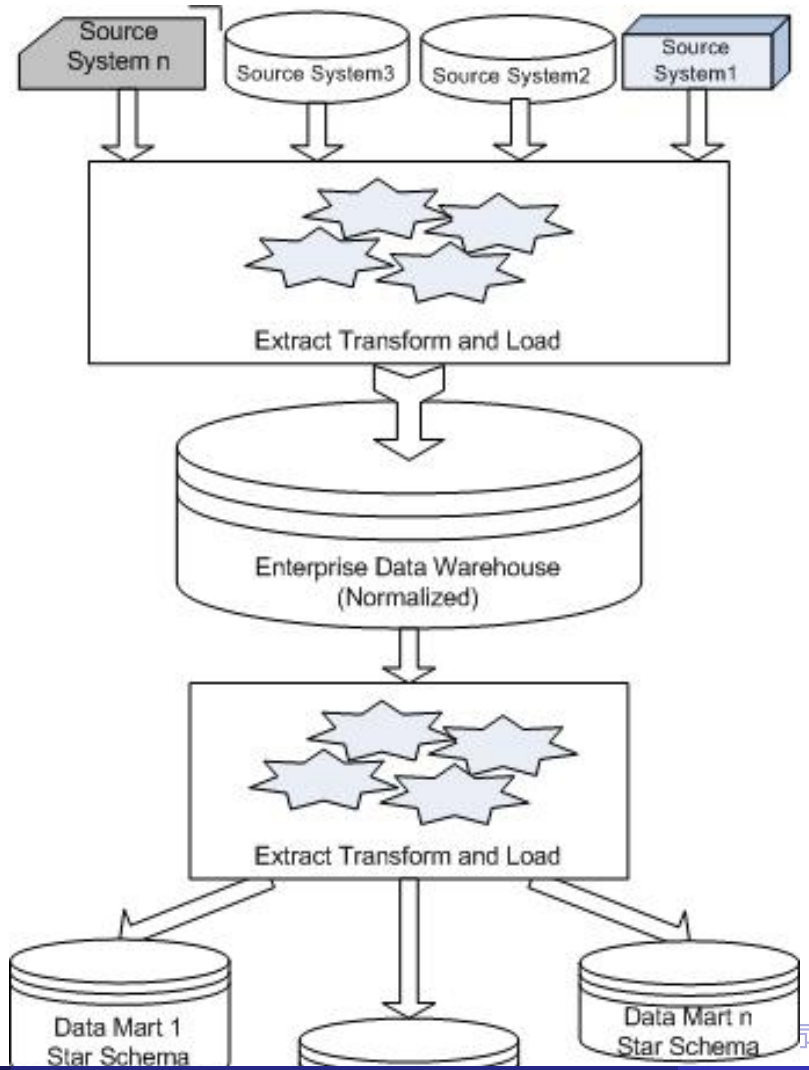
Proceso de diseño de un DWH

Dra. Amparo López Gaona

Fac. Ciencias, UNAM

Desarrollo de un almacén de datos

- Enfoque de arriba a abajo: Analiza las necesidades globales de la organización, y se planea el dwn como un todo.

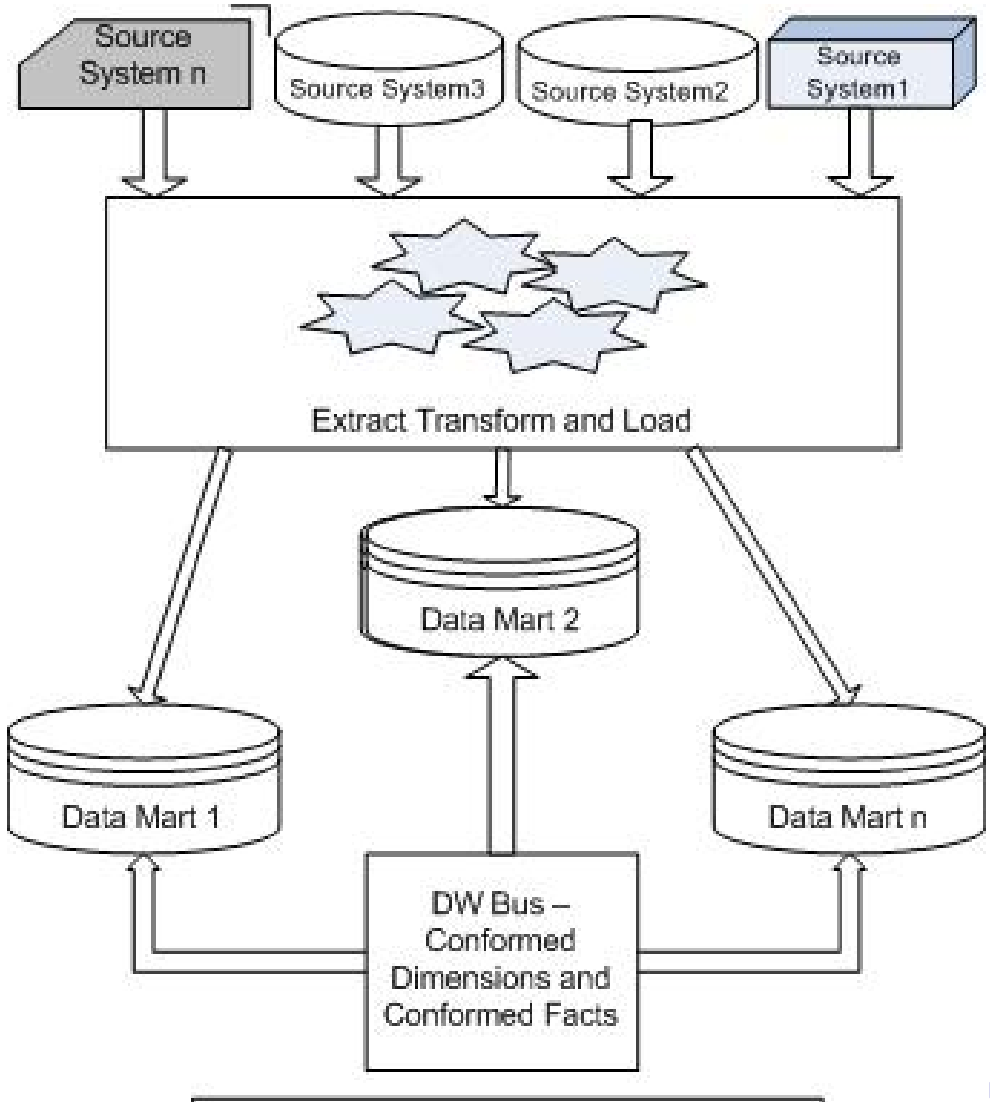


... Desarrollo de arriba a abajo

- Ventajas:
  - Se tiene el panorama general de la organzación.
  - La arquitectaura es integrada, no está formada como la unión de data marts dispares.
  - Es un sólo depósito central de datos consistente e integrado.
  - Se tienen reglas y control centralizados.
- Desventajas:
  - El alto costo estimado y largo tiempo de implementación desanima a los administradores de la compañía.
  - Analizar e integrar **todas** las fuentes relevantes es tarea muy difícil.
  - Es extremadamente dificil separar las necesidades específicas de cada departamento involucrado en el proyecto,
  - Debido a que no hay un sistema trabajando a corto plazo, los usuarios no pueden verificar si el proyecto es útil.
- En pocas palabras, el riesgo de fracasar es alto.

... Desarrollo de un DWH de abajo-arriba

- El dwh se construye de manera incremental creando varios datamarts.



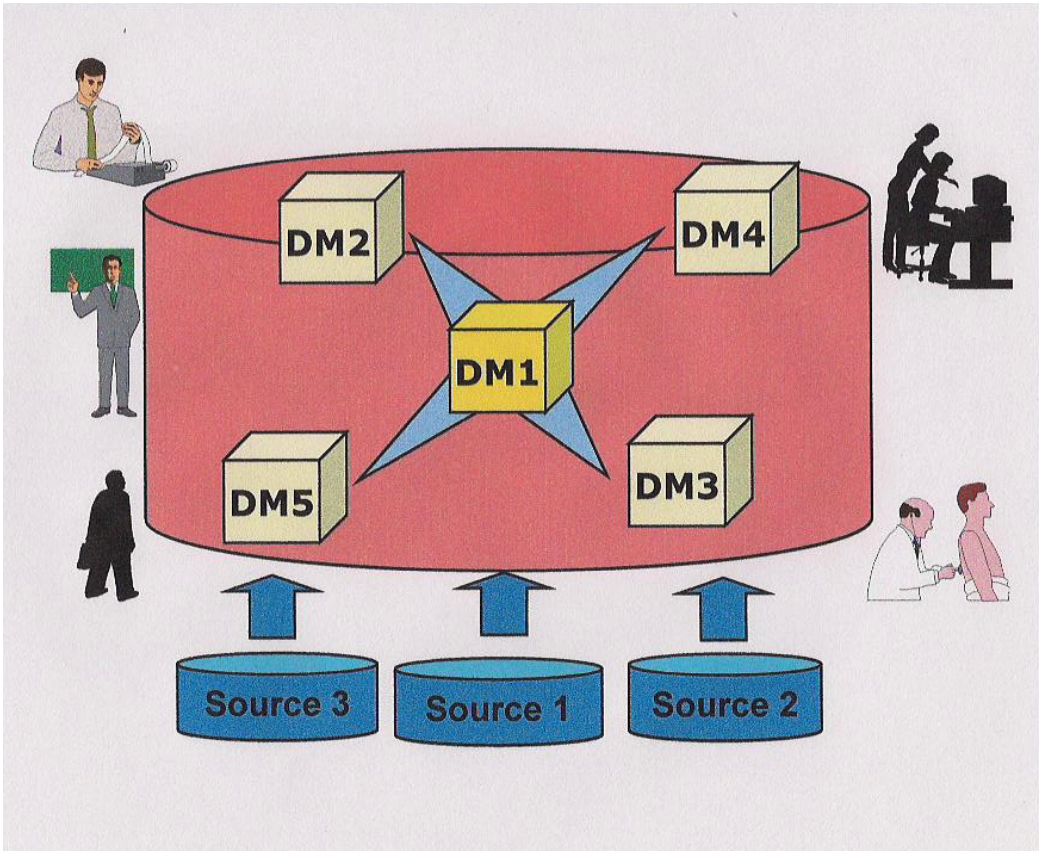
... Desarrollo de un DWH de abajo-arriba

- Ventajas:
  - Se ven resutlados a corto plazo, así que puede ser de interés para la organización.
  - No requiere grandes inversiones.
  - Permite a los diseñadores concentrarse en un área a la vez.
  - Proporciona a los ejecutivos una pronta retroalimentación acerca de los beneficios reales del sistema que se está construyendo.
  - Es inherentemente incremental, por lo tanto puede priorizarse los data marts de acuerdo a su importancia.
  - Menor riesgo de falla.
- Desventajas:
  - Cada datamart tiene su propia vista de datos.
  - Puede haber datos redundantes entre los datamart.

... Desarrollo de abajo a arriba

El primer datamart :

- Es el que juega el papel estratégico en la organización.
- Debería ser la columna vertebral para el DWH completo.
- Debería apoyarse en fuentes de datos disponibles y consistentes.



... Desarrollo de un almacén de datos

- Planeación del proyecto
- Definición de requerimientos
- Diseño
- Construcción
- Utilización/uso
- Mantenimiento

Planeación del proyecto

En esta etapa se determina el propósito del proyecto, objetivos específicos, el alcance, principales riesgos y una aproximación inicial a las necesidades de información.
Esta tarea incluye las siguientes acciones típicas de un plan de proyecto:

- Definir el alcance (entender los requerimientos del negocio).
- Identificar las tareas
- Programar las tareas
- Planificar el uso de los recursos.
- Asignar la carga de trabajo a los recursos
- Elaboración de un documento final que representa un plan del proyecto.

Requerimientos

- Levantamiento de requerimientos en un sistema operacional. Entrevistas con los usuarios.
  - Usuarios listan las funciones necesarias/deseadas del sistema.
- Para un dwh, los usuarios, generalmente, son incapaces de definir claramente sus requerimientos.
- Si la definición de requerimientos para un dwh no es clara, el análisis para el proceso es imposible.
- ¿Cómo construir algo que los usuarios son incapaces de definir claramente?
- Los usuarios pueden no ser capaces de definir lo que quieren en el dwh pero sí pueden dar información de cómo piensan acerca de su negocio.
  - Unidades de medidas importantes para el éxito de su organización
  - Cómo combinan varias piezas de información para la toma de decisiones estratégicas.

... Requerimientos

Ejemplos:

- Ganancias generadas por un nuevo producto.

  mes por mes, en la división sur, por región, por sucursal, con respecto a la versión anterior, etc.

- Estadísticas de sus ventas.

  por productos, resumidas por categoría de productos, diarias, semanales y mensuales, por ventas en cada estado, por canales de distribución.

- Conocer los gastos.

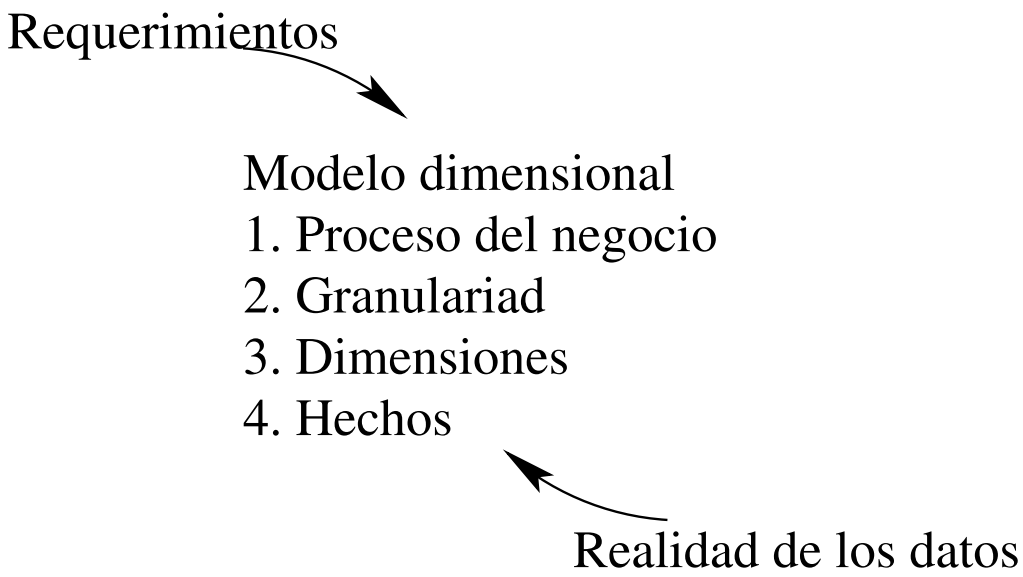
  por mes, trimestre, año, por sucursal, por zona geográfica, comparada con otros periodos de tiempo,o por tipos de artículos, resumida para toda la compañía,

Si los usuarios piensan en términos de dimensiones para tomar sus decisiones, el diseñador del dwh debe pensar en esos términos al recabar los requerimientos.

Diseño dimensional

Kimball propone 4 pasos para el diseño dimensional:

1. Seleccionar el proceso a modelar.
2. Definir la granularidad del proceso.
3. Elegir las dimensiones.
4. Identificar los hechos.



... Diseño dimensional (Selección del proceso)

- Proceso = actividad desarrollada por/en la organización.
  - Ejemplos: tomar ordenes de compra, facturar, registrar alumnos, realizar estudios médicos, procesar reclamaciones, etc.
- Proceso $\neq$ departamento en la organización.
  - Ejemplo: crear un modelo dimensional para manejar información de ordenes de compra, y no construir modelos separados para el departamento de ventas y otro para el de mercadotecnia.
- Crear el almacén enfocándose en un proceso:
  - Ayuda a evitar redundancia.
- Características que ayudan a identificar un proceso:
  - Frecuentemente se expresan con verbos.
  - Son soportados por sistemas operacionales, como el sistema de facturación o el de compras.
  - Generan o capturan las métricas clave.

... Diseño dimensional (Granularidad)

- La granularidad es el nivel de detalle asociado con las medidas en la tabla de hechos.
  - Dado por la combinación de niveles bajos en las jerarquías de dimensiones.
  - Ejemplos: total de ventas por tienda por día por producto.
- Lo común es que se refiera a una transacción del negocio.
  - Ejemplo: una venta.
  - En ocasiones el dato es agregado (total de ventas por día)
- A mayor nivel de detalle, mayores posibilidades analíticas.
  - Los datos con granularidad fina podrán ser resumidos hasta obtener una granularidad media o gruesa.
  - No sucede lo mismo en sentido contrario.
  - Ejemplo: "Rendimiento de un empleado" tiene nivel de granularidad alto, sin embargo "Rendimiento diario de un empleado", puede considerarse de granularidad baja.

... Diseño dimensional (Granularidad)

- Determinada por la realidad física del sistema operacional que captura los eventos del proceso de negocio.
- Generalmente, está dada por una transacción del negocio.
  - Una línea en una factura.
  - Una línea en un pase de abordar.
  - Una línea en un estado de cuenta bancario mensual.
  - Una línena de un reporte diario de los niveles de inventario por cada producto en un almacén.
  - Total de artículos en promoción en cada almacén por día.
- La granularidad ayuda a determinar:
  - las dimensiones que deben incluirse.
  - las jerarquías dentro de cada dimensión.

... Diseño dimensional (Dimensiones)

- Responden la pregunta "¿Cómo describen los hombres de negocio los datos que resultan del proceso de negocio?"
- Las dimensiones describen el contexto para analizar los hechos.
- Si se definió bien la granularidad, entonces este paso es sencillo.
- Las dimensiones se usan para:
  - Seleccionar datos.
  - Agrupar datos en el nivel de detalle deseado.
- Ejemplos de dimensiones: Fecha, Producto, Cliente, Curso, etc.
- Al elegir las dimensiones se debe incluir una lista de sus atributos.
- Los atributos de una dimensión pueden tener un orden jerárquico.
  - Típicamente entre 3 y 6 niveles de detalles.
  - En una dimensión puede haber más de una jerarquía.
  - Puede haber dimensiones sin jerarquía de atributos: Ciudades.
- Regla general: las dimensiones deberían contener mucha información.
- Las dimensiones tienen valores. (generalmente alfanuméricos)
  - La dimensión producto tiene valores "leche","crema".
  - La dimensión fecha "10/10/2016".
- Las tablas de dimensiones son más pequeñas que las de hechos

... Diseño dimensional (Hechos)

- Responden a la pregunta "¿Qué se desea medir en el proceso?"
- Los usuarios (ejecutivos) están interesados en analizar el rendimiento de su negocio de acuerdo a esas medidas.
- Un hecho es identificado por los valores de sus dimensiones:
  - Un hecho es una celda no vacía en el cubo.
  - Ejemplos: cantidades ordenadas, cantidad vendida en pesos, etc.
- Los hechos que pertenecen a diferentes granularidades deben estar en tablas separadas.
- Los hechos pueden ser:
  - Atómicos. Por ejemplo: cantidad vendida, precio, etc.
  - Derivados. Utilizan una fórmula para calcularlos.
    Por ejemplo: PrecioTotal = precio * cantidad_vendida

... Diseño dimensional (Hechos)

- Las medidas en una tabla de hechos caen en tres categorías:
  - Aditivas. Las más flexibles y útiles, son aquellas que se calculan sumando a lo largo de cualquier dimensión asociada con la tabla de hechos. (`COUNT, SUM,` etc.)
    - Ejem. cantidades vendidas, ganancia calculada de ventas y costo.
  - Semi-aditivas. Se obtienen sumando a lo largo de algunas dimensiones, pero no de todas. (`Last, First, Top10, Balance, Average,` etc.)
    - Ejemplo saldos, las dimensiones menos en la de tiempo.
  - No aditivas. No pueden ser agregadas sobre ninguna dimensión. (Proporciones).
    - Precios unitario, promedio de precios.

Ejemplo de una cadena de tiendas

- Se tiene una cadena con 500 mini tiendas de autoservicio distribuidas en 5 áreas geográficas.
- Cada tienda tiene departamentos como abarrotes, alimentos congelados, lácteos, carnicería, panadería, artículos no perecederos, licorería, y farmacia.
- Cada tienda tiene 60,000 productos individuales en sus estantes.
- Cada producto se conoce como SKU (stock keeping unit).
- Los datos se obtienen de:
  - Las cajas registradoras de las compras realizadas.
  - Los pedidos recibidos para entrega a domicilio.

Allstar Grocery
123 Loon Street
Green Prairie, MN 55555
(952) 555-1212

Store: 0022
Cashier: 00245409/Alan

0030503347 Baked Well Multigrain Muffins	2.50
2120201195 Diet Cola 12-pack	4.99
Saved $.50 off $5.49	
0070806048 Sparkly Toothpaste	1.99
Coupon $.30 off $2.29	
2840201912 SoySoy Milk Quart	3.19
TOTAL	12.67
AMOUNT TENDERED	
CASH	12.67
ITEM COUNT: 4	

Transaction: 649 4/15/2013 10:56 AM

Thank you for shopping at Allstar

... Ejemplo de una cadena de tiendas

- La administración se encarga de la logística de las ordenes de compra, las existencias en los estantes, y las ventas de los productos con la finalidad de maximizar la ganancia en cada tienda.
- La ganancia se obtiene de cobrar tanto como sea posible de cada producto, bajar los costos de compra y atraer la mayor cantidad posible de clientes.
- Las decisiones más importantes están relacionadas con precios y promociones. Tanto el administrador de tiendas como el jefe de mercadotecnia emplean mucho tiempo jugando con los precios y diseñando promociones. Las promociones incluyen reducción de precios de artículos de temporada, anuncios en los medios de comunicación, desplegados en la tienda.

Proceso de negocio (tienda)

Paso 1. Elegir el proceso de negocio que será modelado.

- Tomar en cuenta tanto los requerimientos del negocio como los datos disponibles.
- El primer modelo dimensional construido debe ser el mayor impacto posible.
- Debería responder a las cuestiones más apremiantes del negocio y tener accesibles los datos para su extracción.

... Proceso de negocio (tienda)

En el ejemplo de los autoservicios.

- Procesos:
  - Ventas de artículos a clientes.
  - Compra de artículos a proveedores.
  - Envío de artículos a sucursales.
  - Pago a empleados.
  - Recursos humanos: pago de alquiler, pagos de anuncios, etc.
- ¿Cuál proceso es relevante para ser analizado e incrementar las ganancias?
  - El administrador desea entender mejor las compras de los clientes según lo capturado por el sistema POS.
  - Proceso de negocio: ventas al menudeo en POS.
  - Permite analizar:
    - qué productos se están vendiendo.
    - en qué tiendas.
    - en qué días.
    - bajo qué condiciones promocionales.

Granularidad (tienda)

Paso 2: Elegir la granularidad del proceso de negocio.

- Desarrollar modelos dimensionales para la información atómica capturada por un proceso de negocio no por las consultas de reportes individuales.
- En el ejemplo:
  - ¿Una venta?
  - ¿Ventas por sucursal/día?
  - ¿Ventas por ciudad/año?
  - **Información sobre las ventas diarias de cada producto en cada almacén de la cadena**.
- Permite un análisis muy detallado de las ventas:
  - Se puede diferenciar ventas en domingo o en lunes.
  - Permite conocer la venta de ciertos productos en diversos tamaños presentaciones.
  - ¿Cuántos artículos en promoción se vendieron?
  - ¿Impacta en términos de crecimiento de ventas un refresco de dieta?, ¿es promovido fuertemente?, etc.

Identificar las dimensiones (tienda)

- Determinar cuidadosamente la granularidad determina las principales dimensiones de la tabla de hechos.
- Con frecuencia, es posible agregar dimensiones después.
- En ocasiones se requiere una revisión del paso 2.
- Granularidad:
  - Información sobre las ventas diarias de cada producto en cada almacén de la cadena.
- Dimensiones que caracterizan la actividad al nivel de detalle (granularidad) elegido:
  - Fecha (dimensión temporal: ¿cuándo se produce la actividad?)
  - Producto (dimensión ¿cuál es el objeto de la actividad?)
  - Tienda (dimensión geográfica: ¿dónde se produce la actividad?)
  - Promoción. Usada para ver si las promociones son adecuadas. . Anuncios, reducciones de precio, despliegues, cupones.

Versión preliminar del esquema de las tiendas

Identificación de las medidas (tienda)

- Este paso tiene que ver con cuáles atributos se agregarán a la tabla de hechos.
- La granularidad determina las principales medidas a incluir en la tabla de hechos.
- En el ejemplo de las tiendas:
  - Granularidad: Se desea conocer información sobre las ventas diarias de cada producto en cada establecimiento de la cadena.
  - Hechos:
    - Importe total de las ventas ...
    - Cantidad de unidades vendidas a precio regular.
    - Cantidad de unidades vendidas con descuento.
    - Total de unidades vendidas a precio neto.
    - Importe total de ventas con descuento.
    - Ganancias.
    - Cantidad de clientes distintos que han comprado el producto en el día.
    - etc.

... Identificación de las medidas (tienda)

Hechos derivados

- Para obtener la ganancia neta se calcula el total de las ventas menos el total del costo, o ingresos.
- Aunque es calculada también puede ser utilizada para todas las combinaciones de dimensiones.
- ¿Cuando se debe almacenar?
  - Se calcula consistentemente en el proceso de ETL,
  - Se utiliza consistentemente.

Tabla de Hechos Transaccionales

Características:

- Generalmente la granularidad puede ser expresada en el contexto de una transacción.
- Son cubos muy ralos.
- Pueden ser enormes.
- Tienden a ser altamente dimensionales.
- Las métricas resultantes generalmente son aditivas.

Más acerca de las dimensiones

La dimensión Fecha.

- Está presente en todos los DW.
- Se puede crear de antemano.
  Se pueden poner los días de 20 años y sólo son 7,300 renglones lo cual es tabla de dimensión relativamente pequeña.
- Los valores significativos son importantes, por ejemplo, para la generación de reportes.
- Para el ejemplo de las tiendas puede ser:

Date Dimension
Date Key (PK)
Date
Full Date Description
Day of Week
Day Number in Calendar Month
Day Number in Calendar Year
Day Number in Fiscal Month
Day Number in Fiscal Year
Last Day in Month Indicator
Calendar Week Ending Date
Calendar Week Number in Year
Calendar Month Name
Calendar Month Number in Year
Calendar Year-Month (YYYY-MM)
Calendar Quarter
Calendar Year-Quarter
Calendar Year
Fiscal Week
Fiscal Week Number in Year
Fiscal Month
Fiscal Month Number in Year
Fiscal Year-Month
Fiscal Quarter
Fiscal Year-Quarter

... Dimensión fecha

Ejemplo de la tabla para la dimensión Fecha.

Date Key	Date	Full Date Description	Day of Week	Calendar Month	Calendar Quarter	Calendar Year	Fiscal Year-Month	Holiday Indicator	We Ind
20130101	01/01/2013	January 1, 2013	Tuesday	January	Q1	2013	F2013-01	Holiday	We
20130102	01/02/2013	January 2, 2013	Wednesday	January	Q1	2013	F2013-01	Non-Holiday	We
20130103	01/03/2013	January 3, 2013	Thursday	January	Q1	2013	F2013-01	Non-Holiday	We
20130104	01/04/2013	January 4, 2013	Friday	January	Q1	2013	F2013-01	Non-Holiday	We
20130105	01/05/2013	January 5, 2013	Saturday	January	Q1	2013	F2013-01	Non-Holiday	We
20130106	01/06/2013	January 6, 2013	Sunday	January	Q1	2013	F2013-01	Non-Holiday	We
20130107	01/07/2013	January 7, 2013	Monday	January	Q1	2013	F2013-01	Non-Holiday	We
20130108	01/08/2013	January 8, 2013	Tuesday	January	Q1	2013	F2013-01	Non-Holiday	We

... Dimensión fecha

Dimensión fecha vs el tipo `Date` de SQL:

- No se tiene tanta versatilidad en SQL, por ejemplo, no tiene mes fiscal, no distingue entre día hábil y día no hábil.
- Los tomadores de decisiones no saben SQL.
- La lógica de calendario está en la tabla de la dimensión no en el código de aplicación.
- La dimensión Fecha es relativamente pequeña. Por ejm. para 20 años se tienen 7,300 tuplas.

Banderas e indicadores como atributos textuales

¿Porqué para indicar si un día es hábil se implementa con una cadena y no con un booleano?

Monthly Sales

Period: June 2013

Product Baked Well Sourdough

Holiday Indicator	Extended Sales Dollar Amount
N	1,009
Y	6,298

Monthly Sales

Period: June 2013

Product Baked Well Sourdough

Holiday Indicator	Extended Sales Dollar Amount
Holiday	6,298
Non-holiday	1,009

Dimensión Producto

- Contiene la descripción de los productos.
  (Más de 50 atributos)
- Aunque una tienda puede tener hasta 60,000 SKUs al considerar diferentes esquemas de ventas y productos descontinuados, puede crecer aún más.
  (Más de 300 M renglones)
- Se alimenta, generalmente, del archivo maestro de la BD operacional.

Product Key	Product Description	Brand Description	Subcategory Description	Category Description	Department Description	Fat Content
1	Baked Well Light Sourdough Fresh Bread	Baked Well	Fresh	Bread	Bakery	Reduced Fat
2	Fluffy Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Regular Fat
3	Fluffy Light Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Reduced Fat
4	Light Mini Cinnamon Rolls	Light	Pre-Packaged	Sweeten Bread	Bakery	Non-Fat
5	Diet Lovers Vanilla 2 Gallon	Coldpack	Ice Cream	Frozen Desserts	Frozen Foods	Non-Fat
6	Light and Creamy Butter Pecan 1 Pint	Freshlike	Ice Cream	Frozen Desserts	Frozen Foods	Reduced Fat
7	Chocolate Lovers 1/2 Gallon	Frigid	Ice Cream	Frozen Desserts	Frozen Foods	Regular Fat
8	Strawberry Ice Creamy 1 Pint	Icy	Ice Cream	Frozen Desserts	Frozen Foods	Regular Fat
9	Icy Ice Cream Sandwiches	Icy	Novelties	Frozen Desserts	Frozen Foods	Regular Fat

Product Dimension
Product Key (PK)
SKU Number (NK)
Product Description
Brand Description
Subcategory Description
Category Description
Department Number
Department Description
Package Type Description
Package Size
Fat Content
Diet Type
Weight
Weight Unit of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
...

... Dimensión Producto

- Jerarquía de conceptos:
  - departamento > categoría > marca > SKU
  - Muchos de los atributos en esta tabla no son parte de la jerarquía de productos. Ejemplo tipo de envasado.
- Es frecuente que códigos de productos operacionales, identificados con NK (Natural Key) sean atributos cuyas partes tienen significado implícito.
- En ocasiones se tienen valores numéricos que no se sabe si son hechos o atributos de dimensiones. Ej. precio de lista para un producto.
  - Si el valor numérico se utiliza para cálculos, probablemente pertenezca a la tabla de hechos.
  - Si es un valor numérico estable usado para filtrar y agrupar debería tratare como un atributo en la dimensión producto.
  - Si se usa tanto para cálculos como para filtrar/agrupar se debería almacenar en ambas tablas.

... Dimensión Producto (drill-down)

- Ventas por departamento:

Department Name	Sales Dollar Amount
Bakery	12,331
Frozen Foods	31,776

- Ventas por departamento- marca:

Department Name	Brand Name	Sales Dollar Amount
Bakery	Baked Well	3,009
Bakery	Fluffy	3,024
Bakery	Light	6,298
Frozen Foods	Coldpack	5,321
Frozen Foods	Freshlike	10,476
Frozen Foods	Frigid	7,328
Frozen Foods	Icy	2,184

... Dimensión Producto

- Ventas por departamento - contenido calórico.

Department Name	Fat Content	Sales Dollar Amount
Bakery	Nonfat	6,298
Bakery	Reduced fat	5,027
Bakery	Regular fat	1,006
Frozen Foods	Nonfat	5,321
Frozen Foods	Reduced fat	10,476
Frozen Foods	Regular fat	15,979

... Dimensión Producto

Conclusión:

- Incluir en la dimensión producto, (que es común en muchos modelos dimensionales) tantos atributos descriptivos como sea posible.
- Un conjunto de atributos robusto y completo se traduce en posibilidades de análisis robustas y completas para los usuarios.

Dimensión Tienda

- Describe cada tienda en la cadena.
- A diferencia de la dimensión Producto, en ésta no hay un archivo maestro que tenga la información.
- En este caso, se deben recabar los componentes necesarios a partir de diversas fuentes operacionales.
- Cada tienda puede concebirse como una ubicación. Por tanto, País $>$ estado $>$ ciudad $>$ CP
- También pueden agruparse por regiones y municipios.
- Ambas jerarquías pueden estar en la dimensión. Cosa común.
- Los nombres de los atributos y valores deberán ser únicos a lo largo de las diversas jerarquías.


Store Dimension
Store Key (PK)
Store Number (NK)
Store Name
Store Street Address
Store City
Store County
Store City-State
Store State
Store Zip Code
Store Manager
Store District
Store Region
Floor Plan Type
Photo Processing Type
Financial Service Type
Selling Square Footage
Total Square Footage
First Open Date
Last Remodel Date
...

Extensibilidad del esquema

- Pasado el tiempo, se requiere saber de qué manera las promociones afectan las ventas. ¿Cómo implementarlo?
- Agregando otra dimensión.
- Sencillo debido a que se modeló el sistema con una línea de la transacción POS como granularidad, es decir con granularidad muy fina.
- Extensiones factibles:
  - Agregar atributos a una dimensión.
  - Agregar dimensiones.
  - Agregar hechos medibles.

Dimensión promoción

- Describe las condiciones de promoción bajo las cuales se debe vender un producto.
- Esta dimensión suele llamarse **dimensión causal** porque describe los factores que ocasionan un cambio en las ventas de los productos.
- Los tomadores de decisiones están interesados en determinar cuándo una promoción es efectiva.
- Las promociones se juzgan tomando en cuando los siguientes factores:
  - Si las ventas de los productos en promoción (PEP) van en ascenso, durante el periodo promocional.
  - Si los PEP muestran caída en las ventas justo antes o después de la promoción, cancelando la ganancia en ventas durante la promoción.
  - Si los PEP muestran una ganancia pero otros productos cercanos a él, en el aparador, muestran un decremento en ventas (canibalismo).
  - Si todos los productos en la categoría de "en promoción" experimentan una ganancia neta en las ventas tomando en cuenta los tiempos antes, durante y después de la promoción.

... Dimensión promoción

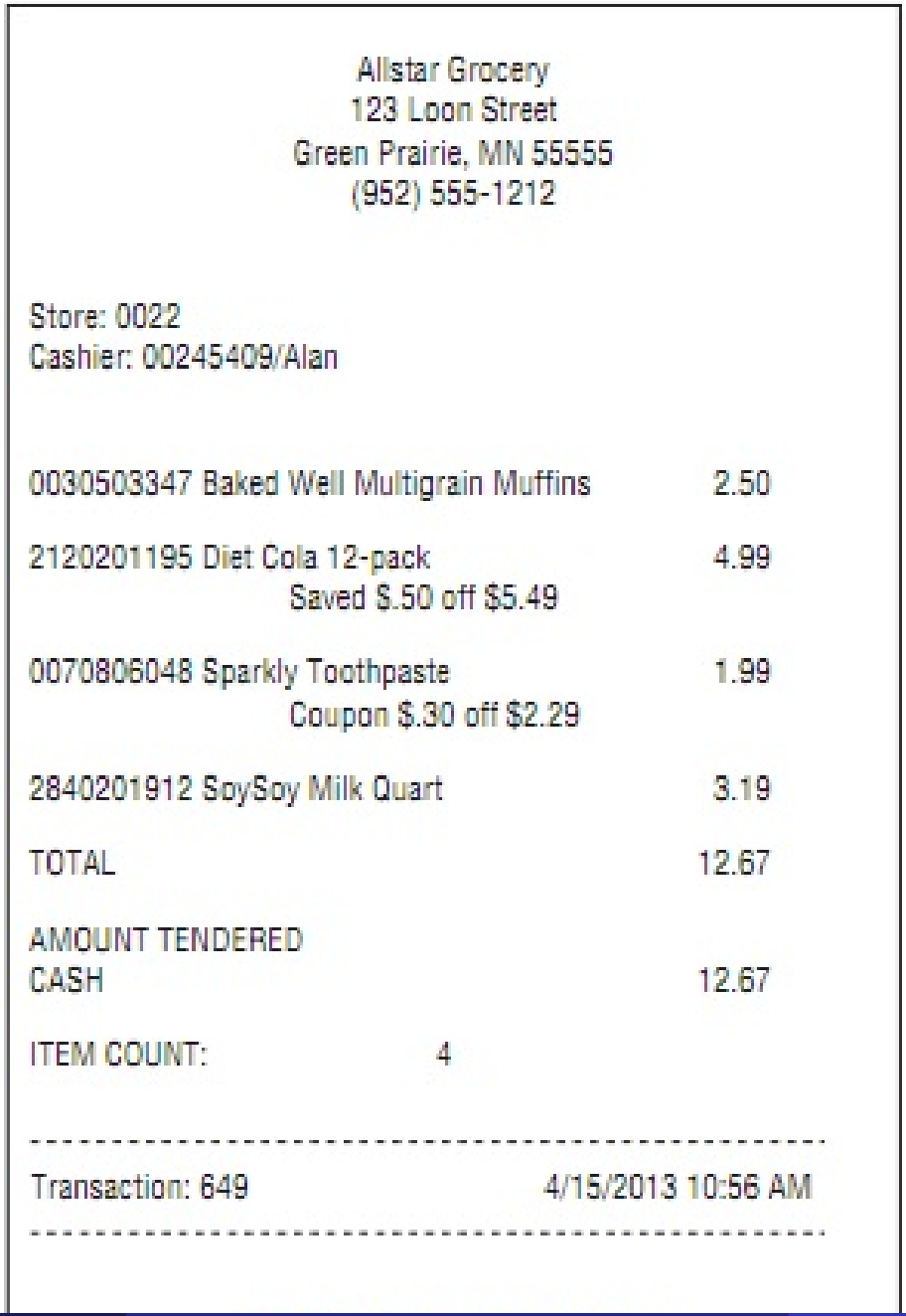
- Las condiciones causales afectan potencialmente una venta pero no necesariamente son registradas por el sistema POS directamente. El sistema transaccional registra la reducción de precio y rebajas.
- La presencia de cupones, también se registra al presentarlos.
- Las condiciones de propagandas en las tiendas pueden ser capturadas de otras fuentes.
- Las diferentes condiciones causales están altamente correlacionadas. Por eso tiene sentido crear un renglón en la dimensión promoción para cada combinación de condiciones de promoción.

... Dimensión promoción

... Dimensión promoción (Nulos)

- ¿Qué pasa con los artículos que no están en promoción? !!!
  - Llaves foráneas nulas.
  - Incluir un renglón en la tabla de dimensión promoción, para identificar que esta dimensión no es aplicable a esta medida.
  - De preferencia no utilizar nulos, mejor poner "desconocido", "no-aplicable", etc.

Otras dimensiones:



Allstar Grocery
123 Loon Street
Green Prairie, MN 55555
(952) 555-1212

Store: 0022
Cashier: 00245409/Alan

0030503347 Baked Well Multigrain Muffins	2.50
2120201195 Diet Cola 12-pack	4.99
Saved $.50 off $5.49	
0070806048 Sparkly Toothpaste	1.99
Coupon $.30 off $2.29	
2840201912 SoySoy Milk Quart	3.19
TOTAL	12.67
AMOUNT TENDERED	
CASH	12.67
ITEM COUNT: 4	

Transaction: 649 4/15/2013 10:56 AM

Dimensiones degeneradas

- Las dimensiones degeneradas son vacías, es decir son llaves a una tabla de dimensión inexistente.
- No son ni hechos ni dimensiones, y provienen de un sistema operacional.
- Ejemplos: los números de control operacional tales como número de orden, número de factura, transacción POS, etc.
  - Contienen información importante.
  - Resultan útiles como parte de la llave primaria en la tabla de hechos o para agrupar.
  - Por ejemplo, agrupar por número de transacción POS para recuperar todos los productos comprados en una sola transacción.

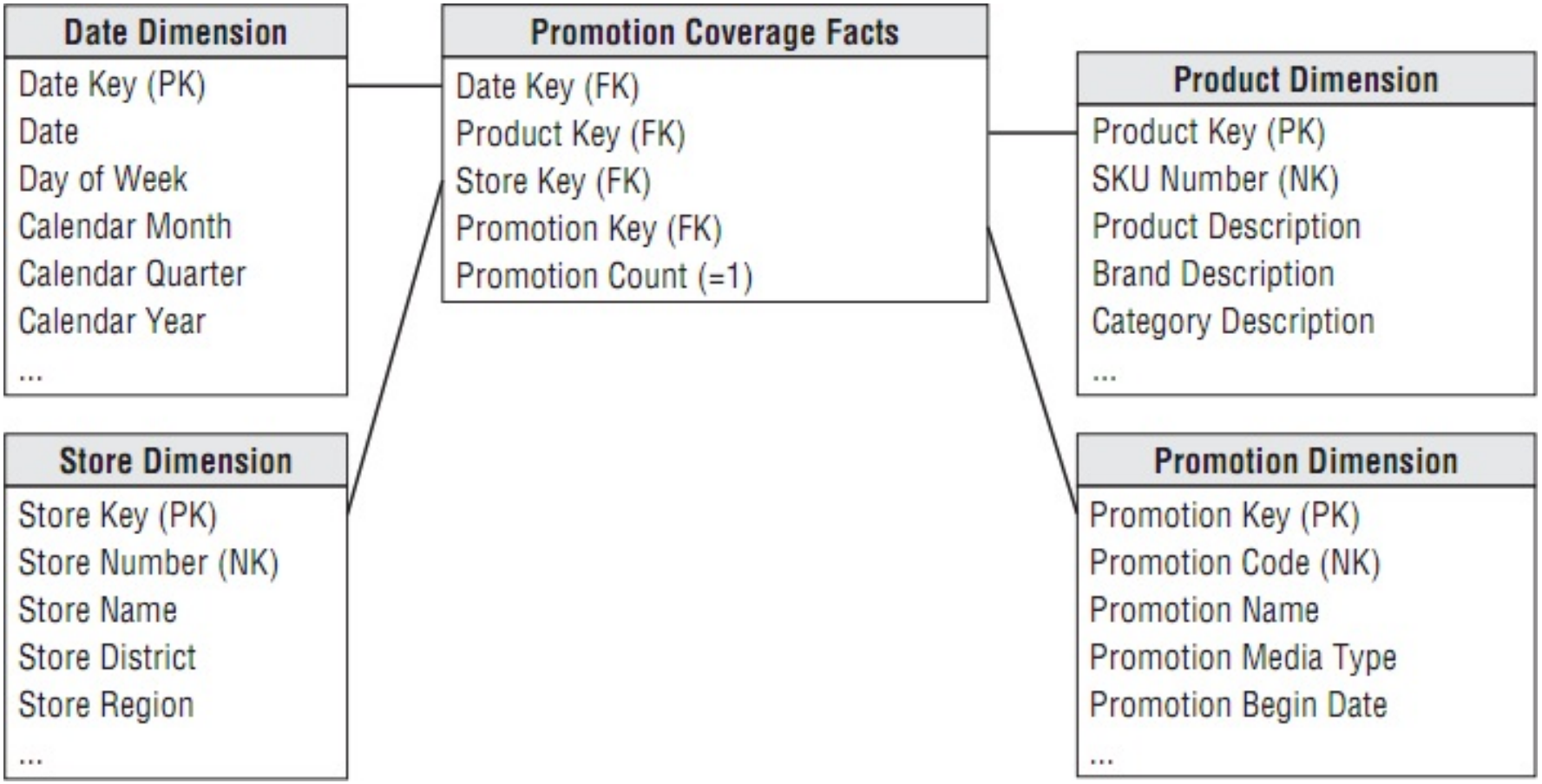
... Dimensiones degeneradas

Pueden utilizarse para formar la llave primaria. Ejemplo, la PK puede formarse con el número de transacción y el identificador del producto.

Tablas de hechos *factless*

- ¿Cómo podemos saber qué productos estuvieron en promoción y no se vendieron?
  Crear una tabla de hechos "Cobertura de promoción": fecha, producto, tienda y promoción.
  - Se puede cargar un renglón por cada producto en promoción en cada almacén cada día o semana, independiente de la venta del producto.
- Esta tabla sólo permite ver la relación entre las llaves definidas por la promoción independiente de cualquier otro evento como por ejemplo venta de productos.
- Se conocen como tablas de hechos sin hechos, debido a que no hay medidas sólo captura la relación entre las llaves involucradas.

... Tablas de hechos *factless*



... Tablas de hechos *factless*

Para la pregunta se requieren los pasos:

- Consultar la tabla factless para obtener el conjunto de productos que estuvieron en promoción en un día dado.
- Consultar la tabla de hechos para determinar los productos que se vendieron.
- Obtener la diferencia de los conjuntos obtenidos en los dos puntos anteriores.

Llaves sustitutas

- La PK de una tabla de dimensión se conoce como **llave sustituta**. Éstas son números enteros consecutivos.
  - No tienen significado para el negocio.
  - Evitan tener llaves reales.
- Las llaves reales se conocen como **llaves de negocio, de producción** u **operacionales**.
  - Se identifican con NK. (Natural key)
  - Lo común es que aparezcan como atributos en alguna dimensión.
  - Deberían utilizarse en lugar de códigos de producción operacionales.
- Ventajas de las llaves sustitutas:
  - Hacen el DWH independiente de cambios operacionales.
  - Evitan el traslape de llaves al consolidar los datos.
  - Rendimiento: enteros pequeños vs. largos códigos alfanuméricos.
- ¿Convenientes para la dimensión Fecha?.
- ¿En la tabla de hechos?

Características de las dimensiones

Cada tabla de dimensión.

- Tiene una llave sustituta.
- Es una tabla ancha.
- Principalmente tiene atributos textuales.
- Es común que los atributos tengan orden jerárquico.
- Tiene pocos registros.
- No está normalizada.

Midiendo la tabla de hechos

- Dimensión tiempo: 2 años = 730 días.
- Dimensión tienda: 300 tiendas reportando diariamente.
- Dimensión producto: 30,000 productos, sólo se venden 3,000 diarios.
- Dimensión promoción: 5,000 combinaciones, pero un producto sólo aparece en una combinación.
- Cantidad de hechos registrados:
  $730 \times 300 \times 3,000 \times 1 = 657,000,000$
- Cantidad de campos: 4 llaves + 7 medidas = 11 campos.
- Tamaño de la tabla:
  $657,000,000 \times 11 campos \times 4 bytes = 28 GB$
- ¿Una tabla **pequeña** para los estándares actuales?
- Huecos en la tabla de hechos.
  - 1 -(Cardinalidad de TH)/(Producto de cardinalidad de dimensiones)
  - (1-(657,000,000)/(730*300*30,000*5,000) = 1 - 0.00002 = 0.99998

  Cubo principalmente vacío, sólo cerca del 1 % tiene valor.

Redundancia en los almacenes de datos

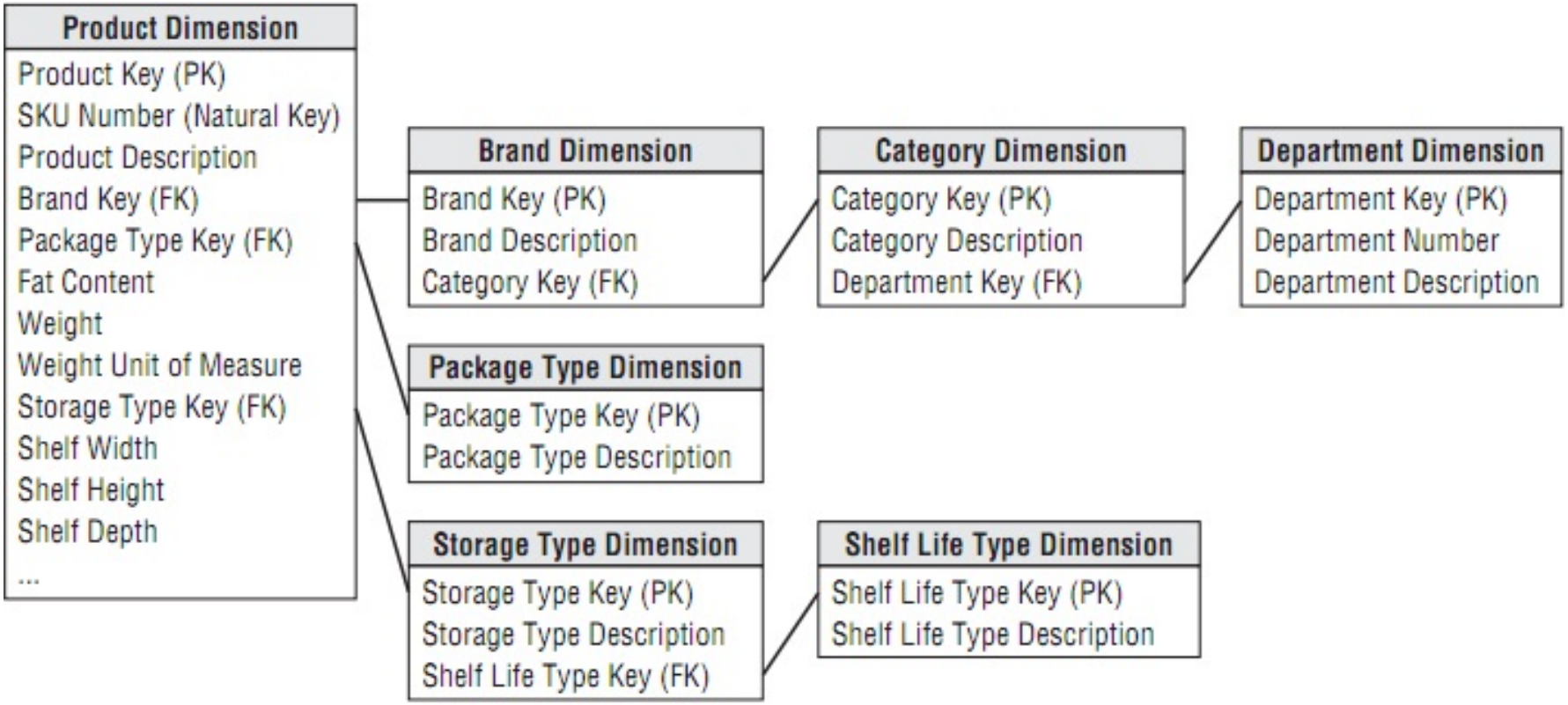
- Es mínima en las tablas de hechos.
  - Un hecho, generalmente, se almacena en una sola tabla de hechos.
- La redundancia principalmente se da en las tablas de dimensión.
  - Las tablas de dimensión tienen entradas redundantes para elementos en los niveles superiores.
  - Ejemplo: Si hay 300,000 productos en 50 departamentos distintos en la dimensión producto.
- ¿Problemas con la redundancia?
  - Inconsistencia en los datos.
  - Tiempo para actualizaciones.
  - Espacio usado.

... Redundancia en los almacenes de datos

La normalización en DWH se conoce como *snowflaking* con ella se eliminan valores redundantes en algún atributo.

Product Key	Product Description	Brand Description	Subcategory Description	Category Description	Department Description	Fat Content
1	Baked Well Light Sourdough Fresh Bread	Baked Well	Fresh	Bread	Bakery	Reduced Fat
2	Fluffy Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Regular Fat
3	Fluffy Light Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Reduced Fat
4	Light Mini Cinnamon Rolls	Light	Pre-Packaged	Sweeten Bread	Bakery	Non-Fat
5	Diet Lovers Vanilla 2 Gallon	Coldpack	Ice Cream	Frozen Desserts	Frozen Foods	Non-Fat
6	Light and Creamy Butter Pecan 1 Pint	Freshlike	Ice Cream	Frozen Desserts	Frozen Foods	Reduced Fat
7	Chocolate Lovers 1/2 Gallon	Frigid	Ice Cream	Frozen Desserts	Frozen Foods	Regular Fat
8	Strawberry Ice Creamy 1 Pint	Icy	Ice Cream	Frozen Desserts	Frozen Foods	Regular Fat
9	Icy Ice Cream Sandwiches	Icy	Novelties	Frozen Desserts	Frozen Foods	Regular Fat

... Redundancia en los almacenes de datos



Ejemplo: Inventarios

- La base de toda empresa comercial es la compra y venta de bienes o servicios. De ahí la importancia de los inventarios.
- Un inventario es
  " Asiento de los bienes y demás cosas pertenecientes a una persona o comunidad, hecho con orden y precisión."
- Dada la importancia que tiene para una empresa un manejo eficiente de su inventario, existen diferentes maneras de realizar un modelo de inventario:
  - Fotografía (snapshot) periódica.
  - Inventario por transacciones.
  - Modelo de fotografía acumulada.

Inventario por fotos periódicas

- Un inventario optimizado puede tener impacto sobre la rentabilidad.
  - Asegurar que cada producto está en la tienda en el momento preciso, minimiza la carencia de ellos y reduce los gastos del inventario.
- El minorista desea analizar diariamente la cantidad disponible de productos por tienda.
- Diseño multidimensional:
  - Proceso: cadena de tiendas.
  - Granularidad:
  - Dimensiones:

Inventario por fotos periódicas

- Un inventario optimizado puede tener impacto sobre la rentabilidad.
  - Asegurar que cada producto está en la tienda en el momento preciso, minimiza la carencia de ellos y reduce los gastos del inventario.
- El minorista desea analizar diariamente la cantidad disponible de productos por tienda.
- Diseño multidimensional:
  - Proceso: Llevar/analizar el inventario periódico de una cadena de tiendas.
  - Granularidad: Inventario diario de cada producto en cada tienda.
  - Dimensiones:
  - Hechos:

... Fotos periódicas

Atributos para las dimensiones:

- Fecha: igual que antes.
- Producto: como antes más atributos necesarios para el inventario, tales como cantidad mínima para solicitar más producto o necesidades de espacio.
- Tienda: características tales como superficie de las áreas de refrigeración y congelación.

... Fotos periódicas (tabla de hechos)

- La tabla de hechos ahora es muy densa, pues no se desea caer en situaciones de carencia/falta de algunos artículos.
- Se podría registrar un renglón en la tabla de hechos para indicar que no hay falta de algún articulo.
  - En la cadena de tiendas, con 60,000 productos en 100 tiendas, se tendrían aproximadamente 6 millones de registros cada noche que se actualice la tabla de hechos.
- Aunque es manejable, se podrían tomar acciones como inventario. Mantener los 60 días recientes de inventario y agruparlas semanalmente para los datos históricos:
  - En tres años diario: 1,095 fotografías
  - 60 diarias y semanales, en tres años = 60+148 = 208
    Problema: granularidad.

... Fotos periódicas

- En la bodega se tiene como una medida la cantidad disponible de producto.
- Puede resumirse/sumarse a lo largo de productos o de tiendas:
  - Dado un producto y una fecha, se puede obtener por tienda.
  - Dada una fecha y una tienda se puede obtener por producto.
  - Dado un producto y una tienda ¿se puede obtener por fecha? los niveles de inventario representan un nivel o balance en un momento dado.
- Por lo tanto, es un hecho/medida semi-aditiva.
- Las medidas que registran un nivel estático (niveles de inventario, saldos en las cuentas, temperatura ambiente, etc.) son medidas no aditivas a lo largo de la dimensión fecha y posiblemente de otras dimensiones.
- La forma más útil de combinar niveles de inventario a lo largo de las fechas es promediándolas.

... Fotos periódicas (Hechos mejorados)

- Puede no ser suficiente conocer la cantidad de artículos disponibles, podrían necesitarse otros hechos.
- Ejemplo medir el nivel de rotación del inventario y la cantidad de días para suministrar productos.
- Si se agrega cantidad vendida como una medida se pueden calcular las dos anteriores.
  - Para este tipo de inventarios diarios, la tasa de rotación se calcula dividiendo la cantidad vendida entre la cantidad disponible.
  - Para inventarios a otro plazo, se calcula como cantidad total vendida dividida entre el promedio diario de cantidad disponible.
  - Días para suministro es un calculo equivalente.
- Otras medidas pueden ser el costo del inventario, así como el precio más reciente de venta.
- Notar que aunque la cantidad de artículos disponibles es semi-aditiva, las otras medidas son aditivas.
- Este es el tipo más común de inventario.

... Fotos periódicas (Esquema del DWH)

Introducción a la cadena de valor

- Todas las organizaciones tienen una cadena de valor de los procesos clave subyacentes.
- Una cadena de valor identifica el flujo lógico, natural de las actividades primarias de una organización.
- Una cadena de valor es el conjunto de actividades desempeñadas internamente por una organización para diseñar, producir, llevar al mercado, entregar y apoyar sus productos.
- Permite identificar y analizar actividades estratégicamente relevantes para obtener alguna ventaja competitiva.

... Cadena de valor (ejemplo)

Orden de compra

Entrega a Almacén minorista

Almacén Individual

Consumidor Compra Producto

Producto desaparece de la Cadena de Valor

... Cadena de valor de la cadena de minoristas

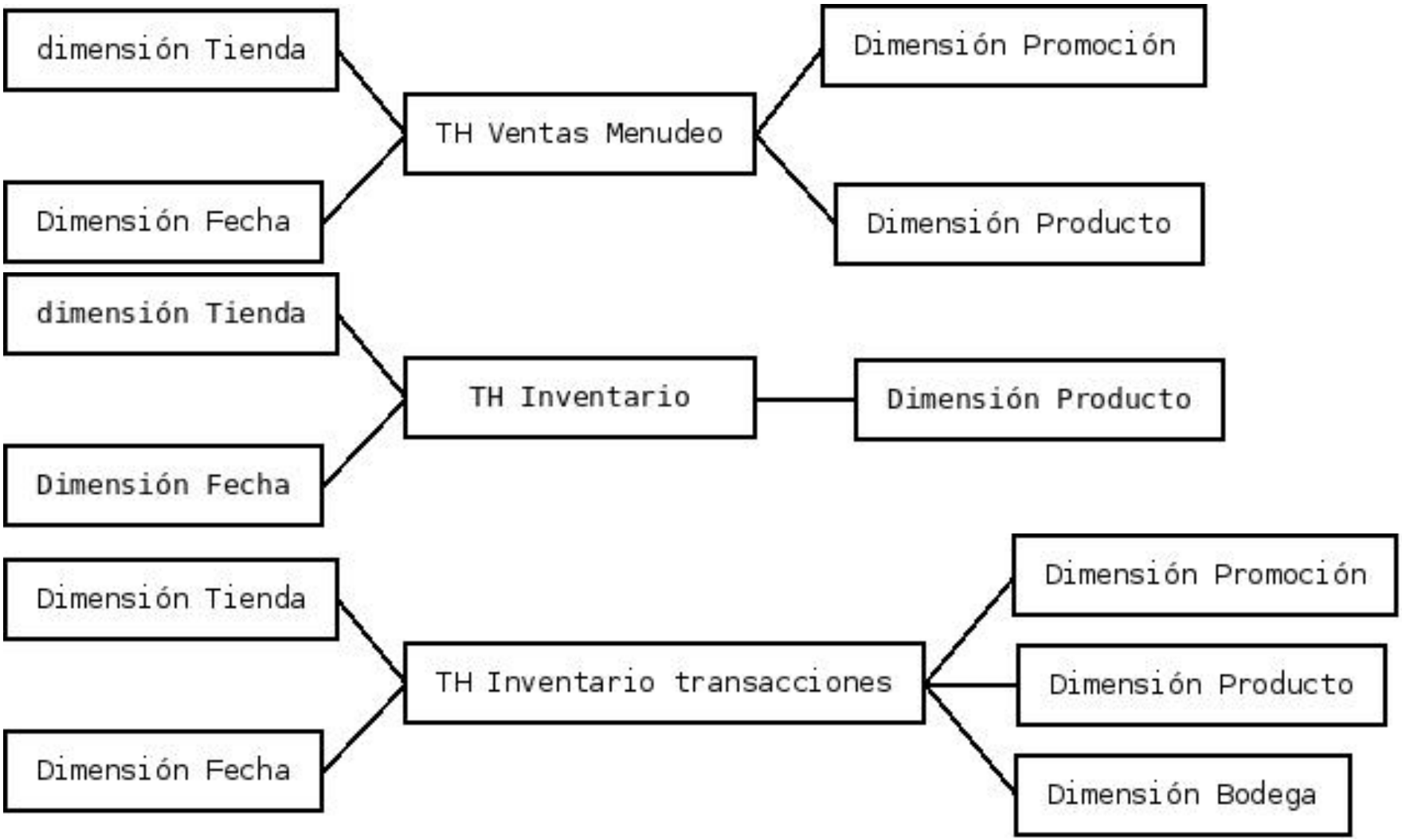
- Orden de compra a mayoristas.
- Entregas en la bodega minorista.
- Inventario de la bodega.
- Entregas en la tienda minorista.
- Inventario de la tienda minorista.
- Ventas de la tienda minorista.

Si la entrega no requiere de pasar por la bodega el camino es más corto.

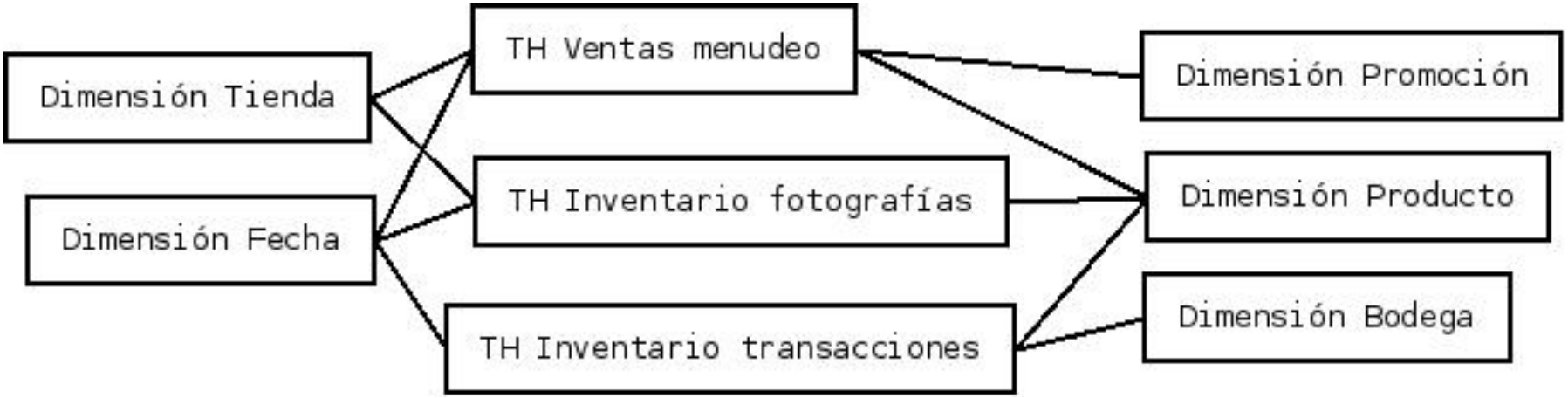
Integración de una cadena de valor en un DWH

- Integración de una cadena de valor en un DWH.
- Como se puede observar la cadena de valor se puede modelar con diferentes modelos dimensionales. Éstos pueden compartir dimensiones.
- Utilizar dimensiones compartidas es fundamental para el diseño de modelos dimensionales que puedan integrarse.
    - Permite combinar las métricas de rendimiento de diferentes procesos en un solo reporte.

... Integración de una cadena de valor en un DWH



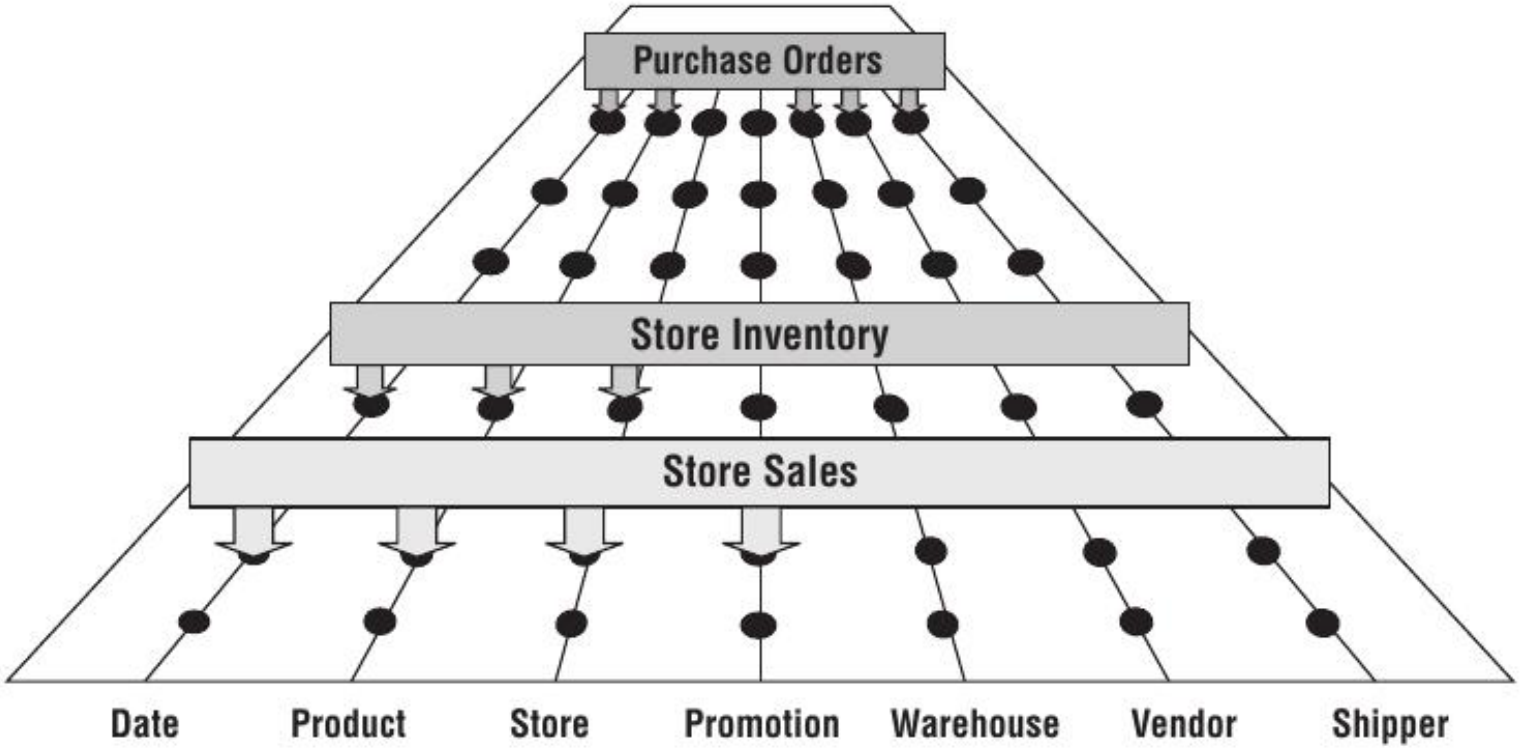
... Integración de una cadena de valor en un DWH



Arquitectura de bus

- La construcción de un DWH es una labor titánica.
- Para facilitar el trabajo se utiliza un enfoque incremental.
- Un bus es una estructura común en la que todo se conecta para derivar poder.
  - Permite que modelos dimensionales puedan ser implementados por diferentes grupos en tiempos diferentes.
  - Estos procesos separados de negocio se conectan entre si y coexisten útilmente si se ajustan a la norma.

... Arquitectura de bus



... Arquitectura de bus

- La arquitectura de bus proporciona un enfoque para descomponer las tareas de planeación de un DWH empresarial.
- El conjunto de dimensiones y hechos tiene una interpretación uniforme a lo largo de la organización.
- Permite emprender la implementación de modelos dimensionales separados dentro de la misma arquitectura.
- Conforme los modelos dimensionales estén disponibles se van uniendo como piezas de rompecabezas.
- Esta arquitectura tiene como ventajas
  - Proporcionar un marco para guiar el diseño del almacén completo dividiendo el problema en pequeños procesos.
  - Equipos de desarrollo independientes pueden trabajar.
  - Esta arquitectura es independiente de tecnologías y plataformas de BD.

Matriz del bus del almacén de datos

La matriz del bus del almacén de datos es una herramienta que sirve para diseñar, documentar y comunicar esta arquitectura.

	COMMON DIMENSIONS						
BUSINESS PROCESSES	*Date*	*Product*	*Warehouse*	*Store*	*Promotion*	*Customer*	*Employee*
Issue Purchase Orders	X	X	X				
Receive Warehouse Deliveries	X	X	X				X
Warehouse Inventory	X	X	X				
Receive Store Deliveries	X	X	X	X			X
Store Inventory	X	X		X			
Retail Sales	X	X		X	X	X	X
Retail Sales Forecast	X	X		X			
Retail Promotion Tracking	X	X		X	X		
Customer Returns	X	X		X	X	X	X
Returns to Vendor	X	X		X			X
Frequent Shopper Sign-Ups	X			X		X	X

... Matriz del bus del almacén de datos

- Las columnas representan las dimensiones de la empresa
- Los renglones son los procesos (datamarts) basados en las actividades de la organización.
- Errores comunes:
  - Diseñar los renglones para departamento.
  - Diseñar los renglones para reportes solicitados.
  - Columnas excesivamente generalizadas.
  - Columnas separadas para cada nivel de una jerarquía.
- Adaptación de los modelos existentes a una matriz de bus.
  - Primero evaluar las estructuras dimensionales no integradas.
  - Desarrollar un plan incremental para convertir los modelos dimensionales aislados a la arquitectura de la empresa.

... Matriz del bus del almacén de datos

- Crear la matriz del bus del almacén de datos:
  - Herramienta de diseño que permite la comunicación en entro o fuera d la empresa.
  - Dispositivo de gran alcance para el planteamiento y comunicación.
  - Permite comunicación dentro y a través de los equipos de desarrollo de los data marts.
  - Crear una matriz es uno de los elementos más importantes para la implementación del DWH.

Dimensiones conformadas

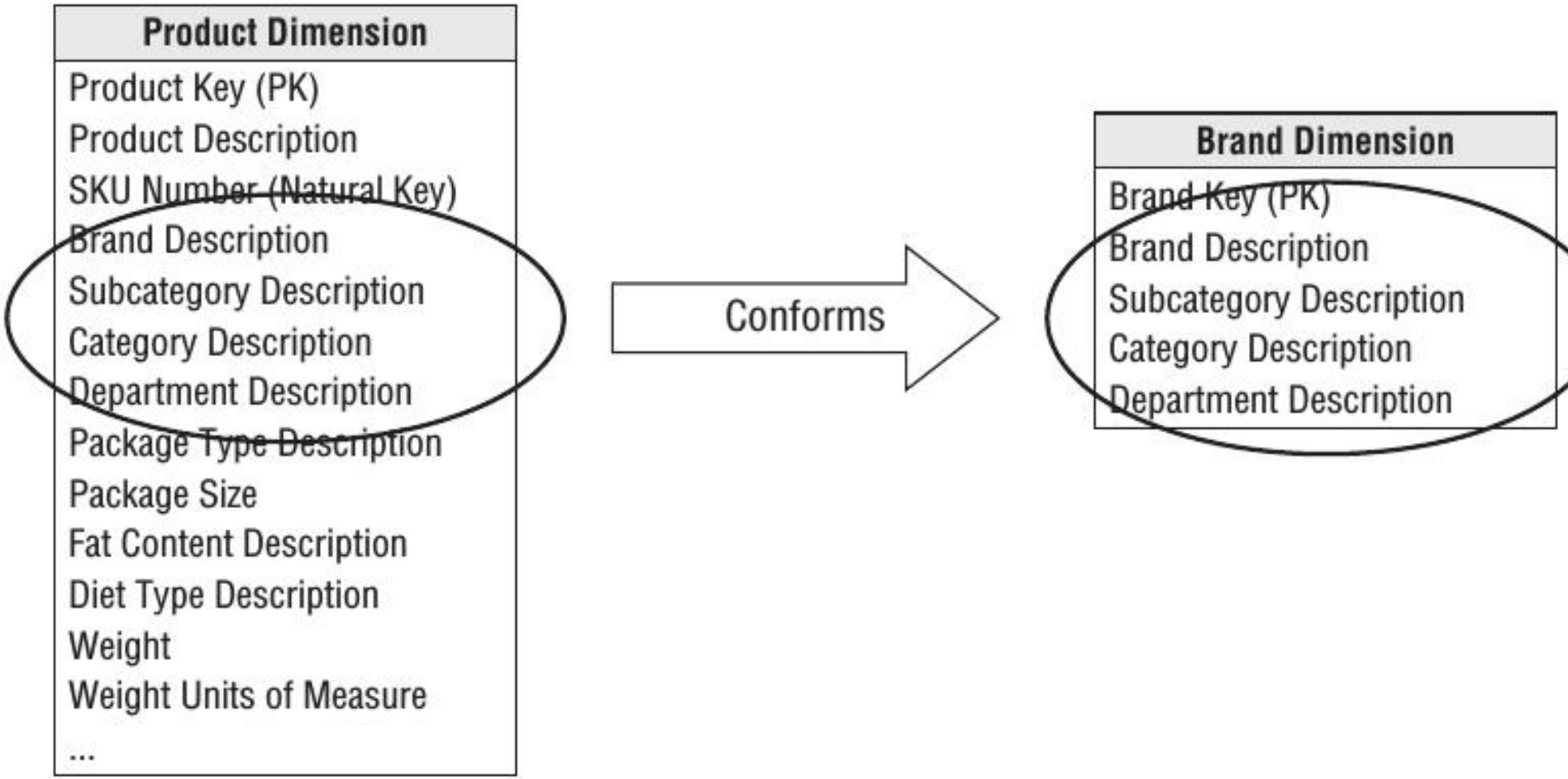
- Dimensiones conformadas son aquellas que:
  - Tienen exactamente los mismos atributos, valores e incluso llaves.
    - Ejemplo:
  - Una es subconjunto de la otra
    - Dimensión Fecha conformada a una dimensión Mes.
- También se conocen como dimensiones comunes, maestras, referencias o compartidas.
- Sirven como piedra angular del bus debido a que son compartidas a lo largo de las tablas de hechos del negocio.
- Las tablas de dimensiones no están conformadas si sus atributos están etiquetados de manera diferente o sus valores son diferentes.

Desglosado en tablas de hechos

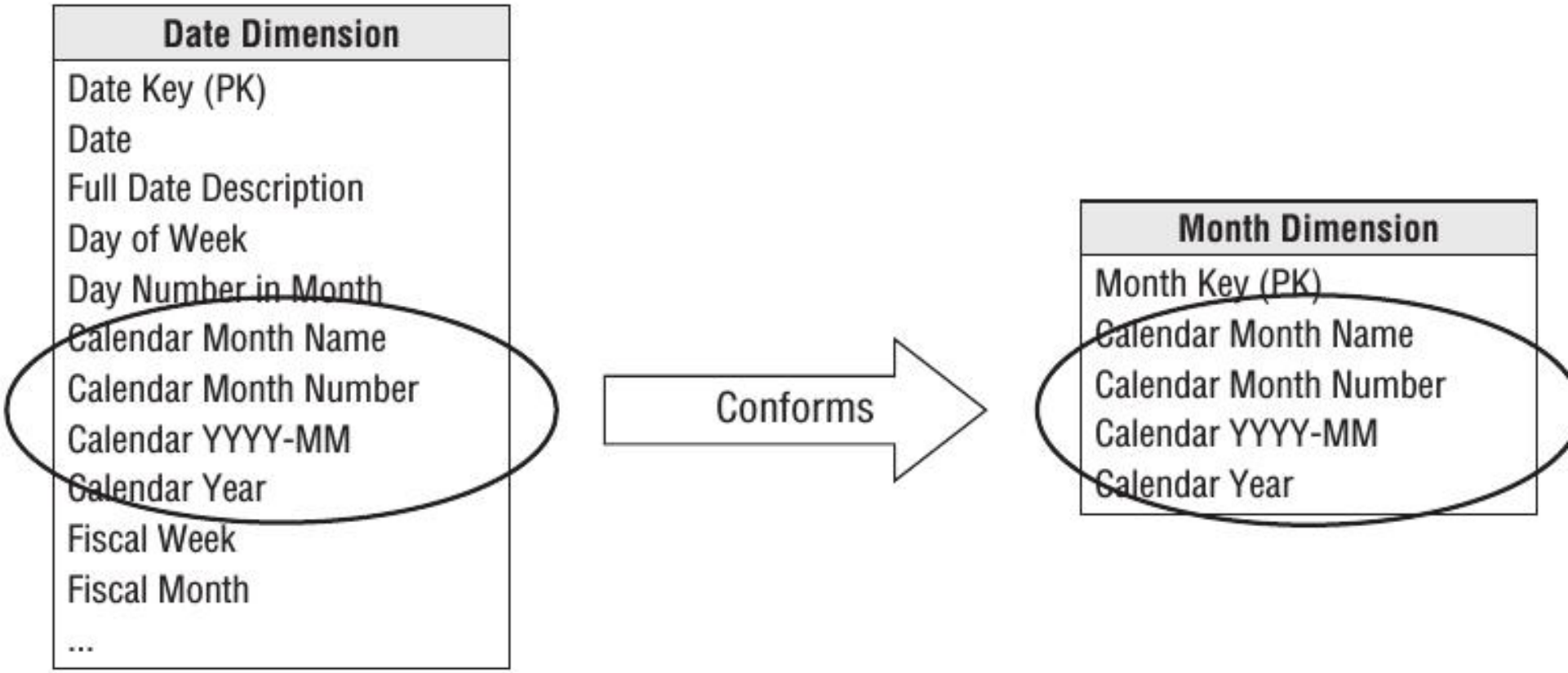
- Además de consistencia y reusabilidad, las dimensiones conformadas permiten combinar medidas de diferentes procesos en un solo reporte.
- Esta operación se conoce como drill-across.
- Ejemplo:

Product Description	Open Orders Qty	Inventory Qty	Sales Qty
Baked Well Sourdough	1,201	935	1,042
Fluffy Light Sliced White	1,472	801	922
Fluffy Sliced Whole Wheat	846	513	368

... Dimensiones conformadas reducidas (ejemplos)



... Dimensiones conformadas reducidas (ejemplos)



Detección de dimensiones conformadas

- En la matriz de bus se pueden detectar fácilmente las dimensiones conformadas idénticas.
- Las dimensiones conformadas reducidas pueden detectarse.
  - Marcar la celda para la dimensión atómica, pero incluir un texto para indicar la granularidad del subconjunto.
  - Subdividir la columna de dimensión para indicar las granularidades del subconjunto.

	Date
Issue Purchase Orders	X
Receive Deliveries	X
Inventory	X
Retail Sales	X
Retail Sales Forecast	X Month

OR

	Date	
	Day	Month
Issue Purchase Orders	X	
Receive Deliveries	X	
Inventory	X	
Retail Sales	X	
Retail Sales Forecast		X