

# Procesos ETL

Tema 5: Introducción a la integración de datos  
mediante ETLs. Parte II

Adquisición y preparación de datos



Universitat d'Alacant  
Universidad de Alicante

# Pentaho Data Integration

Pentaho suite Business Intelligence para los procesos de gestión de datos.

Open source multiplataforma creada en 2004, plataforma Java.

Disponible en dos ediciones: **Community Edition (CE)** y **Enterprise Edition (EE)**.

Múltiples plugins creado por la comunidad de código abierto.

Pentaho Community wiki

<https://pentaho-public.atlassian.net/wiki/spaces/COM/overview?mode=global>

# Pentaho Data Integration

## Pentaho Data Integration (Kettle)

Entorno gráfico para diseño flujos de datos.

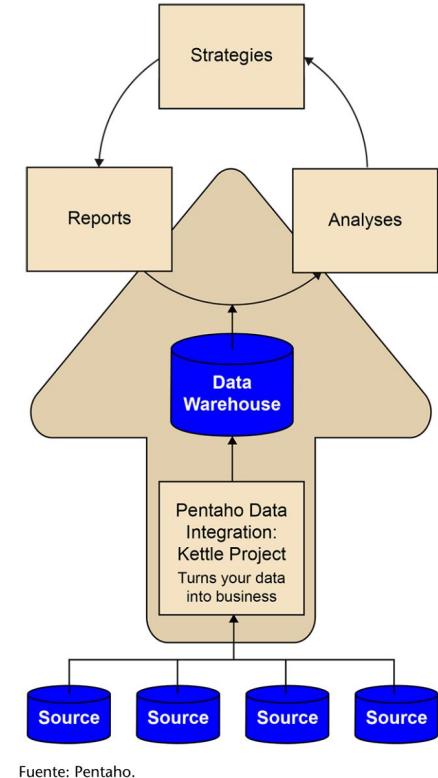
Servicio de integración de datos ETL (extract, transform, load).

Facilita el acceso a una amplia gama de fuentes de datos, incluidas bases de datos de código abierto y propietarias.

Arquitectura extensible mediante plugins y perspectivas

## Algunos usos de Pentaho Data Integration

- Migración de datos entre diferentes sistemas
- Carga de grandes volúmenes de datos
- Limpieza de datos mediante transformaciones
- Integración de datos mediante herramientas ETL en tiempo real



Fuente: Pentaho.

# Pentaho Data Integration

## Componentes clave de Pentaho Data Integration

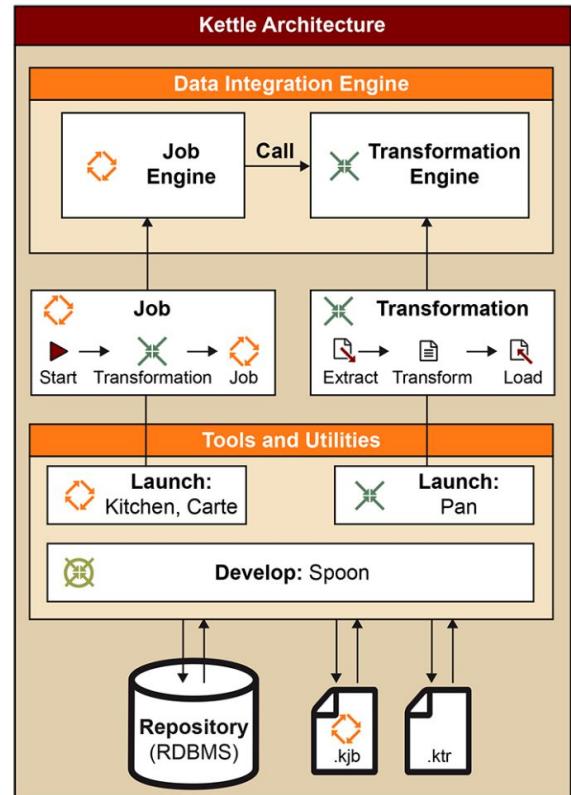
**Spoon:** entorno gráfico para creación de flujos de trabajo, transformaciones y trabajos.

- Realizar data warehouse
- Operaciones, como crear conexiones de datos, transformaciones o inserción de fórmulas, entre otras.
- Múltiples funcionalidades para realizar procesos de ejecución, transformación y carga de datos.

**Pan:** creación líneas de comandos para ejecutar transformaciones

**Kitchen:** ejecución de trabajos por línea de comandos

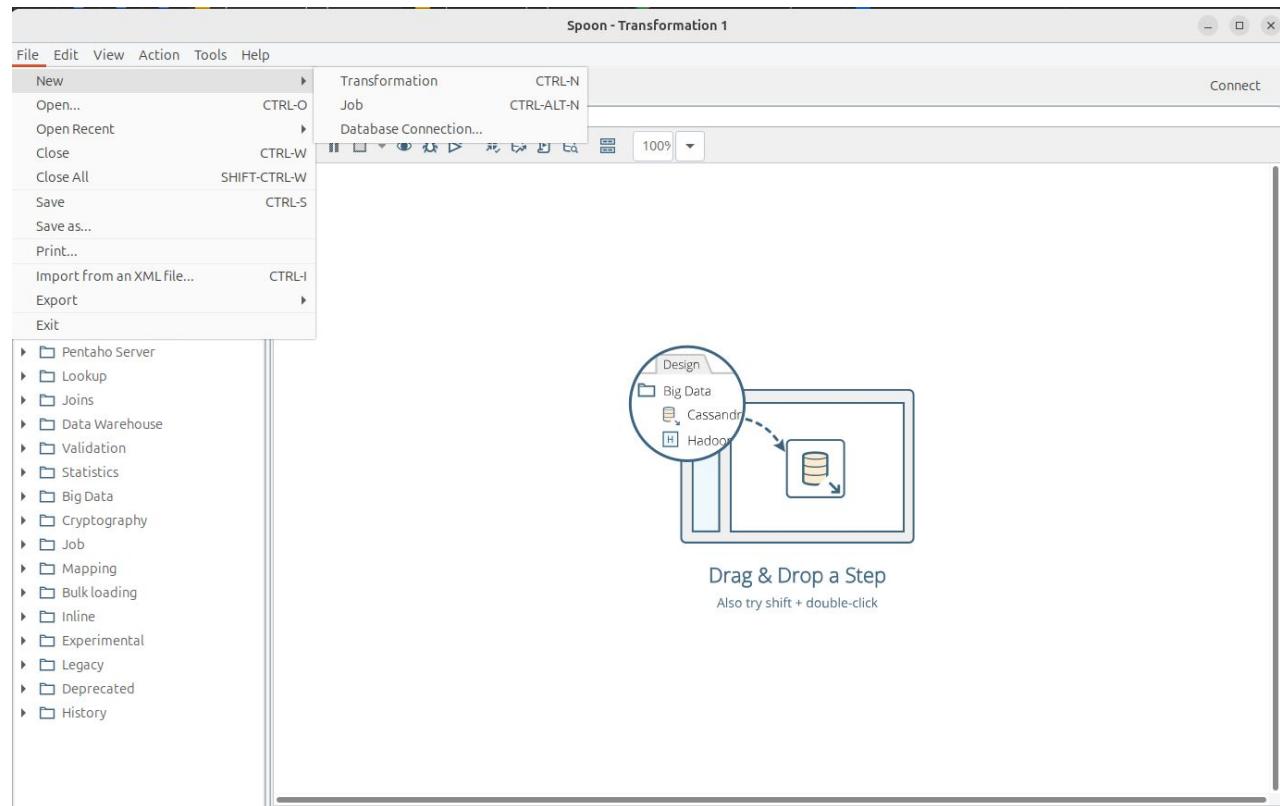
**Carte:** servidor web para ejecutar transformaciones y trabajos



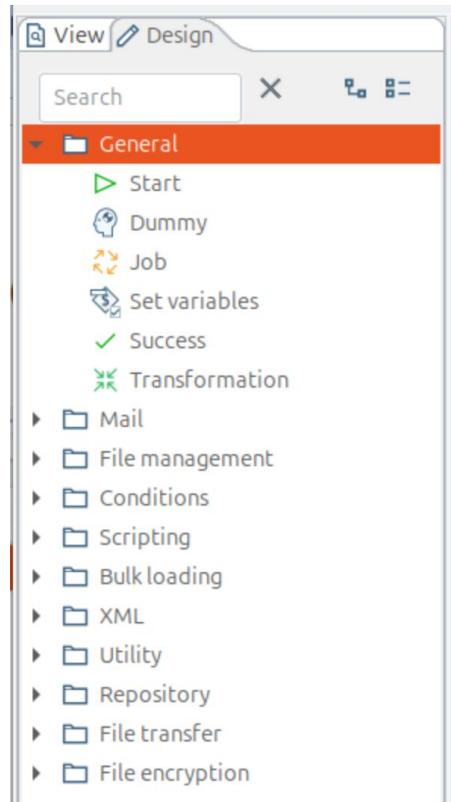
Fuente: Pentaho.

# Pentaho Data Integration

## Entorno visual



# Pentaho Data Integration



## Pasos para el diseño de *jobs*

**General**, pasos de inicio de trabajo, ejecución transformaciones o trabajos entre otras operaciones

**Mail**, envío de correos, recuperación y validación de cuentas

**File management**, gestión de operaciones sobre ficheros y carpetas, HTTP, etc

**Conditions**, comprobaciones sobre conexiones, BD, ficheros, etc

**Scripting**, scripts en JavaScript, Shell y SQL

**Bulk loading**, cargas a MySQL y MSSQL

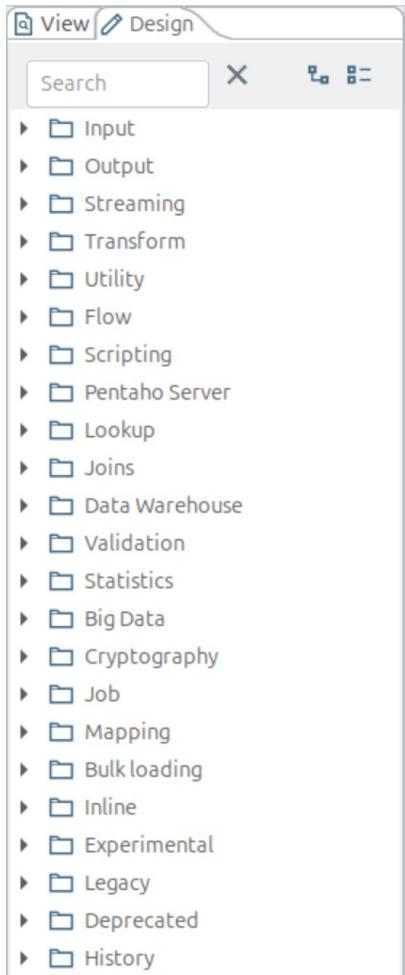
**XML**, validaciones XML, DTD, XSD y XSL

**Utility**, pasos para ejecución de transformaciones como ping, enviar información a syslogs, Nagios, etc

**Repository**, operaciones con el repositorio de transformaciones y trabajos.

**File transfer**, gestión ficheros FTP, FTPS y SFTP

**File encryption** : sistema PGP (Pretty Good Privacy) para envío y recepción de ficheros



# Pentaho Data Integration

## Pasos para el diseño de *transformations*

**Input**, adquisición de datos desde diversas fuentes, BD, CSV, Excel, JSON, repositorios, etc

**Output**, carga datos en diferentes sistemas

**Streaming**, obtener datos de Kafka, JMS (Java Message Service) o MQTT (Standard for IoT Messaging)

**Transform**, ejecución de diversas operaciones con datos, como filtrar, ordenar, partir, añadir nuevos campos, mapear, etc

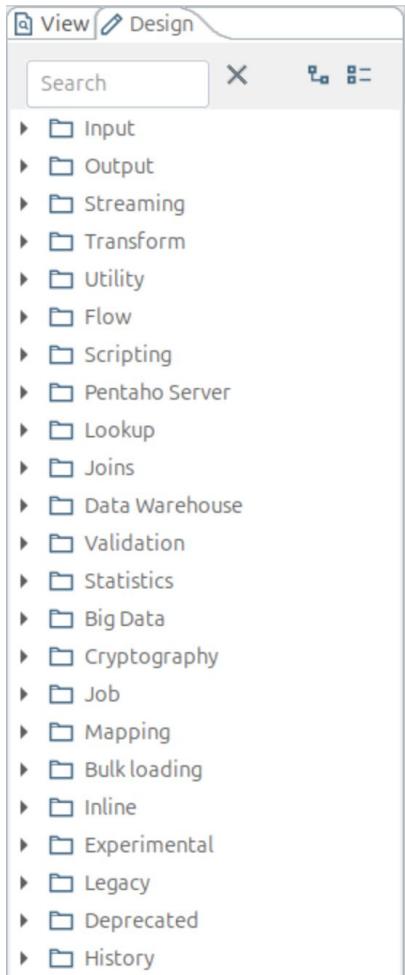
**Utility**, ejecución ssh commands, envío de mensajes syslog, etc

**Flow**, control de flujo de datos en las transformaciones

**Scripting**, SQLScript, JavaScript, expresiones regulares, fórmulas y expresiones java

**Pentaho server**, conexión al servidor de Pentaho para operaciones como la gestión de variables

**Lookup**, pasos para añadir información al flujo de datos mediante búsqueda en fuentes como BD, web services, HTTP, etc



# Pentaho Data Integration

## Pasos para el diseño de *Transformations*

**Joins**, unión de filas en función de diferentes criterios

**Data Warehouse**, para trabajar con dimensiones

**Validation**, validaciones como tarjetas de crédito, datos, email o XSD

**Statistics**, operaciones estadísticas sobre un flujo de datos como group by

**Big Data**, cargar y extraer datos de Avro y MongoDB

**Cryptography**, uso de PGP (Pretty Good Privacy)

**Job**, realizar operaciones propias de un Job

**Mapping**, mapeo entre campos de entrada y salida

**Bulk loading**, cargas desde Vectores, MySQL, Oracle, PostgreSQL entre otras

**Inline**, embebido para operaciones con sockets

**Experimental**, pasos en fase de validación

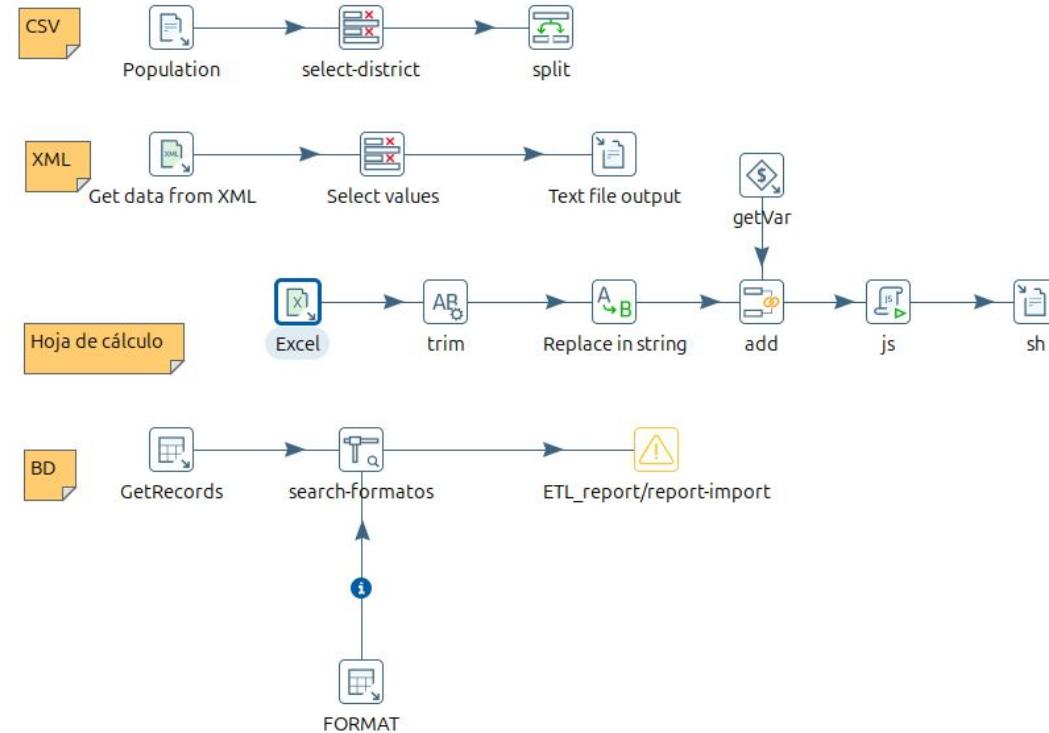
# Pentaho Data Integration

**Extract**, adquisición de datos y distintos conectores para el acceso a las fuentes.

Entrada	
CSV file input	HL7 input
Data grid	Información de sistema
Des-Serializacion desde Fichero	JSON input
ESRI shapefile reader	LDAP input
Email messages input	LDIF input
Entrada Fichero de Texto	Load file content in memory
Entrada Tabla	Microsoft Access input
Entrada XBase	Microsoft Excel input
Fixed file input	Mondrian input
GZIP CSV input	OLAP input
Generar filas	Obtener nombres de fichero
Generate random credit card numbers	Property input
Generate random value	RSS input
Get data from XML	S3 CSV input
Get files rows count	SAS input
Get repository names	Salesforce input
Get subfolder names	XML input stream (StAX)
Get table names	YAML input

# Pentaho Data Integration

Ejemplo 1: conexión a distintas fuentes de datos



Número\_punt\_critic|Codi\_Districte|Nom\_Districte|Codi\_Barri|Nom\_Barri|Nom\_Carrer|Número|Latitud|Longitud|Motiu|Nombre\_visites\_febrer\_2017|  
 Estat\_febrer\_2017|Nombre\_visites\_març\_2017|Estat\_març\_2017|Nombre\_visites\_abril\_2017|Estat\_abril\_2017|Nombre\_visites\_maig\_2017|  
 Estat\_maig\_2017|Nombre\_visites\_juny\_2017|Estat\_juny\_2017|Nombre\_visites\_juliol\_2017|Estat\_juliol\_2017|Nombre\_visites\_agost\_2017|  
 Estat\_agost\_2017|Districte d'alta: Mitjana edat dones|Districte d'alta: Mitjana edat homes|Districte de baixa: Mitjana edat dones|Districte de baixa: Mitjana edat homes|població residente|Índex RFD BCN=100|posició  
 1|1|Ciutat Vella|1|el Raval|Passatge Elisabets|9|41.4|2.2|Mal ús de contenidors i/o papereres|0|No visitat|0|No visitat|11|9.2|20|10.4|19|  
 14.7|21|15|21|22.3|31|32.1|32.4|32.9|47.274|74,6|48  
 2|1|Ciutat Vella|1|el Raval|Carrer de Valldoncella|49|41.4|2.2|Mal ús de la recollida pneumàtica|0|No visitat|0|No visitat|11|16.5|21|6.9|19|  
 16.9|21|11.4|21|15.1|31|32.1|32.4|32.9|47.274|74,6|48  
 3|1|Ciutat Vella|1|el Raval|Carrer d'en Roig|1|41.4|2.2|Efecte "crida"|0|No visitat|0|No visitat|11|7.7|21|9.8|20|11.5|20|11.8|21|9.6|31|32.1  
 32.4|32.9|47.274|74,6|48  
 4|1|Ciutat Vella|1|el Raval|Carrer Cardona|2|41.4|2.2|Mal ús de contenidors i/o papereres|0|No visitat|0|No visitat|11|11.7|20|17.6|20|11.9|  
 21|13.8|21|9.3|31|32.1|32.4|32.9|47.274|74,6|48  
 5|1|Ciutat Vella|1|el Raval|Carrer Robadors |26|41.4|2.2|Efecte "crida"|0|No visitat|0|No visitat|10|6.8|21|5.8|21|8.8|23|5.2|20|7.8|31|32.1  
 32.4|32.9|47.274|74,6|48  
 6|1|Ciutat Vella|1|el Raval|Carrer de la Riera Baixa|26|41.4|2.2|Efecte "crida"|0|No visitat|0|No visitat|11|11.9|21|10.6|20|13|21|10.3|21|  
 7.4|31|32.1|32.4|32.9|47.274|74,6|48  
 7|1|Ciutat Vella|1|el Raval|Plaça Hort de la Bomba|1|41.4|2.2|Mal ús de la recollida pneumàtica|0|No visitat|0|No visitat|14|2.2|22|4.2|21|  
 6.5|21|4|21|3.6|31|32.1|32.4|32.9|47.274|74,6|48  
 8|1|Ciutat Vella|1|el Raval|Carrer de Sant Josep Oriol|13|41.4|2.2|Mal ús de la rec  
 21|4.3|21|2.6|21|1.7|31|32.1|32.4|32.9|47.274|74,6|48

1,"4. Sant Pere, Santa Caterina i la Ribera","32.5","","33.1","","33.1","34.1"  
 2,"5. el Fort Pienc","32.6","33.9","","33.3","34.9"  
 2,"6. la Sagrada Família","33.6",36,"33.2","35.5"  
 2,"7. la Dreta de l'Eixample","36.2","39.6","","34.3","36.7"  
 2,"8. l'Antiga Esquerra de l'Eixample","34.6","36.6","","35.2","35.8"  
 2,"9. la Nova Esquerra de l'Eixample","34.1","35.3","","34","34.6"  
 2,"10. Sant Antoni","34.5","34.9","","33.5",35

```
<collection>
  <record>
    <leader>01481nam a2200385 i 4500</leader>
    <controlfield tag="005">20210908155716.0</controlfield>
    <controlfield tag="007">ta</controlfield>
    <controlfield tag="008">210906s1584 it a r 000 0 ita d</controlfield>
    <controlfield tag="003">ES-GRU</controlfield>
    <controlfield tag="001">991014439313204990</controlfield>
    <datafield tag="040" ind1=" " ind2=" ">
      <subfield code="a">UGR</subfield>
    </datafield>
    <datafield tag="100" ind1="1" ind2=" ">
      <subfield code="a">Cattaneo, Girolamo.</subfield>
    </datafield>
    <datafield tag="245" ind1="1" ind2="0">
      <subfield code="a">Dell'arte militare libro quinto, de gli auvertimenti, et essamini intorno a quelle cose che richiede a
      vn bombardiero, così circa all'Artiglieria, come anco a fuochi artificiati / di M. Girolamo Cataneo nouarese</subfield>
    </datafield>
    <datafield tag="260" ind1=" " ind2=" ">
      <subfield code="a">In Brescia :</subfield>
      <subfield code="b">appresso Thomaso Bozzola,</subfield>
      <subfield code="c">1584</subfield>
    </datafield>
    <datafield tag="300" ind1=" " ind2=" ">
      <subfield code="a">[2], 39, [1] h. :</subfield>
      <subfield code="b">il. ;</subfield>
      <subfield code="c">49</subfield>
    </datafield>
    <datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Marca tip. en port.</subfield>
    </datafield>
    <datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Colofón</subfield>
    </datafield>
```

# Extracción de datos

# Extracción de datos

1. Extracción basada en bases de datos
2. Extracción basada en archivos
  - a. VFS
  - b. Trabajar con CSV/Excel
  - c. Trabajar con archivos XML
3. Extracción basada en web
  - a. Servicios web
  - b. Cliente HTTP
  - c. Extracción web basada en texto

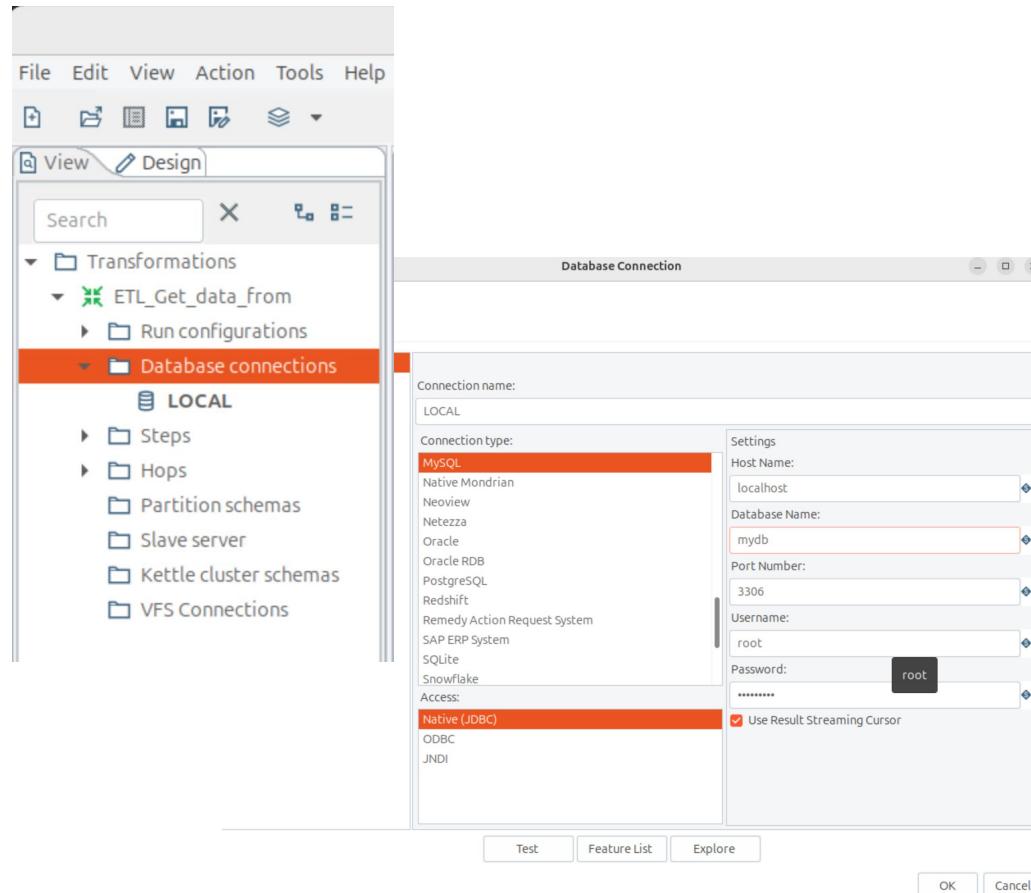
# Extracción desde BD

## *Database Connection*

Son utilizadas tanto por transformations como jobs.

La conexión con el RDBMS se establece en tiempo de ejecución.

Definir conexiones por sí solas no abre una conexión real a la base de datos.



# Extracción desde BD

## Database Connection

The screenshot shows a configuration dialog for a database connection, likely from a tool like MySQL Workbench. The dialog is divided into several sections:

- General Tab:** Contains basic connection parameters and a "Test" button.
- Advanced Tab:** Contains options for identifiers:
  - Supports the boolean data type
  - Supports the timestamp data type
  - Quote all in database
  - Force all to lower case
  - Force all to upper case
  - Preserve case of reserved words
- Options Tab:** Contains parameters:

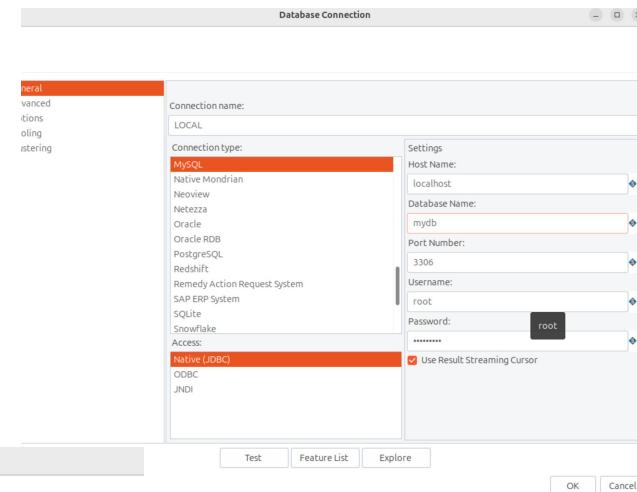
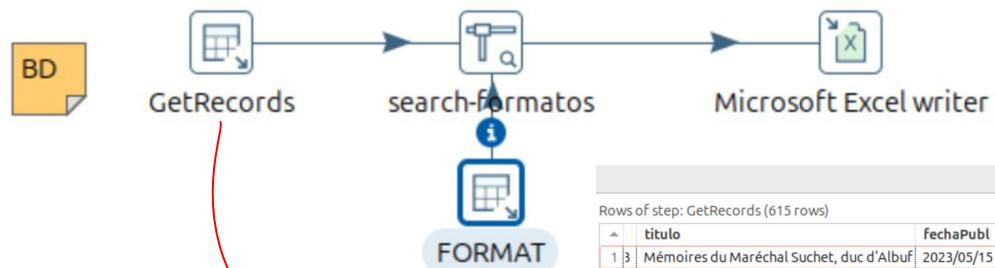
Parameter	Value
defaultFetchSize	500
useCursorFetch	true
- Pooling Tab:** Contains settings for connection pooling:
  - Enable Connection Pooling
    - Initial: [Input Field]
    - Maximum: [Input Field]
    - Pool Size: [Input Field]
  - Parameters:

Name	Value
defaultAutoCommit	true
defaultReadOnly	
defaultTransactionIsolation	
defaultCatalog	
- Clustering Tab:** Contains settings for clustering:
  - Enable Clustering
  - Parameters:

Partition	Host Na	Port	Databas	User Na	Passwor

At the bottom, there are "Test", "Feature", and "Explore" buttons, and "OK" and "Cancel" buttons.

# Extracción desde BD



	título	fechaPubl	lugarPublicacionOriginal
1	Mémoires du Maréchal Suchet, duc d'Albufera	2023/05/15	Anselin, 1834
2	Las mujeres en la prehistoria	2023/05/15	Museu de Prehistòria, 2008
3	Les dones en la prehistòria	2023/05/15	Diputació Provincial de València
4	Libre de memories de diuersos successos en el Comte Dío	2023/05/15	1600
5	[Còpia de documents relatius al Comte Dio]	2023/05/15	1326
6	Otro título: Memorial y proyecto	2023/05/15	1769
7	Executoria del pleito seguido en el Consejo	2023/05/15	1795
8	Recopilación verificada por Felipe Mateu, c.	2023/05/15	1800
9	Dissertación histórica sobre el Pontificado	2023/05/15	1740
10	Memoria histórica sobre el Origen y Visita de	2023/05/15	1880
11	Carta de Sento Beseroles a Pascual Valote	2023/05/15	1845
12	Carta de Bernat i Baldoví a D. Jose Brusca,	2023/05/15	1840
13	In Aristot. logicam commentarius...	2023/05/15	apud, Ioannem Bap. Marçal,
14	Academias morales de las musas...	2023/05/15	por Claudio Macè..., 1647
15	Segunda peregrinacion del dotor D. Pedro	2023/05/15	por layme de Bordazar, 1697
16	Glorias y excelencias del Gran Patriarca Sar...	2023/05/15	1698
17	Las obras de Publio Virgilio Maron...	2023/05/15	en la imprenta de Vicente Ca...
18	Vida del hermano Iuan Berchemans flamenco	2023/05/15	por la viuda de Iuan Crisostomo
19	De morborum et symptomatum differentiis	2023/05/15	typis & expensis Michaelis So...
20	Discursos predicables en las festividades de	2023/05/15	por Iuan Chrysostomo Garriz...
21	El Pastor de Noche Buena : practica breve	2023/05/15	por Silvestre Esparsa..., 1646

# Extracción desde BD

## Conexiones y transacciones

La conexión solo se utiliza durante la ejecución de un *job* o *transformation*.

En un job, cada entrada abre y cierra las conexiones de forma independiente.

En una transformación, debido a la naturaleza paralela, cada paso que utiliza una conexión, se abrirá una conexión a BD independiente e inicia una transacción independiente.

Generalmente ofrece excelente rendimiento para la mayoría de los casos, pero puede causar problemas de bloqueo e integridad referencial cuando se utilizan diferentes pasos de actualización en la misma tabla.

- Para evitar este problema se puede habilitar la opción "*Make transformation database transactional*", de forma que se establece una única conexión a esa BD.
- Se ejecuta un COMMIT cuando la transformación finaliza de manera correcta y un ROLLBACK en otro caso

# Extracción basada en archivos

## *Virtual File Systems*

La gestión flexible y uniforme de archivos es fundamental para cualquier herramienta ETL.

Especifica archivos como URL.

Utiliza backend VFS de Apache Commons que se encargará de la complejidad.

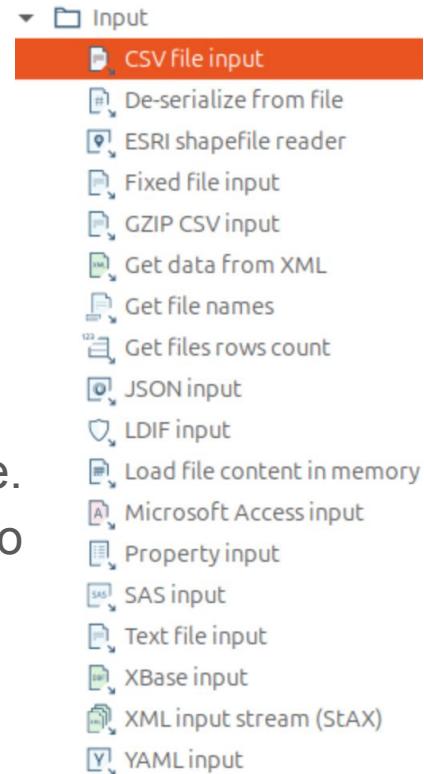
Por ejemplo, con Apache VFS, es posible procesar una selección de archivos dentro de un archivo .zip exactamente de la misma manera que procesaría una lista de archivos en una carpeta local. Para más información sobre cómo especificar archivos VFS,

# Extracción basada en archivos

Nombre del archivo	Descripción
Nombre del archivo: /data/input /clientes.dat	Este archivo se define utilizando el método clásico (no VFS) y se encontrará y leer como tal.
Nombre del archivo: file:///data /input/clientes.dat	Este mismo archivo se leerá desde el sistema de archivos local mediante Apache Controlador VFS.
Trabajo: http://www. hervidor.be /GenerateRows.kjb	Este archivo se puede cargar en Spoon, ejecutar usando Kitchen y referenciar En la entrada de trabajo. Cada vez que se carga el archivo XML de forma transparente. desde el servidor web.
Carpetas: zip:file:///C: /entrada/datosdeventa.zip Comodín: *.txt\$	Esta combinación de carpeta/comodín se puede ingresar en pasos como "Archivo de texto entrada." La especificación del comodín buscará y leerá todos los archivos en el archivo zip especificado que termina en .txt.

# Extracción basada en archivos

- Fáciles de gestionar con una herramienta ETL.
- Las operaciones R/W disponibles como pasos de transformación.
- La gestión de archivos está disponible en los pasos de *job*  
**File management:** mover, copiar, crear, eliminar, comparar, comprimir y descomprimir.
- ***Text file output*** crea el archivo automáticamente si no existe.
- Tipos: ***Delimited*** (CSV recomendado) y ***Fixed*** (cada campo o columna tiene un ancho y una longitud designados).
- Importante configurar la codificación correcta, se puede seleccionar (UTF8 es más o menos el estándar).

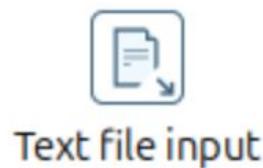


# Extracción basada en archivos

La entrada más básica es **CSV file input**. Permite gestionar un fichero y es necesario especificar el delimitador.

**Text file input**, versión más potente y completa de **CSV file input** y **Fixed file input**:

- Puede obtener el nombre del archivo del paso anterior
- Leer varios archivos en una sola ejecución
- Leer desde archivos comprimidos .zip o .gzip
- Mostrar el contenido del archivo de datos sin especificar la estructura
- Especificar carácter de escape
- Gestionar errores
- Filtrar
- Especificar formato de fecha local



Text file input

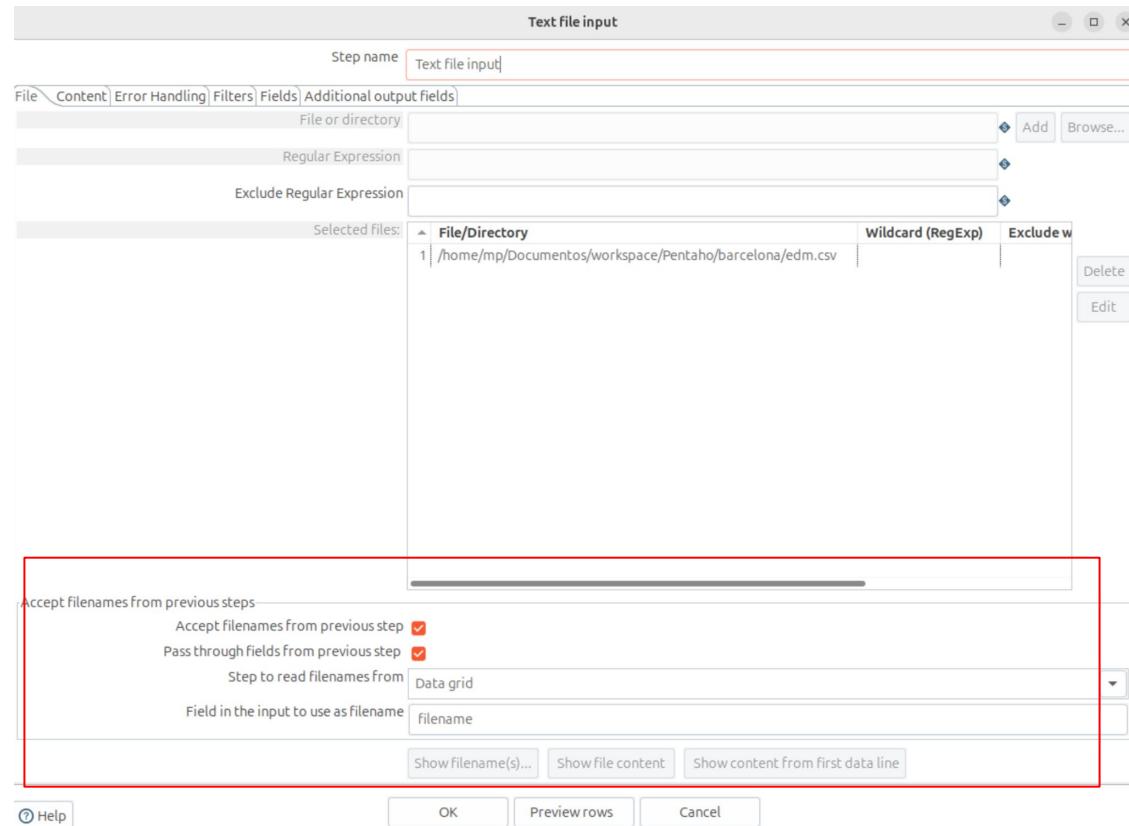


# Extracción basada en archivos

## *Text file input*

Nombre del fichero lo recoge  
del paso anterior

Indicamos el nombre del paso  
y el nombre del campo

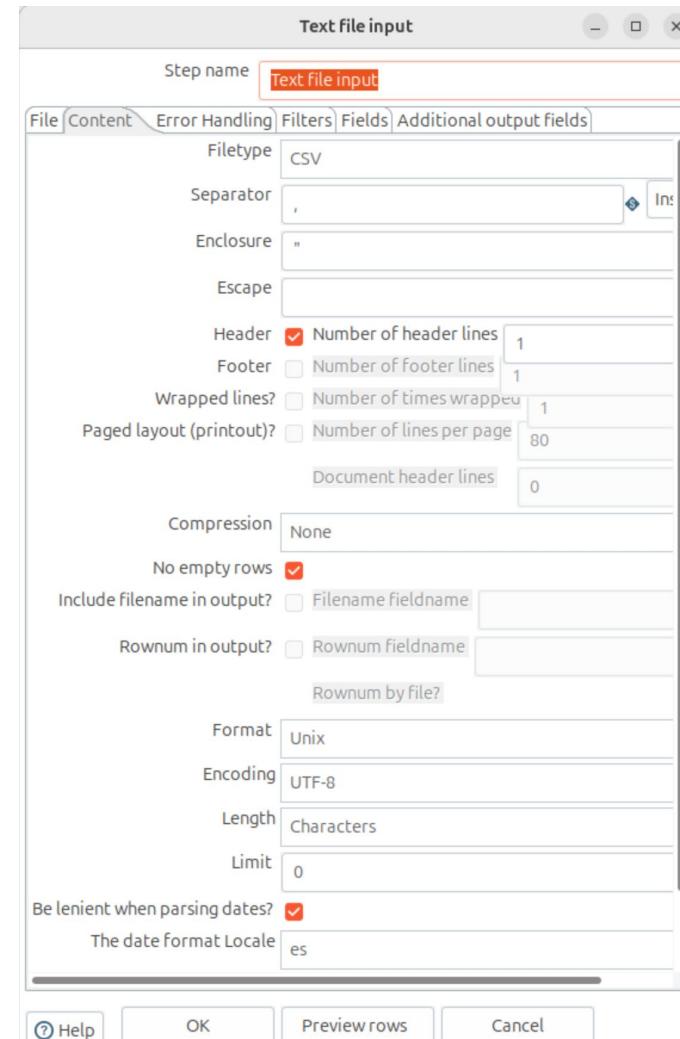


# Extracción basada en archivos

## *Text file input*

Específica:

- Tipo de fichero: CSV
- Carácter de escape: ,
- Formato: Unix
- Encoding: UTF-8
- Especificar formato de fecha local: es



# Extracción basada en archivos

## *Text file input*

Selección de campos con los que trabajar de este archivo

The screenshot shows the 'Text file input' step configuration in the Pentaho Data Integration interface. The 'Step name' is 'Text file input'. The 'Fields' tab is selected, displaying the following field definitions:

Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default	Trim type	Repeat
1 Dte.	String			0		\$	.	,	-		none	N
2 Barri_de_baixa	String			9		\$	.	,	-		none	N
3 Districte_d'alta:_Mitjana_edat_homes	String			48		\$	.	,	-		none	N
4 Districte_d'alta:_Mitjana_edat_dones	BigNumber	#.#		4	1	\$	.	,	-		none	N
5 Districte_de_baixa:_Mitjana_edat_homes	BigNumber	#.#		4	1	\$	.	,	-		none	N
6 Districte_de_baixa:_Mitjana_edat_dones	String			0		\$	.	,	-		none	N

The screenshot shows the preview of the extracted data from the 'Text file input' step. The preview shows 74 rows of data with the following columns:

filename	Dte.	Barri_de_baixa	Districte_d'
/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	BARCELONA	<null>	33.2
/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	1	1. el Raval	32.1
/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	1	2. el Barri Gòtic	33.9
/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	1	3. la Barceloneta	34.6
/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	1	4. Sant Pere, Santa Caterina i la Ribera	32.5
/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	2	5. el Fort Pienc	32.6

# Extracción basada en archivos

Uso del paso CSV input file

Particular de este tipo de fichero pero más limitado



Uso del paso Text input file

Más genérico y potente que los específicos CSV y Fixed



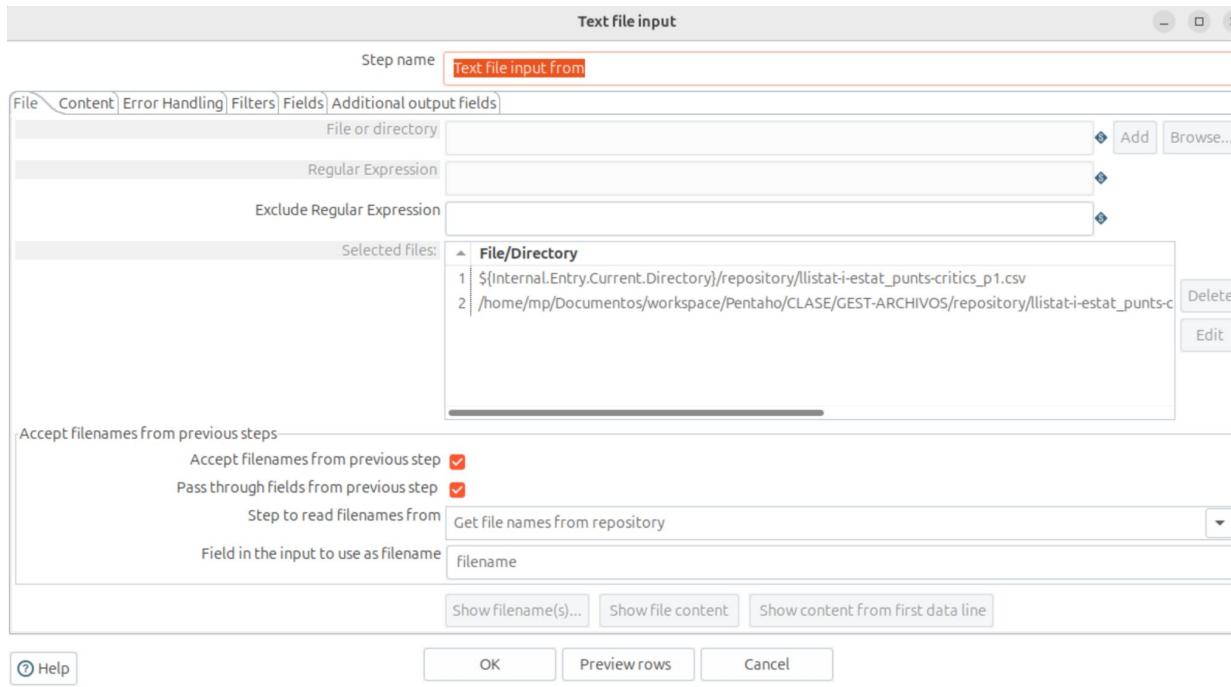
# Extracción basada en archivos

*Reading multiple files in a single run*



# Extracción basada en archivos

*Reading multiple files in a single run*



# Extracción basada en archivos

## ***Working with XML files***

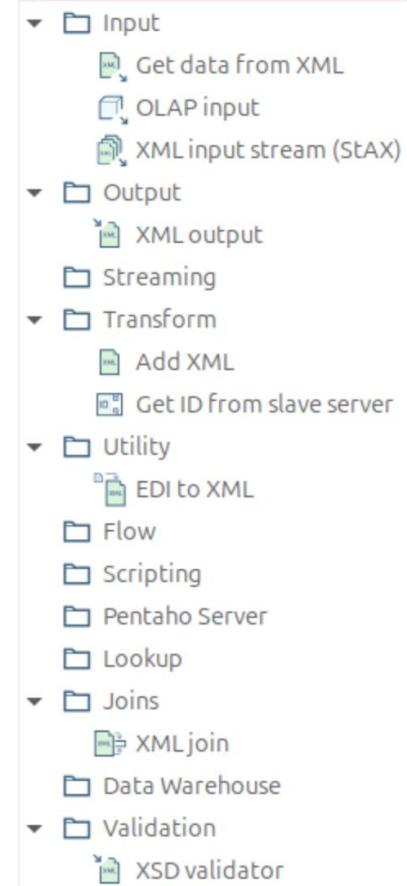
XML no es un formato simple y es necesario cumplir ciertas especificaciones, PDI facilita 4 opciones para validar la estructura:

- Check if XML file is well formed. Comprueba si las etiquetas de apertura y cierre están bien formadas. y la estructura anidada está bien equilibrada.
- DTD Validator. Verifica el contenido del archivo XML en base a un archivo de definición de tipo de datos, que puede ser interno (contenido dentro del archivo XML) o externo (archivo DTD separado).
- XSD Validator (Job). Verifica el contenido del archivo XML basándose en XML Schema
- Definition file
- XSD Validator (Transformation). Igual que el anterior, también puede verificar XML válido dentro de un campo de entrada específico, como una columna de base de datos que contiene datos XML

# Extracción basada en archivos

## *Working with XML files*

- Get data from XML
- XML input stream
- XML output
- XML Join
- XSD validator



# Extracción basada en archivos

## *Working with XML files*



# Extracción basada en archivos

## XML file

```
▼<collection>
  ▼<record>
    <leader>01481nam a2200385 i 4500</leader>
    <controlfield tag="005">20210908155716.0</controlfield>
    <controlfield tag="007">ta</controlfield>
    <controlfield tag="008">210906s1584 it a r 000 0 ita d</controlfield>
    <controlfield tag="003">ES-GrU</controlfield>
    <controlfield tag="001">991014439313204990</controlfield>
    ▼<datafield tag="040" ind1=" " ind2=" ">
      <subfield code="a">UGR</subfield>
    </datafield>
    ▼<datafield tag="100" ind1="1" ind2=" ">
      <subfield code="a">Cattaneo, Girolamo.</subfield>
    </datafield>
    ▼<datafield tag="245" ind1="1" ind2="0">
      <subfield code="a">Dell'arte militare libro quinto, de gli auuertimenti, et essamini intorno  
a quelle cose che richiede a vn bombardiero, cosi circa all'Artiglieria, come anco a fuochi  
arteficiati / di M. Girolamo Cataneo nouarese</subfield>
    </datafield>
    ▼<datafield tag="260" ind1=" " ind2=" ">
      <subfield code="a">In Brescia :</subfield>
      <subfield code="b">appresso Thomaso Bozzola,</subfield>
      <subfield code="c">1584</subfield>
    </datafield>
    ▼<datafield tag="300" ind1=" " ind2=" ">
      <subfield code="a">[2], 39, [1] h. :</subfield>
      <subfield code="b">il. ;</subfield>
      <subfield code="c">4º</subfield>
    </datafield>
    ▼<datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Marca tip. en port.</subfield>
    </datafield>
    ▼<datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Colofón</subfield>
    </datafield>
```

# Extracción basada en archivos

## Working with XML files

### Get data from XML

The screenshot shows the configuration interface for the 'Get data from XML' step. The main window title is 'Get data from XML' with a step name 'Get data from XML'. The 'Content' tab is selected.

**XML source from field:**

- XML source is defined in a field?
- XML source is a filename?
- Read source as Url
- get XML source from a field

**File or directory:**

**Regular Expression:**

**Exclude Regular Expression:**

**Selected files:**

- File/Directory
  - 1 \${Internal.Entry.Current.Directory}/repository/fa\_febrero\_new\_1.xml
  - 2 \${Internal.Entry.Current.Directory}/repository/fa\_febrero\_new\_2.xml
  - 3 \${Internal.Entry.Current.Directory}/repository/fa\_febrero\_new\_3.xml
  - 4 \${Internal.Entry.Current.Directory}/repository/fa\_febrero\_new\_4.xml

**Show filename(s)...**

**Step name:** Get data from XML

**Settings:**

- Loop XPath: /collection/r
- Encoding: UTF-8
- Namespace aware?
- Ignore comments?
- Validate XML?
- Use token
- Ignore empty file
- Do not raise an error
- Limit: 0
- Prune path to handle

**Additional fields:**

- Include filename in  **Filename fieldname:**
- Rownum in output:  **Rownum fieldname:**

**Add to result filename:**   
**Add files to result f:**

Buttons at the bottom: Help, OK, Preview rows, Cancel.

# Extracción basada en archivos

## *Working with XML files*

### Get data from XML

Get data from XML

Step name Get data from XML

	Name	XPath	Element	Result type	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type	Repeat
1	leader	leader	Node	Value of	String							none	N
2	nuevoid_001	controlfield[@tag=001]	Node	Value of	String							none	N
3	OCoLC_ID_035	datafield[@tag=035]/subfield[starts-with(., 'OCoLC'))]	Node	Value of	None							none	N
4	HANLED_URL_856u	datafield[@tag=856]/subfield[@code='u']	Node	Value of	String							none	N
5	antiguoID_URL_962a	datafield[@tag=962]/subfield[@code='a']	Node	Value of	String							none	N
6	antiguoID_035	datafield[@tag=035]/subfield[starts-with(., '(ES-GrU)b'))]	Node	Value of	String							none	N
7	antiguoID_907	datafield[@tag=907]/subfield[@code='a']	Node	Value of	String							none	N
8	nombre_245	datafield[@tag=245]/subfield[@code='a']	Node	Value of	String							none	N

**Get fields**

OK Preview rows Cancel

Help

# Extracción basada en archivos

## Working with XML files

### Get data from XML

#### Preview rows

Rows of step: Get data from XML (1000 rows)									Examine preview data
leader	nuevoID_00	OCoLC_ID_035	HANDLED	antiguoID_URL_962a	anl	antiguoID_907	nombre_245		
1 01481nam a2200385i 4500	99101443931		http://hdl.h				Dell'arte militare libro quinto, de gli auertimenti, et essami		
2 01564nam a2200385i 4500	99101443931		http://hdl.h				Dell'arte militare libro secondo, nel qual si mostra un nouo r		
3 01592nam a2200409i 4500	99101443931		http://hdl.h				Del misurare le muraglie, imbottare grani, vini, fieni, et stran		
4 01516nam a2200385i 4500	99101443871		http://hdl.h				Dell'arte militare libro terzo, come si deve formare con pre:		
5 01612nam a2200385i 4500	99101443871		http://hdl.h				Dell'arte libro quarto, doue con tauole breuissime si mostra		
6 01244nam a2200313i 4500	99101438255		http://hdl.h				Encyclopedie méthodique : histoire naturelle des vers :		
7 01211nam a2200277i 4500	99101434035		http://hdl.h				Maravillosa vida, y feliz muerte de la venerable madre Sor N		
8 01533nam a2200445i 4500	99101412285		http://hdl.h				Voyage pittoresque ou Description des royaumes de Naples		
9 02216nam a2200433 a 4500	99101411135		http://hdl.h				Tabula Smaragdina medico-pharmaceutica; in qua sexcento		
10 01800nam a2200421 a 4500	99101411135		http://hdl.h		(ES-		Enchiridion medicum practicum, sive tractatus de morboru		
11 01688nam a2200373i 4500	99101410275		http://hdl.h				Voyages dans les Alpes :		
12 01401nam a2200337 i 4500	99101410275		http://hdl.h		(ES-		Ornithologia sive Synopsis methodica sistens avium divisione		
13 01554nam a2200361 i 4500	99101410274		http://hdl.h				Tratado Fisico-medico de los baños de Alhama de Guadix qu		
14 02204nam a2200481 a 4500	99100985977	(OCoLC)929319035	http://hdl.h		(ES-	.b19516393	Monarchia ecclesiastica /		
15 02118nam a2200457 a 4500	99100985922		http://hdl.h	000:000:URL:b2651977:012924:0:0:0:0:0	(ES-	.b2651977x	Dissertatio medica inauguralis de Caryophyllis aromaticis. Q		
16 02392nam a2200493 a 4500	99100985223	(OCoLC)934318479	http://hdl.h			.b19512624	Los treynta libros de la Monarchia ecclesiastica, o Historia u		
17 02216nam a2200481 a 4500	99100985084		http://hdl.h	000:000:URL:b2651552:012923:0:0:0:0:0	(ES-	.b26515520	D. D. Arbor toxicaria Macassariensis. Quam venia exp. Fac. M		
18 02233nam a2200481 a 4500	99100985080		http://hdl.h	000:000:URL:b2651550:012922:0:0:0:0:0	(ES-	.b26515507	D. D. Restio, quem dissertatione botanica, consensu exp. Fac		
19 01423nam a2200325 a 4500	99100985067	(OCoLC)934318176	http://hdl.h	000:000:URL:b1951178:012961:0:0:0:0:0	(ES-	.b19511784	Prodigios obrados por el qran patriarca San Felipe Neri en		

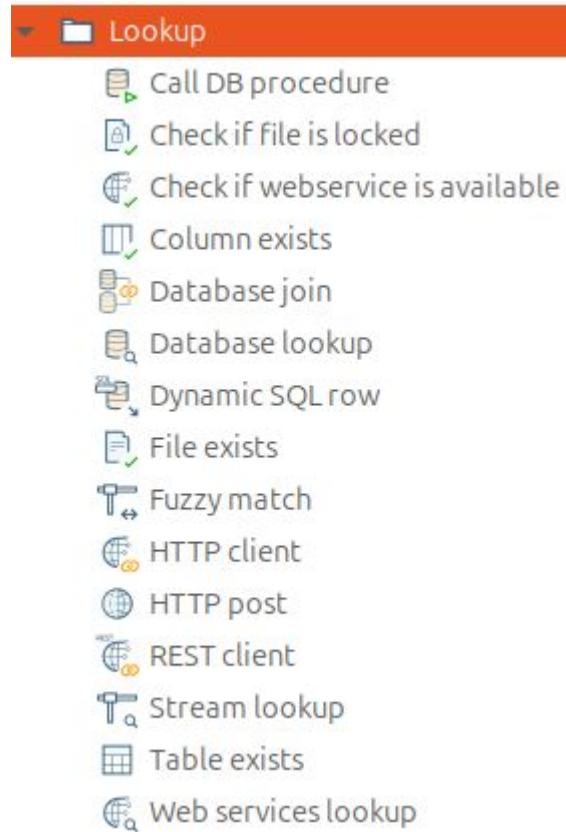
Close

Show Log

# Extracción basada en web

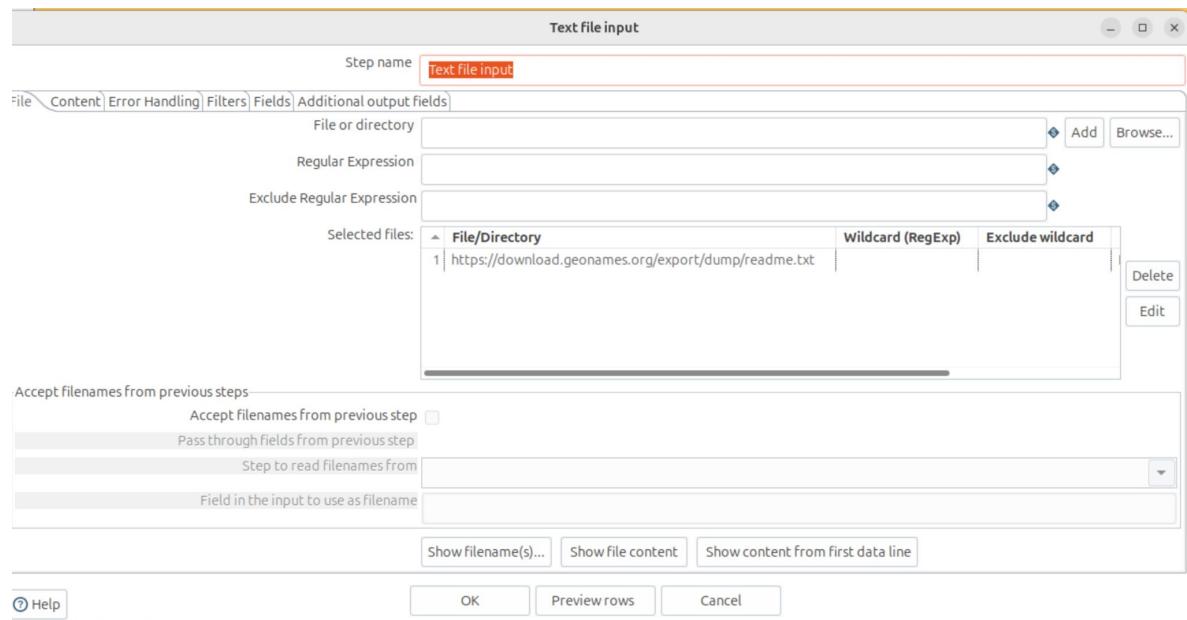
Obtener datos a través de la web mediante:

- *Text-Based Web Extraction*
- *HTTP Client*
- *Web services*
  - *REST client*
  - *Web services lookup - SOAP*



# Extracción basada en web

Desde el paso ***Text input data*** se puede utilizar una URL como nombre de archivo que utiliza Apache VFS.



# Extracción basada en web

## ***HTTP Client***

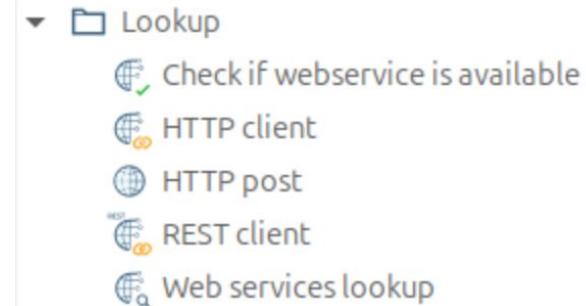
Recuperación de datos estructurados de Internet.

Realiza una llamada a la URL específica y retorna los datos como String.

Este resultado puede ser texto normal que se puede guardar como un archivo *delimited*, por ejemplo CSV o en XML.

El resultado debe ser procesado posteriormente dependiendo del formato de salida.

Es un paso de búsqueda y necesita una entrada para activarse.



# Extracción basada en web

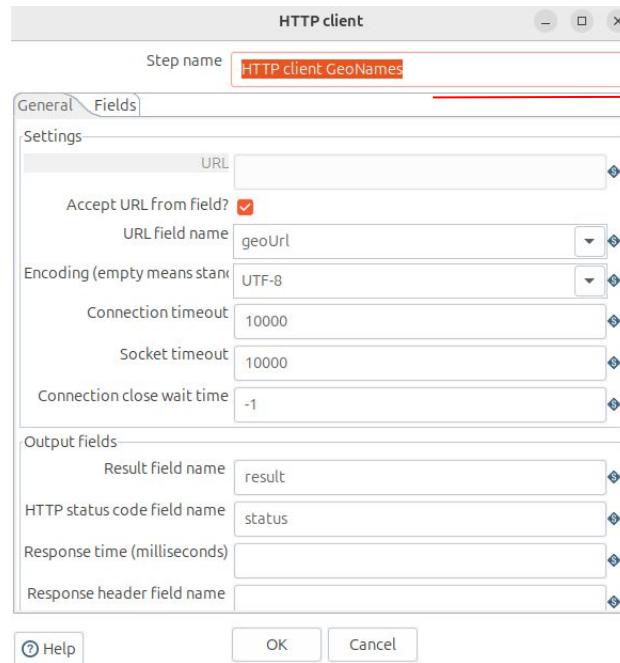
## HTTP Client



f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12
iso alpha2	iso alpha3	iso numeric	fps code	name	capital	arealnSqKm	population	continent	languages	currency	geonameid
ES	ESP	724	SP	Spain	Madrid	504782.0	46723749	EU	es-ES,ca,gl,eu,oc	EUR	2510769

# Extracción basada en web

## HTTP Client



Rows of step: HTTP client GeoNames (1 rows)														Examine preview data
geoUrl	result	status												
1 https://www.geonames.org/countryInfoCSV?country=ES	iso alpha2 iso alpha3 iso numeric fips code name capital areaInSqKm population continent languages currency geonameid	ES ESP 724 SP Spain Madrid 504782.0 46723749 EU es-ES,ca,gl,eu,oc EUR 2510769	200											



HTTP client GeoNames

<https://www.geonames.org/countryInfoCSV?country=ES>

# Extracción basada en web

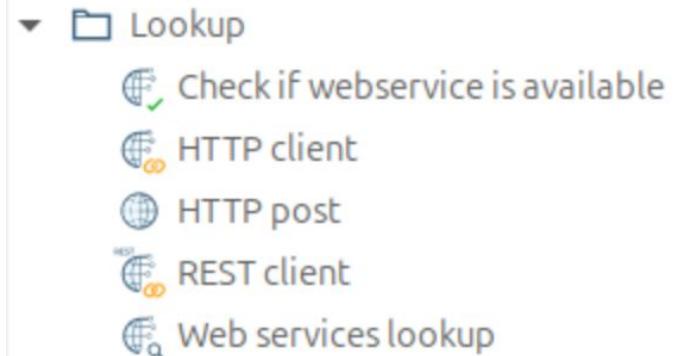
## *Web services*

### *REST client*

- REST (Representational State Transfer)

### *Web services lookup*

- SOAP (Simple Object Access Protocol)
- Utiliza the Web Services Description Language (WSDL)



# Extracción basada en web

## *Web services REST client*

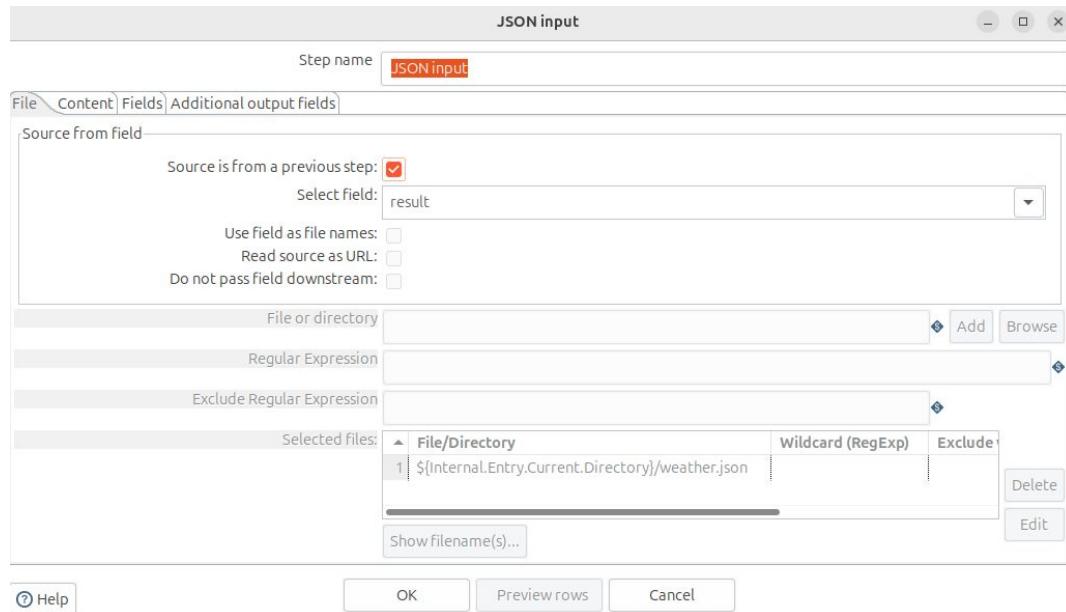


<https://api.openweathermap.org/data/2.5/weather?q=Zaragoza&APPID=>

```
{
  "coord": {
    "lon": -0.8773,
    "lat": 41.6561
  },
  "weather": [
    {
      "id": 800,
      "main": "Clear",
      "description": "clear sky",
      "icon": "01d"
    }
  ],
  "base": "stations",
  "main": {
    "temp": 302.75,
    "feels_like": 301,
    "temp_min": 302.4,
    "temp_max": 303.16,
    "pressure": 1014,
    "humidity": 17,
    "sea_level": 1014,
    "grnd_level": 976
  },
  "visibility": 10000,
  "wind": {
    "speed": 6.69,
    "deg": 310
  },
  "clouds": {
    "all": 0
  },
  "dt": 1757006653,
  "sys": {
    "type": 2,
    "id": 2003310,
    "country": "ES",
    "sunrise": 1756963963,
    "sunset": 1757010766
  },
  "timezone": 7200,
  "id": 3104324,
  "name": "Zaragoza",
  "cod": 200
}
```

# Extracción basada en web

**Web services REST client**



# Extracción basada en web

## *Web services lookup*

Para servicios web basados en Simple Object Access Protocol (SOAP)

Utiliza Web Services Description Language (WSDL) para cargar las operaciones y mapear campos.

En URL se indica la dirección del WSDL

Puede necesitar usuario y contraseña

No necesita ser activado por un paso previo como en el caso de HTTP Client.

# Extracción basada en web

## *Web services lookup*

Web services lookup

Step name: indexFRBR

Web Service [in return]

URL: http://ws.bvmc.es/api/IndexWSService?wsdl  
Operation: index  
Operation request name (optional): index  
The number of rows per call: 1  
Pass input data to output:   
v2.x/3.0 Compatibility mode:   
Repeating element name:   
Return the complete reply from the service as a String:

HTTP authentication

HTTP Login: userEl  
HTTP Password:

Proxy to use

Proxy Host:   
Proxy Port:

Buttons: OK, Add Input, Add Output, Cancel



Web services lookup

Step name: indexFRBR

Web Service [in return]

Name	WS Name	WS Type
1	id	idDocument
		string

Get Fields

Buttons: Help, OK, Add Input, Add Output, Cancel