

DATA WAREHOUSE

**ASIGNATURA: SISTEMA PARA EL SOPORTE A LA
TOMA DE DECISIONES**



Por: MSc. Ing. Edwin Calle Terrazas

DATA WAREHOUSE (DS)

Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia.



DW (Data Warehouse) es una base de datos con estructura multidimensional.

ARQUITECTURA DE UN DATA WAREHOUSE (DW)

1) Data Sources

2) Load Manager

- ✓ Extracción
- ✓ Transformación: Codificación, medida de atributos, fuentes múltiples data cleansing
- ✓ Carga

3) Data Warehouse Manager

- ✓ Base de datos multidimensional
- ✓ Tablas de dimensiones: Tiempo
- ✓ Tablas de hechos: Agregadas y Preagregadas
- ✓ Cubo Multidimensional: Representación matricial
- ✓ Modelos del DW: Estrella, Copo de Nieve y Constelación
- ✓ OLTP vs DW
- ✓ Implementación: ROLAP, MOLAP, HOLAP
- ✓ Cubo Multidimensional: Jerarquías
- ✓ Metadatos

4) Query Manager

DATA SOURCE

Los Data Sources (orígenes de datos) representan los datos transaccionales que genera la empresa en su accionar diario, junto a otros datos internos y/o externos complementarios.

Los Data Sources poseen características muy disímiles entre sí, en formato, procedencia, función, etc.

Los Data Sources, más habituales para extraer datos relevantes son:

- ✓ Archivos de textos
- ✓ Hojas de cálculos
- ✓ Informes semanales, mensuales, anuales, etc.
- ✓ Bases de datos transaccionales, SQL y NoSQL
- ✓ Información no estructurada (páginas web, mails)
- ✓ Redes sociales
- ✓ Web Services

LOAD MANAGER

El componente Load Manager es el encargado de la ejecución y calendarización (scheduling) de los diferentes procesos de Integración de Datos a través de los cuales:

- ✓ Se extraerán los datos desde los Data Sources
- ✓ Serán manipulados, integrados y transformados
- ✓ Para luego cargar los resultados obtenidos en el DW.

El proceso ETL (Extraction, Transformation, Load) es una de las muchas técnicas más importantes de la **Integración de datos**. En este orden, se puede ubicar el resto de las técnicas en las diferentes etapas del ETL:

- ✓ El proceso Extracción incluirá técnicas enfocadas por ejemplo a obtener desde diversas fuentes solamente los datos relevantes
- ✓ El proceso Transformación incluirá técnicas encargadas de compatibilizar formatos, filtrar y clasificar datos, relacionar diversas fuentes, etc;
- ✓ El proceso Carga incluirá técnicas propias de la carga de datos y actualización del DW.

DATA WAREHOUSE MANAGER

El DW Manager está compuesto por una serie de aplicaciones de software dedicadas a gestionar:

- ✓ El DW (SGDB)
- ✓ Las conexiones a base de datos y otros Data Sources
- ✓ Las estructuras de datos (Cubos Multidimensionales, Business Models, etc),
- ✓ Información de autenticación y autorización (credenciales de acceso, users, roles, permisos, etc)
- ✓ Otros metadatos.

BASE DE DATOS MULTIDIMENSIONAL

Un Data Warehouse, es una base de datos con estructura multidimensional, lo cual se traduce en una forma específica de almacenamiento en la cual se definen dos estructuras principales :
tablas de Hechos y tablas de Dimensiones.

La utilización de tablas de Hechos y Dimensiones, facilita la creación de estructuras de datos (Cubos Multidimensionales, Business Models, etc.) y posibilita que las consultas al SGBD sean respondidas con mucha performance.

TIPOS DE MODELOS LÓGICOS

Existen tres formas de modelar las tablas de Hechos y Dimensiones:

- ✓ Esquema en estrella (Star Scheme).
- ✓ Esquema copo de nieve (Snowflake Scheme).
- ✓ Esquema constelación (Starflake Scheme).

QUERY MANAGER

El Query Manager es una pieza fundamental y compleja del proceso de DWH, pues es el encargado de realizar las operaciones necesarias para soportar los procesos de gestión y ejecución de:

- ✓ Consultas relacionales: como JOIN y agregaciones (SUM, COUNT, AVG, etc)
- ✓ Consultas propias del análisis de datos: como DRILL-UP y DRILL-DOWN.

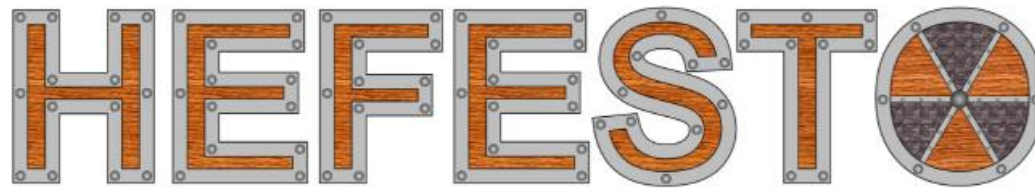
El funcionamiento del Query Manager es el siguiente:

1. Recibe consultas de los usuarios, que en general están escritas en un lenguaje de alto nivel (por ejemplo MDX);
2. Lee los metadatos que describen los mapeos (Cubo Multidimensional, Business Models, etc) y reescribe las consultas para que sean ejecutadas en el sistema destino (por lo general SQL);
3. Una vez que obtiene los datos y utilizando, nuevamente, las estructuras de metadatos, éstos son transformados a un formato final de alto nivel que será interpretado y renderizado por las herramientas de visualización.

DATA MART

Un Data Mart, es una versión específica del almacén de datos centrados en un tema o un área de negocio dentro de una organización. Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.

Como por ejemplo, Ventas, Recursos Humanos u otros sectores en una organización.



HEFESTO es una metodología, cuya propuesta está fundamentada en una extensa investigación, comparación de metodologías existentes y el aporte de experiencias propias en procesos de diseño e implementación de data warehouse.

HEFESTO está compuesto por los siguientes pasos:

1) ANÁLISIS DE REQUERIMIENTOS

1.1) Preguntas del Negocio

1.2) Indicadores y Perspectivas

1.3) Modelo Conceptual

2) ANÁLISIS DE DATA SOURCES

2.1) Hechos e Indicadores

2.2) Mapeo

2.3) Granularidad

2.4) Modelo Conceptual Ampliado

3) MODELO LÓGICO DEL DW

3.1) Tipología

3.2) Tablas de Dimensiones

3.3) Tablas de Hechos

3.4) Uniones

4) INTEGRACIÓN DE DATOS

4.1) Carga Inicial

4.2) Actualización