Tema 3: Análisis y transformación de variables

Minería de Datos

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades

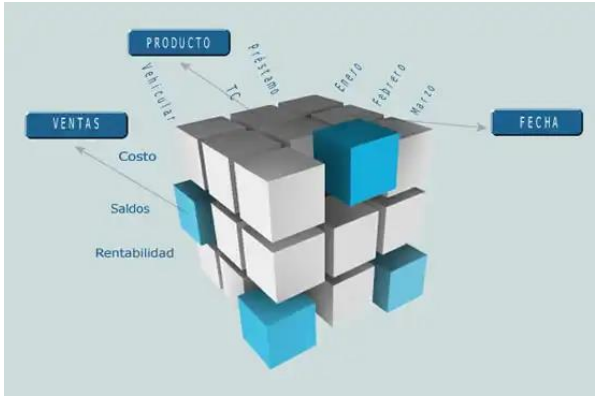
Análisis de datos con cubos y modelos de minería

Introducción

- En el mundo de las **soluciones para Business Intelligence**, una de las **herramientas** más utilizadas por las empresas son las aplicaciones OLAP, ya que las misma han sido creadas en función a bases de datos multidimensionales, que **permiten procesar grandes volúmenes de información,** en campos bien definidos, y con un acceso inmediato a los datos para su consulta y posterior análisis.

OLAP: una base de datos multidimensional

- Utilizan un tipo de base de datos que posee la peculiaridad de ser multidimensional: **cubo OLAP**
- **El Cubo OLAP,** que nació de la mano de Edgar F. Codd, de la compañía EF Codd & Associates, sólo se utilizaban bases de datos relacionales para el proceso de la información, con sistemas tales como el **ROLAP**.

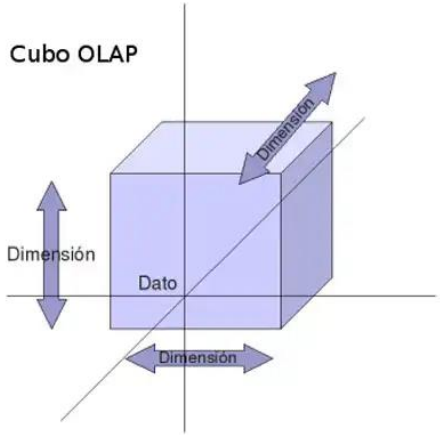


OLAP: una base de datos multidimensional

- Se encuentra ordenada jerárquicamente para poder realizar un análisis rápido de los datos.
- Permite el procesamiento de importantes volumenes de información
- Cada una de las dimensiones que posee la base de datos incorpora un campo determinado para un tipo de dato específico, que luego podrá ser comparado con la información contenida en el resto de dimensiones, para hacer posible la evaluación y posteriores informes de la información realmente relevante para una compañía.
- Una base de datos multidimensional puede contener varios cubos o vectores que extenderán las posibilidades del sistema OLAP.
- El gran fallo reside en la imposibilidad de realizar cambios en su estructura.

OLAP: una base de datos multidimensional

- Si se modifica la estructura de datos, se debe rediseñar el cubo OLAP.
- No se puede reutilizar la estructura en la que se ha trabajado previamente.



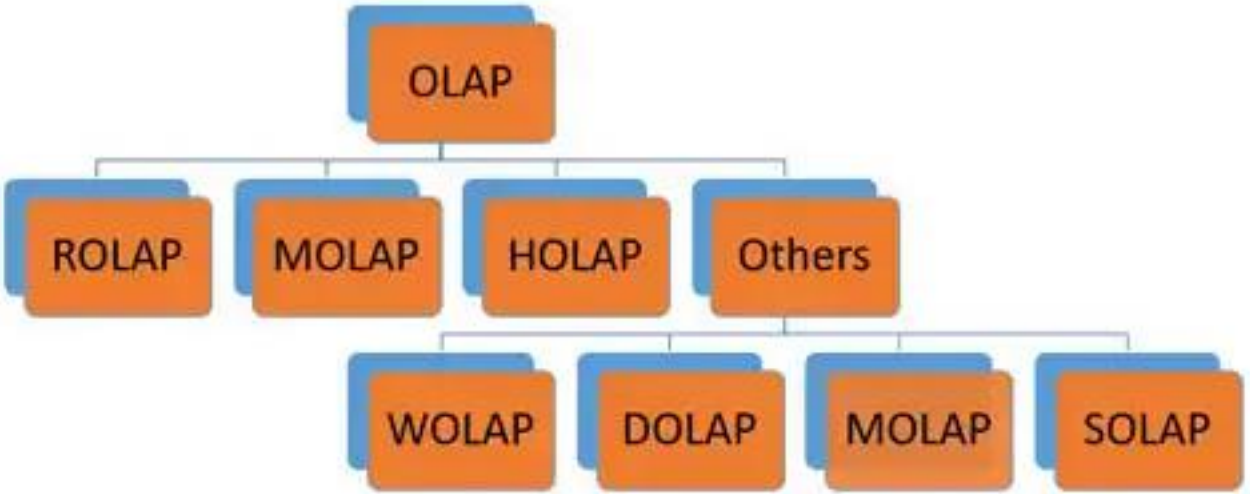
¿Dónde se utiliza OLAP?

- **La herramienta OLAP ha sido ampliamente utilizada durante años en diversos sectores empresariales**, tales como el marketing, ventas, gerencia y demás, permitiendo realizar informes de negocios confiables, que mejoran la competitividad de las organizaciones, tanto a nivel interno como externo.
- Tengamos en cuenta que **una base de datos multidimensional permite disponer de una importante plataforma para contener la información emitida por las distintas áreas de la empresa,** ya que su característica principal reside en que cada dimensión que posee la base de datos tiene su propio campo, y además incluye otro campo por cada hecho, ofreciendo la posibilidad de obtener un registro completo y perfectamente organizado.

¿Dónde se utiliza OLAP?

- **El Cubo OLAP está compuesto por campos numéricos, a los cuales se los denominada medidas**, las que se encuentran clasificadas en tres dimensiones, a diferencia de las conocidas hojas de cálculo, que sólo disponen de dos dimensiones.
- **Para que este sistema funcione, todo el esquema de tablas que son parte del Cubo OLAP se halla sometido a una base de datos relacional**, que permite utilizar información de diferentes sectores y épocas, relacionarlos, para luego poder efectuar un análisis completo de la situación.

Tipos de sistemas OLAP



Tipos de sistemas OLAP

- **ROLAP:** procesamiento analítico online
  - Cualquier usuario puede acceder fácilmente a la información que contiene la base de datos relacional.
  - Logra un tiempo de carga menor que otros métodos basados en OLAP.
- **MOLAP**: multidimensional Online Analytical Processing
  - Almacena todos los datos capturados en una base de datos multidimensional, que ha sido optimizada para ofrecer rapidez de acceso para las cargas y consultas de información, la cual se halla contenida en el denominado Cubo OLAP.
- **HOLAP**: sistema OLAP hibrido (Hybrid Online Analytical Process), que combina ROLAP y MOLAP.
  - Ejemplos: Microsoft Analysis Services, MicroStrategy y SAP AG BI Accelerator.

ALMACENES DE DATOS (DATA WAREHOUSE Y DATA MARTS)

- *Data warehousing*, principalmente para procesamiento de grandes conjuntos de datos.
- *Data marts*, subconjuntos o conjuntos especializados de *data warehouse*.
- Herramientas de ETL, BI, *reporting*, *query*, visualización y analítica.
- Almacenes de datos *columnares*, distribución y compresión por clave.

ALMACENES DE DATOS (DATA WAREHOUSE Y DATA MARTS)

- Las organizaciones tienen grandes inversiones en *data warehouses* y *data marts* que se pueden basar en:
  - Bases de datos relacionales (tales como Oracle Database 11g y 12, IBM DB2 o SQL Server de Microsoft).
  - Bases de datos *columnares* (tales como SAP Sybase IQ, HP Vertica y Par Accel).
  - *Appliances* (máquinas *hardware/software*) de *data warehousing* (tales como Oracle Exadata, Oracle Exalytics *in-memory* Machine, IBM Netezza, HP Vertica, y EMC Greenplum. InfoBright (basada en MySQL), InfiniDB (*open source*), Teradata, etcétera).

ALMACENES DE DATOS (DATA WAREHOUSE Y DATA MARTS)

- Las empresas tienen la opción de seleccionar múltiples categorías de bases de datos:
  - relacionales,
  - NoSQL,
  - *in-memory* (“en memoria”),
  - bases de datos heredadas (*legacy*)

ALMACENES DE DATOS (DATA WAREHOUSE Y DATA MARTS)

- A estas bases de datos hay que añadirles las que funcionan en la nube y que están dando lugar a la tendencia DBaaS (*Database as a Service*). Algunos modelos de bases de datos en la nube:
  - Amazon RDS, DynamoDB, SimpleDB, PostgreSQL.
  - Xeround (MySQL).
  - Microsoft SQL Azure Database (SQL Server).
  - Google App Engine (NoSQL).
  - Salesforce Database.com (Oracle).
  - ClearDB (MySQL).
  - Cloudant (CouchDB).

HADOOP

- Apache Hadoop es una biblioteca de software de código abierto (*open source*) que soporta el procesamiento distribuido de grandes conjuntos de datos a través de miles de computadoras ordinarias. El proyecto Apache Hadoop ha nacido de la mano de las dos grandes empresas de la Web, Google y Yahoo!, cuyos investigadores trabajaron con grandes volúmenes de datos en grandes *clusters* de computadoras.
- *Hadoop es el líder en plataformas de Big Data* y su uso crece de modo espectacular, por no decir exponencial. Hadoop consta de tres componentes principales: Hadoop Distributed File System (HDFS), MapReduce y Hadoop Common. Además existen otras tecnologías complementarias como HBase, Hive, Pig, y otras con la misma filosofía tales como IMPALA de Cloudera, DRILL o Google Big Query

HADOOP - Plataformas

- Apache Hadoop es la plataforma de software de código abierto de mayor impacto en Big Data, pero como sucede con otras soluciones de software abierto no suelen ofertarse con soporte de productos. Por esta razón, han surgido un gran número de vendedores que han lanzados sus propias distribuciones de Apache Hadoop. La mayoría de las empresas que han desplegado Hadoop para uso comercial han seleccionado alguna de las distribuciones comerciales de Hadoop.
- La consultora Forrester publicó, en febrero de 2012 el estudio *The Forrester Wave™: Soluciones Hadoop empresariales*, primer trimestre de 2012, donde evaluó las distribuciones comerciales más populares. Destaca como líderes del mercado a Amazon Web Services, IBM, EMC Greenplum, MapR, Cloudera y Hortonworks, junto con otros siete proveedores que prestan sus servicios a nichos clave muy cercanos, Pentaho, DataStax Datameer Platform Computing, Zettaset, Outerthought y HStreaming.

OLAPs Tools

- Dundas BI
- Sisense
- IBM Cognos Analytics
- InetSoft
- SAP Business Intelligence
- Halo

Gracias

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades