CI2355 – Almacenes de datos y OLAP

UNIVERSIDAD DE
COSTA RICA

CI2355 – Almacenes de datos y OLAP

UNIVERSIDAD DE
COSTA RICA

Modelo multidimensional

Modelo multidimensional

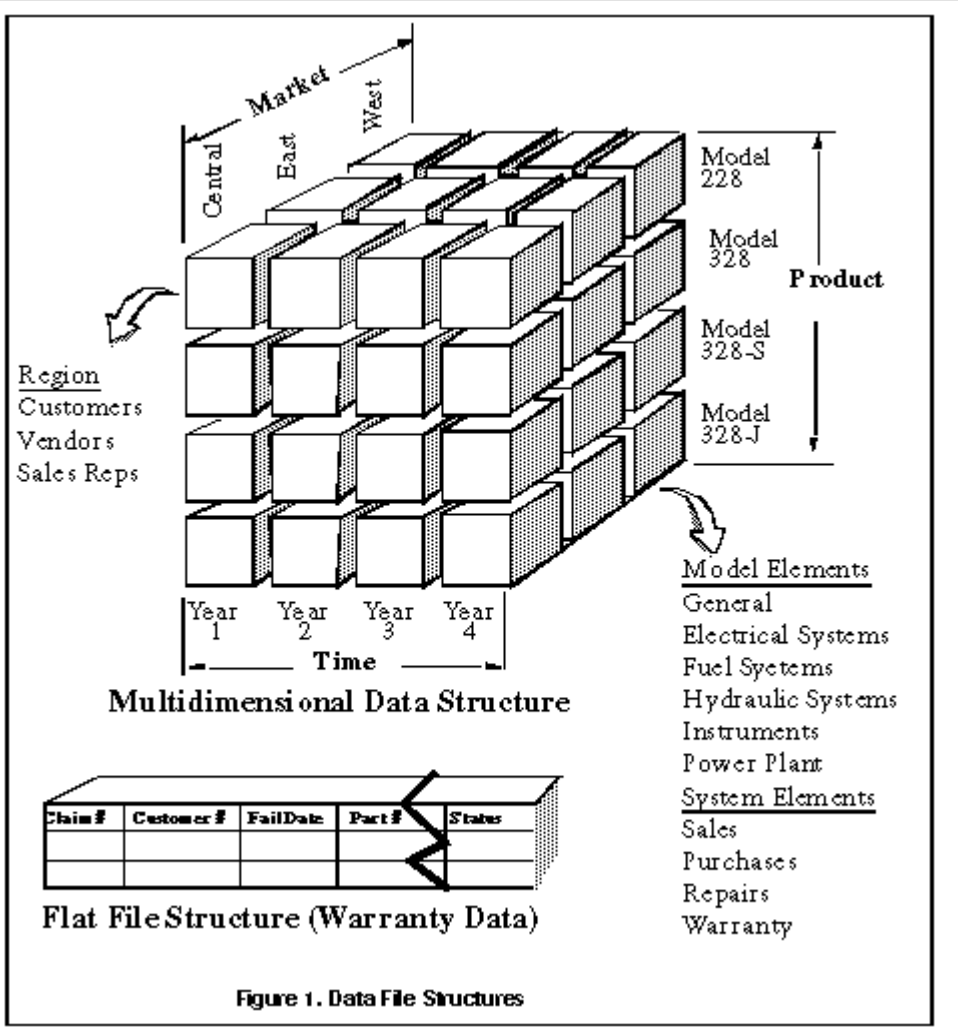
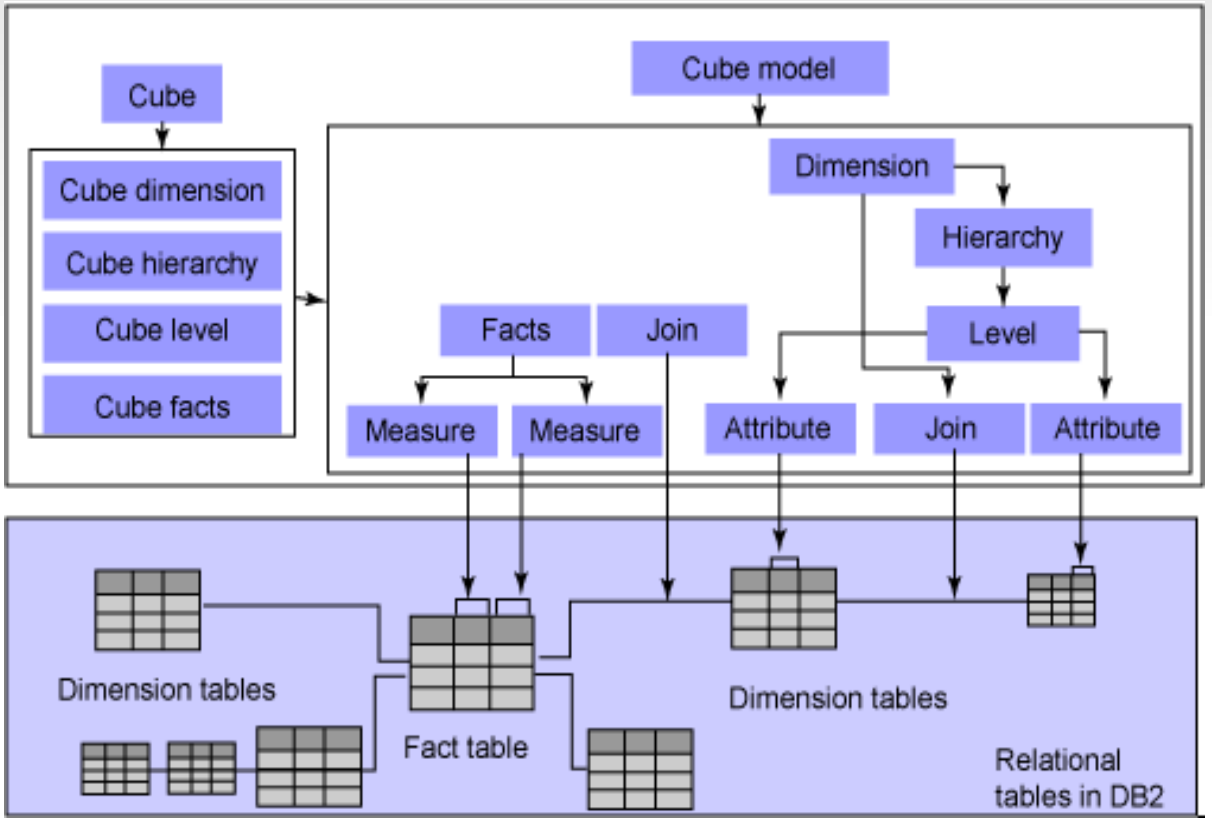


Figure 1. Data File Structures

Modelo multidimensional



Introducción al modelo multidimensional

- Almacenes de datos como disciplina
  - No sólo técnicas y herramientas
  - Antes que nada, deben considerarse las necesidades del negocio
  - Veremos los almacenes de datos como disciplina tanto desde el punto de vista del DBA como del analista de negocios
  - Uno de los activos más importantes de cualquier organización es su información

Dos enfoques diferentes

- Sistemas operacionales
  - Donde los datos son ingresados
  - Usualmente un registro a la vez
  - Operaciones repetitivas
- Almacenes de datos
  - Donde los datos son extraídos
  - Usuarios ven la organización desde otra perspectiva
  - Requieren ver miles de registros a la vez y hacen consultas diferentes

Objetivos de un almacén de datos

- Muchas veces, encontramos almacenes de datos que son simples copias del sistema operacional alojado en diferente hardware
- Lo que requieren las gerencias –temas recurrentes–
  - Poder acceder a los datos directamente
  - Información consistente entre diferentes departamentos
  - Ver sólo lo que es 'importante'
  - Tomar decisiones basadas en hechos

Requerimientos de los AD

1. El AD debe hacer que la información de la organización sea fácilmente accesible
   - Comprensible
   - Datos deben ser obvios para el usuario no informático
   - Comprensibilidad implica legibilidad
     - Contenidos etiquetados apropiadamente
   - Usuarios quieren separar y combinar los datos en innumerables formas diferentes (*slicing* y *dicing*)
   - Interfaces intuitivas y fáciles de usar (y rápidas)

Requerimientos de los AD

2. El AD debe presentar la información de la organización de forma consistente

- Datos deben ser creíbles
- Datos deben ser cuidadosamente ensamblados desde diferentes fuentes
- Información de un proceso de negocio debe 'cuadrar' con la de otro proceso de negocio
- Si dos apreciaciones de rendimiento tienen el mismo nombre, entonces deben ser lo mismo
- Información consistente significa información de alta calidad: detallada y completa

Requerimientos de los AD

3. El AD debe ser adaptativo y elástico
   - El cambio no puede evitarse
     - Necesidades de los usuarios
     - Condiciones de los negocios, datos y tecnologías
   - AD debe diseñarse para manejar estos cambios
     - Cambios en el AD no deben invalidar datos o aplicaciones ya existentes
   - Estas aplicaciones ya existentes no deberían ser afectadas cuando se hacen nuevas consultas o se agregan nuevos datos

Requerimientos de los AD

4. El AD debe ser un bastión seguro que proteje los activos de información de la organización
   - La información más importante de la organización se guarda en el AD
   - Al menos, debe contener qué se le vende a quién y a qué precio
   - Detalles potencialmente nocivos en manos de la gente equivocada
   - El AD debe efectivamente controlar el acceso a la información confidencial de la organización

Requerimientos de los AD

5. El AD debe ser el pilar de mejores tomas de decisiones
   - Debe contener los datos adecuados para apoyar la toma de decisiones
   - Verdadero producto final del AD
     - Las decisiones que se toman luego de que el AD presenta su evidencia
   - Estas decisiones son el impacto en el negocio y el valor agregado atribuible al AD
   - Recordar que estos sistemas eran conocidos como de 'apoyo a la toma de decisiones'

Requerimientos de los AD

6. La comunidad de negocios va a aceptar el AD si es exitoso
   - No importa si el AD es una solución elegante con tecnología de punta y la herramienta más prestigiosa
     - Si la comunidad de negocios no sigue utilizándolo seis meses luego de su implementación
       - Entonces el AD no pasó la prueba de la aceptación
   - Otra diferencia notable con los sistemas OLTP

Metáfora del webmaster (1)

- Principales actividades
  - Identificar a los visitantes demográficamente
  - Encontrar qué quieren los visitantes del sitio web
  - Identificar los 'mejores' visitantes (hacen suscripciones y compran de los anunciantes)
  - Descubrir nuevos visitantes potenciales y hacerles conocer el sitio web
  - Escoger el tipo de contenido del sitio que más atrae a los visitantes

Metáfora del webmaster (2)

- Mantener estándares de contenido y edición de alta calidad, al mismo tiempo que adoptar un estilo de presentación consistente
- Monitorear continuamente la veracidad de los artículos y de los anuncios
- Desarrollar una buena comunidad de escritores y lectores que contribuyen con comentarios
- Conseguir nuevas fuentes de contenido para el sitio
- Atraer y conservar anunciantes

Metáfora del webmaster (3)

- Publicar nuevo contenido a intervalos regulares
- Mantener la confianza de los visitantes
- Mantener felices a los dueños del sitio
- Tomar decisiones sobre el estilo y formato que más agrade a los visitantes

No recomendable

- Depender de alguna tecnología o plataforma determinada
- Dedicar demasiada energía en la parte operacional
- Imponer un estilo de escritura que los lectores no comprenden fácilmente
- Crear un formato complejo e intrincado

Paralelo con los AD

- Administrador de AD
  - Editor de datos correctos
  - De acuerdo con las necesidades del negocio, son responsables publicar datos recogidos de una variedad de fuentes y editados buscando calidad y consistencia
  - Sus lectores (visitantes)
    - Usuarios de negocios (gerentes)
  - Enfoque en los clientes, además de los productos y procesos

Calidades del AAD (1)

- Comprender a los usuarios por área, responsabilidades y tolerancia a la tecnología
- Determinar qué decisiones quieren tomar los usuarios de negocio con ayuda del AD
- Identificar los 'mejores' usuarios, que toman decisiones efectivas y de gran impacto utilizando el AD
- Encontrar nuevos usuarios potenciales y hacerles saber del AD
- Escoger el subconjunto de datos más efectivo para presentar en el AD del vasto universo de datos posibles en la organización

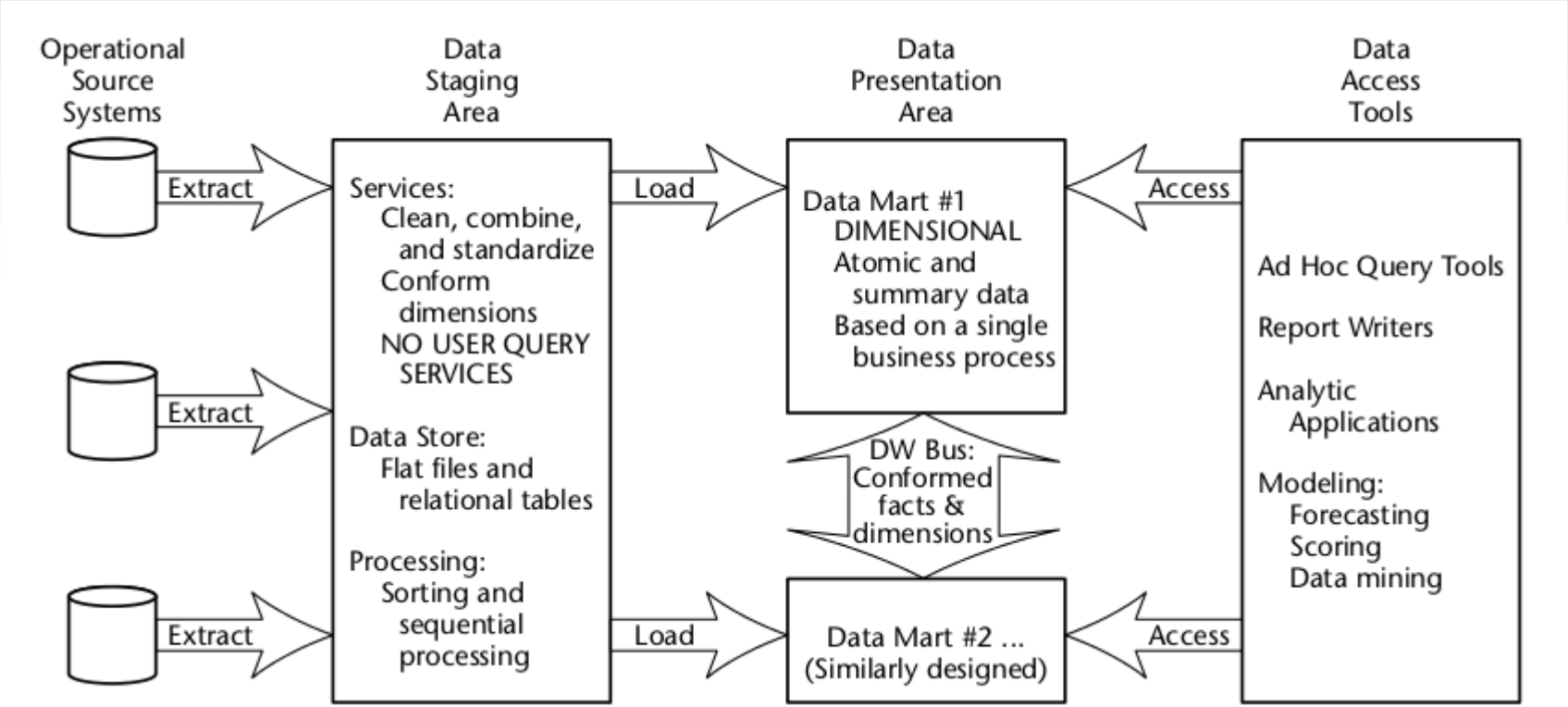
Calidades del AAD (2)

- Hacer las interfaces de usuario y las aplicaciones simples y basadas en plantillas, conciliarlas explícitamente con los perfiles de procesamiento de los usuarios
- Asegurarse de la exactitud y confiabilidad de los datos, etiquetándola consistentemente (documentación)
- Monitorear continuamente la exactitud de los datos y del contenido de los reportes
- Buscar nuevas fuentes de datos y adaptar continuamente el AD a perfiles cambiantes, requerimiento de reportes y prioridades del negocio

Calidades del AAD (3)

- Tomar parte del crédito por las decisiones de negocios tomadas utilizando el AD y utilizar los éxitos para justificar las inversiones en personal, software y hardware requeridos por el AD
- Publicar datos a intervalos regulares
- Mantener la confianza de los usuarios de negocios
- Mantener felices a los usuarios, ejecutivos y miembros de junta directiva

Componentes del AD



Sistemas operacionales fuentes

- Usualmente considerados fuera del ambiente del AD porque se tiene poco o ningún control sobre el formato y contenido de los datos en estos sistemas
- Principales prioridades
  - Rendimiento
  - Disponibilidad
- Pocos datos históricos
- Si existe un esfuerzo tipo EAI (*Enterprise Application Integration*), la tarea de diseño del AD se facilita

Área de preparación de datos

- Área de almacenamiento y de procesamiento tipo ETL
- Análogo a la cocina de un restaurante
  - Materia prima (ingredientes) es transformada en deliciosos platillos
  - Sólo accesible a profesionales entrenados
  - No se atiende aquí directamente a los clientes
- **El requerimiento arquitectónico clave es que está fuera del alcance de los usuarios del negocio y no provee servicios de consulta y presentación**

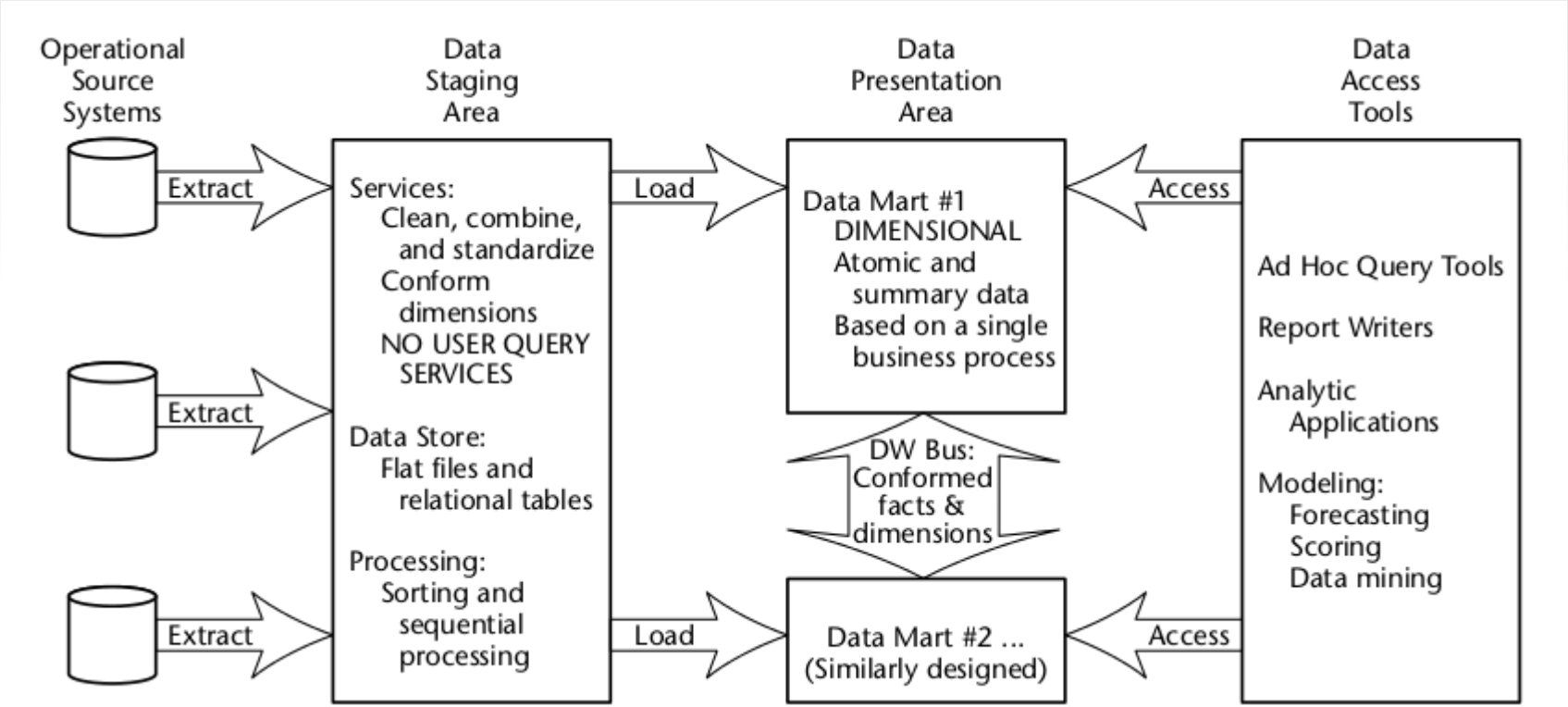
Área de preparación de datos (cont.)

- Primera etapa
  - Extraer datos
    - Leer y comprender
    - Copiar lo que sea necesario
  - BD relacionales o archivos planos
- Segunda etapa
  - Transformación
    - Correcciones, resolución de conflictos, manejar elementos faltantes, formatos, etc.

Área de preparación de datos (cont.)

- Tercera etapa
  - Entregar los datos depurados a las herramientas de carga de los mercados de datos
  - Cada mercado de datos debe indexar los datos recientemente recibidos, calcular agregados y volver a validar
  - Comunicar a los usuarios sobre la reciente actualización

Componentes del AD



Área de presentación de datos

- Área donde los datos son organizados, guardados y puestos a disposición para las consultas de los usuarios, herramientas de reportes y otras aplicaciones analíticas
- Para la comunidad de negocios, esta área ***es*** el AD
  - Todo lo que pueden ver y tocar del AD con sus herramientas de acceso a los datos
- Puede verse como una serie de mercados de datos
  - Cada mercado de datos presenta los datos de un único proceso de negocios

Área de presentación de datos (cont.)

- Modelo dimensional
  - Técnica 'vieja' utilizada para hacer las bases de datos simples y comprensibles
- Gerente:
  - 'Vendemos **productos** en varios **mercados** y medimos nuestros resultados cada ciertos intervalos de **tiempo**'
- → Cubo

Modelo multidimensional

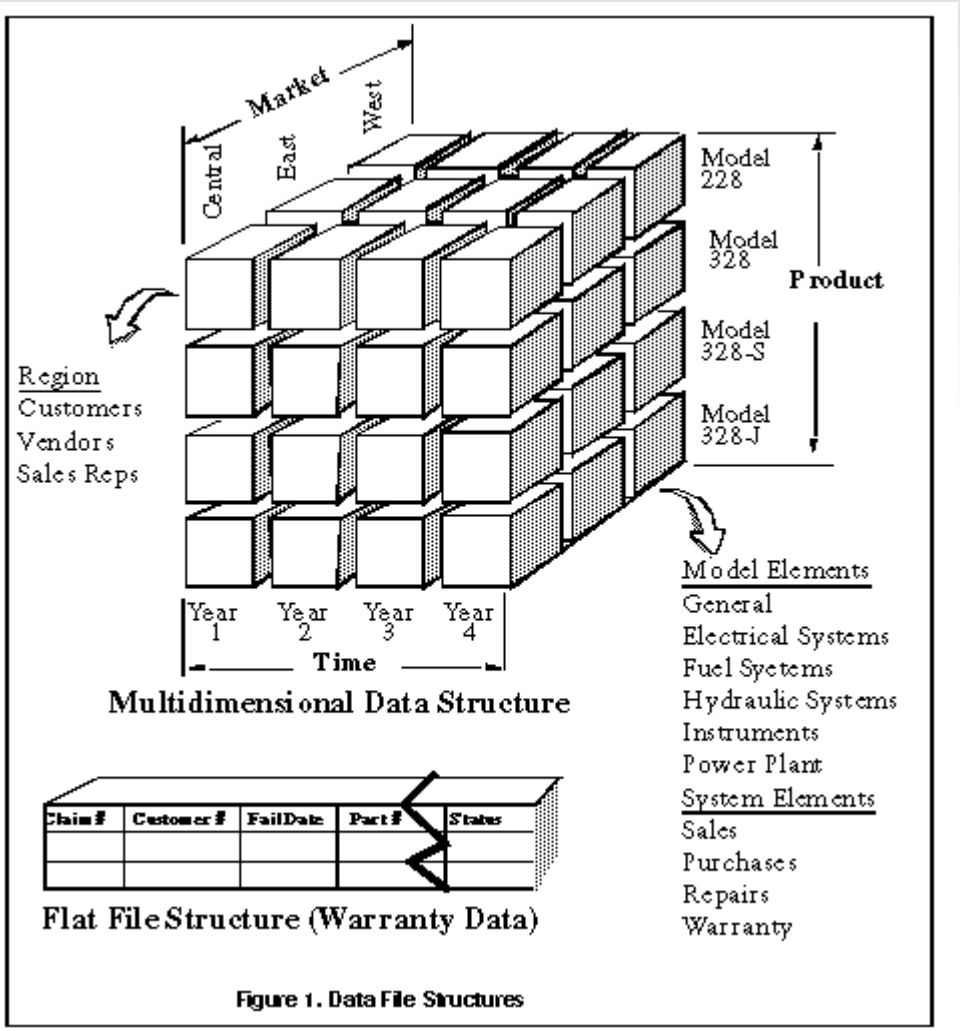


Figure 1. Data File Structures

Modelo de datos

- La habilidad de poder visualizar algo tan abstracto como un conjunto de datos de forma concreta y tangible es la base de la comprensibilidad
- Un modelo simple redunda en un diseño simple
- Un diseño complejo correrá más lento y no será aceptado por los usuarios
- El modelo dimensional es completamente diferente a un modelo 3NF

Modelo 3NF

- Trata de evitar redundancia de datos
- A veces se lo confunde con un modelo ER
- Un modelo dimensional también puede ser expresado con un diagrama *entidad-relación*
- Principal diferencia entre los modelos dimensionales y los modelos 3NF
  - Grado de normalización
- Modelo 3NF → modelo normalizado
- Principales ventajas ?

Desventajas de los modelos normalizados para AD

- Demasiado complejos para el tipo de consultas requerido
- Difícil de comprender para los usuarios
- Bajo rendimiento
- Contrario a los objetivos del AD
  - Obtener información de los datos de forma intuitiva y con un alto rendimiento
- Problemas para el departamento de TI
  - Subutilización de los recursos, consultas delegadas a TI, consultas complejas, usuarios descontentos

Modelo dimensional en el área de presentación

- Contiene la misma información que un modelo normalizado
- Sólo que empaca los datos en un formato cuyos objetivos de diseño son:
  - Comprensibilidad
  - Rendimiento (consultas)
  - Elasticidad

Datos detallados y atómicos

- Aunque los mercados de datos pueden contener datos resumidos y agregados
  - Es necesario conservar la granularidad de los datos

- ¿Por qué?
  - No podemos predecir todas las necesidades presentes y futuras de los usuarios

Mercados de datos

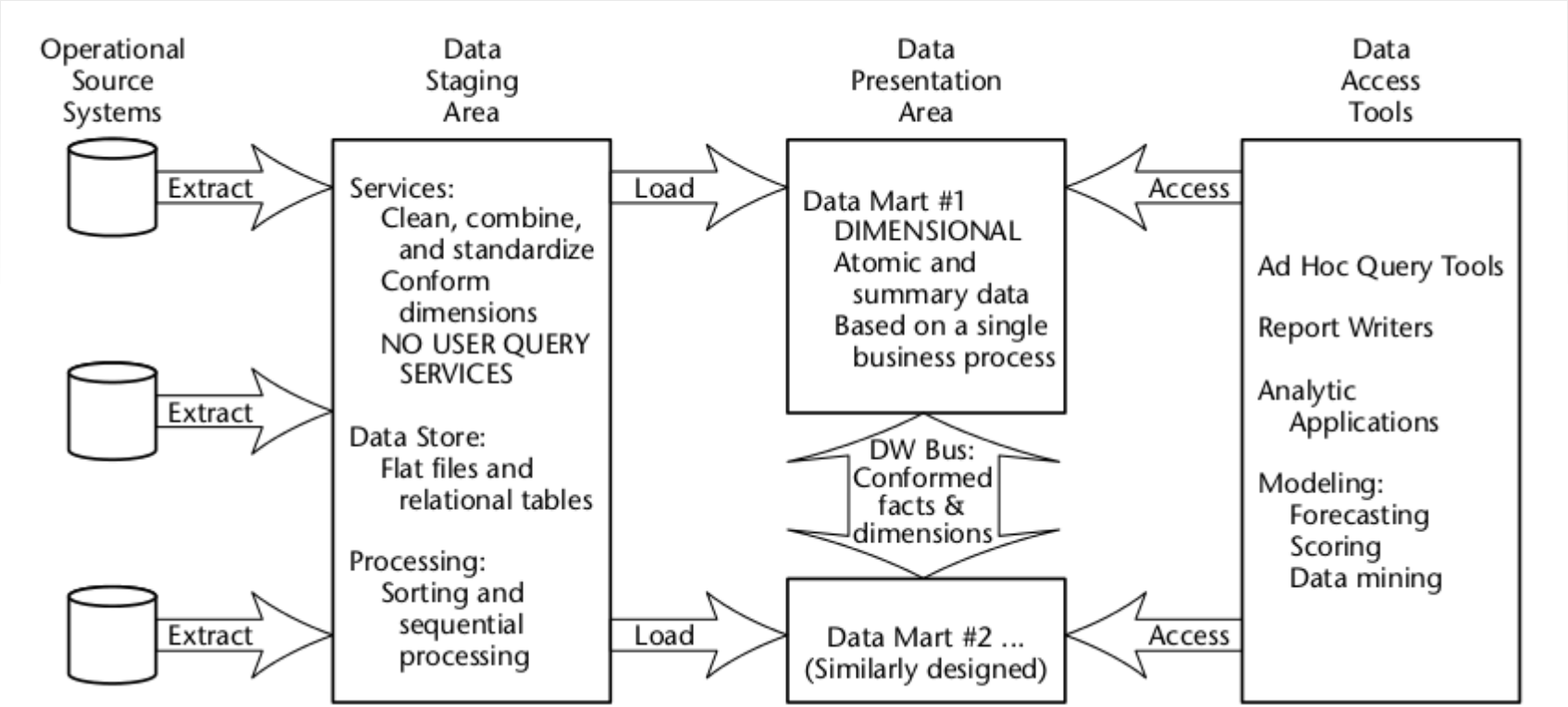
- Hechos y dimensiones comunes en toda la organización
- Base para la construcción de un AD distribuido

- **Los datos en el área de presentación (consultas) del AD deben ser dimensionales, atómicos y uniformes en todos los mercados de datos (*datamarts*)**

Modelado multidimensional

- Si el área de presentación está basada en una base de datos relacional, entonces a las tablas modeladas de forma dimensional se les conoce como **esquema de estrella**
- Si el área de presentación está basada en una base de datos multidimensional o tecnología OLAP, entonces se dice que los datos se guardan en **cubos**
- El modelado multidimensional se aplica tanto a las bases de datos relacionales como a las multidimensionales

Componentes del AD



Herramientas de acceso a datos

- Función principal
  - Consultas al área de presentación
- Puede ser tan simple como un gestor de consultas o tan complejo como una aplicación de modelado o de minería de datos
- La mayoría de los usuarios acceden a los datos por medio de consultas preconstruidas y parametrizadas (el usuario sólo cambia los parámetros)
  - Aproximadamente 80-90 % de los usuarios

Vocabulario del modelado dimensional

- Hechos (facts) y dimensiones
- Origen:
  - Proyecto de General Mills y Dartmouth University en los '60s
- Tablas de hechos
- Tablas de dimensiones

Tabla de hechos

Tabla de hechos

- Tabla primaria en un modelo dimensional donde se guardan las mediciones de rendimiento del negocio
- Las mediciones de un proceso de negocio idealmente se guardan un un solo mercado de datos
- Componentes que más espacio demandan en un mercado de datos
- La lista de dimensiones definen el *grano* de la tabla de hechos y nos dice cuál es el rango de la medición
- Cada fila corresponde a una medición. Una medición es una fila en la tabla de hechos. Todas las mediciones en una tabla de hechos deben ser del mismo grano

Características

- Los hechos más útiles son numéricos y aditivos
- La aditividad es muy importante porque las aplicaciones en AD raramente regresan una sola fila
- Teóricamente, un hecho podría ser textual; sin embargo, esta condición sólo se presenta raramente
- En el caso del ejemplo anterior, el diseñador debe hacer su mejor esfuerzo por poner las mediciones textuales como atributos de dimensiones
- Al menos que el texto sea diferente en cada fila, debe incluirse en una tabla de dimensiones

Características (2)

- En el ejemplo anterior, si un día dado no hay actividad para cierto producto en determinada tienda, la fila no se inserta (no es conveniente insertar ceros)
- Consecuencia:
  - Tablas de hechos son ralas
- Sin embargo,
  - Suelen contener el 90 % o más del espacio consumido por la BD
- Tablas de hechos tienden a tener muchas filas y pocas columnas

Características (3)

- La granularidad de las tablas de hechos puede ser de tres categorías diferentes:
  - Transacción
  - 'Fotografía' periódica
  - 'Fotografía' acumulativa
- Las del primer tipo tienden a ser las más comunes
- Tienen dos o más llaves foráneas que las conectan con las llaves primarias de las tablas de dimensiones
  - Recordar *integridad referencial*
  - Son accedidas haciendo *joins* con las tablas de dimensiones

Características (4)

- Llave primaria de una tabla de hechos es un subconjunto de sus llaves foráneas (llaves compuestas)
- Cada tabla de hecho en un modelo dimensional tiene una llave compuesta
- Si una tabla tiene una llave compuesta, es una tabla de hechos
- Todas las demás tablas son tablas de dimensiones
- **Las tablas de hechos expresan las relaciones de tipo *muchos-a-muchos* entre dimensiones en los modelos dimensionales**

Tabla de dimensiones

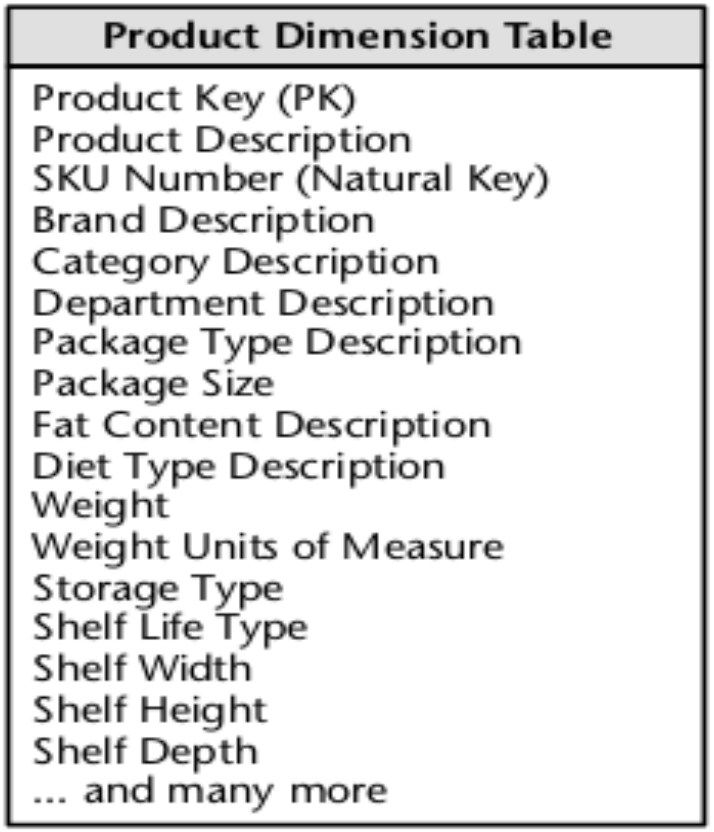


Tabla de dimensiones

- Descriptores textuales del negocio
- En un buen diseño dimensional, tienen muchas columnas o atributos
- Cada dimensión se define por una sola llave primaria
- Atributos son la fuente primaria de restricciones y agrupamientos en las consultas y las etiquetas de los reportes
- Si un usuario quiere ver ventas *por* semana *por* marca, semana y marca deben ser atributos de dimensiones

Características

- Atributos son la clave para que el AD sea usable y comprensible
- **Las tablas de dimensiones son el punto de entrada a las tablas de hechos**
- **Atributos de dimensiones robustos producen capacidades analíticas robustas (para *slicing* y *dicing*)**
- **Las dimensiones implementan la interfaz de usuario del AD**

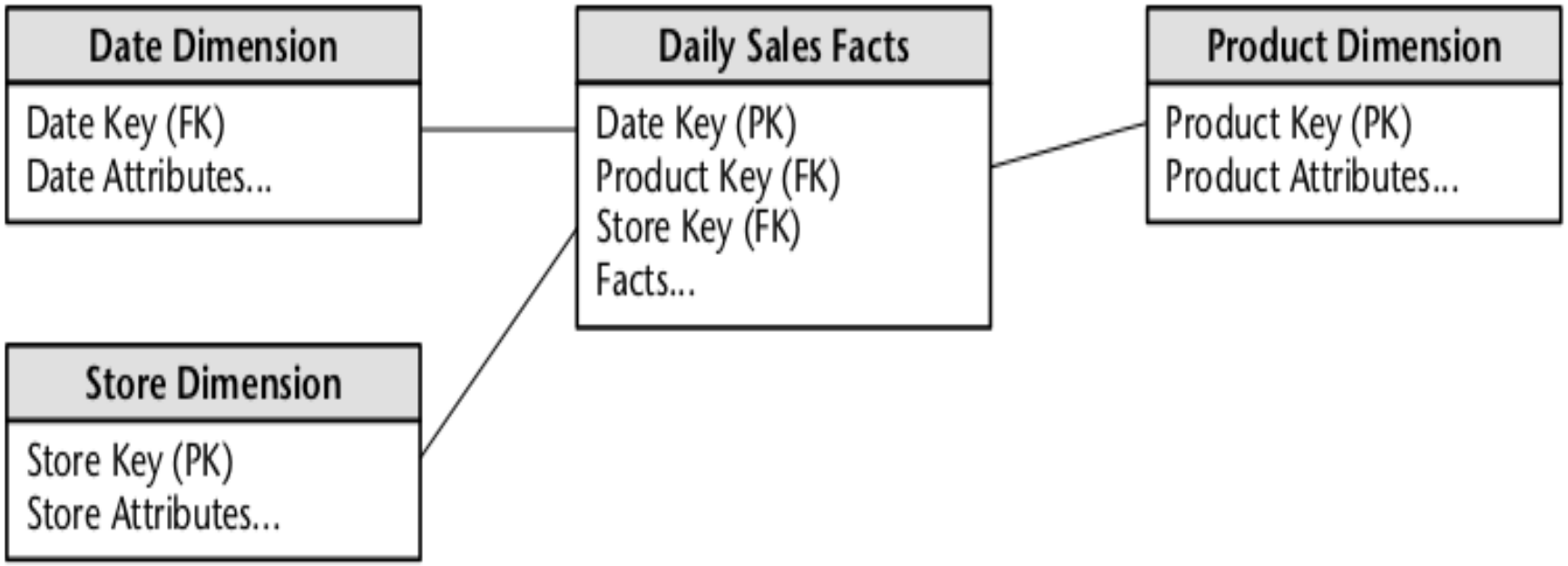
Características (2)

- Los mejores atributos son textuales y discretos
- Deben ser palabras reales
- Ejemplos para un producto:
  - Descripciones corta y larga, marca, categoría, tipo de empaque, tamaño, etc.
- Aunque el tamaño normalmente es numérico, se comporta como una descripción textual y no como una medición numérica
  - Es una descriptor constante y discreto de un producto específico

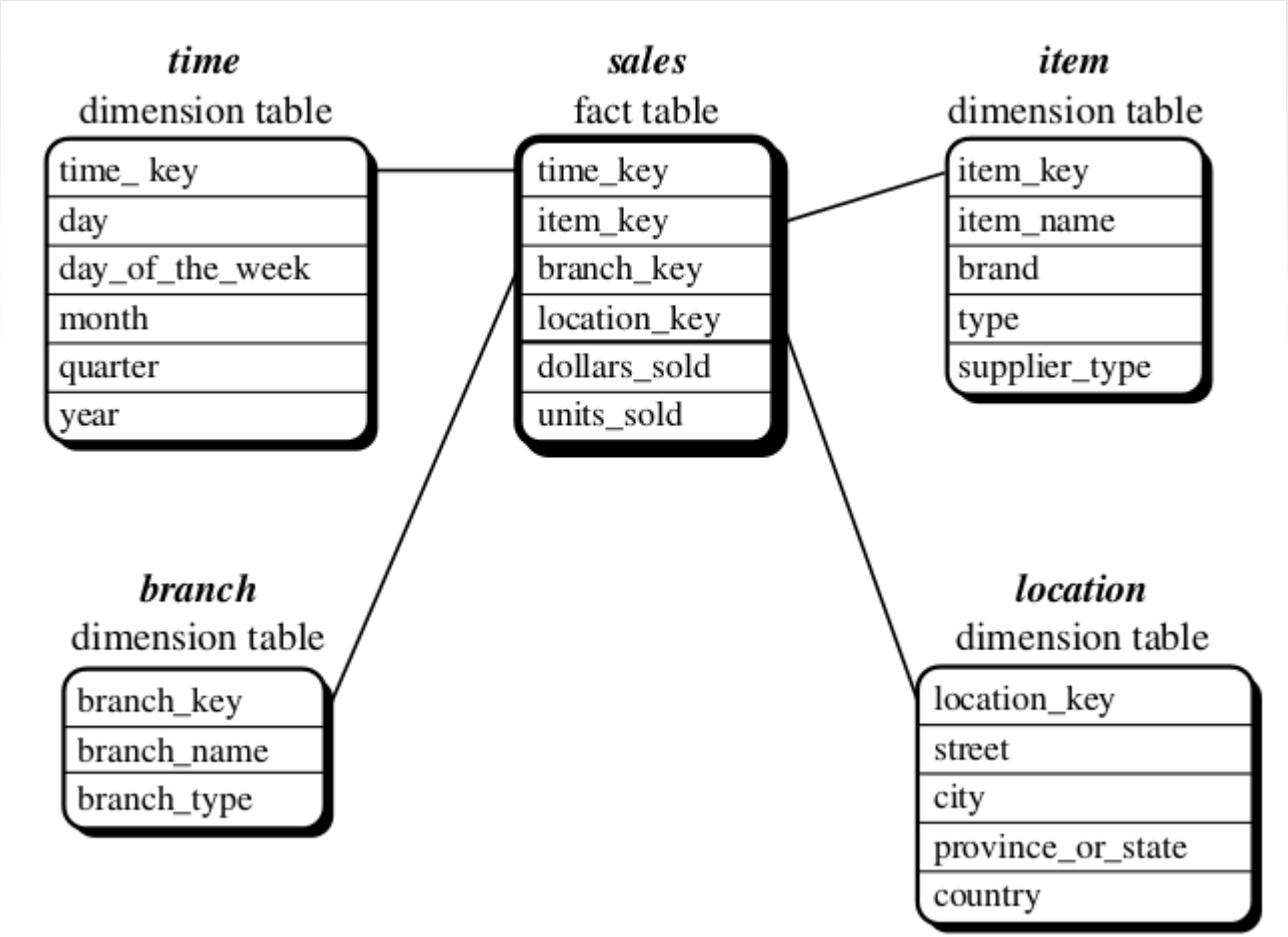
Características (3)

- Muchas veces representan relaciones jerárquicas del negocio
  - Productos incluyen marca y una categoría
  - Esto a veces produce redundancia que es tolerada por facilidad de uso y mejor rendimiento
- Alternativamente, podría utilizarse un código de marca y otra tabla con los nombres de las marcas
  - Esquema de copo de nieve (snowflake)
- Típicamente, altamente denormalizadas
- Usualmente, 10 % del espacio total de almacenaje

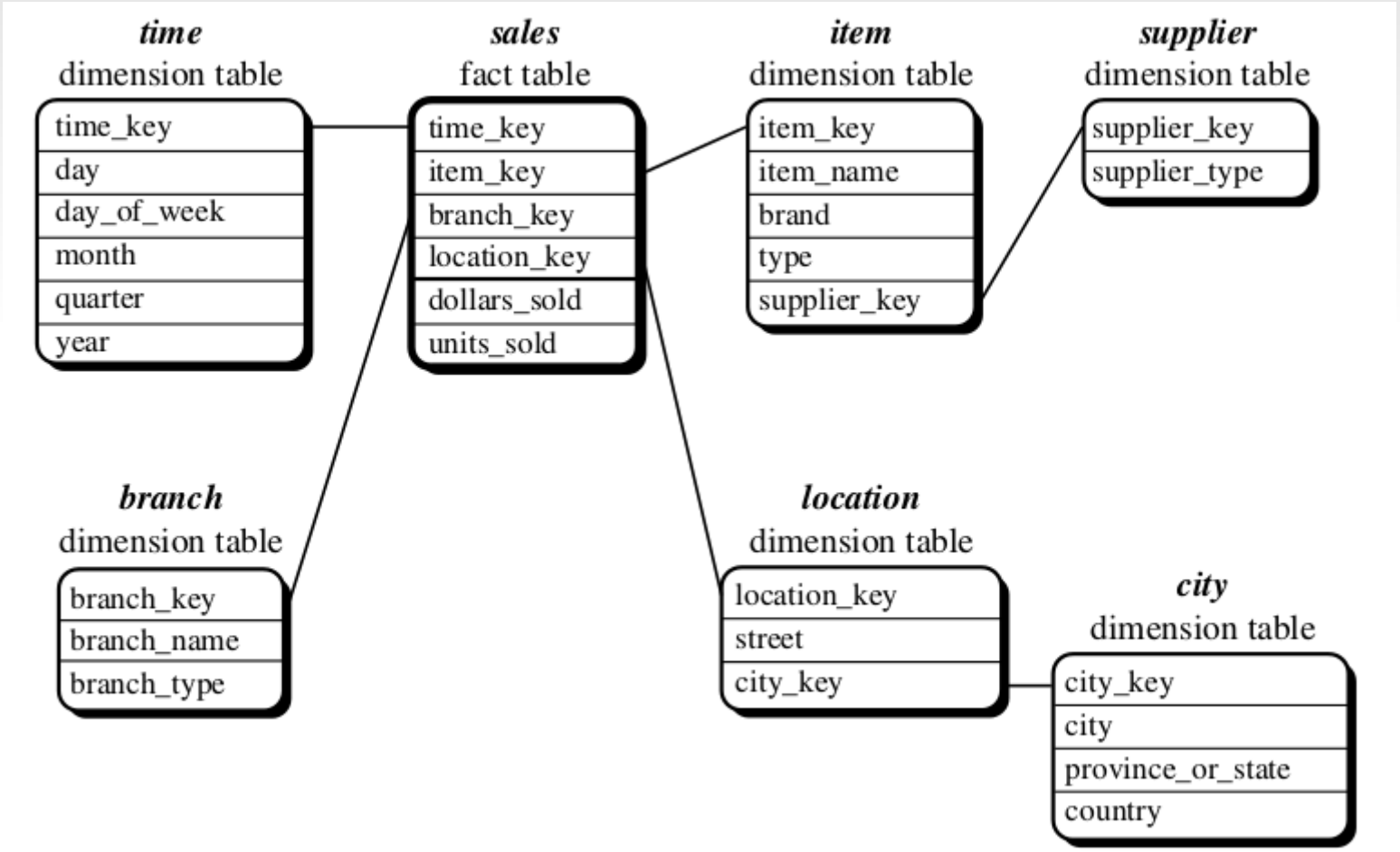
Tablas en un modelo multidimensional



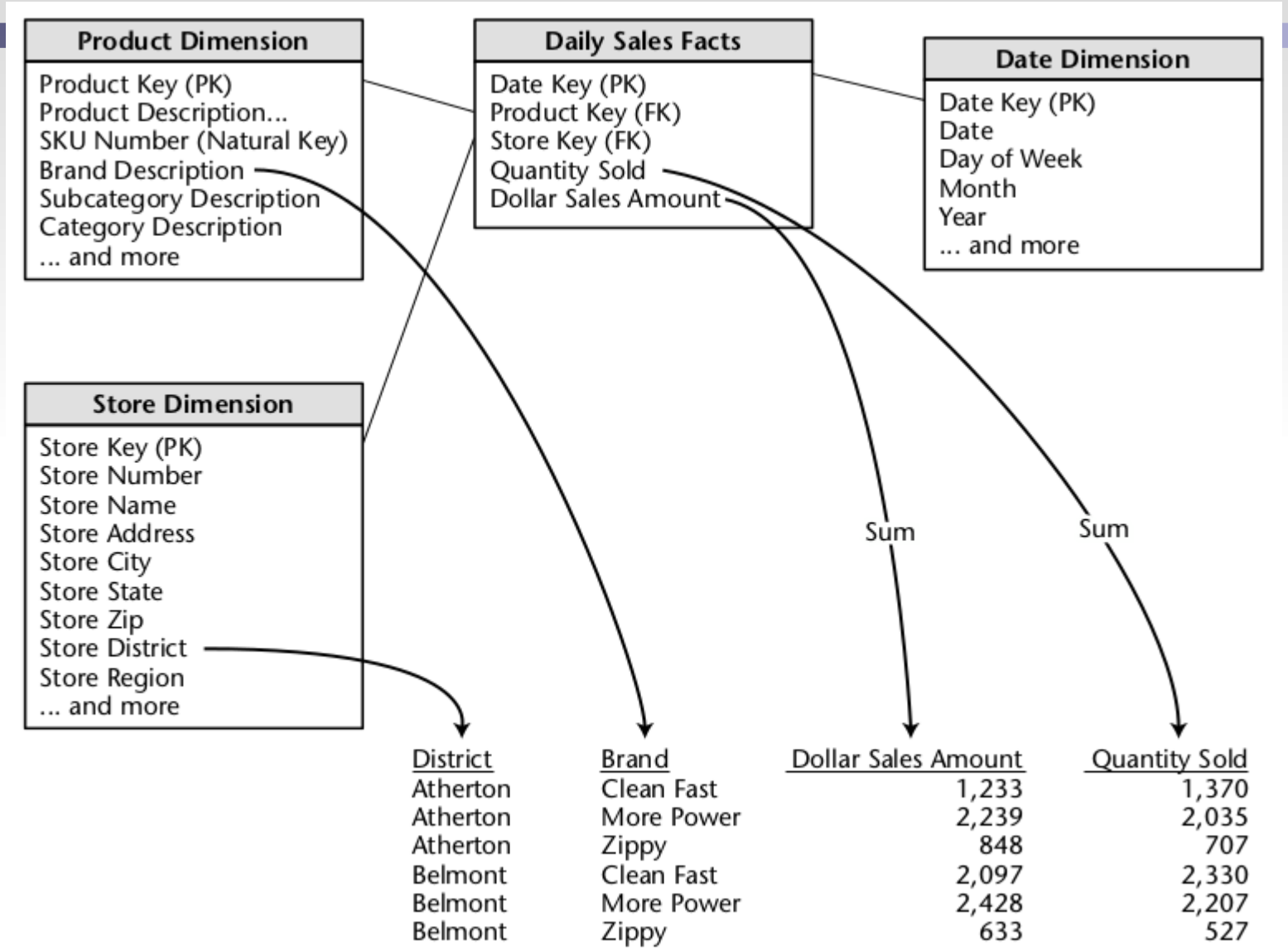
Esquema de estrella



Esquema de copo de nieve



Atributos y hechos en un reporte



Referencias

- Golfarelli, M., Rizzi, S. *Data Warehouse Design: Modern Principles and Methodologies.* McGraw-Hill, 2009
- Jiawey, H., Kamber, M. *Data Mining: Concepts and Techniques (Second Edition)*. Morgan-Kaufmann, 2006
- Kimball, R., Ross, M. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition)*. John Wiley & Sons, 2002