INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Alumno: De Luna Ocampo Yanina

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

**Práctica no. 2: Aplicación del proceso ETL a un modelo multidimensional**

**Número de Práctica**: 2

**Nombre de la Práctica**: Aplicación del proceso ETL a un modelo multidimensional

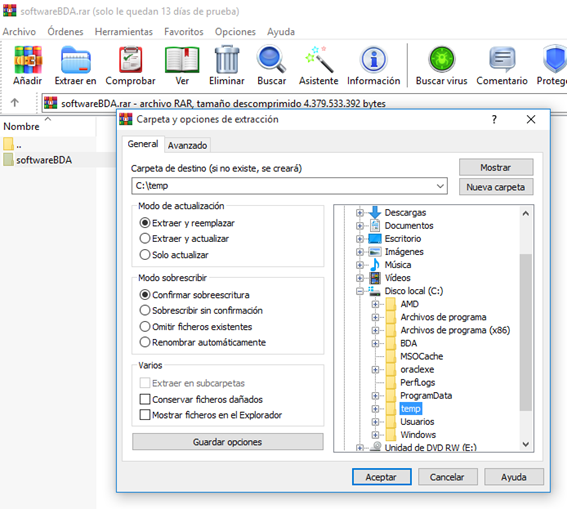
REQUISITOS PREVIOS

Para la realización de la práctica se deberá tener instalado en la computadora el software:

\* Pentaho Data Integration

que viene dentro del archivo comprimido *softwareBDA.rar* proporcionado por el profesor.

El proceso de instalación solo requiere descomprimir la carpeta llamada *data-integration* en algún espacio dentro del almacenamiento de la computadora local, por ejemplo en C:\temp\



Posteriormente se debe abrir una terminal de ejecución de comandos del Sistema Operativo en la ruta en donde se descomprimió la carpeta y ejecutar las sentencias mostradas a continuación:

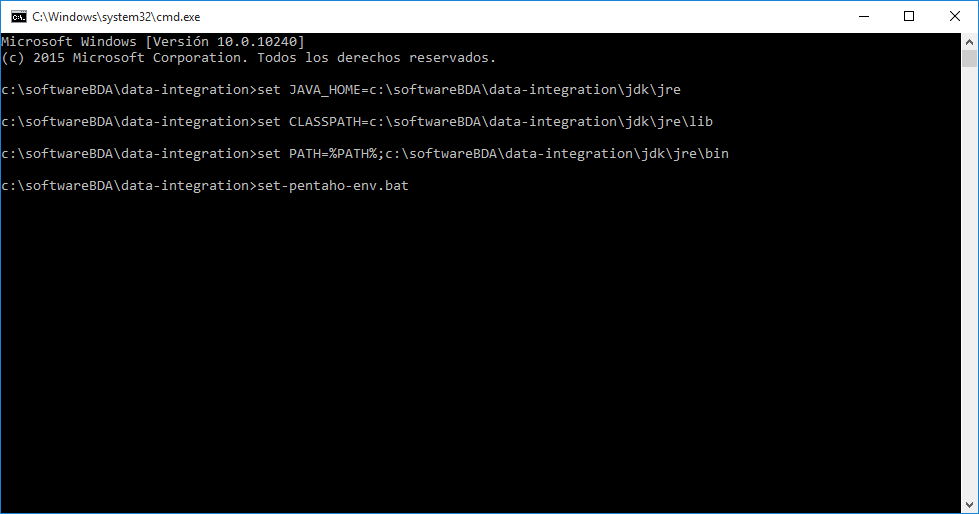
C:\[carpeta\_de\_instalacion]\data-integration\> **set JAVA\_HOME=c:\ [carpeta\_de\_instalacion]\data-integration\jdk\jre**

C:\[carpeta\_de\_instalacion]\data-integration\> **set CLASSPATH=c:\ [carpeta\_de\_instalacion]\data-integration\jdk\jre\lib**

C:\[carpeta\_de\_instalacion]\data-integration\> **set PATH=%PATH%;c:\ [carpeta\_de\_instalacion]\data-integration\jdk\jre\bin**

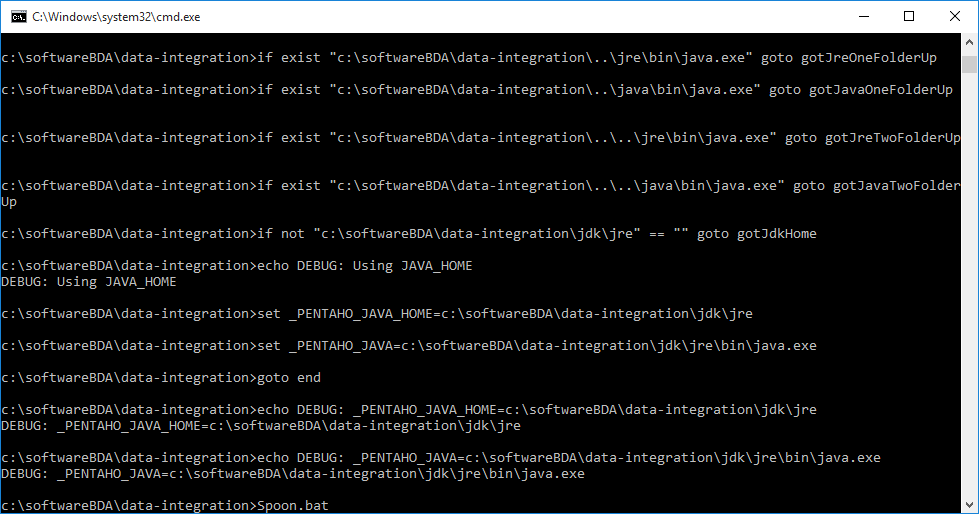
C:\[carpeta\_de\_instalacion]\data-integration\>**set-pentaho-env.bat**

Tome en consideración la ruta en que se descomprimió el software para no tener problemas en el arranque:

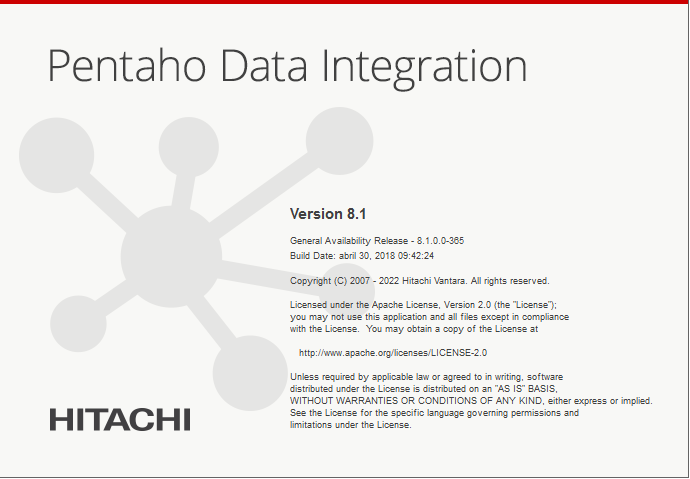
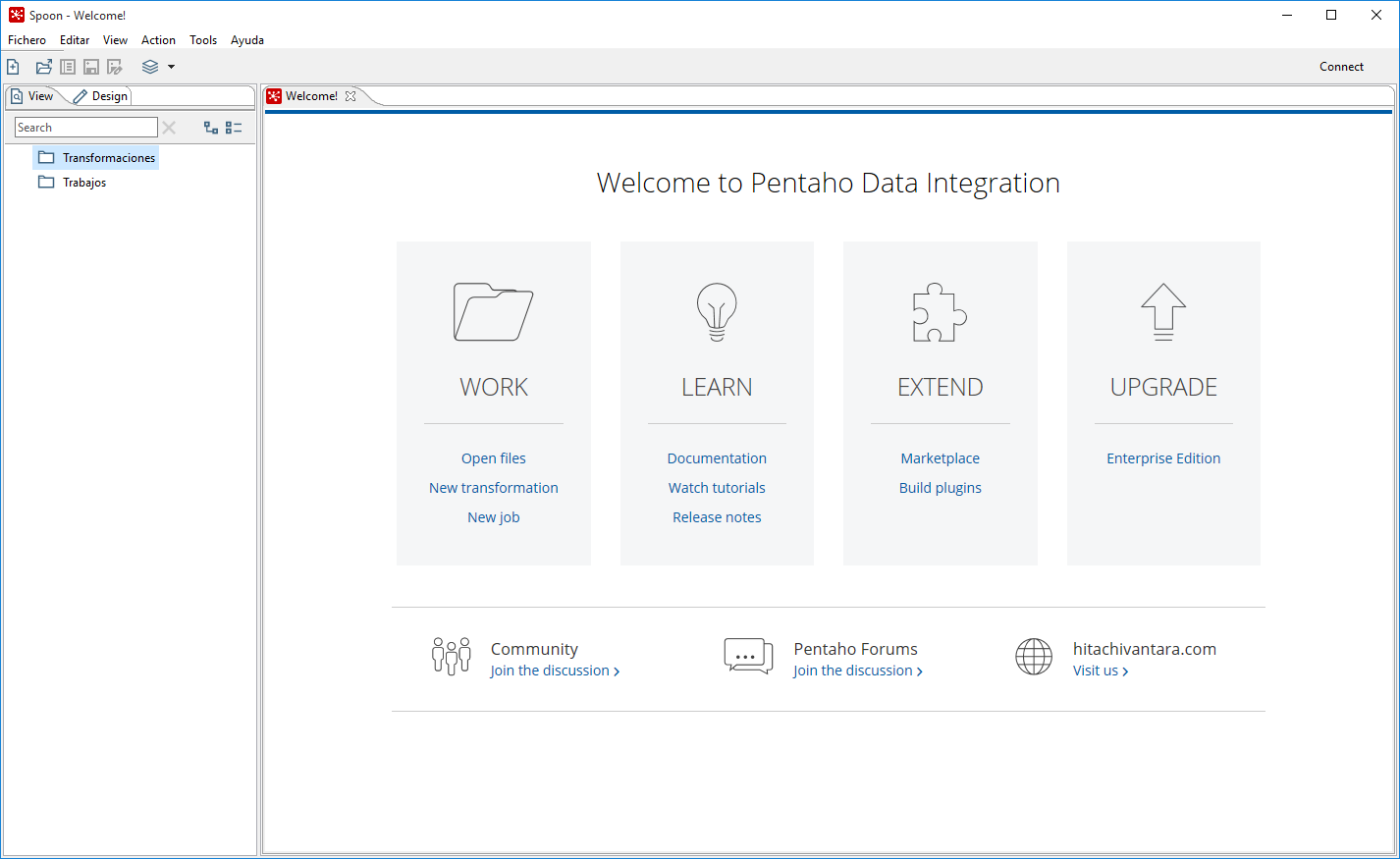


En este paso se configura el ambiente y puede tardar unos momentos, para posteriormente ejecutar propiamente el software con la siguiente instrucción:

C:\[carpeta\_de\_instalacion]\data-integration\>**Spoon.bat**



Una vez realizado el paso anterior, el software inicia mostrando la pantalla de bienvenida y el panel de la aplicación:

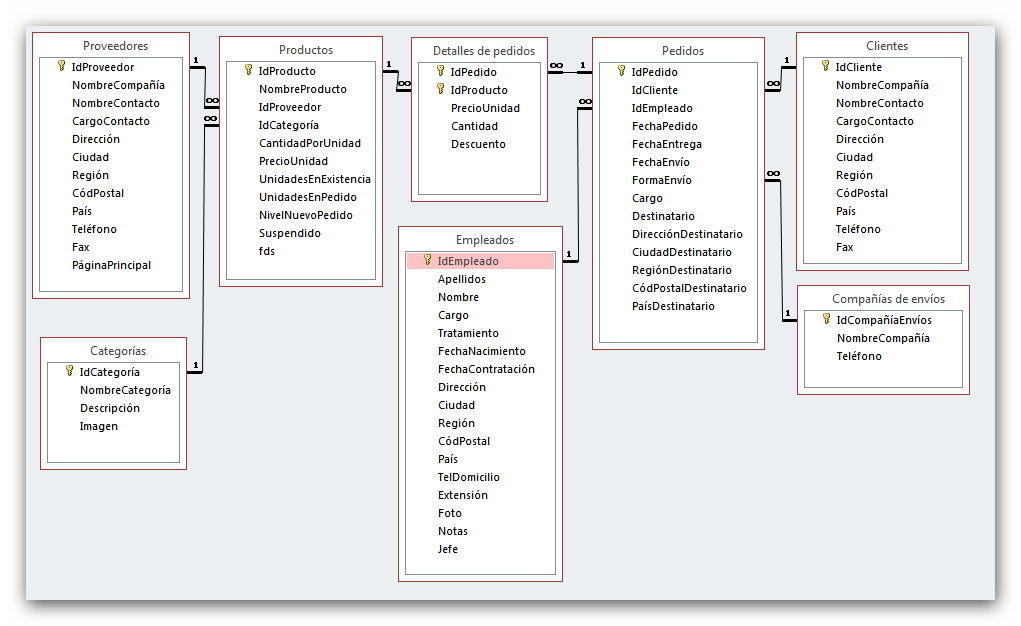


**MARCO TEÓRICO**

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:**

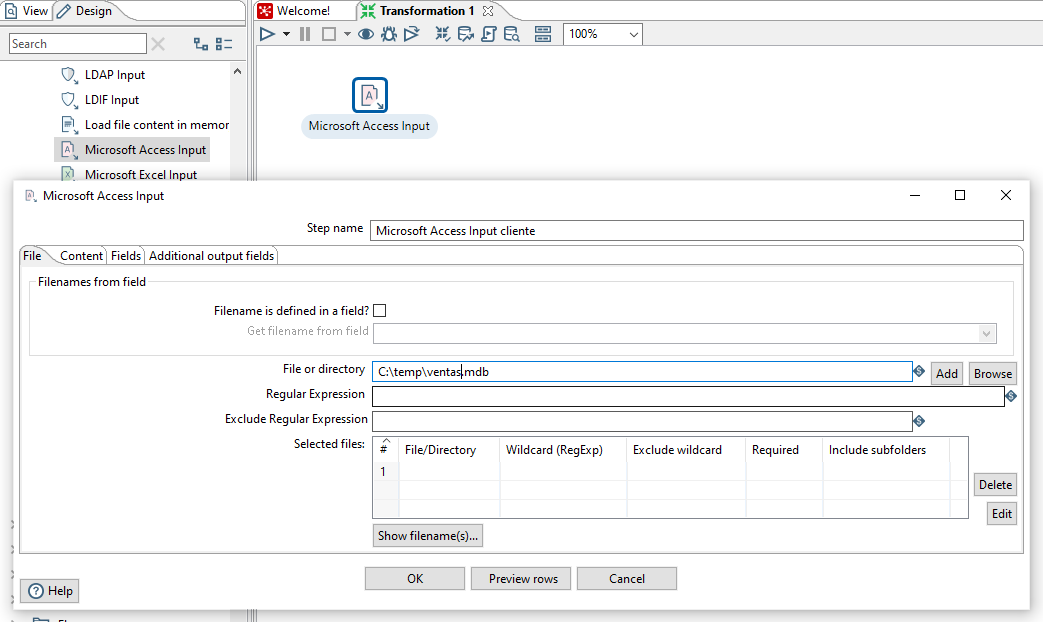
1.- Descargue el archivo *ventas.mdb* que representa una base de datos operacional de Access, y guárdelo en una carpeta accesible en la computadora local.

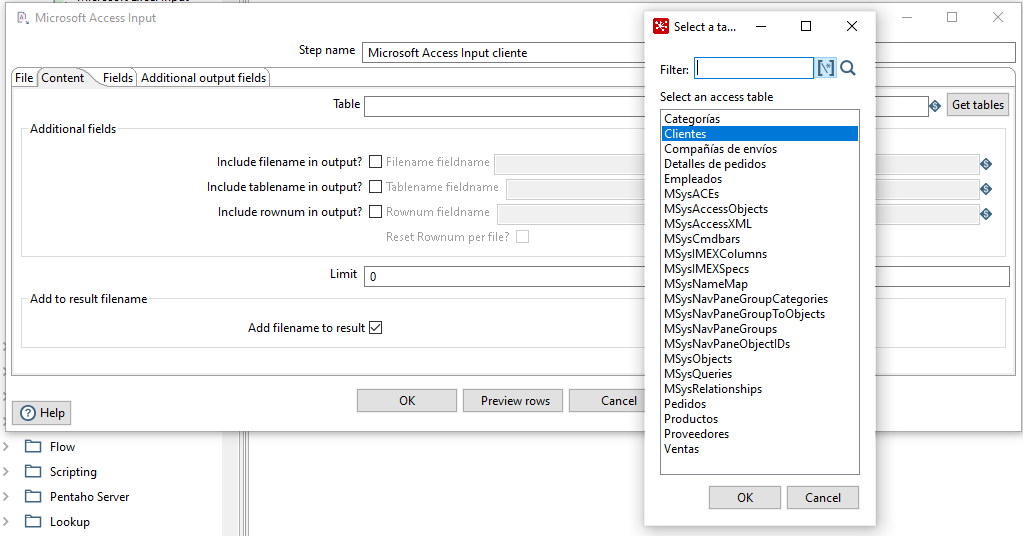


2.- Inicie el servidor Oracle XE en su computadora local. Verifique que tenga conexión y que los procesos ejecuten correctamente. Conéctese con el usuario creado en la práctica anterior. Si aún no tiene el esquema multidimensional, es necesario que lo cree, usando las sentencias incluidas en el Anexo 1.

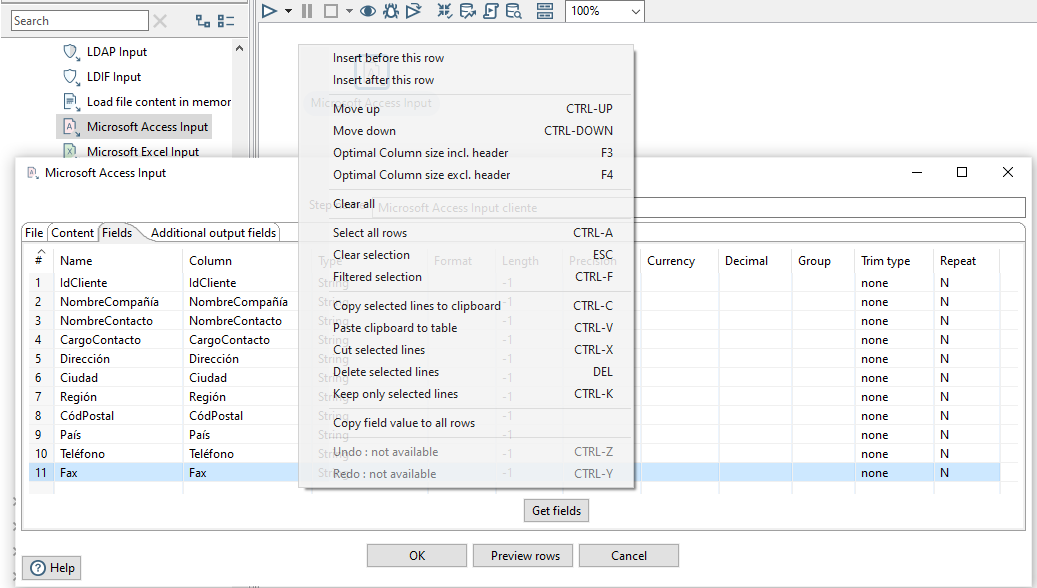
3.- Una vez que la aplicación esté lista y en ejecución, realice los siguientes pasos:

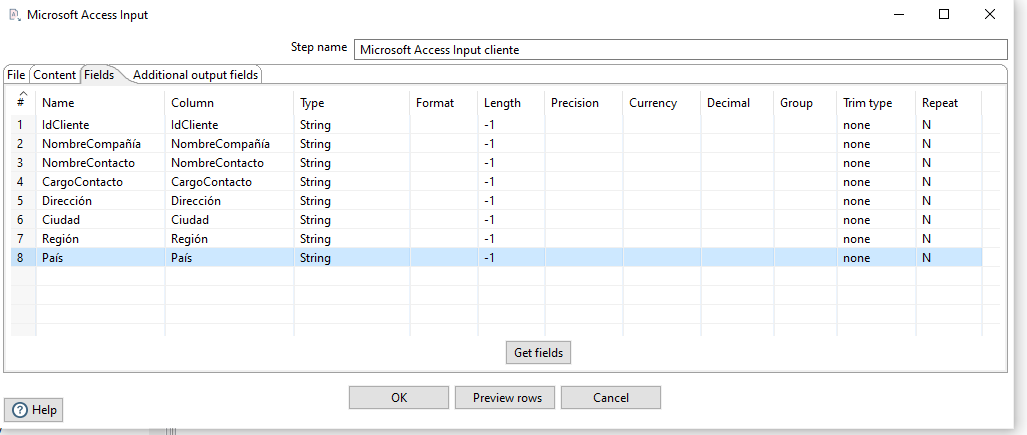
a) Cree una nueva transformación y agregue los procesos de tipo **Input -> Microsoft Access Input,** arrastrando el icono correspondiente al panel principal, y editando sus propiedades para que se pueda conectar a la fuente de datos *Ventas* de la base de datos de Access descargada en el paso 1. Compruebe la conexión obteniendo algunos datos de previsualización de la tabla *Cliente*.



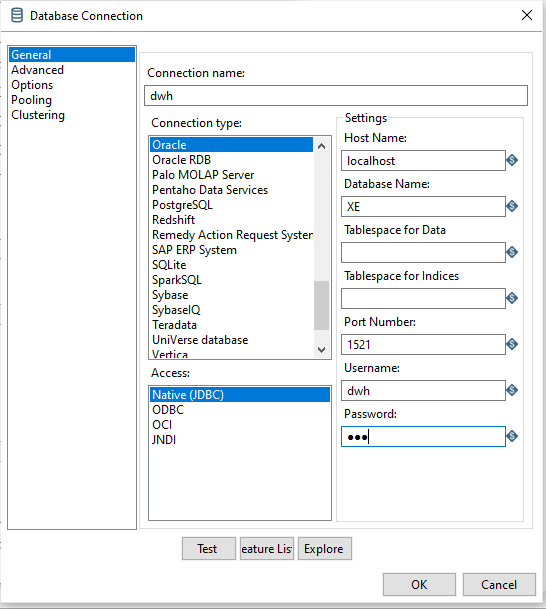


b) De las columnas de la tabla *Cliente*, descarte aquellas que no son necesarias para el modelo multidimensional. Mantenga únicamente las columnas de *IDCliente, NombreCompañia, NombreContacto, CargoContacto, Dirección, Ciudad, Región,* y *País.*

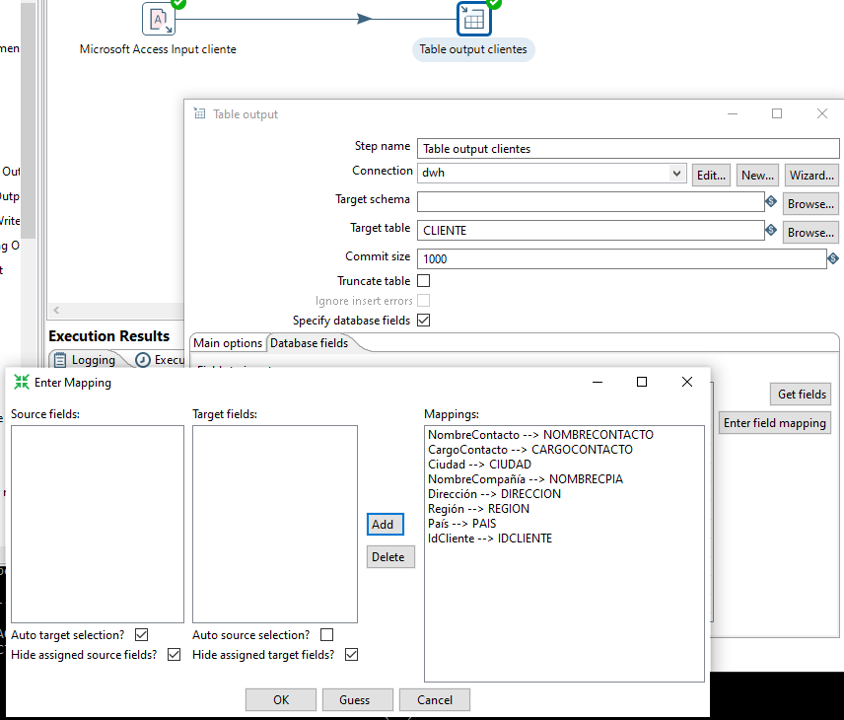




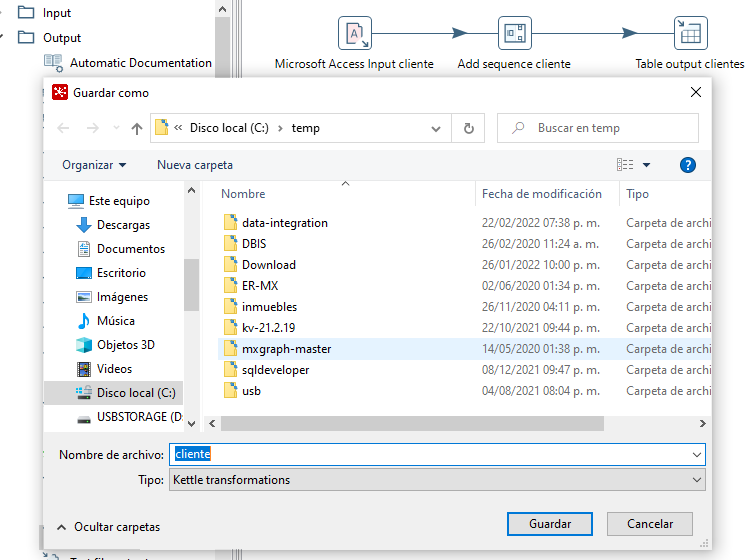
c) Agregue un icono de *Table output* para enviar el proceso como salida a la conexión del servidor Oracle (en la base de datos multidimensional creada de la práctica anterior), usando el controlador JDBC correspondiente y con los valores de conexión apropiados.

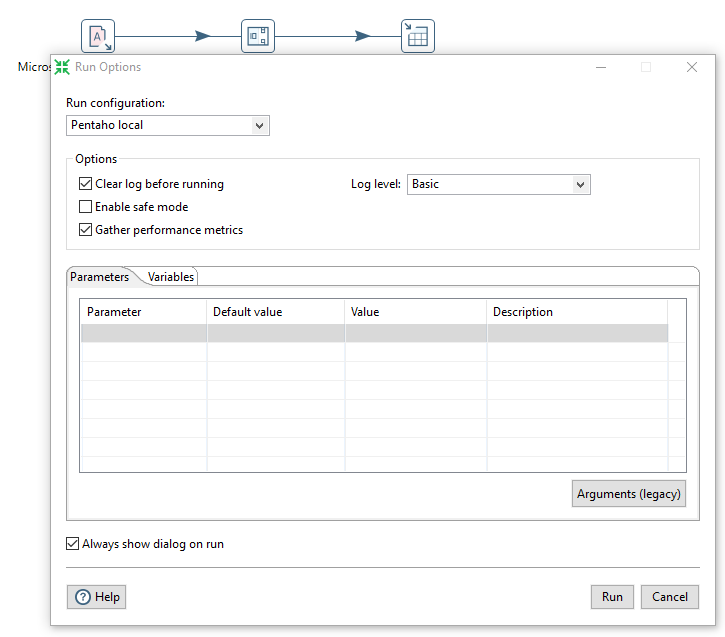






d) Guarde el proceso de transformación como un archivo *clientes.ktr* y ejecútelo; observe los resultados

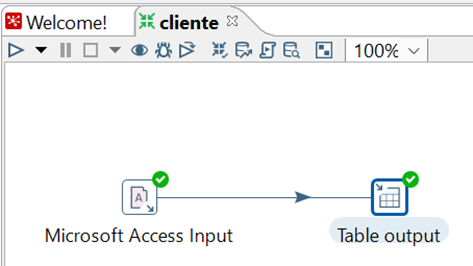




Incluya una captura de pantalla del proceso creado, incluyendo el *preview* de los datos generados de forma correcta. Ejecute un comando ***select idCliente, nombrecontacto from cliente;*** desde la ventana SQL de Oracle. Incluya los resultados obtenidos.

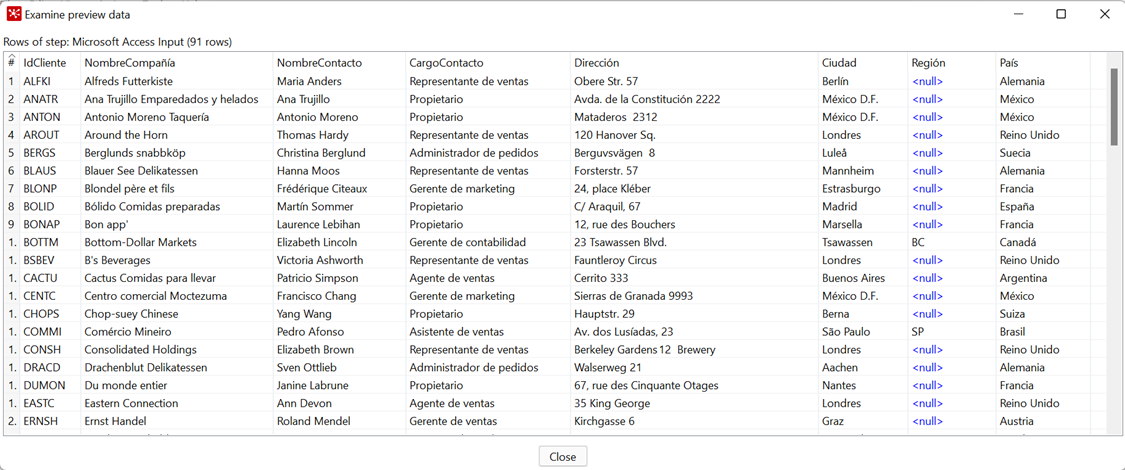
Cliente

1. Proceso creado

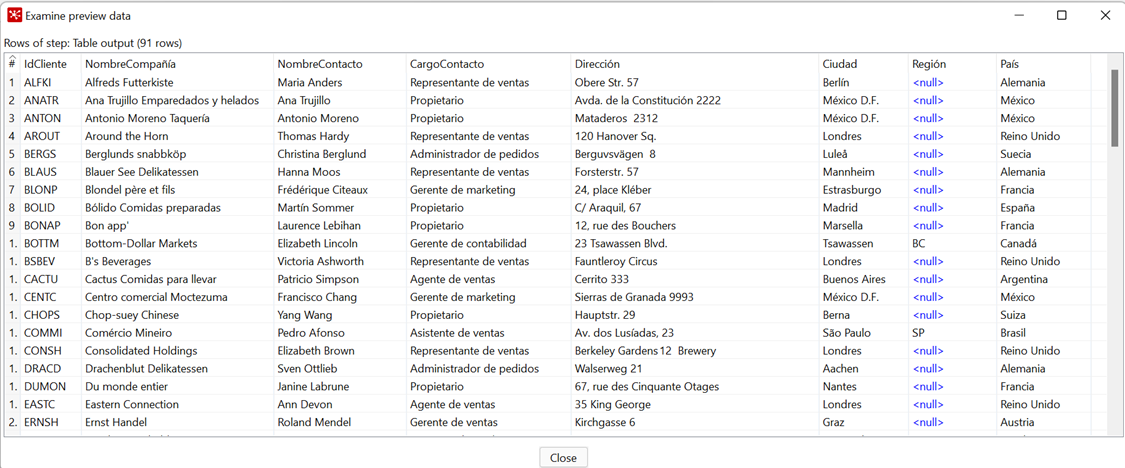


1. Preview

* Microsoft Access Input

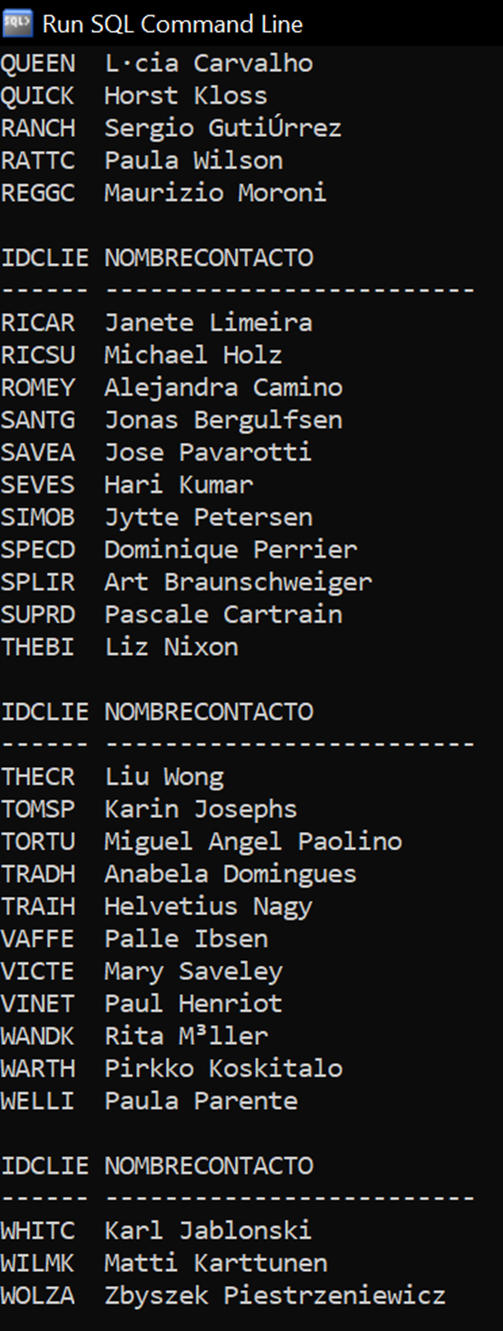


* Table output

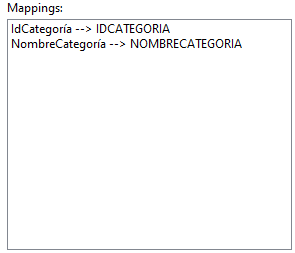
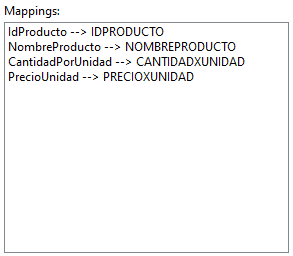


1. select idCliente, nombrecontacto from cliente;





e) Repita el proceso, de manera similar, para las tablas de *producto* y *categoría,* considerando los esquemas de las tablas origen (Access) y destino (Oracle). Realice los siguientes mapeos:

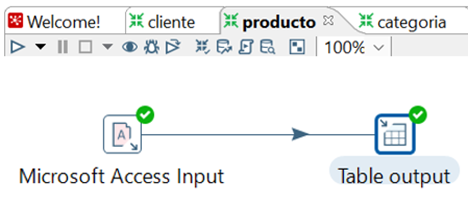


Guarde con el mismo nombre cada uno de los procesos generados.

Incluya una captura de pantalla de los procesos creados, incluyendo el *preview* de los datos generados de forma correcta.

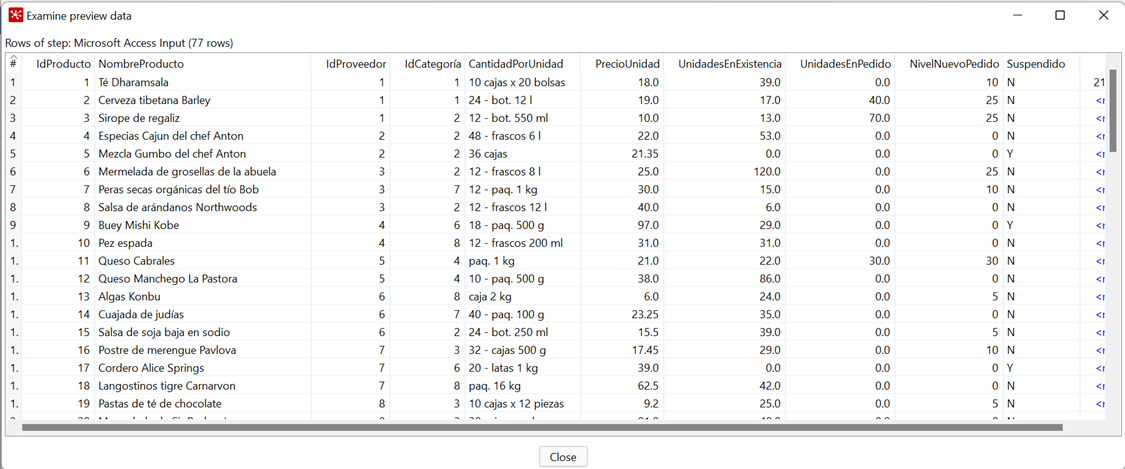
Producto:

1. Proceso creado

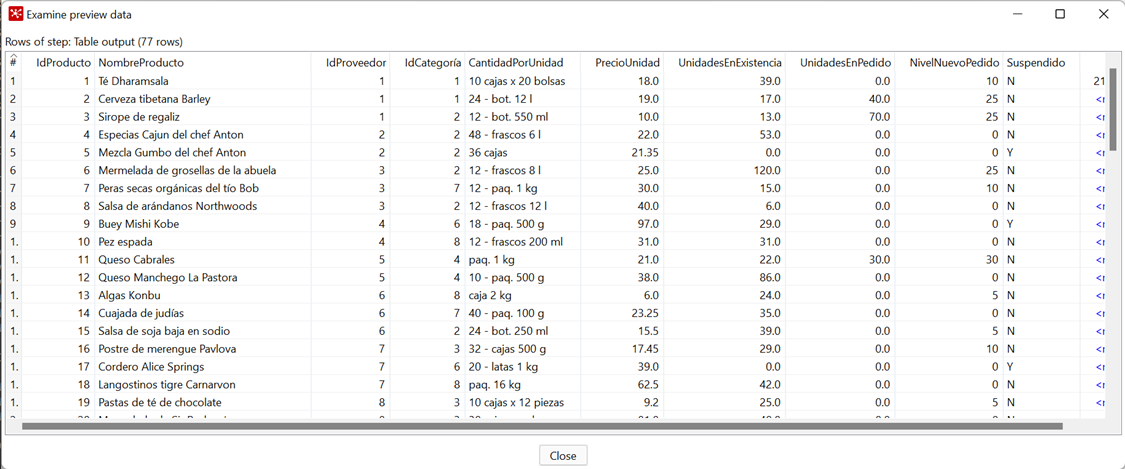


1. Preview

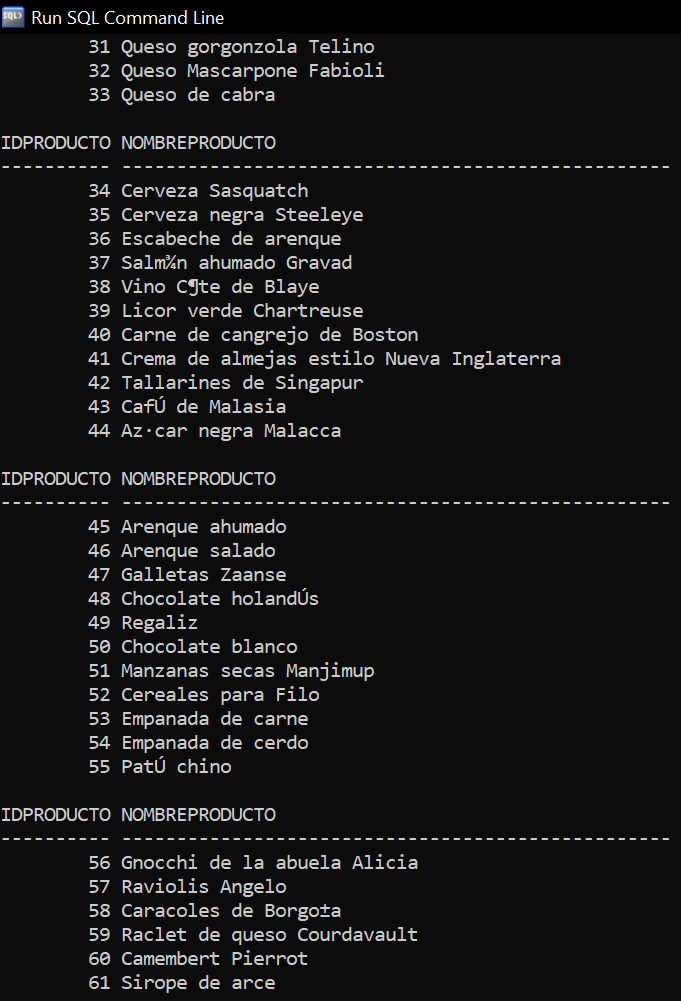
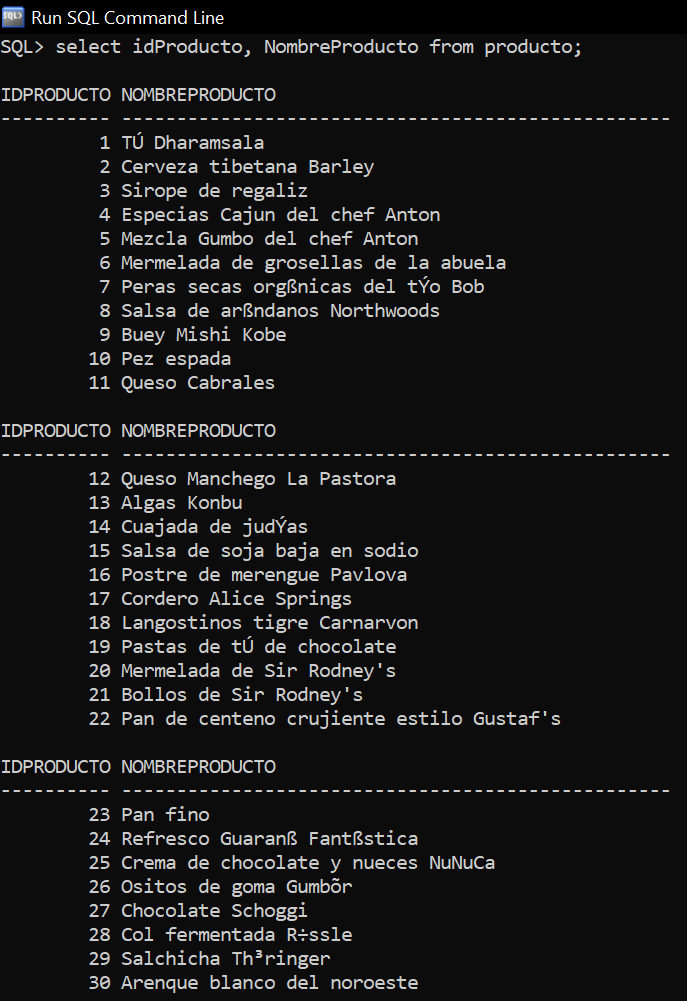
* Microsoft Access Input

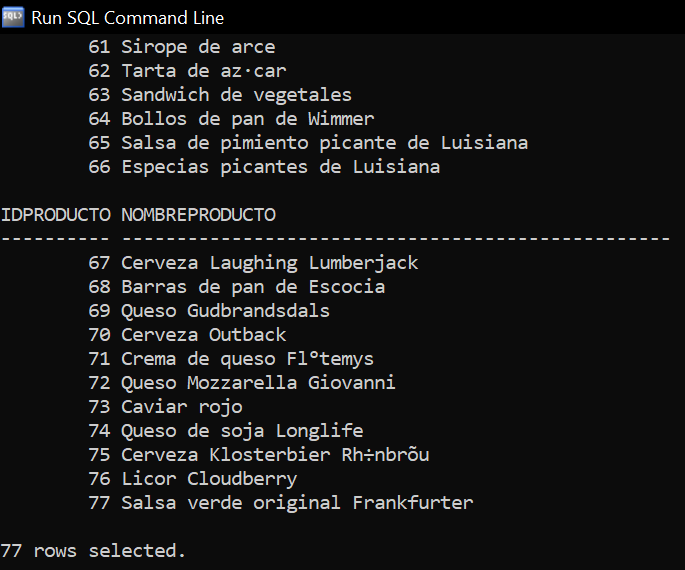


* Table output



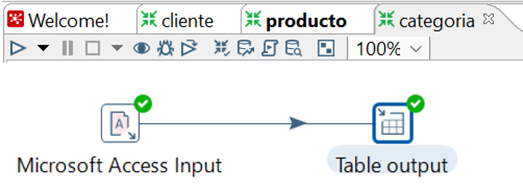
1. select idProducto, NombreProducto from producto;





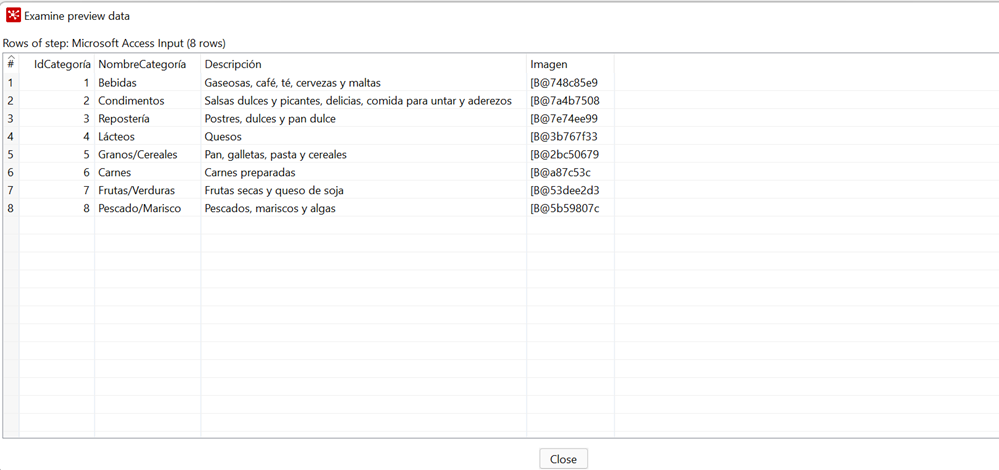
Categoría

1. Proceso creado

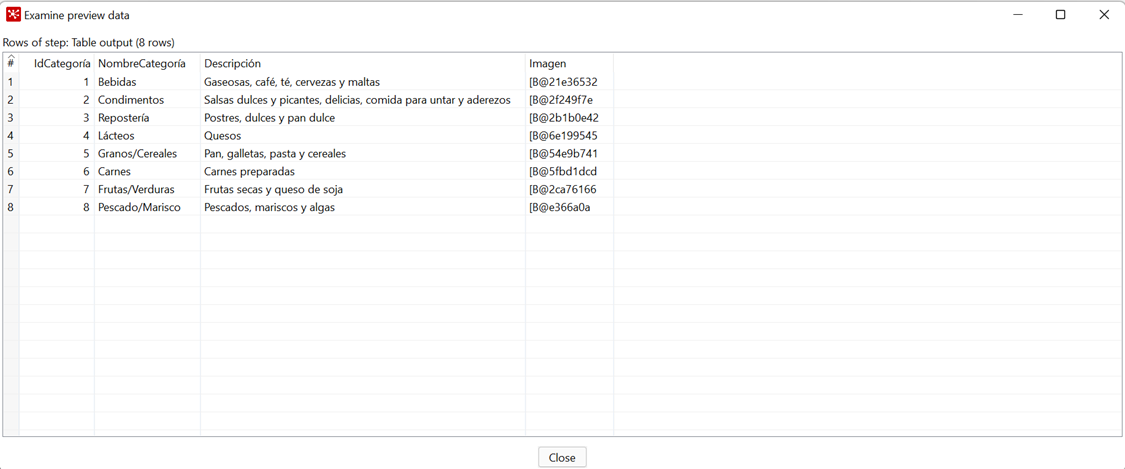


1. Preview

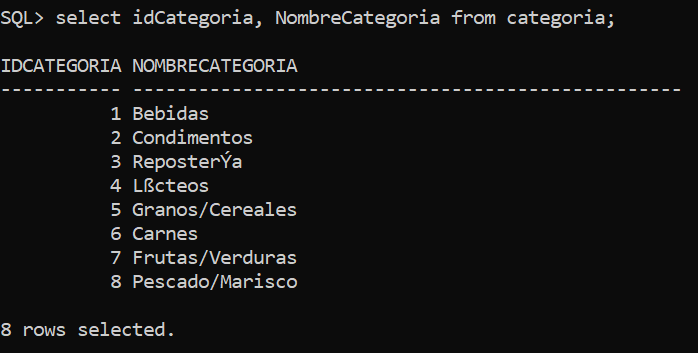
* Microsoft Access Input



* Table output

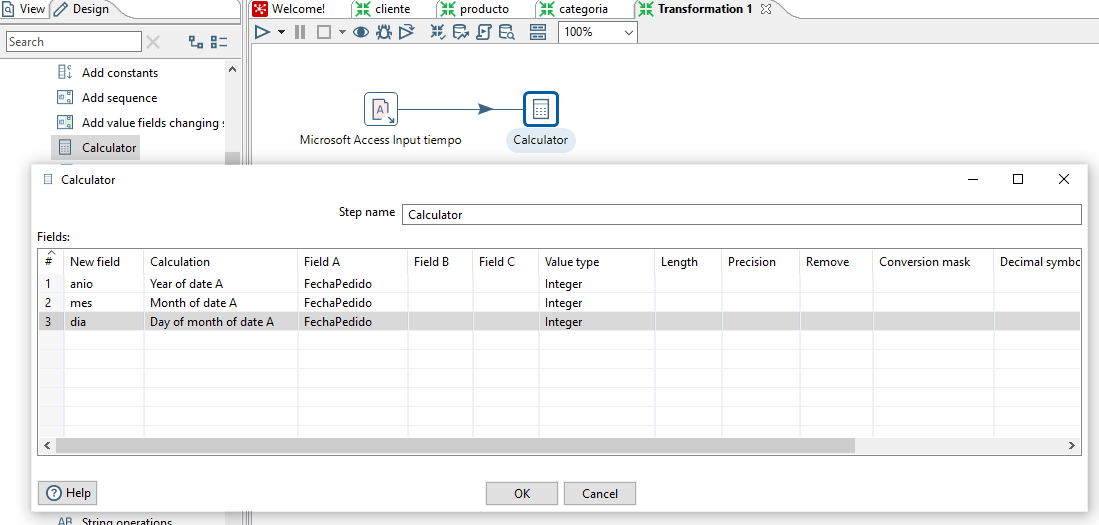


1. select idCategoria, NombreCategoria from categoria;



f) 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 obtener el año, el mes y el día de la fecha para convertirlo en el identificador 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

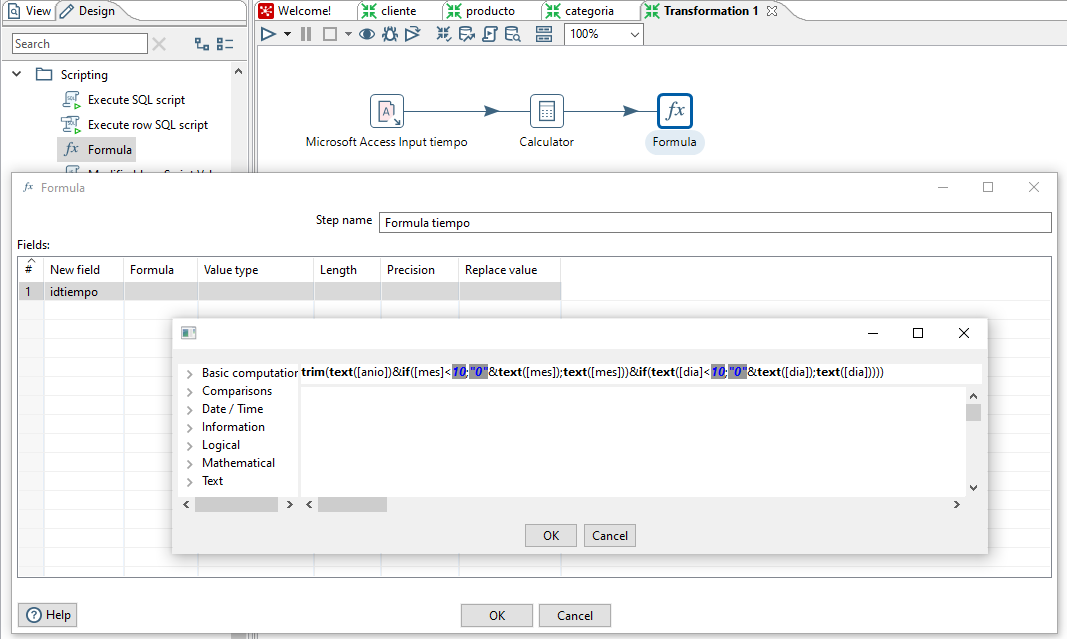
g) Incluya un paso *Calculator* y líguelo con el de **Input** para la tabla *Pedidos* de **Access**. Cree las columnas *anio*, *mes*, *dia* y aplique el cálculo correspondiente a la extracción del año, del mes y del día, como se observa en la siguiente figura. El atributo de entrada debe ser el mismo,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



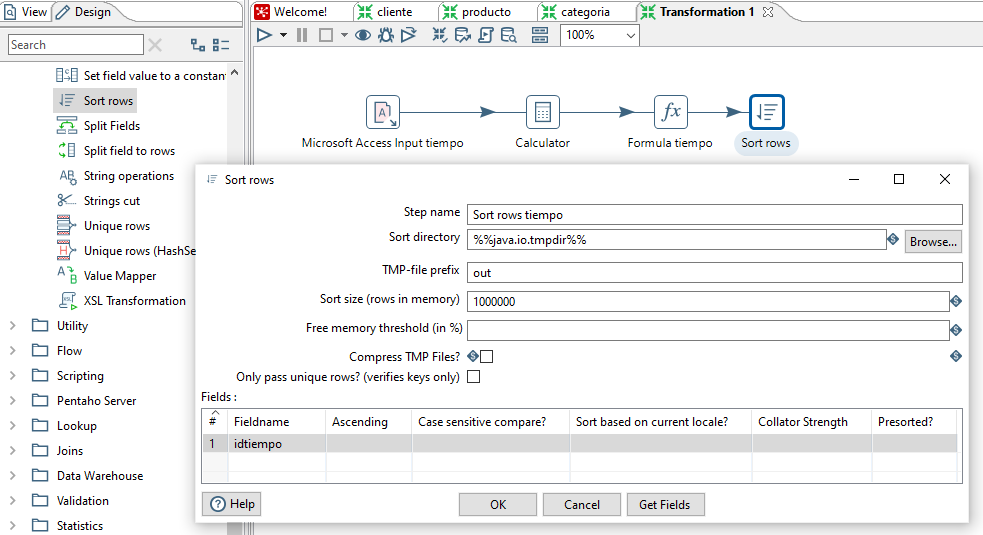
h) Incluya un operador *Formula* en el proceso, ligado al anterior. Cree un nuevo campo llamado *idTiempo* y en la formula escriba la siguiente expresión:

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

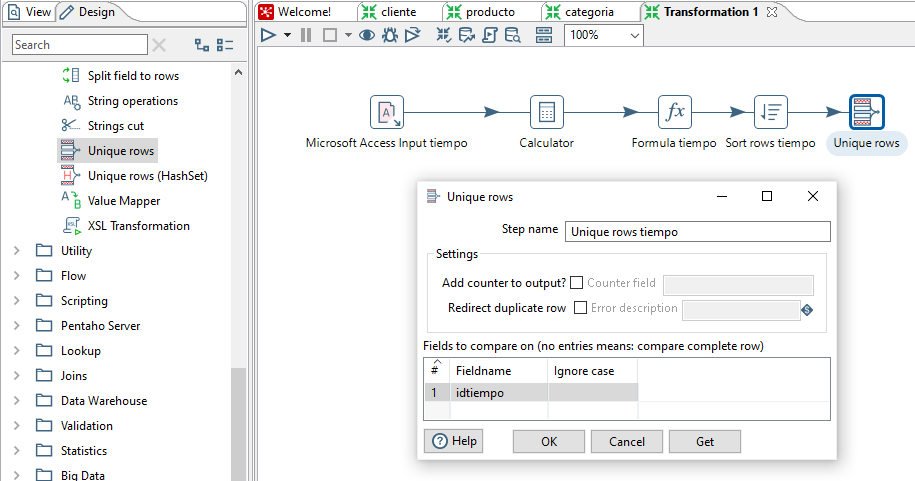
Establezca el tipo de dato a *String* y acepte los cambios.



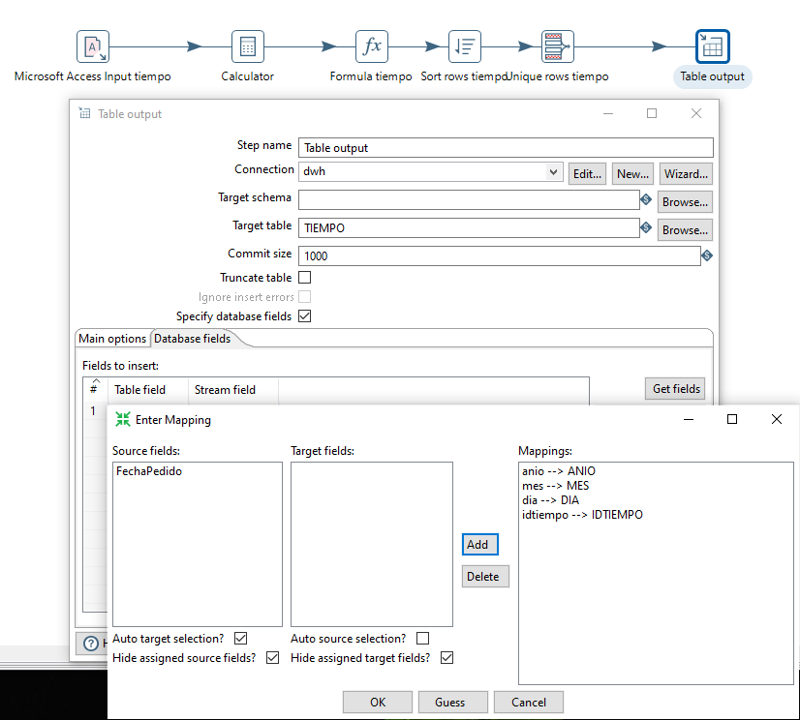
i) Incluya un operador *Sort Rows* y líguelo con el anterior. Seleccione el campo *idTiempo* como criterio de ordenamiento, y configure para que sea ascendente. Guarde los cambios.



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



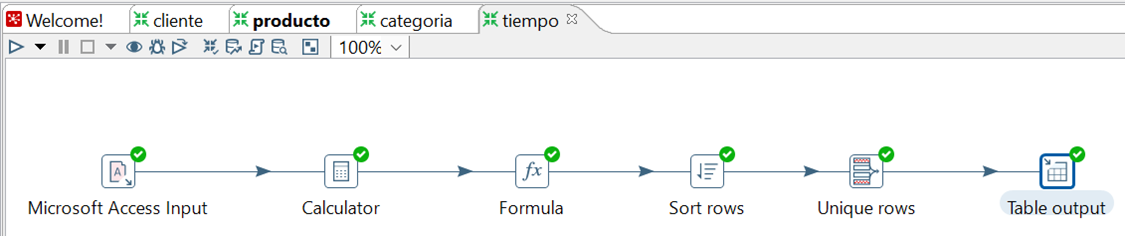
k) Finalmente agregue un paso *Table Output* y establezca una conexión con la tabla *Tiempo* de la base de datos multidimensional (Oracle). Guarde las modificaciones como un proceso (*Tiempo.ktr*).



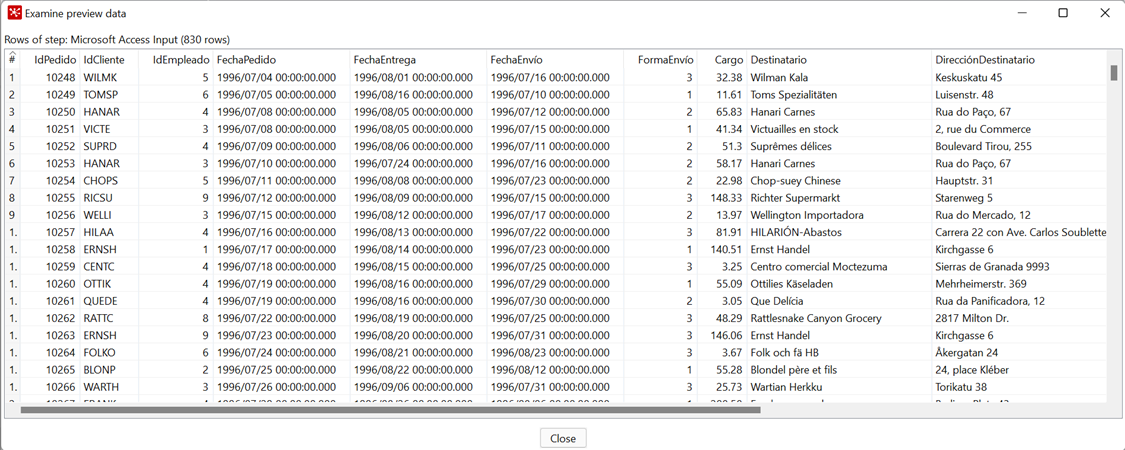
Incluya una captura de pantalla del proceso *Tiempo* creado, incluyendo el *preview* de los datos generados de forma correcta.

Tiempo

1. Proceso creado

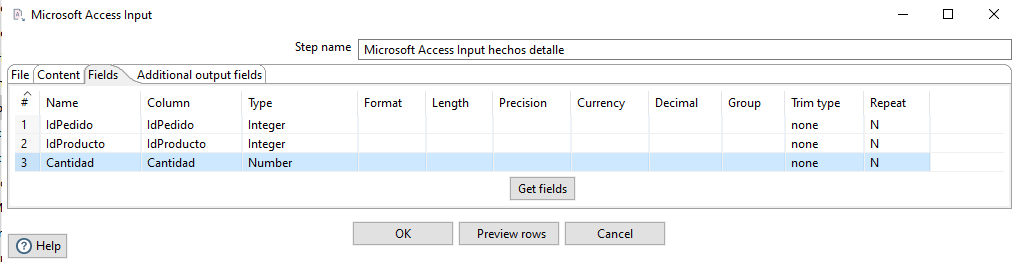


1. Preview

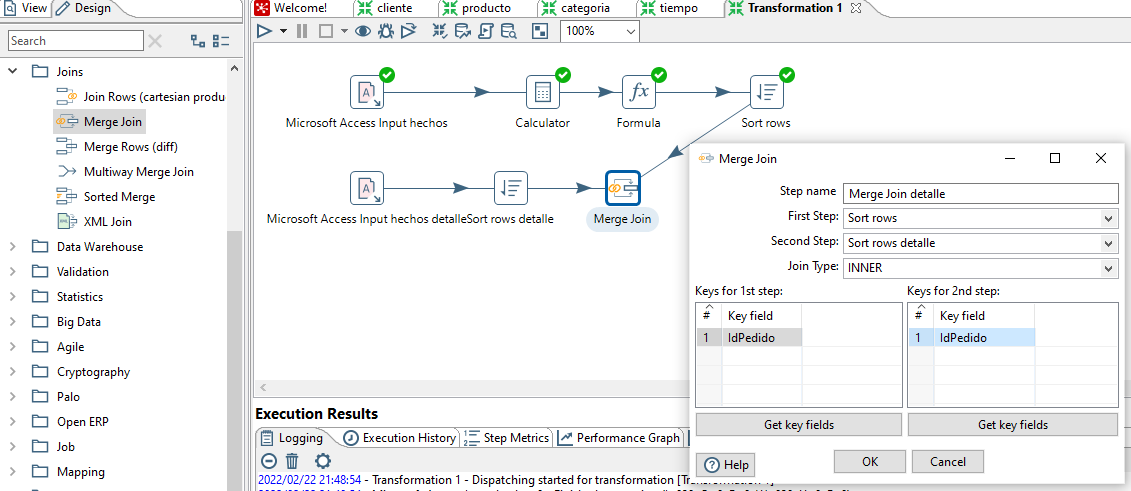


l) Abra una nueva Transformación en el panel del proceso. Realice el mismo procedimiento para obtener las fechas de los pedidos y transformarla para la tabla de hechos del esquema multidimensional, siguiendo los mismos pasos de los incisos f), g), h) y i). Se deben incluir de la tabla *Pedidos* los campos *IdPedido, IdCliente, fechaPedido y Cargo.*

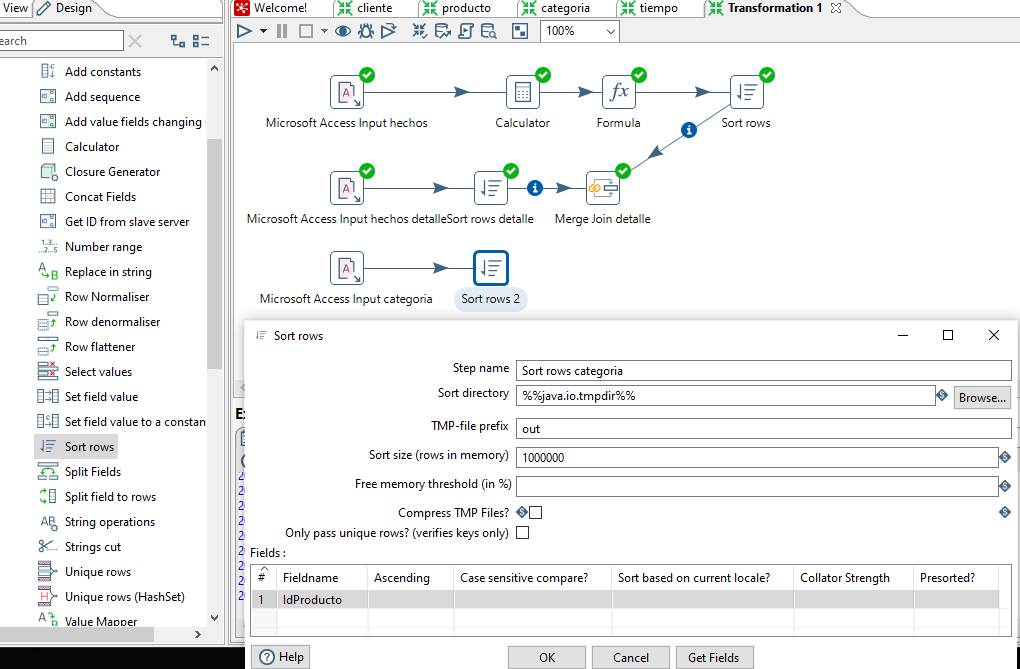
m) Para relacionar los datos del cliente con sus compras, es necesario hacer una combinación para cada pedido con los datos obtenidos hasta el momento de las fechas de los pedidos. Incluya una conexión de entrada hacía la tabla de *detalles de pedidos* en las columnas *IdPedido, IdProducto y Cantidad*



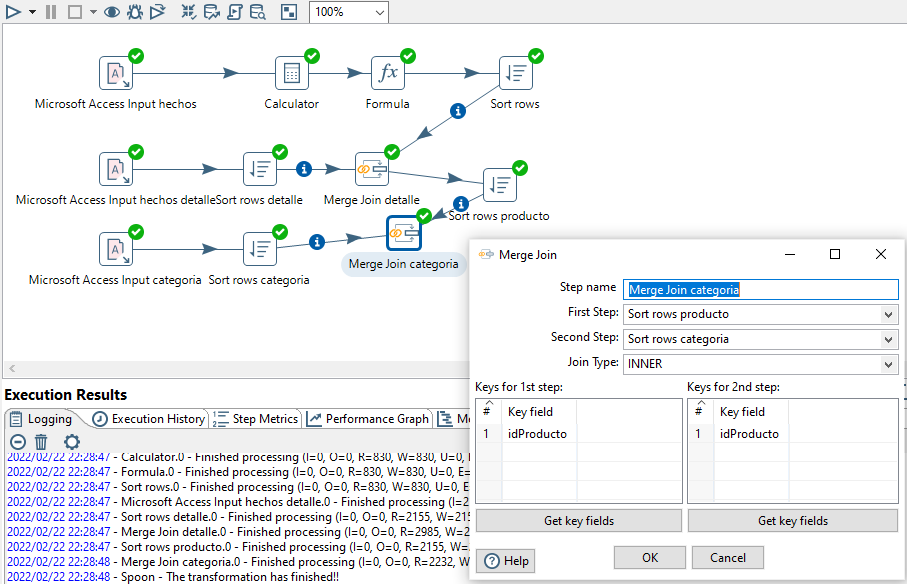
n) Incluya un componente de ordenamiento para la columna *IdPedido*. Incluya un paso de **Merge Join** entre las salidas de los ordenamientos anteriores. Establezca el tipo de reunión como INNER. En las llaves para realizar la reunión establezca para ambas la columna *idPedido*. Guarde los cambios y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



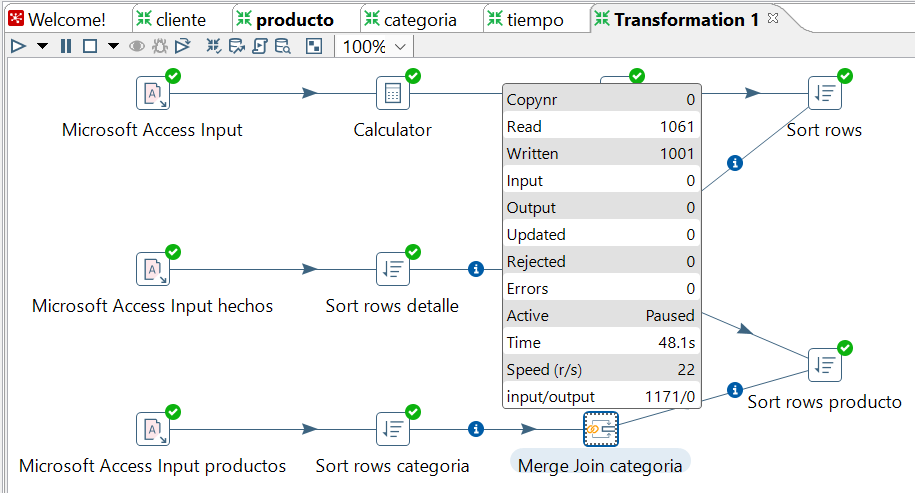
o) Incluya nuevamente un componente de entrada para la base de datos operacional de Access en la tabla de *Productos* incluyendo las columnas *IdProducto* e *IdCategoria*. Agregue un paso de ordenamiento en la columna *idProducto*.

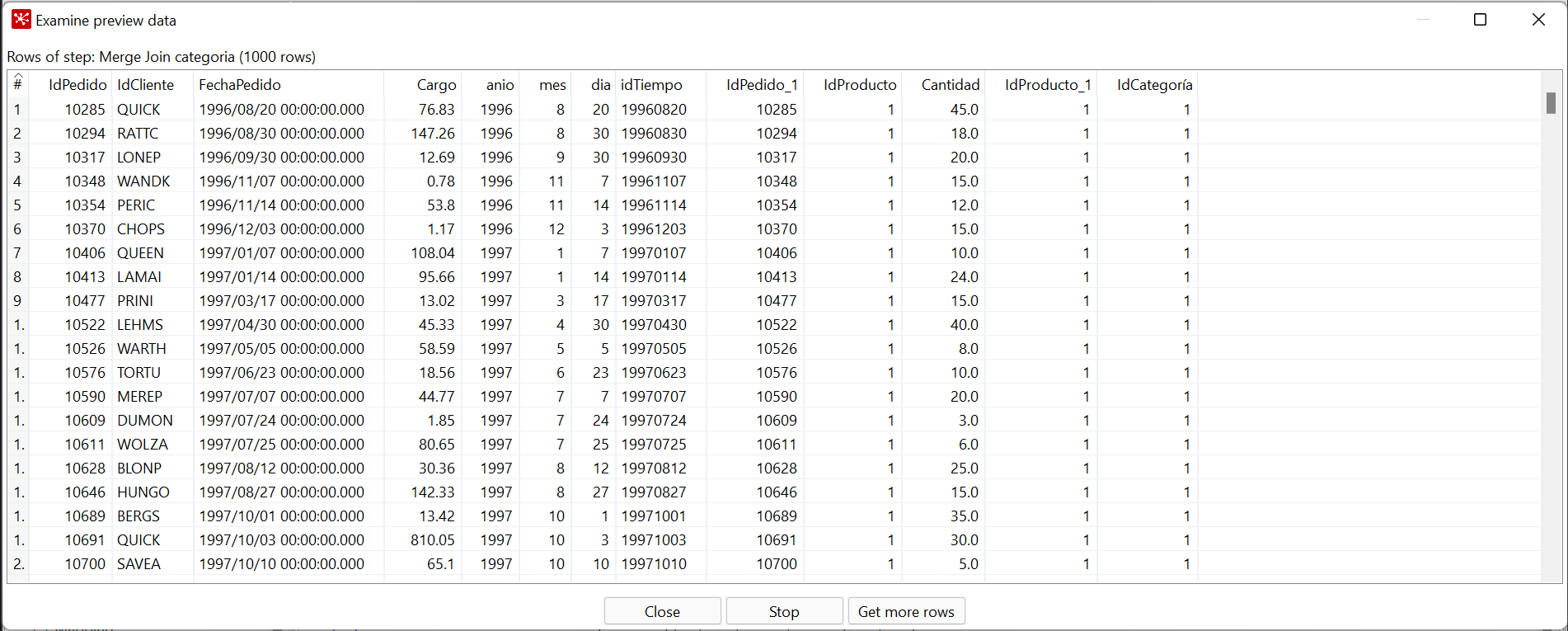


p) Antes de incluir un **Merge Join** entre los pedidos y las categorías, es necesario incluir en el flujo intermedio un ordenamiento por *IdProducto*. Ahora incluya el paso de **Merge Join** para los resultados intermedios empleando como clave el *IdProducto* en ambas entradas con el tipo *INNER Join.* Observe el flujo en la siguiente figura.

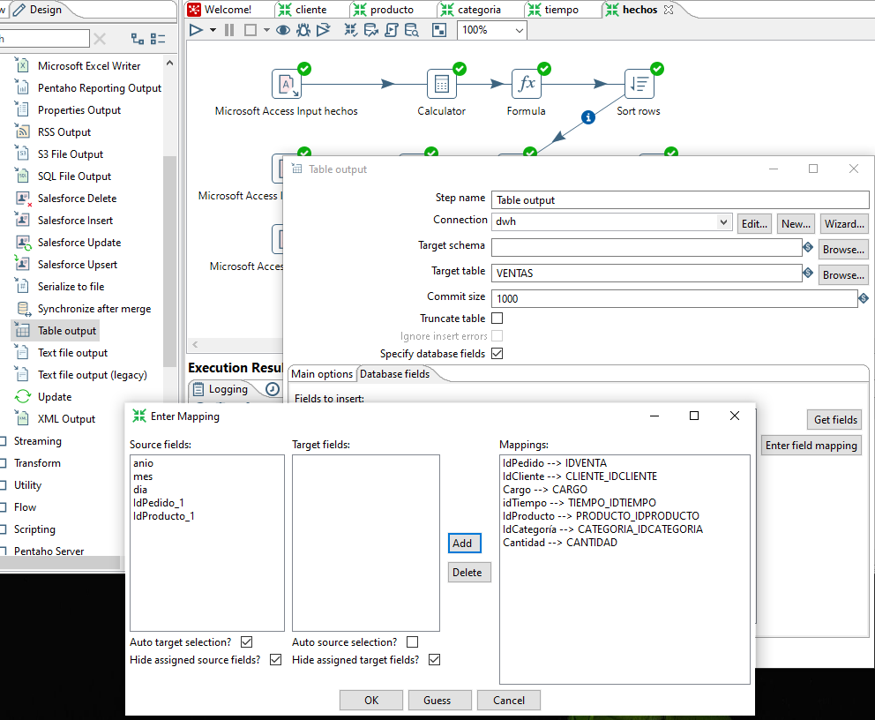


q) Ejecute el flujo hecho hasta el momento y muestre los resultados obtenidos.

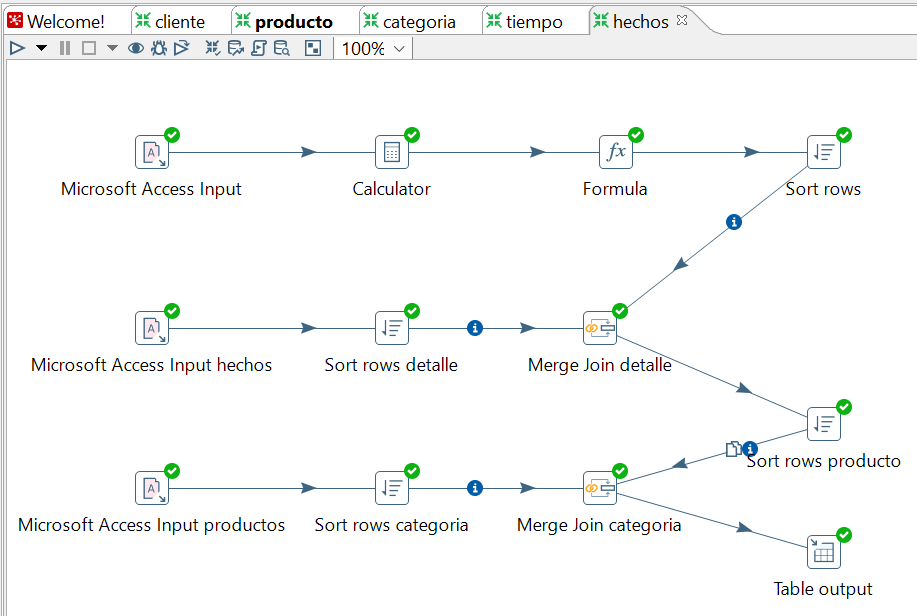




r) Finalmente agregue el componente de salida hacía la tabla de Ventas (hechos) del modelo multidimensional. Configure la conexión como se ha realizado en los ejercicios anteriores para la base de datos de Oracle. Incluya los siguientes mapeos:



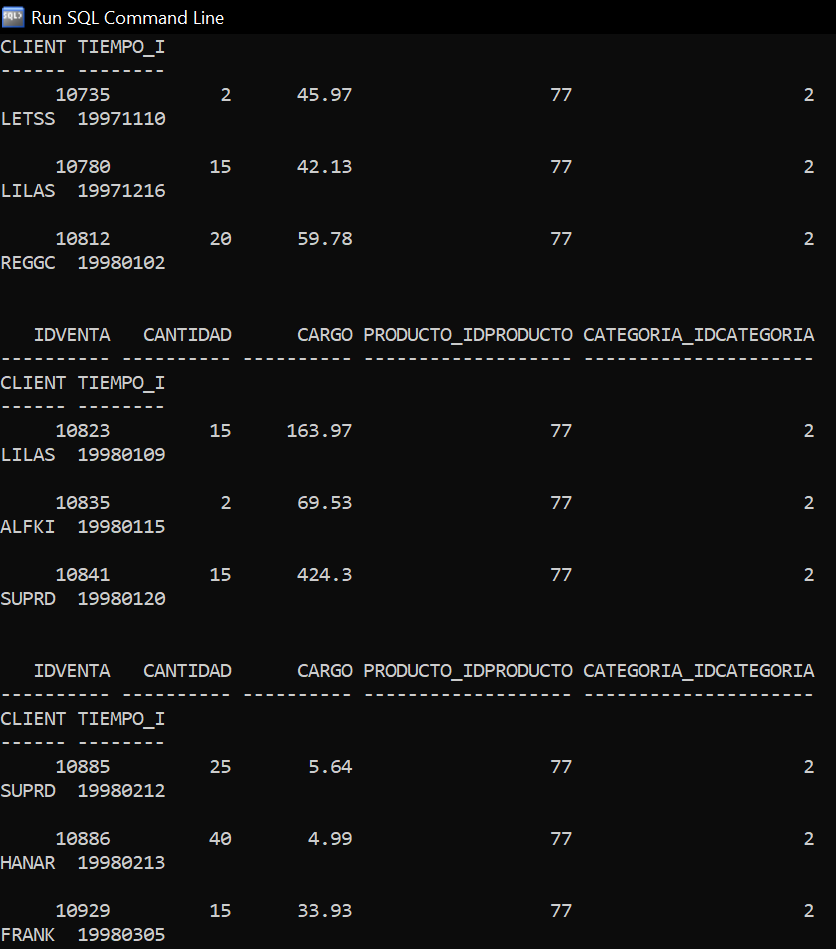
s) Los resultados de aplicar el proceso anterior deberán ser mostrados, así como el diagrama del flujo de pasos (como imágenes). Tome en consideración que este proceso se deberá realizar una sola vez, ya que se generarán errores de duplicados para varias ejecuciones. Verifique que el esquema multidimensional tenga los datos correctos e imprima capturas de pantalla de su contenido (en la tabla de hechos).



Escribimos el siguiente comando en SQL:



Y obtenemos, por poner un ejemplo de toda la impresión:



Vemos que al final de todo nuestro proceso son:



CUESTIONARIO

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

R:

Extracción: aquí los datos se extraen de las fuentes internas y externas que estén disponibles. La selección de los datos a importar, se basa en el diseño del almacén de datos, que de igual forma depende de la información que se necesite para el análisis de inteligencia empresarial y los DSS.

Transformación: el objetivo de ésta, que abarca limpieza y transformación, es mejorar la calidad de los datos que se han extraído de las diferentes fuentes, a través de la corrección de incoherencias, inexactitudes y valores faltantes. En la fase de limpieza, aplicamos reglas preestablecidas para modificar los errores, en la mayoría de los casos, los diccionarios con términos válidos se utilizan para sustituir los que están supuestamente incorrectos, en función del nivel de similitud. Durante la parte de transformación, se realizan conversiones de datos para garantizar la homogeneidad de estos, con el fin de obtener resúmenes que reducirán el tiempo de respuesta.

Cargar: por último, aquí es en donde se cargan los datos en las tablas de DW para que estén a disposición del analista y las aplicaciones de apoyo a las tomas de decisiones. Debemos considerar la carga de datos como escribiré a continuación: replicar las tablas de dimensiones, replicar la tabla de hechos, realizar una carga automática, indexar los datos para un mayor rendimiento.

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?

R:

Debe tener tabla de hechos (para las transacciones) y de tiempo (para revisar las diferentes fechas en las reglas del negocio.)

CONCLUSIONES

Aprendimos a utilizar los diferentes componentes de Oracle, Spoon, en donde creamos diagramas con el fin de visualizar las diferentes consultas que podemos hacer o tener conforme a las reglas del negocio. Mediante diferentes componentes de entrada de datos, ordenamiento, salida, merge join, calculator, fórmula, entre muchos más, se va creando el modelo ideal. No olvidemos que para crear estos modelos se necesita tener la visualización de estas tablas, ya que sin ellas no podríamos guiarnos para lo que lleguemos a necesitar.

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

**Anexo 1**

**Esquema de la base de datos multidimensional.**

Las siguientes son las sentencias SQL empleadas en Oracle para la creación de la estructura de la bodega de datos, incluyendo las restricciones de integridad referencial, para ser usadas en la práctica.

CREATE TABLE "CATEGORIA"

("IDCATEGORIA" NUMBER(2,0) NOT NULL,

"NOMBRECATEGORIA" VARCHAR2(50),

CONSTRAINT "CATEGORIA\_PK" PRIMARY KEY ("IDCATEGORIA")

);

CREATE TABLE "CLIENTE"

("IDCLIENTE" VARCHAR2(6) NOT NULL,

"NOMBRECONTACTO" VARCHAR2(25),

"CARGOCONTACTO" VARCHAR2(40),

"NOMBRECPIA" VARCHAR2(40),

"DIRECCION" VARCHAR2(60),

"CIUDAD" VARCHAR2(15),

"REGION" VARCHAR2(40),

"PAIS" VARCHAR2(15),

CONSTRAINT "CLIENTE\_PK" PRIMARY KEY ("IDCLIENTE")

);

CREATE TABLE "PRODUCTO"

("IDPRODUCTO" NUMBER(8,0) NOT NULL,

"NOMBREPRODUCTO" VARCHAR2(50),

"CANTIDADXUNIDAD" VARCHAR2(50),

"PRECIOXUNIDAD" NUMBER(10,2),

CONSTRAINT "PRODUCTO\_PK" PRIMARY KEY ("IDPRODUCTO")

);

CREATE TABLE "TIEMPO"

("IDTIEMPO" VARCHAR2(8) NOT NULL,

"ANIO" NUMBER(4,0),

"MES" NUMBER(2,0),

"DIA" NUMBER(2,0),

CONSTRAINT "TIEMPO\_PK" PRIMARY KEY ("IDTIEMPO")

);

CREATE TABLE "VENTAS"

("IDVENTA" NUMBER(8,0) NOT NULL,

"CANTIDAD" NUMBER(5,0),

"CARGO" NUMBER(10,2),

"PRODUCTO\_IDPRODUCTO" NUMBER(8,0),

"CATEGORIA\_IDCATEGORIA" NUMBER(2,0),

"CLIENTE\_IDCLIENTE" VARCHAR2(6),

"TIEMPO\_IDTIEMPO" VARCHAR2(8)

);

ALTER TABLE "VENTAS" ADD CONSTRAINT "VENTAS\_CATEGORIA\_FK" FOREIGN KEY ("CATEGORIA\_IDCATEGORIA") REFERENCES "CATEGORIA" ("IDCATEGORIA");

ALTER TABLE "VENTAS" ADD CONSTRAINT "VENTAS\_CLIENTE\_FK" FOREIGN KEY ("CLIENTE\_IDCLIENTE") REFERENCES "CLIENTE" ("IDCLIENTE");

ALTER TABLE "VENTAS" ADD CONSTRAINT "VENTAS\_PRODUCTO\_FK" FOREIGN KEY ("PRODUCTO\_IDPRODUCTO") REFERENCES "PRODUCTO" ("IDPRODUCTO");

ALTER TABLE "VENTAS" ADD CONSTRAINT "VENTAS\_TIEMPO\_FK" FOREIGN KEY ("TIEMPO\_IDTIEMPO") REFERENCES "TIEMPO" ("IDTIEMPO");

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.