



Fundamentos Tecnologías Emergentes con Python



@SENAComunica

www.sena.edu.co



Limpieza de datos con Python

José Fernando Galindo Suárez

jgalindos@sena.edu.co

2023

Fundamentos Tecnologías Emergentes con Python



PROCESO DE GESTIÓN DE FORMACIÓN PROFESIONAL INTEGRAL

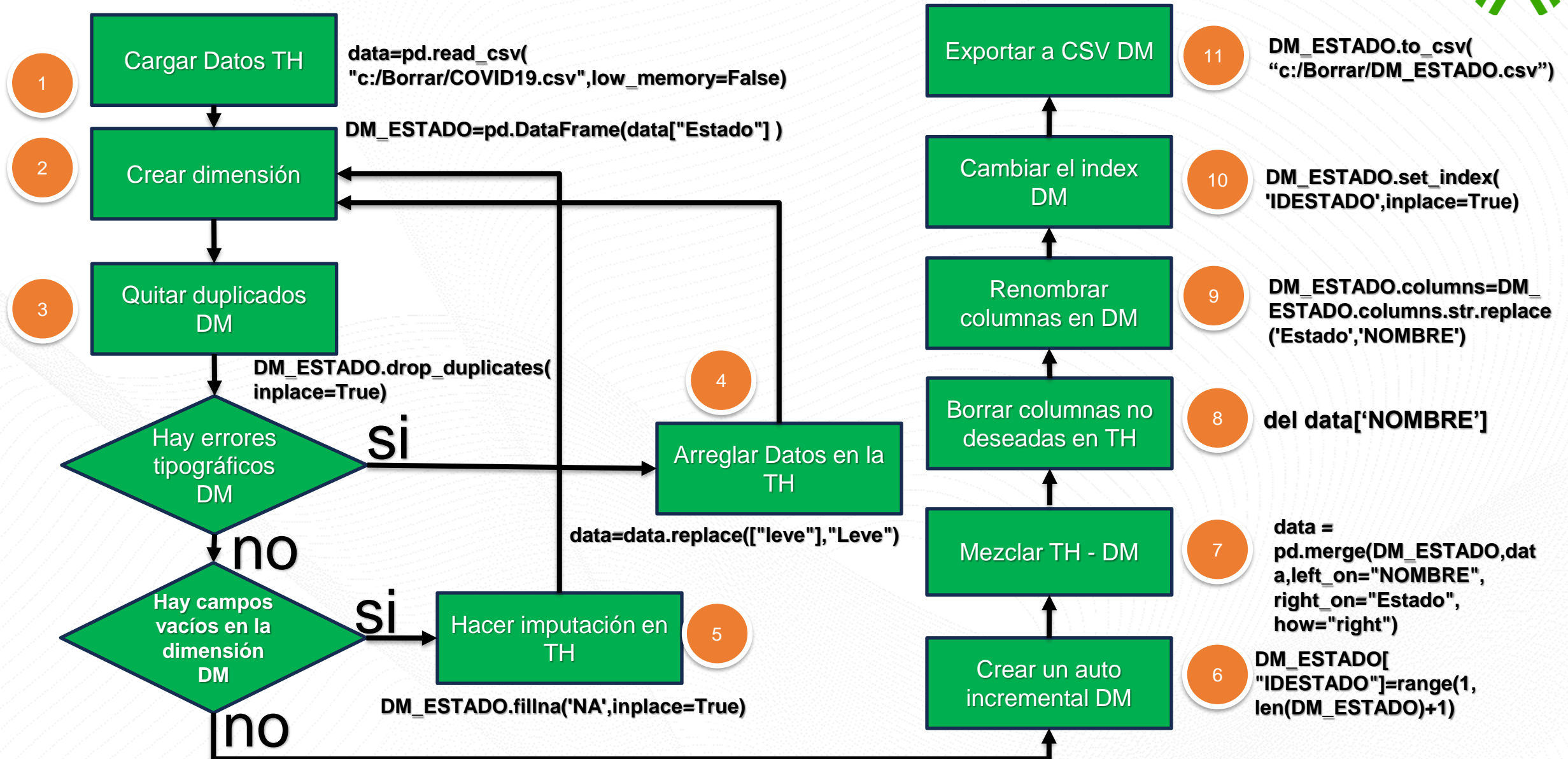
FORMATO GUÍA DE APRENDIZAJE

FUNDAMENTOS DE TECNOLOGÍAS EMERGENTES (PYTHON)

IDENTIFICACIÓN DE LA GUÍA DE APRENDIZAJE

- Denominación del Programa de Formación: ANÁLISIS Y DESARROLLO DE SOFTWARE.
- Código del Programa de Formación: 228118
- Nombre del Proyecto: 2605752 DESARROLLO DE SOFTWARE PARA EL CONTROL DE PROCESOS ORIENTADOS A MICROSERVICIOS.
- Fase del Proyecto: EVALUACIÓN
- Actividad de Proyecto: REALIZAR ACTIVIDADES DE VERIFICACIÓN DE CALIDAD DEL SOFTWARE
- Competencia: 220501098 - ADOPCIÓN DE BUENAS PRÁCTICAS EN EL PROCESO DE DESARROLLO DE SOFTWARE.
- Resultados de Aprendizaje Alcanzar:
- 220501098 02 VERIFICAR LA CALIDAD DEL SOFTWARE DE ACUERDO CON LAS PRÁCTICAS ASOCIADAS EN LOS PROCESOS DE DESARROLLO.
- Duración de la Guía: 66 Horas (DIRECTO + INDEPENDIENTE)

Fundamentos Tecnologías Emergentes con Python



Cargando el dataset



```
import pandas as pd
ruta="c:/Borrar/"
# Cargar los datos
data=pd.read_csv("c:/Borrar/COVID19.csv",low_memory=False)
```

Descargar DATASET **COVID19**

Cargando el dataset



Spyder (Python 3.10)

File Edit Search Source Run Debug Consoles Projects Tools View Help

C:\Users\Administrador\dos.py

temp.py x PRUEBA.py x ManejoFechas.py x dos.py* x uno.py x

```
1 # -*- coding: utf-8 -*-
2 """
3 Created on Thu Oct 12 20:27:23 2023
4
5 @author: Administrador
6 """
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
```

data - DataFrame

Index	IDTIPO	ATENCIO	IDPAIS	DESTADC	ID	FECHA	DCIUAD
0	1	1	1	1	1	2020-03-02T00:00:00.000	11001
1	1	1	2	1	2	2020-03-06T00:00:00.000	76111
2	1	1	2	1	3	2020-03-07T00:00:00.000	5001
3	2	1	3	1	4	2020-03-09T00:00:00.000	5001
4	2	1	3	1	5	2020-03-09T00:00:00.000	5001
5	2	1	3	1	6	2020-03-10T00:00:00.000	5360
6	1	1	4	1	7	2020-03-08T00:00:00.000	13001
7	1	1	2	1	8	2020-03-09T00:00:00.000	11001
8	1	1	2	1	9	2020-03-08T00:00:00.000	11001

Format Resize Background color Column min/max Save and Close Close

C:\Users\Administrador

Name	Type	Size	Value
data	DataFrame	(400489, 9)	Column names: IDTIPO, IDATENCION, IDP...
DM_ATENCION	DataFrame	(6, 2)	Column names: NOMBRE, IDATENCION
DM_CIUAD	DataFrame	(937, 2)	Column names: NOMBRE, IDDPOT
DM_DEPARTAMENTO	DataFrame	(400489, 1)	Column names: NOMBRE
DM_ESTADO	DataFrame	(7, 1)	Column names: NOMBRE
DM_PAIS	DataFrame	(50, 1)	Column names: NOMBRE
DM_TIPO	DataFrame	(3, 2)	Column names: Tipo, IDTIPO
ruta	str	10	c:/Borrar/
TH_CASOS	DataFrame	(400489, 6)	Column names: IDATENCION, IDCIUAD, ID, IDESTADO, FECHA, IDPAIS

Help Variable Explorer Plots Files

Console 1/A x

```
c:\users\administrador\dos.py:36: FutureWarning: Passing a set as an indexer is deprecated and will raise in a future version. Use a list instead.
DM_CIUAD=pd.DataFrame(data[{"Código DIVIPOLA","Ciudad de ubicación"}] )
c:\users\administrador\dos.py:47: FutureWarning: Passing a set as an indexer is deprecated and will raise in a future version. Use a list instead.
DM_DEPARTAMENTO=pd.DataFrame(data[{'IDCIUAD',"Departamento o Distrito"}] )
c:\users\administrador\dos.py:98: FutureWarning: Passing a set as an indexer is deprecated and will raise in a future version. Use a list instead.
TH_CASOS=pd.DataFrame(data[{"ID","FECHA","IDCIUAD","IDESTADO","IDATENCION","IDPAIS"}])

In [67]:
```

IPython Console History

conda: base (Python 3.10.9) Completions: conda LSP: Python Line 14, Col 1 UTF-8 CRLF RW Mem 35%

Cambiar el nombre de una columna



```
data.columns=data.columns.str.replace('ID de caso','ID')
```

A screenshot of a Jupyter Notebook interface showing a DataFrame viewer. The window title is "data - DataFrame". The table has 9 columns: Index, IDTIPO, ATENCIO, IDPAIS, DESTADC, ID, FECHA, DCIUDAE, and an unlabeled column on the far right. The 'ID' column is highlighted with a red box. A red arrow points from the 'DESTADC' column header to the 'ID' column header. The table contains 9 rows of data.

Index	IDTIPO	ATENCIO	IDPAIS	DESTADC	ID	FECHA	DCIUDAE	
0	1	1	1	1	1	2020-03-02T00:00:00.000	11001	19
1	1	1	2	1	2	2020-03-06T00:00:00.000	76111	34
2	1	1	2	1	3	2020-03-07T00:00:00.000	5001	50
3	2	1	3	1	4	2020-03-09T00:00:00.000	5001	55
4	2	1	3	1	5	2020-03-09T00:00:00.000	5001	25
5	2	1	3	1	6	2020-03-10T00:00:00.000	5360	27
6	1	1	4	1	7	2020-03-08T00:00:00.000	13001	85
7	1	1	2	1	8	2020-03-09T00:00:00.000	11001	22
8	1	1	2	1	9	2020-03-08T00:00:00.000	11001	28

At the bottom of the viewer, there are buttons for "Format", "Resize", "Background color", "Column min/max", "Save and Close", and "Close".

Grabar la dimensión DM_ESTADO en un archivo CSV

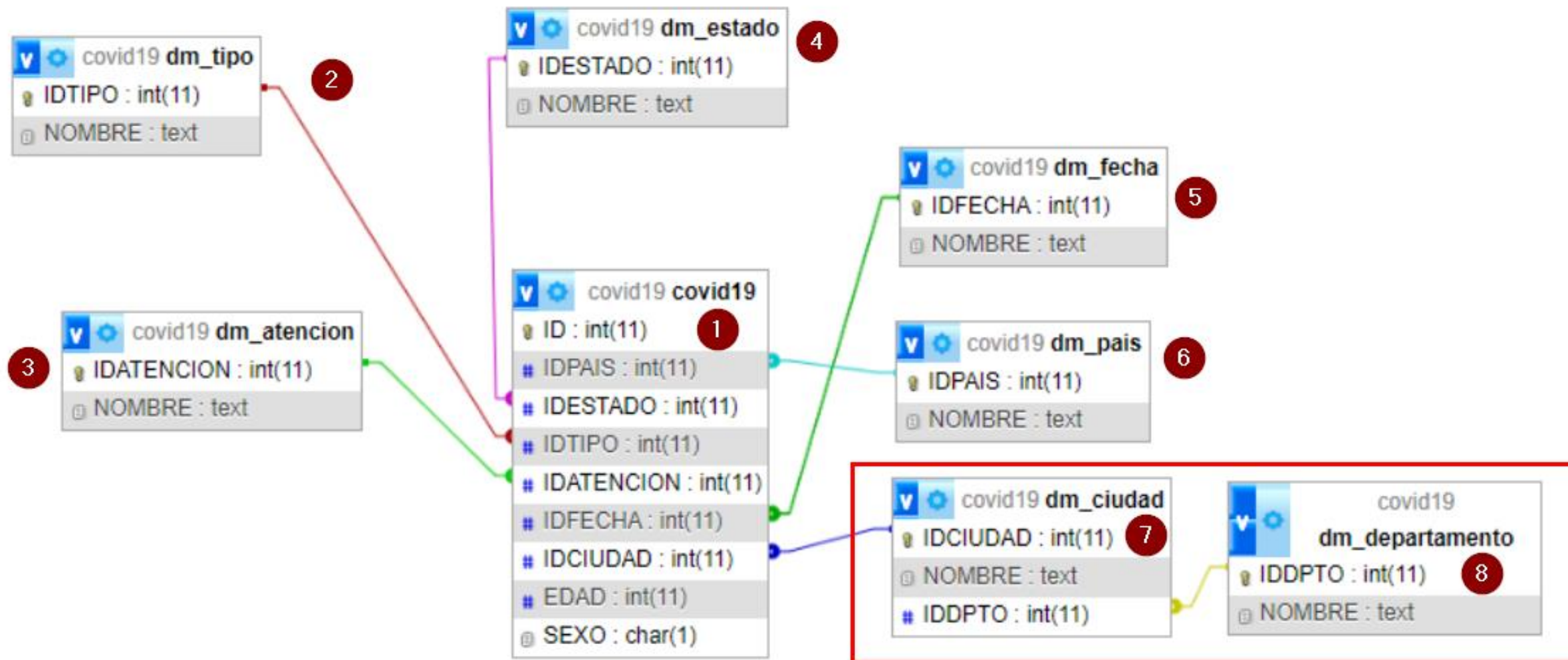


```
# Define a 'IDESTADO' como índice
DM_ESTADO.set_index('IDESTADO',inplace=True)
# Guarda a archivo plano
DM_ESTADO.to_csv(ruta+"DM_ESTADO.csv")
```

C: > Borrar >  DM_ESTADO.csv

```
1  IDESTADO,NOMBRE
2  1,Leve
3  2,Asintomático
4  3,Fallecido
5  4,NA
6  5,Moderado
7  6,Grave
```


Modelo Estrella y Copo de Nieve



Modelo Fisico de la Dimension y carga de datos



1

```
CREATE TABLE DM_TIPO(  
IDTIPO INTEGER PRIMARY KEY AUTO_INCREMENT,  
NOMBRE VARCHAR(30)  
);
```

2

```
ALTER TABLE COVID19 ADD CONSTRAINT TIPOFK FOREIGN  
KEY(IDTIPO)  
REFERENCES DM_TIPO(IDTIPO);
```

3

```
LOAD DATA INFILE 'c:/Borrar/DM_TIPO.csv' INTO TABLE DM_TIPO  
FIELDS TERMINATED BY ','  
LINES TERMINATED BY '\n'  
IGNORE 1 LINES  
(IDTIPO,NOMBRE);
```

VALIDAR EL MODELO DE BASE DE DATOS



```
MariaDB [covid19]> SELECT * FROM VCOVID19 ;
```

DM_ATENCION	FILAS
DM_ATENCION	6
DM_CIUDAD	937
DM_DEPARTAMENTO	33
DM_ESTADO	6
DM_PAIS	44
DM_TIPO	3
DM_FECHA	149
TH_COVID19	400489

```
8 rows in set (0.185 sec)
```

```
CREATE OR REPLACE VIEW VCOVID19 AS
SELECT 'DM_ATENCION',COUNT(*) FILAS FROM DM_ATENCION
UNION
SELECT 'DM_CIUDAD',COUNT(*) FILAS FROM DM_CIUDAD
UNION
SELECT 'DM_DEPARTAMENTO',COUNT(*) FILAS FROM DM_DEPARTAMENTO
UNION
SELECT 'DM_ESTADO',COUNT(*) FILAS FROM DM_ESTADO
UNION
SELECT 'DM_PAIS',COUNT(*) FILAS FROM DM_PAIS
UNION
SELECT 'DM_TIPO',COUNT(*) FILAS FROM DM_TIPO
UNION
SELECT 'DM_FECHA',COUNT(*) FILAS FROM DM_FECHA
UNION
SELECT 'TH_COVID19',COUNT(*) FILAS FROM COVID19
;
```

VALIDAR CONSTRAINT EN LA BASE DE DATOS



```
CREATE OR REPLACE VIEW ESQUEMA AS
SELECT
CONSTRAINT_NAME, TABLE_SCHEMA, TABLE_NAME, CONSTRAINT_TYPE
FROM INFORMATION_SCHEMA.TABLE_CONSTRAINTS
WHERE CONSTRAINT_SCHEMA='COVID19';
```

```
MariaDB [covid19]> SELECT * FROM ESQUEMA;
```

CONSTRAINT_NAME	TABLE_SCHEMA	TABLE_NAME	CONSTRAINT_TYPE
PRIMARY	covid19	covid19	PRIMARY KEY
ATENCIONFK	covid19	covid19	FOREIGN KEY
CIUDADFK	covid19	covid19	FOREIGN KEY
ESTADOFK	covid19	covid19	FOREIGN KEY
PAISFK	covid19	covid19	FOREIGN KEY
TIPOFK	covid19	covid19	FOREIGN KEY
PRIMARY	covid19	dm_atencion	PRIMARY KEY
PRIMARY	covid19	dm_ciudad	PRIMARY KEY
DPTOFK	covid19	dm_ciudad	FOREIGN KEY
PRIMARY	covid19	dm_departamento	PRIMARY KEY
PRIMARY	covid19	dm_estado	PRIMARY KEY
PRIMARY	covid19	dm_fecha	PRIMARY KEY
PRIMARY	covid19	dm_pais	PRIMARY KEY
PRIMARY	covid19	dm_tipo	PRIMARY KEY

```
14 rows in set (0.043 sec)
```

RETO EVIDENCIA



DESARROLLE LAS SIGUIENTES ACTIVIDADES:

- **LA DIMENSIONES CON PYTHON DE ACUERDO AL MODELO DE DATOS**
- **EXPORTE LOS DATOS DE CADA DIMENSION A UN ARCHIVO PLANO**
- **CONSTRUYE LA BASE DE DATOS COVID19 EN MYSQL**
- **CONSTRUYA LAS TABLAS DE CADA DIMENSION Y LA TABLA DE HECHO EN MYSQL**
- **CARGUE LOS DATOS DE CADA DIMENSION Y LA TABLA DE HECHO EN MYSQL**

ENTREGUE AL INSTRUCTOR EL SCRIPT DE PYTHON, SQL Y LA GRAFICA DEL MODELO DE DATOS DESDE PHPMYADMIN EN UN ARCHIVO COMPRIMIDO LLAMADO “EVIDENCIA1.zip”



GRACIAS

Línea de atención al ciudadano: 01 8000 910270
Línea de atención al empresario: 01 8000 910682



www.sena.edu.co