Unidad 3

Actividad:

Evidencia de aprendizaje 3. Staging _Jardineria, proceso ETL

Julio César Cárdenas Veloth

Módulo – Base de datos II

Grupo 94

Profesor

Victor Hugo Mercado

Institución Universitaria Digital de Antioquia

Ingeniería de Software y Datos

Medellín

2024

1 Introducción

Teniendo en cuenta que una de las fases de gran importancia en la conformación de un data warehouse, y específicamente de un data mart, consiste en el proceso de extracción de la información necesaria desde las bases de datos. Esta información permitirá realizar la debida estructuración de las tablas de acuerdo a los requerimientos del cliente, las cuales posteriormente alimentarán cada una de las dimensiones que darán respuesta al modelo de negocio del cliente. Es así como en el presente trabajo se describe el proceso llevado a cabo de forma automatizada (ETL) para realizar este tipo de procedimientos para la base de datos jardineria.

1 Objetivo

Desarrollar un proceso de transformación y carga de datos desde la base de datos origen, pasar Staging y luego hasta el data mart final, utilizando la base de datos de staging previamente creada.

2 Planteamiento del problema

Una empresa de jardinería requiere poder obtener datos estadísticos de forma eficiente y ágil a partir de la base de datos que poseen actualmente, para esto requieren la implementación de un datamarts que les permita contestar las siguientes preguntas: ¿Cuál es el producto más vendido?, ¿Cuál es la categoría con más productos? y ¿Cuál es el año con más ventas?

3 Análisis del problema

De acuerdo a la necesidad que plantea la empresa de jardinería y lo costoso que puede resultar implementar un datamarts con toda la información que contiene la base de datos actual, es importante tener en cuenta lo siguientes elementos para dar respuesta a las preguntas plateadas en el problema:

Solo se tendrán en cuenta las tablas y campos que ayuden a resolver cada una de las preguntas planteadas, esto permitirá invertir menos horas en el levantamiento y análisis de requerimientos.

En el modelo estrella del datamarts solo se crearan las dimensiones que contengan la información para dar respuestas a las preguntas planteadas.

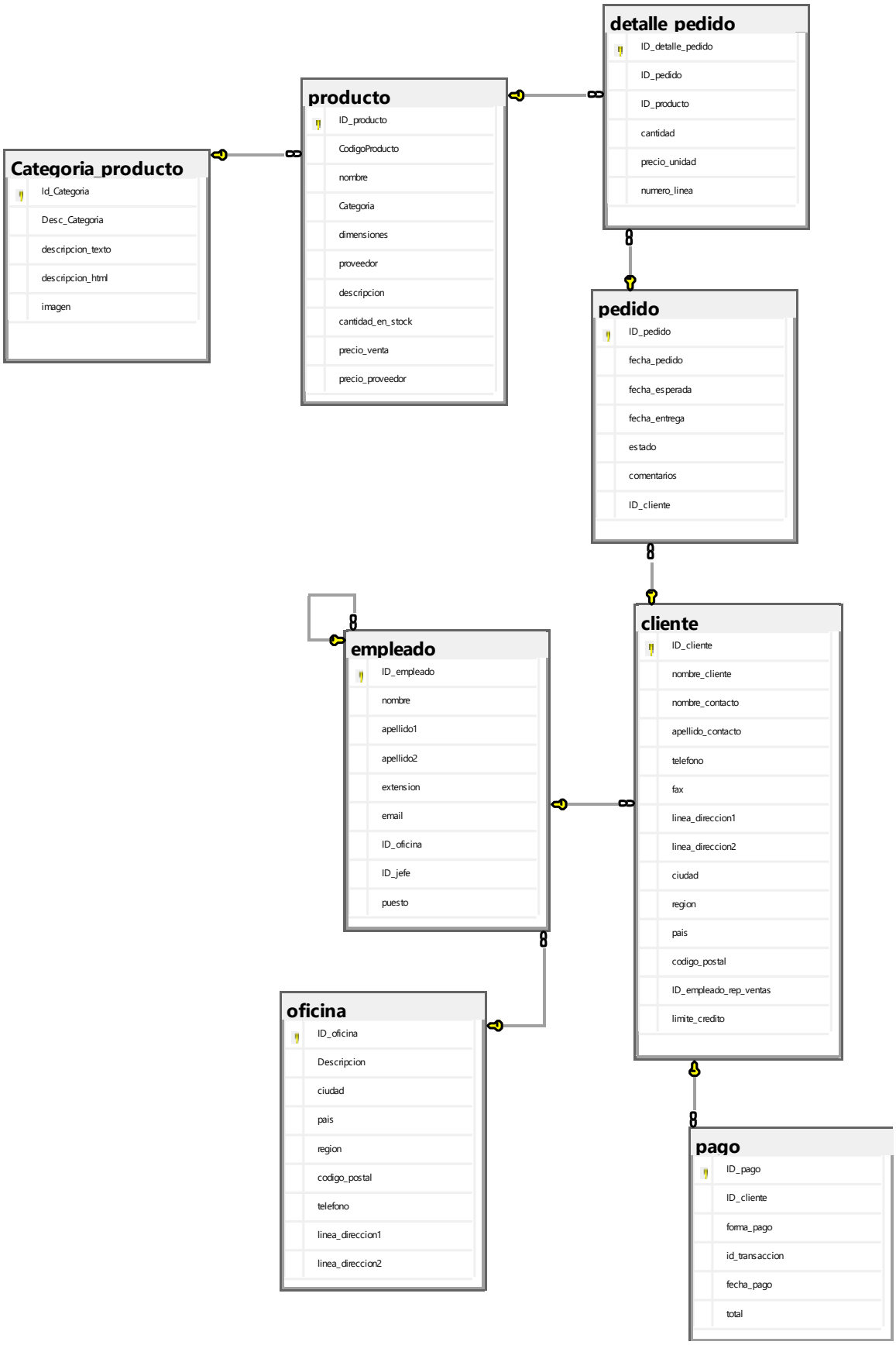


Figura 1. Modelo actual de la base de datos jardineria

4 Propuesta de la solución.

4.1 Correcciones a la entrega 1.

De acuerdo con las observaciones y sugerencias realizadas en la primera entrega del desarrollo del data marts, a continuación se realiza la descripción del modelo estrella corregido y los pasos detallados para la realización de la extracción de los campos e información necesaria para la conformación del data marts mediante procesos ETL.

4.2 Descripción del modelo estrella propuesto.

Para la construcción del modelo estrella del datamarts se tuvieron en cuenta los siguientes elementos:

Dimensiones.

Dimensión productos: en esta se realizó la integración de los campos necesarios de la tabla categoría_productos y los campos necesarios de la tabla productos.

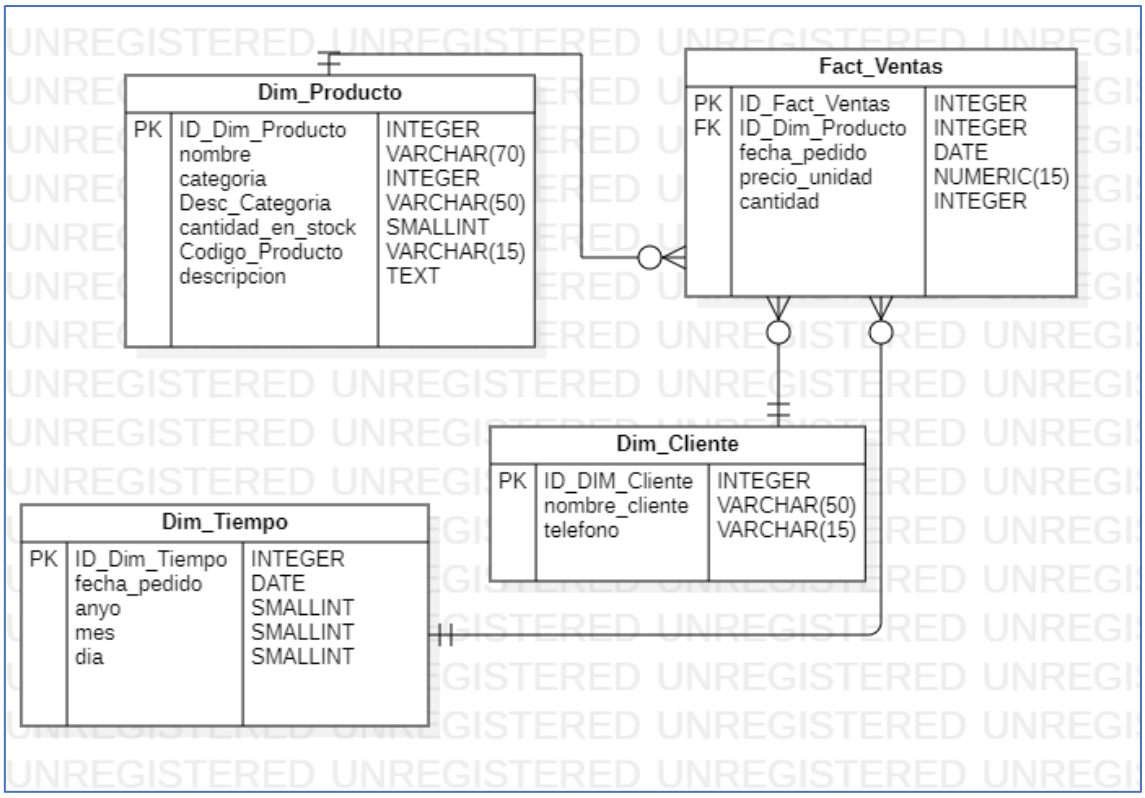
Dimensión tiempo: esta dimensión contiene el campo correspondiente al año, el cual servirá como filtro para seleccionar los grupos de datos del datamarts.

Tabla de hechos

Fact_Ventas: en esta tabla se integran las tablas de la base de datos trasaccional pedido y detalle_pedido.

Todas las tablas dimensiones presentan una relación de uno a muchos con la tabla de hechos.

4.2.1 Modelo estrella datamarts



4.2.2 Listado de campos por componente del datamart

Dim_producto

- ID_Dim_Producto (INTEGER); clave primaria que permite relacionar la dimensión con la tabla de hechos
- nombre (VARCHAR); este campo permitiría identificar cada uno de los productos mediante su nombre comercial.
- Categoría(int): Campo que funciona como clave foránea para conectar con la tabla Categoria_producto
- Desc_Categoria (INTEGER): este campo permite tener el detalle descriptivo de la categoría a la cual pertenece el producto, se tiene en cuenta para realizar el agrupamiento por categoría de los productos en stock y luego realizar el conteo de los productos agrupados, esto permitirá luego visualizar todos las categorías en orden ascendente e identificar cual es la que mayor productos posee.

- cantidad_en_stock (VARCHAR); permite realizar la consulta para conocer la categoría con mayor cantidad de productos, este campo proviene de la tabla producto.
- Codigo_Producto (VARCHAR); este campo permite realizar la identificación del producto mediante un código alfanumérico y proviene de la tabla producto.
- Descripción (TEXT): permite realizar una descripción del producto y proviene de la tabla producto.

Dim_Tiempo

- ID_Dim_Tiempo (INTEGER): clave primaria, funciona como indice.
- Fecha_pedido (DATE): Campo tipo fecha proveniente de la tabla pedido
- Anyo (INTEGER). Campo calculado que permite realizar los filtros por año de acuerdo a como lo solicita el cliente, el valor se extrae de la fecha pedido.
- Mes (INTEGER): Campo calculado que permite realizar filtros por mes.
- Día(INTEGER): Campo calculado que permite realizar filtros por día.

Dim_Cliente

ID_Dim_Cliente (INTEGER): Clave primaria que conecta con la tabla de hechos Fact_Ventas, provienen de la tabla cliente.

Nombre_cliente (VARCHAR): nombre del cliente, proviene de la tabla cliente

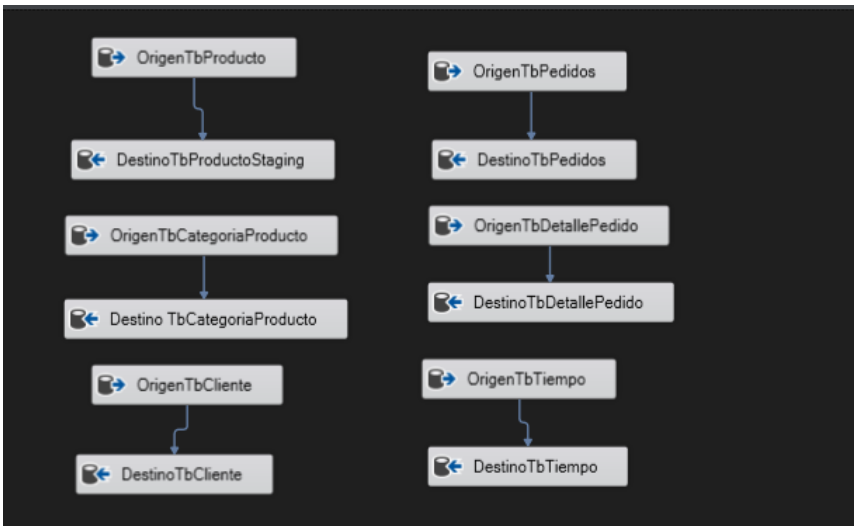
Teléfono (VARCHAR): Numero de contacto del cliente, proviene de la tabla cliente.

Fact_Ventas (Tabla de hechos)

- ID_Fact_Ventas (INTEGER); clave primaria de la tabla de hechos
- ID_Dim_Producto (INTEGER); clave foránea que trae los valores cualitativos de la tabla Dim_Productos
- fecha_pedido (DATE): clave foránea que permite relacionar la tabla hechos con la tabla Dim_Tiempo para aplicar filtros temporales de acuerdo a requerimiento del cliente, para este caso para año.
- precio_unidad (NUMERIC); se emplea para conocer el año con más ventas al realizar la multiplicación con el campo cantidad, en este caso se tendría el valor en pesos.
- cantidad (INTEGER); se utiliza para realizar una sumatoria total por año y conocer basado en la cantidad de productos vendidos el año con mayor ventas.

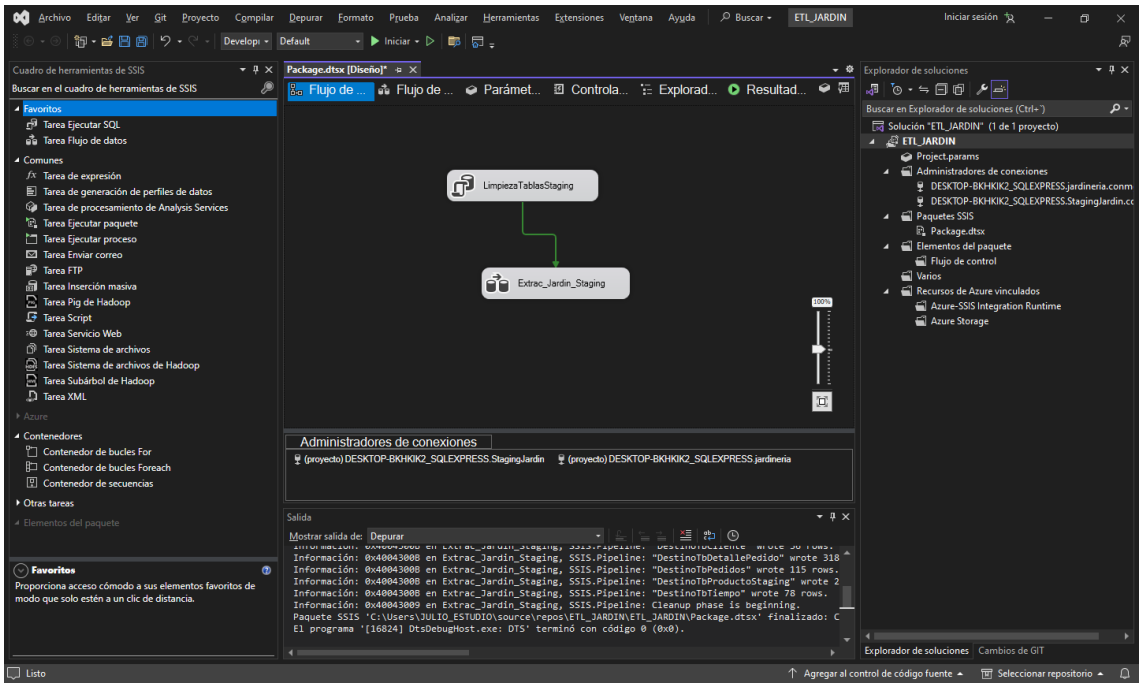
4.3 Descripción del análisis realizado a los datos Jardinería y cómo estos se trasladaron a la base de datos Staging .

Teniendo en cuenta la herramienta Integration Services Project, la cual funciona dentro de Microsoft Visual Studio se procedió a elaborar una tarea de flujo de datos de forma automatizada, definiendo un origen y un destino para la información a procesar relacionada con el data marts, para este caso se establecieron 6 procesos automatizados como se muestran en la siguiente imagen.



Para cada uno de los procesos automatizados de extracción se definió como origen la base de datos jardineria y como destino la base de datos StagingJardin.

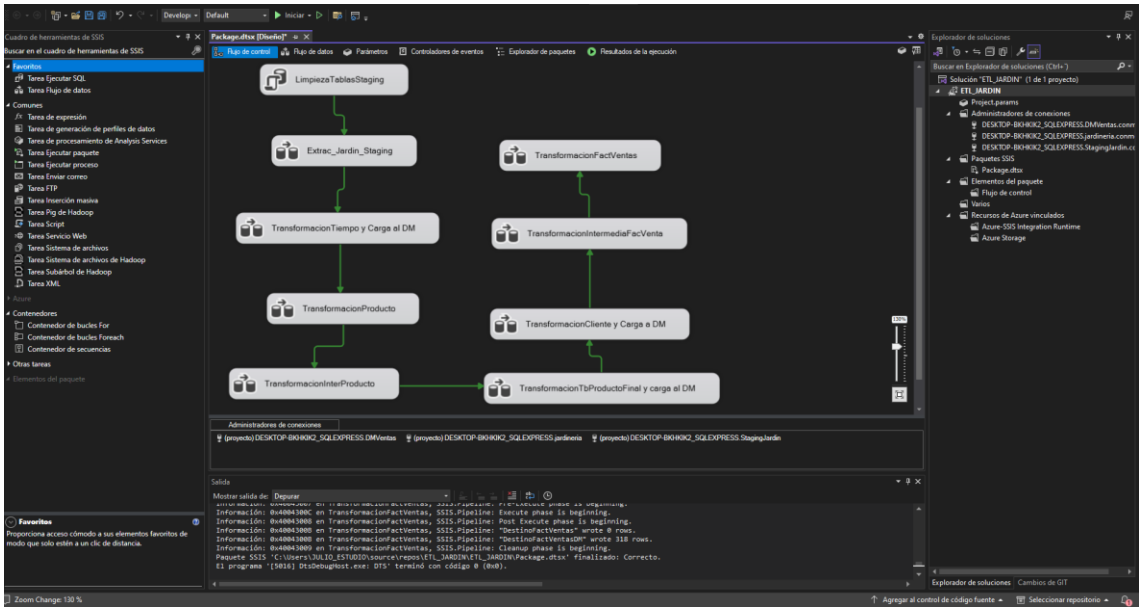
Con el fin de garantizar que, al correr la tarea automatizada, orientada a realizar la extracción de cada uno de los campos y la información respectiva contenida en la base de datos jardineria y que no se presentaran duplicados en la base de datos StagingJardin, se implementó una tarea de SQL automatizada, encargada de realizar una limpieza de las tablas antes de volverlas a poblar con información, en la siguiente imagen se observan los dos flujos automatizados definidos.



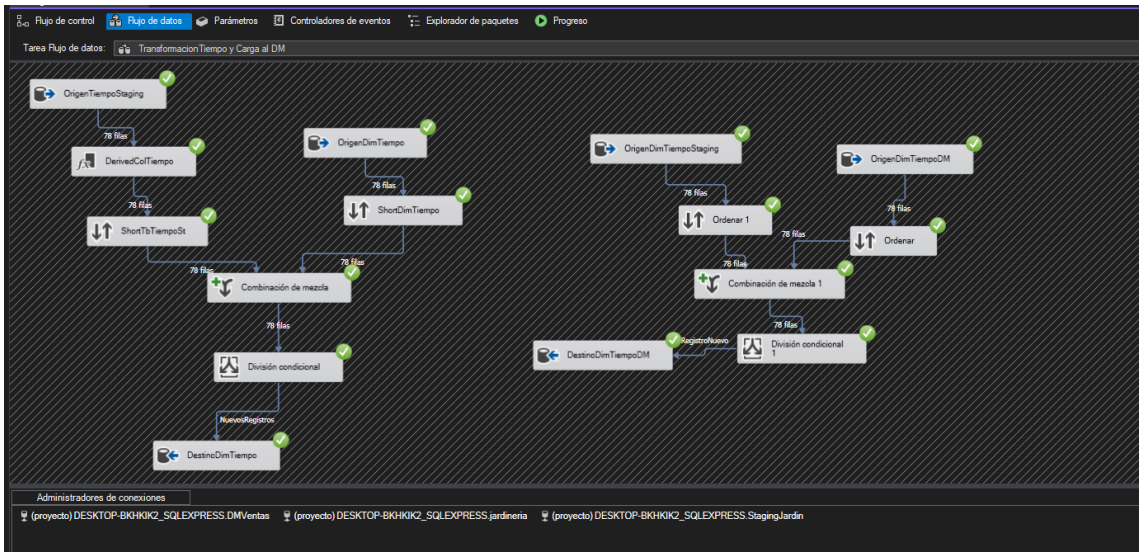
El contenido de las consultas elaboradas para realizar los procesos de extracción, se adjuntan a este documento.

4.4 Proceso de transformación y carga de datos al Data Marts.

Con los campos y tablas debidamente extraídos en el ejercicio anterior y almacenados en la base de datos intermedia StagingJardin, se procedió a realizar los procesos ETL para realizar la conformación de las dimensiones y la tabla de hechos como se muestra en las siguientes imágenes:

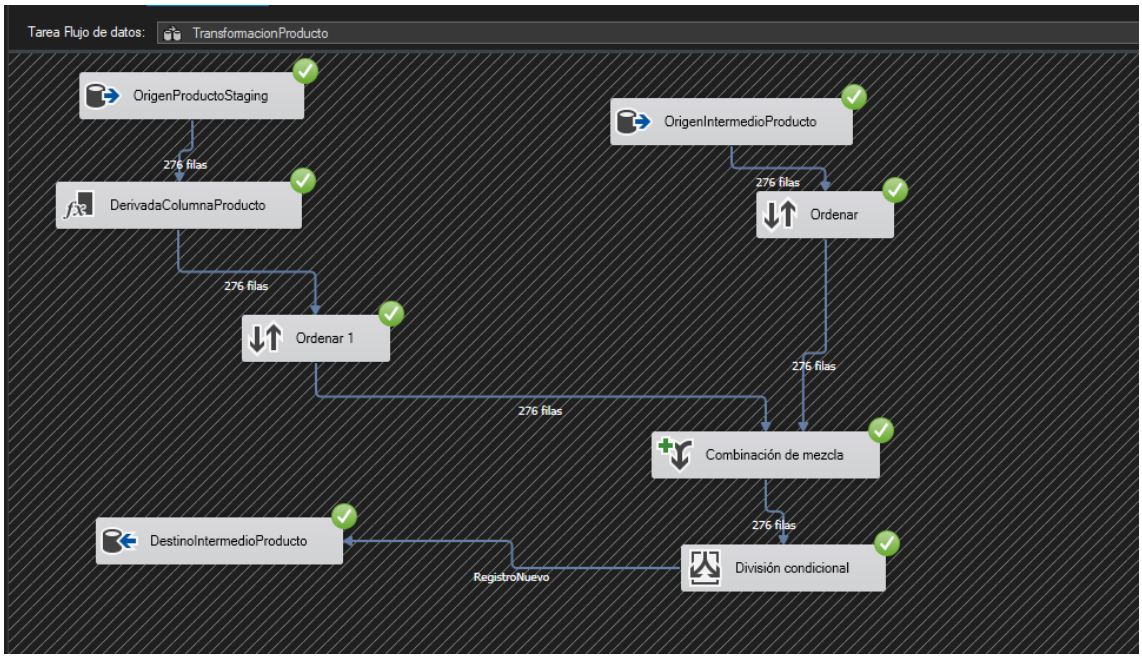


En la siguiente imagen se muestra el proceso de transformación y carga de la dimensión tiempo.

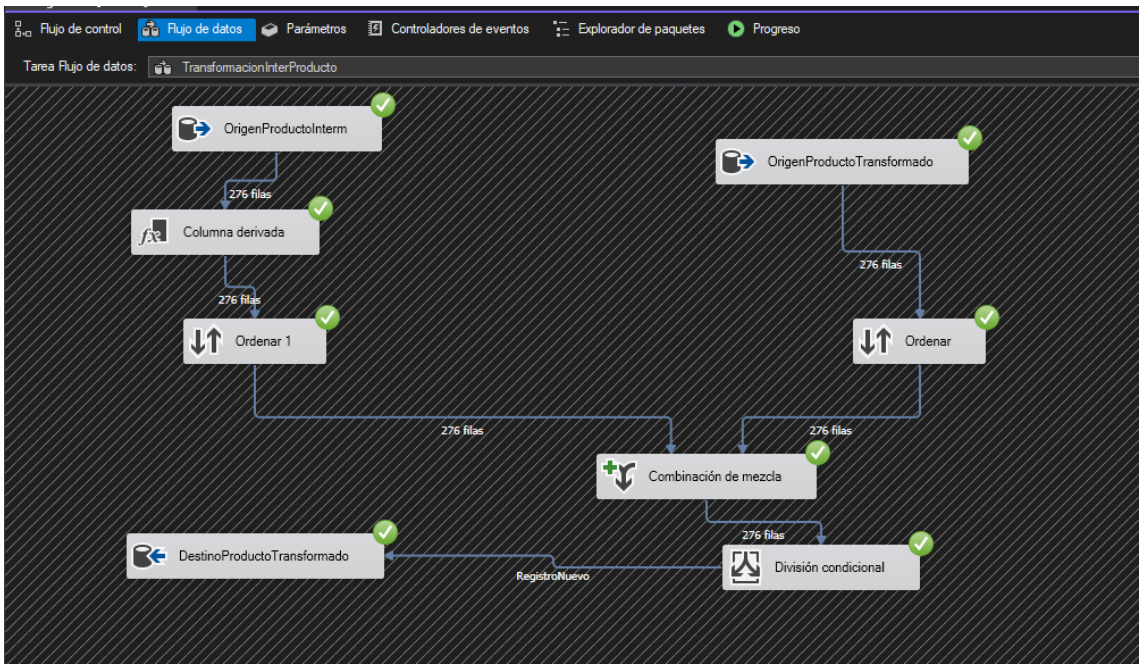


Proceso de transformación y carga de la dimensión producto, para esta fue necesario realizar algunos pasos intermedio, ya que el campo descripción de los productos era de valores tipo DT_TEXT y estos no son soportados en los procesos de la herramienta SSIS, por lo que se convirtieron a valores tipo DT_WSTR y luego de esto si realizaron los pasos para emplear las herramientas de columna derivadas.

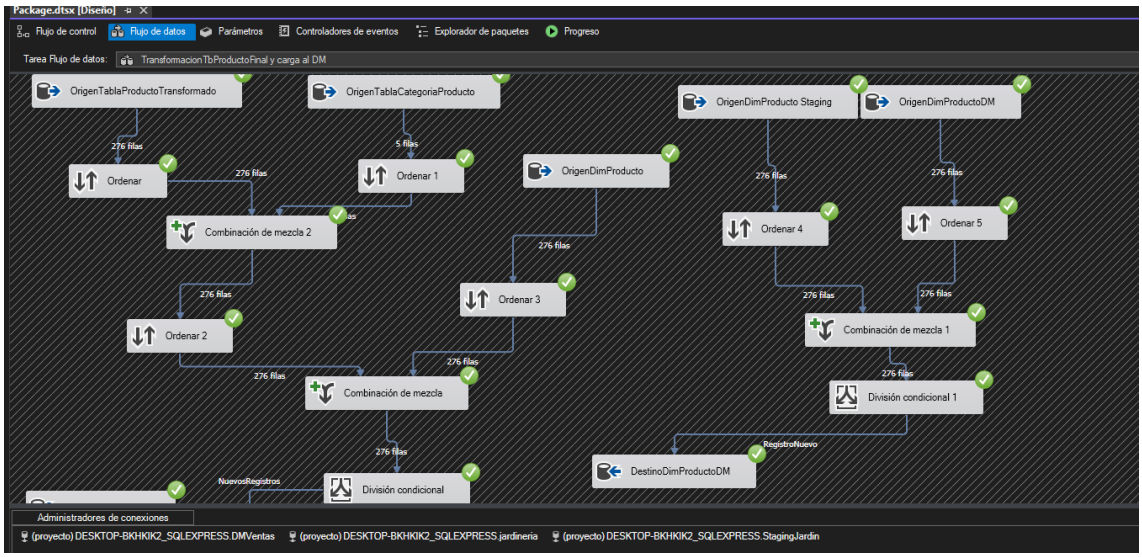
En la siguiente imagen se muestra el proceso realizado para la conversión de DT_TEXT a DT_WSTR del campo descripción y se crea un atabla intermedia con el nombre TabIntermediaProducto.



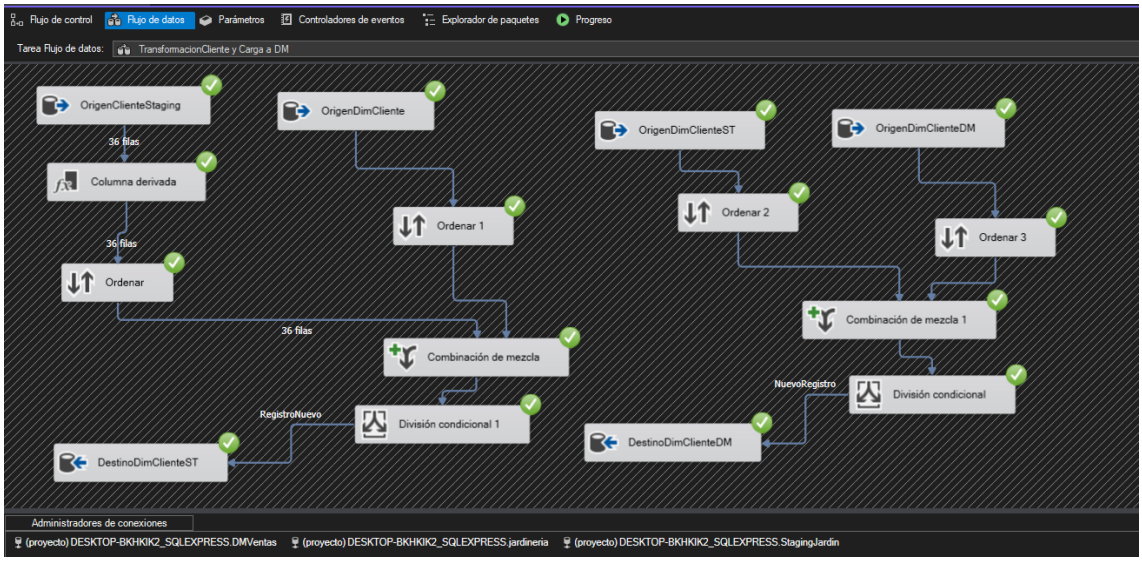
En la siguiente imagen se muestra el proceso en el que se utilizó la herramienta columna derivada para realizar el llenado de los campos en blanco que tenían la columna descripción de productos por la palabra “SIN_DESCRIPCIÓN”.



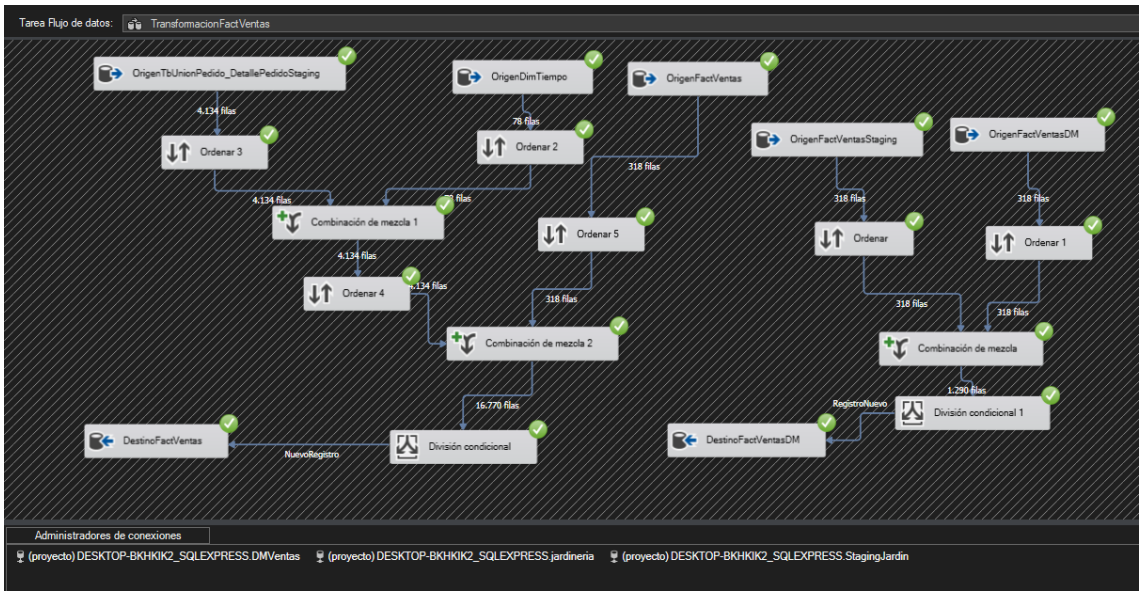
Y en la siguiente imagen ya se puede observar el proceso de transformación y creación de la dimensión producto y el proceso de carga hacia el Data Marts.



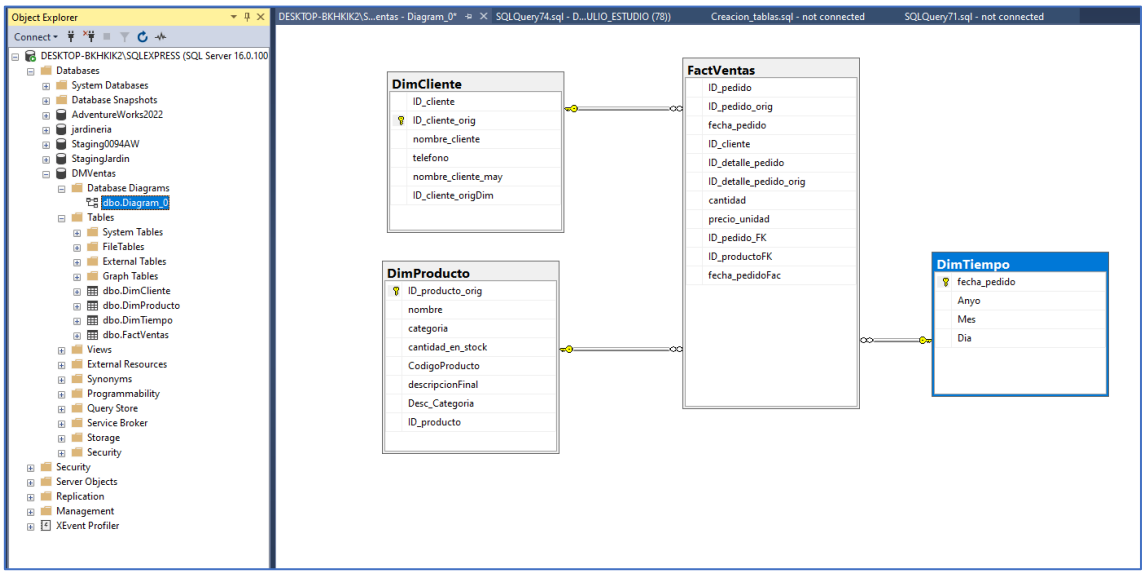
En la siguiente imagen se muestra el proceso de transformación y carga al Data Marts de la dimensión cliente.



En la siguiente imagen se observa los procesos para la creación de la tabla de hechos FactVentas y su proceso de carga al Data Marts.



Una vez se realizó todo el proceso ETL, se corroboró que el respectivo Data Marts quedará talmente relacionado como se observa en la siguiente imagen.



5 Conclusiones

Como conclusión del ejercicio se pude mencionar lo siguiente:

Antes de iniciar cualquier proceso extracción y carga de información automatizado para la creación de un data marts, es importante tener definido de forma coherente el modelo de datos, a partir del cual se realizará la estructuración de las consultas SQL orientadas a la selección especifica de los campos que se tendrán en cuenta para la conformación de los data marts.

La implementación de procesos automatizados para la extracción, transformación y carga de la información hacia los data marts, permite disminuir los errores en los procesos de migración de la información, de igual forma permiten el ahorro de esfuerzos, tanto en recursos económicos como humanos al no depender de personas que realicen los procesos de actualización, cada vez que produzcan cambios en la base de datos origen.

De acuerdo a los procesos de transformación realizado a las diferentes tablas extraídas para conformar el Data Marts se pudo evidenciar, que algunos procesos de transformación no son soportados por las herramientas SSIS y se deben establecer procesos intermedios o emplear consultas SQL directamente.

6 Bibliografía

IUdigital [WEB] (s.f.). Recuperado de https://iudigital.instructure.com/courses/15609