MODELADO DE DATOS

DATA WAREHOUSE

Data Warehouse - Inmon

- Orientado al Sujeto (entidades).
  - Los datos de las bases de datos están organizados de manera que todos los elementos de los datos relativos al mismo evento u objeto del mundo real están unidos entre sí
- Variable en el Tiempo:
  - Los cambios en los datos en la BD son registrados y se les da seguimiento para que los informes que se generen muestren los cambios en el tiempo
- No-Volatil
  - Los datos en las Bases de datos nunca debe ser sobre-escritos o borrados, una vez que se registraron, los datos deben ser estáticos, de solo lectura, pero conservarlos para reportes futuros

Data Warehouse - Inmon

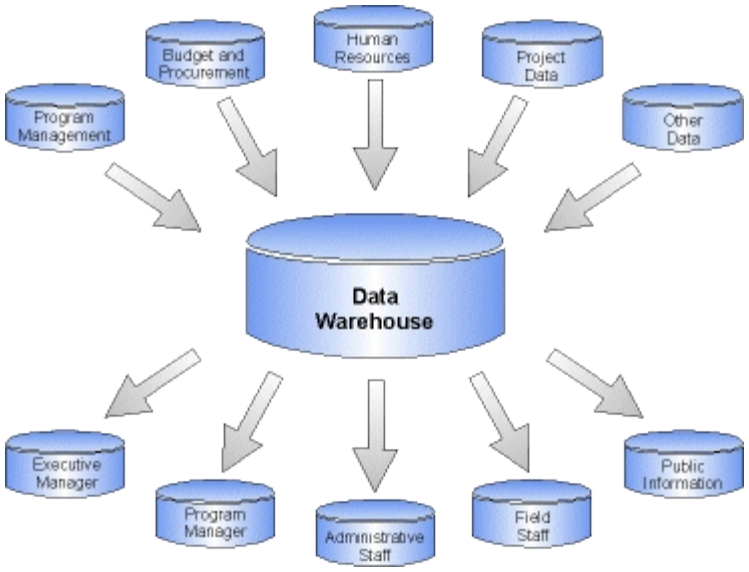
- Integrado:
  - La base de datos contiene datos de la mayoría o la totalidad de las aplicaciones operativas de un organización y estos datos deben ser consistentes.
- No Virtual:
  - El datawarehouse debe existir físicamente, no deben ser un conjunto de vistas.

Data Warehouse – Kimball

- El data warehouse consiste en la unión de todos los data marts.
- Indica una metodología Bottom-Up de Data warehousing
  - **Paso 1**. Identificar Áreas de Negocio. Identificar departamentos
  - **Paso 2**. Entrevista con los jefes de cada área, para levantamiento de requerimientos
  - (Necesidades de negocio muy específicas por cada área) hacer entrevistas no solo con jefes sino con los expertos del área.
  - Anotar todas las preguntas que se hace cada área, qué es lo que quisieran conocer.
    - Se requiere que toda la información que se desea exista en Bases de datos, o en archivos de texto

Data Warehouse – Definición

Repositorio único de la organización que contiene todos los datos de la empresa a través de muchas o todas las líneas de negocio.



¿Porqué se necesita un DWH?

- Toda la información este en un solo lugar.
- Información actualizada de la operación de la empresa.
- Acceso rápido
- Sin limitantes de tamaño
- Toda la historia disponible
- Facilidad de entendimiento de la información.
- Definiciones Uniformes y claras: Simplifica la comunicación, se eliminan discusiones acerca de lo que los datos significan.
- Estandarización de datos.

¿Qué es un Data Mart?

- Es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica.

Data Marts - Ventajas

- Poco volumen de datos
- Mayor rapidez de consulta
- Consultas SQL y/o MDX sencillas
- Validación directa de la información
- Facilidad para la validar información histórica

Data Warehouse VS Data Marts

Data Warehouses	Data Marts
Alcance	**Alcance**
* Centralizada o Empresa	* Descentralizada por grupo (Departamento)
Datos	**Datos**
* Histórica, Detallada, resumida	* Algo de historia, Detallada, resumida
Origen	**Origen**
* Muchos origenes internos y externos	* Pocos origenes internos y externos
Otros	**Otros**
* Felxible	* Restrictivo
* Orientado a datos	* Orientado a Proyecto
* Pocas estructuras complejas	* Múltiples estructuras simples que pueden formar una estructura compleja

Modelado Inmon

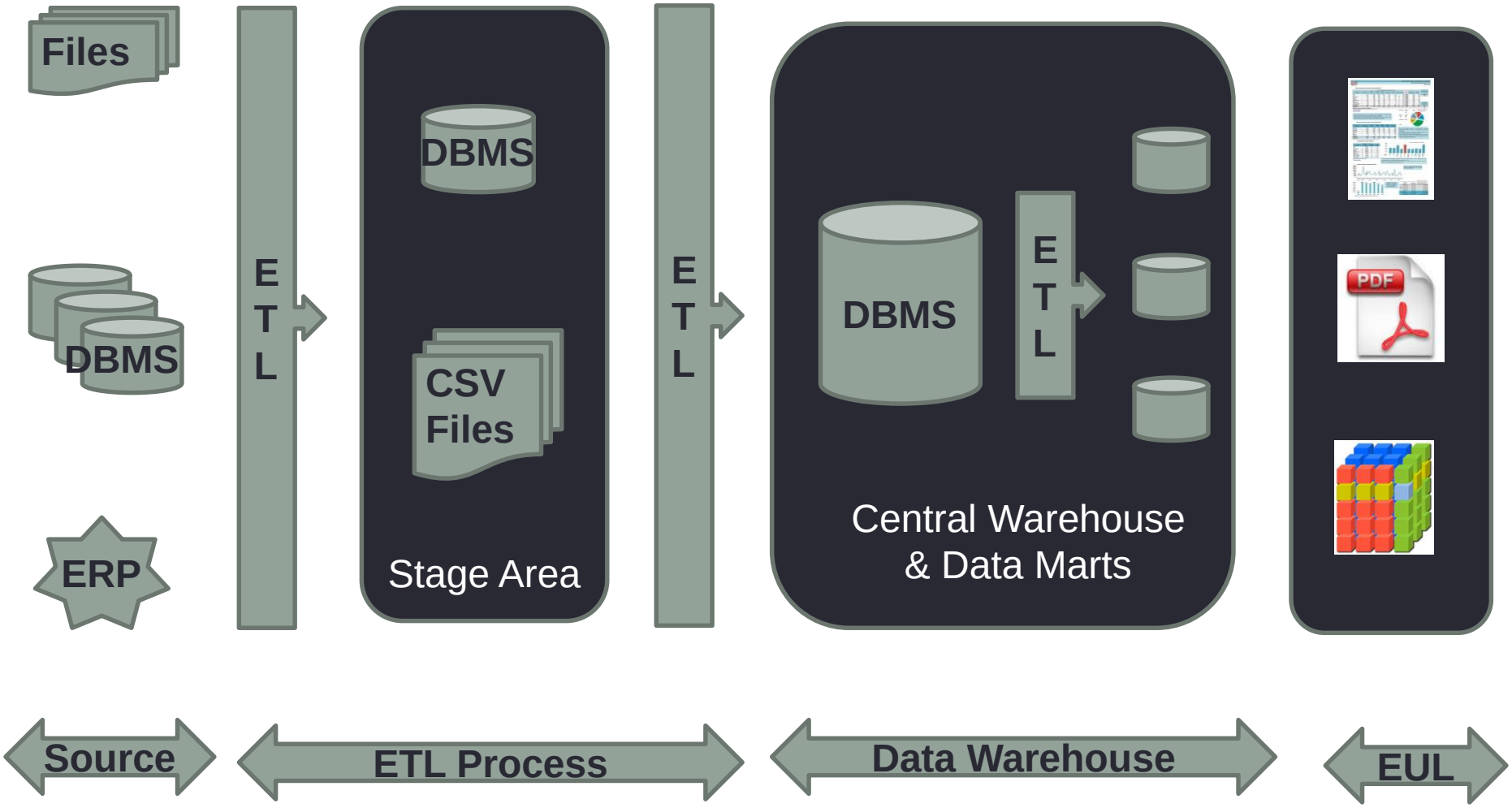
Tres niveles de modelado de datos:

- ERD (Entity Relationship Diagram)
  - Entidades, atributos y relaciones
- Modelo de capa media
  - Conjunto de datos por departamento
  - Cuatro constructores:
    - Agrupación de datos primarios
    - Agrupación de datos secundarios
    - Conectores
    - Tipos de datos
- Modelado Físico de datos
  - Optimización para performance (Normalización)

Modelado de Datos - Kimball

- Iniciar con tablas
  - Hechos
  - Dimensiones
- Hechos contienen las métricas
- Dimensiones contienen atributos
  - Pueden contener grupos repetidos
- De-normalización
- Accesible a Usuarios

Diagrama Arquitectura General DWH



Staging Area

- Cada solución de Data Warehouse debería utilizar una área de Stage donde los datos sean extraídos y transformados antes de ser cargados al Warehouse Central.

Motivos para usar un Stage:

1) Los tiempos de carga del sistema origen debería mantenerse en un mínimo absoluto, por lo que cuando los datos se extrae de las tablas del sistema de origen, es una buena práctica para copiar los datos "tal cual" con las tablas de stage tan pronto como sea posible.

Staging Area

- El uso de un área de stage independiente permite trabajar en un subconjunto específico de los datos, o para ser más específicos, sólo los datos que se requiere para la ejecución actual.
- Un Schema dedicado permite un ordenamiento específico para la optimización y soporte del proceso de ETL

Staging Area

- Unos autores recomiendan realizar una copia idéntica del OLTP
- Otros recomiendan una copia de OLAP
- Otros dicen que no es ni la OLTP ni la OLAP sino un diseño especial llamado Data Vault,

“Dependerá del diseño, arquitectura e infraestructura de la empresa”

Data Warehouse - Desafíos

a) Calidad de los datos:

- Datos duplicados:
- Datos incompletos
- Datos incorrectos
- Datos en conflicto
- Datos perdidos.
- Valores Nulos (No aplicable, desconocido, erróneo)

Data Warehouse - Desafíos

b) Data Vault:

Data Vault es una técnica de modelado de dimensional, basado en el concepto que cualquiera de los datos pertenece a uno de los tres tipos de entidades:

- Hubs (Atributos llave de las entidades del negocio)
- Links (relaciones entre los hubs)
- Satellites (los atributos relacionados a hubs y links)

Data Warehouse - Desafíos

b) Data Vault – Ejemplo:

Empleados	
empleado_id	int(8)
nombre	varchar(30)
ape_pat	varchar(30)
ape_mat	varchar(30)
ciudad_nombre	varchar(64)
fec_nac	date
genero	smallint
region_id	int(8)
domic1	varchar(50)
domic2	varchar(50)
cod_postal	varchar(10)
email	varchar(64)
tel	char(12)

Almacen	
almacen_id	smallint
ciudad_nombre	varchar(64)
domic1	varchar(50)
domic2	varchar(50)
cod_postal	varchar(10)
email	varchar(64)
tel	char(12)

Tareas	
tarea_id	smallint
descripcion	varchar(64)
titulo	varchar(50)

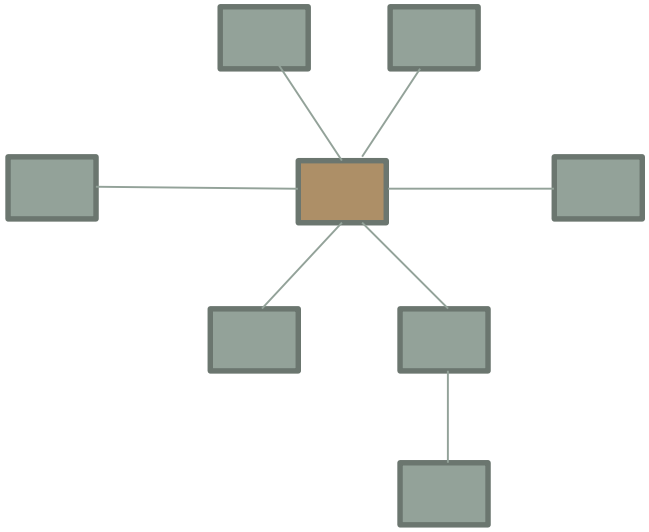
Data Warehouse - Desafíos

c) Volumen de Datos y Performance:

- Data Warehouse pueden llegar a contener Petabyte de información (1024 terabytes = 1 Petabyte)
- El desafío consiste en tratar de analizar esta cantidad de datos por medio de consultas que no afecten el performance.
- Soluciones:
  - Índices
  - Particionamiento
  - Tablas de agregados
  - Vistas Materializadas
  - Funciones de Ventana
  - Archiving

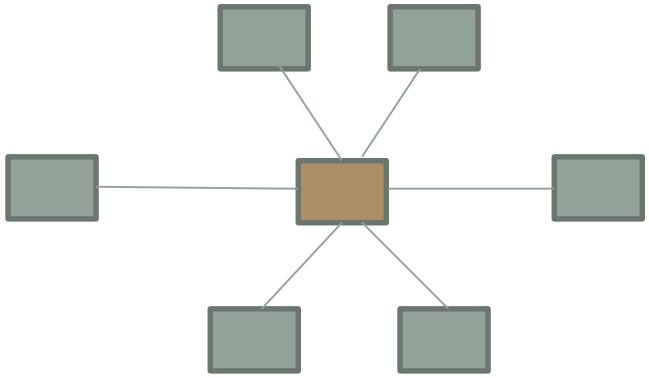
Inmon

- Propone concepto de Data Warehouse
- Maneja la tabla de Hechos (Denormalizada)
- Maneja las tablas de dimensiones (Normalizadas)
- Sugiere el uso de esquema Copo de Nieve

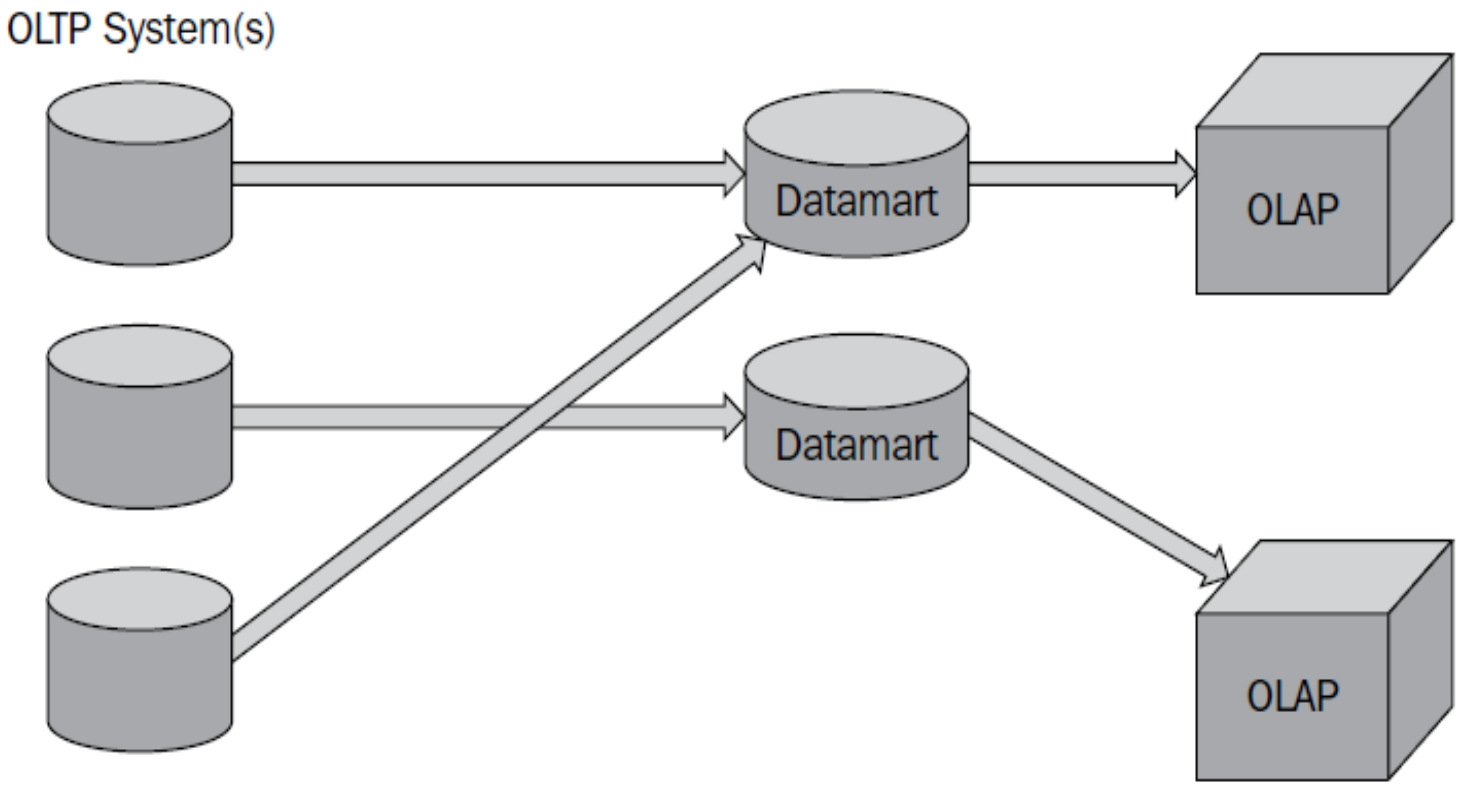


Kimball

- Propone concepto de Data Marts con dimensiones conformadas
- Divide la información en áreas de negocio
- Propone denormalización
- Propone modelo estrella

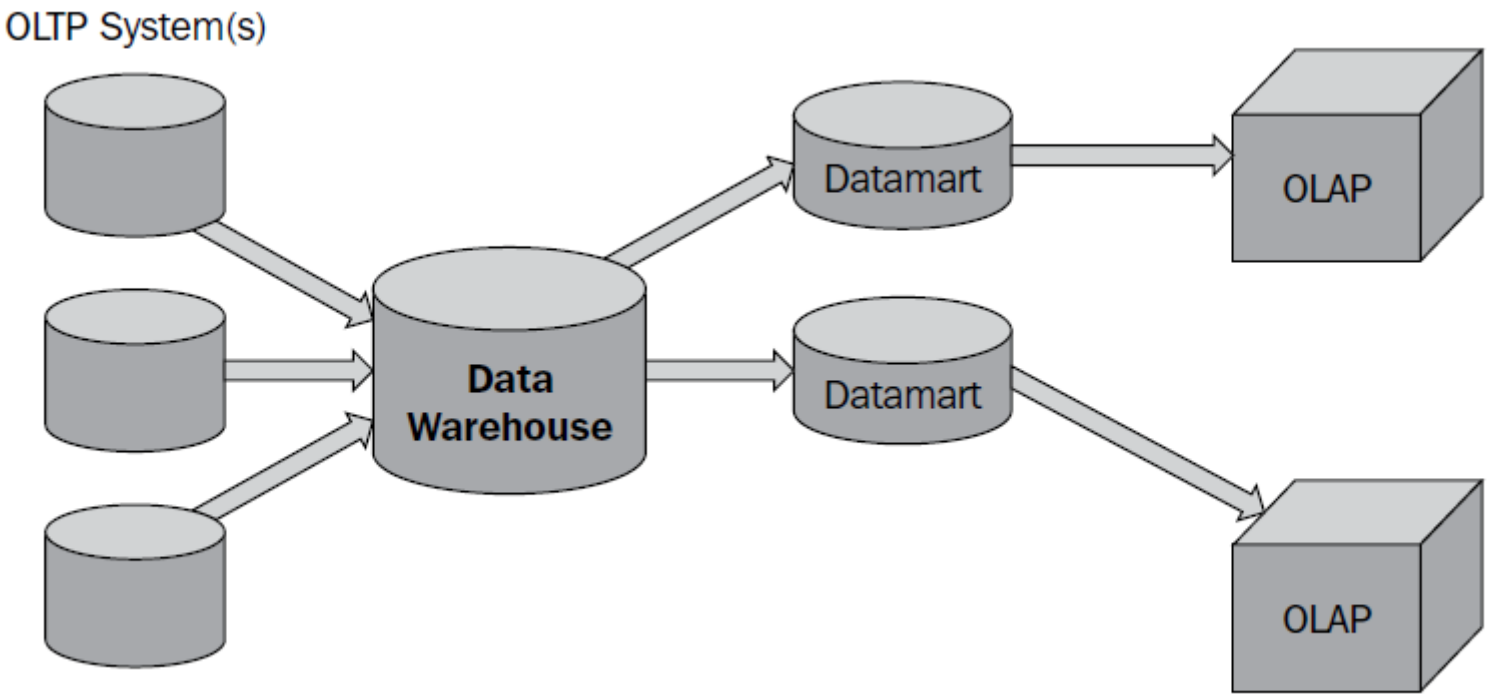


Modelo de Kimball



Arquitectura de BUS - Kimball

Modelo de Inmon



Principios de Diseño

- Claves subrogadas
- Nombres y tipos de convenciones
- Granularidad y agregación
- Fecha y Tiempo
- Llave desconocida en dimensiones.

Principios de Diseño – Claves Subrogadas

- Cada tabla usualmente tiene una PK, la cual es único identificador de un registro.
- La llave primaria puede consistir de uno o más atributos.
- Las llaves usadas como un sistema origen para un data warehouse a menudo contienen PK consistentes de múltiples columnas.
- El término utilizado para referirse a las llaves de un sistema origen es llaves naturales.

Principios de Diseño – Claves Subrogadas

- Las llaves naturales a menudo contienen información acerca de la naturaleza del registro. Ejemplo: RFC, CURP
- Una llave subrogada es un identificador sin un inherente significado. Es un campo numérico de una tabla cuyo único requisito es almacenar un valor numérico único para cada fila de la tabla.

Principios de Diseño – Nombres y Convenciones

- Todas las tablas deberán llevar un prefijo:
  - STG_ Tablas de Staging
  - HIS_ Tablas históricas
  - DIM_ Tablas de dimensión
  - FACT_ Tablas de Hechos
  - AGG_ Tablas de agregados
- Las columnas llave llevan el postfijo _KEY

Principios de Diseño – Nombres y Convenciones

- Todas la columnas llave en una dimensión son del tipo de dato más pequeño y sin signo.
- Usar nombres significativos para las columnas
- Usar nombres estándar para columnas de auditoria

Principios de Diseño – Granularidad y Agregación

Granularidad:

Nivel de detalle en el cual los datos son almacenados en el data warehouse.

Agregación:

Incrementa el performance en las consultas

Datos por mes, región y grupo de producto

Principios de Diseño – Columnas de Auditoría

- Habilitan la traza de los datos.
- Se recomiendan las siguientes columnas:
  - Insert timestamp
  - Update timestamp
  - Delete timestamp

Principios de Diseño – Modelando Fecha y Tiempo

- * Granularidad de la dimensión tiempo

  Ejemplo: Dimensión tiempo- fecha a nivel segundo.

  24 x 60 x 60 - > cada día

  x 365 -> por año

  31,536,000 registros en una tabla de dimensión

Principios de Diseño – Llave desconocida en dimensión

- Se recomienda generar una llave desconocida en las tablas de dimensión para hacer referencia a registros de las tablas de hechos que no tiene referencia.

Key	Source_id	name	address	phone
0	0	Unknown	Unknown	Unknown

Tabla de Hechos:

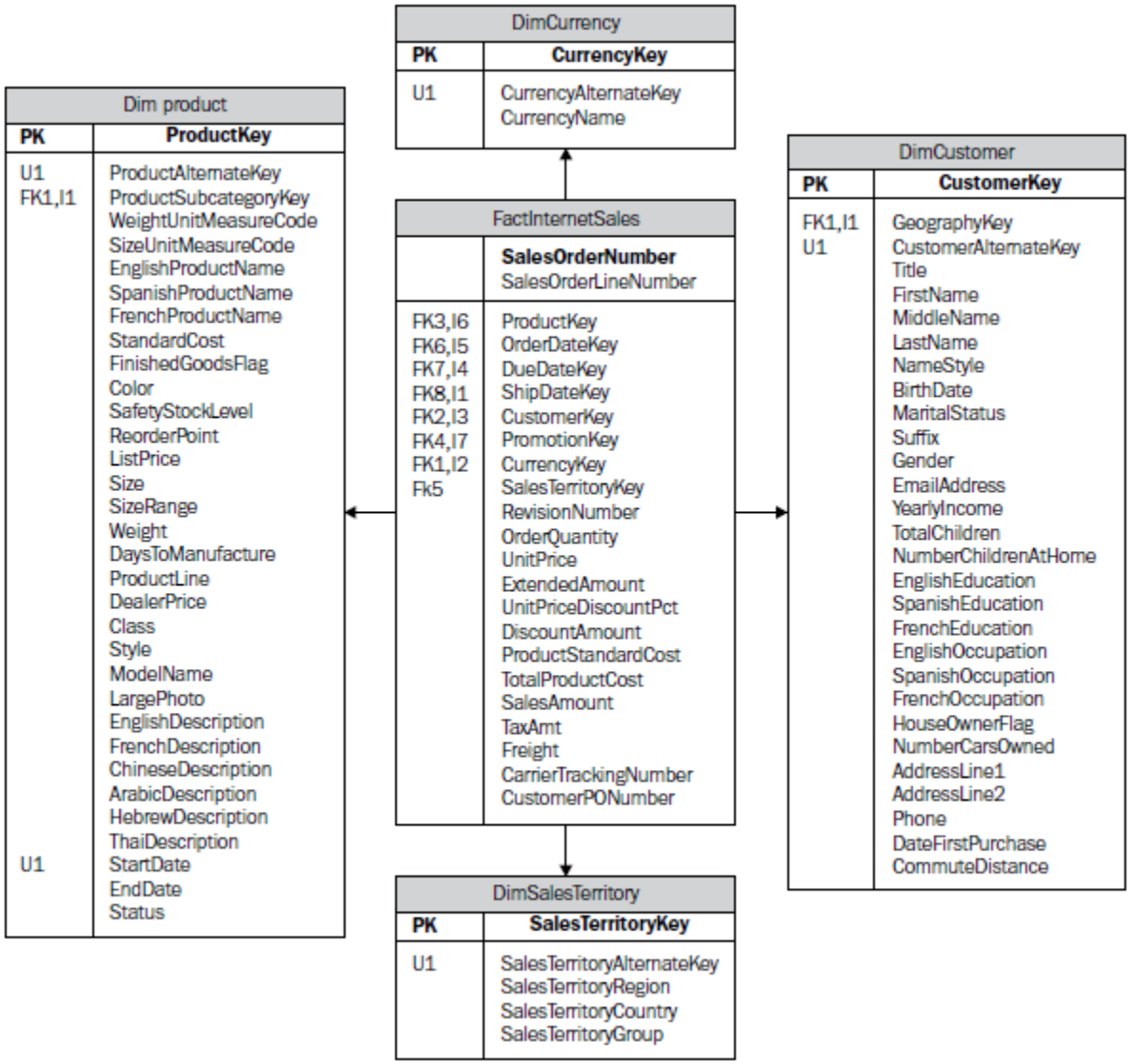
- Consiste de eventos del negocio que pueden ser registrados sobre el tiempo, tales como transacciones bancarias, ventas, órdenes, embarques, devoluciones y visitas al sitio web.
- Ellas son normalmente generadas por llaves foráneas a las tablas de dimensiones y un conjunto de valores numéricos.
- La información almacenada en las tablas de hechos es usualmente estática debido a que es histórica.
- El más común ejemplo de una tabla de hechos en el esquema estrella son las ventas.

Dimensiones

- Consisten principalmente de información textual ligada a los registros de hechos, tales como nombres de clientes, descripciones de productos, proveedores y vendedores.
- Estas tablas contendrán menos registros que las tablas de hechos y no son estáticas. Los registros en las tablas de dimensiones pueden ser actualizables.

P. ejemplo: Las direcciones de clientes podrían modificarse en el sistema origen.

Esquema Estrella



Esquema Copo de Nieve