**Actividad XI**

Data warehouse

Alumno: Juan Pablo Padilla Martin

217294261



**Datawarehouse definición y ejemplificación**

Un data warehouse es un repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización, independiente de cómo se vayan a utilizar posteriormente por los consumidores o usuarios, con las propiedades siguientes: estable, coherente, fiable y con información histórica.

Al abarcar el ámbito global de la organización y un amplio alcance histórico, el volumen de datos de un data warehouse puede ser muy grande (centenas de terabytes). Las bases de datos relacionales han sido el medio más usado para almacenar los data warehouse. No obstante, en la actualidad existen otras opciones que ofrecen modelos de datos más cercanos a las necesidades analíticas, como por ejemplo, las bases de datos orientadas a columnas o incluso las basadas en lógica asociativa.

Resumiendo, el data warehouse presenta las siguientes características:

• Orientado a un tema: organiza una colección de información alrededor de un tema central.

• Integrado: incluye datos de múltiples orígenes y presenta consistencia de datos.

• Variable en el tiempo: se realizan fotografías de los datos basadas en fechas o hechos a lo largo de periodos de tiempo prefijados.

• No volátil: sólo de lectura para los usuarios finales.

Debemos tener en cuenta que existen otros elementos en el contexto de un data warehouse. Podemos diferenciar estos elementos en los almacenes de datos, procesos y metadatos:

a) Almacenes de datos: permiten definir versiones más actuales, sectoriales o versiones intermedias del data warehouse. De entre ellos podemos destacar:

• Data Mart: contiene un subconjunto de los datos del DataWarehouse con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica. Al igual que en un data warehouse, los datos están estructurados en modelos de estrella o copo de nieve. Los datamarts pueden ser dependientes o independientes de un datawarehouse. El data mart está pensado para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de la organización.

• Operational Data Store: es un tipo de almacén de datos que tiene como objetivo permitir realizar analítica sobre los últimos valores de los datos no disponibles en el data warehouse. Para ello, los Operational Data Stores proporcionan los últimos valores de los datos, pero no su historial. Permiten generalmente un pequeño desfase o retraso sobre los datos operacionales.

• Staging Area: es un área de almacenaje intermedia que se encuentra entre las fuentes de datos y el data warehouse. Sus principales objetivos son:

– Facilitar la extracción de datos desde fuentes de origen que tengan una gran heterogeneidad y complejidad.

– Mejorar la calidad de datos.

– Ofrecer una cache de datos operacionales que de soporte al proceso de data warehousing.

– Permitir al acceso a información no contenida en el datawarehouse.

b) Procesos: permiten el análisis, filtraje, mejora, integración y carga de los datos en los distintos almacenes de datos. Podemos destacar:

• Data warehousing: es el proceso de diseño, creación, mantenimiento y ampliación de la arquitectura de un data ware-house. Dentro de los procesos de data warehousing destacan aquellos que permiten extraer y filtrar datos de las operaciones comunes de la organización, procedentes de los distintos sistemas de información operacionales y/o sistemas externos, para transformarlos, integrarlos y almacenarlos en un almacén de datos con el fin de acceder a ellos para dar soporte en el proceso de toma de decisiones de una organización.

• Procesos de ETL: es una de las principales tecnologías de integración de datos basada en la consolidación de datos que se usa tradicionalmente para alimentar los distintos almacenes de datos (data warehouse, data mart, staging area y operational datastore). Usualmente se combina con otras técnicas de consolidación de datos. Es la tecnología que usaremos en este libro.

c) Metadatos: Datos estructurados y codificados que describen características de los datos y los procesos que participan en los procesos de data warehousing.

2.1. Elementos de un data warehouse

Es de esperar que la estructura relacional de una base de datos operacional siga las formas normales de Codd en su diseño. En un datawarehouse no debe seguirse ese patrón de diseño y se permite desnormalizar la información con el objetivo de optimizar las consultas.

Para identificar la información a tener en cuenta en el data warehouse, debemos identificar, en el seno de nuestra organización, los procesos de negocio, las vistas para el proceso de negocio y las medidas cuan tificables asociadas a los mismos. De esta manera hablaremos de los siguientes elementos del data warehouse:

• Hechos: permiten representar los procesos de negocio de la organización. Por ejemplo, una venta puede identificarse como un proceso de negocio de manera que es factible, si corresponde en nuestra organización, considerar el hecho ventas.

• Dimensiones: permiten representar las distintas vistas para un cierto proceso de negocio (o los distintos puntos de vista desde los cuales se puede analizar). Si regresamos al ejemplo de una venta, podemos estudiarla a partir del cliente que la ha realizado, la fecha en la que se ha realizado, los productos que se han venido... Estos conceptos pueden ser considerados como vistas para este proceso de negocio. Puede ser interesante recuperar todas las compras realizadas por un cliente. Ello nos hace entender por qué la identificamos como una dimensión.

• Métricas: son los indicadores de negocio de un proceso de negocio, es decir, aquellos conceptos cuantificables que permiten medir un proceso de negocio. Están asociados a las tablas de hecho. Por ejemplo, en una venta tenemos el importe de la misma.

Existen principalmente dos tipos de esquemas para estructurar los datos en un almacén de datos:

• Esquema en estrella: permite estructurar la información de procesos, vistas y métricas mediante una estructura de estrella. A nivel de diseño, consiste en una tabla de hecho (también denominada a veces fact table) en el centro para el hecho objeto de análisis y una o varias tablas de dimensión por cada punto de vista de análisis que participa de la descripción de ese hecho. En la tabla de hecho encontramos los atributos destinados a medir (cuantificar): sus métricas. La tabla de hechos sólo presenta uniones con dimensiones.

• Esquema en copo de nieve: es un esquema de representación derivado del esquema en estrella, en el que las tablas de dimensión se normalizan en múltiples tablas. Por esta razón, la tabla de hechos deja de ser la única tabla del esquema que se relaciona con otras tablas, y aparecen nuevas uniones.

Debido a su importancia, creemos que es conveniente profundizar en los conceptos de tabla de hecho, dimensión y métrica.

**Bibliografías**