Taller sobre integración de datos (abiertos)

Uso de Pentaho Data Integration

Jose Norberto Mazón
Twitter: @jnmazon

Grupo de investigación WaKe
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

lsi Departamento de Lenguajes y Sistemas Informáticos

wake Web and Knowledge Group
Universidad de Alicante

Datos abiertos

- Datos libremente **accesibles y reutilizables**
  - Única condición de la **atribución**
- Características **legales**
  - **Licencia**, protección de datos, etc.
- Características **tecnológicas**
  - Formato, calidad, etc.

Licencias

- Condiciones que regulan el uso de los datos abiertos
- Licencias recomendadas por OKFN
  - Creative Commons Attribution 4.0
    - Reconocer explícitamente al autor
  - ODC Open Database License (ODbL)
    - Posibilidad de tener una versión cerrada de los datos siempre y cuando se sumisnistre una versión abierta
- También se pueden crear licencias ad-hoc
  - Por ejemplo, Ayuntamiento de Zaragoza
  - http://www.zaragoza.es/ciudad/servicios/avisolegal.htm#condiciones

Licencias

- El creador del material **cede algunos de sus derechos** a terceras personas bajo ciertas condiciones
  - **Reconocimiento** (attribution): reconocer explícitamente al autor
  - **No comercial** (non commercial): no hacer usos comerciales.
  - **Sin obra derivada** (no derivate works): prohíbe la modificación del material.
  - **Compartir igual** (share alike): la obra creada a partir de la original, debe tener la misma licencia Creative Commons

Licencias

Reconocimiento (by): Se permite cualquier explotación de la obra, incluyendo una finalidad comercial, así como la creación de obras derivadas, la distribución de las cuales también está permitida sin ninguna restricción.

Reconocimiento - NoComercial (by-nc): Se permite la generación de obras derivadas siempre que no se haga un uso comercial. Tampoco se puede utilizar la obra original con finalidades comerciales.

Reconocimiento - NoComercial - CompartirIgual (by-nc-sa): No se permite un uso comercial de la obra original ni de las posibles obras derivadas, la distribución de las cuales se debe hacer con una licencia igual a la que regula la obra original.

Reconocimiento - NoComercial - SinObraDerivada (by-nc-nd): No se permite un uso comercial de la obra original ni la generación de obras derivadas.

Reconocimiento - CompartirIgual (by-sa): Se permite el uso comercial de la obra y de las posibles obras derivadas, la distribución de las cuales se debe hacer con una licencia igual a la que regula la obra original.

Reconocimiento - SinObraDerivada (by-nd): Se permite el uso comercial de la obra pero no la generación de obras derivadas.

Formatos

- Favorecer la **reutilización**

★ Los datos están disponibles en la Web, independientemente del formato utilizado.

★★ Los datos se publican en la Web en un formato estructurado.

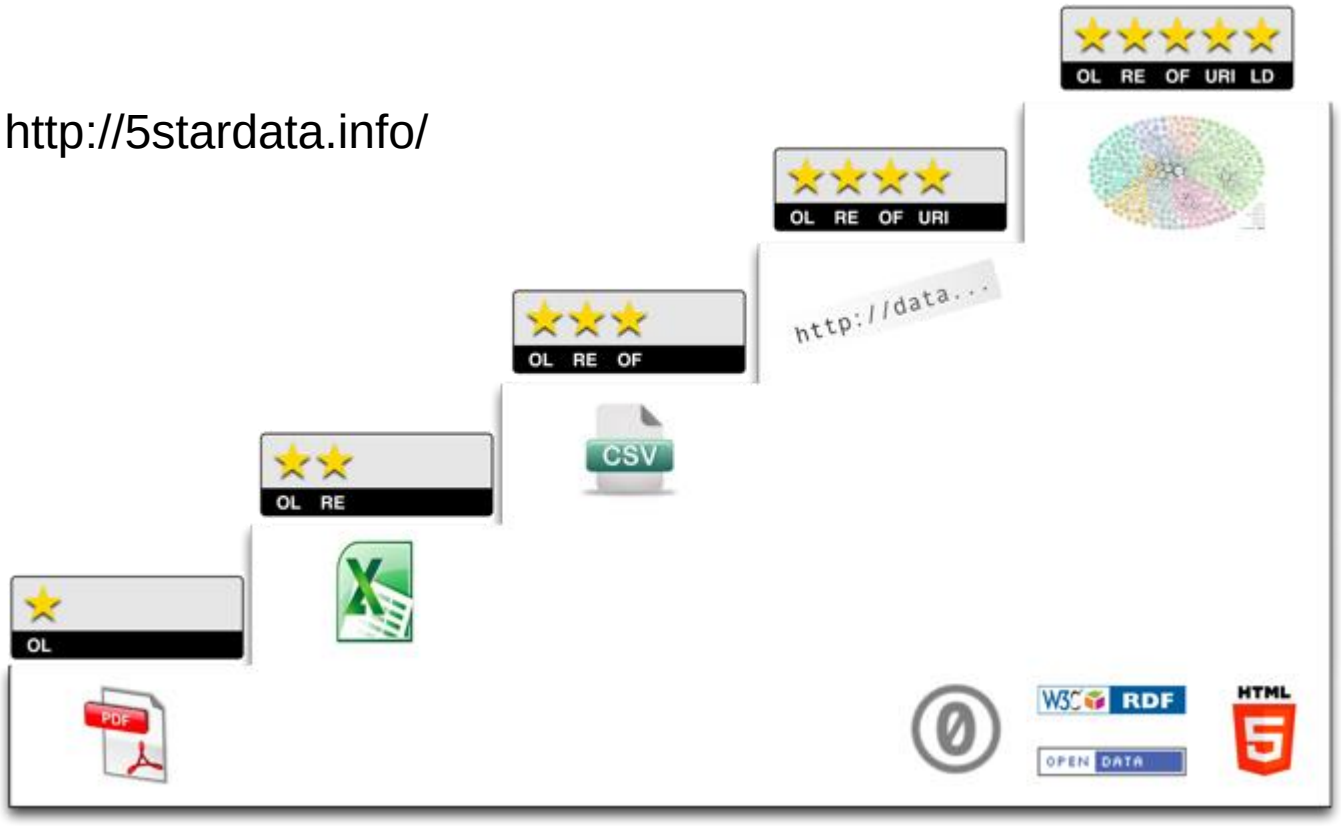
★★★ Los datos están publicados bajo un formato no propietario.

★★★★ Los datos se identifican mediante URLs de manera que sean fácilmente interpretables.

★★★★★ Los datos están vinculados con otros datos de manera que se encuentran contextualizados.

Formatos

http://5stardata.info/



Datos como materia prima

Datos como materia prima

- http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
- http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide.html

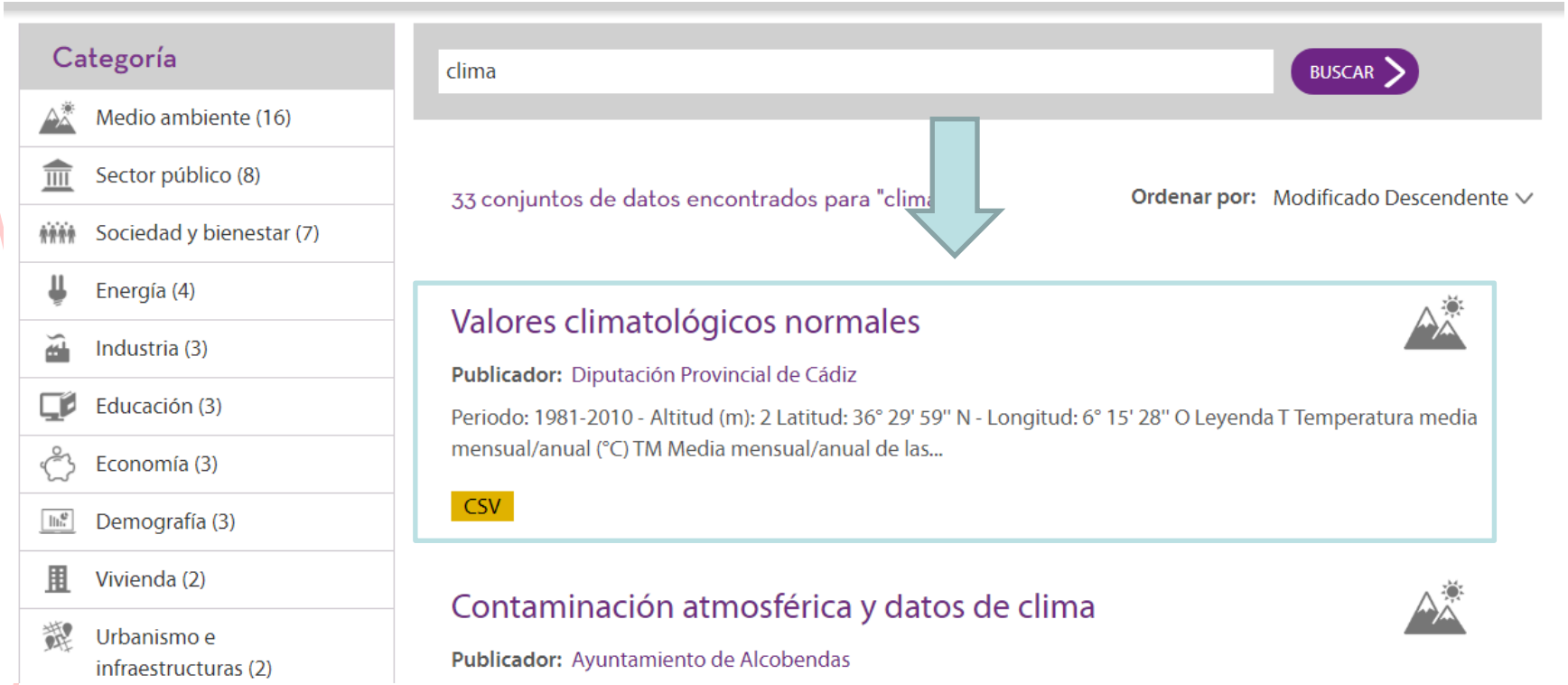
Portal de datos abiertos

- Sitio web donde una organización publicará todos sus datos
  - **Facilidad** para que terceras personas consulten y reutilicen los datos
    - Infomediarios y ciudadanía en general
  - **Enfocados en el dato** como unidad principal de interacción

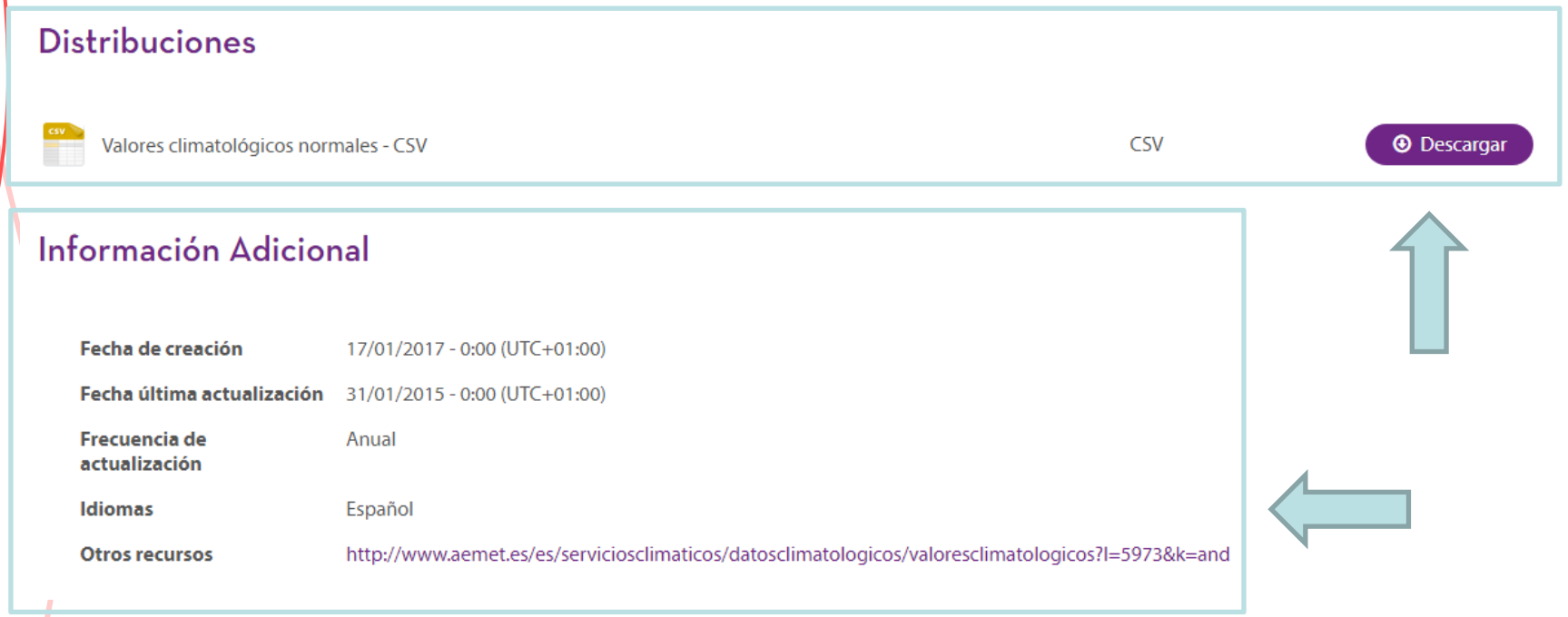
Portales de datos abiertos

Portales de datos abiertos

Portales de datos abiertos



Portales de datos abiertos



Portales de datos abiertos

- Plataformas de publicación
  - Data Management System (DMS)
- CKAN - https://ckan.org/
  - Código abierto (proyecto en Github)
  - PostgreSQL & Python

- Socrata - https://socrata.com/
  - Publica Open Data
  - Potente API – SODA (Socrata Open Data API)

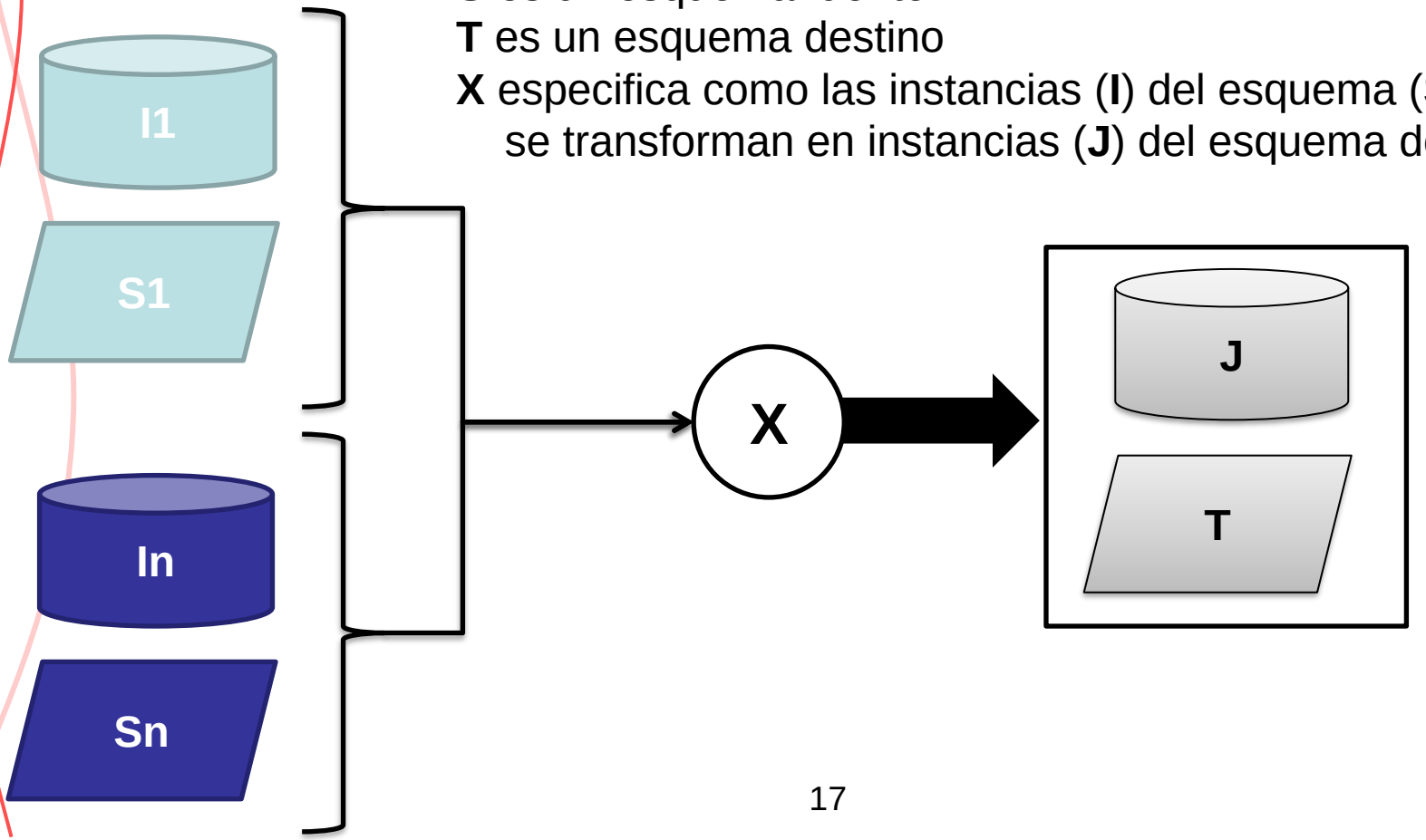
Ejemplo de uso de CKAN API

- Conjunto de datos en el catálogo
  - http://demo.ckan.org/api/3/action/package_list
- Obtener un conjunto de datos en formato JSON
  - http://demo.ckan.org/api/3/action/package_show?id=adur_district_spending
- Buscar un conjunto de datos según palabras clave
  - http://demo.ckan.org/api/3/action/package_search?q=spending

Integración de datos

- Transformar las fuentes de datos en un destino de datos

Transformación de datos
S es un esquema fuente
T es un esquema destino
X especifica como las instancias (**I**) del esquema (**S**)
se transforman en instancias (**J**) del esquema destino (**T**)



Integración de datos

- Acceso uniforme a fuentes de datos heterogéneas
  - Diferentes **formatos**
    - CSV vs base de datos relacional vs XML vs JSON …
  - Diferentes **tecnologías**
    - Oracle vs SQL Server vs SQLite …
  - Diferentes **accesos**
    - Servicios web vs JDBC …
  - Diferentes **esquemas**
    - Asignatura(código, nombre, titulación)

      Titulación (id_titulación, nombre)
    - Asignatura (id_asignatura, nombre_asignatura, id_titulación, nombre_titulación)

Formatos de datos

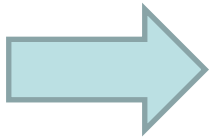
- Archivo **CSV** (Comma Separated Values)
  - Formato para representar datos en forma de tabla en un **fichero de texto**
  - Cada línea en el fichero es una **fila de datos**
  - Cada valor se separa por comas/puntos y comas/otro símbolo representando **columnas**

```
Pepe,34,03181
María,32,03690
Ana,45,03080
```

Integración de datos
Formatos

- Archivo **CSV** (Comma Separated Values)
  - La primera fila puede contener el nombre de las columnas

```
Pepe,34,03181
María,32,03690
Ana,45,03080
```



```
Nombre,Edad,CP
Pepe,34,03181
María,32,03690
Ana,45,03080
```

Integración de datos
Formatos

- Archivo **CSV** (Comma Separated Values)
  - Si los valores contienen comas, se usa un delimitador, p.e. comillas

```
“Nombre”,”Edad”,”CP”,”Dirección”
“Pepe”,”34”,”03181”,”Gran Vía, 16”
“María”,”32”,”03690”,”Plaza Mayor  8”
“Ana”,”45”,”03080”,”Gran Vía, 45, 2ºB”
```

Integración de datos
Formatos

- **XML** (eXtensible Markup Language)
  - Lenguaje de etiquetas utilizado para almacenar datos de forma estructurada (legible por máquinas)
  - Estándar para el intercambio de información estructurada entre diferentes plataformas.
    - Se puede usar en bases de datos, editores de texto, hojas de cálculo y casi cualquier cosa imaginable

Integración de datos
Formatos

- **XML** (eXtensible Markup Language)
  - Separación de contenido y maquetación
  - Las etiquetas representan significado del contenido pero no la maquetación

    <titulo>El Quijote</titulo>

    <autor>Cervantes>/autor>
  - Estructura en árbol

    <libro>

        <titulo> El Quijote</titulo>

        <autor>Cervantes</autor>

    </libro>

Integración de datos
Formatos

- **HTML** (HiperText Markup Language)
  - XML usado para crear páginas Web
  - Los significados de las etiquetas se refieren a las partes de un sitio Web
    - Párrafo
    - Título
    - Tabla
    - Enlace
    - etc.

 Hola mundo!

 Hola mundo!

Integración de datos
Formatos

- **JSON** (JavaScript Object Notation)
  - Mismo propósito que XML
    - Intercambio de datos
  - Pesa menos

```
{
    libro:
    {
      titulo: "El Quijote",
      autor: "Cervantes"
    }
}
```

Integración de datos

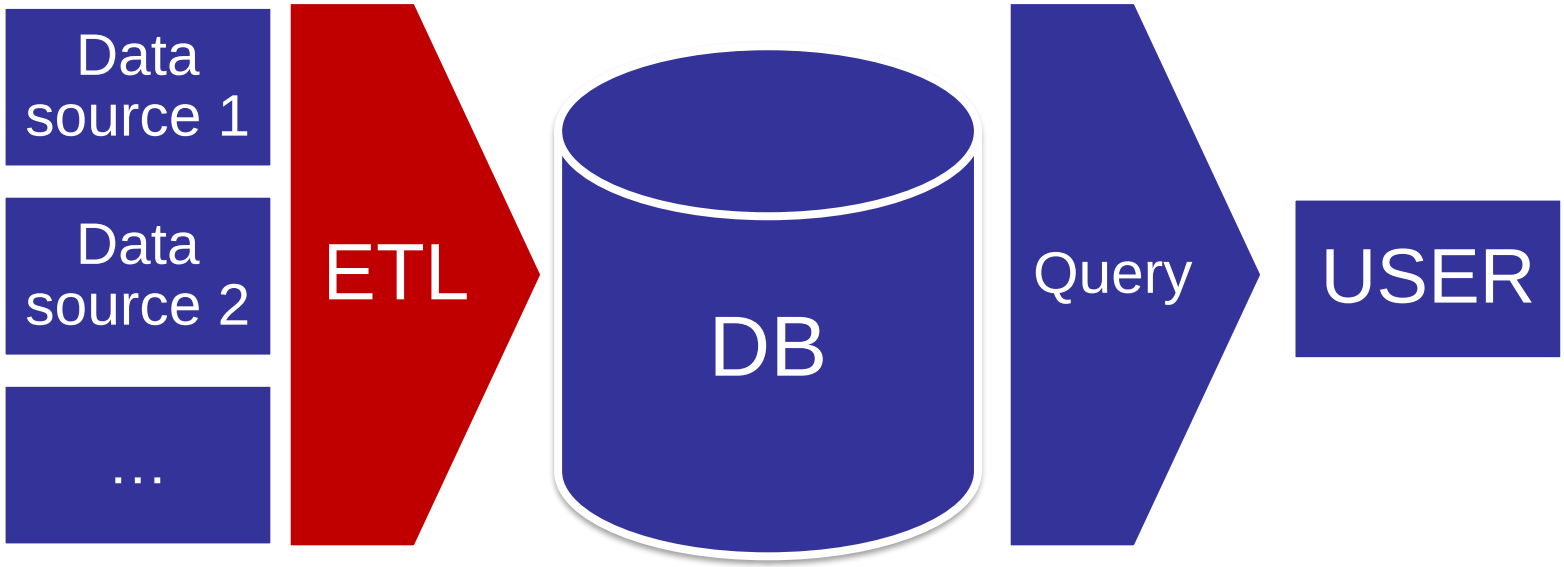
- **Calidad** de datos
  - **Limpieza** de datos
    - Generación de claves
      1. Masculino, Femenino
      2. 0, 1
      3. Hombre, Mujer

      H, M
    - Conversión
      - Fechas
      - Unidades de medida
      - Etc.
    - Normalización
      - C/ Vicente Blasco Ibáñez 18
      - Calle Blasco Ibáñez nº 18
      - Blasco Ibanez 18

Calle	Número
Vicente Blasco Ibáñez	18
  - Filtrado, Unión, etc.

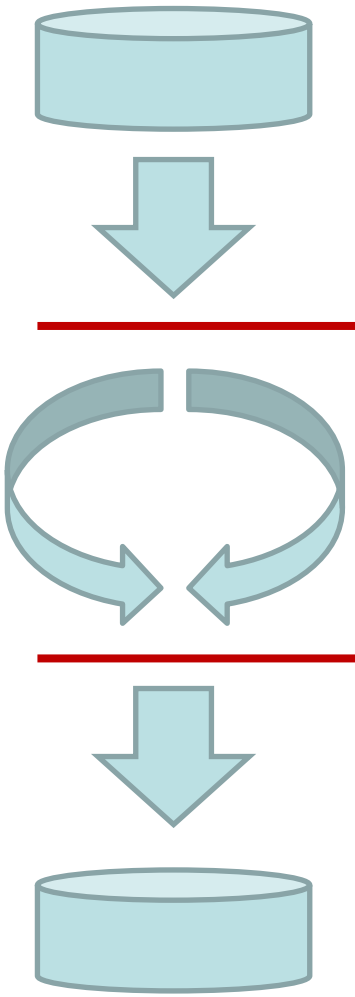
Integración de datos

- Procesos **ETL**
  - Extraction / Transformation/ Load
  - **Transformaciones** que preparan datos para una tarea concreta (e.g. solución *business intelligence*)



Integración de datos

Etapas



EXTRAER

Recolectar datos de diferentes fuentes de datos

TRANSFORMAR

Modificar datos (limpiar, agregar, enriquecer, etc.)

CARGAR

Almacenar datos

¿Cómo especificar transformaciones de datos?

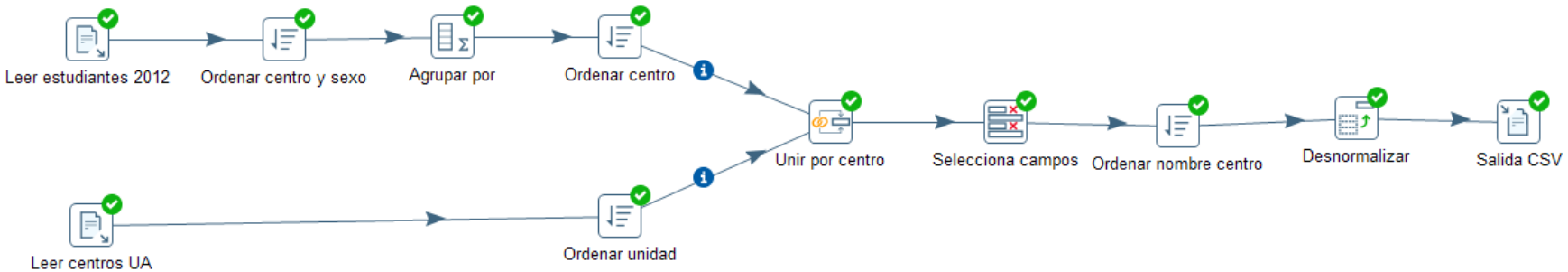
- Directamente programando (código)
  - SQL, Java, Pig, etc.

```
SELECT price.col1 AS col1, price.col2 AS col2 , price.col3 AS col3, MAX(price.col4) AS col4, MAX(price.col
MAX(price.col6) AS col6, MAX(price.col7) AS col7

FROM  table_1 t1, table_2 t2 WHERE col1 = col2 AND column_1 = small_column AND column_3411 <= column_12_su
'Test Run' AND column_4532 = c1.dert UNION SELECT price.col1 AS col1, price.col2 AS col2 , price.col3 AS c
(price.col4) AS col4, MAX(price.col5) AS col5, MAX(price.col6) AS col6, MAX(price.col7) AS col7 FROM  (SEL
store.column1, CAST (store.column2 AS INTEGER) AS column2, store.columnwe34r3 AS column3, store.column4_pr
store.column5_pre_prod_first AS column5 , SUBSTR(store.column6,11,1) AS column6, store.column7 AS column7

FROM  (SELECT library.column1, library.column2, library.column3 , CASE library.column4 WHEN cheap THEN dig
(library.column27) concat library.column28 ELSE 123456 END AS column4, CASE library.column5 WHEN expensive
(library.column27) concat library.column28 ELSE 123456 END AS library.column6, CASE column7 WHEN free THEN
(library.column27) concat library.column28 ELSE 123456 END AS column7, FROM
```

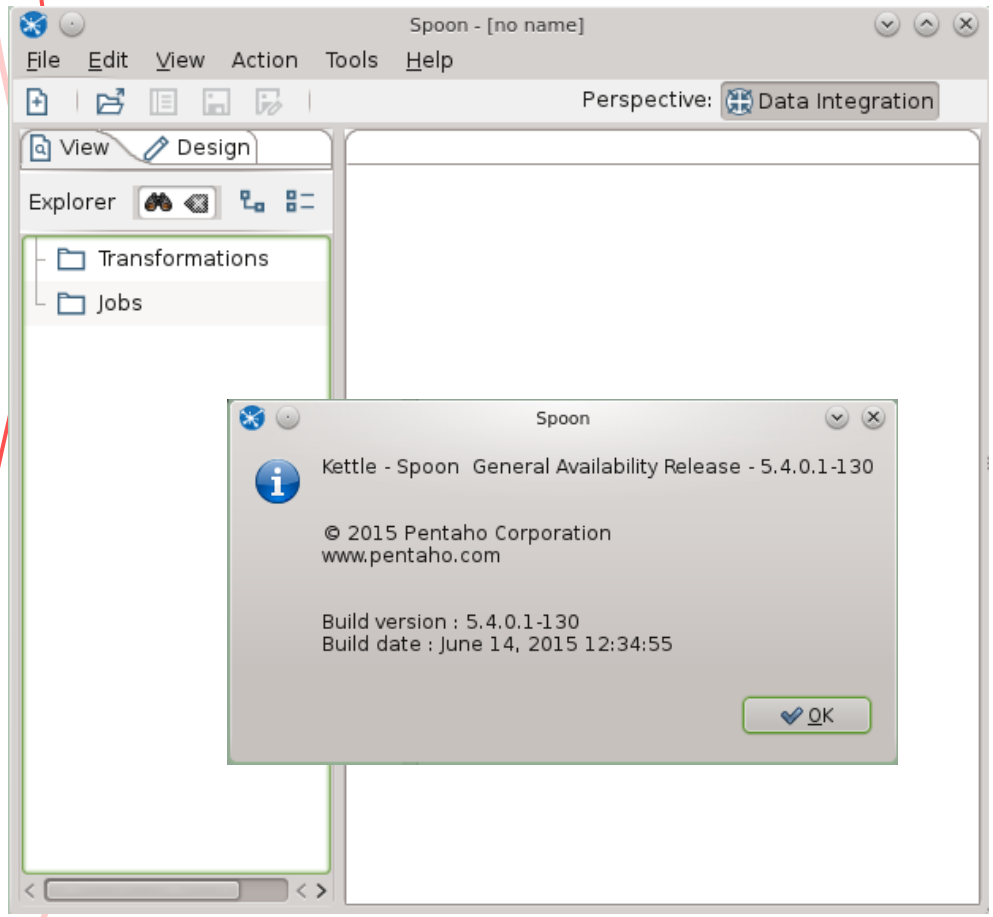
- Modelos mediante interfaz visual
  - El código se genera a partir del modelo



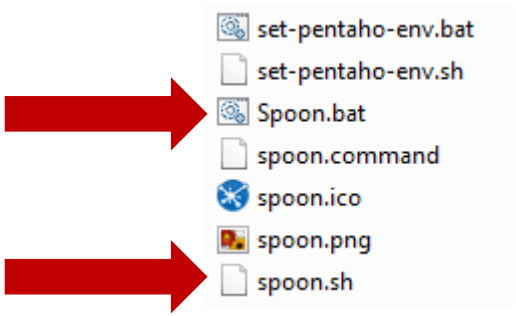
Pentaho Data Integration

- AKA Pentaho Kettle
- https://community.hitachivantara.com/s/article/data-integration-kettle
- Conjunto de herramientas para diseñar transformaciones para integración de datos
  - Editor gráfico para modelar transformaciones y trabajos (**Spoon**)
  - Ejecución de transformaciones vía línea de comandos (**Pan**)
  - Ejecución vía servidor (**Carte**)
  - Ejecución de trabajos vía línea de comandos (**Kitchen**)

Pentaho Data Integration



- Ejecutar
  - **Spoon.bat**
  - **spoon.sh**

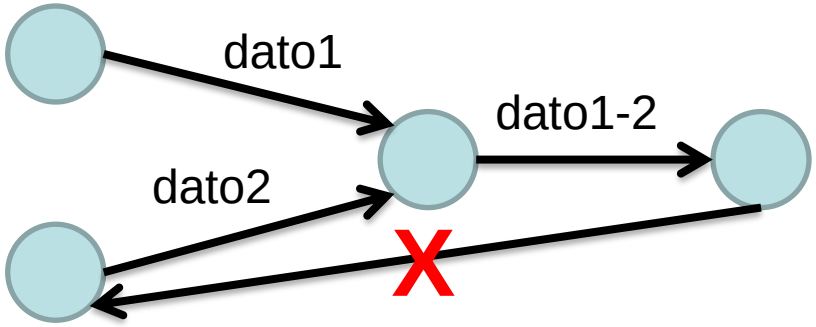


Pentaho Data Integration

- Transformación
  - **Transformation**
  - Conjunto de **pasos**
    - Ejecución secuencial
    - *Row-oriented*
- Trabajo
  - **Job**
  - Conjunto de **transformaciones**
    - Gestión
    - Manejo de errores

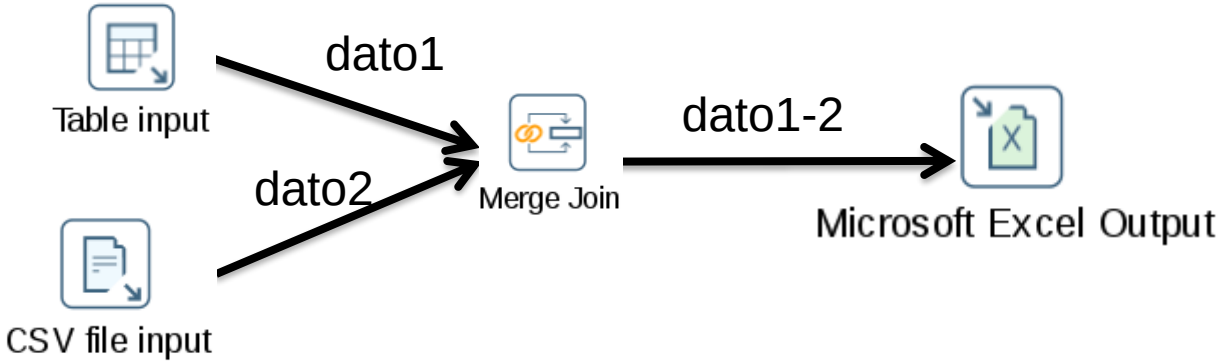
Pentaho Data Integration
Transformaciones

- **Grafo acíclico dirigido**
  - Sin ciclos
  - Nodos → pasos (**steps**)
    - Origen y destino
  - Aristas → saltos (**hops**)
    - Conforman un flujo de datos



Pentaho Data Integration
Transformaciones

- Pasos son **funciones** a realizar en los datos
  - Entrada y salida
    - Situar el puntero encima del paso
  - Edición de funcionalidad
    - Doble clic
  - Configuración de ejecución
    - Clic en botón derecho

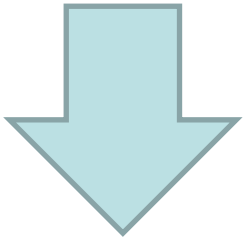


Pentaho Data Integration
Transformaciones

- Ejecución secuencial orientada a filas

“Nombre”,”Fecha nacimiento”,”CP”
“Pepe”,”12/08/1984”,”03181”
“María”,”13/04/1990”,”03690”
“Ana”,”24/02/1971”,”03080”

Fecha de nacimiento sólo debe contener el año



Fila 1: cabecera
Fila 2: datos
Fila 3: datos
Fila 4: datos

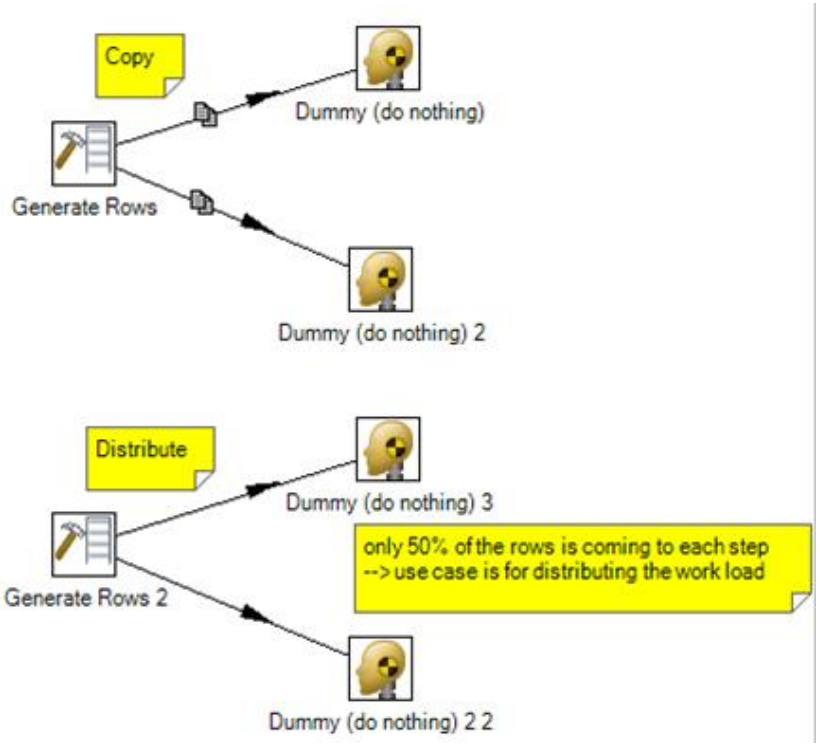
“Nombre”,”Fecha nacimiento”,”CP”
“Pepe”,”**1984**”,”03181”
“María”,”**1990**”,”03690”
“Ana”,”**1971**”,”03080”

Pentaho Data Integration
Transformaciones

- **Saltos**
  - **Conectan pasos** (origen y destino)
  - Permiten el flujo de datos y metadatos
  - Saltos determinan el flujo de datos
    - Cada paso se ejecuta en su propio hilo
    - La secuencia de ejecución la determina en propio PDI
  - Pueden habilitarse o deshabilitarse
    - Clic encima del salto

Pentaho Data Integration
Transformaciones

- Flujo de datos con **dos o más pasos destino**
  - **Copiar** todos los datos desde un paso hacia todos los siguientes pasos
  - **Distribuir** datos desde un paso hacia todos los siguientes pasos

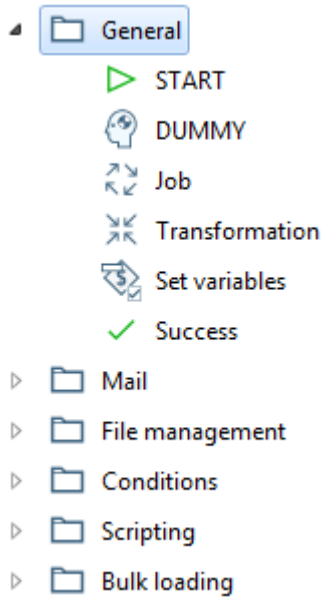


Pentaho Data Integration
Trabajos

- **Coordinación** de la ejecución de varias trasformaciones
  - Incluye **gestión de la ejecución** (por ejemplo, cuándo se realiza la ejecución)
- Adición de funcionalidad de **gestión**
  - Comprobar **condiciones previas**
    - Por ejemplo, existencia de determinada tabla en la base de datos origen o destino
  - Gestión de **logs**
  - Gestión de **errores**
    - Por ejemplo, envío de correo electrónico si ocurre un fallo

Pentaho Data Integration
Trabajos

- **Entradas** de un trabajo
  - Son las **partes elementales** de un trabajo
  - Proveen la **funcionalidad** del trabajo
    - Ejecutar una transformación, ejecutar otro trabajo, comprobar si existe algún recurso, enviar emails, etc.



Pentaho Data Integration
Trabajos

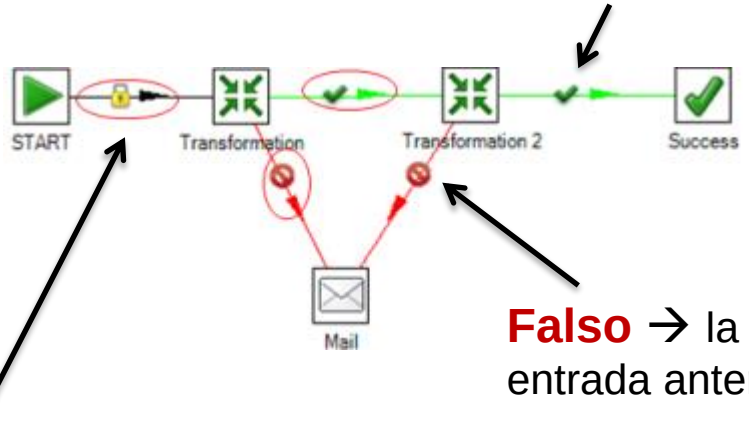
- **Saltos** en un trabajo
  - Son **flujos de control**, no de datos
    - Para pasar datos de una entrada de trabajo a otra entrada hay que usar la variable global resultado (**Result**)
    - Copia filas a resultado (**Copy Rows to Result**)
      - Paso que permite transferir filas de datos a la siguiente entrada de trabajo

    - Obtener filas de resultado anterior (**Get Files From Result**)
      - Paso que permite obtener datos de la variable resultados

Pentaho Data Integration
Trabajos

- Se debe indicar la **condición** bajo la cual se ejecuta la siguiente entrada del trabajo en dependencia del resultado de la entrada anterior
  - Clic en el salto del trabajo

Verdadero → la siguiente entrada se ejecuta sólo si la anterior termina correctamente



Falso → la siguiente entrada se ejecuta si la entrada anterior termina de manera errónea

Incondicional → la siguiente entrada siempre se ejecuta

Pentaho Data Integration
Transformaciones y trabajos

- Ficheros XML
  - Transformaciones con extensión **.ktr**
  - Trabajos con extensión **.kjb**

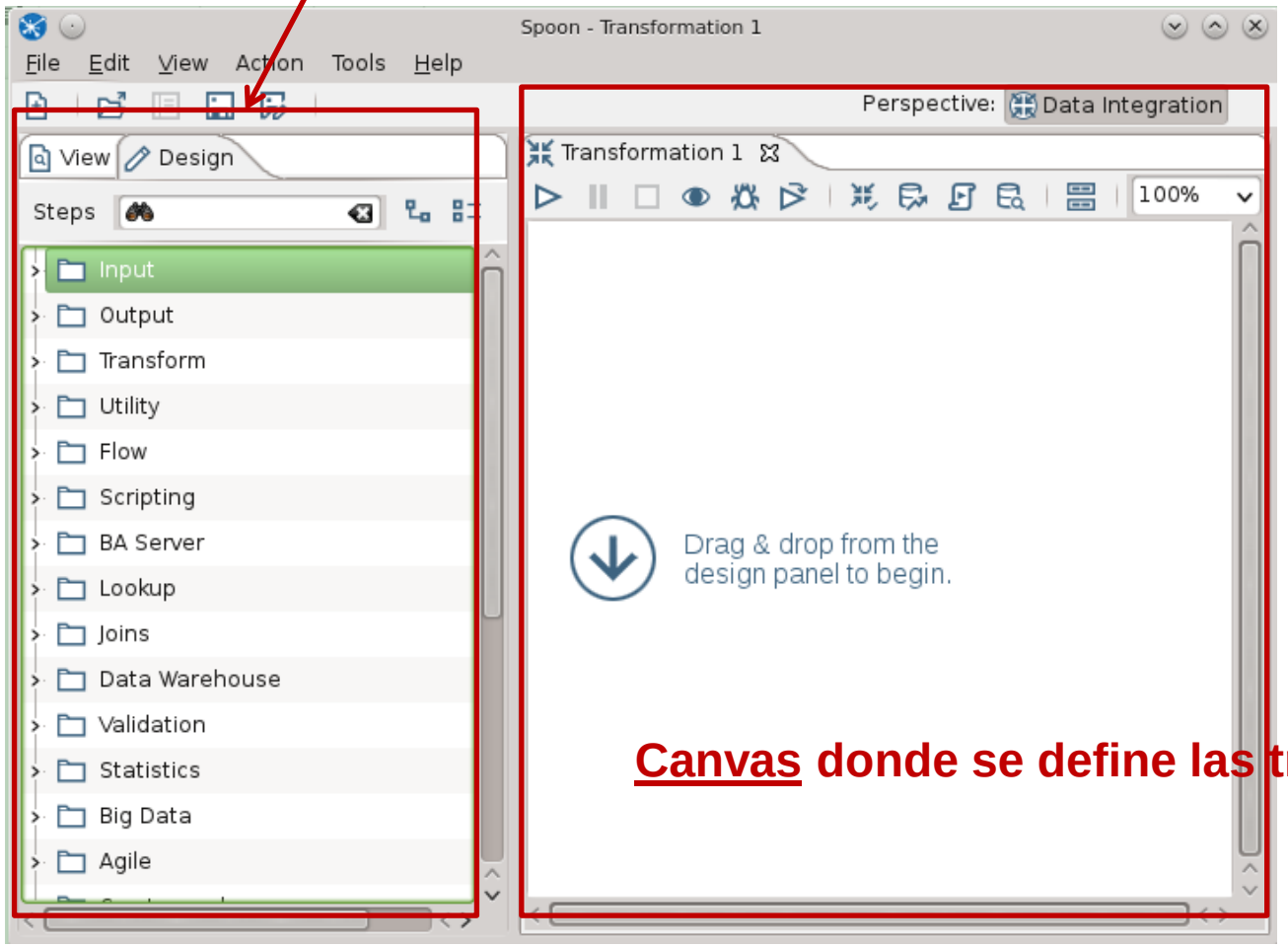
```
<order>
<hop> <from>Entrada Fichero de Texto</from><to>Json Input</to><enabled>Y</enabled> </hop>
<hop> <from>Json Input</from><to>Filtrar filas</to><enabled>N</enabled> </hop>
<hop> <from>Json Input</from><to>Filtrar filas 2</to><enabled>Y</enabled> </hop>
<hop> <from>Generar Filas</from><to>A&#xfffd;adir secuencia</to><enabled>Y</enabled> </hop>
<hop> <from>montar URL para obtener datos</from><to>obtener datos</to><enabled>N</enabled> </hop>
<hop> <from>A&#xfffd;adir secuencia</from><to>add 0</to><enabled>Y</enabled> </hop>
<hop> <from>add 0</from><to>montar URL para obtener datos</to><enabled>Y</enabled> </hop>
<hop> <from>Json Input</from><to>Filtrar filas 3</to><enabled>Y</enabled> </hop>
</order>
<step>
  <name>A&#xfffd;adir secuencia</name>
  <type>Sequence</type>
```

- Se genera código JAVA para su ejecución

Pentaho Data Integration

Interfaz de Spoon

Diseño donde se muestran los tipos de pasos



Canvas donde se define las transformaciones

Pentaho Data Integration
Interfaz de Spoon

- Los **pasos** de las transformaciones y las **entradas** de los trabajos se añaden al **canvas** mediante **drag & drop** desde la pestaña de diseño

- Se puede **abrir** haciendo **doble clic**
  - Aparece un **diálogo para parametrizar el paso** y obtener el comportamiento deseado

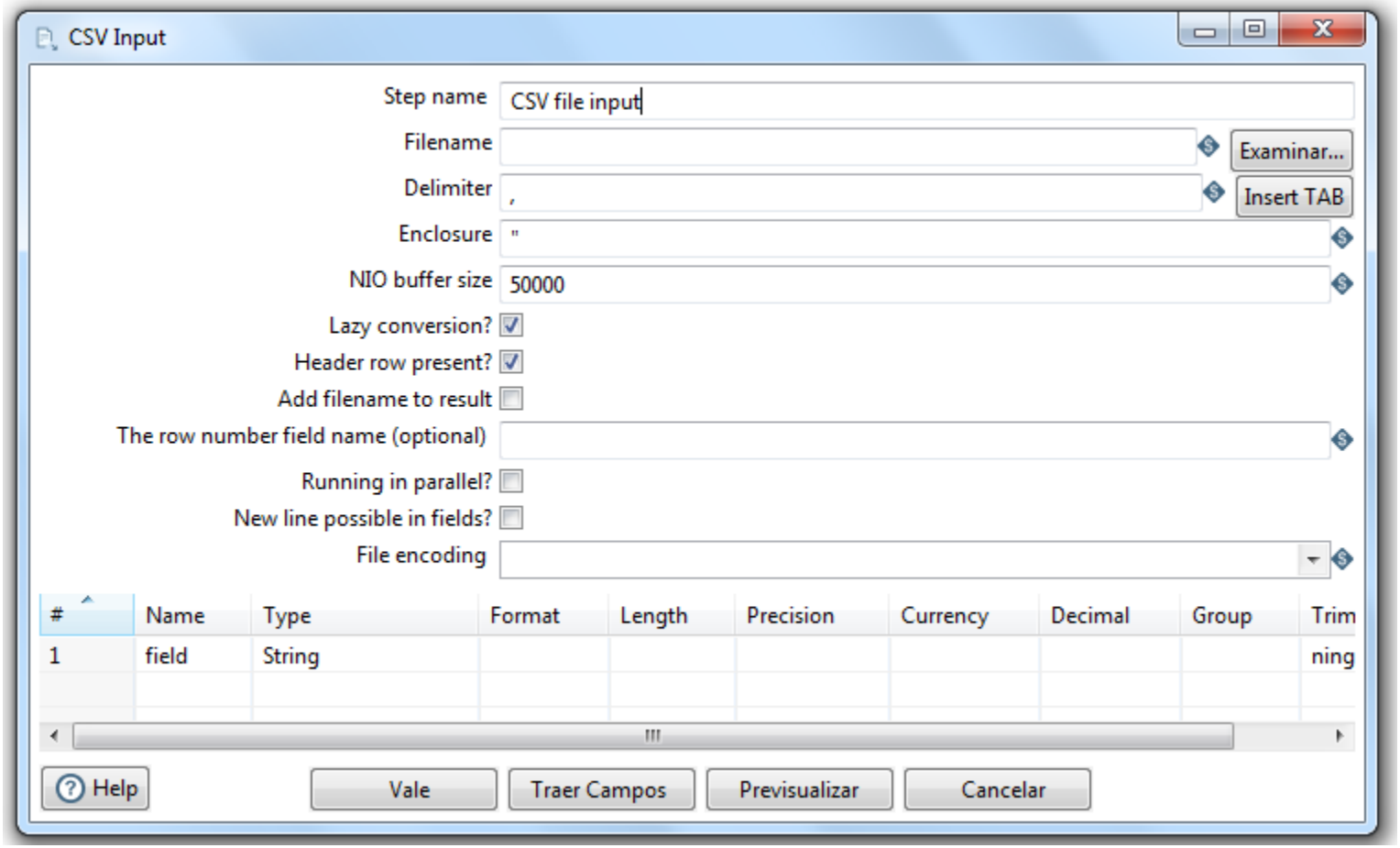
- También se puede **cambiar el nombre** del paso o entrada

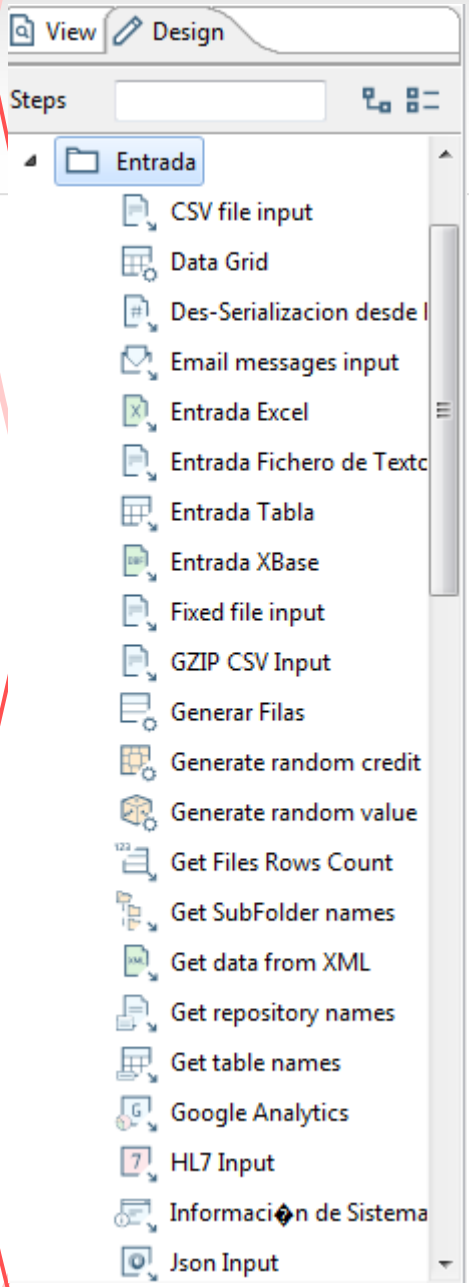
Pentaho Data Integration

Interfaz de Spoon

CSV file input

doble clic

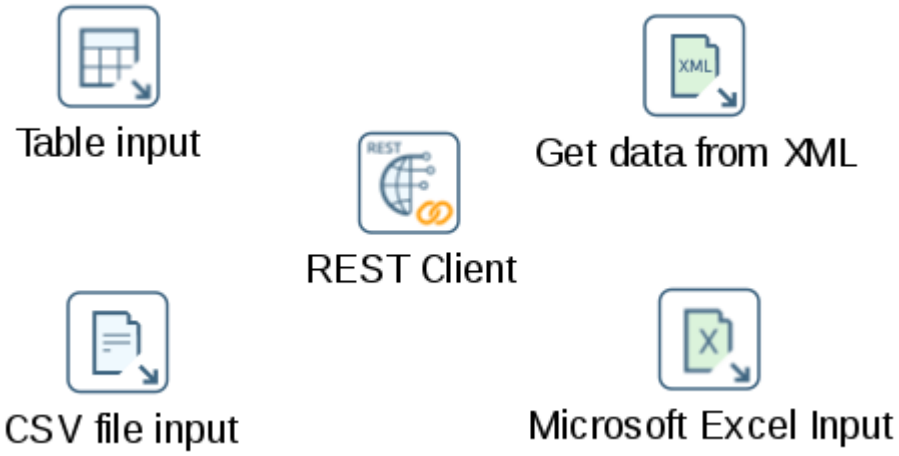


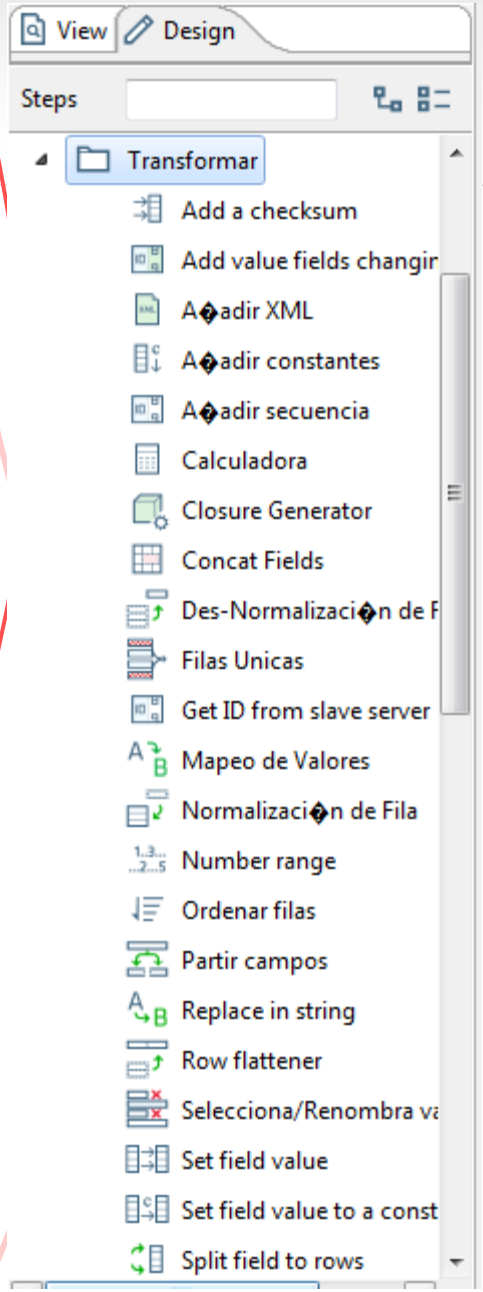


Pentaho Data Integration

Pasos EXTRACCIÓN

- Entrada (**input**)
  - Permiten acceder a recursos para **leer datos**
    - Ficheros, bases de datos, etc.
  - Crean un **flujo de salida** con los datos leídos

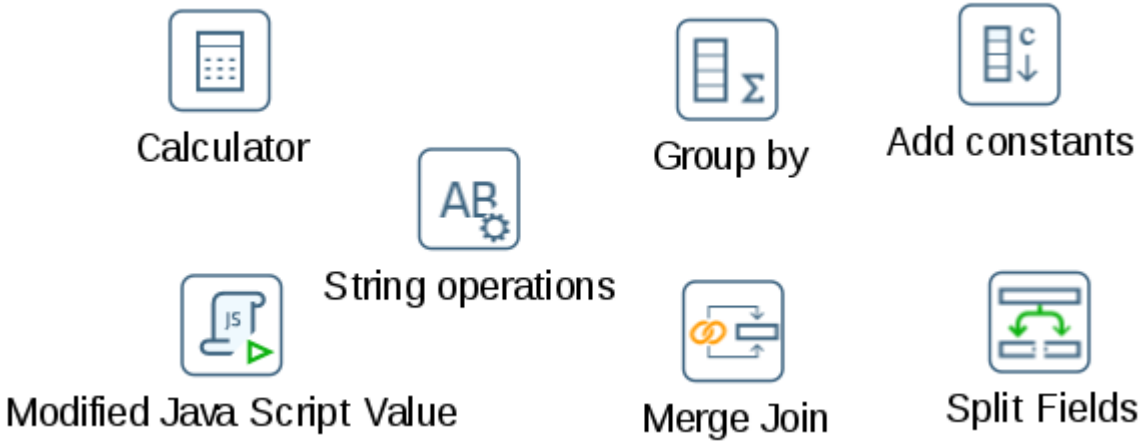


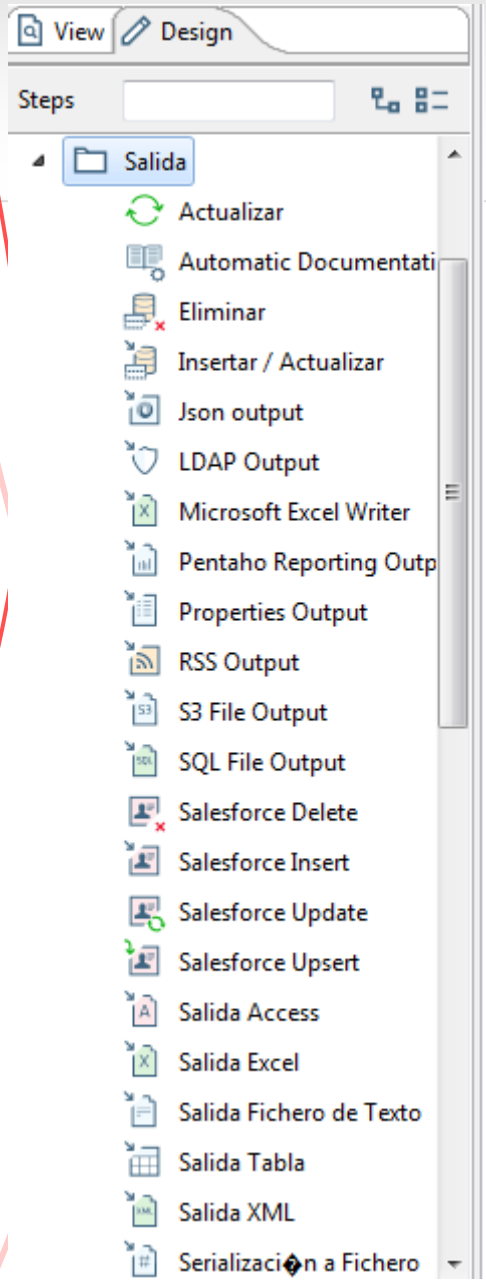


Pentaho Data Integration

Pasos TRANSFORMACIÓN

- Transformar (**transforming**)
  - Permiten desarrollar una **acción** concreta en el flujo de datos de entrada

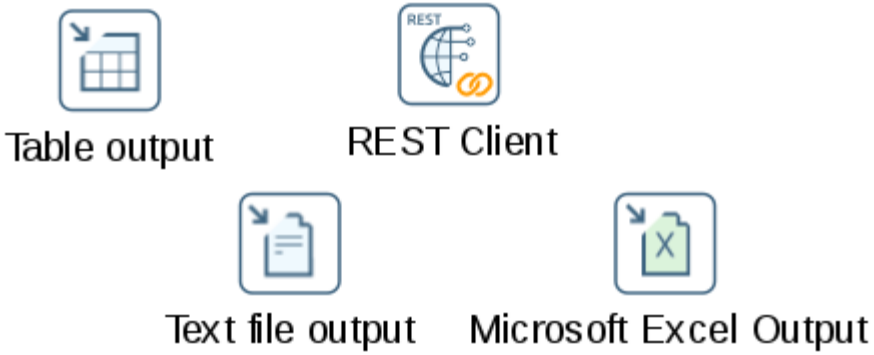




Pentaho Data Integration

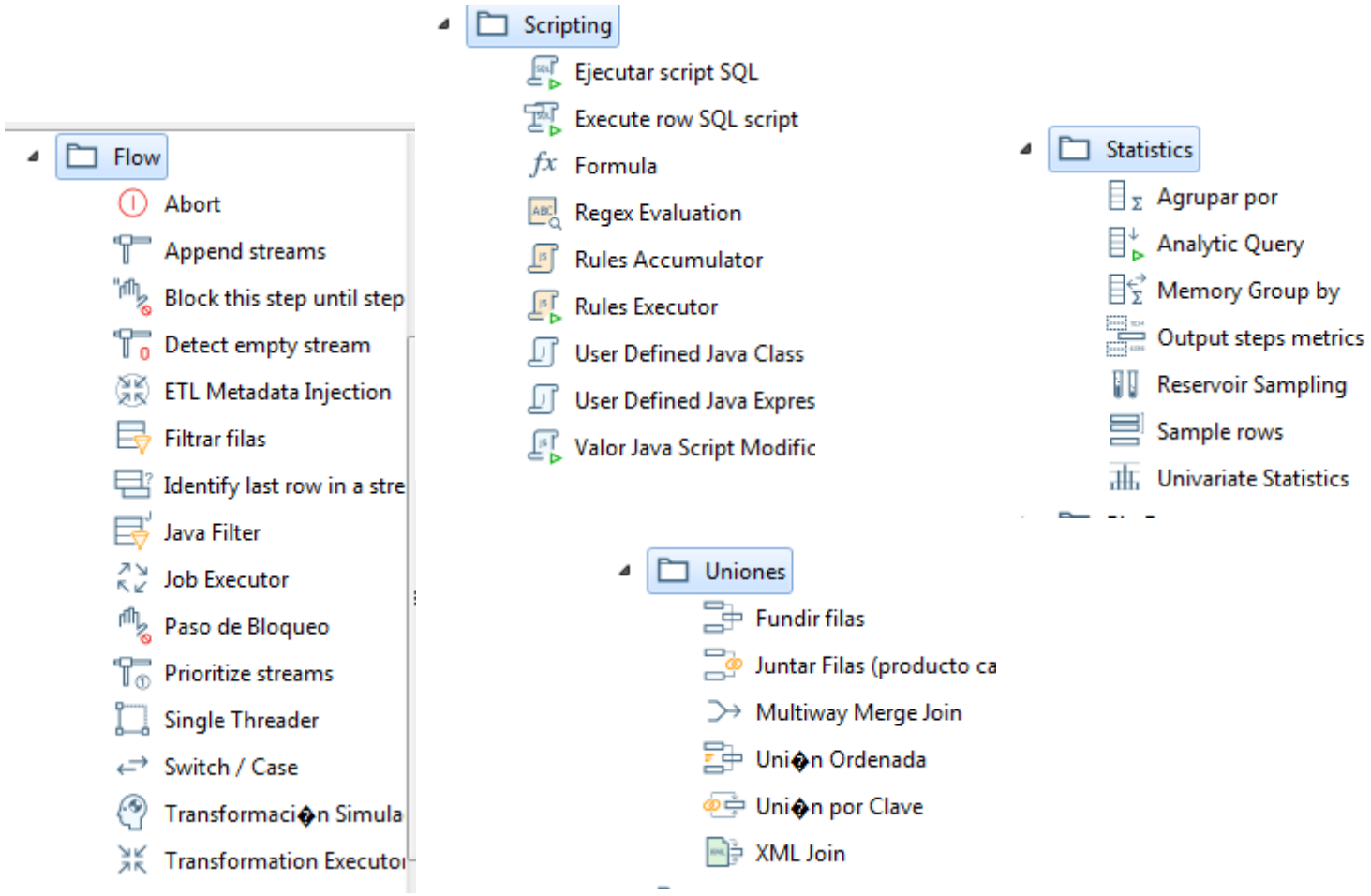
Pasos CARGA

- Salida (**output**)
  - Permiten **leer de un flujo de datos** y almacenarlos en un **recurso externo**
    - Fichero, base de datos, etc.



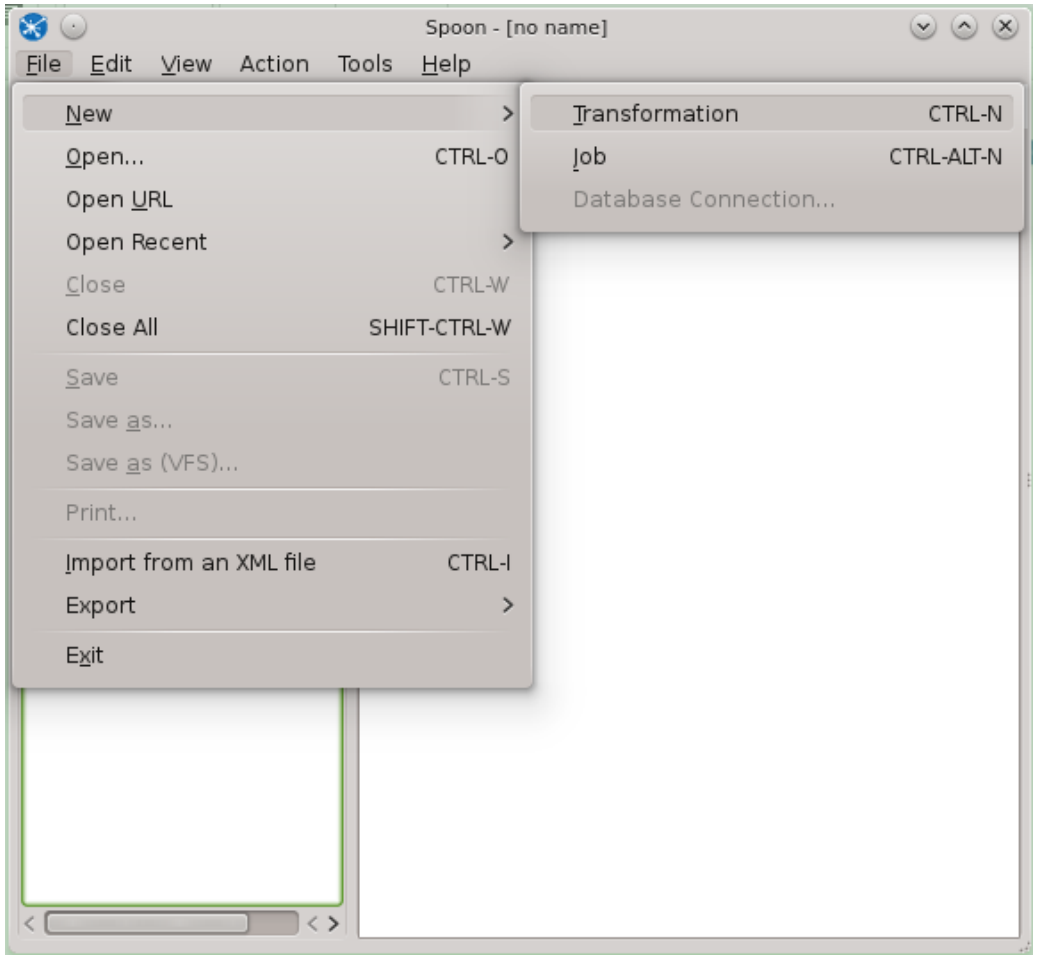
Pentaho Data Integration

Más tipos de pasos



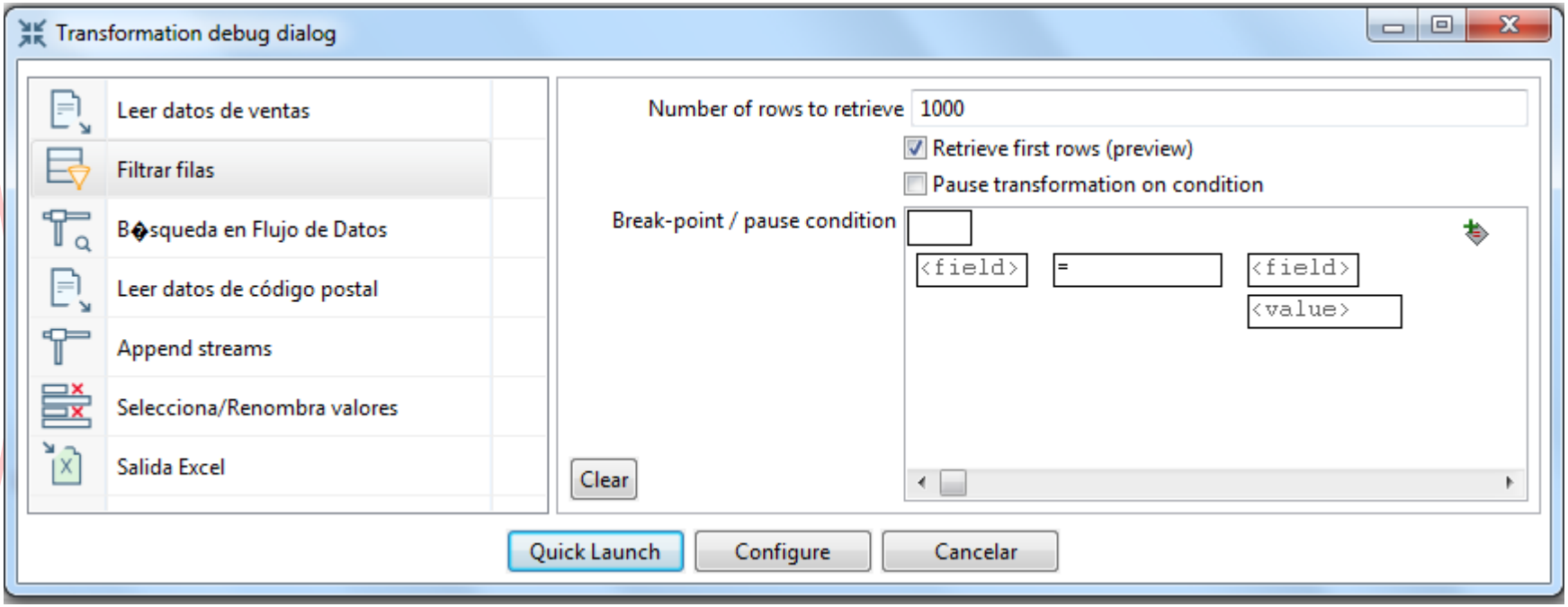
Pentaho Data Integration

Crear nueva transformación

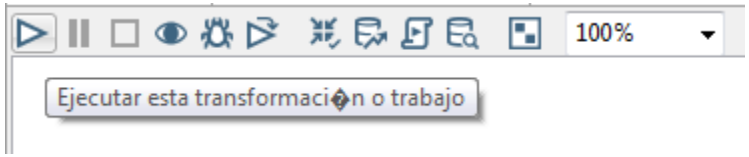


Ejecutar transformación

- Previsualizar (**preview**)
  - Clic derecho encima del paso y opción “preview”



Ejecutar transformación



Ejecutar una transformaci&#65533;n

Ejecuci&#65533;n local, remota o clustered

Ejecuci&#65533;n local

Ejecuci&#65533;n remota

Servidor remoto

Pass export to remote server

Ejecuci&#65533;n clustered

Enviar transformaci&#65533;n

Preparar ejecuci&#65533;n

Iniciar ejecuci&#65533;n

Mostrar transformaciones

Details

Habilitar modo seguro

Gather performance metrics

Clear the log before execution

Nivel de registro: Basic logging

Fecha de Ejecuci&#65533;n (yyyy/MM/dd HH:mm:ss)

Parameters

#	Parameter	Value	Default value
1			

Par&#65533;metros

#	Par&#65533;metro	Valor
1		

Variables

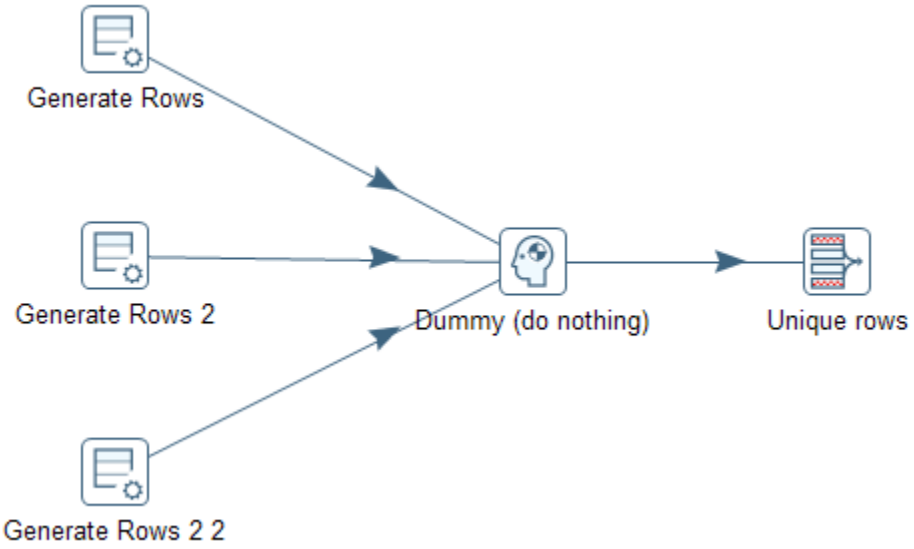
#	Variable	Valor
1	Internal.Job.Filename.Directory	Parent Job File Directory
2	Internal.Job.Filename.Name	Parent Job Filename
3	Internal.Job.Name	Parent Job Name
4	Internal.Job.Repository.Directory	Parent Job Repository Directory

Ejecutar

Cancelar

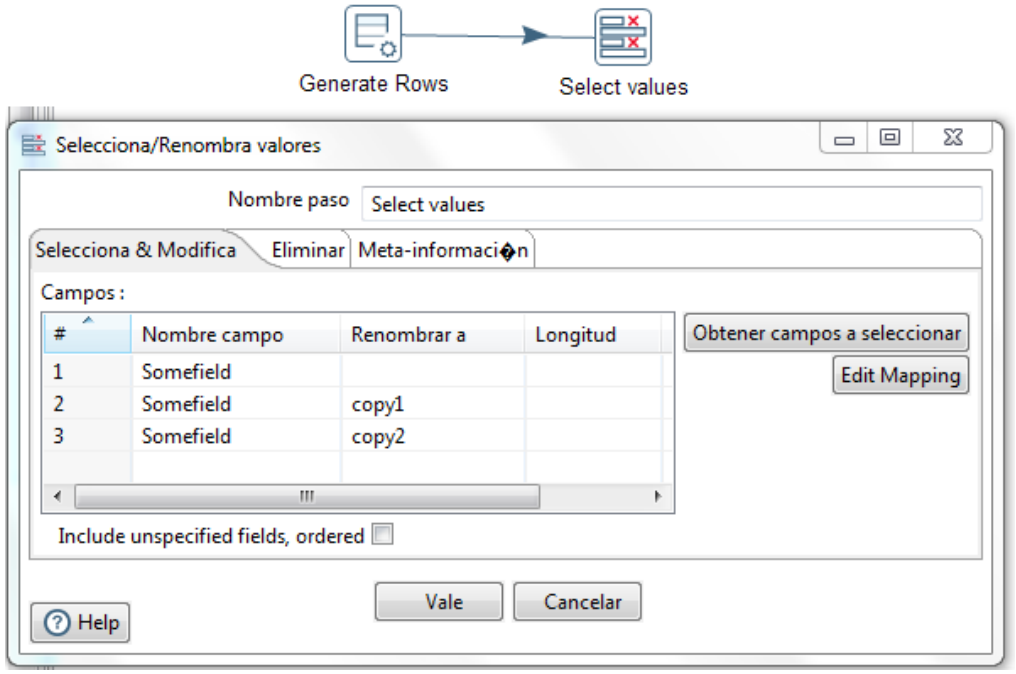
Filas únicas

- Elimina las filas duplicadas de entrada
- **samples\transformations\Unique - Case insensitive unique.ktr**



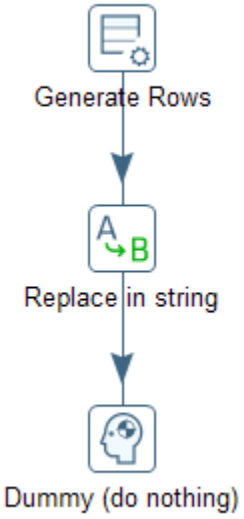
Seleccionar valores

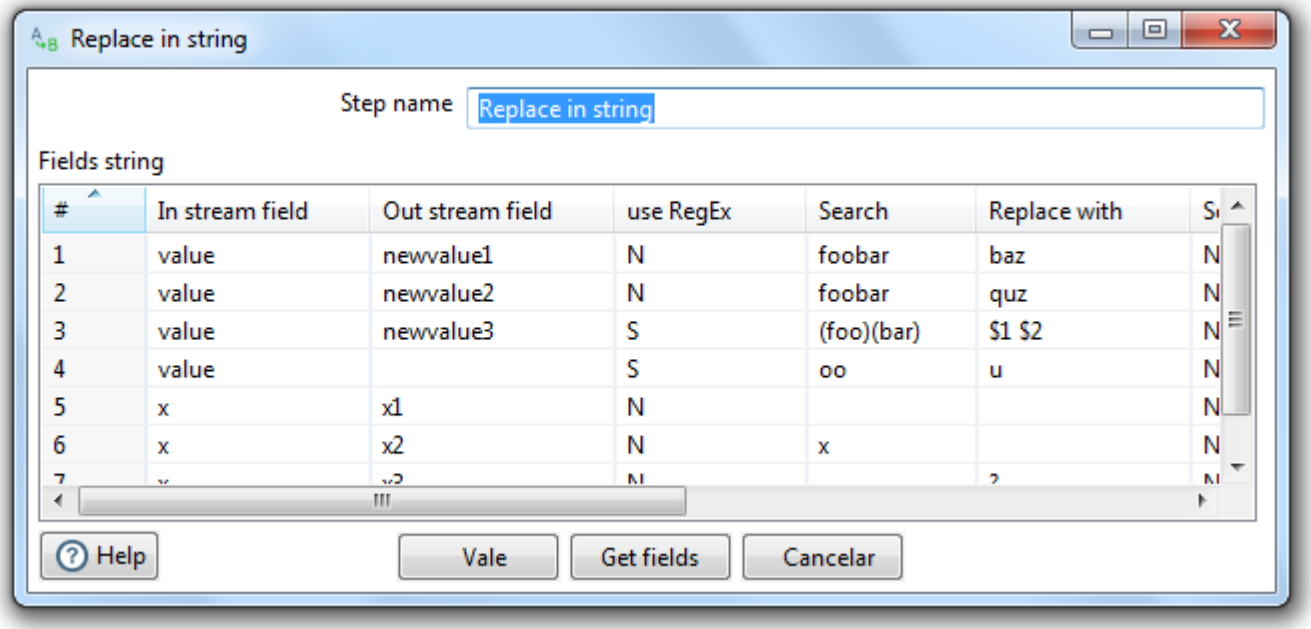
- Obtiene el valor de un subconjunto de campos
- **samples\transformations\Select Values - copy field values to new fields.ktr**



Reemplazar en cadena de caracteres

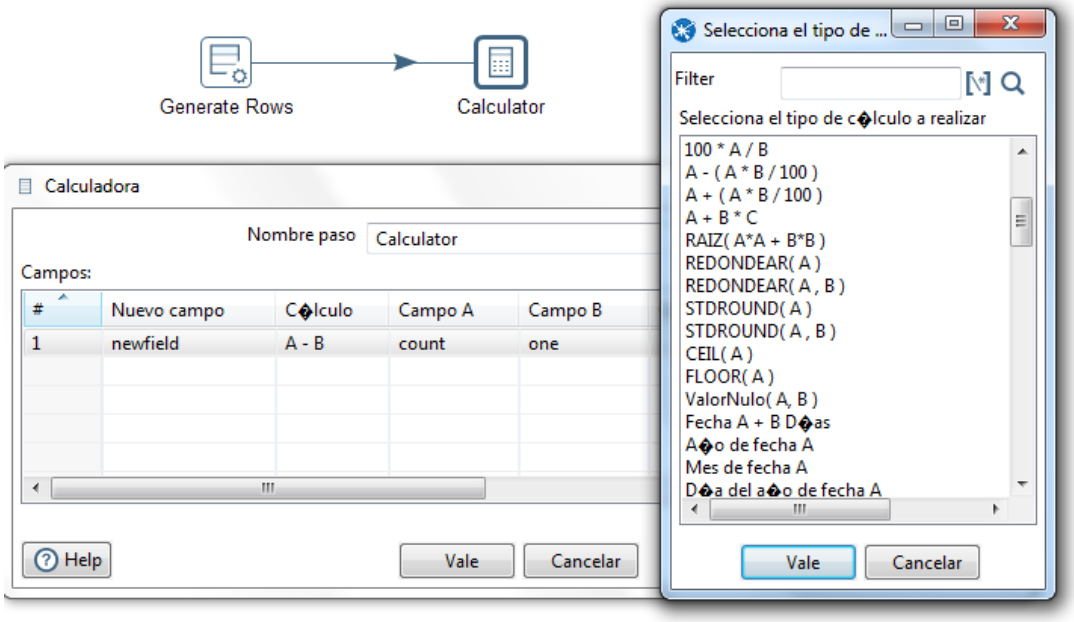
- Permite cambiar unos caracteres por otros
- **samples\transformations\Replace in string - Simple example.ktr**





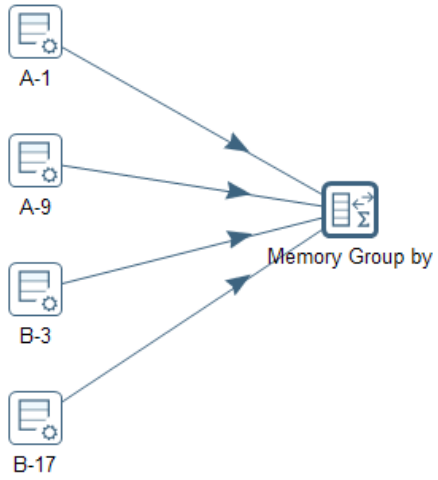
Calculadora

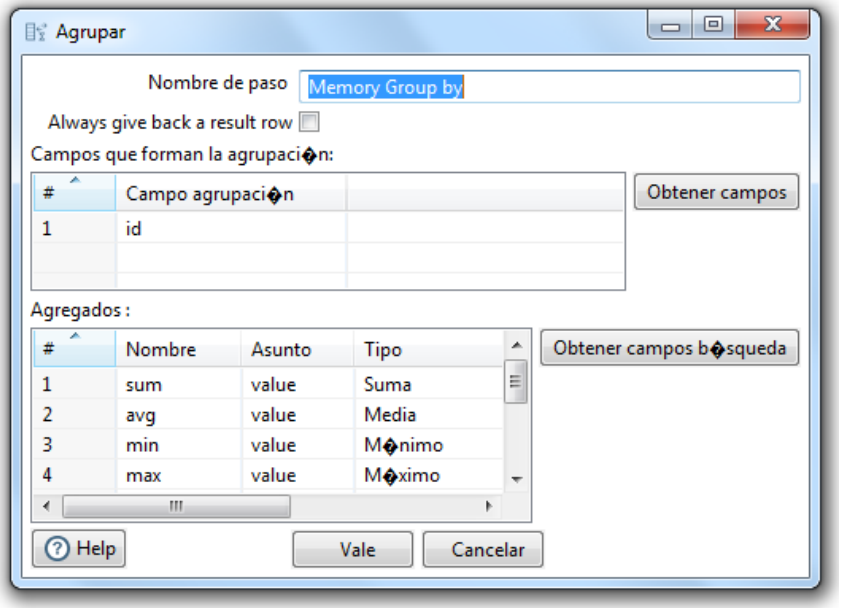
- Suministra funciones predefinidas que pueden ser ejecutadas sobre los campos de entrada
- **samples\transformations\Calculator.ktr**



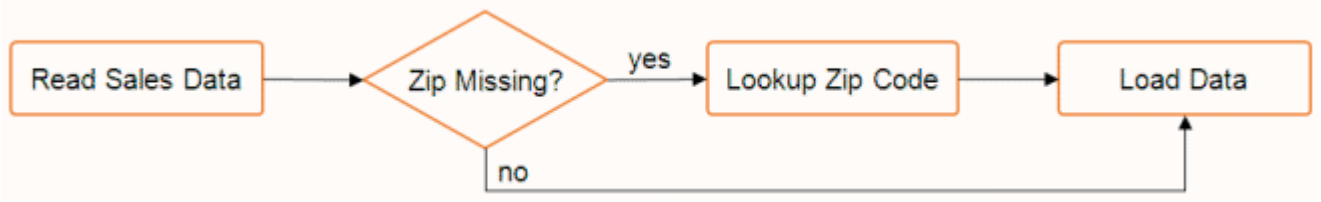
Agrupar

- Permite calcular valores a partir de los valores definidos en un campo
- **samples\transformations\Memory Group By - simple example.ktr**



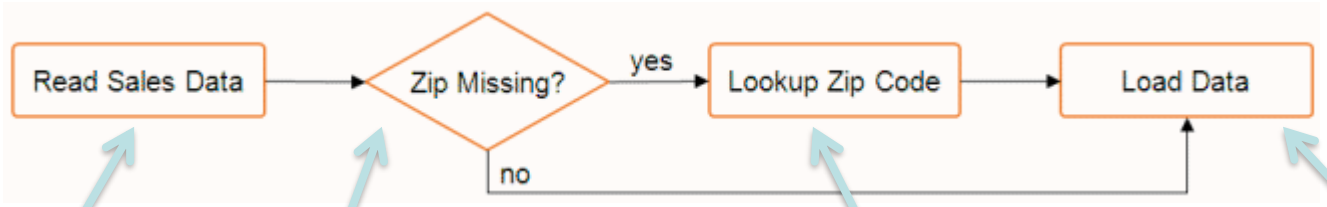


Ejemplo



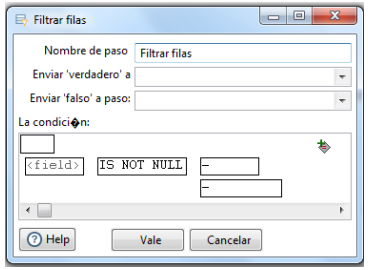
- **Lectura de datos**
  - Desde fichero **samples\transformations\files\sales_data.csv**
- **Filtrado**
  - Determinar si existe código postal
- **Búsqueda de datos**
  - Obtener código postal de **samples\transformations\files\Zipssortedbycitystate.csv**
- **Carga de datos**
  - Fichero MS Excel

Ejemplo



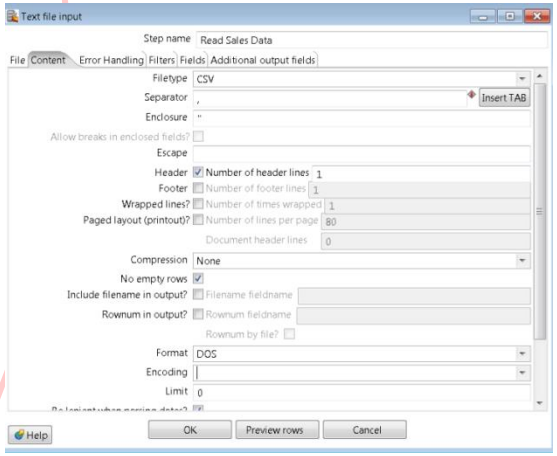
Entrada Fichero de Texto

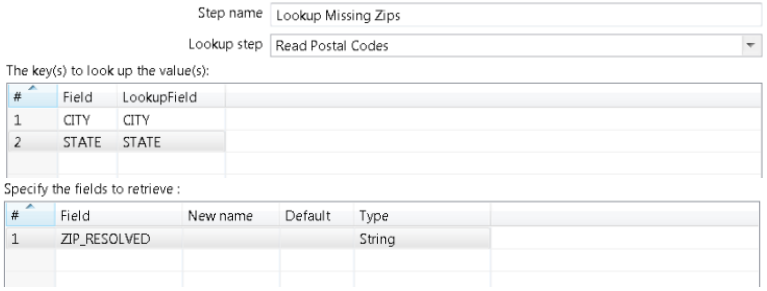
Filtrar filas

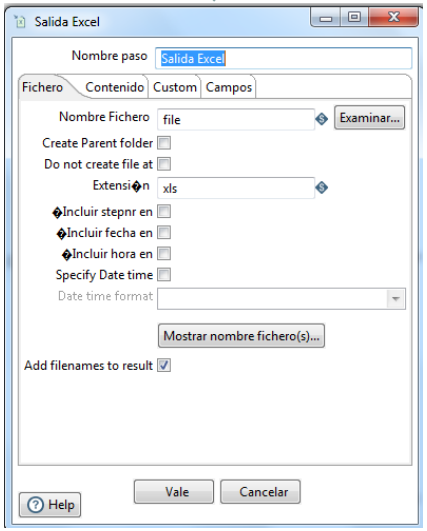


B&#65533;squeda en Flujo de Datos

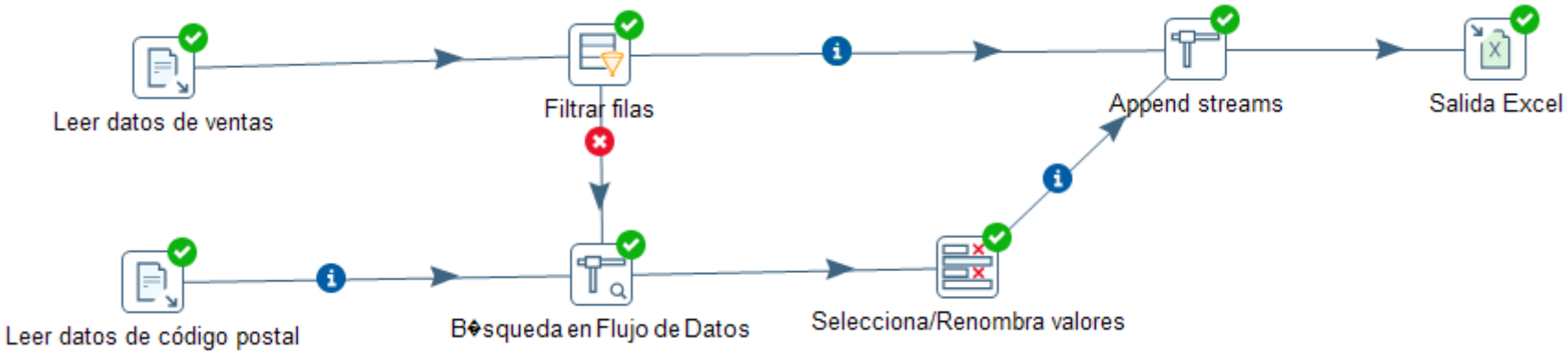
Salida Excel



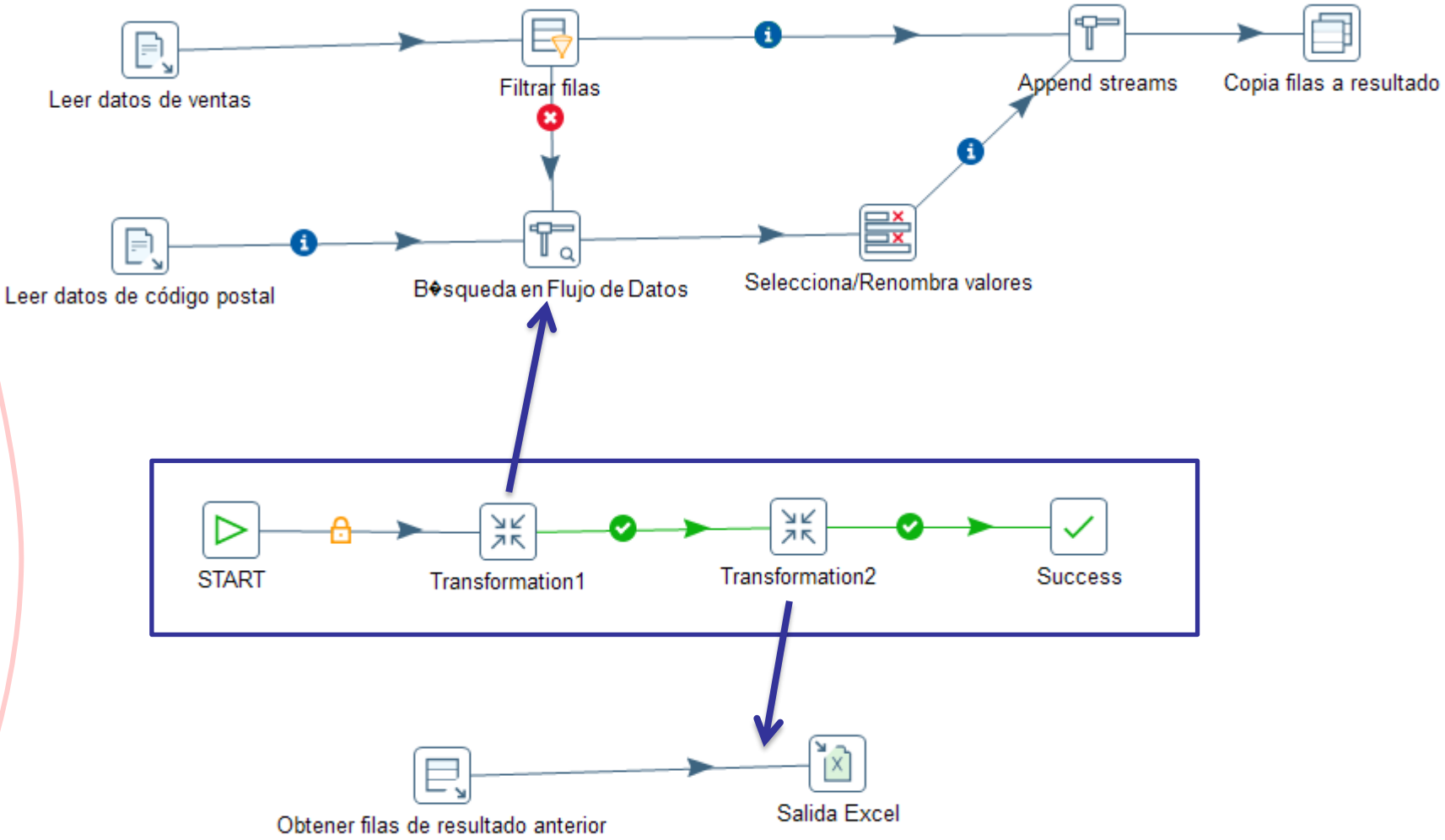




Ejemplo


Leer datos de ventas
Filtrar filas
Append streams
Salida Excel
Leer datos de código postal
B♦squeda en Flujo de Datos
Selecciona/Renombra valores

Ejemplo de trabajo



Ejercicio

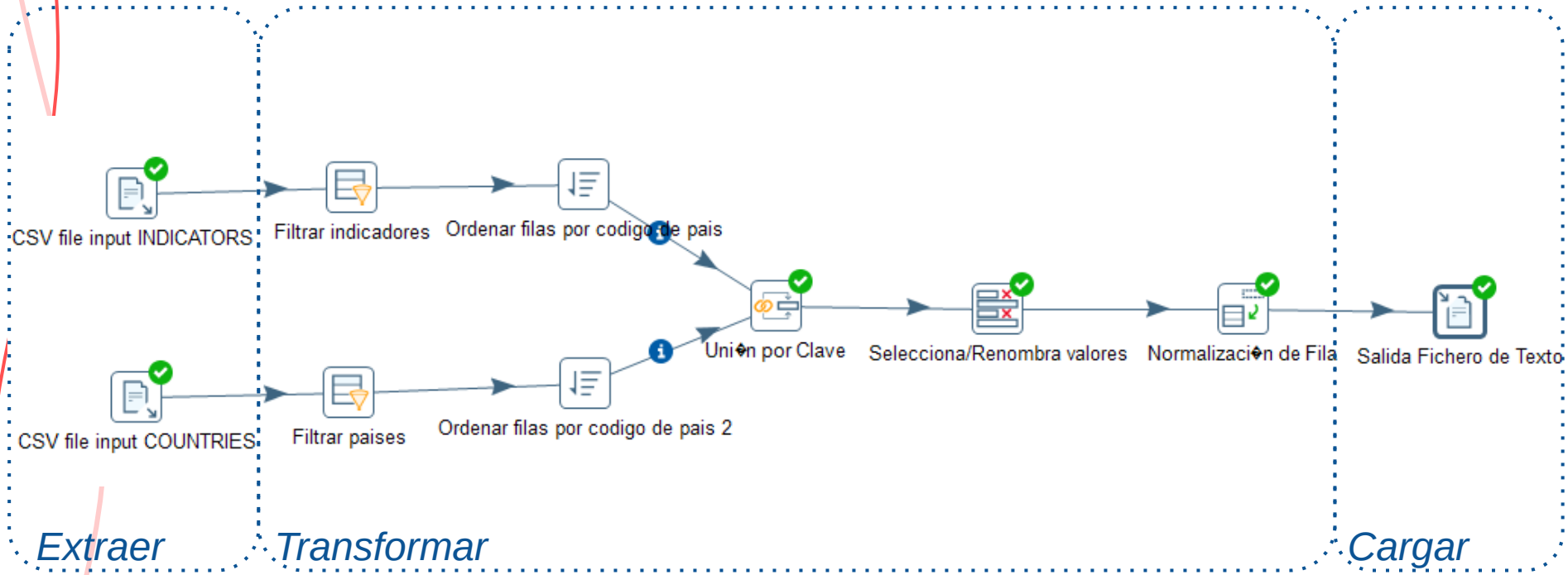
- Fuentes de datos
  - Portal de datos abiertos del Banco Mundial
    - http://data.worldbank.org/
    - Estadísticas de educación
      - http://data.worldbank.org/data-catalog/ed-stats
- Objetivo
  - Conseguir un fichero CSV
    - Nombre del indicador, código de indicador, país, código del país, región, año y valor
  - Sólo ciertos indicadores
  - Sólo países de Europa
  - Sólo ciertos años

Ejercicio

```
ï»¿"Country Name","Country Code","Indicator Name","Indicator Code","1977","1978","1979","1980","1981",'
91","1992","1993","1994","1995","1996","1997","1998","1999","2000","2001","2002","2003","2004","2005",'
15","2016","2020","2025","2030","2035","2040","2045","2050","2055","2060","2065","2070","2075","2080",'
"Arab World","ARB","Adjusted net enrolment rate, lower secondary, both sexes (%)","UIS.NERA.2","","","
","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","",
"Arab World","ARB","Adjusted net enrolment rate, lower secondary, female (%)","UIS.NERA.2.F","","","",'
"","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","",
"Arab World","ARB","Adjusted net enrolment rate, lower secondary, gender parity index (GPI)","UIS.NERA
","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","
"Arab World","ARB","Adjusted net enrolment rate, lower secondary, male (%)","UIS.NERA.2.M","","","",""
,"","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","",
"Arab World","ARB","Adjusted net enrolment rate, primary, both sexes (%)","SE.PRM.TENR","61.25255","62
.9397","68.54954","69.57877","70.82125","71.56102","71.66595","72.10472","72.80175","70.32265","70.873
,"78.48897","79.95569","80.95473","81.33537","81.76691","82.22437","82.82338","82.81864","83.70538","8
","","","","","","","","","","","","","","","",
"Arab World","ARB","Adjusted net enrolment rate, primary, female (%)","SE.PRM.TENR.FE","50.58012","51.
6525","60.05314","61.34173","62.89461","63.99371","64.15573","65.30609","66.56963","62.65608","63.5247
","74.32056","75.98818","77.18511","77.79193","78.41859","78.71033","79.64724","79.75937","80.86663","
","","","","","","","","","","","","","","","",
```

```
Country;Country.code;Indicator;Indicator.code;Region;year;value
Andorra;ADO;GDP per capita (current US$);NY.GDP.PCAP.CD;Europe & Central Asia;2000;21432.9600712136
Andorra;ADO;GDP per capita (current US$);NY.GDP.PCAP.CD;Europe & Central Asia;2001;21897.6629394768
Andorra;ADO;GDP per capita (current US$);NY.GDP.PCAP.CD;Europe & Central Asia;2002;24175.3727542522
Andorra;ADO;GDP per capita (current US$);NY.GDP.PCAP.CD;Europe & Central Asia;2003;31742.9925847537
Andorra;ADO;GDP per capita (current US$);NY.GDP.PCAP.CD;Europe & Central Asia;2004;37235.4500323146
Andorra;ADO;GDP per capita (current US$);NY.GDP.PCAP.CD;Europe & Central Asia;2005;39990.3304085679
```

Solución Ejercicio



Taller sobre integración de datos (abiertos)

Uso de Pentaho Data Integration

Jose Norberto Mazón
Twitter: @jnmazon

Grupo de investigación WaKe
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

lsi Departamento de Lenguajes y Sistemas Informáticos

wake Web and Knowledge Group
Universidad de Alicante