Práctica 1: Exportación de datos y cobertura de espacio y tiempo.

Introducción

En la presente práctica se hace uso de registros de accidentes viales de diferentes tipos, todos ellos sucedidos en la CDMX. Durante la práctica se hará uso del motor de SQL Server Development 2019, utilizando el sistema operativo Windows 10. El dataset ha sido obtenido mediante distintas fuentes a lo largo de todas las alcaldías de la Ciudad de México, en total se cuenta con 33071 registros, sin embargo, durante el análisis de los datos se podrá ver mejor la distribución y significado de los datos

Limpieza de los datos

Los datos recibidos mediante el archivo CSV fueron analizados por encima mediante el uso de la herramienta **EXCEL**, lo que nos permitió importar los datos dentro de la base de datos sin mayor problema, al asignar el tipo de dato columna por columna, nos aseguramos de que no haya problemas de regionalización entre el archivo CSV y el manejador.

Durante la importación de los datos tambien realizaron algunas modificaciones, cambiando todos los campos nvarchar[50] a text, lo que en sentencias posteriores generó problemas para aplicar la sentencia DISTINCT durante las consultas, por lo que se tuvo que aplicar un procesamiento posterior durante la ejecución.

Para consultar la estructura de los datos dentro de la base de datos, puede consultar el siguiente enlace: https://drive.google.com/file/d/1Qp56m44c0X3L4A8XuEt6HtDNIzZ3OK61/view?usp=sharing (https://drive.google.com/file/d/1Qp56m44c0X3L4A8XuEt6HtDNIzZ3OK61/view?usp=sharing)

Desarrollo

1.- Descargue el dataset de incidentes viales, que corresponde al último semestre del 2020.

2.- Exporte el archivo CSV en el manejador de base de datos seleccionado (nota: estudie los videos de instalación de los posibles manejadores)

3.- Indique el número de registros del dataset en el manejador.

In [2]:
```
SELECT COUNT(folio) FROM incidentevial2dsem2020;
```

(1 row affected)

Total execution time: 00:00:00.083

Out[2]:

(No column name)
33071

4.- ¿Cuál es el rango de los campos relacionados (valor minimo y maximo)? Respecto a:

- “Fecha” (todos los relacionados)
- Latitud y longitud
- Año_cierre y hora_cierre (todos los relacionados al cierre”)

Lo primero que procedemos a hacer es buscar todas las columnas que tengan algo que ver con la palabra "Fecha"

In [3]:
```
SELECT Table_Name, Column_Name
FROM INFORMATION_SCHEMA.COLUMNS
WHERE TABLE_CATALOG = 'Practica1'
AND COLUMN_NAME LIKE '%fecha%'
```

(2 rows affected)

Total execution time: 00:00:00.018

Out[3]:

Table_Name	Column_Name
incidentevial2dsem2020	fecha_cierre
incidentevial2dsem2020	fecha_creacion

Hacemos lo mismo para los campos relacionados a "cierre"

In [4]:
```
SELECT Table_Name, Column_Name
FROM INFORMATION_SCHEMA.COLUMNS
WHERE TABLE_CATALOG = 'Practica1'
AND COLUMN_NAME LIKE '%cierre%'
```

(6 rows affected)

Total execution time: 00:00:00.018

Out[4]:

Table_Name	Column_Name
incidentevial2dsem2020	año_cierre
incidentevial2dsem2020	codigo_cierre
incidentevial2dsem2020	delegacion_cierre
incidentevial2dsem2020	fecha_cierre
incidentevial2dsem2020	hora_cierre
incidentevial2dsem2020	mes_cierre

Procedemos a ver el rango máximo y mínimo de valores en los campos correspondientes:

Nota: Para el caso de fecha cierre solo se realizará una vez.

- Para la columna fecha cierre :

In [9]:
```
SELECT MIN(fecha_cierre) Fecha_Minima,
       MAX(fecha_cierre) Fecha_Maxima FROM incidentevial2dsem2020
```

(1 row affected)

Total execution time: 00:00:00.007

Out[9]:

Fecha_Minima	Fecha_Maxima
2020-06-01 00:00:00.0000000	2020-11-26 00:00:00.0000000

Podemos observar que el rango va del primero de mayo hasta el 26 de noviembre en la fecha de cierre.

- Para la columna fecha creación :

In [10]:
```
SELECT MIN(fecha_creacion) Fecha_Minima,
       MAX(fecha_creacion) Fecha_Maxima FROM incidentevial2dsem2020
```

(1 row affected)

Total execution time: 00:00:00.059

Out[10]:

Fecha_Minima	Fecha_Maxima
2020-02-08 00:00:00.0000000	2020-11-26 00:00:00.0000000

Podemos observar que el rango va del 8 de febrero hasta el 26 de noviembre. Siendo este último dato curioso, debido a que la fecha de creación será la misma que la fecha de cierre.

- Para la columna latitud :

In [12]:
```
SELECT MIN(latitud) Latitud_Minima,
       MAX(latitud) Latitud_Maxima FROM incidentevial2dsem2020
```

(1 row affected)

Total execution time: 00:00:00.059

Out[12]:

Latitud_Minima	Latitud_Maxima
19.095427	19.57671

Vemos que los datos no varian demasiado debido a que la extensión territorial de la CDMX no es tán grande.

- Para la columna longitud :

In [13]:
```
SELECT MIN(longitud) Longitud_Minima,
       MAX(longitud) Longitud_Maxima FROM incidentevial2dsem2020
```

(1 row affected)

Total execution time: 00:00:00.053

Out[13]:

Longitud_Minima	Longitud_Maxima
-99.348434	-98.94764

- Para la columna año cierre :

In [11]:
```
SELECT MIN(año_cierre) Año_Minimo,
       MAX(año_cierre) Año_Maximo FROM incidentevial2dsem2020
```

(1 row affected)

Total execution time: 00:00:00.060

Out[11]:

Año_Minimo	Año_Maximo
2020	2020

- Para la columna código cierre : En este caso el tipo de dato no se presta para la operación.
- Para la columna delegación cierre : En este caso el tipo de dato no se presta para la operación.
- Para la columna mes cierre : En este caso el tipo de dato no se presta para la operación.
- Para la columna hora cierre :

In [19]:
```
SELECT MIN(hora_cierre) Hora_Minima,
       MAX(hora_cierre) Hora_Maxima FROM incidentevial2dsem2020
```

(1 row affected)

Total execution time: 00:00:00.052

Out[19]:

Hora_Minima	Hora_Maxima
2021-09-01 00:00:00.0000000	2021-09-01 23:59:59.0000000

5. ¿Cuáles son los valores que toman las siguientes columnas (rango, i.e. valores posibles no repetidos) y que significado tiene (https://datos.cdmx.gob.mx/dataset/incidentes-viales-c5 (https://datos.cdmx.gob.mx/dataset/incidentes-viales-c5))?

- Incidente_c4
- Tipo_entrada
- Clas_con_f_alarma
- Delegación

In [4]:
```
SELECT DISTINCT CONVERT(VARCHAR(MAX), incidente_c4) Rango_Incidente_c4
FROM incidentevial2dsem2020
```

(19 rows affected)

Total execution time: 00:00:00.165

Out[4]:

Rango_Incidente_c4
accidente-choque sin lesionados
cadáver-atropellado
detención ciudadana-accidente automovilístico
accidente-vehículo atrapado-varado
accidente-vehiculo desbarrancado
cadáver-accidente automovilístico
mi ciudad-taxi-incidente de tránsito
detención ciudadana-atropellado
accidente-ciclista
sismo-persona atropellada
accidente-choque con lesionados
accidente-motociclista
accidente-persona atrapada / desbarrancada
accidente-otros
lesionado-atropellado
mi ciudad-calle-incidente de tránsito
accidente-volcadura
sismo-choque con lesionados
accidente-choque con prensados

Estos datos representan los tipos de incidentes reportados en la CDMX, en total son 19 tipos distintos, además, con estos datos podemos además cuantos incidentes resultaron en lesionados, prensados, atropellados, atrapados, muertos, etcétera. Aunque esto no forma parte de la práctica, resulta interesante notar la utilidad de esta columna para obtener más información.

In [5]:
```
SELECT DISTINCT CONVERT(VARCHAR(MAX), tipo_entrada) Rango_tipo_entrada
FROM incidentevial2dsem2020
```

(7 rows affected)

Total execution time: 00:00:00.150

Out[5]:

Rango_tipo_entrada
BOTÓN DE AUXILIO
LLAMADA APP911
APLICATIVOS
RADIO
LLAMADA DEL 911
REDES
CÁMARA

In [6]:
```
SELECT DISTINCT CONVERT(VARCHAR(MAX), clas_con_f_alarma) Rango_clas_con
_f_alarma FROM incidentevial2dsem2020
```

(4 rows affected)

Total execution time: 00:00:00.147

Out[6]:

Rango_clas_con_f_alarma
EMERGENCIA
DELITO
FALSA ALARMA
URGENCIAS MEDICAS

In [1]:
```
SELECT DISTINCT CONVERT(NVARCHAR(MAX), delegacion_inicio) Rango_delegac
ion_inicio FROM incidentevial2dsem2020
```

(17 rows affected)

Total execution time: 00:00:01.479

Out[1]:

Rango_delegacion_inicio
MIGUEL HIDALGO
MILPA ALTA
AZCAPOTZALCO
CUAJIMALPA
TLALPAN
NULL
VENUSTIANO CARRANZA
MAGDALENA CONTRERAS
IZTACALCO
CUAUHTEMOC
XOCHIMILCO
TLAHUAC
GUSTAVO A. MADERO
BENITO JUAREZ
IZTAPALAPA
ALVARO OBREGON
COYOACAN

In [8]:
```
SELECT * FROM incidentevial2dsem2020 WHERE delegacion_inicio IS NULL;
```

(0 rows affected)

Total execution time: 00:00:00.003

Out[8]:

folio	fecha_creacion	hora_creacion	dia_semana	codigo_cierre	fecha_cierre	año_cierre	mes

En la consulta anterior se puede observar que se detecta una delegación NULL, sin embargo, al buscarla, esta no aparece, lo cual puede ser resultado de convertir el texto a varchar

In [2]:
```
SELECT DISTINCT CONVERT(NVARCHAR(MAX), delegacion_cierre) Rango_delegac
ion_cierre FROM incidentevial2dsem2020
```

(17 rows affected)

Total execution time: 00:00:00.076

Out[2]:

Rango_delegacion_cierre
MIGUEL HIDALGO
MILPA ALTA
AZCAPOTZALCO
CUAJIMALPA
TLALPAN
NULL
VENUSTIANO CARRANZA
MAGDALENA CONTRERAS
IZTACALCO
CUAUHTEMOC
XOCHIMILCO
TLAHUAC
GUSTAVO A. MADERO
BENITO JUAREZ
IZTAPALAPA
ALVARO OBREGON
COYOACAN

In [11]:
```
SELECT * FROM incidentevial2dsem2020 WHERE delegacion_cierre IS NULL;
```

(0 rows affected)

Total execution time: 00:00:00.002

Out[11]:

folio	fecha_creacion	hora_creacion	dia_semana	codigo_cierre	fecha_cierre	año_cierre	mes

6. Contar la cantidad de NULL o NULOS encontrados en las 4 columnas anteriores del punto 5.

In [27]:
```
SELECT SUM(CASE WHEN incidente_c4 IS NULL THEN 1 ELSE 0 END) Incidente_
c4,
       SUM(CASE WHEN tipo_entrada IS NULL THEN 1 ELSE 0 END) Tipo_entra
da,
       SUM(CASE WHEN clas_con_f_alarma IS NULL THEN 1 ELSE 0 END) Clas_
con_f_alarma,
       SUM(CASE WHEN delegacion_inicio IS NULL THEN 1 ELSE 0 END) Deleg
acion_inicio,
       SUM(CASE WHEN delegacion_cierre IS NULL THEN 1 ELSE 0 END) Deleg
acion_cierre FROM incidentevial2dsem2020
```

(1 row affected)

Total execution time: 00:00:00.361

Out[27]:

Incidente_c4	Tipo_entrada	Clas_con_f_alarma	Delegacion_inicio	Delegacion_cierre
0	0	0	0	0

7. En las columnas analizadas (del punto 4 al 6) ¿encontró alguna anomalía en los valores?. Si, sí realice una discusión de las columnas en cuestión.

Solo se pudieron ver anomalías al obtener las delegaciones, pareciendo que existe un registro nulo en el sistema, sin embargo, parece ser un error de conversión, porque al almacenarse los datos se especificó que ningún campo aceptara nulos.