



# Pentaho


Big Data



accenture

# Introducción

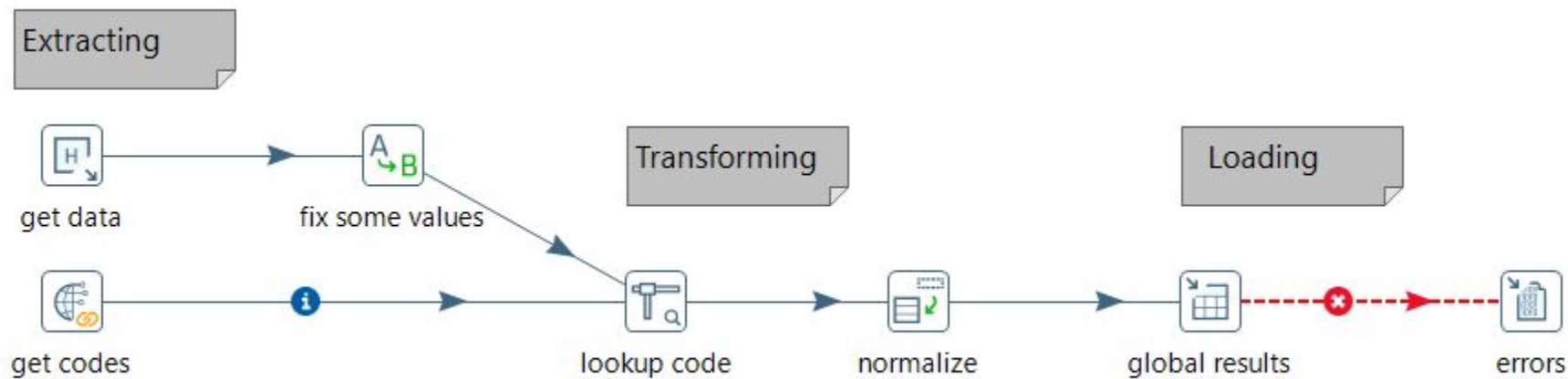
**Kettle** es un componente de **Pentaho Data Integration** que a su vez contiene a **Spoon**. Mediante Spoon se pueden realizar procesos ETL de manera visual, de forma muy fácil y rápida, como por ejemplo:

- Conexiones a los datos.
  - Transformaciones (filtrado, limpieza, formateado, ... y almacenamiento en diferentes formatos y destinos).
  - Inserción de fórmulas.
  - Desarrollo de data warehouses
- 

# Introducción

- Y todo esto sin necesidad de programar directamente con código y sin necesidad de instalar o configurar nada para poder empezar a usarla.
- Por esto, este tipo de herramientas se conocen como herramientas de **metadatos**, ya que trabajan a nivel de definición diciendo qué hay que hacer, pero no el detalle del cómo se hace.

# Introducción

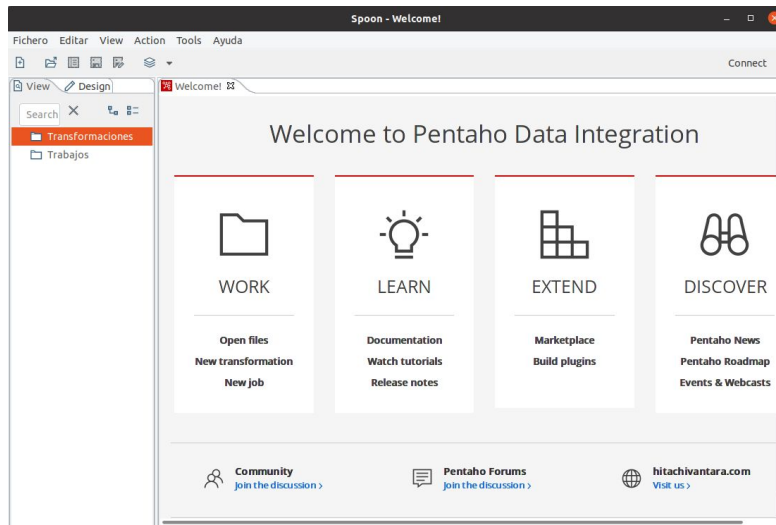


# Introducción

- Se trata de una herramienta open source multiplataforma que también tiene su soporte comercial. La versión open source se puede descargar desde <https://sourceforge.net/projects/pentaho/>
- Es importante destacar como requisito que necesitamos tener instalado en el sistema la versión 8 de Java.


# Instalación

- Una vez descargado el archivo y tras descomprimirlo, mediante el archivo **spoon.bat** (o spoon.sh) lanzaremos la aplicación.




# Dentro de Spoon

**Spoon** permite diseñar mediante un interfaz gráficos las transformaciones y trabajos que se ejecutan con las siguientes herramientas:

- **Pan** es un motor de transformación de datos que facilita la lectura, manipulación, y escritura de datos hacia y desde varias fuentes de datos. Ejecuta archivos ktr.
  - **Kitchen** es un programa que ejecuta los trabajos (jobs) diseñados por Spoon en XML o en una base de datos. Ejecuta archivos kjb.
- 

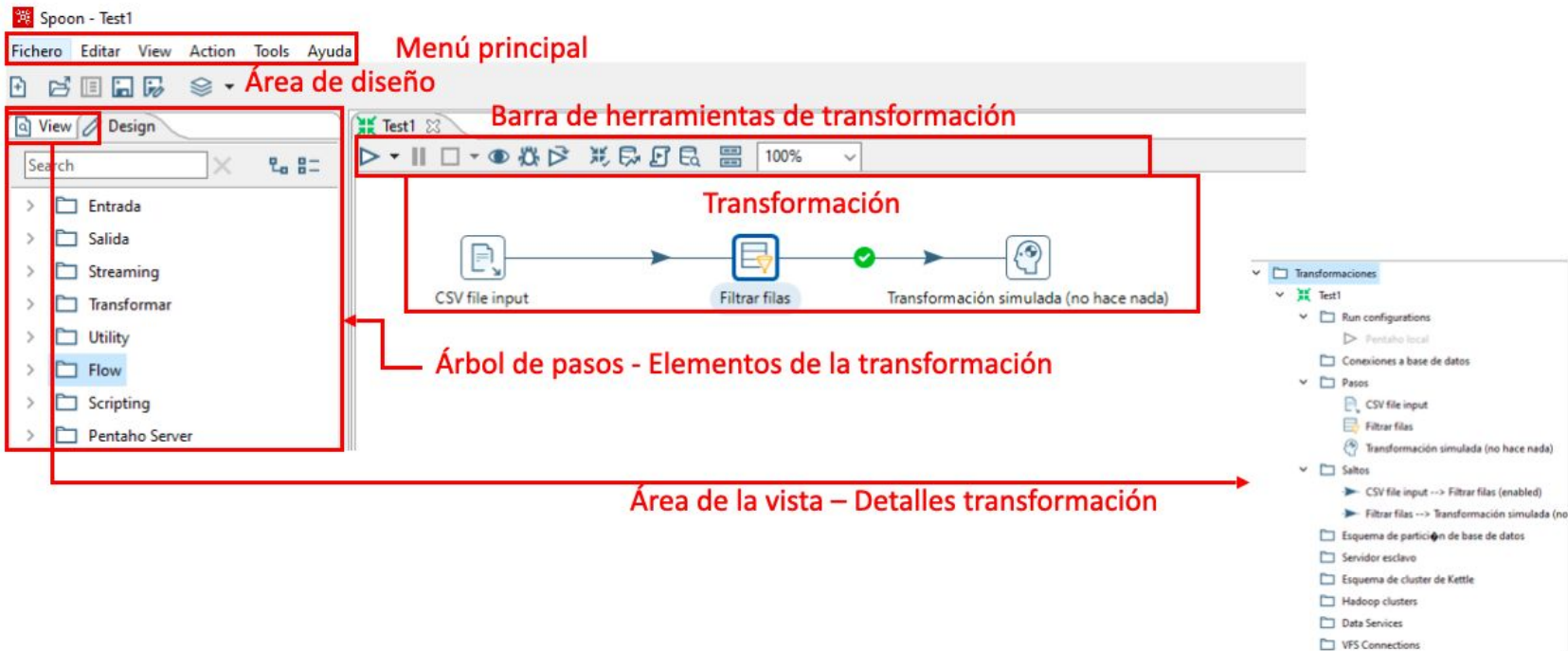
# Elementos

En PDI hay dos tipos de elementos: Transformations y Jobs.

- Se definen **Transformations** para transformar los datos, describiendo flujos de datos para ETL como leer desde una fuente, transformar los datos y cargarlos en un nuevo destino.
  - Se definen **Jobs** para organizar tareas y transformaciones estableciendo su orden y condiciones de ejecución (del tipo ¿existe el fichero X en origen? ¿existe la tabla X en mi base de datos?).
  - Tanto las transformaciones como las tareas, cuando se definen, se almacenan como archivos.
- 



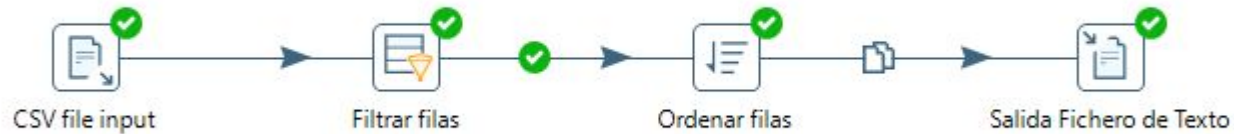
# Interfaz de Spoon



# Ejemplo 1



# Ejemplo 2



<https://www.kaggle.com/datasets/themrityunjaypathak/imdb-top-100-movies?resource=download>