# limpieza de datos con SQL

create database if not exists clean;

use clean;

select \* from limpieza limit 10;

# para evitar copiar el select \* from limpieza, creo un STORE PROCEDURE.

select \* from limpieza;

DELIMITER //

CREATE PROCEDURE limp()

BEGIN

select \* from limpieza;

END//

DELIMITER;

CALL limp(); #cada vez que haga este llamado me va a salir el select \* from limpieza;

#UNA DE LAS TAREAS QUE DEBO REALIZAR ES RENOMBRAR LOS NOMBRE DE LAS COLUMNAS

#cambio el ID

ALTER TABLE limpieza CHANGE COLUMN `ï»¿Id?empleado` ID\_emp varchar(20) null;

#cambio GENERO

ALTER TABLE limpieza CHANGE COLUMN `gÃ©nero` Genero varchar(20) null;

#Me fijo si hay valores duplicados, en ese caso los remuevo -- USO GROUP BY Y HAVING

SELECT ID\_emp, count(\*) as cantidad\_duplicados

from limpieza

group by id\_emp

having count(\*) >1;

#cuento cuanto registros hay duplicados

select count(\*) as cantidad\_duplicados

from (

SELECT ID\_emp, count(\*) as cantidad\_duplicados

from limpieza

group by id\_emp

having count(\*) >1

) as subquery;

#ELIMINO DUPLICADOS: 1) renombro la tabla, 2) creo una tabla temporal, 3) convierto la tabla temporal en permanente

rename table limpieza to conduplicados;

create temporary table temp\_limpieza as

select distinct \* from conduplicados;

select count(\*) as original from conduplicados;

#en esta tabla me arroja ('22223') registros

select count(\*) as original from temp\_limpieza;

#esta consulta me arroja ('22214'), siendo esta la que esta sin duplicados y la voy a pasar de forma permanente

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

call limp();

select count(\*) from limpieza;

# Elimino la tabla con duplicados

drop table conduplicados;

set sql\_safe\_updates = 0;

set sql\_safe\_updates = 1;

alter table limpieza change column `Apellido` last\_name varchar(50) null;

alter table limpieza change column `star\_date`start\_date varchar(50) null;

alter table limpieza change column `Genero` Gender varchar (50) null;

#Propiedades de la tabla

DESCRIBE limpieza;

# las columnas en formato texto las paso a formato varchar, y luego las columnas que van con fecha que estan en texto las paso a formato fecha

call limp();

#TRIM Funciona para eliminar los espacios que estan antes o despues del texto

select name from limpieza

where length(name) - length(trim(name)) > 0;

select name, trim(name) as name

from limpieza

where length(name) - length(trim(name)) > 0;

update limpieza set name = trim(name)

where length(name) - length(trim(name)) > 0;

select last\_name from limpieza

where length(last\_name) - length(trim(last\_name)) > 0;

select last\_name, trim(last\_name) as last\_name

from limpieza

where length(last\_name) - length(trim(last\_name)) > 0;

update limpieza set last\_name = trim(last\_name)

where length(last\_name) - length(trim(last\_name)) > 0;

# VOY A INTRODUCIR ESPACIOS A PROPOSITO PARA LUEGO REMOVERLOS

update limpieza set area = replace(area," "," ");

call limp();

# para identificar si hay dos o mas espacios entre las palabras

select area from limpieza where area regexp '\\s{2,}'; #el 2 indica la cantidad de espacios

select area, trim(regexp\_replace(area,'\\s+',' ')) as ensayo from limpieza;

#actualizo mi tabla de manera definitiva

update limpieza set area = trim(regexp\_replace(area,'\\s+',' '));

call limp();

# todo mi set esta en ingles, pero genero me aparece en español. por lo tanto lo voy a pasar a ingles.

select gender,

case

when gender = "hombre" then "male"

when gender = "mujer" then "female"

else "other"

end as gender1 from limpieza;

update limpieza set gender = CASE

when gender = "hombre" then "male"

when gender = "mujer" then "female"

else "other"

END;

call limp();

# LA COLUMNA TYPE MUESTRA VALORES BOOLEANOS, ALGUIEN EXTERNO QUE NO SABE NADA ACERCA DE ESTA BASE DE DATOS ESTA COLUMNA LE VA A RESULTAR IRRELEVANTE

# TYPE HACE REFERENCIA AL TIPO DE CONTRATO QUE TIENE EL EMPLEADO

# O = TIENE UN TRABAJO HIBRIDO

# 1 = TIENE TRABAJO DE TIPO REMOTO

DESCRIBE limpieza;

#type esta en formato int, lo voy a pasar a texto.

alter table limpieza modify column type text;

select type,

case

when type = 1 then "Remote"

when type = 0 then "Hybrid"

else "other"

end as ejemplo

from limpieza;

update limpieza

set type =

case

when type = 1 then "Remote"

when type = 0 then "Hybrid"

else "other"

end;

call limp();

describe limpieza;

#dar formato de numer a un texto, ya que salario esta en text y es int

#Ensayo

select salary,

cast(trim(replace(replace(salary, "$",""),",","")) as decimal (15,2)) as salary1 from limpieza;

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

call limp();

alter table limpieza modify column salary int null;

describe limpieza;

#voy a trabjar con las fechas, birth date, start date y finish date estan en text y tienen que ir en formato fecha.

select birth\_date from limpieza;

/\*lo primero que hago es usar la funcion str\_to\_date que lo que hace es convertir el valor de una cadena en formato mes año a un formato de fecha aceptado por sql

la otra funcion es date\_format que formatea el objeto de fecha en una cadena con el formato año, mes, dia.\*/

select birth\_date, case

when birth\_date like "%/%" then date\_format(str\_to\_date(birth\_date,"%m/%d/%y"),"%Y-%m-%d")

when birth\_date like "%-%" then date\_format(str\_to\_date(birth\_date,"%m-%d-%y"),"%Y-%m-%d")

else null

end as new\_birth\_date

from limpieza;

UPDATE limpieza

SET birth\_date =

CASE

WHEN birth\_date LIKE "%/%" THEN DATE\_FORMAT(STR\_TO\_DATE(birth\_date, "%m/%d/%Y"), "%Y-%m-%d")

WHEN birth\_date LIKE "%-%" THEN DATE\_FORMAT(STR\_TO\_DATE(birth\_date, "%m-%d-%Y"), "%Y-%m-%d")

ELSE NULL

END;

select brith\_date from limpieza

where birth\_date = "06/04/1991";

call limp();

alter table limpieza modify column birth\_date date;

#hago lo mismo para las columnas START\_DATE Y FINISH\_DATE

select start\_date, case

when start\_date like "%/%" then date\_format(str\_to\_date(start\_date,"%m/%d/%y"),"%Y-%m-%d")

when start\_date like "%-%" then date\_format(str\_to\_date(start\_date,"%m-%d-%y"),"%Y-%m-%d")

else null

end as new\_start\_date

from limpieza;

update limpieza

set start\_date =

case

when start\_date like "%/%" then date\_format(str\_to\_date(start\_date,"%m/%d/%y"),"%Y-%m-%d")

when start\_date like "%-%" then date\_format(str\_to\_date(start\_date,"%m-%d-%y"),"%Y-%m-%d")

else null

END;

UPDATE limpieza

SET start\_date =

CASE

WHEN start\_date LIKE "%/%" THEN DATE\_FORMAT(STR\_TO\_DATE(start\_date, "%m/%d/%Y"), "%Y-%m-%d")

WHEN start\_date LIKE "%-%" THEN DATE\_FORMAT(STR\_TO\_DATE(start\_date, "%m-%d-%Y"), "%Y-%m-%d")

ELSE NULL

END;

alter table limpieza modify column start\_date date;

describe limpieza;

select finish\_date, STR\_TO\_DATE(finish\_date, '%Y-%m-%d %H:%i:%s') as fecha from limpieza;

select finish\_date, date\_format(STR\_TO\_DATE(finish\_date, '%Y-%m-%d %H:%i:%s'),'%Y-%m-%d') as fecha from limpieza;

update limpieza set finish\_date = str\_to\_date(finish\_date, '%Y-%m-%d %H:%i:%s UTC')

where finish\_date <> '';

alter table limpieza

add column fecha date,

add column hora time;

call limp();

update limpieza

set fecha = date(finish\_date),

hora = time(finish\_date)

where finish\_date is not null and finish\_date <> '';

update limpieza set finish\_date = null where finish\_date = '';

alter table limpieza modify column finish\_date datetime;

describe limpieza;

alter table limpieza add column age int;

call limp();

update limpieza

set age = timestampdiff(year, birth\_date, curdate()); #calculo la edad de los empleados