

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.

pablomr-rrhh Información

Objetos | Metadatos | Propiedades | Permisos | Métricas | Acciones

Objetos (1)

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

Buscar objetos por prefijo

Nombre	Tipo
Hr1m.csv	CSV

- 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 MiB	Proyectos asociados databrew-rrhh	Trabajos asociados -
Origen de datos 53	Ubicación de S3 s3://pablomr-rrhh/Hr1m.csv		

Vista previa del conjunto de datos

Nombre de la columna	Primeras 5 filas de datos	Cuadrícula	Esquema	Texto	Árbol
# Emp ID	549821, 429350, 702166, 982838, 565681				
ABC Name Prefix	Mrs., Mr., Mrs., 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, Christofferso				

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

pablomr-rrhh Información

Objetos Metadatos Propiedades Permisos

Objetos (1/2)

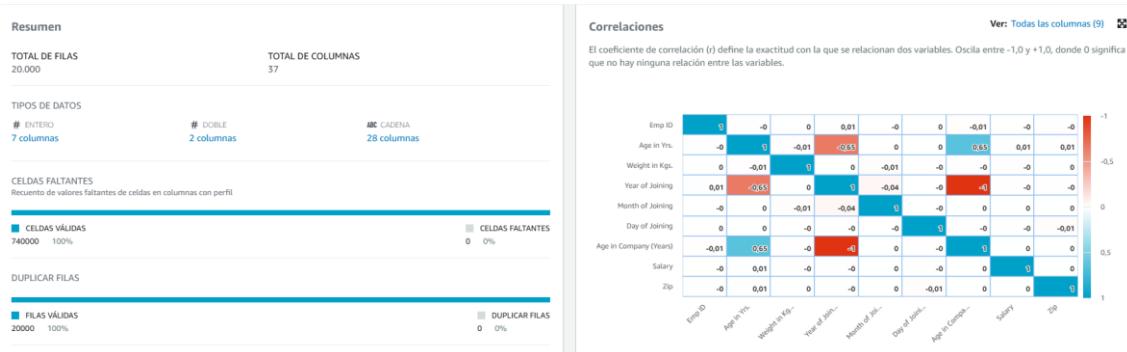
Los objetos son las entidades fundamentales que se almacenan en Archivo y que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

Nombre	Tipo
<input type="checkbox"/> Hr1m.csv	CSV
<input checked="" type="checkbox"/> 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 qué 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%

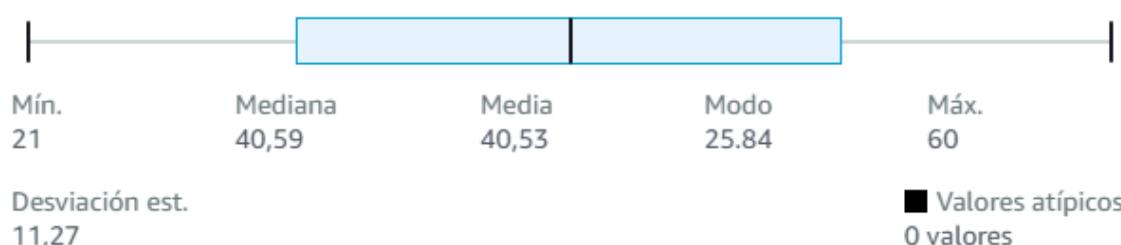
d) Analizando el diagrama de cajas de los salarios ¿En qué horquilla se mueven? ¿Cuál es la media, mediana y moda? ¿Están distribuidos simétricamente?

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



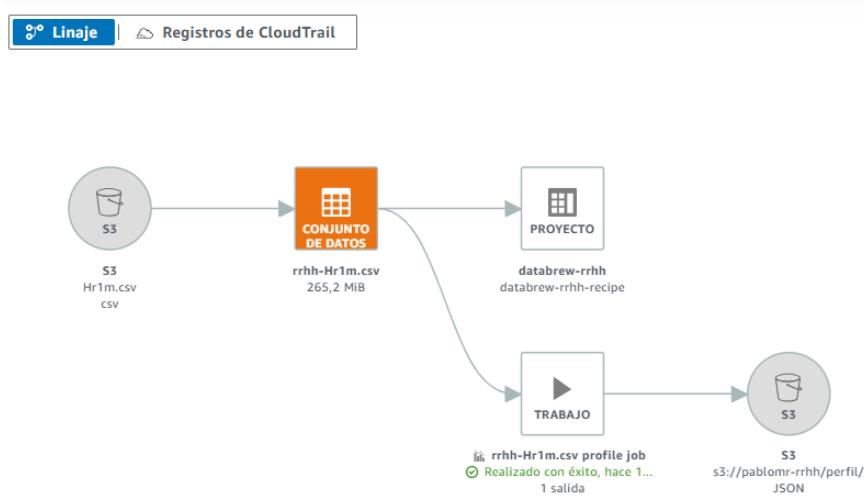
e) Busca la misma información para el campo Age in years.

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.

Visualizando 37 columnas ▾ 500 filas									
MUESTRA									
#	Emp ID	ABC Name Prefix	ABC First Name	ABC Middle Initial	ABC Last Name	Total	Distinctive	Única	0
100	500	Única 500	Total 500						
112,38 K	565,2 K	Media 572,02 K	Modo 999,46 K						
112,38 K	565,2 K	Media 572,02 K	Modo 999,46 K						
149821		Mrs.	Jeffrey	C	Murakami				
129350		Mr.	Shelby	D	Davidson				
702166		Drs.	Wen	P	Russo				
382838		Prof.	Aaron	Q	Delima				
165681		Mr.	Frederic	M	Christofferso				
331466		Hon.	Billie	M	Lachapelle				
135618		Mrs.	Roseline	M	Bach				
351970		Ms.	Jerlene	H	Chalk				
705378		Drs.	Marcella	Q	Payan				
193321		Mrs.	Sylvie	X	Pautz				
140547		Mr.	Osvaldo	S	Swayne				
212154		Mr.	Devon	E	Kehoe				
167533		Ms.	Brigitte	G	Tong				
763494		Dr.	Mauricio	F	Ryles				
137884		Hon.	Fidela	N	Norden				
250120		Hon.	Isalias	A	Dibenedetto				
329416		Mr.	Landon	K	Stoltz				

- 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
Seleccione dos o más columnas en el orden de fusión

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

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

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 lo apliques, ya que si no un paso posterior que tenemos que hacer nos dará un error)

The screenshot shows the AWS DataBrew interface. On the left, there are two preview tables: 'ORIGEN ABC SSN' and 'ORIGEN ABC Phone No.'. Both tables show sample data with columns 'Nombre' and 'Apellido'. The 'Nombre' column contains values like 'Juan', 'Pedro', etc., and the 'Apellido' column contains values like 'Pérez', 'García', etc. On the right, there is a configuration panel titled 'Celdas de origen' (Origin Cells). It has a dropdown menu 'Nombre de la columna' (Column Name) set to 'ABC SSN' and 'ABC Phone No.'. Below it, there is a section for 'Reemplazar opciones' (Replace options) with a dropdown 'Reemplazar por valor personal...' (Replace by personal value) and a text input field containing '#####'. At the bottom, there are sections for 'Aplicar la sustitución a' (Apply substitution to) and 'Aplicar transformación a' (Apply transformation to), both with radio buttons for 'Todas las filas' (All rows).

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

The screenshot shows the 'Cifrado de datos' (Data Encryption) configuration screen. It starts with a 'Columnas de origen' (Origin Columns) section where 'Nombre de la columna' (Column Name) is set to 'E Mail' and 'Date of Birth'. Below this is an 'Opciones de cifrado' (Encryption options) section with two radio buttons: 'Cifrado determinista' (Deterministic encryption) and 'Cifrado probabilístico' (Probabilistic encryption). The 'Cifrado determinista' option is selected, with a note explaining it encrypts data while maintaining the same resulting value for each distinct value. A callout box provides information about using AWS Secrets Manager to decrypt deterministic ciphertext. Finally, there is a 'Seleccionar secreto' (Select secret) section with a dropdown set to 'CREATE_DEFAULT_SECRET' and two buttons: 'Seleccionar secreto' (Select secret) and 'Crear nuevo secreto' (Create new secret).

- 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

3) Publica la receta.

Publicar receta X

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.

pablomr-rrhh [Información](#)

- [Objetos](#)
- [Metadatos](#)
- [Propiedades](#)
- [Permisos](#)
- [Métodos](#)

Objetos (1/3)

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

Buscar objetos por prefijo	Nombre	Tipo
<input type="checkbox"/>	Hr1m.csv	CSV
<input type="checkbox"/>	perfil/	Carpeta
<input checked="" 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.

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"/> Amazon S3	<input type="button"/> CSV	<input type="button"/> Coma (,)	<input type="button"/> None

Cuenta de AWS del propietario del bucket de S3

- Cuenta de AWS actual
017319017365
- Otra cuenta de AWS

Ubicación de S3
El formato es \$3://bucket/folder/

s3://pablomr-rrhh/transformado/ Explorar

Resumen de configuración

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

Vista previa de la ruta de salida
s3://pablomr-rrhh/transformado/_26ene.2026_timestamp_part00000.csv

Agregar otra salida

Configuración

Configuración adicional - opcional

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

[Comenzamos ejecución.](#)

Historial de la ejecución del trabajo

Buscar por ID de ejecución de trabajo Mostrar

ID de ejecución de trabajo	Estado de la última ejecución del trabajo	Tiempo de ejecución	Salida
<input checked="" type="radio"/> Guardar rrhh a csv_2026-01-26-08:57:04	<input type="radio"/> En ejecución	7 segundos	1 salida

Termino por fin

Historial de la ejecución del trabajo

Buscar por ID de ejecución de trabajo			Mostrar todo ▾	
ID de ejecución de trabajo ▾	Estado de la última ejecución del trabajo ▾	Tiempo de ejecución ▾	Salida ▾	Resumen

Observamos el ficherillo, está perfectamente transformado.

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 transformado

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

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

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

Guardar rrhh a csv_26Jan2026_17694143126

Seleccionar toda la carpeta

Buscar objetos de S3 por nombre

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.

The screenshot shows the 'Regla 1' configuration screen. At the top right is a 'Habilitar regla' (Enable rule) button. The 'Nombre de regla' (Rule name) is set to 'Valida el recuento de filas'. Under 'Ámbito de comprobación de calidad de los datos' (Scope of data quality check), 'Comprobación individual de cada columna' (Check each column individually) is selected, and the status is 'Se cumplen todas las comprobaciones' (All checks are met). The 'Comprobaciones de calidad de los datos' (Data quality checks) section contains one entry: 'Comprobación 1' (Check 1) for 'Número de filas' (Number of rows), which checks if the count equals 1000000. The 'Resumen de Reglas' (Rule summary) at the bottom states: 'La regla pasará si conjunto de datos tiene recuento de filas == 1000000' (The rule will pass if the dataset has a row count == 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.

The screenshot shows two 'Comprobación' (Check) sections. The first section, 'Comprobación 2', is for the 'E Mail' column, checking if there are 100 unique values. The second section, 'Comprobación 3', is for the 'SSN' column, also checking for 100 unique values. Both sections follow the same structure: 'Comprobación de la calidad de los datos' (Data quality check) for 'Valores únicos' (Unique values), 'Condición' (Condition) set to 'Es igual' (Is equal), and 'Valor' (Value) set to 100. The 'Resumen de Reglas' (Rule summary) at the bottom states: 'La regla pasará si Emp ID, E Mail, SSN tiene valores únicos == 100%' (The rule will pass if Emp ID, E Mail, SSN have unique values == 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.*

The screenshot shows the Databrew interface with two validation rules defined:

- Comprobación 1** (Validation 1):
 - Condition: Valores faltantes (Missing values)
 - Column: Emp ID
 - Condition: No es igual (Is not equal)
 - Value: 100
 - Unit: % (porcentaje) filas (Percentage of rows)
- Comprobación 2** (Validation 2):
 - Condition: Valores faltantes (Missing values)
 - Column: E Mail
 - Condition: No es igual (Is not equal)
 - Value: 100
 - Unit: % (porcentaje) filas (Percentage of rows)

- 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 (`^\d{3}-\d{2}-\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 comprobaciones

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

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

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

Conjuntos de reglas de calidad de datos (2)						
<input type="button" value="Crear trabajo de perfil con conjunto de reglas"/>		<input type="button" value="Acciones"/>		<input type="button" value="Crear un conjunto de reglas"/>		
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)

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

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

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

- 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	Tipo de archivo
<input checked="" type="radio"/> Cuenta de AWS actual 017319017365	Formato de salida
<input type="radio"/> Otra cuenta de AWS	JSON
Ubicación de S3	
El formato es s3://bucket/folder/	<input type="text" value="s3://pablomr-rrhh/calidad/"/> <input type="button" value="X"/> <input type="button" value="Explorar"/>
Cifrado	
<input type="checkbox"/> 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	
Elige el rol que tiene acceso para conectarse a los datos. Actualice para ver las últimas actualizaciones.	<input type="text" value="LabRole"/> <input type="button" value="▼"/> <input type="button" value="C"/>

- 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="button" value="C"/>	<input type="button" value="Ver detalles"/>	<input type="button" value="Ejecutar trabajo"/>	<input type="button" value="Acciones ▾"/>	<input type="button" value="Crear trabajo"/>
	<input type="text" value="Buscar trabajos"/>	<input type="button" value="Mostrar todo"/>									
<input type="checkbox"/>	Nombre del trabajo	Estado de la última ejecución del trabajo	Conjunto de datos	Perfil de datos	Última ejecución	Creado el					
<input checked="" type="checkbox"/>	rrhh transformado profile job	-	rrhh transform	Ver perfil de datos	-	hace 2 minutos 27 de enero de 2026, 8:44:27 am	Creado por	vocabs	Etiquetas	-	

Lo ejecutamos:

<input type="checkbox"/>	rrhh transformado profile job	<input type="button" value="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](#)

Buscar

[TODOS \(5\)](#) (✓) REALIZADO CON ÉXITO (2) (✗) FALLO (3) (⚠) ERROR (0) (⊖) DESACTIVADO (0)

rrhh transformado reglas 5 reglas

✗ Fallo

✗ Valida el recuento de filas

Comprobar si **conjunto de datos** tiene recuento de filas == **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 válidos != **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 >= **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 **^\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)



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)



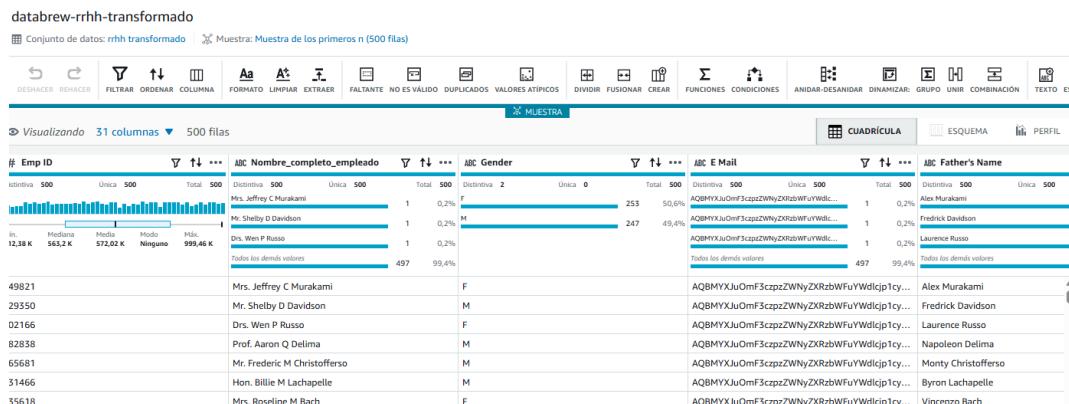
Copiar URI de S3

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

Buscar objetos por prefijo

<input type="checkbox"/>	Nombre	Tipo	Última modificación
<input type="checkbox"/>	calidad/	Carpeta	-
<input checked="" type="checkbox"/>	curated/	Carpeta	-
<input type="checkbox"/>	Hr1m.csv	csv	22
<input type="checkbox"/>	perfil/	Carpeta	-
<input type="checkbox"/>	transformado/	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.



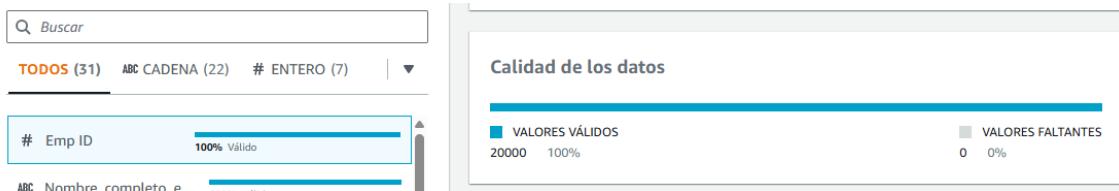
Creamos el trabajo

Trabajos de recetas (3)

<input type="checkbox"/>	Nombre del trabajo	Estado	Entrada del trabajo
<input checked="" type="checkbox"/>	curriendo dia y tarde increible	En ejecución	databrew-rrhh... (rrhh transformado + databrew-rrhh...) Proyecto Conjunto de datos Receta

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

- 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)

- 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-curated

3) Ejecuta dicho trabajo contra todo el dataset.

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

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)						
<input type="button"/> Expandir todo <input type="button"/> Contraer todo						
<input type="button"/> TODOS (5) <input checked="" type="button"/> REALIZADO CON ÉXITO (5) <input type="button"/> FALLO (0) <input type="button"/> ERROR (0) <input type="button"/> DESACTIVADO (0)						
<input type="checkbox"/> rrhh curado reglas 5 reglas						
<table border="1"> <tr> <td><input checked="" type="checkbox"/> Valida el recuento de filas Comprobar si conjunto de datos tiene recuento de filas >= 1000000</td> </tr> <tr> <td><input checked="" type="checkbox"/> 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%</td> </tr> <tr> <td><input checked="" type="checkbox"/> El ID de empleado y la dirección de correo electrónico no deben ser nulos Comprobar si Emp ID, E Mail tiene valores faltantes != 100%</td> </tr> <tr> <td><input checked="" type="checkbox"/> 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</td> </tr> <tr> <td><input checked="" type="checkbox"/> 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</td> </tr> </table>		<input checked="" type="checkbox"/> Valida el recuento de filas Comprobar si conjunto de datos tiene recuento de filas >= 1000000	<input checked="" type="checkbox"/> 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%	<input checked="" type="checkbox"/> El ID de empleado y la dirección de correo electrónico no deben ser nulos Comprobar si Emp ID, E Mail tiene valores faltantes != 100%	<input checked="" type="checkbox"/> 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	<input checked="" type="checkbox"/> 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
<input checked="" type="checkbox"/> Valida el recuento de filas Comprobar si conjunto de datos tiene recuento de filas >= 1000000						
<input checked="" type="checkbox"/> 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%						
<input checked="" type="checkbox"/> El ID de empleado y la dirección de correo electrónico no deben ser nulos Comprobar si Emp ID, E Mail tiene valores faltantes != 100%						
<input checked="" type="checkbox"/> 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						
<input checked="" type="checkbox"/> 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						