

CC5608 - Inteligencia de Negocios

Clase 3 – Modelamiento dimensional
Arquitectura general de soluciones analíticas

Marcela Calderón Corail - Semestre Otoño 2011

Diferentes mundos de información

Uno de los activos más importantes de una organización es la información.

Este activo es al menos siempre resguardado de dos maneras: Los sistemas operacionales de registro y el Data Warehouse.

Los sistemas operacionales son donde ponemos datos, el Data Warehouse es de donde sacamos datos.

Diferentes mundos de información

Los usuarios de un sistema operacional son el motor de la organización.

- Éstos toman órdenes, ingresan nuevos usuarios y registran quejas.
- Los usuarios siempre trabajan con un registro a la vez
- Trabajan repetidamente realizando las mismas tareas operacionales una y otra vez.

Diferentes mundos de información

Los usuarios de un Data Warehouse observan cómo funciona el motor de la organización.

- Éstos cuentan las nuevas órdenes y las comparan con las de la última semana, preguntan por qué ingresaron nuevos clientes y cuáles eran sus quejas.
- Los usuarios casi nunca trabajan con un registro a la vez
- Frecuentemente sus consultas requieren que cientos o miles de registros sean buscados y comprimidos a un conjunto de respuestas.
- Y para complicarlo más, estos usuarios continuamente cambian el tipo de consultas que hacen.

¿Qué es un Data Warehouse?

Un Data Warehouse es una base de datos utilizada para generar reportes. Es un repositorio que contiene información recolectada proveniente de datos generalmente transaccionales.

- Los datos provienen de una o más fuentes
- Se almacenan centralizadamente para ser posteriormente reportados de forma agregada.
- La función del Data Warehouse es el almacenamiento pero a veces se habla “Sistemas de Data Warehouse” para referirse no sólo al almacén de datos si no que a las herramientas de staging, integración, y acceso y las herramientas de Inteligencia de Negocios (reportes, Data Mining, tableros, cubos, etc).

¿Qué es un Data Mart?

Los Data Marts son colecciones de datos más reducidas o subconjuntos del Data Warehouse que sirven el propósito de implementar análisis más específicos sobre los datos.

- Son subconjuntos de datos centrados en un tema o un área de negocio específico

Corrientes diferentes para DW

Kimball

v/s

Inmon

Enfoque dimensional, el DW debe ser modelado dimensionalmente o con un esquema estrella	Enfoque normalizado, los datos en un DW son almacenados siguiendo reglas de normalización en un modelo E-R
Los datos transaccionales son particionados en “hechos” o “dimensiones”	Las tablas se agrupan por temas , los datos se separan en entidades que crean diversas tablas en una base de datos relacional
Es un modelo más simple de comprender y usar	Resulta en muchas tablas unidas por una red de uniones
La recuperación de datos es veloz	Es muy simple agregar nueva información
Es complejo de cargar con datos de diferentes sistemas operacionales	Dado el alto número de tablas dificulta unir diferentes fuentes hacia información útil
Es difícil modificar su estructura	Es difícil acceder información sin entender el modelo

Corrientes diferentes para DW

Kimball

v/s

Inmon

Diseño bottom-up	Diseño top-down
Se crean primero los data marts para proveer de reportería y capacidades analíticas a procesos de negocio específicos	Las tablas se agrupan por temas , los datos se separan en entidades que crean diversas tablas en una base de datos relacional.
Los Data Marts tienen principalmente hechos y dimensiones. Cada uno modela un área específica del negocio	Es un gran repositorio central con la información de la empresa completa.
Los Data Marts se integran en una colección de dimensiones conformadas.	El DW contiene los datos atómicos. Se crean Data Marts a partir del DW.
El DW termina siendo segmentado en un conjunto de data marts consistentes y autocontenidos.	El DW es orientado a temas, es no volátil, es integrado y varía en el tiempo.
Se obtienen resultados en etapas tempranas.	Representa un proyecto muy amplio y de largo alcance.

Objetivos de un Data Warehouse

Los objetivos para un DW se pueden desarrollar caminando por los pasillos de cualquier organización escuchando de gestión del negocio. Los temas recurrentes son:

- “Tenemos montones de datos en esta compañía pero no tenemos acceso a ellos”.
- “Necesitamos cortarlos y combinarlos (slice and dice)”
- “La gente de negocio necesita llegar directamente a los datos”
- “Sólo necesito ver lo importante”
- “Me vuelve loco dos personas en una reunión que presentan las mismas métricas del negocio con números diferentes”

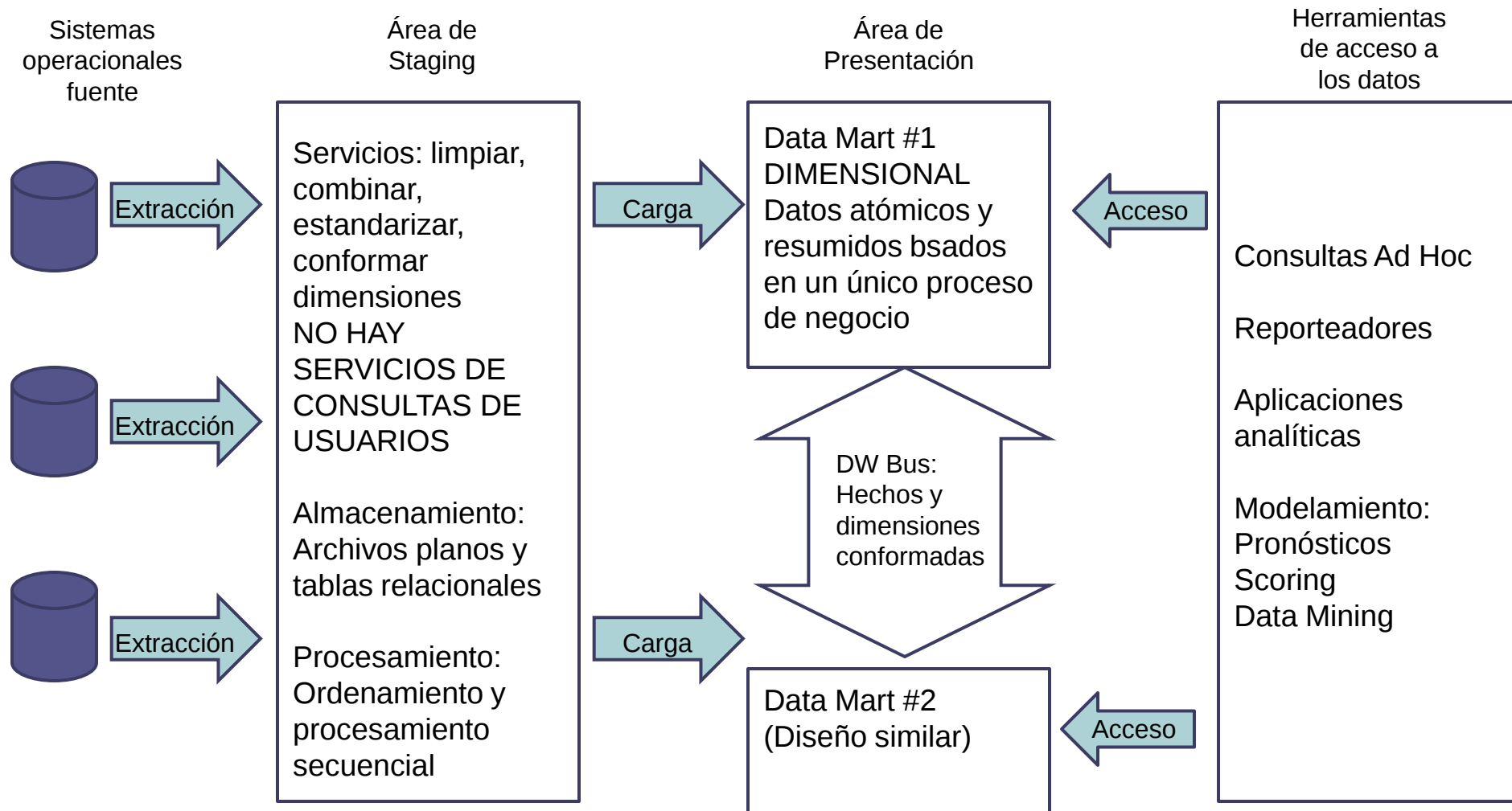
Características que debe cumplir un DW

- El DW debe hacer que la información de la organización sea fácilmente accesible.
 - El contenido debe ser entendible, los datos deben ser intuitivos y obvios para el usuario de negocio. Las herramientas para acceder al DW deben ser simples y permitirle al usuario realizar todas las combinaciones que desee y responder muy rápidamente.
- El DW debe presentar la información de la organización de forma consistente.
 - Los datos deben ser confiables, cuidadosamente compuestos desde una variedad de fuentes, limpiados, revisados.
- El DW debe ser adaptativo y resiliente
 - Los cambios en el DW deben incorporarse sin impacto en los datos existentes y las aplicaciones.

Características que debe cumplir un DW

- El DW debe resguardar la información
 - Debe controlar el acceso a los datos de manera efectiva. Son datos generalmente confidenciales.
- El DW debe servir como la base para la mejora en la toma de decisiones.
 - Debe contener los datos correctos para esto. El resultado de un DW, su output, son las decisiones que se toman en base a la evidencia que éste entrega.
- El DW debe ser aceptado por el negocio
 - Si los usuarios de negocio no lo usan activamente, es un fracaso. El DW generalmente es opcional, su aceptabilidad tiene que ver con su simplicidad generalmente.

Componentes de un DW



Modelamiento dimensional

- **Tabla de Hechos:** Es la tabla principal del modelo donde se almacenan las métricas numéricas de desempeño.
 - Las métricas sólo se almacenan acá.
 - Se usa la palabra Hecho para representar una medida del negocio.
 - En la feria podemos registrar la fecha, unidades vendidas y su precio en pesos para cada producto vendido en cada puesto. Una medida se toma en una intersección de todas las dimensiones (día, producto y puesto), esto determina el nivel de detalle de la tabla de hechos.
- Una fila en la tabla de hechos es una medida. Una medida es una fila en la tabla de hechos. Todas las medidas en una tabla de hechos deben estar el mismo nivel de detalle.
- Las métricas más útiles son numéricas y aditivas y responden a hechos reales.
- Las tablas de hechos expresan las relaciones muchos a muchos entre dimensiones en los modelos dimensionales.

Modelamiento dimensional

- **Tabla de Hechos:** Un ejemplo

Tabla de Hecho de Ventas Diarias

Fecha Id (FK)
Producto Id (FK)
Tienda Id (FK)
Unidades vendidas
Monto en pesos

Modelamiento dimensional

- **Tablas de Dimensiones:** Acompañan a la tabla de hechos y le dan el contexto a la información.
 - Contienen los descriptores textuales del negocio.
 - Tienen muchas columnas o atributos que describen a las filas en la tabla de dimensión.
 - Los atributos sirven como fuente principal de restricciones para las consultas, agrupaciones y etiquetas de reportes.
 - Son los “por”. Ventas por semana, por marca, etc. Deben ser atributos de dimensiones.
- Los atributos de las tablas de dimensiones son vitales para los DW, proveen todas las restricciones interesantes y la información para reportes, hacen que la información sea usable y comprensible.
- Es muy importante darle a los atributos nombres verbosos, poblarlos con valores y asegurar su calidad.

Modelamiento dimensional

Tablas de Dimensiones

- Las tablas de dimensiones son los puntos de entrada a la tabla de hechos. Si sus atributos son robustos, permiten análisis robustos.
- Los mejores atributos son textuales y discretos.

Modelamiento dimensional

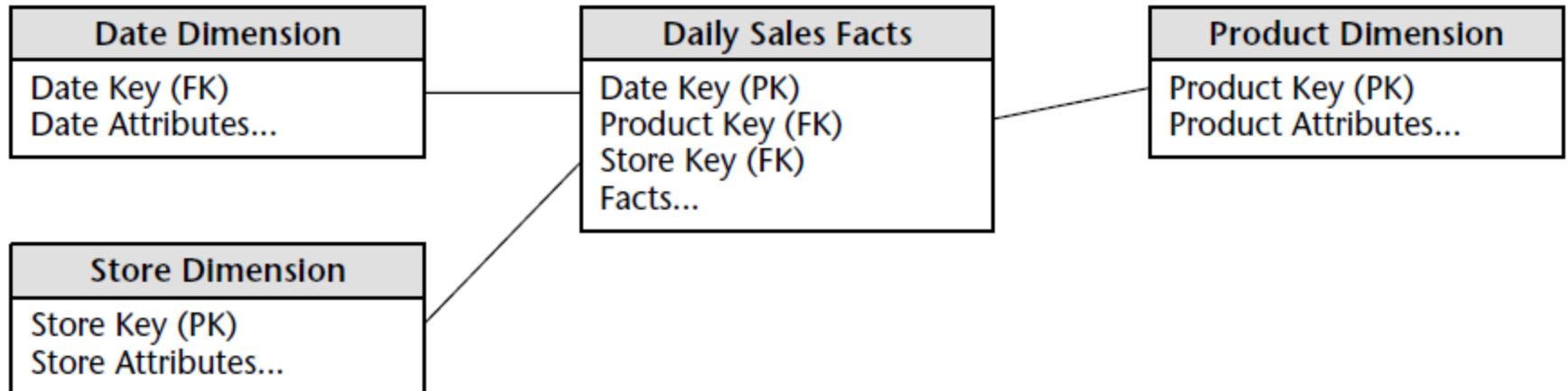
Tablas de Dimensiones: Un ejemplo

Tabla de Dimension Producto

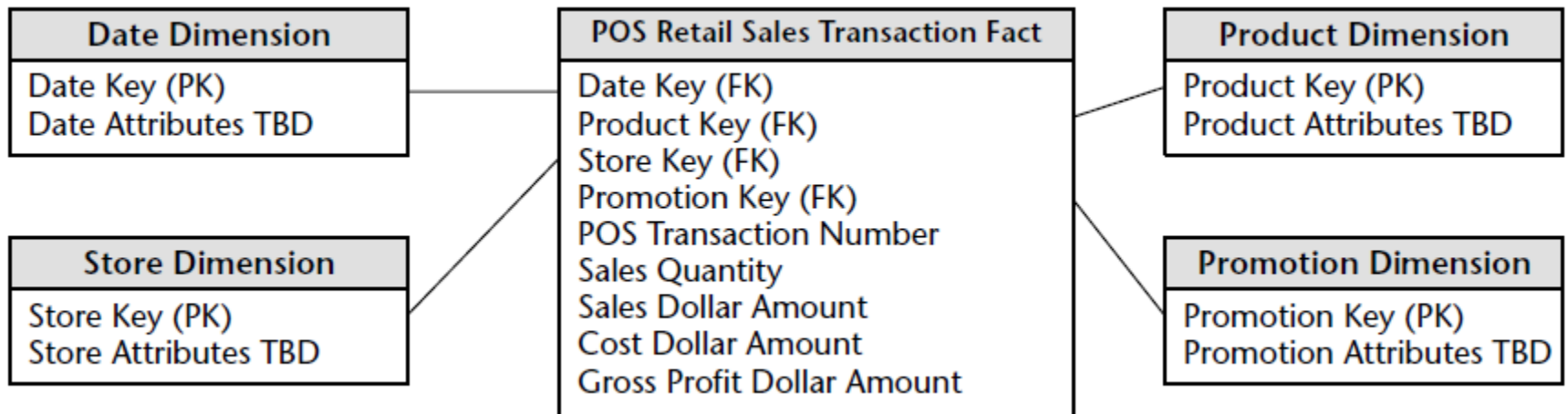
Producto Id (PK)
Descripción
Código SKU (llave natural)
Marca
Categoría
Departamento
Tipo de empaque
Tamaño
Contenido graso
Peso
Tipo de almacenamiento
Tipo de estante
y muchos más

Modelamiento dimensional

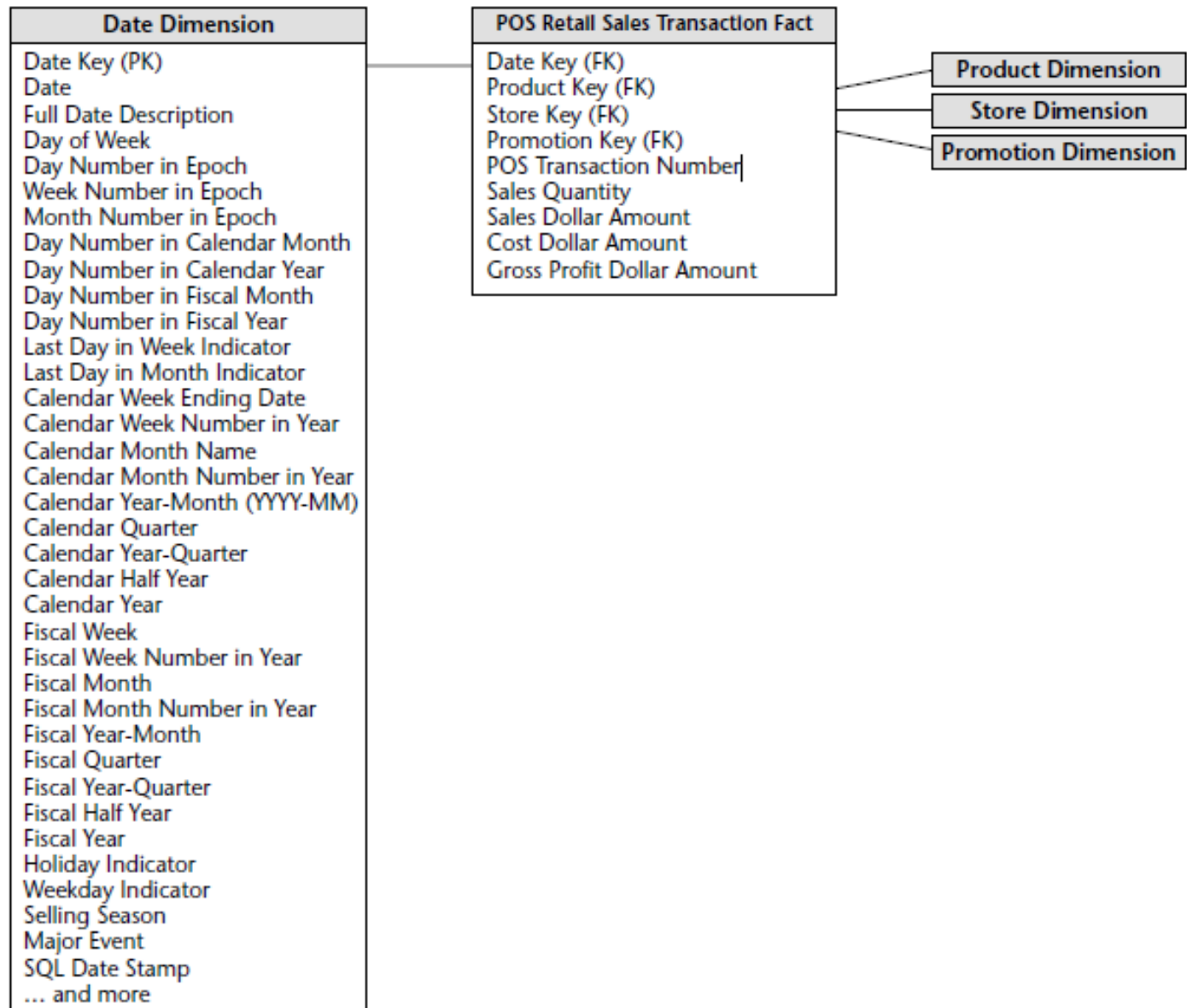
Juntando hechos y dimensiones



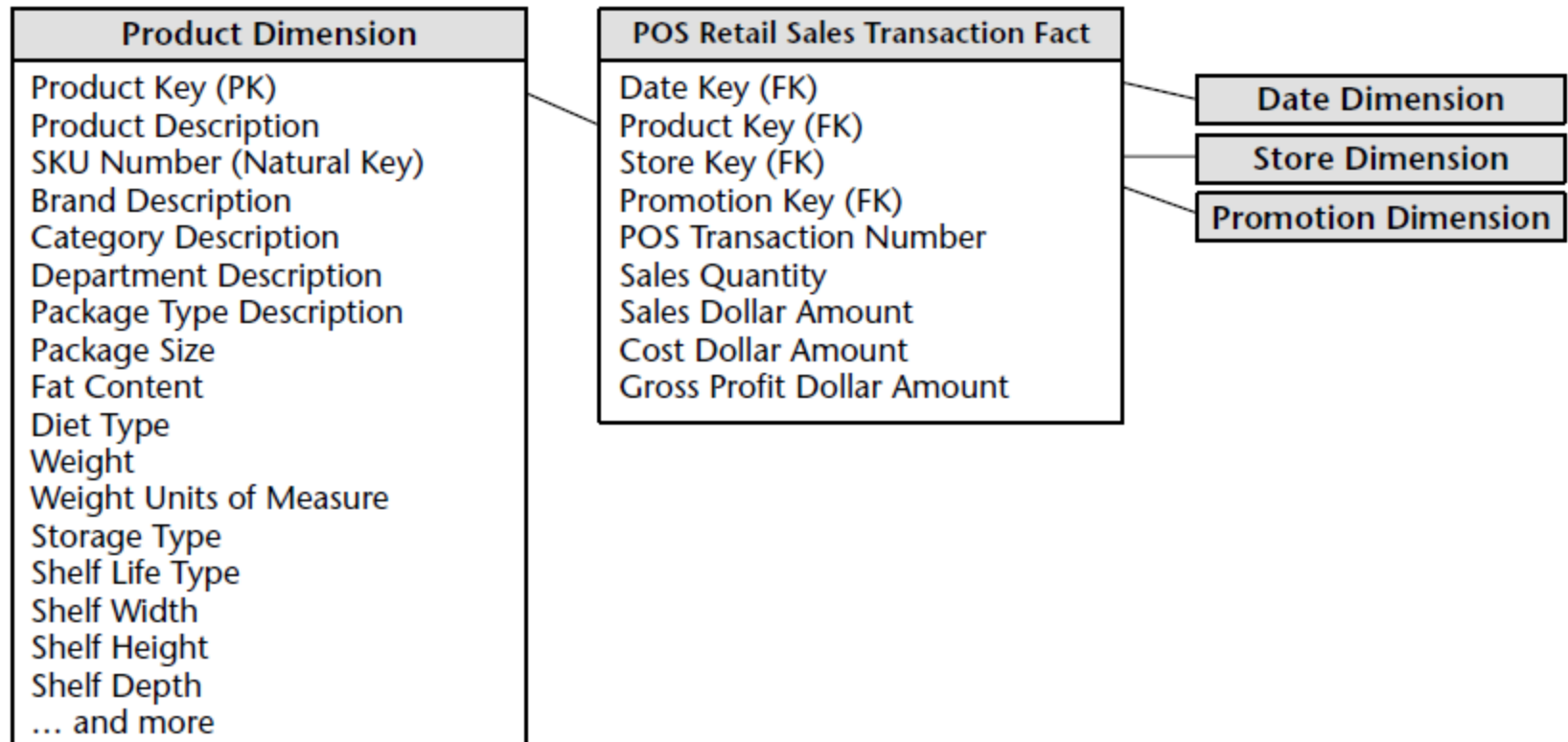
Modelamiento dimensional, ventas de retail



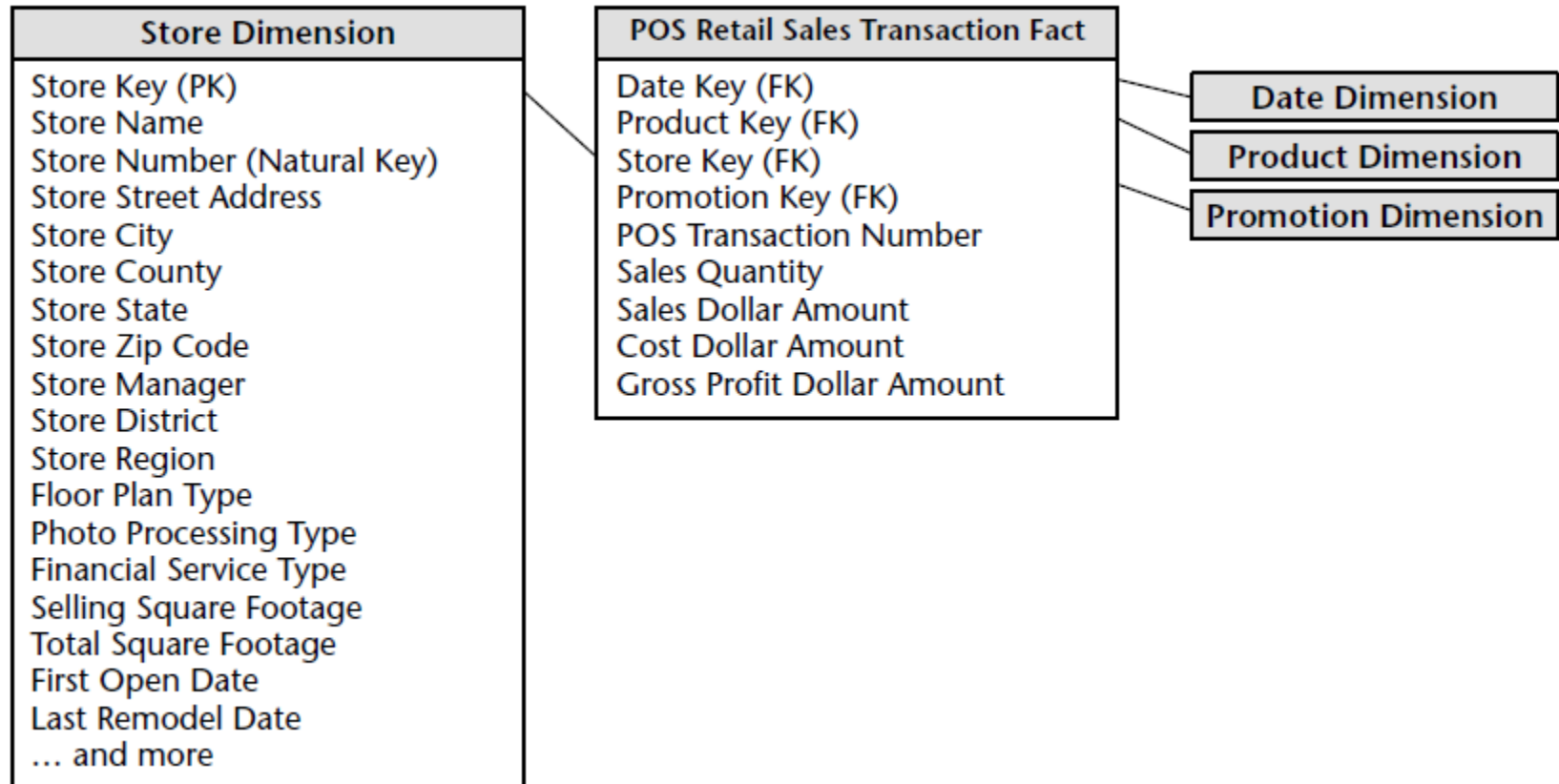
Modelamiento dimensional, ventas de retail



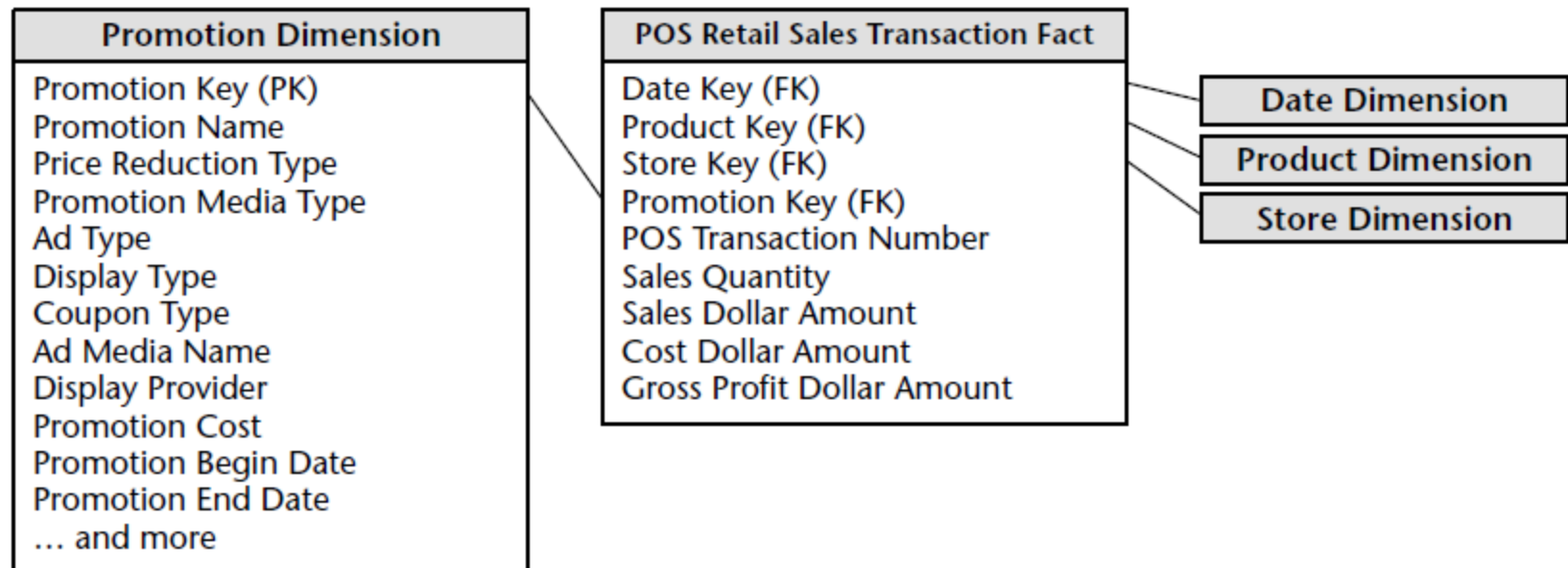
Modelamiento dimensional, ventas de retail



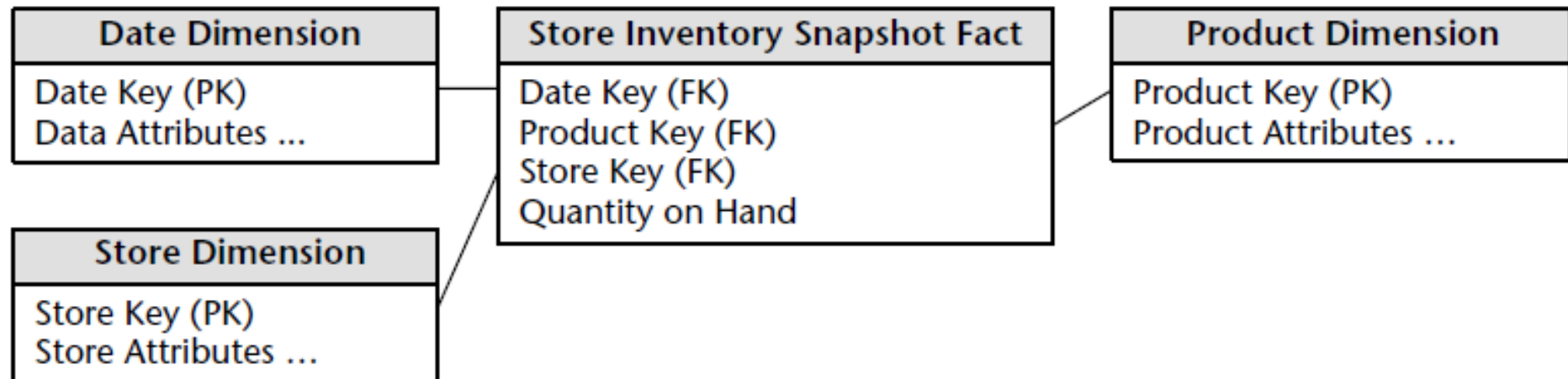
Modelamiento dimensional, ventas de retail



Modelamiento dimensional, ventas de retail



Modelamiento dimensional, otro problema: Inventario



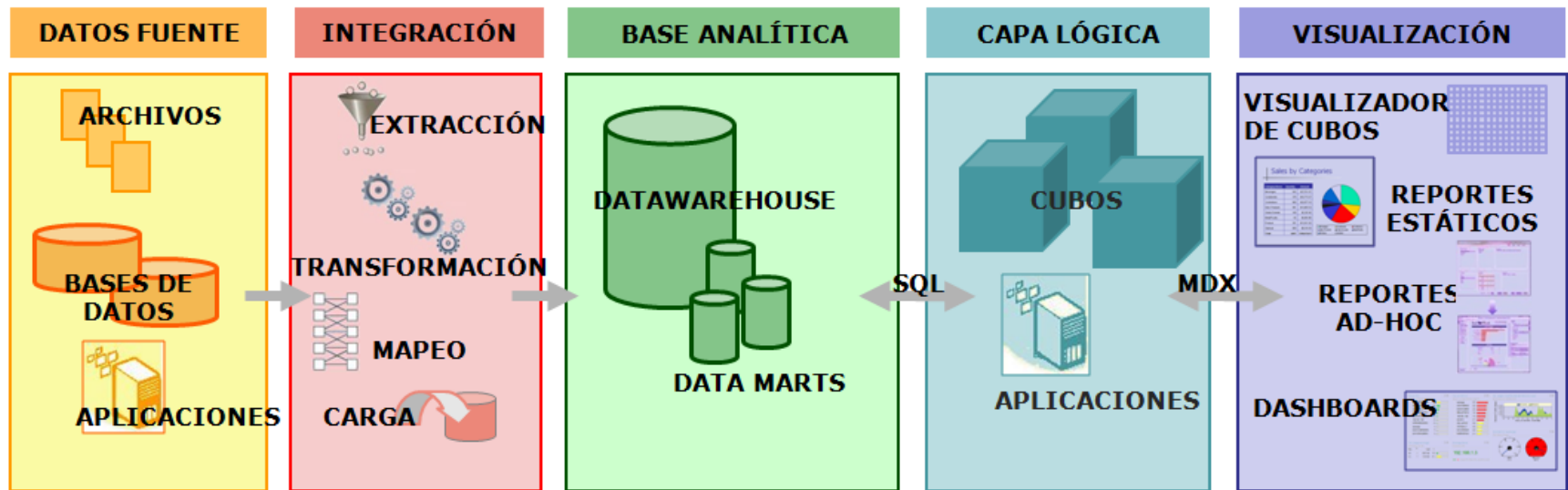
Modelamiento dimensional

Otros:

- Adquisiciones
- Gestión de órdenes
- Gestión de clientes
- Contabilidad
- Servicios financieros
- Telecomunicaciones
- Recursos humanos
- Educación
- Seguros
- etc

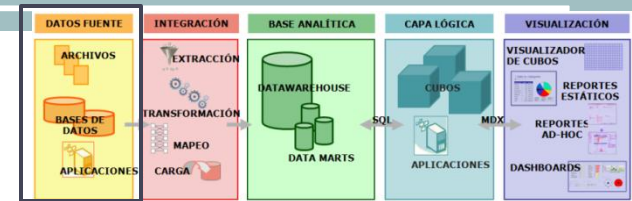
Arquitectura general de soluciones analíticas

Las soluciones de Business Intelligence se construyen sobre una arquitectura general básica que comprende las siguientes componentes:



Datos → Información → Conocimiento → Decisión

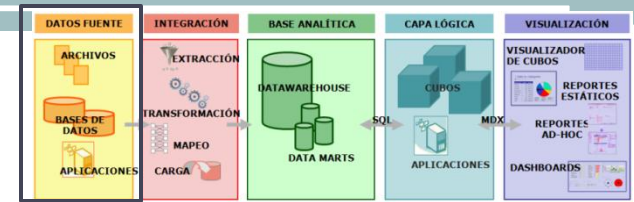
Datos fuente



La primera capa considera todas las fuentes de datos relevantes para el sistema de BI a implementar.

1. Las fuentes de datos pueden tener variado origen:
 - ✓ Datos operacionales procedentes de Bases de Datos corporativas. Generalmente provenientes de ambientes transaccionales OLTP (On-Line Transaction Processing)
 - ✓ Datos departamentales en Bases de Datos propietarias
 - ✓ Datos generados por sistemas de información
 - ✓ Datos privados
 - ✓ Sistemas externos (Internet, Bases de Datos comerciales, de clientes, información económica, etc..)

Datos fuente



La primera capa considera todas las fuentes de datos relevantes para el sistema de BI a implementar.

2. Estas pueden encontrarse en diversos medios y formatos

- ✓ Archivos
- ✓ Bases de Datos
- ✓ Internet
- ✓ Salida de datos de aplicaciones que además pueden estar en ambientes, servidores y lugares físicos distintos.
- ✓ Otros.