

Práctica 2: Carga de datos

Patricia Lázaro Tello

Índice

1. Introducción	2
2. Identificación de los procesos ETL	7
2.1. Bloque IN	7
2.2. Bloque TR	8
3. Diseño de los procesos ETL	10
3.1. <i>Staging area</i>	10
3.2. Modelo multidimensional	12
3.3. Creación de los procesos ETL	17
3.4. Volumetría	52
4. Implementación de los <i>jobs</i> con ETL	53
4.1. <i>Job</i> IN	53
4.2. <i>Job</i> TR_DIM	54
4.3. <i>Job</i> TR_FACT	55
4.4. <i>Job</i> DW	56

1. Introducció

Este documento recoge implementación del modelo multidimensional y la carga de datos de las diferentes fuentes a dicho modelo. Para ello se ha tenido que configurar el entorno de trabajo *Pentaho Data Integration* (PDI), diseñar e implementar los distintos procesos de extracción, transformación y carga (ETL), incluyendo una *staging area*.

Parte de la base de la solución de la práctica 1, donde se detallaban las dimensiones y hechos del modelo multidimensional, así como las fuentes de datos que se iban a utilizar para la primera carga del modelo.

El documento se divide en los siguientes apartados, correspondientes a los pasos necesarios para la carga de datos en las tablas del modelo multidimensional:

- **Identificación de los procesos ETL**, creando los vínculos entre fuentes y tablas del *staging area*, así como la relación entre los datos y las tablas del modelo multidimensional.
- **Diseño de los procesos ETL**, traduciendo las relaciones identificadas en el paso anterior a instrucciones concretas en PDI.
- **Implementación de los *jobs***, aglutinando las transformaciones creadas en el paso anterior según su función para el modelo multidimensional.

2. Matriz de dimensiones y métricas

Environmental Measurements				
Proceso	STG_AmbientalProtection	STG_ProtectedAreas	STG_Residuos	
DIM_Métrica	Inversión en protección ambiental (€)	Áreas terrestre y marina protegidas	Gestión de residuos y reciclaje	
DIM_Date	X	X	X	
DIM_Region	X	Relación indirecta a través de los países	Relación indirecta a través de los países	
DIM_EconomicActivitySector	X	N/A	N/A	
DIM_TypeEquipmentInstallation	X	N/A	N/A	
DIM_SDG	'Ambito' es el código de las áreas objetivo	'Unit' es el código de las áreas objetivo	'Var' es el código de las áreas objetivo	
DIM_Measurement	'Ambito' es el código de la measurement	'Unit' es el código de la measurement	'Var' es el código de la measurement	
	La unidad de la measurement es €	'Unit' permite obtener la ud. de la measurement	'Unit' es la unidad de la measurement	
DIM_Product	N/A	N/A	N/A	
DIM_Country	Relación indirecta a través de la C.C.A.A.	'Geo' es el código ISO-2 del país	'Cou' es el código ISO-2 del país	

Environmental Measurements				
Proceso	STG_EnergyBalance	STG_Objectives	STG_ObjectivesAreas	STG_Countries
DIMMétrica	-	-	-	-
DIM_Date	N/A	N/A	N/A	N/A
DIM_Region	N/A	N/A	N/A	Relación indirecta a través de los países
DIM_Economic ActivitySector	N/A	N/A	N/A	N/A
DIM_TypeEquipment Installation	N/A	N/A	N/A	N/A
DIM_SDG	N/A	X	'ODS principal' hace referencia a pk_sdg	N/A
DIM_Measurement	N/A	'Objetivo' es FK de la tabla	X	N/A
DIM_Product	N/A	N/A	N/A	N/A
DIM_Country	N/A	N/A	N/A	N/A

Proceso	Energy Balances		
	STG_AmbientalProtection	STG_ProtectedAreas	STG_Residuos
DIM_Métrica	-	-	-
DIM_Date	N/A	N/A	N/A
DIM_Region	N/A	N/A	N/A
DIM_Economic ActivitySector	N/A	N/A	N/A
DIM_TypeEquipment Installation	N/A	N/A	N/A
DIM_SDG	N/A	N/A	N/A
DIM_Measurement	N/A	N/A	N/A
DIM_Product	N/A	N/A	N/A
DIM_Country	N/A	N/A	N/A

Energy Balances				
Proceso	STG_EnergyBalance	STG_Objectives	STG_ObjectivesAreas	STG_Countries
DIMMétrica	Balance energético por países	-	-	-
DIM_Date	X	N/A	N/A	N/A
DIM_Region	N/A	N/A	N/A	N/A
DIM_Economic ActivitySector	N/A	N/A	N/A	N/A
DIM_TypeEquipment Installation	N/A	N/A	N/A	N/A
DIM_SDG	'flow' es el código de las áreas objetivo	X	'ODS principal' hace referencia a pk_sdg	N/A
DIM_Measurement	'flow' contiene el código y la unidad de la measurement	'Objetivo' es FK de la tabla	X	N/A
DIM_Product	X	N/A	N/A	N/A
DIM_Country	N/A	N/A	N/A	X

3. Identificación de los procesos ETL

Una vez se han analizado las fuentes de datos y se ha diseñado un modelo multidimensional con sus hechos y dimensiones, es momento de crear dicho modelo y proceder a la carga de datos.

Para tal efecto, se estructurarán los procesos ETL según la función que cumplan dentro del proceso de carga de datos, tomando en cuenta las características de la misma:

- Tipo de carga: carga inicial
- Uso de *staging area* permitido

Dadas estas características, se encuentran 2 bloques bien diferenciados dentro de los procesos ETL identificados:

- **Bloque IN:** corresponde con la carga de las fuentes de datos a las tablas intermedias en el *staging area*.
- **Bloque TR:** corresponde con la transformación y carga de datos de las tablas intermedias del *staging area* a las tablas finales del modelo multidimensional.

En este bloque se encuentran los procesos ETL que corresponden con la carga de datos en las dimensiones y los que corresponden con la carga de datos en los hechos.

A continuación se definen los procesos de ETL de los bloques identificados, adjuntando información del nombre del proceso, una descripción sucinta, el origen y el destino de los datos.

3.1. Bloque IN

Cuadro 5: Relación origen-destino

Nombre	Origen → Destino
IN_AMBIENTALPROTECTION	02002.xlsx → STG_AmbientalProtection
IN_COUNTRIES	Countries.json → STG_Countries
IN_ENVBIO1	env_bio1.tsv → STG_ProtectedAreas
IN_OBJECTIVES	ODS.xlsx → STG_Objectives
IN_OBJECTIVESAREAS	ODS.xlsx → STG_ObjectivesAreas

IN_RESIDUOS	DataGeneric.xml → STG_Residuos
IN_WORLD_ENERGY_BALANCES	WorldEnergyBalancesHighlights_final.xlsx → STG_EnergyBalance

Cuadro 6: Descripción de los procesos ETL

Nombre	Descripción
IN_AMBIENTALPROTECTION	Carga de los datos de inversión en protección ambiental por Comunidad Autónoma en la <i>staging area</i>
IN_COUNTRIES	Carga de los datos relativos a países en la <i>staging area</i>
IN_ENVBIO1	Carga de los datos de áreas protegidas por países en la <i>staging area</i>
IN_OBJECTIVES	Carga de los objetivos de desarrollo sostenible en la <i>staging area</i>
IN_OBJECTIVESAREAS	Carga de las relaciones entre objetivos de desarrollo sostenible y áreas en la <i>staging area</i>
IN_RESIDUOS	Carga de los datos de residuos en la <i>staging area</i>
IN_WORLD_ENERGY_BALANCES	Carga de los datos de balance energético en la <i>staging area</i>

3.2. Bloque TR

Dentro del bloque de procesos ETL correspondientes a cargar los datos finales en las tablas del modelo multidimensional se encuentran otros 2 bloques: los correspondientes a la población de las tablas de dimensiones (**TR_DIM**) y los correspondientes a la población de las tablas de hechos (**TR_FACT**).

3.2.1. Bloque TR_DIM

Cuadro 7: Relación origen-destino

Nombre	Origen → Destino
TR_DIM_Country	STG_Countries → DIM_Country

TR_DIM_Date	entrada manual → DIM_Date
TR_DIM_EconomicActivitySector	STG_AmbientalProtection → DIM_EconomicActivitySector
TR_DIM_Measurement	STG_AmbientalProtection, STG_EnergyBalance, STG_ProtectedAreas, STG_Residuos, STG_ObjectivesAreas → DIM_Measurement
TR_DIM_Product	STG_EnergyBalance → DIM_Product
TR_DIM_Region	STG_AmbientalProtection → DIM_Region
TR_DIM_SDG	STG_Objectives → DIM_SDG
TR_DIM_TypeEquipmentInstallation	STG_AmbientalProtection → DIM_TypeEquipmentInstallation

Cuadro 8: Descripción de los procesos ETL

Nombre	Descripción
TR_DIM_Country	Carga de los datos relativos a países en el modelo
TR_DIM_Date	Carga de las fechas en el modelo
TR_DIM_EconomicActivitySector	Carga de los tipos de sectores de actividad económica en el modelo
TR_DIM_Measurement	Carga de los tipos de medidas en el modelo
TR_DIM_Product	Carga de los tipos de productos en el modelo
TR_DIM_Region	Carga de los datos de regiones por países en el modelo
TR_DIM_SDG	Carga de los objetivos de desarrollo sostenible en el modelo
TR_DIM_TypeEquipmentInstallation	Carga de los tipos de instalaciones y equipamiento en el modelo

3.2.2. Bloque TR_FACT

Cuadro 9: Relación origen-destino

Nombre	Origen → Destino
TR_FACT_EnergyBalances	STG_EnergyBalance → FACT_EnergyBalance

TR_FACT_EnvironmentalMeasurements	STG_AmbientalProtection, STG_Residuos, STG_ProtectedAreas → FACT_EnvironmentalMeasurements
-----------------------------------	--

Cuadro 10: Descripción de los procesos ETL

Nombre	Descripción
TR_FACT_EnergyBalances	Carga de los datos relativos al balance de energía en el modelo
TR_FACT_EnvironmentalMeasurements	Carga de las medidas ambientales en el modelo

4. Diseño de los procesos ETL

Después de identificar los procesos de extracción, transformación y carga de datos, se ha de proceder a diseñar e implementar las tablas SQL en las que los datos estarán contenidos, así como los procesos ETL para poblarlas.

4.1. *Staging area*

La creación de las tablas intermedias del *staging area* se realizará dentro de un *job* de PDI, antes de comenzar la transformación y carga de datos en las mismas. El proceso de creación de las tablas también podría llevarse a cabo una única vez a través de la ejecución de *scripts* en la base de datos SQL Server.

Script 1: STG_AmbientalProtection

```

1 CREATE TABLE [dbo].[STG_AmbientalProtection](
2     [Periodo] [int] NOT NULL,
3     [Sector] [varchar](100) NOT NULL,
4     [EquipoInstalacion] [varchar](200) NULL,
5     [Ambito] [varchar](200) NULL,
6     [ComunidadAutonoma] [varchar](200) NULL,
7     [Inversion] [decimal](19,4) NULL,
8 ) ON [PRIMARY]
```

Script 2: STG_Objectives

```

1 CREATE TABLE [dbo].[STG_Objectives](
2     [Objetivo] [int] NOT NULL,
```

```

3      [Nombre] [varchar](100) NULL,
4      [Descripcion] [varchar](512) NULL
5  ) ON [PRIMARY]

```

Script 3: STG_ObjectivesAreas

```

1  CREATE TABLE [dbo].[STG_ObjectivesAreas](
2      [Codigo] [varchar](100) NOT NULL,
3      [AmbitoVarFlow] [varchar](200) NULL,
4      [ODS principal] [int] NOT NULL
5  ) ON [PRIMARY]

```

Script 4: STG_EnergyBalance

```

1  CREATE TABLE [dbo].[STG_EnergyBalance](
2      [country] [varchar] (255) NULL,
3      [product] [varchar] (255) NULL,
4      [flow] [varchar] (255) NULL,
5      [year] [int] NULL,
6      [value] [float] NULL) ON [PRIMARY]

```

Script 5: STG_Countries

```

1  CREATE TABLE [dbo].[STG_Countries](
2      [Nombre] [varchar](50) NOT NULL,
3      [Name] [varchar](50) NOT NULL,
4      [Nom] [varchar](50) NULL,
5      [Iso2] [varchar](2) NOT NULL,
6      [Iso3] [varchar](3) NOT NULL,
7      [Phone] [varchar](10) NULL
8  ) ON [PRIMARY]

```

Script 6: STG_ProtectedAreas

```

1  CREATE TABLE [dbo].[STG_ProtectedAreas](
2      [Geo][varchar](10) NOT NULL,
3      [Year] [int] NOT NULL,
4      [Unit][varchar](15) NOT NULL,
5      [Value][decimal](19,4) NULL,
6  ) ON [PRIMARY]

```

Script 7: STG_Residuos

```

1  CREATE TABLE [dbo].[STG_Residuos](
2      [Cou][varchar](10) NOT NULL,
3      [Var][varchar](25) NULL,
4      [TimeFormat][varchar](5) NULL,

```

```

5      [Unit][varchar](15) NULL,
6      [Powercode][int] NULL,
7      [Year][int] NULL,
8      [Obs][decimal](19,4) NULL,
9      [Status][varchar](5) NULL
10 ) ON [PRIMARY]

```

Las tablas intermedias se han creado sin índices para facilitar la carga de datos desde las fuentes de origen. Todas las tablas de la *staging area* tienen el prefijo «STG_» en su nombre.

4.2. Modelo multidimensional

El modelo multidimensional se ha estructurado de acuerdo a la división entre dimensiones y hechos. Los hechos tienen el prefijo «FACT_» mientras que las dimensiones cuentan con el prefijo «DIM_» en su nombre de tabla.

4.2.1. Dimensiones

Para las tablas de las dimensiones, se han creado claves primarias (*primary keys*) de tipo entero para permitir ser referenciadas de forma sencilla por las tablas de hechos.

Script 8: DIM_Measurement

```

1  CREATE TABLE [dbo].[DIM_Measurement](
2      [pk_measurement] [int] NOT NULL,
3      [measurement_code] [varchar](100) NULL,
4      [measurement_name] [varchar](200) NULL,
5      [unit] [varchar](25) NULL,
6      [fk_sdg] [int] NULL,
7      CONSTRAINT [PK_DIM_Measurement] PRIMARY KEY CLUSTERED (
8          [pk_measurement] ASC
9      )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
10     OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]

```

Script 9: DIM_Date

```

1  CREATE TABLE [dbo].[DIM_Date](
2      [pk_date] [int] NOT NULL,
3      [date_year] [int] NOT NULL,
4      [date_month] [int] NOT NULL,

```

```

5      [date_day] [int] NOT NULL,
6      [date_date] [datetime] NOT NULL,
7      CONSTRAINT [PK_DIM_Date] PRIMARY KEY CLUSTERED (
8          [pk_date] ASC
9      )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
10     ) ON [PRIMARY]

```

Script 10: DIM_EconomicActivitySector

```

1  CREATE TABLE [dbo].[DIM_EconomicActivitySector](
2      [pk_activitysector] [int] NOT NULL,
3      [activitysector_name] [varchar](100) NULL,
4      CONSTRAINT [PK_DIM_EconomicActivitySector] PRIMARY KEY CLUSTERED (
5          [pk_activitysector] ASC
6      )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
7      ) ON [PRIMARY]

```

Script 11: DIM_TypeEquipmentInstallation

```

1  CREATE TABLE [dbo].[DIM_TypeEquipmentInstallation](
2      [pk_typeeequipinstall] [int] NOT NULL,
3      [typeeequipinstall_name] [varchar](100) NULL,
4      CONSTRAINT [PK_DIM_TypeEquipmentInstallation] PRIMARY KEY CLUSTERED (
5          [pk_typeeequipinstall] ASC
6      )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
7      ) ON [PRIMARY]

```

Script 12: DIM_Product

```

1  CREATE TABLE [dbo].[DIM_Product](
2      [pk_product] [int] NOT NULL,
3      [product_name] [varchar](100) NULL,
4      CONSTRAINT [PK_DIM_Product] PRIMARY KEY CLUSTERED (
5          [pk_product] ASC
6      )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
7      ) ON [PRIMARY]

```

Script 13: DIM_Region

```

1  CREATE TABLE [dbo].[DIM_Region](
2      [pk_region] [int] NOT NULL,
3      [region] [varchar](100) NULL,

```

```

4      [country_code2] [varchar](2) NULL,
5      [country_code3] [varchar](3) NULL,
6      [country_name] [varchar](100) NULL,
7      CONSTRAINT [PK_DIM_Region] PRIMARY KEY CLUSTERED (
8          [pk_region] ASC
9      )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
10     ) ON [PRIMARY]

```

Script 14: DIM_SDG

```

1      CREATE TABLE [dbo].[DIM_SDG](
2          [pk_sdg] [int] NOT NULL,
3          [sdg_name] [varchar](50) NULL,
4          [sdg_description] [varchar](500) NULL,
5          CONSTRAINT [PK_DIM_SDG] PRIMARY KEY CLUSTERED (
6              [pk_sdg] ASC
7          )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
8      ) ON [PRIMARY]

```

Script 15: DIM_Country

```

1      CREATE TABLE [dbo].[DIM_Country](
2          [pk_country] [int] NOT NULL,
3          [country_code] [varchar](2) NULL,
4          [country_code3] [varchar](3) NULL,
5          [country_name_sp] [varchar](100) NULL,
6          [country_name_en] [varchar](100) NULL,
7          [country_name_fr] [varchar](100) NULL,
8          [country_phone_code] [varchar](5) NULL,
9          CONSTRAINT [PK_DIM_Country] PRIMARY KEY CLUSTERED (
10             [pk_country] ASC
11         )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
12     ) ON [PRIMARY]

```

4.2.2. Hechos

Para las tablas de hechos también se han definido claves primarias (*primary keys*) según el diseño propuesto en la solución de la práctica 1.

Script 16: FACT_EnvironmentalMeasurements

```

1 CREATE TABLE [dbo].[FACT_EnvironmentalMeasurements](
2     [pk_id] [int] NOT NULL,
3     [fk_date] [int] NOT NULL,
4     [fk_region] [int] NOT NULL,
5     [fk_activitysector] [int] NOT NULL,
6     [fk_typeequipinstall] [int] NOT NULL,
7     [fk_measurement] [int] NOT NULL,
8     [value] [decimal](19,4) NULL,
9     CONSTRAINT [PK_FACT_EnvironmentalMeasurements] PRIMARY KEY CLUSTERED (
10         [pk_id] ASC
11     )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
12         OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
13 ) ON [PRIMARY]

```

Script 17: FACT_EnergyBalances

```

1 CREATE TABLE [dbo].[FACT_EnergyBalances](
2     [pk_fk_date] [int] NOT NULL,
3     [pk_fk_country] [int] NOT NULL,
4     [pk_fk_product] [int] NOT NULL,
5     [pk_fk_measurement] [int] NOT NULL,
6     [value] [decimal](19,4) NULL,
7     CONSTRAINT [PK_FACT_EnergyBalances] PRIMARY KEY CLUSTERED (
8         [pk_fk_date] ASC,
9         [pk_fk_country] ASC,
10        [pk_fk_product] ASC,
11        [pk_fk_measurement] ASC
12     )WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
13         OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
14 ) ON [PRIMARY]

```

4.2.3. Foreign keys

Después de crear todas las tablas de dimensiones y hechos se han definido las restricciones (*constraints*) para las claves foráneas (*foreign keys*) de todas las tablas que poseen una.

Script 18: Foreign keys

```

1 ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] WITH CHECK ADD
2     CONSTRAINT [FK_FACT_EnvironmentalMeasurements_DIM_Date] FOREIGN KEY([
3         fk_date])
4     REFERENCES [dbo].[DIM_Date] ([pk_date])

```

```

4  ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] CHECK CONSTRAINT [
    FK_FACT_EnvironmentalMeasurements_DIM_Date]
5
6  ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] WITH CHECK ADD
    CONSTRAINT [FK_FACT_EnvironmentalMeasurements_DIM_EconomicActivitySector]
    FOREIGN KEY([fk_activitysector])
7  REFERENCES [dbo].[DIM_EconomicActivitySector] ([pk_activitysector])
8
9  ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] CHECK CONSTRAINT [
    FK_FACT_EnvironmentalMeasurements_DIM_EconomicActivitySector]
10
11 ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] WITH CHECK ADD
    CONSTRAINT [FK_FACT_EnvironmentalMeasurements_DIM_Measurement] FOREIGN
    KEY([fk_measurement])
12 REFERENCES [dbo].[DIM_Measurement] ([pk_measurement])
13
14 ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] CHECK CONSTRAINT [
    FK_FACT_EnvironmentalMeasurements_DIM_Measurement]
15
16 ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] WITH CHECK ADD
    CONSTRAINT [FK_FACT_EnvironmentalMeasurements_DIM_Region] FOREIGN KEY([
    fk_region])
17 REFERENCES [dbo].[DIM_Region] ([pk_region])
18
19 ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] CHECK CONSTRAINT [
    FK_FACT_EnvironmentalMeasurements_DIM_Region]
20
21 ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] WITH CHECK ADD
    CONSTRAINT [
    FK_FACT_EnvironmentalMeasurements_DIM_TypeEquipmentInstallation] FOREIGN
    KEY([fk_typeequipinstall])
22 REFERENCES [dbo].[DIM_TypeEquipmentInstallation] ([pk_typeequipinstall
    ])
23
24 ALTER TABLE [dbo].[FACT_EnvironmentalMeasurements] CHECK CONSTRAINT [
    FK_FACT_EnvironmentalMeasurements_DIM_TypeEquipmentInstallation]
25
26 ALTER TABLE [dbo].[FACT_EnergyBalances] WITH CHECK ADD CONSTRAINT [
    FK_FACT_EnergyBalances_DIM_Country] FOREIGN KEY([pk_fk_country])
27 REFERENCES [dbo].[DIM_Country] ([pk_country])
28
29 ALTER TABLE [dbo].[FACT_EnergyBalances] CHECK CONSTRAINT [
    FK_FACT_EnergyBalances_DIM_Country]
30

```



```

31  ALTER TABLE [dbo].[FACT_EnergyBalances] WITH CHECK ADD CONSTRAINT [
    FK_FACT_EnergyBalances_DIM_Date] FOREIGN KEY([pk_fk_date])
32  REFERENCES [dbo].[DIM_Date] ([pk_date])
33
34  ALTER TABLE [dbo].[FACT_EnergyBalances] CHECK CONSTRAINT [
    FK_FACT_EnergyBalances_DIM_Date]
35
36  ALTER TABLE [dbo].[FACT_EnergyBalances] WITH CHECK ADD CONSTRAINT [
    FK_FACT_EnergyBalances_DIM_Measurement] FOREIGN KEY([pk_fk_measurement])
37  REFERENCES [dbo].[DIM_Measurement] ([pk_measurement])

```

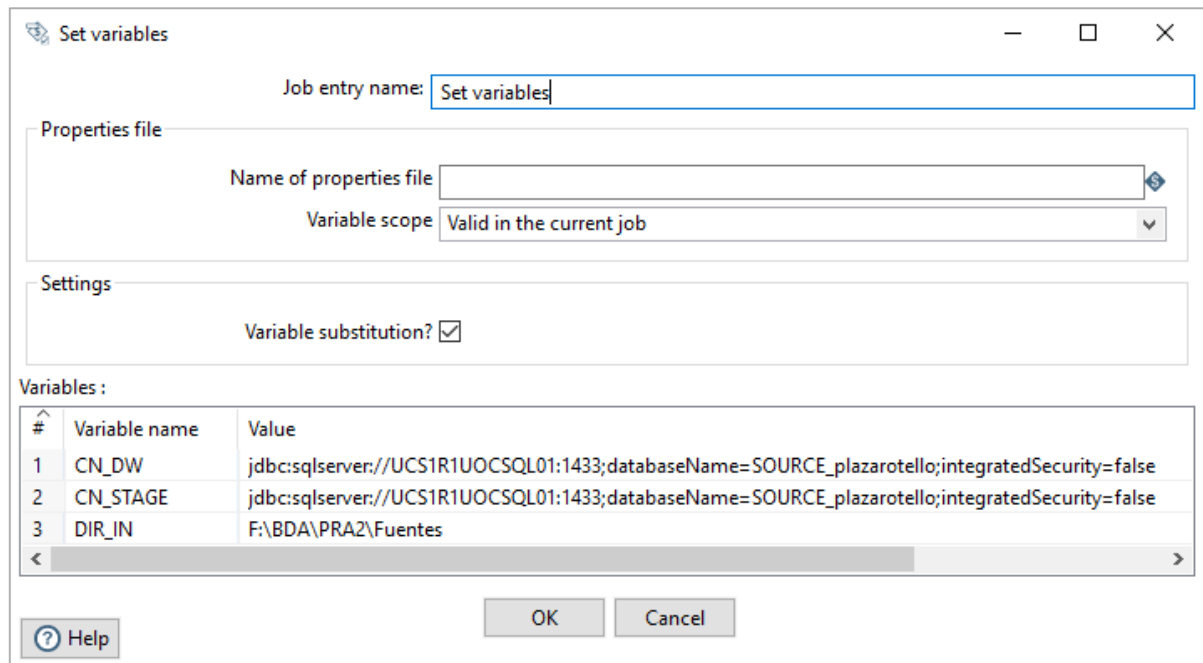
4.3. Creación de los procesos ETL

Una vez ejecutados los *scripts* anteriores, la base de datos habrá de tener las tablas creadas y listas para ser pobladas. Los siguientes pasos incluyen configurar el entorno PDI para conectar con las bases de datos y crear los procesos ETL para cargar los datos en las tablas.

4.3.1. Variables de entorno y *setup*

En primer lugar es necesario cargar las variables de entorno en PDI. Estas variables de entorno incluyen la localización de las fuentes y la conexión a las bases de datos del *data warehouse* y *staging area*.

Estas variables pueden ser cargadas en cada *job* ejecutado mediante el nodo «Set variables» como se muestra en la figura, o se pueden cargar globalmente mediante la adición de propiedades en *kettle.properties*, o se pueden cargar para la sesión mediante el menú *Edit > Set Environment Variables...*



Set variables

Job entry name: Set variables

Properties file

Name of properties file

Variable scope: Valid in the current job

Settings

Variable substitution? ☒

Variables :

#	Variable name	Value
1	CN_DW	jdbc:sqlserver://UCS1R1UOCSQL01:1433;databaseName=SOURCE_plazarotello;integratedSecurity=false
2	CN_STAGE	jdbc:sqlserver://UCS1R1UOCSQL01:1433;databaseName=SOURCE_plazarotello;integratedSecurity=false
3	DIR_IN	F:\BDA\PRA2\Fuentes

Help OK Cancel

Figura 1: Valores de las variables de entorno

4.3.2. Conexión a la base de datos

Aunque la base de datos del modelo multidimensional y la del *staging area* son la misma, es necesario crear una conexión a la base de datos para cada una de ella. Esta separación de bases de datos final e intermedia se utiliza para poder hacer el sistema más flexible y escalable en un futuro.

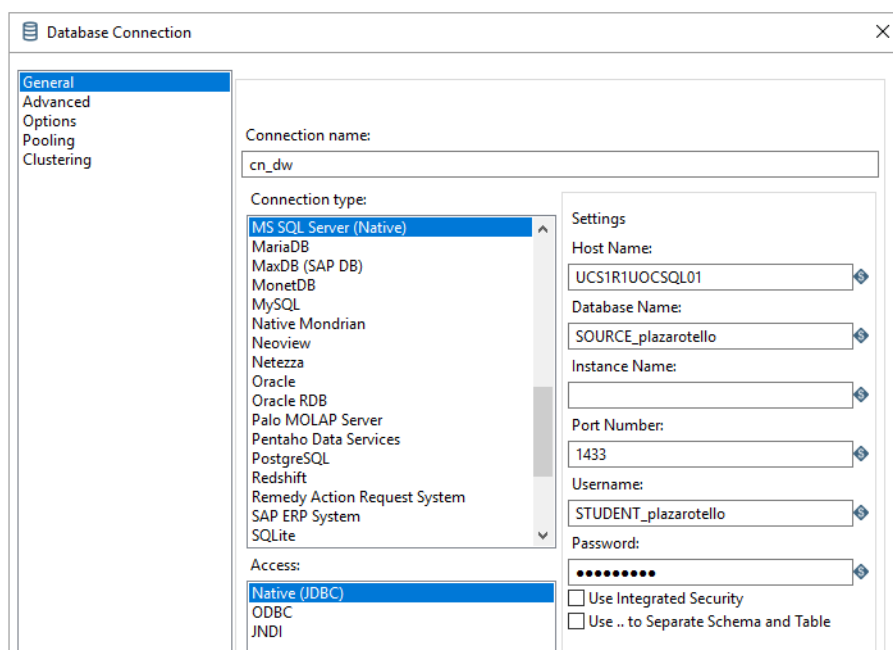


Figura 2: Conexión a la base de datos del modelo multidimensional

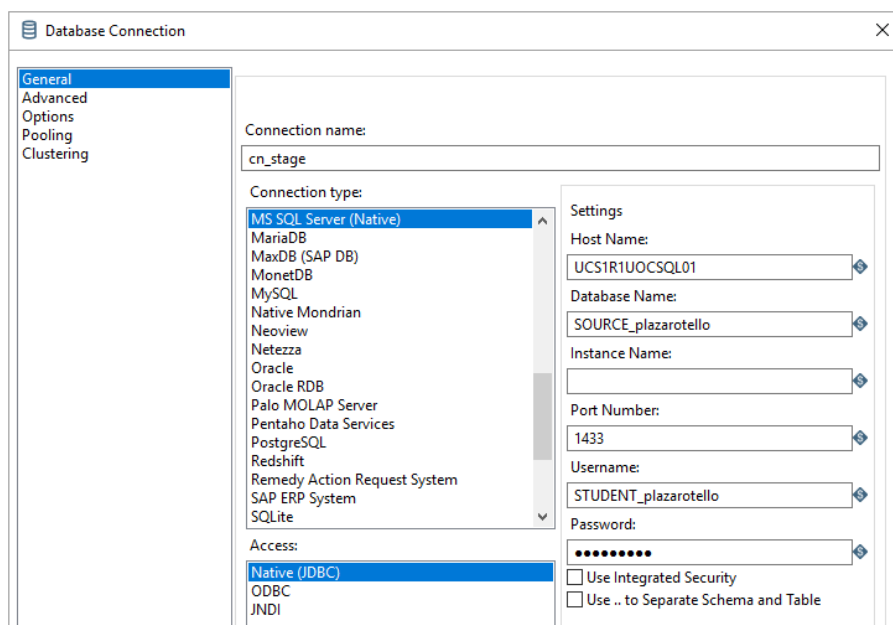


Figura 3: Conexión a la base de datos intermedia

4.3.3. Bloque IN

El bloque IN corresponde con las transformaciones que tienen como origen las fuentes de datos en bruto o *raw* y como destino la base de datos intermedia. A continuación se detallan los pasos seguidos para cada uno de los procesos.

IN_AMBIENTALPROTECTION

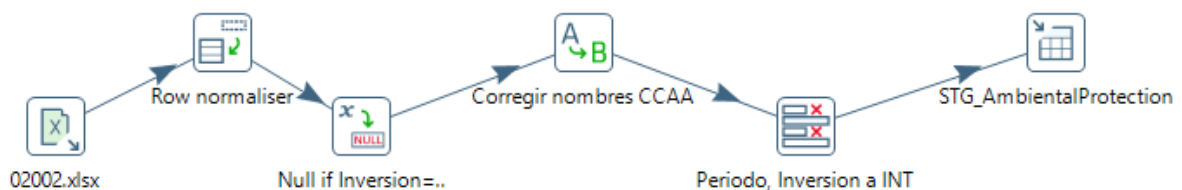


Figura 4: IN_AMBIENTALPROTECTION

En primer lugar, se procede a leer los datos del fichero fuente, “02002.xlsx”. Se leen todos los campos de la hoja “tabla-0” y para cada campo se recortan todos sus valores, para evitar posibles caracteres invisibles.

1

3

#	Name	Type	Length	Precision	Trim type
1	Periodo	String	-1	-1	both
2	Sector	String	-1	-1	both
3	Tipo	String	-1	-1	both
4	Ámbito	String	-1	-1	both
5	Andalucía	String	-1	-1	both
6	Aragón	String	-1	-1	both
7	Asturias, Principado de	String	-1	-1	both
8	Baleares, Illes	String	-1	-1	both
9	Canarias	String	-1	-1	both
10	Cantabria	String	-1	-1	both
11	Castilla y León	String	-1	-1	both
12	Castilla-La Mancha	String	-1	-1	both
13	Cataluña	String	-1	-1	both
14	Comunitat Valenciana	String	-1	-1	both
15	Extremadura	String	-1	-1	both
16	Galicia	String	-1	-1	both
17	Madrid, Comunidad de	String	-1	-1	both
18	Murcia, Región de	String	-1	-1	both
19	Navarra, Comunidad Foral de	String	-1	-1	both
20	País Vasco	String	-1	-1	both
21	Rioja, La	String	-1	-1	both
22	Total nacional	String	-1	-1	both

2

#	Sheet name	Start row	Start column
1	tabla-0	7	0

A continuación se transforman las columnas correspondientes con las Comunidades Autónomas.

Se crea un campo “ComunidadAutonoma” y los valores de las distintas columnas se almacenan en “Inversion”.

Por último, se cambian a «null» los valores de “Inversion” no numéricos, se cambia el formato de los valores de las Comunidades Autónomas, se cambian los tipos de “Inversion” y “Periodo” a entero y se vuelca el resultado en la base de datos del *staging area*.

Row normaliser

Step name: Row normaliser

Type field: ComunidadAutonoma

#	Fieldname	Type	new field
1	Andalucía	Andalucía	Inversion
2	Aragón	Aragón	Inversion
3	Asturias, Principado de	Asturias, Principado de	Inversion
4	Balears, Illes	Balears, Illes	Inversion
5	Canarias	Canarias	Inversion
6	Cantabria	Cantabria	Inversion
7	Castilla y León	Castilla y León	Inversion
8	Castilla-La Mancha	Castilla-La Mancha	Inversion
9	Cataluña	Cataluña	Inversion
10	Comunitat Valenciana	Comunitat Valenciana	Inversion
11	Extremadura	Extremadura	Inversion
12	Galicia	Galicia	Inversion
13	Madrid, Comunidad de	Madrid, Comunidad de	Inversion
14	Murcia, Región de	Murcia, Región de	Inversion
15	Navarra, Comunidad Foral de	Navarra, Comunidad Foral de	Inversion
16	País Vasco	País Vasco	Inversion
17	Rioja, La	Rioja, La	Inversion
18	Total nacional	Total nacional	Inversion

Buttons: Help, OK, Cancel, Get Fields

Null if

Step name: Null if Inversion=.

#	Name	Value to turn to NULL
1	Inversion	..

Buttons: Help, OK, Cancel, Get Fields

Replace in string

Step name: Corregir nombres CCAA

#	In stream field	Out stream field	use RegEx	Search	Replace with
1	ComunidadAutonoma		Y	^([^\,]*), ([^\,]*)\$	\$2 \$1

Buttons: Help

Select values

Step name: Periodo, Inversion a INT

Select & Alter | Remove | Meta-data

#	Fieldname	Rename to	Type
1	Periodo		Integer
2	Inversion		Integer

Buttons: Help, OK, Cancel

Table output

Step name: STG_AmbientalProtection

Connection: cn_stage

Target schema: dbo

Target table: STG_AmbientalProtection

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

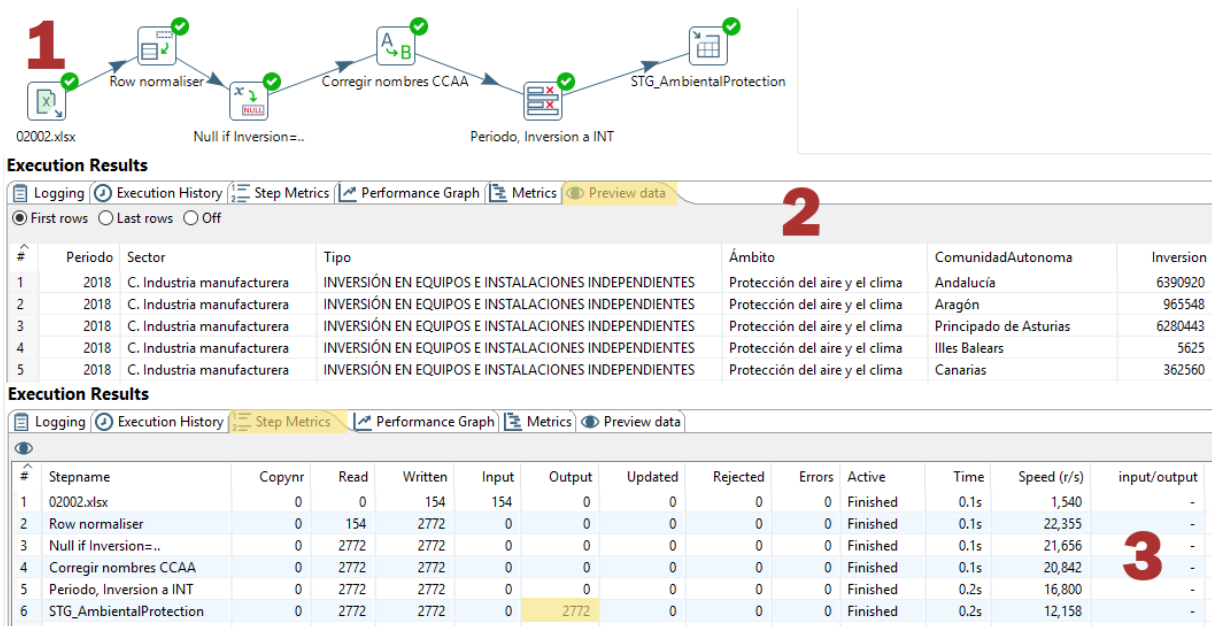
Specify database fields: ☒

Main options: Database fields

#	Table field	Stream field
1	Periodo	Periodo
2	Sector	Sector
3	Equipolnstalacion	Tipo
4	Ambito	Ámbito
5	ComunidadAutonoma	ComunidadAutonoma
6	Inversion	Inversion

Buttons: Get fields, Enter field mapping

Los resultados de ejecutar la transformación se encuentran en la siguiente figura:



Se han cargado 2.772 registros en STG_AmbientalProtection.

IN_COUNTRIES



Figura 5: IN_COUNTRIES

El fichero JSON con los datos de los países ya se encuentra correctamente formateado, por lo que el único paso necesario es leerlo y escribir en la tabla intermedia.

1

JSON input

Step name: Countries.json

File | Content | Fields | Additional output fields

Exclude Regular Expression:

Selected files:

#	File/Directory	Wildcard (Re)
1	\${DIR_IN}\Countries.json	

Show filename(s)...

3

Table output

Step name: STG_Countries

Connection: cn_stage

Target schema: dbo

Target table: STG_Countries

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

Specify database fields: ☒

Main options | Database fields

Fields to insert:

#	Table field	Stream field
1	Nombre	nombre
2	Name	name
3	Nom	nom
4	Iso2	iso2
5	Iso3	iso3
6	Phone	phone_code

Get fields

Enter field mapping

2

JSON input

Step name: Countries.json

File | Content | Fields | Additional output fields

#	Name	Path	Type	Trim type	Repeat
1	nombre	\$.nombre	String	both	N
2	name	\$.name	String	both	N
3	nom	\$.nom	String	both	N
4	iso2	\$.iso2	String	both	N
5	iso3	\$.iso3	String	both	N
6	phone_code	\$.phone_code	String	both	N

A continuación se muestran los resultados de la ejecución:

1

Countries.json

STG_Countries

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	nombre	name	nom	iso2	iso3	phone_code
1	Afganistán	Afghanistan	Afghanistan	AF	AFG	93
2	Albania	Albania	Albanie	AL	ALB	355
3	Alemania	Germany	Allemagne	DE	DEU	49
4	Algeria	Algeria	Algérie	DZ	DZA	213
5	Andorra	Andorra	Andorra	AD	AND	376

2

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Countries.json	0	0	246	246	0	0	0	0	Finished	0.1s	1,757	-
2	STG_Countries	0	246	246	0	246	0	0	0	Finished	0.5s	478	-

3

Se han introducido 246 registros en STG_Countries.

IN_ENVBIO1

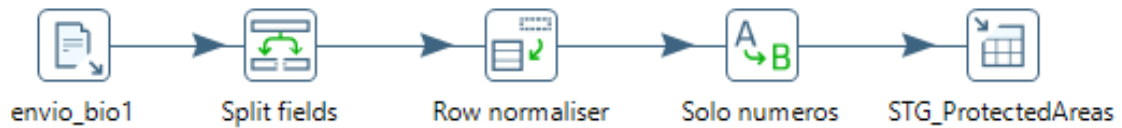


Figura 6: IN_ENVBIO1

Se obtienen los campos del TSV delimitados por tabuladores; se corta el campo areaprot_geo\time y se normalizan los años. Después se obtienen solo los valores numéricos y se introducen los datos transformados en la tabla STG_ProtectedAreas.

1 CSV file input

Step name: envio_bio1

Filename: \$(DIR_IN)\env_bio1.tsv

Delimiter: Insert TAB

Enclosure:

NIO buffer size: 50000

Lazy conversion? ☒

Header row present? ☒

Add filename to result? ☐

The row number field name:

Running in parallel? ☐

New line possible in fields? ☐

Format: mixed

File encoding:

#	Name	Type	Length	Decimal	Group	Trim type
1	areaprot_geo\time	String	18	,	.	both
2	2019	String	15	,	.	both
3	2018	String	15	,	.	both
4	2017	String	8	,	.	both
5	2016	String	8	,	.	both
6	2015	String	8	,	.	both
7	2014	String	8	,	.	both
8	2013	String	8	,	.	both
9	2012	String	7	,	.	both
10	2011	String	7	,	.	both

2 Split fields

Step name: Split fields

Field to split: areaprot_geo\time

Delimiter: ,

Enclosure:

#	New field	ID	Remove ID?	Type	Trim type
1	areaprot		N	String	none
2	geo		N	String	none

3 Row normaliser

Step name: Row normaliser

Type field: Year

#	Fieldname	Type	new field
1	2019	2019	Value
2	2018	2018	Value
3	2017	2017	Value
4	2016	2016	Value
5	2015	2015	Value
6	2014	2014	Value
7	2013	2013	Value
8	2012	2012	Value
9	2011	2011	Value

4 Replace in string

Step name: Solo numeros

#	In stream field	Out stream field	use RegEx	Search	Replace with	Whole Word
1	Value		Y	[^\d.]+/g		N

5 Table output

Step name: STG_ProtectedAreas

Connection: cn_stage

Target schema: dbo

Target table: STG_ProtectedAreas

Commit size: 1000

Truncate table: ☒

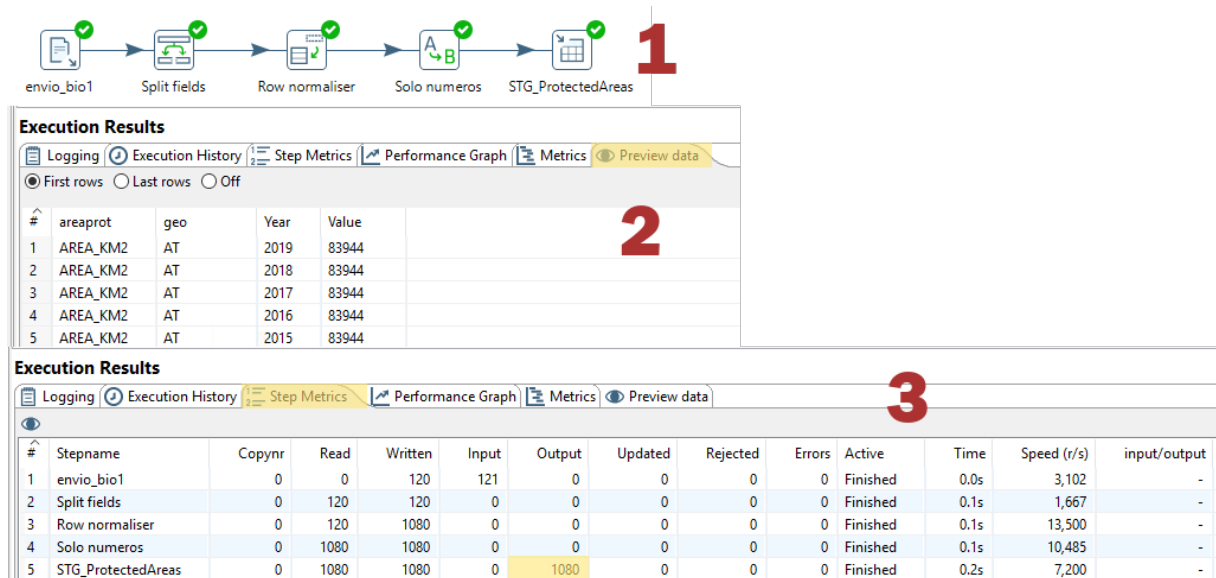
Ignore insert errors: ☐

Specify database fields: ☒

Main options: Database fields

#	Table field	Stream field
1	Unit	areaprot
2	geo	geo
3	Year	Year
4	Value	Value

Se muestran los resultados de la ejecución de la transformación:



Se han insertado 1.080 registros en la tabla STG_ProtectedAreas.

IN_OBJECTIVES



Figura 7: IN_OBJECTIVES

Los datos del fichero fuente ya se encuentran correctamente formateados; solo hay que cargarlos e insertarlos en la base de datos intermedia.

1 Microsoft Excel input

Step name: Objetivos Desarrollo sostenibles

Files | Sheets | Content | Error Handling | Fields | Additional output fields

Spread sheet type (engine): Excel 2007 XLSX (Apache POI)

File or directory: Add Browse...

Regular Expression:

Exclude Regular Expression:

Password:

Selected files: 1 \$DIR_IN\ODS.xlsx

4 Table output

Step name: Table output (STG_Objectives)

Connection: cn_stage Edit... New... Wizard...

Target schema: dbo Browse...

Target table: STG_Objectives Browse...

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

Specify database fields: ☒

Main options: Database fields

#	Table field	Stream field
1	Objetivo	Objetivo
2	Nombre	Nombre
3	Descripción	Descripción

Get fields Enter field mapping

2 Microsoft Excel input

Step name: Objetivos Desarrollo sostenibles

Files | Sheets | Content | Error Handling | Fields | Additional output fields

List of sheets to read: 1 ODS Sheet name Start row Start column

3 Microsoft Excel input

Step name: Objetivos Desarrollo sostenibles

Files | Sheets | Content | Error Handling | Fields | Additional output fields

#	Name	Type	Length	Precision	Trim type	Repeat	Format	C
1	Objetivo	Number	-1	-1	both	N		
2	Nombre	String	-1	-1	both	N		
3	Descripción	String	-1	-1	both	N		

Se muestran los resultados de la ejecución:

Objetivos Desarrollo sostenibles

→

Table output (STG_Objectives)

1

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

First rows Last rows Off

#	Objetivo	Nombre	Descripción
1	1.0	Fin de la pobreza	Para lograr este Objetivo de acabar con la pobreza, el crecimiento económico debe ser inclusivo, con el fin de crear empleos sostenibles y de promover la igualdad.
2	2.0	Hambre cero	El sector alimentario y el sector agrícola ofrecen soluciones claves para el desarrollo y son vitales para la eliminación del hambre y la pobreza.
3	3.0	Salud y bienestar	Para lograr los Objetivos de Desarrollo Sostenible, es fundamental garantizar una vida saludable y promover el bienestar universal.
4	4.0	Educación de calidad	La educación es la base para mejorar nuestra vida y el desarrollo sostenible.
5	5.0	Igualdad de género	La igualdad entre los géneros no es solo un derecho humano fundamental, sino la base necesaria para conseguir un mundo pacífico, próspero y sostenible.

2

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Objetivos Desarrollo sostenibles	0	0	17	17	0	0	0	0	Finished	0.1s	274	-
2	Table output (STG_Objectives)	0	17	17	0	17	0	0	0	Finished	0.1s	136	-

3

Se han insertado 17 registros en la tabla STG_Objectives.

IN_OBJECTIVESAREAS

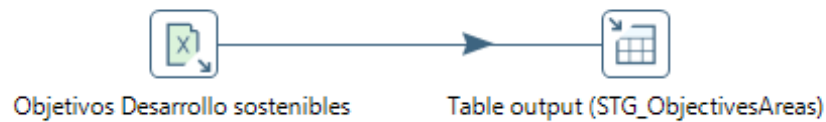


Figura 8: IN_OBJECTIVESAREAS

Los datos del fichero fuente ya se encuentran correctamente formateados; solo hay que cargarlos e insertarlos en la base de datos intermedia.

1 Microsoft Excel input

Step name:

#	Name	Type	Length	Precision	Trim type
1	Codigo	String	-1	-1	both
2	Ambito/VAR/Flow	String	-1	-1	both
3	ODS principal	Integer	-1	-1	both

2 Table output

Step name:

Connection:

Target schema:

Target table:

Commit size:

☒ Truncate table

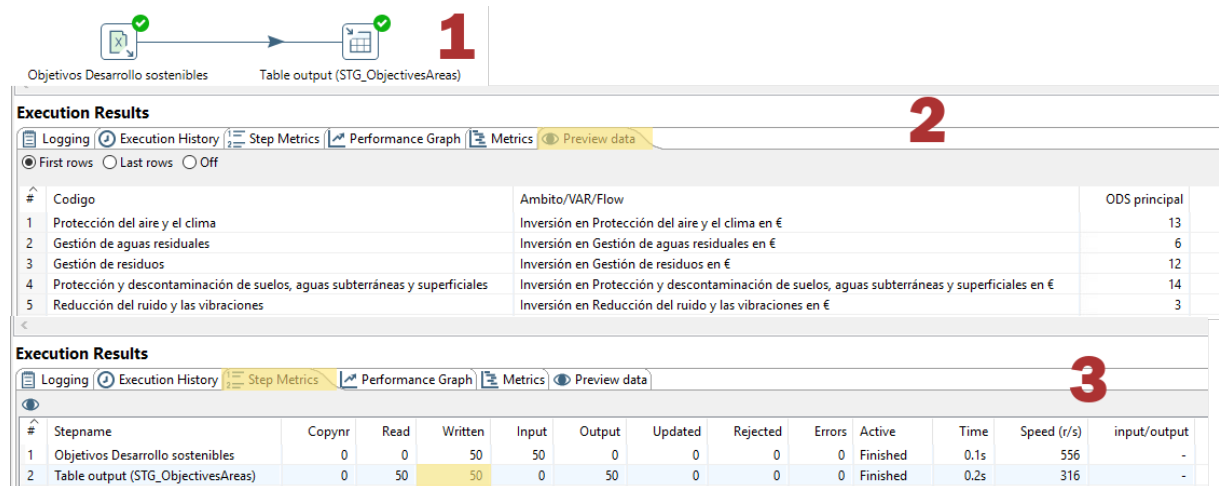
☐ Ignore insert errors

☒ Specify database fields

Main options: **Database fields**

#	Table field	Stream field
1	Codigo	Codigo
2	AmbitoVarFlow	Ambito/VAR/Flow
3	ODS principal	ODS principal

Se muestran los resultados de la ejecución:



Se han insertado 50 registros en la tabla STG_ObjectivesAreas.

IN_RESIDUOS



Figura 9: IN_RESIDUOS

Los datos del fichero fuente no se encuentran correctamente formateados; hay que cargarlos, eliminar las observaciones no numéricas e insertar los datos transformados en la base de datos intermedia.

1 Get data from XML

Step name: Get data from XML

#	Name	XPath	Element	Result type	Type	Format
1	COU	../[name()='SeriesKey']/[name()='Value' and @concept='COU']/value	Attribute	Value of	String	
2	VAR	../[name()='SeriesKey']/[name()='Value' and @concept='VAR']/value	Attribute	Value of	String	
3	TIME_FORMAT	../[name()='Attributes']/[name()='Value' and @concept='TIME_FORMAT']/value	Attribute	Value of	String	
4	UNIT	../[name()='Attributes']/[name()='Value' and @concept='UNIT']/value	Attribute	Value of	String	
5	POWERCODE	../[name()='Attributes']/[name()='Value' and @concept='POWERCODE']/value	Attribute	Value of	Integer	
6	Time	*[name()='Time']	Node	Value of	Integer	
7	Obs	*[name()='ObsValue']/value	Attribute	Value of	String	
8	OBS_STATUS	*[name()='Attributes']/[name()='Value' and @concept='OBS_STATUS']/value	Attribute	Value of	String	

2 Replace in string

Step name: Replace in string

#	In stream field	Out stream field	use RegEx	Search	Replace with
1	Obs		Y	[a-zA-Z\s+]/g	

3 Select values

Step name: Select values

#	Fieldname	Rename to	Type	Length	Precision	Binary
1	COU		None			N
2	VAR		None			N
3	TIME_FORMAT		None			N
4	UNIT		None			N
5	POWERCODE		Integer		0	N
6	Time		Integer		0	N
7	Obs		Number			N
8	OBS_STATUS		None			N

4 Table output

Step name: Table output

Connection: cn_stage

Target schema: dbo

Target table: STG_Residuos

Commit size: 1000

☒ Truncate table

☐ Ignore insert errors

☒ Specify database fields

Main options: Database fields

#	Table field	Stream field
1	Cou	COU
2	Var	VAR
3	Unit	UNIT
4	Powercode	POWERCODE
5	Obs	Obs
6	TimeFormat	TIME_FORMAT
7	Year	Time
8	Status	OBS_STATUS

Se muestran los resultados de la ejecución:

1

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

☒ First rows ☐ Last rows ☐ Off

#	COU	VAR	TIME_FORMAT	UNIT	POWERCODE	Time	Obs	OBS_STATUS
1	AUS	MUNICIPAL	P1Y	TONNE	3	1992	12000.0	E
2	AUS	MUNICIPAL	P1Y	TONNE	3	2000	13200.0	E
3	AUS	MUNICIPAL	P1Y	TONNE	3	2007	12873.0	
4	AUS	MUNICIPAL	P1Y	TONNE	3	2008	13096.5	
5	AUS	MUNICIPAL	P1Y	TONNE	3	2009	13320.0	

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

☒ First rows ☐ Last rows ☐ Off

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Get data from XML	0	0	20779	20779	0	0	0	0	Finished	7.5s	2,766	-
2	Replace in string	0	20779	20779	0	0	0	0	0	Finished	7.5s	2,764	-
3	Select values	0	20779	20779	0	0	0	0	0	Finished	7.5s	2,762	-
4	Table output	0	20779	20779	0	20779	0	0	0	Finished	7.6s	2,745	-

Se han insertado 20.779 registros en la tabla STG_Residuos.

IN_WORLD_ENERGY_BALANCES



Figura 10: IN_WORLD_ENERGY_BALANCES

Los datos del fichero fuente no se encuentran correctamente formateados; hay que cargarlos, normalizar las columnas de los años, eliminar las observaciones no numéricas e insertar los datos transformados en la base de datos intermedia. Además, se han transformado 2 valores perdidos de países.

1 Row normaliser

Step name: Normalización filas

Type field: Year

#	Fieldname	Type	new field
47	2017	2017	Value
48	2018	2018	Value
49	2019 Provisional	2019	Value

2 Null if

Step name: Nulo SI

#	Name	Value to turn to NULL
1	Value	..

3 Replace in string

Step name: Eliminar c

#	In stream field	Out stream field	use RegEx	Search	Replace with	Set e
1	Value		Y	[a-zA-Z/s+/g]		N

4 Value mapper

Step name: Value mapper

Fieldname to use: Country

Target field name:

Default upon:

#	Source value	Target value
1	United States	United States of America
2	Slovak	Slovakia

5 Table output

Step name: Salida Tabla

Connection: cn_stage

Target schema: dbo

Target table: STG_EnergyBalance

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

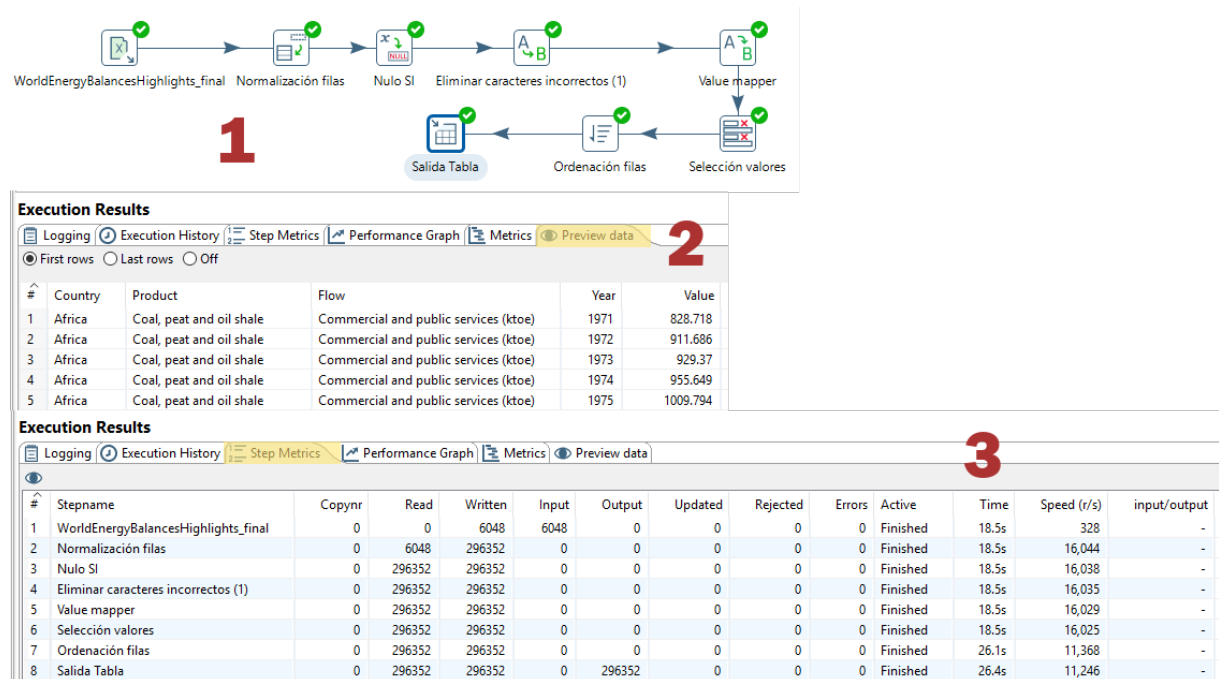
Specify database fields: ☒

Main options: Database fields

#	Table field	Stream field
1	country	Country
2	product	Product
3	flow	Flow

Buttons: OK, Cancel, SQL

Se muestran los resultados de la ejecución:



Se han insertado 296.352 registros en la tabla STG_EnergyBalance.

4.3.4. Bloque TR_DIM

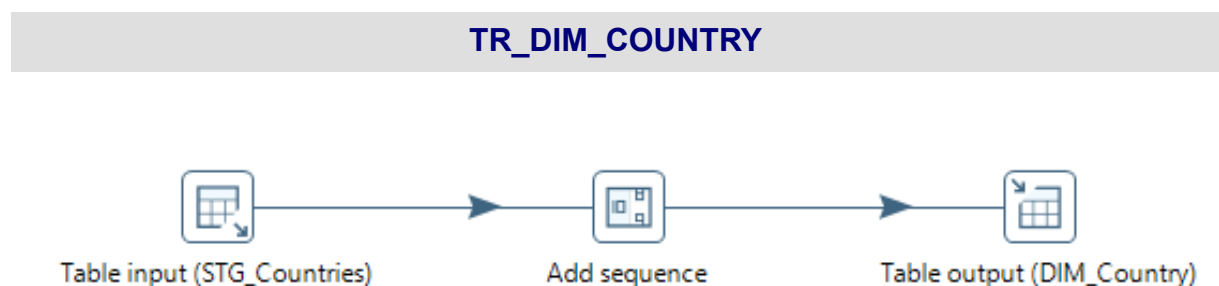


Figura 11: TR_DIM_COUNTRY

Los datos de la tabla intermedia se encuentran correctamente formateados. El único campo a crear es la clave primaria única, que se construirá con un nodo «Add sequence».

Table input

Step name

Table input (STG_Countries)

Connection

cn_s

Edit...

New...

Wizard...

SQL

Get SQL select statement...

```

SELECT
  Nombre
, Name
, Nom
, Iso2
, Iso3
, Phone
FROM dbo.STG_Countries
order by Iso2

```

Line 1 Column 0

Store column info

Enable lazy

Disable variables in

1

Add sequence

Step name

Add sequence

Name of value

pk

Use a database to generate the sequence

Use DB to get

Connection

Edit...

New...

Wizard...

Schema name

Schema...

Sequence

SEQ

Sequences...

Use a transformation counter to generate the s

Use counter to

Counter name

Start at value

Increment by

Maximum

2

Table output

Step name

Table output (DIM_Country)

Connection

cn_dw

Edit...

New...

Wizard...

Target schema

dbo

Browse...

Target table

DIM_Country

Browse...

Commit size

1000

Truncate table

Ignore insert errors

Specify database fields

Main options

Database fields

Fields to insert:

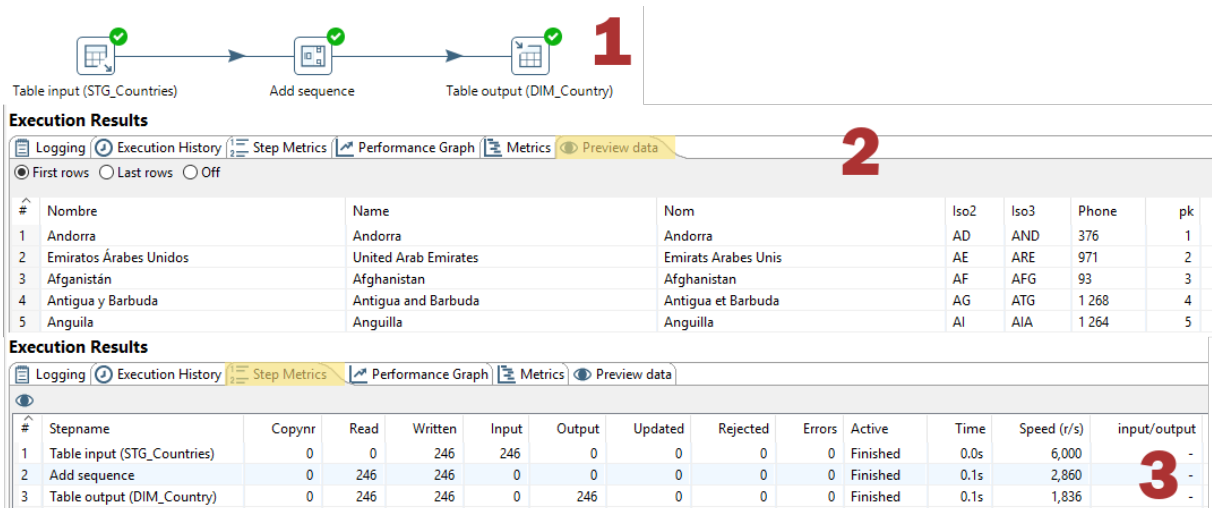
#	Table field	Stream field
1	country_phone_code	Phone
2	pk_country	pk
3	country_code	Iso2
4	country_code3	Iso3
5	country_name_fr	Nom
6	country_name_en	Name
7	country_name_sp	Nombre

Get fields

Enter field mapping

3

Se muestran los resultados de la ejecución:



Se han insertado 246 registros en la tabla DIM_Country.

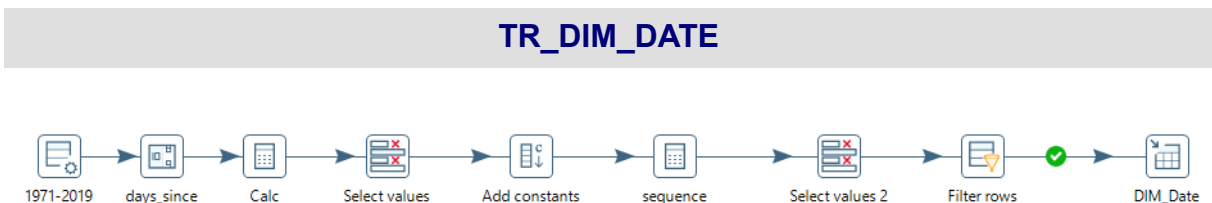
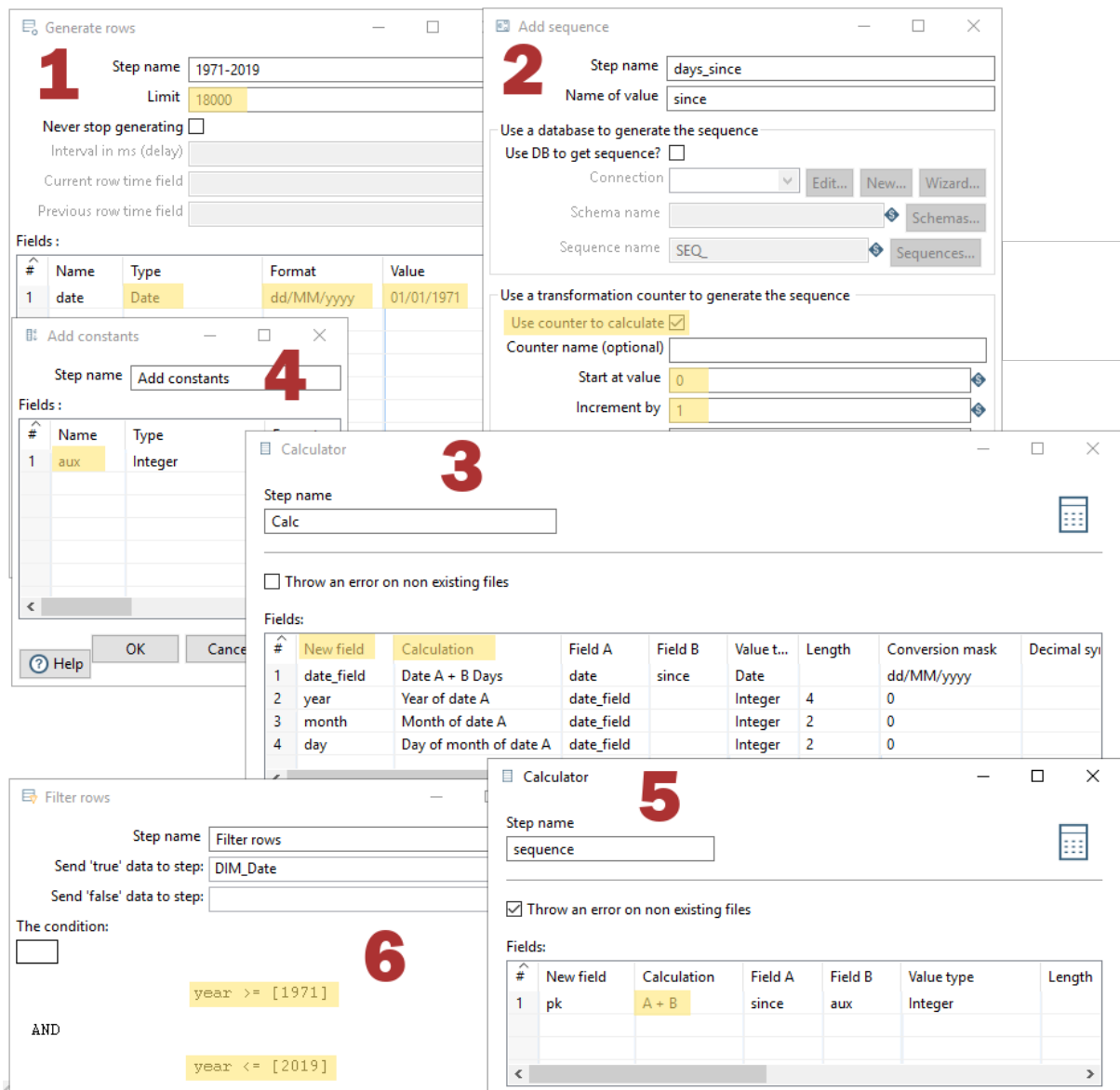


Figura 12: TR_DIM_DATE

No se cuenta con datos en las tablas del *staging area*, sino que hay que crear las fechas manualmente. Se sabe que los datos están disponibles en el rango de años 1971-2019; por tanto, hay que generar registros para cada día entre esos años.

En primer lugar se crean $(2019 - 1971) \times (365 + 1) = 17,568 \approx 18,000$ registros con la fecha del primer día, 1 de enero de 1971. A cada registro se le añade una secuencia, que comienza en 0, para después sumar a la fecha la secuencia y obtener el día, mes y año de la fecha final.

Por último, para crear la clave primaria se utiliza la secuencia que comienza en 0 y se le suma 1; así se obtiene un número positivo único para cada registro; y se seleccionan solo los valores en el rango 1971-2019.



1 Generate rows

Step name: 1971-2019
Limit: 18000
Never stop generating: ☐
Interval in ms (delay):
Current row time field:
Previous row time field:

Fields:

#	Name	Type	Format	Value
1	date	Date	dd/MM/yyyy	01/01/1971

2 Add sequence

Step name: days_since
Name of value: since

Use a database to generate the sequence
Use DB to get sequence? ☐
Connection:
Schema name:
Sequence name: SEQ

Use a transformation counter to generate the sequence
Use counter to calculate: ☒
Counter name (optional):
Start at value: 0
Increment by: 1

3 Calculator

Step name: Calc

☐ Throw an error on non existing files

Fields:

#	New field	Calculation	Field A	Field B	Value t...	Length	Conversion mask	Decimal sy
1	date_field	Date A + B Days	date	since	Date		dd/MM/yyyy	
2	year	Year of date A	date_field		Integer	4	0	
3	month	Month of date A	date_field		Integer	2	0	
4	day	Day of month of date A	date_field		Integer	2	0	

4 Add constants

Step name: Add constants

Fields:

#	Name	Type
1	aux	Integer

5 Calculator

Step name: sequence

☒ Throw an error on non existing files

Fields:

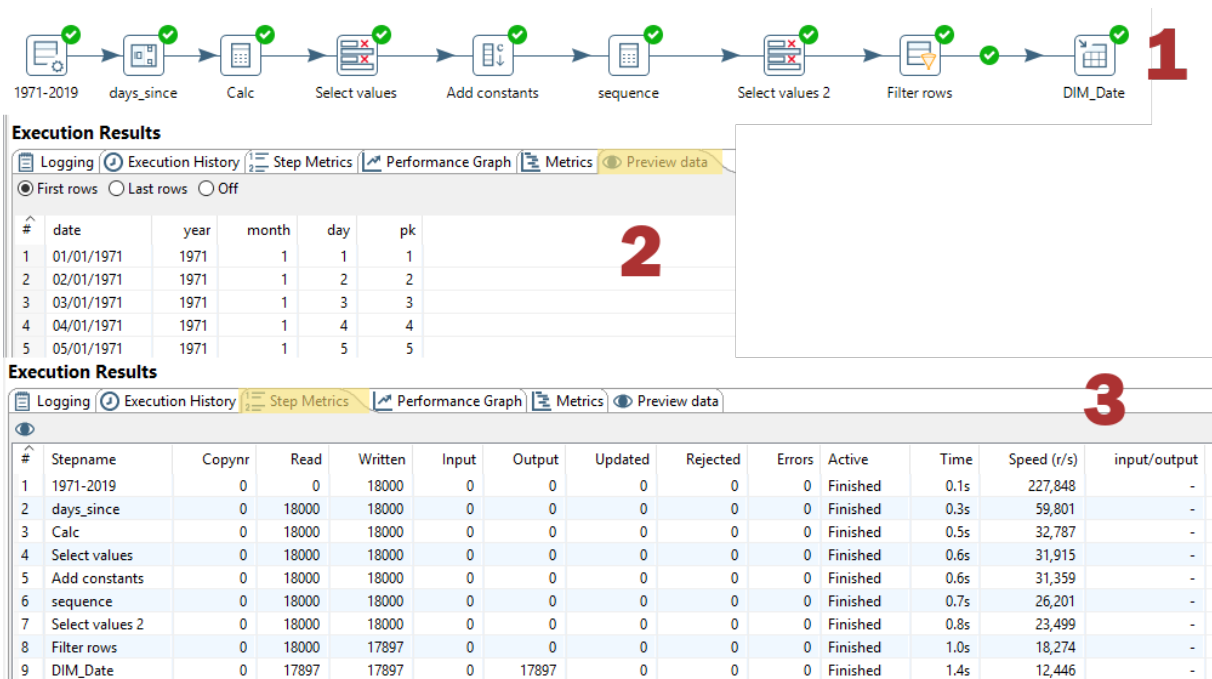
#	New field	Calculation	Field A	Field B	Value type	Length
1	pk	A + B	since	aux	Integer	

6 Filter rows

Step name: Filter rows
Send 'true' data to step: DIM_Date
Send 'false' data to step:

The condition:
☐ year >= [1971]
AND
☐ year <= [2019]

Se muestran los resultados de la ejecución:



Se han insertado 17.897 registros en la tabla DIM_Date.

TR_DIM_ECONOMICACTIVITYSECTOR

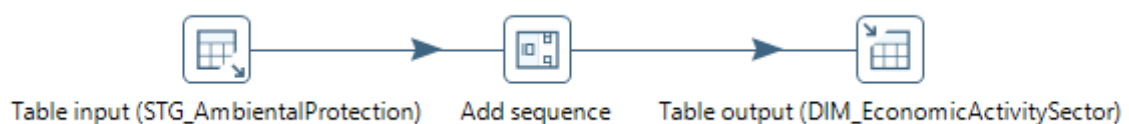


Figura 13: TR_DIM_ECONOMICACTIVITYSECTOR

Se procede a extraer los distintos sectores de la tabla intermedia correspondiente a la inversión en protección ambiental por las Comunidades Autónomas. Se añade también un registro “NA” para los hechos que no tengan sector asociado.

1

Step name: ble input (STG_AmbientalProtection)

Connection: cn_stage

SQL

```
SELECT distinct Sector
FROM dbo.STG_AmbientalProtection
union
select 'NA'--, 'NOT AVAILABLE'
```

2

Step name: Add sequence

Name of value: pk

Use a database to generate the sequence

Use DB to get: ☐

Connection: Edit... New... Wizard...

Schema name: Schemas...

Sequence name: SEQ_ Sequences...

Use a transformation counter to generate the sequence

Use counter to: ☒

Counter name:

Start at value: 1

Increment by: 1

3

Step name: Table output (DIM_EconomicActivitySector)

Connection: cn_dw Edit... New... Wizard...

Target schema: dbo Browse...

Target table: DIM_EconomicActivitySector Browse...

Commit size: 1000

Truncate table: ☐

Ignore insert errors: ☐

Specify database fields: ☒

Main options: Database fields

#	Table field	Stream field
1	activitysect...	Sector
2	pk_activitys...	pk

Get fields

Enter field mapping

Se muestran los resultados de la ejecución:

1

Table input (STG_AmbientalProtection) Add sequence Table output (DIM_EconomicActivitySector)

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

First rows Last rows Off

2

#	Sector	pk
1	C. Industria manufacturera	1
2	NA	2

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

3

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Table input (STG_AmbientalProtection)	0	0	2	2	0	0	0	0	Finished	0.0s	87	-
2	Add sequence	0	2	2	0	0	0	0	0	Finished	0.0s	45	-
3	Table output (DIM_EconomicActivitySector)	0	2	2	0	2	0	0	0	Finished	0.1s	27	-

Se han insertado 2 registros en la tabla DIM_EconomicActivitySector.

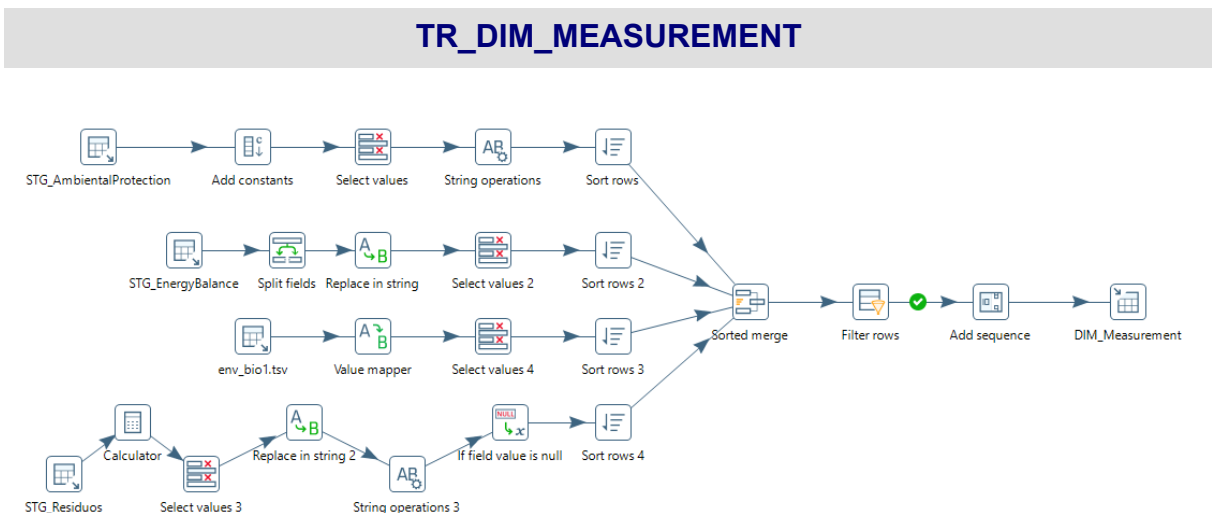
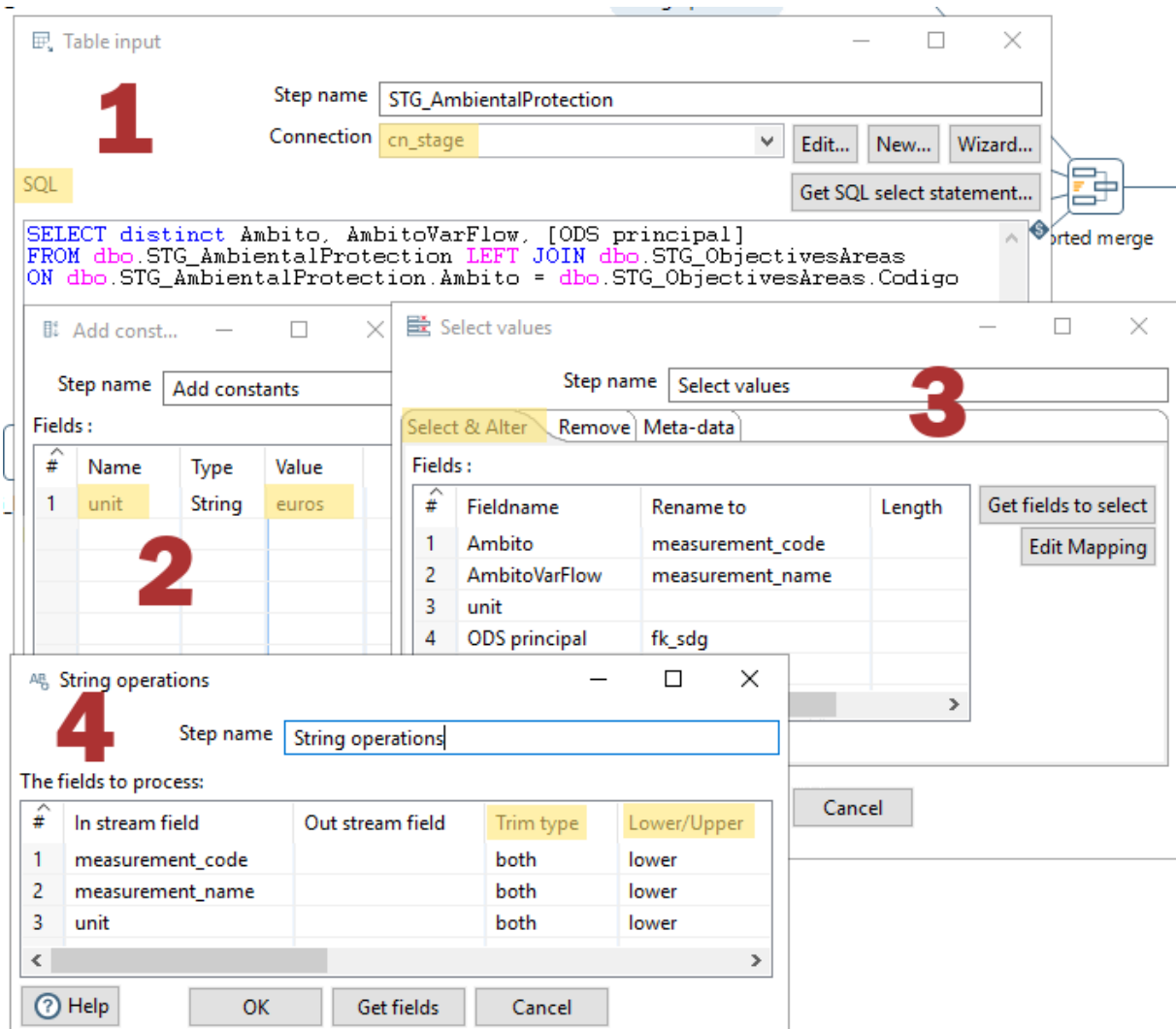


Figura 14: TR_DIM_MEASUREMENT

Para obtener los tipos de mediciones que registra el modelo multidimensional, es necesario cruzar los datos de varias tablas intermedias, concretamente las que tienen datos de mediciones y unidades de las mismas.

Para el flujo de datos de la tabla de protección ambiental por Comunidad Autónoma, se han de cruzar sus datos con las áreas de los Objetivos de Desarrollo Sostenible (ODS). Todas las medidas se encuentran en €, por lo que se añade también una constante para representarlo. Por último, se limpian las cadenas de texto y se convierten en minúsculas.



1 Step name: STG_AmbientalProtection
Connection: cn_stage

SQL

```
SELECT distinct Ambito, AmbitoVarFlow, [ODS principal]
FROM dbo.STG_AmbientalProtection LEFT JOIN dbo.STG_ObjectivesAreas
ON dbo.STG_AmbientalProtection.Ambito = dbo.STG_ObjectivesAreas.Codigo
```

2 Step name: Add constants

#	Name	Type	Value
1	unit	String	euros

3 Step name: Select values

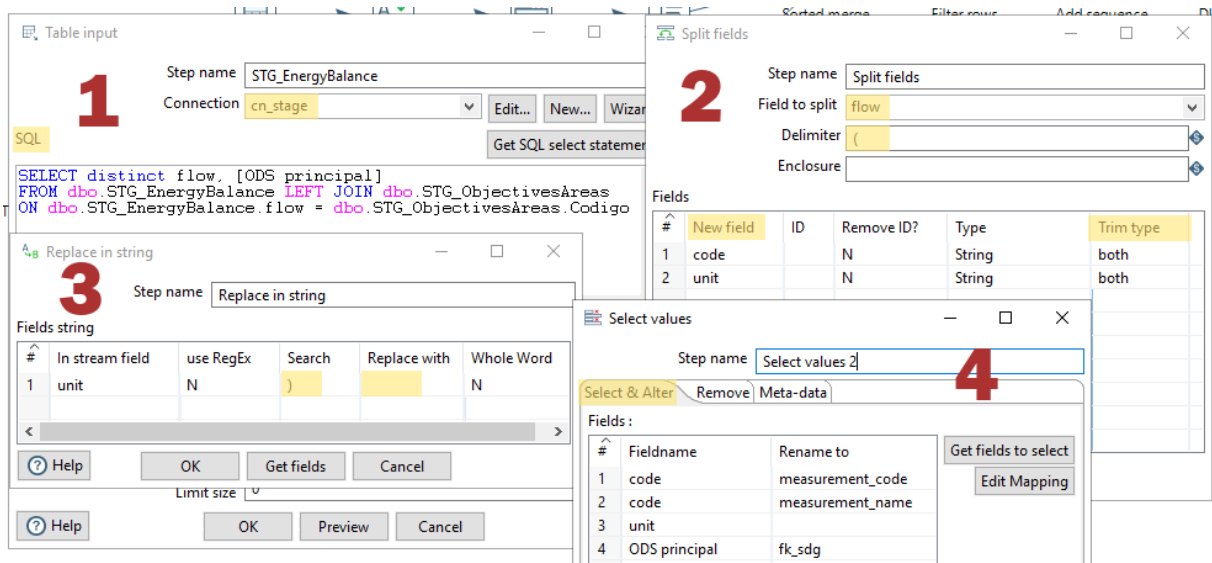
#	Fieldname	Rename to	Length
1	Ambito	measurement_code	
2	AmbitoVarFlow	measurement_name	
3	unit		
4	ODS principal	fk_sdg	

4 Step name: String operations

The fields to process:

#	In stream field	Out stream field	Trim type	Lower/Upper
1	measurement_code		both	lower
2	measurement_name		both	lower
3	unit		both	lower

Para el flujo de datos de la tabla de balances energéticos, se ha de desglosar el campo “flow” en el código del ODS y las unidades. Este campo presenta la siguiente estructura: “código (unidades)”. Se separa el campo por “(” y se elimina de las unidades “)”.



1 Step name: STG_EnergyBalance
Connection: cn_stage
SQL: `SELECT distinct flow, [ODS principal]
FROM dbo.STG_EnergyBalance LEFT JOIN dbo.STG_ObjectivesAreas
ON dbo.STG_EnergyBalance.flow = dbo.STG_ObjectivesAreas.Codigo`

2 Step name: Split fields
Field to split: flow
Delimiter: (
Enclosure:

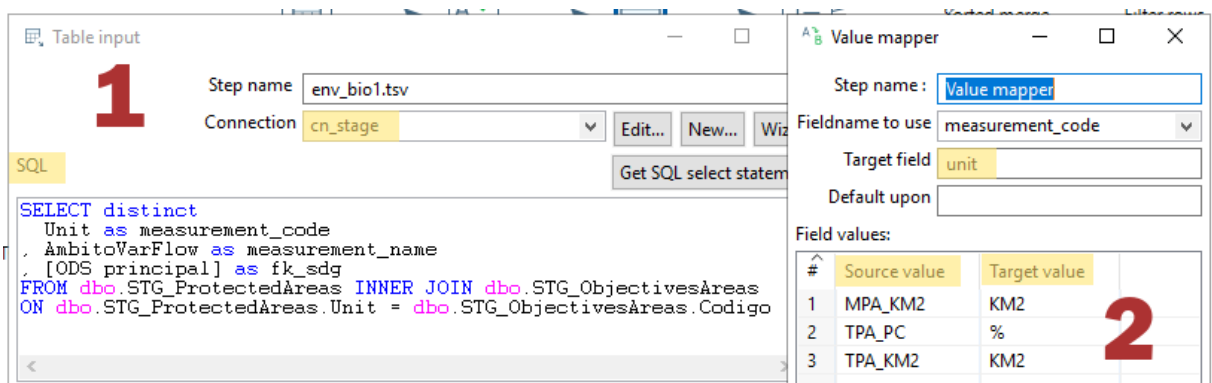
3 Step name: Replace in string
Fields string:

#	In stream field	use RegEx	Search	Replace with	Whole Word
1	unit	N)		N

4 Step name: Select values 2
Fields:

#	Fieldname	Rename to
1	code	measurement_code
2	code	measurement_name
3	unit	
4	ODS principal	fk_sdg

Para el flujo de datos de áreas protegidas se han de mapear los códigos del ODS, que contienen embebida la información de la unidades, con las unidades.



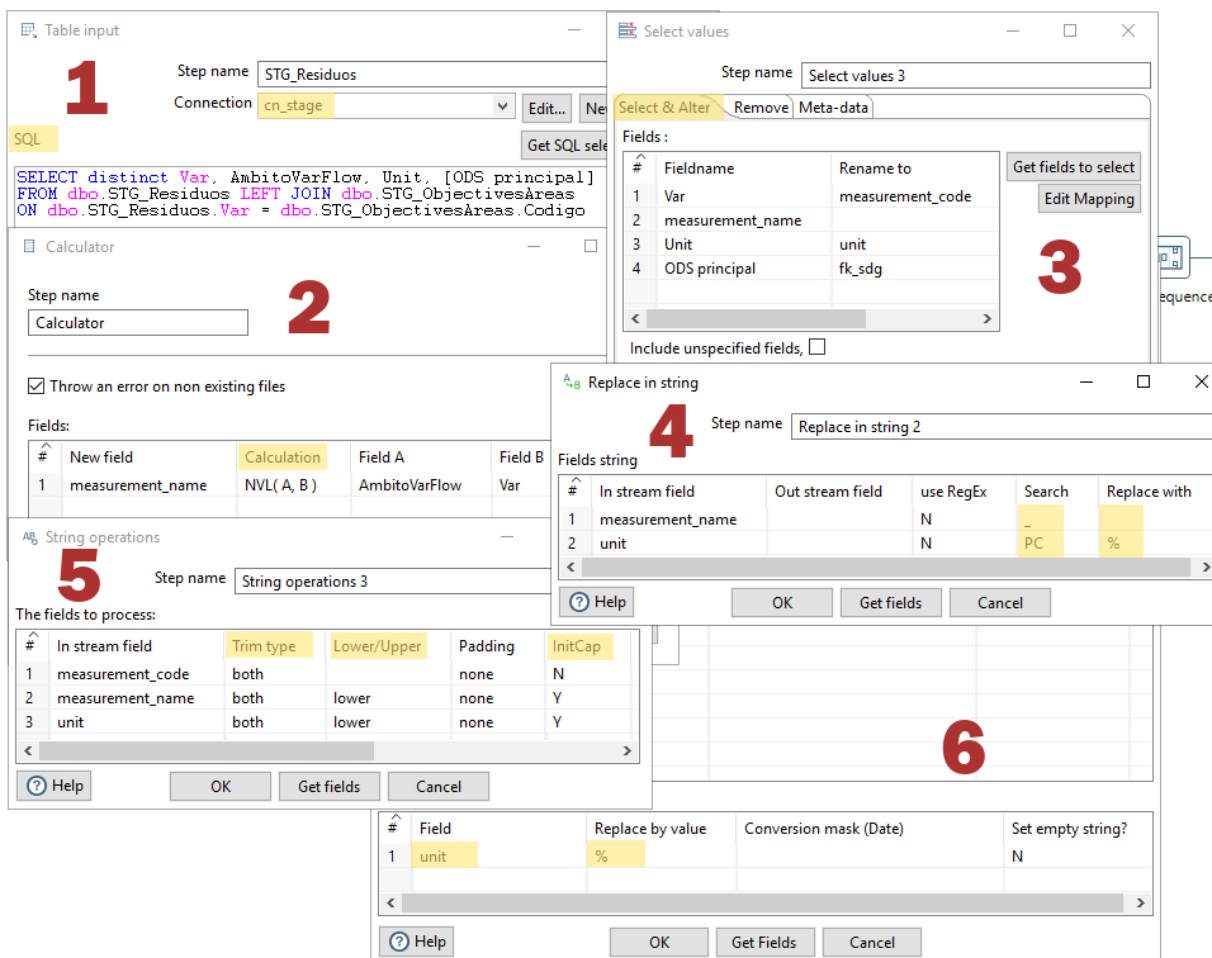
1 Step name: env_bio1.tsv
Connection: cn_stage
SQL: `SELECT distinct
Unit as measurement_code
, AmbitoVarFlow as measurement_name
, [ODS principal] as fk_sdg
FROM dbo.STG_ProtectedAreas INNER JOIN dbo.STG_ObjectivesAreas
ON dbo.STG_ProtectedAreas.Unit = dbo.STG_ObjectivesAreas.Codigo`

2 Step name: Value mapper
Fieldname to use: measurement_code
Target field: unit
Field values:

#	Source value	Target value
1	MPA_KM2	KM2
2	TPA_PC	%
3	TPA_KM2	KM2

Por último, el flujo de datos de residuos tiene que cruzarse con las áreas del ODS de la *staging area*. Algunos valores de la tabla de residuos no tienen contraparte en la tabla de áreas del ODS; para rellenar estos valores restantes se ejecuta una operación «NVL» que rellenará “AmbitoVarFlow” con “Var” en caso de que no exista la primera.

El siguiente paso consiste en renombrar los campos del *stream* y cambiar algunos valores de “measurement_name” (debido al «NVL» se han insertado caracteres incorrectos) y transformar “PC” a “%” en las unidades. Por último se normalizarán las cadenas de texto y se cambiarán los valores nulos en las unidades a “%”.



The screenshot displays the Alteryx Designer interface with six numbered steps in a workflow:

- Table input:** Step name 'STG_Residuos', Connection 'cn_stage'. The SQL query is:


```
SELECT distinct Var, AmbitoVarFlow, Unit, [ODS principal]
FROM dbo.STG_Residuos LEFT JOIN dbo.STG_ObjectivesAreas
ON dbo.STG_Residuos.Var = dbo.STG_ObjectivesAreas.Codigo
```
- Calculator:** Step name 'Calculator'. The 'Fields' table is:

#	New field	Calculation	Field A	Field B
1	measurement_name	NVL(A, B)	AmbitoVarFlow	Var
- Select values:** Step name 'Select values 3'. The 'Fields' table is:

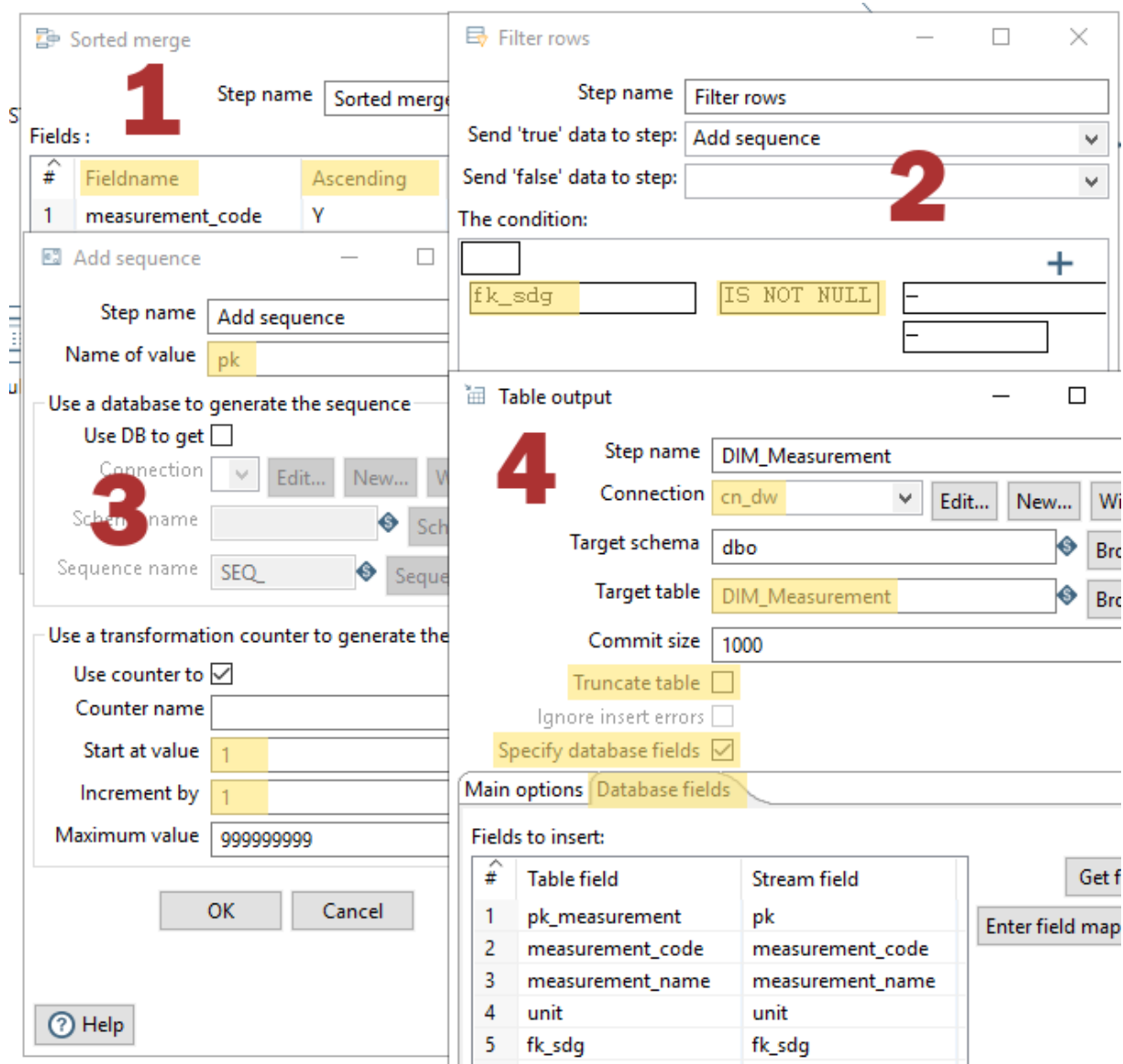
#	Fieldname	Rename to
1	Var	measurement_code
2	measurement_name	unit
3	Unit	fk_sdg
4	ODS principal	
- String operations:** Step name 'String operations 3'. The 'Fields to process' table is:

#	In stream field	Trim type	Lower/Upper	Padding	InitCap
1	measurement_code	both		none	N
2	measurement_name	both	lower	none	Y
3	unit	both	lower	none	Y
- Replace in string:** Step name 'Replace in string 2'. The 'Fields string' table is:

#	In stream field	Out stream field	use RegEx	Search	Replace with
1	measurement_name		N	-	
2	unit		N	PC	%
- Unlabeled step:** A table with the following data:

#	Field	Replace by value	Conversion mask (Date)	Set empty string?
1	unit	%		N

En último lugar se procede a unir los flujos de datos creados anteriormente, ordenándolos por código de medida, se filtran aquellas filas que no tengan ODS asociado, se crea la clave primaria única con una secuencia y se introduce el resultado en la tabla.



1 Sorted merge

Step name: Sorted merge

Fields:

#	Fieldname	Ascending
1	measurement_code	Y

2 Filter rows

Step name: Filter rows

Send 'true' data to step: Add sequence

Send 'false' data to step:

The condition:

fk_sdg	IS NOT NULL	-
--------	-------------	---

3 Add sequence

Step name: Add sequence

Name of value: pk

Use a database to generate the sequence

Use DB to get ☐

Connection: Edit... New... W

Schema name: \$ Sch

Sequence name: SEQ_ \$ Seque

Use a transformation counter to generate the

Use counter to ☒

Counter name:

Start at value: 1

Increment by: 1

Maximum value: 99999999

OK Cancel

4 Table output

Step name: DIM_Measurement

Connection: cn_dw Edit... New... Wi

Target schema: dbo \$ Bro

Target table: DIM_Measurement \$ Bro

Commit size: 1000

Truncate table ☐

Ignore insert errors ☐

Specify database fields ☒

Main options Database fields

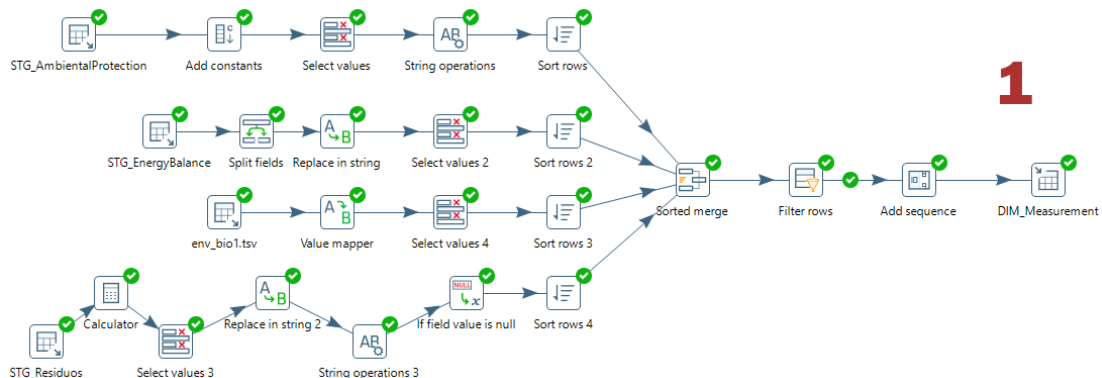
Fields to insert:

#	Table field	Stream field
1	pk_measurement	pk
2	measurement_code	measurement_code
3	measurement_name	measurement_name
4	unit	unit
5	fk_sdg	fk_sdg

Get f

Enter field map

Se muestran los resultados de la ejecución:



1

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

First rows Last rows Off

#	measurement_code	measurement_name	unit	fk_sdg	pk
1	BULKY	Bulky Waste	Tonne	9	1
2	Commercial and public services	Commercial and public services	ktoe	12	2
3	COMPOST	Composting	Tonne	7	3
4	COMPOST_SHARE	% Composting	%	7	4
5	DISP_SHARE	% Disposal	%	7	5

2

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	STG_Residuos	0	0	29	29	0	0	0	0	Finished	0.1s	244	-
2	Calculator	0	29	29	0	0	0	0	0	Finished	0.1s	284	-
3	Select values 3	0	29	29	0	0	0	0	0	Finished	0.1s	242	-
4	STG_AmbientalProtection	0	0	7	7	0	0	0	0	Finished	0.1s	89	-
5	STG_EnergyBalance	0	0	13	13	0	0	0	0	Finished	0.3s	49	-
6	env_bio1.tsv	0	0	3	3	0	0	0	0	Finished	0.1s	38	-
7	Add constants	0	7	7	0	0	0	0	0	Finished	0.1s	85	-
8	Split fields	0	13	13	0	0	0	0	0	Finished	0.3s	47	-
9	Select values	0	7	7	0	0	0	0	0	Finished	0.1s	66	-
10	Replace in string 2	0	29	29	0	0	0	0	0	Finished	0.1s	234	-
11	String operations 3	0	29	29	0	0	0	0	0	Finished	0.1s	216	-
12	Value mapper	0	3	3	0	0	0	0	0	Finished	0.1s	38	-
13	String operations	0	7	7	0	0	0	0	0	Finished	0.1s	58	-
14	Sort rows	0	7	7	0	0	0	0	0	Finished	0.1s	57	-
15	Replace in string	0	13	13	0	0	0	0	0	Finished	0.3s	46	-
16	Select values 2	0	13	13	0	0	0	0	0	Finished	0.3s	46	-
17	If field value is null	0	29	29	0	0	0	0	0	Finished	0.1s	212	-
18	Sort rows 2	0	13	13	0	0	0	0	0	Finished	0.3s	41	-
19	Select values 4	0	3	3	0	0	0	0	0	Finished	0.1s	29	-
20	Sort rows 3	0	3	3	0	0	0	0	0	Finished	0.1s	28	-
21	Sort rows 4	0	29	29	0	0	0	0	0	Finished	0.2s	172	-
22	Sorted merge	0	52	52	0	0	0	0	0	Finished	1.1s	49	-
23	Filter rows	0	52	50	0	0	0	0	0	Finished	1.1s	48	-
24	Add sequence	0	50	50	0	0	0	0	0	Finished	1.1s	46	-
25	DIM_Measurement	0	50	50	0	50	0	0	0	Finished	1.1s	44	-

3

Se han insertado 50 registros en la tabla DIM_Measurement.

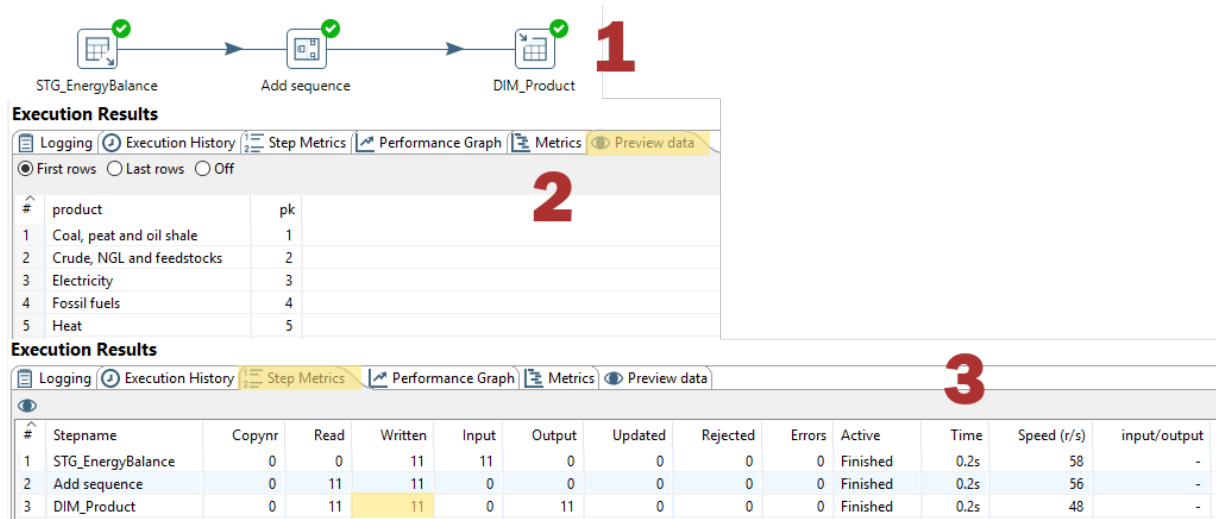
TR_DIM_PRODUCT



Figura 15: TR_DIM_PRODUCT

Se recogen los productos distintos de la tabla de balance energético, se añade una clave primaria numérica con «Add sequence» y se introducen en la tabla de la dimensión del producto en el modelo multidimensional. A continuación se muestran los resultados de la ejecución:

1



2

3

#	product	pk
1	Coal, peat and oil shale	1
2	Crude, NGL and feedstocks	2
3	Electricity	3
4	Fossil fuels	4
5	Heat	5

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	STG_EnergyBalance	0	0	11	11	0	0	0	0	Finished	0.2s	58	-
2	Add sequence	0	11	11	0	0	0	0	0	Finished	0.2s	56	-
3	DIM_Product	0	11	11	0	11	0	0	0	Finished	0.2s	48	-

Se han insertado 11 registros en la tabla DIM_Product.

TR_DIM_REGION

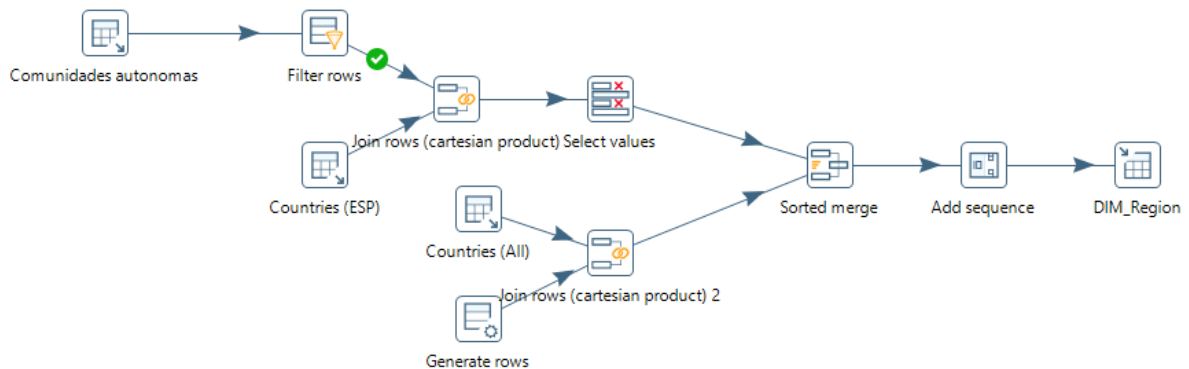
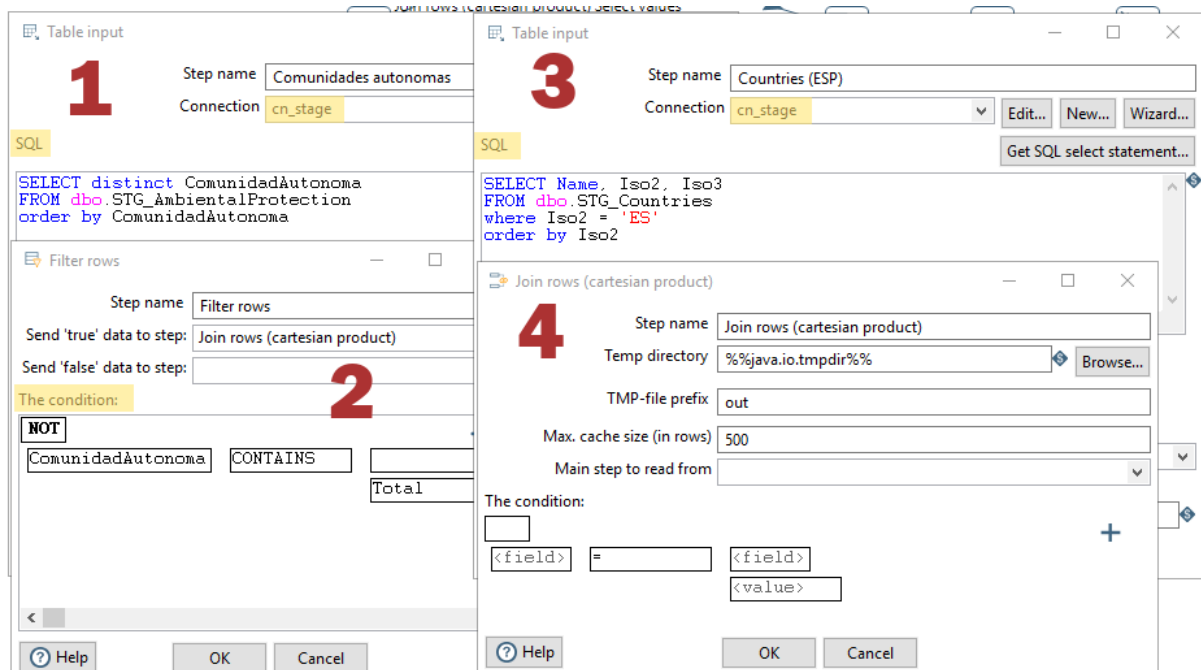


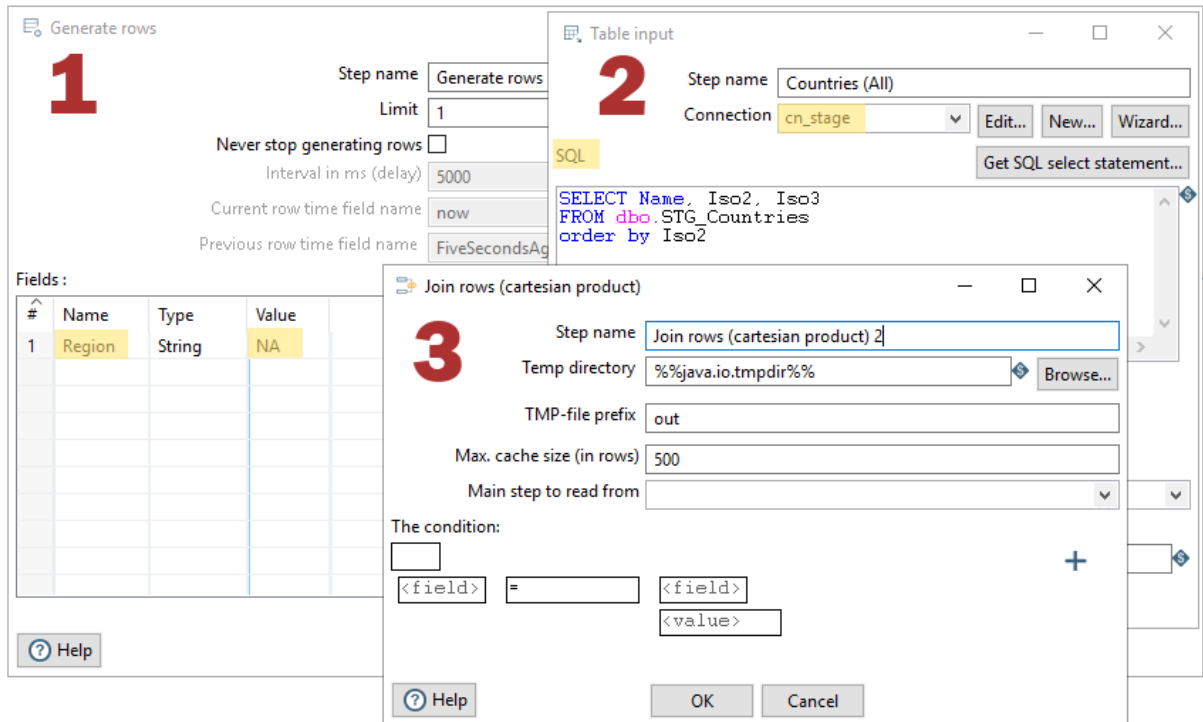
Figura 16: TR_DIM_REGION

Para completar la tabla de regiones, se necesita la información de las Comunidades Autónomas y rellenar con “NA” el resto de países. En primer lugar se procede con el flujo de datos correspondiente a las Comunidades Autónomas, rellenando los campos restantes con la información de España de la tabla intermedia STG_Countries.



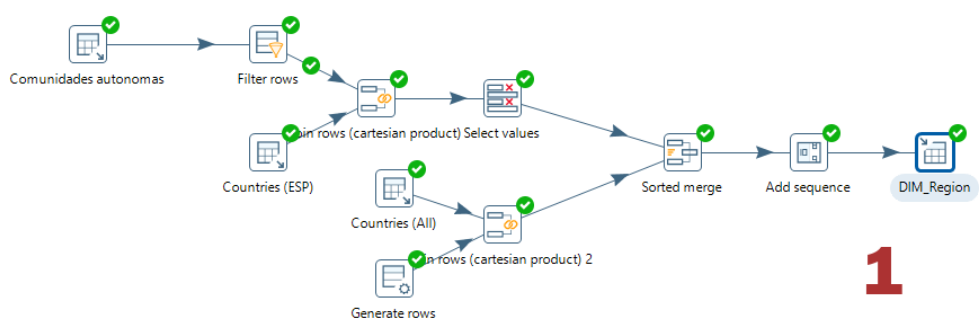
Para el resto de países, se les añade solo la región “NA”, dado que no se tiene más

información de regiones que la de las Comunidades Autónomas.



Por último se unen ambos flujos y se les añade la clave primaria única numérica; después se introducen los nuevos datos en la base de datos del modelo multidimensional.

A continuación se muestran los resultados de la ejecución:



Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

☒ First rows ☐ Last rows ☐ Off

#	Name	Iso2	Iso3	Region	pk
1	Andorra	AD	AND	NA	1
2	United Arab Emirates	AE	ARE	NA	2
3	Afghanistan	AF	AFG	NA	3
4	Antigua and Barbuda	AG	ATG	NA	4
5	Anguilla	AI	AIA	NA	5

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Countries (ESP)	0	0	1	1	0	0	0	0	Finished	0.0s	50	-
2	Comunidades autonomas	0	0	18	18	0	0	0	0	Finished	0.0s	857	-
3	Generate rows	0	0	1	0	0	0	0	0	Finished	0.0s	67	-
4	Filter rows	0	18	17	0	0	0	0	0	Finished	0.1s	340	-
5	Countries (All)	0	0	246	246	0	0	0	0	Finished	0.0s	17,571	-
6	Join rows (cartesian product) 2	0	247	246	0	0	0	0	0	Finished	0.4s	668	-
7	Join rows (cartesian product)	0	18	17	0	0	0	0	0	Finished	0.4s	49	-
8	Select values	0	17	17	0	0	0	0	0	Finished	0.4s	44	-
9	Sorted merge	0	263	263	0	0	0	0	0	Finished	0.7s	374	-
10	Add sequence	0	263	263	0	0	0	0	0	Finished	0.7s	371	-
11	DIM_Region	0	263	263	0	263	0	0	0	Finished	0.7s	355	-

Se han insertado 263 registros en la tabla DIM_Region.

TR_DIM_TYPEEQUIPMENTINSTALLATION



Figura 17: TR_DIM_TYPEEQUIPMENTINSTALLATION

La transformación de la tabla intermedia a la tabla del modelo multidimensional es trivial; solo hay que recuperar los registros del *staging area*, añadir la clave primaria única numérica como en transformaciones anteriores e introducirlos en la tabla de la dimensión.

1

Table input (STG_Investments) Add sequence Table output (DIM_TypeEquipmentInstallation)

2

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

First rows Last rows Off

#	Equipoinstalacion	pk
1	INVERSIÓN EN EQUIPOS E INSTALACIONES INDEPENDIENTES	1
2	INVERSIÓN EN EQUIPOS E INSTALACIONES INTEGRADOS	2
3	NA	3

3

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Table input (STG_Investments)	0	0	3	3	0	0	0	0	Finished	0.0s	167	-
2	Add sequence	0	3	3	0	0	0	0	0	Finished	0.0s	130	-
3	Table output (DIM_TypeEquipmentInstallation)	0	3	3	0	3	0	0	0	Finished	0.1s	54	-

Se han insertado 3 registros en la tabla DIM_TypeEquipmentInstallation.

4.3.5. Bloque TR_FACT

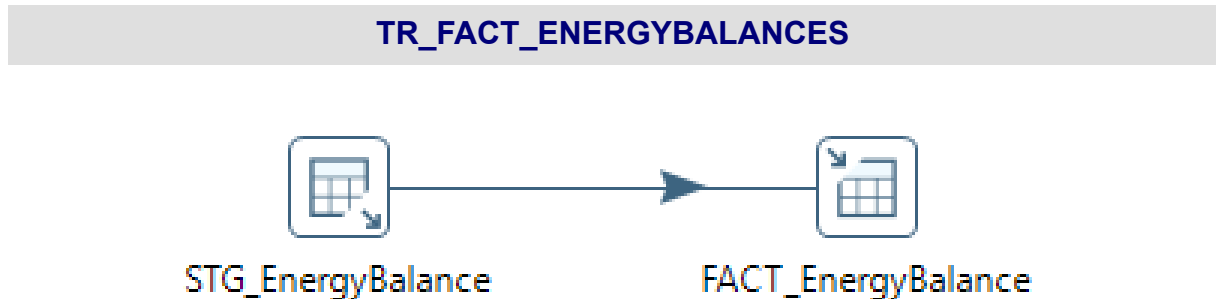


Figura 18: TR_FACT_ENERGYBALANCES

La transformación de la tabla intermedia a la tabla del modelo multidimensional corresponde con una consulta SQL que mezcla tablas del *staging area* con las tablas de dimensiones ya creadas del modelo multidimensional.

Table input
— □ ×

Step name STG_EnergyBalance

Edit...
New...
Wizard...

Connection cn_stage

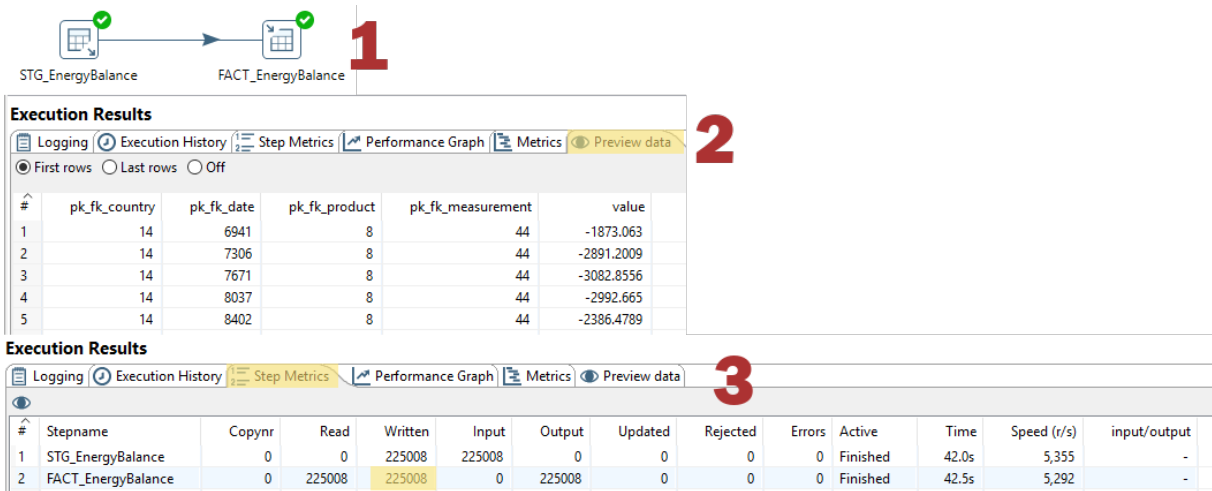
Get SQL select statement...

SQL

```

SELECT
  , dbo.DIM_Country.pk_country as pk_fk_country
  , dbo.DIM_Date.pk_date as pk_fk_date
  , dbo.DIM_Product.pk_product as pk_fk_product
  , dbo.DIM_Measurement.pk_measurement as pk_fk_measurement
  , "value"
FROM dbo.STG_EnergyBalance
INNER JOIN dbo.DIM_Country on dbo.STG_EnergyBalance.country=dbo.DIM_Country.country_name_en
LEFT JOIN dbo.DIM_Date on dbo.STG_EnergyBalance."year"=dbo.DIM_Date.date_year
LEFT JOIN dbo.DIM_Product on dbo.STG_EnergyBalance.product = dbo.DIM_Product.product_name
LEFT JOIN dbo.DIM_Measurement on dbo.STG_EnergyBalance.flow LIKE '%'+ dbo.DIM_Measurement.measurement_code + '%'
where dbo.DIM_Date.date_day=1 and dbo.DIM_Date.date_month=1 and DIM_Country.pk_country IS NOT NULL
and DIM_Date.pk_date IS NOT NULL and DIM_Product.pk_product IS NOT NULL and DIM_Measurement.pk_measurement IS NOT NULL
order by dbo.STG_EnergyBalance.country
        
```

Los resultados de ejecutar la transformación:



Se han insertado 225.008 registros en la tabla FACT_EnergyBalances.

TR_FACT_ENVIRONMENTALMEASUREMENTS

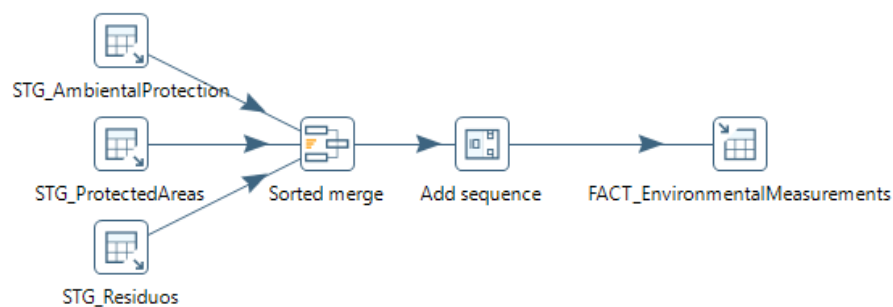


Figura 19: TR_FACT_ENVIRONMENTALMEASUREMENTS

La transformación de la tabla intermedia a la tabla del modelo multidimensional corresponde con varias consultas SQL que mezcla tablas del *staging area* con las tablas de dimensiones ya creadas del modelo multidimensional.

Table input

Step name:

Connection:

SQL

```

SELECT
    DIM_Date.pk_date as fk_date
    , DIM_Region.pk_region as fk_region
    , DIM_EconomicActivitySector.pk_activitysector as fk_activitysector
    , DIM_TypeEquipmentInstallation.pk_typeequipinstall as fk_typeequipinstall
    , DIM_Measurement.pk_measurement as fk_measurement
    , Inversion as "value"
FROM dbo.STG_AmbientalProtection
left join dbo.DIM_Date on Periodo=DIM_Date.date_year and DIM_Date.date_day=1 and DIM_Date.date_month=1
left join dbo.DIM_EconomicActivitySector on dbo.DIM_EconomicActivitySector.activitysector_name=Sector
left join dbo.DIM_TypeEquipmentInstallation on dbo.DIM_TypeEquipmentInstallation.typeequipinstall_name=EquipoInstalacion
left join dbo.DIM_Measurement on dbo.DIM_Measurement.measurement_code=Ambito
left join dbo.DIM_Region on dbo.DIM_Region.region=ComunidadAutonoma
where DIM_Date.pk_date IS NOT NULL and DIM_EconomicActivitySector.pk_activitysector IS NOT NULL and
DIM_TypeEquipmentInstallation.pk_typeequipinstall IS NOT NULL and
DIM_Measurement.pk_measurement IS NOT NULL and DIM_Region.pk_region IS NOT NULL
order by fk_date, fk_region, fk_measurement, fk_activitysector, fk_typeequipinstall

```

Table input

Step name:

Connection:

SQL

```

SELECT
    DIM_Date.pk_date as fk_date
    , DIM_Region.pk_region as fk_region
    , DIM_EconomicActivitySector.pk_activitysector as fk_activitysector
    , DIM_TypeEquipmentInstallation.pk_typeequipinstall as fk_typeequipinstall
    , DIM_Measurement.pk_measurement as fk_measurement
    , "Value" as value
FROM dbo.STG_ProtectedAreas
left join dbo.DIM_Date on "Year"=DIM_Date.date_year and DIM_Date.date_day=1 and DIM_Date.date_month=1
left join dbo.DIM_Measurement on dbo.DIM_Measurement.measurement_code=dbo.STG_ProtectedAreas.Unit
left join dbo.DIM_Region on dbo.DIM_Region.country_code2=dbo.STG_ProtectedAreas.Geo
left join dbo.DIM_EconomicActivitySector on dbo.DIM_EconomicActivitySector.activitysector_name='NA'
left join dbo.DIM_TypeEquipmentInstallation on dbo.DIM_TypeEquipmentInstallation.typeequipinstall_name='NA'
where DIM_Date.pk_date IS NOT NULL and DIM_EconomicActivitySector.pk_activitysector IS NOT NULL and
DIM_TypeEquipmentInstallation.pk_typeequipinstall IS NOT NULL and
DIM_Measurement.pk_measurement IS NOT NULL and DIM_Region.pk_region IS NOT NULL
order by fk_date, fk_region, fk_measurement, fk_activitysector, fk_typeequipinstall

```

Table input

Step name:

Connection:

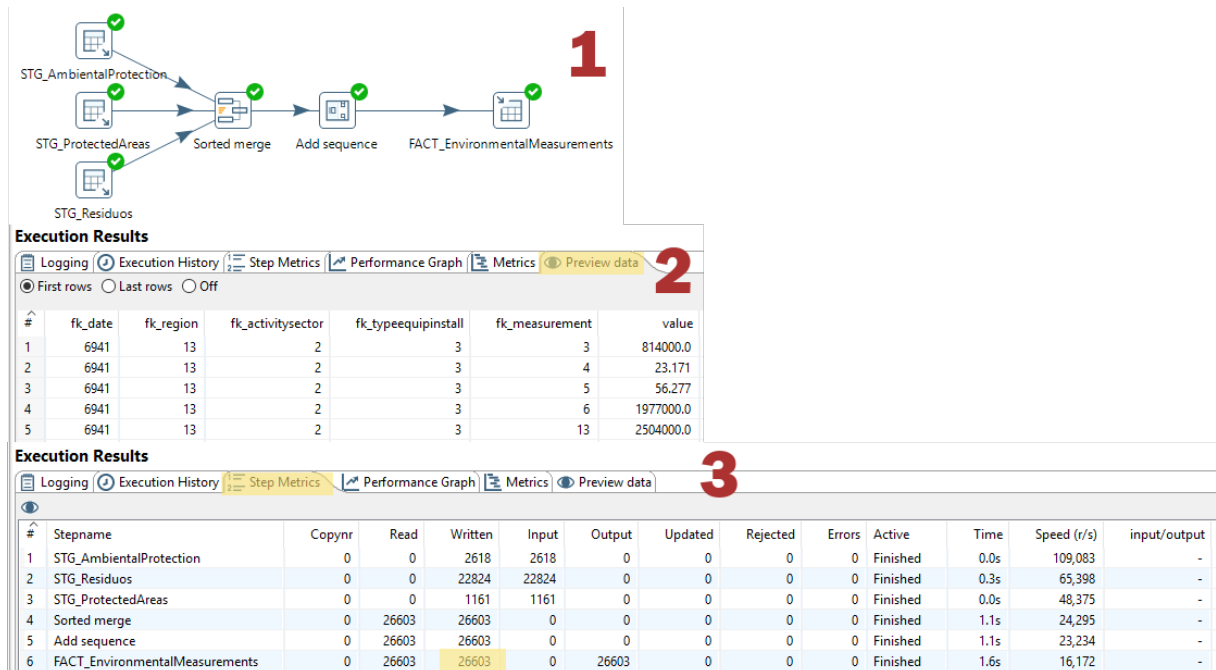
SQL

```

SELECT
    DIM_Date.pk_date as fk_date
    , DIM_Region.pk_region as fk_region
    , DIM_EconomicActivitySector.pk_activitysector as fk_activitysector
    , DIM_TypeEquipmentInstallation.pk_typeequipinstall as fk_typeequipinstall
    , DIM_Measurement.pk_measurement as fk_measurement
    , Obs*POWER(10, Powercode) as value
FROM dbo.STG_Residuos
left join dbo.DIM_Date on "Year"=DIM_Date.date_year and DIM_Date.date_day=1 and DIM_Date.date_month=1
left join dbo.DIM_Measurement on dbo.DIM_Measurement.measurement_code=dbo.STG_Residuos."Var"
left join dbo.DIM_Region on dbo.DIM_Region.country_code3=dbo.STG_Residuos.Cou
left join dbo.DIM_EconomicActivitySector on dbo.DIM_EconomicActivitySector.activitysector_name='NA'
left join dbo.DIM_TypeEquipmentInstallation on dbo.DIM_TypeEquipmentInstallation.typeequipinstall_name='NA'
where DIM_Date.pk_date IS NOT NULL and DIM_EconomicActivitySector.pk_activitysector IS NOT NULL and
DIM_TypeEquipmentInstallation.pk_typeequipinstall IS NOT NULL and
DIM_Measurement.pk_measurement IS NOT NULL and DIM_Region.pk_region IS NOT NULL
order by fk_date, fk_region, fk_measurement, fk_activitysector, fk_typeequipinstall

```

Los resultados de ejecutar la transformación:



Se han insertado 26.603 registros en la tabla FACT_EnvironmentalMeasurements.

4.4. Volumetría

4.4.1. Staging area

Nombre	Registros cargados
IN_AMBIENTALPROTECTION	2.772
IN_COUNTRIES	246
IN_ENVBIO1	1.080
IN_OBJECTIVES	17
IN_OBJECTIVESAREAS	50
IN_RESIDUOS	20.779
IN_WORLDENERGYBALANCES	296.352
Total	321.296

4.4.2. Modelo multidimensional

Nombre	Registros cargados
TR_DIM_COUNTRY	17.897
TR_DIM_ECONOMICTIVITYSECTOR	2
TR_DIM_MEASUREMENT	50
TR_DIM_PRODUCT	11
TR_DIM_REGION	263
TR_DIM_TYPEEQUIPMENTINSTALLATION	3
Total	18.226

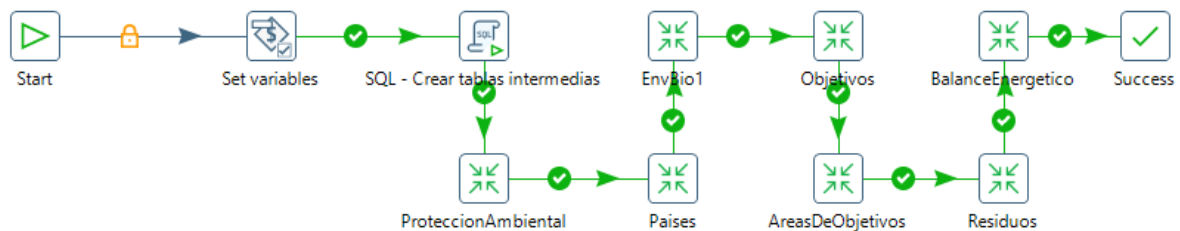
Nombre	Registros cargados
TR_FACT_ENERGYBALANCES	225.008
TR_FACT_ENVIRONMENTALMEASUREMENTS	26.603
Total	251.611

5. Implementación de los *jobs* con ETL

Los trabajos que ejecutarán las transformaciones siguen la misma estructura que las propias transformaciones.

- Se utilizará un *job* para cargar las fuentes de datos a la *staging area* (*job IN*).
- Se utilizará un *job* para cargar las tablas correspondientes a las dimensiones desde la *staging area* (*job TR_DIM*).
- Se utilizará un *job* para cargar las tablas de los hechos (*job TR_FACT*).
- Se utilizará un *job* para cargarlo todo en un único paso (*job DW*).

5.1. *Job IN*

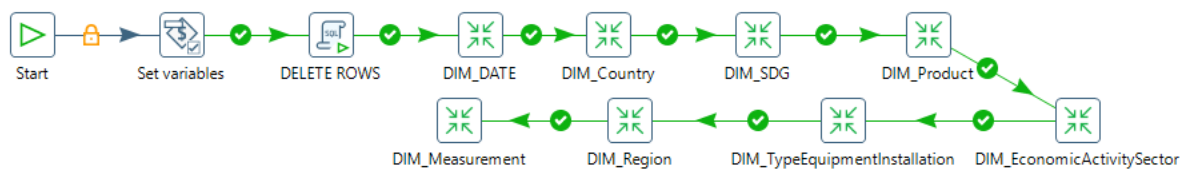


El *job* configura el entorno del PDI, recrea las tablas intermedias del *staging area* y ejecuta secuencialmente las transformaciones «IN_». El resultado de la ejecución es:

Execution Results

Job / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job: JOB_IN						
Job: JOB_IN	Start of job execution		start			2021/12/11 12:36:36
Start	Start of job execution		start			2021/12/11 12:36:36
Start	Job execution finished	Success			0	2021/12/11 12:36:36
Set variables	Start of job execution		Followed unconditional link			2021/12/11 12:36:36
Set variables	Job execution finished	Success			0	2021/12/11 12:36:36
SQL - Crear tablas intermedias	Start of job execution		Followed link after success			2021/12/11 12:36:36
SQL - Crear tablas intermedias	Job execution finished	Success			0	2021/12/11 12:36:36
ProteccionAmbiental	Start of job execution		Followed link after success			2021/12/11 12:36:36
ProteccionAmbiental	Job execution finished	Success			3	2021/12/11 12:36:38
Países	Start of job execution		Followed link after success			2021/12/11 12:36:38
Países	Job execution finished	Success			4	2021/12/11 12:36:38
EnvBio1	Start of job execution		Followed link after success			2021/12/11 12:36:38
EnvBio1	Job execution finished	Success			5	2021/12/11 12:36:38
Objetivos	Start of job execution		Followed link after success			2021/12/11 12:36:38
Objetivos	Job execution finished	Success			6	2021/12/11 12:36:39
AreasDeObjetivos	Start of job execution		Followed link after success			2021/12/11 12:36:39
AreasDeObjetivos	Job execution finished	Success			7	2021/12/11 12:36:39
Residuos	Start of job execution		Followed link after success			2021/12/11 12:36:39
Residuos	Job execution finished	Success			8	2021/12/11 12:36:44
BalanceEnergetico	Start of job execution		Followed link after success			2021/12/11 12:36:44
BalanceEnergetico	Job execution finished	Success			9	2021/12/11 12:37:09
Success	Start of job execution		Followed link after success			2021/12/11 12:37:09
Success	Job execution finished	Success			9	2021/12/11 12:37:09
Job: JOB_IN	Job execution finished	Success	finished		9	2021/12/11 12:37:09

5.2. Job TR_DIM

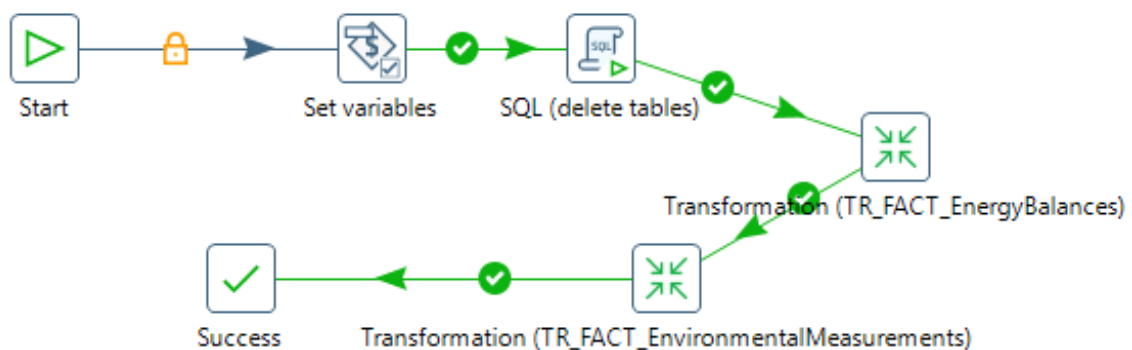


El *job* configura el entorno del PDI, elimina los registros de las tablas de dimensiones del modelo multidimensional y ejecuta secuencialmente las transformaciones «TR_DIM_». El resultado de la ejecución es:

Execution Results

Job / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job: JOB_TR_DIM						
Start	Start of job execution		start			2021/12/11 12:43:29
Start	Start of job execution		start			2021/12/11 12:43:29
Set variables	Job execution finished	Success			0	2021/12/11 12:43:29
Set variables	Start of job execution		Followed unconditional link			2021/12/11 12:43:29
DELETE ROWS	Job execution finished	Success			0	2021/12/11 12:43:29
DELETE ROWS	Start of job execution		Followed link after success			2021/12/11 12:43:29
DIM_DATE	Job execution finished	Success			3	2021/12/11 12:43:30
DIM_Date	Start of job execution		Followed link after success			2021/12/11 12:43:30
DIM_Country	Job execution finished	Success			4	2021/12/11 12:43:30
DIM_Country	Start of job execution		Followed link after success			2021/12/11 12:43:30
DIM_SDG	Job execution finished	Success			5	2021/12/11 12:43:31
DIM_SDG	Start of job execution		Followed link after success			2021/12/11 12:43:31
DIM_Product	Job execution finished	Success			6	2021/12/11 12:43:31
DIM_Product	Start of job execution		Followed link after success			2021/12/11 12:43:31
DIM_EconomicActivitySector	Job execution finished	Success			7	2021/12/11 12:43:31
DIM_EconomicActivitySector	Start of job execution		Followed link after success			2021/12/11 12:43:31
DIM_TypeEquipmentInstallati	Job execution finished	Success			8	2021/12/11 12:43:31
DIM_TypeEquipmentInstallati	Start of job execution		Followed link after success			2021/12/11 12:43:31
DIM_Region	Job execution finished	Success			9	2021/12/11 12:43:32
DIM_Region	Start of job execution		Followed link after success			2021/12/11 12:43:32
DIM_Measurement	Job execution finished	Success			10	2021/12/11 12:43:33
DIM_Measurement	Start of job execution		Followed link after success			2021/12/11 12:43:33
Job: JOB_TR_DIM	Job execution finished	Success	finished		10	2021/12/11 12:43:33

5.3. Job TR_FACT



El *job* configura el entorno del PDI, elimina los registros de las tablas de hechos del modelo multidimensional y ejecuta secuencialmente las transformaciones «TR_FACT_». El resultado de la ejecución es:

Execution Results

Job / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job: JOB_TR_FACT	Start of job execution		start			2021/12/11 12:45:25
Start	Start of job execution		start			2021/12/11 12:45:25
Start	Job execution finished	Success			0	2021/12/11 12:45:25
Set variables	Start of job execution		Followed unconditional link			2021/12/11 12:45:25
Set variables	Job execution finished	Success			0	2021/12/11 12:45:25
SQL (delete tables)	Start of job execution		Followed link after success			2021/12/11 12:45:25
SQL (delete tables)	Job execution finished	Success			0	2021/12/11 12:45:25
Transformation (TR_FACT_Env)	Start of job execution		Followed link after success			2021/12/11 12:45:25
Transformation (TR_FACT_Env)	Job execution finished	Success			3	2021/12/11 12:46:02
Transformation (TR_FACT_Env)	Start of job execution		Followed link after success			2021/12/11 12:46:02
Transformation (TR_FACT_Env)	Job execution finished	Success			4	2021/12/11 12:46:04
Success	Start of job execution		Followed link after success			2021/12/11 12:46:04
Success	Job execution finished	Success			4	2021/12/11 12:46:04
Job: JOB_TR_FACT	Job execution finished	Success	finished		4	2021/12/11 12:46:04

5.4. Job DW



El *job* configura el entorno del PDI, crea las tablas del modelo multidimensional y ejecuta secuencialmente los anteriores trabajos. Por último, elimina las tablas del *staging area* cuando ya no son necesarias. El resultado de la ejecución es:

Execution Results

Job / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
Job: JOB_DW	Start of job execution		start			2021/12/11 12:47:33
Start	Start of job execution		start			2021/12/11 12:47:33
Start	Job execution finished	Success			0	2021/12/11 12:47:33
Set variables	Start of job execution		Followed unconditional link			2021/12/11 12:47:33
Set variables	Job execution finished	Success			0	2021/12/11 12:47:33
CREATE DIM/FACT	Start of job execution		Followed link after success			2021/12/11 12:47:33
CREATE DIM/FACT	Job execution finished	Success			0	2021/12/11 12:47:34
JOB_IN	Start of job execution		Followed link after success			2021/12/11 12:47:34
JOB_IN	Job execution finished	Success			3	2021/12/11 12:48:03
JOB_DIM	Start of job execution		Followed link after success			2021/12/11 12:48:03
JOB_DIM	Job execution finished	Success			4	2021/12/11 12:48:07
JOB_FACT	Start of job execution		Followed link after success			2021/12/11 12:48:07
JOB_FACT	Job execution finished	Success			5	2021/12/11 12:48:41
Delete staging	Start of job execution		Followed link after success			2021/12/11 12:48:41
Delete staging	Job execution finished	Success			5	2021/12/11 12:48:42
Job: JOB_DW	Job execution finished	Success	finished		5	2021/12/11 12:48:42

El tiempo total de la carga inicial del *data warehouse* es de algo más de un minuto.