Almacenes de datos (DWH- Data Warehouse)

ITAM - Bases de Datos

Dr. Felipe López G.

Introducción

- Las bases de datos relacionales:
  1. Apoyan las funciones diarias de las empresas con aplicaciones de negocios que almacenan y analizan datos con confiabilidad y precisión.
  2. Los sistemas que conforman también se conocen como de OLTP (On Line Transaction Processing).
  3. Son las fuentes de datos con las que se lleva el día a día de las empresas, sin las cuales éstas no funcionarían.
  4. Sin embargo, los sistemas OLTP no son muy apropiados para realizar análisis de datos.

Introducción

- DWH (también conocido como OLAP - On Line Analytic Processing):

1. Es un conjunto de conceptos, lenguajes y productos cuyo objetivo central es facilitar el análisis de datos.
2. Ejemplos: planeación de recursos, presupuestos, análisis de ventas, análisis financieros, etc.
3. En general los usuarios de DWH no trabajan con una transacción de datos a la vez, sino con cientos de ellas.
4. Normalmente el interés no está en saber, p. ej., qué producto compró un cliente, sino cuáles fueron las ventas del producto la última semana o mes.

Comparación entre actividades OLTP vs. OLAP

Actividades operacionales	Actividades de análisis
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

Una arquitectura para DWH

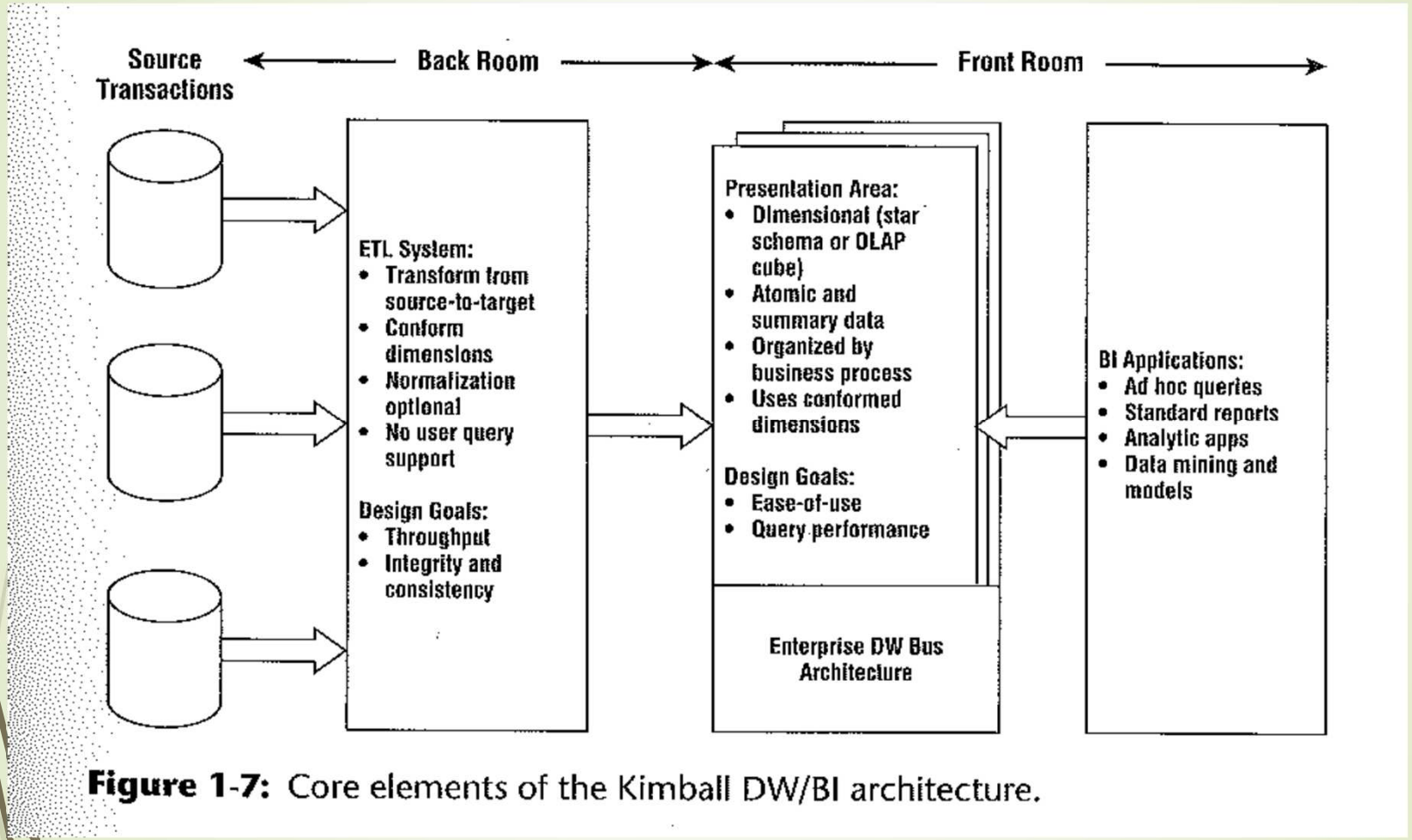


Figure 1-7: Core elements of the Kimball DW/BI architecture.

Figura tomada de [4].

Una arquitectura para DWH

- **Sistemas fuente operacionales:** son los que capturan las transacciones del negocio.
- **Sistema ETL (Extracción, Transformación y Carga):** es cualquier (sub) sistema entre los fuente y el área de presentación del DWH.
- **Área de presentación del DWH:** es dónde los datos son organizados y almacenados, quedando disponibles para las labores de análisis.
- **Aplicaciones de BI:** es la parte relacionada con los procesos y la presentación de los datos, los cuales sirven para hacer las decisiones analíticas.

Una arquitectura alterna para DWH

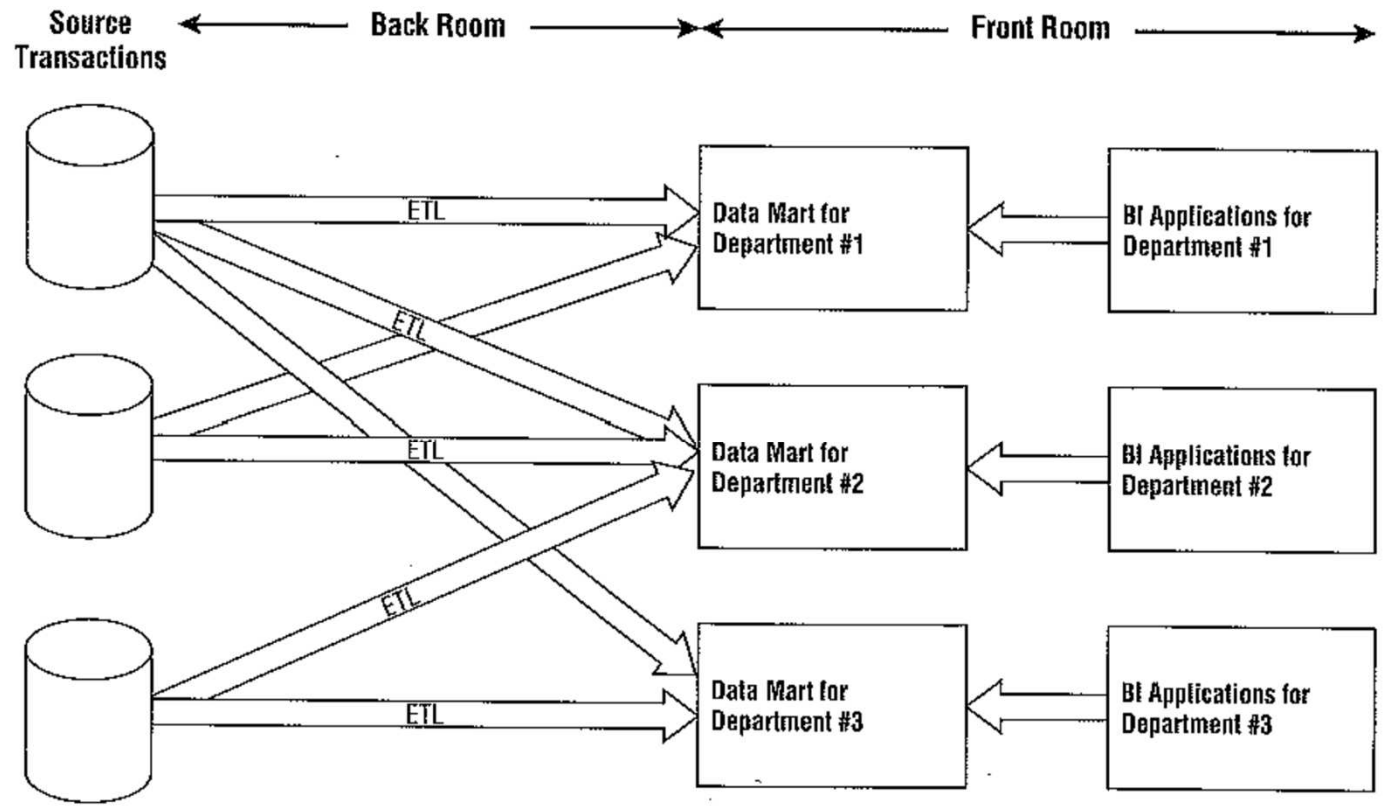


Figure 1-8: Simplified illustration of the independent data mart "architecture."

Figura tomada de [4].

Una arquitectura alterna para DWH

- **Arquitectura de Data Marts independientes**
  - En este enfoque los datos son desplegados tomando como base la organización departamental de las empresas.
  - El Data Mart satisface los requerimientos analíticos de un departamento.
  - Normalmente no se consideran aspectos de compartición e integración de información a través de la empresa.
  - Es muy difícil el integrar la información de la empresa con Data Marts independientes.

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 datos numéricos como: valores, medidas y ocurrencias.

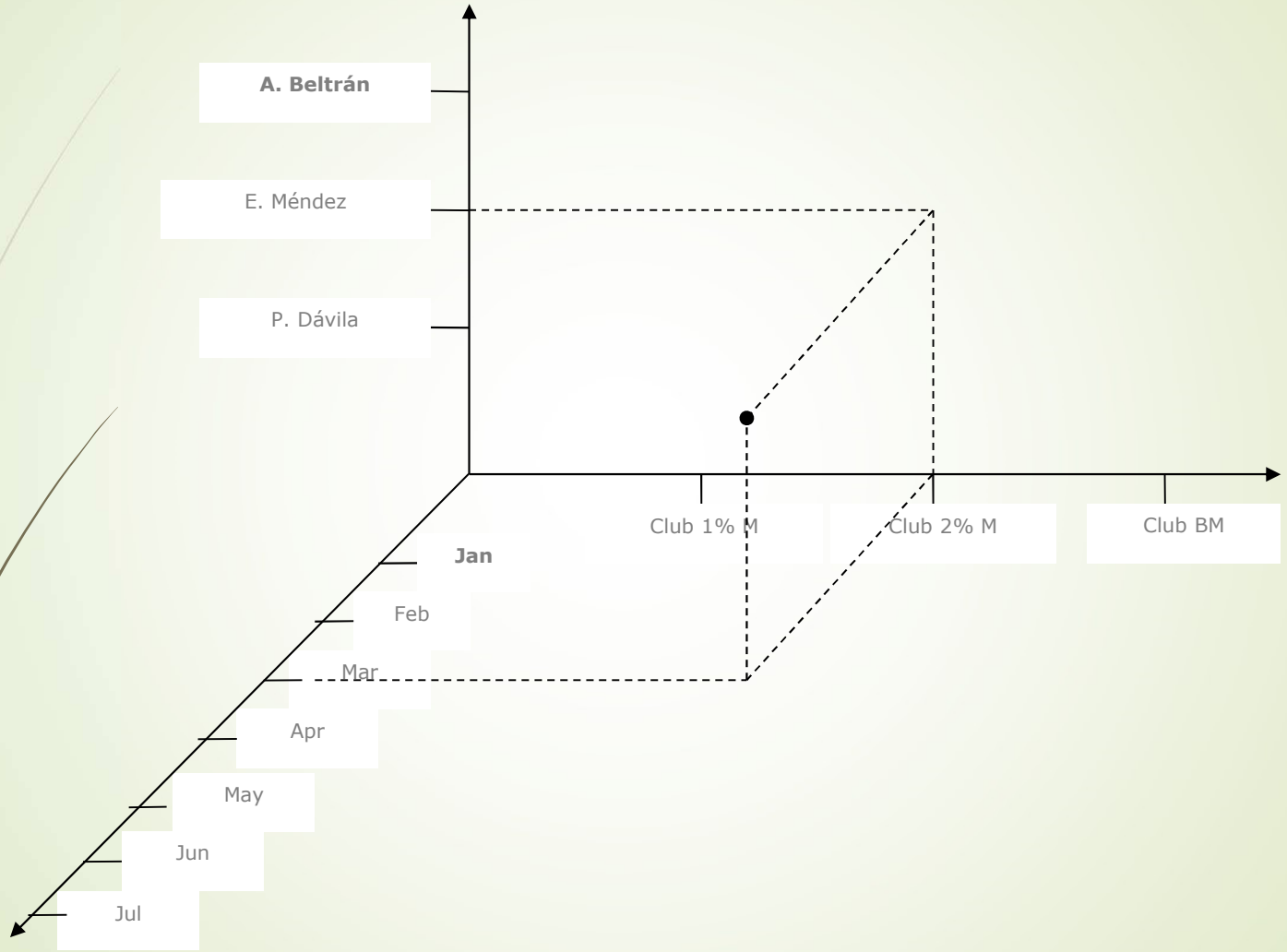
Espacio multidimensional

- Se usa el término *hipercubo* (o *cubo*, de manera abreviada) para describir un espacio de datos multidimensional.
- El cubo puede tener cualquier cantidad de dimensiones, cada una, posiblemente, con un tamaño distinto.
- El cubo contiene una cantidad discreta (esto es, no continua) de valores en cada dimensión.

Definiciones relacionadas con los cubos

- Una *dimensión* describe algún elemento en los datos que el negocio quiere analizar.
- Un *elemento (o miembro)* corresponde a un punto dentro de una dimensión.
- Un *atributo* es una colección completa de elementos.
- Una dimensión puede tener varios atributos; entre éstos, se escoge uno como *atributo clave*.
- El *tamaño*, o *cardinalidad*, de un atributo es la cantidad de elementos que contiene.

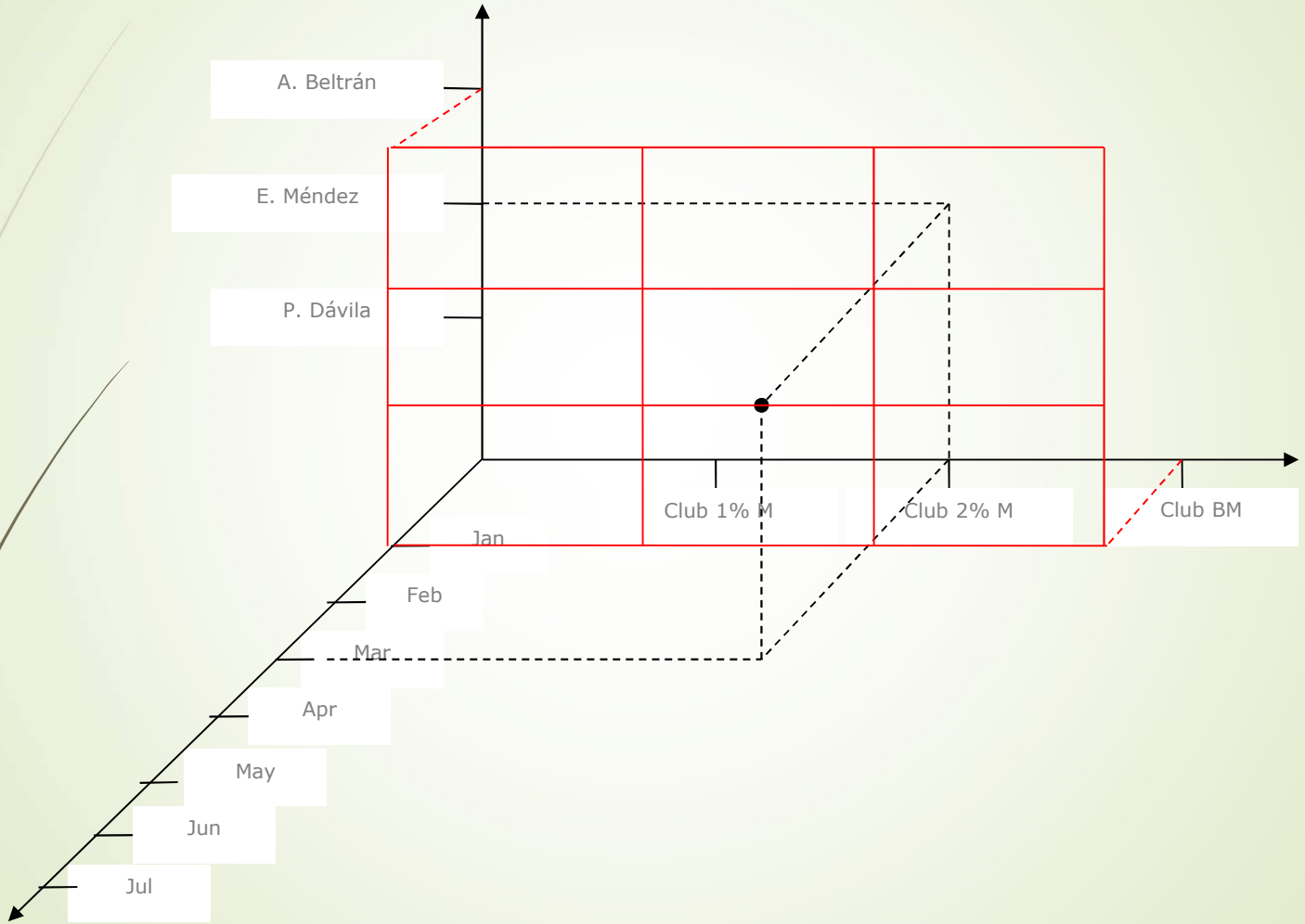
Ejemplo de un cubo (fig. 1)



Más definiciones sobre cubos

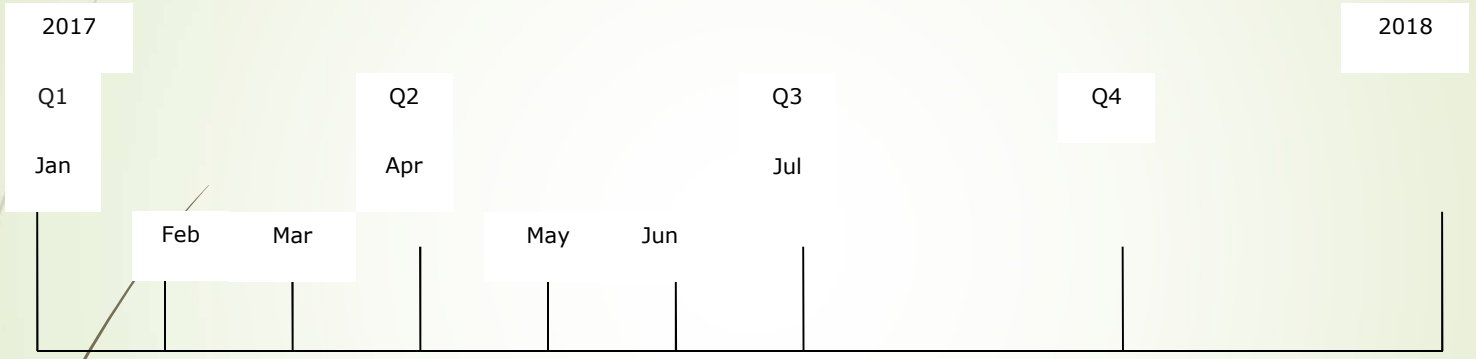
- Un *espacio de hechos*, *datos de hechos* o, simplemente, *hechos*, es el conjunto de puntos en el espacio de datos. Por ejemplo, una venta realizada sería un hecho.
- El tamaño máximo del espacio de hechos es la multiplicación del tamaño de cada dimensión.
- Una *tupla* es una coordenada en el espacio multidimensional.
- Una *rebanada* (del inglés, *slice*) es una sección del espacio multidimensional (sinónimo: subcubo).

Ejemplo de una rebanada (fig. 2)



Más definiciones sobre cubos

- Una *jerarquía* de una dimensión es un tipo de agrupación de sus atributos. Una jerarquía normalmente tiene varios niveles (fig. 3).

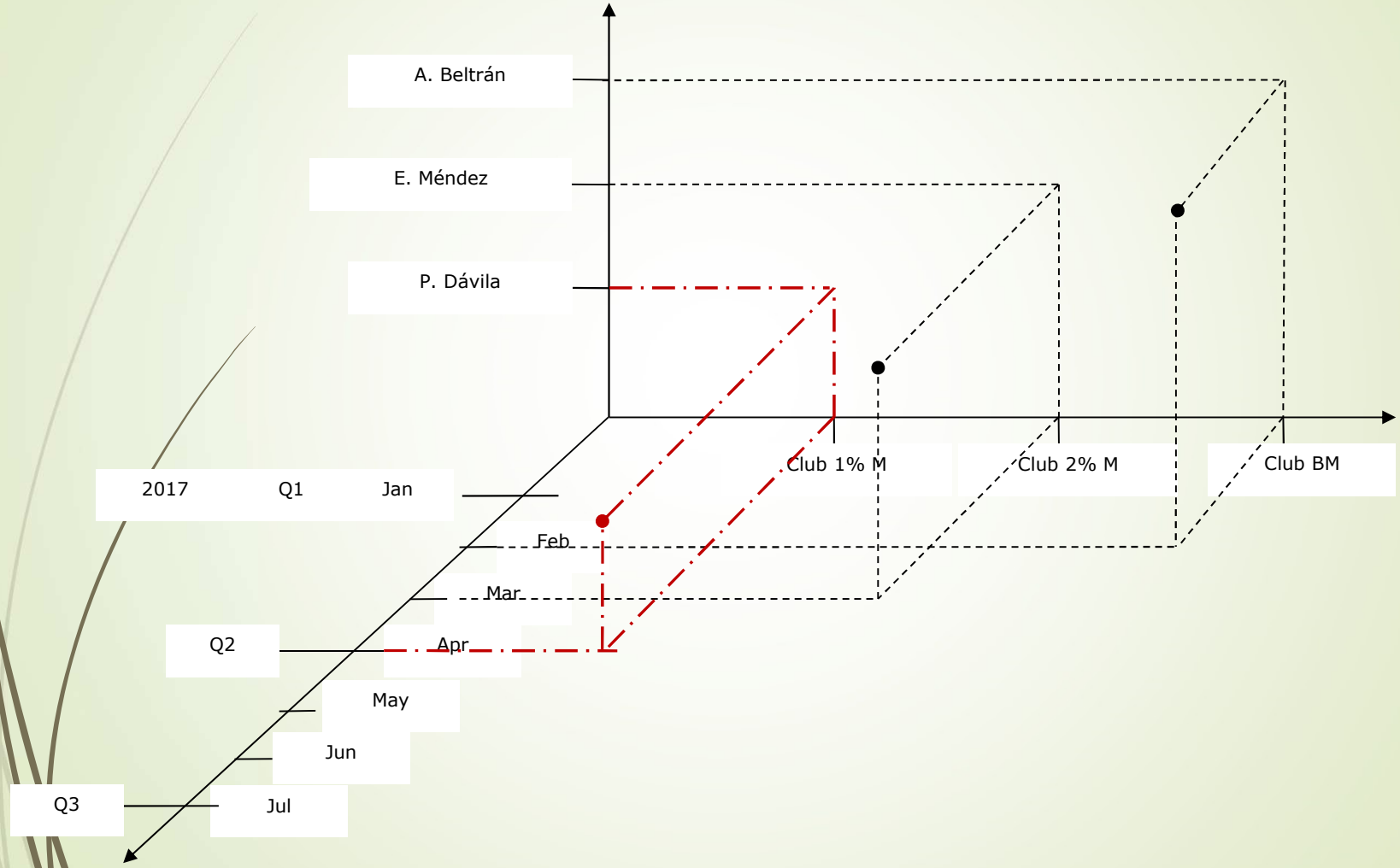


- Una dimensión puede tener más de una jerarquía, aunque todas usar el mismo atributo clave (por ejemplo, los días o los meses del año). Ejemplo (si el atributo clave fueran los días):
  - Jerarquía 1: años, trimestres, meses, días (cuatro niveles).
  - Jerarquía 2: años, semanas, días (tres niveles).

Más definiciones sobre cubos

- 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*.

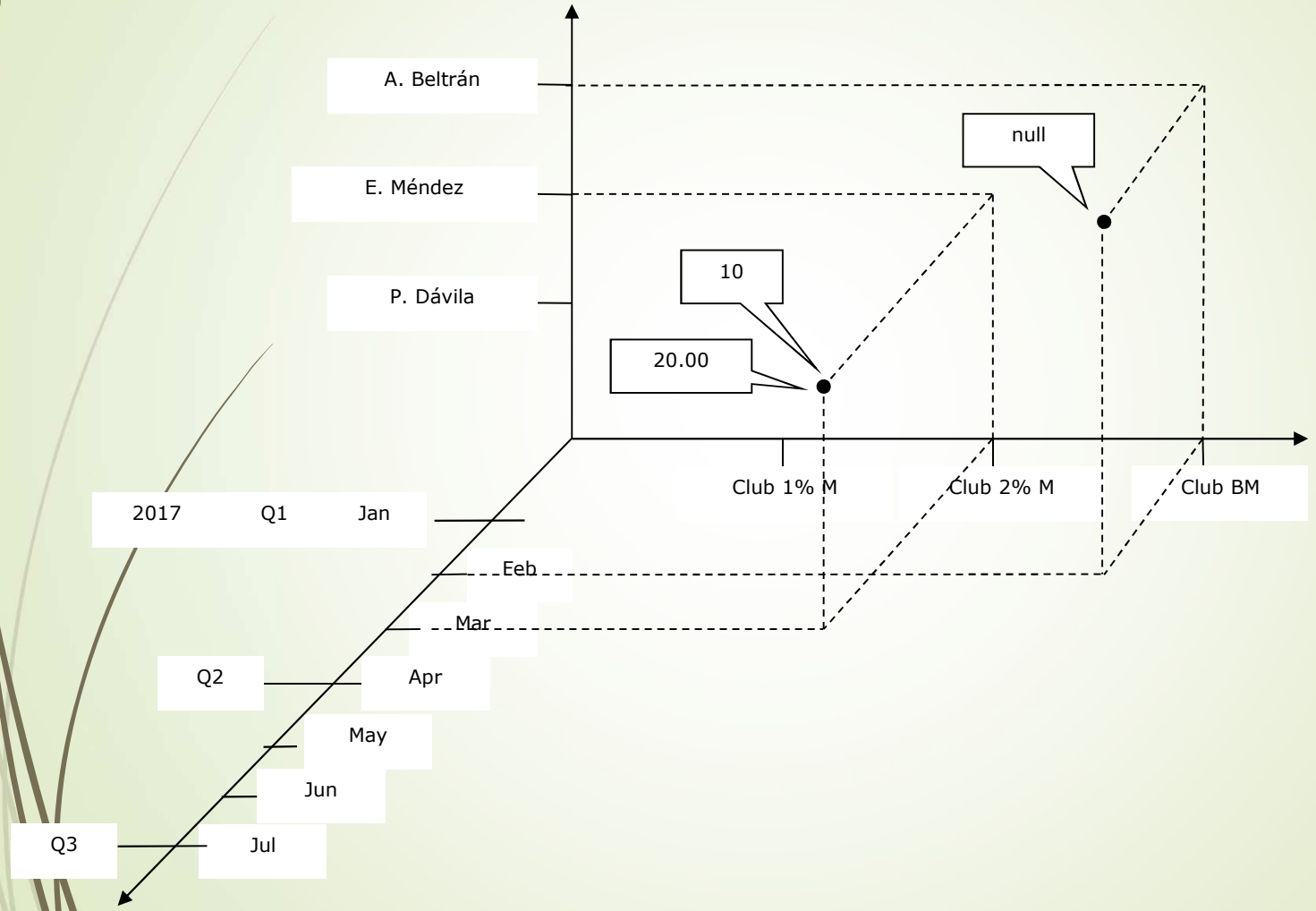
Espacio “completo” de datos (fig. 4)



Más definiciones sobre cubos

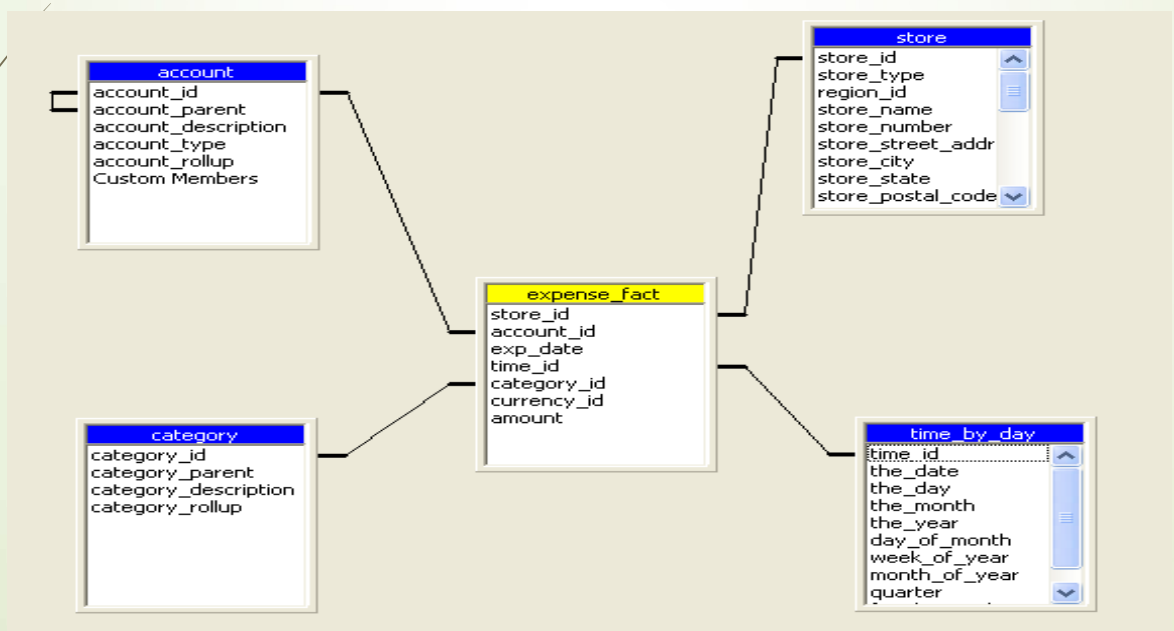
- Una *medida* es el valor de una celda. Una celda puede tener varias medidas, por ejemplo: el monto de una venta o la cantidad de unidades vendidas de un producto.
- Estas medidas pueden verse como una *dimensión de medidas*, cada medida con tipo de datos, unidad, etc.
- Las *funciones de agregación*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

Medidas de un celda de hechos (fig. 5)



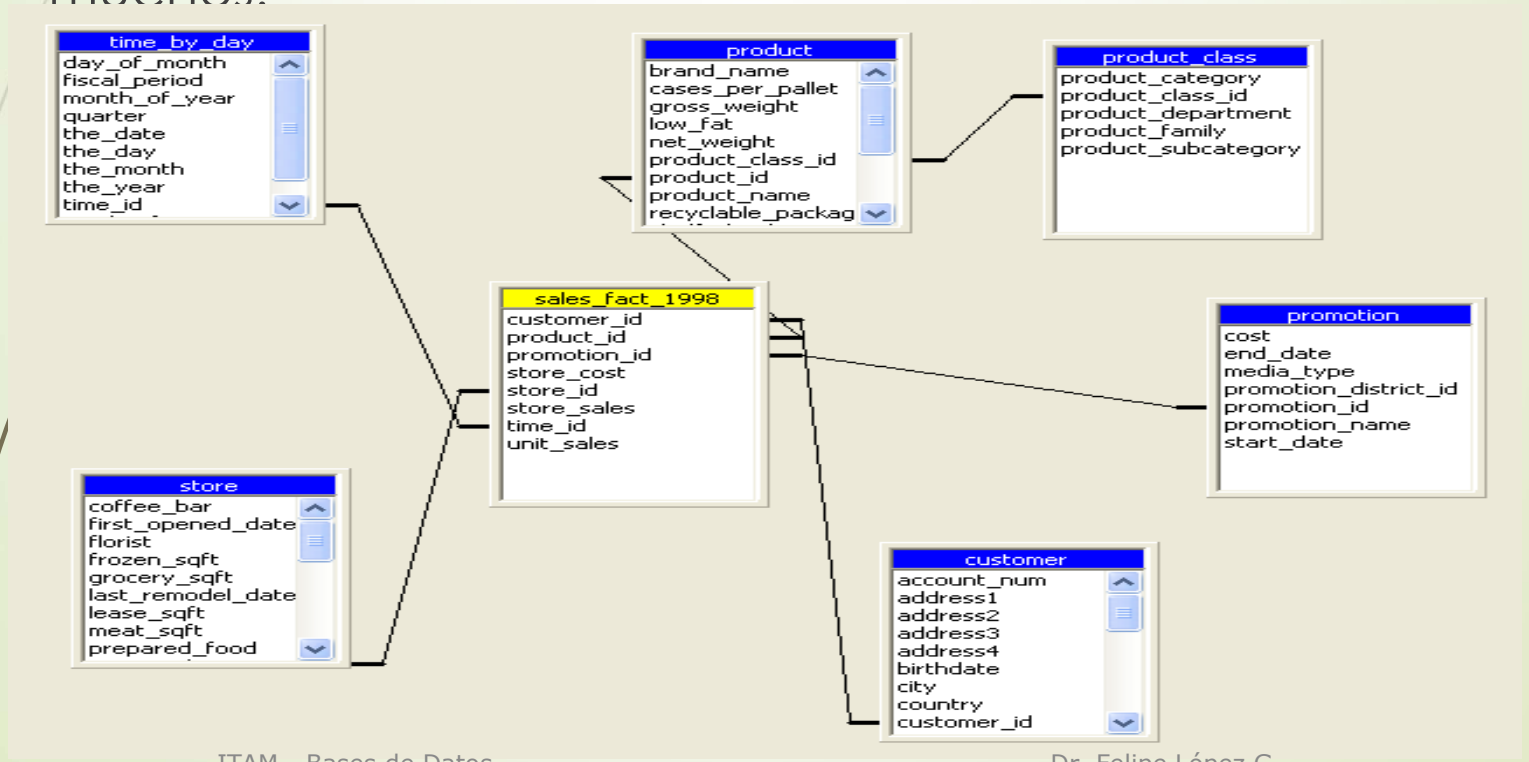
Modelos de diseño para los cubos

- Modelo de Estrella (*Star*).
  - Es la estructura básica para un cubo.
  - Se compone de una gran tabla central (llamada tabla de hechos) y de un conjunto de tablas más pequeñas (las tablas de las dimensiones) concentradas alrededor de la tabla de hechos.



Modelos de diseño para los cubos

- 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.



Operaciones básicas con los cubos

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