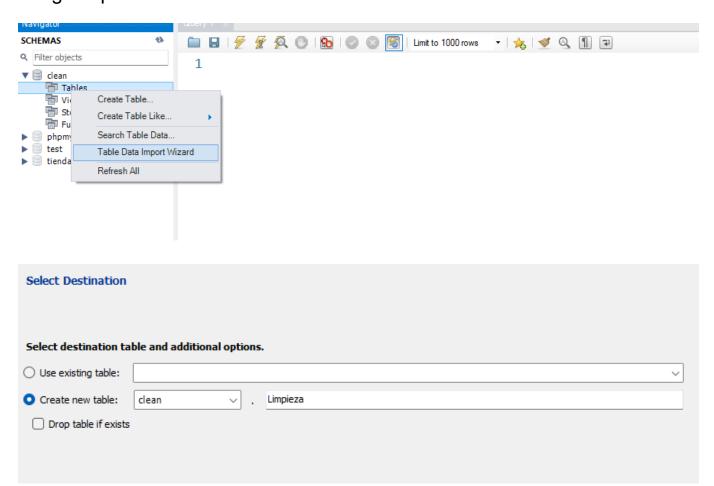
## Limpieza de datos

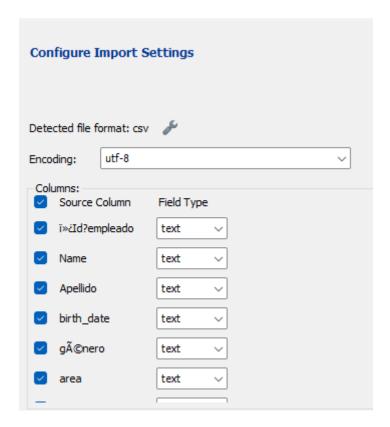
Crear una base de datos:

```
create database if not exists clean;
```

Luego importamos el archivo:



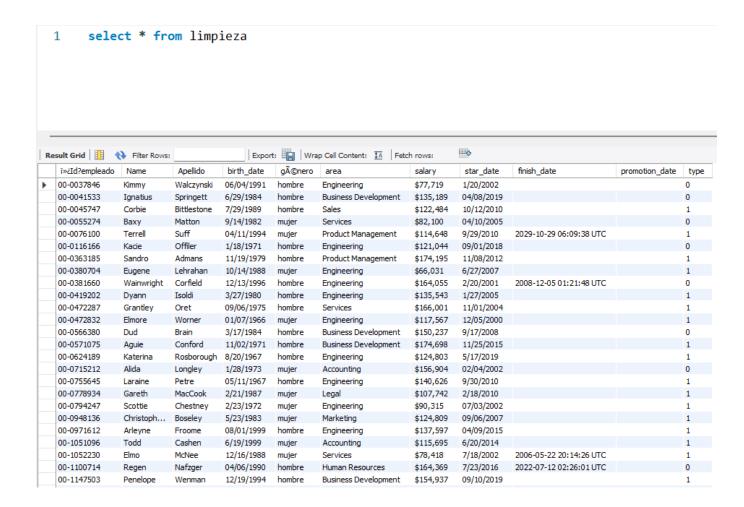
Como vemos los datos vienen sucios:



## **Objetivos**

- Estandarizar del idioma: convertir todos los registros al idioma inglés para mantener la coherencia lingüística.
- Corrección de encabezados: revisar y corregir los nombres de los encabezados para asegurar que sean claron y descriptivos
- Formateo de fechas: ajustar las fechas para que estén en un formato de fecha adecuado en lugar de estar en formato de texto.
- Formato del salario: asegurarse de que el campo de salario esté en el formato numérico adecuado, eliminando cualquier formato de texto que pueda existir.
- Eliminación de espacios extras en nombres: detectar y eliminar los espacios adicionales en los nombres para mantener la consistencia y precisión en los registros.

Información de la tabla:



# Renombrar los nombres de las columnas con caracteres especiales

Para renombrar una columna:

```
ALTER TABLE limpieza CHANGE COLUMN `Id?empleado` Id_emp varchar (20) null;
```

## Verificar si hay registros duplicados

```
select Id_emp, count(*) as cantidad_duplicados
from limpieza
group by Id_emp
having count(*) > 1;
```

si queremos contar la cantidad de valores duplicados podemos hacer una subconsulta

```
select count(*) as cantidad_duplicados
from (x) as subquery
```

En x tenemos que copiar la consulta anterior sin ;:

```
select Id_emp, count(*) as cantidad_duplicados
from limpieza
group by Id_emp
having count(*) > 1
```

# Crear una tabla temporal con valores únicos y luego hacerla permanente

Para renombrar la tabla:

```
rename table limpieza to conduplicados;
```

### Creación de una tabla temporal (sin datos nulos)

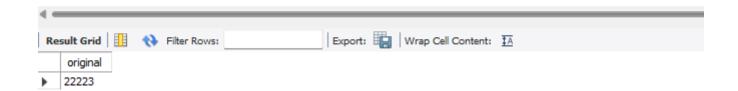
Una tabla temporal se caracteriza por su creación y existencia temporal durante la sesión actual de la base de datos. Al cerrar el programa o la sesión de la base de datos, esta tabla se elimina automáticamente.

```
create temporary table Temp_limpieza as select distinct * from conduplicados;
```

Al seleccionar los distintos, va a seleccionar todos los valores no duplicados. Abajo selecciono los valores con duplicados.

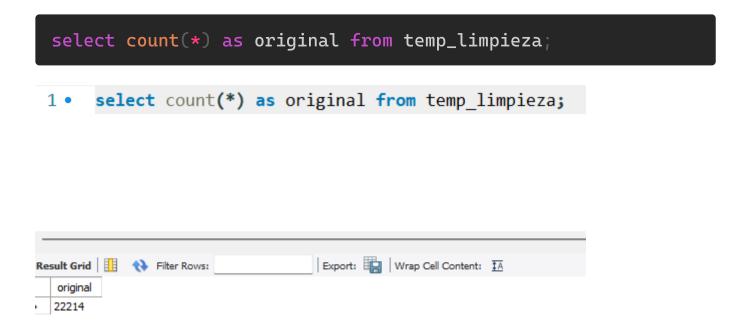
```
select count(*) as original from conduplicados;
```

1 • select count(\*) as original from conduplicados;



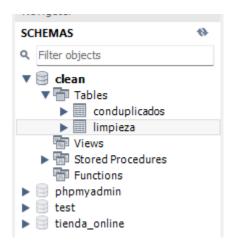
Seleccionamos nuestra tabla temporal y observamos los siguientes puntos:

- 1. Los valores duplicados han sido eliminados correctamente.
- 2. Faltan nueve filas en la tabla.



### Convertir la tabla temporal a permanente

create table limpieza as select \* from temp\_limpieza;



Verificamos si aún hay duplicados nuevamente:

```
SELECT COUNT(*) AS cantidad_duplicados
FROM (
    SELECT Id_emp
    FROM conduplicados
    GROUP BY Id_emp
    HAVING COUNT(*) > 1
) AS subquery;
```

Eliminar tabla que contiene los duplicados:

```
drop table conduplicados;
```

## Activar/Desactivar modo seguro

Para desactivar el modo seguro que SQL trae por defecto y permitir realizar modificaciones, puedes utilizar el siguiente código:

```
set sql_safe_updates = 0;
```

#### Renombrar los nombres de las columnas

```
ALTER TABLE limpieza CHANGE COLUMN 'gÃ@nero' Gender varchar(20) null;
ALTER TABLE limpieza CHANGE COLUMN Apellido Last_name varchar(50) null;
ALTER TABLE limpieza CHANGE COLUMN star_date Start_date varchar(50) null;
```

## Revisar los tipos de datos de la tabla

# describe limpieza;

	Field	Type	Null	Key	Default	Extra
•	Id_emp	varchar(20)	YES		NULL	
	Name	text	YES		NULL	
	Last_name	varchar(50)	YES		NULL	
	birth_date	text	YES		NULL	
	Gender	varchar(20)	YES		NULL	
	area	text	YES		NULL	
	salary	text	YES		NULL	
	start_date	varchar(50)	YES		NULL	
	finish_date	text	YES		NULL	
	promotion_date	text	YES		NULL	
	type	int(11)	YES		NULL	

Hay fechas con tipo de dato texto.

## Trabajando con texto (strings)

### Identificar espacios extra

```
select name from limpieza
where length(name) - length(trim(name)) > 0;
```

• **trim** se utiliza para eliminar espacios en blanco o caracteres específicos al inicio y/o al final de una cadena de texto.

```
select name, trim(name) as name
from limpieza
where length(name) - length(trim(name)) > 0;
```



#### Eliminar los espacios extra

Modificando nombres: ahora para modicar la tabla con los nombres sin espacios en blanco podemos:

```
update limpieza set name = trim(name)
where length(name) - length(trim(name)) > 0;
```

Apellido con espacios:

```
SELECT last_name, TRIM(Last_name) AS Last_name
FROM limpieza
WHERE LENGTH(last_name) - LENGTH(TRIM(last_name)) > 0;
```

Modificando apellidos:

```
UPDATE limpieza
SET last_name = TRIM(Last_name)
WHERE LENGTH(Last_name) - LENGTH(TRIM(Last_name)) > 0;
```

# ¿Qué sucede si tenemos más de un espacio entre dos palablas?

Esto código es para agregar más espacios:

```
update limpieza set area = replace(area, ' ', ' ');
```

Este codigo es para ver que casillas tiene más de un espacio entre dos palabras:

```
select area from limpieza
where area regexp '\\s{2,}'
```

#### regexp : expresion regular



#### Código de ensayo de eliminación de los espacios

```
select area, trim(regexp_replace(area, '\\s+', ' ')) as
ensayo from limpieza;
```

Engineering		Engineering
Business	Development	Business Development
Sales		Sales
Services		Services
Product	Management	Product Management
Engineering		Engineering
Product	Management	Product Management
Engineering		Engineering
Engineering		Engineering
Engineering		Engineering
Services		Services
Engineering		Engineering
Business	Development	Business Development
Business	Development	Business Development
Engineering		Engineering
Accounting		Accounting

#### Aplicar los cambios a la tabla

## **Buscar y reemplazar (textos)**

- 1. Ensayar
- 2. Actualizar tabla
- 3. Modificar propiedad (si es necesario)

Vamos a reemplazar todo al inglés

### **Ensayo**

```
select gender,
case
    when gender = 'hombre' then 'male'
    when gender = 'mujer' then 'female'
```

```
else 'other'
end as gender1
from limpieza;
```

gender	gender 1
hombre	male
hombre	male
hombre	male
mujer	female
mujer	female
hombre	male
hombre	male
mujer	female
hombre	male
hombre	male
hombre	male
mujer	female
hombre	male
hombre	male
hombre	male
mujer	female
hombre	male
mujer	female

#### Actualizar la tabla

## Cambiar la propiedad de una columna

Ahora hay que modificar esta columna type y que admita texto no números:

```
type
0
0
0
0
1
0
1
                                         NULL
Id_emp
               varchar(20) YES
                                         NULL
Name
                         YES
               text
                                          NULL
               varchar(50) YES
Last_name
                                         NULL
birth_date
                            YES
               text
                                          NULL
Gender
               varchar(20)
                           YES
```

YES

YES

YES

YES

YES

#### **CONSEJO**

promotion\_date text

text

text

text

int(11)

varchar(50) YES

area

salarv

type

start\_date

finish\_date

ALTER TABLE limpieza CHANGE COLUMN: se utiliza para cambiar el nombre de una columna y su tipo de dato.

ALTER TABLE limpieza MODIFY COLUMN: se utiliza para modificar el tipo de datos y otras propiedades de una columna.

```
alter table limpieza modify column type text;
```

NULL

NULL

NULL

NULL

NULL

NULL

#### **Ensayo**

type	ejemplo
0	Hybrid
0	Hybrid
1	remote
0	Hybrid
1	remote
0	Hybrid
1	remote
1	remote
0	Hvbrid

#### Actualizar la tabla

### Ajustar formato números

```
$77,719
$135,189
$122,484
$82,100
$114,648
$121,044
$174,195
$66,031
```

#### **Ensayo**

consultar: reemplazar \$ por un vacío y cambiar el separador de mil por vacío.

cast : es para agregar decimales

- cantidad de digitos: 15
- cantidad de decimales: 2
- trim: si es que hay espacios de más
- replace
  - para eliminar \$
  - para eliminar la coma

salary	salary
\$77,719	77719.00
\$135,189	135189.00
\$122,484	122484.00
\$82,100	82100.00
\$114,648	114648.00
\$121,044	121044.00
\$174,195	174195.00
\$66,031	66031.00

#### Actulizar la tabla

```
update limpieza set salary =
cast(trim(replace(replace(salary, '$', ''), ',', '')) as
decimal (15,2));
```

#### Modificar el tipo de dato

```
alter table limpieza modify column salary int null;
```

## Trabajando con fechas

#### Dar formato a la fecha

```
birth_date

06/04/1991

6/29/1984

7/29/1989

9/14/1982

04/11/1994

1/18/1971

11/19/1979

10/14/1988

12/13/1996

3/27/1980
```

#### **Ensayo**

- El símbolo se utiliza porque, por ejemplo, el día puede tener uno o dos dígitos.
- Se utiliza WHEN dos veces porque puede haber filas donde la fecha esté separada por un guion u otro delimitador.

new_birth_date
2019-04-11
2019-03-11
2019-06-07
2019-03-22
2019-07-23
2019-06-02
2019-03-11
2019-06-18
2019-10-25
2019-06-18

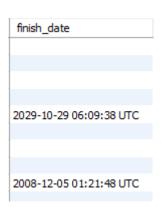
#### Actualizar la tabla

```
update limpieza
set birth_date = case
when birth_date like '%/%'then
```

#### Cambiar el tipo de datos de la columna

alter table limpieza modify column birth\_date date;

## Explorando otras funciones de fecha



#### Objetivo:

- convertirlo al formato de fecha
- Y sacar UTC

### **Prototipo**

```
select finish_date, str_to_date(finish_date, '%Y-%m-%d%H:%i:%s') as fecha from limpieza;
```

finish_date	fecha
	NULL
	NULL
	NULL
	NULL
2029-10-29 06:09:38 UTC	2029-10-29 06:09:38
	NULL
	NULL
	NULL
2008-12-05 01:21:48 UTC	2008-12-05 01:21:48
	NULL
	NULL

### **Prototipo 2**

Para quedarme con el año, mes y dia, nada mas.

```
select finish_date, date_format(str_to_date(finish_date, '%Y-%m-%d %H:%i:%s'), '%Y-%m-%d') as fecha from limpieza;
```

fecha
NULL
NULL
NULL
NULL
2029-10-29
NULL
NULL
NULL
2008-12-05 NULL

### Para separar solo la fecha

select finish\_date, str\_to\_date(finish\_date, '%Y-%m-%d') as
fd from limpieza;

finish_date	fd
	NULL
	NULL
	NULL
	NULL
2029-10-29 06:09:38 UTC	2029-10-29
	NULL
	NULL
	NULL
2008-12-05 01:21:48 UTC	2008-12-05

### Separar solo la hora

```
select finish_date, date_format(finish_date, '%H:%i:%s') as
hour_stamp from limpieza;
```

Para obtener solo la hora hay que usar date\_format, si utilizamos
 str\_to\_data, no va a funcionar

finish_date	hour_stamp
	NULL
	NULL
	NULL
	NULL
2029-10-29 06:09:38 UTC	06:09:38
	NULL
	NULL
	NULL
2008-12-05 01:21:48 UTC	01:21:48

# Dividiendo los elementos de la hora en en distintas columnas

finish_date	hora	minutos	segundos	hour_stamp
	NULL	NULL	NULL	NULL
	NULL	NULL	NULL	NULL
	NULL	NULL	NULL	NULL
	NULL	NULL	NULL	NULL
2029-10-29 06:09:38 UTC	06	09	38	06:09:38
	NULL	NULL	NULL	NULL
	NULL	NULL	NULL	NULL
	NULL	NULL	NULL	NULL
2008-12-05 01:21:48 UTC	01	21	48	01:21:48
	NULL	NULL	NULL	NULL

# Hacer una copia de seguridad de una columna

```
alter table limpieza add column date_backup text;
```

Para que los elementos sean los mismos

```
update limpieza set date_backup = finish_date

date_backup

2029-10-29 06:09:38 UTC

2008-12-05 01:21:48 UTC
```

# Renombramos y cambiamos el formato Prototipo

```
select finish_date, str_to_date(finish_date, '%Y-%m-%d%H:%i:%s') as fecha from limpieza;
```

fecha
NULL
NULL
NULL
NULL
2029-10-29 06:09:38
NULL
NULL
NULL
2008-12-05 01:21:48
NULL

#### Actualizar la tabla

```
update limpieza set finish_date = str_to_date(finish_date, '%Y-%m-%d %H:%i:%s UTC')
where finish_date <>'';
```

- <> : significa diferente
- El segundo argumento que termina en UTC tiene que ser igual a la columna original.



# Ahora separar en una columna la fecha y en otra la hora

Primero creamos las columnas:

```
alter table limpieza

add column fecha date,

add column hora time;
```

NOLL NOLL
NOLL NOLL
NOLL NOLL
NOLL NOLL
NOLL NOLL
NOLL NOLL

Para evitar errores debemos llenar las casillas sin nada con valores null:

```
update limpieza set finish_date = null where finish_date =
```

#### 

Todas estas operaciones se realizaron utilizando finish\_date, cuyo formato siempre ha estado en texto y no en datetime. Por lo tanto, ahora vamos a cambiar su formato a datetime, que es el correcto.

```
alter table limpieza modify column finish_date datetime;
```

#### Calculos con fechas

# Conocer la edad de ingresos de nuestros empleados

Primero añadimos una columna para las edades:

```
alter table limpieza add column age int;
```

Se calculo: birth\_date - start\_date

```
select name, birth_date, start_date, timestampdiff(year,
birth_date, start_date) as edad_de_ingreso from limpieza;
```

	name	birth_date	start_date	edad_de_ingreso
•	Kimmy	1991-06-04	2002-01-20	10
	Ignatius	1984-06-29	2019-04-08	34
	Corbie	1989-07-29	2010-10-12	21
	Baxy	1982-09-14	2005-04-10	22
	Terrell	1994-04-11	2010-09-29	16
	Kacie	1971-01-18	2018-09-01	47
	Sandro	1979-11-19	2012-11-08	32
	Eugene	1988-10-14	2007-06-27	18
	Wainwright	1996-12-13	2001-02-20	4
	Dyann	1980-03-27	2005-01-27	24 4
	Grantley	1975-09-06	2004-11-01	29
	Elmore	1966-01-07	2000-12-05	34
	Dud	1984-03-17	2008-09-17	24
	Aguie	1971-11-02	2015-11-25	44
	Katerina	1967-08-20	2019-05-17	51

• timestampdiff() : no toma en cuenta los meses y días.

### Calcular la edad de los empleados actualmente

```
update limpieza
set age = timestampdiff(year, birth_date, curdate())
```

#### Funciones de texto

Si queremos crear una columna email donde tome las iniciales del nombre, del apellido y del tipo (hibrido o remoto):

```
select concat(substring_index(name, ' ', 1), '_',
substring(last_name, 1,2), '.', substring(type, 1, 1),
'@consulting.com') as email from limpieza;
```

- En (name, '', 1), se toma el nombre hasta que se encuentra un espacio. El tercer argumento, que es 1, indica que se detiene la búsqueda en la primera ocurrencia de dicho espacio.
- En (last\_name, 1, 2), se captura el apellido hasta el segundo carácter.
- SUBSTRING(type, 1, 1): Solo se extrae el primer carácter del campo "type".

```
email

Kimmy_Wa.H@consulting.com

Ignatius_Sp.H@consulting.com

Corbie_Bi.r@consulting.com

Baxy_Ma.H@consulting.com

Terrell_Su.r@consulting.com

Kacie_Of.H@consulting.com

Sandro_Ad.r@consulting.com

Eugene_Le.r@consulting.com

Wainwright_Co.H@consulting.com

Dyann_Is.r@consulting.com
```

Creamos la columna email:

```
alter table limpieza add column email varchar(100);
```

#### Actualizamos la tabla:

```
update limpieza set email = concat(substring_index(name, '
', 1), '_', substring(last_name, 1,2), '.', substring(type,
1, 1), '@consulting.com');
```

# Seleccionamos las columnas que deseamos conservar.

```
select Id_emp, name, last_name, age, gender, area, salary,
email, finish_date from limpieza
where finish_date ≤ curdate() or finish_date is null
order by area, name;
```

#### curdate() : fecha actual

	Id_emp	name	last_name	age	gender	area	salary	email	finish_date
•	15-6583103	Abagael	Aindrais	NULL	male	Accounting	173307	Abagael_Ai.r@consulting.com	HULL
	17-1220615	Abagael	Wadeling	NULL	male	Accounting	179857	Abagael_Wa.r@consulting.com	2024-03-04 11:25:30
	91-5463311	Abba	Stanlock	NULL	male	Accounting	125438	Abba_St.r@consulting.com	2022-07-17 14:08:57
	67-6411440	Abbey	Tracey	NULL	male	Accounting	145739	Abbey_Tr.r@consulting.com	NULL
	01-8402838	Abbott	Springham	NULL	male	Accounting	151323	Abbott_Sp.H@consulting.com	NULL
	04-9122557	Abbott	Despenser	NULL	male	Accounting	65890	Abbott_De.H@consulting.com	NULL
	28-5044439	Abelard	Woodcroft	NULL	male	Accounting	143330	Abelard_Wo.r@consulting.com	2023-02-26 03:16:53
	02-3241469	Abey	Cornuau	NULL	female	Accounting	112070	Abey_Co.r@consulting.com	NULL
	87-7466058	Abigael	Stonestreet	NULL	male	Accounting	107350	Abigael_St.r@consulting.com	NULL
	27-8525957	Abrahan	Ducastel	NULL	female	Accounting	77429	Abrahan_Du.H@consulting.com	NULL
				MILIT					MITT

# Contar la cantidad de empleados que hay en cada area

```
select area, count(*) as cantidad_empleados from limpieza group by area order by cantidad_empleados desc;
```

area	cantidad empleados
Engineering	6686
Accounting	3333
Sales	1832
Human Resources	1807
Training	1692
Services	1686
Business Development	1642
Research and Development	1084
Support	954
Product Management	641
Marketing	494
Legal	311
Auditing	52

# **Exportar datos**

Para exportar la configuración anterior o la de arriba de los empleados, procedemos a:

