Diseño e implementación de un proceso ETL en Python para la consolidación y análisis de datos empresariales, orientado al apoyo a la toma de decisiones mediante Power BI

Jeison Andres Gamba Alvarado

Asesor

Lina Rocio Rivadeneira Muñoz

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Resumen

El proyecto aplicado tiene como objetivo diseñar e implementar un proceso ETL en Python (Pandas) para integrar datos empresariales provenientes de múltiples archivos CSV, con el fin de identificar duplicidades, inconsistencias y errores de digitación, y presentar los resultados mediante un dashboard en Power BI orientado a la gerencia. La iniciativa se desarrolla en la empresa Gestiones y Representaciones Chía S.A.S., que gestiona información de 43 empresas clientes y actualmente enfrenta fragmentación de datos, redundancias y dificultades para el análisis global. El proyecto definirá estándares técnicos de calidad de datos, documentará el flujo ETL a través de técnicas de profiling data y consolidará la información para evaluar su estado. Finalmente, el resultado se enfoca en visualizar y dimensionar el esfuerzo de corrección requerido, establecer pautas de mejora y fortalecer la confiabilidad de la información, contribuyendo a una mayor eficiencia operativa y calidad en la gestión de datos a través de visualizaciones que permitan apoyar la toma de decisiones estratégicas.

***Palabras clave*:** ETL, Data Profiling, Calidad de datos, Integración de datos, Inteligencia de negocios.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Ápendices

Introducción

En la actualidad, las organizaciones gestionan grandes volúmenes de información provenientes de múltiples fuentes, lo que supone una oportunidad clave para la toma de decisiones, pero también un desafío cuando los datos presentan problemas de calidad. Errores de digitación, valores nulos, duplicidades e inconsistencias pueden afectar la confiabilidad de los análisis y limitar el aprovechamiento de la información disponible. Por ello, resulta fundamental contar con mecanismos que permitan evaluar de forma objetiva el estado de los datos antes de su uso analítico.

En este contexto, los procesos ETL (Extract, Transform, Load) juegan un papel central en la integración y análisis de la información empresarial, no solo como herramienta de consolidación, sino también como apoyo para la identificación y cuantificación de problemas de calidad. Complementariamente, las técnicas de data profiling permiten analizar la estructura y el contenido de los datos, facilitando la definición de métricas relacionadas con su completitud, unicidad, consistencia y validez.

Este proyecto tiene como objetivo diseñar e implementar un proceso ETL en Python orientado al diagnóstico de la calidad de datos empresariales, a partir de la integración de 43 bases de datos en formato CSV. El enfoque se centra en la identificación y clasificación de errores mediante reglas de perfilamiento, sin realizar correcciones automáticas, con el fin de proporcionar un diagnóstico claro que sirva como insumo para la toma de decisiones. Los resultados se presentan mediante un dashboard en Power BI, facilitando la visualización de los indicadores de calidad y apoyando la gestión de la información dentro de la organización.

Justificación

El proyecto aplicado tiene como objetivo diseñar e implementar un proceso ETL en Python, utilizando la librería Pandas, para consolidar 43 bases de datos empresariales en formato CSV gestionadas por la empresa Gestiones y Representaciones Chía S.A.S., clasificar errores de calidad de datos como duplicados, inconsistencias, valores faltantes y formatos inválidos y cuantificar su frecuencia y proporción. Los resultados del proceso ETL se presentarán mediante un dashboard en Power BI orientado a la gerencia, con el fin de apoyar la toma de decisiones para la mejora de la calidad de la información. A partir de la definición de estándares técnicos de calidad y la documentación del flujo ETL, el proyecto busca mitigar la fragmentación de datos existente, fortalecer la confiabilidad de la información y contribuir a una mayor eficiencia operativa en la gestión de datos empresariales.

Objetivos

Objetivo General

Diseñar e implementar un proceso ETL en Python que permita integrar y analizar la información empresarial proveniente de múltiples fuentes, con el propósito de identificar errores, duplicidades y problemas de digitación en los datos, generando visualizaciones en Power BI que apoyen la toma de decisiones orientadas a la mejora de la calidad de la información.

Objetivos Específicos

Implementar un proceso ETL en Python que permita consolidar las 43 bases de datos empresariales y clasificar los errores de los registros mediante reglas de perfilamiento de datos, garantizando la trazabilidad por fuente

Analizar y cuantificar los errores identificados en el proceso ETL mediante el cálculo de indicadores de calidad de datos, tales como frecuencia y proporción de errores por tipo y por base de datos.

Diseñar un dashboard en Power BI que visualice los indicadores de calidad de los datos

Marco de Referencia

Estado del Arte

La literatura reciente en inteligencia de negocios y gestión de datos coincide en que la calidad de la información es un factor crítico para la toma de decisiones organizacionales. Desde los primeros enfoques de integración de datos, los procesos ETL han sido reconocidos como el eje central para consolidar información proveniente de múltiples fuentes heterogéneas (Kimball & Caserta, 2004). En este contexto, diversos estudios han señalado que los problemas de calidad como duplicidades, inconsistencias y errores de digitación que no suelen detectarse adecuadamente sin un análisis previo del contenido de los datos.

El data profiling surge como una respuesta a esta necesidad diagnóstica. Müller et al. (2003) lo definen como un conjunto de técnicas orientadas a examinar los datos para descubrir su estructura real, patrones recurrentes y anomalías. Posteriormente, Abedjan, Golab y Naumann (2015) sistematizan estas técnicas en su estudio sobre profiling de datos relacionales, proponiendo una taxonomía ampliamente aceptada que distingue entre profiling a nivel de columna, multicolumna, tabla y basado en reglas. Estos trabajos evidencian que el profiling constituye una etapa previa indispensable para evaluar la calidad de los datos antes de su integración y análisis.

En contextos de grandes volúmenes de información, algunos autores han explorado el uso del muestreo para reducir costos computacionales en tareas de profiling. Zhang et al. (2018) analizan esta aproximación y concluyen que, aunque el muestreo puede mejorar la eficiencia, el valor principal del data profiling reside en su capacidad diagnóstica más que en la optimización del rendimiento. En proyectos de alcance empresarial moderado, el profiling completo sigue siendo una práctica viable y recomendable.

Adicionalmente, estudios recientes en gobernanza de datos advierten que la falta de control sobre la calidad y estructura de la información incrementa riesgos operativos, financieros y reputacionales para las organizaciones (Alhassan & Sammon, 2020). En este sentido, la visualización de métricas de calidad mediante herramientas de BI ha sido identificada como un mecanismo efectivo para comunicar hallazgos técnicos a niveles directivos (von Enzberg et al., 2024).

Marco Contextual

El proyecto se desarrolla en el contexto de la empresa Gestiones y Representaciones Chía S.A.S., organización que administra información proveniente de 43 empresas clientes. Actualmente, dicha información se encuentra distribuida en múltiples archivos CSV, generados de forma independiente y sin estándares homogéneos de calidad o estructura. Esta fragmentación dificulta el análisis global de los datos y limita la capacidad de la gerencia para tomar decisiones informadas.

Situaciones similares han sido documentadas en múltiples organizaciones que gestionan datos empresariales desde diversas fuentes operativas, donde la ausencia de procesos formales de integración y diagnóstico de calidad deriva en redundancias, inconsistencias y pérdida de confianza en la información (Redman, 2013). En este escenario, la implementación de un proceso ETL acompañado de técnicas de data profiling se presenta como una solución adecuada para consolidar la información y evaluar su estado real antes de cualquier iniciativa de mejora.

Marco Teórico

Desde el enfoque de Business Intelligence, la información adquiere valor cuando es transformada en conocimiento útil para la toma de decisiones estratégicas. El BI se apoya en

procesos de integración de datos que permiten consolidar información dispersa y garantizar su coherencia interna (Kimball & Caserta, 2004).

El proceso ETL constituye el componente técnico central de esta integración. Su función no se limita a la extracción y carga de datos, sino que incluye actividades de transformación orientadas a la depuración, estandarización y validación de la información. La literatura especializada destaca que los procesos ETL bien diseñados permiten identificar problemas de calidad y documentar su impacto en los sistemas de información (Airbyte, 2025; EM360Tech, 2025).

La calidad de los datos se aborda desde modelos formales que definen dimensiones como completitud, unicidad, consistencia y validez (Batini & Scannapieco, 2016; ISO/IEC 25012, 2008). Estas dimensiones permiten evaluar de manera objetiva el estado de la información y establecer criterios medibles para su análisis. La ausencia de calidad en los datos puede derivar en decisiones erróneas, reprocesos y pérdidas económicas, tal como lo señalan organizaciones especializadas en ciencia de datos (DASCA, 2025).

En este marco, el data profiling se concibe como una técnica analítica que apoya los procesos ETL, proporcionando un diagnóstico previo del contenido y la estructura de los datos. Abedjan et al. (2015) señalan que el profiling no sustituye al ETL, sino que lo fortalece al orientar la definición de reglas de transformación y validación

Marco Conceptual

Para efectos de este proyecto, se adoptan los siguientes conceptos operativos:

1. Business Intelligence (BI): conjunto de metodologías y herramientas orientadas a transformar datos empresariales en información útil para la toma de decisiones estratégicas.

2. ETL (Extract, Transform, Load): proceso mediante el cual se extraen datos desde múltiples fuentes, se transforman mediante reglas de limpieza y estandarización, y se cargan en un repositorio consolidado.

3. Calidad de datos: grado en que la información es precisa, completa, consistente, válida y confiable para su uso previsto.

4. Data profiling: conjunto de técnicas analíticas utilizadas para examinar los datos con el fin de identificar patrones, anomalías y problemas de calidad. En este proyecto se aplican:

a. Profiling a nivel de columna, para detectar valores nulos, tipos de datos reales y formatos inválidos.

b. Profiling a nivel multicolumna, para identificar coherencia entre campos relacionados.

c. Profiling a nivel de tabla, para evaluar volumen de registros y duplicidades.

d. Profiling basado en reglas, para validar formatos y restricciones definidas.

Estas técnicas se implementan en Python mediante la librería Pandas, alineándose con los objetivos del proyecto de identificar, clasificar y cuantificar errores sin realizar correcciones automáticas.

5. Visualización de datos: proceso de representación gráfica de los resultados analíticos mediante dashboards, que facilita la interpretación gerencial y apoya la toma de decisiones (von Enzberg et al., 2024).

Marco Normativo

Aunque el proyecto utiliza datos simulados con fines académicos, su contexto real implica el tratamiento de información sensible de proveedores y clientes. En Colombia, la Ley

1581 de 2012 establece los principios que rigen el tratamiento de datos personales, incluyendo legalidad, finalidad, confidencialidad y seguridad.

La utilización de datos artificiales en este proyecto responde a la necesidad de cumplir con dichos principios y minimizar riesgos legales y reputacionales. Estudios recientes en gobernanza de datos resaltan que el manejo inadecuado de información personal puede generar consecuencias financieras y legales significativas para las organizaciones (Alhassan & Sammon, 2020). Por ello, el proyecto contextualiza sus resultados dentro de un marco normativo que promueve el uso responsable de la información y sienta las bases para una futura implementación en entornos reales bajo condiciones de cumplimiento regulatorio.

Metodología

Método

El presente proyecto se desarrolla bajo un enfoque metodológico aplicado, sistemático y estructurado, fundamentado en la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), ampliamente reconocida en proyectos de analítica y gestión de datos por su flexibilidad y claridad en la organización de las actividades (Chapman et al., 2000; Acuña-Cid et al., 2025). Si bien CRISP-DM se originó en el ámbito de la minería de datos, su estructura resulta pertinente para proyectos de integración y diagnóstico de calidad de datos, dado que enfatiza la comprensión profunda de la información antes de cualquier proceso de transformación o explotación analítica.

En este proyecto, CRISP-DM se adapta al contexto de un proceso ETL orientado al diagnóstico de la calidad de datos empresariales provenientes de múltiples fuentes. Las fases de comprensión del negocio y comprensión de los datos permiten identificar las problemáticas asociadas a la fragmentación de información, tales como duplicidades, errores de digitación e inconsistencias estructurales, así como definir criterios formales de calidad basados en dimensiones como completitud, unicidad, consistencia y validez. Posteriormente, las fases de preparación de los datos y transformación se enfocan en la aplicación controlada de reglas de validación, estandarización e integración, siempre precedidas por un análisis detallado mediante técnicas de data profiling.

El flujo metodológico seguido a lo largo del proyecto, alineado con las fases de CRISP-DM y adaptado al diagnóstico de calidad de datos, se presenta en la Figura 1, donde se ilustra la secuencia lógica de actividades desde la comprensión del negocio hasta el despliegue de resultados.

Figura 1

Flujo Metodológico del Proyecto

El método se apoya en el uso de Python y la librería Pandas para la implementación del flujo ETL y el perfilamiento de los datos, dada su capacidad para manejar estructuras tabulares, calcular métricas y garantizar trazabilidad en los procesos analíticos (McKinney, 2022). Finalmente, la fase de despliegue se realiza a través de visualizaciones en Power BI, herramienta que facilita la comunicación de los resultados a nivel gerencial y refuerza el carácter aplicado del proyecto.

Tipo de Estudio

De acuerdo con sus objetivos y alcance, el proyecto se clasifica como un estudio aplicado, ya que busca resolver una problemática concreta de gestión de datos en un entorno empresarial real, aportando una solución técnica orientada al diagnóstico y mejora de la calidad de la información. Asimismo, corresponde a un estudio descriptivo, dado que su propósito principal es identificar, clasificar y cuantificar los errores presentes en las bases de datos, sin intervenir directamente en su corrección automática.

Desde el enfoque metodológico, el estudio es cuantitativo, pues se apoya en el cálculo de métricas objetivas como frecuencias y proporciones de valores nulos, duplicados e inconsistencias, las cuales permiten dimensionar el estado de la calidad de los datos. Adicionalmente, se trata de un estudio no experimental, ya que no se manipulan variables ni se introducen tratamientos controlados, sino que se analizan los datos tal como se encuentran en su estado inicial.

Esta clasificación es coherente con el carácter diagnóstico del proyecto, el cual no pretende desarrollar modelos predictivos ni aplicar técnicas de aprendizaje automático, sino generar un análisis estructurado que sirva de base para futuras decisiones de mejora y gobernanza de datos

Recolección de Datos

La recolección de datos se realiza a partir de 43 archivos en formato CSV, cada uno correspondiente a una empresa distinta gestionada por la organización objeto de estudio. Con el fin de preservar la trazabilidad por fuente y evitar la pérdida de contexto, los archivos son cargados de manera individual durante la fase de comprensión de los datos. Como control inicial de calidad, se verifica explícitamente que el número de archivos cargados coincida con el total esperado, asegurando la integridad del insumo analítico antes de continuar con el proceso.

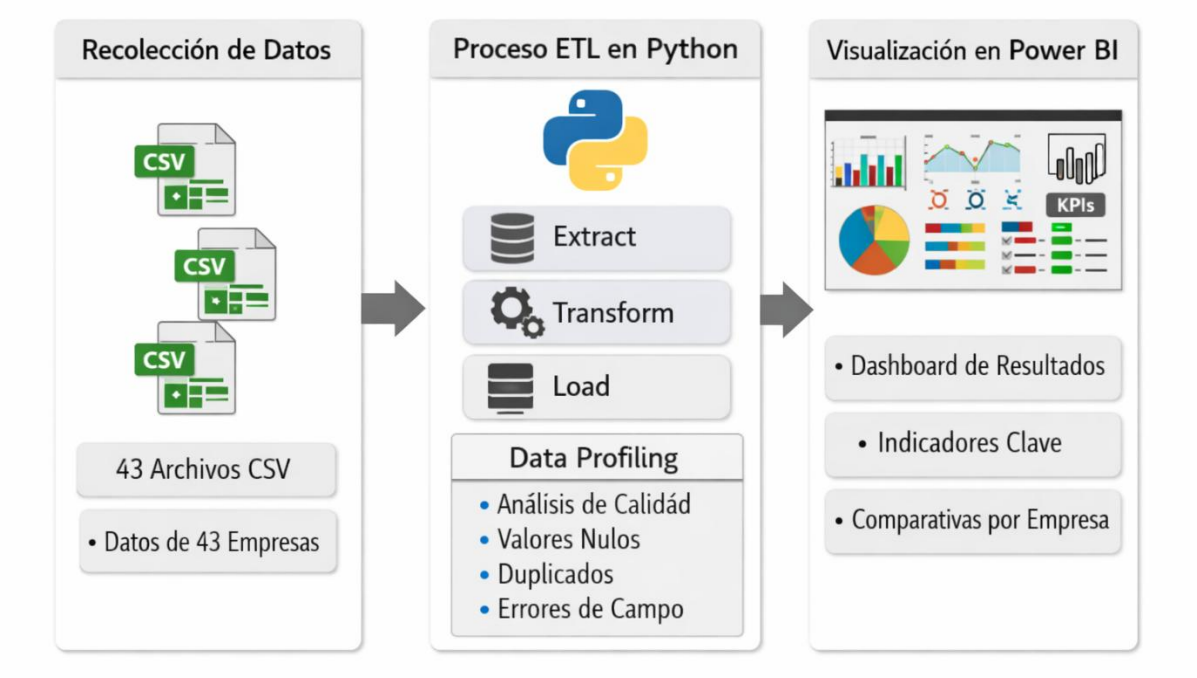
Posteriormente, se ejecuta un perfilamiento estructural de las bases de datos, orientado a validar la homogeneidad de los archivos en términos de número de columnas, nombres de los campos y cantidad de registros. Esta actividad permite confirmar la compatibilidad técnica de las bases para un proceso de integración posterior y descartar riesgos asociados a desalineaciones de esquema o inconsistencias estructurales.

Una vez validada la estructura, se desarrolla un perfilamiento de calidad por columna, en el cual se cuantifican valores nulos, vacíos y duplicados, tanto en términos absolutos como porcentuales. Este análisis permite dimensionar la magnitud de los problemas de calidad y establecer una línea base objetiva para la evaluación posterior. Adicionalmente, se aplican reglas cruzadas de consistencia lógica entre variables relacionadas, con el fin de detectar inconsistencias que no son evidentes mediante validaciones simples de formato o nulidad.

El esquema general del proceso ETL seguido desde la recolección de los archivos CSV hasta la preparación de la información para su visualización se presenta en la Figura 2, donde se resume la arquitectura técnica del proyecto. De igual manera, la implementación del proceso ETL y de las reglas de perfilamiento se realizó mediante scripts desarrollados en Python, los cuales se presentan en detalle en el Apéndice A.

Figura 2

Esquema del Proceso ETL y Visualización de Datos del Proyecto



Preparación de los Datos

Como parte del proceso de preparación y diagnóstico de los datos, se definieron reglas explícitas de perfilamiento, orientadas a identificar errores frecuentes en los registros. Estas reglas se establecen antes del análisis con el objetivo de garantizar criterios homogéneos para la evaluación de la calidad de la información en todas las bases de datos analizadas. Las reglas de perfilamiento aplicadas en el proceso ETL son las siguientes:

1. Regla NAME: Identificación de registros que presentan un número de identificación válido, pero cuyo campo de nombre o razón social contiene el valor “0”. Esta regla permite detectar errores de digitación o fallas en los procesos de captura que afectan la completitud y validez del dato.

2. Regla ID: Detección de identificaciones asociadas a más de un nombre o razón social dentro de una misma base de datos. Esta regla busca identificar inconsistencias lógicas que comprometen la unicidad y coherencia de los registros, y que pueden generar duplicidades o errores en procesos posteriores de análisis o consolidación.

3. Regla EMAIL: Identificación de registros que no cuentan con información válida en Email 1 y posee un registro en Email 2. Esta regla permite evaluar la completitud de los datos de contacto y detectar registros con información incompleta o insuficiente para fines operativos y comerciales.

Estas reglas se implementan mediante funciones de validación desarrolladas en Python utilizando la librería Pandas, y sus resultados se expresan en métricas cuantitativas por base de datos. Desde una perspectiva de negocio, la definición de reglas explícitas de perfilamiento resulta fundamental, ya que permite identificar de manera temprana errores que afectan directamente procesos operativos, comerciales y administrativos.

En particular, las inconsistencias entre identificaciones y nombres comprometen la unicidad de los registros, generando duplicidades que pueden derivar en reprocesos, errores de facturación o distorsiones en los análisis gerenciales. De igual forma, la ausencia o inconsistencia en los datos de contacto, como los correos electrónicos, limita la capacidad de comunicación con clientes y proveedores, afectando la eficiencia de los procesos comerciales y de servicio.

La literatura señala que la aplicación de reglas de validación orientadas a atributos críticos del negocio constituye una práctica esencial para la evaluación de la calidad de los datos, ya que permite reducir riesgos operativos, mejorar la confiabilidad de la información y apoyar la toma de decisiones basada en datos confiables (Batini & Scannapieco, 2016; Redman, 2013). Asimismo, estudios en gobernanza de datos destacan que la identificación temprana de errores mediante reglas de perfilamiento contribuye a minimizar costos asociados a correcciones posteriores y a fortalecer los procesos de control en el origen de la información (Alhassan & Sammon, 2020; DASCA, 2025).

El flujo de aplicación de estas reglas y la generación de indicadores de calidad se ilustra en la Figura 3, donde se presenta el proceso de perfilamiento desde la carga de los archivos hasta la obtención de métricas diagnósticas, evidenciando su papel como herramienta de diagnóstico para la gestión y mejora de la calidad de los datos empresariales.

Figura 3

Flujo de Perfilamiento de Datos en Python

Es importante resaltar que el objetivo de estas reglas no es la corrección automática de los errores, sino su identificación, clasificación y cuantificación, como insumo para la toma de decisiones gerenciales y la planificación de estrategias de mejora en la calidad de los datos.

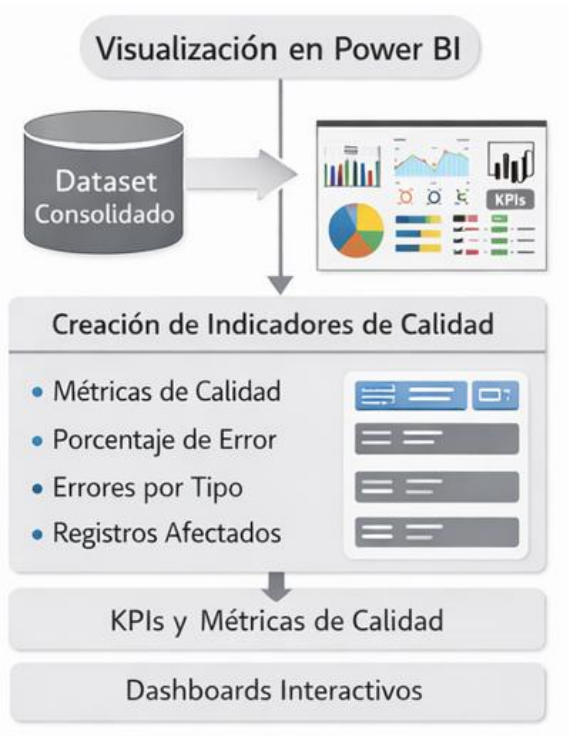
Los resultados del perfilamiento se consolidan en una tabla de métricas por base de datos, donde cada fila representa una empresa y cada columna un indicador de calidad. En esta etapa no se consolidan los registros, sino únicamente los indicadores diagnósticos, lo que permite comparar el estado de la calidad entre las distintas fuentes sin alterar los datos originales.

Finalmente, con base en el diagnóstico estructural y de calidad, se deja preparada la información para la consolidación final en un archivo maestro único. Esta consolidación se realiza de manera controlada y consciente de los errores identificados, respetando la secuencia metodológica de CRISP-DM y asegurando que el proceso de integración no oculte problemas de

calidad preexistentes. Los datos consolidados se preparan para su posterior visualización en Power BI, donde se presentan los indicadores de calidad como apoyo a la toma de decisiones gerenciales, tal como se muestra en la Figura 4.

Figura 4

Proceso de Visualización de Indicadores de Datos en Power Bi



Resultados

El proceso ETL implementado permitió integrar y analizar la información proveniente de 43 bases de datos empresariales en formato CSV, alcanzando un total de 1.137.456 registros evaluados. A partir del perfilamiento estructural aplicado en Python, se obtuvieron métricas que permiten dimensionar el estado actual de la calidad de los datos y comparar el comportamiento de los errores entre las distintas fuentes.

Resultados del Perfilamiento Estructural

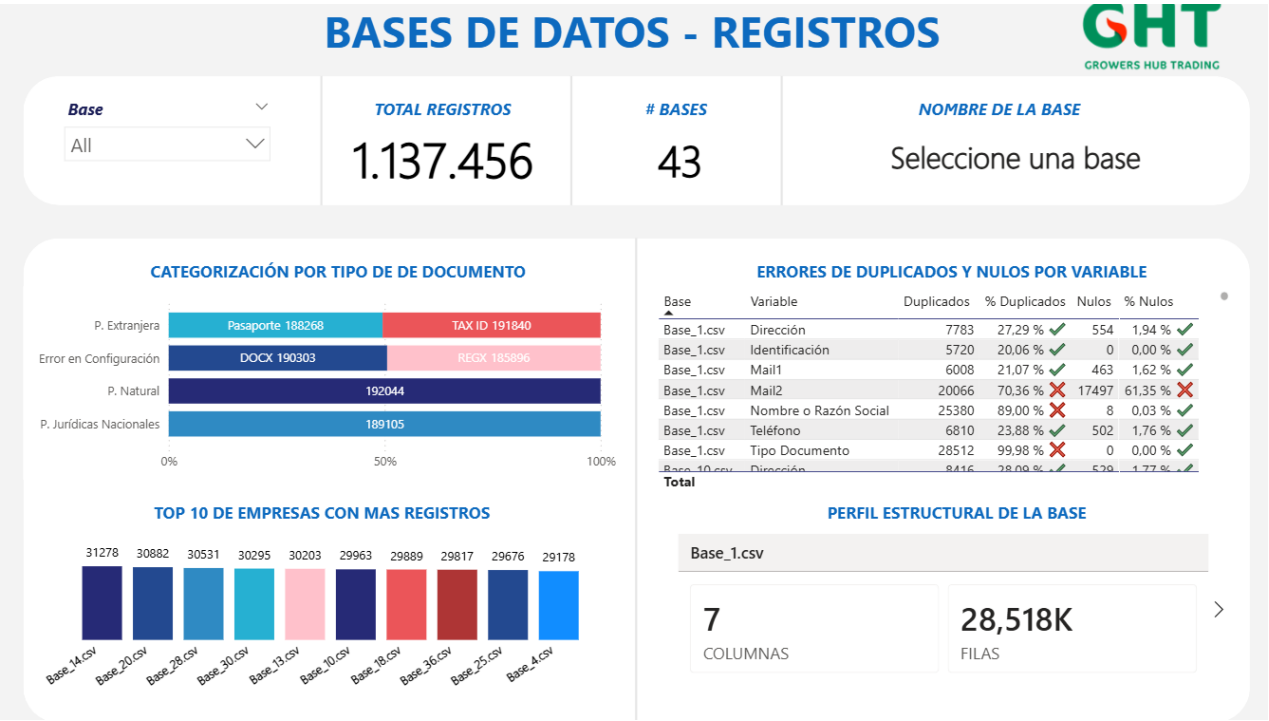
El análisis estructural confirmó que las 43 bases de datos presentan un esquema homogéneo, tanto en el número de columnas como en los nombres de los campos, lo que permitió su integración técnica sin introducir inconsistencias adicionales. Cada base contiene 7 columnas, lo cual facilitó la aplicación de las reglas de perfilamiento y la comparación directa de indicadores entre empresas.

Como indicador estructural relevante, se identificó que el número de registros por base presenta variaciones moderadas, con valores cercanos a los 28.000–31.000 registros por empresa, lo que permitió realizar análisis comparables sin sesgos extremos asociados al tamaño de las fuentes como se observa en la Figura 5. Entre esto se destacan los siguientes Indicadores estructurales:

- Número total de bases analizadas: 43
- Número total de registros evaluados: 1.137.456
- Número de columnas por base: 7

Figura 5

Perfilamiento Estructural de las 43 Bases de Datos



Resultados del Perfilamiento de Calidad por Reglas

La aplicación de las reglas de perfilamiento definidas en la metodología permitió identificar y cuantificar errores asociados a completitud, unicidad y consistencia lógica de los datos. A nivel global, se identificó que aproximadamente el 8,05 % de los registros presentan al menos un error, lo que evidencia deficiencias relevantes en los procesos de captura y control de la información.

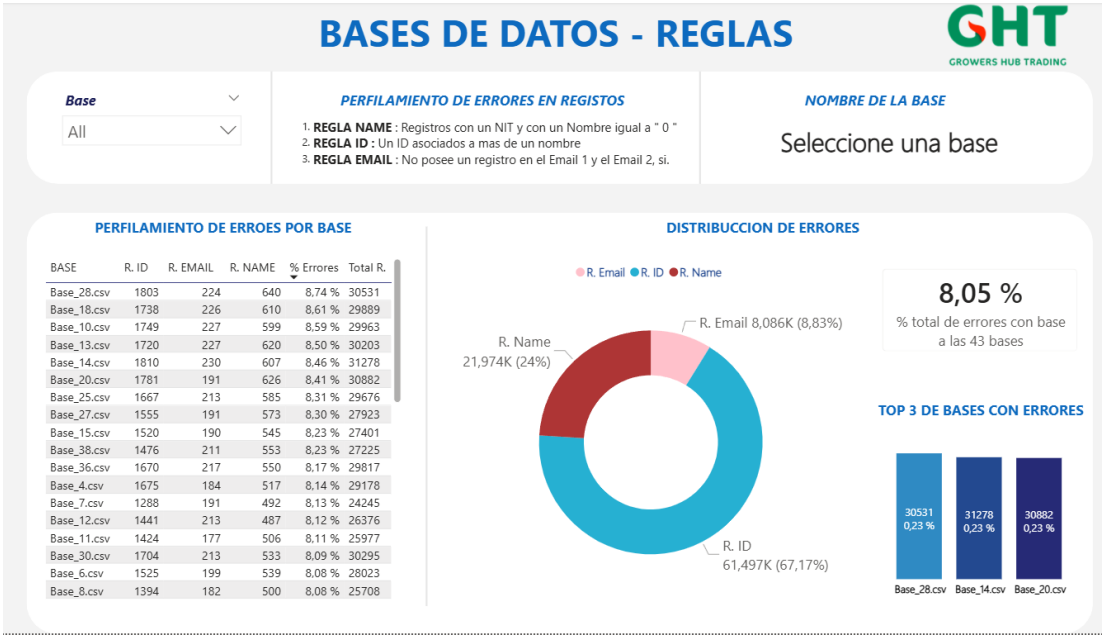
El análisis de los errores identificados a partir del proceso de perfilamiento evidencia que estos no se distribuyen de manera homogénea entre los distintos tipos de validación aplicados. En particular, las inconsistencias asociadas a la Regla ID concentran la mayor proporción de errores, representando el 61,77 % del total, lo que refleja una alta recurrencia de identificaciones

vinculadas a múltiples nombres o razones sociales y compromete directamente la unicidad y confiabilidad de los registros. En segundo lugar, los errores asociados a la Regla NAME corresponden al 24 %, relacionados con la presencia de valores no informativos o genéricos en el campo de nombre pese a contar con una identificación registrada. Finalmente, los errores vinculados a la Regla EMAIL representan el 8,83 %, evidenciando deficiencias en la completitud de la información de contacto. Esta distribución permite identificar los tipos de error con mayor impacto relativo en la calidad de los datos, los cuales se visualizan de forma comparativa en la Figura 6 y se resumen en los siguientes KPI's:

- Porcentaje global de registros con errores: 8,05 %
- Participación de errores por Regla ID: 61,77 %
- Participación de errores por Regla NAME: 24 %
- Participación de errores por Regla EMAIL: 8,83 %

Figura 6

Reglas de Validación



Resultados del Análisis de Impacto Operativo y Económico

A partir del número de errores identificados y considerando un tiempo promedio estimado de 2 minutos por corrección, una jornada laboral efectiva de 7 horas diarias y un costo diario de $67.000 COP con base al salario mínimo actual vigente, se estimó el impacto operativo asociado a la corrección manual de los errores detectados.

El análisis permitió estimar un tiempo total de inversión de 436 días laborales, lo que se traduce en una inversión económica aproximada de $29.211.043 COP para la corrección de los errores identificados en las 43 bases de datos de manera manual, tal como se evidencia en la Figura 7 y en los KPI's recopilados en la tabla 1.

Tabla 1

Indicadores Clave de Calidad de Datos

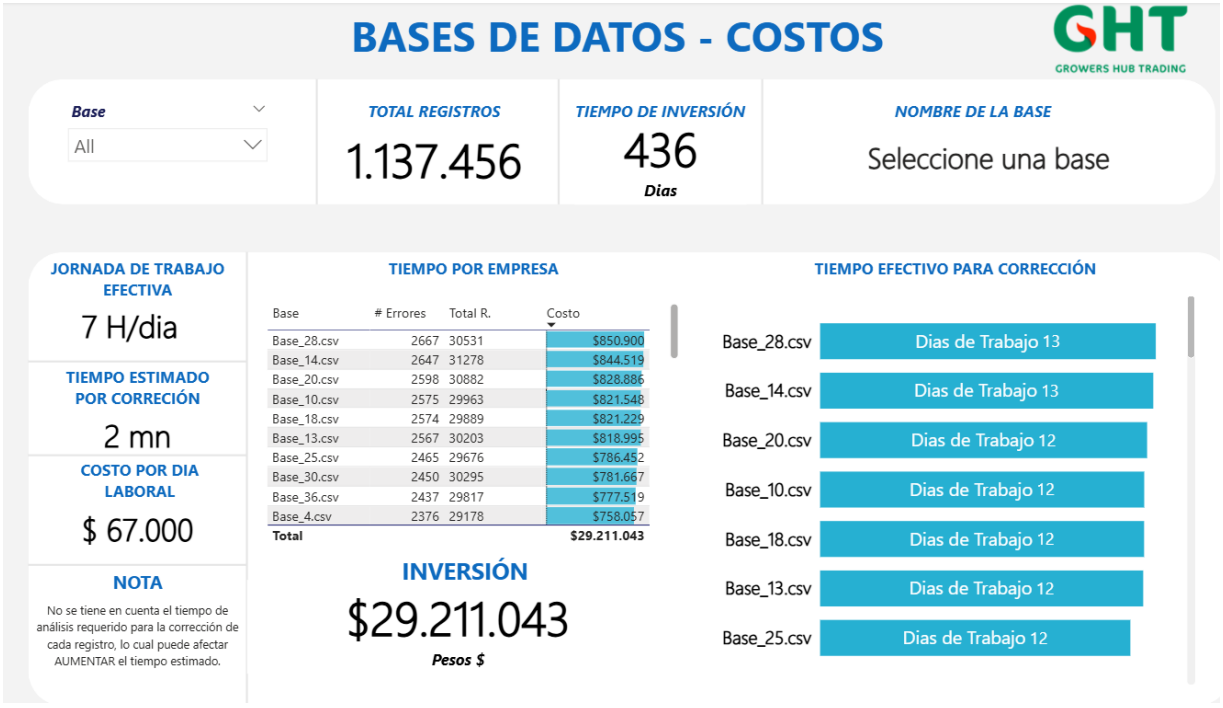
Categoría	KPI	Descripción	Valor
Estructural	Número de bases analizadas	Total de archivos CSV analizados	43
Estructural	Número total de registros	Registros evaluados en el proceso ETL	1.137.456
Estructural	Número de columnas por base	Campos analizados por cada fuente	7
Calidad global	Porcentaje total de registros con errores	Registros que presentan al menos un error	8,05 %
Calidad por regla	Participación errores Regla ID	Identificaciones asociadas a más de un nombre	61,77 %
Calidad por regla	Participación errores Regla NAME	Nombres o razones sociales no informativos	24,00 %

Categoría	KPI	Descripción	Valor
Calidad por regla	Participación errores Regla EMAIL	Registros sin correo electrónico válido	8,83 %
Impacto operativo	Tiempo estimado de corrección	Días laborales requeridos	436 días
Impacto económico	Costo total estimado de corrección	Inversión asociada a la corrección de errores	$29.211.043 COP

Nota. Indicadores (KPI's) establecidos para el análisis de las 43 bases de datos.

Figura 7

Inversión Económica para la Corrección de Bases de Datos



Este resultado cuantifica de manera objetiva el costo potencial de la mala calidad de los datos y proporciona un insumo clave para la toma de decisiones estratégicas relacionadas con la priorización de acciones de mejora. Cabe resaltar que gracias a la accesibilidad y dinámica de visualizaciones en Power Bi, se puede acceder de manera especifica a la empresa o base de datos de nuestro interés.

Conclusiones

El desarrollo del proceso ETL en Python permitió integrar de manera controlada la información proveniente de 43 fuentes de datos empresariales, garantizando la trazabilidad de los registros mediante la identificación de su base de origen. La validación estructural confirmó la homogeneidad del esquema de las bases, lo que facilitó la aplicación uniforme de las reglas de perfilamiento y aseguró la coherencia del análisis realizado.

El perfilamiento de los datos evidenció la presencia de problemas relevantes de calidad asociados principalmente a valores nulos, inconsistencias lógicas y errores de digitación. En particular, las identificaciones asociadas a múltiples nombres o razones sociales se establecieron como el principal tipo de error detectado, seguido por la presencia de nombres no informativos y la ausencia de información válida de contacto, lo que afecta directamente la confiabilidad de la información empresarial.

La cuantificación de los errores permitió dimensionar objetivamente el impacto de la mala calidad de los datos, cumpliendo con el objetivo de medir su frecuencia y proporción dentro del conjunto analizado. Este diagnóstico proporciona una base sólida para comprender el estado actual de la información y para estimar el esfuerzo requerido en procesos de limpieza y normalización.

La aplicación de reglas de consistencia lógica complementó el análisis tradicional por columna, permitiendo identificar inconsistencias que no serían detectables mediante validaciones simples. Este enfoque aportó una visión más profunda del estado de los datos y permitió diferenciar errores técnicos de fallas en los procesos operativos de captura.

Finalmente, la integración de las métricas de calidad en un dashboard desarrollado en Power BI facilitó la visualización clara de los resultados del proceso ETL, convirtiéndose en una

herramienta efectiva para apoyar la toma de decisiones orientadas a la mejora de la calidad de la información y al fortalecimiento de la gobernanza de datos

Recomendaciones

Se recomienda fortalecer los mecanismos de validación en el origen de los datos, especialmente en campos como identificación, nombre y correos electrónicos, con el fin de reducir la generación de errores de digitación, inconsistencias lógicas y registros incompletos.

Es aconsejable priorizar la atención de los errores relacionados con identificaciones inconsistentes, dado su impacto directo sobre la unicidad y confiabilidad de los registros, antes de abordar procesos generales de limpieza de datos.

Se sugiere utilizar el dashboard desarrollado en Power BI como una herramienta de seguimiento continuo de la calidad de la información, permitiendo monitorear periódicamente los indicadores de calidad y evaluar el efecto de las acciones de mejora implementadas.

Asimismo, se recomienda definir un plan progresivo de mejora de la calidad de los datos, iniciando con reglas de fácil implementación, y avanzando posteriormente hacia estrategias más complejas de normalización solo cuando el diagnóstico lo justifique.

Finalmente, se recomienda documentar formalmente el proceso ETL y las reglas de perfilamiento aplicadas, con el fin de facilitar su reutilización, mantenimiento y posible ampliación a nuevas fuentes de información dentro de la organización.

Referencias Bibliográficas

Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data: A survey. *The VLDB Journal, 24*(4), 557–581.https://doi.org/10.1007/s00778-015-0389-y

Acuña-Cid, H. A., Paredes, J., & Carrillo, C. (2025). Integration of the CRISP-DM model with network analysis. *Journal of Data Science Applications, 7*(3), 101. MDPI.https://www.mdpi.com/2504-4990/7/3/101

Airbyte. (2025, septiembre 2). *ETL data quality testing: Tips for cleaner pipelines*. Airbyte.https://airbyte.com/data-engineering-resources/etl-data-quality

Alhassan, I., & Sammon, D. (2020). Data governance and data quality: The role of data protection regulations in corporate environments. *Journal of Information Systems, 34*(3), 67–81.https://doi.org/10.2308/isys-18-053

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

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Colombia. (2012). *Ley 1581 de 2012: Por la cual se dictan disposiciones generales para la protección de datos personales*. Diario Oficial No. 48.587. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

DASCA. (2025, octubre 3). *Data quality in ETL: Essential checks and implementation strategies*. Data Science Council of America.https://www.dasca.org/world-of-data-science/article/data-quality-in-etl-essential-checks-and-implementation-strategies

EM360Tech. (2025, julio 31). *What is ETL testing, and how does it impact data quality?* EM360Tech.https://em360tech.com/tech-articles/etl-testing-data-quality

International Organization for Standardization. (2008). *ISO/IEC 25012:2008 Software engineering — Software product quality requirements and evaluation (SQuaRE) — Data quality model*. ISO.

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

McKinney, W. (2022). *Python for data analysis* (3rd ed.). O'Reilly Media.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Bulletin of the Technical Committee on Data Engineering, 23*(4), 3–13.

Redman, T. C. (2013). *Data driven: Profiting from your most important business asset*. Harvard Business Review Press.

von Enzberg, S., Hohmann, L., & Wischnewski, C. (2024). On the current state of industrial data science. *Procedia CIRP, 126*, 456463.https://www.sciencedirect.com/science/article/pii/S2212827124014239

Zhang, Y., He, Y., & Liu, G. (2018). Sampling for big data profiling: A survey. *IEEE Access, 6*, 45765–45778.https://doi.org/10.1109/ACCESS.2018.2865804

Apéndices

Apéndice A

Código del Proceso ETL en Python y Reglas de Perfilamiento

```
#Librerias
import os # Acceso al sistema operativo
import pandas as pd # Manipular tablas y series
from time import time # Permite trabajar con funciones de tiempo
import re # Regula expresiones de texto
import openpyxl # Manipular archivos en Excel
import xlsxwriter # Creación de archivos en Excel
```

1. Definir ruta y cargar las 43 bases

```
# Cargue de las 43 bases de datos

# Ruta donde están los CSV
RUTA_DATOS = r"C:\Users\jeck1\OneDrive\Documentos\JEISON DOCUMENTOS\UNAD CIENCIA DE DATOS\TERCER SEMESTRE\PROYECTO DE GRADO 2\Proyecto Escrito\Proyecto aplicado\Base de datos"

# Listas contenedoras
dfs = []               # DataFrames individuales
nombres_bases = []     # Nombre de cada archivo

# Carga de archivos
for archivo in os.listdir(RUTA_DATOS):
    if archivo.lower().endswith(".csv"):
        df = pd.read_csv(os.path.join(RUTA_DATOS, archivo))
        dfs.append(df)
        nombres_bases.append(archivo)

# Verificación crítica
total_bases = len(dfs)
print(f"Bases cargadas: {total_bases}")
assert total_bases == 43, "No se cargaron las 43 bases"
```

```
Bases cargadas: 43
```

2. Perfilamiento estructural (verificación de homogeneidad)

```
perfil_estructural = []

for df, nombre_base in zip(dfs, nombres_bases):
    perfil_estructural.append({
        "Base": nombre_base,
        "Numero_columnas": df.shape[1], #Numero de columnas
        "Numero_registros": df.shape[0], # Numero de filas
    })

df_perfil_estructural = pd.DataFrame(perfil_estructural)

df_perfil_estructural.head()
```

	Base	Numero_columnas	Numero_registros
0	Base_1.csv	7	28518
1	Base_10.csv	7	29963
2	Base_11.csv	7	25977
3	Base_12.csv	7	26376
4	Base_13.csv	7	30203

3. Perfilamiento de calidad: nulos y duplicados por columna

```
perfil_calidad = []

for df, nombre_base in zip(dfs, nombres_bases):

    total_registros = len(df)

    for columna in df.columns:

        nulos = (
            df[columna].isna().sum()
            + (df[columna].astype(str).str.strip() == "").sum()
        )

        duplicados = df[columna].duplicated().sum()

        perfil_calidad.append({
            "Base": nombre_base,
            "Columna": columna,
            "Total_registros": total_registros,
            "Valores_nulos": nulos,
            "Duplicados": duplicados,
            "Pct_nulos": round((nulos / total_registros) * 100, 2),
            "Pct_duplicados": round((duplicados / total_registros) * 100, 2)
        })

df_calidad_por_columna = pd.DataFrame(perfil_calidad)

df_calidad_por_columna
```

	Base	Columna	Total_registros	Valores_nulos	Duplicados	Pct_nulos	Pct_duplicados
0	Base_1.csv	Nombre o Razón Social	28518	8	25380	0.03	89.00
1	Base_1.csv	Tipo Documento	28518	0	28512	0.00	99.98
2	Base_1.csv	Identificación	28518	0	5720	0.00	20.06
3	Base_1.csv	Mail1	28518	463	6008	1.62	21.07
4	Base_1.csv	Mail2	28518	17497	20066	61.35	70.36
...	...	...	...	...	...	...	...

4. Reglas de perfilamiento

```
# Nombre = "0" pero identificación existe

regla_nombre_cero = []

for df, nombre_base in zip(dfs, nombres_bases):
    mask = (
        df["Nombre o Razón Social"].astype(str).str.strip() == "0"
    ) & (
        df["Identificación"].notna()
    )

    regla_nombre_cero.append({
        "Base": nombre_base,
        "Registros_afectados": mask.sum()
    })


df_regla_nombre_cero = pd.DataFrame(regla_nombre_cero)

df_regla_nombre_cero.head()
```

	Base	Registros_afectados
0	Base_1.csv	536
1	Base_10.csv	599
2	Base_11.csv	506
3	Base_12.csv	487
4	Base_13.csv	620

```python
# Identificación duplicada con nombres distintos

regla_id_conflicto = []

for df, nombre_base in zip(dfs, nombres_bases):

    conflicto = (
        df.groupby("Identificación")["Nombre o Razón Social"]
        .nunique()
        .reset_index()
    )

    conflicto = conflicto[conflicto["Nombre o Razón Social"] > 1]

    regla_id_conflicto.append({
        "Base": nombre_base,
        "Identificaciones_en_conflicto": len(conflicto)
    })

df_regla_id_conflicto = pd.DataFrame(regla_id_conflicto)

df_regla_id_conflicto.head()
```

	Base	Identificaciones_en_conflicto
0	Base_1.csv	1541
1	Base_10.csv	1749
2	Base_11.csv	1424
3	Base_12.csv	1441
4	Base_13.csv	1720

```python
# Mail1 vacío pero Mail2 informado

regla_mail = []

for df, nombre_base in zip(dfs, nombres_bases):
    mask = (
        (df["Mail1"].isna() | (df["Mail1"].astype(str).str.strip() == ""))
        &
        (df["Mail2"].notna())
    )

    regla_mail.append({
        "Base": nombre_base,
        "Registros_afectados": mask.sum()
    })

df_regla_mail = pd.DataFrame(regla_mail)

df_regla_mail.head()
```

	Base	Registros_afectados
0	Base_1.csv	191
1	Base_10.csv	227
2	Base_11.csv	177
3	Base_12.csv	213
4	Base_13.csv	227