

Recuperado de:

 $\underline{https://www.btelligent.com/en/blog/staging-area-potential-in-comparison-to-source-system/}$ 

# IMPLEMENTACIÓN DE MODELO DIMENSIONAL EN PENTAHO DATA INTEGRATION (PDI)

Por:

Mg. Esther Balbuena

Copyright 2019 Mg. Esther Balbuena Alarcón. Se otorga permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.3 o cualquier otra versión posterior publicada por la Free Software Foundation. Una copia de la licencia se puede encontrar en el siguiente enlace: http://www.gnu.org/licenses/fdl-1.3.html

Fecha	Versión	Autor	Detalle del cambio
14/04/2019	0.1	Mg. Esther Balbuena	Versión Inicial.

Para ver la traducción no oficial de la Licencia de Documentacion Libre de GNU, seguir el siguiente enlace: <a href="http://stuff.danexnow.org/gfdl">http://stuff.danexnow.org/gfdl</a> es.html

### IMPLEMENTACIÓN DE MODELO DIMENSIONAL

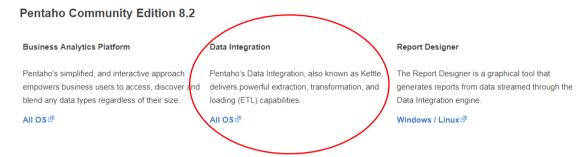
**Objetivo:** Implementar el modelo dimensional de una datawarehouse usando las bases de datos dimensionales y Pentaho Data Integration para los procesos ETL.

### 1) Requisitos previos

a) Tener instalado Pentaho Data Integration versión Community.

Descarga desde:

https://community.hitachivantara.com/docs/DOC-1009931-downloads



Una vez descargado el archivo, procedemos a descomprimirlo en algún lugar apropiado (sugerido: C:\)

b) Tener instalado MySql Server y Workbench
 Descargar desde el siguiente enlace:
 https://dev.mysql.com/downloads/mysql/

Instalar MySql Server y MySql Workbench para la versión de Sistema Operativo que cuente en su máquina.

- c) Configuración:
- Descargar de: <a href="https://dev.mysql.com/downloads/connector/j/">https://dev.mysql.com/downloads/connector/j/</a>
- Descomprimir el archivo zip (in my case it was mysql-connector-java-5.1.31.zip)
- Copiar el archivo .jar (mysql-connector-java-5.1.45.-bin.jar) y pegarlo en el folder Lib del PDI:
- PC: C:\pdi-ce-8.2.0.0-342\data-integration\lib , o la dirección donde se descomprimió PDI
- Mac: /Applications/data-integration/lib
- d) Reconocer la estructura de datos de la fuente de datos fichero\_book.csv

"user\_code"|"user"|user\_id|"producto"|nota|texto|resaltar|subrayar|tachar|manoalzada|rectangulo|elipse|"nive"
"000007b7-a199-47a4-8288-441880f97a11"|"Usuario Usuario 55000"|55000|"ESA172025"|1|0|0|0|0|0|0|1|"Secundaria
e historia. 4 ESO. Savia"|"Alumno"|"Castilla-La Mancha"

"0000c53a-899d-4051-a579-e74cfd5e4d1f"|"Usuario Usuario 55001"|55001|"ESA146820"|0|0|0|0|0|0|0|0|"Secundaria

- 2) Creación del staging area para guardar los datos desde la fuente de datos.
  - Se debe tener el servicio de MySql arriba.
  - Abrir MySql Worbench y crear la Base de Datos stg\_apellido (en mi caso stg\_ebalbuena)
  - Crear la tabla f\_salida que almacenará los datos que vienen de la fuente de datos fichero book.csv

```
CREATE DATABASE IF NOT EXISTS 'stg_ebalbuena' /*!40100 DEFAULT CHARACTER SET utf8 */;
USE `stg_ebalbuena`;
-- Table structure for table `f_salida`
DROP TABLE IF EXISTS `f_salida`;
/*!40101 SET @saved cs client = @@character set client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `f_salida` (
 'user code' varchar(100) NOT NULL,
 'user' varchar(100) NOT NULL,
 `user_id` int(11) DEFAULT NULL,
 'producto' varchar(100) DEFAULT NULL,
 `nota` int(11) DEFAULT NULL,
 `texto` int(11) DEFAULT NULL,
 'resaltar' int(11) DEFAULT NULL,
 `subrayar` int(11) DEFAULT NULL,
 `tachar` int(11) DEFAULT NULL,
 'manoalzada' int(11) DEFAULT NULL,
 'rectangulo' int(11) DEFAULT NULL,
 'elipse' int(11) DEFAULT NULL,
 'nivel' varchar(100) DEFAULT NULL,
 `curso` varchar(100) DEFAULT NULL,
 `asignatura` varchar(100) DEFAULT NULL,
 'titulo' varchar(250) DEFAULT NULL,
'rol' varchar(100) DEFAULT NULL,
 'zona' varchar(100) DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='tabla f salida';
/*!40101 SET character set client = @saved cs client */;
/*!40103 SET TIME_ZONE=@OLD_TIME_ZONE */;
```

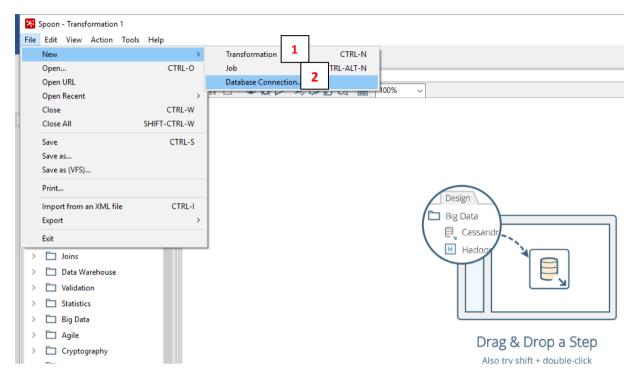
### 3) Creación de una conexión de Pentaho Data Integration y MySql.

a) Iniciar Pentaho Data Integration ejecutando:

Spoon.bat	14/11/2018 17:21	Archivo por lotes	5 KB
spoon.command	14/11/2018 17:21	Archivo COMMA	2 KB
🔀 spoon.ico	14/11/2018 17:21	Icono	362 KB

b) Crear una conexión a base de datos:

- 1. Crear nueva transformación
- 2. Crear nueva conexión a base de datos



### Configurar la conexión:

### Parámetros:

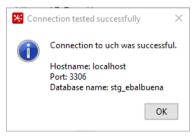
✓ Connection name: uch✓ Connection type: MySQL

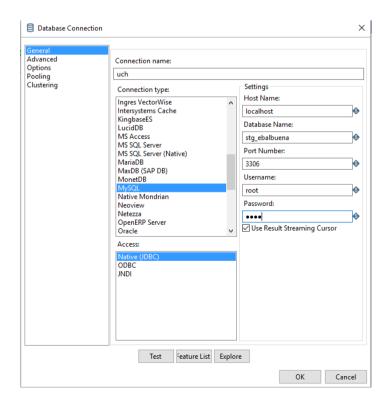
✓ Access: Native (JDBC)✓ Host name: localhost

✓ Database name: stg ebalbuena

✓ Port number: 3306 ✓ User name: root ✓ Password: root

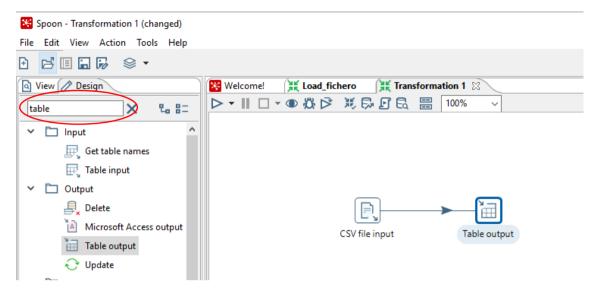
c) Presionar botón test:



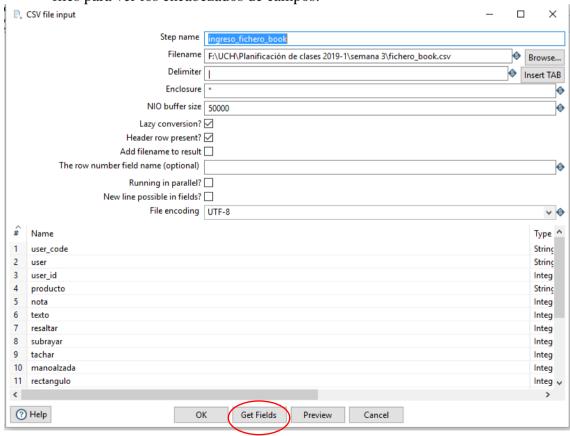


### 4) Cargar los datos desde la fuente de datos al staging área.

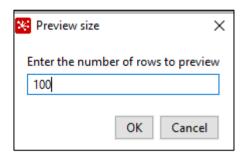
Usaremos dos pasos: CSV file input y Table output. Puede usar el buscador para hallarlos con facilidad

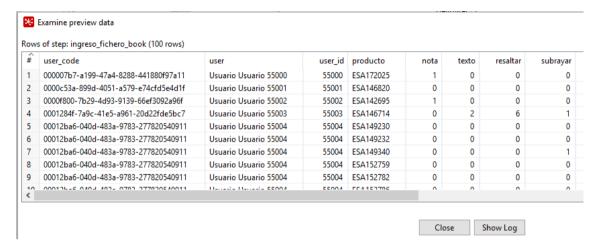


a) Configuramos el paso CSV file input. Le damos doble click al paso CVS file input y aparecerá la siguiente ventana donde completamos la dirección de la fuente de datos, es decir, la dirección de fichero\_book, completamos el delimitador de campos que en este caso es | y con qué carácter están cerrados los campos. Escogemos UTF-8 como codificador de archivo. Le damos click a Get files para ver los encabezados de campos:



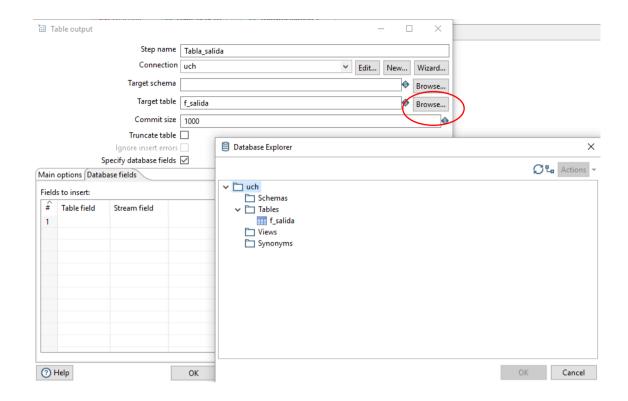
También podemos tener una vista previa de los datos presionando el botón Preview en la pantalla anterior:



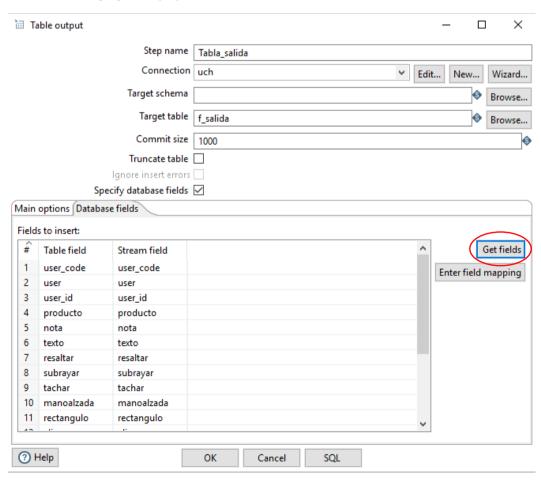


Cerramos está pantalla y le damos Ok a la pantalla emergente principal.

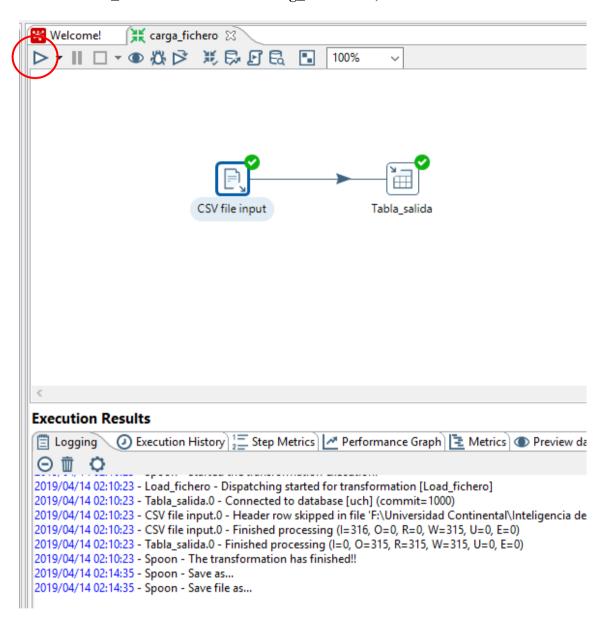
b) Configuramos el paso Table output. Le damos doble click al Table output y aparecerá la siguiente ventana donde elegiremos la conexión uch que se configuró en un paso anterior. El target table lo buscamos en la base datos MySql que es donde está el staging área dando click al botón Browse.



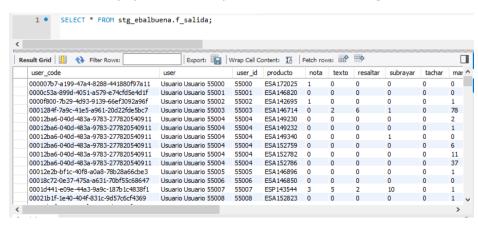
Presionamos el botón get filed para obtener todos los campos de la entrada y que pasarán a la tabla del staging en MySql.



c) Ejecutamos la transformación haciendo click en RUN, nos pedirá que lo grabemos (nombre.ktr), grabe en su carpeta de trabajo. Con lo cual se habrá almacenado los datos de la fuente (fichero\_book.csv) al staging área (tabla f\_salida en base de datos stg\_ebalbuena)



### Verificamos en Mysql Workbench que los datos han sido cargados:



## 5) Implementación del modelamiento dimensional correspondiente al datawarehouse.

Tabla de hecho	fact_fichero
Tablas de dimensiones	dim_asignatura dim_rol dim_nivel dim_usuario dim_producto

Creamos una base de datos usando MySql WorkBench, que corresponde al datawarehouse que almacenará los datos para el análisis tomando el esquema estrella.

```
CREATE DATABASE IF NOT EXISTS `stg_ebalbuena` /*!40100 DEFAULT CHARACTER SET utf8 */; USE `stg_ebalbuena`;
```

Script de creación de tabla dimensión (Usar este script como modelo para crear las demás tablas de dimensiones.

```
-- Table structure for table `dim_asignatura`

DROP TABLE IF EXISTS `dim_asignatura`;

CREATE TABLE `dim_asignatura` (
   `id` int(11) NOT NULL AUTO_INCREMENT,
   `desc` varchar(50) NOT NULL DEFAULT '0',
   PRIMARY KEY (`id`)

) ENGINE=InnoDB AUTO_INCREMENT=604 DEFAULT CHARSET=utf8;
```

### Script de creación de la tabla de hecho: fact\_fichero:

```
CREATE TABLE `fact_fichero` (
   `id` int(11) NOT NULL AUTO_INCREMENT,
   `id_asignatura` int(11) NOT NULL,
   `id_nivel` int(11) NOT NULL,
   `id_rol` int(11) NOT NULL,
   `id_usuario` int(11) NOT NULL,
   `id_producto` int(11) NOT NULL,
   `c_notas` int(11) NOT NULL DEFAULT '0',
   `c_texto_libre` int(11) NOT NULL DEFAULT '0',
   `c_resaltar_texto` int(11) NOT NULL DEFAULT '0',
   `c_subrayar_texto` int(11) NOT NULL DEFAULT '0',
   `c_tachar_texto` int(11) NOT NULL DEFAULT '0',
```

```
`c_mano_alzada` int(11) NOT NULL DEFAULT '0',
    `c_rectangulo` int(11) NOT NULL DEFAULT '0',
    `c_elipse` int(11) NOT NULL DEFAULT '0',
    PRIMARY KEY (`id`),
    UNIQUE KEY `id_unique` (`id_usuario`,`id_producto`),
    KEY `id_asignatura` (`id_asignatura`),
    KEY `id_nivel` (`id_nivel`),
    KEY `id_rol` (`id_rol`),
    KEY `id_producto` (`id_producto`),
    KEY `IDX` (`id_usuario`,`id_producto`)
) ENGINE=InnoDB AUTO_INCREMENT=836001 DEFAULT CHARSET=utf8;
```

#### Verificamos que se haya creado la estructura deseada:

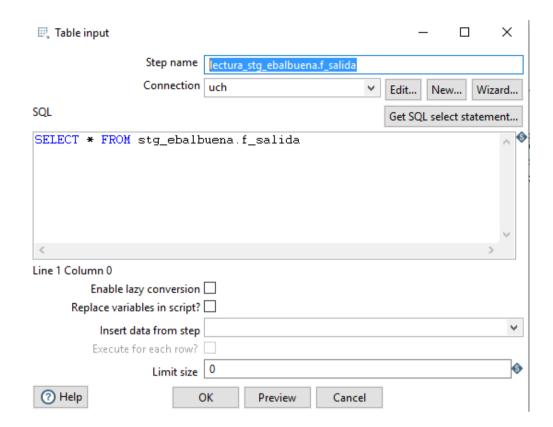


- 6) Carga de datos desde el staging área (stg\_ebalbuena.f\_salida) a nuestro Datawarehouse (olap\_ebalbuena)
- a) Carga de datos a las tablas dimensiones. Abrimos una nueva transformación y realizaremos los procesos ETL para cada dimensión usando los siguientes pasos:

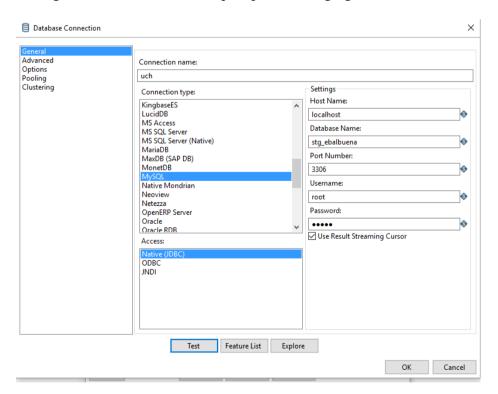


Canfiguramos cada paso, lo cual debe ser realizado para cada dimensión.

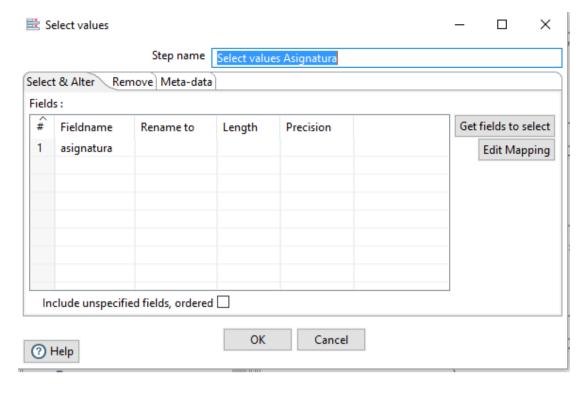
1a) PASO: lectura\_stg\_ebalbuena.f\_salida



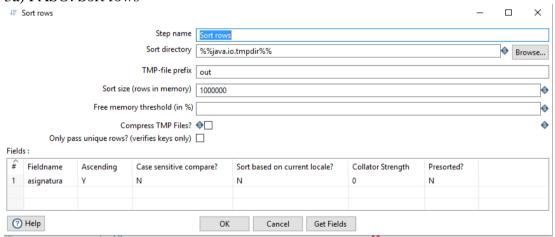
Configuramos la conexión uch que apunta al staging área.



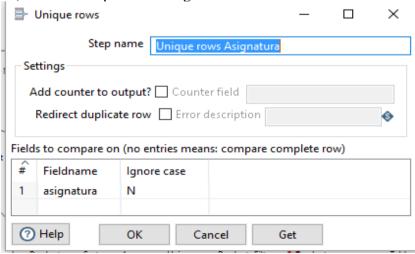
2a) PASO: Select values Asignatura.



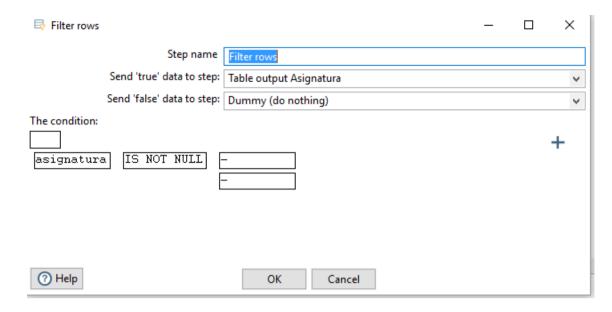
3a) PASO: Sort rows



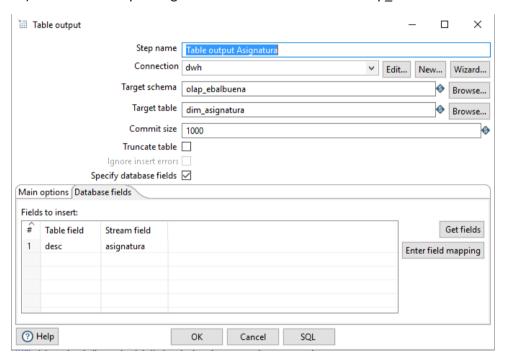
4a) PASO: Unique rows Asignatura

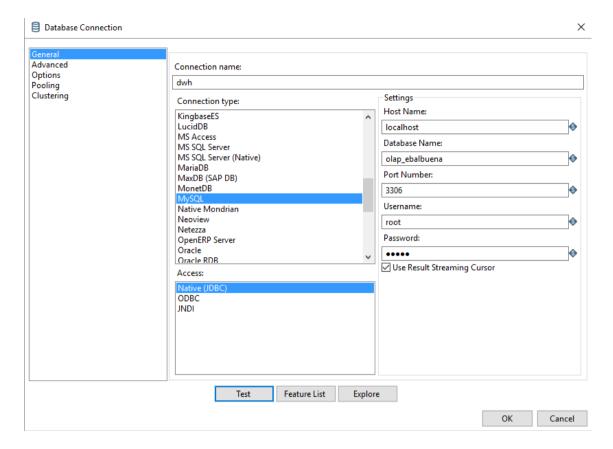


### 5a) Paso: Filter rows

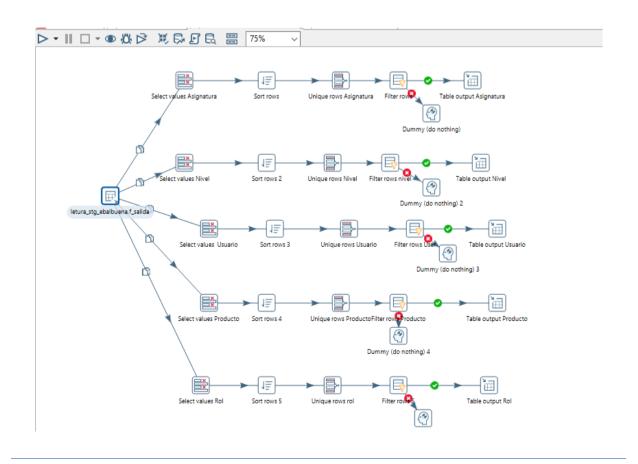


### 6a) PASO: Table output Asignatura con la conexión dwh a olap\_ebalbuena

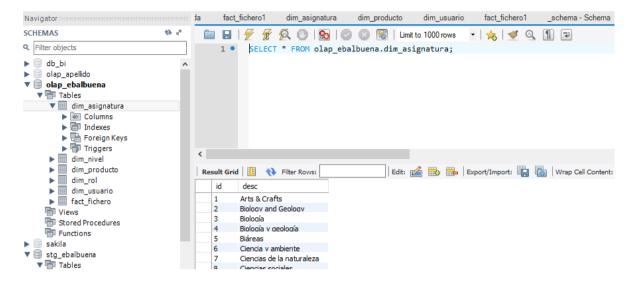




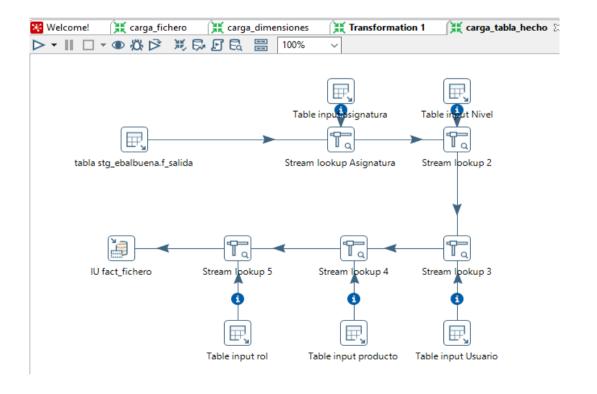
Usar los mismos pasos de 2a a 6a para cada dimensión. El modelo final debe tener todas las dimensiones que se definieron en el modelo:



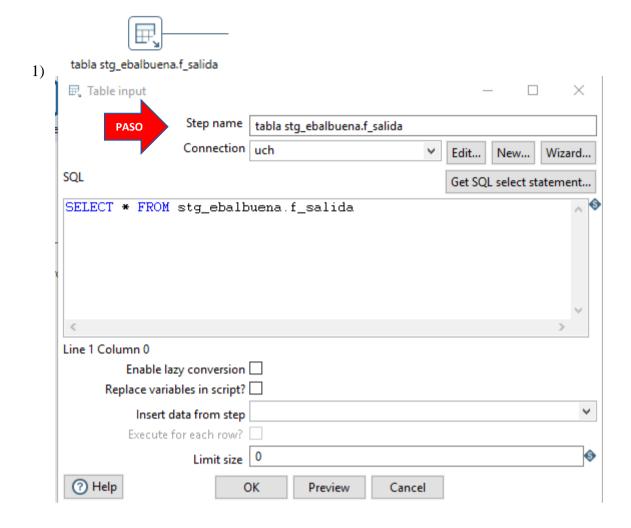
Corremos la actual transformación y luego verificamos que los datos se hayan cargado en nuestra base de datos.



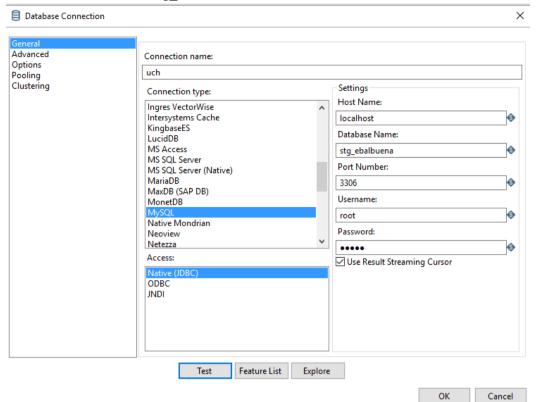
b) Carga de datos a la tabla de hechos: fact\_fichero. Abrimos una nueva transformación y realizaremos los procesos ETL para cargar la tabla de hechos usando los siguientes pasos:



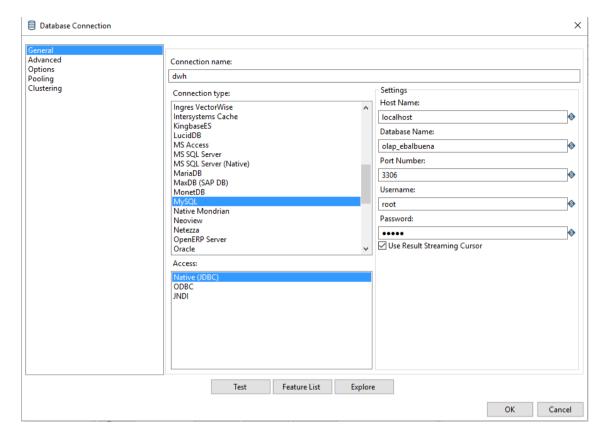
### Configuración de cada paso:



### Usamos la conexión a stg\_ebalbuena



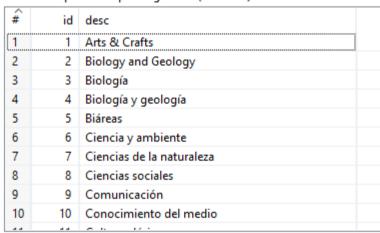
2) PASO: Table input\_asignatura × 🖳 Table input PASO Step name Table input asignatura Connection Edit.. Wizard... SQL Get SQL select statement... SELECT \* FROM olap\_ebalbuena.dim\_asignatura Line 1 Column 0 Enable lazy conversion Replace variables in script? v Insert data from step Execute for each row? Limit size 0 ᢒ ? Help OK Preview Cancel



### Examinamos la data con una vista previa:

Examine preview data

Rows of step: Table input asignatura (100 rows)



3) PASO: Stream lookup Asignatura Ta Stream lookup Х **PASO** Step name Stream lookup Asignatura Lookup step Table input asignatura The key(s) to look up the value(s): Field LookupField asignatura desc Specify the fields to retrieve: Field New name Default Type 1 id id\_asignatura -1 Integer Preserve memory (costs CPU) Key and value are exactly one integer field ✓se sorted list (i.s.o. hashtable) 

✓

asignatura, es el campo que viene de la tabla f\_salida (staging área)

Cancel

desc, es el campo de la tabla de hecho fact\_fichero donde se guardará ese dato.

4) Los pasos 2 y 3 deben configurarse de manera similar para cada dimensión.

Get Fields

5) PASO:

OK

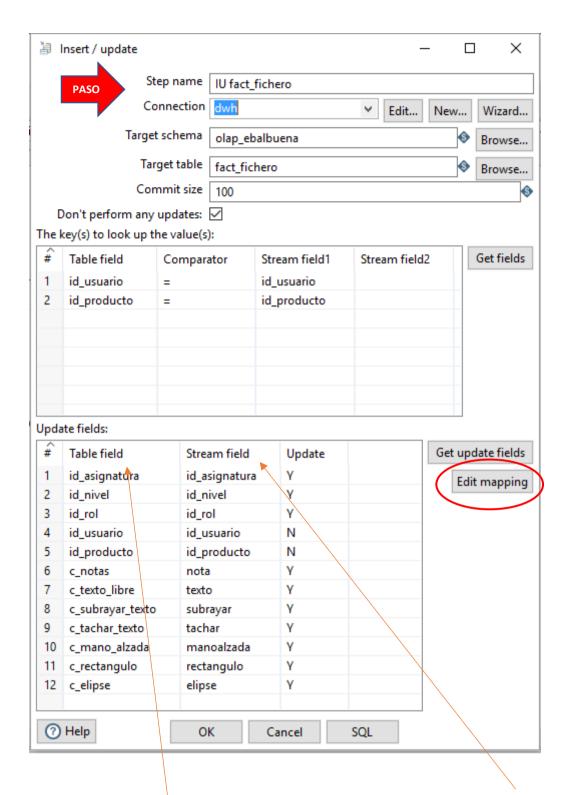
(7) Help



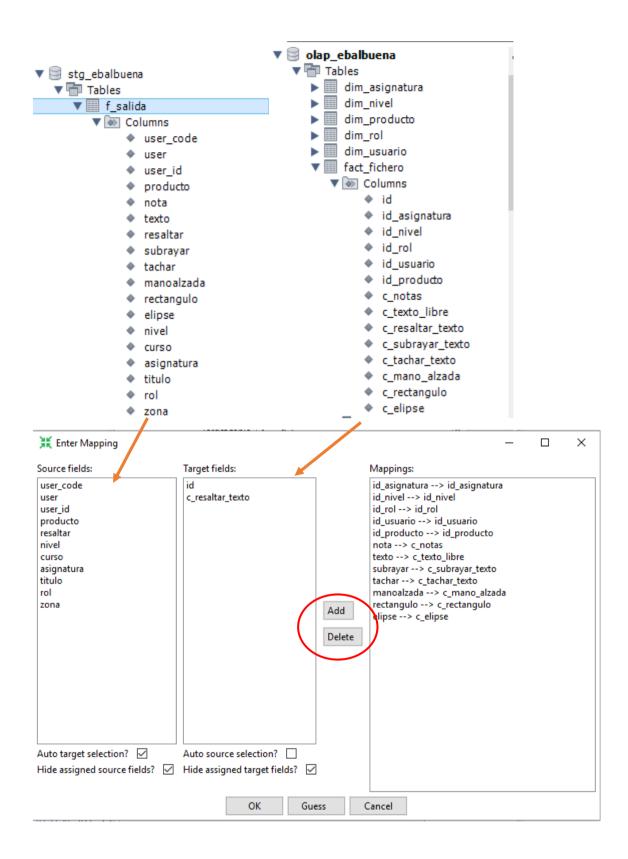
IU fact\_fichero

En este paso se realiza la carga de datos a la tabla de hechos en el Datawarehouse que está dado la tabla olap\_ebalbuena.fact\_fichero.

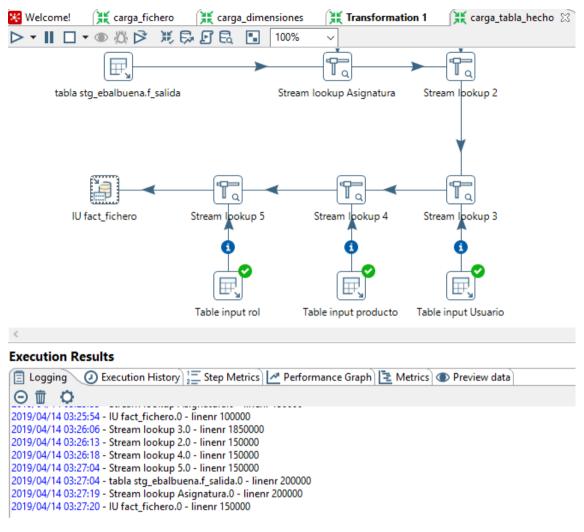
Get lookup fields



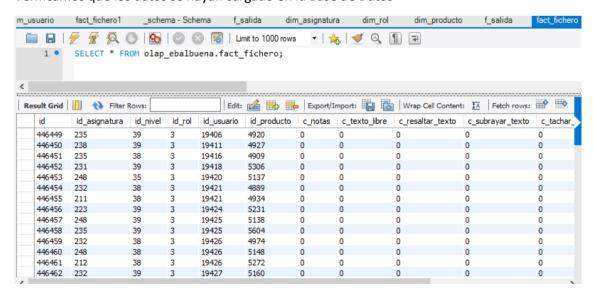
Usamos la opción Edit mapping para relacionar los campos que vienen del **staging** : **f\_salida** con los campos de **la tabla de hechos** olap\_ebalbuena.fact\_fichero.



Corremos la transformación y veremos en la zona de Logging como avanza la ejecución hasta llegar al último paso.



Verificamos que los datos se hayan cargado en la Base de Datos



**¡Muy bien!** Ya tenemos nuestra datawarehouse con los datos cargados. Ya estamos listos para explotar los datos.