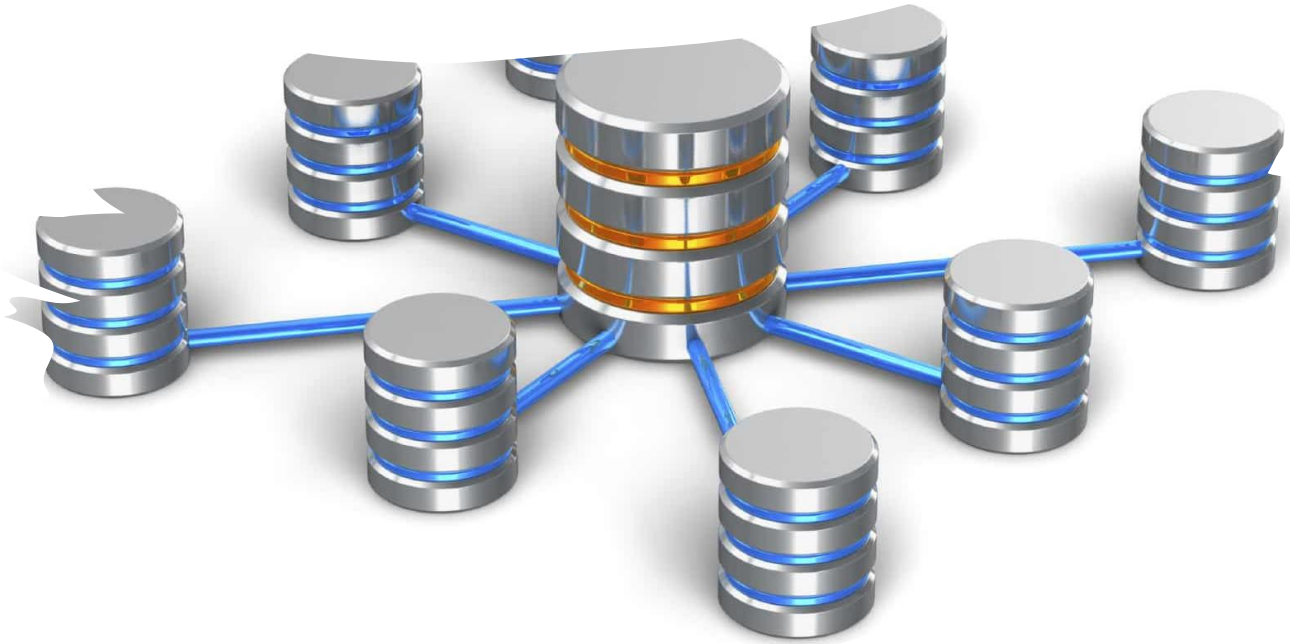


Almacenes de datos (Data Warehouse - DW)

Dr. Luis Gustavo Esquivel Quirós

Este material está basado en documentos desarrollados por Elzbieta Malinowski y Esteban Zimányi

Método de diseño

- El desarrollo de un almacén de datos es una tarea compleja y costosa
- Un proyecto de almacén de datos es similar en muchos aspectos a cualquier proyecto de desarrollo de software y requiere la definición de las diversas actividades que se deben realizar:
  - Recopilación de requisitos
  - Diseño
  - Implementación
  - Y otros

Método de diseño

- Muchos escritos sobre desarrollo de almacenes de datos (Kimball, Inmon, Imhoff y otros)
  - Se basan en su experiencia en la construcción de almacenes de datos
  - Carecen de formalidad y no consideran todas las fases de diseño
- Otros son propuestos por la comunidad científica (Böehnlein et al., Bonifati et al., Carneiro et al., Luján-Mora et al., Prakash et al., entre otros)
  - Están dirigidos a un modelo conceptual específico
  - Suelen ser demasiado complejos

Método de diseño

- Se requiere un marco metodológico que pueda guiar a los desarrolladores en las distintas etapas del proceso de desarrollo de DW
  - Motivo: la necesidad de construir sistemas DW surgió antes de la definición de enfoques formales para el desarrollo de DW, como fue el caso de los DB operativos
- La propuesta de un método general para el diseño de un almacén de datos convencional que unifica los enfoques existentes permite que:
  - Los diseñadores y desarrolladores puedan comprender mejor los enfoques alternativos que se pueden utilizar para el diseño de DW
  - Pueden elegir el que mejor se adapte a sus necesidades
- El método propuesto aquí no cubre el proceso general de desarrollo de DW, sino que se centra en el diseño de DW

Diseño *"top-down"* vs *"bottom-up"*

- De manera similar al diseño de bases de datos transaccionales, existen dos métodos principales para el diseño de un almacén de datos y sus data marts relacionados.
- Diseño *"top-down"*: los requisitos de los usuarios en diferentes niveles organizacionales se fusionan antes de que comience el proceso de diseño y se crea un esquema para todo el almacén de datos. Posteriormente, los data marts separados se adaptan de acuerdo con las características particulares de cada área de negocio, proceso o requerimiento
- Diseño *"bottom-up"*: se crea un esquema separado para cada data mart, teniendo en cuenta los requisitos de los usuarios responsables de la toma de decisiones del área de negocio o proceso específico correspondiente. Posteriormente, estos esquemas se fusionan, formando un esquema global para todo el almacén de datos

Diseño *"top-down"* vs *"bottom-up"*

- La planificación e implementación de un almacén de datos en toda la empresa utilizando *"top-down"* es una tarea abrumadora para la mayoría de las organizaciones en términos de costo y duración
  - También es una actividad desafiante para los diseñadores debido a su tamaño y complejidad
- Por otro lado, el tamaño reducido de los data marts permite que una empresa recupere el costo de construirlos en un período de tiempo más corto y facilita los procesos de diseño e implementación

Diseño *"top-down"* vs *"bottom-up"*

- Los profesionales suelen utilizar el enfoque *"bottom-up"* en el desarrollo de almacenes de datos.
  - Sin embargo, esto requiere que se establezca un marco de almacenamiento de datos global para que los *"data marts"* se construyan considerando su integración futura en un almacén de datos completo
- Se pueden aplicar varios marcos de desarrollo para lograrlo
  - Imhoff propone crear un esquema global del almacén de datos y luego, se construye un prototipo de cada *"data mart" y posteriormente* su estructura se mapea en el esquema global del almacén de datos; el proceso de mapeo se repite para cada *"data mart"* construido

Diseño *"top-down"* vs *"bottom-up"*

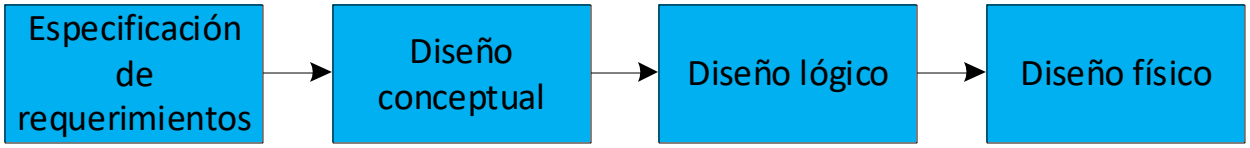
- Se pueden aplicar varios marcos de desarrollo para lograrlo
  - Kimball propone un marco de trabajo llamado arquitectura de bus de almacén de datos
    - Las dimensiones y los hechos compartidos entre diferentes *"data marts"* deben ser integrados
    - Una dimensión se integra cuando es idéntica en cada *"data mart"* que la utiliza
    - Un hecho se integra si tiene la misma semántica (por ejemplo, la misma terminología, granularidad y unidades para sus medidas) en todos los *"data marts"*
    - Los nuevos *"data marts"* pueden incorporarse al almacén de datos de manera incremental, asegurando su compatibilidad con los *"data marts"* ya existentes

Un método para el diseño de DW

- Un método general para el diseño de DW que abarca los diversos enfoques.
- El marco de trabajo:
  - Permite describir con precisión las ventajas y desventajas de las distintas opciones
  - Permite a los diseñadores elegir la opción que mejor se adapte a sus necesidades y a las particularidades del proyecto DW en cuestión

Un método para el diseño de DW

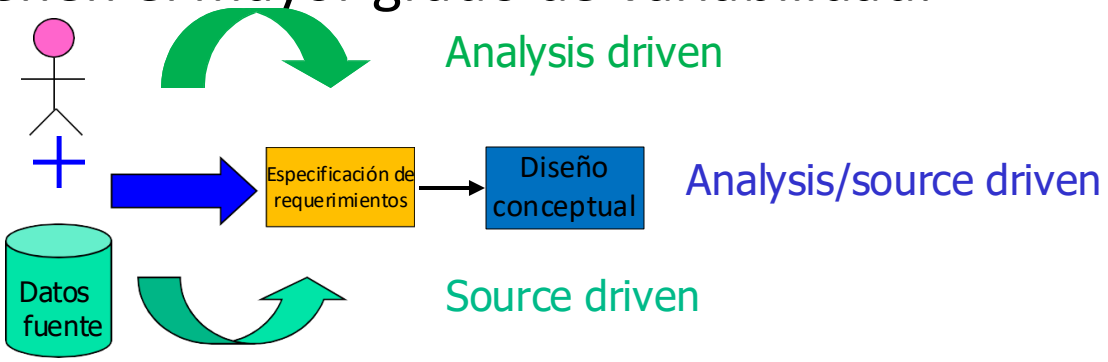
- Dado que los DW son un tipo particular de bases de datos dedicadas a fines analíticos, su diseño debe seguir las fases tradicionales de diseño de bases de datos



- Las fases no dependen de si se utiliza el enfoque de arriba hacia abajo o de abajo hacia arriba

Un método para el diseño de DW

- Las dos primeras fases, especificación de requerimientos y diseño conceptual, son las más críticas:
  - Pueden afectar considerablemente la aceptación del sistema por parte de los usuarios
  - Determinan si la relación entre el mundo real y el mundo del software, es decir, entre las necesidades de los usuarios y lo que ofrecerá el sistema modelado, será adecuada
  - Tienen el mayor grado de variabilidad.



Un método para el diseño de DW

- Las siguientes fases, diseño lógico y físico, son fases más técnicas que traducen sucesivamente el esquema conceptual obtenido en la fase anterior a las estructuras de implementación objetivo de una herramienta DW en particular

Un Estudio de Caso de la Universidad

- En las últimas décadas, las universidades de todo el mundo han enfrentado cambios significativos como resultado de varios factores.
  - Disminuciones en la financiación
  - Cambios políticos y organizativos
  - Globalización
  - Aumento de la competencia
- En particular, debido a la globalización de la educación superior
  - Los estudiantes buscan en el extranjero las mejores oportunidades educativas
  - Las empresas recorren el mundo para establecer contratos de investigación
  - Académicos altamente calificados buscan mejores condiciones para la investigación

Un Estudio de Caso de la Universidad

- Atraer a muchos estudiantes, contratos de investigación y académicos bien preparados ayuda a las universidades a mejorar su situación económica general
- El ranking de universidades en varias escalas (mundial, continental, nacional, regional, etc.) se ha convertido en un factor importante para establecer la reputación de una universidad a nivel internacional
- Los rankings publicados por la Universidad Jiao Tong de Shanghái desde 2003 y por The Times desde 2004 han atraído una gran atención en todo el mundo
- Estas clasificaciones pueden brindar información importante con respecto a las instituciones competidoras

Un Estudio de Caso de la Universidad

- Supongamos que una universidad quiere determinar qué acciones debe tomar para mejorar su posición en el ranking del Times
- Los criterios de evaluación de este ranking se refieren a las dos principales áreas de actividad de las universidades
  - Investigación - 60% de los criterios
  - Educación – 40% de los criterios
- Los usuarios decisores optaron inicialmente por analizar la situación relacionada con las actividades de investigación

Un Estudio de Caso de la Universidad

- Las universidades generalmente se dividen en facultades que representan campos generales de conocimiento, por ejemplo, medicina, ingeniería y ciencias
- Estas facultades comprenden varios departamentos diferentes dedicados a dominios más especializados, por ejemplo, la facultad de ingeniería puede incluir departamentos de ingeniería civil, ingeniería mecánica e ingeniería informática, etc
- El personal universitario (es decir, profesores, investigadores, asistentes de enseñanza, personal administrativo, etc.) está adscrito administrativamente a los departamentos

Un Estudio de Caso de la Universidad

- Esta estructura organizativa tradicional no se adapta bien a las actividades de investigación multidisciplinarias, que requieren experiencia en varios dominios, posiblemente en diferentes facultades.
- Estructuras autónomas llamadas centros de investigación apoyan este tipo de investigación multidisciplinar.
- A estos centros de investigación puede pertenecer personal universitario de diversas facultades o departamentos.
- Los proyectos de investigación son realizados por uno o varios organismos de investigación, que pueden ser departamentos o centros de investigación.

Un Estudio de Caso de la Universidad

- El departamento de investigación es el órgano administrativo que coordina todas las actividades de investigación en la universidad.
- Sirve como puente entre ejecutivos de alto nivel e investigadores, así como entre investigadores y organizaciones externas.
- En particular, el establecimiento de áreas estratégicas de investigación se basa en las fortalezas y ambiciones centrales de la universidad, teniendo en cuenta el potencial y la relevancia a largo plazo.
- Estas áreas son el foco de iniciativas e inversiones institucionales.
- Sobre la base de la estrategia de investigación institucional, las facultades, departamentos y centros de investigación establecen sus propias prioridades de investigación.

Un Estudio de Caso de la Universidad

- Determine
  - Qué datos deberían estar disponibles
  - Cómo se debe organizar
  - Qué consultas son de interés
  - Otro
- Es uno de los primeros pasos en el desarrollo del sistema.
- Puede acarrear problemas importantes si está defectuoso o incompleto.
- Debería llamar especialmente la atención.
- Debe estar ampliamente respaldado por métodos efectivos.

Un Estudio de Caso de la Universidad

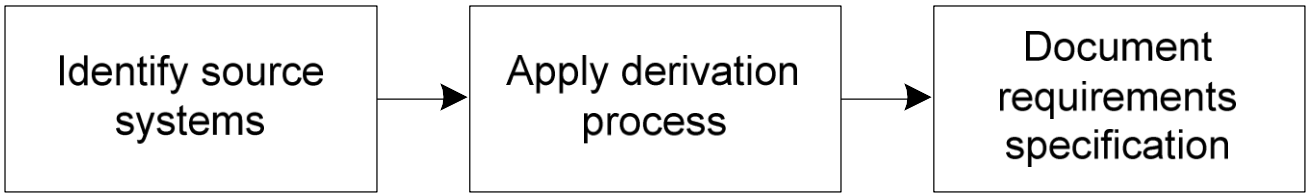
- Sin embargo
  - No se ha prestado mucha atención a la fase de análisis de requisitos en el desarrollo de DW
  - La variedad de enfoques existentes para la especificación de requisitos ha llevado a que muchos proyectos DW se salten esta fase
  - Se concentran en cuestiones técnicas, por ejemplo, modelado DW o rendimiento de consultas.
  - Se estima que más del 80% de los proyectos de DW no satisfacen las necesidades de los usuarios y no brindan el apoyo esperado para el proceso de toma de decisiones.

Especificación de requerimientos
Source driven

- Se basa en los datos disponibles en los sistemas de origen
- Su objetivo es identificar todos los esquemas multidimensionales candidatos que se pueden implementar de manera realista además de las bases de datos operativas disponibles
- Estas bases de datos se analizan exhaustivamente para descubrir los elementos que pueden representar hechos con medidas asociadas y dimensiones con jerarquías
- La identificación de estos elementos conduce a un esquema DW inicial que puede corresponder a varios propósitos de análisis diferentes

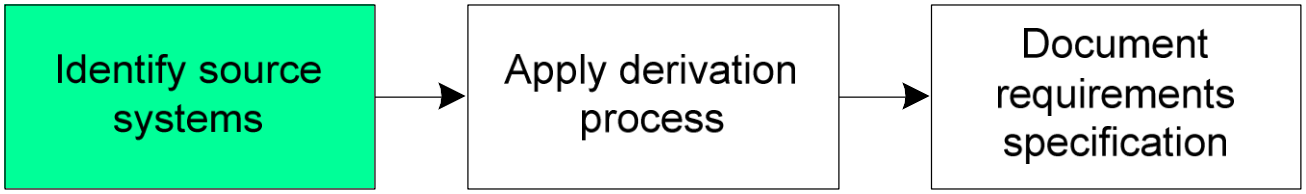
Especificación de requerimientos
Source driven

- Varios pasos a realizar:



Especificación de requerimientos
Identificar sistema fuente

- Varios pasos a realizar:

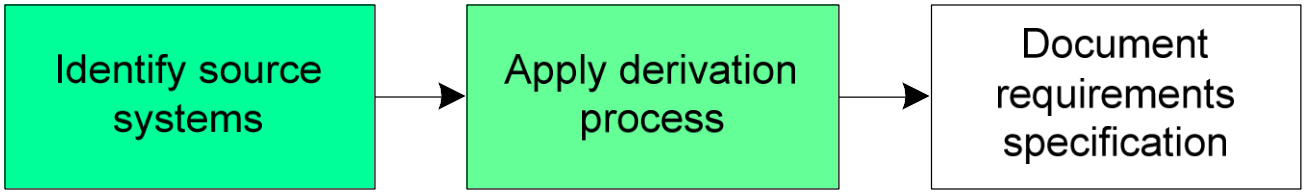


- Determinar los sistemas operativos existentes que pueden servir como proveedores de datos para DW
- Las fuentes externas no se consideran en esta etapa; se pueden incluir más adelante
- En presencia de varios sistemas operativos, se deben seleccionar aquellos que proporcionen mayor calidad de datos y estabilidad de sus esquemas

Especificación de requerimientos
Identificar sistema fuente

- Este paso se basa en la documentación del sistema representada por:
  - El modelo ER
  - Tablas relacionales
- Esta puede ser una tarea difícil si
  - Las fuentes de datos incluyen estructuras implícitas que no se declaran a través de DDL
  - Se agregaron estructuras redundantes y no normalizadas para mejorar el tiempo de respuesta de las consultas.
  - Las bases de datos han sido creadas por desarrolladores novatos o no capacitados
  - Los DB residen en sistemas heredados cuya inspección es una tarea difícil

Especificación de requerimientos
Aplicar proceso de derivación



- Se pueden aplicar varias técnicas para derivar elementos multidimensionales de bases de datos operativas
- En el primer paso, se determinan las relaciones de hecho y sus medidas asociadas
- Luego, se derivan la dimensión y las jerarquías

Especificación de requerimientos
Aplicar proceso de derivación

- Identificar relaciones de hechos y medidas es el aspecto más importante de este enfoque, ya que son la base para construir esquemas multidimensionales
- Las relaciones de hecho y las medidas son elementos que corresponden a eventos que ocurren dinámicamente en la organización, es decir, que se actualizan con frecuencia

Especificación de requerimientos
Aplicar proceso de derivación

- Si las bases de datos operativas son relacionales:
  - Los hechos pueden corresponder a tablas y/o atributos
  - Las medidas corresponden a los atributos
- Si las bases de datos operativas se representan utilizando el modelo ER
  - Los hechos pueden ser tipos de entidad o de relación
  - Las medidas pueden ser atributos de estos elementos
- Una opción alternativa puede ser involucrar a los usuarios, quienes entienden los sistemas 'operativos' y pueden ayudar a determinar qué datos se pueden considerar como medidas.

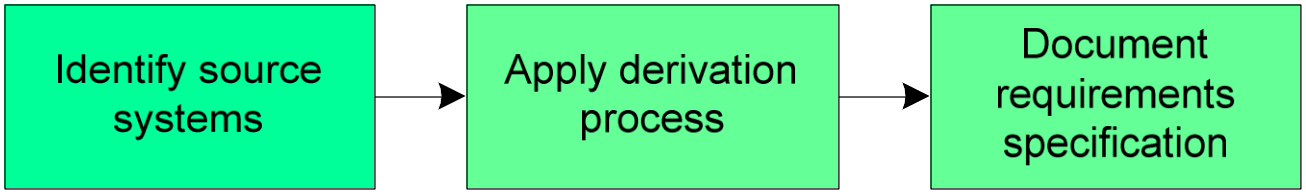
Especificación de requerimientos
Aplicar proceso de derivación

- Se pueden aplicar varios procedimientos para derivar dimensiones y jerarquías.
- Pueden ser
  - Automático
  - Semiautomático
  - Manual
- El proceso de descubrir una dimensión o un nivel de hoja de una jerarquía generalmente comienza con la identificación de los elementos estáticos (que no se actualizan con frecuencia) que están relacionados con los hechos.
- Luego, se realiza una búsqueda de otros niveles de jerarquía.
  - Comenzando con un nivel de hoja de una jerarquía, se revisa cada relación de uno a muchos

Especificación de requerimientos
Aplicar proceso de derivación

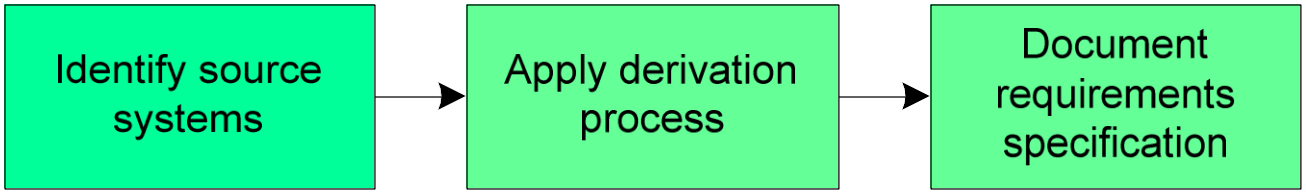
- Los procedimientos automáticos o semiautomáticos requieren conocimientos sobre los modelos conceptuales específicos que se utilizan para el esquema inicial y sus transformaciones posteriores.
- Los procedimientos manuales permiten a los diseñadores encontrar jerarquías incrustadas dentro de la misma entidad o tabla, por ejemplo, para encontrar atributos de ciudad y provincia en un tipo de entidad de tienda
  - Se requiere la presencia de expertos en sistemas que entiendan los datos en las bases de datos operativas
  - O el diseñador debe tener un buen conocimiento sobre el dominio comercial y los sistemas subyacentes.

Especificación de requerimientos
Documentación de requerimientos



- Al igual que en el enfoque basado en el análisis, la fase de especificación de requisitos debe documentarse
- La documentación debe describir aquellos elementos de los sistemas fuente que pueden considerarse como relaciones de hechos, medidas, dimensiones y jerarquías
- Esto estará contenido en los metadatos técnicos
- Es deseable involucrar en esta etapa a un experto en el dominio que pueda ayudar a definir la terminología comercial para estos elementos

Especificación de requerimientos
Documentación de requerimientos



- Al igual que en el enfoque basado en el análisis, la fase de especificación de requisitos debe documentarse
- La documentación debe describir aquellos elementos de los sistemas fuente que pueden considerarse como relaciones de hechos, medidas, dimensiones y jerarquías
- Esto estará contenido en los metadatos técnicos
- Es deseable involucrar en esta etapa a un experto en el dominio que pueda ayudar a definir la terminología comercial para estos elementos