

Existen dos formas de tomar decisiones:



Decisiones *lentas*, basadas sólo en la *experiencia*, prediciendo el futuro en base a la *intuición*.



Decisiones *rápidas y oportunas*, tomadas sobre *cimientos firmes*, prediciendo el futuro en base a *información* del pasado.

¿Usted cómo lo haría en su empresa?



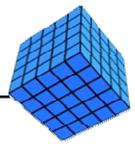
Microsoft
DATA WAREHOUSE

SISTEMAS
DE GESTIÓN II

DOCENTES: Ing. Bigatti, Cristian G.
Est. Grasso, Mónica

COLABORADORES: Alvarez, Ezequiel
Gorza, Luciano
Leala, Guillermo

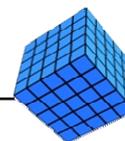




TUTORIAL DATA WAREHOUSE SQL Server: **ETL – Parte 1**

Este documento consta de:

- Descripción general del Tutorial
- Prerrequisitos
- Pasos generales a seguir
- Pasos específicos a seguir
- Anexo I – Instalación de Microsoft SQL Server
- Anexo II - Almacenamiento y recuperación

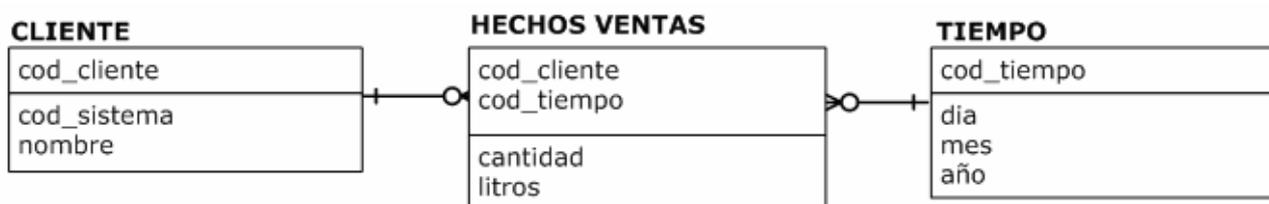


Descripción general del Tutorial

Nuestro objetivo es la construcción de un Data Warehouse para que TDC (The Drinking Company) pueda atender las siguientes cuestiones:

- o Cantidad de litros consumidos y de productos adquiridos por cliente en el tiempo.
- o Compra promedio en litros por cliente en el tiempo.

El esquema del Data Warehouse que responderá esos interrogantes es el siguiente:



El proceso de construcción del Data Warehouse será dividido en dos grandes etapas:

- o Procesos de extracción, transformación y carga (ETL).
- o Procesamiento analítico de datos (OLAP).

Este tutorial tiene como objetivo la creación y población de la Base de Datos del Data Warehouse, es decir, los procesos ETL.

Esta Base de Datos poseerá las siguientes tablas:

Cliente
Cod_cliente
Cod_sistema
Nombre

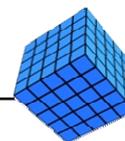
Ventas
Cod_cliente
Cod_tiempo
Cantidad
Litros

No es necesario crear la tabla Tiempo debido a que se generará en la etapa OLAP.

Para llegar a construir esta Base de Datos, los analistas determinaron que serán necesarias las siguientes Fuentes de Datos:

Products.txt (Productos)

CAMPO
PRODUCT_ID
DETAIL
PACKAGE



Customer_R.txt (Clientes Minoristas)

CAMPO
CUSTOMER_ID
FULL_NAME
BIRTH_DATE
CITY
STATE
ZIPCODE

Customer_W.txt (Clientes Mayoristas)

CAMPO
CUSTOMER_ID
FULL_NAME
BIRTH_DATE
CITY
STATE
ZIPCODE

Sales.mdb (Ventas)

Billing (Facturas)

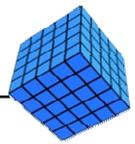
BILLING_ID (ENTERO LARGO)
CUSTOMER_ID (ENTERO LARGO) DATE (FECHA - HORA) EMPLOYEE_ID (ENTERO)



Billing_Detail (Detalle de Facturas)

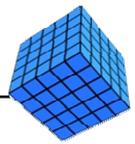
BILLING_ID (ENTERO LARGO) PRODUCT_ID (ENTERO LARGO)
QUANTITY (ENTERO LARGO)

La Base de Datos History_Sales no será utilizada en este Tutorial.



Prerrequisitos

- Microsoft SQL Server 2000: Enterprise Manager
- Fuentes de datos del Sistema de Gestión de Ventas de TDC:
 - ✓ Customer_W.txt (Clientes Mayoristas)
 - ✓ Customer_R.txt (Clientes Minoristas)
 - ✓ Sales.mdb (Ventas)
- Carpeta de trabajo definida, por ejemplo: C:\SGII2005. Usaremos este nombre de carpeta para el presente trabajo. En esta carpeta deberán estar copiadas las fuentes de datos.



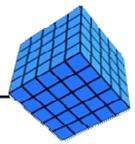
Pasos generales a seguir

Parte I

1. Crear la Base de Datos de la Staging Area.
2. Cargar los datos desde las Fuentes de Datos a la Staging Area.

Parte II

3. Crear la Base de Datos del Data Warehouse.
4. Realizar los procesos de Extracción, Transformación y Carga (ETL).



Pasos específicos a seguir

1. Crear la Base de Datos de la Staging Area.

Un Staging Area es un área de almacenamiento intermedio, que nos permitirá una mejor manipulación de los datos a la hora de realizar los procesos ETL subsiguientes.

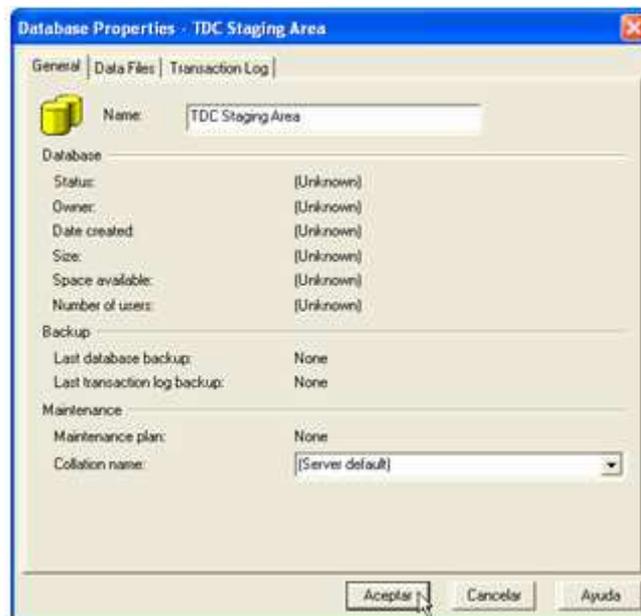
- a. Ingresaremos al Enterprise Manager y al área de trabajo.

Clic en **Inicio** → **Programas** → **Microsoft SQL Server** → **Enterprise Manager**. En el árbol de carpetas de la Consola, doble clic en **Microsoft SQL Servers**. Doble clic en **SQL Server Group** para ver la lista de servidores. Seleccionaremos el servidor local, que en nuestro caso es **localhost** o el nombre de la PC local.

Doble clic en el servidor seleccionado para establecer una conexión con SQL Server. Luego de que la conexión se ha establecido, aparecerá una lista de carpetas.

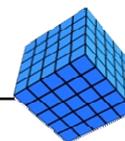
- b. Crearemos la Base de datos de la Staging Area y le daremos un nombre.

En **Enterprise Manager**, clic con el botón derecho en **Databases** → **New Database....** Escribir el nombre de la base de datos "**TDC Staging Area**". Clic en **Aceptar**.

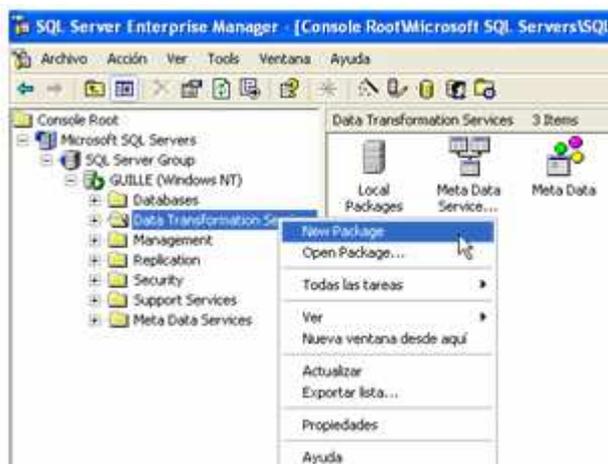


2. Cargar los datos desde las Fuentes de Datos a la Staging Area.

- a. Ingresaremos en el DTS Package Designer.



Clic con el botón derecho en **Data Transformation Services** → **New Package**.

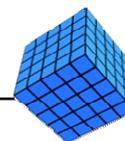


- b. Crearemos una conexión para las fuentes de datos del Sistema de Gestión de Ventas que sean archivos de texto: **Products.txt**, **Customer_R.txt** y **Customer_W.txt**.

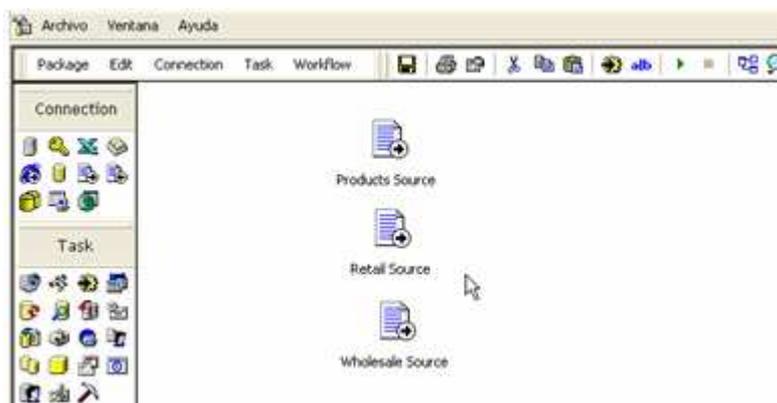
Para **Products.txt**: En **DTS Package Designer** → Menú **Connection** → **Text File (Source)**. En el cuadro de texto **New connection** vamos a escribir un nombre para la conexión: "Products Source". Clic en el botón situado a la derecha del cuadro de texto **File Name ([...])** y seleccionamos el archivo fuente: **C:\SGII2005\Fuentes de datos\Stock\Products.txt**. Una vez seleccionado el archivo y cerrado el cuadro de diálogo de selección, clic en **Properties...**, clic en **Siguiente >**, verificar que la opción de delimitador sea **Other**, y en el cuadro de texto aparezca el símbolo pipe (|), clic en **Finalizar**. Clic en **Aceptar**.

Para **Customer_R.txt**: En **DTS Package Designer** → Menú **Connection** → **Text File (Source)**. En el cuadro de texto **New connection** vamos a escribir un nombre para la conexión: "Retail Source". Clic en el botón situado a la derecha del cuadro de texto **File Name ([...])** y seleccionamos el archivo fuente: **C:\SGII2005\Fuentes de datos\Ventas\Customer_R.txt**. Una vez seleccionado el archivo y cerrado el cuadro de diálogo de selección, clic en **Properties...**, tildar la opción **First row has column names**, clic en **Siguiente >**, verificar que la opción de delimitador sea **Tab**, clic en **Finalizar**. Clic en **Aceptar**.

Para **Customer_W.txt**: En **DTS Package Designer** → Menú **Connection** → **Text File (Source)**. En el cuadro de texto **New**



connection vamos a escribir un nombre para la conexión: "Wholesale Source". Clic en el botón situado a la derecha del cuadro de texto **File name ([...])** y seleccionamos el archivo fuente: **C:\SGII2005\Fuentes de datos\Ventas\Customer_W.txt**. Una vez seleccionado el archivo y cerrado el cuadro de diálogo de selección, clic en **Properties...**, tildar la opción **First row has column names**, clic en **Siguiente >**, verificar que la opción de delimitador sea **Tab**, clic en **Finalizar**. Clic en **Aceptar**.



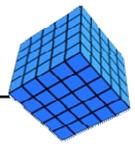
- c. Crearemos un archivo UDL que nos servirá de vínculo de datos.

*Un archivo de vínculos de datos se utiliza para crear y administrar mediante una interfaz OLE DB las conexiones entre el equipo cliente y los datos almacenados en un origen de información. Se trata de un archivo de texto cuya extensión es *.UDL (Universal Data Link) y que se puede tener como referencia para futuras conexiones a bases de datos. El archivo generalmente tiene la extensión UDL, aunque puede tener la extensión que deseemos, pero en éste último caso, no se abrirá el cuadro de diálogo Propiedades de vínculo de datos cuando se haga doble clic sobre el archivo en el Explorador de Windows.*

Si se usan conexiones a vínculos de datos (data link connections) es posible mantener la información de la conexión para un origen de datos OLE DB en un archivo externo UDL, y luego leer esta información dentro de la conexión a vínculos de datos cada vez que se ejecuta el proceso.

Implementar un vínculo de datos implica:

- ✓ Crear un archivo UDL
- ✓ Almacenar la información de la conexión en el archivo UDL
- ✓ Crear una conexión a vínculos de datos que use el archivo UDL



En el **Explorador de Windows** → carpeta **C:\SGII2005\Fuentes de datos\Ventas**, botón derecho → **Nuevo** → **Documento de texto**. Renombrar "Nuevo Documento de texto.txt" a **Sales Source.udl**.

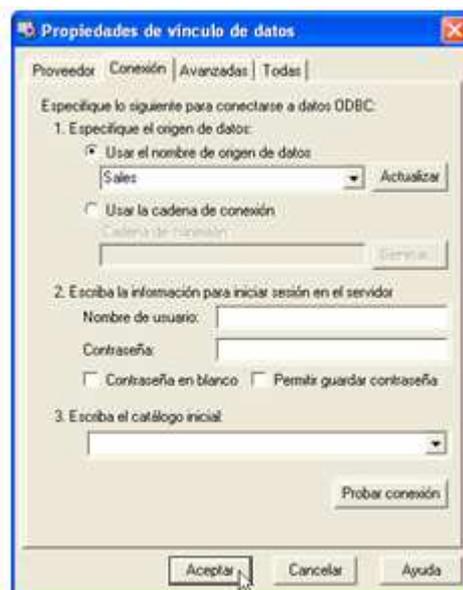
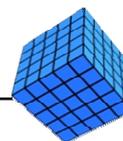
Importante: antes de renombrar el documento de texto y cambiar la extensión del archivo, verificar que el checkbox **Ocultar las extensiones de archivo para tipos de archivo conocidos** esté destildado. Esto lo hacemos en el **Explorador de Windows**, **Menú Herramientas** → **Opciones de carpeta...** → solapa **Ver**.

d. Configuraremos el archivo **Sales Source.udl**.

Para ello deberemos definir primero un origen de datos mediante ODBC. Con este fin, abriremos el **Administrador de orígenes de datos ODBC**, accesible desde el **Panel de Control**, o desde la carpeta **Herramientas Administrativas** dentro del **Panel de Control**.

Dentro del **Administrador de orígenes de datos ODBC**, solapa **DSN de Sistema (System DSN)**, botón **Agregar...** → **Microsoft Access Driver (*.mdb)**, botón **Finalizar**. En el cuadro de texto **Nombre del origen de datos**, colocar "Sales". Botón **Seleccionar**, seleccionamos el archivo **C:\SGII2005\Fuentes de datos\Ventas\Sales.mdb**. Clic en **Aceptar**, clic en **Aceptar**.

Ya tenemos el origen de datos ODBC. Ahora vamos a trabajar con el archivo UDL. En el **Explorador de Windows**, situados en la carpeta **C:\SGII2005\Fuentes de datos\Ventas**, doble clic en **Sales Source.udl**. En la solapa **Proveedor**, seleccionar **Microsoft OLE DB Provider for ODBC Drivers**. En la solapa **Conexión**, seleccionar la opción **Usar el nombre de origen de datos**, en el menú desplegable seleccionar **Sales**, botón **Probar conexión**. Con los pasos realizados correctamente, la prueba de conexión resultará satisfactoria. Clic en **Aceptar**, clic en **Aceptar**.

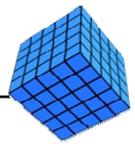


- e. Crearemos una conexión para la fuente de datos restante del Sistema de Gestión de Ventas de TDC, que es un archivo de Microsoft Access: **Sales.mdb**. Vamos a crear una conexión para cada tabla.

Para la tabla **Billing**: En **DTS Package Designer** → Menú **Connection** → **Microsoft Data Link**. En el cuadro de texto **New connection** vamos a escribir un nombre para la conexión: "**Sales Source - Billing**".

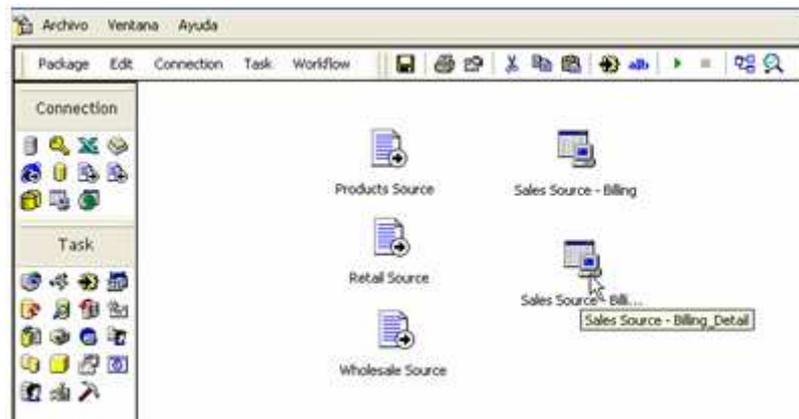


Clic en el botón situado a la derecha del cuadro de texto **UDL file name** ([...]) y seleccionamos el archivo fuente: **C:\SGII2005\Fuentes de datos\Ventas\Sales Source.udl**. Una vez seleccionado el archivo y



cerrado el cuadro de diálogo de selección, tildar el checkbox **Always read properties from UDL file**. Con esto, cualquier cambio en el archivo UDL podrá leerse dentro de la conexión. Clic en **Aceptar**.

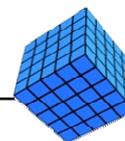
Procederemos de forma análoga con la tabla **Billing_Detail**. El nombre de la conexión será **"Sales source - Billing_Detail"**



- f. Crearemos una conexión para el destino de los datos, que en nuestro caso es el Staging Area.

En **DTS Package Designer** → Menú **Connection** → **Microsoft OLE DB Provider for SQL Server...**. En el cuadro de texto **New connection** vamos a escribir un nombre para la conexión: **"TDC Staging"**. En el menú desplegable **Database**, seleccionar la base de datos que hemos creado en el punto 1., **TDC Staging Area**. Clic en **Aceptar**.

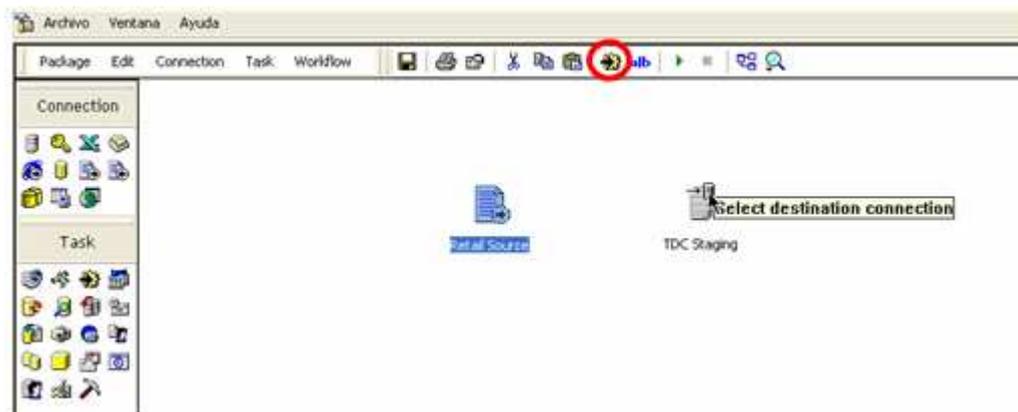




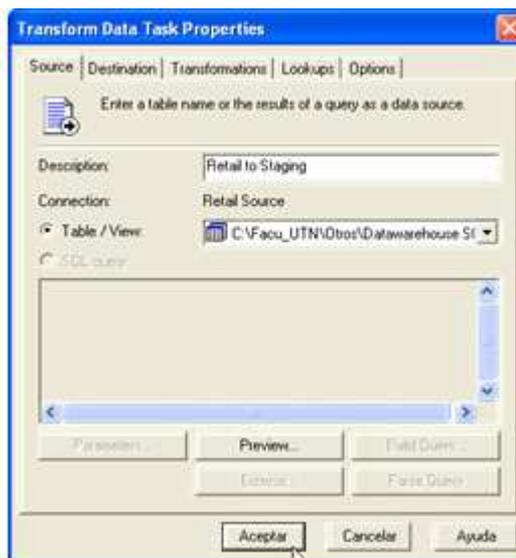
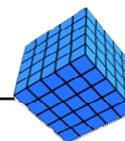
- g. Uniremos a las conexiones fuente con la conexión destino mediante transformaciones simples, de tipo Copy column.

Una transformación Copy column consiste en encontrar las columnas de las fuentes de datos y mapearlas a columnas en el destino de los datos. Un ejemplo es tomar un archivo separado por comas (.csv) como fuente de datos, y especificar los separadores de campo (coma) y de registro (retorno de carro). Suponiendo que nuestro destino de datos es una base de datos de SQL Server, el proceso Copy column consistirá en mapear cada renglón del archivo *.csv a un registro en una tabla de la base de datos de SQL Server.*

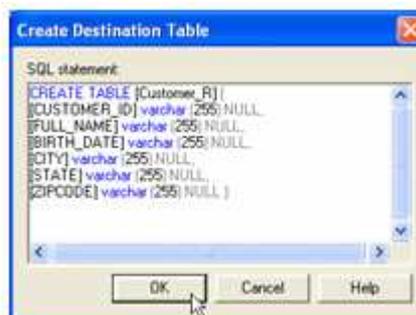
Transformación **Retail Source** a **TDC Staging**: En **DTS Package Designer** → Clic en **Retail source**, clic en el botón **Transform Data Task** (representado en la barra de herramientas superior por una flecha ingresando a un engranaje, señalado con rojo en la figura siguiente) luego clic en **TDC Staging**



Doble clic sobre la flecha gris que aparece luego de ejecutar el paso anterior. Aparece el cuadro de diálogo Transform Data Task Properties:

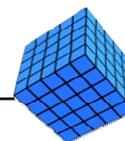


En la solapa **Source**, en el cuadro de texto **Description** escribiremos el nombre de la transformación, en nuestro caso **"Retail to Staging"**. Clic en la solapa **Destination**. En el caso de que no existan tablas en la Base de Datos, aparece el cuadro de diálogo **Create Destination Table**. Caso contrario accedemos a él mediante clic en **Create...**



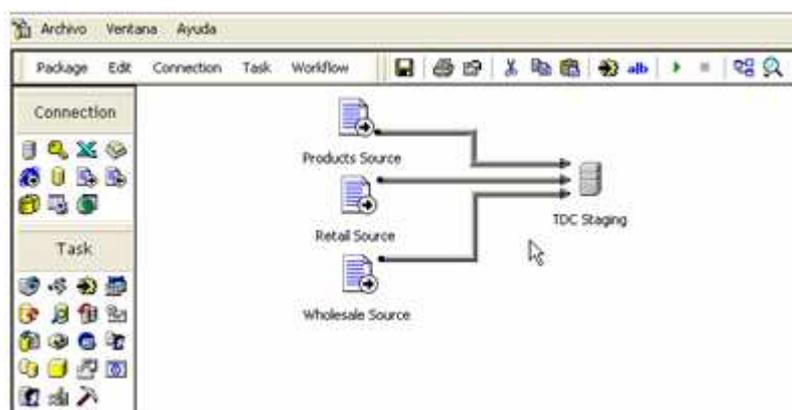
Dejamos la estructura que nos propone SQL Server. Clic en **OK**. Clic en la solapa **Transformations**. Lo que muestra el cuadro de diálogo es lo siguiente: a la izquierda, los atributos de la fuente de datos; a la derecha, los atributos del destino de datos. Cada flecha indica una transformación desde uno o más atributos de origen hacia uno o más atributos de destino. Por defecto, estas flechas indican un mapeo o correspondencia uno a uno entre los atributos fuente y destino (transformación Copy column). Clic en **Aceptar**.

Procederemos análogamente con la transformación **Wholesale Source** a **TDC Staging**. El nombre de la transformación será **"Wholesale to**

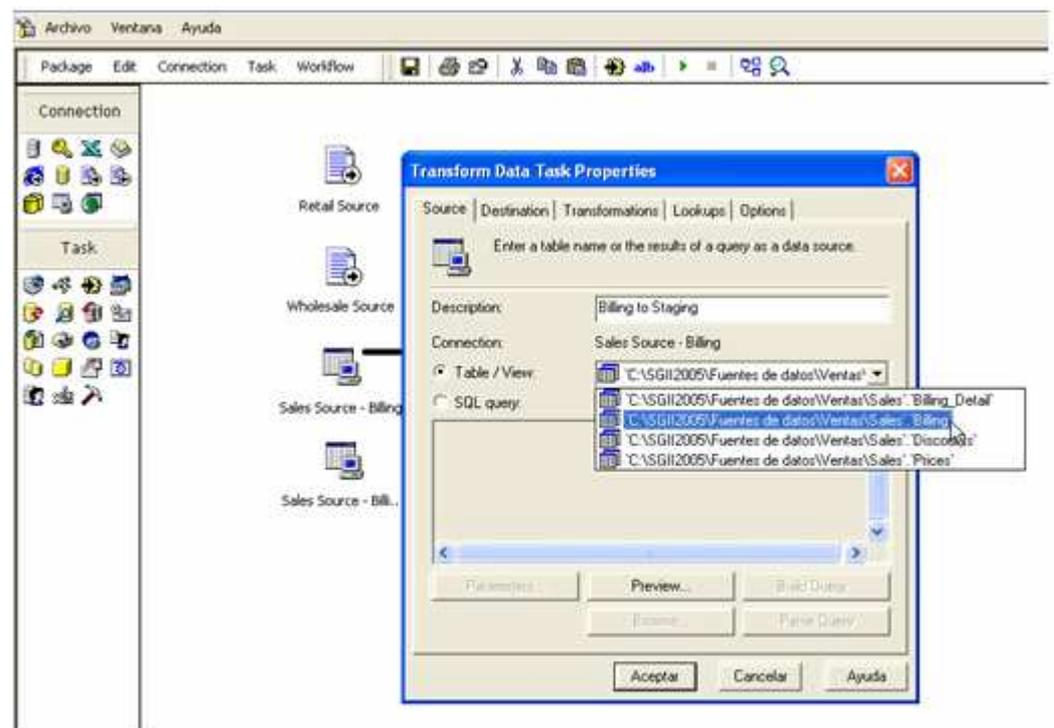
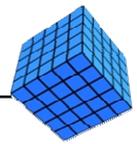


Staging". No olvidar presionar **Create...** en la solapa **Destination** para crear una tabla nueva. No usemos la que nos propone SQL Server.

Para la transformación **Products Source** a **TDC Staging** el procedimiento es el mismo, usando el nombre "**Products to Staging**". La diferencia aquí es que la fuente de datos no posee una fila de encabezado, por lo tanto cuando presionemos **Create...** debemos renombrar las columnas que nos propone SQL Server a sus respectivos nombres (**PRODUCT_ID, DETAIL, PACKAGE**).

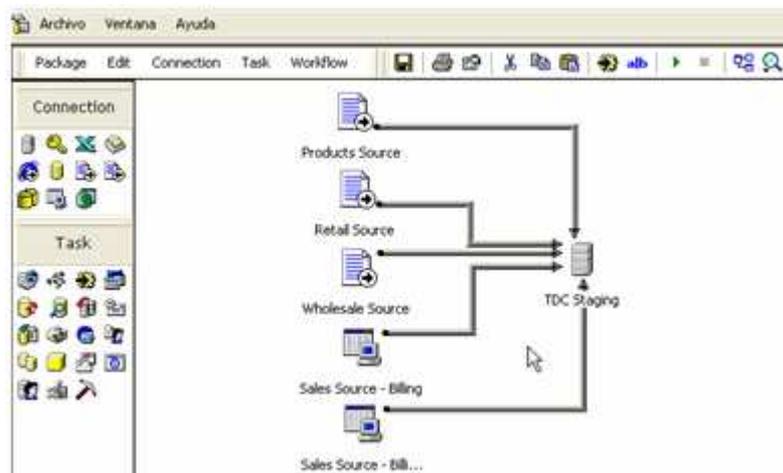


Transformación **Sales Source – Billing** a **TDC Staging**: En **DTS Package Designer** → Clic en **Sales Source – Billing**, clic en el botón **Transform Data Task**, luego clic en **TDC Staging**. Doble clic sobre la flecha gris que une las dos conexiones. En la solapa **Source**, en el cuadro de texto **Description**, escribimos "**Billing to Staging**". En el menú desplegable a la derecha de la opción **Table/View**, seleccionamos ``C:\SGII2005\Fuentes de datos\Ventas\Sales`.`Billing``.

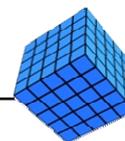


Clic en la solapa **Destination**. Clic en **Create...**. Clic en **OK**. Clic en la solapa **Transformations**, Clic en **Aceptar**.

Procederemos de forma análoga con la transformación **Sales Source – Billing_Detail** a **TDC Staging**. El nombre de la transformación correspondiente será "**Billing_Detail to Staging**". No olvidar presionar **Create...** en la solapa **Destination** para crear una tabla nueva. No usemos la que nos propone SQL Server.



h. Por último, vamos a crear un **Execute SQL Task**, o tarea de ejecución SQL. Vamos a utilizarla para borrar los datos que el proceso almacenará



en la Base de Datos **TDC Staging Area**, de tal forma que cada vez que este proceso se corra, se haga una actualización completa (eliminación y reemplazo de los datos) de dicha Base de Datos, a partir de las Fuentes de Datos.

Un Execute SQL Task es una tarea que permite ejecutar una sentencia SQL sobre una conexión.

En **DTS Package Designer**, menú **Task** → **Execute SQL Task....** En el cuadro de texto **Description**, escribimos "**Delete Staging**". En el menú desplegable **Existing connection** seleccionamos **TDC Staging**. En el cuadro de texto **SQL Statement** escribimos la siguiente sentencia SQL:

DELETE FROM Products;

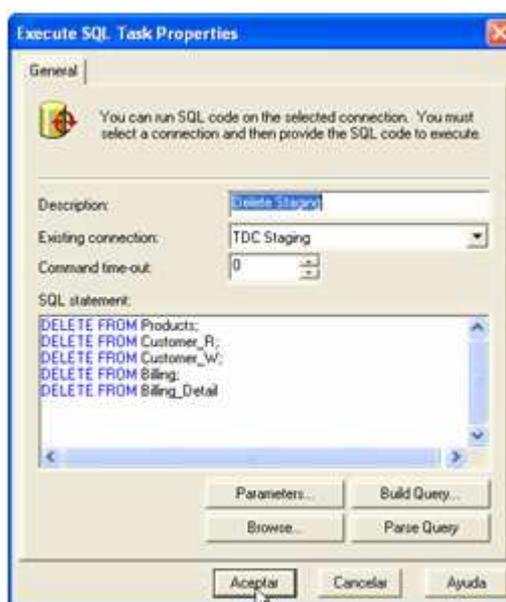
DELETE FROM Customer_R;

DELETE FROM Customer_W;

DELETE FROM Billing;

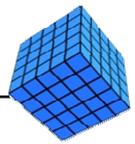
DELETE FROM Billing_Detail

Clic en **Parse Query** para verificar la sintaxis de la sentencia. Si es correcta se mostrará un mensaje de éxito. Clic en **OK**, clic en Aceptar.



Uniremos esta tarea con el resto del proceso mediante condiciones de éxito.

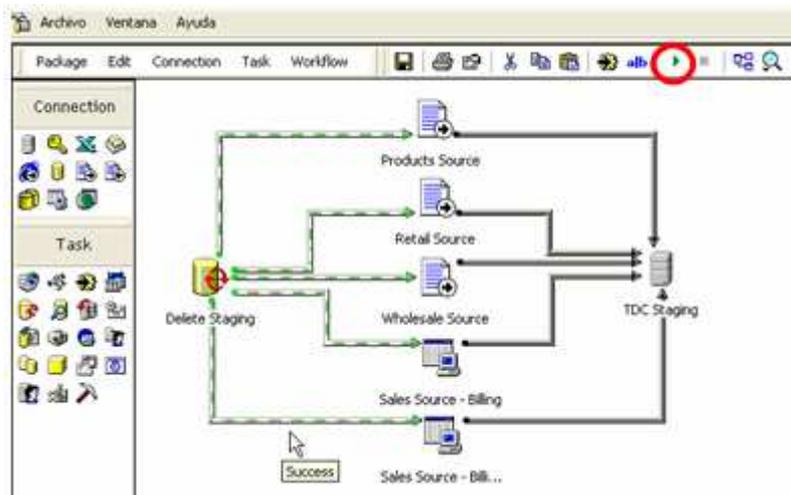
Una condición de éxito (On Success) se representa por una flecha verde y blanca, y valida que una parte del proceso (la cola de la flecha) se ha



llevado a cabo satisfactoriamente antes de ejecutar el resto (la punta de la flecha). En caso contrario la ejecución se detiene.

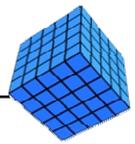
Clic en **Delete Staging**, ctrl+clic en **Retail Source**, de tal forma que queden los dos elementos seleccionados. Clic en el menú **Workflow** → **On Success**.

Análogamente con **Delete Staging** y **Wholesale Source**; **Delete Staging** y **Sales Source - Billing**; **Delete Staging** y **Sales Source - Billing_Detail**.



- i. Correremos el proceso. Clic en el botón **Play** (representado en la barra de herramientas superior por un triángulo verde apuntando hacia la derecha, señalado con rojo en la figura anterior).

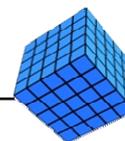




Si el proceso está correctamente armado, cada una de sus partes se ejecutará en forma satisfactoria. Clic en **Aceptar**, clic en **Done**



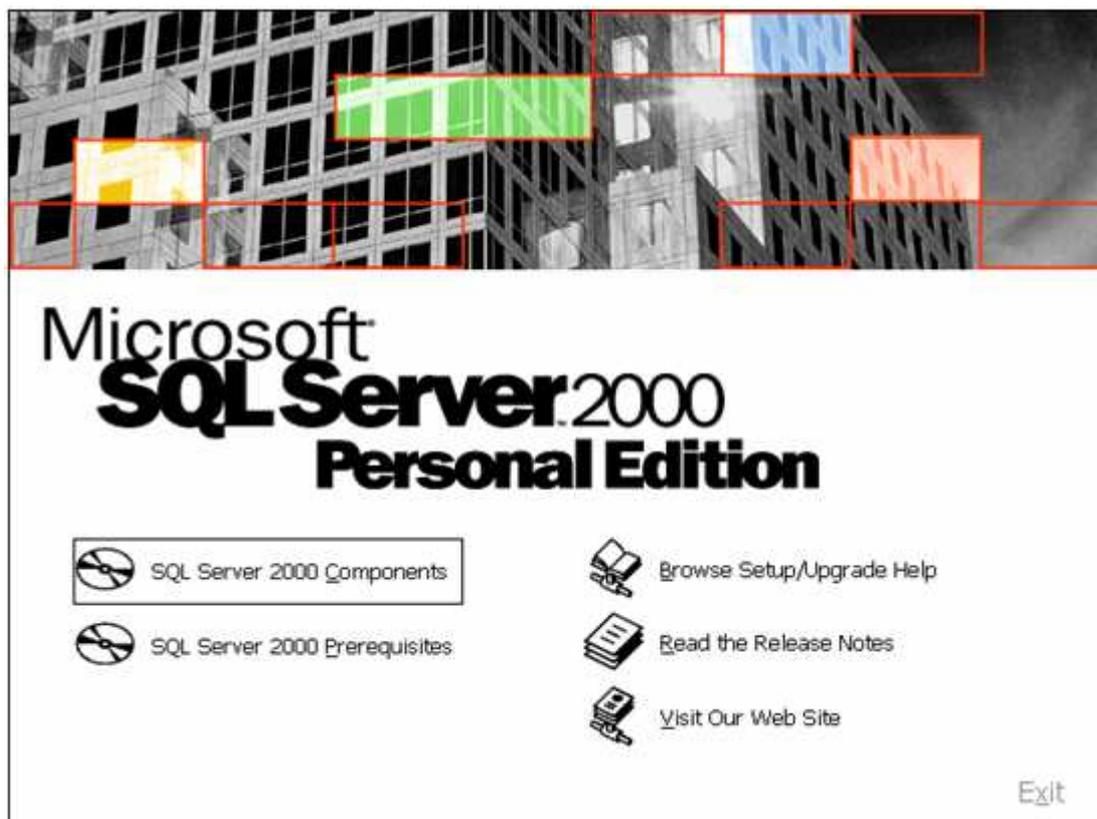
En este momento tenemos las tablas de la Base de Datos de la Staging Area cargadas con los datos operacionales.



Anexo I - Instalación de Microsoft SQL Server

1. Instalar SQL Server

Si la computadora reconoce archivos de autoarranque, se mostrará la **Pantalla de Inicio** al insertar el CD, caso contrario ejecutar **Autorun.exe** desde la carpeta raíz del CD



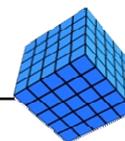
Clic en **SQL Server 2000 Components**, clic en **Install Database Server**. Seguir las opciones por defecto, salvo la siguiente: cuando aparezca la pantalla que incluye la **selección del modo de autenticación (authentication mode)**, seleccionar **Use Windows Authentication**.

2. Instalar Analysis Services

En la **Pantalla de Inicio**, clic en **SQL Server 2000 Components**, clic en **Install Análisis Services**. Seguir las opciones por defecto.

3. Instalar Service Pack 1

En la carpeta **service pack 1**, ejecutar **setup.bat**. Seguir las opciones por defecto.



Anexo II - Almacenamiento y recuperación

- Grabar un DTS en el servidor local

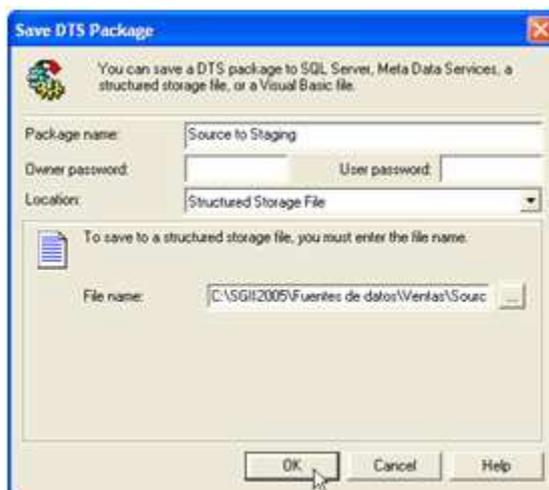
Una vez creado un proceso en **DTS Package Designer**, clic en menú **Package** → **Save As....** Aparece la pantalla **Save DTS Package**

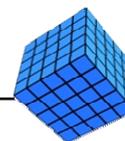


En el cuadro de texto **Package name** escribimos el nombre del proceso, en este caso **“Source to Staging”**. En el menú desplegable **Location**, seleccionar **SQL Server**. Clic en **OK**.

- Grabar un DTS en un archivo *.dts

Una vez creado un proceso en **DTS Package Designer**, clic en menú **Package** → **Save As....** Aparece la pantalla **Save DTS Package**

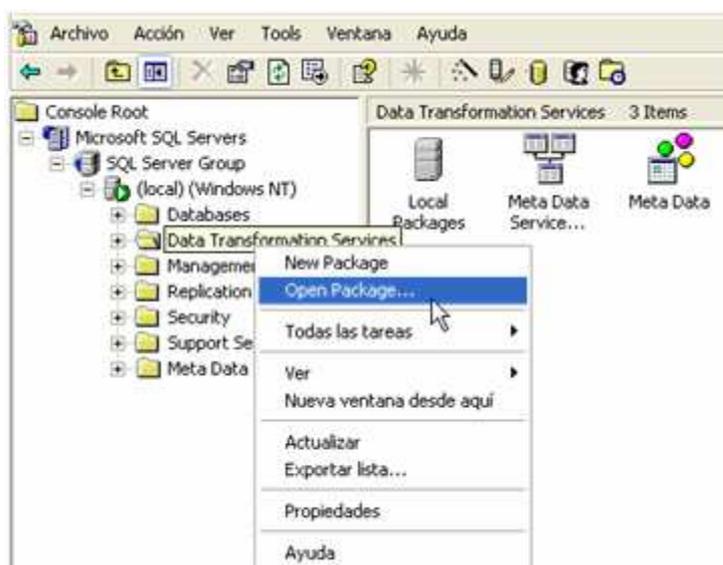




En el cuadro de texto **Package name** escribimos el nombre del proceso, en este caso **"Source to Staging"**. En el menú desplegable **Location**, seleccionar **Structured Storage File**. En el cuadro de texto **File Name**, seleccionar la carpeta de almacenamiento, en este caso **C:\SGII2005\Fuentes de datos\Ventas\Source to Staging.dts**. Clic en **OK**.

- o Recuperar un archivo *.dts

En **Enterprise Manager**, clic con el botón derecho en **Data Transformation Services**, clic en **Open package....**



Aparecerá un cuadro de diálogo para seleccionar el archivo *.dts que deseamos recuperar. Seleccionar el archivo correspondiente. Clic en **Abrir**.