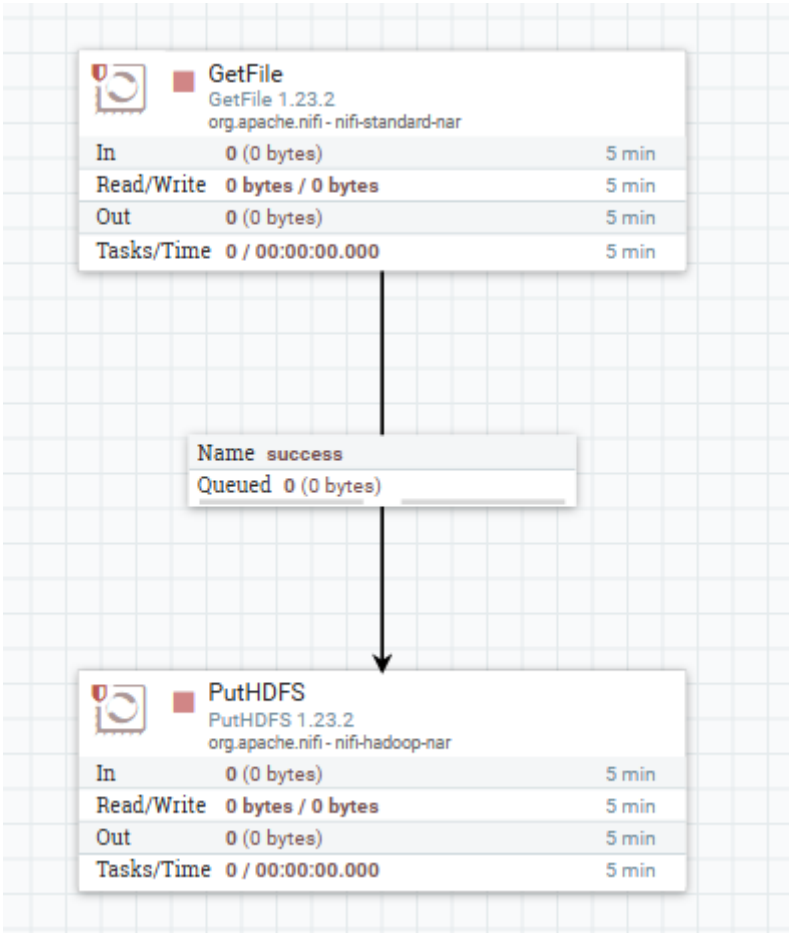


Ilustración 1: Flujo de ingesta en NiFi: captura de ficheros y escritura en HDFS con enrutado dinámico.

5.4 ORQUESTACIÓN DEL PIPELINE Y CONTROL DE ERRORES

El segundo flujo de NiFi se encarga de orquestar el procesamiento completo. Su ejecución se inicia manualmente mediante un disparador que genera un único evento y, a partir de ahí, encadena en secuencia seis etapas: la verificación de integridad y los cinco jobs de Spark. Esta separación respecto al flujo de ingesta es intencionada, ya que el procesamiento solo debe ejecutarse cuando todos los datos necesarios están disponibles y han sido validados, y no cada vez que se incorpora un fichero.

Cada etapa se implementa mediante un procesador que lanza un comando externo. Aquí aparece una de las principales restricciones del diseño: el contenedor de NiFi no dispone del entorno de Spark y, por tanto, no puede ejecutar los trabajos directamente. Para resolverlo, se establece una conexión por SSH desde el contenedor hasta el nodo Edge, donde sí está disponible el entorno completo. Cada etapa invoca así un script envoltorio alojado en ese nodo, que fija las variables de entorno necesarias y lanza el job correspondiente con spark-submit en modo YARN. De este modo, NiFi actúa como capa de orquestación, mientras que la ejecución real tiene lugar en el clúster.

El control de errores se basa en un mecanismo de gating apoyado en el código de salida de cada etapa. Cada procesador solo da paso a la siguiente fase si el comando termina correctamente. Si una etapa devuelve un error, el flujo se detiene en ese punto y no se ejecutan las posteriores.

La consecuencia práctica es que el pipeline no genera resultados parciales ni inconsistentes. Si la verificación detecta un problema en los datos de entrada, o si uno de los jobs de Spark falla, el resto de las fases no llega a ejecutarse.

Esta estrategia de parada temprana se complementa con una vía de depuración manual. Dado que, antes de incorporarse al pipeline, el procesamiento se desarrolló y probó en notebooks, cuando un script falla en un punto intermedio y el error no puede localizarse con suficiente precisión en la ejecución orquestada, es posible volver a su notebook equivalente y ejecutarlo de forma secuencial, celda a celda, en Jupyter. Esto permite aislar el paso problemático con

más detalle y facilita la identificación de la causa del fallo antes de reintegrar la corrección en el script automatizado.

La primera etapa, la verificación de integridad, actúa precisamente como barrera de entrada. Antes de iniciar el procesamiento, comprueba que los ficheros depositados en la capa raw se ajustan a lo esperado en número de registros y estructura, escribe un informe del resultado en HDFS y devuelve un código de éxito o de error que condiciona el arranque del resto del pipeline. Solo cuando esta comprobación finaliza correctamente comienza la cadena de jobs de Spark.

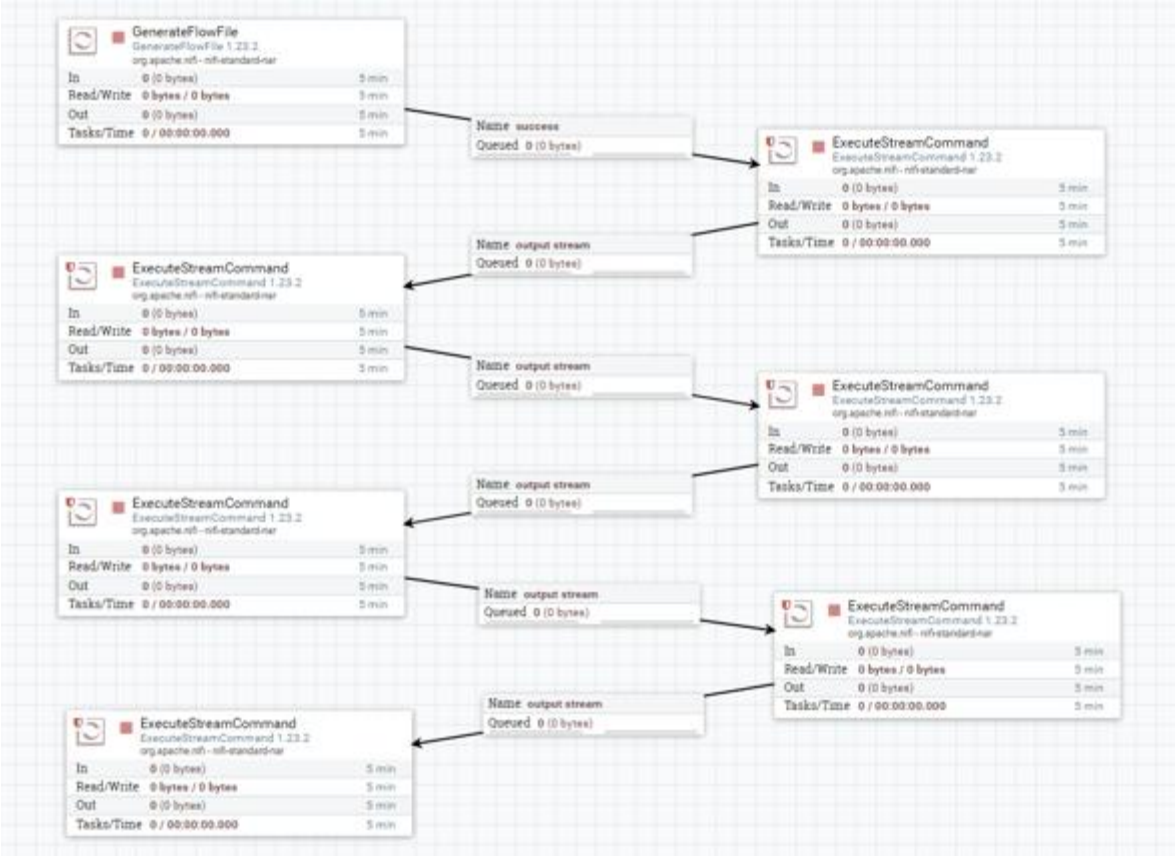


Ilustración 2: Flujo de orquestación del pipeline en NiFi: verificación y Jobs 1-5 encadenados con gating por código de salida.

5.5 JOBS DE PROCESAMIENTO EN SPARK

El núcleo de transformación del sistema se organiza en cinco jobs de Spark. Cada uno tiene una responsabilidad concreta y genera una salida persistida en HDFS. Esta división facilita la depuración y permite reejecutar una fase concreta sin necesidad de repetir las anteriores.

Job	**Entrada**	**Salida**	**Función**
Job 1	Ficheros raw	curated/licitaciones_age (123.966 reg.)	Filtrado al subconjunto AGE y limpieza
Job 2	Salida de Job 1 + maestro de órganos	curated/licitaciones_age_enriched	Enriquecimiento con los datos de órgano
Job 3	Salida de Job 2	analytics/features (104.909 contratos, 4.077 órgano-año)	Deduplicación, red flags y agregaciones
Job 4	Agregados de Job 3	analytics/cri_organo_anyo (4.077 reg.)	Construcción del Composite Risk Index
Job 5	Salida de Job 4	Índice en Elasticsearch (4.077 docs.)	Indexación de resultados

Tabla 1: Resumen de los cinco jobs de Spark del pipeline: entrada, salida y función.

El Job 1 aplica los filtros que delimitan el ámbito de trabajo, como el tipo de administración, el estado de la licitación, el tipo de contrato y la existencia de presupuesto positivo. Además, normaliza los tipos de datos y genera un conjunto curated particionado por año.

El Job 2 enriquece ese conjunto mediante el cruce con el maestro de órganos de contratación. De este modo, incorpora la denominación y el ministerio de adscripción de cada órgano.

El Job 3 concentra la parte analítica más importante del pipeline. En esta fase se corrigen los duplicados estructurales de los snapshots anuales, se calculan los cinco indicadores de riesgo a nivel de licitación y se generan las agregaciones a nivel de órgano-año.

El Job 4 toma esos indicadores agregados, los normaliza y construye el índice de riesgo según los tres esquemas de ponderación definidos en el proyecto. Por último, el Job 5 vuelca la tabla resultante en Elasticsearch.

La lógica interna de la deduplicación, el cálculo de cada indicador y la construcción del índice se desarrolla con más detalle en el Capítulo 6.

Como cierre de esta sección, conviene señalar que dos pasos auxiliares quedan deliberadamente fuera de la cadena orquestada. El primero es la preparación de los catálogos de órganos de contratación, que transforma el fichero maestro original en formato de hoja de cálculo en los dos catálogos utilizados después en la fase de enriquecimiento. El segundo es el análisis exploratorio de los datos curated, que se realiza en modo de solo lectura.

Ambos pasos se implementan en pandas sobre un único nodo y no se integran en NiFi por dos motivos. Por un lado, los nodos de cómputo del clúster no disponen del entorno necesario para ejecutar este tipo de procesos. Por otro, se trata de tareas que solo se repiten de forma puntual, ya sea cuando cambia el catálogo de órganos o cuando se documenta la calidad del dataset, y no en cada ejecución completa del pipeline.

5.6 INDEXACIÓN Y VISUALIZACIÓN

El Job 5 cierra el pipeline indexando en Elasticsearch la tabla de riesgo a nivel de órgano-año. La indexación se realiza con un mapeo explícito de los campos, de forma que Kibana interprete correctamente cada tipo de dato. Los identificadores y categorías se tratan como texto exacto, mientras que los indicadores y puntuaciones se almacenan como valores numéricos.

El resultado es un índice con 4.077 documentos, uno por cada combinación de órgano y año. Cada registro queda además clasificado según la calidad de su estimación.

Sobre este índice se construye el dashboard de Kibana, que actúa como interfaz del sistema para el usuario final. El panel reúne varias visualizaciones, entre ellas el ranking de órganos por nivel de riesgo, la distribución del índice, su evolución temporal, el riesgo agregado por ministerio y el desglose de los indicadores que lo componen.

El dashboard incorpora también tres filtros globales: año, ministerio y calidad de la estimación. Este último resulta especialmente útil, ya que permite limitar el análisis a los órganos cuyo tamaño muestral hace más fiable la estimación y separarlos de aquellos con muy pocos contratos.

El objetivo de esta capa es hacer accesibles los resultados a perfiles no técnicos, como analistas de integridad, y facilitar la priorización de revisiones sobre los órganos con mayor riesgo relativo.

Capítulo 6. COMPOSITE RISK INDEX

Este capítulo desarrolla la lógica analítica del sistema, que en el Capítulo 5 se dejó en segundo plano para centrar la atención en la arquitectura. Su objetivo es explicar cómo se pasa de los datos curated al Composite Risk Index siguiendo la secuencia de transformación implementada en los jobs de Spark.

A lo largo del capítulo se describen las decisiones metodológicas que más influyen en la validez de los resultados. Entre ellas se encuentran el tratamiento de los valores extremos, la deduplicación de los snapshots acumulativos, la definición operativa de cada indicador, la agregación a nivel de órgano y la construcción del índice bajo distintos esquemas de ponderación.

Antes de describir la implementación, conviene justificar tres decisiones de diseño que condicionan el resto del trabajo: el enfoque metodológico adoptado, la unidad de análisis elegida y el tratamiento de la fiabilidad estadística.

La primera decisión se refiere al tipo de enfoque. Como se argumentó en la sección 3.3.1, la ausencia de un conjunto de contratos etiquetados como irregulares descarta una aproximación de aprendizaje supervisado, que requiere una variable objetivo de la que aquí no se dispone. También se valoraron otras alternativas no supervisadas, pero finalmente se descartaron. La detección de anomalías permite identificar observaciones estadísticamente atípicas, aunque un valor atípico no equivale necesariamente a un mayor riesgo de irregularidad: un órgano grande o muy especializado puede comportarse como un outlier por razones legítimas, sin que ello diga nada sobre su integridad. Las técnicas de pseudoetiquetado se rechazaron igualmente por su carácter circular, ya que convertirían en verdad aproximada los mismos supuestos que el sistema pretende medir. La opción adoptada es, por tanto, la construcción de un índice compuesto a partir de indicadores definidos conceptualmente desde la literatura, en la tradición de Fazekas, Tóth y King, porque es el único enfoque que combina tres propiedades necesarias para una herramienta de

priorización: se apoya en una definición interpretable del riesgo, no depende de etiquetas y produce un orden de los órganos, que es precisamente el resultado que se busca.

La segunda decisión afecta a la unidad de análisis. El sistema calcula las señales a nivel de licitación-lote, pero la unidad principal del análisis y del ranking es el órgano-año. Esta elección responde a tres razones. En primer lugar, es la unidad accionable para un organismo de control, que prioriza revisiones sobre órganos y no sobre licitaciones aisladas. En segundo lugar, permite agregar el ruido propio de las señales binarias a nivel de contrato en tasas y promedios más estables, de modo que el riesgo se mida sobre el comportamiento sostenido de un órgano y no sobre un caso individual. En tercer lugar, coincide con la lógica de agregación a nivel de comprador empleada en la literatura de referencia sobre indicadores objetivos de riesgo en contratación pública. La licitación-lote conserva, un doble papel: constituye el nivel intermedio sobre el que se calculan las señales y sigue siendo la capa de detalle a la que se desciende mediante el drill-down del dashboard.

La tercera decisión se refiere a la fiabilidad estadística. Como el sistema produce un índice descriptivo y no un modelo predictivo, los umbrales y percentiles que definen las señales, en particular el percentil del plazo de presentación, se calculan sobre la población observada completa y se interpretan como estadísticos descriptivos, no como parámetros aprendidos. Por esta razón, y según se justificó en la sección 4.4, no procede dividir los datos en conjuntos de entrenamiento, validación y test, y la cuestión del data leakage propia de los modelos supervisados no se plantea. La fiabilidad se garantiza por otra vía: el umbral mínimo de treinta contratos por órgano-año que delimita la población de referencia del ranking.

6.1 FILTRADO Y CURACIÓN DEL SUBCONJUNTO (JOB 1)

El primer job parte del CSV de licitaciones ya restringido al subconjunto de la Administración General del Estado definido en el Capítulo 4. Sobre él aplica un filtro temporal al periodo 2018-2023 que, además de acotar la ventana de análisis, elimina un pequeño número de registros corruptos generados durante el parsing de campos de texto multilínea del CSV. Como resultado, el conjunto se reduce de 124.761 a 123.966 filas. A

continuación, se descartan los registros con presupuesto base nulo o no positivo, ya que esta variable actúa como denominador en el cálculo del precio relativo de adjudicación.

Sobre este subconjunto se aplica una winsorización del importe de adjudicación en el percentil 99, estratificada por año y tipo de contrato. Esta operación consiste en truncar los valores que superan ese umbral y sustituirlos por el propio valor del percentil, en lugar de eliminarlos. La decisión responde a que el importe de adjudicación presenta una cola derecha muy marcada: unos pocos contratos de cuantía muy elevada, o incluso posibles errores de carga, pueden distorsionar de forma desproporcionada las medias y los ratios calculados después, especialmente el precio relativo de adjudicación.

La estratificación por año y tipo de contrato resulta necesaria porque las escalas de importe son diferentes de la forma notable entre categorías contractuales. Aplicar un único umbral global mezclaría esas diferencias estructurales con valores atípicos. Al acotar los extremos dentro de cada estrato homogéneo, se reduce su influencia sin eliminar observaciones y se preserva mejor la forma general de la distribución.

Por último, el conjunto curated se persiste en formato Parquet y particionado por año.

6.2 ENRIQUECIMIENTO CON EL MAESTRO DE ÓRGANOS (JOB 2)

El cruce se apoya en dos catálogos derivados del maestro de órganos de PLACSP mediante un script de preparación previo, ejecutado fuera de la cadena orquestada. El primero es el catálogo completo de la AGE y el segundo un catálogo de respaldo restringido a los NIF que aparecen una sola vez dentro de la AGE. Esta restricción en el catálogo de respaldo evita asignaciones erróneas en los casos en que un mismo NIF agrupa a varios órganos distintos.

El segundo job incorpora la identidad de cada órgano de contratación mediante un cruce con el maestro de órganos publicado por PLACSP. Este proceso se realiza en dos pasos, debido a una limitación de los datos de origen. El identificador unívoco de órgano no está disponible en los ficheros de 2018 y 2019, y su cobertura es solo parcial en 2020 y 2021.

Por ello, el primer cruce se realiza utilizando el identificador de plataforma cuando está disponible. Después, los registros que no han encontrado coincidencia se resuelven mediante un segundo cruce, usando el NIF del órgano como clave alternativa. Esta estrategia permite conservar los años más antiguos sin excluirlos del análisis.

El enriquecimiento añade la denominación del órgano y su ministerio de adscripción. Estos campos se utilizan más adelante en las visualizaciones y en los filtros del dashboard.

El uso del NIF como clave secundaria introduce, no obstante, una limitación de granularidad ya señalada en la sección 4.3.2. En los años más antiguos, puede agrupar bajo un mismo identificador a varios órganos dependientes de un mismo organismo padre. Su impacto está acotado por dos motivos. Por un lado, solo afecta al periodo 2018-2021, ya que a partir de 2022 el identificador de plataforma está disponible y la granularidad es completa. Por otro, la agrupación se produce únicamente entre órganos que comparten NIF, es decir, dependientes de un mismo organismo, de modo que reduce el nivel de detalle pero no llega a mezclar entidades administrativas no relacionadas. Estos casos coinciden además, en buena parte, con los órganos cuyo nombre no se resuelve, que no son interpretables sin una identificación previa.

6.3 CORRECCIONES ESTRUCTURALES Y DEDUPLICACIÓN

Antes de calcular los indicadores, el tercer job aplica tres correcciones estructurales sobre el conjunto enriquecido.

La primera es la limpieza del identificador de órgano. Este identificador se construye en el Job 1 a partir del identificador de plataforma del órgano, o del NIF cuando aquel no está disponible. El Job 3 normaliza su formato eliminando un sufijo residual que de otro modo dividiría a un mismo órgano en claves distintas según el año.

La segunda, y más importante para la validez del análisis, es la deduplicación de los snapshots anuales. Como se indicó en la sección 4.3.2, los ficheros anuales de PLACSP son acumulativos, de modo que un contrato adjudicado en un año puede volver a aparecer en el

fichero del año siguiente con información actualizada. La deduplicación se realiza en dos pasos. El primero elimina las filas estrictamente idénticas, reduciendo el conjunto de 123.966 a 123.238 registros. El segundo, para cada par formado por el identificador de la licitación y el número de lote, conserva únicamente el snapshot más reciente, descartando 18.329 registros adicionales y dejando 104.909 contratos únicos. Sin este tratamiento, los indicadores agregados quedarían inflados por duplicados estructurales.

Tras deduplicar, el job descarta los contratos cuyo órgano no ha podido identificarse por ninguna de las dos vías, para no formar un grupo espurio de "órgano desconocido" que contaminaría el ranking. Los contratos que sí tienen identificador, pero cuyo nombre no se resolvió en el enriquecimiento se conservan: constituyen un grupo válido para el cálculo, aunque no interpretable sin trabajo manual adicional.

La tercera corrección consiste en una segunda winsorización al percentil 99, aplicada en este caso sobre el presupuesto base y con la misma lógica de estratificación utilizada en el Job 1. Dado que el presupuesto base actúa como denominador en el cálculo del precio relativo de adjudicación, acotar sus valores extremos permite evitar que presupuestos anómalos distorsionen el indicador.

6.4 DEFINICIÓN Y CÁLCULO DE LOS INDICADORES DE RIESGO (JOB 3)

Sobre el conjunto deduplicado se calculan los cinco indicadores de riesgo. Su fundamento teórico ya se presentó en la sección 3.4, aquí se describe su definición operativa a partir de los campos disponibles en PLACSP. Cuatro de ellos se calculan a nivel de licitación-lote y uno a nivel de órgano-año.

El indicador de single bidding toma valor uno cuando una licitación adjudicada ha recibido un único licitador. Se trata del caso que la literatura identifica como el proxy más sólido de restricción de la competencia.

El indicador de procedimientos no competitivos marca las licitaciones tramitadas mediante figuras que evitan la convocatoria pública abierta, como el negociado sin publicidad, el contrato menor o la adjudicación directa. La categoría derivada de acuerdo marco se trata como procedimiento competitivo y no se computa como señal de riesgo, ya que la competencia se produce en la fase previa de licitación del propio acuerdo. Todos los contratos permanecen en el denominador del indicador.

El precio relativo de adjudicación se define como el cociente entre el importe de adjudicación y el presupuesto base. Un valor próximo a uno refleja la falta de descuento competitivo, mientras que valores más bajos apuntan a una mayor competencia en precio. Los registros con un cociente superior a 1,5 se excluyen del cálculo por considerarse poco creibles, ya que una adjudicación que supera el presupuesto base en más de la mitad sugiere un posible error en los datos.

La concentración de proveedores se calcula directamente a nivel de órgano-año, y no a nivel de licitación, mediante el índice de Herfindahl-Hirschman y la cuota del proveedor principal. Al trabajar sobre la unidad final de análisis, este indicador permite captar la falta de competencia sostenida en el tiempo con menos ruido que los indicadores binarios.

Por último, el plazo de presentación de ofertas marca como anómalas las licitaciones cuyo plazo efectivo, calculado como la diferencia en días entre la fecha de publicación y la fecha límite, se sitúa por debajo del percentil 10 de la distribución observada para su tipo de contrato. Ese percentil se calcula sobre la población deduplicada completa y se interpreta como un estadístico descriptivo de la población observada, no como un parámetro derivado de un modelo predictivo.

6.5AGREGACIÓN A NIVEL DE ÓRGANO-AÑO Y CALIDAD DE LA ESTIMACIÓN

Los indicadores calculados a nivel de licitación se agregan después a la unidad de análisis del proyecto, el órgano-año, lo que da lugar a 4.077 grupos. La agregación adopta formas

distintas según el tipo de indicador: tasas para los indicadores binarios, promedios para el precio relativo e índices de concentración para los proveedores. Un aspecto importante es que cada tasa se calcula sobre su propio denominador de registros válidos, que no siempre coincide entre indicadores debido a las diferencias de cobertura. Los registros sin valor computable se excluyen tanto del numerador como del denominador, lo que evita sesgos, aunque hace que algunas tasas se estimen sobre muestras más pequeñas.

Para tratar de forma explícita la diferencia de tamaño entre órganos, cada grupo recibe una etiqueta de calidad de la estimación en función de su número de contratos. La calidad se considera alta a partir de treinta contratos, media entre diez y veintinueve, y baja por debajo de diez. El umbral de treinta delimita la población incluida en el ranking principal y responde al error estándar admisible de una proporción. De los 4.077 grupos generados, 743 alcanzan calidad alta, 720 calidad media y 2.614 calidad baja.

Este mecanismo constituye la traducción operativa del sesgo por tamaño señalado en la sección 4.3.2. Los grupos con pocas observaciones no se eliminan del sistema, pero quedan fuera del ranking interpretable.

La tabla siguiente resume las principales variables del conjunto final a nivel de órgano-año:

Variable	**Descripción**
id_organo, anyo	Clave de la unidad de análisis, formada por órgano y año
nombre_organo, ministerio	Identidad del órgano tras el enriquecimiento
n_contratos	Número de contratos únicos del grupo
sb_rate	Tasa de single bidding
ncb_rate	Tasa de procedimientos no competitivos
avg_relative_price	Precio relativo de adjudicación medio
stp_rate	Tasa de plazos de presentación anormalmente cortos
hhi, share_top1	Concentración de proveedores, medida mediante HHI y cuota del principal
n_proveedores_distintos	Número de proveedores distintos
calidad_estimacion	Etiqueta de fiabilidad según el tamaño muestral
cri_ew, cri_cf, cri_vw	Índice de riesgo bajo los tres esquemas de ponderación

rank_ew, rank_cf, rank_vw	Posición en el ranking para cada esquema

Tabla 2: Principales variables del conjunto de datos a nivel de órgano-año.

6.6CONSTRUCCIÓN DEL COMPOSITE RISK INDEX (JOB 4)

El cuarto job construye el índice compuesto a partir de los cinco indicadores agregados (single bidding, procedimientos no competitivos, precio relativo, concentración medida mediante el HHI y plazo de presentación). El proceso se organiza en dos fases: normalización y ponderación.

La normalización se realiza mediante un escalado min-max calculado sobre la población de referencia, es decir, sobre los órganos con calidad de estimación alta. Tomar como referencia este subconjunto, y no el conjunto completo, evita que los grupos con muy pocos contratos distorsionen la escala con valores extremos debidos al azar.

Una vez normalizados los indicadores, se aplican los tres esquemas de ponderación definidos en el OE3. El primero asigna pesos iguales a los cinco indicadores y actúa como línea base. El segundo, orientado a la competencia, da mayor peso al single bidding y a los procedimientos no competitivos, que son los indicadores con mayor respaldo en la literatura. El tercero asigna los pesos en proporción a la varianza observada, de modo que los indicadores con mayor capacidad de discriminación tienen más influencia en el resultado.

Cada esquema produce un valor de índice y una posición en el ranking. Esto permite comparar los tres resultados de forma directa.

Cuando a un grupo le falta alguna señal por ausencia de datos para calcularla, el peso de esa señal se redistribuye proporcionalmente entre las disponibles, de modo que el índice se calcula sobre el peso efectivamente presente y sigue siendo comparable entre grupos.

La existencia de estos tres esquemas no es un elemento accesorio, sino una parte central de la aportación metodológica del trabajo. En lugar de asumir que un único esquema produce un resultado robusto, el sistema permite comprobarlo de forma explícita.

6.7 ESTRATEGIA DE VALIDACIÓN DE LA ESTABILIDAD

En ausencia de un ground truth de corrupción, la validez del índice se evalúa a partir de su estabilidad. La estrategia adoptada combina dos análisis complementarios.

El primero calcula la correlación de Spearman entre los rankings generados por los tres esquemas en cada año. Con ello se mide hasta qué punto la posición de los órganos depende de la elección de los pesos.

El segundo analiza la coincidencia de los órganos situados en las primeras posiciones del ranking para distintos puntos de corte. En concreto, compara qué proporción de los órganos con mayor riesgo permanece en el grupo de cabeza con independencia del esquema utilizado.

La lógica es sencilla. Si los mismos órganos ocupan las posiciones extremas bajo distintos esquemas de ponderación, el índice está captando una señal estructural y no un efecto del método de agregación.

Capítulo 7. ANÁLISIS DE RESULTADOS

Este capítulo presenta los resultados obtenidos por el sistema para el periodo 2018-2023. La exposición avanza de lo general a lo particular. En primer lugar, se describe la distribución global del riesgo, después, el ranking de órganos, su evolución temporal y su reparto por ministerio, a continuación, se analiza la estabilidad del índice, y por último, los resultados se traducen en tipologías de riesgo y en casos concretos junto con sus posibles pautas de actuación.

Antes de entrar en los resultados, conviene recordar la advertencia metodológica que atraviesa todo el trabajo. El índice no prueba la existencia de irregularidades, sino que identifica patrones estadísticos asociados a un mayor riesgo. En consecuencia, debe interpretarse como una herramienta de priorización de revisiones y no como una clasificación de los órganos según su integridad.

7.1 VISIÓN GENERAL

El pipeline procesa 124.761 registros de entrada que, tras el filtrado y la limpieza, se reducen a 123.966 licitaciones. Después, la deduplicación de los snapshots acumulativos deja un total de 104.909 contratos únicos. La agregación a nivel de órgano-año produce 4.077 grupos, de los cuales 743 alcanzan la calidad de estimación alta, es decir, al menos treinta contratos, y constituyen la población de referencia del ranking. Estos 743 grupos concentran en torno al ochenta por ciento del volumen contratado, de modo que la restricción por calidad no compromete la representatividad del análisis: excluye grupos pequeños y estadísticamente frágiles, no la parte principal de la contratación.

La distribución de los grupos por calidad y año, mostrada en la visualización V6 del dashboard, indica que la mayoría de los órgano-año presentan calidad baja en todos los ejercicios, con un máximo de volumen en 2021. Este predominio de grupos pequeños resulta esperable en una administración compuesta por numerosos órganos con una actividad

contractual reducida. Precisamente por eso, el filtro de calidad resulta necesario, sin él, el ranking quedaría dominado por órganos con muy pocos contratos, cuyos indicadores reflejarían más ruido estadístico que una señal interpretable.

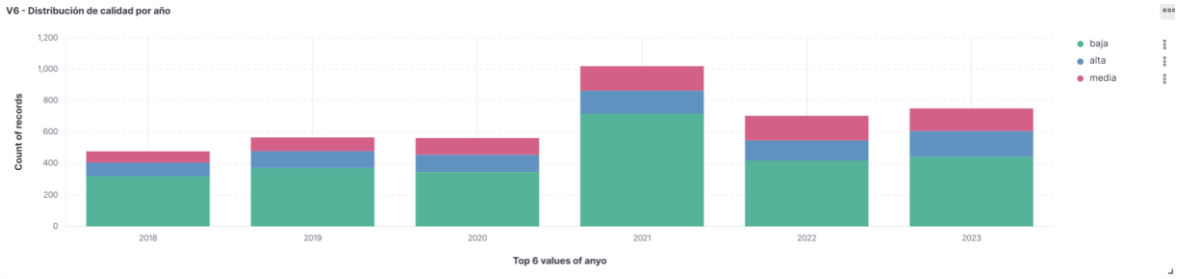


Ilustración 3: V6 - Distribución de calidad por año

7.2DISTRIBUCIÓN DEL ÍNDICE DE RIESGO

Sobre la población de calidad alta, el índice con pesos iguales presenta una media de 0,310 y una mediana de 0,304. La mayor parte de los órganos se concentra entre 0,250, correspondiente al percentil 25, y 0,365, correspondiente al percentil 75. El percentil 95 se sitúa en 0,456 y el valor máximo alcanza 0,753. La distribución, representada en la visualización V2 del dashboard, adopta la forma de una campana ligeramente asimétrica centrada en torno a 0,30, con una cola derecha estrecha y un valor extremo aislado en 0,75.

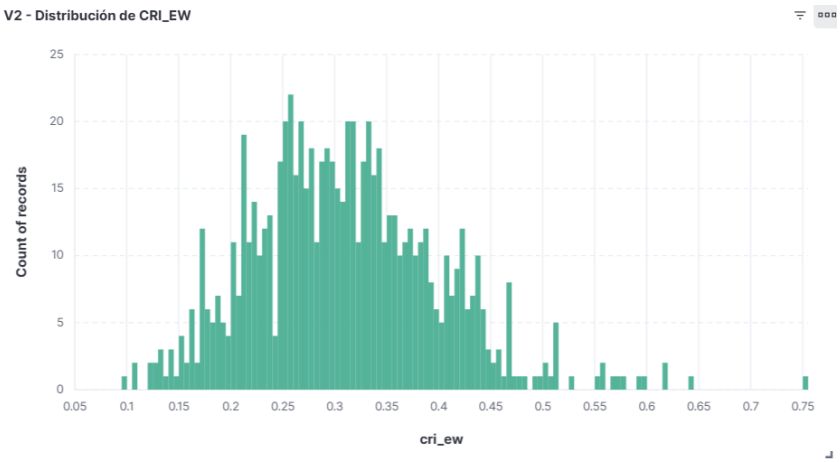


Ilustración 4: V2 - Distribución de CRI_EW

Esta configuración admite una lectura operativa directa. El riesgo no se distribuye de manera uniforme entre los órganos, sino que se concentra en una cola reducida. La mayoría se agrupa en una franja intermedia relativamente homogénea, mientras que solo unas pocas decenas se separan con claridad del conjunto. Para un organismo de control, este resultado es relevante, porque indica que el sistema no apunta a un problema difuso y generalizado, sino a un grupo acotado y manejable de órganos prioritarios.

7.3 RANKING DE ÓRGANOS DE MAYOR RIESGO

La tabla siguiente recoge los doce órgano-año con mayor valor del índice de riesgo bajo el esquema de pesos iguales. Las cinco señales se expresan como tasas entre 0 y 1.

Año	Órgano	Ministerio	n	sb	ncb	precio	hhi	stp	CRI (EW)
2023	(sin nombre)	-	36	0,97	0,00	1,00	0,54	1,00	0,753
2019	(sin nombre)	-	39	0,44	0,54	0,81	0,30	1,00	0,645
2020	INGESA	Sanidad	161	0,91	0,85	0,86	0,05	0,38	0,618
2021	IMSERSO	Derechos Sociales	67	0,95	0,96	1,00	0,10	0,00	0,615
2019	AECID	Asuntos Exteriores	216	0,29	0,87	0,89	0,53	0,17	0,599
2020	(sin nombre)	-	58	0,78	0,78	0,88	0,12	0,33	0,592
2021	DGAM Armamento	Defensa	41	0,80	0,98	0,93	0,11	0,00	0,578
2022	DGAM Armamento	Defensa	65	0,88	0,91	0,88	0,12	0,00	0,572

2019	(sin nombre)	-	58	0,88	0,90	0,91	0,09	0,00	0,568
2018	(sin nombre)	-	38	0,47	0,58	0,82	0,35	0,40	0,557
2023	INGESA Melilla	Sanidad	87	0,67	0,59	0,88	0,24	0,28	0,556
2023	DGAM Armamento	Defensa	69	0,96	0,61	0,83	0,07	0,13	0,526

Tabla 3: Órganos-año con mayor índice de riesgo bajo el esquema de pesos iguales.

anyo	cri_ew	ministerio	nombre_organo	n_contrat...	sb_rate	ncb_rate	hhi
2023	0.753	-	-	36	0.972	0	0.535
2019	0.645	-	-	39	0.436	0.538	0.299
2020	0.618	Ministerio de Sanidad	Dirección del Instituto Nacional de Gestión Sanitaria (INGESA)	161	0.912	0.851	0.046
2021	0.615	Ministerio de Derechos Sociales, Consumo y Agenda 2030	Dirección General del Instituto de Mayores y Servicios Sociales	67	0.952	0.955	0.101
2019	0.599	Ministerio de Asuntos Exteriores, Unión Europea y ...	Dirección de la Agencia Española de Cooperación Internacional para el ...	216	0.294	0.866	0.528
2020	0.592	-	-	58	0.776	0.776	0.122
2021	0.578	Ministerio de Defensa	Subdirección General de Adquisiciones de Armamento y Material DGAM	41	0.805	0.976	0.108
2022	0.572	Ministerio de Defensa	Subdirección General de Adquisiciones de Armamento y Material DGAM	65	0.879	0.908	0.115
2019	0.568	-	-	58	0.879	0.897	0.088
2018	0.557	-	-	38	0.472	0.579	0.345
2023	0.556	Ministerio de Sanidad	Gerencia de Atención Sanitaria de Melilla - Instituto Nacional de Gestión ...	87	0.674	0.586	0.237
2018	0.551	-	-	40	0.351	0.325	0.623
2023	0.526	Ministerio de Defensa	Subdirección General de Adquisiciones de Armamento y Material DGAM	69	0.956	0.609	0.066

Ilustración 5: V1 - Top ranking órganos por CRI_EW

La primera observación remite a una limitación del sistema más que a un hallazgo sustantivo. Varios de los órganos situados en las primeras posiciones, incluido el primero, carecen de nombre. Se trata de órganos de los años 2018-2021 identificados por NIF, debido a la ausencia o cobertura incompleta del identificador de plataforma en esos ejercicios, cuyo nombre no pudo resolverse en el cruce con el maestro. El sistema los conserva porque sus indicadores son válidos para el cálculo, pero su interpretación exige un trabajo manual previo de identificación mediante el cruce del NIF con otras fuentes administrativas.

La segunda observación sí es directamente interpretable y constituye el resultado más sólido de la tabla. Entre los órganos con nombre, varios reaparecen en posiciones altas a lo largo de distintos años. La Subdirección General de Adquisiciones de Armamento y Material del Ministerio de Defensa figura en el grupo de cabeza en 2021, 2022 y 2023. El Instituto

Nacional de Gestión Sanitaria aparece en 2020 y 2023, y además su patrón de riesgo se prolonga en una de sus unidades territoriales. La Agencia Española de Cooperación Internacional para el Desarrollo vuelve a situarse entre los casos destacados en más de un ejercicio.

Esta persistencia temporal es relevante desde el punto de vista analítico. Un valor elevado en un solo año puede responder a circunstancias coyunturales o al azar propio de una muestra concreta, pero su repetición sostenida sugiere una característica más estable del órgano. Por ello, la prioridad de revisión no debería definirse solo por la posición puntual en el ranking, sino también por la continuidad del patrón de riesgo en el tiempo.

7.4 EVOLUCIÓN TEMPORAL

La evolución de las señales medias sobre la población de calidad alta se resume en la tabla siguiente:

Año	**sb**	**ncb**	**stp**	**precio**	**hhi**	**CRI (EW)**
2018	0,415	0,173	0,179	0,665	0,126	0,322
2019	0,403	0,116	0,141	0,659	0,131	0,300
2020	0,403	0,139	0,133	0,668	0,124	0,303
2021	0,457	0,115	0,161	0,699	0,131	0,323
2022	0,424	0,100	0,151	0,699	0,132	0,311
2023	0,410	0,087	0,145	0,682	0,132	0,301

Tabla 4: Evolución temporal de las señales de riesgo y del índice sobre la población de calidad alta (2018-2023).

El patrón más claro es el de los procedimientos no competitivos, que parten de su valor más alto en 2018 (0,173) y descienden de forma sostenida hasta 0,087 en 2023. Esta evolución es coherente con la entrada en vigor de la LCSP: 2018 constituye todavía un año de transición, en el que conviven expedientes tramitados bajo el marco anterior, mientras que

en los ejercicios posteriores el recurso a procedimientos sin concurrencia abierta pierde peso. El índice medio, en cambio, se mantiene notablemente estable a lo largo de todo el periodo, con una oscilación muy contenida entre 0,300 y 0,323.

Conviene subrayar, además, un matiz sobre 2020. En el agregado, el año de la pandemia no se traduce en un pico del riesgo medio. El efecto asociado a la COVID-19 no aparece como una elevación generalizada del índice, sino concentrado en órganos concretos, especialmente del ámbito sanitario, como se verá en el análisis de casos. Esto refuerza la utilidad de no quedarse en la media: el aumento del riesgo vinculado a la emergencia existe, pero adopta una forma localizada y se diluye cuando se promedia sobre el conjunto de órganos.

La visualización V3 del dashboard, que representa el índice medio anual bajo los tres esquemas de ponderación, muestra además trayectorias muy próximas entre sí. Esa cercanía constituye una primera evidencia visual de la estabilidad del índice, ya que el nivel medio apenas varía en función de los pesos elegidos. Las cifras absolutas del gráfico pueden diferir ligeramente de las de la tabla, porque la visualización agrega sobre el conjunto indexado mientras que aquí se trabaja solo con la población de calidad alta, pero el patrón temporal de fondo es el mismo.

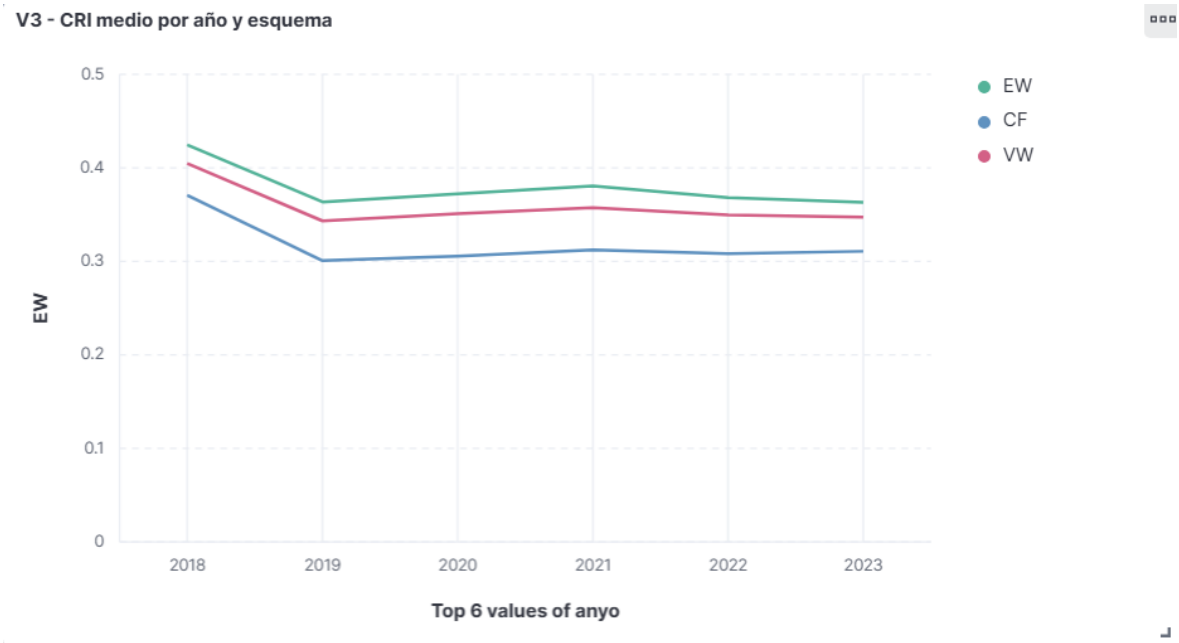


Ilustración 6: V3 - CRI medio por año y esquema

7.5RIESGO POR MINISTERIO

La agregación por ministerio sobre la población de calidad alta sitúa a Asuntos Exteriores como el departamento con mayor índice medio, con un valor de 0,423 sobre 10 grupos. Le siguen Sanidad, con 0,394 sobre 14 grupos, y Derechos Sociales, con 0,356 sobre 10 grupos. El Ministerio de Defensa, por su parte, aporta con diferencia el mayor volumen a la población de referencia, con 255 grupos, y presenta un índice medio de 0,323, ligeramente superior a la media global.

Un aspecto importante para la interpretación es que un mismo valor medio del índice no tiene por qué responder al mismo patrón de riesgo. Asuntos Exteriores encabeza la clasificación pese a mostrar una tasa media de single bidding relativamente baja, de 0,345, de modo que su posición se explica sobre todo por los procedimientos no competitivos y por la concentración de proveedores. Sanidad, en cambio, combina un índice elevado con una tasa de single bidding claramente más alta, de 0,575. Esta diferencia confirma que el índice agregado debe leerse junto con el perfil de sus componentes y no de forma aislada. La

conveniencia de interpretar un índice compuesto junto con sus componentes individuales está bien establecida en la literatura metodológica sobre composite indicators.

La lectura del riesgo por ministerio exige, además, dos cautelas. La primera es el distinto tamaño de los agregados: ministerios como Igualdad aparecen con un único órgano-año de calidad alta, por lo que su media no es directamente comparable con la de Defensa. La segunda es que la visualización V4 del dashboard se calcula para un año concreto, mientras que las cifras anteriores resumen el conjunto del periodo. Aun así, ambas lecturas coinciden en situar a Sanidad y Asuntos Exteriores entre los ministerios con mayor nivel de riesgo relativo.

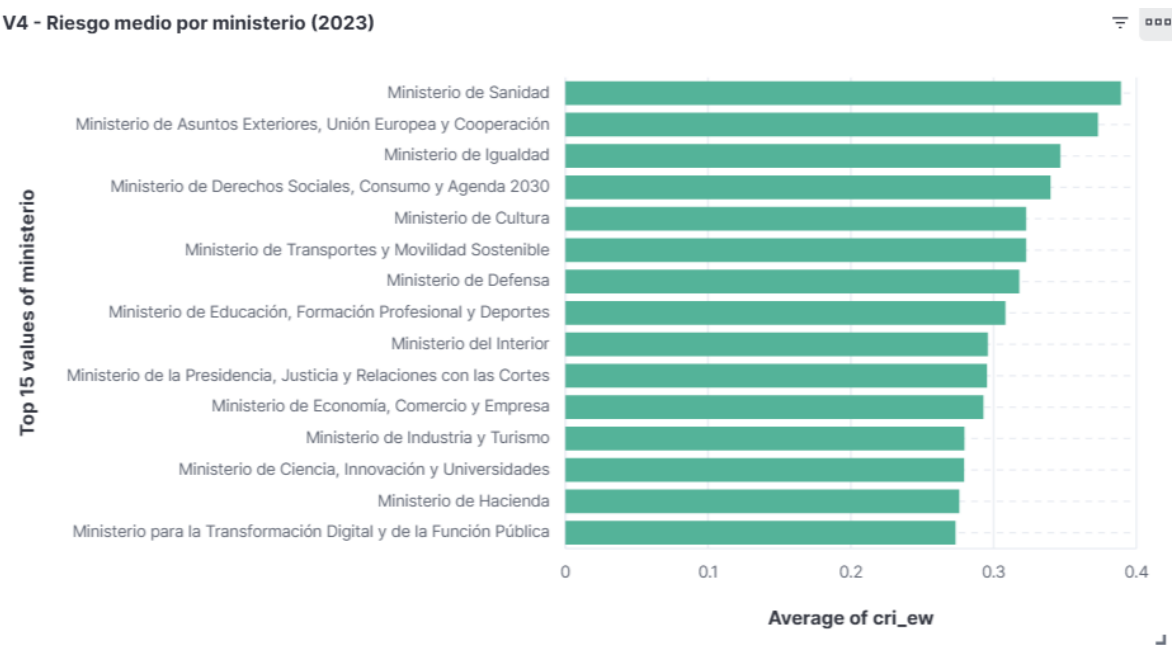


Ilustración 7: V4 - Riesgo medio por ministerio

7.6ESTABILIDAD DEL ÍNDICE

En ausencia de un ground truth de corrupción, la validez del índice se evalúa a partir de su estabilidad. El resultado es claro y constituye el principal argumento de robustez del trabajo: la correlación de Spearman entre los rankings generados por los tres esquemas de ponderación es muy alta, con valores medios de 0,929 entre pesos iguales y competition-

focused, 0,989 entre pesos iguales y variance-weighted, y 0,961 entre los dos esquemas teóricos. Esto significa que el orden de los órganos apenas varía según los pesos utilizados y que las primeras posiciones del ranking tienden a estar ocupadas por los mismos casos con independencia del esquema aplicado.

La estabilidad del grupo de mayor riesgo refuerza esa misma lectura. La coincidencia de órganos en el top entre esquemas alcanza el 68 por ciento en los diez primeros puestos, el 76 por ciento en los veinte primeros y el 84 por ciento en los cincuenta primeros. En consecuencia, los resultados no parecen depender de una elección arbitraria de ponderaciones, sino de una señal subyacente que se mantiene al modificar la forma de agregación.

La estabilidad entre años consecutivos es más moderada, con una correlación de Spearman mediana en torno a 0,63 para los tres esquemas. Esta diferencia no debe interpretarse como una debilidad del índice, sino como una propiedad esperable del fenómeno observado: la composición de los contratos de un órgano cambia de un ejercicio a otro y, además, dos de los años analizados, 2018, por la transición normativa, y 2020, por la emergencia sanitaria, presentan circunstancias singulares. Que el ranking sea muy robusto frente a la elección de pesos y solo moderadamente estable entre años indica, por tanto, que el índice captura una señal estructural consistente en su construcción y, al mismo tiempo, sensible a la variación real de cada ejercicio.

7.7 TIPOLOGÍAS DE RIESGO

Más allá del ranking, los resultados permiten agrupar los órganos según el patrón de señal que activan, una lectura más útil para la acción que una simple lista ordenada. Sobre la población de calidad alta, 122 órgano-año presentan una tasa elevada de plazos cortos de presentación, 85 una tasa elevada de procedimientos no competitivos, 20 una tasa de single bidding superior al ochenta por ciento, 12 una concentración alta de proveedores y solo 2 una adjudicación sistemáticamente próxima al presupuesto base.

El rasgo más relevante es que estas tipologías rara vez se superponen. Solo un órgano-año combina simultáneamente un single bidding alto y una concentración alta, que es precisamente el caso aislado que aparece en la celda roja del mapa de calor del dashboard. La mayoría de los órganos eleva una sola dimensión, no varias a la vez. Esto tiene una consecuencia operativa clara: las elevaciones en un único eje son relativamente frecuentes y a menudo pueden admitir explicaciones legítimas, mientras que la coincidencia de varias señales es poco frecuente y debe situarse en la parte más alta de la cola de revisión.

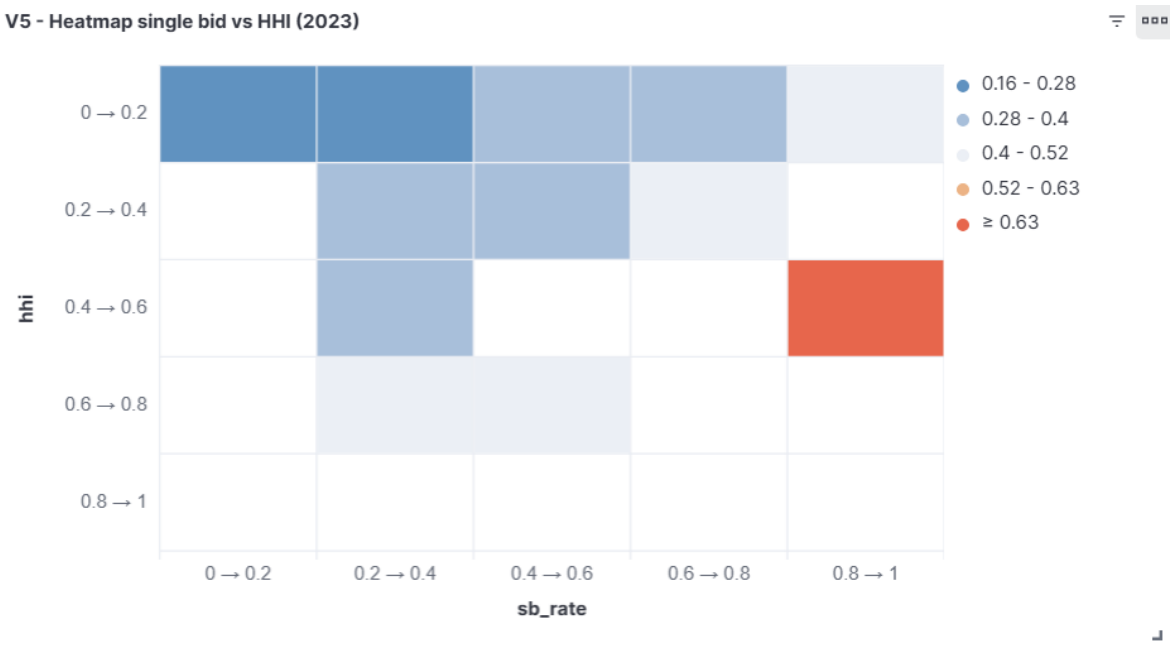


Ilustración 8: V5 - Heatmap single bid vs HHI

A partir de estas tipologías puede proponerse una pauta básica de actuación para el analista de integridad. Cuando un órgano presenta single bidding alto pero baja concentración, la revisión debe orientarse a las condiciones de los pliegos para comprobar si los requisitos técnicos restringen de hecho la concurrencia, dado que los contratos no recaen siempre en los mismos proveedores. Cuando la señal dominante es una tasa elevada de procedimientos no competitivos, lo procedente es verificar que las justificaciones invocadas están documentadas y se ajustan a los supuestos legales. Si el problema principal es la concentración de proveedores, la atención debe centrarse en la relación sostenida con el adjudicatario principal y en la existencia o no de rotación. Si aparecen plazos de presentación

sistemáticamente cortos, conviene comprobar si responden a una urgencia justificada o si, por el contrario, limitan el acceso de nuevos licitadores. Y cuando coinciden varias señales, que es el supuesto menos frecuente, la revisión debe ser prioritaria y abarcar el expediente en su conjunto. En todos los casos, el sistema indica dónde conviene mirar primero, la valoración sobre la existencia o no de una irregularidad corresponde a la revisión humana posterior.

7.8 ANÁLISIS DE CASOS

Cuatro casos del ranking permiten ilustrar cómo un perfil de señales se traduce en una hipótesis de revisión y muestran, al mismo tiempo, que un índice elevado no equivale por sí solo a una conclusión.

El primero es la DGAM de Defensa, que aparece en posiciones altas en 2021, 2022 y 2023, con una tasa de single bidding creciente, 0,80, 0,88 y 0,96, y con procedimientos no competitivos también elevados, pero con una concentración de proveedores baja, medida por un HHI en torno a 0,10. Este perfil, caracterizado por una competencia formal reducida en el procedimiento pero sin dependencia de un único adjudicatario, resulta coherente con la contratación de armamento y material de defensa, un ámbito con pocos suministradores cualificados y con procedimientos negociados amparados en razones de seguridad. Se trata, por tanto, de un caso paradigmático de señal alta con explicación plausible. La actuación recomendable no es presumir irregularidad, sino comprobar que la justificación de los procedimientos no competitivos está debidamente documentada y resulta proporcionada.

El segundo caso es INGESA en 2020, con una tasa de single bidding de 0,91 y una tasa de procedimientos no competitivos de 0,85 sobre 161 contratos. El año y el ámbito remiten de forma inmediata a la contratación sanitaria durante la pandemia. En este contexto, unos valores anómalamente altos pueden reflejar adaptaciones legítimas a un marco de emergencia más que un patrón irregular en sentido estricto. La revisión adecuada pasa, por ello, por contextualizar los resultados a la luz de la normativa excepcional vigente en ese ejercicio antes de extraer cualquier valoración.

El tercer caso es la AECID en 2019, que presenta un patrón distinto. Aquí la tasa de single bidding es baja, 0,29, pero los procedimientos no competitivos alcanzan 0,87 y la concentración de proveedores es elevada, con un HHI de 0,53, sobre una muestra amplia de 216 contratos. En este caso, el riesgo no parece proceder de la falta de ofertas, sino de la combinación entre adjudicaciones no competitivas y dependencia de un número reducido de adjudicatarios. Es un perfil coherente con determinadas dinámicas de la cooperación al desarrollo. La revisión debería centrarse, por tanto, en la concentración de adjudicaciones y en la naturaleza de la relación con los principales proveedores.

El cuarto caso es el órgano de mayor riesgo de todo el periodo, en 2023, que activa cuatro de las cinco señales: una tasa de single bidding de 0,97, un precio relativo prácticamente igual al presupuesto base, plazos cortos en su valor máximo y una concentración de 0,54. Sin embargo, el sistema no puede identificarlo nominalmente. Este ejemplo resume bien el estado de los resultados: el perfil de señales lo convierte en el candidato más claro a revisión, pero la limitación de identificabilidad impide saber con precisión qué órgano se encuentra detrás. En este caso, la primera actuación necesaria no es sustantiva, sino previa: resolver su identidad.

7.9 VALIDACIÓN CUALITATIVA CON FUENTES EXTERNAS

Como se anticipó en la sección 4.3.2, la ausencia de un ground truth de corrupción impide validar el índice frente a una verdad conocida. Por ello, la estrategia adoptada combina criterios internos de robustez, desarrollados en la sección 7.6, con una revisión cualitativa de los casos extremos contrastada con fuentes públicas de control. Esta sección recoge ese contraste para los tres órganos con nombre que aparecen de forma recurrente en las posiciones de cabeza del ranking. Conviene subrayar, además, una condición que refuerza el valor de este ejercicio: el sistema construyó el ranking exclusivamente a partir de los datos de PLACSP, sin incorporar ninguno de los informes que se citan a continuación.

INGESA, incluida su Gerencia de Melilla (2020, 2021, 2023).

El Tribunal de Cuentas ha documentado de forma reiterada contratación al margen de procedimiento: compra directa mayoritaria en las adquisiciones de Ceuta y Melilla en 2016 [23], incidencias en la contratación de emergencia de 2020 [24] y un porcentaje muy elevado de expedientes "sin procedimiento" en 2022-2023 [25]. La correspondencia con el índice es directa, ya que las señales más elevadas del caso son precisamente el single bidding y los procedimientos no competitivos.

Subdirección General de Adquisiciones de Armamento y Material, Ministerio de Defensa (2021, 2022, 2023).

El Tribunal de Cuentas detectó deficiencias de justificación y garantías insuficientes en contratos de Defensa del ejercicio 2022 [26], y la CNMC sancionó en 2023 dos cárteles de reparto de licitaciones de material militar [27]. La relación con el índice es de coherencia del sector: el perfil de baja competencia resulta consistente con un ámbito en el que se han documentado deficiencias procedimentales y conductas colusorias, aunque las fuentes no individualicen expresamente al órgano y la sanción de la CNMC recaiga sobre las empresas adjudicatarias.

AECID, Asuntos Exteriores (2019, 2022).

El Tribunal de Cuentas señaló deficiencias de control y concurrencia en las subvenciones de ayuda al desarrollo [28], y distintos análisis de prensa sobre datos de PLACSP describieron la articulación de un programa en contratos independientes con funciones recurrentes [29]. La correspondencia con el índice es parcial pero apreciable, y encaja con un perfil caracterizado por procedimientos no competitivos y concentración, aunque parte de los hallazgos se refieran a subvenciones más que a contratos.

El caso más concluyente es el de INGESA. Un órgano de control independiente ha señalado de forma sostenida, a lo largo de varios ejercicios, justamente el tipo de comportamiento que el índice detecta: contratación al margen de procedimientos competitivos. Además, lo ha hecho sobre la misma unidad, incluida su gerencia de Melilla, que el sistema sitúa de forma reiterada entre los casos de mayor riesgo.

En el caso de Defensa, la correspondencia es menos directa, pero sigue siendo relevante. El perfil de baja competencia que refleja el índice resulta coherente con un ámbito en el que se han documentado tanto deficiencias procedimentales como prácticas colusorias entre empresas adjudicatarias. En AECID, la coincidencia es más parcial, pero también apreciable, y el patrón de fragmentación descrito por la prensa conecta de forma natural con la línea de trabajo futuro basada en similitud textual.

Esta validación debe interpretarse, no obstante, con tres cautelas. En primer lugar, se trata de una validación de coherencia, no de causalidad: muestra que el sistema señala órganos sobre los que las instituciones de control ya habían llamado la atención, pero no demuestra que las señales identifiquen por sí mismas irregularidades. En segundo lugar, el índice mide el comportamiento del órgano comprador, mientras que algunos de los hallazgos citados afectan al lado de la oferta, como los cárteles sancionados por la CNMC, o a instrumentos distintos del contrato, como las subvenciones en el caso de AECID. En tercer lugar, el contraste se limita a un número reducido de casos extremos y tiene un carácter ilustrativo, por lo que no valida el ranking en su conjunto.

Con estas reservas, el ejercicio aporta un indicio razonable de validez externa. El sistema, construido únicamente a partir de datos abiertos y sin conocimiento previo de los informes de fiscalización, vuelve a señalar órganos sobre los que el control externo ya había puesto el foco.

Capítulo 8. CONCLUSIONES Y TRABAJOS FUTUROS

Este capítulo recapitula el trabajo realizado, valora el grado de cumplimiento de los objetivos planteados, sintetiza las principales aportaciones del proyecto y, por último, plantea las líneas en las que el sistema puede ampliarse en el futuro.

8.1 CONCLUSIONES

El proyecto ha permitido diseñar e implementar un sistema completo de análisis de riesgo en la contratación pública española que transforma los datos abiertos de PLACSP en un ranking de órganos de la Administración General del Estado ordenados por nivel de riesgo relativo. El resultado se concibe como una herramienta de apoyo a la priorización de auditorías. En este sentido, el objetivo general puede considerarse cumplido: el sistema recorre todo el ciclo, desde la ingesta de los ficheros brutos hasta la exploración interactiva de los resultados, y lo hace de forma reproducible y auditable.

El grado de cumplimiento de los cinco objetivos específicos también es satisfactorio. El OE1, relativo al pipeline de datos, se ha cubierto mediante un flujo distribuido sobre HDFS, Spark y YARN, orquestado de extremo a extremo con Apache NiFi a través de un mecanismo de gating que impide ejecutar fases posteriores sobre datos no validados. Además, la orquestación con NiFi, inicialmente planteada como una extensión condicionada, se completó dentro del propio proyecto. El OE2 se ha cumplido con la definición y el cálculo de las cinco red flags a partir de las variables de PLACSP, siguiendo el marco de Fazekas y de la Open Contracting Partnership. El OE3 se ha alcanzado mediante la construcción del Composite Risk Index bajo los tres esquemas de ponderación previstos: pesos iguales, jerarquía empírica de Fazekas y varianza observada. El OE4 se ha satisfecho evaluando la estabilidad de los rankings mediante correlaciones de Spearman entre esquemas y entre años, además del análisis de coincidencia en las primeras posiciones. Por último, el

OE5 se ha completado con un dashboard interactivo en Kibana que permite explorar los resultados por parte de perfiles no técnicos.

Los resultados obtenidos confirman la viabilidad y la solidez del enfoque adoptado. El sistema procesa más de cien mil contratos únicos y produce un índice cuya principal propiedad es su robustez frente a la elección de pesos: las correlaciones de Spearman entre esquemas, muy próximas a la unidad, muestran que el orden de los órganos apenas depende del método de agregación, lo que constituye la validación interna más relevante en ausencia de un ground truth. A ello se añade que el riesgo no se distribuye de forma difusa, sino que se concentra en una cola estrecha de órganos, lo que convierte el resultado en algo directamente utilizable desde el punto de vista de la supervisión. Finalmente, el contraste cualitativo de los casos extremos con fuentes públicas de control ofrece un indicio razonable de validez externa, ya que el sistema, construido únicamente a partir de datos abiertos, vuelve a señalar órganos sobre los que el control institucional ya había puesto el foco. Estas conclusiones deben leerse, no obstante, dentro de los límites ya documentados en la sección 4.3.2, que acotan el alcance del análisis sin invalidar sus resultados.

8.2 APORTACIONES

El trabajo realiza tres aportaciones principales. La primera es de alcance empírico y geográfico: hasta donde alcanza el conocimiento del autor, constituye el primer sistema de análisis de riesgo construido sobre datos primarios de PLACSP y centrado en la Administración General del Estado, cubriendo así un espacio que la literatura internacional no había abordado para el caso español. La segunda es de carácter técnico y metodológico: el proyecto diseña y documenta un pipeline reproducible con tecnologías Big Data de código abierto, organizado en capas, con jobs independientes, auditables y orquestados automáticamente, que podría ser reejecutado por cualquier institución con acceso a los datos públicos. La tercera se refiere a la robustez del índice: frente a los trabajos que adoptan un único esquema de ponderación, este proyecto aplica tres y mide explícitamente la estabilidad de los rankings, de modo que la robustez no se presupone, sino que se comprueba.

A estas aportaciones se suma un resultado transversal que recorre todo el trabajo: la disciplina interpretativa con la que se ha construido y presentado el sistema. En todo momento se ha planteado como una herramienta que identifica patrones estadísticos asociados a un mayor riesgo, y no como un mecanismo capaz de detectar o probar irregularidades. Esta distinción no reduce el valor del resultado; al contrario, es lo que lo hace metodológicamente defendible y operativamente útil para un organismo de control.

8.3 LÍNEAS DE TRABAJO FUTURO

El sistema admite varias extensiones naturales que parten directamente de su diseño actual.

La primera es la incorporación de un análisis de tipologías mediante clustering como complemento al índice. El Composite Risk Index permite ordenar a los órganos por nivel de riesgo, pero no describe de forma explícita la estructura de los perfiles que subyacen a ese resultado. Un clustering no supervisado sobre las cinco señales normalizadas, por ejemplo mediante k-medias o métodos jerárquicos, permitiría agrupar los órganos según el tipo de patrón que presentan y formalizar cuantitativamente las tipologías que en este trabajo se han identificado de forma descriptiva en la sección 7.7. Esto añadiría una segunda dimensión de lectura, complementaria al ranking: no solo qué órganos presentan mayor riesgo, sino también qué clase de riesgo concentran. Su carácter no supervisado, además, lo hace coherente con la naturaleza del problema, ya que no requiere etiquetas de las que no se dispone.

La segunda línea es la incorporación de una red flag de fragmentación de contratos, RF6, basada en la similitud semántica del objeto del contrato para detectar posibles divisiones artificiales de adjudicaciones. Su inclusión estaba prevista desde el inicio como una extensión condicionada, pero requiere técnicas de procesamiento de lenguaje natural y depende de la calidad del campo textual en PLACSP. Por ello, se reserva como continuación natural del trabajo.

La tercera es la extensión del sistema más allá de la Administración General del Estado, incorporando la contratación de comunidades autónomas y entidades locales. En ese escenario, el volumen de datos se multiplicaría varias veces. Es precisamente ahí donde la arquitectura distribuida elegida dejaría de ser una opción ventajosa para convertirse en una necesidad práctica, y donde el pipeline podría escalar sin necesidad de una reescritura sustancial.

La cuarta línea consiste en resolver la identidad de los órganos no identificados en los primeros años mediante el cruce del NIF con registros administrativos externos. Esta mejora permitiría recuperar para la interpretación varios de los órganos que hoy aparecen en posiciones altas del ranking sin nombre asociado y reduciría una de las principales limitaciones de interpretabilidad del sistema.

En conjunto, estas extensiones comparten un rasgo importante: ninguna exige rediseñar el sistema desde cero, sino ampliarlo a partir de la base ya construida. Esa capacidad de crecer sin rehacerse constituye, en sí misma, una confirmación de que las decisiones de arquitectura adoptadas a lo largo del proyecto fueron acertadas.

Capítulo 9. BIBLIOGRAFÍA

[1] Oficina Independiente de Regulación y Supervisión de la Contratación, «Informe Anual de Supervisión de la Contratación Pública de España 2024,» Ministerio de Hacienda, 2024.

[2] Jefatura del Estado (España), «Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público,» Boletín Oficial del Estado, núm. 272 (BOE-A-2017-12902), 2017.

[3] U. Menéndez, «Principales novedades de la Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público,» 2017. [En línea]. Available: https://www.uria.com/es/prensa/1093-principales-novedades-de-la-ley-92017-de-8-de-noviembre-de-contratos-del-sector-publico.

[4] Ministerio de Hacienda, Dirección General del Patrimonio del Estado, «Plataforma de Contratación del Sector Público: datos de licitaciones,» 2024. [En línea]. Available: https://www.hacienda.gob.es/en-GB/GobiernoAbierto/Datos%20Abiertos/Paginas/LicitacionesContratante.aspx.

[5] M. Fazekas y G. Kocsis, «Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data,» *British Journal of Political Science,* vol. 50, nº 1, pp. 155-164, 2020.

[6] Partnership, Open Contracting, «Red Flags in Public Procurement: A Guide to Using Data to Detect and Mitigate Risks,» 2024. [En línea]. Available: https://www.open-contracting.org/resources/red-flags-in-public-procurement-a-guide-to-using-data-to-detect-and-mitigate-risks/.

[7] M. Bauhr, Á. Czibik, J. de Fine Licht y M. Fazekas, «Lights on the shadows of public procurement: Transparency as an antidote to corruption,» *Governance,* vol. 33, nº 3, pp. 495-523, 2020.

[8] K. Shvachko, H. Kuang, S. Radia y R. Chansler, «The Hadoop Distributed File System,» de *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, Incline Village, NV (EE. UU.), 2010.

[9] Foundation, Apache Software, «Apache NiFi Overview,» 2024. [En línea]. Available: https://nifi.apache.org/docs/nifi-docs/html/overview.html.

[10] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker y I. Stoica, «Apache Spark: A Unified Engine for Big Data Processing,» *Communications of the ACM,* vol. 59, nº 11, pp. 56-65, 2016.

[11] Elastic, «Elasticsearch Documentation,» 2024 . [En línea]. Available: https://www.elastic.co/docs.

[12] M. J. García Rodríguez, V. Rodríguez Montequín, F. Ortega Fernández y J. M. Villanueva Balsera, «Spanish Public Procurement: legislation, open data source and extracting valuable information of procurement announcements,» *Procedia Computer Science,* vol. 164, pp. 441-448, 2019.

[13] M. Muñoz Plá, «A Decade of Public Procurement in Spain: A Longitudinal Open Dataset from the BOE (2014-2024),» arXiv, 2025.

[14] C. Kenny y M. Musatova, «'Red Flags of Corruption' in World Bank Projects: An Analysis of Infrastructure Contracts,» World Bank (Policy Research Working Paper 5243), Washington, DC, 2010.

[15] M. Fazekas, I. J. Tóth y L. P. King, «An Objective Corruption Risk Index Using Public Procurement Data,» *European Journal on Criminal Policy and Research,* vol. 22, nº 3, pp. 369-397, 2016.

[16] O. Basdevant, A. Abdou, M. Fazekas y E. David-Barrett, «Assessing Vulnerabilities to Corruption in Public Procurement and Their Price Impact,» International Monetary Fund (IMF Working Paper 2022/094), Washington, DC, 2022.

[17] F. Decarolis y C. Giorgiantonio, «Corruption red flags in public procurement: new evidence from Italian calls for tenders,» *EPJ Data Science,* vol. 11, nº 1, p. 16, 2022.

[18] E. S. Santos, M. M. Santos, M. Castro y J. T. Carvalho, «Detection of fraud in public procurement using data-driven methods: a systematic mapping study,» *EPJ Data Science,* vol. 14, 2025.

[19] OECD; Union, European; EC-JRC, Handbook on Constructing Composite Indicators: Methodology and User Guide, París: OECD Publishing, 2008.

[20] F. Caires, S. Peralta y D. Mendes, «Contract Splitting in Public Procurement,» SSRN (Working Paper 4599419), 2023.

[21] Gubernamentales, Red Interamericana de Compras, «Guía para la identificación de riesgos de corrupción en contratación pública utilizando ciencia de datos,» 2023. [En línea]. Available: https://ricg.org/es/publicaciones/lanzamiento-guia-para-la-identificacion-de-riesgos-de-corrupcion-en-contratacion-publica-utilizando-ciencia-de-datos/.

[22] M. Gnaldi y S. Del Sarto, «Measuring Corruption Risk in Public Procurement over Emergency Periods,» *Social Indicators Research,* vol. 172, pp. 859-877, 2024.

[23] Cuentas, Tribunal de, «Informe de fiscalización de la actividad asistencial del Instituto Nacional de Gestión Sanitaria, ejercicio 2016,» Tribunal de Cuentas, 2018.

[24] Cuentas, Tribunal de, «Informe de fiscalización de la contratación de emergencia (COVID-19), ejercicio 2020, ámbito socio-laboral y de la Seguridad Social,» Tribunal de Cuentas, 2022.

[25] Cuentas, Tribunal de, «Informe sobre la gestión de la farmacia de los hospitales del INGESA de Ceuta y Melilla (2022-2023),» Tribunal de Cuentas, 2025.

[26] Cuentas, Tribunal de, «Informe de fiscalización de los contratos de la AGE vinculados a las políticas de gasto 11 (Justicia), 12 (Defensa) y 14 (Política exterior y cooperación), ejercicio 2022,» Tribunal de Cuentas, 2025.

[27] Competencia, Comisión Nacional de los Mercados y la, «Resolución del expediente S/0008/21, licitaciones de material militar,» CNMC, 2023.

[28] Cuentas, Tribunal de, «Informe de fiscalización de las subvenciones de Ayuda Oficial al Desarrollo gestionadas por la AECID, ejercicio 2016,» Tribunal de Cuentas.

[29] E. Morales, «Exteriores destina más de 5,2 millones a contratos contra la corrupción en Mozambique,» *The Objective,* 8 mayo 2026.

[30] C. G. Revillas, «Github,» TFM_BIGDATA, 2026. [En línea]. Available: https://github.com/carlosgarcirev/TFM_BIGDATA.

ANEXO I. GESTIÓN DE CÓDIGO Y CONTROL DE VERSIONES

1. Introducción

Este anexo se describe la estrategia de versionado del código del proyecto mediante Git y GitHub, siguiendo el modelo de ramificación Gitflow. Todo el código, tanto los notebooks de desarrollo como los scripts de producción, se aloja en un repositorio público [30], lo que permite mantener la trazabilidad del historial completo de cambios y de las versiones publicadas.

Al tratarse de un proyecto individual, la aplicación de este flujo ha sido necesariamente secuencial. Las ramas se integran una tras otra, siguiendo el orden en que se completa cada parte del pipeline, sin desarrollo simultáneo sobre componentes independientes. Aun así, el modelo Gitflow sigue aportando una estructura útil, porque separa el código estable del código en integración y del trabajo en curso, y ordena la publicación de versiones mediante releases y etiquetas.

2. Modelo de ramificación

El repositorio se organiza en torno a cuatro tipos de rama.

- La rama main contiene únicamente versiones estables y publicables. Solo recibe código a través de ramas de release, y cada incorporación queda marcada con su correspondiente etiqueta de versión.
- La rama develop actúa como rama de integración. En ella se acumula el trabajo ya completado antes de su promoción a una versión estable.
- Las ramas feature/* se emplean para el desarrollo de cada notebook o job del pipeline. Cada una se crea a partir de develop y, una vez finalizado el trabajo correspondiente, se fusiona de nuevo en esa misma rama.

- Por último, las ramas release/* se utilizan de forma temporal para preparar una versión, incorporando ajustes finales y cambios de documentación antes de su integración en main.

Esta estructura puede observarse con claridad en el grafo de red del repositorio, donde se aprecia cómo las ramas se separan de develop, se reintegran posteriormente en ella y, finalmente, cómo develop se fusiona en main en cada publicación de versión.

3. Flujo de trabajo y commits

Para cada notebook se ha seguido un mismo ciclo de trabajo.

1. En primer lugar, se parte de una rama develop actualizada y se crea una rama feature/NN-nombre.
2. A continuación, el desarrollo se realiza mediante varios commits organizados por bloques lógicos, como la configuración inicial, las transformaciones, la escritura de resultados o las verificaciones intermedias, utilizando mensajes descriptivos.
3. Una vez completado el trabajo, la rama se publica en el repositorio remoto y se fusiona en develop sin avance rápido, de forma que el historial conserve de manera explícita el ciclo completo de cada feature.
4. Finalmente, la rama se elimina tras su integración.

El carácter secuencial de este flujo viene impuesto por la propia naturaleza del pipeline. Cada job consume la salida del anterior, por lo que las distintas features no se desarrollan en paralelo, sino de forma consecutiva, en el mismo orden lógico del procesamiento.

4. Versiones publicadas

A lo largo del proyecto se han publicado tres versiones principales, cada una integrada en main a través de una rama de release y marcada con su correspondiente etiqueta.

- La versión v1.0 constituye la primera versión usable del sistema e incluye los jobs 1 a 3, es decir, la carga, la limpieza, el enriquecimiento y el cálculo inicial de indicadores.
- La versión v1.1 amplía esa base hasta completar el pipeline de notebooks, incorporando los jobs 4 y 5, correspondientes a la construcción del índice compuesto de riesgo y a la indexación en Elasticsearch.
- La versión v2.0 corresponde a la versión final del proyecto y añade los scripts de producción preparados para su ejecución automática dentro del pipeline orquestado.

En todos los casos, el proceso ha seguido el mismo patrón: la rama de release se crea a partir de develop, se actualiza la documentación, se fusiona en main con su etiqueta de versión y, posteriormente, se reintegra en develop para mantener sincronizadas ambas ramas.

5. Estructura del repositorio y buenas prácticas

El repositorio diferencia de forma explícita el desarrollo analítico del código de producción. Los notebooks recogen el trabajo exploratorio y de construcción progresiva del pipeline, preservando la trazabilidad de cómo se desarrolló cada fase. La carpeta scripts/ contiene los scripts de Python derivados de esos notebooks una vez validados, que son los que emplea el pipeline automatizado. Junto a ello, el archivo README documenta el proyecto y el fichero .gitignore excluye los datos y los artefactos temporales del control de versiones.

Durante la gestión del repositorio se han seguido varias buenas prácticas. En primer lugar, se han excluido mediante .gitignore tanto los ficheros de datos como los artefactos generados por Jupyter, de manera que el versionado se limita al código y a la documentación relevante. En segundo lugar, las credenciales se han mantenido fuera del código fuente y se han leído desde variables de entorno. Además, se han utilizado mensajes de commit descriptivos para facilitar la comprensión del historial. Cuando, de forma puntual, una credencial se incorporó por error al repositorio, el historial completo se depuró para eliminar cualquier rastro de ella en todos los commits afectados.

6. Integración con el pipeline de producción

La carpeta scripts/ contiene una copia versionada del código que se ejecuta en producción. Conviene precisar, no obstante, que NiFi no toma los scripts directamente de este repositorio. El código que el pipeline ejecuta de forma efectiva reside en el clúster, en HDFS, y es desde allí desde donde se lanza. El repositorio conserva una segunda copia de esos mismos scripts con un único propósito: garantizar su control de versiones y su trazabilidad. En consecuencia, no forma parte de la ruta real de ejecución.

La orquestación en producción se realiza mediante Apache NiFi, que lanza de forma secuencial los scripts de Python previamente desplegados en el clúster. La salida final se publica en un índice de Elasticsearch, sobre el que Kibana construye el dashboard de indicadores por órgano-año.

De este modo, el repositorio cumple una doble función documental, ajena en ambos casos a la ejecución directa del sistema. Por un lado, conserva los notebooks de desarrollo, que recogen el proceso analítico seguido durante la construcción del pipeline. Por otro, almacena una copia versionada de los scripts de producción, de manera que el código ejecutado en el clúster queda también registrado, fechado y trazable en el historial de Git.

7. Conclusión

El uso de Git y GitHub bajo el modelo Gitflow ha permitido mantener el código del proyecto organizado, versionado y completamente trazable. Aunque se trata de un trabajo individual y, por tanto, desarrollado de forma secuencial, la separación entre ramas estables, de integración y de trabajo en curso, junto con la publicación ordenada de versiones mediante etiquetas, aporta estructura y reproducibilidad y deja un historial limpio y comprensible del proceso seguido.

ANEXO II. PLANIFICACIÓN DEL PROYECTO

1. Metodología de planificación

La planificación del proyecto se ha estructurado mediante un esquema de paquetes de trabajo, cada uno de ellos descompuesto en actividades y tareas con una estimación específica de horas, un perfil responsable y una naturaleza asociada, ya sea de desarrollo técnico, análisis o documentación. El calendario de trabajo abarca del 11 de mayo al 11 de junio de 2026, con una dedicación ordinaria concentrada en jornadas parciales de tarde.

Al tratarse de un trabajo individual, los perfiles definidos no corresponden a personas distintas, sino a los diferentes roles asumidos por el autor a lo largo del proyecto. Esta diferenciación permite ordenar mejor la planificación y sirve, además, como base de la estimación económica desarrollada en el anexo correspondiente.

2. Paquetes de trabajo

PT	Paquete	Periodo	Contenido
PT1	EDA y validación de datos	11-15 may	Descarga de PLACSP con OpenPLACSP, validación de cobertura de variables, análisis exploratorio y verificación del entorno del clúster.
PT2	Ingesta y primeros jobs	16-21 may	Flujo de ingesta de NiFi a HDFS, organización en capas raw/curated/analytics, Job 1 de filtrado y curación del subconjunto AGE y Job 2 de enriquecimiento con

			el maestro de órganos.
PT3	Red flags (Job 3)	22-27 may	Cálculo de los indicadores RF1 a RF5, deduplicación de snapshots y agregación a nivel de órgano-año.
PT4	Composite Risk Index y estabilidad	28 may-1 jun	Normalización min-max, construcción del CRI bajo tres esquemas de ponderación, análisis de estabilidad de rankings y caracterización descriptiva de tipologías de riesgo.
PT5	Resultados, Kibana y validación	2-6 jun	Indexación en Elasticsearch, construcción del dashboard de Kibana y validación cualitativa de los casos recurrentes mediante fuentes externas.
PT6	Redacción de la memoria	15 may-10 jun	Redacción de los capítulos, revisión bibliográfica, elaboración del resumen y cierre formal de la memoria.

Tabla 5: Paquetes de trabajo de la Planificación del Proyecto

El esfuerzo total planificado asciende a 136 horas. Los paquetes técnicos, PT1 a PT5, siguen una lógica predominantemente secuencial, ya que cada fase consume la salida de la anterior. La redacción de la memoria, agrupada en PT6, se desarrolló en paralelo desde la mitad del calendario, de forma que la documentación avanzase a medida que se completaban los distintos bloques del sistema.

3. Cronograma

El plan de trabajo se ha desagregado a nivel de tarea, indicando para cada una de ellas la duración estimada en horas, el perfil responsable y la naturaleza de la actividad. Este desglose completo se recoge en la Ilustración 9.

TFM - Análisis de Riesgo en Contratación Pública Española - Plan de Trabajo 11/05/2026 - 11/06/2026						
Paquete	**Actividad**	**Tarea**	**Horas**	**Perfil**	**Naturaleza**	**Fechas**
PT1	ACT1.1 - Descarga y Prep. Dataset	TR1.1.1 - Descarga PLACSP 2018-2023 con OpenPLACSP	4	Ing. de Datos	Desarrollo Técnico	2026-05-11,2026-05-11
PT1	ACT1.1 - Descarga y Prep. Dataset	TR1.1.2 - Validación estructura y cobertura de columnas críticas	4	Analista de Datos	Análisis	2026-05-12,2026-05-12
PT1	ACT1.2 - Análisis Exploratorio (EDA)	TR1.2.1 - Estadísticas descriptivas y distribuciones de variables	4	Analista de Datos	Análisis	2026-05-13,2026-05-13
PT1	ACT1.2 - Análisis Exploratorio (EDA)	TR1.2.2 - Análisis nulos, outliers y calidad del dataset	4	Analista de Datos	Análisis	2026-05-14,2026-05-14
PT1	ACT1.3 - Configuración Entorno	TR1.3.1 - Verificación recursos clúster Hadoop/Spark	4	Ing. de Sistemas	Desarrollo Técnico	2026-05-15,2026-05-15
		SUBTOTAL PT1	**20**			
PT2	ACT2.1 - Ingesta NiFi + HDFS	TR2.1.1 - Config. NiFi: flujo de ingesta a HDFS con enrutado dinámico	8	Ing. de Datos	Desarrollo Técnico	2026-05-16,2026-05-17
PT2	ACT2.1 - Ingesta NiFi + HDFS	TR2.1.2 - Estructura capas HDFS (raw / curated / analytics) en Parquet	4	Ing. de Datos	Desarrollo Técnico	2026-05-18,2026-05-18
PT2	ACT2.2 - Spark Job 1: Filtrado	TR2.2.1 - Filtros subset AGE, estado, tipo contrato y presupuesto positivo	8	Desarrollador de Datos	Desarrollo Técnico	2026-05-19,2026-05-20
PT2	ACT2.3 - Spark Job 2: Enriquecimiento	TR2.3.1 - Join en dos pasos con maestro de órganos (plataforma y NIF); Parquet curated	4	Desarrollador de Datos	Desarrollo Técnico	2026-05-21,2026-05-21
		SUBTOTAL PT2	**24**			
PT3	ACT3.1 - Spark Job 3: Red Flags	TR3.1.1 - RF1 Single Bidding: binario por licitación y tasa órgano-año	4	Científico de Datos	Desarrollo Técnico	2026-05-22,2026-05-22
PT3	ACT3.1 - Spark Job 3: Red Flags	TR3.1.2 - RF2 Procedimiento No Competitivo (NCB_rate)	4	Científico de Datos	Desarrollo Técnico	2026-05-23,2026-05-23
PT3	ACT3.1 - Spark Job 3: Red Flags	TR3.1.3 - RF3 Precio relativo (importe/presupuesto; exclusión de ratios > 1,5)	4	Científico de Datos	Desarrollo Técnico	2026-05-24,2026-05-24
PT3	ACT3.1 - Spark Job 3: Red Flags	TR3.1.4 - RF4 Concentración de Proveedores (HHI y cuota del proveedor principal)	4	Científico de Datos	Desarrollo Técnico	2026-05-25,2026-05-25
PT3	ACT3.1 - Spark Job 3: Red Flags	TR3.1.5 - RF5 Plazo de Presentación (STP, percentil P10 por tipo contrato)	4	Científico de Datos	Desarrollo Técnico	2026-05-26,2026-05-26
PT3	ACT3.2 - Particionado y Validación	TR3.2.1 - Deduplicación de snapshots y agregación a features_organo_anyo	4	Científico de Datos	Análisis	2026-05-27,2026-05-27
		SUBTOTAL PT3	**24**			
PT4	ACT4.1 - Normalización y Agregación	TR4.1.1 - Min-max sobre población de calidad alta (3 esquemas: iguales, Fazekas, varianza)	8	Científico de Datos	Análisis	2026-05-28,2026-05-29
PT4	ACT4.2 - Estabilidad de Rankings	TR4.2.1 - Correlaciones de Spearman entre los 3 rankings resultantes	4	Analista de Datos	Análisis	2026-05-30,2026-05-30
PT4	ACT4.2 - Estabilidad de Rankings	TR4.2.2 - Top-K stability (top 10/20/50) y coincidencia entre esquemas	4	Analista de Datos	Análisis	2026-05-31,2026-05-31
PT4	ACT4.3 - Tipologías de Riesgo	TR4.3.1 - Mapa de calor single bidding x concentración y perfiles de riesgo	4	Científico de Datos	Análisis	2026-06-01,2026-06-01
		SUBTOTAL PT4	**20**			
PT5	ACT5.1 - Indexación y Dashboard	TR5.1.1 - Indexación resultados en Elasticsearch	4	Ing. de Datos	Desarrollo Técnico	2026-06-02,2026-06-02
PT5	ACT5.1 - Indexación y Dashboard	TR5.1.2 - Dashboard Kibana: ranking y red flags por órgano, ministerio, calidad y año	8	Desarrollador de Datos	Desarrollo Técnico	2026-06-03,2026-06-04
PT5	ACT5.2 - Validación Cualitativa	TR5.2.1 - Spot-check de órganos recurrentes en cabeza (Tribunal de Cuentas, CNMC, prensa)	4	Investigador	Análisis	2026-06-05,2026-06-05
		SUBTOTAL PT5	**16**			
PT6	ACT6.1 - Redacción Caps. 1-4	TR6.1.1 - Caps. 1-2: introducción, motivación y marco tecnológico	4	Redactor Técnico	Documentación	2026-05-15,2026-06-10
PT6	ACT6.1 - Redacción Caps. 1-4	TR6.1.2 - Cap. 3: marco teórico y estado del arte	8	Investigador	Documentación	2026-05-15,2026-06-10
PT6	ACT6.1 - Redacción Caps. 1-4	TR6.1.3 - Cap. 4: definición, alcance y metodología	8	Redactor Técnico	Documentación	2026-05-15,2026-06-10
PT6	ACT6.2 - Redacción Caps. 5-8	TR6.2.1 - Caps. 5-6: sistema desarrollado y Composite Risk Index	4	Redactor Técnico	Documentación	2026-05-15,2026-06-10
PT6	ACT6.2 - Redacción Caps. 5-8	TR6.2.2 - Caps. 7-8: resultados, conclusiones y trabajo futuro	4	Redactor Técnico	Documentación	2026-05-15,2026-06-10
PT6	ACT6.3 - Revisión y Entrega Final	TR6.3.1 - Bibliografía, resumen (ES/EN) y revisión de formato	4	Redactor Técnico	Documentación	2026-05-15,2026-06-10
		SUBTOTAL PT6	**32**			
		TOTAL	**136**			

Ilustración 9: Planificación del Proyecto.

El plan se compone de veintiocho tareas agrupadas en seis paquetes de trabajo. Para cada una de ellas se especifica la estimación horaria, el perfil que la asume y su naturaleza, ya sea de desarrollo técnico, análisis o documentación. Los subtotales por paquete y el total de 136

horas permiten dimensionar el esfuerzo previsto y sirven, además, como base para la estimación de costes desarrollada en el anexo correspondiente.

Sobre esta base, la distribución temporal del proyecto se representa en el diagrama de Gantt de la Ilustración 10, que abarca el periodo comprendido entre el 11 de mayo y el 11 de junio de 2026, con una dedicación de referencia de cuatro horas diarias.

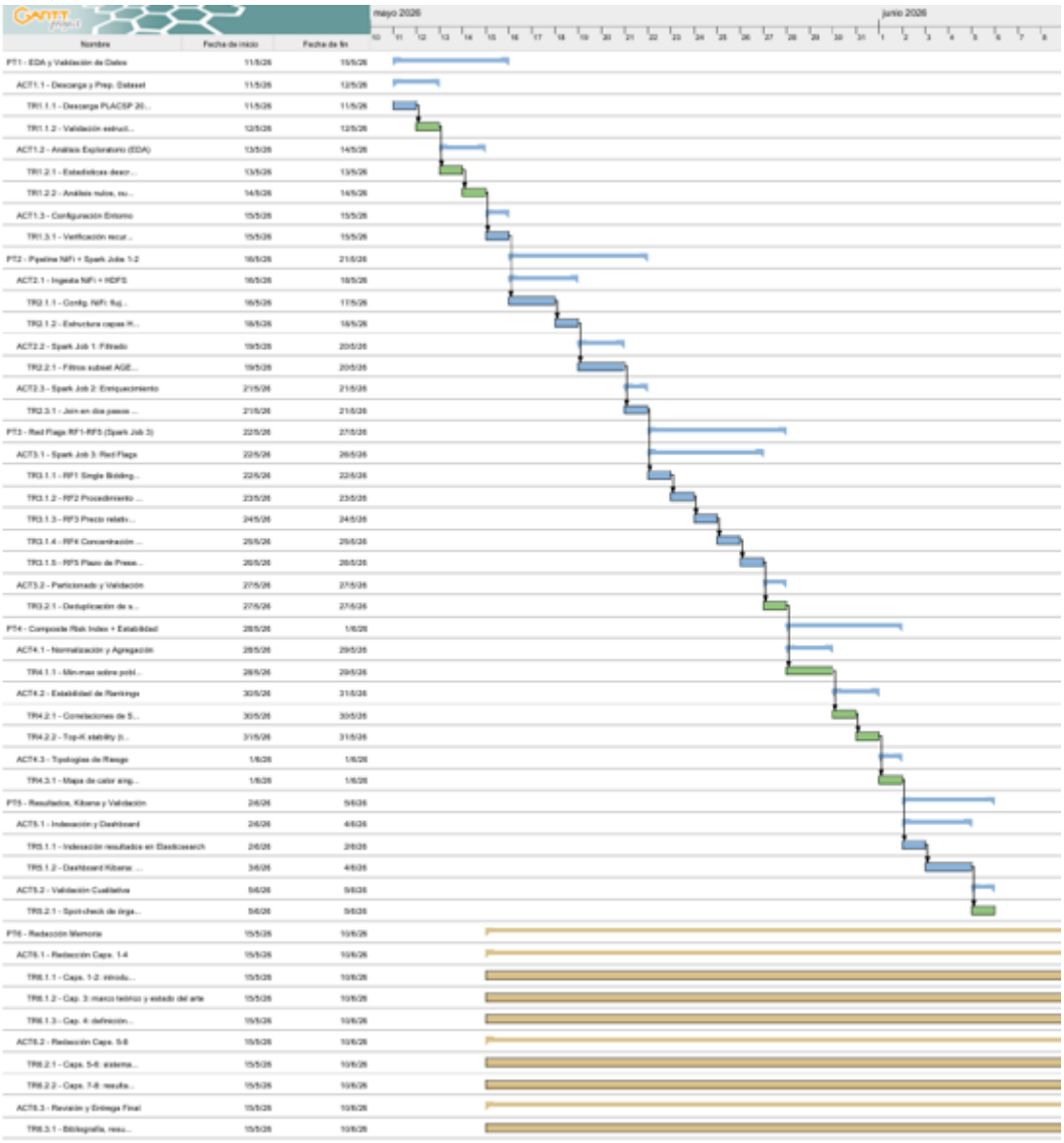


Ilustración 10: Diagrama de Gantt de la planificación del proyecto.

Los paquetes PT1 a PT5 siguieron una secuencia técnica encadenada, de modo que cada uno partía de la salida generada por el anterior. En paralelo, el PT6, correspondiente a la redacción de la memoria, se desarrolló durante buena parte del periodo con el fin de documentar el trabajo a medida que este avanzaba.

4. Perfiles y dedicación

La planificación distingue siete perfiles en función de la naturaleza del trabajo. El ingeniero de datos asume las tareas ligadas al pipeline de NiFi, HDFS y la conectividad con Spark. El ingeniero de sistemas se ocupa de la verificación del clúster y del entorno de ejecución. El desarrollador de datos se centra en los jobs de Spark orientados al filtrado, los joins y el feature engineering. El analista de datos aborda el análisis exploratorio y la evaluación de rankings. El científico de datos se encarga del cálculo de las red flags, la normalización, la construcción del índice y la elaboración de tipologías de riesgo. El investigador se ocupa del contraste cualitativo y del soporte bibliográfico. Finalmente, el redactor técnico concentra la elaboración formal de la memoria.

La carga de trabajo se concentra en los perfiles técnicos durante la primera mitad del calendario y se desplaza progresivamente hacia la documentación y la validación en la segunda. Esta distribución responde a la propia lógica del proyecto: primero se construye y valida el pipeline, y después se consolidan los resultados y su interpretación.

5. Desviaciones respecto a la planificación inicial

El seguimiento del plan permitió ajustar algunos elementos metodológicos y técnicos a medida que avanzaba el desarrollo. El primero afecta a los esquemas de ponderación. Finalmente se implementaron pesos iguales, la jerarquía empírica de Fazekas y la varianza observada, en lugar de los esquemas igual, PCA y entrópico contemplados en una formulación inicial.

El segundo ajuste se refiere a la validación. Una vez confirmado el carácter no supervisado del enfoque, se descartó la división en conjuntos de entrenamiento, validación y test, prevista en etapas preliminares, y la fiabilidad del análisis se articuló mediante el umbral

mínimo de treinta contratos por órgano-año. El tercero afecta al clustering secundario. El análisis de tipologías mediante clustering se reclasificó como línea de trabajo futuro y, en esta versión del proyecto, las tipologías se abordaron de forma descriptiva a través del mapa de calor de single bidding frente a concentración.

Por último, la orquestación también se concretó de forma más precisa. La conversión de los ficheros ATOM se mantuvo como paso previo mediante OpenPLACSP, mientras que NiFi quedó centrado en la ingesta a HDFS y en la orquestación del pipeline con control de errores. Ninguno de estos ajustes alteró el calendario general ni los objetivos del proyecto. Se trata, más bien, de refinamientos metodológicos resueltos dentro de los plazos previstos.

ANEXO III. ESTIMACIÓN ECONÓMICA

1. Metodología

La estimación económica del proyecto se ha realizado considerando los recursos humanos, técnicos y de servicios necesarios para su desarrollo. Los costes se clasifican en gastos de capital (CAPEX) y gastos operativos (OPEX), de acuerdo con la distinción habitual en la gestión de proyectos tecnológicos. Sobre el coste base se ha aplicado un margen de beneficio del 30%, con el fin de simular un escenario de prestación en condiciones de mercado profesional, y posteriormente el Impuesto sobre el Valor Añadido (IVA) del 21%. Las horas imputadas a cada perfil proceden de la planificación detallada en el anexo anterior.

2. Resumen de costes

Categoría	Descripción	Tipo	Coste base (€)	Margen (€)	IVA (€)	Total (€)
Personal	Equipo del proyecto (136 h)	OPEX	4.808,00	1.442,40	1.312,58	7.562,98
Infraestructura Big Data	Clúster Spark/Hadoop con HDFS, Elasticsearch, NiFi y Kibana (equivalente cloud)	OPEX	750,00	225,00	204,75	1.179,75
Software y licencias	Stack de código abierto (Spark, Hadoop, NiFi, Elasticsearch, Kibana, Python)	OPEX	0,00	0,00	0,00	0,00
Suministros	Electricidad e internet (reparto alícuota)	OPEX	50,00	15,00	13,65	78,65

Varios	Gestión bibliográfica y herramientas visuales	OPEX	30,00	9,00	8,19	47,19
Hardware	Portátil de soporte local	CAPEX	850,00	255,00	232,05	1.337,05
TOTAL OPEX			5.638,00€	1.691,40€	1.539,17€	**8.868,57€**
TOTAL CAPEX			850,00€	255,00€	232,05€	**1.337,05€**
TOTAL GENERAL			**6.488,00€**	**1.946,40€**	**1.771,22€**	**10.205,62€**

Tabla 6: Resumen de costes por categoría.

3. Coste de la infraestructura

El componente de coste más característico del proyecto es la infraestructura. A diferencia de un desarrollo convencional ejecutado sobre un único equipo, este sistema se apoya en un stack completo de tecnologías Big Data, que incluye almacenamiento distribuido en HDFS, planificación de recursos con YARN, procesamiento en Spark, un motor Elasticsearch y contenedores para NiFi y Kibana. Replicar este entorno sin disponer previamente de un clúster requeriría, bien la contratación de servicios gestionados en la nube, bien el despliegue sobre servidores propios. En ambos casos, la configuración y el mantenimiento del entorno suponen un esfuerzo técnico especializado.

La cifra recogida en la tabla, 750 euros para el periodo del proyecto, representa una estimación del coste equivalente de mercado de esa infraestructura en un entorno cloud. En la práctica, esta infraestructura fue proporcionada por ICAI como entidad colaboradora, por lo que no supuso un desembolso efectivo para el proyecto. No obstante, se incorpora a la estimación como una contribución en especie con valor económico, dada su relevancia para el desarrollo del sistema.

Conviene señalar, además, que todo el stack empleado es software de código abierto. Por ello, el coste de licencias es nulo y el peso económico del proyecto se concentra fundamentalmente en el personal y en la infraestructura. Esta distribución es coherente con el carácter reproducible y abierto que persigue la solución desarrollada.

4. Desglose de costes de personal

Los costes de personal se han estimado a partir de los distintos perfiles asumidos durante el proyecto, asignando a cada uno un coste horario neto representativo de valores habituales en el mercado profesional. El total de horas coincide con el recogido en la planificación.

Perfil	**Horas**	**Coste neto/hora (€)**	**Coste neto (€)**
Científico de Datos	36	45	1.620,00
Ingeniero de Datos	20	40	800,00
Desarrollador de Datos	20	38	760,00
Analista de Datos	20	28	560,00
Redactor Técnico	24	26	624,00
Investigador	12	25	300,00
Ingeniero de Sistemas	4	36	144,00
Total	**136**		**4.808,00**

Tabla 7: Desglose de costes de personal.