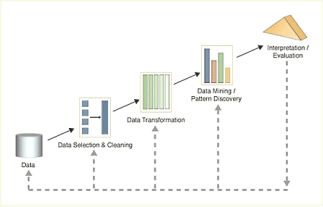
Universidad del Valle

# MODELO MULTIDIMENSIONAL

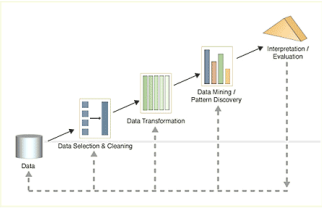
Fuente: The Data Warehouse Toolkit

Ralph Kimbal



# Modelo Dimensional

- Una técnica para **diseñar el modelo lógico** de la bodega de datos
- Permite **alto rendimiento** en el momento de acceder a los datos (orientado a consultas)
- Dimensional (orientado al negocio)
- Diferente del modelo entidad/relación



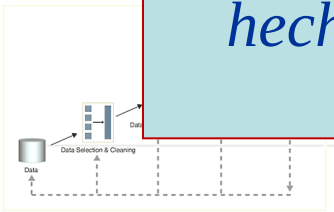
# Modelo Dimensional

- Tablas de hechos => Contienen *medidas numéricas* de los procesos del negocio.

- Tablas de dimensión => Contienen *atributos descriptivos* que proveen contexto para las mediciones almacenadas en los hechos .



*«Quién , cuándo, dónde, cómo sucedieron los hechos»*



# MODELO MULTIDIMENSIONAL

## Tablas de hechos

Almacenan las mediciones de un proceso del negocio. Los mejores hechos son **aditivos, numéricos**, evaluados continuamente.

- Todas las mediciones en una tabla de hechos deben estar en el mismo *nivel de granularidad*, contienen dos o más llaves foráneas.

# MODELO MULTIDIMENSIONAL

## Tablas de dimensión

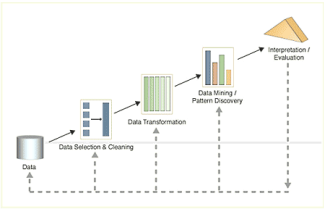
- Contienen los descriptores textuales de los atributos del proceso de negocio. Puede tener muchas columnas o atributos.
- Los mejores atributos de las dimensiones son *textuales y discretos.*

- Los atributos de las dimensiones servirán como restricciones en las consultas

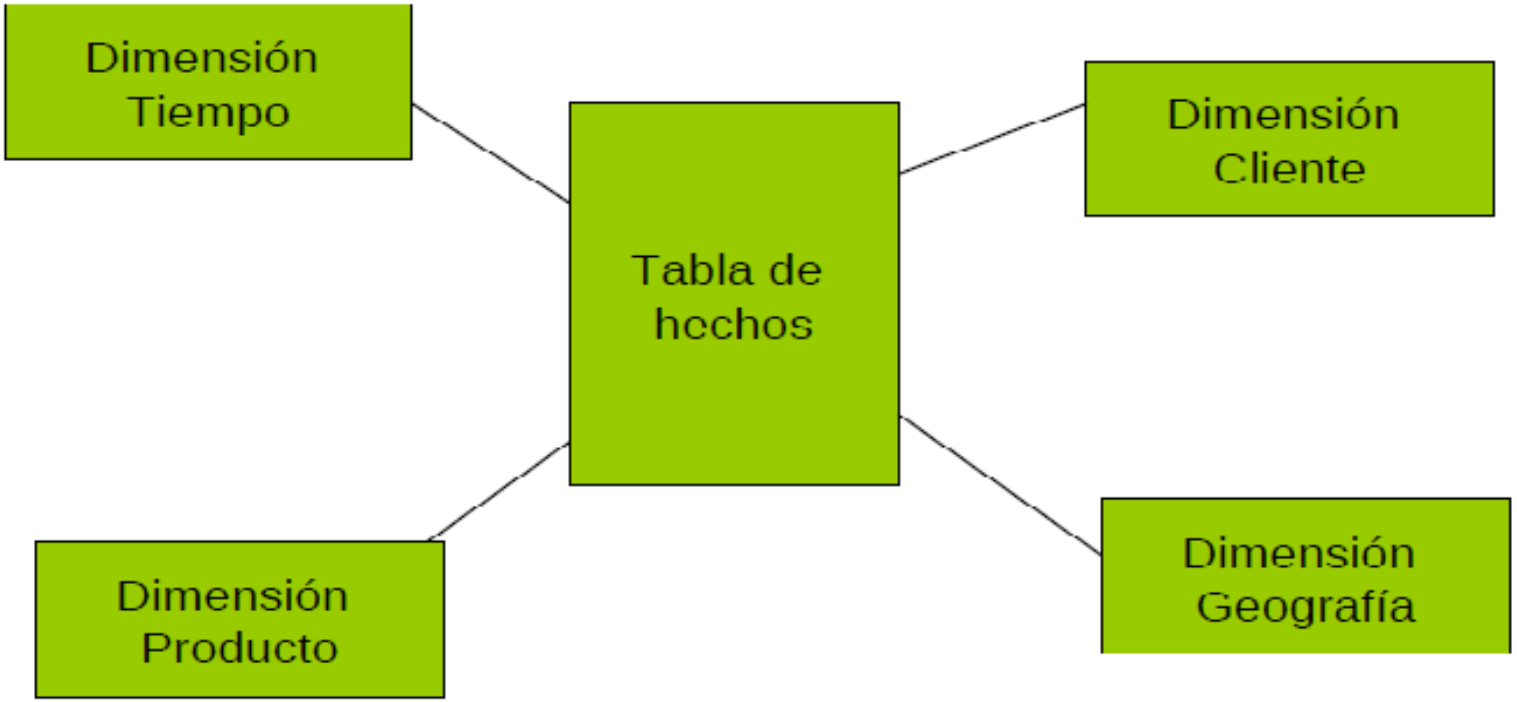
# MODELO DIMENSIONAL

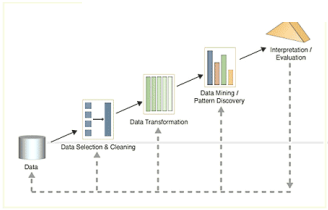
- “Un *hecho* toma muchos valores y cambia frecuentemente, un *atributo de dimensión* en cambio tiene valores más o menos constantes.

- Un *hecho* participa en cálculos, en cambio un *atributo de dimensión* es una restricción para una consulta”.

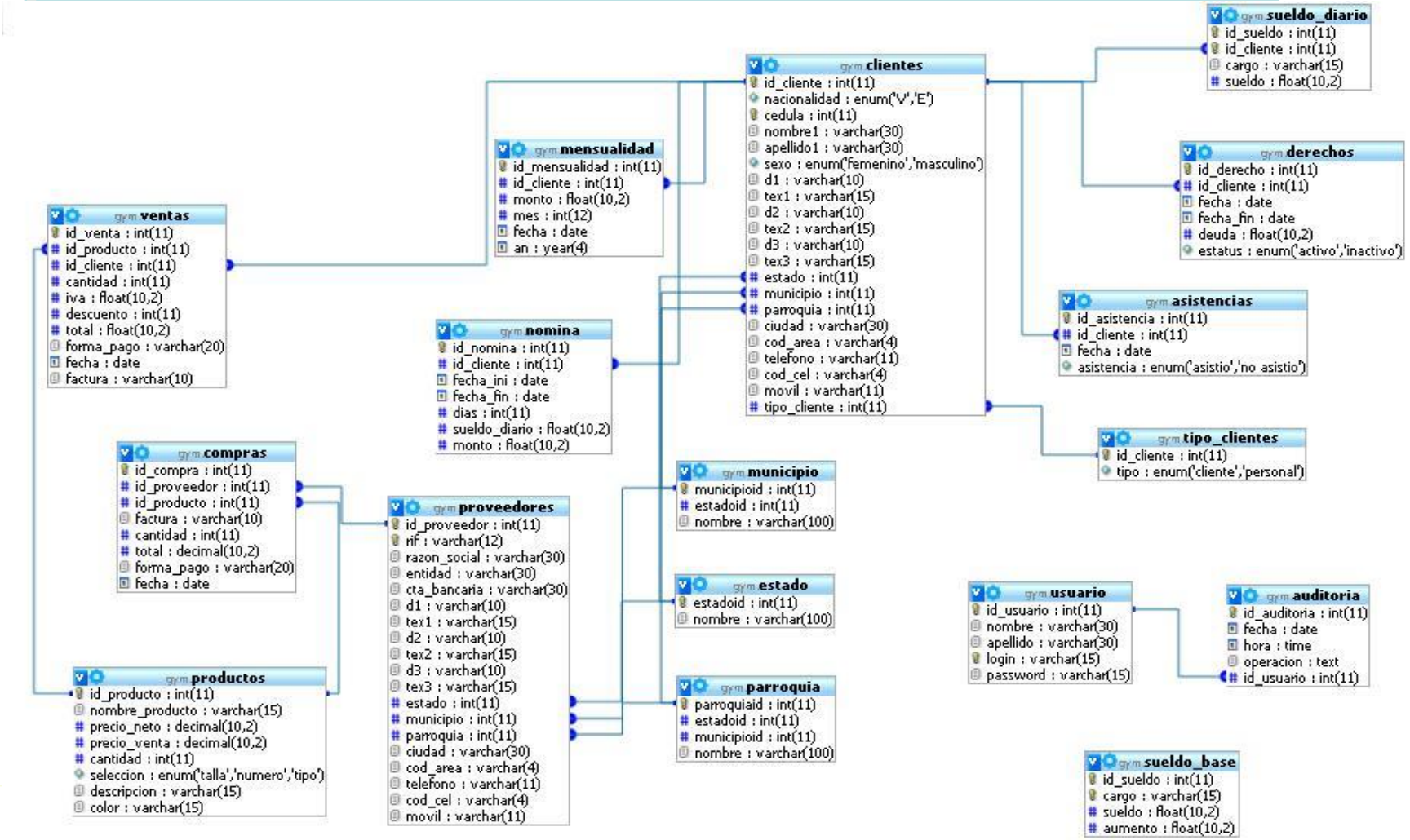


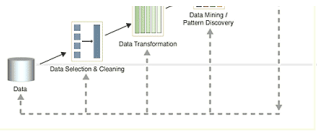
# MODELO MULTIDIMENSIONAL



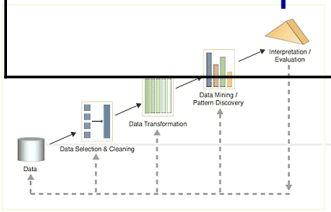
# MODELO RELACIONAL



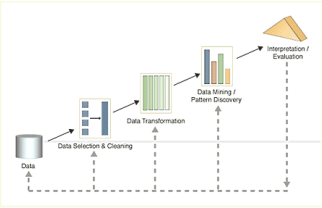
# Operacional Vs Dimensional

| Operacional | BI (Dimensional) |
|---|---|
| Enfocado a la actualización, elimina redundancia, muchas actualizaciones y repite el mismo tipo de operaciones diariamente | Enfoque a la consulta |
| Altamente normalizadas para soportar actualizaciones consistentes y mantenimiento de la integridad referencial | Altamente desnormalizada, se requiere disminución de tiempos en la obtención de grandes cantidades de datos |
| Tiempos de respuesta en segundos o inferior | Tiempos de respuesta aceptables pueden ser segundos, minutos, Horas. |
| Pocos datos agregados | Agregación: Varios niveles de datos precalculados |
| Almacenan pocos datos derivados | Gran cantidad de datos derivados . Redundancia |



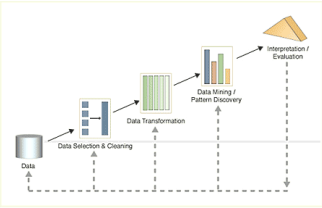
# Pasos proceso de diseño

1. Seleccionar un proceso de negocio a modelar
2. Escoger el nivel de granularidad del proceso
3. Seleccionar las dimensiones que se aplicarán a los hechos
4. Escoger los hechos medibles que poblarán cada tabla de hechos



# 1. Selección Proceso de Negocio a Modelar

- Proceso operacional importante
- Soportado en un sistema (legacy) fuente de datos
- Ej. Órdenes, facturación, envios (empresa)
- Ej. Matriculas, Modificaciones a la matricula



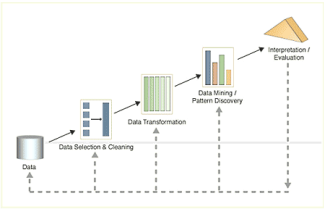
# 2. Escoger la granularidad del proceso

- **Nivel atómico** de los datos que representan el hecho en tabla de hechos

- **Nivel de detalle** que contienen las medidas almacenadas en la tabla de hechos

- **Impacto** en almacenamiento

  Ej. Transacciones individuales, diarias, mensuales, etc.



# 3. Seleccionar las dimensiones

- Formas de ver y analizar hechos
- El tiempo: dimensión estándar

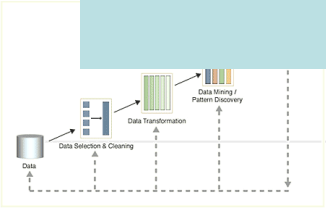
- **Descriptivas**

  Ej. Producto, almacén, bodega, tipo transacción, asignatura, cliente, estudiante, Programa

# 4. Seleccionar los hechos medibles

- Hechos que poblarán tabla de hechos
- Medidas de interés para el análisis
- Valor intersección de dimensiones
- Valor no conocido anticipadamente

***«La granularidad define el nivel de detalle de las medidas»***



# Ejemplo

1. Identificar el proceso a modelar
   - Clave: entender negocio y datos
   - Movimientos de items diario

2. Grano: SKU Xsupermercado X promoción X día
   - Determina tamaño de bd: movimiento diario de productos.

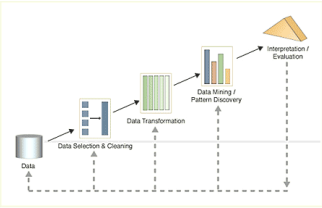
# Ejemplo

- Supermercado tiene 100 tiendas en todo el país, cada tienda o sucursal está dividida en departamentos: comestibles, lácteos, granos, congelados,  etc.

- Cada tienda tiene alrededor de 60.000  productos en su estantería. Cada producto se identifica mediante un código SKU.

- El supermercado maneja promociones  como reducciones temporales de precios, pague 1 y lleve 2,  manejo de cupones, etc.

- El supermercado necesita analizar el impacto que tienen estas promociones en el nivel de ventas y utilidades



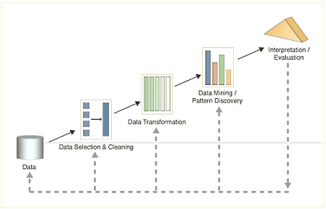
# Ejemplo

- Qué pasa si pensamos en semanas o meses ?
- Almacenar transacciones por cliente?
- Almacenar ventas por marca?
- Almacenar ventas por paquete?

# 1. Seleccionar  proceso del negocio

- Analizar que productos se están vendiendo, en qué almacenes, en qué días, y en que condiciones de promoción.

- Proceso del negocio: Ventas realizadas en el POS



# 2. Definir nivel de granularidad

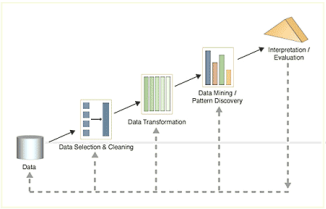
¿Cuál es el nivel de detalle de los datos?

Posibles análisis:

- Diferencia de ventas entre Lunes y Viernes
- Existencia de productos en ciertos almacenes.

Ej: cereales

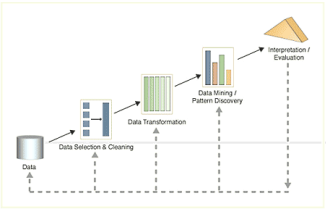
«Entender por qué ciertos compradores tomaron la promoción del “shampoo”»



# 2. Definir nivel de granularidad

- Si se elige el nivel  más bajo del grano, la mayoría atómica tiene sentido en varios frentes, los datos atómicos son  altamente dimensionales.

- Cuanto más detallado es  hecho más cosas se pueden conocer.  En este caso la granularidad es una transacción  individual en un punto de venta.



# 3. Elegir las dimensiones

- Fecha
- Producto
- Tienda
- Promoción

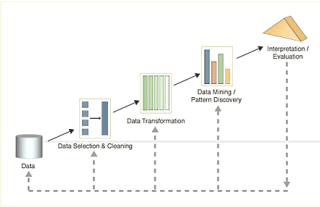
La dimensión promoción describe condiciones de promoción en que los productos son vendidos

# *Surrogate Keys (Claves suplentes)*

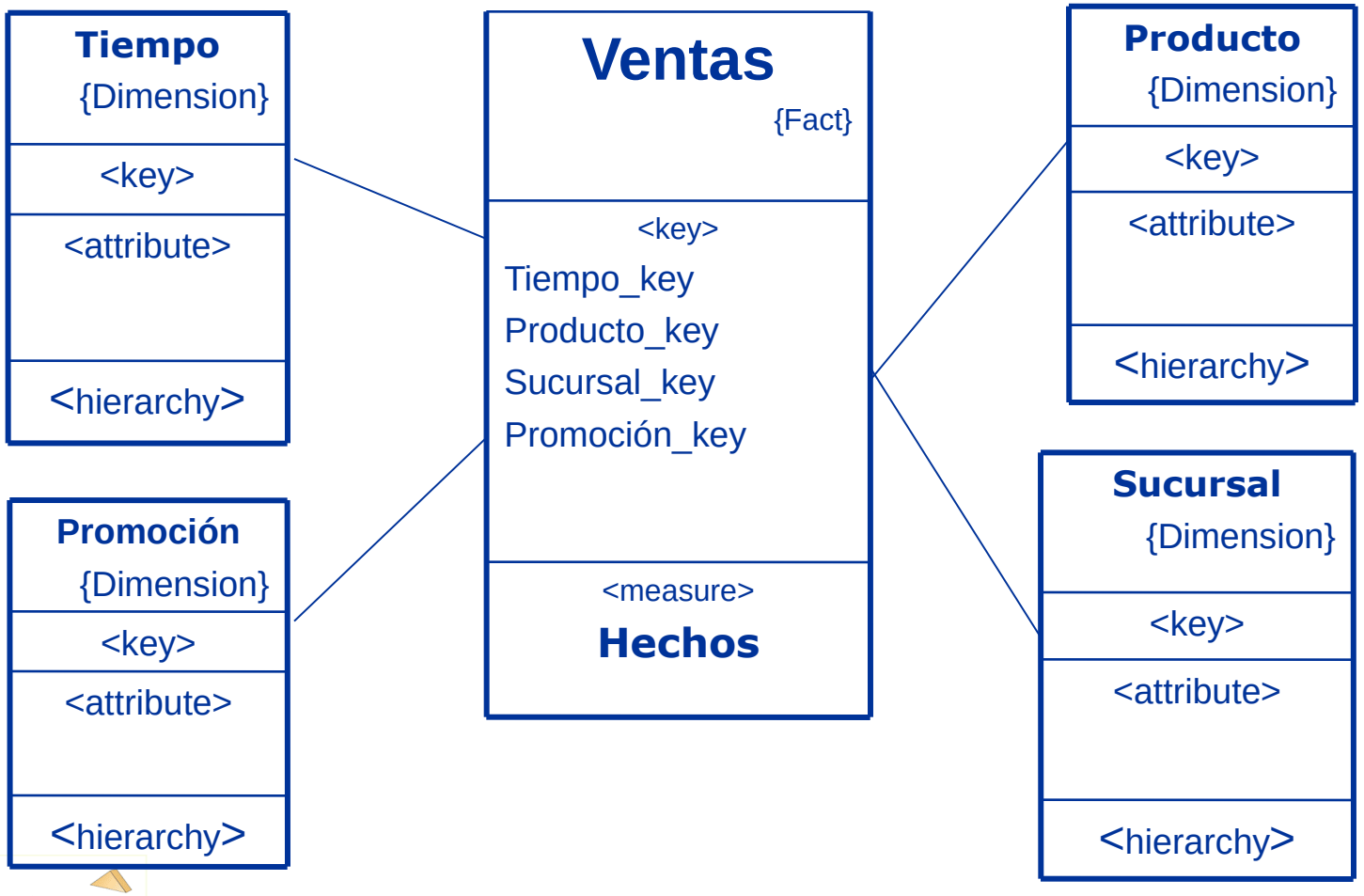
- Es una clave usada como un sustituto de una clave natural. Las claves artificiales por lo general toman valores numéricos enteros. Cada *join* entre tablas de dimensiones y tablas de hechos en un entorno de un *data warehouse* debe basarse en claves artificiales, no en claves naturales.

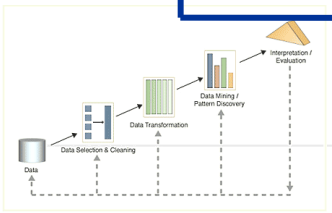
# *Surrogate Keys: Razones*

- Las tablas de datos de varios sistemas de origen OLTP pueden utilizar distintas claves para la misma entidad.
- Las claves suplentes proporcionan el medio para mantener la información del DW  cuando cambian las dimensiones.
- Las claves del sistema OLTP naturales pueden cambiar o ser reutilizadas en los sistemas de datos de origen.
- Las claves suplentes  pueden mejorar el rendimiento de las consultas al hacer las operaciones de *join.*



# Diagrama Inicial



# 4. Identificar los hechos

Medidas:

- Cantidad vendida
- Valor de la venta
- Costo de la venta
- Utilidad.  (Valor venta – Costo venta)

Otras medidas como el porcentaje de utilidad  no son aditivas.

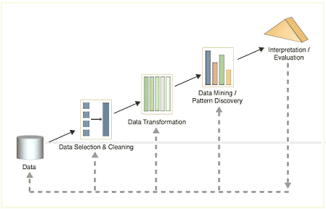
Los porcentajes y proporciones, como el margen bruto, son no aditiva. El numerador y denominador debe ser almacenado en la tabla de hechos y la relación entre estos se puede calcular

# 4. Identificar los hechos

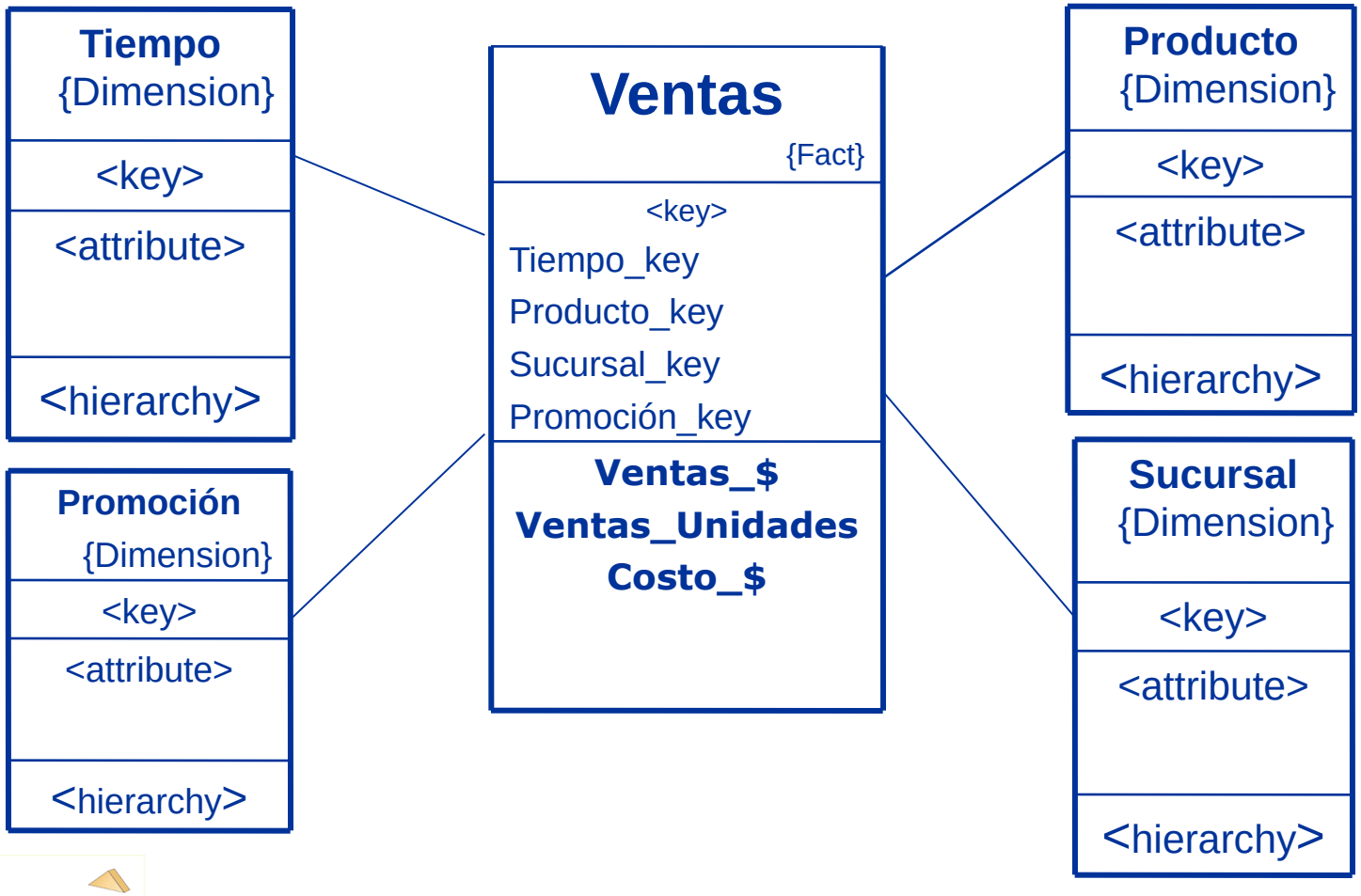
Estimar el número de filas que se almacenarán anualmente a la tabla de hechos.

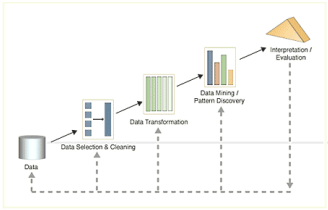
Estimar si es razonable.

Si dos o más hechos tienen un nivel de granularidad diferente se deben almacenar en tablas de hechos separadas.



# Diagrama Inicial



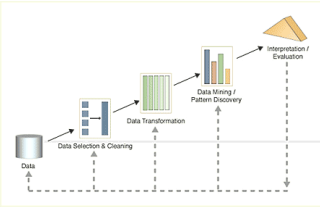



# La dimensión fecha

- La dimensión fecha es una de las más importantes ya que estará presente en casi todos los *data marts.*

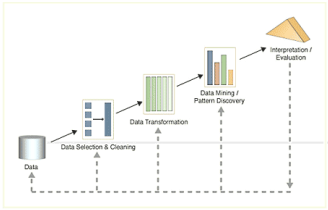
- Se puede construir la dimensión fecha previamente. Se pueden cargar 5 o 10 años en la tabla fecha

- Para almacenar los datos de un año se requerirán 365 filas en una tabla.

# La dimensión fecha

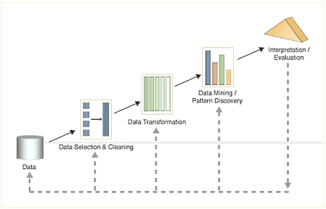
- Conocer el comportamiento de los hechos en ciertas franjas horarias?

# La dimensión fecha

- Conocer el comportamiento de los hechos en ciertas franjas horarias?
- Lo más recomendable es manejar el tiempo en una dimensión separada.

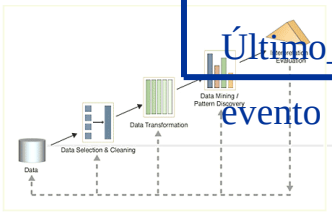

- Es preferible crear 365 filas para los días del año en una tabla y 1440 filas para representar los minutos de un día en otra, que una sola tabla con 525600 registros (365*1440)



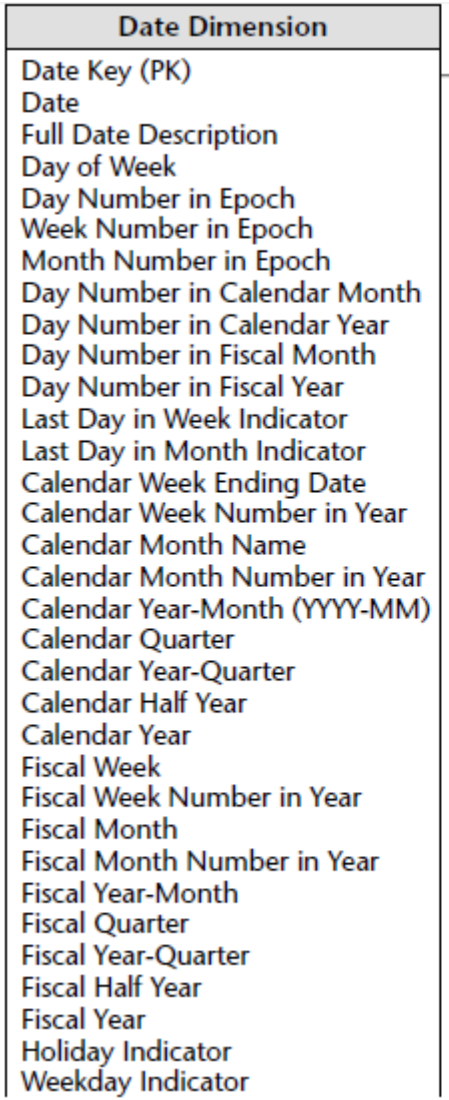
# La dimensión fecha

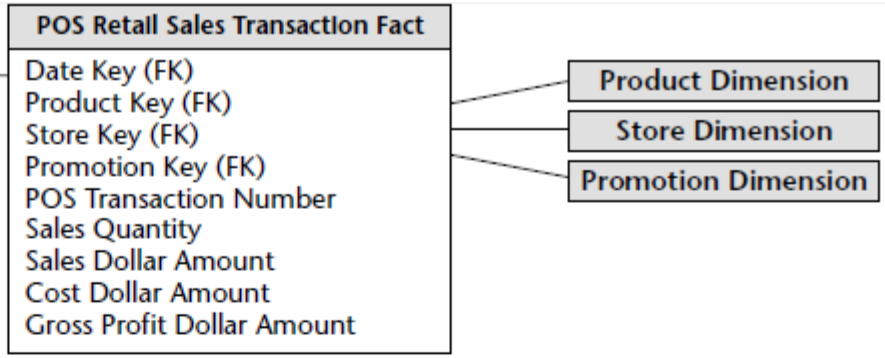


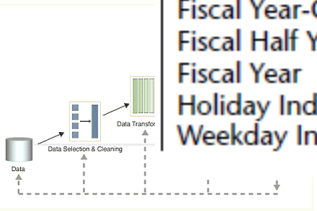


# La dimensión fecha



**Producto**

{Dimension}

Producto_key

SKU_descripción

SKU_número

Tamaño_paquete

Marca

Subcategoria

Categoria

Departamento

Tipo_dietetico

Peso

Unidad_de_medida

Unidades_por_caja

Altura_caja_empaque

**Ventas**

{Fact}

<key>

Tiempo_key

Producto_key

Sucursal_key

Promoción_key

**Ventas_$**

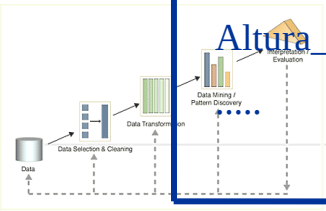
**Ventas_Unidades**

**Costo_$**

**Tiempo**

**Sucursal**

**Promoción**



**Sucursal**

{Dimension}

Sucursal_key

Nombre_sucursal

Número_sucursal

Dirección_sucursal

Ciudad_sucursal

Depto_sucursal

Región_sucursal

Teléfono_sucursal

Fax_sucursal

E-mail_sucursal

Fecha_apertura_inicial

Metros_2_almacen

Metros_2_sucursal

Metros_2_congelados....

**Ventas**

{Fact}

<key>

Tiempo_key

Producto_key

Sucursal_key

Promoción_key

**Ventas_$**

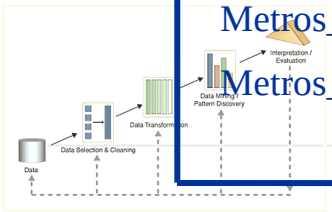
**Ventas_Unidades**

**Costo_$**

**Tiempo**

**Producto**

**Promoción**



**Promoción**

{Dimension}

Promoción_key

Nombre_promoción

Tipo_reducción_precio

Tipo_publicidad

Tipo_cupon

Costo_promoción

Fecha_inicio_promoción

Fecha_fin_promoción

Proveedor_publicidad

....

**Ventas**

{Fact}

<key>

Tiempo_key

Producto_key

Sucursal_key

Promoción_key

**Ventas_$**

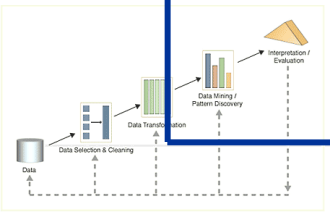
**Ventas_Unidades**

**Costo_$**

**Tiempo**

**Producto**

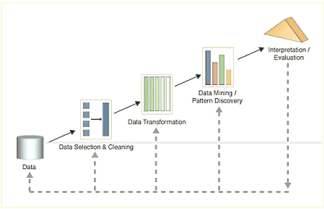
**Sucursal**



# Dimension degenerada

- En la tabla de hechos de ventas se almacena el número de transacción el cual se usa para identificar todos los productos que se compran juntos en una transacción.

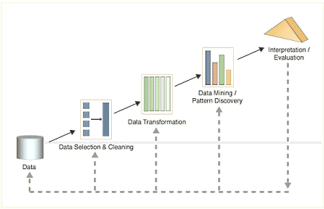
- Una clave de dimensión a la cual no le corresponde ninguna tabla de dimensión.



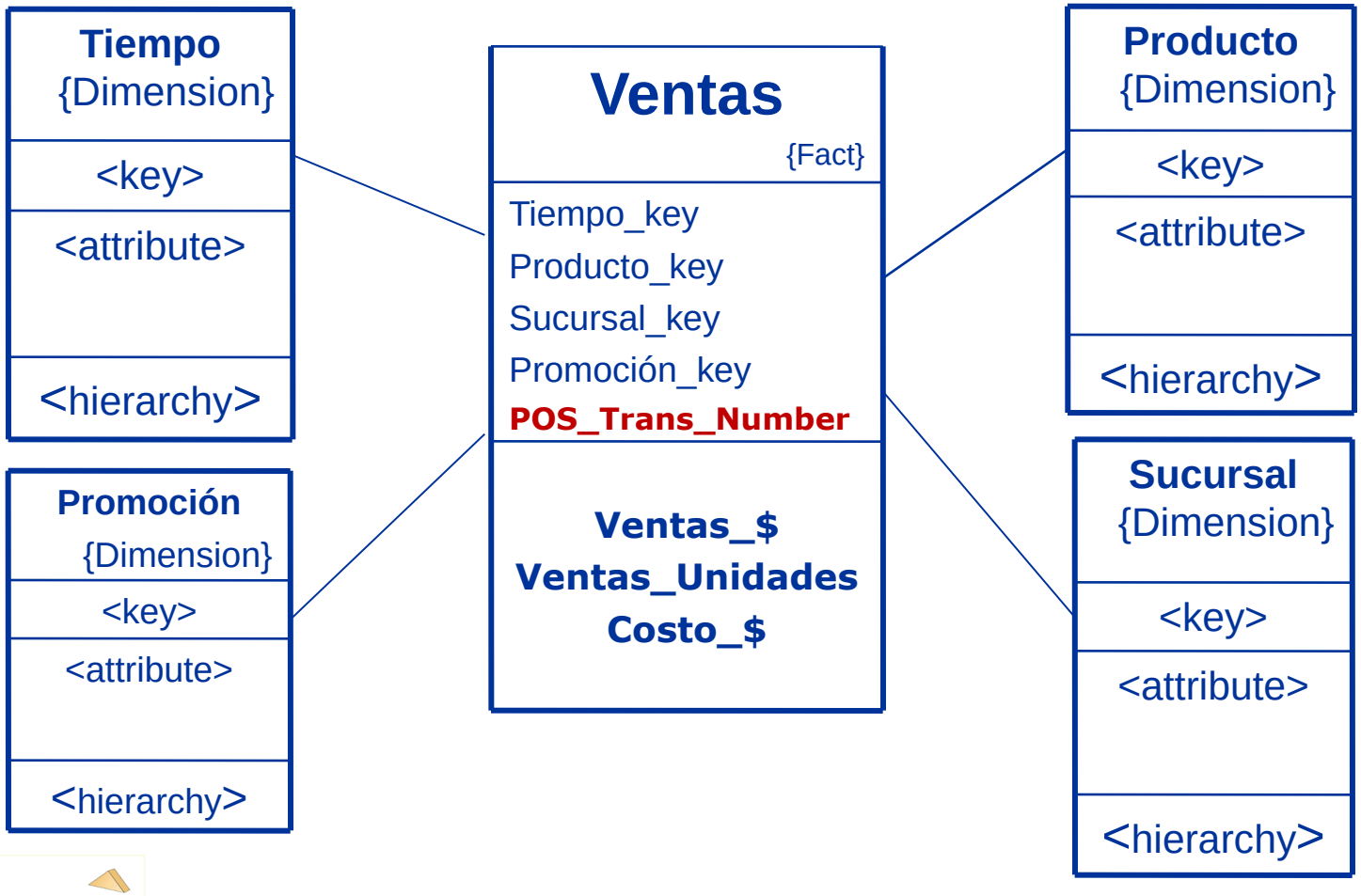
# Dimensión degenerada

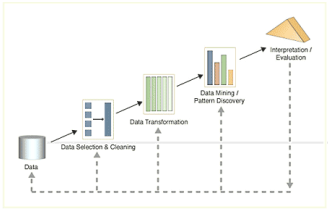
- Las dimensiones degeneradas aparecen cuando la granularidad de la tabla de hechos es el mismo documento que agrupa varios ítems en una transacción.

- Algunos atributos como números de pedido, número de factura y número de embarque suelen dar lugar a dimensiones vacías y se representan como dimensiones degeneradas.



# Dimension degenerada



# Consulta estándar

```
select p.marca, sum(h.pesos), sum(h.unidad)
from ventashecho h, producto p, tiempo t

where h.productokey = p.productokey
        and h.timekey   = t.timekey
        and t.semestre  = '1 S 1999'
group by p.marca
order by p.marca
```

⇐ select list

⇐ from clause con alias h,p y t

⇐ join constraint

⇐ join constraint

⇐ application constraint

⇐ group by clause

⇐ group by clause

# Bodegas de Datos

- Dos arquitecturas de acuerdo con la normalización de sus dimensiones:

**Estrella:**
Desnormalizado

**Copo de Nieve (*snowflake*)**
Normalizado

# Estrella Vs Copo de nieve

- **Estrella:** Desnormalizado

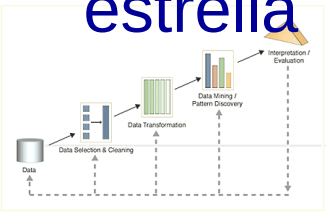
  Habilidad para análisis dimensional

- **Copo de nieve**

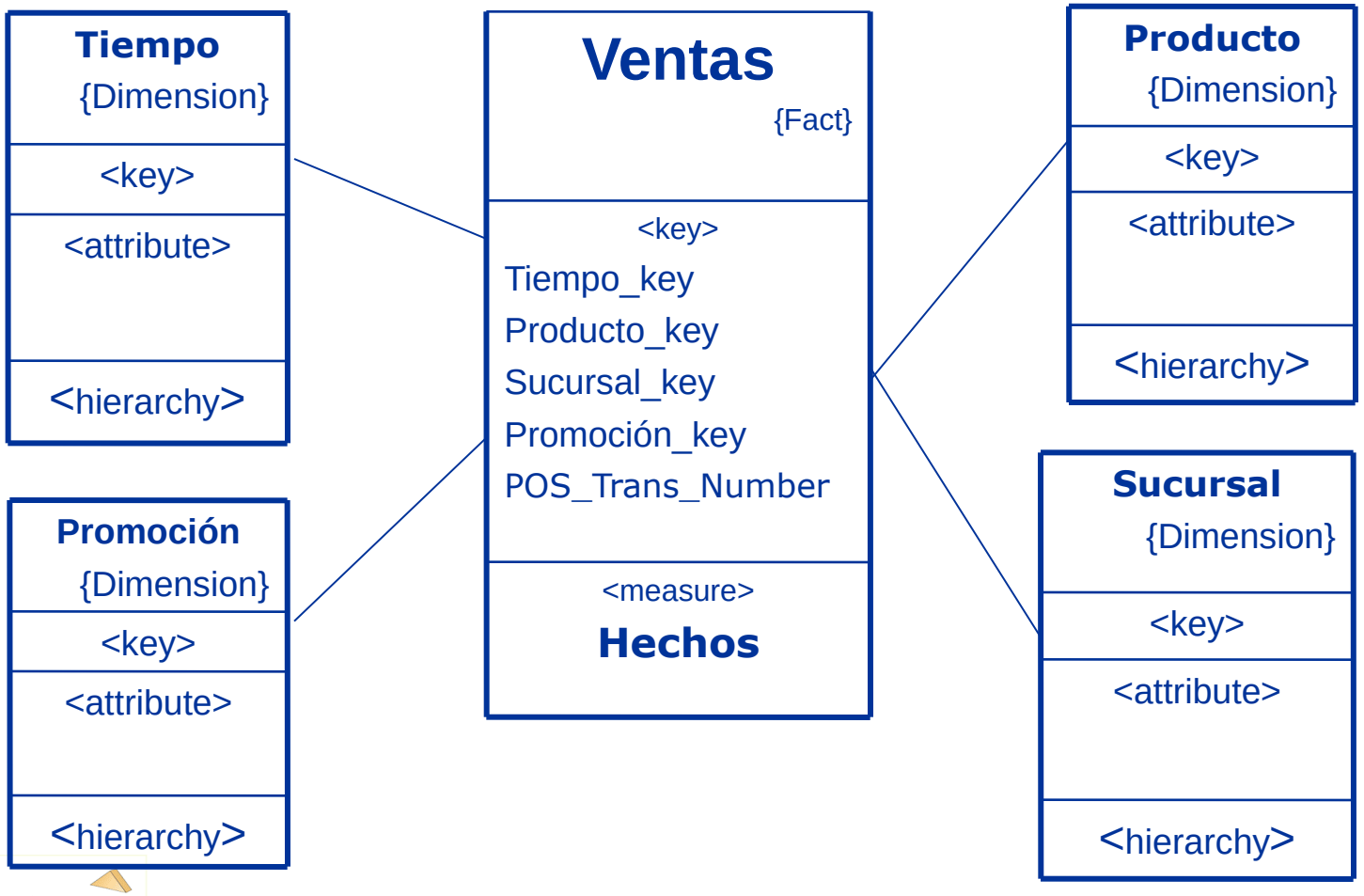
  Variación del modelo estrella

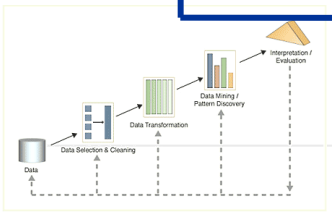
  Forma normalizada de las dimensiones (*solo las dimensiones primarias están enlazadas con la tabla de hechos*)

Se usa cuando no se puede implementar un modelo estrella



# Diagrama en Estrella



# Diagrama Copo de nieve

**Tiempo**
{Dimension}
<key>
<attribute>
<hierarchy>

**Promoción**
{Dimension}
<key>
<attribute>
<hierarchy>

**Ventas**
{Fact}
<key>
Tiempo_key
Producto_key
Sucursal_key
Promoción_key
POS_Trans_Number
<measure>
**Hechos**

**Producto**
{Dimension}
<key>
<attribute>
<hierarchy>

**Clase de producto**

**Categoría de producto**


Extensión del modelo dimensional

# Copo de Nieve

En el diseño del modelo se debe tener en cuenta: **uso** y **desempeño**

- Consideraciones de este modelo
- Múltiples tablas aumentan la complejidad de uso
- Mas tablas y joins afectan el desempeño de las consultas
- Navegar a través de las dimensiones puede ser más lento (cruce de dimensiones)