Proceso de limpieza y re ordenamiento de los datos del sistema de expediente SIE, usando Power BI

R

Limpieza de datos con Power BI

Diplomatura en Ciencias de Datos - 2021

Autora: Ing. Romina C. Scarazzini

Tutora: Ing. Layla Scheli

Tabla de contenido

[Limpieza de Datos con Power BI 2](#_Toc82000700)

[Objetivo 2](#_Toc82000701)

[Resumen de actividades realizadas 2](#_Toc82000702)

[Origen de los datos 2](#_Toc82000703)

[Extracción de Datos desde Power Bi 2](#_Toc82000704)

[Primeras transformaciones de los datos 3](#_Toc82000705)

[DIM\_Oficina - Secuencia de pasos 3](#_Toc82000706)

[DIM\_Tema – Secuencia de pasos 4](#_Toc82000707)

[Terminales\_Habilitadas – Secuencia de pasos 5](#_Toc82000708)

[DIM\_Fecha – Secuencia de pasos 5](#_Toc82000709)

[FAC\_Ult\_Mov\_Expte – Secuencia de pasos 6](#_Toc82000710)

[Relaciones entre la tabla de hechos y las dimensiones 9](#_Toc82000711)

[Siguientes transformaciones de datos 10](#_Toc82000712)

[Armado de Dataset para Python 12](#_Toc82000713)

# Limpieza de Datos con Power BI

## Objetivo

Este documento incluye el detalle de los pasos que se realizaron para limpiar y re ordenar los datos del sistema de expedientes de la provincia de Santa FE (SIE), con el objetivo de usarlos para análisis de Business Inteligence (BI) y Maching Learning.

## Resumen de actividades realizadas

1. Extracción de datos desde archivos .txt
2. Configuración de los tipos de datos
3. Configuración de fechas
4. Eliminación de registros con datos nulos y con errores
5. Para las oficinas: División de columna que contiene el nombre de la oficina para obtener: Tipo de Organismo, Organismo y Nombre de la oficina
6. Generación de índice
7. Eliminación de columnas que no son necesarias para el análisis que se va a realizar
8. Eliminación de expedientes agregados a otros expedientes
9. Filtrado de registro: Fecha de inicio de expediente desde 01/01/2011 al 31/12/2020

## Origen de los datos

Diariamente se generan cuatro archivos en formato .txt que contienen el último movimiento realizado de todos los expedientes del sistema SIE. Con dicha información se actualiza diariamente la Base de Datos de SIE:

* ispyc.diario: último movimiento realizado de todos los expedientes del sistema SIE
* ispyc.ofi: oficinas del sistema de expedientes SIE
* ispyc.tem: temas del sistema de expedientes SIE. Cada expediente tiene un tema asignado
* ispyc.tha: terminales habilitadas sistema de expedientes SIE (puestos de trabajo)
* Fechas: Planilla generada con formato “.calc”, que contiene una columna con todas las fechas desde el año 1900 al 2030

Nota: Por otro lado, hay un webservice soap online que obtiene la información de SIEi en tiempo real contra el mainframe

## Extracción de Datos desde Power Bi

Desde Power Query en Power BI se extraen los archivos necesarios más una planilla de cálculo que contiene las fechas desde 1900 al 2030. Dichos archivos pasan a tener los siguientes nombres:

|  |  |  |  |
| --- | --- | --- | --- |
| **Archivo** | **Tabla** | **Origen** | **Formato** |
| ispyc.diario | FAC\_Ult\_Mov\_Expte | SIE | .txt |
| ispyc.ofi | DIM\_Oficinas | SIE | .txt |
| ispyc.tem | DIM\_Temas | SIE | .txt |
| ispyc.tha | DIM\_Terminales | SIE | .xtx |
| Fechas | DIM\_Fechas | Generación propia | .calc |

## Primeras transformaciones de los datos

### DIM\_Oficina - Secuencia de pasos

1. Renombrar columnas

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre actual** | **Nombre a cambiar** | **Tipo de dato que corresponde** | **Descripción** |
| Column1 | Cod\_Oficina | String (contiene valores)alfanuméricos) | Código de la oficina. |
| Column2 | Nom\_Oficina | String | Nombre de la oficina, contiene el nombre del Ministerio u Organismo y el nombre de la oficina dentro del mismo. |
| Column3 | Baja | Boolean. Valores: “V”:False, “\*: True” | Indica si la oficina fue dada de baja o no |

1. La columna 4 debe ser eliminada.
2. Eliminar las filas con valores nulos y las que contienen errores.
3. Columna Baja:
   1. Reemplazar los “\*” por True y “V” por False
   2. Cambiar a tipo datos “Booleano”
4. Agregar columnas: Id\_Tipo\_Organismo
   1. Se genera con el primer dígito del campo “Cod\_Oficina”. Por esto, hay que usar la función “Dividir columna”
   2. Reemplazar valores: 0 por 1, 1 por 2 y 2 por 3
   3. Convertir a tipo integer
5. Duplicar columna Id\_Tipo\_Organismo y renombrar como Nom\_Tipo\_Organismo
   1. Reemplazar valores: 1 por “Administración Central, 2 por Organismos Centralizados y 3 por Organismos Desentralizados
6. Agregar columnas: Cod\_Organismo
   1. Se genera con el 2, 3 y 4 dígito del campo “Cod\_Oficina”. Por esto, hay que usar la función “Dividir columna”
   2. Este campo aún no se pasa al tipo interger porque contiene código de oficinas alfanuméricos
7. Duplicar columna Cod\_Organismo y renombrar como Nom\_Organismo
   1. Reemplazar valores por los nombres de la Ministerio/Organismos principales que figuran en el campo “Nom\_Oficina
8. Generar índice automática: Id\_Oficina
9. La tabla DIM\_Oficina finalmente quedará como sigue:

|  |  |  |
| --- | --- | --- |
| **Columna** | **Tipo de dato** | **Descripción** |
| Id\_Oficina | Integer | Identificador de la oficina |
| Cod\_Oficina | String | Código de la oficina. |
| Id\_Tipo\_Organismo | Integer | Identificación del tipo de organismo al cual pertenece la oficina |
| Nom\_Tipo\_organismo | String | Nombre del tipo de organismo al cual pertenece la oficina |
| Cod\_Organismo | String | Identificación del organismo al cual pertenece la oficina |
| Nom\_Organismo | String | Nombre del organismo al cual pertenece la oficina |
| Nom\_Oficina | String | Nombre de la oficina, contiene el nombre del Ministerio u Organismo y el nombre de la oficina dentro del mismo. |
| Baja | Boolean | Indica si la oficina fue dada de baja o no |

1. Renombrar la tabla DIM\_Oficina por DIM\_Oficina\_Remitente
2. Duplicar la tabla DIM\_Oficina\_Remitente y renombrarla por DIM\_Oficina\_Destino

### DIM\_Tema – Secuencia de pasos

1. Renombrar columnas

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre actual** | **Nombre a cambiar** | **Tipo de dato que corresponde** | **Descripción** |
| Column1 | Cod\_Tema | String (contiene valores alfanuméricos) | Código de tema |
| Column2 | Nom\_Tema | String | Nombre del tema |
| Column3 | Baja | Booleano (Valores: “\*”:True, 0 o Null: False) | Indica si el tema fue dado de baja o no |

1. La columna 4 debe ser eliminada
2. Eliminar las filas con valores nulos y las que contienen errores
3. Columna Baja:
   1. Reemplazar los “\*” por True y “null o 0” por False
   2. Cambiar a tipo datos “Booleano”
4. Generar un índice automático: Id\_Tema
5. La tabla DIM\_Tema finalmente quedará como sigue:

|  |  |  |
| --- | --- | --- |
| **Columna** | **Tipo de dato** | **Descripción** |
| Id\_Tema | Integer | Identificador del tema |
| Cod\_Tema | String | Código de tema |
| Nom\_Tema | String | Nombre del tema |
| Baja | Boolean | Indica si el tema fue dado de baja o no |

### Terminales\_Habilitadas – Secuencia de pasos

1. Renombrar columnas

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre actual** | **Nombre a cambiar** | **Tipo de dato que corresponde** | **Descripción** |
| Column1 | Cod\_Terminal | String | Código de terminal |
| Column2 | Cod\_Oficina (Mesa) | String | Código de oficina a la cual fue asignada la terminal |
| Column3 | Nom\_Oficina (\*) | String | Nombre de la Oficina (\*) |
| Column4 | Cod\_Impresora | String | Código de la impresora a signada a la terminal |

(\*): El nombre de lo oficina se ingresa a mano, por lo cual no es igual al nombre de oficina de la tabla Oficinas, pero es la misma oficina, por lo cual este dato será eliminado

1. Explicación de los datos
   1. Las Mesa (Oficinas) pueden tener o no asignada una terminal. Esta asociación es a través del IP (Terminal/IP). es generada por la gente de teleproceso de la STG. La relación Terminal/Mesa la maneja SIE (el IP no viene en los datos del archivo)
   2. No todas las mesas (oficinas) tienen una terminal asignada. Existe también un pool de terminales que no tienen relación con ninguna IP. Son terminales libres y las mesas (oficinas) que no tiene terminal, van accediendo a ellas, pueden generar expedientes, son solo para consultar expedientes
   3. Mesa tiene 5 posiciones, una mesa puede tener más de una terminal asociada
   4. El código de oficina es el mismo que el de mesa
2. Fuera de alance de esta primer fase
   1. El análisis con los datos de la tabla terminales no se realizará en esta primer fase del proyecto
3. La tabla Terminales\_Habilitadas finalmente quedará como sigue:

|  |  |
| --- | --- |
| **Columna** | **Tipo de dato** |
| Cod\_Terminal | Integer |
| Cod\_Oficina | String |
| Nom\_Oficina2 | String |
| Cod\_Impresora | String |

### DIM\_Fecha – Secuencia de pasos

1. Renombrar columnas

|  |  |  |
| --- | --- | --- |
| **Columna** | **Tipo de dato que corresponde** | **Descripción** |
| Fecha | Date | Fecha |

1. Generar columna Id\_Fecha:
   1. Crear columna personalizada con lo siguiente: Date.Year([Fecha])\*10000+ Date.Month([Fecha])\*100 + Date.Day([Fecha])
   2. Cambiar tipo de dato a Integer
2. La tabla DIM\_Fecha finalmente quedará como sigue:

|  |  |  |
| --- | --- | --- |
| **Columna** | **Tipo de dato** | **Descripción** |
| Id\_Fecha | Integer | Identificador de la fecha |
| Fecha | Date | Fecha |

1. Renombrar la tabla DIM\_Fecha por DIM\_Fecha\_Ini\_Expte
2. Duplicar la tabla DIM\_Fecha\_Ini\_Expte y renombrarla por DIM\_Fecha\_Ult\_Mov\_Expte

### FAC\_Ult\_Mov\_Expte – Secuencia de pasos

1. Renombrar columnas

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre actual** | **Nombre a cambiar** | **Tipo de dato que corresponde** | **Descripción** |
| Column1 | Extpe\_Mesa | String (\*6) | Parte del identificador de un expediente. Número de mesa: es el mismo que el código de oficina (Origen), el cual está formado como se muestra en figura 1 |
| Column2 | Expte\_Nro | Integer | Número de expediente: Parte del identificador de un expediente |
| Column3 | Expte\_Dig | String | Dígito verificado. Puede tener valores entre 0 y 9 y además puede tener el valor “V”, que es cuando el expediente es generado por una entidad externa, ya que para generar un dígito verificador entre 0 y 9, hay un algoritmo interno de la provincia |
| Column4 | Cod\_Tema | String | Código del tema del expediente (es alfanumérico). Este campo debe ser cambiado a String. Power BI lo pone como numérico porque solo evalúa los primeros 1000 registros para definir el tipo de dato de cada campo |
| Column 5 | Iniciador | String | Iniciador: Nombre libre. Persona, Organismo, etc. que inició el expediente (\*1) |
| Column 6 | Doc\_Tipo | String | Tipo de documento de la persona que inicio el trámite (\*1) |
| Column 7 | Doc\_Numero | Integer | Número de documento de la persona que inicio el trámite (\*1) |
| Column 8 | Fec\_Ini\_Anio | Integer | Año de inicio del expediente |
| Column 9 | Fec\_Ini\_Mes | String(\*6) | Mes de inicio del expediente |
| Column 10 | Fec\_Ini\_Dia | String(\*6) | Día de inicio del expediente |
| Column 11 | Concluido | Boleano | Indica si el expediente está o no concluido. En general no se carga porque no es obligatorio, igualmente no vamos a dejar para hacer análisis de expedientes que están como no concluidos pero que realmente lo están |
| Column 12 | Urgente | Boleano | Indica si el expediente es o no urgente. Tiene 3 valores: 0, 1 y 2. El 2 probablemente es un error de migración (\*2) |
| Column 13 | Reservado | Boleano | Indica si el expediente está o no reservado (\*2) |
| Column 14 | Privado | Boleano | Indica si el expediente está o no privado. Tiene 3 valores: 0, 1 y 2. El 2 probablemente es un error de migración (\*2) |
| Column 15 | Ubi\_Anio (Anio\_Pase\_Archivo) | Integer | Año en el cual el expediente pasó a archivo (\*3) |
| Column 16 | Ubi\_Tiempo (Cant\_Anios\_Archivo) | Integer | Cantidad de años que el expediente debe permanecer en el archivo, antes de su destrucción (\*3) |
| Column 17 | Ubi\_Caja (Ubi\_Fis\_Archivo) | String | Número de carpeta, caja o toma, etc. donde se archiva el expediente |
| Column 18 | Fec\_Mov\_Anio | Integer | Año del último movimiento del expediente |
| Column 19 | Fec\_Mov\_Mes | String (\*6) | Mes del último movimiento del expediente |
| Column 20 | Fec\_Mov\_Dia | String (\*6) | Día del último movimiento del expediente |
| Column 21 | Cod\_Ofi\_Remitente | String | Código de la oficina que remite el expediente en el último pase |
| Column 22 | Cod\_Ofi\_Destino | String | Código de la oficina destino del expediente en el último pase |
| Column 23 | Folio (Cant\_Hojas) | Integer | Cantidad de hojas del expediente. Es el número del último folio (hoja) agregado al expediente |
| Column 24 | Expte\_Agre\_Mesa | Integer | Parte del identificador de un expediente. Número de mesa del expediente al cual se agregó este expediente. Si es un expediente agregado a otro, el último pase válido es el del expediente cabecera (\*5) |
| Column 25 | Expte\_Agre\_Nro | Integer | Parte del identificador de un expediente. Número del expediente al cual se agregó este expediente. Si es un expediente agregado a otro, el último pase válido es el del expediente cabecera (\*5) |
| Column 26 | Expte\_Agre\_Dig | String | Parte del identificador de un expediente. Dígito verificador del expediente al cual se agregó este expediente. Si es un expediente agregado a otro, el último pase válido es el del expediente cabecera (\*5) |
| Column 27 | Demora |  | Indica si el expediente tiene no demora (\*2) |
| Column 28 | Concepto |  | Descripción del pase, es un texto libre, por lo cual no se utilizará para el análisis de dat |

(\*1) Estos campos no se cargan por cuestión de confidencialidad

(\*2) Estos campos en general no se cargan porque no son obligatorios, por lo cual no se cargan porque no podemos usarlos para hacer análisis de datos

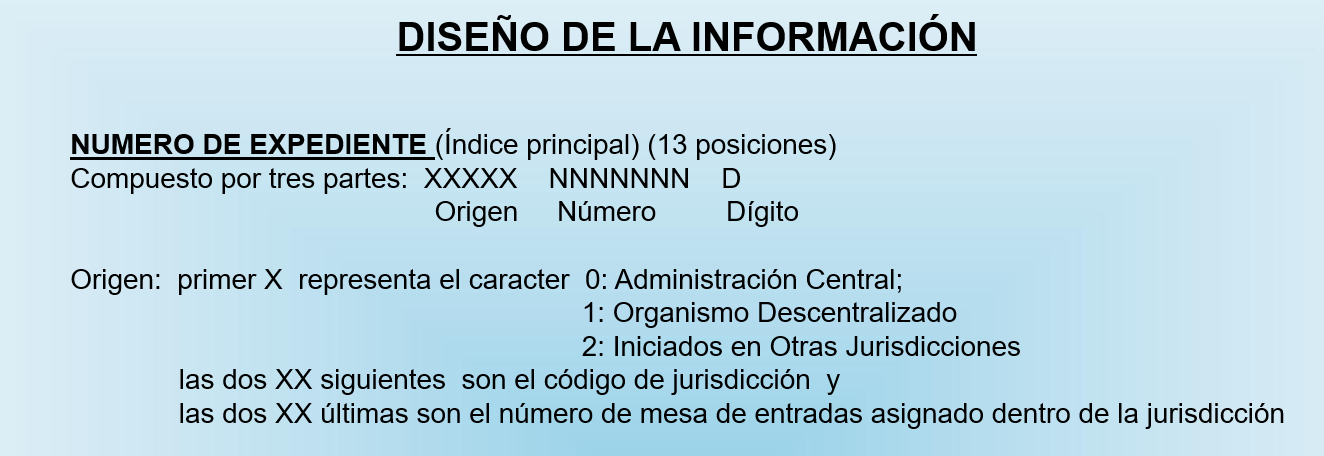
(\*3) Estos campos los usa solo la Caja de Jubilaciones y tampoco los usa mucho, por lo cual no se cargan porque no podemos usarlos para hacer análisis de datos

(\*4): Para más información sobre el desglose de estos campos, ver archivo: “Lista\_Ofi\_SIE”

(\*5): Estos campos se usarán para localizar los expedientes asignados a otros y luego serán eliminados

(\*6): Estos campos se le deben cambiar el tipo de datos a String, sino Power Bi importa los datos como números y quita las ceros a la izquierda, los cuales vamos a necesitar

1. Figura 1



1. Las columnas 5, 6 y 7 deben ser eliminadas corresponden a Iniciador, tipo y nro. de documento porque son datos confidenciales.
2. Las columnas de la 12 a la 17 y 27, deben ser eliminadas, ya que no son muy utilizadas en el sistema, por lo cual no se pueden realizar análisis de los datos
3. La columna 28 y 29 deben ser eliminadas
4. Quitar filas con errores de toda la tabla
5. Filtrar nulo de las columnas: Cod\_Ofi\_Remitente
6. Cambiar el tipo de datos del campo Concluido a booleano

#### Expedientes agregados

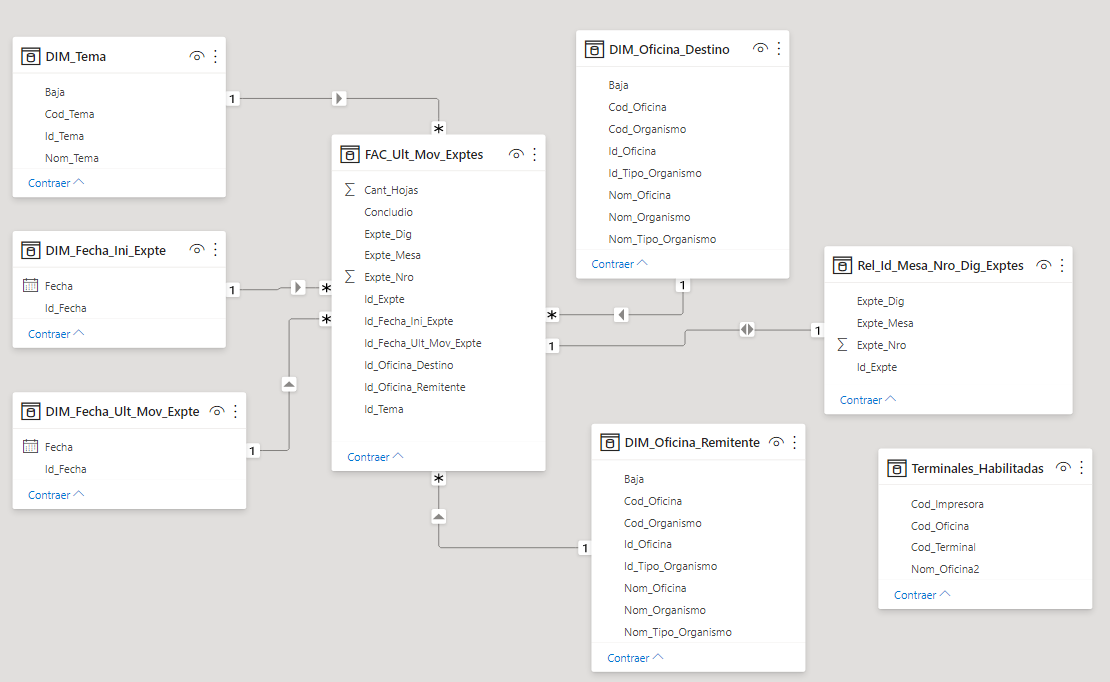
1. Filtrar y eliminar filas de los expedientes que fueron agregados a otros expedientes
   1. Si el campo: Expte\_Agre\_Mesa es 0, entonces los campos Expte\_Agre\_Nro y Expte\_Agre\_Dig son 0 o nulo. Si el campo: Expte\_Agre\_Mesa no 0, entonces los campos Expte\_Agre\_Nro es distinot de 0 y Expte\_Agre\_Dig tiene un valor que puede ser entre 0 y 9 o V o puede ser nulo.
   2. Por lo cual vamos a eliminar los expedientes agregados a otros expedientes, filtrando solo por el campo Expte\_Agre\_Mesa
2. Se eliminar las columnas: Expte\_Agre\_Mesa, Expte\_Agre\_Nro. Expte\_Agre\_Dig, ya que ahora no tienen valores

#### Fechas

1. Filtrar Fec\_Ini\_Anio: quitar valores menores a 1900
2. Quitar nulos del campo Fec\_Ul\_Mov\_Expte.
3. Luego de aplicar el filtro anterior, se ve que todos los campos de fechas están bien. No hace falta revisarlos
4. Los campos: Fec\_Ini\_Anio, y Fec\_Mov\_Anio, cambiarlos a tipo Texto
5. Combinar columnas: Fec\_Ini\_Anio, Fec\_Ini\_Mes, Fec\_Ini\_Dia para crear Id\_Fecha\_Ini\_Expte. Esto elimina los 3 primeros campos y crea el campo combinado
6. Combinar columnas: Fec\_Mov\_Anio, Fec\_Mov\_Mes, Fec\_Mov\_Dia para crear Id\_Fecha\_Ult\_Mov\_Expte. Esto elimina los 3 primeros campos y crea el campo combinado
7. Cambiar el tipo de dato de las columnas: Id\_Fecha\_Ini\_Expte y Id\_Fecha\_Ult\_Mov\_Expte a Integer
8. Guardar y regresar a Power BI

## Relaciones entre la tabla de hechos y las dimensiones

1. Las relaciones entre la tabla de hechas y sus dimensiones, deben quedar como sigue:



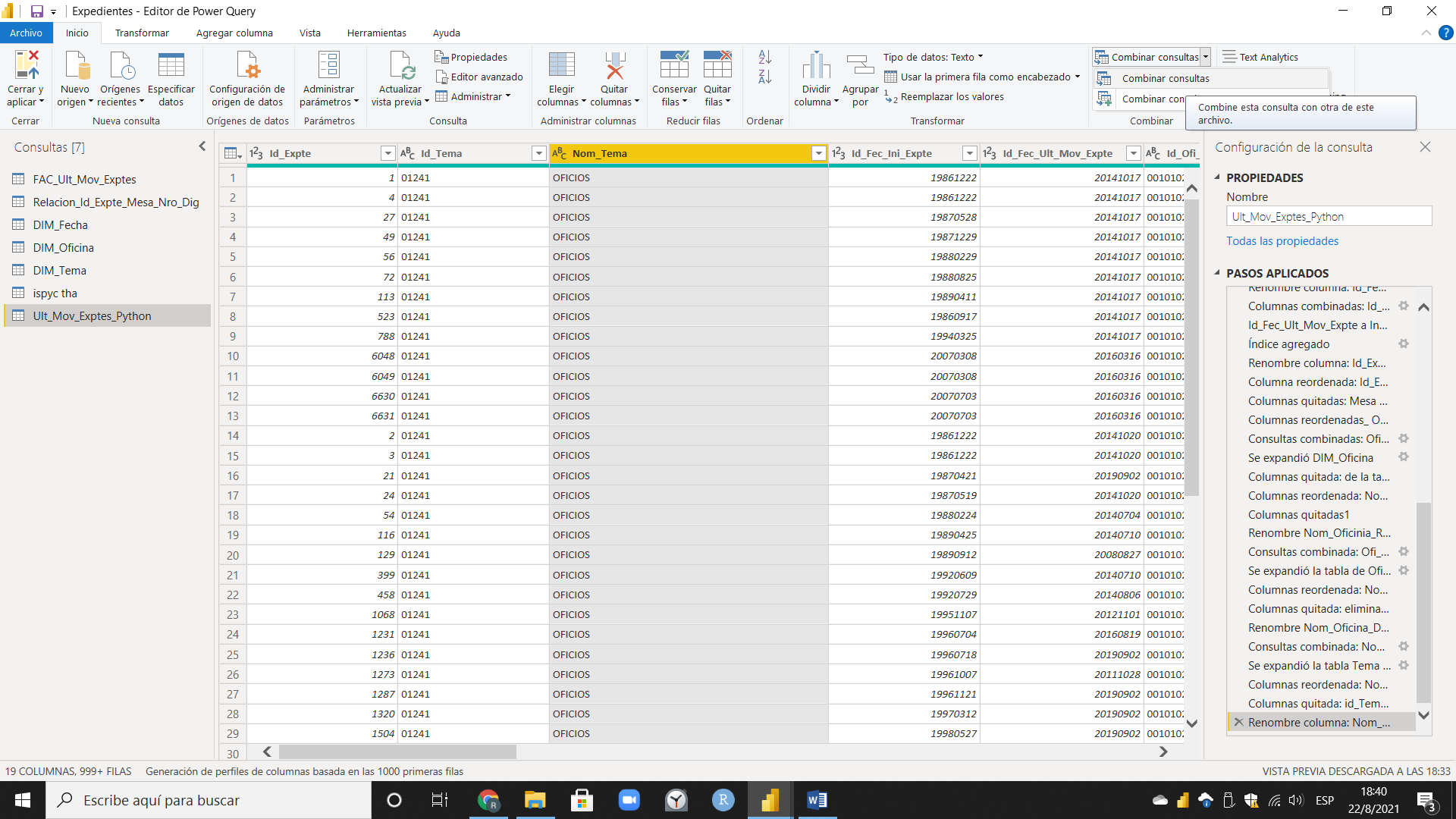
1. Guardar y volver a Power Query

## Siguientes transformaciones de datos

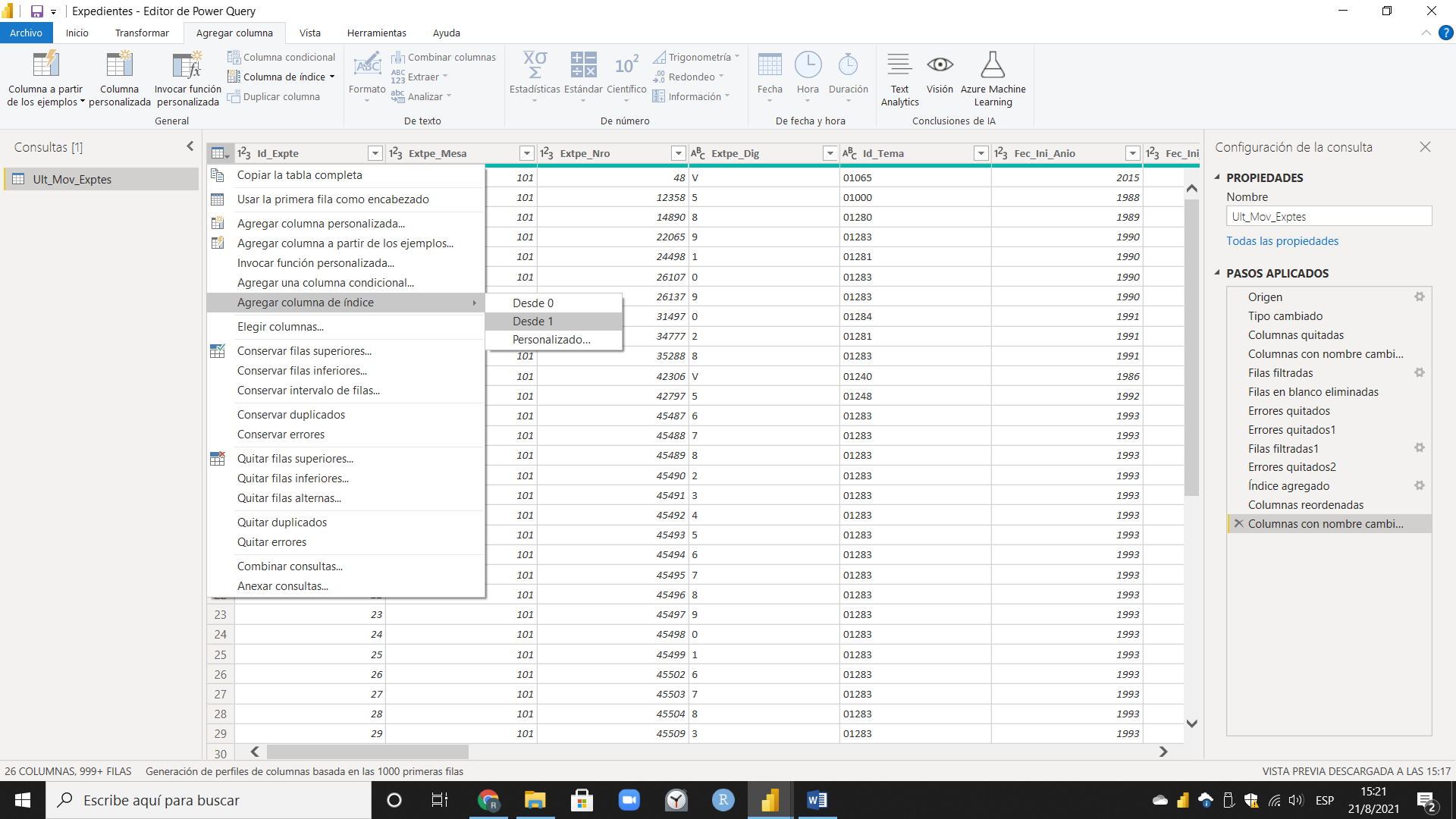
1. En la tabla: FAC\_Ult\_Mov\_Exptes, se debe aplicar el proceso de “Combinar tablas”, para los siguientes campos:

|  |  |  |
| --- | --- | --- |
| **Columna** | **Dimensión y Columna a relacionar** | **Columnas a agregar de la dimensión** |
| Id\_Fecha\_Ini\_Expte | DIM\_Fecha\_Ini\_Expte.Id\_Fecha | DIM\_Fecha\_Ini\_Expte.Fecha |
| Id\_Fecha\_Ult\_Mov\_Expte | DIM\_Fecha\_Ult\_Mov\_Expte.Id\_Fecha | DIM\_Fecha\_Ult\_Mov\_Expte.Fecha |
| Cod\_Oficinia\_Remitente | DIM\_Oficina\_Remitente.Cod\_Oficina | DIM\_Oficina\_Remitente.Id\_Oficina |
| Cod\_Oficinia\_Destino | DIM\_Oficina\_Destino.Cod\_Oficina | DIM\_Oficina\_Destino.Id\_Oficina |
| Cod\_Tema | DIM\_Tema.Cod\_Tema | DIM\_Tema.Id\_Tema |

1. Para cada combinación realizada, se deben filtrar y eliminar las filas que no se pudieron relacionar. Podemos filtrar buscado nulos o 0 para cada columna agregada
2. Eliminar de la tabla FAC\_Ult\_Mov\_Exptes las columnas: DIM\_Fecha\_Ini\_Expte.Fecha, DIM\_Fecha\_Ult\_Mov\_Expte.Fecha, DIM\_Oficina\_Remitente.Cod\_Oficina, DIM\_Oficina\_Destino.Cod\_Oficina y DIM\_Tema.Cod\_Tema
3. Re nombrar las columnas de id recuperadas de las dimensiones, quitando el nombre de la dimensión
4. Nota: Con las fechas, también se debe aplicar “Combinar tablas” y luego se deben buscar las filas con valores nulos porque hay fechas mal cargadas, por ejemplo días 31 de meses que tienen 30 días o día 29 de algún febrero que tuvo 28 días
5. Proceso para combinar tablas en Power Query
6. Estos se hace, posicionándome en la tabla Ult\_Mov\_Exptes\_Python y yendo a la opción “Combinar consulta/Combinar consulta



1. Luego de cada proceso, debemos expandir las campos de la tabla combinadas, seleccionar solo la columna que queremos agregar, renombrarlo y ordenarlas donde corresponda
2. Generar índice automática y renombrarlo Id\_Expte.



1. Crear tabla de relación: “Rel\_Id\_Expte\_Mesa\_Nro\_Dig ” con los campos: Id\_Expte, Extpe\_Mesa, Extpe\_Nro y Extpe\_Dig, duplicando la tabla FAC\_Ult\_Mov\_Exptes
2. Borrar los campos. Extpe\_Mesa, Extpe\_Nro y Extpe\_Dig de la tabla FAC\_Ult\_Mov\_Exptes
3. Nota: Esto se hace para quitar el identificador real de los expedientes, por confidencialidad de los datos
4. La tabla quedará como sigue

|  |  |
| --- | --- |
| **Columna** | **Tipo de dato** |
| Id\_Expte | Integer |
| Expte\_Mesa | String |
| Expte\_Nro | Integer |
| Expte\_Dig | String |
| Id\_Tema | Integer |
| Id\_Fecha\_Ini\_Expte | Integer |
| Id\_Fecha\_Ult\_Mov\_Expte | Integer |
| Id\_Oficina\_Remitente | Integer |
| Id\_Oficina\_Destino | Integer |
| Concluido | Boleano |
| Cant\_Hojas | Integer |

## Armado de Dataset para Python

1. Crear tabla: Ult\_Mov\_Exptes\_Python, duplicando la tabla: FAC\_Ult\_Mov\_Exptes
2. Filtrar por el campo: Fec\_Ini\_Expte mayor a 31/12/2010, para usar solo los exedientes desde el año 2010 en adelante
3. La tabla tendrá los siguientes campos:

|  |  |
| --- | --- |
| **Columna** | **Tipo de dato** |
| Id\_Expte | Integer |
| Id\_Tema | Integer |
| Nom\_Tema | String |
| Id\_Fecha\_Ini\_Expte | Integer |
| Fecha\_Ini\_Expte | Integer |
| Id\_FechaUlt\_Mov\_Expte | Integer |
| Fecha\_Ult\_Mov\_Expte | Date |
| Id\_Oficina\_Remitente | Integer |
| Id\_Tipo\_Org\_Remit | Integer |
| Nom\_Tipo\_Org\_Remit | String |
| Cod\_Org\_Remit | String |
| Nom\_Org\_Remit | String |
| Id\_Oficina\_Destino | Integer |
| Id\_Tipo\_Org\_Destino | Integer |
| Nom\_Tipo\_Org\_Destino | String |
| Cod\_Org\_Destino | String |
| Nom\_Org\_Destino | String |
| Concluido | Boleano |
| Cant:Hojas | Integer |

1. Exportarla a .csv usando DAX Studio
   1. Abrir DAX Studio, sin cerrar el proyecto en Power BI
   2. Seleccionar la primer opción de conexión
   3. En Output seleccionar: Fila
   4. Escribir “EVALUATE Ult\_Mov\_Exptes\_Python” y presionar el ícono de ejecutar. Esto descarga la atabla en un formato .csv
2. Levantar en Python y verificar los datos