

UF1890: Desarrollo de componente software y
consultas dentro del sistema de almacén de datos

Elaborado por: Jorge Martínez Mostazo

Edición: 5.0

EDITORIAL ELEARNING S.L.

ISBN: 978-84-16557-59-2

No está permitida la reproducción total o parcial de esta obra bajo cualquiera de sus formas gráficas o audiovisuales sin la autorización previa y por escrito de los titulares del depósito legal.

Impreso en España - Printed in Spain

Presentación

Identificación de la Unidad Formativa

Bienvenido a la Unidad Formativa **UF1890: Desarrollo de componente software y consultas dentro del sistema de almacén de datos**. Esta Unidad formativa pertenece al Módulo Formativo **MF1215_3: Creación y mantenimiento de componentes software en sistemas de planificación de recursos empresariales y de gestión de relaciones con clientes**, que forma parte del Certificado de Profesionalidad **IFCT0610: Administración y programación en sistemas de planificación de recursos empresariales y de gestión de relaciones con clientes**, de la familia de **Informática y Comunicaciones**.

Presentación de los contenidos

La finalidad de esta Unidad Formativa es enseñar al alumno a desarrollar componentes y consultas dentro del sistema de almacén de datos para almacenar y recopilar información, de acuerdo a especificaciones de diseño establecidas. Para ello, se analizará la carga de datos, la extracción de datos, y se profundizará en las herramientas de obtención de información.

Objetivos de la Unidad Formativa

Al finalizar esta Unidad Formativa aprenderás a:

- Especificar las estructuras y desarrollar componentes para la manipulación y recopilación de información del sistema de almacén de datos en sistemas ERP-CRM, siguiendo especificaciones técnicas y funcionales dadas.

Área: informática y comunicaciones

Índice

UD1. Carga de datos	7
1.1. Exploración del sistema de almacén de datos. Estructuras de información cubos y multicubos.....	9
1.1.1. Identificación de tipos de estructuras de información y sus relaciones para almacenar información	52
1.2. Procesos de carga de datos al sistema de almacén de datos.....	64
1.2.1. Identificación de orígenes de datos para la carga de datos	67
1.2.2. Creación de componentes de software para extraer información de un sistema de almacén de datos	73
UD2. Extracción de datos (data warehouse)	85
2.1. Herramientas para la carga y extracción de datos de sistemas de almacén de datos.....	87
2.1.1. Mecanismos que se utilizan para la extracción de datos .	90
2.1.2. Estructuración de la información para adecuarse a las necesidades de la empresa.....	95
2.2. Creación de extractores de datos.....	96
2.2.1. Obtención de información de fuentes internas o externas	105
2.2.2. Agrupación, transformación y homogeneización de la información para su posterior estudio.....	111

UD3. Herramientas de obtención de información	125
3.1. Herramientas de obtención de información	127
3.2. Herramientas de visualización y difusión	142
Supuesto Práctico 1	155
Supuesto Práctico 2	165
Glosario	169
Soluciones	171

UD1

Carga de datos

- 1.1. Exploración del sistema de almacén de datos. Estructuras de información cubos y multicubos
 - 1.1.1. Identificación de tipos de estructuras de información y sus relaciones para almacenar información
- 1.2. Procesos de carga de datos al sistema de almacén de datos
 - 1.2.1. Identificación de orígenes de datos para la carga de datos
 - 1.2.2. Creación de componentes de software para extraer información de un sistema de almacén de datos

1.1. Exploración del sistema de almacén de datos. Estructuras de información cubos y multicubos

ETL: Extraer Transformar y Cargar. Es un proceso donde se obtienen datos de una fuente, los datos son cambiados y cargados en un almacenamiento. Este tipo de proceso es muy usado en distintos tipos de situaciones:

- **Datawarehouse.** Creación de un almacén de datos, nos va a facilitar acceder a la información mediante informes. Este es el caso que nos ocupa.
- **Migración de datos.** Pasar los datos de una base de datos a otra en el mismo sistema.
- **Migración de datos a otro sistema.** Cuando se crea un ETL, se crea un vínculo con dicho ETL, siendo parte de la empresa de forma que cualquier proceso queda recogido y formará parte de los negocios.

El uso de una herramienta ETL es un movimiento estratégico importante independientemente del tamaño del proyecto.

Se pueden considerar varias áreas que nos pueden ayudar a elegir nuestra herramienta ETL:

1. **Conexión de datos.** Debería ser versátil y ser capaz de conectarse con cualquier fuente de datos. La fuente de datos puede cambiar en nuestra empresa más adelante.
2. **Performance.** El movimiento y el cambio de datos son procesos potentes que pueden consumir recursos. Es importante disponer de los recursos necesarios para afrontar dichos procesos.
3. **Transformación de datos.** Sencillez y simplicidad a la hora de transformar los datos.

4. **Calidad datos.** Una herramienta ETL debe permitir comprobar y validar los datos con los que se va trabajar. Los datos deben ser consistentes.
5. **Obtención de datos.** Debe ser flexible a la hora de obtener los datos ya cambiados:
 - Obtener un subconjunto de datos.
 - Obtener datos en un intervalo de tiempo.
 - Obtener informes.

Una vez decidido es importante tener en cuenta la empresa que nos va a suministrar la herramienta, de forma que siempre tengamos el soporte adecuado.

Los procesos ETL también se pueden utilizar para la integración con sistemas antiguos de la empresa. Microsoft SQL Server Integration Services (SSIS) permite utilizar ETL para el almacenar datos.

Proceso SSIS.



ETL en SQL Server

El proceso de ETL en SQL Server se realiza mediante la herramienta SSIS (Integration Services).



Un **paquete** es una unidad de trabajo en la que vamos a indicar los distintos elementos que van a participar en el proceso ETL. Los paquetes se crean junto con un proyecto en SSIS.

El primer paso al crear un paquete en Integration Services es crear un proyecto que incluye las plantillas para los objetos que se utilizan en una solución de transformación de datos.

La implementación del proceso ETL puede ser realizada de dos formas distintas:

- **Un paquete por cada fuente de datos.** Cada paquete maneja de forma independiente los datos de la fuente. Es recomendable cuando:
 - La fuente de datos es independiente.
 - Fuentes de datos que cambian.
- **Paquetes separados en cada paso (ETL).** Requiere una base de datos staging. Es una buena opción cuando:
 - Tenemos poco tiempo para la carga de datos.
 - Fuentes de datos diferentes que necesitan ser integradas.

Proyecto SSIS.

- Paquete por cada fuente de datos.
- Paquetes separados por cada paso.
- ETL. *Extract Transform Load*.

Carga

La fase de carga se produce cuando los datos son cargados en el sistema destino. Dependiendo de las necesidades de la organización, este proceso puede llevar una amplia variedad de diferentes procesos.

Existen dos formas de desarrollar un proceso de carga:

- **Single Accumulation.** Es la más utilizada. Realiza un extracto de todas las transacciones que se encuentran en una fase de tiempo adecuado y lleva el resultado como una transacción al **DW**, guardando un valor calculado que será en un sumatorio o un valor medio.
- **Rolling.** Se utiliza en los casos en que se desea mantener varios grados de granularidad. Para ello se guardan datos resumidos en distintos niveles, que se corresponden a distintos grupos de la unidad de tiempo o distintos niveles subordinados.

La fase de carga actúa de forma directa con la base de datos final. Realizando esta operación se aplican todas las constraints y disparadores que se hayan definido en ella. Estos disparadores consiguen que se obtenga la calidad de los datos en el proceso de ETL.

Proceso de carga.

- Acumulación. Resumen de todas las transacciones y las envía al DW.
- Rolling. Resume por niveles.
- Interacciona con la BD destino.

Sistema de almacén de datos. Introducción

DW o sistema de almacén, es el eje de la construcción de los SI. Permite realizar un análisis en base a unos datos antiguos o históricos DW. Permite la unión de sistemas no integrados. Dispone y guarda los datos necesarios para el procesamiento ordenado en función del tiempo.

Un DW es un conjunto de información orientado a temas, integrado, no variable, de tiempo variable, que se usa para sostener la fase de toma de resoluciones en una empresa.

Se puede caracterizar un data warehouse haciendo un contraste de cómo los datos de un negocio que son almacenados en un data warehouse, difieren de los datos operacionales usados por las aplicaciones de producción.

DW	Base de Datos Actual
Datos de negocio para Información	Datos del sistema de operaciones
Orientado a la persona	Orientado al software
Especificada y más sintetizada	Especificada
Fija	Cambio constante

La inserción de datos en el DW proviene principalmente del sistema de operaciones. DW posee información transformada y en un lugar físico diferente al que se encuentran los datos del sistema de operación.



El sistema de almacén se utiliza en la toma de decisiones en la empresa.

Definición de Data warehouse

Un almacén de datos es un sistema que recupera y consolida datos periódicamente de los sistemas de fuente en un almacén de datos dimensionales o normalizados. Generalmente se mantiene años de historia y se consulta para inteligencia de negocios u otras actividades analíticas. Normalmente se actualiza en lotes, no cada vez que ocurre una transacción en el sistema de origen.

El ETL (extract, transform, load) aporta datos de varios sistemas de fuente en un área. ETL es un sistema que tiene la capacidad para conectarse a los sistemas de fuente, leer los datos, transformar los datos y cargarlos en un sistema de destino (el sistema de destino no tiene que ser un almacén de datos). El sistema ETL entonces integra, transforma y carga los datos en un almacén de datos dimensionales (DDS).

Un DDS es una base de datos que almacena los datos en un formato diferente de OLTP. La razón para meter los datos desde el sistema de origen en el DDS y luego consultar la DDS en lugar de consultar directamente el sistema de origen es que en un DDS se arreglan los datos en un formato tridimensional que es más conveniente para el análisis. La segunda razón es porque un DDS contiene datos integrados de varios sistemas de origen.

Sistema Data warehouse.



El sistema de almacén de datos contiene solamente un sistema ETL y un almacén de datos dimensionales. El sistema de origen no es parte del sistema del almacén de datos. Este es básicamente la unidad mínima. Si se toma sólo un componente más, no se puede llamar **sistema de almacén de datos**. Aunque no hay ninguna aplicación front-end como informes o aplicaciones analíticas, los usuarios todavía pueden consultar los datos de la DDS mediante la emisión de directas instrucciones select de SQL usando herramientas de consulta de base de datos genéricos como la que organizó en SQL Server Management Studio.

Obtención de datos en DW

La recuperación de datos se realiza mediante un conjunto de rutinas ampliamente conocido como un sistema ETL, que es una abreviatura de extracto, transformación y carga. El sistema ETL es un conjunto de procesos que recuperan datos de los sistemas de fuente, transforman los datos y carga en un sistema de destino. La transformación puede utilizarse para cambiar los datos según el formato y criterios del sistema de destino, para derivar nuevos valores para ser cargados en el sistema de destino, o para validar los datos desde el sistema de origen. Sistemas ETL no sólo se utilizan para cargar los datos en el almacén de datos. Son ampliamente utilizados para cualquier tipo de movimientos de datos.

La mayoría de los sistemas ETL también tienen mecanismos para limpiar los datos desde el sistema de origen antes de ponerlo en el almacén. Limpieza de datos es el proceso de identificar y corregir datos sucios. Esto es implementado usando reglas de calidad de datos que definición están qué datos sucios.

Después se extrajeron los datos desde el sistema de origen, pero antes de carguen los datos en el almacén, los datos se examinan por medio de estas reglas. Si la regla determina los datos son correctos, entonces está cargado en el almacén. Si la regla determina que la información es incorrecta, entonces hay tres opciones: puede ser rechazado, corregido o puede ser cargado en el almacén.

No hay otra alternativa a ETL, conocido como extraer, cargar y transformar (ELT). En este enfoque, los datos se cargan en el almacén de datos. Las transformaciones, las búsquedas, y así sucesivamente, se realizan dentro del almacén de datos. A diferencia del enfoque ETL, el enfoque ELT no necesita un servidor de ETL. Este enfoque es generalmente aplicado para aprovechar las ventajas del almacén de datos potentes motores de base de datos tales como sistemas de procesamiento masivamente paralelo (MPP).

ETL:

- Recupera datos de un sistema fuente y lo envía a un sistema destino.
- Corrección y limpieza de datos.
- Si los datos son correctos se cargan en el almacén.

Consolidación de datos

Una empresa puede tener muchos sistemas transaccionales. Por ejemplo, un banco puede utilizar 15 diferentes aplicaciones para sus servicios, uno para

procesamiento de préstamos, uno para el servicio al cliente, uno para cajeros / cajeras, uno de los cajeros automáticos, uno para bonos para ISA, uno para el ahorro, uno para privatebanking, uno para el parquet, uno para seguros de vida, seguros de hogar, uno de las hipotecas, uno para la llamada centro, uno para las cuentas internas y otro para la detección de fraude. Sería muy difícil realizar (por ejemplo) Análisis de rentabilidad de clientes a través de estas aplicaciones.

Un almacén de datos consolida muchos sistemas transaccionales. La diferencia clave entre un almacén de datos y un sistema transaccional de front-office es que los datos en el almacén de datos están integrados. Esta consolidación o integración debe tener en cuenta la disponibilidad de los datos (algunos datos están disponibles en varios sistemas pero no en otros), el tiempo de respuesta (datos en distintos sistemas tienen períodos de validez diferentes), diversas definiciones (el término semanal los ingresos totales en un sistema que tenga un significado diferente del total de los ingresos semanal en otros sistemas), conversión (diferentes sistemas que tenga una unidad de medida o moneda diferente) y congruentes (fusión de datos basados en identificadores comunes entre los distintos sistemas).

Conceptos a tener en cuenta para la consolidación de datos:

- **Disponibilidad de datos.** Rango de tiempo.
- **Conversión de datos.** Comparación de datos.

Periodicidad en un DW

La consolidación y la recuperación de datos no ocurren sólo una vez; pasan muchas veces y generalmente a intervalos regulares, como todos los días o un par de veces al día. Si la recuperación de datos ocurre sólo una vez, entonces los datos serán obsoletos, y después de algún tiempo no será útil. Puede determinar el período de recuperación de datos y basado en los requerimientos del negocio y la frecuencia de las actualizaciones de datos en los sistemas de fuente de consuelo. El intervalo de recuperación de datos debe ser la misma frecuencia de actualización de datos del sistema de la fuente. Si el sistema de origen se actualiza una vez al día, necesitas la recuperación de datos una vez al día. No hace falta extraer los datos de ese sistema de fuente varias veces al día.

Por otro lado, para asegurarse de que el intervalo de recuperación de datos cumple con los requerimientos del negocio, por ejemplo, si el negocio necesita el informe de rentabilidad del producto una vez a la semana, entonces los datos de varios sistemas de fuente necesitan ser consolidados por lo menos una vez por semana. Otro ejemplo es cuando una empresa informa a su clien-

te que habrá que esperar 24 horas para cancelar las suscripciones de comercialización. Entonces los datos en el almacén de datos CRM necesitan ser actualizados varias veces al día; de lo contrario, se arriesga a enviar campañas de marketing a los clientes que ya han cancelado su suscripción.

Periodicidad:

- Datos son añadidos al almacén.
- Datos nunca son alterados físicamente.
- Datos nunca son eliminados físicamente.

Almacén de datos dimensional

Un almacén de datos es un sistema que recupera datos desde sistemas de origen y la pone en un almacén de datos dimensionales o un almacén de datos normalizado. Sí, algunos almacenes de datos están en formato tridimensional, pero algunos almacenes de datos están en formato normalizado. Vamos a repasar las diferencias entre ellos y formatos.

Un DDS es una o varias bases de datos que contiene una colección de puestos de datos dimensionales. Un datamart dimensional es un grupo de tablas de hechos relacionados y sus correspondientes tablas de dimensiones que contienen las mediciones de eventos empresariales categorizados por sus dimensiones.

Un almacén de datos dimensionales es desnormalizado, y las dimensiones están conformadas. Dimensiones conformadas son exactamente la misma tabla de dimensión o uno es el subconjunto de la otra. Dimensión A se dice que es un subconjunto de dimensión B cuando todas las columnas de existir dimensión A en la dimensión B y todas las filas de la existencia de la dimensión A en la dimensión B.

Un almacén de datos dimensionales puede implementarse físicamente en forma de varios esquemas diferentes. Ejemplos de esquemas de almacén de datos dimensionales son un esquema en estrella, un esquema de copo de nieve. En un esquema en estrella, una dimensión no tiene una subtabla (una subdimensión). En un esquema de copo de nieve, una dimensión puede tener una subdimensión. El propósito de tener una subdimensión es minimizar los datos redundantes. Un esquema de estrella es también conocido como un esquema de constelación de hecho. En un esquema de la estrella, tienes dos o más tablas de hechos rodeadas de dimensiones comunes relacionadas. El beneficio de tener un esquema en estrella es que es más simple que copo

de nieve, lo que facilita los procesos ETL cargar los datos en DDS. El beneficio de un esquema de copo de nieve es que algunas aplicaciones analíticas funcionan mejor con un esquema de copo de nieve en comparación con un esquema en estrella o estrella esquema. El otro beneficio de un esquema de copo de nieve es menos redundancia de datos, así que se requiere menos espacio en disco. La ventaja del esquema de la estrella es la habilidad para modelar con mayor precisión los eventos de negocios mediante el uso de varias tablas de hechos.

Almacén de datos dimensional Estrella:

- Es simple.
- Diseño lógico relacional de la base de datos.
- Solo existe una tabla de dimensiones para cada dimensión.

Almacén de datos normalizado

Otros tipos de almacenes de datos ponen los datos no en un almacén de datos dimensionales sino en un almacén de datos normalizado. Un **almacén de datos normalizado** es una o más bases de datos relacionales con poca o ninguna redundancia de datos. Una base de datos relacional es una base de datos que consiste en tablas de entidad con relación parent-child (padre-hijo) entre ellos.

Normalización es un proceso de eliminación de redundancia de datos mediante la aplicación de **las reglas de normalización**. Hay cinco grados de formas normales, desde la primera forma normal a la Quinta forma normal. Un almacén de datos normalizado suele ser en tercera forma normal o superior, como la cuarta o quinta forma normal.

Un almacén de datos dimensionales es un mejor formato para almacenar los datos en el almacén con el fin de realizar consultas y análisis de datos de un almacén de datos normalizado. Esto es porque es más simple (un nivel profundo en todas las direcciones en esquema en estrella) y le da mejor rendimiento de consultas. Un almacén de datos normalizado es un mejor formato para integrar datos de varios sistemas de fuente, especialmente en tercera forma normal y superior. Esto es porque hay sólo un lugar para actualizar sin redundancia de datos como en un almacén de datos dimensionales.

El almacén de datos normalizado se utiliza generalmente para un almacén de datos de la empresa; desde allí los datos luego se cargan en los almacenes de datos dimensionales para fines de consulta y análisis.

Algunas aplicaciones se ejecutan en un DDS, es decir, una base de datos relacional que consiste en tablas con filas y columnas. Algunas aplicaciones se ejecutan en una base de datos multidimensional que consiste en cubos con celdas y dimensiones.

DW Normalizado.

- Son bases de datos relacionales con poca o ninguna redundancia de datos.
- Normalización es un proceso de eliminación de redundancia de datos.
- El almacén de datos normalizado se utiliza generalmente para un almacén de datos de la empresa.

Historial DW

Una de las diferencias clave entre un sistema transaccional y un sistema de almacén de datos es la capacidad para almacenar historial. Los sistemas transaccionales guardan algo más de historia, pero los sistemas de almacén de datos almacenan historia durante más tiempo. En mi experiencia, los sistemas transaccionales almacenan solamente de uno a tres años de datos; más allá de eso, los datos se depuran. Por ejemplo, echemos un vistazo a un sistema de procesamiento de pedidos – ventas. El propósito de este sistema es procesar pedidos de clientes. Una vez que un pedido es enviado y pagado, está cerrado, y después de dos o tres años, que desee purgar los pedidos cerrados fuera del sistema activo y archivarlos para mantener el rendimiento del sistema.

Quizá quieras guardar los registros, por ejemplo, dos años, en caso de que el cliente quiera consultar sus pedidos, pero no quieres mantener el espacio que ocupan diez años de datos en el sistema activo, pues se ralentiza el sistema. Algunas regulaciones (que difieren de país a país) requieren, para mantener datos hasta cinco o siete años, fines impositivos o adherirse a las normas de la bolsa de valores. Pero esto no significa que debes mantener los datos en el sistema activo. Puede archivarlos por medios offline. Eso es lo que hace un sistema de transacciones típicas: mantiene sólo dos o tres años de datos en el sistema activo y el resto a medios offline o a una base secundaria sólo de lectura de los archivos.

En un almacén de datos, por otro lado, almacenando la historia de los datos maestros es una de las principales características. Esto se conoce como una dimensión lentamente cambiante (SCD).



Definición

Una **dimensión lentamente cambiante** es una técnica utilizada en el modelado dimensional para preservar la información histórica sobre datos dimensionales. En tipo SCD-2, te quedas con la información histórica en filas; mientras que en tipo SCD 3, te quedas con la información histórica en columnas. En el tipo 1 de SCD, no mantienes la información histórica.

También relacionado con la historia, un almacén de datos almacena capturas periódicas de operaciones de sistemas de origen. Una **captura** es una copia de una o más tablas maestras tomada en un momento determinado. Una captura periódica es una instantánea que se toma a intervalos regulares; por ejemplo, la industria bancaria toma capturas de las tablas de cuenta del cliente todos los días. Los datos y aplicaciones de almacén luego comparan las capturas diarias para analizar al cliente bidones, los saldos de cuenta y condiciones inusuales. Si el tamaño de un sistema de fuente es, digamos, 100 MB, en un año habría acumulado 37 GB.

Esquema de tipos SCD:

- **SCD1.** Sobrescribir la información.
- **SCD2.** Añadir fila indicando información de modificación.
- **SCD3.** Añadir columna de información por cada columna.

Consultas en DW

Consulta es el proceso de obtener datos de un almacén de datos, que cumple con ciertos criterios. Aquí hay un ejemplo de una consulta simple: "¿Cuántos clientes tienes ahora?" Un ejemplo de una consulta compleja: "Muéstrame los nombres y los ingresos de todas las líneas de producto que tuvo una pérdida de 10 por ciento o mayor en 2008, categorizados por salida."

Un almacén de datos está construido para ser consultados. Es el objetivo número uno de su existencia. Los usuarios no pueden actualizar el almacén de datos. Los usuarios sólo pueden consultar el almacén de datos. Se permite solamente el sistema ETL para actualizar el almacén de datos. Esta es una de las principales diferencias entre un almacén de datos y un sistema de transacciones.

El sistema de origen es generalmente un sistema transaccional, utilizado por muchos usuarios. Una característica importante de un sistema transaccional es la capacidad para permitir que muchos usuarios puedan actualizar y seleccionar en el sistema al mismo tiempo. Para ello, debe ser capaz de realizar muchas transacciones de bases de datos (update, insert, delete y seleccione) en un período relativamente corto de tiempo. En otras palabras, debe ser capaz de realizar las transacciones de bases de datos muy rápidamente. Si se guarda la misma pieza de datos — digamos, precio unitario — en muchos lugares diferentes en el sistema, tomaría mucho tiempo para actualizar los datos y para mantener la consistencia de estos. Si lo almacenamos en un sólo lugar, será más rápido actualizar los datos, y no habría que preocuparse por mantener la consistencia de los datos entre diferentes lugares. Además, sería más fácil mantener la concurrencia y mecanismo de bloqueo para permitir a muchas personas a trabajar juntos en la misma base de datos. Por lo tanto, uno de los principios fundamentales de un sistema de transacción es eliminar la redundancia de datos.

Realizar una consulta compleja en una base de datos normalizada (tales como sistemas transaccionales) es más lenta que la realización de una consulta compleja en una base de datos desnormalizada (por ejemplo, un almacén de datos), porque en una base de datos normalizada, necesita unirse a muchas tablas. Una base de datos normalizada no es adecuada para ser utilizada para cargar los datos en una base de datos multidimensional con el propósito de corte-y-corte en cuadrillos de análisis. A diferencia de una base de datos relacional que contiene las tablas con dos dimensiones (filas y columnas), una base de datos multidimensional consiste en cubos que contienen celdas con más de dos dimensiones. Luego cada celdas está asignada a un miembro de cada dimensión. Para cargar una base de datos multidimensional de una base de datos normalizada, necesitas hacer una consulta multijoin para transformar datos a formato tridimensional.

La segunda razón por qué no consulta los sistemas de fuente directamente es porque una empresa puede tener que muchos fuentes o recepción transaccional. Un almacén de datos, por otro lado, consolida los datos de muchos sistemas de fuente.

Consultas:

- Obtención de datos conforme a unos criterios.
- Una consulta compleja es más rápida en un DW.
- DW consolida los datos de muchos sistemas fuente.