XUNTA DE GALICIA

ESCOLA GALEGA DE ADMINISTRACIÓN PÚBLICA

22. DATAWAREHOUSE. DATA MARTS. ARQUITECTURA. ANÁLISIS MULTIDIMENSIONAL Y ARQUITECTURAS OLAP. ROLAP/MOLAP/HOLAP. MINERÍA DE DATOS. GENERACIÓN DE INFORMES A LA DIRECCIÓN.

Fecha de actualización 26 de julio de 2011

XUNTA DE GALICIA

ESCOLA GALEGA DE ADMINISTRACIÓN PÚBLICA

Tema 22: Datawarehouse. Data Marts. Arquitectura. Análisis multidimensional y arquitecturas OLAP. ROLAP/MOLAP/HOLAP. Minería de datos. Generación de informes para la dirección.

ÍNDICE

22.1 DATAWAREHOUSE

22.1.1 Estructura Multidimensional

La estructura multidimensional de bases de datos es una variante del modelo relacional, la cual hace uso de estructuras multidimensionales en las cuales mantiene la información organizada y en las que es capaz de expresar a su vez las relaciones entre los datos contenidos en ellas. Estas estructuras multidimensionales son visualizadas como cubos datos que a su vez pueden contener otros cubos datos, considerando cada una de las caras de los cubos como una dimensión de los datos.

Las celdas que conforman la estructura multidimensional contienen información agregada que se relaciona con los elementos a lo largo de cada una de sus dimensiones. Es decir, una única celda puede llegar a contener las ventas totales de un determinado artículo en una zona geográfica específica para un tipo de venta concreto en un período determinado (ventas del procesador XMX-01, en Galicia, a través del portal web, en el mes de abril).

El principal aporte que ofrece esta estructuración multidimensional es que supone un modelo compacto y fácilmente comprensible, lo cual permite la manipulación y visualización de elementos de datos que poseen un elevado número de interrelaciones.

Debido a todo esto las estructuras multidimensionales han pasado a formar parte de las estructuras de bases de datos más utilizadas, llegando a convertirse en las estructuras más importantes para las bases de datos analíticas que soportan las aplicaciones que llevan a cabo *procesamiento analítico en línea, OLAP,* en donde es vital obtener una respuesta rápida al realizarse una serie de consultas de elevada complejidad.

Generalmente una organización o entidad almacena su datos en bases de datos diseñadas para introducir y almacenar datos mediante el proceso OLTP (On-Line Transaction Process, Proceso de Transacciones On-Line). Este proceso realiza de una manera idónea las tareas de inserción, modificación o borrado de registros, pero resulta ineficiente a la hora de realizar consultas complejas. Los Datawarehouse surgen como solución a los

problemas que plantea el realizar análisis de datos sobre una base de datos OLTP.

El término *Datawarehouse* o almacén de datos representa un conjunto o compendio de datos atendiendo a una temática, no volátil, integrado, de tiempo variable, que es utilizado para aportar mayor información y en menor tiempo al proceso de toma de decisiones en el ámbito de la gerencia empresarial.

Desde un punto de vista más concreto un *Datawarehouse* es una base de datos de carácter corporativo. Esta base de datos se caracteriza por la integración y depuración de información procedente de una o múltiples fuentes de datos, con el fin de procesarla y así ofrecer la posibilidad de analizarla desde un mayor número de perspectivas y a una mayor velocidad de respuesta sobre las posibles consultas que sobre ella se realicen.

La ventaja principal que aporta un Datawarehouse tiene su origen en las estructuras o modelos en los que organiza y almacena la información, los principales son:

- Modelo de tablas en estrella.
- Modelo de tablas en copo de nieve.
- Cubos relacionales.

Debido a esto, la persistencia de la información de un Datawarehouse es homogénea y fiable, aportando a su vez la posibilidad de realizar consultas y de tratar la información de una manera jerárquica.

22.1.2 Características de un Datawarehouse

Los aspectos por los cuales se caracteriza un Datawarehouse son los siguientes:

- **Temático:** orientado sobre la información que resulta relevante para la organización. El proceso de desarrollo del Datawarehouse, se lleva a cabo con el fin de realizar de una manera eficiente las consultas sobre la información que ofrece mayor interés a las

actividades esenciales de la entidad, compras, ventas, producción, etc. En ningún momento se tienen en cuenta otro tipo de procesos como gestión o facturación.

Los datos son organizados en temas, facilitando así su acceso y entendimiento por parte de los usuarios. De esta forma también se ven beneficiadas las consultas, puesto que toda la información referente a una temática se encontrará agrupada.

- **Integrado:** es capaz de incorporar en una sola solución integral datos recopilados de diversos sistemas operacionales y fuentes de información, las cuales pueden ser de carácter interno a la organización como externo. Además permite que esas fuentes de información externa contengan los datos en distintos formatos. La estructura en la cual se realiza la integración de los datos ha de ser consistente, para lo cual se deben eliminar las inconsistencias que existen entre los distintos sistemas operacionales.

- **Histórico:** la variable temporal forma parte de la información implícita contenida en un Datawarehouse. En contraposición con los sistemas operacionales, los cuales siempre son reflejo del estado de la actividad en el momento presente, es decir no reflejan la temporalidad de la información, un Datawarehouse proporciona la capacidad de analizar tendencias que se produzcan a lo largo de una etapa y compararlos con otros anteriores. En definitiva, almacena los datos como si realizara fotos que se corresponden con los distintos momentos o períodos de tiempo.

- **No volátil:** el repositorio de información contenida dentro de un Datawarehouse sólo existe para ser consultada, no para modificar su contenido, lo cual proporciona a esta información carácter permanente. Esto significa que cuando se realiza una actualización del Datawarehouse, únicamente se incorporan los últimos valores

que tomaron las variables contenidas en él, pero no se realiza ningún tipo de acción que altere los valores ya existentes.

22.1.3 Los Metadatos

Otro valor añadido que aporta un Datawarehouse son los metadatos, datos que describen a otros datos. Estos metadatos aportan nueva información sobre los valores existentes en el Datawarehouse, mejorando la descripción de los datos y aportando nuevo conocimiento permitiendo, por ejemplo, conocer la procedencia de la información, su grado de fiabilidad, su periodicidad de refresco o incluso el algoritmo utilizado para calcularla.

Los metadatos, además de ampliar información permiten al Datawarehouse realizar una serie de procesos que simplifican y posibilitan obtener la información de una manera automática desde los sistemas operacionales a los sistemas informacionales.

Los objetivos principales que siguen los metadatos son:

- **Ofrecer soporte al usuario final:** gracias a su propio lenguaje de negocio, facilitan el acceso al Datawarehouse indicando qué información está contenida y el significado que esta aporta. También pueden ser utilizados por otras herramientas para construir informes, consultas, etc.
- **Ofrecer soporte técnico:** suponen un gran punto de apoyo para los técnicos que gestionan el Datawarehouse ya que son de gran ayuda en aspectos de auditoría, en la gestión de la información histórica, en la propia administración del Datawarehouse, etc.

22.1.4 Elementos que componen un Datawarehouse

Los elementos más importantes que conforman un Datawarehouse son:

1. **Fuentes de datos:** las fuentes de datos forman parte del Datawarehouse desde el principio, puesto que son el origen de los datos que contendrá. Las fuentes de datos pueden pertenecer a

diversos ámbitos tanto dentro como fuera de la organización, desde sistemas operacionales propios a fuentes externas.

2. **ETL (Extracción, transformación y carga):** representa la parte del sistema que realiza el proceso de construcción del Datawarehouse. Para ello hace uso de las fuentes de datos desde las cuales extrae la información para luego procesarlas y almacenarlas.

    - *Extracción:* recuperación de la información procedente de las distintas fuentes de datos.
    - *Transformación:* proceso mediante el cual se realizan tareas de filtrado, limpieza, depuración, homogeneización y agrupación de la información.
    - *Carga:* este proceso se encarga de la organización y la actualización de los datos y de los metadatos del Datawarehouse.

3. **Servidor de datos:** componente que realiza las labores de gestión del Datawarehouse. Para llevar a cabo estas tareas suele hacer uso de los recursos ofrecidos por el sistema operativo y por el gestor de base de datos.

    Para el almacenamiento de los datos se pueden diferenciar dos posibilidades en función del tipo de bases de datos y gestor de la misma empleados:

    - Bases de datos relacionales y un sistema gestor de base de datos relacional o SGBDR.
    - Bases de datos multidimensionales, con un gestor de base de datos multidimensional o SGBDM.

    Ha de ofrecer:

    - Servicio de mantenimiento.

- Servicio de distribución para poder exportar los datos hacia otros servidores de bases de datos descentralizadas y a otros sistemas de soporte de decisiones.
- Servicio de seguridad, archivo, backup, recuperación, etc.

4. **Herramientas de acceso:** estas herramientas aportan técnicas para la captura de datos de una manera rápida para poder ser analizados desde distintos puntos de vista. También realizan tareas de transformación de los datos en información útil para el usuario. Este tipo de herramientas se denominan business intelligence tools y se sitúan a nivel conceptual sobre el Datawarehouse. Alguna de estas herramientas son:
   - Consultas SQL.
   - Herramientas MDA, Multidimensional Analysis.
   - Herramientas OLAP, On-Line Analytical Processing.
   - Herramientas ROLAP, Relational On-Line Analytical Processing.
   - Herramientas MOLAP, Multidimensional On-Line Analytical Processing.
   - Herramientas HOLAP, Hybrid On-Line Analytical Processing.
   - Herramientas de Minería de Datos.
5. **Repositorio/Metadatos:** el repositorio ayuda a los usuarios a saber qué es lo que hay almacenado en el Datawarehouse y como pueden acceder a lo que quieren. Además realiza diversas funcionalidades como:
   - Catalogar y describir la información disponible.
   - Especificar el propósito de la información.
   - Reflejar las relaciones de los datos.

- Indicar el propietario de la información.
- Relacionar las estructuras técnicas de datos con la información de negocio
- Especificar las relaciones entre los datos operacionales y las reglas de transformación.
- Limitar la validez de la información.

22.1.5 Ventajas principales de un Datawarehouse

Las aportaciones más importantes que ofrece un Datawarehouse son las siguientes:

- Facilita la implantación de sistemas de gestión integral de la relación con el cliente desde el núcleo de una organización.
- Posibilita la utilización de técnicas de modelización y análisis estadístico para la búsqueda de relaciones ocultas entre los datos almacenados, lo cual ofrece un valor añadido al sistema de gestión de la información.
- Ofrece una herramienta de apoyo a la toma de decisiones en cualquier área funcional, gracias la información integrada y global que proporciona.
- Aporta la capacidad de aprendizaje sobre los datos pasados para la predicción de posibles situaciones futuras.

22.2 DATA MARTS

La solución a los problemas relacionados con el análisis de datos sobre una base de datos OLTP son solucionados con la creación de los datawarehouse (base de datos independiente orientada a consultas). Sin embargo, cuando los datawarehouse aumentan su tamaño se vuelven cada vez más complejos lo que provoca un decrecimiento en el rendimiento de las consultas, dejando de ser útil el modelo centralizado. Como respuesta a esta bajada de rendimiento surgen los Data Marts, que son almacenes de datos que se especializan por áreas o temáticas como pueden ser ventas o compras.

Los Data Marts suelen recibir la información desde el datawarehouse, almacén de datos centralizado, y pueden estar ubicados en máquinas distintas, en otras BBDD, redes, etc. También pueden integrar la información desde distintas fuentes. Según estos dos modelos de extracción de la información, existen dos tipos de Data Mart:

- *Data Mart dependiente*, cuyos datos vienen proporcionados desde un datawarehouse. En la ilustración 9 se puede observar el funcionamiento de un Data Mart con un datawarehouse.
- *Data Mart independiente,* donde los datos son extraídos de diversas fuentes de información de los sistemas operacionales.

Un Data Mart representa una pequeña porción de un datawarehouse, con lo que soporta un número de usuarios más reducido, con lo cual se pueden optimizar para realizar el proceso de recuperación de la información de una manera más rápida.

Un Data Mart en realidad es una base de datos específica de un departamento o sección, dedicada únicamente a los datos relevantes que se producen en ese ámbito. Debido a esta especialización disponen de una estructura óptima de datos adaptada al análisis de la información desde todas las perspectivas que afecten a los procesos de ese ámbito.

Para la creación de un Data Mart es necesario encontrar la estructura

montada sobre un sistema OLTP como un datawarehouse, o por el contrario, puede sostenerse sobre un sistema OLAP. Por ello se pueden establecer dos tipos de Data Marts:

- **Data Mart OLTP:** son Data Marts basados en un datawarehouse, pero es habitual que incorporen mejoras para ofrecer un mayor rendimiento adaptando las necesidades de cada área al Data Mart.

  En este tipo de Data Marts las estructuras más comunes son:

  - *Tablas report*, que son tablas de hechos reducidas que agregan las dimensiones oportunas.
  - *Vistas materializadas*, que se construyen con la misma estructura que las tablas report para explotar la reescritura de las consultas. Este tipo de estructura es dependiente del SGBD.

- **Data Mart OLAP:** se basan en cubos OLAP que se generan en función de los requisitos de cada área, agregando las dimensiones y los indicadores necesarios de cada cubo relacional. Los modos de creación, explotación y mantenimiento de este tipo de estructura es muy dependiente de la herramienta que se utilice para su manejo.

Los Data Marts gracias a este tipo de estructuras óptimas para el análisis ofrecen una serie de ventajas como:

- Elevada rapidez de consulta de la información.
- Reducido conjunto de datos.
- La información se valida directamente.
- Realizar fácilmente históricos de los datos.
- Posibilidad de Consultas SQL y MDX sencillas.

22.3

22.4 ANÁLISIS MULTIDIMENSIONAL Y ARQUITECTURAS OLAP.

La primera aparición del término OLAP (On-Line Analytical Processing) fue publicada en 1993 por Edgar F. Codd. Sin embargo, en 1970 ya existían productos que realizaban consultas OLAP. Codd definió OLAP como un tipo de procesamiento de datos caracterizado por permitir el análisis multidimensional.

22.4.1 Análisis Multidimensional y OLAP

La multidimensionalidad desde el punto de vista de un proceso analítico en línea consiste en transformar los datos procedentes desde varias fuentes, tablas de una base de datos, archivos,... y convertirlos en una estructura donde estos estén agrupados en dimensiones separadas y heterogéneas. Estas estructuras se denominan *cubos*.

Las dimensiones constituyen las perspectivas de alto nivel de los datos que representan la información más importante de un negocio. Estas dimensiones en una solución OLAP tienden a ser invariantes.

El análisis multidimensional se fundamenta en modelar la información en dimensiones, hechos y medidas.

- **Medidas:** es un tipo de dato que contiene información que utilizan los usuarios en sus consultas con las que son capaces de medir el grado de rendimiento de un proceso.
- **Dimensiones:** entidades o colección de entidades que se encuentran relacionadas y que son usadas para determinar o identificar el contexto de las medidas.

  El tipo y el número de dimensiones para cada una de las medidas del modelo es un proceso que ha de realizarse cuidadosamente, puesto que al definir las dimensiones, el añadir, eliminar o cambiar propiedades particulares de las dimensiones candidatas varía el contexto y también el significado de la medida candidata.

Una dimensión tiene componentes denominados *miembros* (dimensión Tiempo, miembro trimestre) y entre los miembros pueden existir jerarquías (un mes puede considerarse dentro de un trimestre).

Las dimensiones contienen:

- Entidades de dimensión.
- Atributos de dimensión.
- Jerarquías de dimensión.
- Niveles de agregación.

Para referenciar a las dimensiones se utilizan las *llaves de dimensión.*

- **Hechos:** identifican la existencia de valores específicos de una o más medidas para una combinación concreta de dimensiones. Mediante un hecho se puede representar desde un objeto de negocio hasta una transacción e incluso un evento utilizado por los usuarios.

  Los hechos contienen:

  - Un identificador para cada hecho.
  - Llaves de dimensión, que lo enlazan con las dimensiones.
  - Medidas.
  - Tipos de atributos normalmente derivados de otros datos del modelo.

Una característica fundamental y muy importante de este modelo es que tiene la capacidad de representarse de manera vectorial. Los hechos se sitúan de manera lógica en una celda, la cual se encuentra en la intersección de ciertas coordenadas según el modelo (x, y, z,...), donde

además cada una de las coordenadas que se encuentran en las celdas representan una dimensión.

La utilización de la correspondencia entre los elementos del modelo, es decir, los hechos y las coordenadas, y los de la base de datos, la tabla de hechos y dimensiones, es fundamental para poder llevar a cabo el análisis multidimensional en una base de datos. En una base de datos se pueden implementar los hechos y las dimensiones en una tabla y debido a esto es posible utilizar el lenguaje SQL para la definición de un modelo multidimensional en una base de datos relacional. A pesar de esto, fue necesario realizar una serie de extensiones del modelo relacional para poder dar soporte a las funcionalidades y necesidades propias del análisis multidimensional. Estas funcionalidades son:

1. Declaración de Dimensiones y Jerarquías. El modelo relacional no incorporaba ni trataba con anterioridad estos conceptos.
2. Acceso más rápido a los datos. Para añadir esta mejora se utilizaron métodos de generación de índices para datos espaciales desde el punto de vista multidimensional.
3. Cálculo de valores previamente agrupados para la optimización de consultas.
4. Definición de operaciones de navegación en las dimensiones y de agrupación de medidas como:
    - Slice-and-dice:
    - Drill-down
    - Roll-up
    - Pivot
    - Drill-across
    - Drill-through

Partiendo de las primeras propuestas, el modelo multidimensional no precisa de un almacenaje previo en una base de datos multidimensional, sino que propone que el acceso a la información puede hacerse directamente a múltiples fuentes, bases de datos (ya sean relacionales o

multidimensionales), hojas de cálculo, archivos e incluso permite que los datos puedan proceder directamente de los usuarios finales.

A pesar de estas primeras ideas, se ha determinado a través de la experiencia de que el análisis OLAP tiene un mejor desempeño si la fuente de datos es única y aún mejor si esa fuente de información es a su vez una base de datos multidimensional, como por ejemplo un Datawarehouse.

22.4.2 Sistemas OLAP

Los sistemas OLAP son un conjunto de métodos que permiten consultar la información contenida en los datos de diversas maneras. Esta versatilidad y multiplicidad de opciones de visualización viene producida por la clasificación de los datos en diferentes dimensiones que pueden ser visualizadas unas con otras combinándolas para obtener diferentes análisis de la información.

La información de un modelo OLAP es vista como un cubo, los cuales son determinados pro categorías descriptivas o dimensiones y valores cuantitativos, las medidas. Este modelo de datos multidimensional simplifica mucho las tareas que el usuario puede realizar sobre los datos, consultas complejas, filtrar en subconjuntos,...

22.4.3 OLAP como sistema de información ejecutiva

Como clasificación de los sistemas OLAP se pueden considerar dentro del grupo de aplicaciones o sistemas de información para ejecutivos (EIS), que se emplean para proporcionar al nivel estratégico información que resulte relevante a la hora de tomar decisiones.

Si comparamos los sistemas OLAP con el resto de EIS podemos afirmar que las herramientas OLAP ofrecen una opción mucho más general, es decir son más genéricas:

- Funcionan sobre un sistema de información como los datawarehouse.
- Permiten la creación de agregaciones y combinaciones de los datos de muchas maneras posibilitando realizar análisis más estratégicos de datos.

- Posee operadores para realizar tareas especificas (Drill, Roll, Slice-and-Dice,...).
- El resultado puede ser expresado de manera matricial o hibrida.

22.4.4 Operadores OLAP

Estas herramientas de las soluciones OLAP permiten al usuario tener una visión multidimensional de la información para cada una de las actividades de análisis. Con los operadores se realizan consultas simplemente seleccionando atributos del esquema multidimensional sin tener que tener conocimiento de es la estructura interna en la que se almacenan los datos, puesto que la propia herramienta OLAP se encarga de generar la consulta y enviarla al sistema de gestión de consultas.

Una consulta consiste en la obtención de medidas sobre los hechos parametrizadas por los atributos de las dimensiones y limitadas por las condiciones impuestas sobre las dimensiones. Las herramientas OLAP ofrecen una serie de nuevos operadores que refinan esas consultas. Los operadores son los ya mencionados anteriormente en el punto (Análisis multidimensional y OLAP).

- **Drill o disgregación:** posibilita introducir un nuevo criterio de agrupación en el análisis, disgregando los grupos actuales. Actúa sobre el operador original *informa* con lo cual no es necesario crear o realizar un nuevo informe. Existen varias variantes:
    - *Drill-down,* permite visualizar los datos del nivel inferior de la dimensión actual dentro de una jerarquía definida. Muestra los datos detallados que en conjunto determinan el valor.
    - *Drill-across,* visualiza la información contenida en otro modelo multidimensional, sin detallar ni consolidar la información cambia el modelo multidimensional que se está consultando. Para realizar esta operación ambos modelos han de tener una dimensión común.
    - *Drill-through*, similar a drill-dow consulta la información del nivel inferior a la dimensión actual. Sin embargo drill-throug

navega por fuera del modelo multidimensional estableciendo un enlace entre este y el sistema fuente, sobre el cual consulta los datos del nivel detallado directamente. Para poder utilizar este operador se debe establecer acceso al sistema fuente desde el sistema OLAP.

- **Roll o agregación:** permite que se elimine un criterio de agrupación en el análisis, agregando los grupos actuales. Actúa sobre el informe ya creado y no es preciso realizar uno nuevo. Variantes:
  - *Roll-up*, también conocido como drill-up se encarga de pasar al nivel superior de la jerarquía de la dimensión actual. Para ello consolida los datos del nivel actual y muestra el valor consolidado que corresponde con el nivel superior de la dimensión.
  - *Roll-across*, funciona de una manera parecida al *Roll-up* salvo que no se realiza sobre jerarquías de una dimensión, sino que elimina un criterio de análisis eliminando de la consulta una dimensión.

- **Slice-and-dice**: permite seleccionar y proyectar datos en el informe. Selecciona la información de un miembro de una dimensión, se trabaja con un subconjunto de los datos para un valor determinado de un nivel en una dimensión. Con frecuencia este operador es empleado sobre un eje temporal para poder analizar tendencias y encontrar patrones.

- **Pivot:** con este operador se permite cambiar la orientación de las dimensiones en un informe. Selecciona el orden de visualización

de las dimensiones con el fin de analizar los datos desde distintas perspectivas.

22.5 ROLAP, MOLAP Y HOLAP

22.5.1 ROLAP

Es un tipo de organización de la información a nivel físico que se implementa sobre tecnología relacional, pero incorpora de algunas facilidades que incrementan su rendimiento.

ROLAP (Relational On-Line Analytic Processing) posee las virtudes de un sistema gestor de bases de datos relacional sobre el cual se le incorporan una serie de herramientas y extensiones para poder ser utilizado como un datawarehouse o almacén de datos. Las principales características de los sistemas ROLAP son:

- Almacena los datos en una base de datos relacional.
- Utilización de índices de mapas de bits.
- Utilización de índices de join.
- Técnicas de particionamiento de datos.
- Optimizadores de consultas.
- Extensiones de SQL (drill, roll, etc).

22.5.1.1 Ventajas de los sistemas ROLAP

- Utilización completa de la integridad y seguridad que ofrecen las bases de datos relacionales.
- Es escalable para volúmenes grandes.
- Los datos pueden ser compartidos con otras aplicaciones que utilicen el lenguaje SQL.
- Datos y estructuras más dinámicas.

22.5.1.2 Desventajas de los sistema ROLAP

- Las consultas resultan más lentas.
- Su construcción suele resultar costosa.
- Los índices no se mantienen de manera automática.
- Los cálculos se encuentran limitados por las funciones de la base de datos.

22.5.2 MOLAP

La función principal de los sistemas MOLAP (Multidimensional On-Line Analytic Processing) es la de almacenar físicamente los datos en estas estructuras especificas de tipo multidimensional haciendo coincidir la representación interna de los datos con la representación que las capas superiores dan a la información.

Poseen estructuras específicas para el almacenamiento de la información, aportando también técnicas para la compactación de los datos lo cual mejora el rendimiento del almacén de datos.

Las características más importantes de los sistemas MOLAP son:

- Incorporan tecnología optimizada para la realización de las consultas y del análisis, la cual está fundamentada en el modelo multidimensional.
- Tiene un motor especializado.
- Construye los datos y los almacena en estructuras multidimensionales.

22.5.2.1 Ventajas de los sistemas ROLAP

- Mayor rendimiento a la hora de ejecutar las consultas.
- Poco tiempo de cálculos realizados en el momento.
- Puede realizar la escritura de manera directa en la base de datos.
- Ofrece la posibilidad de implementar cálculos más sofisticados.

22.5.2.2 Desventajas de los sistema ROLAP

- El tamaño viene limitado por la arquitectura del cubo.
- Sólo es capaz de gestionar los datos si estos se encuentran almacenado en un cubo.
- Procesos de mantenimiento y de copias de seguridad limitados.
- No explota la capacidad de paralelismo que ofrecen las bases de datos.
- Introduce redundancia de datos.

22.5.3 HOLAP

Este tipo de sistemas HOLAP (Hybrid On-Line Analytic Processing) están considerados como sistemas híbridos entre los ROLAP y los MOLAP, puesto que incorpora características de ambos. Para ello utiliza un motor relacional para almacenar parte de los datos en una base de datos de tipo relacional y utiliza una base de datos multidimensional para otra parte de la información.

22.5.3.1 Particionamiento vertical

En este modo, HOLAP mejora la velocidad de las consultas almacenando las agregaciones en un sistema MOLAP, mientras que para optimizar el tiempo, los datos son detallados en un sistema ROLAP.

22.5.3.2 Particionamiento horizontal

Un sistema HOLAP en modo de particionamiento horizontal almacena parte de los datos, normalmente los más recientes particionados por una de las dimensiones (dimensión tiempo por ejemplo) en modo MOLAP, con lo que consigue un aumento en la velocidad de respuesta de las consultas. Por otra parte mantiene en un sistema ROLAP los datos más antiguos. También este modo permite que los cubos se almacenen unos en sistemas MOLAP y otros en sistemas ROLAP.

22.6 MINERÍA DE DATOS

El concepto de Data Mining o Minería de datos viene determinado por el método de extracción de la información contenida en los datos. La minería de datos obtiene información contenida en los datos, pero de una forma indirecta, puesto que esta se encuentra implícita en los datos y no se puede acceder a ella directamente.

La minería de datos *prepara, sondea* y *explora* el conjunto de datos con el fin de conseguir información que de algún modo se encuentra oculta. Esta ocultación de la información se debe a que normalmente para un experto lo que resulta relevante es la información contenida en las relaciones, fluctuaciones y dependencias de los datos, no los datos en sí. Esta información es por lo general desconocida, lo cual ofrece un valor muy importante, puesto que puede resultar de gran utilidad a los procesos de una organización.

Está formada por un conjunto de técnicas dirigidas a la obtención del conocimiento procesable oculto en las bases de datos. Estas técnicas fundamentan su base en la inteligencia artificial y en el análisis estadístico para generar modelos con el fin de poder abordar la solución a problemas de predicción, clasificación y segmentación.

La minería de datos es un proceso que invierte la dinámica del método científico puesto que en este, primero se formulan las hipótesis y luego se desarrolla un experimento para la obtención de los datos que las confirmen o refuten, obteniendo así nuevo conocimiento. En la minería de datos se recaba una colección de datos con la intención de que de estos surjan hipótesis. Se espera que de los propios datos describan o indiquen como son para poder validar las hipótesis aparecidas con los datos mismos. Es por este motivo que la minería de datos ha de realizarse con un enfoque exploratorio y no confirmador

22.6.1 Características principales

Las principales características que determinan un sistema de minería de datos son:

- Trabaja con la información contenida en lo más oculto de las bases de datos o almacenes de datos analizando información almacenada durante años.
- Suelen ser soluciones con una arquitectura cliente-servidor.
- Poseen gran variedad de herramientas para la extracción de la información.
- Las herramientas son fácilmente combinables entre sí.
- Son los usuarios finales los que hacen uso de las herramientas para indagar en el conjunto de datos y obtener respuestas rápidas.
- Es habitual hacer uso de un procesamiento paralelo que acelere el proceso debido a la existencia de una gran cantidad de datos.
- Produce cinco tipos de información:
    - Asociaciones
    - Secuencias
    - Clasificaciones
    - Agrupamientos
    - Pronósticos.

22.6.2 Técnicas principales

Las técnicas más importantes que se utilizan para llevar a cabo este proceso son:

- *Redes Neuronales:* es una técnica que proviene de la inteligencia artificial para la detección de categorías comunes en los datos ya que es capaz de detectar y aprender complejos patrones y características de los datos.

  Un punto fuerte de las redes neuronales es que son capaces de trabajar con conjuntos incompletos de datos e incluso con algunos

paradójicos que en función del problema pueden ser ventajosos o resultar un inconveniente.

- *Árboles de Decisión:* esta técnica se representa en forma de árbol siendo cada nodo una decisión, los cuales generan una serie de reglas mediante las cuales clasifican los datos.

  Son sencillos de utilizar, admiten tanto atributos discretos como continuos y tratan bien tanto los atributos no significativos como los valores faltantes. Además son fácilmente interpretables.

- *Algoritmos Genéticos:* son técnicas que imitan la evolución de las especies mediante la generación de mutaciones, la reproducción y selección. Aportan herramientas para integrar en la construcción y entrenamiento de otras estructuras como por ejemplo las redes neuronales. Están basados en el principio de supervivencia de los más aptos.

- *Clustering (Agrupamiento):* técnica que agrupa datos dentro de una serie de clases, que pueden ser predefinidas o no, siguiendo los criterios de distancia o similitud de modo que los datos contenidos dentro de una clase son similares entre sí y distintos con los contenidos en las otras clases. Es un método muy flexible y es fácilmente combinable con otras técnicas de minería de datos.

- *Aprendizaje Automático:* técnica procedente de la inteligencia artificial en la que se trata de inferir conocimiento partiendo del resultado obtenido mediante alguna de las otras técnicas anteriormente mencionadas.

22.6.3 Algoritmos empleados

Los algoritmos utilizados en la minería de datos se pueden clasificar en:

- **Supervisados**
  - Predicen el valor de un atributo de un conjunto de datos una vez conocidos otros atributos.

- Partiendo de los datos cuyos atributos son conocidos, se inducen nuevas relaciones entre atributos.
- Constan de dos fases:
  - *Entrenamiento*, en la cual se construye un modelo usando un subconjunto de datos conocidos.
  - *Prueba*, se prueba el modelo con el resto de los datos.

- **No Supervisados**
  - Se utilizan cuando una aplicación no se encuentra lo suficientemente madura o no tiene las capacidades necesarias para realizar una solución predictiva.
  - Descubren patrones y tendencias en los datos.
  - Con el descubrimiento de la información se pueden llevar a cabo acciones que reporten en un beneficio.

22.7

22.8 GENERACIÓN DE INFORMES A LA DIRECCIÓN

Los aplicaciones para la generación de informes a la dirección o *Sistemas de Información para Ejecutivos (EIS)*, son herramientas software que se basan en sistemas de apoyo a las decisiones (DSS Decision Support System) proporcionando a la gerencia de una organización acceso fácil y sencillo a la información que resulta clave para el éxito de su compañía, ya sea interna o externa.

El objetivo principal de este tipo de aplicaciones es poner a disposición de los ejecutivos una serie de herramientas que muestren el abanico completo del estado de los indicadores de negocio que le interesan en tiempo real, ofreciendo a su vez la capacidad de un análisis detallado de aquellos que no se estén consiguiendo las expectativas o las planificaciones establecidas a priori.

Este tipo de sistemas se pueden definir como soluciones para mostrar informes y listados (query & reporting) de los distintas áreas de negocio de una forma consolidada facilitando una monitorización completa y real de una organización.

Ofrecen además un acceso rápido y efectivo a la información compartida, para lo cual hacen uso de interfaces gráficas muy visuales e intuitivas. Incorporan también incluyen alertas e informes basados en excepción, así como históricos y análisis de tendencias.

Mediante estos sistemas el seguimiento del comportamiento de una organización o de un área de negocio se hace de una manera fácil y comparable a través del tiempo.

Dentro de estos sistemas, lo más común es encontrar los términos de Informes (Reports), Cuadro de Mando (Dashboard) y Cuadro de Mando Integral (Balanced ScoreCard).

Informes

Los informes son la herramienta más común de transmitir toda la información obtenida de un sistema de business intelligence. Un informe se puede describir como un documento, o conjunto de documentos, que

contiene datos utilizados para su estudio y análisis por parte de la dirección. Pueden estar compuestos desde una simple tabla de datos o un vista más compleja con datos agregados, con datos transformados mediante la aplicación de fórmulas o con sistemas de navegación interactiva a través de los datos (habitualmente ampliando la vista de cada fila en la tabla).

La característica principal de un informe es que no ofrece al lector del mismo ningún tipo de conclusión o visión predefinida de los datos. Aunque un informe incluya datos analíticos, datos agregados, datos calculados o algún gráfico es el propio lector el que debe extraer conclusiones o determinar las próximas acciones en base a los datos presentados en el informe.

Dashboard

Un cuadro de mando es una interfaz de carácter visual que ofrece en cada momento diferentes vistas o perspectivas de las diferentes métricas o indicadores (también denominados KPI Key Performance Indicators)que se hayan considerado como relevantes para un proceso de negocio o los objetivos de una empresa. Un KPI es un indicador de la ejecución y el rendimiento de una tarea o actividad diaria que se considera fundamental desde el punto de vista de la dirección para su seguimiento. La idea que subyace a un KPI es que no es una métrica simple del negocio, sino que está diseñado de forma que describe y alerta sobre distintas circunstancias permitiendo detectar e intervenir en aquellas situaciones que así lo requieran.

Un dashboard presenta tres características diferenciadoras:

Muestra los datos de forma gráfica. Esto proporciona una visión mucha más centrada en los indicadores de rendimiento, en las posibles comparaciones entre datos y aquellos datos que sean una excepción o que identifiquen una anomalía.

Sólo muestran aquellos datos que son necesarios para un determinado objetivo empresarial.

Además, incluye conclusiones predefinidas que son relevantes para los objetivos del cuadro de mando y que ayudan al lector a realizar su propio análisis.

Cuadro de Mando Integral

Un cuadro de mando integral (Balanced Scorecard) es una representación visual de la estrategia de la empresa. El cuadro de mando integral permite de una forma sencilla presentar los indicadores o métricas críticas para el negocio y contrastarlas con la estrategia de negocio que se pretende para la organización.

El cuadro de mando integral ha de mostrarse de una forma visual y ser la referencia a cualquier persona de la organización para ver:

El rendimiento de las iniciativas específicas a distintas unidades de negocio o desde un punto de vista global a toda la compañía.

Los objetivos individuales referenciados al contexto global de la compañía mediante una representación visual.

Los cuadros de mando integral se diseñan siempre con el fin de aumentar la productividad de toda la organización, porque indica en tiempo real como se está comportando un empleado, un equipo, un departamento o toda la empresa, de acuerdo a los objetivos definidos en el plan estratégico. Esto lo convierte en un sistema de gestión estratégica de la empresa que permite:

Formular estrategias consistentes y que estas sean transparentes a toda la organización.

Comunicar las estrategias definidas por la dirección a través de toda la organización.

Coordinar los objetivos de las diversas unidades organizacionales (equipos, departamentos, secciones, etc.) de acuerdo al mismo plan estratégico.

Conectar los objetivos de cada unidad organizacional con la planificación financiera y presupuestaria de la organización.

Medir de un modo sistemático la realización, proponiendo acciones correctivas oportunas por parte de la dirección o por cada uno de las unidades organizacionales implicadas.

22.9 BIBLIOGRAFÍA

- *“Building the Data Warehouse”*. Inmon, W.H.
- "*Sistemas de Información Para la Toma de Decisiones*". Cohen K. Daniel, Ed. Mc Graw Hill, 1996.
- *“OLAP Solutions: Building Multidimensional Information Systems”.* Erik Thomsen. Ed. Wiley, 2002. ISBN: 04 714 0030 0
- *“State of the Art: Data Mining”*. S. R. Hedberg, K. Watterson y C. D. Krivda. Publicado en BYTE (10-95)
- *“MOLAP, ROLAP, Overlap”*. Jeff Stamen. Publicado en BYTE (8-96)
- State of the Art: Data Warehouses. Autor: J. L. Weldon, A. Simon y M. Hurwicz. Publicado en BYTE (1-97)
- "*Introducción a la Minería de Datos”.* José Hernández Orallo, M.José Ramírez Quintana, César Ferri Ramírez. Ed. Pearson, 2004. ISBN: 84 205 4091 9.

Autor: Francisco Javier Rodriguez Martínez

Subdirector de Sistemas Escola Superior Enxeñaría Informática Ourense

Colegiado del CPEIG