UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

# Diseño, implementación y explotación de un almacén de datos

Trabajo Fin de Grado

## Grado en Ingeniería Informática

**Autor**: Julen Ortí Rodríguez

**Tutor**: Laura Mota Herranz

Curso 2020-2021

# Resumen

Los Sistemas de Información Estratégicos tienen como objetivo aportar a una o varias organizaciones la información necesaria para el cumplimiento de sus fines. Estos sistemas permiten a éstas almacenar gran cantidad de información histórica con el propósito de extraer conocimiento de cara a realizar un análisis sobre el estado actual de la entidad o previsiones futuras de la misma.

Los almacenes de datos son sistemas tienen un gran coste monetario mensual, lo que conlleva un coste acumulado anual elevado, siendo para las pequeñas empresas inviable cubrir dichos gastos.

Para ello, en este trabajo final de grado se presenta una solución de Almacén de datos más económica que no implique a compañías externas para su construcción, implementación y mantenimiento. Además, se indicarán las operaciones de extracción, transformación y carga (*más conocidas como ETL*) sobre la fuente de datos original.

Esta solución no está enfocada exclusivamente a pequeñas entidades, sino que su uso puede ser extendido al ámbito personal o incluso universitario.

**Palabras clave:** Almacén de datos, análisis, base de datos, información histórica, ETL.

# Abstract

The objective of Strategic Information Systems is to provide one or more organisations with the necessary information to fulfil their purposes. These systems allow them to store a large amount of historical information with the purpose of extracting knowledge in order to carry out an analysis of the current state of the entity or its future forecasts.

Data warehouse is the best known concept. These systems have a high monthly monetary cost, which leads to a high annual accumulated cost, making it unfeasible for small companies to cover these expenses.

For this reason, this final project presents a more economical Data Warehouse solution that does not involve external companies for its construction, implementation and maintenance. In addition, the extraction, transformation and loading operations (better known as ETL) on the original data source will be indicated.

This solution is not exclusively focused on small entities, but its use can be extended to personal or even university environments.

**Keywords :** Data warehouse, analysis, database management, historical information, ETL.

# Índice

# Índice de ilustraciones





# Índice de tablas

# 1. Introducción

## 1.1. Objetivos

El objetivo de este trabajo es el diseño, implementación y explotación de un almacén de datos basado en información sobre colisiones de tráfico en el estado de California, Estados Unidos entre los meses de enero del año dos mil uno y diciembre del año dos mil veinte. Este almacén de datos será usado para la docencia de las asignaturas SIE (*Sistemas de Información Estratégicos*) y GDA (*Gestión de Bases de Datos*) del DSIC (*Departamento de Sistemas Informáticos y Computación*) de la UPV (*Universitat Politècnica de València*).

Evaluando lo que ofrecen los sistemas de información estratégicos actuales, se ha concluido que se debe ofrecer al usuario una experiencia ágil y sencilla en la obtención de resultados. Para ello se optimizará los tiempos de acceso utilizando técnicas de programación con el fin de minimizar el procesamiento de las consultas. En función de obtener estos resultados, se seguirán los siguientes caminos:

- Análisis del software actual que pueda concebir un almacén de datos.
- Estudio y modelado de la información.
- Análisis temporal de la solución propuesta respecto a las alternativas actuales.
- Explotación de los datos y generación de informes.

La información objeto de su extracción, transformación y explotación será el registro oficial de las colisiones de tráfico en el estado de California, Estados Unidos. La fuente original es la base de datos *Statewide Integrated Traffic Records System* (*SWITRS*) y es gestionada, entre otros, por la patrulla de tráfico de California, más conocida como *California Highway Patrol (CHP)*.

Con la explotación de dichos datos podremos obtener informes, en base a cuantiosas características, sobre las casuísticas más comunes entre los accidentes de tráfico.

## 1.2. Motivación

Desde mis primeras clases en la asignatura Bases de Datos y Sistemas de Información, con código 11548, me ha atraído el mundo de las bases de datos relacionales. Esta atracción fue creciendo cuando, en mi trabajo actual, me sumergí por completo en los algoritmos programados en SQL tanto en la variante de Microsoft en SQL Server como en Oracle con su software Oracle Developer.

Contemplar el desarrollo de algoritmos matemáticos, estudiarlos y entender cómo unas cuantas líneas de código pueden producir resultados tan importantes para mi empresa me hizo valorar que mi futuro como desarrollador y arquitecto en este ámbito era posible.

Recibí una propuesta de título por parte de mi tutora en la cual me proponía un reto personal: mi primer desarrollo desde cero de un almacén de datos. Todo esto conllevaba la búsqueda de una fuente abundante de datos y realizar todo el tratamiento y operaciones sobre la información de cara a obtener un resultado coherente y consistente sin perder su significado original.

## 1.3. Metodología

La metodología utilizada en este desarrollo desde el primer día ha sido *Waterfall*, conocido como cascada en español. Esta metodología se basa en etapas que se desarrollan de arriba abajo validando en cada punto si está preparado el desarrollo para pasar a la siguiente fase.

De cara a validar cada etapa, se han hecho reuniones donde valorábamos la calidad de la solución propuesta. Esto ha sido posible gracias a que los requisitos han sido definidos desde el momento en el cual se decidió la fuente de los datos y se llevó a cabo su análisis y en ningún momento han variado.

En caso de que la solución no fuese la indicada, se volvía al inicio de la etapa valorando dónde se había fallado y los pasos a seguir para solucionar los errores detectados.

## 1.4. Estructura

La memoria se ha organizado en los siguientes apartados:

En el primer apartado, *Introducción*, se marcan los objetivos, motivación y metodología del trabajo desarrollado.

En el segundo apartado, *Contexto tecnológico*, se introduce el concepto de Almacén de Datos, así como sus propiedades y metodología de diseño. Se realiza una introducción al software utilizado para la explotación de los datos (*Power BI*).

En el tercer apartado, *Análisis del problema*, se explicará el origen del problema planteado. Asimismo, se realiza un análisis con las distintas líneas de trabajo a seguir a lo largo del desarrollo.

En el cuarto apartado*, Desarrollo de la solución*, se detallará cada una de las fases definidas en el plan de trabajo.

En el quinto apartado, *Implantación*, se describirá cómo se ha cargado la información en *Power BI Desktop*, incluyendo la creación de dimensiones y generación de informes, entre otros aspectos.

Por último, en los apartados seis y siete, *Conclusiones* y *Referencias* respectivamente, se expondrán las conclusiones alcanzas tras acabar el trabajo y las referencias a las fuentes externas utilizadas.

# 2. Contexto tecnológico

En este apartado se va a llevar a cabo una explicación detallada sobre qué son los almacenes de datos, el tipo de arquitectura en el cual están basados y la definición de cómo se estructura la información.

## 2.1. ¿Qué es un almacén de datos?

Se ha planteado la definición técnica de un **sistema de información** como “un conjunto de componentes interrelacionados que recolectan (o recuperan), procesan, almacenan y distribuyen información para apoyar los procesos de toma de decisiones y de control en una organización. Además de apoyar la toma de decisiones, la coordinación y el control, los sistemas de información también pueden ayudar a los gerentes y trabajadores del conocimiento a analizar problemas, visualizar temas complejos y crear nuevos productos." (Laudon y Laudon, 2012, p.15)

Dentro de estos sistemas de información estratégicos se puede encontrar a los conocidos como Almacenes de Datos (*Data Warehouse*). Un almacén de datos es una colección de datos de distintos orígenes estructurada y orientada al uso en un determinado ámbito empresarial siendo integrada, no volátil y variable en el tiempo.

Poniendo en contexto sus características, se dice que es destinada al uso en el ámbito empresarial debido a que se diseña específicamente con la información relativa a las actividades de la organización.

Se trata de una colección de datos integrada debido a que aúna datos recolectados de diferentes software utilizados dentro la propia empresa u obtenidos a través de orígenes de datos externos a la misma. Este proceso se considera el más costoso en términos monetarios y temporales a la hora de construir un almacén de datos.

De cara a obtener un estudio con resultados coherentes, deberán realizarse tareas en el procesamiento de la información para ser capaces de compatibilizar los distintos formatos en los cuales una organización, ya sea la cliente o un proveedor externo, aporta los datos. Todo usuario desea poder separar y combinar sin límites datos de forma analítica.

Tras su procesamiento, es indispensable realizar una limpieza de todo aquello que no sea consistente ni tenga valor alguno. Los usuarios necesitan poder confiar en aquello que se le está mostrando en base a sus conocimientos.

Se considera variable en el tiempo y no volátil ya que la información almacenada no varía con el paso del tiempo, sino que se incrementa periódicamente en base a la frecuencia de envío de la información por parte de los orígenes de datos. Por tanto, **se trata de integrar datos históricos** como pueden llegar a ser las ventas de un negocio, sus anteriores clientes o el estado respecto a las competencias en el mercado.

Tras realizar un estudio del tipo de sistema de los almacenes de datos, hay que aclarar que no se deben confundir los sistemas de procesamiento analítico en línea, más conocidos como sistemas OLAP, con el uso de sistemas de procesamiento transaccional en línea, llamados sistemas OLTP.

Los sistemas OLTP modifican la base de datos mediante transacciones realizadas por el usuario sobre un determinado conjunto de datos. Es muy común la realización de operaciones de inserción, actualización o borrado de datos. En lenguaje SQL corresponden a operaciones INSERT, UPDATE, DELETE respectivamente. Para este tipo de sistemas es esencial que se cumplan las cuatro propiedades ACID.

La información manejada por las aplicaciones OLTP tiende a ser relativamente actual y no con valor histórico. Esto conlleva que la información sea volátil. Otra característica de gran importancia de estos sistemas es la velocidad transaccional, la cual es de alta velocidad por el bajo volumen de datos por transacción. Estos sistemas son los utilizados en bases de datos al uso.

En contraposición, se encuentran los sistemas OLAP. Este tipo de aplicaciones se basan en la consulta, SELECT en lenguaje SQL, de grandes cantidades de datos para realizar un análisis de cara a la toma de decisiones, siendo su naturaleza histórica. Al igual que el sistema OLTP, es necesaria que la velocidad transaccional sea la mayor posible a pesar de que el usuario realice consultas ad-hoc.

Dentro de los sistemas OLAP se encuentra su variación relacional, denominada ROLAP, al ser la tecnología relacional la que soporta el AD. Éstos, se encuentran almacenados dentro de bases de datos relaciones, aunque se evita la normalización de las tablas que la conforman a pesar de mantener la agregación de los datos.

En referencia a la arquitectura, la más utilizada es un almacén de datos basado en una arquitectura de tres niveles. Para ello, vamos a mostrar los fundamentos de la metodología de diseño de un almacén de datos.

Dichos sistemas no se consideran una base de datos estándar al uso, sino un conjunto de colecciones de bases de datos estructuradas. En consecuencia, el

modelado de los AD se basa en la visión multidimensional de la información donde los datos objeto de estudio se representan mediante una tabla llamada *tabla de hechos* junto con los indicadores que interesa analizar acompañado de tablas que caracterizan la actividad que son las *tablas de dimensiones*.

Este tipo de estructuración de la información lleva a la simplicidad de las bases de datos siendo muy eficientes en los tiempos en los cuales el software de inteligencia empresarial devuelve los resultados de los informes solicitados por el usuario.

## 2.2. Metodología de diseño de un almacén de datos

Dentro de la metodología de diseño de una almacén de datos se encuentran tres etapas diferenciadas que deben ejecutarse en orden secuencial para la obtención del modelo multidimensional deseado. Cabe mencionar, que el software escogido para el desarrollo inicial del almacén de datos es *PostgreSQL*, posteriormente los datos generados se cargarán al *Power BI*.

### 2.2.1. Diseño conceptual

Tras el análisis y la recogida de requisitos, se debe hacer un diseño conceptual multidimensional sobre los datos que recibimos como *input.*

En un *esquema multidimensional* se representa una actividad que es objeto de análisis (*hecho*) y las dimensiones que caracterizan la actividad (*dimensiones*).

La información relevante sobre el hecho se representa por un conjunto de indicadores (*medidas o atributos de hecho*).

La información descriptiva de cada dimensión se representa por un conjunto de atributos (*atributos de dimensión*).

Tal y como se ha comentado anteriormente, el esquema utilizado en las construcciones de almacenes de datos es el ROLAP y se puede encontrar varios tipos de representaciones del esquema multidimensional. Comenzamos presentando el *esquema en estrella*, en este esquema, como se puede ver en la figura, la actividad objeto de estudio se representa en el centro y las dimensiones en las puntas de la estrella.

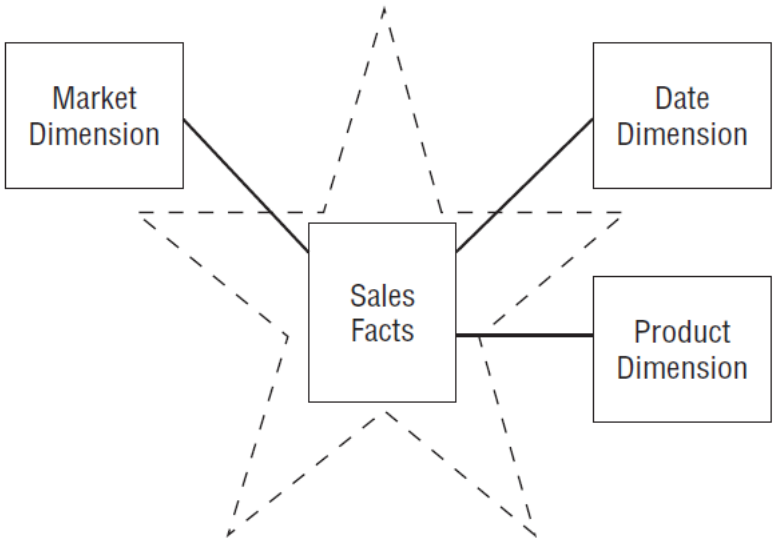



*Ilustración 1. Esquema en estrella, tomado de The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (p. 9), de Ralph Kimball & Margy Ross, 2013, John Wiley & Sons, Inc.*

Se debe tener presente una característica importante en los esquemas relacionales que no se produce dentro de los almacenes de datos. Esto es que las tablas finales del modelo propuesto, tras la transformación de los datos, no se normalizan a tercera forma normal, notación 3NF. Para que una tabla esté en tercera forma normal no se deben producir dependencias funcionales entre los atributos de la tabla y los atributos que conforman la clave. Es decir, que no se produzcan redundancias de significado y contenido en la tabla diseñada. El hecho de no normalizar las tablas es debido a que en un almacén de datos no se busca eliminar la redundancia de la información.

Las estructuras en tercera forma normal son inmensamente útiles en el procesamiento operacional porque una inserción o actualización toca una sola parte de la base de datos. (…) la complejidad de las impredecibles consultas de los usuarios abruma los optimizadores de bases de datos, resultando en un rendimiento desastroso de las consultas. El uso de un modelo normalizado en los Almacenes de Datos/Inteligencia Empresarial derrota el gran rendimiento y la intuitiva recuperación de datos. (Kimball & Ross, 2013, p.8)

Como alternativa al esquema en estrella, el *esquema en copo de nieve* (“*Snowflake Schema”*) contradice los principios definidos para un almacén de datos. Este tipo de modelo se basa en la normalización de las tablas del esquema relacional eliminando redundancia, aumentando la complejidad de las consultas a la base de datos pudiendo afectar al rendimiento de éstas y convirtiéndose en un desarrollo más difícil tanto para el programador como para aquél que necesite entender el modelo de datos.

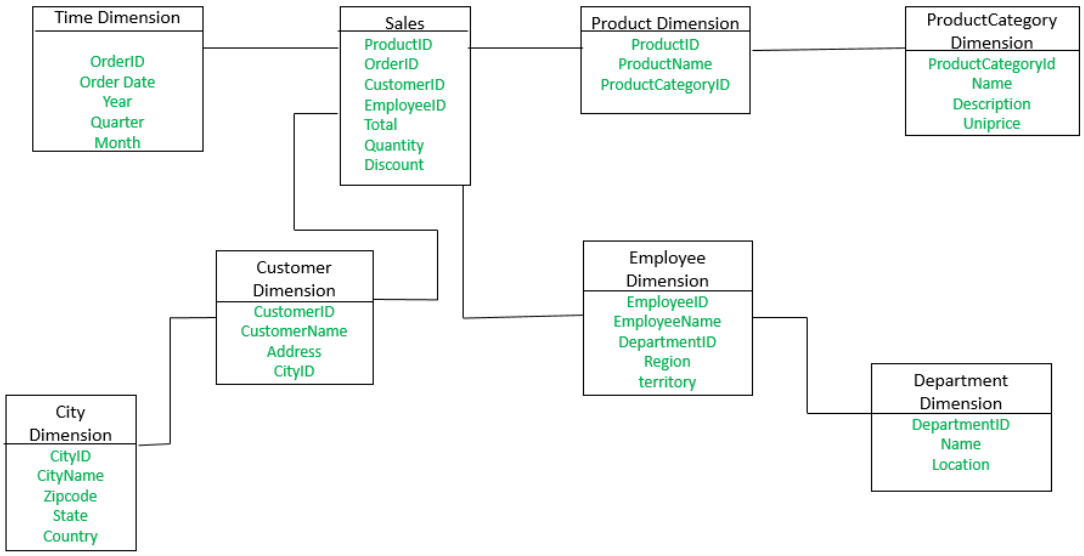



*Ilustración 2. Esquema en copo de nieve, Adaptado de Maurya, P. (2021). Snowflake Schema in Data Warehouse Model. Noida, Uttar Pradesh: GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/snowflake-schema-in-data-warehouse-model/*

Ciertas dimensiones tienen características almacenadas en subdimensiones. Desde el punto de vista de un esquema en estrella, la dimensión se consideraría el hecho objeto de estudio y las subdimensiones serían las dimensiones del hecho. Las relaciones entre estas dimensiones son de uno a muchos, relacionadas mediante claves ajenas con los identificadores de cada casuística de la dimensión.

El modelo resultante del análisis y diseño se representará mediante un diagrama de clases UML donde el hecho será una clase débil y las dimensiones clases relacionadas con ella mediante las cardinalidades definidas.

### 2.2.2. Diseño lógico

Una vez definido el modelo multidimensional con el esquema que se haya seleccionado, se transformará el diagrama de clases UML a un diagrama relacional sobre el que trabajaremos y hemos comentado anteriormente, ROLAP.

Las dimensiones pasarán a ser tablas de una base de datos relacional donde se definirá un identificador único por cada casuística que se produzca en la dimensión. Este identificador será utilizado más adelante por la tabla de hechos.

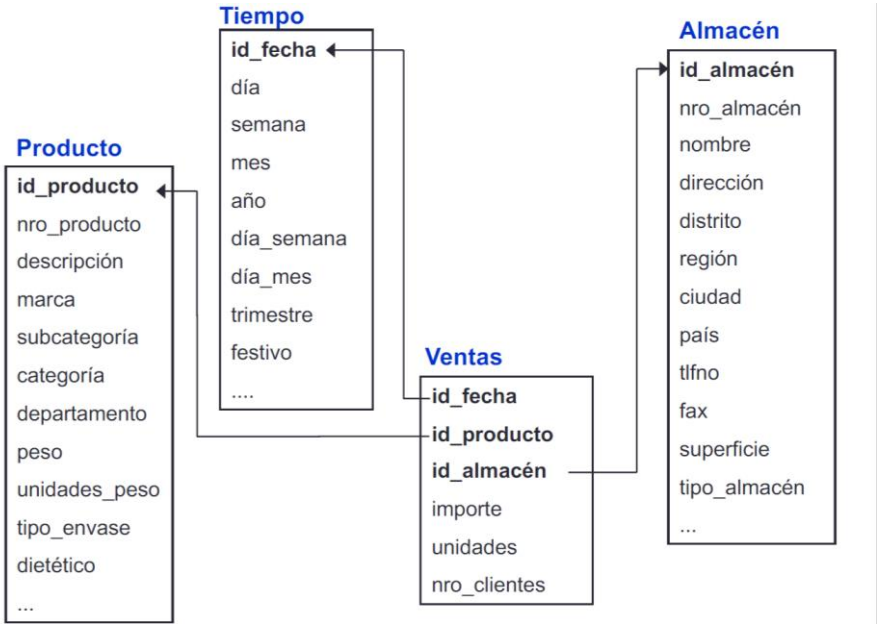



*Ilustración 3. Casamayor, J. (2020 - 2021). Sistemas de Información Estratégicos Parte I: Almacenes de Datos Tema 4: Diseño de Almacenes de Datos (maestría). Universitat Politècnica de València, Valencia*

El hecho, al igual que las dimensiones, será una tabla de una base de datos relacional la cual contendrá los identificadores de cada casuística producida en las dimensiones y que estén relacionadas entre ellas. Se definirán claves ajenas a las dimensiones por dichos identificadores. También, se definirá una clave primaria compuesta por los identificadores.

### 2.2.3. Diseño físico

La fase de diseño físico trata de buscar una optimización de los tiempos de consulta sobre las tablas de hechos y dimensiones. Tal y como se ha comentado anteriormente, es muy importante que el usuario obtenga una respuesta rápida a pesar de tener que consultar una cantidad cuantiosa de datos, del orden de cientos de millones de registros.

En la tabla de hechos se van a realizar acciones de creación de índices sobre los identificadores de las dimensiones. El orden de las columnas al crear el índice es muy importante. Además, se valorará la posibilidad de particionar la tabla a partir de las propias claves, creando un índice local para cada partición.

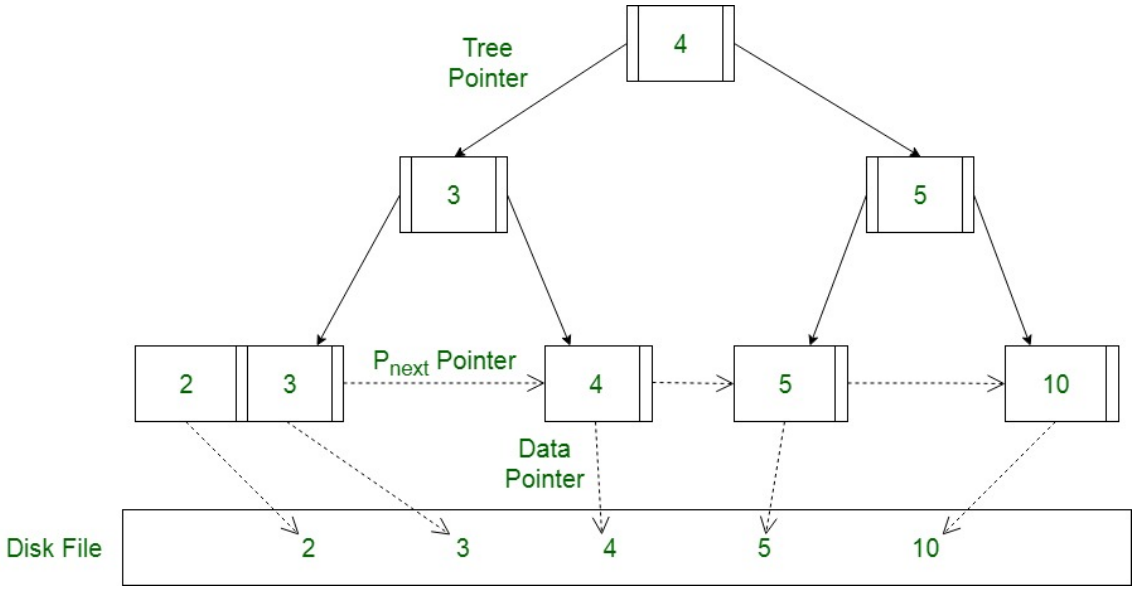



*Ilustración 4. Árbol B+, Adaptado de Adhikary, S. (2020). Introduction of B+ Tree. Noida, Uttar Pradesh: GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/introduction-of-b-tree/*

En las tablas de dimensiones el procedimiento será similar que en la tabla de hechos. Se crearán índices por el identificador de cada casuística y, en caso de ser necesario, por columnas que se utilicen recursivamente en consultas.

## 2.3. Herramienta Power BI

Power BI Desktop es una aplicación que permite conectarse a múltiples orígenes de datos, transformarlos y visualizarlos en base a un modelo de datos. Los usos más comunes de este software son:

- Conectar datos.
- Transformar y limpiar datos con el fin de realizar un modelado de estos.
- Crear objetos visuales, como gráficos o grafos, que proporcionan representaciones visuales de la información.
- Crear informes.

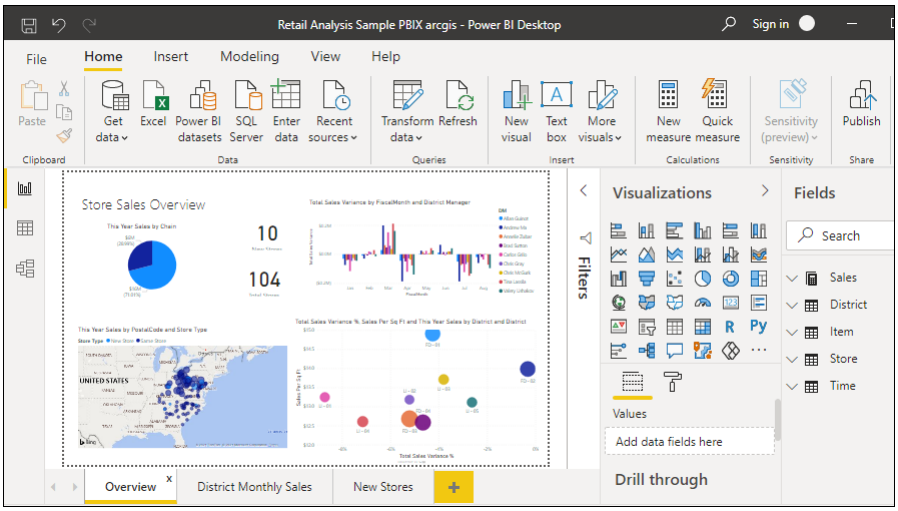



*Ilustración 5. Interfaz gráfica de Power BI Desktop. Iseminger, D. (2021). What is Power BI Desktop? Microsoft Docs. Recuperado de https://docs.microsoft.com/es-es/power-bi/fundamentals/desktop-what-is-desktop*

Esta herramienta ha sido la seleccionada para realizar la explotación de la información una vez terminadas las fases análisis, extracción, transformación y carga de los datos. Con ella obtendremos informes y estadísticas sobre las tendencias, en base a distintos factores, de las colisiones de tráfico en el estado de California, Estados Unidos. Esto es debido a que la Universitat Politècnica de València, hace uso de este software en la docencia a los alumnos.

# 3. **Análisis del problema**

En este apartado se va a exponer la necesidad del proyecto al igual que el por qué se ha utilizado como fuente de datos las colisiones de tráfico.

Actualmente en las asignaturas Sistemas de Información Estratégicos y Gestión de Bases de Datos, en el departamento DSIC, utilizan un caso de estudio basado en las ventas por productos de un cierto grupo de tiendas en distintas ciudades de Estados Unidos. Dicho caso lleva siendo impartido en la docencia durante varios años. Por tanto, nació la necesidad de encontrar una problemática diferente y con un volumen de datos mayor al actual.

Por ello, ante mi necesidad y requerimiento de hacer un trabajo final de grado que estuviese dentro del marco de las bases de datos, especialmente trabajar con el lenguaje de programación SQL, y la necesidad del departamento para darle un lavado de cara a las prácticas de la asignatura entramos en contacto para ofrecerles un almacén de datos diseñado, implementado y listo para ser explotado por los alumnos de las asignaturas.

En base a obtener una fuente de información pública, se ha indagado en distintos portales web, tales como el Instituto Nacional de Estadística (*INE*) o Kaggle (*la comunidad de científicos más conocida de Estados Unidos, subsidiaria de Google LLC*). Dentro de esta última se ha localizado la base de datos con la información de las colisiones de tráfico en el estado de California.

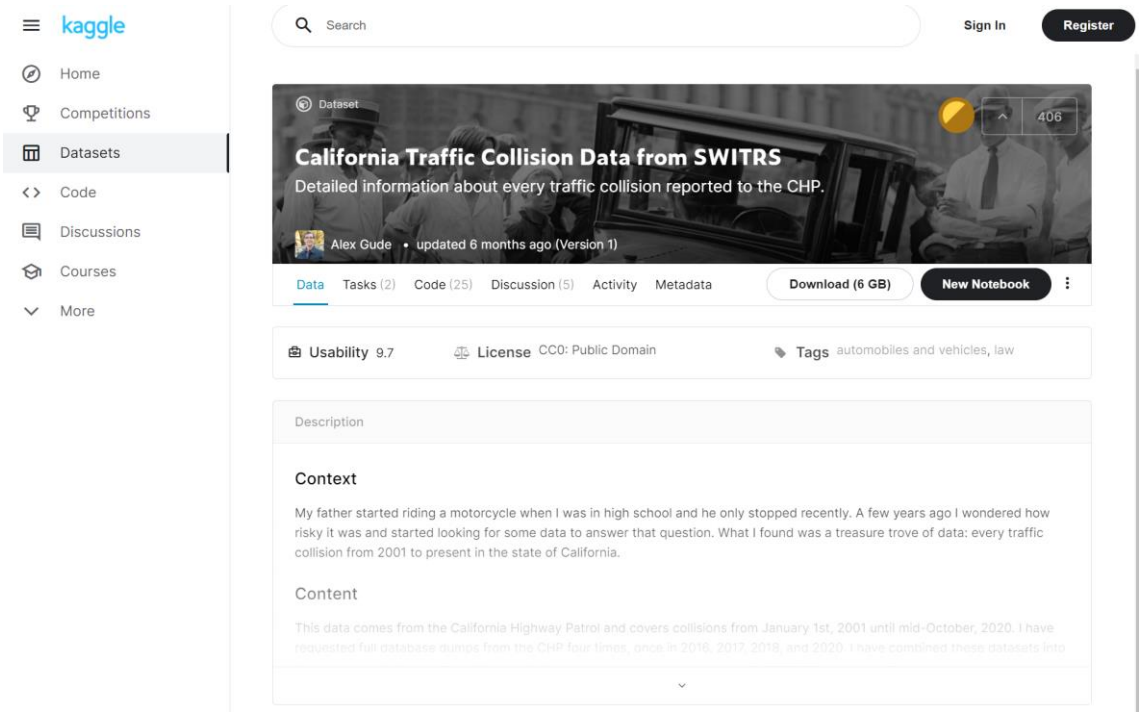



*Ilustración 6. Portal web de Kaggle.com, en concreto, el origen de la base de datos sobre las colisiones de tráfico en el estado de California. Recuperado de https://www.kaggle.com/alexgude/california-traffic-collision-data-from-switrs*

Dicha base de datos contiene aproximadamente de diez millones de colisiones de tráfico entre los años dos mil uno y dos mil veinte. En ella se pueden encontrar información de los vehículos infractores, vehículos implicados, peatones implicados, el tiempo, lugar, meteorología, estado de la vía y referencias a las personas infractoras e implicadas.

## 3.1. Plan de trabajo

Todo proyecto tiene sus fases de captación, desarrollo y utilización y en este apartado vamos a describir cuáles son y la estimación en coste temporal de cada una de ellas. Se acompañará de un diagrama de Gantt ilustrativo con el desglose de tareas, las dependencias entre ellas y la duración.

### 3.1.1. Captación de la información

Se investigará dentro de aquellos portales web, que ofrezcan fuentes de información gratuitas, un objeto de estudio óptimo para las necesidades de la docencia. Además, se realizará la extracción de los datos de dicha fuente. Se estima que estas tareas puedan abarcar tres jornadas.

### 3.1.2. Diseño conceptual multidimensional

Una vez obtenida la información, se organizará la información realizando un estudio de esta redistribuyéndola y eliminando aquella que no aporte valor tenerla. Tras ello de diseñará en un diagrama de clases UML las dimensiones con sus atributos y el hecho con sus atributos. Se estima la duración de estas tareas en diez jornadas.

### 3.1.3. Diseño lógico

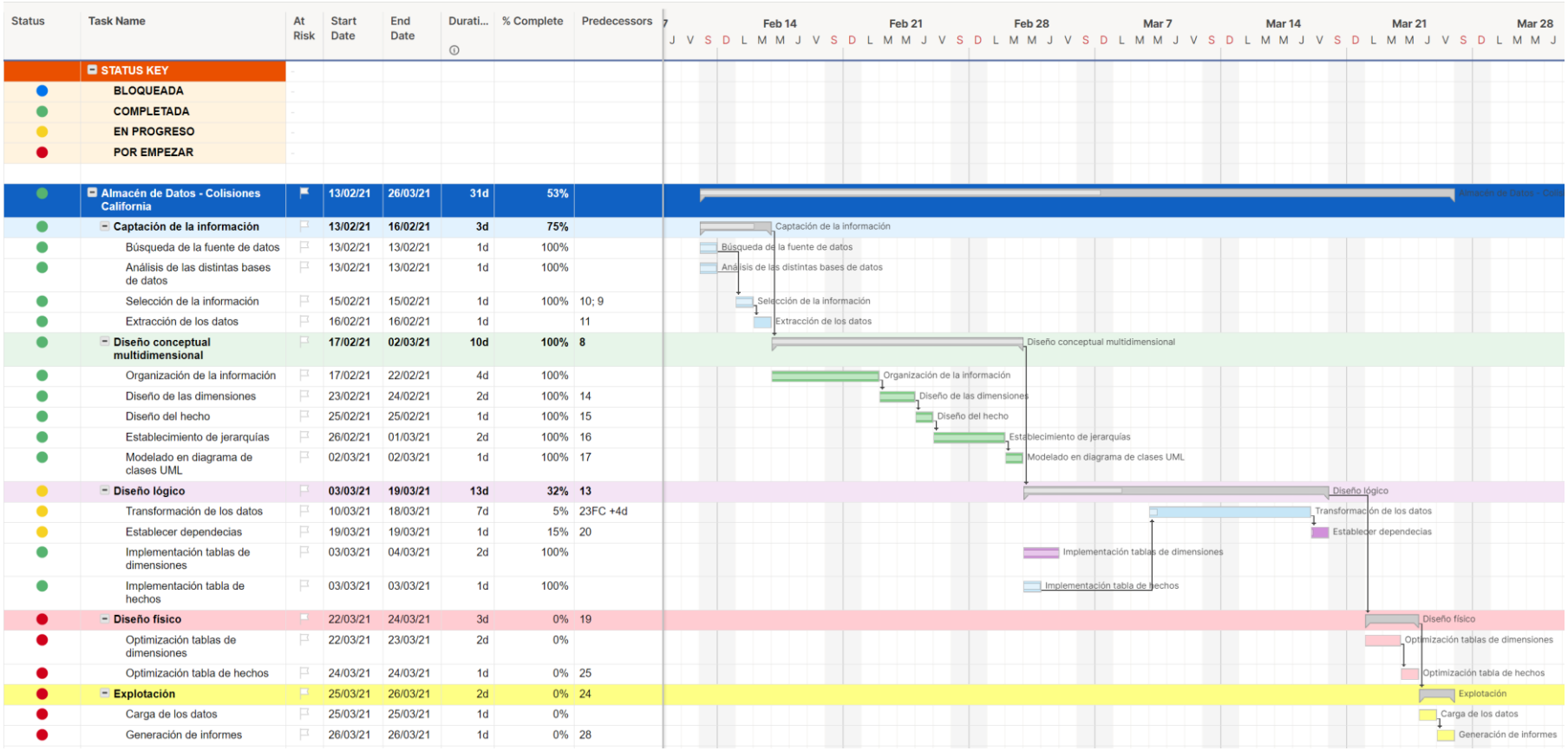
Tras el modelado de la información en un modelo multidimensional, se transformará este diagrama de clases a un diseño en una base de datos relacional y se implementarán las tablas de dimensiones y la tabla de hechos. Además, se transformarán los datos necesarios para su correcto entendimiento y se establecerán las dependencias mediante claves ajenas. Se estima esta fase en trece jornadas de trabajo.

### 3.1.4. Diseño físico

Con la implementación del modelo realizada, optimizaremos los tiempos de acceso a los datos mediante la creación de índices en las tablas de dimensiones y en la tabla de hechos. En caso de que la volumetría sea muy grande en la tabla de hechos, se valorará la posibilidad de particionarla. Se estima esta fase en tres jornadas.

### 3.1.5. Explotación

Por último, se cargará el almacén de datos diseñado en el software Power BI Desktop donde se ultimará el diseño. Una vez finalizado el desarrollo, se obtendrán informes generados por la herramienta. Se estima la duración de esta fase en dos jornadas.



| Status | Task Name | At Risk | Start Date | End Date | Durati... | % Complete | Predecessors |
|---|---|---|---|---|---|---|---|
| | STATUS KEY | | | | | | |
| | BLOQUEADA | | | | | | |
| | COMPLETADA | | | | | | |
| | EN PROGRESO | | | | | | |
| | POR EMPEZAR | | | | | | |
| | Almacén de Datos - Colisiones California | | 13/02/21 | 26/03/21 | 31d | 53% | |
| | Captación de la información | | 13/02/21 | 16/02/21 | 3d | 75% | |
| | Búsqueda de la fuente de datos | | 13/02/21 | 13/02/21 | 1d | 100% | |
| | Análisis de las distintas bases de datos | | 13/02/21 | 13/02/21 | 1d | 100% | |
| | Selección de la información | | 15/02/21 | 15/02/21 | 1d | 100% | 10; 9 |
| | Extracción de los datos | | 16/02/21 | 16/02/21 | 1d | | 11 |
| | Diseño conceptual multidimensional | | 17/02/21 | 02/03/21 | 10d | 100% | 8 |
| | Organización de la información | | 17/02/21 | 22/02/21 | 4d | 100% | |
| | Diseño de las dimensiones | | 23/02/21 | 24/02/21 | 2d | 100% | 14 |
| | Diseño del hecho | | 25/02/21 | 25/02/21 | 1d | 100% | 15 |
| | Establecimiento de jerarquías | | 26/02/21 | 01/03/21 | 2d | 100% | 16 |
| | Modelado en diagrama de clases UML | | 02/03/21 | 02/03/21 | 1d | 100% | 17 |
| | Diseño lógico | | 03/03/21 | 19/03/21 | 13d | 32% | 13 |
| | Transformación de los datos | | 10/03/21 | 18/03/21 | 7d | 5% | 23FC +4d |
| | Establecer dependecias | | 19/03/21 | 19/03/21 | 1d | 15% | 20 |
| | Implementación tablas de dimensiones | | 03/03/21 | 04/03/21 | 2d | 100% | |
| | Implementación tabla de hechos | | 03/03/21 | 03/03/21 | 1d | 100% | |
| | Diseño físico | | 22/03/21 | 24/03/21 | 3d | 0% | 19 |
| | Optimización tablas de dimensiones | | 22/03/21 | 23/03/21 | 2d | 0% | |
| | Optimización tabla de hechos | | 24/03/21 | 24/03/21 | 1d | 0% | 25 |
| | Explotación | | 25/03/21 | 26/03/21 | 2d | 0% | 24 |
| | Carga de los datos | | 25/03/21 | 25/03/21 | 1d | 0% | |
| | Generación de informes | | 26/03/21 | 26/03/21 | 1d | 0% | 28 |



*Ilustración 7. Diagrama de Gantt. Tareas por fases, estimación temporal y estado actual*

## 3.2. Presupuesto

Una vez estimadas las jornadas totales de trabajo que serán necesarias para el desarrollo de este proyecto, obtenemos el coste total del software utilizado y del caudal humano, tomando como referencia el sueldo medio de un desarrollador SQL en España.

El coste del software utilizado será:

| Software | Descripción | Coste mensual (€) |
|---|---|---|
| PostgreSQL | Es un potente software de código abierto para sistemas de bases de datos relacionales. | 0,00 |
| Power BI | Es una aplicación que permite conectarse a múltiples orígenes de datos, transformarlos y visualizarlos en base a un modelo de datos. Su uso particular es gratuito durante un período de ciento veinte días | 4.212,30 |
| | | 4.212,30 |

*Tabla 1. Costes del software.*

El coste del caudal humano (*con el salario obtenido de https://es.indeed.com/career/desarrollador-pl%2Fsql/salaries)*:

| Caudal humano (*personas*) | Duración (jornadas laborales) | Coste mensual (€) |
|---|---|---|
| 1 | 31 | 1.761,00 |

*Tabla 2. Costes caudal humano*

Teniendo en cuenta las estimaciones anteriores, el coste del desarrollo y explotación de este proyecto se dividirá:

| Coste inicial (€) | Coste mensual (€) | Coste total (*€, tras un mes*) |
|---|---|---|
| 4.526 | 4.212,30 | 8.738,30 |

*Tabla 3. Coste total del proyecto*

# 4. Diseño y desarrollo de la solución

En este apartado, se va a realizar la especificación de cada una de las fases descritas anteriormente, tras la obtener la información relacionada con los accidentes de tráfico en el estado de California, Estados Unidos.

## 4.1. Diseño conceptual multidimensional

Para poder realizar el modelado multidimensional, se ha realizado previamente un análisis de la información objeto de estudio. De esta forma se ha podido reducir la volumetría de columnas, siendo las descartadas aquellas que no aportaban valor tener por su significado o debido a que el dato significante no era consistente.

Se va a mostrar el estado original de las tablas de la base de datos obtenida:

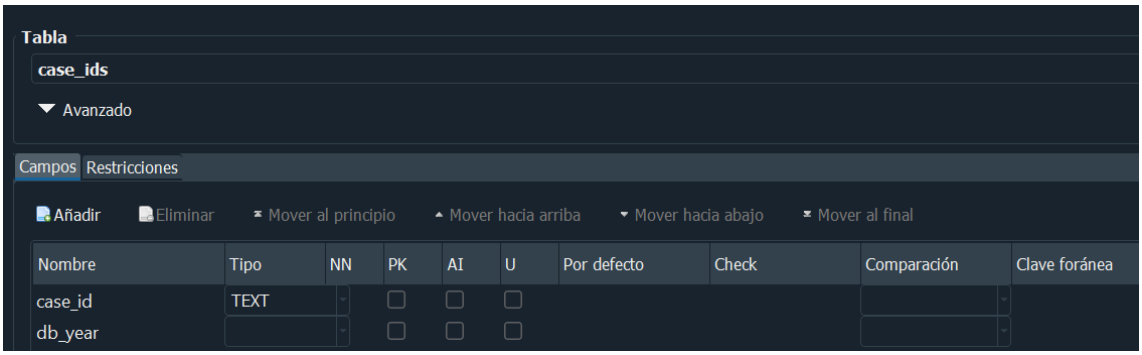
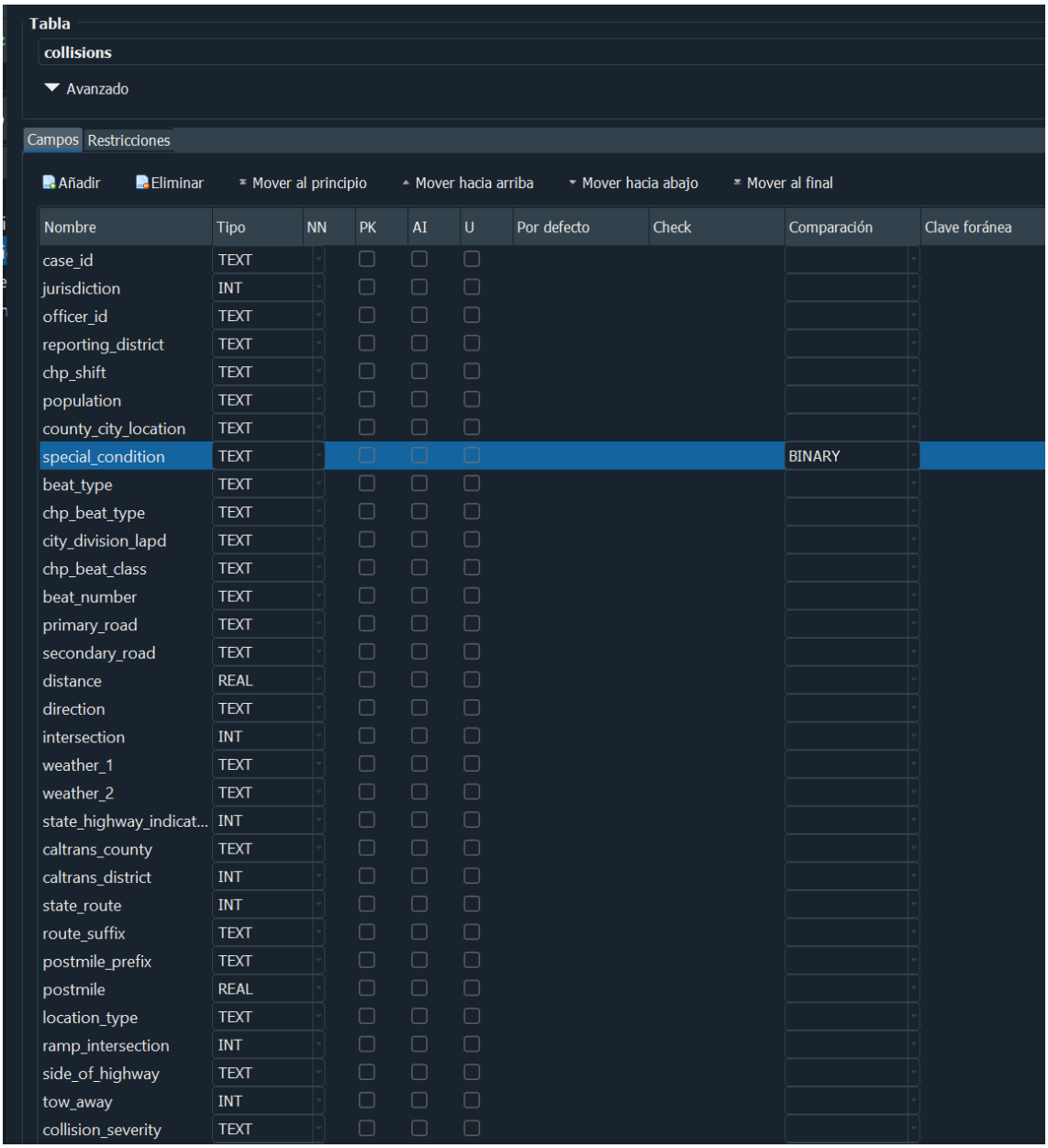



*Ilustración 8. Tabla case_ids*




| Nombre | Tipo | NN | PK | AI | U | Por defecto | Check | Comparación | Clave foránea |
|---|---|---|---|---|---|---|---|---|---|
| case_id | TEXT | | | | | | | | |
| jurisdiction | INT | | | | | | | | |
| officer_id | TEXT | | | | | | | | |
| reporting_district | TEXT | | | | | | | | |
| chp_shift | TEXT | | | | | | | | |
| population | TEXT | | | | | | | | |
| county_city_location | TEXT | | | | | | | | |
| special_condition | TEXT | | | | | | | BINARY | |
| beat_type | TEXT | | | | | | | | |
| chp_beat_type | TEXT | | | | | | | | |
| city_division_lapd | TEXT | | | | | | | | |
| chp_beat_class | TEXT | | | | | | | | |
| beat_number | TEXT | | | | | | | | |
| primary_road | TEXT | | | | | | | | |
| secondary_road | TEXT | | | | | | | | |
| distance | REAL | | | | | | | | |
| direction | TEXT | | | | | | | | |
| intersection | INT | | | | | | | | |
| weather_1 | TEXT | | | | | | | | |
| weather_2 | TEXT | | | | | | | | |
| state_highway_indicat... | INT | | | | | | | | |
| caltrans_county | TEXT | | | | | | | | |
| caltrans_district | INT | | | | | | | | |
| state_route | INT | | | | | | | | |
| route_suffix | TEXT | | | | | | | | |
| postmile_prefix | TEXT | | | | | | | | |
| postmile | REAL | | | | | | | | |
| location_type | TEXT | | | | | | | | |
| ramp_intersection | INT | | | | | | | | |
| side_of_highway | TEXT | | | | | | | | |
| tow_away | INT | | | | | | | | |
| collision_severity | TEXT | | | | | | | | |

*Ilustración 9. Tabla collisions pt.1*

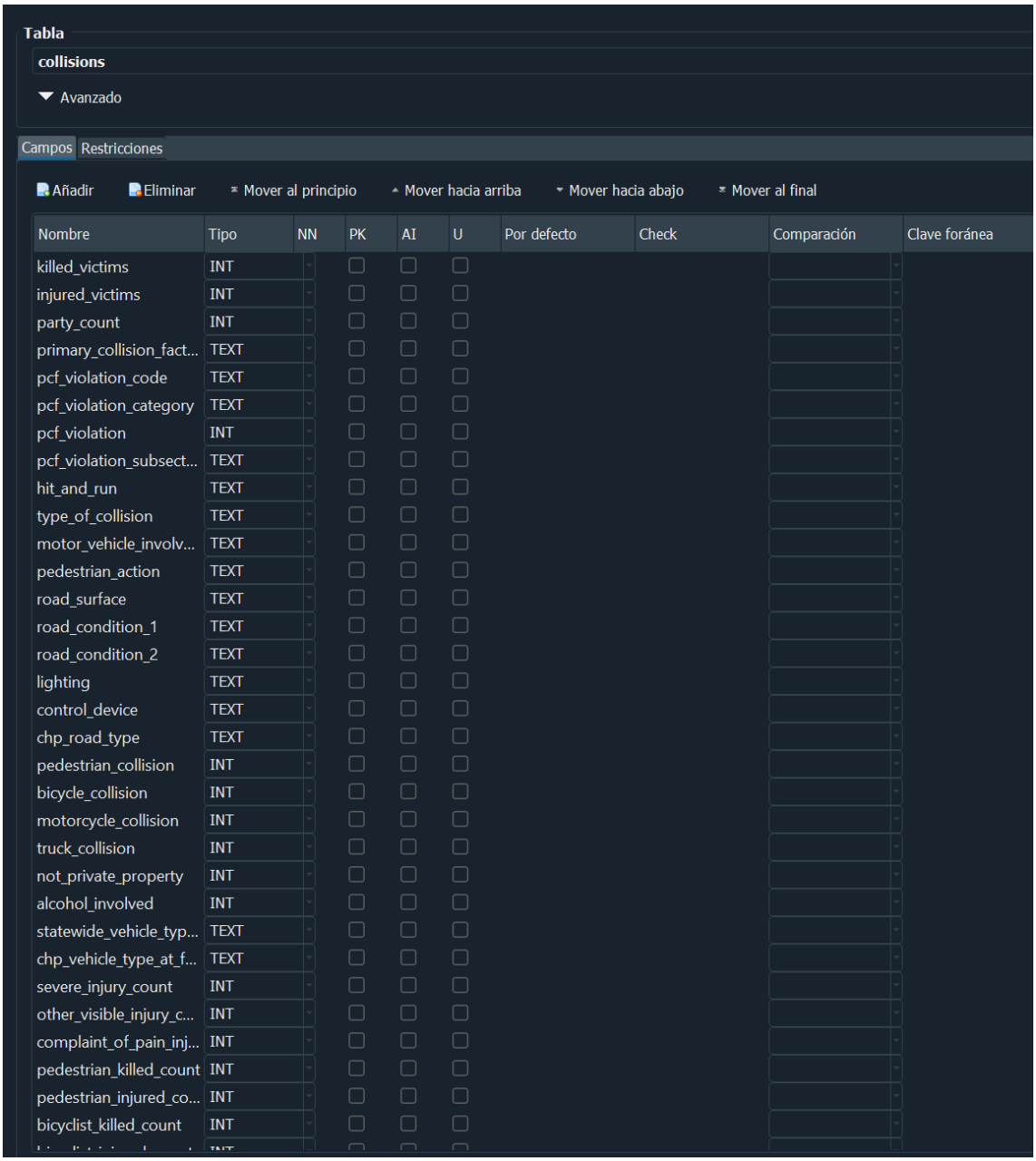
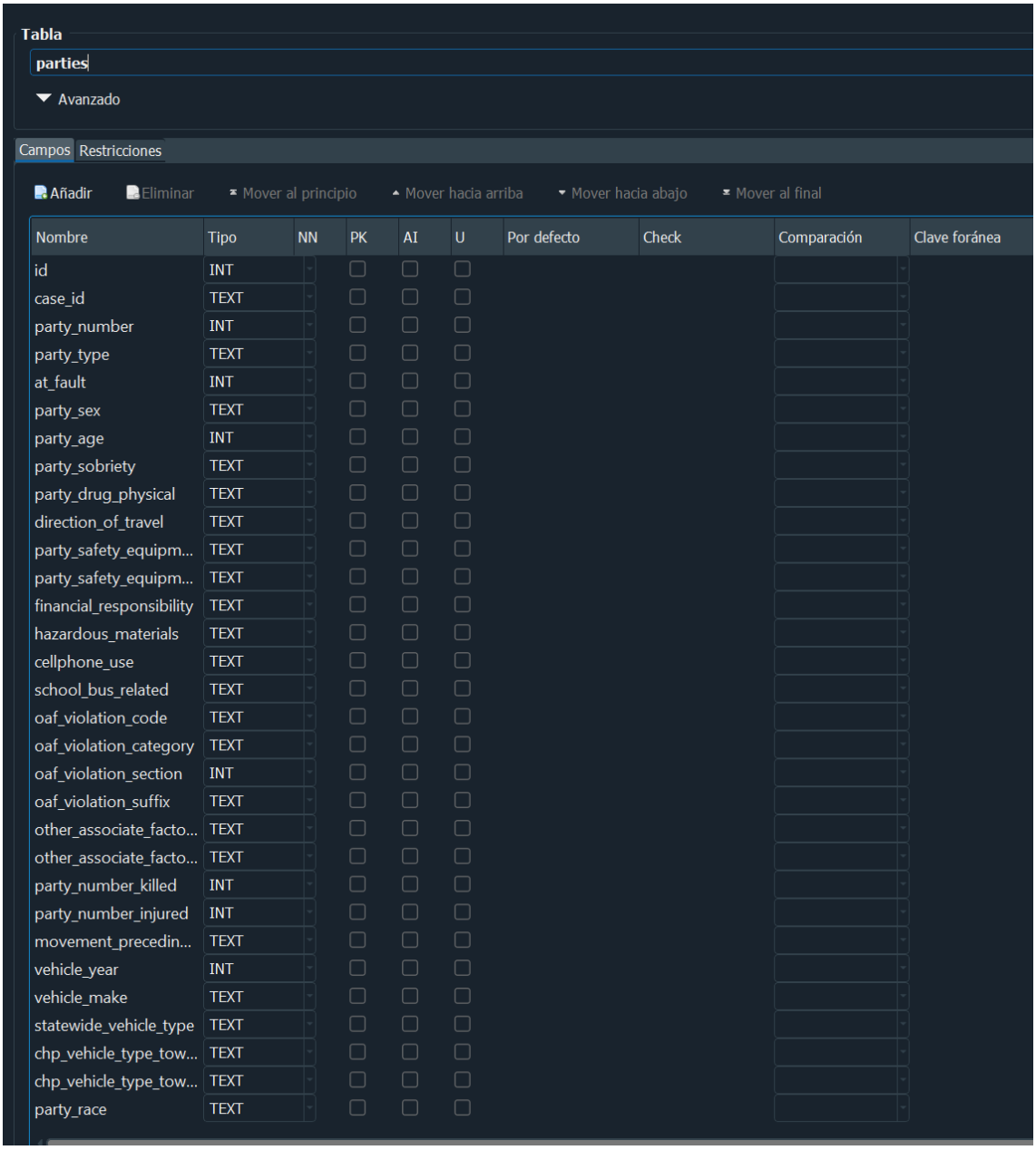



*Ilustración 10. Tabla collisions pt.2*


*Ilustración 11. Tabla collisions pt.3*




| Nombre | Tipo | NN | PK | AI | U | Por defecto | Check | Comparación | Clave foránea |
|---|---|---|---|---|---|---|---|---|---|
| id | INT | | | | | | | | |
| case_id | TEXT | | | | | | | | |
| party_number | INT | | | | | | | | |
| party_type | TEXT | | | | | | | | |
| at_fault | INT | | | | | | | | |
| party_sex | TEXT | | | | | | | | |
| party_age | INT | | | | | | | | |
| party_sobriety | TEXT | | | | | | | | |
| party_drug_physical | TEXT | | | | | | | | |
| direction_of_travel | TEXT | | | | | | | | |
| party_safety_equipm... | TEXT | | | | | | | | |
| party_safety_equipm... | TEXT | | | | | | | | |
| financial_responsibility | TEXT | | | | | | | | |
| hazardous_materials | TEXT | | | | | | | | |
| cellphone_use | TEXT | | | | | | | | |
| school_bus_related | TEXT | | | | | | | | |
| oaf_violation_code | TEXT | | | | | | | | |
| oaf_violation_category | TEXT | | | | | | | | |
| oaf_violation_section | INT | | | | | | | | |
| oaf_violation_suffix | TEXT | | | | | | | | |
| other_associate_facto... | TEXT | | | | | | | | |
| other_associate_facto... | TEXT | | | | | | | | |
| party_number_killed | INT | | | | | | | | |
| party_number_injured | INT | | | | | | | | |
| movement_precedin... | TEXT | | | | | | | | |
| vehicle_year | INT | | | | | | | | |
| vehicle_make | TEXT | | | | | | | | |
| statewide_vehicle_type | TEXT | | | | | | | | |
| chp_vehicle_type_tow... | TEXT | | | | | | | | |
| chp_vehicle_type_tow... | TEXT | | | | | | | | |
| party_race | TEXT | | | | | | | | |

*Ilustración 12. Tabla parties*

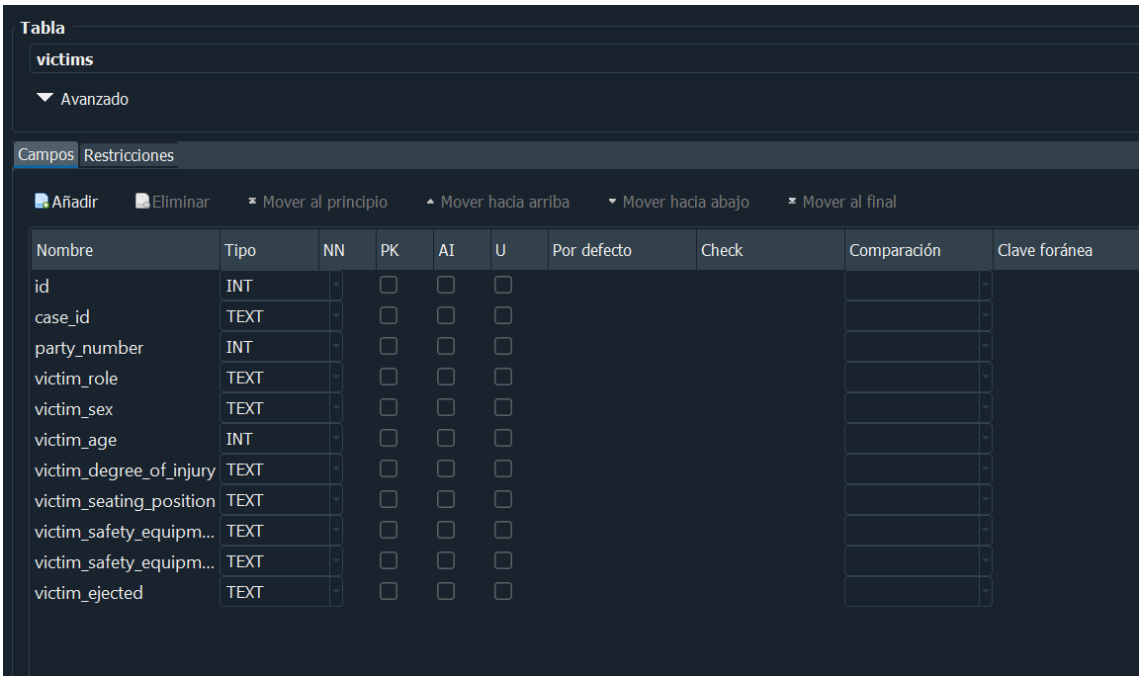



*Ilustración 13. Tabla victims*

Con dicho análisis ha obtenido como objeto de interés de estudio la colisión teniendo muchas posibles dimensiones que caracterizasen cada caso, tales como información sobre las personas implicadas, las víctimas, los vehículos implicados, la condición meteorológica en el momento del accidente, la localización geográfica, el estado de la vía por la cual se circulaba, el momento del percance, entre otros atributos.

Una vez obtenidos los datos, se ha decidido modelarlos creando un diagrama de clases UML. Buscando una jerarquía que permitiese obtener un modelo lo más óptimo posible, se han agrupado las columnas de las tablas de la base de datos original en una clase definida como un componente participante en la colisión de tráfico.

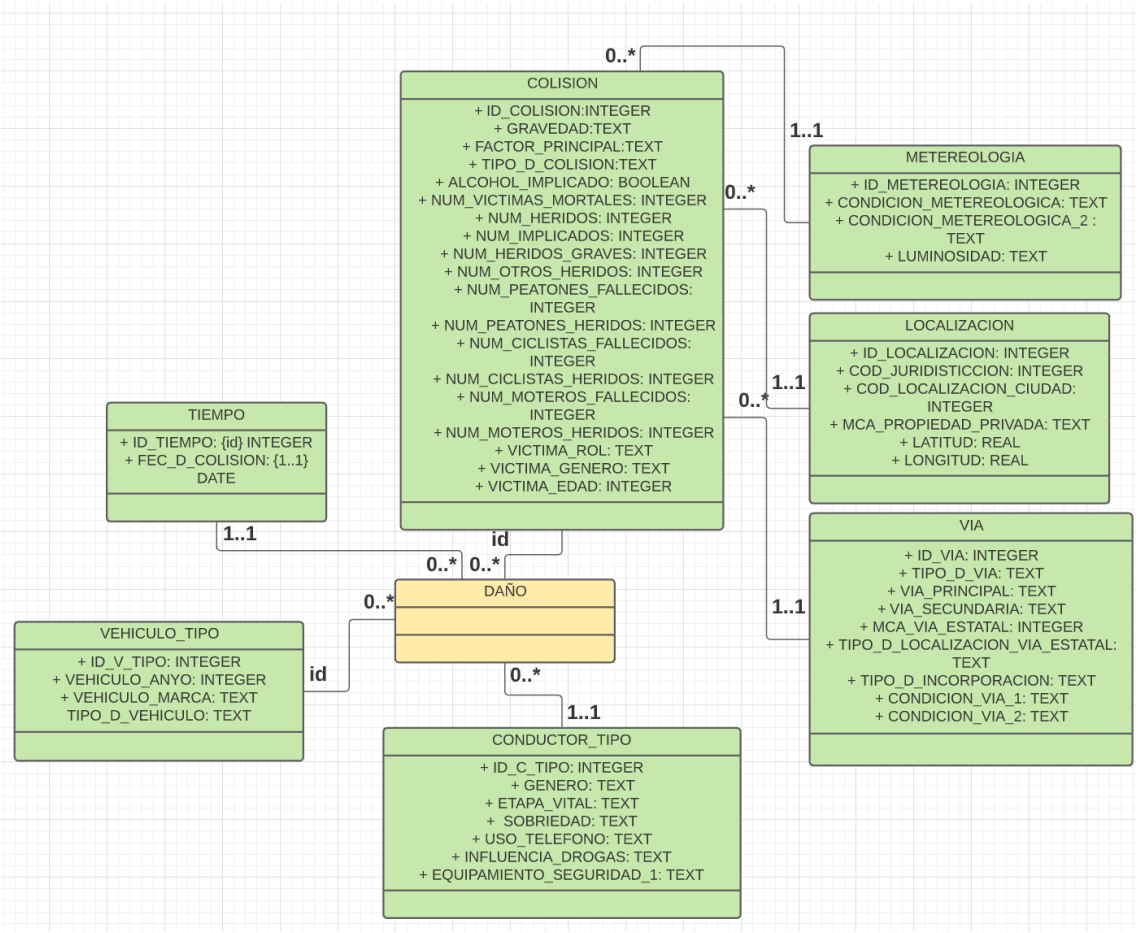



*Ilustración 14. Modelo conceptual multidimensional de las colisiones de tráfico representado en un diagrama de clases UML*

A continuación se describe cada una de las clases definidas:

| Clase | Descripción |
|---|---|
| DAÑO | Hecho que relaciona la información temporal, el tipo de vehículo y tipo de conductor de la colisión. |
| TIEMPO | Dimensión temporal con la fecha en la que se produce la colisión. |
| VEHICULO_TIPO | Dimensión que contiene cada tipo de vehículo que se han visto implicados en una colisión. |
| CONDUCTOR_TIPO | Dimensión que contiene cada casuística del tipo de conductor que se han visto implicados en una colisión. |
| COLISION | Dimensión que contiene la información relativa a la colisión. |
| METEREOLOGIA | Dimensión que contiene la información de la climatología en el momento de una colisión. |
| VIA | Dimensión que contiene la información del tipo vía en la cual se produjo una colisión. |
| LOCALIZACION | Dimensión que contiene la información del punto geográfico donde se produjo una colisión. |

*Tabla 4. Descripción de las clases*

Para cada clase, se va a describir cada uno de sus atributos:

| **TIEMPO** | |
|---|---|
| Atributo | Descripción |
| id_tiempo | Identificador único de cada casuística temporal. |
| Fec_d_colision | Fecha en la que se produjo la colisión. |

*Tabla 5. Descripción de los atributos de la dimensión TIEMPO*

| **VEHICULO_TIPO** | |
|---|---|
| Atributo | Descripción |
| id_v_tipo | Identificador único de cada casuística de tipo de vehículo. |
| vehiculo_anyo | Año del vehículo. |
| vehiculo_marca | Marca del vehículo. |
| tipo_d_vehiculo | Tipo de vehículo motorizado. |

*Tabla 6. Descripción de los atributos de la dimensión VEHICULO_TIPO*

| CONDUCTOR_TIPO | |
|---|---|
| Atributo | Descripción |
| id_c_tipo | Identificador único de cada casuística de tipo de conductor. |
| genero | Género del conductor. |
| etapa_vital | Etapa vital del conductor. |
| sobriedad | Indica si el conductor estaba sobrio en el momento de la colisión. |
| uso_telefono | Indica si el conductor estaba usaba el teléfono en el momento de la colisión. |
| influencia_drogas | Indica si el conductor estaba bajo los efectos de las drogas en el momento de la colisión. |
| equipamiento_seguridad_1 | Equipamiento de seguridad que utilizaba el conductor en el momento de la colisión. |

*Tabla 7. Descripción de los atributos de la dimensión CONDUCTOR_TIPO*

| COLISION | |
|---|---|
| Atributo | Descripción |
| id_colision | Identificador único de cada casuística de una colisión. |
| gravedad | Daños producidos y gravedad de la colisión. |
| factor_principal | Motivo principal por la que se produjo la colisión. |
| tipo_d_colision | Tipo de impacto. |
| alcohol_implicado | Indica si alguno de los implicados estaba bajo los efectos del alcohol. |
| num_victimas_mortales | Número de víctimas mortales implicadas en la colisión. |
| num_heridos | Número de heridos implicados en la colisión. |
| num_implicados | Número de implicados en la colisión. |
| num_heridos_graves | Número de heridos graves implicados en la colisión. |
| num_otros_heridos | Número de heridos implicados en la colisión. |
| num_peatones_fallecidos | Número de peatones fallecidos implicados en la colisión. |
| num_peatones_heridos | Número de peatones heridos implicados en la colisión. |
| num_ciclistas_fallecidos | Número de ciclistas fallecidos implicados en la colisión. |
| num_ciclistas_heridos | Número de ciclistas heridos implicados en la colisión. |
| num_moteros_fallecidos | Número de motoristas fallecidos implicados en la colisión. |
| num_moteros_heridos | Número de motoristas heridos implicados en la colisión. |
| victima_rol | Acción de la víctima implicada |
| victima_genero | Género de la víctima implicada. |
| victima_edad | Edad de la víctima implicada. |

*Tabla 8. Descripción de los atributos de la dimensión COLISION*

| METEREOLOGIA | |
|---|---|
| Atributo | Descripción |
| id_metereologia | Identificador único relacionado con la condición meteorológica de cada casuística de una colisión. |
| condición_metereologica | Condición meteorológica en el momento de la colisión. |
| condición_metereologica_2 | Descripción de la condición meteorológica en el momento de la colisión. |
| luminosidad | Cantidad de luz natural y artificial en el momento de la colisión. |

*Tabla 9. Descripción de los atributos de la dimensión METEREOLOGIA*

| LOCALIZACION | |
|---|---|
| Atributo | Descripción |
| id_localizacion | Identificador único de la localización geográfica de cada casuística de una colisión. |
| cod_juridisticcion | Código de la jurisdicción policial en el estado de California. |
| cod_localizacion_ciudad | Código de la ciudad en el estado de California. |
| mca_propiedad_privada | Marca que indica la colisión afecta a una propiedad privada. |
| latitud | Latitud geográfica. |
| longitud | Longitud geográfica. |

*Tabla 10. Descripción de los atributos de la dimensión LOCALIZACION*

| VIA | |
|---|---|
| Atributo | Descripción |
| id_via | Identificador único de la vía de cada casuística de una colisión. |
| tipo_d_via | Tipo de vía en la que se produjo la colisión. |
| via_principal | Nombre de la vía principal. |
| via_secundaria | Nombre de la vía secundaria. |
| mca_via_estatal | Marca si la colisión se produjo en vía estatal. |
| tipo_d_localizacion_via_estatal | Punto de la vía estatal en el que se produjo la colisión. |
| tipo_d_incoporacion | Indica el tipo de incorporación en caso de haber ocurrido el accidente en dicho punto. |
| condicion_meteo_via | Estado climatológico de la vía. |
| condicion_via_1 | Estado de la vía. |
| condicion_via_2 | Especificación del estado de la vía. |

*Tabla 11. Descripción de los atributos de la dimensión VIA*

En el diagrama se puede observar que, a diferencia de lo esperado, no se ha definido como hecho relevante una colisión. Esto es debido a que en una colisión se pueden ver implicados muchos vehículos y muchas personas lo que no permitía definir el esquema en estrella en el que un hecho se conecta con una instancia de cada una de las dimensiones como mucho. Con el diseño propuesto, se considera como hecho, lo que se ha llamado *Daño* y que representa, el efecto que en una determinada colisión se vio implicado un determinado tipo de vehículo que era conducido por un determinado tipo de conductor. A partir de la colisión, se tiene acceso al tipo de vía en el que se produjo la colisión, al punto donde sucedió y en qué condiciones meteorológicas. Con relación a la dimensión temporal, siempre presente en un almacén de datos, se ha enlazado con la clase *Daño* en lugar de con la clase *Colisión,* para optimizar las consultas. Es obvio que esto introduce un cierto grado de redundancia, pero esta redundancia será controlada en el proceso de introducción de datos.

## 4.2. Diseño lógico

Una vez definido el modelo multidimensional con el esquema copo de nieve, presentando las peculiaridades anteriormente comentadas, se ha transformado el diagrama de clases UML a un modelo relacional:

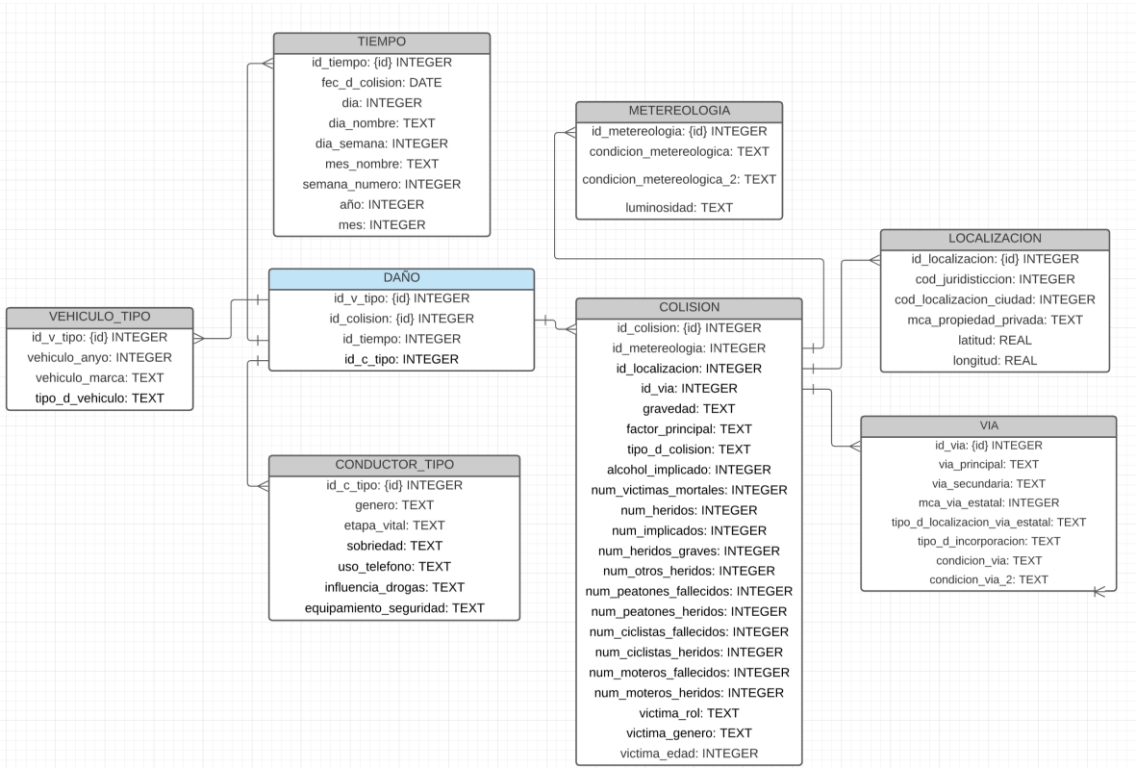



*Ilustración 15. Diagrama relacional*

**Hecho:**

**daño** (
id_tiempo: INTEGER,
id_colision: INTEGER,
id_v_tipo: INTEGER,
id_c_tipo: INTEGER
)
CP: {id_v_tipo, id_colision}
VNN: {id_tiempo, id_c_tipo}
CAj: {id_tiempo} → Tiempo
CAj: {id_colision} → Colision
CAj: {id_v_tipo} → Vehiculo_tipo
CAj: {id_c_tipo} → Conductor_tipo

**Dimensiones:**

**tiempo** (
id_tiempo: INTEGER,
fec_d_colision: DATE
)
CP: {id_tiempo}
VNN: {fec_d_colision}
**vehiculo_tipo** (
id_v_tipo: INTEGER,
vehiculo_anyo: INTEGER,
vehiculo_marca: TEXT,
tipo_d_vehiculo: TEXT
)
CP: {id_v_tipo}
**condutor_tipo** (
id_c_tipo: INTEGER,
genero: TEXT,
etapa_vital: TEXT,
sobriedad: TEXT,
uso_telefono: TEXT,
influencia_drogas: TEXT,
equipamiento_seguridad: TEXT
)
CP: {id_c_tipo}

**colision** (
id_colision: INTEGER,
id_metereologia: INTEGER,
id_localizacion: INTEGER,
id_via: INTEGER,
gravedad: TEXT,
factor_principal: TEXT,
tipo_d_colision: TEXT,
alcohol_implicado: INTEGER,
num_victimas_mortales: INTEGER,
num_heridos: INTEGER,
num_implicados: INTEGER,
num_heridos_graves: INTEGER,
num_otros_heridos: INTEGER,
num_peatones_fallecidos: INTEGER,
num_peatones_heridos: INTEGER,
num_ciclistas_fallecidos: INTEGER,
num_ciclistas_heridos: INTEGER,
num_moteros_fallecidos: INTEGER,
num_moteros_heridos: INTEGER,
victima_rol: TEXT,
victima_genero: TEXT,
victima_edad: INTEGER
)
CP: {id_colision}
CAj: {id_metereologia} → Metereologia
CAj: {id_localizacion} → Localización
CAj: {id_via} → Via

**metereologia** (
id_metereologia: INTEGER,
condicion_metereologica: TEXT,
condicion_metereologica_2: TEXT,
luminosidad: TEXT
)
CP: {id_metereologia}
**localizacion** (
id_localizacion: INTEGER,
cod_juridisticcion: INTEGER,
cod_localizacion_ciudad: INTEGER,
mca_propiedad_privada: TEXT,
latitud: REAL,
longitud: REAL
)
CP: {id_localizacion}
**via** (
id_via: INTEGER,
via_principal: TEXT,
via_secundaria: TEXT,
mca_via_estatal: INTEGER,
tipo_d_localizacion_via_estatal: TEXT,
tipo_d_incorporacion: TEXT,
condicion_via: TEXT,
condicion_via_2: TEXT)
CP: {id_via}

## 4.3. Diseño físico

Tal y como se ha descrito en el diseño lógico, se está utilizando un sistema relacional para representar el modelo multidimensional. En los almacenes de datos es más conocido como Sistemas ROLAP y, con el objetivo de obtener resultados óptimos en las respuestas tras la evaluación de las consultas, permiten el uso de índices generados a partir de las claves primarias.

Se han definido claves primarias para cada una de las dimensiones sobre el identificador de cada casuística presentada. Tal y como hace Oracle, PostgreSQL crea implícitamente un índice sobre la clave ajena que se ha creado. En este sistema el acceso a los datos se realizará mayoritariamente por este campo por lo que no aportaría, en relación con el valor esfuerzo, un análisis con las posibles combinaciones de columnas para nuevos índices.

En la tabla del hecho, la clave primaria se trata de una clave compuesta por los identificadores que la conforman. En ella también se han definido las claves ajenas al resto de dimensiones de las cuales dependen sus atributos. Esto permite, al cargarlo en el Power BI, tener el modelo perfectamente establecido.

Por último, se ha realizado un estudio de la volumetría y el espacio ocupado en memoria por todas aquellas tablas que han participado en el proceso:

| TABLA | TIPO | VOLUMETRÍA (unidades) | ESPACIO EN DISCO (MB) |
|---|---|---|---|
| S_CASO_COLISION | Asociación | 13.652.976 | 577,00 |
| S_CASO_CONDUCTOR | Asociación | 39.879.138 | 1684,00 |
| S_CASO_LOCALIZACION | Asociación | 9.172.565 | 387,00 |
| S_CASO_METEREOLOGIA | Asociación | 9.172.565 | 387,00 |
| S_CASO_TIEMPO | Asociación | 9.172.565 | 387,00 |
| S_CASO_VEHICULO | Asociación | 17.632.271 | 745,00 |
| S_CASO_VIA | Asociación | 9.172.565 | 387,00 |
| COLISION | Dimensión | 12.754.614 | 2.182,00 |
| CONDUCTOR_TIPO | Dimensión | 90.012 | 6,208,00 |
| LOCALIZACION | Dimensión | 2.331.377 | 134,00 |
| METEREOLOGIA | Dimensión | 167 | 0,02 |
| TIEMPO | Dimensión | 7.235 | 0,32 |
| VEHICULO_TIPO | Dimensión | 65.565 | 4,46 |
| VIA | Dimensión | 2.826.824 | 340,00 |
| DAÑO | Hecho | 69.488.295 | 3.992,00 |

*Tabla 12. Volumetría y ocupación de memoria de las tablas del esquema relacional*

## 4.4. ETL

### 4.4.1. Extracción

Con los datos obtenidos tras su selección, se ha realizado una carga total de la base de datos en el software SQLite debido a que la extensión del archivo que contiene la información es .sqlite (*switrs.sqlite*).

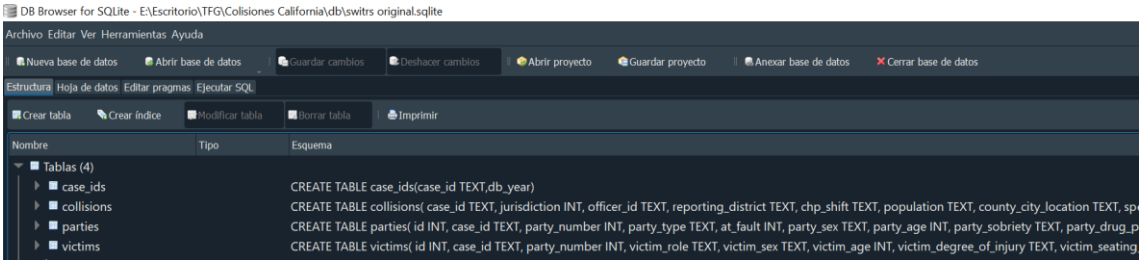



*Ilustración 16. Base de datos original en SQLite*

### 4.4.2. Transformación

Para poder conocer el significado de cada una de las columnas integrantes de las tablas, la fuente original *Statewide Integrated Traffic Records System* tiene documentación pública que permite entender la razón de cada valor por atributo. Con ello ha sido posible transformar la información codificada por su significado real, más entendible y amigable para el usuario.

**SWITRS Codebook**

**SWITRS Collision Raw Data**

| Item Name | Variable Name | Description | Label | Possible Values |
|---|---|---|---|---|
| Case Id | **CASEID** | the unique identifier of the collision report (barcode beginning 2002; 19 digit code prior to 2002) | | |
| X-Coordinate Location | **POINT_X** | The longitude of the geocoded location; uses the World Geodetic System from 1984 (WGS84). | | |
| Y-Coordinate Location | **POINT_Y** | The latitude of the geocoded location; uses the World Geodetic System from 1984 (WGS84). | | |
| Collision Year | **YEAR_** | the year when the collision occurred | | |
| County City Location | **LOCATION** | the location code of where the collision occurred | | Data may appear with no leading zero. |
| CHP Beat Type | **CHPTYPE** | | **0 "Not CHP"<br>1 "Interstate"<br>2 "US Highway"<br>3 "State Route"<br>4 "County Road Line"<br>5 "County Road Area"<br>6 "US Highway"<br>7 "State Route"<br>8 "County Road Line"<br>9 "County Road Area"<br>10 "Safety Services Program Beats"<br>11 "Administrative Beats (900's)"** | 1 - Interstate<br>2 - US Highway<br>3 - State Route<br>4 - County Road Line<br>5 - County Road Area<br>A - Safety Services Program Beats<br>S - Administrative Beats (900's)<br>0 - Not CHP<br>Contract City:<br>6 - US Highway<br>7 - State Route<br>8 - County Road Line<br>9 - County Road Area |
| Day of Week | **DAYWEEK** | the code for the day of the week when the collision occurred | | 1 - Monday<br>2 - Tuesday<br>3 - Wednesday<br>4 - Thursday<br>5 - Friday<br>6 - Saturday<br>7 - Sunday |
| Collision Severity | **CRASHSEV** | the injury level severity of the collision | | 1 - Fatal<br>2 - Injury (Severe) |

1

*Ilustración 17. Documentación oficial de la base de datos Statewide Integrated Traffic Records System*

Para la creación de tablas intermedias que permitiesen hacer de puente al modelo diseñado, se han reconstruido los identificadores de caso ya que originalmente eran campos de tipo *TEXT* haciendo que las concatenaciones fuesen más difíciles, así como, ser una mala práctica de programación tener campos por los cuales se van a concatenar tablas como no numéricos. Además, se deben crear índices sobre ellos siendo más eficiente si el campo es un *INTEGER*. Como solución a dicho inconveniente, se ha utilizado el *rowid* para sustituir el código original por el nuevo utilizando un campo número auto incremental en la nueva tabla de identificadores de caso.

Se han utilizado instrucciones *CREATE TABLE nombre_tabla [(columna_1 tipo,…)]* para poder crear las tablas deseadas y se ha insertado la información mediante sentencias *INSERT INTO nombre_tabla* concatenando las tablas originales necesarias para obtener el subconjunto de columnas deseado.

Una vez obtenidas las tablas intermedias, se ha decidido utilizar otro software de base de datos relacional, PostgreSQL, ya que se ha encontrado muchos problemas de rendimiento con el uso de SQLite. Se han exportado cada una de las tablas a ficheros .csv para poder importarlas dentro de nuestro nuevo servidor. El procedimiento de carga de esta información consiste en crear previamente la tabla e importar el fichero, teniendo especial atención que las columnas deben coincidir y el tipo de datos debe ser el mismo.

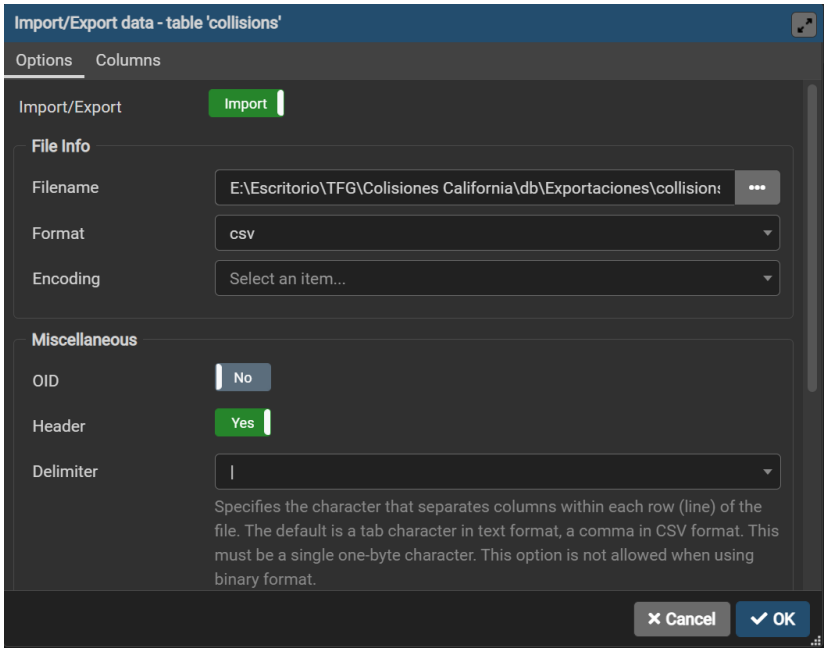



*Ilustración 18. Proceso de carga por fichero en PostgreSQL*

A partir de las tablas intermedias, se ha realizado una agrupación por cada casuística presentada de cara a la construcción de las dimensiones del modelo

relacional anteriormente comentado. Cabe realizar una mención especial al tratado de los valores nulos en los registros. En caso de no controlarlos, al realizar la concatenación de tablas por dichos valores de las columnas no coincidirán y, por tanto, perderemos casos de colisiones y las relaciones con el resto de los atributos.

En base a generar un nuevo identificador único por caso se han utilizado, junto al tratamiento de los nulos, sentencias *GROUP BY* y secuencias que generasen dicho identificador. Éste, será insertado en la tabla de hechos junto al resto de identificadores del resto de dimensiones. Sin embargo, para no perder la trazabilidad del Daño (*tabla de hechos que relaciona las dimensiones de la colisión*) se han creado tablas de asociación que vinculan el nuevo identificador a los originales. Se ha obtenido un total de siete tablas de asociación: caso – vehículo, caso – conductor, caso – tiempo, caso – colisión, caso – localización, caso – meteorología, caso – vía.

Con la ayuda de cuatro de las siete las tablas de asociación, se ha realizado una concatenación aquellas que participan directamente en la relación con el Daño insertando en la tabla del Hecho los nuevos códigos cruzándolos por las relaciones del caso originales.

### 4.4.3. Transporte

Una vez realizada la implementación del almacén de datos, con las fases que conlleva, se ha realizado la carga en Power BI. Gracias a la integración de este software con muchos sistemas de almacenamiento de bases de datos, la integración con PostgreSQL es muy sencilla.

Se ha indicado únicamente el servidor que contiene las tablas creadas en el sistema relacional, el nombre de la base de datos y el usuario propietario de ésta junto con su contraseña.

En este caso, debido a que se ha diseñado e implementado en un entorno local se ha indicado como servidor *localhost,* como nombre de la base de datos *postgres*, usuario propietario *postgres* y contraseña la definida. Se ilustra el proceso a seguir:

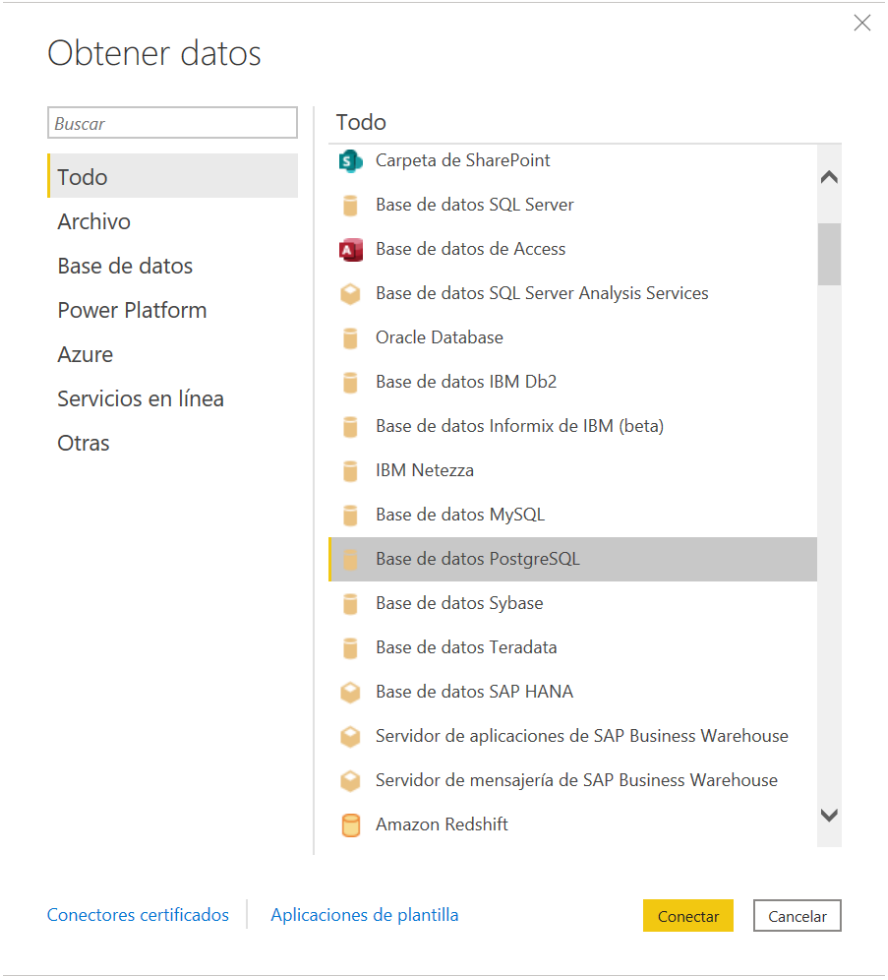



*Ilustración 19. Seleccionar origen de los datos*

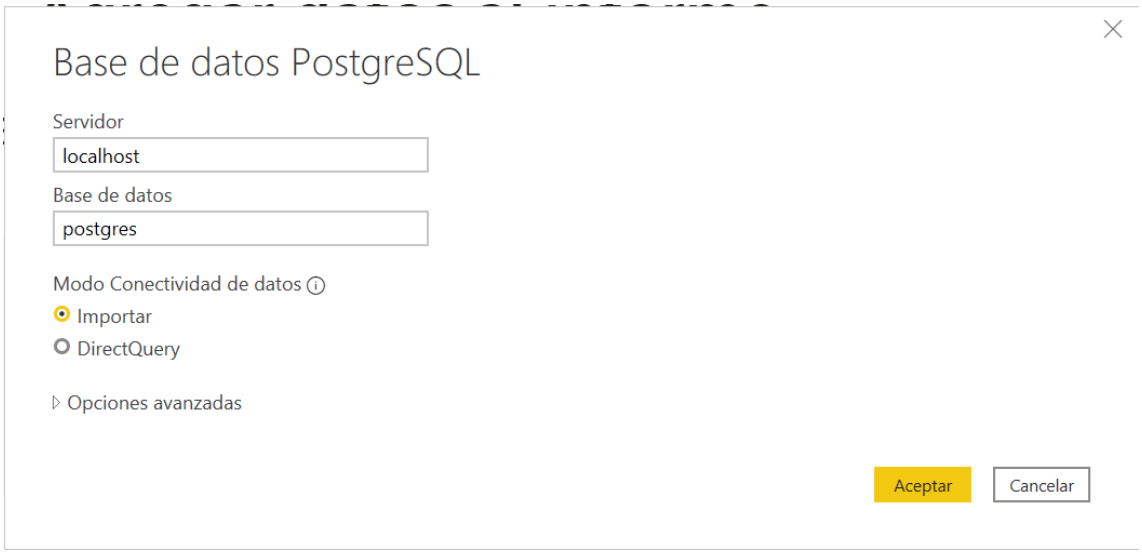



*Ilustración 20. Indicar servidor y base de datos a la cual realizar la conexión*

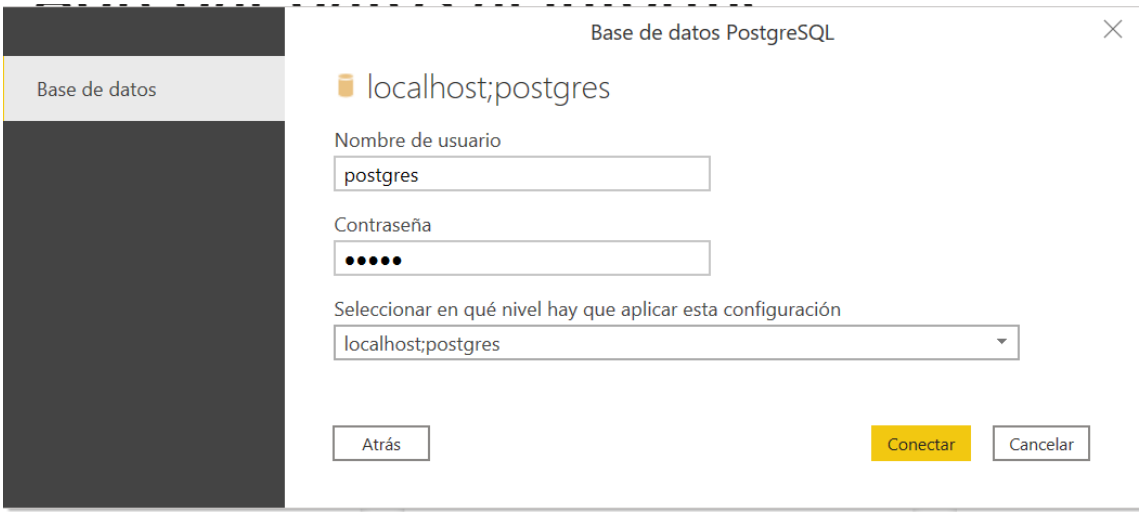



*Ilustración 21. Indicar credenciales*

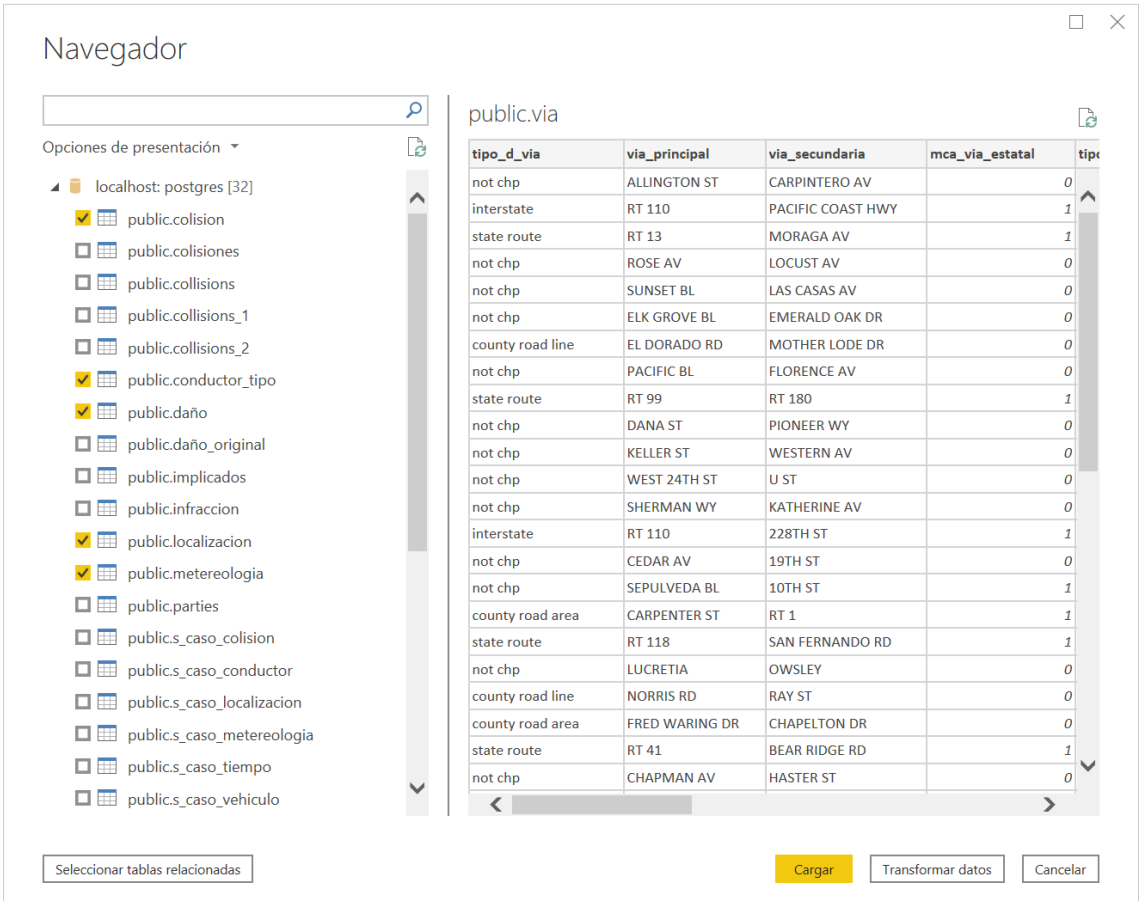



*Ilustración 22. Seleccionar tablas a importar a Power BI*

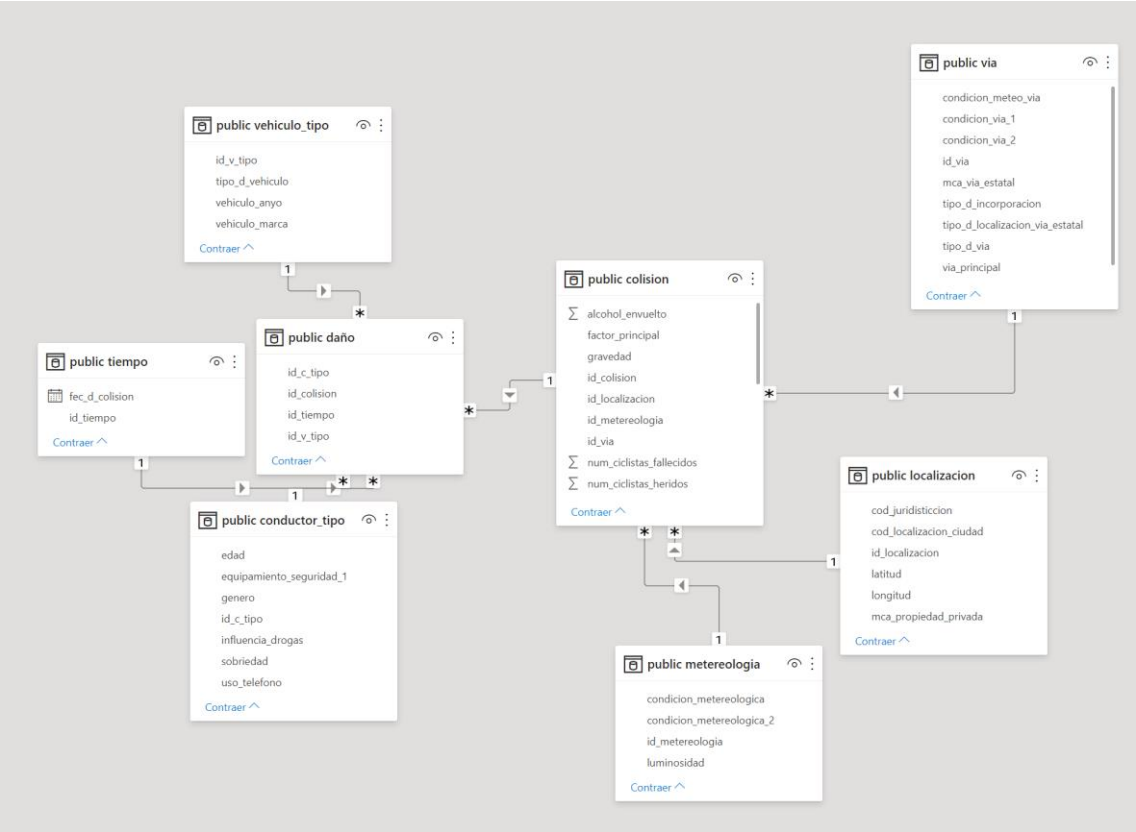



*Ilustración 23. Modelo importado en Power BI*

# 5. Implantación

## 5.1. Creación en Power BI

Una vez realizada la carga en Power BI, se ha estudiado la posibilidad de crear nuevas medidas en las dimensiones con el fin de poder crear jerarquías. Éstas, están definidas sobre los atributos de una dimensión con el fin de poder navegar a través de ellas en la generación de los informes, pudiendo cambiar el nivel de agregación en el cual se presentan los datos.

En esta navegación, encontramos dos operaciones: agregación (*ROLL*) y disgregación (*DRILL*). La agregación permite sustituir el criterio de agrupación por uno de mayor granularidad. La disgregación permite reemplazar el criterio de agrupación por uno de menor granularidad.

Por ello, se han definido para la dimensión Tiempo nuevas medidas que nos servirán para definir jerarquías sobre ellas. A partir de la fecha de colisión, se han generado columnas nuevas indicando el día, el nombre del día, el número de día dentro de una semana, el mes, el nombre del mes, el número de la semana en el año y el año. Todas estas medidas se han obtenido gracias a funciones propias de Power BI sobre el tipo de datos DATE.

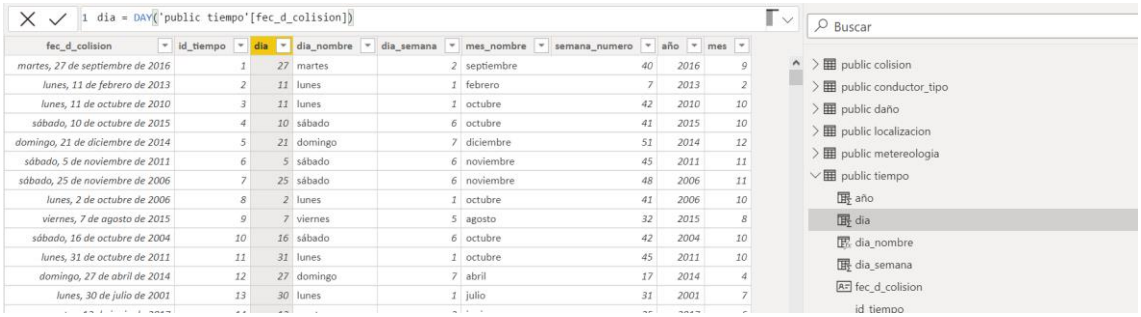



*Ilustración 24. Función de obtención del día del mes sobre un campo DATE*

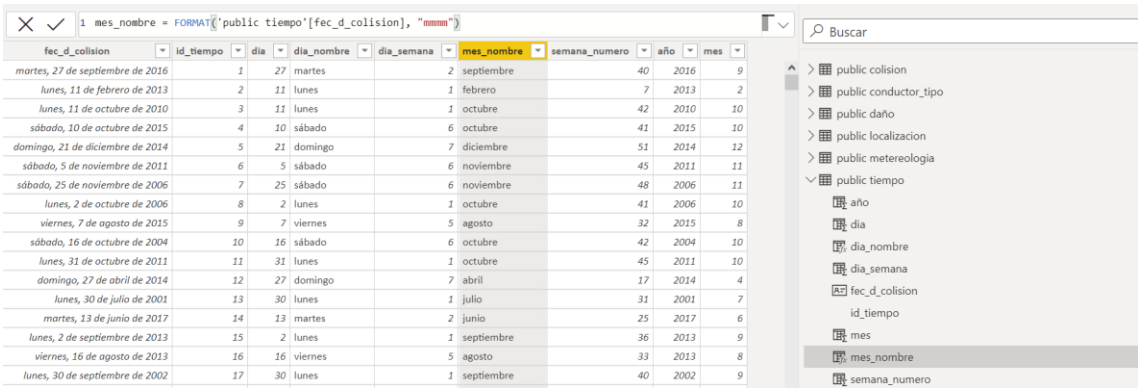



*Ilustración 25. Función de obtención del nombre del mes a partir de un campos DATE*

Con estas nuevas medidas, se ha definido una jerarquía que agrupa año, el nombre del mes y el día. Está definida desde un nivel de granularidad mayor (año) hasta un grano mucho más fino (día).

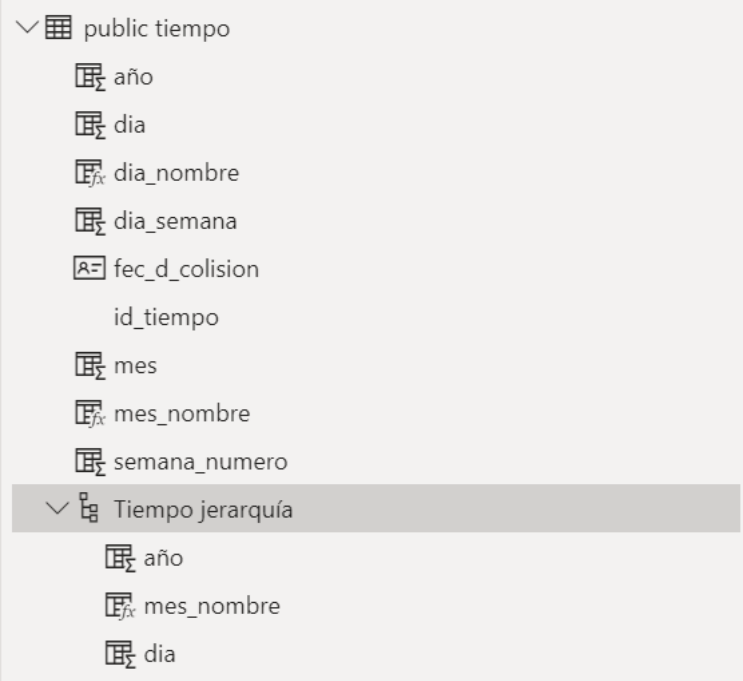



*Ilustración 26. Jerarquía en la dimensión Tiempo*

## 5.2. Diseño de informes

Para el diseño de informes dentro de la herramienta, se han generado mediante el uso de los elementos disponibles en la interfaz gráfica de usuario. Cabe destacar que también sería posible retocarlos y modificar las cargas y transformación de datos modificando los scripts que genera el software en el lenguaje de programación M.

El primer informe que se ha obtenido se trata de un gráfico circular con la cantidad de víctimas mortales por accidente de tráfico. Teniendo esto en mente, se ha indicado que los valores debe obtenerlos de la columna *NUM_VICTIMAS_MORTALES* de la tabla de dimensiones *COLISION*. Como ayuda para el lector de este gráfico, se ha incluido una leyenda con los valores participantes.

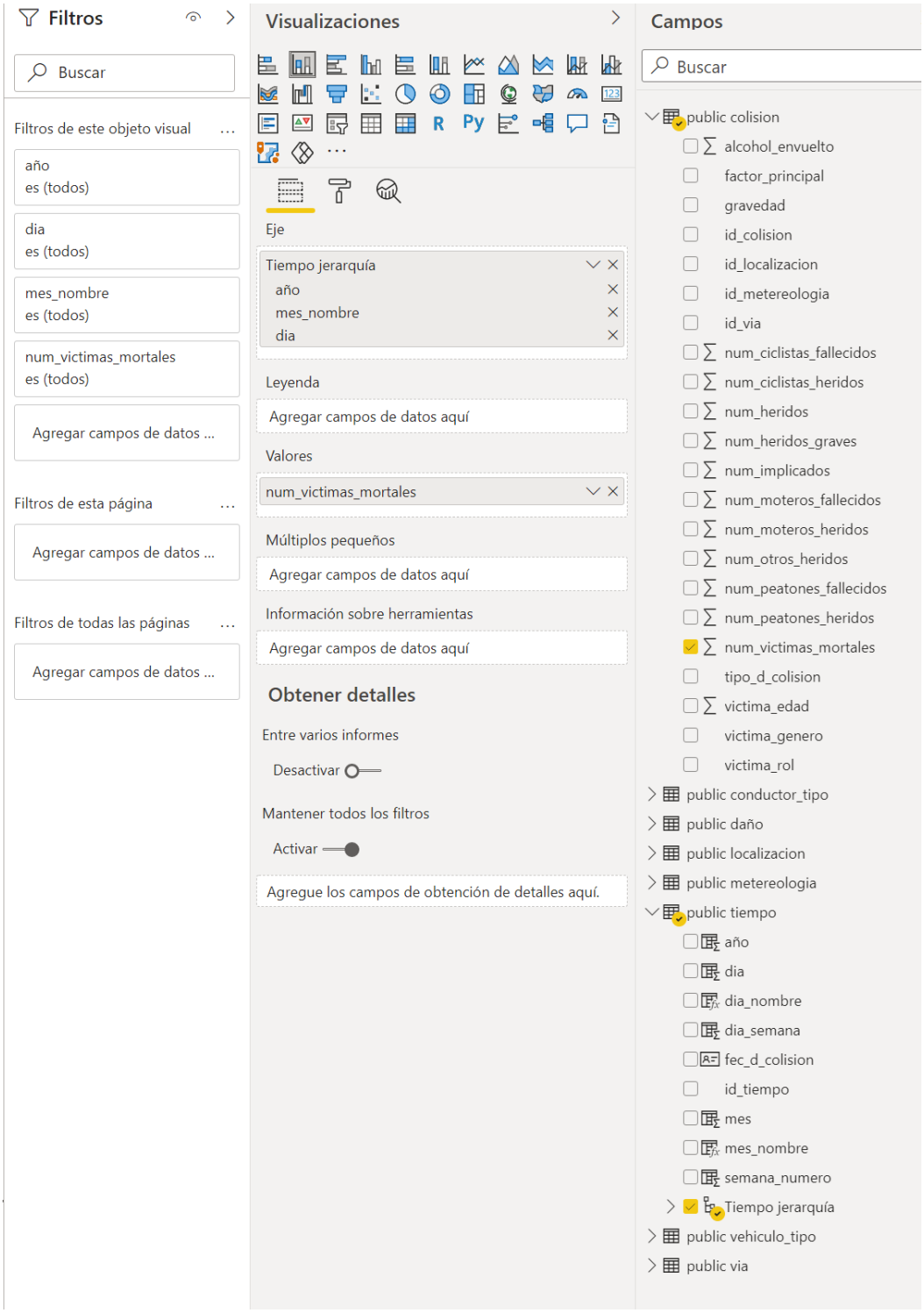



*Ilustración 27. Configuración del informe*

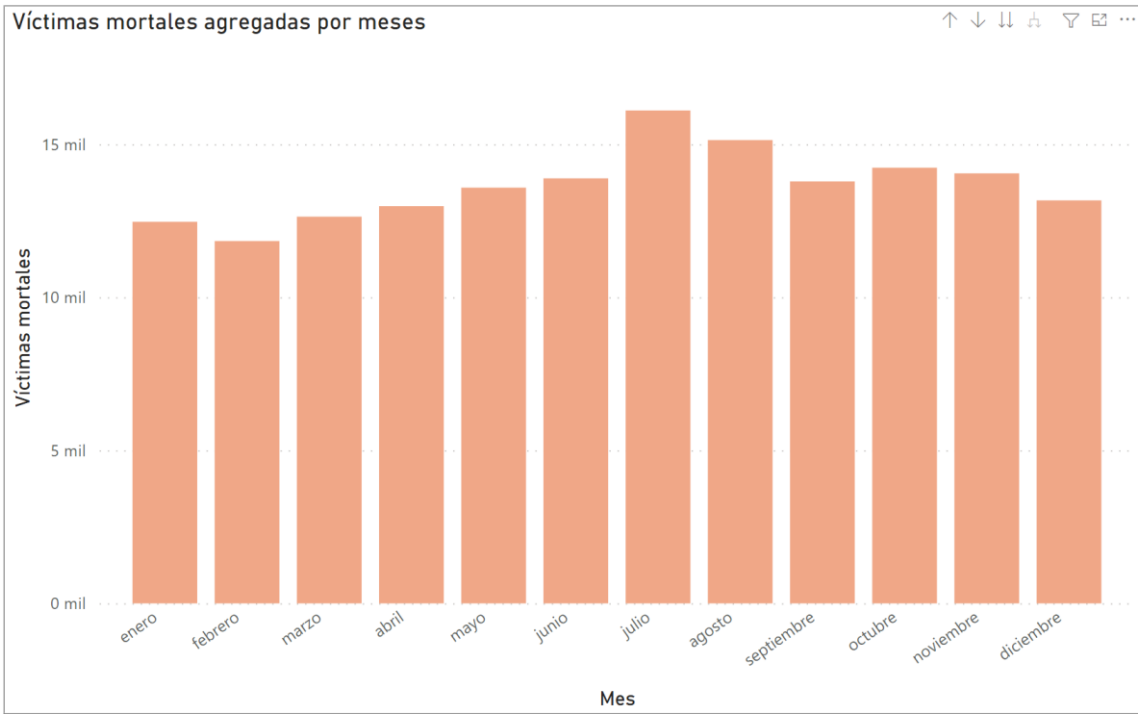



*Ilustración 28. Gráfico de barras con el número total de víctimas mortales agregado por meses*

Tal y como podemos observar, la mayor tendencia en las colisiones de tráfico es al fallecimiento de una persona o personas en los meses de verano. Sin embargo, el resto de los meses del año no se produce una gran disminución de las muertes por accidente debido a que el clima del estado de California varía entre árido y subártico dependiendo de su proximidad al norte del país. Esto produce que en sur encontremos veranos secos e inviernos frescos, mientras que en el norte se puede encontrar inviernos fríos con grandes nevadas y veranos suaves.

Un segundo informe generado es la cantidad de vehículos por marca involucrados en las colisiones. En esta ocasión se ha seleccionado un diagrama de barras para poder representar de manera más visual cuáles son las marcas comerciales de vehículos motorizados más accidentadas en el estado de california en los últimos 20 años.

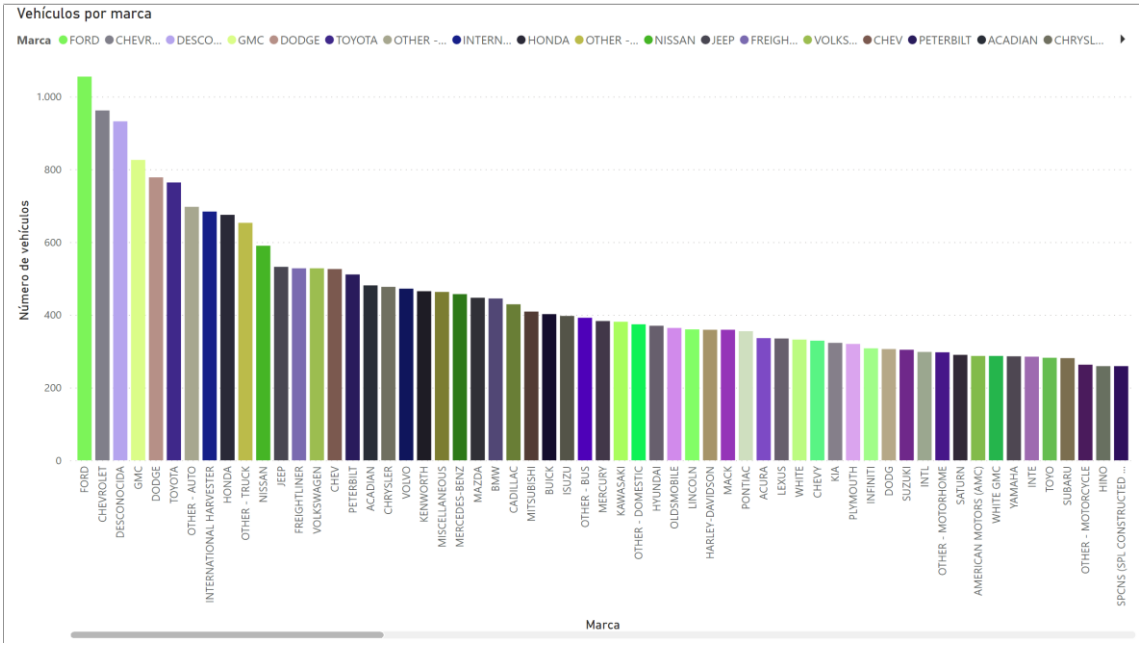



*Ilustración 29. Número de vehículos por marca involucrados*

Se puede observar una tendencia clara sobre la nacionalidad del fabricante de los automóviles. Las cuatro firmas automovilísticas con mayor participación son todas de origen estadounidense: Ford, Chevrolet, General Motors Company y Dodge. Tras el dominio norteamericano, se encuentran tres marcas japonesas mundialmente reconocidas y extendidas: Toyota, Honda y Nissan.

# 6. **Conclusiones**

En este apartado se va a valorar si los objetivos definidos al inicio de esta memoria han sido alcanzados.

Se ha conseguido diseñar e implementar un almacén de datos basado en la información de las colisiones de tráfico en el estado de California utilizando software gratuito y sin coste alguno. Tras modelar el problema se ha optimizado el acceso a los registros siendo prácticamente instantáneo el acceso a estos. Sin embargo, no se han optimizado aquellas consultas en las cuales se busca por algún otro atributo de la dimensión que no sea la clave primaria.

La extracción, transformación y carga de los datos se ha completado exitosamente, al igual que la generación de informes estadísticos sobre los datos en Power BI. La generación de estos no requiere de tiempos de espera.

Por último, falta por consensuar cómo se va a llevar a cabo el uso de este problema para la docencia debido a su complejidad de uso dentro del software de análisis. La extracción se realizará mediante ficheros .csv para su carga en la herramienta utilizada en la *Universitat Politècnica de València,* Oracle Developer.

A nivel personal, me ha permitido experimentar desde el inicio el proceso de creación de un almacén de datos. Me ha enseñado la importancia que tiene conocer al detalle los datos sobre los cuales se va a trabajar y el hecho de tener un modelo multidimensional diseñado correctamente. Además, he podido conocer la herramienta Power BI y generar informes con mis propios datos. La cantidad de funcionalidades que tiene puede resultar difícil de utilizar pero sigo considerando que es un software muy potente.

# 7. **Referencias**

Kimball, R., Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Indianapolis, Estados Unidos: John Wiley & Sons, Inc.

Laudon, K. C., Laudon, J. P. (2012). *Management Information Systems*. Nueva York, Estados Unidos: Pearson Education, Inc.

Maurya, P. (2021). *Snowflake Schema in Data Warehouse Model*. Noida, Uttar Pradesh: GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/snowflake-schema-in-data-warehouse-model/

Casamayor, J. (2020 - 2021). *Sistemas de Información Estratégicos Parte I: Almacenes de Datos Tema 4: Diseño de Almacenes de Datos (maestría)*. Universitat Politècnica de València, Valencia

Árbol B+, Adaptado de Adhikary, S. (2020). *Introduction of B+ Tree*. Noida, Uttar Pradesh: GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/introduction-of-b-tree/

Iseminger, D. (2021). *What is Power BI Desktop?* Microsoft Docs. Recuperado de https://docs.microsoft.com/es-es/power-bi/fundamentals/desktop-what-is-desktop