INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

ACADEMIA DE CIENCIA DE DATOS

**Profesor:** M. en C. Alejandro Botello Castillo

**Práctica no. 8: Aplicación del proceso ETL a un modelo multidimensional (tabla de tiempo y de hechos)**

**Número de Práctica**: 8

**Nombre de la Práctica**: Aplicación del proceso ETL a un modelo multidimensional (tabla de tiempo y de hechos)

REQUISITOS PREVIOS

Para la realización de la práctica se deberán tener los mismos requerimientos de la práctica 7.

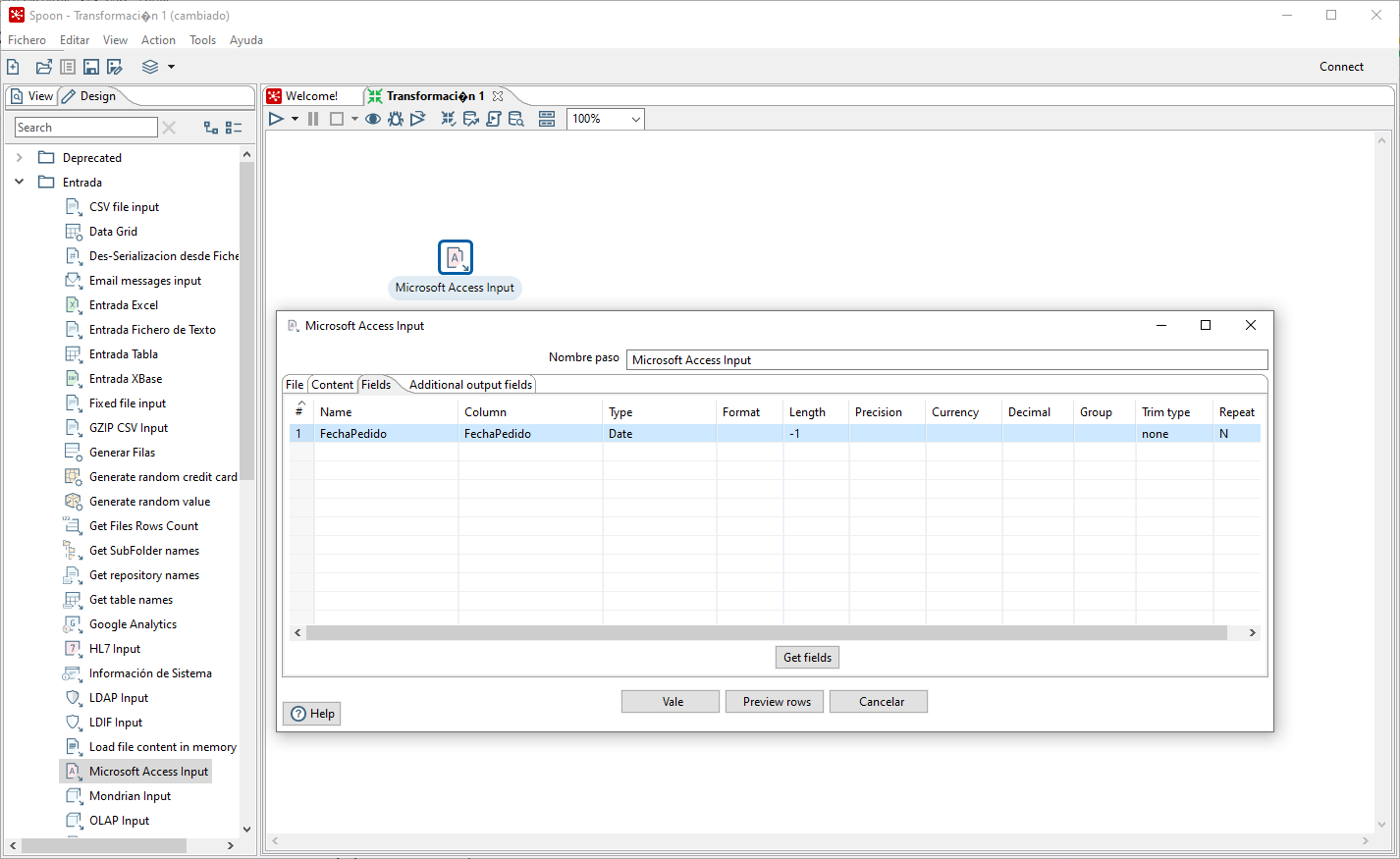
**MARCO TEÓRICO**

De acuerdo con lo visto en clase, se emplearán los conceptos del proceso de Extracción, Transformación y Carga (ETL) aplicado a un modelo multidimensional, mediante una herramienta CASE de generación de flujos de trabajo.

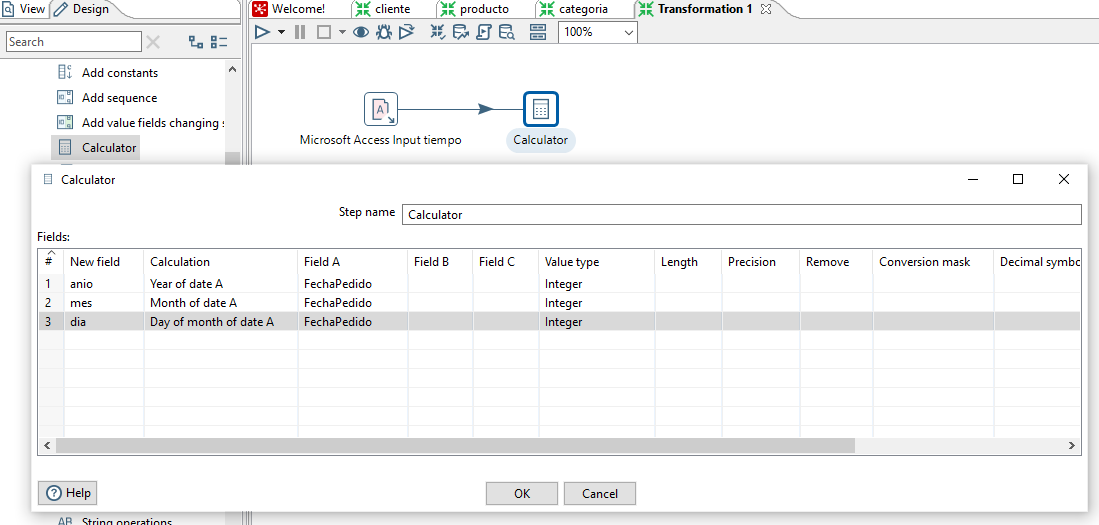
**DESARROLLO DE LA PRÁCTICA (CONTINUACIÓN DE LA PRÁCTICA 7):**

1.- Para llenar la tabla de tiempo de la bodega de datos, es necesario obtener los datos de la fecha de venta (f*echaPedido*) de la tabla *Pedidos* de la fuente operacional, y posteriormente separar el año, el mes y el día de la fecha de pedido para convertirlo en el identificador primario de la tabla *Tiempo* (concatenando el año, el mes y el día, y convirtiéndolo en un entero). Los procesos involucrados se muestran a continuación, así como los valores que se tienen que establecer para configurar el paso

a) Cree un nuevo flujo. Agregue un componente **Access Input**  (como en la práctica 7) y obtenga la tabla *Pedidos*. Únicamente seleccione la columna *fechaPedido*.



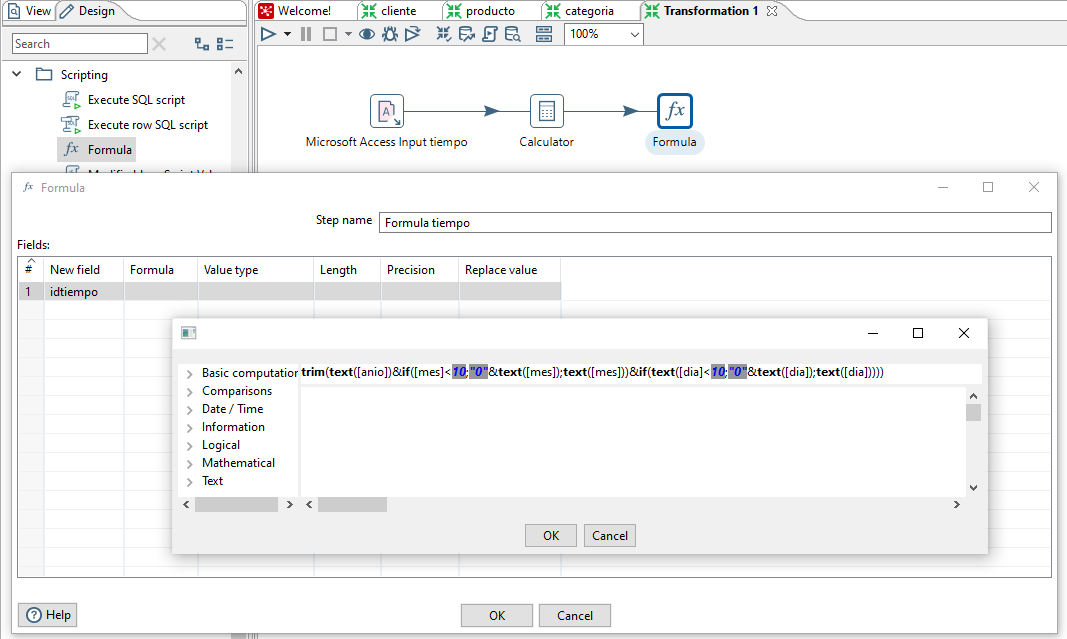
b) Incluya un componente *Calculator* y líguelo con el de **Input** anterior. Cree las columnas *anio*, *mes*, *dia* y aplique el cálculo correspondiente a la extracción del año, del mes y del día, como se observa en la siguiente figura. El atributo de entrada debe ser el mismo, *fechaPedido*, de la fuente operacional. Establezca los valores de tipo entero.



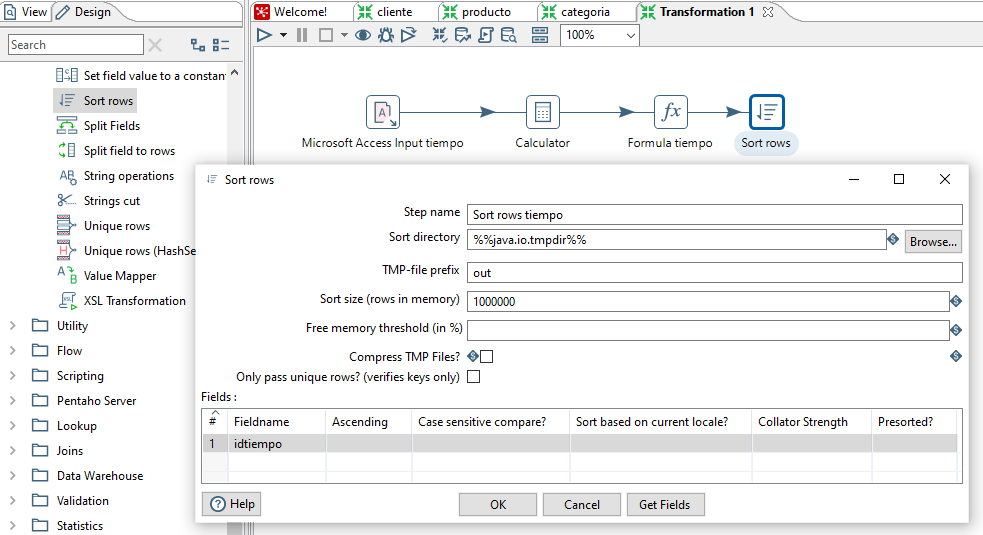
c) Incluya un operador *Formula* en el proceso, ligado al anterior. Cree un nuevo campo llamado *idTiempo* y en la fórmula escriba la siguiente expresión:

text([anio])&if([mes]<10;"0"&text([mes]);text([mes]))&if([dia]<10;"0"&text([dia]);text([dia]))

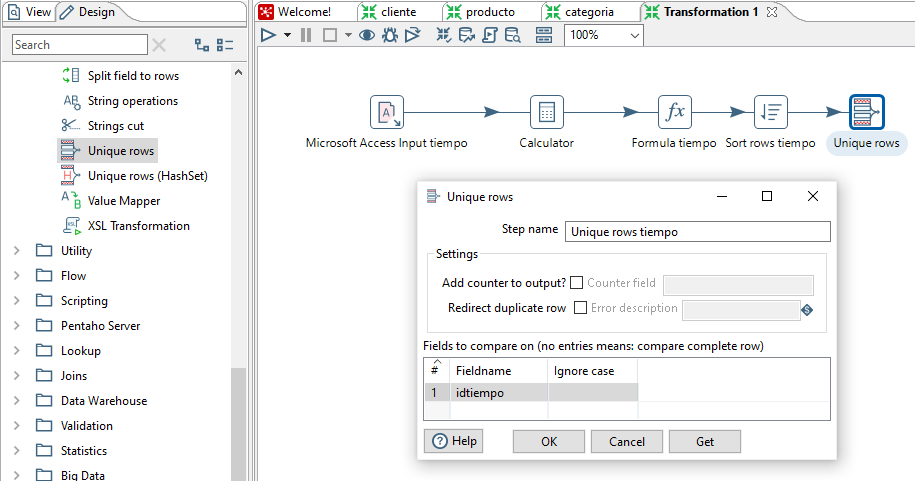
Establezca el tipo de dato a *String* y acepte los cambios.



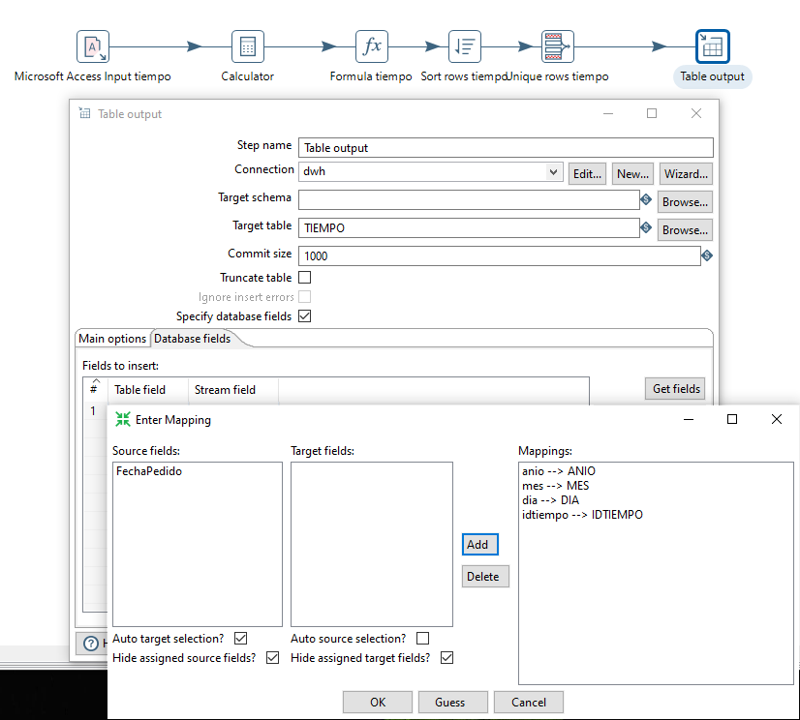
d) Incluya un operador *Sort Rows* y líguelo con el anterior. Seleccione el campo *idTiempo* como criterio de ordenamiento, y configure para que sea ascendente. Guarde los cambios.



e) Incluya un paso *Unique row*s y líguelo al anterior. Seleccione el campo *idTiempo* como criterio de eliminación de duplicados y guarde los cambios.



f) Agregue un paso *Table Output* y establezca una conexión con la tabla *Tiempo* de la base de datos multidimensional (Oracle). Guarde las modificaciones como un proceso (*Tiempo.ktr*).



Incluya una captura de pantalla del proceso *Tiempo* creado, incluyendo el *preview* de los datos generados de forma correcta.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

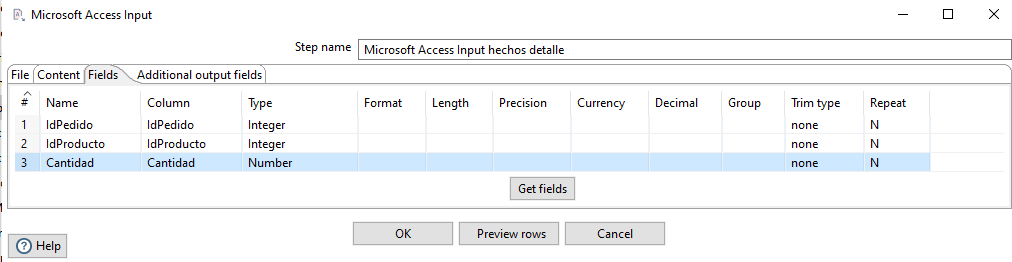
Incluya los resultados de la consulta SELECT \* FROM tiempo, de la base de datos multidimensional en Oracle (solo los últimos 30 renglones).

Imagen que contiene Calendario

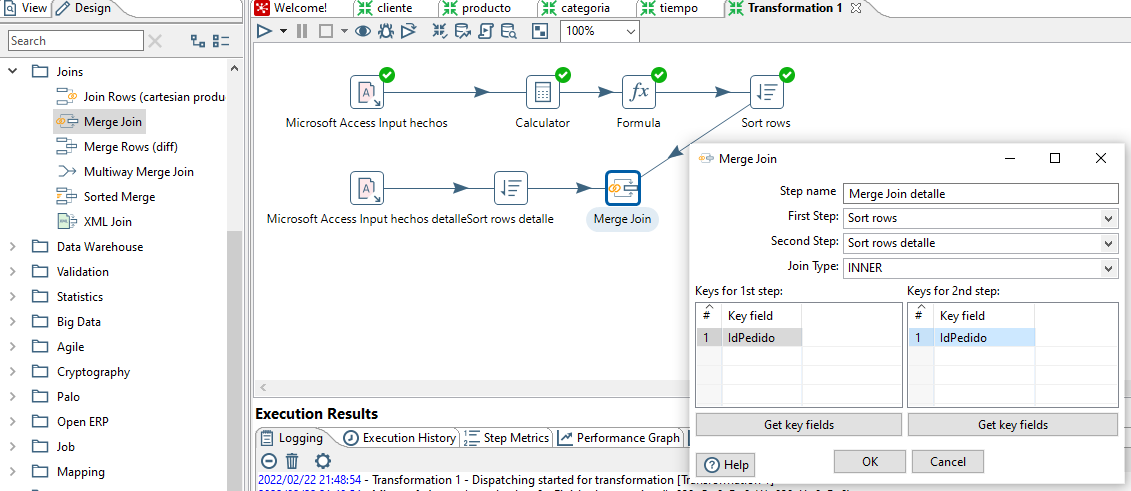
Descripción generada automáticamente

g) Cree un nuevo flujo. Realice el mismo procedimiento para obtener las fechas de los pedidos y transformarla para la tabla de hechos del esquema multidimensional, siguiendo los mismos pasos de los incisos a), b), c) y d) (no incluya el componente de eliminación de duplicados) Se deben incluir de la tabla *Pedidos* los campos *IdPedido, IdCliente, fechaPedido y Cargo.*

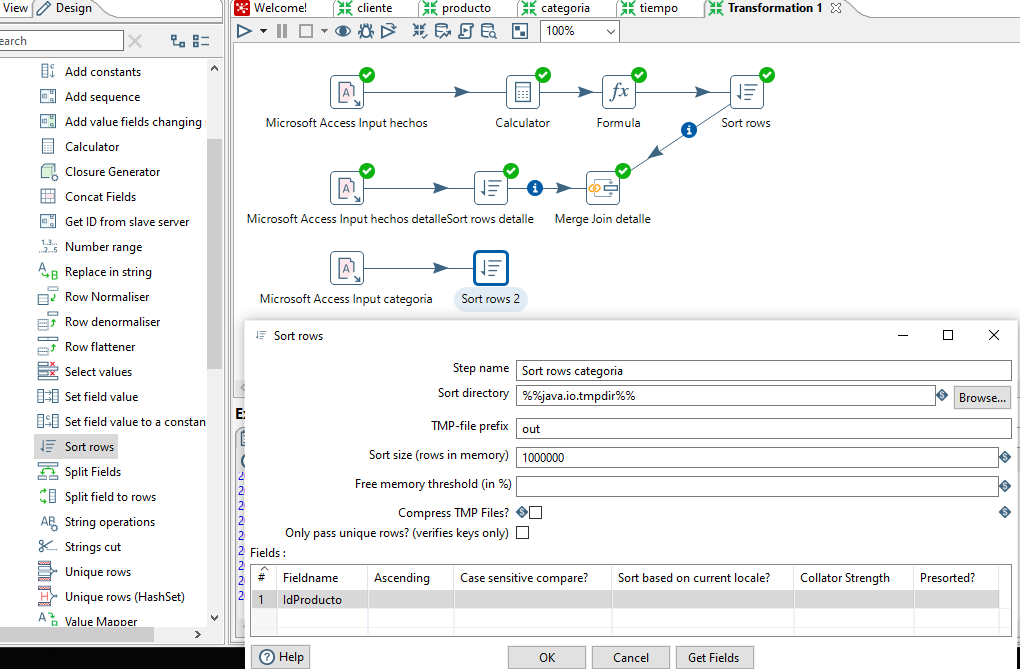
h) Para relacionar los datos del cliente con sus compras, es necesario hacer una combinación para cada pedido con los datos obtenidos hasta el momento de las fechas de los pedidos. Incluya en el mismo flujo, una conexión de entrada (**Access Input**) hacía la tabla de *detalles de pedidos* en las columnas *IdPedido, IdProducto y Cantidad*



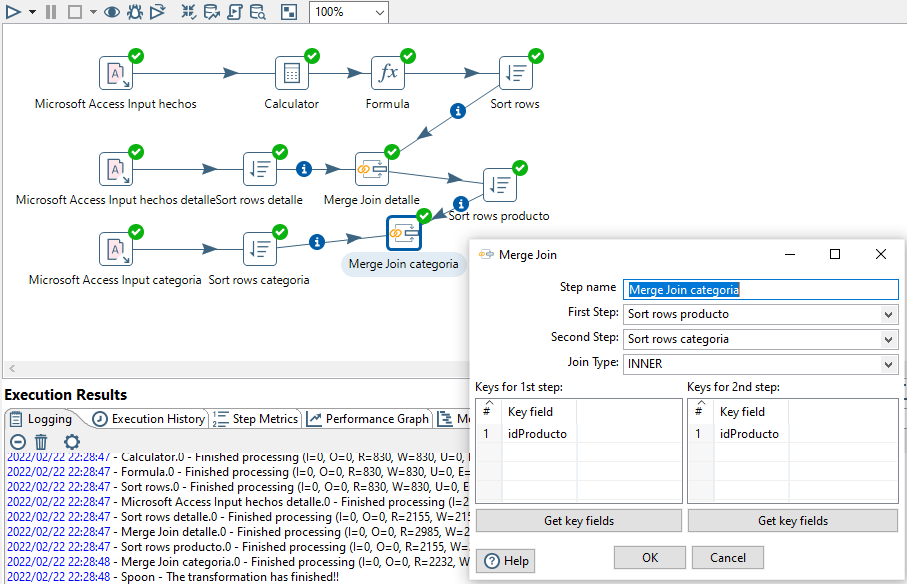
i) Incluya un componente de ordenamiento para la columna *IdPedido*. Incluya un paso de **Merge Join** entre las salidas de los ordenamientos anteriores. Configure este componente y establezca el tipo de reunión como INNER. En las llaves para realizar la reunión establezca para ambas la columna *idPedido*. Guarde los cambios y observe los resultados obtenidos mediante un preview.



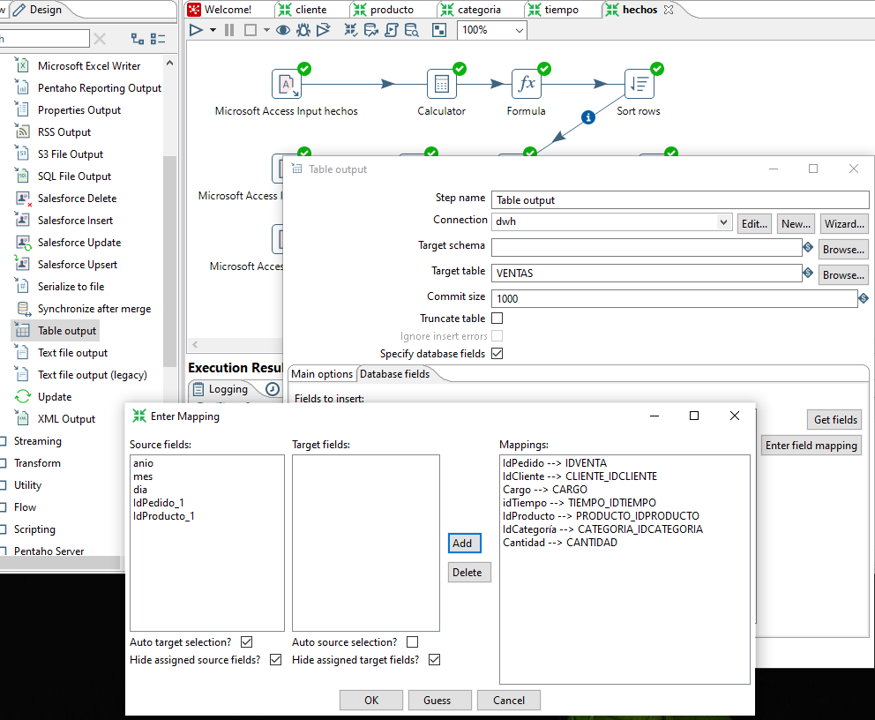
j) Incluya nuevamente un componente de entrada (**Access Input**) para la base de datos operacional de Access en la tabla de *Productos* incluyendo las columnas *IdProducto* e *IdCategoria*. Agregue un componente de ordenamiento (**Sorts Rows**) en la columna *idProducto*.



k) Del primer componente **Merge Join**, agregue un componente de ordenamiento (**Sorts Rows**) en la columna *idProducto.* Ahora incluya el paso de **Merge Join** para los resultados intermedios empleando como clave el *IdProducto* en ambas entradas con el tipo *INNER Join.* Observe el flujo en la siguiente figura.



l) Finalmente agregue el componente de salida hacía la tabla de *Ventas* (tabla de hechos) del modelo multidimensional. Configure la conexión como se ha realizado en los ejercicios anteriores para la base de datos de Oracle. Incluya los siguientes mapeos:



m) Incluir una captura con los resultados de aplicar el proceso anterior, así como del diagrama del flujo. NOTA: tome en consideración que este proceso se deberá realizar una sola vez, ya que se generarán errores de duplicados para varias ejecuciones.

Tabla

Descripción generada automáticamente

Incluya los resultados de los primeros 30 registros de la consulta SELECT \* FROM ventas del esquema multidimensional en Oracle (en la tabla de hechos).

Texto

Descripción generada automáticamente

CUESTIONARIO

1.- Describa claramente cada uno de los procesos realizados por una herramienta ETL

Paso 1: Extract

Extraer datos desde cualquier fuente, ya sea archivos planos, binarios, bases de datos o servicios cloud. Pueden ser fuentes operacionales internas (es decir, que recopilan información de transacciones en nuestra empresa), o bien fuentes externas. Esto en función de los análisis que se quieran llevar a cabo.

Paso 2: Transform

Transformar, limpiar o enriquecer la información extraída sin modificar la fuente. Es en este paso que se ajustan los datos según el modelo de datos (el cual se diseña previo a la creación del ETL).

Paso 3: Load

Cargar los datos ya transformados al modelo de datos.

2.- Con respecto a las transformaciones de un modelo relacional a un multidimensional, ¿qué consideraciones se deben de tomar en cuenta para una correcta carga en la base de datos?

Las restricciones correspondientes a las llaves primarias y llaves foreanas

CONCLUSIONES

Escriba sus conclusiones (no personales) con respecto al desarrollo de esta práctica.

El uso de herramientas etl para el procesamiento y visualización nos da una claridad especifica de los datos que utilizaremos en el procedimiento. Considerando nuevamente que el uso del modelo multidimensional , genera una mayor eficacia al momento de generar consultas referentes basadas en las preguntas del negocio y sin la necesidad de generar demasiados joins.

REFERENCIAS BIBLIOGRÁFICAS (APA, IEEE, ETC.)

1.- Ralph Kimball: “The Data Warehouse Toolkit:

Practical Techniques for Building Dimensional Data Warehouses”

John WIley & Sons, 1996. ISBN 0-471-15337-0

- Jiawei Han & Micheline Kamber “Data Mining: Concepts and Techniques”

Morgan Kaufmann, 2000. ISBN 1-55860-489-8

- C.J. Date: “Introducción a los sistemas de bases de datos”

Pearson Educación, 2001. ISBN 968-444-419.

- Ramez A. Elmasri & Shamkant B. Navathe: “Fundamentos de Sistemas de Bases de Datos”.

Addison-Wesley, 2002 [3ª edición]. ISBN 84-782-9051-6.

- Henry F. Korth, Abraham Silberschatz & S. Sudarshan: “Fundamentos de Bases de Datos”.

Mc-Graw Hill, 2002 [4ª edición]. ISBN 84-481-3654-3.