



GFPI-F-135 REALIZA EL PROCESO DE LIMPIEZA DE DATOS  
Extracción Transformación y Carga

ACTIVIDADES POR DESARROLLAR:

1. ¿Qué es un ETL?

La palabra ETL corresponde a las siglas en inglés de:

- Extraer: **extract**.
- Transformar: **transform**.
- Y Cargar: **load**.

Descargar [Pentaho Data Integration](#)

## PRACTICA ETL CON PENTAHO

Una vez que se instale la ETL Pentaho Data Integration, se debe ejecutar spoon.bat pero para que funcione se deben configurar las siguientes variables de entorno:

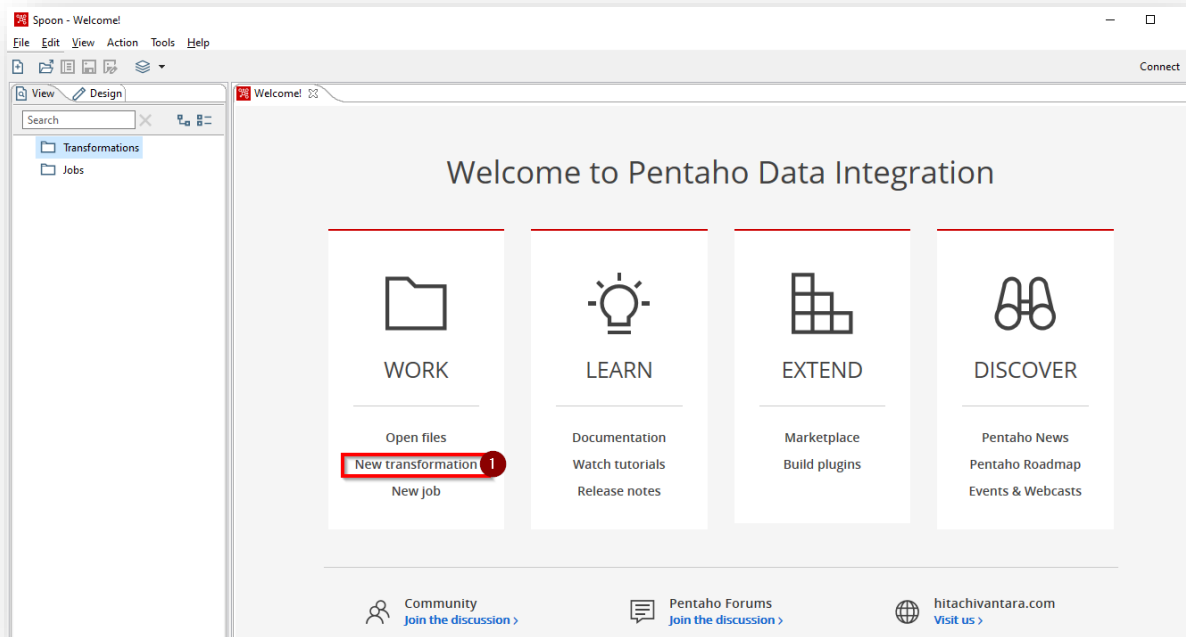
```
PENTAHO_JAVA_HOME C:\Program Files\Java\jdk1.8.0_202  
JRE_HOME C:\Program Files\Java\jdk1.8.0_202\bin
```

Antes de arrancar la aplicación se debe editar “spoo.bat” y agregar las siguientes líneas:

```
set PENTAHO_OPTS=-Dfile.encoding=UTF-8  
set JAVA_OPTS=-Dfile.encoding=UTF-8
```



Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.



Se crea una nueva transformación que llamaremos “paso1”  
Hace un input de archivo plano CSV



Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.

Descargar el **DATASET** de trabajo "[SB11-20121-RGSTRO-CLFCCN-V1-0-txt.csv](#)" y cargarlo en la configuración input

The screenshot shows the 'CSV file input' configuration window. The fields are as follows:

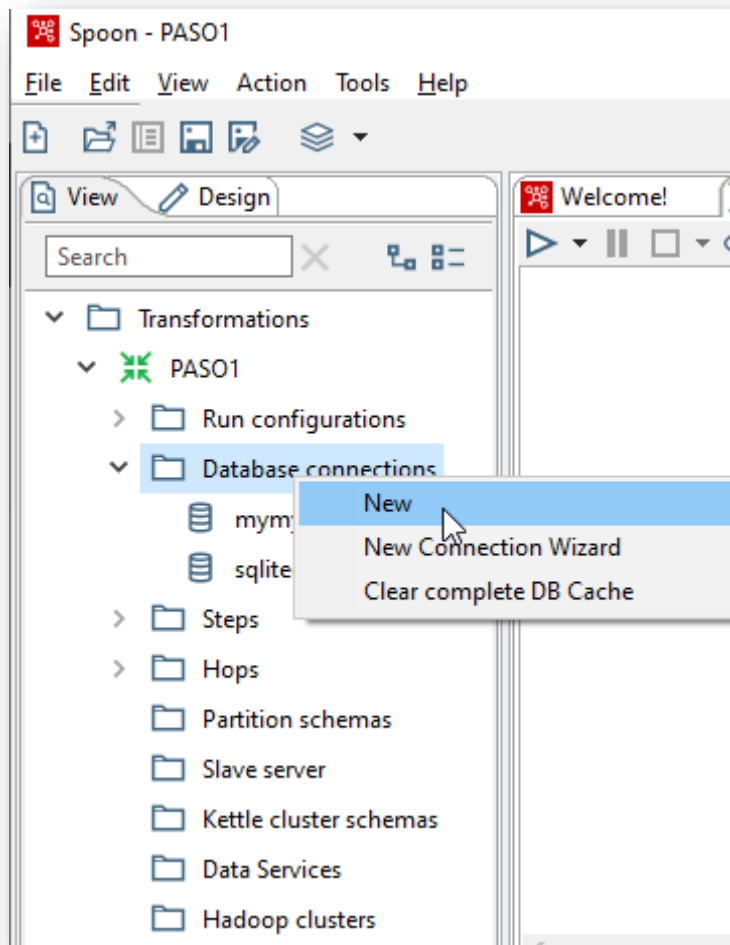
- Step name: CSV file input
- Filename: C:\Borrar\SB11-20121-RGSTRO-CLFCCN-V1-0-txt.csv (Annotation 1)
- Delimiter: , (Annotation 2)
- Enclosure: "
- NIO buffer size: 50000
- Lazy conversion? ☒
- Header row present? ☒
- Add filename to result ☐
- The row number field name (optional):
- Running in parallel? ☐
- New line possible in fields? ☐
- Format: mixed
- File encoding:

At the bottom, there is a table with the following data:

#	Name	Type	Format	Length	Precision	Currency
4	ESTU_GENERO	String		1		\$
5	ESTU_NACIMIENTO_DIA	Integer	#	15	0	\$
6	ESTU_NACIMIENTO_MES	Integer	#	15	0	\$
7	ESTU_NACIMIENTO_ANNO	Integer	#	15	0	\$
8	ESTU_EDAD	Integer	#	15	0	\$
9	FECHA_ANO	Date	dd/MM/yyyy			\$

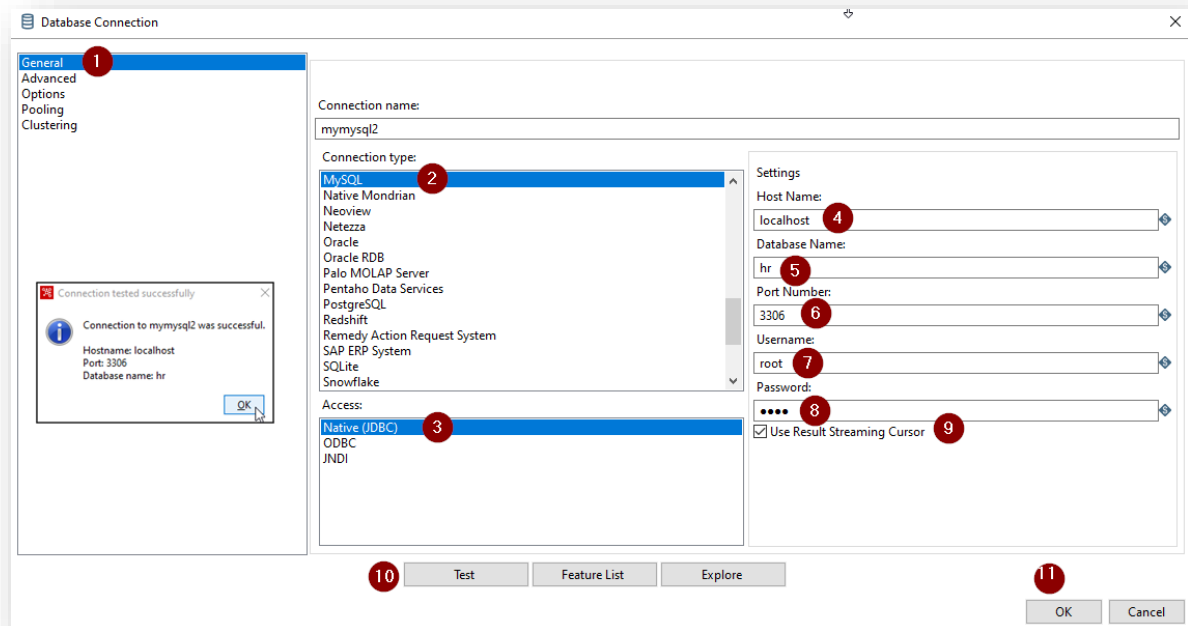
At the bottom of the window, there are buttons: Help, OK (Annotation 5), Get Fields (Annotation 3), Preview (Annotation 4), and Cancel.

Creamos la base de datos [HR](#) creamos la conexión a MYSQL desde Pentahoo

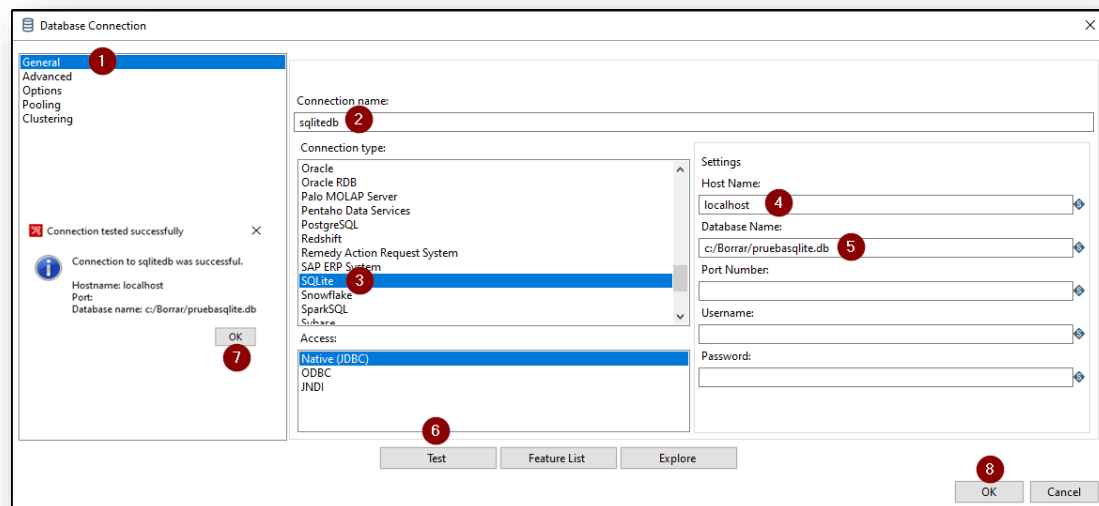




Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.

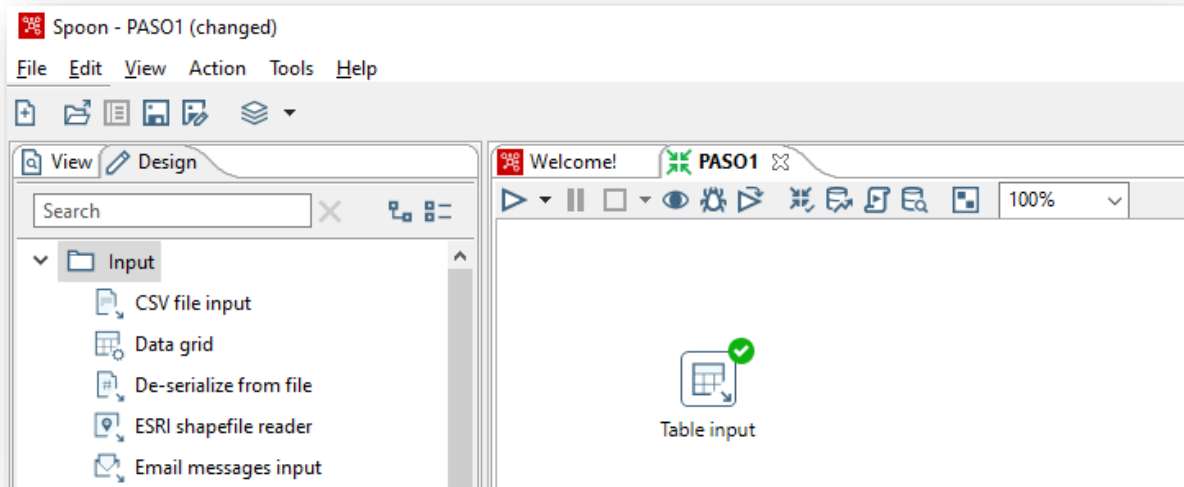


Creamos el enlace hacia SQLITE





Se crea la input “Table input”



Configure a “Table input” y cámbielo por “EMP”



Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.

Table input

Step name: EMP

Connection: mymysql2

SQL

```
SELECT
  department_id
, department_name
, manager_id
, location_id
FROM hr.departments
```

Line 1 Column 0

Store column info in step meta ☐

Enable lazy conversion ☐

Replace variables in script? ☐

Insert data from step

Execute for each row? ☐

Limit size: 0

Help OK Preview Cancel

Si le damos "Preview"



Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.

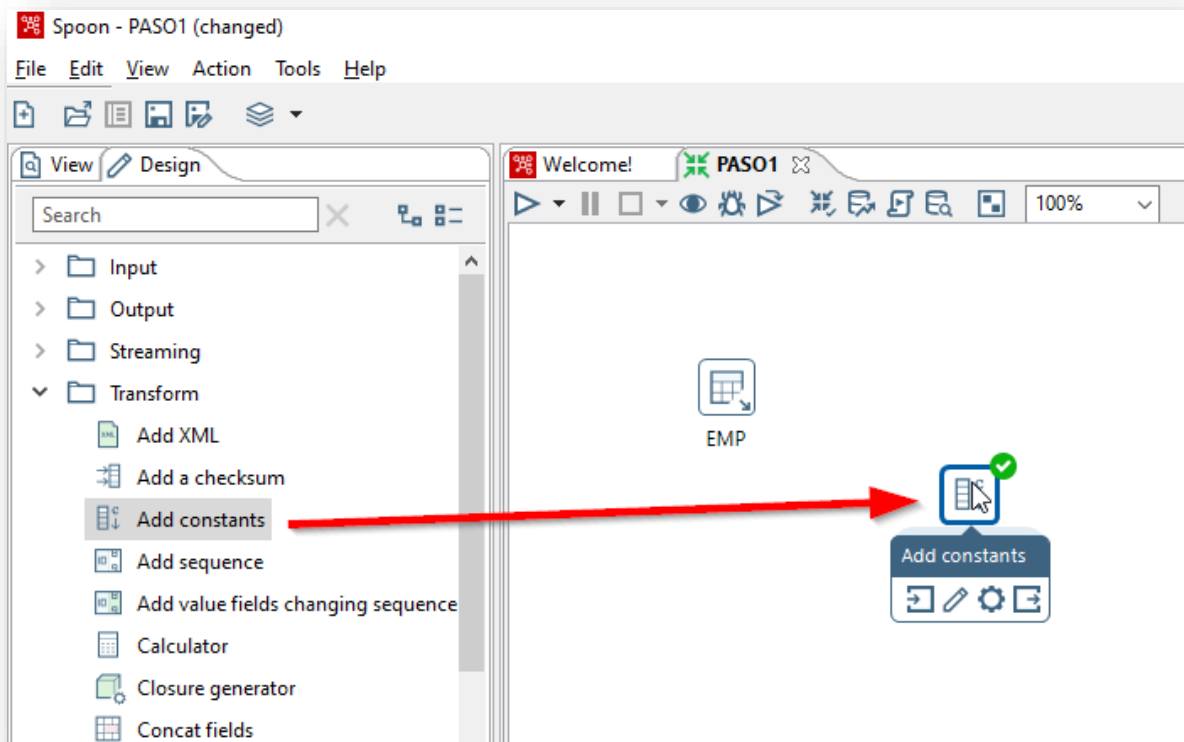
Examine preview data

Rows of step: EMP (27 rows)

#	department_id	department_name	manager_id	location_id
1	10	Administration	200	1700
2	20	Marketing	201	1800
3	30	Purchasing	114	1700
4	40	Human Resources	203	2400
5	50	Shipping	121	1500
6	60	IT	103	1400
7	70	Public Relations	204	2700
8	80	Sales	145	2500
9	90	Executive	100	1700
10	100	Finance	108	1700
11	110	Accounting	205	1700
12	120	Treasury	<null>	1700
13	130	Corporate Tax	<null>	1700
14	140	Control And Credit	<null>	1700
15	150	Shareholder Services	<null>	1700
16	160	Benefits	<null>	1700
17	170	Manufacturing	<null>	1700
18	180	Construction	<null>	1700
19	190	Contracting	<null>	1700

Close Show Log

Adicione “Add constants” de la sección “Transform”







Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.

Configurar para adicionar un campo llamado “ESTADO” con un valor de 1 para todos los registros.

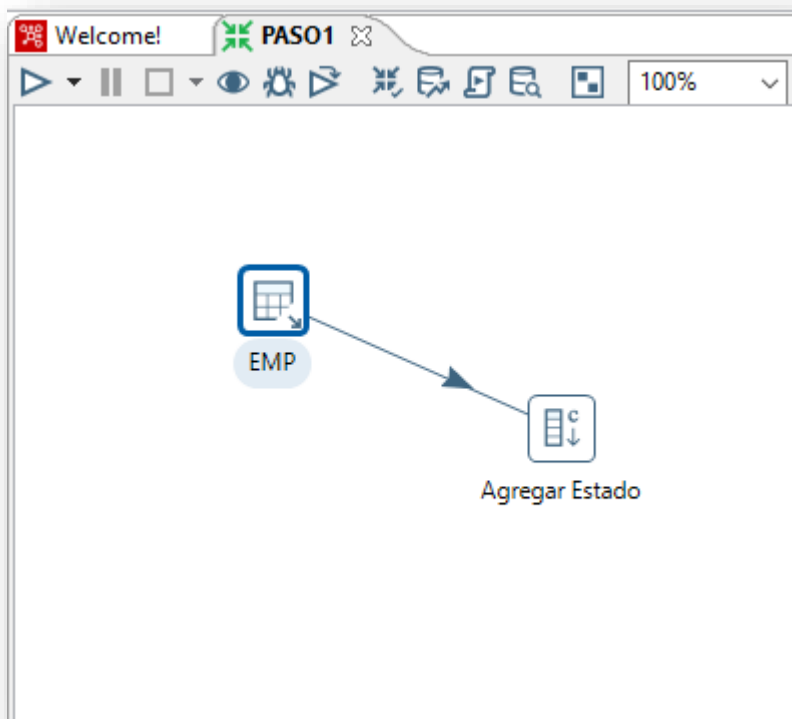
Step name:

Fields:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Value	Set empty string?
1	ESTADO	Integer		1					1	

Buttons: ? Help, OK, Cancel

Enlazar el nuevo paso



Ejecutamos la secuencia de pasos



Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.

**Run Options**

Run configuration:  
Pentaho local

Options

☒ Clear log before running      Log level: Basic

☐ Enable safe mode

☒ Gather performance metrics

Parameters   Variables

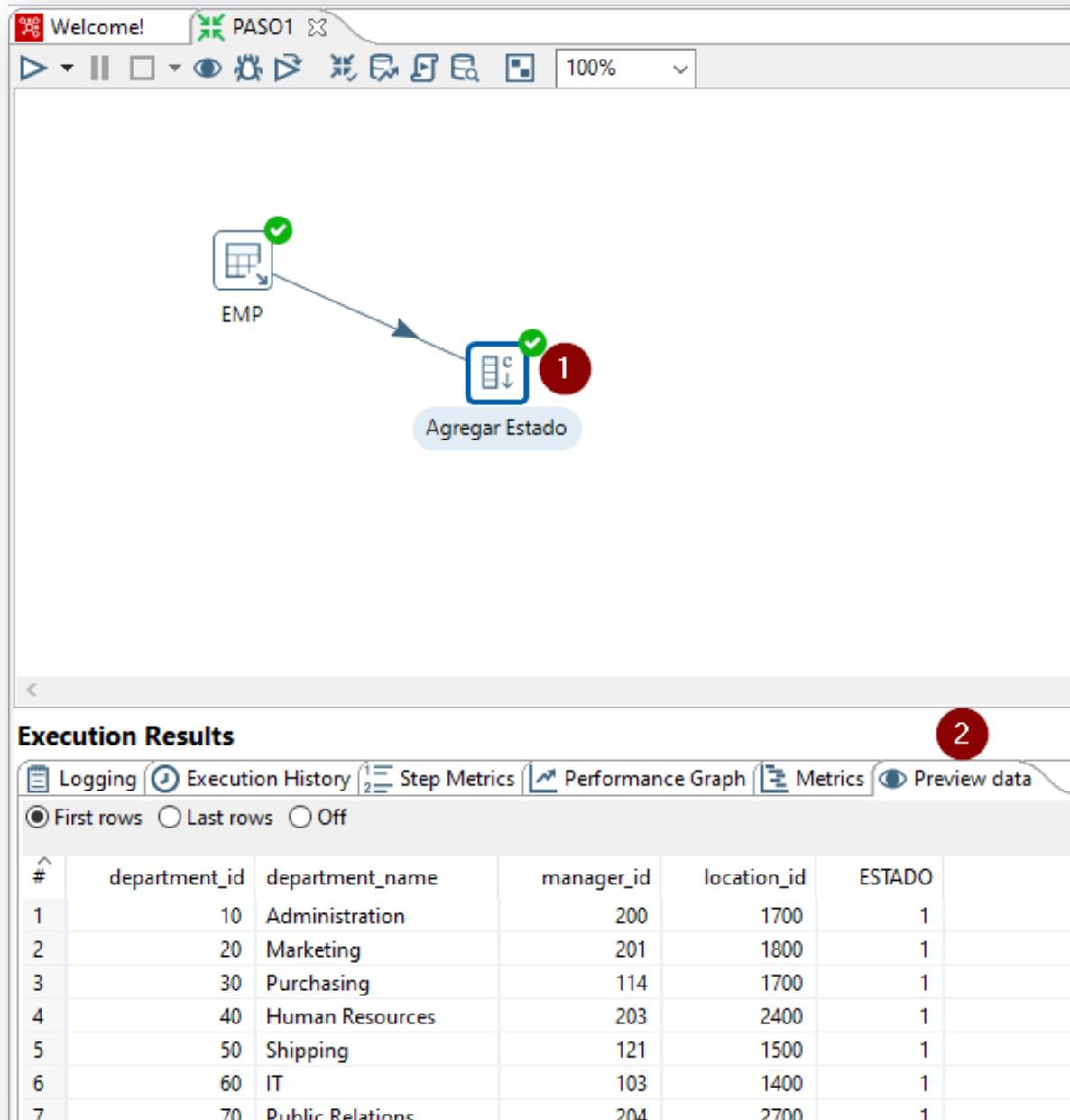
Parameter	Default value	Value	Description

Arguments (legacy)

☒ Always show dialog on run

Help      Run      Cancel

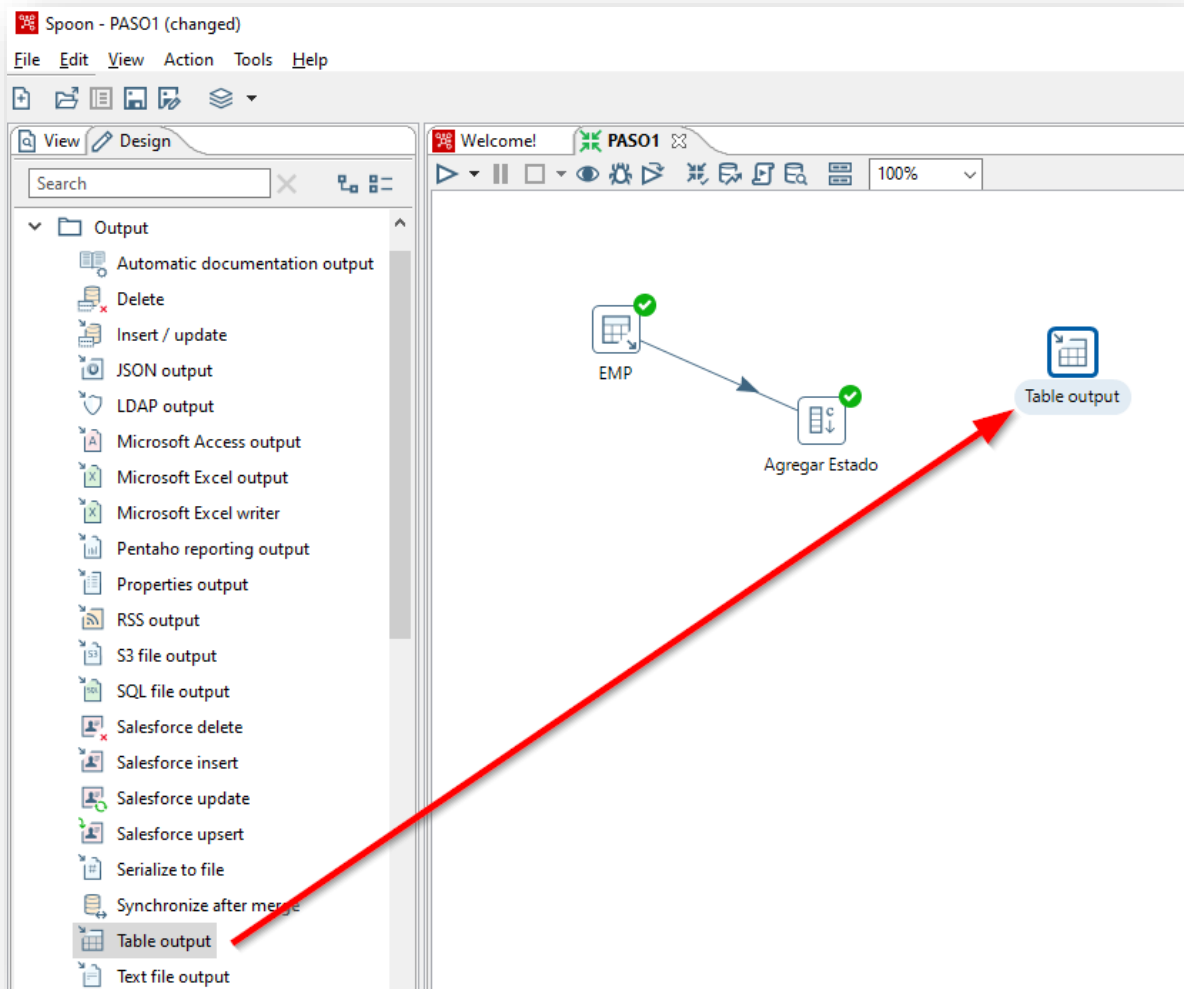
Revisamos los resultados del proceso



The screenshot shows a software interface with a workflow diagram and an execution results table. The workflow diagram has two steps: 'EMP' (marked with a green checkmark) and 'Agregar Estado' (marked with a green checkmark and a red circle with the number 1). An arrow points from 'EMP' to 'Agregar Estado'. Below the diagram is the 'Execution Results' section, which includes tabs for 'Logging', 'Execution History', 'Step Metrics', 'Performance Graph', 'Metrics', and 'Preview data'. The 'First rows' radio button is selected. The table below shows the execution results for 7 rows.

#	department_id	department_name	manager_id	location_id	ESTADO
1	10	Administration	200	1700	1
2	20	Marketing	201	1800	1
3	30	Purchasing	114	1700	1
4	40	Human Resources	203	2400	1
5	50	Shipping	121	1500	1
6	60	IT	103	1400	1
7	70	Public Relations	204	2700	1

Se crea una “Table output” de la sección “Output”



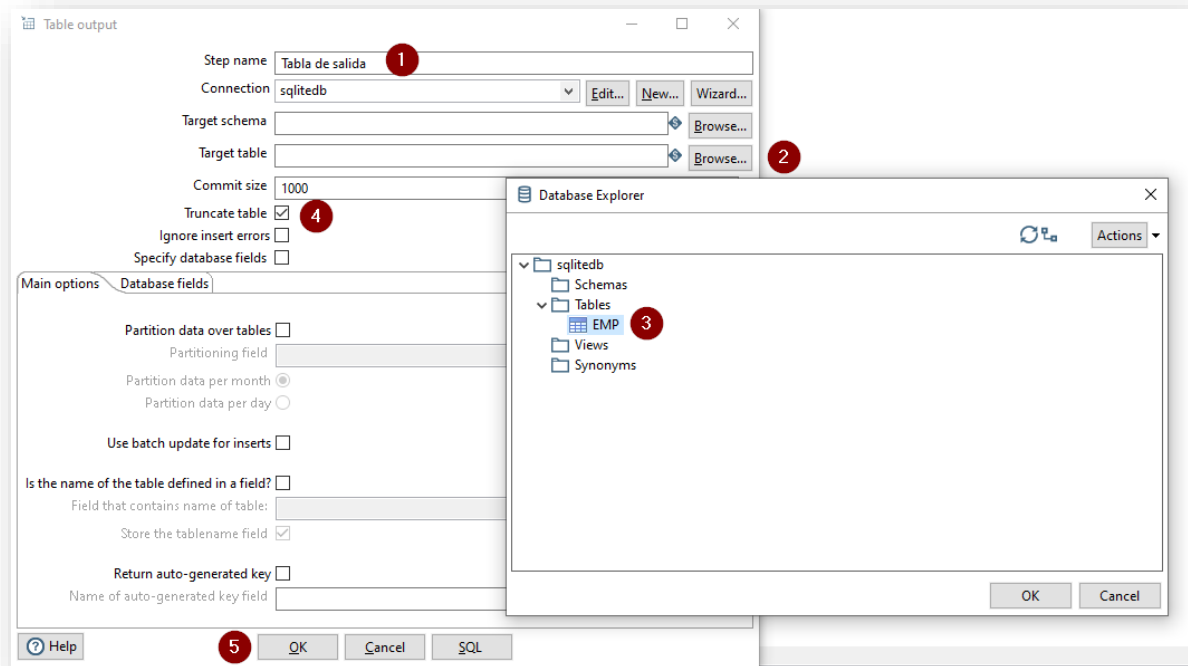
Debemos crear la tabla “EMP” en SQLite

```
CREATE TABLE EMP( department_id integer primary key  
, department_name text  
, manager_id integer  
, location_id INTEGER  
)
```

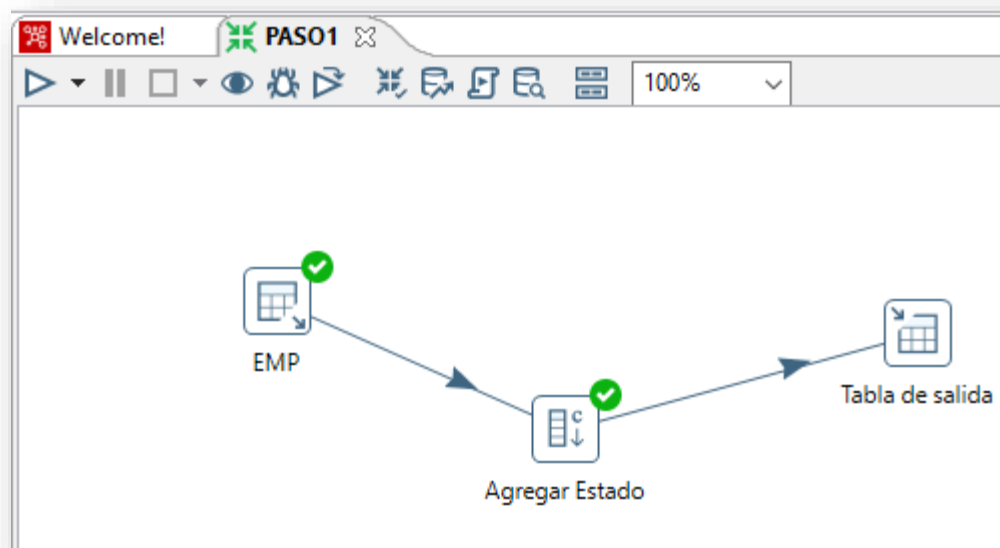
Configuramos el paso “Table output” y le damos el nombre de “Tabla de salida”



Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.



Realizamos el enlace y ejecutamos.





Miramos los datos migrados

Welcome! PASO1

100%

EMP

Agregar Estado

Tabla de salida

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

First rows Last rows Off

#	department_id	department_name	manager_id	location_id	ESTADO
1	10	Administration	200	1700	1
2	20	Marketing	201	1800	1
3	30	Purchasing	114	1700	1
4	40	Human Resources	203	2400	1
5	50	Shipping	121	1500	1
6	60	IT	103	1400	1
7	70	Public Relations	204	2700	1



Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.

Revisamos desde el visor “SQLite”

The screenshot shows the DB Browser for SQLite application. The 'Browse Data' tab is selected, and the 'EMP' table is chosen. The table contains 19 rows of data. A red circle with the number '1' is placed over the 'Open Database' button in the toolbar.

	department_id	department_name	manager_id	location_id	estado
1	10	Administration	200	1700	1
2	20	Marketing	201	1800	1
3	30	Purchasing	114	1700	1
4	40	Human Resources	203	2400	1
5	50	Shipping	121	1500	1
6	60	IT	103	1400	1
7	70	Public Relations	204	2700	1
8	80	Sales	145	2500	1
9	90	Executive	100	1700	1
10	100	Finance	108	1700	1
11	110	Accounting	205	1700	1
12	120	Treasury	NULL	1700	1
13	130	Corporate Tax	NULL	1700	1
14	140	Control And Credit	NULL	1700	1
15	150	Shareholder ...	NULL	1700	1
16	160	Benefits	NULL	1700	1
17	170	Manufacturing	NULL	1700	1
18	180	Construction	NULL	1700	1
19	190	Contracting	NULL	1700	1

**EVIDENCIA(S) A ENTREGAR:**

1. Cargar el DATASET que eligió en el primer trimestre en el aplicativo y aplicar las transformaciones de limpieza de datos correspondientes.



Servicio Nacional de Aprendizaje  
Formato Taller  
Centro de Gestión de Mercados, Logística y Tecnologías de la Información.

**CONTROL DEL DOCUMENTO**

	Nombre	Cargo	Dependencia	Fecha
<b>Autor (es)</b>	José Fernando Galindo Suarez	Instructor		05/04/2023

**CONTROL DE CAMBIOS** (diligenciar únicamente si realizan ajustes al taller)

	Nombre	Cargo	Dependencia	Fecha	Razón del Cambio
<b>Autor (es)</b>					