Fase 1, Proyecto de Catedra

Autor: BA1819, CG181933, GA181935, PG120321

2020

Universidad Don Bosco

Datawarehouse y Minería de Datos

Proyecto de catedra, fase 1

Integrantes:

Bonilla Avilés, David Alejandro-BA1819

Cruz González, José Roberto - CG181933

Garay Alvarado, Bryan Walberto - GA181935

Portal Gómez, Roberto José - PG120321

Catedrático: Ing. Carlos Filiberto Alfaro Castro

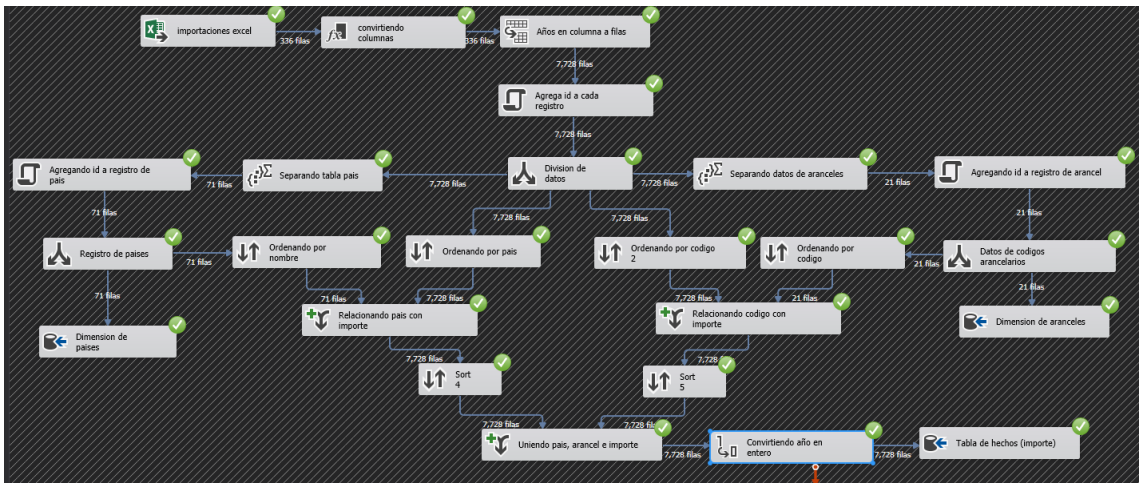
Soyapango, 7 de septiembre del 2020

Contenido

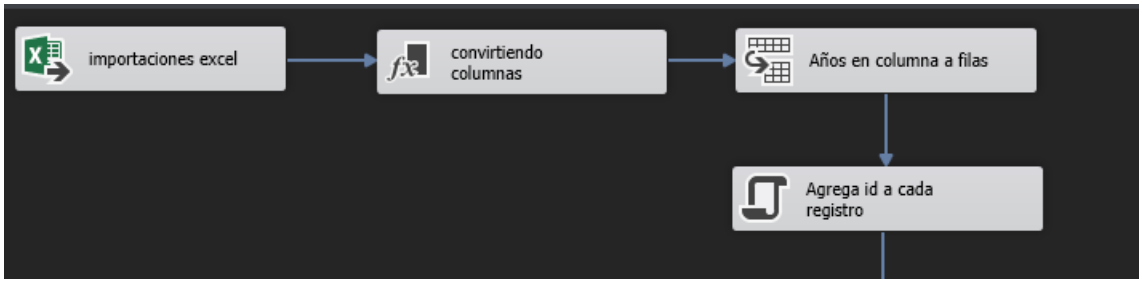
Autor: BA1819, CG181933, GA181935, PG120321

Descripción general

Debido a la complejidad del problema, el ETL resultante del análisis y procesamiento del mismo resulto bastante largo, es por esto que como primera instancia damos una vista general del mismo para proceder a la explicación de cada una de sus partes:

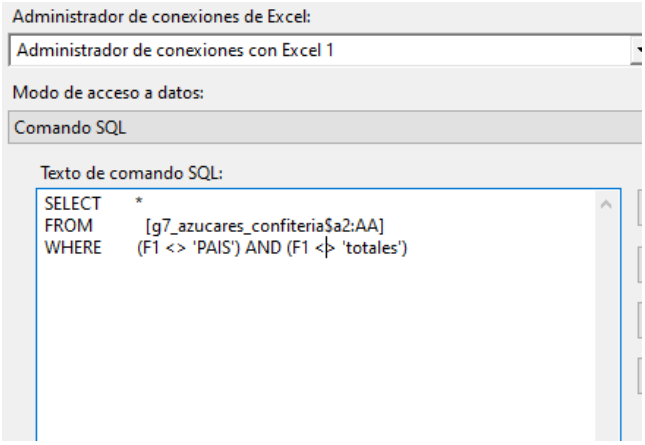


Extracción de datos:

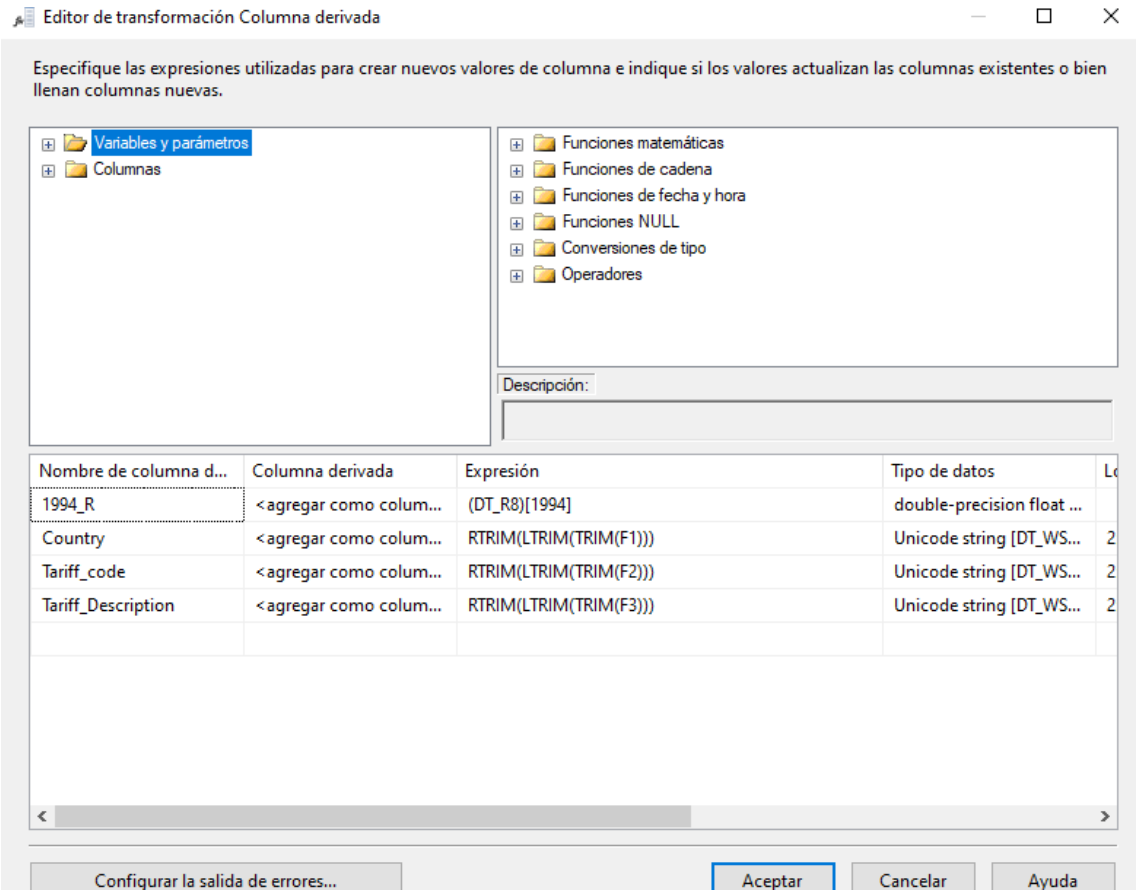


Como primera parte se obtienen los datos del origen, que en este caso en particular es un documento en Excel con los registros solicitados, es importante recalcar que se realizó una consulta SQL dentro del Excel para indicar desde que columna y fila se obtendrían los datos ya que era necesario evadir la primera fila que contiene el título de la hoja, también se evita la última fila que contiene el total por año (Este dato se puede obtener posteriormente con una consulta SQL dentro de la base de datos resultante).

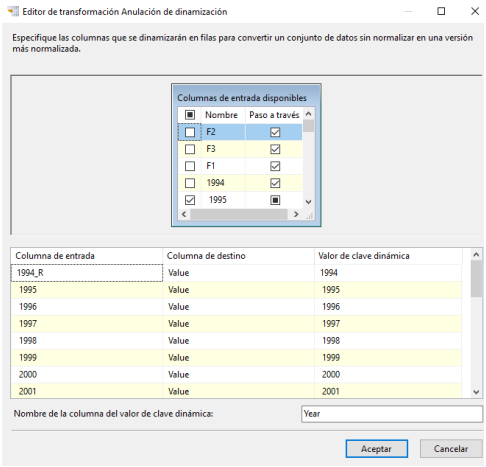
Autor: BA1819, CG181933, GA181935, PG120321



Dentro de la transformación de la columna derivada se procede a sustituir el valor de 1994_R por 1994, ya que es un año se desea procesar como un numero dentro del ETL, así como se realizan otros cambios necesarios.



En el paso siguiente se opta por hacer un pívot de los datos, ya que se tenía un formato de registro en filas por año, lo que dificulta el procesamiento, entonces todas estas se les realiza un pívot para tener más registros relacionados con cada importación

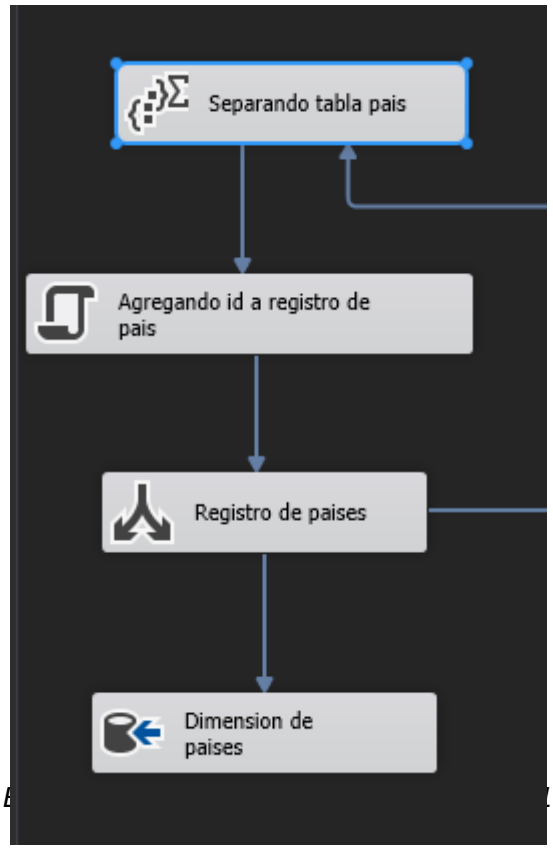


En este punto no se cuenta con un id único para cada registro por este motivo, dentro de un componente de script, se añade una nueva columna con un identificador único para cada registro de importación:

```
2 referencias
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    // Agregando uid para cada importacion
    Row.id = System.Guid.NewGuid();
}
```

Dimensión País

Se debe obtener una tabla solo con los países para posteriormente relacionarlo con las importaciones, de eso se encarga este proceso.



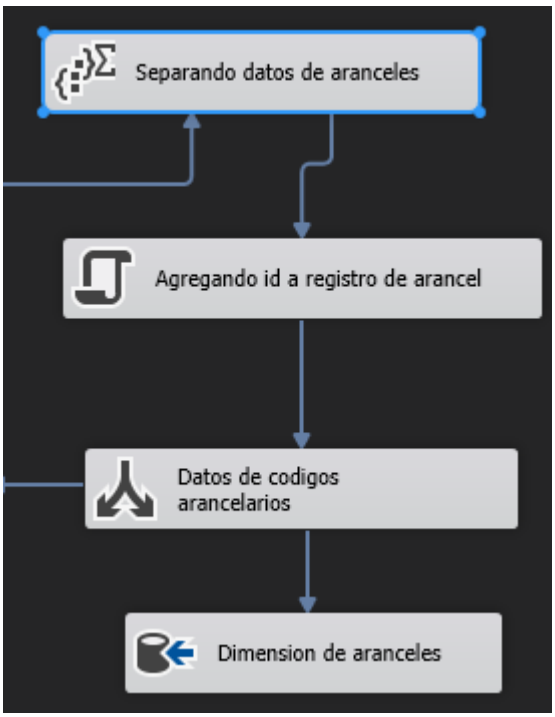
En primer lugar, se obtienen solo los países sin repetición, para posteriormente agregarle un identificador único, dentro del componente de script también se hace una nueva columna con el nombre del país sin espacios en blanco.

```
// Generar un nuevo guid para cada pais
Row.idcountry = System.Guid.NewGuid();
Row.name = Row.Country.Trim();
```

Después se obtienen dichos países, se realiza un multicast, en uno de los extremos se guarda dentro de la base de datos en la tabla

dbo.DimCountry perteneciente a la dimensión del país.

Dimensión Códigos Arancelarios con descripción



Este proceso es bastante parecido al anterior, con la salvedad que en este caso se obtienen los códigos arancelarios junto a una descripción.

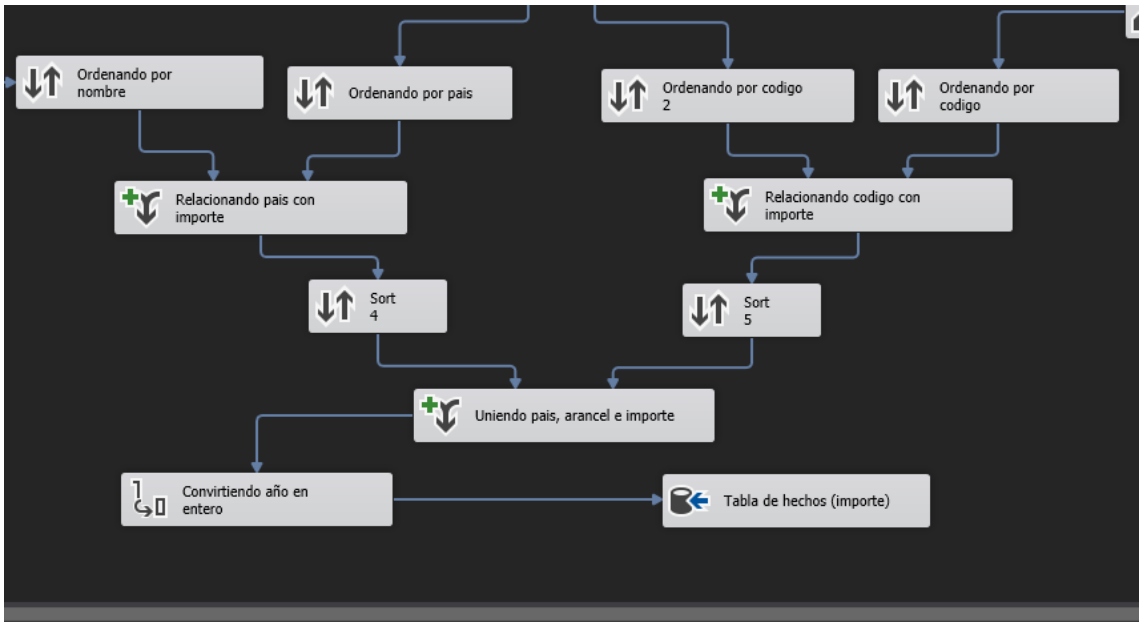
Como primer paso se filtran los códigos para que no estén repetidos, ya que como origen tenemos una serie de importaciones, en las cuales cada una tiene un código arancelario asignado, es decir el código se puede repetir en dos o más importaciones, ya que funciona como una categoría para cada importación.

Se procede a agregar a cada código un identificador único para ser asignado a cada importación posteriormente.

```
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    // Creando Guid para id de codigo arancelario
    Row.idtariff = System.Guid.NewGuid();
    Row.code = Row.Tariffcode.Trim();
}
```

Como último paso la dimensión código arancelario se almacena dentro de la tabla dbo.DimTariff.

Tabla de Hechos



En esta tabla central, donde se recogen las relaciones entre todos los datos anteriormente procesados, se empieza por relacionar el país correspondiente con cada importación. Después se realiza la misma operación pero con el código de la importación (Código arancelario), para posteriormente relacionar todo , convertir el año que hasta este momento es de tipo string a integer y posteriormente insertar todo en la tabla de hechos dbo.Fact_imports.