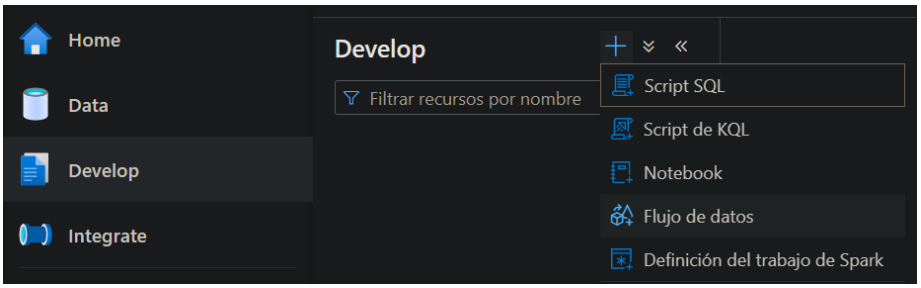
BIG DATA
ACADEMY

LABORATORIO 45
PROCESO TO_SILVER PARA REGLAS DE CALIDAD CON DATAFLOW

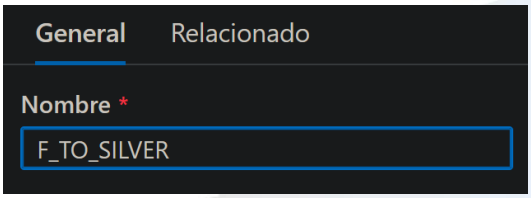
FORMADOR: ALONSO MELGAREJO
alonsoraulmgs@gmail.com

LABORATORIO 45 - PROCESO TO_SILVER PARA REGLAS DE CALIDAD CON DATAFLOW

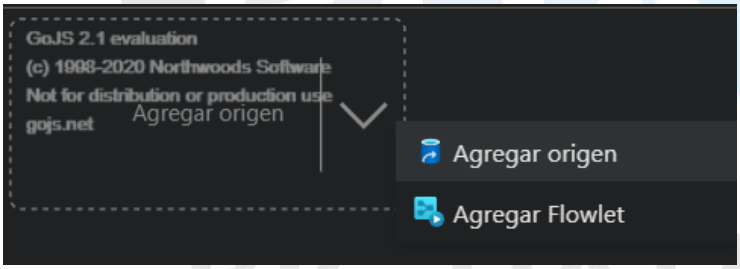
1. Crearemos un flujo de limpieza de datos con Dataflow. Seleccionamos la opción “Develop / Flujo de datos”



2. De nombre de flujo ingresamos “F_TO_SILVER”



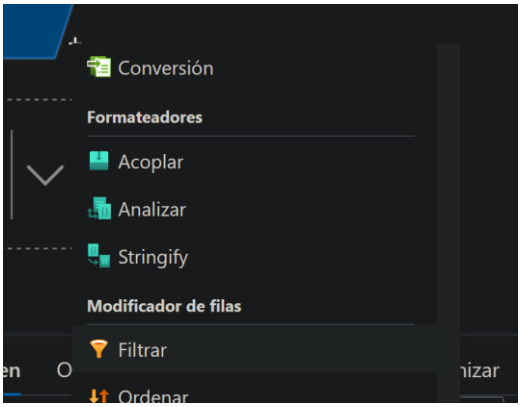
3. En “Agregar origen”, seleccionamos “Agregar origen”



Configuramos:

Tipo de origen	Conjunto de datos de integración
Conjunto de datos	BRONZE_PERSONA

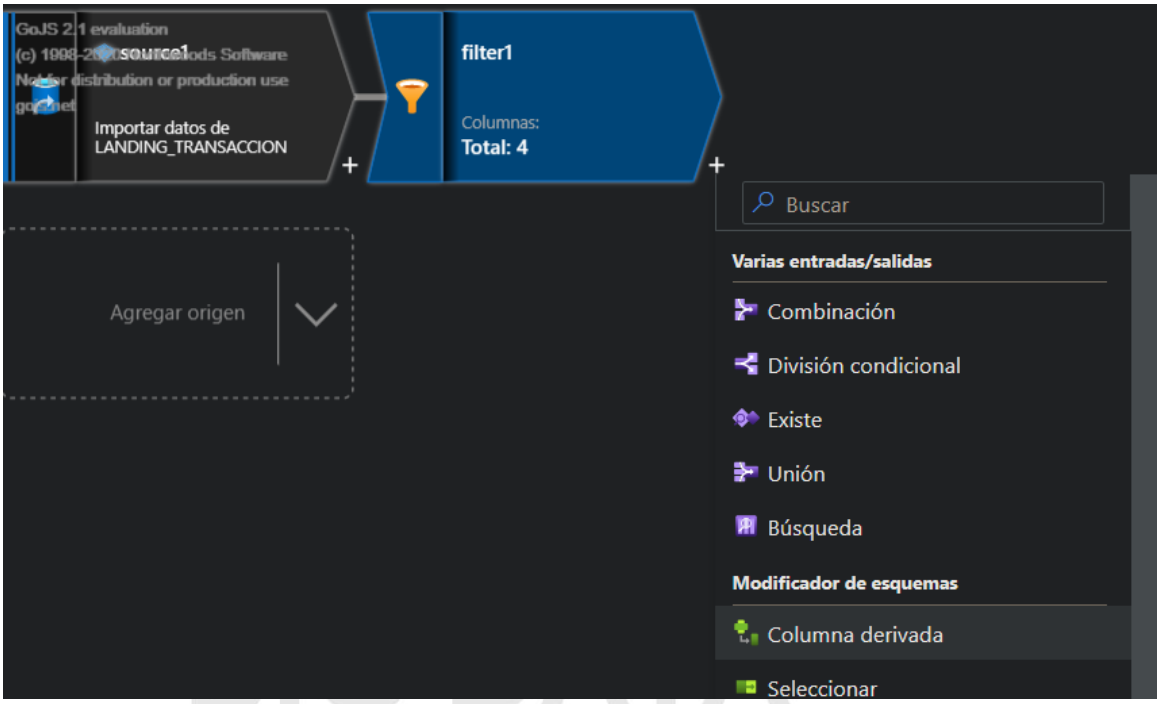
4. Agregaremos un filtro para limpiar los registros, seleccionamos “+ / FIltrar”:



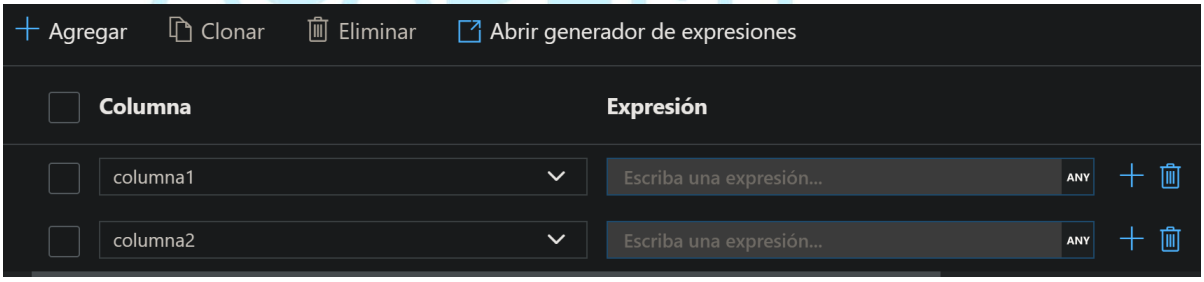
Configuramos

Filtro activado	!isNull(ID) && !isNull(ID_EMPRESA) && isInteger(EDAD) && isDecimal(SALARIO) && toInteger(EDAD) > 0 && toDecimal(SALARIO) > 0

5. Castearemos los datos a los tipos de datos correctos, seleccionamos "+ / columna derivada"



6. Desde la opción "+ Agregar", agregamos dos columnas:



Y colocamos los siguientes nombres a cada columna:

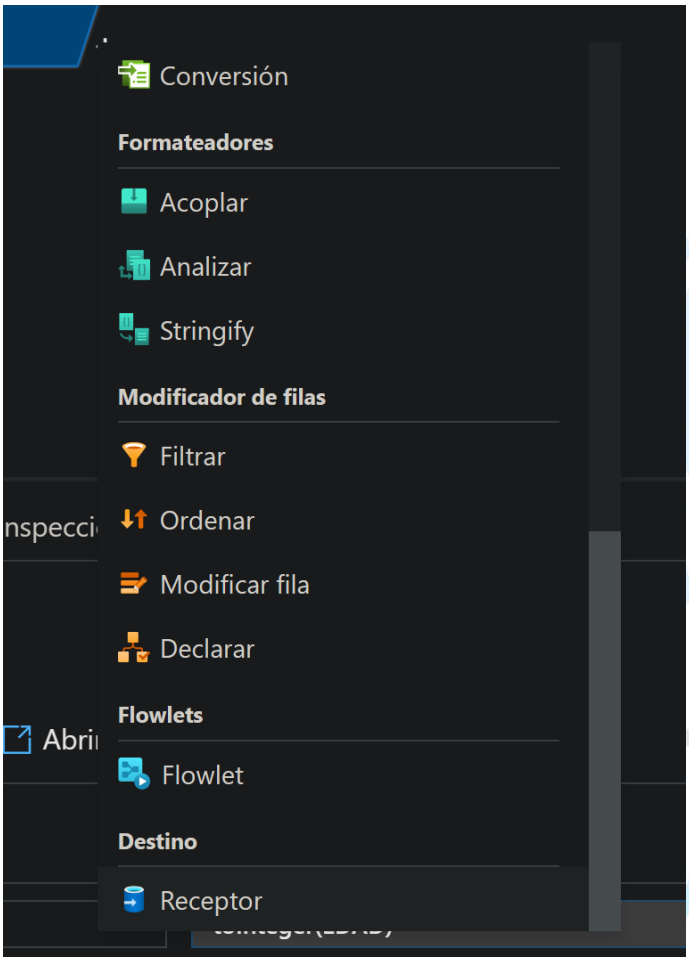
COLUMNA
EDAD
SALARIO

7. En cada columna colocamos el siguiente código

EDAD	toInteger(EDAD)
SALARIO	toDecimal(SALARIO)

Seleccionamos “Guardar y finalizar”

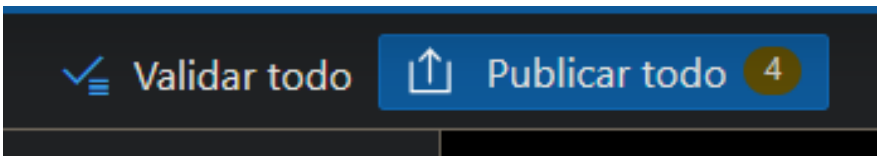
8. Agregaremos el receptor en donde se escribirán los registros limpios y casteados. Damos clic en “+ / Receptor”



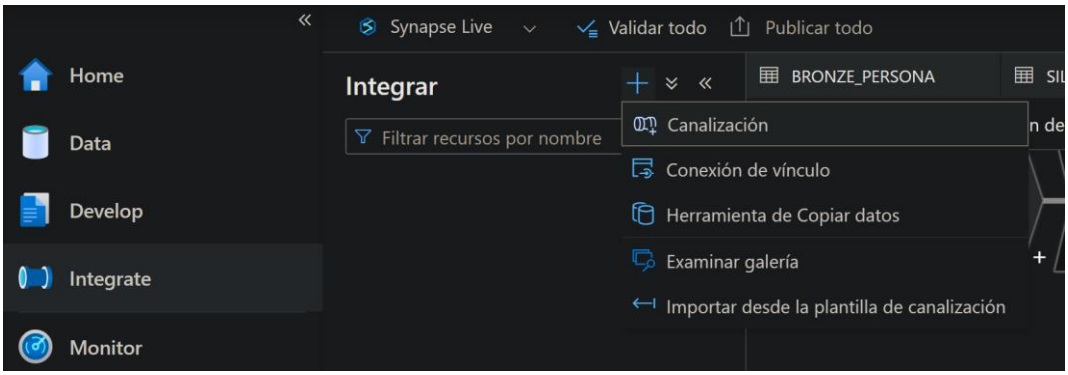
Configuramos:

Tipo de receptor	Conjunto de datos de integración
Conjunto de datos	SILVER_PERSONA

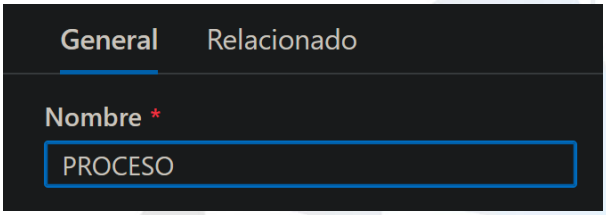
9. Validamos y publicamos los cambios



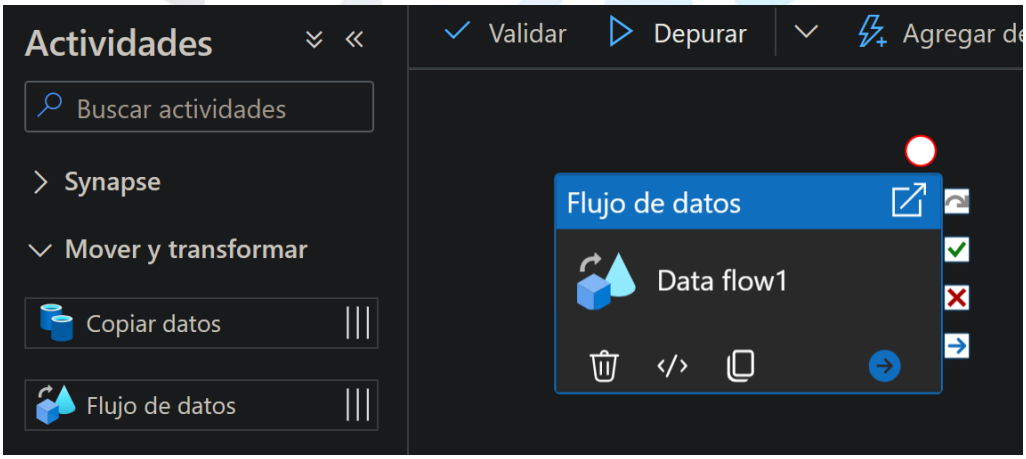
10. Crearemos el pipeline que ejecute el proceso, cerramos la pestaña “F_TO_SILVER”. Seleccionamos “Integrate / + / Canalización”



11. De nombre de pipeline ingresamos “PROCESO”



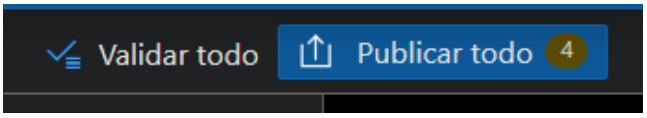
12. Desde el conjunto de actividades “Mover y transformar” agregamos la actividad “Flujo de datos”



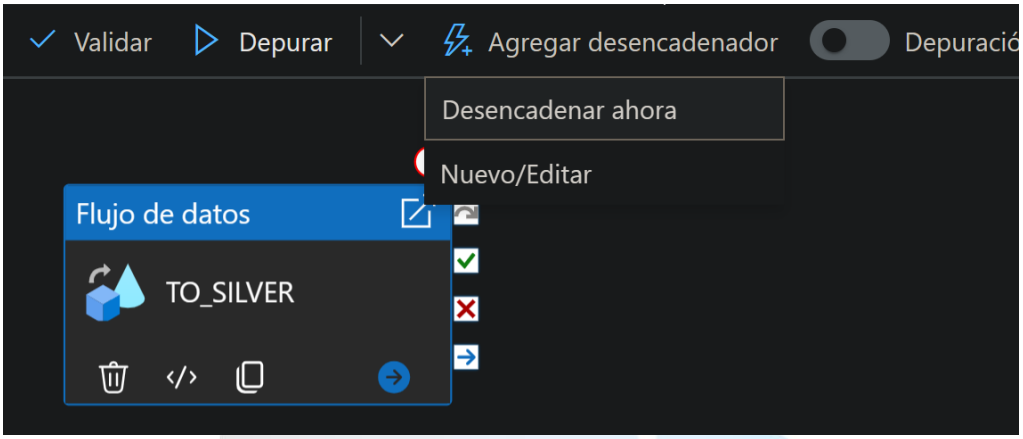
13. Configuramos:

Pestaña	**Opción**	**Valor**
General	Nombre	TO_SILVER
Configuración	Flujo de datos	F_TO_SILVER
Configuración	Tamaño de proceso	Personalizada
Configuración	Tipo de proceso	Básico (de uso general)
Configuración	Recuento de núcleos	4 (+ 4 núcleos de controlador)

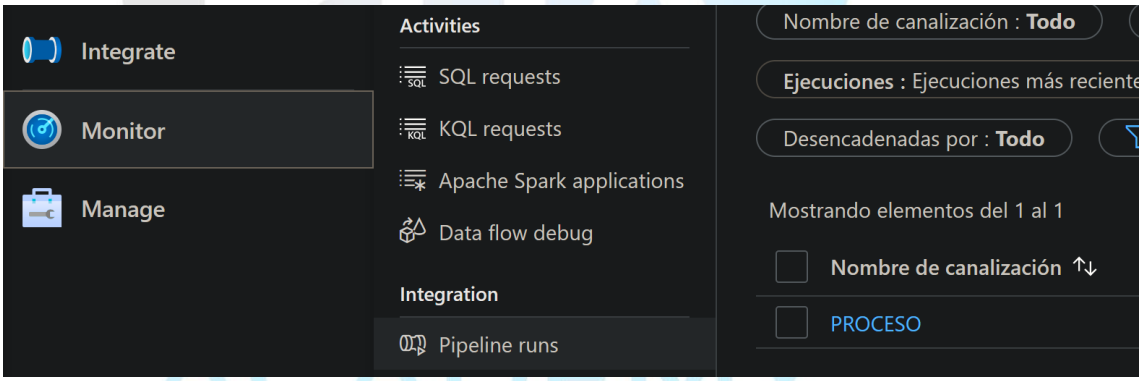
14. Validamos y publicamos los cambios



15. Ejecutamos el proceso seleccionando “Agregar desencadenador / Desencadenar ahora”



16. Para monitorear el proceso damos clic en “Monitor / Pipeline runs” y damos clic sobre “PROCESO”



TIEMPO: 5 MINUTOS