**EVIDENCIA DE APRENDIZAJE 4. INTEGRACIÓN FINAL**

**OSSMAN MEJÍA GUZMÁN**

**GRUPO PREICA2401B010076**

**OLIMPO DE JESÚS MACEA HERNÁNDEZ**

**GRUPO** [**PREICA2401B010094**](https://iudigital.instructure.com/courses/15609)

**INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA**

**INGENIERIA DE SOFTWARE Y DATOS**

**BASES DE DATOS II**

**VICTOR HUGO MERCADO**

**MEDELLÍN**

**MAYO DE 2024**

**INTRODUCCIÓN**

Esta evidencia de aprendizaje es la continuación de las evidencias de aprendizaje 1, 2 y 3 que responde a los siguientes objetivos de aprendizaje:

1. Da cuenta del diseño de un modelo estrella para un data mart. El modelo estrella es una técnica de modelado de datos que busca, especialmente, optimizar los datos correspondientes a almacenes de datos y data mart. El esquema lógico de un modelo de estrella consta de una tabla central denominada hechos y diversas tablas de dimensiones. La tabla central de hechos, almacena datos susceptibles de análisis, mientras que las tablas de dimensiones almacenan datos utilizados para agrupar, clasificar o filtrar las medidas de la tabla de hechos. El buen diseño de los data mart, asegura la eficiencia de los procesos de inteligencia de negocios, por lo tanto, esta evidencia de aprendizaje permite desarrollar las competencias necesarias para su análisis y diseño dependiendo de las necesidades de la empresa.
2. Incluye el proceso de diseño y creación de la base de datos staging basada en el modelo relacional de la base de datos jardinería. Esto incorpora las siguientes tareas:
   1. Análisis de los datos almacenados en la base de datos Jardinería.
   2. Revisión de los datos almacenados en el Jardinería para identificar cuáles son relevantes y cuáles se deben trasladar a la base de datos Staging.
   3. Construcción de la base de Datos Staging
   4. Diseño de la estructura tablas que estarán en la base de datos Staging.
   5. Construcción de las consultas que permitan traer los registros de Jardinería a la Base de Datos Staging.
   6. Ejecución de las consultas y validar que los datos queden almacenados correctamente en la base de datos Staging.
   7. Construcción del backup de ambas bases de datos.
3. Desarrollar un proceso de transformación desde la base de datos origen, pasar Staging y luego hasta el data mart final, utilizando la base de datos de staging previamente creada. El objetivo es asegurar que los datos estén preparados para el análisis, incluyendo la identificación del producto más vendido y otros análisis relevantes para la toma de decisiones empresariales.

**OBJETIVO GENERAL**

Unificar todo el trabajo desarrollado en las demás evidencias en una sola entrega.

**OBJETIVOS ESPEFICOS**

* Implementar un flujo de ETL (Extracción, Transformación y Carga) que extraiga datos de la base de datos origen y los traslade al entorno de staging, asegurando la integridad y calidad de los datos durante todo el proceso.
* Diseñar y ejecutar un plan de extracción de datos desde la base de datos origen.
* Aplicar transformaciones necesarias en el entorno de staging para limpiar y normalizar los datos.
* Realizar validaciones y controles de calidad para asegurar que los datos en el entorno de staging sean precisos y consistentes.

**INSTRUCCIONES**

* Creación de repositorio de Git para todas las actividades
* Reúne todos los archivos y documentos relacionados con cada una de las actividades desarrolladas, incluyendo los informes, consultas SQL, scripts, diagramas y, ahora, también las pruebas de verificación de calidad de datos.
* Diseña una estructura de carpetas clara y organizada para almacenar todos los archivos en el repositorio de Git.
* Agrega archivos al repositorio
* Confirma los cambios

**PLANTEAMIENTO DEL PROBLEMA**

En las evidencias de aprendizaje 1, 2 y 3 se tomó como caso de estudio, una empresa de jardinería que carece de un sistema que permita realizar el análisis de sus datos de ventas y entender, por ejemplo, las tendencias clave que impulsan su negocio. En la evidencia de aprendizaje 1 se planteó el análisis y diseño de un data mart que dé respuesta a las necesidades del negocio para comprender procesos propios de las ventas como el producto más vendido, la categoría con más productos vendidos y el año con más ventas. Esta información ayudará al desarrollo de la inteligencia de negocios en la empresa.

Por su parte, en la evidencia de aprendizaje 2 surgió la necesidad de realizar el análisis y migración de los datos almacenados en la base de datos transaccional hacia una base de datos Staging, esta última definida en pocas palabras como un sistema que se encuentra entre las fuentes de datos y el data mart.

En esta evidencia de aprendizaje 3, se continuó en proceso de ETL específicamente en la transformación de los datos, se realizó el proceso de enfocándose específicamente en la transformación de los datos. Se realizó el proceso de limpieza de datos, que implicó eliminar duplicados y corregir valores inconsistentes y, de normalización y estandarización que consistió en convertir los datos a formatos consistentes y estandarizados.

Finalmente, en esta evidencia de aprendizaje se describe el proceso de ajustes de los datos, la carga en el datamart final y se adjunta el repositorio de evidencias en GitHub.

**ANÁLISIS DEL PROBLEMA**

Luego de diseñar el modelo de estrella para el data mart de la empresa de jardinería, en el cual se realizó un análisis detallado de los requisitos y las necesidades del negocio y del modelo relacional, fue necesario construir la base de datos staging que permitiera un correcto proceso de ETL hacia el data mart. En este sentido, se llevó a cabo el análisis de los datos almacenados en la base de datos Jardinería con el fin de identificar cuáles eran relevantes y cuáles debían ser trasladados a la base de datos Staging, según el modelo de estrella planteado.

Posteriormente, se realizó la construcción de la base de datos, teniendo en cuenta la estructura de sus tablas, el diseño de consultas SQL para trasladar los registros desde el origen y la ejecución de consultas y validaciones para certificar que los datos quedaran almacenados correctamente en la base de datos staging. Luego, se dio cuenta del proceso de transformación de los datos disponibles en la base de datos staging.

Finalmente, en esta evidencia de aprendizaje se describe el proceso de ajustes de los datos, la carga en el datamart final y se adjunta el repositorio de evidencias en GitHub.

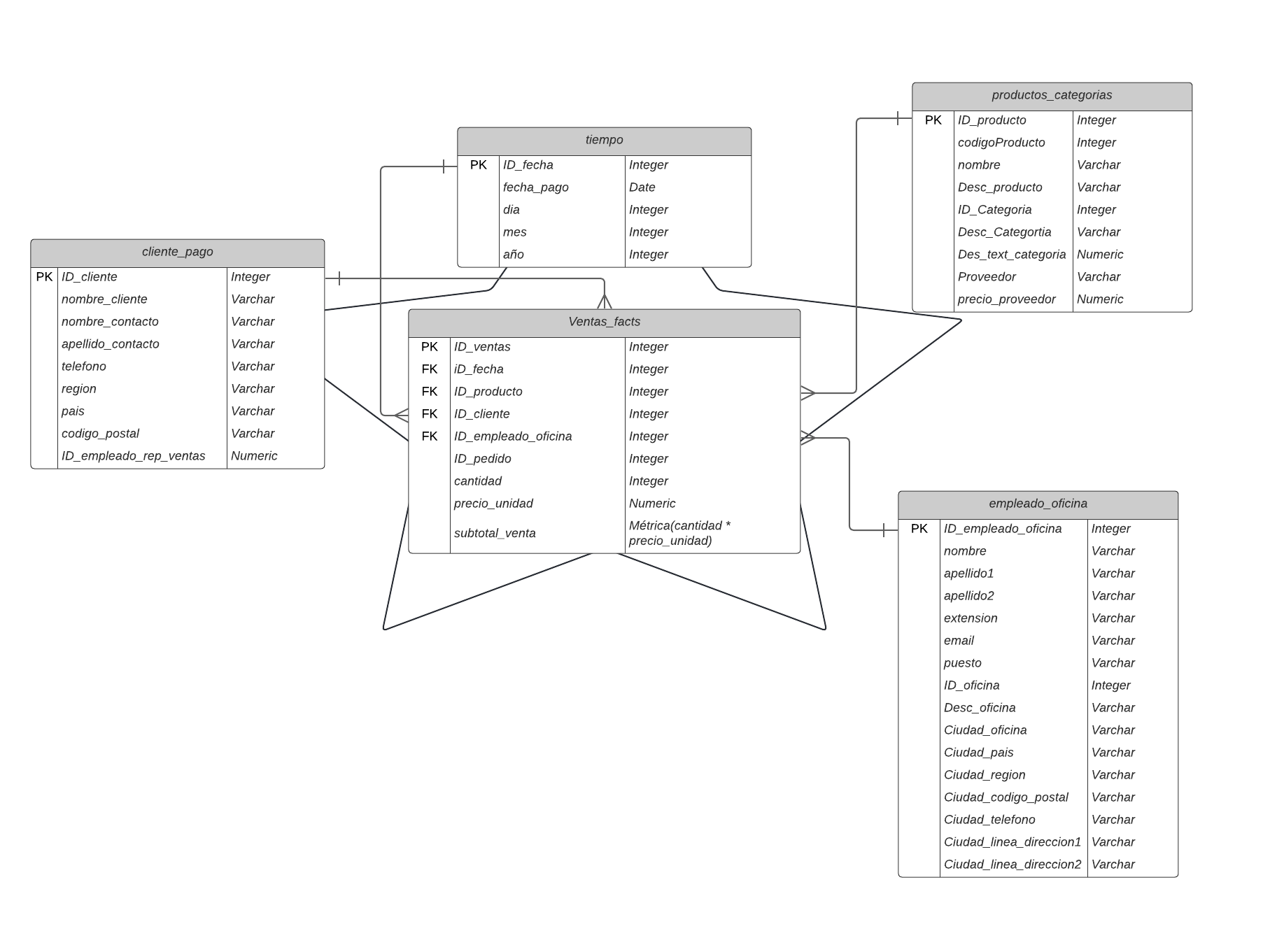
**DESCRIPCIÓN Y AJUSTES FINALES AL MODELO DE ESTRELLA**

Durante el proceso de carga de datos, se evidenció que no era necesario incluir la tabla de pagos lo que implicó modificar el modelo estrella planteado como se describe a continuación:

Este modelo cuenta con una tabla de hechos basada en las tablas *pedidos* y *detalles\_pedidos* del modelo relacional, ya que son las tablas que contienen la información de los pedidos generados con sus diferentes productos y cantidades. Adicional a ello, la tabla de hechos contiene dos campos cuantificables, cantidad y precio unidad para realizar las métricas de productos más vendidos y categoría más vendida.

Las dimensiones del modelo son: *tiempo* que permite consultar las ventas por periodos de tiempo específico, la dimensión *productos\_categoría* que agrupa los datos de las tablas *producto* y *categoría de producto*, la dimensión *cliente* que facilita la consulta de los clientes a los cuales están asociadas los pedidos y, la dimensión *empleado\_oficina* que agrupa los datos de las tablas *empleado* y *oficina.*

**DISEÑO DEL MODELO DE ESTRELLA FINAL**



**CARGA DE DATOS**

Para el proceso de carga de datos se crearon herramientas de *Data Flow task* para cada una de las dimensiones y la tabla de hechos. La siguiente ilustración presenta el Flujo de control final de todo el proceso ETL.



Ilustración 1. Flujo de control del proceso ETL de Jardinería

Al interior de cada Data Flow para la carga de las dimensiones y la tabla de hechos en el datamart, se realizaron tareas de orígenes de datos, ordenamiento (sort), merge join, división condicional (conditional split) y finalmente destino de datos a las respectivas tablas del datamart para las dimensiones y tabla de hechos. En la siguiente ilustración se ejemplifica este proceso para la dimensión de tiempo.

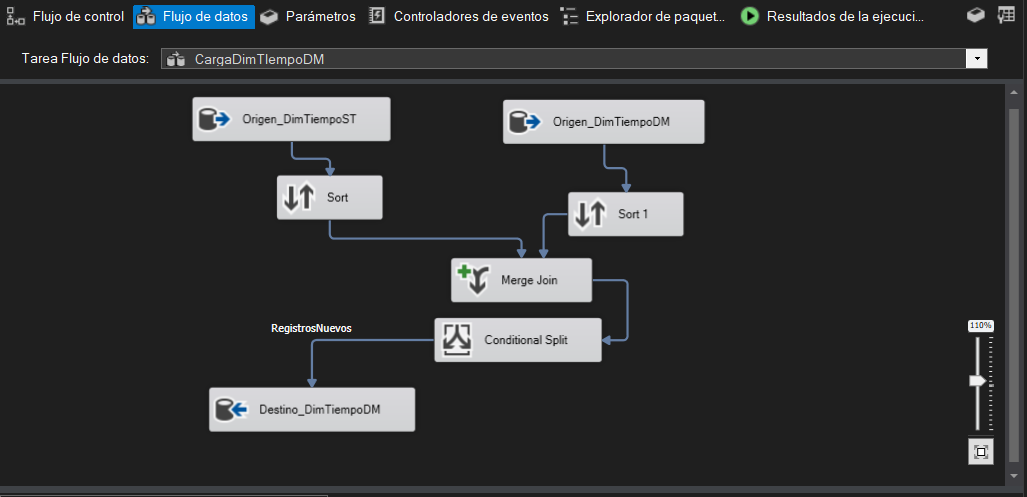


Ilustración 2. Flujo de datos para la carga de la dimensión de tiempo en el datamart.

Posterior a la carga de las dimensiones y la tabla de hechos, se establecieron las relaciones entre las tablas a través de Microsoft SQL Management Studio. La siguiente ilustración muestra el diagrama final del datamart incluyendo la relación entre las tablas de dimensiones y la tabla de hechos teniendo en cuenta las llaves.

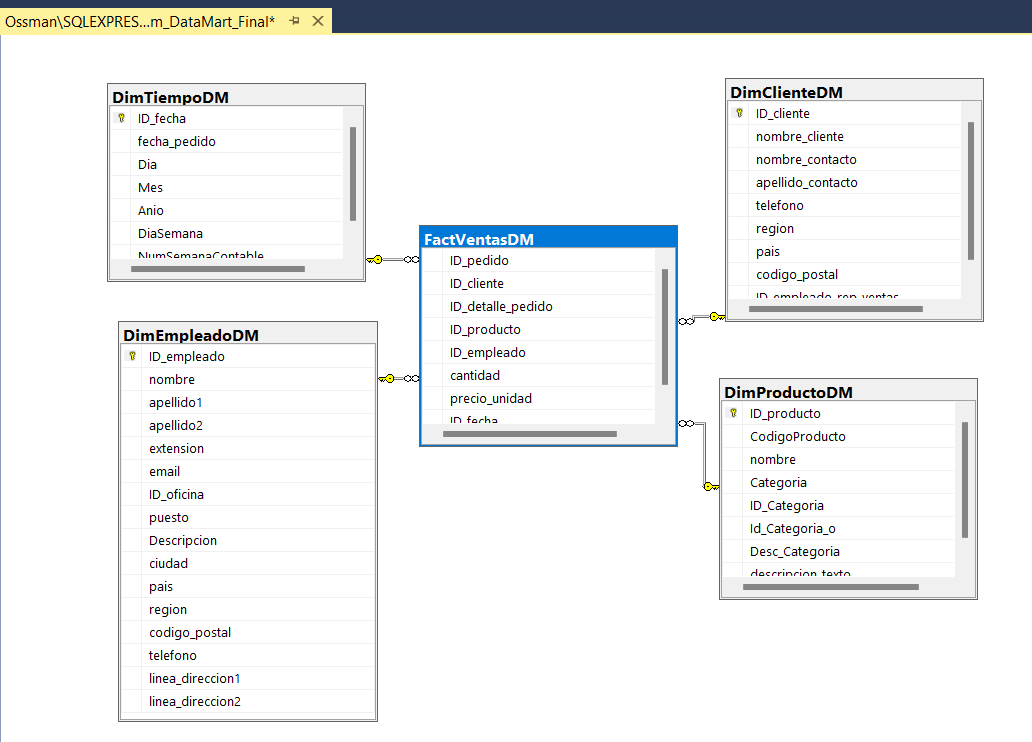


Ilustración 3. Diagrama del datamart de ventas de jardinería

**CONCLUSIONES**

El desarrollo del proceso de transformación de datos abordado en esta evidencia de aprendizaje ha permitido consolidar los conocimientos y habilidades necesarias para el diseño y ejecución de un flujo ETL (Extracción, Transformación y Carga). A través de las actividades realizadas, se han alcanzado los objetivos propuestos y se han puesto en práctica conceptos fundamentales del manejo de bases de datos y la inteligencia de negocios.

En primer lugar, se logró diseñar un modelo estrella para un data mart, lo que permitió optimizar el almacenamiento y acceso a los datos relevantes para el análisis empresarial. Este modelo no solo facilita la organización y estructuración de los datos, sino que también mejora la eficiencia de los procesos analíticos.

Asimismo, se desarrolló la base de datos staging, un paso intermedio crucial que garantiza la correcta preparación y limpieza de los datos antes de su carga en el data mart. Este proceso incluyó un análisis de los datos originales, la identificación de aquellos que eran relevantes para el negocio, y la implementación de técnicas de normalización y estandarización para asegurar su consistencia y precisión.

El proceso de transformación de datos desde la base de datos staging hasta el data mart final fue especialmente significativo, este paso no solo implicó la limpieza y corrección de los datos, sino también su reestructuración para que se ajustaran al esquema del modelo estrella. La creación de dimensiones como Tiempo, Cliente, Producto y Empleado, y la fact hechos de ventas, ejemplifican la capacidad de transformar datos en bruto en información útil para el negocio de jardinería.

En cuanto al proceso de carga de los datos, es importante destacar la relevancia de esta etapa dentro del flujo ETL. La carga final en el data mart ya se ha completado, y se puede observar que la carga de datos garantizó la integridad y disponibilidad de la información en el data mart. Esto implicó la implementación de estrategias de carga incremental asegurando que los datos más recientes estuvieran siempre disponibles para el análisis.

**REPOSITORIO EN GITHUB**

* **Jardineria**: Carpeta que contiene todo el proyecto de Integration Services de Microsoft Visual Community.
* **BD\_Backup**: Carpeta que contiene la copia de seguridad de las tres bases de datos utilizadas:
  + **Jardineria.bak**: Archivo de respaldo de la base de datos de jardinería.
  + **Staging\_jardineria.bak**: Archivo de respaldo de la base de datos staging para el proceso ETL del datamart.
  + **Jardineria\_DMVentasDataWirehouse**: Archivo de respaldo de la base de datos del datamart de ventas.
* **Evidencias\_Aprendizaje:** Carpeta que contiene los documentos que dan cuenta de las evidencias de aprendizaje 1, 2, 3 y 4. Con los respectivos anexos de consultas SQL, scripts y diagramas.
* **Enlace a repositorio GitHub del proyecto**: <https://github.com/ossmanmejia/BI_Jardineria>