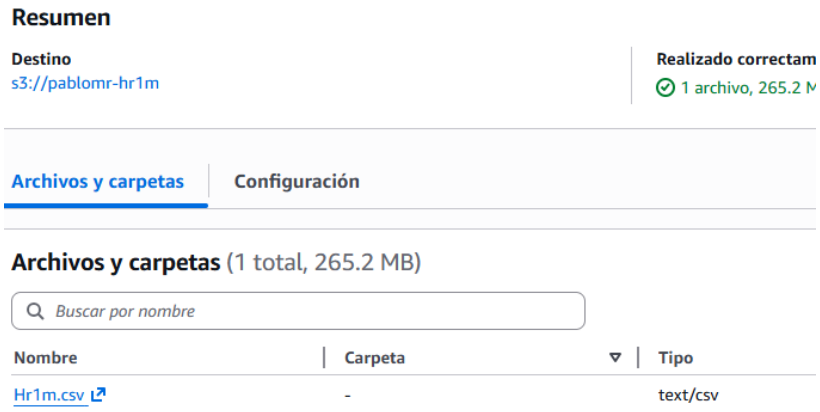
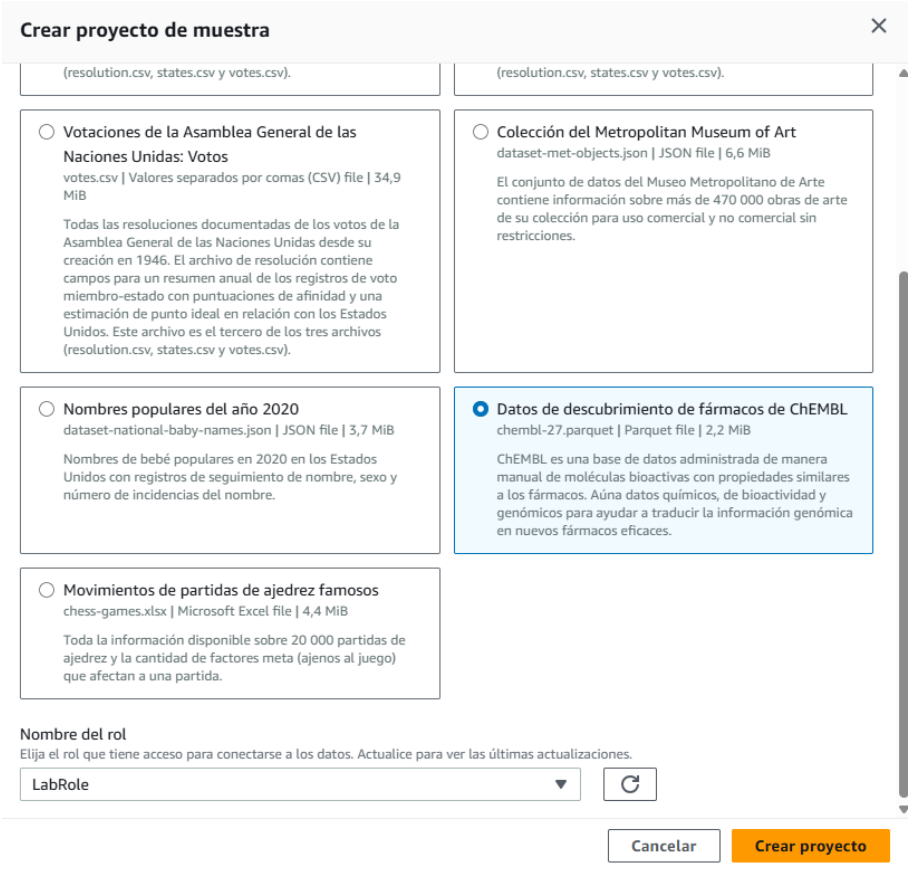
DataBrew I

1) **En este caso, para utilizar un conjunto de datos más voluminoso y con una casuística más amplia, nos centraremos en un dataset de descubrimiento de fármacos de ChEMBL.**



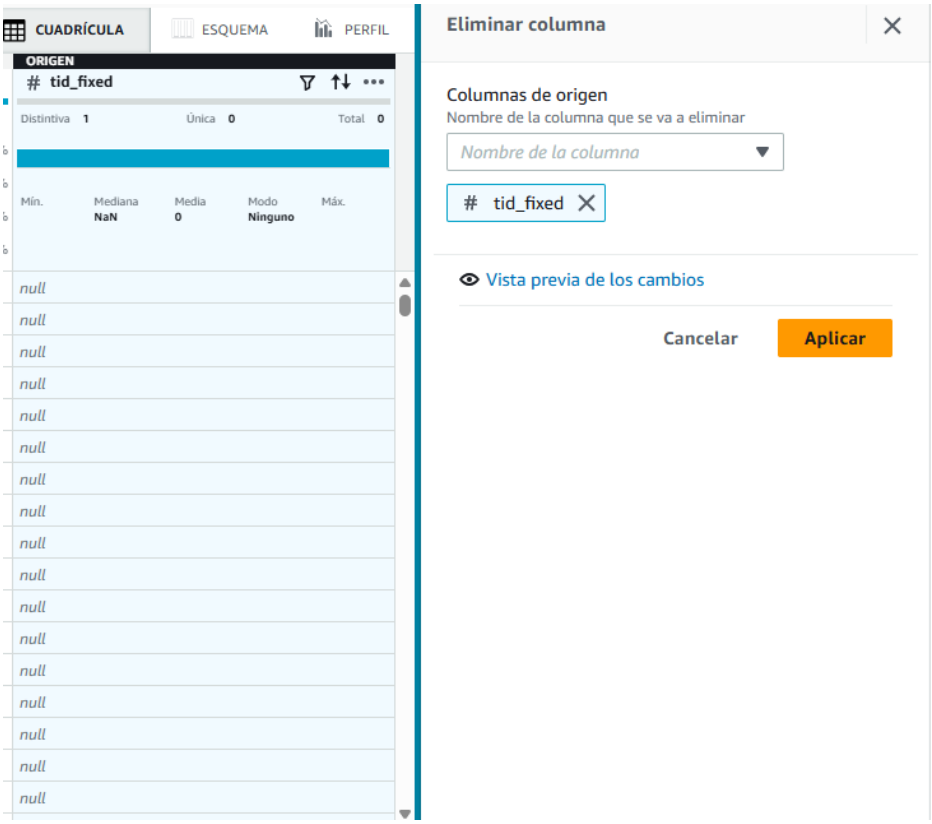
1) El primer paso es entrar a Glue DataBrew y Crear el proyecto de muestra con los datos de CHEMBL, utilizando el LabRole de AWS Academy



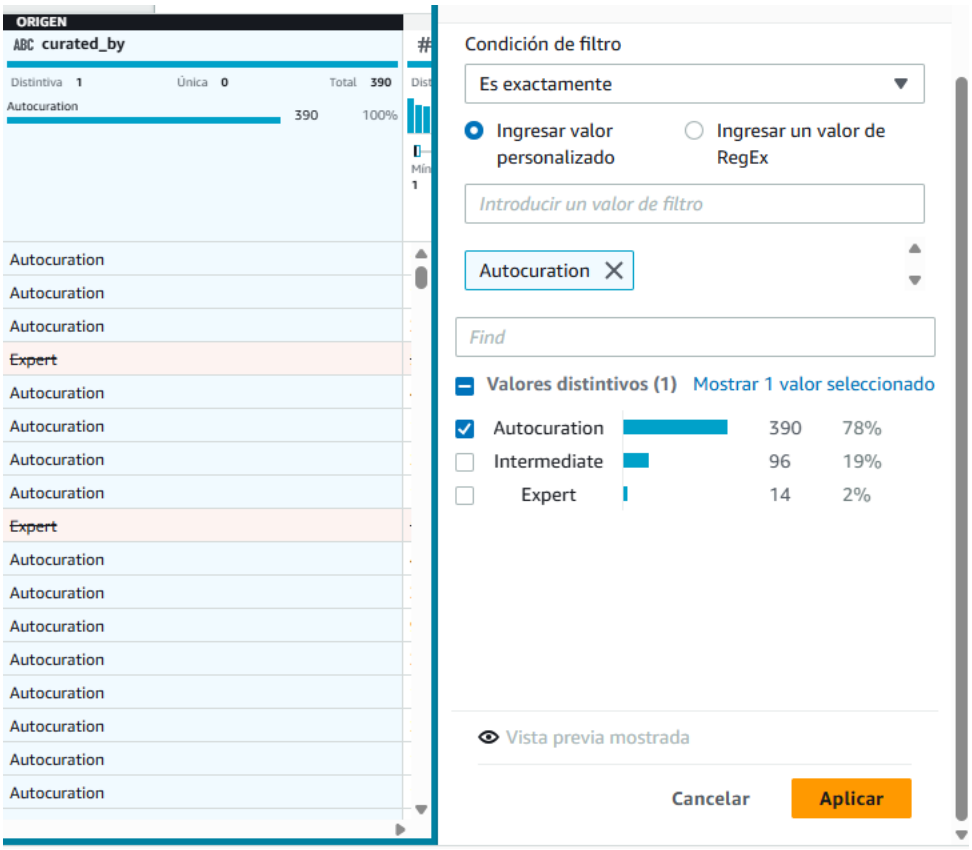
2) **Creando la receta**

Una vez tenemos el entorno listo, vamos a realizar un conjunto de transformaciones que añadiremos a nuestra receta:

1) El primer paso será eliminar la última columna, tid_fixed que tiene todos los valores nulos. Para ello, bien desde el menú Columna, seleccionamos la opción de Eliminar.

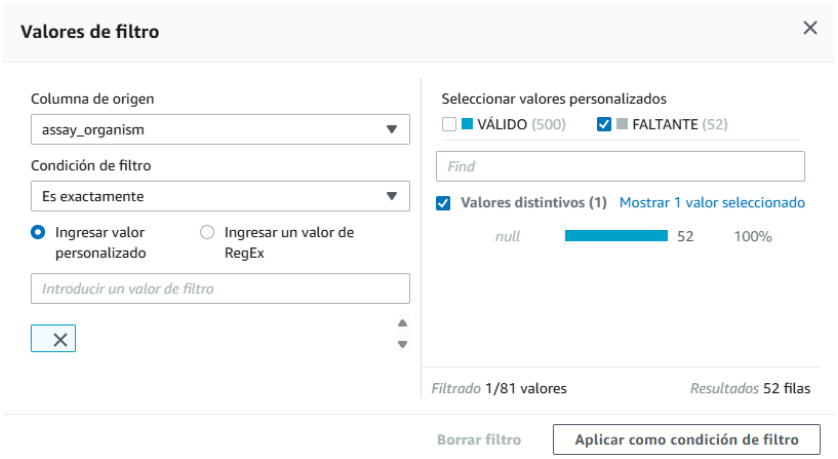


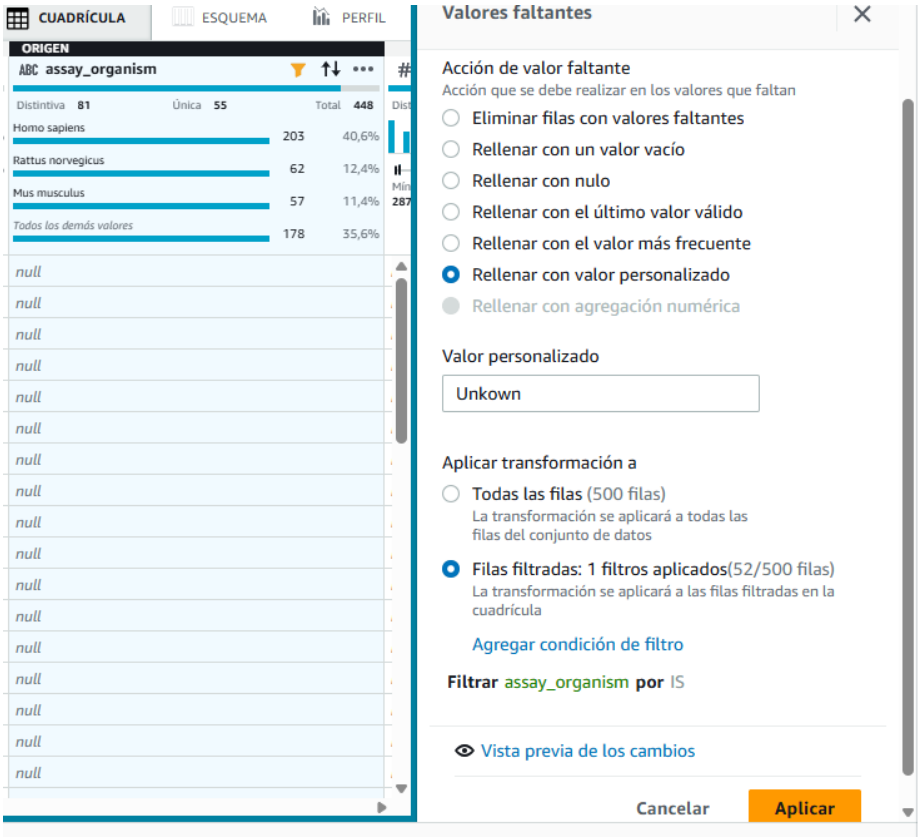
2) Tras aplicar los cambios, en la zona de la receta, aparecerá el paso aplicado. A continuación, vamos a filtrar datos. Por ejemplo, seleccionamos la columna curated_by y seleccionamos para que sea exactamente Autocuration. En la parte derecha podremos ver una pequeña estadística de los valore existentes y si pulsamos sobre Vista previa, se marcarán en rojo las filas que se eliminarán.



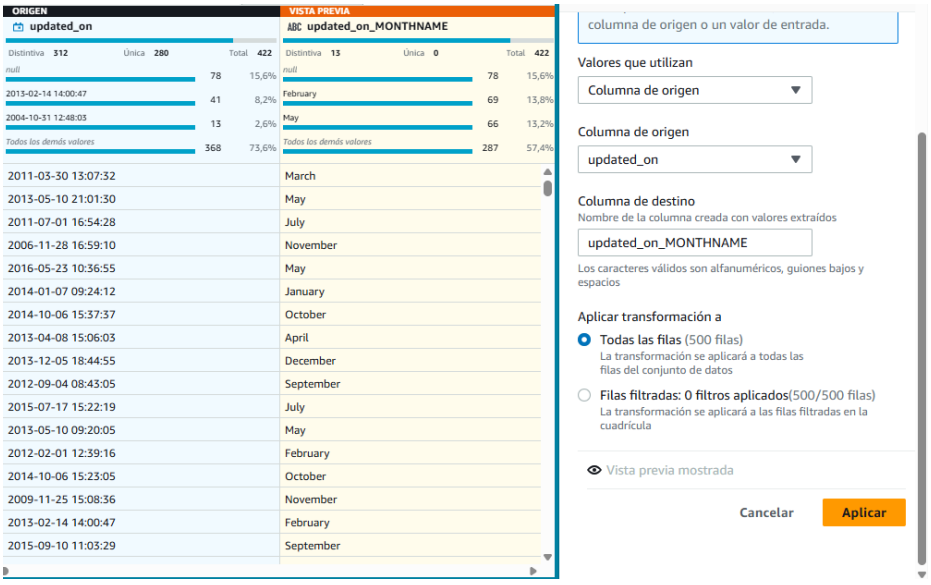
3) Ahora nos vamos a centrar en la gestión de los valores nulos. Para ello, en la columna assay_organism cambiaremos los nulos por Unknown, utilizando el menú Faltante y la opción de Rellenar con valor personalizado:

Filtro:

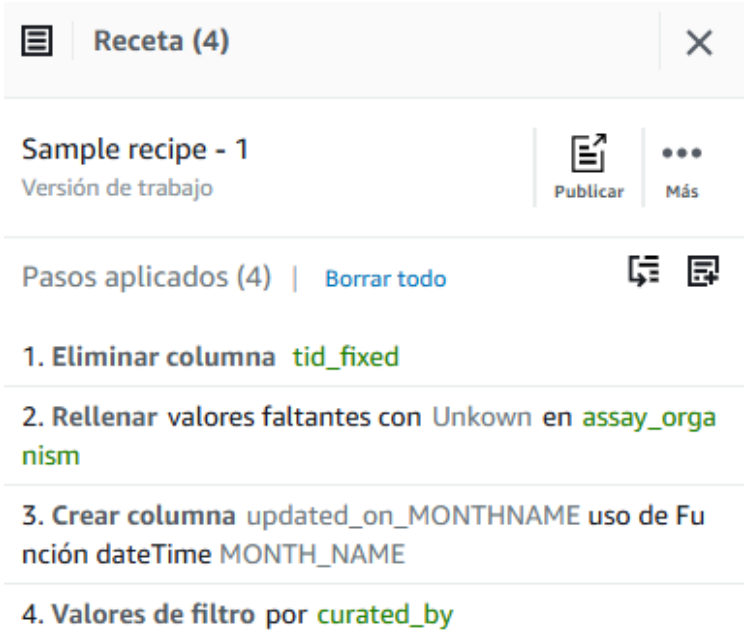




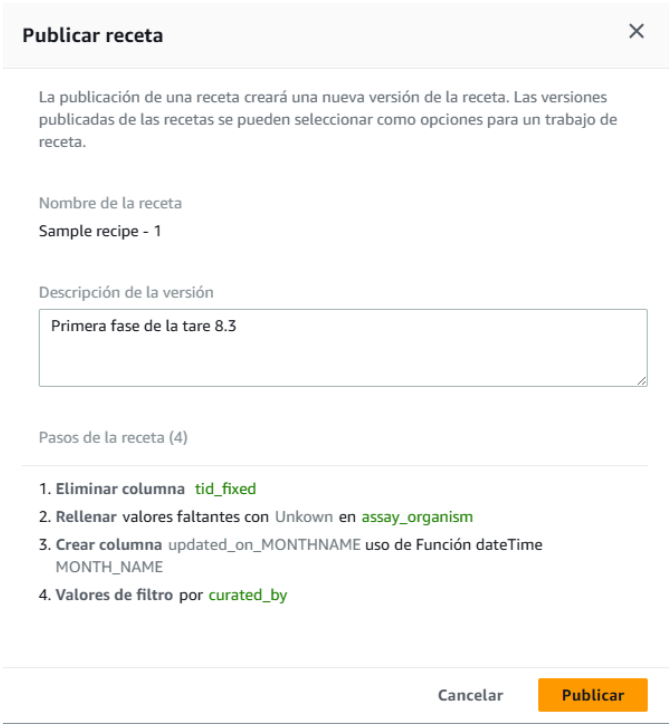
4) Si trabajamos con fechas es muy común crear columnas nuevas con información más útil. En nuestro caso, vamos a añadir una columna que llamaremos Mes con el nombre del mes que conseguimos con la función MONTHNAME sobre la columna updated_on. Para ello, desde el menú Funciones seleccionamos la función de fecha que nos interesa y configuramos los valores



5) Una vez ya tenemos nuestra receta completa con todos los pasos necesarios en nuestra transformación, llega el momento de publicarla para crear una versión de esta y posteriormente poder reutilizarla.



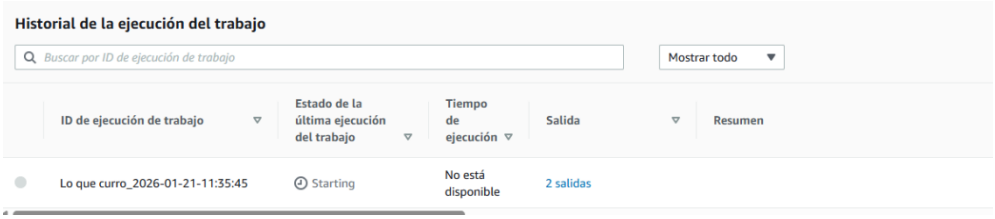
La publicamos



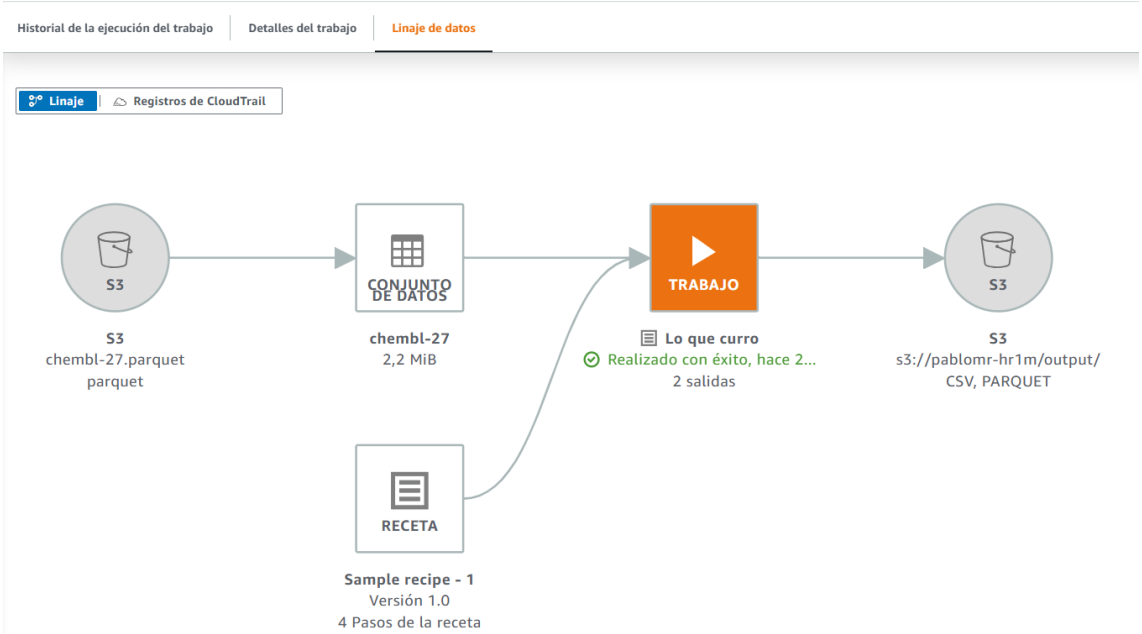
3) De la receta al Job.

1) Si vamos al menú de las recetas, seleccionamos la receta recién publicada, en nuestro caso Sample recipe-2, y creamos un trabajo (job) con

la misma, en el cual, tras darle un nombre y seleccionar el dataset, vamos a guardar el resultado en S3 tanto en formato CSV como en formato Parquet particionado por la columna Mes, y finalmente seleccionamos el rol LabRole



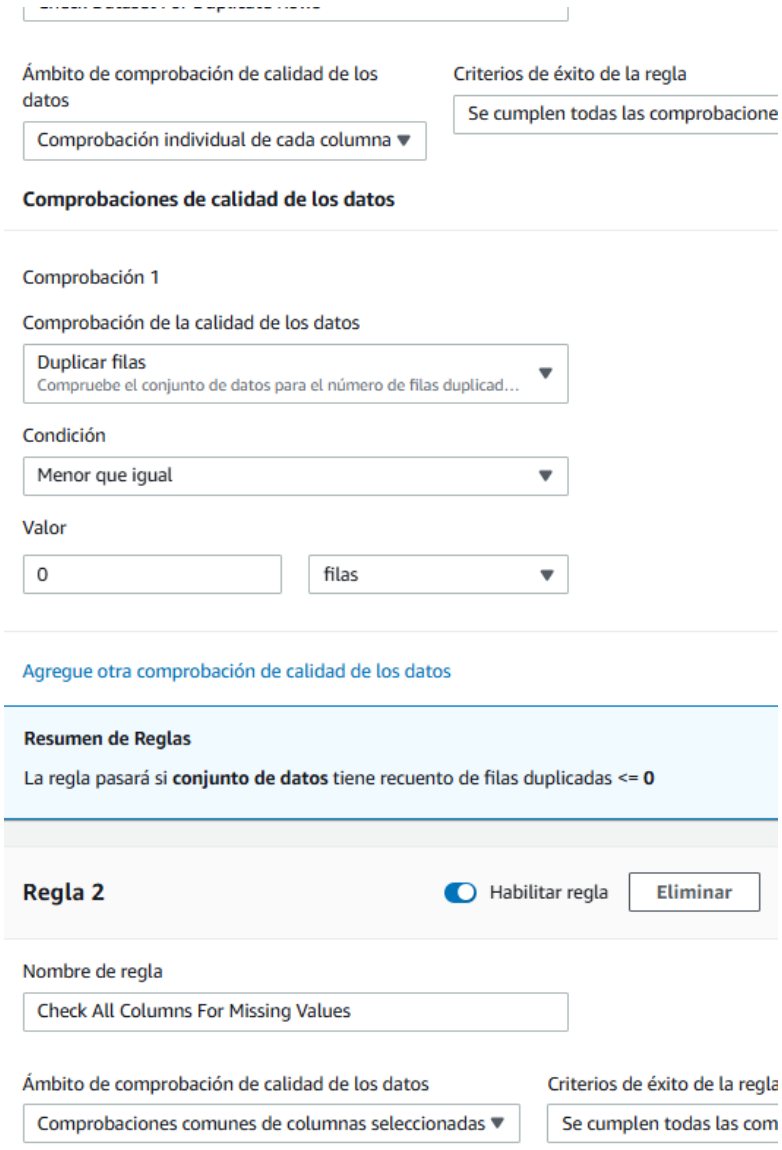
En linaje.



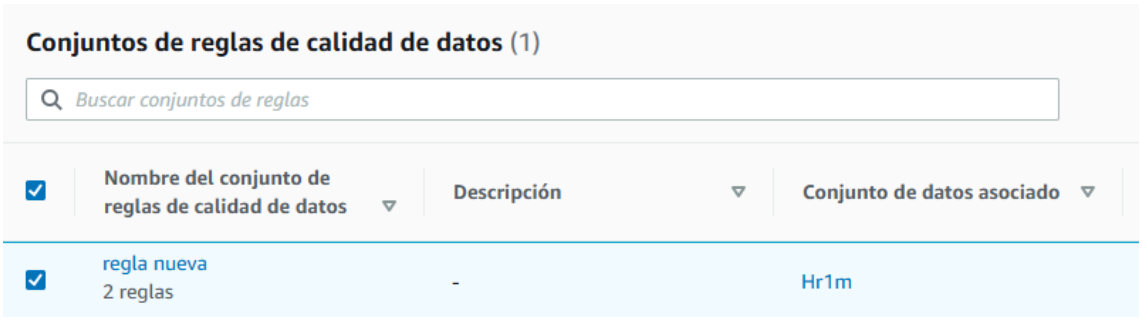
4) Calidad de datos con AWS Glue DataBrew

1) Para crear reglas de calidad de datos, siga los pasos que se mencionan a continuación: Haga clic en la opción DQ Rules. Proporcione un nombre para su conjunto de reglas de calidad de datos. Por ejemplo, puede llamarlo calidad de datos-recurso humano. En la sección Elegir conjunto de datos, seleccione el conjunto de datos “Hr1m.csv”. Una vez seleccionado el conjunto de datos, el sistema le ofrecerá recomendaciones para las reglas de calidad de datos.

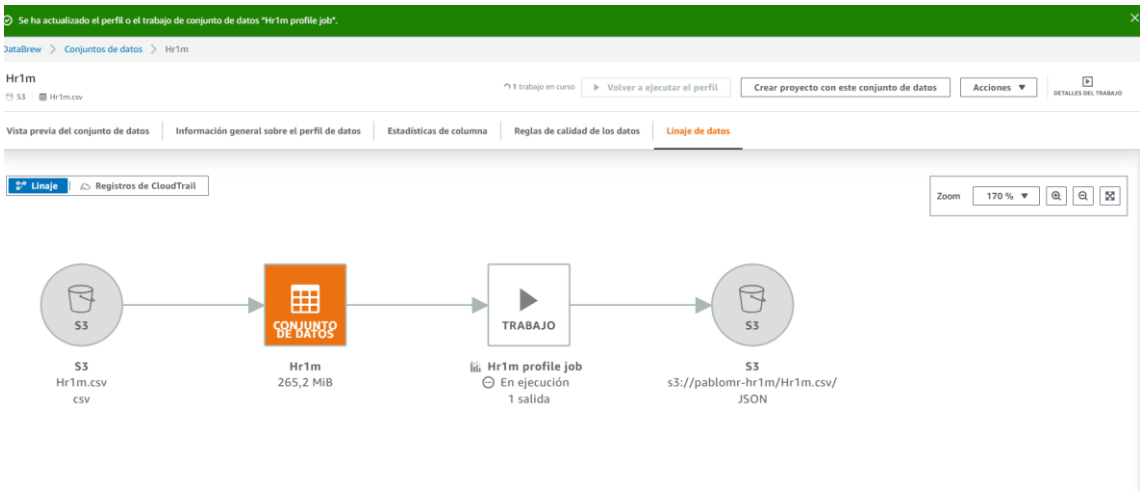
   Seleccionamos las reglas recomendadas por el sistema de Amazon Web Service.



Regla perfectamente creada



Creamos un trabajo y lo ejecutamos



Los pasó todos, están muy bien hechos