Databrew II

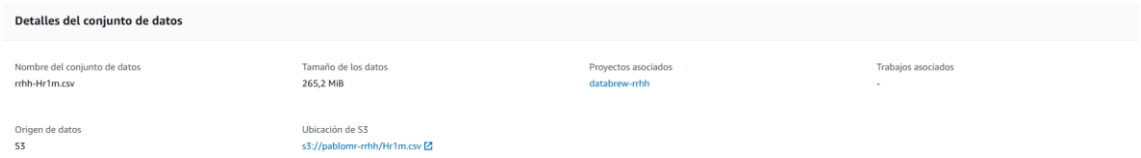
1) Apartado A

1) Carga el CSV anterior en una carpeta llamada raw dentro de un bucket nombrado con algo similar a rrhh.

2) Desde Databrew crea una conexión de datos a dicha carpeta.

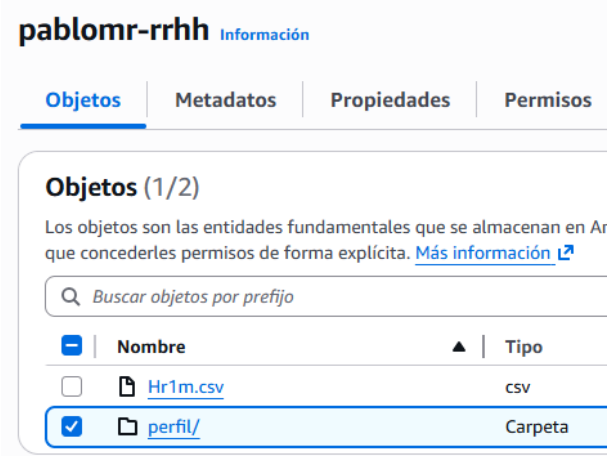
   a) ¿Cuánto ocupa el archivo?

   En ocupa 265.2 MB



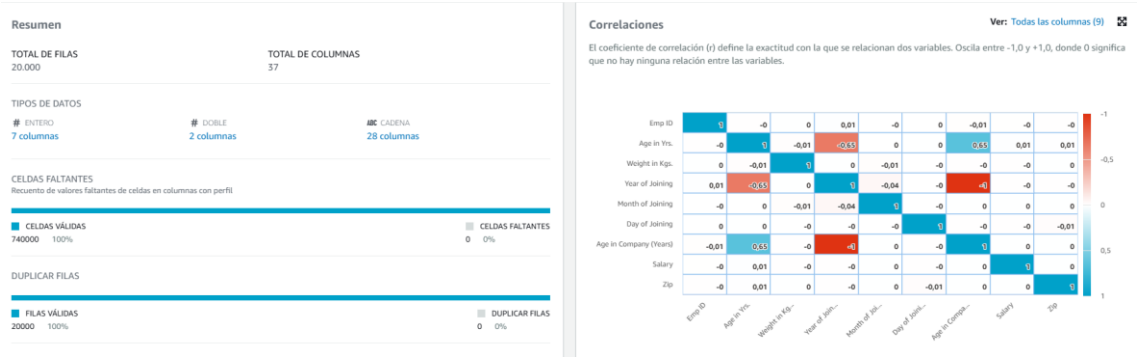
   b) Haz una captura que nos muestre el tipo y contenido de las 5 primeras filas de algunas de las columnas.

3) Crea una carpeta dentro del bucket anterior llamada perfil.

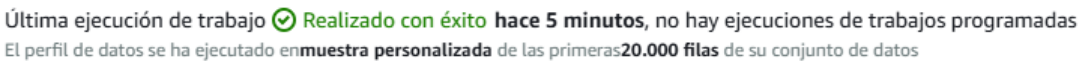


4) Genera el perfil de datos de dicho conjunto de datos. Deja la configuración por defecto. ¿Cuántas filas utiliza por defecto para el análisis?

Perfil de datos:



Se utilizaron 20.000 filas.



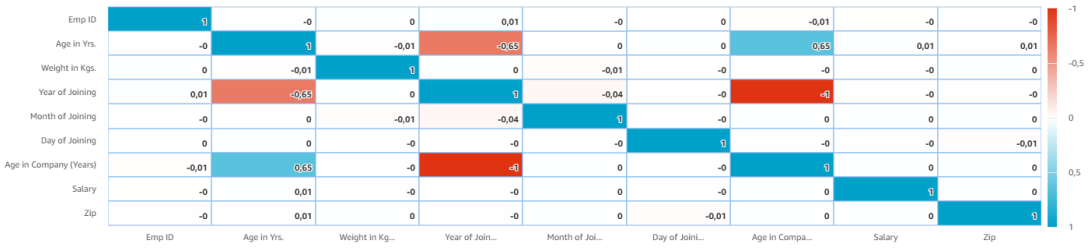
5) Analiza los datos obtenidos.

   a) *¿Cuántas columnas y de que tipo tiene el conjunto de datos?*

Son 37 columnas, 28 son de tipo cadena de carácter y 9 de números enteros.

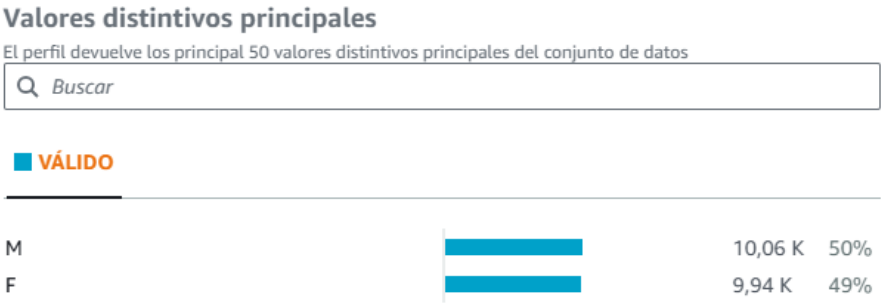
   b) *¿Hay alguna correlación positiva o negativa que te llame la atención?*

En cuanto a correlación positiva, destacan “Age in Years” y “Age in company”. Por otro lado, en correlación negativa destacan “Years of Joining” y “Age in company”



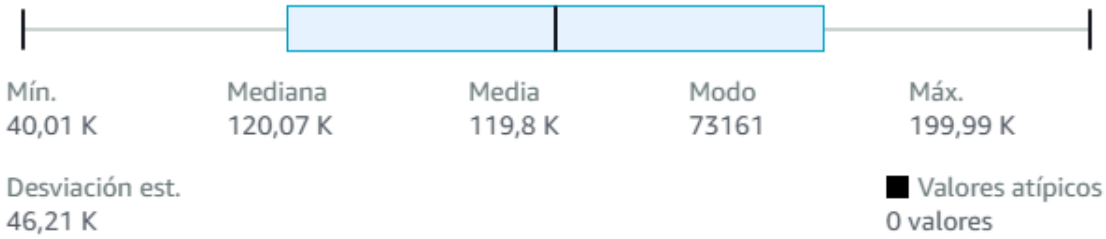
c) *¿Qué porcentaje de hombres y mujeres hay?*

En la sección estadísticas de columnas seleccionamos “gender” y observamos:



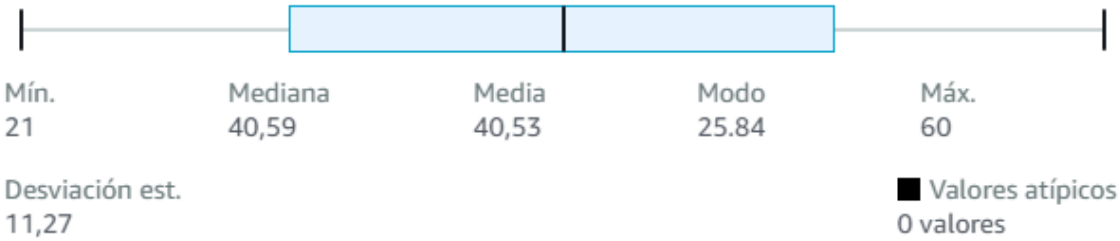
d) *Analizando el diagrama de cajas de los salarios ¿En qué horquilla se mueven? ¿Cuál es la media, mediana y moda? ¿Están distribuidos simétricamente?*

Se mueven entre 40.01 y 199.99. La media es 119.8, la mediana 120.07 y la moda 73161. Están distribuidos simétricamente.



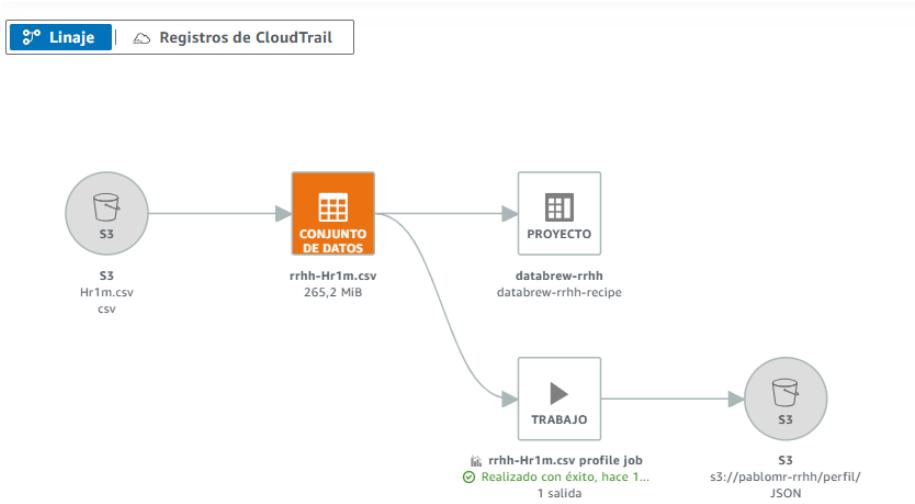
e) *Busca la misma información para el campo Age in years.*

Se mueven entre 21 y 60. La media es 40.53, la mediana 40.59 y la moda 25.84. Están distribuidos simétricamente.



6) Haz una captura del contenido de la pestaña Linaje de datos. ¿Qué se muestra en ella?

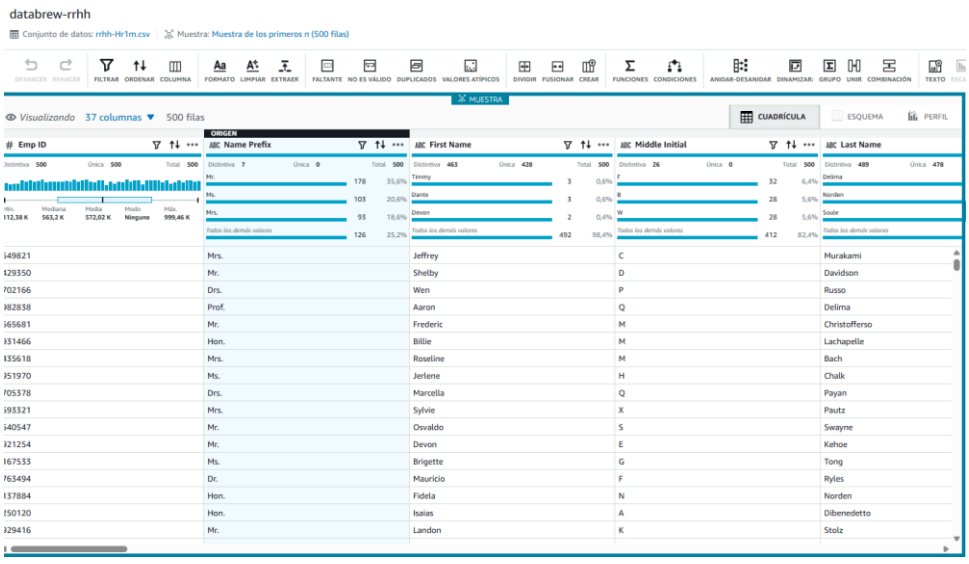
Se muestra con que elementos de AWS se relaciona el dataset.



2) APARTADO B

1) Crea un proyecto con el conjunto de datos del apartado anterior. Deja los valores por defecto. ¿Cuántos registros utiliza por defecto para el muestreo?

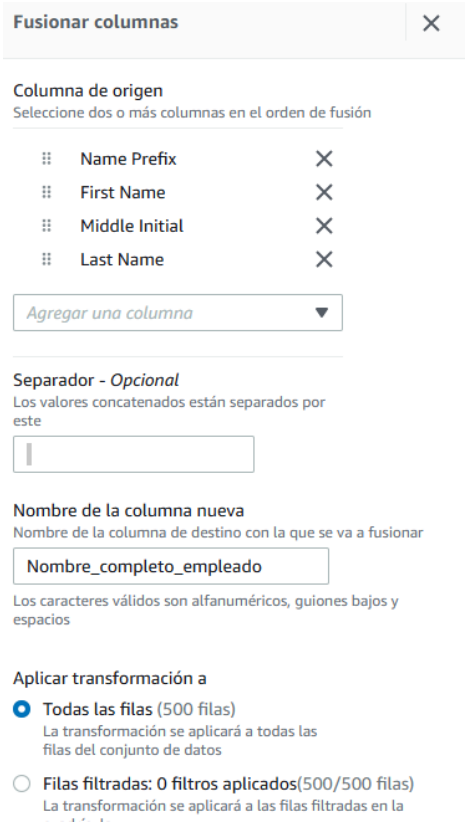
Se han utilizado 500 registros.



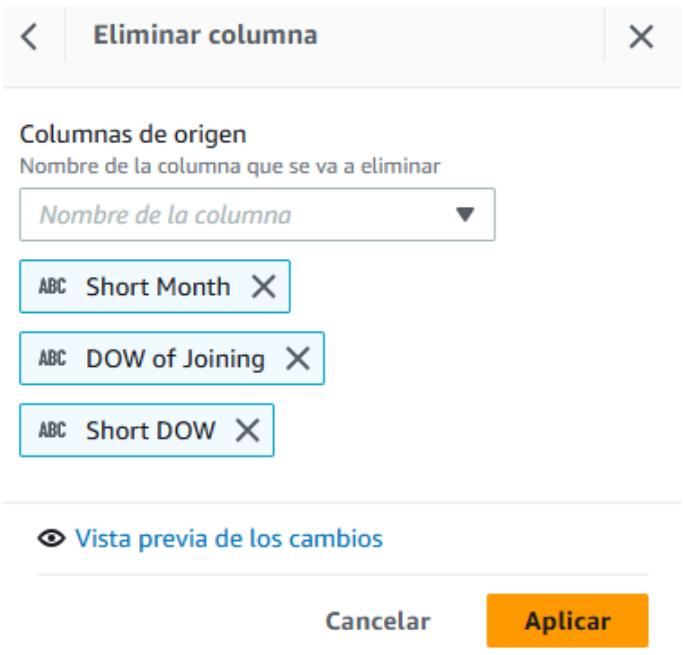
2) Generaremos una receta para realizar diferentes transformaciones a los datos:

   a) *Fusión de varias columnas en una sola. Selecciona las columnas Name Prefix, First Name, Middle Initial y Last Name como*

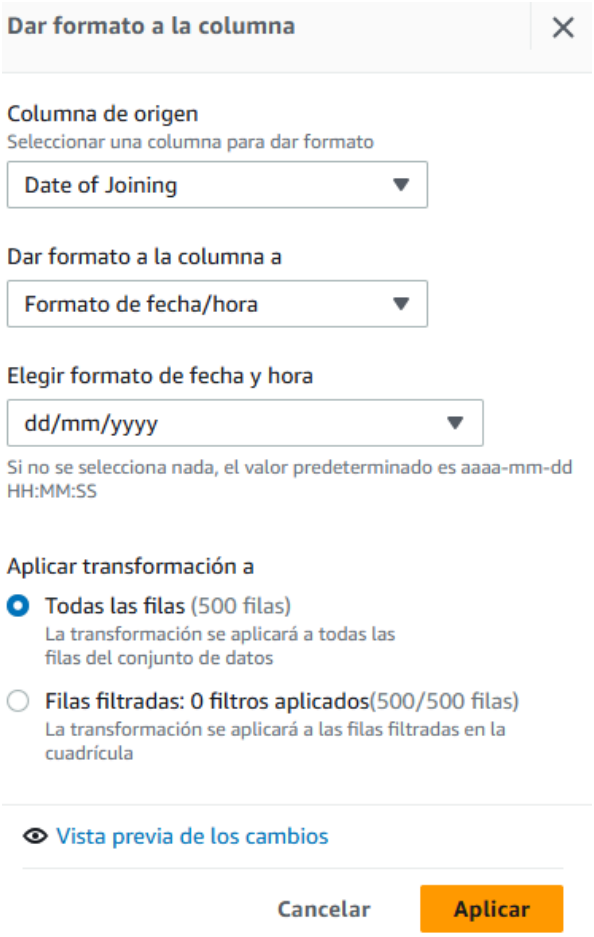
columnas de origen. Añade un espacio como separador. Como nuevo nombre de columna pondremos, por ejemplo, Nombre_completo_empleado.



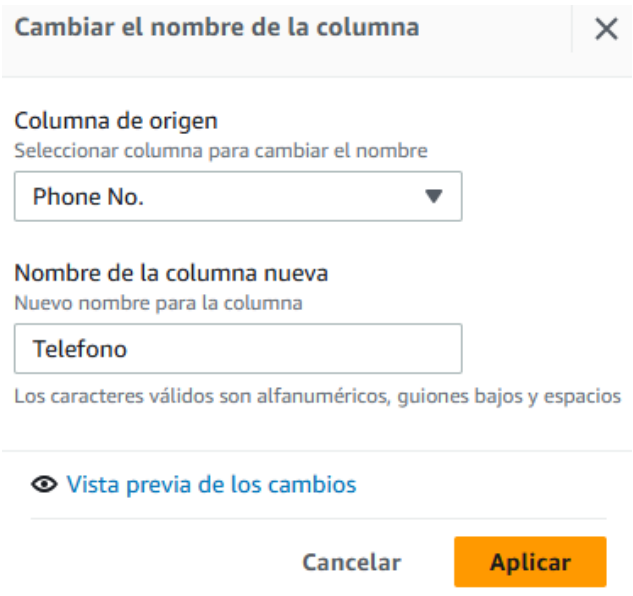
b) *Elimina las columnas Short Month, DOW of Joining y Short DOW.*



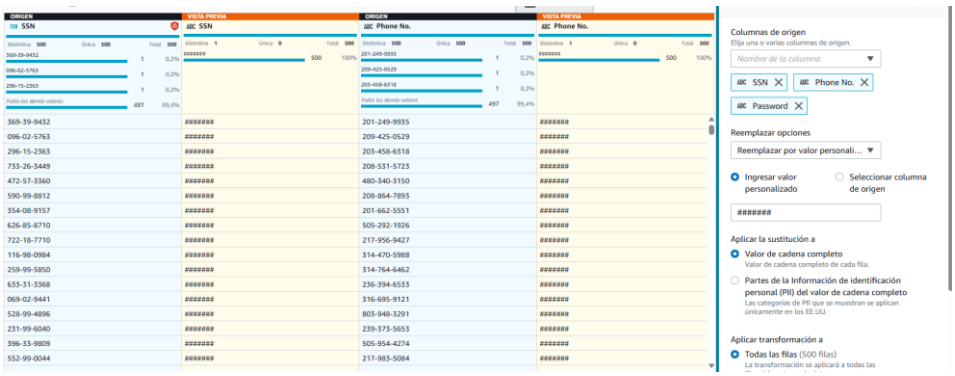
c) *Formatea la columna Date of Joining a la forma utilizada en España dd/mm/yyyy.*



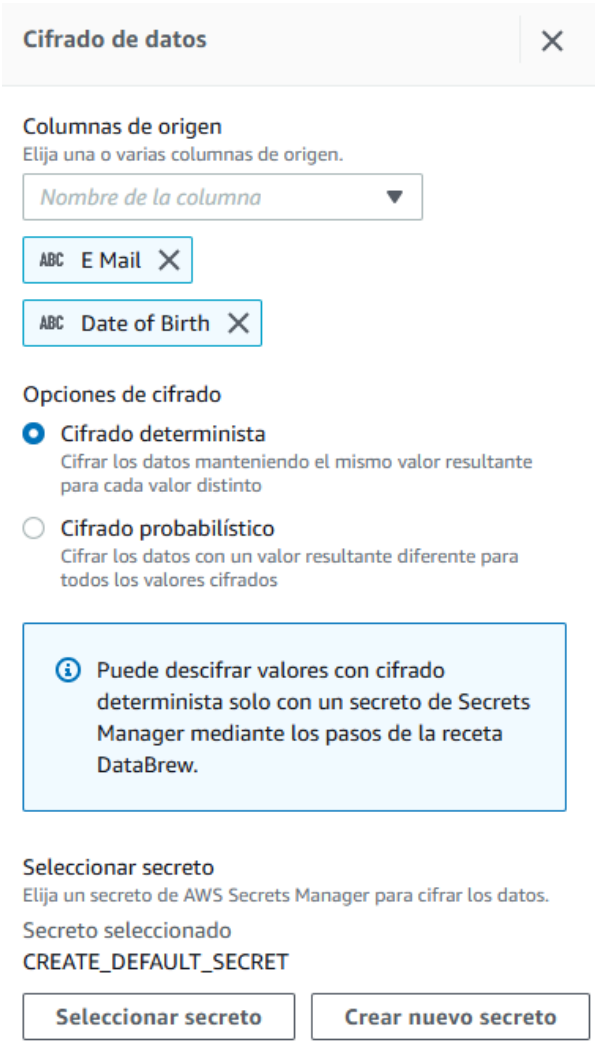
d) *Renombra la columna Phone No. a Telefono.*



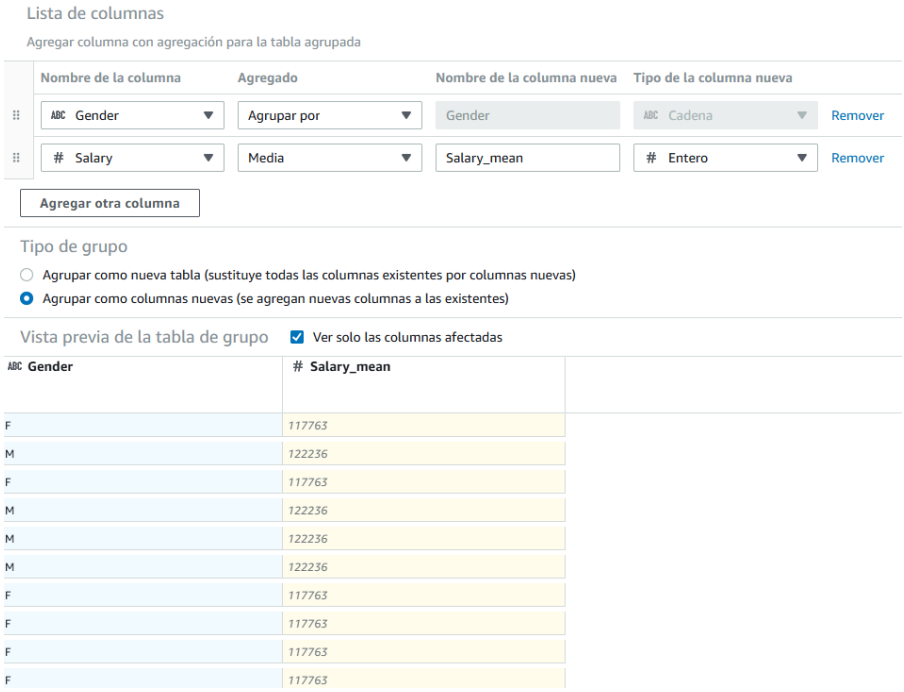
e) *Para enmascarar columnas confidenciales, cambiaremos el contenido de los campos número de la seguridad social, teléfono y contraseña por almohadillas. (Muestra cómo se hace, pero no la apliques, ya que si no un paso posterior que tenemos que hacer nos dará un error)*



f) *Realiza un cifrado determinista de los campos E Mail y Date of Birth.*

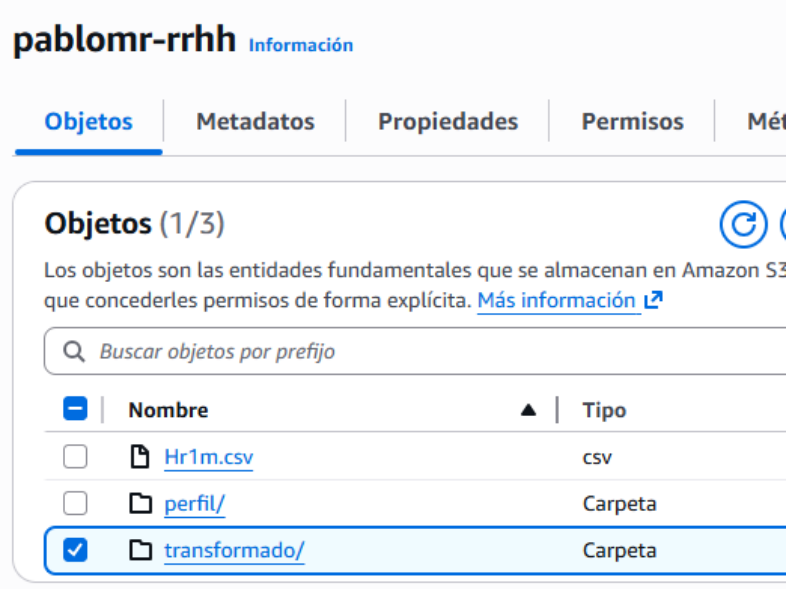


g) *Agrupar por sexo. Agrupa los datos en función del sexo y calcula cuál es el salario medio de hombres y mujeres. Una vez hechos los cálculos elimina este paso.*

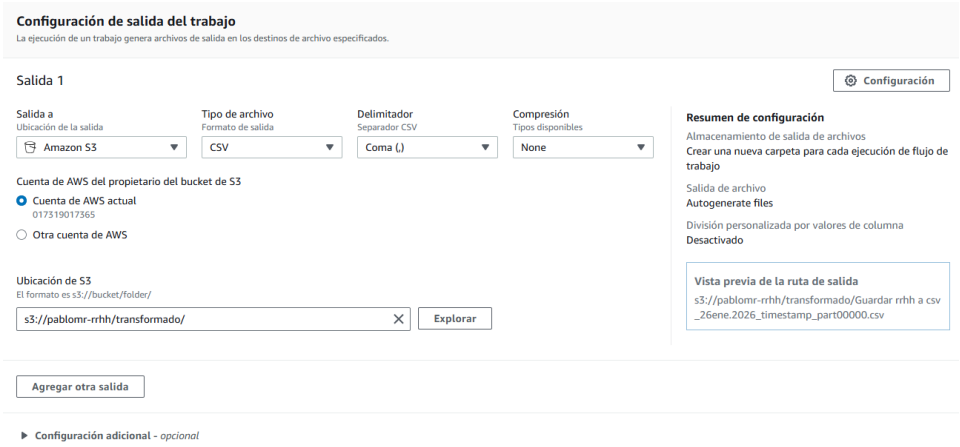


3) Publica la receta.

4) Crea en el bucket en el que estábamos trabajando una nueva carpeta llamada transformado.

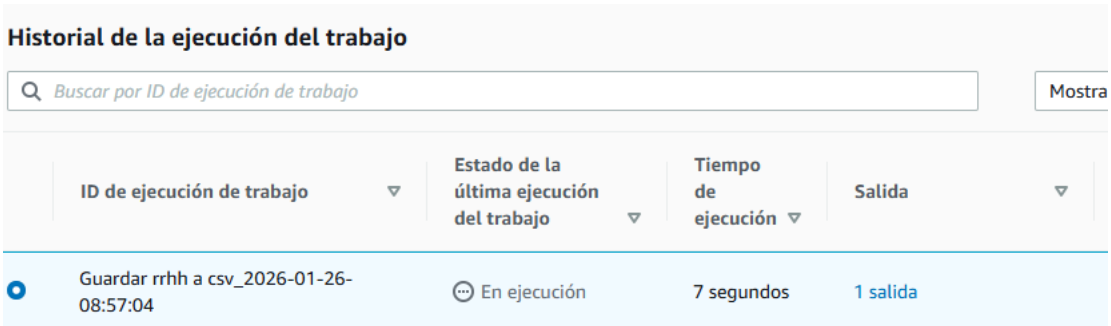


5) A partir de la receta anterior, crea un nuevo trabajo que nos deje los datos en formato CSV con comas en la carpeta del punto anterior.

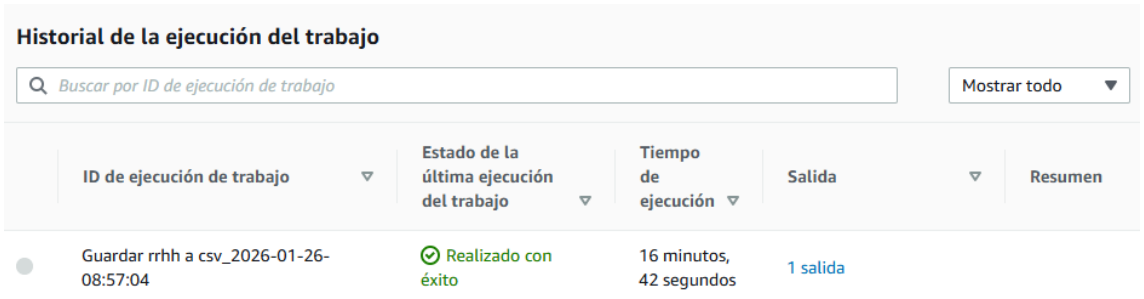


6) Descarga el CSV obtenido y échale una ojeada para verificar que se han realizado las transformaciones.

Comenzamos ejecución.



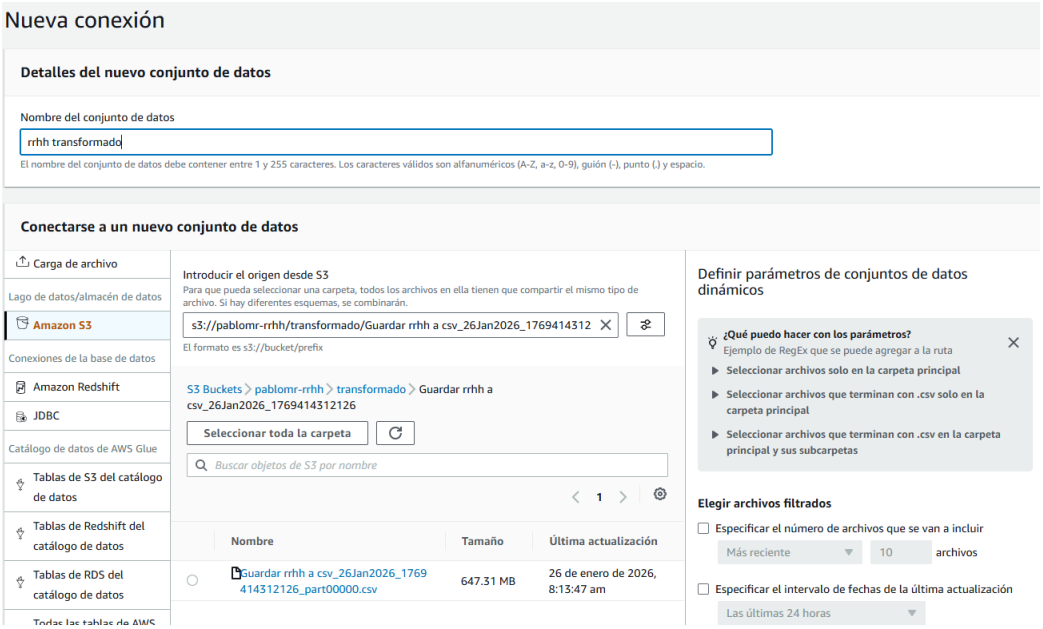
Termino por fin



Observamos el ficherillo, está perfectamente transformado.

3) APARTADO C

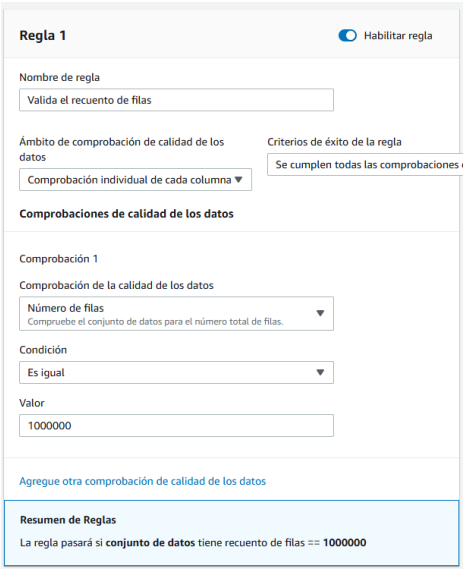
1) Crea en Databrew un conjunto de datos asociado al CSV de la carpeta transformado.



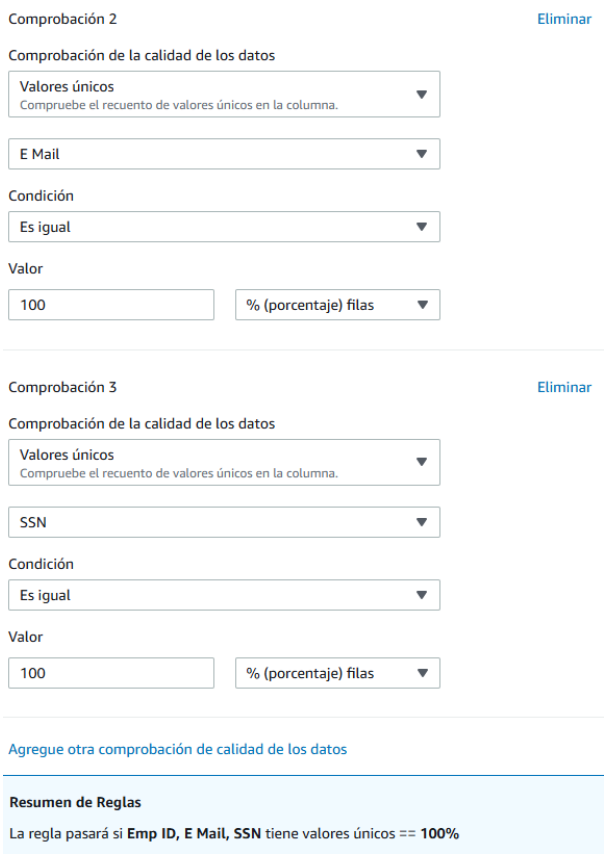
2) Crea un conjunto de reglas de calidad de los datos asociado al dataset anterior.

Añade las siguientes reglas:

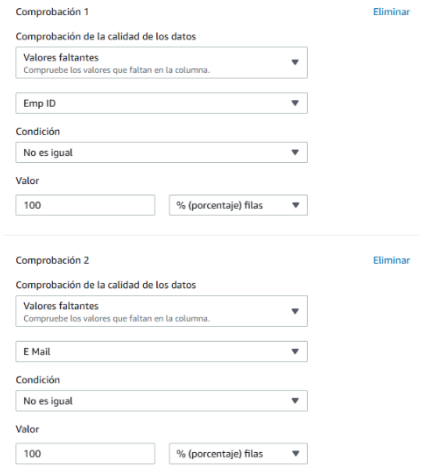
a) *Valida el recuento de filas: Hemos utilizado un conjunto de datos de 1 millón de registros. Vamos a validar si el recuento coincide.*



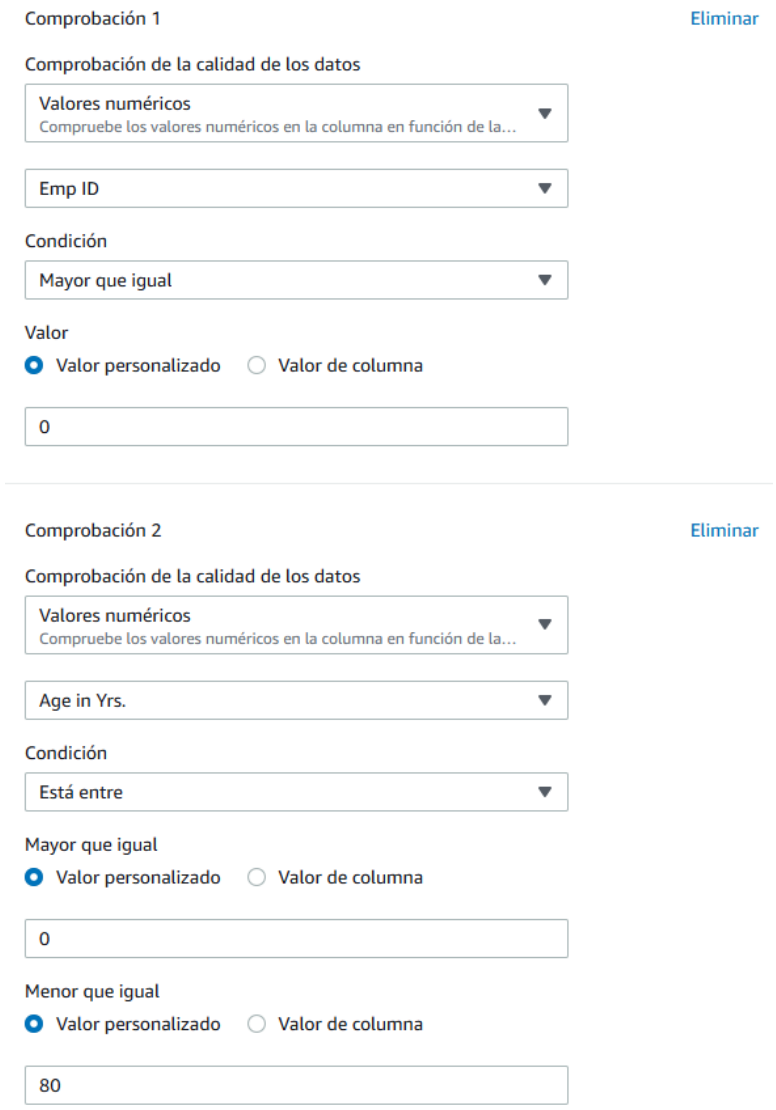
b) *El ID de empleado, la dirección de correo electrónico y el SSN deben ser únicos: Estos valores deben ser siempre únicos en el 100% de las filas.*



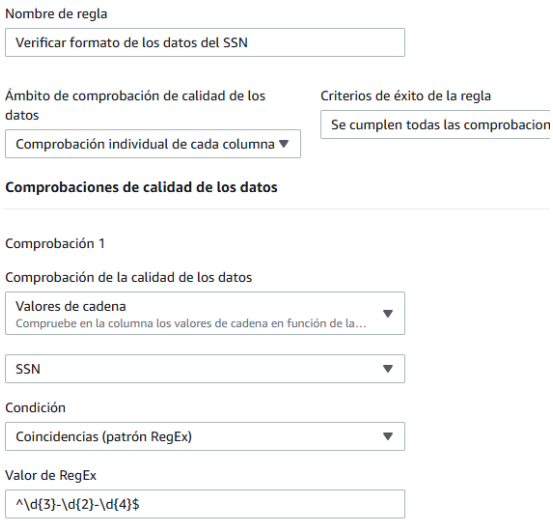
c) *El ID de empleado y la dirección de correo electrónico no deben ser nulos: Normalmente, no queremos que estos valores sean nulos en el 100% de las filas.*



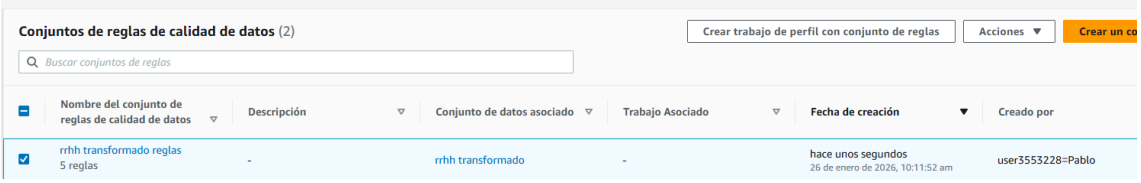
d) *El ID del empleado y la edad del empleado en años no deben tener valores negativos y además la edad debe de estar entre 0 y 80: Para ello tienes que seleccionar al crear la regla la opción de la imagen para que te permita aplicar dos comprobaciones distintas.*



e) *Verificar mediante una expresión regular (^\d{3}-\d{2}-\d{4}$) que el formato de los datos del SSN debe tener ser del tipo xxx-xx-xxxx).*

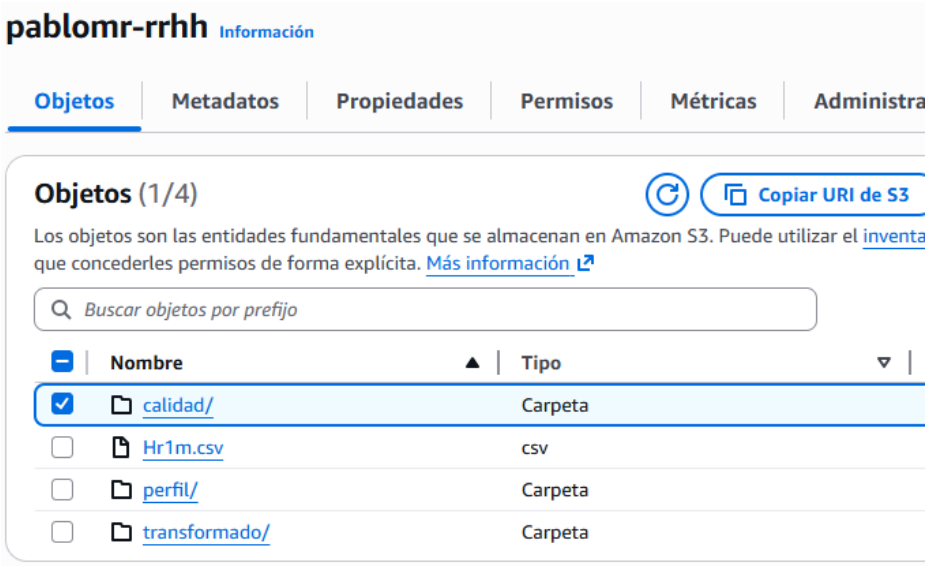


3) Crea el conjunto de reglas sin asociarlo a ningún trabajo.



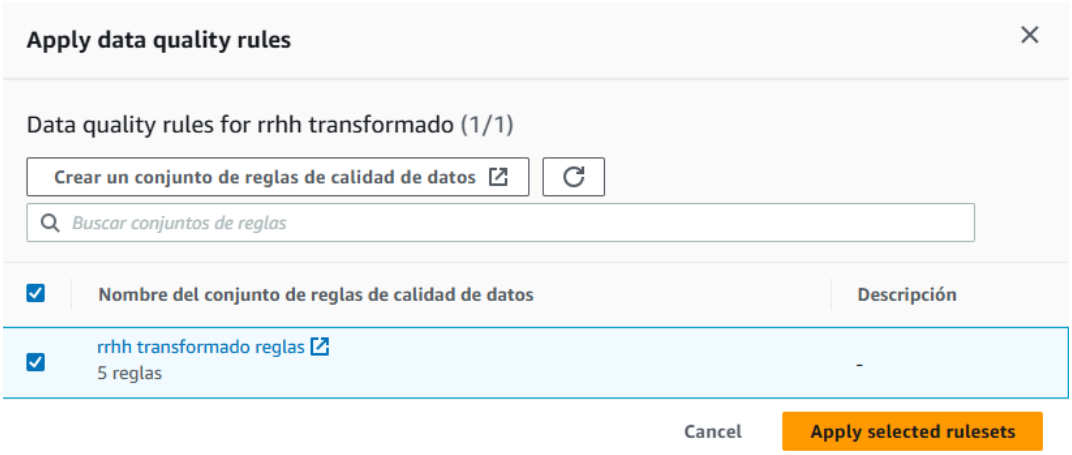
4) APARTADO D

1) Dentro del bucket de la práctica, crea una nueva carpeta llamada calidad que utilizaremos posteriormente para almacenar la salida del análisis de calidad que vamos a realizar.

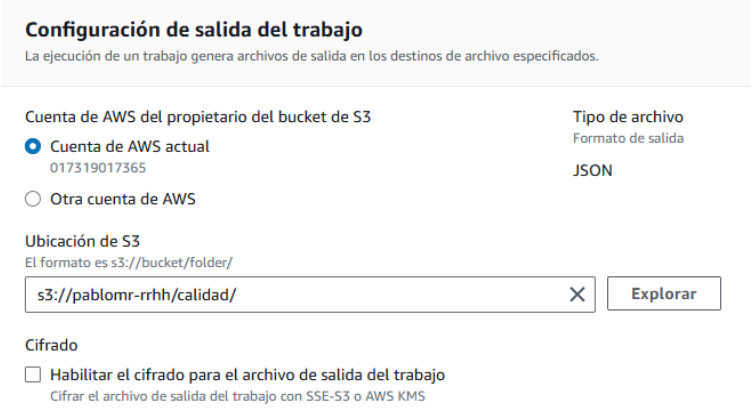


2) Vete al apartado de reglas de calidad en Databrew y asocia las reglas creadas a un trabajo de perfil.

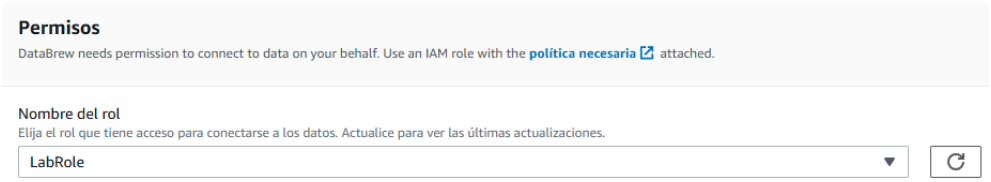
    a) *Aplica el trabajo a todo el dataset.*



b) *Configura el bucket de salida del análisis en la carpeta del punto anterior.*

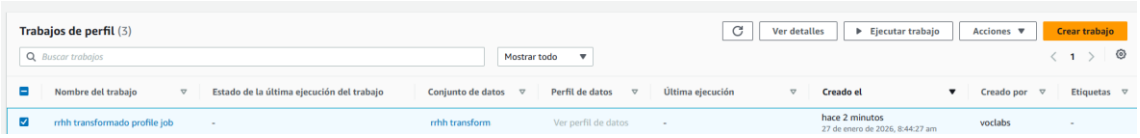


c) *Verifica en rol adecuado en el apartado de Permisos (Labrole). Crea el trabajo.*

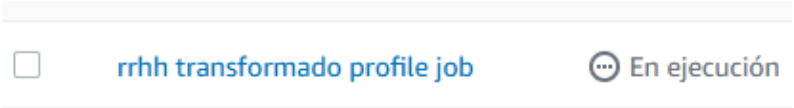


3) Verifica el nombre del trabajo de perfil asociado a las reglas y posteriormente vete a trabajos de perfil y ejecútalo.

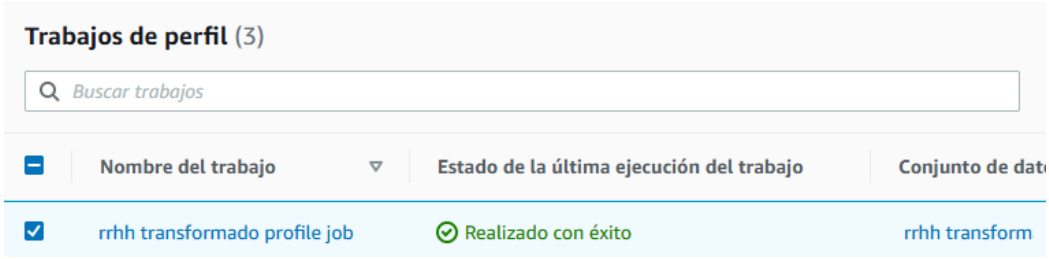
Aquí está el trabajo:



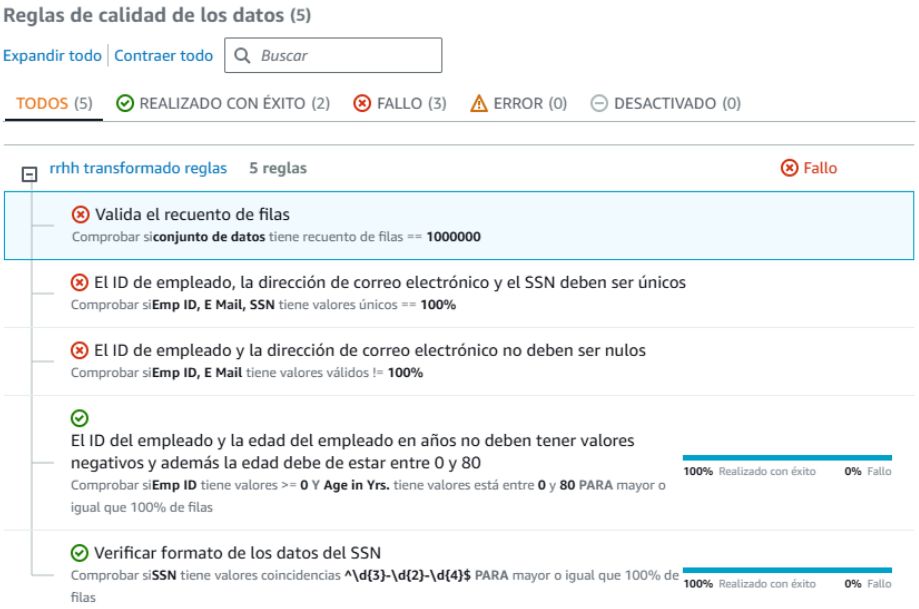
Lo ejecutamos:



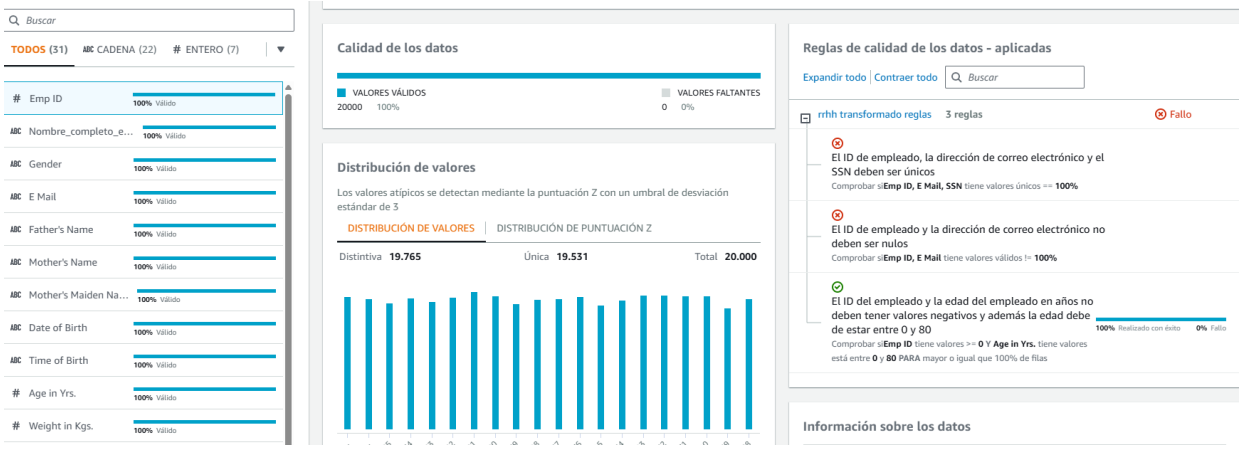
Gestionado:



4) Una vez ejecutado accede al enlace Ver perfil de datos y dentro de la pestaña Reglas de calidad de datos verifica el resultado de la comprobación de las reglas configuradas.

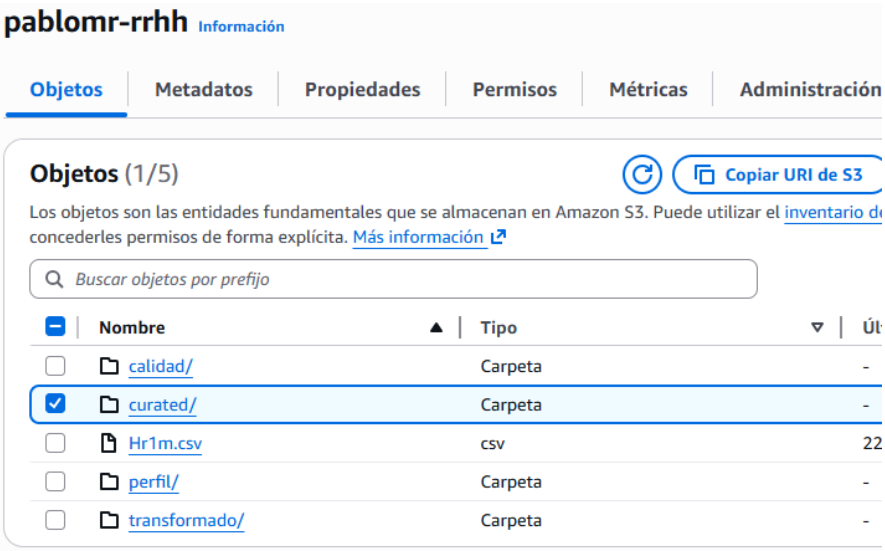


5) De aparecer algún error (por ejemplo, en la imagen me dice que hay SSN y Emp ID repetidos), vete a la pestaña de Estadísticas de columna y comprueba que es cierto el error de las reglas (por ejemplo, en la imagen puedo ver los Emp ID repetidos)

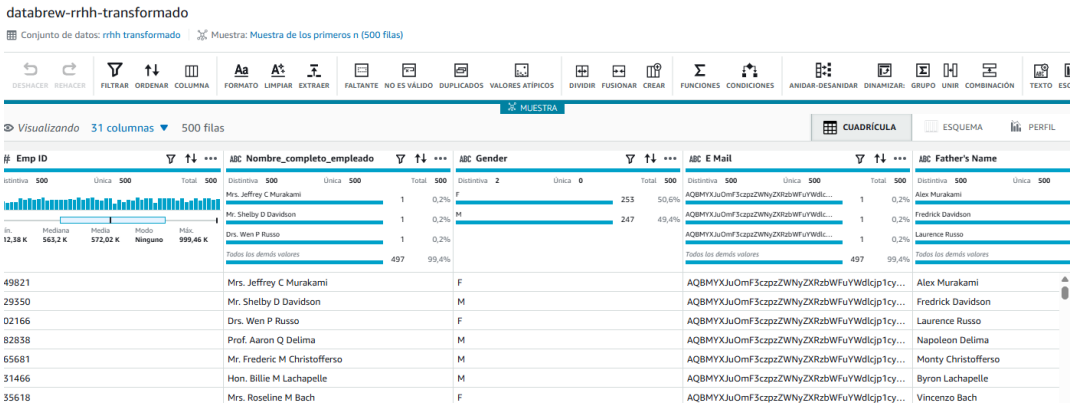


5) APARTADO E

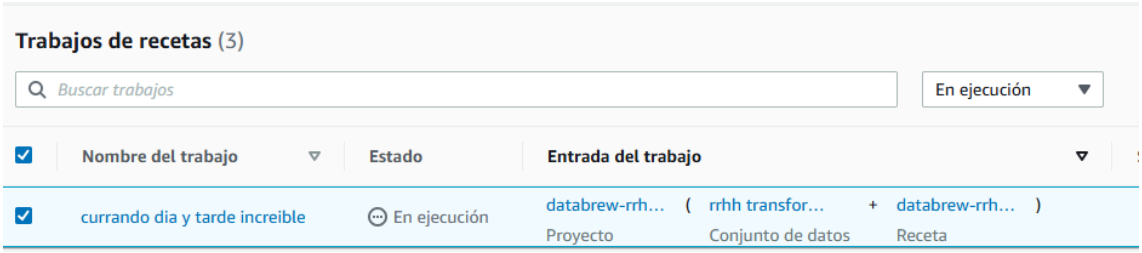
1) Crea en el bucket en el que estábamos trabajando una nueva carpeta llamada curated.



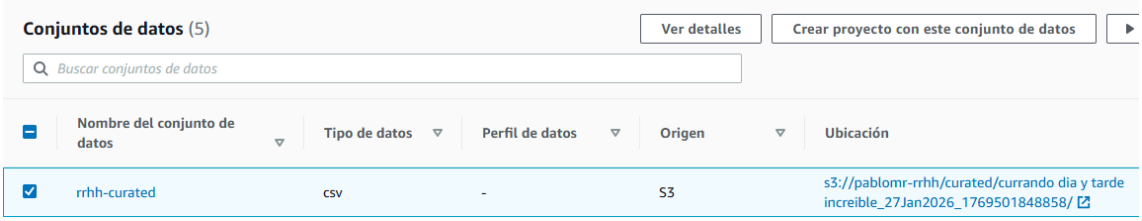
2) De modo similar a como hicimos en el apartado B, crea un nuevo proyecto que a partir del conjunto de datos que tenemos en la carpeta transformado y mediante una nueva receta y un nuevo trabajo intenta corregir los errores aparecidos en el ejercicio anterior (por ejemplo, eliminando filas con campos duplicados). El resultado del trabajo almacénalo en formato parquet comprimido en la carpeta curated.



Creamos el trabajo



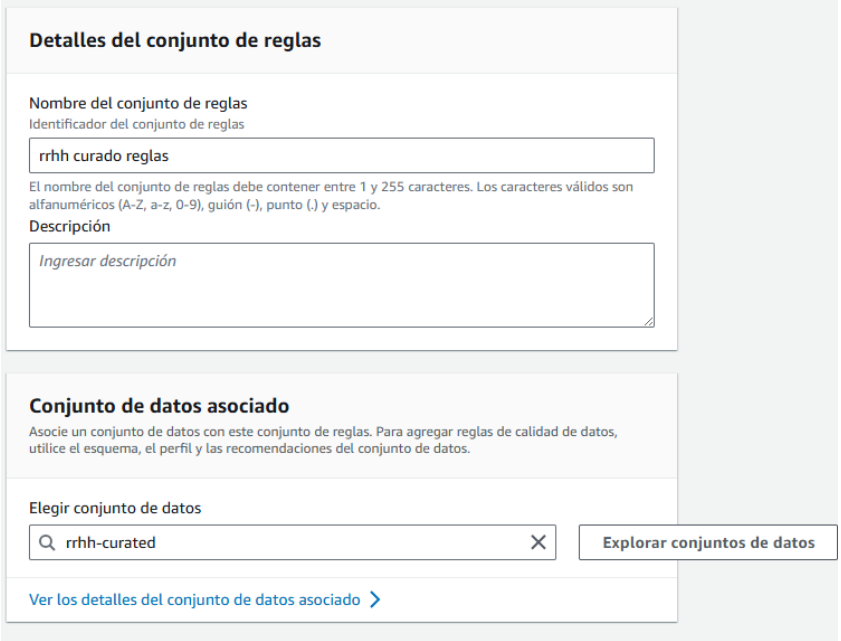
3) Crea un nuevo conjunto de datos en Databrew que apunte al archivo de curated.



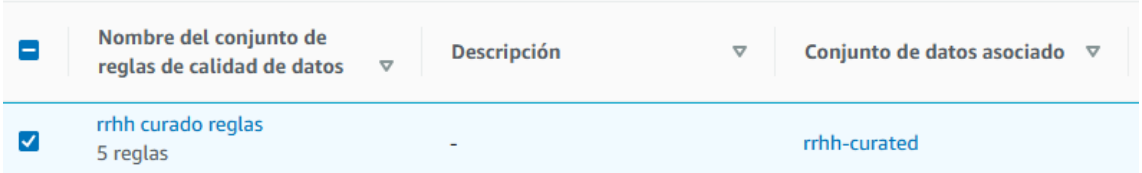
4) Verifica ahora las estadísticas de las columnas que has modificado para asegurarnos que todo ha ido bien (por ejemplo, que no haya datos repetidos en Emp ID). ¿Cuántas filas tiene ahora el archivo resultante?

6) APARTADO F

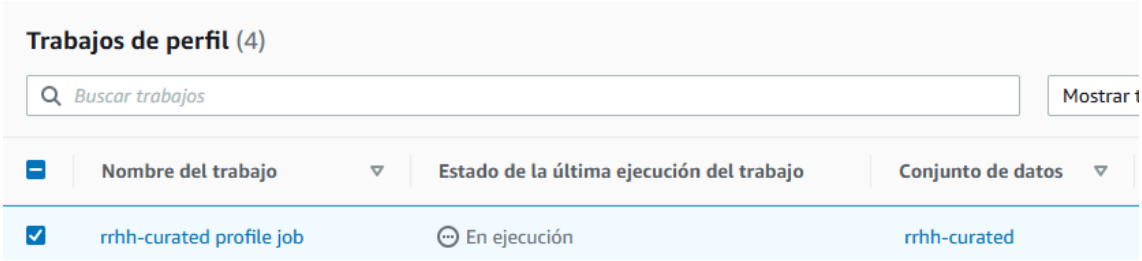
1) Duplica el conjunto de reglas de calidad del Apartado C, pero ahora hazlo apuntar al dataset de la carpeta curated. Modifica alguna regla si fuese necesario, por ejemplo, la que nos contaba el número de filas. (Puede ser que los nombres de los campos hayan cambiado respecto a los originales, si es así modifícalos en el conjunto de reglas)



2) Asocia dicho conjunto de reglas al trabajo de perfil.



3) Ejecuta dicho trabajo contra todo el dataset.



4) Verifica en el perfil de datos, apartado Reglas de calidad que se han pasado correctamente todas las comprobaciones.