

# **PR\_08.4 Dani Gayol Rodríguez**

PR_08.4 Dani Gayol Rodríguez.....	1
Apartado A.....	1
1.) Carga el CSV anterior en una carpeta llamada raw dentro de un bucket nombrado con algo similar a rrhh .....	2
2.) Desde Databrew crea una conexión de datos a dicha carpeta.....	5
a) ¿Cuánto ocupa el archivo?.....	6
b) Haz una captura que nos muestre el tipo y contenido de las 5 primeras filas de algunas de las columnas. Se ven en la pestaña: .....	7
3.) Crea una carpeta dentro del bucket anterior llamada perfil.....	8
4.) Genera el perfil de datos de dicho conjunto de datos. Deja la configuración por defecto. ¿Cuántas filas utiliza por defecto para el análisis?.....	9
5.) Analiza los datos obtenidos.....	10
a) ¿Cuántas columnas y de qué tipo tiene el conjunto de datos? .....	11
b) ¿Hay alguna correlación positiva o negativa que te llame la atención? .....	11
c) ¿Qué porcentaje de hombres y mujeres hay? .....	11
d) Analizando el diagrama de cajas de los salarios ¿En qué horquilla se mueven? ¿Cuál es la media, mediana y moda? ¿Están distribuidos simétricamente?.....	11
e) Busca la misma información para el campo Age in years. ....	12
6.) Haz una captura del contenido de la pestaña Linaje de datos. ¿Qué se muestra en ella?.....	13
Apartado B.....	14
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? .....	15
2.) Generaremos una receta para realizar diferentes transformaciones a los datos: 16	
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. ....	17
b) Elimina las columnas Short Month, DOW of Joining y Short DOW .....	18

c) Formatea la columna Date of Joining a la forma utilizada en España dd/mm/yyyy .....	19
d) Renombra la columna Phone No. a Telefono .....	21
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) .....	22
f) Realiza un cifrado determinista de los campos E Mail y Date of Birth.....	23
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. ....	24
3.) Publica la receta.....	26
4.) Crea en el bucket en el que estábamos trabajando una nueva carpeta llamada transformado.....	26
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.....	27
6.) Descarga el CSV obtenido y échale una ojeada para verificar que se han realizado las transformaciones. ....	30
Apartado C .....	31
1.) Crea en Databrew un conjunto de datos asociado al CSV de la carpeta transformado.....	32
2.) Crea un conjunto de reglas de calidad de los datos asociado al dataset anterior.33	
3.) Añade las siguientes reglas: .....	35
a) Valida el recuento de filas: Hemos utilizado un conjunto de datos de 1 millón de registros. Vamos a validar si el recuento coincide.....	35
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. ....	36
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. ....	37
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.....	38
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). ....	40

4.) Crea el conjunto de reglas sin asociarlo a ningún trabajo .....	41
Apartado D .....	42
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.....	43
2.) Vete al apartado de reglas de calidad en Databrew y asocia las reglas creadas a un trabajo de perfil. ....	44
a) Aplica el trabajo a todo el dataset: .....	44
b) Configura el bucket de salida del análisis en la carpeta del punto anterior.....	45
c) Verifica en rol adecuado en el apartado de Permisos (Labrole). Crea el trabajo.	45
3.) Verifica el nombre del trabajo de perfil asociado a las reglas:.....	46
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. .....	46
5.- De aparecer algún error vete a la pestaña de Estadísticas de columna y comprueba que es cierto el error de las reglas. ....	47
Apartado E .....	48
1.) Crea en el bucket en el que estábamos trabajando una nueva carpeta llamada curated. ....	49
2.) 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. El resultado del trabajo almacénalo en formato parquet comprimido en la carpeta curated. ....	49
3.) Crea un nuevo conjunto de datos en Databrew que apunte al archivo de curated. ....	54
4.) Verifica ahora las estadísticas de las columnas que has modificado para asegurarnos que todo ha ido bien. ¿Cuántas filas tiene ahora el archivo resultante?55	
Apartado F .....	57
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. ....	58
2.) Asocia dicho conjunto de reglas al trabajo de perfil. ....	58
3.) Ejecuta dicho trabajo contra todo el dataset.....	58
4.) Verifica en el perfil de datos, apartado Reglas de calidad que se han pasado correctamente todas las comprobaciones.....	58

# Apartado A

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

Para cargar el “csv”, entramos en “S3” y creamos un bucket:

The screenshot shows the AWS search bar with "S3" selected. Below it, the "Servicios" section is expanded, and the "S3" service card is highlighted with a red box. The main content area displays the "Buckets de uso general" tab, which lists one existing bucket: "databrew-practica-danigayol". The "Crear bucket" button is highlighted with a red box.

Nombre	Región de AWS	Fecha de creación
databrew-practica-danigayol	EE.UU. Este (Norte de Virginia) us-east-1	26 Jan 2026 10:04:11 AM CET

## Crear bucket Información

Los buckets son contenedores de datos almacenados en S3.

### Configuración general

#### Región de AWS

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

#### Tipo de bucket Información

##### Uso general

Recomendado para la mayoría de los casos de uso y patrones de acceso. Los buckets de uso general son del tipo de bucket de S3 original. Permiten una combinación de clases de almacenamiento que almacenan objetos de forma redundante en múltiples zonas de disponibilidad.

#### Nombre del bucket Información

rrhh-databrew-danigayol

Los nombres de los buckets deben tener entre 3 y 63 caracteres y ser únicos dentro del espacio de nombres global. Los nombres de los buckets t

#### Copiar la configuración del bucket existente: *opcional*

Solo se copia la configuración del bucket en los siguientes ajustes.

[Elegir el bucket](#)

Formato: s3://bucket/prefijo

Ahora, una vez creado, entramos dentro y creamos la carpeta “raw”:

El bucket “rrhh-databrew-danigayol” se creó correctamente  
Para cargar archivos y carpetas, o para configurar ajustes adicionales del bucket, elija [Ver detalles](#).

Buckets de uso general Todas las regiones de AWS

Buckets de directorio

Buckets de uso general (2) Información

Los buckets son contenedores de datos almacenados en S3.



[Copiar ARN](#)

[Vaciar](#)

[Eliminar](#)

[Crear bucket](#)

Buscar buckets por nombre

< 1 >

Nombre	Región de AWS	Fecha de creación
<a href="#">databrew-practica-danigayol</a>	EE.UU. Este (Norte de Virginia) us-east-1	26 Jan 2026 10:04:11 AM CET
<a href="#">rrhh-databrew-danigayol</a>	EE.UU. Este (Norte de Virginia) us-east-1	27 Jan 2026 9:23:36 AM CET

## Crear carpeta Información

Utilice carpetas para agrupar los objetos en buckets. Al crear una carpeta, S3 creará un ob

**ⓘ Su política de bucket podría bloquear la creación de carpetas**

Si su política de bucket impide cargar objetos sin etiquetas, metadatos o beneficiar cargar una carpeta vacía y especificar la configuración adecuada.

### Carpeta

#### Nombre de la carpeta

raw

Los nombres de las carpetas no pueden contener "/". [Consulte las reglas de nomenclatura ↗](#)

ⓘ Se creó correctamente la carpeta "raw"

## rrhh-databrew-danigayol Información

Objetos Metadatos Propiedades Permisos Métricas Administración Puntos de acceso

### Objetos (1)



Copiar URI de S3

Copiar URL

Descargar

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 y concederles permisos de forma explícita. [Más información ↗](#)

Buscar objetos por prefijo

Nombre

▲ | Tipo

▼ | Última modificación

raw/

Carpeta

### Ahora vamos a entrar dentro de la carpeta y cargar el “csv”:

#### Cargar Información

Agregue los archivos y las carpetas que desea cargar en S3. Para cargar un archivo de más de 160 GB, utilice la CLI de AWS, los SDK de AWS o la API REST de Amazon S3. [Más información ↗](#)

Arrastre y suelte aquí los archivos y carpetas que deseé cargar, o seleccione Add files (Agregar archivos) o Add folder (Agregar carpeta).

#### Archivos y carpetas (1 total, 265.2 MB)

Se cargarán todos los archivos y las carpetas de esta tabla.

Eliminar

Agregar archivos

Agregar carpeta

Buscar por nombre

< 1 >

Nombre

▼ | Carpeta

▼ | Tipo

▼ | Tamaño

Hr1m.csv

-

text/csv

265.2 MB

#### Destino Información

Destino

s3://rrhh-databrew-danigayol/raw/ ↗

► Detalles del destino

Los ajustes del bucket que afectan a los objetos nuevos almacenados en el destino especificado.

► Permisos

Conceder acceso público y acceso a otras cuentas de AWS.

► Propiedades

Especifique la clase de almacenamiento, los ajustes de cifrado, las etiquetas y mucho más.

Cancelar

Cargar

Se ha realizado la carga correctamente  
Para obtener más información, consulte la tabla Archivos y carpetas.

Cargar: estado

Después de salir de esta página, la siguiente información ya no estará disponible.

**Resumen**

Destino: s3://rrhh-databrew-danigayol/raw/ | Realizado correctamente: 1 archivo, 265.2 MB (100.00%) | Con errores: 0 archivos, 0 B (0%)

**Archivos y carpetas** | Configuración

**Archivos y carpetas** (1 total, 265.2 MB)

Nombre	Carpetas	Tipo	Tamaño	Estado	Error
Hr1m.csv	-	text/csv	265.2 MB	Realizado correctamente	-

Amazon S3 > Buckets > rrhh-databrew-danigayol > raw/

**raw/**

**Objetos** | Propiedades

**Objetos (1)**

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
Hr1m.csv	csv	27 Jan 2026 9:28:48 AM CET	265.2 MB	Estándar

**Copiar URI de S3**

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

Para crear una conexión hacia esta carpeta, entramos primeramente en “AWS Glue DataBrew”

AWS Glue DataBrew

Servicios (159)

**AWS Glue DataBrew**  
Herramienta de preparación de datos visuales para limpiar y normalizar datos para a...

**AWS Glue**  
AWS Glue es un servicio de integración de datos sin servidor.

**AWS Private Certificate Authority**  
Servicio de la entidad de certificación privada administrada

Conjuntos de datos

Conjuntos de datos | Conexiones

Conjuntos de datos (1)

Nombre del conjunto de datos	Tipo de datos	Perfil de datos	Origen	Ubicación	Fecha de creación	Creado por	Etiquetas
chembl-27	parquet	-	S3	s3://databrew-public-datasets-us-east-1/chembl-27.parquet	hace un día 26 de enero de 2026, 9:20:54 am	voclabs	-

Conectar nuevo conjunto de datos

**Conectarse a un nuevo conjunto de datos**

- 
- Lago de datos/almacén de datos
- Amazon S3**
- Conexiones de la base de datos
- Amazon Redshift
- JDBC
- Catálogo de datos de AWS Glue
- Tablas de S3 del catálogo de datos
- Tablas de Redshift del catálogo de datos
- Tablas de RDS del catálogo de datos
- Todas las tablas de AWS Glue

Introducir el origen desde S3  
Para que pueda seleccionar una carpeta, todos los archivos en ella tienen que compartir el mismo tipo de archivo. Si hay diferentes esquemas, se combinarán.

s3://rrhh-databrew-danigayol/raw/

El formato es s3://bucket/prefix

**Se seleccionan todos los archivos de la carpeta raw y sus subcarpetas**

S3 Buckets > rrhh-databrew-danigayol Seleccionar toda la carpeta C

Buscar objetos de S3 por nombre

Nombre	Tamaño	Última actualización
<input checked="" type="radio"/> raw	-	-

**Configuraciones adicionales**

Tipo de archivo seleccionado  
Formato del archivo seleccionado  CSV Coma (,)

Valores de encabezado de columna  
 Considera la primera fila como encabezado  
La primera fila del conjunto de datos se tratará como valores de encabezado de columna  
 Agregar encabezado predeterminado  
Los encabezados predeterminados se agregarán con los valores Column\_1, Column\_2...

**Etiquetas:**opcional

Metadatos que puede definir y asignar a los recursos de AWS. Cada etiqueta es una etiqueta sencilla que consta de una clave definida por el cliente (nombre) y un valor opcional. El uso de etiquetas puede facilitarle la administración, la búsqueda y el filtrado de recursos por finalidad, propietario, entorno u otros criterios.

Cancelar Crear conjunto de datos

Conjuntos de datos (2)							<a href="#">Ver detalles</a>	<a href="#">Crear proyecto con este conjunto de datos</a>	<a href="#">Ejecutar perfil de datos</a>	Acciones	Conectar nuevo conjunto de datos
	Nombre del conjunto de datos	Tipo de datos	Perfil de datos	Origen	Ubicación	Fecha de creación	Creado por	Etiquetas			
<input type="checkbox"/>	rrhh-databrew-raw-hr1m	csv	-	S3	s3://rrhh-databrew-danigayol/raw/ <span style="border: 2px solid red; padding: 2px;">🔗</span>	hace unos segundos 27 de enero de 2026, 9:58:14 am	voclabs	-			
<input type="checkbox"/>	chembl-27	parquet	-	S3	s3://databrew-public-datasets-us-east-1/chembl-27.parquet <span style="border: 2px solid red; padding: 2px;">🔗</span>	hace un día 26 de enero de 2026, 9:20:54 am	voclabs	-			

a) ¿Cuánto ocupa el archivo?

El archivo ocupa **265.2 MB**

**Hr1m.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

27 Jan 2026 9:28:48 AM CET

**Tamaño**

265.2 MB

### Tipo

CSV

### Clave

raw/Hr1m.csv

También lo podemos mirar desde el “databrew”:

Archivos	
<input type="text"/>	
Archivo	Tamaño
raw/Hr1m.csv	265,2 MiB

b) Haz una captura que nos muestre el tipo y contenido de las 5 primeras filas de algunas de las columnas. Se ven en la pestaña:

**Lo podemos comprobar directamente desde “AWS Glue Databrew”, entramos dentro del conjunto de datos y nos aparece ahí:**

Vista previa del conjunto de datos

Nombre de la columna	Primeras 5 filas de datos
# Emp ID	549821, 429350, 702166, 982838, 565681
ABC Name Prefix	Mrs., Mr., Drs., Prof., Mr.
ABC First Name	Jeffrey, Shelby, Wen, Aaron, Frederic
ABC Middle Initial	C, D, P, Q, M
ABC Last Name	Murakami, Davidson, Russo, Delima, Christofferson
ABC Gender	F, M, F, M, M
ABC E Mail	jeffrey.murakami@gmail.com, shelby.davidson@verizon.net, wen.russo@yahoo.com, aaron.delima@gmail.com, frederic.christofferson@microsoft.com
ABC Father's Name	Alex Murakami, Frederick Davidson, Laurence Russo, Napoleon Delima, Horst Christofferson
ABC Mother's Name	Kimberly Murakami, Briana Davidson, Jeri Russo, Marketta Delima, Sarah Christofferson
ABC Mother's Maiden Name	Bona, Whitford, Montero, McAfee, Zullo
ABC Date of Birth	10/7/1991, 10/21/1964, 9/4/1994, 3/6/1968, 7/28/1975

Cuadrícula **Esquema** Texto Árbol

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

Para ello, tenemos que volver a “S3” y crear del mismo modo una carpeta nueva:

rrhh-databrew-danigayol Información

Objetos Metadatos Propiedades Permisos Métricas Administración Puntos de acceso

Objetos (1) Copia URI de S3 Copiar URL Descargar Abrir Eliminar Acciones Crear carpeta Cargar

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: raw/

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
raw/	Carpeta	-	-	-

### Crear carpeta Información

Utilice carpetas para agrupar los objetos en buckets. Al crear una carpeta, S3 creará un obj

**i Su política de bucket podría bloquear la creación de carpetas**  
Si su política de bucket impide cargar objetos sin etiquetas, metadatos o beneficiarios, [carga](#) para cargar una carpeta vacía y especificar la configuración adecuada.

### Carpeta

Nombre de la carpeta

perfil

Los nombres de las carpetas no pueden contener “/”. [Consulte las reglas de nomenclatura](#)

✓ Se creó correctamente la carpeta "perfil"

## rrhh-databrew-danigayol Información

Objetos Metadatos Propiedades Permisos Métricas Administración

### Objetos (2)



Copiar URI de S3

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario](#) que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

<input type="checkbox"/>   Nombre	▲   Tipo	▼
<input type="checkbox"/> perfil/	Carpeta	
<input type="checkbox"/> raw/	Carpeta	

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

Tenemos que volver a “AWS Glue Databrew”, una vez dentro, entramos en el menú de “trabajos” y hacemos clic en “crear trabajo” y lo ponemos de la siguiente manera:

DataBrew > Trabajos > Crear trabajo

## Crear trabajo

**Detalles del trabajo**

**Nombre del trabajo**  
Identificador del trabajo  
**perfil-datos-databrew**

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.

Elegir conjunto de datos  
**s3://rrhh-databrew-raw-hr1m**

## Configuración de salida del trabajo

La ejecución de un trabajo genera archivos de salida en los destinos de archivo especificados.

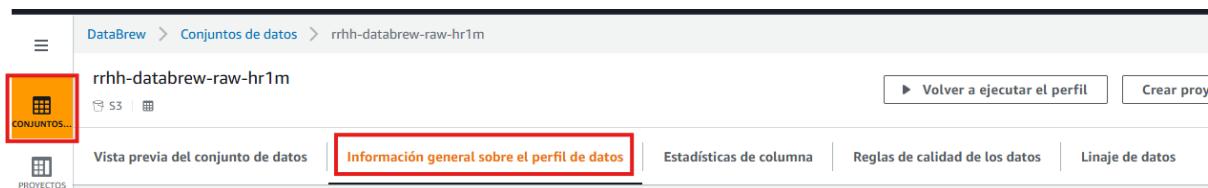
<b>Cuenta de AWS del propietario del bucket de S3</b>	<b>Tipo de archivo</b>
<input checked="" type="radio"/> Cuenta de AWS actual 043356869404	Formato de salida
<input type="radio"/> Otra cuenta de AWS	JSON
<b>Ubicación de S3</b> El formato es s3://bucket/folder/	
<b>s3://rrhh-databrew-danigayol/perfil/</b> <input type="button" value="X"/> <input type="button" value="Explorar"/>	
<b>Cifrado</b>	
<input type="checkbox"/> <b>Habilitar el cifrado para el archivo de salida del trabajo</b> Cifrar el archivo de salida del trabajo con SSE-S3 o AWS KMS	

Historial de la ejecución del trabajo					
ID de ejecución de trabajo		Estado de la última ejecución del trabajo	Tiempo de ejecución	Salida	Iniciado por
perfil-datos-databrew_2026-01-27-10:00:30		Realizado con éxito	2 minutos, 55 segundos	1 salida	user3935192-Daniel_Gayol_Rodríguez

**Por defecto, utiliza todas las filas del dataset para generar el perfil**

## 5.) Analiza los datos obtenidos.

Para ver los datos, nos vamos al menú “conjunto de datos” y entramos en el que creamos para esta práctica, una vez dentro nos vamos al apartado de “Información general sobre el perfil de datos”:



The screenshot shows the AWS DataBrew interface. At the top, there's a navigation bar with 'DataBrew > Conjuntos de datos > rrhh-databrew-raw-hr1m'. Below it, the dataset name 'rrhh-databrew-raw-hr1m' is displayed along with S3 and Glue icons. On the right, there are buttons for 'Volver a ejecutar el perfil' and 'Crear proy'. The main content area has tabs: 'Vista previa del conjunto de datos' (highlighted with a red box), 'Información general sobre el perfil de datos' (also highlighted with a red box), 'Estadísticas de columna', 'Reglas de calidad de los datos', and 'Linaje de datos'.

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



The screenshot shows the 'Resumen' section of the dataset profile. It displays the total number of rows (20.000) and columns (37). Under the 'TIPOS DE DATOS' section, it shows the distribution of data types: 7 integer columns, 2 double columns, and 28 string columns.

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

Hay una correlación en función de los años y el salario, a mayor edad, el salario es más elevado



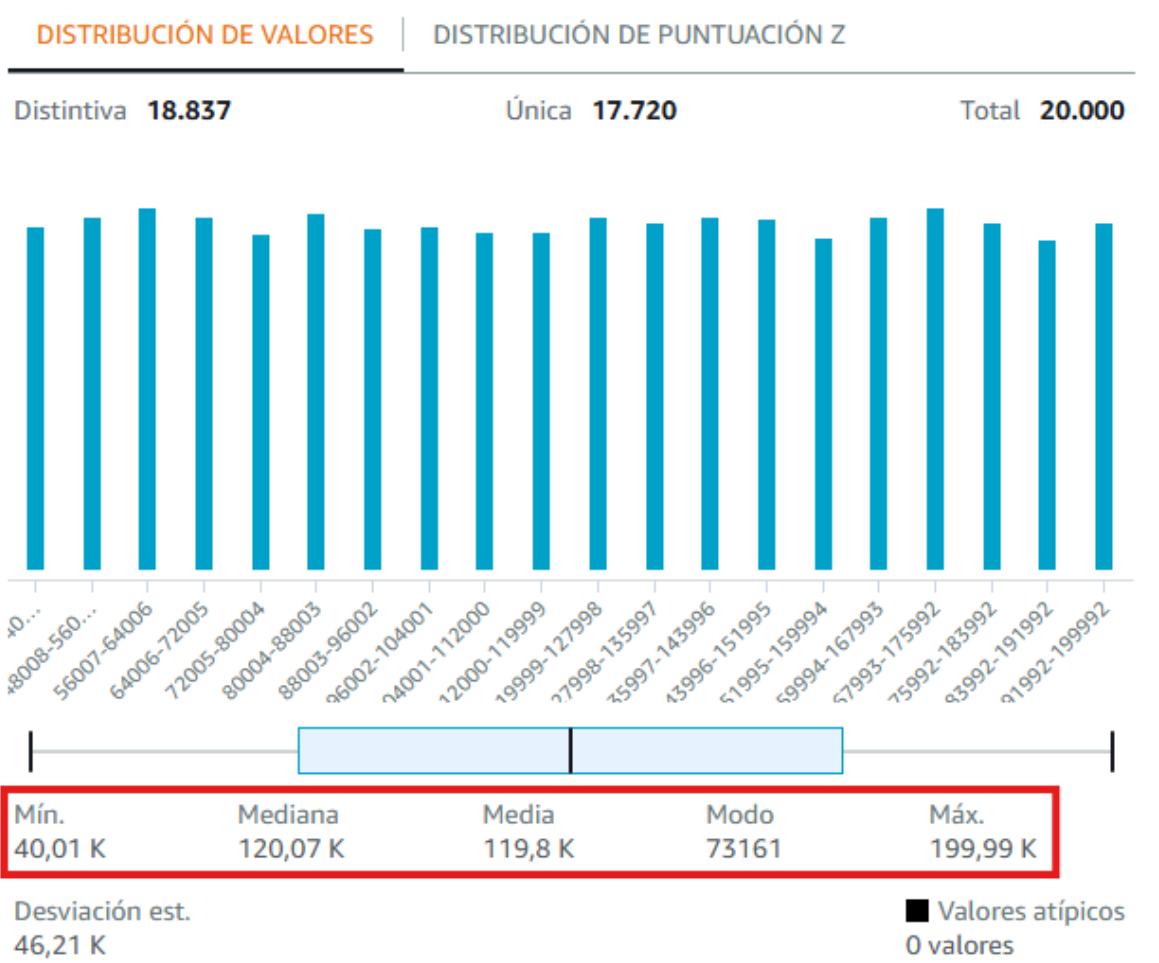
c) ¿Qué porcentaje de hombres y mujeres hay?

Para ello, entramos en la columna de “gender” y nos aparecerá ahí

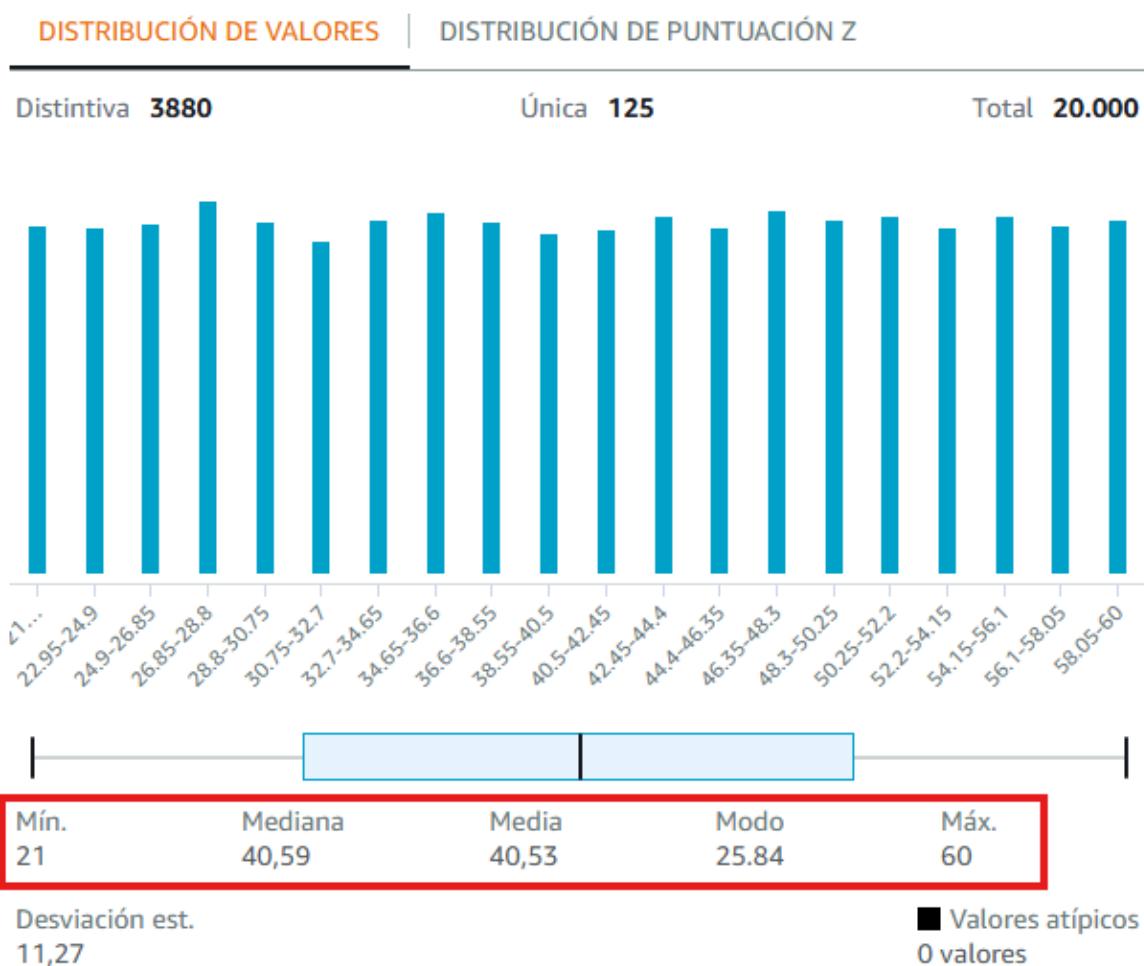
 VÁLIDO

M		10,06 K	50%
F		9,94 K	49%

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

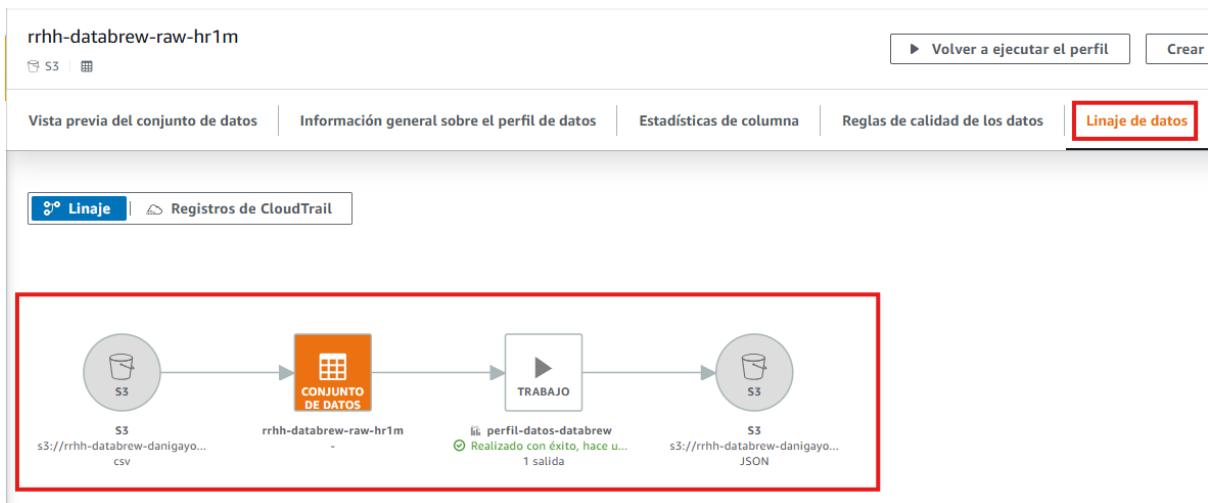


e) Busca la misma información para el campo Age in years.



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

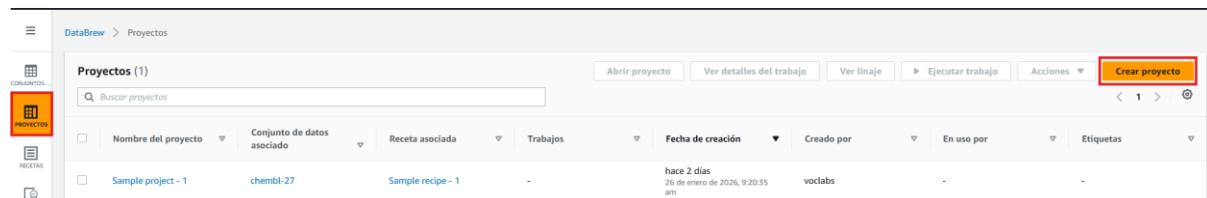
En la pestaña de “Linaje de datos” podemos ver dataset en S3, luego, los procesos aplicados en el “Databrew” y la ubicación de los resultados del perfil



# 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?**

Para crear el proyecto, tenemos que entrar en “AWS Glue Databrew”, una vez dentro, nos vamos al apartado de “proyectos” y lo creamos por defecto:



## Crear proyecto

### Detalles del proyecto

Nombre del proyecto

**rrhh-databrew-proyecto**

El nombre del proyecto debe contener entre 1 y 255 caracteres. Los caracteres válidos son alfanuméricos (A-Z, a-z, 0-9), guiones (-) y espacios.

### Detalles de la receta

Los pasos de limpieza de datos en DataBrew se almacenan como una receta. Una receta está conectada a un proyecto de forma predeterminada.

Receta asociada

**Crear nueva receta ▾**

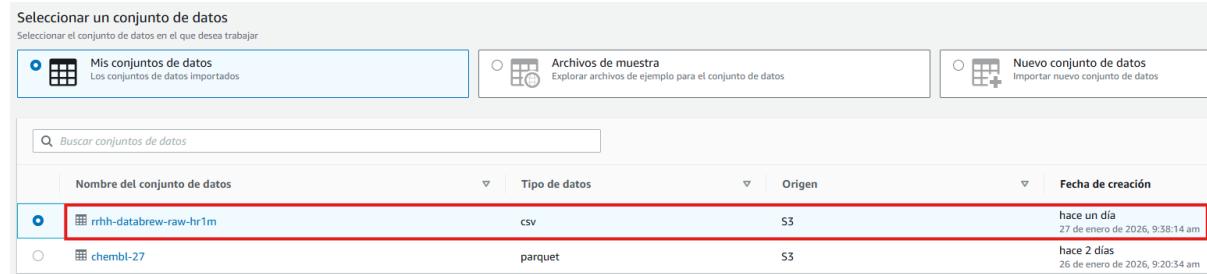
Nombre de la receta

**rrhh-databrew-proyecto-recipe**

El nombre de la receta debe contener entre 1 y 255 caracteres. Los caracteres válidos son alfanuméricos (A-Z, a-z, 0-9), guiones (-) y espacios.

Importar pasos de la receta

Importe los pasos de una receta existente al proyecto. La receta elegida existente no se editará.



**Por defecto:**

## ▼ Muestreo : opcional

Seleccionar el tipo y el tamaño de la muestra

### Tipo

Primeras n filas

¿Cuántas filas desea muestrear?

500

1000

2500

Tamaño personalizado

## 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

Proyectos (2)								
	Nombre del proyecto	Conjunto de datos asociado	Receta asociada	Trabajos	Fecha de creación	Creado por	En uso por	Etiquetas
<input type="checkbox"/>	rrhh-databrew-proyecto	rrhh-databrew-raw-hr1m	rrhh-databrew-projecto-recipe	-	hace 15 minutos 28 de enero de 2026, 11:30:11 am	voclabs	voclabs/user3935192:Dani el_Gayol_Rodr_guez	-
<input type="checkbox"/>	Sample project - 1	chembl-27	Sample recipe - 1	-	hace 2 días 26 de enero de 2026, 9:20:35 am	voclabs	-	-

Por defecto, como ya señalé antes, utiliza 500 registros

## rrhh-databrew-proyecto

Conjunto de datos: rrhh-databrew-raw-hr1m | Mues

DESHACER REHACER

FILTRAR ORDENAR COLUMNA

FORMATO

Visualizando 37 columnas ▾ 500 filas

## 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.

The screenshot shows the 'Fusionar columnas' (Merge columns) dialog box. At the top, it says 'Columna de origen' (Source column) and 'Seleccione dos o más columnas en el orden de fusión' (Select two or more columns in the order of fusion). Below this, a list of columns is shown in a red-bordered box:

Name Prefix	X
First Name	X
Middle Initial	X
Last Name	X

Below the list is a button 'Agregar una columna' (Add a column) with a dropdown arrow. The next section is 'Separador - Opcional' (Separator - Optional), which says 'Los valores concatenados están separados por este' (The concatenated values are separated by this) and has a red-bordered input field. The final section is 'Nombre de la columna nueva' (New column name), which says 'Nombre de la columna de destino con la que se va a fusionar' (Name of the destination column to merge with) and has a red-bordered input field containing 'Nombre\_completo\_empleado'. Below this is a note 'Los caracteres válidos son alfanuméricos, guiones bajos y espacios' (Valid characters are alphanumeric, underscores, and spaces). At the bottom, it says 'Aplicar transformación a' (Apply transformation to) and has a radio button 'Todas las filas (500 filas)' (All rows (500 rows)) selected, with the note 'La transformación se aplicará a todas las filas del conjunto de datos' (The transformation will be applied to all rows of the data set).

ORIGEN			
ABC	Nombre_completo_empleado	▼	↑
Distintiva	500	Única	500
Mrs. Jeffrey C Murakami		1	0,2%
Mr. Shelby D Davidson		1	0,2%
Drs. Wen P Russo		1	0,2%
Todos los demás valores	497		99,4%
Mrs. Jeffrey C Murakami			
Mr. Shelby D Davidson			
Drs. Wen P Russo			
Prof. Aaron Q Delima			
Mr. Frederic M Christofferso			
Hon. Billie M Lachapelle			
Mrs. Roseline M Bach			
Ms. Jerlene H Chalk			
Drs. Marcella Q Payan			
Mrs. Sylvie X Pautz			
Mr. Osvaldo S Swayne			
Mr. Devon E Kehoe			
Ms. Brigette G Tong			
Dr. Mauricio F Ryles			
Hon. Fidela N Norden			
Hon. Isaias A Dibenedetto			
Mr. Landon K Stolz			

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

< Eliminar columna X

Columnas de origen  
Nombre de la columna que se va a eliminar

Nombre de la columna ▼

ABC Short Month X

ABC DOW of Joining X

ABC Short DOW X

---

👁 Vista previa de los cambios

Cancelar Aplicar

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

< Dar formato a la columna X

Columna de origen  
Seleccionar una columna para dar formato  
▼

Dar formato a la columna a  
▼

Elegir formato de fecha y hora  
▼

Si no se selecciona nada, el valor predeterminado es aaaa-mm-dd  
HH:MM:SS

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 de los cambios](#)

[Cancelar](#) [Aplicar](#)

ORIGEN					
ABC Date of Joining		Y	↑↓	...	
Distintiva	484	Única	469	Total	500
26/09/2018			3		0,6%
24/12/2018			2		0,4%
28/12/2018			2		0,4%
Todos los demás valores			493		98,6%
15/01/2014					
22/01/1989					
24/05/2019					
01/09/2007					
19/06/2008					
25/12/1997					
20/07/2017					
11/05/2017					
24/12/2018					
10/02/1985					
26/01/2013					
30/03/2009					
27/03/2006					
10/09/2018					
26/01/2003					
19/07/2017					
16/03/2009					

d) Renombra la columna Phone No. a Telefono

Cambiar el nombre de la columna X

Columna de origen  
Seleccionar columna para cambiar el nombre

Phone No. ▼

Nombre de la columna nueva  
Nuevo nombre para la columna

Telefono ▼

Los caracteres válidos son alfanuméricos, guiones bajos y espacios

[Vista previa de los cambios](#)

[Cancelar](#) Aplicar

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)

**Valores de Redact**

**Columnas de origen**  
Elija una o varias columnas de origen.

*Nombre de la columna*

ABC SSN X   ABC Telefono X  
ABC Password X

**Valor que se va a redactar**  
Elija la opción para ocultar los datos originales con caracteres especificados.

Todos los caracteres

**Símbolo Redact**

#

El símbolo de Redact debe ser un carácter único.

Mantenga el formato de los datos  
Los caracteres no alfanuméricos se conservarán después de la transformación.

**Aplicar la redacción a**

Valor de cadena completo  
Valor de cadena completo de cada fila.

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

**Cifrado de datos**

**Columnas de origen**  
Elija una o varias columnas de origen.

*Nombre de la columna* ▾

**ABC E Mail X**

**ABC Date of Birth X**

**Opciones de cifrado**

**Cifrado determinista**  
Cifrar los datos manteniendo el mismo valor resultante para cada valor distinto

**Cifrado probabilístico**  
Cifrar los datos con un valor resultante diferente para todos los valores cifrados

**i** Puede descifrar valores con cifrado determinista solo con un secreto de Secrets Manager mediante los pasos de la receta DataBrew.

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.

**Grupo**

**Lista de columnas**  
Agregar columna con agregación para la tabla agrupada

Nombre de la columna	Agregado	Nombre de la columna nueva	Tipo de la columna nueva
ABC Gender	Agrupar por	Gender	ABC Cadena
# Salary	Desviación absoluta m...	Salario_medio	# Entero

**Remover** **Remover**

**Agregar otra columna**

**Tipo de grupo**

- Agrupar como nueva tabla (sustituye todas las columnas existentes por columnas nuevas)
- Agrupar como columnas nuevas (se agregan nuevas columnas a las existentes)

**Vista previa de la tabla de grupo**

ABC Gender	# Salario_medio
F	38077
M	39803

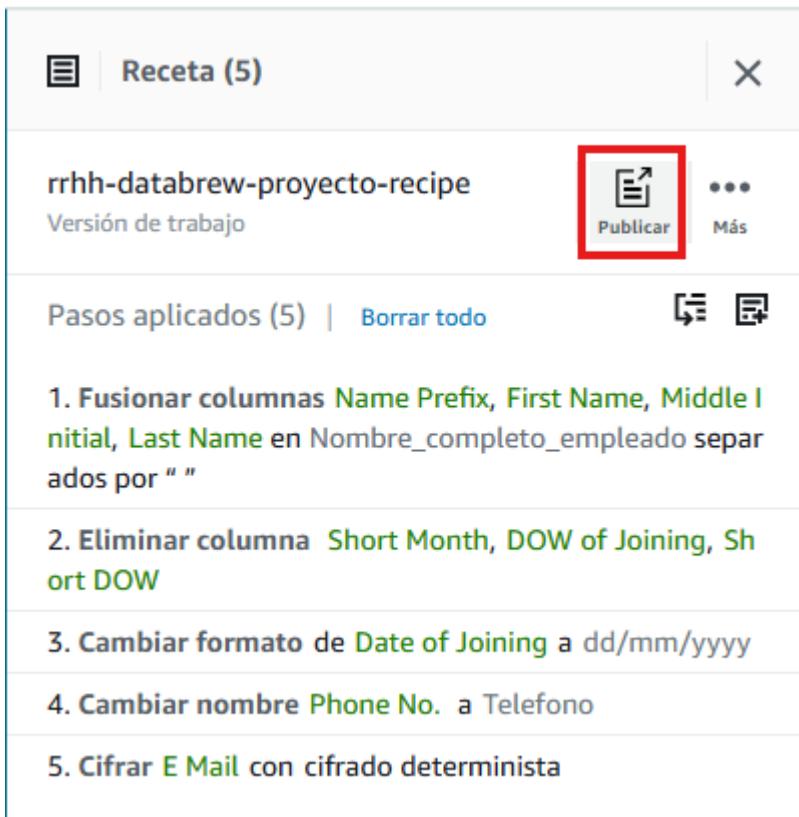
**Receta (6)**

rrhh-databrew-proyecto-recipe | Versión de trabajo | Publicar | Más

Pasos aplicados (6) | Borrar todo

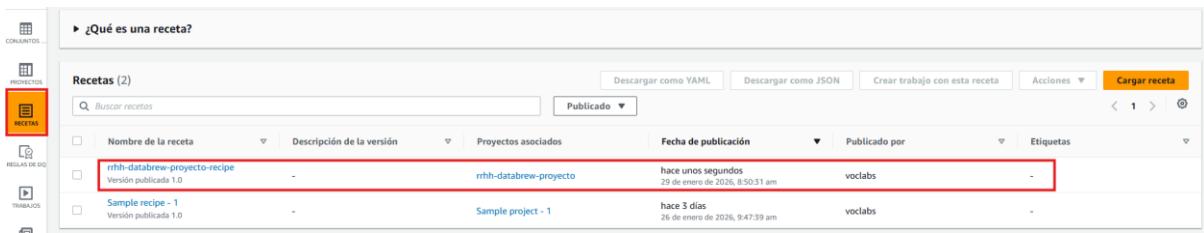
1. Fusionar columnas Name Prefix, First Name, Middle Initial, Last Name en Nombre\_completo\_empleado separados por " "
2. Eliminar columna Short Month, DOW of Joining, Short DOW
3. Cambiar formato de Date of Joining a dd/mm/yyyy
4. Cambiar nombre Phone No. a Telefono
5. Cifrar E Mail con cifrado determinista
6. Agrupar por Gender y crear S AN\_ABSOLUTE\_DEVIATION(Salary) Editar Eliminar

### 3.) Publica la receta.



The screenshot shows the AWS DataBrew Recipe Editor interface. At the top, it displays "Receta (5)" and the name of the recipe: "rrhh-databrew-proyecto-recipe". Below this, there's a section for "Versión de trabajo". On the right side of the screen, there are three buttons: "Publicar" (highlighted with a red box), "Más", and "..." (three dots). Underneath the recipe name, it says "Pasos aplicados (5) | Borrar todo". A list of five steps is provided:

1. Fusionar columnas Name Prefix, First Name, Middle Initial, Last Name en Nombre\_completo\_empleado separados por " "
2. Eliminar columna Short Month, DOW of Joining, Short DOW
3. Cambiar formato de Date of Joining a dd/mm/yyyy
4. Cambiar nombre Phone No. a Telefono
5. Cifrar E Mail con cifrado determinista

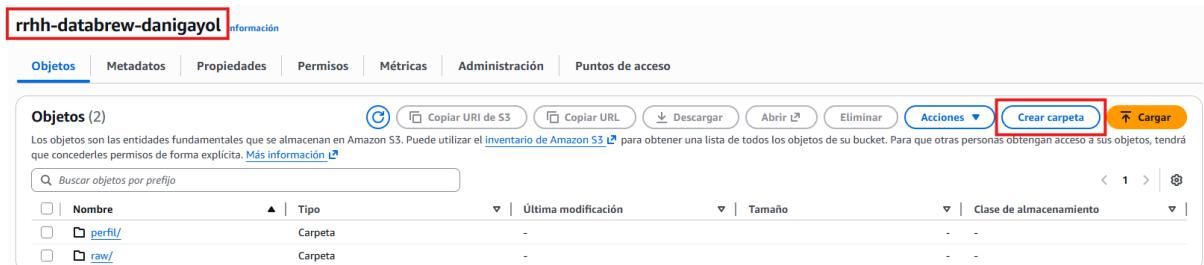


The screenshot shows the "Recetas" (Recipes) page in the AWS DataBrew console. On the left, there's a sidebar with icons for CONJUNTOS, PROYECTOS (highlighted with a red box), REGLAS DE DO, and TRABAJOS. The main area has a header "¿Qué es una receta?". Below it, a table lists two recipes:

Nombre de la receta	Descripción de la versión	Proyectos asociados	Fecha de publicación	Publicado por	Etiquetas
rrhh-databrew-proyecto-recipe Versión publicada 1.0	-	rrhh-databrew-proyecto	hace unos segundos 29 de enero de 2026, 8:50:31 am	voclabs	-
Sample recipe - 1 Versión publicada 1.0	-	Sample project - 1	hace 3 días 26 de enero de 2026, 9:47:59 am	voclabs	-

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

Para este paso, nos tenemos que ir hasta “S3” y desde ahí, crear la carpeta al igual que en pasos anteriores



The screenshot shows the "Objetos" (Objects) page in the AWS S3 console. At the top, it shows the bucket name: "rrhh-databrew-danigayol". Below the navigation tabs (Objetos, Metadatos, Propiedades, Permisos, Métricas, Administración, Puntos de acceso), there are several buttons: "Copiar URI de S3", "Copiar URL", "Descargar", "Abrir", "Eliminar", "Acciones", and "Crear carpeta" (highlighted with a red box). There's also a "Cargar" button. A note below the buttons states: "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](#)". The main table lists two objects:

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
perfil/	Carpeta	-	-	-
raw/	Carpeta	-	-	-

## Crear carpeta Información

Utilice carpetas para agrupar los objetos en buckets. Al crear una carpeta, S3 creará un objeto con el nombre de la carpeta.

### ⓘ Su política de bucket podría bloquear la creación de carpetas

Si su política de bucket impide cargar objetos sin etiquetas, metadatos o beneficiarios específicos, [carga](#) para cargar una carpeta vacía y especificar la configuración adecuada.

## Carpeta

### Nombre de la carpeta

transformado

Los nombres de las carpetas no pueden contener "/". [Consulte las reglas de nomenclatura ↗](#)

✓ Se creó correctamente la carpeta "transformado"

## rrhh-databrew-danigayol Información

Objetos Metadatos Propiedades Permisos Métricas Administración Puntos de acceso

### Objetos (3)



[Copiar URI de S3](#)

[Copiar URL](#)

[Descargar](#)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3 ↗](#) para obtener una lista de los objetos y concederles permisos de forma explícita. [Más información ↗](#)

Buscar objetos por prefijo

<input type="checkbox"/>   Nombre	▲   Tipo	▼   Última modificación	▼
<input type="checkbox"/> perfil/	Carpeta	-	
<input type="checkbox"/> raw/	Carpeta	-	
<input type="checkbox"/> transformado/	Carpeta	-	

**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.**

Nos volvemos otra vez a “AWS Glue Databrew” y desde ahí nos vamos al apartado de “trabajos”:

The screenshot shows the AWS DataBrew interface. In the top navigation bar, 'DataBrew' is selected. Below it, there are three tabs: 'Trabajos de recetas' (selected), 'Trabajos de perfil', and 'Programaciones'. On the left, a sidebar has several icons: 'CONSULTAS', 'PROYECTOS', 'RECETAS', 'REGLAS DE DQ' (highlighted with a red box), and 'NOVEDADES'. The main content area displays a table titled 'Trabajos de recetas (1)'. The table columns are: Nombre del trabajo, Estado, Entrada del trabajo, Salida del trabajo, Última ejecución, Creado el, Creado por, and Etiquetas. One row is shown: 'DataBrew-danigayol' (Realizado con éxito), 'chembl-27', 'Sample recip...', '2 salidas', 'hace 3 días', '26 de enero de 2026, 10:08:37 am', 'vocabs', and '-'.

## Crear trabajo

### Detalles del trabajo

#### Nombre del trabajo

Identificador del trabajo

DataBrew-transformado-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 espacio:

### 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

rrhh-databrew-raw-hr1m



Explorar conjuntos de datos

Conectar nuevo conjunto de datos

#### Seleccionar una receta

#### Versión de la receta

rrhh-databrew-proyecto-recipe



Versión 1.0

Examinar recetas

## Configuración de salida del trabajo

La ejecución de un trabajo genera archivos de salida en los destinos de archivo especificados.

### Salida 1

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

► Configuración adicional - *opcional*

## 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.

DataBrew-transformado-danigayol

Conjunto de datos: rrhh-databrew-raw-hr1m Receta: rrhh-databrew-proyecto-receta

Acciones ▾

[Historial de la ejecución del trabajo](#)

[Detalles del trabajo](#)

[Línea de datos](#)

Última ejecución de trabajo 19 minutos, no hay ejecuciones de trabajos programadas

### Historial de la ejecución del trabajo

Acciones ▾

< 1 > ⌂

ID de ejecución de trabajo	Estado de la última ejecución del trabajo	Tiempo de ejecución	Salida	Resumen	Iniciado por	Iniciado el	Finalizado el
DataBrew-transformado-danigayol_2026-01-29-09:00:21	<span style="color: green;">Realizado con éxito</span>	17 minutos, 7 segundos	1 salida		user3935192=Daniel_Gayol_Rodr_guez	hace 19 minutos 29 de enero de 2026, 9:00:21 am	hace unos segur 29 de enero de 20 am

Ahora verificamos rápidamente que aparece el archivo en “S3”:

The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with navigation links like 'Amazon S3', 'Buckets', 'Seguridad y administración de acceso', 'Información y administración de almacenamiento', and 'AWS Marketplace para S3'. The main area displays a file named 'DataBrew-transformado-danigayol\_29Jan2026\_1769673762474\_part00000.csv'. A red box highlights the file name at the top. Below it, there are four buttons: 'Copiar URI de S3', 'Descargar' (which is also highlighted with a red box), 'Abrir L', and 'Acciones de objetos'. Underneath these buttons, there are three tabs: 'Propiedades' (selected), 'Permisos', and 'Versiones'. A large red box encloses the 'Información general sobre el objeto' section, which contains details such as Propietario, Región de AWS, Última modificación, Tamaño, Tipo, and Clave. To the right of this section, there are several metadata fields with their values: URI DE S3, Nombre de recurso de la entidad, Etiqueta de entidad (Entidad), and URL del objeto.

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

Para descargar el “csv”, desde la pestaña donde lo dejamos en el ejercicio anterior, nos aparece un botón que pone “descargar” y simplemente es darle ahí y se nos descarga directamente

This screenshot shows the same AWS S3 file properties as the previous one, but with a different focus. The 'Descargar' button is highlighted with a red box again. The rest of the interface is identical, including the sidebar, the file name at the top, the other buttons, the tabs, and the 'Información general sobre el objeto' section with its various details and metadata fields.

**Una vez descargado, abrimos el archivo para comprobar que esta todo correcto:**

## Apartado C

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

Para crear el conjunto de datos, nos vamos a “AWS Glue Databrew” y una vez ahí, nos vamos al apartado de “conjuntos de datos” para crearlo:

CONJUNTOS	DataBrew > Conjuntos de datos						
PROYECTOS	<a href="#">Conjuntos de datos</a>   <a href="#">Conexiones</a>						
RECETAS							
REGISTROS DE DQ							
TRABAJOS							
Conjuntos de datos (2)							
<div style="display: flex; justify-content: space-between;"> <span>Ver detalles</span> <span>Crear proyecto con este conjunto de datos</span> <span>Ejecutar perfil de datos</span> <span>Acciones</span> <span>Conectar nuevo conjunto de datos</span> </div>							
<div style="display: flex; align-items: center;"> <span>Buscar conjuntos de datos</span> </div>							
Nombre del conjunto de datos	Tipo de datos	Perfil de datos	Origen	Ubicación	Fecha de creación	Creado por	Etiquetas
<a href="#">hrm-databrew-raw-hr1m</a>	csv	<a href="#">Ver perfil de datos</a>	S3	s3://rhm-databrew-danigayl/s3raw/	hace 2 días 27 de enero de 2026, 9:38:14 am	voclabs	-
<a href="#">chembi-27</a>	parquet	-	S3	s3://databrew-public-datasets-us-east-1/chembi-27.parquet/	hace 3 días 26 de enero de 2026, 9:20:34 am	voclabs	-

**Le damos al botón de “conectar nuevo conjunto de datos” y lo configuramos de la siguiente manera:**

# Nueva conexión

## Detalles del nuevo conjunto de datos

**Conectarse a un nuevo conjunto de datos**

↑ Carga de archivo	Introducir el origen desde S3 Para que pueda seleccionar una carpeta, todos los archivos en ella tienen que compartir el mismo tipo de archivo. Si hay diferentes esquemas, se combinarán.
Lago de datos/almacén de datos	<input type="text" value="s3://rrhh-databrew-danigayol/transformado/"/> <span>X</span> <span>🔗</span>
Amazon S3	El formato es s3://bucket/prefix
Conexiones de la base de datos	<b>Se seleccionan todos los archivos de la carpeta transformado y sus subcarpetas</b>
Amazon Redshift	S3 Buckets > rrhh-databrew-danigayol <span>Seleccionar toda la carpeta</span> <span>C</span>
JDBC	<input type="text" value="Buscar objetos de S3 por nombre"/> <span>&lt;</span> <span>1</span> <span>&gt;</span> <span>①</span>
Catálogo de datos de AWS Glue	
Tablas de S3 del catálogo de datos	
Tablas de Redshift del catálogo de datos	
Tablas de RDS del catálogo de datos	
Todas las tablas de AWS Glue	

▼ Configuraciones adicionales

Tipo de archivo seleccionado	Delimitador CSV
Formato del archivo seleccionado	<input type="text" value="Coma (,)"/> <span>▼</span>
<input checked="" type="radio"/> CSV	
<input type="radio"/> JSON	
<input type="radio"/> PARQUET	
<input type="radio"/> EXCEL	
<input type="radio"/> ORC	

Valores de encabezado de columna

- Considerar la primera fila como encabezado  
La primera fila del conjunto de datos se tratará como valores de encabezado de columna
- Agregar encabezado predeterminado  
Los encabezados predeterminados se agregarán con los valores Column\_1, Column\_2...

Le damos al botón de “crear conjunto de datos” y se nos creará:

Conjuntos de datos		Conexiones			
<b>Conjuntos de datos (3)</b>					
<input type="text" value="Buscar conjuntos de datos"/> <span>Ver detalles</span> <span>Crear proyecto con este conjunto de datos</span> <span>Ejecutar perfil de datos</span> <span>Acciones</span> <span>Conectar nuevo conjunto de datos</span>					
Nombre del conjunto de datos	Tipo de datos	Perfil de datos	Origen		
rrhh-databrew-transformado-hr1m	csv	-	S3 s3://rrhh-databrew-danigayol/transformado/ <span>🔗</span>		
rrhh-databrew-raw-hr1m	csv	Ver perfil de datos	S3 s3://rrhh-databrew-danigayol/raw/ <span>🔗</span>		
chembl-27	parquet	-	S3 s3://databrew-public-datasets-us-east-1/chembl-27.parquet <span>🔗</span>		
hace unos segundos 29 de enero de 2026, 9:44:52 am voclabs hace 2 días 27 de enero de 2026, 9:38:14 am voclabs hace 3 días 26 de enero de 2026, 9:20:54 am voclabs					

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

Ahora, nos vamos al apartado de “conjunto de reglas de calidad de los datos” y creamos uno nuevo:

The screenshot shows the AWS DataBrew console interface. On the left, there is a sidebar with the following navigation options: CONJUNTOS ..., PROYECTOS, RECETAS, REGLA DE DQ (which is highlighted with a red box), TRABAJOS, and NOVEDADES. The main content area has a breadcrumb navigation path: DataBrew > Conjuntos de reglas de calidad de datos. Below this, there is a section titled "¿Qué son los conjuntos de reglas de calidad de datos?" (What are quality rule sets?). Underneath, it says "Conjuntos de reglas de calidad de datos (0)". In the center, there is a large icon of a document with a checkmark inside a ribbon. To the right of the icon, the text reads: "Crear conjuntos de reglas para ejecutarlos en el conjunto de datos" (Create rule sets to run them in the data set). Below this, a descriptive text states: "Varias reglas relacionadas forman un conjunto de reglas que ayuda a validar los datos del conjunto de datos. Comience a crear conjuntos de reglas para ejecutarlos en sus conjuntos de datos ahora." At the bottom right, there is a yellow button with the text "Crear un conjunto de reglas de calidad de datos".

# Crear un conjunto de reglas de calidad de datos

## Detalles del conjunto de reglas

Nombre del conjunto de reglas

Identificador del conjunto de reglas

reglas\_calidad\_rrhh\_databrew\_transformado

El nombre del conjunto de reglas debe contener entre 1 y 255 caracteres. Los caracteres válidos son alfanuméricos (A-Z, a-z, 0-9), guión (-), punto (.) y espacio.

Descripción

Ingresar descripción

## Conjunto de datos asociado

Asocie un conjunto de datos con este conjunto de reglas. Para agregar reglas de calidad de datos, utilice el esquema, el perfil y las recomendaciones del conjunto de datos.

Elegir conjunto de datos

rrhh-databrew-transformado-hr1m



Explorar conjuntos de datos

[Ver los detalles del conjunto de datos asociado >](#)

Ahora saltamos al siguiente paso para añadirle las reglas

**3.) 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.

**Regla 1**  Habilitar regla

Nombre de regla

Ámbito de comprobación de calidad de los datos

Criterios de éxito de la regla

**Comprobaciones de calidad de los datos**

Comprobación 1

Comprobación de la calidad de los datos

Número de filas

Condición

Valor

Agregue otra comprobación de calidad de los datos

**Resumen de Reglas**

La regla pasará si **conjunto de datos** tiene recuento de filas == **1000000**

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.

**Hacemos lo mismo para los otros dos campos:**

## Regla 2

Habilitar regla

[Eliminar](#)

Nombre de regla

Valores\_unicos

Ámbito de comprobación de calidad de los datos

Criterios de éxito de la regla

Se cumplen todas las comprobaciones de calidad de los datos (Y) ▾

Comprobación individual de cada columna ▾

### Comprobaciones de calidad de los datos

#### Comprobación 1

##### Comprobación de la calidad de los datos

Valores únicos

Compruebe el recuento de valores únicos en la columna.

Emp ID

##### Condición

Es igual

##### Valor

100

% (porcentaje) filas

[Agregue otra comprobación de calidad de los datos](#)

### Resumen de Reglas

La regla pasará si Emp ID tiene valores únicos == 100%

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.

**Regla 3**  Habilitar regla [Eliminar](#)

Nombre de regla  
**Valores\_no\_nulos**

Ámbito de comprobación de calidad de los datos  
Criterios de éxito de la regla  
**Se cumplen todas las comprobaciones de calidad de los datos (Y)**

Comprobación individual de cada columna

**Comprobaciones de calidad de los datos**

Comprobación 1

Comprobación de la calidad de los datos

Valores faltantes  
Compruebe los valores que faltan en la columna.  
**Emp ID**

Condición  
**Es igual**

Valor  
**0** % (porcentaje) filas

Agregue otra comprobación de calidad de los datos

**Resumen de Reglas**  
La regla pasará si **Emp ID** tiene valores faltantes == **0%**

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.

**Regla 4**       Habilitar regla      [Eliminar](#)

Nombre de regla  
**Valores\_positivos\_y\_rango\_edad**

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

**Comprobaciones de calidad de los datos**

Comprobación 1

Comprobación de la calidad de los datos

**Valores válidos**  
Compruebe si hay valores válidos en la columna.

**Emp ID**

Condición

**Mayor que igual**

Valor

**0**      filas

Agregue otra comprobación de calidad de los datos

**Resumen de Reglas**  
La regla pasará si **Emp ID** tiene recuento de valores válidos  $\geq 0$

#### Comprobación de la calidad de los datos

Valores válidos

Compruebe si hay valores válidos en la columna.



Age in Yrs.



#### Condición

Mayor que igual



#### Valor

0

filas



#### Comprobación 3

[Eliminar](#)

#### Comprobación de la calidad de los datos

Valores válidos

Compruebe si hay valores válidos en la columna.



Age in Yrs.



#### Condición

Menor que igual



#### Valor

80

filas



[Agregue otra comprobación de calidad de los datos](#)

#### Resumen de Reglas

La regla pasará si **Age in Yrs.** tiene recuento de valores válidos  $\geq 0$  Y tiene recuento de valores válidos  $\leq 80$  Y **Emp ID** tiene recuento de valores válidos  $\geq 0$

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).

The screenshot shows the configuration of a data quality rule. The rule is named "Verificar\_expresion\_regular". It is set to check individual columns and requires all data quality checks to be successful. The first check is for SSN values matching the regex `^\d{3}-\d{2}-\d{4}$`. The threshold is set to 100%.

Nombre de regla	Verificar_expresion_regular
Ámbito de comprobación de calidad de los datos	Comprobación individual de cada columna
Criterios de éxito de la regla	Se cumplen todas las comprobaciones de calidad de los datos (Y)
<b>Comprobaciones de calidad de los datos</b>	
Comprobación 1	
Comprobación de la calidad de los datos	Valores de cadena Compruebe en la columna los valores de cadena en función de la...
SSN	
Condición	Coincidencias (patrón RegEx)
Valor de RegEx	<code>^\d{3}-\d{2}-\d{4}\$</code>
<b>Agregue otra comprobación de calidad de los datos</b>	
<b>Umbral</b>	
Definir el umbral de filas que deben cumplir las comprobaciones antes de que se supere la regla	
Condición	Umbral
Mayor que igual	100 % (porcentaje) filas

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

Una vez configurado todo, le damos al botón de “crear” y nos aparecerá de la siguiente manera, sin tener asociado ningún trabajo:

The screenshot shows the management of data quality rule sets. A single rule set named "reglas-calidad-rrhh-databrew-transformado" is listed, created by "user3935192=Daniel\_Gayol\_Rodríguez" just now. It is associated with the dataset "rrhh-databrew-transformado" and has 5 rules.

Conjuntos de reglas de calidad de datos (1)						
Nombre del conjunto de reglas de calidad de datos		Descripción		Conjunto de datos asociado	Trabajo Asociado	Fecha de creación
<input type="checkbox"/>	reglas-calidad-rrhh-databrew-transformado	-	-	rrhh-databrew-transformado-hr1m	-	hace unos segundos 29 de enero de 2020, 10:50:52 am



# 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.**

Para ello, nos vamos a “S3” y como hicimos en anteriores pasos, creamos otra carpeta nueva con el nombre de “calidad”:

## Crear carpeta Información

Utilice carpetas para agrupar los objetos en buckets. Al crear una carpeta, S3 creará un ob]

**ⓘ Su política de bucket podría bloquear la creación de carpetas**

Si su política de bucket impide cargar objetos sin etiquetas, metadatos o beneficiari [carga](#) para cargar una carpeta vacía y especificar la configuración adecuada.

### Carpeta

Nombre de la carpeta

Los nombres de las carpetas no pueden contener "/". [Consulte las reglas de nomenclatura ↗](#)

✓ Se creó correctamente la carpeta “calidad”

## rrhh-databrew-danigayol [Información](#)

Objetos Metadatos Propiedades Permisos Métricas Administración

### Objetos (4)



Copiar URI de S3

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario](#) que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

<input type="checkbox"/>	Nombre	<input type="checkbox"/>	Tipo
<input type="checkbox"/>	<a href="#">calidad/</a>	<input type="checkbox"/>	Carpeta
<input type="checkbox"/>	<a href="#">perfil/</a>	<input type="checkbox"/>	Carpeta
<input type="checkbox"/>	<a href="#">raw/</a>	<input type="checkbox"/>	Carpeta
<input type="checkbox"/>	<a href="#">transformado/</a>	<input type="checkbox"/>	Carpeta

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

Ahora nos vamos a “AWS Glue Databrew” y entramos en el apartado de “reglas de calidad” y asociamos las reglas a un trabajo de perfil

La captura de pantalla muestra la interfaz de AWS Glue DataBrew. En el menú lateral, el ícono de 'Reglas de DQ' (que tiene un cuadro rojo) está resaltado. La sección central titulada 'Conjuntos de reglas de calidad de datos' muestra una lista con un solo elemento: 'reglas-calidad-rrhh-databrew-transformado'. A la derecha de la lista, hay un botón 'Crear trabajo de perfil con conjunto de reglas' que también tiene un cuadro rojo. Los demás íconos en el menú lateral ('CONJUNTOS...', 'PROYECTOS', 'RECETAS', 'TRABAJOS', 'NOVEDADES') no están resaltados.

a) Aplica el trabajo a todo el dataset:

Crear trabajo

**Detalles del trabajo**

Nombre del trabajo  
Identificador del trabajo

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.

**Ejemplo de ejecución de trabajo**  
Un trabajo se puede ejecutar en todo el conjunto de datos o en una muestra personalizada del conjunto de datos.

Muestra de datos  
Definir el ámbito del conjunto de datos en el que se va a ejecutar el trabajo

Conjunto de datos completo  
 Ejemplo personalizado

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

**Configuración de salida del trabajo**  
La ejecución de un trabajo genera archivos de salida en los destinos de archivo especificados.

Cuenta de AWS del propietario del bucket de S3

Cuenta de AWS actual  
043356869404

Otra cuenta de AWS

Tipo de archivo  
Formato de salida  
JSON

Ubicación de S3  
El formato es s3://bucket/folder/  
 X Explorar

Cifrado

Habilitar el cifrado para el archivo de salida del trabajo  
Cifrar el archivo de salida del trabajo con SSE-S3 o AWS KMS

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

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.

### 3.) Verifica el nombre del trabajo de perfil asociado a las reglas:

DataBrew > Conjuntos de reglas de calidad de datos

▶ ¿Qué son los conjuntos de reglas de calidad de datos?

Conjuntos de reglas de calidad de datos (1)

Nombre del conjunto de reglas de calidad de datos	Descripción	Conjunto de datos asociado	Trabajo Asociado	Fecha de creación
reglas-calidad-rrhh-databrew-transformado-hr1m 5 reglas	-	rrhh-databrew-transformado-hr1m	rrhh-databrew-transformado-hr1m profile job	hace 4 días 29 de enero de 2026, 10:30:32 am

Ahora nos vamos a “trabajos de perfil” y lo ejecutamos:

DataBrew > Trabajos

Trabajos de recetas | **Trabajos de perfil** | Programaciones

Trabajos de perfil (2)

Nombre del trabajo	Estado de la última ejecución del trabajo	Conjunto de datos	Perfil de datos	Última ejecución	Creado el	Creado por	Etiquetas
rrhh-databrew-transformado-hr1m profile job	-	rrhh-databrew-	Ver perfil de datos	-	hace 2 minutos 2 de febrero de 2026, 9:01:30 am	voclabs	-
perfil-datos-databrew	Realizado con éxito	rrhh-databrew-	Ver perfil de datos	hace 6 días 27 de enero de 2026, 10:05:17 am	hace 6 días 27 de enero de 2026, 10:00:28 am	voclabs	-

Nombre del trabajo	Estado de la última ejecución del trabajo	Conjunto de datos	Perfil de datos	Última ejecución	Creado el	Creado por	Etiquetas
rrhh-databrew-transformado-hr1m profile job	Realizado con éxito	rrhh-databrew-	Ver perfil de datos	hace un minuto 2 de febrero de 2026, 9:08:14 am	hace 8 minutos 2 de febrero de 2026, 9:01:30 am	voclabs	-
perfil-datos-databrew	Realizado con éxito	rrhh-databrew-	Ver perfil de datos	hace 6 días 27 de enero de 2026, 10:05:17 am	hace 6 días 27 de enero de 2026, 10:00:28 am	voclabs	-

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.

Trabajos de perfil (2)							
Buscar trabajos		Mostrar todo		Acciones		Crear trabajo	
Nombre del trabajo	Estado de la última ejecución del trabajo	Conjunto de datos	Perfil de datos	Última ejecución	Creado el	Creado por	Etiquetas
<input checked="" type="checkbox"/> rrhh-databrew-transformado-hr1m profile job	Realizado con éxito	rrhh-databrew-	Ver perfil de datos	hace 2 minutos 2 de febrero de 2026, 9:08:14 am	hace 9 minutos 2 de febrero de 2026, 9:01:30 am	voclabs	-
<input type="checkbox"/> perfil-datos-databrew	Realizado con éxito	rrhh-databrew-	Ver perfil de datos	hace 6 días 27 de enero de 2026, 10:05:17 am	hace 6 días 27 de enero de 2026, 10:00:28 am	voclabs	-

rrhh-databrew-transformado-hr1m

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

### Reglas de calidad de los datos (5)

Expandir todo | Contraer todo | Buscar

**TODOS** (5) **REALIZADO CON ÉXITO** (2) **FALLO** (3) **ERROR** (0) **DESACTIVADO** (0)

**reglas-calidad-rrhh-databrew-transformado** 5 reglas **Fallo**

- ☒ Recuento\_filas**  
Comprobar si **conjunto de datos** tiene recuento de filas == **1000000**
- ☒ Valores\_unicos**  
Comprobar si **Emp ID, E Mail, SSN** tiene valores únicos == **100%**
- ✓ Valores\_no\_nulos**  
Comprobar si **Emp ID** tiene valores faltantes == **0%**
- ☒ Valores\_positivos\_y\_rango\_edad**  
Comprobar si **Age in Yrs.** tiene recuento de valores válidos >= **0** Y tiene recuento de valores válidos <= **80** Y **Emp ID** tiene recuento de valores válidos >= **0**
- ✓ Verificar\_expresion\_regular**  
Comprobar si **SSN** tiene valores coincidencias **^\d{3}-\d{2}-\d{4}\$** PARA mayor o igual que 100% de **100%** Realizado con éxito **0%** Fallo

## 5.- De aparecer algún error vete a la pestaña de Estadísticas de columna y comprueba que es cierto el error de las reglas.

rrhh-databrew-transformado-hr1m

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

## Valores distintivos principales

Profile devuelve principal 50 valores distintivos y principal 50 valores atípicos en el conjunto de datos

 Buscar

 VÁLIDO  VALORES ATÍPICOS

255047		3	<1%
372733		2	<1%
954733		2	<1%
389783		2	<1%
407135		2	<1%
974833		2	<1%
174305		2	<1%
377519		2	<1%
283219		2	<1%
Otros		19,98 K	99%

[Ver los principales 50 valores distintivos](#)

# Apartado E

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

## Crear carpeta Información

Utilice carpetas para agrupar los objetos en buckets. Al crear una carpeta, S3 creará un objeto

**(i) Su política de bucket podría bloquear la creación de carpetas**

Si su política de bucket impide cargar objetos sin etiquetas, metadatos o beneficiarios [carga](#) para cargar una carpeta vacía y especificar la configuración adecuada.

### Carpeta

Nombre de la carpeta

curated

Los nombres de las carpetas no pueden contener "/". [Consulte las reglas de nomenclatura](#) ↗

✓ Se creó correctamente la carpeta "curated"

## rrhh-databrew-danigayol Información

Objetos

Metadatos

Propiedades

Permisos

Métricas

Administración

Puntos de

### Objetos (5)



[Copiar URI de S3](#)

[Copiar URL](#)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) ↗ para concederles permisos de forma explícita. [Más información](#) ↗

Buscar objetos por prefijo

<input type="checkbox"/>	Nombre	▲   Tipo	▼   Última modificación
<input type="checkbox"/>	<a href="#">calidad/</a>	Carpeta	-
<input type="checkbox"/>	<a href="#">curated/</a>	Carpeta	-
<input type="checkbox"/>	<a href="#">perfil/</a>	Carpeta	-
<input type="checkbox"/>	<a href="#">raw/</a>	Carpeta	-
<input type="checkbox"/>	<a href="#">transformado/</a>	Carpeta	-

**2.) 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. El resultado del trabajo almacénalo en formato parquet comprimido en la carpeta curated.**

Ahora para crear un proyecto, nos vamos a “AWS Glue Databrew” y en el apartado de “proyectos le damos a crear un nuevo proyecto:

The screenshot shows the AWS Glue DataBrew interface. On the left, there's a sidebar with icons for 'CONJUNTOS DE DATOS' (Data Sets), 'RECETAS' (Recipes), and 'REGLAS DE DQ' (DQ Rules). The main area is titled 'Proyectos (2)' and lists two existing projects: 'rrhh-databrew-proyecto' and 'Sample project - 1'. A red box highlights the 'Crear proyecto' (Create project) button at the top right of the table header.

**Crear proyecto**

**Detalles del proyecto**

**Nombre del proyecto**  
rrhh-databrew-proyecto-correcionErrores

El nombre del proyecto debe contener entre 1 y 255 caracteres. Los caracteres válidos son alfanuméricos (A-Z, a-z, 0-9), guiones (-) y espacios.

**Detalles de la receta**

Los pasos de limpieza de datos en DataBrew se almacenan como una receta. Una receta está conectada a un proyecto de forma predeterminada. L

**Receta asociada** Crear nueva receta ▼ **Nombre de la receta** rrhh-databrew-proyecto-correcionErrores-recipe

El nombre de la receta debe contener entre 1 y 255 caracteres. Los caracteres válidos son alfanuméricos (A-Z, a-z, 0-9), guiones (-) y espacios.

**Importar pasos de la receta**  
Importe los pasos de una receta existente al proyecto. La receta elegida existente no se editará.

**Seleccionar un conjunto de datos**

Seleccionar el conjunto de datos en el que desea trabajar

Mis conjuntos de datos Los conjuntos de datos importados

Archivos de muestra Explorar archivos de ejemplo para el conjunto de datos

Nuevo conjunto de datos Importar nuevo conjunto de datos

Buscar conjuntos de datos

Nombre del conjunto de datos	Tipo de datos	Origen	Fecha de creación
rrhh-databrew-transformado-hr1m	csv	S3	hace 4 días 29 de enero de 2026, 9:44:52 am
rrhh-databrew-raw-hr1m	csv	S3	hace 6 días 27 de enero de 2026, 9:38:14 am
chembl-27	parquet	S3	hace 7 días 26 de enero de 2026, 9:20:34 am

**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.  
   C

ⓘ En cuanto cree un proyecto DataBrew, se abrirá el proyecto y los costos comenzarán a acumularse en su cuenta de AWS. [Información sobre precios](#)

Cancelar Crear proyecto

## Ahora en la receta, vamos a corregir los errores:

### Eliminar duplicados

**Columna de origen**  
Seleccionar una columna de origen para eliminar duplicados

**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 cuadricula

---

[Vista previa de los cambios](#)

Cancelar Aplicar

**Receta (3)**

rrhh-databrew-proyecto-correcionErrores-recipe  
Versión de trabajo

**Publicar** (button highlighted with a red box)

**Más**

Pasos aplicados (3) | Borrar todo

1. Eliminar duplicados de Emp ID  
2. Eliminar duplicados de SSN  
3. Eliminar duplicados de E Mail

CONSUMOS	PROYECTOS	REGLAS DE DÍA	TRABAJOS	Trabajos de recetas	Trabajos de perfil	Programaciones	Ver detalles	Ejecutar trabajo	Acciones	Crear trabajo
				Trabajos de recetas (2)						

Buscar trabajos Mostrar todo ▾

Nombre del trabajo	Estado	Entrada del trabajo	Salida del trabajo	Última ejecución	Creado el	Creado por	Etiquetas
DataBrew-transformado-danigayol	Realizado con éxito	rrhh-databre... Conjunto de datos	rrhh-databre... Receta	1 salida 29 de enero de 2026, 9:18:32 am	hace 4 días 29 de enero de 2026, 9:00:16 am	voclabs	-
DataBrew-danigayol	Realizado con éxito	chembi-27 Conjunto de datos	Sample recip... Receta	2 salidas 26 de enero de 2026, 10:08:57 am	hace 7 días 26 de enero de 2026, 10:06:50 am	voclabs	-

## Crear trabajo

### Detalles del trabajo

Nombre del trabajo  
Identificador del trabajo

**DataBrew-correcionErrores-danigayol** (input field highlighted with a red box)

El nombre del trabajo debe contener entre 1 y 240 caracteres. Los car

### Tipo de trabajo

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

<input checked="" type="radio"/> <b>Crear un trabajo de receta</b> Ejecuta las transformaciones de la receta asociada en la población del conjunto de datos asociado.	<input type="radio"/> <b>Crear un trabajo de perfil</b> 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

<input checked="" type="radio"/> <b>Conjunto de datos</b> Ejecute el trabajo en un conjunto de datos DataBrew existente o nuevo.	<input type="radio"/> <b>Proyecto</b> Ejecute el trabajo en un proyecto sin trabajo asociado.
---	--

Elegir conjunto de datos

Seleccionar una receta

Versión de la receta

### Configuración de salida del trabajo

La ejecución de un trabajo genera archivos de salida en los destinos de archivo especificados.

Salida 1

Salida a Ubicación de la salida	Tipo de archivo Formato de salida	Delimitador Separador CSV	Compresión Tipos disponibles
<input type="button" value="Amazon S3"/> <input type="button" value="▼"/>	<input type="button" value="PARQUET"/> <input type="button" value="▼"/>	<input type="button" value="Coma (,)"/> <input type="button" value="▼"/>	<input type="button" value="Snappy"/> <input type="button" value="▼"/>

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/

### 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.

Historial de la ejecución del trabajo						Detener la ejecución del trabajo	Acciones
ID de ejecución de trabajo	Estado de la última ejecución del trabajo	Tiempo de ejecución	Salida	Resumen	Iniciado por	Iniciado el	
DataBrew-correcionErrores-danigayol_2026-02-02-09:55:10	Realizado con éxito	2 minutos, 36 segundos	1 salida		user3935192:Daniel_Gayol_Rodríguez	hace 4 minutos 2 de febrero de 2026, 9:55:10 am	

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

Conjuntos de datos (3)								Ver detalles	Crear proyecto con este conjunto de datos	Ejecutar perfil de datos	Acciones	Conectar nuevo conjunto de datos
	Nombre del conjunto de datos	Tipo de datos	Perfil de datos	Origen	Ubicación	Fecha de creación	Creado por	Etiquetas				
<input type="checkbox"/>	rrhh-databrew-transformado-hr1m	csv	Ver perfil de datos	S3	s3://rrhh-databrew-danigayol/transformado/	hace 4 días 29 de enero de 2026, 9:44:52 am	voclabs	-				
<input type="checkbox"/>	rrhh-databrew-raw-hr1m	csv	Ver perfil de datos	S3	s3://rrhh-databrew-danigayol/raw/	hace 6 días 27 de enero de 2026, 9:58:14 am	voclabs	-				
<input type="checkbox"/>	chembl-27	parquet	-	S3	s3://databrew-public-datasets-us-east-1/chembl-27.parquet	hace 7 días 26 de enero de 2026, 9:20:54 am	voclabs	-				

## Nueva conexión

### Detalles del nuevo conjunto de datos

Nombre del conjunto de datos

rrhh-databrew-correcionErrores-hr1m

El nombre del conjunto de datos debe contener entre 1 y 255 caracteres. L

**Conectarse a un nuevo conjunto de datos**

- 
- Lago de datos/almacén de datos
- Amazon S3**
- Conexiones de la base de datos
- Amazon Redshift
- JDBC
- Catálogo de datos de AWS Glue
- Tablas de S3 del catálogo de datos
- Tablas de Redshift del catálogo de datos
- Tablas de RDS del catálogo de datos

Introducir el origen desde S3  
Para que pueda seleccionar una carpeta, todos los archivos en ella tienen que compartir el mismo tipo de archivo. Si hay diferentes esquemas, se combinarán.

**s3://rrhh-databrew-danigayol/curated/DataBrew-correcionErrores-danigayol\_02Feb2026\_1770022600535/**

El formato es s3://bucket/prefix

**Se seleccionan todos los archivos de la carpeta DataBrew-correcionErrores-danigayol\_02Feb2026\_1770022600535 y sus subcarpetas**

S3 Buckets > rrhh-databrew-danigayol > curated

Nombre	Tamaño	Última actualización
<b>DataBrew-correcionErrores-danigayol_02Feb2026_1770022600535</b>	-	-

▼ Configuraciones adicionales

Tipo de archivo seleccionado  
Formato del archivo seleccionado

CSV

JSON

**PARQUET**

EXCEL

ORC

asignar a los recursos de AWS. Cada etiqueta es una etiqueta sencilla que consta de una clave definida por el cliente (nombre) y un valor opcional. El uso de etiquetas puede facilitarle la filtrado de recursos por finalidad, propietario, entorno u otros criterios.

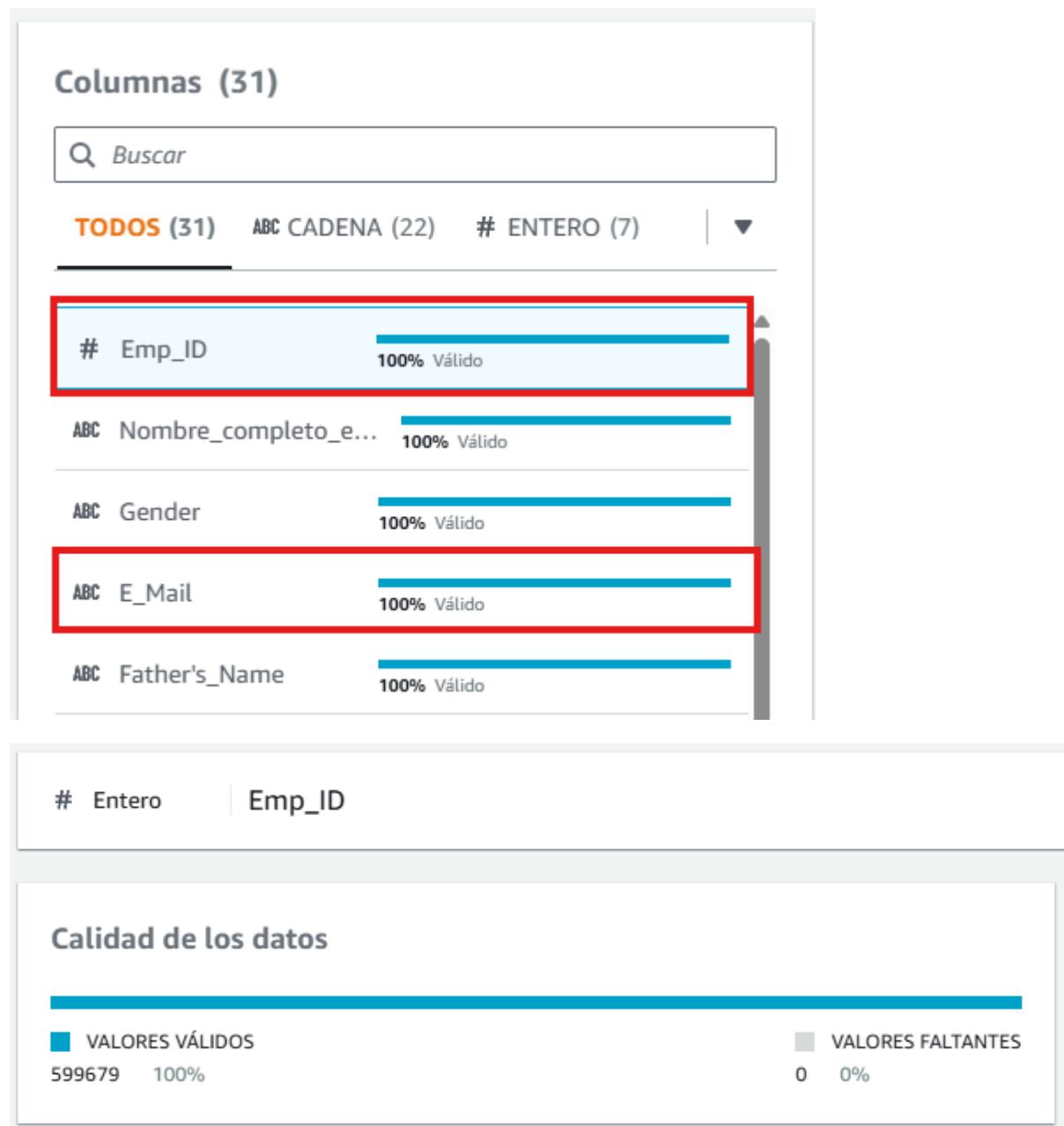
## 4.) Verifica ahora las estadísticas de las columnas que has modificado para asegurarnos que todo ha ido bien. ¿Cuántas filas tiene ahora el archivo resultante?

DataBrew > Conjuntos de datos

**Conjuntos de datos** (4)

Nombre del conjunto de datos	Tipo de datos	Perfil de datos	Origen	Ubicación	Fecha de creación	Creado por	Etiquetas
<input checked="" type="checkbox"/> <b>rnh-databrew-correcionErrores-hr1m</b>	parquet	-	S3	s3://rrhh-databrew-danigayol/curated/DataBrew-correcionErrores-danigayol_02Feb2026_1770022600535/	hace 3 minutos 2 de febrero de 2026, 10:06:27 am	voclabs	-
<input type="checkbox"/> rnh-databrew-transformado-hr1m	csv	Ver perfil de datos	S3	s3://rrhh-databrew-danigayol/transformado/	hace 4 días 29 de enero de 2026, 9:44:52 am	voclabs	-
<input type="checkbox"/> rnh-databrew-raw-hr1m	csv	Ver perfil de datos	S3	s3://rrhh-databrew-danigayol/raw/	hace 6 días 27 de enero de 2026, 9:38:14 am	voclabs	-
<input type="checkbox"/> chembl-27	parquet	-	S3	s3://databrew-public-datasets-us-east-1/chembl-27.parquet	hace 7 días 26 de enero de 2026, 9:20:54 am	voclabs	-

Una vez ejecutado, ya podemos revisar los datos obtenidos:



## Valores distintivos principales

Profile devuelve principal 50 valores distintivos y principal 50 valores atípicos en el conjunto de datos

 Buscar

 VÁLIDO  VALORES ATÍPICOS

879388		1	<1%
894921		1	<1%
506112		1	<1%
254705		1	<1%
715766		1	<1%
663710		1	<1%
545982		1	<1%
979199		1	<1%
143535		1	<1%
Otros		599,67 K	99%

[Ver los principales 50 valores distintivos](#)

Ahora el número total de filas es inferior, debido a la eliminación de valores duplicados

## Resumen

TOTAL DE FILAS  
599.679

TOTAL DE COLUMNAS  
31

# 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.**

**Entramos en el apartado de “reglas de calidad de datos” y duplicamos el conjunto de reglas del apartado C:**

Nombre del conjunto de reglas de calidad de datos	Descripción	Conjunto de datos asociado	Trabajo Asociado	Fecha de creación	Última ejecución	Propietario
reglas-calidad-rrhh-databrew-transformado	-	rrhh-databrew-transformado-hr1m	rrhh-databrew-transformado-hr1m profile job	hace 6 días 29 de enero de 2026, 10:30:32 am	use3935192-Daniel_Gayol_Ro	dr_guez

**Ahora lo hacemos apuntar hacia los datos de la carpeta “curated”:**

## Crear un conjunto de reglas de calidad de datos

### Detalles del conjunto de reglas

Nombre del conjunto de reglas

Identificador del conjunto de reglas

**reglas-calidad-rrhh-databrew-transformado-curated**

El nombre del conjunto de reglas debe contener entre 1 y 255 caracteres. Los caracteres válidos son alfanuméricos (A-Z, a-z, 0-9), guión (-), punto (.) y espacio.

Descripción

Ingresar descripción

### Conjunto de datos asociado

Asocie un conjunto de datos con este conjunto de reglas. Para agregar reglas de calidad de datos, utilice el esquema, el perfil y las recomendaciones del conjunto de datos.

Elegir conjunto de datos

**rrhh-databrew-correcionErrores-hr1m**

Explorar conjuntos de datos

[Ver los detalles del conjunto de datos asociado >](#)

**Y ahora cambiamos algunos valores para adaptarlo:**

**Regla 1**

Habilitar regla  Eliminar

Nombre de regla  
Recuento\_filas

Ámbito de comprobación de calidad de los datos  
Comprobación individual de cada columna ▾

Criterios de éxito de la regla  
Se cumplen todas las comprobaciones de calidad de los datos (Y) ▾

**Comprobaciones de calidad de los datos**

Comprobación 1

Comprobación de la calidad de los datos

Número de filas  
Compruebe el conjunto de datos para el número total de filas.

Condición  
Es igual

Valor  
599679

Agregue otra comprobación de calidad de los datos

**Resumen de Reglas**

La regla pasará si **conjunto de datos** tiene recuento de filas == **599679**

**Y los nombres de los campos siguen siendo los mismo, por lo tanto, la podemos crear sin ningún tipo de problema:**

Conjuntos de reglas de calidad de datos (2)						
<input type="text"/> Buscar conjuntos de reglas		Crear trabajo de perfil con conjuntos de reglas		Acciones		Crear un conjunto de reglas de calidad de datos
Nombre del conjunto de reglas de calidad de datos	Descripción	Conjunto de datos asociado	Trabajo Asociado	Fecha de creación	Creado por	Etiquetas
reglas-calidad-rrhh-databrew-transformado-curated 5 reglas	-	rrhh-databrew-correccionErros-hr1m	-	hace unos segundos 4 de febrero de 2026, 11:06:09 am	user3935192:Daniel_Gayol_Ro dr_guez	-
reglas-calidad-rrhh-databrew-transformado 5 reglas	-	rrhh-databrew-transformado-hr1m	rrhh-databrew-transformado-hr1m profile job	hace 6 días 29 de enero de 2026, 10:30:32 am	user3935192:Daniel_Gayol_Ro dr_guez	-

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

Ahora tenemos que asociarle un conjunto de reglas al trabajo de perfil:

Conjuntos de reglas de calidad de datos								
Conjuntos de reglas de calidad de datos (2)								
Nombre del conjunto de reglas de calidad de datos	Descripción	Conjunto de datos asociado	Trabajo Asociado	Fecha de creación	Creado por	Etiquetas	Acciones	Crear un conjunto de reglas de calidad de datos
<input checked="" type="checkbox"/> reglas-calidad-rrhh-databrew-transformado-curated 5 reglas	rrhh-databrew-correcionErrores-hr1m	-	-	hace unos segundos 4 de febrero de 2026, 11:06:09 am	user3935192=Daniel_Gayol_Rodríguez	-		
<input type="checkbox"/> reglas-calidad-rrhh-databrew-transformado 5 reglas	rrhh-databrew-transformado-hr1m	rrhh-databrew-transformado-hr1m profile job	-	hace 6 días 29 de enero de 2026, 10:30:52 am	user3935192=Daniel_Gayol_Rodríguez	-		

## Editar rrhh-databrew-correcionErrores-hr1m profile job

### Tipo de trabajo



#### Trabajo de perfil

Un trabajo de perfil genera un resumen y estadísticas que le dan la forma de los datos.

### Conjunto de datos asociado

#### rrhh-databrew-correcionErrores-hr1m

S3 | s3://rrhh-databrew-danigayol/curated/DataBrew-correcionErrores-danigayol\_02Feb2026\_1770022600535/

### Ejemplo de ejecución de trabajo

Un trabajo se puede ejecutar en todo el conjunto de datos o en una muestra personalizada del conjunto de datos.

### Muestra de datos

Definir el ámbito del conjunto de datos en el que se va a ejecutar el trabajo

- Conjunto de datos completo
- Ejemplo personalizado

### Configuración de salida del trabajo

La ejecución de un trabajo genera archivos de salida en los destinos de archivo especificados.

#### Cuenta de AWS del propietario del bucket de S3

- Cuenta de AWS actual

043356869404

- Otra cuenta de AWS

#### Tipo de archivo

Formato de salida

JSON

#### Ubicación de S3

El formato es s3://bucket/folder/

s3://rrhh-databrew-danigayol/curated/DataBrew-correcionErrores-danigayol\_02Feb2026\_1770022600535/

Explorar

- Habilitar el cifrado para el archivo de salida del trabajo

Cifrar el archivo de salida del trabajo con SSE-S3 o AWS KMS

**Reglas de calidad de los datos - opcional**  
Valide su perfil de datos sobre conformidad y reglas empresariales.

**Conjuntos de reglas de calidad de datos aplicados**

Nombre del conjunto de reglas de calidad de datos	Descripción	Remover
reglas-calidad-rrhh-databrew-transformado-curated	-	Remover
5 reglas		

**Aplicar otro conjunto de reglas de calidad de datos**

**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	
---------	--

**Cancelar** **Guardar**

### 3.) Ejecuta dicho trabajo contra todo el dataset.

Ahora vamos a ejecutar el trabajo:

Nombre del trabajo	Estado de la última ejecución del trabajo	Conjunto de datos	Perfil de datos	Última ejecución	Creado el	Creado por	Etiquetas
<input checked="" type="checkbox"/> rrhh-databrew-correcionErrores-hr1m profile job	Realizado con éxito	rrhh-databrew	Ver perfil de datos	hace 2 días 2 de febrero de 2026, 10:14:55 am	hace 2 días 2 de febrero de 2026, 10:11:10 am	voclabs	-
<input type="checkbox"/> rrhh-databrew-transformado-hr1m profile job	Realizado con éxito	rrhh-databrew	Ver perfil de datos	hace 2 días 2 de febrero de 2026, 9:08:14 am	hace 2 días 2 de febrero de 2026, 9:01:30 am	voclabs	-
<input type="checkbox"/> perfil-datos-databrew	Realizado con éxito	rrhh-databrew	Ver perfil de datos	hace 8 días 27 de enero de 2026, 10:05:17 am	hace 8 días 27 de enero de 2026, 10:00:28 am	voclabs	-

**Ejecutar trabajo: rrhh-databrew-correcionErrores-hr1m profile job**

**Ubicaciones de salida**  
La ejecución de un trabajo genera archivos de salida en los siguientes destinos:

Cuenta de AWS del propietario del bucket de S3	Tipo de archivo
	Formato de salida
<input checked="" type="radio"/> Cuenta de AWS actual 043356869404	JSON
<input type="radio"/> Otra cuenta de AWS	

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

s3://rrhh-databrew-danigayol/curated/DataBrew-correcionErrores-dani |

**Muestra de datos**  
Definir el ámbito del conjunto de datos en el que se va a ejecutar el trabajo

Conjunto de datos completo  
 Ejemplo personalizado  
El valor debe ser mayor que cero

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