Evidencia de aprendizaje 3. Proceso de transformación de datos y carga en el data mart final

Carlos Dueiner Castaño Rodríguez

Facultad de Ingeniería y ciencias agropecuarias

Bases de datos II

Docente: Víctor Hugo Mercado

Ingeniería de Software y Datos

IU Digital de Antioquía

Marzo 2025

Introducción

En el contexto actual donde la inteligencia empresarial (Business intelligence) es el motor de crecimiento de una organización, los datos se convierten en la fuente de dicha inteligencia, siendo el eje mediante el cual, tras su recolección, limpieza, organización y selección, se toman las decisiones que marcarán fundamentalmente el rumbo de dicha organización. De este modo, la manera en la que se presentan estos datos tras ser recolectados, es vital para lograr la obtención de información valiosa que ayude a tomar las medidas correspondientes frente a las necesidades o demandas de la empresa.

Es claro que las aplicaciones o plataformas manejadas por una organización están soportadas por una base de datos tradicional, evitando redundancias, por lo que están altamente normalizadas, estas, junto con otras fuentes de utilidad para la organización arrojan estos datos esenciales, dichos datos usualmente son de grandes volúmenes y necesitan ser procesados de manera correcta y ordenada de manera que generen información de valor que soporten las decisiones tomadas para el crecimiento esperado.

En este contexto, uno de los modelos más utilizados para estructurar estos datos y lograr los objetivos indicados es el modelo estrella, el cual permite consultar y analizar dichos datos por medio de una tabla central de hechos vista o analizada desde distintas dimensiones, que a su vez son otras tablas conectadas a esta que aportan detalles relevantes adicionales a los hechos. Para el propósito de esta actividad se diseñará un modelo estrella para un data mart que permita dar respuesta a tres categorías de la base de datos llamada **Jardinería,** dichas categorías son:

- Identificar el producto más vendido
- Identificar la categoría con más productos
- Identificar el año con más ventas.

Objetivos

Objetivo general:

La idea es que con la ejecución de esta actividad, creando el data mart de la base de datos jardinería, se demuestre la comprensión de los temas relacionados al modelo de estrella, definiendo y diferenciando la tabla de hechos y su relación con las dimensiones desde la que será evaluada, identificando los campos y tipos de datos que son relevantes en mi base de datos tradicional que nos ayudarán a lograr los requerimientos que senos han realizado, asimismo como la creación de estos mismos campos pero ahora en la tabla de hechos y las dimensiones correspondientes. De este modo se podrá reconocer la adquisición de habilidades de detalle y análisis que son esenciales para la correcta resolución de la problemática abordada.

Además de este análisis, se debe confirmar el aprendizaje del proceso de extracción de las tablas que podrán ayudar a crear las dimensiones por medio de una base de datos temporal o Staging, a la cual pasarán las tablas correspondientes a las cuales posteriormente se la hará el proceso de transformación.

Objetivos específicos:

- Lograr identificar correctamente la tabla de hechos y las dimensiones realmente necesarias para abordar la problemática de estudio.
- Detallar de forma correcta todos los campos y su tipo de dato de la tabla de hechos y las tablas de dimensiones que serán indispensables para lograr dar respuesta a cada requerimiento.
- Establecer las relaciones lógicas entre la tabla de hechos y las dimensiones, que permitan cumplir con las categorías solicitadas: identificar el producto más

vendido, la categoría con más productos y el año con más ventas, utilizando la base de datos Jardinería como fuente de datos.

- Diseñar de manera elegante y entendible la gráfica del modelo de estrella donde los anteriores objetivos se vean plasmados de forma correcta.
- Realizar de manera correcta todo el proceso de extracción desde la base de datos de Jardinería usando Visual Studio Code, estableciendo la conexión con SQL Server.
- Identificar las tablas que tienen los campos necesarios en cada dimensión del modelo estrella, de modo que el proceso de extracción cumpla con todos los datos necesarios para hacer posible la creación de las dimensiones.
- Realizar el análisis correcto para el proceso de transformación de los datos, de modo que podamos tener información de valor, eliminando datos nulos, reemplazando valores, cambiando a mayúscula, ordenando las columnas.
- Ejecutar todo el proceso de carga al data mart, permitiendo así tener todo el análisis para su presentación final por medio de dashboards fáciles de leer.

Planteamiento del problema

El software de la empresa Jardinería, está soportado por una base de datos transaccional, la cual tiene almacenada información relevante acerca de su servicio en las siguientes tablas: producto, categoría_producto, cliente, pedido, detalle_pedido, estas tablas están correctamente normalizadas y relacionadas de manera apropiada para el sostenimiento de la aplicación que administra la empresa, almacenando así las operaciones diarias; pero dicha base de datos no está diseñada para realizar análisis profundos y eficientes que conlleven a la toma de decisiones estratégicas en la búsqueda del exitoso crecimiento de la organización.

Frente a esta realidad, se ha planteado la necesidad de crear un data mart, usando el modelo de estrella, en el que se pueda dar respuesta a los siguientes requerimientos específicamente:

- Identificar el producto más vendido
- Señalar la categoría con más productos
- Traer el año con más ventas

Para ello, debe hacerse un análisis profundo que permita identificar las tablas de la base de datos transaccional que conformarán la tabla de hechos y a su vez, relacionarla con las tablas de dimensiones con las que se logrará dar respuesta a cada requerimiento, identificando los campos que deben estar presente en cada tabla; de modo que no solamente se de respuesta a los tres requerimientos planteados, sino que también se logre dar una base sólida para posibles futuras categorías adicionales que puedan necesitarse en pro del crecimiento de la jardinería; por ejemplo, en el caso que no solo pregunten por el año con más ventas, sino el mes, semana o día del año con más ventas.

Una vez realizado el proceso de modelado del diagrama de estrella, se debe proceder con el análisis de las tablas para realizar todo el proceso de ETL, realizando las consultas correspondientes en las tablas de las que necesitamos extraer los datos necesarios para crear las dimensiones correspondientes posteriormente, realizando las conexiones correctas entre la base de datos de Jardinería en SQL Server y Visual Studio Code en este caso, para que dichas consultas pasen los datos de la base de datos transaccional a la base de datos temporal “Staging” para su uso posterior en las dimensiones.

Posteriormente, luego de tener nuestras tablas de Staging, debemos crear un nuevo paquete SSIS de transformación que se conectará con la base de datos de Staging y allí mismo haremos el proceso de transformación de las tablas de Staging, de modo que podamos crear las dimensiones propiamente, allí mismo en la tabla Staging pero con los datos transformados y organizados, de modo que se tenga información contundente y de valor.

Finalmente, se crea un nuevo paquete SSIS llamado carga en el que podremos finalmente crear las tablas de las dimensiones finales y la tabla de hechos con los datos ya transformados, ordenados y limpios; con esta es que finalmente se podrá tener respuesta a las preguntas planteadas al iniciar el proyecto.

Análisis del problema

Como es sabido, las bases de datos transaccionales son excelentes para servir como soporte de almacenamiento y consistencia de los datos que por medio de la plataforma se ingresa, haciendo posible la ejecución de distintos procesos para los cuales la aplicación fue diseñada en beneficio de la organización, no obstante, este tipo de base de datos transaccionales, aunque por medio del lenguaje de consulta DQL es posible extraer información de utilidad para ciertas decisiones a tomar, estos no nos dan un detalle profundo o amplio de distintos enfoques desde los que podríamos soportar decisiones de gran importancia dentro de la empresa.

Si se intentara trabajar solo con consultas SQL, se requeriría hacer las mismas consultas cada vez que se requiere nueva información o se actualizan los datos, lo cual genera una mayor carga operativa y mayor extensión en el tiempo requerido para integrar los datos manualmente, además dado que es muy manual dicho proceso, podría desaprovecharse información de gran relevancia, maximizando la posibilidad de tomar decisiones sin mucho fundamento que podrían generar confusión en lugar de dar soluciones.

Para superar dichas limitaciones, se hace necesaria la creación de un data mart, en el que podamos por medio de un modelo de estrella, reorganizar todo este gran volumen de datos para facilitar la identificación de métricas clave, consolidando las ventas en una tabla de hechos y sus correspondientes descripciones en las tablas de dimensiones, con esto logramos:

- Reducir la complejidad de cada consulta
- Permitir acceso de madera directa a información que es clave.
- Escalabilidad, pues no solo se da respuesta a las preguntas actuales sino a posibles consultas futuras que vayan surgiendo a medida que crece el negocio.

Además de los procesos correspondientes de análisis y modelado de diagrama estrella, se debe realizar un proceso de extracción que permita por medio de una base de datos de Staging, crear las tablas con los datos necesarios para la posterior creación de las tablas de dimensiones identificadas al crear el modelo de estrella.

Propuesta de la solución

Descripción del modelo de estrella:

El modelo de estrella creado, tiene una estructura que se basa en una tabla de hechos central y varias tablas de dimensiones que nos dan contexto y hace posible el análisis de los datos. Cada tabla de dimensión tiene su clave primaria que pasa como clave foránea en la table de hechos, haciendo así una relación de uno a muchos, conectando las dimensiones con la tabla de hechos. De este modo es posible analizar cada una de las ventas desde distintas dimensiones, además, es posible agruparlas según cada dimensión y dar respuesta a los requerimientos que el dueño del producto demanda para su análisis y toma de decisiones.

Tabla de hechos: Llamada FAC_Ventas; Contiene su clave principal id_venta y las claves foráneas de cada una de las tablas de dimensiones. Además, cuenta con otros campos propios como la forma de pago, la cantidad y el total de la venta

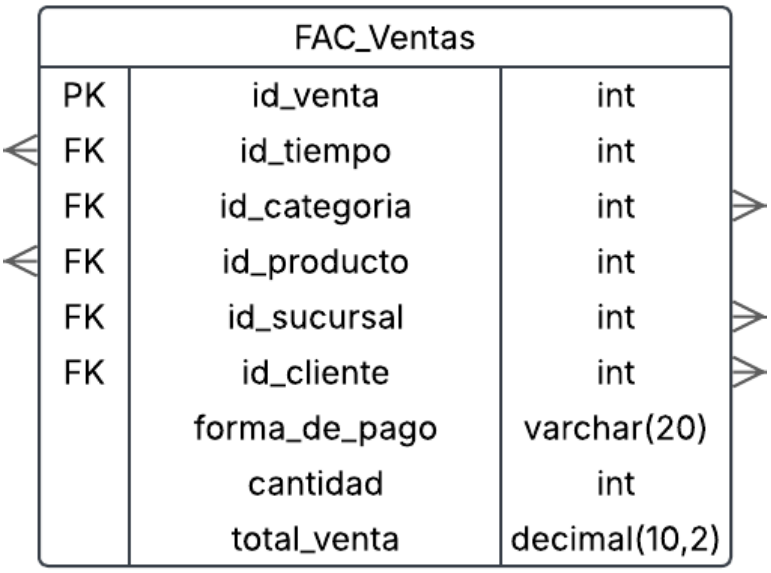


Tabla de Dimensión de tiempo; DIM_Tiempo: En esta tabla, encontramos campos relacionados a detalles sobre el tiempo asociado a las ventas, cuya clave principal es el id_tiempo y demás campos como fecha, año, semestre, mes_texto, num_semana, fin_de_semana (booleano), día_semana, y hora.

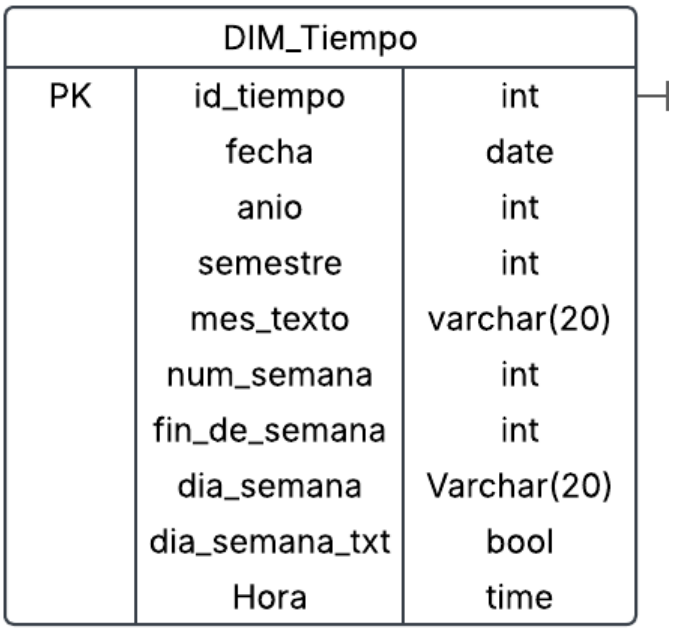
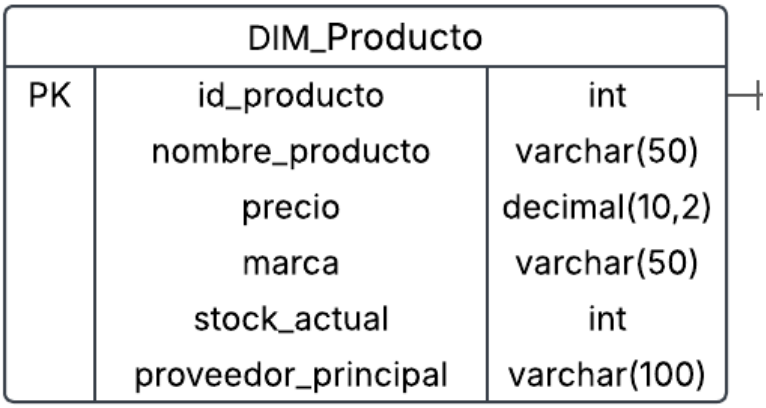


Tabla de dimensión de producto; DIM_Producto: Aquí se proporciona detalles sobre los productos vendidos y cuenta con su clave primaria id_producto, además de atributos de valor como nombre_producto, precio, marca, stock_actual y proveedor_principal



Además de estas tablas, con las cuales de por si ya es posible dar respuesta a las categorías que el usuario de la Jardinería demanda, se han añadido dos tablas adicionales de gran importancia, dado el caso que en el futuro se demanden más categorías, dentro de estas tablas tenemos:

Tabla de dimensión de sucursal; DIM_Sucursal; Contiene detalles de las sucursales donde se realizan las ventas. Incluye:su clave primaria id_sucursal y demás atributos como nombre_sucursal, país, ciudad y dirección.

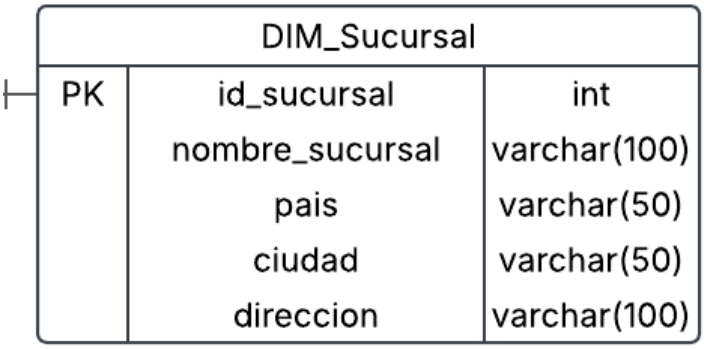
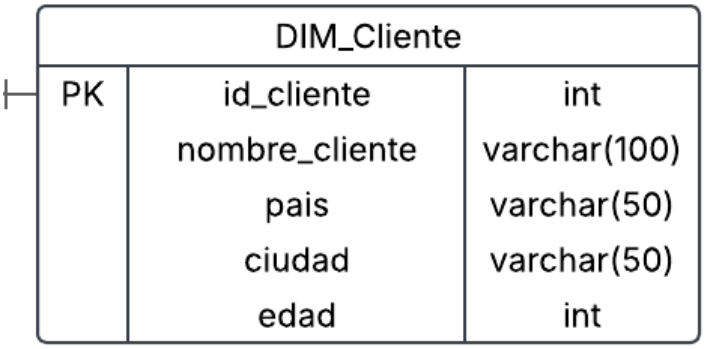
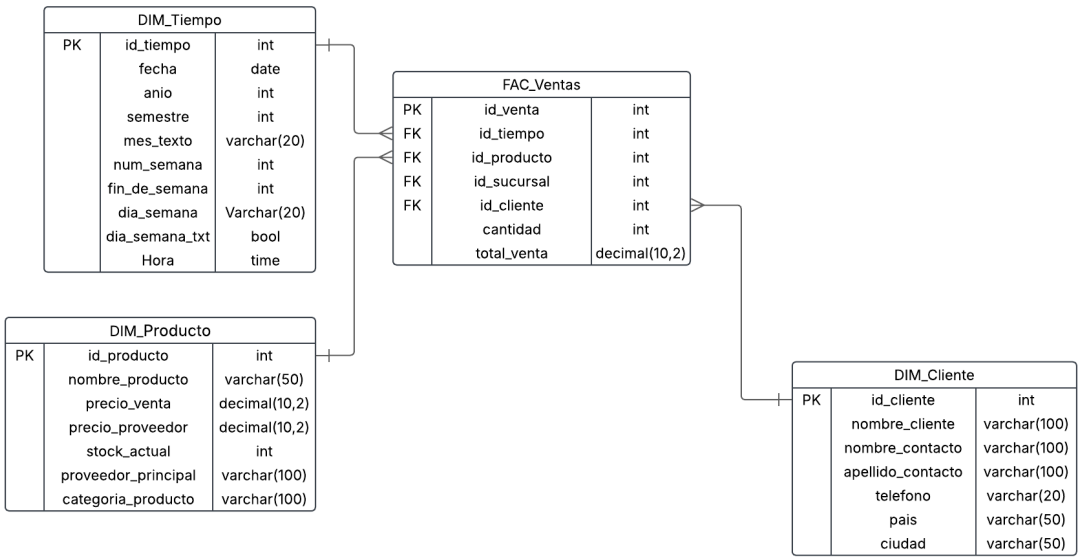


Tabla de dimensión de cliente; DIM_Cliente: Proporciona información sobre los clientes. Incluyendo la clave principal id_cliente y demás atributos como nombre_cliente, país, ciudad y edad.



A continuación, se comparte la gráfica del modelo de estrella, tanto la imagen, como su respectivo link de la herramienta Lucidchart.



https://lucid.app/lucidchart/c6a1db98-295d-4e54-bde1-3f3c343e90aa/edit?viewport_loc=-90%2C-55%2C1862%2C787%2C0_0&invitationId=inv_987a1269-c012-4257-96b8-1ec8f577ce5f

Proceso de extracción de las tablas de la base de datos jardinería a la base de datos Staging.

Una vez identificadas las dimensiones, con ayuda del diagrama del modelo de estrella, se realiza el análisis correspondiente a la base de datos Jardinería, fijándonos en el diagrama relacional de dicha base de datos, y de esta manera se puede identificar que con el fin de crear las dimensiones correspondientes, primero se deben hacer la extracción de las tablas que necesitamos con los datos que harán posible la lógica de dichas dimensiones, lo cual fue posible realizando los siguientes procesos.

Establecimiento de conexión:

Primero se creó una base de datos en SQL Server llamada StagingJardinería la cual está vacía, luego se estableció en Visual Studio Code la conexión a la base de datos Jardinería y a la nueva base StagingJardinería, definiendo así un origen y un destino; una vez hecho esto, se crearon las tareas de flujo de datos cada una con su respectiva ejecución de limpieza de tablas, donde se truncaron las tablas de modo que no se repitan las entidades, de esta manera las tareas creadas fueron:

- ExtraccionJardineriaProducto
- ExtraccionJardineriaCliente
- ExtraccionJardineriaTiempo
- ExtraccionFAC

De la misma manera, para cada una de estas tareas, se crearon los flujos de datos correspondientes, siendo las siguientes las tablas extraídas y las creadas en la base de datos StagingJardineria respectivamente:

Tabla Producto:

- **Consulta SQL para la extracción:**

```
SELECT [ID_producto]
      ,[CodigoProducto]
      ,[nombre]
      ,[Categoria]
      ,[proveedor]
      ,[cantidad_en_stock]
      ,[precio_venta]
      ,[precio_proveedor]
  FROM [jardineria].[dbo].[producto]
```

- **Creación de la tabla en Staging:**

```
CREATE TABLE "StagingProducto" (
    "ID_producto" int,
    "codigo_Producto" nvarchar(15),
    "nombre_producto" nvarchar(70),
    "categoria_producto" int,
    "proveedor" nvarchar(50),
    "cantidad_en_stock" smallint,
    "precio_venta" numeric(15,2),
    "precio_proveedor" numeric(15,2)
)
```

Tabla CategoríaProducto:

- **Consulta SQL para la extracción**

```
SELECT  [Id_Categoria]
      ,[Desc_Categoria]
      ,[descripcion_texto]
  FROM [jardineria].[dbo].[Categoria_producto]
```

- **Creación de la tabla en Staging:**

```
CREATE TABLE "StagingCategoriaProducto" (
    "Id_Categoria" int,
    "Desc_Categoria" nvarchar(50),
    "descripcion_texto" nvarchar(max)
)
```

Tabla Cliente:

- **Consulta SQL para la extracción:**

```
SELECT [ID_cliente]
      ,[nombre_cliente]
      ,[nombre_contacto]
      ,[apellido_contacto]
      ,[telefono]
      ,[pais]
      ,[ciudad]
      ,[codigo_postal]
  FROM [jardineria].[dbo].[cliente]
```

- **Creación de la tabla en Staging:**

```
CREATE TABLE "StagingCliente" (
    "ID_cliente" int,
    "nombre_cliente" nvarchar(50),
    "nombre_contacto" nvarchar(30),
    "apellido_contacto" nvarchar(30),
    "telefono" nvarchar(15),
    "pais" nvarchar(50),
    "ciudad" nvarchar(50),
    "codigo_postal" nvarchar(10)
)
```

Tabla FechaVenta:

- **Consulta SQL para la extracción:**

```
SELECT [fecha_pedido]
  FROM [jardineria].[dbo].[pedido]
  union all

SELECT [fecha_esperada]
  FROM [jardineria].[dbo].[pedido]
  union all

SELECT [fecha_entrega]
  FROM [jardineria].[dbo].[pedido]
  order by fecha_pedido
```

- **Creación de la tabla en Staging:**

```
CREATE TABLE "StagingFechaVenta" (
    "fecha_pedido" date
)
```

Tabla de pedido y detalle pedido para la creación de la tabla FAC

Tabla pedido:

- **Consulta SQL para la extracción:**

```
SELECT [ID_pedido]
      ,[fecha_pedido]
      ,[fecha_esperada]
      ,[fecha_entrega]
      ,[estado]
      ,[ID_cliente]
  FROM [jardineria].[dbo].[pedido]
```

- **Creación de la tabla en Staging:**

```
CREATE TABLE "StagingPedido" (
    "ID_pedido" int,
    "fecha_pedido" date,
    "fecha_esperada" date,
    "fecha_entrega" date,
    "estado" nvarchar(15),
    "ID_cliente" int
)
```

Tabla DetallePedido:

- **Consulta SQL para la extracción:**

```
SELECT [ID_detalle_pedido]
      ,[ID_pedido]
      ,[ID_producto]
      ,[cantidad]
      ,[precio_unidad]
      ,[numero_linea]
  FROM [jardineria].[dbo].[detalle_pedido]
```

- **Creación de la tabla en Staging:**

```
CREATE TABLE "StagingDetallePedido" (
    "ID_detalle_pedido" int,
    "ID_pedido" int,
    "ID_producto" int,
    "cantidad" int,
    "precio_unidad" numeric(15,2),
    "numero_linea" smallint
)
```

Proceso de transformación de las tablas de la base de datos StagingJardinería a las tablas de dimensiones dentro de esa misma base de datos:

Es claro que con los datos extraídos, por lo que es necesario realizar un proceso de transformación de los datos de cada una de estas tablas, de manera que se puedan transformar los datos nulos por datos más claros como N/A o “Sin detalle”, además, se puede poner en mayúscula datos que queramos resaltar y por medio de Visual Studio, por medio de los objetos sort, merge join, Derived Column, y Conditional Split, llegar al destino final que sería la Dimensión de la tabla.

Tabla Producto:

```
CREATE TABLE "DimProducto" (
    "ID_producto" int identity(1,1) primary key,
    "codigo_Producto" nvarchar(15),
    "nombre_producto" nvarchar(70),
    "Categoria" nvarchar(50),
    "categoria_producto" int,
    "proveedor" nvarchar(50),
    "cantidad_en_stock" smallint,
    "precio_venta" numeric(15,2),
    "precio_proveedor" numeric(15,2)
   )
```

Tabla Cliente:

```
CREATE TABLE "DimCliente" (
    "ID_cliente" int,
    "nombre_cliente" nvarchar(50),
    "nombre_contacto" nvarchar(30),
    "apellido_contacto" nvarchar(30),
    "telefono" nvarchar(15),
    "pais" nvarchar(50),
    "ciudad" nvarchar(50),
    "codigo_postal" nvarchar(10)
)
```

Tabla de tiempo:

```
CREATE TABLE "DimDate" (
    IDDate int identity(1,1) primary key,
    "fecha_pedido" date,
    "Day" int,
    "Month" int,
    "Year" int,
    "NumWeek" int,
    "Quarter" int,
    "DayWeek" int
)
```

Tabla de hechos:

```
SELECT p.ID_pedido, dp.ID_detalle_pedido, fp.IDDate as fecha_pedido, p.estado,
dimCl.ID_cliente, dimP.ID_producto,
dp.cantidad, dp.precio_unidad, dp.cantidad * dp.precio_unidad AS total

FROM [dbo].[StagingPedido] p
inner join [dbo].[StagingDetallePedido] dp on p.ID_pedido=dp.ID_pedido
inner join [dbo].[DimDate] fp on p.fecha_pedido=fp.fecha_pedido
inner join [dbo].[DimProducto] dimP on dp.ID_producto=dimP.ID_producto
inner join [dbo].[DimCliente] dimCl on p.ID_cliente=dimCl.ID_cliente
```

Imagen de la estructura de la base de datos StagingJardineria

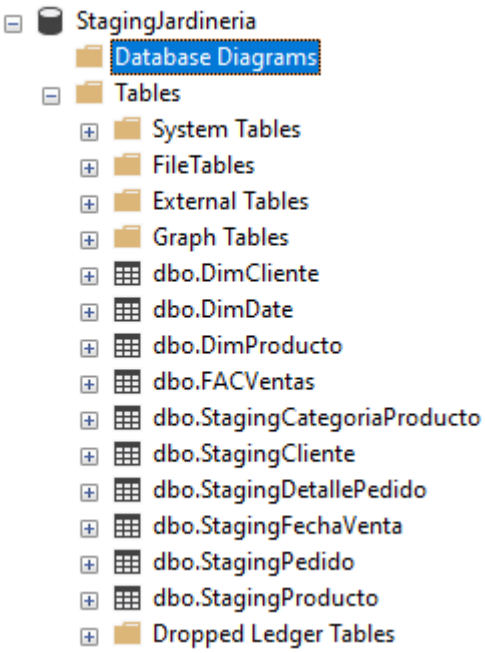
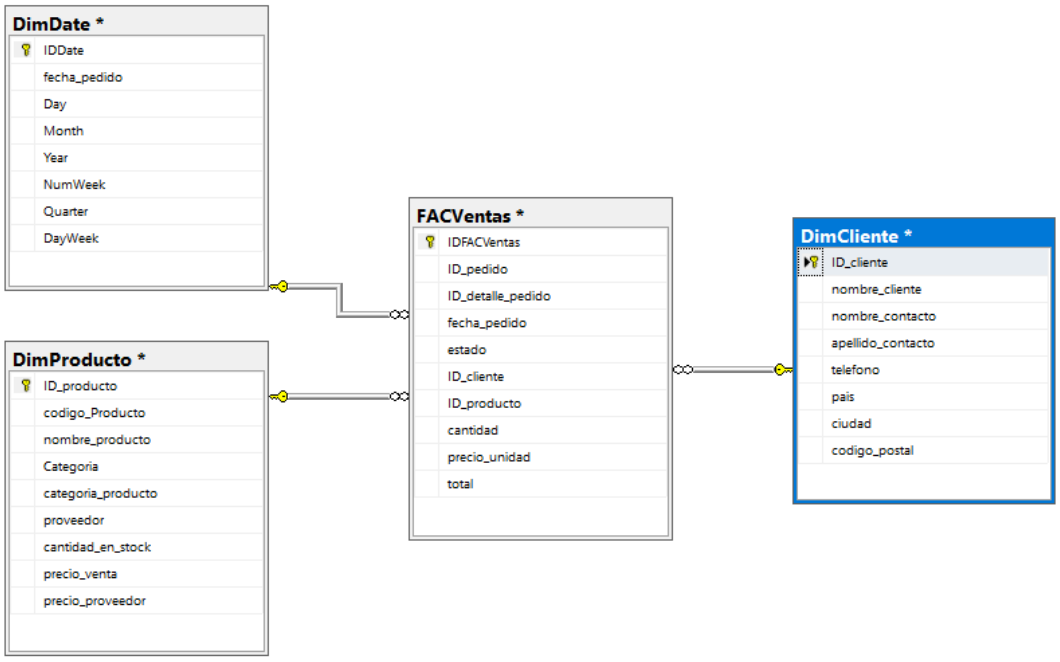


Diagrama del modelo dimensional en SSMS



Proceso de carga de las tablas de dimensión de la base de datos StagingJardinería a las tablas de dimensiones dentro de la base de datos Data Mart:

Dimensión Tiempo

```
CREATE TABLE "DimDate" (
    "IDDate" int primary key,
    "fecha_pedido" date,
    "Day" int,
    "Month" int,
    "Year" int,
    "NumWeek" int,
    "Quarter" int,
    "DayWeek" int
)
```

Dimensión Producto

```
CREATE TABLE "DimProducto" (
    "ID_producto" int primary key,
    "codigo_Producto" nvarchar(15),
    "nombre_producto" nvarchar(70),
    "Categoria" nvarchar(50),
    "categoria_producto" int,
    "proveedor" nvarchar(50),
    "cantidad_en_stock" smallint,
    "precio_venta" numeric(15,2),
    "precio_proveedor" numeric(15,2)
)
```

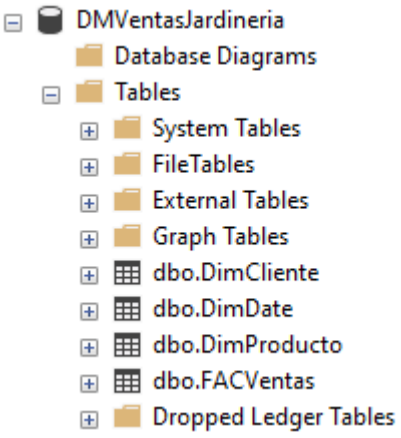
Dimensión Cliente

```
CREATE TABLE "DimCliente" (
    "ID_cliente" int primary key,
    "nombre_cliente" nvarchar(50),
    "nombre_contacto" nvarchar(30),
    "apellido_contacto" nvarchar(30),
    "telefono" nvarchar(15),
    "pais" nvarchar(50),
    "ciudad" nvarchar(50),
    "codigo_postal" nvarchar(10)
)
```

Tabla de hechos FAC

```
CREATE TABLE "FACVentas" (
    "IDFACVentas" int primary key,
    "ID_pedido" int,
    "ID_detalle_pedido" int,
    "fecha_pedido" int,
    "estado" nvarchar(15),
    "ID_cliente" int,
    "ID_producto" int,
    "cantidad" int,
    "precio_unidad" numeric(15,2),
    "total" numeric(26,2)
)
```

Imagen de la estructura de la base de datos DMVentasJardinería



Bibliografía:

Herramienta externa Lucidchart: https://www.lucidchart.com

Guía para la construcción del modelo de estrella: Zerpa, H., García, R., & Izquierdo, H. (2020). DATAMART BASADO EN EL MODELO STAR PARA LA IMPLEMENTACIÓN DE INDICADORES CLAVE DE DESEMPEÑO COMO SALIDA DE BIG DATA. *Universidad Ciencia Y Tecnología* , *24* (102), 47-54. https://doi.org/10.47460/uct.v24i102.342