ZAGAL FLORES ROBERTO ESWART

DATA MINING

ALUMNO:
LÓPEZ MORALES MIGUEL ÁNGEL

3CV19

Práctica 6:

Diseño estructural del datawarehouse y desarrollo del ETL del proyecto semestral

CONTENIDO

1. OBJETIVO

Desarrollar la lattice de cubos de datos principales para construir el data warehouse el proyecto semestral.

2. INTRODUCCIÓN

Esta práctica se desarrollará con el lenguaje de programación Python, en la cual la principal librería que se utiliza es la librería Pandas.

Las principales características de esta librería son:

- Define nuevas estructuras de datos basadas en los arrays de la librería NumPy, pero con nuevas funcionalidades.
- Permite leer y escribir fácilmente ficheros en formato CSV, Excel y bases de datos SQL.
- Permite acceder a los datos mediante índices o nombres para filas y columnas.
- Ofrece métodos para reordenar, dividir y combinar conjuntos de datos.
- Permite trabajar con series temporales.
- Realiza todas estas operaciones de manera muy eficiente.

Esta librería se utilizará principalmente para la extracción de datos de los archivos xls.
MySQL Workbench es un software creado por la empresa Sun Microsystems, esta herramienta permite modelar diagramas de Entidad-Relación para bases de datos MySQL.

Con esta herramienta se puede elaborar una representación visual de las tablas, vistas, procedimientos almacenados y claves foráneas de la base de datos. Además, es capaz de sincronizar el modelo en desarrollo con la base de datos real. Se puede realizar una ingeniería directa e ingeniería inversa para exportare e importar el esquema de una base de datos ya existente el cual haya sido guardado o hecho copia de seguridad con MySQL Administrador.

MySQL Workbench puede generar también el guion necesario para crear la base de datos que se ha dibujado en el esquema; es compatible con los modelos de base de datos de DBDesigner 4 y soporta las novedades incorporadas en MySQL 5.x

La versión utilizada en esta práctica es la 8.0.12.

Tableau.

Tableau es una herramienta de visualización de datos potente utilizada en el área de la Inteligencia de negocios (más conocida como Business Intelligence). Simplifica los datos en bruto en un formato muy fácil de entender.

La esencia de Tableau es simple y a la vez muy relevante: ayudar a las personas y empresas a ver y comprender todos sus datos. Y esto lo consigue ofreciendo a los usuarios toda una selección de herramientas útiles e intuitivas de inteligencia de negocios.
A través de funciones simples como la de arrastrar y soltar, cualquier persona puede acceder y analizar de forma sencilla datos, e incluso, crear informes y compartir esta información con otros usuarios.

3. DESARROLLO

Procedimiento: Construya la arquitectura de datos para responder a preguntas de minería, del proyecto semestral seleccionado

1. Retome y explique algunas preguntas de minería que desea responder, bajo el siguiente formato:
    a. Título del proyecto

- Tasa de delincuencia en la delegación GAM

    b. Objetivo del proyecto

- Conocer la tasa de delincuencia de la alcaldía en los alrededores de ESCOM, nuestra escuela, es decir, la alcaldía Gustavo A. Madero.

    c. Entendimiento del negocio:

        i. Explique las preguntas de minería que responderá

- ¿Cuál es la alcaldía con más delitos cometidos?
  ¿Cuál es la alcaldía con la categoría más grave de delitos?
  ¿Cuál es la tasa de delincuencia de estas delegaciones?
  Entre otras.

        ii. Explique con base en el diccionario de datos el temas o temas de su fuente de datos

- Carpetas de investigación PGJ de la Ciudad de México

Esta base de datos contiene las Carpetas de investigación de delitos a nivel de calle de la Procuraduría General de Justicia de la Ciudad de México desde enero de 2016 hasta junio de 2019. Esta base de datos dejó de actualizarse a partir de julio de 2019 debido a que contiene la información de carpetas de investigación de 2018 previa a la reclasificación que realizó la PGJ-CDMX con supervisión de la Oficina de Naciones Unidas contra la Droga y el Delito (UNDOC) y la validación del Centro Nacional de Información (CNI) del Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP)

d. Entendimiento de los datos
   i. Explique la estructura original de la fuente de datos

- El dataset contiene varios campos de datos con diferentes funciones, los cuales se podrían dividir en grupos, los campos de “ao_hechos”, “mes_hechos” y “fecha_hechos” Son los campos en donde se nos muestra información de la fecha en la que sucedió algún delito.

  Los campos de “delito” y “categoría_delito” muestran datos acerca de los acontecimientos delictivos que sucedieron en algún lugar.

  Los campos “fiscalía”, “agencia” y “unidad_investigacion” muestran información de los departamentos encargados de la investigación y registro de los hechos.

  Los campos de “fecha_inicio”, “mes_inicio” y “ao_inicio” refieren a datos de la fecha en la que se abrió la carpeta de investigación o realización de la denuncia de algún delito.

  Los campos restantes muestran información del lugar donde ocurrieron los hechos delictivos.

Columns:

id	int(11)
ao_hechos	int(11)
mes_hechos	text
fecha_hechos	text
delito	text
categoria_delito	text
fiscalia	text
agencia	text
unidad_investigacion	text
colonia_hechos	text
alcaldia_hechos	text
fecha_inicio	text
mes_inicio	text
ao_inicio	int(11)
calle_hechos	text
calle_hechos2	text
longitud	double
latitud	double
geopoint	text

Figura 1. Dataset de Carpetas de investigación PGJ de la Ciudad de México

   ii. Explique su estrategia de Extracción de los datos

- Se extraerán los datos por medio de un programa realizado en Python con el cual realizaremos una lectura del csv obtenido de los conjuntos de datos abiertos de la Ciudad de México con la librería Pandas la cual nos ayudara, además de extraer los datos, a formar dataframes para una manipulación de los datos óptima.

2. Diseñe la estructura de su tabla de hechos y el modelado conceptual del data lake.
   a. Explique cómo organizará las dimensiones

- El dataset contiene varios campos de datos con diferentes funciones, los cuales se podrían dividir en grupos, los campos de “ao_hechos”, “mes_hechos” y “fecha_hechos” Son los campos en donde se nos muestra información de la fecha en la que sucedió algún delito.

  Los campos de “delito” y “categoría_delito” muestran datos acerca de los acontecimientos delictivos que sucedieron en algún lugar.

  Los campos “fiscalía”, “agencia” y “unidad_investigacion” muestran información de los departamentos encargados de la investigación y registro de los hechos.

  Los campos de “fecha_inicio”, “mes_inicio” y “ao_inicio” refieren a datos de la fecha en la que se abrió la carpeta de investigación o realización de la denuncia de algún delito.

  Los campos restantes muestran información del lugar donde ocurrieron los hechos delictivos.

  b. Diseñe una implementación SIMPLE (la solución) tomando como base el modelo multidimensional EN ESTRELLA, es decir puede plantear un diseño teórico sencillo. Puede ser separado en catálogos las dimensiones y agregando información adicional para hacer análisis.

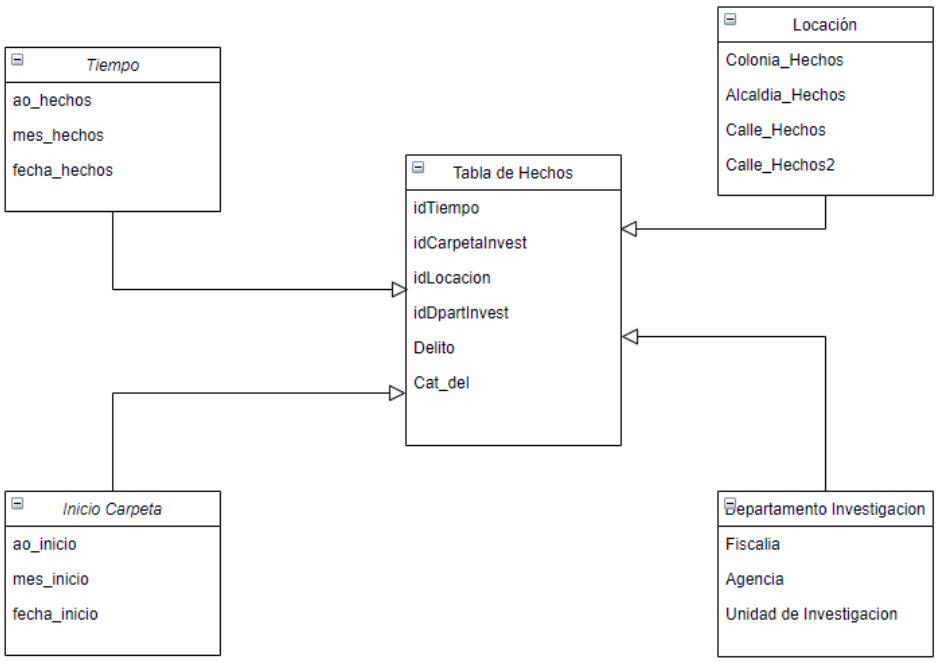

Figura 2. Implementación de modelo en estrella

3. Explique las fases del ETL o integración de datos de su proyecto.
   a. Estrategia de extracción de datos

- Como anteriormente se hizo en prácticas anteriores, se extraerán los datos de un archivo CSV por medio de lenguaje de programación Python y su librería Pandas, para poder extraer de una manera óptima los datos.

b. Estrategia de transformación

- Una vez cargados los datos del CSV se agruparán los datos conforme al diagrama propuesto en tablas para así mandarse a la base de datos por medio de dataframes realizados con la librería Pandas.

c. Estrategia de carga de datos a su modelo del data lake

- Dichos dataframes se mandarán directamente a MySQL interpretados como tablas y después de eso se mandarán sentencias SQL para la creación de llaves primarias y foráneas.

4. Implemente su proceso de ETL, y explique las tecnologías usadas para desarrollarlo.

- El código se implementó como se planeaba con la diferencia de que, como los id's de cada una de las tablas eran los mismos, se decidió agregar una letra a dicho id para diferenciarlos del resto (Ejemplo Fiscalia letra "F", ejemplo de id: "F14").

5. Documente con capturas de pantalla que los modelos del data lake mostrando que tenga datos cargados, explique si se cargó la totalidad de datos originales incluyendo nulos**. Documente con captura de pantalla el modelo relacional desarrollado en la plataforma de base de datos seleccionada.**

- Se cargó el CSV al que se le aplico un proceso de limpieza de valores nulos generado en la práctica 3, en seguida se muestran capturas de la base de datos ya cargada después del proceso ETL.

idHechos	idInicio	idFiscalia	idLocacion	delito	categoria_delito
H0	IC0	F0	L0	LESIONES CULPOSAS	DELITO DE BAJO IMPACTO
H2	IC2	F2	L2	ROBO A TRANSEUNTE EN VIA PUBLICA CON VI...	ROBO A TRANSEUNTE EN VÃA PÃšBLICA CON Y...
H3	IC3	F3	L3	FRAUDE	DELITO DE BAJO IMPACTO
H4	IC4	F4	L4	ROBO A REPARTIDOR CON VIOLENCIA	ROBO A REPARTIDOR CON Y SIN VIOLENCIA
H5	IC5	F5	L5	ROBO A TRANSEUNTE DE CELULAR CON VIOLE...	DELITO DE BAJO IMPACTO
H7	IC7	F7	L7	ROBO DE DINERO	DELITO DE BAJO IMPACTO
H8	IC8	F8	L8	LESIONES INTENCIONALES	DELITO DE BAJO IMPACTO
H10	IC10	F10	L10	ABUSO DE AUTORIDAD	DELITO DE BAJO IMPACTO
H11	IC11	F11	L11	DAÃ'O EN PROPIEDAD AJENA CULPOSA POR T...	DELITO DE BAJO IMPACTO
H12	IC12	F12	L12	ROBO DE VEHICULO DE SERVICIO PÃšBLICO SI...	ROBO DE VEHÃCULO CON Y SIN VIOLENCIA
H15	IC15	F15	L15	ROBO DE OBJETOS	DELITO DE BAJO IMPACTO
H17	IC17	F17	L17	ROBO DE OBJETOS	DELITO DE BAJO IMPACTO
H18	IC18	F18	L18	VIOLENCIA FAMILIAR	DELITO DE BAJO IMPACTO
H19	IC19	F19	L19	NARCOMENUDEO POSESION SIMPLE	DELITO DE BAJO IMPACTO
H21	IC21	F21	L21	ROBO DE ACCESORIOS DE AUTO	DELITO DE BAJO IMPACTO

Figura 3. Tabla de hechos.

id	fiscalia	agencia	unidad_investigacion
F0	INVESTIGACIÃ"N EN GUSTAVO A. MADERO	GAM-6	UI-1CD
F10	INVESTIGACIÃ"N EN AGENCIAS DE ATENCIÃ"N ...	STCMOB	UI-1CD
F100	INVESTIGACIÃ"N EN MIGUEL HIDALGO	MH-5	UI-3CD
F100001	INVESTIGACIÃ"N EN BENITO JUÃREZ	BJ-3	UI-3SD
F100003	INVESTIGACIÃ"N EN AGENCIAS DE ATENCIÃ"N ...	STCMZV	UI-2CD
F100004	INVESTIGACIÃ"N EN IZTAPALAPA	IZP-6	UI-1SD
F100005	INVESTIGACIÃ"N EN XOCHIMILCO	XO-2	UI-1SD
F100006	INVESTIGACIÃ"N EN BENITO JUÃREZ	BJ-3	UI-2SD
F100008	INVESTIGACIÃ"N EN TLALPAN	TLP-1	UI-2CD
F100009	INVESTIGACIÃ"N DE LOS DELITOS COMETIDOS...	B	UI-2CD
F10001	INVESTIGACIÃ"N EN GUSTAVO A. MADERO	GAM-5	UI-2CD
F100014	INVESTIGACIÃ"N EN MIGUEL HIDALGO	MH-1	UI-1SD
F100015	INVESTIGACIÃ"N DE LOS DELITOS COMETIDOS...	B	UI-1CD
F100016	INVESTIGACIÃ"N EN IZTAPALAPA	IZP-6	UI-1SD
F100017	INVESTIGACIÃ"N PARA LA ATENCIÃ"N DEL DEL...	E	2 CON DETENIDO 2 ...

Figura 4. Tabla de Fiscalia.

id	fecha_inicio	mes_inicio	ao_inicio
IC0	05/01/2016 18:35	Enero	2016
IC10	01/11/2017 19:21	Noviembre	2017
IC100	16/01/2016 17:15	Enero	2016
IC100001	09/03/2018 13:28	Marzo	2018
IC100003	07/11/2016 16:39	Noviembre	2016
IC100004	13/10/2016 18:27	Octubre	2016
IC100005	07/11/2016 18:37	Noviembre	2016
IC100006	24/02/2018 16:16	Febrero	2018
IC100008	24/02/2018 16:24	Febrero	2018
IC100009	07/11/2016 20:45	Noviembre	2016
IC10001	31/08/2016 00:59	Agosto	2016
IC100014	13/10/2016 20:50	Octubre	2016
IC100015	13/03/2018 23:18	Marzo	2018
IC100016	13/10/2016 20:58	Octubre	2016
IC100017	07/11/2016 21:59	Noviembre	2016

Figura 5. Tabla de Inicio_Carpeta.

id	colonia_hechos	alcaldia_hechos	calle_hechos	calle_hechos2	longitu
L0	GRANJAS MODERNAS	GUSTAVO A MADERO	SAN JUAN DE ARAGON	CALLE ANZAR	-99.103
L10	ROMA SUR	CUAUHTEMOC	AV. INSURGENTES SUR	BAJA CALIFORNIA	-99.168
L100	ARGENTINA PONIENTE	MIGUEL HIDALGO	LAGO MUSTER	AV. SAN BARTOLO NAUCALPAN Y CALZ. MEXIC...	-99.204
L100001	NARVARTE	BENITO JUAREZ	OBRERO MUNDIAL	AVENIDA CUAUHTEMOC	-99.155
L100003	GUERRERO	CUAUHTEMOC	METRO HIDALGO	A BORDO DE VAGON DEL METRO HIDALGO	-99.147
L100004	CHINAMPAC DE JUÃREZ	IZTAPALAPA	CALZ. IGNACIO ZARAGOZA	AVENIDA GUELATAO	-99.043
L100005	SANTA CRUZ ACALPIXCA - PUEBLO	XOCHIMILCO	LA GALLERA	AV. MEXICO-TULYEHUALCO	-99.068
L100006	DEL LAGO	BENITO JUAREZ	CALZADA DE TLALPAN	LAGO	-99.140
L100008	PEDREGAL DE SAN NICOLÃS 3A SECCIÃ"N	TLALPAN	HOMUN	MANI	-99.237
L100009	ZACATEPEC(SAN MATEO XALPA)	XOCHIMILCO	CIRCUITO JAVIER PIÃ'A Y PALACIOS	BENITO JUAREZ (SIN REGISTRO DEL SAP)	-99.121
L10001	SAN FELIPE DE JESÃšS	GUSTAVO A MADERO	ATIZAPAN DE ZARAGOZA	SAN JUAN DE LOS LAGOS	-99.070
L100014	ANAHUAC	MIGUEL HIDALGO	MARINA NACIONAL	LAGO SILVERIO	-99.180
L100015	SANTA MARIA NONOALCO	BENITO JUAREZ	AVENIDA REVOLUCION	BOTTICELLI	-99.186
L100016	SANTA MARTHA ACATITLA NORTE	IZTAPALAPA	AVENIDA TEXCOCO	MARCOS LOPEZ JIMENEZ	-99.014

Figura 6. Tabla de Locación.

id	fecha_hechos	mes_hechos	ao_hechos
H0	05/01/2016 17:00	Enero	2016
H10	01/11/2017 09:00	Noviembre	2017
H100	16/01/2016 15:45	Enero	2016
H100001	08/03/2018 22:00	Marzo	2018
H100003	03/11/2016 15:00	Noviembre	2016
H100004	13/10/2016 09:00	Octubre	2016
H100005	07/11/2016 10:30	Noviembre	2016
H100006	24/02/2018 13:00	Febrero	2018
H100008	11/02/2018 13:30	Febrero	2018
H100009	21/10/2016 12:30	Octubre	2016
H10001	30/08/2016 22:00	Agosto	2016
H100014	12/10/2016 20:40	Octubre	2016
H100015	08/03/2018 13:36	Marzo	2018
H100016	13/10/2016 11:15	Octubre	2016
H100017	07/11/2016 17:00	Noviembre	2016

Figura 7. Tabla de Tiempo_Hechos

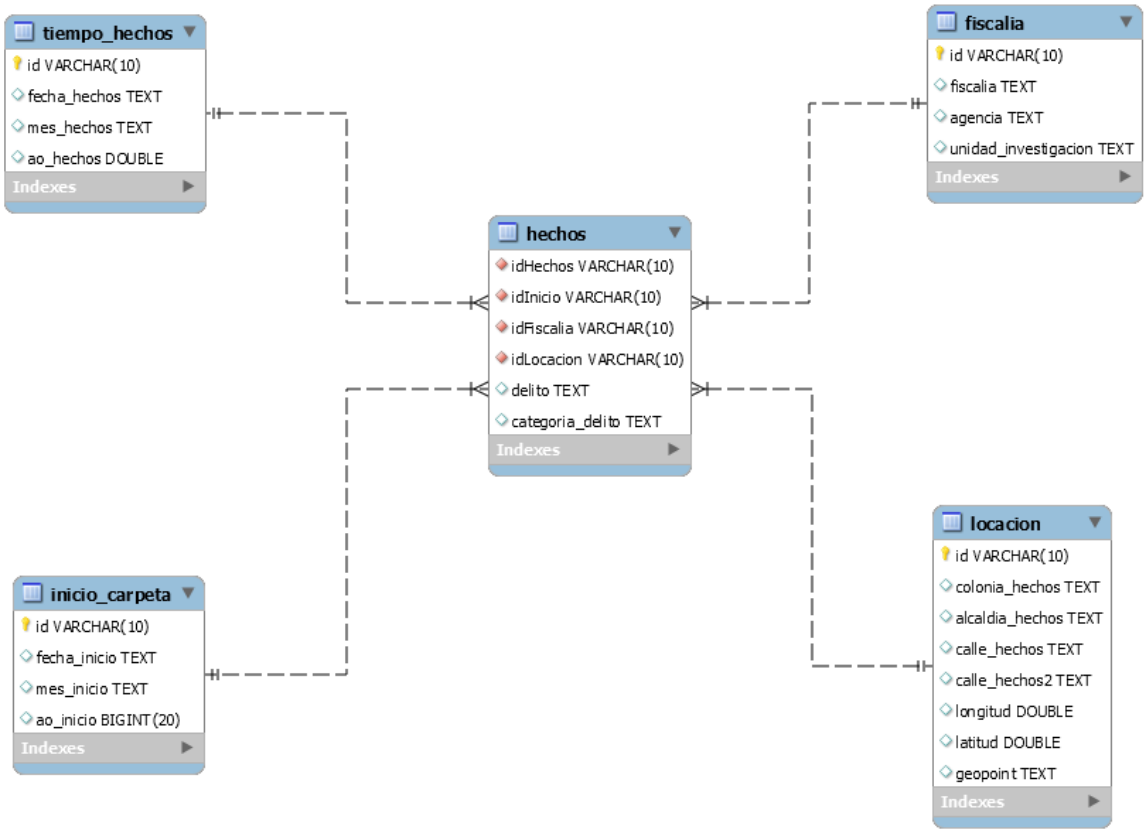

Figura 8. Modelo entidad relación generado en Workbench

6. Desarrolle los cubos de datos (los más importantes para usted) tipo roll-up y drill-down donde muestre con tableau el uso de su esquema de datos. Proporcione una interpretación de los resultados

- Se representaron a los cubos de datos por medio de vistas y se decidieron armar 3 cubos, los cuales se muestran a continuación con el código de su armado en SQL.

```
CREATE
    ALGORITHM = UNDEFINED
    DEFINER = `root`@`localhost`
    SQL SECURITY DEFINER
VIEW `c1` AS
    SELECT
        `h`.`categoria_delito` AS `categoria_delito`,
        `l`.`alcaldia_hechos` AS `alcaldia_hechos`,
        `t`.`mes_hechos` AS `mes_hechos`,
        `t`.`ao_hechos` AS `ao_hechos`
    FROM
        ((`hechos` `h`
        JOIN `locacion` `l`)
        JOIN `tiempo_hechos` `t`)
    WHERE
        ((`h`.`idLocacion` = `l`.`id`)
            AND (`h`.`idHechos` = `t`.`id`))
```

Figura 9. Estructura del cubo 1.

```
CREATE
    ALGORITHM = UNDEFINED
    DEFINER = `root`@`localhost`
    SQL SECURITY DEFINER
VIEW `c2` AS
    SELECT
        `h`.`delito` AS `delito`,
        `h`.`categoria_delito` AS `categoria_delito`,
        `l`.`alcaldia_hechos` AS `alcaldia_hechos`,
        `t`.`mes_hechos` AS `mes_hechos`,
        `t`.`ao_hechos` AS `ao_hechos`
    FROM
        ((`hechos` `h`
        JOIN `locacion` `l`)
        JOIN `tiempo_hechos` `t`)
    WHERE
        ((`h`.`idLocacion` = `l`.`id`)
            AND (`h`.`idHechos` = `t`.`id`))
```

Figura 10. Estructura del cubo 2.

```
CREATE
    ALGORITHM = UNDEFINED
    DEFINER = `root`@`localhost`
    SQL SECURITY DEFINER
VIEW `c3` AS
    SELECT
        `h`.`delito` AS `delito`,
        `t`.`mes_hechos` AS `mes_hechos`,
        `t`.`ao_hechos` AS `ao_hechos`
    FROM
        (`hechos` `h`
        JOIN `tiempo_hechos` `t`)
    WHERE
        (`h`.`idHechos` = `t`.`id`)
```

Figura 11. Estructura del cubo 3.

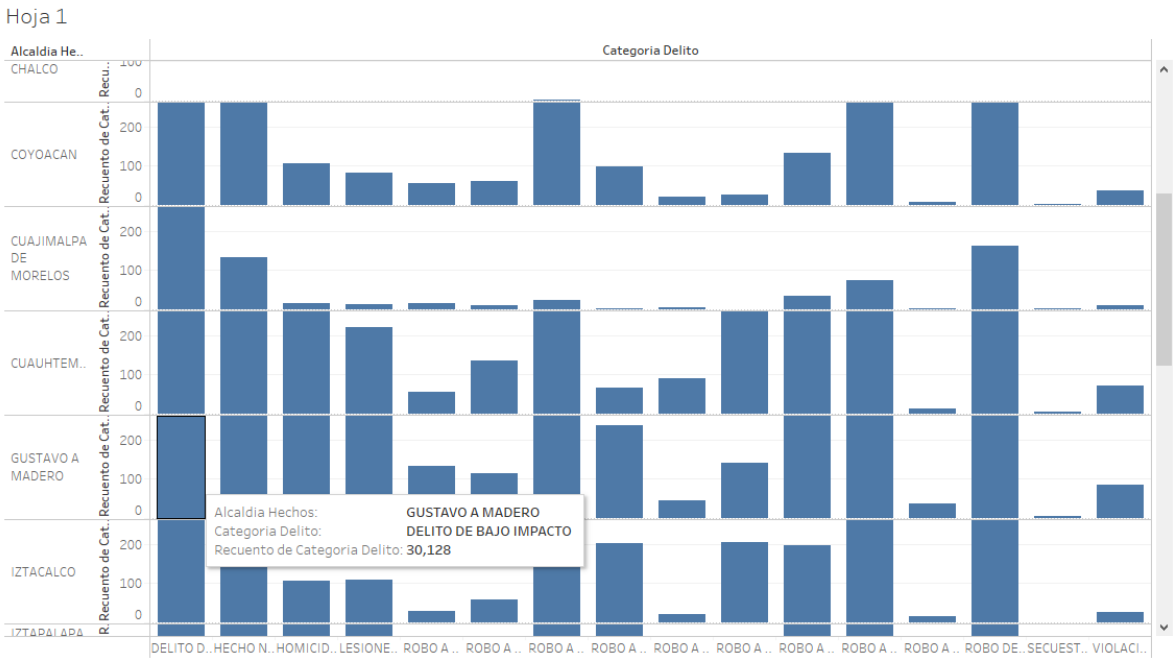

Figura 12. Representación del cubo 1.

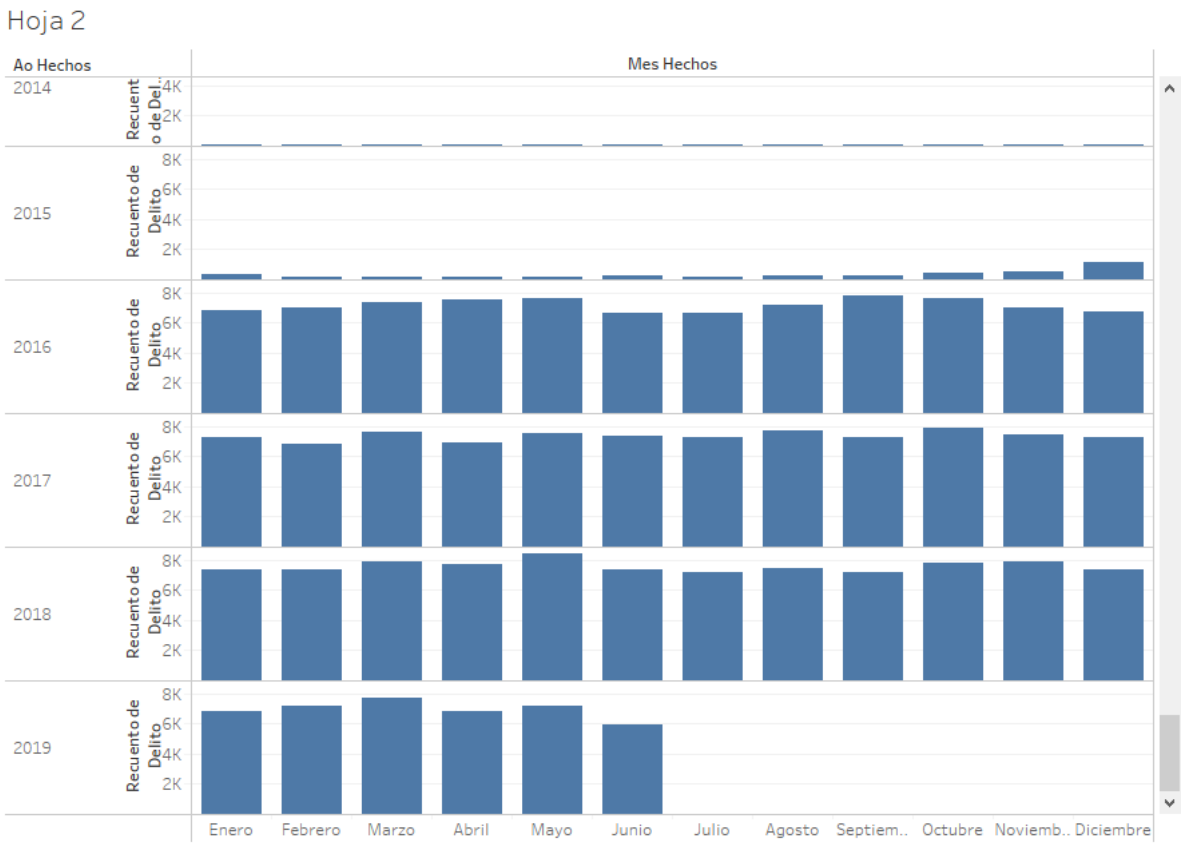

Figura 13. Representación del cubo 2.

4. CONCLUSIÓN

"En esta práctica pudimos trabajar y poner en práctica lo aprendido en clases teóricas con el dataset que elegimos en prácticas anteriores para nuestro proyecto, al cual le aplicamos la técnica ETL para recopilar y manejar directamente los datos para poderlos utilizar, además de la creación de unos cuantos cubos de datos que nos ayudaran a obtener la información para responder las preguntas que queremos responder con nuestro proyecto"

5. REFERENCIAS

[1]. Alberca, A. S. (2020, 4 octubre). La librería Pandas. Aprende con Alf. https://aprendeconalf.es/docencia/python/manual/pandas/

[2]. Pública, A. D. D. I. (s. f.). Portal de Datos Abiertos de la CDMX. Portal de datos Abiertos de la CDMX. Recuperado 23 de mayo de 2021, de https://datos.cdmx.gob.mx/dataset/carpetas-de-investigacion-pgj-cdmx

[3]. Nelson Aranibar, Monografias.com. (s. f.). MySQL WorkBench. Monografias.com. Recuperado 3 de marzo de 2021, de https://www.monografias.com/trabajos88/mysql-worckbench/mysql-worckbench.shtml

[4]. Tableau, una de las principales herramientas de Inteligencia de Negocios. (s. f.). SPnet. Recuperado 28 de marzo de 2021, de https://softwarepara.net/tableau

[5]. Pública, A. D. D. I. (s. f.). Portal de Datos Abiertos de la CDMX. Portal de datos Abiertos de la CDMX. Recuperado 23 de mayo de 2021, de https://datos.cdmx.gob.mx/dataset/carpetas-de-investigacion-pgj-cdmx