GUÍA PRÁCTICA
HERRAMIENTAS BÁSICAS DE OBTENCIÓN Y VISUALIZACIÓN DE DATOS

1. Datos Generales

Carrera:	Tecnología Superior en Big Data
Período académico:	Abril – Agosto 2024
Asignatura:	Marcos de Referencia a la Big Data
Unidad Nº:	1. Conceptos Básicos
Tema:	Herramientas básicas de obtención y visualización de datos
Ciclo-Paralelo:	M3A
Fecha de inicio de la Unidad:	15/04/2024
Fecha de fin de la Unidad	13/05/2024
Práctica Nº:	1
Horas:	9
Docente:	Mgtr. Verónica Paulina Chimbo Coronel
Elaborado por:	Victor Manuel Cabrera Barbecho

2. Contenido

2.1 Fundamentos

La Extracción, Transformación y Carga (ETL) de datos es un proceso esencial en el mundo de la gestión de datos, permitiendo la migración de información desde diversas fuentes hacia un sistema de almacenamiento centralizado. En esta guía, exploraremos cómo llevar a cabo un proceso ETL específico utilizando Pentaho Data Integration (también conocido como Kettle) para extraer datos de un archivo Excel que contiene información sobre taxis y cargarlos en una base de datos MySQL.

2.2 Objetivos de la Guía

- El objetivo principal de esta guía es proporcionar una visión detallada y paso a paso del proceso ETL, desde la configuración de Pentaho Data Integration hasta la carga efectiva de datos en una base de datos MySQL. En particular, nos enfocaremos en la conversión de datos provenientes de un archivo Excel, donde cada columna representa una categoría específica de información sobre los taxis.

2.3 Evaluación del Aprendizaje

Rúbrica de Evaluación de la Guía Práctica

Criterios de Evaluación	Puntuación Máxima
Base de Datos	3/3

Extracción de datos	1.5/3
Limpieza de datos	0,5/1
Llenado de Datos	0,5/1
Script	1,5/2
Puntuación Total	/10

2.4 Preparación previa, materiales, herramientas, equipos y software

- Pentaho Data Integration: Asegúrate de tener Pentaho Data Integration instalado en tu sistema.
- Base de Datos MySQL: Contar con una base de datos MySQL disponible para la carga de datos. Si aún no tienes MySQL instalado.
- Conocimientos Básicos de SQL: Será útil tener un entendimiento básico de SQL para configurar la conexión a la base de datos y realizar ajustes según sea necesario.

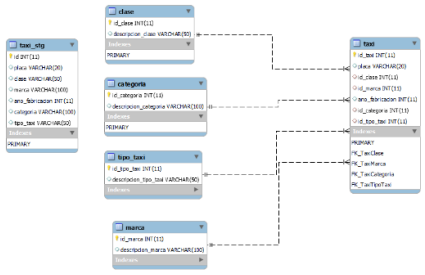
2.5 Procedimientos a emplear

En esta guía, abordaremos el proceso ETL para procesar datos de las Cooperativas de Taxis de Cuenca relacionados con la flota de taxis. Estos datos contienen información clave sobre los taxis, incluyendo detalles como la placa, clase, marca, año de fabricación, categoría y tipo de taxi. El objetivo es realizar la extracción, transformación y carga de estos datos en una base de datos MySQL.

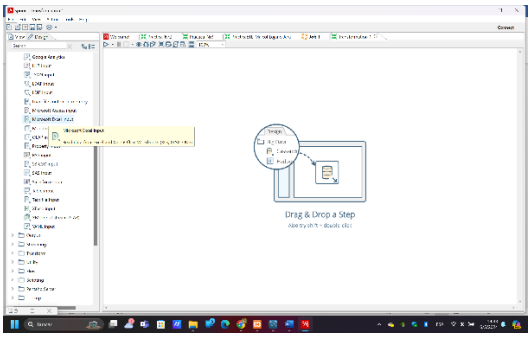
1. **Descargarse los datos del siguiente repositorio: https://docs.google.com/spreadsheets/d/1qDOA-ubd-2BNzR5qsxpa8z5pZ_JHKi3f/edit?usp=sharing&ouid=101015197702841688612&rtpof=true&sd=true y colocar en una carpeta denominada Practicas donde se desarrollará la Practica.**

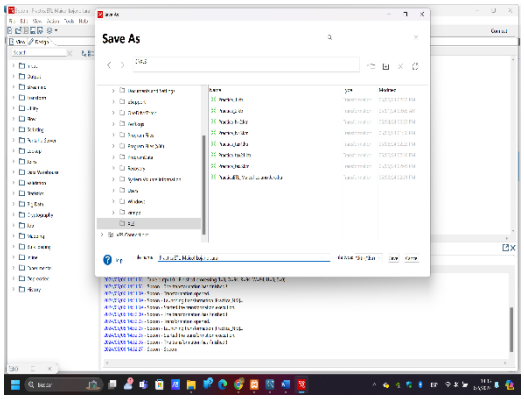
	A	B	C	D	E	F
1	Placa	Clase	Marca	Año	Categoría	Tipo taxi
2	AP-000024	Taxi Ejecutivo	Toyota	1990	Automovil	Sedan
3	AP-000033	Taxi Ejecutivo	Nissan	1990	Automovil	Sedan
4	AP-000035	Taxi Ejecutivo	Nissan	1990	Automovil	Sedan
5	AP-000049	Taxi Ejecutivo	Volvo	1990	Automovil	Sedan
6	AP-000074	Taxi Ejecutivo	Nissan	1990	Automovil	Sedan
7	AP-000080	Taxi Ejecutivo	Toyota	1989	Automovil	Noid
8	AP-000119	Taxi Ejecutivo	Toyota	1987	Automovil	Sedan
9	AP-000136	Taxi Ejecutivo	Nissan	1990	Automovil	Sedan
10	AP-000145	Taxi Ejecutivo	Toyota	1990	Automovil	Sedan
11	AP-000158	Taxi Ejecutivo	Lada	1985	Categoria no registrada	Sedan
12	AP-000160	Taxi Ejecutivo	Toyota	1990	Automovil	Noid
13	AP-000190	Taxi Ejecutivo	Zil	1980	Categoria no registrada	Noid
14	AP-000201	Taxi Ejecutivo	Toyota	1990	Automovil	Sedan
15	AP-000213	Taxi Ejecutivo	Toyota	1990	Automovil	Noid
16	AP-000216	Taxi Ejecutivo	Toyota	1990	Automovil	Sedan
17	AP-000229	Taxi Ejecutivo	Toyota	1986	Automovil	Noid

2. **Crear la base de datos denominada bd_taxis_cuenca con la siguiente estructura**

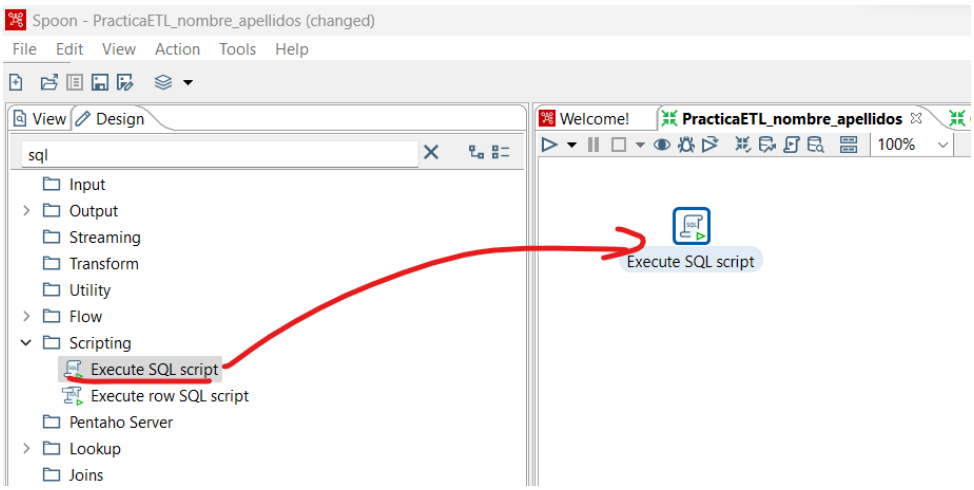


3. **Crear un proyecto nuevo para ello ir a file, nuevo, transformación con el nombre PracticaETL_nombres_apellidos**

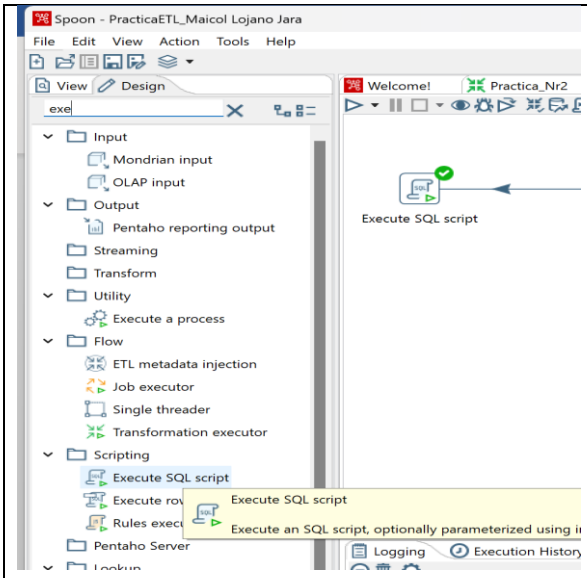




3.1. Agregar al área de trabajo el elemento Execute SQL script de la categoría Scripting para limpiar los datos de la base de datos.

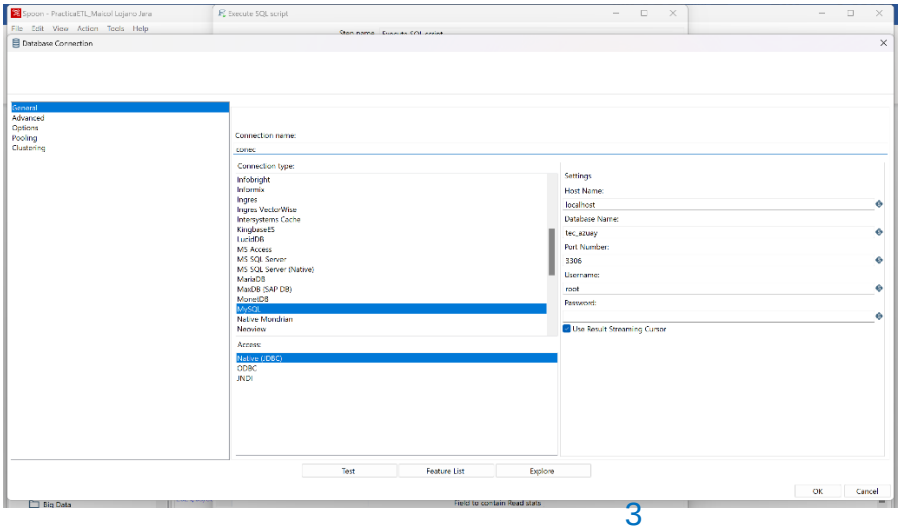


De igual manera arrastramos hacia la derecha.



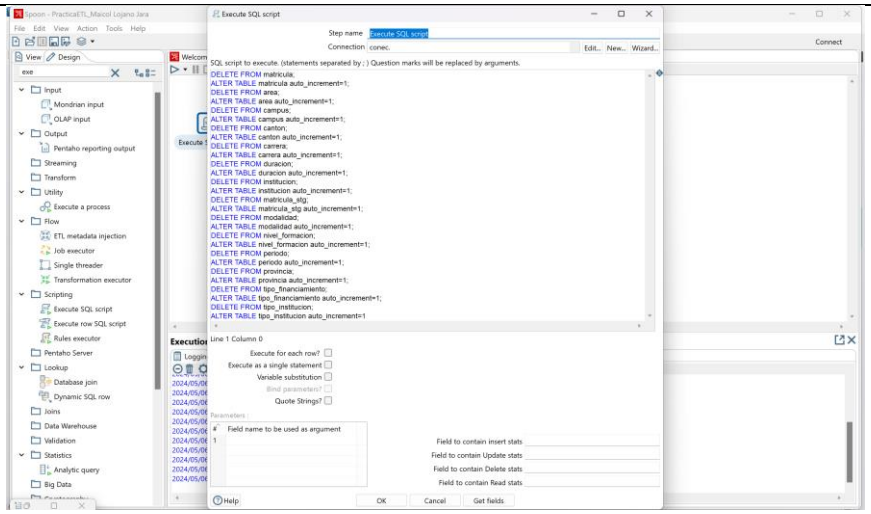
Configurar el **Execute SQL script** dando doble click sobre el icono.

- **Dar click en testear para verificar la conexion.**

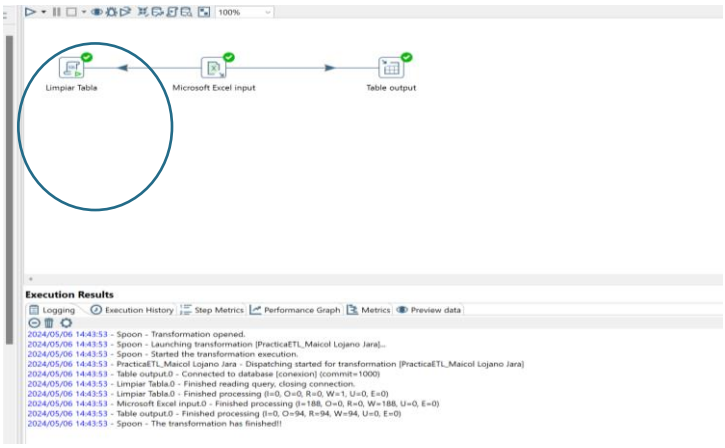


3

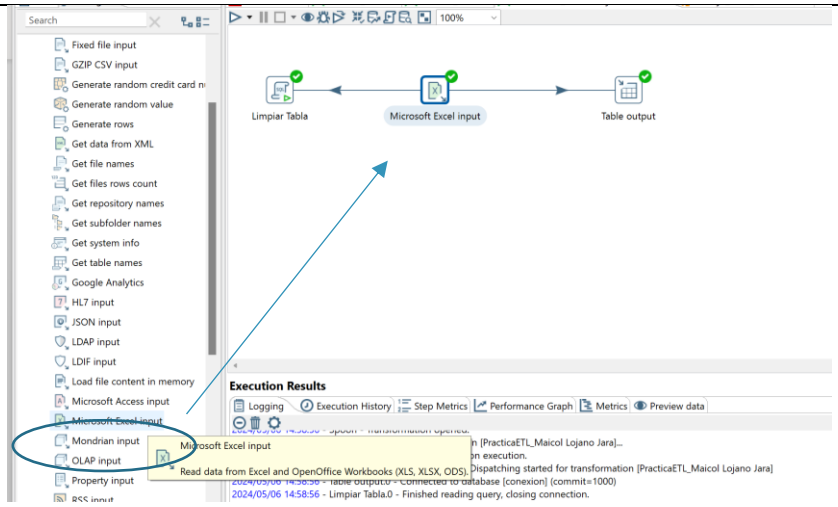
1. Renombrar el Excute con Limpiar Tablas
2. Crear la conexión a la base de datos y asegurarse de llenar todos los campos requeridos.
3. Realizar un test de conexión
4. Realizar las sentencias sql para eliminar los datos de las tablas y asignarles el **AUTO_INCREMET** en 1



5. Comprobar que funcione correctamente la configuración dar click en ejecutar y verificar que se marque en color verde.



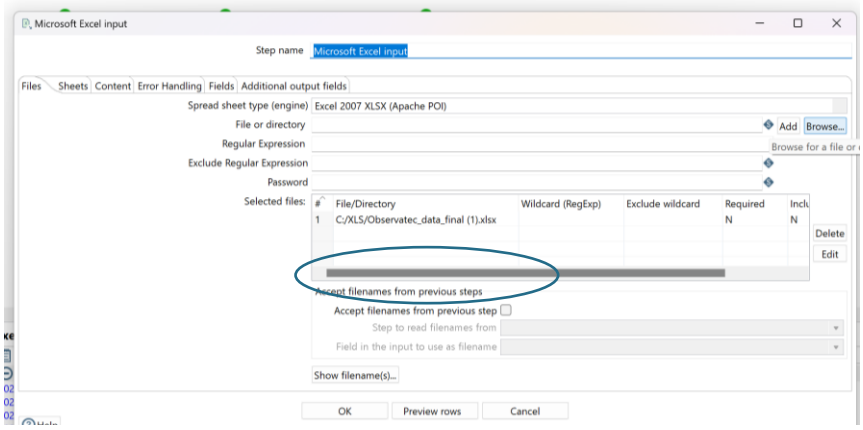
3.2. Cargar los datos del Excel Datos_taxi_Cuenca.xls para ello seleccionaremos de la categoría input e elemento Microsoft Excel input



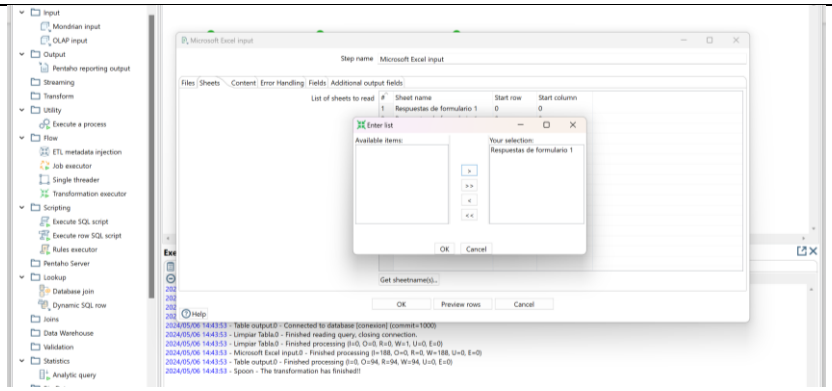
1.- Dar click derecho sobre el icono y realizar las configiuraciones, asignarle el nombre de Extracción de Datos del Dataset.

2.- Buscar el archivo de Excel de la data

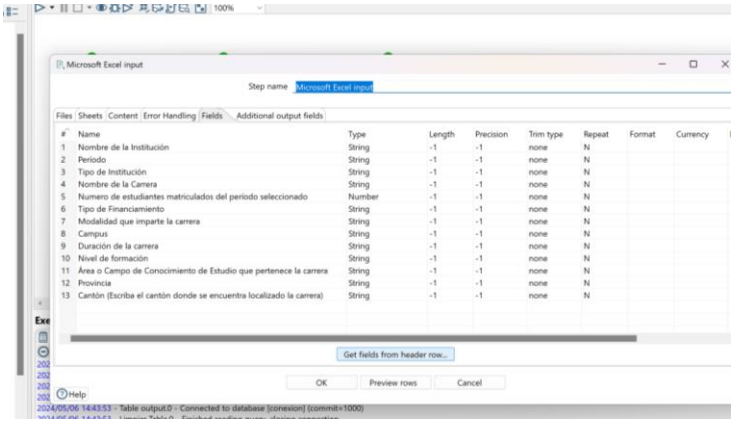
3.- Agregar al panel de selección de files



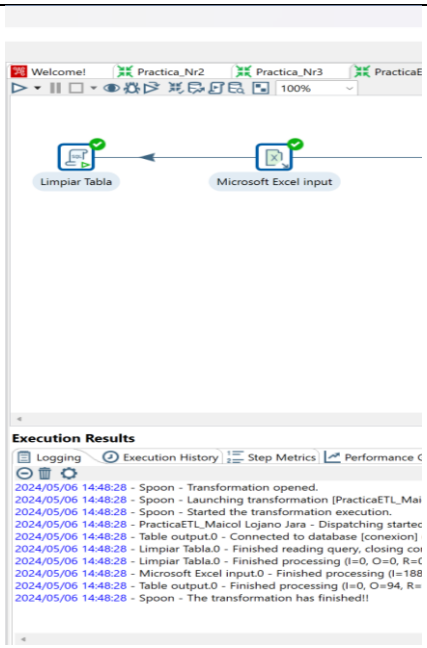
4. Agregar los sheets



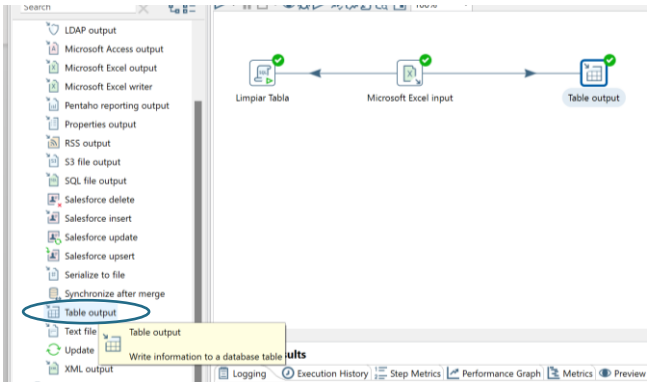
5.- Agregar los fields



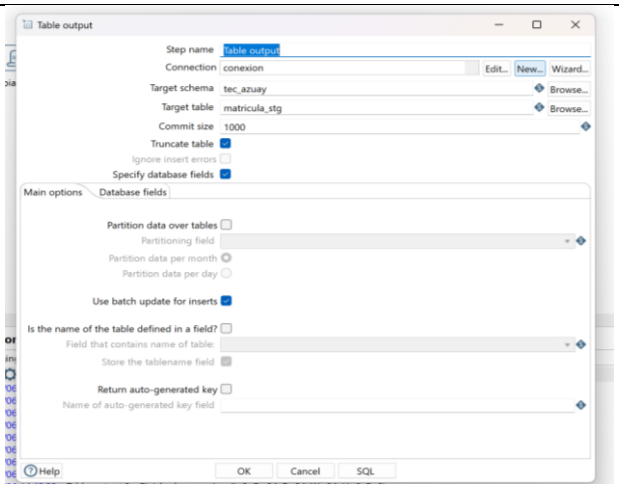
6.- Conectar y probar



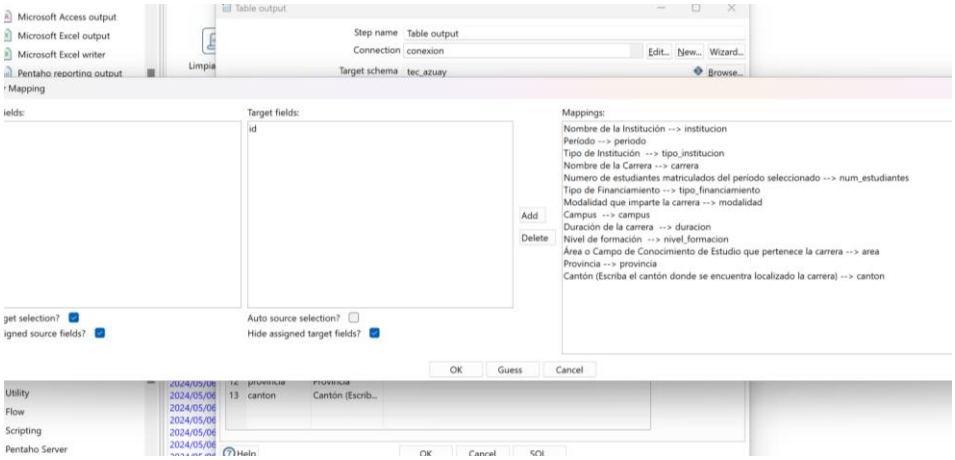
3.3. Pasar los datos cargados del Excel a la tabla taxi_stg para ello utilizaremos el elemento Table output de la categoría Output.



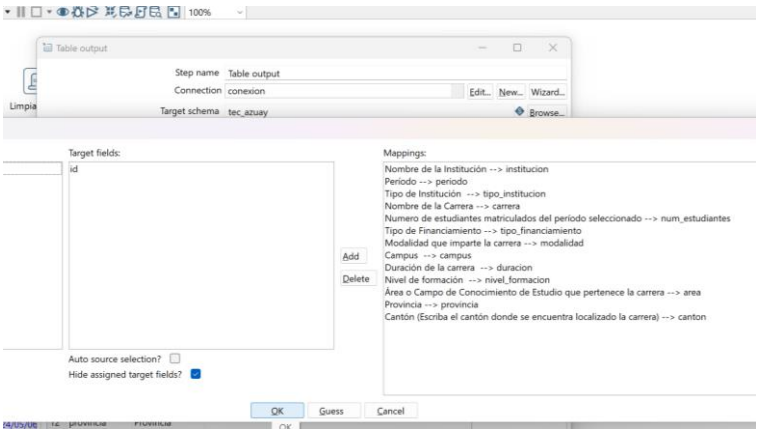
1. Dar doble click y configurar los datos de la conexión, base de datos y la tabla donde se guardrán los datos.



2. Seleccionar los fields y relacionar

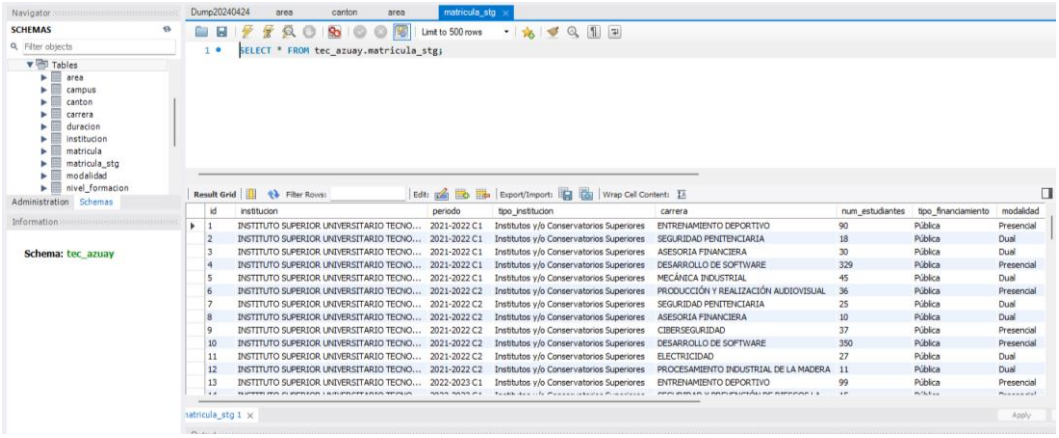


3. Una vez relacionada los fieds de la data con los de la tabla. Finalmente presionar ok

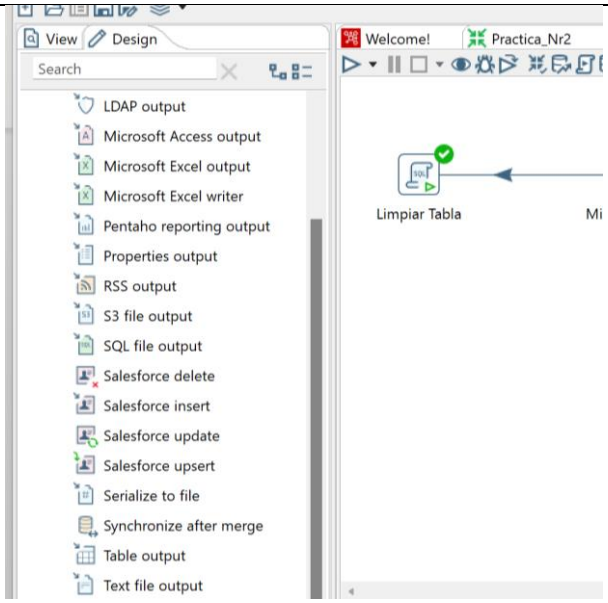


4. Comprobar enlazando los elementos y ejecutando para comrpbar que el trabajo es exitoso se debe de hacer un select a la tabla en la herramienta visual de la base de datos en nuestro caso estamos utilizando MysqlWorkbeanch.

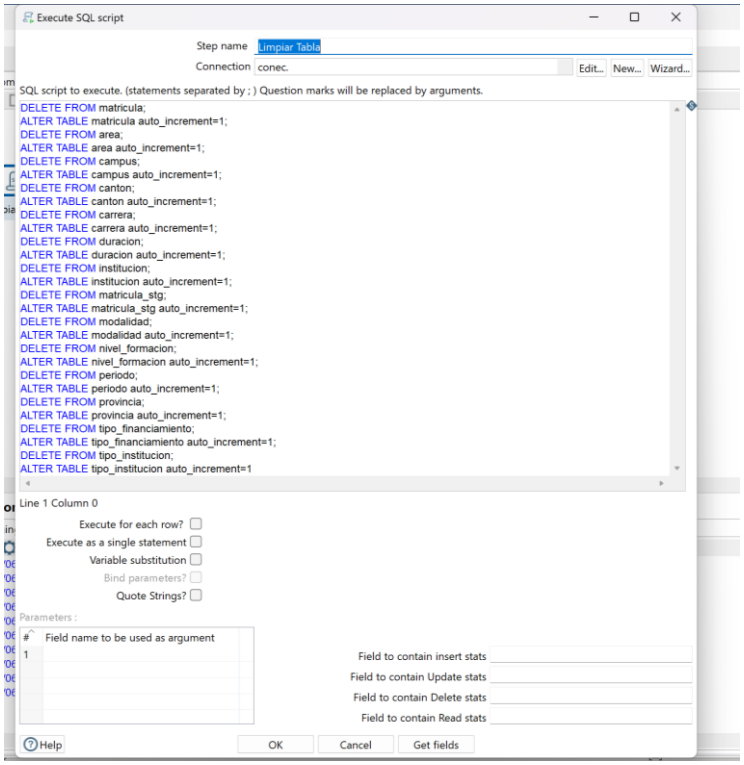
Verificación en la base de datos



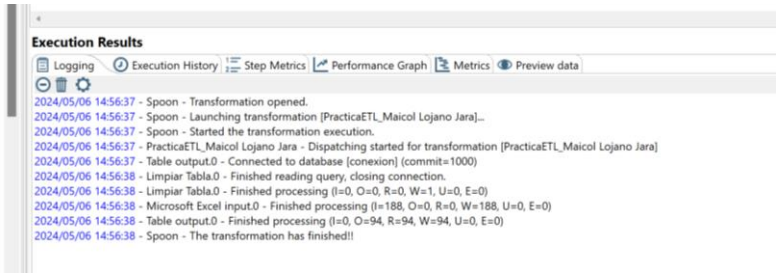
3.4. Limpiar información inválida para ello se debe de crear un archivo nuevo con nombre Limpieza de la data y seleccionar el elemento Execute SQL script de la categoría Scripting.



Configurar el elemento para ello dar doble click renombrar el elemento, crear la conexión a la base de datos y agregar las sentencias sql donde se elimine los valores inválidos como: donde el año de fabricación sea 0, la marca tenga como valor Marca no registrada, en categoría Categoria no registrada y en tipo de taxi Noid.



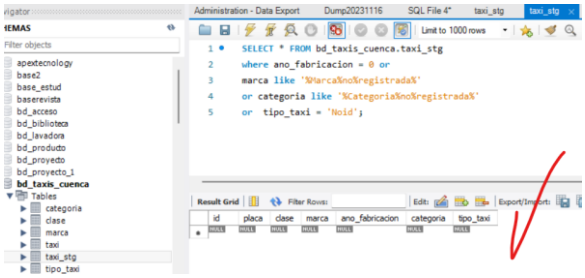
Ejecutar la tarea.



Para verificar realizar lo siguiente en el gestor de base de datos ejecutar la siguiente sentencia:

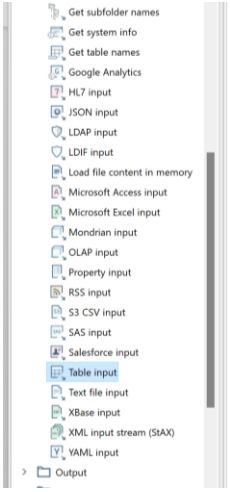
```
SELECT * FROM bd_taxis_cuenca.taxi_stg
where ano_fabricacion = 0 or
marca like '%Marca%no%registrada%'
or categoria like '%Categoria%no%registrada%'
or  tipo_taxi = 'Noid';
```

Si se ejecuto correctamente debe de salir

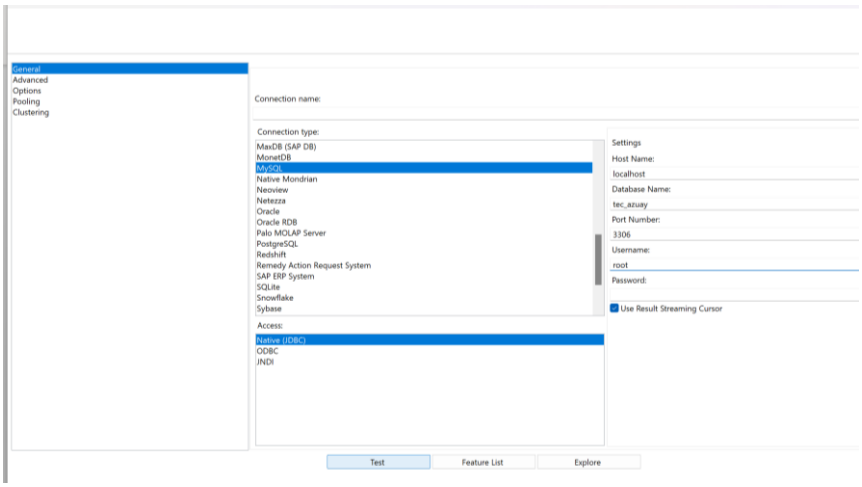


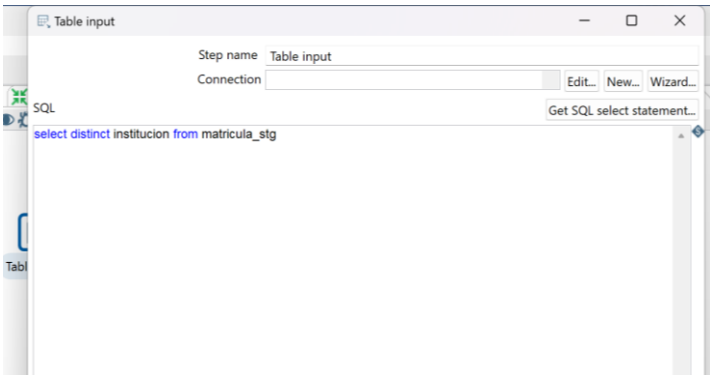
3.5. Llenar los datos en las tablas de la base de datos para ello crear un nuevo archivo con el nombre Llenar tablas de la base de datos y realizar los siguientes pasos:

Agregar un elemento Table input de la categoría input al área de trabajo

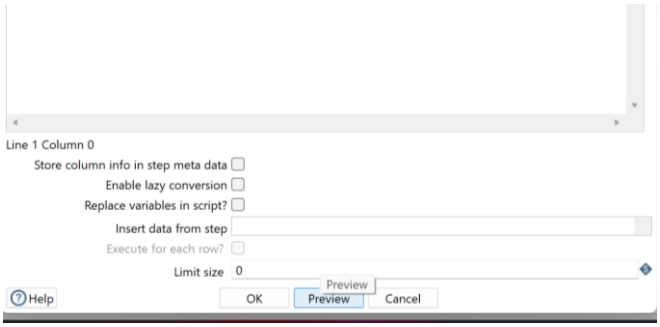


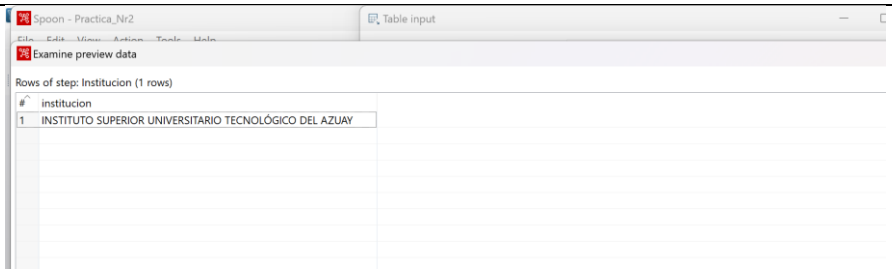
Dar doble click sobre el elemento, renombrar con el nombre Seleccionar clases, verificar la conexión a la base de datos y agregar la siguiente sentencia sql.



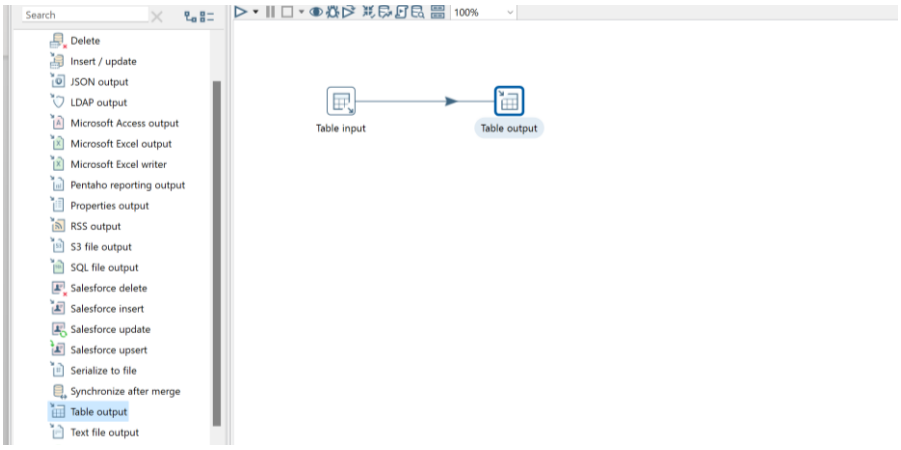


Verificar que funcione correctamente que funcione la sentencia sql para seleccionar los valore únicos de clases de taxis.

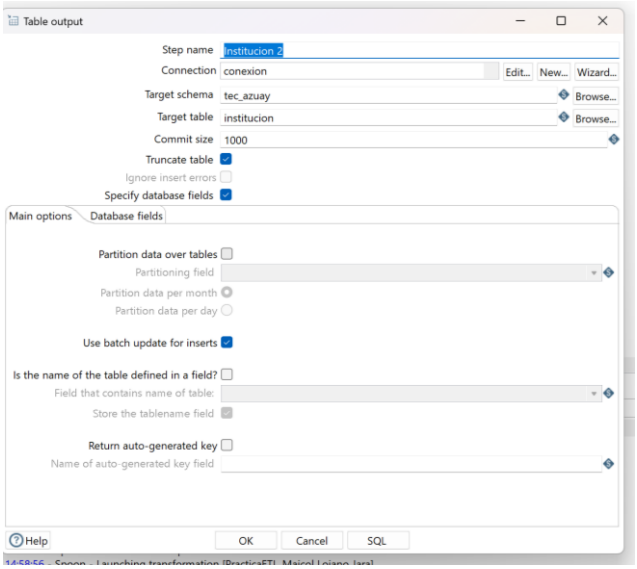




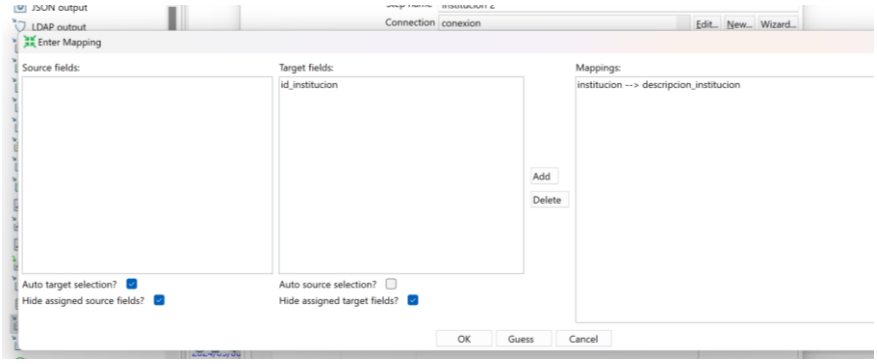
Luego crear las clases en la tabla clase para ello agregar al área de trabajo un Table output y conectar al elemento anterior.



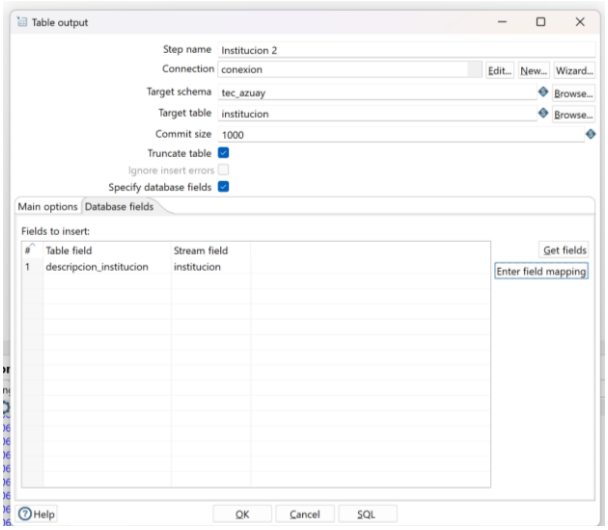
Configurar el Table output llenar los campos requeridos renombrar con el nombre Llenar clases y verificar los demás campos así como la conexión a la base de datos.



Mapear los campos como lo muestra la siguiente imagen

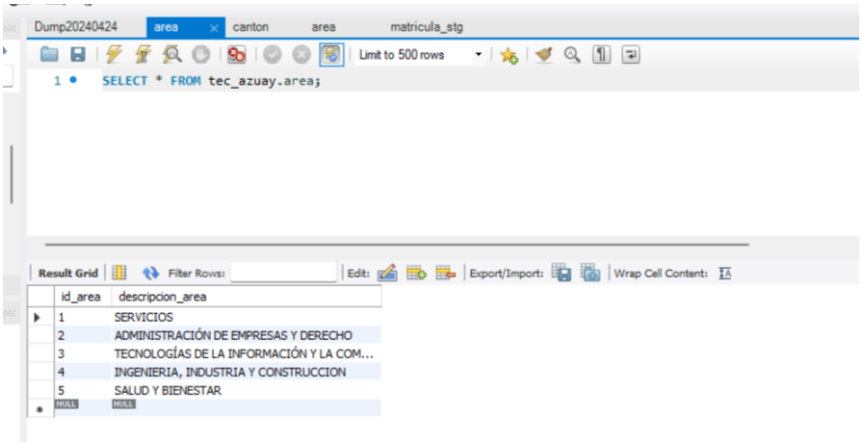


Verificar que se haya cargado el campo a guardar y finalizar.



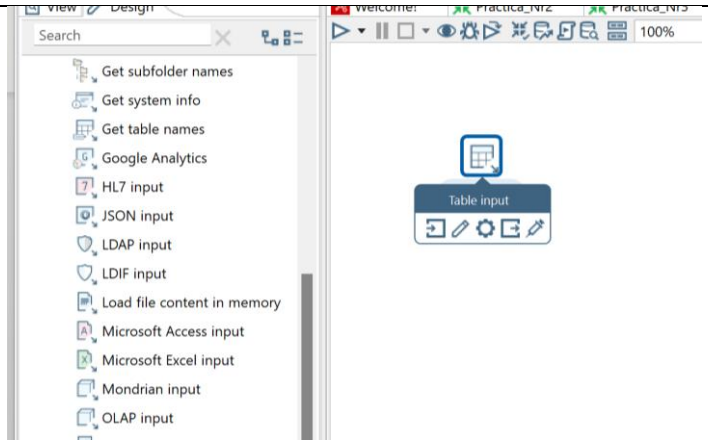
Realizar el mismo procedimiento cambiando las sentencias según corresponda de cada tabla.

Ejecutar y comprobar que se hayan llenado las tablas en la base de datos.



3.4. Cargar los dato de taxi para ello crear un nuevo archivo con el nombre Llenar Tabla Taxi y realizar las siguientes acciones:

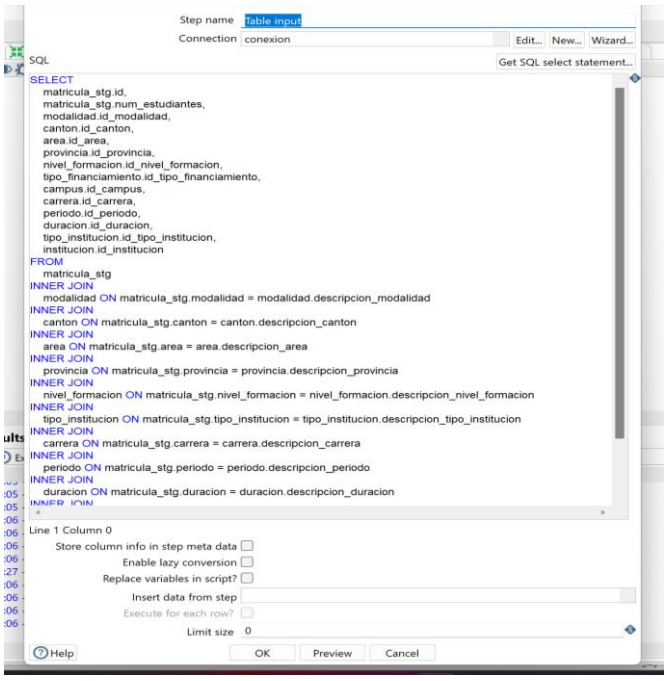
Arrastrar al área de trabajo le elemento table input en el cual seleccionaremos los datos a guardar en la tabla taxi.

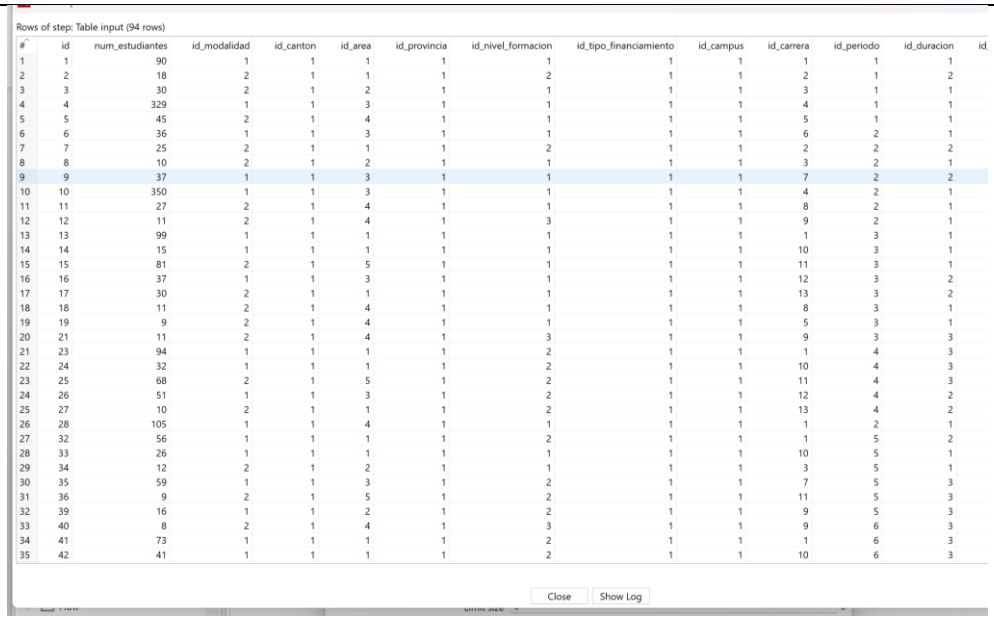


Renombrar con el nombre Seleccionar taxis, crear la conexión a la base de datos y agregar la siguiente sentencia.

```
select placa, id_clase, id_marca, ano_fabricacion, id_categoria, id_tipo_taxi from taxi_stg
join clase on taxi_stg.clase = clase.descripcion_clase
join marca on taxi_stg.marca = marca.descripcion_marca
join categoria on taxi_stg.categoria = categoria.descripcion_categoria
join tipo_taxi on taxi_stg.tipo_taxi = tipo_taxi.descripcion_tipo_taxi
```

Luego finalmente colocar un Preview de los datos y dar click en ok.



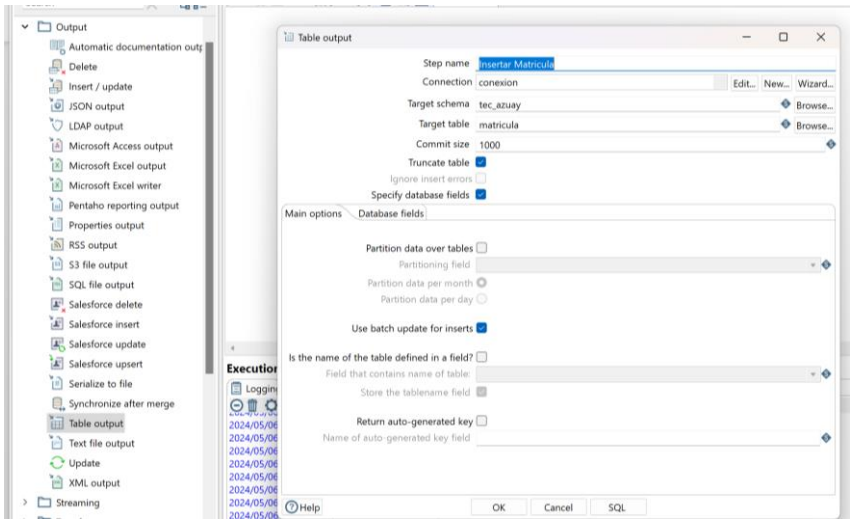


Rows of step: Table input (94 rows)

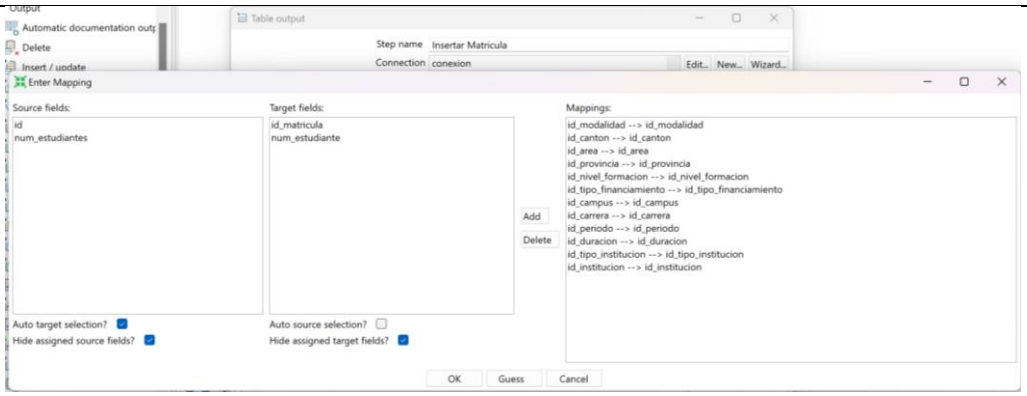
#	id	num_estudiantes	id_modalidad	id_canton	id_area	id_provincia	id_nivel_formacion	id_tipo_financiamiento	id_campus	id_carrera	id_periodo	id_duracion	id
1	1	90	1	1	1	1	1	1	1	1	1	1	
2	2	18	2	1	1	1	2	1	1	2	1	2	
3	3	30	2	1	2	1	1	1	1	3	1	1	
4	4	329	1	1	3	1	1	1	1	4	1	1	
5	5	45	2	1	4	1	1	1	1	5	1	1	
6	6	36	1	1	3	1	1	1	1	6	2	1	
7	7	25	2	1	1	1	2	1	1	2	2	2	
8	8	10	2	1	2	1	1	1	1	3	2	1	
9	9	37	1	1	3	1	1	1	1	7	2	2	
10	10	350	1	1	3	1	1	1	1	4	2	1	
11	11	27	2	1	4	1	1	1	1	8	2	1	
12	12	11	2	1	4	1	3	1	1	9	2	1	
13	13	99	1	1	1	1	1	1	1	1	3	1	
14	14	15	1	1	1	1	1	1	1	10	3	1	
15	15	81	2	1	5	1	1	1	1	11	3	1	
16	16	37	1	1	3	1	1	1	1	12	3	2	
17	17	30	2	1	1	1	1	1	1	13	3	2	
18	18	11	2	1	4	1	1	1	1	8	3	1	
19	19	9	2	1	4	1	1	1	1	5	3	1	
20	21	11	2	1	4	1	3	1	1	9	3	3	
21	23	94	1	1	1	1	2	1	1	1	4	3	
22	24	32	1	1	1	1	2	1	1	10	4	3	
23	25	68	2	1	5	1	2	1	1	11	4	3	
24	26	51	1	1	3	1	2	1	1	12	4	2	
25	27	10	2	1	1	1	2	1	1	13	4	2	
26	28	105	1	1	4	1	1	1	1	1	2	1	
27	32	56	1	1	1	1	2	1	1	1	5	2	
28	33	26	1	1	1	1	1	1	1	10	5	1	
29	34	12	2	1	2	1	1	1	1	3	5	1	
30	35	59	1	1	3	1	2	1	1	7	5	3	
31	36	9	2	1	5	1	2	1	1	11	5	3	
32	39	16	1	1	2	1	2	1	1	9	5	3	
33	40	8	2	1	4	1	3	1	1	9	6	3	
34	41	73	1	1	1	1	2	1	1	1	6	3	
35	42	41	1	1	1	1	2	1	1	10	6	3	

Close Show Log

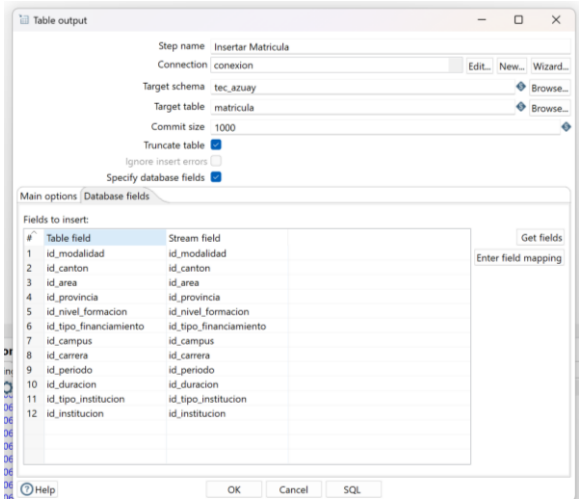
El siguiente paso final es cargar los datos en la tabla taxi pare ello arrastrar el elemento Table output y cargar los datos en la tabla taxi.



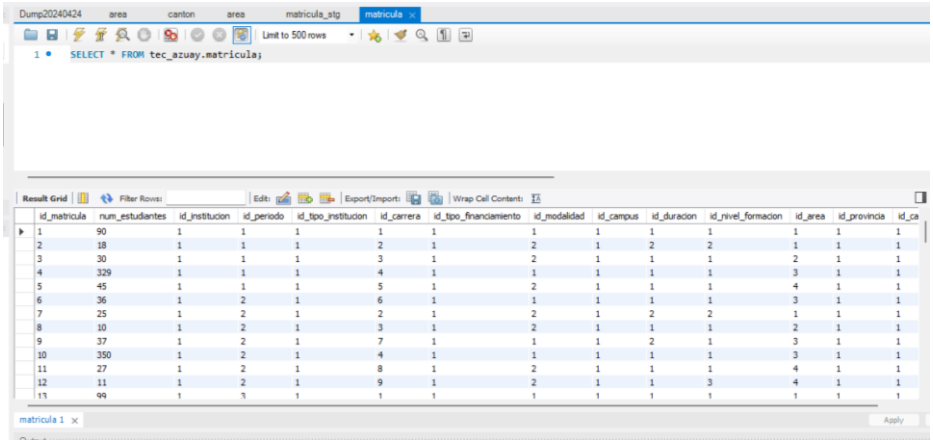
Realizadas las configuraciones pertinentes seleccionar lo fields para ello se debe de mapear los datos:

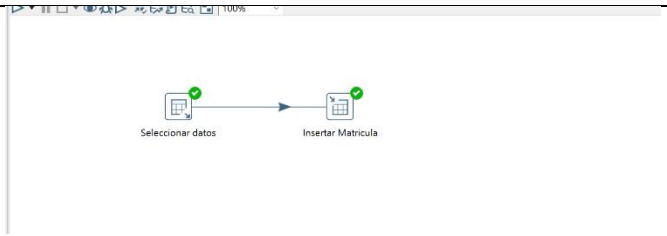


Verificar que se hayan agregado los fiedls a insertar y guardar la configuración.



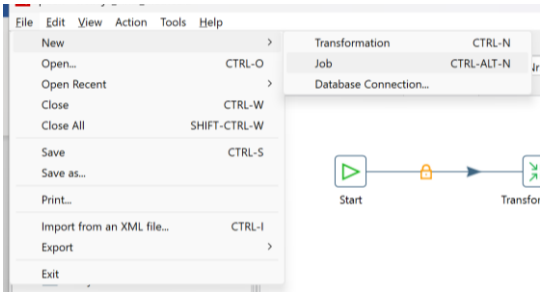
Finalmente ejecutar y verificar que se hayan guardado los datos en la tabla realizando un select sobre taxi.



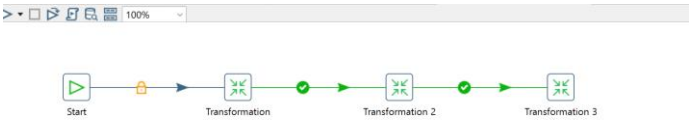


Para no estar ejecutando uno a uno los archivos de transformación se puede adjuntarlos en un archivo de trabajo para ello.

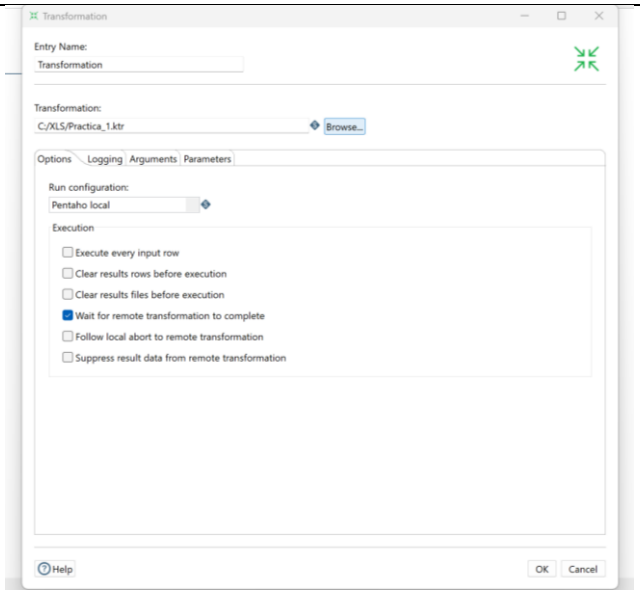
Crear un archivo de trabajo



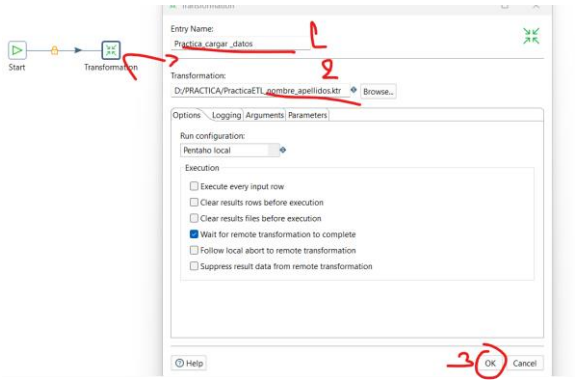
Guardarlo con el nombre Trabajo Final



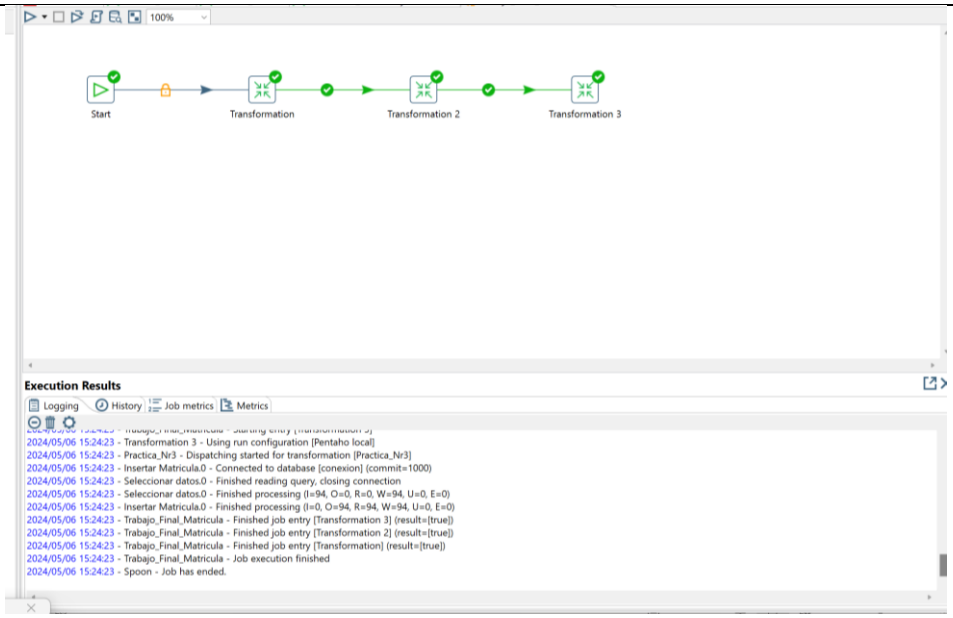
Agregar al área de trabajo los elementos start y configurar el transformation para realizar la secuencia de ejecución.



Renombrar con nombre Cargar datos y hacer referencia al primer archivo creado en la sección 3. De la misma manera agregar mas transformaciones y hacer referencia a los archivos creados en el mismo orden de creación.



Finalmente ejecutar y verificar los datos



2.6 Normas de Seguridad

Las normas de seguridad se han tomado del reglamento general de seguridad para el uso de los talleres, aulas y laboratorios del Instituto Superior Universitario Tecnológico del Azuay.
El estudiante, al ingresar a los talleres o laboratorios, está sujeto a este reglamento; y, tendrá la supervisión del profesor y del personal técnico; será responsable de:
a) Usar los EquipoS de Protección Personal(EPP) de acuerdo con lo establecido en la “Matriz de equipos de protección individual (EPP`s) requeridos para el ingreso de estudiantes y profesores a los laboratorios y talleres del INSTITUTO”;
b) Al inicio de cada práctica, recibir y revisar el material y herramientas requeridas para la Práctica, serán responsables de su buen uso.
c) La operación de los equipos por los estudiantes deberá ser con el conocimiento de su funcionamiento y bajo las directrices del profesor o personal técnico del laboratorio o taller; bajo ninguna circunstancia el estudiante podrá trabajar solo y sin vigilancia;
d) Seguir las instrucciones dadas por el docente o el personal técnico de apoyo;
e) Al término de la práctica, entregar limpio tanto el material como su área de trabajo;
f) Informar inmediatamente al profesor o personal técnico de apoyo, cualquier desperfecto que se localice en los equipos e instalaciones.

2.7 Resultados esperados

- El estudiante realiza paso a paso del proceso ETL, desde la configuración de Pentaho Data Integration hasta la carga efectiva de datos en una base de datos MySQL.

2.8 Bibliografía

Descripción en norma APA
• Borland, B. (Año de publicación). Pentaho Analytics for MongoDB. Editorial.

- Ramazzina, S. (Año de publicación). Pentaho Business Analytics Cookbook. Editorial.
- Casters, M., Bouman, R., & van Dongen, J. (Año de publicación). Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration. Editorial.
- Roldán, M. C. (Año de publicación). Pentaho Data Integration Beginner's Guide - Second Edition. Editorial.

3. Firmas de Responsabilidad

ESTUDIANTE	DOCENTE	DIRECTORA DE CARRERA
Nombre: **Firma**	**Nombre:** Mgtr. Verónica Chimbo **Firma**	**Nombre:** Mgtr. Verónica Chimbo **Firma**
Fecha: ( )	**Fecha: (11/04/2024)**	**Fecha: ( )**