

# **PR\_08.3 Dani Gayol Rodríguez**

PR_08.3 Dani Gayol Rodríguez.....	1
Apartado A.....	1
1.) Realiza el ejemplo de Proyecto Databrew que se muestra en el tema UT_08.4_AWS DataBrew entre las páginas 1 y 7.....	2
2,) Muestra también el linaje resultante de los datos. .....	16

# Apartado A

1.) Realiza el ejemplo de Proyecto Databrew que se muestra en el tema UT\_08.4\_AWS DataBrew entre las páginas 1 y 7.

El primer paso va a ser crear el proyecto de muestra con los datos de CHEMBL

The image consists of two screenshots from the AWS website. The top screenshot shows the search results for 'AWS Glue Databrew'. The search bar contains 'AWS Glue Databrew'. Below it, under the heading 'Servicios', there is a card for 'AWS Glue DataBrew' which is highlighted with a red box. The card describes it as a visual tool for preparing data. The bottom screenshot shows the 'AWS Glue DataBrew' landing page. It features the title 'AWS Glue DataBrew' and a sub-headline 'Limpie y normalice datos hasta un 80 % más rápido'. Below this is a detailed description of the service. To the right, there is a 'Crear proyecto' section with a button labeled 'Crear proyecto' and another section with a button labeled 'Crear proyecto de muestra', which is also highlighted with a red box.

## 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



Cerrar

**Crear proyecto**

## Una vez creado el proyecto, nos manda a interfaz del Databrew:

The screenshot shows the Databrew interface with the following details:

- Top Bar:** Includes "Crear trabajo", "LUAJE", "ACCIONES", and a "RECETA" button.
- Left Sidebar:** Shows sections for "PROYECTOS", "RECETAS", "REGLES DE DATO", and "TRABAJOS".
- Data Preview:** A large table titled "Muestra" showing 39 columns and 500 rows. The columns include "# assay\_id", "# doc\_id", "assay\_description", "assay\_type", and "assay\_test\_type". The data includes various biological activity records like "Antifungal activity against Candida glabrata isol..." and "Antimicrobial activity against Listeria monocytogenes".
- Right Side:** A "RECETA" panel titled "Sample recipe - 1" with a sub-section "Cree su receta" containing a "Agregar paso" button.

Ahora vamos a seguir con los pasos que nos indica la práctica, el primer paso es eliminar la columna “tid\_fixed”

The screenshot illustrates the process of deleting the 'tid\_fixed' column from a dataset. On the top right, the 'COLUMNAS' (Columns) icon is highlighted with a red box. A context menu is open, showing options like 'Cambiar nombre' (Change name), 'Duplicar' (Duplicate), 'Tipo de cambio' (Type conversion), 'Trasladar columna' (Move column), and 'Eliminar' (Delete). The 'Eliminar' option is also highlighted with a red box. In the main workspace, the 'CUADRÍCULA' (Grid) view shows the 'ORIGEN' (Origin) section. The '# tid\_fixed' column is selected, indicated by a red border around its header. The 'ESQUEMA' (Schema) tab is visible above the grid. On the right, a 'Eliminar columna' (Delete Column) dialog box is open, showing a list of columns under 'Columnas de origen'. The '# tid\_fixed' column is listed with its checkbox checked, and the 'Aplicar' (Apply) button is highlighted with a red box.

**Receta (1)**

Sample recipe - 1

Versión de trabajo

Publicar

Más

**El paso Eliminar columna se ha aplicado correctamente.**

Pasos aplicados (1) | Borrar todo

1. Eliminar columna `tid_fixed`

Ahora, vamos a filtrar los datos:

The screenshot shows a data analysis interface with a toolbar at the top and a main content area below.

**Toolbar:**

- DESHACER
- REHACER
- FILTRAR (highlighted with a red box)
- ORDENAR
- COLUMNA
- FORMATO
- LIMPIAR
- EXTRAER
- FALTANTE
- NO ES VÁLIDO
- DUPLOCADOS

**Left Panel (Visualizando 3):**

- ABC curated\_by
  - Distintiva 3
  - Autocuration
  - Intermediate
  - Expert

**Right Panel (Filtros condicionados aplicados):**

No hay filtros aplicados

Agregar nuevo filtro

Valores faltantes

Es válido

Por condición >

Es exactamente (highlighted with a red box)

Other filter options listed on the right:
 

- No es
- Contiene
- No contiene
- Comienza por
- Termina con
- Menor que
- Menor o igual que
- Mayor que
- Mayor o igual que
- Está entre

ABC curated_by	Count	...	ABC ass
Única	13	Total	470
Distintiva			
ORIGINAL			
Autocuration	1		
Autocuration	1		
Autocuration	3		
Expert	23		
Autocuration	48		
Autocuration	1		
Autocuration	22		
Autocuration	1		
Expert	1		
Autocuration	48		

**Valores de filtro**

NOMBRE DE LA COLUMNA QUE SE VA A FILTRAR  
**curated\_by**

CONDICIÓN DE FILTRO  
**Es exactamente**

Ingresar valor personalizado       Ingresar un valor de RegEx

*Introducir un valor de filtro*

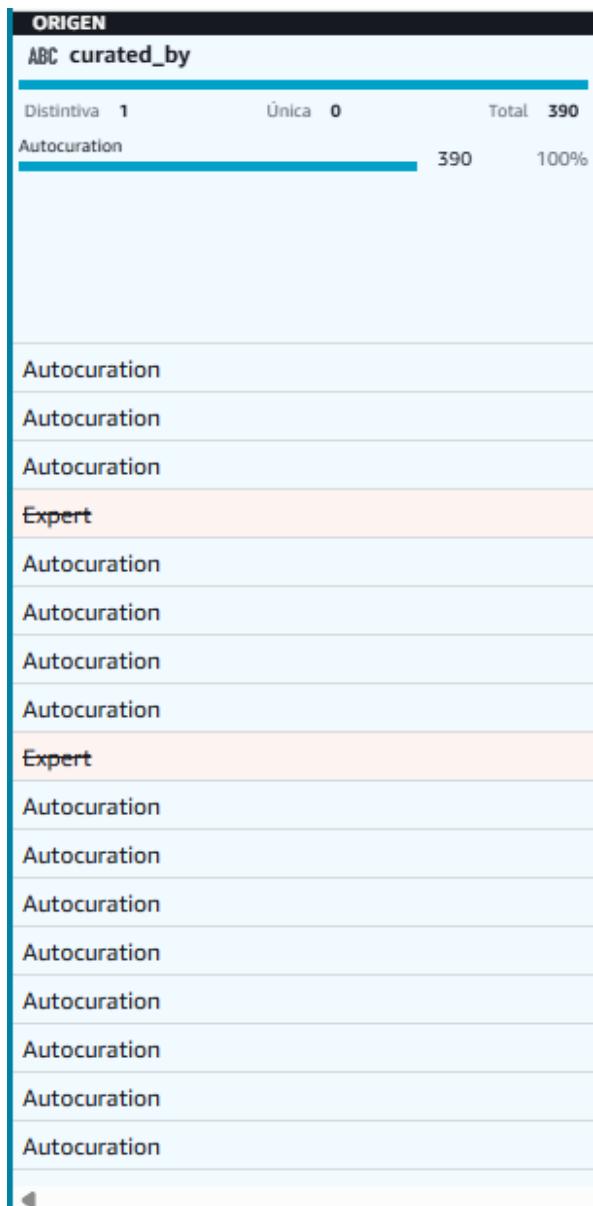
**Autocuration X**

*Find*

**- Valores distintivos (1) Mostrar 1 valor seleccionado**

<input checked="" type="checkbox"/> Autocuration	390	78%
<input type="checkbox"/> Intermediate	96	19%
<input type="checkbox"/> Expert	14	2%

**Vista previa de los cambios**



**El siguiente paso es llenar los valores nulos:**

The screenshot shows the Data Studio interface with various tools and a context menu open. The 'FALTANTE' button in the top right toolbar is highlighted with a red box. A context menu is open on the right side, listing options for handling missing values: 'Eliminar filas faltantes', 'Rellenar con valor personalizado' (which is also highlighted with a red box), 'Rellenar con un valor vacío', 'Rellenar con nulo', 'Rellenar con el valor más frecuente', 'Rellenar con el último valor válido', and 'Rellenar con agregación numérica'.

**Valores faltantes**

**Columna de origen**  
Nombre de la columna con 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**

**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

ABC assay_organism	▼	↑	...
Distintiva 81	Única 55	Total 500	
Homo sapiens	203	40,6%	
Rattus norvegicus	62	12,4%	
Mus musculus	57	11,4%	
Todos los demás valores	178	35,6%	
 Homo sapiens			
 Mus musculus			
 Homo sapiens			
 Homo sapiens			
 Mus musculus			
 Unknown			
 Homo sapiens			
 Rattus norvegicus			
 Mus musculus			
 Homo sapiens			
 Unknown			
 Rattus norvegicus			
 Unknown			
 Homo sapiens			
 Haemophilus influenzae			
 Homo sapiens			
 Oryctolagus cuniculus			

List icon Receta (2) X

Sample recipe - 1 | Versión de trabajo Publicar ... Más

Pasos aplicados (2) | Borrar todo Print icon Download icon

1. Eliminar columna `tid_fixed`

2. Rellenar valores faltantes con Unknown en `assay_organism`

Ahora, vamos a crear una columna nueva:

The screenshot shows a software interface with a toolbar at the top containing icons for DIVIDIR, FUSIONAR, CREAR, FUNCIONES, CONDICIONES, ANIDAR-DESANIDAR, DINAMIZAR:, GRUPO, UNIR, COMBINACIÓN, TEXTO, and E. The FUNCIONES icon is highlighted with a red box.

A dropdown menu is open under the FUNCIONES icon, showing a search bar with "Buscar" and a list of function categories:

- Funciones matemáticas
- Funciones de agregación
- Funciones de texto
- Funciones de fecha**
- Funciones de ventana
- Funciones web
- Otras funciones
- Expresión personalizada

The "Funciones de fecha" category is also highlighted with a red box. To its right, a list of date-related functions is displayed:

- FECHA
- HORA
- DATETIME
- DATEADD
- DATEDIFF
- DATEFORMAT
- MES
- MONTHNAME**
- AÑO
- TRIMESTRE
- DÍA
- WEEKDAY
- WEEKNUMBER
- HORA
- MINUTO
- SEGUNDO
- MILISEGUNDOS

The word "MONTHNAME" is also highlighted with a red box.

In the main area, there is a chart with a legend and some data points labeled "null".

CUADRÍCULA    ESQUEMA    PERFIL

ORIGEN

	updated_on	Total
Distintiva	312	422
null	78	15,6%
2013-02-14 14:00:47	41	8,2%
2004-10-31 12:48:03	13	2,6%
Todos los demás valores	368	73,6%
2016-12-20 11:12:26		
2009-11-25 15:08:27		
2011-06-09 17:19:30		
null		
2013-02-14 14:00:11		
2010-10-26 17:10:29		
null		
2013-04-08 15:03:54		
2013-12-05 14:18:48		
2016-12-19 13:30:13		
2011-04-08 10:46:23		
2014-01-07 13:09:06		
2014-02-14 00:00:00		
2016-05-09 15:29:11		
2015-09-10 13:00:16		
2012-02-01 12:46:11		
2014-01-07 15:51:20		

### Crear columna

Opciones de creación de columnas

Basado en funciones

Seleccionar una función

Crear una columna basada en la función seleccionada

MONTHNAME

**MONTHNAME**  
Devuelve en una nueva columna el nombre que corresponde al número del mes, en función de una columna de origen o un valor de entrada.

Valores que utilizan

Columna de origen

Columna de origen

updated\_on

Columna de destino

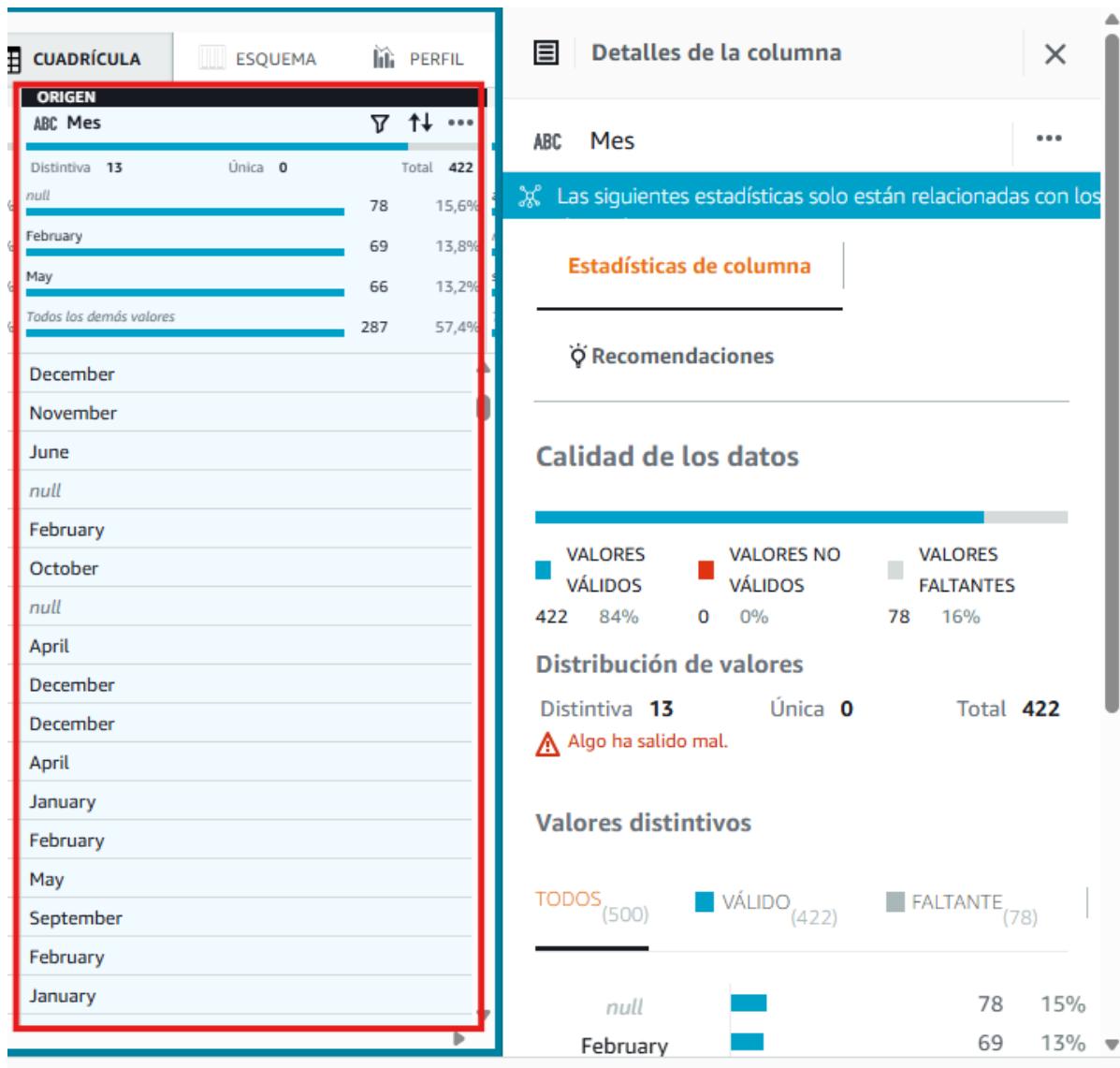
Nombre de la columna creada con valores extraídos

Mes

Los caracteres válidos son alfanuméricos, guiones bajos y espacios

Aplicar transformación a

Todas las filas (500 filas)



El siguiente paso es publicar la receta, para ello le tenemos que dar al botón de publicar y ya nos aparecerá en el menú de la izquierda de “Recetas”

Receta (3)

Sample recipe - 1

Versión de trabajo

Publicar

Pasos aplicados (3) | Borrar todo

1. Eliminar columna `tid_fixed`

2. Rellenar valores faltantes con Unknown en `assay_organism`

3. Crear columna Mes uso de Función `dateTime MONTH_NAME`

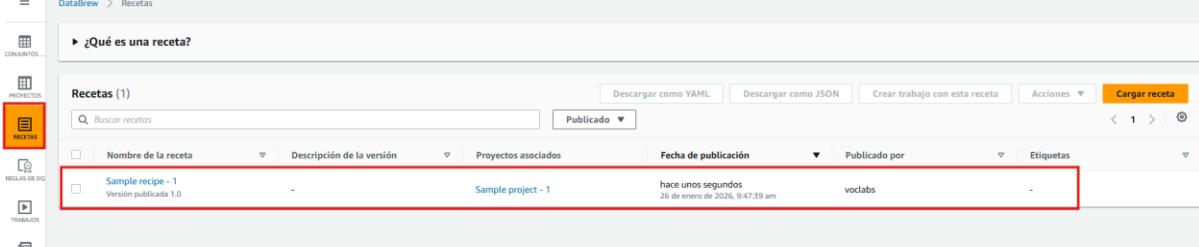


DataBrew > Recetas

¿Qué es una receta?

Recetas (1)

Nombre de la receta	Descripción de la versión	Proyectos asociados	Fecha de publicación	Publicado por	Etiquetas
Sample recipe - 1 Versión publicada 1.0	-	Sample project - 1	hace unos segundos 26 de enero de 2026, 9:47:39 am	vocabs	-



Una vez dentro de la “receta”, le daremos al botón para crear un trabajo:

Cargar receta

Crear trabajo con esta receta

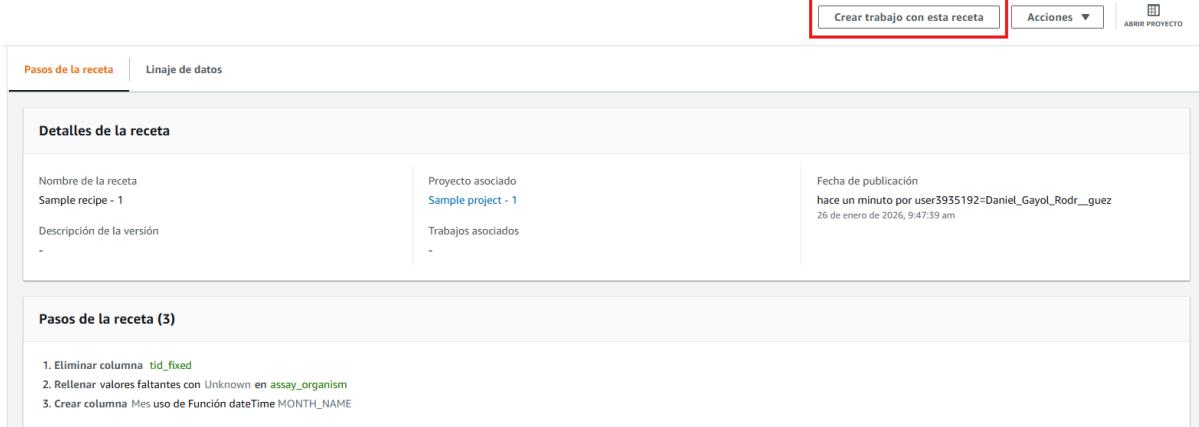
Pasos de la receta | Linaje de datos

Detalles de la receta

Nombre de la receta Sample recipe - 1	Proyecto asociado Sample project - 1	Fecha de publicación hace un minuto por user3935192=Daniel_Gayol_Rodríguez 26 de enero de 2026, 9:47:39 am
Descripción de la versión -	Trabajos asociados -	

Pasos de la receta (3)

1. Eliminar columna `tid_fixed`
2. Rellenar valores faltantes con Unknown en `assay_organism`
3. Crear columna Mes uso de Función `dateTime MONTH_NAME`



## Crear trabajo

### Detalles del trabajo

**Nombre del trabajo**  
Identificador del trabajo  
**DataBrew-danigayol**

El nombre del trabajo debe contener entre 1 y 240 caracteres. Los caracteres válidos son alfanuméricos (A-Z, a-z, 0-9), guiones (-), puntos (.) y espacios.

### Tipo de trabajo

Tipo de trabajo que se va a ejecutar en el conjunto de datos

- Crear un trabajo de receta**  
Ejecuta las transformaciones de la receta asociada en la población del conjunto de datos asociado.
- Crear un trabajo de perfil**  
Genera un resumen y estadísticas que le dan la forma de los datos.

### Entrada del trabajo

El conjunto de datos de entrada para el trabajo y la receta que se le aplicará.

**Ejecutar en**

- Conjunto de datos**  
Ejecute el trabajo en un conjunto de datos DataBrew existente o nuevo.
- Proyecto**  
Ejecute el trabajo en un proyecto sin trabajo asociado.

**Elegir conjunto de datos**

**Seleccionar una receta**  
  Versión 1.0

### Salida 1

<b>Salida a</b> Ubicación de la salida	<b>Tipo de archivo</b> Formato de salida	<b>Delimitador</b> Separador CSV	<b>Compresión</b> Tipos disponibles
<input checked="" type="button" value="Amazon S3"/>	<input checked="" type="button" value="CSV"/>	<input type="button" value="Coma (,)"/>	<input type="button" value="None"/>

Cuenta de AWS del propietario del bucket de S3

- Cuenta de AWS actual  
043356869404
- Otra cuenta de AWS

**Ubicación de S3**  
El formato es s3://bucket/folder/

**Resumen de configuración**  
Almacenamiento de salida de archivos  
Crear una nueva carpeta para cada ejecución de flujo de trabajo  
Salida de archivo  
Autogenerate files  
División personalizada por valores de columna  
Desactivado

**Vista previa de la ruta de salida**  
s3://databrew-practica-danigayol/DataBrew-danigayol\_26ene.2026\_timestamp\_part00000.csv

### Salida 2

<b>Salida a</b> Ubicación de la salida	<b>Tipo de archivo</b> Formato de salida	<b>Delimitador</b> Separador CSV	<b>Compresión</b> Tipos disponibles
<input checked="" type="button" value="Amazon S3"/>	<input checked="" type="button" value="PARQUET"/>	<input type="button" value="Coma (,)"/>	<input type="button" value="None"/>

Cuenta de AWS del propietario del bucket de S3

- Cuenta de AWS actual  
043356869404
- Otra cuenta de AWS

**Ubicación de S3**  
El formato es s3://bucket/folder/

**Resumen de configuración**  
Almacenamiento de salida de archivos  
Crear una nueva carpeta para cada ejecución de flujo de trabajo  
Salida de archivo  
Autogenerate files  
División personalizada por valores de columna  
Desactivado

**Vista previa de la ruta de salida**  
s3://databrew-practica-danigayol/DataBrew-danigayol\_26ene.2026\_timestamp\_part00000.parquet

### Permisos

DataBrew needs permission to connect to data on your behalf. Use an IAM role with the [política necesaria](#) attached.

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

Se ha creado el trabajo de receta "DataBrew-danigayol".

DataBrew > Trabajos > DataBrew-danigayol

DataBrew-danigayol  
Conjunto de datos: chemb1-27 Receta: Sample recipe - 1

Ejecutar trabajo Acciones

Historial de la ejecución del trabajo Detalles del trabajo Linaje de datos

Última ejecución de trabajo 2 minutos, no hay ejecuciones de trabajos programadas

Historial de la ejecución del trabajo

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	Iniciado por	Iniciado el	Finalizado el
DataBrew-danigayol_2026-01-26-10:06:52	Realizado con éxito	1 minuto, 30 segundos	2 salidas		user3935192:Daniel_Gayol_Rodríguez	hace 2 minutos 26 de enero de 2026, 10:06:52 am	hace unos segundos 26 de enero de 2026, 10:06:52 am

Finalmente, nos vamos a “S3” y ya nos saldrán las carpetas:

databrew-practica-danigayol Información

Objetos Metadatos Propiedades Permisos Métricas Administración Puntos de acceso

Objetos (2)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
DataBrew-danigayol_26Jan2026_1769418489718/	Carpetas	-	-	-
DataBrew-danigayol_26Jan2026_1769418502926/	Carpetas	-	-	-

**DataBrew-danigayol\_26Jan2026\_1769418489718\_part00000.csv** Información

Propiedades Permisos Versiones

### Información general sobre el objeto

**Propietario**  
d1efcee244946868bbfc7ed543f4012c35bbed396dc1f3889c27be6adc76c1e3

**Región de AWS**  
EE.UU. Este (Norte de Virginia) us-east-1

**Última modificación**  
26 Jan 2026 10:08:23 AM CET

**Tamaño**  
3.2 MB

**Tipo**  
CSV

**Clave**  
DataBrew-danigayol\_26Jan2026\_1769418489718/DataBrew-danigayol\_26Jan2026\_1769418489718\_part00000.cs

## DataBrew-danigayol\_26Jan2026\_1769418502926\_part00000.parquet Inform

Propiedades

Permisos

Versiones

### Información general sobre el objeto

**Propietario**

d1efcee244946868bbfc7ed543f4012c35bbed396dc1f3889c27be6adc76c1e3

**Región de AWS**

EE.UU. Este (Norte de Virginia) us-east-1

**Última modificación**

26 Jan 2026 10:08:26 AM CET

**Tamaño**

2.2 MB

**Tipo**

parquet

**Clave**

 DataBrew-danigayol\_26Jan2026\_1769418502926/DataBrew-danigayol\_26Jan2026\_1769418502926\_part00000.parquet

## 2,) Muestra también el linaje resultante de los datos.

Para mostrar el Linaje de Datos, nos vamos al “job” ejecutado anteriormente y entramos en el menú de “Linaje de Datos”:

