

# 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://pablmr-hr1m

Realizado correctamente  
1 archivo, 265.2 MB

## Archivos y carpetas

## Configuración

Archivos y carpetas (1 total, 265.2 MB)

Buscar por nombre

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.

☒ 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.

☐ 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.

☐ 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

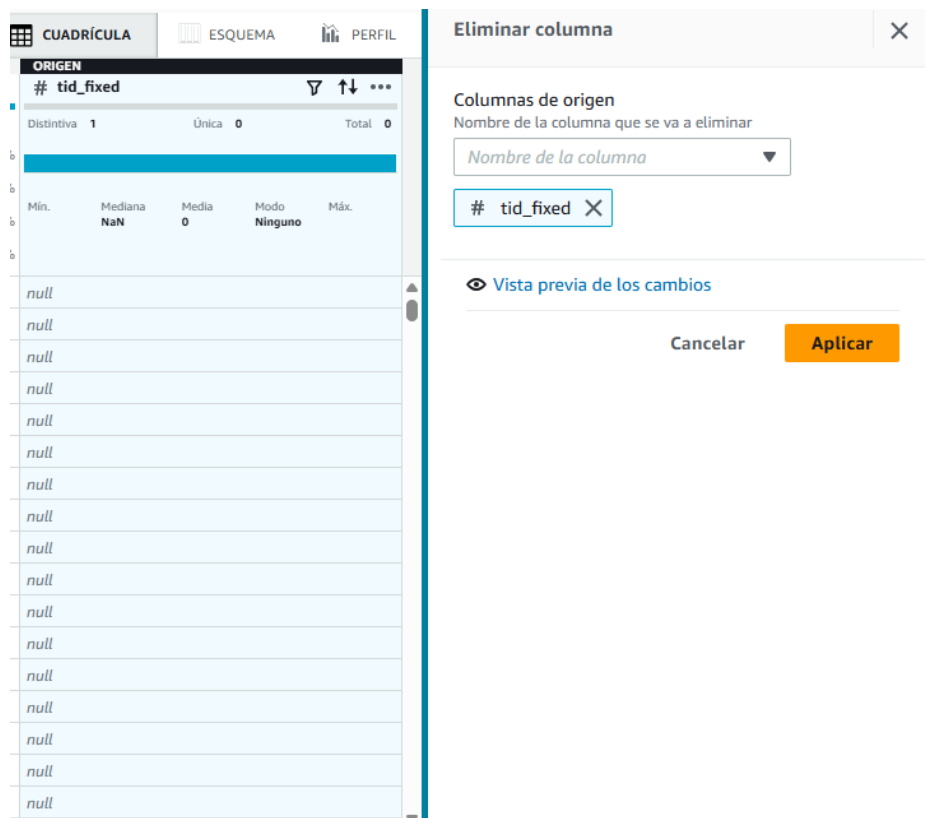
Cancelar

Crear proyecto

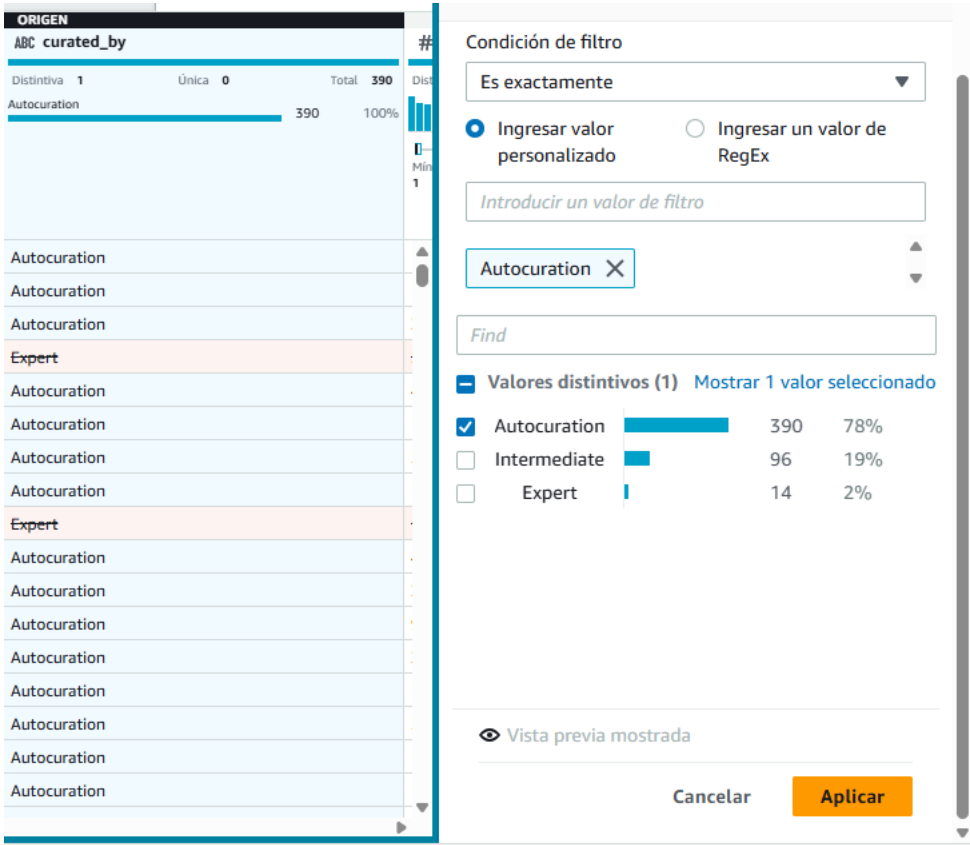
- 2) Creando la receta

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.



- 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 valores 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:

Valores de filtro

Columna de origen  
assay\_organism  
Condición de filtro  
Es exactamente  
☒ Ingresar valor personalizado  
☐ Ingresar un valor de RegEx  
Introducir un valor de filtro  

X

Seleccionar valores personalizados

☐ VÁLIDO (500) ☒ FALTANTE (52)

Find

☒ Valores distintivos (1) [Mostrar 1 valor seleccionado](#)

null

52100%

Filtrado 1/81 valoresResultados 52 filas

Borrar filtro

Aplicar como condición de filtro

**CUADRÍCULA** **ESQUEMA** **PERFIL**

**ORIGEN**

ABC assay\_organism

Distintiva	Única	Total	Porcentaje
81	55	448	40,6%
Homo sapiens		203	40,6%
Rattus norvegicus		62	12,4%
Mus musculus		57	11,4%
Todos los demás valores		178	35,6%

**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

**ORIGEN** **VISTA PREVIA**

ABC updated\_on

Distintiva	Única	Total	Porcentaje
312	280	422	78,2%
2015-02-14 14:00:47		41	8,2%
2004-10-31 12:48:03		13	2,6%
Todos los demás valores		368	73,6%

**ABC updated\_on\_MONTHNAME**

Distintiva	Única	Total	Porcentaje
13	0	422	3,1%
February		69	13,8%
May		66	13,2%
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.

**Receta (4)**

**Sample recipe - 1**  
Versión de trabajo

**Publicar**

**Más**

---

**Pasos aplicados (4)** | [Borrar todo](#)

---

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

**Publicar receta**

La publicación de una receta creará una nueva versión de la receta. Las versiones publicadas de las recetas se pueden seleccionar como opciones para un trabajo de receta.

Nombre de la receta  
**Sample recipe - 1**

Descripción de la versión

Primera fase de la tare 8.3

Pasos de la receta (4)

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**

Cancelar

Publicar

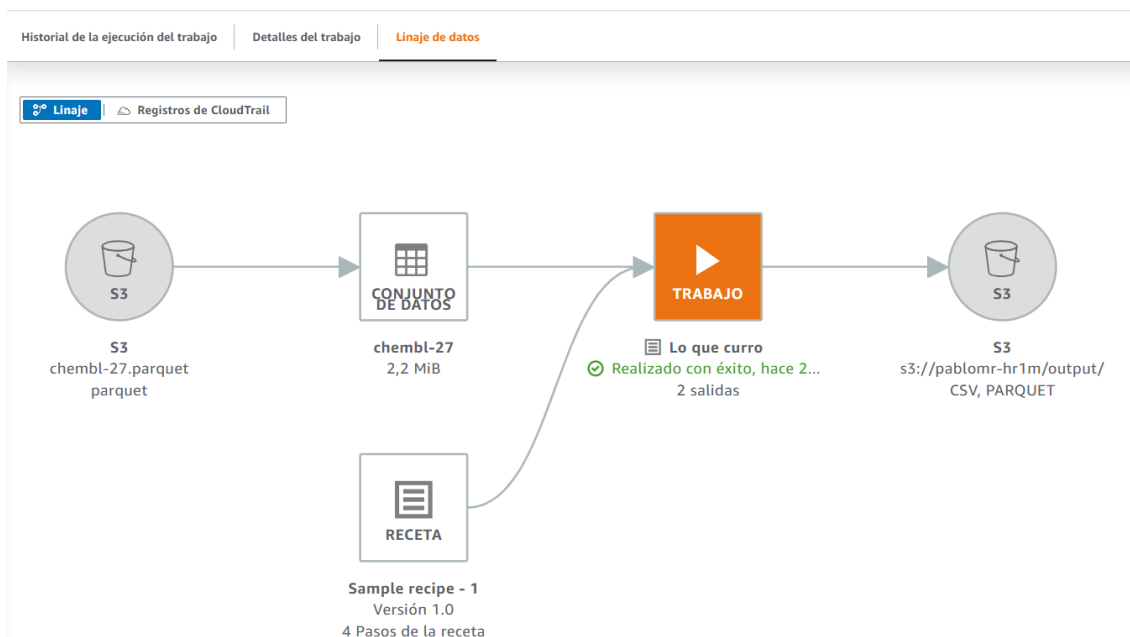
## 3) De la receta al Job.

- 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

Historial de la ejecución del trabajo				
<input type="text" value="Buscar por ID de ejecución de trabajo"/>				Mostrar todo
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

- 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.

Ámbito de comprobación de calidad de los datos

Comprobación individual de cada columna ▼

Criterios de éxito de la regla

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 duplicadas...

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

Comprobaciones comunes de columnas seleccionadas ▼

Criterios de éxito de la regla

Se cumplen todas las comprobaciones

### Regla perfectamente creada

Conjuntos de reglas de calidad de datos (1)		
<input type="text" value="Buscar conjuntos de reglas"/>		
<input checked="" type="checkbox"/>	Nombre del conjunto de reglas de calidad de datos ▾	Descripción ▾
<input checked="" type="checkbox"/>	regla nueva 2 reglas	-

## Creamos un trabajo y lo ejecutamos

Se ha actualizado el perfil o el trabajo de conjunto de datos "Hr1m profile job".

Databrew > Conjuntos de datos > Hr1m

Hr1m

1 trabajo en curso ▶ Volver a ejecutar el perfil Crear proyecto con este conjunto de datos Acciones

Vista previa del conjunto de datos Información general sobre el perfil de datos Estadísticas de columna Reglas de calidad de los datos **Linaje de datos**

Linaje Registros de CloudTrail

Zoom 170 %

```

graph LR
    S3_1[S3 Hr1m.csv csv] --> CD[CONJUNTO DE DATOS Hr1m 265,2 MiB]
    CD --> T[TRABAJO Hr1m profile job En ejecución 1 salida]
    T --> S3_2[S3 s3://pablomr-hr1m/Hr1m.csv/JSON]
    
```

Los pasó todos, están muy bien hechos

Vista previa del conjunto de datos Información general sobre el perfil de datos Estadísticas de columna **Reglas de calidad de los datos** Linaje de datos

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 Contraer todo

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

regla nueva 2 reglas **Realizado con éxito**

Regla	REALIZADO CON ÉXITO	FALLO	ERROR
<b>Check Dataset For Duplicate Rows</b> Comprobar si <b>conjunto de datos</b> tiene recuento de filas duplicadas <= 0	37	0 columnas	0 columnas
<b>Check All Columns For Missing Values</b> Comprobar si <b>todas las columnas</b> tiene valores faltantes == 0%	37 columnas	0 columnas	0 columnas

**Check Dataset For Duplicate Row**  
Comprobar si **conjunto de datos** tiene recuento de

**Realizado con éxito**