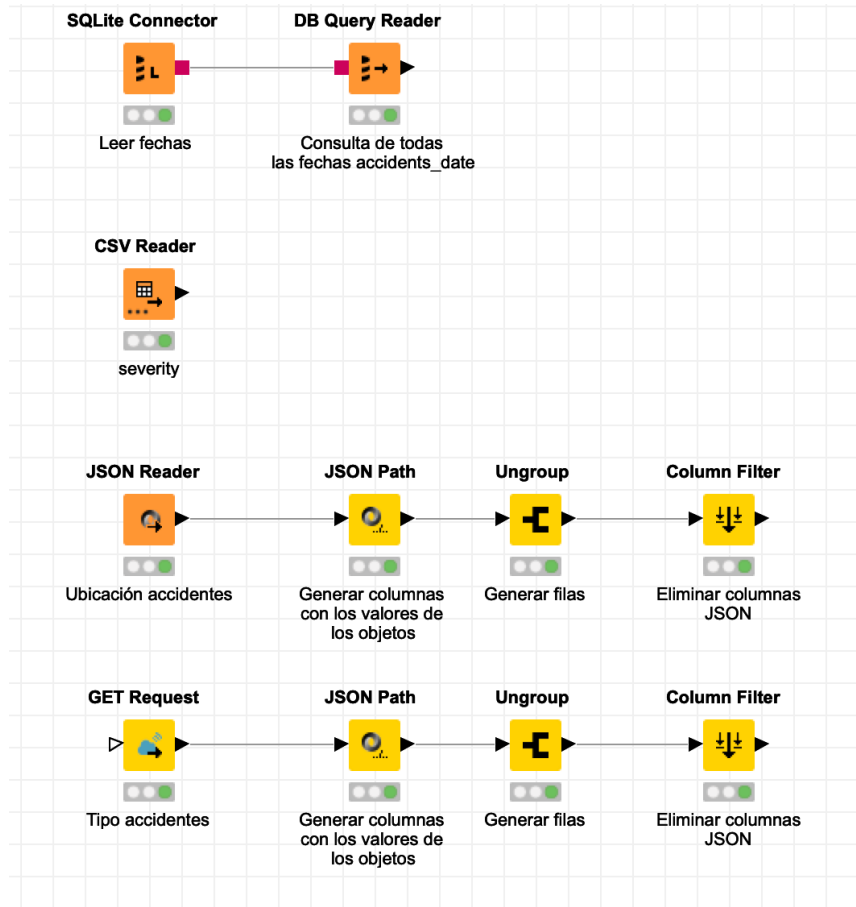


Parcial 2

Laura Juliana Mora Páez

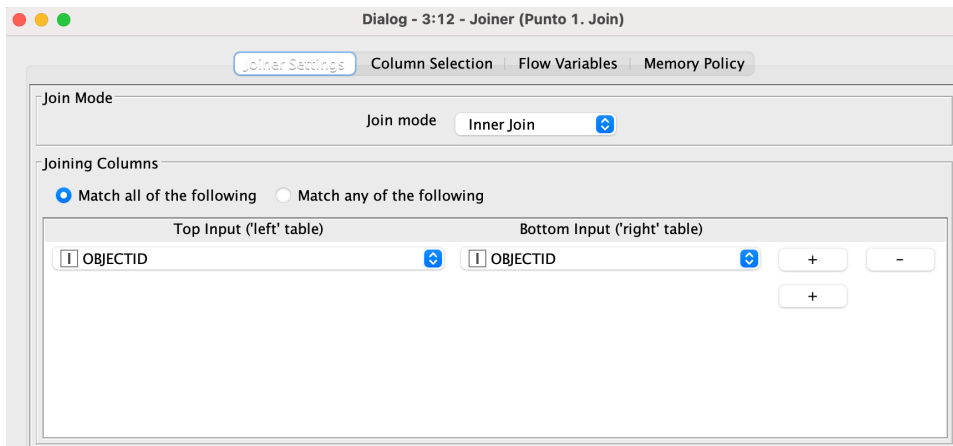
Actividad: Construcción de un flujo de datos (ETL)

Lo primero que se hizo fue leer los datos de las diferentes fuentes y en los casos de los tipo JSON transfórmalos a tablas.

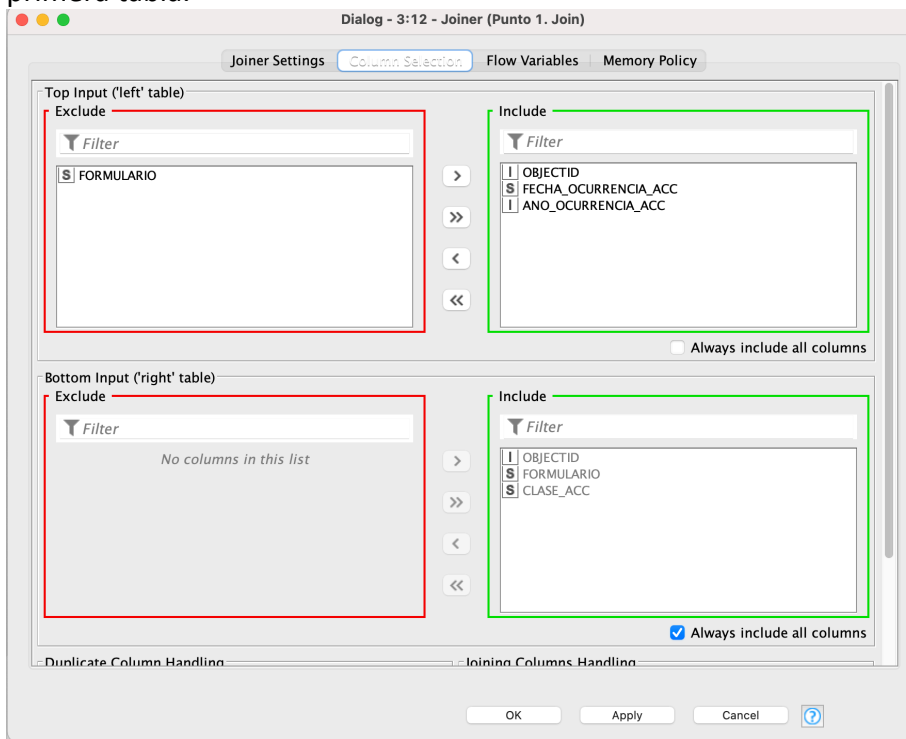


- Realice la unión de las fuentes de fecha y tipo por medio de una llave en común y almacene el resultado en un archivo CSV (el archivo resultante debe estar separado por punto y coma (;))

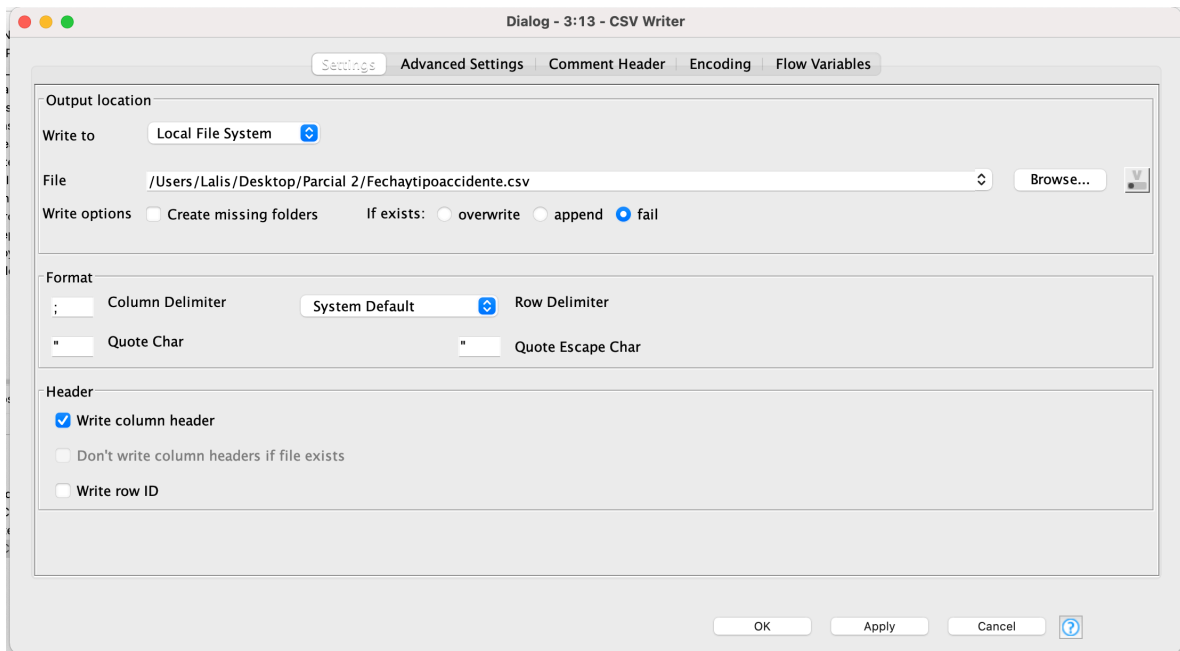
Se unen las tablas según el ID del objeto.



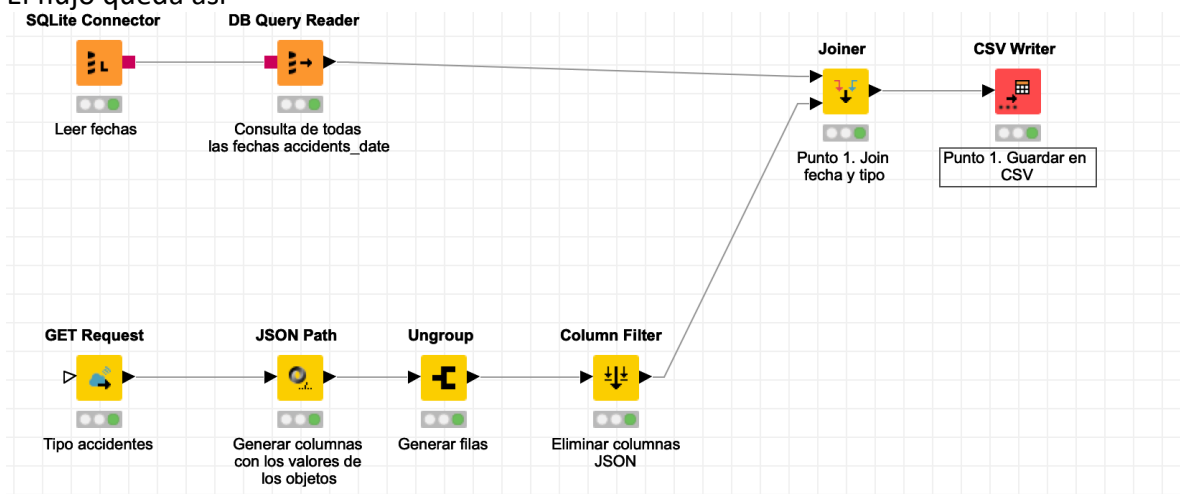
Con el fin de que no queden atributos repetidos no se pasa el atributo FORMULARIO de la primera tabla.



Se guarda la tabla en un csv.

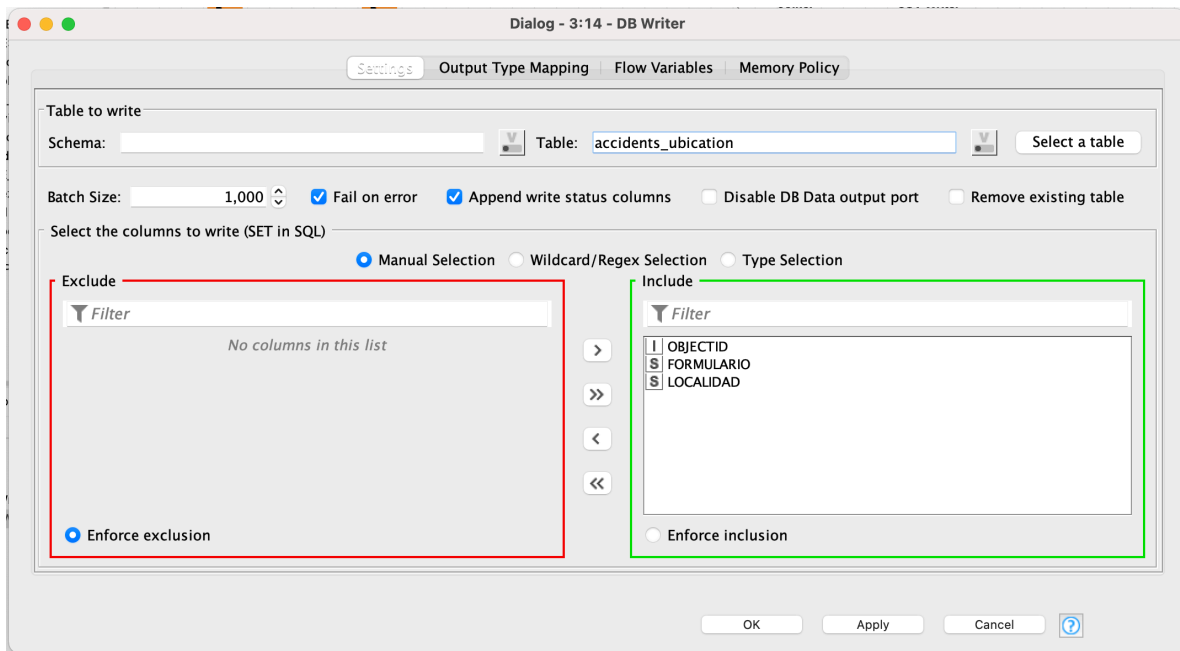


El flujo queda así

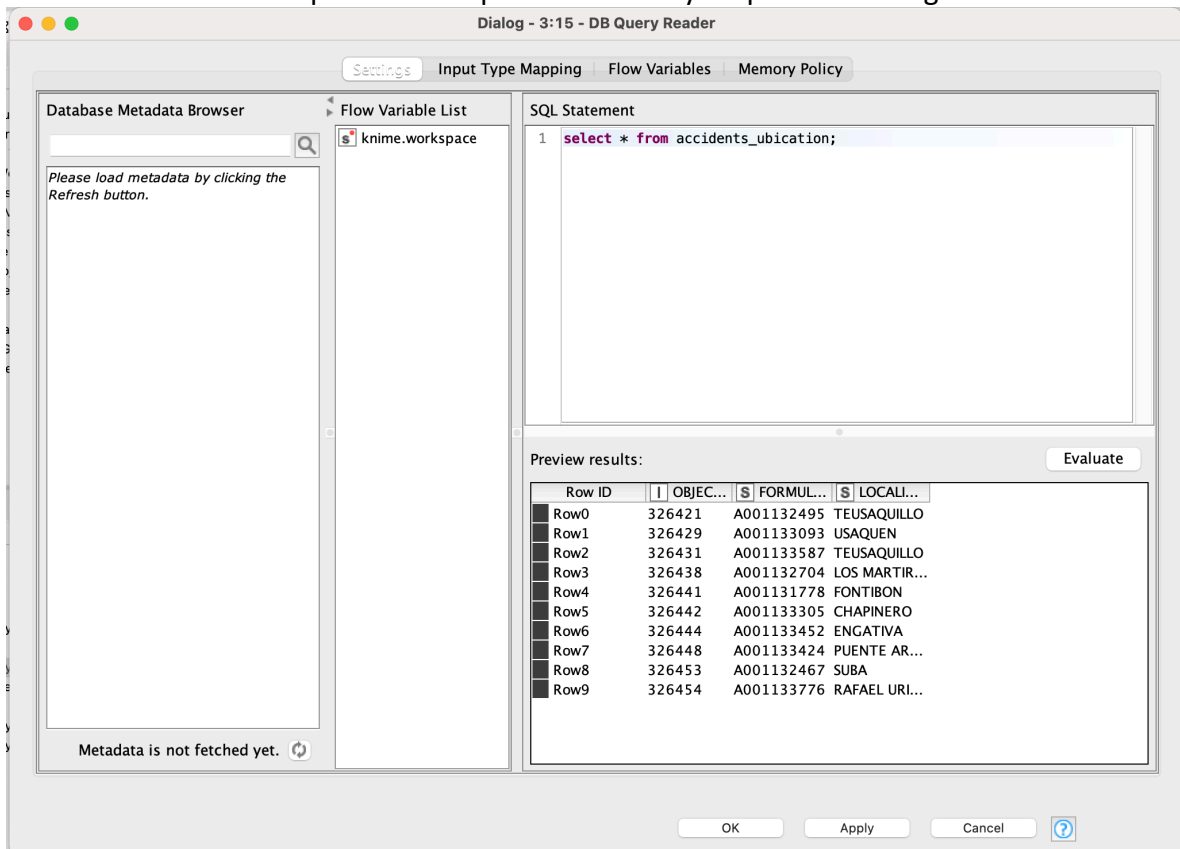


- Inserte el contenido de la fuente de ubicación en la base de datos (la tabla resultante debe tener el nombre accidents_ubication)

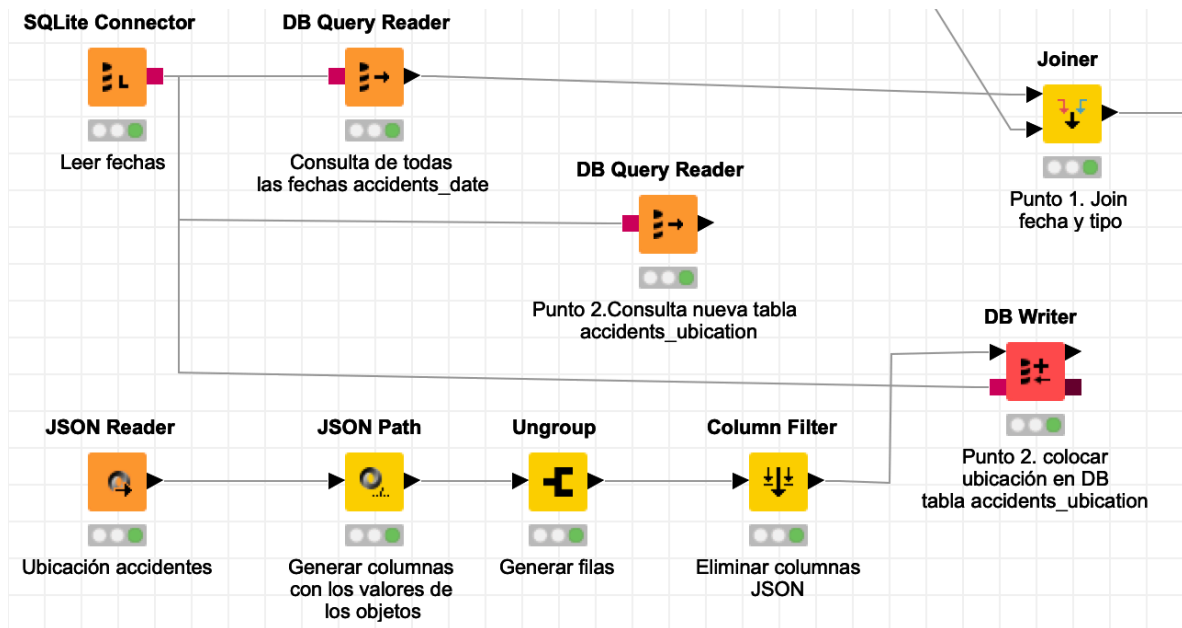
Se pasan todos los atributos a la nueva tabla en la DB, a través del DB writer.



Se realiza una consulta para validar que los datos hayan quedado bien guardados.



El flujo queda así:



- Realice la unión de la fuente ubicación (importante: la fuente debe ser tomada de la base de datos (punto anterior) y no del archivo json) y severidad y almacene el resultado en un archivo XML

Se unen las tablas según el ID del objeto.

Dialog - 3:16 - Joiner

Joiner Settings | Column Selection | Flow Variables | Memory Policy

Join Mode

Join mode: Inner Join

Joining Columns

☒ Match all of the following ☐ Match any of the following

Top Input ('left' table)	Bottom Input ('right' table)		
OBJECTID	OBJECTID	+	-
		+	

Performance Tuning

Maximum number of open files: 200

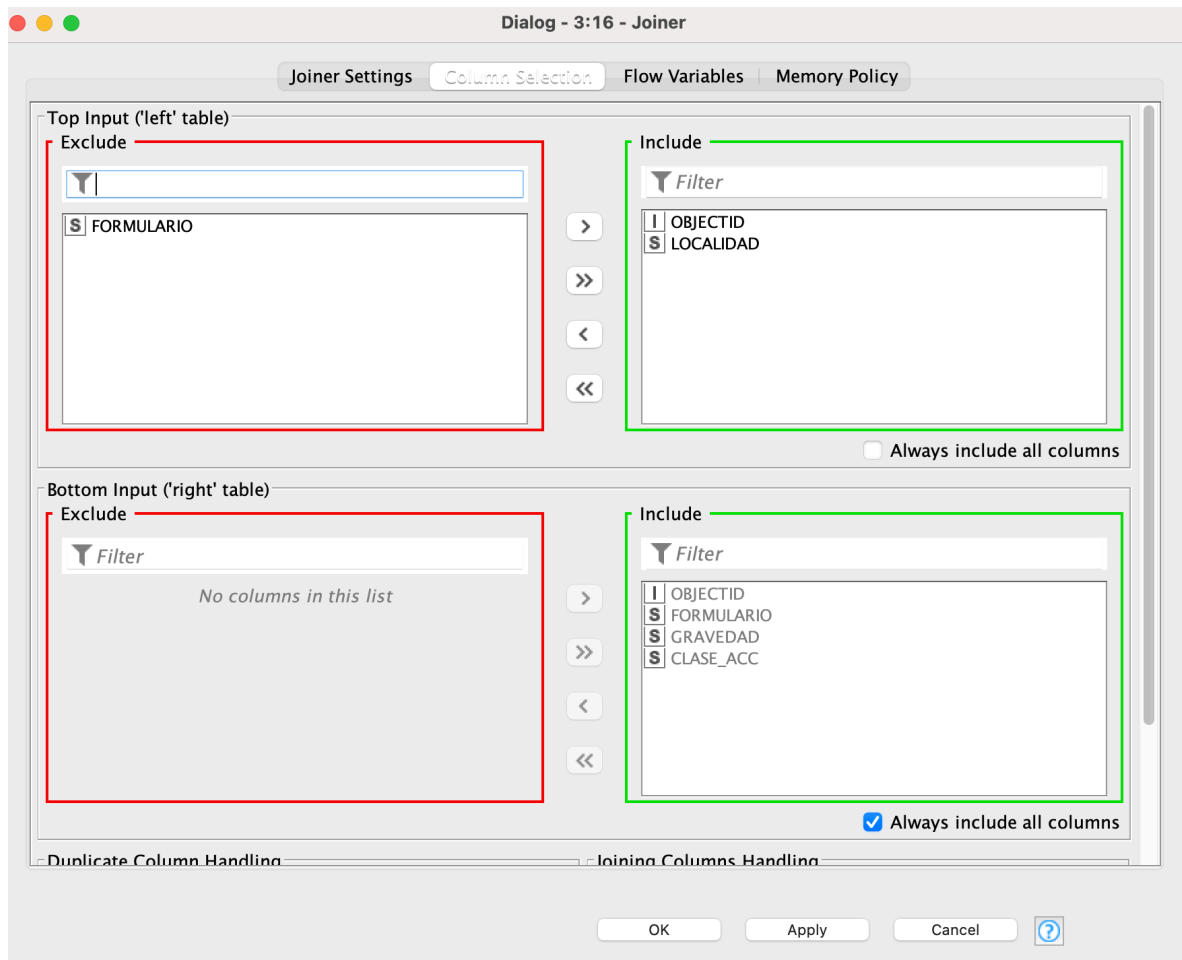
☐ Enable hlliting

Row IDs

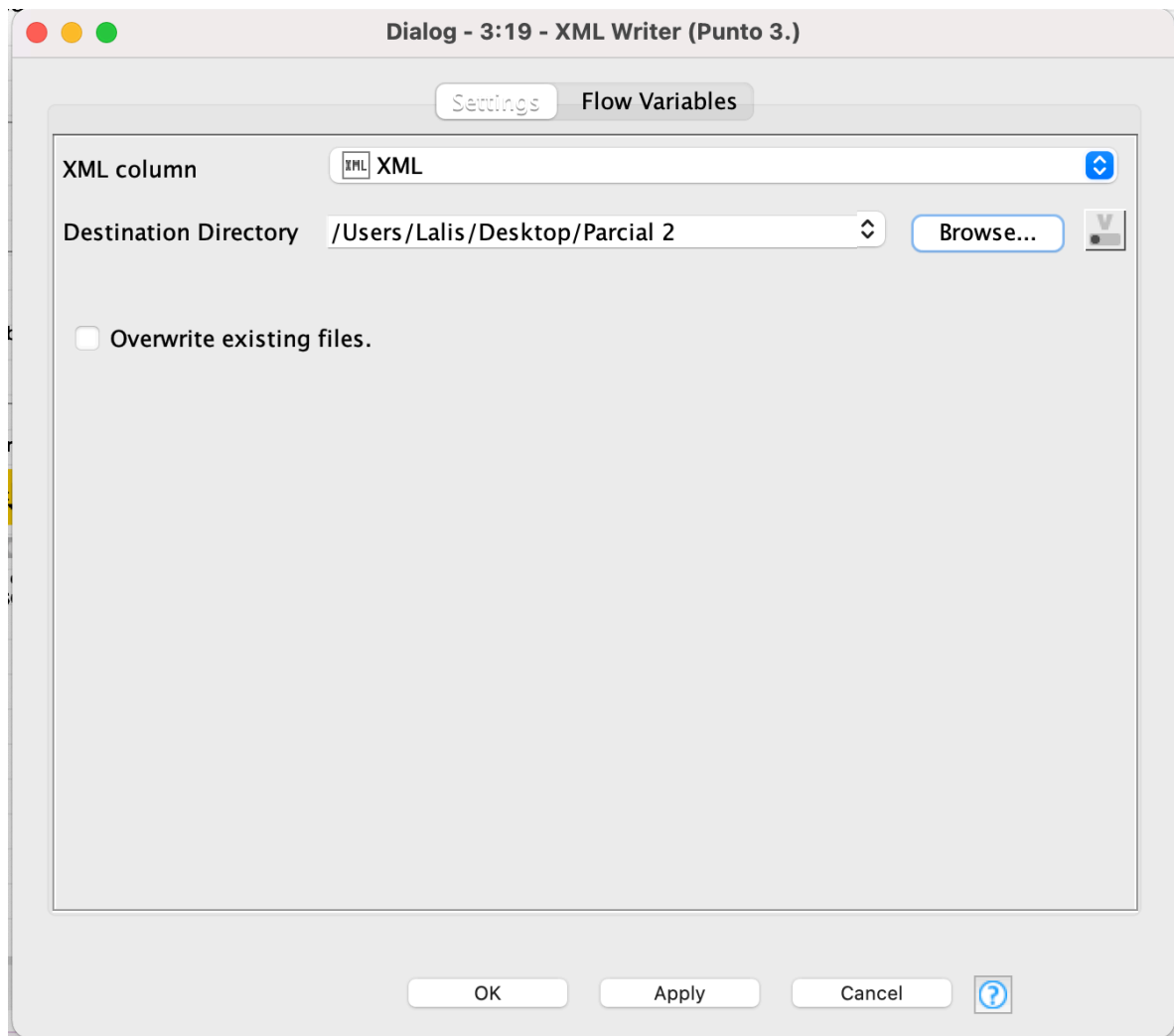
Row ID separator in joined table: _

OK Apply Cancel ?

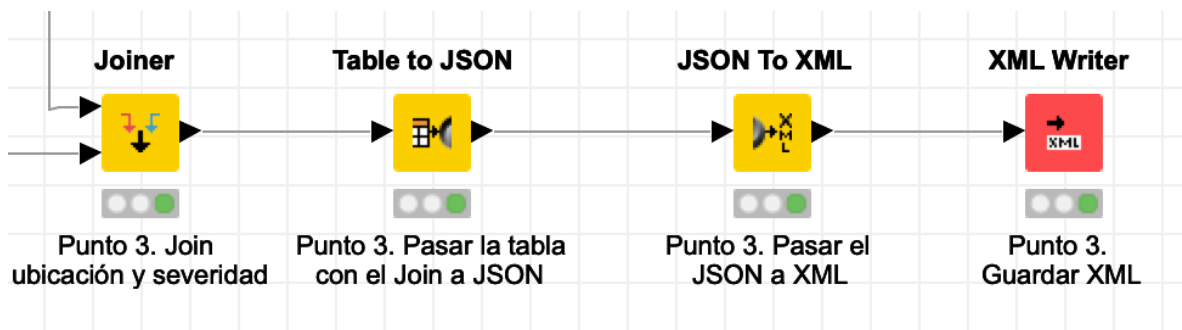
Con el fin de que no queden atributos repetidos no se pasa el atributo FORMULARIO de la primera tabla.



Se crea el archivo XML

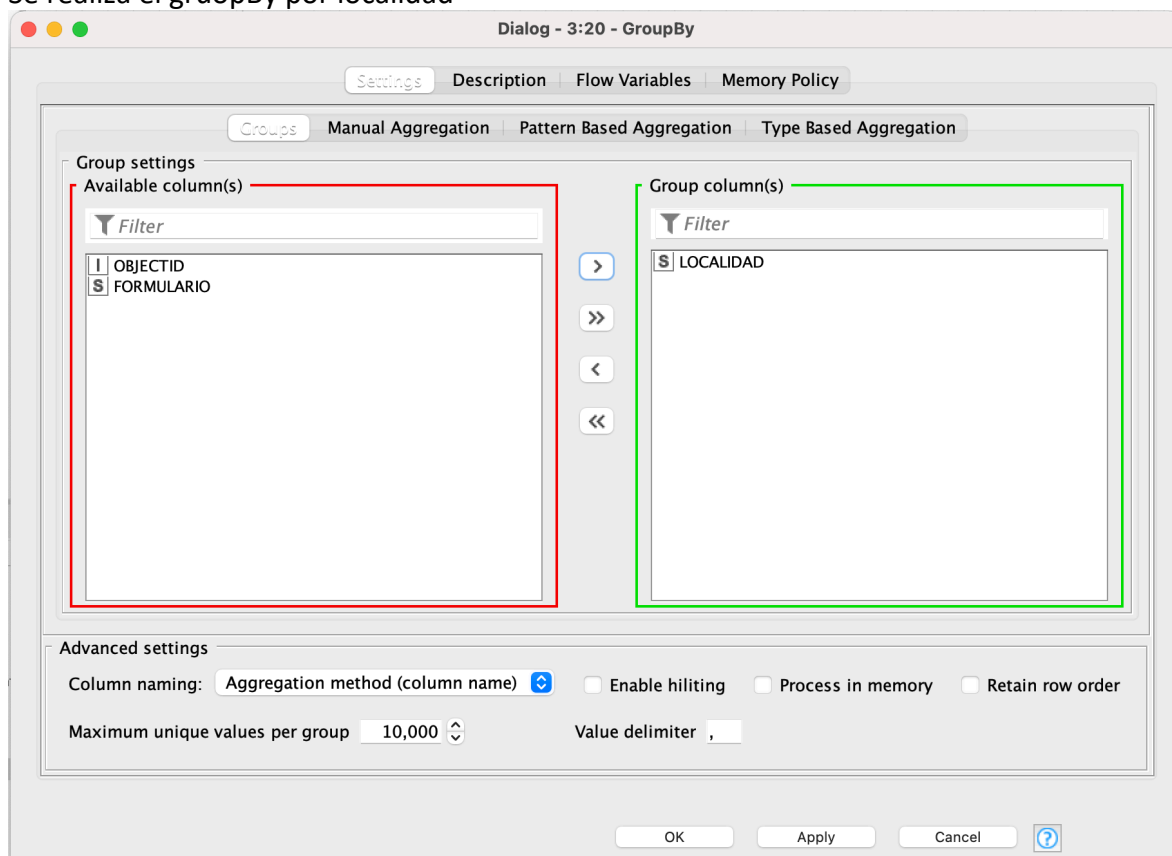


El flujo queda así, con el fin de pasarlo a formato XML primero se pasa a formato JSON y posteriormente a XML

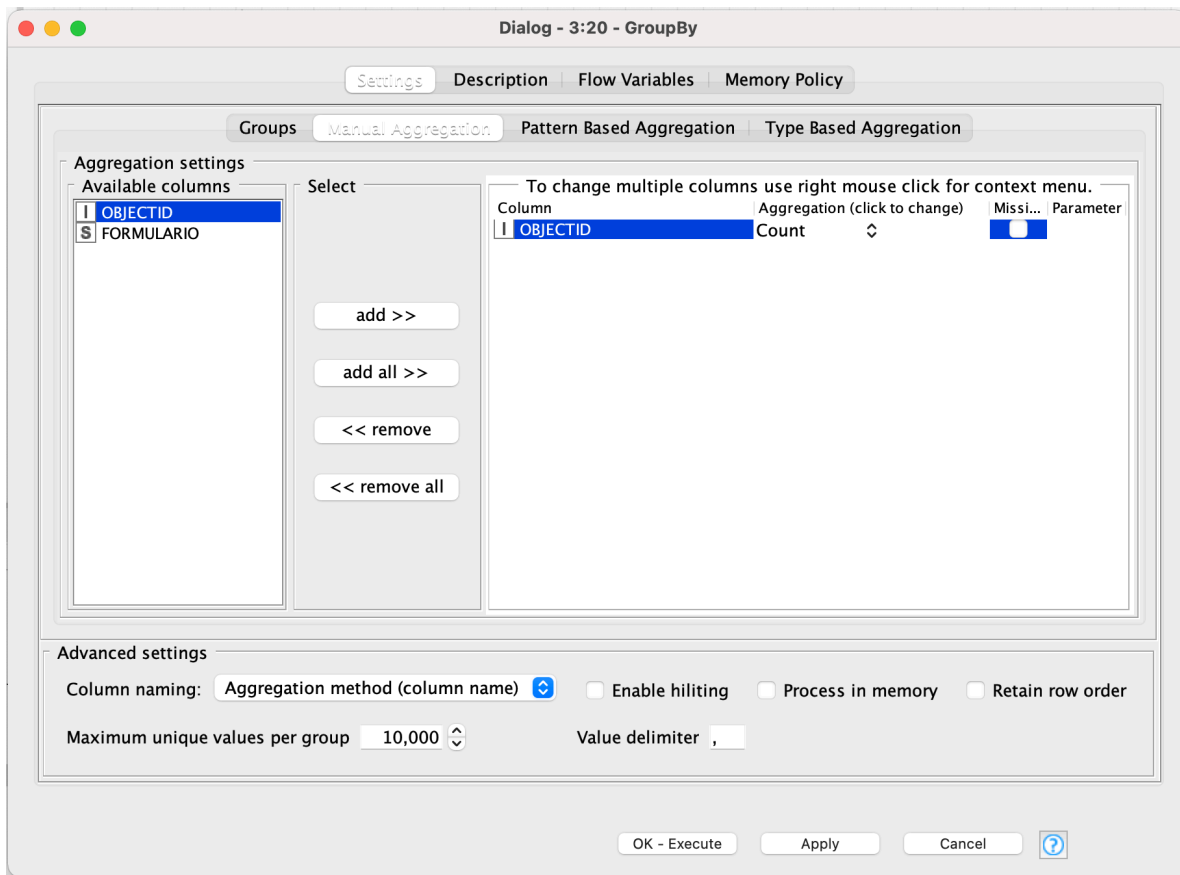


- Realice el cálculo de la cantidad de accidentes ocurridos por localidad (consulte la agregación Count en el módulo GroupBy) y almacene el resultado como un archivo TSV (archivo separado por tabs, puede emplear el módulo CSV Writer para esta tarea)

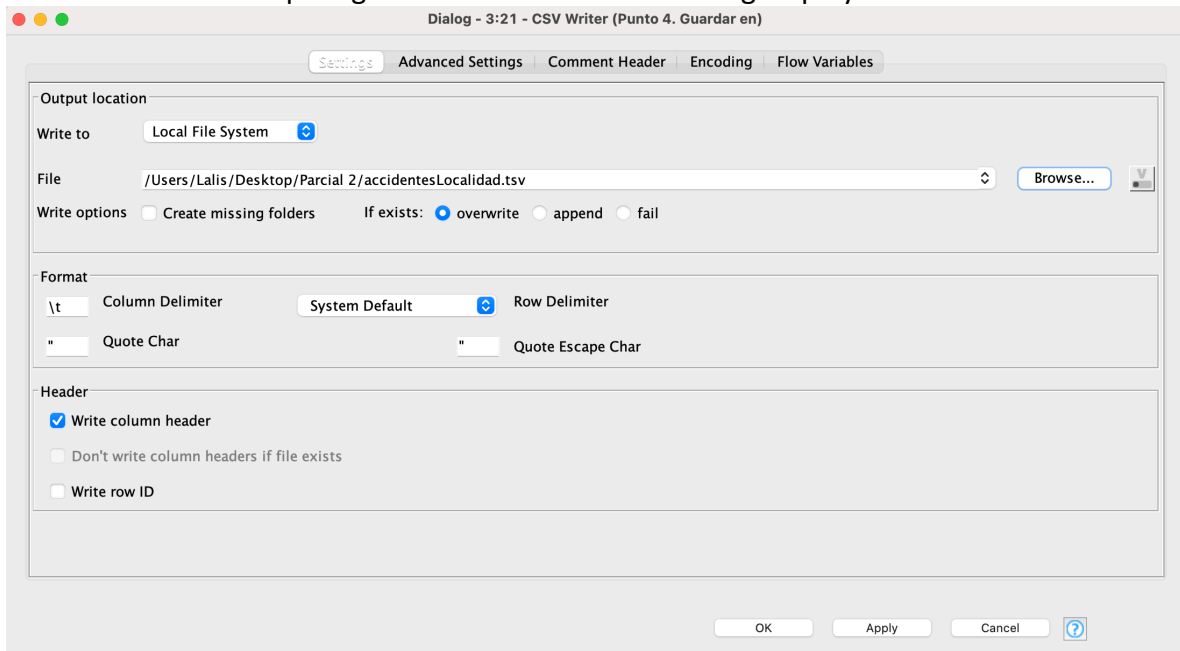
Se realiza el groupBy por localidad



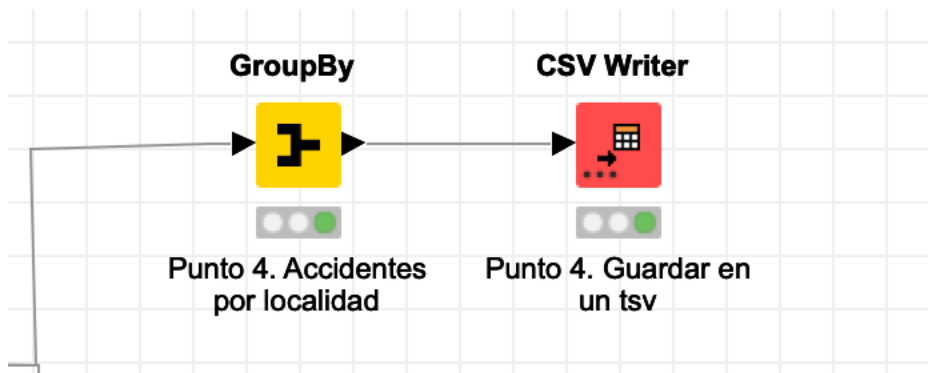
Se cuentan los objetos de cada localidad



Se crea el archivo tsv para guardar la tabla resultado del group by

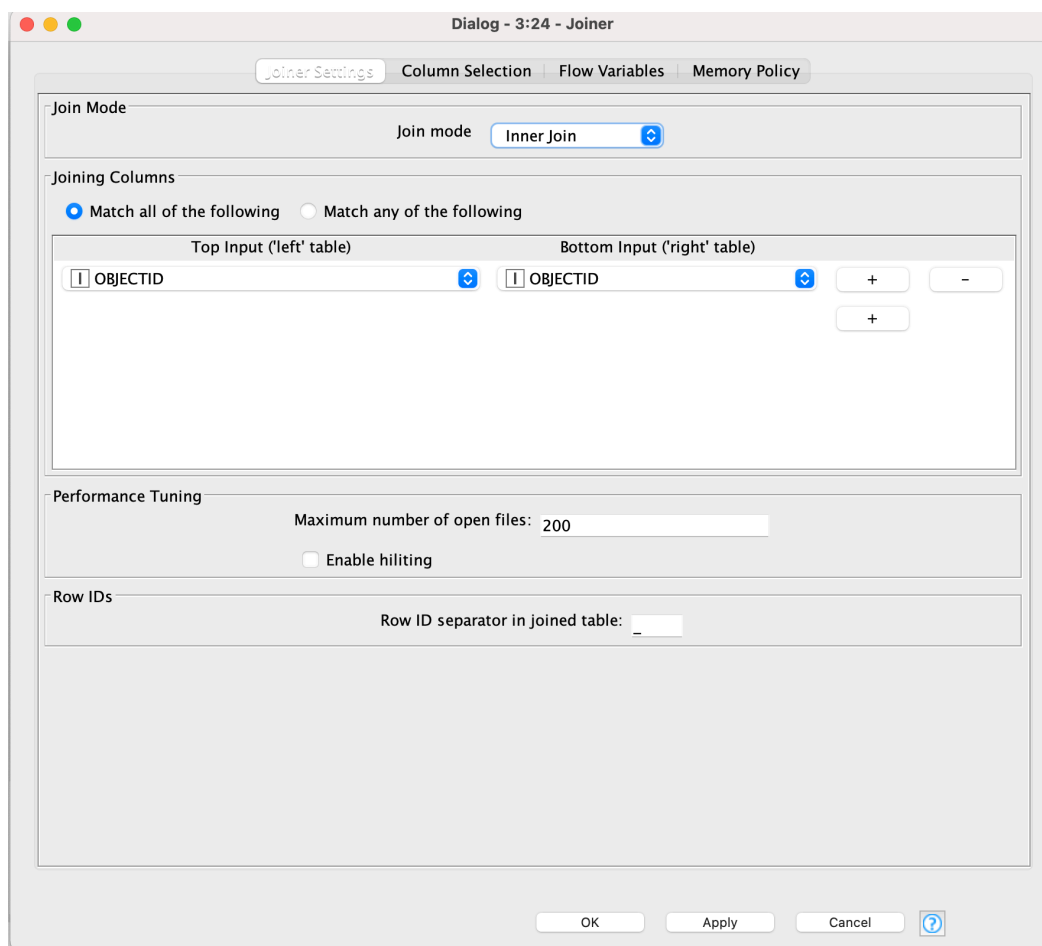


El flujo queda así:

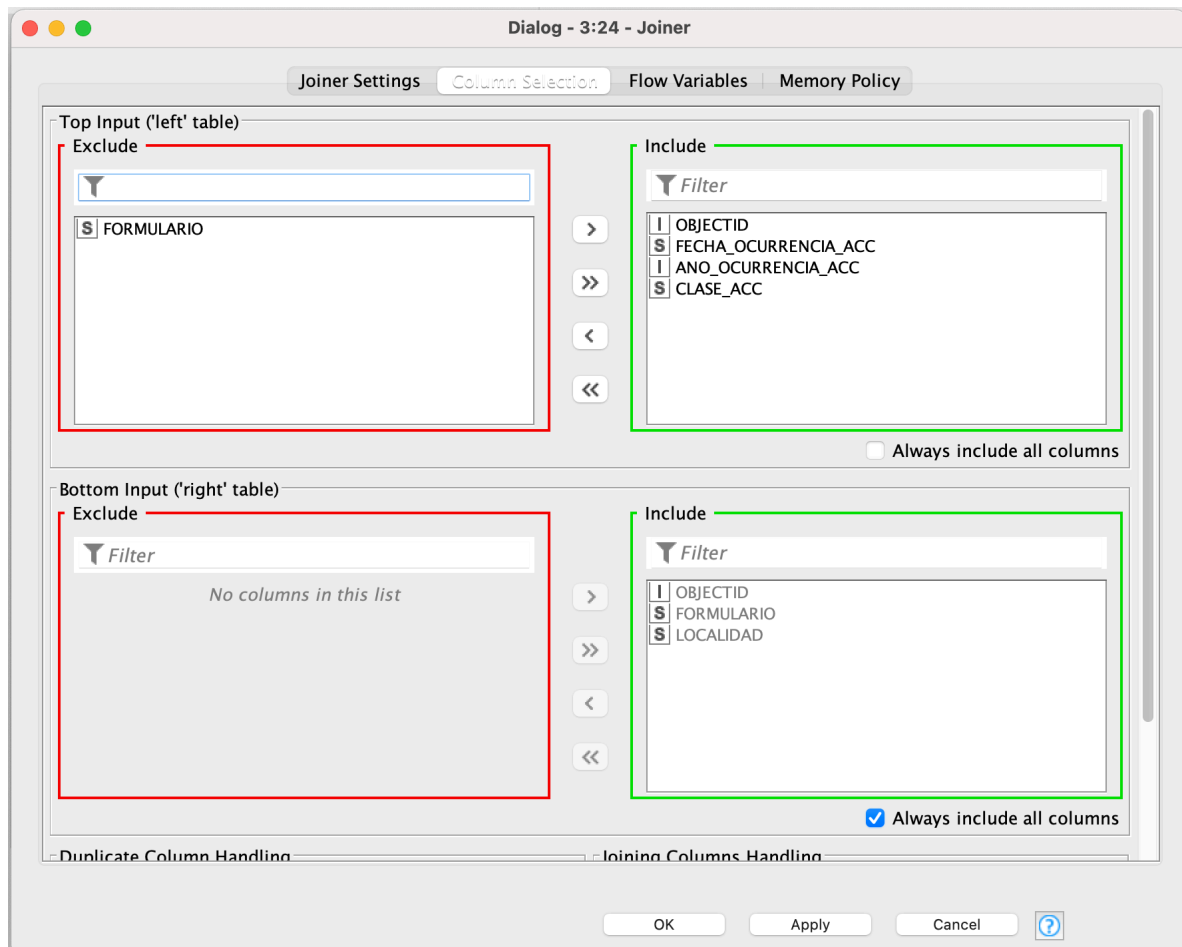


- Realice el cálculo de la cantidad de accidentes ocurridos por tipo (únicamente en la localidad de Suba) y almacene el resultado en un archivo JSON. Pista: para seleccionar únicamente los accidentes ocurridos en Suba puede utilizar una consulta en SQL (SELECT * FROM localidades WHERE LOCALIDAD="SUBA") o el módulo Row Filter

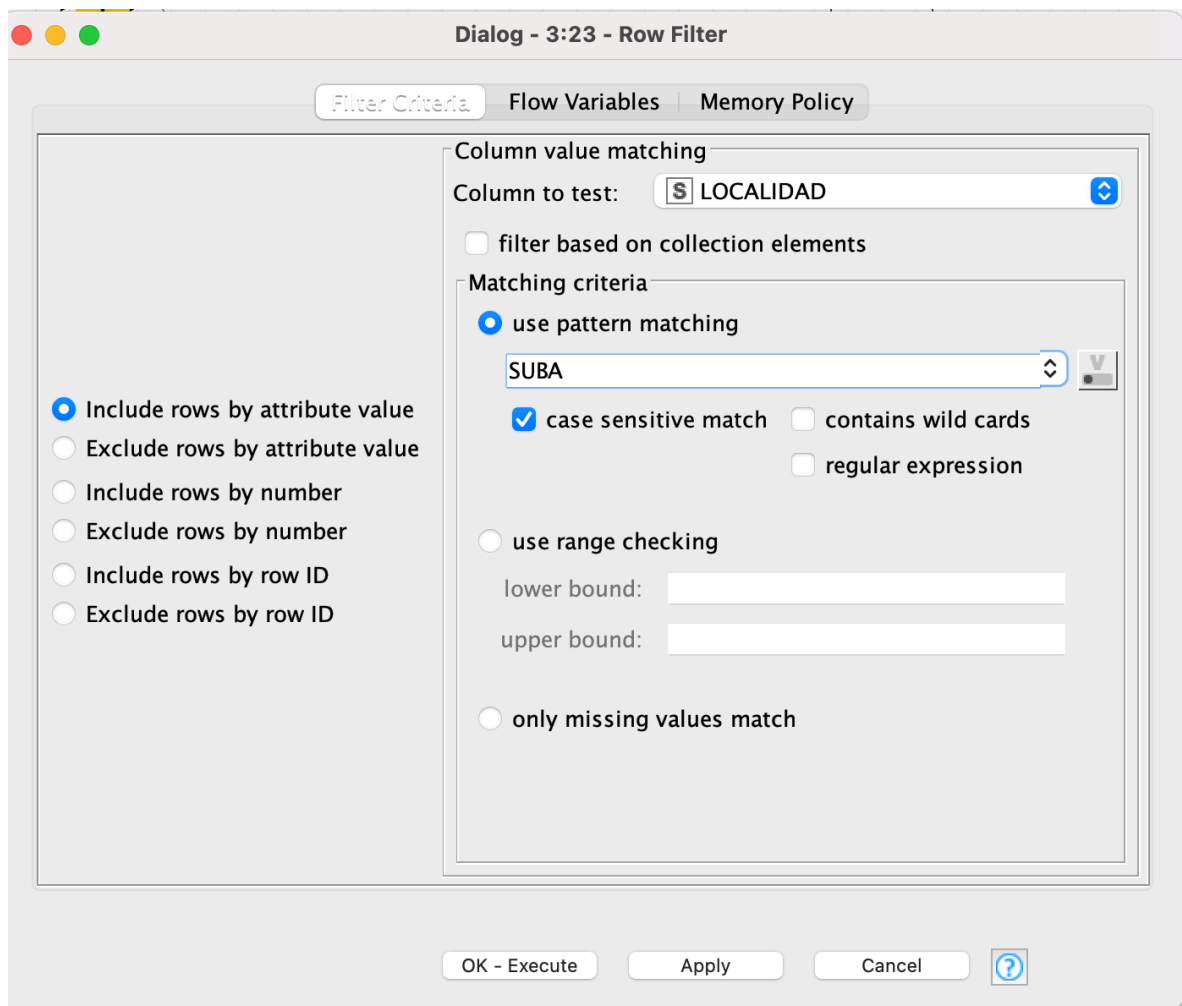
Se realiza un Join entre la tabla con fecha y tipo y la tabla con ubicación.



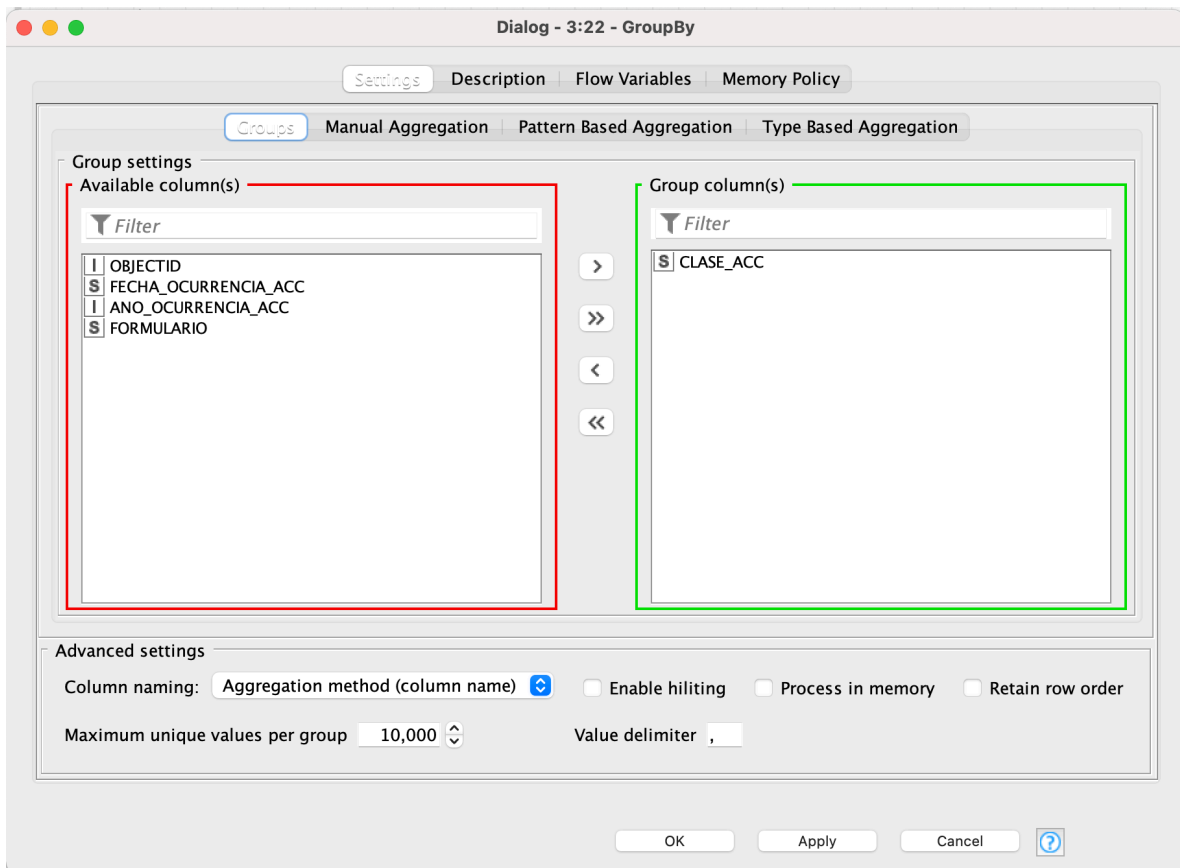
Con el fin de que no queden atributos repetidos no se pasa el atributo formulario de la primera tabla.



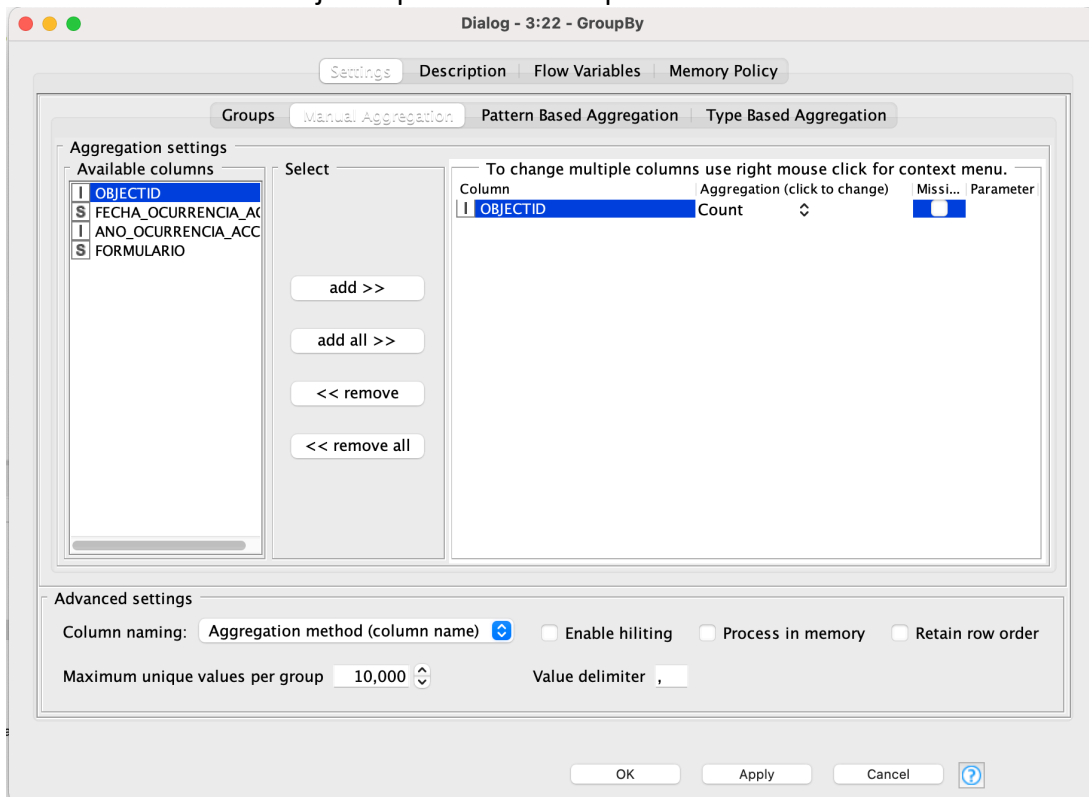
Se realiza un filtro por localidad para que solo se tomen los datos donde la localidad es suba.



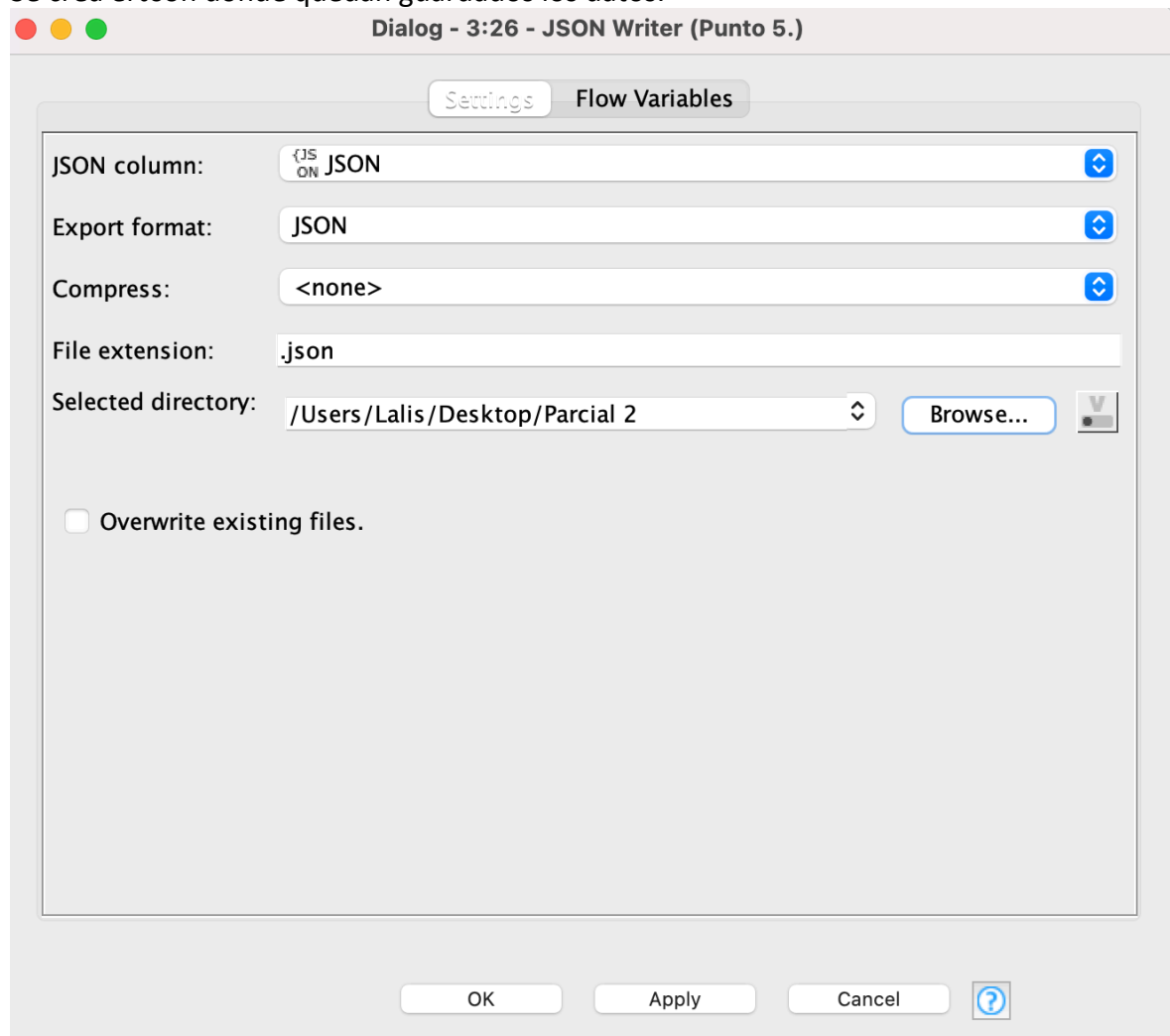
Se agrupan los datos según el tipo



Se cuentan todos los objetos que tienen ese tipo



Se crea el Json donde quedan guardados los datos.



Así queda el filtro, tras realizar el agrupamiento se pasa la tabla a JSON para poder guardarlo



Actividad: Identificación de tipos de datos

Para los siguientes contextos de negocio, indique cuáles podrían ser las (a) fuentes de datos, (b) uso de estas fuentes de datos y (c) almacenamiento de las fuentes de datos para los dominios de (I) metadatos, (II) datos maestros, (III) datos operacionales, (IV) datos no estructurados y (V) datos analíticos:

(1) Hospital

Las fuentes de datos en un hospital dependiendo de los recursos asignados a tecnología pueden estar en una DB centralizada, en un Excel o en un ARP grande como SAP, también se cuentan con sensores médicos, cámaras de seguridad, entre otros. El uso de estos datos depende de la funcionalidad que se requiera realizar por ejemplo la remisión a un especialista involucra el paso del paciente a nivel de nombre, identificación y diagnóstico o en el caso para la aplicación de la vacuna se requiere el nombre del paciente, su fecha de nacimiento y las comorbilidades que padece.

Entre los **metadatos** que se pueden manejar en el hospital se encuentran la fecha de modificación de las fuentes de datos, entre los **datos maestros** se destacan los pacientes (con su nombre, fecha de nacimiento, etc...), los médicos con los que cuenta el hospital, eps, proveedores de insumos. Para el caso de **datos operativos** se tienen todas las facturas generadas a los pacientes por diferentes tratamientos realizados, las facturas de los proveedores, las historias clínicas, las proyecciones financieras del hospital; por el lado de los **datos no estructurados** se tienen las radiografías, exámenes médicos, videos de seguridad del hospital. Y finalmente los **datos analíticos** interpretación de los exámenes médicos y radiografías, conceptos de especialistas.

(2) Universidad

Las fuentes de datos en una universidad dependiendo de los recursos asignados a tecnología pueden estar en una DB centralizada, en un Excel, en un ARP grande como SAP, cámaras de seguridad, sensores y diferentes sistemas de información. Estos datos se pueden manejar de diferentes formas como por ejemplo para saber quien entra y quien sale de las instalaciones, control de las calificaciones de los estudiantes, saber las asignaturas que cursan los estudiantes, etc..

Entre los **metadatos** que se pueden manejar en una universidad están la fecha de modificación de las fuentes de datos, la fecha en la que se obtuvieron los diferentes datos. Entre los **datos maestros** se destacan los estudiantes, profesores, asignaturas, facultades, carreras, aspirantes, entre otros; para los **datos operativos** se encuentran los recibos de matrícula, parciales, talleres, proyectos, investigaciones, entre otros. Para los **datos no estructurados** videos de cámaras de seguridad, datos captados a través de sensores como los de las entradas a los edificios donde se coloca el carnet para entrar. Finalmente los **datos analíticos** se tienen como ejemplo los resultados de consejería académica, estadísticas académicas de cada estudiante, tendencias académicas.