ETL
ENTREGA PROYECTO 2

Maestría en Ciencia de Datos e Inteligencia Artificial

Nelcy Lucia Zapata Gil– 22502267
Valentina Isaza Ospina - 22502266

Resumen— El Fondo Mixto de Etnocultura y Desarrollo Social - FONPACÍFICO enfrenta dificultades en su gestión financiera debido a la ausencia de un flujo de caja proyectado, lo que impide tomar decisiones informadas sobre la disponibilidad de recursos para su sostenibilidad y la cofinanciación de proyectos. La dispersión de información financiera en múltiples fuentes y la dependencia de procesos manuales agravan esta situación, aumentando el riesgo de errores y limitando la generación de reportes fiables en tiempo real.

Este proyecto propone diseñar, preparar y construir flujos de datos eficientes mediante procesos de Extracción, Transformación y Carga (ETL) para integrar diversas fuentes de datos. Que permita a FONPACIFICO desarrollar un modelo predictivo para el análisis avanzado de la información y optimice su gestión financiera, mejorando la toma de decisiones estratégicas y operativas. La implementación de procesos ETL en la entidad, facilitará la consolidación de datos financieros, mejorará la precisión de los reportes y aumentará la capacidad de análisis en tiempo real, utilizando herramientas como Python, Pandas y MySQL, y establecerá las bases para el uso futuro de modelos predictivos.

1. DESCRIPCIÓN DEL PROBLEMA

FONPACÍFICO es una entidad pública sin ánimo de lucro que gerencia proyectos de interés público a nivel nacional. En la actualidad, enfrenta un desafío crítico en su gestión financiera debido a la falta de un flujo de caja proyectado, lo que genera incertidumbre sobre la sostenibilidad y disponibilidad de recursos para la ejecución de proyectos. Esta situación impacta la toma de decisiones estratégicas y operativas, y la capacidad de la entidad para anticiparse a periodos de déficit o superávit de liquidez y gestionar de manera eficiente los recursos.

La entidad opera en un entorno altamente regulado, donde los ciclos de contratación estatal, las restricciones normativas y los cambios en la administración pública influyen directamente en el flujo financiero. La dinámica presupuestaria gubernamental, marcada por la apertura y cierre de vigencias fiscales, así como las limitaciones impuestas en períodos electorales (Ley de Garantías), o las alteraciones (provocados por factores externos como inestabilidad social, condiciones climáticas adversas o emergencias nacionales) de cronograma de flujo de los proyectos financiados por el Sistema General de Regalías (SGR), compromete la estabilidad financiera de la entidad e impactan la capacidad de planificación y operación administrativa, así como la ejecución de los proyectos a su cargo.

Desde una perspectiva técnica, la ausencia de un sistema automatizado de Extracción, Transformación y Carga de datos (ETL) dificulta la consolidación de información financiera dispersa en diferentes fuentes. La dependencia de procesos manuales aumenta la probabilidad de errores y dificulta la generación de reportes confiables en tiempo real, lo que limita la capacidad de análisis y toma de decisiones fundamentadas.

El desarrollo de un modelo de flujo de caja proyectado, mediante procesos de Extracción, Transformación y Carga (ETL) para integrar diversas fuentes de datos, permitiría a FONPACIFICO desarrollar un modelo predictivo para el análisis avanzado de la información y optimice su gestión financiera, mejorando la toma de decisiones estratégicas y operativas. La implementación de procesos ETL en la entidad, facilitará la consolidación de datos financieros, mejorará la precisión de los reportes y aumentará la capacidad de análisis en tiempo real, utilizando herramientas como Python, Pandas y MySQL, y establecerá las bases para el uso futuro de modelos predictivos.

1.1. Contribución al Desarrollo de Habilidades en ETL y Ciencia de Datos

La implementación de procesos ETL en la entidad, facilitará la consolidación de datos financieros, mejorará la precisión de los reportes y aumentará la capacidad de análisis en tiempo real, utilizando herramientas como Python, Pandas y MySQL, permitirá aplicar conocimientos clave en ETL y Ciencia de Datos para la estructuración y análisis de datos financieros. Además, servirá como base para futuras implementaciones de modelos de Machine Learning para predicción financiera, generando un impacto significativo en la eficiencia operativa de la organización.

Este proyecto no solo aborda un problema de gestión financiera real, sino que también representa una aplicación práctica de la Ciencia de Datos en la optimización de procesos organizacionales, fortaleciendo la capacidad analítica y de predicción financiera de FONPACÍFICO.

2. JUSTIFICACIÓN DEL PROYECTO

La implementación de procesos de Extracción, Transformación y Carga (ETL) permitirá a FONPACÍFICO optimizar la gestión financiera y mejorar la toma de decisiones. Este modelo proporcionará una visión consolidada y automatizada de los ingresos y egresos, alineándose con la planificación comercial y las restricciones estatales.

Desde la perspectiva de la Ciencia de Datos, este proyecto permite aplicar técnicas de ETL para estructurar la información financiera dispersa en diversas fuentes, reduciendo errores manuales y mejorando la eficiencia operativa. Además, facilitará el análisis de datos para anticipar posibles riesgos financieros.

Los beneficios clave incluyen:

a. Mejor previsión financiera, permitiendo la identificación temprana de periodos críticos.
b. Mayor transparencia y cumplimiento normativo.
c. Automatización del análisis financiero, integrando datos heterogéneos en un sistema centralizado.
d. Optimización de la planeación estratégica, comercial, y operativa considerando los ciclos de contratación estatal.

3. IDENTIFICACIÓN DE FUENTES

Para el proceso de Extracción, Transformación y Carga (ETL) de datos financieros, se identificarán y utilizarán diversas fuentes de datos, clasificadas según su tipo, formato y volumen:

3.1. Fuentes Internas

a. *Presupuesto 20204 y 2025:* Archivo Excel (.xls) con proyecciones de ingresos y egresos.
b. *Sistema Financiero y Contable Soul SIFP:* Documento Word (.docx) Políticas y criterios contables.
c. *Historial de Contratos*: Base de datos interna con registros de contratación pública y privada.
d. *Estados Financieros*: Documentos financieros en formato Excel y PDF (.xlsx, .pdf).
e. *Matriz de información Consolidada:* Base de datos interna con registros de contratación pública y privada formato Excel y PDF (.xlsx, .pdf).

3.2. Fuentes Externas

a. *Sistema de Contratación Estatal (SECOPI y II):* Datos abiertos (.csv, API) sobre contratos y licitaciones.
b. *GESPROY del DNP*: Plataforma tecnológica que permite a las entidades ejecutoras gestionar los proyectos financiados con recursos del Sistema General de Regalías (SGR) (.csv, .xlsx).
c. *SPGR del DNP*: el Sistema de Presupuesto y Giro de Regalías, es un sistema de información que gestiona el presupuesto y el giro de regalías del SGR (Sistema General de Regalías) administrado por el Ministerio de Hacienda y Crédito Público. (.csv, .xlsx).
d. *Normatividad Estatal:* Documentación en PDF (.pdf) sobre Ley de Garantías y restricciones presupuestales.
e. *Calendario Electoral y de Vigencias*: Datos estructurados en formato CSV (.csv) con fechas clave del ciclo gubernamental.

a. Tipo de Dato

Fuente: Excel (Drive)
Tipo de Dato: Semi- Estructurado
Información: Columnas con valores numéricos, fechas y textos sobre contratos y pagos.

Fuente: MySQL
Tipo de Dato: Estructurado
Información: Tablas con registros normalizados de contratos, pagos y estados financieros.

b. Formato del Dato

Fuente: Excel (Drive)
Formato: .xlsx
Detalles: Archivo de hoja de cálculo con varias hojas.

Fuente: MySQL
Formato: Tabla de MySQL
Detalles: Base de datos relacional con múltiples tablas organizadas.

c. Volumen de los Datos

Fuente: Excel (Drive)
Volumen Aproximado: Medio (MB - GB)
Frecuencia de Actualización: Actualización manual mensual o trimestral.

Fuente: MySQL
Volumen Aproximado: Medio (MB - GB)
Frecuencia de Actualización: Actualización frecuente, dependiendo de la operación diaria.

Excel: 1,12 MB

MySQL:

Base de datos	Tamaño (MB)	Tamaño (GB)
viaticos	0.27	0.00
soul_red_2024	26.58	0.03
soul_fon_2023	66.24	0.06
soul_fon_2022	86.39	0.08
soul_fon_2025	79.86	0.08
soul_fon_2024	98.11	0.10

4. PROBLEMA

La falta de un sistema ETL para integrar y analizar la información financiera impide a FONPACÍFICO tener una visión clara y oportuna de su situación financiera, lo que afecta su capacidad para planificar y ejecutar sus proyectos de manera eficiente.

5. EXPLORACIÓN INICIAL DE DATOS

5.1. DATOS INTERNOS

5.1.1. Archivo "Matriz de Gestión de Proyectos 04-03-2025"

a. Estructura del Archivo

- Contiene 243 columnas y varias filas con datos nulos.
- Muchas columnas están etiquetadas como "Unnamed", lo que indica un problema de encabezados o estructura del archivo. Las primeras filas contienen categorías como:
  Estado Principal del Proyecto: "Sin Contratar", "Terminado", en ejecución, suspendido, etc.
  Entidad Auditora y datos administrativos de seguimiento.
  Información Financiera, incluyendo compromisos, órdenes de pago, deducciones y comisiones.

b. Problemas identificados

- **Datos dispersos**: Se observan muchas columnas con nombres genéricos como "Unnamed", lo que puede deberse a una mala estructuración del archivo.
- **Valores faltantes**: Existen múltiples valores vacíos, especialmente en campos de auditoría y seguimiento financiero.
- **Formato irregular**: Algunas columnas parecen ser una combinación de etiquetas y valores, lo que requiere limpieza antes de analizarlas.

c. Acciones de Transformación de Datos

Para mejorar la usabilidad de este archivo en el proceso **ETL**:

- **Renombrar columnas** para mejorar la comprensión de los datos.
- **Eliminar columnas irrelevantes** o sin datos significativos.
- **Estandarizar estados de proyectos** para facilitar su clasificación.
- **Transformar la estructura del archivo** en un formato más relacional para facilitar análisis en una base de datos.

5.1.2. Archivo "SOUL. PROYECTOS 04-03-2025.xlsx"
1. Estructura del Archivo

- Contiene 14 columnas y 398 filas, con información sobre proyectos financiados dentro del sistema SOUL.
- Algunas variables contienen datos en formato numérico (valores financieros) y otras en texto (códigos de proyectos y programas).
- Existen valores faltantes en varias columnas, especialmente en nombres de proyectos y sectores.

a. Principales Variables en el Archivo

Identificación de Proyectos:

- Código del Proyecto (codigo).
- Nombre del Proyecto (nombre), aunque muchos valores están vacíos.
- Sector y Nombre del Sector (sector, nombre_sec).
- Código BPIN (cbnp), vinculado a la planeación de inversión pública.

Información Financiera:

- Valor del Proyecto (valor), indicando los montos asignados a cada iniciativa.
- Contratos Asociados (contratos), aunque casi todos están en 0.
- Disponibilidades, Compromisos, Obligaciones, IVA, Deducciones y Pagos, con información detallada sobre la ejecución presupuestal.

b. Problemas Identificados

Valores Faltantes:

- Muchas filas tienen **nombres de proyectos vacíos** (nombre).
- Algunas celdas de **sector y nombre del sector** están sin información.

Valores Numéricos con Cero:

- La mayoría de las columnas financieras como **Contratos, Disponibilidades, Compromisos, Obligaciones y Pagos** están en **cero**, lo que indica que los datos pueden no estar actualizados o representar proyectos inactivos.

Formato de Datos:

- La columna **Código BPIN (cbnp)** tiene registros mezclados entre números y texto (BPIN2023000060014 vs 2021000030042), lo que indica una posible inconsistencia en los formatos.
- Algunas columnas que deberían ser categóricas (nombre_sec) tienen valores faltantes.

c. Acciones de Transformación de Datos

Para mejorar la usabilidad de este archivo en el proceso **ETL**

- **Llenar valores faltantes** en nombre y nombre_sec con información de referencia si es posible.
- **Revisar los valores en cero**, para determinar si corresponden a datos no actualizados o proyectos cancelados.
- **Estandarizar el código BPIN**, asegurando que todos tengan el mismo formato.
- **Transformar columnas categóricas**, asignando códigos únicos donde falten valores.
- **Generar una variable de estado del proyecto**, según si tiene asignación presupuestal activa o está en cero.

5.2. DATOS EXTERNOS

5.2.1. Archivo "SECOP 1. CONSULTA PRINCIPALES"

a. Estructura del Archivo

- Contiene 75 columnas y 114 filas, con una gran cantidad de datos administrativos y financieros sobre contratación estatal.
- Se enfoca en procesos de compra pública registrados en el SECOP I (Sistema Electrónico de Contratación Pública).
- La información está organizada por entidades contratantes, modalidades de contratación, montos y fechas clave.

b. Principales Variables en el Archivo

- **Identificación de la Contratación**:
  UID: Código único de cada proceso de compra.
  Año de Firma del Contrato y Año de Cargue en el SECOP.
  Nivel de la Entidad: Territorial o Nacional.
  Nombre de la Entidad y su NIT.
- **Modalidad y Tipo de Contratación**:
  Código y Descripción de Modalidad de Contratación: Ej., Contratación Directa, Régimen Especial.
  Orden de la Entidad: Centralizado o descentralizado.
- **Ubicación Geográfica**:
  Departamento y Municipio de la Entidad que ejecuta la contratación.
- **Estado y Cumplimiento Normativo**:
  Fecha de Última Actualización.
  Fecha de Liquidación del Contrato (NaN en muchos casos).
  Cumplimiento del Decreto 248 (indica si cumple requisitos de contratación pública).
- **Otros Indicadores**:
  Si la empresa es una MiPyme.
  Si cumple con sentencias de contratación pública.

c. Problemas identificados

- **Valores Faltantes**:

La columna "Fecha de Liquidación" tiene muchos valores nulos, lo que puede indicar que los contratos aún están activos o que faltan registros.
Variables como "Cumple Decreto 248" y "Es MiPyme" tienen valores genéricos como "No definido", lo que sugiere que estos datos no están siendo bien reportados.

- **Formato y Normalización**:

Existen columnas con nombres en mayúsculas y caracteres especiales, lo que puede dificultar su procesamiento.
Algunas variables están en formato texto cuando deberían ser numéricas o categóricas.

- **Fechas No Estructuradas Correctamente**:

"Fecha de Última Actualización" está bien estructurada, pero "Fecha de Liquidación" tiene muchos valores en blanco o sin definir.

- **Consistencia en Modalidades de Contratación**:

Algunos registros de "Modalidad de Contratación" pueden ser normalizados, ya que contienen diferentes formas de escritura para el mismo tipo de contratación.

d. Acciones de Transformación de Datos

Para mejorar la usabilidad de este archivo en el proceso **ETL**

- **Estandarizar nombres de columnas** eliminando caracteres especiales y mayúsculas inconsistentes.
- **Manejar valores faltantes**, imputando fechas faltantes o marcándolas como contratos en proceso.
- **Convertir datos categóricos en variables codificadas**, especialmente en modalidades de contratación.
- **Normalizar la información financiera**, verificando si existen inconsistencias en los datos de contratación.
- **Crear variables derivadas**, como "Tiempo de Ejecución del Contrato", basado en la diferencia entre firma y liquidación.

5.2.2. Archivo "SECOP 2. CONSULTA PRINCIPALES"

a. Estructura del Archivo

- Contiene 59 columnas y 103 filas con información detallada sobre procesos de contratación pública en SECOP II.
- Se enfoca en datos de licitaciones, adjudicaciones y contratación estatal, con un alto nivel de detalle.
- Algunas columnas contienen valores repetidos, lo que puede indicar datos duplicados o contratos con múltiples registros.

b. Principales Variables en el Archivo

- **Identificación de la Entidad y el Contrato**:
  Entidad Contratante (Entidad, Nit Entidad).
  Departamento y Ciudad de la Entidad.
  ID del Proceso (ID del Proceso): Código único del proceso de compra.
  Referencia del Proceso: Nombre del contrato.
- **Detalles de Contratación**:
  Estado del Proceso (Estado de Apertura del Proceso), puede ser "Abierto", "Adjudicado", etc.
  Tipo de Contrato y **Subtipo de Contrato**.
  Código de Categoría de Compra.
  Nombre del Proveedor Adjudicado y **NIT del Proveedor**.
- **Información de Publicación y Adjudicación**:
  URL del Proceso (URLProceso), lo que permite acceder a detalles en SECOP II.
  Código de la Entidad y estado de adjudicación (Estado Resumen).

c. Problemas identificados

- **Valores Repetidos**:
  Existen registros duplicados de adjudicación para el mismo contrato (ID del Proceso y Referencia del Proceso aparecen varias veces para el mismo proveedor).
  Puede ser necesario **de duplicar los datos**, dejando solo la adjudicación final.
- **Valores Faltantes**:
  Algunas columnas como **"Subtipo de Contrato"** y **"Categorías Adicionales"** tienen valores en **"No definido"**, lo que sugiere datos incompletos en el sistema de contratación.
- **Formato de Datos**:
  "Código Entidad" está separado por comas, lo que puede requerir limpieza y conversión a formato numérico.
  "URLProceso" es un enlace funcional, pero podría ser estructurado para consulta más eficiente.
- **Datos No Normalizados**:
  Nombres de proveedores pueden estar en diferentes formatos (mayúsculas/minúsculas).
  Estado del proceso tiene diferentes categorizaciones que podrían ser unificadas.

d. Acciones de Transformación de Datos

Para mejorar la usabilidad de este archivo en el proceso **ETL**

- **Eliminar duplicados**, dejando solo la adjudicación final de cada contrato.
- **Completar valores faltantes**, especialmente en tipos de contrato y categorías de compra.
- **Convertir "Código Entidad" a formato numérico** eliminando comas.
- **Estandarizar los nombres de proveedores y entidades**, asegurando consistencia en el análisis.
- **Generar un indicador de estado de contratación**, para clasificar contratos en "Adjudicado", "En proceso" o "Cancelado".

5.2.3. Archivo "GESPROY. CG-CG-CG-proy_04_marzo_2025_09-38-10.xlsx.xlsx"

a. Estructura del Archivo

- Contiene 22 columnas y 168 filas.
- La mayoría de las columnas tienen nombres genéricos como "Unnamed", lo que sugiere un problema con los encabezados o la estructura de los datos.
- Incluye información sobre proyectos financiados a través del Sistema General de Regalías (SGR).

b. Principales Variables en el Archivo

- **Ubicación Geográfica**: Departamento, región y entidad territorial que gestiona cada proyecto.
- **Datos del Proyecto**:
  BPIN: Código del proyecto en el Banco de Programas y Proyectos de Inversión Nacional.
  Nombre del Proyecto.
  Fecha de Acuerdo: Fecha en que se aprobó el proyecto.
  Estado del Proyecto: Ej., "Cerrado", "Contratado en Ejecución", "Terminado".
- **Financiamiento**:
  Valor SGR: Recursos asignados por el Sistema General de Regalías.
  Valor Nación y Otros: Recursos adicionales.
  Valor Total del Proyecto.
- **Ejecución**:
  Avance Físico (%): Grado de avance de la obra o ejecución del proyecto.
  Avance Financiero (%): Relación entre el dinero ejecutado y el presupuesto total.
  Valor Pagos: Dinero ya desembolsado en el proyecto.

c. Problemas identificados

- **Encabezados mal estructurados**: Muchas columnas tienen espacios o caracteres especiales (\t).
- **Valores faltantes en algunas variables**:
  Algunas celdas de "Subsector" están vacías.
  En algunas filas, la columna de "Valor Otros" está en blanco.
- **Valores inconsistentes en Avance Físico y Financiero**:
  Algunos proyectos muestran **0% de avance físico**, pero con desembolsos financieros significativos.
  Se requiere validar si estos valores reflejan correctamente la ejecución de cada proyecto.
- **Conversión de Formatos**:
  Fechas almacenadas como texto deben convertirse a formato YYYY-MM-DD.
  Valores numéricos como **"Valor Total del Proyecto"** tienen decimales que deben revisarse.

d. Acciones de Transformación de Datos

Para mejorar la usabilidad de este archivo en el proceso **ETL**

- **Estandarizar nombres de columnas** eliminando espacios y caracteres no deseados.
- **Rellenar valores faltantes** con estrategias de interpolación o categorización.
- **Validar la consistencia de avances físicos y financieros**, detectando posibles errores.
- **Convertir datos financieros a formato numérico adecuado** para cálculos precisos.
- **Optimizar la base de datos**, eliminando columnas irrelevantes o redundantes.

5.2.4. Archivo "SPGR. Exportar CRONOGRAMA DE FLUJOS 04-03-2025.xlsx"

a. Estructura del Archivo

- Contiene 38 columnas y 79 filas, con datos sobre la programación financiera de proyectos de inversión gestionados a través del Sistema de Presupuesto y Giro de Regalías (SPGR).
- La información está organizada por entidades ejecutoras, proyectos, códigos presupuestales y programación de recursos a lo largo del tiempo.

b. Principales Variables en el Archivo

- **Identificación del Proyecto**:
  Vigencia Presupuestal (2025-2026 en este caso).
  Código y Nombre de la Entidad (Nombre de la Entidad, Identificación de la Entidad).
  NIT del Ejecutor del Proyecto (NIT del Ejecutor del proyecto de Inversión).
  Código y Nombre del Proyecto (Identificación del proyecto de Inversión, Nombre del proyecto de Inversión).
- **Información Presupuestaria**:
  Código del recurso presupuestal (Código del recurso presupuestal).
  Código y Descripción del Fondo (Código del Fondo, Descripción del Fondo).
  Tipo de recurso (Descripción del recurso presupuestal).
- **Cronograma de Flujos Financieros**:
  Fechas en formato **"AAAA-MM-DD"** que representan meses específicos desde el **2002** hasta **2025**.
  Montos programados en cada fecha, indicando desembolsos o giros esperados para los proyectos.

c. Problemas Identificados

- **Valores Faltantes**:
  En algunas columnas, especialmente en los códigos presupuestales, hay valores vacíos o incompletos.
- **Formato de Fechas y Desembolsos**:
  Hay múltiples columnas con **fechas en formato de texto** (ejemplo: 2002-04-01 00:00:00), lo que puede dificultar el análisis temporal.
  La mayoría de los valores en estas fechas están en **cero (0)**, lo que podría indicar falta de programación efectiva o datos incompletos.
- **Nombres de Variables Largos e Inconsistentes**:
  Algunos nombres de columnas como **"Identificación del proyecto de Inversión"** pueden acortarse para mayor claridad.
  Los nombres de entidades y fondos deben estandarizarse para evitar inconsistencias en el análisis.

d. Acciones de Transformación de Datos

Para mejorar la usabilidad de este archivo en el proceso **ETL**

- **Eliminar columnas irrelevantes** y dejar solo aquellas con valores financieros relevantes.
- **Convertir columnas de fechas a formato adecuado (YYYY-MM-DD)** y organizarlas en un formato más estructurado.
- **Filtrar datos con montos programados diferentes de cero**, para enfocarse en proyectos con ejecución presupuestal real.
- **Normalizar nombres de entidades y códigos presupuestales**, asegurando consistencia con otros archivos financieros.
- **Transformar el dataset a un modelo relacional**, con una tabla para **entidades, proyectos y desembolsos por fecha**.

5.2.5. Archivo "SPGR. ExportarPAGOS 04-03-2025.xlsx"

a. Estructura del Archivo

- Contiene 46 columnas y varios registros de pagos, con información sobre documentos de pago, fechas, montos y beneficiarios.
- Se enfoca en pagos realizados bajo el Sistema de Presupuesto y Giro de Regalías (SPGR).
- Los pagos incluyen detalles sobre deducciones, compromisos, cuentas por pagar y reintegros.
- Principales Variables en el Archivo

Identificación del Pago:

Número de Documento (Numero Documento).
Fecha de Registro y **Fecha de Pago**, que indican el ciclo de ejecución del desembolso.
Estado del Pago (Estado), donde la mayoría están marcados como **"Pagada"**.

Valores Monetarios:

Valor Bruto (Valor Bruto).
Valor Deducciones (Valor Deducciones).
Valor Neto (Valor Neto), que indica el monto final después de descuentos.

Información del Beneficiario:

Tipo de Beneficiario (Tipo Beneficiario), usualmente **"Beneficiario final"**.
Tipo de Identificación y **Número de Identificación** del beneficiario.

Estructura Presupuestaria:

Compromisos y Obligaciones (Compromisos, Obligaciones).
Cuentas por Pagar y Órdenes de Pago (Cuentas por Pagar, Órdenes de Pago).
Reintegros (Reintegros), que aparecen en **cero (0)** en la mayoría de registros.

b. Problemas identificados

- **Valores Faltantes**:
  Algunas columnas como **"Concepto Pago"** y **"Documento Masivo"** están en **blanco** en varios registros.
  Existen valores nulos en los números de cuenta bancaria.
- **Formato de Fechas y Valores Monetarios**:
  Las fechas están en formato de texto (YYYY-MM-DD HH:MM:SS), lo que puede requerir conversión a datetime.
  Algunos valores numéricos tienen espacios adicionales (Valor Bruto, Valor Neto), lo que puede generar problemas al hacer cálculos.
- **Datos Repetidos**:
  Algunos registros tienen **idénticos valores en varias columnas**, lo que puede indicar duplicación de pagos.
- **Información Financiera Incompleta**:
  Cuentas por Pagar y Obligaciones tienen valores que podrían no coincidir con los registros de compromisos.
  Reintegros aparecen en **cero en todos los registros**, lo que puede indicar que la información de devoluciones no está siendo registrada.

c. Acciones de Transformación de Datos

Para mejorar la usabilidad de este archivo en el proceso **ETL**

- **Limpiar y convertir las fechas** a un formato de fecha legible.
- **Eliminar espacios y caracteres adicionales** en valores monetarios.
- **Identificar registros duplicados y consolidarlos** si es necesario.
- **Completar valores faltantes** en Concepto Pago y otros campos críticos.
- **Estandarizar las cuentas por pagar, compromisos y obligaciones**, verificando consistencia con los valores de pago.
- **Crear una tabla de beneficiarios única** con identificadores limpios.

6. EXTRACCION DE DATOS

En primera instancia enlistamos los dataframe en formatos. excel y .csv que han identificado previamente como insumo para el desarrollo del flujo de caja, para ello, se optó por dividirlo en 5 categorías según la fuente desde la cual se extrajo, quedando de la siguiente forma:

FUENTE	DATA BASE
GESPROY	GESPROY. CG-CG-CG-proy_04_marzo_2025_09-38-10.xlsx.xlsx
PROYECTOS	MATRIZ GESTIÓN DE PROYECTOS 04-03-2025.csv
SECOP	SECOP 1. CONSULTA PRINCIPALES.csv SECOP 2. CONSULTA PRINCIPALES.csv
SOUL	EXCEL_FACTURACION_2024.csv RECAUDO2.csv
SPGR	SPGR. ExportarCRONOGRAMA DE FLUJOS 04-03-2025 SPGR. ExportarPAGOS 04-03-2025.csv

Seguidamente cada una pasa por un proceso de ingesta, selección y limpieza, como se presenta a continuación:

- *SECOP*

Para los archivos pertenecientes a esta categoría se lleva a cabo el mismo proceso, donde una vez sea importada se procede a imprimir el listado de las columnas de cada una, con el fin de seleccionar únicamente las de interés, siendo divididas de la siguiente manera

SECOP 1

- 'Anno Cargue SECOP',
- 'Nombre Entidad',
- 'NIT de la Entidad',
- 'Tipo De Contrato','Municipios Ejecucion',
- 'Numero de Contrato',
- 'Cuantia Proceso',
- 'Cuantia Contrato',
- 'Valor Total de Adiciones',
- 'Valor Contrato con Adiciones',
- 'Municipio Entidad',
- 'Departamento Entidad'

SECOP 2

- 'Entidad',
- 'Nit Entidad',
- 'Fecha de Publicacion del Proceso'

Teniendo presente que solo se necesita información del 2024 para cada una utilizamos las columnas; Anno Cargue SECOP y Fecha de Publicacion del Proceso para filtrar la información solo correspondiente a esa vigencia. Por último, se procede a guardar los datos de nuestro interés en las siguientes rutas:

"SECOP/df_filtrado_secop1_2024.csv"
"SECOP/df_filtrado_secop2_2024.csv"

- *MATRIZ PROYECTOS*

Para este archivo en excel se realizo un proceso mas arduo, debido a que se presentaba muchos datos en formato “NaN” e incluso columnas vacías, una vez en listada las columnas, se seleccionan las siguiente para un posterior procesamiento:

- 'Estado Principal',
- 'Estado Derivados',
- ' Valor Aporte de la Entidad ($)',
- 'N Proyecto / Contrato / Convenio / ',
- 'Entidad Contratante',
- 'Departamento',
- 'Año',
- ' Valor Aporte de la Entidad ($).1',
- 'Porcentaje Aporte de la Entidad (%)',
- 'Valor Aporte de Fonpacífico ($)',
- 'Porcentaje Aporte de Fonpacífico (%)',
- 'Entidad del Derivado'

Como se logra observar incluso los nombres de las columnas cuenta con un espacio antes o después del nombre lo que puede presentar errores y confusiones al momento de procesar, por lo cual se opta por renombrar dichas columnas más adelante, siguiendo con este proceso al igual que los .csv anteriores, este también se filtra por vigencia 2024 a través de la columna: 'Año' y se guarda de la siguiente forma:"MATRIZ_PROYECTOS/df_filtrado_matrizgestion proyectos_2024.csv"

Llevando a cabo el mismo proceso se realiza lo mismo para las plataformas de “***GESPROY***” y “***SPGR***”, sin embargo, los datos proporcionados representan vigencia 2025, por lo cual, para este primer análisis, se opta por no tomar en cuenta dicha información, hasta que no se analice dicho año.

Finalmente, la ultima fuente que se procesa es la de SOUL, donde, se seleccionan las siguientes columnas para procesar:

FACTURACION

- 'codigo',
- 'fecha',
- 'subtotal',
- 'total',
- 'saldo',
- 'nom_tercero',
- 'abonos',

RECAUDOS

- 'fecha',
- 'factura',
- 'nombre',
- 'consignacion',
- 'deducciones',
- 'nombre'

Y se almacenan en las siguientes rutas:
"SOUL/df_filtrado_facturacion.csv"
"SOUL/df_filtrado_recaudo.csv"

7. TRANSFORMACION

El principal objetivo de este primer análisis es concentrarse en el **flujo de caja de 2024**, lo cual requiere datos directamente relacionados con las transacciones financieras, pagos, aportes y facturación ocurridos en ese año. Los archivos provenientes de **SECOP**, aunque contienen información sobre los procesos de contratación pública, no tienen un impacto directo en el flujo de caja para el análisis de 2024, ya que los datos sobre contratos y procesos anteriores no reflejan transacciones o pagos realizados en dicho año.

Por lo tanto, se decide **descartar el uso de SECOP en este primer acercamiento**, para evitar introducir datos

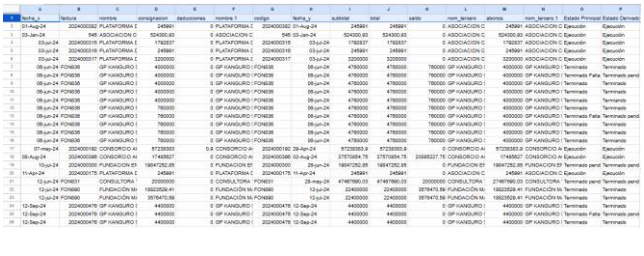
irrelevantes que puedan complicar el análisis y dificultar la identificación de los flujos de caja específicos de 2024. Esta decisión asegura que el análisis se enfoque de manera más directa en los aspectos financieros que son fundamentales para el análisis de flujo de caja.

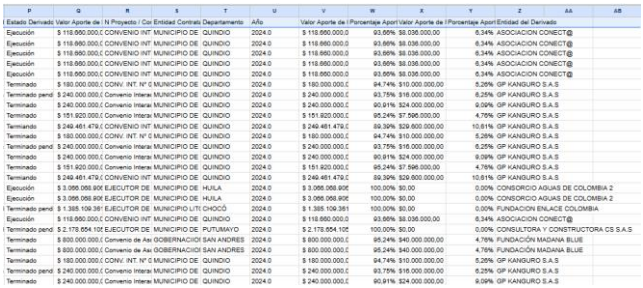
Teniendo en cuenta lo anterior la transformación se llevará a cabo con las siguientes fuentes;

- "SOUL/df_filtrado_facturacion.csv"
- "SOUL/df_filtrado_recaudo.csv"
- "MATRIZ_PROYECTOS/df_filtrado_matrizgestionproyectos_2024.csv"

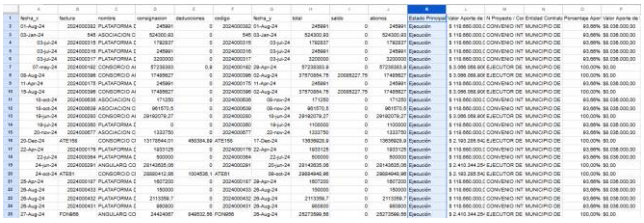
Como primera etapa de la transformación se procede a realizar merge entre los dos DataFrames de facturación y recaudo; Esto a través de un inner mediante las llaves: “factura” y “codigo”, obteniendo asi la siguiente tabla denominada: df_combinado_facturacion_recaudos.csv

Una vez realizado dicho cruce se hace el mismo procedimiento, pero ahora para incluir la matriz de proyectos, adquiriendo como nombre: df_combinado_facturacion_recaudos_matriz.csv

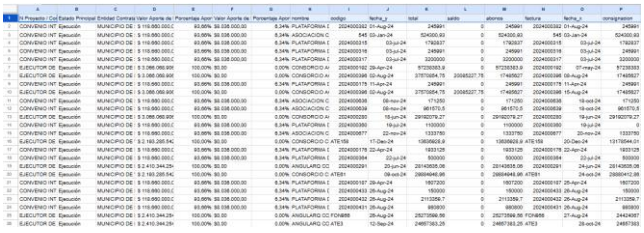




Luego de ello, y de eliminar las columnas que cuentan con la misma información, haciendo que la matriz sea redundante, se identifica que es necesarios los proyectos que se encuentren en estado de ejecución, por lo cual, haciendo uso de esta variable ‘’Estado Principal’’ se trae únicamente los proyectos que correspondan;



Seguidamente, se re ordenan las columnas, con el fin de que sea más fácil su visualización e interpretación de datos interrelacionados



Una vez lista la información se verifica el tipo de formato en que se encuentran los datos numéricos; Los valores numéricos están almacenados como **cadenas de texto (object)** en lugar de números (int o float). Esto ocurre porque los valores incluyen **símbolos de moneda, comas como separadores de miles y puntos como separadores decimales**. Para realizar cálculos correctamente, es necesario **convertirlos a números** eliminando los caracteres no numéricos.

Como se oberva en cuanto a las columnas de valores numéricos, cada una cuenta con un formato diferente, por lo cual, se identifican cuales comparten el mismo, y se agrupan, como se muestra a continuación:

"Valor Aporte de la Entidad ($)" y "Valor Aporte de Fonpacífico ($)"

- Ambos valores incluyen el signo $, tienen separador de miles con punto (.) y separador decimal con coma (,).
- Ejemplo: $ 118.660.000,00 y $8.036.000,00.

"total", "abonos", "consignacion", "deducciones"

- Estos valores parecen ser numéricos, pero algunos registros tienen coma como separador decimal (524300,93), mientras que otros solo tienen números enteros (245991).
- En algunos casos, los valores no tienen separadores de miles, lo que indica que hay inconsistencias en el formato.
- Ejemplo: 245991, 524300,93, 1792837.

"saldo"

- En la muestra proporcionada, todos los valores de la columna son 0, lo que sugiere que podrían ser enteros o ser tratados como string.

Para eliminar los formatos iniciales, se aplica el siguiente procedimiento;

Para estandarizar los valores de "**Valor Aporte de la Entidad ($)**" y "**Valor Aporte de Fonpacífico ($)**", eliminaremos el formato monetario, es decir:

1. Se suprime el símbolo $.
2. Eliminaremos los puntos que separan los miles.
3. Reemplazaremos la coma decimal por un punto (para convertirlo en un formato numérico estándar).

```
🔹 Limpieza de montos:
   Valor Aporte de la Entidad ($)  Valor Aporte de Fonpacífico ($)
0                   118660000.0                         8036000.0
1                   118660000.0                         8036000.0
2                   118660000.0                         8036000.0
3                   118660000.0                         8036000.0
4                   118660000.0                         8036000.0
```

Para el segundo grupo de columnas: "**total**", "**abonos**", "**consignacion**", "**deducciones**", vamos a:

1. Eliminar los separadores de miles (si existen).
2. Reemplazar la coma decimal por un punto para mantener un formato numérico estándar.
3. Convertir los valores a tipo float.

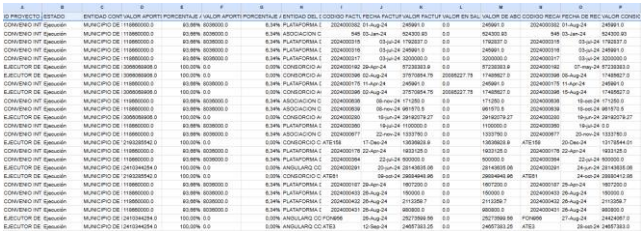
```
🔹 Limpieza de números:
      total      abonos  consignacion  deducciones
0    245991.00   245991.00     245991.00          0.0
1    524300.93   524300.93     524300.93          0.0
2   1792837.00  1792837.00    1792837.00          0.0
3    245991.00   245991.00     245991.00          0.0
4   3200000.00  3200000.00    3200000.00          0.0
```

Para garantizar que la columna "**saldo**" tenga el mismo formato que los demás valores numéricos (dos decimales), aplicaremos una limpieza específica asegurando que todos los valores sean tratados como float con dos decimales.

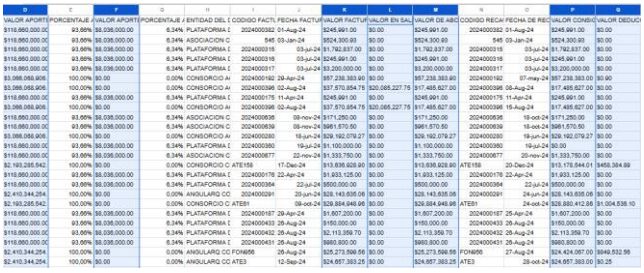
```
🔹 Limpieza de saldo:
0    0.0
1    0.0
2    0.0
3    0.0
4    0.0
Name: saldo, dtype: float64
```

El proceso de limpieza de datos aplicado a la matriz de facturación y recaudos consistió en la sustitución de valores incorrectos o mal formateados en las columnas monetarias y numéricas, asegurando la coherencia y precisión de la información. Para ello, se creó una copia del DataFrame original con el fin de preservar la estructura de los datos y evitar modificaciones no deseadas. Posteriormente, se reemplazaron las columnas monetarias con montos corregidos, las columnas numéricas con valores limpios y la columna de saldo con datos ajustados. Finalmente, el DataFrame procesado se exportó en formato CSV con codificación UTF-8 y sin incluir índices, almacenándose en un repositorio en Google Drive.

Luego de ello renombramos las columnas con el fin de evitar el error mencionado previamente y ayudar a una mejor interpretación de los datos presentes



El proceso de tratamiento de datos en la matriz de facturación y recaudos consistió en la carga de un archivo CSV previamente renombrado, seguido de la eliminación de columnas irrelevantes para el análisis. Posteriormente, se aplicó un formato monetario a las columnas financieras para garantizar la legibilidad de los valores numéricos. Se generó una copia del DataFrame original para evitar modificaciones sobre los datos de origen y se transformaron los valores en las columnas correspondientes mediante una función que les asignó el formato de moneda con separación de miles y dos decimales. Finalmente, los datos formateados se almacenaron en un nuevo archivo CSV y en un archivo Excel, donde además se aplicó alineación a la derecha y un formato específico de moneda en las celdas numéricas.



Como se logra observar ya las columnas;

- 'VALOR APORTE DE LA ENTIDAD',
- 'VALOR APORTE FONPACIFICO',
- 'VALOR FACTURACION',
- 'VALOR EN SALDO',
- 'VALOR DE ABONOS',
- 'VALOR CONSIGNADO',
- 'VALOR DEDUCCIONES'

Presentan el mismo formato, lo que permite poder Generar nuevas columnas basadas en los datos existentes, con cálculos matemáticos;

En este fragmento realiza varios cálculos sobre los datos de facturación y recaudos, organizándolos en un resumen mensual. Primero, convierte la columna de fecha a un formato de año y mes (%Y-%m) para agrupar los datos por período. Luego, calcula tres métricas clave para cada mes: **total facturado** (suma de VALOR FACTURACION), **total recaudado** (suma de VALOR CONSIGNADO) y **saldo pendiente** (suma de VALOR EN SALDO). Después de obtener estos valores agregados, se les aplica un formato de moneda para mejorar la legibilidad. Finalmente, el resumen mensual se guarda en archivos CSV y Excel, aplicando formato numérico y alineación a la derecha en las celdas monetarias del archivo Excel para una mejor presentación de los datos.

MES	TOTAL_FACTURAD	TOTAL_RECAUDAD	SALDO_PENDIENT
2024-01	$524,300.93	$524,300.93	$0.00
2024-04	$61,024,699.90	$61,024,699.00	$0.00
2024-06	$57,335,714.33	$57,335,714.33	$0.00
2024-07	$7,084,819.00	$5,984,819.00	$0.00
2024-08	$104,651,450.76	$63,631,462.70	$40,170,455.50
2024-09	$24,657,383.25	$24,657,383.00	$0.00
2024-10	$29,884,948.96	$28,880,412.86	$0.00
2024-11	$2,958,551.50	$2,958,551.50	$0.00
2024-12	$13,636,928.90	$13,178,544.01	$0.00

8. VALIDACION DE DATOS

Para validar la correcta generación del archivo de resumen mensual de facturación, se utilizó un procedimiento de carga y verificación de datos en un entorno de Python con la librería Pandas. Primero, se definió la ruta del archivo CSV almacenado en Google Drive y se montó el entorno para permitir su acceso. Posteriormente, el archivo fue leído utilizando pd.read_csv(), transformándolo en un DataFrame para su manipulación y análisis. Finalmente, se visualizaron las primeras filas del conjunto de datos mediante df_resumen.head(), lo que permitió verificar la estructura y coherencia de la información. Esta validación asegura que el archivo se generó correctamente, sin errores en la conversión ni en la integridad de los datos.

	MES	TOTAL_FACTURADO	TOTAL_RECAUDADO	SALDO_PENDIENTE
0	2024-01	$524,300.93	$524,300.93	$0.00
1	2024-04	$61,024,699.90	$61,024,699.00	$0.00
2	2024-06	$57,335,714.33	$57,335,714.33	$0.00
3	2024-07	$7,084,819.00	$5,984,819.00	$0.00
4	2024-08	$104,651,450.76	$63,631,462.70	$40,170,455.50

Este fragmento de código realiza una validación de la calidad de los datos en el DataFrame df_resumen. Primero, utiliza df_resumen.isnull().sum() para contar y mostrar la cantidad de valores nulos en cada columna, lo que permite identificar posibles inconsistencias o datos faltantes. Luego, con df_resumen.dtypes, se verifica el tipo de dato de cada columna para asegurarse de que los valores tienen el formato esperado (por ejemplo, números en columnas monetarias y fechas en formato datetime).

```
Valores nulos por columna:
MES                 0
TOTAL_FACTURADO     0
TOTAL_RECAUDADO     0
SALDO_PENDIENTE     0
dtype: int64

Tipos de datos de cada columna:
MES                 object
TOTAL_FACTURADO     object
TOTAL_RECAUDADO     object
SALDO_PENDIENTE     object
dtype: object
```

En este fragmento se implementa validaciones adicionales para verificar la consistencia de los datos en el resumen mensual. Primero, df_resumen["MES"].unique() muestra los valores únicos en la columna "MES" para asegurarse de que los datos están correctamente agrupados por mes y no contienen valores atípicos. Luego, se recalculan las sumas totales de "TOTAL_FACTURADO", "TOTAL_RECAUDADO" y "SALDO_PENDIENTE" en el resumen mensual y se comparan con los valores originales del dataset cargado desde df_combinado_facturacion_recaudos_matriz_renombrado. csv. Esto permite detectar discrepancias y validar que no haya errores en el procesamiento de datos.

```
Valores únicos en la columna MES:
['2024-01' '2024-04' '2024-06' '2024-07' '2024-08' '2024-09' '2024-10'
 '2024-11' '2024-12']
Suma de valores en el resumen mensual:
TOTAL_FACTURADO    $524,300.93$61,024,699.90$57,335,714.33$7,084,...
TOTAL_RECAUDADO    $524,300.93$61,024,699.00$57,335,714.33$5,984,...
SALDO_PENDIENTE    $0.00$0.00$0.00$0.00$40,170,455.50$0.00$0.00$0...
dtype: object

Suma de valores en el dataset original:
VALOR FACTURACION    3.017588e+08
VALOR CONSIGNADO     2.581759e+08
VALOR EN SALDO       4.017046e+07
dtype: float64
```

Este fragmento de código tiene como objetivo eliminar el formato de moneda de las columnas financieras y convertirlas a valores numéricos para facilitar cálculos posteriores. La función limpiar_moneda(valor) elimina los símbolos de dólar y las comas en caso de que los valores sean cadenas de texto, convirtiéndolos en números flotantes. Luego, se aplica esta función a las columnas "TOTAL_FACTURADO", "TOTAL_RECAUDADO" y "SALDO_PENDIENTE" para asegurarse de que los datos sean de tipo numérico. Finalmente, df_resumen.dtypes permite verificar que la conversión se haya realizado correctamente.

```
MES                 object
TOTAL_FACTURADO    float64
TOTAL_RECAUDADO    float64
SALDO_PENDIENTE    float64
dtype: object
```