**CARGA DE DATOS DEL ALMACÉN DE DATOS PARA EL ANÁLISIS DEL IMPACTO AMBIENTAL Y CONSUMO ENERGÉTICO DERIVADOS DE LA ACTIVIDAD ECONÓMICA**



**EDUARDO MORA GONZÁLEZ**

Tabla de contenido

[**1.** **Identificación de los procesos ETL** 3](#_Toc90024691)

[**1.1.** **Bloque IN (de las fuentes a tablas intermedias)** 3](#_Toc90024692)

[**1.2.** **Bloque TR (poblar las tablas de nuestro almacén)** 4](#_Toc90024693)

[**2.** **Diseño y desarrollo de los procesos ETL** 5](#_Toc90024694)

[**2.1.** **Creación de tablas intermedias (staging area)** 5](#_Toc90024695)

[**2.1.1.** **IN\_OBJECTIVES** 5](#_Toc90024696)

[**2.1.2.** **IN\_INVESTMENTS** 6](#_Toc90024697)

[**2.1.3.** **IN\_COUNTRIES** 6](#_Toc90024698)

[**2.1.4.** **IN\_PROTECTEDAREAS** 7](#_Toc90024699)

[**2.1.5.** **IN\_ ENERGYBALANCES** 7](#_Toc90024700)

[**2.1.6.** **IN\_URBANWASTES** 8](#_Toc90024701)

[**2.2.** **Creación del modelo multidimensional** 8](#_Toc90024702)

[**2.3.** **Base de datos final** 8](#_Toc90024703)

[**2.4.** **Creación del proceso de extracción, transformación y carga** 9](#_Toc90024704)

[**2.4.1.** **Crear repositorio de trabajo** 9](#_Toc90024705)

[**2.4.2.** **Conexión a la base de datos SQL Server** 9](#_Toc90024706)

[**2.4.3.** **Bloque IN** 10](#_Toc90024707)

[**2.4.4.** **Bloque TR** 27](#_Toc90024708)

[**2.4.5.** **Bloque TR\_DIM** 27](#_Toc90024709)

[**2.4.6.** **Bloque TR\_FACT** 37](#_Toc90024710)

[**3.** **Implementación de los trabajos con procesos ETL** 44](#_Toc90024711)

[**3.1.** **JOB\_IN** 44](#_Toc90024712)

[**3.2.** **JOB\_TR\_DIMS** 46](#_Toc90024713)

[**3.3.** **JOB\_TR\_FACTS** 47](#_Toc90024714)

[**3.4.** **JOB\_CARGA\_DW** 48](#_Toc90024715)

# **Identificación de los procesos ETL**

No existe una estrategia única a la hora de diseñar y analizar el proceso de carga de la base de datos. Es muy común construir un proceso ETL basado en las entidades de datos que se van a actualizar, porque existe una diferencia conceptual en la actualización de la tabla de dimensiones en comparación con la tabla de hechos.

Dividir el proceso de carga inicial en diferentes bloques de actualización ayudará al diseño, la secuencia de ejecución y la gestión de dependencias. Cada uno de estos bloques de actualización se dividirá en las correspondientes fases de extracción, transformación y carga.

Se identifican los dos bloques siguientes:

* **Bloque IN**: procesos de carga de los datos desde las fuentes a las tablas intermedias en el área de maniobras (staging area).
* **Bloque TR**: procesos de transformación para cargar los datos desde las tablas intermedias hasta nuestro almacén de datos, según el modelo multidimensional diseñado.

## **Bloque IN (de las fuentes a tablas intermedias)**

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE DEL ETL | DESCRIPCIÓN | ORÍGENES DE LOS DATOS | TABLA DESTINO |
| IN\_OBJECTIVES | Carga de los datos correspondientes a los Objetivos de Desarrollo Sostenible y su relación con los ámbitos medioambientales | ODS.xlsx | STG\_OBJECTIVES  STG\_OBJECTIVESAREAS |
| IN\_INVESTMENTS | Carga de datos correspondientes a la evolución de la inversión en protección ambiental por tipo de equipo e instalación, ámbito medioambiental y sector de actividad económica. | 02002.xlsx | STG\_INVESTMENTS |
| IN\_COUNTRIES | Carga de datos correspondientes a los nombres de los distintos países en orden alfabético y los elementos de código ISO 3166-1-alpha-2 y alpha-3 | Countries.json | STG\_COUNTRIES |
| IN\_PROTECTEDAREAS | Carga de datos correspondientes a las áreas protegidas tanto marinas como terrestres | env\_bio1.tsv | STG\_ PROTECTEDAREAS |
| IN\_ ENERGYBALANCES | Carga de datos correspondientes a todos aquellos aspectos relevantes del balance energético mundial | WorldEnergy  Balances  Highlights\_final.xlsx | STG\_ ENERGYBALANCES |
| IN\_URBANWASTES | Carga de datos correspondientes a generación y tratamiento de residuos urbanos | DataGeneric.xml | STG\_ URBANWASTES |

## **Bloque TR (poblar las tablas de nuestro almacén)**

El bloque «TR\_» de procesos de ETL para poblar el modelo multidimensional del almacén tiene dos partes:

* Los procesos de carga y transformación de las dimensiones.
* Los procesos de las tablas de hechos.

El orden de ejecución es importante para que la carga de datos sea correcta. Las dimensiones se cargarán primero y, después, las tablas de hechos, si no ha habido errores.

Los procesos del bloque de carga y transformación de las dimensiones son los siguientes:

|  |  |  |  |
| --- | --- | --- | --- |
| NOMBRE DEL ETL | DESCRIPCIÓN | TABLA DE ORIGEN | TABLA DESTINO |
| TR\_DIM\_Date | Carga y transformación de la dimensión temporal (fecha en la que se realiza la medición). | SQL | DIM\_Date |
| TR\_DIM\_SDG | Carga y transformación de la dimensión de Objetivos de Desarrollo Sostenible. | STG\_Objectives | DIM\_SDG |
| TR\_DIM\_Country | Carga de la dimensión de países donde se localiza la medición del balance energético. | STG\_Countries | Dim\_Country |
| TR\_DIM\_EconomicActivitySector | Carga y transformación de la dimensión con datos del sector de la actividad económica al que se clasifica la medición. | STG\_Investments | Dim\_EconomicActivitySector |
| TR\_DIM\_TypeEquipmentInstallation | Carga de la dimensión con los tipos de equipamientos o instalación de la medición. | STG\_Investments | Dim\_TypeEquipmentInstallation |
| TR\_DIM\_Product | Carga de la dimensión con los productos que intervienen en la medición del balance energético. | STG\_ EnergyBalances | DIM\_Product |
| TR\_DIM\_Region | Carga de la dimensión de la región se localiza la medición. | STG\_Investments  STG\_ProtectedAreas  STG\_EnergyBalance  STG\_Urbanwastes | DIM\_Region |
| TR\_DIM\_Measurement | Carga de la dimensión con información de las unidades de medida. | STG\_ObjectivesAreas  STG\_Urbanwastes  STG\_Investments  STG\_ProtectedAreas | DIM\_Measurement |

Los procesos del bloque de carga y transformación de las tablas de hechos son:

|  |  |  |
| --- | --- | --- |
| NOMBRE DEL ETL | DESCRIPCIÓN | TABLA DE ORIGEN |
| TR\_FACT\_  EnvironmentalMeasurements | Carga y transformación de la tabla de hechos «FACT\_EnvironmentalMeasurements» | STG\_Investments  STG\_ProtectedAreas  STG\_EnergyBalance  STG\_Urbanwastes |
| TR\_FACT\_  EnergyBalances | Carga y transformación de la tabla de hechos  «FACT\_EnergyBalances» | STG\_ EnergyBalances |

# **Diseño y desarrollo de los procesos ETL**

En este apartado, combinado con las consideraciones anteriores, usaremos Pentaho Data Integration (PDI) para diseñar e implementar los procesos de carga.

Los procesos de ETL que diseñaremos en PDI consistirán en la definición de trabajos y transformaciones.

## **Creación de tablas intermedias (staging area)**

El primer paso para la implementación del proceso de ETL consiste en la creación de las tablas intermedias en la staging area. Esta se llevará a cabo una única vez, mediante scripts sobre la base de datos, en nuestro caso SQL Server.

Las tablas intermedias se utilizarán en los procesos IN, que permitirán cargar los datos desde las fuentes de datos.

### **IN\_OBJECTIVES**

Inicialmente añadimos los objetivos ODS:

CREATE TABLE [dbo].[STG\_Objectives](

[Objetivo] [float] NULL,

[Nombre] [varchar](100) NULL,

[Descripción] [varchar](512) NULL

) ON [PRIMARY]

Una vez añadido los objetivos se añaden los ámbitos a áreas que abarcan estos objetivos:

CREATE TABLE [dbo].[STG\_ObjectivesAreas](

[Codigo] [varchar](100) NULL,

[Ambito/VAR/Flow] [varchar](512) NULL,

[ODS principal] [float] NULL

) ON [PRIMARY]

Con estas dos tablas ya se ha completado la estructura donde se va a cargar las dos pestañas del fichero ODS.xlsx

### **IN\_INVESTMENTS**

CREATE TABLE [dbo].[STG\_Investments](

[Periodo] [int] NOT NULL,

[Inversion] [float] NULL,

[sector\_economico] [varchar](100) NULL,

[tipo\_instalacion] [varchar](100) NULL,

[ambito\_medioambiental] [varchar](100) NULL,

[comunidad\_autonoma] [varchar](27) NULL

) ON [PRIMARY]

Con esta tabla ya se ha completado la estructura donde se va a cargar el fichero 02002.xlsx

### **IN\_COUNTRIES**

CREATE TABLE [dbo].[STG\_Countries](

[Nombre] [varchar](100) NOT NULL,

[Name] [varchar](100) NOT NULL,

[Nom] [varchar](100) NOT NULL,

[ISO2] [varchar](2) NULL,

[ISO3] [varchar](3) NULL,

[phone\_code] [varchar](100) NULL

) ON [PRIMARY]

A nivel de seguridad en la carga, en la ISO2 e ISO3 se le ha puesto el tamaño máximo que puede tener ambas.

Con esta tabla ya se ha completado la estructura donde se va a cargar el fichero Countries.json

### **IN\_PROTECTEDAREAS**

CREATE TABLE [dbo].[STG\_ProtectedAreas](

[Year] [varchar](4) NULL,

[areaprot] [varchar](100) NULL,

[value] [bigint] NULL,

[geoTime] [varchar](2) NULL

) ON [PRIMARY]

Con esta tabla ya se ha completado la estructura donde se va a cargar el fichero env\_bio1.tsv

### **IN\_ ENERGYBALANCES**

Esta tabla ya ha sido cargada en la BBDD en la PEC2, la sentencia SQL usada fue:

CREATE TABLE [dbo].[STG\_EnergyBalance](

[country] [varchar] (255) NULL,

[product] [varchar] (255) NULL,

[flow] [varchar] (255) NULL,

[year] [int] NULL,

[value] [float] NULL

) ON [PRIMARY]

Con esta tabla ya se ha completado la estructura donde se va a cargar el fichero WorldEnergyBalancesHighlights\_final.xlsx

### **IN\_URBANWASTES**

CREATE TABLE [dbo].[STG\_Urbanwastes](

[Cou] [varchar](100) NULL,

[Var] [varchar](100) NULL,

[Time\_Format] [varchar](100) NULL,

[Unit] [varchar](100) NULL,

[PowerCode] [int] NULL,

[OBS\_Status] [varchar](100) NULL,

[Time] [int] NULL,

[ObsValue] [float] NULL

) ON [PRIMARY]

Con esta tabla ya se ha completado la estructura donde se va a cargar el fichero DataGeneric.xml

## **Creación del modelo multidimensional**

Las tablas relativas a las Dimensiones y las Tablas de hechos se han añadido gracias al script proporcionado en el enunciado de la práctica.

## **Base de datos final**

Finalmente, y para comprobar que se han añadido todas las tablas a la BBDD, se mostrará una captura en donde aparece el nombre de la BBDD y todas las tablas:

Tabla

Descripción generada automáticamente con confianza bajaTabla

Descripción generada automáticamente con confianza baja

## **Creación del proceso de extracción, transformación y carga**

Una vez creado el modelo físico del almacén, se va a pasar a crear y diseñar los procesos ETL que permitirán poblar las tablas intermedias del área intermedia (staging area) y las tablas de dimensiones y de hechos del Data Mart.

### **Crear repositorio de trabajo**

La primera cosa que se debe realizar es crear un nuevo repositorio en donde se va a almacenar todas las transformaciones y trabajos de nuestro almacén de datos.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Como se puede comprobar en la imagen, se ha sincronizado el repositorio y se muestra la ruta donde estarán todos los elementos necesarios.

### **Conexión a la base de datos SQL Server**

Otro paso previo que se debe realizar es crear las conexiones a las bases de datos que se usan en todas las transformaciones y trabajos de los procesos de carga.

Para facilitar las conexiones, se van a crear dos conexiones diferentes, una para la base de datos del modelo multidimensional y otra para el área intermedia.

En la creación de la conexión al «STAGE», el nombre usado es «cn\_stage»:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

En la creación de la conexión al «DW», el nombre usado es «cn\_dw»:

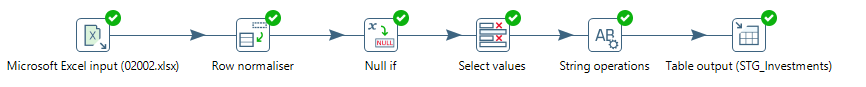
Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

### **Bloque IN**

#### Transformación IN\_INVESTMENTS

La carga de la fuente 02002.xlsx era una de las que se adjuntaban junto el enunciado, pero se hará un análisis para ver cómo funciona y explicar el paso que se ha añadido. La transformación completa es la siguiente:



Como se puede ver, consta de 6 pasos:

* **Carga del fichero**: en donde se selecciona el fichero a cargar:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Y como se deben cargar los campos:

Tabla

Descripción generada automáticamente con confianza media

* **Normalización de filas:** se normalizan las filas relativas a las comunidades autónomas:

Tabla

Descripción generada automáticamente con confianza media

* **Tratamientos de nulos:** como se ha mencionado antes, hay comunidades que no han realizado alguna inversión en un sector o tipo, por eso en el fichero fuente aparecen “..”, y ahora se consideran como nulos:

Tabla

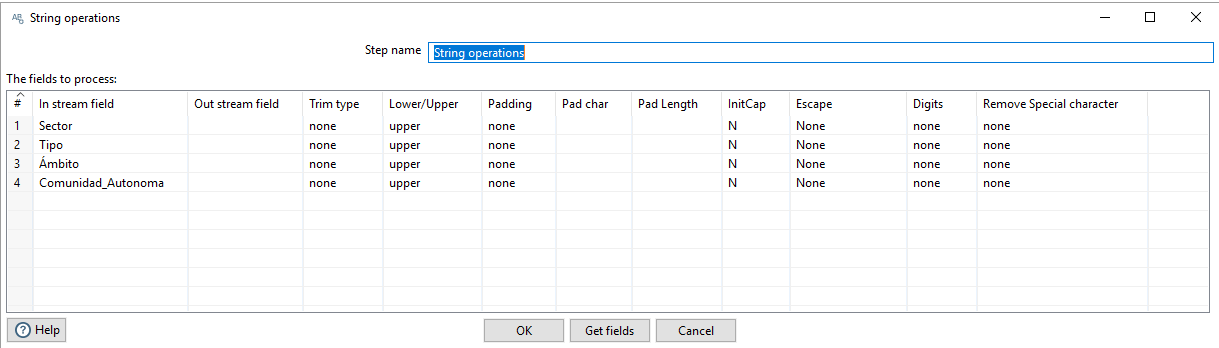
Descripción generada automáticamente

* **Selección de valores:** antes de añadir los valores a la BBDD se formalizan y se preparan:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

* **Operaciones con cadenas:** El siguiente paso de la transformación es asegurar la calidad de los datos mediante la normalización de los valores de los campos de tipo String que se consideren necesarios, es este caso es poner todas las cadenas en mayúsculas:



* **Carga en la BBDD:** una vez terminado todas las transformaciones, se añaden a la BBDD, para ello se usa la conexión de nombre <<cn\_stage>>:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que todo ha funcionado de manera correcta, se mostraran las estadísticas una vez completado todo el proceso:

Captura de pantalla de computadora

Descripción generada automáticamente

#### Transformación IN\_ ENERGYBALANCES

Esta carga se realizó en la PCE2, pero al igual que en la transformación anterior se ha modificado para obtener todos los campos String en mayúsculas.

Como es un paso bastante sencillo, se va a mostrar solamente las estadísticas del proceso terminado:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

#### Transformación IN\_OBJECTIVES

Como en el modelo de la BBDD, el fichero ODS.xlsx tiene dos pestañas distintas, por lo que se debe realizar dos transformaciones distintas.

##### STG\_OBJECTIVES

La transformación completa es la siguiente:

Diagrama

Descripción generada automáticamente

Como se puede observar en la imagen anterior, esta transformación contiene cuatro pasos:

* **Lectura del fichero XLSX:** donde seleccionamos el fichero:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

La hoja:

Imagen que contiene Aplicación

Descripción generada automáticamente

y la forma de importar los campos:

Imagen que contiene Tabla

Descripción generada automáticamente

Para comprobar que se han cargado de manera correcta, se hará una visualización previa de los datos:

Texto

Descripción generada automáticamente

Como se han cargado de una manera correcta, podemos pasar al siguiente paso.

* **Operaciones con cadenas:** El siguiente paso de la transformación es asegurar la calidad de los datos mediante la normalización de los valores de los campos de tipo String que se consideren necesarios, que en este caso es poner todos los caracteres en mayúsculas:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

* **Ordenación de las filas**: A continuación, hay que ordenar de manera ascendente el campo objetivo:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* **Carga a la tabla intermedia «STG\_Objectives»:** El paso de cargar los datos a la tabla intermedia del stage:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

##### STG\_OBJECTIVESAREAS

La transformación completa es la siguiente:

Escala de tiempo

Descripción generada automáticamente con confianza media

Como se puede observar la transformación contiene los siguientes pasos:

* **Lectura del fichero XLSX:** donde seleccionamos el fichero:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

La hoja:

Una captura de pantalla de una red social

Descripción generada automáticamente

Y la forma de importar los campos:

Aplicación

Descripción generada automáticamente con confianza baja

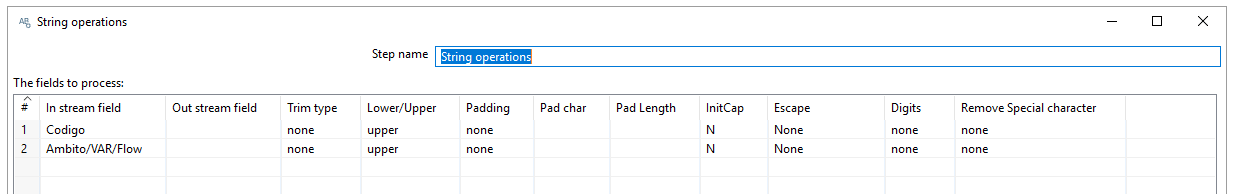
Para comprobar que se han cargado de manera correcta, se hará una visualización previa de los datos:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Como se han cargado de una manera correcta, podemos pasar al siguiente paso.

* **Operaciones con cadenas:** El siguiente paso de la transformación es asegurar la calidad de los datos mediante la normalización de los valores de los campos de tipo String que se consideren necesarios, que será poner los campos String en mayúsculas:

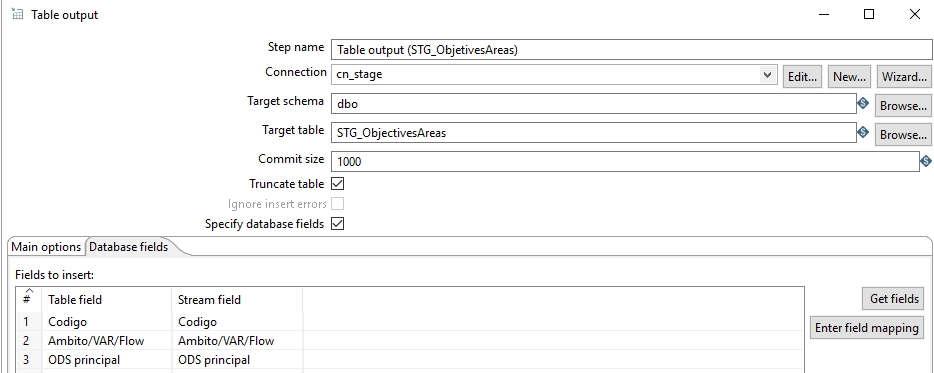


* **Ordenación de las filas**: A continuación, hay que ordenar de manera ascendente el campo referente al ODS principal:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* **Carga a la tabla intermedia STG\_ObjectivesAreas:** El paso de cargar los datos a la tabla intermedia del stage:



Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

#### Transformación IN\_COUNTRIES

La transformación completa es la siguiente:

Diagrama

Descripción generada automáticamente

La transformación consta de los siguientes pasos:

* **Lectura del fichero JSON:** donde seleccionamos el fichero:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

y los campos:

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Para comprobar que se han cargado de manera correcta, se hará una visualización previa de los datos:

Tabla

Descripción generada automáticamente

Como se han cargado de una manera correcta, podemos pasar al siguiente paso.

* **Operaciones con cadenas:** El siguiente paso de la transformación es asegurar la calidad de los datos mediante la normalización de los valores de los campos de tipo String que se consideren necesarios, que en este caso es poner todos los campos en mayúsculas:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente con confianza media

* **Carga a la tabla intermedia «STG\_Countries»:** Como el fichero ya venía ordenado alfabéticamente, procedemos a insertar los datos en la tabla intermedia del stage:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

#### Transformación IN\_PROTECTEDAREAS

La transformación completa es la siguiente:

Diagrama

Descripción generada automáticamente

La transformación consta de los siguientes pasos:

* **Lectura del fichero tsv:** donde seleccionamos el fichero y los delimitadores (un espacio en blanco):

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Para comprobar que se han cargado de manera correcta, se hará una visualización previa de los datos:

Tabla

Descripción generada automáticamente

Aunque se han cargado los datos hay que hacer ciertas transformaciones para obtener los datos de manera correcta.

* **Dividir campos delimitados por comas:** el campo areaprot,geo\time se dividirá en dos variables separadas:

Interfaz de usuario gráfica, Aplicación

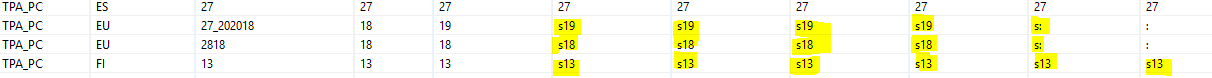
Descripción generada automáticamente

Como salida de esta transformación, vemos que tenemos ahora dos campos separados:

Tabla

Descripción generada automáticamente

* **Eliminar caracteres mal ingresados:** hay algunos casos en donde había una “s” después del número, y se ha añadido de la siguiente forma:



Además, se han añadido valores numéricos en el campo *GeoTime*:

Tabla

Descripción generada automáticamente

Haciendo un análisis del fichero, nos damos cuenta de que siempre son los mismos valores “27\_2020” y “28 “.

Para todos estos casos se hará las siguientes operaciones:

* + Para los campos referente a los años solamente se permitirán caracteres numéricos.
  + Para el campo *GeoTime* solamente nos quedamos con los 2 primeros dígitos.

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

* **Seleccionar los campos y el formato**: se le cambiaran el formato a los campos para que sean los correctos antes de añadirlos a la BBDD:

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

* **Normalización de columnas:** ahora que tenemos todos los datos de una manera correcta, vamos a normalizar los años y sus valores de la siguiente manera:

Tabla

Descripción generada automáticamente

Gracias a esta operación los campos vacíos se han añadido como nulos:

Tabla

Descripción generada automáticamente

* **Carga a la tabla intermedia «STG\_PROTECTEDAREAS»:** procedemos a insertar los datos en la tabla intermedia del stage:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

#### Transformación IN\_URBANWASTES

La transformación completa es la siguiente:

Escala de tiempo

Descripción generada automáticamente con confianza media

La transformación consta de los siguientes pasos:

* **Lectura del fichero XML:** donde seleccionamos el fichero a cargar:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Seleccionamos el *Loop Xpath* para saber dónde iterar:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Seleccionamos los campos a extraer:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Para comprobar que se han cargado de manera correcta, se hará una visualización previa de los datos:

Tabla

Descripción generada automáticamente

Aunque se han cargado los datos hay que hacer ciertas transformaciones para obtener los datos de manera correcta.

* **Operaciones con cadenas:** El siguiente paso de la transformación es asegurar la calidad de los datos mediante la normalización de los valores de los campos de tipo String que se consideren necesarios:

Tabla

Descripción generada automáticamente con confianza baja

* **Reemplazar los caracteres no numéricos:** para preparar los datos antes de transformar el tipo:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

* **Cambio de formato:** transformaos el campo *ObsValue* de String a numérico.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

* **Ordenación de las filas**: A continuación, hay que ordenar de manera ascendente los campos:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

* **Carga a la tabla intermedia «STG\_URBANWASTES»:** procedemos a insertar los datos en la tabla intermedia del stage:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Captura de pantalla de computadora

Descripción generada automáticamente

### **Bloque TR**

El bloque TR contiene los procesos de ETL, que se encargan de la carga inicial de datos, desde las tablas intermedias pobladas con los procesos del bloque IN, al modelo multidimensional del almacén, compuesto por dimensiones y tablas de hechos.

Este bloque se divide, a su vez, en dos bloques: por un lado, los procesos para la carga de dimensiones y, por el otro, los procesos para la carga de tablas de hechos.

### **Bloque TR\_DIM**

Las dimensiones:

* TR\_DIM\_Country
* TR\_DIM\_EconomicActivitySector
* TR\_DIM\_TypeEquipmentInstallation

Venían incluidas entre los ficheros del enunciado, por lo que se obviaran su diseño, tratando solamente las nuevas dimensiones.

#### TR\_DIM\_Date

La carga de esta dimensión es diferente a la carga de las dimensiones de datos, la opción elegida es a través de un script SQL para generar todos los registros necesarios.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

El script SQL completo es el siguiente:

-- Declaración y establecimiento de fecha inicio y fecha fin (fecha actual)

DECLARE @FechaInicio datetime

DECLARE @FechaFin datetime

SET @FechaInicio = '01/01/1970'

SET @FechaFin = GETDATE()

-- Declaración y establecimiento de fecha ciclo

DECLARE @FechaCiclo datetime

SET @FechaCiclo = @FechaInicio

-- Bucle hasta fecha fin

WHILE @FechaCiclo <= @FechaFin

BEGIN

-- Insertar un registro en la dimensión Fecha

INSERT INTO DIM\_Date VALUES (

cast(cast(Year(@FechaCiclo) as

varchar(4))+right('0'+cast(Month(@FechaCiclo) as

varchar(2)),2)+right('0'+cast(Day(@FechaCiclo) as varchar(2)),2)

as int),

Year(@FechaCiclo),

Month(@FechaCiclo),

Day(@FechaCiclo),

@FechaCiclo

)

-- Incrementar la FechaCiclo en un día

SET @FechaCiclo = DateAdd(d, 1, @FechaCiclo)

END

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

#### TR\_DIM\_SDG

La transformación completa es la siguiente:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Esta transformación consta de los siguientes pasos:

* **Obtención de los valores:** Mediante una sentencia «SELECT» de SQL (donde se trata el diseño de los registros NA)

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* **Carga los datos:** a la tabla de dimensión DIM\_SDG del modelo dimensional en el DataMart:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

Notad que no se añade la secuencia, ya que el código de SDG será el identificador.

#### TR\_DIM\_Region

La transformación completa es la siguiente:

Diagrama

Descripción generada automáticamente

Esta transformación consta de los siguientes pasos:

* **Obtención de los valores:** Mediante una sentencia «SELECT» de SQL en cada una de las 4 tablas en donde se especifican los países o regiones. La sentencia es la siguiente:

-- TABLA STG\_Investments

SELECT DISTINCT

t1.comunidad\_autonoma as region,

t2.country\_code as country\_code2,

t2.country\_code3 as country\_code3,

t2.country\_name\_en as country\_name

FROM [dbo].[STG\_Investments] t1, [dbo].DIM\_Country t2

WHERE t2.country\_name\_en = 'Spain'

UNION

SELECT DISTINCT -- TABLA STG\_ProtectedAreas

'NA' as region,

t1.geoTime as country\_code2,

t2.country\_code3 as country\_code3,

t2.country\_name\_en as country\_name

FROM [dbo].STG\_ProtectedAreas t1, [dbo].DIM\_Country t2

WHERE t2.country\_code = t1.geoTime

UNION

SELECT DISTINCT -- TABLA STG\_EnergyBalance

'NA' as region,

t2.country\_code as country\_code2,

t2.country\_code3 as country\_code3,

t2.country\_name\_en as country\_name

FROM [dbo].STG\_EnergyBalance t1, [dbo].DIM\_Country t2

WHERE t2.country\_name\_en = t1.country

UNION

SELECT DISTINCT -- TABLA STG\_Urbanwastes

'NA' as region,

t2.country\_code as country\_code2,

t1.Cou as country\_code3,

t2.country\_name\_en as country\_name

FROM [dbo].STG\_Urbanwastes t1, [dbo].DIM\_Country t2

WHERE t2.country\_code3 = t1.Cou

UNION

SELECT -- NULOS

'NA' as region,

'NA' as country\_code2,

'NA' as country\_code3,

'NA' as country\_name

Como se puede observar, solamente en la tabla *STG\_Investments* se especifica la comunidad autónoma, en el resto solamente el país, por eso ese valor se pone a nulo en el resto de los casos. Además, solamente se añadirán los distintos valores así evitando añadir repetidos.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* **Creación de una secuencia:** que hará las funciones de clave primaria de incremento automático.
* **Carga los datos:** a la tabla de dimensión DIM\_Region del modelo dimensional en el DataMart. Se ha modificado el campo Región para que permita más caracteres. La sentencia usada ha sido la siguiente:

ALTER TABLE dbo.DIM\_Region ALTER COLUMN region VARCHAR(27)

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Aplicación

Descripción generada automáticamente con confianza media

#### TR\_DIM\_Measurement

La transformación completa es la siguiente:

Diagrama

Descripción generada automáticamente con confianza media

En esta transformación cargaremos los datos de distintas tablas, y siempre comparándolo con las distintas áreas de los objetivos SDG.

* **Obtención de los valores:** Mediante una sentencia «SELECT» de SQL

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

La sentencia SQL es la siguiente:

SELECT DISTINCT

t2.Codigo as measurement\_code,

t2.[Ambito/VAR/Flow] as measurement\_name,

t1.areaprot as Unit,

t2.[ODS principal] as fk\_sdg

FROM dbo.STG\_ProtectedAreas t1, dbo.STG\_ObjectivesAreas t2

WHERE t1.areaprot = t2.Codigo

UNION

SELECT DISTINCT

t1.Codigo as measurement\_code,

t1.[Ambito/VAR/Flow] as measurement\_name,

t2.Unit as Unit,

t1.[ODS principal] as fk\_sdg

FROM dbo.STG\_ObjectivesAreas t1, dbo.STG\_Urbanwastes t2

WHERE t1.Codigo = t2.Var

UNION

SELECT DISTINCT

t1.Codigo as measurement\_code,

t2.tipo\_instalacion as measurement\_name,

'euros' as Unit,

t1.[ODS principal] as fk\_sdg

FROM dbo.STG\_ObjectivesAreas t1, dbo.STG\_Investments t2

WHERE t1.Codigo = t2.ambito\_medioambiental

UNION

SELECT

'NA' as measurement\_code,

'NA' as measurement\_name,

'NA' as Unit,

0 as fk\_sdg --CAMPO NULO DE LA TABLA SDG

ORDER BY fk\_sdg

* **Creación de una secuencia:** que hará las funciones de clave primaria de incremento automático.
* **Carga los datos:** a la tabla de dimensión *DIM\_Measurement* del modelo dimensional en el *DataMart*.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Captura de pantalla de computadora

Descripción generada automáticamente

#### TR\_DIM\_Product

La transformación completa es la siguiente:

Diagrama

Descripción generada automáticamente

Esta transformación consta de los siguientes pasos:

* **Obtención de los valores:** Mediante una sentencia «SELECT» de SQL de los productos.

Texto

Descripción generada automáticamente

* **Creación de una secuencia:** que hará las funciones de clave primaria de incremento automático.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* **Carga los datos:** a la tabla de dimensión DIM\_Product del modelo dimensional en el DataMart.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostrara las estadísticas:

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

### **Bloque TR\_FACT**

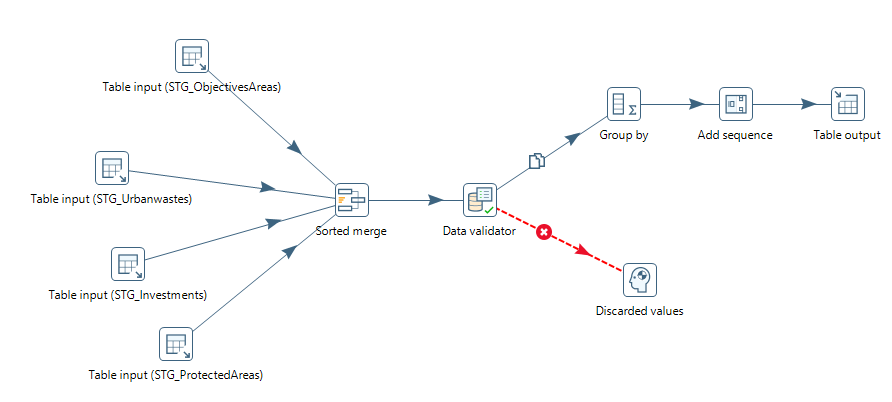
Este bloque contiene las transformaciones para la carga inicial de las tablas de hecho al almacén desde las tablas intermedias «STG\_» del staging area.

La parte principal en la carga de las tablas de hechos es la búsqueda de los valores de las claves foráneas en las tablas de dimensiones cargadas anteriormente.

#### TR\_FACT\_EnvironmentalMeasurements

En esta transformación se integrarán diversas tablas en una única tabla de hechos del almacén de datos.

La transformación completa es la siguiente:



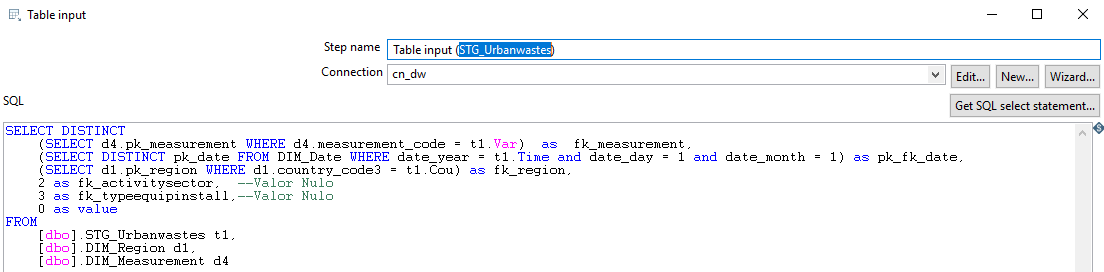
Esta transformación consta de los siguientes pasos:

* **Obtención de los valores:** Mediante sentencias «SELECT» de SQL de distintas tablas:
  + STG\_ObjectivesAreas: para los campos que no existen valores en esa tabla, se ha asignado el valor nulo por defecto.

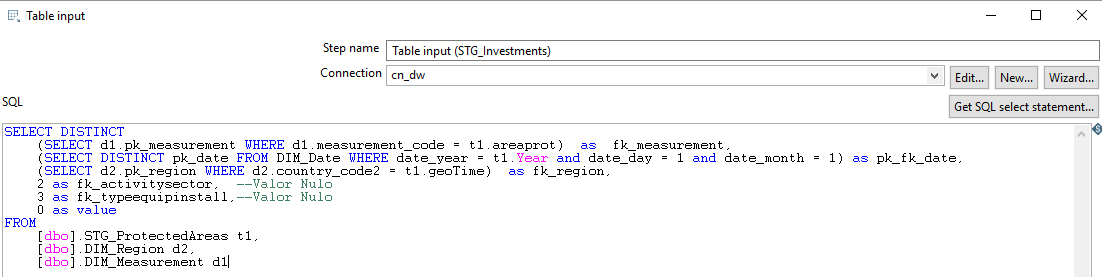
Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

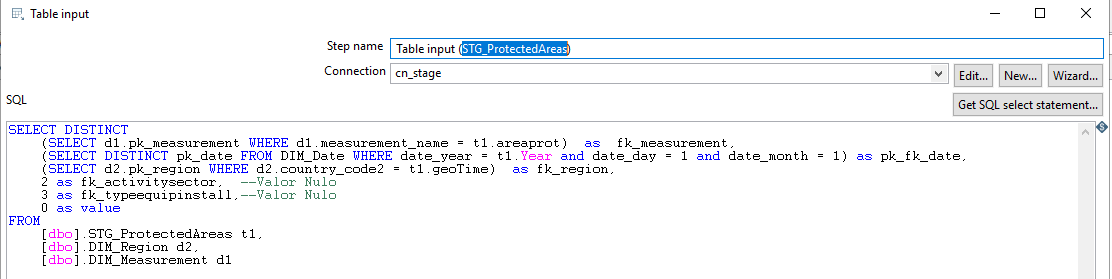
* + STG\_Urbanwastes: para los campos que no existen valores en esa tabla, se ha asignado el valor nulo por defecto.



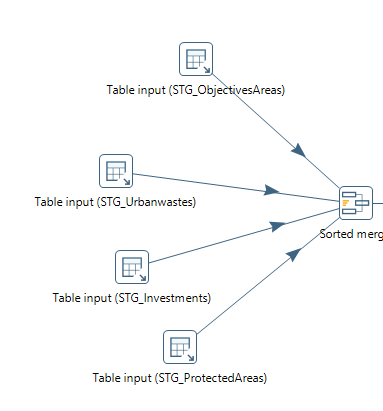
* + STG\_Investments: para los campos que no existen valores en esa tabla, se ha asignado el valor nulo por defecto.



* + STG\_ProtectedAreas: para los campos que no existen valores en esa tabla, se ha asignado el valor nulo por defecto.



* **Unión ordenada:** una vez obtenido los valores, a través de esta funcionalidad unimos los valores de todas las tablas:



* **Validador de datos:** haciendo un análisis exploratorio de los datos obtenidos, veo que hay muchos campos nulos, por lo que he tomado la decisión de quedarme solamente con los datos completos, ya que no hay ninguna regla que especifique que se debe hacer en estos casos, si rellenar los campos o descartarlos.

Las validaciones son por todos los campos y para todos se sigue la misma forma comprobar de que el campo tenga un valor numérico:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Para ver los valores descartados se ha añadido un dummy.

* **Agrupar campos:** la otra opción del validador es la funcionalidad que nos permite agrupar los campos, esto se hace para evitar tener valores repetidos, para ello he decidido agruparlos por todas las claves foráneas y sumar los valores de cada una de ellas:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* **Creación de una secuencia:** que hará las funciones de clave primaria de incremento automático.
* **Carga los datos:** a la tabla de hechos FACT\_EnvironmentalMeasurements:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostraran las estadísticas:

Interfaz de usuario gráfica

Descripción generada automáticamente

Además, se van a mostrar el resultado de la carga de datos en la BBDD:

Tabla

Descripción generada automáticamente

#### TR\_FACT\_EnergyBalances

En esta transformación se integrarán diversas tablas en una única tabla de hechos del almacén de datos.

La transformación completa es la siguiente:

Gráfico de líneas

Descripción generada automáticamente con confianza media

Esta transformación consta de los siguientes pasos:

* **Obtención de los valores:** Mediante sentencias «SELECT» de SQL de distintas tablas, al tratar solamente de una tabla IN, la consulta queda mas sencilla:

Texto

Descripción generada automáticamente

* **Validador de datos:** haciendo un análisis exploratorio de los datos obtenidos, veo que hay muchos campos nulos, por lo que he tomado la decisión de quedarme solamente con los datos completos, ya que no hay ninguna regla que especifique que se debe hacer en estos casos, si rellenar los campos o descartarlos.

Las validaciones son por todos los campos y para todos se sigue la misma forma comprobar de que el campo tenga un valor numérico:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Para ver los valores descartados se ha añadido un dummy como en el caso anterior.

* **Ordenar los datos:** de manera ascendente todos los campos:

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

* **Agrupar campos:** la otra opción del validador es la funcionalidad que nos permite agrupar los campos, esto se hace para evitar tener valores repetidos, para ello he decidido agruparlos por todas las claves foráneas y sumar los valores de cada una de ellas:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* **Carga los datos:** a la tabla de hechos:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para comprobar que se funciona de manera correcta, se mostraran las estadísticas:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Además, se van a mostrar el resultado de la carga de datos en la BBDD:

Tabla

Descripción generada automáticamente

# **Implementación de los trabajos con procesos ETL**

Habrá que tener en cuenta los siguientes bloques de procesos implementados:

* Bloque «IN\_»: procesos de ETL de transformación y carga al área intermedia.
* Bloque «TR\_DIM»: procesos de ETL de transformación y carga de dimensiones.
* Bloque «TR\_FACT»: procesos de ETL de transformación y carga de hechos.

### **JOB\_IN**

El trabajo (job) «JOB\_IN» procesa todas las transformaciones del bloque «IN\_» para la carga de datos desde las fuentes de datos proporcionadas al área intermedia (staging area).

El diseño completo del trabajo (job) «JOB\_IN» es el siguiente:

Escala de tiempo

Descripción generada automáticamente

Los pasos incluidos en el trabajo «JOB\_IN» son:

* Inicio del job.
* Configuración de las variables de entorno.
* Ejecución de las transformaciones «IN\_» de carga del staging area.
* Finalización del job.

El resultado de la ejecución de la transformación completa es el siguiente:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Se observa el procesamiento con éxito de todos los pasos del «JOB\_IN» correspondientes a la ejecución de todas las transformaciones que están incluidas en el trabajo.

### **JOB\_TR\_DIMS**

El trabajo (job) «JOB\_TR\_DIMS» procesa todas las transformaciones del bloque «TR\_DIMS» para la carga de datos, desde las tablas intermedias hasta las tablas de dimensiones del almacén.

El diseño completo del trabajo (job) «JOB\_TR\_DIMS» es el siguiente:Diagrama, Escala de tiempo

Descripción generada automáticamente

Los pasos incluidos en el trabajo «JOB\_TR\_DIMS» son:

* Inicio del job.
* Carga de variables de entorno (path de orígenes de datos y conexiones).
* Borrado de todas las tablas. Esto permite la recarga inicial en caso de ser necesario. Aunque es importante respetar el orden de borrado, según las relaciones definidas entre tablas, en este caso en particular, dado que no hay relaciones entre tablas de dimensión, no influirá el orden.
* Ejecución secuencial de todas las transformaciones «TR\_DIM» (extracción, transformación y carga de dimensiones).
* Finalización del job.

El resultado de la ejecución de la transformación completa es el siguiente:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Se observa el procesamiento con éxito de todos los pasos del «JOB\_TR\_DIMS», correspondientes a la ejecución de todas las transformaciones que están incluidas en el trabajo.

### **JOB\_TR\_FACTS**

El trabajo (job) «JOB\_TR\_FACTS» procesa todas las transformaciones del bloque «TR\_FACT» para la carga de datos desde las tablas intermedias a las tablas de hechos del almacén.

El diseño completo del trabajo (job) «JOB\_TR\_FACTS» es el siguiente:

Escala de tiempo

Descripción generada automáticamente

Los pasos incluidos en el trabajo «JOB\_TR\_FACTS» son:

* Inicio del job.
* Carga de variables de entorno.
* Eliminación de tablas.
* Ejecución de las transformaciones «TR\_FACT».
* Finalización del job.

El resultado de la ejecución de la transformación completa es el siguiente:

Captura de pantalla de computadora

Descripción generada automáticamente

### **JOB\_CARGA\_DW**

El trabajo «JOB\_CARGA\_DW» orquesta todos los trabajos anteriores en un único proceso.

El diseño completo del trabajo «JOB\_ CARGA\_DW» es el siguiente:Imagen que contiene Diagrama

Descripción generada automáticamente

Los pasos incluidos en el trabajo «JOB\_ CARGA\_DW» son:

* Inicio del job.
* Ejecución orquestada de los jobs de carga de todas las transformaciones.
* Finalización del job.

El resultado de la ejecución de la transformación completa es el siguiente:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Se observa el procesamiento con éxito de todos los pasos del «JOB\_CARGA\_DW», correspondientes a la ejecución de todas las transformaciones que están incluidas en el trabajo.

El tiempo total de la carga inicial del data Warehouse es de aproximadamente de 3 minutos y medio:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente