UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CIENCIAS

ALMACENES Y MINERÍA DE DATOS

Extracción, Transformación y Carga

Gerardo Avilés Rosas
gar@ciencias.unam.mx

Datos y sus problemas

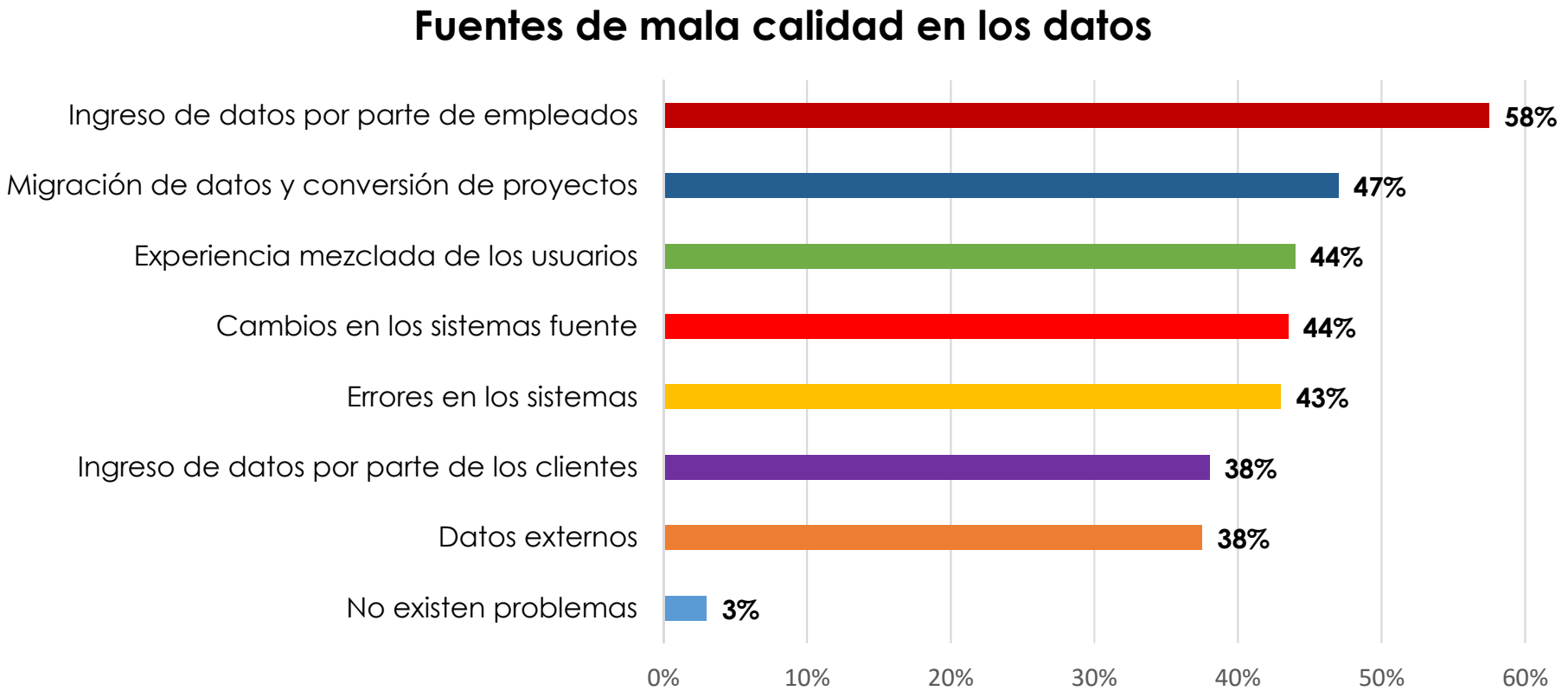
- Las organizaciones siempre se han visto atraídas por la evolución de la tecnología.
- En el inicio, el uso que se le daba principalmente a las computadoras era el de procesamiento de transacciones:

Disminuir el tiempo de los procesos internos del negocio

- Sin embargo, se ignoraba el valor que tiene los datos y se desaprovechaba la ganancia que brinda su análisis.
- Hoy en día, se busca a partir de la información crear **ventajas competitivas**.
- Para darle un uso productivo a la información se deben identificar:
  1. ***Utilidad que se desea dar a la información***
  2. ***Fuente de datos de la cual se extraerá***
  3. ***Ubicación que tendrá la misma dentro de la organización.***

…Datos y sus problemas

- Las bases de datos (hoy en día) son altamente susceptibles de tener **datos inconsistentes**, con **ruido** o bien tener datos faltantes (**missing values**).
- Las razones son variadas, pero principalmente se deben a:



Lehmann, C., Roy K., Wintere B., The State of Enterprise Data Quality: 2016. 451 Research, 2016

…Datos y sus problemas

- Se tienen grandes volúmenes de datos almacenados en las organizaciones, los cuales tienen un papel central en la toma de decisiones, requerimos entonces que estén libes de errores:

  Si vamos a confiar en la información de nuestra organización para tomar decisiones de negocio, debemos estar seguros que los datos sobre los cuales estamos tomado estas decisiones críticas son: ***exactos, completos y relevantes.***

- Hoy en día las organizaciones están tomando sus decisiones con base en el conocimiento derivado de los datos almacenados en sus **BD** o **DWH**: ***Inteligencia de Negocios (BI)***.

  "No es poco frecuente que las operaciones de bases de datos tengan del 60% al 90% de problemas de calidad de datos"

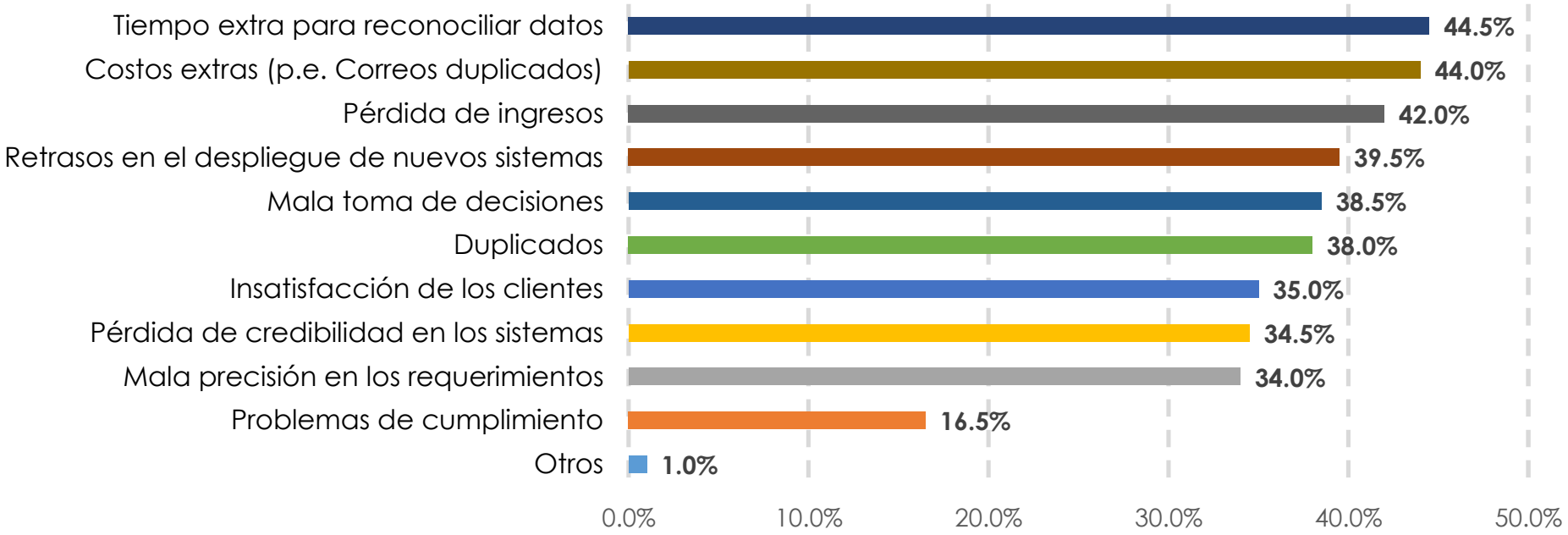
  Dasu, T., Vesonder, G. T., y Wright, J. R., Data quality through knowledge engineering, 2003

…Datos y sus problemas

- Los **datos sucios** pueden conducir a decisiones erróneas, ocasionando pérdidas de tiempo, dinero y credibilidad:

The Data Warehousing Institute (TDWI) *estima que la mala calidad de los datos del cliente cuesta a empresas en EUA* ***alrededor de $611 billones*** *al año.*

Problemas debido a la mala calidad de datos



Lehmann, C., Roy K., Wintere B., The State of Enterprise Data Quality: 2016. 451 Research, 2016

Un estudio indica que el **2%** de los datos de contacto se vuelven obsoletos cada mes, lo cual cuesta a una organización entre el **15-20%** de sus ingresos operativos.

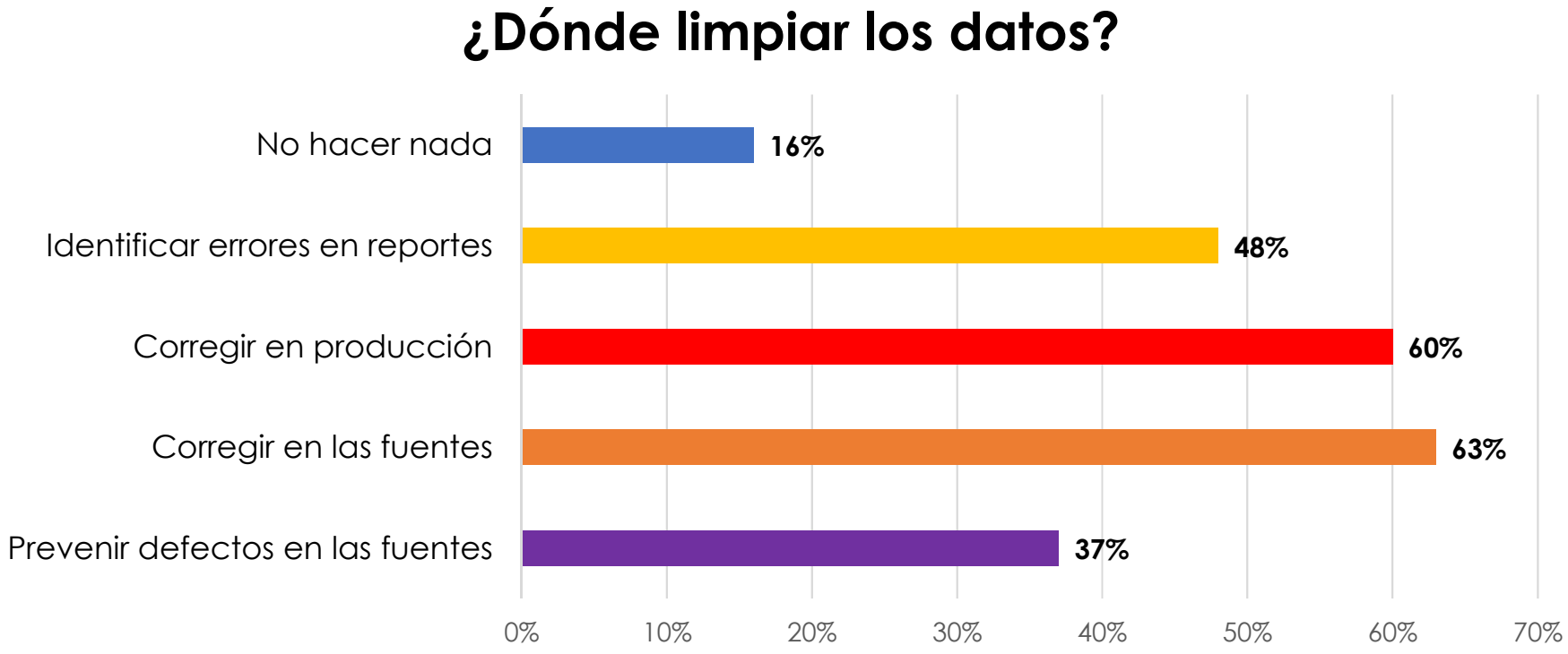
…Datos y sus problemas

“La mala calidad de los datos de los clientes, lleva a costos importantes, como el sobreestimar el volumen de ventas de los clientes, el exceso de gastos en los procesos de contacto con los clientes y pérdida de oportunidades de ventas”

Gartner. *Dirty Data is a Business Problem, Not an IT Problem,* 2007

¿Qué hacer?

- Derivado de lo anterior, muchas compañías están aplicado limpieza de datos para combatir los problemas que se han descrito:
  - ❑ ***Los mismos datos se limpian múltiples veces → diferentes aplicaciones.***
  - ❑ ***Esfuerzos duplicados y altos costos.***



Eckerson, W., Data Quality and the Bottom Line. The Data Warehouse Institute, 2002

…¿Qué hacer?

- En la mayoría de los casos, los proyectos de **"calidad de datos"** surgen cuando ocurre una crisis o algún proyecto clave corre peligro

  Se desarrollan proyectos especiales que no son permanentes ni consistentes con la organización.

- En la mayoría de los casos, no se cuenta con una guía que indique la metodología a seguir para lograr la mejor limpieza de los datos:

  La elección de la técnica está íntimamente ligada con la naturaleza de los datos específicos sobre los que se está trabajando.

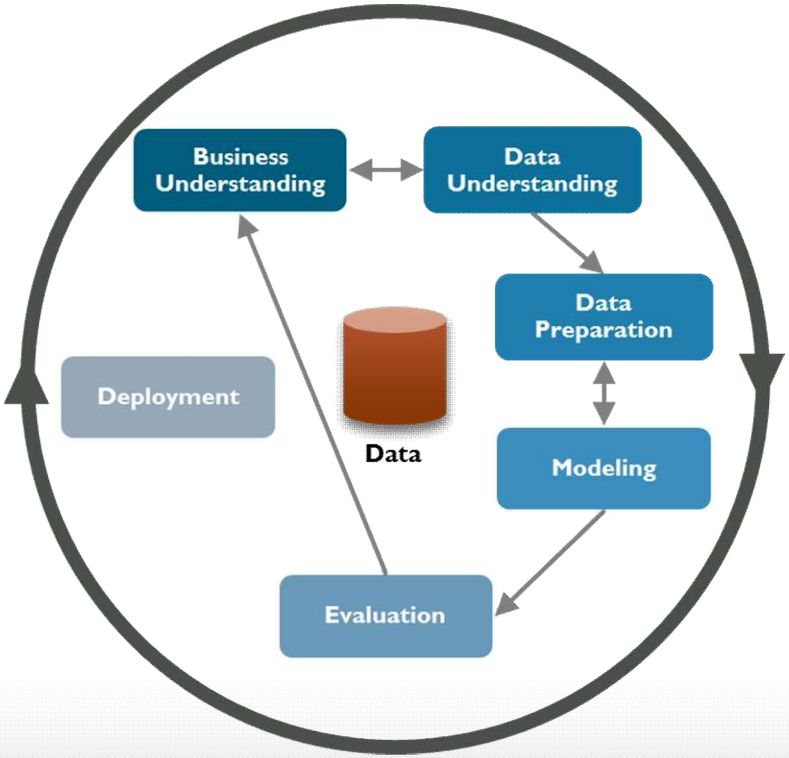
- Las herramientas comerciales que realizan limpieza de datos, en general, no realizan el trabajo de forma autónoma ni automática → ***Intervención del usuario***:
  - ***Se ofrecen un conjunto de opciones***
  - ***Se debe elegir la técnica a ser aplicada a los datos***
  - ***Demanda altos conocimientos técnicos.***

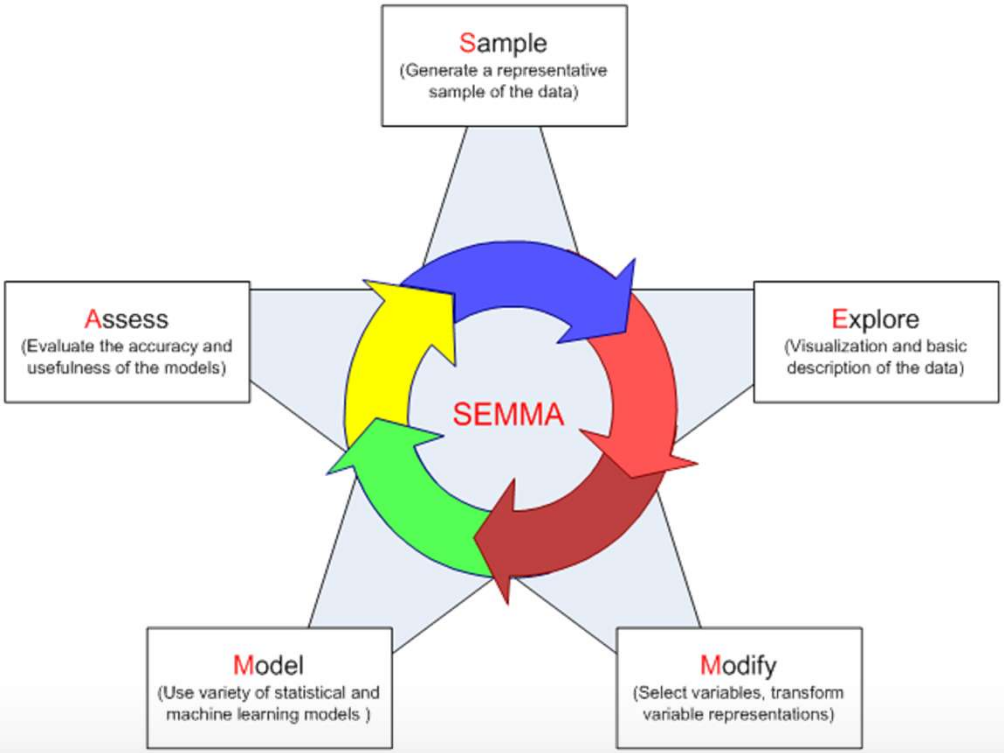
…¿Qué hacer?

- Una metodología ampliamente conocida y usada en proyectos del proceso **KDD** es **CRISP-DM (2000)**:

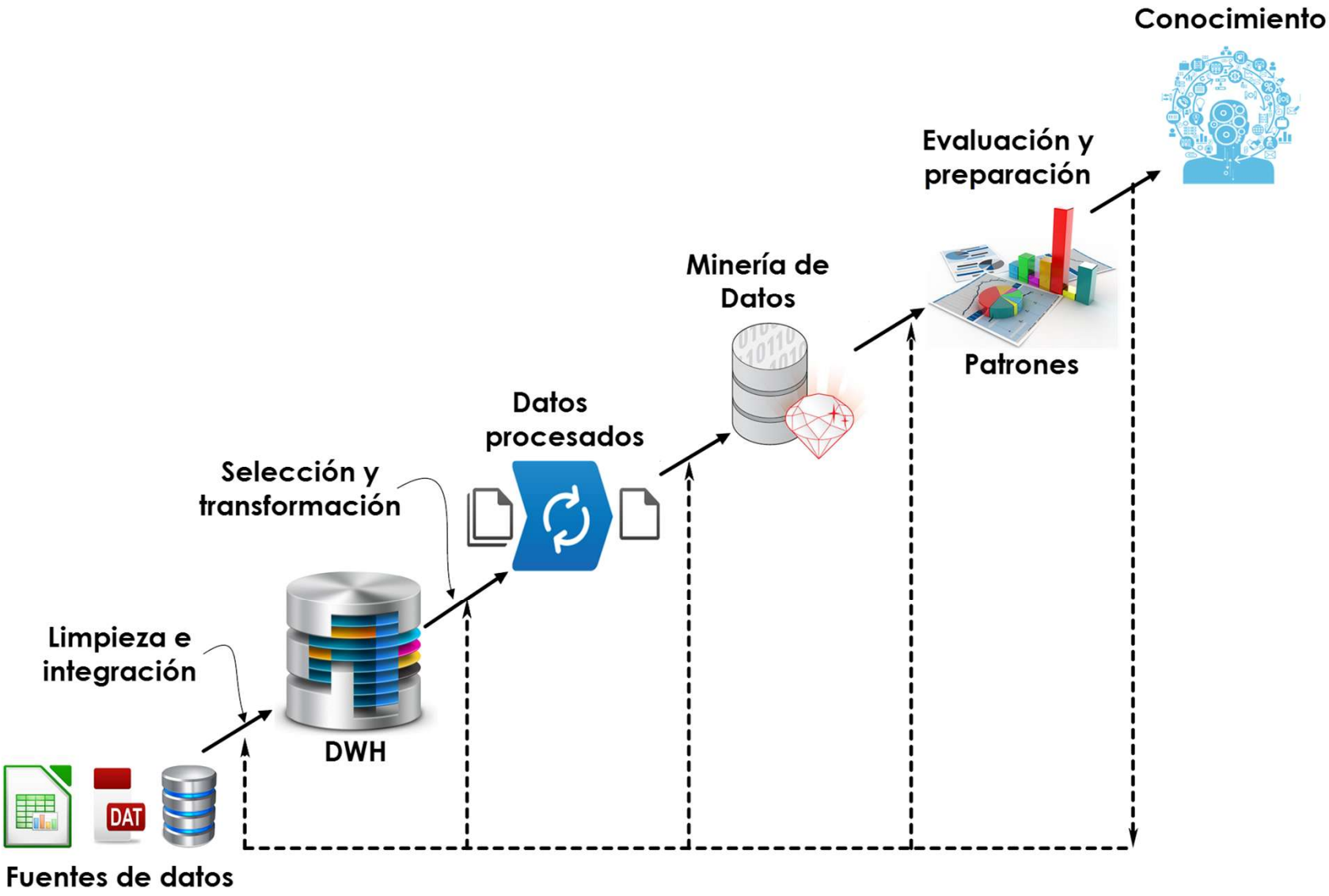
  En su fase de preparación de los datos se ocupa de la transformación y limpieza de los datos pero no llega hasta el nivel de recomendar técnicas específicas dependiendo de la naturaleza de los datos.

- Similar situación sucede con **SEMMA (SAS, 2003).**

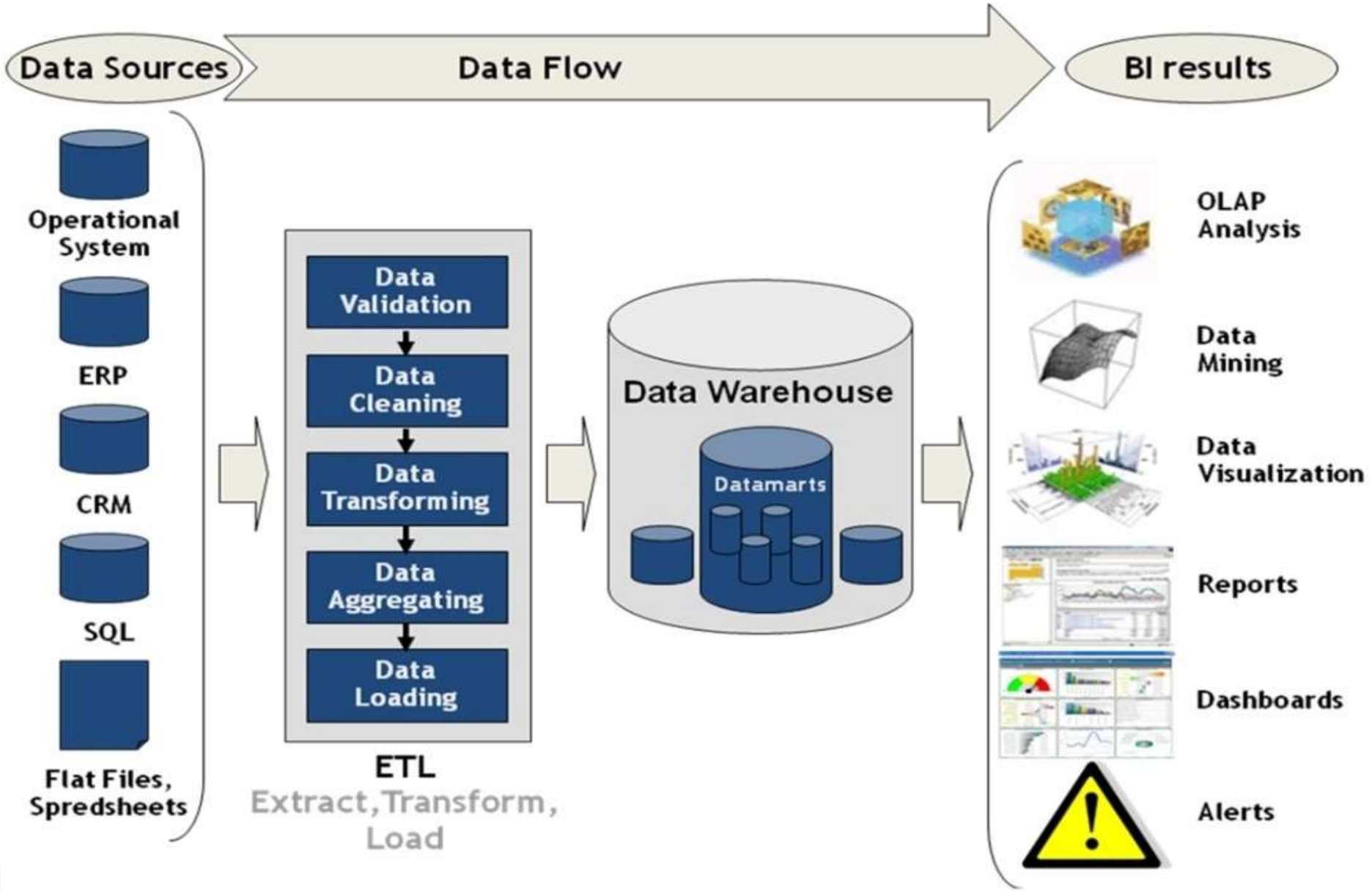




Knowledge Discovery in Databases



Almacenes de datos



… Almacenes de datos

Un almacén de datos es una colección de datos, orientados a un tema, integrados, variante en el tiempo y no volátiles, para apoyar el proceso de toma de decisiones de los ejecutivos" – W. H. Inmon

- **Orientado a un tema.** Colección de información relacionada y organizada alrededor de un ***tema central***: *cliente que realiza una compra, un pedido que contiene artículos.*
- **Integrada.** Usualmente se construye integrando ***múltiples fuentes heterogéneas***. Se requiere técnicas de limpieza e integración de datos para asegurar la consistencia entre los datos.
- **Variante en el tiempo.** Los datos se almacenan para proporcionar información desde una ***perspectiva histórica***. Cada elemento clave contiene explícita o implícitamente un elemento de tiempo.
- **No volátil.** No requiere mecanismos para procesamiento de transacciones, recuperación y control de concurrencia. Sólo requiere dos operaciones para acceder a los datos: ***carga inicial y acceso de datos***.

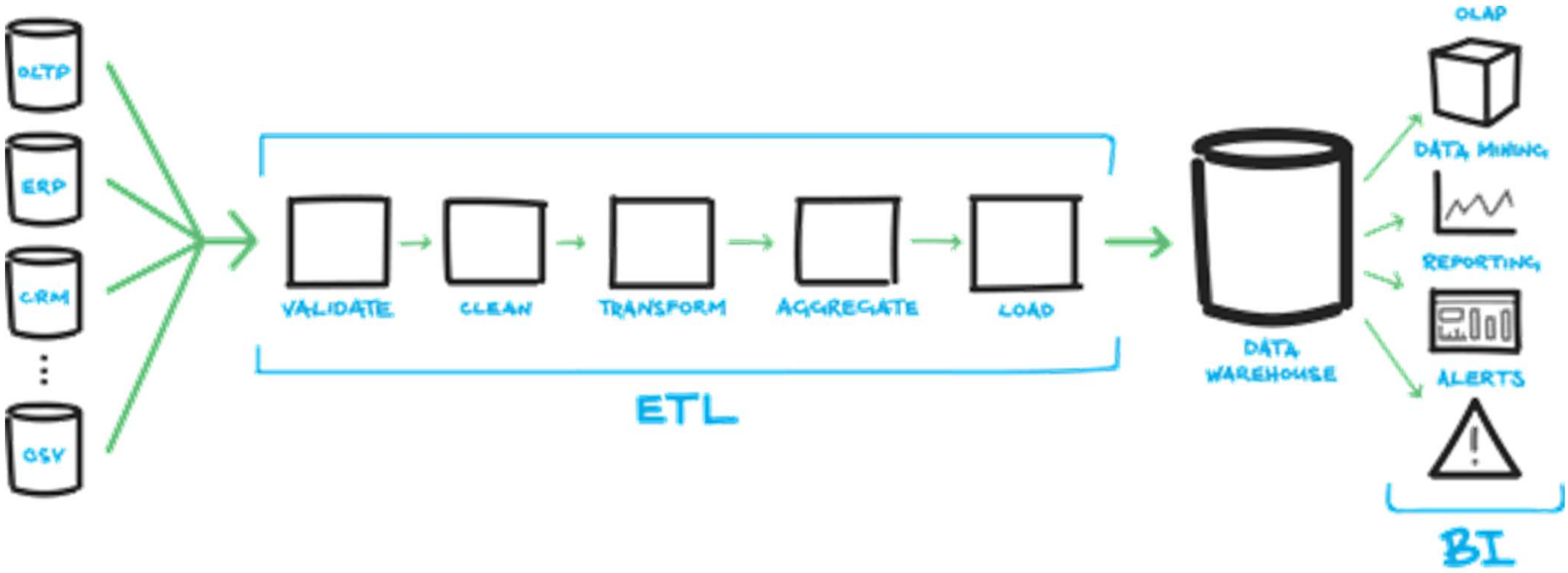
… Almacenes de datos

Un almacén de **datos** es:

- Un ***modelo de datos*** para dar soporte a la toma de decisiones, el cual representa la información que una compañía necesita para tomar **buenas decisiones** estratégicas.
- Basado en la estructura de un ***sistema manejador de bases de datos*** relacional, puede ser usado para interrelacionar los datos contenidos en él.
- Tiene el propósito de proporcionar a los usuarios finales un ***acceso sencillo*** a la información.

Extracción, transformación y carga

- Es el proceso que permite obtener datos desde **fuentes de datos heterogéneas** y cargarlos dentro del **almacén de datos**:



*Los datos se **extraen** de sistemas **OLTP** (no necesariamente), son **transformados** para que coincidan con el esquema del almacén de datos y se **cargan** dentro de la base de datos del almacén.*

…Extracción, transformación y carga

- Se trata del proceso **más subestimado** de todo el desarrollo del DWH y por ende el que consume más tiempo:

Extracción:

Determinar los datos relevantes que se enviarán al DWH.

Transformación

- Transformar los datos para ajustarlos al esquema del DWH.
- Construir llaves
- Limpiar los datos
- Etcétera.

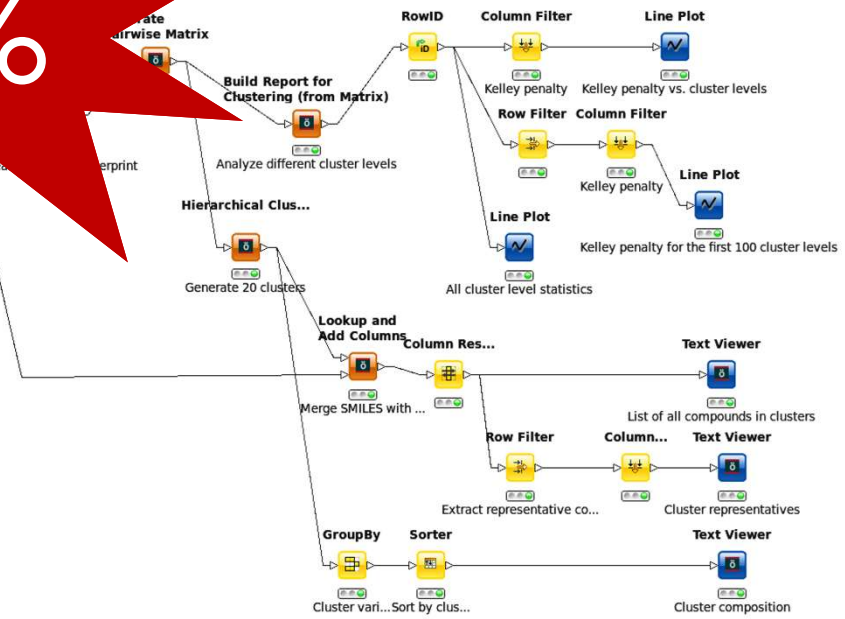
Carga:

- Cargar los datos en el DWH.
- Construir agregaciones
- Etc.

…Extracción, transformación y carga

- Se trata de una combinación compleja de procesos y tecnologías que consumen una porción significativa de los esfuerzos del desarrollo del almacén de datos:

≈80%

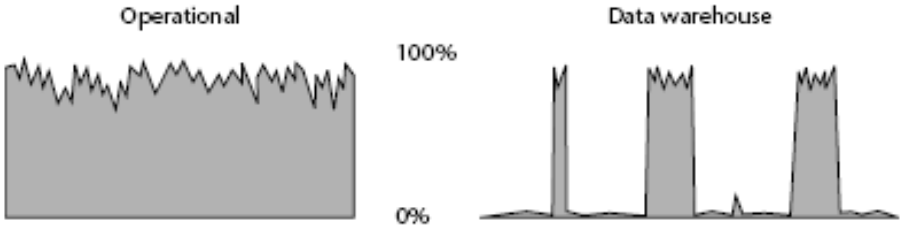


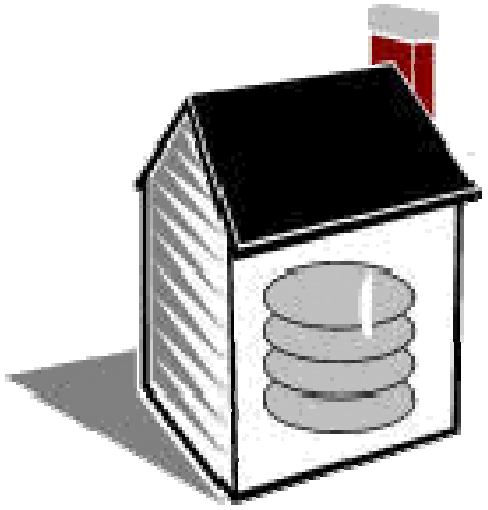
- Analistas de negocio
- Diseñadores de BD
- Desarrolladores de aplicaciones
- Etc.

…Extracción, transformación y carga

- No se trata de un evento que ocurra una vez en el tiempo, ya que la organización evoluciona y continuamente se generan nuevos datos:

  Los nuevos datos se deben añadir al almacén de datos con cierta regularidad: cada hora, cada día, cada mes, etc.

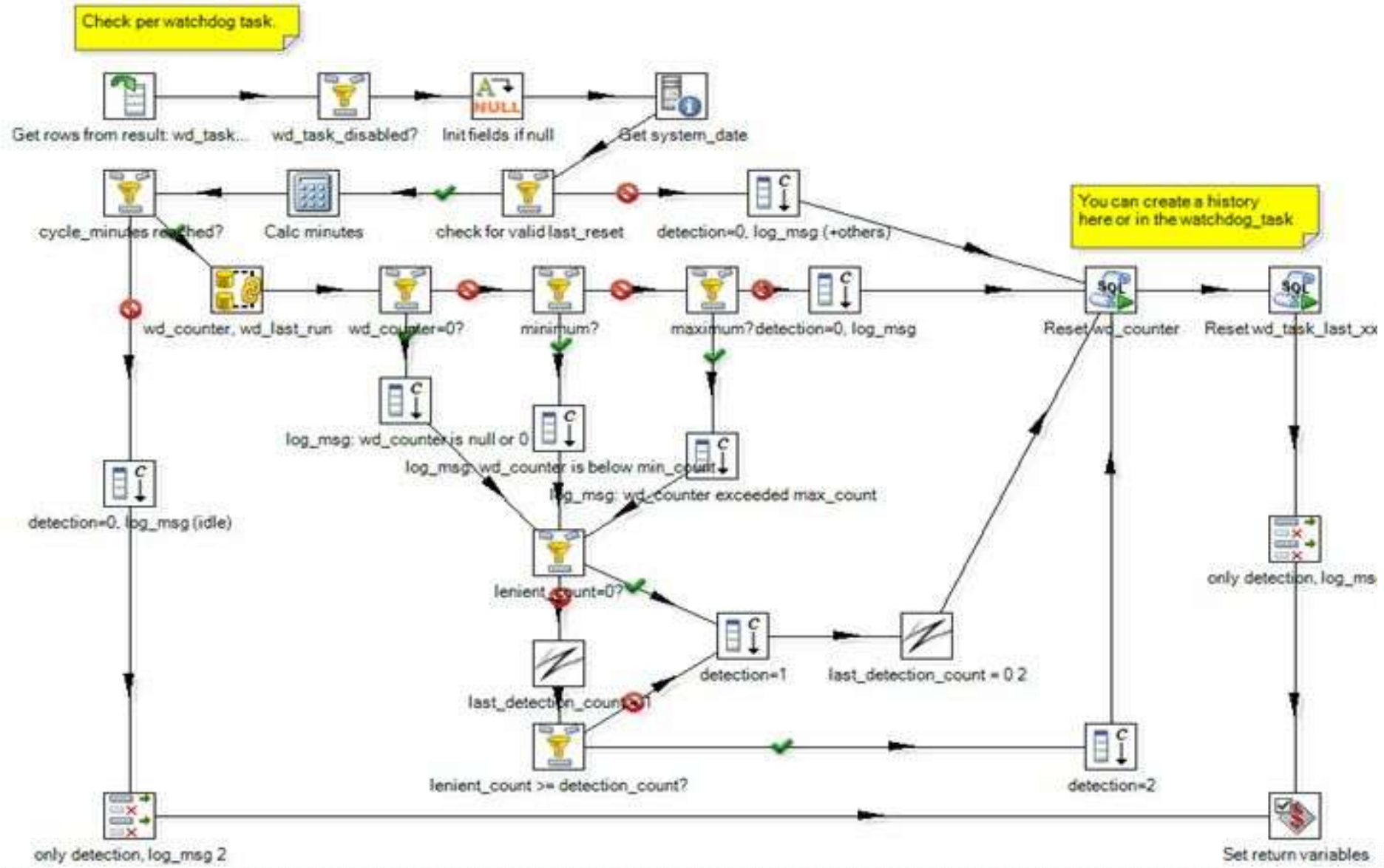




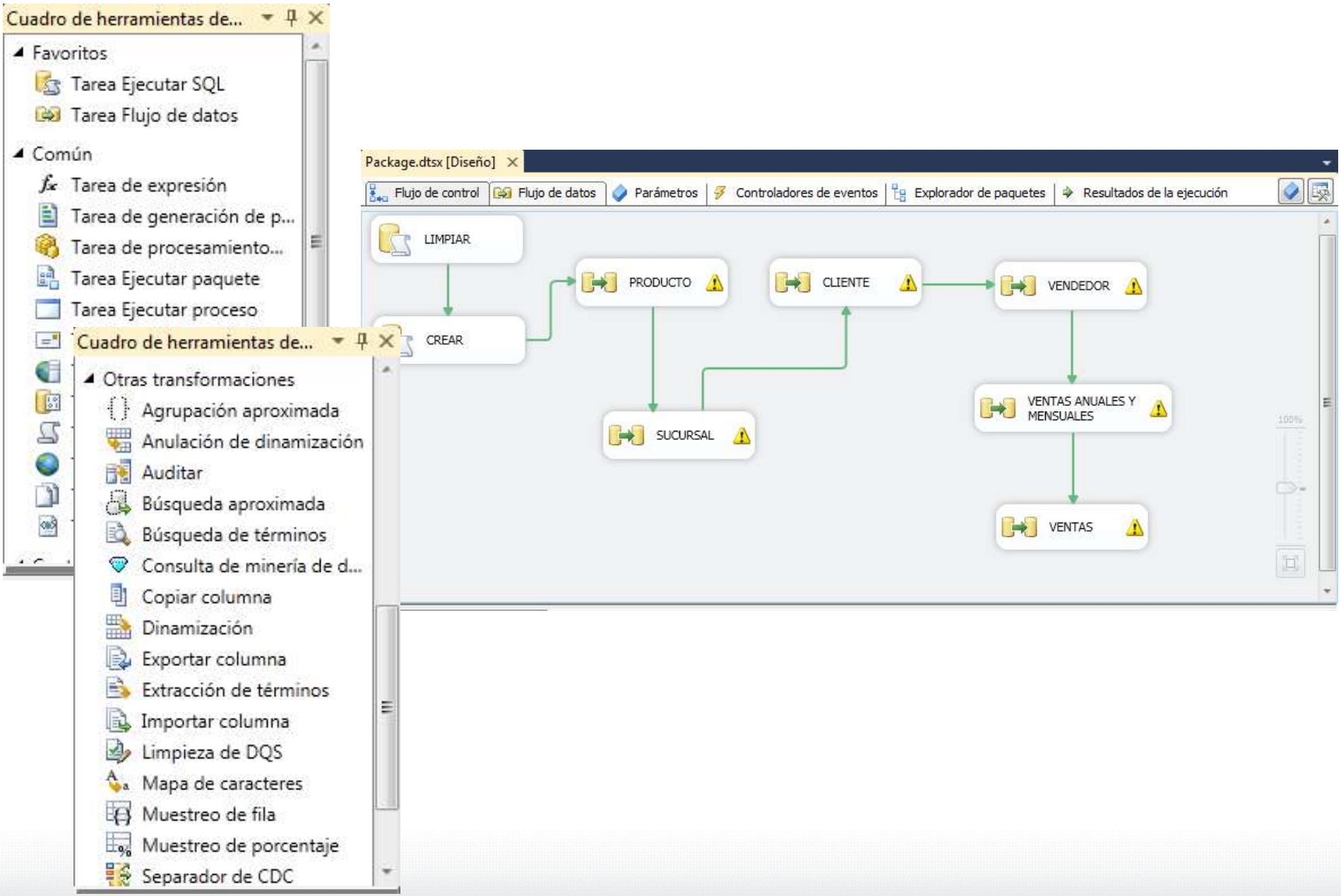
Proceso ETL

- Se define como una **gráfica acíclica dirigida** que se compone de:
  - ***Nodos:*** *actividades, conjuntos de registros, operaciones, etc.*
  - ***Aristas:*** *permiten establecer relaciones de entrada salida entre los nodos.*
- Permite modelar un flujo de actividades para realizar:
  - ***Filtros apropiados***
  - ***Preparación intermedia de los datos***
  - ***Transformaciones***
  - ***Carga***
  - ***Etcétera.***

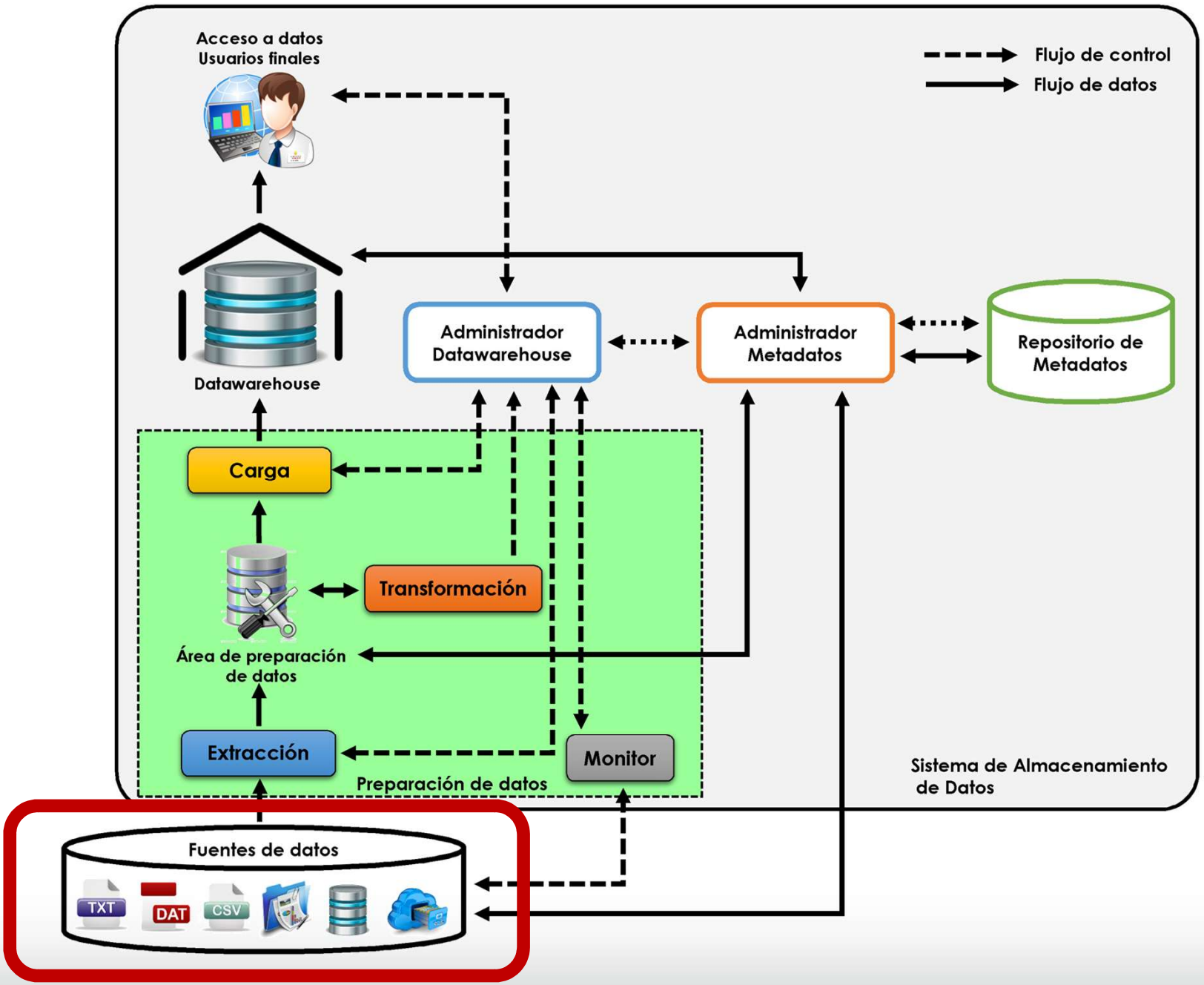
…Proceso ETL


Check per watchdog task.
Get rows from result: wd_task...
wd_task_disabled?
Init fields if null
Get system_date
cycle_minutes reached?
Calc minutes
check for valid last_reset
detection=0, log_msg (+others)
You can create a history here or in the watchdog_task
wd_counter, wd_last_run
wd_counter=0?
minimum?
maximum?
detection=0, log_msg
Reset wd_counter
Reset wd_task_last_xx
log_msg: wd_counter is null or 0
log_msg: wd_counter is below min_count
log_msg: wd_counter exceeded max_count
detection=0, log_msg (idle)
only detection, log_ms
lenient_count=0?
detection=1
last_detection_count = 0 2
last_detection_count=1
lenient_count >= detection_count?
detection=2
only detection, log_msg 2
Set return variables

…Proceso ETL


Cuadro de herramientas de...
Favoritos
Tarea Ejecutar SQL
Tarea Flujo de datos
Común
Tarea de expresión
Tarea de generación de p...
Tarea de procesamiento...
Tarea Ejecutar paquete
Tarea Ejecutar proceso
Cuadro de herramientas de...
Otras transformaciones
Agrupación aproximada
Anulación de dinamización
Auditar
Búsqueda aproximada
Búsqueda de términos
Consulta de minería de d...
Copiar columna
Dinamización
Exportar columna
Extracción de términos
Importar columna
Limpieza de DQS
Mapa de caracteres
Muestreo de fila
Muestreo de porcentaje
Separador de CDC
Package.dtsx [Diseño]
Flujo de control
Flujo de datos
Parámetros
Controladores de eventos
Explorador de paquetes
Resultados de la ejecución
LIMPIAR
CREAR
PRODUCTO
CLIENTE
VENDEDOR
SUCURSAL
VENTAS ANUALES Y MENSUALES
VENTAS

Arquitectura de un DWH

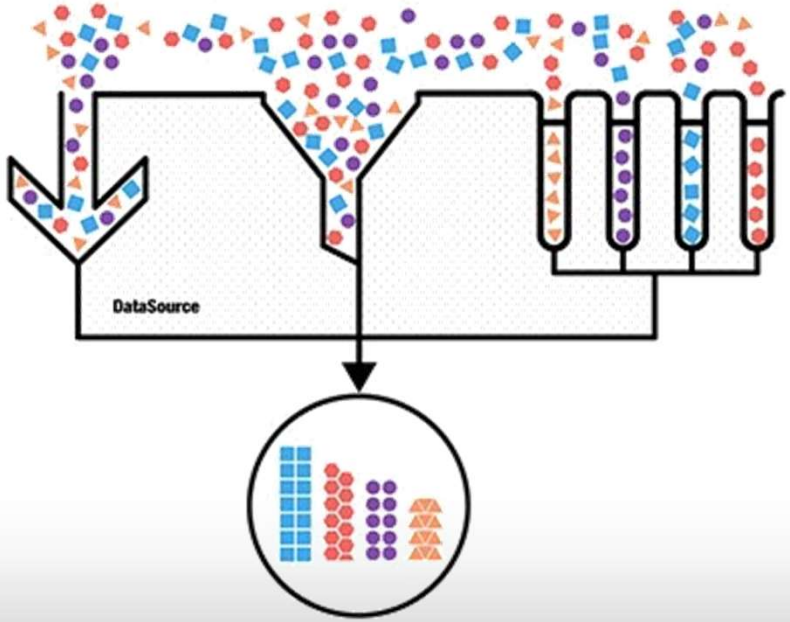


Fuentes de datos

- Se conforma a partir de las fuentes de las cuales se extrae la información.
- En la mayoría de los casos las fuentes son sistemas **OLTP** (sistemas que son diseñados para trabajar de forma independiente):

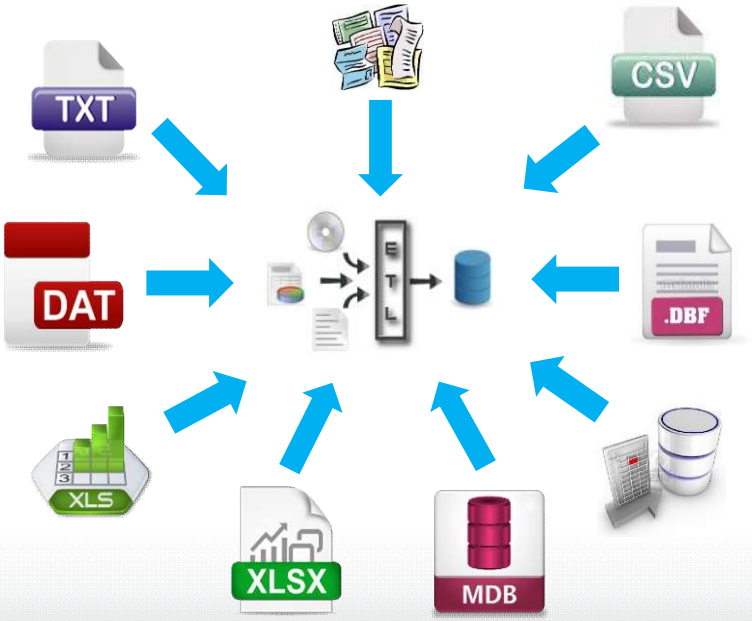
  *Muchos **DWH** incorporan además, datos que no provienen de sistemas **OLTP**: **archivos de texto, sistemas heredados, hojas de cálculo, archivos en papel**.*

- En la mayoría de las organizaciones hay **miles de archivos** que coexisten en un sinnúmero de aplicaciones con una gran redundancia en sus datos y distintos formatos.

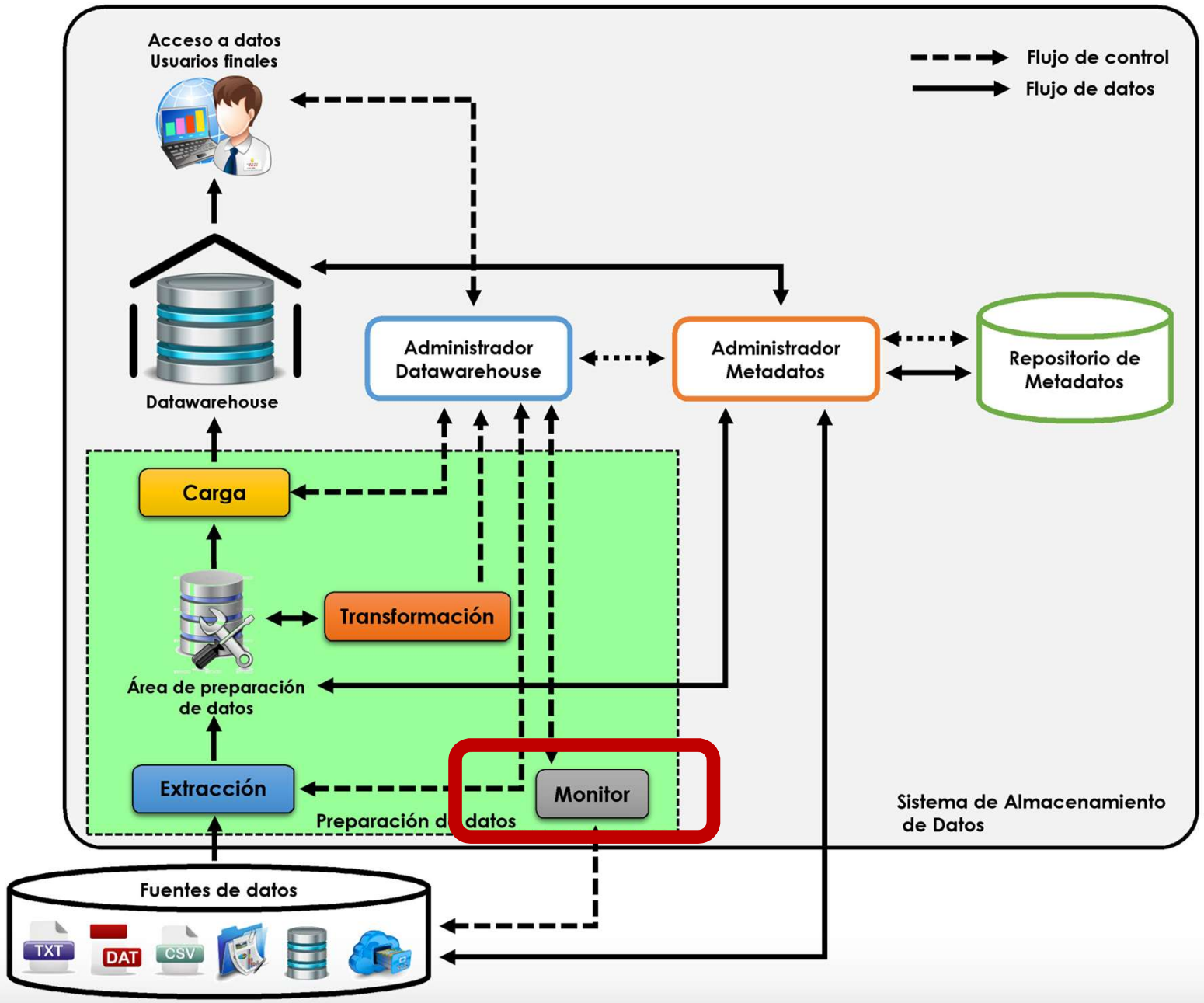


...Fuentes de datos

- Se pueden agrupar en **4 categorías:**
  - **Datos de producción.** Sistemas de procesamiento diario, engloban a todas las transacciones de la organización, manejan grandes volúmenes de información.
  - **Datos internos.** Se trata de sistemas de oficina y bases de datos no convencionales pero que contienen información relevante.
  - **Datos archivados.** Se trata de datos antiguos a los cuales no se accede frecuentemente y se almacenan a un nivel de detalle consistente con los datos actuales.
  - **Datos externos.**



Determinar cambios en los datos



…Determinar cambios en los datos

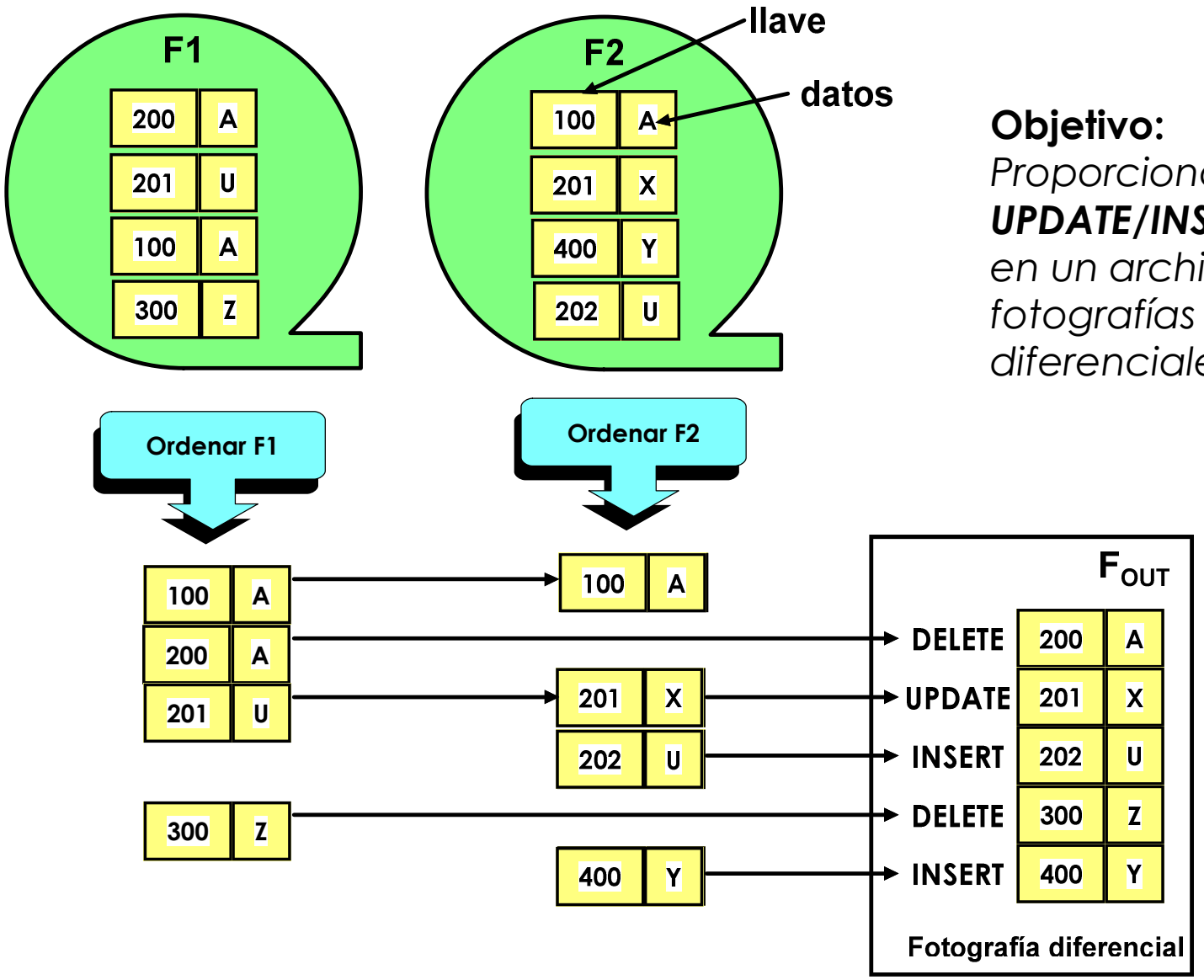
- Durante la **carga inicial**, **no es importante** capturar de cambios en el contenido de los datos en los sistemas de origen, ya que es muy probable que se extraiga toda la fuente de datos (*o una poción*) en un punto predeterminado en el tiempo.
- Posteriormente, se vuelve una **prioridad** la capacidad de identificar cambios en los datos en el sistema de origen: *el equipo de ETL es responsable de la captura de cambios durante la* ***carga incremental***.
- **Auditoría de columnas:** enfoque utilizado por las **BD** y actualizadas por **disparadores**:
  - Las **columnas auditadas** añaden al final de cada tabla la **fecha y la hora** en que se añadió/modificó un registro.
  - Se deben analizar y probar cada una de las columnas para asegurarse de que es una **fuente confiable** para indicar cambios en los datos. Si se encuentran valores **NULL**, se debe encontrar un enfoque alternativo para detectar el cambio (p.e. **outer joins**).
- **Fotografía periódica**

Monitor

- **Objetivo:** *Descubrir cambios en las fuentes de datos de forma incremental.*

Enfoque	Basado en	Cambios identificados por
Disparadores	Definidos en el SMBD	Escriben una copia de los datos modificados en tablas de históricos.
Replicación	Soporte de replicación del SMBD (BD distribuidas)	La replicación proporciona filas modificadas en una tabla separada.
Timestamp	Marcas de tiempo asignadas a cada fila	Se utilizan las marcas de tiempo para identificar los cambios.
Log	Archivo log del SMBD	Leer el archivo log.
Fotografías	Fotografías periódicas de la fuente de datos	Comparar las fotografías.

Monitor: fotografía diferencial



Objetivo:
Proporcionar ***UPDATE/INSERT/DELETE*** *en un archivo de fotografías diferenciales*

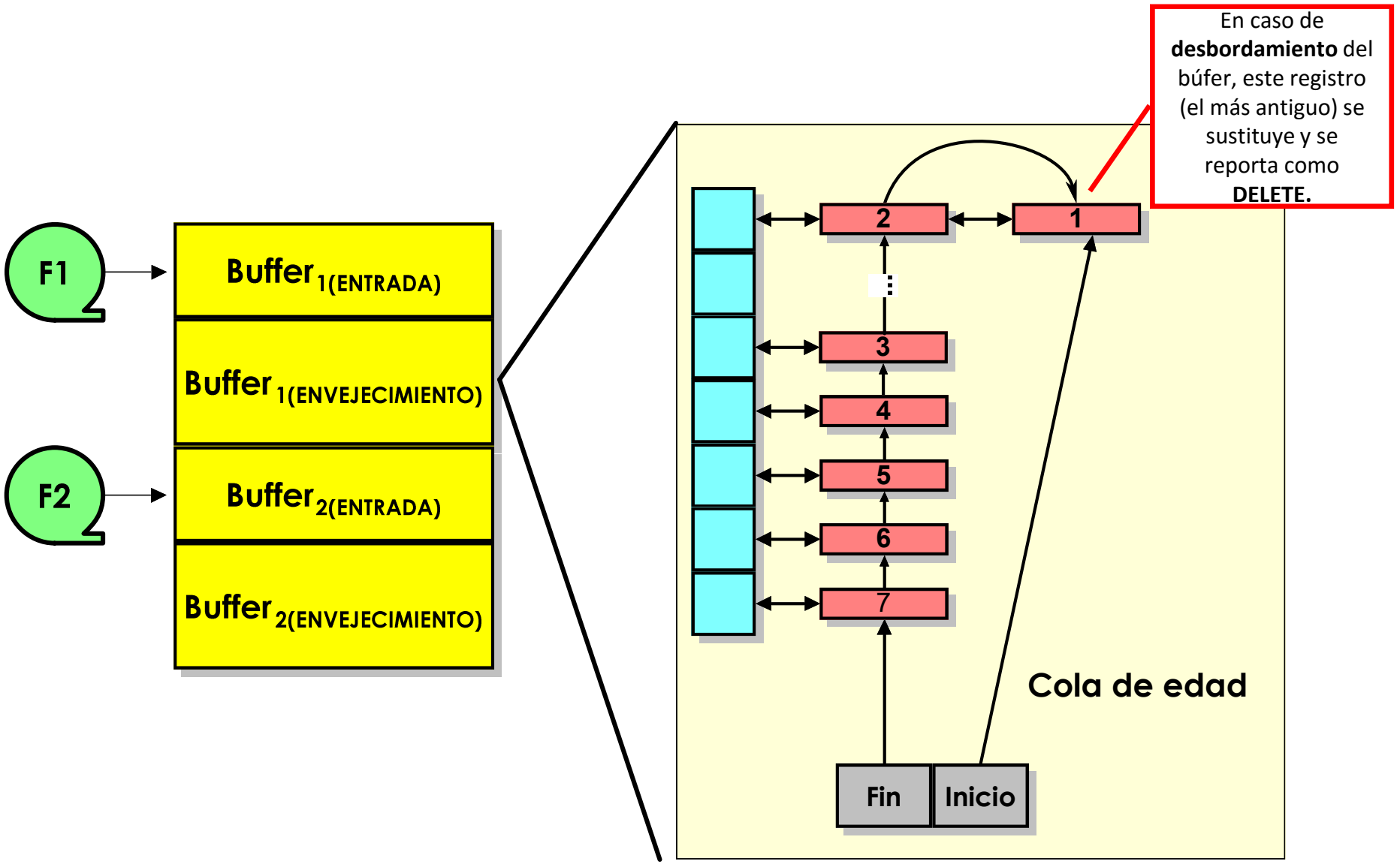
Monitor: algoritmo de ventana

Entrada: **F_1, F_2, n**

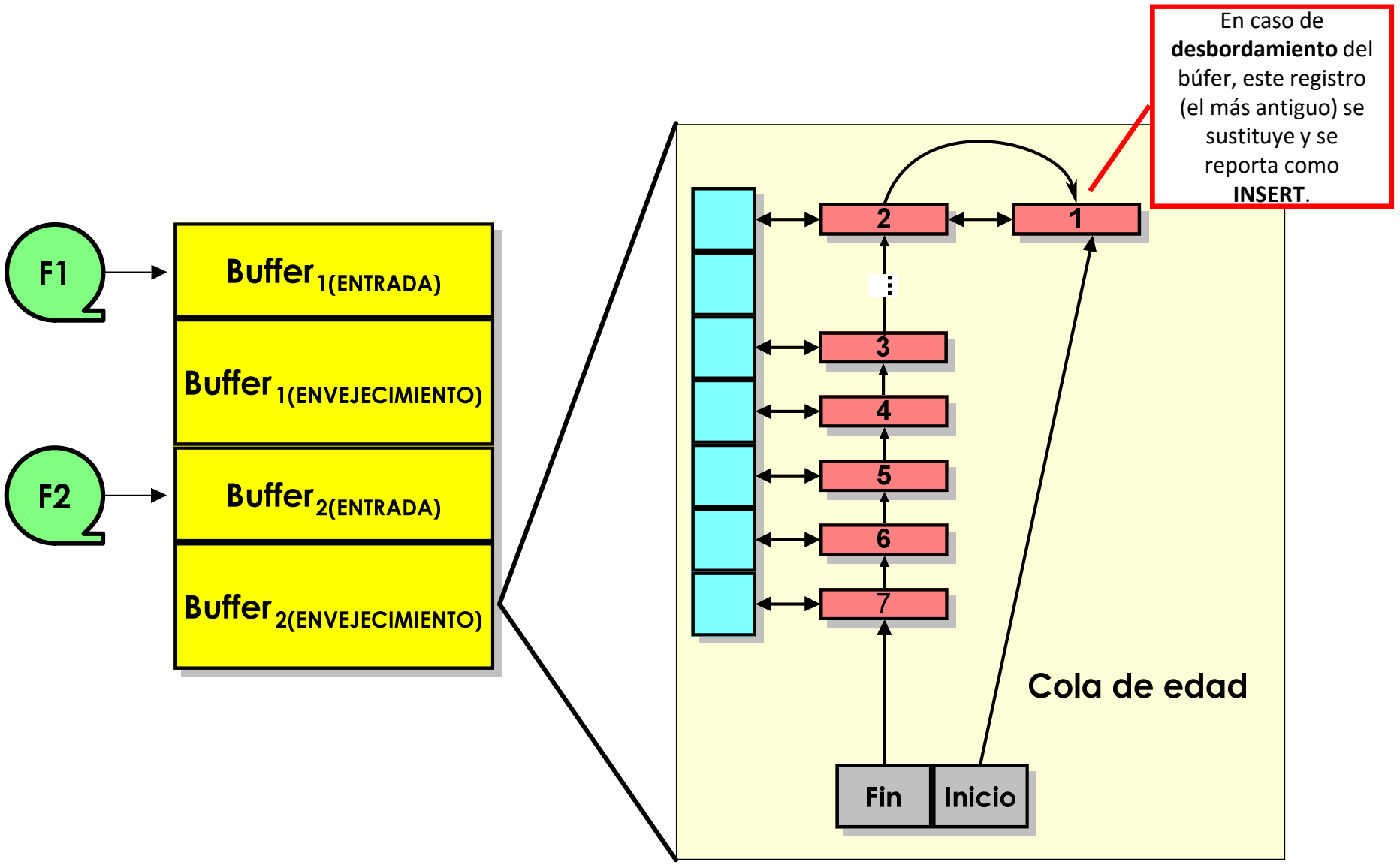
Salida: **F_{out} /*Fotografía diferencial*/**

```
1)  Buffer1(ENTRADA) ← Leer n bloques de F1
2)  Buffer2(ENTRADA) ← Leer n bloques de F2
3)  Mientras ((Buffer1(ENTRADA) ≠ Vacío) y (Buffer2(ENTRADA) ≠ Vacío))
4)       Comparar Buffer1(ENTRADA) vs Buffer2(ENTRADA)
5)       Comparar Buffer1(ENTRADA) vs Buffer2(ENVEJECIMIENTO)
6)       Comparar Buffer2(ENTRADA) vs Buffer1(ENVEJECIMIENTO)
7)       Buffer1(ENTRADA) → Buffer1(ENVEJECIMIENTO)
8)       Buffer2(ENTRADA) → Buffer2(ENVEJECIMIENTO)
9)       Buffer1(ENTRADA) ← Leer n bloques de F1
10)      Buffer2(ENTRADA) ← Leer n bloques de F2
11) Reportar registros en Buffer1(ENVEJECIMIENTO) como borrados (DELETE).
12) Reportar registros en Buffer2(ENVEJECIMIENTO) como inserciones
    (INSERT).
```

…Monitor: algoritmo de ventana



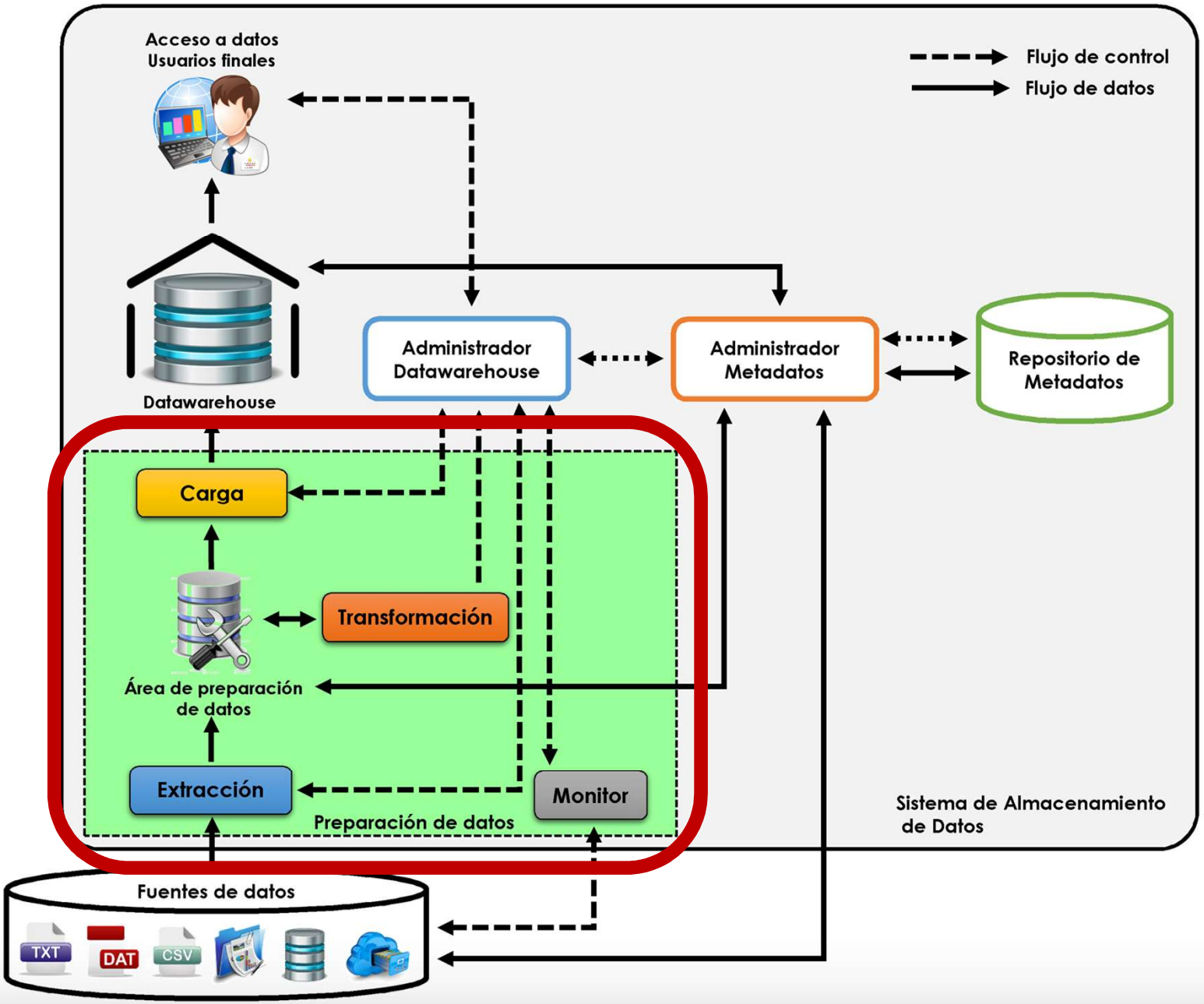
…Monitor: algoritmo de ventana



Carga inicial e incremental

- Crear dos tablas: una para la **carga anterior** y una para la **carga actual**.
- El **proceso inicial de carga masiva** se almacena en la tabla de **carga actual**. Dado que la detección de cambios es irrelevante durante la carga inicial, los datos se transforman y se carga en la **tabla de hechos destino**.
- Cuando el proceso se haya completado, se cambia el nombre de la tabla de **carga actual** a **carga anterior** y se deja a la tabla de carga actual vacía (operaciones muy rápidas)
- La próxima vez que el proceso de carga se ejecute, la tabla de carga actual se poblará con los nuevos datos.
- Se selecciona la tabla de **carga actual** y se resta la tabla de **carga anterior**. Transformar y cargar el conjunto resultante en el **DWH**.

Preparación de datos



…Preparación de datos

- Es una área intermedia en la arquitectura del **DWH**, donde las fases de **transformación**, **integración** y **limpieza** tienen lugar.
- Permanece entre las fuentes de datos y el **DWH** con el fin de:
  - *Facilitar la extracción de datos desde las fuentes de origen realizando un **pre-procesamiento**.*
  - *Realizar **limpieza de datos** (data cleaning)*
  - *Mejorar y asegurar la calidad de datos.*
  - *Permite acceder en detalle a información no contenida (aún) en el **DWH***
- Es recomendable que las operaciones **ETL** se realicen en un servidor de **BD relacional** separado (física y lógicamente) de las fuentes de datos y del **DWH**:

  Minimizar el impacto de la actividad periódica de los procesos ETL

- Las **herramientas ETL**, son **piezas de software** responsables de la extracción de los datos de distintas fuentes, su depuración, personalización e inserción dentro de un **DWH.**

…Preparación de datos

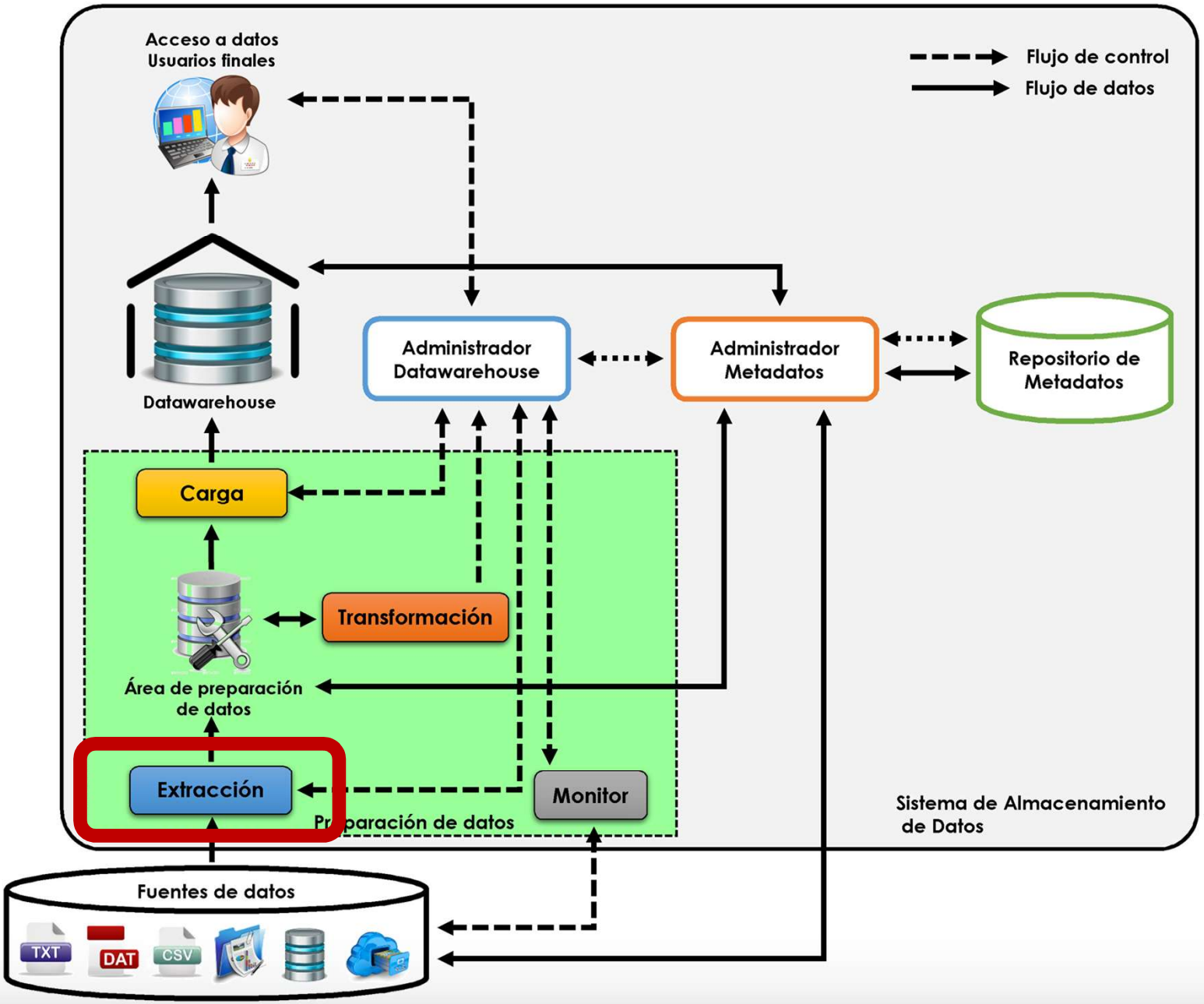
- Se encarga de:
  - *Tomar **fotografías** (periódicas), de las fuentes de datos (tan rápido como sea posible) a fin de comparar con las versionas previas y de esta forma detectar los valores que serán **actualizados/insertados** en el **DWH**.*
  - *Almacenar en disco los datos nuevos, a fin de asegurar que el proceso no comience desde cero en caso de que el sistema falle.*
  - *Aplicar múltiples transformaciones y filtros a los datos.*
  - *Restaurar el sistema sin repetir los procesos desde el inicio*

- Características:
  - **Almacena** los datos tan pronto como son extraídos de las fuentes de datos e inmediatamente después de que los datos han sufrido procesos mayores (*limpieza, transformaciones, etc.*)
  - **Capacidad** de **re-cargar** los datos que llegaron al área de preparación, sin necesidad de ir hasta las fuentes de datos operacionales.
  - **Realiza auditorías** entre los datos de origen y las transformaciones realizadas, antes de la carga al **DWH**.

…Preparación de datos

- Debe asegurar que :
  - ***Sea administrada por el equipo ETL:*** *no hay índices, no hay agregaciones, sin acceso para presentación, de datos, sin consultas.*
  - ***Los usuarios no tengan permitido utilizar esta área por ningún motivo:*** *se trata de sitio de "construcción".*
  - ***Los reportes no tengan acceso a datos en el área de presentación:*** *se pueden añadir/borrar tablas sin intervención de los usuarios.*
  - ***Solo los procesos ETL puedan leer/escribir en esta área:*** *los desarrolladores* ***ETL*** *son los responsables de definir el esquema, estrategias de actualización, frecuencia de carga, flujos* ***ETL****, etc. .*

Extracción


Acceso a datos
Usuarios finales
Flujo de control
Flujo de datos
Datawarehouse
Administrador
Datawarehouse
Administrador
Metadatos
Repositorio de
Metadatos
Carga
Transformación
Área de preparación
de datos
Extracción
Monitor
Preparación de datos
Sistema de Almacenamiento
de Datos
Fuentes de datos
TXT
DAT
CSV

…Extracción

- Objetivo:

  Identificar el subconjunto correcto de fuentes de datos que serán enviadas a los flujos de trabajo ETL para su posterior procesamiento y extracción rápida.

- Este proceso se realiza **sin detener** los sistemas **OLTP** o el **DWH**:

  Se lleva a cabo en los tiempos de inactividad de los sistemas fuente, normalmente por la noche.

- Debe asegurar:
  - Los sistemas fuente sufran sobrecarga mínima.
  - Interferencia mínima en la configuración de SW del lado de la fuente.
- Requiere **integrar de forma efectiva** sistemas que tienen diferentes: ***SMBD****, sistemas operativos, hardware, protocolos de comunicación, etc.*

…Extracción

- Se requiere tener un **mapa lógico** de los datos antes de que los datos físicos puedan ser transformados:
  - ***Especificar*** *dónde están los datos y a dónde los queremos llevar.*
  - ***Determinar*** *las estructuras de almacenamiento físico.*
  - ***Indicar*** *la mejor forma de obtener los datos y llevarlos al área de preparación.*
  - ***Planear*** *cuando deberá comenzar y cuánto tomará el finalizar el proceso.*

- Este mapa describe la relación entre los puntos de **inicio** y **fin** del sistema **ETL** (*suele ser una tabla u hoja de cálculo*):
  - Permite planificar eficientemente los procesos **ETL**: ***dimensiones → hechos***.
  - La columna **Transformación** puede contener cualquier cosa: a menudo, la transformación se puede expresar en **SQL** (*puede o no ser la declaración completa*).

Objetivo			**Fuente**			**Transformación**
Nombre de la tabla	Nombre de la columna	Tipo de dato	Nombre de la tabla	Nombre de la columna	Tipo de dato	

Fase de descubrimiento de datos

- El criterio clave para el éxito del **DWH** es la **limpieza** y la **cohesión** de los datos dentro de éste.
- El equipo de **ETL** debe determinar todos los **requerimientos de datos:** cada sistema de origen, tabla y atributo necesario para cargar el **DWH**:
  - **Recolectar y documentar** todos los sistemas fuente:

    Importante porque en la mayoría de las empresas, los datos se almacenan redundantemente a través de sistemas diferentes:

    *La mayoría de las empresas hacen esto para que los **sistemas no integrados** puedan compartir datos. Es muy común que el mismo dato se copie, se mueva, se manipule, se transforme, se altere, se limpie, o se corrompa a través de toda la empresa, lo que resulta en **diferentes versiones** del mismo dato.*

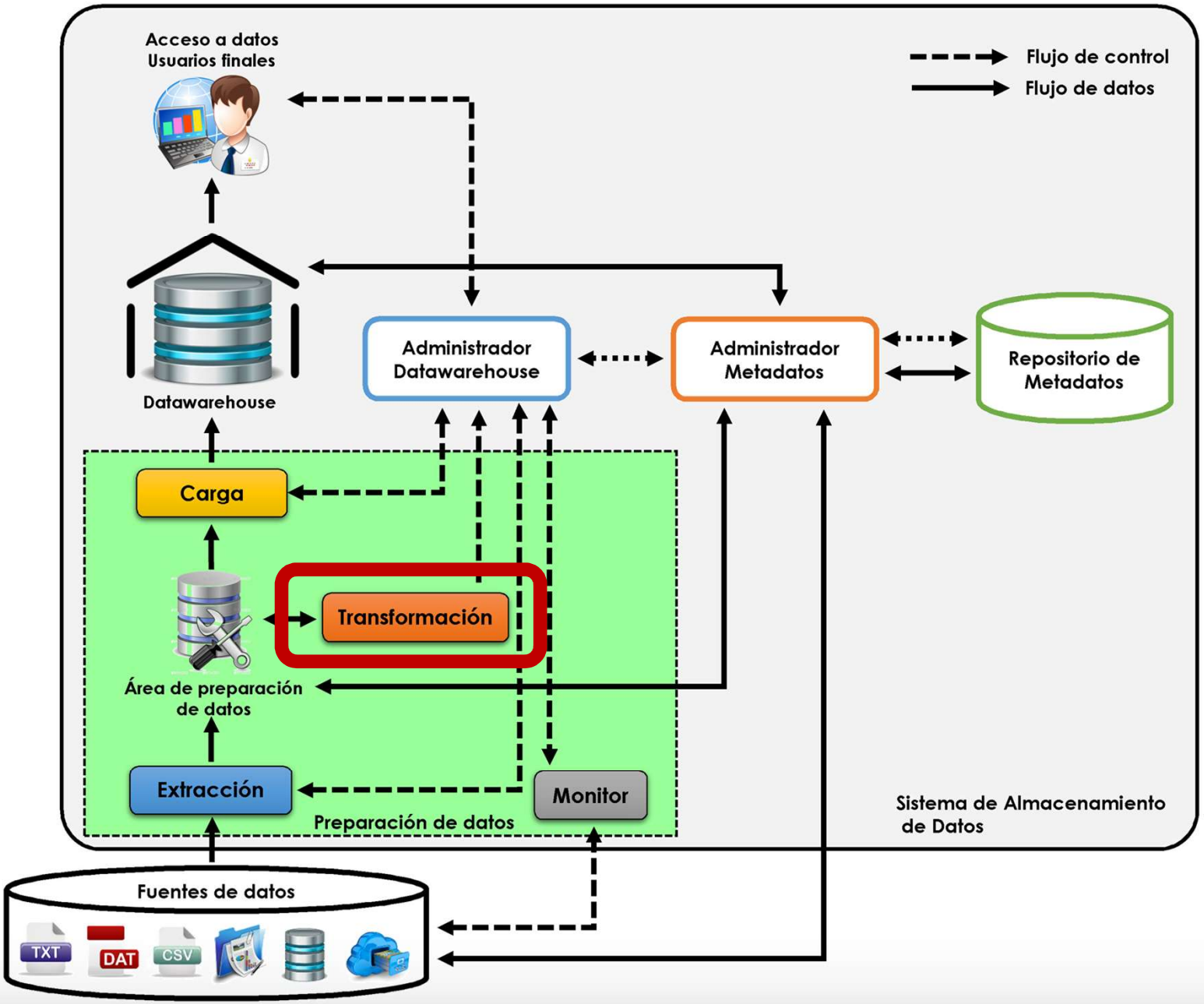
Análisis del contenido de los datos

- Comprender el contenido de los datos es crucial para determinar el mejor enfoque para la recuperación

  - **Valores NULL.** Los valores **NULL** no controlados puede destruir cualquier proceso **ETL** y plantean el mayor riesgo cuando están en las columnas **llave foránea** ya que causan pérdida de datos. Es necesario verificar si hay valores **NULL** en todas llaves foráneas en la base de datos de origen, si es el caso, realizar **join externos** en las tablas.

  - **Fechas.** Los campos para fecha son elementos muy peculiares, ya que son los únicos elementos lógicos que pueden venir en **varios formatos**, que contienen valores diferentes y que tienen el **mismo significado**. Afortunadamente, la mayoría de los sistemas de bases de datos soportan la mayor parte de los distintos formatos con fines de exhibición, pero los almacenan en un solo formato estándar (**yyyy-mm-dd**).

Transformación



…Transformación

- Es el paso más importante en todo el proceso **ETL** ya que agrega valor a los datos.
- Convierte los datos en algo que sea representable y con valor para el negocio.
- Aplica para la **semántica**, **estructura** y **datos**.
- Involucra:
  - **Limpieza de Datos (Data Cleaning)**
  - **Datos no existentes (Missing values)**
  - **Datos extremos (Outliers)**
  - **Integración de esquemas**

Limpieza de datos

- El objetivo es lograr construir un **DWH** que refleje una imagen válida y consistente del negocio para el cual se está implementando: *poner en práctica los procesos de* ***limpieza y mejoramiento de calidad de los datos***.
- Definiremos limpieza de datos como:

La actividad de convertir datos de origen en datos de destino, sin errores, sin duplicados, sin inconsistencias, discrepancias.

"This is not what I meant when I said 'we need better data cleansing!'"

www.iwaysoftware.com/go/dataquality

…Limpieza de datos

Los datos en el mundo real están sucios, debido a que:

- **están incompletos.** Carecen de valores en algunos atributos, falta de ciertos atributos de interés o que contiene sólo datos agregados:

  p.e. **ocupación: , delegación: "Coyoacán", C.P.:**

- **tiene ruido.** Contienen errores o valores atípicos (*errores ortográficos, fonéticos y de captura, transposición de palabras, varios valores en un solo campo de forma libre*)

  p.e. **nombre: "Pteer", "salario: "- 10"**

  p.e. **nombre: "CDMX", ciudad: "Carlos Sánchez"**

- **tienen inconsistencias.** Contienen discrepancias en códigos o nombres (*sinónimos, homónimos, parónimos, apodos, alias, abreviaturas, variaciones de prefijo y sufijo, truncamiento e iniciales*)

  p.e. **edad: 42, f_nac: '03/07/1998' o delegación: "Coyoacán", C.P.:06760**

  p.e. **la calificación era 1,2,3 → A, B, C**

  p.e. **clave: 100, del: "Cuauhtémoc", C.P.: 06760** vs.

  clave: 2, del: "Cuauhtémoc", C.P.: 06760

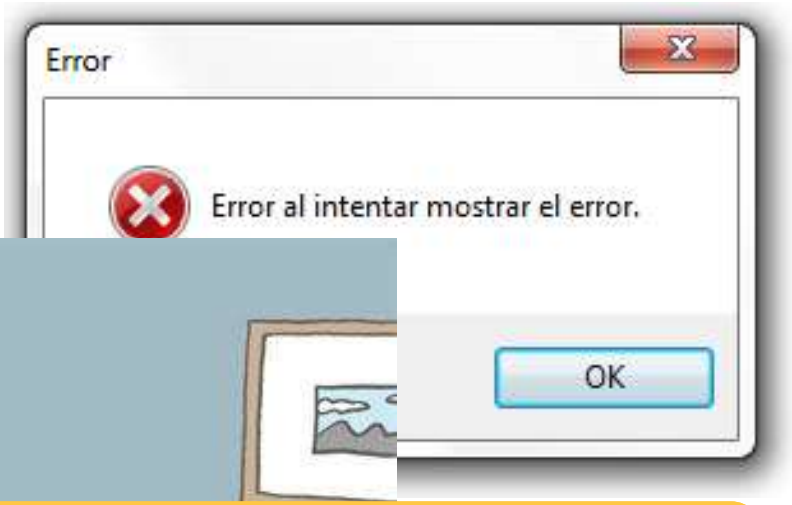
…Limpieza de datos

Problema	Errores
Falta de estándares	Los datos se representan en múltiples formatos. **p.e. diferentes formas de escribir nombres o fechas.**
Información perdida en campos de texto	Información decisiva no es ingresada en los campos de texto correspondientes.
Información no consolidada	Múltiples identificadores de una misma entidad. **p.e.: un mismo cliente con distintos códigos.**
Comparación compleja y consolidación	Entidades de negocio representadas en una amplia variedad de formas, dificulta relacionar todas las instancias o condensarlas en una representación única
Sorpresas de datos dentro de campos individuales	Valores en los datos que escapan de las descripciones de los campos y de las reglas de negocio. **p.e.: nombres comerciales mezclados con nombres personales, indicaciones de ubicación en campos de direcciones, uso inconsistente del espacio en blanco, caracteres especiales y límites de campos (cortar una palabra en un campo y continuar en el siguiente).**

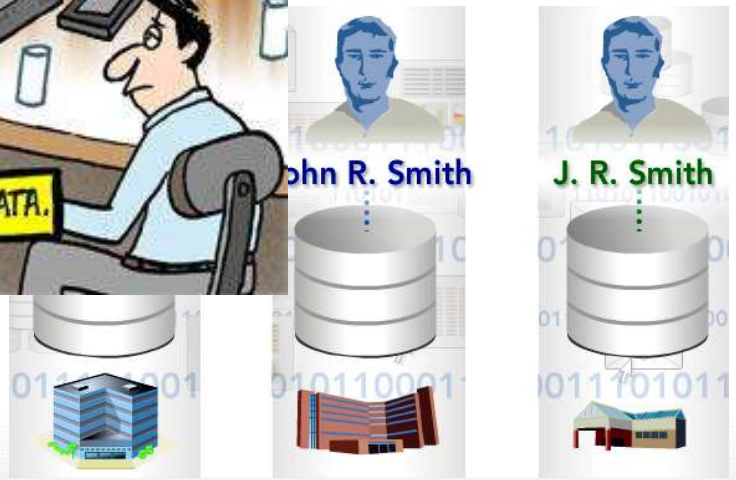
…Limpieza de datos

Problema	Errores
Errores	Errores tipográficos, faltas de ortografías, valores fuera de rango y tipos de datos incorrectos.
Homónimos	Palabras que se escriben igual y tienen diferente significados, a veces hasta sin relación o conflictivos, y su significado correcto depende del contexto.
Datos que faltan o datos invisibles	Datos con estructura y valor apropiados pueden omitir información inadvertidamente. **p.e. la dirección de una persona "J.M. Perez 324" puede ser valida pero si es un edificio se estaría omitiendo el numero interior.**
Datos Fantasma	En ocasiones se ingresan valores especiales indicando que el campo tiene valor desconocido o que no se utiliza más. **p.e. en un campo de fecha se puede encontrar el valor 99/99/9999**

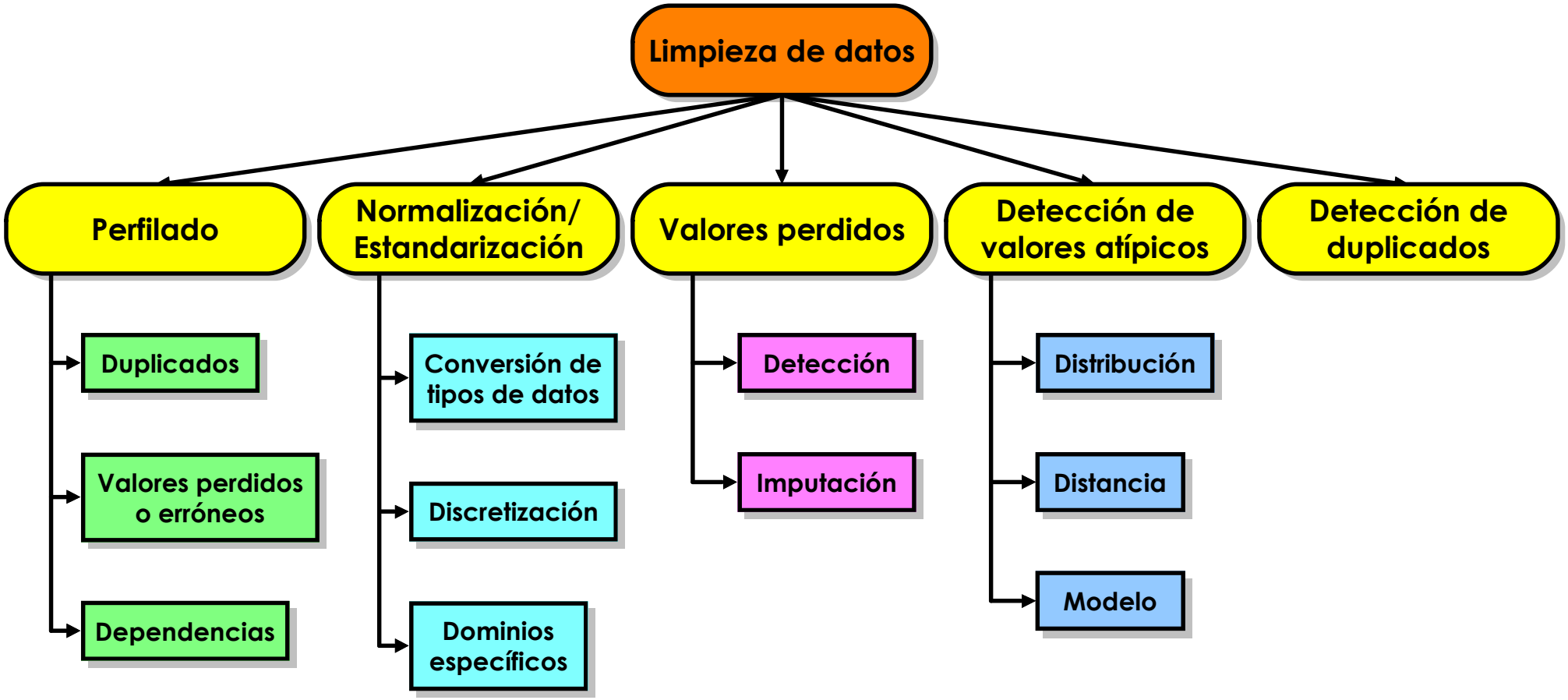
…Limpieza de datos



¡Bienvenidos al mundo real!



Tareas de limpieza de datos



Valores perdidos

Las técnicas que se pueden utilizar cuando nos percatamos que en una tabla no tenemos registros en varias tuplas:

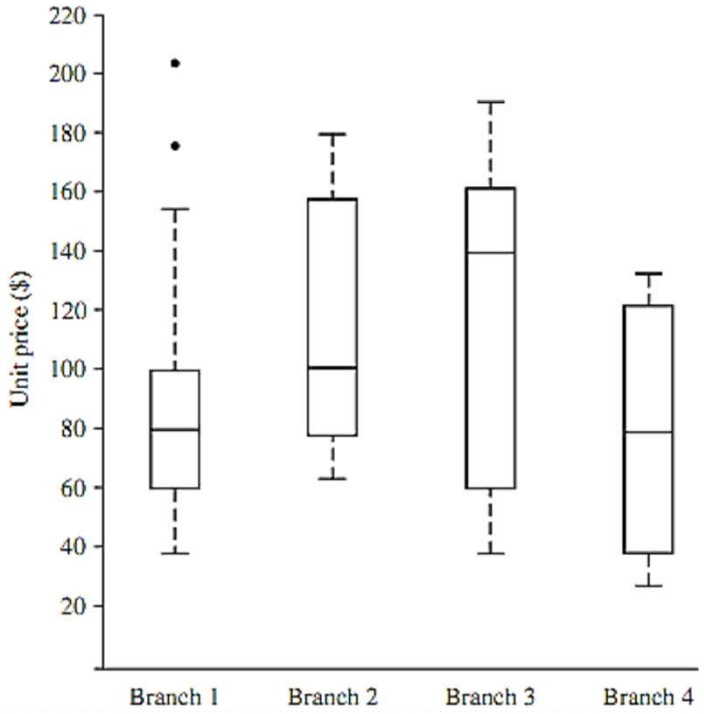
1. **Ignorar la tupla**
2. **Llenar el valor manualmente**
3. **Usar una constante global**
4. **Utilizar una medida de tendencia central**
5. **Utilizar la media (o mediana) para todas las muestras que pertenezcan a la misma clase**
6. **Utilizar el valor más probable**

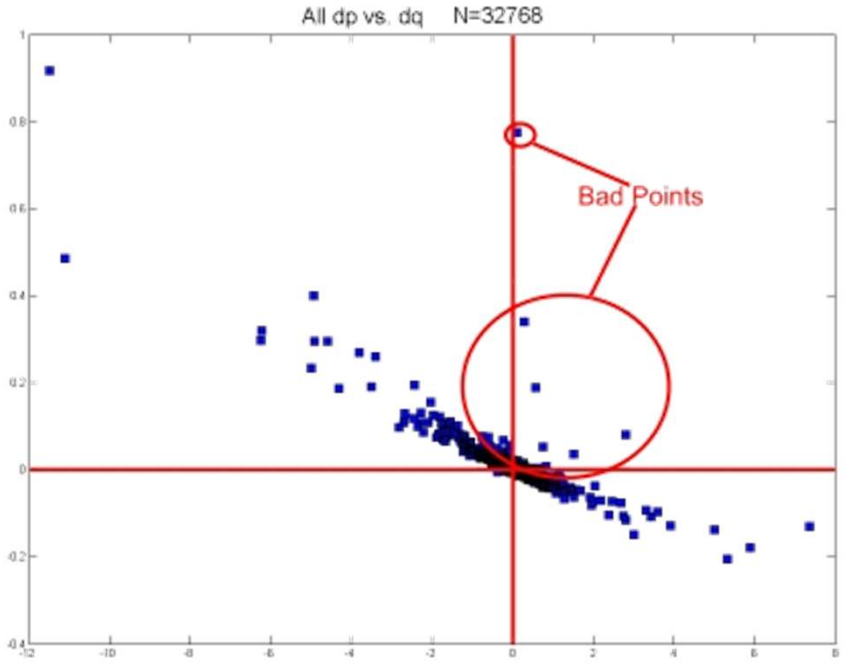
Datos ruidosos

- ¿Qué es el ruido?

  El ruido es un error aleatorio o variación en una medida.

- En los datos, se suelen emplear técnicas de **estadística descriptiva** o **métodos de visualización** para identificar dichos valores





Atenuar o remover el ruido

- **Binning**
  - *Es un método que atenúa el ruido en un conjunto ordenado, consultando los valores alrededor de él. Los valores son distribuidos cubetas a través de un suavizado local.*
- **Análisis de Regresión**
  - Técnica que ajusta los datos a una función. Involucra encontrar la **mejor línea** que ajusta dos atributos (o variables) de manera que un atributo puede ser utilizado para predecir otro.
  - La **regresión múltiple** es una extensión de la regresión lineal en donde se involucran más de dos atributos que se ajustan a una superficie multidimensional.
- **Análisis de datos atípicos (Outliers)**
  - *Los valores atípicos pueden detectarse utilizando técnicas de agrupación (clúster), donde los valores son organizados en grupos de valores similares. Intuitivamente los valores que caigan fuera de esos grupos se consideran como valores atípicos.*


Después de
la limpieza …
¿qué sigue?

Esquemas heterogéneos

- Fuentes con diferentes esquemas almacenan información redundante
- Queremos ser capaces de trasladar los datos de un esquema hacia diferentes esquemas.
- Queremos ser capaces de traducir consultas que se realizan en un esquema a consultas similares en otros esquemas.
- ¿Qué necesitamos?
  - *Conocer los elementos de diferentes esquemas que están relacionados.*
  - *Coincidencia de esquemas*
  - *Mapeo de esquemas*

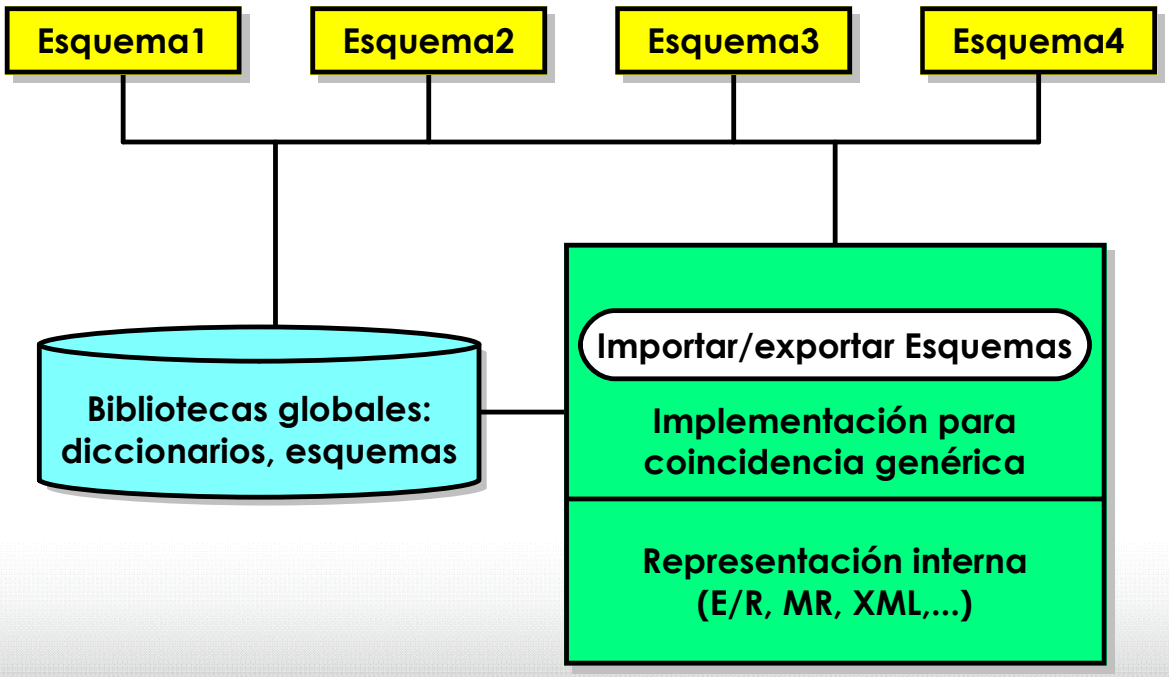
…Integración

Se analizan los distintos esquemas y se deciden las políticas generales de integración

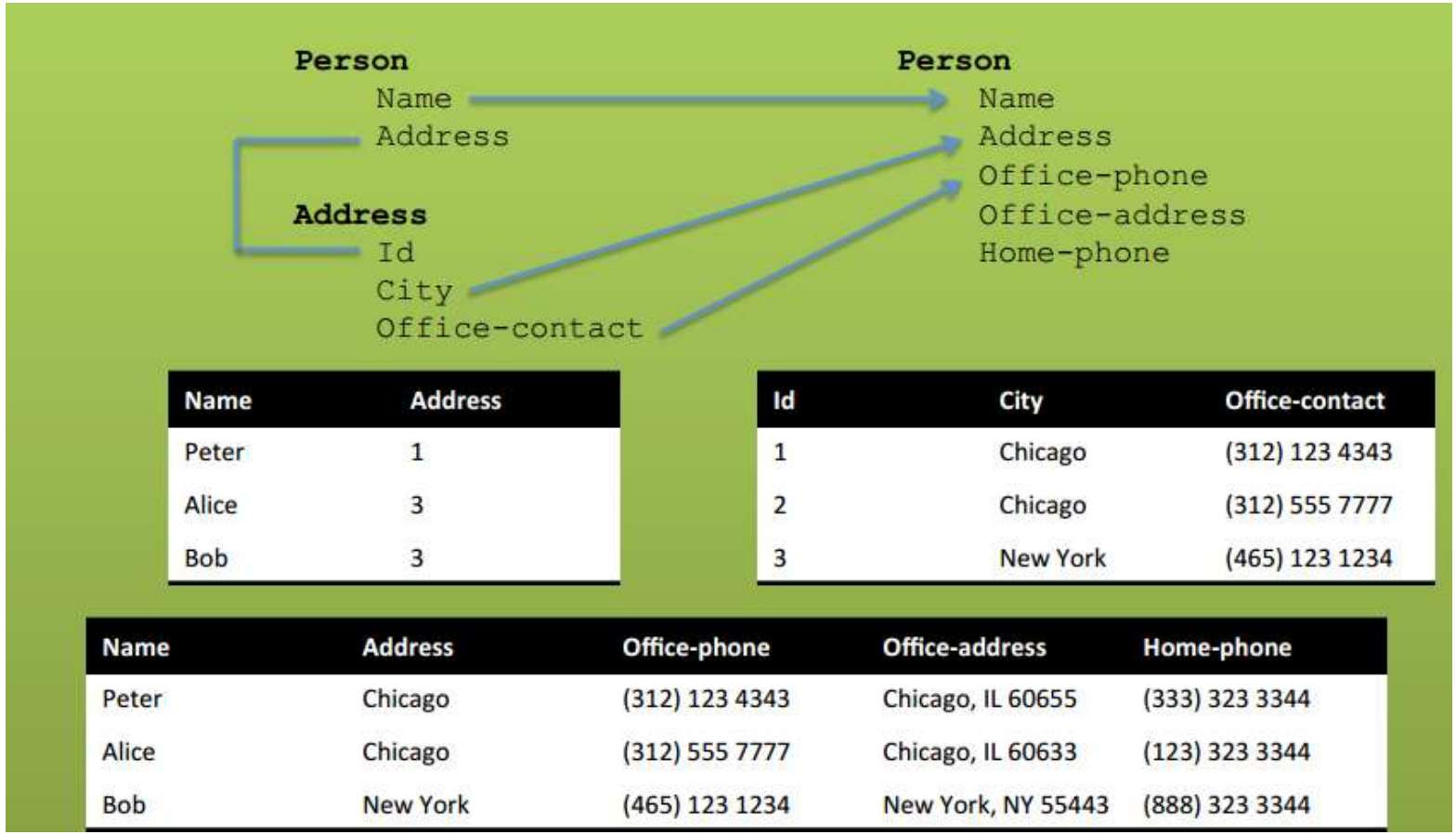
- **Qué esquemas serán integrados**
- **Orden de integración/proceso de integración**
- **Preferencias**

Coincidencia de esquemas

- Su objetivo es **tomar dos (o más) esquemas** como entrada (fuente) y generar un **mapeo (destino)** entre los elementos de los esquemas que corresponden semánticamente.
- Es un proceso **típicamente manual** (se puede apoyar con interfaces gráficas): *tedioso consume tiempo, propenso a errores, caro*.
- Determinar de qué forma los elementos están relacionados: atributos que representan la misma información, atributos que pueden traducirse (salario mensual a salario anual).
- **Arquitectura para coincidencia genérica:**



...Coincidencia de esquemas



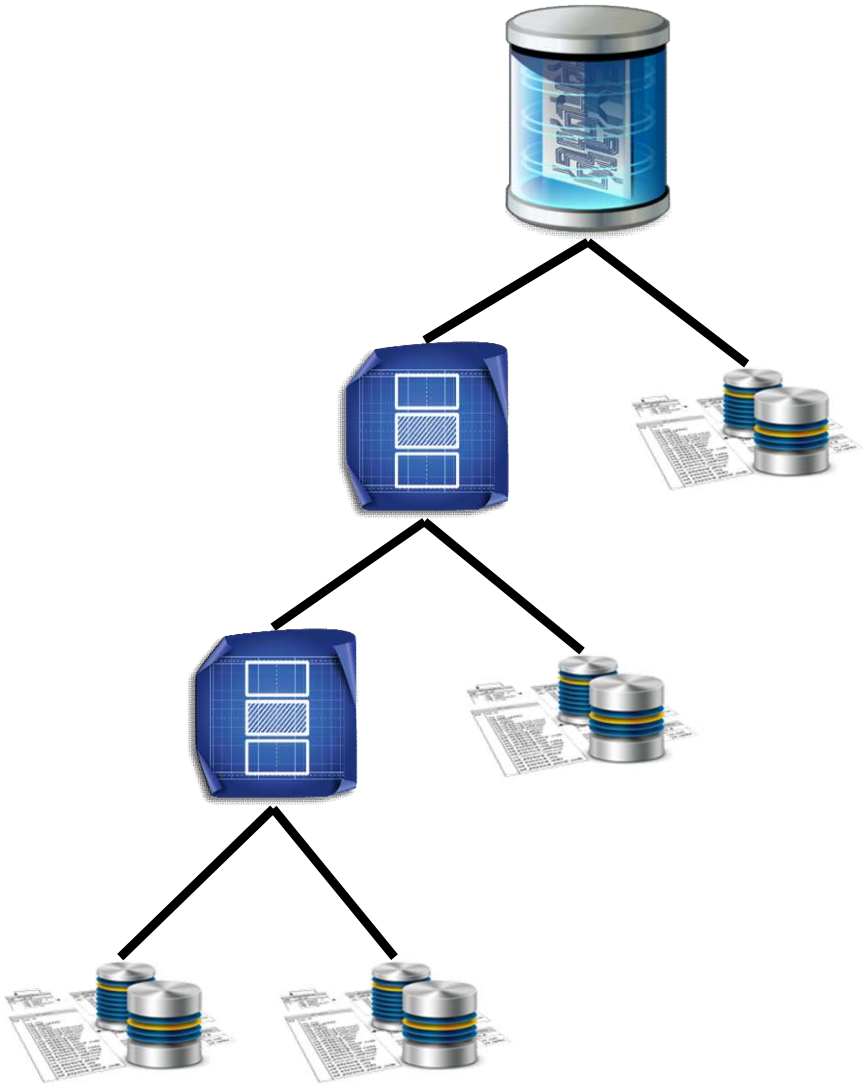
Name	Address
Peter	1
Alice	3
Bob	3

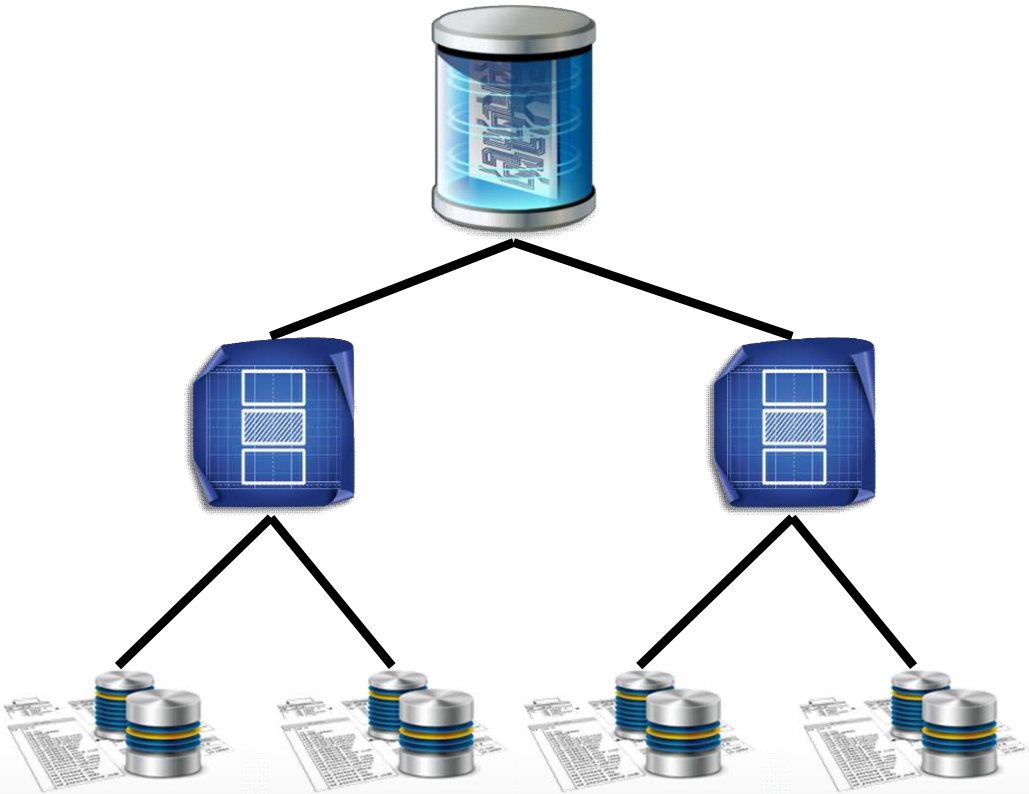
Id	City	Office-contact
1	Chicago	(312) 123 4343
2	Chicago	(312) 555 7777
3	New York	(465) 123 1234

Name	Address	Office-phone	Office-address	Home-phone
Peter	Chicago	(312) 123 4343	Chicago, IL 60655	(333) 323 3344
Alice	Chicago	(312) 555 7777	Chicago, IL 60633	(123) 323 3344
Bob	New York	(465) 123 1234	New York, NY 55443	(888) 323 3344

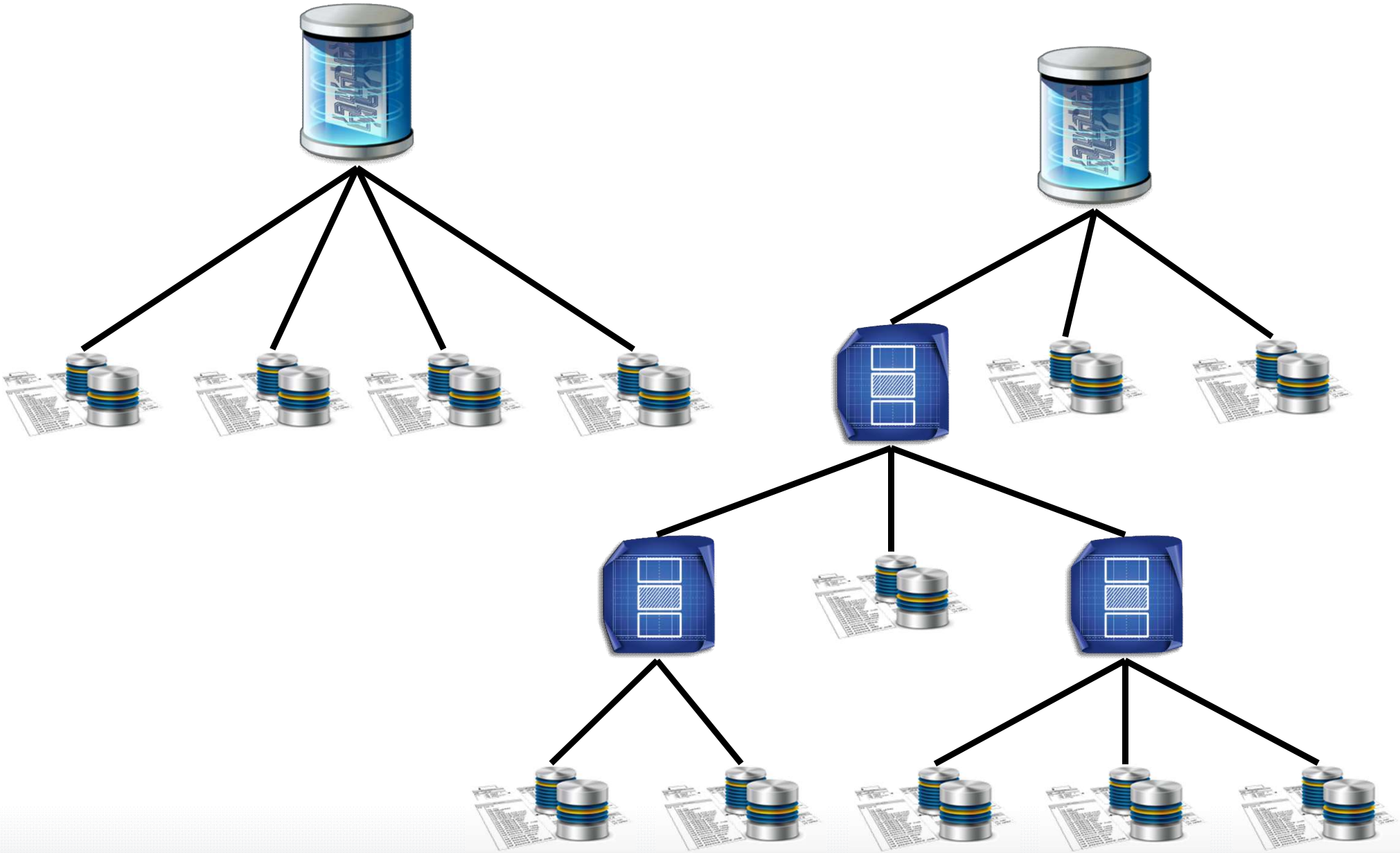
	Name	Address	Office-phone	Office-address	Home-phone
Name	**1**	0	0	0	0
Address	0	**1**	0	**0.4**	0
Id	0	0	0	0	0
City	0	0	0	0	0
Office-contact	0	0	**0.5**	**0.5**	0

…Coincidencia de esquemas





...Coincidencia de esquemas



Conformación de esquemas

- Su objetivo es determinar las correlaciones entre conceptos de diferentes esquemas y detectar posibles conflictos.

ESQUEMA

Tabla/Tabla

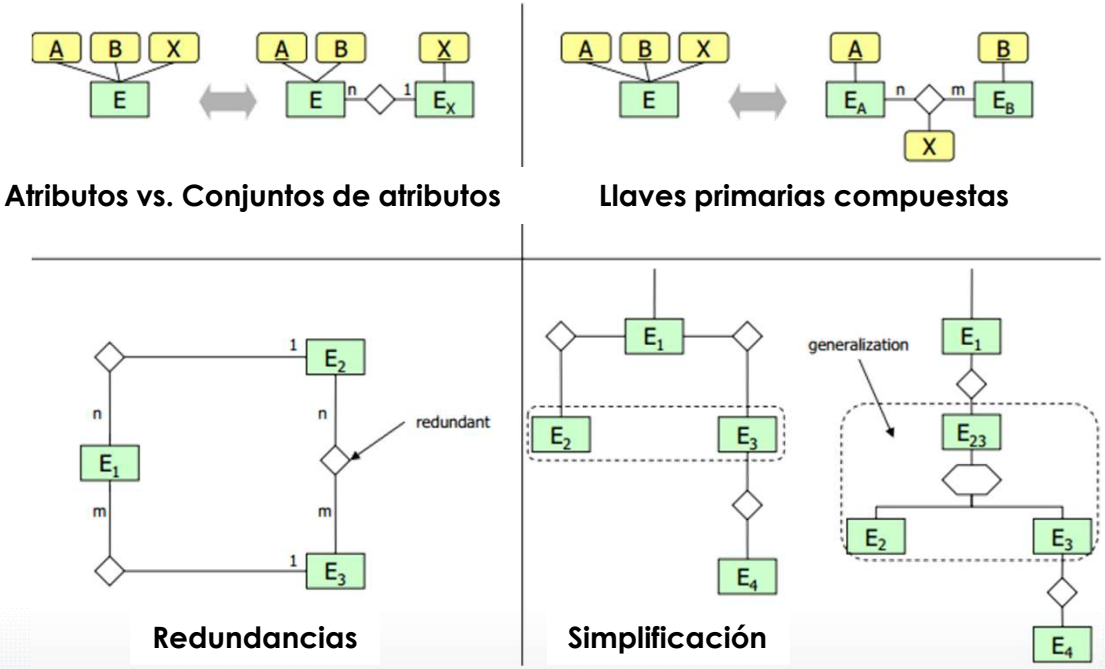
- **Nombrado**
  - ✓ Sinónimos
  - ✓ Homónimos
  - ✓ Parónimos
- **Estructura**
  - ✓ Valores perdidos
  - ✓ Atributos implicados
- **Restricciones de Integridad**

Atributo/Atributo

- **Nombrado**
  - ✓ Sinónimos
  - ✓ Homónimos
  - ✓ Parónimos
- **Valores por omisión**
- **Restricciones de Integridad**
  - ✓ Tipos de datos
  - ✓ Rango de valores

Comparación de esquemas

- El objetivo es **conformar y alinear** los esquemas para hacerlos **compatibles** para la integración.
- Solución de conflictos:
  - Basado en el contexto de la aplicación
  - No puede ser completamente automatizado
  - Se requiere la intervención del equipo, apoyados por herramientas gráficas.



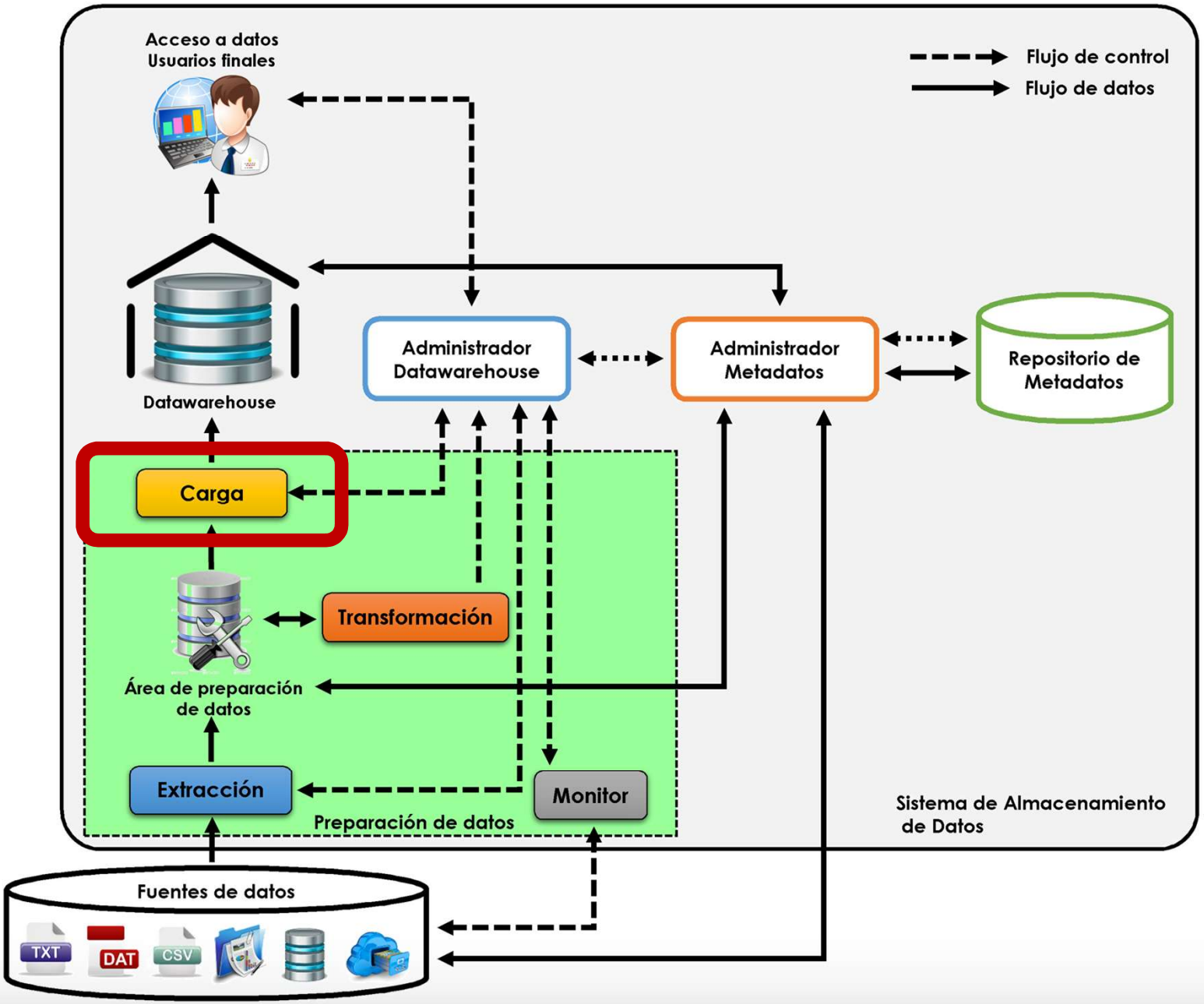
Transformaciones comunes

- **Selección de datos del Origen**, representa la consulta o primer filtrado/ordenación de los datos origen
- **Normalización.**
- **Cálculo de Expresiones/Nuevos Campos,** realiza cálculos a nivel de campo.
- **Filtro,** funciona como un filtro condicional de los registros procesados.
- **Agregación**, realiza cálculos agregados (totales o incrementales).
- **Rango,** limita los registros a los primeros o últimos de un rango.
- **Estrategia de Actualización,** para marcar cada registro como inserción, actualización, borrado, o registro rechazado.
- **Lookup,** busca valores complementarios y los pasa a otros objetos.
- **Procedimientos Externos/Almacenados,** llama a programas desarrollados en otros lenguajes o en la base de datos.
- **Generador de secuencia**, genera nuevos identificadores únicos.

…Transformaciones comunes

- **Estandarización de códigos y formatos de representación**
  - *Pasar información **EBCDIC** a **ASCII** o **UNICODE***
  - *Convertir números cardinales a ordinales*
  - *Separar fecha y hora*
  - *Poner textos descriptivos*
  - *Unificar códigos*
  - *Unificar estándares (tiempo, medida, moneda, etc.)*
- **Corrección (si no se pudo hacer en el origen)**
  - *Errores tipográficos*
  - *Datos que no tienen sentido*
  - *Resolver conflictos de dominio*
  - *Aclarar datos ambiguos*
  - *Asignar valores a datos nulos*
- **Resolución de llaves foráneas** *(datos procedentes de varias fuentes)*
- **Eliminación de registros duplicados**
- **Utilización de conversiones o combinaciones para generar nuevos campos**

Carga


Acceso a datos
Usuarios finales
Flujo de control
Flujo de datos
Datawarehouse
Administrador
Datawarehouse
Administrador
Metadatos
Repositorio de
Metadatos
Carga
Transformación
Área de preparación
de datos
Extracción
Monitor
Preparación de datos
Sistema de Almacenamiento
de Datos
Fuentes de datos
TXT
DAT
CSV

Carga

Objetivo:

Cargar de forma rápida los datos en el DWH. La carga de los cambios es mucho más rápida que la carga total de los datos.

- **Actualizaciones basadas en SQL son lentas:** *gran sobrecarga (optimización, bloqueo, etc.) por cada llamada SQL.*
- **Carga masiva utilizando SMBD específicos es mucho más rápida:** *algunas herramientas de carga también puede realizar actualizaciones.*
- **Índices en las tablas ralentizan cargas de gran volumen:**
  - *Borrar los índices y reconstruirlos después de la carga*
  - *Pueden ser hecho por partición*
- **Paralelización:** *Las dimensiones, tablas de hechos y particiones se pueden cargar al mismo tiempo.*

Planificación

- **Orden de carga**, cargar primero las tablas independientes.
- Determinar la **ventana necesaria** de carga: utilizar las horas de inicio y final para determinar el tiempo necesario para las cargas
- Ejecutar **cargas en paralelo:**
  - Ejecución concurrente.
  - Uso de hilos, desarrollos multiproceso, paralelización de base de datos
  - No sobrecargar los sistemas origen o destino.
- **Cargar en paralelo un mismo destino**.
  - Datos de sistemas independientes que van al mismo destino
- **Cargar múltiples destinos en paralelo.**
  - Datos del mismo origen que vayan a diferentes destinos, ahorrar accesos de lectura.

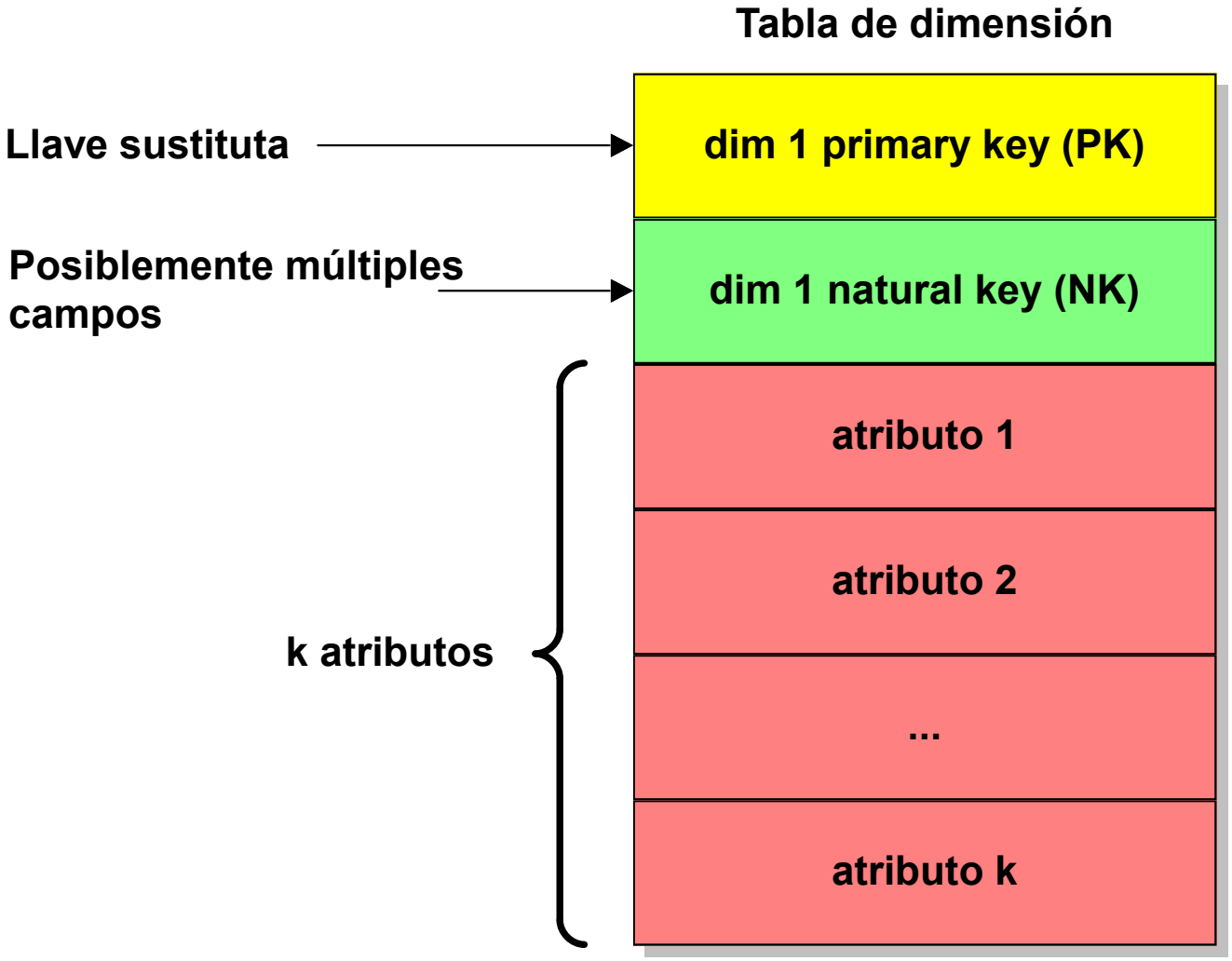
Construcción de dimensiones

- **Tabla de dimensión**
  - Asignación de llaves, llaves de producción.
  - Combinación de fuentes de datos: encontrar la llave común
- **Manejo de cambios de dimensión**
  - Dimensiones lentamente cambiantes
  - Encontrar las nuevas llaves del DWH para una llave de producción indicada.
  - Tabla para mapear las llaves de producción a las llaves del DWH que debe actualizarse.
- **Carga de dimensiones**
  - Pequeñas dimensiones: **reemplazar**
  - Grandes dimensiones: **carga sólo cambios**

Estructura básica de una dimensión

- **Llave primaria (PK)**
  - Sin significado, entero único (**surrogate key → llaves sustitutas**)
  - Permite vincular a la tabla de hechos (llave foránea)
- **Llave natural (NK)**
  - Mismo significado que el extraído del sistemas de origen
  - Relación **1:1** con la **PK** de una dimensión estática
  - Relación **1:N** con la **PK** para dimensiones que cambian lentamente, permite llevar un registro histórico de todos los cambios en la dimensión.
- **Atributos descriptivos**
  - Principalmente texto, es posible encontrar números, pero no números que representen cantidades
  - ≈100 atributos es un número normal
  - Estáticos o de lento cambio.
- **Tabla denormalizadas**

…Estructura básica de una dimensión



Llaves sustitutas (surrogate key)

Se pueden generar

- **A través de disparadores en el SMBD**
  - Leer la última llave artificial creada y generar el siguiente valor.
  - Desventajas: cuellos de botella en el rendimiento
- **A través del proceso de ETL, una herramienta ETL o una aplicación de terceros generan los números únicos**
  - Se crea la llave sustituta a partir de un contador; una por dimensión
  - Mantener la consistencia entre las llaves sustitutas de desarrollo, prueba y producción.
- **Uso de Llaves inteligentes**
  - Concatenar la clave natural de la dimensión en la fuente con la marca de tiempo del registro en la fuente o el **DWH**.
  - Tentador, pero incorrecto.

Llaves inteligentes

- **Por definición**
  - Las llaves sustitutas no tienen ningún significado.
  - ¿Se actualiza la llave inteligente concatenada si cambia la llave natural?
- **Rendimiento**
  - Las llaves naturales pueden ser caracteres no enteros.
  - Agregar una marca de tiempo hace que la llave sea muy grande
    - La dimensión es más grande
    - Las tablas de hechos que contienen la llave foránea son más grandes
    - Unir la tabla de hechos con las dimensiones es ineficiente.
- **Fuentes heterogéneas**
  - Llaves inteligentes "trabajan" en entornos homogéneos. En fuentes que son heterogéneas, cada una tiene su propia definición.
  - ¿Cómo cambia la definición de la llave inteligente cuando se añade otra fuente? No escala muy bien.
- **Ventaja:** simplifica el proceso **ETL**.

Carga de dimensiones

- Cuando un **DWH** recibe la notificación de que un registro en la dimensión ha cambiado, es posible tener **tres escenarios**:

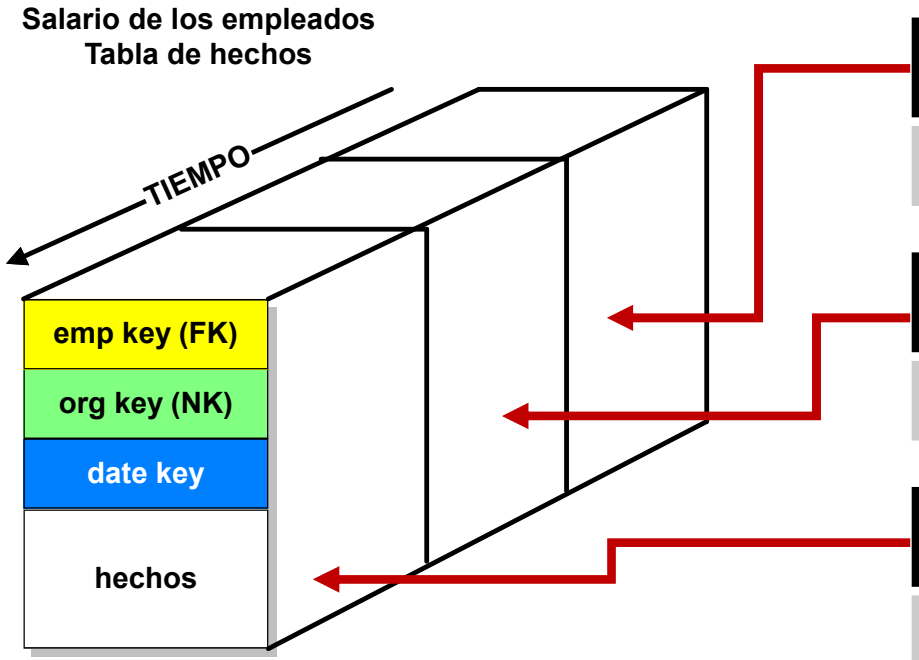
Dimensión TIPO 1

Llave primaria	Llave natural	Prod_name	Categoría	Empacado
45789	IV22	Coca Cola 125 ml	Bebidas sin alcohol	Vidrio

Se transforma en:

Llave primaria	Llave natural	Prod_name	Categoría	Empacado
45789	IV22	Coca Cola 125 ml	Bebidas sin alcohol	Plástico

Dimensión TIPO 2



Emp key (PK)	Llave natural	nombre	trabajo
2811	GE2080	Carlos Lara	Entrenamiento

Emp key (PK)	Llave natural	nombre	trabajo
2811	GE2080	Carlos Lara	Jefe depto.

Emp key (PK)	Llave natural	nombre	trabajo
2811	GE2080	Carlos Lara	Administrador

Dimensión TIPO 3

Prod key	Prod id	Prod_name	Talla	Categoría	Color	…
1123456	A 457U	Pantalón de mezclilla	38	Caballeros	Azul	

Se agrega un nuevo campo

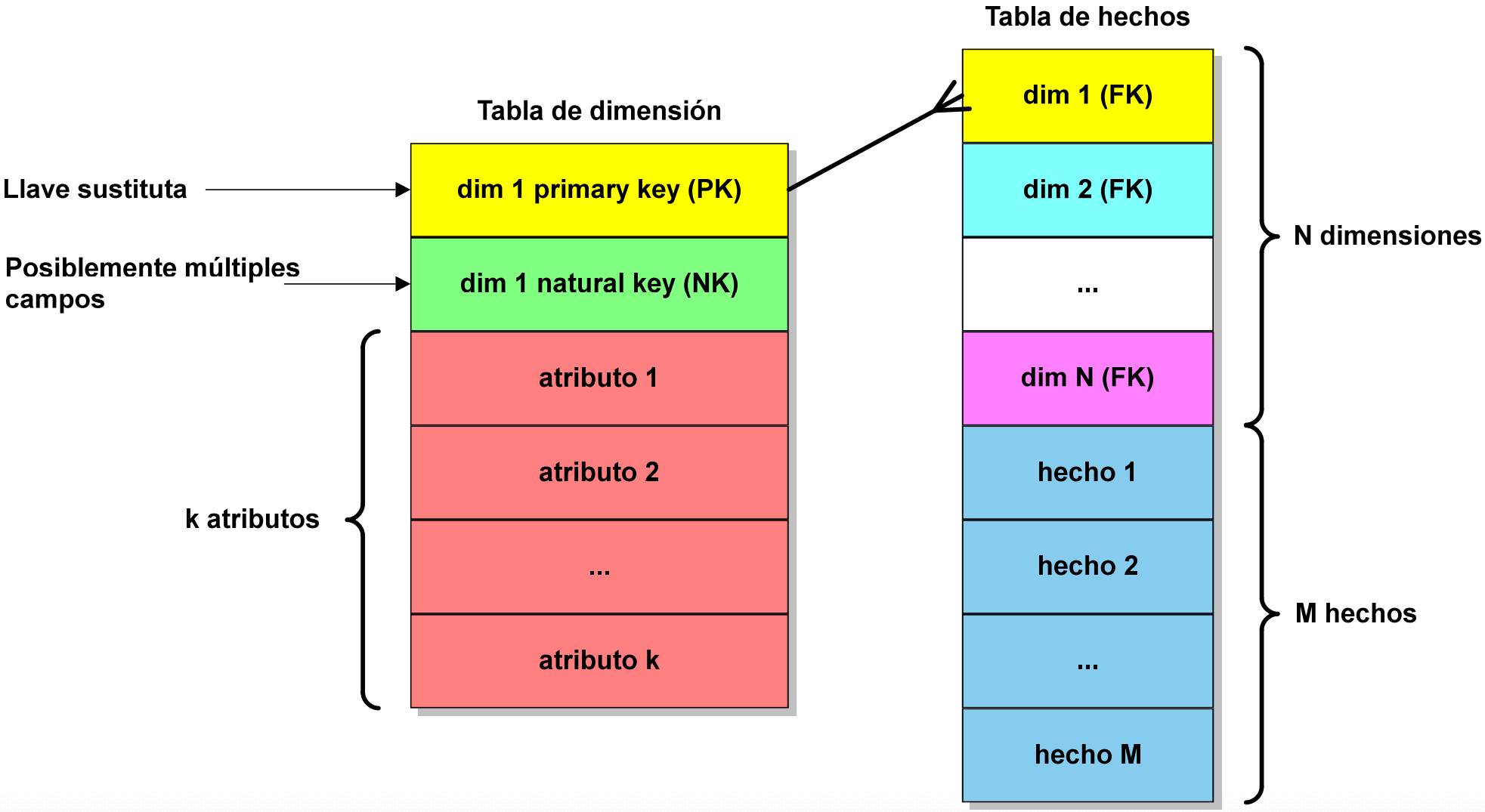
Prod key	Prod id	Prod_name	Talla	Categoría	Categoría anterior	Color	…
1123456	A 457U	Pantalón de mezclilla	38	Ropa informal	Caballeros	Azul	

Se sobrescribe con el nuevo valor

Construyendo tablas de hecho

- Las tablas de hechos tienen las **mediciones de una empresa**.
- Relación entre las tablas de hechos y medidas es extremadamente simple.
  - Si existe una medida, se puede modelar como una fila de tabla de hechos.
  - Si existe una fila de tabla hecho, es una medida.
- Cuando se construye una **tabla de hechos**, la etapa final del proceso ETL es la conversión de las llaves naturales en llaves sustitutas.
- El proceso ETL mantiene una tabla de búsqueda para las llaves sustitutas de cada tabla de dimensión.
  - Esta tabla se actualiza cada vez que se crea una nueva entidad de dimensión y siempre que un cambio de tipo 2 se produce. Se mantiene en memoria.

…Estructura de una tabla de hechos




¡Gracias!