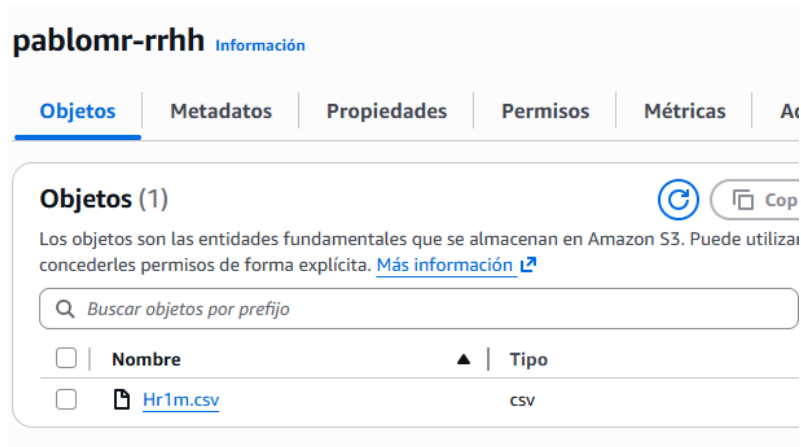


# Databrew II

## 1) Apartado A

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



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

a) ¿Cuánto ocupa el archivo?

En ocupa 265.2 MB

Detalles del conjunto de datos			
Nombre del conjunto de datos rrhh-Hr1m.csv	Tamaño de los datos 265.2 MB	Proyectos asociados databrew-rrhh	Trabajos asociados -
Origen de datos S3	Ubicación de S3 s3://pablomr-rrhh/Hr1m.csv		

b) Haz una captura que nos muestre el tipo y contenido de las 5 primeras filas de algunas de las columnas.

Vista previa del conjunto de datos		<div> Cuadrícula Esquema Texto Árbol </div>			
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				

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

**pablomr-rrhh** Información

**Objetos** | Metadatos | Propiedades | Permisos

**Objetos (1/2)**

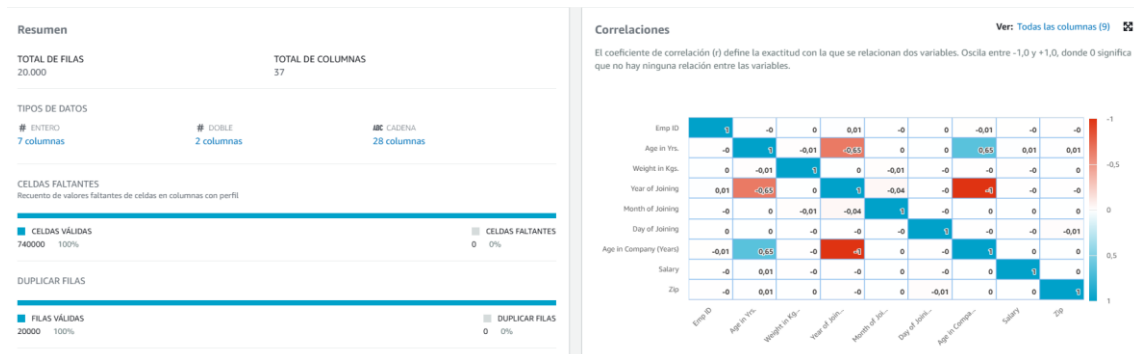
Los objetos son las entidades fundamentales que se almacenan en Ar que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

Nombre	Tipo
Hr1m.csv	csv
perfil/	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?

Perfil de datos:



Se utilizaron 20.000 filas.

Última ejecución de trabajo ✔ Realizado con éxito hace 5 minutos, no hay ejecuciones de trabajos programadas  
El perfil de datos se ha ejecutado en muestra personalizada de las primeras 20.000 filas de su conjunto de datos

- 5) Analiza los datos obtenidos.

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

Son 37 columnas, 28 son de tipo cadena de carácter y 9 de números enteros.

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

En cuanto a correlación positiva, destacan “Age in Years” y “Age in company”. Por otro lado, en correlación negativa destacan “Years of Joining” y “Age in company”



c) ¿Qué porcentaje de hombres y mujeres hay?

En la sección estadísticas de columnas seleccionamos “gender” y observamos:

### Valores distintivos principales

El perfil devuelve los principal 50 valores distintivos principales del conjunto de datos

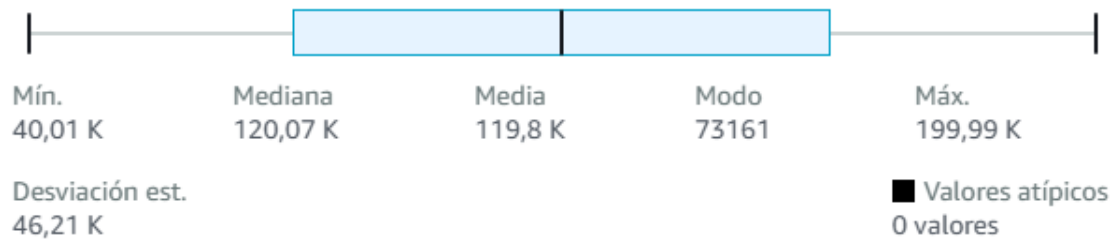
Buscar

VÁLIDO

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

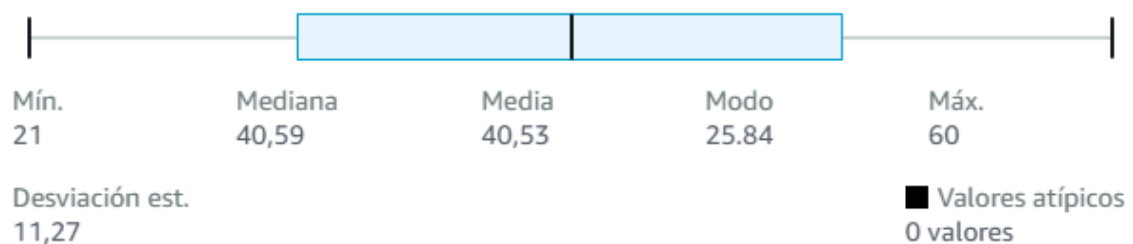
d) Analizando el diagrama de cajas de los salarios ¿En qué horquilla se mueven? ¿Cuál es la media, mediana y moda? ¿Están distribuidos simétricamente?

Se mueven entre 40.01 y 199.99. La media es 119.8, la mediana 120.07 y la moda 73161. Están distribuidos simétricamente.



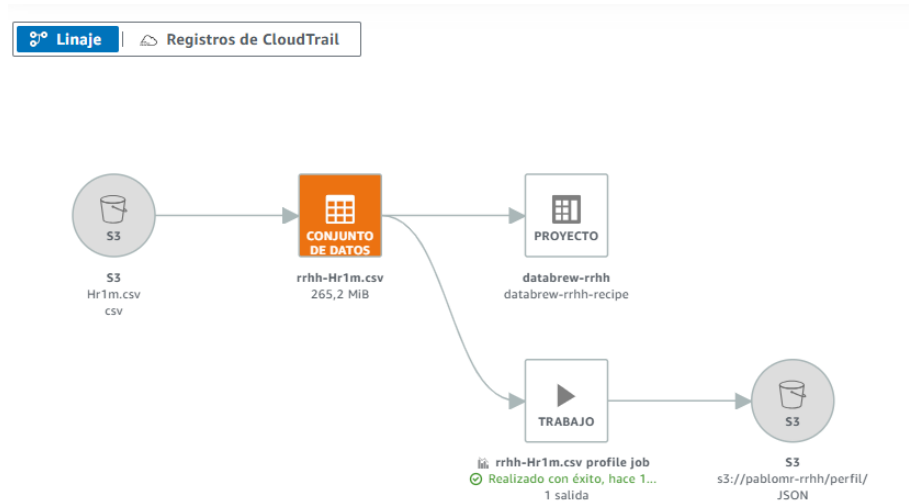
e) Busca la misma información para el campo Age in years.

Se mueven entre 21 y 60. La media es 40.53, la mediana 40.59 y la moda 25.84. Están distribuidos simétricamente.



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

Se muestra con que elementos de AWS se relaciona el dataset.



## 2) 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?

Se han utilizado 500 registros.

databrew-rrhh

Conjunto de datos: rrhh-Hr1m.csv Muestra: Muestra de los primeros n (500 filas)

Visualizando 37 columnas 500 filas

#	Emp ID	ABC Name Prefix	ABC First Name	ABC Middle Initial	ABC Last Name
1	149821	Mrs.	Jeffrey	C	Murakami
2	129350	Mr.	Shelby	D	Davidson
3	702166	Dr.	Wien	P	Russo
4	182838	Prof.	Aaron	Q	Delima
5	165681	Mr.	Frederic	M	Christofferso
6	131466	Hon.	Billie	M	Lachapelle
7	135618	Mrs.	Roseline	M	Bach
8	151970	Ms.	Jerlene	H	Chalk
9	705378	Dr.	Marcella	Q	Payan
10	193321	Mrs.	Sylvie	X	Pautz
11	140547	Mr.	Oswaldo	S	Swayne
12	121254	Mr.	Devon	E	Kehoe
13	167533	Ms.	Brigitte	G	Tong
14	763494	Dr.	Mauricio	F	Ryles
15	1137884	Hon.	Fidela	N	Norden
16	150120	Hon.	Isalias	A	Dibenedetto
17	129416	Mr.	Landon	K	Stolz

- 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*.

Fusionar columnas

Columna de origen  
Selecione dos o más columnas en el orden de fusión

⋮	Name Prefix	×
⋮	First Name	×
⋮	Middle Initial	×
⋮	Last Name	×

Agregar una columna

Separador - Opcional  
Los valores concatenados están separados por este

Nombre de la columna nueva  
Nombre de la columna de destino con la que se va a fusionar

Nombre\_completo\_empleado

Los caracteres válidos son alfanuméricos, guiones bajos y espacios

Aplicar transformación a

☒ Todas las filas (500 filas)  
La transformación se aplicará a todas las filas del conjunto de datos

☐ Filas filtradas: 0 filtros aplicados(500/500 filas)  
La transformación se aplicará a las filas filtradas en la

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

Columna de origen

Seleccionar una columna para dar formato

Date of Joining

Dar formato a la columna a

Formato de fecha/hora

Elegir formato de fecha y hora

dd/mm/yyyy

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

- d) *Renombra la columna Phone No. a Telefono.*

Cambiar el nombre de la columna

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)

Columna	SSN	Phone No.	Password
096-02-5763	096-02-5763	201-249-9955	*****
296-15-2963	296-15-2963	209-425-0529	*****
753-26-5449	753-26-5449	209-458-4318	*****
472-57-5360	472-57-5360	209-511-5725	*****
590-99-8812	590-99-8812	480-540-3150	*****
354-08-9157	354-08-9157	208-864-7893	*****
626-85-8710	626-85-8710	201-662-5551	*****
722-18-7710	722-18-7710	505-292-1926	*****
116-98-0984	116-98-0984	217-956-9427	*****
259-99-5850	259-99-5850	314-470-5988	*****
653-31-3368	653-31-3368	314-764-6462	*****
069-02-9441	069-02-9441	236-394-6533	*****
528-99-4896	528-99-4896	316-695-9121	*****
231-99-6040	231-99-6040	803-948-3291	*****
386-33-8050	386-33-8050	239-375-5853	*****
552-99-0044	552-99-0044	505-964-4274	*****
		217-983-5084	*****

- 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

ABC Date of Birth

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

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

Seleccionar secreto

Elija un secreto de AWS Secrets Manager para cifrar los datos.

Secreto seleccionado

CREATE\_DEFAULT\_SECRET

Seleccionar secreto

Crear nuevo secreto

PABLO MENÉNDEZ DE LA ROSA

7

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.*

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	Remover
::	# Salary	Media	Salary_mean	# Entero	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 ☒ Ver solo las columnas afectadas

ABC Gender	# Salary_mean
F	117763
M	122236
F	117763
M	122236
M	122236
M	122236
F	117763
F	117763
F	117763
F	117763

3) Publica la receta.

Publicar receta

×

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

Nombre de la receta

databrew-rrhh-recipe

Descripción de la versión

Pasos de la receta (5)

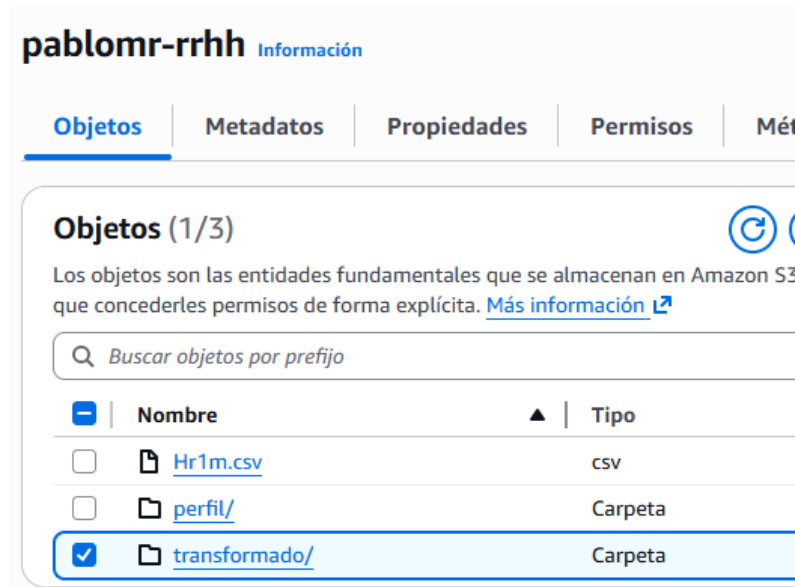
1. Fusionar columnas **Name Prefix**, **First Name**, **Middle Initial**, **Last Name** en **Nombre\_completo\_empleado** separados por " "
2. Cifrar **E Mail** con cifrado determinista
3. Eliminar columna **Short Month**, **DOW of Joining**, **Short DOW**
4. Cambiar formato de **Date of Joining** a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

Cancelar

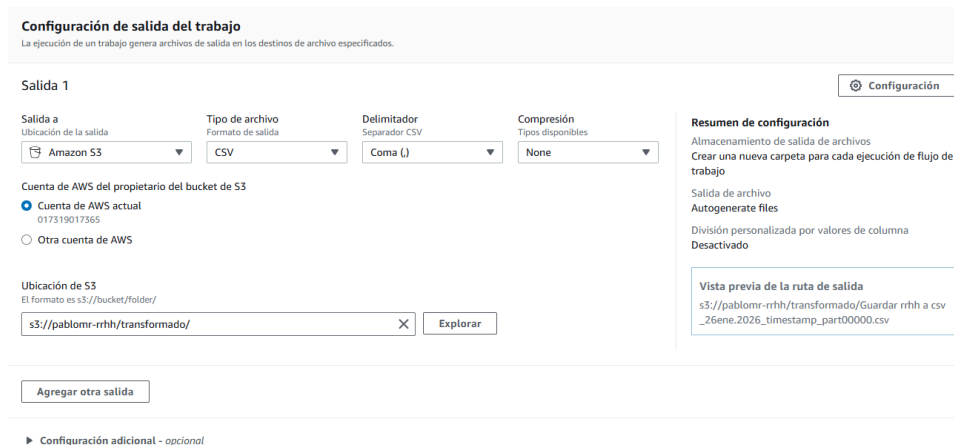
Publicar



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

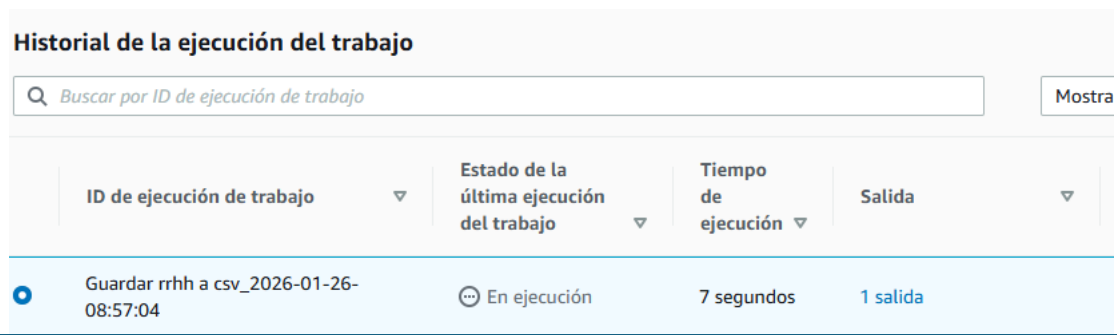


- 5) A partir de la receta anterior, crea un nuevo trabajo que nos deje los datos en formato CSV con comas en la carpeta del punto anterior.



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

Comenzamos ejecución.



Termino por fin

### Historial de la ejecución del trabajo

ID de ejecución de trabajo	Estado de la última ejecución del trabajo	Tiempo de ejecución	Salida	Resumen
----------------------------	---	---------------------	--------	---------

Guardar rrhh a csv_2026-01-26-08:57:04	✓ Realizado con éxito	16 minutos, 42 segundos	1 salida
--	-----------------------	-------------------------	----------

Observamos el ficherillo, está perfectamente transformado.

[illegible]

### 3) APARTADO C

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

Nueva conexión

Detalles del nuevo conjunto de datos

Nombre del conjunto de datos

rrhh transformad

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

Conectarse a un nuevo conjunto de datos		
Carga de archivo		El sistema se conectará automáticamente al nuevo conjunto de datos.

Carga de archivo

Largo 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://pablomr-rhth/transformado/Guardar rhth a csv\_26Jan2026\_1769414312

×

⌵

El formato es s3://bucket/prefix

S3 Buckets > pablomr-rhth > transformado > Guardar rhth a csv\_26Jan2026\_1769414312126

Seleccionar toda la carpeta

↻

🔍 Buscar objetos de S3 por nombre

<

1

>

⌵

	Nombre	Tamaño	Última actualización
<input type="radio"/>	<div>📄 Guardar rhth a csv_26Jan2026_1769414312126_part00000.csv</div>	647.31 MB	26 de enero de 2026, 8:13:47 am

Definir parámetros de conjuntos de datos dinámicos

¿Qué puedo hacer con los parámetros?

Ejemplo de RegEx que se puede agregar a la ruta

▶ Seleccionar archivos solo en la carpeta principal

▶ Seleccionar archivos que terminan con .csv solo en la carpeta principal

▶ Seleccionar archivos que terminan con .csv en la carpeta principal y sus subcarpetas

Elegir archivos filtrados

☐ Especificar el número de archivos que se van a incluir
 

Más reciente

10

archivos

☐ Especificar el intervalo de fechas de la última actualización
 

Las últimas 24 horas

⌵

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

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

Valida el recuento de 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

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

1000000

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.

Comprobación 2

Eliminar

Comprobación de la calidad de los datos

Valores únicos

Compruebe el recuento de valores únicos en la columna.

E Mail

Condición

Es igual

Valor

100

% (porcentaje) filas

Agregue otra comprobación de calidad de los datos

Comprobación 3

Eliminar

Comprobación de la calidad de los datos

Valores únicos

Compruebe el recuento de valores únicos en la columna.

SSN

Condición

Es igual

Valor

100

% (porcentaje) filas

Agregue otra comprobación de calidad de los datos

Resumen de Reglas

La regla pasará si Emp ID, E Mail, SSN 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.*

Comprobación 1

Eliminar

Comprobación de la calidad de los datos

Valores faltantes

Compruebe los valores que faltan en la columna.

Emp ID

Condición

No es igual

Valor

100

% (porcentaje) filas

---

Comprobación 2

Eliminar

Comprobación de la calidad de los datos

Valores faltantes

Compruebe los valores que faltan en la columna.

E Mail

Condición

No es igual

Valor

100

% (porcentaje) filas

- 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.*

Comprobación 1

[Eliminar](#)

Comprobación de la calidad de los datos

Valores numéricos

Compruebe los valores numéricos en la columna en función de la...

Emp ID

Condición

Mayor que igual

Valor

☒ Valor personalizado ☐ Valor de columna

0

Comprobación 2

[Eliminar](#)

Comprobación de la calidad de los datos

Valores numéricos

Compruebe los valores numéricos en la columna en función de la...

Age in Yrs.

Condición

Está entre

Mayor que igual

☒ Valor personalizado ☐ Valor de columna

0

Menor que igual

☒ Valor personalizado ☐ Valor de columna

80

- e) Verificar mediante una expresión regular ( $^{\backslash}d\{3\}-\backslash d\{2\}-\backslash d\{4\}\$$ ) que el formato de los datos del SSN debe tener ser del tipo xxx-xx-xxxx).

Nombre de regla

Verificar formato de los datos del SSN

Ámbito de comprobación de calidad de los datos

Comprobación individual de cada columna

Criterios de éxito de la regla

Se cumplen todas las comprobacion

Comprobaciones de calidad de los datos

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

$^{\backslash}d\{3\}-\backslash d\{2\}-\backslash d\{4\}\$$

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


Conjuntos de reglas de calidad de datos (2)						
<input type="text" value="Buscar conjuntos de reglas"/>				Crear trabajo de perfil con conjunto de reglas	Acciones ▾	Crear un co
<input type="checkbox"/>	Nombre del conjunto de reglas de calidad de datos ▾	Descripción ▾	Conjunto de datos asociado ▾	Trabajo Asociado ▾	Fecha de creación ▾	Creado por
<input checked="" type="checkbox"/>	rrhh transformado reglas 5 reglas	-	rrhh transformado	-	hace unos segundos 26 de enero de 2026, 10:11:52 am	user3553228=Pablo

#### 4) 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.

**pablomr-rrhh** Información

Objetos Metadatos Propiedades Permisos Métricas Administra


**Objetos (1/4)**  [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](#)


<input type="checkbox"/>	Nombre	Tipo
<input checked="" type="checkbox"/>	<a href="#">calidad/</a>	Carpeta
<input type="checkbox"/>	<a href="#">Hr1m.csv</a>	csv
<input type="checkbox"/>	<a href="#">perfil/</a>	Carpeta
<input type="checkbox"/>	<a href="#">transformado/</a>	Carpeta

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

a) *Aplica el trabajo a todo el dataset.*

**Apply data quality rules** 

Data quality rules for rrhh transformado (1/1)

[Crear un conjunto de reglas de calidad de datos](#) 

<input checked="" type="checkbox"/>	Nombre del conjunto de reglas de calidad de datos	Descripción
<input checked="" type="checkbox"/>	<a href="#">rrhh transformado reglas</a> 5 reglas	-

Cancel [Apply selected rulesets](#)

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  
017319017365

☐ 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 y posteriormente vete a trabajos de perfil y ejecútalo.

Aquí está el trabajo:

Trabajos de perfil (3)							
<input type="text" value="Buscar trabajos"/>				Mostrar todo			
<input checked="" type="checkbox"/>	Nombre del trabajo	Estado de la última ejecución del trabajo	Conjunto de datos	Perfil de datos	Última ejecución	Creado el	Creado por
<input checked="" type="checkbox"/>	rrhh transformado profile job	-	rrhh transform	Ver perfil de datos	-	hace 2 minutos 27 de enero de 2026, 9:44:27 am	voclabs

Lo ejecutamos:

☐

rrhh transformado profile job

En ejecución

Gestionado:

Trabajos de perfil (3)			
<input type="text" value="Buscar trabajos"/>			
<input checked="" type="checkbox"/>	Nombre del trabajo	Estado de la última ejecución del trabajo	Conjunto de datos
<input checked="" type="checkbox"/>	rrhh transformado profile job	Realizado con éxito	rrhh transform

- 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.

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

Expandir todo | Contraer todo

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

---

☒ rrhh transformado reglas 5 reglas Fallo

- Fallo **Valida el recuento de filas**  
Comprobar si **conjunto de datos** tiene recuento de filas == 1000000
- Fallo **El ID de empleado, la dirección de correo electrónico y el SSN deben ser únicos**  
Comprobar si **Emp ID, E Mail, SSN** tiene valores únicos == 100%
- Fallo **El ID de empleado y la dirección de correo electrónico no deben ser nulos**  
Comprobar si **Emp ID, E Mail** tiene valores válidos != 100%
- Realizado con éxito **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**  
Comprobar si **Emp ID** tiene valores >= 0 Y **Age in Yrs.** tiene valores está entre 0 y 80 PARA mayor o igual que 100% de filas 100% Realizado con éxito 0% Fallo
- Realizado con éxito **Verificar formato de los datos del SSN**  
Comprobar si **SSN** tiene valores coincidencias ^\d(3)-\d(2)-\d(4)\$ PARA mayor o igual que 100% de filas 100% Realizado con éxito 0% Fallo

- 5) De aparecer algún error (por ejemplo, en la imagen me dice que hay SSN y Emp ID repetidos ), vete a la pestaña de Estadísticas de columna y comprueba que es cierto el error de las reglas (por ejemplo, en la imagen puedo ver los Emp ID repetidos)

TODOS (31) ABC CADENA (22) # ENTERO (7)

#	Emp ID	100% Valido
ABC	Nombre_completo_e...	100% Valido
ABC	Gender	100% Valido
ABC	E Mail	100% Valido
ABC	Father's Name	100% Valido
ABC	Mother's Name	100% Valido
ABC	Mother's Maiden Na...	100% Valido
ABC	Date of Birth	100% Valido
ABC	Time of Birth	100% Valido
#	Age in Yrs.	100% Valido
#	Weight in Kgs.	100% Valido

**Calidad de los datos**

VALORES VÁLIDOS 20000 100% VALORES FALTANTES 0 0%

**Distribución de valores**

Los valores atípicos se detectan mediante la puntuación Z con un umbral de desviación estándar de 3

DISTRIBUCIÓN DE VALORES | DISTRIBUCIÓN DE PUNTUACIÓN Z

Distintiva 19.765 Única 19.531 Total 20.000

**Reglas de calidad de los datos - aplicadas**

Expandir todo | Contraer todo

☒ rrhh transformado reglas 3 reglas Fallo

- Fallo **El ID de empleado, la dirección de correo electrónico y el SSN deben ser únicos**  
Comprobar si **Emp ID, E Mail, SSN** tiene valores únicos == 100%
- Fallo **El ID de empleado y la dirección de correo electrónico no deben ser nulos**  
Comprobar si **Emp ID, E Mail** tiene valores válidos != 100%
- Realizado con éxito **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**  
Comprobar si **Emp ID** tiene valores >= 0 Y **Age in Yrs.** tiene valores está entre 0 y 80 PARA mayor o igual que 100% de filas 100% Realizado con éxito 0% Fallo

**Información sobre los datos**

## 5) APARTADO E

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



**pablomr-rrhh** Información

**Objetos** | Metadatos | Propiedades | Permisos | Métricas | Administración

**Objetos (1/5)** 🔄 Copiar URI de S3

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

🔍 *Buscar objetos por prefijo*

	Nombre	Tipo	Últ
<input type="checkbox"/>	<a href="#">calidad/</a>	Carpeta	-
<input checked="" type="checkbox"/>	<a href="#">curated/</a>	Carpeta	-
<input type="checkbox"/>	<a href="#">Hr1m.csv</a>	csv	22
<input type="checkbox"/>	<a href="#">perfil/</a>	Carpeta	-
<input type="checkbox"/>	<a href="#">transformado/</a>	Carpeta	-

- 2) De modo similar a como hicimos en el apartado B, 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 (por ejemplo, eliminando filas con campos duplicados). El resultado del trabajo almacénalo en formato parquet comprimido en la carpeta curated.

**databrew-rrhh-transformado**

Conjunto de datos: **rrhh transformado** 🔍 Muestra: Muestra de los primeros n (500 filas)

DESARROLLAR | REVISAR | FILTRAR | ORDENAR | COLUMNA | FORMATEO | LIMPIAR | EXTRAER | FALTANTE | NO ES VÁLIDO | DUPLICADOS | VALORES ATÍPICOS | DIVIDIR | FUSIONAR | CREAR | FUNCIONES | CONDICIONES | ANIDAR-DESANIDAR | DINAMIZAR | GRUPO | UNIR | COMBINACIÓN | TEXTO | ESC

Visualizando 31 columnas 500 filas

#	Emp ID	Nombre_completo_employeado	Gender	E Mail	Father's Name
49821	Mrs. Jeffrey C Murakami	F	AQBMXXJuOmF3cpzZWnyZKRzbWfuYWdlcjp1cy...	Alex Murakami	
29350	Mr. Shelby D Davidson	M	AQBMXXJuOmF3cpzZWnyZKRzbWfuYWdlcjp1cy...	Fredrick Davidson	
02166	Drs. Wen P Russo	F	AQBMXXJuOmF3cpzZWnyZKRzbWfuYWdlcjp1cy...	Laurence Russo	
82838	Prof. Aaron Q Delima	M	AQBMXXJuOmF3cpzZWnyZKRzbWfuYWdlcjp1cy...	Napoleon Delima	
65681	Mr. Frederic M Christofferso	M	AQBMXXJuOmF3cpzZWnyZKRzbWfuYWdlcjp1cy...	Monty Christofferso	
31466	Hon. Billie M Lachapelle	M	AQBMXXJuOmF3cpzZWnyZKRzbWfuYWdlcjp1cy...	Byron Lachapelle	
35618	Mrs. Roseline M Bach	F	AQBMXXJuOmF3cpzZWnyZKRzbWfuYWdlcjp1cy...	Vincenzo Bach	

## Creamos el trabajo

**Trabajos de recetas (3)**

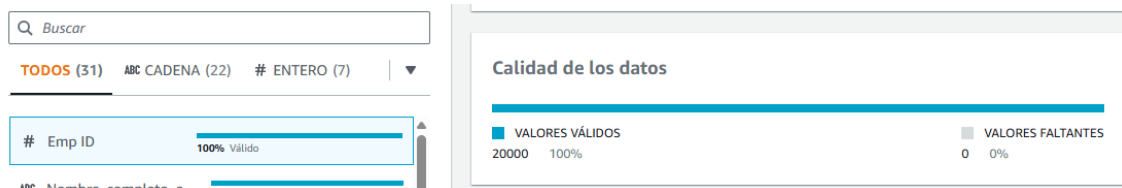
🔍 *Buscar trabajos* En ejecución

<input checked="" type="checkbox"/>	Nombre del trabajo	Estado	Entrada del trabajo
<input checked="" type="checkbox"/>	currando día y tarde increíble	En ejecución	databrew-rrh... ( rrhh transfor... + databrew-rrh... ) Proyecto Conjunto de datos Receta

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

Conjuntos de datos (5)						Ver detalles	Crear proyecto con este conjunto de datos
<input type="text" value="Buscar conjuntos de datos"/>							
<input type="checkbox"/>	Nombre del conjunto de datos	Tipo de datos	Perfil de datos	Origen	Ubicación		
<input checked="" type="checkbox"/>	rrhh-curated	csv	-	S3	s3://pablomr-rrhh/curated/currando día y tarde increíble_27Jan2026_1769501848858/		

- 4) Verifica ahora las estadísticas de las columnas que has modificado para asegurarnos que todo ha ido bien (por ejemplo, que no haya datos repetidos en Emp ID). ¿Cuántas filas tiene ahora el archivo resultante?



## 6) 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, por ejemplo, la que nos contaba el número de filas. (Puede ser que los nombres de los campos hayan cambiado respecto a los originales, si es así modifícalos en el conjunto de reglas)

### Detalles del conjunto de reglas

Nombre del conjunto de reglas

Identificador del conjunto de reglas

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

### 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

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

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

	Nombre del conjunto de reglas de calidad de datos ▾	Descripción ▾	Conjunto de datos asociado ▾
<input checked="" type="checkbox"/>	rrhh curado reglas 5 reglas	-	rrhh-curved

3) Ejecuta dicho trabajo contra todo el dataset.

Trabajos de perfil (4)			
<input type="text" value="Buscar trabajos"/>			
<input type="button" value="Mostrar 1"/>			
	Nombre del trabajo ▾	Estado de la última ejecución del trabajo	Conjunto de datos ▾
<input checked="" type="checkbox"/>	rrhh-curved profile job	En ejecución	rrhh-curved

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

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

[Expandir todo](#) | [Contraer todo](#)

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

☒ rrhh curado reglas 5 reglas

##### ☒ Valida el recuento de filas

Comprobar si **conjunto de datos** tiene recuento de filas  $\geq$  1000000

##### ☒ El ID de empleado, la dirección de correo electrónico y el SSN deben ser únicos

Comprobar si **Emp ID, E Mail, SSN** tiene valores únicos == 100%

##### ☒ El ID de empleado y la dirección de correo electrónico no deben ser nulos

Comprobar si **Emp ID, E Mail** tiene valores faltantes != 100%



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

Comprobar si **Emp ID** tiene valores  $\geq$  0 Y **Age in Yrs.** tiene valores está entre 0 y 80 PARA mayor o igual que 100% de filas

100% Realizado con éxito 0% Fallo

##### ☒ Verificar formato de los datos del SSN

Comprobar si **SSN** tiene valores coincidencias  $^{\wedge}d\{3\}-\backslash d\{2\}-\backslash d\{4\}$  PARA mayor o igual que 100% de filas

100% Realizado con éxito 0% Fallo