RESUMEN TEMA 72: GESTIÓN DE LOS DATOS CORPORATIVOS. ALMACÉN DE DATOS (DATA-WAREHOUSE). ARQUITECTURA OLAP. MINERÍA DE DATOS.

En el presente resumen, se pretende condensar las principales ideas recogidas en la documentación de la carpeta de contenidos del tema 72. Se recomienda la lectura de la documentación.

DEFINICIONES

Data Warehouse (DW) o almacenes de datos: base de datos corporativa caracterizada por la integración y depuración de información procedente de múltiples fuentes de datos tanto internas como externas a la organización. Su fin es el de procesar la información para poder analizarla.

Data Mart: almacenes de datos especializados por áreas o temas.

Minería de datos: se encarga de la extracción de patrones o de información implícita u oculta, contenida en los datos.

DATA WAREHOUSE:

Los almacenes de datos se caracterizan por ser:

- Orientados a temas o Temático: el almacén se orienta a información relevante para la organización.
- Variantes en el tiempo o históricos: toda la información se almacena con una referencia temporal.
- No volátiles: la información en el DW sólo puede ser consultada, no podrá ser ni borrada ni modificada.
- Integrados: incluyen en una solución integral, información procedente de diversas fuentes.

APROVISIONAMIENTO DE UN DATA WAREHOUSE

Los DW han de recopilar la información de diferentes fuentes para ello hace uso de herramientas ETL.

Las **herramientas ETL** se encargan de:

- Extracción de la información: recuperación de la información de diferentes fuentes de datos.
- Transformación: conjunto de tareas sobre los datos como depuración, filtrado, convención de nombres, homogeneización.
- Carga en el almacén: proceso de organización y actualización de datos en el DW.

En todo este proceso son importantes los **METADATOS**: datos que describen a datos y que son empleados por el DW para simplificar y posibilitar la obtención de información.

Cuando la información almacenada en el DW es voluminosa y variada se puede aprovisionar el DW desde dos posibles enfoques:

- Bottom Up: Se aprovisionan almacenes temáticos (DATAMARTS) y del conjunto de ellos se crea el DW.
- Top Down: Se aprovisiona el DW y si se requiere especialización temática se aprovisionan a partir del DW los DATAMARTS. Este enfoque es poco práctico.

EXPLOTACIÓN DEL DATA WAREHOUSE

Estructuras multidimensionales: los DW a diferencia de los sistemas transaccionales (OLTP) se basan en el uso de estructuras multidimensionales que permiten la manipulación y visualización de los datos de manera más eficiente. Son una variante de los modelos relacionales tradicionales y se componen de:

- **Tablas de hechos**: donde se almacena la información propiamente dicha.
- **Tablas de dimensiones**: perspectivas de alto nivel acerca de los datos.

Este modelo permite representarse de forma vectorial. En cada eje se representa las dimensiones requeridas para las búsquedas. Así surge la idea de los **Cubos dimensionales**.

Las operaciones habituales que se pueden hacer en este modelo son:

- Slice-and-dice: permite obtener los datos seleccionando un valor fijo de una dimensión.

- Drill-down: permite ver datos de nivel inferior (aumenta nivel de detalle).
- Roll-up: permite ver datos con mayor nivel de agregación (disminuye nivel de detalle).
- Pivot: permite cambiar los ejes.
- Drill-across: permite ver datos de otro modelo multidimensional.
- Drill-through: similar a drill down pero navegando por fuera del modelo.

Para explotar la información de esta manera se emplean **sistemas OLAP** (Online Analytical Processing) que son un conjunto de métodos que permiten consultar la información contenida en los datos de diversas maneras basándose en el modelo de los Cubos dimensionales.

Tipos de sistemas OLAP:

- **ROLAP**: a nivel físico la información se almacena de forma relacional pero para su explotación se construyen cubos dinámicamente.
- **MOLAP**: la información se almacena directamente de forma multidimensional (cubos) .
- **HOLAP**: mezcla de los dos anteriores.

MINERÍA DE DATOS

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.

Principales características

- Trabaja con la información oculta.
- Suelen ser soluciones con una arquitectura cliente-servidor.
- Poseen gran variedad de herramientas para la extracción de la información.
- 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.

Se basa en el uso de diferentes técnicas:

- Redes neuronales.
- Árboles de decisión.
- Algoritmos genéticos
- Clustering o agrupamiento
- Aprendizaje automático

Estas técnicas se basa en el uso de algoritmos los cuales se pueden clasificar en:

- Supervisados: cuentan con una fase de entrenamiento para construir el modelo.
- No supervisados.: no cuentan con esa fase de entrenamiento.