Universidad Tecnológica de Puebla

Ingeniería en Desarrollo y Gestión de Software

9° Cuatrimestre

Grupo: “D”

PRODUCTO 2

Materia: “Extracción de conocimiento en bases de datos”

Docente:

José Francisco Espinoza Garita

Evaluado:

Salma Cid Morales

Periodo:
Mayo – Agosto 2022

Contenido

Tabla de ilustraciones

Introducción

En el siguiente documento se mostrará la forma de realizar un análisis el cual es una ciencia que se encarga de examinar un conjunto de datos con el que se podrán comentar y mostrar diversos temas como lo son:

- Data warehouse
- Tipos y fuentes de datos.
- Técnicas de limpieza de datos.
- Parámetros de configuración del data warehouse.

Mostrando que cada uno de esos temas nos llevaran a un mejor análisis de datos e información para poder comprender dichos datos de la forma mas correcta, con el propósito de sacar conclusiones sobre la información para poder tomar decisiones, o simplemente ampliar los conocimientos sobre diversos temas.

Caso de estudio

Desde el año 2019 a finales del mes de Marzo comenzó en nuestro país una pandemia con un virus llamado Covid-19 el cual se había extendido por todo el mundo, los efectos de la pandemia y el virus que en ese tiempo era considerado como un virus “mortal” en los primeros meses y el año ya que no había ninguna información realmente certera de este virus.

A lo largo de estos años y a lo largo de todos los casos que se presentaron en nuestra república se realizaron informes y recopilación de todos los casos que se presentaron, de esa forma se logró crear un archivo .csv con todos esos datos recabados hasta el mes de Abril del presente año.

Dichos estos aspectos, se necesita hacer un análisis de la información recabada, debido a que es una gran cantidad de datos por analizar, se ha decidido realizar el análisis solo a un par de estados, en este caso a Chiapas e Hidalgo.

Objetivo y alcance

El objetivo de este caso de estudio es realizar una extracción de conocimientos basándonos en un análisis de datos de los casos de covid que se lograron recabar en la “base de datos”.

El alcance de dicho análisis a realizar solo llegará por el momento a los estados de Chiapas e Hidalgo, ya que al ser una gran cantidad de información almacenada no se podría hacer un análisis adecuado, y al solo tomar en cuenta a esos estados el análisis podrá ser mas exacto y preciso, pero sin olvidar que los datos dentro de los estados de Chiapas e Hidalgo no son pocos.

Esquema de data warehouse

Un data warehouse es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso.

Data Warehouse es una arquitectura de almacenamiento de datos que **permite a los ejecutivos de negocios organizar, comprender y utilizar sus datos para**

tomar decisiones estratégicas. Un data warehouse es una arquitectura conocida ya en muchas empresas modernas.

Esquema de estrella.

En el esquema de estrella, el centro de la estrella puede tener una tabla de hechos y varias tablas de dimensiones asociadas. Se conoce como esquema estelar ya que su estructura se asemeja a una estrella. El esquema en estrella es el tipo más simple de esquema de Data Warehouse. También se conoce como Star Join Schema y está optimizado para consultar grandes conjuntos de datos.

Características del esquema estelar:

- Cada dimensión en un esquema de estrella se representa con la única tabla de una dimensión.
- La tabla de dimensiones debe contener el conjunto de atributos.
- La tabla de dimensiones se une a la tabla de hechos utilizando una clave foránea.
- Las tablas de dimensiones no están unidas entre sí.
- La tabla de hechos contendría clave y medida.
- El esquema Star es fácil de entender y proporciona un uso óptimo del disco.
- Las tablas de dimensiones no están normalizadas. Por ejemplo, en la figura anterior, Country_ID no tiene la tabla de búsqueda Country como lo tendría un diseño OLTP.
- El esquema es ampliamente compatible con BI Tools.

El esquema a utilizar es el esquema de estrella ya que no tiene mucha complejidad al ser realizado y analizado, el cual se muestrea en la siguiente imagen.

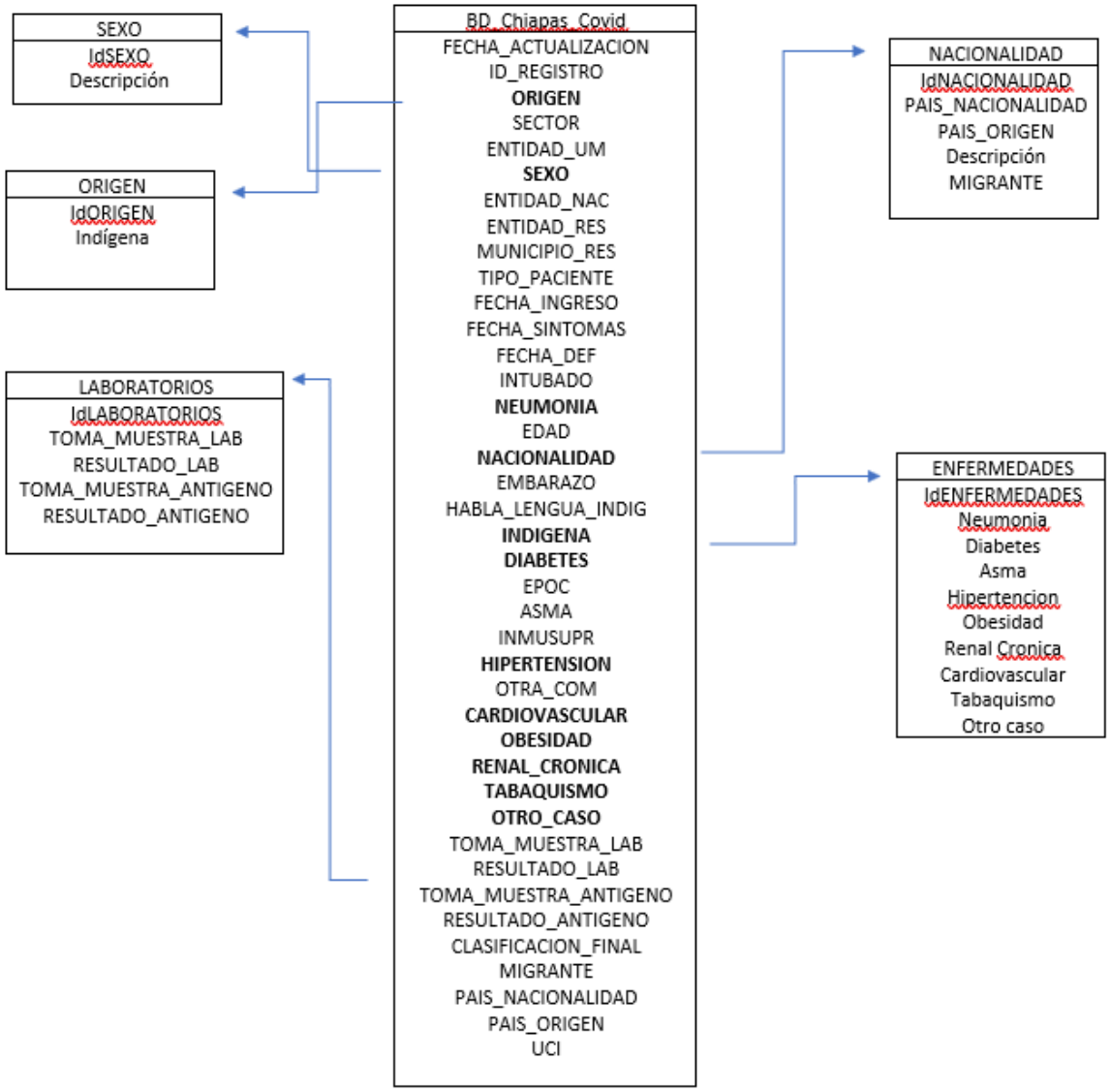

Ilustración 1 Esquema de estrella

Tipos de datos

Un dato es la representación de una variable que puede ser cuantitativa o cualitativa que **indica un valor que se le asigna a las cosas y se representa a través de una secuencia de símbolos, números o letras**.

Tomaremos como referencia dos tipos de datos:

- **Cuantitativos:** Comprenden etiquetas o nombres que se usan para identificar un atributo de cada elemento.
    - Continuos: Cuando, entre dos valores cualesquiera, puede haber valores intermedios. Es decir, se toman todos los valores de un determinado intervalo. Por ejemplo: peso de las personas, nivel sobre el mar en que se encuentra tu ciudad, medida del perímetro torácico.
    - **Discretos:** Cuando se toman valores aislados. Por ejemplo: número de amigos de tu pandilla, número de veces que vas al cine al mes, número de coches que tiene tu familia.
- **Cualitativos:** Los datos cuantitativos requieren valores numéricos que indiquen cuanto o cuantos.
    - Binominales: Los datos binarios colocan las cosas en una de dos categorías mutuamente
    - Nominales: Los **datos nominales** son **datos** "etiquetados" o "nombrados" que pueden dividirse en varios grupos que no se traslapan. En este caso, los **datos** no se miden ni se evalúan, sino que se asignan a varios grupos. Estos grupos son únicos y no tienen elementos comunes.
    - Ordinales: Los datos ordinales son un tipo de datos estadísticos categóricos donde las variables tienen categorías naturales ordenadas y no se conocen las distancias entre las categorías.

Fuentes de datos

Una fuente de datos carga datos estáticos o casi en tiempo real en un análisis en tiempo real o de big data.

Las fuentes de datos cargan datos utilizados junto con herramientas que requieren un dataset espacial o tabular auxiliar para enriquecer, filtrar, unir o calcular la distancia entre eventos. En los análisis de big data, las fuentes de datos cargan los datos que se procesarán con las herramientas y se escribirán en las salidas.

Categorías de las fuentes de origen de datos, cada una de las cuales tiene a su vez un número de fuentes que recolectan, almacenan, procesan y analizan.

- **Biometría**
  - Reconocimiento facial.
  - Genética.
- **Web y medios sociales**
  - Datos de flujos de clicks.
  - Entradas de Facebook.
  - Contenido web.
- **Transacciones de grandes datos**
  - Demandas de salud
  - Llamadas de telecomunicaciones.
  - Registro de detalles
  - Registro de facturaciones
- **Generado por los humanos**
  - Registros de voz en centros de llamadas.
  - Correos electrónicos.
  - Registros médicos electrónicos.
- **Maquina a maquina**
  - Lectura de medidores inteligentes.
  - Lectura de RFID.
  - Lecturas de sensores en plataformas petroleras.
  - Señales de GPS.

Las transacciones de grandes datos, en mi opinión, se reflejan en el archivo .cvs de la "base de datos" con la que se está trabajando, ya que cuenta con toda la información recabada durante los últimos años de pandemia desde el primer caso que se registró de COVID hasta la última actualización de información que se registro dentro del archivo .csv que se está trabajado.

Aunque también creo que los datos generados por los humanos también tienen mucho que ver para la creación de la "base de datos" que se esta trabajando, ya que estos datos incluyen:

- **Documentos electrónicos.**
- **Estudios y registros médicos electrónicos.**
- **Recetas médicas.**

Para la creación del archivo .csv que se esta trabajando obviamente se necesito realizar un estudio y recabar información utilizando varias fuentes de datos para llegar hasta el punto el que se está trabajando con dicha información, por lo que en

mi caso la fuente de datos que se está “ocupando” es el archivo 220430COVID19MEXICO.csv.

Técnica de limpieza de datos

La limpieza de datos es la depuración de datos erróneos en una tabla o base de datos. Esta acción permite identificar datos incorrectos, incompletos o poco relevantes para tu empresa. Después de la limpieza, se sustituyen, modifican o eliminan por completo los datos inservibles.

La limpieza de datos intenta resolver la problemática de la detección y corrección de errores e inconsistencias que ocurren en los datos, con el fin de mejorar su calidad.

Limpieza de datos

La limpieza de datos la comencé al separar los datos de los estados con los que en mi caso se están trabajando:

- **Chiapas.**
- **Hidalgo.**

Se comenzó descargando el archivo datos_abiertos_covid.zip el cual contenía el archivo 220430COVID19MEXICO.csv que tiene la información de todos los estados de la república.

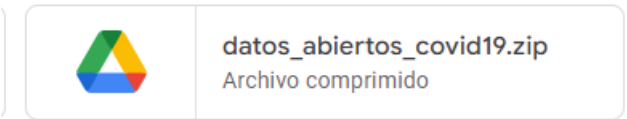

Ilustración 2 datos abiertos covid

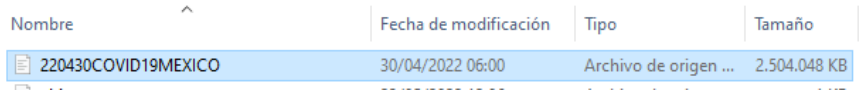

Ilustración 3Archivo covid19

La herramienta de software que se utiliza el ANACONDA Jupyter

Se abre el software y nos aparece una ventana con la que podemos abrir la extensión de Jupyter.
Se selecciona “Launch”.

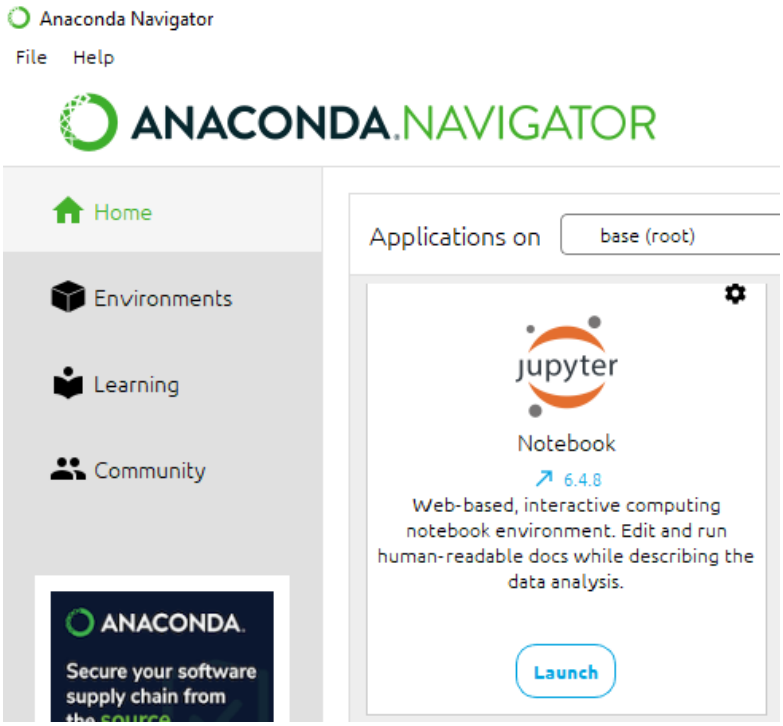

Ilustración 4 Anaconda page home

Se abre una nueva página de navegación, mostrando lo que se encuentra en la página principal de la carpeta de Anaconda.

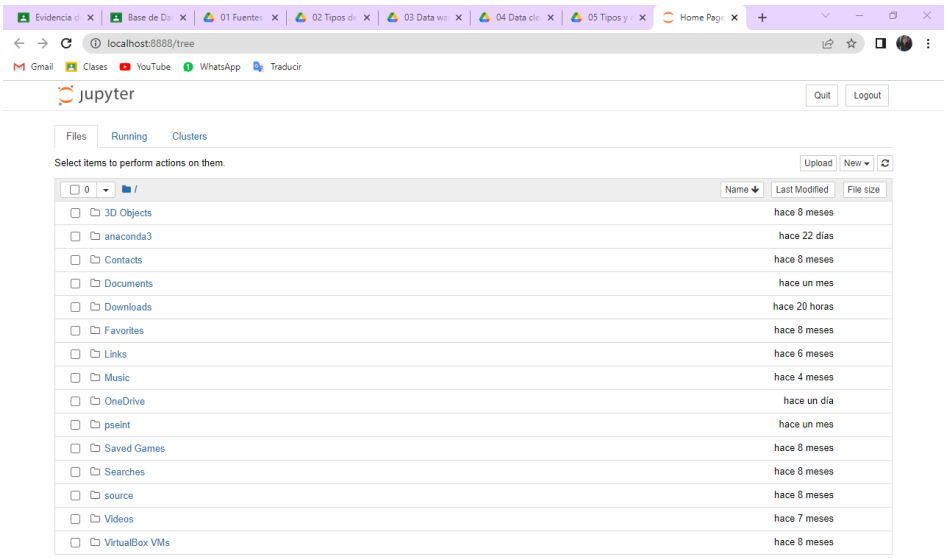

Ilustración 5 Anaconda page inicio Jupyter

En esta parte puedes crear capetas y archivos para trabajar con la información que necesites, en mi caso dentro de la carpeta anaconda3 se hizo otra carpeta "BDCovid", en esa carpeta se guardaran los archivos que se realicen.

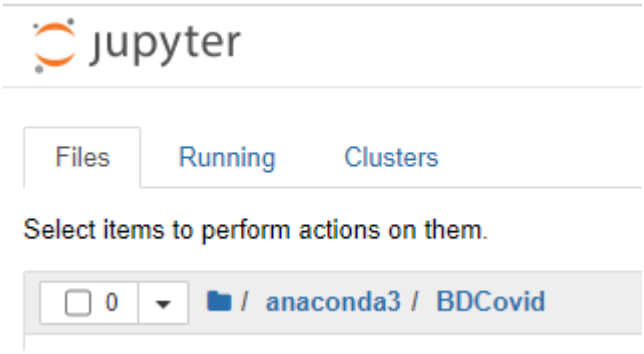

Ilustración 6 Anaconda Ruta de archivos

Se crea un nuevo archivo .ipynb

Ilustración 7 Anaconda creación de documento

En este primer archivo, se separará la información del estado de Chiapas e Hidalgo que son los dos estados con los que se estará trabajando.

Paso 1.

En mi caso necesito comenzar a trabajar el archivo 220430COVID19MEXICO.csv

```
#se establece la ruta donde se sencuetra el archivo csv

In [8]: path = 'C:/covid2020/'
```

Ilustración 8 Extracción de datos1

En mi caso el archivo 220430COVID19MEXICO.csv se encuentra dentro de mi disco C en una carpeta llamada covid2020.

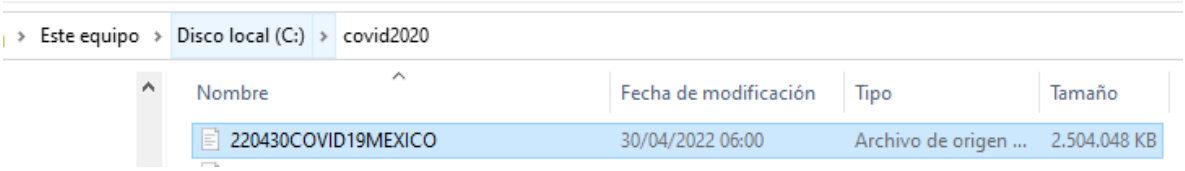

Ilustración 9 Extracción de datos2

Paso 2.

```
#se hace la importacion de las librerias de csv y para obtener el tiempo

import csv
from datetime import datetime
```

Ilustración 10 Extracción de datos3

Paso 3.

se obtiene la fecha actual y se despliega

```
startTime = datetime.now()
print("Fecha: ", startTime)
```

```
Fecha:  2022-05-22 23:49:48.969944
```

Ilustración 11 Extracción de datos4

Paso 4.

#Se establece a cero a la variable para contar el numero de registros tranferidos

```
rowNumber = 0
```

Ilustración 12 Extracción de datos5

Paso 5.

Se establecen los encabezados, se crea una lista y se agrega el encabezado a la lista. La lista va a tener todo el contenido que se vuelca al nuevo archivo.

```
csvHeader =[
    'FECHA_ACTUALIZACION','ID_REGISTRO','ORIGEN','SECTOR','ENTIDAD_UM','SEXO','E
    'ENTIDAD_RES','MUNICIPIO_RES','TIPO_PACIENTE','FECHA_INGRESO','FECHA_SINTOMA
    'INTUBADO','NEUMONIA','EDAD','NACIONALIDAD','EMBARAZO','HABLA_LENGUA_INDIG',
    'EPOC','ASMA','INMUSUPR','HIPERTENSION','OTRA_COM','CARDIOVASCULAR',
    'OBESIDAD','RENAL_CRONICA','TABAQUISMO','OTRO_CASO','TOMA_MUESTRA_LAB','RESU
    'TOMA_MUESTRA_ANTIGENO','RESULTADO_ANTIGENO','CLASIFICACION_FINAL','MIGRANTE
    'PAIS_NACIONALIDAD','PAIS_ORIGEN','UCI']

csvList = []
csvList.append(csvHeader)
```

Ilustración 13 Extracción de datos6

Paso 6.

Se abre el archivo de lectura “covid.csv”. Se lee cada línea y se identifica si el registro pertenece a “Chiapas”. De ser así, se agrega a la lista.

```
with open(path + '220430COVID19MEXICO.csv', newline='') as File:
    reader = csv.reader(File)
    for row in reader:
        if row[4] == '07' or row[6] == '07' or row[7] == '07':
            rowNumber += 1
            csvList.append(row)
```

Ilustración 14 Extracción de datos7

Paso 7.

Se crea el archivo de escritura. Se agrega la lista al archivo. La lista ya trae los registros que pertenecen al estado de “Chiapas”.

```
csvFiltered = open(path + 'chiapas01.csv', 'w',newline="")
with csvFiltered:
    writer = csv.writer(csvFiltered)
    writer.writerows(csvList)
```

Ilustración 15 Extracción de datos8

Paso 8.

Para los datos de estado de Hidalgo se repiten los pasos 6 y 7 con el numero de estado correspondiente y el nombre que se le quiera dar al archivo obtenido.

```
with open(path + '220430COVID19MEXICO.csv', newline='') as File:
    reader = csv.reader(File)
    for row in reader:
        if row[4] == '13' or row[6] == '13' or row[7] == '13':
            rowNumber += 1
            csvList.append(row)
```

Ilustración 16 Extracción de datos9

```
csvFiltered = open(path + 'hidalgo01.csv', 'w',newline="")
with csvFiltered:
    writer = csv.writer(csvFiltered)
    writer.writerows(csvList)
```

Ilustración 17 Extracción de datos10

De esta forma los archivos obtenidos se guardan en la ruta que se estableció en el path.

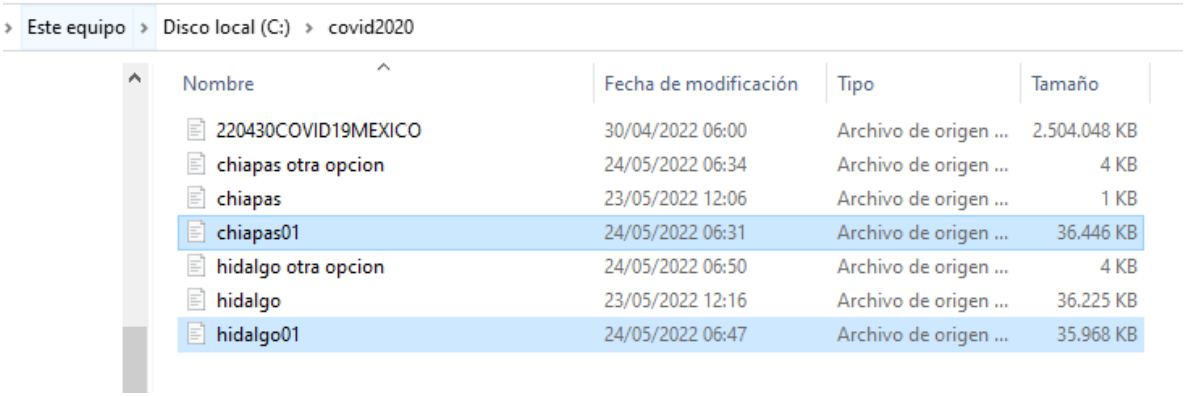

Ilustración 18 Extracción de datos11

Una vez que se hizo la separación de datos se puede comenzar a trabajar con los archivos creados, en un nuevo archivo .ipynb.

En el caso del archivo chiapas01.csv en el apartado de “SEXO” el “femenino y masculino” se establece con los numero 1 y 2. Al realizar un análisis de datos se estableció que los valores de 1 y 2 no son los mas adecuados por lo que se procedió a hacer un cambio de datos y valores.

```
print(dfChiapas)
```

```
       FECHA_ACTUALIZACION ID_REGISTRO  ORIGEN  SECTOR  ENTIDAD_UM  SEXO  \
0               2022-04-30      z59dea       1      12           7     2
1               2022-04-30      z1f605       2      12           7     2
2               2022-04-30      z3f33c       1      12           7     1
3               2022-04-30      z3be8c       2      12           7     1
4               2022-04-30      z21f6f       2      12           7     1
...                    ...         ...     ...     ...         ...   ...
265565          2022-04-30     m110eb5       2      12          15     1
265566          2022-04-30     m1e1d97       2       3          15     1
265567          2022-04-30     m0bc59b       2      12          15     2
265568          2022-04-30     m0a48c3       2      12          15     1
265569          2022-04-30     m1e4cd7       2      12          15     1
```

Ilustración 19 Tabla de datos

Cambio de tipo de datos.

```
Se establece la ruta en la que se encuentra el archivo chiapas01.csv
```

```
path = 'C:/covid2020/'
```

```
Se hace la importanción de la librería de Pandas.
```

```
import pandas as pd
```

```
Se cargan los datos del archivo en el dataframe.
```

```
dfChiapas = pd.read_csv(path + "chiapas01.csv",low_memory=False)
```

Ilustración 20 Cambio de tipo de datos.1

```
Mostramos los tipos de datos.
```

```
dfChiapas.dtypes
```

```
FECHA_ACTUALIZACION      object
ID_REGISTRO              object
ORIGEN                    int64
SECTOR                    int64
ENTIDAD_UM                int64
SEXO                      int64
ENTIDAD_NAC               int64
ENTIDAD_RES               int64
MUNICIPIO_RES             int64
TIPO_PACIENTE             int64
FECHA_INGRESO            object
FECHA_SINTOMAS           object
```

Ilustración 21 Cambio de tipo de datos.2

```
Lo primero es cambiar el tipo de datos y posteriormente los valores.
```

```
dfChiapas['SEXO'] = dfChiapas['SEXO'].astype('string')
dfChiapas['SEXO'] = dfChiapas['SEXO'].replace(["1"], ["Femenino"])
dfChiapas['SEXO'] = dfChiapas['SEXO'].replace(["2"], ["Masculino"])
```

```
Mostramos el tipo de dato asociado a sexo.
```

```
dfChiapas['SEXO'].dtype
```

```
string[python]
```

Ilustración 22 Cambio de tipo de datos.3

```
Finalmente mostramos el contenido del data frame.
```

```
print(dfChiapas)
```

```
       FECHA_ACTUALIZACION ID_REGISTRO  ORIGEN  SECTOR  ENTIDAD_UM       SEXO  \
0               2022-04-30      z59dea       1      12           7  Masculino
1               2022-04-30      z1f605       2      12           7  Masculino
2               2022-04-30      z3f33c       1      12           7   Femenino
3               2022-04-30      z3be8c       2      12           7   Femenino
4               2022-04-30      z21f6f       2      12           7   Femenino
...                    ...         ...     ...     ...         ...        ...
265565          2022-04-30     m110eb5       2      12          15   Femenino
265566          2022-04-30     m1e1d97       2       3          15   Femenino
265567          2022-04-30     m0bc59b       2      12          15  Masculino
265568          2022-04-30     m0a48c3       2      12          15   Femenino
265569          2022-04-30     m1e4cd7       2      12          15   Femenino
```

Ilustración 23 Cambio de tipo de datos.4

De esta forma se logro realizar el cambio de tipo de datos junto con el valor asignado. Demostrando también una limpieza de datos en el campo “SEXO”.

Parámetros de configuración del data warehouse.

Real Time D

La migración gradual hacia la operación en tiempo real es la tendencia actual en la gestión de datos. Esto incluye las disciplinas de gestión de calidad de datos, integración de datos, gestión maestra de datos y el procesamiento de eventos complejos.

Perfilado

Mejorar continuamente la calidad de los datos es un reto cuando no se sabe el estado actual de los datos y su uso. Se deben comprender los datos empresariales a través de profiling, es un punto de partida para decidir qué datos necesitan especial atención.

Técnicas D

DQ es una familia de ocho o más técnicas relacionadas entre sí. La estandarización de datos es el método más comúnmente usado, seguido de verificaciones, validaciones, monitoreo, profiling, matching, y así sucesivamente.

Integrado: los datos almacenados en el Data Warehouse deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.

Temático: sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del Data Warehouse. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.

Histórico: el tiempo es parte implícita de la información contenida en un Data Warehouse. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información

almacenada en el Datawarehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el Data Warehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

No volátil: el almacén de información de un Data Warehouse existe para ser leído, y no modificado. La información es por tanto permanente, significando la actualización del DWH la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

Conclusión

Durante la elaboración de este documento se estudiaron nuevos temas como: Data Warehouse, Tipos y fuentes de datos, Técnicas de limpieza de datos, los cuales me ayudaron a entender mejor como realizar un Análisis de datos mediante la extracción de conocimientos en una “Base de datos”, basándonos en un caso de estudio sobre el covid.

Pude darme cuenta que se necesitan bastantes conocimientos para realizar dicho análisis, que no solamente es leer e interpretar, si no que es saber reflexionar y analizar la información de forma mas exacta que nos permita manipular todos los datos de la forma mas correcta posible.

Bibliografías

Etecé, E. (1 de Octubre de 2020). *concepto*. Obtenido de concepto: https://concepto.de/dato/

Santos, D. (2021). *HubSpot*. Obtenido de HubSpot: https://blog.hubspot.es/marketing/limpieza-de-datos

Software, E. (17 de Mayo de 2021). *evaluandosoftware*. Obtenido de evaluandosoftware: https://www.evaluandosoftware.com/abc-del-data-warehouse/

Technologies, G. K. (10 de Junio de 2020). *grupokorporate*. Obtenido de grupokorporate: https://grupokorporate.com/los-cinco-tipos-de-fuentes-de-datos/

Valbuena, D. F. (s.f.). *datamanagement*. Obtenido de datamanagement: https://datamanagement.es/2020/04/03/esquemas-data-warehousing/