**TUTORIAL DE PROCESO ETL SIMPLE**

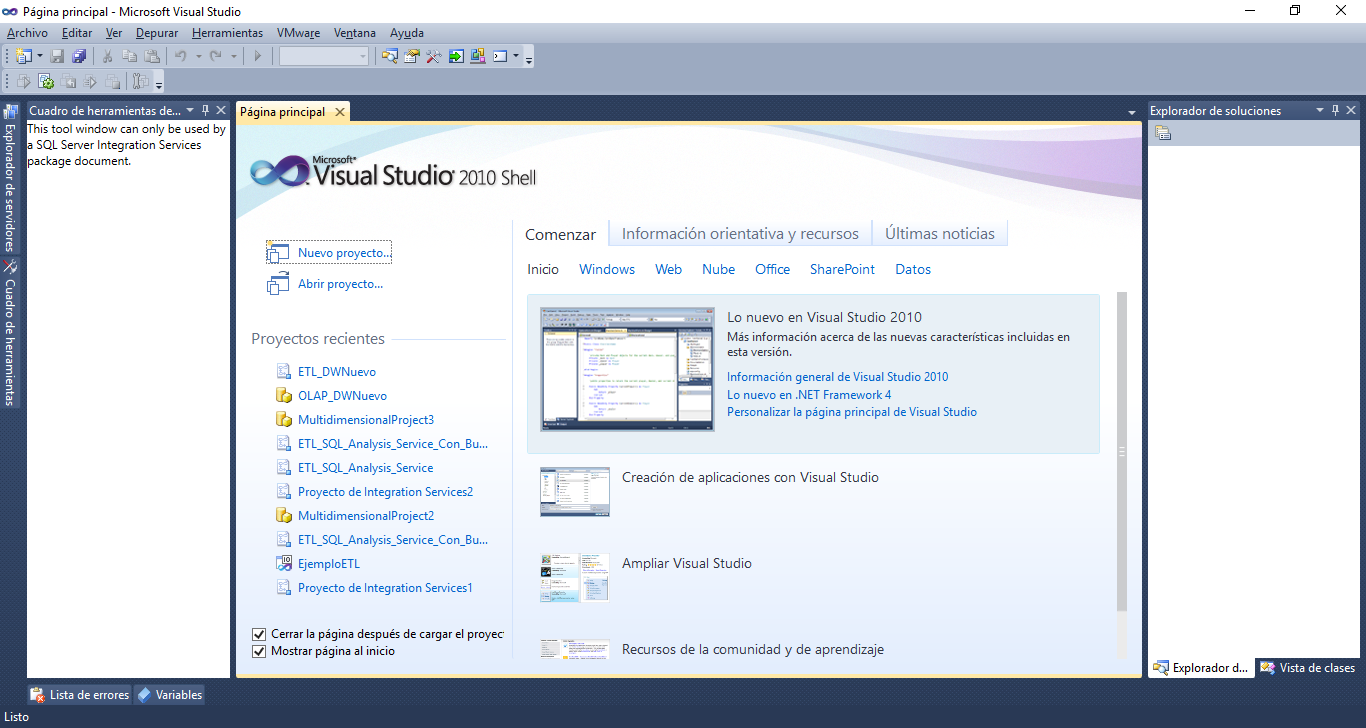
# INTRODUCCIÓN

En este tutorial desarrollaremos un proceso ETL cuyo propósito es extraer datos desde una base de datos de un sistema OLTP y cargarlos en un Data Warehouse. Para llevar a cabo el desarrollo, utilizaremos como herramienta SQL Data Tools que viene con la versión SQL Server 2012.

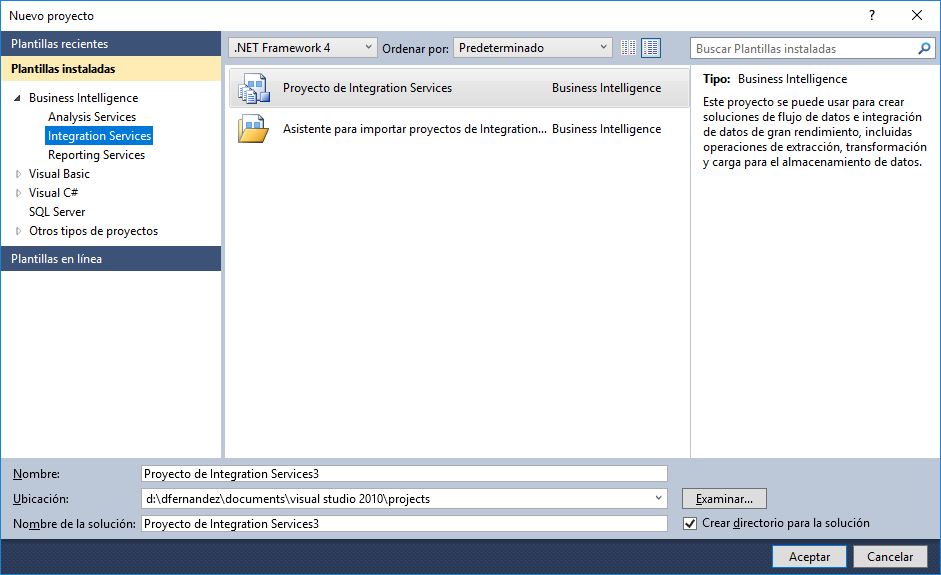
Como estrategia utilizaremos la de cargar todo cada vez que se ejecuta el proceso ETL. Elegimos esta estrategia dado que es la más sencilla con el propósito de que se conozca la herramienta y un proceso sencillo. Luego, pueden optar por utilizar estrategias más complejas, como la actualización incremental.

# PREPARACIÓN DEL AMBIENTE

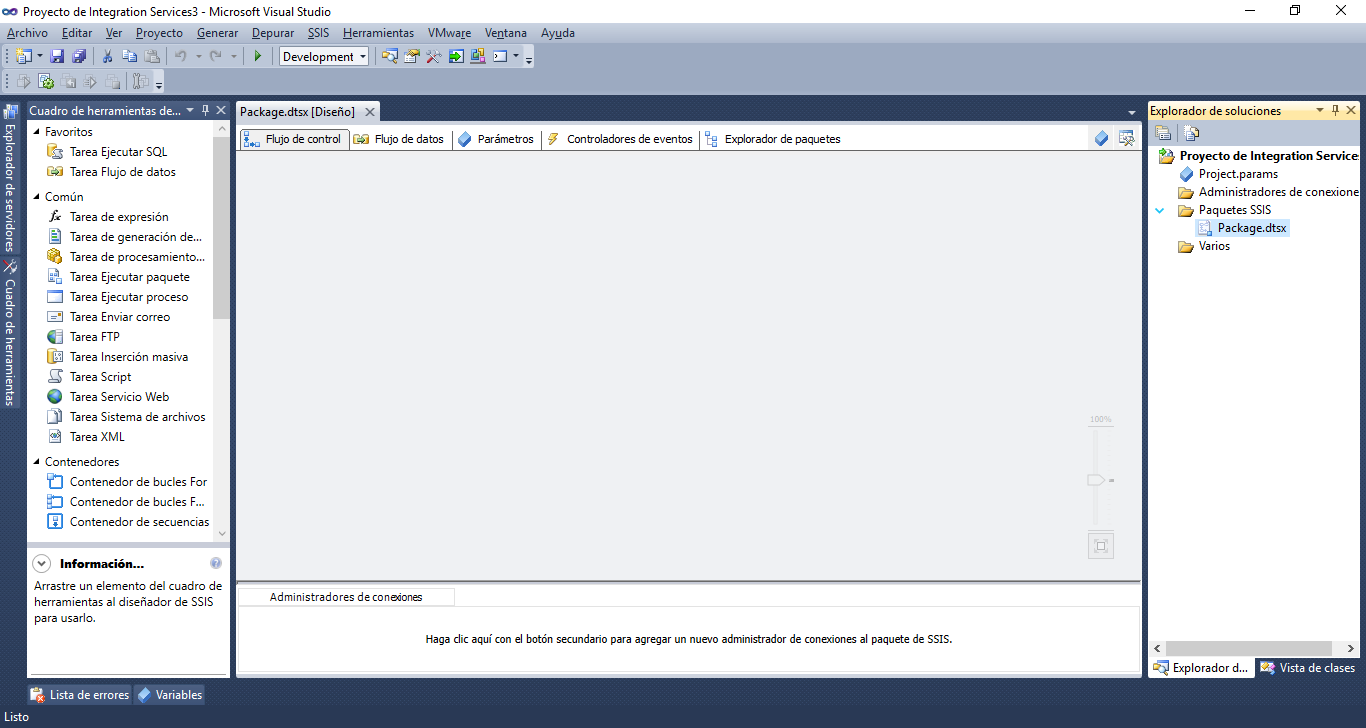
1. Para comenzar es necesario crear dos bases de datos. Una que hará de fuente de datos (sistema OLTP) y otra que servirá para nuestro Data Warehouse.
   1. Ingresar a SQL Server Management Studio.
   2. Ejecutar (en este orden) los siguientes scripts: SISTEMAOLTP\_Estructura.sql, SISTEMAOLTP\_Datos.sql y SISTEMAOLAP\_Estructura.sql
   3. Los scripts crearan las dos bases de datos necesarias (SistemaOLTP y SistemaOLAP), y completará la base de datos SistemaOLTP con datos.
2. Una vez que las bases de datos están creadas, acceder a SQL Data Tools, donde verán una pantalla como la siguiente:



1. Ir a “Nuevo proyecto…”, de las planillas seleccionar “Integration Services” y “Proyecto de Integration Services” y asignar un nombre al proyecto.



1. Si todo se realizó correctamente, deberían ver la siguiente pantalla:



1. Por último, hay que crear una conexión a cada base de datos, o sea, una conexión para la fuente de datos (SistemaOLTP) y otra para el destino de los datos (SistemaOLAP):
   1. En el “administrador de conexiónes” (en la parte de debajo de la pantalla), hacer click derecho y seleccionar “Nueva conexión OLE DB…”



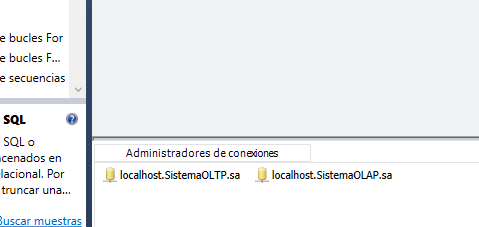
* 1. Se abrirá una nueva pantalla, allí click en el botón “Nueva…”.



* 1. Otra pantalla aparecerá, y allí podremos configurar la conexión con la base de datos “SistemaOLTP” complentando el “Nombre del servidor” con los mismos datos que utilizamos para conectarnos con el “SQL Server Management Studio”. Una vez completo el “Nombre del servidor”, se listarán las bases de datos disponibles. Una vez configurada, click en “Aceptar”.



* 1. Realizar el mismo procedimiento para conectarse a la base de datos “SistemaOLAP”. Al finalizar deberán verse ambas conexiones de la siguiente manera (la denomincación “localhost” varía según el “Nombre del servidor” que hayan ingresado):



1. Una vez finalizado estos pasos, ya tenemos todo el entorno listo para poder comenzar a desarrollar el proceso ETL.

# ELIMINACIÓN DE DATOS EN DATA WAREHOUSE

1. En primer lugar, debemos eliminar los datos del Data Warehouse para luego volver a cargarlos. Para ellos, en la pestaña “Flujo de Control”, arrastrar el componente “Tarea Ejecutar SQL”. Renombrarla como “Eliminar Datos DW” y dentro de sus propiedades generales, seleccionar como valor de “Connection” a la base OLAP y cargar como valor de “SQLStatement”:

truncate table FactVentaOrden;

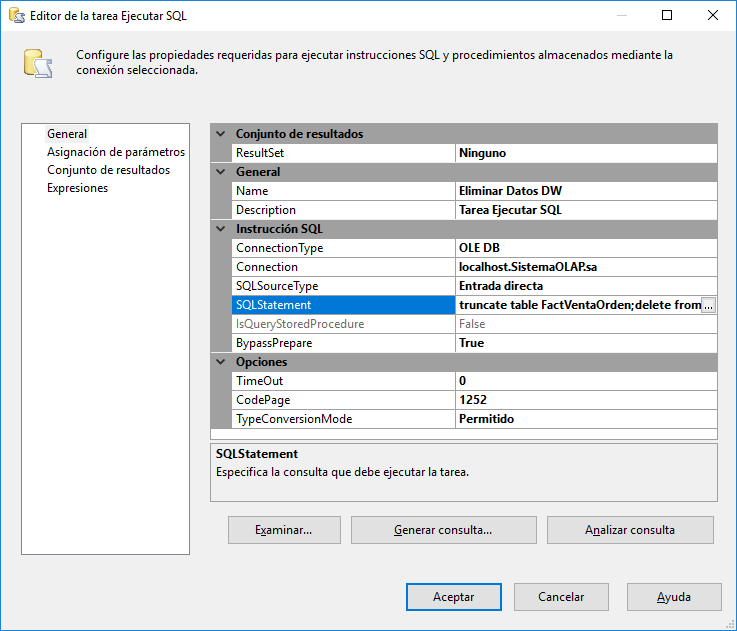
delete from DimCliente;

delete from DimEmpleado;

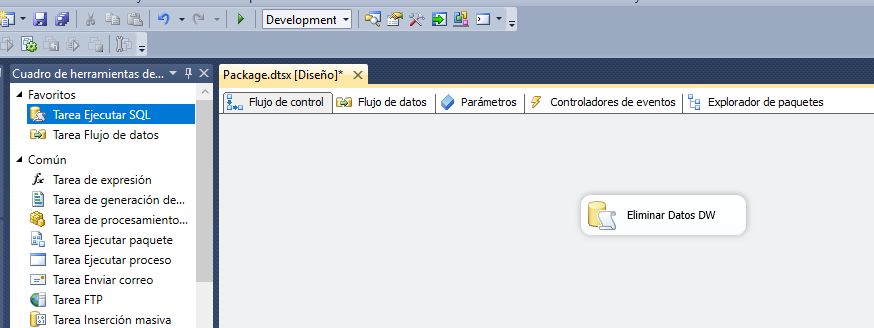
delete from DimProducto;

delete from DimTiempo;

delete from DimUbigeo;



1. Click en el botón “Aceptar” y quedará configurada la primera tarea del proceso ETL.



# CARGA DE DIMENSIÓN CLIENTE, EMPLEADO, PRODUCTO Y UBICACIÓN GEOGRÁFICA

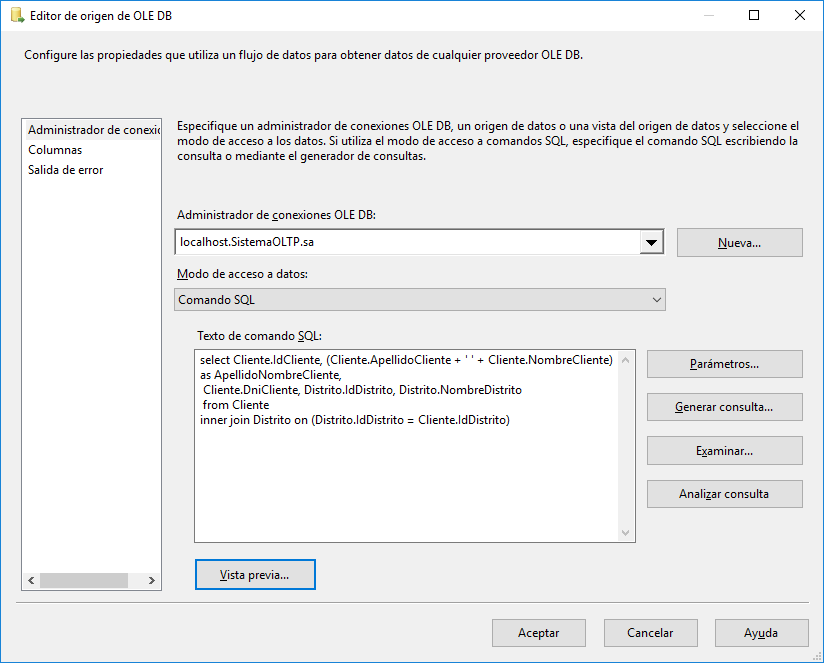
1. En la pestaña “Flujo de Control”, arrastrar el componente “Tarea Flujo de datos”.  
   Renombrarla como “Carga DimCliente”.
2. Hacer doble click sobre “Carga DimCliente”, nos llevará al Flujo de datos de esa tarea.
   1. Allí dentro arrastrar el componente “Origen de OLEDB” y doble click sobre él para editarlo.
   2. Como administrador de conexiones seleccionar la conexión hacia la base OLTP (SistemaOLTP). Como “Modo de acceso a datos” seleccionar “Comando SQL” y en “Texto de comando SQL” ingresar:

select Cliente.IdCliente, (Cliente.ApellidoCliente + ' ' + Cliente.NombreCliente) as ApellidoNombreCliente,

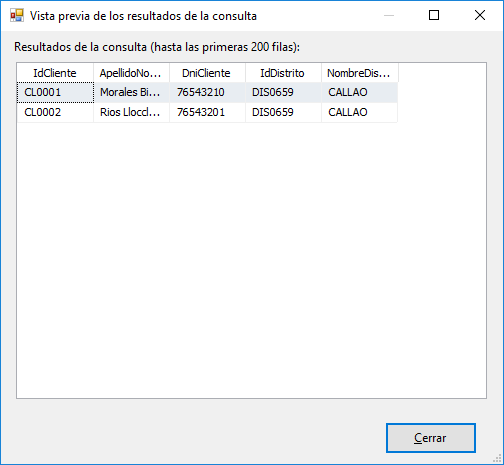
Cliente.DniCliente, Distrito.IdDistrito, Distrito.NombreDistrito

from Cliente

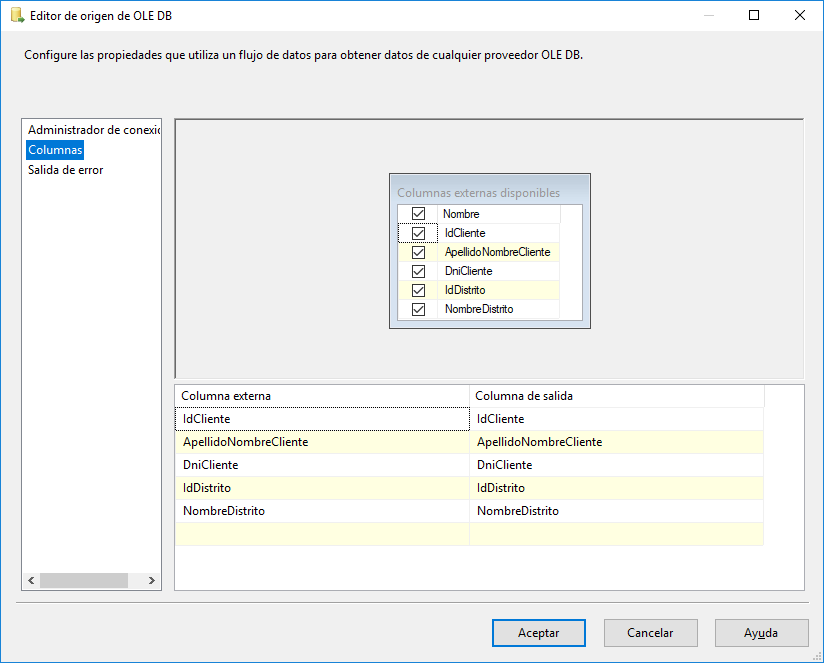
inner join Distrito on (Distrito.IdDistrito = Cliente.IdDistrito)



* 1. Para validar que todo esté funcionando, hacer click en el botón “Vista previa…”, deberíamos ver allí el resultado de la consulta.



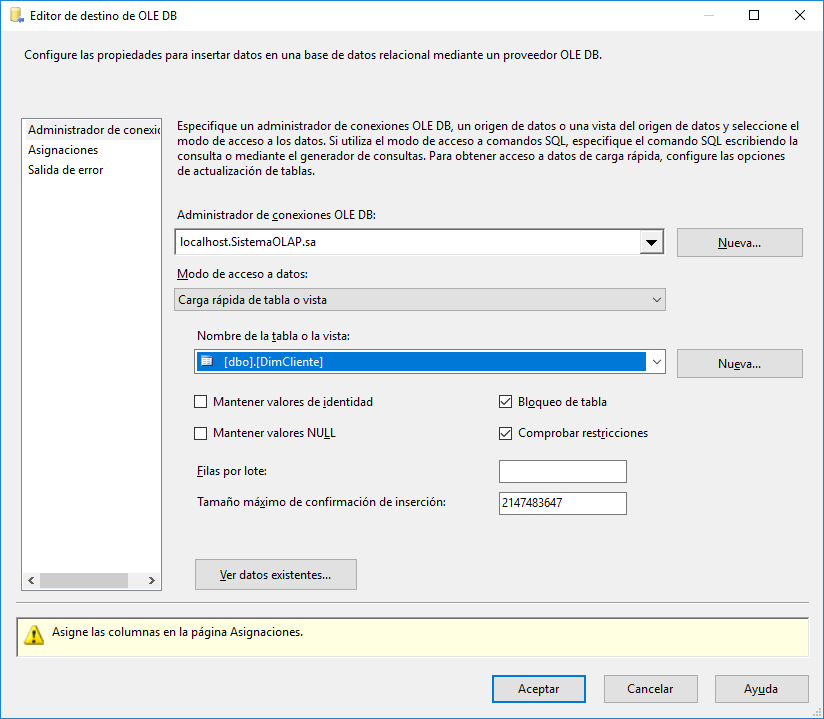
* 1. En la sección “Columnas” tildar todas. Las columnas que tildamos son la continuaran existiendo para otras tareas del proceso.



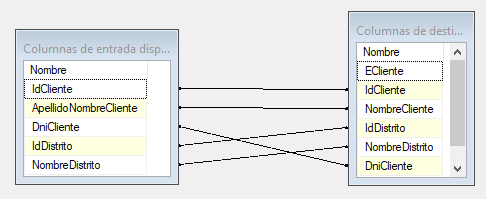
* 1. Click en el botón “Aceptar” para volver al “Flujo de datos”. Con esto hemos configurado la parte del proceso que se encarga de extraer los datos desde el origen de datos “SistemaOLTP”, particularmente los datos necesarios para cargar la dimensión cliente (“DimCliente”).
  2. Arrastrar un nuevo componente “Destino de OLEDB”.
  3. Unir la flecha azul desde “Origen de OLE DB” hasta “Destino de OLE DB” de tal forma que quede de la siguiente manera:



* 1. Editar el paso “Distino de OLE DB”. Como administrador de conexiones seleccionar la conexión hacia la base OLAP. Como “Modo de acceso a datos” seleccionar “Carga rápida de tabla o vista” y en “Nombre de la tabla o vista” seleccionar “DimCliente”.

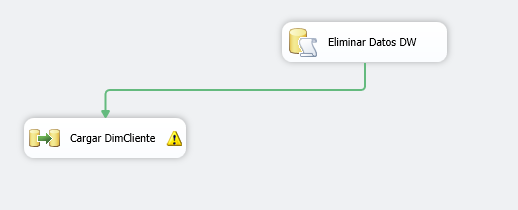


* 1. En la sección “Asignaciones”, relacionar los atributos extraidos desde la base OLTP con los atributos de la base OLAP. De esta manera estamos indicando al proceso que atributos del origen se deben insertar en qué atributos del destino.

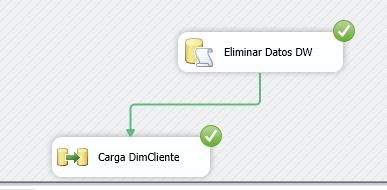


* 1. Click en “Aceptar” para finalizar.

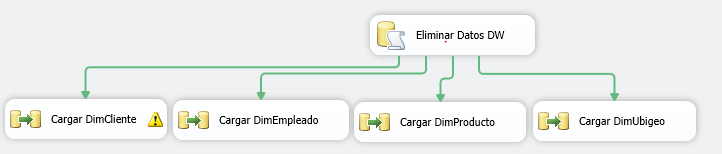
1. Volver al flujo de control y unir ambos componentes quedando de esta manera:



1. Ejecutar el proceso y ver si se completa la tabla DimCliente.
2. Una vez finalizado estos pasos, podemos probar la tarea llendo a “Depurar 🡪 Iniciar depuración”. Si todo salió bien debería verse de la siguiente manera:



1. Una vez que todo funcione correctamente, realizar la misma acción para cargar DimEmpleado, DimProducto y DimUbigeo. Prestar atención en que cambiará la consulta SQL para obtener los datos indicada en el paso “10.b” según cada origen y destino de cada dimensión.
2. El resultado final debería verse de la siguiente manera:



1. Asegurarse que todas las dimensiones se carguen correctamente. Para ello ejecute el proceso y luego puede consultar las dimensiones cargadas utilizando “SQL Server Management Studio”.

# CARGA DE DIMENSIÓN TIEMPO

1. La dimensión tiempo (DimTiempo) es una dimensión que se carga de una manera particular, dado que no solo hay que consultar el origen de datos, sino que hay que realizar un proceso de transformación para poder cargar todos sus atributos.
2. Arrastar un nuevo componente “Tarea de Flujo de datos”, renombrarlo a “Cargar DimTiempo” y acceder al mismo.
   1. Allí dentro arrastrar el componente “Origen de OLEDB”. Como administrador de conexiones seleccionar la conexión hacia la base OLTP. Como “Modo de acceso a datos” seleccionar “Comando SQL” y en “Texto de comando SQL” ingresar:

SELECT DISTINCT

YEAR(FechaPedido) \* 10000 + MONTH(FechaPedido) \* 100 + DAY(FechaPedido) AS CodigoFecha,

FechaPedido, YEAR(FechaPedido) AS Año,

DATEPART(QUARTER,FechaPedido) AS Trimestre,

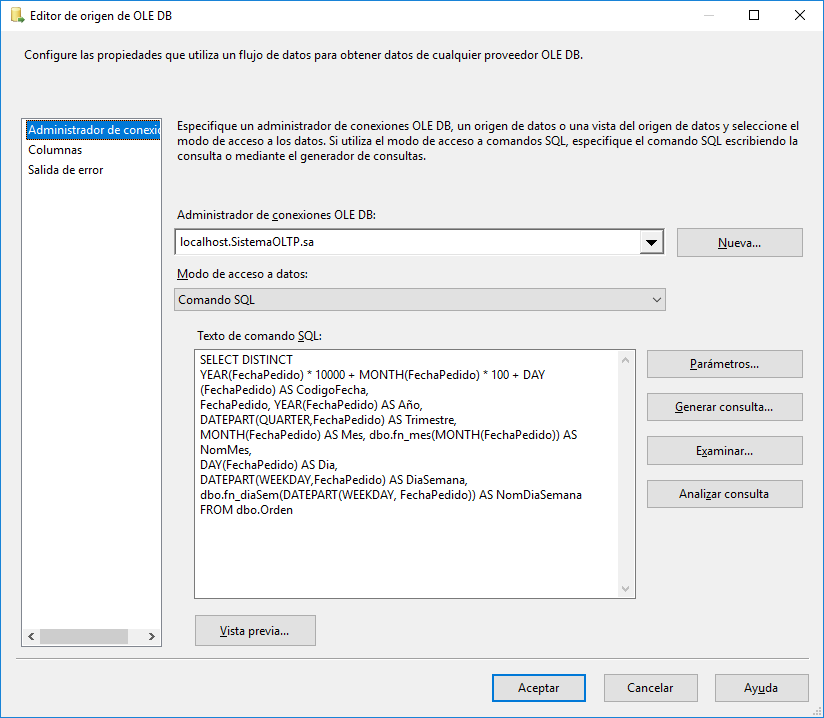
MONTH(FechaPedido) AS Mes, dbo.fn\_mes(MONTH(FechaPedido)) AS NomMes,

DAY(FechaPedido) AS Dia,

DATEPART(WEEKDAY,FechaPedido) AS DiaSemana,

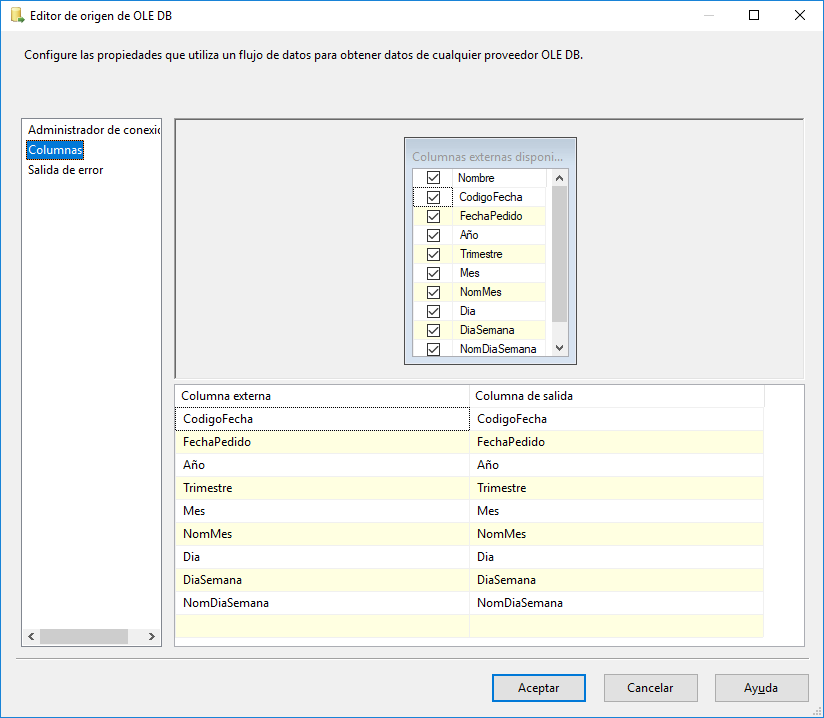
dbo.fn\_diaSem(DATEPART(WEEKDAY, FechaPedido)) AS NomDiaSemana

FROM dbo.Orden



Notese que la consulta propuesta utiliza funciones declaradas por el usuario (funciones que no son propias de SQL Server), dichas funciones son “fn\_mes()” y “fn\_diaSem()”. Estas funciones están declaradas en el script “SistemaOLAP.sql” donde podrán verla.

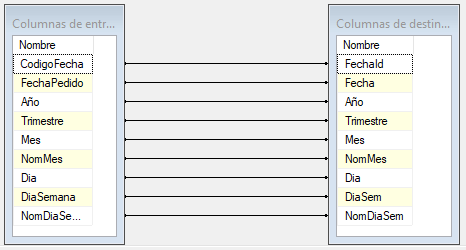
* 1. En la sección “Columnas” tildar todas.



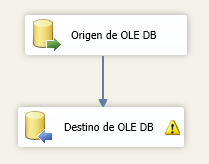
* 1. Arrastrar un nuevo componente “Destino de OLEDB”.
  2. Unir la flecha azul desde “Origen de OLE DB” hasta “Destino de OLE DB” de tal forma que quede de la siguiente manera:



* 1. Arrastrar un nuevo componente “Destino de OLEDB”. Como administrador de conexiones seleccionar la conexión hacia la base OLAP. Como “Modo de acceso a datos” seleccionar “Carga rápida de tabla o vista” y en “Nombre de la tabla o vista” seleccionar “DimTiempo”.
  2. En la sección “Asignaciones”, relacionar los atributos extraidos desde la base OLTP con los atributos de la base OLAP.

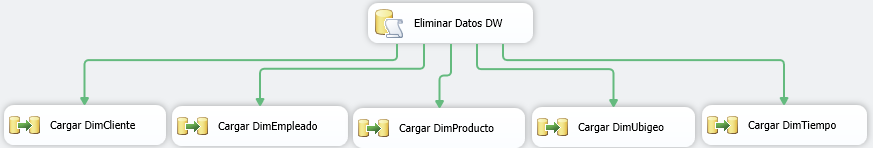


* 1. El flujo de tarea de DimCliente debe quedar así:



* 1. Click en “Aceptar” para finalizar.

1. El resultado final debería verse de la siguiente manera:



1. Asegurarse que todas las dimensiones se carguen correctamente. Para ello ejecute el proceso y luego puede consultar las dimensiones cargadas utilizando “SQL Server Management Studio”.

# CARGA DE TABLA DE HECHO

1. Por último, solo resta cargar la tabla de hecho “FactVentaOrden”. Esta tabla puede cargarse una vez que se hayan cargado todas las dimensiones, dado que la misma poseé claves foráneas de todas las dimensiones.
2. En la pestaña “Flujo de Control”, arrastrar el componente “Tarea Ejecutar SQL”. Renombrarla como “Cargar FactVentaOrden” y dentro de sus propiedades generales, seleccionar como valor de “Connection” a la base OLAP y cargar como valor de “SQLStatement”:

insert into FactVentaOrden

(

ECliente,EEmpleado,Ubigeo\_SKey,EProducto,FechaId,Total

)

SELECT distinct

ECliente,EEmpleado,Ubigeo\_SKey,EProducto,FechaId,(dord.PrecioUnidad\*dord.Cantidad) as Total

FROM SistemaOLTP..Orden ord

inner join SistemaOLTP..DetalleOrden dord on ord.IdOrden=dord.IdOrden

inner join DimCliente dc on dc.IdCliente=ord.IdCliente

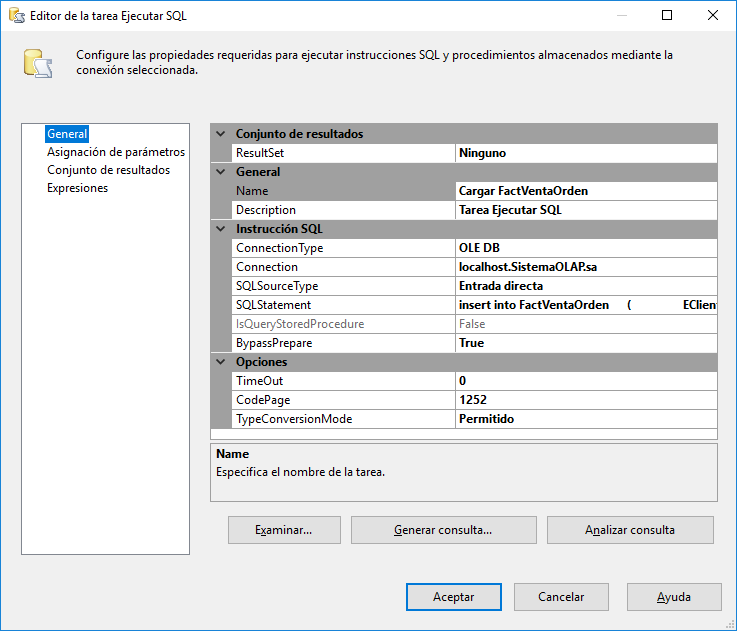
inner join DimEmpleado de on de.IdEmpleado=ord.IdEmpleado

inner join DimUbigeo du on du.IdDistrito=ord.IdDistrito

inner join DimProducto dp on dp.IdProducto=dord.IdProducto

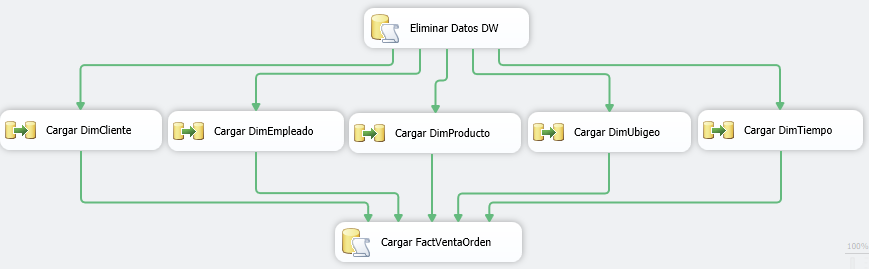
inner join DimTiempo dti on dti.Fecha=ord.FechaPedido

WHERE (ord.FechaPedido IS NOT NULL)



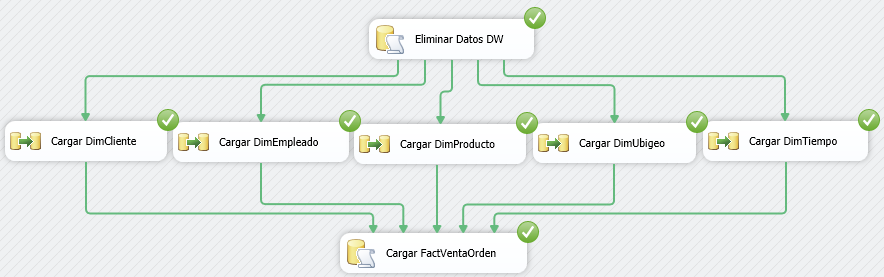
Nótese que se utiliza una consulta SQL compleja para tomar las claves primarias de cada dimensión y poder cargar la tabla de hechos. Este paso puede realizarse de diversas maneras en el proceso ETL, esta es solo una de esas maneras.

1. Es importante que esta tarea se ejecute al finalizar las otras, es por eso que se unirán de la siguiente manera:



1. Para asegurarse que todo el proceso funciona correctamente, ejecutarlo y luego consultar las dimensiones y la tabla de hecho cargadas utilizando “SQL Server Management Studio”.

Un proceso que funcione correctamente debería verse de la siguiente manera:



# CONCLUSION

Todo el proceso ETL propuesto en este tutorial solo indica una de las tantas formas de realizarlo. Solo hemos utilizado los componentes básicos, pero existen muchos otros componentes que nos permitirás desarrollar el proceso de una manera mucho más controlada y disgregada. Alguno de los componentes no vistos, pero más utilizados son: Búsqueda, Columna derivada, Ordenar y Conversión de datos.

Por otro lado, recuerden que un proceso ETL puede integrar fuentes heterogéneas para cargar un Data Warehouse, es por eso que la herramienta que utilizamos no solo permite conexiones OLE DB, sino que también permite conectarse a archivos planos (txt, csv, etc.), archivos de Excel, y otras fuentes existentes.

Espero que este tutorial haya sido de utilidad y tómenlo solo como un punto de partida a muchas otras formas de diseñar un proceso ETL.