

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.**

Resumen

Destino s3://pablomr-hr1m	Realizado correctam ✓ 1 archivo, 265.2 M
------------------------------	---

Archivos y carpetas Configuración

Archivos y carpetas (1 total, 265.2 MB)

Nombre	Carpeta	Tipo
Hr1m.csv	-	text/csv

- 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**

Crear proyecto de muestra

(resolution.csv, states.csv y votes.csv). (resolution.csv, states.csv y votes.csv).

- Votaciones de la Asamblea General de las Naciones Unidas: Votos
votes.csv | Valores separados por comas (CSV) file | 34,9 MiB
Todas las resoluciones documentadas de los votos de la Asamblea General de las Naciones Unidas desde su creación en 1946. El archivo de resolución contiene campos para un resumen anual de los registros de voto miembro-estado con puntuaciones de afinidad y una estimación de punto ideal en relación con los Estados Unidos. Este archivo es el tercero de los tres archivos (resolution.csv, states.csv y votes.csv).
- Colección del Metropolitan Museum of Art
dataset-met-objects.json | JSON file | 6,6 MiB
El conjunto de datos del Museo Metropolitano de Arte contiene información sobre más de 470 000 obras de arte de su colección para uso comercial y no comercial sin restricciones.
- Nombres populares del año 2020
dataset-national-baby-names.json | JSON file | 3,7 MiB
Nombres de bebé populares en 2020 en los Estados Unidos con registros de seguimiento de nombre, sexo y número de incidencias del nombre.
- Datos de descubrimiento de fármacos de ChEMBL
chembl-27.parquet | Parquet file | 2,2 MiB
ChEMBL es una base de datos administrada de manera manual de moléculas bioactivas con propiedades similares a los fármacos. Aúna datos químicos, de bioactividad y genómicos para ayudar a traducir la información genómica en nuevos fármacos eficaces.
- Movimientos de partidas de ajedrez famosos
chess-games.xlsx | Microsoft Excel file | 4,4 MiB
Toda la información disponible sobre 20 000 partidas de ajedrez y la cantidad de factores meta (ajenos al juego) que afectan a una partida.

Nombre del rol
Elija el rol que tiene acceso para conectarse a los datos. Actualice para ver las últimas actualizaciones.
LabRole

Cancelar **Crear proyecto**

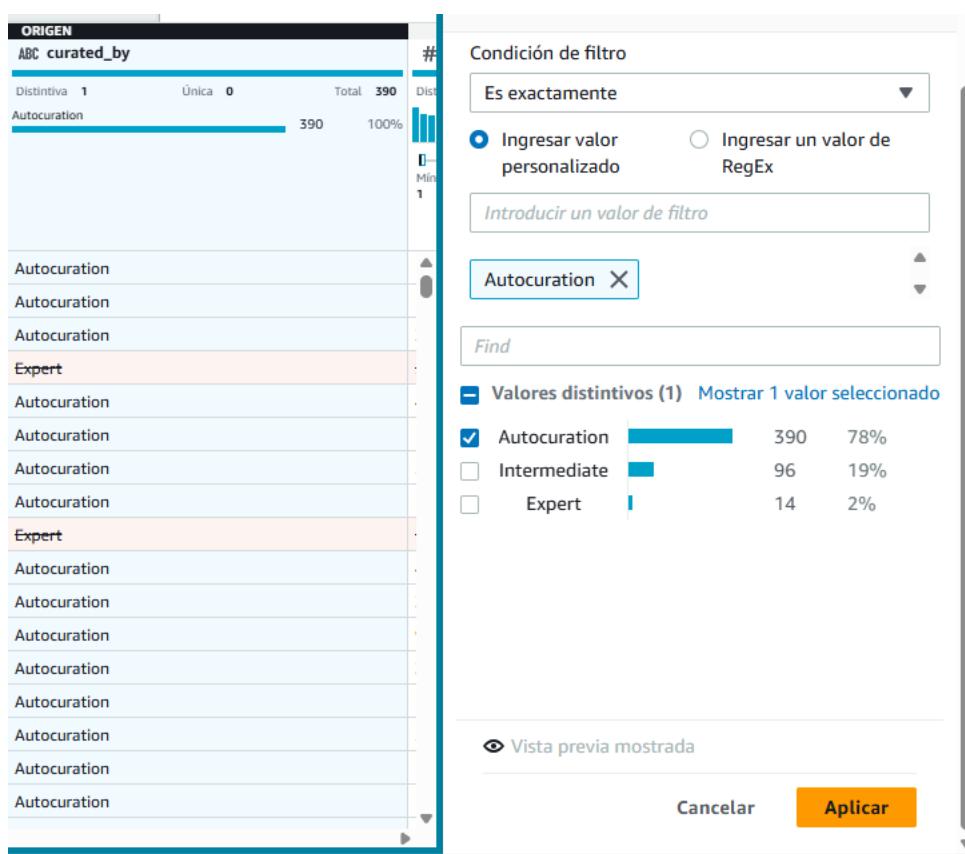
- 2) Creando la receta**

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.

The screenshot shows the Databrew interface with two main windows. On the left is the 'CUADRÍCULA' (Grid) view, which displays a single column named '# tid_fixed' with 1 distinct value ('NaN') and 0 unique values. Below this, there are summary statistics: Min. (null), Media (0), Modo (Ninguno), and Máx. (null). The data preview below shows 15 rows, all of which are 'null'. On the right is the 'Eliminar columna' (Delete column) dialog box. It has a dropdown menu 'Nombre de la columna' (Column name) set to '# tid_fixed'. At the bottom are 'Cancelar' (Cancel) and 'Aplicar' (Apply) buttons. A preview section at the bottom of the dialog shows the first few rows of the dataset with red highlights on the 'null' entries, indicating they are selected for deletion.

- 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:

The 'Valores de filtro' dialog is shown for the 'assay_organism' column. The 'Condición de filtro' (Filter Condition) is set to 'Es exactamente' (Is exactly). The input field contains 'assay_organism'. In the 'Seleccionar valores personalizados' (Select personalized values) section, 'FALTANTE' is checked. The summary table shows 'Valores distintivos (1)' (Distinct values (1)) with one value selected: 'null' (52, 100%).

At the bottom, it says 'Filtrado 1/81 valores' (Filtered 1/81 values) and 'Resultados 52 filas' (Results 52 rows). Buttons for 'Borrar filtro' (Delete filter) and 'Aplicar como condición de filtro' (Apply as filter condition) are at the bottom right.

Valores faltantes

Acción de valor faltante
Acción que se debe realizar en los valores que faltan

- Eliminar filas con valores faltantes
- Rellenar con un valor vacío
- Rellenar con nulo
- Rellenar con el último valor válido
- Rellenar con el valor más frecuente
- Rellenar con valor personalizado
- Rellenar con agregación numérica

Valor personalizado
Unknown

Aplicar transformación a

- Todas las filas (500 filas)
La transformación se aplicará a todas las filas del conjunto de datos
- Filas filtradas: 1 filtros aplicados(52/500 filas)
La transformación se aplicará a las filas filtradas en la cuadrícula

Agregar condición de filtro
Filtrar assay_organism por IS

Vista previa de los cambios

Cancelar Aplicar

- 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

VISTA PREVIA
ABC updated_on_MONTHNAME

Distinct	Unique	Total	Distinct	Unique	Total
312	280	422	13	0	422
null	78	15,6%	null	78	15,6%
2015-02-14 14:00:47	41	8,2%	February	69	13,8%
2004-10-31 12:48:03	13	2,6%	May	66	13,2%
Todos los demás valores	368	73,6%	Todos los demás valores	287	57,4%

columna de origen o un valor de entrada.

Valores que utilizan

Columna de origen

Columna de origen
updated_on

Columna de destino
Nombre de la columna creada con valores extraídos
updated_on_MONTHNAME

Los caracteres válidos son alfanuméricos, guiones bajos y espacios

Aplicar transformación a

- Todas las filas (500 filas)
La transformación se aplicará a todas las filas del conjunto de datos
- Filas filtradas: 0 filtros aplicados(500/500 filas)
La transformación se aplicará a las filas filtradas en la cuadrícula

Vista previa mostrada

Cancelar Aplicar

- 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.

The screenshot shows a 'Receta (4)' (Recipe (4)) window. At the top, there's a title 'Sample recipe - 1' and a subtitle 'Versión de trabajo' (Working version). To the right are 'Publicar' (Publish) and 'Más' (More) buttons. Below the title, there's a section titled 'Pasos aplicados (4)' (Applied steps (4)) with a 'Borrar todo' (Delete all) link. A list of four steps is shown:

1. Eliminar columna `tid_fixed`
2. Rellenar valores faltantes con `Unkown` en `assay_organism`
3. Crear columna `updated_on_MONTHNAME` uso de Función `dateTime MONTH_NAME`
4. Valores de filtro por `curated_by`

[La publicamos](#)

The screenshot shows a 'Publicar receta' (Publish recipe) dialog box. It contains the following fields:

- Nombre de la receta:** Sample recipe - 1
- Descripción de la versión:** Primera fase de la tare 8.3
- Pasos de la receta (4):** The same four steps as in the previous screenshot are listed here.

At the bottom, there are 'Cancelar' (Cancel) and 'Publicar' (Publish) buttons, with 'Publicar' being highlighted.

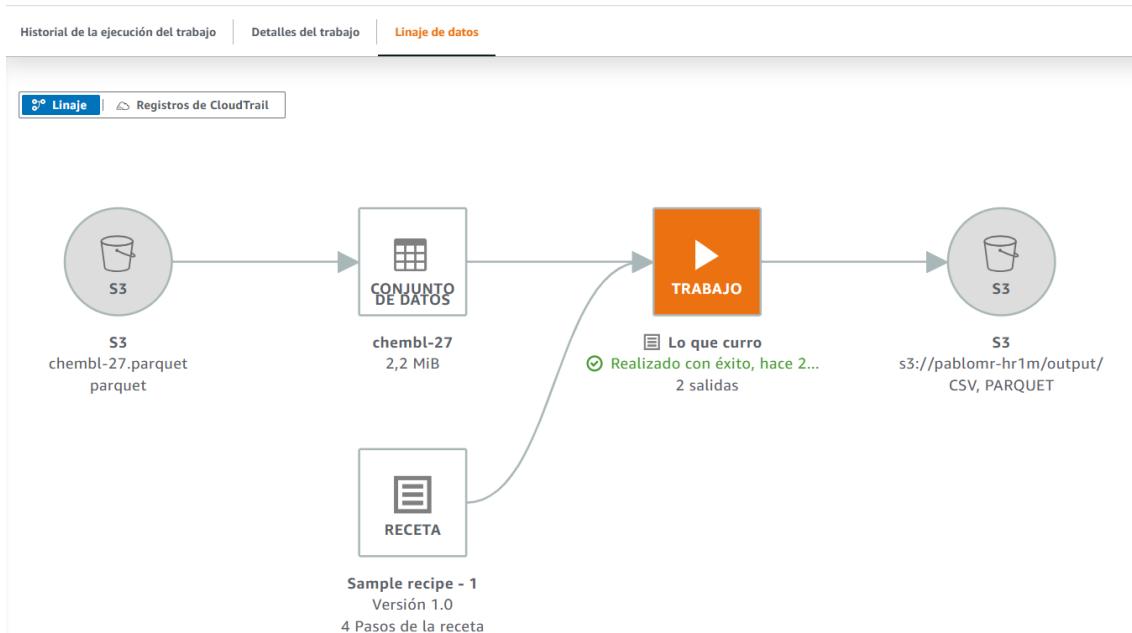
3) De la receta al Job.

- 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

ID de ejecución de trabajo	Estado de la última ejecución del trabajo	Tiempo de ejecución	Salida	Resumen
Lo que curro_2026-01-21-11:35:45	Starting	No está disponible	2 salidas	

En linaje.



4) Calidad de datos con AWS Glue 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.

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

Ámbito de comprobación de calidad de los datos Criterios de éxito de la regla
Comprobación individual de cada columna Se cumplen todas las comprobaciones

Comprobaciones de calidad de los datos

Comprobación 1

Comprobación de la calidad de los datos

Duplicar filas Compruebe el conjunto de datos para el número de filas duplicad...

Condición Menor que igual

Valor 0 filas

Agregue otra comprobación de calidad de los datos

Resumen de Reglas
La regla pasará si **conjunto de datos** tiene recuento de filas duplicadas <= 0

Regla 2 Habilitar regla Eliminar

Nombre de regla Check All Columns For Missing Values

Ámbito de comprobación de calidad de los datos Criterios de éxito de la regla
Comprobaciones comunes de columnas seleccionadas Se cumplen todas las com...

Regla perfectamente creada

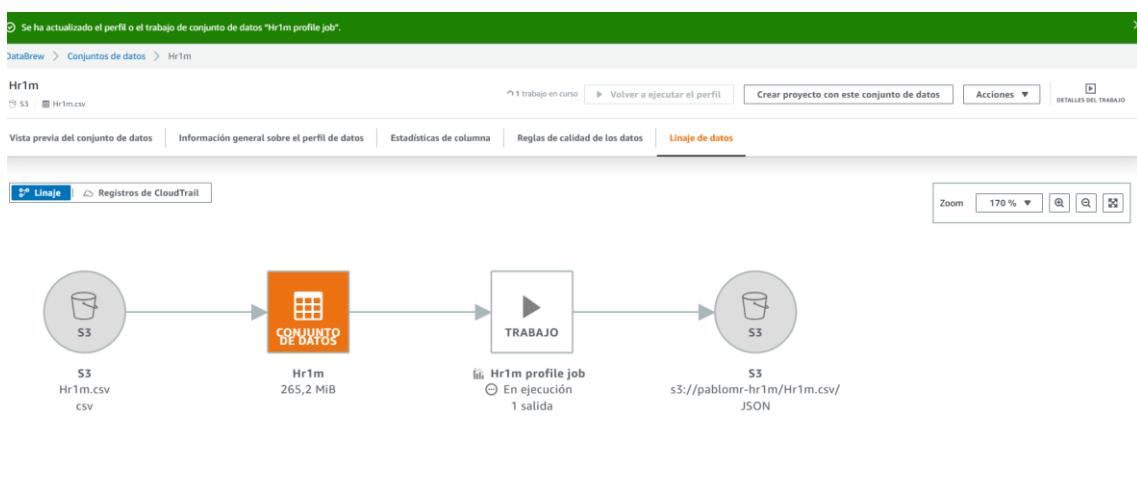
Conjuntos de reglas de calidad de datos (1)

Buscar conjuntos de reglas

<input checked="" type="checkbox"/>	Nombre del conjunto de reglas de calidad de datos	Descripción	Conjunto de datos asociado
<input checked="" type="checkbox"/>	regla nueva 2 reglas	-	Hr1m

Creamos un trabajo y lo ejecutamos

DATABREW I



Los pasó todos, están muy bien hechos

Última ejecución de trabajo Realizado con éxito hace 8 minutos, no hay ejecuciones de trabajos programadas
El perfil de datos se ha ejecutado en muestra personalizada de las primeras 20.000 filas de su conjunto de datos

Reglas de calidad de los datos (2)

Expandir todo | Contrair todo | Buscar

TODOS (2) REALIZADO CON ÉXITO (2) FALLO (0) ERROR (0) DESACTIVADO (0)

regla nueva 2 reglas Realizado con éxito

Check Dataset For Duplicate Rows
Comprobar si **conjunto de datos** tiene recuento de filas duplicadas <= 0

Check All Columns For Missing Values
Comprobar si **todas las columnas** tiene valores faltantes == 0%

REALIZADO CON ÉXITO	FALLO	ERROR
37	0 columnas	0 columnas