



Alumno:	García González Aarón Antonio
Grupo:	3CV19
Unidad de Aprendizaje:	Data Mining
Profesor:	Zagal Flores Roberto Eswart
Parcial #1 Practica #1	Exportación de datos y cobertura espacio – tiempo
Fecha:	Lunes 01 de marzo de 2021

Índice

Introducción	2
Base de datos y SQL.....	2
Sistema de administración de bases de datos.....	2
Minería de datos	2
Diferencias entre Data Mining y Big Data	3
Desarrollo.....	4
1. Descargue el dataset de incidentes viales, que corresponde al último semestre del 2020.....	4
2. Exporte el archivo CSV en el manejador de base de datos seleccionado (nota: estudie los videos de instalación de los posibles manejadores).....	4
3. Indique el número de registros del dataset en el manejador.	4
4. ¿Cuál es el rango de los campos relacionados (valor mínimo y máximo) con?	4
⇒ “Fecha” (todos los relacionados)	4
⇒ Codigo_cierre	4
⇒ Año_cierre,mes_cierre y hora_cierre (todos los relacionados al cierre”).....	5
5. ¿Cuales son los valores que definen el dominio de la columna de los siguientes campos?	6
⇒ Incidente_c4.....	6
⇒ Tipo_entrada.....	6
⇒ Clas_con_f_alarma.....	7
⇒ Delegación.....	7
Conclusiones.....	8
Referencias.....	8

Introducción

Base de datos y SQL

Una base de datos es un conjunto de datos que se encuentran organizados, estructurados y almacenados para su acceso y manipulación. SQL es un lenguaje estándar para almacenar, manipular y recuperar datos en bases de datos.

Sistema de administración de bases de datos

Entre las principales características de los sistemas de base de datos podemos mencionar:

- Independencia lógica y física de los datos.
- Redundancia mínima.
- Acceso concurrente por parte de múltiples usuarios.
- Integridad de los datos.
- Consultas complejas optimizadas.
- Seguridad de acceso y auditoría.
- Respaldo y recuperación.
- Acceso a través de lenguajes de programación estándar.

Sistema de administración de bases de datos: consiste en un conjunto de programas utilizados para definir, administrar y procesar una base de datos y sus aplicaciones. A los sistemas de administración de bases de datos también se les llama Sistemas de Gestión de Bases de Datos (SGBD). Un sistema de administración de bases de datos es una herramienta de propósito general que permite crear bases de datos de cualquier tamaño y complejidad y con propósitos específicos distintos.

El objetivo principal de un sistema de administración de bases de datos es proporcionar una forma de almacenar y recuperar la información de una base de datos de manera que sea tanto práctica como eficiente. Los SGBD se diseñan para gestionar grandes cantidades de información. La gestión de los datos implica tanto la definición de estructuras para almacenar la información como la provisión de mecanismos para la manipulación de la información. Además, los sistemas de bases de datos deben proporcionar la fiabilidad de la información almacenada, a pesar de las caídas del sistema o los intentos de acceso sin autorización. Si los datos van a ser compartidos entre varios usuarios, el sistema debe evitar posibles datos contradictorios.

La realización de base de datos se ha vuelto una acción fundamental para las empresas, ya que les permiten crear estrategias para conseguir nuevos clientes o fidelizar a los habituales. Pero a consecuencia de la generación masiva de datos, nos encontramos frente a un problema, la infoxicación, disponemos de tanta información, que a veces es imposible organizarla con efectividad. Por ello, la clave está en descubrir patrones o algoritmos para sacarle el máximo partido, y aquí es donde entra en juego el Data Mining o minería de datos.

Minería de datos

La minería de datos es la práctica de buscar automáticamente grandes almacenes de datos para descubrir patrones y tendencias que van más allá del simple análisis. La minería de datos utiliza sofisticados algoritmos matemáticos para segmentar los datos y evaluar la probabilidad de eventos futuros. La minería de datos también se conoce como descubrimiento de conocimiento en datos (KDD).

Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

- Pronóstico: cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
- Riesgo y probabilidad: elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.
- Recomendaciones: determinación de los productos que se pueden vender juntos y generación de recomendaciones.
- Búsqueda de secuencias: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.
- Agrupación: distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

Para asegurar resultados de minería de datos significativos, debe comprender sus datos. Los algoritmos de minería de datos suelen ser sensibles a características específicas de los datos, valores atípicos (valores de datos que son muy diferentes de los valores típicos en su base de datos), columnas irrelevantes, columnas que varían juntas (como edad y fecha de nacimiento), codificación de datos y datos que elige incluir o excluir. Minería de datos, puede automáticamente realizar gran parte de la preparación de datos requerida por el algoritmo. Pero parte de la preparación de datos suele ser específica del dominio o del problema de minería de datos. En cualquier caso, debe comprender los datos que se utilizaron para construir el modelo a fin de interpretar correctamente los resultados cuando se aplica el modelo.

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los seis pasos básicos siguientes:

- Definir el problema
- Preparar los datos
- Explorar los datos
- Crear modelos
- Explorar y validar modelos
- Implementar y actualizar los modelos

Diferencias entre Data Mining y Big Data

El Big Data es una tecnología que tiene la capacidad de capturar, gestionar y procesar de forma veraz todo tipo de datos, utilizando herramientas o softwares que identifican patrones comunes. Estos patrones podrían ser características específicas de los consumidores, generación de parámetros, métricas, entre muchos otros. Y, tienen la capacidad de cambiar la manera de hacer negocios, ya que permiten aumentar la rentabilidad y productividad de las compañías.

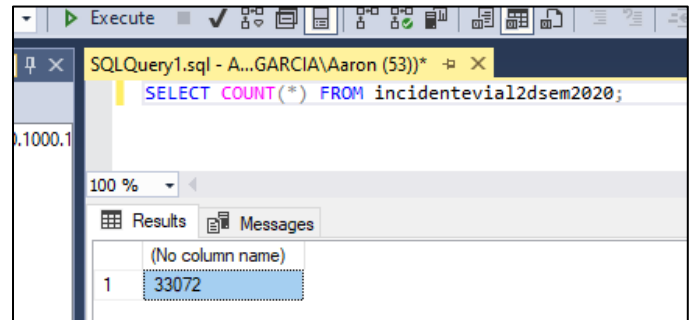
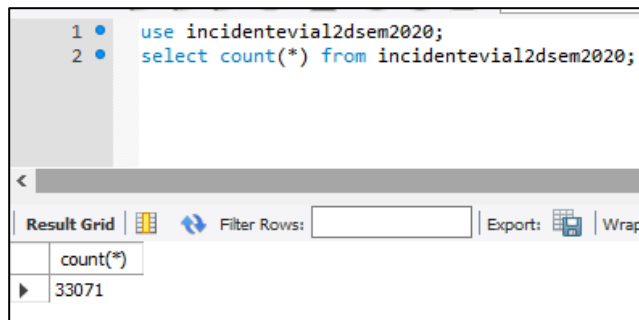
A diferencia del Big Data, tal y como se ha comentado anteriormente, cuando hablamos de Data Mining nos referimos al análisis de los grandes datos o Big Data para buscar y obtener una información concreta, y así, poder ofrecer resultados que sirvan como solución para optimizar las actividades de una empresa.

En resumen, Big Data y Minería de datos podrían ser definidos como el “activo” y el “manejo”, respectivamente.

Para esta primera practica introductoria únicamente nos metemos a la parte de exploración y preparación de datos, en donde importaremos datos en formato CSV a la base de datos en el motor de mysql o sql server, nos centraremos en especial en los rangos de fechas y horas, los posibles valores clave que podremos encontrar por cada atributo en algunos de estos.

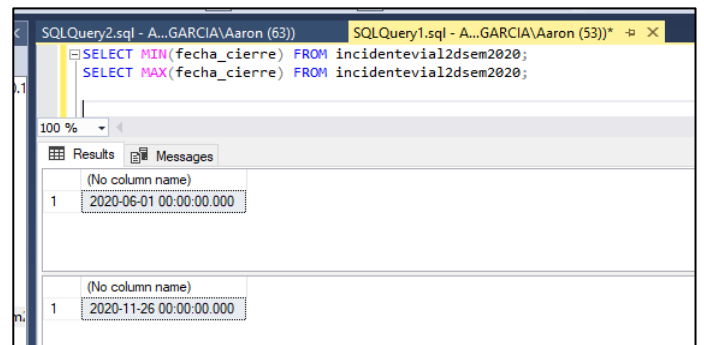
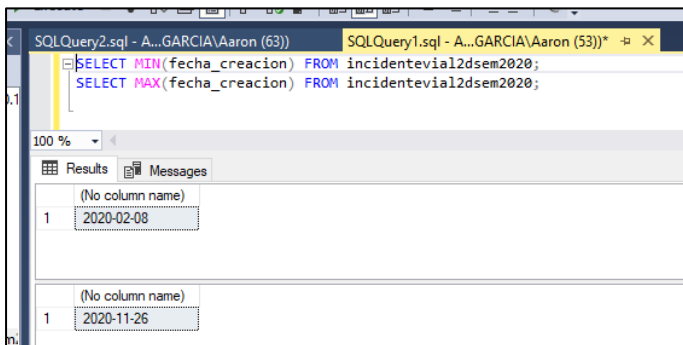
Desarrollo

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.
 - ⇒ 33072 en Excel (para comprobación)
 - ⇒ 33,071 en mysql
 - ⇒ 33/072 en SQL server



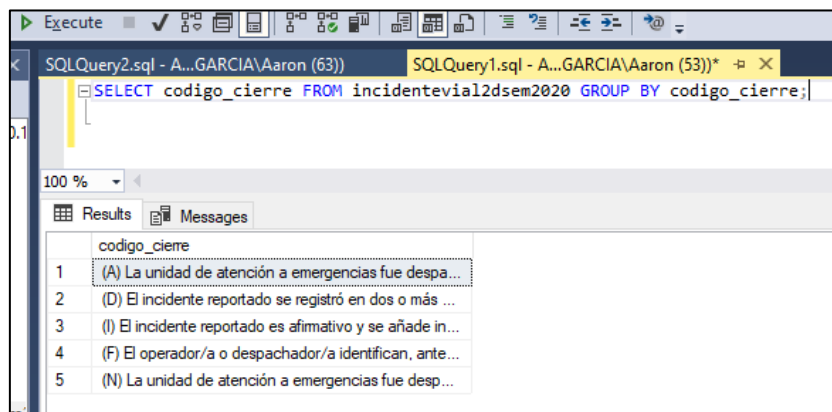
4. ¿Cuál es el rango de los campos relacionados (valor mínimo y máximo) con?

- ⇒ “Fecha” (todos los relacionados)
 - Rango de fecha de creación: 2020-02-08 al 2020-11-26
 - Rango de fecha de cierre: 2020-06-01 al 2020-11-26



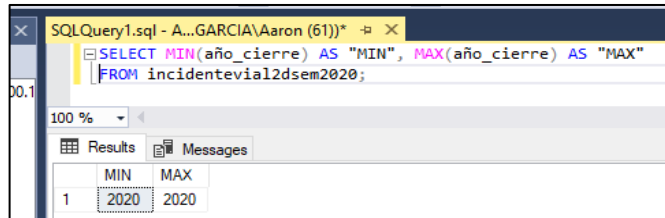
- ⇒ Codigo_cierre

Aquí quiero suponer que se refiere a la gama de valores que toma este atributo en la tabla de hechos.



⇒ Año_cierre,mes_cierre y hora_cierre (todos los relacionados al cierre")

Año cierre

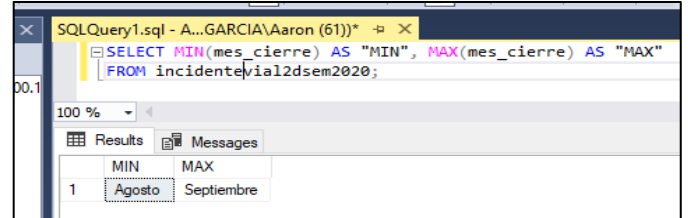


SQLQuery1.sql - A...GARCIA\Aaron (61))*

```
SELECT MIN(año_cierre) AS "MIN", MAX(año_cierre) AS "MAX"
FROM incidente12dsem2020;
```

MIN	MAX
1	2020

Mes cierre

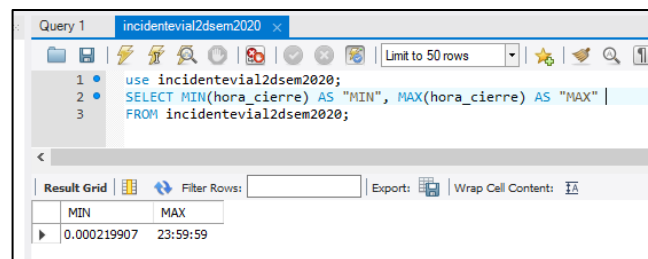


SQLQuery1.sql - A...GARCIA\Aaron (61))*

```
SELECT MIN(mes_cierre) AS "MIN", MAX(mes_cierre) AS "MAX"
FROM incidente12dsem2020;
```

MIN	MAX
1	Agosto Septiembre

Año cierre



Query 1 incidente12dsem2020

```
use incidente12dsem2020;
SELECT MIN(hora_cierre) AS "MIN", MAX(hora_cierre) AS "MAX"
FROM incidente12dsem2020;
```

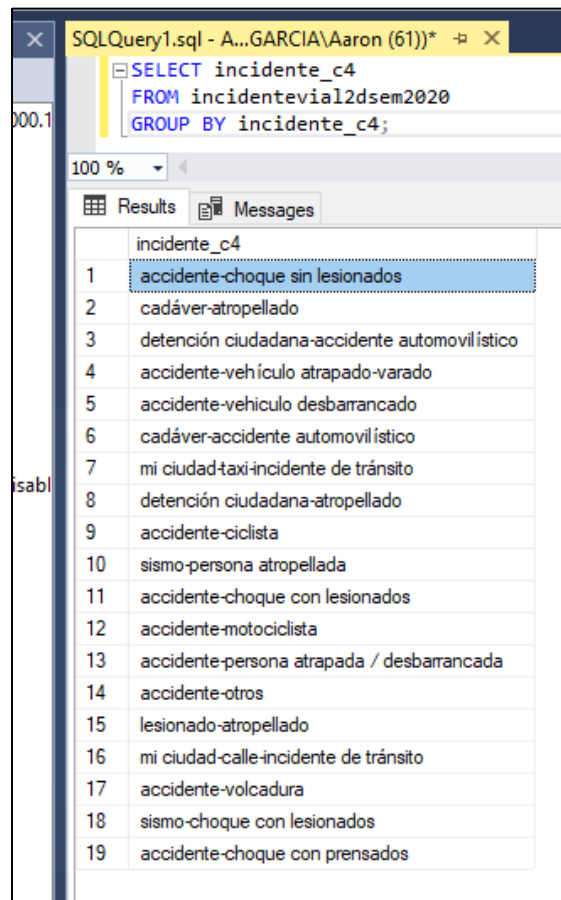
MIN	MAX
0.000219907	23:59:59

Reporte los valores obtenidos, copie los valores y anexe una captura de imagen, también agregue las consultas SQL usadas. Puede usar <https://www.w3schools.com/sql/>

5. ¿Cuales son los valores que definen el dominio de la columna de los siguientes campos?

⇒ Incidente_c4

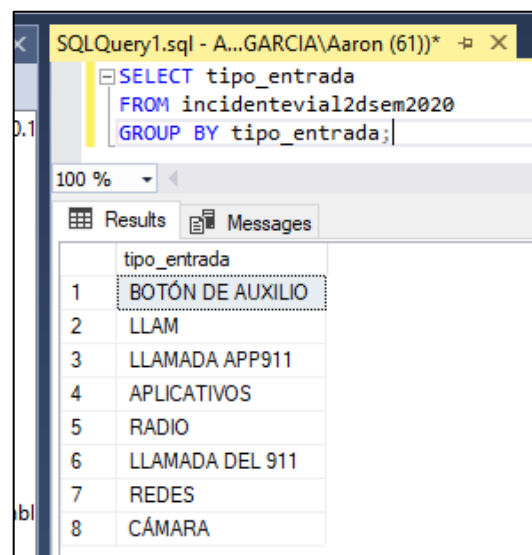
En total hay 19 tipos de incidente C4, los cuales son:



The screenshot shows a SQL query window titled 'SQLQuery1.sql - A...GARCIA\Aaron (61))' with the following query: `SELECT incidente_c4 FROM incidente_vial2dsem2020 GROUP BY incidente_c4;`. The 'Results' tab is active, displaying a table with 19 rows. The first row is highlighted. The table has two columns: 'incidente_c4' and an unlabeled column.

	incidente_c4	
1	accidente-choque sin lesionados	
2	cadáver-atropellado	
3	detención ciudadana-accidente automovilístico	
4	accidente-vehículo atrapado-varado	
5	accidente-vehículo desbarrancado	
6	cadáver-accidente automovilístico	
7	mi ciudad-taxi-incidente de tránsito	
8	detención ciudadana-atropellado	
9	accidente-ciclista	
10	sismo-persona atropellada	
11	accidente-choque con lesionados	
12	accidente-motociclista	
13	accidente-persona atrapada / desbarrancada	
14	accidente-otros	
15	lesionado-atropellado	
16	mi ciudad-calle-incidente de tránsito	
17	accidente-volcadura	
18	sismo-choque con lesionados	
19	accidente-choque con prensados	

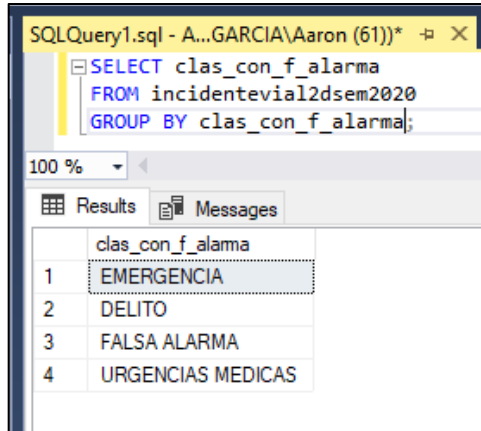
⇒ Tipo_entrada



The screenshot shows a SQL query window titled 'SQLQuery1.sql - A...GARCIA\Aaron (61))' with the following query: `SELECT tipo_entrada FROM incidente_vial2dsem2020 GROUP BY tipo_entrada;`. The 'Results' tab is active, displaying a table with 8 rows. The first row is highlighted. The table has two columns: 'tipo_entrada' and an unlabeled column.

	tipo_entrada	
1	BOTÓN DE AUXILIO	
2	LLAM	
3	LLAMADA APP911	
4	APLICATIVOS	
5	RADIO	
6	LLAMADA DEL 911	
7	REDES	
8	CÁMARA	

⇒ Clas_con_f_alarma



SQLQuery1.sql - A...GARCIA\Aaron (61))*

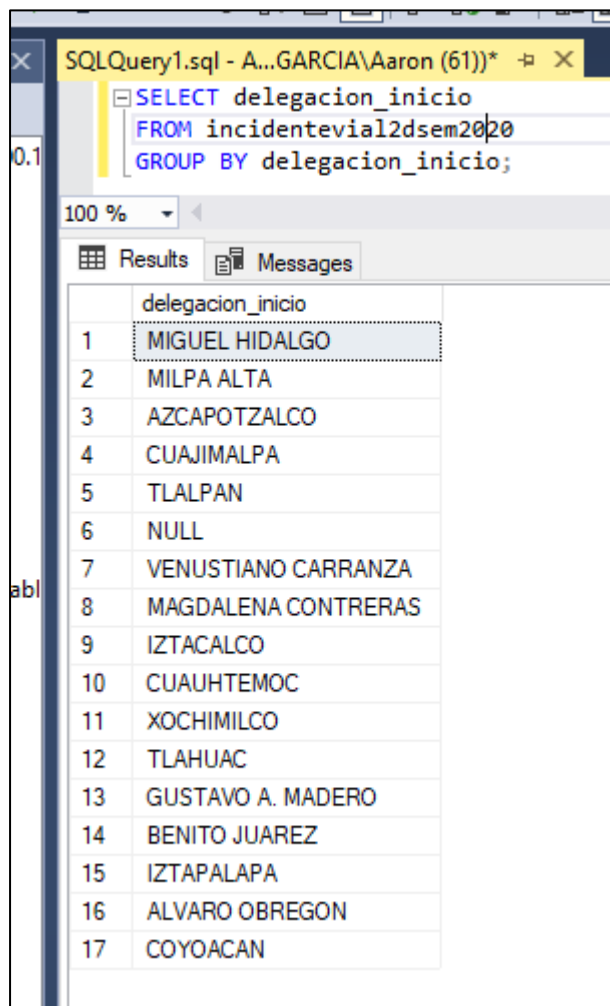
```
SELECT clas_con_f_alarma
FROM incidente2dsem2020
GROUP BY clas_con_f_alarma;
```

Results

	clas_con_f_alarma
1	EMERGENCIA
2	DELITO
3	FALSA ALARMA
4	URGENCIAS MEDICAS

⇒ Delegación

Podemos encontrar dos atributos asociados con la delegación, delegacion_inicio y delegación_cierre:

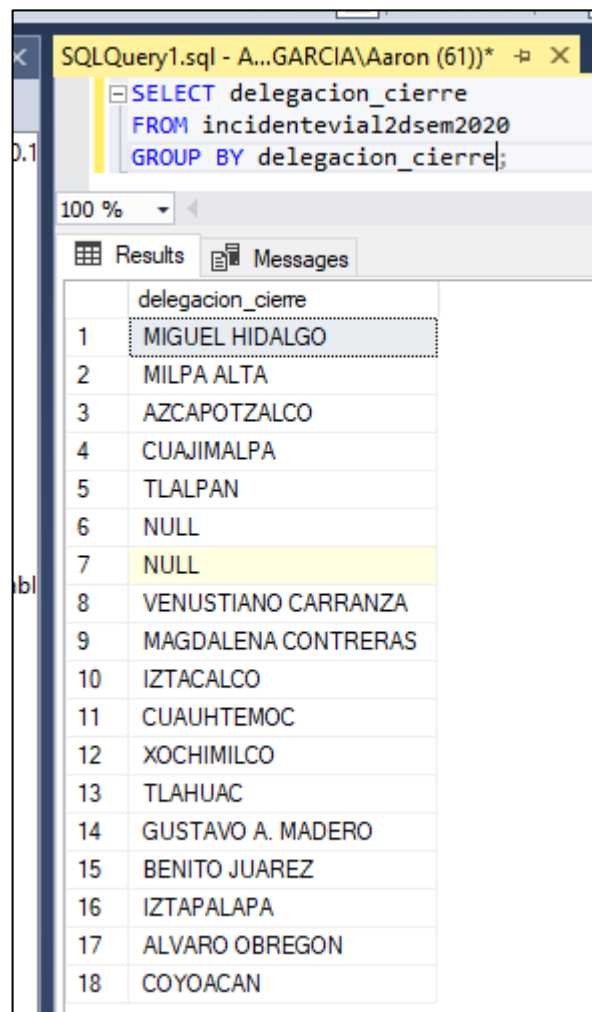


SQLQuery1.sql - A...GARCIA\Aaron (61))*

```
SELECT delegacion_inicio
FROM incidente2dsem2020
GROUP BY delegacion_inicio;
```

Results

	delegacion_inicio
1	MIGUEL HIDALGO
2	MILPA ALTA
3	AZCAPOTZALCO
4	CUAJIMALPA
5	TLALPAN
6	NULL
7	VENUSTIANO CARRANZA
8	MAGDALENA CONTRERAS
9	IZTACALCO
10	CUAUHTEMOC
11	XOCHIMILCO
12	TLAHUAC
13	GUSTAVO A. MADERO
14	BENITO JUAREZ
15	IZTAPALAPA
16	ALVARO OBREGON
17	COYOACAN



SQLQuery1.sql - A...GARCIA\Aaron (61))*

```
SELECT delegacion_cierre
FROM incidente2dsem2020
GROUP BY delegacion_cierre;
```

Results

	delegacion_cierre
1	MIGUEL HIDALGO
2	MILPA ALTA
3	AZCAPOTZALCO
4	CUAJIMALPA
5	TLALPAN
6	NULL
7	NULL
8	VENUSTIANO CARRANZA
9	MAGDALENA CONTRERAS
10	IZTACALCO
11	CUAUHTEMOC
12	XOCHIMILCO
13	TLAHUAC
14	GUSTAVO A. MADERO
15	BENITO JUAREZ
16	IZTAPALAPA
17	ALVARO OBREGON
18	COYOACAN

Reporte los valores obtenidos, copie los valores y anexe una captura de imagen, también agregue las consultas SQL usadas.

Conclusiones

Esta es la primera vez que hago la inserción de datos desde un archivo con extensión CSV, en mis cursos previos lo realice desde scripts en consola o scripts que se ejecutan al iniciar una aplicación, es decir datos precargados. El profesor nos comentó que al cargar datos por ejemplo de este tipo, podrían darse ciertas anomalías, que dependerían en su mayoría del sistema gestor de base de datos a utilizar, para esta primera práctica utilice workbench y Microsoft SQL Server Management Studio para realizar la importación, efectivamente en workbench fue más “sencillo” ya que no es tan específico con los tipos de datos, olvide revisar el tipo de dato de fecha de cierre, supuse que sería un “date” por su nombre, pero al terminar la importación de datos me salió una cadena hexadecimal en ese atributo para todos los registros, dado que en mi computadora con Windows se tardó aproximadamente 30 minutos en realizar la importación, decidí dejarlo así e intentar con Microsoft, aquí la cosa fue un poco distinta, aquí si me fijé en el tipo de dato que mencione anterior, pero aquí los campos de tipo “time” no me dejaba avanzar, probé varias opciones que trae el manejador para este tipo de dato, me dejó con timestamp pero únicamente permite tener un atributo de este tipo de dato por tabla, por lo que se lo asigne a la consulta que si ocupamos en la especificación de la práctica, el otro lo deje en varchar y si había que hacer alguna consulta sobre este campo, lo podría revisar en workbench ya que ahí no hubo problema con los atributos de tiempo.

Decidí probar primero con mi computadora Windows de 8 RAM, pero considero que fue demasiado lo que tardó, optare por trabajar en mi MAC de 8RAM que siento como si fuera el doble de RAM en Windows, no lo quería hacer debido a la limitante que tengo en MAC por el almacenamiento.

Referencias

W3C Schools. (s. f.). SQL Tutorial. [www.w3schools.com](https://www.w3schools.com/sql/). Recuperado 1 de marzo de 2021, de <https://www.w3schools.com/sql/>

Oracle. (s. f.). What Is Data Mining? Conceptos de minería de datos. Recuperado 1 de marzo de 2021, de https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON046