Instituto Politécnico Nacional

Escuela Superior de Cómputo

Data Mining

Profesor: Zagal Flores Roberto Eswart

Autor: Velasco Huerta Angel Eduardo

Práctica no. -1

03/09/2021

Ambiente de desarrollo:

Para poder realizar esta práctica, es necesario instalar un servidor de bases de datos, así como un cliente, en este caso, por preferencia personal (uso previo y comodidad) he instalado MariaDB, que es un gestor de base de datos que remplaza a MYSQL, pero que es prácticamente idéntico, pues fue diseñado por su mismo creador.

La instalación es muy simple, seleccionas las claves de acceso para el root, el puerto donde va a estar activo ese servidor, y pregunta si se quiere instalar HeidiSQL, que es el equivalente al MySQL Workbench, es decir, una interfaz gráfica para el manejo de bases de datos. En este caso, si se va a utilizar HeidiSQL, pues para importar archivos CSV puede resultar un poco mas sencillo realizarlo en la interfaz gráfica, pero los querys se correrán desde la terminal de MariaDB.

HeidiSQL:

Es una interfaz grafica bastante intuitiva, donde en un costado izquierdo se almacenan todas las bases de datos existentes en nuestro servidor, es sencillo crear tablas dentro de las bases, así como los campos, pues todo es cuestión de dar clic en las opciones que el programa va arrojando.

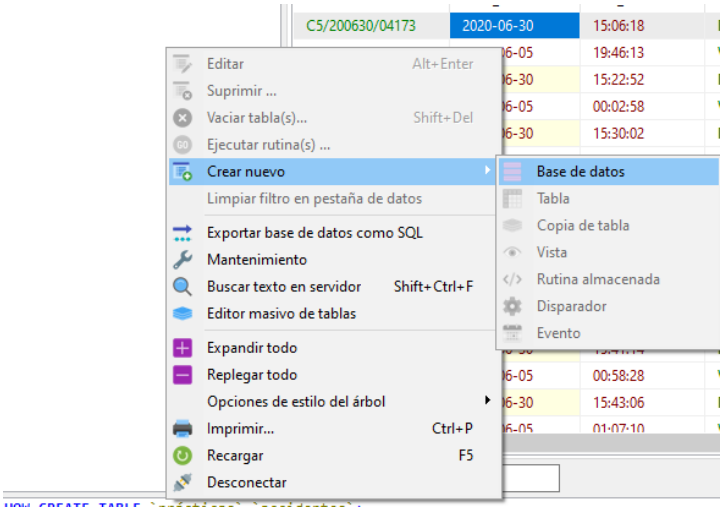
MariaDB:

Como tal, es una terminal de Windows, si bien no es tan intuitivo el manejar bases de datos, es rápido realizar alguna consulta por comandos, aunque la presentación de los datos puede no ser tan vistosa como en HeidiSQL.

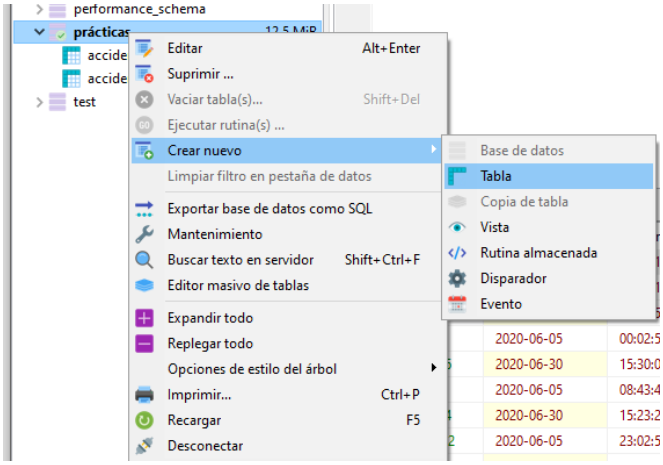
El objetivo principal de la práctica es comenzar a detectar y analizar los datos presentes en una base de datos, en el curso de minería de datos, trabajaremos mucho con la representación de estos, por eso mismo, debemos comprender muy bien la estructura de los datos, tales como los formatos, los problemas que pudieran existir,etc.

Importación de CSV y creación de base.

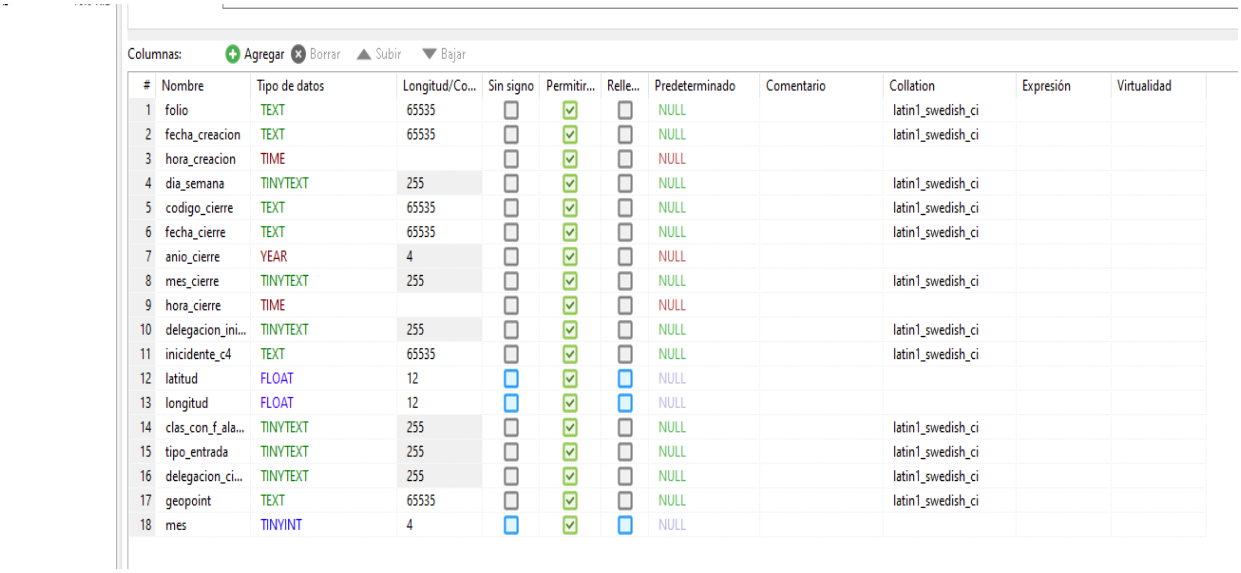
Lo primero a realizar es crear una base de datos donde se trabajará, en este caso, se llamará prácticas, para ello, desde HeidiSQL, damos clic derecho en la parte izq. y creamos la base.



Ahora entramos en la base dando clic en ella, y repetimos el proceso anterior, pero esta vez, crearemos una tabla.



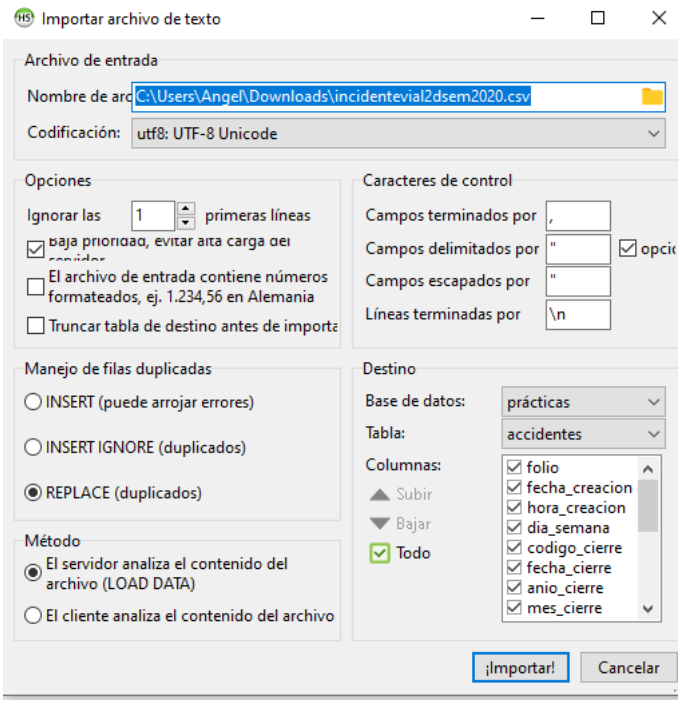
Nos saldrá una pestaña, donde crearemos los campos de la tabla, en MariaDB, es necesario crear los campos de la tabla antes de importar un CSV, los campos los podemos ver abriendo el documento en Excel:



Columnas: Agregar Borrar Subir Bajar

#	Nombre	Tipo de datos	Longitud/Co...	Sin signo	Permitir...	Relle...	Predeterminado	Comentario	Collation	Expresión	Virtualidad
1	folio	TEXT	65535	☐	☑	☐	NULL		latin1_swedish_ci		
2	fecha_creacion	TEXT	65535	☐	☑	☐	NULL		latin1_swedish_ci		
3	hora_creacion	TIME		☐	☑	☐	NULL				
4	dia_semana	TINYTEXT	255	☐	☑	☐	NULL		latin1_swedish_ci		
5	codigo_cierre	TEXT	65535	☐	☑	☐	NULL		latin1_swedish_ci		
6	fecha_cierre	TEXT	65535	☐	☑	☐	NULL		latin1_swedish_ci		
7	anio_cierre	YEAR	4	☐	☑	☐	NULL				
8	mes_cierre	TINYTEXT	255	☐	☑	☐	NULL		latin1_swedish_ci		
9	hora_cierre	TIME		☐	☑	☐	NULL				
10	delegacion_ini...	TINYTEXT	255	☐	☑	☐	NULL		latin1_swedish_ci		
11	inicidente_c4	TEXT	65535	☐	☑	☐	NULL		latin1_swedish_ci		
12	latitud	FLOAT	12	☐	☑	☐	NULL				
13	longitud	FLOAT	12	☐	☑	☐	NULL				
14	clas_con_f_ala...	TINYTEXT	255	☐	☑	☐	NULL		latin1_swedish_ci		
15	tipo_entrada	TINYTEXT	255	☐	☑	☐	NULL		latin1_swedish_ci		
16	delegacion_ci...	TINYTEXT	255	☐	☑	☐	NULL		latin1_swedish_ci		
17	geopoint	TEXT	65535	☐	☑	☐	NULL		latin1_swedish_ci		
18	mes	TINYINT	4	☐	☑	☐	NULL				

Guardamos, y una vez creada la tabla, lo siguiente fue importar el CSV con la opción especifica para eso en HeidiSQL:



Una vez finalizado, después de realizar algunas modificaciones (se hablará de ello en las conclusiones) tenemos todos los registros, que podemos ver con el query:

Select * from (Nombre de tabla)

Desarrollo:

Numero de registros: Par ver el numero de registros usaremos: `SELECT COUNT(*) FROM accidentes_viales;`

```
Your MariaDB connection id is 9
Server version: 10.5.8-MariaDB mariadb.org binary distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [(none)]> use prácticas;
Database changed
MariaDB [prácticas]> SELECT COUNT(*) FROM accidentes_viales;
+----------+
| COUNT(*) |
+----------+
|    33071 |
+----------+
1 row in set (0.107 sec)

MariaDB [prácticas]>
```

33071 registros en total.

¿Cual es el rango de los campos relacionados (valor minimo y maximo) con?:
Usaremos el query: `SELECT MAX (nombre del campo) from (nomnre de la tabla);` y `SELECT MIN (nombre del campo) from (nomnre de la tabla);` para así, determinar el rango.

Fechas:

```
MariaDB [prácticas]> SELECT MAX(fecha_creacion), MAX(fecha_cierre) from final;
+---------------------+--------------------+
| MAX(fecha_creacion) | MAX(fecha_cierre) |
+---------------------+--------------------+
| 2020-11-26          | 2020-11-26         |
+---------------------+--------------------+
1 row in set (0.038 sec)

MariaDB [prácticas]> SELECT MIN(fecha_creacion), MIN(fecha_cierre) from final;
+---------------------+--------------------+
| MIN(fecha_creacion) | MIN(fecha_cierre) |
+---------------------+--------------------+
| 2020-02-08          | 2020-06-01         |
+---------------------+--------------------+
1 row in set (0.038 sec)

MariaDB [prácticas]>
```

PD. Se cambió la tabla, pues se tuvo que corregir el formato de algunos datos.

Latitud y longitud:

```
MariaDB [prácticas]> SELECT MAX(latitud),MIN(latitud) from final;
+--------------+--------------+
| MAX(latitud) | MIN(latitud) |
+--------------+--------------+
|      19.5767 |      19.0954 |
+--------------+--------------+
1 row in set (0.017 sec)

MariaDB [prácticas]> SELECT MAX(longitud),MIN(longitud) from final;
+---------------+---------------+
| MAX(longitud) | MIN(longitud) |
+---------------+---------------+
|      -98.9476 |      -99.3484 |
+---------------+---------------+
1 row in set (0.017 sec)

MariaDB [prácticas]>
```

Año_cierre y hora_cierre:

```
MariaDB [prácticas]> SELECT MAX(hora_cierre),MIN(hora_cierre) from final;
+------------------+------------------+
| MAX(hora_cierre) | MIN(hora_cierre) |
+------------------+------------------+
| 23:59:59         | 00:00:00         |
+------------------+------------------+
1 row in set (0.019 sec)

MariaDB [prácticas]> SELECT MAX(anio_cierre),MIN(anio_cierre) from final;
+------------------+------------------+
| MAX(anio_cierre) | MIN(anio_cierre) |
+------------------+------------------+
|             2020 |             2020 |
+------------------+------------------+
1 row in set (0.029 sec)

MariaDB [prácticas]>
```

DOMINIO: Usaremos el query Select DISTINCT (valor) AS DOMINIO FROM (tabla);

Incidente_c4:

```
ERROR 1054 (42S22): Unknown column 'inicidente_c4' in 'field list'
MariaDB [prácticas]> SELECT DISTINCT(inicidente_c4) AS DOMINIO FROM final;
+-----------------------------------------------+
| DOMINIO                                       |
+-----------------------------------------------+
| accidente-choque con lesionados               |
| accidente-motociclista                        |
| accidente-choque sin lesionados               |
| lesionado-atropellado                         |
| accidente-persona atrapada / desbarrancada    |
| accidente-ciclista                            |
| sismo-choque con lesionados                   |
| accidente-volcadura                           |
| accidente-vehículo atrapado-varado            |
| detención ciudadana-atropellado               |
| cadáver-accidente automovilístico             |
| accidente-vehiculo desbarrancado              |
| cadáver-atropellado                           |
| accidente-choque con prensados                |
| accidente-otros                               |
| detención ciudadana-accidente automovilístico |
| sismo-persona atropellada                     |
| mi ciudad-calle-incidente de tránsito         |
| mi ciudad-taxi-incidente de tránsito          |
+-----------------------------------------------+
19 rows in set (0.357 sec)

MariaDB [prácticas]>
```

Tipo_entrada:

```
MariaDB [prácticas]> SELECT DISTINCT(Tipo_entrada) AS DOMINIO FROM final;
+------------------+
| DOMINIO          |
+------------------+
| LLAMADA DEL 911  |
| BOTÓN DE AUXILIO |
| REDES            |
| RADIO            |
| LLAMADA APP911   |
| CÁMARA           |
| APLICATIVOS      |
+------------------+
7 rows in set (0.268 sec)

MariaDB [prácticas]>
```

Clas_con_f_alarma:

```
MariaDB [prácticas]> SELECT DISTINCT(clas_con_f_alarma) AS DOMINIO FROM final;
+--------------------+
| DOMINIO            |
+--------------------+
| URGENCIAS MEDICAS  |
| EMERGENCIA         |
| DELITO             |
| FALSA ALARMA       |
+--------------------+
4 rows in set (0.376 sec)

MariaDB [prácticas]>
```

Delegación:

```
MariaDB [prácticas]> SELECT DISTINCT(delegacion_cierre) AS DOMINIO FROM final;
+----------------------+
| DOMINIO              |
+----------------------+
| GUSTAVO A. MADERO    |
| ALVARO OBREGON       |
| XOCHIMILCO           |
| TLALPAN              |
| IZTAPALAPA           |
| AZCAPOTZALCO         |
| CUAUHTEMOC           |
| MIGUEL HIDALGO       |
| VENUSTIANO CARRANZA  |
| BENITO JUAREZ        |
| COYOACAN             |
| IZTACALCO            |
| MAGDALENA CONTRERAS  |
| TLAHUAC              |
| MILPA ALTA           |
| CUAJIMALPA           |
| NULL                 |
+----------------------+
17 rows in set (0.222 sec)

MariaDB [prácticas]>
```

Finalmente Contar la cantidad de NULL o NULOS encontrados en las 4 columnas anteriores del punto 5.
Para ello, usaremos el siguiente query:
Select * FROM (tabla) WHERE (valor) is NULL;
Originalmente, el query, no me arrojaba nada, sin embargo, me di cuenta de que por alguna razón, el manejador de datos que use, ahora aparecía como valores vacíos en vez de NULL, motivo por el cual use el query que se muestra a continuación.
Solo hay nulos en las delegaciones:

```
MariaDB [incidentes_viales]> select folio,delegacion_ini,delegacion_cierre from incidentes where delegacion_ini = "";
+-----------------+----------------+-------------------+
| folio           | delegacion_ini | delegacion_cierre |
+-----------------+----------------+-------------------+
| AO/200820/01262 |                |                   |
| C5/200626/01619 |                |                   |
| AO/200820/01262 |                |                   |
| C5/200626/01619 |                |                   |
+-----------------+----------------+-------------------+
4 rows in set (0.111 sec)

MariaDB [incidentes_viales]>
```

Conclusiones:
EL desarrollo de esa práctica, aunque sencilla, implicó muchos detalles importantes del manejo de datos, lo primero que note, fue que al momento de querer importar el CSV, tuve muchísimos problemas pues el formato de los datos que contenía el CSV, no eran del todo compatibles con los formatos prestablecidos de MariaDB como lo son las fechas y las horas, por lo que antes de hacer la importación, modifique las fechas para tener formato de año/mes/día, y en las fechas de cierre, venia una hora que siempre era igual 00:00.0, que complicaba la correcta inserción (Por lo menos en el gestor que utilice).
Todas las modificaciones, fueron hechas en EXCEL, pues así podía cambiar el formato de las 33072 filas.

Posteriormente, tras importar los datos satisfactoriamente, aun note algunos problemas en algunos campos, que el gestor de bases los llenaba como NULL, o un dato que no era, afortunadamente, estos no afectaron tanto en las consultas. Al ser una base de datos de una cantidad de datos algo grande, era imposible verificar que todos los datos estuvieran llenos, por eso, fue importante, permitir NULLS en la importación.

Note que, por alguna razón, el manejador de bases no estaba mostrando NULL sino “ ”, aunque esto no complico la práctica, es importante tomarlo en cuenta, en la próxima práctica se probara con SQL Server, para probar si se tiene el mismo detalle.

Finalmente, la parte de las consultas no resultó tan complicada, la realicé en la terminal, pues siento que aprendo un poco más asiéndolo así, bastó con buscar algunas de las funciones de SQL en https://www.w3schools.com/sql/ y aplicarlas a las tablas y campos que tenia mi base, en conclusión, hay que ser muy cuidadoso con el formato de los datos al momento de crear una base de datos, pues estas pueden complicarse si no se manejó correctamente la creación de los campos o la inserción de la información, y en un posterior análisis de datos, o transformación, si no se tiene en cuenta el formato o los errores antes de limpiar nuestra base, podríamos terminar con datos o conclusiones erróneas.