ECCI
Escuela de Ciencias de la Computación e Informática

ETL

Dr. Luis Gustavo Esquivel Quirós
Escuela de Ciencias de la Computación e Informática
Universidad de Costa Rica

ETL

- ETL es un proceso en los almacenes de datos (Data Warehouse) y significa Extraer (Extract), Transformar(Transform) y Cargar(Load).
- Es un proceso en el que se extraen los datos de varias fuentes (que no están optimizadas para análisis), se transforman en un área de preparación (implica conversiones, limpieza, etc.) y finalmente se cargan en el sistema de almacenamiento de datos.

Extraction, Transformation, and Loading (ETL)

- **Extracción de datos**
  - Obtener datos de múltiples fuentes , heterogéneas y externas
- **Limpieza de datos**
  - Detectar errores en los datos y rectificarlos cuando sea posible
- **Transformación de datos**
  - Convertir datos de formato heredado o específico a formato del almacén
- **Carga**
  - Ordenar, resumir, consolidar, calcular vistas, verificar la integridad y crear índices y particiones
- **Actualizar**
  - Propagar las actualizaciones de las fuentes de datos al almacén

ETL

Entender requerimientos

Identificar los datos y las herramientas requeridas

Construir

Terminamos!!

ETL

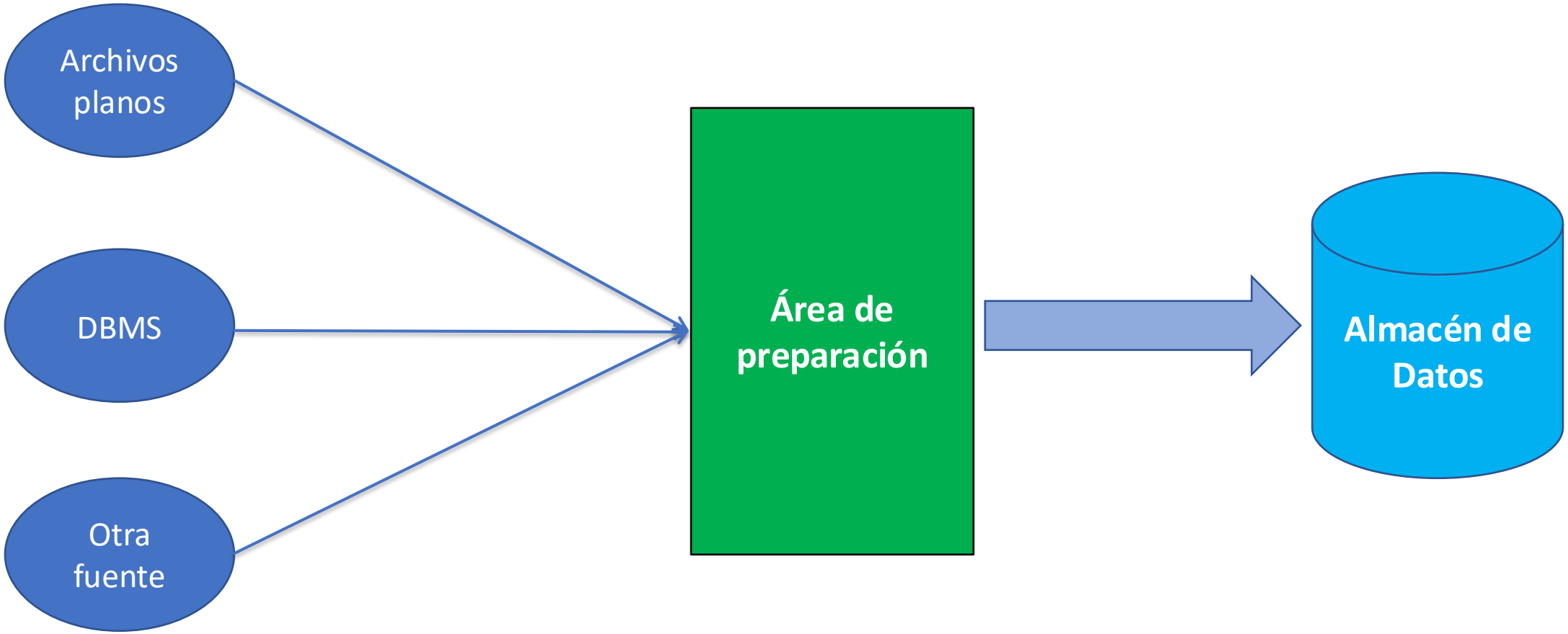
Expectativa

Realidad

Proceso ETL

- La especificación de un ETL se puede desarrollar de forma manual utilizando programas específicos o utilizando herramientas especializadas.
- Un proceso ETL bien diseñado extrae datos de las fuentes de datos y hace cumplir estándares de calidad.
- Su finalidad es que los datos puedan ser utilizados para las aplicaciones y así los usuarios finales puedan tomar decisiones estratégicas.

Proceso ETL


Archivos planos
DBMS
Otra fuente
Área de preparación
Almacén de Datos

Extracción

- El primer paso del proceso ETL es la extracción.
- En este paso, se extraen datos de varios sistemas de origen, estos datos pueden estar en varios formatos, como bases de datos relacionales, NoSQL, XML o archivos planos.
- Es importante primero extraer los datos y almacenarlos en el área de preparación. No se recomienda hacerlo directamente en el almacén de datos porque los datos extraídos por lo general están en varios formatos y también pueden corromperse.

Estrategias de extracción

- Una fase:

  Los datos se cargan directamente en la tabla de destino realizando pruebas de aseguramiento de la calidad y realizando modificaciones sobre la marcha.

- Dos fases:
  - Los datos se cargan primero en un almacenamiento temporal (área de preparación) y ahí se aplican las pruebas de aseguramiento de la calidad y se realizan las modificaciones sobre los datos.
  - Se finaliza copiando los datos desde el área de preparación al almacenamiento destino (almacén de datos).

Extracción de datos

- Se recomienda documentar:
  - Fuentes de datos (tipo, ubicación, acceso, etc.).
  - Dominio (actividad relacionada).
  - Cuales herramientas se utilizarán en el procesamiento.
  - Usuarios de los datos.
  - Importancia de los datos para los usuarios del Almacén de Datos.
  - Dónde se almacenarán (detalles de hardware y software).
  - Volumen de datos.
  - Esquemas o descripción detallada de los datos.
  - Número de transacciones y/o usuarios por día.

Transformación

- En este paso, se aplica a los datos extraídos un conjunto de reglas o funciones para convertirlos en un formato estándar único. Puede implicar los siguientes procesos / tareas:
  1. Filtrado: cargar solo ciertos atributos en el almacén de datos.
  2. Limpieza: completar los valores NULL con algunos valores predeterminados, mapear, entre otros.
  3. Unión: se agregan o unen múltiples atributos en uno.
  4. División: separar un solo atributo en múltiples atributos.
  5. Clasificación: clasificar tuplas en función de algún atributo (generalmente, atributo clave).

Carga de datos

- En este paso, los datos transformados finalmente se cargan en el almacén de datos.
- A veces, los datos se actualizan cargándolos en el almacén de datos con mucha frecuencia y, en otras ocasiones la actualización se realiza después de intervalos largos pero regulares.
- La velocidad y el período de carga dependen únicamente de los requisitos y varían de un sistema a otro.
- El proceso ETL también puede utilizar el concepto de canalización, es decir, tan pronto como se extraen algunos datos, se pueden transformar y durante ese período se pueden extraer algunos datos nuevos. De esa forma mientras los datos transformados se cargan en el almacén de datos, los datos ya extraídos se pueden ir transformando.

Diseño de un proceso ETL

- El diseño de un proceso ETL se compone generalmente de seis tareas:

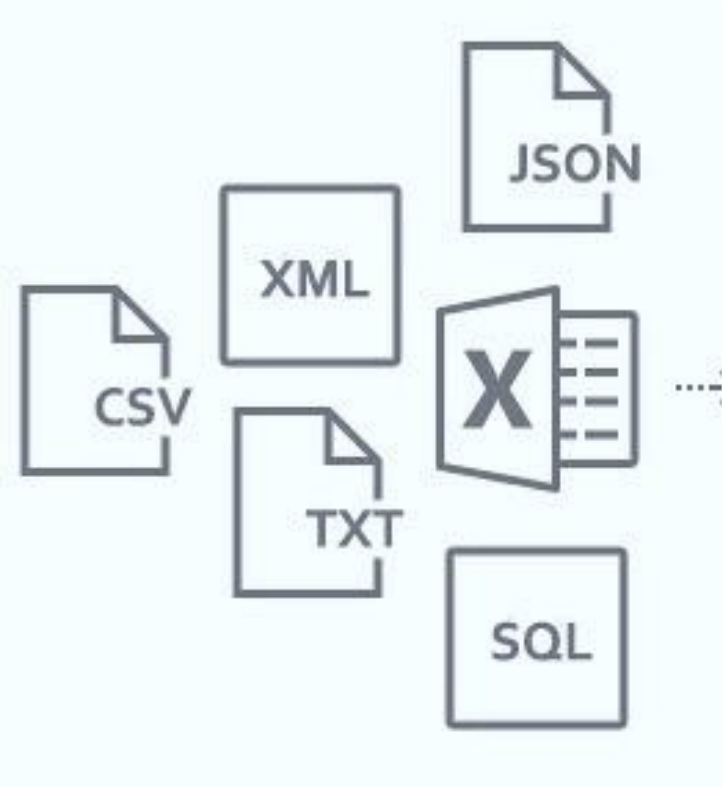
1. Seleccionar los datos para la extracción.
2. Transformar las fuentes.
3. Unir las fuentes.
4. Seleccionar el destino para la carga.
5. Unir los atributos de las fuentes de datos con los atributos del destino.
6. Cargar los datos.

Puesta en escena de datos

- A menudo se usa como un paso intermedio entre la extracción de datos y los pasos posteriores.
- Acumula datos de fuentes asincrónicas utilizando interfaces nativas, archivos planos, sesiones FTP u otros procesos.
- En un tiempo límite predefinido, los datos en el archivo de preparación se transforman y se cargan en el almacén.
- Por lo general, no hay acceso de usuario final al archivo de preparación.
- Se puede usar un almacén de datos operativo para la preparación de datos.

Seleccionar los datos para la extracción

- Se definen los datos de las fuentes (generalmente provienen de diversas fuentes heterogéneas).

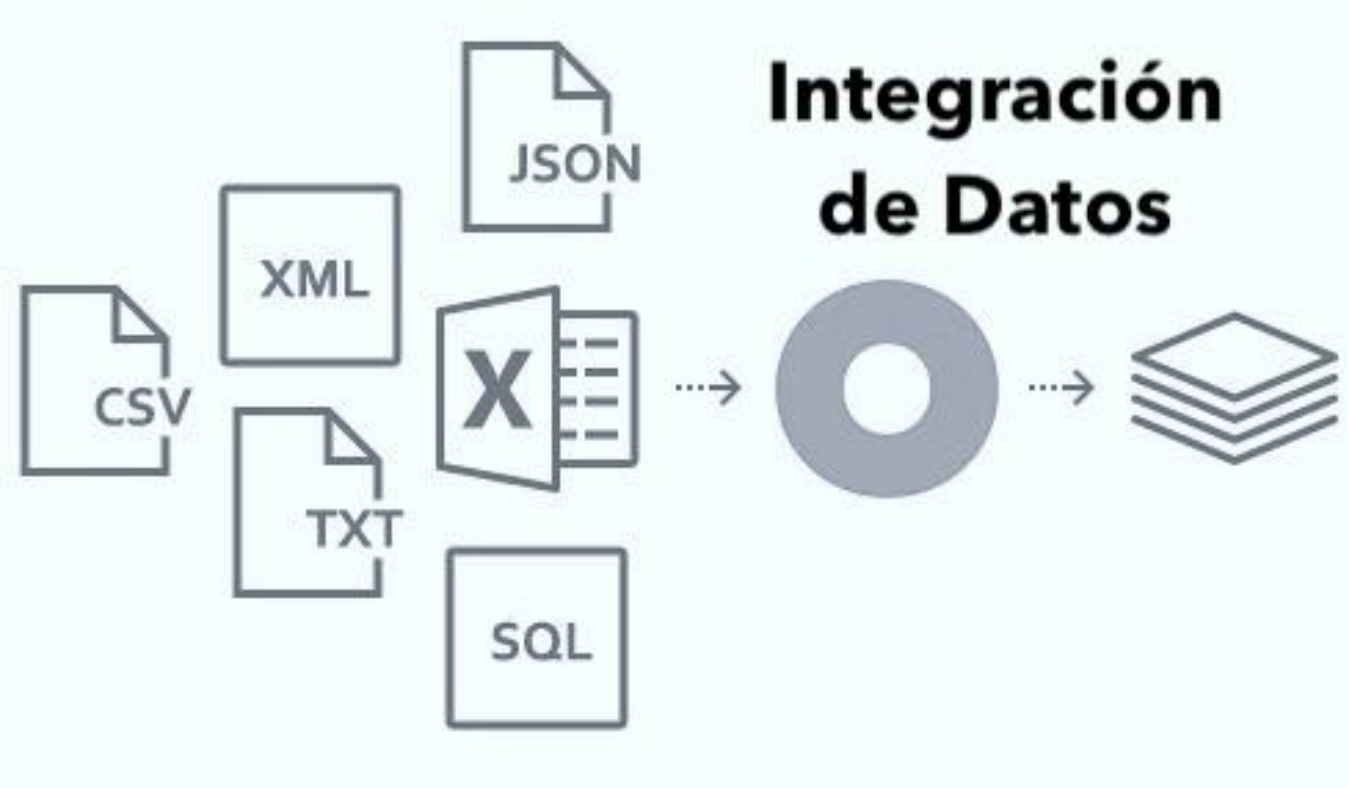


Transformar las fuentes

- Una vez que los datos se hayan extraído de las fuentes de datos pueden ser transformados o esos nuevos datos pueden ser derivados.
- Algunas de las tareas más comunes de este paso son:
  - Filtrado de datos
  - Conversión de códigos
  - Cálculos de valores derivados
  - Transformación entre diversos formatos de datos
  - Generación automática de números secuenciales (llaves derivadas)

Unir las fuentes

- Las diversas fuentes pueden unirse para ser cargadas al almacén como una sola fuente



Razones para los datos "sucios"

- Valores ficticios
- Ausencia de datos
- Campos Multipropósito
- Datos crípticos
- Datos contradictorios
- Uso inapropiado de líneas de dirección
- Violación de las reglas comerciales
- Claves primarias reutilizadas
- Identificadores no únicos
- Problemas de integración de datos

Limpieza de datos

- Los sistemas de origen contienen "datos sucios" que deben limpiarse
- El software ETL contiene capacidades rudimentarias de limpieza de datos
- A menudo se utiliza un software de limpieza de datos especializado. Es importante para realizar correcciones de nombres y direcciones y funciones domésticas (propias del contexto o aplicación)

Pasos en la limpieza de datos

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating

Parsing

- Analiza, localiza e identifica elementos de datos individuales en los archivos de origen y luego aísla estos elementos de datos en los archivos de destino
- Ejemplos:
  - Analizar el nombre, el segundo nombre y el apellido
  - El número de calle y nombre de la calle
  - La ciudad y el estado.

Correcting

- Corrige los componentes de datos individuales analizados utilizando algoritmos de datos sofisticados y fuentes de datos secundarias
- Ejemplo:
  - Reemplazar una dirección personalizada y agregar un código postal

Standardizing

- La estandarización aplica rutinas de conversión para transformar los datos en su formato preferido (y consistente) utilizando reglas comerciales estándar y personalizadas
- Ejemplos:
  - Agregar un nombre previo
  - Reemplazar un apodo y usar un nombre de una calle preferida

Matching

- Buscar y hacer coincidir registros dentro y entre los datos analizados, corregidos y estandarizados en función de reglas comerciales predefinidas para eliminar duplicaciones
- Ejemplos:
  - Identificación de nombres similares
  - Identificación de direcciones similares

Consolidating

- Analizar e identificar relaciones entre registros coincidentes y consolidarlos/fusionarlos en **una** representación.

Limpieza de datos: Faltan datos

- Faltan datos: muchas tuplas no tienen valor registrado para algunos atributos debido a:
  - El valor de datos N/A cuando se recopilan
  - Mal funcionamiento del equipo
  - Inconsistencia con otros datos registrados y, por lo tanto, fue eliminado
  - Datos no ingresados por malentendidos
  - Ciertos datos pueden no ser considerados importantes en el momento de entrada
  - No registrar el historial o los cambios de datos
- Es posible que sea necesario inferir los datos faltantes

Limpieza de datos: Faltan datos

- Manejo de datos faltantes:
  - Ignorar la tupla: no se puede hacer análisis
  - Rellene el valor que falta manualmente: ¿tedioso + inviable?
  - Rellénelo automáticamente con
    - Una constante global, por ejemplo, "desconocido": aproximación de resultados
    - La media del atributo: mejor solución que antes
    - La media del atributo para todas las muestras pertenecientes a la misma clase: más inteligente
    - El valor más probable según algunas inferencias basadas en fórmulas bayesianas, árbol de decisión u otras técnicas: la mejor solución

Limpieza de datos: datos ruidosos

- Datos ruidosos: error aleatorio o varianza en una variable medida
  - Los valores de atributo incorrectos pueden deberse a
    - Instrumentos de recolección de datos defectuosos
    - Problemas de entrada de datos
    - Problemas de transmisión de datos

Limpieza de datos: datos ruidosos

- Manejo de datos ruidosos
  - Agrupación
    - Detectar y eliminar valores atípicos
  - Inspección combinada por computadora y humana
    - Detectar valores sospechosos y verificarlos por humanos (por ejemplo, tratar con posibles valores atípicos)
    - Usar diccionarios y repositorios de datos disponibles para corregir palabras mal escritas
    - Usar herramientas existentes con búsqueda difusa
  - Regresión
    - Suavice ajustando los datos en funciones de regresión

Transformación de datos

- Transforma los datos de acuerdo con las reglas y estándares de negocio que se han establecido
- Ejemplos:
  - Cambio de formato
  - Deduplicación
  - División de campos
  - Reemplazo de códigos
  - Valores derivados y agregados

Transformación de datos

- Los datos integrados y reducidos pueden requerir otras transformaciones
  - División en campos separados, p. ej., nombre separado en el primer nombre, segundo nombre y apellido o dirección en componentes correspondientes
  - Cambiar la fecha de forma abreviada a diferentes componentes, por ejemplo, nombre de semana, indicador de fin de semana, etc.

Transformación de datos

- Representar la codificación en una forma más descriptiva, por ejemplo, en un banco el estado de la cuenta se puede representar como 1, 2, ... Requiriendo cambiar 1 a activa, 2 cerrada, ...
- Crear un nivel de jerarquía para los valores cardinales y ordinales, si no existen, por ejemplo, la clasificación de diferentes marcas de automóviles en mayores niveles de jerarquía como malo, bueno, excelente
- Estandarización de valores para establecer formatos de fechas, dinero, minúsculas mayúsculas, abreviaturas consistentes, ...

Seleccionar el destino para la carga

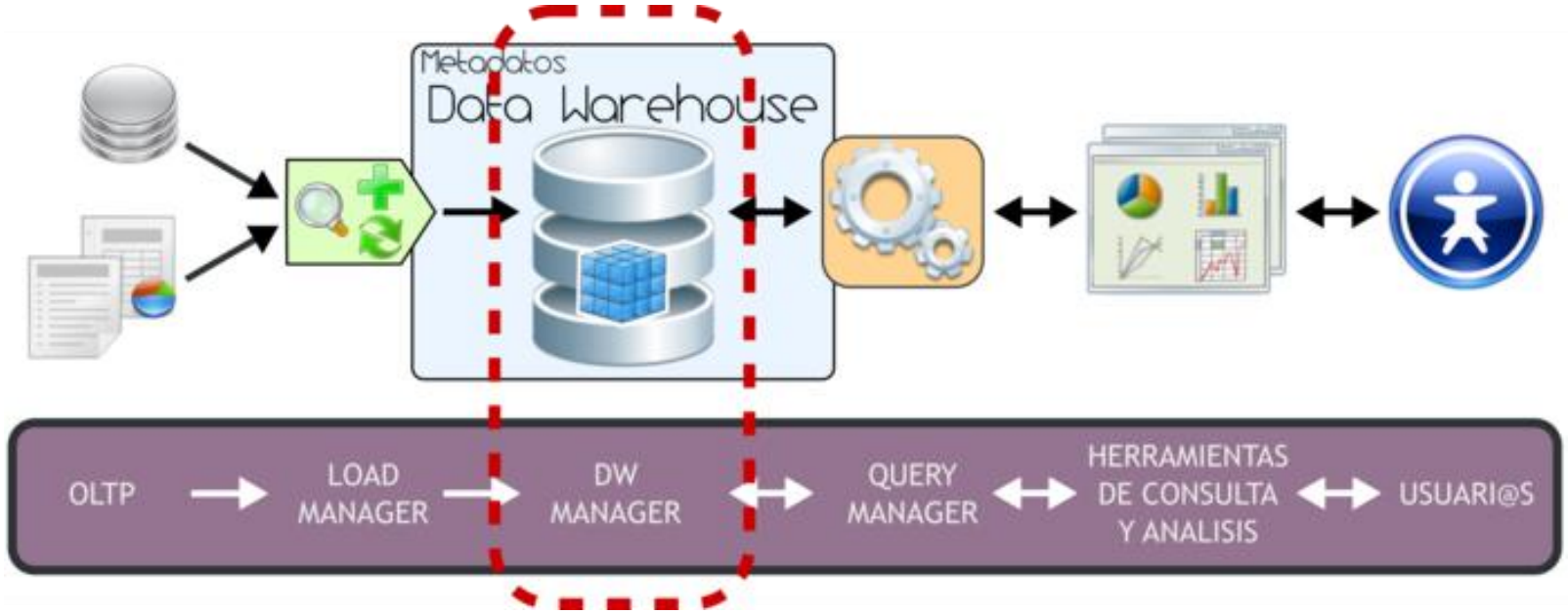
- El destino o los destinos son seleccionados para cargar los datos posteriormente.

Unir los atributos de las fuentes de datos con los atributos del destino

- Los atributos (campos) que se obtuvieron de las fuentes de datos pueden ser mapeados con los correspondientes destinos.

Cargar los datos

- El almacén poblado con los datos transformados



Carga de datos

- Los datos se mueven físicamente al almacén de datos
- La carga tiene lugar dentro de una “ventana de carga”
- La tendencia es hacia actualizaciones casi en tiempo real del almacén de datos, ya que el almacén se utiliza cada vez más para aplicaciones operativas

Carga de datos

- Proceso de carga: alimenta el almacén de datos con los
- datos transformados
- Por lo general, las utilidades de carga por lotes se utilizan para cargar
  - Debe manejar grandes volúmenes de datos
  - Debe permitir que el administrador controle el estado, cancele, suspenda y reanude una carga y reiniciar después de una falla sin pérdida de integridad de datos
  - La carga secuencial puede llevar mucho tiempo
    - Carga paralela

Carga de datos

- El proceso de carga puede incluir las siguientes actividades:
  - Desactivar el registro y las restricciones
  - Establecer un nivel de concurrencia bajo para evitar el uso de bloqueos
  - Eliminar índices (si ya existen)
  - Use instrucciones o utilidades de carga masiva, por ejemplo, bulk insert
  - Rellenar tablas de dimensiones asignándoles claves sustitutas
  - Rellenar tablas de hechos con los sustitutos y las medidas correspondientes

Metadatos

- “Datos sobre datos”
- Son necesarios tanto para el personal especializado como para los usuarios
- El personal especializado necesita conocer las fuentes y los objetivos de los datos; nombres de bases de datos, tablas y columnas; actualizar horarios; medidas de uso de datos; entre otros
- Los usuarios necesitan conocer las definiciones de entidad/atributo; informes/herramientas de consulta disponibles; información de distribución de informes; información de contacto de la mesa de ayuda, entre otros