

Solución PRA2- Carga de datos

Caso Práctico: Sistema Integrado de Egresados Universitarios

En esta parte del caso práctico, se realiza el diseño y desarrollo de los procesos de carga del almacén de datos diseñado en el apartado anterior, así como su carga efectiva.

A partir de las fuentes de datos proporcionadas, se deberán identificar los procesos necesarios para la extracción, transformación y carga de los datos. A continuación, se describirán las acciones a llevar a cabo en cada uno de los procesos con su implementación correspondiente, mediante las herramientas de diseño proporcionadas. Finalmente se realizará la carga de datos.

Los apartados que se han desarrollado en esta solución son los siguientes:

1. Revisión del diseño conceptual, lógico y físico del modelo multidimensional.
2. Identificación de los procesos necesarios para la extracción, transformación y carga de datos (ETL's).
3. Diseño de los procesos ETL's.
4. Implementación de trabajos (*jobs*) de procesos ETL's.

1) Revisión del diseño conceptual, lógico y físico.

En este apartado se deberían incorporar aquellas diferencias detectadas al comparar nuestro modelo de *data warehouse* con el que se propone en la solución oficial del caso.

Por ejemplo: si nos falta alguna dimensión; si no hemos identificado correctamente la granularidad de la tabla de hechos; si optamos por alguna dimensión degenerada; si no hemos definido las claves foráneas...

Para esta implementación no se ha modificado el diseño multidimensional propuesto en la práctica anterior.

2) Identificación de los procesos ETL's.

A la hora de diseñar los procesos de carga de un DW, no hay una única estrategia que sirva para todos los casos, pero es habitual estructurar los tratamientos en diferentes bloques para garantizar la integridad en la carga de datos y facilitar la ejecución de los mismos.

En nuestro caso, cada uno de estos bloques se corresponde con una transformación de *Pentaho Data Integration* (PDI, en adelante). En nuestro caso identificamos dos bloques y utilizaremos un prefijo en el nombre para identificarlos:

- **Bloque IN:** procesos de carga de los datos desde las fuentes a tablas intermedias en el área intermedia (*staging area*). Estos procesos se distinguen por el prefijo: IN_ en el nombre.
- **Bloque TR:** procesos de transformación para la carga de datos desde tablas intermedias a nuestro almacén según el modelo multidimensional diseñado. Se diferencian los procesos ETL de transformación para la carga de dimensiones, de los procesos de transformación para la carga de las tablas de hecho. Estos procesos se distinguen con el prefijo TR_ en el nombre.

Esta separación permite la ejecución de los bloques de forma consecutiva (lo más habitual) y a la vez, de forma "aislada" en el caso de que ser requiera reprocesar alguno de los bloques por cualquier incidencia se pueda producir en la ejecución. Así mismo, la evolución de los bloques para incorporar nuevos requerimientos, se puede llevar a cabo de forma independiente optimizando el tiempo requerido a largo plazo.

Finalmente, la utilización de bloques de procesos permite que el mantenimiento de las ETLs será más sencilla cuando en un futuro se tengan que modificar. Normalmente, el personal responsable de la creación de las ETLs no coincide con el personal responsable de la ejecución de las mismas.

Consideramos el uso de un área intermedia (*staging area*) como una buena práctica para la extracción de los datos desde las fuentes origen y la carga al modelo multidimensional, las ventajas que se obtienen con su uso entre otras son:

- Facilitar la extracción de datos (con procesos ETL) desde las fuentes de origen realizando un pretratamiento, si es necesario.
- Realizar lo que se conoce como *Data Cleansing* (detección, corrección y eliminación de datos erróneos).
- Mejorar la calidad de los datos.

Veamos a continuación los procesos de los 2 bloques identificados:

A. Bloque IN (de las fuentes a tablas intermedias)

Nombre ETL	Descripción
IN_RAMA	Carga los registros de la fuente de ISCED_2013 a la tabla intermedia IN_RAMA.
IN_SEGR1_N	Carga de los datos de las fuentes SEGR1.csv y SEGR2.csv, correspondientes a egresados en universidades públicas y privadas por curso académico, rama de enseñanza y universidad a la tabla intermedia IN_SEGR_N.

IN_EGR_C16_17	Carga los datos de egresados de grado y máster por ámbito de Enseñanza, Sexo y Grupos de Edad del curso académico 2016-2017 a la tabla intermedia IN_EGR_C16_17.
IN_EIL_03003	Carga los datos de titulados universitarios según su situación laboral en 2014 por sexo, universidad y rama de conocimiento del curso académico 2009-2010 a la tabla IN_EIL_03003.
IN_EGR_EUR	Carga de los datos de las fuentes edat_lfse_03.xls y educ_uee_grad01.xls, correspondientes a egresados universitarios de diferentes países europeos a la tabla intermedia IN_EGR_EUR.

B. Bloque TR (poblar las tablas de nuestro almacén)

El bloque TR_ de procesos ETL para poblar el modelo multidimensional del almacén, tiene dos partes diferenciadas. Los procesos de carga y transformación de las dimensiones y los de 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 sino ha habido errores.

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

Nombre ETL	Descripción
TR_DIM_ANIO	Carga y transformación de la dimensión año (DIM_ANIO).
TR_DIM_MODALIDAD	Carga y Transformación de la dimensión modalidad de impartición (DIM_MODALIDAD).
TR_DIM_TIPO_UNIV	Carga y transformación de la dimensión tipo de universidades (DIM_TIPO_UNIV).
TR_DIM_UNIVERSIDAD	Carga y transformación de la dimensión universidad (DIM_UNIVERSIDAD).
TR_DIM_RAMA	Carga y transformación de la dimensión ramas de enseñanza (DIM_RAMA).
TR_DESCONOCIDOS	Carga de valores desconocidos a la tabla de dimensión DIM_RAMA.
TR_DIM_PERFIL	Carga y transformación de las dimensiones sexo (DIM_SEXO intervalos de edad (DIM_EDAD)).

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

Nombre ETL	Descripción
TR_FACT_PEGR_EVOLUTIVO	Carga y transformación de la tabla de hechos FACT_PEGR_EVOLUTIVO.
TR_FACT_PEGR_PERFIL	Carga y transformación de la tabla de hechos FACT_PEGR_PERFIL.
TR_FACT_PEGR_INSERTADAS	Carga y transformación de la tabla de hechos FACT_PEGR_INSERTADAS

TR_FACT_PEGR_EUR	Carga y transformación de la tabla de hechos FACT_PEGR_EUR
------------------	--

Existen otras estrategias válidas que nos permitirán cargar los datos, ya sea organizando los procesos de otra forma o fusionándose en un único proceso que lleve a cabo todas las tareas. La opción de una única ETL, podría ser de aplicación en nuestro caso, aunque no es recomendable en cargas más complejas y cambiantes.

En el diseño de los procesos ETL se han tenido en cuenta las siguientes consideraciones:

- Desconocemos la ventana de tiempo disponible (y no es una condición para esta actividad), pero en el contexto de producción, es un factor muy relevante a tener en cuenta en el diseño y ejecución de los procesos para que las cargas no impacten en ningún sistema productivo.
- Son procesos ETL de cargas iniciales con el objetivo de ejecutarse para la carga inicial de datos al almacén. No son procesos ETL para cargas incrementales de datos dado que éstas tienen como el objetivo repetirse de manera periódica.
- No se han incluido los procesos específicos de control de errores, de generación de logs, de metadatos...
- Tal y como se definió en la fase de diseño, usaremos un área intermedia (*staging area*) para cargar los datos desde las fuentes origen antes de poblar el almacén.
- Los procesos podrán ejecutarse tantas veces como sea necesario garantizando la calidad y la integridad de los datos del almacén.

3) Diseño de los procesos ETL's.

En este apartado, y teniendo en cuenta las consideraciones anteriores, vamos a diseñar e implementar los procesos de carga mediante la herramienta de diseño proporcionada: *Pentaho Data Integration* (PDI). Y en particular, el programa de escritorio llamado *Spoon*, que corresponde al entorno gráfico (IDE) de desarrollo de ETL's.

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

Las transformaciones contienen la operativa de bajo nivel, con las acciones con los datos y los trabajos son procesos de alto nivel compuestas por flujos de transformaciones.

A. 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* área. Ésta se llevará a cabo una única vez, mediante scripts sobre la base de datos proporcionada, en nuestro caso, SQL Server. Las tablas intermedias se utilizarán en los procesos IN que permitirán cargar los datos desde las fuentes de datos.

Tabla Intermedia IN_RAMA

```
CREATE TABLE [dbo].[IN_RAMA](
    [COD_RAMA] [varchar](1) NOT NULL,
    [NOM_RAMA] [varchar](50) NOT NULL,
    [COD_RAMA_N2] [varchar](2) NOT NULL,
    [NOM_RAMA_N2] [varchar](100) NOT NULL,
    [COD_RAMA_N3] [varchar](3) NOT NULL,
    [NOM_RAMA_N3] [varchar](100) NOT NULL,
    [COD_RAMA_N4] [varchar](4) NOT NULL,
    [NOM_RAMA_N4] [varchar](150) NOT NULL,
    [COD_RAMA_N5] [varchar](7) NOT NULL,
    [NOM_RAMA_N5] [varchar](200) NOT NULL
) ON [PRIMARY]
```

GO

Tabla Intermedia IN_SEGR_N

```
CREATE TABLE [dbo].[IN_SEGR_N](
    [TIPO_UNIVERSIDAD] [varchar](25) NOT NULL,
    [MODALIDAD] [varchar](50) NOT NULL,
    [UNIVERSIDAD] [varchar](100) NOT NULL,
    [RAMA_ENSEÑANZA] [varchar](100) NOT NULL,
    [CURSO] [varchar](50) NOT NULL,
    [NEGR] [numeric](8, 0) NULL
) ON [PRIMARY]
```

GO

Tabla Intermedia IN_EGR_2016_2017

```
CREATE TABLE [dbo].[IN_EGR_2016_2017](
    [COD_AMBITO] [varchar](200) NOT NULL,
    [SEXO] [varchar](10) NOT NULL,
    [EDAD] [varchar](50) NOT NULL,
    [NUM_EGR_NV1] [numeric](8, 0) NULL,
    [NUM_EGR_NV2] [numeric](8, 0) NULL
) ON [PRIMARY]
```

GO

Tabla Intermedia IN_EIL_03003

```
CREATE TABLE [dbo].[IN_EIL_03003](
    [TIPO_UNIVERSIDAD] [varchar](25) NOT NULL,
    [SEXO] [varchar](20) NOT NULL,
    [RAMA_ENSEÑANZA] [varchar](100) NOT NULL,
    [TRABAJANDO] [numeric](18, 0) NULL,
    [DESEMPLEO] [numeric](18, 0) NULL,
    [INACTIVO] [numeric](18, 0) NULL
) ON [PRIMARY]
```

GO

Tabla Intermedia IN_EGR_EUR

```
CREATE TABLE [dbo].[IN_EGR_EUR](
```

```
[PAIS] [varchar](50) NOT NULL,  
[ANIO] [numeric](4, 0) NOT NULL,  
[NEGR] [numeric](6, 0) NULL,  
[PEGR] [numeric](8, 2) NULL  
) ON [PRIMARY]  
GO
```

Las tablas intermedias se han creado sin restricciones ni índices para facilitar la carga de datos desde las fuentes de origen.

B. Creación del modelo multidimensional.

En este punto veremos los scripts de creación del modelo físico multidimensional que hemos diseñado para el almacén de egresados universitarios, compuesto por las dimensiones y las tablas de hechos. En la creación, además de atributos y métricas, se crearán también las restricciones definidas y que son propias del modelo multidimensional, las claves primarias de las dimensiones y las foráneas de las tablas de hechos.

a) Dimensiones

Dimensión Año (DIM_ANIO)

```
CREATE TABLE [dbo].[DIM_ANIO](  
    [SK_DIM_ANIO] [numeric](5, 0) NOT NULL,  
    [DESC_ANIO] [varchar](10) NOT NULL,  
    CONSTRAINT [PK_DIM_ANIO] PRIMARY KEY CLUSTERED  
(  
        [SK_DIM_ANIO] ASC  
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY  
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]  
) ON [PRIMARY]  
GO
```

Dimensión Tipo de Universidad (DIM_TIPO_UNIV)

```
CREATE TABLE [dbo].[DIM_TIPO_UNIV](  
    [SK_DIM_TIPO_UNIV] [numeric](1, 0) NOT NULL,  
    [DESC_TIPO_UNIV] [varchar](30) NOT NULL,  
    CONSTRAINT [PK_DIM_TIPO_UNIV] PRIMARY KEY CLUSTERED  
(  
        [SK_DIM_TIPO_UNIV] ASC  
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY  
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]  
) ON [PRIMARY]  
GO
```

Dimensión Rama de Enseñanza (DIM_RAMA)

```
CREATE TABLE [dbo].[DIM_RAMA](  
    [SK_DIM_RAMA] [numeric](5, 0) NOT NULL,  
    [COD_RAMA] [varchar](1) NOT NULL,
```

```
[NOM_RAMA] [varchar](50) NOT NULL,  
[COD_RAMA_N2] [varchar](2) NOT NULL,  
[NOM_RAMA_N2] [varchar](100) NOT NULL,  
[COD_RAMA_N3] [varchar](3) NOT NULL,  
[NOM_RAMA_N3] [varchar](100) NOT NULL,  
[COD_RAMA_N4] [varchar](4) NOT NULL,  
[NOM_RAMA_N4] [varchar](150) NOT NULL,  
[COD_RAMA_N5] [varchar](7) NOT NULL,  
[NOM_RAMA_N5] [varchar](200) NOT NULL,  
CONSTRAINT [PK_DIM_RAMA] PRIMARY KEY CLUSTERED  
(  
    [SK_DIM_RAMA] ASC  
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY  
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]  
) ON [PRIMARY]  
GO
```

Dimensión Tipo de Modalidad de Impartición (DIM_MODALIDAD)

```
CREATE TABLE [dbo].[DIM_MODALIDAD](  
    [SK_DIM_MODALIDAD] [numeric](1, 0) NOT NULL,  
    [DESC_MODALIDAD] [varchar](50) NOT NULL,  
    CONSTRAINT [PK_DIM_MODALIDAD] PRIMARY KEY CLUSTERED  
(  
        [SK_DIM_MODALIDAD] ASC  
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY  
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]  
) ON [PRIMARY]  
GO
```

Dimensión Universidades (DIM_UNIVERSIDADES)

```
CREATE TABLE [dbo].[DIM_UNIVERSIDAD](  
    [SK_DIM_UNIVERSIDAD] [numeric](3, 0) NOT NULL,  
    [DESC_UNIVERSIDAD] [varchar](100) NOT NULL,  
    [SK_DIM_MODALIDAD] [numeric](1, 0) NOT NULL,  
    CONSTRAINT [PK_DIM_UNIVERSIDAD] PRIMARY KEY CLUSTERED  
(  
        [SK_DIM_UNIVERSIDAD] ASC  
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY  
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]  
) ON [PRIMARY]  
GO  
  
ALTER TABLE [dbo].[DIM_UNIVERSIDAD] WITH CHECK ADD CONSTRAINT  
[FK_DIM_UNIV_DIM_MODALIDAD] FOREIGN KEY([SK_DIM_MODALIDAD])  
REFERENCES [dbo].[DIM_MODALIDAD] ([SK_DIM_MODALIDAD])  
GO  
  
ALTER TABLE [dbo].[DIM_UNIVERSIDAD] CHECK CONSTRAINT  
[FK_DIM_UNIV_DIM_MODALIDAD]  
GO
```

Dimensión Genero (DIM_SEXO)

```
CREATE TABLE [dbo].[DIM_SEXO](  
    [SK_DIM_SEXO] [numeric](1, 0) NOT NULL,
```

```
[DESC_SEXO] [varchar](10) NOT NULL,  
CONSTRAINT [PK_DIM_SEXO] PRIMARY KEY CLUSTERED  
(  
    [SK_DIM_SEXO] ASC  
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY  
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]  
    ) ON [PRIMARY]  
GO
```

Dimensión Edad (DIM_EDAD)

```
CREATE TABLE [dbo].[DIM_EDAD](  
[SK_DIM_EDAD] [numeric](1, 0) NOT NULL,  
[DESC_INT_EDAD] [varchar](50) NOT NULL,  
CONSTRAINT [PK_DIM_EDAD] PRIMARY KEY CLUSTERED  
(  
    [SK_DIM_EDAD] ASC  
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY  
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]  
    ) ON [PRIMARY]  
GO
```

Dimensión Países (DIM_PAIS)

```
CREATE TABLE [dbo].[DIM_PAIS](  
[SK_DIM_PAIS] [numeric](3, 0) NOT NULL,  
[DESC_PAIS_ES] [varchar](50) NOT NULL,  
[DESC_PAIS_EN] [varchar](200) NOT NULL,  
CONSTRAINT [PK_DIM_PAIS] PRIMARY KEY CLUSTERED  
(  
    [SK_DIM_PAIS] ASC  
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY  
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]  
    ) ON [PRIMARY]  
GO
```

b) Tablas de Hechos

Tabla de Hechos FACT_PEGR_EVOLUTIVO

```
CREATE TABLE [dbo].[FACT_PEGR_EVOLUTIVO](  
[SK_DIM_ANIO] [numeric](5, 0) NOT NULL,  
[SK_DIM_TIPO_UNIV] [numeric](1, 0) NOT NULL,  
[SK_DIM_UNIVERSIDAD] [numeric](3, 0) NOT NULL,  
[COD_RAMA] [varchar](1) NOT NULL NOT NULL,  
[PERSONAS_EGRESADAS] [numeric](8, 0) NOT NULL  
) ON [PRIMARY]  
GO  
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] WITH CHECK ADD CONSTRAINT  
[FK_ANIO] FOREIGN KEY([SK_DIM_ANIO])  
REFERENCES [dbo].[DIM_ANIO] ([SK_DIM_ANIO])  
GO  
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] CHECK CONSTRAINT [FK_ANIO]  
GO
```

```
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] WITH CHECK ADD CONSTRAINT
[FK_TIPO_UNIV] FOREIGN KEY([SK_DIM_TIPO_UNIV])
REFERENCES [dbo].[DIM_TIPO_UNIV] ([SK_DIM_TIPO_UNIV])
GO
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] CHECK CONSTRAINT
[FK_TIPO_UNIV]
GO

ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] WITH CHECK ADD CONSTRAINT
[FK_UNIVERSIDAD] FOREIGN KEY([SK_DIM_UNIVERSIDAD])
REFERENCES [dbo].[DIM_UNIVERSIDAD] ([SK_DIM_UNIVERSIDAD])
GO
ALTER TABLE [dbo].[FACT_PEGR_EVOLUTIVO] CHECK CONSTRAINT
[FK_UNIVERSIDAD]
GO
```

Tabla de Hechos FACT_PEGR_PERFIL

```
CREATE TABLE [dbo].[FACT_PEGR_PERFIL](
[SK_DIM_ANIO] [numeric](5, 0) NOT NULL,
[SK_DIM_SEXO] [numeric](1, 0) NOT NULL,
[SK_DIM_EDAD] [numeric](1, 0) NOT NULL,
[COD_RAMA_N4] [varchar](4) NOT NULL,
[PERSONAS_EGRESADAS] [numeric](8, 0) NOT NULL
) ON [PRIMARY]
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] WITH CHECK ADD CONSTRAINT
[FK_FACT_PEGR_PERFIL_ANIO] FOREIGN KEY([SK_DIM_ANIO])
REFERENCES [dbo].[DIM_ANIO] ([SK_DIM_ANIO])
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] CHECK CONSTRAINT
[FK_FACT_PEGR_PERFIL_ANIO]
GO

ALTER TABLE [dbo].[FACT_PEGR_PERFIL] WITH CHECK ADD CONSTRAINT
[FK_FACT_PEGR_PERFIL_EDAD] FOREIGN KEY([SK_DIM_EDAD])
REFERENCES [dbo].[DIM_EDAD] ([SK_DIM_EDAD])
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] CHECK CONSTRAINT
[FK_FACT_PEGR_PERFIL_EDAD]
GO

ALTER TABLE [dbo].[FACT_PEGR_PERFIL] WITH CHECK ADD CONSTRAINT
[FK_FACT_PEGR_PERFIL_SEXO] FOREIGN KEY([SK_DIM_SEXO])
REFERENCES [dbo].[DIM_SEXO] ([SK_DIM_SEXO])
GO
ALTER TABLE [dbo].[FACT_PEGR_PERFIL] CHECK CONSTRAINT
[FK_FACT_PEGR_PERFIL_SEXO]
GO
```

Tabla de Hechos FACT_PEGR_INSERTADAS

```
CREATE TABLE [dbo].[FACT_PEGR_INSERTADAS](
[SK_DIM_ANIO] [numeric](5, 0) NOT NULL,
[SK_DIM_TIPO_UNIV] [numeric](1, 0) NOT NULL,
[SK_DIM_SEXO] [numeric](1, 0) NOT NULL,
[COD_RAMA] [varchar](1) NOT NULL,
```

```
[PEGR_TRABAJANDO] [numeric](8, 0) NOT NULL,  
[PEGR_DESEMPELADOS] [numeric](8, 0) NOT NULL,  
[PEGR_INACTIVOS] [numeric](8, 0) NOT NULL  
) ON [PRIMARY]  
GO  
  
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_INSERTADAS_ANIO] FOREIGN KEY([SK_DIM_ANIO])  
REFERENCES [dbo].[DIM_ANIO] ([SK_DIM_ANIO])  
GO  
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] CHECK CONSTRAINT  
[FK_FACT_PEGR_INSERTADAS_ANIO]  
GO  
  
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_INSERTADAS_SEXO] FOREIGN KEY([SK_DIM_SEXO])  
REFERENCES [dbo].[DIM_SEXO] ([SK_DIM_SEXO])  
GO  
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] CHECK CONSTRAINT  
[FK_FACT_PEGR_INSERTADAS_SEXO]  
GO  
  
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_INSERTADAS_TIPO_UNIV] FOREIGN KEY([SK_DIM_TIPO_UNIV])  
REFERENCES [dbo].[DIM_TIPO_UNIV] ([SK_DIM_TIPO_UNIV])  
GO  
ALTER TABLE [dbo].[FACT_PEGR_INSERTADAS] CHECK CONSTRAINT  
[FK_FACT_PEGR_INSERTADAS_TIPO_UNIV]  
GO
```

Tabla de Hechos FACT_PEGR_COMPARATIVA

```
CREATE TABLE [dbo].[FACT_PEGR_COMPARATIVA]  
[SK_DIM_ANIO] [numeric](5, 0) NOT NULL,  
[SK_DIM_PAIS] [numeric](3, 0) NOT NULL,  
[PERSONAS_EGRESADAS] [numeric](8, 0) NOT NULL,  
[PORCENTAJE_PEGR_JOVENES] [numeric](8, 0) NOT NULL  
) ON [PRIMARY]  
GO  
  
ALTER TABLE [dbo].[FACT_PEGR_COMPARATIVA] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_COMPARATIVA_ANIO] FOREIGN KEY([SK_DIM_ANIO])  
REFERENCES [dbo].[DIM_ANIO] ([SK_DIM_ANIO])  
GO  
  
ALTER TABLE [dbo].[FACT_PEGR_COMPARATIVA] CHECK CONSTRAINT  
[FK_FACT_PEGR_COMPARATIVA_ANIO]  
GO  
  
ALTER TABLE [dbo].[FACT_PEGR_COMPARATIVA] WITH CHECK ADD CONSTRAINT  
[FK_FACT_PEGR_COMPARATIVA_PAIS] FOREIGN KEY([SK_DIM_PAIS])  
REFERENCES [dbo].[DIM_PAIS] ([SK_DIM_PAIS])  
GO  
  
ALTER TABLE [dbo].[FACT_PEGR_COMPARATIVA] CHECK CONSTRAINT  
[FK_FACT_PEGR_COMPARATIVA_PAIS]  
GO
```

C. Creación del proceso de extracción, transformación y carga (ETL).

Una vez que tenemos implementado el modelo físico del almacén, pasaremos a diseñar los procesos ETL que permitirán poblar las tablas intermedias del área intermedia (*staging área*) y las tablas dimensiones y de hechos del *data mart* que hemos diseñado.

Antes del diseño de las transformaciones definiremos en PDI las variables de entorno que usaremos en la implementación de los procesos ETL, así como la conexión a la base de datos que utilizaremos en todos ellos.

a) Variables de entorno

Es una buena práctica utilizar variables de entorno, para evitar introducir errores en definiciones repetitivas durante la implementación de los procesos. PDI nos permite añadir variables personalizadas y propias de nuestros desarrollos el archivo *kettle.properties*.

En nuestro caso utilizaremos dos variables. Una para almacenar la ruta de las fuentes de datos.

Variable name: DIR_ENT
Value: C:\ proyecto_egresados\fuentes

Y otra dos para almacenar las cadenas de conexión a la base de datos.

Variable name: STAGE
Value: jdbc:sqlserver://UCS1R1UOCSQL01:1433;databaseName= DW_DB_loginuoc;integratedSecurity=false

Variable name: BBDD
Value: jdbc:sqlserver://UCS1R1UOCSQL01:1433;databaseName= DW_DB_loginuoc;integratedSecurity=false

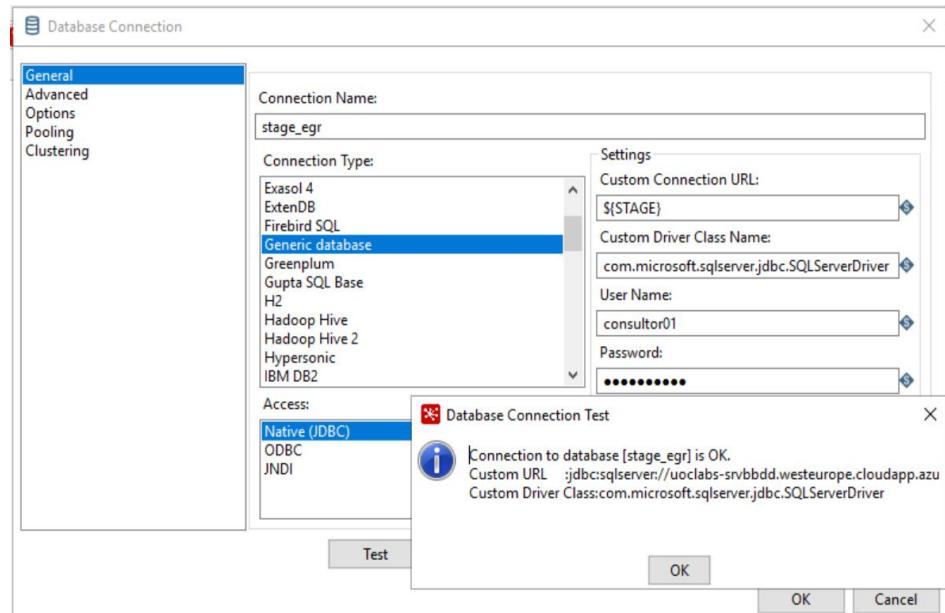
La referencia a las variables de entorno durante la implementación de los procesos se realiza mediante llaves, de esta manera: {DIR_ENT}, {STAGE}, {BBDD}.

b) Conexión Base de Datos SQL Server

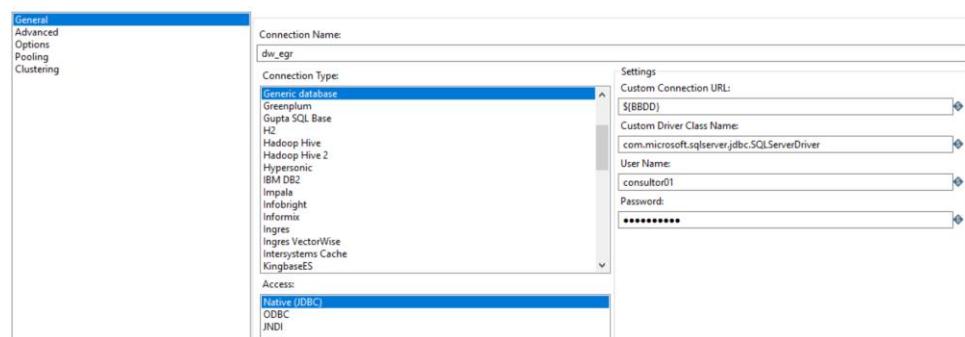
Otro paso previo que se debe realizar, es crear las conexiones a las bases de datos que se usan en todas las transformaciones y trabajos de los procesos de carga.

Se han definido dos conexiones diferentes, una para la base de datos del modelo multidimensional (BBDD) y otra para el área intermedia (STAGE), de esta manera diferenciamos claramente su uso.

Creación de la conexión al STAGE, el nombre que utilizamos es stage_egr:



Creación de la conexión al DW, el nombre que le damos es dw_egr:



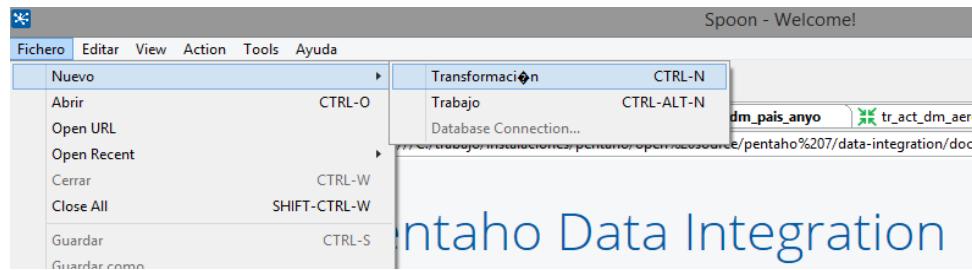
c) Bloque IN

Transformación IN_RAMA

El primer proceso a desarrollar es la carga de la fuente con clasificación normalizada internacional de educación a la tabla intermedia IN_RAMA.

Partimos del fichero: ISCED_2013.csv

El proceso IN_RAMA contiene tres transformaciones: Lectura del fichero csv, Ordenación de los datos y Carga a la tabla intermedia IN_RAMA



Este es el primer paso de la transformación como se trata de un fichero plano utilizaremos como Entrada el tipo “**CSV file input**”. Y realizaremos la parametrización cambiando los valores por defecto para que se puedan capturar los datos correctamente del fichero de entrada.

- La ruta del fichero utilizaremos la variable de entorno DIR_ENT, previamente creada.
- Separador de campos: ";" (cambiaremos porque por defecto se indica ',')
- La primera fila son los nombres de los campos.
- Quitar espacio de origen en los campos tipo string por la derecha y por la izquierda, con la función Trim.

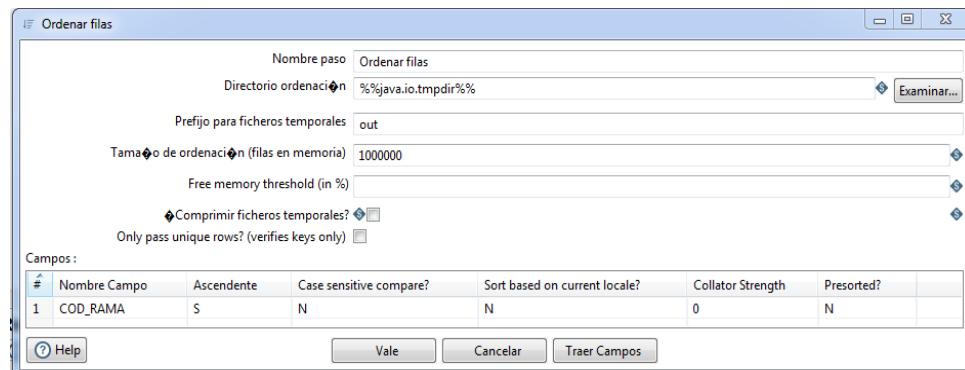
Una vez definido el fichero de entrada podemos obtener los datos de dicho fichero utilizando el botón “Traer Datos” del paso de CSV Input de la transformación.

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	COD_RAMA	String	#	15	0	€	,	.	ambos
2	NOM_RAMA	String		29		€	,	.	ambos
3	COD_RAMA_N2	String	#	15	0	€	,	.	ambos
4	NOM_RAMA_N2	String		46		€	,	.	ambos
5	COD_RAMA_N3	String	#	15	0	€	,	.	ambos
6	NOM_RAMA_N3	String		51		€	,	.	ambos
7	COD_RAMA_N4	String	#	15	0	€	,	.	ambos
8	NOM_RAMA_N4	String		70		€	,	.	ambos
9	COD_RAMA_N5	String	#	15	0	€	,	.	ambos
10	NOM_RAMA_N5	String		50		€	,	.	ambos

Para realizar una visualización previa de los datos que se cargarán se utiliza el botón Previsualizar.

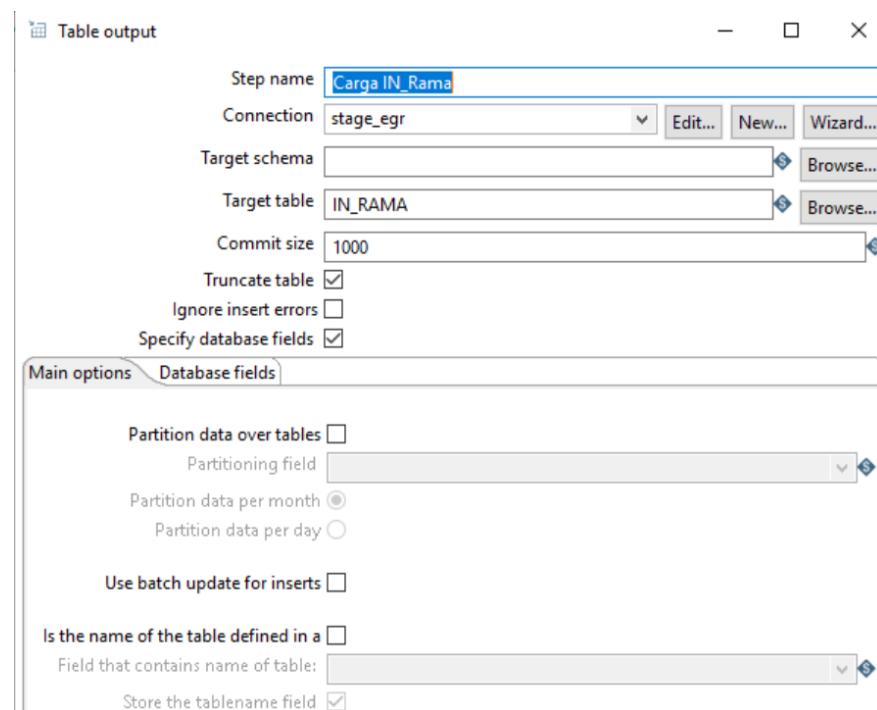
Examine preview data								
Rows of step: CSV file input (562 rows)								
#	COD_RAMA	NOM_RAMA	COD_RAMA_N2	NOM_RAMA_N2	COD_RAMA_N3	NOM_RAMA_N3	COD_RAMA_N4	NOM_RAMA_N4
1	1	Artes y humanidades	01	Educación	011	Educación	0111	Ciencias de la educación
2	1	Artes y humanidades	01	Educación	011	Educación	0111	Ciencias de la educación
3	1	Artes y humanidades	01	Educación	011	Educación	0112	Formación de docentes de enseñanza
4	1	Artes y humanidades	01	Educación	011	Educación	0113	Formación de docentes de educación
5	1	Artes y humanidades	01	Educación	011	Educación	0114	Formación de docentes de educación
6	1	Artes y humanidades	01	Educación	011	Educación	0114	Formación de docentes de educación
7	1	Artes y humanidades	01	Educación	011	Educación	0114	Formación de docentes de educación
8	1	Artes y humanidades	01	Educación	011	Educación	0119	Educación (Otras enseñanzas)
9	1	Artes y humanidades	01	Educación	011	Educación	0129	Técnicas audiovisuales y medios
10	1	Artes y humanidades	02	Artes	021	Artes	0211	Diseño de moda e interiores
11	1	Artes y humanidades	02	Artes y humanidades	021	Artes	0212	Bellas artes
12	1	Artes y humanidades	02	Artes y humanidades	021	Artes	0213	Bellas artes
13	1	Artes y humanidades	02	Artes y humanidades	021	Artes	0213	Conservación, restauración y atención
14	1	Artes y humanidades	02	Artes y humanidades	021	Artes	0214	Muñecos o actos del esoterismo
15	1	Artes y humanidades	02	Artes y humanidades	021	Artes		

El siguiente paso de la transformación sería la ordenación ascendente por el campo COD_RAMA. Para ello utilizaremos la función “Ordenar Filas” de las posibles transformaciones disponibles.



Por último, cargamos los datos en la tabla intermedia, utilizando el paso ‘Salida Tabla’ de la carpeta ‘Salida’. Este paso necesita especificar la conexión de base de datos, utilizaremos la variable de entorno BBDD que hemos definido.

El paso de carga de datos a la tabla intermedia lo configuramos como sigue en el menú principal:

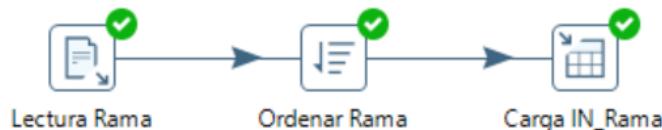


Para dejar la transformación preparada para posibles reprocesos, es necesario realizar un borrado previo para actualizar los datos en el caso de reproceso. Para esto, activaremos el check 'Vaciar tabla'.

Y en los campos de la base de datos:

Fields to insert:		
#	Table field	Stream field
1	COD_RAMA	COD_RAMA
2	NOM_RAMA	NOM_RAMA
3	COD_RAMA_...	COD_RAMA_N2
4	NOM_RAMA_...	NOM_RAMA_N2
5	COD_RAMA_...	COD_RAMA_N3
6	NOM_RAMA_...	NOM_RAMA_N3
7	COD_RAMA_...	COD_RAMA_N4
8	NOM_RAMA_...	NOM_RAMA_N4
9	COD_RAMA_...	COD_RAMA_N5
10	NOM_RAMA_...	NOM_RAMA_N5

La transformación completa es la siguiente:



El resultado de la ejecución es el siguiente:

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	Lectura Rama	0	0	161	162	0	0	0	0	Finished	0.0s	6.750
2	Ordenar Rama	0	161	161	0	0	0	0	0	Finished	0.1s	2.268
3	Carga IN_Rama	0	161	161	0	161	0	0	0	Finished	0.4s	456

Como se observa en las métricas se cargan los 161 registros del fichero de entrada.

Transformación IN_SEGR_N

Extrae la información de 2 archivos csv's y carga los datos normalizados, transformándolos de columnas a filas, en la tabla intermedia IN_SEGR_N del staging área.

Partimos de los ficheros: SEGR1.csv y SEGR2.csv.

La transformación IN_SEGR_N contiene 5 pasos: Lectura de los 2 ficheros csv, Unión de los datos, Normalización de los datos en columnas a filas y Carga a la tabla intermedia IN_SEGR_N.

- Paso **CSV file input** Lectura Segr1, indicamos el archivo a cargar y determinamos con la ayuda de PDI los campos y tipo de dato, además quitamos espacios en ambos lados en los campos tipo *String*.

CSV file input

Step name: Lectura Segr1

Filename: \${DIR_ENT}\SEGR1.csv

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion?

Header row present?

Add filename to result?

The row number field name (optional)

Running in parallel?

New line possible in fields?

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	TIPO_UNIVERSIDAD	String		22		€	,	.	ambos
2	MODALIDAD	String		10		€	,	.	ambos
3	UNIVERSIDAD	String		39		€	,	.	ambos
4	RAMA_ENSEÑANZA	String		29		€	,	.	ambos
5	EGR_C16_17	Integer	#,##0.###	15	0	€	,	.	ninguno
6	EGR_C15_16	Integer	#,##0.###	15	0	€	,	.	ninguno
7	EGR_C14_15	Integer	#,##0.###	15	0	€	,	.	ninguno
8	EGR_C13_14	Integer	#,##0.###	15	0	€	,	.	ninguno
9	EGR_C12_13	Integer	#,##0.###	15	0	€	,	.	ninguno
10	EGR_C11_12	Integer	#,##0.###	15	0	€	,	.	ninguno
11	EGR_C10_11	Integer	#	15	0	€	,	.	ninguno
12	EGR_C09_10	Integer	#	15	0	€	,	.	ninguno

- Paso **CSV file input** Lectura Segr2 con las mismas especificaciones.

CSV file input

Step name: Lectura Segr2

Filename: \${DIR_ENT}\SEGR2.csv

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion?

Header row present?

Add filename to result?

The row number field name (optional)

Running in parallel?

New line possible in fields?

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	TIPO_UNIVERSIDAD	String		22		€	,	.	ambos
2	MODALIDAD	String		10		€	,	.	ambos
3	UNIVERSIDAD	String		21		€	,	.	ambos
4	RAMA_ENSEÑANZA	String		29		€	,	.	ambos
5	EGR_C16_17	Integer	#	15	0	€	,	.	ninguno
6	EGR_C15_16	Integer	#	15	0	€	,	.	ninguno
7	EGR_C14_15	Integer	#	15	0	€	,	.	ninguno
8	EGR_C13_14	Integer	#	15	0	€	,	.	ninguno
9	EGR_C12_13	Integer	#	15	0	€	,	.	ninguno
10	EGR_C11_12	Integer	#	15	0	€	,	.	ninguno
11	EGR_C10_11	Integer	#	15	0	€	,	.	ninguno
12	EGR_C09_10	Integer	#	15	0	€	,	.	ninguno

- Paso **Unión Ordenada** para obtener la suma de los registros de ambos ficheros. Esta transformación exige que ambos ficheros tengan la misma estructura, que podemos obtener con la ayuda del botón "Traer Campos"

Unión ordenada

Nombre de paso: Unión ordenada

Campos:

#	Nombre campo	Ascendente
1	TIPO_UNIVERSIDAD	N
2	MODALIDAD	N
3	UNIVERSIDAD	N
4	RAMA_ENSEÑANZA	N
5	EGR_C16_17	N
6	EGR_C15_16	N
7	EGR_C14_15	N
8	EGR_C13_14	N
9	EGR_C12_13	N
10	EGR_C11_12	N
11	EGR_C10_11	N
12	EGR_C09_10	N

- Paso **Normalización de fila** para convertir cada una de las columnas que contienen los datos de los 8 cursos académicos en las 8 filas correspondientes.

Normalizar filas

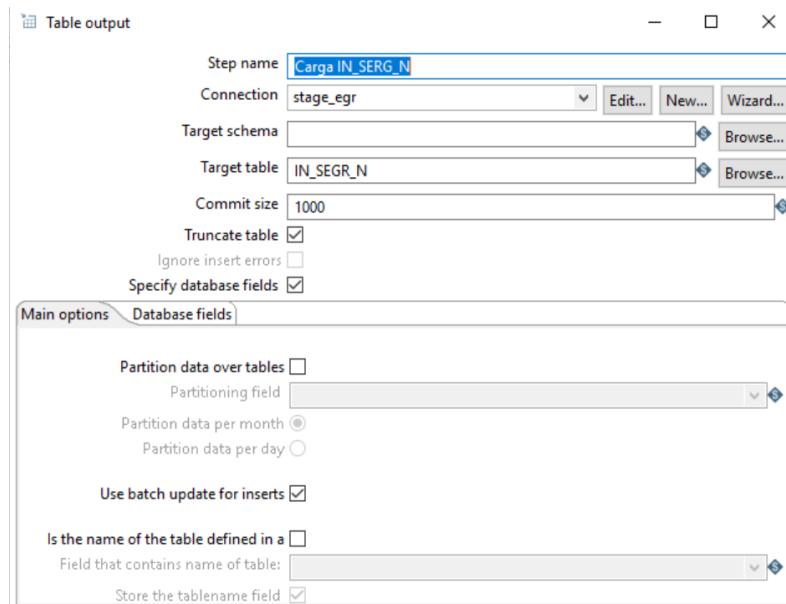
Nombre de paso: Normalización de fila

Tipo de campo: curso

Campos

#	Nombre campo	Tipo	campo nuevo
1	EGR_C16_17	EGR_C16_17	negr
2	EGR_C15_16	EGR_C15_16	negr
3	EGR_C14_15	EGR_C14_15	negr
4	EGR_C13_14	EGR_C13_14	negr
5	EGR_C12_13	EGR_C12_13	negr
6	EGR_C11_12	EGR_C11_12	negr
7	EGR_C10_11	EGR_C10_11	negr
8	EGR_C09_10	EGR_C09_10	negr

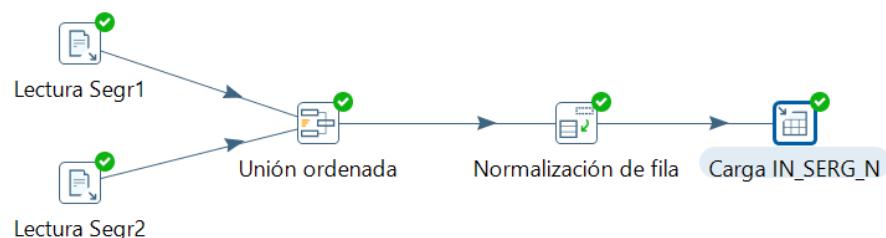
- Paso **Salida de Tabla Carga IN_SEGR_N**, cargamos los datos de los archivos a la tabla intermedia IN_SEGR_N. Activaremos el check 'Vaciar tabla'.



Indicamos la relación entre los campos de la tabla y obtenidos en la transformación.

Fields to insert:		Get fields	Enter field mapping
#	Table field	Stream field	
1	TIPO_UNIVERSIDAD	TIPO_UNIVERSIDAD	
2	MODALIDAD	MODALIDAD	
3	UNIVERSIDAD	UNIVERSIDAD	
4	RAMA_ENSEÑANZA	RAMA_ENSEÑANZA	
5	CURSO	curso	
6	NEGR	negr	

La transformación completa es la siguiente:



El resultado de la ejecución es el siguiente:

Execution Results

#	Nombre paso	Numero Copia	Lecto	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Lectura Segr2	0	0	250	251	0	0	0	0	Finalizado	0.0s	13.210	-
2	Lectura Segr1	0	0	160	161	0	0	0	0	Finalizado	0.0s	10.733	-
3	Unión ordenada	0	410	410	0	0	0	0	0	Finalizado	0.3s	1.258	-
4	Normalización de fila	0	410	3280	0	0	0	0	0	Finalizado	0.3s	9.762	-
5	Carga IN_SEGR_N	0	3280	3280	0	3280	0	0	0	Finalizado	0.4s	8.677	-

Como se observa en las métricas, la transformación convierte 410 registros extraídos de los ficheros en 3280 registros que se cargan en la tabla intermedia IN_SEGR_N.

Transformación IN_EGR_C16_17

Extrae la información del archivo csv y carga los datos en la tabla intermedia IN_EGR_2016_2017 del *staging* área.

Partimos del fichero: grad_5sc.csv

El proceso IN_EGR_C16_17 contiene 3 pasos: Lectura del fichero csv, Ordenación de datos y Carga a la tabla intermedia IN_SEGR_N.

- **Paso CSV file input** Lectura grad_5sc indicamos el archivo a cargar y determinamos con la ayuda de PDI los campos y tipo de dato, además quitamos espacios en ambos lados en los campos tipo String.

CSV file input

Step name: Lectura_grad_5sc

Filename: \${DIR_ENT}\grad_5sc.csv

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion?

Header row present?

Add filename to result

The row number field name (optional)

Running in parallel?

New line possible in fields?

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	COD_AMBITO	String		84		€	,	.	ambos
2	SEXO	String		7		€	,	.	ambos
3	EDAD	String		20		€	,	.	ambos
4	NUM_EGR_NV1	Integer	#####.##	15	0		,		ninguno
5	NUM_EGR_NV2	Integer	#####.##	15	0		,		ninguno

- **Paso Ordenar Filas** Ordenar Egr_c16_17 para ordenación ascendente por el campo COD_AMBITO.

Ordenar filas

Nombre paso	Ordenar Egr_c16_17						
Directorio ordenación	%%java.io.tmpdir%%						
Prefijo para ficheros temporales	out						
Tamaño de ordenación (filas en memoria)	1000000						
Free memory threshold (in %)							
<input checked="" type="checkbox"/> Comprimir ficheros temporales?	<input type="checkbox"/>						
<input type="checkbox"/> Only pass unique rows? (verifies keys only)	<input type="checkbox"/>						
Campos :							
#	Nombre Campo	Ascendente	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?	
1	COD_AMBITO	S	N	N	0	N	

- **Paso Salida de Tabla** Carga IN_EGR_C16_17, cargamos los datos de los archivos a la tabla intermedia IN_EGR_2016_2017, previamente. Activaremos el check 'Vaciar tabla' para eliminar los datos antes de cargar en la ficha Main Options.

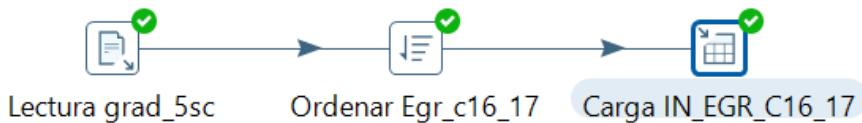
Table output

Step name	Carga IN_EGR_C16_17
Connection	stage_egr
Target schema	
Target table	IN_EGR_2016_2017
Commit size	1000
Truncate table	<input checked="" type="checkbox"/>
Ignore insert errors	<input type="checkbox"/>
Specify database fields	<input checked="" type="checkbox"/>
Main options	
Partition data over tables	<input type="checkbox"/>
Partitioning field	
Partition data per month	<input checked="" type="radio"/>
Partition data per day	<input type="radio"/>
Use batch update for inserts	<input checked="" type="checkbox"/>
Is the name of the table defined in a field?	<input type="checkbox"/>
Field that contains name of table:	
Store the tablename field	<input checked="" type="checkbox"/>
Return auto-generated key	<input type="checkbox"/>
Name of auto-generated key field	

Indicamos la relación entre los campos de la tabla y obtenidos en la transformación.

Main options	Database fields	
Fields to insert:		
#	Table field	Stream field
1	COD_AMBITO	COD_AMBITO
2	SEXO	SEXO
3	EDAD	EDAD
4	NUM_EGR_NV1	NUM_EGR_NV1
5	NUM_EGR_NV2	NUM_EGR_NV2

La transformación completa es la siguiente:



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

#	Nombre paso	Numero Copia	Lecto	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Lectura grad_5sc	0	0	712	713	0	0	0	0	Finalizado	0.0s	47.533	-
2	Ordenar Egr_c16_17	0	712	712	0	0	0	0	0	Finalizado	0.1s	13.185	-
3	Carga IN_EGR_C16_17	0	712	712	0	712	0	0	0	Finalizado	0.1s	5.235	-

Como se observa en las métricas se cargan los 712 registros del fichero csv a la tabla intermedia.

Transformación IN_EIL_03003

Extrae la información del archivo xls con datos de inserción de 2014 de las personas egresadas y los carga en la tabla intermedia IN_EIL_03003 del *staging* área.

Partimos del fichero: 03003.xls

Para este caso práctico hemos utilizado fuentes externas *open data* que utilizaremos para el descubrimiento de conocimiento realizando análisis de los datos. No se utilizan de fuentes operacionales, en cuyo caso hay una etapa muy importante de preparación de las fuentes para dejarlas listas para su tratamiento con la herramienta ETL. Es muy habitual manipular los ficheros, concretamente sobre los ficheros tipo xls que se utilizan en esta transformación se realizarán una serie de acciones de preparación antes de su procesamiento.

Para el fichero con datos de personas egresadas según situación profesional en 2014 (03003.xls) se dispone de 5 bloques: 1.-bloque de totales; 2.-bloque de universidades públicas y hombres; 3.-bloque de universidades públicas y mujeres; 4.-bloque de universidades privadas y hombres, bloque de universidades privadas y mujeres. Se manipularán los ficheros para simplificarlo a un solo bloque.

Eliminamos las 6 primeras filas antes de la cabecera.

	A	B	C	D	E
1		Encuesta de inserción laboral de titulados universitarios 2014. Cifras absolutas			
2		Situación laboral en 2014 de los titulados universitarios del curso 2009-2010. Cifras absolutas			
3					
4		Titulados universitarios según su situación laboral en 2014 por sexo, universidad y campo de estudio			
5		Unidades:			
6		Total	Trabajando	En desempleo	
7	Ambos性				
8	Total				
9	Total	197.535	149.395	35.530	12.610
10	Ciencias sociales y jurídicas	104.836	77.778	19.967	7.091
11	Ingeniería y arquitectura	44.448	35.923	6.556	1.969
12	Artes y humanidades	13.253	8.522	3.320	1.411
13	Ciencias de la salud	23.744	19.295	3.185	1.265
14	Ciencias	11.254	7.878	2.502	874
15	Total universidades públicas	169.410	126.401	32.052	10.957
16	Total	90.166	65.745	18.162	6.260
17	Ciencias sociales y jurídicas	36.717	29.580	5.621	1.516
18	Ingeniería y arquitectura				
19					

Eliminamos todas las filas que están al final del fichero (porque existe unas filas al final con notas y la fuente de datos), concretamente las últimas 9 filas.

67	A	B	C	D	E	65
68		Ciencias	487	381		
69						
70		Notas:				
71		1) Los titulados en más de una titulación se han contabilizado una vez en cada una de las titulaciones.				
72						
73		No se dispone de información de la Universidad Pablo de Olavide.				
74						
75		Fuente:				
76		Instituto Nacional de Estadística				
77						
78						
79						
80						
81						
82						

Como siguiente paso, insertaremos 2 columnas antes de la primera columna.

	A	B	C	D	E	F
1	Ambos性		Trabajando	En desempleo	Inactivo	
2	Total					
3	Total		35	149.395	35.530	12.610
4	Ciencias sociales y jurídicas		36	77.778	19.967	7.091
5	Ingeniería y arquitectura		48	35.923	6.556	1.969
6	Artes y humanidades		53	8.522	3.320	1.411
7	Ciencias de la salud		44	19.295	3.185	1.265
8	Ciencias		54	7.878	2.502	874
9	Total universidades públicas					
10	Total		10	126.401	32.052	10.957
11	Ciencias sociales y jurídicas		90.166	65.745	18.162	6.260
12	Ingeniería y arquitectura					

A las columnas que acabamos de insertar, les pondremos su correspondiente etiqueta. La primera TIPO_UNIVERSIDAD y a la segunda SEXO. También añadimos el título la tercera columna, será RAMA.

A	B	C		D	E	F	G
1	TIPO_UNIVERSIDAD	SEXO	RAMA	Total	Trabajando	En desempleo	Inactivo
2			Ambos sexos				
3			Total				
4			Total	197.535	149.395	35.530	12.610
5			Ciencias sociales y jurídicas	104.836	77.778	19.967	7.091
6			Ingeniería y arquitectura	44.448	35.923	6.556	1.969
7			Artes y humanidades	13.253	8.522	3.320	1.411
8			Ciencias de la salud	23.744	19.295	3.185	1.265
9			Ciencias	11.254	7.878	2.502	874
10			Total universidades públicas				
11			Total	169.410	126.401	32.052	10.957
12			Ciencias sociales y jurídicas	90.166	65.745	18.162	6.260
13			Ingeniería y arquitectura	36.717	29.580	5.621	1.516
14			Artes y humanidades	12.594	8.018	3.238	1.339
15			Ciencias de la salud	19.397	15.720	2.645	1.032
16			Ciencias	10.535	7.339	2.385	811

El siguiente paso para la preparación de este fichero, será eliminar el bloque de totales dado que son valores acumulados, derivados de la suma del resto de bloques, es decir, son redundantes. Descartamos las filas 2:23, de 25:31 y de 47:53. Haciendo clic + CTRL se pueden borrar bloques de filas en una sola operación.

A	B	C		D	E	F	G
1	TIPO_UNIVERSIDAD	SEXO	RAMA	Total	Trabajando	En desempleo	Inactivo
2			Ambos sexos				
3			Total				
4			Total	197.535	149.395	35.530	12.610
5			Ciencias sociales y jurídicas	104.836	77.778	19.967	7.091
6			Ingeniería y arquitectura	44.448	35.923	6.556	1.969
7			Artes y humanidades	13.253	8.522	3.320	1.411
8			Ciencias de la salud	23.744	19.295	3.185	1.265
9			Ciencias	11.254	7.878	2.502	874
10			Total universidades públicas				
11			Total	169.410	126.401	32.052	10.957
12			Ciencias sociales y jurídicas	90.166	65.745	18.162	6.260
13			Ingeniería y arquitectura	36.717	29.580	5.621	1.516
14			Artes y humanidades	12.594	8.018	3.238	1.339
15			Ciencias de la salud	19.397	15.720	2.645	1.032
16			Ciencias	10.535	7.339	2.385	811
17			Total universidades privadas				
18			Total	28.125	22.994	3.478	1.653
19			Ciencias sociales y jurídicas	14.669	12.033	1.804	832
20			Ingeniería y arquitectura	7.730	6.343	935	453
21			Artes y humanidades	658	504	82	72
22			Ciencias de la salud	4.348	3.575	539	234
23			Ciencias	719	539	117	63

El siguiente paso es quitar la propiedad de celdas combinadas en las filas 2 y 17, para poder moverlas a la columna de sexo, que hemos creado anteriormente.

A	B	C	D	E	F	G
TIPO_UNIVERSIDAD	SEXO	RAMA	Total	Trabajando	En desempleo	Inactivo
	Hombres					
		Total universidades públicas				
		Total	66.186	50.796	11.441	3.949
		Ciencias sociales y jurídicas	27.915	20.637	5.323	1.956
		Ingeniería y arquitectura	26.033	21.411	3.570	1.052
		Artes y humanidades	4.254	2.674	1.122	458
		Ciencias de la salud	4.183	3.510	570	102
		Ciencias	3.801	2.565	856	380
		Total universidades privadas				
		Total	12.450	10.513	1.306	631
		Ciencias sociales y jurídicas	5.135	4.395	503	237
		Ingeniería y arquitectura	5.547	4.649	595	303
		Artes y humanidades	242	175	32	35
		Ciencias de la salud	1.295	1.136	124	35
		Ciencias	232	159	52	22
	Mujeres					
		Total universidades públicas				
17						
18						

Copiamos los valores a todas las filas vacías de la columna B según el sexo. De la B2 a la B16, las celdas corresponden al sexo, “Hombres” y de la B17 a la B31, las celdas corresponden al sexo, “Mujeres”.

A	B	C	D	E	F	G
TIPO_UNIVERSIDAD	SEXO	RAMA	Total	Trabajando	En desempleo	Inactivo
	Hombres					
	Hombres	Total universidades públicas				
	Hombres	Total	66.186	50.796	11.441	3.949
	Hombres	Ciencias sociales y jurídicas	27.915	20.637	5.323	1.956
	Hombres	Ingeniería y arquitectura	26.033	21.411	3.570	1.052
	Hombres	Artes y humanidades	4.254	2.674	1.122	458
	Hombres	Ciencias de la salud	4.183	3.510	570	102
	Hombres	Ciencias	3.801	2.565	856	380
	Hombres	Total universidades privadas				
	Hombres	Total	12.450	10.513	1.306	631
	Hombres	Ciencias sociales y jurídicas	5.135	4.395	503	237
	Hombres	Ingeniería y arquitectura	5.547	4.649	595	303
	Hombres	Artes y humanidades	242	175	32	35
	Hombres	Ciencias de la salud	1.295	1.136	124	35
	Hombres	Ciencias	232	159	52	22
	Mujeres					
	Mujeres	Total universidades públicas				
	Mujeres	Total	103.224	75.605	20.610	7.009
	Mujeres	Ciencias sociales y jurídicas	62.251	45.108	12.840	4.304
	Mujeres	Ingeniería y arquitectura	10.685	8.169	2.051	464
	Mujeres	Artes y humanidades	8.340	5.344	2.116	881
	Mujeres	Ciencias de la salud	15.214	12.210	2.075	929
	Mujeres	Ciencias	6.734	4.774	1.529	431
	Mujeres	Total universidades privadas				
	Mujeres	Total	15.675	12.481	2.172	1.022
	Mujeres	Ciencias sociales y jurídicas	9.535	7.638	1.302	595
	Mujeres	Ingeniería y arquitectura	2.184	1.694	340	150
	Mujeres	Artes y humanidades	417	329	50	37
	Mujeres	Ciencias de la salud	3.053	2.439	415	199
	Mujeres	Ciencias	487	381	65	41

Ahora pasamos a completar el valor de las celdas de la columna A, con el TIPO_UNIVERSIDAD. El primer bloque por cada sexo se corresponde a datos de las universidades públicas, entonces escribimos en las celdas A2 y A17, el dato “Universidades Públicas” y en las celdas A10 y A25, el dato “Universidades Privadas”, y luego copiamos a las celdas vacías de la columna A con el valor correspondiente. Quedando las celdas de A2 a la A23, con el valor “Universidades Públicas” y de la celda A24 a la celda A45 tendría el valor “Universidades Privadas”.

A	B	C	D	E	F	G
TIPO_UNIVERSIDAD	SEXO	RAMA	Total	Trabajando	En desempleo	Inactivo
Universidades Públicas	Hombres					
Universidades Públicas	Hombres	Total universidades públicas				
Universidades Públicas	Hombres	Total	66.186	50.796	11.441	3.949
Universidades Públicas	Hombres	Ciencias sociales y Jurídicas	27.915	20.637	5.323	1.956
Universidades Públicas	Hombres	Ingeniería y Arquitectura	26.033	21.411	3.570	1.052
Universidades Públicas	Hombres	Humanidades	4.254	2.674	1.122	458
Universidades Públicas	Hombres	Ciencias de la salud	4.183	3.510	570	102
Universidades Públicas	Hombres	Ciencias	3.801	2.565	856	380
Universidades Privadas	Hombres	Total universidades privadas				
Universidades Privadas	Hombres	Total	12.450	10.513	1.306	631
Universidades Privadas	Hombres	Ciencias sociales y Jurídicas	5.135	4.395	503	237
Universidades Privadas	Hombres	Ingeniería y Arquitectura	5.547	4.649	595	303
Universidades Privadas	Hombres	Humanidades	242	175	32	35
Universidades Privadas	Hombres	Ciencias de la salud	1.295	1.136	124	35
Universidades Privadas	Hombres	Ciencias	232	159	52	22
Mujeres		Total universidades públicas				
Universidades Públicas	Mujeres	Total	103.224	75.605	20.610	7.009
Universidades Públicas	Mujeres	Ciencias sociales y Jurídicas	62.251	45.108	12.840	4.304
Universidades Públicas	Mujeres	Ingeniería y Arquitectura	10.685	8.169	2.051	464
Universidades Públicas	Mujeres	Humanidades	8.340	5.344	2.116	881
Universidades Públicas	Mujeres	Ciencias de la salud	15.214	12.210	2.075	929
Universidades Públicas	Mujeres	Ciencias	6.734	4.774	1.529	431
Universidades Privadas	Mujeres	Total universidades privadas				
Universidades Privadas	Mujeres	Total	15.675	12.481	2.172	1.022
Universidades Privadas	Mujeres	Ciencias sociales y Jurídicas	9.535	7.638	1.302	595
Universidades Privadas	Mujeres	Ingeniería y Arquitectura	2.184	1.694	340	150
Universidades Privadas	Mujeres	Humanidades	417	329	50	37
Universidades Privadas	Mujeres	Ciencias de la salud	3.053	2.439	415	199
Universidades Privadas	Mujeres	Ciencias	487	381	65	41

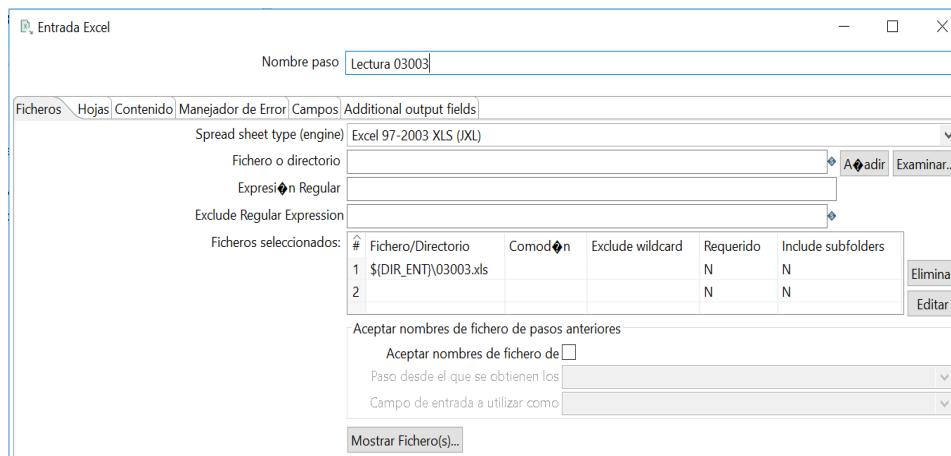
El siguiente paso es eliminar las filas vacías y con totales. Concretamente, los grupos de filas 2:4; 10:11; 17:19; 25:26. Y la columna D (Total). Quedando el fichero depurado y con los datos en un solo bloque, preparado para cargar.

TIPO_UNIVERSIDAD	SEXO	RAMA	Trabajando	En desempleo	Inactivo
Universidades Públicas	Hombres	Ciencias sociales y Jurídicas			
Universidades Públicas	Hombres	Ingeniería y Arquitectura			
Universidades Públicas	Hombres	Artes y Humanidades			
Universidades Públicas	Hombres	Ciencias de la salud			
Universidades Públicas	Hombres	Ciencias			
Universidades Privadas	Hombres	Ciencias sociales y Jurídicas			
Universidades Privadas	Hombres	Ingeniería y Arquitectura			
Universidades Privadas	Hombres	Artes y Humanidades			
Universidades Privadas	Hombres	Ciencias de la salud			
Universidades Privadas	Hombres	Ciencias			
Universidades Públicas	Mujeres	Ciencias sociales y Jurídicas			
Universidades Públicas	Mujeres	Ingeniería y Arquitectura			
Universidades Públicas	Mujeres	Artes y Humanidades			
Universidades Públicas	Mujeres	Ciencias de la salud			
Universidades Públicas	Mujeres	Ciencias			
Universidades Privadas	Mujeres	Ciencias sociales y Jurídicas			
Universidades Privadas	Mujeres	Ingeniería y Arquitectura			
Universidades Privadas	Mujeres	Artes y Humanidades			
Universidades Privadas	Mujeres	Ciencias de la salud			
Universidades Privadas	Mujeres	Ciencias			

Después de todos los pasos realizados anteriormente, ya tenemos el fichero 03003.xls preparado para cargar mediante transformaciones de PDI.

La transformación IN_EIL_03003 contiene 4 pasos: Lectura del fichero xls, Conversión de cadenas a Mayúsculas, Ordenación de datos y Carga a la tabla intermedia IN_EIL_03003.

- **Paso Entrada Excel** Lectura 03003, en la pestaña Ficheros añadimos el fichero Excel desde donde extraemos los datos para ello utilizamos la variable de entorno DIR_ENT.



En la pestaña Hojas, indicamos el nombre de la hoja de cálculo que queremos procesar e indicamos desde qué fila del archivo se comienzan a leer los datos.



Indicamos en la siguiente pestaña que existe una fila de encabezados de los campos.



Le indicamos que recupere los campos que vamos a tratar mediante el botón que nos muestra el paso y completamos la definición de los campos, especificando la precisión de los campos numéricos y eliminando espacios en ambos lados de los campos de tipo string.

Campos											Additional output fields	
#	Nombre	Tipo	Longitud	Precisión	Tipo de poda	Repetir	Formato	Moneda	Decimal	Agrupamiento		
1	TIPO_UNIVERSIDAD	String	-1	-1	ambos	N						
2	SEXO	String	-1	-1	ambos	N						
3	RAMA	String	-1	-1	ambos	N						
4	Trabajando	Number	-1	0	ninguno	N			0			
5	En desempleo	Number	-1	0	ninguno	N			0			
6	Inactivo	Number	-1	0	ninguno	N			0			

- Paso **Ordenar Filas** para ordenación ascendente por los campos TIPO_UNIVERSIDAD, SEXO y RAMA.

Ordenar filas

Nombre paso	Ordenar filas						
Directorio ordenación	%%java.io.tmpdir%% <input type="button" value="Examinar..."/>						
Prefijo para ficheros temporales	out						
Tamaño de ordenación (filas en memoria)	1000000						
Free memory threshold (in %)							
Comprimir ficheros temporales?	<input type="checkbox"/>						
Only pass unique rows? (verifies keys only)	<input type="checkbox"/>						
Campos :							
#	Nombre Campo	Ascendente	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?	
1	TIPO_UNIVERSIDAD	S	N	N	0	N	
2	SEXO	S	N	N	0	N	
3	RAMA	S	N	N	0	N	

- Paso **Salida de Tabla** Cargar IN_EIL_03003 para cargar los datos a la tabla intermedia IN_EIL_03003. En la ficha Main options, activaremos el check 'Vaciar tabla' para eliminar los datos antes de cargar y evitar duplicados cuando se tengan que realizar reprocesos.

Table output

Step name	Cargar IN_EIL_03003
Connection	stage_egr <input type="button" value="Edit..."/> <input type="button" value="New..."/> <input type="button" value="Wizard..."/>
Target schema	
Target table	IN_EIL_03003 <input type="button" value="Browse..."/>
Commit size	1000
Truncate table	<input checked="" type="checkbox"/>
Ignore insert errors	<input type="checkbox"/>
Specify database fields	<input checked="" type="checkbox"/>
Main options	Database fields
Partition data over tables	<input type="checkbox"/>
Partitioning field	
Partition data per month	<input checked="" type="radio"/>
Partition data per day	<input type="radio"/>
Use batch update for inserts	<input checked="" type="checkbox"/>
Is the name of the table defined in a	<input type="checkbox"/>
Field that contains name of table:	
Store the tablename field	<input checked="" type="checkbox"/>

En la ficha Database fields, indicamos la relación entre los campos de la tabla y obtenidos en la transformación.

Main options		Database fields	
Fields to insert:			
#	Table field	Stream field	
1	TIPO_UNIVERSIDAD	TIPO_UNIVERSIDAD	
2	SEXO	SEXO	
3	RAMA_ENSEÑANZA	RAMA	
4	TRABAJANDO	Trabajando	
5	DESEMPLÉO	En desempleo	
6	INACTIVO	Inactivo	

La transformación completa es la siguiente:



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

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	Lectura 03003	0	0	20	20	0	0	0	0	Finished	1.3s	16
2	String operations	0	20	20	0	0	0	0	0	Finished	1.3s	15
3	Ordenar filas	0	20	20	0	0	0	0	0	Finished	1.3s	15
4	Cargar IN_EIL_03003	0	20	20	0	20	0	0	0	Finished	1.4s	15

Como se observa en las métricas se cargan los 20 registros del fichero a la tabla intermedia.

Transformación IN_EGR_EUR

La transformación extrae la información de 2 archivos xls y carga los datos normalizados, transformándolos de columnas a filas, en la tabla intermedia IN_EGR_EUR del *staging* área.

Partimos de los ficheros: edat_lfse_03.xls y educ_uee_grad01.xls

Lo primero que haremos es la preparación de los ficheros antes de su procesamiento. Para el fichero con datos del número de egresados universitarios por años entre España y otros Países (educ_uee_grad01.xls) eliminamos las hojas de cálculo Data2 y Data3, dejando sólo la primera (Data).

En la hoja Data con la que nos hemos quedado, eliminamos las primeras filas hasta la cabecera, desde 1:11, y también las 3 últimas filas. Dejaremos sólo las filas con los datos y sus cabeceras.

Esto mismo, se puede hacer también directamente en la transformación Entrada Excel, indicando donde se encuentran los datos a tratar.

A	B	C	D	E	F	G	H
1 Graduates by education level, programme orientation, completion, sex and age [educ_uee_grad01]							
2							
3 Last update 05.04.19							
4 Extracted on 09.04.19							
5 Source of data Eurostat							
6							
7	UNIT	Number					
8	AGE	Total					
9	SEX	Total					
10	ISCED11	Tertiary education (levels 5-8)					
11	GEO/TIME	2013	2014	2015			
12	Denmark	66.467	70.245	74.428			
13	Germany (u)	495.808	521.845	544.743			
14	Ireland	61.297	64.955	67.303			
15	Spain	407.036	443.321	438.616			
16	France	727.011	742.565	752.068			
17	Italy	361.907	374.353	370.298			
18	Netherlands	138.287	141.270	148.942			
19	Finland	52.730	53.878	56.829			
20	Sweden	72.782	74.736	78.244			
21	United Kingdom	791.945	772.362				
22	Norway	44.753	47.742	48.212	49.010	53.085	
23	Switzerland	81.909	85.750	86.178	87.479		
24							
25							
26	Special value:						
27	:	not available					

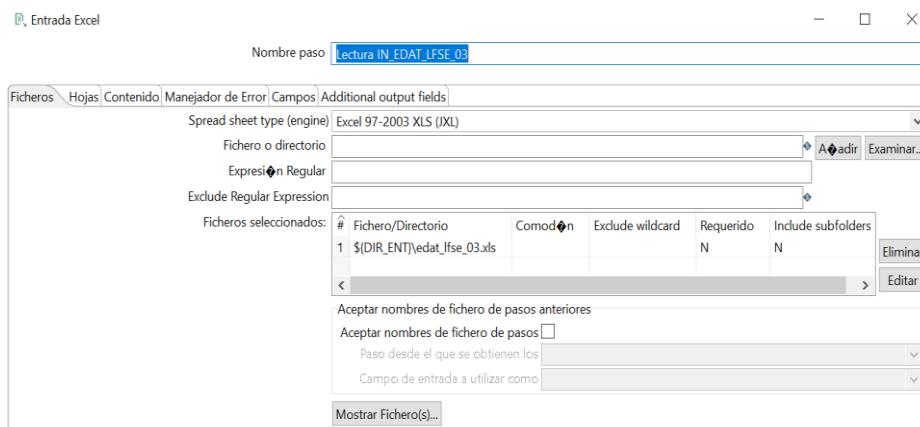
Realizaremos la misma preparación, es decir eliminar hojas y filas para preparar el fichero con datos del porcentaje de egresados universitarios jóvenes (edat_lfse_03.xls). El primer paso será eliminar la hoja de cálculo Data, dejando sólo Data2. Sobre esta hoja, eliminamos las primeras filas hasta la cabecera, desde 1:10 y las 3 últimas filas, dejando sólo las filas con los datos a cargar y sus cabeceras.

A	B	C	D	E	F	G	H	I
1 Graduates in tertiary education by age groups - per 1000 of population								
2								
3 Last update 05.04.19								
4 Extracted on 09.04.19								
5 Source of data Eurostat								
6								
7	UNIT	Per thousand inhabitants						
8	ISCED11	Tertiary education (levels 5-8)						
9	AGE	From 20 to 29 years						
10	GEO/TIME	2013	2014	2015	2016	2017	2018	2019
11	Denmark	74,2	76,6	79,7	90,1			
12	Germany (u)	:	:	45,9	45,7			
13	Ireland	72,6	79,6	84,9	76,0			
14	Spain	55,9	64,3	66,4	68,1			
15	France	:	:	:	:			
16	Italy	:	41,4	43,4	45,0			
17	Netherlands	58,4	59,6	62,7	63,3			
18	Finland	50,6	50,8	54,4	53,9			
19	Sweden	37,6	37,7	37,4	38,2			
20	United Kingdom	67,0	67,7		65,3			
21	Norway	48,3	50,7	50,9	51,7	56,1		
22	Switzerland	53,5	55,5	55,5	55,6			
23								
24								
25	Special value:							
26	:	not available						

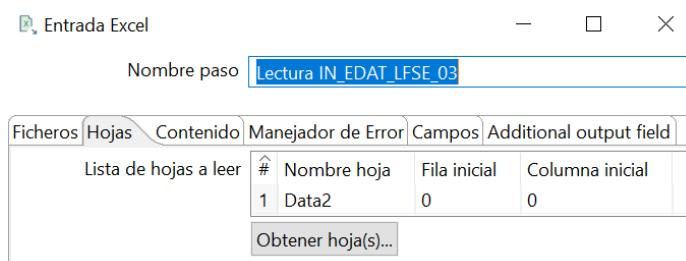
Ahora pasamos a PDI para crear las ETL's que procesará todos los ficheros.

La transformación IN_EGR_EUR contiene 9 pasos: Lectura de los 2 ficheros xls's, 2 Reemplazamientos de cadenas, 2 Normalizaciones de los datos de columnas a filas, Unión de datos por clave, Select Values y Carga a la tabla intermedia IN_EGR_EUR.

- Paso **Entrada Excel** Lectura IN_EDAT_LFSE_03, en la pestaña Ficheros añadimos el fichero Excel desde donde extraemos los datos para ello utilizamos la variable de entorno DIR_ENT.



En la pestaña Hojas, indicamos el nombre de la hoja de cálculo, Data2, que queremos procesar e indicamos desde qué fila del archivo se comienzan a leer los datos, en nuestro caso después de la preparación la fila 0. En el caso de no haber preparado previamente la fuente de datos, en esta ficha indicaremos en qué fila y en qué columna comenzarán los datos a procesar.



Indicamos en la siguiente pestaña que existe una fila de encabezados de los campos.



Le indicamos que recupere los campos que vamos a tratar mediante el botón que nos muestra el paso y completamos la definición de los campos,

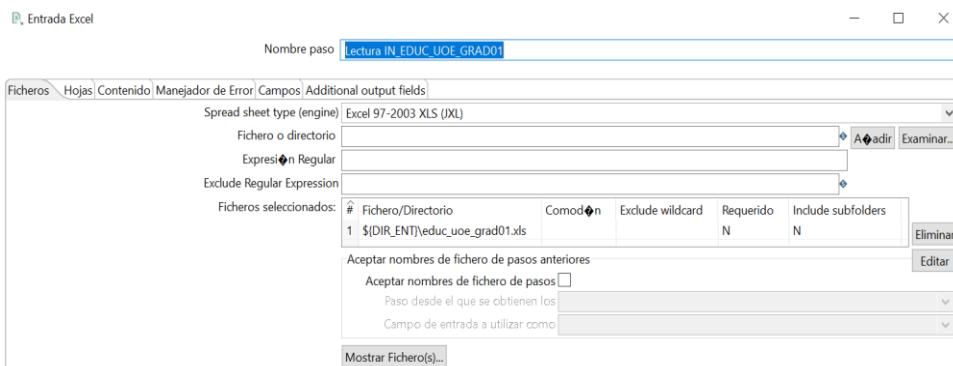
especificando la precisión de los campos numéricos y eliminando espacios en ambos lados de los campos de tipo *string*.

Entrada Excel

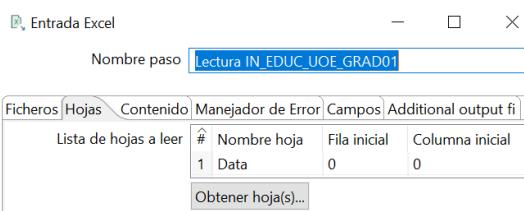
Nombre paso **Lectura IN_EDAT_LFSE_03**

Ficheros Hojas Contenido Manejador de Error Campos Additional output fields						
#	Nombre	Tipo	Longitud	Precisión	Tipo de poda	Repetir
1	GEO/TIME	String	-1	-1	ambos	N
2	2013	String	-1	-1	ninguno	N
3	2014	String	-1	-1	ninguno	N
4	2015	String	-1	-1	ninguno	N
5	2016	String	-1	-1	ninguno	N
6	2017	String	-1	-1	ninguno	N

- Paso **Entrada Excel** Lectura IN_EDUC_UOE_GRAD01, en la pestaña Ficheros añadimos el fichero Excel desde donde extraemos los datos para ello utilizamos la variable de entorno DIR_ENT.



En la pestaña Hojas, indicamos el nombre de la hoja de cálculo, Data, que queremos procesar e indicamos desde qué fila del archivo se comienzan a leer los datos, en nuestro caso después de la preparación la fila 0.



Al igual que en el paso anterior, indicamos que existe una fila de encabezados de los campos y que recupere los campos que vamos a tratar.

#	Nombre	Tipo	Longitud	Precisión	Tipo de poda	Repetir
1	GEO/TIME	String	-1	-1	ambos	N
2	2013	String	-1	-1	ninguno	N
3	2014	String	-1	-1	ninguno	N
4	2015	String	-1	-1	ninguno	N
5	2016	String	-1	-1	ninguno	N
6	2017	String	-1	-1	ninguno	N

Las fuentes de datos utilizan dos puntos ":" como indicativo de que son valores desconocidos. Estos valores, si no los transformamos, pueden dar errores al cargarlos. Lo que haremos en los dos pasos siguientes es reemplazarlos por 0 en cada uno de los ficheros, utilizando la función **Replace in string**. También cambiaremos el valor del nombre largo "Germany (until 1990 former territory of the FRG)" por el nombre corto "Germany".

- Paso **Replace in string** para los datos del fichero EDAT_LFSE_03.xls y reemplazar en las columnas 2013:2017 el carácter ":" por 0. Y en la columna GEO/TIME la cadena "Germany (until 1990 former territory of the FRG)" por "Germany".

#	In stream field	Out stream field	use RegEx	Search	Replace with
1	GEO/TIME		N	Germany (until 1990 former territory of the FRG)	Germany
2	2013		N	:	0
3	2014		N	:	0
4	2015		N	:	0
5	2016		N	:	0
6	2017		N	:	0

- Paso **Replace in string** para los datos del fichero educ_uee_grad01.xls y reemplazar en las columnas 2013:2017 el carácter ":" por 0. Y en la columna GEO/TIME la cadena "Germany (until 1990 former territory of the FRG)" por "Germany".
- Paso **Normalización de fila** sobre los datos de la primera fuente para convertir cada una de las columnas que contienen los datos de los 5 cursos académicos en las 5 filas correspondientes.

Normalizar filas

Nombre de paso: Normalización de fila

Tipo de campo: anio

Campos

#	Nombre campo	Tipo	campo nuevo
1	2013	2013	pegr
2	2014	2014	pegr
3	2015	2015	pegr
4	2016	2016	pegr
5	2017	2017	pegr

- Paso **Normalización de fila** sobre los datos de la segunda fuente para convertir cada una de las columnas que contienen los datos de los 5 cursos académicos en las 5 filas correspondientes.

Normalizar filas

Nombre de paso: Normalización de fila 2

Tipo de campo: anio

Campos

#	Nombre campo	Tipo	campo nuevo
1	2013	2013	negr
2	2014	2014	negr
3	2015	2015	negr
4	2016	2016	negr
5	2017	2017	negr

- Paso **Unión por clave**, para unir los datos de las 2 fuentes de datos normalizadas, utilizando como clave el nombre de los países, del campo GEO/TIME y el campo año (anio) obtenido de la normalización de ambas fuentes.

Unión por clave

Nombre de: Unión por clave

Primer Paso: Normalización de fila

Segundo Paso: Normalización de fila 2

Tipo Unión: INNER

Claves de primer paso: Claves de segundo paso

#	Campo clave	#	Campo clave
1	GEO/TIME	1	GEO/TIME
2	anio	2	anio

- Paso **Select Values** Convertir a numérico.

Select / Rename values

Step name: Select values

Select & Alter | Remove | Meta-data

Fields to alter the meta-data for:

#	Fieldname	Rename to	Type	Length	Precision	Binary to Normal?	Format	Date Format Lenient?
1	negr		Number		N	N	N	
2	pegr		Number		N	N	N	

- Paso **Salida de Tabla** Cargar IN_EGR_EUR para cargar los datos normalizados del número de egresados y el porcentaje de egresados jóvenes con estudios completos en la tabla intermedia IN_EGR_EUR. En la ficha *Main options*, activaremos el check 'Vaciar tabla'.

Table output

Step name: Carga IN_EGR_EUR

Connection: stage_egr

Target schema: []

Target table: IN_EGR_EUR

Commit size: 1000

Truncate table:

Ignore insert errors:

Specify database fields:

Main options | Database fields

Partition data over tables:
Partitioning field: []

Partition data per month: Partition data per day:

Use batch update for inserts:

Is the name of the table defined in a:
Field that contains name of table: []
Store the tablename field:

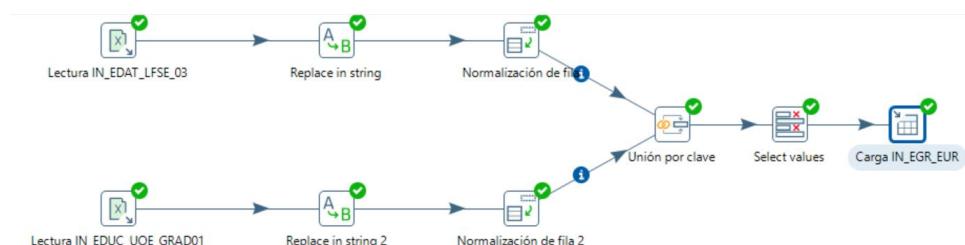
Y en la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Main options | Database fields

Fields to insert:

#	Table field	Stream field
1	PAIS	GEO/TIME
2	ANIO	anio
3	NEGR	negr
4	PEGR	pegr

La transformación completa es la siguiente:



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

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time
1	Lectura IN_EDAT_LFSE_03	0	0	12	12	0	0	0	0	Finished	2.1s
2	Replace in string	0	12	12	0	0	0	0	0	Finished	2.1s
3	Lectura IN_EDUC_UOE_GRAD01	0	0	12	12	0	0	0	0	Finished	1.2s
4	Normalización de fila	0	12	60	0	0	0	0	0	Finished	2.1s
5	Replace in string 2	0	12	12	0	0	0	0	0	Finished	2.1s
6	Normalización de fila 2	0	12	60	0	0	0	0	0	Finished	2.1s
7	Unión por clave	0	120	60	0	0	0	0	0	Finished	2.5s
8	Select values	0	60	60	0	0	0	0	0	Finished	2.5s
9	Carga IN_EGR_EUR	0	60	60	0	60	0	0	0	Finished	2.5s

Como se observa en las métricas, la transformación convierte 12 registros extraídos de los ficheros en 60 registros que se cargan en la tabla intermedia IN_EGR_EUR.

d) Bloque TR

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

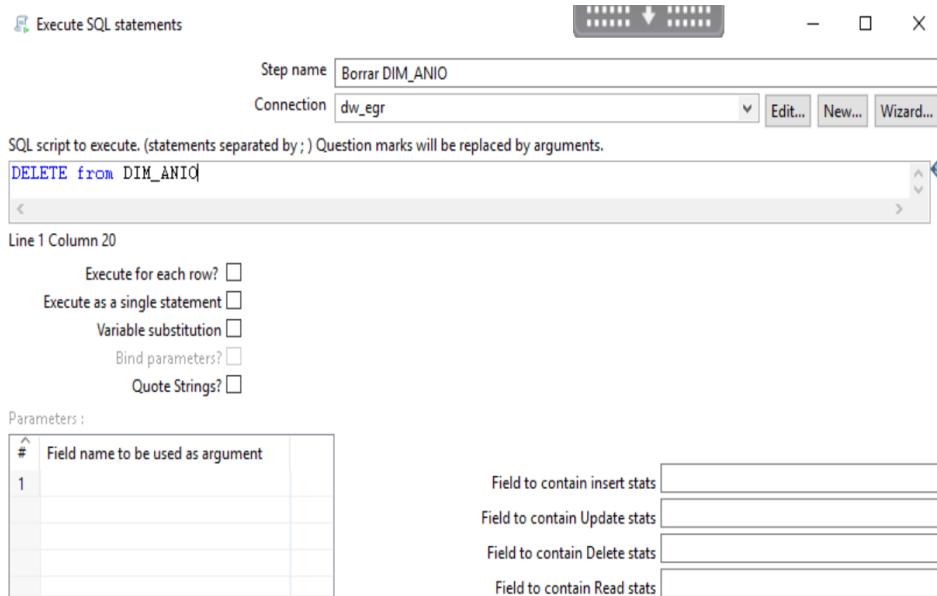
Este bloque se divide a su vez en dos sub-bloques; por un lado, los procesos para la carga de dimensiones y por otro, los procesos para la carga de tablas de hechos. Esta división permite el reprocesamiento de dichos procesos en caso de error y un mejor entendimiento de la implementación de los procesos.

1. Transformaciones Bloque TR_DIM

Este bloque, contiene las transformaciones para la carga inicial de las dimensiones al almacén desde las tablas intermedias IN_ del **Staging Area**.

Además de la carga inicial de las dimensiones del modelo multidimensional, se tendrá en cuenta que se pueda ejecutar las transformaciones las veces que sean necesarias, por esto se incluirá como primer paso de las transformaciones del bloque TR_DIM un borrado de las tablas de dimensiones. Consistirá en la ejecución de la sentencia DDL “delete from *tabla_dimension*”, donde tabla dimensión es la tabla de dimensión del modelo.

A continuación, se muestra el borrado de la dimensión DIM_ANIO como primer paso.

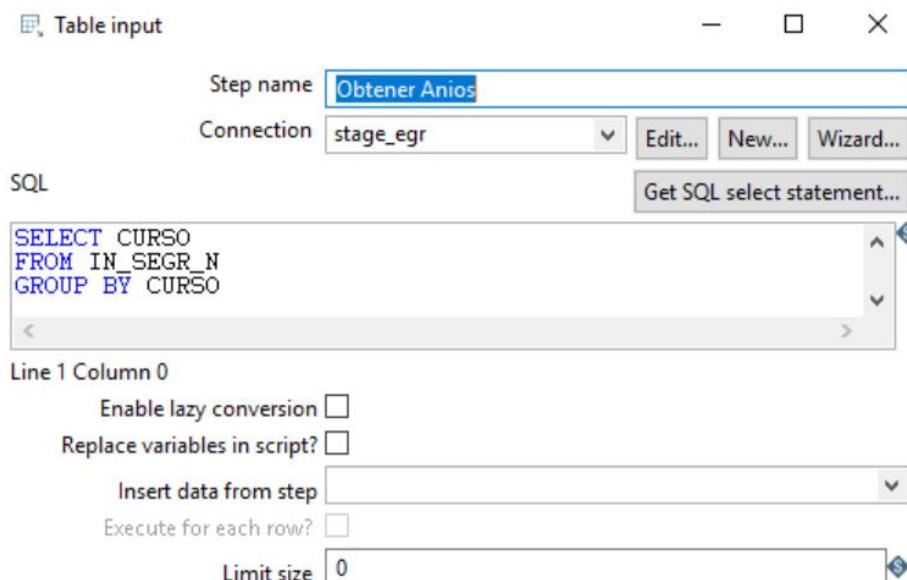


Transformación TR_DIM_ANIO

Mediante esta transformación obtendremos los valores de la dimensión temporal DIM_ANIO, utilizando los datos cargados en la tabla intermedia IN_SEGR_N.

La transformación TR_DIM_ANIO contiene 5 pasos: Borrado de dimensión, Lectura tabla intermedia, Mapeo de valores para obtención de años, Mapeo de valores para obtención de descripciones de cursos académicos y Carga de la dimensión a la tabla DIM_ANIO.

- **Paso Entrada Tabla Obtener Anios.** Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_SEGR_N para la obtención de los valores distintos en el campo CURSO.



- Paso **Mapeo de valores** para realizar una correspondencia entre los datos del campo CURSO y el nuevo campo ANIO que creamos con los valores de los años para la dimensión temporal. Introducimos los valores, tal y como se muestra a continuación.

The screenshot shows the 'Mapeo de valores' (Mapping values) dialog box. It has fields for 'Nombre de paso:' (Step name) set to 'Mapeo de valores', 'Nombre de campo' (Source field) set to 'CURSO', and 'Nombre de campo' (Target field) set to 'ANIO'. Below these are sections for 'Valores de campo:' (Mapping table) and 'Default upon'. The mapping table contains the following data:

#	Valor origen	Valor destino
1	EGR_C09_10	2011
2	EGR_C10_11	2012
3	EGR_C11_12	2013
4	EGR_C12_13	2014
5	EGR_C13_14	2015
6	EGR_C14_15	2016
7	EGR_C15_16	2017
8	EGR_C16_17	2018

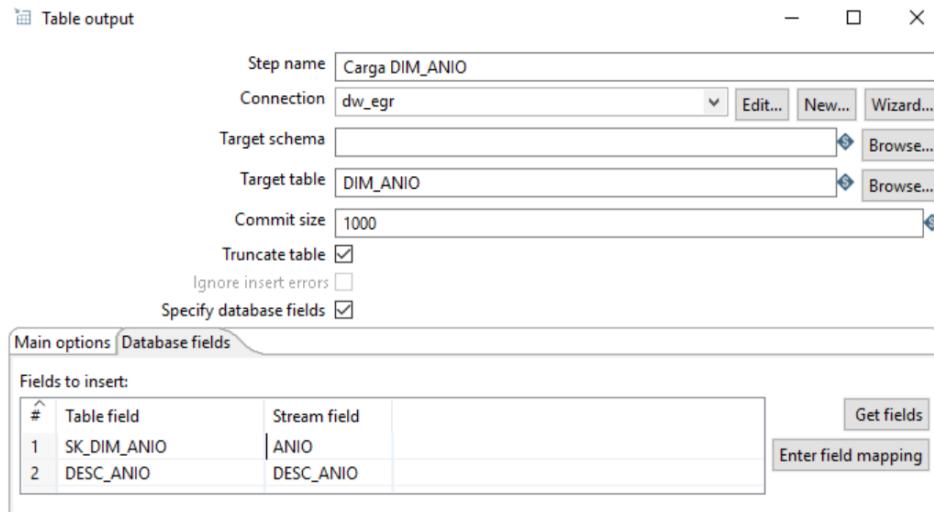
- Paso **Mapeo de valores 2** para realizar una correspondencia entre los datos del campo ANIO del paso anterior y el nuevo campo DESC_ANIO que creamos con los valores de las descripciones de los cursos académicos para la dimensión temporal. Introducimos los valores, tal y como se muestra a continuación.

The screenshot shows the 'Mapeo de valores 2' (Mapping values 2) dialog box. It has fields for 'Nombre de paso:' (Step name) set to 'Mapeo de valores 2', 'Nombre de campo' (Source field) set to 'ANIO', and 'Nombre de campo' (Target field) set to 'DESC_ANIO'. Below these are sections for 'Valores de campo:' (Mapping table) and 'Default upon'. The mapping table contains the following data:

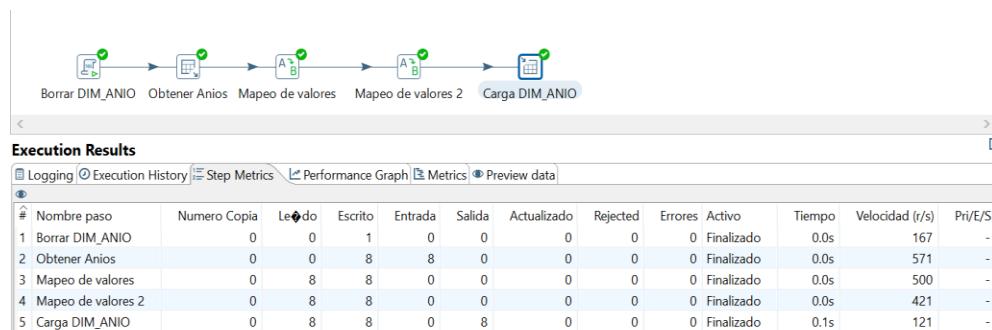
#	Valor origen	Valor destino
1	2011	C09_10
2	2012	C10_11
3	2013	C11_12
4	2014	C12_13
5	2015	C13_14
6	2016	C14_15
7	2017	C15_16
8	2018	C16_17

La creación de los campos de la dimensión DIM_ANIO planteada, mediante mapeo de valores del campo CURSO de la tabla intermedia IN_SEGR_IN es una posible opción, pero podemos utilizar otras técnicas, como inserción de valores con la función Ejecutar Sentencia SQL o directamente en la base de datos.

- Paso **Salida de Tabla** Cargar DIM_ANIO, para cargar los datos para la dimensión temporal en la tabla DIM_ANIOS del almacén, correspondiente a la conexión de base de datos dw_egr definida anteriormente. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.



La transformación completa es la siguiente y los resultados de la ejecución de la transformación es la siguiente:



Como se observa en las métricas se cargan 8 registros correspondientes a los años de los cursos académicos de los que disponemos datos en la dimensión DIM_ANIO.

Transformación TR_DIM_MODALIDAD

Mediante esta transformación obtendremos los valores para la dimensión DIM_MODALIDAD, utilizando los datos cargados en la tabla intermedia IN_SEGR_N.

La transformación TR_DIM_MODALIDAD contiene 4 pasos: Borrado de dimensión, Lectura tabla intermedia, Creación de Secuencia, Carga de la dimensión a la tabla DIM_ANIO.

- Paso **Entrada Tabla** Obtener Modalidad. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_SEGR_N para la obtención de los valores distintos en el campo MODALIDAD.

Table input

Step name: Obtener Modalidad

Connection: stage_dw

SQL:

```
SELECT MODALIDAD
FROM IN_SEGR_N
GROUP BY MODALIDAD
```

Line 1 Column 0

Enable lazy conversion:

Replace variables in script:

Insert data from step:

Execute for each row?:

Limit size: 0

incremento: 1

Valor máximo: 999999999

- Paso **Obtener valor** de la secuencia de la base de datos para obtener un campo id_modalidad mediante una secuencia de valores que comienza por 1, y que servirá de campo clave de la dimensión DIM_MODALIDAD. Otra técnica que se puede utilizar para la creación de una clave primaria, es utilizar una secuencia de base de datos.
- Paso **Salida de Tabla** Cargar DIM_MODALIDAD, para cargar los datos para la dimensión al almacén. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Table output

Step name: DIM_MODALIDAD

Connection: dw_egr

Target schema:

Target table: DIM_MODALIDAD

Commit size: 1000

Truncate table:

Ignore insert errors:

Specify database fields:

Main options Database fields

Fields to insert:

#	Table field	Stream field
1	SK_DIM_MODALIDAD	id_modalidad
2	DESC_MODALIDAD	MODALIDAD

Get fields Enter field mapping

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



Las métricas del resultado de la ejecución reflejan que se han cargado los 3 registros en la dimensión DIM_MODALIDAD, con los tipos de modalidad de impartición obtenidos de los datos que disponemos.

Execution History													
#	Nombre paso	Numero Copia	Lectura	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Borrado DIM_MODALIDAD	0	0	1	0	0	0	0	0	Finalizado	0.0s	200	-
2	Obtener Modalidad	0	0	3	3	0	0	0	0	Finalizado	0.0s	500	-
3	Secuencia Modalidad	0	3	3	0	0	0	0	0	Finalizado	0.0s	214	-
4	DIM_MODALIDAD	0	3	3	0	3	0	0	0	Finalizado	0.1s	52	-

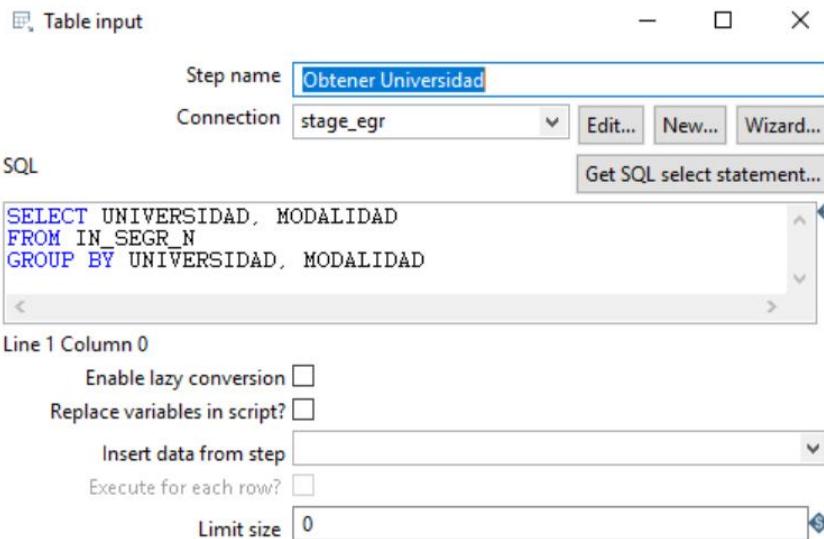
Recordar que en el diseño multidimensional, existe un modelo copo de nieve entre la dimensión DIM_UNIVERSIDAD y DIM_MODALIDAD, es por esto, que se deberá cargar la dimensión modalidad antes de la dimensión universidad.

Transformación TR_DIM_UNIVERSIDAD

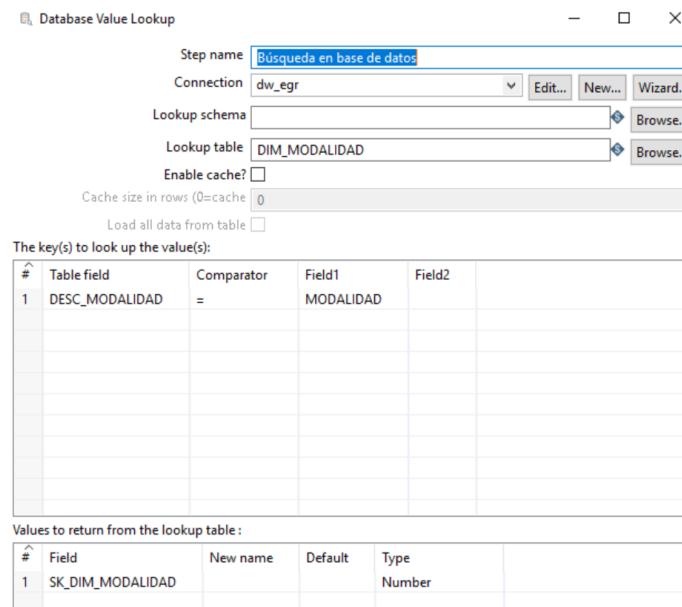
Mediante esta transformación obtendremos los valores para la dimensión DIM_UNIVERSIDAD, utilizando los datos cargados en la tabla intermedia IN_SEGR_N.

La transformación TR_DIM_UNIVERSIDAD contiene 5 pasos: Borrado de dimensión, Lectura tabla intermedia, Búsqueda en la base de datos para obtener la modalidad, Creación de Secuencia, Carga de la dimensión a la tabla DIM_ANIO.

- Paso **Entrada Tabla** Obtener Universidad. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_SEGR_N para la obtención de los valores distintos en los campos UNIVERSIDAD y MODALIDAD.



- Paso **Búsqueda en base de datos** para obtener la clave principal de la modalidad de impartición de cada universidad de la dimensión DIM_MODALIDAD creada anteriormente. Para la búsqueda se utiliza la correspondencia entre el campo DESC_MODALIDAD de la dimensión y el campo MODALIDAD obtenido en el primer paso de la transformación. Así construiremos el modelo copo de nieve entre la dimensión DIM_UNIVERSIDAD y DIM_MODALIDAD que hemos definido en el diseño.



- Paso **Obtener valor** de la secuencia de la base de datos para obtener un campo *id_universidad* mediante una secuencia de valores que comienza por 1, y que servirá de campo clave de la dimensión DIM_UNIVERSIDAD.

Obtener valor de la secuencia de la base de datos

Nombre de paso **Añadir secuencia**

Nombre de valor **ID_UNIVERSIDAD**

- Utilizar una base de datos para generar la secuencia

Utilizar base datos para obtener secuencia?

Conexión **ora_dw_egr**

Nombre de esquema

Nombre de secuencia **SEQ_**

- Utilizar un contador de la transformación para generar la secuencia

Utilizar contador para calcular secuencia?

Nombre contador (opcional)

Valor inicial **1**

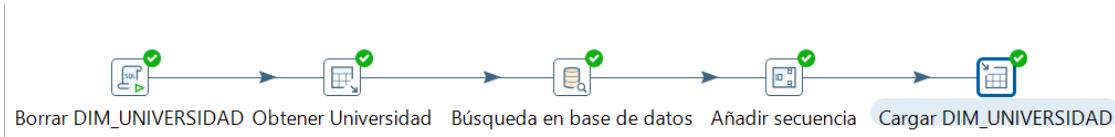
Incremento **1**

Valor máximo **999999999**

- **Paso Salida de Tabla** Cargar DIM_UNIVERSIDAD, para cargar los datos para la dimensión. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Main options		Database fields
Fields to insert:		
#	Table field	Stream field
1	SK_DIM_UNIVERSIDAD	ID_UNIVERSIDAD
2	DESC_UNIVERSIDAD	UNIVERSIDAD
3	SK_DIM_MODALIDAD	SK_DIM_MODALIDAD

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



Las métricas del resultado de la ejecución reflejan que se han cargado los 82 registros en la dimensión DIM_UNIVERSIDAD, con las universidades de las que disponemos datos de egresados.

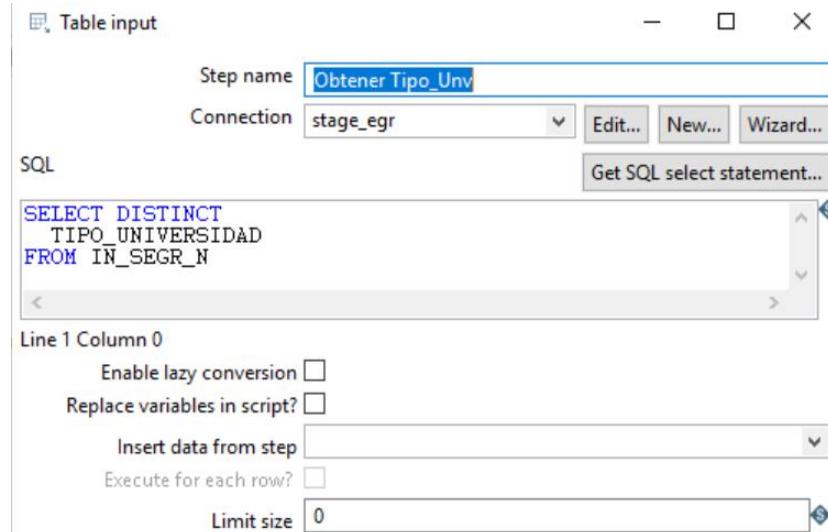
Execution History													
#	Nombre paso	Numero Copia	Lectura	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Borrar DIM_UNIVERSIDAD	0	0	1	0	0	0	0	0	Finalizado	0.0s	333	-
2	Obtener Universidad	0	0	82	82	0	0	0	0	Finalizado	0.0s	9.111	-
3	Búsqueda en base de datos	0	82	82	82	0	0	0	0	Finalizado	0.0s	1.907	-
4	Añadir secuencia	0	82	82	0	0	0	0	0	Finalizado	0.0s	1.864	-
5	Cargar DIM_UNIVERSIDAD	0	82	82	0	82	0	0	0	Finalizado	0.1s	845	-

Transformación TR_DIM_TIPO_UNIV

Mediante esta transformación obtendremos los valores para la dimensión DIM_TIPO_UNIV, utilizando los datos cargados en la tabla intermedia IN_SEGR_N.

La transformación TR_DIM_TIPO_UNIV contiene 4 pasos: Borrado de dimensión, Lectura tabla intermedia, Creación de Secuencia, Carga de la dimensión a la tabla DIM_ANIO.

- Paso **Entrada Tabla** Obtener Tipo Univ. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_SEGR_N para la obtención de los valores distintos en el campo TIPO_UNIVERSIDAD.



- Paso **Obtener valor** de la secuencia de la base de datos para obtener un campo *id_tipo_univ* mediante una secuencia de valores que comienza por 1, y que servirá de campo clave de la dimensión DIM_TIPO_UNIV.
- Paso **Salida de Tabla** Cargar DIM_TIPO_UNIV, para cargar los datos en la tabla de la dimensión DIM_TIPO_UNIV. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Fields to insert:		
#	Table field	Stream field
1	SK_DIM_TIPO_UNIV	ID_TIPO_UNIV
2	DESC_TIPO_UNIV	TIPO_UNIVERSIDAD

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



Las métricas del resultado de la ejecución reflejan que se han cargado los 2 registros en la dimensión DIM_TIPO_UNIV, con los tipos de universidad obtenidos de los datos que disponemos.

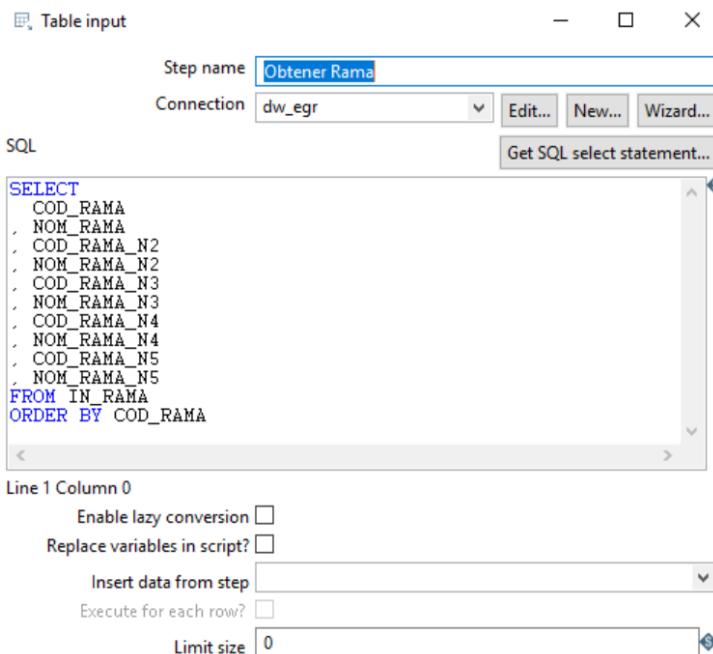
4 Cargar DIM_TIPO_UNIV	0	2	2	0	2	0	0	0 Finalizado	0.1s	27	-
------------------------	---	---	---	---	---	---	---	--------------	------	----	---

Transformación TR_DIM_RAMA

Mediante esta transformación obtendremos los valores para la dimensión DIM_RAMA, utilizando los datos cargados en la tabla intermedia IN_RAMA.

La transformación TR_DIM_RAMA contiene 4 pasos: Borrado de dimensión, Lectura tabla intermedia, Conversión a Mayúsculas, Creación de Secuencia, Carga de la dimensión a la tabla DIM_ANIO.

- **Paso Entrada Tabla** Obtener Rama. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_RAMA para la obtención de los registros. Podemos utilizar el botón *Obtener consulta SQL* para recoger la sentencia que permite extraer todos los registros con todos sus campos de la tabla IN_RAMA de la base de datos.



- **Paso String operations** para conseguir estandarizar los datos de las ramas de impartición y almacenar todos los campos en mayúsculas. En diferentes fuentes pueden venir los mismos datos escritos unas veces en mayúsculas y otras en minúsculas. Es una buena práctica estandarizar y aplicar un criterio de que todos los datos de las tablas se almacenen cumpliendo ciertas reglas y así garantizar que las búsquedas de información obtienen los resultados esperados.

String operations											
Step name String operations											
The fields to process:											
#	In stream field	Out stream field	Trim type	Lower/Upper	Padding	Pad char	Pad Length	InitCap	Escape	Digits	Remove Special character
1	COD_RAMA		none	none				N	None	none	none
2	NOM_RAMA		none	upper	none			N	None	none	none
3	COD_RAMA_N2		none	none	none			N	None	none	none
4	NOM_RAMA_N2		none	upper	none			N	None	none	none
5	COD_RAMA_N3		none	none	none			N	None	none	none
6	NOM_RAMA_N3		none	upper	none			N	None	none	none
7	COD_RAMA_N4		none	none	none			N	None	none	none
8	NOM_RAMA_N4		none	upper	none			N	None	none	none
9	COD_RAMA_N5		none	none	none			N	None	none	none
10	NOM_RAMA_N5		none	upper	none			N	None	none	none

- **Paso Obtener valor** de la secuencia de la base de datos para obtener un campo ID_RAMA mediante una secuencia de valores que comienza por 1, y que servirá de campo clave de la dimensión DIM_RAMA.
- **Paso Salida de Tabla** Cargar DIM_RAMA, para cargar los datos en la tabla de la dimensión DIM_RAMA. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Main options		Database fields
Fields to insert:		
#	Table field	Stream field
1	COD_RAMA	COD_RAMA
2	NOM_RAMA	NOM_RAMA
3	COD_RAMA_N2	COD_RAMA_N2
4	NOM_RAMA_N2	NOM_RAMA_N2
5	COD_RAMA_N3	COD_RAMA_N3
6	NOM_RAMA_N3	NOM_RAMA_N3
7	COD_RAMA_N4	COD_RAMA_N4
8	NOM_RAMA_N4	NOM_RAMA_N4
9	COD_RAMA_N5	COD_RAMA_N5
10	NOM_RAMA_N5	NOM_RAMA_N5
11	SK_DIM_RAMA	ID_RAMA

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



Las métricas del resultado de la ejecución reflejan que se han cargado los 161 registros en la dimensión DIM_RAMA, con la clasificación de ramas de estudios obtenidos de fuente de datos proporcionada.

Execution History													
#	Nombre paso	Numero Copia	Lectura	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Borrar DIM_RAMA	0	0	1	0	0	0	0	0	Finalizado	0.0s	200	-
2	Obtener Rama	0	0	161	161	0	0	0	0	Finalizado	0.0s	12.385	-
3	String operations	0	161	161	0	0	0	0	0	Finalizado	0.0s	10.062	-
4	Añadir secuencia	0	161	161	0	0	0	0	0	Finalizado	0.0s	8.050	-
5	Cargar DIM_RAMA	0	161	161	0	161	0	0	0	Finalizado	0.1s	2.477	-

Transformación TR_DESCONOCIDOS

Esta transformación se utiliza para añadir un registro en la dimensión DIM_RAMA para garantizar que no se pierden datos cuando no existen y así evitar pérdida de datos. Será un registro con clave 99999.

La transformación TR_DESCONOCIDOS contiene 3 pasos: Borrado de registro de desconocidos, Definición de registro desconocido y Carga de registro a la tabla DIM_RAMA.

- Paso **Data Grid** Cargar Valores Desconocidos. En el “Data Grid” definimos cuáles serán estos datos desconocidos que nos servirán de referencia.

The screenshot shows the 'Data grid' interface with the following details:

Nombre paso: Cargar Valores Desconocidos

Meta

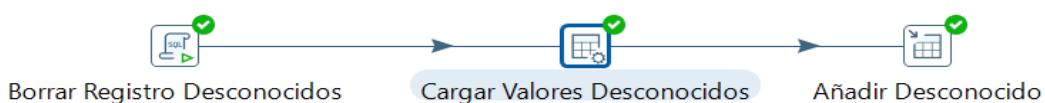
#	Name	Type
1	ID_RAMA	Integer
2	COD_RAMA	String
3	NOM_RAMA	String
4	COD_RAMA_N2	String
5	NOM_RAMA_N2	String
6	COD_RAMA_N3	String
7	NOM_RAMA_N3	String
8	COD_RAMA_N4	String
9	NOM_RAMA_N4	String
10	COD_RAMA_N5	String
11	NOM_RAMA_N5	String

Data

#	ID_RAMA	COD_RAMA	NOM_RAMA	COD_RAMA_N2	NOM_RAMA_N2	COD_RAMA_N3	NOM_RAMA_N3	COD_RAMA_N4	NOM_RAMA_N4	COD_RAMA_N5	NOM_RAMA_N5
1	99999	9	Desconocido	99	Desconocido	999	Desconocido	99999	Desconocido	99999	Desconocido

- Paso **Salida de Tabla** Añadir Desconocido, para el registro de Desconocidos en la tabla de la dimensión DIM_RAMA.

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



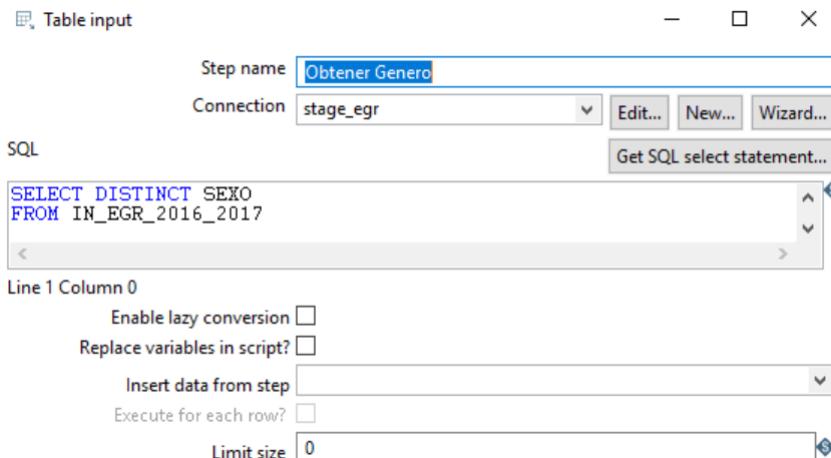
Transformación TR_DIM_PERFIL

Mediante esta transformación obtendremos los valores para las dimensiones DIM_SEXO y DIM_EDAD, utilizando los datos cargados en la tabla intermedia IN_EGR_2016_2017.

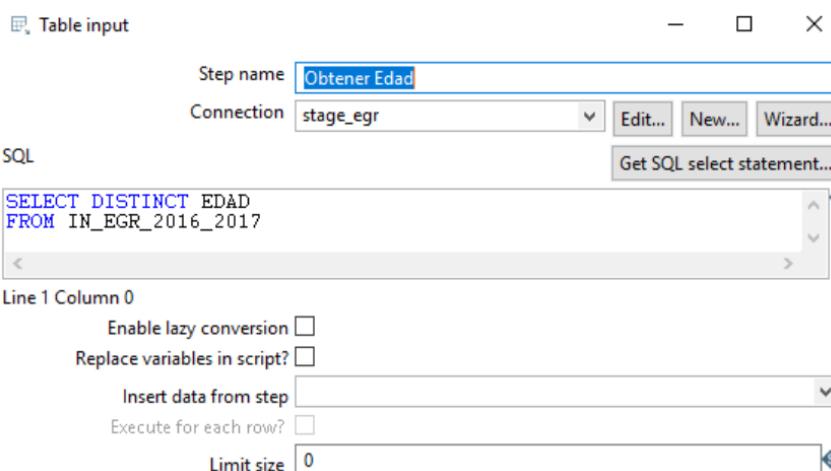
En lugar de una única transformación, pueden crearse dos transformaciones distintas. Se ha optado por implementar una única por simplicidad.

La transformación TR_DIM_PERFIL contiene 8 pasos: 2 Borrados de dimensiones, 2 Lecturas de tabla intermedia, 2 Creaciones de Secuencia y 2 Cargas de las dimensiones DIM_SEXO y DIM_EDAD.

- **Paso Entrada Tabla** Obtener Sexo, lectura de la tabla intermedia IN_EGR_2016_2017 para la obtención de los distintos valores del campo SEXO.



- **Paso Entrada Tabla** Obtener Edad, lectura de la tabla intermedia IN_EGR_2016_2017 para la obtención de los distintos valores del campo EDAD.



- Paso **Obtener valor** de la secuencia de la base de datos para obtener un campo ID_SEXO mediante una secuencia de valores que comienza por 1, y que servirá de campo clave de la dimensión DIM_SEXO.
- Paso **Value Mapper** para obtener un campo ID_EDAD mediante la asignación de valores según el intervalo de edad, para que queden los valores ordenados, y que servirá de campo clave de la dimensión DIM_EDAD.

The screenshot shows the 'Value Mapper' step configuration. The 'Step name:' field is set to 'Value Mapper'. The 'Fieldname to use:' dropdown is set to 'EDAD'. The 'Target field name (empty=overwrite):' input is set to 'ID_EDAD'. The 'Default upon non-matching:' field is empty. Below these settings is a table titled 'Field values:' with four rows:

#	Source value	Target value
1	Menos de 25 años	1
2	De 25 a 30 años	2
3	De 31 a 40 años	3
4	Más de 40 años	4

- Paso **Salida de Tabla** Cargar DIM_SEXO, para cargar los datos en la tabla de la dimensión DIM_SEXO. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

The screenshot shows the 'Database fields' configuration for the 'SK_DIM_SEXO' field. The 'Table field' is 'SK_DIM_SEXO' and the 'Stream field' is 'ID_SEXO'. There is another row for 'DESC_SEXO' and 'SEXO'.

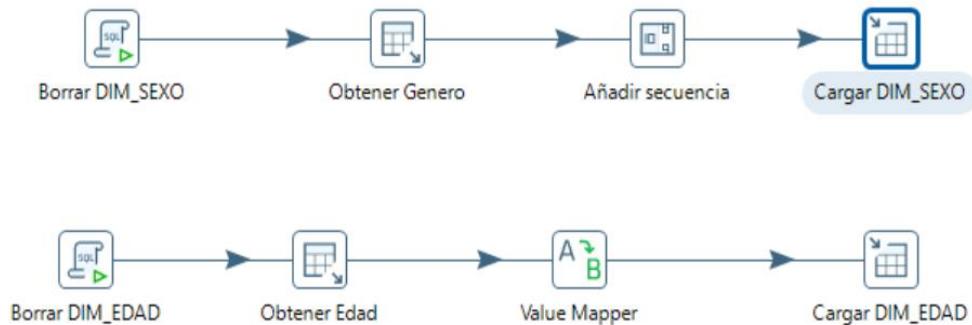
#	Table field	Stream field
1	SK_DIM_SEXO	ID_SEXO
2	DESC_SEXO	SEXO

- Paso **Salida de Tabla** Cargar DIM_EDAD, para cargar los datos en la tabla de la dimensión DIM_EDAD. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

The screenshot shows the 'Database fields' configuration for the 'SK_DIM_EDAD' field. The 'Table field' is 'SK_DIM_EDAD' and the 'Stream field' is 'ID_EDAD'. There is another row for 'DESC_INT_EDAD' and 'EDAD'.

#	Table field	Stream field
1	SK_DIM_EDAD	ID_EDAD
2	DESC_INT_EDAD	EDAD

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



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

#	Nombre paso	Numero Copia	Lectura	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Borrar DIM_SEXO	0	0	1	0	0	0	0	0	Finalizado	0.0s	200	-
2	Borrar DIM_EDAD	0	0	1	0	0	0	0	0	Finalizado	0.0s	200	-
3	Obtener Genero	0	0	2	2	0	0	0	0	Finalizado	0.0s	333	-
4	Añadir secuencia	0	2	2	0	0	0	0	0	Finalizado	0.0s	125	-
5	Obtener DIM_SEXO	0	2	2	0	2	0	0	0	Finalizado	0.1s	28	-
6	Obtener Edad	0	0	4	4	0	0	0	0	Finalizado	0.0s	667	-
7	Añadir secuencia 2	0	4	4	0	0	0	0	0	Finalizado	0.0s	235	-
8	Obtener DIM_EDAD	0	4	4	0	4	0	0	0	Finalizado	0.1s	58	-

Se observa que se han insertado 2 registros en la DIM_SEXO y 4 en la DIM_EDAD.

Transformación TR_DIM_PAIS

Mediante esta transformación obtendremos los valores de la dimensión temporal DIM_PAIS, utilizando los datos de países que contiene la tabla intermedia IN_EGR_EUR.

La transformación TR_DIM_PAIS contiene 4 pasos: Lectura tabla intermedia, Mapeo de valores para obtención de país, Obtención de Secuencia y Carga de la dimensión a la tabla DIM_PAIS.

- Paso **Entrada Tabla** Obtener País. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_EGR_EUR para la obtención de los valores distintos en el campo PAIS.

Table input

Step name: Entrada Tabla

Connection: stage_egr

SQL:

```
SELECT DISTINCT PAIS
FROM IN_EGR_EUR
```

Line 1 Column 0

Enable lazy conversion

Replace variables in script?

Insert data from step

Execute for each row?

Limit size: 0

- Paso **Mapeo de valores** para realizar una correspondencia entre los datos del campo PAIS y el nuevo campo PAIS_ES que creamos con los nombres de los países en castellano. Introducimos los valores, tal y como se muestra a continuación.

Value Mapper

Step name: Mapeo de valores

Fieldname to use: PAIS

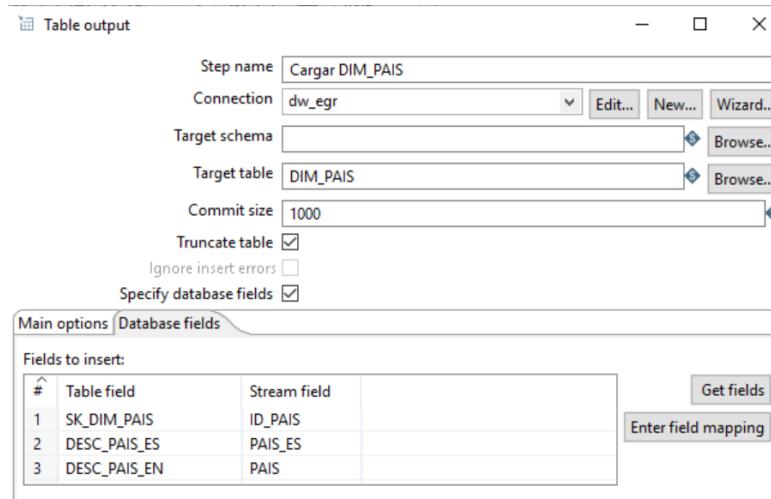
Target field name (empty=overwrite): PAIS_ES

Default upon non-matching:

Field values:

#	Source value	Target value
1	Denmark	Dinamarca
2	Germany	Alemania
3	Ireland	Irlanda
4	Spain	España
5	France	Francia
6	Italy	Italia
7	Netherlands	Paises Bajos
8	Finland	Finlandia
9	Sweden	Suecia
10	United Kingdom	Reino Unido
11	Norway	Noruega
12	Switzerland	Suiza

- Paso **Obtener valor** de la secuencia de la base de datos para obtener un campo ID_PAIS mediante una secuencia de valores que comienza por 1, y que servirá de campo clave de la dimensión DIM_PAIS.
- Paso **Salida de Tabla** Cargar DIM_PAIS, para cargar los datos en la tabla DIM_PAIS del almacén. En la ficha Database fields, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.



La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



2. Transformaciones Bloque TR_FACT

Este bloque, contiene las transformaciones para la carga inicial de las tablas de hecho al almacén desde las tablas intermedias /N_ del **Staging Area**.

Con la implementación y ejecución de los procesos de carga de dimensiones tendremos pobladas de datos nuestro modelo dimensional y podemos ahora pasar a poblar el modelo de hechos, haciendo referencia a las dimensiones disponibles, mediante sus claves foráneas.

Transformación TR_FACT_PEGR_EVOLUTIVO

Mediante esta transformación poblaremos la tabla de hechos FACT_PEGR_EVOLUTIVO según el modelo multidimensional definido y que nos permitirá realizar un análisis evolutivo de las personas egresadas por año, tipo de universidad, universidad y rama de estudios.

La transformación TR_FACT_PEGR_EVOLUTIVO contiene 7 pasos: Lectura tabla intermedia IN_SEGR_N, Obtención de clave foránea DIM_ANIO, Obtención clave foránea DIM_TIPO_UNIV, Obtención clave foránea DIM_UNIVERSIDAD, Obtener clave foránea DIM_RAMA, Tratamiento de Desconocidos y Carga FACT_PEGR_EVOLUTIVO.

- Paso **Entrada Tabla** Lectura IN_SEGR_N. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_SEGR_N para la obtención de todos los registros y todos los campos de la tabla.

Se utiliza la función SQL `UPPER(RAMA_ENSEÑANZA)`, para extraer los datos de las ramas de estudio en mayúsculas, tal y como tenemos la información en la dimensión `DIM_RAMA`.

Se utiliza la función `SUBSTRING(CURSO,5,10)`, para extraer la subcadena desde el carácter 5 hasta el 10 del campo `CURSO`, y así obtener la cadena con el curso académico. Por ejemplo, el registro con valor del campo `CURSO` “`EGR_C16_17`”, con esta función obtenemos la subcadena “`C16_17`”, que se corresponden con el curso académico.

Table input

Step name: Lectura IN_SEGR_N

Connection: stage_egr

SQL:

```
SELECT
    TIPO_UNIVERSIDAD
, MODALIDAD
, UNIVERSIDAD
, UPPER(RAMA_ENSEÑANZA) RAMA_ENSEÑANZA
, SUBSTRING(CURSO,5,10) CURSO
, NEGR
FROM IN_SEGR_N
```

Line 1 Column 0

Enable lazy conversion

Replace variables in script?

Insert data from step

Execute for each row?

Limit size: 0

- Paso **Búsqueda en base de datos** Obtener `ANIO`, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión `DIM_ANIO`. Para la búsqueda se utiliza la correspondencia entre el campo `DESC_ANIO` de la dimensión y el campo `CURSO` obtenido en el primer paso de la transformación y obtendremos el campo `SK_DIM_ANIO`, que es la clave principal de la dimensión `DIM_ANIO`.

Database Value Lookup

Step name: Obtener ANIO

Connection: dw_egr

Lookup schema:

Lookup table: DIM_ANIO

Enable cache?

Cache size in rows (0=cache): 0

Load all data from table

The key(s) to look up the value(s):

#	Table field	Comparator	Field1	Field2
1	DESC_ANIO	=	CURSO	

Values to return from the lookup table :

#	Field	New name	Default	Type
1	SK_DIM_ANIO			None

- Paso **Búsqueda en base de datos** Obtener TIPO_UNIV, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_TIPO_UNIV. Para la búsqueda se utiliza la correspondencia entre el campo DESC_TIPO_UNIV de la dimensión y el campo TIPO_UNIVERSIDAD obtenido en el primer paso de la transformación y obtendremos el campo SK_DIM_TIPO_UNIV, que es la clave principal de la dimensión DIM_TIPO_UNIV.

Database Value Lookup

Step name: Obtener TIPO_UNIV

Connection: dw_egr

Lookup schema:

Lookup table: DIM_TIPO_UNIV

Enable cache?

Cache size in rows (0=cache)

Load all data from table

The key(s) to look up the value(s):

#	Table field	Comparator	Field1	Field2
1	DESC_TIPO_UNIV	=	TIPO_UNIVERSIDAD	
2				

Values to return from the lookup table :

#	Field	New name	Default	Type
1	SK_DIM_TIPO_UNIV			None

- Paso **Búsqueda en base de datos** Obtener Universidad, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_UNIVERSIDAD. Para la búsqueda se utiliza la correspondencia entre el campo *DESC_UNIVERSIDAD* de la dimensión y el campo *UNIVERSIDAD* obtenido en el primer paso de la transformación y obtendremos el campo *SK_DIM_UNIVERSIDAD*, que es la clave principal de la dimensión DIM_UNIVERSIDAD.

Database Value Lookup

Step name: Obtener Universidad

Connection: dw_egr

Lookup schema:

Lookup table: DIM_UNIVERSIDAD

Enable cache?

Cache size in rows (0=cache): 0

Load all data from table:

The key(s) to look up the value(s):

#	Table field	Comparador	Field1	Field2
1	DESC_UNIVERSIDAD	=	UNIVERSIDAD	

Values to return from the lookup table:

#	Field	New name	Default	Type
1	SK_DIM_UNIVERSIDAD			None

- Paso **Búsqueda en base de datos** Obtener Rama, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_RAMA. Para la búsqueda se utiliza la correspondencia entre el campo *NOM_RAMA* de la dimensión y el campo *RAMA_ENSEÑANZA* obtenido en el primer paso de la transformación y obtendremos el campo *COD_RAMA*, que corresponde al primer nivel de la dimensión DIM_RAMA.

The key(s) to look up the value(s):

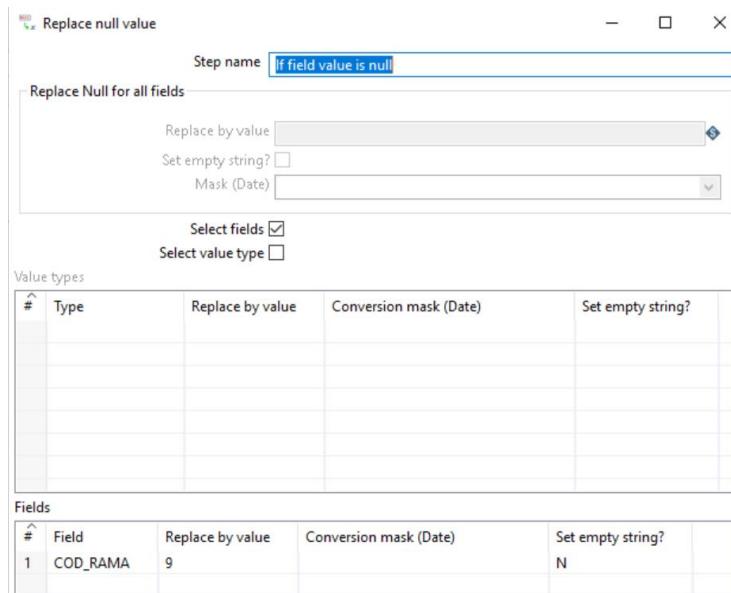
#	Table field	Comparador	Field1	Field2
1	NOM_RAMA	=	RAMA_ENSEÑANZA	

Values to return from the lookup table:

#	Field	New name	Default	Type
1	COD_RAMA			None

- Paso **If field value is null**, para tratamiento de valores desconocidos de datos relativos a las ramas de conocimiento que pueden venir en las fuentes de datos de origen y que no se encuentran en la dimensión DIM_RAMA. Cuando existan registros con referencias a ramas de conocimiento inexistentes (valor *null*), no descartaremos los datos en la tabla de hechos, pero los distinguiremos porque tendrán en el atributo

COD_RAMA el valor 9, que corresponde al registro de desconocidos, que inserta la transformación TR_DESCONOCIDOS.



- Paso **Salida de Tabla** Cargar FACT_PEGR_EVOLUTIVO, para cargar los datos en la tabla de hechos FACT_PEGR_EVOLUTIVO. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Main options		Database fields	
Fields to insert:			
#	Table field	Stream field	
1	SK_DIM_ANIO	SK_DIM_ANIO	
2	SK_DIM_TIPO_UNIV	SK_DIM_TIPO_UNIV	
3	SK_DIM_UNIVERSIDAD	SK_DIM_UNIVERSIDAD	
4	COD_RAMA	COD_RAMA	
5	PERSONAS_EGRESADAS	NEGR	

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



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

#	Nombre paso	Numero Copia	Lecto	Escrito	Entrada	Salida	Actualizado	Rejected	Errores Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Lectura IN_SEGR_N	0	0	3280	3280	0	0	0	0 Finalizado	0.2s	15.115	-
2	Obtener ANIO	0	3280	3280	3280	0	0	0	0 Finalizado	1.4s	2.384	-
3	Obtener TIPO_UNIV	0	3280	3280	3280	0	0	0	0 Finalizado	1.4s	2.308	-
4	Obtener Universidad	0	3280	3280	3280	0	0	0	0 Finalizado	1.5s	2.257	-
5	Obtener Rama	0	3280	3280	3280	0	0	0	0 Finalizado	1.5s	2.215	-
6	If field value is null	0	3280	3280	0	0	0	0	0 Finalizado	1.5s	2.209	-
7	Cargar FACT_PEGR_EVOLUTIVO	0	3280	3280	0	3280	0	0	0 Finalizado	1.5s	2.158	-

De las métricas se desprende que se han insertado 3.280 registros en la tabla de hechos FACT_PERG_EVOLUTIVO, siguiendo el modelo multidimensional diseñado que nos permitirá realizar análisis de las personas egresadas desde varios puntos de vista como tiempo, rama de estudios y universidades.

Transformación TR_FACT_PEGR_PERFIL

Mediante esta transformación poblaremos la tabla de hechos FACT_PEGR_PERFIL según el modelo multidimensional definido y que nos permitirá realizar un análisis de las características personales de las personas egresadas como la edad y género.

La transformación TR_FACT_PEGR_PERFIL contiene 6 pasos: Lectura tabla intermedia IN_EGR_2016_2017, Obtener clave primaria DIM_RAMA, Obtención clave foránea DIM_SEXO, Obtención clave foránea DIM_EDAD, Tratamiento de Desconocidos y Carga FACT_PEGR_PERFIL.

- **Paso Entrada Tabla** Lectura IN_EGR_2016_2017. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_EGR_2016_2017 para la obtención de los registros de la tabla.

Como primer campo a extraer, será el valor ANIO=2018 por tratarse de datos de las personas egresadas del curso académico 2016-2017.

Se utiliza la función *SUBSTRING(COD_AMBITO,1,4)*, para extraer la subcadena con el código de la rama de estudios del campo COD_AMBITO. Por ejemplo, el registro con valor del campo COD_AMBITO “0831 - Pesca”, con esta función obtenemos la subcadena “0831”.

Se añade un campo calculado NUM_EGR con la suma del número de personas egresadas del nivel grado universitario y del número de personas egresadas del nivel master universitario.

Table input

Step name: Lectura EGR_C2016-2017

Connection: stage_egr

SQL:

```
SELECT
  '2018' ANIO
, SUBSTRING(COD_AMBITO,1,4) COD_RAMA_N4
, SEXO
, EDAD
, (NUM_EGR_NV1 + NUM_EGR_NV2) NUM_EGR
FROM IN_EGR_2016_2017
```

Line 1 Column 0

Enable lazy conversion

Replace variables in script?

Insert data from step:

Execute for each row?

Limit size: 0

- Paso **Búsqueda en base de datos** Obtener Rama, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_RAMA. Para la búsqueda se utiliza la correspondencia entre el campo COD_RAMA_N4 de la dimensión y el campo COD_RAMA_N4 obtenido en el primer paso de la transformación y obtendremos el campo SK_DIM_RAMA, que es la clave principal de la dimensión DIM_RAMA.

Database Value Lookup

Step name: Obtener Rama

Connection: dw_egr

Lookup schema:

Lookup table: DIM_RAMA

Enable cache?

Cache size in rows (0=cache): 0

Load all data from table

The key(s) to look up the value(s):

#	Table field	Comparator	Field1	Field2
1	COD_RAMA_N4	=	COD_RAMA_N4	

Values to return from the lookup table:

#	Field	New name	Default	Type
1	COD_RAMA_N4			None

- Paso **Búsqueda en base de datos** Obtener Género, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_SEXO. Para la búsqueda se utiliza la correspondencia entre el campo DESC_SEXO de la dimensión y el campo SEXO obtenido en el

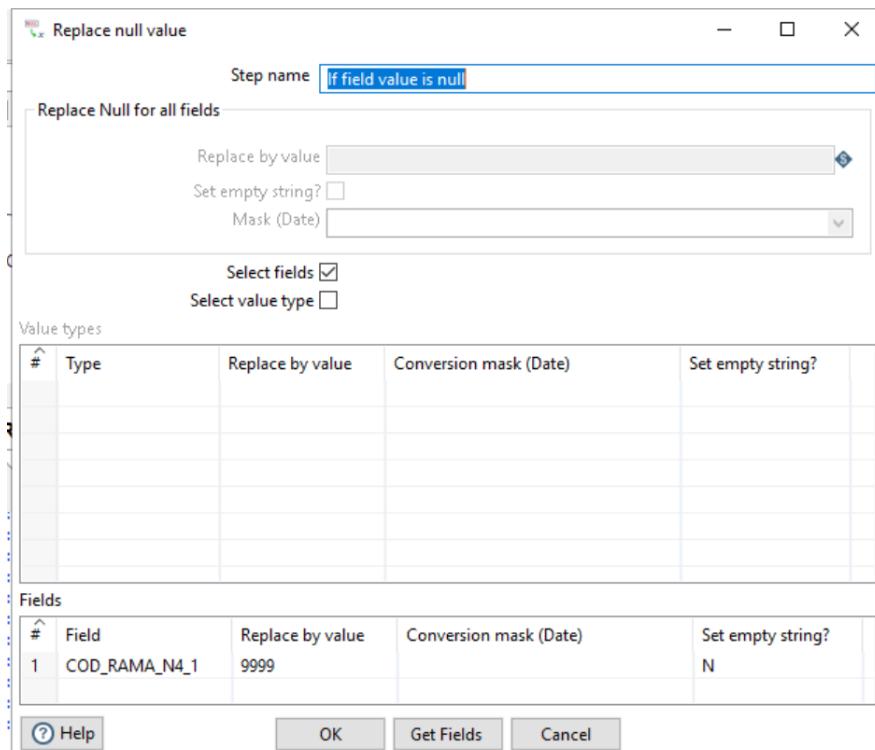
primer paso de la transformación y obtendremos el campo **SK_DIM_SEXO**, que es la clave principal de la dimensión DIM_SEXO.

The screenshot shows the configuration for a Database Value Lookup step named 'Obtener Genero'. The connection is set to 'dw_egr'. The lookup schema is empty, and the lookup table is 'DIM_SEXO'. The 'Enable cache?' checkbox is unchecked. The 'Cache size in rows (0=cache)' field contains '0'. The 'Load all data from table' checkbox is unchecked. Below these settings, there are two tables. The first table, 'The key(s) to look up the value(s):', has one row with '#': 1, 'Table field': 'DESC_SEXO', 'Comparator': '=', 'Field1': 'SEXO', and 'Field2': empty. The second table, 'Values to return from the lookup table:', has one row with '#': 1, 'Field': 'SK_DIM_SEXO', 'New name': empty, 'Default': empty, and 'Type': 'Integer'.

- **Paso Búsqueda en base de datos** Obtener Edad, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_EDAD. Para la búsqueda se utiliza la correspondencia entre el campo *DESC_INT_EDAD* de la dimensión y el campo *EDAD* obtenido en el primer paso de la transformación y obtendremos el campo **SK_DIM_EDAD**, que es la clave principal de la dimensión DIM_EDAD.

The screenshot shows the configuration for a Database Value Lookup step named 'Obtener Edad'. The connection is set to 'dw_egr'. The lookup schema is empty, and the lookup table is 'DIM_EDAD'. The 'Enable cache?' checkbox is unchecked. The 'Cache size in rows (0=cache)' field contains '0'. The 'Load all data from table' checkbox is unchecked. Below these settings, there are two tables. The first table, 'The key(s) to look up the value(s):', has one row with '#': 1, 'Table field': 'DESC_INT_EDAD', 'Comparator': '=', 'Field1': 'EDAD', and 'Field2': empty. The second table, 'Values to return from the lookup table:', has one row with '#': 1, 'Field': 'SK_DIM_EDAD', 'New name': empty, 'Default': empty, and 'Type': 'None'.

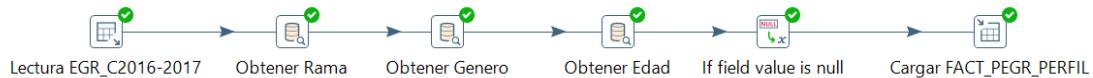
- Paso **If field value is null**, para tratamiento de valores desconocidos de en el nivel 4 de la clasificación de las ramas de conocimiento que pueden venir en las fuentes de datos de origen y que no se encuentran en la dimensión DIM_RAMA. Cuando existan registros con referencias al nivel 4 de la clasificación de las ramas de conocimiento inexistentes (valor *null*), no descartaremos los datos en la tabla de hechos, pero los distinguiremos porque tendrán en el atributo COD_RAMA_N4 el valor 9999, que corresponde al registro de desconocidos, que inserta la transformación TR_DESCONOCIDOS.



- Paso **Salida de Tabla** Cargar FACT_PEGR_PERFIL, para cargar los datos en la tabla de hechos FACT_PEGR_PERFIL. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Fields to insert:		
#	Table field	Stream field
1	SK_DIM_ANIO	ANIO
2	PERSONAS_EGRESADAS	NUM_EGR
3	SK_DIM_EDAD	SK_DIM_EDAD
4	COD_RAMA_N4	COD_RAMA_N4_1
5	SK_DIM_SEXO	SK_DIM_SEXO

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



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

#	Nombre paso	Numero Copia	Lectdo	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Lectura EGR_C2016-2017	0	0	712	712	0	0	0	0	Finalizado	0.0s	22.968	-
2	Obtener Rama	0	712	712	704	0	0	0	0	Finalizado	0.2s	4.368	-
3	Obtener Genero	0	712	712	712	0	0	0	0	Finalizado	0.2s	3.828	-
4	Obtener Edad	0	712	712	712	0	0	0	0	Finalizado	0.2s	3.473	-
5	If field value is null	0	712	712	0	0	0	0	0	Finalizado	0.2s	3.440	-
6	Cargar FACT_PEGR_PERFIL	0	712	712	0	712	0	0	0	Finalizado	0.2s	2.979	-

De las métricas se desprende que se han insertado 712 registros en la tabla de hechos FACT_PEGR_PERFIL, siguiendo el modelo multidimensional diseñado que nos permitirá realizar análisis de las características de las personas egresadas como el sexo y la edad.

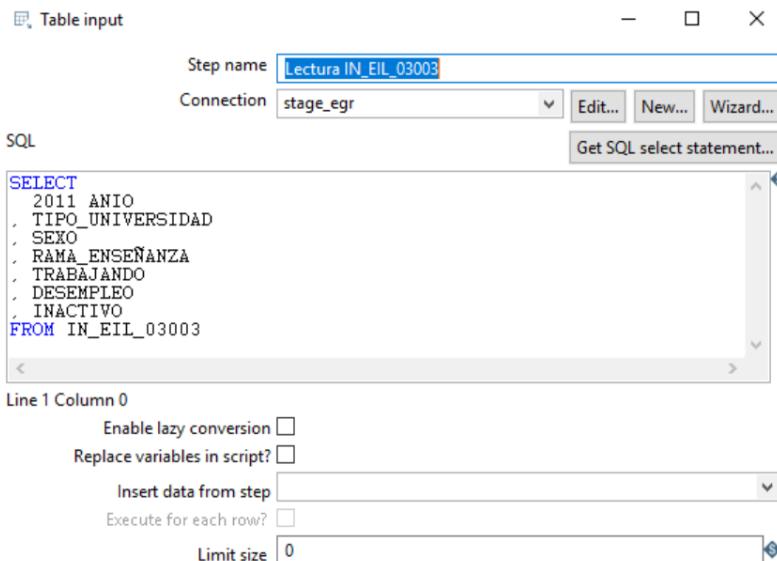
Transformación TR_FACT_PEGR_INSERTADAS

Mediante esta transformación poblaremos la tabla de hechos FACT_PEGR_INSERTADAS según el modelo multidimensional definido y que nos permitirá realizar un análisis evolutivo de las personas egresadas por año, tipo de universidad, universidad y rama de estudios.

La transformación FACT_PEGR_INSERTADAS contiene 6 pasos: Lectura tabla intermedia IN_EIL_03003; Obtención clave foránea DIM_TIPO_UNIV; Obtención clave foránea DIM_SEXO; Obtener DIM_RAMA, Tratamiento de Desconocidos y Carga FACT_PEGR_INSERTADAS.

- **Paso Entrada Tabla** Lectura IN_EIL_03003. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_EIL_03003 para la obtención de los registros de la tabla.

Como primer campo a extraer será el valor ANIO=2011 por tratarse de datos de las personas egresadas insertadas en 2014 del curso académico 2019-2010.



- **Paso Búsqueda en base de datos** Obtener TIPO_UNV, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_TIPO_UNIV. Para la búsqueda se utiliza la correspondencia entre el campo DESC_TIPO_UNIV de la dimensión y el campo TIPO_UNIVERSIDAD obtenido en el primer paso de la transformación y obtendremos el campo SK_DIM_TIPO_UNIV, que es la clave principal de la dimensión DIM_TIPO_UNIV.
- **Paso Búsqueda en base de datos** Obtener Género, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_SEXO. Para la búsqueda se utiliza la correspondencia entre el campo DESC_SEXO de la dimensión y el campo SEXO obtenido en el primer paso de la transformación y obtendremos el campo SK_DIM_SEXO, que es la clave principal de la dimensión DIM_SEXO.
- **Paso Búsqueda en base de datos** Obtener Rama, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_RAMA. Para la búsqueda se utiliza la correspondencia entre el campo NOM_RAMA de la dimensión y el campo RAMA_ENSEÑANZA obtenido en el primer paso de la transformación y obtendremos el campo COD_RAMA, que corresponde al primer nivel de la dimensión DIM_RAMA.
- **Paso If field value is null**, para tratamiento de valores desconocidos de datos relativos a las ramas de conocimiento que pueden venir en las fuentes de datos de origen y que no se encuentran en la dimensión DIM_RAMA. Cuando existan registros con referencias a ramas de conocimiento inexistentes (valor null), no descartaremos los datos en la tabla de hechos, pero los distinguiremos porque tendrán en la clave foránea SK_DIM_RAMA el valor 99999, que corresponde al registro de desconocidos, que inserta la transformación TR_DESCONOCIDOS.
- **Paso Salida de Tabla** Cargar FACT_PERG_INSERTADAS, para cargar los datos en la tabla de hechos FACT_PERG_INSERTADAS. En la ficha

Database fields, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Fields to insert:		
#	Table field	Stream field
1	SK_DIM_ANIO	ANIO
2	SK_DIM_TIPO_UNIV	SK_DIM_TIPO_UNIV
3	SK_DIM_SEXO	SK_DIM_SEXO
4	COD_RAMA	COD_RAMA
5	PEGR_TRABAJANDO	TRABAJANDO
6	PEGR_DESEMPLEADOS	DESEMPLEO
7	PEGR_INACTIVOS	INACTIVO

La transformación completa es la siguiente y los resultados de la ejecución de la transformación es la siguiente:



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

#	Nombre paso	Numero Copia	Lecto	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Lectura IN_EIL_03003	0	0	30	30	0	0	0	0	Finalizado	0.0s	1.875	-
2	Obtener Tipo_Univ	0	30	30	30	0	0	0	0	Finalizado	0.0s	1.000	-
3	Obtener Genero	0	30	30	30	0	0	0	0	Finalizado	0.0s	714	-
4	Obtener Rama	0	30	30	30	0	0	0	0	Finalizado	0.1s	588	-
5	If field value is null	0	30	30	0	0	0	0	0	Finalizado	0.1s	545	-
6	Cargar FACT_PEGR_INSERTADAS	0	30	30	0	30	0	0	0	Finalizado	0.1s	341	-

De las métricas se desprende que se han insertado 30 registros en la tabla de hechos FACT_PEGR_INSERTADAS, siguiendo el modelo multidimensional diseñado que nos permitirá realizar análisis de las personas egresadas insertadas por sexo, tipo de universidad y rama de estudios.

Transformación TR_FACT_PEGR_EUR

Mediante esta transformación poblaremos la tabla de hechos FACT_PEGR_EUR según el modelo multidimensional definido y que nos permitirá realizar un análisis comparativo de las de las personas egresadas entre países europeos.

La transformación TR_FACT_PEGR_EUR contiene 3 pasos: Lectura tabla intermedia IN_EGR_EUR, Obtener clave primaria DIM_PAIS, y Carga FACT_PEGR_EUR.

- **Paso Entrada Tabla** Lectura IN_EGR_EUR. Este primer paso de la transformación consiste en la lectura de la tabla intermedia IN_EGR_EUR para la obtención de los registros de la tabla.

Table input

Step name: Lectura EGR_EUR

Connection: stage_egr

SQL:

```
SELECT
    PAIS
, ANIO
, NEGR
, PEGR
FROM IN_EGR_EUR
```

Line 1 Column 0

Enable lazy conversion

Replace variables in script?

Insert data from step

Execute for each row?

Limit size: 0

- **Paso Búsqueda en base de datos** Obtener País, para obtener de cada registro de la tabla intermedia, la clave principal de la dimensión DIM_PAIS. Para la búsqueda se utiliza la correspondencia entre el campo DESC_PAIS_EN de la dimensión y el campo PAIS obtenido en el primer paso de la transformación y obtendremos el campo SK_DIM_PAIS, que es la clave principal de la dimensión DIM_PAIS.

Database Value Lookup

Step name: Obtener País

Connection: dw_egr

Lookup schema:

Lookup table: DIM_PAIS

Enable cache?

Cache size in rows (0=cache): 0

Load all data from table

The key(s) to look up the value(s):

#	Table field	Comparator	Field1	Field2
1	DESC_PAIS_EN	=	PAIS	

Values to return from the lookup table:

#	Field	New name	Default	Type
1	SK_DIM_PAIS			None

- **Paso Salida de Tabla** Cargar FACT_PEGR_EUR, para cargar los datos en la tabla de hechos FACT_PEGR_EUR. En la ficha *Database fields*, indicamos la relación de campos entre la tabla y los campos obtenidos en la transformación.

Fields to insert:		
#	Table field	Stream field
1	SK_DIM_ANIO	ANIO
2	SK_DIM_PAIS	SK_DIM_PAIS
3	PERSONAS_EGRESADAS	NEGR
4	PORCENTAJE_PEGR_JOVENES	PEGR

La transformación completa y los resultados de la ejecución de la transformación son los siguientes:



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

Step Metrics													
#	Nombre paso	Numero Copia	Leido	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)	Pri/E/S
1	Obtener EGR_EUR	0	0	60	60	0	0	0	0	Finalizado	0.0s	3.529	-
2	Obtener Pais	0	60	60	60	0	0	0	0	Finalizado	0.0s	1.463	-
3	Cargar FACT_PEGR_COMPARATIVA	0	60	60	0	60	0	0	0	Finalizado	0.1s	833	-

De las métricas se desprende que se han insertado 60 registros en la tabla de hechos FACT_PERG_EUR, siguiendo el modelo multidimensional diseñado que nos permitirá realizar un análisis comparativo de las personas egresadas por países y año.

4) Implementación de trabajos con procesos ETL's.

Teniendo en cuenta los bloques de procesos implementados:

- Bloque IN_: Procesos ETL de transformación y carga al área intermedia.
- Bloque TR_DIM: Procesos ETL de transformación y carga de dimensiones.
- Bloque TR_FACT: Procesos ETL de transformación y carga de hechos

Vamos a diseñar los trabajos (*jobs*) mediante PDI que van a permitir la ejecución secuencial de todos los procesos ETL's incluidos en cada bloque definido.

Cada trabajo contiene como pasos cada una de las transformaciones implementadas en el apartado anterior de Diseño de ETL's.

A. Trabajo 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 la siguiente:

Los pasos incluidos en el trabajo JOB_IN son:

- Start: Componente General de diseño de trabajos, que marca el **Inicio** del trabajo.
- Set variables. Componente General de diseño de trabajos, que permite la definición de variables de entorno. En nuestro caso definimos las variables de entorno DIR_ENT, con el directorio donde se encuentran las fuentes proporcionadas y BBDD, con la cadena de conexión a la base de datos.
- IN_RAMA: Ejecución de transformación IN_RAMA.
- IN_SEGR_N: Ejecución de transformación IN_SEGR_N.
- IN_EGR_C16_17: Ejecución de transformación IN_EGR_C16_17.
- IN_EIL_3003: Ejecución de transformación IN_EIL_3003
- IN_EGR_EUR: Ejecución de transformación IN_EGR_EUR
- Success: Componente General de diseño de trabajos, que marca la **Finalización** del trabajo.

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

Trabajo / Entrada de T...	Comentario	Resultado	Razón	Nºm	Fecha registro
▼ job_in_local					
Trabajo: job_in_local	Start of job execution	start			2019/05/12 23:11:15
Start	Start of job execution	start			2019/05/12 23:11:15
Start	Job execution finished	Exito		0	2019/05/12 23:11:15
Set variables	Start of job execution		Followed unconditional link		2019/05/12 23:11:15
Set variables	Job execution finished	Exito		0	2019/05/12 23:11:15
IN_RAMA	Start of job execution		Followed link after success		2019/05/12 23:11:15
IN_RAMA	Job execution finished	Exito		2	2019/05/12 23:11:16
IN_SEGR_N	Start of job execution		Followed link after success		2019/05/12 23:11:16
IN_SEGR_N	Job execution finished	Exito		3	2019/05/12 23:11:16
IN_EGR_C16_17	Start of job execution		Followed link after success		2019/05/12 23:11:16
IN_EGR_C16_17	Job execution finished	Exito		4	2019/05/12 23:11:16
IN_EIL_03003	Start of job execution		Followed link after success		2019/05/12 23:11:16
IN_EIL_03003	Job execution finished	Exito		5	2019/05/12 23:11:17
IN_EGR_EUR	Start of job execution		Followed link after success		2019/05/12 23:11:17
IN_EGR_EUR	Job execution finished	Exito		6	2019/05/12 23:11:19
Success	Start of job execution		Followed link after success		2019/05/12 23:11:19
Success	Job execution finished	Exito		6	2019/05/12 23:11:19
Trabajo: job_in_local	Job execution finished	Exito	finished	6	2019/05/12 23:11:19

Se observa el procesamiento con éxito de todos los pasos del JOB_IN, correspondiente a la ejecución de todas las transformaciones que están incluidas trabajo.

B. Trabajo JOB_TR_DIM

El trabajo (*job*) JOB_TR_DIM procesa todas las transformaciones del bloque TR_DIM para la carga de datos desde las tablas intermedias a las tablas de dimensiones del almacén.

El diseño del trabajo (*job*) JOB_TR_DIM es la siguiente:



Los pasos incluidos en el trabajo JOB_TR_DIM son:

- Start: Componente General de diseño de trabajos, que marca el **Inicio** del trabajo.
- Set variables. Componente General de diseño de trabajos, que permite la definición de variables de entorno. En nuestro caso definimos las variables de entorno DIR_ENT, con el directorio donde se encuentran las fuentes proporcionadas y BBDD, con la cadena de conexión a la base de datos.
- TR_DIM_ANIO: Ejecución de transformación para la carga de la dimensión DIM_ANIO.
- TR_DIM_MODALIDAD: Ejecución de transformación para la carga de la dimensión DIM_MODALIDAD.
- TR_DIM_TIPO_UNIV: Ejecución de transformación para la carga de la dimensión DIM_TIPO_UNIV.
- TR_DIM_UNIVERSIDAD: Ejecución de transformación para la carga de la dimensión DIM_UNIVERSIDAD.
- TR_DIM_RAMA: Ejecución de transformación para la carga de la dimensión DIM_RAMA.
- TR_DESCONOCIDOS: Ejecución de transformación para la carga de registro de desconocidos en la dimensión DIM_RAMA.
- TR_DIM_PERFIL: Ejecución de transformación para la carga datos a las dimensiones DIM_SEXO y DIM_EDAD.
- TR_DIM_PAIS: Ejecución de transformación para la carga datos a la dimensión DIM_PAIS.
- Success: Componente General de diseño de trabajos, que marca la **Finalización** del trabajo.

Se observa del resultado de la ejecución del trabajo, que todos los pasos se procesan con éxito. Con la finalización de este proceso, se realizarán las cargas de todas las dimensiones del modelo multidimensional del almacén integrado de personas egresadas.

Trabajo / Entrada de T...	Comentario	Resultado	Razón	Nºm	Fecha registro
job_tr_dim_local					
Trabajo: job_tr_dim_	Start of job execution	start			2019/05/12 23:13:13
Start	Start of job execution	start			2019/05/12 23:13:13
Start	Job execution finished	Exito		0	2019/05/12 23:13:13
Set variables	Start of job execution		Followed uncondition...		2019/05/12 23:13:13
Set variables	Job execution finished	Exito		0	2019/05/12 23:13:13
TR_DIM_ANIO	Start of job execution		Followed uncondition...		2019/05/12 23:13:13
TR_DIM_ANIO	Job execution finished	Exito		2	2019/05/12 23:13:13
TR_DIM_MODALIDA	Start of job execution		Followed link after suc...		2019/05/12 23:13:13
TR_DIM_MODALIDA	Job execution finished	Exito		3	2019/05/12 23:13:13
TR_DIM_TIPO_UNIV	Start of job execution		Followed link after suc...		2019/05/12 23:13:13
TR_DIM_TIPO_UNIV	Job execution finished	Exito		4	2019/05/12 23:13:13
TR_DIM_UNIVERSID	Start of job execution		Followed link after suc...		2019/05/12 23:13:13
TR_DIM_UNIVERSID	Job execution finished	Exito		5	2019/05/12 23:13:14
TR_DIM_RAMA	Start of job execution		Followed link after suc...		2019/05/12 23:13:14
TR_DIM_RAMA	Job execution finished	Exito		6	2019/05/12 23:13:14
TR_DESCONOCIDO!	Start of job execution		Followed link after suc...		2019/05/12 23:13:14
TR_DESCONOCIDO!	Job execution finished	Exito		7	2019/05/12 23:13:14
TR_DIM_PERFIL	Start of job execution		Followed link after suc...		2019/05/12 23:13:14
TR_DIM_PERFIL	Job execution finished	Exito		8	2019/05/12 23:13:14
Success	Start of job execution		Followed link after suc...		2019/05/12 23:13:14
Success	Job execution finished	Exito		8	2019/05/12 23:13:14
Trabajo: job_tr_dim_	Job execution finished	Exito	finished	8	2019/05/12 23:13:14

C. Trabajo JOB_TR_FACT

El trabajo (*job*) JOB_TR_FACT 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 del trabajo (*job*) JOB_TR_FACT es la siguiente:



Los pasos incluidos en el trabajo JOB_TR_FACT son:

- Start: Componente General de diseño de trabajos, que marca el **Inicio** del trabajo.
- Set variables. Componente General de diseño de trabajos, que permite la definición de variables de entorno. En nuestro caso definimos las variables de entorno DIR_ENT, con el directorio donde se encuentran las fuentes proporcionadas y BBDD, con la cadena de conexión a la base de datos.
- TR_FACT_PEGR_EVOLUTIVO: Ejecución de transformación para la carga de la tabla de hechos TR_FACT_PEGR_EVOLUTIVO.
- TR_FACT_PEGR_PERFIL: Ejecución de transformación para la carga de la tabla de hechos TR_FACT_PEGR_PERFIL.

- TR_FACT_PEGR_INSERTADAS: Ejecución de transformación para la carga de la tabla de hechos TR_FACT_PEGR_INSERTADAS.
- TR_FACT_PEGR_EUR: Ejecución de transformación para la carga de la tabla de hechos TR_FACT_PEGR_EUR.
- Success: Componente General de diseño de trabajos, que marca la **Finalización** del trabajo.

Se observa del resultado de la ejecución del trabajo, que todos los pasos se procesan con éxito. Con la finalización de este proceso, se realizarán las cargas de todas las tablas de hechos del modelo multidimensional del almacén integrado de personas egresadas.

Trabajo / Entrada de T...	Comentario	Resultado	Razón	N...	Fecha registro
job_tr_fact_local					
Trabajo: job_tr_fact_l	Start of job execution	start			2019/05/12 23:14:11
Start	Start of job execution	start			2019/05/12 23:14:11
Start	Job execution finished	Exito		0	2019/05/12 23:14:11
Set variables	Start of job execution		Followed unconditional link		2019/05/12 23:14:11
Set variables	Job execution finished	Exito		0	2019/05/12 23:14:11
TR_FACT_PEGR_EVO	Start of job execution		Followed link after success		2019/05/12 23:14:11
TR_FACT_PEGR_EVO	Job execution finished	Exito		2	2019/05/12 23:14:12
TR_FACT_PEGR_PERI	Start of job execution		Followed link after success		2019/05/12 23:14:12
TR_FACT_PEGR_PERI	Job execution finished	Exito		3	2019/05/12 23:14:13
TR_FACT_PEGR_INSE	Start of job execution		Followed link after success		2019/05/12 23:14:13
TR_FACT_PEGR_INSE	Job execution finished	Exito		4	2019/05/12 23:14:13
TR_FACT_PEGR_EUR	Start of job execution		Followed link after success		2019/05/12 23:14:13
TR_FACT_PEGR_EUR	Job execution finished	Exito		5	2019/05/12 23:14:13
Success	Start of job execution		Followed link after success		2019/05/12 23:14:13
Success	Job execution finished	Exito		5	2019/05/12 23:14:13
Trabajo: job_tr_fact_l	Job execution finished	Exito	finished	5	2019/05/12 23:14:13

D. Trabajo JOB_ETL

El trabajo (*job*) JOB_ETL procesa todos los trabajos con todos los procesos de extracción, transformación y carga del almacén de personas egresadas.

El diseño del trabajo (*job*) JOB_ETL es la siguiente:



Los pasos incluidos en el trabajo JOB_ETL son:

- Start: Componente General de diseño de trabajos, que marca el **Inicio** del trabajo.
- Set variables. Definición de las variables de entorno, DIR_ENT y BBDD
- Job_IN: Ejecución de trabajo del Bloque IN_, para la carga al área intermedia.

- Job_TR_DIM: Ejecución de trabajo del Bloque DIM_, para la carga de dimensiones.
- Job_TR_FACT: Ejecución de trabajo del Bloque FACT_, para la carga de tablas de hechos.
- Success: Componente General de diseño de trabajos, que marca la **Finalización** del trabajo.

La ejecución completa y con éxito del JOB_ETL supondrá la ejecución de los procesos ETL para la carga de los datos al almacén integrado de egresados universitarios desde las fuentes origen hasta el modelo multidimensional que se ha diseñado.