

Caso práctico: almacén de datos para el análisis del impacto conductual de la COVID-19 sobre la población

PRA2- Carga de datos

Presentación

A partir de la solución oficial de la primera práctica (PRA1), el estudiante debe diseñar, implementar y ejecutar los procesos de extracción, transformación y carga de los datos de las fuentes de datos proporcionadas.

Así pues, esta actividad tiene como objetivo identificar y desarrollar los procesos de carga del almacén de datos y que esta sea efectiva.

Descripción

Si nos centramos en los subobjetivos, esta segunda parte del caso práctico consiste en lo siguiente:

- Identificar los procesos de extracción, transformación y carga de datos (ETL) hacia el almacén de datos.
- Diseñar y desarrollar los procesos ETL mediante las herramientas de diseño proporcionadas.
- Implementar con los trabajos (*jobs*) los procesos ETL para que su carga planificada sea efectiva.

Además del documento con la solución de la PRA2 que se debe entregar, también se tendrá en consideración la implementación sobre la máquina virtual proporcionada en el curso.

En resumen, el documento de la solución de la PRA2 debe incluir los siguientes aspectos:

- Descripción de todas las acciones que se han realizado.
- Capturas de pantalla que muestren todas las partes significativas del ETL, sus características y su correspondiente explicación.
- Capturas de pantalla que demuestren la correcta ejecución de la ETL y el tiempo de ejecución.
- Capturas de pantalla que demuestren la correcta carga de los datos (cargados en la base de datos).

Guía de muestra

Con el fin de ayudar a alcanzar los objetivos planteados de la PRA2 se desarrolla esta guía de muestra. Esta servirá como ejemplo de cómo realizar alguna de las tareas anteriormente descritas, es decir, el diseño y el desarrollo de los procesos ETL y la carga efectiva al almacén de datos.

1. Identificación de los procesos ETL

A la hora de diseñar los procesos de carga de una base de datos analítica no hay una única estrategia. Es habitual estructurar los procesos ETL sobre la base de las entidades de datos que se deben actualizar, ya que existen diferencias conceptuales en la actualización de una dimensión con respecto a la de una tabla de hechos. La división del proceso de carga inicial en diferentes bloques de actualización facilitará el diseño de un orden de ejecución y la gestión de las dependencias. Cada uno de estos bloques de actualización se dividirá en las correspondientes etapas de extracción, transformación y carga.

Se identifican los dos bloques siguientes:

- **Bloque IN:** procesos de carga de los datos desde las fuentes a las tablas intermedias en el área de maniobras (*staging area*). Estos procesos se distinguen por el prefijo «IN_» en el nombre.
- **Bloque TR:** procesos de transformación para cargar los datos desde las tablas intermedias hasta nuestro almacén, según el modelo multidimensional diseñado. Así pues, son diferentes los procesos ETL de transformación para cargar las dimensiones de aquellos que se realizan para cargar las tablas de hechos. Estos procesos se distinguen con el prefijo «TR_» en el nombre.

A continuación, se identifican **algunos de los procesos** que forman parte de cada uno de los bloques de actualización:

Bloque IN (de las fuentes a las tablas intermedias)

Nombre del ETL	Descripción	Orígenes de los datos	Tabla de destino (<i>stage</i>)
IN_DENUNCIAS_INFRACCIONES	Carga de los datos correspondientes a las estadísticas sobre los expedientes incoados por el artículo 36.6 LOPSC de desobediencia durante el estado de emergencia sanitaria de la COVID-19 en la comunidad de Euskadi.	ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx	STG_Denuncias_Infracciones

Bloque TR (de las tablas intermedias a nuestro almacén)

El bloque TR de procesos ETL para poblar el modelo multidimensional del almacén tiene dos partes diferenciadas. Por un lado, los procesos de carga y transformación de las dimensiones y, por otro, los de las tablas de hechos. El orden de ejecución es importante para que la carga de datos sea la correcta. Las dimensiones se cargarán primero y, después, las tablas de hechos para que no haya errores durante la carga.

Por una parte, algunos de los procesos del bloque TR de carga y transformación de las dimensiones son los siguientes:

Nombre del ETL	Descripción	Tabla de origen	Tabla de destino (dimensión)
TR_DIM_FECHA	Carga y transformación de la dimensión temporal.	SQL	DIM_Fecha
TR_DIM_AMBITO_GEOGRAFICO	Carga y transformación de la dimensión con los datos de los ámbitos geográficos.	STG_Poblacion STG_Llamadas_CAT112 STG_Evitar_Aglomeracion	DIM_Ambito_Geografico

Por otra, el proceso del bloque de carga y transformación de la tabla de hechos es el siguiente:

Nombre del ETL	Descripción	Tabla de origen
TR_FACT_LLAMADAS112	Carga y transformación de la tabla de hechos FACT_Llamadas112.	STG_Llamadas112

En este punto, el estudiante deberá completar la identificación de los procesos de cada uno de los bloques (IN y TR) que desarrollará para cargar las dimensiones y las tablas de hechos del modelo multidimensional del almacén de datos.

2. Diseño y desarrollo de los procesos ETL

En este apartado, se deben diseñar los procesos de carga identificados en el punto anterior con la herramienta de diseño proporcionada. En este caso es Pentaho Data Integration (PDI).

Creación de tablas

El primer paso para la implementación de los procesos ETL consiste en la creación de las tablas. Esto se llevará a cabo una única vez, mediante *scripts*, sobre la base de datos proporcionada (en nuestro caso: SQL Server). Se deberán crear las tablas intermedias y las tablas del modelo dimensional de la solución oficial, es decir, las dimensiones y las tablas de hechos. Para hacerlo, deben utilizarse los *scripts* facilitados junto a la solución de la PRA1.

Una vez que tenéis implementado el modelo físico del almacén, el siguiente paso que hay que realizar es el diseño de los procesos ETL de cada uno de los bloques (IN y TR). Estos procesos permitirán poblar las tablas del área intermedia (*staging area*), las de dimensiones y las de hechos del almacén de datos que habéis diseñado.

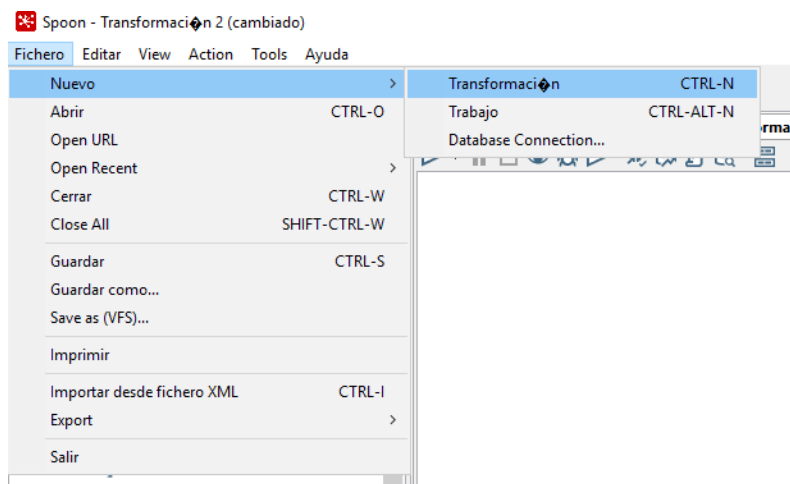
Bloque IN

Transformación de «IN_DENUNCIAS_INFRACCIONES»

A continuación, se describe parte del desarrollo de la transformación de «IN_DENUNCIAS_INFRACCIONES» (identificada en el primer punto de la guía) mediante Spoon. El objetivo es cargar uno de los orígenes de los datos identificados, «ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx», en la tabla «STG_Denuncias_Infracciones» del área intermedia (*staging area*). La tabla intermedia tendrá que haber sido creada con anterioridad en la base de datos analítica, cuyo *script* se habrá escrito en el apartado de «creación».

Para este caso práctico habéis utilizado fuentes externas (no operacionales) que emplearéis para descubrir el conocimiento mediante el análisis de los datos. Es muy habitual manipular los ficheros realizando manualmente una serie de acciones de preparación antes de su procesamiento (preprocesado).

La transformación de «IN_DENUNCIAS_INFRACCIONES» contiene las cuatro alteraciones siguientes: la lectura del fichero XLSX, las operaciones con cadenas, la organización de las filas y la carga a la tabla intermedia «STG_Denuncias_Infracciones».

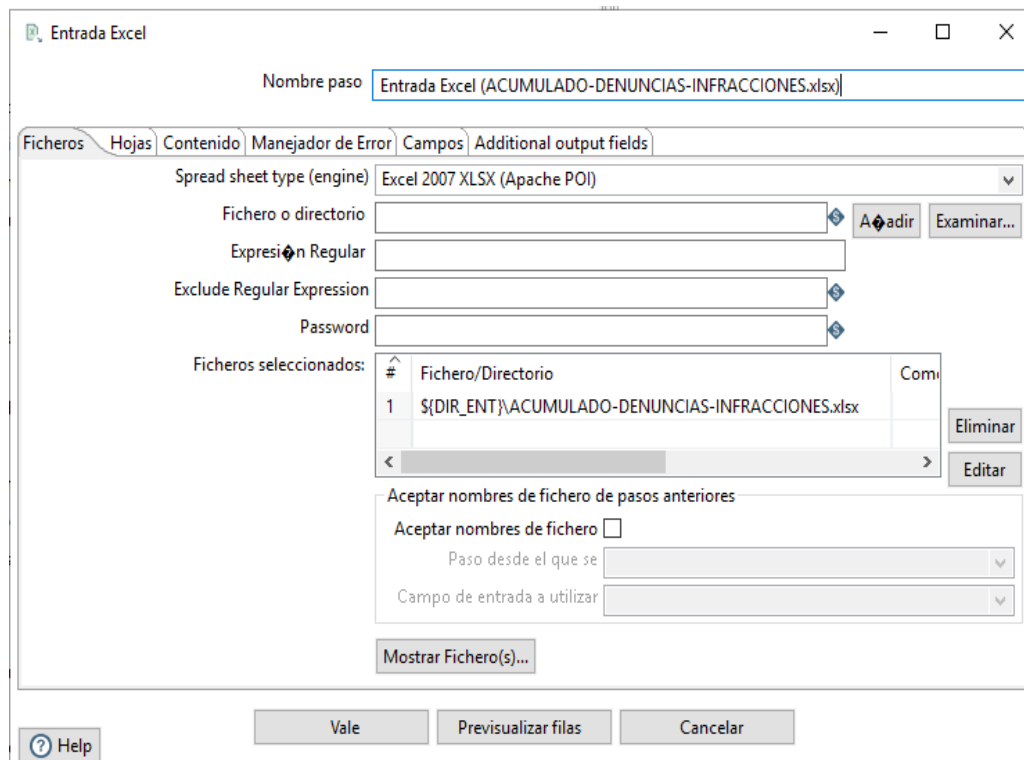


Este es el primer paso de la transformación, en el que realizaréis la entrada del fichero XLSX. Para ello utilizaréis el tipo «Entrada Excel».



Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx)

Aquí indicaréis el fichero desde donde extraéis los datos. Para ello utilizaréis la variable de entorno «DIR_ENT» e indicaréis el tipo de motor que deberá usar, si es para ficheros de tipo XLS o XLSX.



Nombre paso: Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx)

Spread sheet type (engine): Excel 2007 XLSX (Apache POI)

Fichero o directorio:

Expresión Regular:

Exclude Regular Expression:

Password:

Ficheros seleccionados:

#	Fichero/Directorio	Com
1	\${DIR_ENT}\ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx	

Aceptar nombres de fichero de pasos anteriores

Aceptar nombres de fichero: ☐

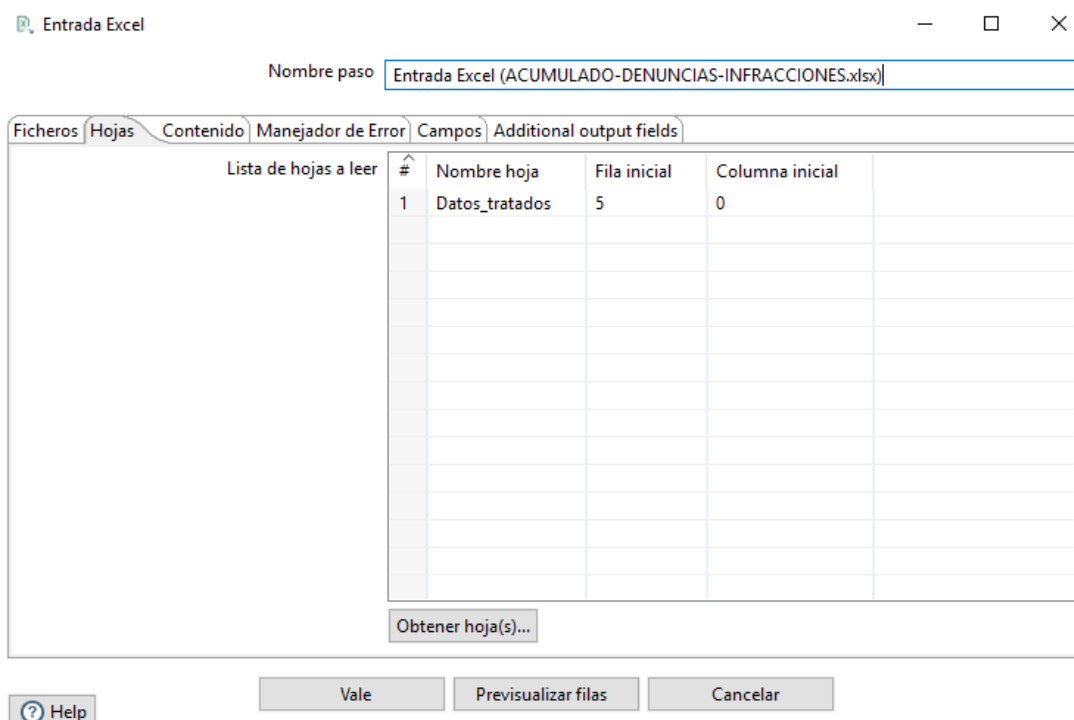
Paso desde el que se:

Campo de entrada a utilizar:

Mostrar Fichero(s)...

Vale Previsualizar filas Cancelar

Le indicáis qué hojas del fichero de origen deberá tener en cuenta y desde qué fila y columna deberá empezar a leer datos, como se muestra a continuación:



Nombre paso: Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx)

Lista de hojas a leer

#	Nombre hoja	Fila inicial	Columna inicial
1	Datos_tratados	5	0

Obtener hoja(s)...

Vale Previsualizar filas Cancelar

Así, tendréis que indicar que recupere los campos que vais a tratar con el botón «Traer campos» y completaréis la definición de estos. Hay que especificar, donde se considere necesario, la precisión y la longitud de los campos e indicar el formato de la fecha en el campo «date».

Si en los campos del tipo «number» se indica una longitud y una precisión con el valor -1 , estos se tratarán como un dato numérico de coma flotante (*float*).

Entrada Excel

Nombre paso: Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx)

Ficheros Hojas Contenido Manejador de Error Campos Additional output fields

#	Nombre	Tipo	Longitud	Precisión	Tipo de poda	Repetir	Formato	Moneda
1	provincia	String	100	-1	ninguno	N		
2	identificados_ertaintza	Number	-1	-1	ninguno	N	#	
3	detenidos_ertaintza	Number	-1	-1	ninguno	N	#	
4	denuncias_ertaintza	Number	-1	-1	ninguno	N	#	
5	vehic_intercept_ertaintza	Number	-1	-1	ninguno	N	#	
6	identificados_ppll	Number	-1	-1	ninguno	N	#	
7	detenidos_ppll	Number	-1	-1	ninguno	N	#	
8	denuncias_ppll	Number	-1	-1	ninguno	N	#	
9	vehic_intercept_ppll	Number	-1	-1	ninguno	N	#	
10	fecha_final	Date	-1	-1	ninguno	N	dd/MM/yyyy	

Obtener campos de cabecera...

Help Vale Previsualizar filas Cancelar

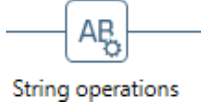
Para realizar una visualización previa de los datos que se cargarán, se utiliza el botón «Previsualizar filas».

Examine preview data

Rows of step: Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx) (219 rows)

#	provincia	identificados_ertaintza	detenidos_ertaintza	denuncias_ertaintza	vehic_intercept_ertaintza	identificados_ppll	detenidos_ppll	denuncias_ppll	vehic_intercept_ppll	fecha_final
1	ARABA	10717	40	2586	15182	21599	29	2748	19250	18/06/2020
2	BIZKAIA	29955	228	6249	56139	27160	65	9209	50539	18/06/2020
3	GIPUZKOA	26051	62	4884	17246	27069	40	4473	37797	18/06/2020
4	ARABA	10708	40	2586	15180	21598	29	2748	19250	17/06/2020
5	BIZKAIA	29867	228	6249	56009	27138	65	9209	50518	17/06/2020
6	GIPUZKOA	26004	62	4882	17183	27059	40	4472	37791	17/06/2020
7	ARABA	10705	40	2585	15176	21598	29	2748	19250	16/06/2020
8	BIZKAIA	29751	228	6249	55859	27124	65	9209	50468	16/06/2020
9	GIPUZKOA	25942	62	4882	17173	27038	38	4465	37782	16/06/2020
10	ARABA	10704	40	2585	15176	21593	29	2746	19250	15/06/2020
11	BIZKAIA	29674	228	6247	55754	27106	65	9203	50436	15/06/2020
12	GIPUZKOA	25872	62	4879	17103	27022	38	4465	37770	15/06/2020
13	ARABA	10700	40	2585	15173	21573	29	2739	19250	14/06/2020
14	BIZKAIA	29597	228	6246	55674	27063	65	9202	50428	14/06/2020
15	GIPUZKOA	25775	62	4873	17003	27015	38	4465	37754	14/06/2020
16	ARABA	10652	40	2572	15134	21572	29	2738	19250	13/06/2020
17	BIZKAIA	29477	228	6236	55470	27055	65	9201	50386	13/06/2020
18	GIPUZKOA	25672	62	4862	16924	27004	38	4464	37744	13/06/2020

El siguiente paso de la transformación es asegurar la calidad de los datos mediante la normalización de los valores de los campos «string». Para ello, convertiréis los datos de las fuentes de origen en mayúsculas y eliminaréis los espacios en blanco que pudiera haber al inicio y al final de la cadena de caracteres con el componente «String operations».



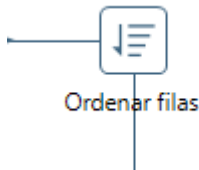
String operations

Step name:

The fields to process:

#	In stream field	Out stream field	Trim type	Lower/Upper	Padding	Pad char
1	provincia		both	upper	none	

A continuación, hay que ordenar de manera ascendente los campos. Para ello, utilizaréis el componente «Ordenar filas» de las posibles transformaciones que tenemos disponibles.



Ordenar filas

Nombre paso:

Directorio ordenación:

Prefijo para ficheros temporales:

Tamaño de ordenación (filas en memoria):

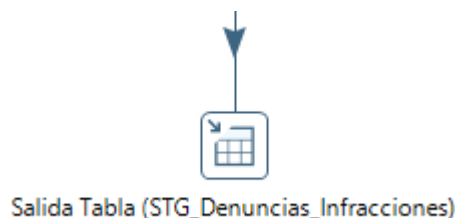
Free memory threshold (in %):

☒ Comprimir ficheros temporales?
 ☐ Only pass unique rows? (verifies keys only)

Campos:

#	Nombre Campo	Ascendente	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
1	provincia	S	N	N	0	N
2	identificados_ertzaintza	S	N	N	0	N
3	detenidos_ertzaintza	S	N	N	0	N
4	denuncias_ertzaintza	S	N	N	0	N
5	vehic_intercept_ertzaintza	S	N	N	0	N
6	identificados_ppll	S	N	N	0	N
7	detenidos_ppll	S	N	N	0	N
8	denuncias_ppll	S	N	N	0	N
9	vehic_intercept_ppll	S	N	N	0	N
10	fecha_final	S	N	N	0	N

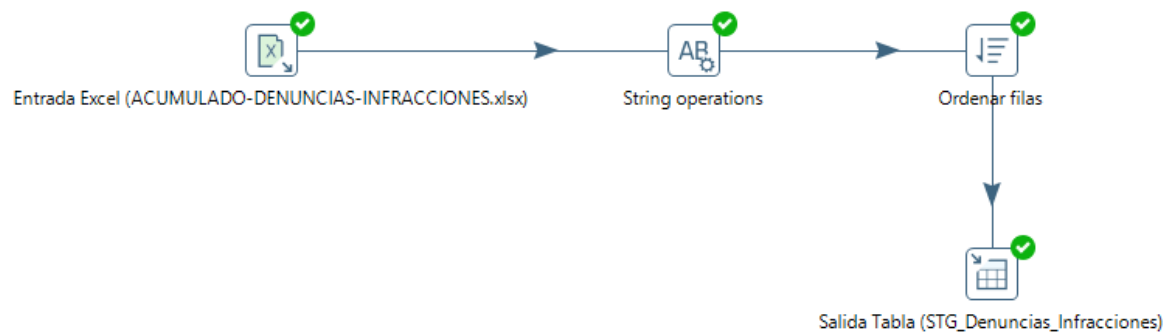
Por último, cargáis los datos en la tabla intermedia del stage, utilizando el paso «Salida Tabla» de la carpeta «Salida». Este paso necesita especificar la conexión de la base de datos y, para ello, utilizaréis la variable de entorno «CN_STAGE» que habéis definido anteriormente.



El paso de cargar los datos a la tabla intermedia del *stage* lo configuraréis como se os indica en el menú principal que veis a continuación:

Para dejar la transformación preparada para posibles reprocesos, es necesario realizar un borrado previo para actualizar los datos en el caso de que tuvierais que efectuarlo. Para esto, activaréis el *check* «Vaciar tabla» situado en los campos de la base de datos, como veis a continuación:

El proceso de la transformación completa es el siguiente:



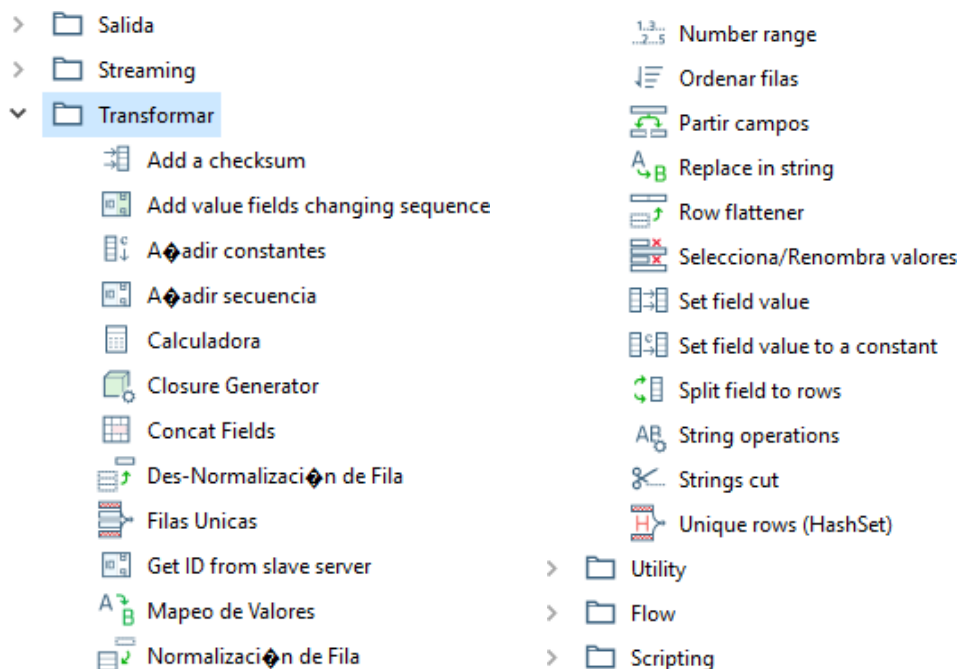
Y el resultado de la ejecución, el que veis a continuación:

Execution Results

Execution Results											
Logging Execution History Step Metrics Performance Graph Metrics Preview data											
#	Nombre paso	Numero Copia	Leído	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo
1	Entrada Excel (ACUMULADO-DENUNCIAS-INFRACCIONES.xlsx)	0	0	219	219	0	0	0	0	Finalizado	2.4s
2	String operations	0	219	219	0	0	0	0	0	Finalizado	2.4s
3	Ordenar filas	0	219	219	0	0	0	0	0	Finalizado	2.4s
4	Salida Tabla (STG_Denuncias_Infracciones)	0	219	219	0	219	0	0	0	Finalizado	2.6s

Como se puede observar en las métricas, se cargan los 219 registros del fichero de entrada.

Para la entrega de la PRA2, el estudiante deberá diseñar todos los procesos ETL de cada uno de los bloques (IN y TR). En este ejemplo se ha mostrado un caso básico de carga de datos, pero, según el formato de origen de los datos y de la calidad de estos, tal vez sea necesario utilizar otras transformaciones. Spoon dispone de una gran cantidad de componentes a los que se puede acceder desde el menú lateral y están organizados por categorías, como se puede observar en el siguiente ejemplo:



3. Implementación de los trabajos con procesos ETL

Los bloques de procesos ETL implementados que hay que tener en cuenta son los siguientes:

- **Bloque IN_:** procesos ETL de transformación y carga al área intermedia.
- **Bloque TR_DIM:** procesos ETL de transformación y carga de dimensiones.
- **Bloque TR_FACT:** procesos ETL de transformación y carga de hechos.

En este punto, para realizar la carga efectiva de los datos, el estudiante debe diseñar mediante PDI los trabajos (*jobs*) que permitan la ejecución secuencial de todos los procesos ETL incluidos en cada bloque. En este apartado se deben incluir también las volumetrías obtenidas (número de registros cargados en cada tabla).

Formato y fecha de entrega

La entrega final de esta actividad debe realizarse enviando un único mensaje al buzón de «Registro de AC» del apartado «Evaluación» del aula. Junto con el mensaje se enviará un único archivo en formato Word o PDF con la solución de la PRA2. El nombre del archivo debe ser la composición del nombre de usuario y «_BDA_PRA2». Por ejemplo, si el nombre de usuario es «bantich», el nombre del archivo debe ser «bantich_BDA_PRA2.pdf».

La fecha máxima de entrega es el 19/05/2021 a las 23:59 horas.