GUION DE CLASE BI.DOCX

# GUION DE CLASE: BI

“BI es un proceso centrado en el usuario que permite explorar datos, relaciones entre datos y tendencias, permitiendo mejorar la toma de decisiones. Esto incluye un proceso interactivo de acceso a los datos y el análisis de los mismos para obtener conclusiones”

## ¿Qué es un Data Warehouse?

Un *data warehouse* es un repositorio de información extraída de otros sistemas de la compañía y se construye duplicando los datos que existen en algún otro lugar

### Ventajas:

* no se utilizan los recursos de los sistemas transaccionales para realizar las consultas, y como consecuencia no se recargan dichos sistemas.
* se obtiene mayor seguridad debido a que los datos de los sistemas de la compañía no pueden ser alterados con consultas realizadas por los usuarios.

### Características particulares:

Están:

* orientados al tema,
* son integrados,
* no volátiles e
* historiados

#### Orientación al tema

Los datos se estructuran por temas, contrariamente a los datos de los sistemas transaccionales, organizados generalmente por proceso funcional

#### Datos integrados

La consolidación de todas las informaciones respecto de un cliente dado es necesaria para dar una visión homogénea de dicho cliente.

#### Datos historiados

El data warehouse almacena el historial, es decir, el conjunto de valores que los datos habrán tenido en su historia.

#### Datos no volátiles

Son vistas o fotos de lo que ya ocurrió. En un transaccional, los datos se actualizan (no así los metadatos) pero en el DW no.

## ¿Qué es un data mart?

Suele ser una parte de un DW

## Estructura de un data warehouse

* Datos detallados (por estructura de BD, suelen ser FK)
* Datos comprimidos o zipeados
* Datos históricos
* Metadatos o definiciones de estos
  + Metadatos de la base de datos
  + Metadatos de aplicaciones (informes)

## Métodos de análisis para la toma de decisiones

* Consultas estándar
* Análisis multidimensional
* Modelización y segmentación
* Descubrimiento del conocimiento



### Consultas estándar

Las consultas, en general, están predefinidas: mientras que los datos pueden ir variando día a día, las consultas estándar no pueden ser redefinidas cada vez que se utilizan.

### Análisis multidimensional

Las dimensiones pueden ser: tiempo, ubicación, producto. Ej.: Clientes por región y producto que compraron en 2014. La diferencia es que mientras las consultas estándar recuperan grandes cantidades de datos cruzados, el análisis multidimensional permite ver los mismos datos en forma diferente llamada OLAP (On Line Analytical Processing)

OLAP es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la Inteligencia de Negocios (Business Intelligence), la cual consiste en consultas a estructuras multidimensionales (o Cubos OLAP y que le permite al usuario tener una visión más rápida e interactiva de los mismos. Se usa en informes de negocios de ventas, mercadeo, informes de dirección, minería de datos y áreas similares.

Este análisis, también conocido como análisis del hipercubo, organiza la información según los parámetros que se consulten, de manera tal que a partir de estructuras multidimensionales que contienen los datos resumidos de Sistemas Transaccionales, conocidos como OLTP (Online Transactional Processing) o de grandes bases, se obtendrá la información requerida.

Es muy utilizado en el área de marketing, ventas, informes, etc., especialmente porque las respuestas a consultas complejas se obtienen muy rápidamente y además porque puede obtener los datos tanto de una fuente externa como de una base interna.

La razón de usar OLAP para las consultas es la velocidad de respuesta. Una base de datos relacional almacena entidades en tablas discretas si han sido normalizadas. Esta estructura es buena en un sistema OLTP pero para las complejas consultas multitabla es relativamente lenta. Un modelo mejor para búsquedas, aunque peor desde el punto de vista operativo, es una base de datos multidimensional. La principal característica que potencia a OLAP, es que es lo más rápido a la hora de hacer SELECT, en contraposición con OLTP que es la mejor opción para INSERT, UPDATE y DELETE.

Existen algunas clasificaciones entre las implementaciones OLAP.

ROLAP es una implementación OLAP que almacena los datos en un motor relacional. Típicamente, los datos son detallados, evitando las agregaciones y las tablas se encuentran normalizadas. Los esquemas más comunes sobre los que se trabaja son estrella ó copo de nieve, aunque es posible trabajar sobre cualquier base de datos relacional. La arquitectura está compuesta por un servidor de banco de datos relacional y el motor OLAP se encuentra en un servidor dedicado. La principal ventaja de esa arquitectura es que permite el análisis de una enorme cantidad de datos.

MOLAP es una implementación OLAP que almacena los datos en una base de datos multidimensional. Para optimizar los tiempos de respuesta, el resumen de la información es usualmente calculado por adelantado. Estos valores pre calculados o agregaciones son la base de las ganancias de desempeño de este sistema. Algunos sistemas utilizan técnicas de compresión de datos para disminuir el espacio de almacenamiento en disco debido a los valores pre calculados.

HOLAP (Hybrid OLAP) almacena algunos datos en un motor relacional y otros en una base de datos multidimensional

DOLAP es un OLAP orientado a equipos de escritorio (Desktop OLAP). Trae toda la información que necesita analizar desde la base de datos relacional y la guarda en el escritorio. Desde ese momento, todas las consultas y análisis son hechas contra los datos guardados en el escritorio.

### Modelización y segmentación

Predecir eventos futuros. El trabajo predictivo puede realizarse utilizando determinadas herramientas de análisis, llamadas MODELOS (una colección de patrones). Ej.: reglas por las cuales un cliente **probablemente** se fidelice o se dé de baja. Esto, es un modelo de comportamiento. Segmentación generalmente refiere a la segmentación de clientes. Por ej.: segmentación geográfica. Consejos: *“Divide & Conqueror”* y *“navaja de Occam”*

SEGMENTAR

Análisis de conjuntos y segmentación de datos. Los usuarios pueden aprovechar los conjuntos de análisis que proporciona para realizar una segmentación sencilla de los datos. Pueden manipular y combinar conjuntos de datos definidos por el usuario para obtener un conjunto de datos depurado para su posterior análisis. El análisis de conjuntos es una parte esencial de la plataforma de y se puede aplicar al almacén de datos completo. Los conjuntos se pueden personalizar por usuario, compartir en un departamento, combinar mediante operadores lógicos y reutilizar en múltiples informes.

FILTRAR

Los usuarios pueden crear informes personalizados seleccionando objetos de datos para mostrar y definir calificaciones como criterios de filtro del informe. Ofrece varios enfoques guiados para crear informes nuevos - desde preguntas paso a paso al usuario para seleccionar y calificar los atributos y métricas empresariales hasta asistentes que incorporan plantillas y filtros existentes.

PROFUNDIZAR

Proporciona el conjunto más amplio y exhaustivo de funciones analíticas disponible en cualquier plataforma. Incluyen desde simples funciones matemáticas, como totales actualizados, a cálculos estadísticos avanzados como los f-tests. Los paquetes analíticos   
disponibles incluyen bibliotecas de funciones estadísticas, financieras y matemáticas. Esto permite a la empresa obtener respuesta a todos sus análisis utilizando una sola plataforma.

SINTETIZAR

La plataforma ofrece análisis SQL iterativo multi-paso, una técnica que combina las funciones analíticas integradas con proceso de base de datos. Preguntas complejas que son imposibles de resolver con otras soluciones obtienen respuesta de forma rápida y fácil. Al combinar potentes funciones analíticas con el motor de generación SQL líder en el sector, pone al alcance de todos los usuarios posibilidades inigualables de análisis.

### Descubrimiento del conocimiento

Uno de los métodos de descubrimiento del conocimiento es el análisis de afinidad, que busca en el data warehouse afinidad entre diferentes comportamientos (ej.: Los padres que llevan pañales en general llevarían cervezas en lata; ergo, hay que dejarlos cercanos). Un ejemplo de amplio uso en la venta cruzada en internet junto con el análisis de afinidad es el de las sugerencias propuestas por Amazon.com:

”Clientes que compraron el libro A también compraron el libro B”, por ejemplo, o,

”Personas que leyeron [Historia de Portugal](http://es.wikipedia.org/wiki/Historia_de_Portugal) también se interesaron en [Historia naval](http://es.wikipedia.org/w/index.php?title=Historia_naval&action=edit&redlink=1).”

## Data mining (minería de datos)

Desde un punto de vista orientado al proceso, existen tres tipos de data mining

* **Descubrimiento**: es el proceso en el cual se buscan patrones ocultos en una base de datos sin una idea predeterminada, o hipótesis acerca de cuáles pueden ser esos patrones
* **Modelado predictivo**: los patrones descubiertos en la base de datos son utilizados para predecir el futuro. Mientras que el descubrimiento descripto anteriormente se encarga de encontrar patrones en los datos, el modelado predictivo aplica los patrones para determinar valores probables
* **Análisis forense**: patrones extraídos para encontrar datos inusuales o anomalías

## Herramientas de visualización

* *Drill down/drill up*: yendo de lo global hacia el detalle (drill down) o viceversa (drill up). También llamada Lógica de zoom
* *Drill across*: Zoom cruzado a través de la multidimensional
* *Slice and dice:* Permite “saltar” de dimensión

Por ejemplo, el siguiente cubo es un modelo de datos multidimensional que contiene información sobre la venta productos en diferentes sucursal, actual y proyectada.



## Herramientas ETL (Extraction, Transformation, Loading)

**Extracción:** Se extraen los datos de su fuente original para determinar los temas (conceptos lógicos) que se deben crear y cargar en el data warehouse. Una parte intrínseca del proceso de extracción es la de analizar los datos extraídos, de lo que resulta un chequeo que verifica si los datos cumplen la pauta o estructura que se esperaba. De no ser así los datos son rechazados.

**Transformación:** La fase de transformación aplica una serie de [reglas de negocio](http://es.wikipedia.org/wiki/Reglas_de_negocio) o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Algunas fuentes de datos requerirán alguna pequeña manipulación de los datos. No obstante en otros casos pueden ser necesarias aplicar algunas de las siguientes transformaciones:

* Seleccionar sólo ciertas columnas para su carga (por ejemplo, que las columnas con valores nulos no se carguen).
* Traducir códigos (por ejemplo, si la fuente almacena una "H" para Hombre y "M" para Mujer pero el destino tiene que guardar "1" para Hombre y "2" para Mujer).
* Codificar valores libres (por ejemplo, convertir "Hombre" en "H" o "Sr" en "1").
* Obtener nuevos valores calculados (por ejemplo, total\_venta = cantidad \* precio, o Beneficio = PVP - Coste).
* Unir datos de múltiples fuentes (por ejemplo, búsquedas, combinaciones, etc.).
* Calcular totales de múltiples filas de datos (por ejemplo, ventas totales de cada región).
* Generación de campos clave en el destino.
* Transponer o pivotar (girando múltiples columnas en filas o viceversa).
* Dividir una columna en varias (por ejemplo, columna "Nombre: García López, Miguel Ángel"; pasar a dos columnas "Nombre: Miguel Ángel", "Apellido1: García" y "Apellido2: López").
* La aplicación de cualquier forma, simple o compleja, de [validación](http://es.wikipedia.org/wiki/Pruebas_de_validaci%C3%B3n) de datos, y la consiguiente aplicación de la acción que en cada caso se requiera:
* Datos OK: Entregar datos a la siguiente etapa (Carga).
* Datos erróneos: Ejecutar políticas de [tratamiento de excepciones](http://es.wikipedia.org/wiki/Manejo_de_excepciones) (por ejemplo, rechazar el registro completo, dar al campo erróneo un valor [nulo](http://es.wikipedia.org/wiki/Null) o un valor centinela).

**Carga:** La fase de carga es el momento en el cual los datos de la fase anterior (**transformación**) son cargados en el sistema de destino

Existen dos formas básicas de desarrollar el proceso de carga:

**Acumulación simple**

La acumulación simple es la más sencilla y común, y consiste en realizar un resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y transportar el resultado como una única transacción hacia el data warehouse, almacenando un valor calculado que consistirá típicamente en un sumatorio o un promedio de la magnitud considerada.

**Rolling**

El proceso de **Rolling** por su parte, se aplica en los casos en que se opta por mantener varios niveles de [granularidad](http://es.wikipedia.org/wiki/Tabla_de_hechos#Granularidad) ([jerarquías](http://es.wikipedia.org/wiki/Tabla_de_dimensi%C3%B3n#Granularidad_de_dimensi.C3.B3n_y_jerarqu.C3.ADas)). Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos en alguna o varias de las dimensiones de la magnitud almacenada (por ejemplo, totales diarios, totales semanales, totales mensuales, etc.).

## Motivaciones y beneficios de los data warehouses

* Homogeniza los datos independientemente de qué metadato sea
* Entorno orientado al usuario final
* Consolidan datos heterogéneos

## Estructura de un sistema BI



## Sistemas transaccionales (OLTP) vs. Sistemas analíticos (OLAP)

(OLTP: *On Line Transaction Processing ;* OLAP: *On Line Analytical Processing*)

* Los sistemas analíticos (OLAP) están orientados al tema y tienen un alcance corporativo.
* Los sistemas analíticos no solo trabajan con datos actuales, sino también históricos
* Los sistemas analíticos no se actualizan con cada transacción en el nivel operativo, sino que lo hacen periódicamente, y son utilizados por analistas y en el nivel gerencial
* Los sistemas transaccionales tienen una estructura bidimensional, mientras que los analíticos tienen una estructura multidimensional
* El tiempo de respuesta en un sistema transaccional es crítico, mientras que en un sistema analítico, cuya función es la de proveer información para análisis, no lo es.

### Sistemas ROLAP y MOLAP

Los términos ***ROLAP*** y ***MOLAP*** corresponden a *Relational* OLAP y *Multidimensional* OLAP respectivamente

MOLAP utiliza tablas multidimensionales en forma de cubos, mientras que ROLAP utiliza tablas bidimensionales, que forman cubos temporales en respuesta a las consultas específicas

## Implementación de un sistema BI

El éxito de un data warehouse depende de que se comience con una identificación de los **requerimientos del negocio**. Estos requerimientos son los que determinarán el diseño del data warehouse y los datos que serán necesarios.