

# Data Warehouse

Javier Bonet  
Joel Catacora

- Base de datos avanzada • 1 de abril del 2015

A large, solid teal circle is centered on a dark, textured background that resembles a scratched or worn surface. The circle is the primary visual element, and the text is centered within it.

**BUSINESS INTELLIGENCE**

En 1989, Howard Dresner (un analista de Gartner Group) propuso la "**inteligencia de negocios**" como un término general para describir "los conceptos y métodos para mejorar la toma de decisiones empresariales mediante el uso de sistemas basados en hechos".

Se denominada **Business intelligence** (BI), al conjunto de estrategias que integran, por un lado el almacenamiento, y por el otro, el procesamiento de grandes cantidades de datos, con el principal objetivo de transformarlos en conocimiento y en decisiones en tiempo real, a través del análisis de los datos existentes en una organización o empresa.

**Dato + Análisis = Conocimiento.**



¿QUÉ ES UN DATA WAREHOUSE?

Definición en términos de las características del DW:

**Data Warehouse** (DW o DWH), es una colección de datos orientada a un ámbito determinado, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.



Una definición más amplia que la anterior:

Un **Data Warehouse**, es un sistema que extrae, limpia, ajusta, y entrega los datos de origen en un almacén de datos dimensional, y luego apoya e implementa consultas y análisis, con el propósito de asistir en la toma de decisiones.



CARACTERÍSTICAS



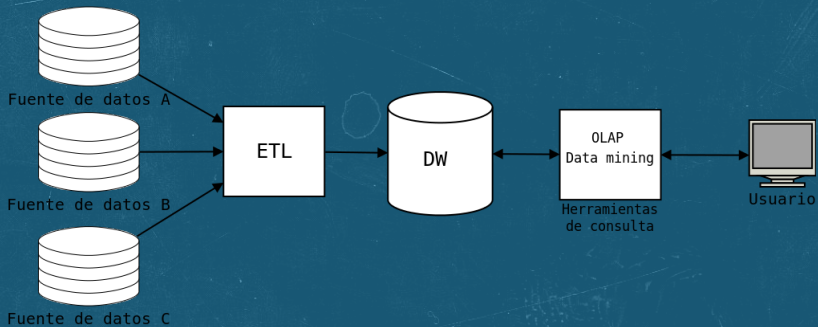
- **Integrado:** Se integran datos provenientes de múltiples fuentes, posiblemente distintas.

- **Integrado:** Se integran datos provenientes de múltiples fuentes, posiblemente distintas.
- **No volátil:** Una vez almacenados los datos en el DW, la información que éstos representan no debe perderse.

- **Integrado:** Se integran datos provenientes de múltiples fuentes, posiblemente distintas.
- **No volátil:** Una vez almacenados los datos en el DW, la información que éstos representan no debe perderse.
- **Variable en el tiempo:** La información histórica se mantiene en el DW a lo largo del tiempo.

A large, solid teal circle is centered on a dark, textured background that resembles a film negative or a heavily scratched surface. The circle is the primary focus of the image.

ARQUITECTURA



A large teal circle is centered on a dark, textured background that resembles a film negative or a scratched surface. The circle is a solid, medium-blue color. Inside the circle, the text "OLTP VS. DW" is written in a white, sans-serif font.

OLTP VS. DW



	OLTP	Data Warehouse
Objetivo	Soportar actividades transaccionales diarias.	Consultar y analizar información estratégica y táctica.
Modelo de datos	Normalizado.	Desnormalizado.
Datos consultados	Actuales.	Actuales e históricos.
Horizonte de tiempo	60 - 90 días.	5 - 10 años.
Tipos de consultas	Repetitivas, predefinidas	No previsibles, dinámicas
Nivel de almacenamiento	Nivel de detalle.	Nivel de detalle y diferentes niveles de sumarización.
Acciones disponibles	Alta, baja, modificación y consulta.	Carga y consulta.
Número de transacciones	Elevado	Medio o bajo
Tamaño	Pequeño - Mediano.	Grande.
Tiempo de respuesta	Pequeño (segundos - minutos).	Variable (minutos - horas).
Sello de tiempo	La clave puede o no tener un elemento de tiempo.	La clave tiene un elemento de tiempo.
Estructura	Generalmente estable.	Generalmente varía de acuerdo a su propia evolución y utilización.

# DATA MARTS

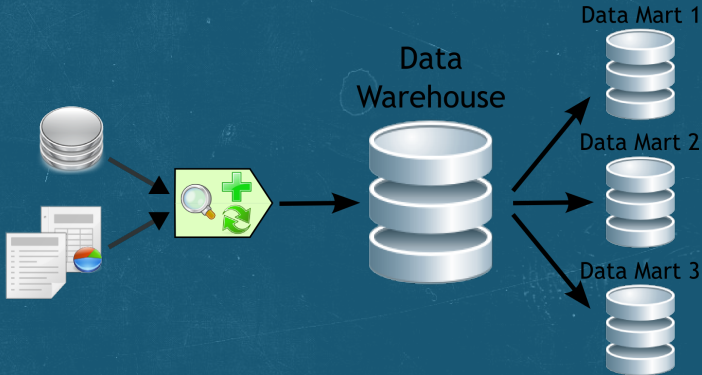
Los **Data marts** (DM), son subconjuntos de datos de un data warehouse para áreas específicas.

Entre las características de un data mart destacan:

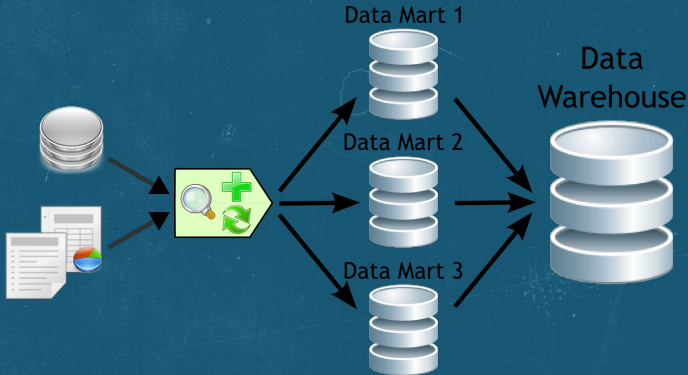
- Usuarios limitados.
- Tiene un propósito específico.
- Tiene una función de apoyo.

# METODOLOGÍAS DE DISEÑO

**Top-Down:** primero se define el DW y luego se desarrollan, construyen y cargan los DM a partir del mismo.



**Bottom-Up:** se definen previamente los DM y luego se integran en un DW centralizado.





# MODELO DIMENSIONAL

El modelado dimensional es una técnica de diseño lógico de una base de datos, útil para el procesamiento analítico en línea (OLAP), que tiene como ideas centrales el rendimiento y lograr la facilidad de comprensión para el usuario.

Hay dos conceptos centrales:

- Hechos (métricas).
- Dimensiones.

# ESQUEMA DIMENSIONAL

Sabemos que la relación entre todas las tablas de una base de datos se denomina esquema de base de datos. Para un cierto grupo de bases de datos, en las cuales se realizan consultas sobre datos históricos, generalmente se utilizan diseños llamados **esquemas dimensionales**.

Un **esquema dimensional** separa físicamente las medidas que cuantifican el negocio (hechos) de los elementos que los describen (dimensiones).

## Esquema estrella

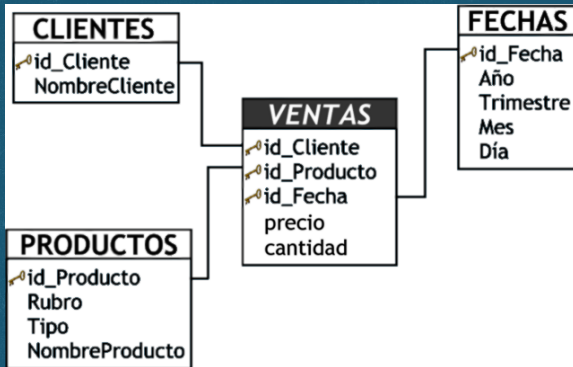
En este tipo de esquemas la idea central es tener,

**Tabla de hechos**

rodeada de

**Tablas de dimensiones**

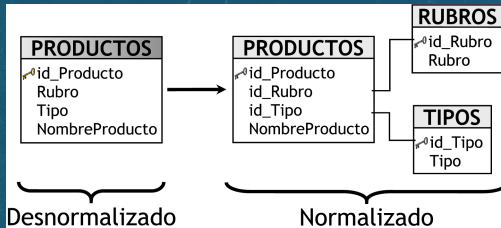
## Esquema estrella





## Esquema estrella

Este modelo debe estar desnormalizado, es decir que no puede presentarse en tercera forma normal (3ra FN). Si se normaliza la tabla "PRODUCTOS", obtendremos lo siguiente:



## Esquema de copo de nieve

Este otro tipo de esquemas, es similar al esquema de estrella, salvo que las dimensiones pueden estar conectadas con otras tablas de dimensiones.

Tendremos,

**Tabla de hechos**

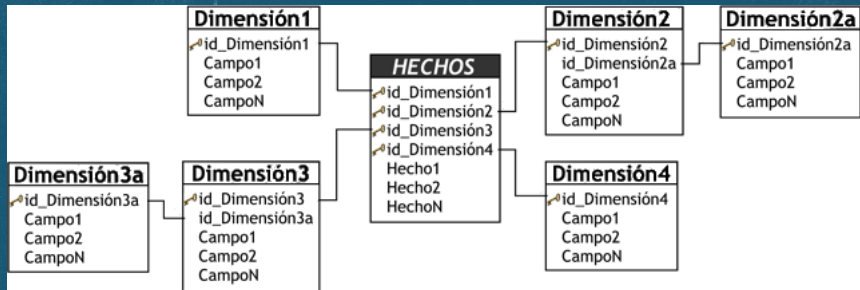
rodeada de

**Tablas de dimensiones**

conectadas con

**Nuevas tablas de dimensiones**

## Esquema de copo de nieve



## Esquema de constelación

Este esquema es una combinación de los dos anteriores, utiliza lo mejor de cada uno, la simplicidad del esquema estrella junto con el cierto nivel de normalización del esquema copo de nieve. Está compuesto de la siguiente forma:

**Una o más tablas de hechos**

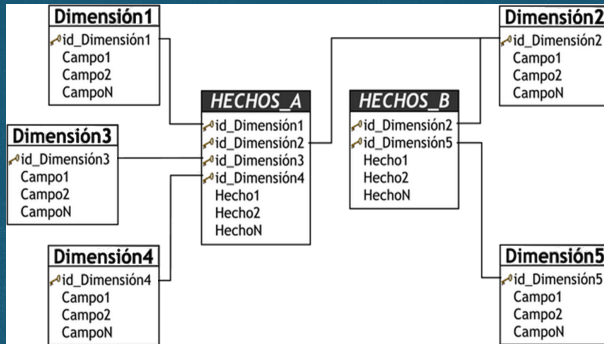
rodeada de

**Tablas de dimensiones**

conectadas (posiblemente) con

**Nuevas tablas de dimensiones**

## Esquema de constelación



A large, solid teal circle is centered on a dark, textured background that resembles a film negative or a scratched surface. The circle is the primary focus of the image.

DIMENSIÓN TIEMPO



Un caso particular entre las dimensiones que se definirán en el data warehouse es el de la **dimensión de tiempo**.

En un DW, la creación y el mantenimiento de una tabla de dimensión tiempo es obligatoria. Por ello, decimos que un DW es una **base de datos temporal**.

Tenemos diferentes elecciones para el tipo del campo clave:

- Tipo Date.
- Tipo Entero:
  - Autoincremental (1,2,3,...,35200,...).
  - yyyyMMdd (por ejemplo 20150310).

VENTAJAS/DESVENTAJAS

## **Ventajas:**

- Proporciona información clave para toma de decisiones.
- Muy útil para el almacenamiento de datos orientados para el análisis, y consultas de datos históricos.
- Permite mayor flexibilidad y rapidez en el acceso a la información.
- Los DW pueden trabajar en conjunto y, por lo tanto, aumentar el valor operacional de las aplicaciones empresariales.

## **Desventajas:**

- Requiere continua limpieza, transformación e integración de datos.
- A lo largo de su vida los almacenes de datos pueden suponer altos costos. El almacén de datos no suele ser estático. Los costos de mantenimiento son elevados.
- Puede presentar ciertas dificultades a la hora de la implementación, debido a los objetivos que pretende la organización.
- A veces, ante una petición de información estos devuelven una información subóptima, lo que supone una pérdida para la organización.