6. Procesamiento analítico en línea (OLAP)

6.1. Introducción

Las bases de datos relacionales tienen como características principales:

1. Proporcionar métodos simples y confiables de almacenamiento y procesamiento de datos, y modos flexibles de acceso a los mismos.
2. Apoyar las funciones diarias de las empresas con aplicaciones de negocios que almacenan y analizan datos con precisión y confiabilidad.
3. Dar soporte a sistemas de información que también se conocen como de OLTP (On Line Transaction Processing).
4. Ser una fuente valiosa de datos para entender las tendencias y el modo en que la empresa funciona.

Ejemplos de sistemas OLTP son: registro de compras, registro de ventas, control de producción. Sin embargo, los sistemas OLTP no son muy apropiados para realizar análisis de datos.

Un sistema **OLAP (On Line Analytic Processing)** está hecho para llevar a cabo, principalmente, análisis de datos. El término OLAP se refiere a un conjunto de conceptos, lenguajes y productos cuyo objetivo central es facilitar el análisis de datos. Ejemplos de sistemas OLAP son: planeación de recursos, presupuestos, análisis de ventas.

Los requerimientos funcionales principales de los sistemas OLAP son:

A) Lógicos

1. Permitir una estructuración dimensional poderosa de la información con manejo de jerarquías. Ésta es una característica fundamental para poder modelar los conjuntos altamente multidimensionales del mundo real, con muchos niveles de datos, detalles o abstracciones.
2. Permitir una especificación eficiente de dimensiones y de cálculos dimensionales. Los análisis de datos realizados, además de incluir sumas o promedios de cantidades grandes de datos, involucran comparaciones inteligentes de valores e inferencias de tendencias sobre el tiempo y otras dimensiones, por lo que estas actividades deben efectuarse con gran eficiencia.
3. Brindar flexibilidad en las operaciones. Esto implica flexibilidad en la visualización de la información (gráficas, matrices, cartas, etc.), en las definiciones (formateo de datos, definiciones de fórmulas, etc.), en el análisis y en las interfaces (navegación en un modelo, ligas a fuentes externas, etc.).

B) Físicos

1. Separar la estructura interna de almacenamiento de los datos de su representación. Para brindar la posibilidad de hacer cambios en las vistas de los datos sin tener que cambiar la estructura interna de almacenamiento de los mismos.
2. Tener la velocidad suficiente para apoyar análisis ad hoc. Éste es un componente crucial de un sistema OLAP, ya que necesita soportar consultas analíticas ad hoc, algunas de las cuales pueden requerir cálculos realizados "al vuelo". Para un análisis dado, puede ser necesario hacer una serie de consultas encadenadas las cuales deben responderse rápidamente para no perder el hilo del análisis.
3. Brindar soporte multi-usuario. Es deseable que un grupo de personas trabaje simultáneamente para que haya coincidencia hacia metas comunes; por ejemplo, una institución bancaria puede proyectar los ingresos anuales que espera obtener por la venta de sus productos (tarjeta, cheques, etc.), tomando como base las estimaciones de los gerentes y ajustando, después, con las expectativas de los agentes de ventas.

Comparación entre actividades OLTP vs. OLAP

La siguiente tabla muestra un comparativo entre las actividades principales en ambos tipos de sistemas:

Actividades operacionales (OLTP)	Actividades de análisis (OLAP)
Más frecuentes	Más frecuentes
Más predecibles	Menos predecibles
Cantidades de datos más pequeñas accedidas por consulta	Grandes cantidades de datos accedidas por consulta
Consultas mayoritarias sobre datos "sin procesar"	Consultas mayoritarias sobre datos derivados
Requieren datos actuales mayoritariamente	Requieren datos pasados, presentes y proyectados
Si hay derivaciones complejas, normalmente son pocas	Muchas derivaciones complejas

6.2. Modelo de datos multidimensional

Es un modelo para conceptualizar y visualizar los datos como un conjunto de variables (dimensiones) que son definidas por aspectos comunes del negocio. Es especialmente útil para resumir y reordenar los datos en diferentes vistas de los mismos con el fin de realizar su análisis. El análisis dimensional se enfoca principalmente en los datos numéricos como: valores, medidas, balances y ocurrencias.

6.2.1. Espacio multidimensional

Se usa el término **hipercubo** (o **cubo**, de manera abreviada) para describir un espacio de datos multidimensional. Hay que aclarar que no es un cubo en el sentido geométrico (el cual sólo tiene tres dimensiones), sino que el hipercubo puede tener cualquier cantidad de dimensiones,

cada una, posiblemente, con un tamaño distinto. El hipercubo contiene una cantidad discreta (esto es, no continua) de valores en cada dimensión.

6.2.2. Definiciones relacionadas con los cubos

A continuación se describen varias definiciones asociadas al manejo de los cubos:

- Una **dimensión** describe algún elemento en los datos que el negocio quiere analizar. Por ejemplo, las ventas de productos o el tiempo son dimensiones bastante comunes.
- Un **elemento** (**miembro**) corresponde a un punto dentro de una dimensión. Por ejemplo, en la dimensión del tiempo, el mes de *marzo* podría ser un elemento.
- Un **atributo** es una colección completa de elementos. Por ejemplo, todos los meses del año serían un atributo de la dimensión del tiempo. Los trimestres del año podrían ser otro atributo de esta dimensión. De los diversos atributos de una dimensión se escoge uno como *atributo clave*.
- El **tamaño**, o **cardinalidad**, de un atributo de una dimensión es la cantidad de elementos que contiene. Por ejemplo, un atributo de la dimensión del tiempo formado por los meses del año tendría una cardinalidad de 12.

La siguiente figura muestra un cubo con tres dimensiones: tiempo (en meses), productos (por su nombre) y clientes (por su nombre). El cubo define un espacio multidimensional de ventas de un producto específico a clientes específicos sobre un período específico de tiempo, medido en meses.

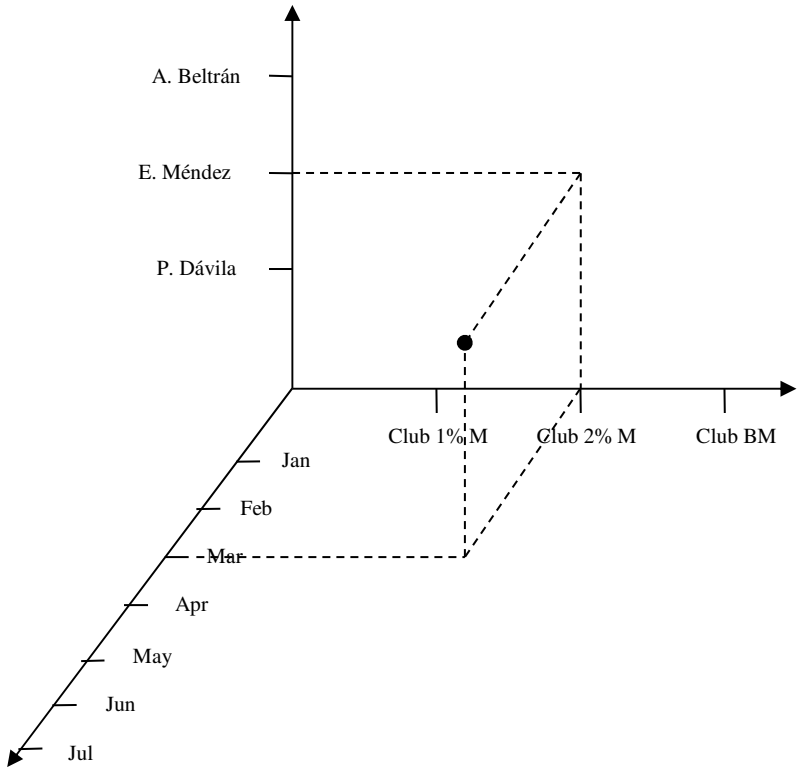


Fig. 6.1. Espacio de datos tri-dimensional que describe ventas de productos a clientes en un período de tiempo.

Más definiciones:

- Un **espacio de hechos**, **datos de hechos** o, simplemente, **hechos**, es el conjunto de puntos en el espacio de datos. Por ejemplo, una venta realizada sería un hecho.
- Una **tupla** es una coordenada en el espacio multidimensional.
- Una **rebanada** (**slice**) es una sección del espacio multidimensional (ver Fig. 6.2). Un término más adecuado para rebanada sería **sub-cubo**.

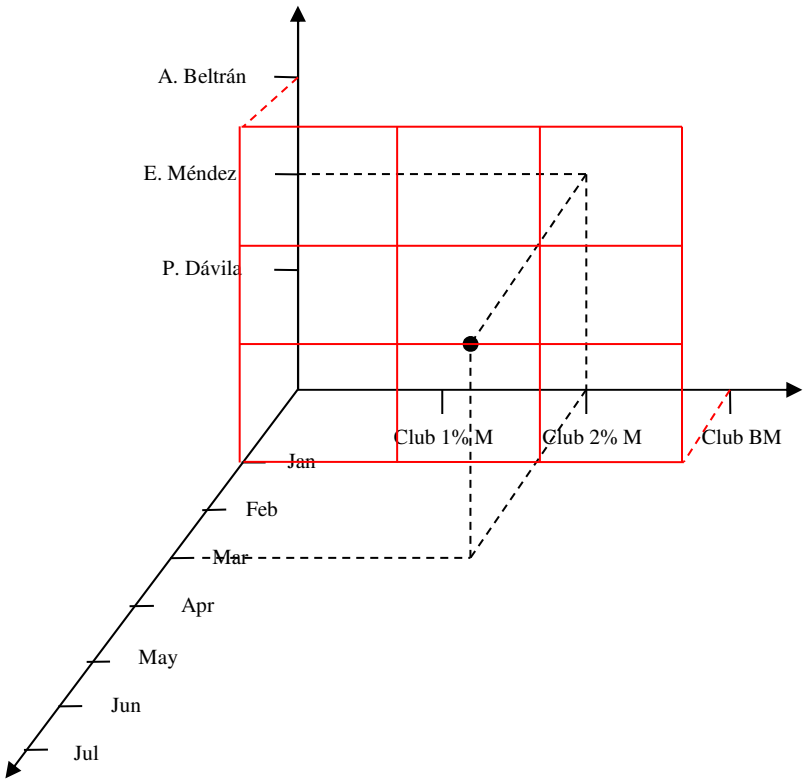


Fig. 6.2. Ejemplo de una *rebanada* de ventas en enero.

- Una **jerarquía** de una dimensión es un tipo de agrupación de sus atributos. Una jerarquía normalmente tiene varios niveles. Ejemplo:

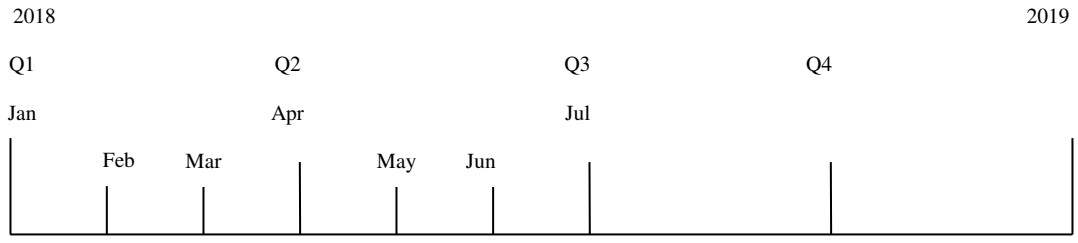


Fig. 6.3. Los atributos relacionados (año, trimestre) son calibrados con respecto al atributo clave (meses).

- Esta jerarquía presenta las siguientes características:
  - Atributo clave de la dimensión: meses.
  - Jerarquía: años, trimestres, meses.
  - Niveles de la jerarquía: 3, que son: años, trimestres, meses.

– Un nivel de la jerarquía es lo mismo que un atributo de la dimensión, aunque también se utiliza el término *atributo de la jerarquía*, como sinónimo de lo anterior.

– Una dimensión puede tener más de una jerarquía, aunque todas usar el mismo atributo clave (por ejemplo, los días o los meses del año).

– Con respecto a la figura anterior, si el atributo clave se cambiara a los días del año, se podrían tener las siguientes jerarquías:
  - Jerarquía 1: años, trimestres, meses, días (cuatro niveles).
  - Jerarquía 2: años, semanas, días (tres niveles).

Más definiciones:

– Cada nivel de una jerarquía define un conjunto de puntos en el espacio multidimensional.

– Los únicos puntos reales (esto es, que existen) son los del **espacio de hechos**.

– Los otros puntos forman un **espacio lógico de datos** y se obtienen sólo por medio de cálculos (por ejemplo, las ventas en un año o en un trimestre).

– El **espacio "completo"** de datos del cubo está formado por el de hechos más el lógico. Cada punto en este espacio se llama **celda**.

La siguiente figura ilustra estos conceptos:

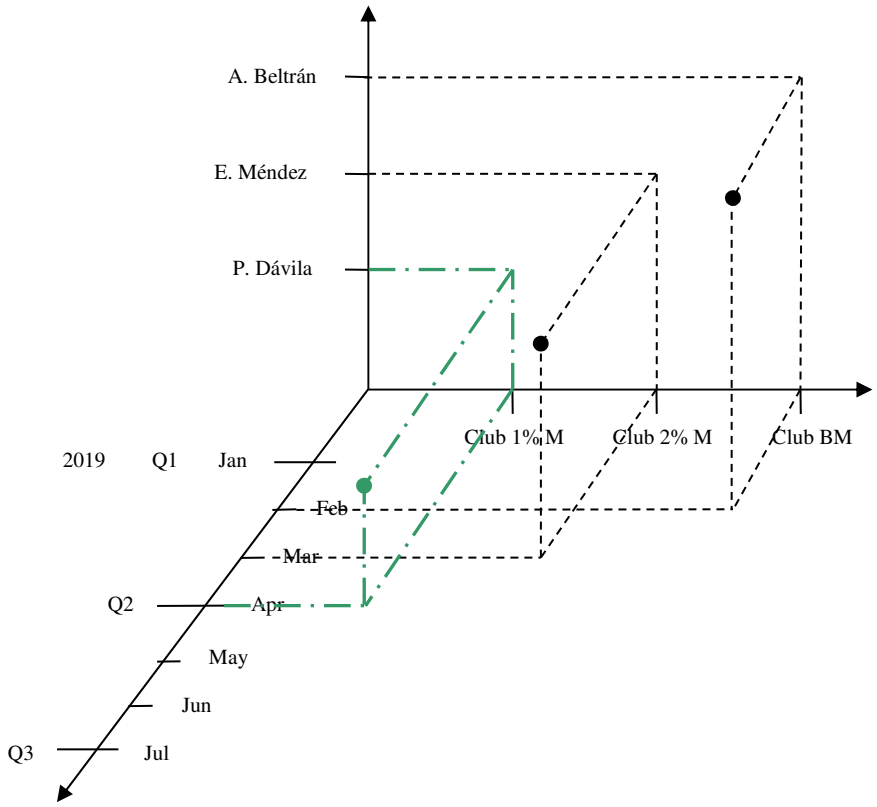


Fig. 6.4. Los atributos relacionados crean nuevos puntos en el espacio multi-dimensional.

Más definiciones:

- Una **medida** es el valor de una celda. Una celda puede tener varias medidas, por ejemplo: el monto de una venta o la cantidad de unidades vendidas de un producto.
- Estas medidas pueden verse como una *dimensión de medidas*, cada medida con tipo de datos, unidad, etc.
- Las **funciones de agregación** son las que calculan los valores de las celdas del espacio lógico de datos, pudiendo ser simples o complejas.

La siguiente figura ilustra lo anterior:

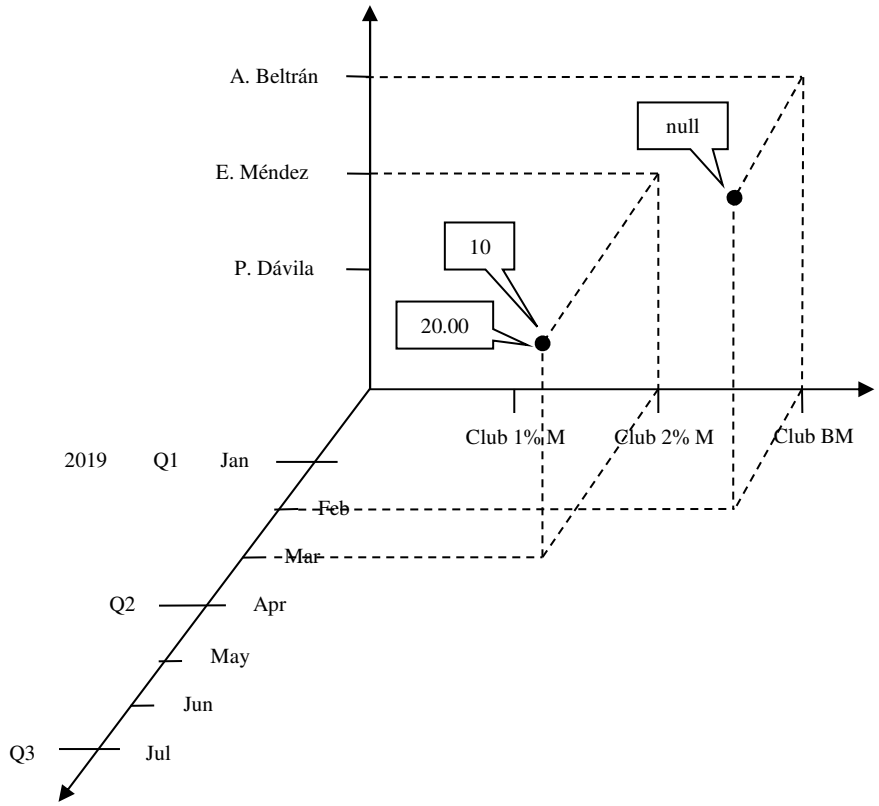


Fig. 6.5. Cada celda de hechos puede tener un valor real o uno potencial.

Cada celda de hechos en el cubo está asociada con una venta real o potencial de un producto a un cliente. Si es una venta potencial, la celda tiene un valor **null**; si es real, tiene un valor distinto de **null**.

Ejemplo de funciones de agregación (para calcular puntos del espacio lógico de datos):

- Simples: suma, cuenta, etc.
- Complejas: fórmulas y algoritmos.

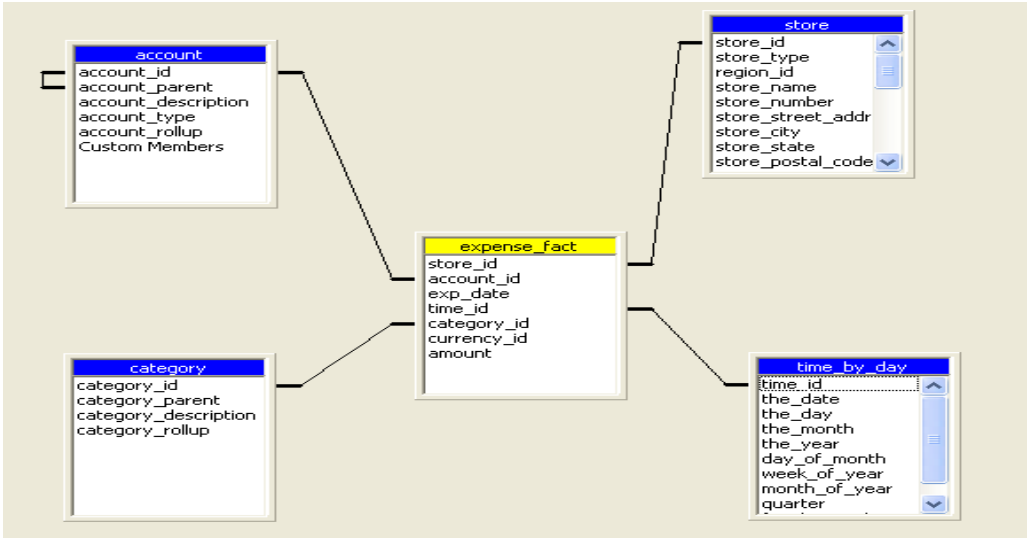
6.2.3. Modelos de diseño para los cubos

Existen dos modelos básicos para el diseño de los cubos:

A) Modelo de Estrella (Star)

Es la estructura básica para un cubo. Se compone de una (o varias) tabla(s) central(es) (llamada(s) tabla(s) de hechos) y de un conjunto de tablas más pequeñas (las tablas de las dimensiones) concentradas alrededor de la(s) tabla(s) de hechos.

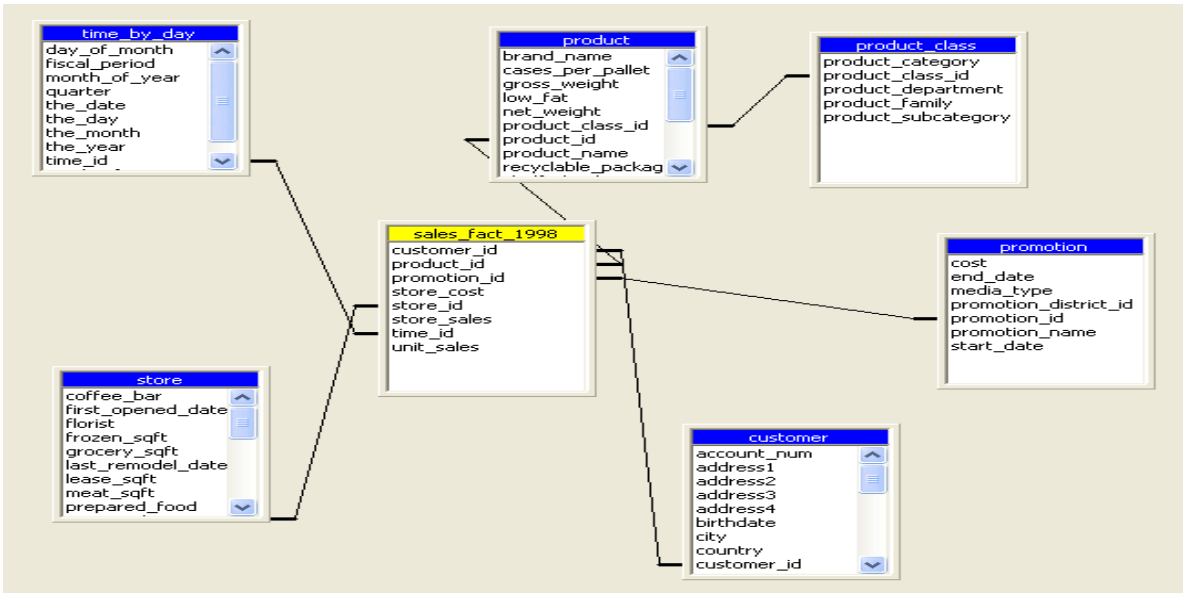
La siguiente figura muestra este diseño:



B) Modelo de Copo de nieve (Snowflake)

Es el resultado de descomponer una jerarquía de una dimensión en una o más tablas. Los vínculos entre estas tablas normalmente serán uno a muchos.

La siguiente figura ilustra el concepto:



6.2.4. Operaciones básicas con los cubos

Son cuatro principalmente:

1. Operación de *Slice*: consiste en mostrar una sección (sub-cubo) del cubo.
2. Operación de *Dice*: consiste en mostrar los datos del cubo desde otra dimensión (también se usa el término "rotar" para esta operación).
3. Operación de *Drill down*: consiste en navegar hacia los niveles inferiores (más detallados) de una jerarquía.
4. Operación de *Drill up*: consiste en navegar hacia los niveles superiores (más agregados) de una jerarquía.