**Documentación del Proceso ETL – CEMPRE**

# 1. Objetivo

Construir un proceso ETL que transforma la base de facturas de venta de materiales aprovechables (hoja "Reporte Inicial") en el formato requerido por la plantilla "Reporte Solicitado", incorporando normalización de municipios, mapeo de regiones, clasificación de familia de material, cálculo de factores de Envases y Empaques (E&E), certificación, y distribución de cantidades por empresas transformadoras según porcentajes definidos. Se generan salidas para consumo analítico y un tablero en Power BI.

# 2. Alcance y supuestos

* Lectura flexible de hojas desde "Copia de Analista de datos\_prueba excel.xlsx" y archivo de soporte "regiones.xlsx" obtenido directamente del DANE.
* Normalización de municipios y cruce con regiones mediante llave mun\_key.
* Clasificación de material en una de 10 familias permitidas; no se producen categorías "Otros", "Maderables", "Textiles" ni "Sin asignar".
* Distribución de cantidades por empresa transformadora de acuerdo con porcentajes definidos por Iniciativa y Familia.
* Cálculo de toneladas y toneladas E&E usando factores de la hoja correspondiente.
* Generación de tres salidas CSV: reporte\_solicitado.csv, agg\_por\_tipo.csv y agg\_por\_familia.csv.

# 3. Fuentes de datos

* Excel principal: Copia de Analista de datos\_prueba excel.xlsx
* Hojas esperadas: "Reporte Inicial", "Iniciativas", "Factor Envases y empaques", "Porcentajes empresas", "Condiciones precertificación"
* Regiones: regiones.xlsx (Municipio ↔ Región)

# 4. Normalización y utilidades

Se implementan utilidades para limpiar texto y construir llaves robustas de cruce:

* Función \_norm(s): normaliza acentos, elimina puntuación, convierte a minúsculas y colapsa espacios.
* Función clean\_text(s): minúsculas, elimina paréntesis/marcas, filtra caracteres no alfanuméricos y normaliza espacios.
* Diccionario NORMALIZACION\_MUNS: corrige casos puntuales (p. ej., 5001→MEDELLÍN, CALI→SANTIAGO DE CALI, CÚCUTA→SAN JOSÉ DE CÚCUTA, TOLU→SANTIAGO DE TOLÚ).
* Función \_normalizar\_municipio\_valor(v): aplica el diccionario y homologa mayúsculas/acentos.
* Función \_agregar\_clave\_municipio(df): crea la llave mun\_key con \_norm sobre la columna Municipio.

# 5. Carga flexible de hojas

La función \_cargar\_hoja busca cada hoja por múltiples alias normalizados para tolerar variaciones de nombre.

# 6. Mapeo de regiones

Se lee regiones.xlsx, se renombran columnas a [Municipio, Región], se crea mun\_key y se deduplican registros por mun\_key conservando la primera región válida. El cruce al reporte se realiza por mun\_key (left join).

# 7. Transformaciones principales

## 7.1 Renombrado de columnas (Reporte Inicial)

Se homologan encabezados a: Gestor, Fecha de transacción, No de factura, Municipio, Tipo de Material, Cantidad entregada (Kg), Intermediario 1, NIT del Comprador. Si faltan columnas requeridas se detiene con error descriptivo.

## 7.2 Tipos y fechas

* Cantidad entregada (Kg): conversión a numérico con coerción.
* Fecha de transacción: parseo a datetime con dayfirst=True.
* Derivados: Año, Periodo (YYYY-MM), Fecha de transacción (formato YYYY/MM/DD).

## 7.3 Iniciativa por Gestor

Se cruza la hoja "Iniciativas" para obtener la columna Iniciativa por Gestor.

## 7.4 Clasificación de Familia

La función map\_family(txt) clasifica el material en una de 10 familias exactas, con prioridad para evitar ambigüedades (p. ej., "kraft" clasifica como Cartón, no como Papel). Categorías permitidas:

* - Vidrio
* - Cartón
* - PET
* - PE - PP Rígido
* - Papel
* - Metales Ferrosos
* - PE - PP Flexible
* - Cartón para bebidas
* - Otros Plásticos
* - Metales No Ferrosos

Reglas clave (resumen):

* Cartón: "carton", "cartón", "kraft", "kraf", "plegad", "corrugado"…
* Papel: "papel", "archivo", "periodico", "pasta", "bond"… (se evalúa tras Cartón).
* PET: "pet", "tereftalato".
* PE/PP Rígido: "polietileno alta", "HDPE", "polipropileno", "pp" (con regex de palabra), "rígido".
* PE/PP Flexible: "polietileno baja", "LDPE", "flexible", "película", "bolsa".
* Vidrio: "vidrio".
* Otros Plásticos: "pvc", "policloruro", "plástico", "acrílico", "ps"…
* Metales No Ferrosos: "aluminio", "cobre", "bronce", "lata".
* Metales Ferrosos: "acero", "chatarra", "hierro", "metal".
* Fallback controlado: asigna Papel si no hay match específico (evitando "Otros").

## 7.5 Factor de E&E

Se cruza la hoja "Factor Envases y empaques" por Tipo de Material. Se calcula:

* Cantidad entregada (Ton) = Kg / 1000
* Cantidad de E&E (Ton) = Ton × Factor Envases y empaques

## 7.6 Certificación

Con la hoja "Condiciones precertificación" se identifica si (Intermediario 1, Tipo de Material) aparece en la lista, generando el booleano Certificado.

## 7.7 Distribución por empresa transformadora

Con la hoja "Porcentajes empresas" se realiza un merge por [Iniciativa, Familia] que expande filas: cada factura se replica por cada empresa con su porcentaje. Se preserva Cantidad\_Original\_Kg y se distribuye así:

* Cantidad entregada (Kg) distribuida = Cantidad\_Original\_Kg × Porcentaje de asignación
* Se recalculan Ton y E&E con la cantidad distribuida

# 8. Columnas del reporte final

|  |  |
| --- | --- |
| Columna | Descripción breve |
| Año | Año calendario de la transacción |
| Periodo | Periodo YYYY-MM |
| Región | Región según regiones.xlsx |
| Municipio | Municipio normalizado |
| Iniciativa | Iniciativa del Gestor |
| Fecha de transacción | YYYY/MM/DD |
| No de factura | Consecutivo de la factura |
| Tipo de Material | Descripción original del material |
| Familia | Familia asignada (una de 10) |
| Cantidad entregada (Kg) | Cantidad distribuida por porcentaje |
| Cantidad entregada (Ton) | Kg / 1000 |
| Cantidad de E&E (Ton) | Ton × Factor E&E |
| Intermediario 1 | Nombre del intermediario/transformador receptor |
| Empresa transformadora | Empresa asignada según porcentaje |
| Certificado | True/False según precertificación |
| Gestor | Nombre del gestor original |
| NIT del Comprador | NIT del intermediario/transformador |

# 9. Agregados y salidas

* agg\_por\_tipo.csv: agregado por [Año, Periodo, Iniciativa, Tipo de Material] con sumatorias de Kg, Ton, E&E Ton.
* agg\_por\_familia.csv: agregado por [Año, Periodo, Iniciativa, Familia] con sumatorias de Kg, Ton, E&E Ton.
* reporte\_solicitado.csv: detalle transaccional listo para análisis/tablero.

# 10. Validaciones y manejo de errores

* Validación de columnas requeridas en "Reporte Inicial"; error explícito si faltan.
* Normalización de municipios y llave mun\_key para minimizar inconsistencias.
* Deduplicación del mapa de regiones por mun\_key conservando el primer match válido.
* Conversión numérica segura de cantidades; fechas con dayfirst=True.
* Fallback controlado en clasificación para evitar categorías no permitidas.

# 11. Recomendaciones y próximos pasos

* Ampliar diccionario de NORMALIZACION\_MUNS con nuevos casos observados.
* Gestionar catálogo de materiales y familias para reducir dependencia de heurísticas.
* Agregar pruebas unitarias de clasificación de familia (casos borde como "cartón kraft").
* Versionar archivos de soporte (regiones.xlsx) para trazabilidad.
* Automatizar ejecución programada y validación de salidas.

# 12. Flujo del proceso (descripción)

Lectura de hojas ⇨ Normalización de municipios ⇨ Cruce de regiones ⇨ Renombrado y tipos ⇨ Iniciativa ⇨ Clasificación de Familia ⇨ Factores E&E ⇨ Certificación ⇨ Distribución por porcentaje ⇨ Cálculos finales ⇨ Exportación de CSVs y agregados.