



# Procesos de Extracción, Transformación y carga (ETL). Introducción al *Pentaho Data Integration* (PDI)

MSc. PA. Lisandra Díaz de la Paz

# Sumario

- Tecnologías de integración de datos
- Procesos ETL. Herramientas
- Introducción al *Pentaho Data Integration*
  - Arquitectura
  - Repositorio de datos y metadatos
  - Transformaciones
  - Trabajos

# Integración de Datos

## **Bases para la integración:**

- Datos heterogéneos y distribuidos.
- Islas de información inconsistentes.
- Sincronización de copia compleja y costosa.
- Datos inconsistentes y de baja calidad.
- Sin retroalimentación de la calidad del servicio.
- Soporte imposible de las transformaciones del negocio.

# Integración de Datos

Tomando en consideración las diferencias en cuanto a:

- calidad de datos.
- procedimientos de expertos en el manejo de los datos.
- formatos de datos.
- Lenguajes, entre otros.

La toma de decisiones sobre estas bases es prácticamente un problema imposible.

**La integración de datos se hace necesaria.**

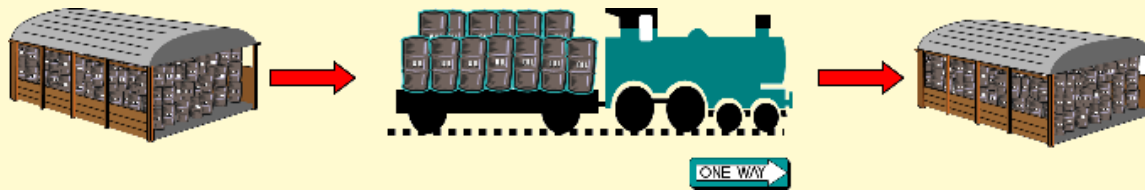
# Formas de integración de datos

- Data Warehousing basado en **ETL** (Extract, Transform, Load)
- **Otras formas de integración:**
  - Replicación de datos
  - **EAI** (Enterprise Application Integration)
  - **EII** (Enterprise Information Integration)



# Replicación de Datos

**Replicación:** Es la creación y mantenimiento de múltiples “*copias*” de la misma base de datos.



*un servidor de base de datos mantiene la copia master y servidores de base de datos adicionales mantienen las copias esclavas de la base de datos.*

- Basado en Tablas
- Viejo estilo de Integración (disaster recovery, site mirroring)
- Unidireccional y en algunos casos bi-direccional
- Modelo Cliente-Servidor
- Fragmentación (Vertical/Horizontal)

# Tecnologías de integración de datos

- **EAI** (aplicación - aplicación)
  - Orientado a Mensajes (queuing system, message switching).
  - Soporta Gestión de Procesos de Negocio (BPM).
  - Tiempo real, bidireccional.

# Enterprise Application Integration (EAI)

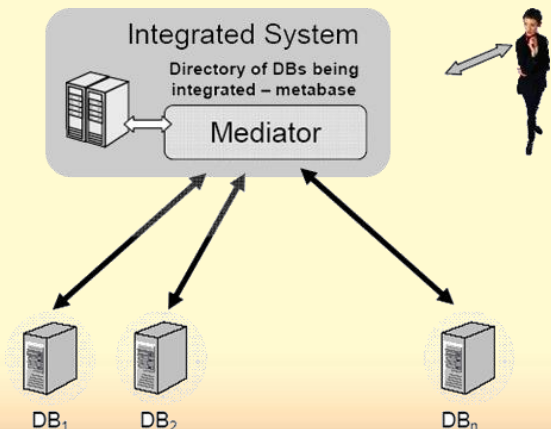




# Enterprise Information Integration (EII)

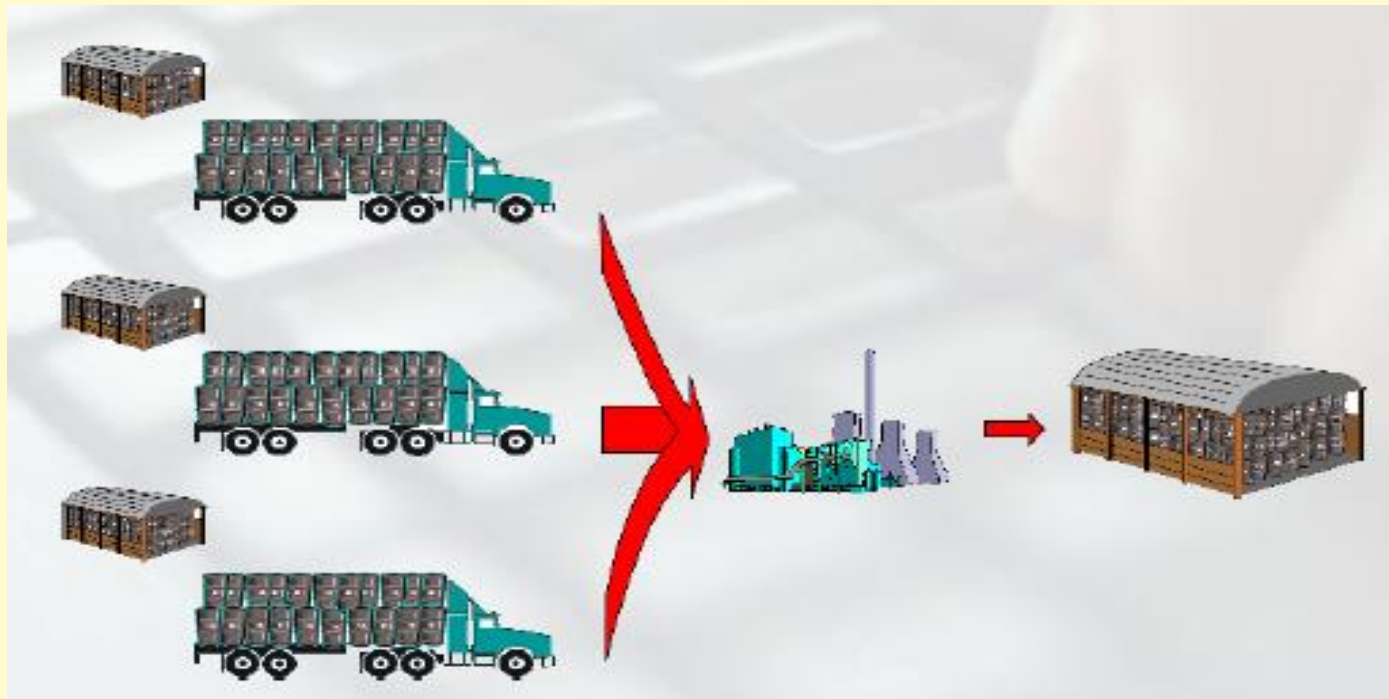
- Esta opción no integra bases de datos.
- Los “datos integrados” no son transferidos a una base de datos central, sino que continúan radicando en sus bases de datos orígenes.
- Se programa una Interfaz para el acceso a datos, que permite acceder a los datos requeridos.

**EII es un Sistema de Información distribuido y heterogéneo virtualmente integrado.**



# Extracción, Transformación, Limpieza y Carga de Datos

- **ETL:** Conjunto de transformaciones para la migración, consolidación y Data Warehousing.

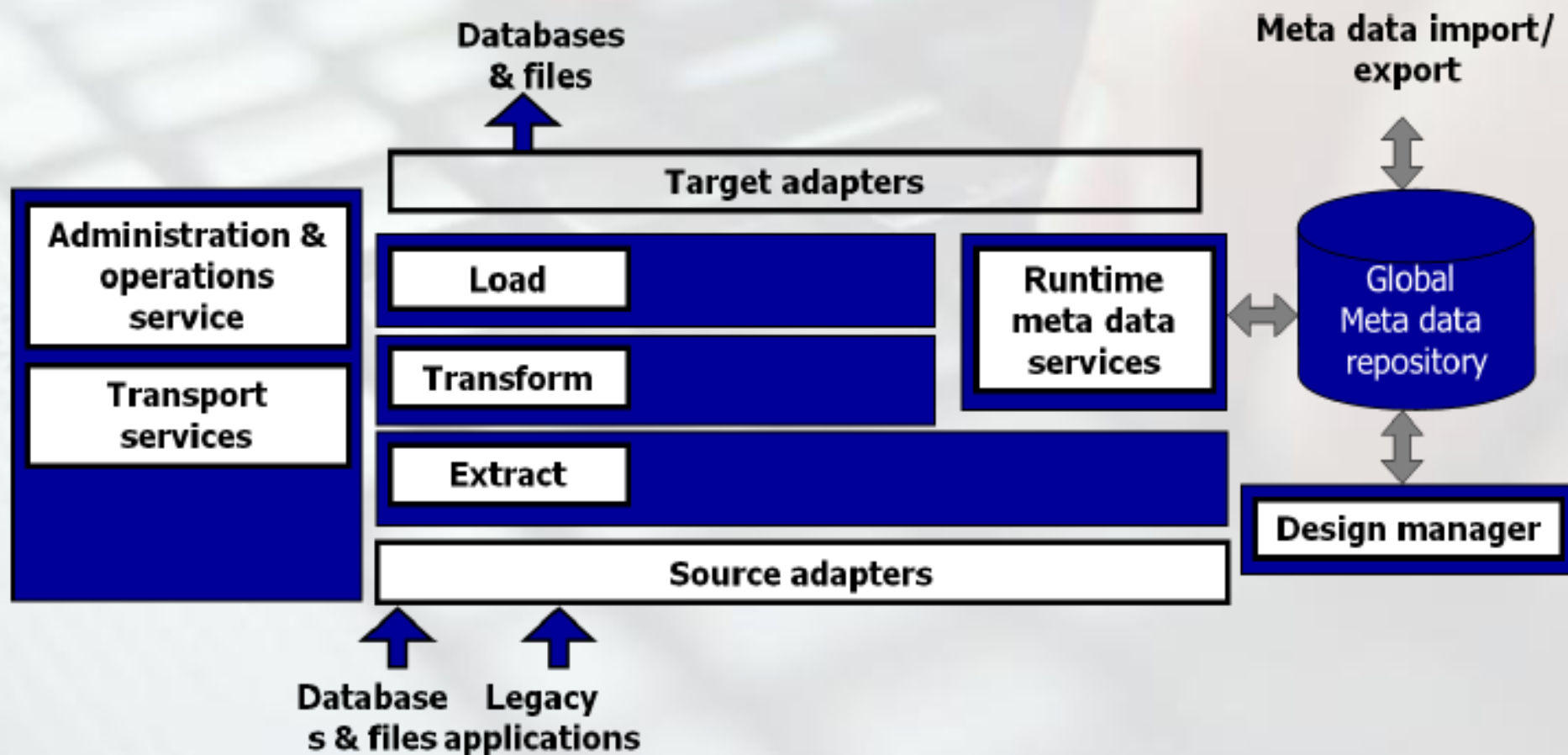


# Extracción, Transformación, Limpieza y Carga de Datos (ETL)

- **Extracción:** La información necesita ser extraída de las fuentes de datos para ser integrada.
- **Transformación:** Estos datos necesitan ser procesados sin inconsistencias (chequeo de discrepancias y obviamente eliminar los datos falsos). Serie de procedimientos especiales que permiten obtener un formato común unificado.
- **Carga:** Solo después de obtener datos limpios y procesados, pueden ser introducidos en el Almacén de Datos.

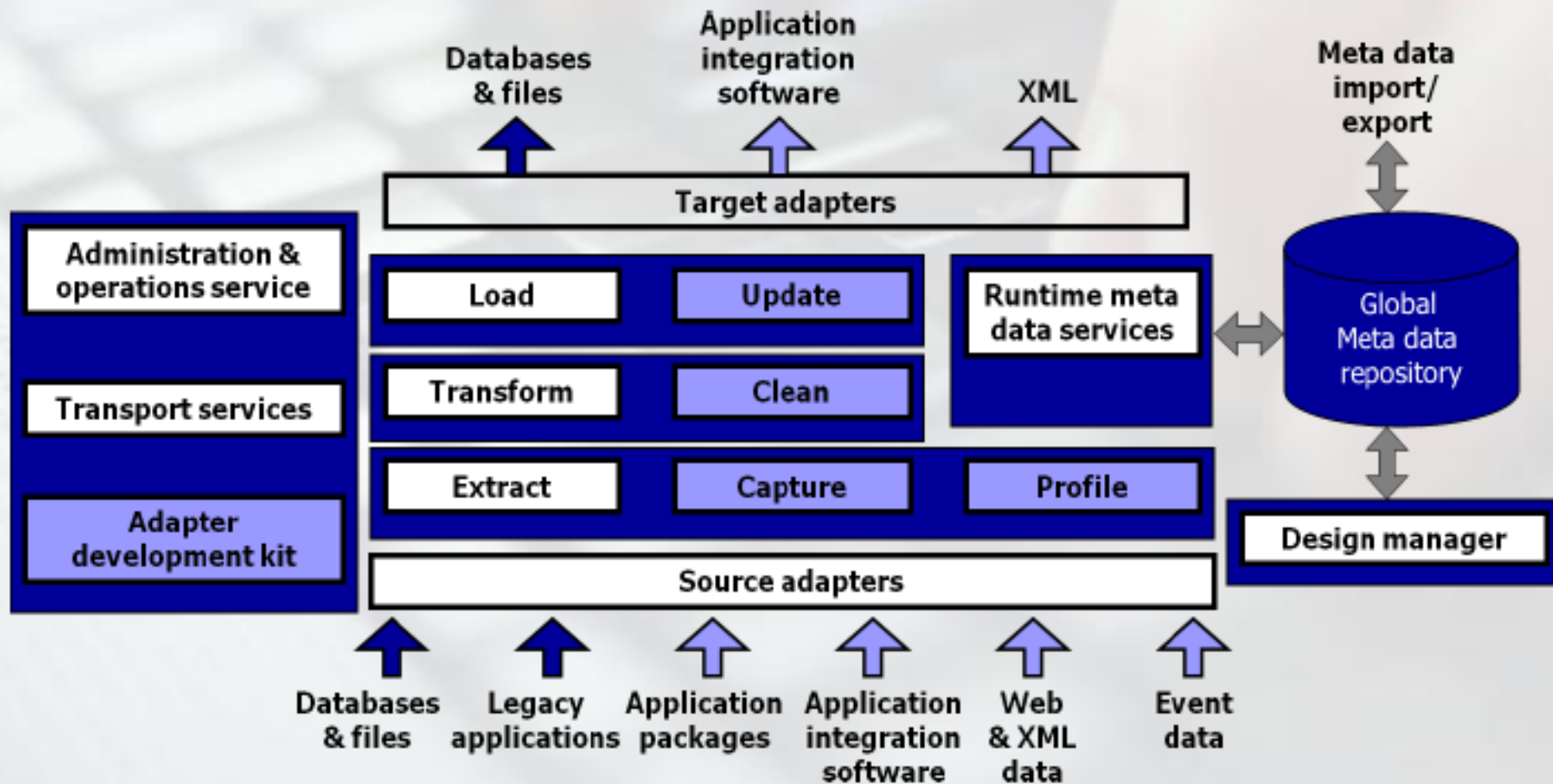
# Extracción, Transformación, Limpieza y Carga de Datos

- ETL Antes:



# Extracción, Transformación, Limpieza y Carga de Datos

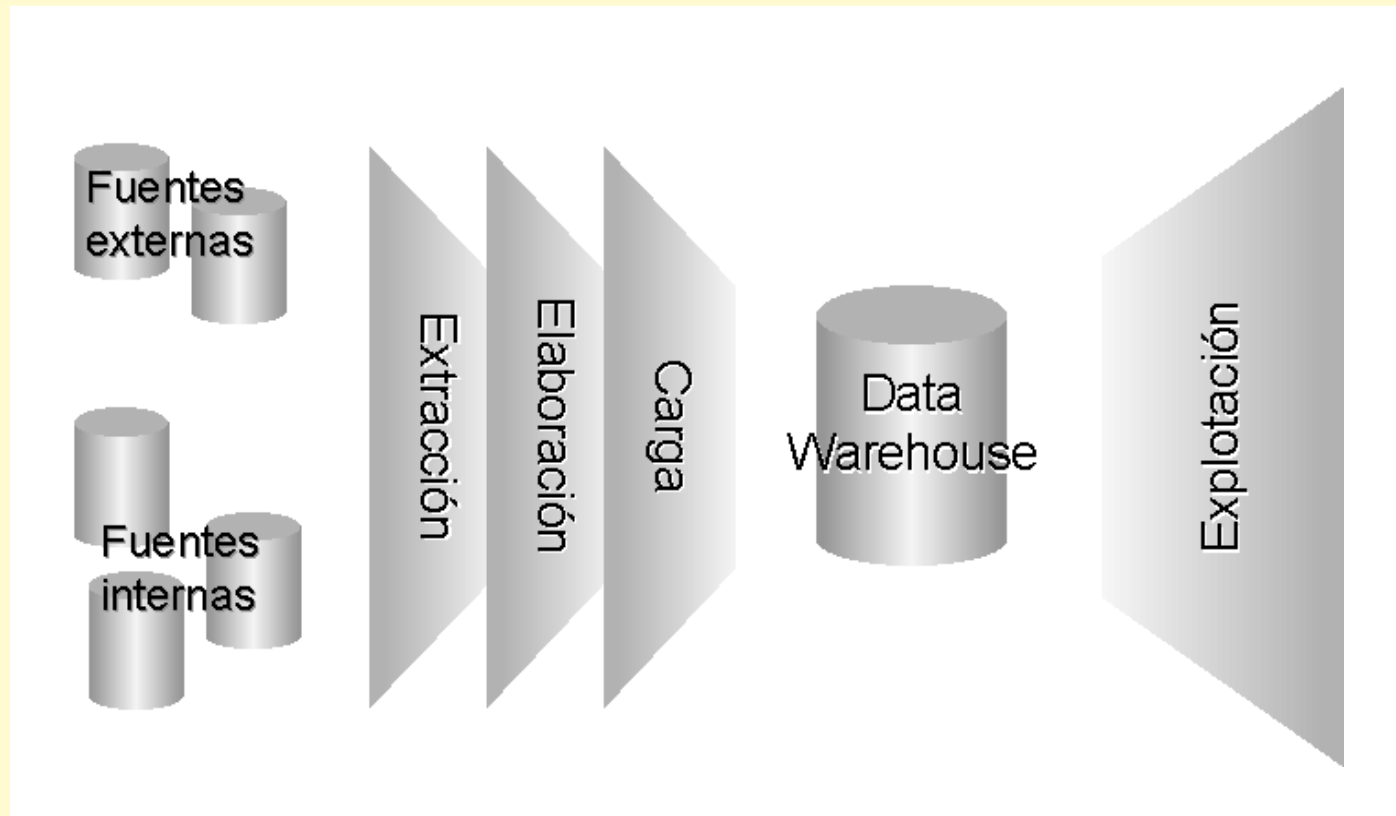
## • ETL Después:



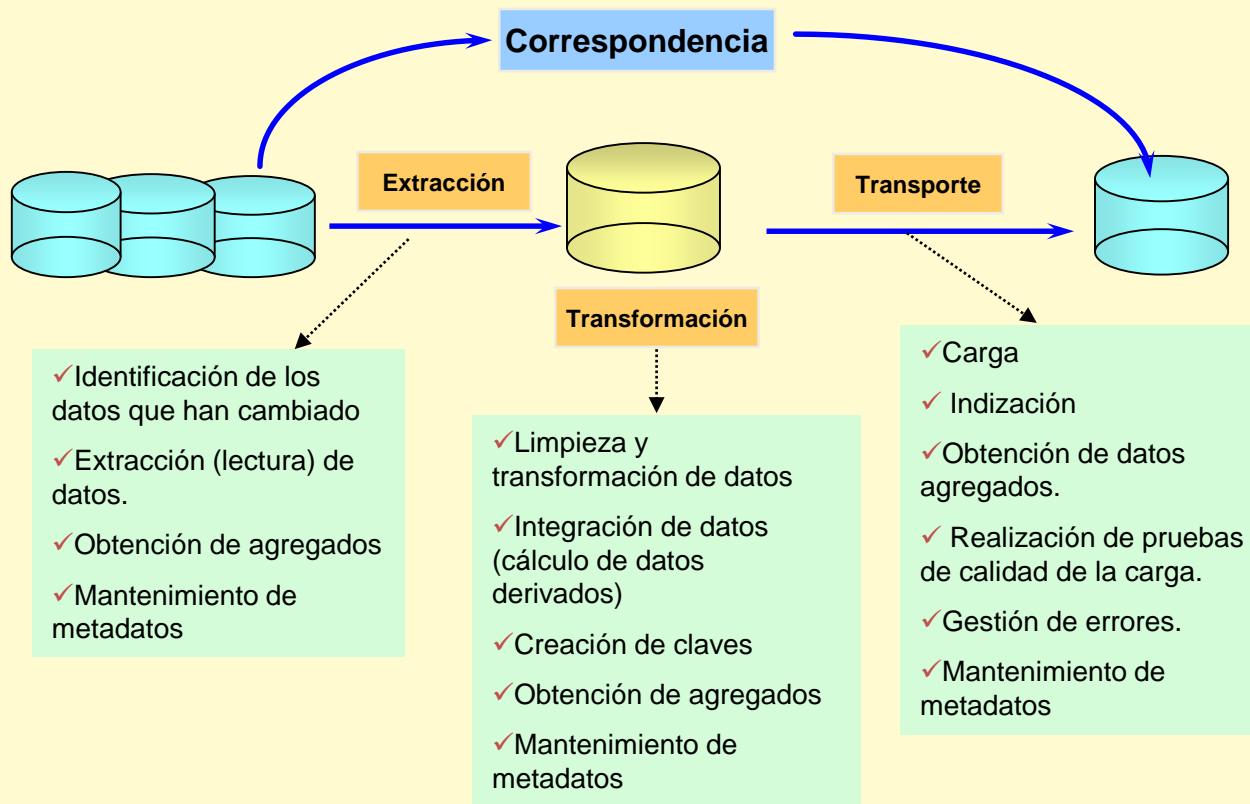


# Arquitectura de un Almacén de Datos

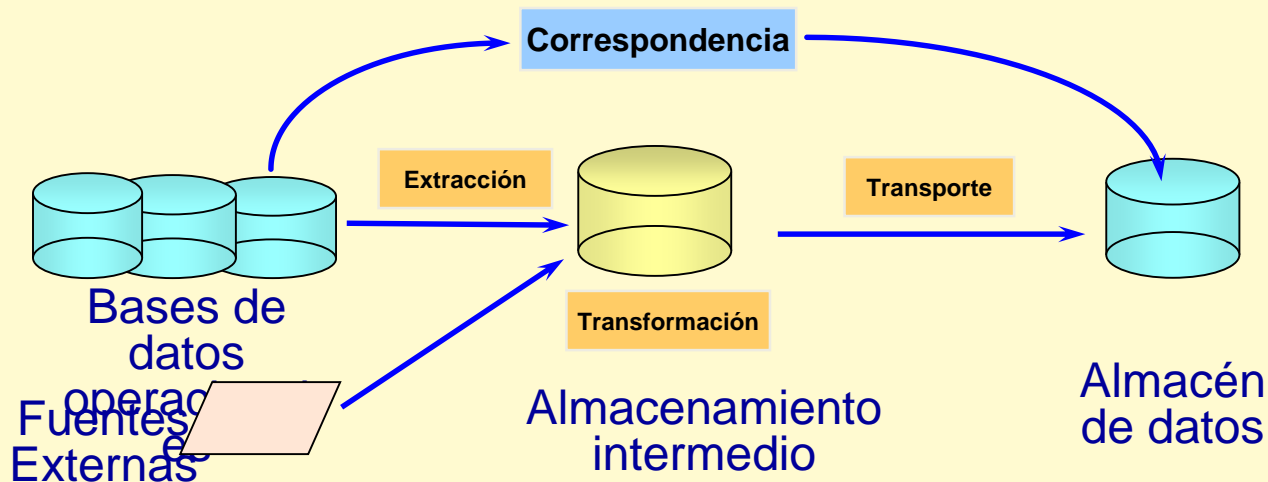
**La Arquitectura de un AD viene determinada por su situación central como fuente de información para las herramientas de análisis**



# ETL



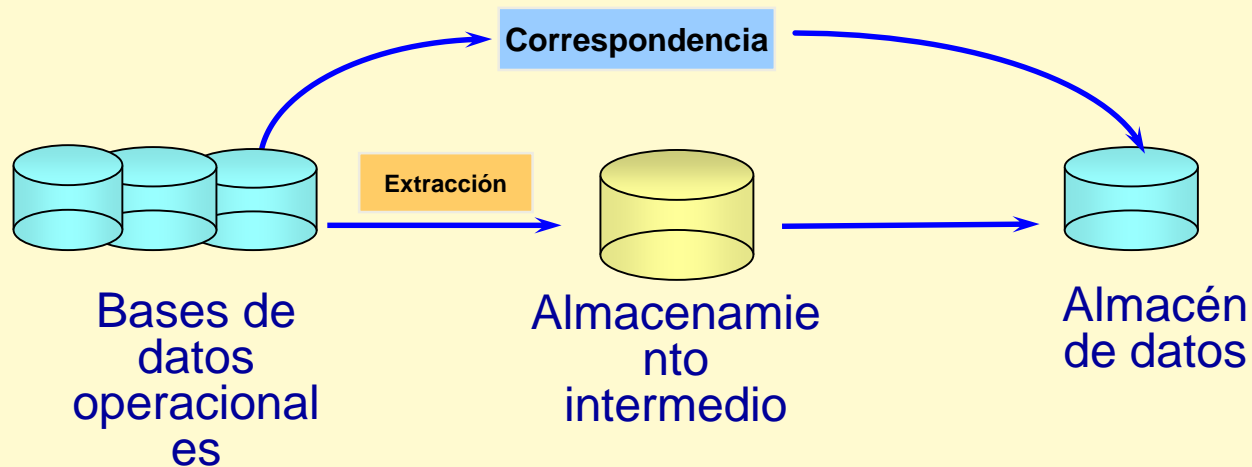
# Área de almacenamiento intermedio (Staging area )



El almacenamiento intermedio permite:

- Realizar transformaciones sin paralizar las bases de datos operacionales y el almacén de datos.
- Almacenar metadatos.
- Facilitar la integración de fuentes externas.

# Extracción



- Programas diseñados para extraer los datos de las fuentes.
- Herramientas: *data migration tools*, *wrappers*, ...



# Extracción

**Extracción:** lectura de datos del sistema operacional.

- a) durante la carga inicial
- b) mantenimiento del AD

**Ejecución de la extracción:**

- a) Datos en **un SGBDR** → **la extracción** de datos se puede reducir a consultas **en SQL** o rutinas programadas.
- b) Datos están en un **sistema propietario** o en **fuentes externas** → **extracción puede ser muy difícil** y puede tener que realizarse a partir de informes o de datos proporcionados por los propietarios que deberán ser procesados posteriormente.

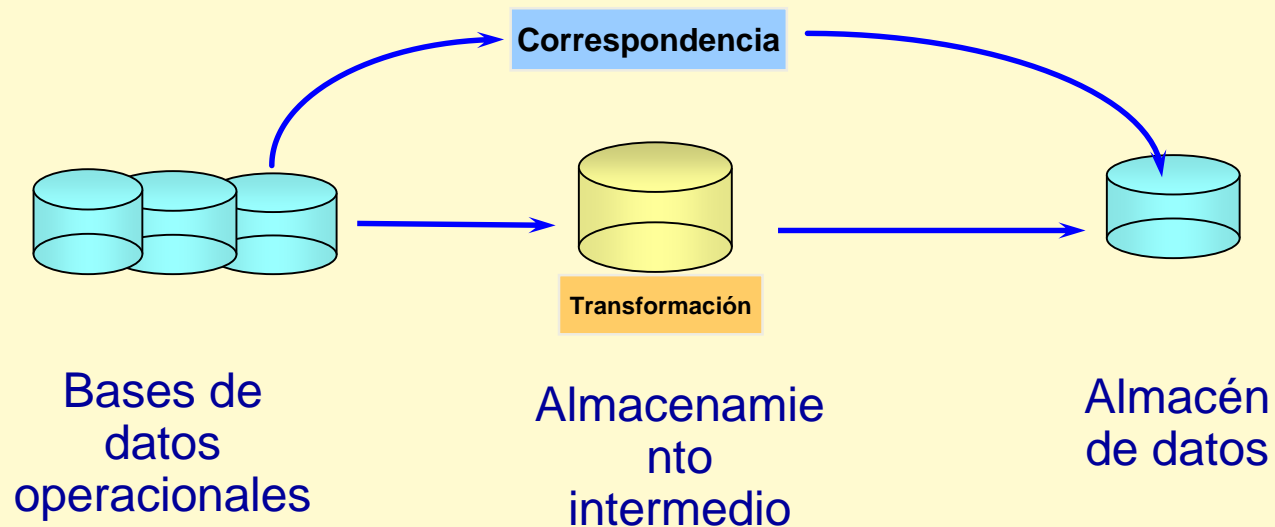
# Extracción

**Extracción:** es el mantenimiento/actualización del AD. Antes de realizar la extracción es preciso **Identificar los Cambios**.

## **Identificación de cambios.**

- Identificar los datos operacionales (relevantes) que han sufrido una modificación desde la fecha del último mantenimiento al AD.
- Métodos
  - Carga total: cada vez se empieza de cero.
  - Comparación de instancias de la base de datos operacional (snapshot).
  - Uso de marcas de tiempo (*time stamping*) en los registros del sistema operacional.
  - Uso de disparadores en el sistema operacional (solo disponible en algunos SGBDR).
  - Uso del fichero de *log* (gestión de transacciones) del sistema operacional.
  - Uso de técnicas mixtas.

# Transformación



- Transformar los datos extraídos de las fuentes operacionales: limpieza, estandarización (**cleansing**)
- Calcular los datos derivados: aplicar las leyes de derivación (**integration**)

# Transformación. Limpieza de datos

- Pocas fuentes de datos controlan adecuadamente la calidad de los datos.
- Los datos requieren frecuentemente de una limpieza antes de que puedan ser introducidos.

# Transformación. Limpieza de datos

- Los datos en el mundo real son:
  - Incompletos: atributos sin valor, falta de atributos interesantes para el contexto o el valor se tiene agregado. (Ej.: ocupación=" ")
  - Con ruido: contienen errores o outliers. (Ej.: Salario= -10)
  - Inconsistentes: contienen discrepancias: (Ej.: edad=60 y fecha\_nacimiento="03-03-2000")



# Transformación. Limpieza de datos

- Incompletos porque:
  - Dato no necesario o desconocido cuando se registra.
- Incorrectos debido a:
  - Error humano o del programa al introducir los datos.
  - Errores en la transmisión de datos.
- Inconsistentes porque:
  - Proviene de diferentes fuentes de datos.
  - Esquemas poco normalizados.

Una causa frecuente: no tener en cuenta la integridad.

# Transformación. Limpieza de datos

- Datos sin calidad ➔ resultados de análisis no fiables (registros duplicados o valores no asignados llevan a estadísticas incompletas).
- Un AD ➔ integración consistente de datos con calidad.

# Transformación. Limpieza de datos

- Las operaciones de limpieza típicas incluyen:
  - El llenado de valores ausentes, la corrección de errores tipográficos y otros de captura de datos.
  - El establecimiento de abreviaturas y formatos estándares.
  - El reemplazo de sinónimos por identificadores estándares.
- Realizar la corrección, si es posible, en cada fuente de datos; si no, en la tarea de transformación.

# Transformación. Limpieza

Se reconocen varios problemas:

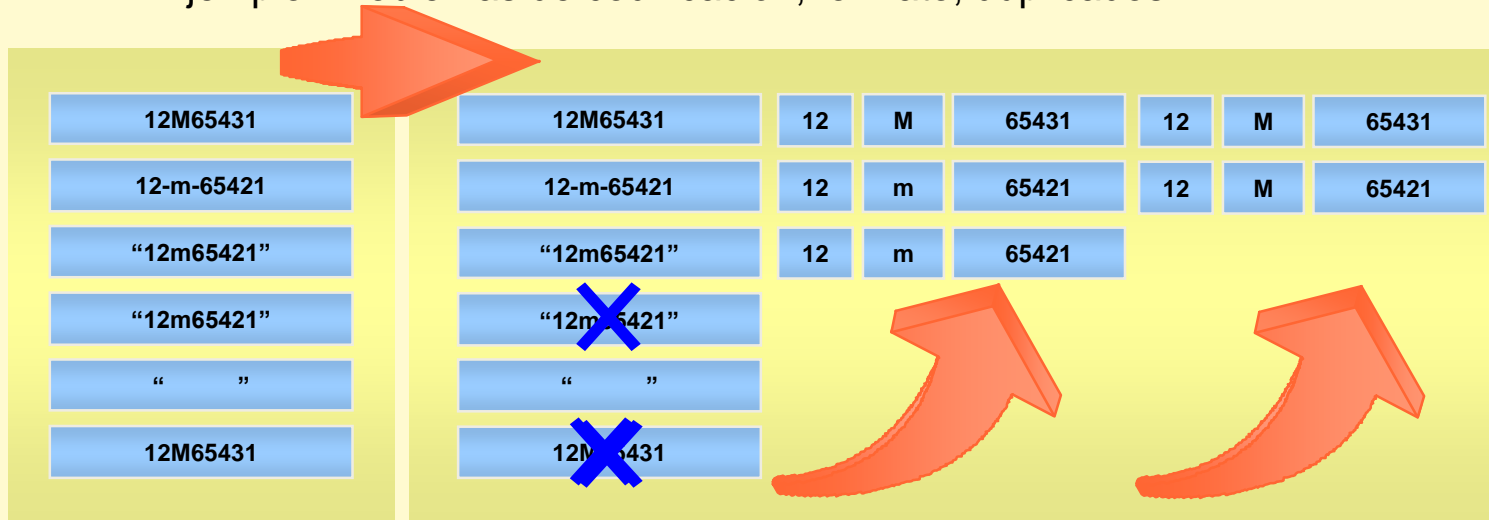
- Valores ausentes. Nulos.
- Descripción, abreviaturas, sinónimos y/o errores.  
Laura P. González. L.P. Glez
- Codificación:  
F,M Hombre, Mujer 1, 0
- Unidades:  
estatura 154? cm o mt?
- Formatos:  
Teléf. (042) 20375, 042-203753.
- duplicados

Eliminar anomalías:

- Limpieza de datos: eliminar datos, corregir y completar datos, eliminar duplicados, ...
- Estandarización: codificación, formatos, unidades de medida, ...

# Transformación

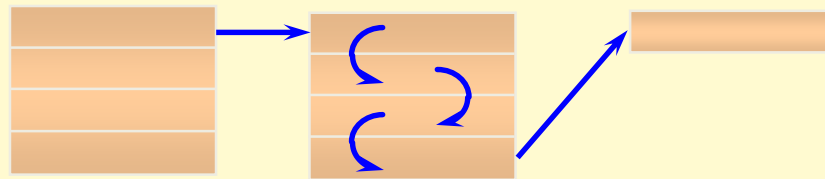
Ejemplo. Problemas de codificación, formato, duplicados.



- En los datos operacionales existen anomalías: desarrollos independientes a lo largo del tiempo, fuentes heterogéneas, ..

# Transformación

Ejemplo de formato (Claves con estructura): descomponer en valores atómicos



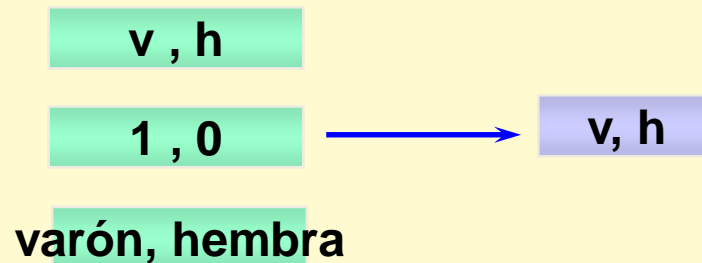
Código de producto = **12M65431345**

**código del país**   **zona de ventas**   **número de productos**   **código de vendedor**

# Transformación

Ejemplo. Codificación.

- Unificar codificaciones: existencia de codificaciones múltiples.

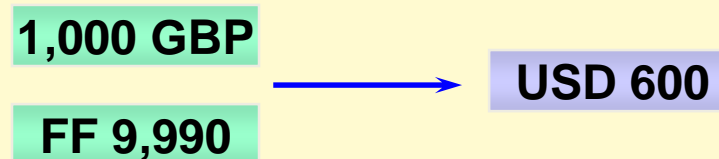
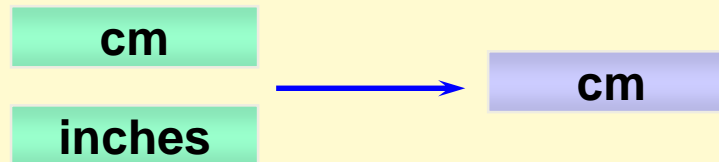


- Deben detectarse los valores erróneos.

# Transformación

Ejemplo. Unidades de medida.

- Unificar estándares: unidades de medida, unidades de tiempo, moneda,...





# Transformación

- Ejemplo. Duplicados, deben ser eliminados.



# Transformación

- Integridad referencial: debe reconstruirse, ahora con llaves sustitutas.

Departamento	Emp	Nombre	Departamento
10	1099	Smith	10
20	1289	Jones	20
30	1234	Doe	50
40	6786	Harris	60

# Transformación

## Transformación. Creación de claves sustitutas.

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

 Claves sin significado

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

# Transporte. Carga

- La fase de **Transporte** consiste en mover los datos desde las fuentes operacionales o el almacenamiento intermedio hasta el almacén de datos y cargar los datos en las correspondientes estructuras de datos.
- La carga puede consumir mucho tiempo.
- En la carga inicial del AD se mueven grandes volúmenes de datos.
- En los mantenimientos periódicos del AD se mueven pequeños volúmenes de datos.
- La frecuencia del mantenimiento periódico está determinada por la granularidad del AD y los requisitos de los usuarios.

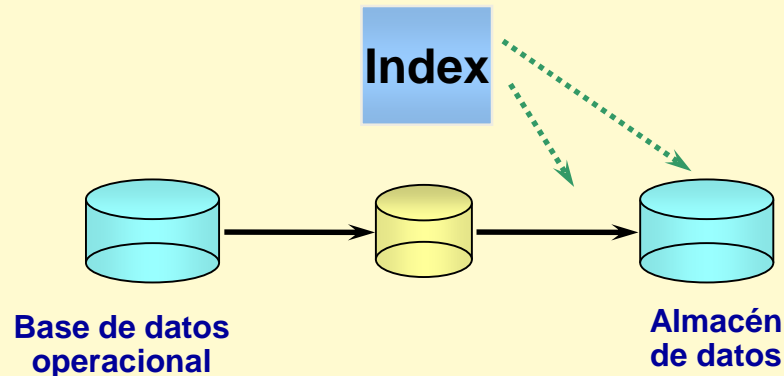
# Carga

- **Problemas:**
  - gran volumen de datos a ser cargados.
  - “Ventanas” de carga pequeñas (usualmente una noche) cuando el AD puede tenerse offline o el acceso es menor
  - Permitir al administrador del sistema monitorear status, cancelar, suspender, reanudar carga, o cambiar razones de carga.
  - Recuperar después de fallas sin perder integridad.
- **Técnicas:**
  - Utilitario de carga en lote: ordenar artículos de entrada sobre llave de clúster y usar E/S secuencial; construir índices y tablas derivadas
  - Usar paralelismo y técnicas incrementales

# Transporte. Indexación

## Proceso posterior a la carga.

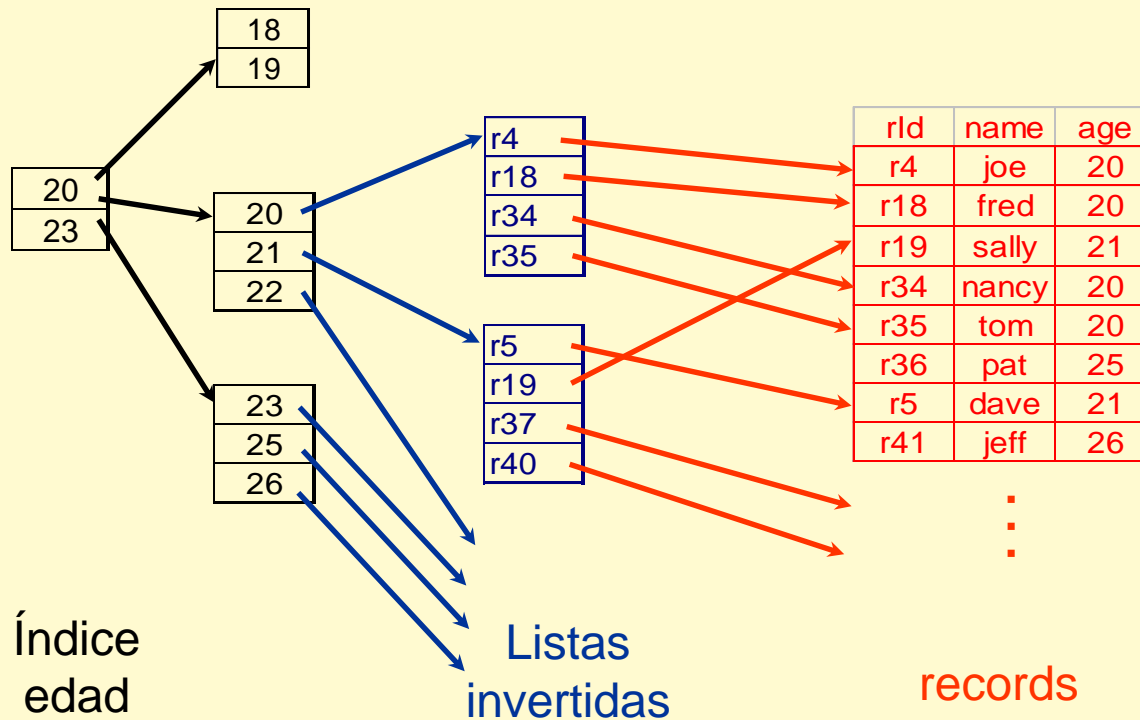
- Indexación durante la carga:
  - carga con el índice habilitado
  - proceso tupla a tupla (lento)
- Indexación después de la carga:
  - carga con el índice deshabilitado
  - creación del índice (total o parcial) (rápido)



# Índices

- Métodos de acceso tradicional
  - B-trees, tablas hash, R-trees, ...
- Popular en almacenes de datos
  - Listas invertidas
  - Índices bit map
  - Índices join

# Listas invertidas

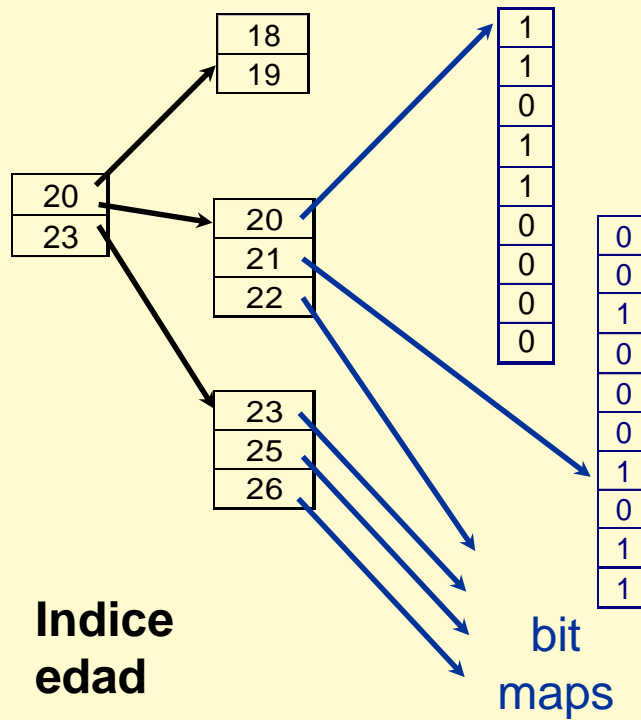




# Usando listas invertidas

- Query:
  - Obtener persona con edad = 20 y nombre = “fred”
- Listar por edad = 20: r4, r18, r34, r35
- Listar por nombre = “fred”: r18, r52
- Respuesta, intersección: r18

# Bit Maps



id	name	age
1	joe	20
2	fred	20
3	sally	21
4	nancy	20
5	tom	20
6	pat	25
7	dave	21
8	jeff	26

⋮  
records

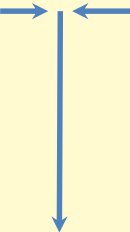
# Usando Bit Maps

- Query:
    - Obtener persona con edad = 20 y nombre = “fred”
  - Listar por edad = 20: 1101100000
  - Listar por nombre = “fred”: 0100000001
- 
- Adecuados para cardinalidad pequeña.
  - vectores de bits pueden comprimirse

# Acople

- “Combinar” relations Venta, Producto
- In SQL: `SELECT * FROM Venta, Producto ...`

venta	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

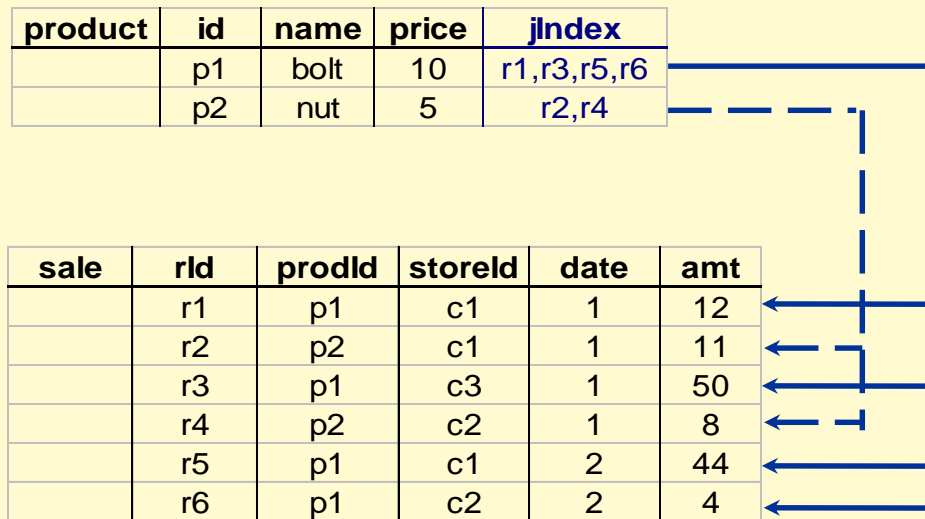


product	id	name	price
	p1	bolt	10
	p2	nut	5

joinTb	prodId	name	price	storeId	date	amt
	p1	bolt	10	c1	1	12
	p2	nut	5	c1	1	11
	p1	bolt	10	c3	1	50
	p2	nut	5	c2	1	8
	p1	bolt	10	c1	2	44
	p1	bolt	10	c2	2	4

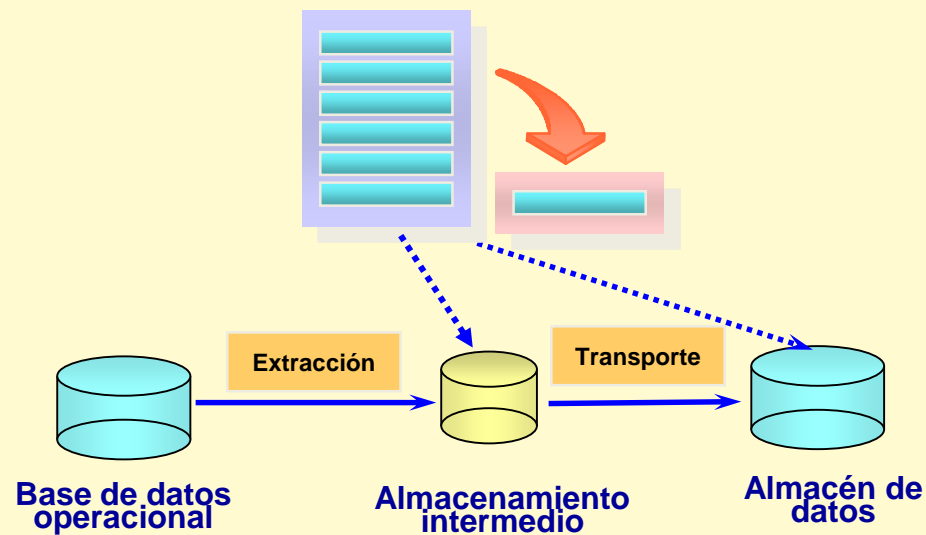
# Indices Join. (Semi-Join)

Indices join



# Transporte. Obtención de agregados

- Durante la extracción.
- Después de la carga (transporte).



# Mantenimiento de un AD (*Refresh*)

## Problemas:

### – Cuándo actualizar

- Por cada actualización: muy costoso, solo necesario si las consultas OLAP necesitan datos actuales.
- periódicamente (ej. cada 24 horas, todas las semanas) o después de eventos significativos.
- Políticas de actualización establecidas por el administrador basadas en necesidades de usuario.
- Diferentes políticas para diferentes fuentes.



# Subsistemas de Kimball

- Son un conjunto de buenas prácticas o plataforma sobre la que toda solución ETL debe construirse (¿Qué debería hacer una solución ETL?)
- Provee de un marco único para la construcción y evaluación de soluciones ETL.
- Dividida en 34 subsistemas, agrupados en:
  - Extracción
  - Limpieza y consolidación
  - Entrega
  - Administración

Si desea conocer los subsistemas de Kimball descargue el documento titulado de igual manera disponible en el Moodle.

# Metadatos

- Están relacionados con el monitoreo y administración del proceso ETL.
- El componente final del AD es el de los metadatos.
- De muchas maneras los metadatos se sitúan en una dimensión diferente al de otros datos del AD, debido a que su contenido no es tomado directamente desde el ambiente operacional.
- Un sistema controla:
  - Valores permitidos en cada campo (ej. UH, UCLV, etc.)
  - Descripción de los contenidos de cada campo (ej. fecha inicio)
  - Fecha de carga de los datos.
  - Fecha de última actualización.
  - Etc.

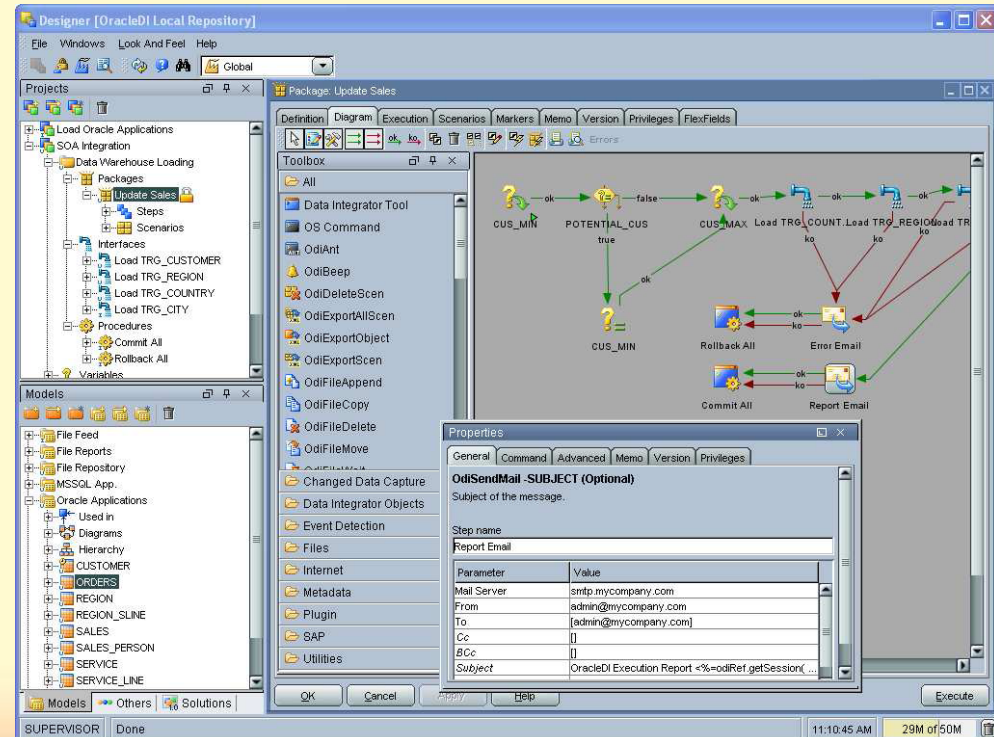
# Metadatos manejados

- BD fuentes y sus contenidos
- Descripciones de las puertas de entradas
- Esquema de DWH, definiciones de datos.
- Dimensiones y jerarquías.
- Consultas y reportes predefinidos.
- Mercados de datos y sus contenidos.
- Particiones de datos.
- Extracción de datos, limpieza, reglas de transformación, valores por defecto
- Reglas para purgar datos y refrescar
- Perfiles de usuario y grupos de usuarios
- Seguridad: autorización de usuario, control de acceso.

# Extracción, Transformación y Carga

Las herramientas de ETL suministran una interfaz gráfica (drag-and-drop) para diseñar los movimientos de datos.

Oracle Data Integrator



# ETL

- Los términos relacionados *ELT* (*extract, load, transform*) y *ETLT* (*extract, transform, load, transform*) son utilizados en dependencia de quien realice las transformaciones:
  - Transformaciones implementadas por el RDBMS (ELT).
  - Transformaciones implementadas por una herramienta especializada independiente del RDBMS (ETL).
  - Transformaciones implementadas por ambos el RDBMS y la herramienta(ETLT).

# Calidad de datos

- El concepto se asocia con frecuencia a sistemas de información con la **necesidad de precisión** en los datos gestionados.
- Los datos son la **materia prima para la toma de decisiones**.
- Existen organizaciones dedicadas al procesamiento de datos (*bancos, compañía de seguros, instituciones estadísticas, etc.*) y otras que sus decisiones y actividades son guiadas por Sistemas de Información.
- La **búsqueda de la calidad** en los datos debe ser un **proceso priorizado** dentro de todo el ciclo de vida de los datos.

# Calidad de datos

- La **Calidad de Datos** es un término que abarca tanto el estado de los datos, así como el conjunto de procesos para lograr dicho estado.
- El objetivo es disponer de datos **libre de duplicados**, errores, omisiones, variaciones e innecesarios, se debe disponer de los datos necesarios y ajustados a la estructura definida.
- Los datos deben ser correctos, inequívocos, coherentes y completos.



# Calidad de datos

- **Correctos:** Los valores y las descripciones de los datos deben describir su **verdadera definición**.
- **Inequívocos:** Los valores y las descripciones de los datos sólo pueden tener un **único significado**.
- **Coherentes:** Los valores y las descripciones de datos deben usar una **notación constante** para transmitir su verdadero significado. *Ejemplo:* para mantener la coherencia de los datos se debe utilizar solo una nomenclatura.
- **Completos:** Se debe garantizar que los valores individuales y las descripciones de los datos, **se definan para cada caso**, permitiendo identificar que valores posibles puede tomar cada dato y se debe asegurar que el número total de registros completados después que se realice el proceso de integración debe ser del 100% completo o por lo menos asegurarse de que no se pierda información en alguna parte del flujo de datos.

# Calidad de Datos

Perfilado de Datos

Limpieza de Datos

Auditoría de Datos

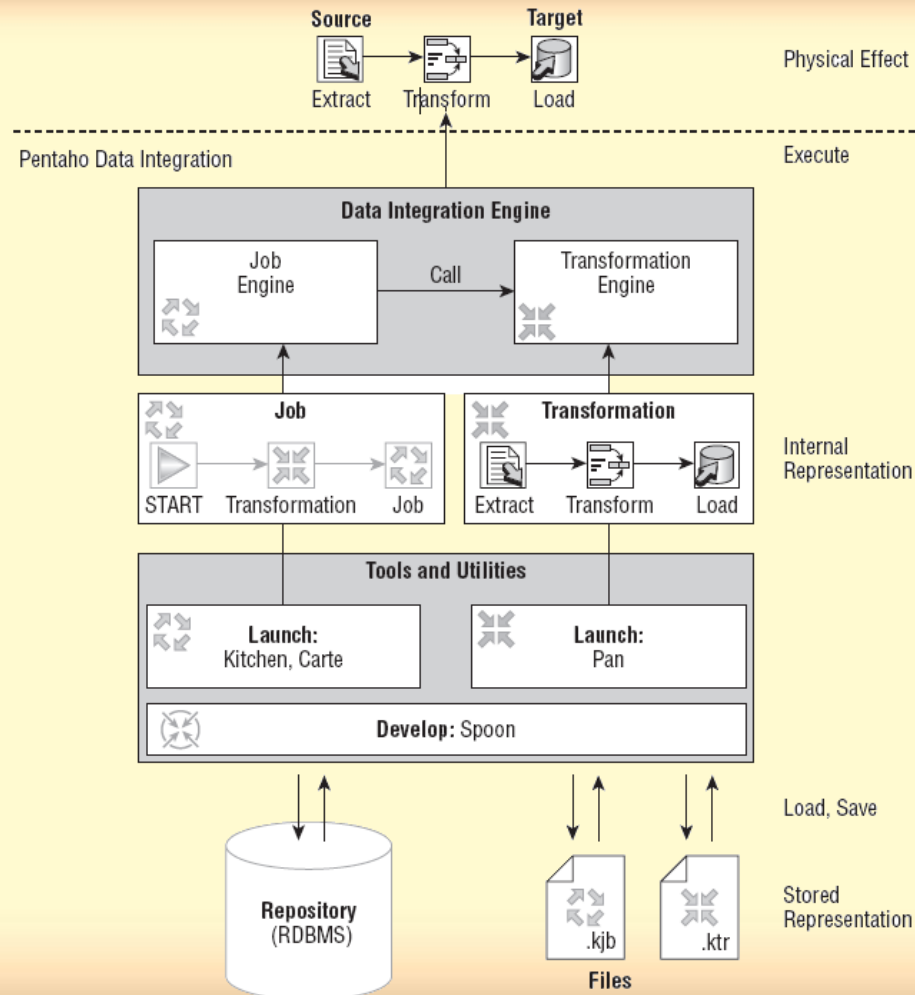
Garantiza la  
calidad de  
los datos.

- **Perfilado de Datos** consiste en el proceso de examinar los datos que existen en una organización y recopilar estadísticas e información sobre los mismos.
- La **Limpieza de Datos** es el proceso de detectar o descubrir, y corregir datos corruptos, incoherentes o erróneos de un conjunto de datos. Permite detectar entradas duplicadas, incompletas y establecer reglas para corregirlas.
- **Auditoría de datos** es el proceso de gestionar cómo los datos se ajustan a los propósitos definidos por la organización.

# Herramientas de ETL más populares

- Oracle Data Integrator
- Oracle Warehouse Builder
- Information Builder
- Talend Open Studio
- **Pentaho Data Integration**
- entre otros ...

# Pentaho Data Integration (Arquitectura)



# Pentaho Data Integration (Kettle)

- Maneja dos tipos diferentes de objetos:
  - Transformaciones: Flujos orientados a los datos (ETL).
  - Trabajos (Jobs): Orquestación de Transformaciones.
- Herramientas del PDI:
  - **Spoon:** Herramienta gráfica para crear transformaciones y trabajos.
  - **Kitchen:** Herramienta a nivel de línea de comandos para ejecutar trabajos.
  - **Pan:** Herramienta a nivel de línea de comandos para ejecutar transformaciones.
  - **Carte:** Un servidor de aplicaciones ligero para ejecutar trabajos y transformaciones en un servidor remoto.



# PDI

- El motor de integración de datos (*job engine + transformation engine*) está físicamente implementado como una librería de Java.
- El *front-end* usa un API para interactuar con el motor de integración en respuesta a la interacción del usuario (No hay restricción del uso del motor mediante la herramienta solamente).

# PDI. Repositorio

- Los trabajos y las transformaciones pueden ser almacenadas en una base de datos o en una carpeta dentro de un directorio de su PC.
- La herramienta de diseño de los trabajos y transformaciones puede conectarse a la base de datos y cargar estos objetos para ser rediseñados y ejecutados.
- Cuando no se utiliza el repositorio, las transformaciones y trabajos pueden ser almacenadas en archivos (formato XML). En este caso, puede utilizarse un mecanismo externo para el control de versiones como Subversion o CVS para facilitar la colaboración.

# PDI. Repositorio

PDI permite seleccionar que tipo de recurso desea utilizar para almacenar su repositorio:

**1.Base de datos:** El repositorio usa una base de datos relacional central para almacenar los metadatos de todo el proceso de las ETL (transformaciones y trabajos).

**2.Archivo:** El repositorio queda almacenado en cierta carpeta en un determinado directorio.



# PDI. Transformaciones

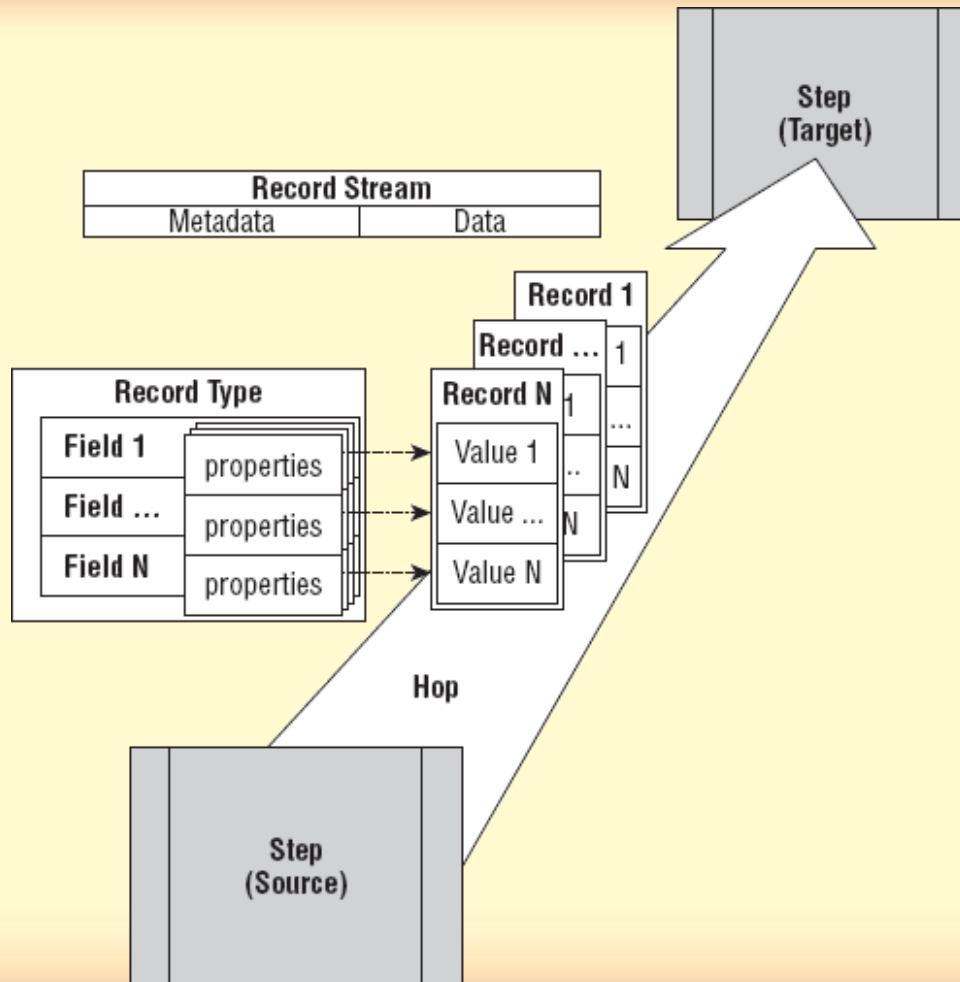
- Colección de pasos (*steps*)
- Un paso denota una operación particular sobre uno o varios bloques de registros (*record streams*)
- Un bloque de registros es una serie de registros (*records*) similar a una tabla de un RDBMS (campos, valores, etc.)
- Los pasos pueden estar conectados por flechas llamadas *hops* ( $\rightarrow$ )
- Un *hop* permite el flujo de un registro de un paso a otro
- Las transformaciones se guardan con extensión .ktr

# PDI (Transformación sencilla)

- **Transformación:** Carga los datos de un archivo, adiciona una constante (*valor*) y lo salva hacia un nuevo archivo.



# PDI. Hop



# PDI. Trabajos

- Los trabajos están compuestos por uno o varias transformaciones. Garantiza el orden adecuado.

Por ejemplo:

**[transf. para extraer datos] -> [transf. para cargar los datos]**

- Los trabajos permiten además, vaciar tablas, mover archivos de un servidor a otro, enviar mensajes por correo y otros.
- Los trabajos se guardan con extensión .kjb

# PDI. Conexiones

- Pentaho Data Integration soporta conexiones diferentes RDBMS (basado en JDBC) tales como:  
IBM DB2, Microsoft SQL Server, MySQL, Oracle, PostgreSQL, JDBC.
- JDBC es un estándar bien establecido y la mayoría de los vendedores de RDBMS suministran manejadores JDBC para sus RDBMS.

# Introducción al PDI

- Herramienta gráfica de Pentaho Data Integration para crear transformaciones y trabajos.
- Interactúa mediante una API (Application Programming Interface) con el Motor de Pentaho Data Integration (Job Engine + Transformation Engine).
- Es independiente del Motor.

# ¿Cómo instalar el PDI?

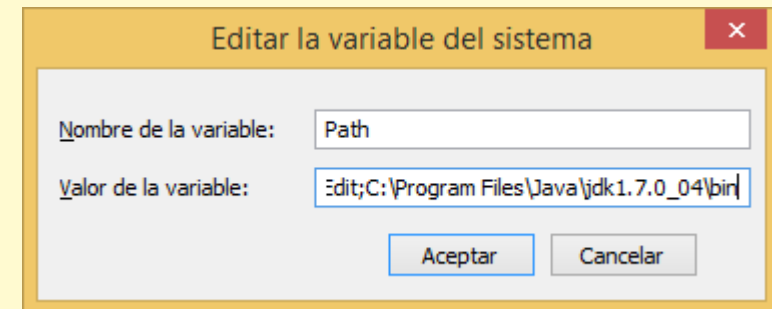
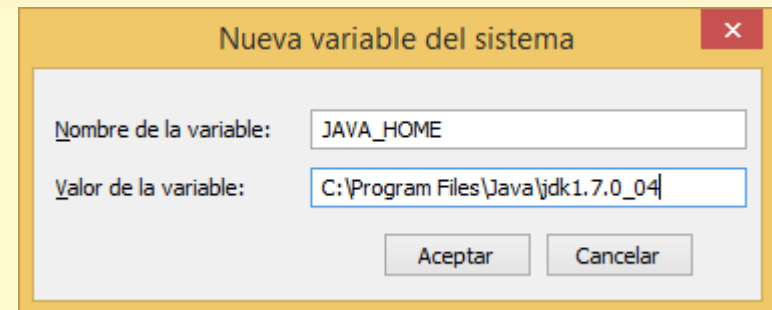
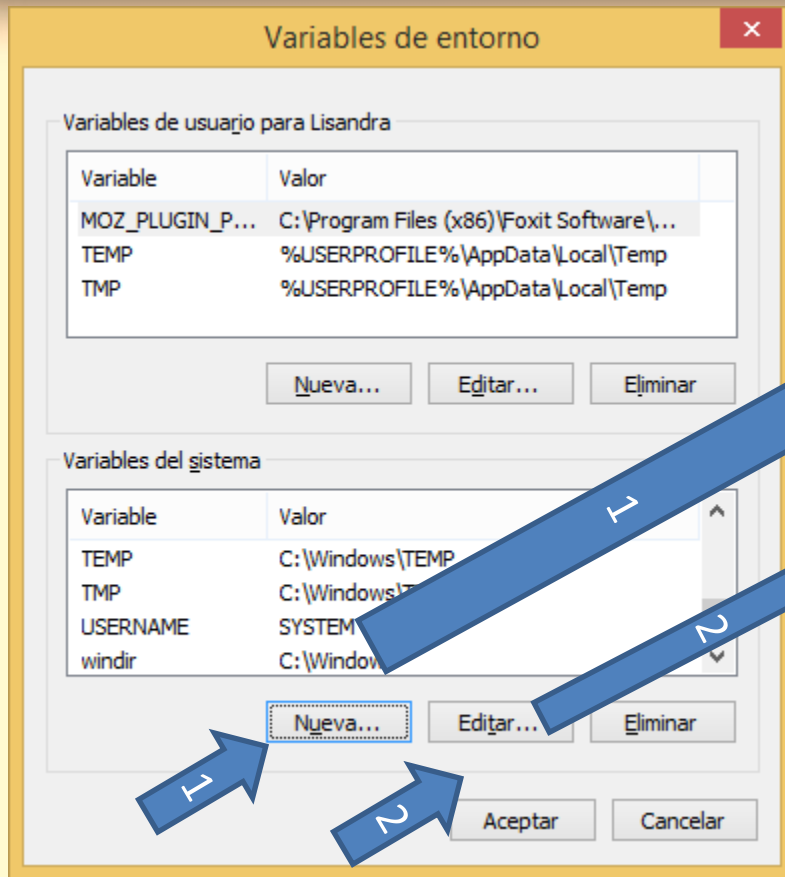
- Todas las herramientas y componentes de **Pentaho Data Integration** se encuentran disponibles para descargar (.zip) en el área de descargas de la página del Proyecto Pentaho en <http://sourceforge.net/projects/pentaho>.
  - En la nube uclv descargar el .rar titulado *pdi-ce-5.0.1-stable*
- **Pentaho Data Integration** no requiere un prodecimiento adicional más allá de descompactar la descarga.
- **Prerequisitos:**
  - JDK según la versión de Pentaho Data Integration (usamos la versión 1.7).
  - Configurar la variable de entorno JAVA\_HOME

# Configurar variable de entorno JAVA\_HOME

- Ir a Mi PC, clic secundario “Propiedades”
- Seleccionar “Configuración avanzada del sistema”
- Luego en la ventana “Propiedades del sistema”, clic en el botón “Variables de entorno”
- Ir a las variables del sistema y crear una Nueva, en el nombre poner exactamente así JAVA\_HOME y editar Path.



# Variables de entorno JAVA\_HOME y Path

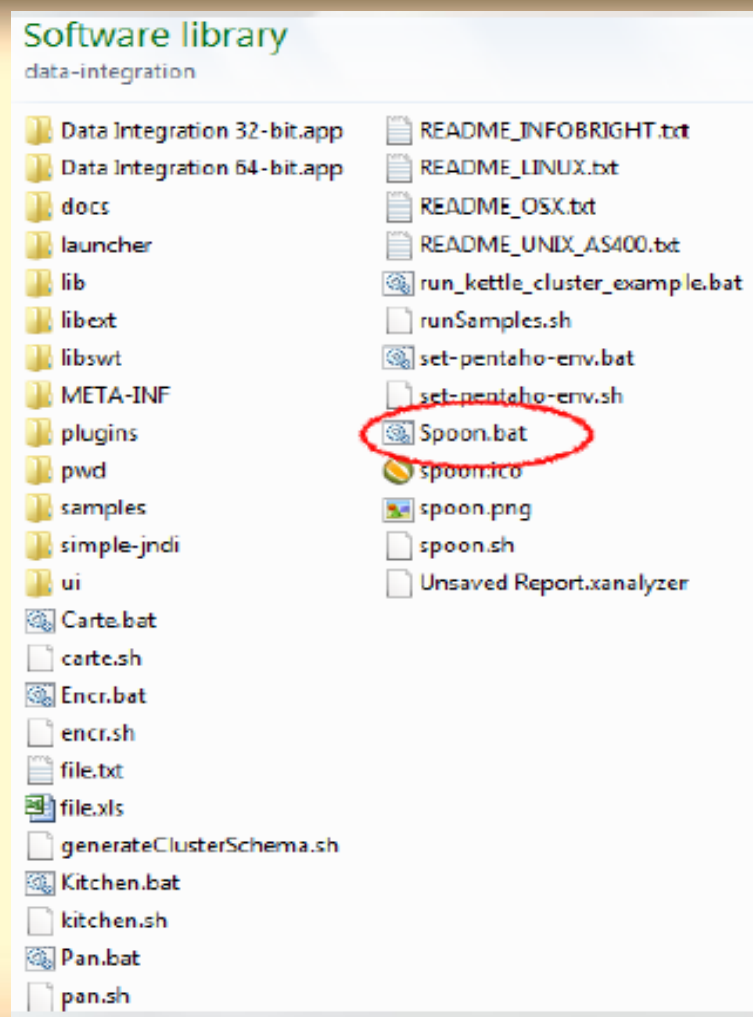


En la variable Path, ir al final poner un punto y coma y luego la dir del jdk /bin

# Introducción al Spoon

Estructura del archivo  
descargado (PDI v5.0.1)

Ejecutar el fichero  
**Spoon.bat** para Windows  
o *spoon.sh* para Linux



# Spoon

- La herramienta de diseño de trabajos y transformaciones (Spoon) puede conectarse a la BD y cargar estos objetos para ser rediseñados y ejecutados.
- Las transformaciones y trabajos también pueden ser almacenadas en archivos (formato XML). En este caso, puede utilizarse un mecanismo externo para el control de versiones como Subversion o CVS para facilitar la colaboración.
- Además PDI permite exportar e importar los repositorios creados.

# Creando el repositorio

**Repository Connection**

Repository:

1

User Name:  
admin

Password:

☒ Show this dialog at startup

OK Cancel

**Select the repository type**

Filter

Select the repository type to create

Kettle database repository : This repository uses a central relational database to store ETL metadata.

Kettle file repository : This is a repository stored in a file in a certain folder.

Vale Cancelar

**File repository settings**

Base directory C:\Users\Lisandra\Desktop\Repo\_ETL Examinar...

Read-only repository? ☐

Hide hidden folders and ☐

ID 1

Name Repo\_ETL

Vale Cancelar

# Repositorio

- Luego hacer clic en el botón Aceptar, nótese que con esta opción no se necesita utilizar contraseña, aquellos que deseen probar el repositorio desde una BD deben:
  - Crear una BD vacía en un SGBD ej. PostgreSQL, deben fijarse que el driver jdbc de este gestor esté en la carpeta *lib* dentro de *data integration*. Viene *por defecto*. Si no esta, lo descarga y lo copia para esta carpeta y luego vuelve a iniciar el Spoon.

# Repositorio

- Despues debe seleccionar la opción 1 donde se refiere a una BD, establecer la conexión con la BD siguiendo los pasos de la ventana de configuración.
- Para acceder al repositorio debe utilizar nombre de usuario *admin* y contraseña *admin*.
- Se crearan 42 tablas con todos los metadatos necesarios.
- OJO esta BD no es el almacén de datos, se usa solo para los metadatos.

# Repositorio

- Una vez creado su repositorio por una de las dos variantes, mejor la de los archivos porque es más portable simplemente copia los ficheros y los puede abrir desde cualquier lugar.
- Le saldrán una ventana con varios consejos (tips) sobre el uso del Spoon, léalos y cuando termine haga clic en Cerrar.

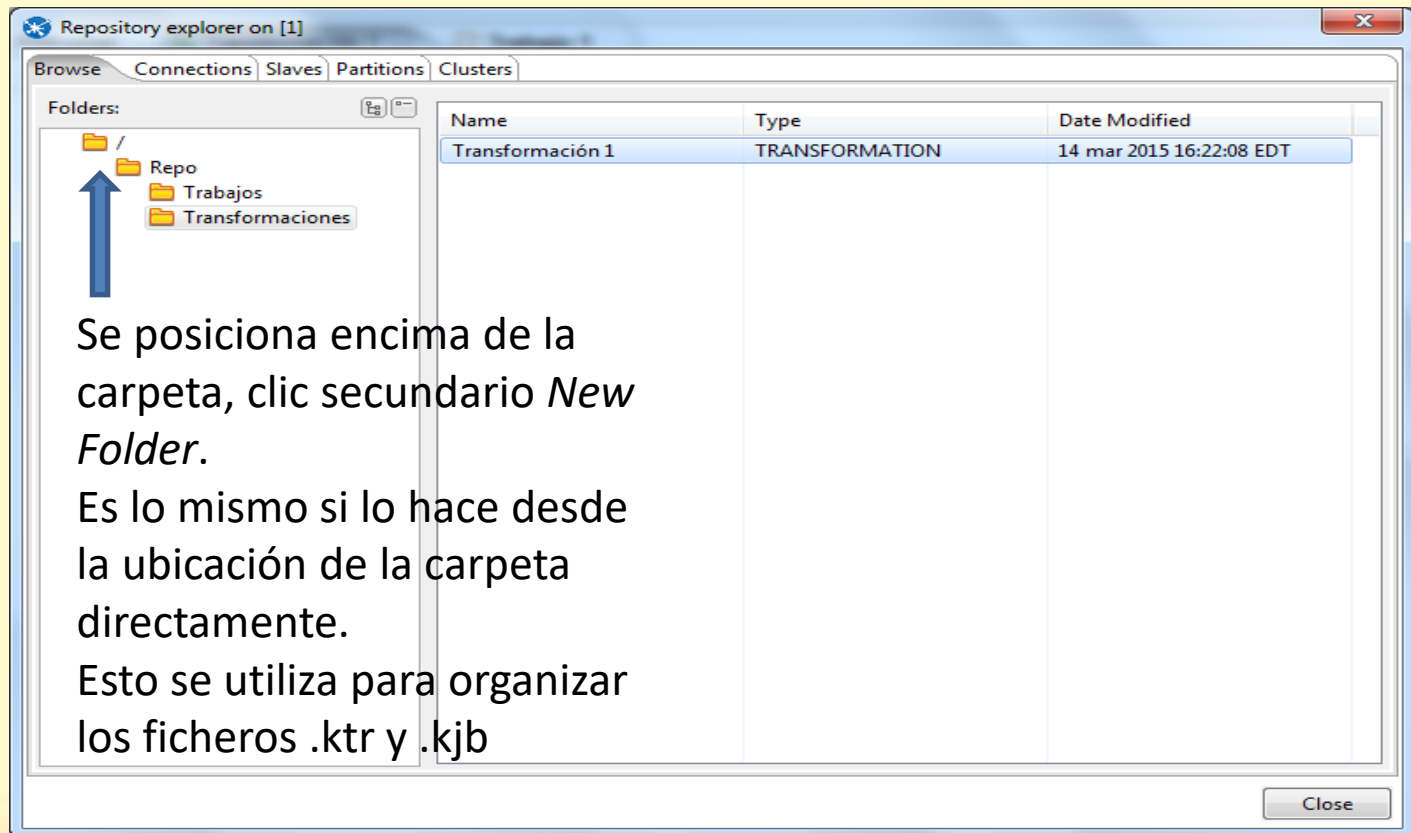
# Ventana de inicio del PDI





# Explorando el catálogo

Ir al menú **Tools – Catálogo – Explorar catálogo** y crear la estructura de carpetas que desee, en este caso se creó la carpeta Repo con las subcarpetas Transformaciones y Trabajos.




# Características de los pasos

1. Todos los pasos se encuentran dentro de la pestaña Diseño (*Design*)
2. Cada paso (step) pertenece a una categoría por ejemplo Input tiene los pasos de entrada como Excel Input, Text Input, Table Input, etc.
3. Para que se activen los pasos deben crear una transformación (*transformation*) o un trabajo (*job*), pero primero deben crear una nueva transformación.

# Características de los pasos

- Los pasos de las Transformaciones y los Trabajos son diferentes aunque existen Categorías en común.
- Cada paso tiene una ventana de configuración diferente, algunas menos intuitivas que otras.
- Aunque no es necesario programar a código las instrucciones, deben aprender a dominar los pasos esenciales y aprender a configurarlos para que el flujo de datos sea lógico y coherente. En dependencia de lo que deseen lograr.

# Vías para crear una nueva transformación

1. Ir al menú File -> New -> Transformation
2. Apretar Ctrl+N
3. Hacer clic en el ícono  y elegir Transformation

Una vez creada, se les activan los pasos en la pestaña Diseño.  
Lo primero que deben hacer es guardar la transformación en su repositorio.

# Guardar la transformación creada

- Ir al menú File – Save as

The screenshot shows a dialog box titled 'propiedades transformación' with several tabs: 'transformación', 'Parameters', 'Archivado', 'Fechas', 'Dependencias', 'Miscelaneos', and 'Monitoring'. The 'transformación' tab is active. It contains the following fields:

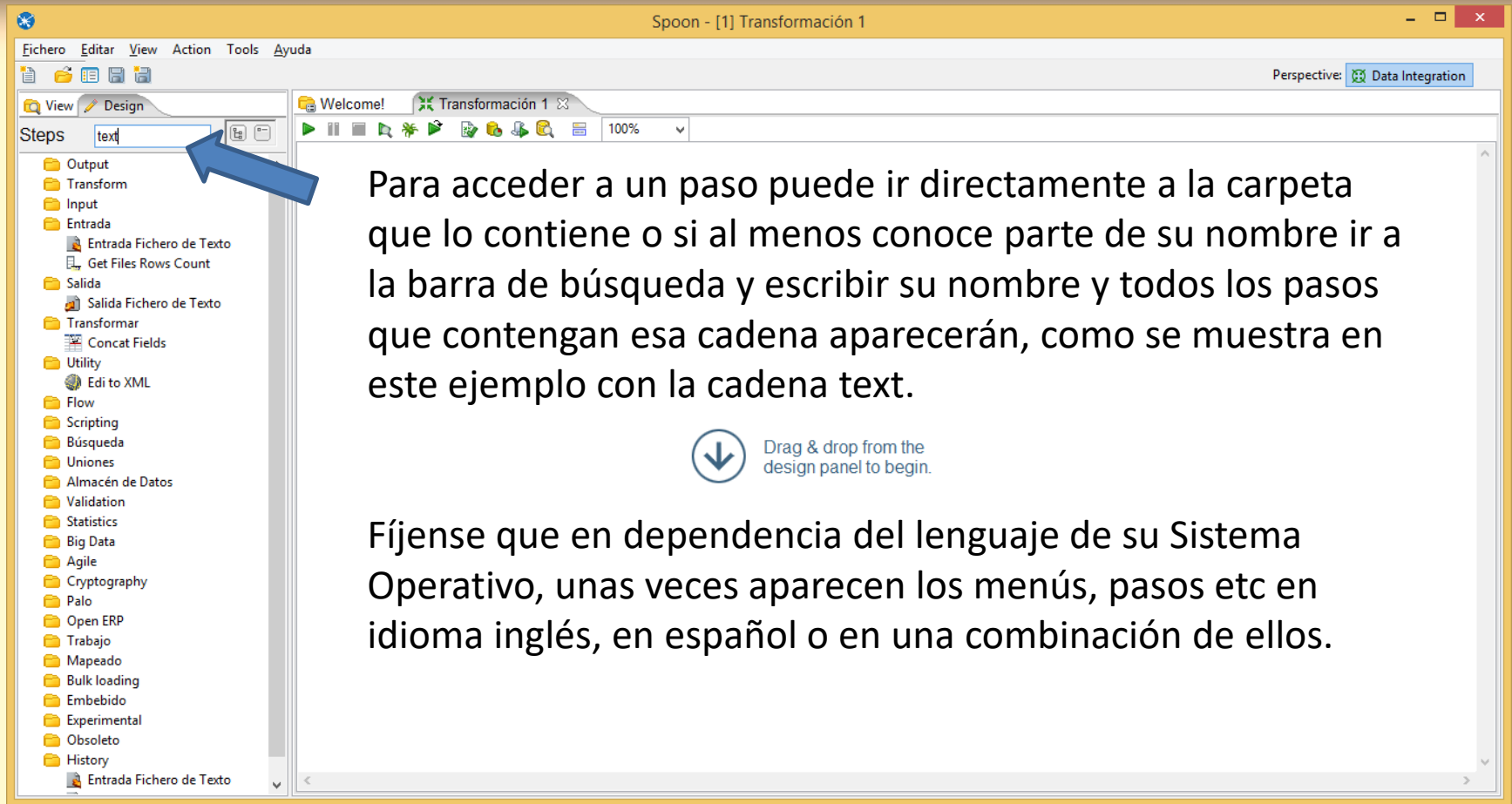
- 'nombre transformación:' with the value 'Transformación 1'. A blue arrow labeled '1' points to this field.
- 'Transformation filename:' (disabled)
- 'Description:' (empty)
- 'Extended description:' (empty text area)
- 'Status:' (dropdown menu)
- 'Version:' (empty)
- 'Directorio:' with the value '/Repo/Transformaciones'. A blue arrow labeled '2' points to the folder icon on the right of this field.
- 'Created by:' with the value '-'
- 'Created at:' with the value 'Mon Mar 15 02:29:42 GMT-11:00 2021'
- 'Ultima modificación por:' with the value '-'
- 'Ultima modificación a:' with the value 'Mon Mar 15 02:29:42 GMT-11:00 2021'

At the bottom, there are three buttons: 'Vale', 'SQL', and 'Cancelar'. A blue arrow labeled '3' points to the 'Vale' button.

Ponerle un nombre a la transformación que tenga relación con lo que debe hacer

Seleccionar el camino dentro del repositorio creado

# Pasos activos



Para acceder a un paso puede ir directamente a la carpeta que lo contiene o si al menos conoce parte de su nombre ir a la barra de búsqueda y escribir su nombre y todos los pasos que contengan esa cadena aparecerán, como se muestra en este ejemplo con la cadena text.

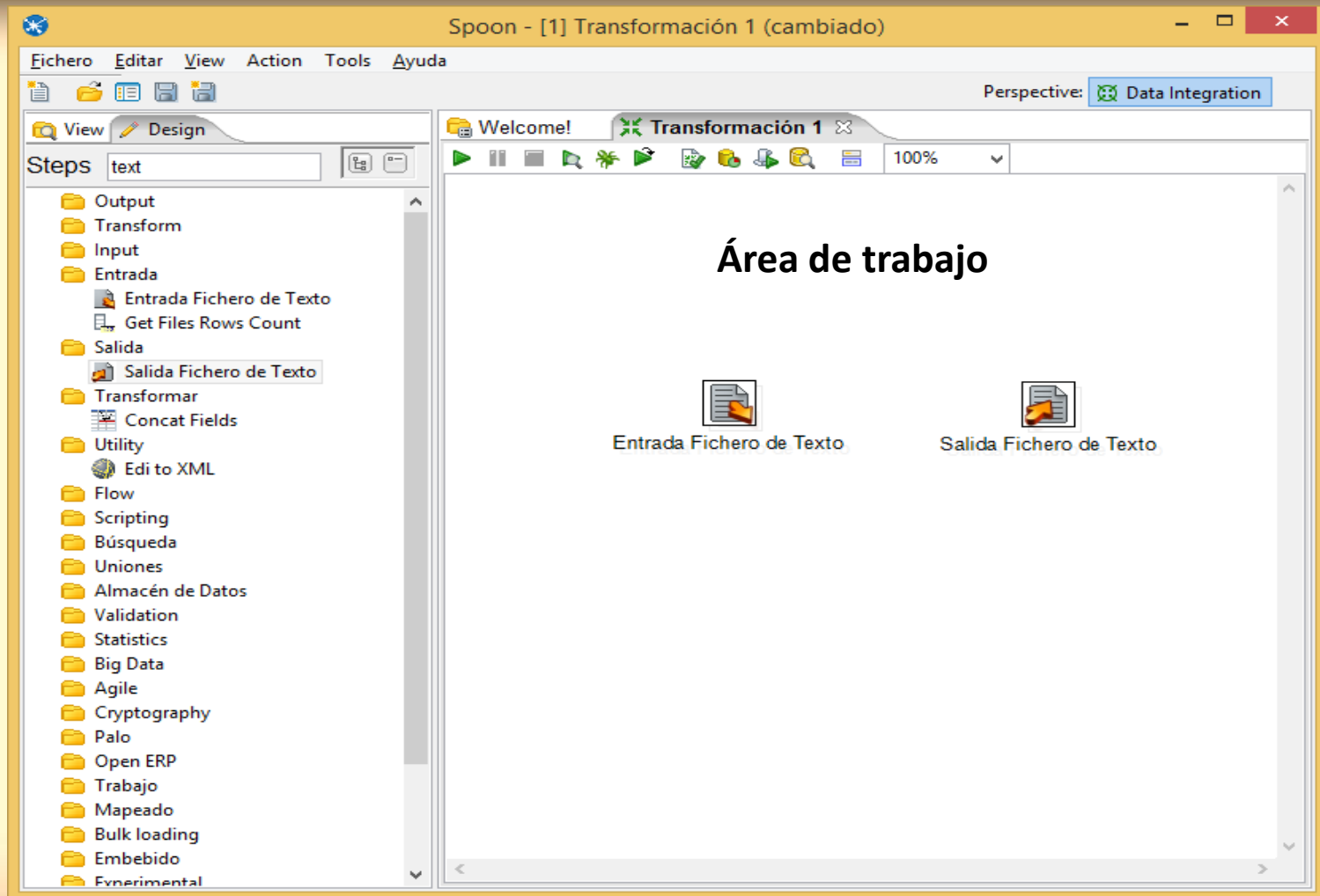
Drag & drop from the design panel to begin.

Fíjense que en dependencia del lenguaje de su Sistema Operativo, unas veces aparecen los menús, pasos etc en idioma inglés, en español o en una combinación de ellos.

# Ejemplo resuelto 1

- Supongamos que se tiene un fichero texto con datos que se desean mover exactamente como aparecen hacia otro fichero texto (copia de ficheros).
- Para ello se necesitan dos pasos uno de entrada y otro de salida para ficheros texto denominados (Entrada Ficheros de Texto y Salida Ficheros de Texto).
- Para que los pasos aparezcan en el área de trabajo posiciónese encima del paso arrástrelo y suéltelo encima del área de trabajo, como se muestra en la próxima diapositiva.

# Ejemplo resuelto 1





# Configurando el paso de entrada texto

The screenshot shows the 'Entrada Fichero de Texto' (Text File Input) dialog box. A blue arrow points to the 'Fichero' tab. Another blue arrow points to the 'Añadir' button. A third blue arrow points to the 'Examinar...' button. A fourth blue arrow points to the first row of the 'Ficheros seleccionados' table.

Nombre paso: Entrada Fichero de Texto

Tab: Fichero

Fichero o directorio:

Expresión Regular:

Exclude Regular Expression:

Ficheros seleccionados:

#	Fichero/Directorio	Comodín	Exclude wildcard	Requerido	Include subfol
1	C:\Users\Lisandra\Desktop\Nombres.txt			N	N

Buttons: Añadir, Examinar..., Eliminar, Editar

Acceptar nombres de fichero de pasos anteriores

Acceptar nombres de fichero de pasos anteriores ☐

Pass through fields from previous step ☐

Paso desde el que se obtienen los nombres de fichero:

Campo de entrada a utilizar como nombre de fichero:

Mostrar Fichero(s)... Ver contenido fichero Mostrar contenido desde la primera línea

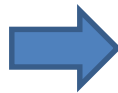
Vale Previsualizar filas Cancelar

Help

1ro ir a Examinar e indique el camino donde se encuentra el fichero. 2do ir a Añadir, vea como aparece en la lista

# Pestaña Contenido

OJO  
especificar el  
separador de  
campos este  
varía de un  
fichero a  
otro, debe  
fijarse antes



Entrada Fichero de Texto

Nombre paso: Entrada Fichero de Texto

Fichero | Contenido | Manejo de Errores | Filtros | Campos | Additional output fields

Tipo de fichero: CSV

Separador de campos: ;

Separador de texto: "

☐ ¿Eliminar saltos de línea en campos con separador de texto?

Escape:

Cabecera ☒ Número de líneas de cabecera: 1

Pie ☐ Número de líneas de pie: 1

¿Líneas cortadas? ☐ Número de veces que se corta: 1

Paginado (impresión)? ☐ Número de líneas por página: 80

Líneas en cabecera documento: 0

Comprimido (Zip): None

Eliminar filas vacías ☒

¿Incluir nombre del fichero en salida? ☐ Campo con el nombre del fichero:

¿Número de fila en salida? ☐ Campo con el número de fila:

¿Número de fila por fichero? ☐

Formato: DOS

Codificación:

Límite: 0

¿Ser flexible al leer fechas? ☒

Utilizar el formato de fecha local(e): es\_ES

Result filenames

Add filenames to result ☒

Help Vale Previsualizar filas Cancelar

# Pestaña Campos

Entrada Fichero de Texto

Nombre paso:

Fichero | Contenido | Manejo de Errores | Filtros | **Campos** | Additional output fields

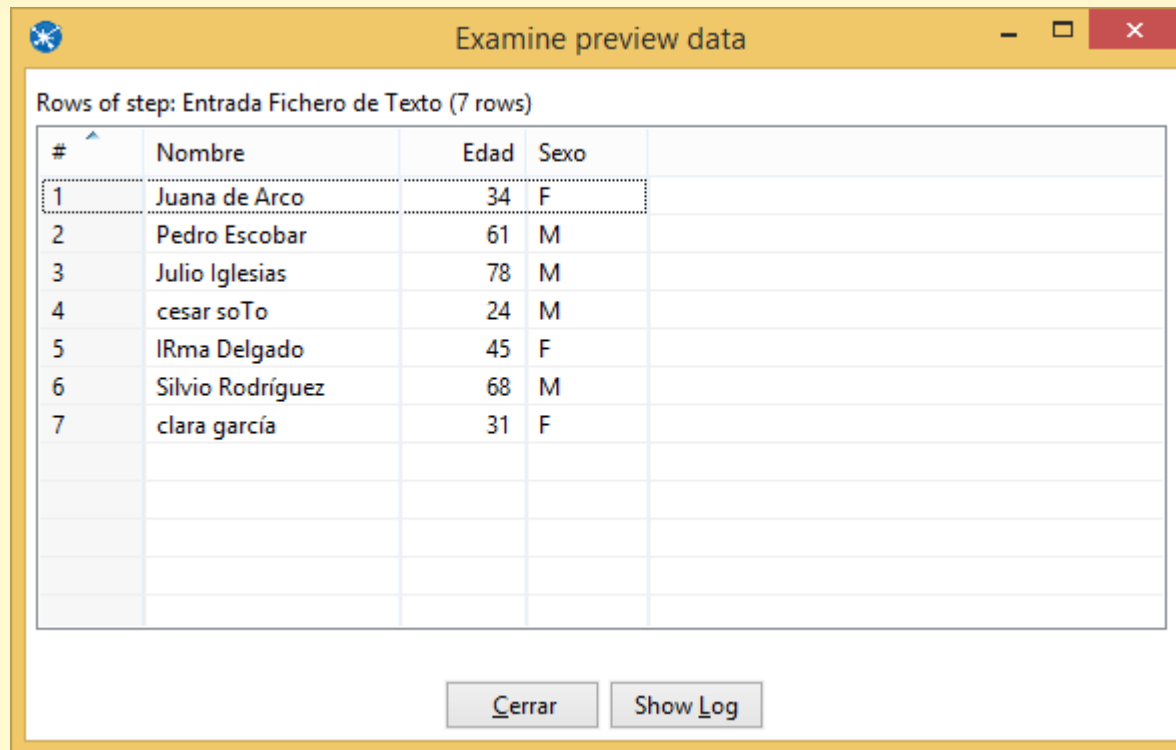
#	Nombre	Tipo	Formato	Posición	Longitud	Precisión	Moneda	Decimal	Grupo	Nulo si	Por defecto	Tipo de poda	Repetir
1	Nombre	String			16		€	,	.	-		ninguno	N
2	Edad	Number	#.#		15		€	.	,	-		ambos	N
3	Sexo	String			1		€	,	.	-		ninguno	N

1 → Ir a Campos

Vale Previsualizar filas Cancelar

Help

# Resultados de Previsualizar filas



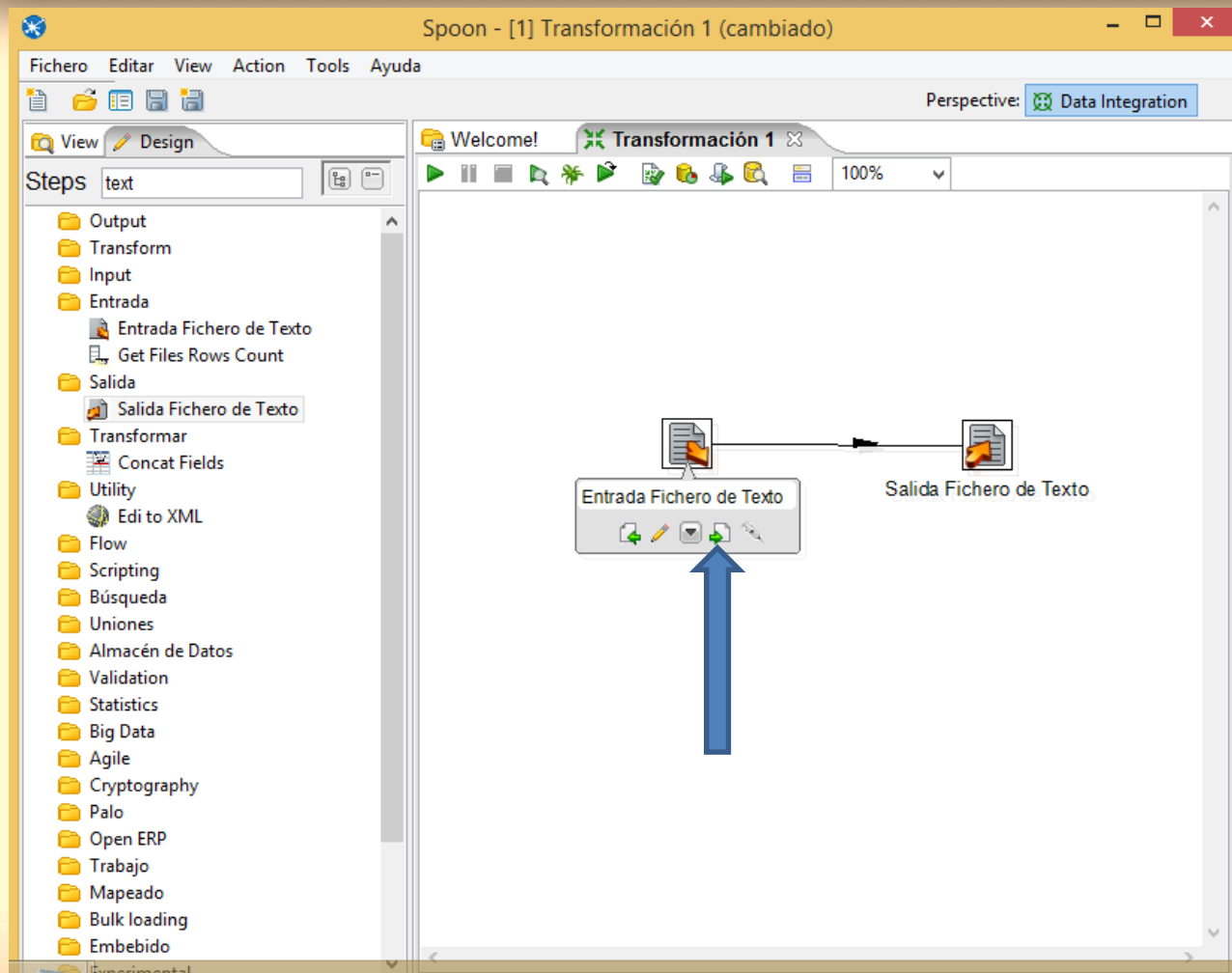
Rows of step: Entrada Fichero de Texto (7 rows)

#	Nombre	Edad	Sexo
1	Juana de Arco	34	F
2	Pedro Escobar	61	M
3	Julio Iglesias	78	M
4	cesar soTo	24	M
5	IRma Delgado	45	F
6	Silvio Rodríguez	68	M
7	clara garcía	31	F

Cerrar Show Log

En esta figura se les muestra el fichero original, o sea se está leyendo directamente de él, si lo desea puede comprobarlo manualmente.

# Agregando el hop para el flujo de trabajo



# Configuración del paso de salida texto

The screenshot shows a configuration window titled "Salida Fichero de Texto". It has three tabs: "Fichero", "Contenido", and "Campos", with "Fichero" currently selected. The "Nombre paso" field is set to "Salida Fichero de Texto". The "Nombre Fichero" field contains the path "C:\Users\Lisandra\Desktop\Salida\_Nombres" and has an "Examinar..." button next to it. Below this are several checkboxes: "¿Ejacular como comando?" (unchecked), "Pass output to servlet" (unchecked), "Create Parent folder" (checked), "Do not create file at start" (unchecked), and "Accept file name from field?" (unchecked). There is a "File name field" dropdown menu. The "Extensión" field is set to "txt". Below these are more checkboxes: "¿Incluir stepnr en nombre fichero?" (unchecked), "¿Incluir número partición en nombre fichero?" (unchecked), "¿Incluir fecha en nombre fichero?" (unchecked), "¿Incluir hora en nombre fichero?" (unchecked), and "Specify Date time format" (unchecked). There is a "Date time format" dropdown menu. A button labeled "Mostrar nombre fichero(s)..." is visible. At the bottom, the "Add filenames to result" checkbox is checked. The window has a "Help" button in the bottom left and "Vale" and "Cancelar" buttons in the bottom right.

Salida Fichero de Texto

Nombre paso: Salida Fichero de Texto

Fichero | Contenido | Campos

Nombre Fichero: C:\Users\Lisandra\Desktop\Salida\_Nombres Examinar...

¿Ejacular como comando? ☐

Pass output to servlet ☐

Create Parent folder ☒

Do not create file at start ☐

Accept file name from field? ☐

File name field: [dropdown]

Extensión: txt

¿Incluir stepnr en nombre fichero? ☐

¿Incluir número partición en nombre fichero? ☐

¿Incluir fecha en nombre fichero? ☐

¿Incluir hora en nombre fichero? ☐

Specify Date time format ☐

Date time format: [dropdown]

Mostrar nombre fichero(s)...

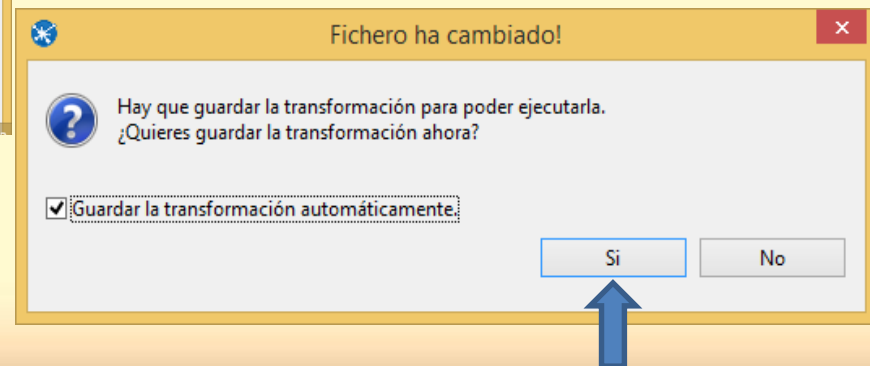
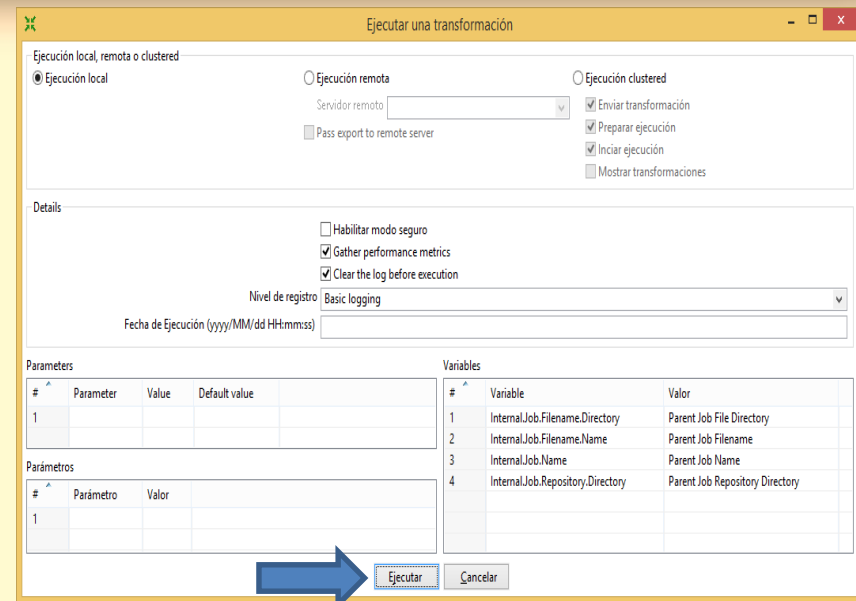
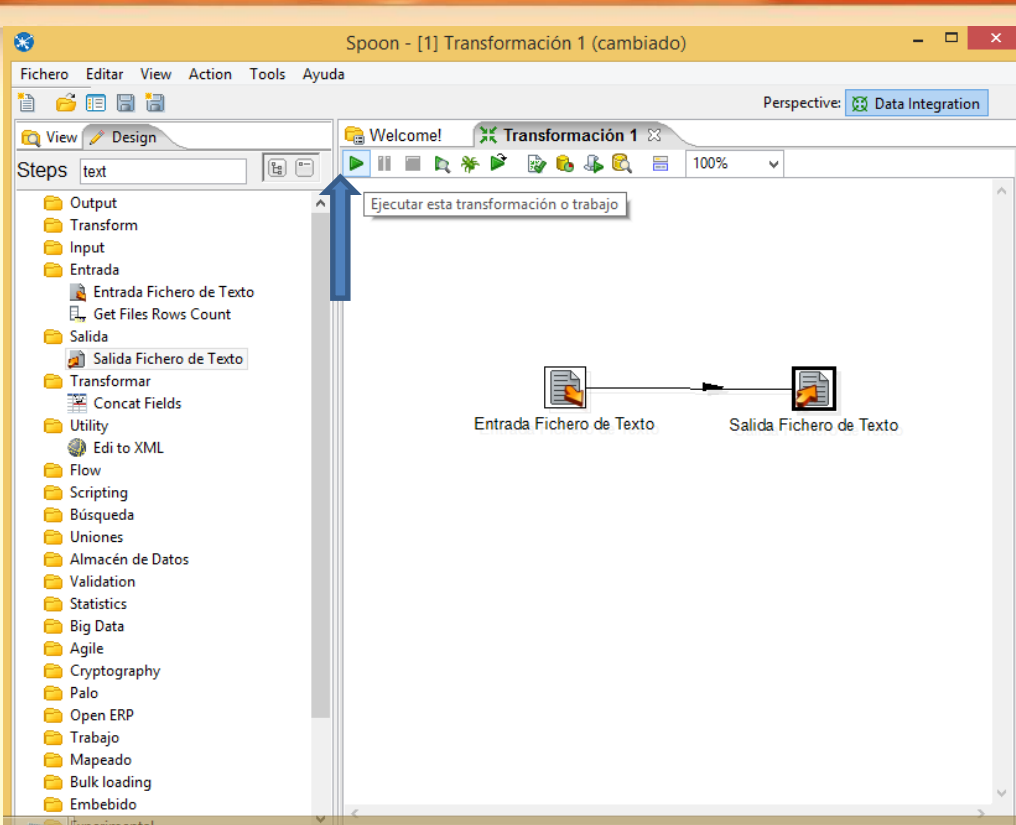
Add filenames to result ☒

Help Vale Cancelar

# Traer campos del flujo

[illegible]

# Ejecutar una transformación





# Resultados

The screenshot shows the Spoon IDE interface for a data integration transformation. The left sidebar contains a tree view of components, including Output, Transform, Input, Entrada, Salida, and various utilities. The main canvas shows a flow from 'Entrada Fichero de Texto' to 'Salida Fichero de Texto'. The bottom panel displays 'Execution Results' with a table showing step metrics.

#	Nombre paso	Numero Copia	Leído	Escrito	Ent
1	Entrada Fichero de Texto	0	0	7	
2	Salida Fichero de Texto	0	7	7	

# Arreglando los nombres

- El ejemplo resuelto es muy sencillo hasta ahora solo se han movido datos de un fichero a otro pero el fichero de entrada tiene nombres que no están escritos con la letra inicial mayúscula, podemos estandarizarlo utilizando el paso *String Operations*.
- Para agregar un paso a la transformación anterior una vez que arrastren el paso hacia el área de trabajo, lo ponen encima del hop cuando este se ponga en negrita entonces lo sueltan y dan Aceptar.

# Configuración del paso String Operations

1ro debe obtener los campos que vienen del flujo, recuerde que debe existir una flecha de entrada hacia el paso, hacer clic en *Get fields*.

2do se obtienen todos los campos tipo *string* del flujo.

3ro se configuran las opciones disponibles, por ejemplo *Trim Type* tiene 3 opciones *left*, *right* y *both*. Significa que puede eliminar los espacios de la izquierda, los de la derecha o ambos, igual al método *trim* de Java. *Lower/Upper* es para convertir toda la cadena a minúsculas o mayúsculas, *InitCap* es para que ponga la letra inicial mayúscula en cada palabra, etc.

#	In stream field	Out stream field	Trim type	Lower/Upper	Padding	Pad char	Pad Length	InitCap	Escape	Digits	Remove Special character
1	Nombre		both	lower	none			S	None	remove	none
2	Sexo		both	upper	none			N	None	none	none

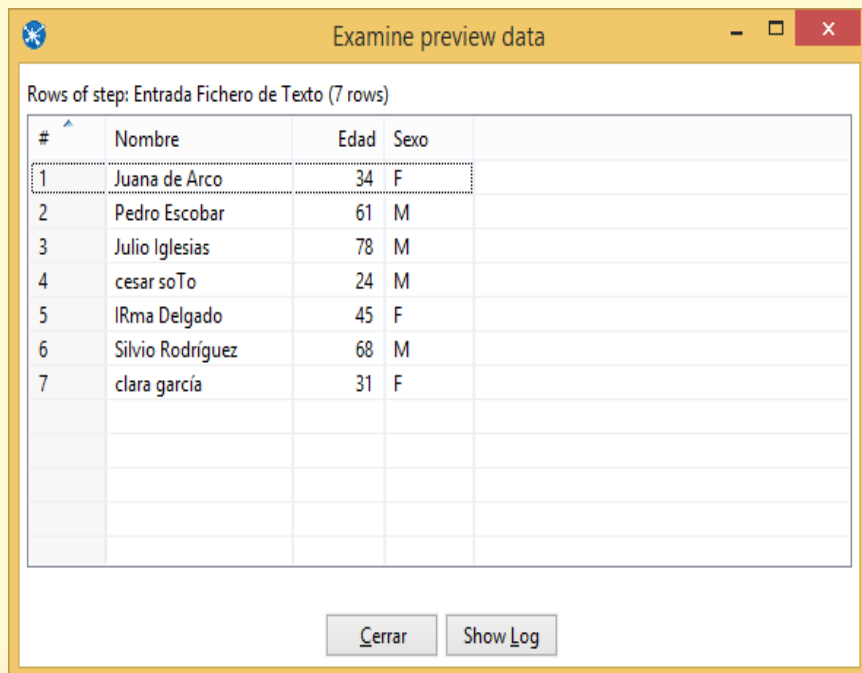
< >

Vale Get fields Cancelar

# Resultados – Ejemplo sencillo de limpieza de datos

Cuando se vuelve a ejecutar la transformación, el resultado es el siguiente. Nótese el cambio en los nombres.

Fichero texto original

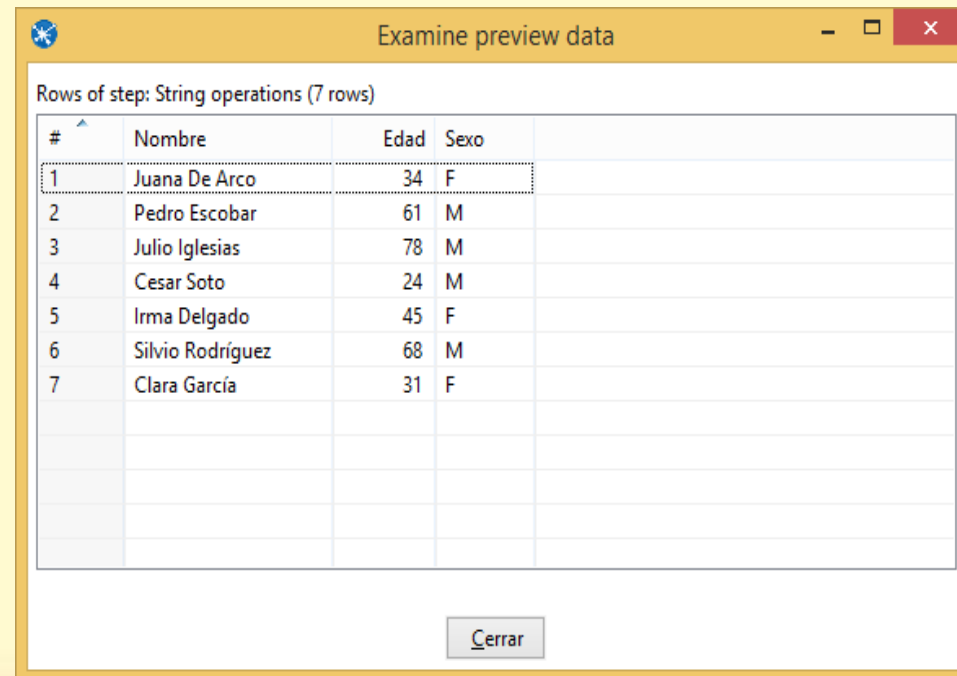


Rows of step: Entrada Fichero de Texto (7 rows)

#	Nombre	Edad	Sexo
1	Juana de Arco	34	F
2	Pedro Escobar	61	M
3	Julio Iglesias	78	M
4	cesar soTo	24	M
5	IRma Delgado	45	F
6	Silvio Rodríguez	68	M
7	clara garcía	31	F

Cerrar Show Log

Fichero texto con transformaciones

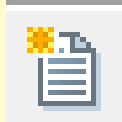


Rows of step: String operations (7 rows)

#	Nombre	Edad	Sexo
1	Juana De Arco	34	F
2	Pedro Escobar	61	M
3	Julio Iglesias	78	M
4	Cesar Soto	24	M
5	Irma Delgado	45	F
6	Silvio Rodríguez	68	M
7	Clara García	31	F

Cerrar

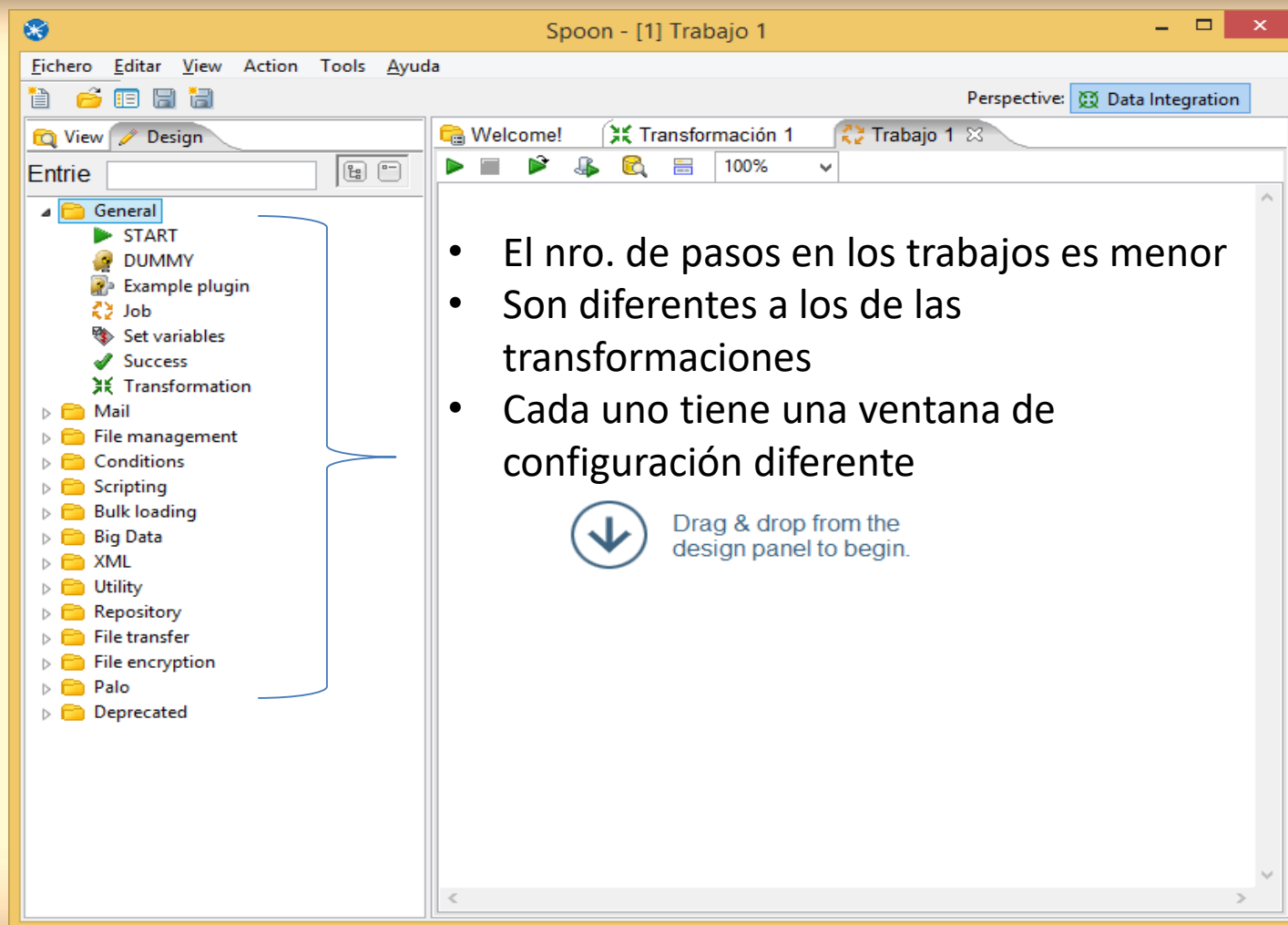
# Vías para crear un trabajo

1. Ir al menú File -> New -> Job
2. Apretar Ctrl+Alt+N
3. Hacer clic en el ícono  y elegir Job

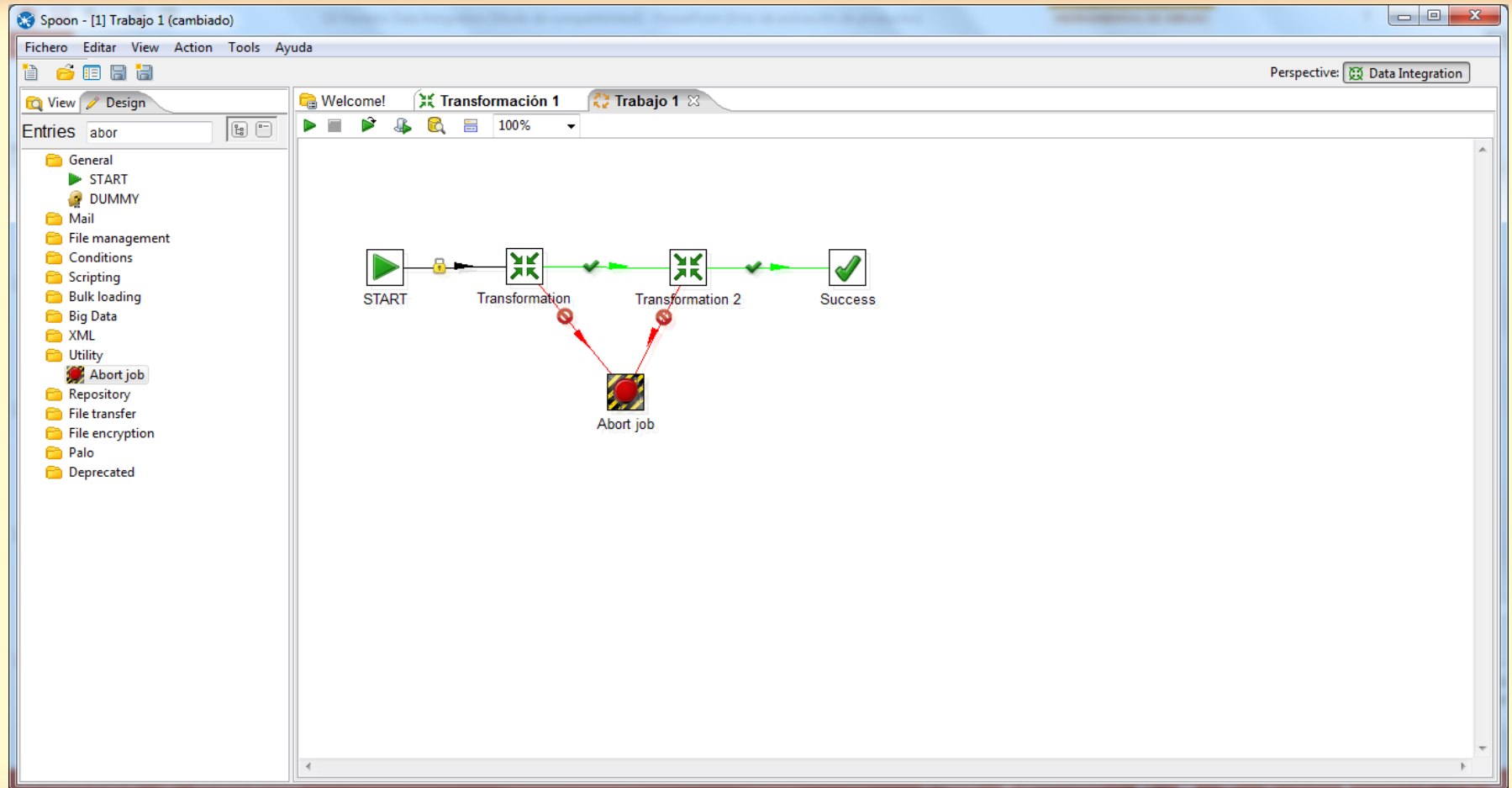
Los trabajos se utilizan principalmente para orquestar las transformaciones, o sea se tienen creadas varias transformaciones y con un trabajo se da un orden lógico de ejecución a modo de flujo de trabajo. Esto permite ejecutar varias transformaciones de una sola vez. Muy utilizado en los almacenes de datos.

Recuerde guardar el trabajo de la misma manera que lo hizo con las transformaciones, pero eligiendo la carpeta de los trabajos.

# Pasos en los trabajos



# Creando un nuevo trabajo



# Configuración de los pasos en el trabajo 1

- Los pasos de inicio (START), Success y Abort Job, no se configuran.
- En el paso Transformation deben especificar la ruta de las transformaciones previamente creadas que desean ejecutar en un determinado orden (puede escoger cualquier variante).

Job entry details for this transformation:

Name of job entry:

Transformation specification | Advanced | Logging settings | Argument | Parameters

☐ Transformation filename:

☒ Specify by name and directory

☐ Specify by reference

New transformation

Help Vale Cancelar



# Ejecutar un trabajo

- El trabajo se ejecuta de igual manera que una transformación, haciendo clic en la flecha verde que indica *Run this job*.
- Se acepta guardar los cambios y listo.
- Si todo está bien configurado se marca con una flecha verde, en caso de error se marca con un círculo rojo y al final de la ventana *Execution results* en la pestaña *Logging* podrán leer los errores en idioma inglés.

# Ejemplo de una ejecución con errores

The screenshot shows the Spoon - [1] Trabajo 1 window. The job flow is: START → Transformación 1 → Success. Transformación 1 is highlighted with a red error icon. The Execution results pane at the bottom shows the following log:

```
2021/03/15 12:10:30 - Trabajo 1 - at org.pentaho.di.core.vfs.KettleVFS.getFileObject(KettleVFS.java:2797)
2021/03/15 12:10:30 - Trabajo 1 - at org.pentaho.di.trans.TransMeta.<init>(TransMeta.java:2797)
2021/03/15 12:10:30 - Trabajo 1 - at org.pentaho.di.trans.TransMeta.<init>(TransMeta.java:2774)
2021/03/15 12:10:30 - Trabajo 1 - at org.pentaho.di.trans.TransMeta.<init>(TransMeta.java:2759)
2021/03/15 12:10:30 - Trabajo 1 - at org.pentaho.di.job.entries.trans.JobEntryTrans.getTransMeta()
2021/03/15 12:10:30 - Trabajo 1 - ... 5 more
2021/03/15 12:10:30 - Spoon - Trabajo ha terminado.
```

De igual manera se muestran los errores en las transformaciones. Debe desplazarse hacia arriba y hacia la derecha para comprender mejor el o los errores.

# Bibliografía

- The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004, R.Kimball and J.Caserta.
- Pentaho Data Integration 4 Cookbook, Packt Publishing, 2011.
- Pentaho Solution, Wiley Publishing, Inc, 2009.
- Enterprise Applications Integration with XML and Java. Prentice Hall PTR; Bk&CD Rom edition, 2000, J.P.Morgenthal.
- Enterprise Information Integration: A Pragmatic Approach. Lulu.com, 2005, J.P.Morgenthal.

# Estudio independiente

- Estudiar por el Pentaho Solution varios pasos y su configuración.
- Realizar la guía de ejercicios de la preparación previa sobre ETL con el PDI (Spoon).