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

Máster Oficial Universitario en
Gestión de la Información
Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Universitat d'Alacant
Universidad de Alicante

lsi Departamento de Lenguajes y Sistemas Informáticos

wake Web and Knowledge Group
Universidad de Alicante

Integración de datos

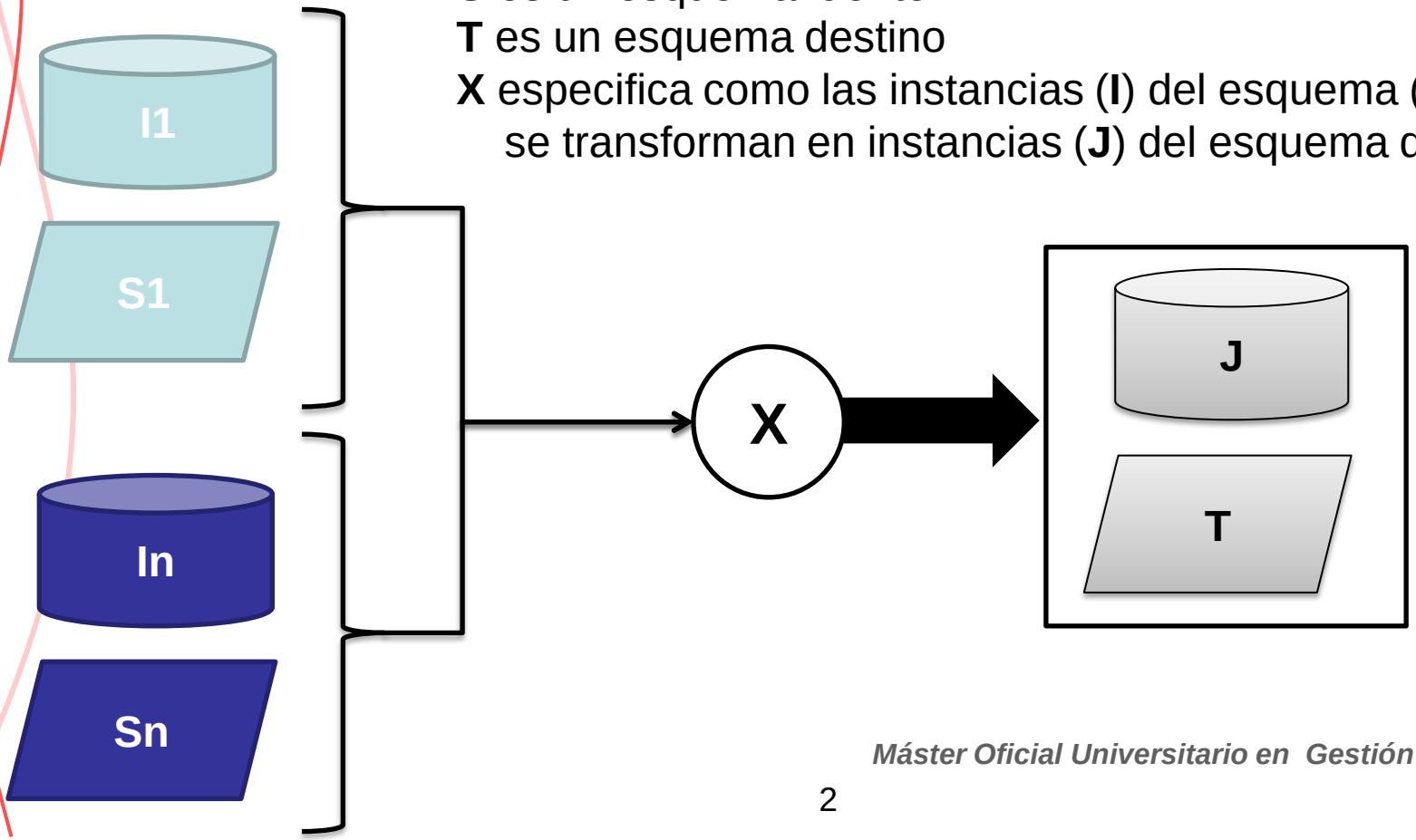
- Transformar las fuentes de datos en un destino de datos

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

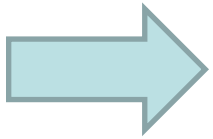
- Archivo **CSV** (Comma Separated Values)
  - Formato para representar datos en forma de tabla en un **fichero de texto**
  - Cada línea en el fichero es una **fila de datos**
  - Cada valor se separa por comas/puntos y comas/otro símbolo representando **columnas**

```
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
    - Conversión
      - Fechas
      - Unidades de medida
      - Etc.
    - Normalización
      - C/ Vicente Blasco Ibáñez 18
      - Calle Blasco Ibáñez nº 18
      - Blasco Ibanez 18
  - Filtrado, Unión, etc.

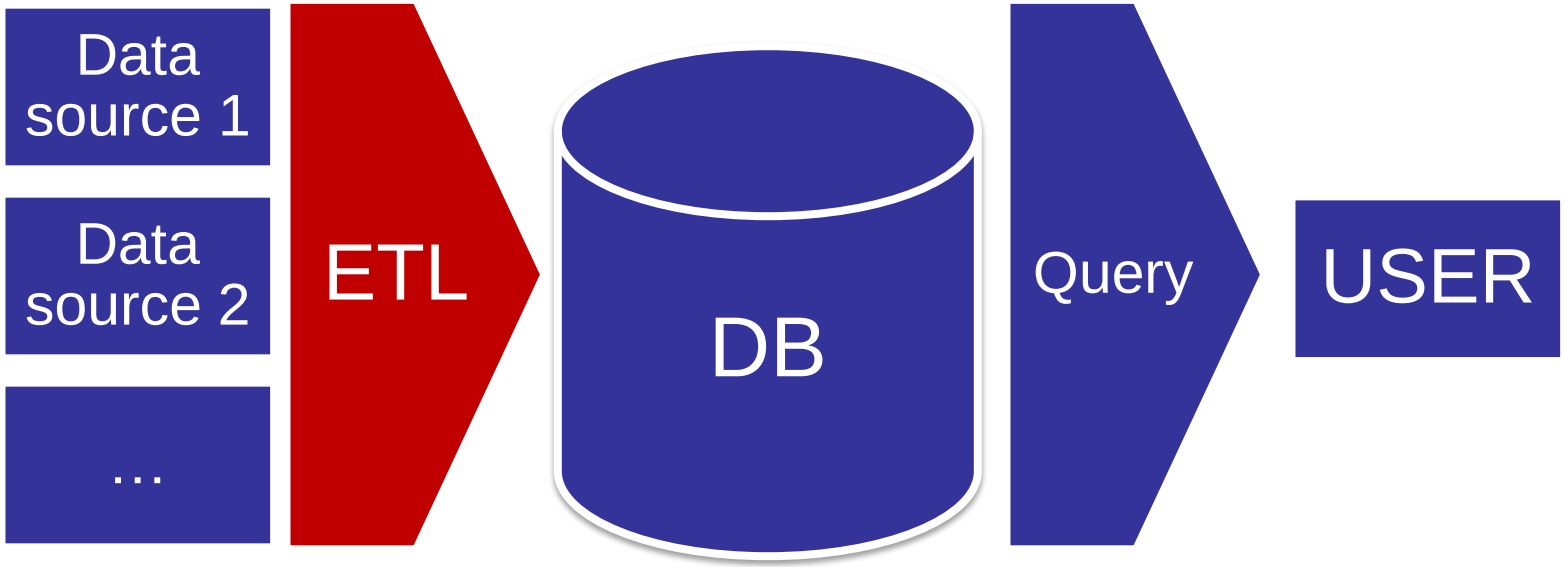
1. Masculino, Femenino
2. 0, 1
3. Hombre, Mujer

H, M

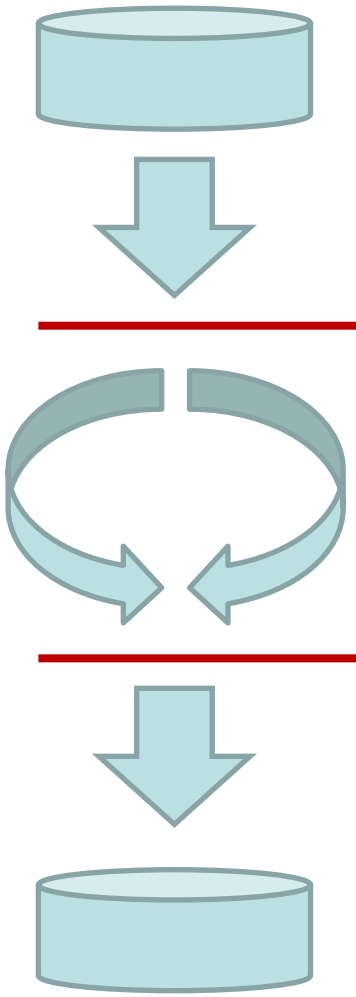
Calle	Número
Vicente Blasco Ibáñez	18

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

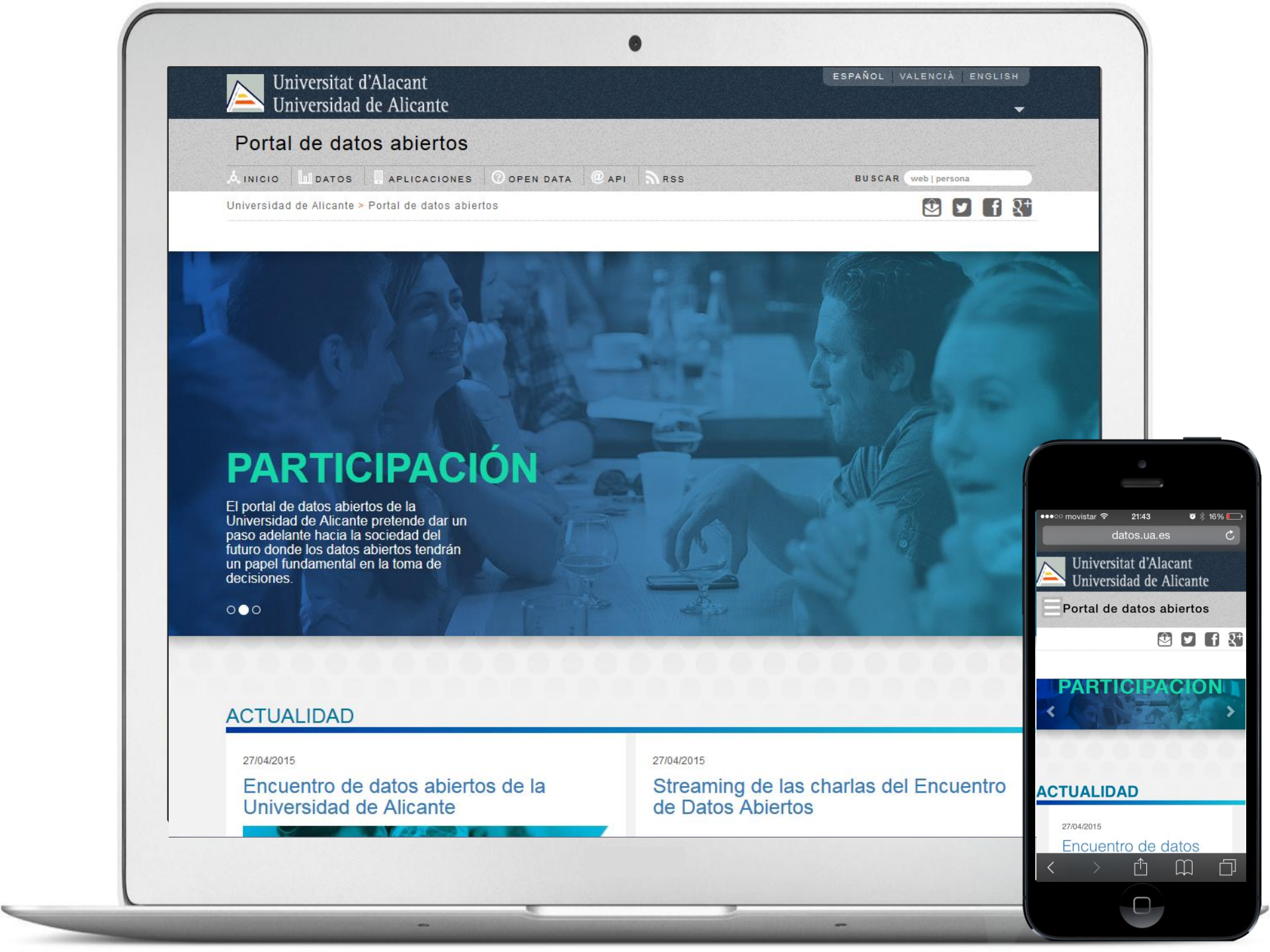


EXTRAER
Recolectar datos de diferentes fuentes de datos

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

CARGAR
Almacenar datos

datos·ua·es

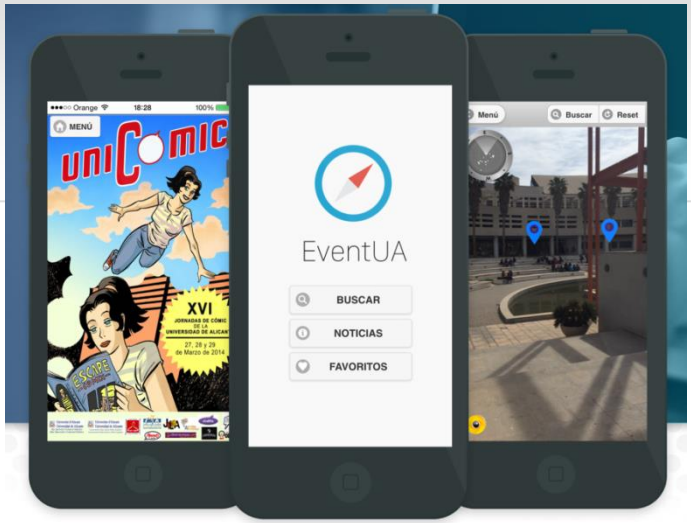

Universitat d'Alacant
Universidad de Alicante
ESPAÑOL | VALENCIÀ | ENGLISH
Portal de datos abiertos
INICIO | DATOS | APLICACIONES | OPEN DATA | API | RSS
BUSCAR web | persona
Universidad de Alicante > Portal de datos abiertos
PARTICIPACIÓN
El portal de datos abiertos de la Universidad de Alicante pretende dar un paso adelante hacia la sociedad del futuro donde los datos abiertos tendrán un papel fundamental en la toma de decisiones.
ACTUALIDAD
27/04/2015
Encuentro de datos abiertos de la Universidad de Alicante
27/04/2015
Streaming de las charlas del Encuentro de Datos Abiertos
movistar 21:43 16%
datos.ua.es
Universitat d'Alacant
Universidad de Alicante
Portal de datos abiertos
PARTICIPACION
ACTUALIDAD
27/04/2015
Encuentro de datos

API datos.ua.es

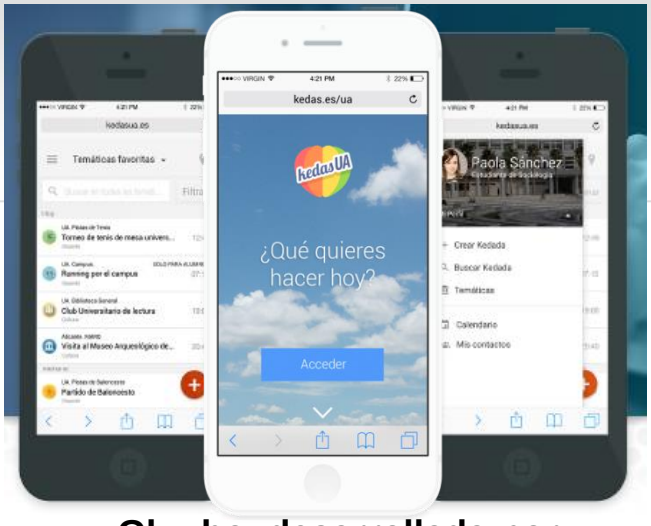
- https://dev.datos.ua.es

Reutilizando datos abiertos

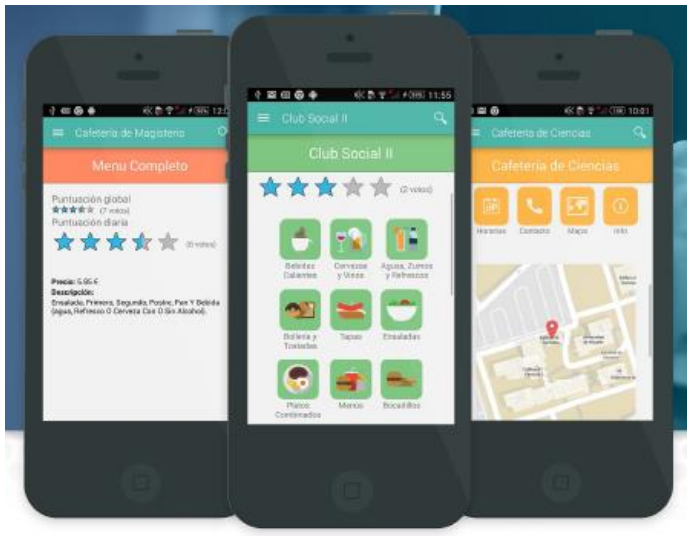
- **2015 → 20 apps presentadas**
  - http://datos.ua.es/es/premios-concurso-aplicaciones.html
- **2016 → 36 apps y visualizaciones presentadas**
  - https://datos.ua.es/es/premios-uabierta-2016.html



EventUA, desarrollada por Alexandre Rubio Alba



Gluubo, desarrollada por Andrea Lluch, Jorge Juan Oliva y Jose Luis Pérez

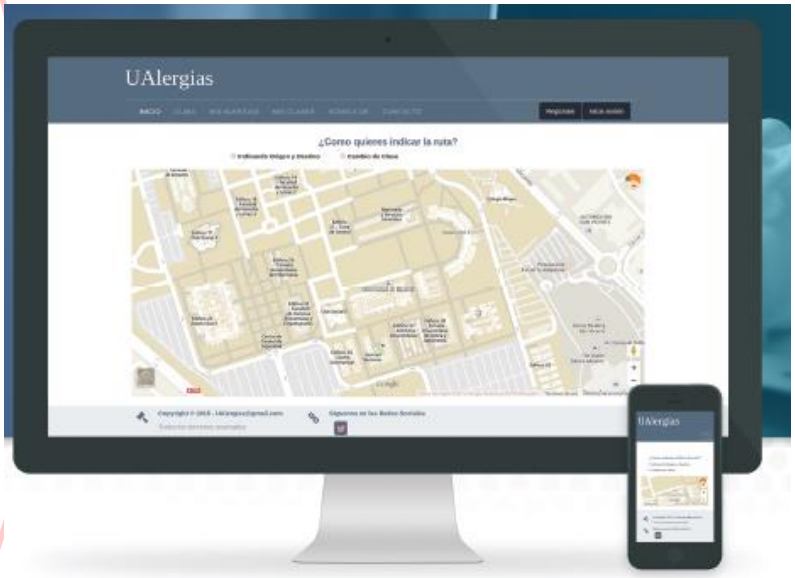


ComerUA, desarrollada por Carolina Prada Hernández

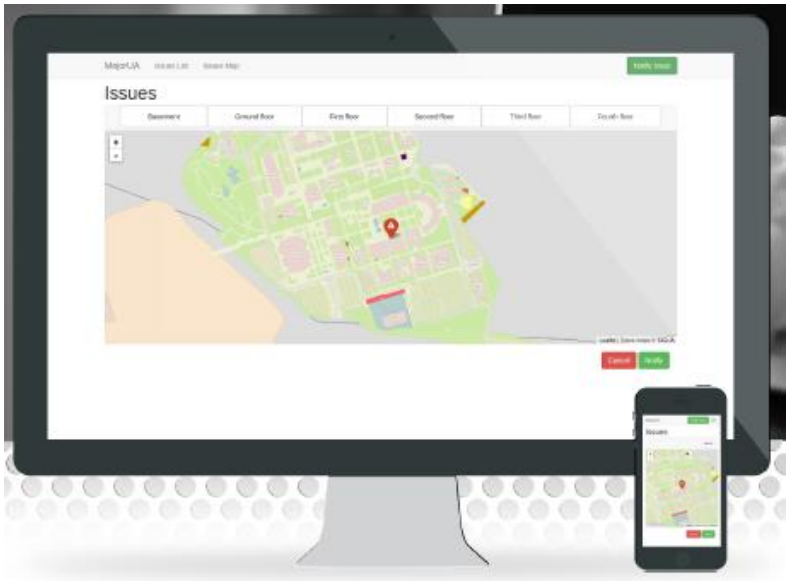
QuizUA, desarrollada por Pablo Marzal Garrigós

BecUAs, desarrollada por Juan Miguel Sánchez Belmonte

CulturAL, desarrollada por Hector Rico García y Jose Antonio Arques Castelló



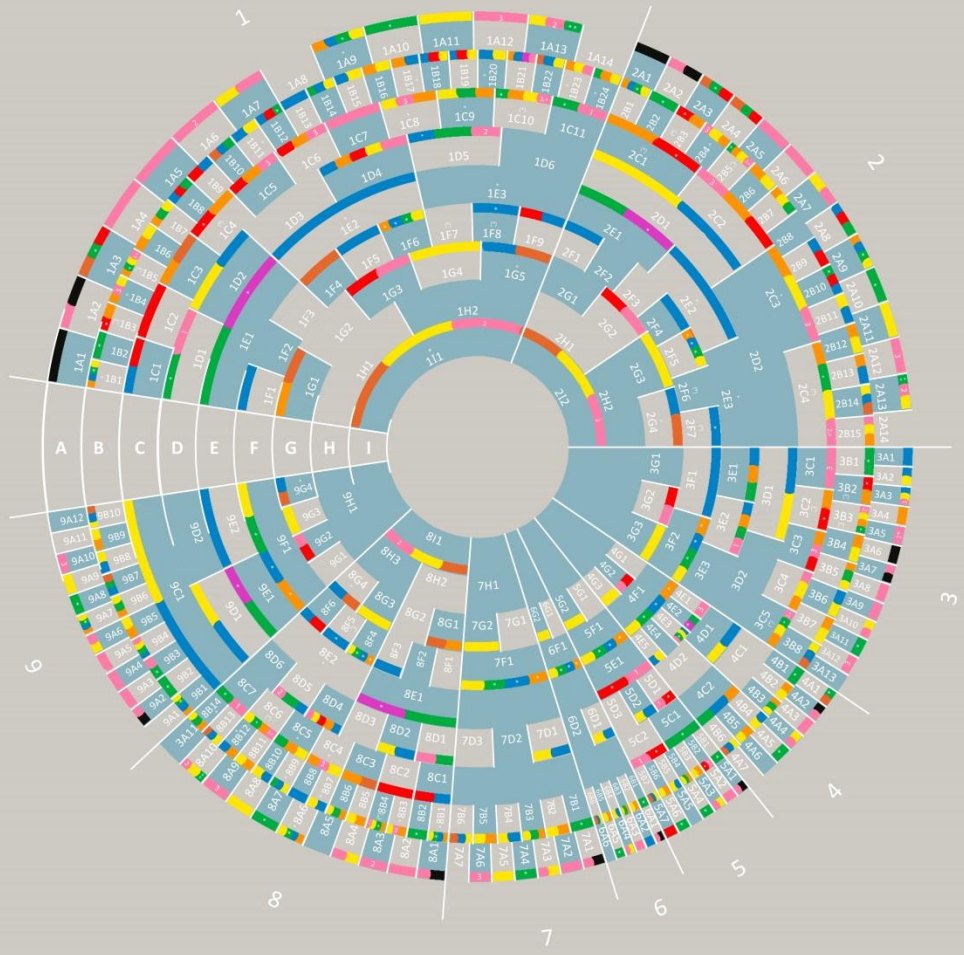
UAlergias, desarrollada por María González Pérez



MejorUA, desarrollada por Samuel Ortiz Reina

GRÁFICO DE BECAS Y AYUDAS UNIVERSIDAD DE ALICANTE

Becatcher UA



LO QUE SOY

1 Grado

2 Master / Postgrado

3 Doctorando / Investigador

4 Profesorado / Docente / Doctor

5 Discapacitado

6 Deportista

7 Futuros estudiantes

8 Estudios a extinguir

9 Otro
Profesional; Lectorado; Formación profesional, grado medio; Organismos, universidades, centros públicos, entidades, instituciones; conservatorio música o danza; Personal administrativo y de servicios de la UA, funcionario, etc.

* (en la casilla): Exige haber finalizado LO QUE SOY (*) (en la casilla): Tanto para los que han finalizado como para los que no

LO QUE BUSCO

A Recursos Económicos
Matrícula, alojamiento, transporte, etc.

B Idiomas / Movilidad
Nacional e internacional

C Prácticas / Formación
Nacional e internacional

D Master / Postgrado
Doctorado, postdoctorado, apoyo

E Investigación

F Colaboración

G Cursos y actividades
Cursos, congresos, actividades

H Otros
Préstamos, voluntariado, solidaridad

I Rendimiento académico

* (Sin especificación): Implica condicionante o excepción de algún otro tipo

REQUISITOS

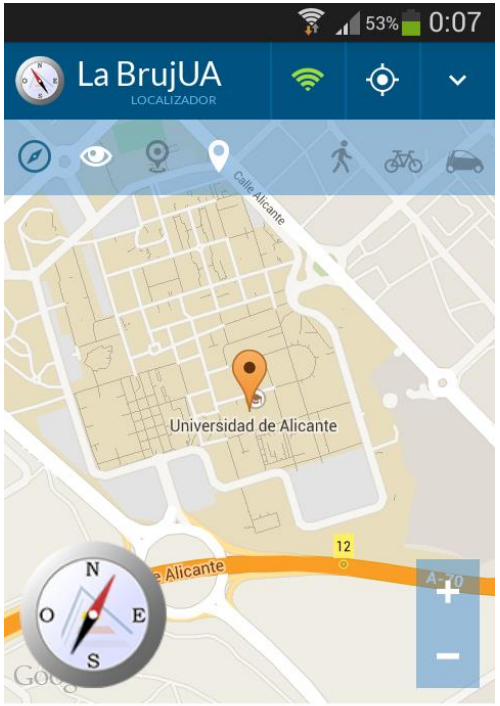
- Rentas
- Idiomas
- Créditos
- Estudios
- Nacionalidad
  Residente (*)
  Empadronamiento (**)
- Experiencia
- Expediente / Méritos
- Edad / Períodos
- Otros
  (1) Estar admitido en una universidad, instituto investigación, etc.
  (2) Localización universidad (de origen)
  (3) Incompatible con otras becas
  Otras

+ : Acumula condiciones de tipo "Otros"

Becatcher, desarrollada por Fernando Colom

Dos ejemplos concretos (1)

- LaBrujUA
  - Desarrollada por Carlos Rafael Constán Nava
  - https://play.google.com/store/apps/details?id=labrujua.labrujua
  - ¿Cómo ir del punto A al punto B en el campus de la UA?
  - ¿Y si estoy fuera del campus?
    - Transporte público
    - Horarios de clase
    - Etc.



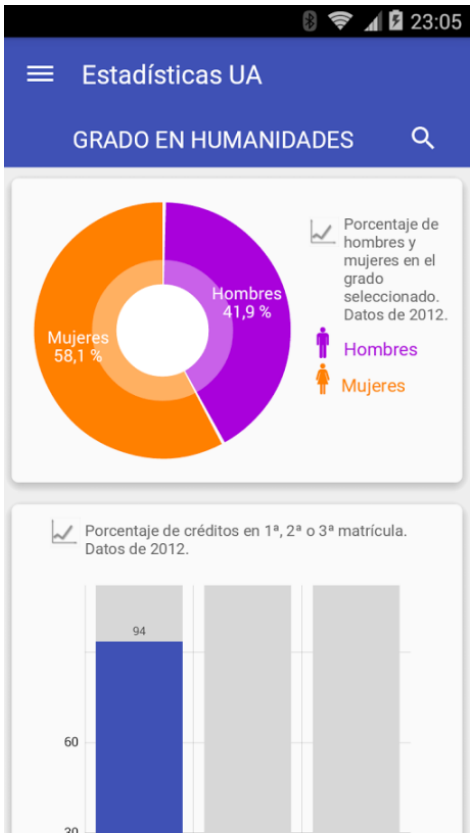
Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

- Necesito **integrar datos**

Dos ejemplos concretos (y 2)

- GradUAte
  - Desarrollada por Pablo González Carrizo
  - https://play.google.com/store/apps/details?id=com.pgonzalezcarrizo.comienzaua
  - ¿A qué titulación me matriculo?
  - Estadísticas de titulaciones
    - Datos de las titulaciones
    - Datos de matriculación en titulaciones
    - Etc.



- Necesito **integrar datos**

¿Cómo especificar transformaciones de datos?

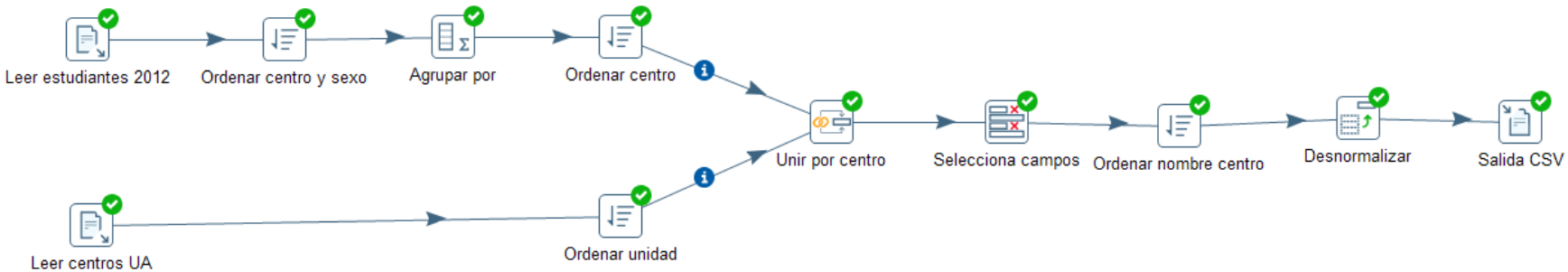
- Directamente programando (código)
  - SQL, Java, Pig, etc.

```
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
```

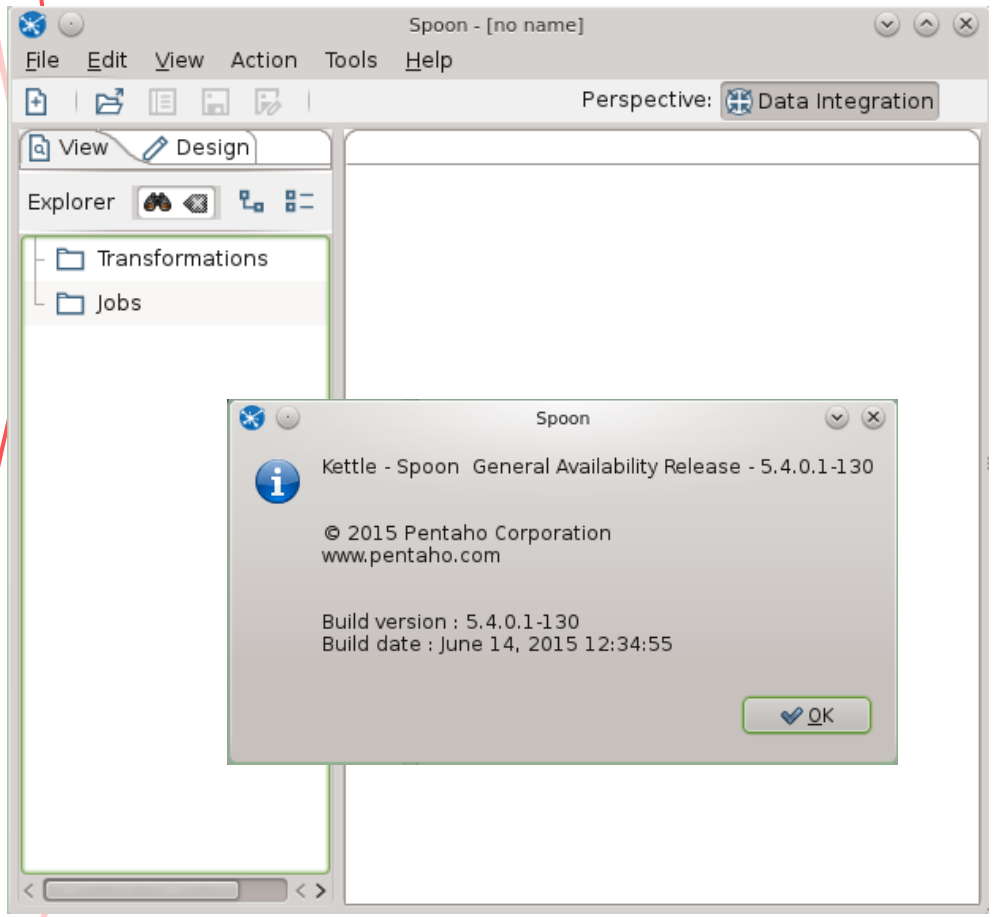
- Modelos mediante interfaz visual
  - El código se genera a partir del modelo



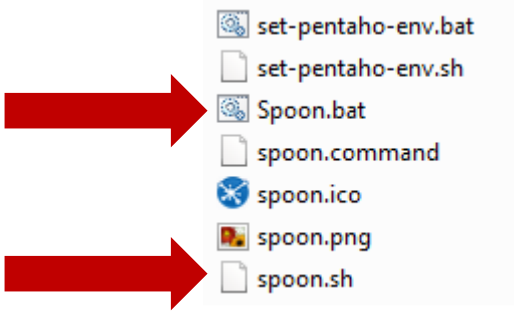
Pentaho Data Integration

- AKA Pentaho Kettle
- http://community.pentaho.com/projects/data-integration/
- 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**

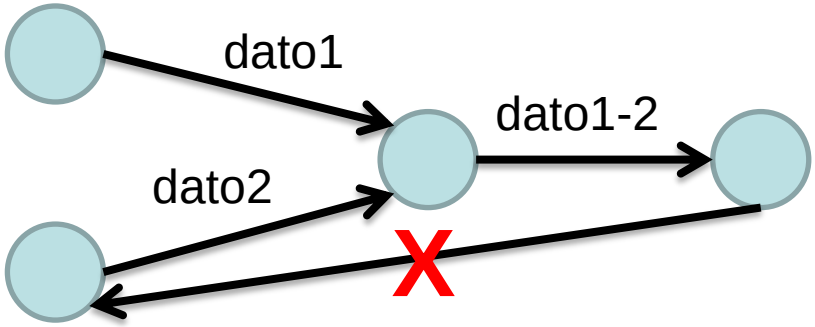


Pentaho Data Integration

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

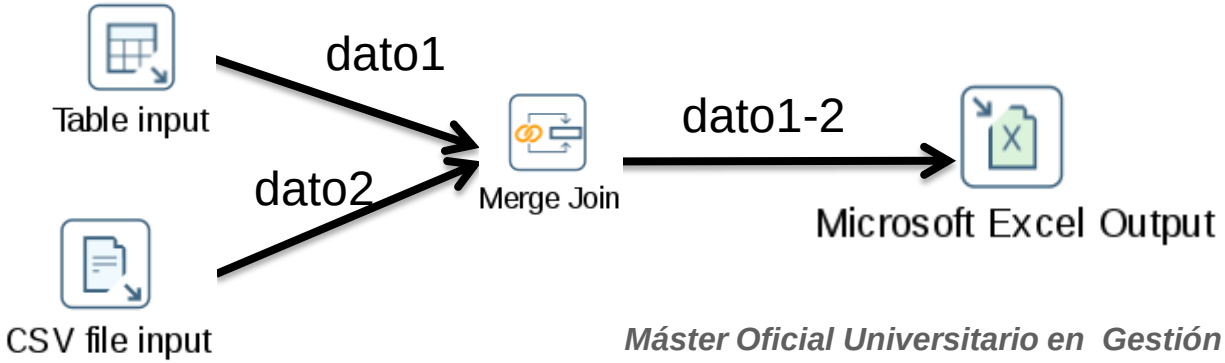
Pentaho Data Integration
Transformaciones

- **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

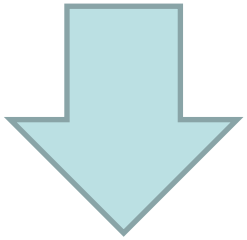


Pentaho Data Integration
Transformaciones

- 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

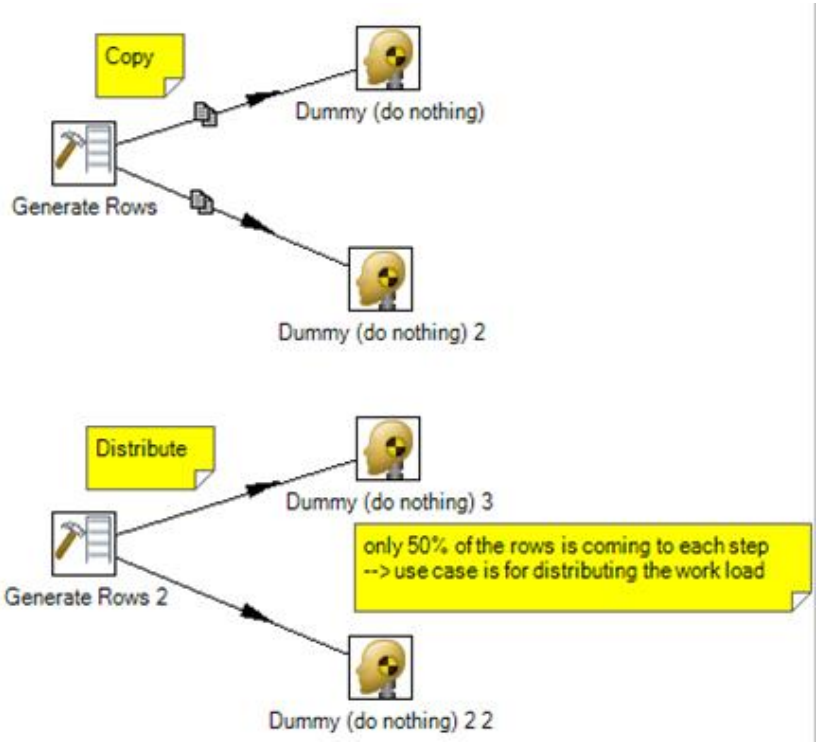
"Nombre","Fecha nacimiento","CP"
"Pepe","**1984**","03181"
"María","**1990**","03690"
"Ana","**1971**","03080"

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

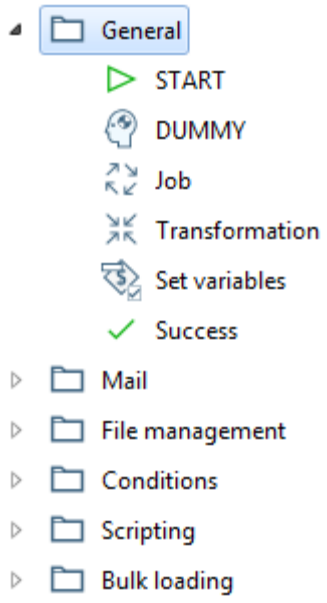


Pentaho Data Integration
Trabajos

- **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

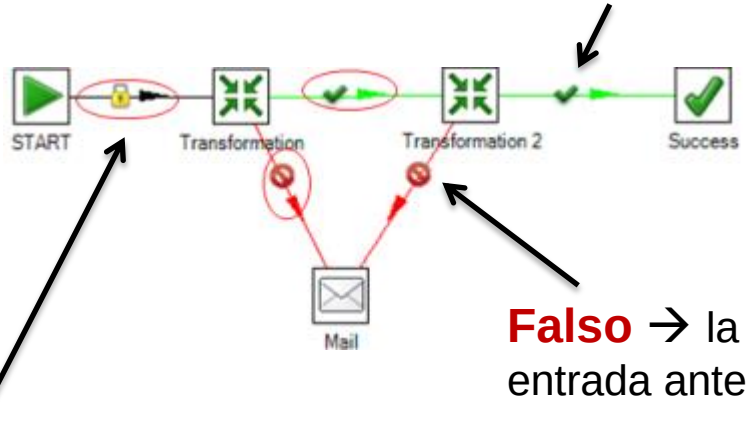
- **Saltos** en un trabajo
  - Son **flujos de control**, no de datos
    - Para pasar datos de una entrada de trabajo a otra entrada hay que usar la variable global resultado (**Result**)
    - Copia filas a resultado (**Copy Rows to Result**)
      - Paso que permite transferir filas de datos a la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

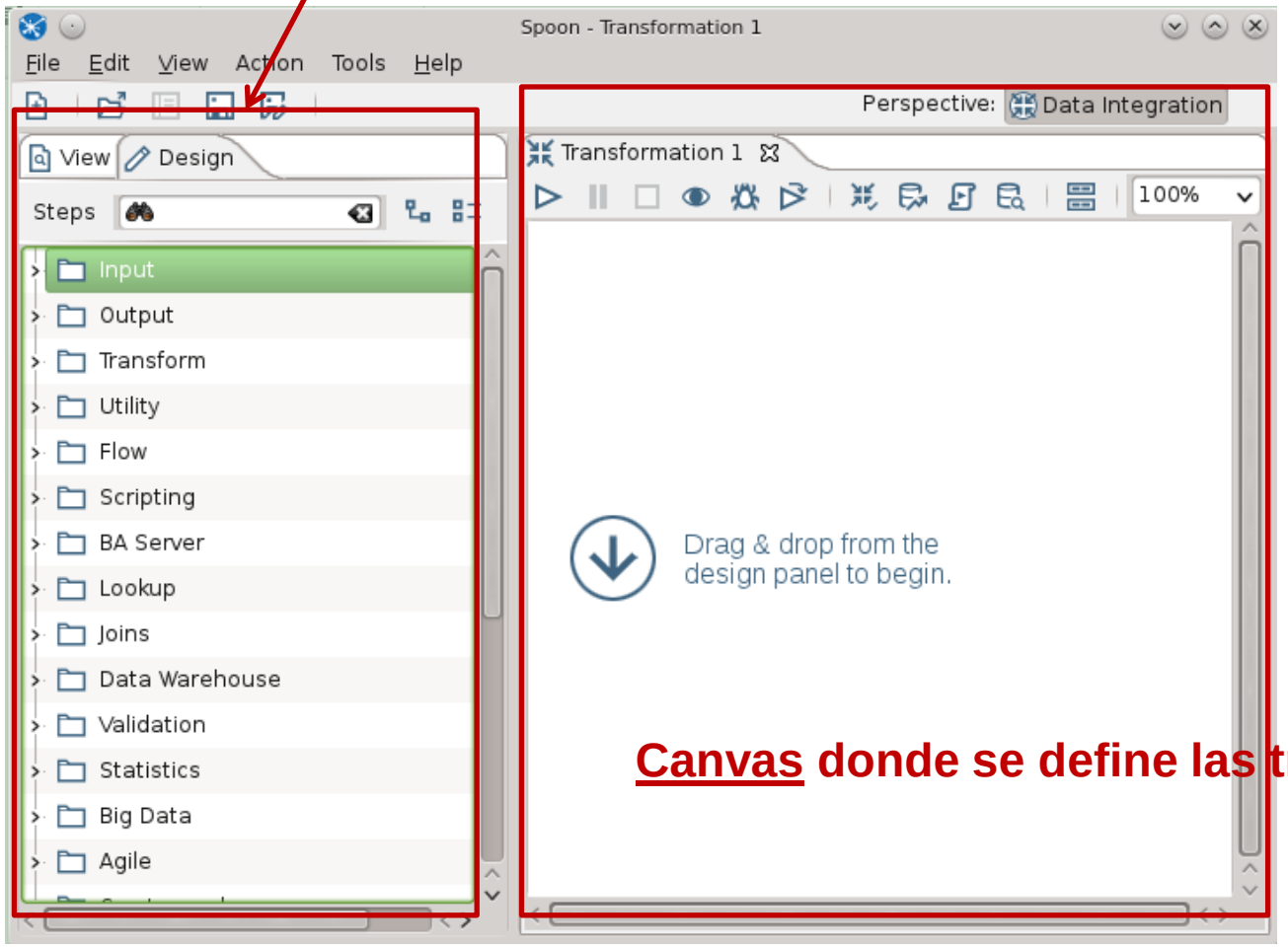
```
<order>
<hop> <from>Entrada Fichero de Texto</from><to>Json Input</to><enabled>Y</enabled> </hop>
<hop> <from>Json Input</from><to>Filtrar filas</to><enabled>N</enabled> </hop>
<hop> <from>Json Input</from><to>Filtrar filas 2</to><enabled>Y</enabled> </hop>
<hop> <from>Generar Filas</from><to>A&#xfffd;adir 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

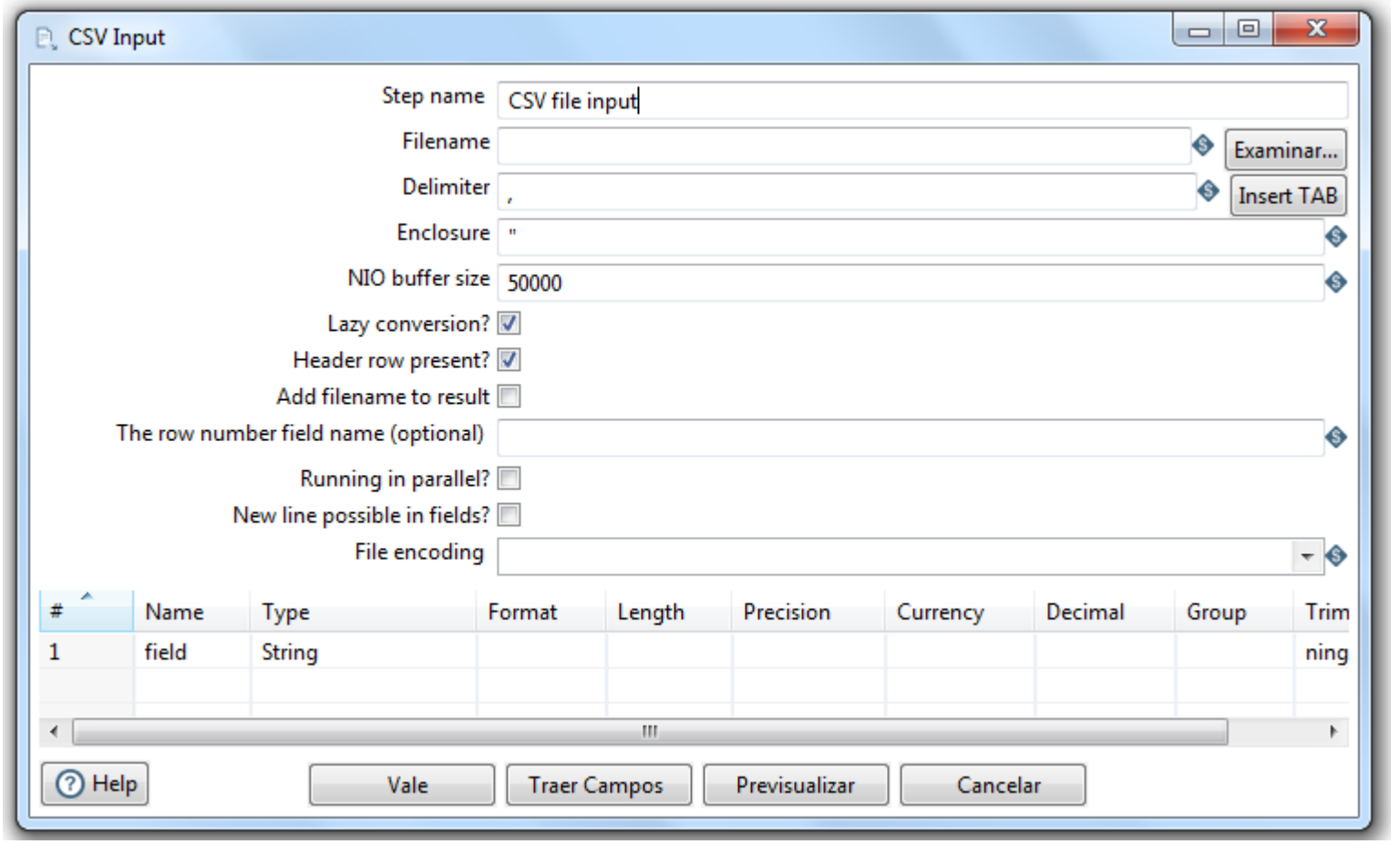
- También se puede **cambiar el nombre** del paso o entrada

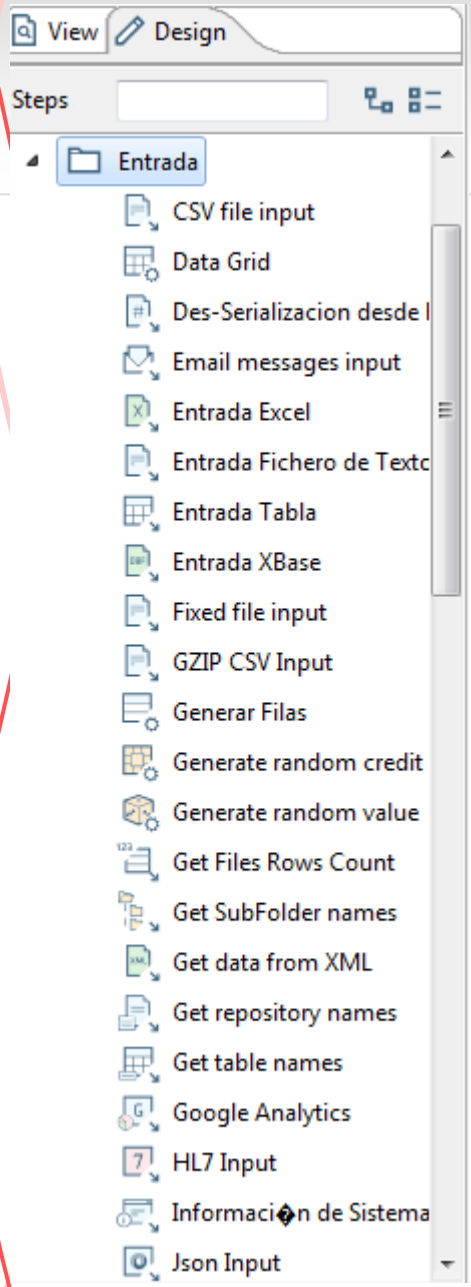
Pentaho Data Integration

Interfaz de Spoon

CSV file input

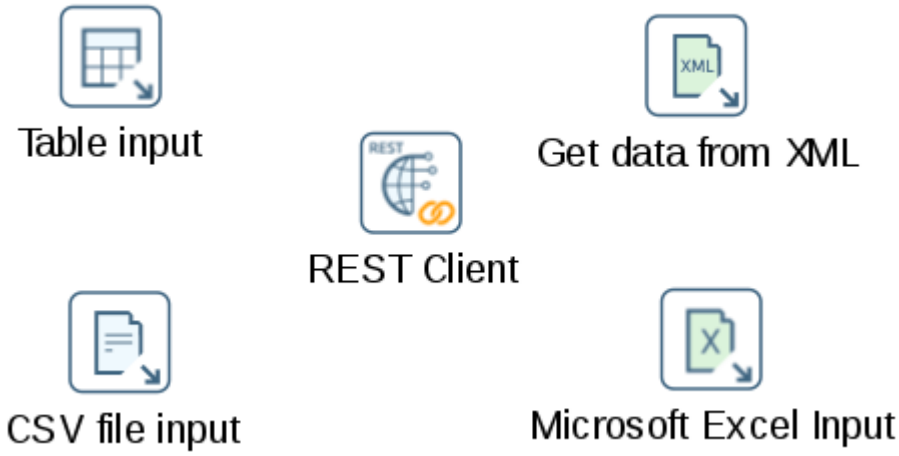
doble clic

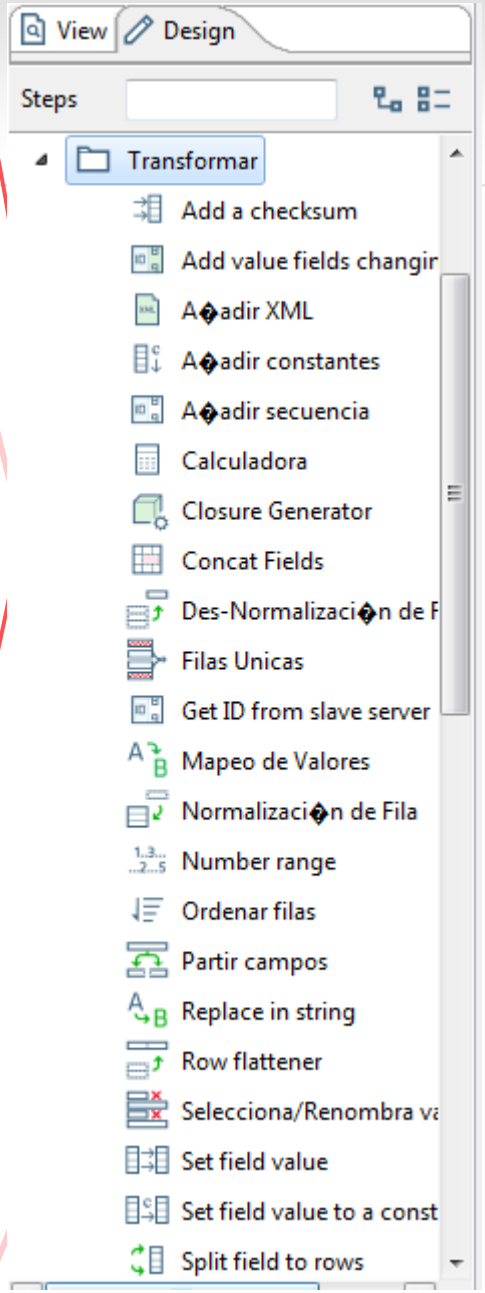




Pentaho Data Integration
Pasos EXTRACCIÓN

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

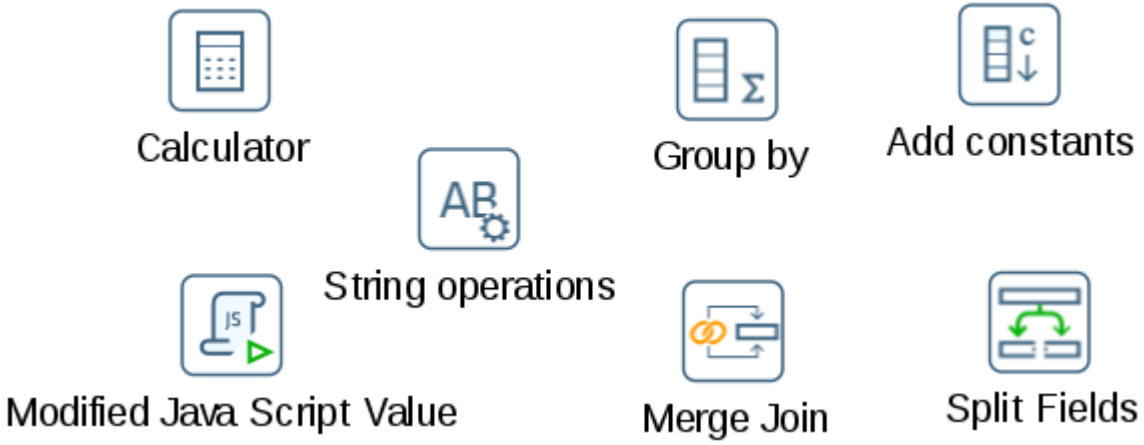


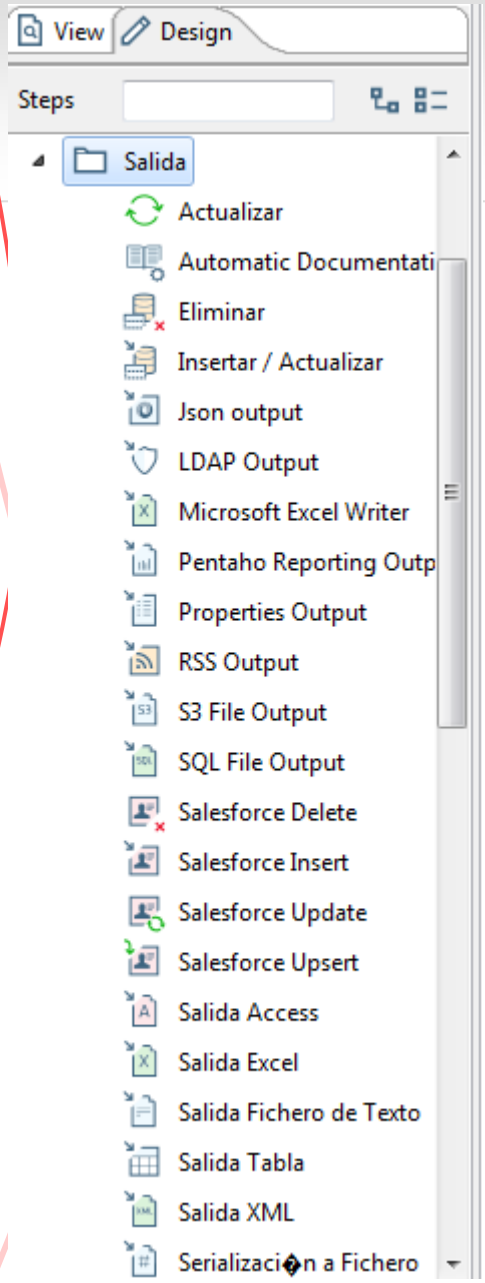


Pentaho Data Integration

Pasos TRANSFORMACIÓN

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

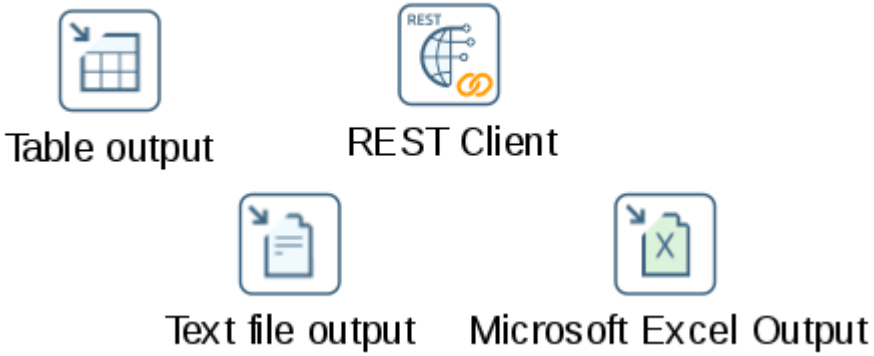




Pentaho Data Integration

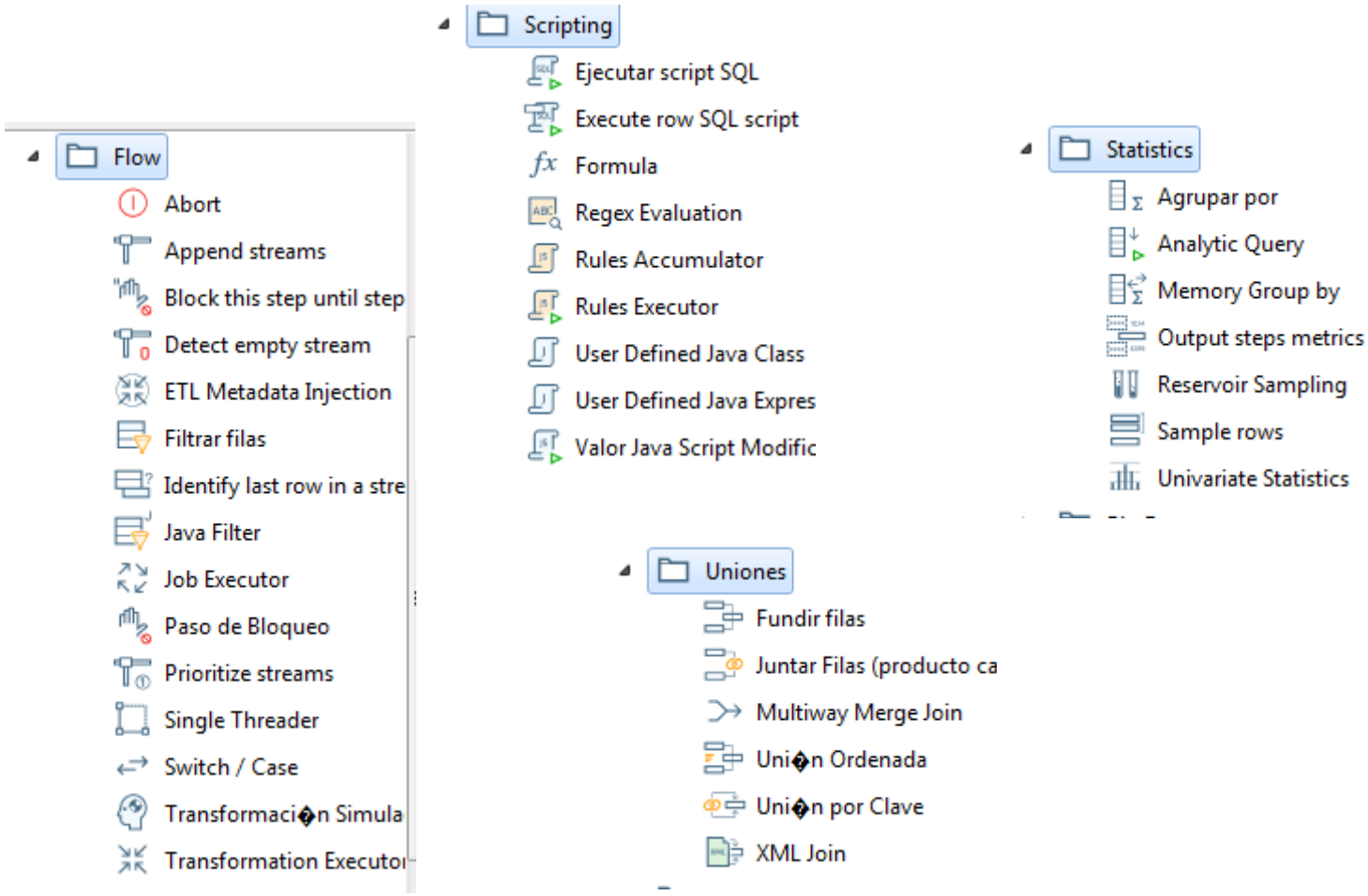
Pasos CARGA

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



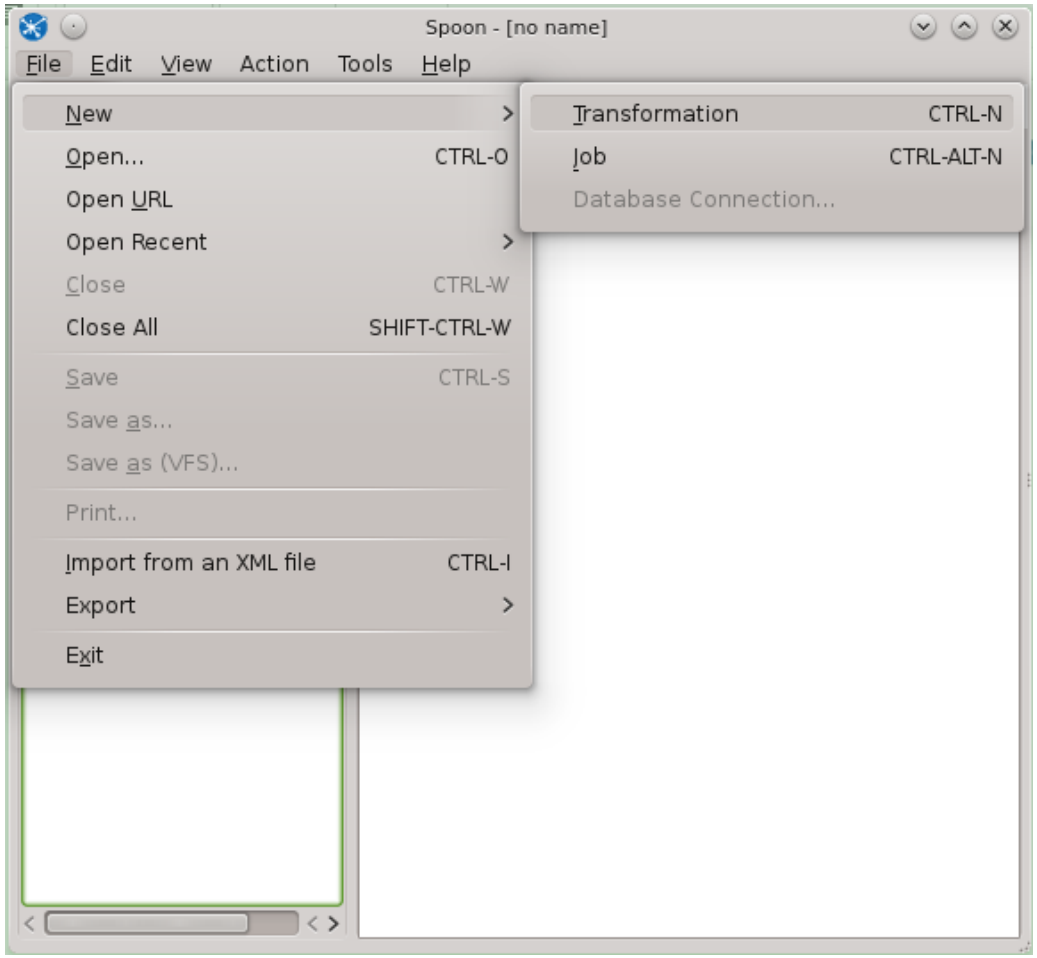
Pentaho Data Integration

Más tipos de pasos

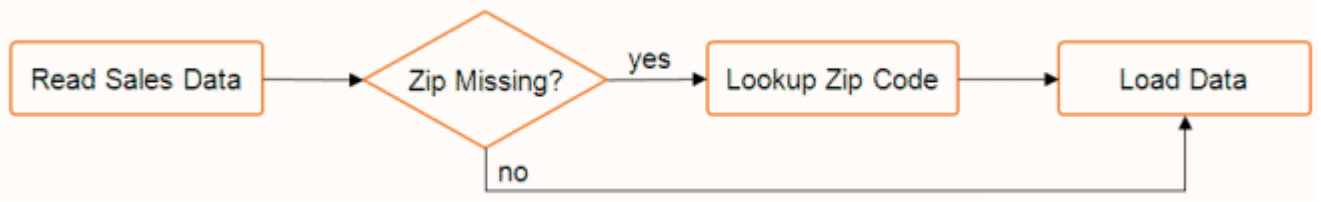


Pentaho Data Integration

Crear nueva transformación

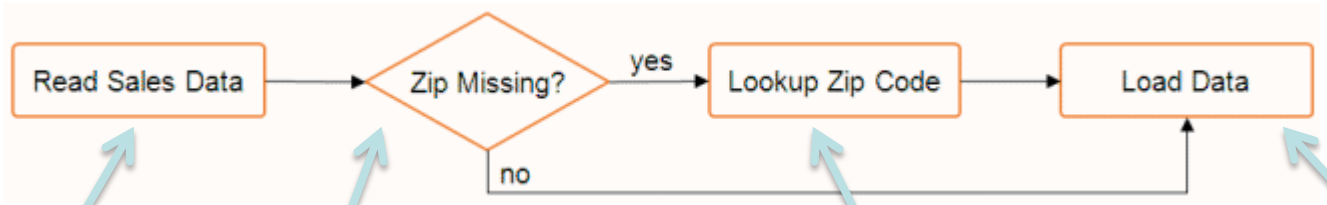


Ejemplo



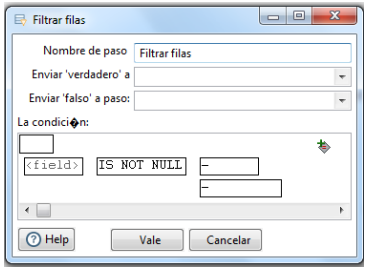
- **Lectura de datos**
  - Desde fichero **samples\transformations\files\sales_data.csv**
- **Filtrado**
  - Determinar si existe código postal
- **Búsqueda de datos**
  - Obtener código postal de **samples\transformations\files\Zipssortedbycitystate.csv**
- **Carga de datos**
  - Fichero MS Excel

Ejemplo



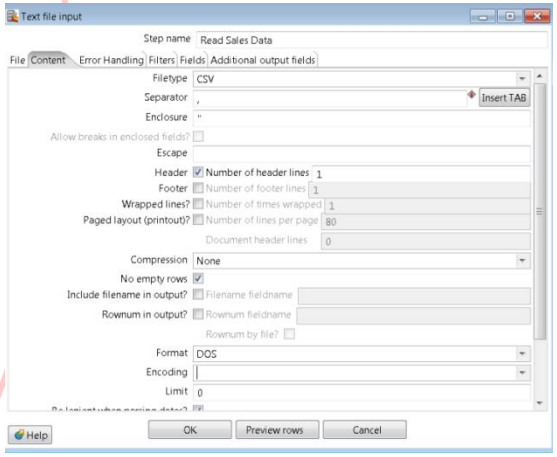
Entrada Fichero de Texto

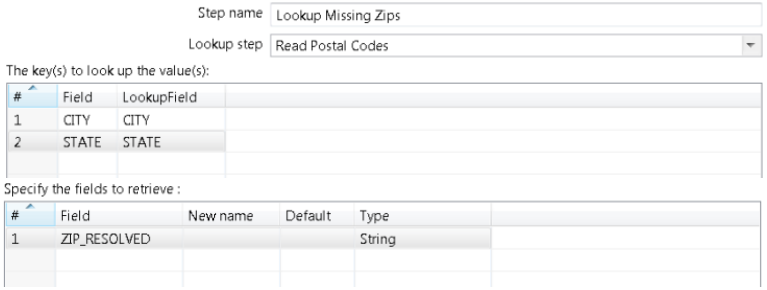
Filtrar filas

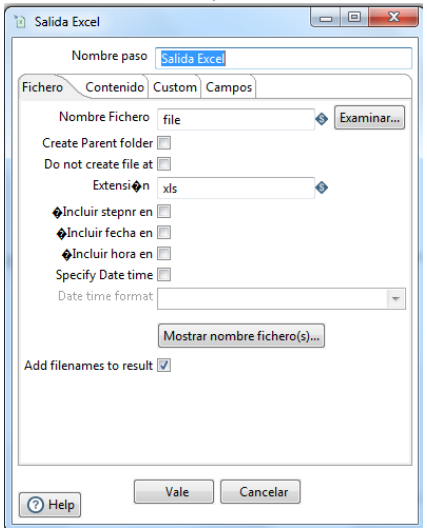


B&#65533;squeda en Flujo de Datos

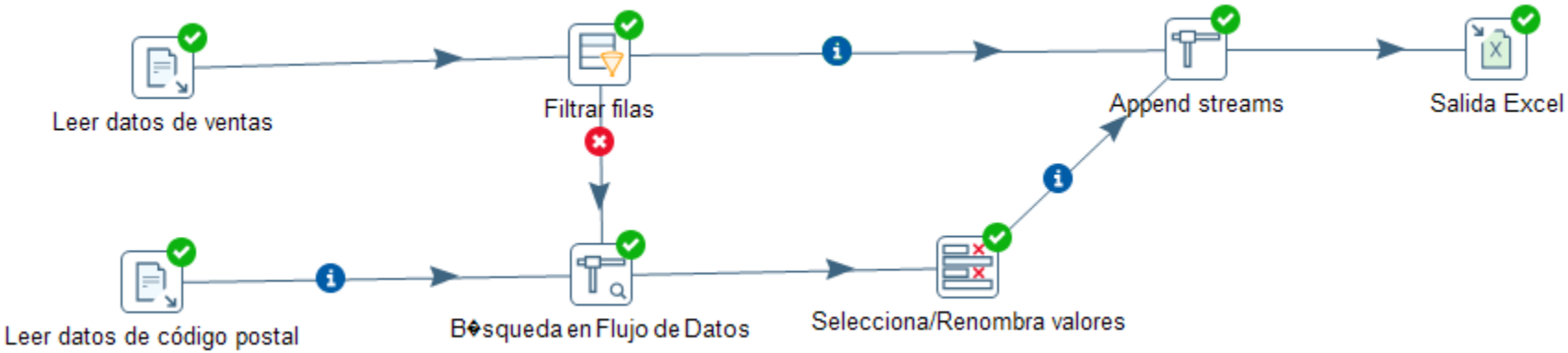
Salida Excel





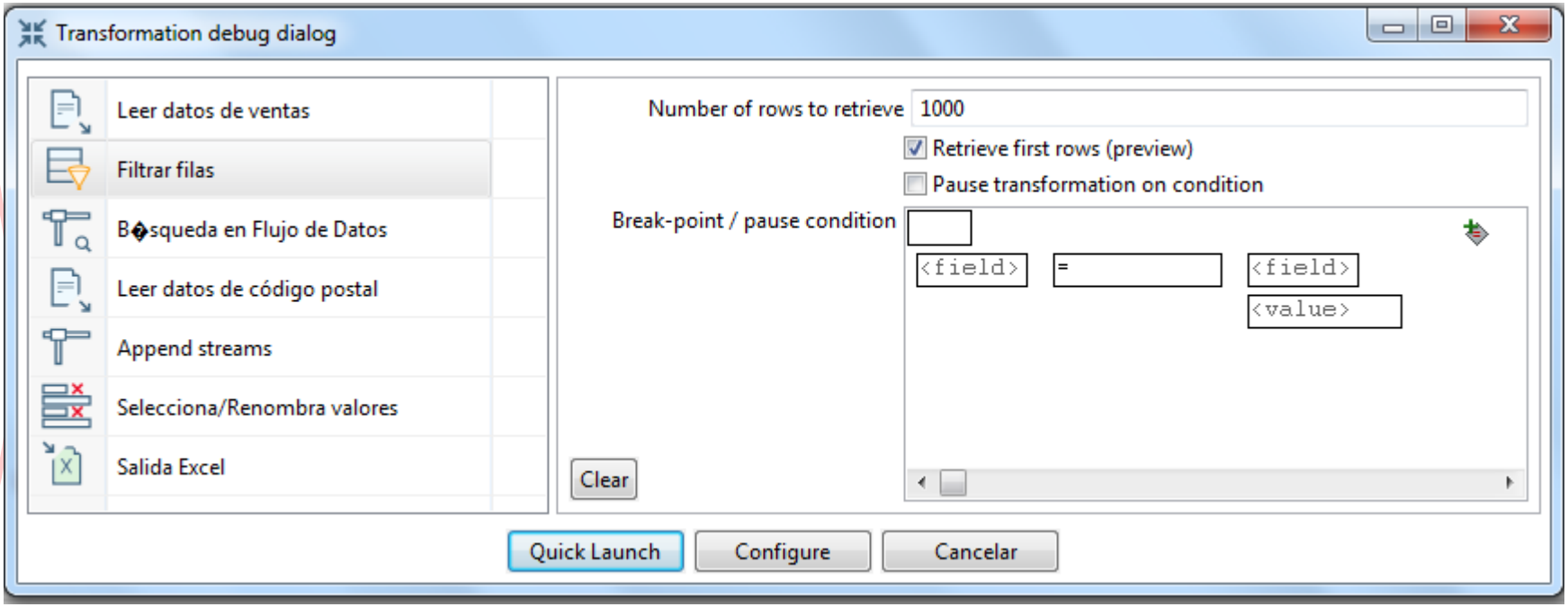


Ejemplo

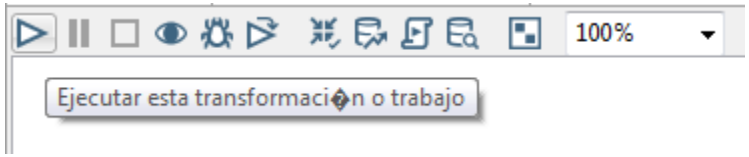

Leer datos de ventas
Filtrar filas
Append streams
Salida Excel
Leer datos de código postal
B♦squeda en Flujo de Datos
Selecciona/Renombra valores

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

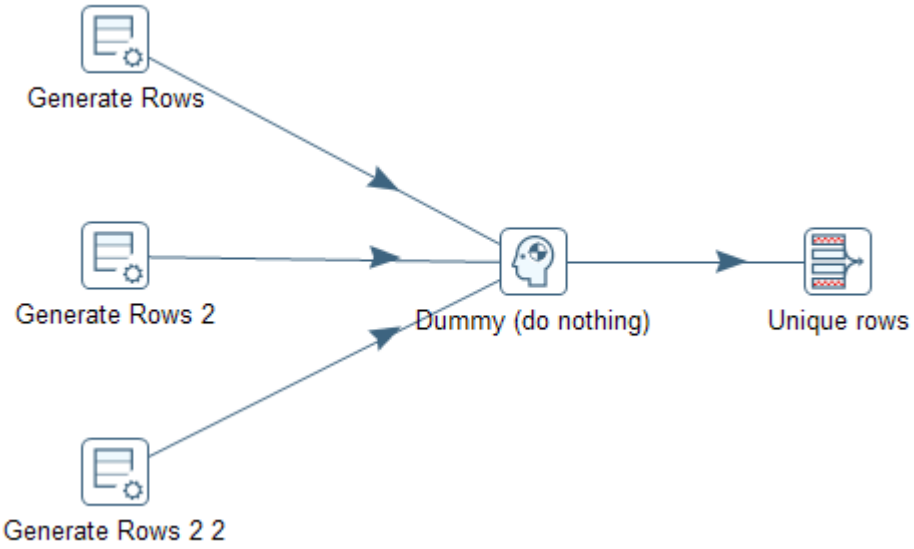
#	Variable	Valor
1	Internal.Job.Filename.Directory	Parent Job File Directory
2	Internal.Job.Filename.Name	Parent Job Filename
3	Internal.Job.Name	Parent Job Name
4	Internal.Job.Repository.Directory	Parent Job Repository Directory

Ejecutar

Cancelar

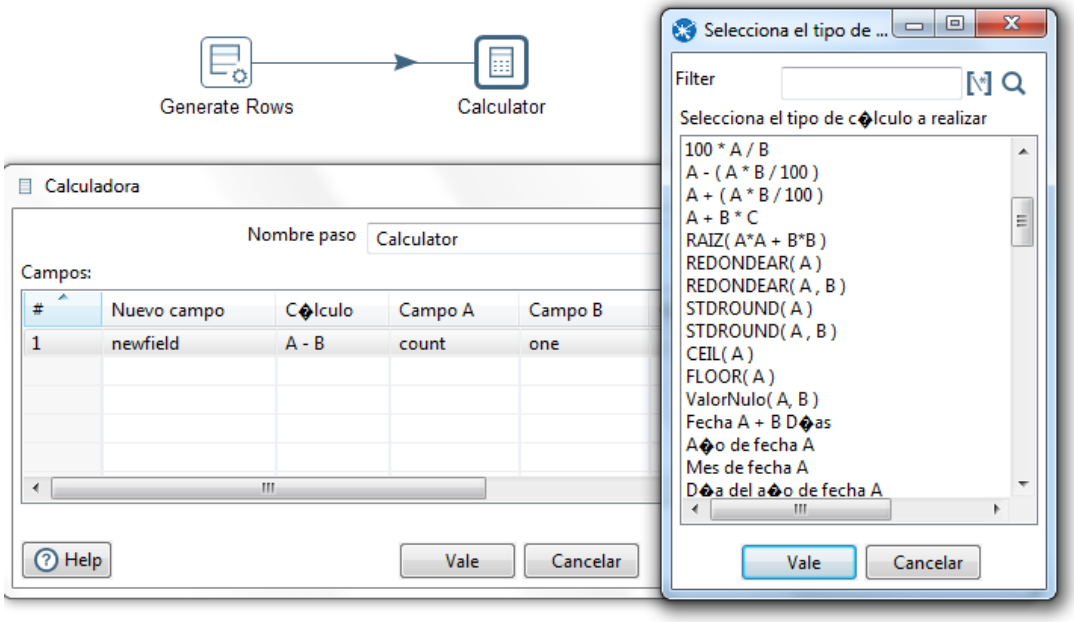
Filas únicas

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



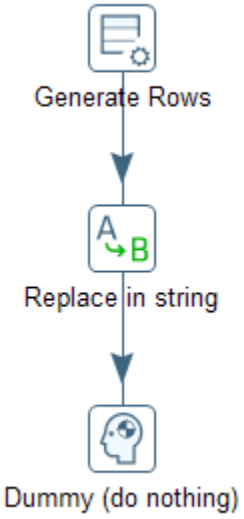
Calculadora

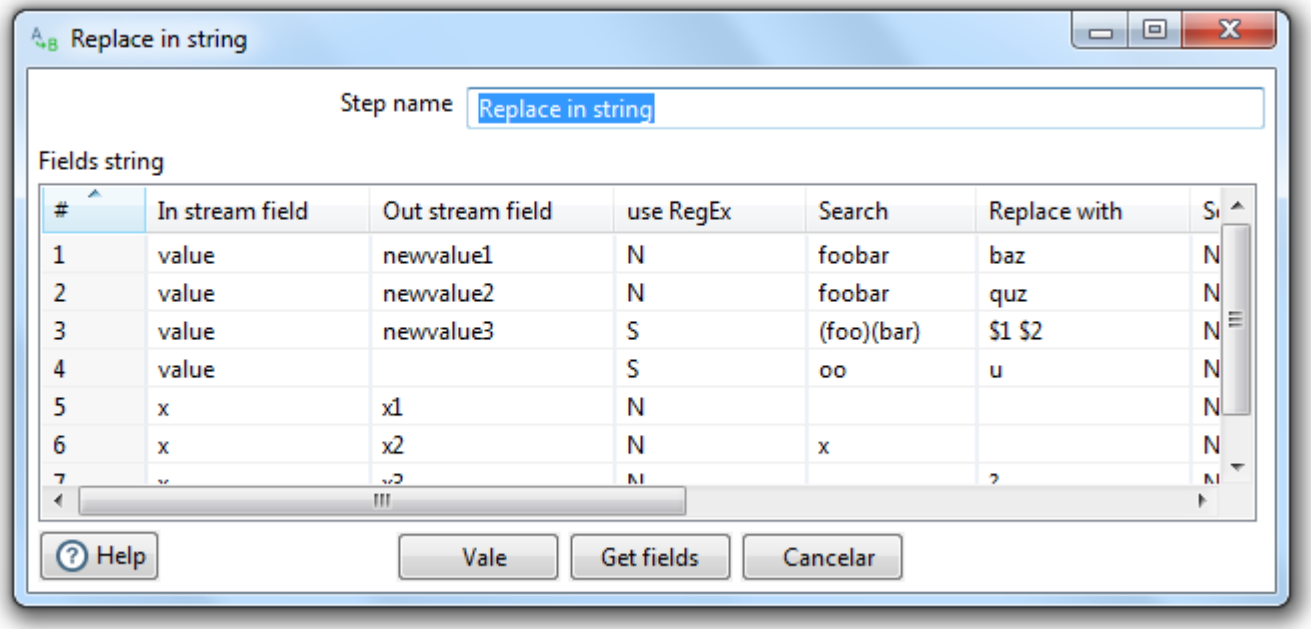
- Suministra funciones predefinidas que pueden ser ejecutadas sobre los campos de entrada
- **samples\transformations\Calculator.ktr**



Reemplazar en cadena de caracteres

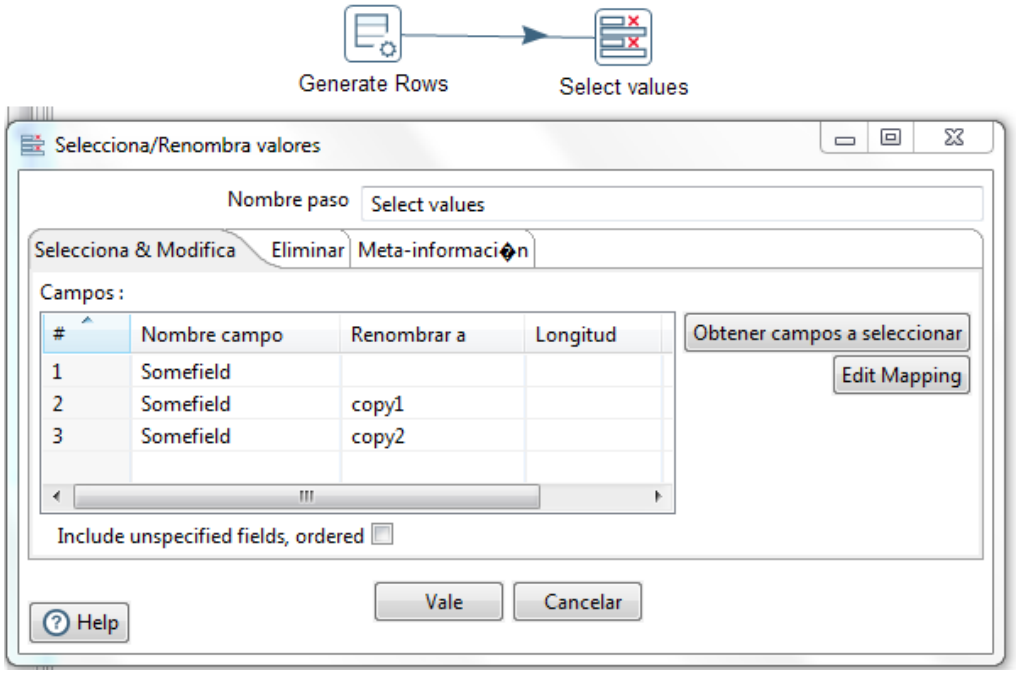
- Permite cambiar unos caracteres por otros
- **samples\transformations\Replace in string - Simple example.ktr**





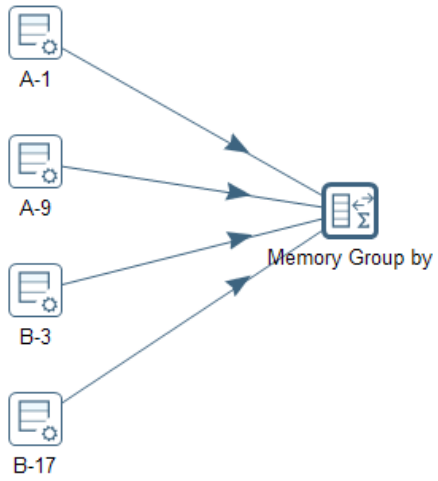
Seleccionar valores

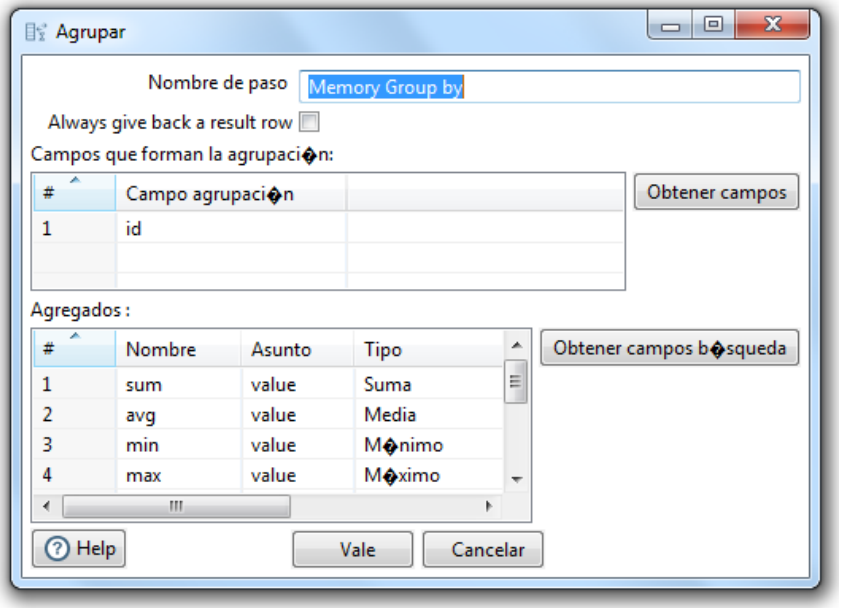
- Obtiene el valor de un subconjunto de campos
- **samples\transformations\Select Values - copy field values to new fields.ktr**



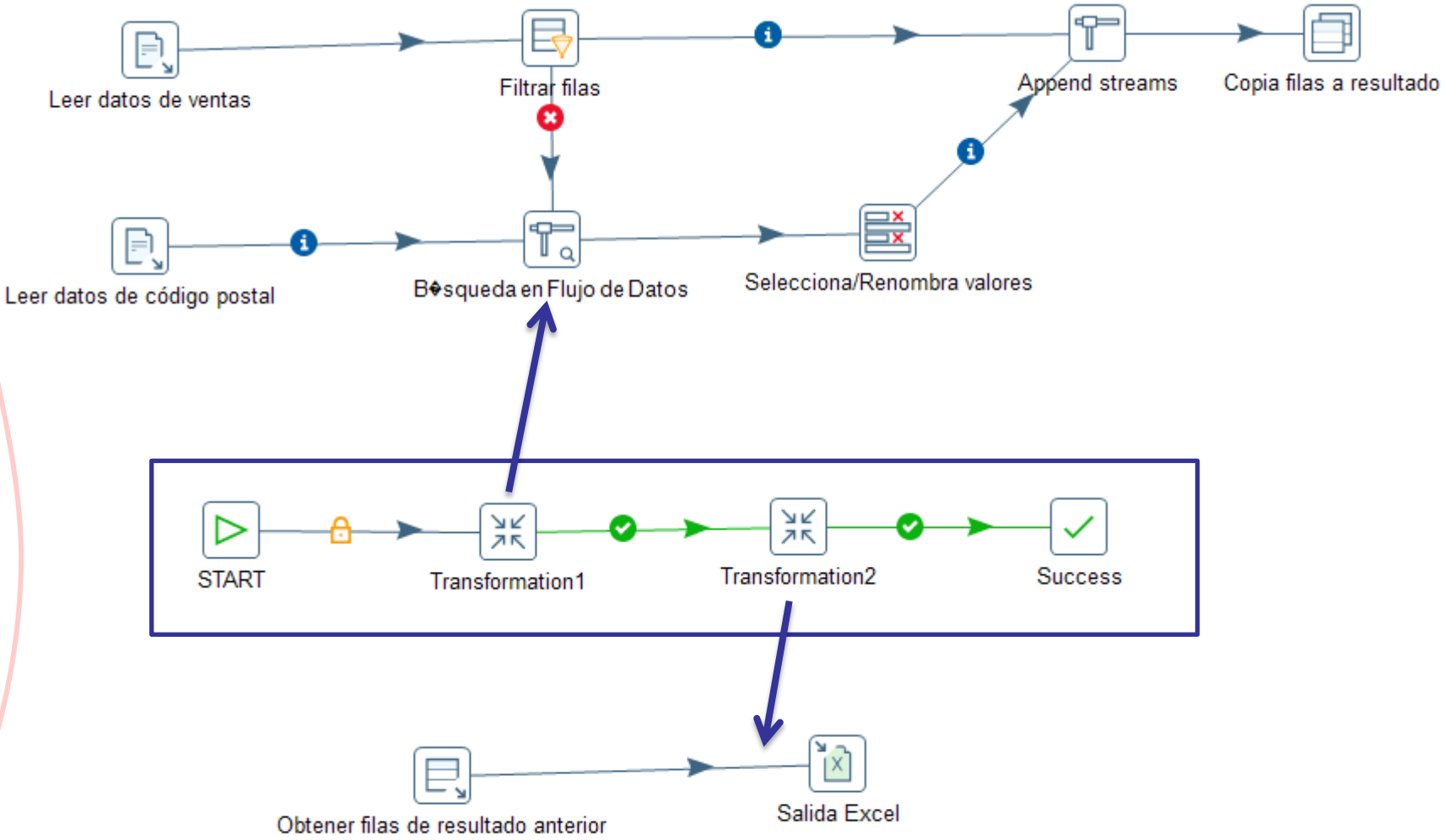
Agrupar

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





Ejemplo de trabajo



Casos de uso

- A partir de datos del portal de datos abiertos de la Generalitat Valenciana

- http://www.dadesobertes.gva.es/

dades Obertes
PORTAL DE TRANSPARÈNCIA DE LA GENERALITAT VALENCIANA

Casos de uso (1)

- Centros educativos 2015/2016
  - http://www.dadesobertes.gva.es/va/dataset/edu-alu-cen-2015-2016
  - Obtener un fichero Excel que contenga una hoja donde por cada población se tengan los siguientes datos
    - Número de centros
    - Número de centros privados
    - Número de centros de secundaria

Casos de uso (2)

- Evolución de centros educativos desde 2008/2009 hasta 2015/2016
  - Obtener un fichero Excel que contenga una hoja donde por cada provincia se tengan la media de centros por población de cada provincia en cada año

Casos de uso (3)

- Resumen de escolarización - 2012-2013
  - http://www.dadesobertes.gva.es/va/dataset/edu-alu-gen-2012-2013

  - Obtener un fichero Excel que contenga una hoja donde por cada provincia se obtengan el centro que más estudiantes tenga y el que menos estudiantes tenga
  - Se debe sustituir “IES” por “Instituto de Educación Secundaria”
  - Se debe sustituir “CEIP” por “Colegio de Educación Infantil y Primaria”

Casos de uso (4)

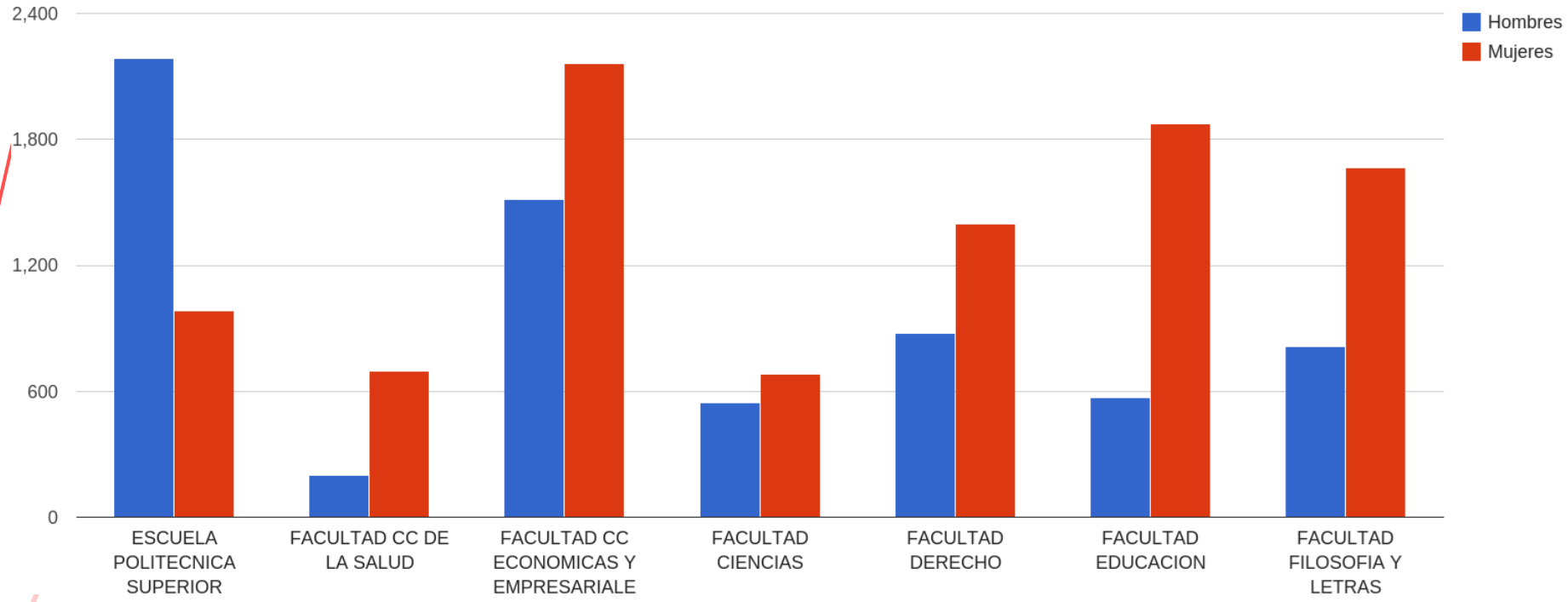
- Obtener un fichero JSON que contenga el número de centros educativos y el número de centros de salud que tengan por cada población de la Comunitat Valenciana

Casos de uso (y 5)

- Fichero CSV con los resultados de las elecciones locales de 2015 por provincias teniendo en cuenta que debe mostrarse los votos por cada partido (y el nombre del partido debe aparecer completo, no sólo las siglas)

Ejercicio

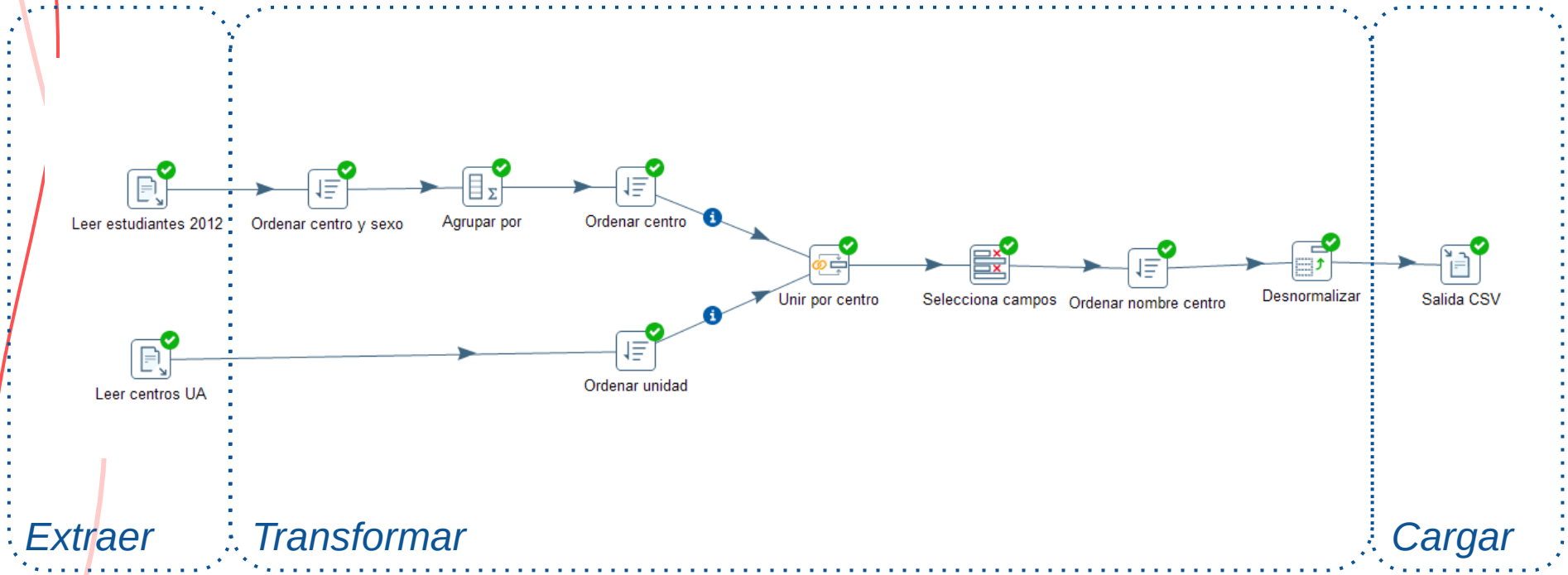
- Número de estudiantes que se matricularon en el 2012 en cada facultad de la UA por género



Ejercicio

- Fuentes de datos
  - Avance de matrícula en estudios de grado - 2012
    - http://datos.ua.es/es/ficha-datos.html?idDataset=868
  - Listado de centros de la UA
    - http://datos.ua.es/es/ficha-datos.html?idDataset=6
- Pasos a realizar
  - Leer archivos CSV
  - Calcular número de estudiantes por centro según sexo
  - Unir conjuntos de datos para obtener el nombre del centro
  - Obtener un archivo CSV
    - Nombre del centro, cantidad de hombres, cantidad de mujeres

Solución


Leer estudiantes 2012
Ordenar centro y sexo
Agrupar por
Ordenar centro
Leer centros UA
Ordenar unidad
Unir por centro
Selecciona campos
Ordenar nombre centro
Desnormalizar
Salida CSV
Extraer
Transformar
Cargar

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

Máster Oficial Universitario en
Gestión de la Información
Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Universitat d'Alacant
Universidad de Alicante

lsi Departamento de Lenguajes y Sistemas Informáticos

wake Web and Knowledge Group
Universidad de Alicante