Temas Generales para la preparación de la Oposición al Cuerpo Superior de Sistemas y Tecnologías de la Información de la Administración del Estado.

Cuerpo Superior de Estadísticos del Estado
Especialidad de Estadística-Ciencia de Datos.

Almacenamiento y modelos de datos

Tema 10. Data Warehousing

AUTOR: **Alexandra Vlad**

Asociación Profesional de Cuerpos Superiores de Sistemas y Tecnologías de la Información de las Administraciones Públicas

Creación: Julio 2021

ÍNDICE

1 Introducción a Data Warehousing

Los Data Warehouses (almacenes de datos) han llegado claramente para quedarse; ya no se consideran una parte opcional del "arsenal" de la base de datos para muchas empresas. La evidencia de la llegada de los data warehouses como un elemento permanente es que los proveedores de bases de datos ahora incluyen capacidades de data warehousing como un servicio central de sus productos de bases de datos.

Los data warehouses no solamente están creciendo en tamaño y prevalencia, sino que también se ha expandido su alcance y complejidad. Se espera que los sistemas actuales de data warehouse no solo sean compatibles con los informes tradicionales, sino que también proporcionen análisis más avanzados, como los análisis multidimensionales y predictivos, y esta gama está destinada a satisfacer las necesidades de un número creciente de diferentes tipos de usuarios. Se espera que los recursos de los data warehouses no solo estén disponibles para un número creciente de usuarios internos, sino que también sea accesible y de utilidad para aquellos usuarios externos a una empresa, como clientes y proveedores. Se cree que la creciente popularidad de los data warehouses está impulsada por una variedad de factores, que incluyen, por ejemplo, el cumplimiento normativo gubernamental que requiere que las empresas mantengan historiales transaccionales e instalaciones de almacenamiento más baratos y fiables a la aparición de data warehouses en tiempo real que satisfacen los requisitos de las aplicaciones de inteligencia empresarial en las cuales el tiempo es crítico.

La evolución de los Data Warehouses

Desde la década de los 70, las organizaciones han centrado principalmente su inversión en nuevos sistemas informáticos que automatizan los procesos de negocio. De esta manera, las empresas obtuvieron una ventaja competitiva a través de sistemas que ofrecían servicios más eficientes y rentables al cliente. Durante este periodo, las empresas acumularon una creciente cantidad de datos almacenados en sus bases de datos operacionales. Sin embargo, en los últimos tiempos, cuando tales sistemas son comunes, las empresas se están enfocando en formas de utilizar los datos operativos para respaldar la toma de decisiones como un medio para recuperar una ventaja competitiva.

Los sistemas operativos nunca se diseñaron para respaldar tales actividades comerciales, por lo que el uso de estos sistemas para la toma de decisiones puede que nunca sea una solución fácil. El legado es que una organización típica puede tener numerosos sistemas operativos con definiciones superpuestas y, a veces, contradictorias, como los tipos de datos. El desafío para una organización es convertir sus archivos de datos en una fuente de conocimiento, de modo que se presente al usuario una única vista integrada / consolidada de los datos de la organización. El concepto de almacén de datos se consideró la solución para cumplir con los requisitos de un sistema capaz de respaldar la toma de decisiones y recibir datos de múltiples fuentes de datos operativos.

2 Conceptos de Data Warehousing

El concepto original de un Data Warehouse fue ideado por IBM como el "almacén de información" y presentado como una solución para acceder a los datos almacenados en sistemas no relacionales. El almacén de información se propuso para permitir que las organizaciones utilicen sus archivos de datos con el fin de ayudarles a obtener una ventaja comercial. Sin embargo, debido a la gran complejidad y los problemas de rendimiento asociados con la implementación de tales soluciones, los primeros intentos de crear un almacén de información fueron en su mayoría rechazados. Desde entonces, el concepto de data warehousing se ha planteado varias veces, pero solo en los últimos años se ha visto el potencial del data warehousing como una solución valiosa y viable. Uno de los primeros promotores del data warehousing es Bill Inmon, quien se ha ganado el título de "padre del data warehousing".

Data Warehousing se puede considerar una colección de datos orientada al tema, integrada, variable con el tiempo y no volátil como apoyo en el proceso de toma de decisiones de la dirección.

En esta primera definición de Inmon, los datos son:

- *Orientados a temas*, ya que el almacén de datos está organizado en torno a los temas principales de la empresa (como clientes, productos y ventas) en lugar de las principales áreas de aplicación (como facturación al cliente, control de existencias y venta de productos). Esto se refleja en la necesidad de almacenar datos de apoyo a la toma de decisiones en vez de datos orientados a aplicaciones.
- *Integrados*, debido a la combinación de datos fuente de diferentes sistemas de aplicaciones de toda la empresa. Los datos fuente suelen ser incoherentes y utilizan, por ejemplo, diferentes formatos. La fuente de datos integrada debe ser coherente para presentar una vista unificada de los datos a los usuarios.
- *Cambiantes con el tiempo*, porque los datos en el warehouse son precisos y válidos solo en algún momento o durante algún intervalo de tiempo. La variación en el tiempo del data warehouse también se muestra en el tiempo extendido que se mantienen los datos, la asociación implícita o explícita de tiempo con todos los datos y el hecho de que los datos representan una serie de instantáneas.
- *No volátiles*, ya que los datos no se actualizan en tiempo real, sino que se actualizan desde los sistemas operativos de forma regular. Los datos nuevos siempre se agregan como un complemento a la base de datos, en lugar de un reemplazo. La base de datos absorbe continuamente estos nuevos datos, integrándolos gradualmente con los datos anteriores.

Existen numerosas definiciones de data warehousing, y las primeras definiciones se centran en las características de los datos almacenados en el warehouse. Las definiciones alternativas y posteriores amplían el alcance de la definición de un data warehouse para incluir el procesamiento asociado con el acceso a los datos desde las fuentes originales hasta la entrega de los datos a los responsables de las tomas de decisiones (Anahory y Murray, 1997).

Cualquiera que sea la definición, el objetivo final del almacenamiento de datos es integrar los datos corporativos de toda la empresa en un único repositorio desde el cual los usuarios pueden ejecutar consultas, generar informes y realizar análisis fácilmente.

Ventajas del Data Warehousing

La implementación exitosa de un data warehouse puede traer grandes beneficios a una organización, que incluyen:

- *Alto potencial del retorno de la inversión*. Una organización debe comprometer una gran cantidad de recursos para garantizar la implementación exitosa de un data warehouse, y el coste puede variar enormemente desde miles a millones de dólares debido a la variedad de soluciones técnicas disponibles. Sin embargo, un estudio de International Data Corporation (IDC) informó que los proyectos de data warehouse arrojaron un retorno de la inversión (ROI) promedio de tres años del 401% (IDC, 1996). Además, un estudio posterior de IDC sobre

análisis de negocios, es decir, herramientas analíticas que acceden a los almacenes de datos, arrojó un ROI promedio de un año del 431% (IDC, 2002).

- *Ventaja competitiva*. El enorme retorno de la inversión para aquellas empresas que han implementado con éxito un data warehouse es una prueba de la enorme ventaja competitiva que acompaña a esta tecnología. La ventaja competitiva se obtiene al permitir que los responsables de las tomas de decisiones accedan a datos que pueden revelar información previamente no disponible, desconocida y sin explotar sobre, por ejemplo, clientes, tendencias y demandas.

- *Mayor productividad de los responsables de las tomas de decisiones corporativas*. El data warehouse mejora la productividad de los responsables de las tomas de decisiones corporativas al crear una base de datos integrada de datos históricos coherentes y orientados a los sujetos. Integra datos de varios sistemas incompatibles en una forma que proporciona una vista coherente de la organización. Al transformar los datos en información significativa, un data warehouse permite a los responsables de las tomas de decisiones corporativos realizar análisis más sustantivos, precisos y consistentes.

Sistemas OLTP (On Line Transaction Processing) vs Data Warehousing

Un DBMS construido para OLTP generalmente se considera inadecuado para un data warehousing, ya que cada sistema está diseñado con un conjunto diferente de requisitos en mente. Por ejemplo, los sistemas OLTP están diseñados para maximizar la capacidad de procesamiento de transacciones, mientras que los data warehouses están diseñados para admitir el procesamiento de consultas ad hoc.

Características	Sistemas OLTP	Sistemas Data Warehousing
Objetivo principal	Ofrece procesamiento operacional.	Ofrece procesamiento analítico.
Edad de los datos	Actual.	Histórico (aunque la tendencia es ir hacia algo que incluya también datos actuales).
Latencia de los datos	Tiempo real.	Depende de la duración del ciclo de los complementos del data warehouse (pero la tendencia es ir hacia los complementos en tiempo real).
Granularidad de los datos	Datos detallados.	Datos detallados, ligeramente resumidos y muy resumidos.
Procesamiento de los datos	Patrones predecibles de inserciones, eliminado y actualizaciones de datos, y consultas. Alto nivel de rendimiento de transacciones.	Patrones menos predecibles de consultas de datos; nivel medio a bajo de rendimiento de transacciones.
Informes	Informes predecibles, unidimensionales, fijos relativamente estáticos.	Informes impredecibles, multidimensionales y dinámicos.
Usuarios	Atiende a un gran número de usuarios operacionales.	Atiende a un menor número de usuarios de la gerencia (aunque la tendencia es ir hacia requisitos analíticos que soporten usuarios operacionales).

La tabla anterior proporciona una comparación de las principales características de los sistemas OLTP y los sistemas de data warehousing. La tabla también indica algunas de las principales tendencias que pueden alterar las características de un data warehousing. Una de esas tendencias es el movimiento hacia un data warehousing en tiempo real.

Una organización normalmente tendrá varios sistemas OLTP diferentes para los procesos comerciales, como el control de inventario, la facturación al cliente y el punto de venta. Estos sistemas generan datos operativos detallados, actualizados y sujetos a cambios. Los sistemas OLTP están optimizados para una gran cantidad de transacciones que son predecibles, repetitivas y de actualización intensiva. Los datos OLTP se organizan de acuerdo con los requisitos de las transacciones asociadas con las aplicaciones comerciales y respaldan las decisiones del día a día de un gran número de usuarios operativos concurrentes.

Por el contrario, una organización normalmente tendrá un único data warehouse, que contiene datos históricos, detallados y resumidos en varios niveles y rara vez están sujetos a cambios (aparte de complementarse con nuevos datos). El almacén de datos está diseñado para admitir un número relativamente bajo de transacciones que son de naturaleza impredecible y requieren respuestas a consultas ad hoc, no estructuradas y heurísticas. Los datos del almacén se organizan de acuerdo con los requisitos de las consultas potenciales y respaldan los requisitos analíticos de un número menor de usuarios.

Aunque los sistemas OLTP y los data warehouses tienen características diferentes y se construyen con diferentes propósitos en mente, estos sistemas están estrechamente relacionados, ya que los sistemas OLTP proporcionan los datos de origen para el warehouse. Un problema importante de esta relación es que los datos en poder de los sistemas OLTP pueden ser inconsistentes, fragmentados y sujetos a cambios, y contienen entradas duplicadas o faltantes. Como tal, los datos operativos deben "limpiarse" antes de que puedan utilizarse en el data warehouse.

Los sistemas OLTP no están diseñados para responder rápidamente consultas ad hoc. También tienden a no almacenar datos históricos, lo cual es necesario para analizar tendencias. Básicamente, OLTP ofrece grandes cantidades de datos sin procesar, que no se analizan fácilmente. El data warehouse permite responder consultas más complejas además de simples agregaciones como por ejemplo, "¿Cuál es el precio de venta promedio de las propiedades en las principales ciudades del Reino Unido?" Los tipos de consultas que se espera que responda un data warehouse van desde los relativamente simples hasta los más complejos y dependen de los tipos de herramientas de acceso del usuario final que se utilicen.

Problemas de Data Warehousing

Los problemas asociados con el desarrollo y la gestión de un data warehouse se enumeran a continuación:

- Infravalorar los recursos para el ETL de datos

  Muchos desarrolladores subestiman el tiempo necesario para extraer, transformar y cargar (ETL) los datos en el warehouse. Este proceso puede representar una proporción significativa del tiempo total de desarrollo, aunque mejores herramientas ETL están ayudando a reducir el tiempo y el esfuerzo necesarios.

- Problemas ocultos con los sistemas fuente

  Es posible que se identifiquen problemas ocultos asociados con los sistemas de origen que alimentan el data warehouse, posiblemente después de años sin ser detectados. El desarrollador debe decidir si arregla el problema en el data warehouse y / o arregla los sistemas fuente.

- Datos requeridos no capturados

Los proyectos de warehouse a menudo destacan un requisito de datos que no ha sido capturado por los sistemas de origen existentes. La organización debe decidir si modifica los sistemas OLTP o crea un sistema dedicado a capturar los datos que faltan.

- Aumento de las demandas de los usuarios finales

  Una vez que los usuarios finales reciben herramientas de consulta e informes, las solicitudes de apoyo del personal de SI pueden aumentar en lugar de disminuir. Esto se debe a una mayor conciencia de los usuarios sobre las capacidades y el valor del data warehouse. Este problema puede aliviarse parcialmente invirtiendo en herramientas más potentes y fáciles de usar, o proporcionando una mejor formación a los usuarios. Otra razón para el aumento de las demandas del personal de SI es que una vez que un datawarehouse está en línea, a menudo ocurre que el número de usuarios y consultas aumenta junto con las solicitudes de respuesta a consultas cada vez más complejas.

- Homogeneización de datos

  El data warehouse a gran escala puede convertirse en un ejercicio de homogeneización de datos que reduce el valor de los datos. Por ejemplo, al producir una vista consolidada e integrada de los datos de la organización, el diseñador del warehouse puede verse tentado a enfatizar las similitudes en lugar de las diferencias en los datos utilizados por diferentes áreas de aplicación.

- Gran demanda de recursos

  El data warehouse puede utilizar una gran cantidad de espacio en disco. Muchas bases de datos relacionales que se utilizan para el apoyo a la toma de decisiones están diseñadas en torno a esquemas de estrella, copo de nieve y copo de estrella. Estos enfoques dan como resultado la creación de tablas de hechos muy grandes. Si hay muchas dimensiones en los datos, la combinación de tablas e índices agregados para las tablas de hechos puede ocupar más espacio que los datos brutos.

- Propiedad de los datos

  El data warehouse puede cambiar la actitud de los usuarios finales hacia la propiedad de los datos. Los datos confidenciales que originalmente se veían y usaban solo por un departamento o área comercial en particular, como ventas o marketing, ahora pueden estar disponibles para otros miembros de la organización.

- Alto mantenimiento

  Los data warehouses son sistemas de alto mantenimiento. Cualquier reorganización de los procesos comerciales y los sistemas de origen puede afectar el data warehouse. Para seguir siendo un recurso valioso, el data warehouse debe ser coherente con la organización a la que da soporte.

- Proyectos de larga duración

  Un data warehouse representa un único recurso de datos para la organización. Sin embargo, la construcción de un warehouse puede llevar varios años, razón por la cual algunas organizaciones están creando data marts. Los data marts solo admiten los requisitos de un departamento o área funcional en particular y, por lo tanto, se pueden construir de forma más rápida.

- Complejidad de integración

  El área más importante para la gestión de un datawarehouse son las capacidades de integración. Esto significa que una organización debe dedicar una cantidad significativa de tiempo a determinar qué tan bien se pueden integrar las distintas herramientas de data warehouse en la solución general que se necesita. Esta puede ser una tarea muy difícil, ya que hay una serie de herramientas para cada operación del data warehouse, que deben integrarse bien para que el warehouse funcione en beneficio de la organización.

Real-Time Data Warehouse

Cuando los data warehouses aparecieron por primera vez en el mercado como las próximas bases de datos "imprescindibles", fueron reconocidos como sistemas que contenían datos históricos. Se aceptó que estos datos podrían tener hasta una semana de antigüedad y en ese momento se consideró suficiente para satisfacer las necesidades de los responsables de la toma de decisiones corporativas. Sin embargo, desde esos primeros días, el ritmo acelerado de las empresas contemporáneas y la necesidad de que los responsables de las tomas de decisiones accedan a los datos actualizados ha requerido una reducción en el tiempo de demora entre la creación de los datos por parte de los sistemas operativos de primera línea y la capacidad de incluir esos datos en cualquier informe y / o aplicaciones analíticas.

En los últimos años, la tecnología de data warehouse se ha desarrollado para permitir una sincronización más estrecha entre los datos operativos y los datos del warehouse y estos sistemas se conocen como data warehouses en tiempo real (RT) o casi en tiempo real (NRT). Sin embargo, intentar reducir el retraso de tiempo (es decir, la latencia de los datos) entre la creación de datos operativos y la inclusión de estos datos en el warehouse ha generado demandas adicionales en la tecnología del data warehouse. Los principales problemas que enfrentan los desarrolladores de almacenes de datos RT / NRT identificados por Langseth (2004) incluyen:

- Habilitar la extracción, transformación y carga (ETL) de RT / NRT de datos de origen. El problema del almacenamiento de datos RT es reducir la ventana ETL para permitir la carga de datos RT / NRT sin tiempo de inactividad o con un tiempo de inactividad mínimo para los usuarios del almacenamiento de datos.
- Modelado de tablas de hechos de RT. El problema con el modelado de datos RT dentro del almacén es cómo integrar los datos RT con los otros datos agregados de diversas formas que ya se encuentran en el almacén.
- Consultas OLAP versus cambio de datos. El problema es que las herramientas OLAP asumen que los datos que se consultan son estáticos y no cambian. Las herramientas no tienen protocolos para tratar con datos de destino que se complementan con nuevos datos durante la vida útil de la consulta.
- Escalabilidad y contención de consultas. El problema es que la escalabilidad y la contención de consultas fue una de las principales razones para separar los sistemas operativos de los sistemas analíticos y, por lo tanto, cualquier cosa que devuelva el problema al entorno del almacén no se concilia fácilmente.

3 ARQUITECTURA DE UN DATA WAREHOUSE

Datos operacionales

La fuente de datos para el data warehouse proviene de:

- Datos operacionales de mainframe almacenados en bases de datos jerárquicas de primera generación y de red.
- Datos departamentales almacenados en sistemas de ficheros propietarios tales como VSAM, RMS y DBMS relacionales como por ejemplo Informix y Oracle.
- Datos privados almacenados en estaciones de trabajo y servidores privados.
- Sistemas externos como Internet, bases de datos disponibles comercialmente o bases de datos asociadas con los proveedores o clientes de una organización.

Almacén operacional de datos

Un almacén operacional de datos (ODS) es un repositorio de datos operacionales actuales e integrados que se utilizan para el análisis. A menudo se estructura y se suministran datos de la misma manera que el data warehouse, pero de hecho puede actuar simplemente como un área de preparación para que los datos se muevan al warehouse.

El ODS se crea a menudo cuando se descubre que los sistemas operacionales heredados son incapaces de cumplir con los requisitos de informes. El ODS proporciona a los usuarios la facilidad de uso de una base de datos relacional mientras permanece alejado de las funciones de soporte de decisiones del data warehouse.

La creación de un ODS puede ser un paso útil hacia la construcción de un data warehuse, porque un ODS puede proporcionar datos que ya se extrajeron de los sistemas de origen y se limpiaron. Esto significa que se simplifica el trabajo restante de integrar y reestructurar los datos para el data warehouse.

Administrador ETL

El administrador de ETL realiza todas las operaciones asociadas con el ETL de datos en el warehouse. Los datos pueden extraerse directamente de las fuentes de datos o, más comúnmente, del almacén operacional de datos.

Administrador Warehouse

El administrador Warehouse lleva a cabo todas las operaciones asociadas con la gestión de los datos en el warehouse. Las operaciones realizadas por el administrador warehouse incluyen:

- análisis de datos para garantizar la coherencia;
- transformación y fusión de datos de origen del almacenamiento temporal en tablas de warehouse;
- creación de índices y vistas en tablas base;
- generación de desnormalizaciones (si es necesario);
- generación de agregaciones;
- copias de seguridad y archivado de datos.

En algunos casos, el administrador warehouse también genera perfiles de consulta para determinar qué índices y agregaciones son apropiados. Se puede generar un perfil de consulta para cada usuario,

grupo de usuarios o el data warehouse, y se basa en la información que describe las características de las consultas, como la frecuencia, la (s) tabla (s) objetivo y el tamaño de los conjuntos de resultados.

Administrador de consultas

El administrador de consultas realiza todas las operaciones asociadas con la gestión de las consultas de los usuarios. La complejidad del administrador de consultas está determinada por las facilidades proporcionadas por las herramientas de acceso del usuario final y la base de datos. Las operaciones realizadas por este componente incluyen dirigir consultas a las tablas apropiadas y programar la ejecución de consultas. En algunos casos, el administrador de consultas también genera perfiles de consulta para permitir que el administrador del warehouse determine qué índices y agregaciones son apropiados.

Datos detallados

Esta área del warehouse almacena todos los datos detallados en el esquema de la base de datos. En la mayoría de los casos, los datos detallados no se almacenan en línea, sino que están disponibles agregando los datos al siguiente nivel de detalle. Sin embargo, de forma regular, se agregan datos detallados en el warehouse para complementar los datos agregados.

Datos muy resumidos y ligeros

Esta área del warehouse almacena todos los datos predefinidos, ligeros y altamente resumidos (agregados) generados por el administrador del warehouse. Esta área del warehouse es transitoria, ya que estará sujeta a cambios de forma continua para responder a los perfiles de consulta cambiantes.

El propósito de la información resumida es acelerar el rendimiento de las consultas. Aunque existen mayores costes operativos asociados con el resumen inicial de los datos, esto se compensa al eliminar el requisito de realizar continuamente operaciones de resumen (como ordenar o agrupar) al responder las consultas de los usuarios. Los datos resumidos se actualizan cuando se cargan nuevos datos en el almacén.

Datos de archivo / copia de seguridad

Esta área del warehouse almacena los datos detallados y resumidos con fines de archivo y respaldo. Aunque los datos resumidos se generan a partir de datos detallados, puede ser necesario realizar una copia de seguridad de los datos resumidos en línea si estos datos se conservan más allá del período de retención de datos detallados.

Metadatos

Esta área del warehouse almacena todas las definiciones de metadatos (datos sobre datos) utilizadas por todos los procesos del almacén. Los metadatos se utilizan para una variedad de propósitos, que incluyen:

- los procesos de extracción y carga: los metadatos se utilizan para mapear las fuentes de datos a una vista común de los datos dentro del almacén;
- el proceso de gestión del almacén: los metadatos se utilizan para automatizar la producción de tablas de resumen;
- como parte del proceso de gestión de consultas: los metadatos se utilizan para dirigir una consulta a la fuente de datos más adecuada.

La estructura de los metadatos difiere entre cada proceso, porque el propósito es diferente. Esto significa que en el data warehouse se almacenan múltiples copias de metadatos que describen el mismo elemento de datos. Además, la mayoría de las herramientas de los proveedores para la gestión

de copias y el acceso a los datos del usuario final utilizan sus propias versiones de metadatos. Específicamente, las herramientas de administración de copias usan metadatos para comprender las reglas de mapeo que se deben aplicar para convertir los datos de origen en una forma común. Las herramientas de acceso del usuario final utilizan metadatos para comprender cómo crear una consulta. La gestión de metadatos dentro del data warehouse es una tarea muy compleja que no debe subestimarse.

Herramientas de acceso para usuarios finales

El propósito principal del data warehousing es ayudar a los usuarios en la toma de decisiones. Estos usuarios interactúan con el warehouse mediante herramientas de acceso de usuario final. El data warehouse debe soportar de manera eficiente análisis ad hoc y de rutina. El alto rendimiento se logra mediante la planificación previa de los requisitos para uniones, sumas e informes periódicos por parte de los usuarios finales.

Aunque las definiciones de las herramientas de acceso del usuario final pueden superponerse, para el propósito de esta discusión, clasificamos estas herramientas en cuatro grupos principales:

- herramientas de informes y consultas;
- herramientas de desarrollo de aplicaciones;
- herramientas OLAP;
- herramientas de minería de datos.

Herramientas de informes y consultas

Las herramientas de informes incluyen herramientas de generación de informes y editores de informes. Las herramientas de generación de informes se utilizan para generar informes operacionales regulares o respaldar trabajos por lotes de gran volumen, como pedidos / facturas de clientes y cheques de pago del personal. Los editores de informes, por otro lado, son herramientas de escritorio económicas diseñadas para usuarios finales.

Las herramientas de consulta para warehouses de datos relacionales están diseñadas para aceptar SQL o generar sentencias SQL para consultar datos almacenados en el warehouse. Estas herramientas protegen a los usuarios finales de las complejidades de las estructuras de bases de datos y SQL al incluir una metacapa entre los usuarios y la base de datos. La metacapa es el software que proporciona vistas orientadas a sujetos de una base de datos y admite la creación de SQL mediante el uso de "apuntar y hacer clic". Un ejemplo de una herramienta de consulta es Query-By-Example (QBE).

Herramientas de desarrollo de aplicaciones

Los requisitos de los usuarios finales pueden ser tales que las capacidades integradas de las herramientas de informes y consultas sean inadecuadas, ya sea porque no se puede realizar el análisis requerido o porque la interacción del usuario requiere un nivel excesivamente alto de experiencia por parte del usuario. En esta situación, el acceso de los usuarios puede requerir el desarrollo de aplicaciones internas que utilicen herramientas gráficas de acceso a datos diseñadas principalmente para entornos cliente-servidor. Algunas de estas herramientas de desarrollo de aplicaciones se integran con herramientas OLAP populares y pueden acceder a los principales sistemas de bases de datos, incluidos Oracle, Sybase e Informix.

Herramientas de procesamiento analítico en línea (OLAP)

Las herramientas OLAP se basan en el concepto de bases de datos multidimensionales y permiten a un usuario sofisticado analizar los datos utilizando vistas complejas y multidimensionales.

Las aplicaciones comerciales típicas de estas herramientas incluyen la evaluación de la eficacia de una campaña de marketing, la previsión de ventas de productos y la planificación de la capacidad. Estas herramientas asumen que los datos están organizados en un modelo multidimensional apoyado por

una base de datos multidimensional especial (MDDB) o por una base de datos relacional diseñada para permitir consultas multidimensionales.

Herramientas de minería de datos

La minería de datos es el proceso de descubrir nuevas correlaciones, patrones y tendencias significativas mediante la extracción de grandes cantidades de datos mediante técnicas estadísticas, matemáticas y de inteligencia artificial. La minería de datos tiene el potencial de reemplazar las capacidades de las herramientas OLAP, ya que el principal atractivo de la minería de datos es su capacidad para construir modelos predictivos en lugar de retrospectivos.

4 HERRAMIENTAS Y TECNOLOGIAS DE DATA WAREHOUSING

EXTRACCIÓN, TRANSFORMACIÓN Y CARGA (ETL, por sus siglas en inglés, Extraction, Transformation and Loading)

Uno de los beneficios más comúnmente citados asociados con los data warehouses empresariales (EDW) es que estos sistemas centralizados proporcionan una visión empresarial integrada de los datos corporativos. Sin embargo, lograr esta valiosa visión de los datos puede resultar muy complejo y llevar mucho tiempo. Los datos destinados a un EDW primero deben extraerse de una o más fuentes de datos, transformarse en un formulario que sea fácil de analizar y coherente con los datos que ya se encuentran en el warehouse y, finalmente, cargarse en el EDW. Todo este proceso se denomina proceso de extracción, transformación y carga (ETL) y es un proceso crítico en cualquier proyecto de data warehouse.

Extracción

El paso de extracción apunta a una o más fuentes de datos para el EDW; estas fuentes suelen incluir bases de datos OLTP, pero también pueden incluir fuentes como bases de datos personales y hojas de cálculo, ficheros de planificación de recursos empresariales (ERP) y ficheros de log del uso de la web. Las fuentes de datos son normalmente internas, pero también pueden incluir fuentes externas, como los sistemas utilizados por proveedores y/o clientes.

La complejidad del paso de extracción depende de cuán similares o diferentes sean los sistemas fuente para el EDW. Si los sistemas de origen están bien documentados, bien mantenidos, se ajustan a los formatos de datos de toda la empresa y utilizan la misma tecnología o una similar, el proceso de extracción debería ser sencillo. Sin embargo, el otro extremo es que los sistemas fuente estén mal documentados y mantenidos utilizando diferentes formatos de datos y tecnologías. En este caso, el proceso ETL será muy complejo. El paso de extracción normalmente copia los datos extraídos en un almacenamiento temporal denominado almacén operacional de datos (ODS) o área de preparación (SA).

Los problemas adicionales asociados con el paso de extracción incluyen establecer la frecuencia para las extracciones de datos de cada sistema fuente al EDW, monitorizar cualquier modificación de los sistemas fuente para garantizar que el proceso de extracción siga siendo válido y monitorizar cualquier cambio en el rendimiento o la disponibilidad de los sistemas fuente que pueda tener un impacto en el proceso de extracción.

Transformación

El paso de transformación aplica una serie de reglas o funciones a los datos extraídos, lo que determina cómo se utilizarán los datos para el análisis y puede involucrar transformaciones como sumas de datos, codificación de datos, fusión de datos, división de datos, cálculos de datos y creación de claves subrogadas. El resultado de las transformaciones son datos limpios y consistentes con los datos que ya se encuentran en el warehouse y, además, están en un formato que está listo para el análisis de los usuarios del warehouse. Aunque las sumas de datos se mencionan como una posible transformación, ahora se recomienda comúnmente que los datos en el warehouse también se mantengan en el nivel más bajo de granularidad posible. Esto permite a los usuarios realizar consultas sobre los datos del EDW que son capaces de profundizar en los datos más detallados.

Carga

La carga de los datos en el warehouse se puede llevar a cabo después de que se hayan realizado todas las transformaciones o como parte del proceso de transformación. A medida que los datos se cargan en el warehouse, se aplicarán restricciones adicionales definidas en el esquema de la base de datos, así como en los triggers (disparadores) activados para la carga de datos (como la unicidad, la integridad referencial y los campos obligatorios), que también contribuyen al rendimiento general de la calidad de los datos del proceso ETL.

En el warehouse, los datos pueden someterse a sumas adicionales y/o reenviarse posteriormente a otras bases de datos asociadas, como data marts, o para alimentar aplicaciones particulares como la gestión de recursos del cliente (CRM). Los problemas importantes relacionados con el paso de carga vienen determinados por la frecuencia de carga y establecen cómo la carga afectará a la disponibilidad del data warehouse.

Herramientas ETL

El proceso ETL puede llevarse a cabo mediante programas personalizados o mediante herramientas ETL comerciales. Al inicio de los data warehouse, no era raro que el proceso ETL se llevara a cabo utilizando programas personalizados, pero el mercado de herramientas ETL ha crecido y ahora hay una gran selección de herramientas ETL. Las herramientas no solo automatizan el proceso de extracción, transformación y carga, sino que también pueden ofrecer características adicionales como la elaboración de perfiles de datos, el control de calidad de datos y la gestión de metadatos.

Perfilado de datos y control de calidad de datos

EL perfilado de datos proporciona información importante sobre la cantidad y la calidad de los datos provenientes de los sistemas fuentes. Por ejemplo, el perfilado de datos puede indicar cuántas filas tienen entradas de datos faltantes, incorrectas o incompletas y la distribución de valores en cada columna. Esta información puede ayudar a identificar los pasos de transformación necesarios para limpiar los datos y/o cambiar los datos a una forma adecuada para la cargar en el warehouse.

Gestión de metadatos

Para comprender completamente los resultados de una consulta, a menudo es necesario considerar el histórico de los datos incluidos en el conjunto de resultados. En otras palabras, ¿qué ha sucedido con los datos durante el proceso ETL? La respuesta a esta pregunta se encuentra en un área de almacenamiento denominada repositorio de metadatos. Este repositorio es administrado por la herramienta ETL y almacena información sobre los datos del warehouse con respecto a los detalles del sistema fuente, detalles de cualquier transformación de los datos y detalles de cualquier fusión o división de datos. Este histórico completo de datos está disponible para los usuarios del data warehouse y puede facilitar la validación de los resultados de la consulta o proporcionar una explicación para alguna anomalía mostrada en el conjunto de resultados que fue causada por el proceso ETL.

DATA WAREHOUSE DBMS

Hay pocos problemas de integración asociados con la base de datos del data warehouse. Debido a la madurez de dichos productos, la mayoría de las bases de datos relacionales se integrarán de manera predecible con otros tipos de software. Sin embargo, existen problemas asociados con el tamaño potencial de la base de datos del data warehouse. El paralelismo en la base de datos se convierte en un tema importante, así como los problemas habituales como el rendimiento, la escalabilidad, la disponibilidad y la capacidad de administración, que deben tenerse en cuenta al elegir un DBMS.

Requisitos para Data Warehouse DBMS

Los requisitos especializados para un DBMS relacional adecuado para el data warehouse están publicados en un white paper (Red Brick Systems, 1996) y se enumeran a continuación:

- ***Rendimiento de carga***. Los data warehouse requieren carga incremental de nuevos datos de forma periódica dentro de ventanas de tiempo estrechas. El rendimiento del proceso de carga debe medirse en cientos de millones de filas o gigabytes de datos por hora y no debe haber un límite máximo que restrinja el negocio.
- ***Procesamiento de carga.*** Se deben llevar a cabo muchos pasos para cargar datos nuevos o actualizados en el almacén de datos, incluidas conversiones de datos, filtrado, reformateo, verificaciones de integridad, almacenamiento físico, indexación y actualización de metadatos. Aunque cada paso puede ser atómico en la práctica, el proceso de carga debería parecer que se ejecuta como una única unidad de trabajo sin fisuras.
- ***Gestión de la calidad de los datos.*** El cambio a la gestión basada en hechos exige la máxima calidad de datos. El warehouse debe garantizar la coherencia local, la coherencia global y la integridad referencial a pesar de las fuentes "sucias" y los tamaños masivos de bases de datos. Si bien la carga y la preparación son pasos necesarios, no son suficientes. La capacidad de responder a las consultas de los usuarios finales es la medida del éxito para una aplicación de data warehouse. A medida que se responden más preguntas, los analistas tienden a hacer preguntas más creativas y complejas.
- ***Rendimiento de las consultas.*** La gestión basada en hechos y el análisis ad hoc no deben ralentizarse ni inhibirse por el rendimiento del DBMS del data warehouse. Las consultas grandes y complejas para operaciones comerciales clave deben completarse en períodos de tiempo razonables.
- ***Altamente escalable***. Los tamaños de los data warehouse están creciendo a un ritmo enorme, con tamaños que normalmente van desde un terabyte (10^{12} bytes) hasta un petabyte (10^{15} bytes). El DBMS no debe tener limitaciones de arquitectura para el tamaño de la base de datos y debe soportar gestión modular y paralela. En caso de fallo, el DBMS debe soportar la disponibilidad continua y proporcionar mecanismos de recuperación. El DBMS debe soportar dispositivos de almacenamiento masivo como discos ópticos y dispositivos de gestión de almacenamiento jerárquico. Por último, el rendimiento de la consulta no debe depender del tamaño de la base de datos, sino de la complejidad de la consulta.
- ***Escalabilidad de usuarios masivos.*** El pensamiento actual es que el acceso a un data warehouse está limitado a un número relativamente bajo de usuarios gerenciales. Es poco probable que esto siga siendo cierto a medida que los data warehouse van ganando más valor. Se prevé que el DBMS del data warehouse debería ser capaz de admitir cientos, o incluso miles de usuarios simultáneos y, al mismo tiempo, mantener un rendimiento de consulta aceptable.
- ***Almacén de datos en red.*** Los sistemas de data warehouse deben ser capaces de cooperar en una red más grande de data warehouse. El data warehouse debe incluir herramientas que coordinen el movimiento de subconjuntos de datos entre warehouses. Los usuarios deben poder examinar y trabajar con múltiples data warehouses desde una sola estación de trabajo cliente.
- ***Administración de warehouse***. La naturaleza cíclica en el tiempo y a muy gran escala del data warehouse exige facilidad y flexibilidad administrativa. El DBMS debe proporcionar controles para implementar límites de recursos, contabilidad de contracargo para asignar costes a los usuarios y priorización de consultas para abordar las necesidades de diferentes clases y actividades de usuarios. El DBMS también debe proporcionar seguimiento y ajuste de la carga de trabajo para que los recursos del sistema puedan optimizarse para obtener el

máximo rendimiento y resultado. El valor más visible y medible de implementar un data warehouse se evidencia en el acceso creativo y desinhibido a los datos que proporciona a los usuarios finales.

- ***Análisis dimensional integrado.*** El poder de las vistas multidimensionales está ampliamente aceptado, y el soporte dimensional debe ser inherente al DBMS del warehouse para proporcionar el mayor rendimiento para las herramientas OLAP relacionales. El DBMS debe soportar la creación rápida y sencilla de resúmenes precalculados comunes en grandes data warehouses y proporcionar herramientas de mantenimiento para automatizar la creación de estos agregados precalculados. El cálculo dinámico de agregados debe ser coherente con las necesidades de rendimiento interactivo del usuario final.
- ***Funcionalidad de consulta avanzada.*** Los usuarios finales requieren cálculos analíticos avanzados, análisis secuencial y comparativo y acceso constante a datos detallados y resumidos. El uso de SQL en un entorno de herramientas cliente-servidor de "point-and-click" puede resultar a veces poco práctico o incluso imposible debido a la complejidad de las consultas de los usuarios. El DBMS debe proporcionar un conjunto completo y avanzado de operaciones analíticas.

DBMS en paralelo

Data warehousing requiere procesar enormes cantidades de datos y la tecnología de bases de datos paralelas ofrece una solución para proporcionar el crecimiento necesario para el rendimiento. El éxito de los DBMS en paralelo depende del funcionamiento eficiente de muchos recursos, incluidos procesadores, memoria, discos y conexiones de red. A medida que el data warehouse crece en popularidad, muchos proveedores están construyendo grandes DBMS de soporte para decisiones utilizando tecnologías paralelas. El objetivo es resolver problemas de apoyo a la toma de decisiones utilizando varios nodos que trabajan en el mismo problema. Las principales características de los DBMS en paralelo son la escalabilidad, la operatividad y la disponibilidad.

Los DBMS en paralelo realiza muchas operaciones de base de datos simultáneamente, dividiendo las tareas individuales en partes más pequeñas para que las tareas se puedan distribuir en varios procesadores. Los DBMS en paralelo deben poder ejecutar consultas en paralelo. En otras palabras, deben poder descomponer consultas grandes y complejas en subconsultas, ejecutar las subconsultas separadas simultáneamente y volver a ensamblar los resultados al final. La capacidad de dichos DBMS también debe incluir la carga de datos en paralelo, el escaneo de tablas y el archivado de datos y copias de resplado. Hay dos arquitecturas de hardware paralelas principales que se utilizan comúnmente como plataformas de servidores de bases de datos para el almacenamiento de datos:

- Multiprocesamiento simétrico (SMP): un conjunto de procesadores estrechamente acoplados que comparten memoria y almacenamiento en disco;
- Procesamiento masivo paralelo (MPP): un conjunto de procesadores débilmente acoplados, cada uno de los cuales tiene su propia memoria y almacenamiento en disco.

DATA WAREHOUSE METADATA

Hay muchos problemas asociados con la integración de los data warehouse; en esta sección nos enfocamos en la integración de metadatos, es decir, “datos sobre datos” (Darling, 1996).

La gestión de los metadatos en el warehouse es una tarea extremadamente compleja y difícil. Los metadatos se utilizan para una variedad de propósitos y la administración de metadatos es un tema crítico para lograr un data warehouse completamente integrado.

El propósito principal de los metadatos es mostrar el camino de regreso al lugar donde comenzaron los datos, para que los administradores del warehouse conozcan el histórico de cualquier item en el warehouse. Sin embargo, el problema es que los metadatos tienen varias funciones dentro del warehouse que se relacionan con los procesos asociados con la transformación y carga de datos, la gestión del data warehouse y la generación de consultas.

Los metadatos asociados con la transformación y carga de datos deben describir los datos de origen y los cambios que se realizaron en los datos. Por ejemplo, para cada campo de origen debe haber un identificador único, el nombre del campo original, el tipo de datos de origen y la ubicación original, incluido el sistema y el nombre del objeto, junto con el tipo de datos de destino y el nombre de la tabla de destino. Si el campo está sujeto a alguna transformación, como un cambio de tipo de campo simple a un conjunto complejo de procedimientos y funciones, esto también debe registrarse.

Los metadatos asociados con la gestión de datos describen los datos tal como se almacenan en el warehouse. Es necesario describir cada objeto de la base de datos, incluidos los datos de cada tabla, índice y vista, y las restricciones asociadas. Esta información se encuentra en el catálogo del sistema DBMS; sin embargo, existen requisitos adicionales para los propósitos del warehouse. Por ejemplo, los metadatos también deben describir cualquier campo asociado con agregaciones, incluida una descripción de la agregación que se realizó. Además, se deben describir las particiones de la tabla, incluida la información sobre la clave de partición y el rango de datos asociado con esa partición.

El administrador de consultas también requiere los metadatos descritos anteriormente para generar las consultas adecuadas. A su vez, el gestor de consultas genera metadatos adicionales sobre las consultas que se ejecutan, que se pueden utilizar para generar un histórico de todas las consultas y un perfil de consulta para cada usuario, grupo de usuarios o el data warehouse. También hay metadatos asociados con los usuarios de las consultas que incluyen, por ejemplo, información que describe lo que significa el término "precio" o "cliente" en una base de datos en particular y si el significado ha cambiado con el tiempo.

Sincronizar metadatos

El principal problema de integración es cómo sincronizar los distintos tipos de metadatos utilizados en todo el data warehouse. Las diferentes herramientas de un data warehouse generan y utilizan sus propios metadatos, y para lograr la integración, necesitamos que estas herramientas sean capaces de compartir sus metadatos. El desafío es sincronizar metadatos entre diferentes productos de diferentes proveedores utilizando diferentes data warehouses. Por ejemplo, es necesario identificar el elemento de metadatos correcto en el nivel de detalle correcto de un producto y asignarlo al elemento de metadatos adecuado en el nivel de detalle correcto de otro producto, luego resolver las diferencias de codificación entre ellos. Esto debe repetirse para todos los demás metadatos que los dos productos tienen en común. Además, cualquier cambio en los metadatos (o incluso meta-metadatos) de un producto debe transmitirse al otro producto. La tarea de sincronizar dos productos es muy compleja y, por lo tanto, repetir este proceso para todos los productos que componen el data warehouse puede requerir muchos recursos. Sin embargo, se debe lograr la integración de los metadatos.

Al principio, había dos estándares principales para metadatos y modelado en las áreas de data warehousing y desarrollo basado en componentes propuestos por Meta Data Coalition (MDC) y Object Management Group (OMG). Sin embargo, estas dos organizaciones de la industria anunciaron conjuntamente que el MDC se fusionaría con el OMG. Como resultado, el MDC interrumpió las operaciones independientes y continuó el trabajo en el OMG para integrar los dos estándares.

La fusión de MDC con OMG marcó un acuerdo de los principales proveedores de data warehousing y metadatos para converger en un estándar, incorporando lo mejor del Modelo de información abierto del MDC con lo mejor del metamodelo común del warehouse de OMG. Un solo estándar permite a los usuarios intercambiar metadatos entre diferentes productos de diferentes proveedores libremente.

El CWM de OMG se basa en varios estándares, incluido el UML (Lenguaje de modelado unificado) de OMG, XMI (Intercambio de metadatos XML) y MOF (Meta Object Facility) y en el OIM de MDC. El CWM fue desarrollado por varias empresas, incluidas IBM, Oracle, Unisys, Hyperion, Genesis, NCR, UBS y Dimension EDI.

Herramientas de administración y gestión

Un data warehouse requiere herramientas para respaldar la administración y gestión de un entorno tan complejo. Estas herramientas deben ser capaces de soportar las siguientes tareas:

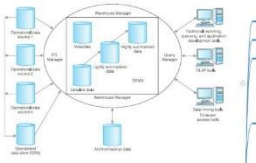
- monitorizar la carga de datos de múltiples fuentes;
- controles de calidad e integridad de los datos;
- gestionar y actualizar metadatos;
- monitorizar el desempeño de la base de datos para asegurar tiempos de respuesta a consultas y utilización de recursos eficientes;
- auditar el uso del data warehouse para proporcionar a los usuarios información de contracargo;
- replicar, dividir en subconjuntos y distribuir datos;
- mantener una gestión eficiente de almacenamiento de los datos;
- depuración de datos;
- archivar y realizar copias de seguridad de los datos;
- implementar la recuperación después de un fallo;
- gestión de seguridad.

5 RESUMEN ESQUEMÁTICO

- **Data Warehousing**
  - **Introducción a Data Warehousing**
    - Implantar un DW tiene una impacto en el desarrollo estratégico de las organizaciones y supone una gran inversión, tanto en la implantación como sobre todo en el mantenimiento
    - DW
      - El DW se hace necesario por 2 aspectos fundamentales:
        - La dispersión de la información dentro de una misma empresa (depts que tienen sus propios SI)
        - Dependencia del resto de depts de un único dept de Sistemas que da soporte a las diversas aplicaciones de los distintos depts para la obtención de la información que necesitan (no hay accesibilidad por parte de todo el mundo)
      - Permite
        - Visión única de la información de una empresa (que es consistente,segura, mantenible, actualizada y fácilmente accesible)
        - No afectar el tiempo de respuesta de los sistemas operacionales (la información se extrae de los sistemas al Almacén de Datos)
        - Reducir las dependencias con el personal de sistemas para realizar consultas (los usuarios son ahora autosuficientes)
    - Implica
      - Aspectos técnicos-económicos...
      - Aspectos organizativos
  - **Conceptos de Data Warehousing**
    - Concepto
      - Son sistemas de apoyo al análisis y a la toma de decisiones.
      - Se enmarcan dentro de los sistemas DSS y EIS (especialización DSS para la Dirección)
    - DW
      - Def
        - Bill Inmon
          - Características de los datos en un DW
            - Orientados a temas
              - los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí
            - Integrados
              - contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes
            - Cambiantes con el tiempo
              - Se registran los cambios producidos en el tiempo por si fuera necesario para un informe que refleje los cambios
            - No volátiles
              - La información no se modifica ni se elimina, una vez almacedado es una información de sólo lectura
          - Enfoque top-down de construcción
            - plantea una metodología descendente para el diseño de un almacén de datos:
              - Primero se consideran todos los datos corporativos (DW completo)
              - Después se plantean Data Marts si fueran necesarios
      - Ventajas
        - Alto potencial del retorno de la inversión
        - Ventaja competitiva
        - Mayor productividad de los responsables de las tomas de decisiones corporativas
      - Sistemas OLTP vs DW

Características	Sistemas OLTP	Sistemas Data Warehousing
Objetivo principal	Ofrece procesamiento operacional	Ofrece procesamiento analítico
Edad de los datos	Actual	Histórico (aunque la tendencia es ir hacia algo que incluya también datos actuales)
Latencia de los datos	Tiempo real	Depende de la duración del ciclo de los complementos de data warehouse (pero la tendencia es ir hacia los complementos en tiempo real)
Granularidad de los datos	Datos detallados.	Datos detallados, ligeramente resumidos y muy resumidos.
Procesamiento de los datos	Pobremas predictivos de inserciones, eliminado y actualizaciones de datos, y consultas. Alto nivel de rendimiento de transacciones.	Tareas muy predictivas de consultas de datos. Nivel medio a bajo de rendimiento de transacciones.
Informes	Informes predictivos, unidimensionales, relativamente estáticos.	Informes impredecibles, multidimensionales y dinámicos.
Usuarios	Atienda a un gran número de usuarios operacionales.	Atienda a un menor número de usuarios de la gerencia (aunque la tendencia es ir hacia requisitos similares a soportes usuarios operacionales)

      - Problemas de DW
        - Infravalorar los recursos para el ETL de datos
        - Problemas ocultos con los sistemas fuente
        - Datos requeridos no capturados
        - Aumento de las demandas de usuarios finales
        - Homogeneización de datos
        - Gran demanda de recursos
        - Propiedad de los datos
        - Alto mantenimiento
        - Proyectos de larga duración
        - Complejidad de integración
      - Real Time DW
        - Sistema que refleja el estado del warehouse en tiempo real. La mayoría de los warehouses contienen datos que tienen una alta latencia. Un data warehouse RT tiene datos de baja latencia y proporciona datos en tiempo real.
  - **Arquitectura de un DW**



    - Datos operacionales
    - Almacen operacional de datos ODS
    - Administrador ETL
    - Administrador Warehouse
      - análisis de datos para garantizar la coherencia:
      - transformación y fusión de datos de origen del almacenamiento temporal en tablas de warehouse:
      - creación de índices y vistas en tablas base:
      - generación de desnormalizaciones (si es necesario):
      - generación de agregaciones:
      - copias de seguridad y archivado de datos.
    - Administrador de consultas
      - realiza todas las operaciones asociadas con la gestión de las consultas de los usuarios
    - Datos detallados
      - almacena todos los datos detallados en el esquema de la base de datos
    - Datos muy resumidos y ligeros
      - almacena todos los datos predefinidos, ligeros y altamente resumidos (agregados)
    - Datos de archivo/backup
    - Metadatos
    - Herramientas de acceso para usuarios finales
      - Herramientas de informes y consultas
      - Herramientas de desarrollo de aplicaciones
      - Herramientas OLAP
      - Herramientas de minería de datos
  - **Herramientas y tecnologías de DW**
    - ETL
      - Es práctica común normalizar los datos antes de combinarlos en el DW mediante herramientas ETL (Extraction, Transformation and Load)
      - Extracción
        - Extrae la informacion de los distintos sistemas
          - Sistemas operacionales: facturación, clientes, ventas, averías, peticiones...
          - Sistemas externos
        - Lenguajes utilizados para la programación de las extracciones: C,PRO*C, PL/SQL, Scripts UNIX...
      - Filtrado
        - Proceso en el que depuran los datos provenientes de los sistemas operacionales o externos (los registros eliminados se registran para su posterior tratamiento de forma manual)
        - Incluye las siguientes tareas:
          - Eliminación de registros defectuosos (quedan registrados en ficheros de log para su posterior análisis)
          - Eliminación de registros duplicados.
          - Verificación de la consistencia de los datos:
          - Verificación de la coherencia de los datos (que un registro está en concordancia con el resto de los registros)
        - Lenguajes utilizados para la programación de las extracciones, filtrado,... C,PRO*C, PL/SQL, Scripts UNIX...
      - Transformación
        - Hay que transformar los datos (según lo def. por los metadatos) para unificar la codificación de los datos, normalizar los datos... (Ej: distintos dept pueden usar distintas notaciones o códigos para los mismos datos)
        - Tras la transformación los datos se dejan listos para que para el servidor de presentación, ya sea un gestor relacional o un motor OLAP
      - Carga
        - Carga final de los datos en el DW
      - Agregación
        - Adicionalmente la agregación permite agilizar las consultas al agregar la información en tablas agregadas (a partir de las tablas detalladadas/tablas de hechos)
    - DW DBMS - Requisitos
      - Rendimiento de carga
      - Procesamiento de carga
      - Gestión de calidad de datos
      - Rendimiento consultas
      - Altamente escalable
      - Escalabilidad de usuarios masivos
      - Almacen de datos en red
      - Administración de warehouse
      - Analisis dimensional integrado
      - Funcionalidad de consulta avanzada
    - DBMS en paralelo
      - Multiprocesamiento simétrico (SMP): un conjunto de procesadores estrechamente acoplados que comparten memoria y almacenamiento en disco:
      - Procesamiento masivo paralelo (MPP): un conjunto de procesadores débilmente acoplados, cada uno de los cuales tiene su propia memoria y almacenamiento en disco.
    - Metadato
      - Es un elemento fundamental de la arquitectura DW
      - Define la estructura de los datos que se van a almacenar y cómo se relacionan.
      - El metadato documenta, entre otras cosas, qué tablas existen en una base de datos, qué columnas posee cada una de las tablas y qué tipo de datos se pueden almacenar.
        - Tablas
        - Columnas de tablas
        - Relaciones entre tablas
        - Jerarquías y Dimensiones de datos
        - Entidades y Relaciones
      - Los datos son de interés para el usuario final, el metadato es de interés para los programas que tienen que manejar estos datos.
      - La función del metadato en entorno DW es recoger todas las definiciones de la organización y el concepto de los datos en el almacén de datos

6 GLOSARIO

OLTP On Line Transaction Processing
DBMS Data Base Management System
ETL Extract, Transform, and Load (extraer, transformar y cargar)
EDW Enterprise Data Warehouse, Data Warehouse empresarial
ODS operational data store
SA Staging area
CRM Customer Resource Management
SMP Symmetric multiprocessing
MMP Massively parallel processing
MDC Meta Data Coalition
OMG Object Management Group

7 BIBLIOGRAFÍA BÁSICA

T. Connolly and C. Begg. Database systems: a practical approach to design, implementation, and management (6th ed.)