Curso Introducción a ETLs con SQL e Integration Services (SSIS)

Danysoft
Haciendo visible lo invisible

ALEDIT
Servicios Profesionales Danysoft

Curso Introducción a ETLs con SQL e Integration Services (SSIS)

Danysoft
Haciendo visible lo invisible

ALEDIT
Servicios Profesionales Danysoft

Agenda

- **Módulo 1 – Proceso ETL**
  - Proceso ETL
  - Herramientas ETL
  - Fuentes de datos y Destinos
- **Módulo 2 – ETL con Programación SQL**
  - Scripts ETL
  - Transformaciones
  - Implementación de ETL con SQL
- **Módulo 3 – Procesamiento ETL con SSIS**
  - Creación de Proyectos SSIS
  - Programación de Paquetes SSIS
  - Ejecución de Tareas SQL
  - Implementación de Paquetes SSIS
- **Módulo 4 – Flujos de Datos SSIS**
  - Creación de Flujos de Datos
  - Transformación de Flujos de Datos
  - Opciones de Rendimiento
- **Módulo 5 – Despliegue y Solución de Problemas**
  - Gestión de Errores
  - Despliegue

Herramientas

- Microsoft SQL Server
  - Necesitará disponer de una instancia de SQL Server instalada, con derechos suficientes para la creación de una base de datos. En caso contrario, se puede instalar la edición Express o Developer, de descarga gratuita desde la web de Microsoft:
  - https://www.microsoft.com/es-es/sql-server/sql-server-downloads
- SQL Server Management Studio y Visual Studio Data Tools
  - Herramientas gratuitas descargables desde la web de Microsoft.
  - https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-2017
- Microsoft Excel o Microsoft Power BI Desktop.
  - Para la visualización de datos. Es posible trabajar con cualquiera de las dos herramientas.
  - Power BI Desktop: herramienta gratuita descargable desde la web de Microsoft. https://powerbi.microsoft.com/es-es/

Extracción Transformación Carga Explotación

Bases de datos
ERPs, CRMs

Hojas de cálculo
Access, O365

Ficheros texto
JSON, XML

Actividad web
SharePoint
Servicios nube
Redes sociales

Validación Datos

Limpieza Datos

Transformaciones

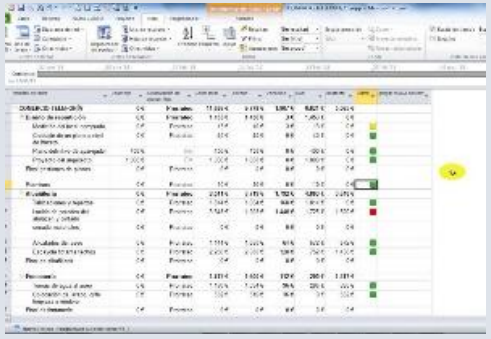
Modelado

Data Warehouse

Data Marts

Análisis de datos

Informes

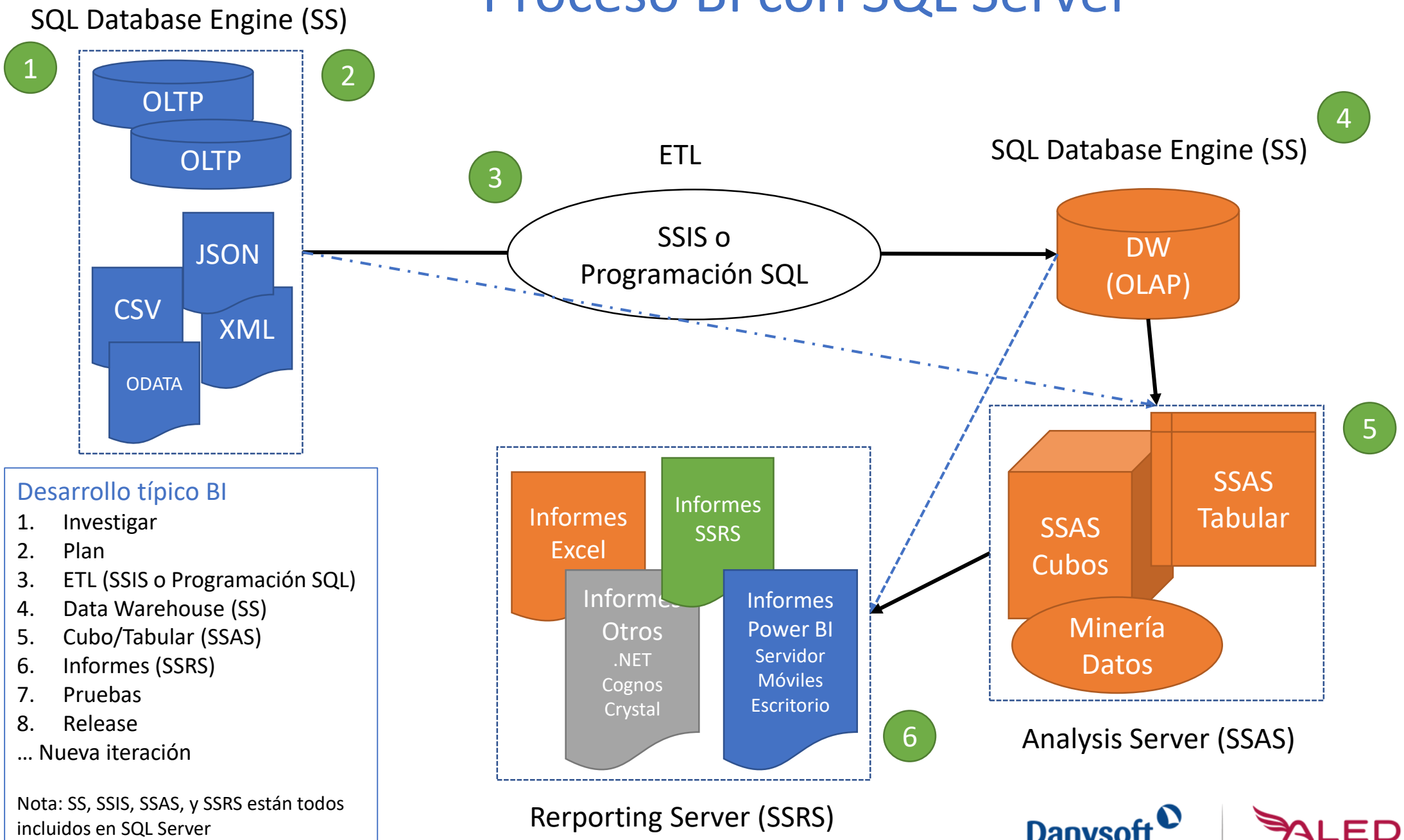


Cuadros de mando

Danysoft
Haciendo visible lo invisible

ALEDIT
Servicios Profesionales Danysoft

Proceso BI con SQL Server



Desarrollo típico BI

1. Investigar
2. Plan
3. ETL (SSIS o Programación SQL)
4. Data Warehouse (SS)
5. Cubo/Tabular (SSAS)
6. Informes (SSRS)
7. Pruebas
8. Release

… Nueva iteración

Nota: SS, SSIS, SSAS, y SSRS están todos incluidos en SQL Server

Danysoft
Haciendo visible lo invisible

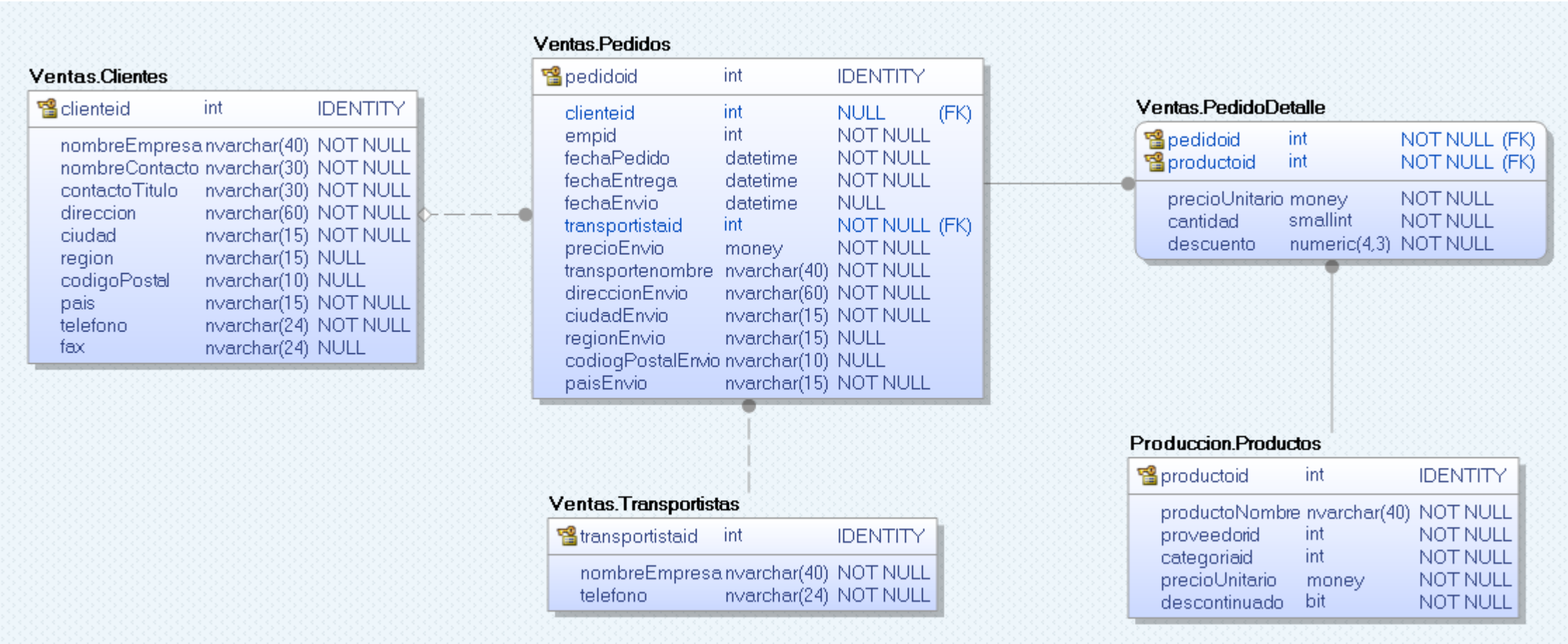
ALEDIT
Servicios Profesionales Danysoft

Documentación

- Diseño de la base de datos
  - Diagramas
    - de la BBDD de origen
    - de la BBDD dimensional
- Excel o World
  - Correlaciones de campos origen y destino

Orígenes de datos

- Base de datos SQL DB01ER



Danysoft Haciendo visible lo invisible
ALEDIT Servicios Profesionales Danysoft

Orígenes de datos

- Base de datos SQL DB01ER
  - Script de creación de DB01ER – 1-DB01ER.sql
    1. Crea la base de datos vacía DB01ER
    2. Crea los esquemas
       1. Ventas
       2. Produccion
    3. Crea las tablas
       1. Ventas.Clientes
       2. Ventas.Transportistas
       3. Ventas.Pedidos
       4. Ventas.PedidoDetalle
       5. Produccion.Productos
    4. Insertar registros en las tablas

Danysoft Haciendo visible lo invisible
ALEDIT Servicios Profesionales Danysoft

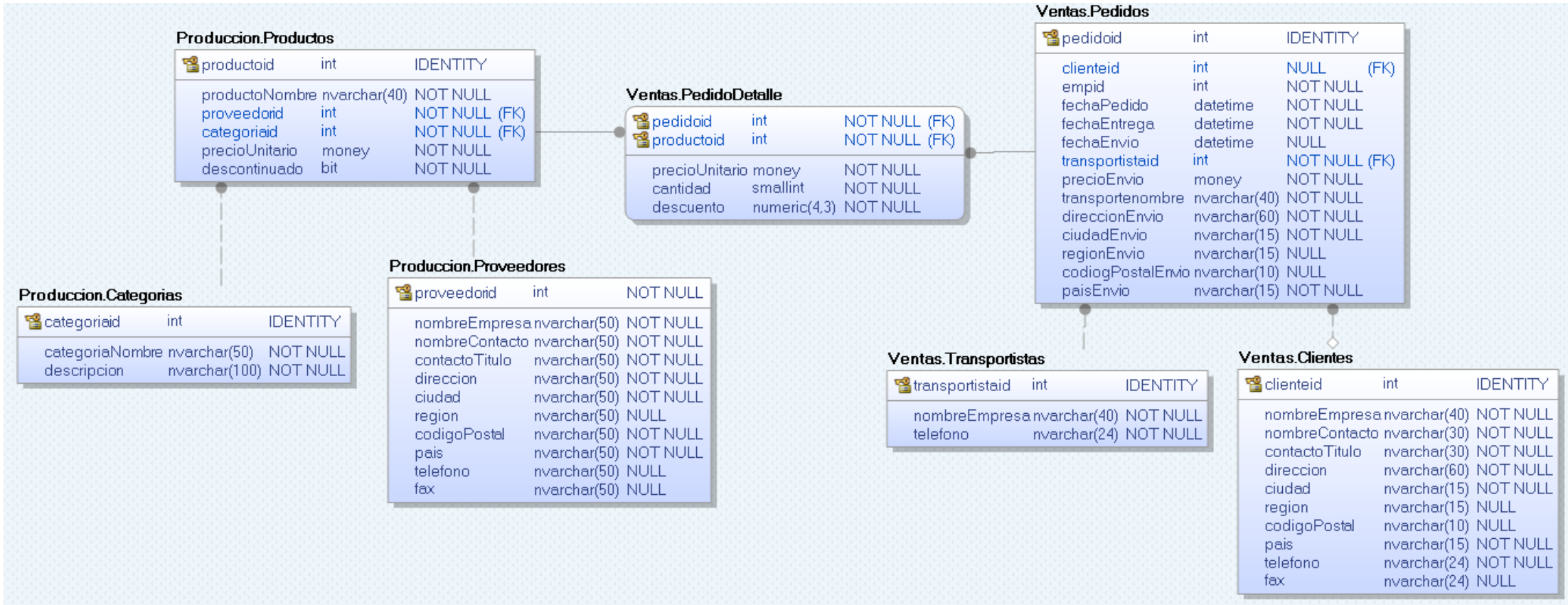
Orígenes de datos

- Importación de ficheros de texto y Excel con DTS
  - Categorias.csv
  - Proveedores.xslx
- Importación Script de carga 2-CargaFicherosExternos.sql
  - Crea las tablas:
    - Produccion.Categoria
    - Produccion.Proveedores
  - Inserta los registros desde los csv
    - Es necesario convertir el Excel en csv

Orígenes de datos

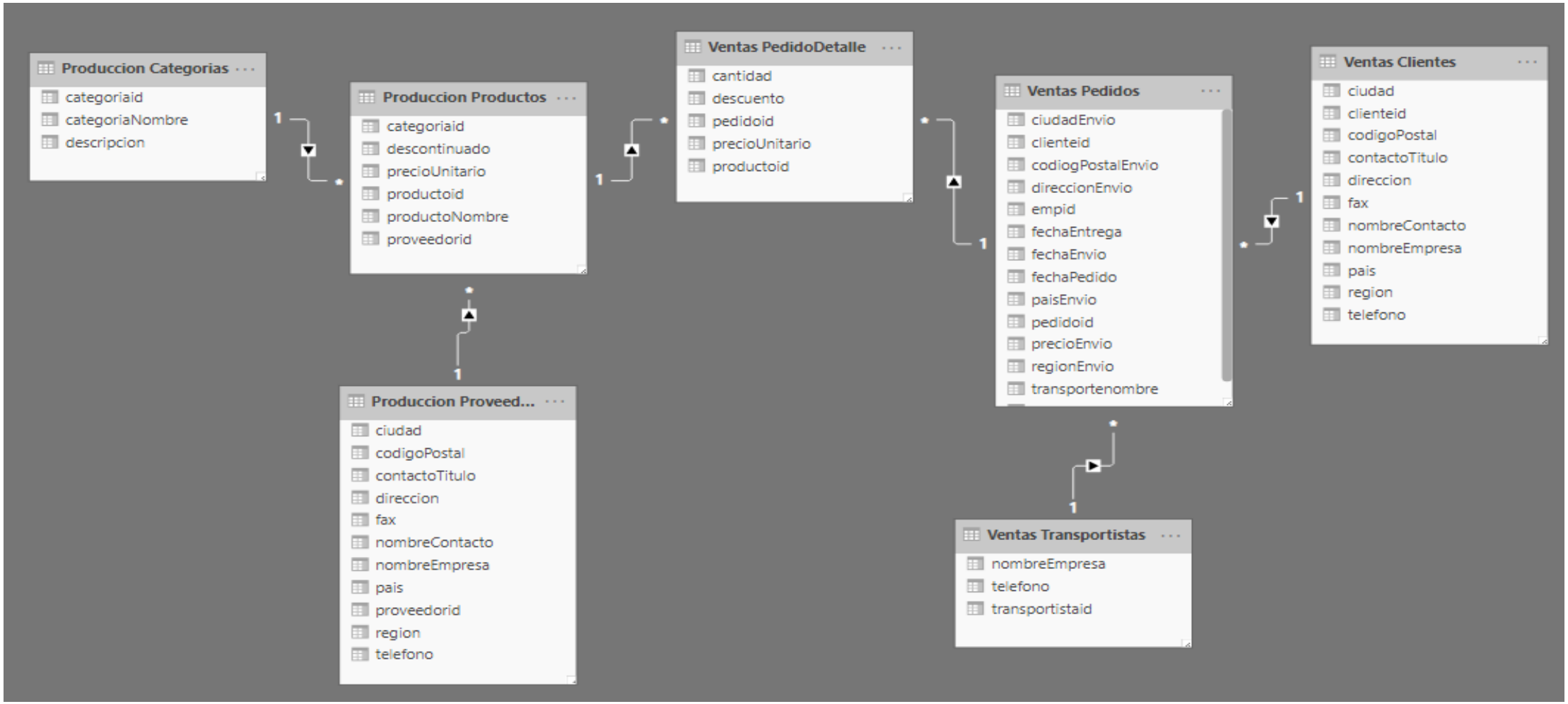
- Relacionar
  - Produccion.Categoria
  - Produccion.Proveedores
- Script “3-Indices y relaciones.sql”
  - Creación de Constraints:
    - Foreign key categoriaid en Productos
    - Foreign key proveedorid en Productos

Base de datos DB01ER



Danysoft
Haciendo visible lo invisible
ALEDIT
Servicios Profesionales Danysoft

Base de datos DB01ER



Danysoft
Haciendo visible lo invisible

ALEDIT
Servicios Profesionales Danysoft

Base de datos DB01DW. Data Warehouse

- Habitualmente se utiliza para representar el acceso a datos de los usuarios finales.
- Se suelen utilizar diagramas dimensionales. Un diagrama E/R representa cada posible proceso de negocio de una empresa, mientras en uno dimensional representa uno proceso de negocio determinado en una tabla de hechos.
- Es un tipo específico de diseño de entidad/relación, optimizado para las consultas utilizadas en bases de datos para el soporte de decisiones (Data Warehouses, Data Marts).

Danysoft Haciendo visible lo invisible
ALEDIT Servicios Profesionales Danysoft

Base de datos DB01DW. Data Warehouse

Modelo dimensional

- Hay varias tablas con las dimensiones y una tabla de hechos.
- La tabla de hechos contiene métricas para el negocio.
- Los hechos relacionados están en la misma tabla de hechos.
- Cada uno puede tener diferente granularidad – es el menor nivel de detalle que hay que tener en cuenta.

Danysoft
Haciendo visible lo invisible
ALEDIT
Servicios Profesionales Danysoft

Base de datos DB01DW. Data Warehouse

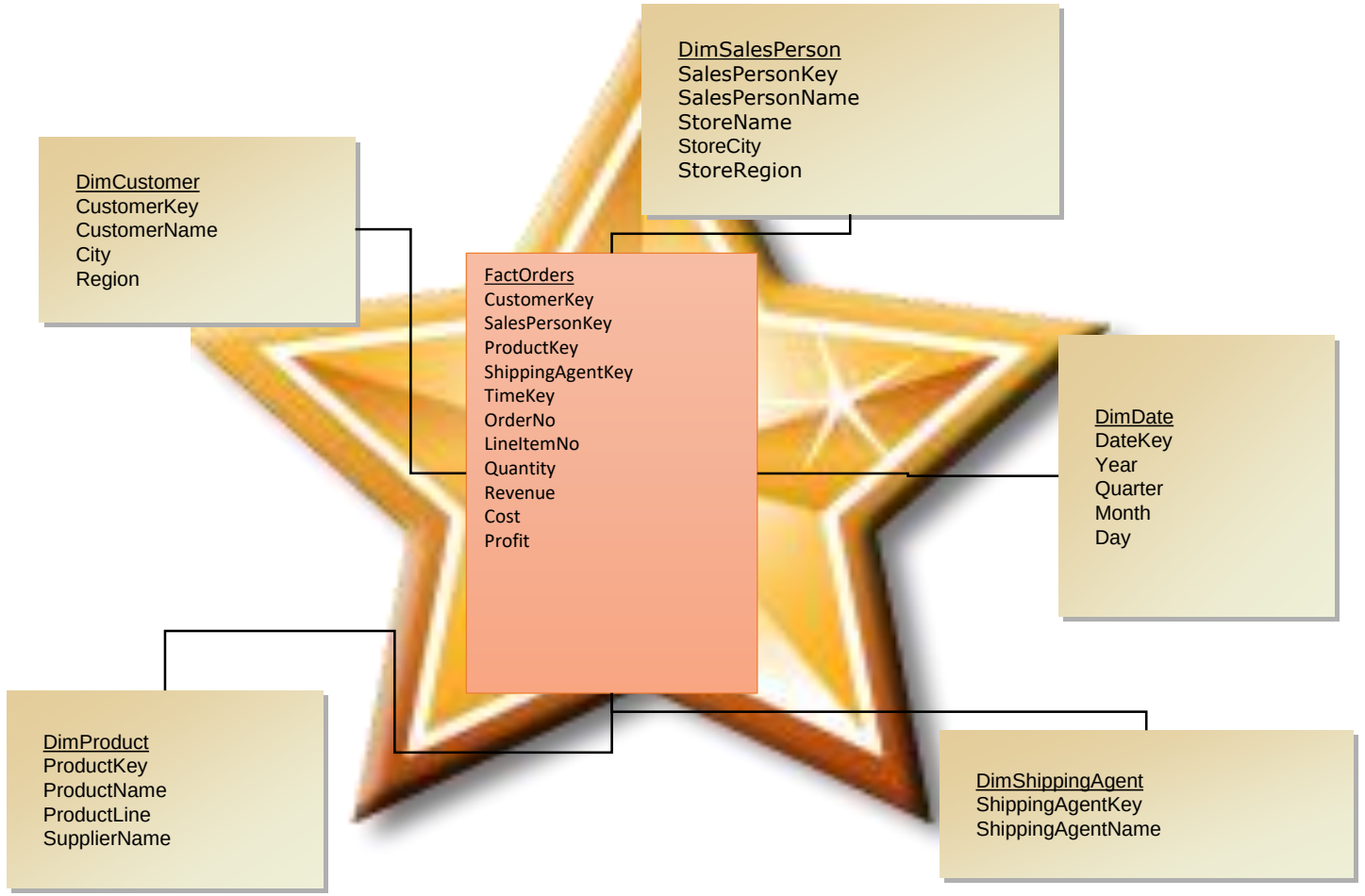
- Tabla de hechos.
  - Son la tabla central del esquema, contienen registros individuales, con lo datos para los cuales se hace el diseño.
  - Tiene una o más claves foráneas y no tiene hijos.
  - Tiene todas las claves que las relacionan con cada dimensión. Y su clave primaria es la suma de todas.
  - Se pueden definir de varios tipos:
    - Aggregate – Contiene información menos detallada. Por ejemplo, en transaccional tienes tablas con toda la información de ventas, y en dimensional una tabla agregada con la información de los totales de ventas por mes y por tipo de artículo.
    - Atomic – Información detallada propia de una tabla transaccional
    - Cumulative – Información acumulativa, como acumulación de tiempos para un proceso.
    - Snapshot – Información relacionada con el tiempo que detalla los pasos de un proceso, por ejemplo, con un pedido, cuándo se creó, se envió, entregó, etc.

Base de datos DB01DW. Data Warehouse

- Tabla de dimensiones
  - Son las tablas padres de la tabla de hechos.
  - Contienen grupos de datos relacionados, representados por una clave
  - Se pueden definir varios tipos:
    - Dimensión fija – no se espera que cambien los datos
    - Degenerada – es una tabla derivada de una de hecho. Tienen un nivel de granularidad parecida a la transaccional.
    - Multi-valor – se utilizan para modelar una situación donde hay múltiples valores para una columna o atributo. En un diseño relacional correcto cada columna ha de tener un solo valor, pero aquí puede ser útil.
    - Desordenada – El padre lógico inmediatamente superior hay veces que no está. Por ejemplo: Comunidad, provincia, ciudad, localidad, pero en algunos casos, hay saltos en la jerarquía.
    - Reducida – es una versión de la tabla de hechos que se adjunta pero solo con unos pocos atributos

Base de datos DB01DW. Data Warehouse

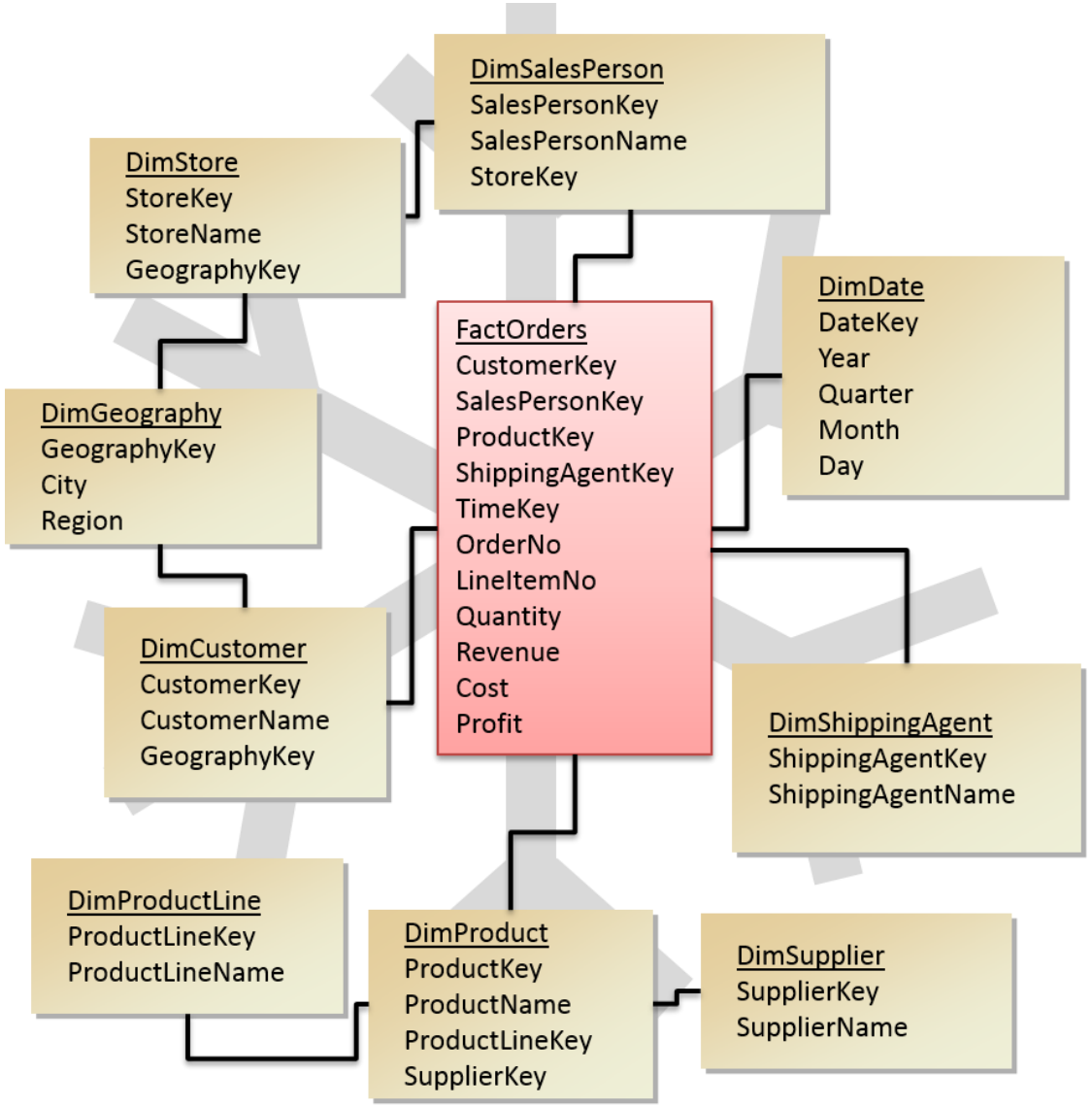
- El diseño de estrella es el de mayor rendimiento
  - Tiene una sola tabla para cada dimensión
  - Cada tabla soporta todos los atributos para esa dimensión
  - Es una solución denormalizada



Danysoft
Haciendo visible lo invisible
ALEDIT
Servicios Profesionales Danysoft

Base de datos DB01DW. Data Warehouse

- El diseño de copo de nieve, es dimensional pero más normalizado, y por tanto tiene una pérdida de rendimiento.
  - Tiene varias tablas para cada dimensión
  - Cada tabla tiene una clave de dimensión, valores, y la clave foránea para la tabla padre.

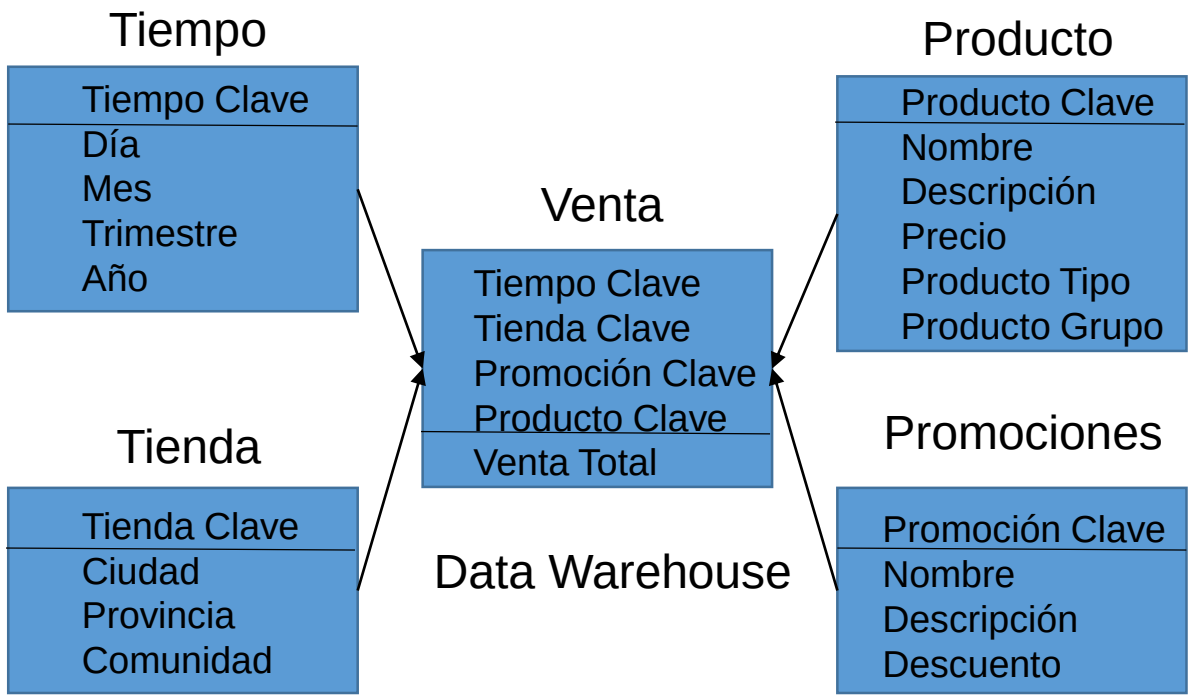


Danysoft
Haciendo visible lo invisible
ALEDIT
Servicios Profesionales Danysoft

Base de datos DB01DW. Data Warehouse

Claves Subrogadas (Surrogate Keys)

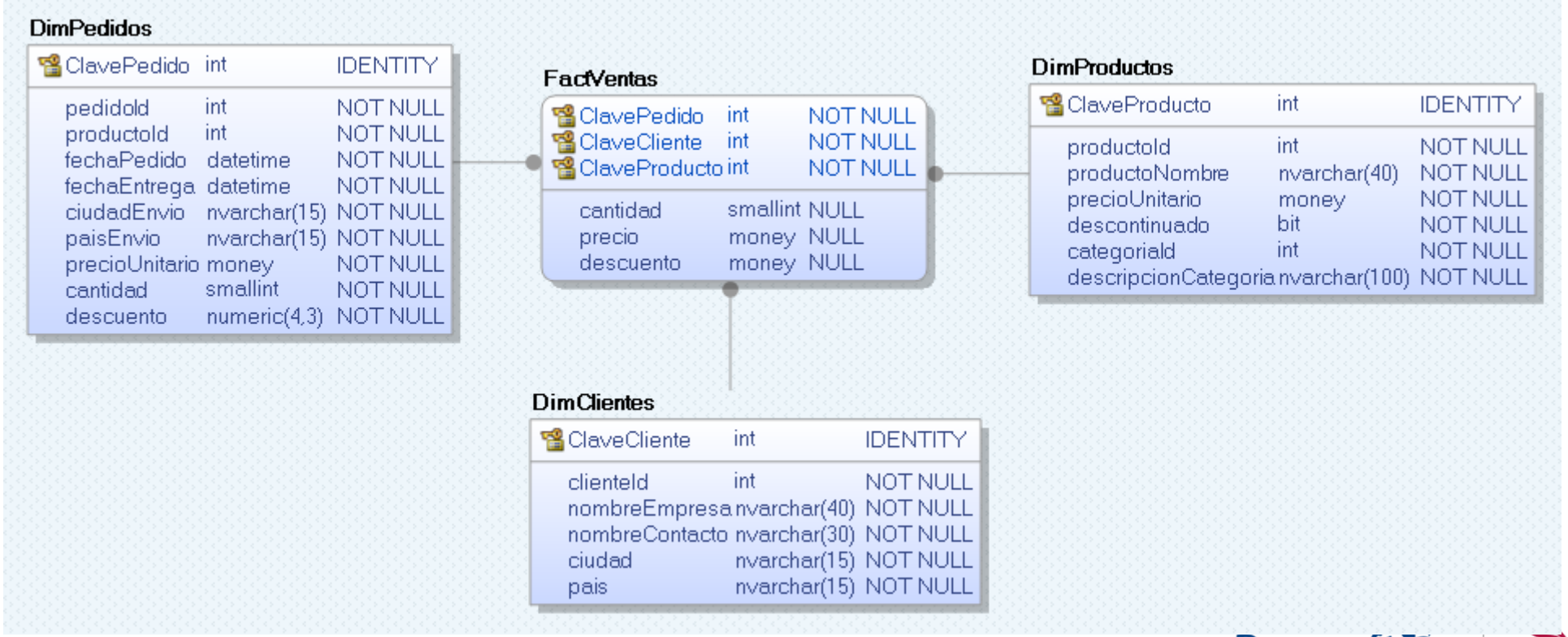
- Es un concepto muy utilizado en el diseño de bases de datos, especialmente en entornos de Data Warehouse (DW) y Business Intelligence (BI).
- Se utilizan en tablas de dimensión versionadas o históricas, conocidas como Slowly Changing Dimension (SCD) de tipo 2, es decir, tablas de dimensión que almacenan tanto los datos actuales (versión actual) como los datos históricos (versiones antiguas).
- La Clave de Negocio tiene un sentido: DNI, email, etc. La Clave Subrogada es un valor numérico único para cada fila de la tabla, actuando como una clave sustituta.



Danysoft Haciendo visible lo invisible
ALEDIT Servicios Profesionales Danysoft

Base de datos DB01DW. Data Warehouse

- Scrip de creación de base de datos DB01DW y tablas: 4-DB01DW.sql
  - Inclusión de una Clave Subrogada que no existe – IDENTITY
  - Tablas de dimensiones desnormalizadas que agrupan datos originarios de diferentes tablas



Danysoft Haciendo visible lo invisible
ALEDIT Servicios Profesionales Danysoft

“Flush and Fill” Versus carga Incremental

Limpiar y rellenar

- Se borran todos los datos de las tablas de destino y se vuelven a rellenar con los datos actuales.
- Script “5-Limpiar y Rellenar.sql” (Vendedor y DimVendedor)

Carga incremental

- Se insertan, actualizan o borran datos en las tablas de destino que han cambiad en el origen, normalmente en base a filas.
- Script “6-Carga Incremental.sql” (Vendedor y DimVendedor)

Danysoft Haciendo visible lo invisible
ALEDIT Servicios Profesionales Danysoft

“Flush and Fill” Versus carga Incremental

Carga Incremental. Script “7-Carga Incremental con Merge.sql”

- Funciona bien con tabla grandes, o cuando quieres preservar los valores originales de los datos
- El método más eficiente es el uso del comando SQL Merge.
  - Merge es más eficiente que el comando Except porque en ambos casos va línea a línea aplicando la lógica. En expect, primero hace todas las líneas para Insert, luego para Update y luego para Delete por lo que hace tres pasadas, y Merge solo hace una pasada evaluando las tres situaciones a la vez.
  - Merge solo está disponible a partir de MSSQL 2012. En otras bases de datos habrá que utilizar Except.
- La carga incremental necesita más programación que Fush and Fill. Si se puede mantener simple, es mejor Limpiar y rellenar.

Base de datos DB01DW. Data Warehouse

Slowly Changing Dimension (SCD):

- SCD Tipo 1. Los cambios se implementan de forma que no se puede realizar un seguimiento de cambios. Se utiliza si el seguimiento de cambios no es relevante. Ejemplo anterior "7-Carga Incremental con Merge.sql"
- SCD Tipo 2 (Tablas Versionadas). Los cambios se implementan manteniendo múltiples filas en la tabla para cada Clave de Negocio, de las cuales sólo una será la fila actual. Es posible realizar un seguimiento de cambios ilimitado, pues podremos insertar ilimitadas filas. "8-SDC Tipos 2 y 3.sql"
- SCD Tipo 3. Los cambios son implementados agregando nuevos campos a la tabla. Para un campo que pueda cambiar, se puede utilizar un campo para el valor actual y otro campo para el valor anterior. En este caso, no se utilizan múltiples filas, aunque por el contrario se utilizarán múltiples campos. Es posible realizar un seguimiento limitado de los cambios, y dicho límite será impuesto por el número de campos utilizados para el seguimiento de cambios.

Preparación de los datos

- Vamos a utilizar objetos de BBDD de SQL para abstraer la capa de datos que seleccionamos de la BBDD de origen, para cargar en la BBDD de destino
  - Los objetos que se utilizarán serán
    - Vistas
    - Procedimientos almacenados
- Primero debemos hacer las Transformaciones, pero para saber cuáles hemos de hacer, antes hay que haber hecho un diseño previo del DW

Transformaciones

Script “9-Transformaciones.sql”

Primero hacemos los Select

- Renombrar columnas para legibilidad
- Dividir y juntar columnas
- Conversión de tipos para consistencia
- Combinar datos desde múltiples tablas
- Etc.. Las posibilidades de programación son las que marcan la cantidad de transformaciones que podemos hacer.

Capas de abstracción

Script “10-CapasAbstraccion.sql”

- Crear capas de abstracción es una buena práctica recomendada para seleccionar y mover datos
- Las capas son objetos de bases de datos:
  - Vistas
  - Procedimientos almacenados
  - Funciones

Danysoft Haciendo visible lo invisible
ALEDIT Servicios Profesionales Danysoft

Limpiar y rellenar con TRUNCATE

- Hay que considerar si utilizar DELETE o TRUNCATE

Script “10-CapasAbstraccion.sql”

  - Si utilizamos Delete, quitando todos los registros de la tabla y volviéndola a llenar, la clave subrogada comenzará no por 1 sino por el número siguiente al último que había. No pasa nada ya que es un número artificial.
  - Pero con Truncate, se resetea el número de la clave.
  - Además Truncate es más rápido.

Automatización de Tareas

- Es necesario activar el SQL Server Agent
  - Se pueden automatizar mediante Jobs para los procedimientos almacenados la carga de datos, programando los tiempos de ejecución

Danysoft
Haciendo visible lo invisible
ALEDIT
Servicios Profesionales Danysoft

Danysoft
Haciendo visible lo invisible

ALEDIT
Servicios Profesionales Danysoft

Contacto:

www.danysoft.com

info@aledit.com | info@danysoft.com

Teléfono: 902 123 146 | [+34] 916 638 683

Avda. de la industria 4, edif. 1 | 28108 Alcobendas | Madrid | España