PROCESOS ETL
(EXTRACCIÓN, TRANSFORMACIÓN Y CARGA)

Tema 5

Profesores:

Juan C. Trujillo, Alejandro Maté

LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

1

Indice

2

Indice

- Introducción
- Extracción
- Transformación
- Carga

3

Introducción

- Bill Inmon (90's): "Un almacén de datos es una colección de datos orientados por temas, integrados, no volátiles y variables en el tiempo en apoyo de la toma de decisiones estratégicas".
  - Datos procedentes de una gran variedad de fuentes
- ETL (Extraction-Transformation-Loading):
  - **Extracción** de datos desde fuentes de datos operacionales y heterogéneas,
  - **Transformación - limpieza** (conversión, limpieza, normalización, etc.), y
  - **Carga - refresco** en el Almacén de datos

4

Introducción

5

- Algunas tareas comunes de procesos ETL:
  - Datos de distintas fuentes se tienen que unir (join)
  - Datos se tienen que agregar
  - Datos se han de convertir a un formato común
  - Generar claves auto generadas
  - Verificar la calidad de los datos
  - Etc.

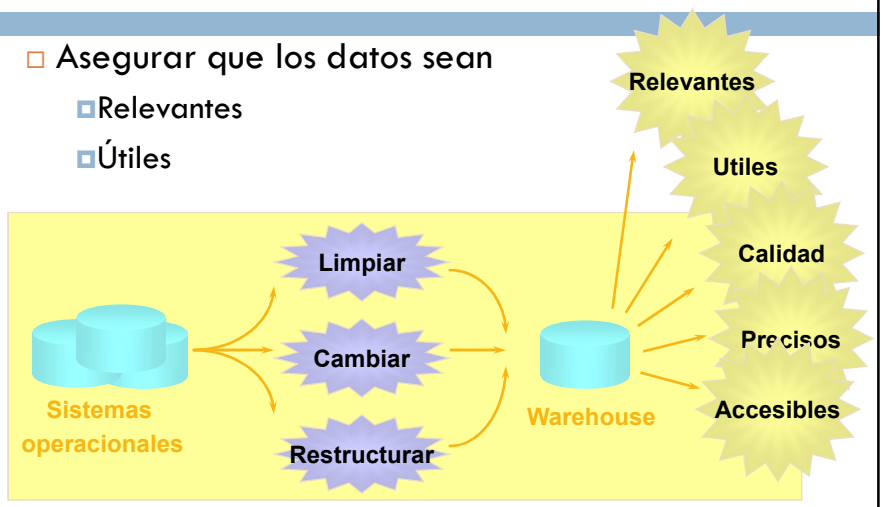
5

Introducción

Importancia de procesos ETL

6

- Asegurar que los datos sean
  - Relevantes
  - Útiles



- **ETL tienen un coste elevado (tiempo, recursos)**

6

Introducción

Algunas consideraciones generales

7

- Definir una estrategia de calidad de datos para la empresa según política de toma de decisiones
- Definir el nivel de calidad óptimo de los datos
- Considerar el modificar reglas de las fuentes de datos operacionales
- Básico → documentar las fuentes
- Diseñar los procesos de limpieza (y sus tareas) de forma muy cuidadosa
- Los procesos de limpieza iniciales puedes variar de los procesos de refresco posteriores

7

Introducción

Algunas consideraciones generales

8

- Cuidado → datos incorrectos o engañosos producirán decisiones estratégicas erróneas
- El mercado de herramientas de ETL en 2001: sobre $667 millones in USA
- Esfuerzo en ETL: aprox. 50% del presupuesto total de los proyectos de DW
- Actualmente el diseño y mantenimiento de procesos ETL es todavía un asunto “pendiente”
- Aunque varias herramientas en mercado, no disponemos de modelo o metodología estándar para su diseño desde primeros pasos de un proyecto de DW

8

Introducción

Soluciones

9

- Rutinas mediante lenguajes de programación
- Herramientas especializadas
- Proceso de conversión personalizada
- Expertos de negocio

9

Introducción

10

- Seis pasos detallados:

1. Seleccionar las fuentes para extraer datos
2. Transformar las fuentes
3. Unir las fuentes
4. Seleccionar las estructuras destino a cargar datos (hechos, dimensiones, etc.)
5. Mapear los atributos de las fuentes en los destinos
6. Cargar los datos

10

Introducción

- El paso de transformación también puede incluir limpieza de datos (detectar y borrar errores e inconsistencias)
- La creación manual y mantenimiento de los procesos ETL aumenta el coste de los DW
- CUIDADO: Documentación con gran cantidad de páginas con código de programas ETL

Introducción

- Tres pasos básicos:

1. Extraer de las fuentes de datos
2. Transformar las fuentes
3. Cargar los datos

Indice

- Introducción
- Extracción
- Transformación
- Carga

13

Indice

- Introducción
- Extracción
- Transformación
- Carga

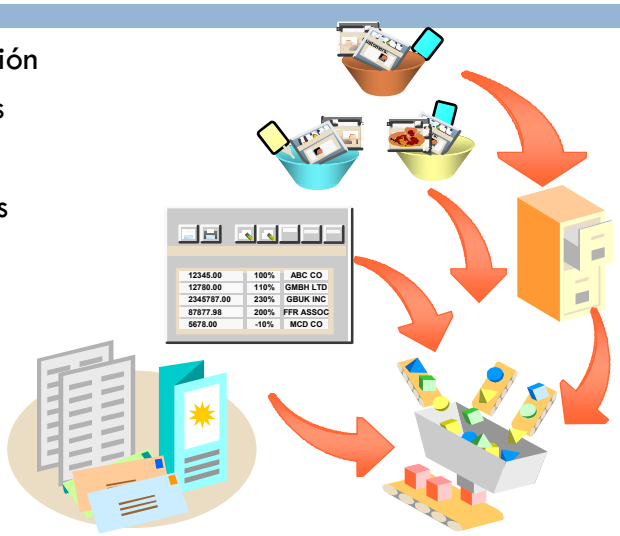
14

Extracción

Fuentes de datos

15

- Producción
- Archivos
- Internas
- Externas



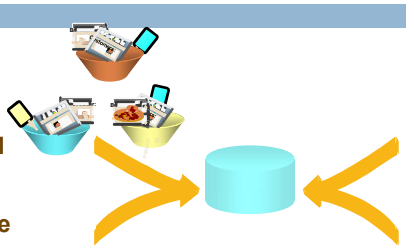
15

Extracción

Fuentes de datos. Producción

16

IMS
DB2
VSAM
SQL
Oracle
Sybase
Rdb



SAP
Sistemas médicos
Predicción financiera
Oracle Financial

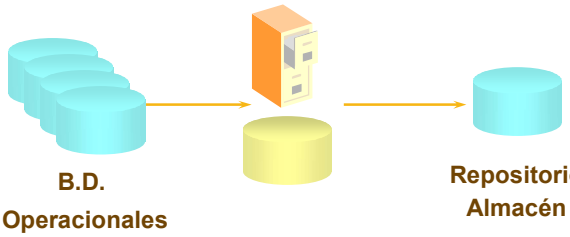
- Distintas plataformas de S.O.
- Plataformas Hardware
- Sistemas de ficheros
- Sistemas de bases de datos

16

Extracción

Fuentes de datos. Archivos

17



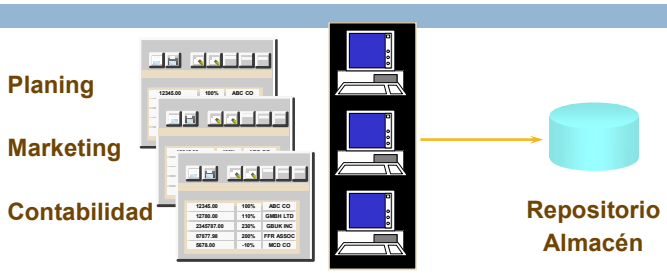
- Datos históricos ya almacenados
- Útiles para análisis de largos periodos de tiempo
- Útiles para primera carga
- Generalmente requerirán transformaciones

17

Extracción

Fuentes de datos. Datos internos

18



- Planning, ventas, y marketing
- Podemos encontrar
  - Spreadsheets - estructurados
  - Documentos – no estructurados
- Todos son fuentes de datos

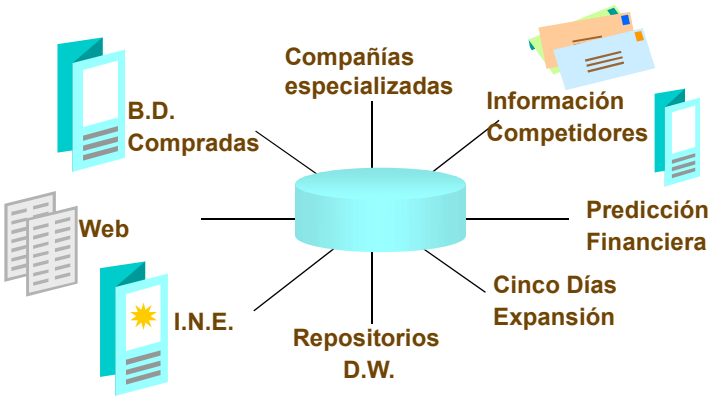
18

Extracción

Fuentes de datos. Datos externos

19

- Información desde fuera de la organización

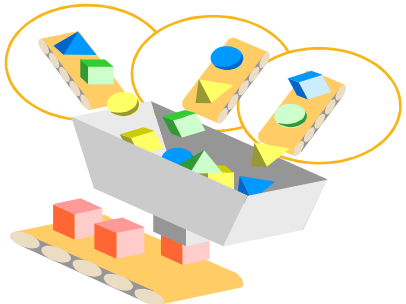


Extracción

Técnicas de extracción

20

- Programas – C#, Java, COBOL, PL/SQL
- Gateways – acceso a b.d. transparentes
- Herramientas
  - Coste inicial muy alto
  - Automatización
  - Limpieza de datos



Indice

21

- Introducción
- Extracción
- Transformación
- Carga

21

Indice

22

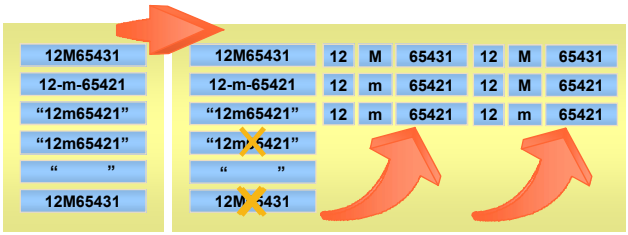
- Introducción
- Extracción
- Transformación
- Carga

22

Transformación

23

- Anomalías existen en fuentes operacionales
  - Limpiar



23

Transformación

Anomalías de fuentes de datos

24

- Normalmente no existe clave única
- Anomalías de instancias y codificado
- Inconsistencias de ortografía

CLINUM	NOMBRE	DIRECCION
90328575	Oracle Corp	100 NE 1st Street, Tampa
90328575	Oracle	100 NE. First St., Tampa
90238475	Oracle Services	100 North East 1st St., FLA
90233479	Oracle Limited	100 N.E. 1st St.
90233489	Oracle Computing	15 Main Road, Ft. Lauderdale
90234889	Oracle Corp. UK	15 Main Road, Ft. Lordadale, FLA
90345672	Oracle Corp UK Ltd	181 North Street, Key West, FLA

24

Transformación

Anomalías de fuentes de datos

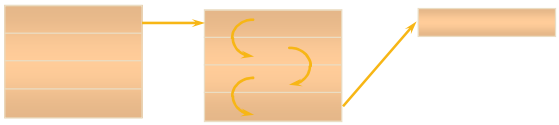
- Wrapper: transformar fuentes de datos nativos en fuentes de datos basadas en registros

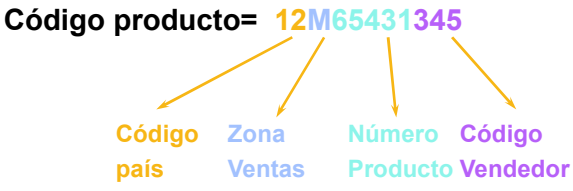
25

Transformación

Algunas transformaciones comunes

- Claves compuestas





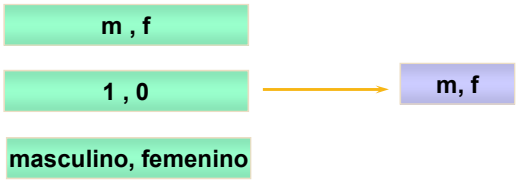
26

Transformación

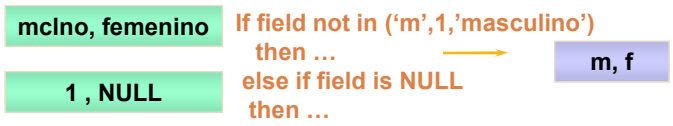
Algunas transformaciones comunes

27

- Codificación múltiple



- Detectar datos erróneos

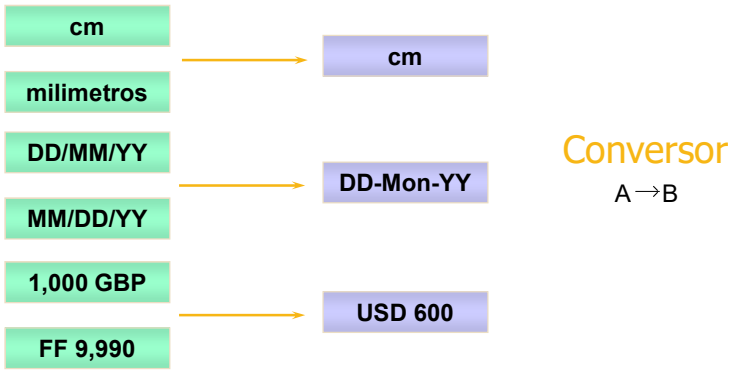


Transformación

Algunas transformaciones comunes

28

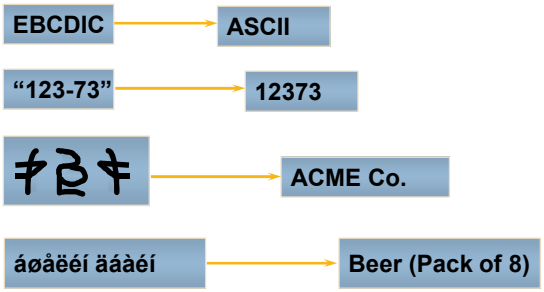
- Varios formatos válidos y estándares
- Herramientas o filtros para pre-procesar



Transformación

Algunas transformaciones comunes

29



29

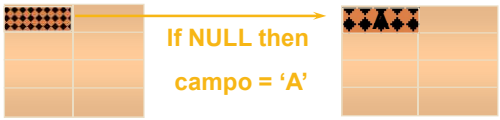
Transformación

Algunas transformaciones comunes

30

- NULL y valores que faltan
  - Ignorar
  - Esperar
  - Marcar las filas
  - Extraer bajo condiciones establecidas

Filter



30

Transformación

Algunas transformaciones comunes

31

- Valores duplicados
  - SQL
  - Server

Join

```
SELECT …
FROM   table_a, table_b
WHERE table_a.key (+) = table_b.key
UNION
SELECT …
FROM   table_a, table_b
WHERE table_a.key = table_b.key (+)
```

ACME Inc
ACME Inc
ACME Inc

→ ACME Inc

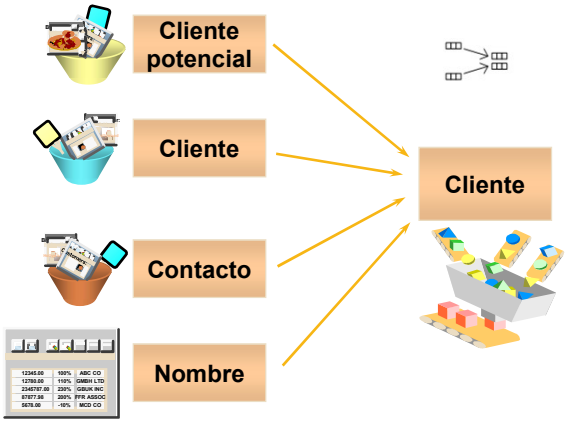
31

Transformación

Algunas transformaciones comunes

32

- Atributos compatibles



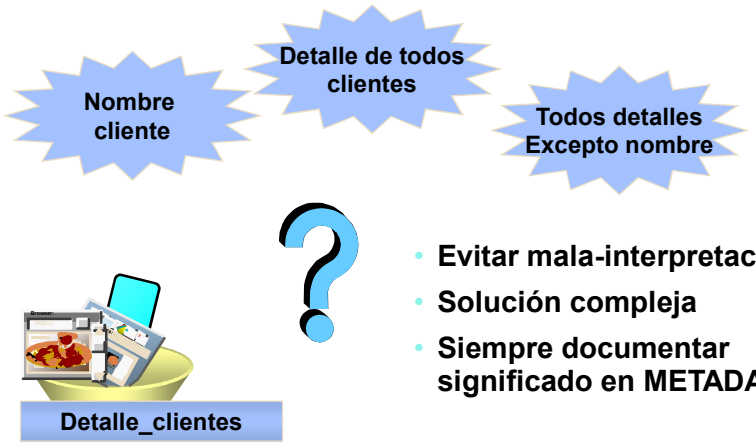
32

Transformación

Algunas transformaciones comunes

33

- Significado correcto de cada elemento



- **Evitar mala-interpretación**
- **Solución compleja**
- **Siempre documentar significado en METADATA**

33

Transformación

Algunas transformaciones comunes. Ejemplo.

34

- No hay clave única
- Valores que faltan
- Nombres personales y comerciales mezclados
- Diferentes direcciones para el mismo miembro
- Diferentes nombres y ortografía para el mismo miembro
- Muchos nombres en la misma linea
- Un nombre en dos líneas

	Nombre	Localización
Database 1	DIANNE ZIEFELD	N100
	HARRY H. ENFIELD	D589
	FRED AND SARA MULLEN	M300
Database 2	ZIEFLED, DIANNE	100
	ENFIELD, HARRY H	589
	MULLEN, SARA AND FRED	300

34

Transformación

Algunas transformaciones comunes. Ejemplo de FUSION

35

- Transacciones operacionales no son un mapeo 1-to-1 con los datos del DW.
- Datos del DW son "fusionados/unidos" para proporcionar información para el análisis.

Merge

Pizza ventas/devoluciones (dia,hora,seg.)

Sale	1/2/98	12:00:01 Ham Pizza	$10.00
Sale	1/2/98	12:00:02 Cheese Pizza	$15.00
Sale	1/2/98	12:00:02 Anchovy Pizza	$12.00
Return	1/2/98	12:00:03 Anchovy Pizza	- $12.00
Sale	1/2/98	12:00:04 Sausage Pizza	$11.00

Transformación

Algunas transformaciones comunes. Ejemplo de FUSION

36

Sale	1/2/98	12:00:01 Ham Pizza	$10.00
Sale	1/2/98	12:00:02 Cheese Pizza	$15.00
Sale	1/2/98	12:00:02 Anchovy Pizza	$12.00
Return	1/2/98	12:00:03 Anchovy Pizza	- $12.00
Sale	1/2/98	12:00:04 Sausage Pizza	$11.00

Sale	1/2/98	12:00:01 Ham Pizza	$10.00
Sale	1/2/98	12:00:02 Cheese Pizza	$15.00
Sale	1/2/98	12:00:04 Sausage Pizza	$11.00

Transformación

Algunas transformaciones comunes. Añadir tiempo.

37

- Permitir análisis del tiempo
- Añadir datos de tiempo en los datos de hechos y dimensiones
  - Añadir *triggers*
  - Aplicaciones de "código"
  - Comparar tablas

37

Transformación

Algunas transformaciones comunes. Claves generadas.

38

Surrogate
123 →

#1	Sale	1/2/98	12:00:01 Ham Pizza	$10.00
#2	Sale	1/2/98	12:00:02 Cheese Pizza	$15.00
#3	Sale	1/2/98	12:00:02 Anchovy Pizza	$12.00
#4	Return	1/2/98	12:00:03 Anchovy Pizza	- $12.00
#5	Sale	1/2/98	12:00:04 Sausage Pizza	$11.00

Valores de datos o claves artificiales

#dw1	Sale	1/2/98	12:00:01 Ham Pizza	$10.00
#dw2	Sale	1/2/98	12:00:02 Cheese Pizza	$15.00
#dw3	Sale	1/2/98	12:00:04 Sausage Pizza	$11.00

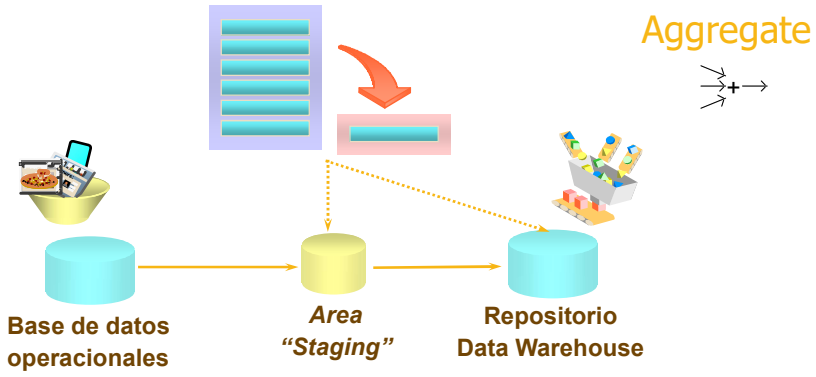
38

Transformación

Algunas transformaciones comunes. Datos agregados/sumados.

39

- Durante extracción o tratamiento (staging)
- Después de cargar los datos en el DW



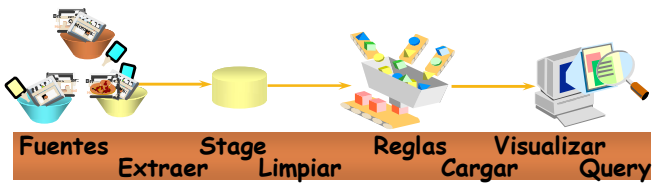
39

Transformación

Algunas transformaciones comunes. Respecto a metadatos.

40

- Extracción y Transformación
  - Reglas
  - Programas y algoritmos
  - Rutinas



40

Transformación

Algunas transformaciones comunes. Herramientas

- Algunos datos para elegir la herramienta
  - Rendimiento
  - Consumo de ancho de banda
  - Espacio de disco
  - Tiempo de la *ventana (window)* de carga
  - Nivel de automatización
  - Monitorización (*log*)
  - Funcionalidad
  - Acceso a metadatos
  - Requerimientos de entrenamiento

Indice

Indice

- Introducción
- Extracción
- Transformación
- Carga

Carga

Objetivos

- Identificar el transporte de datos para la primera vez y refrescos siguientes.
- Describir consideraciones estratégicas e implementar el refresco de datos
- Identificar métodos empleados para capturar cambios en los datos y, aplicarlos en el DW
- Describir técnicas de transporte.
- Identificar las tareas que se llevan a cabo después de que los datos se cargan.

Carga
Planteamiento general

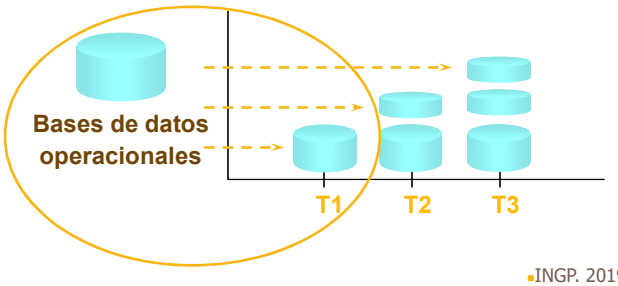
45

- Carga (*loading*) lleva los datos al DW.
- Carga puede necesitar mucho tiempo:
  - Considerar la ventana de carga
  - Planificar → intentar automatizar todos los procesos
- Carga inicial mueve grandes volumen.
- Cargas posteriores mueven volumen de datos más pequeños.
- El "negocio" determinar el ciclo de las cargas.

Carga
Primera carga

46

- Primera carga del DW con datos históricos
- Requiere grandes volúmenes de datos
- Puede emplear distintas tareas ETL
- Requiere grandes cantidades de procesamiento después de la primera carga.

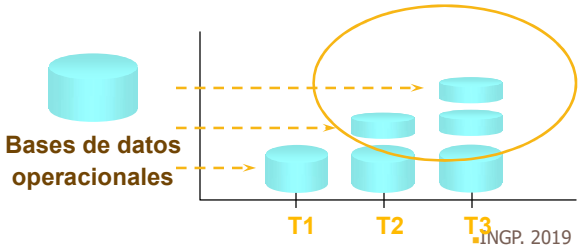


Carga

Refresco

47

- Realizados de acuerdo al ciclo del negocio.
- Es una tarea más simple
- Menos datos para la carga
- ETL menos complejos
- Menos rutinas de procesamiento después de la carga



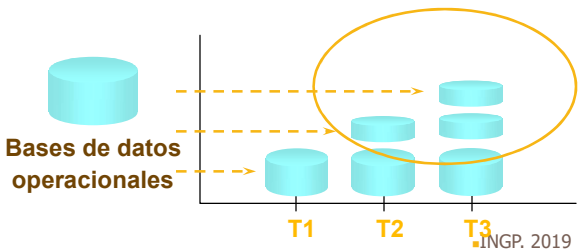
47

Carga

Estrategia de Refresco

48

- Considerar la ventana de carga
- Identificar los volúmenes de datos
- Identificar ciclos
- Conocer la infraestructura técnica
- Planificar un área de "trastienda" (staging)
- Determinar cómo detectar cambios



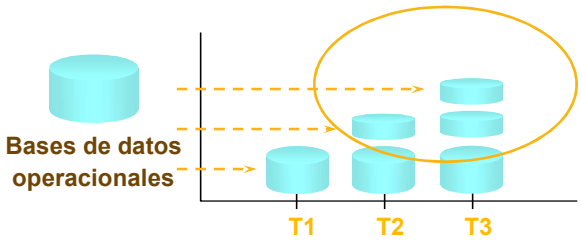
48

Carga

Utilizar requerimientos y usuarios

49

- Usuarios definen también el ciclo de refresco
- Documentar todas las tareas y procesos
- Consultar usuarios expertos



Carga

Construir el proceso de transporte

50

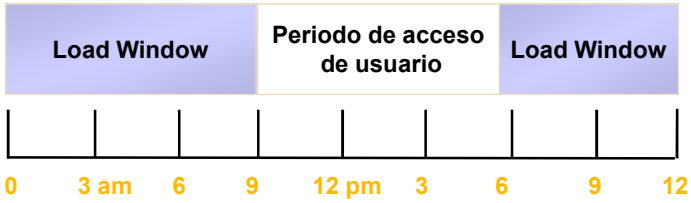
- Especificar
  - Técnicas y herramientas
  - Métodos de transferencia de ficheros
  - La ventana de carga
  - Ventana de tiempo para otras tareas
  - Volúmenes de primera carga y refresco
  - Frecuencia del ciclo de refresco
  - Ancho de banda de conectividad

Carga

Ventana de carga

51

- Tiempo disponible para todo el proceso ETL
- Planificar
- Comprobar
- Probar
- Monitorizar



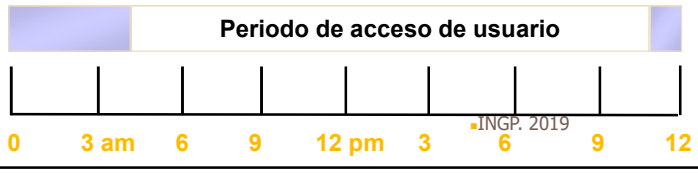
51

Carga

Ventana de carga

52

- Planificar y construir procesos de acuerdo a una estrategia.
- Considerar volúmenes de datos
- Identificar infraestructura técnica
- Asegurar la actualidad de los datos
- Considerar en primer lugar los requerimientos de acceso de usuarios
- Muchos requerimientos puede significar una ventana de carga pequeña



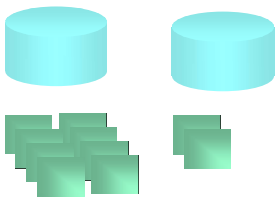
52

Carga

En cuanto a GRANULARIDAD

53

- Importante diseñarla
- Requerimientos de espacio
  - Almacenamiento
  - Copias
  - Recuperación
  - Particionamiento
  - Carga



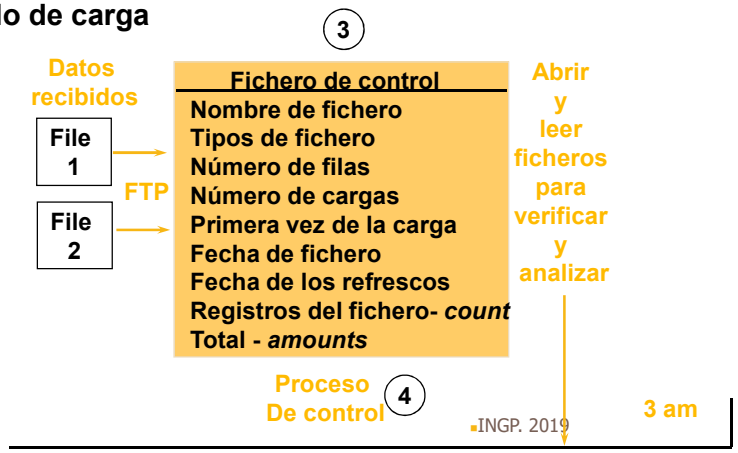
- Nivel de granularidad bajo
  - Caro, alto nivel de procesamiento, más disco, detalle,
- Nivel de granularidad alto
  - Más barato, menos procesamiento, menos disco, poco detalle

Carga

Planificación ventana de carga

54

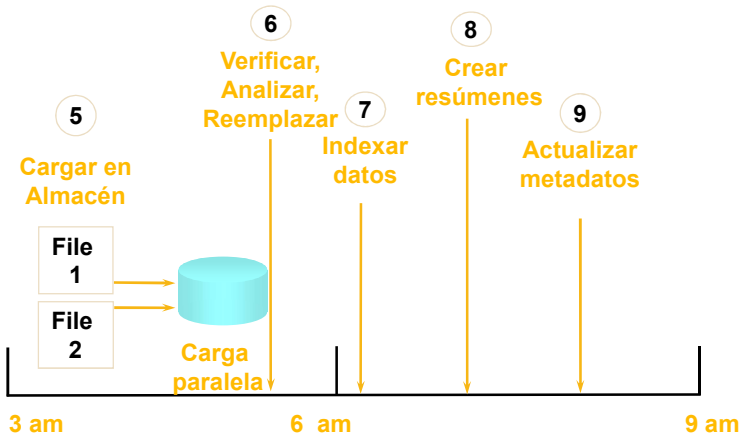
1. **Requerimientos**
2. **Ciclo de carga**



Carga

Planificación ventana de carga

55

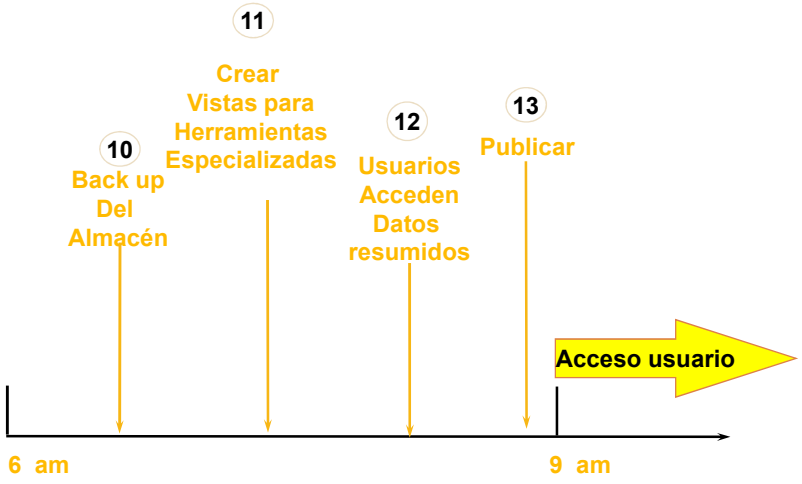


55

Carga

Planificación ventana de carga

56



56

Carga

Capturando los cambios de datos para refrescar

57

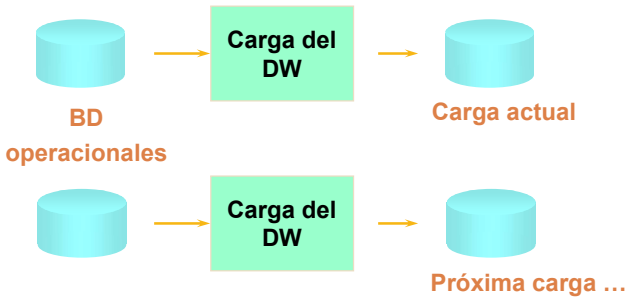
- Capturar nuevos datos de hechos
- Capturar datos de dimensión cambiados
- Determinar método para capturar ambos
- Métodos
  - Remplazar datos a gran escala
  - Comparar instancias de bases de datos
  - Comprobar/escanear tiempo *(Time stamping)*
  - Triggers en bases de datos
  - *Log* de bases de datos
- Técnicas híbridas

57

Carga

Cambios para refrescar. *Remplazar a gran escala.*

58



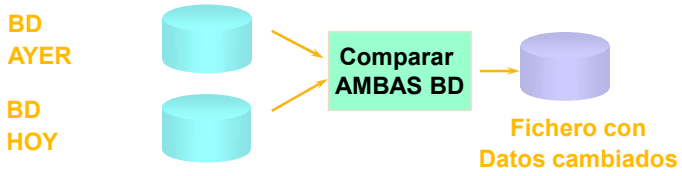
- Cara
- Datos históricos limitados
- Implementaciones de Data mart
- Reemplazar periodo de tiempo

58

Carga

Cambios para refrescar. *Comparar instancias de BD.*

59

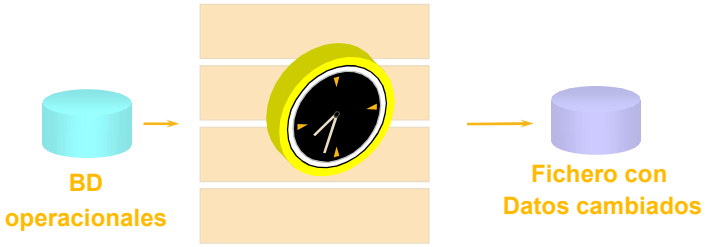


- Simple pero todavía cara
- Fichero con cambios
  - Cambios de datos operacionales desde último refresco
  - Utilizada por varias técnicas

Carga

Cambios para refrescar. *Time y Date stamping*

60

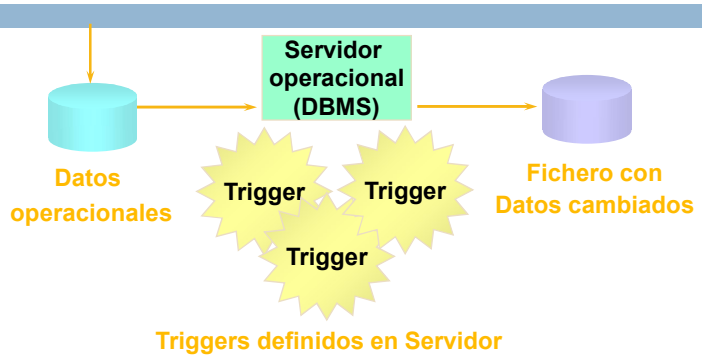


- Rápida comprobación para los registros cambiados desde última extracción
- Fichero actualizado respecto de fechas
- No detecta los datos borrados

Carga

Cambios para refrescar. *Triggers en BD*

61



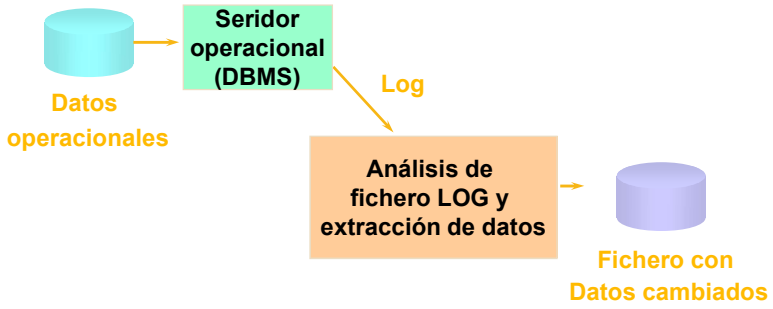
- Datos cambiados interceptados a nivel de servidor
- Uso extra de dispositivos entrada/salida
- Necesidades extras de mantenimiento

61

Carga

Cambios para refrescar. *Log de BD*

62



- Registramos imágenes de antes y después
- Necesita *checkpoint* del sistema
- Una técnica muy común

62

Carga

Cambios para refrescar. *Entonces QUE ???*

63

- Analizar cada método de forma individual
- Considerar una solución híbrida si un solo método no es adecuado
- Considerar elementos como aplicaciones actuales, BD operacionales disponibles y tecnología actual disponible.

63

Carga

Cambios para refrescar. COMO *APLICAR CAMBIOS ?*

64

- Para adoptar soluciones de políticas según cambios en operacionales ver soluciones de *Kimball* para dimensiones y hechos que cambian lenta y rápidamente (Tema 4)

64

Carga

Técnicas de transporte

- Herramientas
- Utilidades y lenguajes de programación
- *Gateways*
- Programas de copias personalizados
- Réplicas
- FTP
- Totalmente manual

Carga

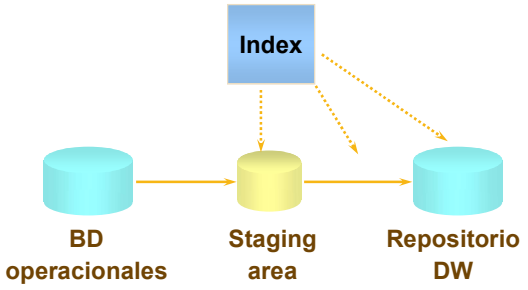
Cambios para refrescar. COMO *APLICAR CAMBIOS ?*

- Herramientas son adecuadas pero CARAS
- Utilidades son rápidas y potentes
- *Gateways* no siempre son los + rápidos
  - Acceso a otras BD
  - Soportar entorno distribuido
  - Proporcionar acceso en tiempo real si necesario
- Siempre se suele necesitar tratamiento posterior de la carga (*Post-processing*)

Carga
Indices

67

- Antes de carga – Indices rápidos
- Durante y después de la carga → añadir tiempo a la ventana de carga

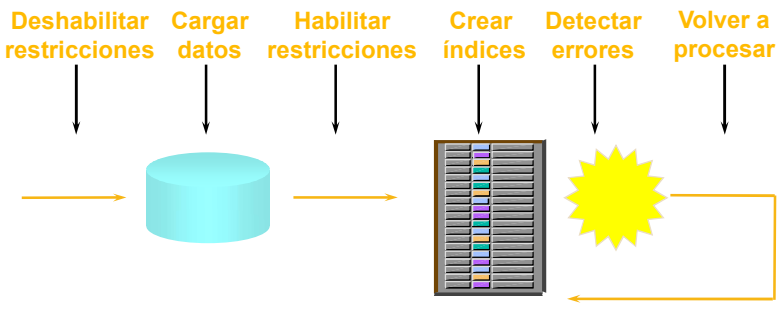


67

Carga
Indices únicos

68

- Deshabilitar restricciones antes de carga
- Habilitar restricciones para crear índices



68

Carga
Crear claves artificiales

69

- Usar claves derivadas o *generadas*
  - Mantener la unicidad de una fila
  - Necesario política y proceso administrativo para asignar claves
- Concatenar claves operacionales con número
  - Fácil de mantener
  - Claves un poco "grandes"

- Yo prefiero *autogeneradas* ...

69

Carga
Crear claves únicas/autogeneradas

70

- Asignar un número de una lista
  - Sin significado semántico
  - Operaciones de extracción deben referenciar a las tablas operacionales para asignar números

- Actualizar metadata
- Comprobar finalmente

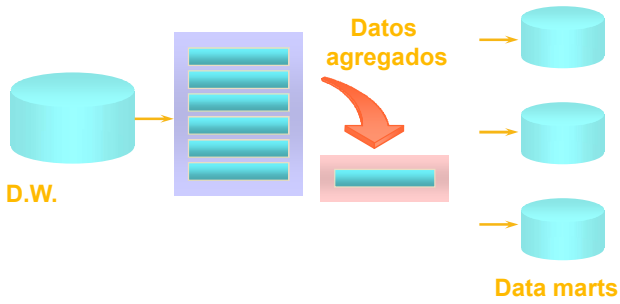
70

Carga

Crear tablas agregadas y cargar DM's

71

- Crear tablas agregadas
- Cargar Data Marts desde DW



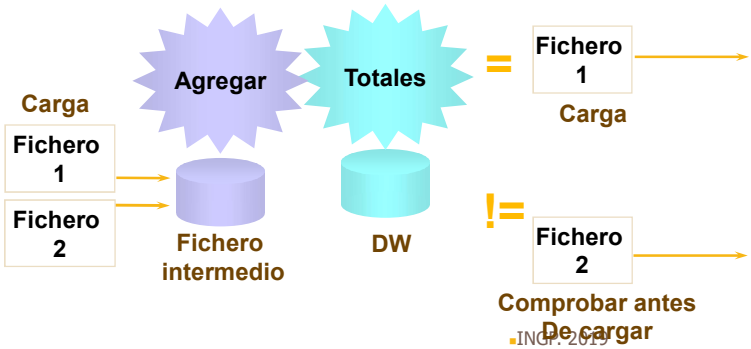
71

Carga

Verificar integridad de datos una vez cargados

72

- Cargar datos en un fichero/tabla intermedia
- Comprobar totales en DW con totales antes de la carga



72

Carga

Comprobar datos cargados

73

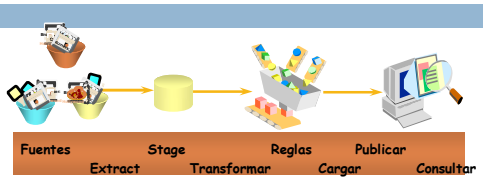
- Estatus de la carga (*log*)
- Proceso finalizado
- Todos datos cargados
- Comprobar violaciones
- Lanzar procesos de nuevo ???
- Comprobar los datos agregados

Carga

Tareas finales después de carga

74

- Actualizar metadata
  - ETL
  - Usuarios
- Publicar nuevos datos
  - Disponibilidad
  - Cambios
  - Vistas de negocio
- Aspectos de seguridad para accesos no deseados



Carga

Disponibilidad de datos

75

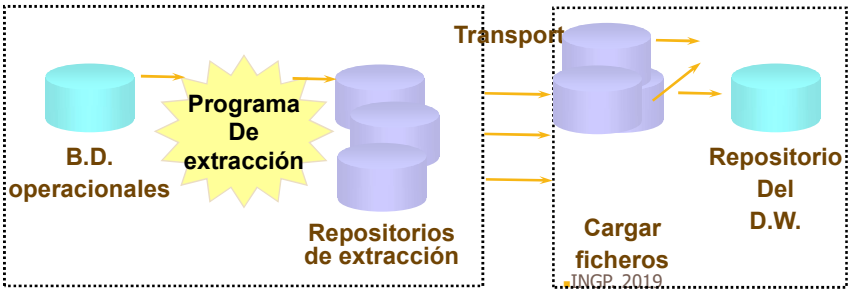
- A veces se requiere 24 horas para realizar todo el proceso de carga
- Compromiso entre carga y acceso de usuarios finales
- Considerar
  - Copias de actualizaciones
  - Tablas temporales
  - Utilizar tablas separadas

Carga

Automatizar proceso

76

- Extracción-transformación y carga
- Permitir procesamiento posterior
- Actualizar metadata
- Publicar cambios después de proceso
- Intervención humana para cuestiones imprevistas



Carga

Diseño de procesos de extracción

77

- Análisis
  - Fuentes, tecnologías
  - Tipos de datos fuentes, calidad, propietarios
- Opciones de diseño
  - Manual, personalizados, *gateway*, terceros
  - Replicar total o parcialmente
- Elementos de diseño
  - Volúmenes de datos, actualizados, copias
  - Automatizar, tecnología disponible

Carga

Diseño de procesos de transformación

78

- Análisis
  - Mapeos de fuentes y destino (DW), reglas de negocio
  - Granularidad, claves, metadatos,...
- Opciones de diseño
  - PL/SQL, replicar, clientes, terceros
- Elementos de diseño
  - Rendimiento
  - Tamaño del pre-procesamiento
  - Manejo de excepción, mantenimiento de integridad

Carga

Diseño de procesos de carga

- Análisis
  - Volúmenes de datos, actualizados
  - Distribución en D.M.
- Opciones de diseño
  - Replicas, personalizados, PL/SQL
  - Herramientas externas
- Elementos de diseño
  - Periodos permitidos (ventanas)
  - Particionamiento, distribución

79

PROCESOS ETL

(EXTRACCIÓN, TRANSFORMACIÓN Y CARGA)

Tema 5

Profesores:

Juan C. Trujillo, Alejandro Maté

LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

80