



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería Electrónica y Eléctrica

**Escuela Profesional de Ingeniería de
Telecomunicaciones**

Diseño e implementación de un servicio digital de valor agregado para identificación de clientes basado en una arquitectura Big Data

TRABAJO DE SUFICIENCIA PROFESIONAL

**Para optar el Título Profesional de Ingeniera de
Telecomunicaciones**

AUTOR

Ruth María TAPIA AGUILAR

ASESOR

Mg. Ana Milagros CORONEL CHAMORRO

Lima, Perú

2026



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Tapia, R. (2026). *Diseño e implementación de un servicio digital de valor agregado para identificación de clientes basado en una arquitectura Big Data*. [Trabajo de Suficiencia Profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería Electrónica y Eléctrica, Escuela Profesional de Ingeniería de Telecomunicaciones]. Repositorio institucional Cybertesis UNMSM.

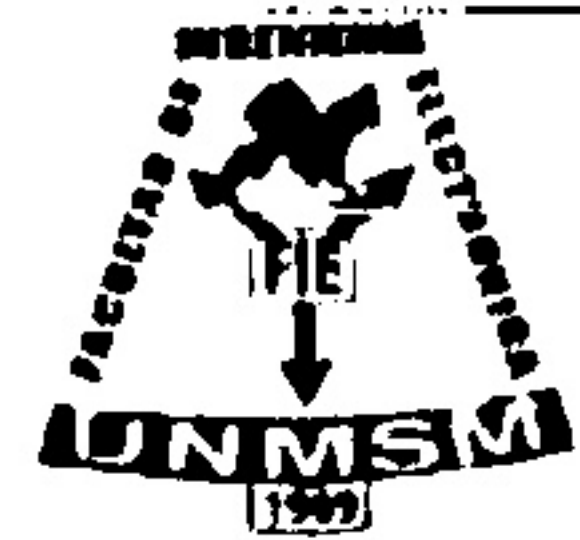
Metadatos complementarios

Datos de autor	
Nombres y apellidos	Ruth María Tapia Aguilar
Tipo de documento de identidad	DNI
Número de documento de identidad	72716166
URL de ORCID	No Aplica
Datos de asesor	
Nombres y apellidos	Ana Milagros Coronel Chamorro
Tipo de documento de identidad	DNI
Número de documento de identidad	09752919
URL de ORCID	https://orcid.org/0000-0003-2165-0742
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Teresa Esther Nuñez Zuñiga
Tipo de documento	DNI
Número de documento de identidad	08006778
Miembro del jurado 1	
Nombres y apellidos	Wilbert Chavez Irazabal
Tipo de documento	DNI
Número de documento de identidad	08121733
Miembro del jurado 2	
Nombres y apellidos	Jean Carlos Malca Fernández
Tipo de documento	DNI
Número de documento de identidad	45535181
Datos de investigación	
Línea de investigación	C.0.3.28. Comunicación/redes y sistemas distribuidos
Grupo de investigación	No Aplica

Agencia de financiamiento	No Aplica
Ubicación geográfica de la investigación	Edificio: Platinum Plaza I País: Perú Departamento: Lima Provincia: Lima Distrito: San Isidro Calle: C. Dean Valdivia 148 Latitud: -12.093480100627184 Longitud: -77.02760447264185
Año o rango de años en que se realizó la investigación	Enero 2024 – Mayo 2024
URL de disciplinas OCDE	Ingeniería, Tecnología https://purl.org/pe-repo/ocde/ford#2.00.00 Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04 Telecomunicaciones https://purl.org/pe-repo/ocde/ford#2.02.05



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
(Universidad del Perú, DECANA DE AMÉRICA)
FACULTAD DE INGENIERÍA ELECTRÓNICA Y ELÉCTRICA
Teléfono 619-7000 Anexo 4226
Calle Germán Amezaga 375 – Lima 1 – Perú



ACTA DE SUSTENTACIÓN
TRABAJO DE
SUFICIENCIA PROFESIONAL
ACTA N°40/FIEE-PPP/2026

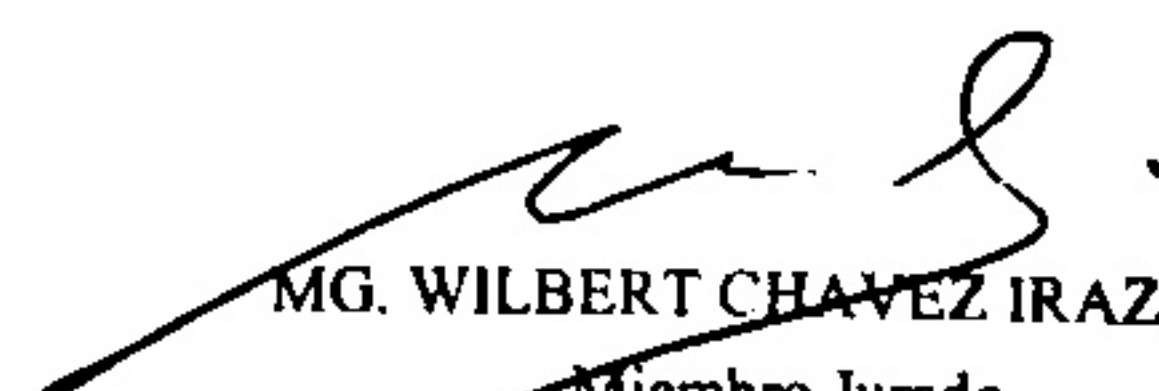
Los suscritos Miembros del Jurado, docentes permanentes de las Escuelas Profesionales de la Facultad de Ingeniería Electrónica y Eléctrica, reunidos en la fecha 19 de Junio del 2026, como presidente de Jurado la **DRA. TERESA ESTHER NUÑEZ ZUÑIGA**, integrado por el Miembro de Jurado el **MG. WILBERT CHAVEZ IRAZABAL**, el Miembro de Jurado el **MG. JEAN CARLOS MALCA FERNÁNDEZ** y Miembro Asesor la **MG. ANA MILAGROS CORONEL CHAMORRO**.

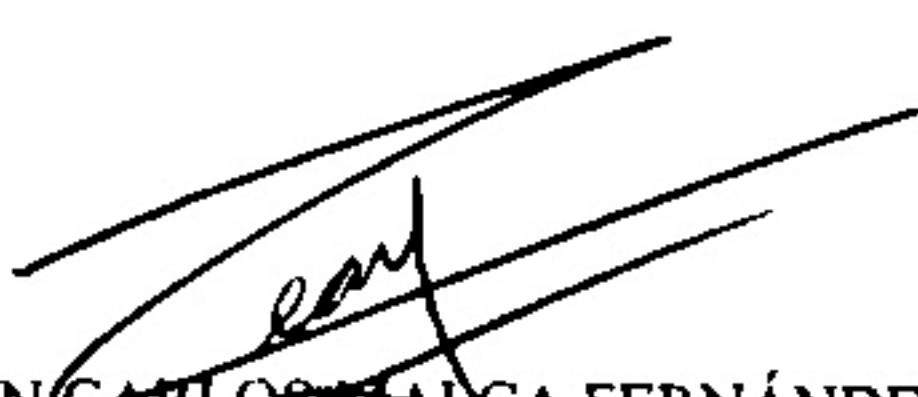
Después de escuchar la Sustentación Presencial del Trabajo de Suficiencia Profesional de la Bachiller **RUTH MARÍA TAPIA AGUILAR** con código 16190027 que para optar el Título Profesional de Ingeniera de Telecomunicaciones sustentó el Trabajo de Suficiencia Profesional titulado **DISEÑO E IMPLEMENTACIÓN DE UN SERVICIO DIGITAL DE VALOR AGREGADO PARA IDENTIFICACIÓN DE CLIENTES BASADO EN UNA ARQUITECTURA BIG DATA**.

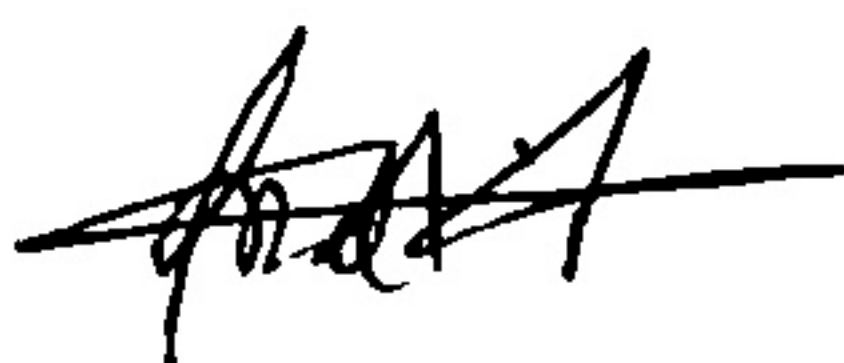
El jurado examinador procedió a formular las preguntas reglamentarias y, luego de una deliberación en privado, decidió otorgarle el calificativo de 17 Diecisiete.

Ciudad Universitaria, 19 de junio del 2026


DRA. TERESA ESTHER NUÑEZ ZUÑIGA
Presidente Jurado


MG. WILBERT CHAVEZ IRAZABAL
Miembro Jurado


MG. JEAN CARLOS MALCA FERNÁNDEZ
Miembro Jurado


MG. ANA MILAGROS CORONEL CHAMORRO
Miembro Asesor



UNIVERSIDAD NACIONAL MAYOR DE
SAN MARCOS
Universidad del Perú, Decana de América

CERTIFICADO DE SIMILITUD

Yo, **ANA MILAGROS CORONEL CHAMORRO** en mi condición de asesor acreditado con el Dictamen N°42/FIEE-PPP/2026 del Trabajo de Suficiencia Profesional titulado **DISEÑO E IMPLEMENTACIÓN DE UN SERVICIO DIGITAL DE VALOR AGREGADO PARA IDENTIFICACIÓN DE CLIENTES BASADO EN UNA ARQUITECTURA BIG DATA**, presentado por la bachiller **RUTH MARÍA TAPIA AGUILAR**, para optar el título profesional de Ingeniera de Telecomunicaciones. CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el 4% de similitud, nivel PERMITIDO para continuar con los trámites correspondientes y para su publicación en el repositorio institucional.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título profesional de Ingeniera de Telecomunicaciones.

Lima, 19 de Junio del 2026

Firma del Asesor

DNI: 09752919

Nombres y apellidos del asesor:

ANA MILAGROS CORONEL CHAMORRO

DEDICATORIA

A Dios, por ser la guía y fortaleza en cada paso de mi camino. A mis padres, cuyo amor, esfuerzo y ejemplo son la base de todo lo que soy. A mis hermanos, compañeros incondicionales que me inspiran a crecer y superarme cada día.

A ustedes, pilares fundamentales de mi vida, dedico este trabajo, como muestra de gratitud por ser mi motor y mi impulso para alcanzar cada meta trazada.

AGRADECIMIENTOS

Agradezco profundamente a mi asesora, Ana Coronel, por su guía constante y por compartir sus conocimientos que fueron fundamentales para el desarrollo de este Trabajo de Suficiencia Profesional.

De igual manera, expreso mi gratitud a mi familia y amigos, quienes con su apoyo, comprensión y aliento incondicional me brindaron la motivación necesaria para continuar y culminar con éxito este proceso.

RESUMEN

El presente Trabajo de Suficiencia Profesional describe el diseño, implementación y validación del servicio digital de valor agregado correspondiente al indicador Marca Pyme, desarrollado para la identificación de clientes del segmento Pyme en una institución financiera del sector bancario peruano. El objetivo del proyecto fue construir un indicador único, trazable y reutilizable a nivel institucional bajo una arquitectura Big Data distribuida. La metodología empleada se sustentó en la ejecución de un ciclo de desarrollo de cinco meses estructurado en seis Sprints bajo el marco Scrum, utilizando procesamiento distribuido con Apache Spark a través de PySpark sobre la capa Master del Data Lake corporativo en HDFS; la lógica de clasificación se estructuró en tres etapas técnicas y tres bloques decisionales alineados con la Recomendación 2003/361/CE de la Unión Europea y el marco regulatorio BIS IV. Como resultado, el campo `marca_pyme` fue integrado exitosamente en la tabla maestra de segmentación, validado funcional y técnicamente en entorno Sandbox, formalizado bajo el gobierno de datos y desplegado en el entorno productivo. Se concluye que el servicio constituye un activo digital escalable, reutilizable y alineado con la arquitectura Big Data distribuida de la institución.

Palabras clave: Big Data, procesamiento distribuido, Data Lake, segmentación de clientes, gobierno de datos.

ABSTRACT

This Professional Sufficiency Work describes the design, implementation and validation of the value-added digital service corresponding to the Marca Pyme indicator, developed for the identification of SME-segment customers in a financial institution of the Peruvian banking sector. The objective of the project was to build a unique, traceable and reusable indicator at the institutional level under a distributed Big Data architecture. The methodology was based on the execution of a five-month development cycle structured into six Sprints under the Scrum framework, using distributed processing with Apache Spark through PySpark on the Master layer of the corporate Data Lake in HDFS; the classification logic was structured into three technical stages and three decisional blocks aligned with Recommendation 2003/361/EC of the European Union and the BIS IV regulatory framework. As a result, the `marca_pyme` field was successfully integrated into the master segmentation table, functionally and technically validated in a Sandbox environment, formalized under data governance and deployed in the production environment. It is concluded that the service constitutes a scalable, reusable digital asset aligned with the institution's distributed Big Data architecture.

Keywords: Big Data, distributed processing, Data Lake, customer segmentation, data governance.

ÍNDICE GENERAL

CARÁTULA	i
RESUMEN	iv
ABSTRACT	v
ÍNDICE GENERAL	vi
LISTA DE FIGURAS	ix
LISTA DE TABLAS	xi
CAPÍTULO I: INTRODUCCION	1
1.1 Objetivos del informe	1
1.2 Estructura del informe	2
CAPÍTULO II: INFORMACION DEL LUGAR DONDE SE DESARROLLO	
LA ACTIVIDAD	4
2.1 Institución – Actividad que desarrolla	4
2.2 Periodo de duración de la actividad	4
2.3 Finalidad y objetivos de la entidad	4
2.4 Razón social	5
2.5 Dirección postal	5
2.6 Nombre y correo electrónico del profesional a cargo	5
CAPÍTULO III:DESCRIPCIÓN DE LA ACTIVIDAD	6
3.1 Organización de la actividad	6
3.2 Finalidad y objetivos de la Actividad	9
3.2.1 <i>Finalidad</i>	9
3.2.2 <i>Objetivos</i>	9

	vii
3.3 Problemática	10
3.3.1 <i>Problema General</i>	12
3.3.2 <i>Problemas Específicos</i>	12
3.3.3 <i>Justificación e importancia de la investigación</i>	13
3.4 Metodología	15
3.4.1 <i>Bases teóricas</i>	15
3.4.2 <i>Marco conceptual</i>	17
3.5 Procedimiento	33
3.5.1 <i>Adaptación del proyecto a la Metodología Scrum</i>	34
3.5.2 <i>Desarrollo del ciclo de Sprints</i>	35
3.6 Resultado de la actividad	75
CAPITULO IV: DESCRIPCIÓN DE RESULTADOS	78
4.1. Descripción de la implementación	78
4.1.1. <i>Etapas de análisis y diagnóstico (Sprint 1)</i>	78
4.1.2. <i>Etapas de diseño lógico del servicio (Sprint 2)</i>	79
4.1.3. <i>Etapas de implementación en entorno sandbox y validación funcional/técnica (Sprint 3 y Sprint 4)</i>	79
4.1.4. <i>Etapas de formalización y gobierno de datos (Sprint 5)</i>	80
4.1.5. <i>Etapas de implementación productiva (Sprint 6)</i>	80
4.1.6. <i>Etapas de cierre y evaluación del servicio</i>	81
4.2. Evaluación Económica	81
4.2.1. <i>Estimación de costos de implementación (CAPEX)</i>	82
4.2.2. <i>Costos operativos (OPEX)</i>	83
4.2.3. <i>Beneficios del servicio digital</i>	83
4.2.4. <i>Análisis costo-beneficio</i>	85
CAPÍTULO V: CONCLUSIONES	87
CAPÍTULO VI: RECOMENDACIONES	89

CAPÍTULO VII: REFERENCIAS BIBLIOGRÁFICAS**90****CAPÍTULO VIII: ANEXOS****94**

LISTA DE FIGURAS

Figura 1 Organigrama de equipos participantes. .	6
Figura 2 Comparación entre enfoques de repositorios de Datos.	20
Figura 3 Estructura interna de un archivo Parquet.	23
Figura 4: Arquitectura por capas del Data Lake corporativo sobre HDFS.	25
Figura 5 Arquitectura general de Apache Spark.	27
Figura 6 Modelo de referencia OSI y posicionamiento del servicio digital de valor agregado.	29
Figura 7 Modelo de referencia del marco de trabajo Scrum	32
Figura 8 Diagrama de proceso	35
Figura 9 Historia de usuario - Jira.	38
Figura 10 Cronograma Gantt de la planificación del ciclo de desarrollo del proyecto.	39
Figura 11 Comando HDFS para verificación de disponibilidad y particionamiento de fuentes.	42
Figura 12 printSchema simulado de la tabla destino.	43
Figura 13 Evidencia de validación de volumetría y unicidad de llaves de la tabla destino.	46
Figura 14 Diagrama lógico de clasificación del indicador Marca Pyme.	51
Figura 15 Lógica condicional final del indicador.	53
Figura 16 Lectura de fuentes.	55
Figura 17 Ejemplo de integración.	56
Figura 18 Transformación monetaria.	57
Figura 19 Bloque 1 - Evaluación de variables económicas.	59
Figura 20 Bloque 2 - Evaluación de variables regulatorias.	59
Figura 21 Bloque 3 - Evaluación de variables estructurales.	60
Figura 22 Asignación final indicador Marca Pyme.	60
Figura 23 Distribución preliminar del indicador Marca Pyme sobre el universo de clientes corporativos.	61

	x
Figura 24 Comparación de registros.	63
Figura 25 HU de elaboración de Dictamen.	68
Figura 26 Vista del indicador Marca Pyme integrado en la tabla maestra de segmentación.	77

LISTA DE TABLAS

Tabla 1. Comparación entre Data Warehouse y Data Lake	20
Tabla 2. Dimensiones de la calidad de datos	28
Tabla 3. Cronograma de desarrollo del Proyecto	35
Tabla 4. Mapeo de fuentes funcionales del indicador Marca Pyme	40
Tabla 5. Diccionario de campos relevantes de la tabla maestra de segmentación	43
Tabla 6. Comparativo de periodicidad de las fuentes del indicador Marca Pyme	44
Tabla 7. Características de las variables de entrada del servicio	47
Tabla 8. Proceso de conversión monetaria del Bloque 2	52
Tabla 9. Estructura de evaluación paramétrica por grupos	52
Tabla 10. Operaciones de cruce aplicadas durante la consolidación	56
Tabla 11. Transformaciones aplicadas durante la normalización	58
Tabla 12. Validación por registro: casos representativos del comportamiento del indicador	64
Tabla 13. Estructura del dictamen técnico del indicador Marca Pyme	67
Tabla 14. Elementos de trazabilidad y versionamiento del indicador	70
Tabla 15. Estimación de esfuerzo e inversión por perfil	82
Tabla 16. Estimación de costos operativos anuales (OPEX)	83
Tabla 17. Resumen del análisis costo-beneficio	85
Tabla 18. Flujo de caja	86

CAPÍTULO I: INTRODUCCIÓN

1.1 Objetivos del informe

1.1.1. Objetivo general

Presentar y documentar de manera clara, ordenada y estructurada el desarrollo del Trabajo de Suficiencia Profesional; referido al diseño, implementación y validación de un servicio digital de valor agregado para la identificación de clientes del segmento Pyme en una institución financiera, en el marco de una arquitectura Big Data distribuida.

1.1.2. Objetivos específicos

- Describir el contexto y los fundamentos técnicos que sustentan el desarrollo del trabajo en el proyecto, así como el rol desempeñado.
- Exponer la problemática abordada y su relevancia dentro del entorno institucional.
- Detallar el enfoque metodológico, así como las herramientas y tecnologías empleadas para el desarrollo del servicio.
- Explicar el proceso de diseño, implementación y validación del servicio digital dentro de una arquitectura Big Data distribuida.
- Presentar de manera estructurada los resultados obtenidos y su validación técnica respectiva.
- Sistematizar la experiencia profesional desarrollada, destacando su aplicabilidad y aporte en el ámbito empresarial.

1.2 Estructura del informe

El presente Trabajo de Suficiencia Profesional describe el desarrollo de un servicio digital de valor agregado orientado a la identificación de clientes del segmento Pyme, bajo una arquitectura Big Data distribuida dentro de una institución financiera. La elección del tema responde a la necesidad de disponer de mecanismos confiables de clasificación de clientes que permitan mejorar la toma de decisiones, optimizar los procesos de segmentación y garantizar el cumplimiento de los requerimientos regulatorios en el sector financiero.

Adicionalmente, la relevancia del tema se sustenta en el aporte del segmento Pyme a la generación de empleo y al Producto Bruto Interno (PBI) del país, en ese sentido, su correcta identificación es un insumo crítico tanto para la institución como para el sistema financiero en su conjunto

El trabajo se sustenta en fundamentos técnicos relacionados con el procesamiento distribuido de datos, la integración de múltiples fuentes de información y el uso de arquitecturas escalables que garantizan la disponibilidad, trazabilidad y consistencia de los datos. Asimismo, se enmarca en el enfoque de servicios digitales de valor agregado, los cuales permiten transformar la infraestructura tecnológica en una plataforma orientada a la generación de valor para el negocio.

Para el desarrollo de la solución se emplea PySpark para el procesamiento distribuido de datos, HDFS para el almacenamiento escalable y Jira para la gestión del ciclo de vida del proyecto, estas herramientas permiten asegurar la trazabilidad del proceso de implementación y facilitan la construcción de un servicio digital integrado dentro del ecosistema de datos institucional.

El informe se organiza de la siguiente manera:

- **Capítulo I: Introducción.** Presenta el contexto general del trabajo, los objetivos y su estructura del informe.

- **Capítulo II: Información del lugar donde se desarrolló la actividad.** Describe la institución, sus características, el periodo de desarrollo y datos generales.
- **Capítulo III: Descripción de la actividad.** Expone la organización del trabajo, la finalidad, los objetivos de la actividad, la problemática, la metodología empleada, el procedimiento seguido y los resultados obtenidos.
- **Capítulo IV: Discusión de resultados.** Presenta el análisis de la implementación del servicio digital, su desempeño técnico y la evaluación económica correspondiente.
- **Capítulo V: Conclusiones.** Sintetiza los hallazgos mediante las conclusiones obtenidas del trabajo realizado.
- **Capítulo VI: Recomendaciones.** Propone acciones de mejora, escalamiento y reutilización del servicio digital implementado.
- **Capítulo VII: Referencias.** Detalla las fuentes bibliográficas utilizadas.
- **Capítulo VIII: Anexos.** Contiene información complementaria que respalda el desarrollo del trabajo.

CAPÍTULO II: INFORMACIÓN DEL LUGAR DONDE SE DESARROLLÓ LA ACTIVIDAD

2.1 Institución – Actividad que desarrolla

En el presente Trabajo de Suficiencia Profesional (TSP), se desarrollan actividades de consultoría profesional en una institución financiera del sector bancario, en las que se hace uso de herramientas de Big Data para el tratamiento de grandes volúmenes de información.

El trabajo fue desarrollado en la empresa Management Solutions, firma internacional de consultoría especializada en asesoramiento en negocios, finanzas, riesgos, procesos, y tecnología.

2.2 Periodo de duración de la actividad

La actividad se desarrolla durante el periodo comprendido entre enero y mayo de 2024, con una duración total de 5 meses.

2.3 Finalidad y objetivos de la entidad

La finalidad de la empresa Management Solutions es generar propuestas de valor para sus clientes comprometiéndose con su implantación efectiva. Su principal objetivo es superar las expectativas de sus clientes y convertirse en su socio de confianza.

Asimismo, su cultura corporativa está basada en el esfuerzo, el compromiso, la búsqueda de la excelencia, la orientación a resultados y el estricto cumplimiento de principios éticos que son la base del modelo de negocio.

Adicionalmente, la empresa cuenta con amplia experiencia en la aplicación de tecnologías de Big Data, alineándose a las necesidades específicas de sus clientes, potenciando, gestionando y optimizando la información para generar mayor valor e impacto en el modelo de negocio.

2.4 Razón social

MANAGEMENT SOLUTIONS PERU S.A.C.

2.5 Dirección postal

Calle Dean Valdivia 148-158, Edificio Centro Empresarial Platinum Plaza – Torre A, San Isidro Lima, Perú.

2.6 Nombre y correo electrónico del profesional a cargo

Sergio Blanco Gutiérrez

sergio.blanco.gutierrez@ms-peru.com

Juan José Aspilcueta

juan.jose.aspilcueta@ms-peru.com

CAPÍTULO III: DESCRIPCIÓN DE LA ACTIVIDAD

3.1 Organización de la actividad

La actividad se desarrolla bajo la participación de los equipos que se muestran en la siguiente figura.

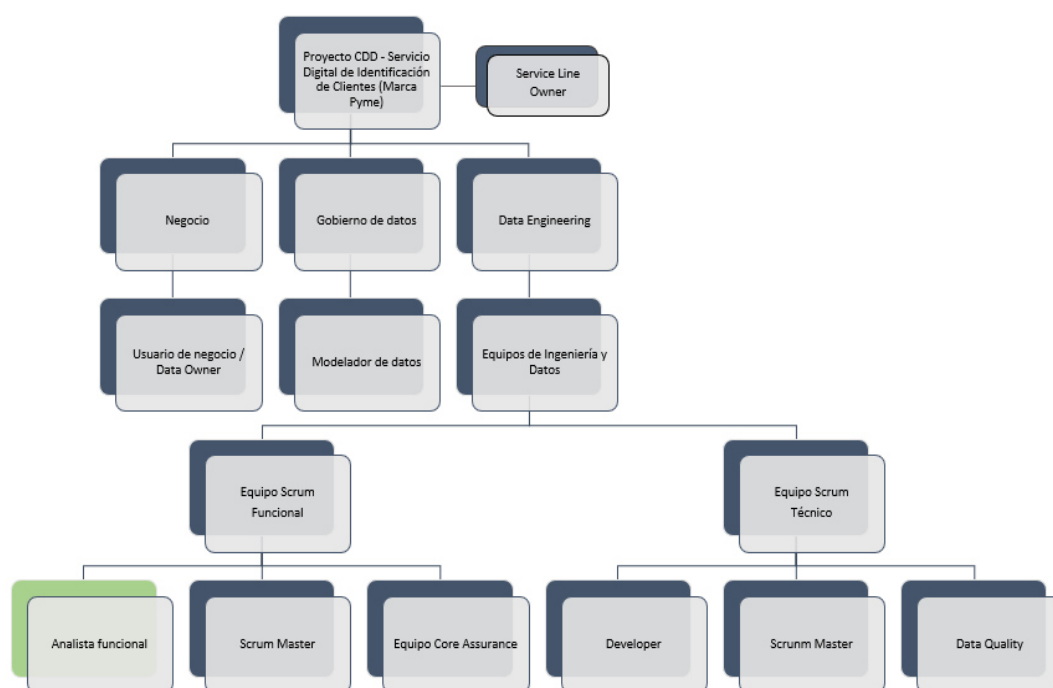


Figura 1 Organigrama de equipos participantes. Fuente. Elaboración propia.

A continuación, se describen las áreas y roles participantes en el proceso de diseño, implementación y puesta en operación del servicio digital de identificación de clientes.

- Encargado del proyecto Service Line Owner: Responsable de garantizar la factibilidad del proyecto y dueño del servicio, su función es garantizar que

se cumplan los objetivos del negocio, conectándolo con las áreas de tecnología y datos.

- **Negocio:** Área responsable de definir la necesidad funcional del servicio digital. En este caso, el usuario de negocio solicita la habilitación de un identificador que permita clasificar a los clientes según su condición como Pequeña y Mediana Empresa (Pyme), de acuerdo con requerimientos normativos y de gestión interna. Esta área está compuesta por:
 - **Usuario de Negocio / Data Owner:** Rol encargado de validar los criterios funcionales del indicador y definir la interpretación del dato. Asimismo, aprueba la lógica de clasificación y brinda el visto bueno funcional para su implementación y uso institucional.
- **Gobierno de Datos:** Área responsable de velar por la correcta definición, estandarización y reutilización del dato dentro de la organización. Participa en la validación del modelo de datos y asegura que el indicador cumpla con los lineamientos de gobierno y calidad establecidos.
 - **Modelador de Datos:** Encargado de revisar la consistencia del indicador dentro del modelo lógico de datos institucional, validando que las fuentes utilizadas, el formato del campo y su documentación permitan su mantenimiento y evolución futura.
- **Datahub – Equipo de Ingeniería y Datos:** Área técnica encargada de la implementación del servicio digital sobre la plataforma Big Data. Centraliza el procesamiento distribuido, almacenamiento y gestión técnica del indicador dentro de la arquitectura de datos institucional. Los equipos en esta área son:
 - **Equipo Scrum Funcional:** Responsable del análisis funcional, levantamiento de requerimientos, definición de la lógica del indicador y ejecución de simulaciones en entornos controlados (sandbox). Este equipo también se encarga de la elaboración de la documentación funcional y de la coordinación con las áreas de negocio y gobierno de datos. Los perfiles más importantes son:
 - **Analista Funcional:** Rol responsable del levantamiento y análisis de los requerimientos del negocio, así como de la definición funcional del indicador de identificación de clientes. Participa en la identificación y validación de las fuentes de datos, en la

construcción de la lógica funcional del indicador y en la ejecución de simulaciones en entornos sandbox utilizando herramientas de procesamiento distribuido. Asimismo, se encarga de la elaboración de la documentación funcional-técnica (dictamen) y de la coordinación con las áreas de negocio, gobierno de datos y el equipo técnico para la validación del servicio digital.

- **Scrum Master (Equipo Funcional):** Rol encargado de facilitar la correcta aplicación de la metodología Scrum dentro del equipo funcional, gestionando el seguimiento de tareas, la resolución de bloqueos y la comunicación con los demás equipos involucrados, asegurando el cumplimiento de los objetivos definidos para cada iteración.
- **Equipo Core Assurance:** Área encargada de revisar y validar el modelado de datos de los diferentes proyectos en la institución. Revisa la documentación mediante lineamientos que garantizan la confiabilidad y valida las fuentes de información propuestas, asegurando que la definición del dato sea consistente con los estándares institucionales, permitiendo su uso transversal por otros equipos.
- **Equipo Scrum Técnico:** Responsable del desarrollo técnico del indicador, implementando la lógica definida en la capa de aplicación mediante herramientas de procesamiento distribuido. Asimismo, gestiona el control de versiones, pruebas técnicas, preparación del componente para su puesta en producción y la correcta transferencia y comunicación de los datos. Los perfiles más importantes son:
 - **Developer - Ingeniero de Desarrollo:** Responsable de la implementación técnica del indicador en la capa de aplicación, traduciendo la lógica funcional definida en código ejecutable sobre la plataforma Big Data. Realiza pruebas técnicas, control de versiones y prepara los componentes necesarios para su posterior despliegue en los entornos productivos. Asimismo, asegura la correcta transferencia y comunicación de los datos.
 - **Scrum Master (Equipo Técnico):** Encargado de coordinar y facilitar las actividades del equipo técnico bajo la metodología Scrum,

asegurando la correcta planificación, ejecución y entrega de las tareas de desarrollo, así como la comunicación con el equipo funcional.

- **Data Quality:** Área encargada de validar la calidad del procesamiento técnico, verificando la consistencia de los datos generados y asegurando que la implementación cumpla con los criterios de calidad antes de su despliegue en el entorno productivo.

3.2 Finalidad y objetivos de la Actividad

3.2.1 Finalidad

Contribuir a la mejora de los procesos de diseño, implementación y validación de indicadores concebidos como servicios digitales, dentro de arquitecturas Big Data distribuidas, mediante la aplicación de prácticas técnicas estandarizadas que garanticen su correcta integración con la infraestructura de comunicaciones, así como su trazabilidad, consistencia y reutilización como fuentes confiables de información.

En este marco, el trabajo se orienta al diseño e implementación de un servicio digital de valor agregado para la identificación de clientes del segmento Pyme, desarrollado sobre una arquitectura Big Data distribuida, con el propósito de fortalecer la disponibilidad de información consistente y oportuna para la toma de decisiones organizacionales.

3.2.2 Objetivos

3.2.2.1 Objetivo General

Diseñar e implementar un servicio digital de valor agregado para la identificación de clientes basado en una arquitectura Big Data distribuida.

3.2.2.2 *Objetivos Específicos*

1. Diagnosticar y caracterizar las fuentes de información de datos institucionales y los requerimientos funcionales necesarios para la identificación del cliente del segmento Pyme, asegurando la calidad y relevancia de los datos disponibles.
2. Diseñar la arquitectura lógica y el modelo de procesamiento del indicador, definiendo el flujo de datos dentro de la infraestructura Big Data distribuida.
3. Desarrollar el servicio digital de clasificación, integrando la lógica de negocio dentro del ecosistema Big Data institucional.
4. Validar la integridad y consistencia técnica del indicador en un entorno de pruebas (sandbox), comparando los resultados obtenidos con métricas de negocio para garantizar la precisión de la segmentación.
5. Implementar el servicio en el entorno productivo, asegurando su correcta integración en la arquitectura Big Data distribuida y la disponibilidad del indicador para los sistemas consumidores de la institución.
6. Documentar y formalizar el indicador de clasificación de clientes, garantizando la trazabilidad de los procesos, el mantenimiento evolutivo y la reutilización del servicio en futuros proyectos.

3.3 Problemática

En las instituciones financieras modernas, la gestión, procesamiento y explotación de grandes volúmenes de datos se apoya en arquitecturas Big Data distribuidas, las cuales integran componentes de almacenamiento, procesamiento y comunicación para atender las necesidades analíticas y operativas del negocio. Estas arquitecturas requieren capacidades de cómputo escalables y mecanismos eficientes de procesamiento distribuido que permitan la integración entre los distintos sistemas y la provisión oportuna de información a las áreas usuarias.

En este contexto, las organizaciones buscan ir más allá de la simple

conectividad, orientándose al diseño de servicios digitales de valor agregado que transformen los datos disponibles en información útil para la toma de decisiones. En el sector financiero, uno de estos servicios se materializa en la construcción de indicadores, los cuales permiten identificar, clasificar y caracterizar a los clientes a partir de múltiples fuentes de información distribuidas en el Data Lake corporativo.

Un caso representativo es la identificación de clientes pertenecientes al segmento de Pequeñas y Medianas Empresas (Pyme), cuya correcta clasificación resulta clave tanto para el cumplimiento normativo como para el diseño de productos financieros, campañas y estrategias de atención. Desde el punto de vista técnico, este requerimiento implica la construcción de un indicador confiable, capaz de integrar datos provenientes de diversas fuentes, ser validado adecuadamente y mantenerse disponible para su uso transversal dentro de la organización. Asimismo, se debe garantizar la confiabilidad, disponibilidad y reutilización de este indicador dentro de la institución.

Durante el desarrollo de este tipo de servicios digitales basados en datos, se evidencian diversos desafíos técnicos, tales como la correcta identificación y análisis de las fuentes de información, la definición de una lógica de procesamiento consistente, la validación del resultado en entornos controlados y la formalización del indicador para garantizar su trazabilidad, mantenimiento evolutivo y reutilización. De igual forma, la integración del servicio dentro de una arquitectura Big Data distribuida exige asegurar que los procesos de procesamiento y disponibilización de la información se encuentren adecuadamente articulados.

Desde una perspectiva de telecomunicaciones, los servicios de valor agregado se ubican en los niveles superiores del modelo de referencia OSI, apoyándose en las capas inferiores responsables del transporte y la transmisión de los datos. Esta visión facilita comprender cómo el servicio digital se sustenta sobre la infraestructura tecnológica de la institución y cómo esta habilita su correcta disponibilidad y consumo por los sistemas.

La problemática abordada en el presente Trabajo de Suficiencia

Profesional se centra, por tanto, en el diseño, implementación y validación técnica de un servicio digital de valor agregado para la identificación de clientes, desarrollado sobre una arquitectura Big Data distribuida.

Para ello, se emplean herramientas de procesamiento y gestión de datos como PySpark, HDFS y Jira, que permiten construir el indicador a partir de múltiples fuentes, validar su correcto funcionamiento en entornos controlados y formalizar su integración dentro de la plataforma de datos, asegurando su confiabilidad, trazabilidad, reutilización y la provisión de información consistente y oportuna para la toma de decisiones organizacionales.

3.3.1 Problema General

¿Cómo diseñar e implementar un servicio digital de valor agregado que permita la identificación de clientes, utilizando una arquitectura Big Data distribuida?

3.3.2 Problemas Específicos

1. ¿Cómo identificar y caracterizar las fuentes de datos institucionales y los requerimientos funcionales necesarios para construir un servicio digital de identificación de clientes del segmento Pyme, asegurando la calidad y relevancia de la información disponible?
2. ¿Cómo diseñar la arquitectura lógica y el modelo de procesamiento de un indicador de identificación de clientes, definiendo adecuadamente el flujo de datos dentro de una infraestructura Big Data distribuida?
3. ¿Cómo desarrollar un servicio digital de clasificación de clientes que integre la lógica de negocio dentro del ecosistema Big Data institucional de forma consistente y reutilizable?
4. ¿Cómo validar la integridad y consistencia técnica del indicador de identificación de clientes en un entorno de pruebas (sandbox), asegurando que los resultados obtenidos sean precisos y coherentes con las métricas de negocio definidas?

5. ¿Cómo desplegar el servicio digital de identificación de clientes en el entorno productivo, garantizando su correcta integración en la arquitectura Big Data distribuida y su disponibilidad para los sistemas consumidores de la institución?
6. ¿Cómo documentar y formalizar el indicador de clasificación de clientes de manera que se asegure la trazabilidad de los procesos, el mantenimiento evolutivo y la reutilización del servicio en futuros proyectos?

3.3.3 Justificación e importancia de la investigación

Justificación social

En el Perú, las Pequeñas y Medianas Empresas (Pyme) constituyen aproximadamente el 99.4 % del tejido empresarial, aportan cerca del 30 % del Producto Bruto Interno (PBI) y generan alrededor del 60 % del empleo formal, consolidándose como uno de los principales motores de la economía nacional. Este segmento, además, representa una proporción significativa de la Población Económicamente Activa, lo que evidencia su impacto no solo económico sino también social, al contribuir a la generación de ingresos y a la reducción de la pobreza. (Ministerio de la Producción, 2022).

Sin embargo, las Pyme enfrentan importantes desafíos en términos de formalización e inclusión financiera. Una gran parte opera fuera del sistema financiero formal, lo que limita su acceso a financiamiento, crecimiento y sostenibilidad. En este contexto, la correcta identificación y clasificación de los clientes pertenecientes al segmento Pyme dentro de las instituciones financieras constituye un elemento clave para diseñar estrategias orientadas a mejorar su acceso a productos financieros y fortalecer su integración al sistema económico.

El desarrollo del presente Trabajo de Suficiencia Profesional contribuye a este propósito, al proponer un servicio digital que permite identificar de manera estructurada y confiable a los clientes Pyme dentro de la plataforma de datos institucional, facilitando así la toma de decisiones orientadas a una mayor inclusión financiera.

Justificación práctica

Desde el punto de vista práctico, las instituciones financieras requieren mecanismos eficientes que les permitan segmentar adecuadamente a sus clientes, optimizar la asignación de recursos y diseñar productos diferenciados según el perfil del usuario.

La implementación de un servicio digital de valor agregado para la identificación de clientes Pyme permite mejorar los procesos internos de clasificación y segmentación, reduciendo ambigüedades en la información y fortaleciendo la consistencia de los datos utilizados por distintas áreas de negocio. Esto impacta directamente en la eficiencia operativa, en la mejora de campañas comerciales y en el cumplimiento de requerimientos regulatorios que se demanden en la institución.

Asimismo, al estar basado en una arquitectura Big Data distribuida, el servicio asegura disponibilidad, escalabilidad y acceso transversal, lo que amplía su aplicabilidad dentro de la organización y favorece su reutilización en futuros proyectos.

Justificación técnica

Desde una perspectiva técnica, la identificación del segmento Pyme requiere la construcción de un indicador confiable integrado dentro de una arquitectura Big Data distribuida y su disponibilidad en línea para las diferentes áreas usuarias. Dicho indicador debe diseñarse considerando aspectos de almacenamiento, procesamiento distribuido, validación técnica y formalización, asegurando su correcta integración con la infraestructura institucional.

El uso de herramientas como PySpark para el procesamiento distribuido, HDFS para el almacenamiento escalable y Jira para la gestión y trazabilidad del flujo técnico, permite estructurar un proceso estandarizado de diseño, implementación y validación del servicio digital. Este enfoque fortalece la confiabilidad del indicador, mejora la consistencia de los datos y garantiza su disponibilidad para los sistemas y usuarios consumidores.

Además, el trabajo visibiliza una sistematización del proceso de construcción de un servicio de valor añadido dentro de una arquitectura Big Data, integrando procesamiento de datos, infraestructura distribuida y soporte en un mismo esquema de implementación.

3.4 Metodología

3.4.1 Bases teóricas

Para la elaboración del presente Trabajo de Suficiencia Profesional se revisaron investigaciones y publicaciones académicas que abordan, desde distintas perspectivas, la aplicación de arquitecturas Big Data en el sector financiero, la gestión de información distribuida y la provisión de servicios digitales sobre redes de comunicaciones. A continuación, se describen los trabajos considerados como antecedentes relevantes.

- Huapaya (2021), analiza cómo la centralización de la información en un entorno Big Data permite optimizar los productos financieros y la toma de decisiones en una entidad bancaria. El autor plantea que la existencia de bases de datos aisladas por cada área de negocio incrementa el consumo de recursos informáticos y ralentiza los procesos, problema que se intensifica en entidades transnacionales. Como respuesta, el área de Arquitectura de Datos implementó una estrategia de centralización en un Data Lake con ingesta mediante procesamiento distribuido en Spark, acompañada de una capa de consumo basada en sandboxes para analítica avanzada y reportería. Los resultados reportados evidencian gobierno de datos, flujos formalizados y mejores tiempos de respuesta. Este trabajo constituye un antecedente directo para el presente TSP, dado que comparte el uso de Data Lake y procesamiento distribuido en Spark como plataforma para habilitar servicios de información en la banca.
- Tapia (2018), analiza cómo las técnicas de Big Data asociadas al tratamiento y análisis de grandes volúmenes de información han

modificado la gestión empresarial, proveyendo a las organizaciones de una herramienta para tomar decisiones ágiles en contextos de incertidumbre. La investigación concluye que la adopción de Big Data impacta favorablemente en los productos de financiamiento dirigidos al segmento Pyme. Este antecedente aporta sustento al presente TSP al vincular el uso de Big Data con la atención específica del segmento Pyme en el sistema financiero peruano, que es precisamente el segmento sobre el cual se construye el indicador diseñado en este trabajo.

- Martínez (2018), sostiene que las entidades financieras tradicionales poseen grandes volúmenes de información provenientes de múltiples fuentes, pero que su complejidad estructural y limitaciones tecnológicas han dificultado su explotación para la generación de valor. Esta brecha ha facilitado la aparición de las fintech, que emplean tecnologías innovadoras y modelos ágiles para ofrecer servicios financieros más eficientes y personalizados. Frente a ello, el autor propone la adopción de un modelo basado en tecnologías Big Data que permita a la banca tradicional gestionar, procesar y analizar grandes volúmenes de datos para mejorar la toma de decisiones y desarrollar productos ajustados al cliente. La relevancia para el presente TSP radica en que el diseño de indicadores como Marca Pyme, contruidos sobre una arquitectura Big Data distribuida, es precisamente uno de los mecanismos a través de los cuales la banca tradicional traduce sus datos en ventajas competitivas sostenibles.
- Keshavarz et al. (2021), analizan cómo la analítica de Big Data se consolida como un pilar estratégico dentro de la industria de las telecomunicaciones, con impacto directo en el desempeño de las organizaciones del sector. A través de un enfoque cualitativo basado en entrevistas semiestructuradas y análisis documental, los autores identifican cinco pilares y un conjunto de subdimensiones que explican cómo las empresas de telecomunicaciones pueden construir ventajas competitivas a partir del aprovechamiento de los datos que fluyen sobre su infraestructura de red. Los autores sostienen que la evolución del sector exige trascender la provisión de conectividad para desarrollar

servicios digitales apalancados en analítica distribuida, apoyados en redes de datos confiables y escalables. Este antecedente es pertinente para el presente TSP porque provee un marco conceptual sobre cómo un servicio digital construido sobre una arquitectura Big Data, como el indicador de identificación de clientes Pyme, constituye un servicio de valor agregado habilitado por la red de telecomunicaciones, en línea con la tendencia internacional del sector.

3.4.2 Marco conceptual

3.4.2.1 Big Data

El término *Big Data* surge en el ámbito de la visualización científica con el trabajo de Cox y Ellsworth (1997), quienes lo introdujeron al identificar un problema recurrente en sus experimentos: la imposibilidad de procesar conjuntos de datos cuya magnitud excedía tanto la memoria principal como el almacenamiento local de las estaciones de trabajo gráficas. Los autores denominaron a este desafío "*el problema de Big Data*" y advirtieron que apoyarse exclusivamente en la memoria virtual del sistema operativo conducía a un rendimiento muy pobre, debido a fenómenos como el *thrashing*, es decir, el intercambio excesivo de páginas que ralentiza el sistema. Con esta caracterización inicial se instaló en la literatura técnica la idea de que el aumento sostenido del volumen de información demandaba soluciones tecnológicas distintas a las disponibles hasta entonces.

Años más tarde, Laney (2001), en un informe publicado para META Group, consolidó un modelo de gestión de datos basado en tres dimensiones críticas, conocido posteriormente como el modelo de las 3V:

- Volumen: cantidad masiva de datos que crece de manera sostenida dentro de las organizaciones.
- Velocidad: ritmo al que se generan, procesan y transmiten los datos.
- Variedad: diversidad de tipos y formatos de información, que incluyen datos estructurados, semiestructurados y no estructurados.

El aporte central de Laney (2001) radica en evidenciar que los desafíos de la

gestión de datos no dependen únicamente de la magnitud del almacenamiento, sino también de la capacidad para integrar fuentes heterogéneas y responder en tiempos adecuados a las necesidades del negocio. De esta manera, cuando una de estas dimensiones no se gestiona de forma integral, los datos pierden utilidad, llegando tarde en formatos difíciles de procesar o con información inconexa. Por el contrario, una administración eficiente de las 3V permite transformar los datos en un recurso estratégico que incrementa la agilidad y precisión en la toma de decisiones organizacionales.

Estas definiciones constituyen la base teórica sobre la cual evolucionaron las arquitecturas y tecnologías de Big Data actualmente vigentes, tales como Apache Hadoop para almacenamiento y procesamiento distribuido, Apache Spark y su API PySpark para cómputo paralelo en memoria, y los ecosistemas de Data Lake para integración de fuentes heterogéneas. En el contexto del presente TSP, el modelo de las 3V sustenta la elección del entorno tecnológico en el que se diseña e implementa el indicador de identificación de clientes Pyme: un volumen considerable de registros provenientes de múltiples fuentes del Datalake corporativo, actualizados con periodicidad mensual, que deben procesarse de manera confiable para habilitar un servicio digital de valor agregado sobre la red de datos de la organización.

3.4.2.2 *Data Warehouse*

Un *Data Warehouse* es un repositorio centralizado concebido para almacenar grandes volúmenes de información histórica y apoyar los procesos de toma de decisiones en las organizaciones. De acuerdo con Inmon (2005), este tipo de repositorio se caracteriza por ser orientado a un tema de negocio, por integrar información proveniente de fuentes heterogéneas, por ser variante en el tiempo (conservando el histórico de los cambios) y no volátil; entendido esto último como la imposibilidad de modificar arbitrariamente los datos una vez cargados. Esta definición ha constituido durante décadas la base conceptual sobre la cual se construyeron las plataformas analíticas tradicionales del sector financiero.

Bani et al. (2018) complementan esta caracterización al señalar que el Data Warehouse se apoya en tecnologías de multiprocesamiento y procesamiento masivo para habilitar consultas analíticas sobre grandes volúmenes de datos históricos, convirtiéndose en un habilitador clave para la identificación de tendencias y la generación de reportería gerencial.

La carga de información en un Data Warehouse, se estructura bajo el proceso conocido como ETL (*Extract, Transform, Load*), mediante el cual los datos son primero extraídos de las fuentes operativas, luego transformados para adecuarlos al modelo analítico definido y finalmente cargados en el repositorio centralizado. En cuanto al modelado de datos, una de las técnicas más utilizadas es el modelado dimensional, cuyo esquema más representativo es el modelo estrella. Santos y Yulia (2017) describen este modelo como una estructura compuesta por una tabla central denominada tabla de hechos, que se encarga de concentrar las métricas cuantitativas sobre las cuales se realiza el análisis, y está rodeada de tablas de dimensiones que contienen los atributos descriptivos asociados a dichas métricas. Los autores destacan que este diseño permite soportar consultas analíticas de alto nivel y facilita la exploración multidimensional de la información.

3.4.2.3 Datalake

En el año 2010, James Dixon, entonces CTO de Pentaho, acuñó el término Data Lake como respuesta al enfoque tradicional de los Data Marts. Dixon utilizó una analogía que hoy se ha vuelto clásica: mientras un Data Mart puede entenderse como una botella de agua limpia, embotellada y lista para un propósito específico, un Data Lake en cambio es un cuerpo de agua natural en el que coexisten múltiples flujos de información en su estado original, disponibles para diversos usos analíticos posteriores (Dixon, 2010).

A diferencia del Data Warehouse, que exige un modelado previo de la información (*schema on write*), el Data Lake aplica un esquema solo al momento de consultar los datos (*schema on read*). Esta característica le permite almacenar información en su formato nativo y soportar tanto datos

estructurados como semiestructurados y no estructurados; lo cual resulta especialmente relevante en escenarios de Big Data (Dixon, 2010; Janelidze, Aptsiauri y Tsitashvili, 2024).

En la Tabla 1 se presenta una síntesis comparativa entre ambos enfoques de repositorio analítico.

Tabla 1. Comparación entre Data Warehouse y Data Lake

Criterio	Data Warehouse	Data Lake
Tipo de datos	Estructurados	Estructurados, semiestructurados y no estructurados
Esquema	Schema on write (definido antes de cargar)	Schema on read (definido al consultar)
Costo	Alto (infraestructura y procesamiento)	Más bajo (almacenamiento en bruto)
Flexibilidad	Limitada (modelado previo)	Alta (uso variado y exploración)
Propósito principal	Análisis empresarial y reportes	Almacenamiento masivo y analítica avanzada

Fuente. Elaboración propia a partir de Dixon (2010) y Janelidze et al. (2024).

Dixon (2010) destacó que los Data Lakes resultan especialmente pertinentes en entornos de Big Data, dado que permiten gestionar grandes volúmenes de información a un costo relativamente bajo, apalancándose en infraestructuras distribuidas como Hadoop. En la Figura 2 se muestra gráficamente esta diferencia de enfoques.

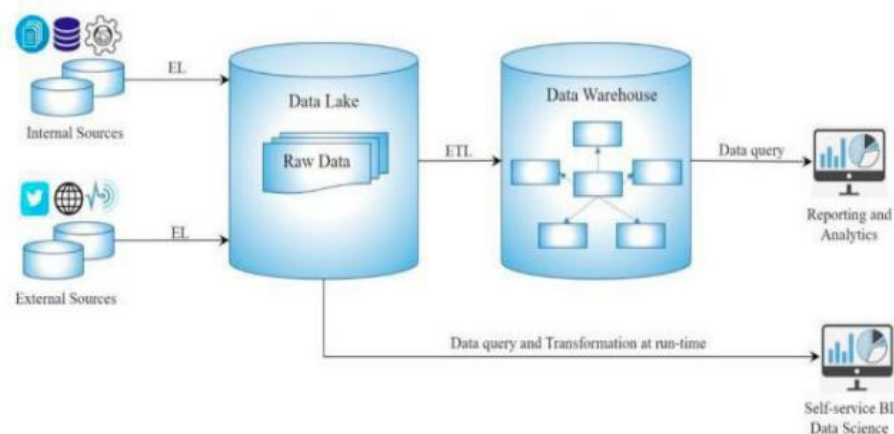


Figura 2 Comparación entre enfoques de repositorios de Datos. Fuente: Llave, M. (2018).

Nota. La imagen contrasta la lógica de un Data Warehouse, orientado a datos procesados y curados, con la de un Data Lake, que conserva la diversidad y el estado original de las fuentes.

Características de un Data Lake

Las principales características funcionales de un Data Lake, de acuerdo con la literatura especializada, son las siguientes:

- **Almacenamiento masivo y flexible.** Permite almacenar grandes volúmenes de datos estructurados y no estructurados en un único repositorio sin requerir un esquema definido previamente (Janelidze et al., 2024).
- **Escalabilidad.** Se adapta al crecimiento de los datos mediante tecnologías distribuidas como Hadoop y Spark, que permiten añadir recursos de manera progresiva sin comprometer el rendimiento (Rao et al., 2021).
- **Integración de datos heterogéneos.** Admite múltiples formatos como texto, imágenes, video, logs transaccionales, datos de sensores IoT, entre otros, lo que facilita el análisis combinado de fuentes diversas (Janelidze et al., 2024).
- **Procesamiento distribuido.** Incorpora motores de cómputo como Apache Spark, que permiten realizar analítica avanzada sobre grandes volúmenes de datos en ventanas de tiempo acotadas (Rao et al., 2021).
- **Gestión de metadatos.** Incluye mecanismos para catalogar y organizar los datos, aspecto que constituye a la vez uno de sus mayores desafíos cuando la gobernanza no está adecuadamente definida (Janelidze et al., 2024).
- **Acceso flexible.** Soporta consultas desde distintos lenguajes y herramientas analíticas, lo que habilita su uso por perfiles técnicos y de negocio (Rao et al., 2021).

Beneficios de un Data Lake

- **Almacenamiento económico y escalable,** apoyado en tecnologías distribuidas y almacenamiento en la nube (Rao et al., 2021).

- **Flexibilidad en el análisis**, al no exigir un esquema predefinido, lo que facilita la exploración y el uso para casos de negocio emergentes (Janelidze et al., 2024).
- **Soporte a múltiples fuentes de datos**, integrando información estructurada, semiestructurada y no estructurada, lo que potencia la analítica avanzada y el aprendizaje automático (Madera y Laurent, 2016).
- **Impulso a la innovación analítica**, al centralizar datos de diversas áreas y habilitar descubrimientos y modelos predictivos (Rao et al., 2021).

Desventajas de un Data Lake

- **Riesgo de *Data Swamp***. En ausencia de una gobernanza adecuada y una gestión rigurosa de metadatos, el Data Lake puede degradarse en un repositorio caótico y de bajo valor (Janelidze et al., 2024).
- **Complejidad en la gobernanza**. Exige estrategias sólidas de seguridad, calidad y control de acceso (Madera y Laurent, 2016).
- **Curva de aprendizaje tecnológica**. El uso de herramientas como Spark, Hive o Presto demanda perfiles especializados, lo que puede limitar su adopción inicial (Rao et al., 2021).
- **Costos ocultos**. Aunque el almacenamiento resulta económico, la gestión, el procesamiento avanzado y la administración generan costos operativos significativos (Janelidze et al., 2024).

3.4.2.4 Formatos de archivo y capas de aprovisionamiento en HDFS

En un entorno Big Data, la elección del formato de archivo constituye una decisión de arquitectura con impacto directo en el rendimiento del almacenamiento y de las cargas analíticas. Esto debido a que no todos los formatos son equivalentes, algunos están diseñados para procesos de ingesta e intercambio rápido entre sistemas, mientras que otros están optimizados para lecturas columnares y analítica a gran escala (Apache Software Foundation, 2024). El Hadoop Distributed File System (HDFS) admite múltiples formatos, pero en la práctica de la mayoría de proyectos de Big Data sobre Hadoop dos de ellos concentran el uso: Avro y Parquet.

Formatos de archivo

- **Avro.** Según la Apache Software Foundation (2022), Avro es un sistema de serialización de datos que ofrece una integración directa con lenguajes dinámicos. Se basa en esquemas descritos en formato JSON, los cuales acompañan siempre a los datos y describen su estructura, lo que permite procesarlos sin necesidad de generar código adicional. Avro aplica compresión por filas, lo que lo vuelve particularmente eficiente para procesos de ingesta, almacenamiento histórico y transferencia entre sistemas.
- **Parquet.** Es un formato columnar de almacenamiento, también perteneciente al ecosistema Hadoop. Soporta esquemas de compresión y codificación de datos que impactan positivamente en el rendimiento de consultas analíticas. Su diseño separa los metadatos de los datos y permite que las columnas se distribuyan en varios archivos o que un único archivo referencie a múltiples archivos Parquet, lo que facilita la lectura selectiva de columnas y optimiza el consumo de recursos en cargas de analítica avanzada (Apache Software Foundation, 2024). En la Figura 3 se muestra la estructura interna de un archivo Parquet.

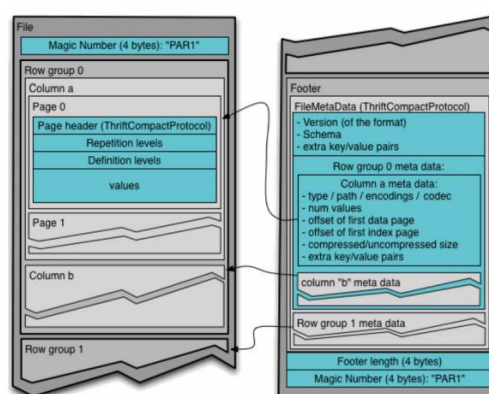


Figura 3 Estructura interna de un archivo Parquet. Fuente: Apache Software Foundation (2022)

Nota. Se observan los bloques de fila (row groups), los grupos de columnas (column chunks) y los metadatos asociados a cada nivel. La convivencia de Avro y Parquet dentro del mismo Data Lake no es redundante, sino complementaria: Avro opera en las capas orientadas al dato crudo y al

histórico, mientras que Parquet habilita el consumo analítico sobre datos ya depurados. Esta lógica se materializa en la estructura por capas del Data Lake corporativo.

Capas de aprovisionamiento en HDFS:

El Data Lake utilizado en el presente trabajo se organiza en tres capas diferenciadas de aprovisionamiento, cada una con un propósito, un formato y un perfil de consumo específicos (Huapaya, 2021).

- **Staging.** Capa temporal que mantiene los archivos en su forma original directamente desde la fuente (archivos CSV, extracciones de tablas u otros formatos planos). Su temporalidad se gestiona al mínimo necesario por dos razones: (i) los archivos pueden contener información sensible que debe ser encriptada en capas posteriores, y (ii) los formatos planos ocupan un volumen de almacenamiento considerablemente mayor que los formatos especializados para Big Data, por lo que prolongar su permanencia degrada la eficiencia del clúster (Huapaya, 2021).
- **Raw.** Capa de datos crudos del entorno Big Data. Contiene archivos con compresión a nivel de filas en formato AVRO, lo que la hace eficiente para transformaciones posteriores, pero no idónea para consultas ni analítica avanzada. Esta capa también cumple el rol de almacenamiento histórico por su ahorro en disco y constituye el respaldo ante eventuales reprocesos de la capa siguiente (Huapaya, 2021).
- **Master.** Capa final del Data Lake. Almacena la información comprimida a nivel columnar en formato PARQUET, con el objetivo de soportar análisis exploratorios y el procesamiento de modelos analíticos. Es la capa que consumen directamente los *Data Scientists* desde sus respectivos *sandboxes*, y sobre la que se construyen los indicadores de negocio. (Huapaya, 2021).

En la Figura 4 se presenta la arquitectura por capas del Data Lake, articulando las tres capas descritas con el motor de procesamiento distribuido que actúa

transversalmente sobre ellas.

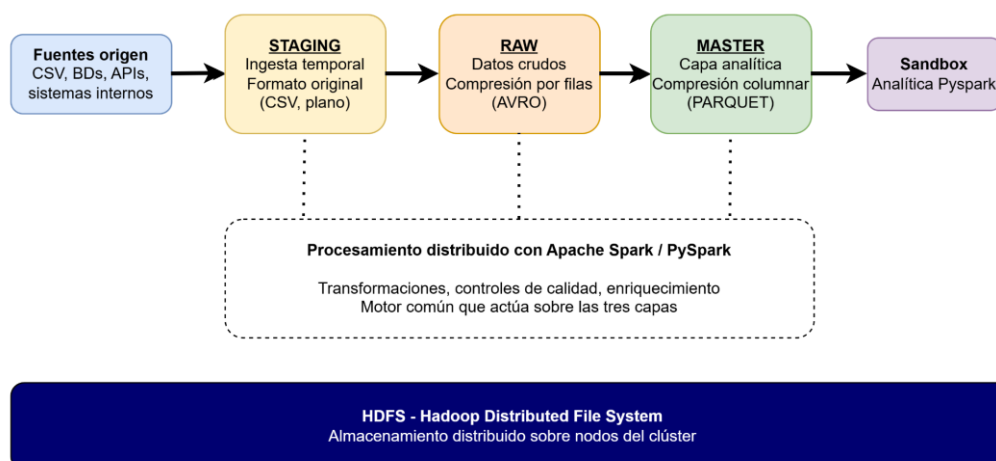


Figura 4: Arquitectura por capas del Data Lake corporativo sobre HDFS.
Fuente. Elaboración propia a partir de Huapaya(2021) y lineamientos internos del Data Lake corporativo

Nota. El diagrama representa las tres capas de aprovisionamiento del Data Lake desplegado sobre HDFS: Staging (ingesta temporal en formato original), Raw (capa histórica en AVRO) y Master (capa analítica en PARQUET). Sobre estas capas actúa Apache Spark / PySpark como motor de procesamiento distribuido, responsable de las transformaciones, controles de calidad y enriquecimiento entre capas.

Esta organización por capas permite que el proceso de construcción del indicador *Marca Pyme* pueda apoyarse en datos trazables, gobernados y en el formato óptimo para cada etapa del pipeline, en la ingesta desde fuentes heterogéneas, la conservación histórica del dato original y el consumo analítico en formato columnar listo para PySpark.

3.4.2.5 Apache Spark y PySpark

Apache Spark es un motor unificado de procesamiento distribuido de código abierto, diseñado para el análisis de datos a gran escala (Zaharia et al., 2016). Surgió como alternativa a Hadoop MapReduce para superar sus limitaciones en cargas iterativas y en la escritura continua a disco. Su principal diferenciador es el procesamiento en memoria (*in-memory computing*), que reduce significativamente los tiempos de ejecución y lo hace idóneo para flujos

analíticos sobre grandes volúmenes de información. Spark mantiene compatibilidad nativa con HDFS y con formatos como AVRO y PARQUET, lo que lo convierte en un componente natural de las arquitecturas Data Lake (Apache Software Foundation, 2024).

Spark opera bajo un modelo maestro–trabajador compuesto por tres elementos: el Driver, que contiene la lógica de la aplicación y coordina la ejecución; el Cluster Manager (YARN, Kubernetes o *stand-alone*), responsable de la asignación de recursos; y los Executors, procesos que corren en los nodos trabajadores y ejecutan las tareas asignadas. Sobre este esquema, Spark organiza la ejecución en un grafo acíclico dirigido (DAG) y aplica evaluación diferida: las transformaciones no se ejecutan hasta que una acción las dispara, lo que permite optimizar el plan físico antes de distribuirlo en el clúster (Karau y Warren, 2017). Su diseño modular —Spark SQL, MLlib, Structured Streaming y GraphX sobre Spark Core— permite cubrir en un mismo entorno transformaciones de datos, consultas analíticas, aprendizaje automático y procesamiento en flujo.

PySpark es la interfaz de Spark para el lenguaje Python, y constituye la vía de entrada más utilizada por perfiles de ingeniería y ciencia de datos. Su abstracción central es el DataFrame, una colección distribuida de datos organizada en columnas tipadas, conceptualmente equivalente a una tabla relacional, pero con capacidad de operar sobre grandes volúmenes de información distribuidos en múltiples nodos del clúster (Apache Software Foundation, 2024).

En el contexto del presente trabajo, PySpark constituye el motor sobre el cual se ejecutan las transformaciones que materializan el indicador Marca Pyme. Su uso resulta pertinente porque opera de forma nativa sobre los archivos PARQUET de la capa Master del Data Lake corporativo, soporta el volumen de información asociado a la cartera de clientes empresariales de la entidad financiera y permite expresar la lógica de clasificación del indicador como un conjunto de transformaciones encadenadas (filtros, joins, agrupamientos y reglas condicionales) que resultan auditables, mantenibles y reutilizables dentro del marco de gobierno de datos.

En la figura 5, se muestra el detalle de la distribución de la arquitectura de Apache Spark.

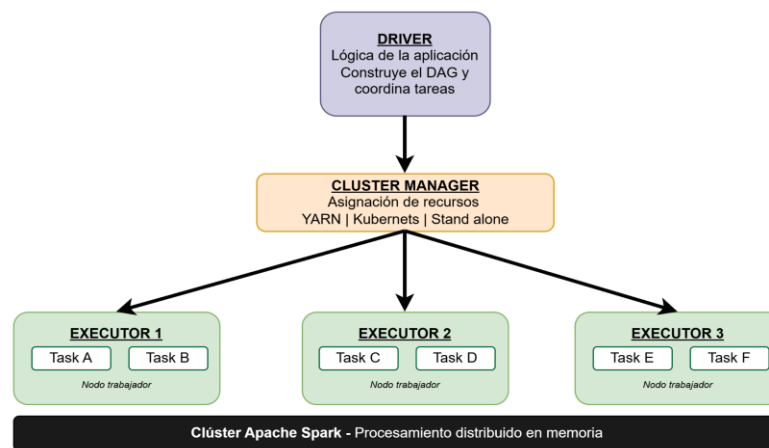


Figura 5 Arquitectura general de Apache Spark. Fuente: Elaboración propia a partir de Apache Software Foundation (2024) y Zaharia et al. (2016).

3.4.2.6 Data Quality

La calidad de los datos se define como el grado en que estos satisfacen el uso previsto. Según Han, Kamber y Pei (2012), esta calidad se articula en torno a seis dimensiones fundamentales: precisión, integridad, consistencia, puntualidad, interpretabilidad y credibilidad. La ausencia de una o varias de ellas compromete la validez de los análisis realizados y, por extensión, la calidad de las decisiones de negocio que se apoyan en esos datos.

En la práctica, la presencia de datos inexactos o incompletos en los repositorios corporativos es un fenómeno frecuente. Entre las causas más comunes se encuentran los errores humanos e informáticos durante el ingreso de la información, la omisión de datos por parte de los usuarios, el uso de instrumentos defectuosos para la recopilación, y la desactualización de registros; por ejemplo, cuando las operaciones se reportan fuera de los plazos establecidos (Han et al., 2012). A ello se suman factores relacionados con la interpretabilidad y la confianza: datos que, aun siendo técnicamente correctos, resultan difíciles de entender o cuya credibilidad es cuestionada por los usuarios, terminan siendo subutilizados.

Tabla 2. Dimensiones de la calidad de datos

Dimensión	Descripción
Precisión	Los datos reflejan correctamente el valor real del atributo que representan.
Integridad	Los datos están completos; no existen registros ni campos faltantes críticos.
Consistencia	Los datos son coherentes entre sí y entre distintas fuentes del repositorio.
Puntualidad	Los datos se encuentran disponibles y actualizados dentro del plazo requerido.
Interpretabilidad	Los datos pueden ser comprendidos y utilizados por los usuarios del negocio.

Fuente. Elaboración propia a partir de Han, Kamber y Pei (2012).

3.4.2.7 *Arquitectura distribuida y red de datos*

Arquitectura distribuida

Una arquitectura distribuida se caracteriza por la integración de múltiples componentes de hardware y software que, interconectados mediante una red, colaboran para ofrecer un servicio de cómputo único a los usuarios finales (Mell y Grance, 2011). Bajo este enfoque, los recursos de almacenamiento y procesamiento no residen en un único equipo, sino que se distribuyen entre varios nodos de un clúster, lo que habilita la escalabilidad, la disponibilidad y el procesamiento paralelo de grandes volúmenes de información.

Mell y Grance (2011), en la definición canónica del NIST sobre computación en la nube, señalan que los servicios digitales que operan sobre infraestructuras distribuidas dependen de redes de comunicaciones capaces de asegurar conectividad, disponibilidad y escalabilidad; de manera que los recursos de procesamiento y almacenamiento puedan ser accedidos de forma eficiente por múltiples usuarios y sistemas. En esta misma línea, Bani et al. (2018) demostraron en un caso aplicado al sector telecomunicaciones cómo el uso de arquitecturas de procesamiento masivamente paralelo (Massively Parallel Processing) sobre Data Warehouses permite procesar volúmenes crecientes de datos transaccionales, manteniendo tiempos de respuesta adecuados para el negocio. Ambos aportes evidencian que la infraestructura

distribuida es hoy un requisito funcional para cualquier servicio digital que opere sobre grandes volúmenes de información.

Modelo de referencia OSI y posicionamiento del servicio

El modelo OSI (Open Systems Interconnection), formalizado por la Organización Internacional de Normalización, constituye el marco de referencia más extendido para describir la operación de los sistemas de comunicaciones. Organiza la comunicación entre sistemas en siete capas jerárquicas: física, enlace de datos, red, transporte, sesión, presentación y aplicación; cada una con funciones y protocolos específicos (Tanenbaum y Wetherall, 2011; Stallings, 2014).

Los servicios digitales de valor agregado, como el desarrollado en el presente trabajo, se ubican en la capa de aplicación del modelo OSI. Desde allí hacen uso de las capas inferiores, principalmente las de transporte y red, para acceder a la información almacenada en la infraestructura distribuida y también para intercambiar datos entre los distintos componentes del sistema. Esta estratificación evidencia que los servicios de valor agregado no sustituyen a la infraestructura de telecomunicaciones, sino que se apoyan en ella: la red de datos corporativa actúa como habilitador del servicio; mientras que la capa de aplicación concentra la lógica de negocio que transforma la infraestructura en valor para la organización. En la Figura 6 se presenta el modelo OSI y el posicionamiento del servicio digital dentro de sus capas.



Figura 6 Modelo de referencia OSI y posicionamiento del servicio digital de valor agregado. Fuente. Elaboración propia a partir de Tanenbaum y Wetherall (2011) y Stallings (2014).

3.4.2.8 Servicios de Valor Agregado (SVA) en Redes de Comunicaciones

En el ámbito de las telecomunicaciones, los Servicios de Valor Agregado (SVA) se definen como aquellos servicios que, ofrecidos sobre una infraestructura básica de comunicaciones, incorporan funcionalidades adicionales más allá de la simple transmisión de información (Stallings, 2014). Un SVA se caracteriza por utilizar la red de datos como medio de transporte, pero añadiendo capacidades de procesamiento, gestión o análisis de la información que generan valor para los usuarios finales. En la misma línea, Tanenbaum y Wetherall (2011) señalan que, en las redes modernas, la diferenciación de los servicios ya no se logra únicamente mediante la conectividad, sino a través de la provisión de servicios digitales que integran procesamiento, almacenamiento y control de la información para resolver necesidades específicas del negocio.

En la práctica del sector, se ha evidenciado una transformación en el modelo de negocio de los operadores de telecomunicaciones. Estudios sectoriales muestran que compañías globales como Telefónica han incrementado significativamente la proporción de ingresos provenientes de servicios digitales, como soluciones en la nube, ciberseguridad, IoT y plataformas de datos, reduciendo su dependencia exclusiva de los servicios tradicionales de conectividad (Analysys Mason, 2024). Esta evolución responde a la necesidad de generar nuevas fuentes de ingresos en un entorno donde la conectividad básica se ha convertido en un servicio altamente competitivo y de márgenes reducidos, e impulsa el posicionamiento de los operadores como habilitadores de soluciones digitales integrales que combinan conectividad, procesamiento distribuido y gestión inteligente de la información.

En el ámbito empresarial y financiero, los SVA se materializan principalmente en soluciones basadas en datos: sistemas de analítica avanzada, monitoreo, clasificación y segmentación de clientes; todas ellas dependientes de arquitecturas capaces de integrar grandes volúmenes de información y procesarlos de manera eficiente (Cisco Systems, 2017). Bajo esta lectura, un

servicio digital construido sobre una arquitectura Big Data distribuida, como el indicador *Marca Pyme* desarrollado en el presente trabajo, constituye un Servicio de Valor Agregado: transforma la infraestructura de comunicaciones en un habilitador de información estratégica para la organización, alineándose con la tendencia global de evolución de los servicios de telecomunicaciones hacia modelos basados en datos y valor agregado.

3.4.2.9 Metodología Scrum y gestión con Jira

Scrum como marco ágil

Scrum es un marco de trabajo ágil, ligero y de propósito general, ampliamente adoptado en proyectos de tecnología de la información para abordar problemas complejos donde los requerimientos pueden evolucionar a lo largo del tiempo (Schwaber y Sutherland, 2020). Su núcleo metodológico se basa en ciclos iterativos e incrementales denominados *Sprints*, de duración acotada (generalmente entre una y cuatro semanas), durante los cuales el equipo se compromete a producir un incremento de valor terminado y verificable.

Sáenz et al. (2018) destacan que Scrum es actualmente uno de los marcos ágiles más populares en el ámbito empresarial, debido a su capacidad para entregar respuestas rápidas, adaptarse a los cambios y mantener la motivación del equipo a través de reuniones periódicas de coordinación. Dichas reuniones constituyen los eventos centrales del marco:

- **Sprint Planning.** Reunión inicial del Sprint en la que el equipo, junto al *Product Owner* y al *Scrum Master*, define la meta de la iteración y prioriza las tareas según el valor que generan al cliente.
- **Daily Scrum.** Reunión diaria breve donde cada miembro del equipo expone los avances del día anterior, los objetivos del día en curso y los impedimentos que requieren atención.
- **Sprint Review.** Reunión de cierre en la que el equipo presenta los incrementos logrados y el *Product Owner* valida los resultados, ajustando el *backlog* si es necesario.

- **Sprint Retrospective.** Reunión de mejora continua en la que el equipo y el *Scrum Master* analizan lo que funcionó, lo que no, y definen acciones de mejora para el siguiente Sprint (Sáenz et al., 2018).

En la figura 7, se muestra modelo de referencia para la comprensión de las actividades en Scrum y su implementación

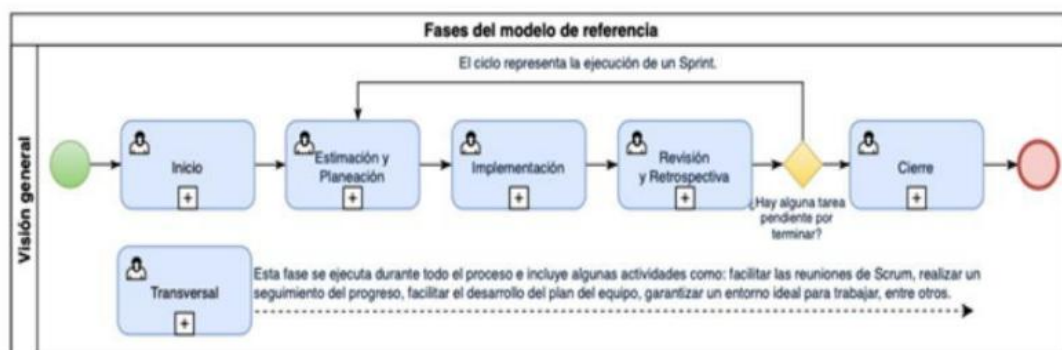


Figura 7 Modelo de referencia del marco de trabajo Scrum. Fuente. Pardo et al. (2020)

Jira como herramienta de gestión y trazabilidad

La aplicación efectiva de Scrum en proyectos de Big Data exige el soporte de una herramienta de gestión que permita registrar requerimientos, organizar el *backlog*, asignar responsabilidades y mantener la trazabilidad del flujo de trabajo entre múltiples equipos (Atlassian, 2024). Para este propósito se utiliza Jira, una plataforma desarrollada por Atlassian ampliamente adoptada en entornos corporativos para la gestión de proyectos ágiles.

Jira permite estructurar el trabajo en *epics*, *stories* y *tasks*, asociarlas a Sprints, asignarlas a perfiles específicos del equipo y dar seguimiento al estado de cada actividad mediante tableros visuales tipo *Kanban*. Adicionalmente, posibilita el registro documental de las decisiones tomadas, la asociación de artefactos técnicos, como dictámenes y validaciones, y la generación de métricas de avance por iteración. En el contexto de proyectos de servicios digitales basados en datos, esta capacidad de trazabilidad resulta esencial para coordinar la interacción entre los equipos de análisis funcional, ingeniería de datos, modelado y gobierno del dato a lo largo del ciclo de vida del servicio.

3.4.2.10 Marco regulatorio para la clasificación Pyme

La clasificación de un cliente como Pequeña o Mediana Empresa se sustenta en dos marcos normativos internacionales que definen los criterios cuantitativos y cualitativos aplicables al segmento. El primero corresponde a la Recomendación 2003/361/CE de la Comisión Europea, que establece la definición armonizada de microempresas, pequeñas y medianas empresas con base en tres umbrales: el número de empleados (inferior a 250), el volumen de negocios anual (igual o inferior a 50 millones de euros) y el balance general anual (igual o inferior a 43 millones de euros). Esta recomendación constituye una referencia ampliamente adoptada por las instituciones financieras internacionales para homogeneizar la categorización del segmento Pyme con fines comerciales, regulatorios y de reporte (Comisión Europea, 2003).

El segundo marco corresponde a los lineamientos del Comité de Supervisión Bancaria de Basilea, en su versión denominada *Basel III: Finalising post-crisis reforms* y comúnmente referida como Basilea IV o BIS IV. Este conjunto de reformas, emitido por el Bank for International Settlements (2017), introduce ajustes en el tratamiento del riesgo de crédito de las exposiciones bancarias, incluyendo disposiciones específicas para las exposiciones frente a Pymes, las cuales reciben ponderaciones de riesgo diferenciadas respecto a las grandes corporaciones. La adopción de estos lineamientos por parte de las entidades financieras impone la necesidad de identificar de manera trazable y consistente al segmento Pyme dentro de sus carteras, condición que motiva la construcción de indicadores estandarizados como el desarrollado en el presente trabajo (Bank for International Settlements, 2017).

3.5 Procedimiento

La necesidad del aprovisionamiento del indicador de identificación del segmento Pyme se origina dentro del flujo formal de gestión de proyectos,

ante un requerimiento del área de Riesgos, la cual solicita segmentar a los clientes según su condición de Pequeña y Mediana Empresa para fines de análisis y gestión.

Para este tipo de desarrollo, la organización utiliza el enfoque ágil bajo la metodología Scrum; el cual regula la forma en que se solicitan, diseñan, desarrollan y formalizan los servicios digitales basados en datos, apoyándose en la herramienta Jira para la gestión del requerimiento, el seguimiento del flujo de trabajo y la trazabilidad continua del ciclo de vida del servicio entre todos los equipos involucrados.

3.5.1 Adaptación del proyecto a la Metodología Scrum

La metodología Scrum fue adaptada al modelo organizacional del proyecto, estructurando el trabajo en un ciclo de desarrollo de veinte semanas (cinco meses) compuesto por seis Sprints consecutivos de tres semanas cada uno.

Si bien el marco Scrum admite Sprints de entre una y cuatro semanas, en este proyecto se optó por una duración de tres semanas por iteración, considerando los tiempos requeridos para las validaciones funcionales con el área de negocio, las revisiones con Aseguramiento de Datos y la formalización bajo gobierno corporativo.

A diferencia del modelo Scrum tradicional, el Sprint Planning se realizó al inicio del ciclo, estableciendo el alcance funcional y técnico del indicador, la identificación de historias de usuario y la planificación por Sprint; por otro lado, el Sprint Review y la Retrospectiva se ejecutaron en la semana final del ciclo, evaluando el cumplimiento de objetivos y formalizando la evaluación de cierre del servicio.

En la Figura 8 se presenta el diagrama del proceso, el cual ilustra la secuencia de las ocho etapas que estructuran el desarrollo del servicio digital, desde la planificación inicial hasta el cierre con la evaluación final.



Figura 8 Diagrama de proceso. Fuente: Elaboración propia

Bajo este esquema, las actividades técnicas del proyecto se organizaron en fases alineadas al flujo iterativo Scrum, abarcando el diagnóstico de fuentes, el diseño lógico del indicador, su implementación en entorno sandbox, la validación tanto funcional como técnica, la formalización bajo lineamientos de gobierno de datos, y finalmente la revalidación e integración en el entorno productivo, tal como se detalla en la Tabla 3.

Tabla 3. Cronograma de desarrollo del Proyecto

Sprint	Fase del proceso	Actividad principal	Semanas
Sprint Planning	Planificación	Definición de alcance	Semana 1
Sprint 1	Diagnóstico	Análisis de requerimientos y fuentes	Semana 2-4
Sprint 2	Diseño lógico	Modelamiento del indicador	Semana 5-7
Sprint 3	Implementación	Desarrollo Pyspark en sandbox	Semana 8-10
Sprint 4	Validación	Pruebas funcionales y técnicas	Semana 11-13
Sprint 5	Gobernanza	Dictamen y formalización	Semana 14-16
Sprint 6	Integración	Revalidación y producción	Semana 17-19
Review y Retrospective	Cierre	Evaluación	Semana 20

Fuente. Elaboración propia.

3.5.2 Desarrollo del ciclo de Sprints

La presente sección describe el detalle de la ejecución del proyecto bajo el marco Scrum adaptado, tal como se expuso en la sección 3.5.1, recorriendo cronológicamente cada una de las ocho etapas que estructuraron el ciclo de

desarrollo de todos los sprints. La narración sigue la secuencia presentada en la Figura 8, iniciando por el Sprint Planning, continuando con los seis Sprints técnicos del proyecto y cerrando con la Sprint Review y la Retrospectiva final.

Para cada Sprint se presentan los objetivos planteados, la descripción de las actividades técnicas ejecutadas, las decisiones de diseño adoptadas y los entregables generados al cierre de cada iteración; mostrando así, la trazabilidad entre la planificación inicial del ciclo y los resultados obtenidos en el desarrollo del indicador Marca Pyme.

3.5.2.1 *Sprint Planning*

El Sprint Planning constituye la fase inicial del ciclo de desarrollo del proyecto, y representa el momento en el que se establece el alcance funcional y técnico del servicio digital a desarrollar a lo largo del ciclo de trabajo. Para el caso del presente Trabajo de Suficiencia Profesional, esta etapa se ejecutó durante la primera semana de enero de 2024, como punto de partida para la construcción del indicador Marca Pyme, en respuesta a un requerimiento formal del área de Riesgos de la institución financiera, orientado a clasificar a los clientes según su pertenencia al segmento de Pequeñas y Medianas Empresas.

Bajo este contexto, durante el Sprint Planning se definieron los lineamientos generales del desarrollo, permitiendo organizar el trabajo iterativo que sería ejecutado a lo largo de los seis Sprints del ciclo; así como articular la coordinación entre los equipos funcionales, técnicos y de gobierno de datos involucrados en la entrega de este indicador.

Objetivo funcional del indicador

El objetivo funcional consiste en disponer de un indicador único, gobernado y reutilizable, capaz de clasificar a los clientes empresariales de la institución según su condición de Pyme, asegurando que el resultado sea consumido por el área solicitante para fines de análisis y gestión, y por las demás áreas usuarias dentro del marco normativo vigente.

Fuentes de información

A fin de garantizar la viabilidad del desarrollo, durante el Sprint Planning se identifican preliminarmente las fuentes de información requeridas para la construcción del indicador. En todos los casos, se trata de datasets previamente ingestados en la capa maestra del Data Lake corporativo, lo cual asegura su disponibilidad inmediata para el procesamiento distribuido. Las fuentes consideradas son las siguientes:

- **Base maestra de clientes**, que contiene los datos generales de las personas, naturales y jurídicas, que han mantenido o mantienen una relación contractual con la institución.
- **Tabla de grupos de clientes**, que registra los Grupos Económicos a nivel local, conformados por al menos dos personas jurídicas relacionadas entre sí.
- **Tabla de relaciones cliente-grupo**, que detalla los vínculos existentes entre un cliente y otros clientes o empresas dentro de un mismo grupo económico.
- **Tabla de información económica del cliente**, que consolida la información económica relevante para la segmentación, como número de empleados y volumen total de ingresos, con visión de fin de mes.
- **Tabla de tipo de cambio**, que provee las divisas y sus correspondientes tipos de cambio, indispensable para homogeneizar los importes considerados en la lógica del indicador.
- **Parametría regulatoria**: contiene los umbrales y categorías vigentes definidos por el regulador y por la institución financiera para la clasificación del segmento Pyme.
- **Indicador Pyme local**: contiene la clasificación local del cliente bajo las categorías Microempresa, Pequeña Empresa y Mediana Empresa.

Alcance técnico del desarrollo

Desde la perspectiva técnica, el alcance del ciclo de desarrollo del proyecto, se concentró en disponibilizar el indicador *Marca Pyme* dentro de la plataforma Big Data institucional, asegurando su procesamiento distribuido

PySpark, la validación en entorno *sandbox*, la documentación técnica y, finalmente, el despliegue productivo del servicio.

Priorización del trabajo del proyecto

Dentro del conjunto de solicitudes activas del período provenientes de las distintas áreas de la institución, la implementación del indicador *Marca Pyme* fue clasificada como prioridad alta (Prioridad 1), estableciéndose como objetivo su implementación y pase a producción para fines del sprint 6. Esta priorización respondió al carácter regulatorio y estratégico del indicador, así como a la necesidad de uniformizar los criterios de segmentación de clientes empresariales a nivel corporativo.

La planificación temporal de las actividades por Sprint se consolidó en un cronograma Gantt, presentado en la Figura 10.

Sprint	SEMANAS																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Planning																				
Sprint 1																				
Sprint 2																				
Sprint 3																				
Sprint 4																				
Sprint 5																				
Sprint 6																				
Review & Retrospective																				

Figura 10 Cronograma Gantt de la planificación del ciclo de desarrollo del proyecto. Fuente. Elaboración propia.

Nota. El cronograma muestra la distribución temporal de las actividades planificadas a lo largo de los seis Sprints ejecutados en 2024.

3.5.2.2 Sprint 1: Definición del requerimiento y diagnóstico de fuentes

Durante el primer Sprint del ciclo de desarrollo del proyecto, se aborda el análisis inicial del requerimiento del indicador *Marca Pyme* y la caracterización técnica de las fuentes de datos disponibles dentro del Data Lake corporativo de la institución financiera. El objetivo central de este Sprint, consiste en validar la viabilidad técnica del desarrollo antes de iniciar con el diseño lógico, asegurando de esta manera que la información requerida exista, este

disponible, sea accesible y resulte consistente con la frecuencia de actualización que demanda el área usuaria solicitante.

En ese sentido, las actividades desarrolladas se estructuran en cinco frentes complementarios: la identificación de las tablas fuente de información, la verificación de su disponibilidad y particionamiento en HDFS, la revisión de la estructura de datos, la validación de la periodicidad respecto del requerimiento funcional y; finalmente, la verificación de la viabilidad técnica del procesamiento distribuido sobre el entorno sandbox.

a) Identificación de tablas fuente

A partir de los atributos definidos en el Sprint Planning, se realiza el levantamiento de las fuentes oficiales del Data Lake corporativo necesarias para la construcción del indicador. Esta identificación se ejecuta en coordinación con el área usuaria, los líderes de gobierno del dato y la documentación funcional de los productos comerciales involucrados, lo que permite delimitar el conjunto mínimo de tablas que sostienen la lógica del *Marca Pyme*.

Como resultado, se mapearon 3 fuentes adicionales, que no habían sido consideradas en el planeamiento, pero son necesarias para la construcción del indicador. Se determinan como fuentes principales la base maestra de segmentación de clientes (tabla destino del indicador), la tabla de información económica del cliente, la tabla de relación cliente–grupo económico, el maestro de grupos económicos, la tabla de tipo de cambio y conversión monetaria oficial, el repositorio paramétrico regulatorio, la tabla de identificación del cliente, la tabla de empleados por cliente y la tabla de ventas consolidadas a nivel grupo, cuyo mapeo funcional se presenta en la Tabla 4. El detalle estructural de la identificación de estas diez fuentes se presenta en el **Anexo B**.

Tabla 4. Mapeo de fuentes funcionales del indicador Marca Pyme

Nombre lógico de la tabla	Capa HDFS	Origen	Periodicidad	Uso dentro del indicador
Base maestra de segmentación de	Master	Banco	Mensual	Tabla destino donde se materializa el campo

clientes (tabla destino del indicador)				Marca Pyme resultado del proceso de clasificación.
Tabla de información económica del cliente	Master	Banco	Mensual	Provee volumen de ventas, número de empleados y demás variables económicas del cliente individual.
Tabla de relación cliente-grupo económico	Master	Banco	Mensual	Permite asociar a cada cliente con su grupo económico para evaluar la condición de pertenencia consolidada.
Maestro de grupos económicos	Master	Banco	Mensual	Aporta el volumen de ventas consolidado del grupo económico al que pertenece el cliente.
Tabla de tipo de cambio y conversión monetaria oficial	Master	SBS	Diaria	Permite expresar los volúmenes de venta en una moneda común (EUR) para la evaluación de los umbrales regulatorios.
Repositorio paramétrico regulatorio	Master	Banco	Mensual	Almacena los umbrales y parámetros regulatorios vigentes que definen las reglas de clasificación del indicador.
Tabla de identificación del cliente	Master	Banco	Mensual	Provee el identificador único del cliente y los atributos básicos requeridos para el cruce entre fuentes.
Tabla de empleados por cliente	Master	Banco	Mensual	Provee el número de empleados (cliente y grupo)
Tabla de ventas consolidadas a nivel grupo	Master	Banco	Mensual	Volumen de ventas del grupo
Tabla de Pyme local	Master	Banco	Mensual	Indicador Pyme local (cruce con parametrías del Repositorio paramétrico regulatorio)

Fuente. Elaboración propia a partir del levantamiento realizado sobre el repositorio oficial de información del Data Lake corporativo de la institución financiera (2024).

b) Verificación de disponibilidad, estructura y particionamiento en HDFS

Al concluir la identificación de fuentes, se valida que se encuentren correctamente ingestadas en el Data Lake y disponibles para su lectura desde

el entorno sandbox. Para ello, se ejecutan comandos `hdfs dfs -ls` sobre las rutas oficiales de cada tabla, lo cual permite confirmar tres aspectos críticos: la existencia física de los archivos sobre HDFS, la consistencia del particionamiento por fecha de proceso (elemento clave para garantizar el acceso histórico requerido por el área de Riesgos) y la habilitación de los permisos de lectura sobre los campos involucrados en la lógica del indicador.

Con la revisión se confirma que la totalidad de las fuentes se encuentran correctamente almacenadas en el sistema distribuido, sin restricciones de acceso ni inconsistencias estructurales que comprometan su procesamiento posterior.

En la figura 11, se puede apreciar un ejemplo de consulta HDFS para revisión de la fuente de la tabla de información económica del cliente.

```
$ hdfs dfs -ls /data/master/clientes/info_economica_cliente/
Found 12 items
drwxr-xr-x - data_user hdfs 0 2024-01-05 03:12 /data/master/clientes/info_economica_cliente/fecha_corte=2023-12-31
drwxr-xr-x - data_user hdfs 0 2023-12-05 03:14 /data/master/clientes/info_economica_cliente/fecha_corte=2023-11-30
drwxr-xr-x - data_user hdfs 0 2023-11-03 03:20 /data/master/clientes/info_economica_cliente/fecha_corte=2023-10-31
drwxr-xr-x - data_user hdfs 0 2023-10-04 03:31 /data/master/clientes/info_economica_cliente/fecha_corte=2023-09-30
drwxr-xr-x - data_user hdfs 0 2023-09-03 03:35 /data/master/clientes/info_economica_cliente/fecha_corte=2023-08-31
drwxr-xr-x - data_user hdfs 0 2023-08-05 03:10 /data/master/clientes/info_economica_cliente/fecha_corte=2023-07-31
drwxr-xr-x - data_user hdfs 0 2023-07-04 03:17 /data/master/clientes/info_economica_cliente/fecha_corte=2023-06-30
...
```

Figura 11 Comando HDFS para verificación de disponibilidad y particionamiento de fuentes. Fuente. Elaboración propia. Las rutas y nombres de tablas se presentan de forma esquemática por motivos de confidencialidad de la institución financiera.

c) Revisión de estructura de datos

Después de realizarse la validación de disponibilidad, se ejecuta un análisis exploratorio de la estructura interna de cada tabla mediante consultas en PySpark sobre el entorno sandbox, complementado con la documentación oficial del repositorio de información del Data Lake.

Esta revisión se focaliza en cuatro dimensiones: los tipos de datos asociados a cada campo, la identificación de los campos obligatorios para definir la lógica del indicador, la presencia de valores nulos y la consistencia del formato de las llaves para el cruce entre tablas.

En la Figura 12, se presenta el detalle de los campos, incluyendo el campo que se busca poblar “marca_pyme”. Asimismo, se puede encontrar mayor detalle sobre las estructuras de las demás fuentes en el **Anexo A**.

```
>>> df_segmentacion = spark.read.parquet("ruta_master/maestro_segmentacion")
>>> df_segmentacion.printSchema()
root
 |-- fecha_cierre:      date      (nullable = true)
 |-- id_cliente:        string    (nullable = true)
 |-- tipo_titular_sra:  string    (nullable = true)
 |-- marca_pyme:        string    (nullable = true)
 |-- id_perfil_cliente: string    (nullable = true)
 |-- id_segmento_actual: string    (nullable = true)
 |-- id_segmento_local: string    (nullable = true)
 |-- tipo_seguimiento:  string    (nullable = true)
 |-- id_segmento_global: string    (nullable = true)
 |-- id_entific:        string    (nullable = true)
 |-- fecha_auditoria:   timestamp (nullable = true)
 |-- id_origen_informacion: string  (nullable = true)
 |-- tipo_vision_riesgo_pyme: string  (nullable = true)
 |-- tipo_pyme_bis_iv:  string    (nullable = true)
```

Figura 12 printSchema simulado de la tabla destino. Fuente. Elaboración propia a partir de la consulta realizada en el entorno PySpark sobre la capa Master del Data Lake corporativo de la institución financiera

En la Tabla 5 se presentan los campos relevantes de la tabla maestra de segmentación de clientes utilizados en el proceso de generación del indicador Marca Pyme. La fila 4 corresponde al campo destino sobre el cual se materializa el resultado del proceso de clasificación. Los nombres técnicos se presentan adaptados con fines de confidencialidad institucional.

Tabla 5. Diccionario de campos relevantes de la tabla maestra de segmentación

Campo	Tipo de dato	Descripción funcional
fecha_cierre	date	Fecha de cierre del periodo de evaluación. Define la partición mensual sobre la cual se procesa el indicador.
id_cliente	string	Identificador único del cliente bajo el modelo global de la institución. Llave primaria para el cruce con las demás fuentes.
tipo_titular_sra	string	Indicador de titularidad del cliente bajo el modelo global.
marca_pyme	string	Campo destino del proceso de clasificación. Toma el valor "Y" cuando el cliente cumple los criterios de Pyme y "N" en caso contrario.
id_perfil_cliente	string	Identificador del perfil del cliente, asociado a su comportamiento y características comerciales.
id_segmento_actual	string	Identificador del segmento comercial vigente del cliente al momento del cierre.
id_segmento_local	string	Identificador del segmento comercial del cliente a nivel local, definido por la institución financiera.

tipo_seguimiento	string	Tipo de seguimiento aplicado al cliente conforme a las políticas de monitoreo de la entidad.
id_segmento_global	string	Identificador de segmentación global utilizado en el reporte hacia la matriz corporativa.
id_entific	string	Identificador entific del cliente bajo el modelo global de identificación corporativa.
fecha_auditoria	timestamp	Fecha y hora del registro de auditoría del proceso de generación del campo.
id_origen_informacion	string	Identificador del origen de la información que alimentó el registro.
tipo_vision_riesgo_pyme	string	Tipo de visión de riesgo del cliente bajo el modelo Pyme corporativo.

Fuente. Elaboración propia a partir del diccionario de datos del Data Lake corporativo de la institución.

d) Validación de periodicidad y alineación con requerimiento funcional

Debido que el área usuaria de Riesgos requiere una clasificación consistente con el cierre mensual, es necesario verificar que la frecuencia de actualización de cada fuente sea compatible con dicha solicitud. De esta manera, se confirma que las fuentes económicas presentan actualización mensual alineada al ciclo contable de la institución financiera, que el tipo de cambio oficial (que se actualiza diariamente) es consumido en su versión correspondiente al último día hábil del mes, y que no existen desfases entre las distintas fuentes que puedan comprometer la consistencia del indicador en el corte mensual solicitado. El detalle comparativo de la periodicidad de las fuentes se presenta en la Tabla 6.

Tabla 6. Comparativo de periodicidad de las fuentes del indicador Marca Pyme

Nombre lógico de la tabla	Periodicidad	Corte
Base maestra de segmentación de clientes	Mensual	Calendario
Tabla de información económica del cliente	Mensual	Calendario
Tabla de relación cliente–grupo económico	Mensual	Calendario
Maestro de grupos económicos	Mensual	Calendario
Tabla de tipo de cambio oficial	Diaria	Hábil
Repositorio paramétrico regulatorio	Mensual	Calendario
Tabla de identificación del cliente	Mensual	Calendario

Tabla de empleados por cliente	Mensual	Calendario
Tabla de ventas consolidadas a nivel grupo	Mensual	Calendario
Tabla de Pyme local	Mensual	Calendario

Fuente. Elaboración propia.

e) Verificación de viabilidad técnica del procesamiento distribuido

Como cierre del Sprint, se ejecutan pruebas preliminares de lectura y conteo en el entorno sandbox utilizando PySpark sobre las diez fuentes identificadas, con el fin de validar dos condiciones técnicas: la correcta carga de los datos, que se verifica mediante el conteo de registros y la unicidad de las llaves primarias en cada tabla, descartando duplicidades en los campos clave considerando el universo total de clientes activos de la institución financiera.

Los resultados de estas pruebas confirman la viabilidad técnica del desarrollo del indicador dentro de la arquitectura Big Data corporativa, sentando las bases para iniciar el diseño lógico del servicio en el Sprint 2.

En la figura 13, se presenta la evidencia de las dos validaciones aplicadas sobre la tabla maestra de segmentación de clientes en el cierre del periodo de evaluación: (a) la validación de volumetría confirma la lectura del universo completo de clientes desde la capa Master del Data Lake corporativo y verifica que el volumen es procesable en entorno distribuido; (b) la validación de unicidad confirma la integridad de la llave primaria conformada por el par (id_cliente, fecha_cierre), condición indispensable para los cruces posteriores con las demás fuentes funcionales sin riesgo de duplicación de registros. El volumen total se presenta redondeado al orden de magnitud con fines de confidencialidad institucional.

a) Validación de volumetría de la tabla destino

```
# Lectura de la tabla maestra de segmentación desde la capa Master
# del Data Lake corporativo, partición correspondiente al cierre.

df_seg = spark.read.parquet("ruta_master/maestro_segmentacion").filter(col("fecha_cierre") == "2023-12-31")
df_seg.count()

12 200 000    # registros aproximados al cierre evaluado
```

b) Validación de unicidad de la llave primaria

```
# Validación de integridad de la llave primaria de la tabla destino:
# el par (id_cliente, fecha_cierre) debe ser único.

total = df_seg.count()
unicos = df_seg.dropDuplicates(["id_cliente", "fecha_cierre"]).count()

print(f"Total registros : {total:,}")
print(f"Registros únicos: {unicos:,}")
print(f"Duplicados      : {total - unicos:,}")

Total registros : 12 200 000
Registros únicos: 12 200 000
Duplicados      : 0      # llave primaria íntegra
```

Figura 13 Evidencia de validación de volumetría y unicidad de llaves de la tabla destino. Fuente. Elaboración propia a partir de la ejecución de consultas en el entorno PySpark de la institución.

Entregables del Sprint 1

Al cierre del Sprint, se obtiene lo siguiente: el inventario validado de las diez fuentes funcionales requeridas para la construcción del indicador, la confirmación documentada de su disponibilidad y particionamiento sobre HDFS, la validación estructural y exploratoria de los datos contenidos en cada tabla, la confirmación de la periodicidad de actualización alineada al cierre mensual exigido por el área de Riesgos y la evidencia preliminar de la viabilidad técnica del procesamiento distribuido en PySpark; lo cual habilita formalmente el inicio del diseño lógico del servicio en el siguiente Sprint.

3.5.2.3 *Sprint 2: Diseño de la arquitectura lógica del servicio*

Una vez confirmada la viabilidad técnica de las fuentes durante el Sprint 1, en el segundo Sprint del ciclo de desarrollo del proyecto, se aborda el diseño de la arquitectura lógica del indicador *Marca Pyme*, esta etapa se orienta a definir el flujo de datos, la estructura de transformación y las reglas de clasificación que serán implementadas posteriormente sobre el entorno distribuido.

El objetivo central de este Sprint consiste en estructurar el servicio digital como un componente reutilizable dentro del ecosistema Big Data de la institución financiera, asegurando coherencia funcional con la lógica de negocio definida por el área de Riesgos, escalabilidad técnica frente al volumen del universo de clientes y compatibilidad con la arquitectura de comunicaciones corporativa descrita en la sección 3.4.2.7.

Considerando este contexto, las actividades desarrolladas se estructuran en cuatro frentes: la definición formal de las variables de entrada del servicio, el diseño del flujo lógico de transformación distribuida, la formulación de la regla condicional final del indicador y la especificación de la estructura de salida del campo a generar.

a) Definición de variables de entrada del servicio

A partir de la lógica funcional levantada con el área usuaria, se establecen formalmente las variables necesarias para la evaluación del indicador, las cuales se agrupan en tres categorías complementarias.

- La primera categoría corresponde a las **variables económicas**, conformadas por el número de empleados, el volumen de ventas y el volumen de activos, tanto a nivel cliente individual como del grupo económico al que pertenece.
- La segunda categoría agrupa las **variables regulatorias**, integradas por la categoría CORASU (identificador de clientes para la gestión de su perfil de riesgo y productos financieros) equivalente al segmento BIS (*Bank for International Settlements*), el tipo de persona y los parámetros del Repositorio paramétrico regulatorio que centraliza los umbrales vigentes definidos por la SBS.
- La tercera categoría comprende las **variables estructurales**, conformadas por los indicadores de grupo corporativo, intergrupo, cámara de compensación e indicador Pyme local, las cuales actúan como filtros de exclusión sobre el universo evaluado.

El detalle de cada variable, su tipo de dato y su procedencia se presenta en la Tabla 7.

Tabla 7. Características de las variables de entrada del servicio

Categoría	Variable	Tipo de dato	Tabla fuente	Nivel	Rol en el indicador
Económica	Número de empleados Total de trabajadores	Numeric	Tabla de información económica del cliente (F2)	Cliente y Grupo	Insumo del Bloque 1 — evaluación contra umbrales

	activos del cliente / grupo.				económicos
Económica	Volumen de ventas Ventas anuales del cliente y de su grupo económico, normalizadas en EUR.	Numeric	Tabla de información económica del cliente (F2) · Maestro de grupos económicos (F4)	Cliente y Grupo	Insumo del Bloque 1 — evaluación contra umbrales económicos
Económica	Volumen de activos Activos totales del cliente y de su grupo económico, normalizados en EUR.	Numeric	Tabla de información económica del cliente (F2) · Maestro de grupos económicos (F4)	Cliente y Grupo	Insumo del Bloque 1 — evaluación contra umbrales económicos
Regulatoria	Tipo de persona Persona natural con negocio / persona jurídica.	String	Tabla de identificación del cliente (F7)	Cliente	Insumo del Bloque 2 — evaluación regulatoria
Regulatoria	Parámetros del repositorio regulatorio Provee los umbrales de comparación dinámicos del indicador (ventas, activos, empleados) por segmento.	Numeric	Repositorio paramétrico regulatorio (F6)	Cliente	Insumo de los Grupos 1 y 2 — provee los umbrales de comparación.
Estructural	Categoría CORASU (BIS) Segmento de cliente equivalente a la categoría BIS estándar Basilea.	String	Tabla de identificación del cliente (F7)	Cliente	Insumo del Bloque 3 — exclusión estructural
Estructural	Indicador de grupo corporativo Marca booleana que identifica clientes pertenecientes a grupos corporativos grandes.	Boolean	Tabla de relación cliente–grupo económico (F3)	Cliente	Insumo del Bloque 3 — exclusión estructural
Estructural	Indicador de intergrupo Identifica clientes que son	Boolean	Tabla de identificación del cliente (F7)	Cliente	Insumo del Bloque 3 — exclusión estructural

	contrapartes del propio grupo financiero.				
Estructural	Indicador de cámara de compensación Identifica entidades que operan como cámara de compensación.	Boolean	Tabla de identificación del cliente (F7)	Cliente	Insumo del Bloque 3 — exclusión estructural
Estructural	Indicador Pyme local Marca local de Pyme proveniente de la clasificación previa del banco.	Boolean	Tabla de Pyme local (F10)	Cliente	Insumo del Bloque 2 — evaluación regulatoria

Fuente. Elaboración propia a partir del diseño funcional del indicador (2024).

Nota. La Tabla presenta la totalidad de las variables de entrada del servicio digital Marca Pyme, agrupadas en tres categorías funcionales: económicas, regulatorias y estructurales. Las referencias (F1) a (F10) corresponden a las fuentes funcionales detalladas en la Tabla 5. La columna "Rol en el indicador" indica el grupo de evaluación paramétrica (Grupo 1, 2 o 3) en el que cada variable participa, según el flujo lógico de transformación descrito en la sección 3.5.1.3.

b) Diseño del flujo lógico de transformación

El procesamiento del indicador se estructura en tres etapas técnicas secuenciales que reflejan el diseño de la lógica de negocio sobre la arquitectura distribuida.

Etapas 1: Consolidación de información cliente-grupo

Se encarga de recuperar el grupo económico activo asociado a cada cliente, filtrando la relación por el tipo y validando que el grupo se encuentre vigente mediante el estatus (activo) del repositorio de grupos económicos; así como de agregar los volúmenes de empleados, ventas y activos a nivel grupo, asegurando que cada cliente sea evaluado bajo la información consolidada que le corresponde, de acuerdo a las consideraciones indicadas.

Etapas 2: Normalización monetaria

En esta etapa se ejecuta la conversión de los volúmenes de ventas y activos a euros utilizando el tipo de cambio oficial SBS del cierre mensual y el factor de conversión paramétrico almacenado en las tablas de tipo de cambio del Data Lake corporativo.

Esta condición es indispensable para la evaluación posterior contra los umbrales europeos de Pyme expresados en una unidad homogénea, a partir de esta se define un criterio importante más que posteriormente delimitará al segmento al que pertenece.

Etapas 3: Evaluación paramétrica

Para esta etapa final se aplica las reglas de clasificación sobre el universo normalizado, estructuradas en los tres bloques decisionales que se presentan en la Figura 14.

El **Bloque 1** evalúa las variables económicas (número de empleados, volumen de ventas y volumen de activos) contra los umbrales europeos parametrizados en el repositorio regulatorio.

El **Bloque 2** valida las variables regulatorias correspondientes al tipo de persona y al indicador Pyme local previamente clasificado por el banco.

El **Bloque 3** aplica las exclusiones estructurales asociadas a las condiciones de intragrupo, cámara de compensación, grupo corporativo y categoría CORASU de exclusión (Soberanos, Multilaterales, Sector Público, Intermediarios de valores y Sistema Financiero).

El resultado parcial de cada bloque se expresa de forma binaria, donde el valor "Y" indica el cumplimiento satisfactorio de los criterios de continuidad y el valor "N" su incumplimiento.

De esta forma con el desarrollo de las 3 etapas, y los 3 bloques definidos en la tercera etapa se cataloga finalmente la clasificación clara de aquellos que pertenecen al segmento pyme, y los que no.

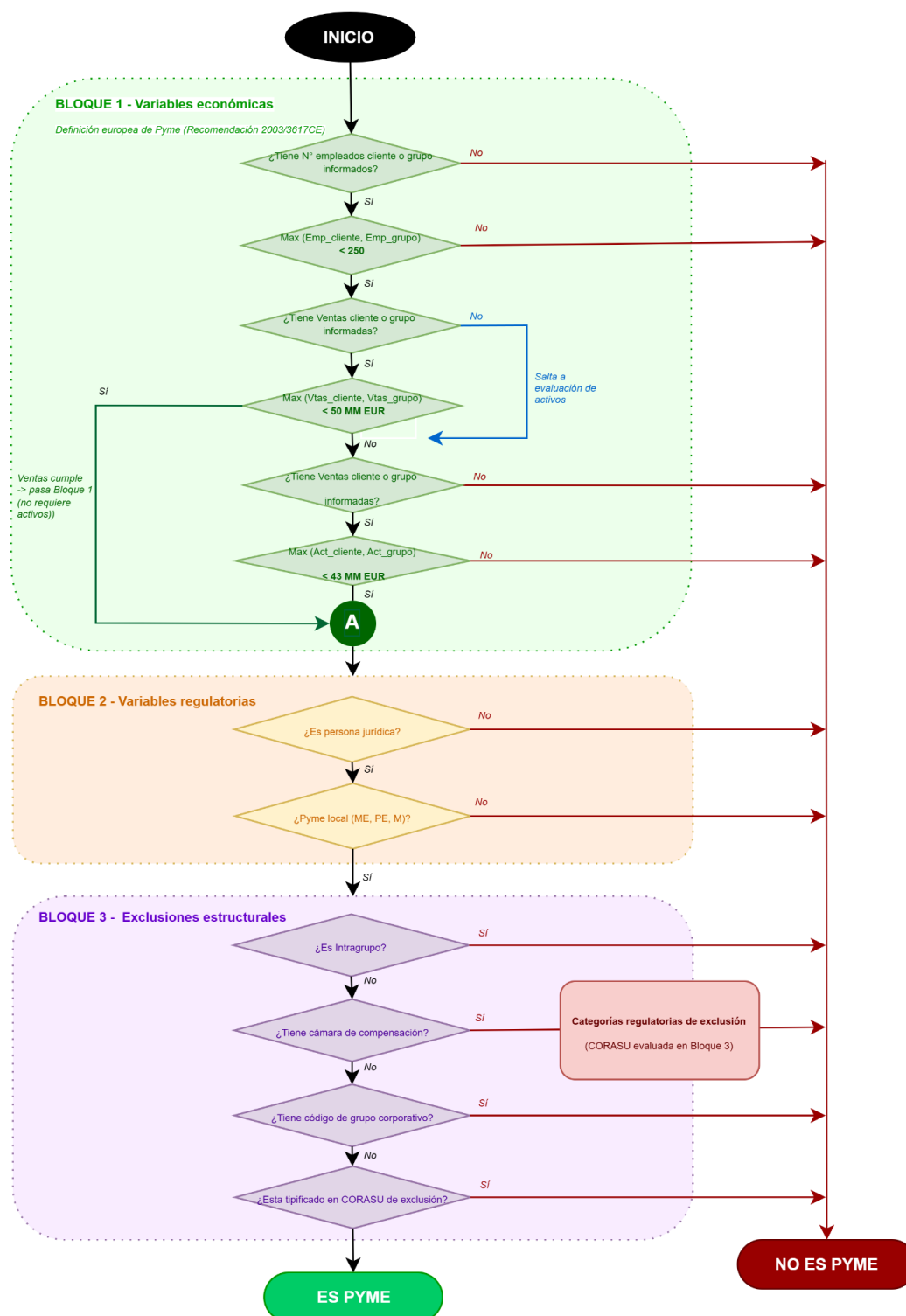


Figura 14 Diagrama lógico de clasificación del indicador Marca Pyme.
Fuente. Elaboración propia a partir de la lógica funcional definida con el área usuaria de Riesgos de la institución financiera y de la Recomendación 2003/361/CE de la Comisión Europea sobre la definición de microempresas, pequeñas y medianas empresas

El detalle del proceso de conversión monetaria de la Etapa 2 y de la estructura de evaluación por bloques de la Etapa 3 se presenta en las Tablas 8 y 9 respectivamente.

Tabla 8. Proceso de conversión monetaria del Bloque 2

Concepto	Fuente	Regla aplicada
Tipo de cambio	Tabla TC oficial	Selección cierre mensual
Factor conversión	Tabla TC oficial	División Valor promedio de cambio/número de unidad de cambio
Ventas EUR	Ventas locales	Ventas/tipo_cambio
Activos EUR	Activos locales	Activos /tipo_cambio

Fuente. Elaboración propia.

En la Tabla 9 se muestran los grupos y las variables evaluadas, cuyo resultado de evaluación se indica en “Resultado parcial”, siendo “Y” cuando pase satisfactoriamente dichos criterios. Y en caso contrario, se asigna “N”, que indica que no cumple los criterios.

Tabla 9. Estructura de evaluación paramétrica por grupos

Grupo	Variables evaluadas	Resultado parcial
Grupo 1(Económico)	Empleados, Ventas, Activos	Y/N
Grupo 2 (Regulatorio)	CORASU, exclusiones BIS	Y/N
Grupo 3 (Estructural)	Tipo persona, relación cliente-grupo	Y/N

Fuente. Elaboración propia

c) Definición de la regla final del indicador

De acuerdo a los resultados parciales obtenidos en cada uno de los tres bloques decisionales de la Etapa 3, se establece la lógica condicional final que determina la asignación del indicador *Marca Pyme* sobre cada cliente. Esta regla considera satisfactoria la clasificación únicamente cuando el cliente cumple simultáneamente con las tres validaciones: cuando supera los umbrales económicos del Bloque 1, satisface los criterios regulatorios del

Bloque 2 y no incurre en ninguna de las exclusiones estructurales del Bloque 3, condición que se expresa en la siguiente sentencia condicional que se muestra en la Figura 15.

```
Bloque1 = (Empleados < 250) AND ((Ventas < 50MM EUR) OR (Activos < 43MM EUR))
Bloque2 = (Es persona jurídica) AND (Es Pyme local: ME, PE o MI)
Bloque3 = NOT(intragrupo) AND NOT(cámara compensación)
          AND NOT(grupo corporativo) AND NOT(CORASU de exclusión)

IF Bloque1 AND Bloque2 AND Bloque3
  THEN marca_pyme = "Y"
  ELSE marca_pyme = "N"
```

Figura 15 Lógica condicional final del indicador. Fuente. Elaboración propia.

En términos de diseño, esta regla condicional constituye la función principal del servicio digital de valor agregado dentro de la capa de aplicación del modelo OSI (capa 7); articulando sobre la información distribuida un criterio de clasificación reproducible, auditable y consistente con la definición europea de Pyme establecida en la Recomendación 2003/361/CE de la Comisión Europea.

d) Definición de estructura de salida del servicio

A partir de la regla final establecida, se especifica formalmente la estructura de salida del servicio digital, definiendo las características técnicas y funcionales del campo a generar y de su tabla destino (segmentación):

- Campo: *marca_pyme*
- Tipo de dato: string (valor binario "Y"/"N")
- Nivel de granularidad: cliente
- Periodicidad: mensual (alineada al cierre de mes del Data Lake corporativo)
- Tabla destino: maestra de segmentación de clientes (capa Master, formato PARQUET)
- Llave primaria: par (id_cliente, fecha_cierre)

Esta especificación técnica formaliza la definición formal entre el servicio y los consumidores (área de Riesgos, modelos de capital regulatorio y reportería corporativa), garantizando que el campo se disponibiliza de forma consistente,

trazable y conforme a la convención de nomenclatura del modelo global de la institución financiera.

Entregables del Sprint 2

Al cierre del Sprint 2, se obtiene lo siguiente: la arquitectura lógica del servicio formalizada sobre las tres etapas técnicas de consolidación, normalización y evaluación; la estructuración de las variables de entrada en sus tres categorías funcionales: económicas, regulatorias y estructurales; la documentación de las reglas paramétricas trazables hacia la Recomendación 2003/361/CE de la Comisión Europea y hacia el repositorio paramétrico regulatorio de la institución financiera; el establecimiento de la lógica condicional final del indicador como la conjunción de los tres bloques decisionales; la especificación formal del campo destino *marca_pyme* y de la tabla maestra de segmentación de clientes; y el diagrama lógico del servicio formalizado en la Figura 14, lo cual habilita formalmente el inicio de la implementación del servicio en el entorno *Sandbox* mediante código PySpark distribuido en el siguiente Sprint.

3.5.2.4 Sprint 3: Implementación inicial del servicio en entorno *Sandbox*

Una vez formalizado el diseño de la arquitectura lógica del servicio durante el Sprint 2, en el tercer Sprint del ciclo de desarrollo del proyecto, se aborda la implementación inicial del indicador *Marca Pyme* dentro del entorno *Sandbox* de la plataforma Big Data corporativa, esta etapa se orienta a traducir el diseño funcional en código PySpark ejecutable sobre el clúster distribuido.

El objetivo central de este Sprint, consiste en validar preliminarmente el comportamiento del servicio sobre data real antes de su formalización productiva, verificando que las reglas paramétricas, las condiciones de clasificación y los criterios de exclusión definidos durante el diseño funcional se ejecuten de forma consistente sobre el volumen completo de clientes corporativos de la institución financiera.

En función de ello, las actividades desarrolladas se estructuran en cinco frentes alineados a las tres etapas técnicas definidas en el Sprint 2: la lectura de las diez fuentes funcionales sobre el entorno distribuido y su integración mediante operaciones de cruce (ambas correspondientes a la etapa de consolidación), la aplicación de las transformaciones y cálculos intermedios propios de la etapa de normalización, la implementación de la lógica de clasificación sobre los tres bloques decisionales (económicas, regulatorias y estructurales) correspondiente a la etapa de evaluación, y la validación preliminar de la distribución del indicador resultante como cierre de la implementación inicial.

a) Lectura de fuentes de datos en entorno distribuido

Se realiza la carga de las diez fuentes funcionales identificadas en la Tabla 5 desde la capa *Master* del Data Lake corporativo hacia el entorno PySpark, tomando para cada fuente la partición mensual correspondiente al cierre de mes (fecha_cierre).

La lectura genera *DataFrames distribuidos* que constituyen la base del procesamiento en Spark y permiten que las transformaciones posteriores se ejecuten en paralelo sobre el clúster.

En la figura 16 se puede observar el ejemplo conceptual de lectura de fuentes.

```
# Se realizan las lecturas y cargas de las fuentes desde la capa Master
# del Data Lake corporativo, usando la partición del cierre de mes.

df_personas          = spark.read.parquet("ruta_master/personas")
df_relacion           = spark.read.parquet("ruta_master/relacion_cliente_grupo")
df_grupos             = spark.read.parquet("ruta_master/maestro_grupos")
df_info_economica     = spark.read.parquet("ruta_master/info_economica_cliente")
df_ventas_grupo       = spark.read.parquet("ruta_master/info_economica_grupo")
df_tipo_cambio        = spark.read.parquet("ruta_master/tipo_cambio")
df_param_regulatorio  = spark.read.parquet("ruta_master/parametria_regulatoria")
df_pyme_local         = spark.read.parquet("ruta_master/indicador_pyme_local")
df_segmentacion       = spark.read.parquet("ruta_master/maestro_segmentacion")
```

Figura 16 Lectura de fuentes. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark

La lectura desde la capa *Master* en formato PARQUET permite aprovechar el almacenamiento columnar y la compresión nativa de Spark, reduciendo significativamente el tiempo de carga frente a otros formatos.

b) Integración de fuentes mediante operaciones de cruce

Una vez cargadas las fuentes, se realizan operaciones de cruce distribuido (*join*) que permiten consolidar, alrededor del cliente individual, toda la información económica, estructural y regulatoria requerida para el cálculo del indicador.

La integración respeta la lógica de cardinalidad definida durante el Sprint 2: la base de personas actúa como tabla pivot y se enriquece mediante *left join* con la información económica del cliente; el vínculo cliente–grupo se resuelve mediante un *left join* contra la tabla de relación, y posteriormente se incorpora la información consolidada del grupo económico mediante un *left join* sobre el identificador de grupo.

En la figura 17 se puede observar el ejemplo conceptual de integración de las fuentes.

```
# Se construye la base consolidada cliente-grupo mediante left joins
# sobre el identificador de cliente y el identificador de grupo.

df_base = (
    df_personas
        .join(df_info_economica, "id_cliente", "left")
        .join(df_relacion, "id_cliente", "left")
        .join(df_grupos, "id_grupo", "left")
        .join(df_ventas_grupo, "id_grupo", "left")
)
```

Figura 17 Ejemplo de integración. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark

Este proceso permite construir un único *DataFrame* consolidado por cliente que contiene todas las variables de entrada necesarias para la evaluación posterior de las reglas paramétricas.

A modo de síntesis, la Tabla 10 resume los cruces principales aplicados en esta etapa.

Tabla 10. Operaciones de cruce aplicadas durante la consolidación

Fuente 1	Fuente 2	Tipo de join	Llave de cruce	Resultado
Personas	Información económica	Left join	id_cliente	Variables económicas individuales
Personas	Relación cliente–grupo	Left join	id_cliente	Identificación de grupo
Relación	Maestro de grupos	Left join	id_grupo	Atributos del grupo económico

Maestro de grupos	Ventas grupo	Left join	id_grupo	Volumen de ventas consolidado
----------------------	--------------	-----------	----------	----------------------------------

Fuente. Elaboración propia.

Nota. Los *left join* preservan la totalidad de los clientes de la base de personas, asignando valores nulos cuando no existe relación con un grupo económico; estos casos se tratan posteriormente en la etapa de normalización.

c) Aplicación de transformaciones y cálculos intermedios

Sobre la base consolidada se aplican las transformaciones necesarias para llevar las variables de entrada a una representación homogénea y comparable contra los umbrales regulatorios.

Las principales transformaciones implementadas son las siguientes:

- Conversión de montos económicos a euros (EUR), aplicando el tipo de cambio oficial publicado por el regulador en la fecha de cierre.
- Imputación del volumen de ventas: se utiliza el volumen consolidado del grupo económico cuando el cliente pertenece a un grupo, y el volumen individual cuando el cliente opera de forma independiente.
- Generación de banderas auxiliares (*flag_**) que materializan las condiciones intermedias requeridas por la lógica de clasificación.
- Normalización de atributos regulatorios y tratamiento de valores nulos.

En la figura 18 se puede observar el ejemplo conceptual de conceptual de transformación monetaria-

```
# Se calcula el tipo de cambio oficial y se convierten las ventas
# y los activos del cliente a la moneda de evaluación regulatoria (EUR).

df_tc = df_tipo_cambio.withColumn(
    "tipo_cambio_eur",
    col("cotizacion_fix") / col("unidad_cotizacion")
).filter(col("moneda") == "EUR")

df_base = (
    df_base
    .join(df_tc, "fecha_cierre", "left")
    .withColumn("ventas_eur", col("ventas_local") / col("tipo_cambio_eur"))
    .withColumn("activos_eur", col("activos_local") / col("tipo_cambio_eur"))
)
```

Figura 18 Transformación monetaria. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark

Estas transformaciones dejan los DataFrames listos para la fase de evaluación, donde se aplican las reglas de clasificación por bloques. En la

Tabla 11, se muestra el detalle de las transformaciones necesarias que fueron aplicadas para cumplir con la lógica definida.

Tabla 11. Transformaciones aplicadas durante la normalización

Transformación	Descripción
Conversión monetaria	Normalización de ventas y activos a EUR
Imputación grupo / individual	Selección del volumen aplicable según pertenencia a grupo
Generación de banderas	Creación de variables intermedias por bloque decisional
Tratamiento de nulos	Manejo de campos no informados en fuentes de origen

Fuente. Elaboración propia.

d) Implementación de la lógica de clasificación

Una vez normalizadas las variables intermedias se implementa la lógica de clasificación definida en el Sprint 2 (Figura 14), evaluando los tres bloques decisionales en PySpark mediante condiciones distribuidas sobre el clúster.

A continuación, se presenta la implementación conceptual de cada uno de los tres bloques, manteniendo la trazabilidad con la regla funcional formalizada al cierre del Sprint 2.

Bloque 1: Variables económicas

En este bloque se evalúa los umbrales cuantitativos de la Recomendación 2003/361/CE de la Comisión Europea: número de empleados estrictamente menor a 250, junto con la condición de ventas anuales menores o iguales a 50 millones de euros o, alternativamente, activos totales menores o iguales a 43 millones de euros.

La disyunción entre ventas y activos se implementa como un operador lógico OR, mientras que la condición sobre empleados se aplica como AND. Esto se puede observar en la Figura 19.

```
# Bloque 1 - Evaluación de variables económicas.
# Cumple si tiene menos de 250 empleados Y además
# (ventas <= 50MM EUR O activos <= 43MM EUR).

UMBRAL_EMPLEADOS = 250
UMBRAL_VENTAS     = 50_000_000 # EUR
UMBRAL_ACTIVOS    = 43_000_000 # EUR

df_base = df_base.withColumn(
  "flag_bloque1",
  when(
    (col("nro_empleados") < UMBRAL_EMPLEADOS) &
    (
      (col("ventas_eur") <= UMBRAL_VENTAS) |
      (col("activos_eur") <= UMBRAL_ACTIVOS)
    ),
    lit("Y")
  ).otherwise(lit("N"))
)
```

Figura 19 Bloque 1 - Evaluación de variables económicas. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark

Bloque 2: Variables regulatorias

Para el segundo bloque se verifica que el cliente sea una persona jurídica (tipo_persona = "GM" y estado activo) y que se encuentre identificado como Pyme local conforme a la parametría regulatoria vigente, abarcando las categorías Microempresa (MI), Pequeña Empresa (PE) y Mediana Empresa (ME) definidas en el repositorio paramétrico de la entidad. Esto se puede observar en la Figura 20.

```
# Bloque 2 - Evaluación de variables regulatorias.
# Cumple si es persona jurídica activa Y está clasificado
# como Pyme local (MI, PE o ME) en la parametría regulatoria.

CATEGORIAS_PYME_LOCAL = ["MI", "PE", "ME"]

df_base = (
  df_base
  .join(df_pyme_local, "id_cliente", "left")
  .withColumn(
    "flag_bloque2",
    when(
      (col("tipo_persona") == "GM") &
      (col("estatus") == "A") &
      (col("categoria_pyme_local").isin(CATEGORIAS_PYME_LOCAL)),
      lit("Y")
    ).otherwise(lit("N"))
  )
)
```

Figura 20 Bloque 2 - Evaluación de variables regulatorias. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark

Bloque 3: Variables estructurales

Finalmente, en el tercer bloque, se definen los criterios de exclusión asociados a la naturaleza estructural del cliente. Se exigen cuatro condiciones simultáneas: no pertenecer a una operación intragrupo, no ser una cámara de compensación, no formar parte de un grupo corporativo de la entidad y no estar tipificado bajo ninguna de las categorías de exclusión CORASU

(sectores soberano, multilateral, público, intermediarios de valores y sistema financiero). Esto se puede observar en la Figura 21.

```
# Bloque 3 – Evaluación de variables estructurales.
# Cumple si el cliente NO está marcado como intragrupo,
# NO es cámara de compensación, NO pertenece a un grupo corporativo
# y NO está tipificado bajo categorías CORASU de exclusión.

CORASU_EXCLUSION = [
    "SOB", # Sector soberano
    "MUL", # Organismos multilaterales
    "SPU", # Sector público
    "INV", # Intermediarios de valores
    "SFI", # Sistema financiero
]

df_base = df_base.withColumn(
    "flag_bloque3",
    when(
        (col("flag_intragrupo") != "Y") &
        (col("flag_camara_comp") != "Y") &
        (col("flag_grupo_corp") != "Y") &
        (~col("categoria_corasu").isin(CORASU_EXCLUSION)),
        lit("Y")
    ).otherwise(lit("N"))
)
```

Figura 21 Bloque 3 - Evaluación de variables estructurales. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark

Asignación final del indicador: Sobre la base de las tres banderas intermedias se aplica la regla compuesta definida en el Sprint 2: el cliente se clasifica como Pyme cuando los tres bloques se cumplen simultáneamente, tal como se observa en la Figura 22.

```
# Asignación final del indicador Marca Pyme.
# El cliente recibe Y cuando se cumplen los tres bloques en simultáneo.

df_resultado = df_base.withColumn(
    "marca_pyme",
    when(
        (col("flag_bloque1") == "Y") &
        (col("flag_bloque2") == "Y") &
        (col("flag_bloque3") == "Y"),
        lit("Y")
    ).otherwise(lit("N"))
)
```

Figura 22 Asignación final indicador Marca Pyme. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark

La trazabilidad entre los tres bloques decisionales y el flujo lógico representado en la Figura 14 permite que cualquier cambio futuro en la parametría regulatoria pueda incorporarse modificando únicamente las constantes asociadas (UMBRAL_VENTAS, UMBRAL_ACTIVOS, CATEGORIAS_PYME_LOCAL, CORASU_EXCLUSION), sin afectar la estructura general del servicio.

e) Validación preliminar del resultado del indicador

Una vez generado el indicador se realizan verificaciones preliminares sobre el *DataFrame* resultante con el fin de asegurar que la lógica implementada es coherente con el comportamiento esperado antes de avanzar a la validación formal del Sprint 4.

Las validaciones preliminares aplicadas incluyen:

- Conteo total de registros generados versus el universo de clientes de entrada, verificando que no se hayan duplicado o descartado registros durante los cruces.
- Distribución del indicador (Y / N) y comparación referencial contra los volúmenes esperados por el área de negocio.
- Inspección manual de muestras representativas, validando casos típicos: cliente Pyme individual, cliente Pyme con grupo económico, cliente excluido por CORASU y cliente excluido por umbrales económicos.

En la Figura 23, se puede apreciar un ejemplo de validación preliminar con el resultado de la distribución del indicador Marca Pyme.

```
# Validación preliminar – distribución del indicador Marca Pyme
# sobre el DataFrame resultante del Sprint 3.

(df_resultado.groupBy("marca_pyme")
               .count()
               .withColumnRenamed("count", "cantidad")
               .orderBy("marca_pyme", ascending=False)
               .show())
```

Out[3]:

marca_pyme	cantidad
Y	75 000
N	12 125 000

clientes clasificados como Pyme
clientes no Pyme

TOTAL UNIVERSO : 12 200 000
PROPORCIÓN Y/N : 0.6 % / 99.4 %

Figura 23 Distribución preliminar del indicador Marca Pyme sobre el universo de clientes corporativos. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark

Estas verificaciones permiten confirmar que la implementación inicial en el entorno *Sandbox* genera resultados consistentes con la regla funcional definida y habilitan el avance hacia las pruebas de validación formal con el área de negocio.

Entregables del Sprint 3

Al cierre del Sprint, se obtiene lo siguiente: la implementación inicial del servicio en PySpark sobre el entorno *Sandbox* del Data Lake corporativo, la lectura, integración y normalización de las diez fuentes funcionales en el entorno distribuido, la implementación parametrizada de la lógica de clasificación sobre los tres bloques decisionales —económicas, regulatorias y estructurales—, la generación preliminar del campo *marca_pyme* sobre el universo de clientes corporativos de la institución financiera y la evidencia preliminar de la consistencia de la distribución Y/N del indicador resultante, lo cual habilita formalmente el inicio de la validación funcional y técnica del servicio con el área de Riesgos en el siguiente Sprint. El detalle del código fuente del servicio implementado se presenta en el **Anexo C**.

3.5.2.5 *Sprint 4: Validación funcional y técnica del indicador*

Una vez completada la implementación inicial del servicio durante el Sprint 3, en el cuarto Sprint del ciclo de desarrollo del proyecto, se aborda la validación funcional y técnica del indicador *Marca Pyme*, etapa orientada a verificar formalmente que los resultados generados sobre el entorno *Sandbox* sean consistentes, correctos y alineados con los criterios definidos por el área usuaria de Riesgos.

El objetivo central de este Sprint consiste en garantizar la calidad del servicio digital antes de su formalización productiva, certificando tanto la integridad técnica de la tabla destino como la coherencia funcional del comportamiento del indicador frente a las reglas de negocio que motivaron su construcción.

En ese sentido, las actividades desarrolladas se estructuran en cuatro frentes complementarios: la validación de integridad estructural de la tabla destino, la validación de consistencia global del indicador a nivel agregado, la validación granular a nivel de registro mediante muestras representativas y la validación funcional formal con el área usuaria de Riesgos como cierre del proceso de aseguramiento de calidad del servicio.

A continuación, se detalla cada actividad desarrollada.

a) Validación de integridad estructural de la tabla destino

La primera validación tiene como objetivo verificar que la incorporación del campo *marca_pyme* no altera la estructura original de la tabla maestra de segmentación de clientes. Se confirma que el conteo total de registros se mantiene constante antes y después de la ejecución del servicio, que ningún campo distinto al campo destino resulta modificado y que la integridad del esquema de la tabla —tipos de dato, particionamiento por fecha de cierre y llave primaria— se preserva tal como fue especificado durante el Sprint 2. En la Figura 24 se muestra esta comparación en conteo de registros.

```
# Comparación de conteo de registros antes y después de la
# incorporación del campo marca_pyme en la tabla destino.

total_antes    = df_segmentacion.count()
total_despues  = df_resultado.count()

print(f"Registros antes : {total_antes:,}")
print(f"Registros después : {total_despues:,}")
print(f"Diferencia : {total_despues - total_antes:,}")
```

Figura 24 Comparación de registros. Fuente. Elaboración propia a partir de la ejecución en el entorno PySpark.

El resultado esperado es una diferencia igual a cero, condición que confirma que el servicio actúa exclusivamente como un proceso de enriquecimiento de la tabla destino y no como un proceso de creación o eliminación de registros.

b) Validación de consistencia del indicador

La segunda validación analiza el comportamiento agregado del indicador sobre el universo total de clientes corporativos, verificando que la distribución de valores generados resulte consistente con las expectativas previas del área usuaria. Se evalúa la cantidad absoluta y la proporción relativa de clientes clasificados como Pyme (Y) y como no Pyme (N), comparando los volúmenes obtenidos contra los rangos referenciales acordados con Riesgos durante el Sprint Planning.

La distribución resultante se presenta en la Figura 22 (sección anterior), donde se evidencia que aproximadamente el 0.6% del universo de clientes corporativos califica bajo el indicador *Marca Pyme*, proporción consistente con

la composición esperada de la cartera de la institución financiera y validada referencialmente con el área de Riesgos.

c) Validación a nivel registro

La tercera validación complementa el análisis agregado mediante una revisión detallada sobre muestras representativas, con el objetivo de asegurar que la lógica aplicada genera el resultado esperado caso a caso, que los valores intermedios coinciden con las reglas paramétricas definidas y que no se presentan inconsistencias en los cruces entre fuentes funcionales. La selección de muestras se realiza junto con el área usuaria, priorizando casos típicos de cada bloque decisional —cliente Pyme individual, cliente Pyme con grupo económico, cliente excluido por umbrales económicos y cliente excluido por categoría CORASU—, así como casos de frontera donde alguno de los criterios opera en su límite.

En la Tabla 12 se presentan cinco casos representativos seleccionados durante esta validación, ilustrando el comportamiento del indicador frente a perfiles distintos de clientes.

Tabla 12. Validación por registro: casos representativos del comportamiento del indicador

Perfil del cliente	Bloque 1	Bloque 2	Bloque 3	Resultado	Criterio determinante
Pyme individual sin grupo económico	Y	Y	Y	Y	Cumple los tres bloques
Pyme con grupo económico (volumen consolidado dentro de umbral)	Y	Y	Y	Y	Cumple por consolidación de grupo
Ciente con número de empleados ≥ 250	N	Y	Y	N	Falla Bloque 1 (umbral económico)
Ciente tipificado como Sector Público (CORASU "SPU")	Y	Y	N	N	Falla Bloque 3 (CORASU de exclusión)

Cliente perteneciente a un grupo corporativo de la entidad	Y	Y	N	N	Falla Bloque 3 (grupo corporativo)
--	---	---	---	---	--

Fuente. Elaboración propia.

Nota. Los casos presentados corresponden a perfiles representativos seleccionados durante la sesión de validación con el área usuaria de Riesgos, anonimizados con fines de confidencialidad institucional. Cada caso ilustra cómo el indicador evalúa los tres bloques decisionales y trazando el criterio determinante que define el resultado final.

Este nivel de validación permite documentar la trazabilidad lógica del indicador y construir un repositorio de casos de prueba que podrá ser reutilizado en futuras revisiones del servicio o en pruebas de regresión ante cambios en la parametría regulatoria.

d) Validación funcional con el área de negocio

Como cierre del proceso de validación se realiza una sesión formal con el área usuaria de Riesgos, en la cual se presentan de manera consolidada los resultados de las tres validaciones previas —integridad estructural, consistencia global y muestras a nivel de registro— y se revisan en conjunto los criterios funcionales que motivaron la construcción del indicador. Durante la sesión se valida la coherencia de la clasificación frente a los casos de uso del área, se confirman los casos representativos del negocio y se verifica el correcto tratamiento de las reglas específicas del modelo, incluyendo los criterios de exclusión CORASU y la disyunción ventas-activos de la Recomendación 2003/361/CE.

Como resultado de esta sesión, el área usuaria de Riesgos otorga la conformidad funcional del indicador, certificando que el comportamiento del servicio responde a los criterios definidos durante el Sprint Planning y habilitando formalmente el avance hacia las etapas posteriores de formalización técnica y despliegue productivo.

Entregables del Sprint 4

Al cierre del Sprint, se obtiene lo siguiente: la validación documentada de la integridad estructural de la tabla maestra de segmentación de clientes, la verificación de la consistencia global del indicador a nivel agregado mediante el análisis de su distribución sobre el universo de clientes corporativos, la validación granular del comportamiento del indicador sobre muestras representativas de los tres bloques decisionales —incluyendo casos típicos y casos de frontera—, y la conformidad funcional formal otorgada por el área usuaria de Riesgos sobre el comportamiento del servicio, lo cual habilita formalmente el inicio de la formalización técnica del indicador y de su gobernanza dentro del modelo de datos corporativo en el siguiente Sprint.

El detalle completo de la suite de validaciones técnicas aplicadas sobre la tabla destino, incluyendo los criterios de aceptación, los resultados obtenidos por dimensión y el estado de conformidad de cada prueba, se presenta en el **Anexo D**.

3.5.2.6 *Sprint 5: Formalización técnica y gobernanza del dato*

Una vez certificada la calidad funcional y técnica del servicio durante el Sprint 4, en el quinto Sprint del ciclo de desarrollo del proyecto, se aborda la formalización técnica del indicador *Marca Pyme* y su alineación con los lineamientos de gobierno de datos de la institución financiera, etapa orientada a transformar la lógica validada en un activo institucional reutilizable, documentado y trazable. El objetivo central de este Sprint consiste en garantizar la correcta interpretación, mantenimiento y futura evolución del indicador dentro de la arquitectura Big Data corporativa, formalizando la transferencia desde el equipo funcional hacia el equipo de desarrollo responsable de su implementación productiva.

Bajo este contexto, las actividades desarrolladas se estructuran en cinco frentes secuenciales: la elaboración del dictamen técnico como documento institucional consolidado del indicador, el registro y gestión formal del requerimiento sobre la herramienta de gestión Jira, la validación del indicador con el equipo de gobierno de datos, la transferencia formal de la lógica al

equipo de desarrollo productivo y, finalmente, la gestión de trazabilidad y versionamiento que asegura el control documental del proceso completo.

A continuación, se detalla cada actividad desarrollada.

a) Elaboración del dictamen técnico

La primera actividad consiste en la elaboración del documento funcional-técnico institucional —denominado dictamen técnico—, en el cual se consolida la totalidad de la información requerida para la implementación productiva del indicador. El dictamen reúne, en un único documento de referencia, la definición funcional del indicador con su nombre lógico y técnico, la descripción detallada de la lógica de negocio sustentada en la Recomendación 2003/361/CE de la Comisión Europea, el inventario de las diez fuentes de datos utilizadas, las reglas de transformación aplicadas durante las etapas de consolidación, normalización y evaluación, las llaves de cruce empleadas en cada operación de *join*, las consideraciones técnicas relativas al manejo de valores nulos, conversiones monetarias y agregaciones por grupo económico, y la especificación del resultado esperado del indicador junto con su tabla destino.

La estructura formal del dictamen técnico se presenta en la Tabla 13.

Tabla 13. Estructura del dictamen técnico del indicador Marca Pyme

Sección	Descripción
Identificación del indicador	Nombre funcional, nombre técnico, código y descripción general
Fuentes de datos	Inventario de las diez fuentes funcionales utilizadas
Lógica de cálculo	Reglas de negocio, bloques decisionales y umbrales regulatorios
Transformaciones	Procesos de consolidación, normalización y evaluación aplicados
Resultado final	Definición del campo destino, tipo de dato y tabla de materialización

Fuente. Elaboración propia a partir del modelo institucional de dictamen técnico utilizado para la formalización de servicios digitales sobre la plataforma Big Data.

b) Registro y gestión del requerimiento en Jira

La segunda actividad corresponde a la gestión formal del requerimiento sobre la herramienta corporativa de gestión Jira, sobre la cual se registra y articula la trazabilidad documental de todo el proceso de desarrollo. Durante esta etapa se realiza la creación o actualización de la historia de usuario asociada al indicador, se adjunta el dictamen técnico como evidencia formal del diseño consolidado, se gestiona el seguimiento del estado de las tareas técnicas a lo largo del ciclo To Do → In Progress → Deployed y se administra la interacción entre los equipos funcionales, técnicos y de gobierno mediante los comentarios y validaciones registradas sobre el ticket. La Figura 25 presenta una vista representativa del ticket utilizado para la gestión del requerimiento.

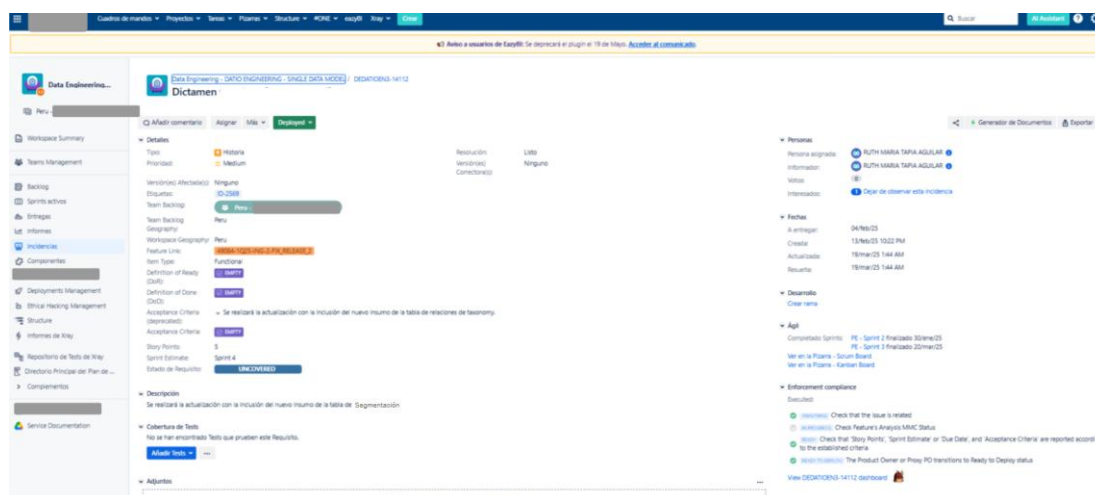


Figura 25 HU de elaboración de Dictamen. Fuente. Elaboración propia a partir de la plataforma institucional.

Nota. La imagen corresponde a una captura de la plataforma Jira utilizada para la gestión del proyecto. Los datos identificatorios de la institución y de los responsables han sido ocultados por razones de confidencialidad.

c) Validación con equipo de Data Governance (Core Assurance)

La tercera actividad consiste en la coordinación con el equipo institucional responsable del gobierno de datos para la validación formal del indicador desde la perspectiva de aseguramiento de calidad documental y modelo. Durante esta validación se evalúa la coherencia entre la lógica funcional

definida con el área usuaria de Riesgos y la lógica técnica implementada en PySpark, el uso adecuado de las fuentes de datos en términos de su origen, periodicidad y nivel de granularidad, el cumplimiento de los estándares institucionales de modelado de datos y la correcta definición del campo destino dentro del modelo de datos corporativo. Como resultado de esta validación, el indicador es aprobado desde la perspectiva de gobierno de datos y queda formalmente apto para su implementación productiva.

d) Transferencia al equipo de desarrollo

Una vez aprobado el dictamen técnico por el equipo de Aseguramiento de Datos, se realiza la transferencia formal de la lógica del indicador al equipo de desarrollo productivo, responsable de su implementación definitiva sobre el entorno de producción de la plataforma Big Data corporativa. La transferencia se materializa mediante una sesión de *handover* en la cual se realiza la explicación detallada de la lógica del indicador y de su trazabilidad con la Recomendación 2003/361/CE, se entrega formalmente el dictamen técnico como documento de referencia, se resuelven en conjunto las dudas técnicas relativas a las fuentes y a las reglas de transformación, y se establecen los acuerdos de alineación sobre la implementación productiva y sus criterios de aceptación. Como resultado, el equipo de desarrollo productivo recibe la especificación completa del indicador y queda habilitado para iniciar el empaquetamiento del proceso en formato ejecutable.

e) Gestión de trazabilidad y versionamiento

La quinta y última actividad del Sprint corresponde a la gestión integral de la trazabilidad documental y al control de versiones del proceso completo, cuya finalidad es garantizar la reproducibilidad del desarrollo y la auditabilidad del indicador a lo largo de su ciclo de vida. La herramienta Jira concentra la trazabilidad operativa, registrando los cambios efectuados sobre la lógica del indicador, las versiones sucesivas del dictamen técnico, las interacciones entre los equipos involucrados y las evidencias formales de validación y aprobación. Los principales elementos de trazabilidad gestionados durante este Sprint se presentan en la Tabla 14.

Tabla 14. Elementos de trazabilidad y versionamiento del indicador

Elemento	Descripción
Ticket Jira	Registro central del requerimiento, estados y asignaciones
Dictamen técnico	Documento institucional base del indicador, sujeto a versionamiento
Comentarios	Interacciones formales entre los equipos funcional, técnico y de gobierno
Versiones	Historial de cambios del dictamen y de la lógica del indicador

Fuente. Elaboración propia a partir del modelo de trazabilidad documental utilizado en la gestión de requerimientos sobre la plataforma Big Data.

Entregables del Sprint 5

Al cierre del Sprint, se obtiene lo siguiente: el dictamen técnico formal del indicador *Marca Pyme* como documento institucional consolidado, el registro completo del requerimiento sobre la herramienta de gestión Jira con la totalidad de evidencias de diseño y validación adjuntas, la aprobación formal del indicador por parte del equipo de Aseguramiento de Datos en términos de gobierno y modelado, la transferencia documentada de la lógica al equipo de desarrollo productivo mediante sesión de *handover* formal, y la gestión integral de la trazabilidad y versionamiento del proceso sobre la herramienta corporativa, lo cual habilita formalmente el inicio de la revalidación técnica del indicador y de su integración productiva dentro de la arquitectura Big Data institucional en el siguiente Sprint.

3.5.2.7 Sprint 6: Revalidación final e integración productiva del indicador

Una vez formalizada la transferencia técnica del servicio hacia el equipo de desarrollo productivo durante el Sprint 5, el sexto y último Sprint del ciclo de desarrollo del proyecto, se enfoca en la revalidación funcional del indicador *Marca Pyme* sobre el entorno productivo de la institución financiera y en su habilitación formal para consumo institucional. A diferencia de los Sprints anteriores —centrados en la construcción y validación del servicio dentro del entorno Sandbox—, este Sprint opera ya sobre el cluster productivo del Data

Lake corporativo, donde la lógica de negocio empaquetada en el archivo ejecutable (.jar) entregado por el equipo de desarrollo productivo es ejecutada por el motor distribuido de la institución bajo las mismas condiciones de scheduling, particionamiento y consumo que tendrá en producción regular.

Bajo este contexto, las actividades desarrolladas se estructuran en cinco frentes complementarios: (a) la recepción y revisión del entregable productivo construido por el equipo de desarrollo, (b) la ejecución controlada del servicio en el entorno productivo sobre la partición de cierre evaluada, (c) la validación de equivalencia funcional entre los resultados obtenidos en Sandbox y los resultados obtenidos en producción, (d) la formalización del despliegue del servicio dentro del flujo de procesamiento institucional y (e) el cierre formal del desarrollo del proyecto mediante la habilitación del indicador para consumo por parte de las áreas usuarias.

a) Recepción del desarrollo productivo (archivo JAR)

El equipo de desarrollo productivo entrega al equipo funcional el archivo ejecutable (.jar) que empaqueta la lógica del servicio *Marca Pyme*, junto con la documentación técnica de despliegue y los parámetros de ejecución requeridos por el *scheduler* corporativo. El archivo es revisado en términos de versión, dependencias y firma técnica, y se coteja contra el dictamen técnico emitido en el Sprint 5 para asegurar que la lógica empaquetada corresponde íntegramente a la versión validada en *Sandbox*. La revisión es registrada en el sistema de gestión de tickets de la institución bajo el mismo identificador del proyecto.

b) Ejecución controlada en entorno productivo

Una vez validada la recepción, el servicio es ejecutado en modo controlado sobre el *cluster* productivo del Data Lake corporativo, tomando como entrada la misma partición de cierre utilizada para la validación preliminar en *Sandbox*. La ejecución se realiza en coordinación con el equipo de Operaciones del Data Lake y bajo la supervisión del equipo de Aseguramiento de Datos, conservando la trazabilidad de los identificadores de ejecución (*application_id*) y los tiempos de procesamiento reportados por el motor distribuido.

c) Validación de equivalencia funcional Sandbox–Producción

Concluida la ejecución productiva, se compara la salida del servicio en producción contra la salida obtenida en *Sandbox* sobre el mismo periodo de cierre, con el objetivo de verificar que ambas ejecuciones producen un resultado funcionalmente equivalente. La comparación se realiza sobre tres dimensiones: (i) el conteo total de registros generados, (ii) la distribución del indicador *Marca Pyme* entre las categorías "Y" y "N" y (iii) la coincidencia registro a registro de la asignación de marca para una muestra estratificada de clientes. La validación arroja una equivalencia funcional plena, sin diferencias atribuibles a la lógica del servicio, lo cual confirma que el comportamiento observado en producción reproduce íntegramente el comportamiento validado durante los Sprints anteriores. La evidencia de esta validación se presenta en la Figura 23.

d) Formalización del despliegue productivo

Verificada la equivalencia funcional, el servicio es formalmente incorporado al flujo de procesamiento institucional, registrándose dentro del scheduler corporativo bajo la frecuencia mensual definida por el área de Riesgos. La formalización incluye la actualización del catálogo técnico del Data Lake corporativo, donde el campo *marca_pyme* queda registrado como atributo oficial de la tabla maestra de segmentación, así como el registro del servicio dentro del repositorio institucional de procesos productivos, con los metadatos asociados a su versión, propietario funcional, propietario técnico y referencia de trazabilidad.

e) Cierre del de desarrollo del proyecto y habilitación institucional

Como cierre del ciclo de desarrollo del proyecto, se realiza la última reunión de revisión con el área usuaria y el equipo de Riesgos, en la cual se presenta la evidencia consolidada de la construcción, validación y despliegue del servicio.

En dicha reunión se confirma la habilitación formal del indicador *Marca Pyme* para consumo por parte de las áreas internas que dependen de su

disponibilidad, entre ellas Riesgos, Reporting Regulatorio y Banca de Empresas, dándose por finalizado el ciclo completo de los seis Sprints conforme a los compromisos establecidos en la planificación inicial.

Entregables del Sprint 6

Al cierre del Sprint, se obtiene lo siguiente: el archivo ejecutable (.jar) del servicio *Marca Pyme* recibido, validado y desplegado en el entorno productivo de la institución financiera; la evidencia de equivalencia funcional entre las ejecuciones realizadas en *Sandbox* y en producción sobre la partición de cierre evaluada; la formalización del campo *marca_pyme* dentro del catálogo técnico del Data Lake corporativo y dentro del repositorio institucional de procesos productivos; y la habilitación formal del indicador para consumo por parte de las áreas usuarias, lo cual habilita formalmente el cierre del ciclo de desarrollo del proyecto y el inicio del régimen regular de procesamiento mensual del servicio dentro del flujo institucional de la entidad.

3.5.2.8 *Sprint Review y Retrospectiva - Evaluación y cierre del ciclo de desarrollo del proyecto*

Al cierre de los seis Sprints planificados, y conforme al marco metodológico Scrum adoptado por la institución financiera para la gestión de proyectos analíticos sobre el Data Lake corporativo, se desarrollan las dos ceremonias formales de cierre del ciclo de desarrollo del proyecto: la *Sprint Review* y la *Sprint Retrospective*. Ambas ceremonias se realizan de manera diferenciada, la primera orientada al producto entregado y la segunda orientada al proceso de trabajo del equipo, y permiten consolidar las lecciones aprendidas durante el ciclo completo de construcción del indicador *Marca Pyme*.

Bajo este contexto, las dos ceremonias se desarrollan de manera secuencial dentro de la última semana del ciclo, con la participación del equipo funcional de Management Solutions, el equipo de Riesgos de la institución financiera en su rol de área usuaria, el equipo de Aseguramiento de Datos y el equipo de desarrollo productivo, en coherencia con la matriz de responsabilidades establecida desde la planificación inicial.

a) Sprint Review

La Sprint Review es conducida sobre la totalidad del ciclo de desarrollo del proyecto y no sobre un Sprint individual, dada la naturaleza acumulativa del entregable construido. Durante la sesión se presenta a las áreas usuarias el recorrido completo del servicio, desde la consolidación inicial de las fuentes funcionales hasta el despliegue productivo del archivo ejecutable (.jar), apoyándose en la evidencia documental generada en cada Sprint, diagramas lógicos, esquemas de arquitectura, validaciones de volumetría y unicidad, distribución preliminar del indicador, dictamen técnico y registros de despliegue. La revisión concluye con la aceptación formal del entregable por parte del área de Riesgos como propietaria funcional del indicador y por parte del equipo de Aseguramiento de Datos como garante de la calidad del proceso.

b) Sprint Retrospectiva

La Sprint Retrospective se desarrolla en una sesión separada con el equipo funcional, orientada a identificar los aciertos y las oportunidades de mejora del proceso de trabajo a lo largo del ciclo de desarrollo del proyecto. Los principales aciertos identificados incluyen la consolidación temprana del catálogo de fuentes en el Sprint 1, lo cual evitó retrabajos en etapas posteriores; el uso de bloques decisionales independientes en el Sprint 3, que facilitó el aislamiento y la trazabilidad de cada criterio de clasificación; y la transferencia técnica formal hacia el equipo de desarrollo productivo en el Sprint 5, que permitió un despliegue productivo sin incidencias funcionales. Como oportunidades de mejora se identifica la conveniencia de reforzar el versionamiento documental desde el inicio del ciclo de desarrollo del proyecto y de anticipar la coordinación con el equipo de Operaciones del Data Lake corporativo para el agendamiento de las ejecuciones controladas.

c) Cierre formal del ciclo de desarrollo del proyecto

Concluidas ambas ceremonias, el ciclo de desarrollo del proyecto se da por formalmente cerrado mediante la actualización del estado de los tickets correspondientes en el sistema de gestión del proyecto y la emisión del acta

de cierre suscrita por el equipo funcional y el área usuaria. El indicador *Marca Pyme* queda incorporado al régimen regular de procesamiento mensual del Data Lake corporativo, con su lógica de negocio empaquetada, validada, documentada, gobernada y desplegada conforme a los estándares técnicos y funcionales de la institución financiera.

Resultado consolidado

Al cierre del ciclo de desarrollo del proyecto, se obtiene lo siguiente: un servicio analítico en producción que materializa el indicador *Marca Pyme* sobre la tabla maestra de segmentación de clientes corporativos del Data Lake institucional; un cuerpo documental completo que comprende el catálogo de fuentes funcionales, los esquemas de arquitectura lógica, los diagramas de flujo, las validaciones técnicas, el dictamen formal y los registros de trazabilidad y versionamiento; y la habilitación efectiva del indicador para consumo por parte de las áreas usuarias dependientes (Riesgos, Reporting Regulatorio y Banca de Empresas), lo cual habilita formalmente la conclusión del proyecto bajo los términos de alcance, calidad y plazo establecidos en la planificación inicial del ciclo de desarrollo del proyecto.

3.6 Resultado de la actividad

Al concluir el proyecto se obtiene como resultado el diseño, implementación y validación del servicio digital de valor agregado constituido por el indicador *Marca Pyme* para la identificación de clientes del segmento Pyme, el cual fue integrado exitosamente dentro de la arquitectura Big Data distribuida de la institución financiera.

El indicador *marca_pyme* clasifica a los clientes en función de criterios económicos, regulatorios y estructurales, utilizando múltiples fuentes de datos integradas desde el Data Lake institucional. Su implementación se realizó mediante procesamiento distribuido con PySpark, garantizando la escalabilidad y eficiencia requeridas para el manejo de grandes volúmenes de información.

A nivel técnico, se validó la viabilidad del procesamiento distribuido desde la lectura de los datos en HDFS hasta la generación del indicador final, asegurando la correcta integración de fuentes, la aplicación de las reglas de negocio definidas y la consistencia de los resultados entre los entornos Sandbox y producción. La comparación entre ambos entornos permitió confirmar la integridad y confiabilidad del servicio digital implementado.

Asimismo, el indicador fue formalizado bajo lineamientos de gobierno de datos mediante la elaboración del dictamen técnico y su gestión a través de la herramienta Jira, lo cual garantiza su trazabilidad, mantenimiento evolutivo y reutilización dentro de la organización. Posteriormente se realizó su integración en la tabla maestra de segmentación, quedando disponible como un atributo transversal para distintos sistemas y áreas de negocio.

Como parte del despliegue final se ejecutó el reprocesamiento histórico de la información, permitiendo que el indicador se encuentre disponible de manera consistente para todos los periodos de análisis, lo cual habilita su uso en procesos analíticos, regulatorios y de toma de decisiones.

Desde la perspectiva de telecomunicaciones, el servicio digital se sustenta sobre la red de datos corporativa de la institución financiera, la cual constituye el sustrato físico y lógico que hace posible la operación del cluster distribuido. Dicha red corresponde a una red corporativa privada de tipo enterprise, organizada bajo una arquitectura jerárquica de tres capas: acceso, distribución y núcleo, con segmentación lógica mediante esquemas de aislamiento de tráfico (VLAN) que separan los flujos de gestión, datos productivos y replicación interna del Data Lake. La interconexión entre los nodos del cluster se realiza sobre enlaces Ethernet de alta capacidad bajo la pila de protocolos TCP/IP, condición indispensable para el desempeño de las operaciones de redistribución de datos (shuffle) que el motor distribuido ejecuta durante las transformaciones del servicio. La conectividad entre sedes y centros de datos se apoya en enlaces WAN gestionados con acuerdos de

nivel de servicio, y la totalidad del entorno opera bajo los controles de seguridad y aislamiento exigidos por la regulación del sector financiero.

Bajo el modelo de referencia OSI, el servicio digital implementado se posiciona en la capa de aplicación, apoyándose en las capas inferiores de transporte y red para asegurar la transmisión eficiente, ordenada y confiable de los datos dentro de la infraestructura distribuida. Esta dependencia evidencia el rol habilitador de la red de comunicaciones como soporte de los servicios basados en Big Data: sin una red corporativa adecuadamente dimensionada, segmentada y disponible, la ejecución coordinada de los procesos distribuidos sobre el Data Lake no resultaría viable.

En consecuencia, se logró no solo la implementación de un indicador funcional, sino la construcción de un servicio digital completo, escalable y alineado con la arquitectura tecnológica institucional, el cual aporta valor a la organización al facilitar la identificación del segmento Pyme y mejorar la disponibilidad de información confiable para la toma de decisiones. En la Figura 26, se puede apreciar como se ve el indicador integrado en la tabla maestra de segmentación.

```
# Lectura de la tabla maestra de segmentación tras la integración productiva del indicador.
# Se selecciona una muestra representativa para evidenciar la presencia del campo marca_pyme.

(df_segmentacion
 .select("id_cliente", "fecha_cierre", "tipo_titular_sra",
        "marca_pyme", "id_segmento_actual", "id_segmento_global")
 .filter(col("fecha_cierre") == "2023-12-31")
 .show(10, truncate=False))
```

id_cliente	fecha_cierre	tipo_titular_sra	marca_pyme	id_segmento_actual	id_segmento_global
CLT-00041287	2023-12-31	T	N	SEG-CORP-02	GLB-CORP-A
CLT-00041288	2023-12-31	T	Y	SEG-PYME-01	GLB-PYME-B
CLT-00041289	2023-12-31	C	N	SEG-CORP-04	GLB-CORP-A
CLT-00041290	2023-12-31	T	Y	SEG-PYME-02	GLB-PYME-B
CLT-00041291	2023-12-31	T	N	SEG-CORP-01	GLB-CORP-A
CLT-00041292	2023-12-31	C	N	SEG-CORP-03	GLB-CORP-A
CLT-00041293	2023-12-31	T	N	SEG-CORP-02	GLB-CORP-A
CLT-00041294	2023-12-31	T	Y	SEG-PYME-01	GLB-PYME-B
CLT-00041295	2023-12-31	T	N	SEG-CORP-05	GLB-CORP-A
CLT-00041296	2023-12-31	C	N	SEG-CORP-02	GLB-CORP-A

only showing top 10 rows

Figura 26 Vista del indicador Marca Pyme integrado en la tabla maestra de segmentación. Fuente. Elaboración propia a partir de la consulta realizada en el entorno productivo del Data Lake.

CAPITULO IV: DISCUSION DE RESULTADOS

4.1. Descripción de la implementación

El presente capítulo discute los resultados obtenidos durante el desarrollo e implementación del indicador Marca Pyme, organizándolos en seis etapas que recogen, de manera consolidada, los hitos alcanzados a lo largo de los seis Sprints descritos en el Capítulo III. El propósito es presentar de forma agregada los logros técnicos, funcionales y de gobierno asociados a la construcción del servicio digital, así como destacar las contribuciones significativas del proyecto a la organización.

4.1.1. Etapa de análisis y diagnóstico (Sprint 1)

- Identificación y caracterización de fuentes de datos: se logra consolidar el catálogo de las diez fuentes funcionales necesarias para la construcción del indicador, validando su disponibilidad, estructura y accesibilidad sobre la capa Master del Data Lake corporativo de la institución financiera. Esto permite asegurar la viabilidad técnica del procesamiento distribuido y eliminar dependencias bloqueantes en etapas posteriores.
- Validación de periodicidad y consistencia de la información: se confirma que las fuentes seleccionadas presentan una periodicidad alineada con el cierre mensual requerido por el área de Riesgos, garantizando la coherencia temporal del indicador y la trazabilidad de los cierres procesados.

4.1.2. Etapa de diseño lógico del servicio (Sprint 2)

- Definición de la arquitectura lógica del indicador: se establece el flujo de procesamiento bajo el modelo Big Data distribuido, definiendo la integración de las fuentes consolidadas, las transformaciones requeridas y la lógica de clasificación del campo *marca_pyme*.
- Estructuración de variables y modelo de clasificación: se organizan los atributos económicos, regulatorios y estructurales del cliente en tres bloques decisionales independientes, lo cual habilita el aislamiento y la trazabilidad de cada criterio dentro del flujo del servicio.
- Formalización de las reglas de negocio: se definen las condiciones de evaluación basadas en parámetros regulatorios (Recomendación 2003/361/CE de la Unión Europea, marco BIS IV) y criterios funcionales, asegurando una clasificación consistente y alineada con los lineamientos del área usuaria.

4.1.3. Etapa de implementación en entorno sandbox y validación funcional/técnica (Sprint 3 y Sprint 4)

- Desarrollo del servicio mediante procesamiento distribuido: se implementa la lógica del indicador utilizando PySpark sobre el entorno Sandbox de la institución, integrando las fuentes funcionales y ejecutando las transformaciones distribuidas asociadas a los tres bloques decisionales.
- Generación preliminar del campo *marca_pyme*: se obtiene la primera versión del indicador sobre la partición de cierre evaluada, permitiendo la verificación de su comportamiento sobre datos reales del banco.
- Verificación de consistencia del indicador: se valida la correcta clasificación de los clientes, tanto a nivel agregado —mediante la distribución Y/N esperada por el área de Riesgos— como a nivel de registros individuales, asegurando coherencia con las reglas de negocio definidas en el Sprint 2.
- Validación funcional con el área usuaria: se obtiene la conformidad del área de Riesgos sobre la lógica del servicio y los resultados

preliminares observados, formalizando el cierre técnico de la etapa de validación en Sandbox.

4.1.4. Etapa de formalización y gobierno de datos (Sprint 5)

- Elaboración del dictamen técnico: se documenta formalmente la lógica del indicador, las fuentes utilizadas, los bloques decisionales y las reglas de transformación, asegurando su comprensión, mantenimiento y reutilización dentro de la organización.
- Gestión del proceso mediante Jira: se registra y gestiona el desarrollo del indicador en el sistema institucional de gestión de tickets, garantizando la trazabilidad documental, el control de versiones y la coordinación entre el equipo funcional, el equipo de Aseguramiento de Datos y el equipo de desarrollo productivo.
- Validación del gobierno de datos: se obtiene la validación correspondiente por parte del equipo de Aseguramiento de Datos, asegurando el cumplimiento de los estándares institucionales y habilitando la transferencia técnica hacia el entorno productivo.

4.1.5. Etapa de implementación productiva (Sprint 6)

- Despliegue del servicio en entorno productivo: se incorpora el indicador *marca_pyme* al flujo de procesamiento institucional mediante la ejecución del archivo ejecutable (.jar) construido por el equipo de desarrollo productivo, bajo el scheduler corporativo del Data Lake.
- Revalidación de equivalencia funcional: se confirma la consistencia entre los resultados obtenidos en Sandbox y en producción sobre la misma partición de cierre, evidenciando la integridad del servicio en el entorno productivo de la institución.
- Reprocesamiento histórico de la información: se habilita el indicador para todos los periodos disponibles del histórico de la tabla maestra de segmentación, permitiendo su uso en análisis retrospectivos, requerimientos regulatorios y procesos analíticos transversales.
- Disponibilización del indicador: el campo *marca_pyme* queda integrado en la tabla maestra de segmentación de clientes corporativos,

disponible para consumo por las áreas dependientes (Riesgos, Reporting Regulatorio y Banca de Empresas) como atributo transversal del Data Lake corporativo.

4.1.6. Etapa de cierre y evaluación del servicio

- Validación final del indicador: se confirma el cumplimiento de los objetivos funcionales y técnicos definidos al inicio del ciclo de desarrollo del proyecto, formalizado mediante el acta de cierre suscrita entre el equipo funcional y el área usuaria.
- Identificación de mejoras y lecciones aprendidas: se reconocen oportunidades de optimización para futuros ciclos, entre ellas el refuerzo del versionamiento documental desde fases tempranas y la coordinación anticipada con el equipo de Operaciones del Data Lake corporativo.
- Consolidación del indicador como servicio digital de valor agregado: el indicador Marca Pyme queda formalizado como un servicio digital reutilizable, alineado con la arquitectura Big Data de la institución y soportado por la infraestructura de comunicaciones corporativa, sobre la cual se sustenta la transmisión, replicación y procesamiento distribuido de los datos del servicio.

4.2. Evaluación Económica

El servicio digital desarrollado corresponde a una solución implementada de manera interna (in-house) dentro de la organización, por lo que no genera ingresos directos derivados de su comercialización. En este contexto, la evaluación económica se orienta al análisis de los costos de implementación y mantenimiento, medidos a través del esfuerzo en horas hombre (HH) de los perfiles involucrados; así como a la identificación y cuantificación referencial de los beneficios operativos y estratégicos que genera su disponibilidad. Cabe destacar que todos los valores presentados son de carácter referencial y estimado.

Al tratarse de un proyecto desarrollado en el marco de una institución financiera privada, bajo el modelo de fábrica de datos de la institución, no se dispone de cifras financieras exactas de carácter interno. Los costos por hora han sido estimados en base al mercado de tecnología y datos en el Perú para el año 2024, y el personal involucrado no contaba con dedicación exclusiva al proyecto.

4.2.1. Estimación de costos de implementación (CAPEX)

Para el desarrollo del indicador Pyme se consideraron distintos perfiles profesionales correspondientes a los equipos funcional y técnico del modelo de fábrica de datos de la institución. El esfuerzo fue estimado considerando una duración de cinco meses (20 semanas hábiles), con porcentajes de dedicación reducidos por perfil, dado que cada profesional participó de manera simultánea en otras iniciativas del portafolio de proyectos de la organización.

El porcentaje de dedicación refleja la proporción de la jornada laboral destinada a este proyecto durante el período de ejecución. A partir de dicho porcentaje y considerando una jornada estándar de 8 horas diarias y 5 días a la semana (≈ 800 HH disponibles por ciclo a dedicación completa), se estimaron las horas hombre efectivas aportadas por cada perfil. En la Tabla 15 se puede observar la estimación de acuerdo al perfil y porcentaje de dedicación.

Tabla 15. Estimación de esfuerzo e inversión por perfil

Perfil	% Dedicación	HH Estimadas	Costo por hora (USD)*	Costo total (USD)
Analista funcional	20%	1600	20	3,200
Data Engineer / Developer	25%	200	25	5,000
Especialista en Gobierno de Datos	10%	80	22	1,760
Scrum Master	8%	64	23	1,472
Total	—	504 HH	—	11,432 USD

Fuente. Elaboración propia.

4.2.2. Costos operativos (OPEX)

Dado que el servicio se implementa sobre una infraestructura Big Data ya existente dentro de la organización, no se generan costos adicionales significativos en términos de hardware, licencias de software o servicios en la nube. En la siguiente Tabla 16, se puede observar que los costos operativos anuales estimados corresponden al esfuerzo de mantenimiento evolutivo y monitoreo técnico del indicador.

Tabla 16. Estimación de costos operativos anuales (OPEX)

Concepto	Descripción	Costo referencial anual (USD)
Mantenimiento evolutivo	Actualización de lógica de negocio y ajustes normativos del indicador (~40 HH/año)	900
Monitoreo y soporte técnico	Supervisión de la ejecución de procesos batch y atención de incidencias (~24 HH/año)	540
Infraestructura y cómputo	Ejecución sobre clúster Big Data institucional ya desplegado. Sin costo incremental.	—
Total OPEX anual estimado	64 HH/año	1,440 USD

Fuente: Elaboración propia.

Nota. El bajo nivel de OPEX se explica por el modelo de reutilización de infraestructura tecnológica existente, característica propia de las soluciones desarrolladas bajo el modelo de fábrica de datos de la institución.

4.2.3. Beneficios del servicio digital

El valor económico del servicio digital no se materializa en ingresos directos, sino en beneficios operativos, estratégicos y de calidad de información que impactan de manera transversal en la organización. A continuación, se describen los principales beneficios identificados, con su respectivo sustento cuantitativo o cualitativo.

a) Reducción del tiempo en procesos de segmentación

Previo a la implementación del indicador, la clasificación de un cliente como Pyme requería procesos de consulta manual sobre múltiples fuentes de datos

dispersas en el Data Lake corporativo, lo que implicaba un esfuerzo considerable por parte de los equipos de negocio y analítica. Con la automatización del proceso mediante el indicador desarrollado, se estima que el tiempo de análisis asociado a esta tarea se reduce en aproximadamente un 70%. Considerando que los equipos usuarios realizan consultas de segmentación de forma recurrente, el ahorro estimado puede alcanzar 40 horas hombre mensuales a nivel organizacional, equivalentes a un ahorro referencial de USD 880 mensuales (a un costo promedio de USD 22/HH), lo que supone USD 10,560 anuales.

b) Disminución de reprocesos y correcciones manuales

La ausencia de un indicador centralizado y confiable generaba inconsistencias en la clasificación de clientes Pyme entre las distintas áreas usuarias, derivando en reprocesos, correcciones y discrepancias que consumían tiempo de los equipos de riesgo, productos y marketing. La estandarización del indicador dentro del Datalake corporativo elimina estas inconsistencias, al establecer una fuente única de verdad. Se estima que esta mejora reduce los reprocesos asociados en un 80%, con un impacto directo en la eficiencia operativa de al menos tres áreas consumidoras del dato (Riesgos, Productos y Marketing).

c) Disponibilidad de un dato único y reutilizable

Al formalizar el indicador como un servicio digital dentro de la plataforma Big Data institucional, se garantiza su disponibilidad para múltiples áreas consumidoras (riesgo crediticio, productos, marketing, cumplimiento normativo) sin necesidad de desarrollos paralelos o duplicados. Esta reutilización evita la inversión en proyectos redundantes: si cada área hubiera desarrollado su propia lógica de clasificación Pyme de manera independiente, el costo equivalente habría sido múltiplos del costo de implementación del indicador centralizado.

d) Reducción del riesgo operacional y cumplimiento normativo

La disponibilidad de un indicador Pyme confiable, trazable y documentado

facilita el cumplimiento de regulaciones aplicables al segmento, tales como las normativas de la Superintendencia de Banca, Seguros y AFP (SBS) sobre clasificación y tratamiento de clientes empresariales. La falta de este indicador representaba un riesgo operacional latente: una clasificación incorrecta o inconsistente podría derivar en observaciones regulatorias, aplicación de condiciones financieras inadecuadas al perfil del cliente, o pérdida de oportunidades comerciales en el segmento Pyme.

e) Soporte a la toma de decisiones estratégicas

Contar con información confiable y actualizada sobre la condición Pyme de los clientes permite a la institución diseñar productos financieros personalizados, campañas dirigidas y estrategias de atención diferenciada para este segmento. Dado que las Pymes representan una proporción significativa del PBI y del empleo en el Perú (Ministerio de la Producción, 2022), la correcta identificación de estos clientes tiene un impacto directo en la rentabilidad del portafolio comercial y en la calidad del servicio ofrecido.

4.2.4. Análisis costo-beneficio

La Tabla 17 resume los principales componentes del análisis económico del servicio digital, contrastando el costo total de implementación y operación frente al ahorro anual estimado derivado de la reducción de horas hombre en procesos manuales de segmentación.

Tabla 17. Resumen del análisis costo-beneficio

Concepto	Monto estimado (USD)
Costo de implementación (CAPEX)	11,432
Costo operativo anual (OPEX)	1,440 / año
Ahorro estimado anual (reducción de HH en análisis manual)	10,560 / año
Período estimado de recuperación de la inversión	< 2 años
Evaluación general	Favorable

Fuente: Elaboración propia.

Nota. Los valores de ahorro son referenciales, estimados en base a la reducción de horas hombre de análisis manual en procesos de segmentación,

considerando un costo promedio de USD 22/HH.

Como se puede observar en la siguiente Tabla 18, se confirma una alta rentabilidad, con una recuperación rápida (entre el año 1 y año 2, el flujo acumulado pasa de negativo a positivo) y bajo riesgo.

Tabla 18. Flujo de caja

Concepto	Año 0 (USD)	Año 1 (USD)	Año 2 (USD)	Año 3 (USD)
Inversión (CAPEX)	-11,432	0	0	0
OPEX	0	-1,440	-1,440	-1,440
Ahorros (beneficio)	0	10,560	10,560	10,560
Flujo Neto	-11,432	9,120	9,120	9,120
Flujo Acumulado	-11,432	-2,312	6,808	15,928

Fuente: Elaboración propia.

Considerando que el costo de implementación es moderado (basado íntegramente en esfuerzo humano interno), que el costo operativo anual es bajo gracias al aprovechamiento de la infraestructura Big Data ya existente, y que los beneficios operativos se extienden simultáneamente a múltiples áreas de la organización. Se concluye que la implementación del servicio digital es económicamente favorable.

El valor generado por la disponibilidad de un indicador confiable, estandarizado y reutilizable supera el costo de desarrollo, con una recuperación estimada de la inversión en un plazo inferior a dos años bajo supuestos conservadores de uso organizacional.

Asimismo, la reutilización del indicador en futuras iniciativas multiplica su valor económico, reduciendo el costo marginal de clasificación Pyme en proyectos subsiguientes.

CAPÍTULO V: CONCLUSIONES

1. El diseño e implementación del servicio digital de valor agregado para la identificación de clientes Pyme, desarrollado sobre una arquitectura Big Data distribuida, resulta técnicamente viable y organizacionalmente valioso. El indicador *Marca Pyme* logró integrarse exitosamente dentro del ecosistema de datos de la institución como un componente reutilizable, confiable y alineado con los criterios funcionales y normativos del negocio.
2. El diagnóstico y caracterización de las fuentes de datos disponibles en el Data Lake corporativo evidenció que la viabilidad del proyecto no depende únicamente de la existencia de la información, sino de la capacidad de garantizar su correcta estructura, accesibilidad, periodicidad y consistencia. En ese sentido, la identificación de valores nulos en datos históricos de información económica y la verificación de los patrones de particionamiento en HDFS permitieron anticipar restricciones técnicas y diseñar las validaciones condicionales antes de la implementación del servicio.
3. El diseño de la arquitectura lógica y del modelo de procesamiento del indicador estructuró el flujo de datos en tres etapas técnicas (consolidación cliente-grupo económico, normalización monetaria y evaluación paramétrica), asegurando una definición rigurosa de entradas, llaves de cruce y salidas; la etapa de evaluación, a su vez, se materializó mediante tres bloques decisionales independientes —económico, regulatorio y estructural— cuya conjunción define la asignación final del campo *marca_pyme*. La implementación del indicador como variable booleana mensual a nivel de cliente garantiza su compatibilidad e integración en la infraestructura Big Data distribuida.

4. La implementación del servicio mediante PySpark sobre HDFS demostró que el procesamiento distribuido es una solución eficiente para integrar fuentes de datos heterogéneas y ejecutar transformaciones complejas sobre grandes volúmenes de información financiera. La arquitectura modular adoptada permitió estructurar y transferir la lógica de negocio de manera consistente hacia el entorno productivo, reduciendo los tiempos de implementación y facilitando el mantenimiento posterior del servicio.
5. La validación funcional y técnica en entorno *Sandbox*, ejecutada tanto a nivel agregado como a nivel de registro individual, permitió comprobar la consistencia e integridad del indicador mediante la contrastación de los resultados con las métricas esperadas por el área de negocio. La coordinación con el área de Riesgos y la conformidad obtenida al cierre del Sprint 4 evidenciaron que la lógica de clasificación cumple con los criterios normativos y funcionales establecidos en la fase de diseño.
6. El despliegue del servicio en entorno productivo y la comparación de resultados con el entorno *Sandbox* confirmaron la integridad de la implementación, evidenciando la consistencia de la lógica de negocio en ambos entornos. Asimismo, el reprocesamiento histórico permitió asegurar la disponibilidad del indicador *Marca Pyme* para distintos periodos de análisis, ampliando su aplicabilidad en procesos regulatorios, analíticos y de toma de decisiones.
7. La formalización del indicador mediante el dictamen técnico, su gestión a través de Jira y la validación por parte del equipo de Aseguramiento de Datos evidenciaron la importancia de un enfoque estructurado de gobernanza para asegurar la sostenibilidad y escalabilidad del servicio. Este proceso permitió consolidar el indicador como un activo organizacional reutilizable y accesible de forma transversal dentro de la institución.

CAPÍTULO VI: RECOMENDACIONES

1. Establecer un catálogo formal de fuentes del Data Lake corporativo que documente la estructura, periodicidad, llaves de cruce y patrones de valores nulos de cada tabla, de modo que los proyectos futuros puedan anticipar restricciones técnicas desde la fase de diagnóstico sin necesidad de relevarlas desde cero en cada iniciativa.
2. Estandarizar la arquitectura de tres etapas técnicas (consolidación, normalización y evaluación) y la separación en bloques decisionales independientes como plantilla reutilizable para el desarrollo de indicadores similares dentro del modelo de fábrica de datos de la institución.
3. Incorporar pruebas unitarias automatizadas sobre las transformaciones PySpark al momento de formalizar el código del indicador, con el fin de detectar regresiones ante ajustes normativos o evolutivos de la lógica sin depender exclusivamente de la comparación manual entre entornos.
4. Formalizar un protocolo de validación funcional con el área de negocio que incluya casos de prueba predefinidos con valores esperados, acordados antes del inicio de la fase Sandbox, para reducir los ciclos de revisión y otorgar mayor trazabilidad al proceso de aprobación.
5. Ampliar el alcance del dictamen técnico para incorporar un análisis de impacto del indicador sobre los procesos y sistemas consumidores, identificando las áreas dependientes del dato y las consecuencias de una eventual modificación de la lógica, con el objetivo de fortalecer la gobernanza y facilitar la gestión de cambios a lo largo del ciclo de vida del servicio.
6. Automatizar la comparación de resultados entre el entorno Sandbox y el entorno productivo mediante un proceso de reconciliación sistemática, a fin de ampliar la cobertura de validación y reducir el tiempo dedicado a esta actividad en futuros despliegues del servicio.

CAPÍTULO VII: REFERENCIAS BIBLIOGRÁFICAS

- aipix.ai. (2025). *Servicios de valor añadido (VAS) para proveedores de servicios de internet y operadores de telecomunicaciones en 2025*. <https://aipix.ai/blog/es/servicios-de-valor-anadido-para-proveedores-de-servicios-de-internet-y-operadores-de-telecomunicaciones-en-2025/>
- Analysys Mason. (2024). *Telefónica's 30% revenue from VAS is not a blueprint for non-telecoms services*. <https://www.analysismason.com/research/content/articles/telefonica-non-telecoms-services-connectivity-rdmm0-rdmb0-rdcs0-rdvs0-ren01-rdmz0-rddg0/>
- Apache Software Foundation. (2024). *Apache Spark documentation*. <https://spark.apache.org/docs/latest/>
- Atlassian. (2024). *Jira Software documentation*. <https://confluence.atlassian.com/jirasoftware/>
- Bani, F. C. D., Suharjito, Diana, & Girsang, A. S. (2018). *Implementation of database massively parallel processing system to build scalability on process data warehouse*. *Procedia Computer Science*, 135, 68–79. <https://doi.org/10.1016/j.procs.2018.08.151>
- Bank for International Settlements. (2017). *Basel III: Finalising post-crisis reforms*. Basel Committee on Banking Supervision. <https://www.bis.org/bcbs/publ/d424.htm>
- Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), 171–209. <https://doi.org/10.1007/s11036-013-0489-0>
- Cisco Systems. (2017). *The next era of digital financial services* [Solution overview]. Cisco. <https://www.cisco.com/c/dam/en/us/solutions/collateral/enterprise-networks/next-era-digital-financial-services.pdf>
- Comisión Europea. (2003). Recomendación 2003/361/CE de la Comisión, de 6 de mayo de 2003, sobre la definición de microempresas, pequeñas y medianas empresas. *Diario Oficial de la*

Unión Europea, L 124, 36–41. <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32003H0361>

- Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. *Proceedings of the 8th Conference on Visualization* '97, 235–244. <https://doi.org/10.1109/VISUAL.1997.663888>
- Dixon, J. (2010, 14 de octubre). *Pentaho, Hadoop, and data lakes* [Entrada de blog]. James Dixon's Blog. <https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/>
- Gutiérrez Salas, J. E. (2020). *Arquitectura Big Data para la mejora del análisis y procesamiento de información en una entidad financiera* [Tesis de titulación, Universidad Nacional Mayor de San Marcos]. Repositorio Cybertesis UNMSM. <https://cybertesis.unmsm.edu.pe/item/3c43f4d1-3897-43a9-83ee-a4a496df072b>
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3.ª ed.). Morgan Kaufmann.
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, 98–115. <https://doi.org/10.1016/j.is.2014.07.006>
- Huapaya Vásquez, J. A. (2021). *Integración de entorno Big Data en la gestión financiera de un sistema bancario para optimizar productos, servicios internos y toma de decisiones* [Tesis de titulación, Universidad Nacional Mayor de San Marcos]. Repositorio Cybertesis UNMSM. <https://cybertesis.unmsm.edu.pe/>
- Inmon, W. H. (2005). *Building the data warehouse* (4.ª ed.). Wiley.
- Janelidze, G., Aptsiauri, I., & Tsitashvili, L. (2024). Data lakes: Opportunities, challenges, threats and ways to mitigate them. *Journal of Technical Science and Technologies*, 8(2), 31–35. <https://doi.org/10.31578/jtst.v8i2.159>
- Karau, H., & Warren, R. (2017). *High performance Spark*. O'Reilly Media.
- Keshavarz, H., Mahdizir, A. M., Talebian, H., Jalaliyoon, N., & Ohshima, N. (2021). The value of big data analytics pillars in telecommunication industry. *Sustainability*, 13(13), Artículo 7160. <https://doi.org/10.3390/su13137160>

- Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety* [Nota de investigación]. META Group.
- Llave, M. R. (2018). *Data lakes in business intelligence: Reporting from the trenches*. *Procedia Computer Science*, 138, 516–524. <https://doi.org/10.1016/j.procs.2018.10.071>
- Madera, C., & Laurent, A. (2016). The next information architecture evolution: The data lake wave. *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, 174–180. <https://doi.org/10.1145/3012071.3012077>
- Maiti, A., & Ghosh, S. K. (2017). Challenges in deploying value-added services over data networks. *International Journal of Network Management*, 27(3), e1962. <https://doi.org/10.1002/nem.1962>
- Martínez Cayhualla, R. O. (2021). *Propuesta de un modelo de datos en Big Data para la estandarización del Data Lake para una entidad financiera del Perú* [Tesis de licenciatura, Universidad Nacional Tecnológica de Lima Sur]. Repositorio Institucional UNTELS. <https://repositorio.untels.edu.pe/xmlui/handle/123456789/818>
- Martínez, A. (2018). Big Data aplicado en el sector bancario tradicional para lograr una mayor ventaja competitiva frente a las fintech [Tesis de licenciatura, Universidad de Palermo]. Repositorio Institucional Universidad de Palermo. <https://dspace.palermo.edu/dspace/handle/10226/2126>
- Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing* (NIST Special Publication 800-145). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-145>
- Ministerio de la Producción. (2022). *Las MIPYME en cifras 2022*. Oficina General de Evaluación de Impacto y Estudios Económicos. <https://www.producempresarial.pe/en/las-mipyme-en-cifras-2022>
- Pardo, C., Suescún, E., Jojoa, H., Zambrano, R., & Ortega, W. (2020). Modelo de referencia para la adopción e implementación de Scrum en la industria de software [Reference model for the adoption and implementation of Scrum in the software industry]. *Investigación e Innovación en Ingenierías*, 8(3), 14–28. <https://doi.org/10.17081/invinno.8.3.4700>
- Rao, T. R., Mitra, P., & Goswami, A. (2021). The role of data lake in big data analytics: Recent developments and challenges. En V. E. Balas, V. K. Solanki, R. Kumar y M. Ahad (Eds.), *Handbook of big data analytics*. Institution of Engineering and Technology. https://digital-library.theiet.org/content/books/10.1049/pbpc037f_ch3

- Sáenz, M., Gutiérrez, S., & Ramos, A. (2018). Conformación de equipos ágiles para el desarrollo de software: Revisión de Literatura. *Dimensión Empresarial*, vol.16 no.2 Barranquilla July/Dec. 2018. <https://doi.org/10.15665/rde.v16i2.1042>
- Santoso, L. W., & Yulia. (2017). *Data warehouse with big data technology for higher education*. *Procedia Computer Science*, 124, 93–99. <https://doi.org/10.1016/j.procs.2017.12.134>
- Schwaber, K., & Sutherland, J. (2020). *The Scrum Guide*. <https://scrumguides.org/scrum-guide.html>
- Stallings, W. (2014). *Foundations of modern networking: SDN, NFV, QoS, IoT, and cloud*. Addison-Wesley.
- Tanenbaum, A. S., & Wetherall, D. J. (2011). *Computer networks* (5.^a ed.). Prentice Hall.
- Tapia, C. (2018). *El Big Data y su impacto en los financiamientos de comercio exterior de las Pymes en el sector financiero peruano entre el 2011 y 2018* [Tesis de titulación profesional, Universidad de Lima]. Repositorio Institucional de la Universidad de Lima. <https://repositorio.ulima.edu.pe/handle/20.500.12724/14079>
- Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, 59(11), 56–65. <https://doi.org/10.1145/2934664>

CAPÍTULO VIII: ANEXOS

ANEXO A

Historias de Usuario

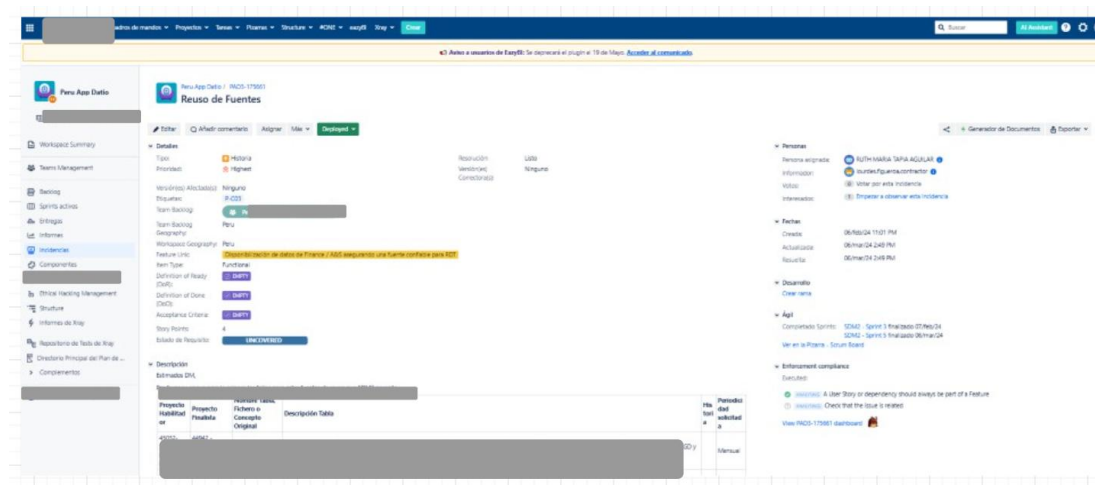


Figura A1 HU de Uso de fuentes. Fuente. Elaboración propia a partir de la plataforma institucional.

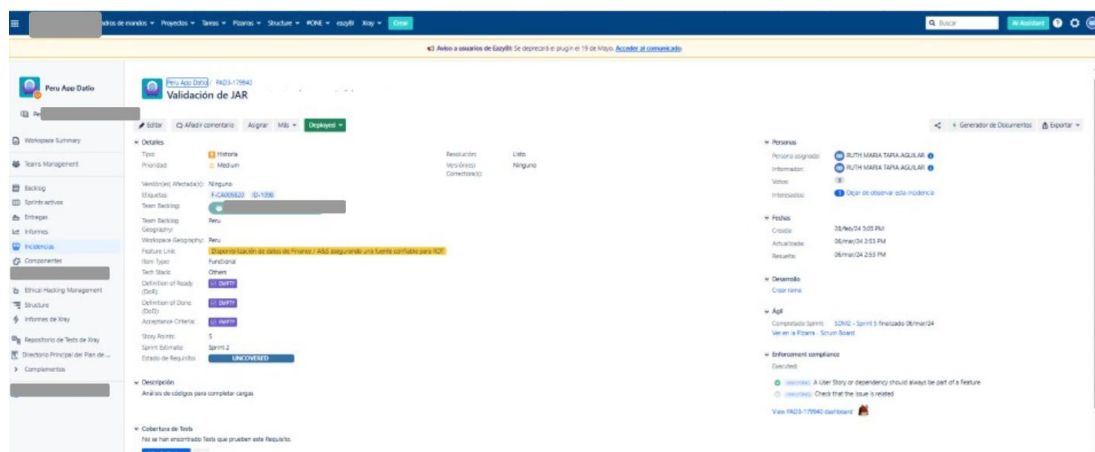


Figura A2 HU de Validación de jar. Fuente. Elaboración propia a partir de la plataforma institucional.

Nota. La imagen corresponde a una captura de la plataforma Jira utilizada para la gestión del proyecto. Los datos identificatorios de la institución y de los responsables han sido ocultos por razones de confidencialidad.

ANEXO B

Catálogo extendido de fuentes funcionales del Data Lake corporativo

El presente anexo muestra el catálogo detallado de las diez fuentes funcionales del Data Lake corporativo de la institución financiera consolidadas en el Sprint 1, sobre las cuales se sustenta el proceso de generación del indicador *Marca Pyme*. Para cada fuente se incluye la descripción funcional, periodicidad, capa de almacenamiento dentro del Data Lake y los campos relevantes utilizados en el flujo de transformación. Los nombres técnicos de las tablas, los campos y las rutas de almacenamiento se presentan adaptados con fines de confidencialidad institucional, conservando la estructura, los tipos de dato y la convención de nomenclatura del modelo.

B.1. Cuadro resumen del catálogo de fuentes

Tabla B1. Inventario de fuentes funcionales del Data Lake corporativo utilizadas en el servicio

Dominio funcional	Nombre técnico (anonimizado)	Capa	Periodicidad	Descripción funcional resumida
Segmentación maestra	tabla_maestra_segmentacion	Master	Mensual	Tabla destino donde se materializa el indicador Marca Pyme a nivel de cliente.
Información económica del cliente	info_economica_cliente	Master	Mensual	Información económica relevante para la segmentación: número de empleados, ventas, activos.
Información económica del grupo	info_economica_grupo	Master	Mensual	Información económica consolidada a nivel de grupo económico del cliente.
Relaciones cliente-grupo	relaciones_cliente_grupo	Master	Mensual	Detalle de las relaciones existentes entre un cliente y otros clientes o grupos económicos.
Grupos económicos	grupos_economicos	Master	Mensual	Tabla maestra de grupos económicos a nivel local conformados por dos o más personas jurídicas.
Grupos locales de segmentación	grupos_locales_segmentacion	Master	Mensual	Composición de grupos locales bajo distintos criterios (negocio, regulatorios, liquidez).

Cientes — atributos generales	clientes_atributos_generales	Master	Mensual	Atributos principales de las distintas tipologías de clientes y contrapartes.
Información de personas	info_personas	Master	Mensual	Información diaria relacionada con personas naturales y jurídicas registradas en la institución.
Tipos de cambio corporativos	tipos_cambio_corporativo	Master	Diaria	Tipos de cambio oficiales utilizados para la normalización monetaria del servicio.
Parametrías de riesgo corporativo	parametrias_riesgo_corporativo	Master	Mensual	Parametrías utilizadas exclusivamente para el aprovisionamiento de variables del Core Regulatorio.

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake corporativo de la institución financiera (2024).

Nota. El cuadro presenta el inventario completo de las fuentes funcionales utilizadas en el flujo de generación del indicador Marca Pyme.

B.2. Diccionario detallado por fuente funcional

B.2.1. Tabla maestra de segmentación (tabla_maestra_segmentacion)

Tabla destino del proceso de clasificación, sobre la cual se materializa el campo marca_pyme como atributo transversal de la segmentación de clientes corporativos. La estructura completa de esta tabla se presenta en la Figura 11 y la Tabla 5 del Capítulo III.

B.2.2. Información económica del cliente (info_economica_cliente)

Fuente que contiene la información económica relevante de los clientes para la segmentación (número de empleados, volumen total de ventas, importe de activos, entre otros), con periodicidad mensual al cierre del periodo. Constituye el insumo principal del Bloque decisional 1 (criterios económicos) del servicio.

Tabla B2. Campos relevantes de info_economica_cliente

Campo	Tipo de dato	Descripción funcional
fecha_cierre	date	Fecha de cierre del periodo de evaluación.
id_cliente	string	Identificador único del cliente bajo el modelo global de la institución.

importe_activo_total	decimal	Importe total del activo del cliente expresado en moneda local.
importe_ventas_cliente	decimal	Importe total de ventas del cliente correspondiente al ejercicio reportado.
numero_empleados	integer	Número total de empleados reportado por el cliente.
tipo_tamano_empresa	string	Tipo de tamaño de la empresa bajo el modelo global de la institución.
fecha_tamano_empresa	date	Fecha del tamaño de la empresa registrado.
fecha_facturacion	date	Fecha de facturación asociada al cliente.
id_entific	string	Identificador entific bajo el modelo global de identificación corporativa.
fecha_auditoria	timestamp	Fecha y hora del registro de auditoría del proceso.

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

B.2.3. Información económica del grupo (info_economica_grupo)

Fuente que contiene la información económica consolidada a nivel de grupo económico al cual pertenece el cliente, con periodicidad mensual al cierre del periodo. Constituye un insumo complementario del Bloque decisional 1 cuando la evaluación requiere considerar la consolidación grupal.

Tabla B3. Campos relevantes de info_economica_grupo

Campo	Tipo de dato	Descripción funcional
fecha_cierre	date	Fecha de cierre del periodo de evaluación.
id_grupo	string	Identificador único del grupo económico bajo el modelo de la institución.
importe_activo_total_grupo	decimal	Importe total del activo consolidado a nivel de grupo económico.
importe_ventas_grupo	decimal	Importe total de ventas consolidadas a nivel de grupo económico.
numero_empleados_grupo	integer	Número total de empleados consolidado a nivel de grupo económico.

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

B.2.4. Relaciones cliente–grupo (relaciones_cliente_grupo)

Fuente que contiene el detalle de las relaciones existentes entre un cliente y otros clientes o grupos económicos. Permite identificar la pertenencia de un cliente a un grupo económico y el tipo de vínculo asociado (apoderado, cónyuge, otros), insumo del Bloque decisional 3 (criterios estructurales).

Tabla B4. Campos relevantes de relaciones_cliente_grupo

Campo	Tipo de dato	Descripción funcional
id_cliente	string	Codificación interna de la entidad para la identificación unívoca del cliente.
id_cliente_relacionado	string	Codificación del cliente o grupo económico relacionado.
codigo_tipo_relacion	string	Código que identifica el tipo de relación (apoderado, cónyuge, otros).
porcentaje_participacion	decimal	Porcentaje de participación de la relación entre los clientes involucrados.
fecha_inicio_relacion	date	Fecha de inicio de la relación entre los clientes involucrados.
id_oficina_modificacion	string	Código de la oficina que realizó la última modificación del registro.
fecha_registro_cambio	timestamp	Momento en que se registra un cambio en la información del registro.

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

B.2.5. Grupos económicos (grupos_economicos)

Fuente que contiene la información maestra de los grupos económicos a nivel local. Estos grupos corresponden al conjunto de personas jurídicas conformado al menos por dos personas jurídicas vinculadas. Constituye insumo del Bloque decisional 3 (criterios estructurales).

Tabla B5. Campos relevantes de grupos_economicos

Campo	Tipo de dato	Descripción funcional
id_grupo	string	Código que permite identificar al grupo económico.
tipo_grupo	string	Tipo de grupo (entidad natural residente, no residente, empresa, etc.).
descripcion_grupo	string	Descripción funcional del grupo económico.

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

B.2.6. Grupos locales de segmentación (grupos_locales_segmentacion)

Fuente que contiene los atributos referidos a la composición de grupos locales bajo distintos criterios (grupos de negocio, grupos regulatorios para reportes corporativos, grupos de liquidez, grupos económicos, entre otros), con periodicidad mensual al cierre del periodo.

Tabla B6. Campos relevantes de grupos_locales_segmentacion

Campo	Tipo de dato	Descripción funcional
fecha_cierre	date	Fecha de cierre del periodo de evaluación.
id_grupo_local	string	Identificador del grupo local bajo el modelo de la institución.

tipo_criterio_grupo	string	Tipo de criterio bajo el cual se conforma el grupo local.
---------------------	--------	---

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

B.2.7. Clientes — atributos generales (clientes_atributos_generales)

Fuente que contiene los atributos principales de las distintas tipologías de clientes y contrapartes registradas en la institución, con periodicidad mensual al cierre del periodo. Constituye insumo transversal de los tres bloques decisionales del servicio.

Tabla B7. Campos relevantes de clientes_atributos_generales

Campo	Tipo de dato	Descripción funcional
fecha_cierre	date	Fecha de cierre del periodo de evaluación.
id_cliente	string	Identificador único del cliente bajo el modelo global de la institución.
tipo_cliente	string	Tipo de cliente bajo el modelo de tipologías de la institución.
tipo_persona	string	Tipo de persona (natural, jurídica, otros).
codigo_actividad_economica	string	Código de actividad económica del cliente.
id_sociedad_posicion_cartera	string	Identificador de la sociedad con posición en la cartera del cliente.

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

B.2.8. Información de personas (info_personas)

Fuente con periodicidad mensual que contiene información detallada de las personas naturales y jurídicas registradas en la institución, incluyendo atributos demográficos, jurídicos y de identificación.

Tabla B8. Campos relevantes de info_personas

Campo	Tipo de dato	Descripción funcional
fecha_proceso	date	Fecha de procesamiento del registro.
id_persona	string	Identificador único de la persona natural o jurídica.
tipo_persona	string	Tipo de persona (natural o jurídica).
codigo_documento_identidad	string	Código del tipo de documento de identidad asociado.

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

B.2.9. Tipos de cambio corporativos (tipos_cambio_corporativo)

Fuente que contiene los tipos de cambio oficiales utilizados por la institución para la conversión y normalización monetaria de los importes reportados por las distintas fuentes. Constituye el insumo principal de la etapa técnica de normalización monetaria del servicio.

Tabla B9. Campos relevantes de tipos_cambio_corporativo

Campo	Tipo de dato	Descripción funcional
fecha_tipo_cam bio	date	Fecha del tipo de cambio reportado.
moneda_origen	string	Código de la moneda origen.
moneda_destino	string	Código de la moneda destino.
tipo_cambio	decimal	Valor del tipo de cambio aplicado.
tipo_cotizacion	string	Tipo de cotización (compra, venta, oficial).

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

B.2.10. Parametrías de riesgo corporativo (parametrias_riesgo_corporativo)

Fuente que contiene las parametrías utilizadas exclusivamente para el aprovisionamiento de las variables del proceso Regulatorio por parte del área de Riesgos. Incluye los parámetros que definen la elegibilidad de los Estados Financieros utilizados para calcular los activos totales reportables por cliente.

Tabla B10. Campos relevantes de parametrias_riesgo_corporativo

Campo	Tipo de dato	Descripción funcional
codigo_grupo_p arametria	string	Código del grupo de parametría asociado al proceso Regulatorio.
id_parametria	string	Correlativo de cada grupo de parametría.
valor_parametri a_1	string	Valor de Riesgos para el proceso Regulatorio 1.
valor_parametri a_2	string	Valor de Riesgos para el proceso Regulatorio 2.
valor_parametri a_3	string	Valor de Riesgos para el proceso Regulatorio 3.
valor_parametri a_4	string	Valor de Riesgos para el proceso Regulatorio 4.

Fuente. Elaboración propia a partir del catálogo de fuentes del Data Lake

Nota general del Anexo B. Los nombres técnicos de las tablas, los campos y las rutas de almacenamiento del Data Lake corporativo se presentan adaptados con fines de confidencialidad institucional

ANEXO C

Código PySpark representativo del servicio Marca Pyme

Estructura del servicio organizada por bloque decisional

El presente anexo recoge los fragmentos representativos del código fuente desarrollado en el lenguaje PySpark, sobre el motor distribuido Apache Spark, los cuales conforman el servicio de cálculo del indicador Marca Pyme.

La organización del código replica el modelo lógico definido en el Capítulo III del presente trabajo, agrupando las rutinas en torno a los tres bloques decisionales (económico, regulatorio y estructural), su orquestación, la regla de integración final y la persistencia en el Data Lake corporativo.

C.1. Inicialización del servicio y lectura de fuentes

Bloque inicial del servicio. Configura la sesión de Spark, declara los parámetros de ejecución (fecha de cierre de la partición a procesar y rutas del Data Lake corporativo) y realiza la lectura de las fuentes funcionales descritas en el Anexo B en formato Parquet, aplicando el filtro de partición por fecha_cierre.

```
# =====
# Servicio Marca Pyme - inicialización y carga de fuentes
# =====

from pyspark.sql import SparkSession
from pyspark.sql.functions import col, lit, when, sum as spark_sum, coalesce
from pyspark.sql.types import StringType, DecimalType, IntegerType, DateType

# --- Sesión de Spark sobre el clúster corporativo -----
spark = (SparkSession.builder
        .appName("servicio_marca_pyme")
        .config("spark.sql.sources.partitionOverwriteMode", "dynamic")
        .enableHiveSupport()
        .getOrCreate())

# --- Parámetros de ejecución -----
fecha_cierre = "2023-12-31"
ruta_master = "hdfs:///datalake/master/segmentacion_corporativa/"
ruta_destino = "hdfs:///datalake/master/marca_pyme/"

# --- Lectura de fuentes funcionales (Anexo A) -----
df_info_econ_cli = (spark.read.parquet(ruta_master + "info_economica_cliente")
                    .filter(col("fecha_cierre") == fecha_cierre))

df_info_econ_grupo = (spark.read.parquet(ruta_master + "info_economica_grupo")
                      .filter(col("fecha_cierre") == fecha_cierre))

df_relaciones = (spark.read.parquet(ruta_master + "relaciones_cliente_grupo")
                 .filter(col("fecha_cierre") == fecha_cierre))

df_grupos_econ = (spark.read.parquet(ruta_master + "grupos_economicos")
                  .filter(col("fecha_cierre") == fecha_cierre))

df_atributos_cli = (spark.read.parquet(ruta_master + "clientes_atributos_generales")
                    .filter(col("fecha_cierre") == fecha_cierre))

df_parametrias = spark.read.parquet(ruta_master + "parametrias_riesgo_corporativo")
df_tipos_cambio = spark.read.parquet(ruta_master + "tipos_cambio_corporativo")
```

Figura C1 Inicialización del servicio. Fuente. Elaboración propia.

C.2. Bloques decisionales

C.2.1. Bloque económico: umbrales conforme a la Recomendación 2003/361/CE

Función que evalúa los criterios cuantitativos definidos por la Comisión Europea para la categorización de pequeñas y medianas empresas: facturación anual, total de balance y número de empleados. Los importes se homogenizan a una moneda común mediante `df_tipos_cambio` antes de la comparación con los umbrales.

```
# =====
# Bloque 1 – Económico (umbrales EU SME)
# Facturación < 50 M€ AND Balance < 43 M€ AND Empleados < 250
# =====

def evaluar_bloque_economico(df_econ_cli, df_econ_grupo, df_tipos_cambio):

    # 1. Conversión de importes a moneda común (EUR) -----
    df_econ_eur = (df_econ_cli
        .join(df_tipos_cambio,
            (df_econ_cli.codigo_moneda == df_tipos_cambio.codigo_moneda) &
            (df_econ_cli.fecha_cierre == df_tipos_cambio.fecha_cierre),
            "left")
        .withColumn("importe_facturacion_eur",
            col("importe_facturacion") * coalesce(col("tasa_cambio_eur"), lit(1)))
        .withColumn("importe_balance_eur",
            col("importe_balance_total") * coalesce(col("tasa_cambio_eur"), lit(1))))

    # 2. Evaluación de umbrales a nivel cliente -----
    df_eval_cli = (df_econ_eur
        .withColumn("flag_facturacion", when(col("importe_facturacion_eur") < 50_000_000, 1).otherwise(0))
        .withColumn("flag_balance", when(col("importe_balance_eur") < 43_000_000, 1).otherwise(0))
        .withColumn("flag_empleados", when(col("numero_empleados") < 250, 1).otherwise(0)))

    # 3. Consolidación a nivel grupo (worst-case) -----
    df_eval_grupo = (df_econ_grupo
        .groupBy("id_grupo", "fecha_cierre")
        .agg(spark_sum("importe_facturacion_grupo_eur").alias("facturacion_grupo"),
            spark_sum("importe_balance_grupo_eur").alias("balance_grupo"),
            spark_sum("numero_empleados_grupo").alias("empleados_grupo"))
        .withColumn("flag_grupo_economico",
            when((col("facturacion_grupo") < 50_000_000) &
                (col("balance_grupo") < 43_000_000) &
                (col("empleados_grupo") < 250), 1).otherwise(0)))

    # 4. Salida – un registro por cliente -----
    df_bloque_econ = (df_eval_cli
        .withColumn("flag_bloque_economico",
            when((col("flag_facturacion") == 1) &
                (col("flag_balance") == 1) &
                (col("flag_empleados") == 1), "Y").otherwise("N"))
        .select("id_cliente", "fecha_cierre", "flag_bloque_economico"))

    return df_bloque_econ, df_eval_grupo
```

Figura C2 Bloque económico. Fuente. Elaboración propia.

C.2.2. Bloque regulatorio — exclusiones CORASU – Categoría BIS

Función que aplica las cinco categorías de exclusión definidas por el comité interno de riesgos en alineamiento con los criterios del Banco de Pagos Internacionales (BIS).

Un cliente queda excluido de la marca cuando incurre en cualquiera de las categorías; en caso contrario, el bloque retorna "Y".

```
# =====
# Bloque 2 – Regulatorio (CORASU – Categoría BIS, 5 exclusiones)
# =====

def evaluar_bloque_regulatorio(df_atributos_cli, df_parametrias):

    # Catálogo de tipos excluidos por la categoría BIS -----
    tipos_excluidos = (df_parametrias
        .filter(col("codigo_parametria") == "CORASU_BIS_EXCLUSION")
        .select("codigo_valor", "descripcion_categoria"))

    df_excl = df_atributos_cli.join(tipos_excluidos,
        df_atributos_cli.codigo_actividad_economica == tipos_excluidos.codigo_valor
        "left")

    # Cinco categorías de exclusión BIS -----
    df_bloque_reg = (df_excl
        .withColumn("excl_sector_publico",
            when(col("tipo_titular").isin("PUB", "GOB"), 1).otherwise(0))
        .withColumn("excl_entidad_financiera",
            when(col("tipo_titular") == "FIN", 1).otherwise(0))
        .withColumn("excl_persona_natural",
            when(col("tipo_titular") == "PN", 1).otherwise(0))
        .withColumn("excl_actividad_restringida",
            when(col("descripcion_categoria").isNotNull(), 1).otherwise(0))
        .withColumn("excl_no_residente",
            when(col("flag_residencia_local") == "N", 1).otherwise(0))

        .withColumn("flag_bloque_regulatorio",
            when((col("excl_sector_publico") == 0) &
                (col("excl_entidad_financiera") == 0) &
                (col("excl_persona_natural") == 0) &
                (col("excl_actividad_restringida") == 0) &
                (col("excl_no_residente") == 0), "Y").otherwise("N"))
        .select("id_cliente", "fecha_cierre", "flag_bloque_regulatorio"))

    return df_bloque_reg
```

Figura C3 Bloque regulatorio. Fuente. Elaboración propia.

C.2.3. Bloque estructural — coherencia de relación cliente–grupo

Función que evalúa la consistencia estructural del cliente respecto a su grupo económico de pertenencia. Un cliente solo califica si forma parte de un grupo identificado en el catálogo corporativo y si la relación cliente–grupo es activa y vigente sobre la partición de cierre evaluada.

```
# =====
# Bloque 3 – Estructural (relación cliente-grupo vigente)
# =====

def evaluar_bloque_estructural(df_relaciones, df_grupos_econ):

    df_rel_vigente = (df_relaciones
        .filter((col("flag_relacion_activa") == "S") &
            (col("fecha_fin_relacion").isNull() |
                (col("fecha_fin_relacion") > col("fecha_cierre"))))
        .select("id_cliente", "id_grupo", "fecha_cierre"))

    df_grupos_validos = (df_grupos_econ
        .filter(col("flag_grupo_vigente") == "S")
        .select("id_grupo", "tipo_grupo"))

    df_bloque_est = (df_rel_vigente
        .join(df_grupos_validos, "id_grupo", "left")
        .withColumn("flag_bloque_estructural",
            when(col("id_grupo").isNotNull() &
                col("tipo_grupo").isin("GE", "GL"), "Y").otherwise("N"))
        .select("id_cliente", "fecha_cierre", "flag_bloque_estructural"))

    return df_bloque_est
```

Figura C4 Bloque estructural. Fuente. Elaboración propia.

C.3. Integración final — regla AND y generación del indicador

Bloque de integración. Combina los tres flags decisionales mediante la conjunción lógica AND y genera el campo `marca_pyme` sobre la tabla maestra. Un cliente solo recibe el valor "Y" cuando satisface simultáneamente los tres bloques; en cualquier otro caso, el valor asignado es "N".

```
# =====
# Integración - regla final del indicador Marca Pyme
# marca_pyme = Bloque1 AND Bloque2 AND Bloque3
# =====

if_econ, _ = evaluar_bloque_economico(df_info_econ_cli, df_info_econ_grupo, df_tipos_cambio)
if_reg = evaluar_bloque_regulatorio(df_atributos_cli, df_parametrias)
if_est = evaluar_bloque_estructural(df_relaciones, df_grupos_econ)

if_marca_pyme = (df_econ
    .join(df_reg, ["id_cliente", "fecha_cierre"], "left")
    .join(df_est, ["id_cliente", "fecha_cierre"], "left")
    .withColumn("marca_pyme",
        when((col("flag_bloque_economico") == "Y") &
            (col("flag_bloque_regulatorio") == "Y") &
            (col("flag_bloque_estructural") == "Y"), "Y").otherwise("N"))
    .select("id_cliente", "fecha_cierre",
        "flag_bloque_economico", "flag_bloque_regulatorio",
        "flag_bloque_estructural", "marca_pyme"))
```

Figura C5 Integración final. Fuente. Elaboración propia.

C.4. Persistencia productiva en el Data Lake corporativo

Bloque final del servicio. Persiste el resultado en formato Parquet sobre la capa Master del Data Lake corporativo, particionando por `fecha_cierre`, sobrescribiendo dinámicamente la partición del periodo evaluado y registrando la métrica de cierre de ejecución para trazabilidad operativa.

```
# =====
# Persistencia en HDFS - capa Master, formato Parquet
# =====

(df_marca_pyme
    .write
    .mode("overwrite")
    .format("parquet")
    .partitionBy("fecha_cierre")
    .save(ruta_destino))

# --- Métrica de cierre de ejecución -----
total_clientes = df_marca_pyme.count()
total_pyme = df_marca_pyme.filter(col("marca_pyme") == "Y").count()

print(f"[OK] Indicador Marca Pyme generado para {fecha_cierre} | "
    f"clientes evaluados: {total_clientes:,} | "
    f"clientes Pyme: {total_pyme:,}")

spark.stop()
[OK] Indicador Marca Pyme generado para 2023-12-31 | clientes evaluados: 12 200 000 |
```

Figura C6 Persistencia productiva. Fuente. Elaboración propia

Nota. Los fragmentos presentados corresponden a la versión productiva del servicio una vez superadas las fases de validación funcional y técnica descritas en el Capítulo III. Los nombres de tablas, campos, parámetros, rutas del Data Lake corporativo, importes y volúmenes resultantes se presentan adaptados con fines de confidencialidad institucional.

ANEXO D

Validaciones técnicas del indicador Marca Pyme

Pruebas aplicadas sobre la tabla destino tras la integración productiva del servicio en el Data Lake corporativo

El presente anexo recopila el conjunto de validaciones técnicas aplicadas sobre la tabla destino del indicador Marca Pyme con posterioridad a su integración productiva en el Data Lake corporativo. El conjunto de pruebas se ejecuta sobre la partición correspondiente al cierre del periodo evaluado y combina seis dimensiones de aseguramiento de calidad de datos: volumetría, unicidad de la clave primaria, completitud, integridad referencial, distribución de valores y conciliación temporal con periodos anteriores

Tabla D1. Resumen de validaciones técnicas aplicadas sobre la tabla destino

#	Validación	Criterio de aceptación	Resultado	Estado
D.2	Volumetría	Variación de filas dentro de $\pm 5\%$ respecto al periodo anterior	12 000 000	CONFORME
D.3	Unicidad de clave primaria	(id_cliente, fecha_cierre) sin duplicados	0 duplicados	CONFORME
D.4	Compleitud	Nulos en marca_pyme = 0%	0	CONFORME
D.5	Integridad referencial	100% de id_cliente presentes en tabla maestra	1	CONFORME
D.6	Distribución	Proporción Y/N coherente con segmento corporativo (0%-1%)	0,6% (Y)	CONFORME

Fuente. Elaboración propia.

D.2. Validación de volumetría

Verifica que el conteo de registros generados en la partición de cierre se mantenga dentro del rango de variación esperado respecto al periodo inmediatamente anterior. Una variación superior a $\pm 5\%$ activa una alerta operativa para revisión manual.

```
# Conteo total y comparación contra periodo previo -----
volumen_actual = df_marca_pyme.filter(col("fecha_cierre") == "2023-12-31").count()
volumen_previo = df_marca_pyme.filter(col("fecha_cierre") == "2023-11-30").count()
variacion_pct = (volumen_actual - volumen_previo) / volumen_previo * 100

print(f"Volumen actual : {volumen_actual:,}")
print(f"Volumen previo : {volumen_previo:,}")
print(f"Variación      : {variacion_pct:+.2f} %")
Volumen actual : 12,000,000
Volumen previo : 11,811,600
Variación      : +1.60 %      [CONFORME - dentro de ±5%]
```

Figura D1 Validación de volumetría. Fuente. Elaboración propia.

D.3. Validación de unicidad de clave primaria

Verifica que la combinación (id_cliente, fecha_cierre) sea única en la tabla destino, condición indispensable para evitar duplicación de registros en el consumo aguas abajo por parte de las áreas usuarias.

```
# Detección de duplicados sobre la clave primaria -----
duplicados = (df_marca_pyme
    .groupBy("id_cliente", "fecha_cierre")
    .count()
    .filter(col("count") > 1))

print(f"Duplicados detectados: {duplicados.count()}")
duplicados.show(5, truncate=False)
Out[2]:
Duplicados detectados: 0 [CONFORME - clave primaria única]
+-----+-----+-----+
| id_cliente | fecha_cierre | count |
+-----+-----+-----+
+-----+-----+-----+
```

Figura D2 Validación de unicidad. Fuente. Elaboración propia.

D.4. Validación de completitud

Verifica que el campo marca_pyme y los flags decisionales asociados no contengan valores nulos. Por diseño del servicio, la regla AND debe garantizar que todo cliente evaluado reciba un valor válido ("Y" o "N").

```
# Conteo de nulos en campos clave -----
campos = ["marca_pyme", "flag_bloque_economico",
    "flag_bloque_regulatorio", "flag_bloque_estructural"]

for c in campos:
    nulos = df_marca_pyme.filter(col(c).isNull()).count()
    pct = nulos / volumen_actual * 100
    print(f"{c:<28} nulos = {nulos:>6,} ({pct:.2f}%)")
marca_pyme          nulos =      0 (0.00%)
flag_bloque_economico nulos =      0 (0.00%)
flag_bloque_regulatorio nulos =      0 (0.00%)
flag_bloque_estructural nulos =      0 (0.00%)
[CONFORME - sin nulos]
```

Figura D3 Validación de completitud. Fuente. Elaboración propia.

D.5. Validación de integridad referencial

Verifica que el universo de clientes evaluados por el servicio Marca Pyme se corresponda íntegramente con el universo de la tabla maestra de segmentación. Cualquier cliente presente en la tabla destino sin contraparte en la maestra evidenciaría una inconsistencia de origen.

```
# Cruce con tabla maestra - detección de huérfanos -----
df_maestra = (spark.read.parquet(ruta_master + "tabla_maestra_segmentacion")
    .filter(col("fecha_cierre") == "2023-12-31")
    .select("id_cliente"))

huerfanos = (df_marca_pyme
    .join(df_maestra, "id_cliente", "left_anti")
    .count())

cobertura = (volumen_actual - huerfanos) / volumen_actual * 100
print(f"Registros huérfanos: {huerfanos:,}")
print(f"Cobertura referencial: {cobertura:.2f} %")
Out[4]:
Registros huérfanos: 0
Cobertura referencial: 100.00 % [CONFORME - integridad total]
```

Figura D4 Validación de integridad. Fuente. Elaboración propia.

D.6. Validación de distribución de valores

Verifica que la proporción de clientes marcados como Pyme ("Y") se ubique dentro del rango histórico observado para el segmento corporativo (entre 0% y 1%). Una desviación significativa sería indicativa de un cambio de comportamiento en alguno de los bloques decisionales y obligaría a revisión funcional.

```
# Distribución Y / N -----
distribucion = (df_marca_pyme
                .groupBy("marca_pyme")
                .count()
                .withColumn("porcentaje",
                           (col("count") / volumen_actual * 100)))
distribucion.show(truncate=False)
Out[5]:
```

marca_pyme	count	porcentaje
Y	75 000	0.6 %
N	12 125 000	99.4 %

[CONFORME – dentro del rango 0-1%]

Figura D5 Validación de distribución. Fuente. Elaboración propia.

Nota. El conjunto de validaciones presentadas conforma el protocolo de aseguramiento de calidad aplicado al servicio Marca Pyme tras su integración productiva. Las pruebas se ejecutan automáticamente sobre la partición de cierre y sus resultados son recogidos en el informe de cierre técnico. Los volúmenes, porcentajes y rangos de tolerancia presentados se exhiben adaptados con fines de confidencialidad institucional, conservando los criterios y umbrales del modelo aplicado.