AWS GLUE III (Databrew)

Glue DataBrew

AWS Glue DataBrew es un herramienta visual de transformación de datos que nos permite aplicar transformaciones a los datos de forma visual, sin necesidad de escribir o administrar código (*serverless*).

Incluye más de 250 transformaciones de datos integradas, que se pueden ensamblar fácilmente de forma gráfica para crear una receta de *DataBrew*, lo que permite aplicar varias transformaciones a un conjunto de datos, como por ejemplo, limpiar y normalizar datos, eliminar o sustituir valores nulos, estandarizar columnas de fecha y hora para que se ajusten a un estándar, crear codificaciones, etc...

DataBrew incluye funcionalidades tanto para la elaboración de perfiles de datos (recopilación de estadísticas sobre las distintas columnas del conjunto de datos) como para la supervisión de la calidad de los datos. También incluye muchos tipos diferentes de transformaciones, como el formateo de datos, la ofuscación de datos personales, la división o unión de columnas, la conversión de zonas horarias, la detección y eliminación de valores atípicos, etc...

Alternativas

Existen multitud herramientas similares, como por ejemplo, Azure Data Factory, Google DataFlow, o PowerQuery dentro de PowerBI

Los diferentes componentes de *DataBrew* son:

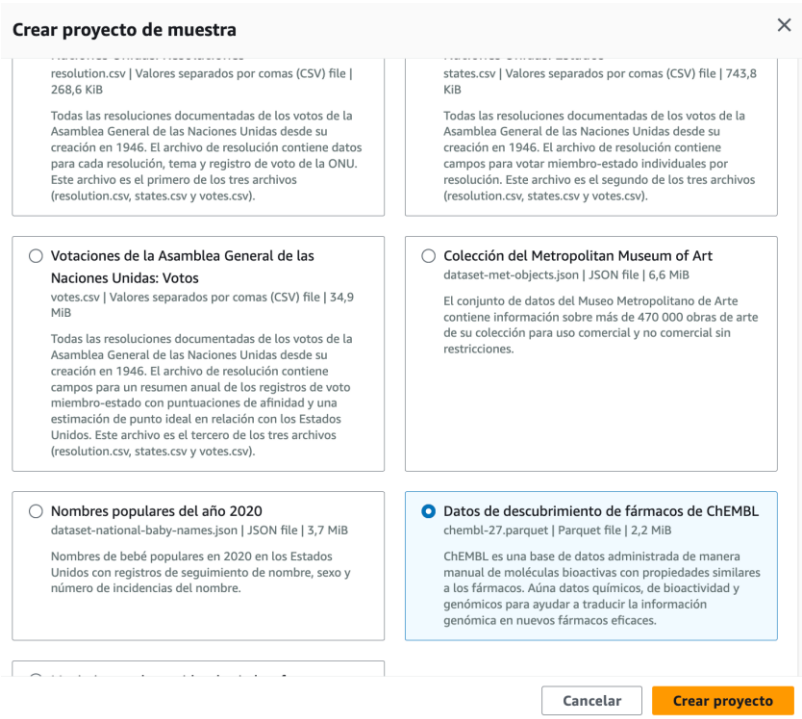
- Proyecto
- Conjunto de datos (*dataset*), que se almacenan en S3.
- Receta (*receipt*), compuestas de uno o varios pasos de transformación. Estas recetas se pueden guardar, publicar, crear versiones, etc. y compartirlas con otros.
- Trabajo (*job*), el cual se puede orquestar mediante *Step Functions*.

Los trabajos de *DataBrew* cuestan 1$ por sesión y luego 0,48$ por hora de nodo empleado.

Ejemplo DataBrew

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.

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:



DataBrew - Creación del proyecto de muestra

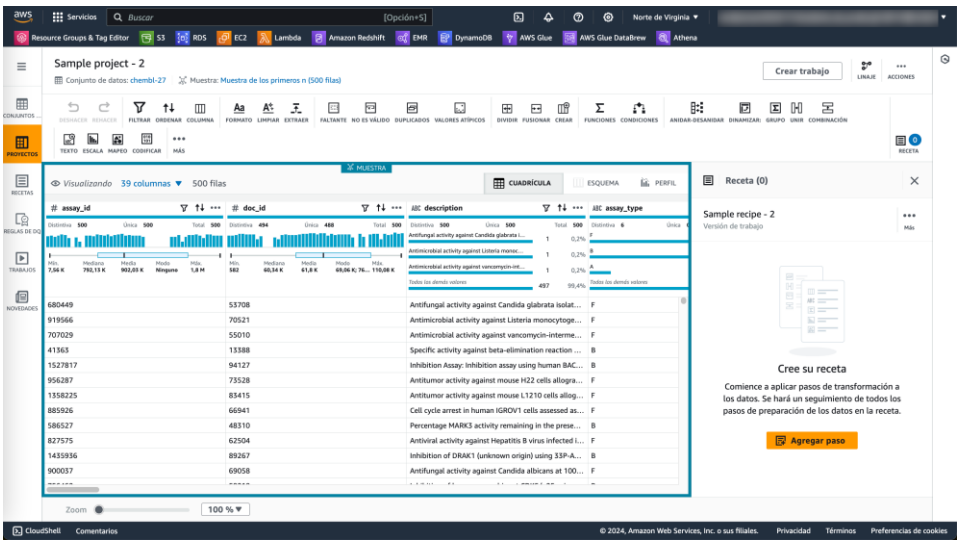
Nada más crear el proyecto, se creará una sesión que tarda alrededor de un minuto en instanciarse.

Cuidado con las sesiones

Cada sesión se factura por franjas de 30 minutos, a 1$. Por ejemplo, si utilizamos 40 minuto serían 2$.

Cabe destacar que las primeras 40 sesiones interactivas son gratuitas para los usuarios de *DataBrew* por primera vez.

Tras crearse la sesión, veremos un interfaz de trabajo con los datos mostrados de forma tabular, con un menú superior con las operaciones/transformaciones disponible, y donde podemos observar cómo tenemos 39 columnas y que sólo se están mostrando 500 filas. Para cada columna, podemos ver un gráfico con diversa información estadística y luego datos de ejemplo de dicha columna. Finalmente, en la parte derecha tenemos los pasos (la receta) que se aplican al *dataset*:

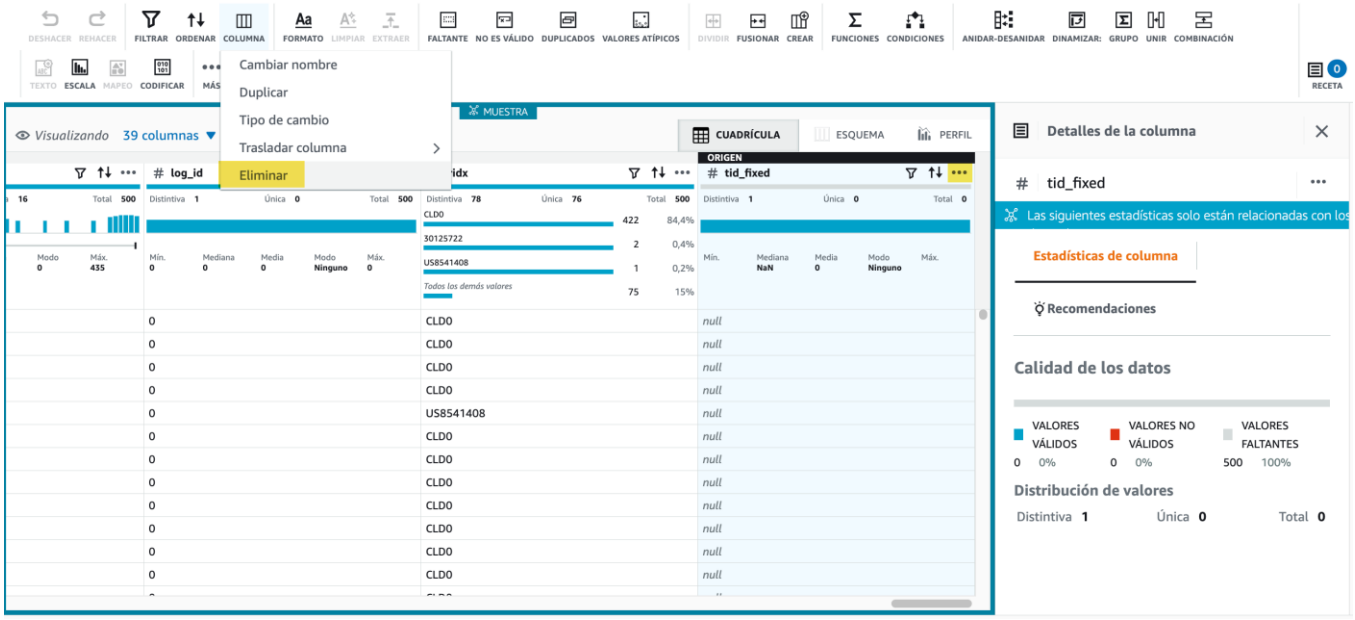


DataBrew - Interfaz de DataBrew

Creando la receta

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

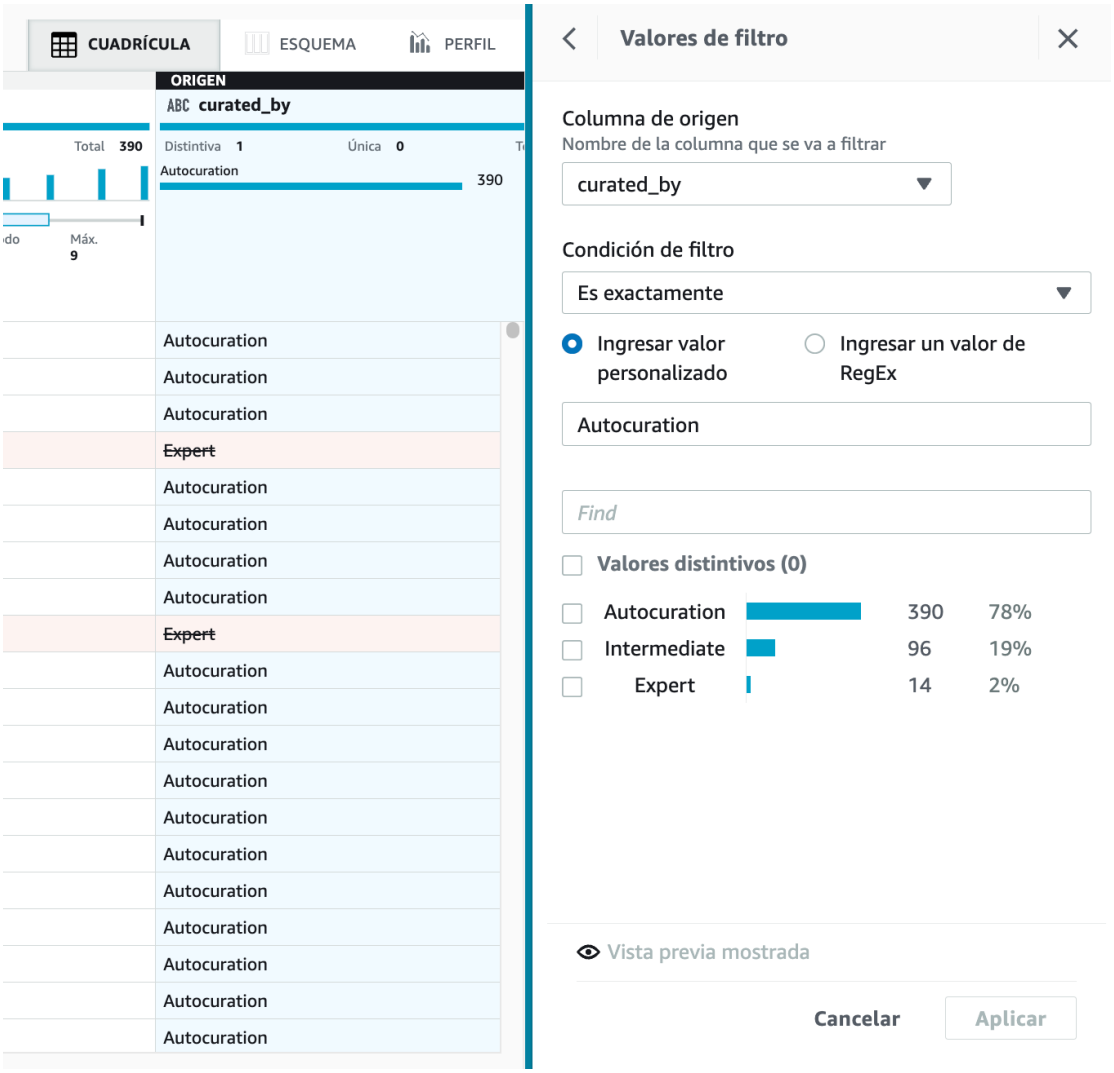
1.- El primer paso será eliminar la última columna, *tid_fixed* que tiene todos los valores nulos. Para ello, bien desde el menú *Columna*, seleccionamos la opción de *Eliminar*.



DataBrew - Borrando una columna

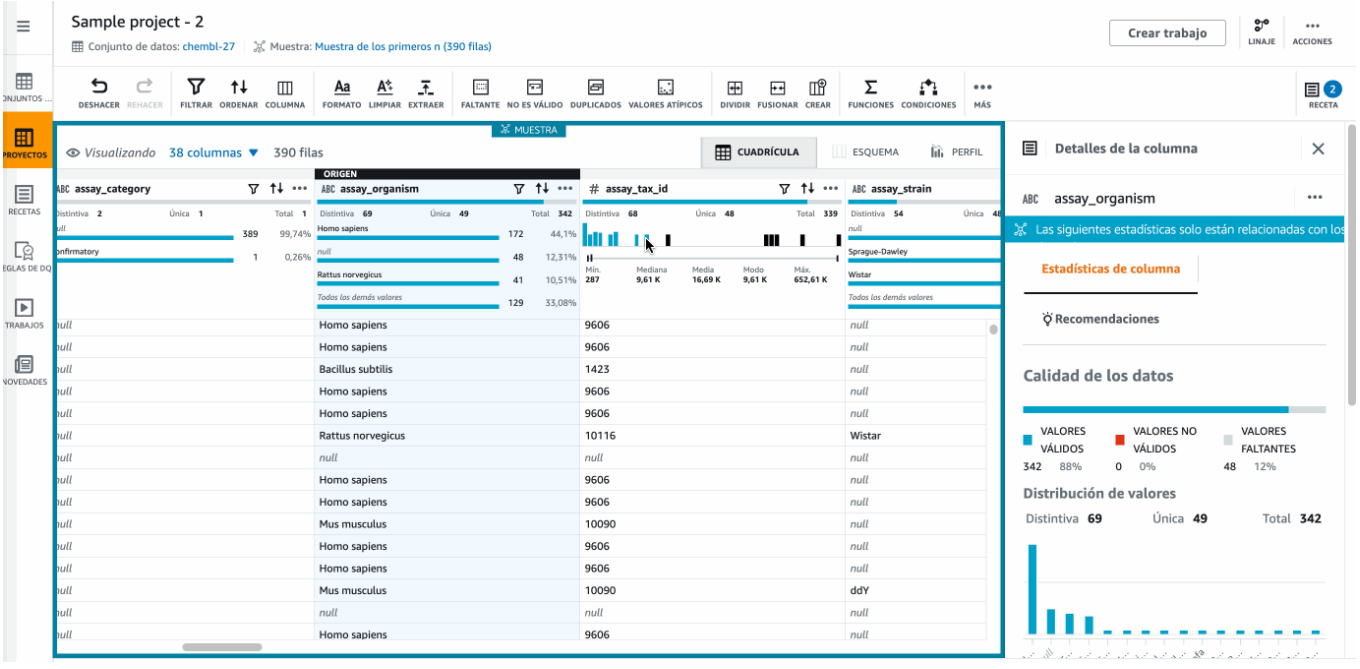
Tras aplicar los cambios, en la zona de la receta, aparecerá el paso aplicado.

2.- 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.



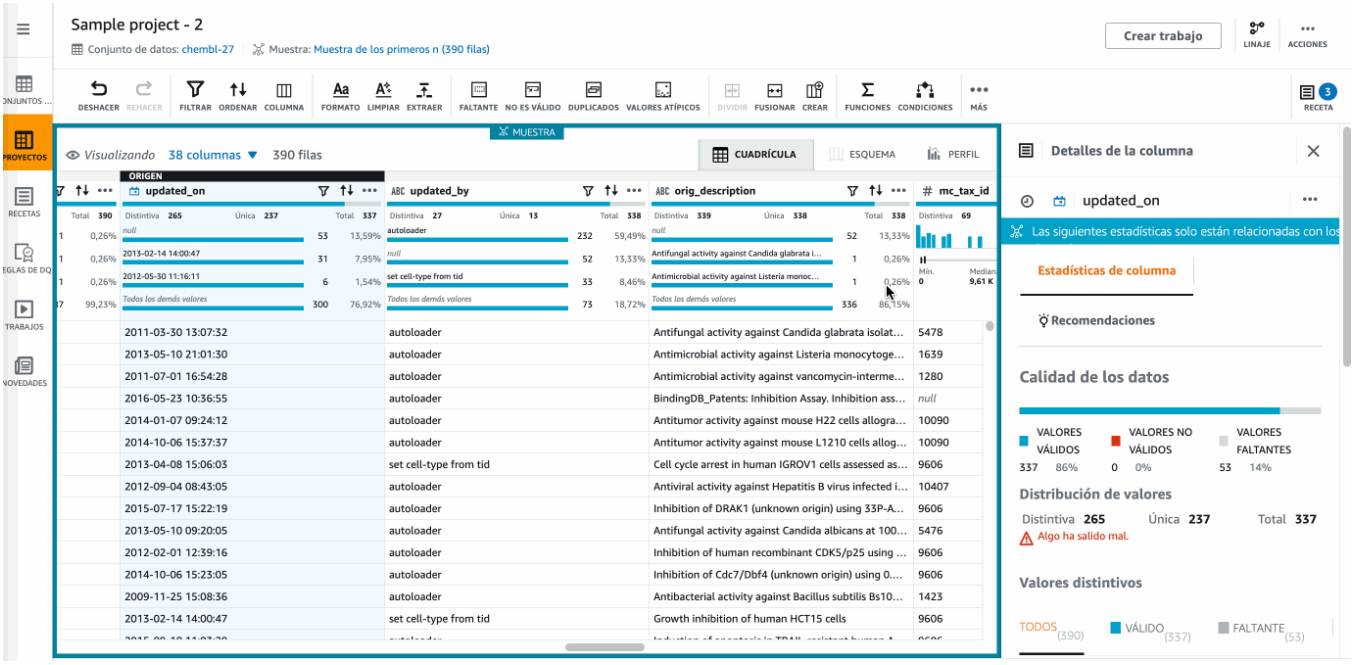
DataBrew - Filtrando datos

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*:



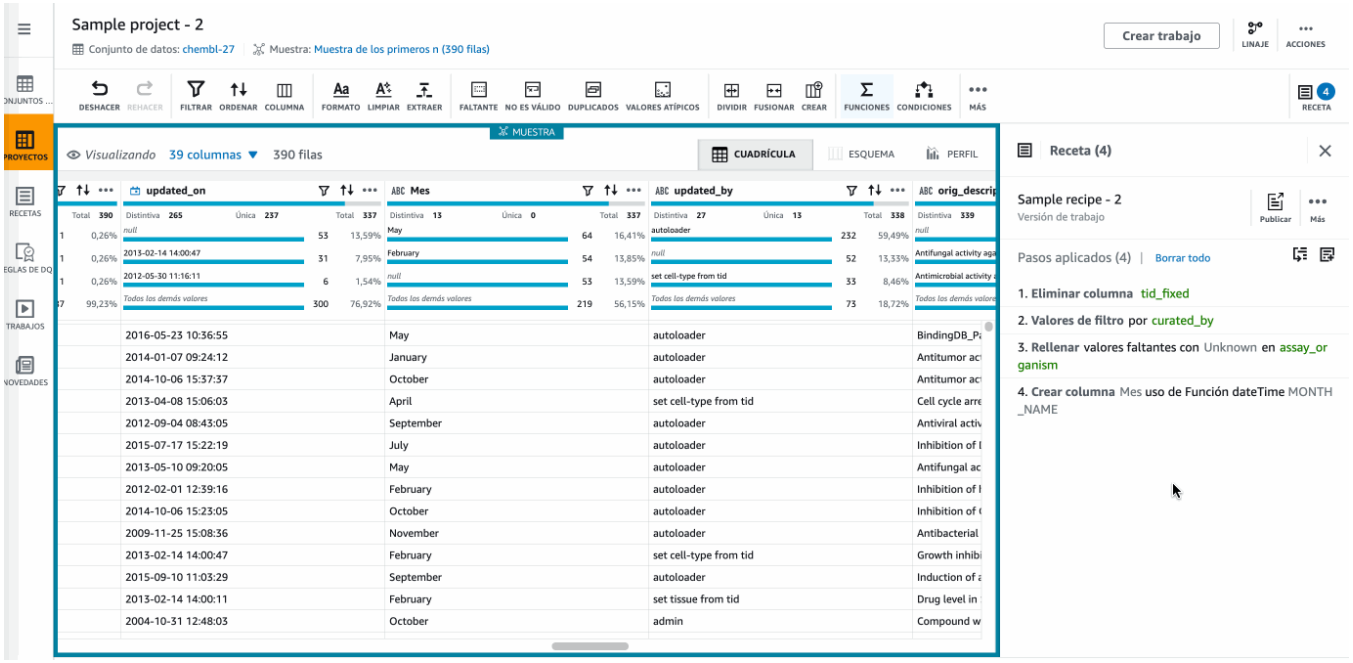
DataBrew - Cambiando los nulos

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:



DataBrew - Utilizando funciones

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.

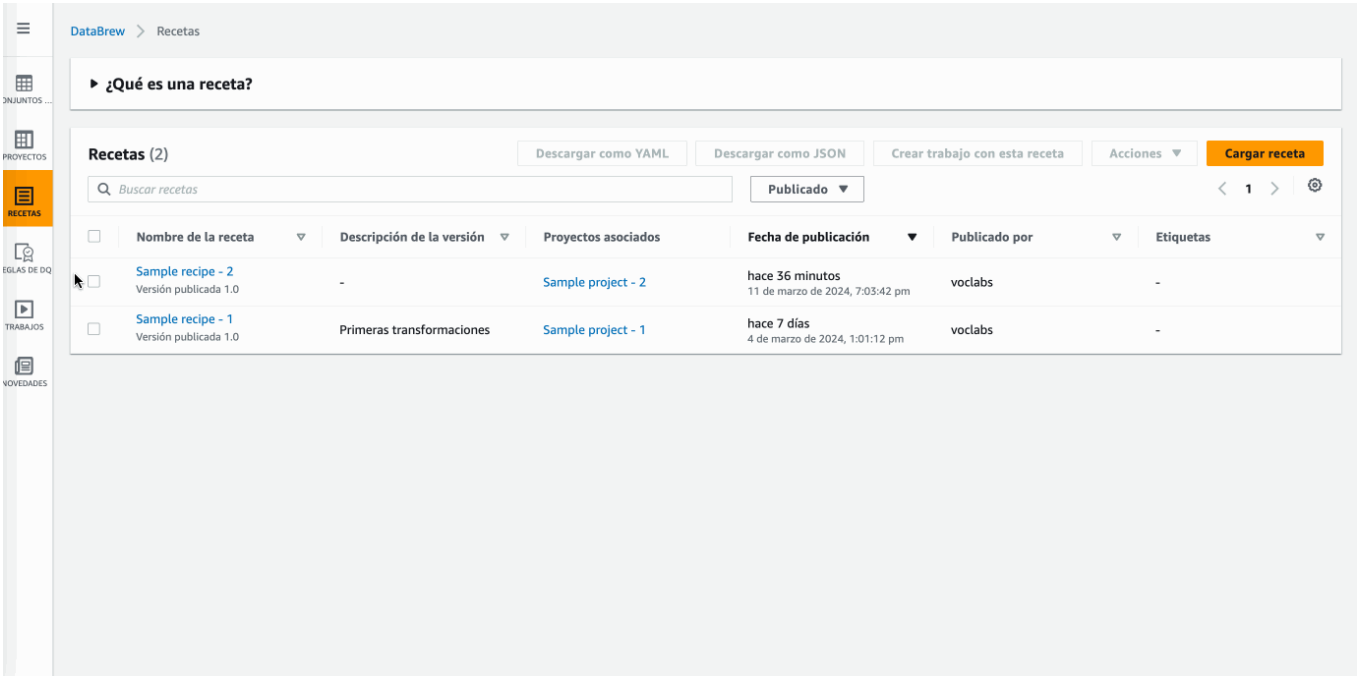


DataBrew - Publicando la receta

De la receta al *job*

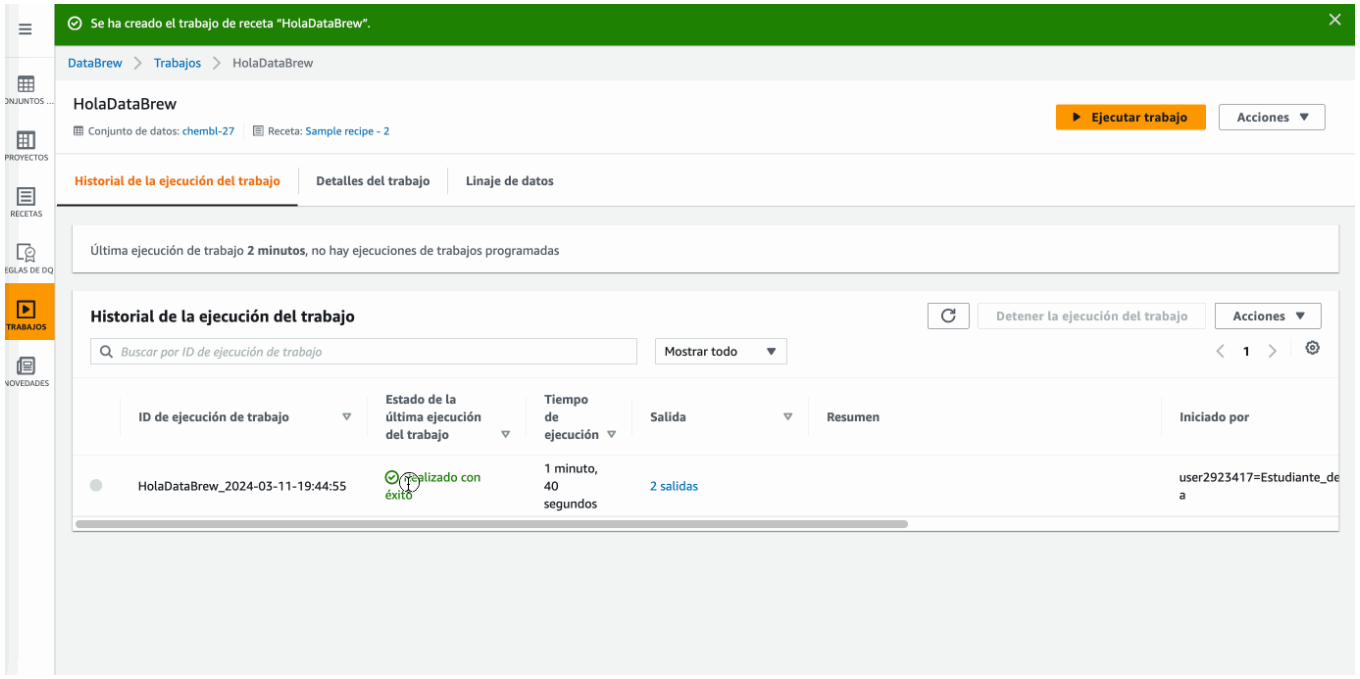
6.- Si vamos al menú de las recetas, seleccionamos la receta recién publicada, en nuestro caso *Sample recipe-2*, y creamos un trabajo (*job*) con la misma, en el cual, tras darle un nombre y seleccionar el *dataset*, vamos a guardar el resultado en S3 tanto en formato

CSV como en formato *Parquet* particionado por la columna *Mes*, y finalmente seleccionamos el rol *LabRole*:



DataBrew - Creando un job

Tras la creación, el *job* se ejecutará automáticamente. Si vamos a S3 veremos cómo ha creado una carpeta por cada salida y dentro estarán los datos transformados.



DataBrew - Resultado del job

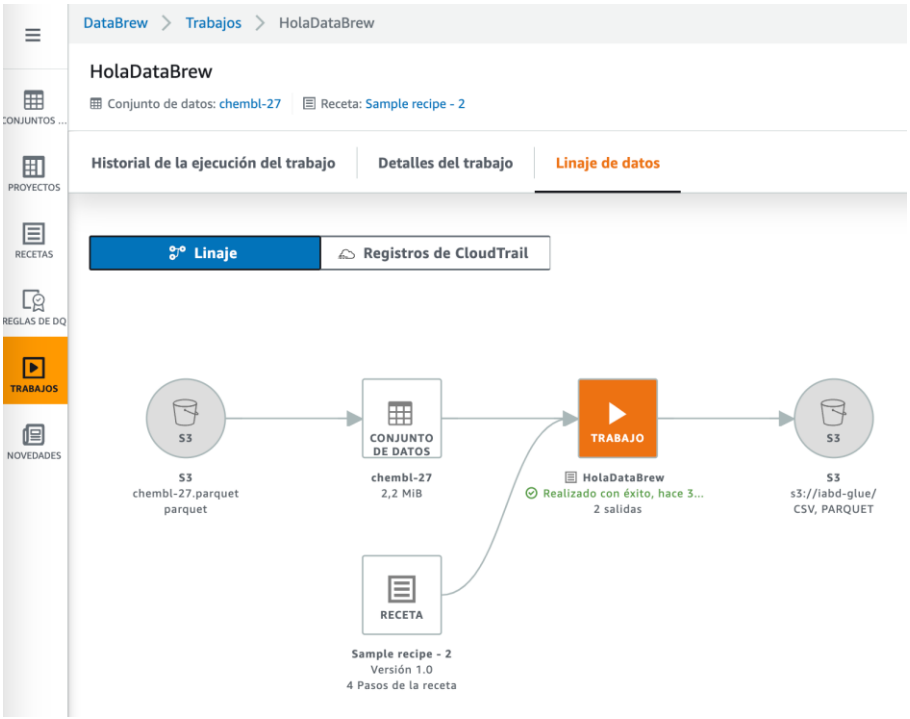
Precio de ejecución

A la hora de diseñar una receta, tenemos que pagar por la sesión. Una vez creada, AWS cobrará por cada ejecución de los *jobs* dependiendo de la cantidad de nodos asignados, a 0,48$ por nodo/hora facturados por minutos. Más información en https://aws.amazon.com/es/glue/pricing/

Entre las diferentes funcionalidades extra que nos ofrece *DataBrew* es poder programar la ejecución de los *jobs* (por ejemplo, determinados días/horas o mediante expresiones CRON), así como visualizar el linaje de los datos de los *job*:

El **linaje de los datos** (conocido en inglés como *data lineage*) se refiere al **seguimiento completo y detallado del recorrido que siguen los datos** a lo largo de su ciclo de vida dentro de un sistema o ecosistema de datos. Es como un "mapa genealógico" que documenta:

- **Origen** (dónde nacen o se capturan los datos: fuentes crudas como flujos de clics, bases de datos, sensores IoT, archivos JSON/XML, etc.).
- **Transformaciones** (cómo se limpian, enriquecen, agregan, unen o modifican en cada paso: por ejemplo, en procesos de limpieza, imputación de valores faltantes, creación de nuevas variables o uniones de tablas).
- **Movimientos** (cómo se mueven entre sistemas: ingestión a un lago de datos como Amazon S3, procesamiento en herramientas ETL, almacenamiento en data *warehouses*, o consumo en modelos de ML/BI).
- **Destino final** (dónde terminan: informes, *dashboards*, modelos predictivos, aplicaciones o decisiones empresariales).



DataBrew - Linaje de un job

Calidad de datos con AWS Glue DataBrew

Nos centraremos aquí en el uso de AWS Glue DataBrew para realizar comprobaciones de validación y calidad de datos con ejemplos reales.

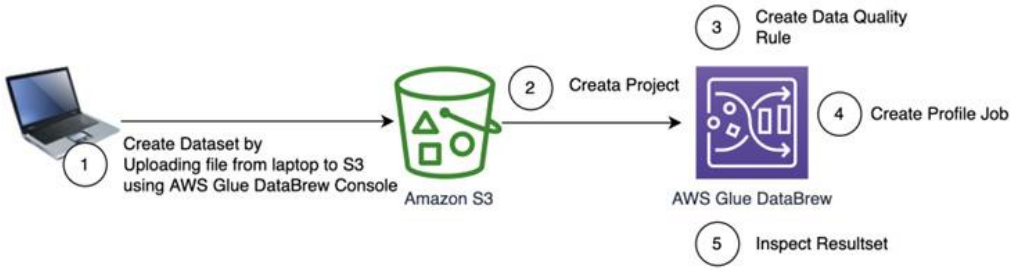
En los nuevos modelos empresariales las organizaciones se esfuerzan por centrarse en los datos y, para ellas, es fundamental abordar los problemas de calidad de datos a los que se enfrentan o disponer de las herramientas y los procesos necesarios para solucionarlos.

La mala calidad de los datos no es sólo un problema tecnológico, ya que puede tener un impacto significativo en las empresas. Según un documento de investigación de Gartner en 2018, las organizaciones creen que la mala calidad de los datos cuesta un promedio de 15 millones de dólares al año en pérdidas.

Es importante señalar que la exactitud de las estimaciones depende en gran medida de la exactitud de los datos subyacentes.

Como ya hemos comentado en el tema anterior, el concepto informático de basura entrante, basura saliente (GIGO) determina la calidad de los datos. La calidad de la salida de datos viene determinada por la calidad de su entrada. En términos sencillos, si proporcionamos datos malos como entrada, obtendremos datos malos como salida. La calidad de los datos está relacionada con su integridad, coherencia, validez, exactitud, unicidad, integridad y puntualidad.

Ahora que entendemos el impacto de la calidad de datos en la tecnología y el negocio, avancemos con nuestro ejemplo de creación de reglas de calidad de datos en AWS Glue DataBrew para identificar y solucionar problemas.

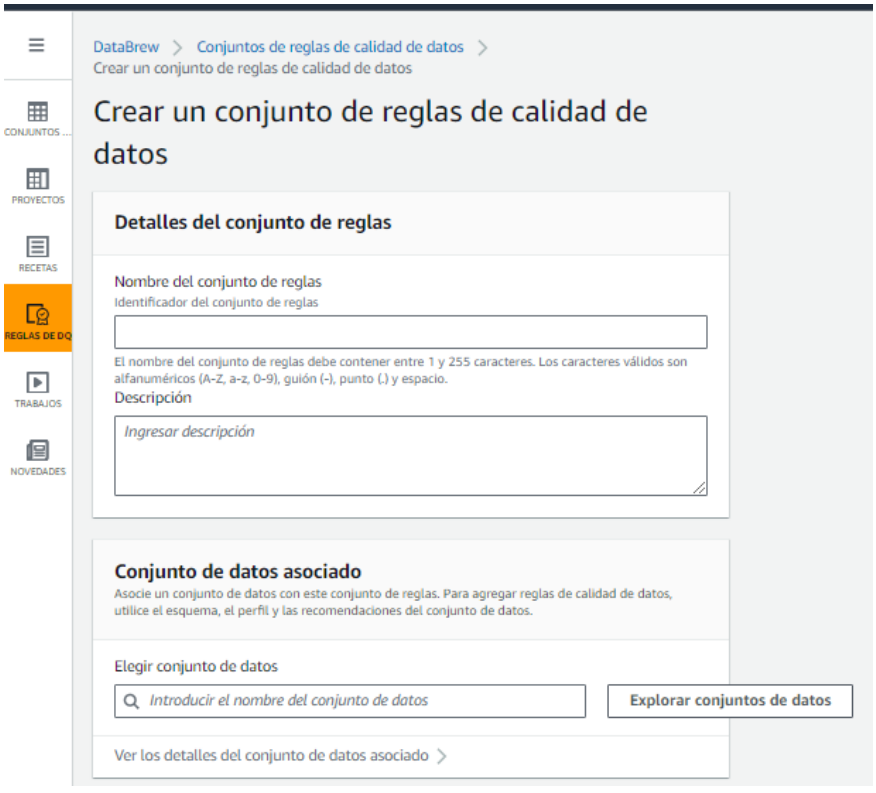


Pasos de una arquitectura de comprobación de la calidad de los datos en AWS Glue DataBrew

Veamos paso a paso como crear reglas de calidad de datos con Databrew.

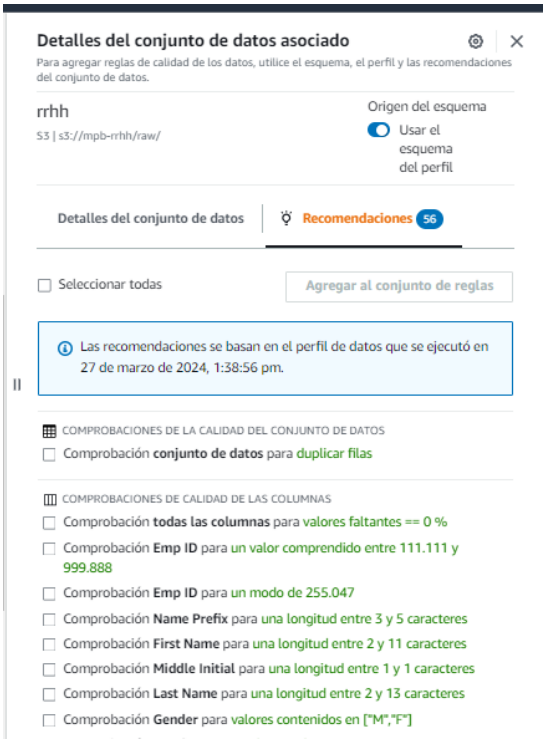
Para crear reglas de calidad de datos, siga los pasos que se mencionan a continuación:

- 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.



Creación de un conjunto de reglas de calidad de datos

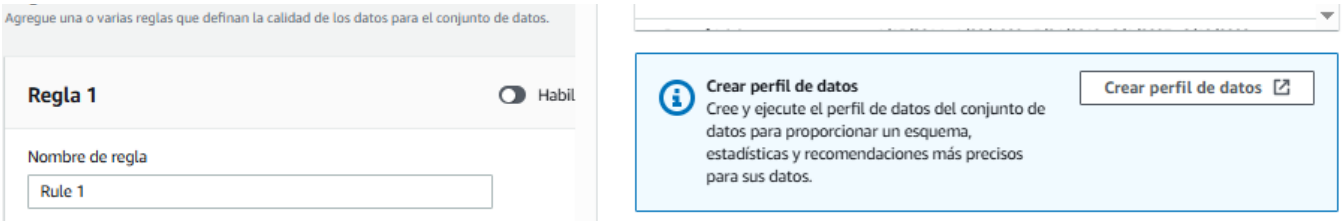
El propio Databrew una vez seleccionado del conjunto de datos nos ofrece un conjunto de comprobaciones a realizar para verificar la calidad (a través del enlace *Ver los detalles del conjunto de datos asociado*)



Por ejemplo, haremos las siguientes verificaciones:

- Comprobar que ciertos campos clave no tienen valores repetidos
- Asegurarse de que no faltan valores en ciertas columnas clave

Lo primero es crear un **perfil de datos.** Para ello seguimos los pasos que se indican a continuación:

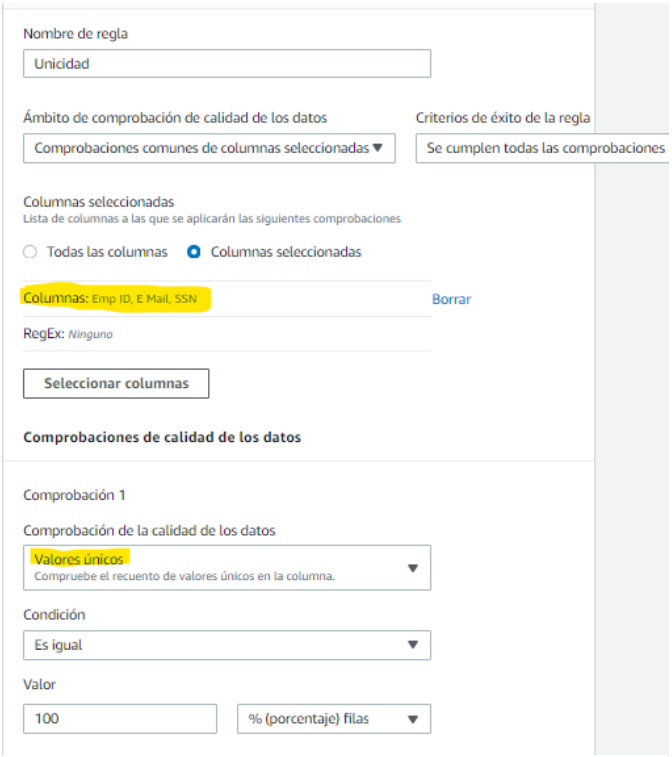


- Proporcionamos un nombre para el trabajo de perfil.
- Seleccionamos un *Conjunto de datos completo* o la opción para *Ejemplo Personalizado*.
- Elegimos la ubicación de S3 en la que se desee almacenar la salida del perfil en la sección Configuración de salida del trabajo.
- Podemos dejar las opciones restantes como predeterminadas.
- En la sección Permisos, acordarse de seleccionar LabRole en el caso de AWS Academy.
- Por último, hacemos clic en *Crear y ejecutar trabajo* para crear el perfil de datos. Tardará algún tiempo en generarse el perfil de datos.
- De las recomendaciones de calidad de datos generadas en la consola de AWS Glue DataBrew puede que no todas las recomendaciones sean útiles, es aconsejable revisarlas todas y añadir las pertinentes al conjunto de reglas.

Para crear reglas de calidad de datos personalizadas mediante la consola Glue DataBrew, primero debemos definir qué reglas queremos crear. Estas son algunas de las reglas que queremos crear

Comprobar que ciertos campos clave no tienen valores repetidos

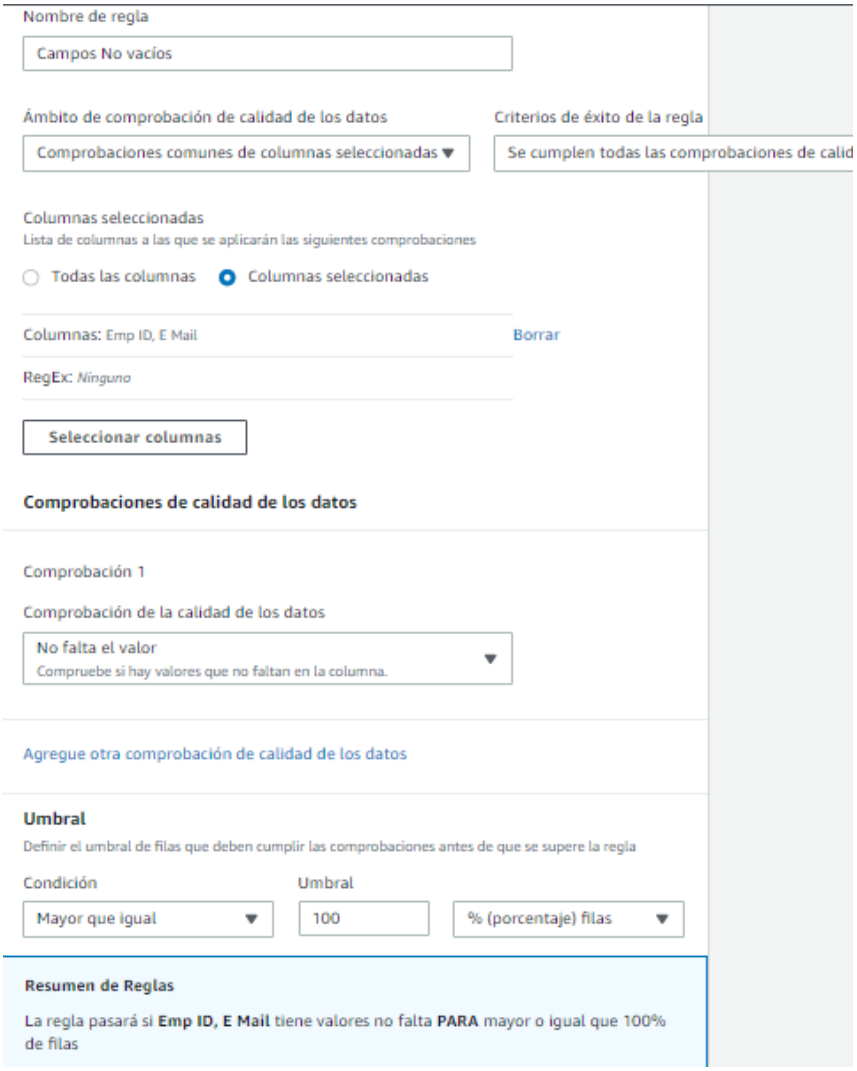
- Seleccionamos Añadir otra regla.
- Se introduce un nombre para la regla, como Comprobar ID de empleado y correo electrónico nulos.
- Seleccionamos *Comprobaciones comunes para columnas seleccionadas* para *Ámbito de la comprobación de calidad de datos*.
- Seleccionamos *Se cumplen todas las comprobaciones de calidad de datos (AND)* en *Criterios de éxito de la regla*.
- Seleccione ID de empleado y Correo electrónico en *Columnas seleccionadas*.
- En *Comprobación de calidad de datos*, seleccionamos Valores único.
- Seleccionamos Mayor que o igual que para *Condición*.



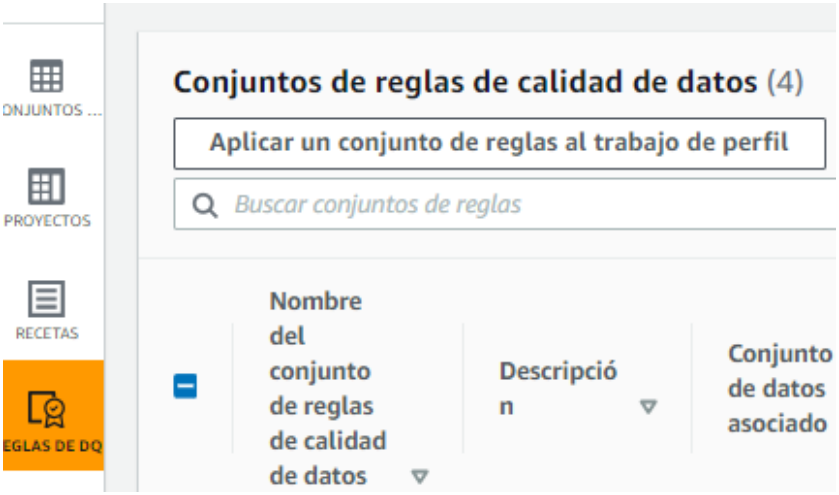
- Introducimos 100 como valor umbral y seleccione %(porcentaje) filas en el menú desplegable.

Asegurarse de que no faltan valores en ciertas columnas clave

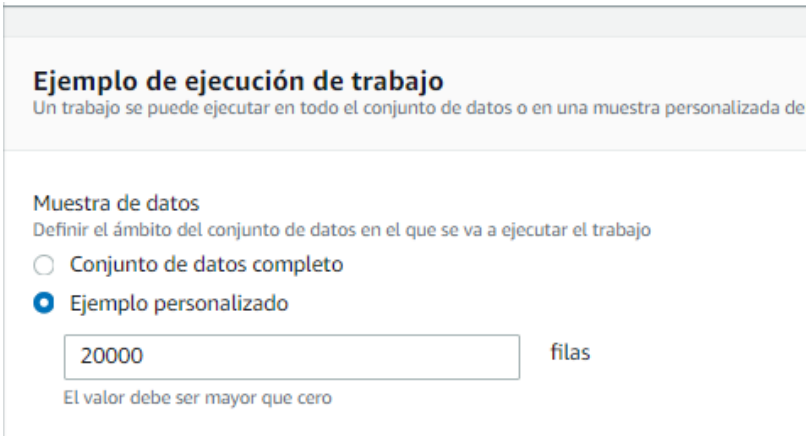
Siguiendo pasos similares al ejemplo anterior, configuraríamos esta regla:



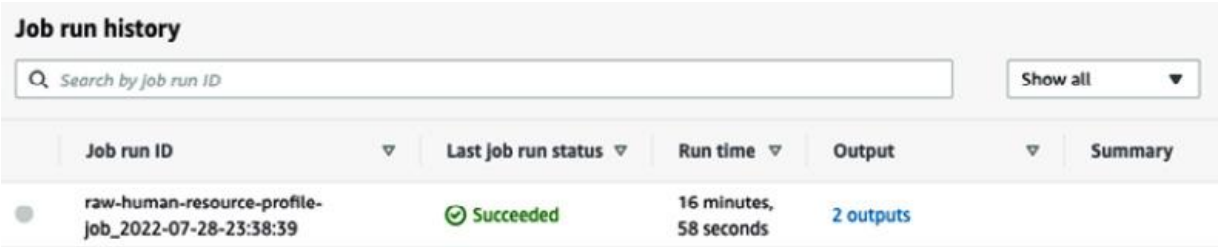
El siguiente paso consiste en ejecutar comprobaciones de calidad de datos aplicando el conjunto de reglas al trabajo de perfil creado anteriormente.



Hacemos clic en *Ejecutar trabajo* y aparecerá una nueva ventana emergente. Es importante decidir si ejecutar el trabajo contra el conjunto de datos completo o contra una muestra del cliente, dependiendo del caso de uso. Si se trata de una aplicación de misión crítica, se recomienda seleccionar el conjunto de datos completo.

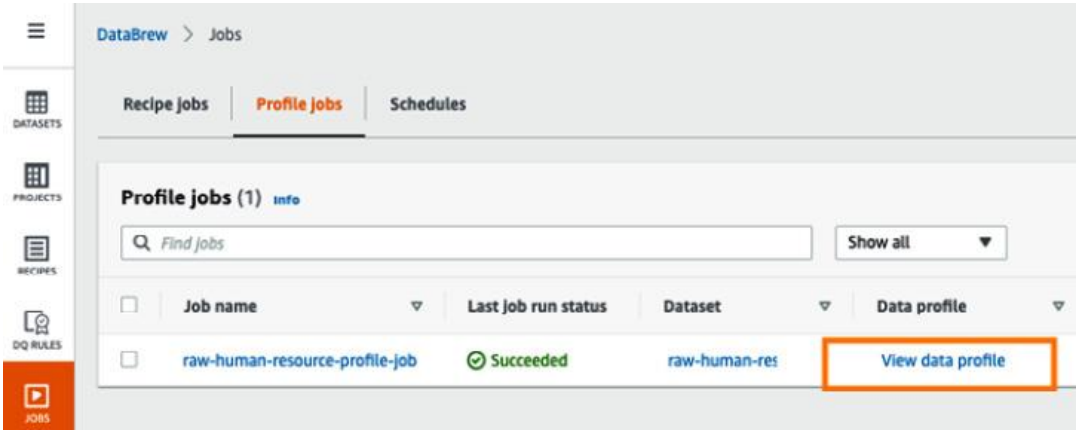


Tras esperar a que finalice el trabajo, que suele tardar varios minutos, el siguiente paso consiste en inspeccionar los resultados de la validación de las reglas de calidad de datos.

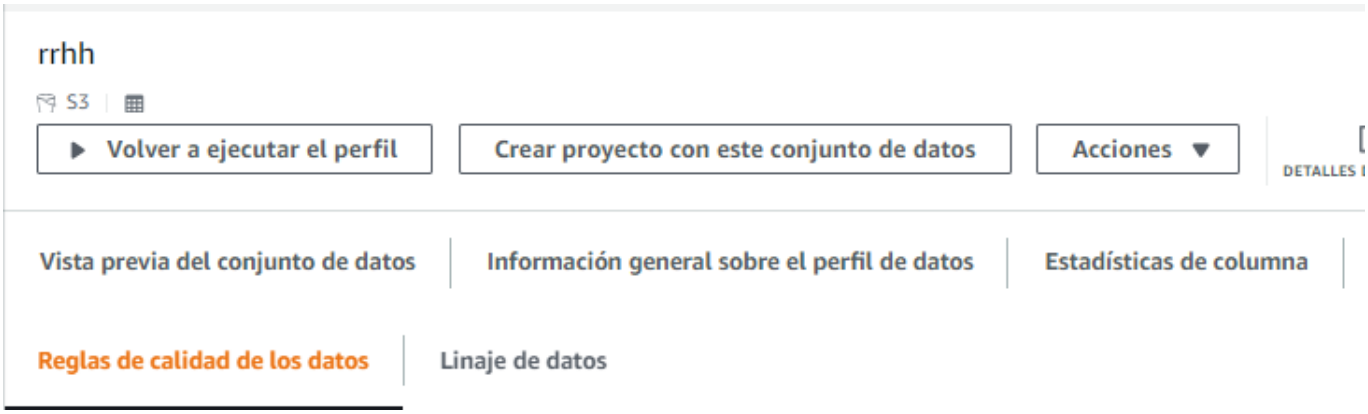


Para ello, seguimos estos pasos:

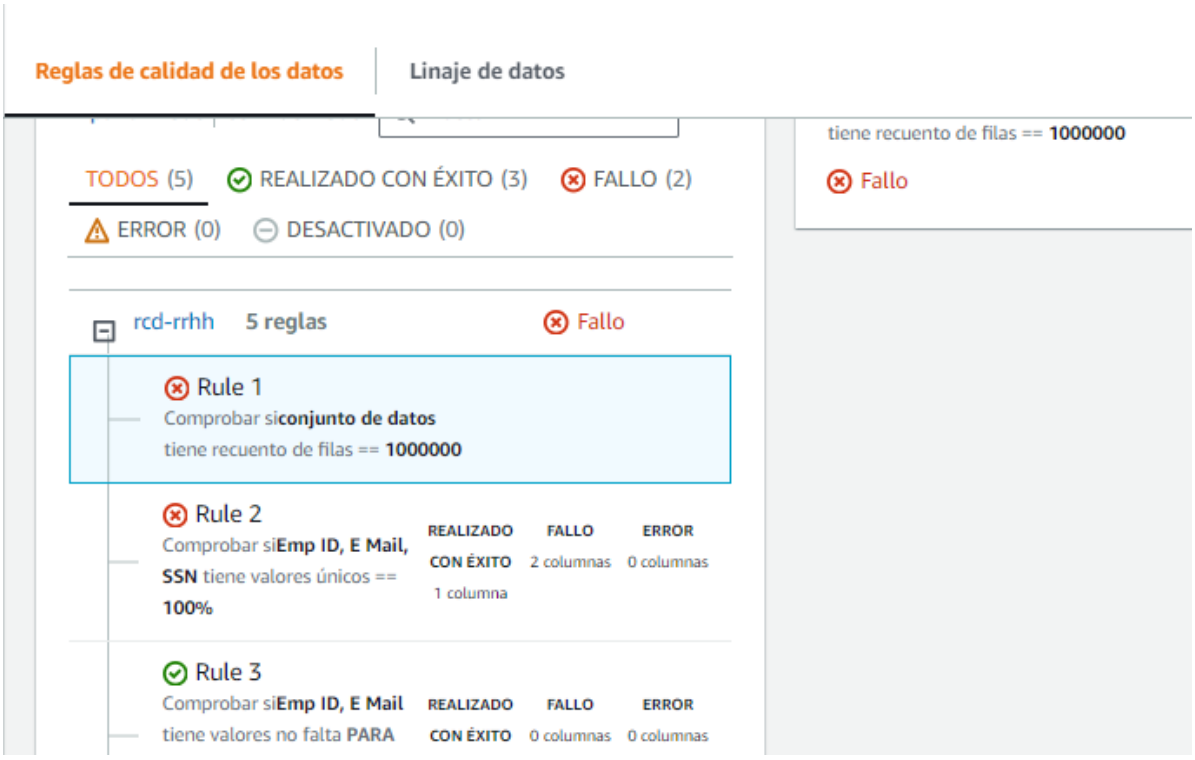
- Vamos a la página Trabajos de la consola DataBrew y seleccione la pestaña *Trabajos de perfil*.
- Espere a que el estado del trabajo de perfil cambie a *Realizado*.
- Una vez completado el trabajo, seleccionamos *Ver perfil de datos*.



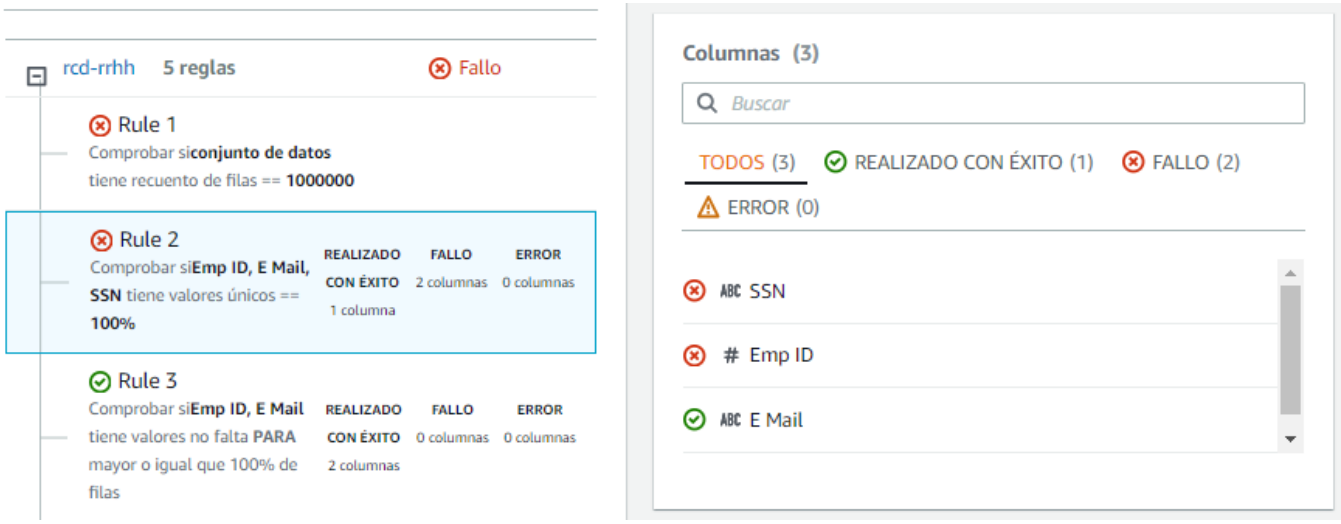
Hacemos clic en la pestaña *Reglas de calidad de datos* para ver el estado de todas las comprobaciones de calidad de datos.



En esta plataforma, se puede ver el estado de todas sus comprobaciones de calidad de datos.

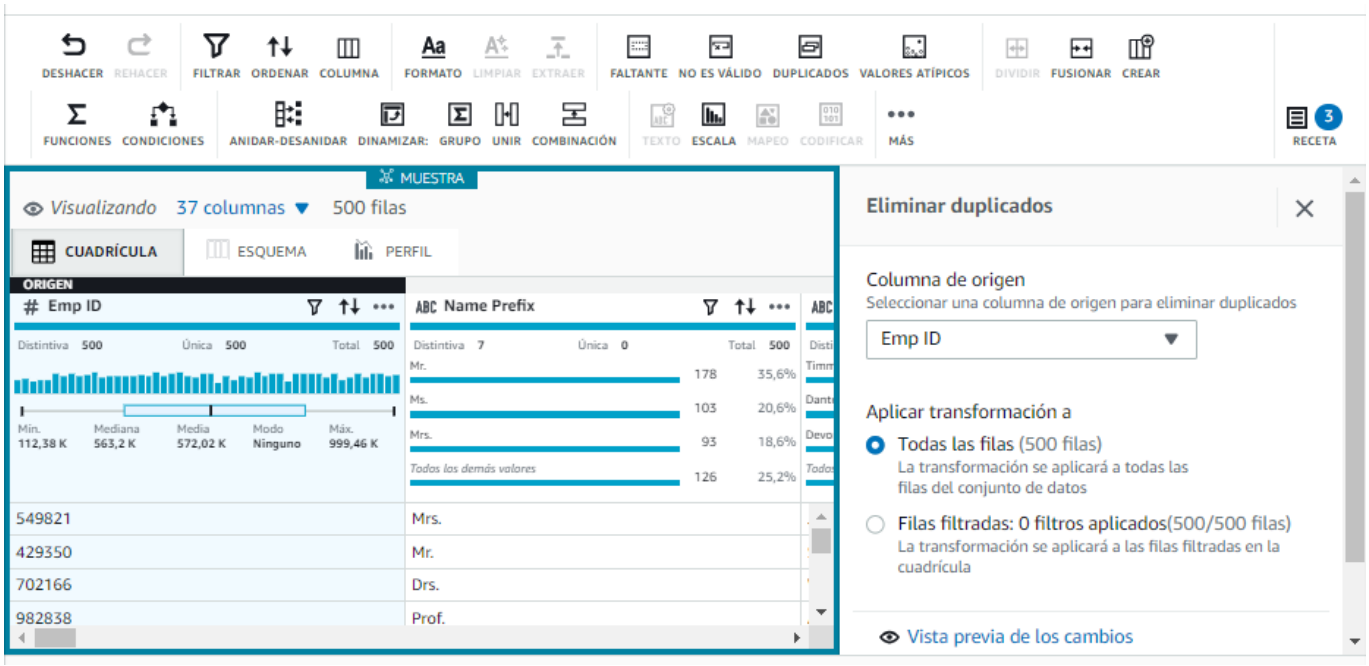


Por ejemplo, en nuestro caso, vemos que el Id de empleado y el número de la seguridad social aparece con duplicados, no así el correo electrónico.



El siguiente paso es abordar los problemas de calidad de los datos. Para ello, el primer paso es crear un proyecto y aplicar las transformaciones pertinentes que nos permitan corregir los errores detectados:

- Por ejemplo, seleccionando la columna *Emp ID*.
- Haciendo clic en el icono *Más opciones* (tres puntos) para ver todas las transformaciones disponibles para esta columna, y seleccionando *Eliminar duplicados*.



Opción Eliminar duplicados

Realizaríamos pasos similares para el resto de los errores detectados al aplicar las reglas de calidad de los datos.