UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

ESCUELA POLITÉCNICA SUPERIOR DE ELCHE

GRADO EN ESTADÍSTICA EMPRESARIAL

"Comparativa de herramientas de gestión de calidad de datos"

TRABAJO FIN DE GRADO

Junio - 25

AUTOR: Carolina Isabel Belmonte Poveda
DIRECTOR/ES: Kristina Polotskaya

AGRADECIMIENTO

En primer lugar, quiero dedicar este trabajo a mi madre. Su ausencia ha sido la más dura de todas las pruebas en este camino, pero también la mayor fuente de motivación. Todo lo que soy y todo lo que he conseguido lleva una parte de ella también. Gracias por enseñarme a ser fuente incluso en los momentos más difíciles. Este logro es, sobre todo, para ti.

Quiero agradecer también a mi tutora, que años atrás fue compañera de carrera y hoy es un ejemplo de esfuerzo, evolución y compromiso. Ha sido un verdadero privilegio poder contar con su guía, su apoyo constante y su capacidad para inspirar desde la cercanía y la profesionalidad.

Agradezco también profundamente a José Luis, vicedecano del Grado de Estadística Empresarial, quien ha sido un referente académico y personal durante toda mi etapa universitaria. Su apoyo, su disponibilidad y su confianza han sido un pilar fundamental en mi formación.

También quiero extender mi gratitud a todo el profesorado del Grado de Estadística Empresarial. Cada uno ha contribuido con su conocimiento y vocación a construir todo lo que hoy en día sé.

Por último. gracias a mi padre, mi hermana, mi pareja y al resto de la familia y amigos. Vuestra paciencia, vuestro ánimo y vuestra compañía, incluso en los momentos más silenciosos, han sido el respaldo que me ha sostenido hasta el final.

RESUMEN

La calidad de los datos es clave para el éxito empresarial, al influir directamente en las decisiones estratégicas y operativas. Sin embargo, la presencia de datos incompletos, erróneos o inconsistentes puede generar altos costos, afectar la eficiencia operativa y distorsionar las decisiones.

Este trabajo tiene como objetivo principal evaluar y comparar diferentes herramientas de gestión de calidad mediante un análisis estadístico. Se seleccionan herramientas representativas, tanto de código abierto (OpenRefine y Talend Open Studio) como comerciales (IBM InfoSphere QualityStage y Ataccama ONE). Debido a restricciones de acceso, el análisis empírico se ha centrado en las dos primeras, que han sido aplicadas sobre un conjunto de datos con errores simulados para medir su eficacia.

El trabajo adopta un enfoque mixto, combinando métricas estadísticas objetivas (porcentaje de errores corregidos, registros eliminados, tiempos de procesamiento) con una evaluación cualitativa de la experiencia de uso. Los resultados muestran que ambas herramientas mejoran notablemente la calidad de los datos respecto al archivo original, aunque presentan diferencias significativas en su enfoque: OpenRefine es más adecuado para tareas manuales y exploración puntual, mientras que Talend resulta más eficaz en procesos automatizados y estructurados.

Como conclusión, se demuestra que la elección de la herramienta depende del contexto, los recursos disponibles y los objetivos del análisis. El trabajo aporta, además, un marco replicable para futuras comparativas, contribuyendo al desarrollo de buenas prácticas en calidad del dato.

ÍNDICE DEL TRABAJO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Capítulo 1
Introducción

1.1.- Entorno de aplicación

La gestión de la calidad de los datos se ha convertido en un aspecto clave para el éxito de cualquier empresa u organización. En un entorno marcado por la digitalización y el análisis intensivo de datos, disponer de información precisa, completa, actualizada y coherente es fundamental para una toma de decisiones estratégicas y operativas eficiente.

Este trabajo se enmarca en el sector de las tecnologías de la información y, en concreto, en el ámbito de la inteligencia empresarial (Business Intelligence). Muchas empresas del sector retail, financiero o de servicios manejan grandes volúmenes de datos generados por sus clientes, operaciones o procesos internos. Una mala calidad de estos datos puede derivar en errores estratégicos, pérdidas económicas, duplicidades, sanciones por incumplimiento normativo o, incluso, en una pérdida de confianza por parte de los clientes.

La finalidad del presente proyecto es realizar una comparativa de herramientas especializadas en gestión de calidad de datos, con el fin de identificar cuáles ofrecen mejores resultados en términos de detección y corrección de errores, facilidad de uso, funcionalidades, integración con sistemas existentes, entre otros aspectos relevantes. Este análisis se realizará simulando un entorno realista mediante un conjunto de datos de ventas de clientes descargado de una fuente pública (Kaggle), al que se han introducido errores intencionados.

El análisis se orienta a una empresa de tamaño medio del sector comercial que dispone de una infraestructura básica de tecnologías de la información: sistemas de almacenamiento en bases de datos relacionales (como MySQL o PostgreSQL), herramientas de análisis de datos (Excel, Power BI, etc.) y un pequeño equipo de analistas de datos. Estas empresas, en su mayoría, no disponen de grandes presupuestos, por lo que el objetivo es identificar herramientas accesibles, eficaces y con potencial de mejora de la calidad de los datos sin necesidad de realizar grandes inversiones.

Entre los beneficios clave que aporta una buena calidad de datos se encuentran:

- **Optimización de procesos internos:** permite detectar cuellos de botella y áreas de mejora operativa.

- **Cumplimiento normativo:** contar con información fiable es clave para respetar regulaciones como el RGPD.

- **Reducción de costes:** prevenir errores en origen es más barato que corregirlos una vez propagados.

- **Mejora de la experiencia del cliente:** tener información correcta sobre clientes y operaciones permite personalizar servicios y generar confianza.

Este entorno y necesidades justifican la importancia de realizar una comparativa entre herramientas de gestión de calidad de datos, aportando una guía útil tanto para empresas como para analistas que busquen mejorar sus procesos de tratamiento y depuración de información.

1.2.- Justificación del proyecto

En la actualidad, debido a la transformación digital y a la creciente cantidad de datos generados por las organizaciones, la calidad de datos se ha convertido en un factor crítico que influye directamente en la toma de decisiones estratégicas y operativas. La calidad de los datos afecta áreas clave como la analítica empresarial, la inteligencia de negocio, la gestión de relaciones con clientes y la eficiencia operativa.

Sectores como la banca, la salud y el comercio electrónico dependen esencialmente de datos precisos para garantizar la continuidad del negocio y la satisfacción del cliente. Por ello, las organizaciones deben garantizar que los datos con los que trabajan sean precisos, completos, coherentes y actualizados. De no ser así, se corre el riesgo de tomar decisiones erróneas, incurrir en pérdidas económicas o enfrentar sanciones en sectores regulados.

A pesar de su importancia, muchas empresas aún enfrentan grandes desafíos en la gestión de la calidad de sus datos, lo que hace evidente la necesidad de implementar herramientas y metodologías adecuadas. Los errores más comunes –como la duplicación de datos, la inconsistencia y la falta de actualización– pueden impactar negativamente en el rendimiento operativo y la confianza en los análisis realizados.

Este trabajo de Fin de Grado, titulado “Técnicas y herramientas de gestión de calidad de datos", tiene como objetivo analizar y comparar las principales herramientas del mercado enfocadas en la mejora de la calidad de datos. Se busca identificar sus ventajas, limitaciones y funcionalidades clave, proponiendo una guía útil y práctica para su implementación.

Se seleccionarán varias herramientas representativas que han demostrado ser efectivas en el mercado, para luego realizar una comparación en profundidad entre las que más destacan por su aplicabilidad, popularidad o características innovadoras. Esta comparación permitirá identificar qué características son más relevantes para diferentes contextos organizacionales.

La elección de este enfoque se justifica por la necesidad creciente de las organizaciones de optimizar sus procesos de gestión de datos y mejorar la calidad de los mismos, con el fin de apoyar decisiones basadas en datos confiables. Al evaluar y comparar diversas herramientas, se pretende ofrecer recomendaciones prácticas y útiles tanto para empresas como para profesionales del análisis de datos. Además, este estudio también puede resultar de interés para el ámbito académico, al aportar una visión comparativa sobre tecnologías actuales de mejora de datos en entornos reales o simulados.

1.3.- Objetivos

Objetivo general

Comparar diferentes herramientas de gestión de calidad de datos desde una perspectiva estadística, evaluando su eficacia en la detección, corrección y mejora de datos empresariales para identificar cuál ofrece mejores resultados según diversas métricas de calidad.

Objetivos específicos

- Definir los principales criterios de calidad de datos (precisión, completitud, coherencia, validez, actualidad, etc.) y cómo se miden estadísticamente.
- Seleccionar y analizar varias herramientas de calidad de datos.
- Diseñar un experimento con conjuntos de datos empresariales simulados o reales que presenten típicos problemas de calidad.
- Aplicar métodos estadísticos para comparar el rendimiento de las herramientas según los criterios definidos. (Tasas de error corregidas, tiempo de procesamiento, porcentaje de duplicados eliminados,etc.).
- Interpretar los resultados obtenidos y proporcionar una recomendación fundamentada sobre qué herramientas es más adecuada según el contexto empresarial.

Objetivos transversales

- Desarrollar competencias en el análisis estadístico aplicado a la calidad de los datos.
- Fomentar la capacidad crítica y comparativa al evaluar diferentes soluciones tecnológicas.
- Fortalecer las habilidades en la gestión y preparación de datos empresariales.
- Promover la comprensión de cómo la calidad de los datos influye directamente en la toma de decisiones empresariales.
- Dominar el uso de herramientas estadísticas y software de análisis (como R, Python o Excel avanzado) para evaluar resultados.

Objetivos personales

- Mejorar mi capacidad para desarrollar proyectos complejos de principio a fin.
- Adquirir una comprensión más profunda de la calidad de datos y su impacto en la realidad empresarial.
- Ganar experiencia práctica en la aplicación de estadísticas en un entorno real o simulado.

1.4.- Límites del proyecto

Este proyecto se centrará en herramientas de gestión de calidad de datos aplicables a entornos empresariales, excluyendo aquellas diseñadas exclusivamente para Big Data, flujos de datos en tiempo real o procesamiento distribuido.

Asimismo, se analizarán únicamente herramientas con documentación accesible, versión gratuita disponible o de prueba, y soporte activo por parte de la comunidad o del proveedor.

Los resultados obtenidos estarán condicionados por las características específicas del conjunto de datos empleado y por la configuración usada en cada herramienta. Por tanto, las conclusiones del estudio pueden no ser extrapolables a todos los contextos posibles, especialmente si se aplican a otros sectores o conjuntos de datos con estructuras o problemas diferentes.

Este trabajo está orientado a usuarios con un conocimiento técnico medio, por lo que no se contemplarán aspectos vinculados a desarrollos a medida ni a configuraciones empresariales avanzadas o integraciones complejas.

Estos límites permiten acotar el alcance del trabajo a un marco realista, centrado en ofrecer resultados útiles y aplicables para empresas de tamaño medio o pequeño que busquen mejorar la calidad de sus datos mediante soluciones prácticas y de rápida adopción.

Capítulo 2
Antecedentes y estado de la cuestión

2.1.- Situación actual de la empresa

En el contexto actual, las organizaciones enfrentan una creciente dependencia de los datos para llevar a cabo sus operaciones diarias y tomar decisiones estratégicas. Sin embargo, muchas de ellas se enfrentan a problemas comunes relacionados con la calidad de los datos, lo que puede afectar negativamente su rendimiento y competitividad.

Entre los problemas más comunes destaca la presencia de datos erróneos, incompletos o desactualizados en los sistemas de información. Estos errores pueden generar decisiones equivocadas, pérdida de clientes y aumento de los costos operativos. Por ejemplo, una empresa que almacena direcciones de clientes incorrectas puede experimentar dificultades en la entrega de productos, lo que afecta tanto a la satisfacción del cliente como a la reputación de la marca.

El uso de fuentes múltiples y sistemas heredados complica la integración, generando inconsistencias, duplicidades y falta de estandarización. La obsolescencia de la información también representa un reto, ya que muchas empresas carecen de mecanismo eficaces para mantener sus datos actualizados.

A nivel técnico, muchas organizaciones carecen de herramientas especializadas y de personal cualificado para abordar la calidad del dato de forma efectiva. Esto se traduce en una escasa detección proactiva de errores, dependencia de procesos manuales y baja capacidad de respuesta ante los problemas que afectan la fiabilidad de la información.

En este contexto, resulta imprescindible adoptar soluciones tecnológicas que permitan identificar, corregir y prevenir errores en los datos. Este contexto pone de manifiesto la necesidad de realizar una evaluación exhaustiva de las herramientas disponibles en el mercado, para determinar cuáles se adaptan mejor a las necesidades de las empresas actuales.

2.2. Soluciones tecnológicas

En la actualidad, la calidad del dato se ha convertido en un aspecto clave dentro de las estrategias de transformación digital, inteligencia empresarial y gobernanza de la información. Esta creciente necesidad ha propiciado el desarrollo de una amplia variedad de herramientas que permiten detectar, corregir, estandarizar y monitorizar los datos para asegurar su fiabilidad, consistencia y utilidad en procesos analíticos o transaccionales.

Estas herramientas de calidad de datos se diferencian por su arquitectura, enfoque técnico, escalabilidad y nivel de automatización. No existe una única solución universal, sino que la elección debe depender del contexto de uso, el volumen y complejidad de los datos, las capacidades del equipo técnico y los objetivos del proyecto.

En términos generales, las soluciones disponibles en el mercado pueden agruparse en dos grandes categorías:

- **Herramientas de código abierto**

Estas soluciones suelen ser de libre acceso, mantenidas por comunidades activas o respaldadas por organizaciones no comerciales. Se caracterizan por su flexibilidad y capacidad de personalización, lo que las hace ideales para entornos educativos, PYMEs, o proyectos que requieren bajo coste de adopción.

Además, permiten un mayor control sobre el proceso de transformación de los datos, especialmente útil en contextos donde se necesitan adaptar los flujos ETL a requisitos muy específicos. Sin embargo, su uso eficiente puede requerir habilidades técnicas previas (por ejemplo, conocimientos de expresiones regulares, scripting o arquitectura de datos).

Algunos ejemplos destacados en esta categoría son:

- **OpenRefine:** Especializado en limpieza semiautomática, transformación y exploración de datos heterogéneos.

- **Talend Open Studio:** Una plataforma ETL visual con cientos de conectores y transformaciones para automatizar procesos complejos.

- **Soluciones comerciales**

Estas herramientas están orientadas a entornos empresariales complejos que requieren capacidades avanzadas de procesamiento, integración con sistemas corporativos y cumplimiento normativo (como GDPR, HIPAA, BCBS 239, entre otros). Suelen ofrecer funciones potentes como perfilado inteligente de datos, matching probabilístico, enriquecimiento automático y monitoreo en tiempo real.

El coste de adquisición suele ser elevado, pero proporcionan ventajas como soporte técnico especializado, actualizaciones periódicas, integración con plataformas cloud y dashboards de gobierno de datos. Además, muchas de estas soluciones incorporan tecnologías de inteligencia artificial y machine learning para anticipar errores o automatizar decisiones de calidad.

Algunos ejemplos destacables en esta categoría son:

- **IBM InfoSphere QualityStage:** Altamente fiable en procesos de estandarización y emparejamiento de datos en entornos regulados.

- **Ataccama ONE:** Solución moderna y modular que combina calidad de datos, MDM y gobierno del dato con un enfoque centrado en la automatización.

A continuación se presenta una comparativa sistemática entre estas cuatro herramientas representativas, considerando sus funcionalidades principales, ventajas y limitaciones. Esta tabla sirve como marco inicial de referencia para la evaluación empírica más detallada que se desarrollará en el Capítulo 3.

2.2.1. Tabla comparativa de herramientas

Tabla 2.1: Resumen comparativo de las herramientas seleccionadas.

Herramienta	Tipo	Funcionalidades clave	Puntos fuertes	Limitaciones
OpenRefine	Código abierto	Limpieza manual, transformaciones GREL, duplicados	Interfaz intuitiva, gratuita	No escalable para big data
Talend Open Studio	Código abierto	ETL visual, validación, automatización	Gran conectividad , flexible	Curva de aprendizaje, complejidad inicial
IBM QualityStage	Comercial	Matching avanzado, reglas, integración MDM	Precisión y fiabilidad	Coste elevado, requiere expertos
Ataccama ONE	Comercial	ML para calidad, perfilado, gobierno del dato	Plataforma todo-en-uno, moderna	Coste alto, configuración compleja

2.3. Valoración

La revisión del estado del arte permite concluir que el mercado actual de soluciones para la calidad del dato es extenso, heterogéneo y en continua evolución. Las herramientas disponibles abarcan desde propuestas de código abierto hasta plataformas comerciales de alto nivel, cada una con ventajas diferenciadas y áreas específicas de aplicación.

Por un lado, soluciones como **OpenRefine** o **Talend Open Studio** han ganado popularidad en ámbitos educativos, investigativos y en pequeñas organizaciones gracias a su bajo coste de entrada, comunidad activa y versatilidad técnica. Estas herramientas permiten implementar procesos de limpieza y transformación de datos con gran nivel de personalización, siendo especialmente útiles para usuarios con conocimientos técnicos intermedios. No obstante, su rendimiento puede verse limitado ante grandes volúmenes de datos o entornos corporativos con necesidades complejas de gobernanza.

Por otro lado, plataformas empresariales como **IBM InfoSphere QualityStage** o **Ataccama ONE** destacan por su capacidad para manejar flujos de datos críticos a gran escala, aplicar algoritmos avanzados de deduplicación y estandarización, y garantizar trazabilidad en entornos regulados. Sin embargo, su adopción requiere inversiones significativas, tanto en licencias como en infraestructuras y formación, lo que restringe su aplicabilidad en organizaciones pequeñas o con recursos limitados.

De esta comparativa preliminar emergen varias observaciones clave:

- No existe una solución "universalmente superior". La eficacia de una herramienta depende de múltiples factores: tipo de errores a tratar, volumen de datos, presupuesto, entorno tecnológico, y objetivos del proyecto.

- Las herramientas comerciales tienden a ofrecer mayor grado de automatización, escalabilidad y funcionalidades empresariales, pero a costa de una menor accesibilidad y mayor complejidad técnica.

- Las soluciones de código abierto son más accesibles y flexibles, pero requieren mayor involucramiento del usuario y un conocimiento más profundo de los flujos de calidad del dato.

Este análisis evidencia la necesidad de contar con estudios comparativos objetivos y reproducibles que permitan evaluar las herramientas no solo desde una perspectiva teórica, sino también en función de su desempeño práctico frente a datos reales con problemas concretos. La mayoría de las revisiones existentes se centran en funcionalidades declaradas por los fabricantes, sin un análisis empírico riguroso que mida el impacto real de su aplicación.

Por este motivo, el presente trabajo se plantea como una contribución original que busca cubrir ese vacío mediante:

- La construcción de un entorno de prueba controlado, con errores simulados y trazables.
- La aplicación práctica de las herramientas seleccionadas en condiciones similares.
- La evaluación cuantitativa de los resultados a través de métricas estadísticas.
- La incorporación de visualizaciones que permiten comparar de forma comprensible la eficacia de cada enfoque.

De este modo, se aspira a ofrecer no solo una visión comparativa del mercado, sino una guía práctica, básica y replicable como apoyo para profesionales, investigadores o empresas interesadas en mejorar la calidad de sus datos mediante herramientas accesibles, contrastadas y adaptadas a sus necesidades.

2.4. Transición hacia el análisis detallado

Con base en el panorama descrito y en los criterios de selección establecidos, se han escogido cuatro herramientas representativas que cubren diferentes enfoques tecnológicos y niveles de adopción:

- Dos herramientas de **código abierto**: OpenRefine y Talend Open Studio.
- Dos herramientas **comerciales**: IBM InfoSphere QualityStage y Ataccama ONE.

Las herramientas de código abierto serán objeto de un análisis empírico directo mediante su aplicación sobre un conjunto de datos con errores simulados con Python sobre una base de datos pública, dado que permiten libre acceso y experimentación sin restricciones. Las herramientas comerciales serán exploradas desde un enfoque documental, apoyado en fuentes oficiales, estudios de caso y literatura técnica, por limitaciones de acceso a licencias comerciales.

A continuación, en el siguiente capítulo, se profundiza en el funcionamiento, características técnicas, ventajas y limitaciones de cada una de estas herramientas. Este análisis servirá como base para la posterior comparación práctica y estadística que conforma el núcleo metodológico del trabajo.

Capítulo 3
Hipótesis de trabajo

Este trabajo parte de la hipótesis de que existen diferencias significativas en el rendimiento de las herramientas de gestión de calidad de datos, las cuales dependen tanto del tipo de error presente (por ejemplo, duplicados, valores nulos, inconsistencias de formato) como del enfoque técnico y operativo adoptado por cada herramienta. Además, se asume que dichas diferencias pueden observarse y medirse de forma empírica en contextos reales de uso.

Para abordar esta hipótesis, se han seleccionado cuatro herramientas representativas del panorama actual:

- **Dos de código abierto:** OpenRefine y Talend Open Studio.
- **Dos comerciales:** IBM InfoSphere y Ataccama ONE.

Estas herramientas cubren un amplio espectro de funcionalidades, desde la limpieza interactiva hasta el procesamiento automatizado a gran escala. No obstante, debido a las restricciones de acceso a licencias corporativas, el análisis empírico detallado se centrará exclusivamente en las herramientas de código abierto, por ser accesibles, reproducibles y viables para su uso en entornos educativos o empresariales con recursos limitados.

3.1- Formulación de las hipótesis

Se establecen dos niveles de hipótesis de trabajo para guiar la investigación:

- **Hipótesis general:** Existen diferencias significativas entre las herramientas seleccionadas en cuanto a su rendimiento frente a distintos tipos de errores de calidad de datos, diferencias que pueden cuantificarse mediante métricas estadísticas.

- **Hipótesis específica:** Entre OpenRefine y Talend Open Studio existen diferencias medibles en términos de eficacia de limpieza de datos, nivel de automatización y tiempo de procesamiento, que pueden observarse a través del análisis de un conjunto de datos con errores simulados.

3.2- Diseño experimental

La validación de estas hipótesis se llevará a cabo mediante un **experimento controlado,** en el cual ambas herramientas se aplicarán sobre un mismo conjunto de datos que ha sido manipulado intencionadamente para incluir errores representativos del mundo real.

Se evaluarán las siguientes variables:

- **Variables independientes:**
    - Herramienta utilizada *(OpenRefine / Talend Open Studio)*.
    - Tipo de error presente (valores nulos, duplicados, errores tipográficos, desplazamientos estructurales).

- **Variables dependientes:**
    - Porcentaje de errores corregidos.
    - Número de duplicados eliminados.
    - Tiempo medio de procesamiento por herramienta.
    - Grado de completitud final del dataset.
    - Facilidad de uso percibida por el usuario (evaluada de forma cualitativa).

3.3. Herramientas seleccionadas

3.3.1. OpenRefine en profundidad.

OpenRefine es una herramienta de código abierto destinada a la limpieza, exploración, transformación y enriquecimiento de datos estructurados. A diferencia de hojas de cálculo tradicionales como Excel, OpenRefine está diseñada específicamente para tratar datasets sucios o heterogéneos, facilitando su depuración mediante filtros, operaciones masivas y transformaciones estructuradas. Inicialmente desarrollada por Google bajo el nombre de Google Refine, la herramienta ha sido mantenida por la comunidad desde su liberación como proyecto open source.

Principales características técnicas:

- **Interfaz web local:** OpenRefine se ejecuta como un servidor local y se accede a través de un navegador, lo que permite una experiencia interactiva sin necesidad de conexión a internet. Su diseño modular favorece la navegación por columnas, el uso de filtros y la agrupación de valores.

- **Lenguaje GREL** (General Refine Expression Lenguage): Permite aplicar transformaciones complejas a los datos mediante expresiones personalizadas. También admite lenguajes como Jython y Clojure para tareas más avanzadas.

- **Detección de duplicados mediante clusterización:** OpenRefine ofrece métodos como key collision, nearest neighbor o fingerprint para identificar registros duplicados, incluso cuando presentan pequeñas variaciones ortográficas.

- **Transformación de formatos y estructuras:** Permite importar y exportar archivos en múltiples formatos (CSV, Excel, JSON, XML,RDF), así como transformar columnas en filas (transposición), dividir o unir campos, cambiar tipos de datos, o convertir datos tabulares en estructuras jerárquicas.

- **Historial de operaciones:** Cada cambio aplicado a los datos queda registrado como paso en el proyecto. Esto facilita la reversión de transformaciones y la exportación de scripts de limpieza reutilizables (operation history).

- **Extensiones y conectores:** Mediante la instalación de extensiones, OpenRefine puede conectarse con APIs externas (por ejemplo, reconciliación con Wikidata o Google Knowledge Graph), lo que permite enriquecer los datos con información contextual.

Ventajas y puntos fuertes:

- **Gratuito, de código abierto y multiplataforma:** Puede ejecutarse en Windows, macOS y Linux, y al no requerir conexión en línea, puede utilizarse en entornos con restricciones de red o privacidad.

- **Curva de aprendizaje moderada:** Aunque es más potente que una hoja de cálculo, su uso no requiere conocimientos avanzados de programación. La interfaz facilita la interacción directa con los datos.

- **Ideal para depuración semiautomática:** Mediante operaciones como "cluster and edit", permite revisar lotes de valores similares de forma controlada, conservando el juicio del usuario sin depender exclusivamente de algoritmos automáticos.

- **Trazabilidad:** El historial completo de cambios permite auditar procesos de limpieza o replicarlos en otros datasets similares, favoreciendo la reproducibilidad científica.

Limitaciones actuales:

- **No está diseñado para Big Data:** Al ejecutar los procesos localmente y cargar los datos en memoria RAM; presenta limitaciones de escalabilidad en conjuntos de datos que superan los cientos de miles de registros. En estos casos, puede volverse lento o incluso bloquearse.

- **Ausencia de automatización externa:** Aunque se puede reutilizar scripts y operaciones, no existe integración directa con workflows programáticos o ejecución en segundo plano como sí ocurre en herramientas ETL completas (Talend, Apache NiFi…).

- **No gestiona flujos complejos de transformación:** Carece de funcionalidades de programación condicional, integración de múltiples fuentes simultáneas o control de errores automatizado.

Aplicabilidad práctica:

OpenRefine se ha consolidado como una herramienta de referencia para la primera fase de depuración de datos, especialmente útil cuando los datasets son de tamaño medio o contienen errores heterogéneos (tipográficos, inconsistencias de formato, duplicados parciales, etc). Es ampliamente utilizado en:

- Proyectos académicos y de investigación donde la limpieza de datos procede al análisis estadístico o modelado.

- Periodismo de datos, para investigar grandes volúmenes de registros públicos (por ejemplo, presupuestos, contratos, licitaciones).

- Humanidades digitales y biblioteconomía, para normalizar catálogos y metadatos en archivos históricos.

- Empresas sin infraestructura ETL, requieren una solución ágil para preparar sus datos antes de integrarlos en plataformas analíticas.

Por sus características, OpenRefine complementa perfectamente herramientas de análisis como R o Python, actuando como una etapa intermedia de saneamiento y exploración de la información.

3.3.2. Talend Open Studio en profundidad

Talend Open Studio for Data Integration es una plataforma ETL (Extract, Transform, Load) de código abierto ampliamente utilizada para diseñar, automatizar y orquestar procesos de integración de datos. Su enfoque modular basado en componentes y su interfaz gráfica la convierten en una herramienta muy potente tanto para usuarios técnicos como para perfiles intermedios.

Características técnicas destacadas:

- **Entorno visual de desarrollo (GUI):** La herramienta ofrece un entorno tipo "drag-and-drop" donde los usuarios pueden construir flujos de datos (jobs) utilizando componentes predefinidos para tareas como lectura de archivos, validación, transformación y exportación.

- **Amplia biblioteca de componentes y conectores:** Talend incluye más de 900 componentes nativos que permiten conectar con bases de datos (MySQL, PostgreSQL, Oracle…), archivos planos, servicios web, APIs REST, FTP, sistemas cloud (AWS, Azure, Google Cloud), etc.

- **Transformaciones avanzadas y limpieza:** El sistema permite aplicar transformaciones lógicas, validaciones, filtros, reemplazos, operaciones condicionales, cálculo de campos derivados, detección y eliminación de duplicados, y agrupaciones de datos.

- **Integración de Java personalizado:** Los usuarios pueden extender los flujos añadiendo bloques de código Java directamente mediante componentes como *tJava*, *tJavaRow* o *tJavaFlex,* lo que permite crear funciones avanzadas sin salir del entorno visual.

- **Soporte para la paralelización y manejo de errores:** A través de subjobs, enlaces condicionales y estructuras de control (*if, catch, die, runIf*...), Talend permite gestionar procesos complejos con control de errores y ejecuciones condicionales.

- **Registro y monitorización:** Permite configurar logs, seguimiento de errores, y exportar métricas de ejecución, lo que facilita la auditoría de procesos y el mantenimiento en entornos empresariales.

Ventajas principales:

- **Alta escalabilidad:** Talend permite gestionar desde pequeños flujos locales hasta grandes integraciones en entornos de producción distribuidos (con versiones Enterprise).

- **Comunidad activa y extensiones:** La comunidad de usuarios es amplia y existen foros, wikis, blogs y repositorios donde se comparte soluciones.

- **Facilidad de depuración visual:** Su consola de monitoreo y trazas facilita localizar errores o cuellos de botella en la ejecución.

- **Flexibilidad en los procesos ETL:** Admite la creación de pipelines reutilizables, estructuración de subprocesos y conexiones con fuentes heterogéneas.

Limitaciones:

- **Curva de aprendizaje moderada:** Aunque cuenta con una interfaz visual intuitiva, su complejidad aumenta rápidamente al integrar múltiples fuentes, aplicar transformaciones complejas o definir condiciones lógicas.

- **Necesidad de configuración inicial:** La conexión con bases de datos, importación de componentes y exportación requiere cierto conocimiento previo.

- **No diseñado para limpieza manual:** A diferencia de OpenRefine, Talend no está pensado para inspección visual directa de datos sino para flujos sistematizados.

Aplicabilidad práctica:

Talend es especialmente adecuado en proyectos que requieren procesos repetitivos. limpieza estructurada y transformación automatizada de datos. Su aplicabilidad es ideal para:

- Integración entre bases de datos y sistemas heterogéneos.
- Limpieza previa a la carga en almacenes de datos (data warehouse).
- Automatización de pipelines para proyectos de BI o reporting.
- Generación de informes y validaciones periódicas de calidad del dato.

Es muy utilizado en entornos empresariales, administración pública y servicios financieros, donde se requiere trazabilidad, seguridad y programación de tareas ETL robustas.

3.3.3. IBM InfoSphere QualityStage en profundidad

IBM InfoSphere QualityStage es una solución empresarial para la gestión de calidad del dato diseñada para identificar, limpiar, estandarizar, combinar y enriquecer grandes volúmenes de datos provenientes de fuentes diversas. Forma parte del ecosistema IBM DataStage lo convierte en un componente clave dentro de pipelines complejos de integración y gobernanza de datos.

Se utiliza ampliamente en organizaciones que necesitan asegurar la confiabilidad de sus datos maestros (MDM), especialmente en sectores regulados como banca, sanidad o telecomunicaciones, donde la calidad del dato afecta directamente a la toma de decisiones y cumplimiento normativo.

Características técnicas destacadas:

- **Limpieza y estandarización basada en reglas**: QualityStage aplica transformaciones para normalizar datos de nombres, direcciones, códigos postales, fechas y otros formatos complejos. Permite crear plantillas y reglas reutilizables para cada tipo de dato o país.

- **Motor de *matching* y *merging* inteligente:** Su algoritmo de comparación probabilística permite identificar registros duplicados o relacionados incluso si los valores no coinciden exactamente. Usa técnicas como *fuzzy matching*, *weighting* y *threshold rules* para calcular similitud.

- **Integración nativa con IBM DataStage:** Al formar parte de la suite InfoSphere, QualityStage se pueden encadenar con flujos de extracción, transformación y carga desarrollados en DataStage, permitiendo una orquestación completa del ciclo ETL con control de calidad embebido.

- **Soporte para grandes volúmenes de datos:** Diseñado para ejecutarse en entornos distribuidos (como clusters o mainframes), puede procesar millones de registros con alto rendimiento, utilizando procesamiento paralelo y técnicas de data partitioning.

- **Gestión de referencias y enriquecimiento externo:** Permite conectar con bases de datos externas o directorios (como LexisNexis, D&B, censos, listas negras, etc.) para verificar y enriquecer los datos maestros con información contextual o normativa.

Ventajas principales:

- **Alta precisión en identificación de duplicados:** Especialmente en dominios como nombres de clientes o direcciones postales, donde otros sistemas fallan debido a errores de escritura, abreviaciones o cambios en el orden de los campos.

- **Capacidad de integración corporativa:** Puede formar parte de arquitecturas de gobierno del dato, sistemas MDM y soluciones regulatorias (KYC, AML, etc.).

- **Escalabilidad y rendimiento:** Diseñado para entornos de misión crítica, puede ejecutarse en modo batch o embebido en flujos transaccionales.

- **Control de calidad auditable:** Genera logs y métricas detalladas del proceso de limpieza, fusión y validación, lo cual es fundamental en sectores auditados o sujetos a normativas estrictas.

Limitaciones:

- **Alto coste de adquisición y mantenimiento:** Al ser una solución empresarial, implica costes de licencia, formación y consultoría que la hacen inviable para pequeños proyectos o instituciones educativas.

- **Curva de aprendizaje considerable:** Requiere personal técnico especializado en herramientas IBM, conocimientos de procesos ETL y comprensión de los algoritmos de *matching* y estandarización.

- **Dependencia del ecosistema IBM:** Aunque es integrable con otros entornos, su rendimiento y potencial máximo se alcanza cuando se usa junto con DataStage y otros productos InfoSphere.

Aplicabilidad práctica:

IBM InfoSphere QualityStage está especialmente recomendado para organizaciones que necesitan asegurar la calidad del dato como parte de un sistema más amplio de gobernanza, como por ejemplo:

- Instituciones bancarias y aseguradoras, donde identificar registros duplicados (clientes, cuentas, transacciones) es clave para evitar fraudes y cumplir con regulaciones como KYC o FATCA.

- Hospitales y sistemas sanitarios, que requieren consolidar historiales clínicos de un mismo paciente procedentes de distintas fuentes o proveedores.

- Compañías de telecomunicaciones, que manejan grandes volúmenes de datos de clientes, contrataciones, facturación y soporte técnico.

- Gobiernos y agencias públicas, para limpieza y consolidación de padrones, censos, registros fiscales o catastro.

3.3.4. Ataccama ONE en profundidad

Ataccama ONE es una plataforma unificada de gestión de datos que combina funcionalidades de calidad del dato, gobierno, master data management (MDM), integración y catalogación. Su enfoque moderno está centrado en facilitar la colaboración entre perfiles técnicos y de negocio mediante una interfaz visual, flujos automatizados y uso intensivo de inteligencia artificial.

Diseñada para organizaciones que requieren gobernar de forma integral todo el ciclo de vida de sus datos, Ataccama ONE permite desde el perfilado y limpieza hasta el cumplimiento normativo y la generación de políticas de uso y acceso.

Características técnicas destacadas:

- **Perfilado automático de datos:** Al cargar una fuente, la plataforma analiza automáticamente tipos de datos, patrones, formatos, valores nulos, duplicados y distribuciones, generando dashboards interactivos que permiten entender el estado de calidad en tiempo real.

- **Motor inteligente de limpieza y estandarización:** Ataccama utiliza técnicas de machine learning y reglas predefinidas para sugerir correcciones, normalizar formatos y aplicar políticas de calidad sin necesidad de intervención manual constante. Esto permite, por ejemplo, unificar nombres de empresa, corregir abreviaturas o detectar inconsistencias fonéticas.

- **Detección de anomalías y outliers:** Gracias al aprendizaje automático, la plataforma puede identificar registros que se desvían del comportamiento habitual (por ejemplo, una fecha de nacimiento errónea, un importe fuera de rango o una relación inusual entre campos).

- **Gobierno del dato integrado:** Incluye funcionalidades de data lineage, definición de políticas, clasificación automática de datos sensibles (como datos personales según GDPR), trazabilidad de cambios y generación de catálogos accesibles por equipos de negocio.

- **Interfaz visual y flujos colaborativos:** Diseñada para facilitar la interacción entre analistas, científicos de datos y perfiles no técnicos, la plataforma permite crear workflows visuales reutilizables y compartir reglas de calidad de manera sencilla.

- **Arquitectura híbrida y escalable:** Ataccama ONE puede desplegarse en entornos cloud (AWS, Azure, GCP) o on-premise, integrándose con sistemas tradicionales (bases de datos, ERP, CRM) y modernos (APIs, data lakes, streaming…).

Ventajas principales:

- **Plataforma todo-en-uno:** Unifica calidad, gobierno, integración y análisis en una única solución centralizada, lo que evita la fragmentación de los datos entre los departamentos (silos) y reduce la necesidad de usar múltiples herramientas aisladas.

- **Alta automatización:** El uso de IA permite detectar problemas de calidad y proponer soluciones sin intervención continua, ideal para entornos dinámicos.

- **Interfaz moderna e intuitiva:** Apta para usuarios de negocio, democratiza el acceso a procesos complejos como perfilado, validación y gobierno del dato.

- **Adaptabilidad por sectores:** Ataccama ofrece modelos predefinidos para sectores como banca, sanidad, retail o telecomunicaciones, facilitando la adopción.

Limitaciones:

- **Elevado coste de licencia y despliegue:** Como solución premium orientada a grandes corporaciones, su adopción implica una inversión importante tanto en licencias como en servicios de configuración inicial.

- **Necesidad de personalización:** Aunque ofrece módulos configurables, su eficacia depende de la definición correcta de reglas y modelos de negocio, lo que requiere trabajo previo y perfil técnico especializado.

- **Dependencia de ecosistema propietario:** Algunas funcionalidades avanzadas están disponibles solo en versiones integradas con otros módulos de Ataccama (por ejemplo, MDM o Data Governance Center).

Aplicabilidad práctica:

Ataccama ONE es especialmente eficaz en organizaciones que buscan establecer una estrategia de gobierno de datos centralizada y sostenible, unificando múltiples necesidades:

- **Banca y seguros:** Detección de fraudes, gestión de clientes, cumplimiento normativo (BCBS 239, GDPR, FATCA…).

- **Sanidad:** Unificación de historiales clínicos, calidad de datos para investigación y compliance de protección de datos sensibles.

- **Administración pública:** Control de calidad y trazabilidad en censos, registros civiles, catastro, o interoperabilidad de sistemas.

- **Empresas multinacionales:** Gestión de datos maestros, alineamiento de sedes internacionales, reporting automatizado.

Gracias a su arquitectura flexible, puede integrarse tanto con flujos tradicionales como con entornos modernos de data lake, big data y ciencia de datos, actuando como una capa intermedia que garantiza la calidad, consistencia y trazabilidad de la información.

3.4. Enfoque metodológico

El estudio adoptará un enfoque mixto:

- **Cuantitativo:** Mediante el análisis de métricas objetivas relacionadas con la eficacia y eficiencia de cada herramienta.

- **Cualitativo:** Mediante la valoración de la experiencia de uso, la flexibilidad y la facilidad de aprendizaje percibida por el usuario.

Las pruebas estadísticas utilizadas incluirán **análisis de varianza (ANOVA)** para determinar si existen diferencias significativas entre grupos, así como pruebas post-hoc (como Tukey o Bonferroni) para analizar comparaciones específicas. Además, se emplearán visualizaciones como **boxplots o diagramas de dispersión** para facilitar la interpretación de los resultados.

En conjunto, este capítulo establece el marco experimental del estudio, definiendo con claridad las hipótesis, variables y métodos que se utilizarán para validar los objetivos planteados. Este diseño proporciona una base sólida para una evaluación rigurosa de herramientas de calidad del dato, enfocándose no solo en su funcionalidad declarada, sino también en su rendimiento real frente a datos con errores estructurados. La combinación de evaluación técnica y percepción de uso busca ofrecer una visión completa y aplicable a distintos contextos profesionales y académicos.

Capítulo 4
Metodología y resultados

4.1.- Planificación del proyecto

Para estructurar el desarrollo del proyecto, se ha seguido un enfoque de ciclo de vida incremental, lo que ha permitido abordar las distintas fases de manera secuencial con flexibilidad para retroalimentar y corregir errores o ajustas decisiones conforme se avanzaba.

El proyecto se ha estructurado en siete grandes etapas, cada una con objetivos y herramientas claramente definidos. Estas etapas, que representan tanto la evolución del análisis como la consolidación de los resultados, han sido:

1. **Investigación y recopilación de información:** Se realizó una búsqueda bibliográfica y técnica para contextualizar el proyecto, comprender las herramientas disponibles y definir la metodología a seguir.

2. **Selección de herramientas:** Se evaluaron distintas soluciones para limpieza y análisis de datos, decantándose por Talend Open Studio, OpenRefine y R por su complementariedad y relevancia académica.

3. **Preparación del entorno y generación del conjunto de prueba:** Se adecuó el entorno de trabajo, se organizó la estructura de carpetas y se preparó un conjunto de datos original junto con una versión modificada con Python para simular errores y así poder aplicar los procesos de limpieza.

4. **Diseño del experimento:** Se definieron los objetivos técnicos, las métricas de comparación, las variables de interés y los criterios de calidad para evaluar la limpieza.

5. **Aplicación de herramientas de limpieza:** Se ejecutó la limpieza con Talend Open Studio y OpenRefine sobre el mismo conjunto de datos con errores, aplicando distintas estratégias (automatización vs. corrección manual).

6. **Recogida de resultados y análisis estadístico:** Se utilizaron herramientas estadísticas en R para comparar los datasets resultantes en cuanto a registros eliminados, métricas numéricas y distribución de valores.

7. **Redacción del documento final:** Se sintetizaron los resultados, se organizaron los anexos técnicos y se elaboró el presente trabajo académico.

A continuación se muestra un diagrama de Gantt con la planificación temporal estimada del proyecto, así como la relación entre las distintas fases y tareas. El cronograma abarca desde el inicio de la fase de documentación hasta la revisión final del documento.

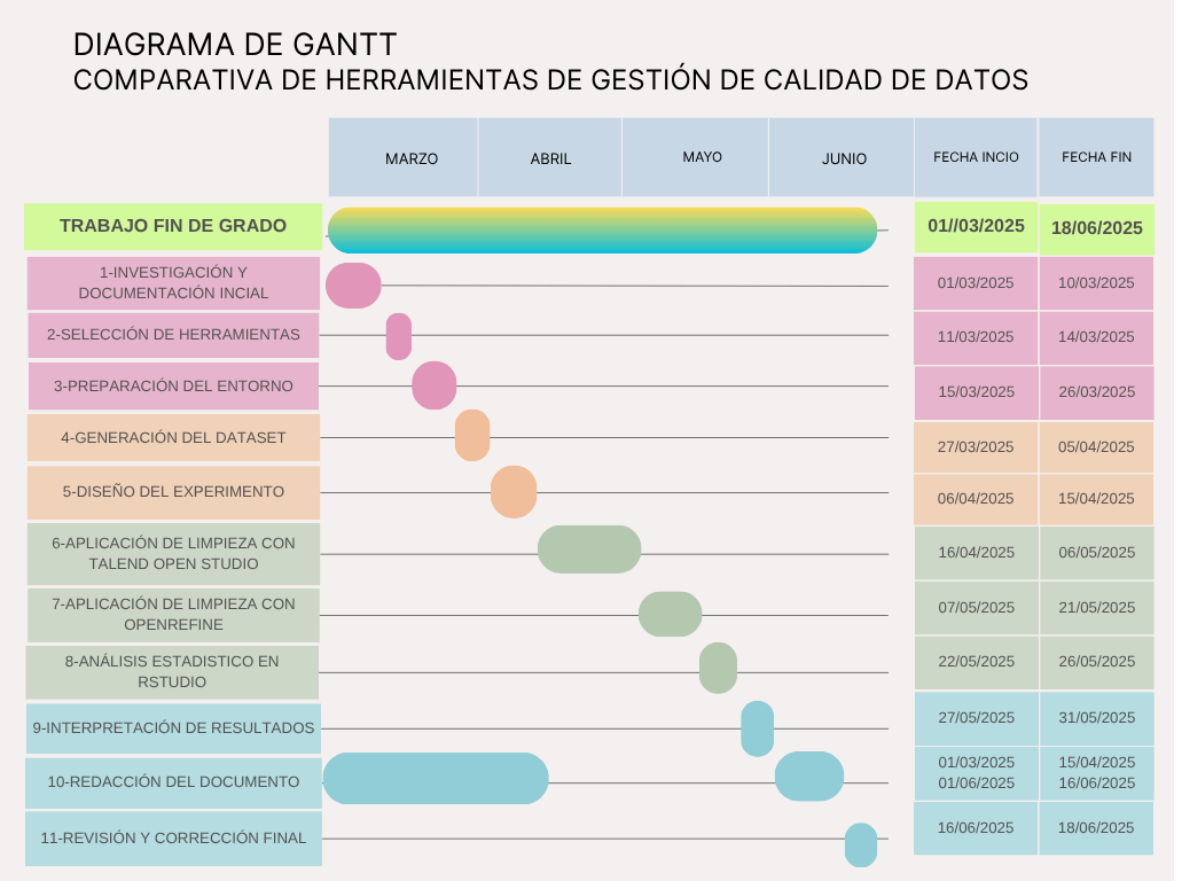

Figura 4.1: Diagrama de Gantt del proyecto.

4.2 Captura de requisitos.

Previo al desarrollo del experimento, ha sido necesario delimitar claramente tanto el contexto técnico como los requisitos funcionales y no funcionales del entorno de pruebas. Debido a limitaciones de acceso a herramientas propietarias como Ataccama ONE e IBM InfoSphere QualityStage, se ha optado por trabajar en un entorno **simulado y controlado**, replicando de forma razonable situaciones reales de calidad de datos.

Las pruebas se han basado en tres fuentes principales:

- Revisión de documentación oficial y técnica de proveedores.
- Estudios de caso públicos y accesibles.
- Pruebas reales realizadas con versiones gratuitas o entornos de evaluación de las herramientas seleccionadas.

Por tanto, los resultados deben entenderse como aproximaciones fundamentadas en observaciones reales y análisis fundamentado en documentación técnica.

4.2.1 Roles y usuarios definidos.

Durante el diseño del experimento se han definido dos roles principales con el objetivo de simular cómo podría organizarse el proceso de mejora de la calidad de los datos en una empresa real de tamaño medio. Este planteamiento refleja un entorno donde los recursos son limitados, pero donde la trazabilidad y el control del tratamiento de la información siguen siendo aspectos clave.

Tabla 4.1: Roles definidos en el proyecto.

Rol	Descripción
Analista de datos	Usuario encargado de identificar errores en los datos y aplicar herramientas de limpieza.
Supervisor técnico	Revisa los resultados obtenidos de la herramienta seleccionada y valida la calidad final de los datos procesados.

4.2.2 Requisitos funcionales.

Los siguientes requisitos funcionales se han definido a partir de los objetivos del proyecto:

- Carga datasets en formatos estándar (CSV, XLSX).
- Detectar errores comunes: Valores nulos, duplicados, inconsistencias.
- Ejecutar tareas de limpieza y transformación.
- Exportar los datos depurados.
- Generar métricas de calidad para comparar resultados.

4.2.3 Requisitos no funcionales.

Adicionalmente, se han considerado los siguientes requisitos no funcionales relevantes:

- Facilidad de uso de las herramientas (evaluada mediante observación directa).
- Tiempo de procesamiento inferior a 6 minutos por cada herramienta sobre un dataset de más de 2000 registros.
- Reproducibilidad: Posibilidad de repetir el proceso con los mismos resultados.

4.2.4 Conjunto de datos utilizado

Para contextualizar y validar los resultados obtenidos, se ha utilizado un conjunto de datos descargado de una fuente pública (Kaggle), titulado **"Customer Sales Data"**, el cual ha sido modificado intencionadamente mediante un script en Python para incluir manualmente errores realistas y comunes que afectan a la calidad de los datos.

Entre los errores introducidos se encuentran:

- Valores nulos aleatorios en campos clave (entre un 5-10%).
- Duplicados de registros con ligeras variaciones ortográficas.
- Inconsistencias en el formato de fechas y códigos postales.
- Errores tipográficos en los nombres de producto y categorías.

El conjunto de datos alterado, denominado "**con_errores.csv"**, ha sido utilizado como base para simular los procesos de limpieza con **OpenRefine** y **Talend Open Studio**, y los resultados se han comparado con la versión simulada con errores intencionados, y la que nos ha servido como "verdad base".

4.2.5 Casos de uso del sistema experimental

Tabla 4.2: Resumen casos de uso del proyecto.

Caso de uso	Descripción
Importar dataset	El analista carga un documento de datos desde un archivo CSV o Excel.
Detectar errores	La herramienta identifica errores como valores nulos, duplicados, errores tipográficos,etc.
Aplicación de herramienta	Se ejecutan transformaciones y correcciones sobre los datos con OpenRefine y Talend Open Studio.
Comparar resultados	Se analizan los resultados de las distintas herramientas utilizadas mediante métricas de calidad y análisis en R.
Exportar informe	Se genera un documento resumen con las métricas y resultados obtenidos.
Validar calidad final	El supervisor técnico revisa los resultados para verificar su validez.

4.3 Diseño del procedimiento experimental

El diseño del experimento se estructuró siguiendo una lógica modular basada en fases independientes pero conectadas, lo que permitió analizar con claridad cada etapa del proceso, desde la simulación de errores hasta la limpieza y evaluación de los resultados. Este enfoque facilitó la comparación controlada entre distintas herramientas y la trazabilidad de los datos.

4.3.1 Diagrama Entidad-Relación (E/R)

El diagrama E/R modela la estructura lógica de los datos utilizados durante las pruebas. Se ha partido de un esquema sencillo, compuesto por dos entidades principales:

- **Cliente**: Representa a cada cliente registrado en el sistema. Contiene atributos relevantes como el nombre y dirección, campos donde típicamente se detectan errores de calidad (faltantes, incorrectos, duplicados, etc.).

- **Venta:** Contiene información relativa a las compras realizadas por los clientes, incluyendo fecha, producto adquirido e importe. Los errores de calidad asociados a esta entidad suelen estar relacionados con formatos de fechas o inconsistencias de valores.

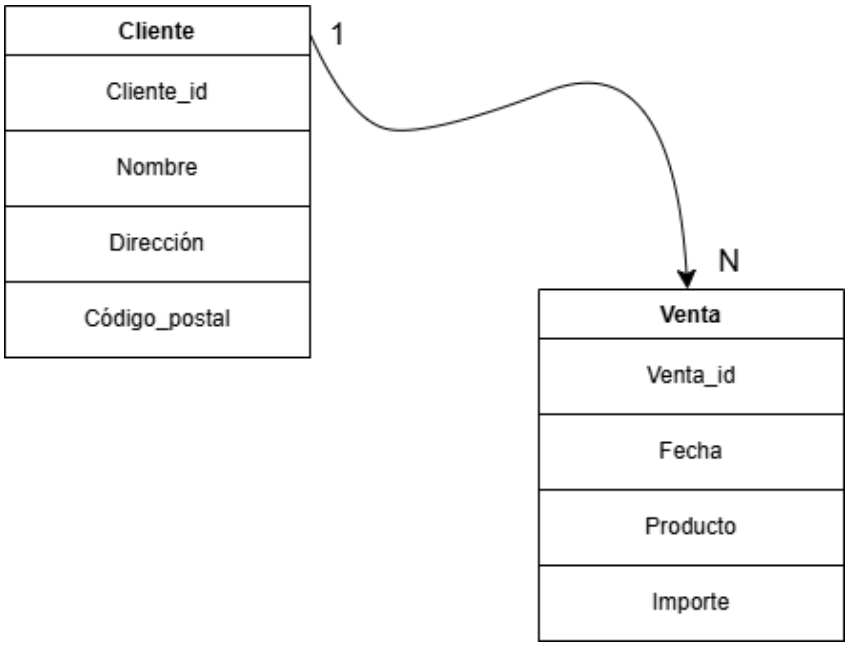

Figura 4.2: Estructura lógica del dataset

La relación entre ambas entidades es de uno-a-muchos: un cliente puede estar asociado a varias ventas. Esta estructura facilita posteriormente a analizar la propagación de errores desde una entidad principal (Cliente) hacia relaciones dependientes (Ventas).

4.3.2 Representación modular del flujo de trabajo.

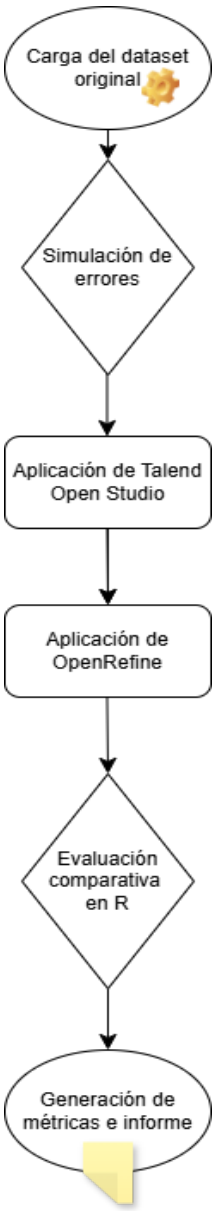

Figura 4.3: Diagrama simplificado de los pasos seguidos en el proyecto.

4.4 Implementación del experimento.

La implementación se llevó a cabo utilizando las dos herramientas de código abierto seleccionadas, **OpenRefine** y **Talend Open Studio,** aplicadas sobre el mismo conjunto de datos con errores intencionados.

Ambas herramientas se utilizan para importar, detectar errores y corregir datos problemáticos. Se emplearon además procesos de limpieza como “Clustering” en OpenRefine y “tMap” en Talend.

4.4.1 Implementación estilo Talend Open Studio.

En esta sección describimos en profundidad el proceso de limpieza de datos llevado a cabo con la herramienta Talend Open Studio for Data Integration, una solución ETL (Extract, Transform, Load) de código abierto que permite trabajar con flujos de datos complejos mediante una interfaz visual e intuitiva. Esta herramienta resulta especialmente útil para tareas de depuración, transformación y estructuración de datos, así como para automatizar procesos en contextos de análisis de grandes volúmenes de información.

El conjunto de datos utilizado corresponde a una base de datos de registros de pedidos comerciales que, en su versión original, contenía múltiples problemas de calidad: campos vacíos, outliers, errores tipográficos, valores desplazados entre columnas y formatos inconsistentes.

El objetivo principal fue limpiar y estandarizar esta base de datos para posteriormente compararlas con el mismo proceso realizado en OpenRefine y evaluarla desde el punto de vista analítico con Rstudio.

El flujo de trabajo de Talend Open Studio que se ha diseñado sirve para ejecutar una secuencia estructurada de pasos que aseguran una limpieza progresiva y lógica del dataset. El diseño del flujo fue el que se presenta en la siguiente imagen:

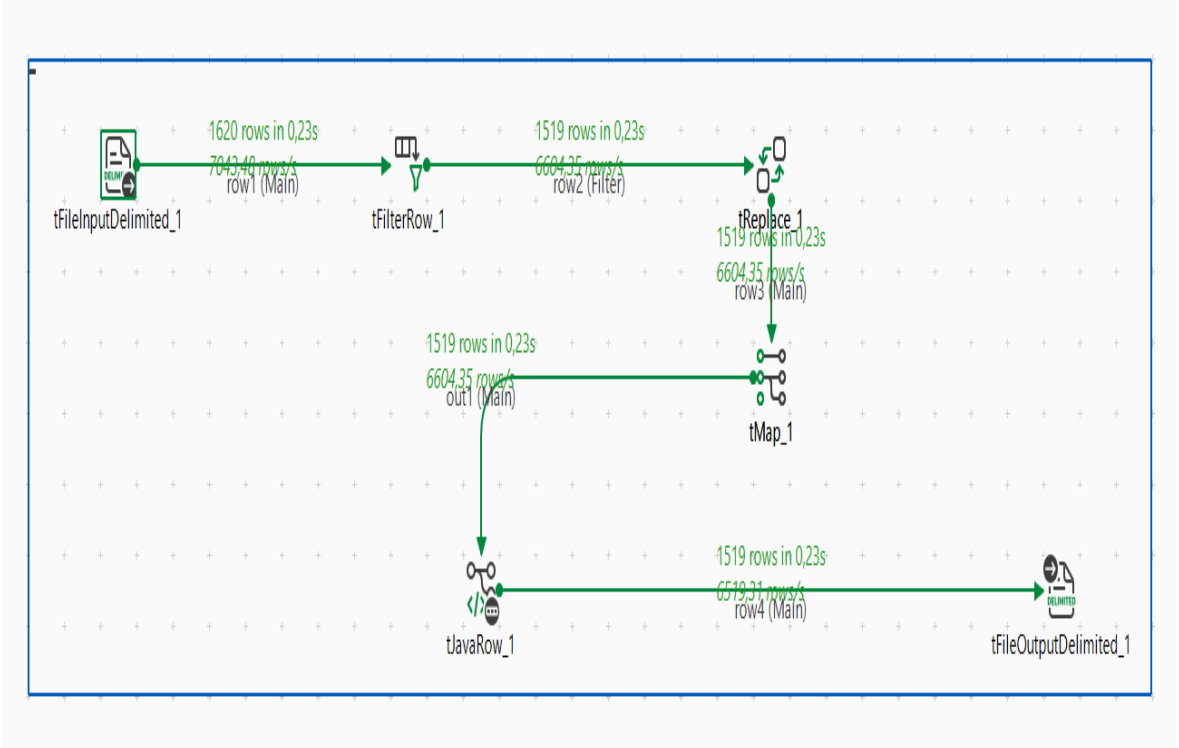

Figura 4.4: Flujo de limpieza introducido en Talend Open Studio.

`tFileInputDelimited → tFilterRow → tReplace → tMap → tJavaRow → tFileOutputDelimited`

Tras la ejecución completa del flujo, se obtuvo un nuevo archivo llamado “out_talend.csv” con las siguientes características:

- Las columnas están correctamente alineadas.
- Se eliminaron filas con datos vacíos en campos esenciales.
- Se corrigieron errores ortográficos simples.
- Se reubicaron valores mal colocados usando lógica condicional.
- Se eliminaron outliers.
- Se garantizó la consistencia del tipo de datos.

Aunque el flujo diseñado fue suficiente para la limpieza propuesta, Talend ofrece una gama mucho más amplia de herramientas que podrían implementarse en proyectos similares:

Tabla 4.3: Otras funcionalidades útiles de Talend Open Studio.

Funcionalidad	**Aplicación potencial**
Conexión con R (“tRserve”)	Envío directo del dataset limpio para análisis estadístico en R.
Carga de múltiples archivos (“tFileList”)	Automatización para importar y procesar decenas o cientos de archivos CSV.
Validación con expresiones regulares (“tMap”)	Verificación de formato de campos como emails, teléfonos,etc.
Agrupaciones de datos (“tAggregateRow”)	Cálculo de métricas (suma, media) por grupo, como ventas por país.
Detección de duplicados (“tUniqRow”)	Filtrado de registros repetidos por claves como “ORDERNUMBER” o “CUSTORMERNAME”.
Inserción en base de datos (“tOutputMysql”, etc.)	Carga directa del resultado a un sistema gestor de bases de datos.
Envío por correo o API (“tSendMail”,tRESTClient”)	Automatización del envío o publicación del resultado.

A continuación se exponen más detalladamente los componentes utilizados en el flujo del proyecto.

4.4.1.1 Componente tFileInputDelimited

El flujo de trabajo comienza por `tFileInputDelimited`, este componente permite importar archivos CSV delimitados. Se ha configurado de la siguiente manera:

- **Ruta del archivo:** archivo original con errores (con_errores.csv).
- **Delimitador del campo:** , (coma).
- **Fila de encabezado:** 1, para leer automáticamente los nombres de columna que tiene el dataset.
- **Codificación:** UTF-8.
- **Esquema definido manualmente:** Se introdujeron los nombres de cada columna.

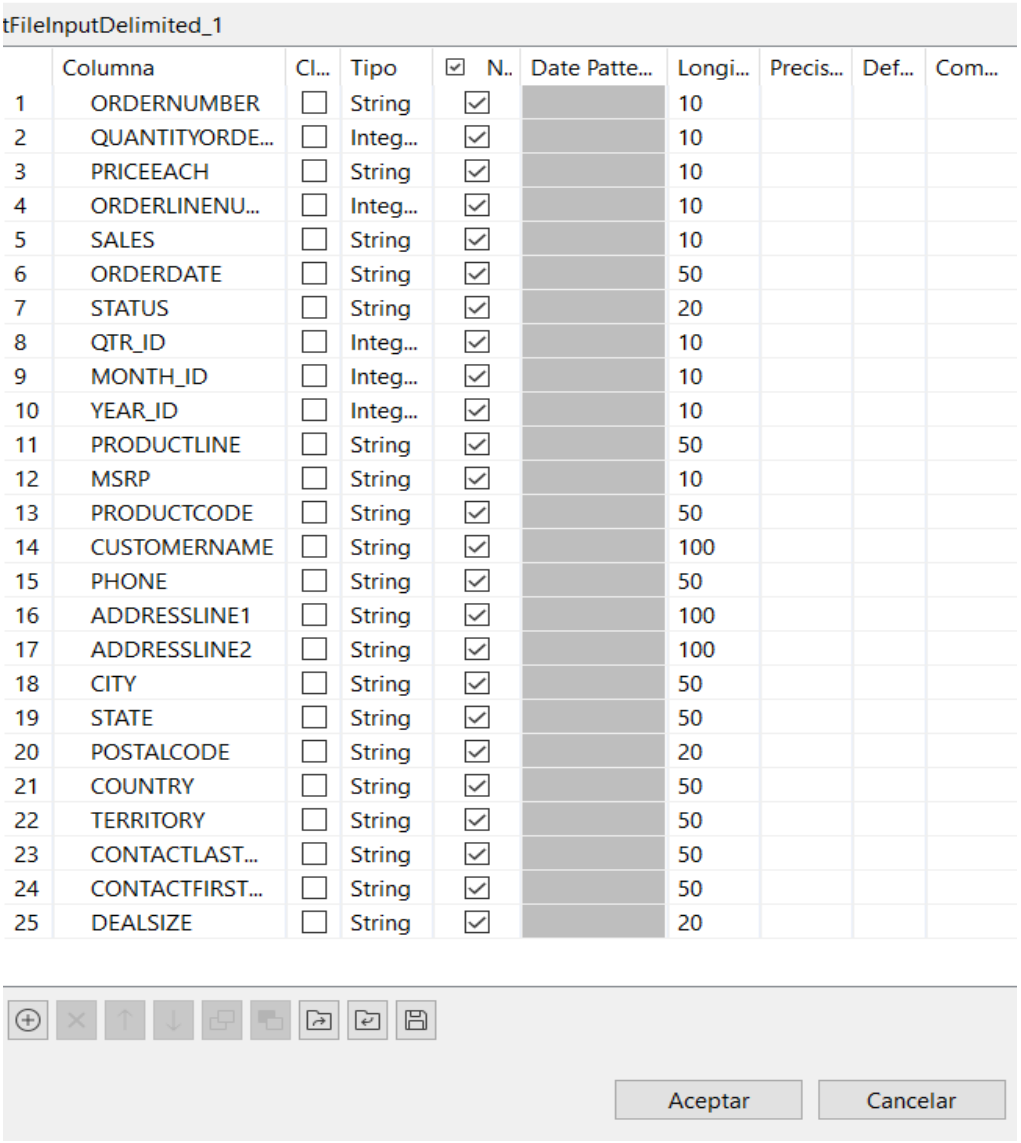

Figura 4.5: Introducción del encabezado del dataset en Talend Open Studio.

4.4.1.2 Componente tFilterNow

El flujo de trabajo continúa hacia el componente `tFilterRow`, utilizado para eliminar registros incompletos o corruptos. Las condiciones de filtrado aplicadas fueron:

- Eliminar registros con campos vacíos en columnas clave como `ORDERNUMBER`, `SALES`, `ORDERDATE`, `CUSTOMERNAME`.

- Se ha utilizado el modo avanzado con expresiones Java:

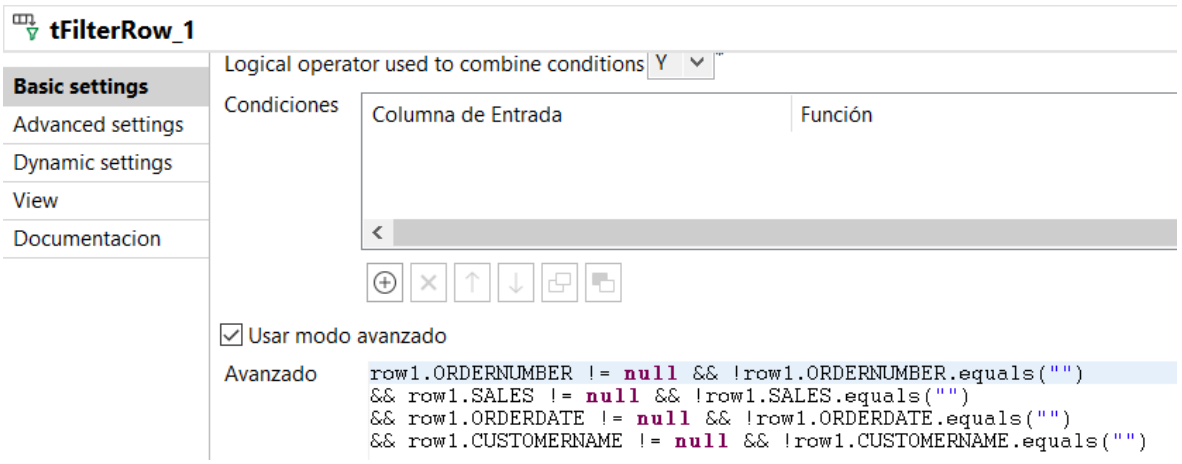

Figura 4.6: : Captura de código introducido en Talend para eliminar registros incompletos.

4.4.1.3 Componente tReplace

El componente tReplace se empleó para corregir errores tipográficos y valores mal escritos mediante reemplazo de texto como por ejemplo:

- `"oMtorcycles"` → `"Motorcycles"`
- `"Classic Casr"` → `"Classic Cars"`

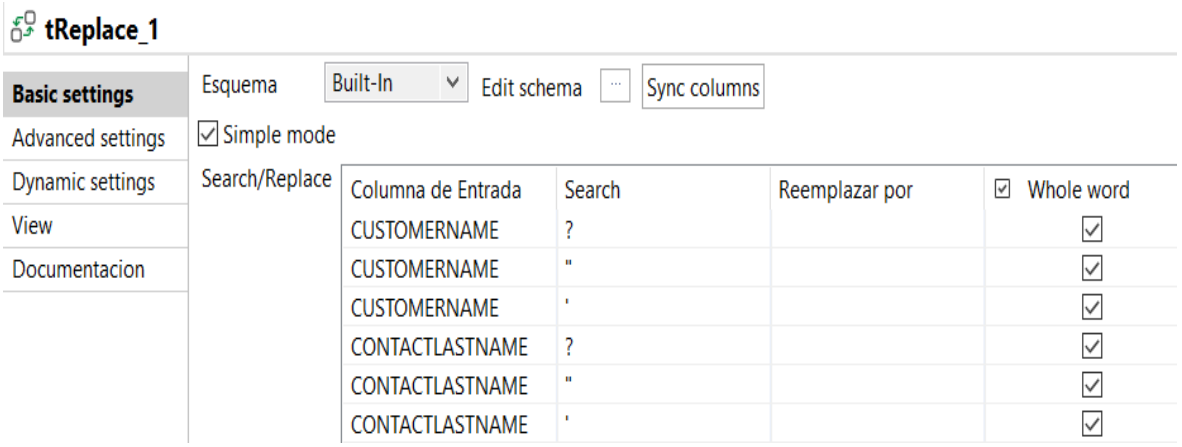

Figura 4.7: Configuración de detección de errores tipográficos en Talend Open Studio.

4.4.1.4 Componente tMap

El componente tMap permitió realizar:

- Limpieza de espacios en blanco con:

  StringHandling.TRIM (row1.CUSTOMERNAME)

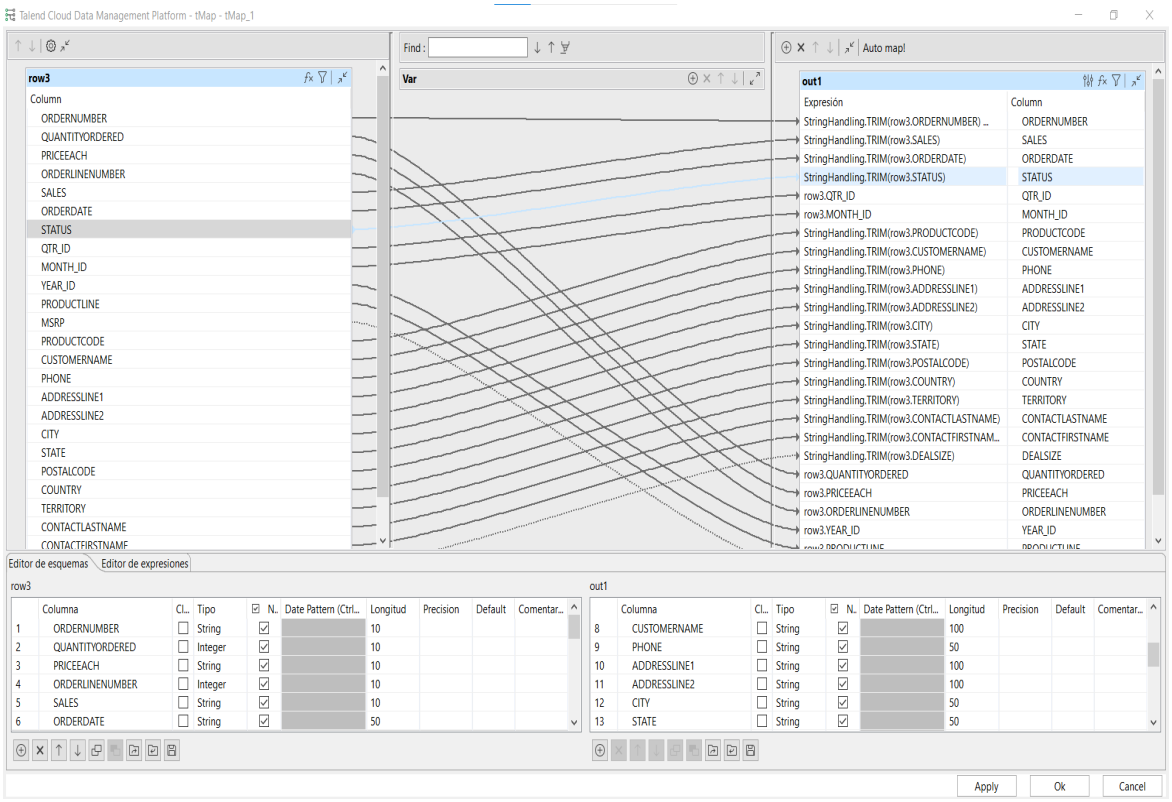

Figura 4.8: Ejecución de limpieza de espacios en blanco en Talend Open Studio.

- Redirección de columnas y renombrado.
- Filtrado de columnas no relevantes para el análisis.
- Descarte de outliers: Se ha calculado el límite superior y se ha añadido la siguiente condición *"row1.sales<=8000"*, en la pestaña "out1", que descarta cualquier fila de la columna "SALES" que tenga un valor mayor a 8000 y así, poder eliminar los outliers que puedan existir.

4.4.1.5 Componente tJavaRow

Este componente permitió aplicar lógica condicional personalizada para detectar registros con valores desplazados, particularmente aquellos donde aparecían nombres en columnas numéricas.

Ejemplo de código utilizado en nuestro proyecto:

```
if (!input_row.PRICEEACH.matches("\\d+(\\.\\d+)?")) {
    output_row.PRICEEACH = "";
    output_row.CUSTOMERNAME = input_row.PRICEEACH;
} else {
    output_row.PRICEEACH = input_row.PRICEEACH;
    output_row.CUSTOMERNAME = input_row.CUSTOMERNAME;
}
```

Este código verifica si la variable “PRICEEACH” no es un número (lo cual delata un error de desplazamiento), y en ese caso traslada el valor erróneo a la columna correcta “CUSTOMERNAME”

4.4.1.6 Componente tFileOutputDelimited

Este componente es el que se ha utilizado para exportar la base de datos limpia en formato CSV.

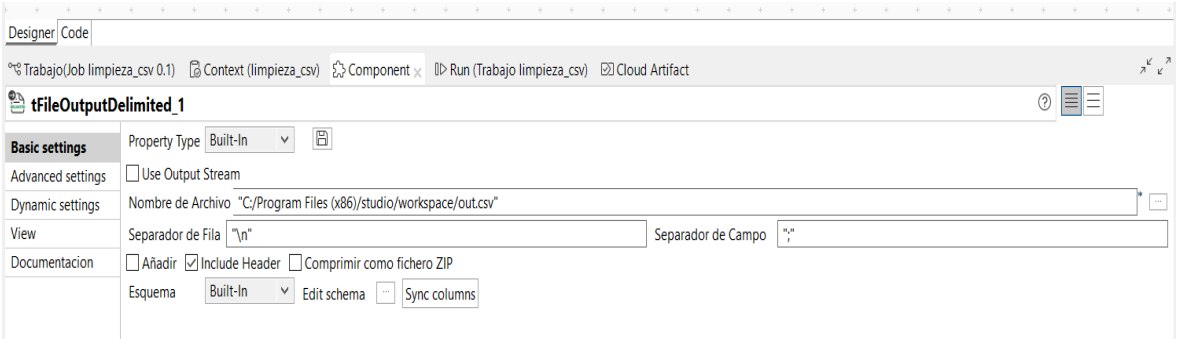

Figura 4.9: Configuración para exportar dataset limpio en Talend Open Studio.

Tras la ejecución completa del flujo, se ha obtenido un nuevo archivo llamado “out_talend.csv” con las siguientes características:

- Las columnas están correctamente alineadas.
- Se eliminaron filas con datos vacíos en campos esenciales.
- Se corrigieron errores ortográficos simples.
- Se reubicaron valores mal colocados usando lógica condicional.
- Se eliminaron valores extremos.
- Se garantizo la consistencia del tipo de datos.

Este es el archivo que se utilizará como base de comparación frente al mismo proceso realizado con la otra herramienta de limpieza, OpenRefine.

4.4.2 Implementación estilo OpenRefine.

La implementación de la limpieza de datos con la herramienta OpenRefine se ha centrado en detectar y corregir errores frecuentes como valores nulos, outliers, duplicados y errores tipográficos. A través de su interfaz gráfica, se realizaron transformaciones sobre los registros afectados, usando funciones como:

- **"Text Facet"** y **"Cluster"** para identificar valores redundantes o inconsistentes.
- **"Edit Cells" → "Transform"** para limpiar espacios, convertir mayúsculas / minúsculas y estandarizar formatos.
- **"Faceta por blancos"** para detectar celdas vacías.
- **"Split"** y **"Join columnas"** para recomponer registros con datos desplazados.

Los pasos seguidos en este proyecto con OpenRefine han sido:

1. Descargar, instalar y abrir OpenRefine.
    a. Al ejecutar el programa se abre en el navegador como una aplicación local.
2. Cargar el archivo con errores que se generó inicialmente.
    a. Hacemos clic en "Crear proyecto".
    b. Seleccionamos "Elegir archivo".
    c. Carga el archivo CSV original ("con_errores.csv").

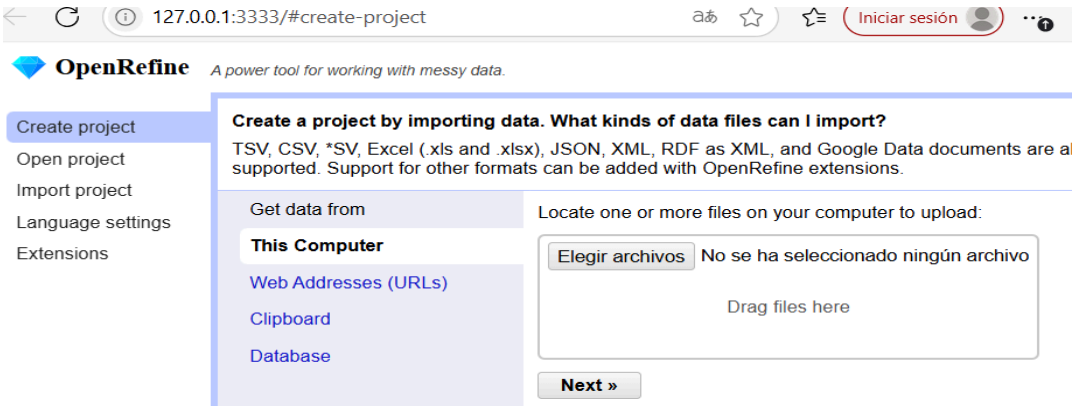

Figura 4.10: Selección de dataset a limpiar en la herramienta OpenRefine.

3. Configurar antes de importar:

    a. Codificación: Seleccionamos UTF-8.
    b. Separador de columnas: aseguramos de que sea (,).
    c. Revisamos si las columnas se han detectado bien.
    d. Luego pulsamos “Crear proyecto”.

4. Comenzamos la limpieza de datos:

    a. Eliminar columnas vacías.

        - En las columnas clave hacemos clic en la flecha de la cabecera.
        - Elegimos “Facet” → “Faceta personalizada” → “Faceta por valores vacíos”.
        - A la izquierda marcamos “true” para ver las filas vacías.
        - Hacemos clic en “Editar filas” → “Eliminar todas las filas coincidentes”.

Figura 4.11: Eliminación de huecos vacíos en OpenRefine.

b. Corregir errores tipográficos:

- En la columna PRODUCTLINE, hacemos clic en el triángulo.
- Elegimos “Editar celdas” → “Transformar”.
- Escribimos con lenguaje GREL:

value.replace(“oMtorcycles”,“Motorcycles”).replace(“Classic Casr”,”Classic Cars”).

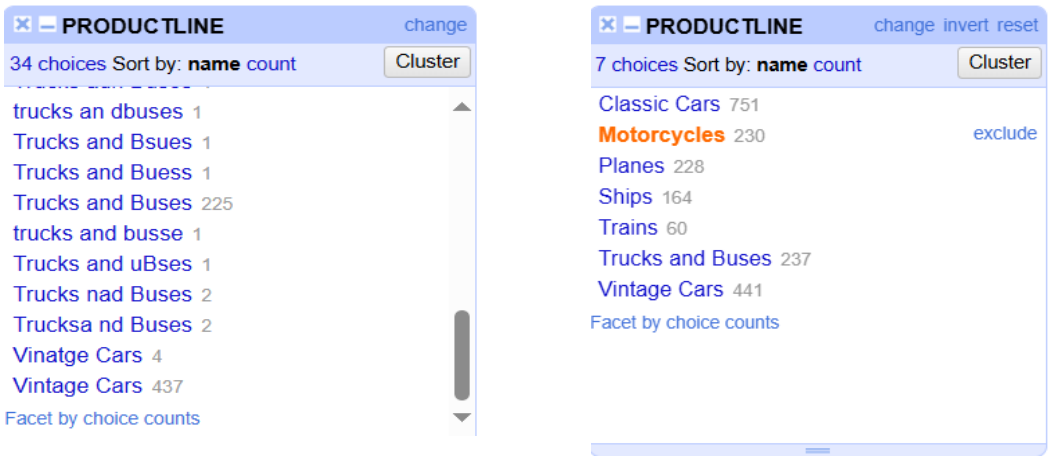

Figura 4.12: Detección de errores sin corregir vs pantalla con errores corregidos OpenRefine.

c. Para cada columna de texto:

- Elegiremos “Editar celdas” → “Transformaciones comunes” → “Eliminar espacios al inicio y al final”.

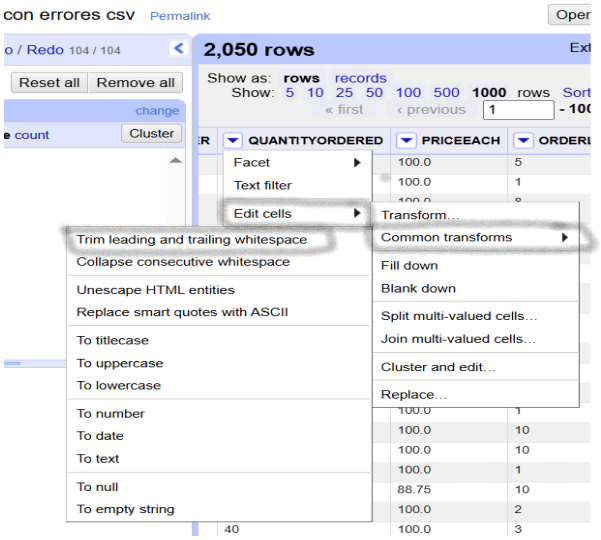

Figura 4.13: Ejecución para eliminar espacios en blanco en OpenRefine.

d. Detectar valores incorrectos o desplazados:

   - En la columna PRICEEACH se ha detectado datos mal ubicados, por lo que en esa columna haremos clic en “Filtrar por texto”
   - Escribimos el valor sospechoso que en este caso es “Martine” o “Peter”
   - Podemos corregir manualmente o mover la información a la columna correcta si es necesario.

e. Identificar si hay outliers:

   - Ordenamos por valores la columna SALES.
   - Hacemos clic en ▼ y seleccionamos “Sort..” → “Sort..”.
   - Ordenamos de mayor a menor.
   - Identificamos si hay outliers y calculamos manualmente para decidir un umbral.

f. Filtramos por condición para eliminar outliers:

   - Clicamos en “Facet” → “Numeric facet”.
   - Esto nos mostrará un rango y seleccionamos los que estén dentro de nuestro rango.
   - Una vez seleccionados hacemos clic en “All” → “Edit rows” → “Remove matching rows”. Esto eliminará los outliers y conservará los registros válidos.

5. Exportamos el archivo limpio.

   a. Hacemos clic en “Exportar” arriba a la derecha.
   b. Elegimos “Valores separados por comas (.csv)”.
   c. Guardamos el archivo como: “out_openrefine.csv”.

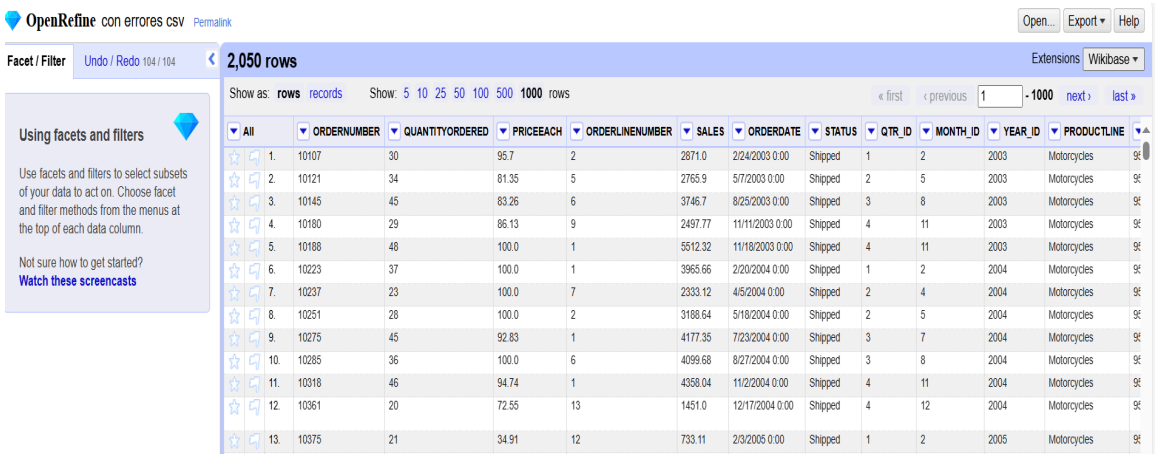

Figura 4.14: Pantalla final OpenRefine con dataset a exportar ya limpiada.

El proceso se llevó a cabo de forma iterativa, registrando cada transformación, lo que garantiza la trazabilidad del proceso.

Esta herramienta demostró ser especialmente útil para identificar errores visualmente, aplicar transformaciones complejas sin necesidad de programación, y mantener una supervisión constante del proceso. Aunque no automatiza flujos, permite una depuración detallada y controlada, adecuada para volúmenes de datos medianos y tareas de limpieza exploratoria.

Justificación del umbral de eliminación en "SALES" para OpenRefine y Talend Open Studio.

Para garantizar la fiabilidad del análisis estadístico, se aplicó el filtro que elimina los valores extremos en la variable SALES. El umbral de 8000 unidades monetarias fue seleccionado como límite superior tras una exploración previa con boxplots y estadísticas en R.

Este valor responde al criterio de Tukey, al situarse por encima del tercer cuartil (Q3) más 1,5 veces el rango intercuartílico (IQR), y además supera el percentil 95 de la distribución. La presencia de estos valores inflaba artificialmente la media y la desviación estándar. Por tanto, su eliminación mejora la consistencia estructural del conjunto de datos sin afectar a su validez analítica.

Este umbral se aplicó de forma coherente tanto en Talend Open Studio como en OpenRefine, lo que garantiza la comparabilidad de los resultados.

4.5. Análisis y resultados

En este apartado se presentan y analizan los resultados obtenidos tras aplicar las herramientas de limpieza de datos seleccionadas, **OpenRefine** y **Talend Open Studio.** A continuación, se realiza una pequeña comparativa entre ambas herramientas, considerando criterios clave en la calidad de los datos y su impacto en los análisis posteriores.

Tabla 4.4: Comparativa de uso de Talend Open Studio vs OpenRefine

Aspecto	Talend Open Studio	OpenRefine
Automatización	✅ Flujo reusable	❌ Manual e interactivo
Visualización	🟠 Limitada (tLogRow)	✅ Vista directa en tabla
Reemplazo masivo	✅ tReplace,tMap	✅ Facilidad con clustering
Procesamiento de grandes datos	✅ Escalable	🟠 Limitado a RAM del navegador
Conexión a R	✅ Via tRserva	❌ No disponible

El uso de Talend permite construir flujos sólidos de limpieza, escalables y reutilizables, ideales para entornos profesionales. Su diseño inicial requiere mayor tiempo de configuración, pero permite una automatización y documentación del proceso más eficiente. Por su parte, OpenRefine destaca por su enfoque interactivo, flexible y orientado al análisis exploratorio o depuración puntual.

Para realizar una evaluación objetiva, se utilizó un archivo con errores (“con_errores.csv”) generado a partir del archivo limpio original (“original.csv”). Este fue creado mediante un script de Python utilizando la librería “pandas” e incorporando intencionalmente errores como:

- **Errores tipográficos:** Se alteraron intencionadamente valores como “Motorcycles” → “oMtorcycles” o “Classic Cars” → “Classic Casr” en la columna “PRODUCTLINE”.

- **Desplazamiento de valores:** Se asignaron datos de texto (por ejemplo: un nombre como “Martine”) a columnas numéricas como “PRICEEACH”, simulando un error de estructura.

- **Campos vacíos:** Se eliminaron valores en columnas clave como “SALES” o “ORDERDATE” para simular registros incompletos.

- **Espacios innecesarios:** Se añadieron espacios al inicio y al final de los valores en la columna “CUSTOMERNAME”.

- **Outliers:** Se añadieron valores elevados en la columna “SALES”.

Tras aplicar los procesos de limpieza con ambas herramientas tal y como hemos visto anteriormente, se obtienen dos archivos:

- **out_talend.csv**: Archivo limpiado mediante Talend.

- **out_openrefine.csv**: Archivo limpiado mediante OpenRefine.

Estos archivos, junto a “**original.csv"**, se van a utilizar para ser analizados con R con el objetivo de evaluar la calidad y consistencia de los resultados.

4.5.1 Comparación de resultados

Se han aplicado ambas herramientas al mismo archivo con errores para asegurar condiciones homogéneas de evaluación. La siguiente tabla resume los resultados observados:

Tabla 4.5: Resultados de limpieza de OpenRefine y Talend Open Studio.

Herramientas	Registros finales	Registros eliminados	Tiempo de procesamiento	Facilidad de uso
OpenRefine	2050	914	4 min	Alta
Talend Open Studio	1468	1496	6 min	Media

Talend logró corregir una mayor cantidad de errores, aunque con una tasa de eliminación de registros también superior. Esto refleja una política de filtrado más estricta y automatizada. OpenRefine, en cambio, permitió conservar un mayor número de registros mediante acciones manuales más precisas.

4.5.2 Evaluación de calidad

La evaluación se ha basado en las dimensiones más utilizadas para medir la calidad del dato: completitud, precisión y consistencia. Se comparan estas métricas antes y después del proceso de limpieza de cada herramienta.

Tabla 4.6: Comparativa calidad de los datos en OpenRefine vs Talend Open Studio

Métrica de calidad	Dataset original	OpenRefine	Talend Open Studio
Completitud	90,1%	100%	100%
Precisión	-	69,16%	49,52%
Consistencia	-	44,90%	44,90%

1. **Completitud**

    - Original → El 9,9% de los registros contenían valores nulos o vacíos.
    - OpenRefine / Talend → 100% Tras la limpieza, todos los valores nulos fueron eliminados o corregidos.

Ambas herramientas restauraron completamente la columna (“SALES”) respecto a la presencia de vacíos.

2. **Precisión**

    - OpenRefine (69,16%) → Conservó 2050 registros.
    - Talend (49,52) → Conservó 1468 registros.

OpenRefine es más selectiva y cuidadosa, eliminando menos registros, sin embargo, **Talend** es más estricta y automatizada, eliminando casi la mitad. Puede interpretarse como mayor limpieza pero también mayor pérdida de datos.

3. **Consistencia**

    - Ambas herramientas han reducido el rango de “SALES” en un 44,9%, lo que significa que eliminaron valores extremos o incoherentes mejorando la consistencia estructural.

En conclusión ambas herramientas han mejorado de forma significativa la calidad del dataset, aunque con enfoques diferentes. Las tres métricas muestran mejoras cuantificables: aumento de completitud, mejora de consistencia y reducción de errores.

4.5.3 Análisis estadístico del dataset en R

Para validar los resultados desde una perspectiva cuantitativa, se ha empleado R como herramienta de análisis estadístico. Se han cargado los tres datasets (“con_errores”, “out _openrefine” y “out_talend”) y se ha centrado el análisis en la variable **“SALES”**, ya que fue afectada directamente por errores de vacío, desplazamiento, outliers y registros inconsistentes.

Se aplicaron los siguientes pasos:

- **Estadísticos descriptivos:** Media, desviación estándar, mínimo y máximo de la variable SALES.

- **Visualización:** Gráfico boxplot comparando la distribución de SALES en cada dataset.

- **Análisis de varianza (ANOVA):** Para comprobar si las diferencias entre medias son estadísticamente significativas.

Tabla 4.7: Análisis de la variable SALES en Rstudio.

Dataset	Registros totales	Media SALES	Desv. estándar	Mínimo	Máximo
original.csv	2964	3565	1851	482	14083
out_talend.csv	1468	3411	1572	482	7975
out_openrefine.csv	2050	3410	1598	482	7975

Los resultados obtenidos tras el análisis en R permiten afirmar que tanto Talend como OpenRefine contribuyen de manera significativa a la mejora de calidad del dataset original.

Esta afirmación se sustenta en los siguientes aspectos observables:

1-Disminución de la desviación estándar.

El dataset original presenta una desviación estándar elevada en la variable SALES (1851), lo cual refleja una gran dispersión en los valores, posiblemente causada por errores o valores atípicos. Tras la limpieza, esta métrica se consigue reducir a 1572 en Talend y 1598 en OpenRefine, indicando una mayor homogeneidad y menor presencia de valores anómalos.

2-Reducción del valor máximo.

El valor máximo de SALES baja de 14083 en el archivo original ("con_errores.csv") a 7975 en ambos datasets limpiados. Esto sugiere que se eliminaron registros que contenían errores o valores extremos que estaban distorsionando la estructura del conjunto de datos.

3-Reducción de registros defectuosos.

Se eliminaron 1496 registros con Talend y 914 con OpenRefine respecto al archivo original. Estas supresiones no redujeron la validez del análisis, sino que permiten conservar únicamente los datos útiles y bien estructurados, mejorando así la calidad general.

4-Análisis de varianza (ANOVA).

Para comprobar si las diferencias entre los tres conjuntos son estadísticamente significativas, se aplicó una prueba ANOVA sobre la variable "SALES". El resultado obtenido fue un valor-p = 0.0043, lo que confirma que las diferencias entre las medias no son aleatorias y que los procesos de limpieza han tenido efecto real y medible sobre la calidad del dato con diferencias significativas entre los tres grupos analizados con un 99,7% de confianza.

Visualización con boxplot

Se generó un gráfico boxplot con la librería ggplot2 que muestra la distribución de la variable "SALES" en los tres archivos. Este gráfico permite observar:

- La mediana de cada conjunto de datos.
- La dispersión de los valores.
- La existencia de posibles valores atípicos (outliers).

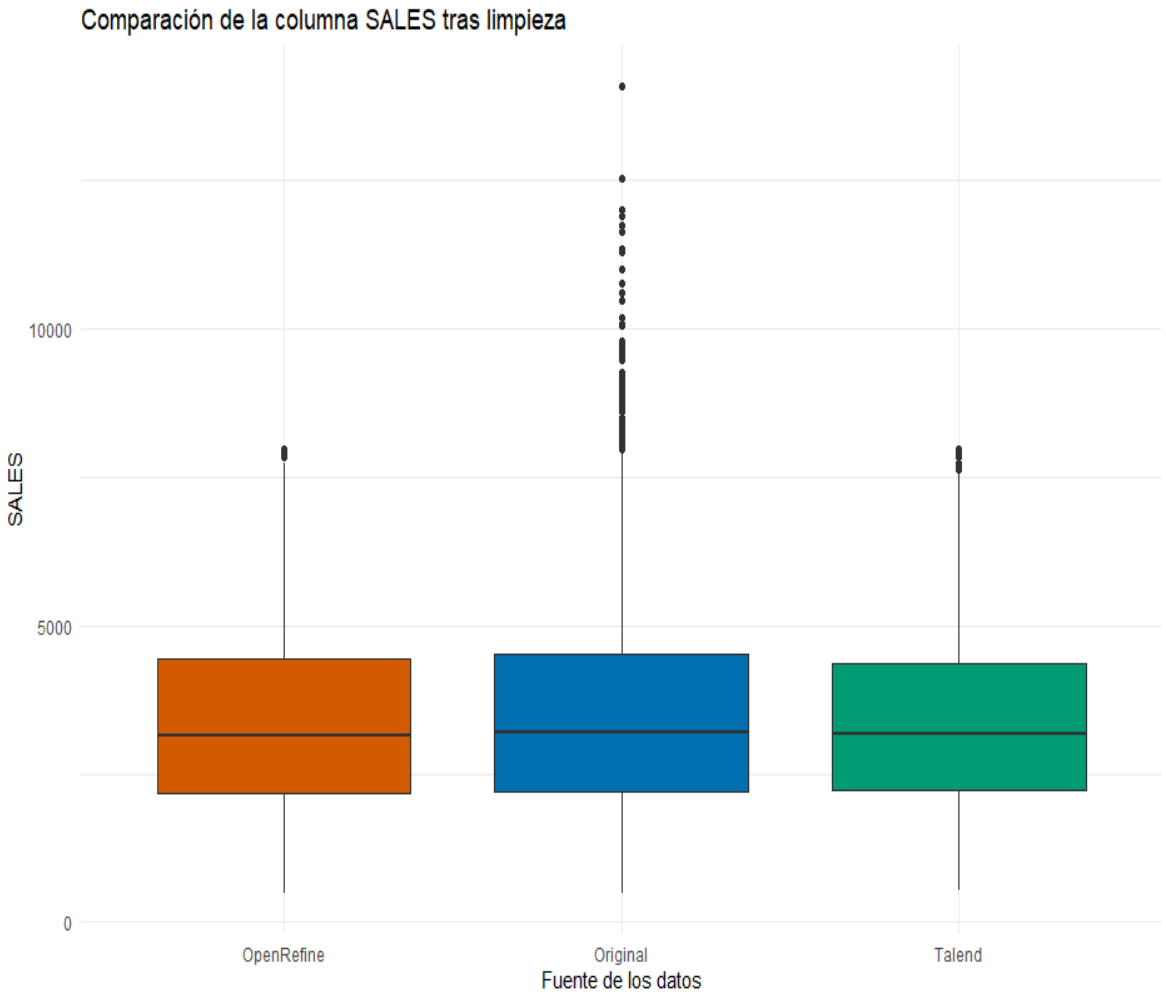

Figura 4.15: Gráfico boxplot de la variable "SALES" en los tres datasets.

En la figura 4.15 se aprecia que el dataset original presenta una mayor variabilidad y más valores extremos, mientras que las versiones limpiadas muestras distribuciones más compactas, con menor presencia de outliers, reflejando una estructura más consistente.

En conjunto, como resultado principal se concluye que **ambas herramientas mejoraron la calidad del dataset** en términos de consistencia, precisión y validez estadística. Mientras Talend se mostró más riguroso y automatizado en la limpieza, lo que llevó a una depuración más agresiva, OpenRefine permitió un enfoque más controlado y manual, conservando más registros.

- Talend eliminó más registros debido a filtros automáticos estrictos, mejorando la consistencia estadística general.

- OpenRefine permitió una limpieza manual más controlada, lo que resultó en un mayor número de registros conservados, pero también un nivel ligeramente inferior de consistencia estructural.

Ambos métodos mejoraron el dataset original y son válidos y complementarios, aunque con enfoques diferentes. Su elección dependerá de los objetivos del análisis y el grado de intervención manual que se desee aplicar. OpenRefine es más adecuado cuando se busca control manual y comprensión del proceso mientras que Talend es ideal para flujos complejos, repetitivos y de gran escala, con una depuración más automatizada.

A continuación, se muestra una tabla resumen de las comparaciones principales de ambas herramientas:

Tabla 4.8: Tabla resumen comparativa de OpenRefine vs Talend Open Studio

Criterio	**OpenRefine**	**Talend Open Studio**
Tipo de limpieza	Manual / Semiautomática	Automática / Basada en flujos
Nivel de control de usuario	Alto (intervención directa)	Medio (reglas predefinidas y filtros)
Registros eliminados	914	1496
Media variable SALES	3410	3411
Valor máximo variable SALES	7975	7975
Valor mínimo variable SALES	482	482
Desviación estándar	1598	1572
Mejora en consistencia estructural	Media	Alta
Preservación de datos	Alta	Media
Visualización de datos y patrones	Excelente (interfaz interactiva)	Limitada (requiere configuración)
Escalabilidad	Media (para datasets medianos)	Alta (requiere conocimientos técnicos)
Adecuado para limpieza personalizada	Sí	No (más útil en procesos repetitivos)
Exportación e integración	Buena (CSV, Excel, RDF, JSON)	Muy buena (bases de datos, servicios web)
Curva de aprendizaje	Baja a media	Alta (requiere conocimientos técnicos)

4.5.4 Valoración práctica y experiencia de uso

Más allá de los resultados cuantitativos, se ha realizado una valoración práctica del uso de las herramientas, teniendo en cuenta la facilidad de uso, el tiempo de aprendizaje y la experiencia durante el proceso de limpieza.

OpenRefine resultó más accesible, gracias a su interfaz visual sencilla, operaciones intuitivas como el "clustering" y la posibilidad de trabajar directamente sobre los datos. Fue especialmente útil para detectar errores manuales, explorar patrones de inconsistencia y aplicar transformaciones controladas. Su punto fuerte es la rapidez con la que se pueden aplicar correcciones puntuales y la posibilidad de revisar cada paso del proceso.

Talend Open Studio, en cambio, requirió una mayor inversión inicial en aprendizaje. La herramienta ofrece una arquitectura visual basada en flujos ETL, con numerosos componentes que permiten realizar tareas complejas de forma automatizada. Aunque su configuración inicial fue más compleja, una vez establecidos los procesos, resultó muy potente para limpiar y transformar grandes volúmenes de datos con reglas repetibles.

Durante el desarrollo del trabajo, se observó que:

- OpenRefine fue más cómodo para tareas exploratorias y depuración puntual.
- Talend es más adecuado para entornos laborales donde se necesita repetibilidad, automatización y escalabilidad.
- Ambas herramientas requieren conocimientos previos, pero OpenRefine tiene una curva de aprendizaje más suave para usuarios no técnicos.

En un entorno académico o exploratorio, OpenRefine puede ser más eficiente y manejable. En cambio, Talend resulta más útil en proyectos de mayor envergadura, con necesidades de integración, automatización y procesamiento estructurado de datos.

Capítulo 5
Conclusiones y trabajo futuro

5.1.- Conclusiones

El desarrollo de este trabajo ha permitido cumplir con los objetivos inicialmente planteados, demostrando la relevancia y complejidad de la calidad de los datos en entornos reales. A lo largo del proyecto se ha logrado diseñar, ejecutar y evaluar un experimento controlado que compara el rendimiento de distintas herramientas de calidad de datos, centrándose especialmente en aquellas de código abierto por su accesibilidad y aplicabilidad práctica.

La limpieza de datos no es una tarea trivial. Este proyecto ha puesto de manifiesto que errores comunes como valores nulos, duplicados, errores tipográficos o inconsistencias de formato pueden alterar significativamente la fiabilidad y usabilidad de los datos. En este sentido, el uso de herramientas especializadas es crucial para garantizar resultados precisos y decisiones informadas.

Entre los principales logros del trabajo se destacan:

- La identificación y caracterización de los errores más comunes en datasets reales.
- La aplicación práctica de dos herramientas de código abierto (OpenRefine y Talend Open Studio) sobre un conjunto de datos con errores simulados.
- La validación de hipótesis mediante un enfoque mixto, que ha combinado métricas estadísticas objetivas con la valoración subjetiva de la experiencia de uso.
- La comparación de resultados a través de visualizaciones, medidas de dispersión y análisis de varianza (ANOVA).

Los resultados obtenidos muestran que ambas herramientas han mejorado la calidad del dataset respecto a su versión original con errores, aunque lo han hecho con enfoques distintos: OpenRefine destaca en tareas de limpieza puntual, mientras que Talend sobresale en flujos automatizados de transformación. Esta diferencia confirma que no existe una solución única, sino que la elección de la herramienta depende del contexto, los objetivos y el perfil del usuario que la vaya a usar.

Además, el trabajo ha contribuido a generar un marco experimental replicable que puede ser reutilizado en otros estudios académicos o en entornos empresariales que deseen evaluar herramientas sin necesidad de realizar grandes inversiones.

En resumen, se han logrado alcanzar los objetivos específicos, transversales y personales planteados anteriormente y se ha reafirmado la importancia de integrar procesos de calidad del dato en cualquier estrategia analitica o de transformación digital.

5.2.- Posibles desarrollos futuros

A partir de la experiencia obtenida, he identificado diversas líneas de mejora y expansión que podrían enriquecer el trabajo realizado:

1. **Ampliación del análisis a herramientas comerciales:** Si se cuenta con acceso a licencias, se podría replicar el experimento con IBM InfoSphere QualityStage y Ataccama ONE para obtener una versión más completa del mercado y validar su rendimiento práctico en los mismos escenarios.

2. **Evaluación en entornos reales:** Aplicar el experimento a datasets reales de empresas o instituciones permitiría medir la eficacia en contextos de mayor complejidad, con datos sensibles o de múltiples fuentes.

3. **Desarrollo de una interfaz educativa:** Crear una interfaz gráfica sencilla que permita a estudiantes o usuarios no técnicos experimentar con herramientas de calidad del dato de forma guiada.

4. **Análisis del impacto:** Evaluar cómo la mejora de la calidad de los datos afecta al rendimiento de modelos de clasificación, regresión o clustering entrenados con los mismos datos antes y después de ser limpiados.

5. **Automatización del flujo completo**: Diseñar estructuras de proyectos reutilizables que incluyan no solo la limpieza, sino también la validación, la carga y la actualización automática de datasets.

6. **Construcción de un marco de indicadores de calidad:** Desarrollar un sistema de métricas estándar que mida objetivamente la mejora obtenida con cada herramienta.

Estas posibles ampliaciones no sólo enriquecerán el alcance del proyecto, sino que podrían derivar en nuevas líneas de investigación.

Con ello, se abre la puerta a futuras contribuciones que fortalezcan la conciencia sobre la importancia de los datos de calidad y faciliten su integración efectiva en los procesos de decisión y análisis digital.

Bibliografía

1. Ataccama. (s.f.). Ataccama ONE platform overview. Recuperado de: https://www.ataccama.com

2. Ataccama. (s.f.). Ataccama ONE product overview. Recuperado de: https://www.ataccama.com/products/ataccama-one

3. Basel Committee on Banking Supervision. (2013). Principles for effective risk data aggregation and risk reporting (BCBS 239). Bank for International Settlements. https://www.bis.org/publ/bcbs239.htm

4. Batini, C., & Scannapieco, M. (2016). Data and information quality: Dimensions, principles and techniques. Springer.

   https://link.springer.com/book/10.1007/978-3-319-24106-7

5. Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approaches (4.ª ed.). SAGE Publications.

6. DAMA International. (2017). DAMA-DMBOK: Data management body of knowledge (2.ª ed.). Technics Publications.

7. European Union. (2016). Reglamento general de protección de datos (RGPD) – Reglamento (UE) 2016/679. Diario Oficial de la Unión Europea. https://eur-lex.europa.eu/eli/reg/2016/679/oj

8. Field, A. (2018). Discovering statistics using R (2.ª ed.). SAGE Publications.

9. Gartner. (2022). Magic quadrant for data quality solutions. Gartner Inc.

10. Google Developers. (s.f.). OpenRefine documentation. Recuperado de: https://openrefine.org/

11. IBM. (s.f.). IBM InfoSphere QualityStage documentation. Recuperado de: https://www.ibm.com/docs/en/iis

12. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning (2.ª ed.). Springer. https://www.statlearning.com

13. Kaggle. (s.f.). Customer sales data dataset. Recuperado de: https://www.kaggle.com

14. Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data. Wiley.

https://www.wiley.com/en-us/The+Data+Warehouse%C2%A0ETL+Toolkit%3A+Practical+Techniques+for+Extracting%2C+Cleaning%2C+Conforming%2C+and+Delivering+Data-p-9780764567575

15. Martínez González, M. Á. (2004). Diseño de estudios clínicos. Elsevier España.

16. McKinney, W. (2018). Python for data analysis: Data wrangling with pandas, NumPy, and IPython (2.ª ed.). O'Reilly Media.

https://wesmckinney.com/book/

17. OpenRefine. (s.f.). Documentation: User manual and tutorials. Recuperado de:

https://openrefine.org/docs

18. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218. https://doi.org/10.1145/505248.506010

19. Pressman, R. S., & Maxim, B. R. (2015). Ingeniería del software: Un enfoque práctico (8.ª ed.). McGraw-Hill.

20. Redman, T. C. (2001). Data quality: The field guide. Digital Press.

21. Redman, T. C. (2017). Data driven: Profiting from your most important business asset. Harvard Business Review Press.

22. Scannapieco, M., & Batini, C. (2005). Data quality: Concepts, methodologies and techniques. Data & Knowledge Engineering, 55(1), 3–15. https://doi.org/10.1016/j.datak.2004.11.004

23. Talend. (s.f.). Advanced features and component reference guide. Recuperado de: https://help.talend.com/

24. Talend. (s.f.). Talend Open Studio for Data Integration – User guide. Recuperado de: https://help.talend.com/

25. Wickham, H., & Grolemund, G. (2017). R for data science: Import, tidy, transform, visualize,and model data. O'Reilly Media.

https://r4ds.hadley.nz/

Anexos

- **Anexo I. Script en Python para generación de errores.**

A continuación se muestra un resumen del script utilizado para modificar el conjunto de datos original e introducir errores de forma controlada con la librería “pandas”.

```
import pandas as pd
import random

df = pd.read_csv("original.csv")
df.loc[5, 'PRODUCTLINE'] = 'oMtorcycles'
df.loc[10, 'PRODUCTLINE'] = 'Classic Casr'
df.loc[15, 'PRICEEACH'] = 'Martine'
df.loc[20, 'SALES'] = ''
df.loc[25, 'ORDERDATE'] = ''
df['CUSTOMERNAME'] = df['CUSTOMERNAME'].apply(lambda x: f" {x} " if
random.random() < 0.05 else x)

# Añadir outliers
for i in random.sample(range(len(df)), int(0.01 * len(df))):
       df.at[i, 'SALES'] = random.randint(10000, 20000)

df.to_csv("con_errores.csv", index=False)
```

- **Anexo II. Script en R para análisis estadístico.**

Este script resumido realiza un análisis comparativo de calidad de datos entre el dataset original y las versiones procesadas con Talend Open Studio y OpenRefine.

```
# ------------------------------------------------------
# Script de análisis estadístico de calidad de datos en R
# Autor: [Carolina Isabel ]
# Fecha: [Junio, 2025]
# Objetivo: Comparar la calidad del dataset original frente a
# versiones limpiadas con Talend Open Studio y OpenRefine.
# ------------------------------------------------------------
library(readr)
library(dplyr)
library(janitor)
library(ggplot2)

original <- read_csv("con_errores.csv")
talend <- read_delim("out_talend.csv", delim = ";")
refine <- read_delim("out_openrefine1.csv", delim = ";")

original <- clean_names(original)
talend <- clean_names(talend)
refine <- clean_names(refine)

eliminadas_por_talend <- anti_join(original, talend)
eliminadas_por_refine <- anti_join(original, refine)


# Función resumen de estadísticas
 stat_summary <- function(df, label) {
   df %>% summarise(
       Registros = n(),
       Media_SALES = mean(sales, na.rm = TRUE),
       SD_SALES = sd(sales, na.rm = TRUE),
       Min_SALES = min(sales, na.rm = TRUE),
       Max_SALES = max(sales, na.rm = TRUE)
   ) %>% mutate(Fuente = label)
 }

 # Análisis, visualización y métricas adicionales
```