Bases de datos II

22/02/2021 - 25/02/2021

Apuntes de la clase 22/02/2021 - 25/02/2021

IS-601

Emilson Acosta

Clase #1

Inteligencia de negocios: Proceso que se realiza para poder obtener, analizar y presentar información a los usuarios, permitiéndoles tomar mejores decisiones que se realizan de formas iterativas, es decir se puede ejecutar continuamente para poder actualizar la información que se obtiene, se analiza y se presenta a los usuarios, se compone de varios elementos:

- Sistemas de información geográfica
- Sistemas de relaciones con el cliente
- Sistemas de administración del conocimiento
- Sistemas de soporte a la toma de decisiones
- Data Mining
- OLAP (Procesamiento analítico en línea) o Cubos OLAP
- Data Warehouse
- Visualización

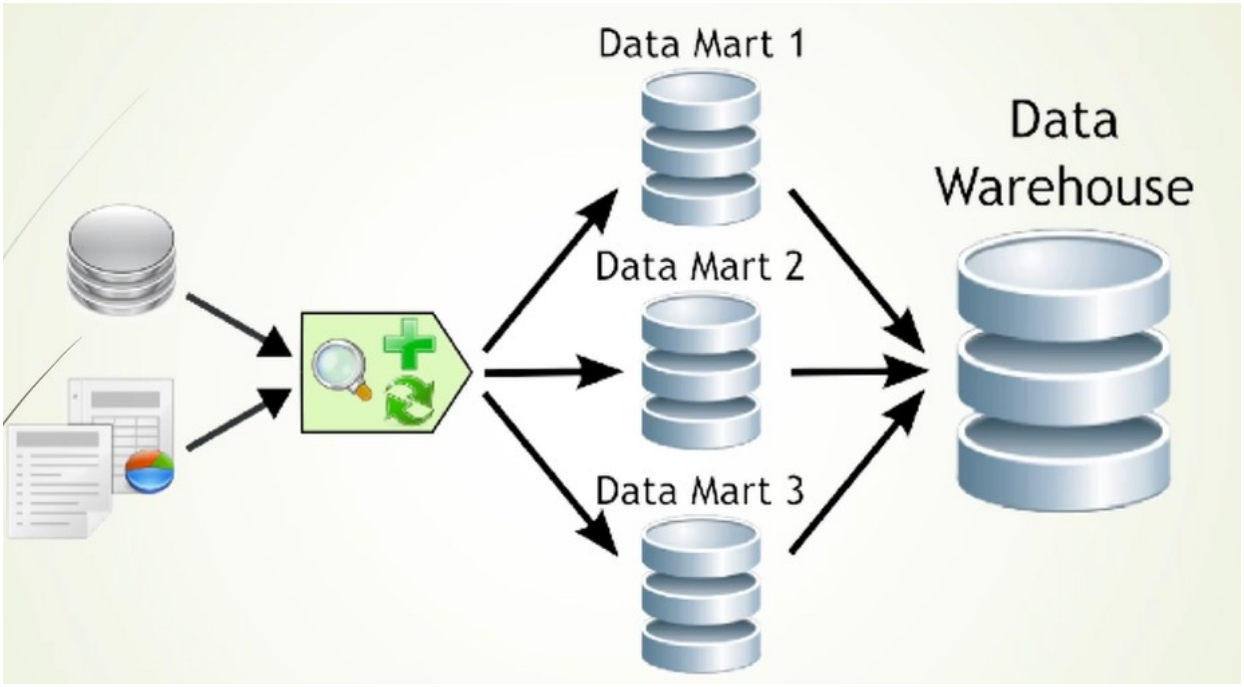
Centrados en Data Warehouse, cubo OLAP, visualización en conjunto de la herramienta DSS

Data WareHouse: Base de datos en la que se almacena una gran cantidad de información, que se extrae de distintas fuentes, como distintos gestores de bases de datos, archivos de texto plano o Excel. Estos datos obtenidos deben ser oportunos, confiables, precisos y concisos para ayudar a la toma de las decisiones. Se compone de uno o distintos data mart.

Data Mart: Aspecto o vista del negocio. Se llenan con los datos obtenidos de las distintas bases de datos. Ejemplo:

- Data mart de ventas
- Data mart de recursos humanos

- Data mart de desempeño organizacional
- Data mart de presupuesto



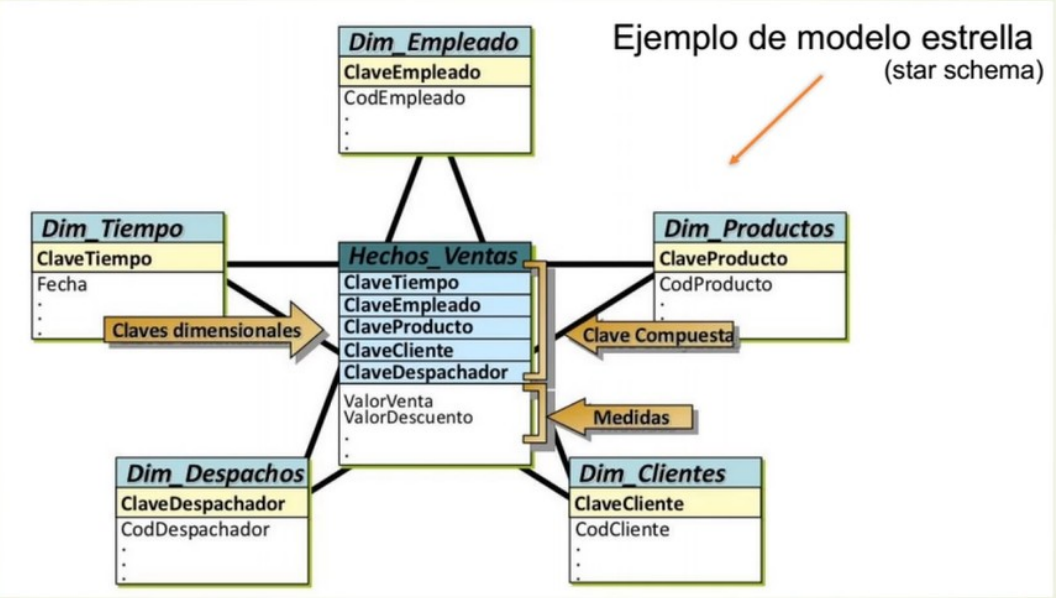
Para construir los data mart se deben tener en cuenta dos conceptos:

- **Tablas de hechos** Tabla en la que se almacenan llaves foráneas que hacen referencia a otras tablas y los otros campos son las métricas, que son valores numéricos que permiten analizar la información y tomar decisiones, Sus únicos campos son **la llave primaria, las llaves foráneas y las métricas.** ejem: cantidad total de ventas de un producto, cantidad de artículos enviados a una región, presupuestos, etc.

- **Tabla de dimensiones** Información que le dará sentido a las métricas. Representa la información a través de la cual se desea medir los hechos, se representan con tablas de dimensiones pueden ser de distintas dimensiones como: Tiempo, regiones, productos, modelos, etc. **Modelos de un Data Warehouse**

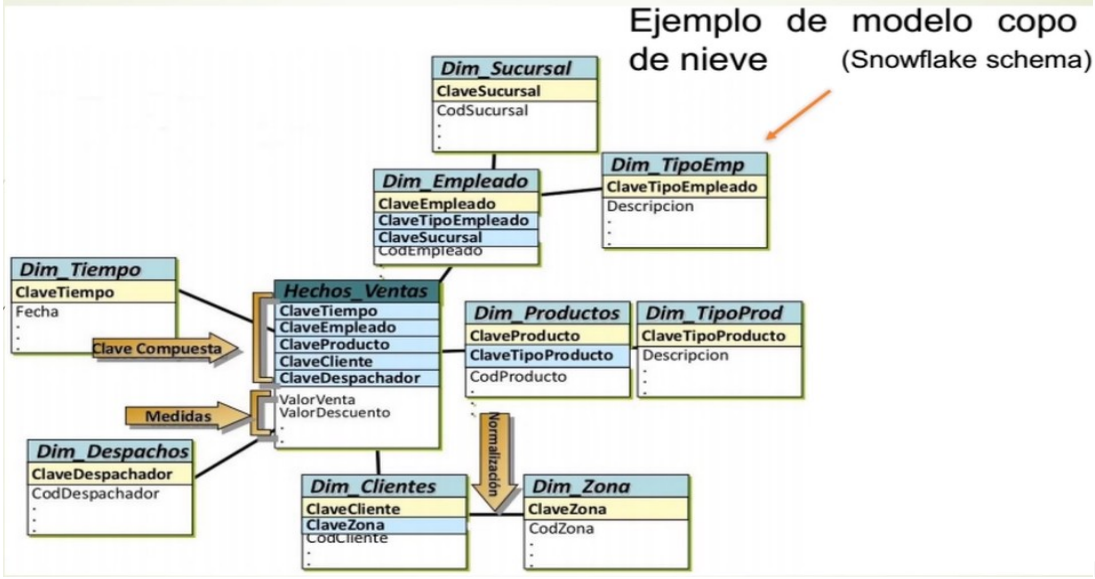
Modelo en estrella (star schema)

La tabla central es la de hechos, las tablas aledañas son las de dimensiones, en este caso la llave primaria se compone de varios campos que a su vez son campos que hacen referencia a las tablas de dimensiones. En las tablas de hechos debe haber al menos una metrica.



Modelo en copo de nieve

La diferencia con el modelo estrella es que las tablas de dimensión pueden estar relacionadas con otra tabla de dimensión sin relacionarse con la tabla de hechos.

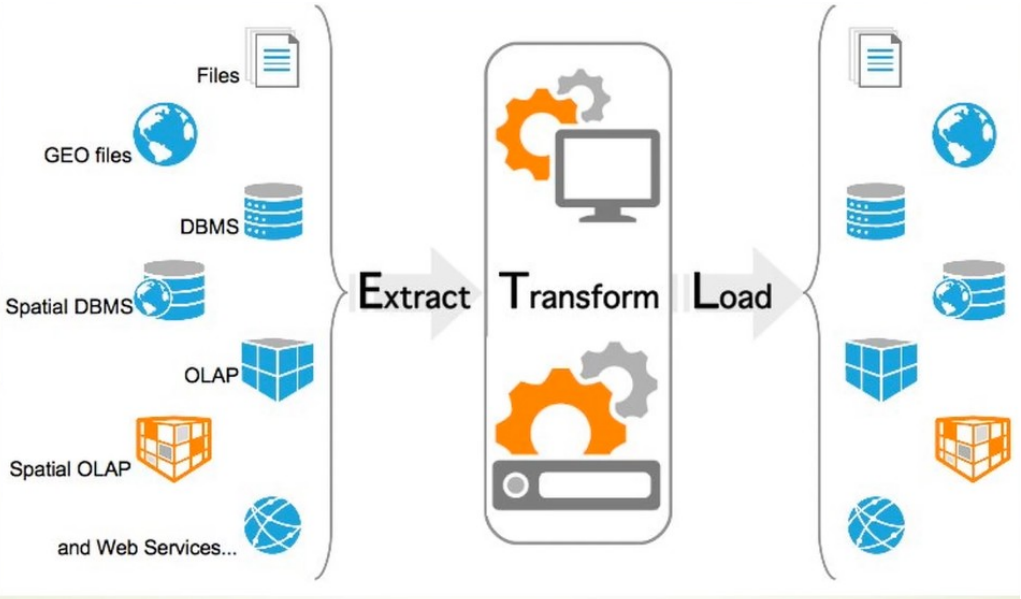


m

En ambos modelos no importa la normalización, es decir no debería de estar normalizado, para pasar del modelo copo de nieve al modelo en estrella lo que se debe hacer es agregar las dimensiones extra en la dimensión principal.

Extracción, transformación y carga de datos – ETL

Es el que se encarga de obtener los datos de las distintas fuentes, transformarlos y poder guardar o insertar los datos en el modelo de estrella o copo de nieve. La transformación de los datos implica poder concatenar un valor con otro, de la fecha extraer el día, mes o año, indicar a que trimestre corresponde, etc.

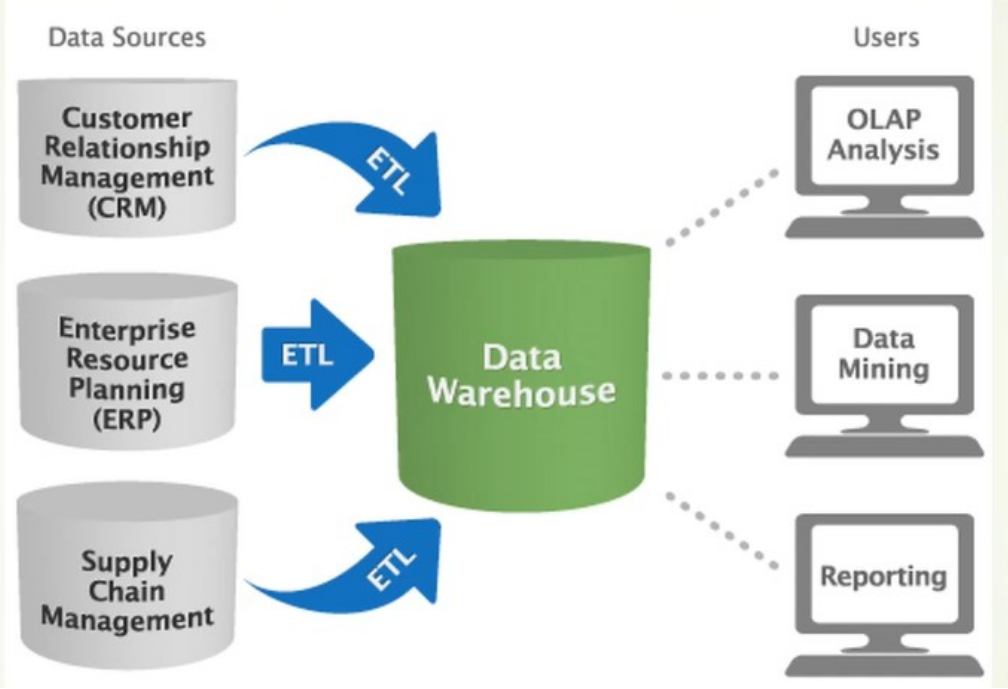


Cubos OLAP

Es una forma de visualizar la información, OLAP significa procesamiento analítico en línea, la información que se tenga se va a interpretar dependiendo de la perspectiva o vista que se tenga del cubo, permite ver la información desde distintas perspectivas, cada celda del cubo representa una métrica.

Herramientas para construir un Data Warehouse

- Integration services → Herramientas para construir ETL
- Analysis Services → Herramientas para construir el cubo OLAP
- Report Services → Herramienta para construir reportes interactivos



Sistema de soporte a la toma de decisiones

Herramienta de BI (Bussines Inteligence) la que permite realizar el análisis de los datos y luego mediante una forma gráfica poder ver la información, interpretarla y tomar decisiones, dicha información será obtenida de los cubos OLAP, los reportes serán interactivos. Un ejemplo de un sistema DSS es **Penataho.** La ventaja de usar Pentaho es que se puede visualizar la información en forma de tabla, en forma de gráficos, se pueden aplicar filtros y distintas configuraciones para que el reporte sea lo más adecuado, preciso, conciso y confiable posible.

Otra herramienta es el **Report Services de SQL Server** se pueden crear una diversidad de reportes, donde se puede escoger el sistema DSS se utilizará para poder crear los reportes y presentarlos a los usuarios.

Clase #2

Pasos para construir un datamart:

1. Obtener las preguntas que se desean saber del negocio
2. Identificar las métricas que generan las preguntas del negocio
3. Identificar las tablas de dimensiones y hechos
4. Decidir que modelo se usará
5. De la base de datos OLTP identificar que tablas servirán para llenar las de hecho y dimensiones, en base a esto se construirá el ETL.
6. Analizar con que campos del OLTP se obtendrán las métricas

- Obtener las preguntas que se desean responder para el negocio
- Identificar la o las métricas que generan las preguntas del negocio
- Identificar las tablas de dimensiones
- Identificar la o las tablas de hechos
- Decidir que modelo se utilizará para diseñar el data mart
- De la base de datos OLTP se debe identificar las tablas que servirán para llenar las tablas de hechos y las tablas de dimensiones
- Analizar con qué campos de la base de datos OLTP se obtendrá la métrica

UNAH

Ejercicio

1- **Preguntas de negocio**

- Se desea analizar cuál es el **total de ventas de los productos** en base a los empleados
- Las **ventas** se deben analizar por año, mes, número de semana del año, trimestre del año y por el nombre del día de la semana.
- Es importante conocer el código y nombre completo del empleado que realiza las ventas
- El **total de ventas** también se puede conocer en base al nombre del producto
- Conocer cuál es el total de ventas que cada empresa transportista ha trasladado

2- **Identificar las métricas**

Total de ventas de productos

Ventas

Total de ventas

3- **Identificar tablas de dimensión**

A partir de las preguntas del negocio, cuando no se especifican campos queda a criterio personal los campos que se utilizaran:

1. Empleados

2. Tiempo

3. **Campos de la dimensión empleado (código y nombre)**
4. **Producto**
5. **Empresa transportista**

4- Campos de las dimensiones

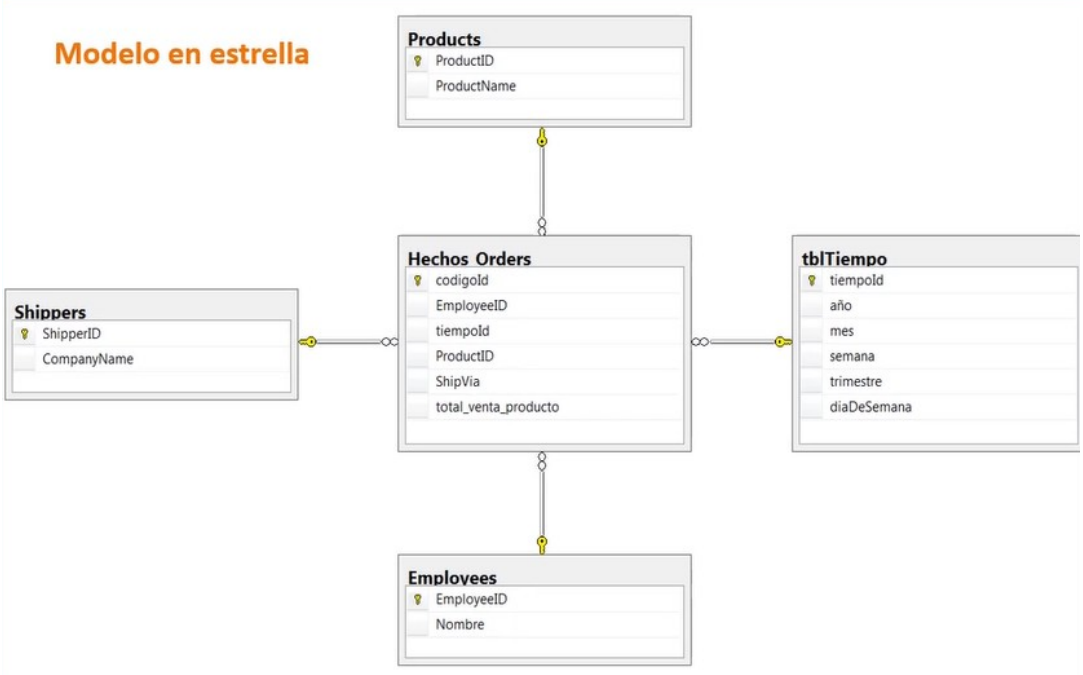
5- Identificar la tabla de hechos

Según las preguntas del negocio la metrica es **Total venta productos,** la tabla se puede llamar hechos ventas o hechos ordenes, en la tabla de hechos se tienen las llaves foráneas de las tablas de dimensiones. Se tomaran los siguientes campos como llaves foráneas:

Campos de la tabla de hechos

- Código único de registro (autoincremental)
- Código producto (llave foránea)
- Código empleado (llave foránea)
- Código transportista (llave foránea)
- Código tiempo (llave foránea)
- Total de venta de los productos (métrica)

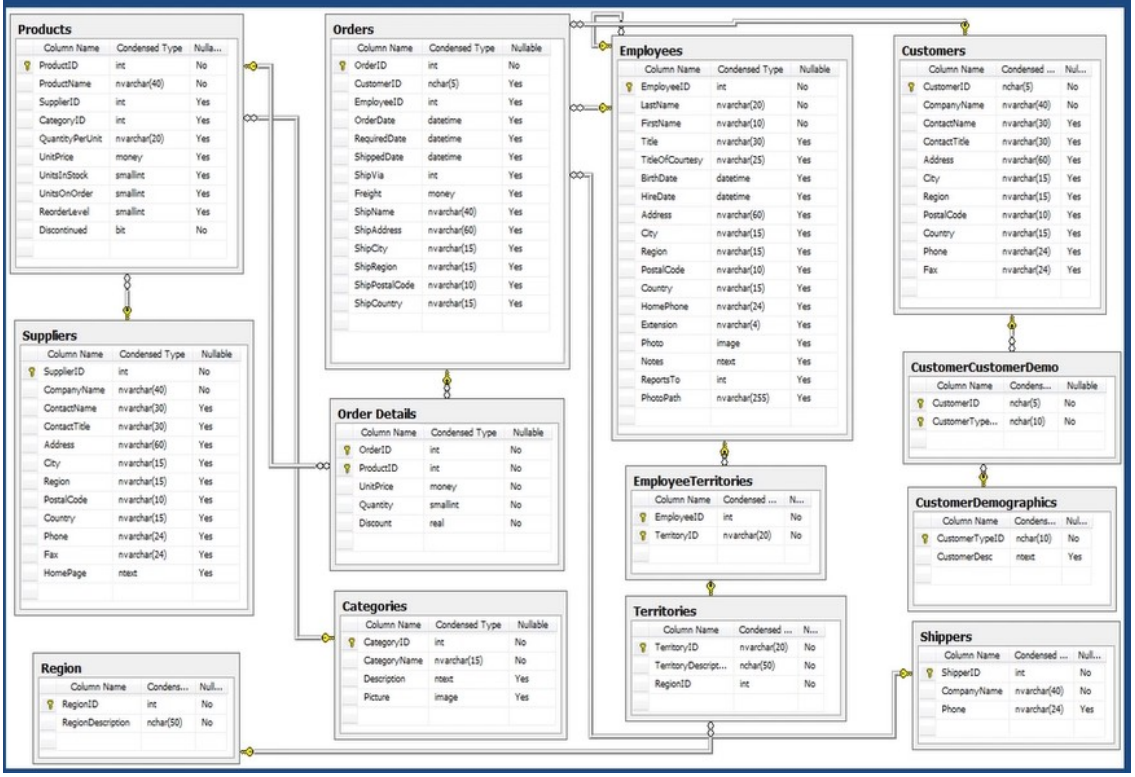
6- Modelo en estrella

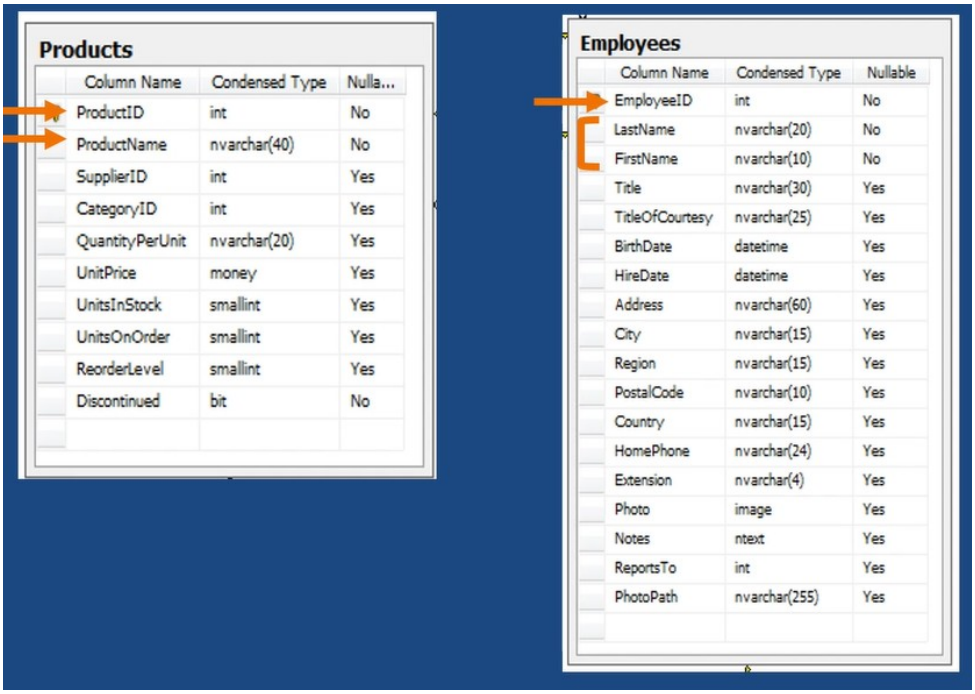


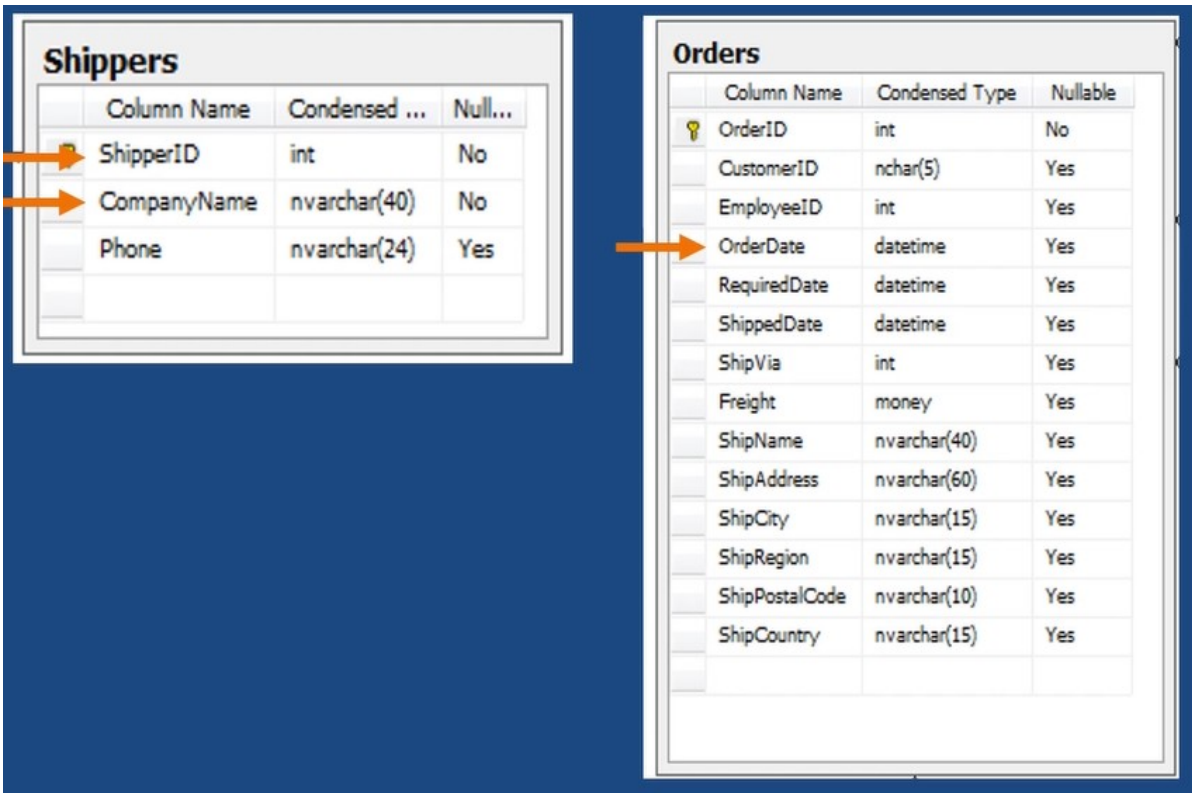
7- Identificar que tablas servirán para llenar las de hecho y dimensiones

Se identifican dimensión por dimensión.

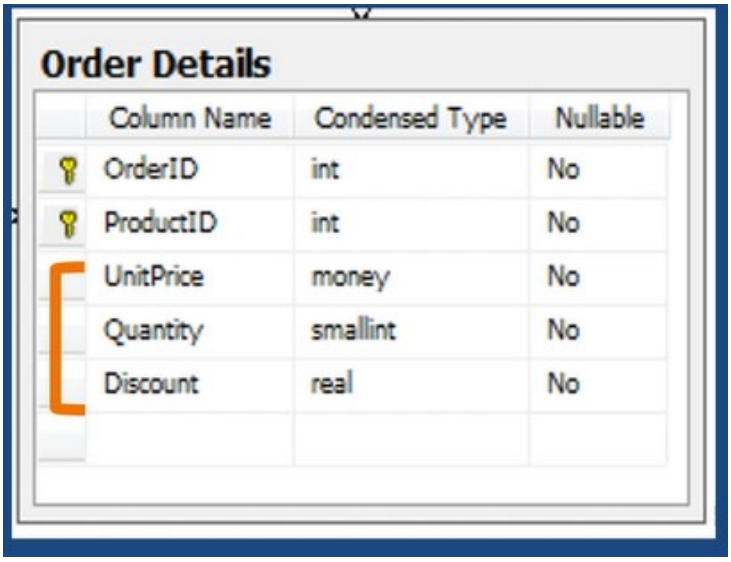
Base de datos OLTP







Para las métricas se utilizaran estos campos



Ya solo queda crear el ETL para cargar los datos en las tablas de dimensiones y la tabla de hechos.

Clase #3

SSMS como gestor de base de datos

Restarurar la base de datos Norrhwind:

1. Clic derecho en base de datos
2. Clic en restaurar base de datos
3. Agregamos y seleccionamos la base de datos a restaurar (extensión .bak)
4. Por último, en opciones habilitar la casilla de reemplazar la base de datos

Restaurar la base de datos OLAP usando script, Al ser una base de datos OLAP el script solo debe contener la creación de tablas, los datos se cargarán en el ETL:

1. Abrir el script con SSMS
2. Crear la base de datos
3. Ejecutar el Script

Antes de empezar con los procesos ETL se debe verificar que las tablas estén vacías, para crear ETL se utiliza Visual Studio SSDT

1. Crear nuevo proyecto
2. En Business intelligence seleccionar Integration service Project
3. Ponerle nombre y crear proyecto

En el espacio de trabajo se agregan todos los objetos se necesitaran para crear el ETL, Control Flow (como será la ejecución del flujo de los objetos) y data Flow son las paginas que mas se usaran.

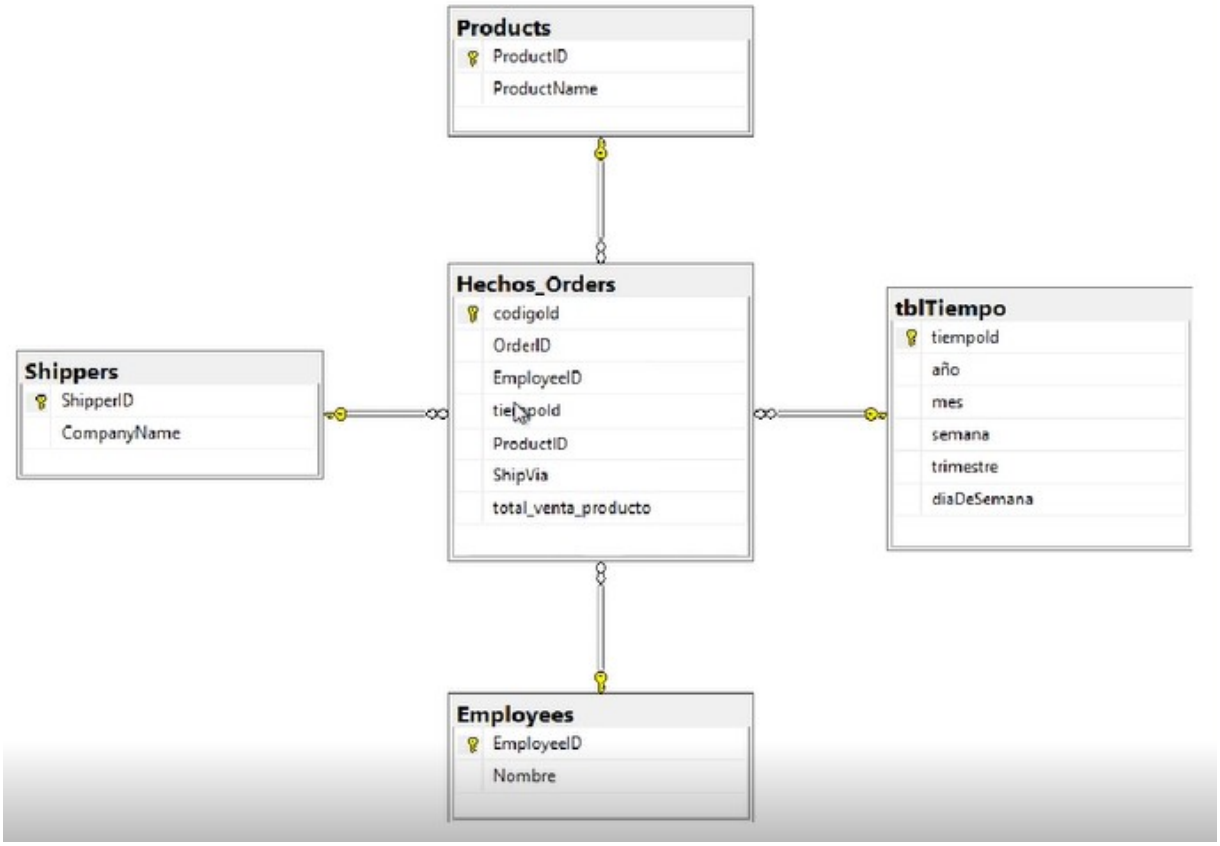
1. Tarea de flujo de datos: Encargada de llenar las tablas de dimensiones

Por cada tabla de dimensión se debe crear un ETL, en cada tabla de dimensión se realiza el proceso de extracción, transformación y carga de datos.

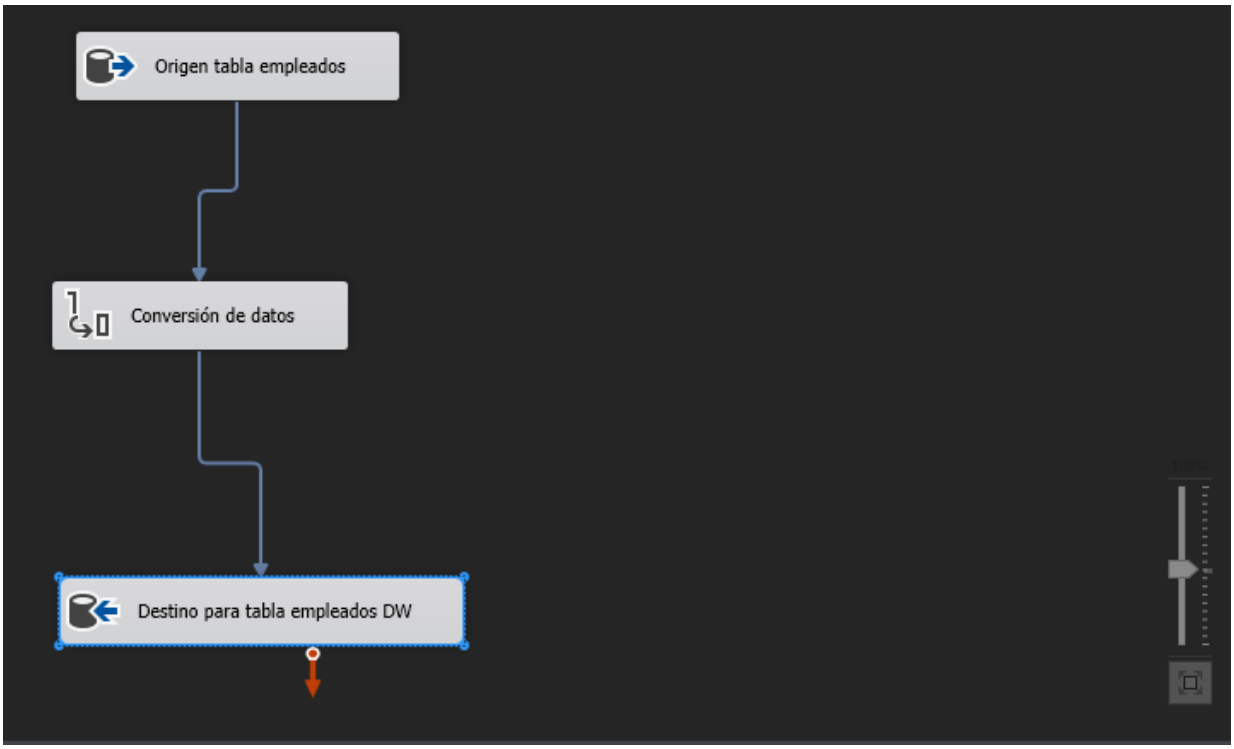
2. Usar Data Flow Task para el llenado de una tabla de dimensión
3. En el flujo de datos de la tarea de flujo de datos (paso anterior) usar OLE DB Source
4. Agregar una conexión, de donde se obtendrán los datos de origen, de que base y de que gestor. Usando el usuario con el que se conecta a Microsoft server managment studio.
5. El modo de acceso a los datos será por SQL Command.
6. La consulta para llenar la tabla empleados de la base de datos NORTHWIND:

a. select EmployeeID, CONCAT(FirstName, ' ', LastName) Nombre from Employees

7. En columnas se pueden escoger las columnas a utilizar.
8. Agregar objeto para transformar datos llamado **Conversión de datos**
9. Seleccionar los campos que se desean transformar
10. Verificar que la longitud y tipo de datos sean correctos con los de la base de datos de origen
11. Pasar los datos al destino llamado OLE DB
12. Se debe crear otra conexión porque se ese objeto se conectará con la base de datos OLAP (DW_NORTHWIND)
13. Se debe escoger tabla o vista y seleccionar la tabla de empleados
14. En la pestaña de asignaciones se indica como se van a llenar los campos en el destino, se especifican los campos de origen.
15. Se debe usar el objeto **Conversión de datos** que lo que hace es buscar el tipo de longitud correcta y el tipo de dato correcto



Así quedaría el primer ETL



Ahora se debe hacer para cada una de las tablas

Para obtener año, mes o día de la fecha se debe usar una función especifica **DATEPART(YEAR, OrderDate)** y para los días se debe usar **DATENAME(WEEKDAY, OrderDate).**

Cuando se genere un error por un tipo de dato Unicode string se debe cambiar el tipo de dato en la configuración de la **conversión de datos** asignándole el tipo de dato cadena (string).

Luego de tener los ETL de todas las tablas de dimensión se debe crear el ETL para la tabla de hechos.

1. Usar la tarea de flujo de datos
2. Seguir procedimiento de creación de ETL

Función de **SUM** para hacer suma de datos numéricos. Ej: SUM([Order Details].Quantity*[Order Details].UnitPrice*(1-[Order Details].Discount)) total_venta_producto

Al terminar los ETL se debe guardar y ejecutar para poder corroborar que funcione todo bien.

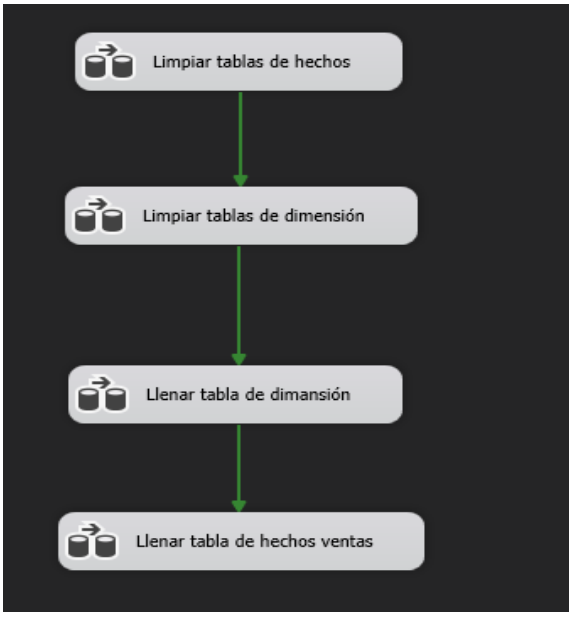
Si se quiere volver a ejecutar tirara error porque los datos ya están cargados en las bases de datos. Para poder corregir el error se agregarán objetos adicionales. Una solución es eliminar los registros que se tienen en la tabla de dimensión y la tabla de hechos antes de que se vuelvan a cargar.

En control Flow se agregará otro flujo de datos llamado **Limpiar tablas de dimensión.**

Se crea un origen de datos OLE usando la base del Data Warehouse (DW) pasando los datos de la tabla que se vaya a usar a un objeto de **comando OLE DB** que se encargará de eliminar los registros que obtenga del origen.

1. Se utiliza la conexión de la base de datos OLAP (DW)
2. En propiedades de la conexión en el apartado de SQLcommand se escribe el comando para limpiar los campos de la tabla: **DELETE FROM EMPLOYEES WHERE EmployeeID=?** donde el signo de interrogación implica un parámetro el cual debe ser asignado.
3. En asignaciones de columna donde aparece Param_0 que es el parámetro asignable, el cual debe ser asociado a alguno de los campos obtenidos por el origen, en este caso EmployeeID.
4. Cada valor que se tenga en employeeID se pasará al parámetro ejecutando de esta forma el DELETE en cada uno de los registros.

Cada cierto tiempo se debe ejecutar este proyecto de ETL, definido en base a las necesidades del negocio (cada mes, cada semana, etc) El orden correcto de las tareas de flujo de datos es:



Configuración SQL SERVER