Construcción y explotación de un almacén de datos para el análisis de información sobre tránsito de vehículos

Memoria del proyecto

Autor: Fernando Mora Pérez
Grado de Ingeniería Informática
Trabajo Fin de Grado

Consultor: Carles Llorach Rius

Fecha de entrega: 6 de enero de 2014

1. RESUMEN Y PALABRAS CLAVE

Como punto de partida de esta memoria se realiza un breve resumen de la misma, así como del proyecto llevado a cabo y la razón del mismo. Además, se expone el conjunto de palabras claves empleadas en este documento.

1.1 Resumen

Esta memoria recoge toda la información relativa al proceso de análisis, diseño e implementación del sistema realizado, el cual está formado por un almacén de datos (Data Warehouse), así como, un conjunto de informes que posibilitan el análisis de información sobre el tránsito de vehículos en Cataluña, durante los años del 2007 al 2012.

Se trata por tanto, de un proyecto enmarcado dentro del área del Data Warehousing. Esta área a su vez, forma parte del área del conocimiento denominada Business Intelligence (inteligencia empresarial, inteligencia de negocios o BI), definida como el conjunto de técnicas y tecnologías que permiten a una empresa recopilar y analizar información con el objetivo de facilitar la definición de estrategias y la toma de decisiones.

En este tipo de proyectos se parte de un conjunto de datos pertenecientes a diferentes bases de datos, los cuales son registrados en el almacén de datos por medio del llamado proceso ETL. Este proceso permite Extraer, Transformar y Cargar (Load en inglés) los datos origen en el almacén. Por ello, en esta memoria también se detalla cómo se ha diseñado e implementado este proceso.

En esta memoria también se incluyen capturas de pantalla de la ejecución de cada uno de los informes contemplados. Lo cual permite obtener una idea de las funcionalidades ofrecidas por el sistema.

Para finalizar, indicar que la realización de este proyecto se corresponde con el Trabajo Fin de Grado (TFG) de los estudios de Grado de Ingeniería de Informática, dentro de la asignatura de Data warehousing. Este trabajo es la conclusión de unos estudios y consiste en la construcción de un sistema informático que solvente un problema planteado, de manera semejante a lo que podríamos encontrarnos en una situación real. Para conseguirlo será necesario emplear los conocimientos y destrezas adquiridos en las diferentes asignaturas cursadas en la carrera. Se trata, por tanto, de un trabajo práctico y de síntesis.

1.2 Abstract

This memorandum contains all the information related to the process of analysis, design and implementation of the system developed, which consists of a Data Warehouse and a set of reports that allow the analysis of information about Cataluña vehicles traffic, during the years from 2007 to 2012.

It is therefore, a project framed within the Data Warehousing area. This area is part of the knowledge area called Business Intelligence, defined as the set of techniques and technologies that allow a company to collect and analyze information in order to facilitate the definition strategies and decision making.

In this type of project is started with a set of data from different databases, which are recorded in the data warehouse through the ETL process. This process allows Extract and Transform data from the origin databases and Load it in the data store. Therefore, this memorandum also details how it was designed and implemented this process.

In this memorandum are also included running screenshots of each of the reports implemented. This allows getting an idea about the functionality offered by the system.

Finally, indicate that this project is corresponded with Degree Project of Computer Science Degree, belonging the subject of Data warehousing. This work is the studies conclusion and involves the construction of a computer system that solves a proposed problem, similarly to what we might find in a actual situation. To achieve this, it is needed use knowledge and skills acquired in the different subjects studied in the career. It is therefore a practical and synthesis work.

1.3 Palabras clave

Almacén de datos, Data Warehouse, Data Warehousing, Business Intelligence, ETL, base de datos multidimensional, cubo multidimensional, Data Marts, OLAP, OLTP, dimensión, atributo, hecho, indicador, jerarquía, dimensión junk, minería de datos, ROLAP, MOLAP, HOLAP, Oracle SQL* Loader, SQLPlus, pl/sql.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS.

2. INTRODUCIÓN.

En esta sección se detallan los pasos preparatorios a la realización del proyecto. Para ello, se razona o justifica la realización del mismo, se exponen los objetivos que se pretenden conseguir, se explica el enfoque y método seguido para la conclusión de cada una de sus fases y se concluye con la planificación del proyecto.

2.1 Justificación del proyecto

La realización de este proyecto surge a petición de la Fundación de Estudios para la Conducción RESponsable (FECRES). Como su nombre indica, es una fundación encargada de realizar diferentes estudios encaminados a conseguir una conducción más responsable y una disminución de los accidentes en carretera.

FECRES detectó que el número de desplazamientos en vehículo a motor había continuado aumentando en el año 2012, por lo que, decidido profundizar en esta evolución y comprobar posibles correlaciones entre medios de locomoción, perfiles de conductores y algunas variables de seguridad vial. Para ello, solicitó a IDESCAT información sobre municipios y vehículos, a la DGT los censos de los conductores y al Servei Català de Trànsit datos relativos a radares fijos.

Tanto la información de los municipios y vehículos, como la información sobre censos de conductores incluyen datos de varios años. El conjunto de datos está organizado en varios ficheros con distintos formatos (csv, excel y txt).

FECRES nos suministró toda esta información con la intención de que creáramos un almacén de datos, así como un conjunto de informes que les permitan obtener la información que necesitan analizar.

2.2 Objetivos del proyecto

Como se ha comentado en el resumen inicial, este proyecto se corresponde con el TFG. De esta forma, podemos diferenciar dos tipos de objetivos distintos, los objetivos propios de la realización del TFG y los objetivos que se pretende conseguir por medio del sistema informático a desarrollar.

2.2.1 Objetivos del TFG

La realización de TFG tiene como objetivo poner en práctica los conocimientos y destrezas adquiridas en las diferentes asignaturas cursadas, mediante de la realización de un proyecto informático concreto. Además, como el TFG está enmarcado dentro de la asignatura Data Warehousing, con la realización del TFG se pretende ampliar los conocimientos adquiridos en las asignaturas de bases de datos, profundizando en las características propias de las bases de datos que dan soporte en la toma de decisiones empresariales.

Así pues, mediante la realización del TFG se demuestran las siguientes capacidades:

- Saber analizar un determinado problema y proponer diferentes alternativas tecnológicas para solventarlo.
- Comprender que es un almacén de datos y en qué circunstancias en ventajoso frente a una base de datos operacional.
- Saber diseñar una base de datos multidimensional, así como, implementarla siguiendo dicho diseño.
- Poder reconocer y solventar los problemas que surgen en la integración, transformación y carga de datos, al crear un almacén de datos a partir de múltiples fuentes.
- Conocer herramientas que faciliten la explotación de un almacén de datos.

2.2.2 Objetivos del proyecto

El objetivo principal del proyecto consiste en la construcción de un sistema informático que permita a FECRES comprobar la evolución del tráfico de vehículos a motor en Cataluña, así como, si existe alguna correlación entre medios de locomoción, perfiles de conductores y algunas variables de seguridad vial.

Para ello, hay que tener en cuenta los siguientes objetivos intermedios:

Creación de un almacén de datos formado a partir de la integración de los datos de varias bases de datos. Estos datos están registrados en ficheros con distinto formato.

Crear formularios sencillos para realizar consultas a partir de los criterios de selección: comarca/provincia, tipo de vehículo, tipo de permiso de conducción. Las consultas a realizar son:

- Total de vehículos
- Total de conductores
- % de vehículos respecto población
- Densidad de población (habitantes/km2) y densidad de tráfico (vehículos/km2)
- Número de vehículos / Número de radares

- % de conductores por radar
- Indicador de conductores vs habitantes por género
- Indicador de radares vs vehículos
- Ratio de vehículos x conductor
- Cantidad de vehículos / superficie del territorio

2.3 Enfoque y método seguido

El desarrollo de proyecto informático suele estar dividido en varias fases o etapas, durante las cuales se generan uno o varios entregables. De esta manera el cliente pueda comprobar, por un lado que se han entendido correctamente sus necesidades, así como, que estas están siendo cubiertas según avanza la realización del proyecto. Y por otro lado, que el grado de avance del proyecto se adecúa a lo indicado en la planificación del mismo. En el caso que nos ocupa, la realización de un TFG, si bien no tenemos un cliente si existen una serie de entregas que realizar con la consiguiente comprobación del trabajo realizado.

La realización de los distintos trabajos comprendidos en cada fase de desarrollo de este proyecto, ha estado enfocada a cumplir con los plazos y especificaciones indicados en el plan docente de la asignatura. La especificación de cada fase, junto con su conjunto de tareas a realizar y la duración de las mismas se detalla en el apartado Planificación del proyecto.

En todo proyecto informático, existe una serie de tareas básicas a realizar, como por ejemplo la planificación. Además, pueden existir tareas propias del área temática a la que pertenecen, en nuestro caso Data Warehousing. Entre las tareas propias de los proyectos de Data Warehousing podríamos destacar la definición del cubo multidimensional. Para este último tipo de tareas se ha empleado como referencia la metodología "*HEFESTO: Metodología propia para la Construcción de un Data Warehouse*" de Bernabéu Ricardo Darío.

2.4 Planificación del proyecto

En esta sección se expone la planificación del proyecto en la cual se definen las fases del mismo, los hitos del proyecto y una relación detallada de las distintas fases del proyecto junto con el desglose de sus tareas. Para finalizar, se muestra un diagrama de Gantt.

2.4.1 Fases del proyecto.

La realización del proyecto comprende cuatro fases diferenciadas:

- Plan de trabajo. El objetivo principal de esta fase es la elaboración de un plan de trabajo que comprenda la planificación del mismo y el análisis de los riesgos que puedan comprometer el cumplimiento de dicha planificación. Para poder concretar estas dos tareas principales, es necesario determinar el objetivo del proyecto, y hacer un análisis previo de los requerimientos y de las distintas fuentes de datos origen.
- Análisis y Diseño. En esta fase se realiza un análisis pormenorizado de los diferentes requerimientos funcionales y no funcionales que debe cumplir el sistema. Además, se analiza el modelo de datos conceptual necesario para cumplir con los requerimientos identificados y que permita recoger los datos origen. En esta misma fase y partiendo del análisis anterior, se realiza el diseño del sistema definiendo la arquitectura hardware y software del mismo, y el modelo físico de datos. Por otro lado, para trasladar los datos de las bases de datos origen al almacén de datos del sistema es necesario emplear un proceso ETL. El diseño lógico de este proceso se realiza, igualmente, en esta fase.
- Implementación. Una vez que se tiene detallado el sistema el siguiente paso es su implementación. Esta fase incluye, y siguiendo este orden, la creación del la base de datos físicamente, la implementación de los scripts del proceso ETL y, por último, la generación de los informes del sistema.
- Memoria y presentación virtual. Esta fase se corresponde con la entrega final y en ella se realiza la memoria del proyecto, así como, una presentación virtual grabada en vídeo con sonido en la cual el alumno expone el trabajo realizado.

2.4.2 Hitos del proyecto.

Los principales hitos del proyecto coinciden con la conclusión de cada fase y conllevan la entrega de uno o varios productos. Son los siguientes:

Actividad	Fecha inicio	Fecha fin
PEC1 - Plan de trabajo	19/09/2013	01/10/2013
PEC2 - Análisis y Diseño	02/10/2013	05/11/2013
PEC3 - Implementación	06/11/2013	18/12/2013
PEC4 - Entrega Final y Defensa	19/12/2013	06/01/2014

2.4.3 Planificación detallada.

Como se ha indicado en el apartado anterior, cada hito se corresponde con el final de una fase y conlleva una entrega. De esta forma, descomponiendo cada fase en las distintas tareas que la componen y teniendo en cuenta los periodos de tiempo disponibles para llevar a cabo cada fase, podemos realizar una planificación detallada en la que se muestre la secuencia de realización de

cada tarea, su fecha de inicio y su fecha de finalización. Dicha planificación detallada queda de la siguiente manera:

Tarea	Fecha inicio	Fecha fin	Días
1. PEC1 - Plan de trabajo	**19/09/2013**	**01/10/2013**	**13**
1.1. Descarga y lectura de la documentación de la PEC	19/09/2013	19/09/2013	1
1.2. Descarga e instalación de software	20/09/2013	20/09/2013	1
1.3. Búsqueda y consulta de información	21/09/2013	24/09/2013	4
1.4. Análisis inicial de requerimientos	25/09/2013	26/09/2013	2
1.5. Modelado inicial de datos	27/09/2013	28/09/2013	2
1.6. Redacción del Plan de trabajo	29/09/2013	01/10/2013	3
1.7. Entrega PEC1	01/10/2013	01/10/2013	-
2. PEC2 - Análisis y Diseño	**02/10/2013**	**05/11/2013**	**35**
2.1. Descarga y lectura de la documentación de la PEC	02/10/2013	02/10/2013	1
2.2. Búsqueda y consulta de información	03/10/2013	06/10/2013	4
2.3. Aprendizaje de uso de herramientas software	07/10/2013	08/10/2013	2
2.4. Análisis detallado de requerimientos	09/10/2013	15/10/2013	7
2.5. Diseño conceptual del modelo de datos	16/10/2013	22/10/2013	7
2.6. Diseño físico del modelo de datos	23/10/2013	28/10/2013	6
2.7. Redacción del documento de entrega	29/10/2013	05/11/2013	8
2.8. Entrega PEC2	05/11/2013	05/11/2013	-
3. PEC3 - Implementación	**06/11/2013**	**18/12/2013**	**43**
3.1. Descarga y lectura de la documentación de la PEC	06/11/2013	06/11/2013	1
3.2. Búsqueda y consulta de información	07/11/2013	10/11/2013	4
3.3. Aprendizaje de uso de herramientas software	11/11/2013	12/11/2013	2
3.4. Creación física del almacén de datos	13/11/2013	14/11/2013	2
3.5. Diseño de la integración de los datos	15/11/2013	20/11/2013	6
3.6. Diseño de la extracción y transformación de datos	21/11/2013	25/11/2013	5
3.7. Diseño de la carga de datos	26/11/2013	30/11/2013	5
3.8. Pruebas de la carga de datos	01/12/2013	02/12/2013	2
3.9. Construcción de las consultas e informes	03/12/2013	10/12/2013	8
3.10. Pruebas de los informes	11/12/2013	14/12/2013	4
3.11. Redacción del documento de entrega	15/12/2013	18/12/2013	4
3.12. Entrega PEC3	18/12/2013	18/12/2013	-
4. PEC4 - Entrega Final y Defensa	**19/12/2013**	**06/01/2014**	**19**
4.1. Descarga y lectura de la documentación de la PEC	19/12/2013	19/12/2013	1
4.2. Instalación y aprendizaje de “Present” TODO	20/12/2013	20/12/2013	1
4.3. Elaboración de la memoria	21/12/2013	27/12/2013	7
4.4. Elaboración de la presentación	28/12/2013	06/01/2014	10
4.5. Entrega PEC4	06/01/2014	06/01/2014	-

2.4.4 Diagrama de Gantt.

Basándonos en la planificación detallada, el diagrama de Gantt queda de la siguiente manera[1]:

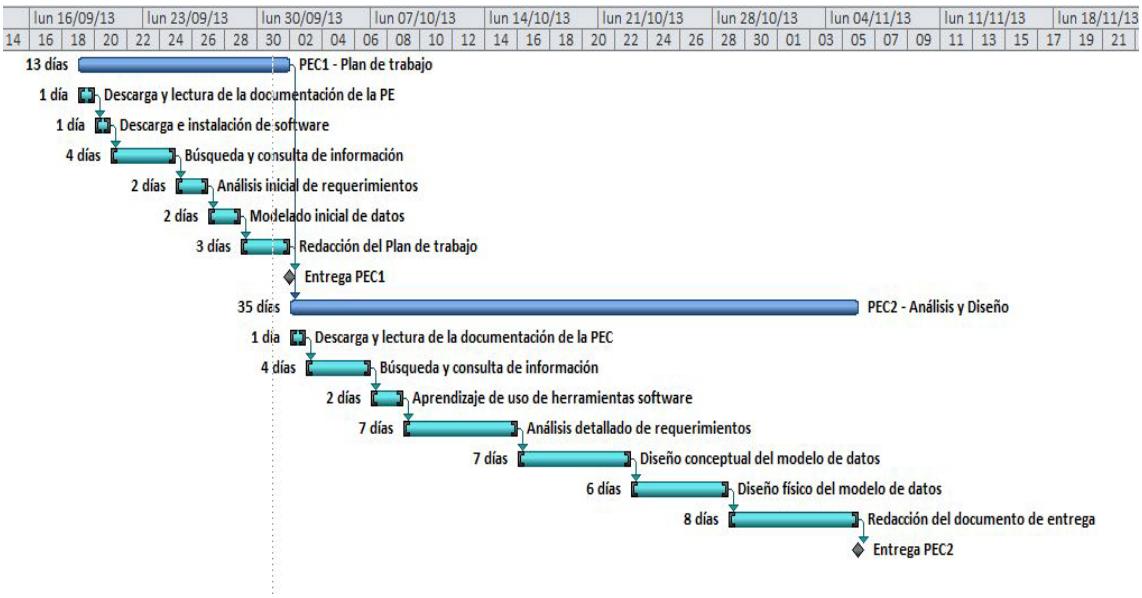


Figura 1: Diagrama de Gantt; Tareas correspondientes a las entregas Plan de trabajo y, Análisis y Diseño

[1] Para facilitar la visualización se divide en diagrama en dos ilustraciones.

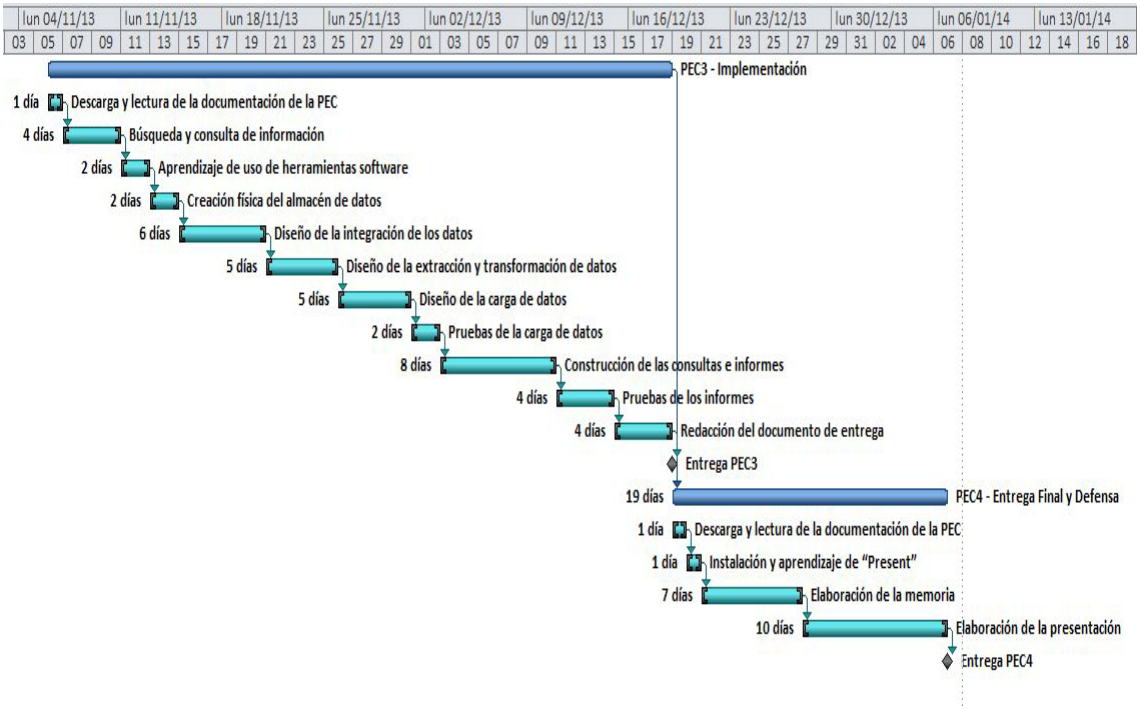


Figura 2: Diagrama de Gantt; Tareas correspondientes a las entregas Implementación y, Entrega Final y Defensa

2.4.5 Análisis de riesgos.

A continuación se detallan los inconvenientes que podrían darse durante la realización de este proyecto. Para cada uno de ellos se exponen las acciones encaminadas a prevenirlos, en caso que sea posible, así como las acciones a realizar en caso que se produzcan.

2.4.5.1 Problemas técnicos

En este tipo de problemas se incluyen tanto los problemas de hardware como de software y podrían producirse tanto en la máquina donde se realiza la documentación, como en la máquina donde se implementará el almacén de datos. Por claridad, se expone cada caso por separado.

- Máquina para la documentación. El software empleado está formado por herramientas para el diseño de diagramas, un procesador de texto para generar documentos, así como, editores de hojas excel y editores de texto para el análisis de los datos recogidos en este tipo de ficheros. En caso que fallara alguno de ellos se procedería a su reinstalación o a sustitución por una herramienta análoga. Para salvaguardar los ficheros generados con estas herramientas, documentos, imágenes, etc, se realizan copias de seguridad en

discos externos y en "la nube" empleando Ubuntu One. En caso de producirse problemas con el hardware se procedería a su reparación o sustitución lo antes posible.

- Máquina para la implementación. Se trata de un entorno virtual proporcionado por la UOC y alojado en Amazon Web Services. Por ello, asumimos que el propio servicio ofrecido por Amazon realizará copias de seguridad de los datos y aplicaciones alojados en él.

  En caso que se produjera un error en alguno de los servicios que presta este entorno, se comunicaría al consultor con el tiempo suficiente. Igualmente, se vería la posibilidad de comunicárselo a la propia administración de estos servicios Amazon. Además, es posible instalar en local las aplicaciones existentes en el entorno virtual, por lo que, en caso que se produjera un fallo de este tipo con una cierta duración, se pasaría a emplear el entorno local.

  Por otro lado, se guardará una copia de la base de datos y del trabajo alojado en este entorno, en discos externos y en “la nube” empleando Ubuntu One.

2.4.5.2 Contingencias relativas al personal

Las contingencias que podrían afectar al personal involucrado en el proyecto, es decir, el alumno serían, enfermedad grave e imposibilidad de realizar las entregas en plazos por carencia de tiempo.

Como medida preventiva, de cara a ambas situaciones, se realizará un trabajo continuo y diario intentando ajustarlo a las estimaciones de tiempos realizadas en el “Plan de trabajo”. Igualmente, se preguntarán las dudas que surjan lo antes posible para que la continuación del trabajo no se vea afectada.

En caso de producirse retrasos en los días próximos a la fecha de alguna entrega, se aumentaría el número de horas empleadas en el desarrollo del proyecto, priorizándolo con respecto a la realización de otras tareas (actualmente, estoy cursando una asignatura más, con sus PECs, etc). Si el retraso es elevado y se tienen causas justificadas, se comentaría al consultor para ver si es posible retrasar la fecha de entrega. En caso contrario, se entregaría el trabajo realizado para su valoración por parte del consultor.

Dada la importancia de este proyecto, no está contemplada la realización de vacaciones u otro tipo de descansos durante el periodo de realización del mismo.

2.4.5.3 Familiarización con las herramientas

Para el desarrollo de este proyecto, en cualquiera de sus etapas, se emplearán herramientas para las cuales se necesita un cierto aprendizaje y familiarización. Se ha planificado un cierto tiempo para ello en la planificación original. Además, se intentarán buscar ejemplos y/o tutoriales de

estas herramientas con el fin de minimizar el tiempo de aprendizaje y familiarización. En caso que alguna de ellas resulte de difícil uso, se buscará alguna herramienta alternativa. Igualmente, se preguntará al consultor de la asignatura si conoce alguna herramienta alternativa.

2.5 Productos obtenidos

Durante la realización del proyecto se han realizado varias entregas en las cuales se han obtenido los siguientes productos:

- Plan de trabajo. Documento donde se realiza un análisis de los requerimientos del sistema, así como, de los riesgos que se pueden producir durante la realización del proyecto, y se concluye con la planificación de las distintas tareas a realizar incluyendo un diagrama de Gantt.
- Análisis y diseño. Documento que recoge el análisis del sistema, tanto desde el punto de vista funcional como de modelo de datos, y el diseño del mismo, incluyendo en este último apartado la arquitectura hardware y software del sistema, el modelo físico de datos y el diseño lógico del proceso ETL.
- Implementación del sistema. Esta entrega incluye los siguientes productos:
    - Scripts para la creación de la base de datos física.
    - Scripts para la ejecución del proceso ETL
    - Documento explicativo sobre cómo acceder a la base de datos del almacén de datos y a los informes. En este documento se recogen los distintos usuario/password que permiten acceder al esquema creado. Igualmente, se recogen los distintos usuario/password y url necesarios para acceder a los informes del sistema por medio de un navegador web.
    - Documento explicativo del trabajo realizado. Este documento recoge la explicación de todo el trabajo realizado en la fase de implementación, scripts de creación de la base de datos, scripts del proceso ETL e implementación de los distintos informes.
    - Documento con capturas de pantalla de la ejecución de los distintos informes implementados en el sistema.
- Memoria y presentación virtual. Se corresponden con dos productos:
    - Memoria. Documento actual. Recoge la información relativa al análisis, diseño e implementación del sistema.
    - Presentación virtual. Video que incluye imagen y sonido donde el estudiante realiza una exposición sobre la visión general del TFG.

2.6 Breve explicación del resto de los apartados.

Esta memoria está organizada en diferentes apartados o secciones, el resto de los cuales se puede resumir de la siguiente manera:

- Análisis. Apartado que recoge todos los aspectos propios del análisis del sistema. Se analizan tanto los requerimientos como el modelo conceptual de datos.
- Diseño. En estas sección se detalla la arquitectura software y hardware, el modelo físico de datos y el diseño del proceso ETL
- Implementación. Explicación de los detalles más significativos de esta fase.
- Informes del sistema. Descripción de cada uno de los informes contemplados por el sistema junto con una captura de pantalla de su ejecución que permite comprobar el diseño de los mismos.
- Conclusiones. Conjunto de conclusiones a las que se ha llegado una vez completado el proyecto.
- Líneas de evolución futura. En este apartado se analizan posibles mejoras, así como, nuevas funcionalidades a incorporar el sistema.

3. ANÁLISIS

En este apartado se realiza un análisis del sistema a desarrollar el cual permite identificar los requerimientos, tanto funcionales como no funcionales, y definir el modelo conceptual de la base de datos multidimensional.

3.1 Requerimientos funcionales

El conjunto de requisitos funcionales de un sistema informático determina el comportamiento esperado de ese sistema, es decir, las funciones u operaciones que debe contemplar.

En nuestro caso, este comportamiento estará formado tanto por las operaciones que el sistema debe ofrecer a los usuarios, es decir, un conjunto de informes, como por las operaciones necesarias para que el almacén de datos pueda ser utilizado, es decir, la operativa propia del proceso ETL. Por tanto, podemos identificar las siguientes funcionalidades:

- Cada uno de los informes que debe ofrecer el sistema se corresponde con un requisito funcional. Estos informes se podrán ejecutar de manera agregada por comarca/provincia, tipo de vehículo y tipo de permiso de conducción. Además, hay que tener en cuenta que la temporalidad de los datos será a nivel de año. Los informes en cuestión son las siguientes:

  - Total de vehículos
  - Total de conductores
  - Porcentaje de vehículos respecto a la población
  - Densidad de población (habitantes/km2)
  - Densidad de tráfico (vehículos/km2)
  - Número de vehículos respecto al número de radares
  - Porcentaje de conductores por radar
  - Indicador de conductores vs habitantes por género
  - Indicador de radares vs vehículos
  - Ratio de vehículos por conductor
  - Cantidad de vehículos por superficie del territorio
- Tal como se ha comentado, las operaciones necesarias para que el sistema se encuentre operativo también son considerados requisitos funcionales[2]. Son estos:
  - Administración del sistema. El sistema debe permitir su propia administración, de forma que se contemple la modificación de la base de datos multidimensional, así como realizar la gestión de usuarios y sus permisos.
  - Ejecución del proceso ETL. Para que el almacén de datos esté operativo es necesario que el sistema contemple un proceso ETL encargado de extraer y transformar los datos recogidos en los diferentes ficheros de los que se dispone, y cargarlos en la base de datos multidimensional. Este proceso estará diseñado para una carga inicial y podría ser modificado si en el futuro se deseara registrar nuevos datos en el almacén y estos datos tuvieran un formato distinto del actual, o se encontraran recogidos en ficheros con formatos distintos al actual.
  - Administración de informes. Para poder ejecutar los distintos informes contemplados por el sistema se hace necesaria la administración de los mismos. Esta administración incluye creación, modificación y eliminación de informes.

3.1.1 Diagrama de casos de uso

Podemos identificar dos actores distintos: El usuario, que ejecuta los informes para realizar los análisis pertinentes, y el administrador, encargado de mantener el sistema en estado operativo. En el caso del administrador también podrá ejecutar informes, bien para realizar pruebas o porque realice funciones de usuario.

[2] Si bien alguna de estas funcionalidades podría estar implementada o parcialmente implementada por las herramientas de la base de datos, no dejan de ser por ello requisitos funcionales.

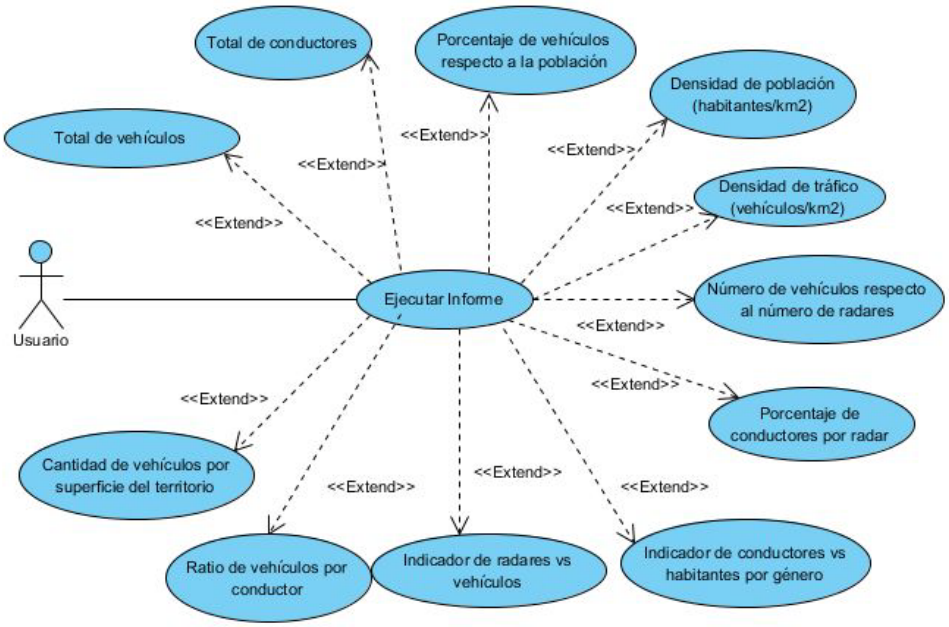


Figura 3: Casos de uso del actor Usuario

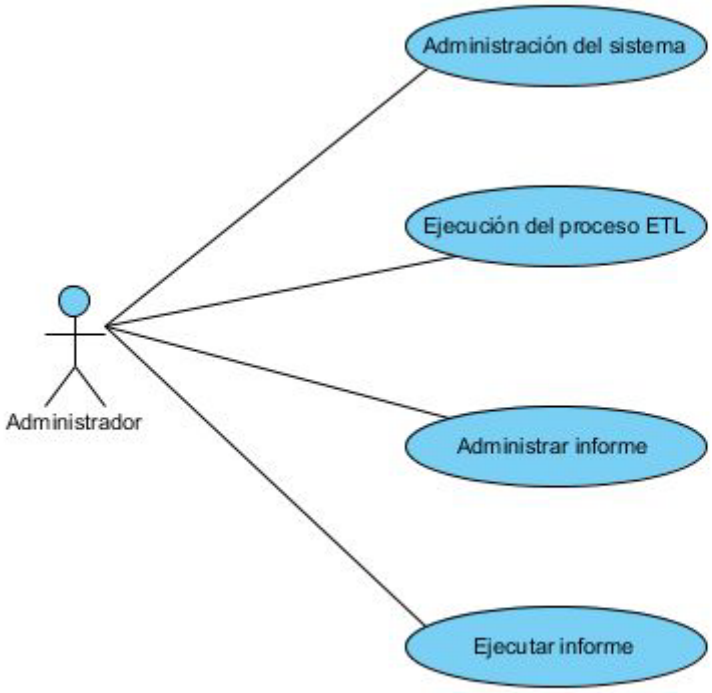


Figura 4: Casos de uso del actor Administrador

3.2 Requerimientos no funcionales

Además del comportamiento antes indicado, el sistema debe contemplar un conjunto de características y/o restricciones en la forma de llevar a cabo este comportamiento. Son los denominados requerimientos no funcionales. Nuestro sistema debe contemplar los siguientes:

3.2.1 Usabilidad

La ejecución de los distintos informes debe de resultar fácil para los usuarios. Igualmente, la visualización de los datos de estos informes debe tener un diseño que facilite el análisis de los mismos. La facilidad de uso debe ser contemplada en toda la aplicación, si bien, se hace especial hincapié en las funcionalidades orientadas a los usuarios, ya que podrían carecer de conocimientos técnicos.

3.2.2 Seguridad

Es necesario que el sistema contemple los mecanismos de seguridad pertinentes para evitar posibles robos de datos, o acceso y manipulación de los mismos por parte de personas no autorizadas. Dichos mecanismos de seguridad tendrán en cuenta tanto los datos del almacén de datos como los recogidos en fichero.

3.2.3 Rendimiento

Los almacenes de datos están pensados para guardar y analizar un alto volumen de datos. Se hace necesario, por tanto, que todos los procesos sobre dichos datos no supongan un consumo excesivo de recursos ni de tiempo de ejecución. Esta restricción es aplicable tanto al proceso ETL como a la ejecución de consultas/informes.

3.2.4 Mantenibilidad

En el futuro, podría ser necesaria la modificación de algún informe, la inclusión de nuevos o realizar nuevas cargas de datos en el almacén. Por tanto, el sistema debe estar diseñado de forma que sea fácil la corrección de errores, así como, la modificación de cualquiera de sus funcionalidades y la inclusión de otras nuevas.

3.2.5 Fiabilidad

Ya que, el objetivo principal del proyecto es la realización de análisis de datos, empleando para ello un almacén de datos, se hace imprescindible asegurar que la información generada por el sistema sea correcta. Para ello, el sistema debe establecer los mecanismos pertinentes que le permitan tolerar errores generados por el usuario o el hardware, o que avisen de dichos errores, en caso de no poder gestionarlos.

3.3 Modelo de datos conceptual

A continuación se detalla el modelo conceptual de la base de datos multidimensional, describiendo su estructura de datos de manera independiente de la tecnología empleada para su implementación. Para ello, inicialmente se indica el enfoque adoptado para la realización del modelo. Posteriormente, se detallan los distintos elementos que forman el modelo y se muestra el diagrama del mismo. Finalmente, se indican una serie de comentarios breves acerca del modelo conceptual.

3.3.1 Enfoque adoptado

Si bien, como se ha indicado anteriormente, el modelo de datos conceptual es independiente de la tecnología empleada para su implementación, la base de datos a modelar se corresponde con un almacén de datos (Data Warehouse), por lo tanto, los pasos a seguir para conseguir este modelo difieren de los propios de una base de datos convencional.

En nuestro caso se ha seguido el paradigma de Ralph Kimball. Según la definición de Ralph Kimball un Data Warehouse es *"una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis"*. De una manera un poco más técnica, Ralph Kimball también indica que un Data Warehouse es *"la unión de todos los Data marts de una entidad"*. Un Data Mark es un conjunto de datos de un área específica dentro de una organización (p.e. ventas, producción, etc) cuyo propósito es ayudar a tomar mejores decisiones. Por esta razón, un Data Mark es un sistema orientado a la consulta por medio de herramientas OLAP, sobre el que se realizan cargas de datos con muy baja frecuencia. Se puede concluir que un Data Mark es un pequeño Data Warehouse centrado en un tema o área de negocio específico.

3.3.2 Dimensiones y atributos.

- INF_GEOGRAFICA (denominada MUNICIPIO en la PEC1). Recoge los datos propios de la información geográfica. Sus atributos son los siguientes:
    - Provincia
    - Demarcación
    - Comarca
    - Municipio
    - Cod_INE. En principio podría no ser necesario para la realización de los informes solicitado, se añade para facilitar el proceso ETL.
    - Vía. Este atributo no es necesario para la realización de los informes solicitados, se añade pensando en posibles informes futuros, p.e. Número de radares por vía, teniendo en cuenta que una misma vía puede discurrir por varias demarcaciones geográficas (municipio, provincia, etc).

  Los atributos "Número de habitantes" y "Extensión" se han eliminado de esta dimensión, ya que se encuentran incluidos en la tabla de hechos
- TIEMPO. Determina cuando sucedió un determinado hecho. Como la temporalidad de los datos es a nivel de año sólo tiene un atributo.
    - Año

3.3.3 Jerarquías

Una jerarquía representa una relación 1-N (o padre-hijo) entre dos o más atributos de una dimensión. Cabe decir, que en una misma dimensión pueden darse varias jerarquías y, que están compuestas por dos o más niveles.

Las jerarquías detectadas son:

- INF_GEOGRAFICA. Provincia → Demarcación → Comarca → Municipio
  Podría incluirse el atributo vía como último elemento de esta jerarquía, de momento no se ha hecho así, ya que este atributo se ha añadido pensando en posibles informes futuros.

3.3.4 Diagrama del modelo conceptual

Abajo se indica el diagrama del modelo conceptual, en el cual se puede observar una única tabla de hechos, llamada "EVOLUCIÓN TRÁFICO", relacionada con cinco tablas de dimensiones. Además, vemos que las tablas de dimensiones sólo se relacionan con la tabla de hechos, por lo tanto seguimos un esquema en estrella.

En caso que algunas tablas de dimensiones estuvieran organizadas jerárquicamente, es decir, de alguna tabla de dimensiones colgara una o más tablas de dimensiones (en una relación padre-hijo) estaríamos hablando de un esquema en copo de nieve.

Nos encontraríamos en un tercer caso si tuviéramos varias tablas de dimensiones, con lo que, nuestro modelo se ajustaría a un esquema en constelación.

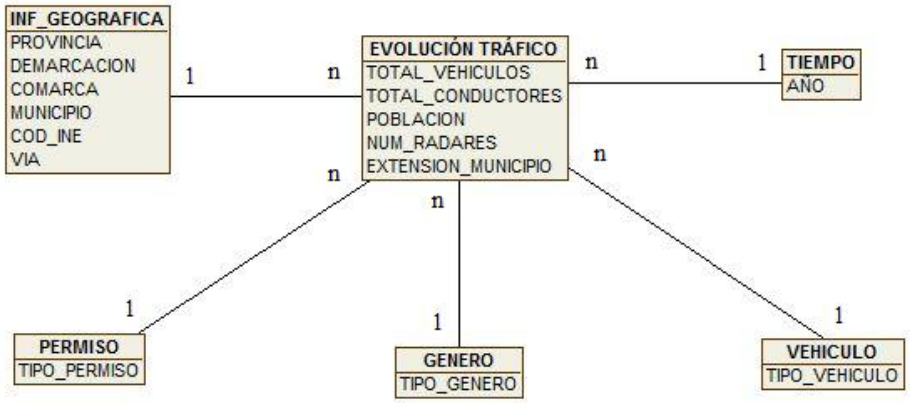


Figura 5: Diagrama del modelo conceptual

3.3.5 Otros comentarios sobre el modelo

Para facilitar el entendimiento del modelo conceptual, así como, el proceso seguido para su definición, se exponen los siguientes comentarios:

- Las dimensiones PERMISO, GENERO y VEHICULO, pueden considerarse Dimensiones Junk, puesto que sus atributos son indicadores o flags con una cardinalidad baja.
- Se considera que los datos de entrada están lo suficientemente desnormalizados, por tanto, no se ha planteado ningún proceso de desnormalización.
- Una vez determinado el modelo de datos y, si tenemos en cuenta que el objetivo de dicho modelo es la implementación de una almacén de datos para la generación de informes que permitan el análisis de un gran volumen de datos acumulados correspondientes a varios años, podemos indicar que este modelo podría ser empleado como soporte para implementar técnicas de minería de datos que permitan extraer conocimiento oculto, así como, inferir posibles comportamientos futuros basados en patrones o comportamientos repetidos en el pasado. Si bien es cierto que las técnicas de minería de datos pueden implementarse en un OLTP, emplear un almacén de datos permite contar con los beneficios del proceso ETL para mejorar la calidad de los datos y el hecho de que un almacén de datos agrupe datos provenientes de varias bases de datos. Para finalizar con este comentario, un posible comportamiento a inferir con minería de datos sería determinar el posible aumento de tráfico en una determinada demarcación geográfica. Este aumento estaría determinado por la evolución del número de conductores y la evolución del número de vehículos en dicha demarcación geográfica. Es necesario tener en cuenta ambas variables, ya que si aumentan los conductores pero no los vehículos el tráfico podría no verse afectado. Lo mismo ocurriría si aumenta el número de vehículos pero no el número de conductores, porque un conductor sólo puede circular con un vehículo a la vez.

4. DISEÑO

En esta sección se expone el diseño del sistema. Inicialmente, se detallan los elementos que forman la arquitectura del sistema desde el punto de vista software y hardware, en ambos casos, se muestra su respectivo diagrama. Posteriormente, se expone el modelo de datos físico, por lo que se indican sus tablas, así como, la estructura de estas (campos, claves, etc). Además, se realiza un diseño de alto nivel del proceso ETL.

4.1 Arquitectura software

La arquitectura software del sistema está formada por los siguientes elementos:

- Fuentes de datos. Las fuentes de datos constituyen las distintas bases de datos origen a partir de las cuales poblaremos el almacén de datos. Físicamente, están formados por un conjunto de ficheros con diferentes formatos (excel, txt, csv).

- Proceso ETL. Proceso de Extracción, Transformación y Carga de los datos de entrada en el almacén de datos. Este proceso se realiza empleando dos series de scripts a ejecutar con dos herramientas. Primero se ejecutan una serie de scripts por medio de la herramienta "Oracle SQL*Loader", la cual consiste en un motor de parseo de datos. Y después se ejecuta otra serie de scripts por medio de la herramienta "SQLPlus". Ambas herramientas están integradas en el SGBD Oracle. La primera serie de scripts toma los datos de las fuentes de datos origen y registra una parte de estos datos en la base de datos del almacén, y otra parte en un área temporal denominada staging area[3]. Después, la segunda serie de scripts trata los datos recogidos en el staging area para terminar de poblar el almacén de datos.
- Almacén de datos. El almacén de datos se implementa por medio de una base de datos relacional alojada en el SGBD "Oracle 11g R2 XE". Esta base de datos recoge tanto las tablas propias del almacén, como las tablas de trabajo pertenecientes al staging area.
- Generación de informes. Para generación de los distintos informes se emplean herramientas pertenecientes al paquete software "MS SQL Server 2012". Principalmente[4], se emplean dos herramientas, "Visual Studio 2010" y "Report Server". La primera de estas herramientas consiste en un entorno de desarrollo que nos permite implementar los distintos informes del sistema. Una vez que tenemos los informes construidos se hace un despliegue de los mismos (desde el propio Visual Studio 2010) en el servidor de informes "Report Server", el cual ofrece una URL para poder ejecutar los informes desde un navegador web. Durante la ejecución de los informes, "Report Server" accede a la base de datos relacional, por tanto, se trata de un sistema ROLAP (Relational On Line Analytic Processing). Este tipo de sistemas realizan procesamiento analítico OnLine accediendo a bases de datos relacionales, además, generan dinámicamente cubos multidimensionales cuando se genera un informe. Para ello, determinan los indicadores, atributos, jerarquías, etc, que formarán el cubo multidimensional y realizan una consulta en el almacén de datos sobre dichos elementos. La generación dinámica de los cubos hace el proceso transparente al usuario, el inconveniente es que se deben generar cada vez que se ejecuta un informe.

  Una alternativa a las herramientas ROLAP, son las de tipo MOLAP (Multidimentional On Line Analytic Processing). En este caso, los cubos multidimensionales que permitirán generar los informes deben ser calculados de forma previa a dicha generación de informes. Con ello se consigue ganar velocidad en la ejecución de los informes, a costa de un mayor espacio de almacenamiento y la necesidad de recalcular un cubo totalmente cada vez que cambie.

[3] En la nomenclatura de "Oracle SQL*Loader" se le denomina tablas de trabajo.
[4] No es el objeto de esta memoria entrar en detalle de todas las herramientas que incluye el paquete "MS SLQ Server 2012". Si bien, en el apartado sobre la implementación del sistema se verán detalles de su configuración necesarios para la puesta en funcionamiento de los informes.

Una tercera alternativa serían los sistemas HOLAP (Hybrid On Line Analytic Processing) que podrían considerarse un híbrido entre MOLAP y ROLAP.

- Navegador Web. Empleando un navegador Web los usuarios podrán acceder a los distintos informes ofrecidos por el sistema, a partir de la URL ofrecida por el servidor "Report Server".

En el siguiente diagrama podemos observar la arquitectura software de nuestro sistema.

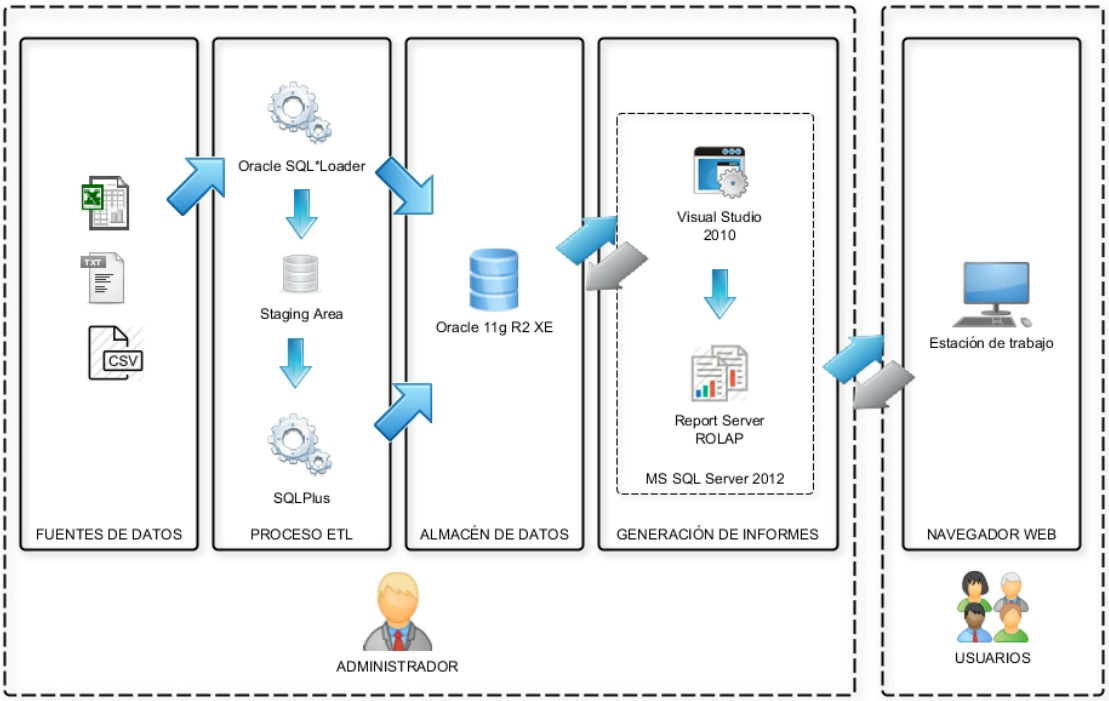


Figura 6: Diagrama de la arquitectura software

Además, se puede observar a que elementos del sistema accede el administrador y a cuales los usuarios.

4.2 Arquitectura hardware

La arquitectura hardware del sistema está formada por los siguientes elementos:

- Servidor de ficheros. Se trata de un ordenador donde residen los ficheros que contienen los datos de entrada del sistema.
- Servidor de BB.DD Oracle. Servidor con un gestor de bases de datos instalado, en nuestro caso será Oracle, que mantiene y permite la gestión del almacén de datos.

Si bien el servidor de ficheros y el servidor de BB.DD. Oracle se muestran como dos máquinas distintas, perfectamente podría tratarse de una sola donde estuvieran los ficheros con los datos de entrada y el SGBD Oracle.

- Red de ordenadores. Esta red permite la conexión de las distintas máquinas.
- Estación de trabajo del administrador del sistema.
- Estaciones de trabajo de los usuarios.

El siguiente diagrama muestra la arquitectura hardware del sistema.

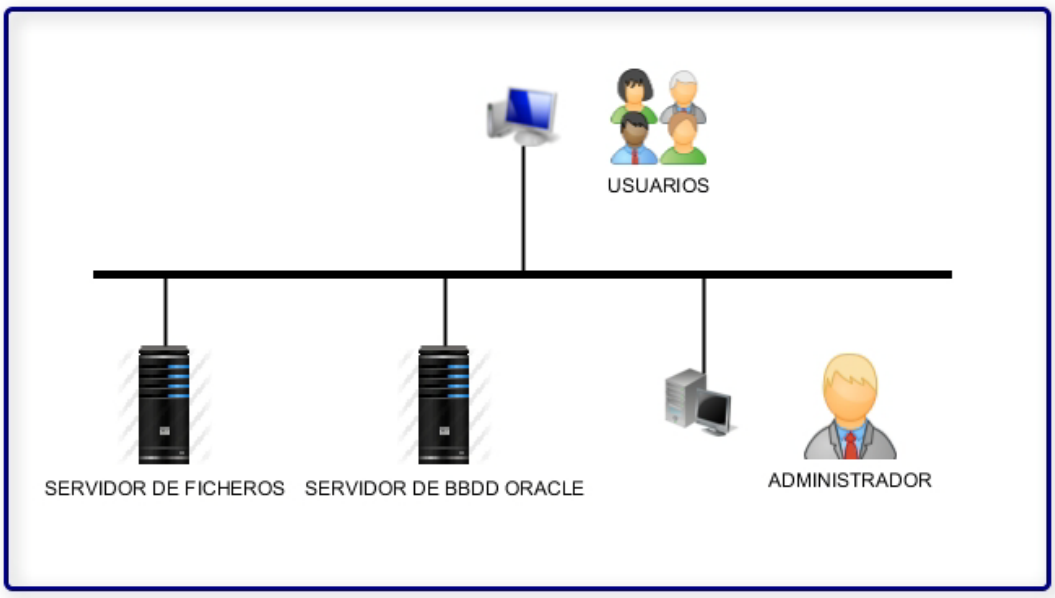


Figura 7: Diagrama de la arquitectura hardware

4.3 Modelo de datos físico

Seguidamente, se detalla el modelo de datos físico diseñado a partir del modelo de datos conceptual. Para ello, inicialmente se muestra el diagrama del modelo y después se detallan cada una de las tablas del modelo indicando sus campos y las características de estos.

4.3.1 Diagrama del modelo físico

El siguiente diagrama muestra el modelo de datos físico. Los nombres de las tablas han sido precedidos por los prefijos TH_ en la tabla de hechos, o TD_ en las tablas de dimensiones. Los

nombres de los campos correspondientes a claves contienen el sufijo PK_ cuando son claves primarias, o FK_ cuando son claves foráneas.

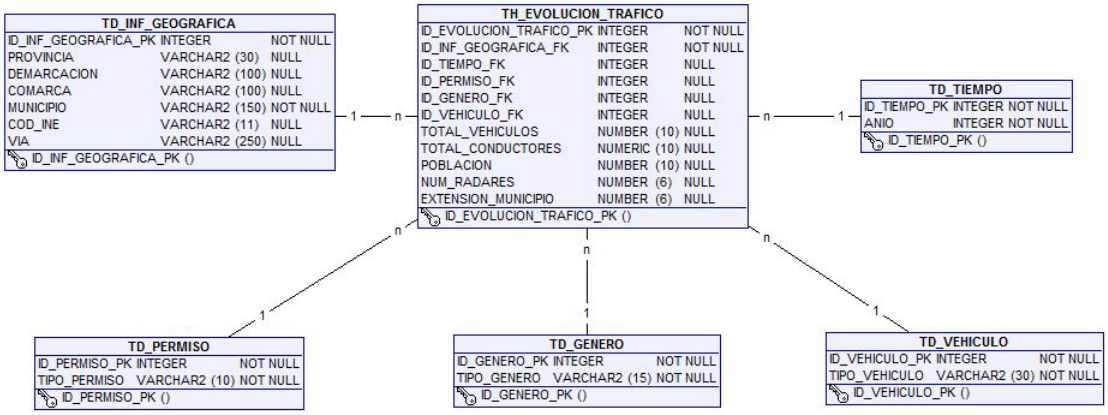


Figura 8: Diagrama del modelo físico

4.3.1.1 Tabla TH_EVOLUCION_TRAFICO

Tabla que recoge los datos propios de los hechos a analizar. El detalle de sus campos es el siguiente:

Nombre	Tipo	Tamaño	Dec.	Nullable	Descripción
ID_EVOLUCION_TRAFICO_PK	INTEGER	-	-	NO	Clave primaria
ID_INF_GEOGRAFICA_FK	INTEGER	-	-	NO	Clave foránea con la tabla TD_INF_GEOGRAFICA
ID_TIEMPO_FK	INTEGER	-	-	SI	Clave foránea con la tabla TD_TIEMPO
ID_PERMISO_FK	INTEGER	-	-	SI	Clave foránea con la tabla TD_PERMISO
ID_GENERO_FK	INTEGER	-	-	SI	Clave foránea con la tabla TD_GENERO
ID_VEHICULO_FK	INTEGER	-	-	SI	Clave foránea con la tabla TD_VEHICULO
TOTAL_VEHICULOS	NUMBER	10	0	SI	Número total de vehículos
TOTAL_CONDUCTORES	NUMBER	10	0	SI	Número total de conductores
POBLACION	NUMBER	10	0	SI	Número de habitantes
NUM_RADARES	NUMBER	6	0	SI	Número de radares
EXTENSION_MUNICIPIO	NUMBER	6	0	SI	Extensión de un municipio en km^2

Las claves foráneas, a excepción de ID_INF_GEOGRAFICA_FK, y los campos de datos están definidos como Nullable porque durante el proceso ETL las tablas se irán poblando poco a poco, según se procesen los distintos ficheros de entrada. De esta manera, es posible que se

inserte un registro en la tabla de hechos con valores en alguna de sus columnas y, posteriormente, se de valor a otras de sus columnas. Esta situación se puede dar incluso con las claves foráneas, ya que al tratar un fichero de entrada, este podría tener algunos de los datos de la tabla TH_EVOLUCION_TRAFICO, por ejemplo la extensión de un municipio, pero no contener los datos de las tablas con las que se relaciona la tabla de hechos, por ejemplo género. La única excepción en este sentido el la clave foránea ID_INF_GEOGRAFICA_PK (que relaciona la tabla TH_EVOLUCION_TRAFICO con la tabla TD_INF_GEOGRAFICA), esto se debe a que todos los ficheros de entrada tiene el nombre del municipio, y por tanto, si se guardan datos en la tabla TH_EVOLUCION_TRAFICO estos tendrán relación con algún registro de la tabla TD_INF_GEOGRAFICA.

Los campos TOTAL_VEHICULOS, TOTAL_CONDUCTORES, POBLACION, NUM_RADARES, EXTENSION_MUNICIPIO, tienen como valor inicial 0. Esto a así porque existen datos para los que no hay valor en todos los años que se están tramitando. Por ejemplo, la información referente al número de conductores (la cual se encuentra registrada en los ficheros Dades_conductorsXXXX.txt) no está disponible para el año 2012, sin embargo si existe información para el año 2012 referente a la población de los municipios. De esta forma, en la tabla TH_EVOLUCION_TRAFICO existirán registros cuyo ID_TIEMPO_PK hagan referencia al año 2012 (registrado en TD_TIEMPO) algunos de los campos indicados al comienzo del párrafo tendrán valor y otros estarán a 0. Además, como se ha indicado la carga de datos de esta tabla se realizará en varios pasos. Ocurrirá que algunos campos estén a 0 inicialmente y posteriormente se les establezca valor.

4.3.1.2 Tabla TD_INF_GEOGRAFICA

Tabla que recoge los datos de la dimensión INF_GEOGRAFICA. El detalle de sus campos es el siguiente:

Nombre	Tipo	Tamaño	Dec.	Nullable	Descripción
ID_INF_GEOFRAFICA_PK	INTEGER	-	-	NO	Clave primaria
PROVINCIA	VARCHAR2	30	-	SI	Nombre de la provincia
DEMARCACION	VARCHAR2	100	-	SI	Nombre de la demarcación
COMARCA	VARCHAR2	100	-	SI	Nombre de la comarca
MUNICIPIO	VARCHAR2	150	-	SI	Nombre del municipio
COD_INE	VARCHAR2	11	-	SI	Código INE del municipio
VIA	VARCHAR2	250	-	SI	Nombre de la vía

4.3.1.3 Tabla TD_TIEMPO

Tabla que recoge los datos de la dimensión temporal. El detalle de sus campos es el siguiente:

Nombre	Tipo	Tamaño	Dec.	Nullable	Descripción
ID_TIEMPO_PK	INTEGER	-	-	NO	Clave primaria
ANIO	INTEGER	-	-	NO	Año

4.3.1.4 Tabla TD_PERMISO

Tabla que recoge los datos de la dimensión PERMISO, que indica los diferentes tipos de permiso. El detalle de sus campos es el siguiente:

Nombre	Tipo	Tamaño	Dec.	Nullable	Descripción
ID_PERMISO_PK	INTEGER	-	-	NO	Clave primaria
TIPO_PERMISO	VARCHAR2	10	-	NO	Tipo de permiso de conducir. Los posibles valores son: Permiso; Licencia.

4.3.1.5 Tabla TD_GENERO

Tabla que recoge los datos de la dimensión GÉNERO, que indica los diferentes géneros. El detalle de sus campos es el siguiente:

Nombre	Tipo	Tamaño	Dec.	Nullable	Descripción
ID_GENERO_PK	INTEGER	-	-	NO	Clave primaria
TIPO_GENERO	VARCHAR2	10	-	NO	Tipo de género. Los posibles valores son: Masculino, femenino.

4.3.1.6 Tabla TD_VEHICULO

Tabla que recoge los datos de la dimensión VEHICULO, que indica los diferentes tipos de vehículos. El detalle de sus campos es el siguiente:

Nombre	Tipo	Tamaño	Dec.	Nullable	Descripción
ID_VEHICULO_PK	INTEGER	-	-	NO	Clave primaria
TIPO_VEHICULO	VARCHAR2	10	-	NO	Tipo de vehículo. Los posibles valores son: Vehículos de motor; Automóviles; Camiones y furgonetas; Otros vehículos de motor; Motocicletas; Autobuses; Tractores industriales; Resto vehículos a motor.

4.4 Diseño de alto nivel del proceso ETL

Antes de poder emplear el almacén de datos para ejecutar informes sobre él, es necesario realizar el proceso ETL (Extract, Transform and Load), extracción, transformación y carga. Este proceso, que consta de tres pasos, permite extraer datos desde diferentes fuentes de datos, transformarlos en caso que fuera necesario para adecuarlos a los tipos de datos del almacén destino y, finalmente, cargarlos en dicho almacén. En nuestro caso, tenemos distintas fuentes de datos recogidas en ficheros de distintos formatos (csv, excel y txt). Por otro lado, para realizar el proceso ETL nos apoyaremos en la herramienta Oracle SQL*Loader que se encuentra en el SGBD Oracle, el cual a su vez soportará el almacén de datos. Esta herramienta tiene sus limitaciones, por lo que, como se verá en los siguientes apartados, ha sido necesario emplear tablas de trabajo y, posteriormente, ejecutar scripts pl/sql que traten los datos registrados en estas tablas de trabajo y pueblen o terminen de poblar ciertas tablas del almacén de datos.

En esta sección se detalla el diseño de alto nivel del proceso ETL.

4.4.1 Estrategia adoptada.

En base al análisis de las fuentes de datos origen realizado al principio del proyecto, la estrategia adoptada consiste en cargar primero las dimensiones junk (tablas TD_PERMISO, TD_GENERO y TD_VEHICULO), así como la dimensión tiempo (tabla TD_TIEMPO). Esto se debe a que sus datos son conocidos y no es necesario obtenerlos desde ningún fichero, con lo cual el proceso de carga es más rápido. Después, se realiza el proceso ETL a cada uno de los ficheros de entrada correspondientes a las fuentes de datos origen. Para realizar esta parte se emplea la herramienta Oracle SQL*Loader, que lee los datos de los ficheros de entrada y guarda una parte de ellos en el almacén y el resto los registra en tablas de trabajo. Posteriormente, se ejecutan scripts pl/sql para poblar las tablas del almacén de datos con los datos de las tablas de trabajo. En este sentido, es necesario tener en cuenta que las tablas de dimensiones deben ser pobladas antes que la tabla de hechos (TH_EVOLUCION_TRAFICO), porque en esta última existen claves foráneas a las primeras. De esta forma, se realiza la población de la dimensión INF_GEOGRAFICA y por último la tabla de hecho EVALUACION_TRAFICO.

A continuación se detalla el proceso ETL de acuerdo a esta estrategia.

4.4.2 Ejecución coordinada del proceso ETL.

Como se ha podido comprobar en el apartado anterior, es necesario que la población de las distintas tablas del almacén de datos se realice en un determinado orden, por ello, la ejecución de los distintos scripts del proceso ETL también debe llevar el mismo orden. Si bien, en el apartado dedicado a exponer los distintos detalles de la implementación se indicará el nombre

concreto de estos scripts, así como la manera de ejecutarlos, en este apartado se detalla el orden de ejecución de los distintos scripts:

- Scripts de carga de las tablas de dimensiones junk y la tabla de dimensión tiempo. El orden concreto en la ejecución de estos scripts es indiferente, ya que estas tablas son independientes entre ellas.
- Scripts del proceso ETL sobre los ficheros de entrada. El orden de ejecución del proceso ETL sobre los distintos ficheros de entrada es: Dades_municipis.xls, Dades_vehicles.xls, Dades_conductos_XXXX.txt y Radars_SCT.txt. Este orden se debe a que inicialmente se pensó que el proceso ETL podría llevarse a cabo empleando únicamente la herramienta Oracle SQL*Loader. Posteriormente se comprobó que no era posible dadas las limitaciones de esta herramienta, sin embargo se mantuvo este orden.
    - El primer fichero a tratar es Dades_municipis.xls porque se supone que tiene registrados todos los municipios de Cataluña, ya que la información que contiene (población y extensión) es intrínseca a cualquier municipio, con lo que poblamos completamente la dimensión TD_INF_GEOGRAFICA con el nombre del municipio y su código INE. Esta asunción podría no ser correcta por lo que hay que tenerlo en cuenta en el resto de scripts que se encarguen de poblar esta tabla. Además del tratamiento anterior, todos los datos de este fichero se registran en la tabla de trabajo WORK_TABLE_DADES_MUNICIPIS, para su posterior uso.
    - Fichero Dades_vehicles.xls que, al igual que el primero, contiene el código INE de los municipios. Los datos de este fichero se registran en la tabla de trabajo WORK_TABLE_DADES_VEHICLES, para su posterior uso.
    - Ficheros Dades_conductos.txt. Los datos de estos ficheros se guardan en la tabla de trabajo WORK_TABLE_DADES_CONDUCTOS, para su posterior uso.
    - Fichero Radars_SCT.txt. Los datos de este fichero se guardan en la tabla de trabajo WORK_TABLE_RADARS_SCT, para su posterior uso.
- Scripts del proceso ETL sobre las tablas de trabajo (scripts pl/spl). Está formado por dos scripts. El primero trata los datos necesarios para terminar de poblar la tabla TD_INF_GEOGRAFICA, el segundo se encarga de poblar la tabla TH_EVOLUCION_TRAFICO.

Siguiendo este orden, se detalla el proceso ETL a realizar.

4.4.3 Carga de la dimensión PERMISO

Esta dimensión sólo tiene un campo cuyos valores son conocidos, por lo que se pueden insertar en su tabla correspondiente TD_PERMISO. Los valores son “Permiso” y “Licencia”, y el campo es TIPO_PERMISO.

4.4.4 Carga de la dimensión GENERO

Esta dimensión sólo tiene un campo cuyos valores son conocidos, por lo que se pueden insertar en su tabla correspondiente TD_GENERO. Los valores son “Masculino” y “Femenino”, y el campo es TIPO_GENERO.

4.4.5 Carga de la dimensión VEHICULO

Esta dimensión sólo tiene un campo cuyos valores son conocidos, por lo que se pueden insertar en su tabla correspondiente TD_VEHICULO. Los valores son “Vehículos de motor”, “Automóviles”, “Camiones y furgonetas”, “Otros vehículos de motor”, “Motocicletas”, “Autobuses”, “Tractores industriales” y “Resto vehículos a motor”, y el campo es TIPO_GENERO.

4.4.6 Carga de la dimensión TIEMPO

Esta dimensión sólo tiene un campo cuyos valores son conocidos, por lo que se pueden insertar en su tabla correspondiente TD_TIEMPO. Los valores en cuestión se corresponden con los años para los que existen datos en alguno de los ficheros de entrada, es decir, los años desde el 2007 al 2012 y el campo es ANIO. Esta tabla tiene un identificador que funciona como clave primaria y, a su vez, es clave foránea en la tabla de hechos (TH_EVOLUCION_TRAFICO). Para facilitar la relación entre cada registro de la tabla de hechos con el correspondiente registro de la dimensión TIEMPO que identifica el año de estos hechos, se ha empleado el siguiente truco, el campo ID_TIEMPO_PK tendrá el mismo valor que el campo ANIO. Como el número de registros totales de esta tabla no será muy cuantioso esta idea no supone ningún problema.

4.4.7 Carga del fichero Dades_municipis.xls

Se trata de un único fichero con la evolución de la población por año de los municipios de Cataluña y la extensión de estos. Concretamente los datos son: el nombre de cada municipio, su código INE, su extensión en km^2 y la población en los años 2012 a 2007. Como este fichero tiene formato excel y para el proceso ETL vamos a apoyarnos en Oracle SQL*Loader, es necesario guardar el fichero con formato csv separado por comas (aunque excel indica este carácter como separador de datos, realmente emplea el punto y coma), antes de realizar este cambio de formato eliminamos la fila de las cabeceras. Además, guardamos este fichero en formato UTF-8 para evitar problemas con los caracteres especiales. El tratamiento a realizar sobre cada línea de este fichero es el siguiente:

1) Los dos primeros columnas, nombre del municipio y código INE se guardan en la tabla TD_INF_GEOGRAFICA como un registro nuevo.

2) Se guarda un nuevo registro en la tabla de trabajo WORK_TABLE_DADES_MUNICIPIS con todos los datos existentes en la línea del fichero.

Excepto las dos primeras columnas, el resto recogen datos numéricos aunque podrían contener el valor “n.d.” (No disponible). El tratamiento que se realiza a este tipo de valores es común en todos los ficheros, por lo que se detallará en otro apartado.

4.4.8 Carga del fichero Dades_vehicles.xls

Se trata de un único fichero con información del número de vehículos de distintas categorías de los municipios de Cataluña. Como este fichero tiene formato excel y para el proceso ETL vamos a apoyarnos en Oracle SQL*Loader, es necesario guardar el fichero con formato csv separado por comas (aunque excel indica este carácter como separador de datos, realmente emplea el punto y coma), antes de realizar este cambio de formato eliminamos la fila de las cabeceras. Además, guardamos este fichero en formato UTF-8 para evitar problemas con los caracteres especiales. Los datos de este fichero se registran en la tabla de trabajo WORK_TABLE_DADES_VEHICLES, para su posterior uso.

Excepto las dos primeras columnas, el resto recogen datos numéricos aunque podrían contener el valor “n.d.” (No disponible).

4.4.9 Carga del fichero Dades_conductors XXXX.txt

Se trata de 5 ficheros, uno por cada año desde el 2007 al 2011, donde se registran el número de permisos y licencias, agrupado por género y municipio. Todos ellos son ficheros de texto que incluyen líneas al principio y al final que no aportan datos. Para facilitar el proceso ETL se eliminan estas líneas y se guardan los ficheros en formato UTF-8, para evitar problemas con los caracteres especiales.

En todos los ficheros excepto “Dades_conductors 2011.txt” los valores están separados por tabuladores, y los valores numéricos se encuentran delimitados por comillas dobles. En el caso de “Dades_conductors 2011.txt”, podría interpretarse como separador de campos la secuencia de caracteres punto y coma seguido de comillas dobles (;”). Si bien, no queda del todo claro cuál es el carácter o secuencia de caracteres separadores. Se decide realizar la siguiente modificación sobre él, con vistas a facilitar el proceso ETL, así como, unificarlo en la medida de lo posible con el resto de ficheros de la misma fuente de datos. La modificación es esta:

- Eliminar todas las comillas dobles
- Reemplazar el carácter punto y coma (;) por un tabulador.

De esta forma, todos los ficheros quedan con el mismo formato, si bien, al tratar los valores numéricos habrá que comprobar si están iniciados y finalizados por el carácter comillas dobles, en cuyo caso habrá que eliminar estos caracteres antes de convertir el valor a numérico. Los datos de estos ficheros se guardan en la tabla de trabajo WORK_TABLE_DADES_CONDUCTOS, para su posterior uso.

4.4.10 Carga del fichero Radars_SCT.txt

Archivo de texto que recoge para cada radar la vía, el municipio, la comarca y la demarcación donde se encuentra ubicado. Pueden existir líneas repetidas (todos sus datos son iguales), esto se debe a que en una misma vía de un mismo municipio puede haber ubicados varios radares. Sus datos están separados por tabuladores. Al principio y al final del mismo, existen una líneas que no aportan información Para facilitar el proceso ETL se eliminan estas líneas y se guardan los ficheros en formato UTF-8, para evitar problemas con los caracteres especiales. Los datos de este fichero se guardan en la tabla de trabajo WORK_TABLE_RADARS_SCT, para su posterior uso.

4.4.11 Carga de la tabla TD_INF_GEOGRAFICA con los datos de las tablas de trabajo.

Esta carga se realiza por medio de un script a ejecutar con SQLPlus, el cual se encarga de terminar de poblar la tabla TD_INF_GEOGRAFICA, empleando algunas de las tablas de trabajo. Concretamente, su ejecución realiza los siguientes pasos:

1) Se recorre la tabla de trabajo WORK_TABLE_DADES_VEHICLES y para cada uno de sus registros, se toma el valor del campo COD_INE y se comprueba si el municipio correspondiente está registrado en la tabla TD_ING_GEOGRAFICA. De no ser así se registra en TD_ING_GEOGRAFICA y se indica en el fichero log, ya que de este municipio no se tendrá ni su población ni su extensión.
2) Se recorre la tabla de trabajo WORK_TABLE_DADES_CONDUCTORS y para cada uno de sus registros, se toma el valor del campo MUNICIPIO y se comprueba si ya está registrado en la tabla TD_ING_GEOGRAFICA. Para llevar a cabo esta comprobación se realiza una búsqueda por proximidad. En caso de no estar registrado se da de alta en la tabla TD_INF_GEOGRAFICA y se indica en el fichero log, ya que de este municipio no se tendrá ni su población ni su extensión. En caso de estar registrado, se actualiza el registro correspondiente en TD_INF_GEOGRAFICA añadiendo la provincia.
3) Se recorre la tabla de trabajo WORK_TABLE_RADARS_SCT y para cada uno de sus registros, se toma el valor del campo MUNICIPIO y se comprueba si ya está registrado en la tabla TD_ING_GEOGRAFICA. Para llevar a cabo esta comprobación se realiza una búsqueda por proximidad. En caso de no estar registrado se da de alta en la tabla TD_INF_GEOGRAFICA y se indica en el fichero log, ya que de este municipio no se

tendrá ni su población ni su extensión. En caso de estar registrado, se actualiza el registro correspondiente en TD_INF_GEOGRAFICA añadiendo la vía, la comarca y la demarcación.

4.4.12 Carga de la tabla TH_EVOLUCION_TRAFICO con los datos de las tablas de trabajo.

Esta carga se realiza por medio de un script a ejecutar con SQLPlus, el cual se encarga de poblar la tabla TH_EVOLUCION_TRAFICO, empleando las tablas de trabajo y el resto de tablas del almacén de datos. Concretamente, su ejecución realiza los siguientes pasos:

1) Inserta en la tabla TH_EVOLUCION_TRAFICO tantos registros como el producto cartesiano del resto de tablas. De esta forma, se consiguen todos los registros necesarios en la tabla TH_EVOLUCION_TRAFICO y que todos ellos tengan valor en aquellos campos que representan claves foráneas (así como en la clave primaria). Dada esta situación, el resto del script sólo tendrá que hacer actualizaciones en los registros de la tabla TH_EVOLUCION_TRAFICO.
2) Se recorre la tabla de trabajo WORK_TABLE_DADES_MUNICIPIS y para cada uno de sus registros, se toma el valor del campo COD_INE y se obtiene el valor del campo ID_INF_GEOGRAFICA_PK del registro cuyo campo COD_INE coincida en valor con el COD_INE antes indicado. Después se actualiza la tabla TH_EVOLUCION_TRAFICO en aquellos registros donde el valor del campo ID_IND_GEOGRAFICA_FK coincida con el valor del campo ID_INF_GEOGRAFICA_PK antes obtenido y el valor del campo ID_TIEMPO_FK coincida con el valor del campo ANIO de la tabla de trabajo. En esta actualización se establece valor a los campos POBLACION y EXTENSION_MUNICIPIO con el valor de los campos correspondientes de la tabla WORK_TABLE_DADES_MUNICIPIS.
3) Se recorre la tabla de trabajo WORK_TABLE_DADES_VEHICLES y para cada uno de sus registros, se toma el valor del campo COD_INE y se obtiene el valor del campo ID_INF_GEOGRAFICA_PK del registro cuyo campo COD_INE coincida en valor con el COD_INE antes indicado. Después se actualiza la tabla TH_EVOLUCION_TRAFICO en aquellos registros donde el valor del campo ID_IND_GEOGRAFICA_FK coincida con el valor del campo ID_INF_GEOGRAFICA_PK antes obtenido, el valor de los campos ID_TIEMPO_FK y ID_VEHICULO_FK coincidan, respectivamente, con los valores de los campos ANIO y TIPO_VEHICULO de la tabla de trabajo. En esta actualización se establece valor a campo TOTAL_VEHICULOS con el valor del campo correspondiente de la tabla WORK_TABLE_DADES_VEHICLES.

   Como en esta tabla de trabajo se encuentran recogidos los datos origen correspondientes al fichero Dades_vehicles.xls es necesario tener en cuenta la siguiente circunstancia, para obtener los valores de “Motocicletas”, “Autobuses”, “Tractores industriales” y

"Resto vehículos de motor" correspondientes a los años 2011 al 2007 hay que tener en cuenta que la suma del valor de estas 4 columnas es igual al valor de la columna "Otros vehículos de motor 2012". Podemos asumir que para el resto de años los valores se distribuyen siguiendo la misma proporción y de esta forma inferir los valores de "Motocicletas", "Autobuses", "Tractores industriales" y "Resto vehículos de motor" correspondientes a los años 2011 al 2007, a partir de los valores de las columnas "Otros vehículos de motor 2011", "Otros vehículos de motor 2010", "Otros vehículos de motor 2009", "Otros vehículos de motor 2008" y "Otros vehículos de motor 2007". Es decir, para cada registro de esta tabla, además de lo indicado hay que hacer lo siguiente:

a) Se obtienen los porcentajes de participación para los tipos de vehículos "Motocicletas", "Autobuses", "Tractores industriales" y "Resto vehículos de motor", tal como se ha indicado anteriormente.
b) Para los años del 2007 al 2011 se obtiene el valor de cada uno de estos tipos de vehículos, a partir de su porcentaje de participación y el valor total, que se corresponde con el valor del tipo vehículo "Otros vehículos de motor" del año que se esté tratando. Con este valor, se actualiza la tabla TH_EVOLUCION_TRAFICO, teniendo en cuenta los valores de los campos ID_INF_GEOGRAFICA_FK, ID_VEHICULO_FK y ID_TIEMPO_FK, de la misma forma que se indicó en el primer párrafo del punto 3).

4.4.13 Tratamiento general de los datos erróneos.

En el tratamiento de los distintos ficheros de entrada se ha indicado que en algunas columnas de tipo numérico podrían encontrarse el valor "n.d." (No disponible). Esta circunstancia tiene dos posibles soluciones modificar estos valores en el fichero .csv o tratarla en el proceso de extracción, ejecutado en Oracle SQL*Loader. Para todos los ficheros en esta situación, se ha decidido emplear esta segunda opción, por lo tanto cuando desde el proceso de extracción se detecte el valor "n.d." se desechará el registro correspondiente, con la intención de que este valor no afecte al resultado de los informes. Además, en el informe de errores se indicará la información necesaria para localizar estos valores.

Para todos los ficheros con datos numéricos, si al tratar alguno de ellos se obtuviera una cadena de caracteres en lugar de un número, la línea donde se encuentre este valor se desecharía y se indicaría esta situación en el informe de errores.

4.4.14 Informe de errores

Tal como se ha comentado anteriormente, para realizar el proceso ETL nos apoyaremos en las herramientas Oracle SQL*Loader y SQLPlus. La primera de estas herramientas genera un fichero Log para cada script que ejecuta. En este log se indica, entre otros datos, mensajes de

error de los registros erróneos, los nombres de los ficheros de entrada que han generado algún error, el número de registros tratado y el número de registros erróneos.

En el caso de los scripts a ejecutarse con SQLPlus, el fichero log consiste en un fichero generado por cada script y contiene información de los registros tratados, registros erróneos y las circunstancias descritas en los apartados Carga de la tabla TD_INF_GEOGRAFICA con los datos de las tablas de trabajo y Carga de la tabla TH_EVOLUCION_TRAFICO con los datos de las tablas de trabajo.

Los detalles más precisos sobre el formato de los distintos ficheros log, así como, su ubicación física una vez generados se exponen en el apartado dedicado a tal efecto dentro del capítulo Implementación del sistema.

5. IMPLEMENTACIÓN DEL SISTEMA

En este apartado se presentan aspectos propios de la implentación del sistema que merecen ser destacados en la memoria, o bien añaden más detalle a elementos del diseño antes expuesto.

5.1 Creación de la base de datos.

En este apartado de describen los pasos necesarios para la creación física de la base de datos, tanto la creación de tablas como la de usuarios, esquemas, etc.

5.1.1 Tareas previas a la creación de la base de datos.

La base de datos empleada es Oracle 11g XE. En su instalación se crean por defecto los usuarios SYS y SYSTEM a los que se les ha asignado la password mmmd2013.

Con Oracle 11g XE se pueden crear workspaces dentro de los cuales crear esquemas, usuarios, etc. Por defecto, en la instalación se crea un workspace interno (llamado INTERNAL) y un usuario administrador del mismo. Este usuario es ADMIN y su password hay que resetearla[5], estableciendo su valor a Adm1n1$tr@d0r. Una vez hecho esto, es posible emplear este usuario para entrar en la aplicación web que incluye Oracle 11g XE y realizar tareas de administración. Esta aplicación está disponible a través de la URL:

http://localhost:8080/apex/apex_admin

[5] Para resetear esta password se han seguido los pasos indicados en https://plus.google.com/+DinoLopez/posts/aUfSA4CdrM7

Concretamente, las tareas de administración que se realizan son:

4) Crear el workspace AEW. Mediante este proceso se crean además los siguientes elementos.
    c) Creación del workspace se crea el esquema EstudiantDW, con el password DW2013.
    d) Se crea como administrador del workspace el usuario ADMIN, con el password ADMIN. La primera vez que se emplea este usuario se debe resetear su password estableciéndose al valor Adm1n1$tr@d0r

Además de la aplicación antes mencionada, Oracle 11g XE incluye una aplicación web que permite gestionar los distintos elementos de un workspace. Esta aplicación está disponible en la dirección:

http://localhost:8080/apex

Se accede a esta aplicación con el usuario administrador del workspace AEW para crear un usuario del workspace con perfil de desarrollador. Dicho usuario es AEW_EstudiantDW, con password AEW_DW2013 (la primera vez que se emplea es necesario resetear el password y se establece con el valor 4Ew_D%013). Una vez creado, el usuario de se emplea para acceder a esta misma aplicación y gestionar las tablas propias de la base de datos de nuestro sistema.

Resumiendo todo tenemos:

Elemento	Valor
Nombre de la base de datos	XE
Usuarios administradores del sistema	SYS y SYSTEM / mmmd2013
Usuario administrador de workspaces	ADMIN / Adm1n1$tr@d0r
Workspace	AEW
Esquema	EstudiantDW / DW2013
Administrador del workspace AEW	ADMIN / Adm1n1$tr@d0r
Usuario de desarrollo del workspace AEW	AEW_EstudiantDW / 4Ew_D%013

5.1.2 Creación de las tablas de la base de datos.

Una vez creado un esquema y un usuario en base de datos, el siguiente paso consiste en la creación de las tablas del modelo de datos del almacén. Además, como se vio en el apartado dedicado al diseño del proceso ETL, se han empleado una serie de tablas de trabajo con el objeto de facilitar la población de las tablas del almacén de datos durante dicho proceso. Para todo ello, se han creado una serie de scripts que permiten crear y eliminar tablas de la base de datos. Los ficheros correspondientes a estos scripts son los siguientes:

- script_creacion_bbdd.sql. Crea el modelo físico de datos formado por sus tablas, así como, las claves primarias y foráneas, y las distintas restricciones de cada tabla.

Además, en este script se crean las secuencias SEQ_INF_GEOGRAFICA y SEQ_EVOLUCION_TRAFICO, empleadas para dar valor a las claves primarias de las tablas TD_INF_GEOGRAFICA y TH_EVOLUCION_TRAFICO, respectivamente. Dentro de esta creación del modelo físico de datos podemos destacar los siguientes aspectos:

- Orden de creación de las tablas. En base al modelo de datos diseñado, es necesario que físicamente se creen primero las tablas de dimensiones (cuyo nombre empieza por TD_), puesto sólo tienen claves primarias y ninguna clave foránea. Por último, se crea la tabla de hechos (TH_EVOLUCION_TRAFICO) porque, además de tener clave primaria, tiene una clave foránea a cada una de las tablas de dimensiones.
- Índices y constraints. Los únicos índices y constraints creados en las tablas son los correspondientes a sus claves primarias. En el caso de la tabla TH_EVOLUCION_TRAFICO, existen además, las constraints propias de sus claves foráneas.

- script_creacion_work_tables.sql. Crea las tablas de trabajo empleadas en el proceso ETL. En este caso, a diferencia del script de creación de las tablas propias del almacén de datos, no es necesario establecer ningún orden en la creación de las tablas, ya que se trata de tablas temporales totalmente independientes, las cuales tampoco tienen índices ni constraints. La única excepción en este sentido es la tabla WORK_TABLE_DADES_CONDUCTORS, sobre la cual se crea una clave primaria formada por los campos PROVINCIA, MUNICIPIO y ANIO. Esta excepción se debe a que esta tabla se puebla mediante la ejecución de varios scripts y es necesario asegurar que no se incluyan en ella varios registros con el mismo valor en estos campos.
- script_destruccion_bbdd.sql. Elimina las tablas del modelo de datos y las secuencias SEQ_INF_GEOGRAFICA y SEQ_EVOLUCION_TRAFICO. Este script se ha creado por si fuera necesario crear la base de datos desde el principio.
- script_eliminacion_work_tables.sql. Elimina las tablas de trabajo. Este script puede ejecutarse una vez realizado el proceso ETL, ya que en ese momento las tablas de trabajo dejan de tener utilidad. De esta forma, se libera espacio en la base de datos.

Para poder ejecutar estos scripts desde la línea de comando es necesario situarse en el directorio donde se encuentran[6], y después ejecutar los siguientes comandos:

5) sqlplus /nolog
6) connect EstudiantDW/DW2013
7) @nom_fich /
8) disconnect
9) exit

[6] En el entorno virtual estos scripts se encuentran en el directorio c:\fmoraper\BBDD

El primer comando se ejecuta sobre la línea de comandos del sistema, el resto sobre la línea de comandos de SQLPlus. Los comandos 1) y 2) nos permiten conectarnos a la base de datos. El comando 3) se corresponde con la invocación al script y habría que indicar el nombre del fichero del script a ejecutar y finalizar la línea con el carácter "/". Los comandos 4) y 5) permiten terminar la conexión con la base de datos y salir de la línea de comandos de SQLPlus.

5.2 Implementación del proceso ETL.

En esta sección se detalla la implementación del proceso ETL basada en el diseño anteriormente detallado. Además, en esta sección se detalla la estructura de directorios empleada para guardar los distintos ficheros empleados en el proceso ETL. Se explican cada uno de los scripts creados, así como, la razón de su existencia. También se explica cómo deben ejecutarse los distintos scripts, todos juntos o bien por separado. Y para finalizar, se expone el formato de los ficheros .log y .bad, resultado de la ejecución de los scripts.

5.2.1 Estructura de directorios empleada.

Los ficheros empleados proceso ETL, así como, los generados en su ejecución se registran en la siguiente estructura de directorios:

- ETL. Directorio raíz desde donde cuelgan el resto de directorios. Además, en este directorio se encuentran los scripts ejecutados por "Oracle SQL* Loader" (ficheros control), así como, los scripts ejecutados por "SQLPlus" (scripts pl/sql). En este directorio, también se encuentra el fichero ETL.bat, que permite ejecutar todo el proceso ETL.
- ETL\data. En este directorio se encuentran los ficheros correspondientes a las bases de datos origen. También se encuentra el fichero "tipos_vehiculos.txt". En este fichero se registran los distintos tipos de datos, así como, un identificador único para cada uno de ellos. Estos datos sirven para poblar la tabla TD_VEHICULO y se encuentran en un fichero porque contienen acentos, por lo que, se necesita guardarlos en un fichero con formato UTF-8 (se dará más información al explicar el script que carga estos datos).
- ETL\log. Este directorio se emplea para guardar los ficheros .log generados por la ejecución de los distintos scripts.
- ETL\bad. Como se ha comentado anteriormente, "Oracle SQL*Plus" genera un fichero .bad al ejecutar un script, con los datos erróneos encontrados en dicha ejecución. Estos ficheros se guardan en este directorio.

5.2.2 Explicación de los scripts empleados.

El proceso ETL consta de scripts a ejecutar con "Oracle SQL* Loader" y scripts a ejecutar con "SQLPlus". Son los siguientes:

- scripts "Oracle SQL* Loader":
    - carga_tabla_TD_GENERO.ctl. Se encarga de registrar en la tabla TD_GENERO sus posibles valores (Masculino y Femenino). Estos datos están incluidos en el propio script. Este script se ejecuta en modo "REPLACE"[7], de forma que si la tabla tuviera datos, estos serían reemplazados.
    - carga_tabla_TD_VEHICULO.ctl. Se encarga de registrar en la tabla TD_VEHICULO sus posibles valores. Estos datos están incluidos en el fichero tipos_vehiculos.txt, ya que contienen acentos y es necesario registrarlos en un fichero con formato UTF-8. Se realizaron pruebas con los datos guardados en el propio script y al cual se le dio formato UTF-8. Sin embargo, estas pruebas no fueron satisfactorias, por lo que se optó por emplear un fichero separado con los datos. Este script se ejecuta en modo "REPLACE", de forma que si la tabla tuviera datos, estos serían reemplazados.
    - carga_tabla_TD_PERMISO.ctl. Se encarga de registrar en la tabla TD_PERMISO sus posibles valores (Permiso y Licencia). Estos datos están incluidos en el propio script. Este script se ejecuta en modo "REPLACE", de forma que si la tabla tuviera datos, estos serían reemplazados.
    - carga_tabla_TD_TIEMPO.ctl. Se encarga de registrar en la tabla TD_TIEMPO sus posibles valores, es decir, los años para los que tenemos datos (del 2007 al 2012). Estos datos están incluidos en el propio script. Este script se ejecuta en modo "REPLACE", de forma que si la tabla tuviera datos, estos serían reemplazados.
    - carga_fichero_Dades_municipis.ctl. Se encarga de registrar los datos del fichero "Dades_municipis.csv", el cual se encuentra en el subdirectorio "data". De los datos de este fichero, el script registra el nombre del municipio y el código INE (las dos primeras columnas) en la tabla TD_INF_GEOGRAFICA y el resto de datos en la tabla WORK_TABLE_DADES_MUNICIPIS. Este script se ejecuta en modo "REPLACE", de forma que si algún de las tablas tuviera datos, estos serían reemplazados.
    - carga_fichero_Dades_vehicles.ctl. Se encarga de registrar los datos del fichero "Dades_vehicles.csv", el cual se encuentra en el subdirectorio "data". Los datos de este fichero se registran en la tabla WORK_TABLE_DADES_VEHICLES. Este script se ejecuta en modo "REPLACE", de forma que si la tabla tuviera datos, estos serían reemplazados.

[7] "Oracle SQL* Loader" soporta varios tipos de ejecución de los scripts.

  - carga_ficheros_Dades_conductors_XXXX.ctl. Existe un fichero de control para cada uno de los años del 2011 al 2007. Cada uno de estos ficheros obtiene los datos del fichero Dades_conductors_XXXX.txt correspondiente, el cual se encuentra en el subdirectorio "data", y guarda los datos en la tabla de trabajo WORK_TABLE_DADES_CONDUCTOS. Además de estos datos, en la tabla de trabajo se registra el año correspondiente. La necesidad de tener un fichero de control (script) por cada fichero de datos, se debe a que es necesario registrar el año al que corresponden dichos datos. Podría haberse creado un único fichero de control y ejecutarlo varias veces, pasándole por parámetro el nombre del fichero de datos, pero no sería posible pasarle por parámetro el valor del año. La ejecución de estos script se realiza en modo APPEND, ya que al ser varios, es necesario que se vayan añadiendo los datos que obtiene cada uno. Para que no se registren datos repetidos en la tabla de trabajo (WORK_TABLE_DADES_CONDUCTORS), se han establecido como clave primaria de esta tabla los campos PROVINCIA, MUNICIPIO y ANIO.
  - carga_fichero_Radars_SCT.ctl. Se encarga de registrar los datos del fichero "Radars_SCT.txt", el cual se encuentra en el subdirectorio "data". Los datos de este fichero se registran en la tabla WORK_TABLE_RADARS_SCT. Este script se ejecuta en modo "REPLACE", de forma que si la tabla tuviera datos, estos serían reemplazados.
- scripts "SQLPlus":
  - carga_TD_INF_GEOGRAFICA_desde_WT.sql. Este script lee de distintas tablas de trabajo para terminar de poblar la tabla TD_INF_GEOGRAFICA. Genera un fichero log carga_TD_INF_GEOGRAFICA_desde_WT.log con información del resultado de la ejecución. Este fichero se guarda en el subdirectorio log.
  - carga_TH_EVOLUCION_TRAFICO_desde_WT.sql. Este script lee de distintas tablas para poblar la tabla TH_EVOLUCION_TRAFICO. Genera un fichero log carga_ TH_EVOLUCION_TRAFICO _desde_WT.log con información del resultado de la ejecución. Este fichero se guarda en el subdirectorio log.

5.2.3 Ejecución de los scripts.

Es necesario ejecutar los scripts "Oracle SQL* Loader" antes que los de "SQLPlus", porque los primeros registran datos en las tablas de trabajo y los segundos los obtienen de estas.

Para evitar problemas con los datos origen que incluyen acentos u otros caracteres especiales, los ficheros que almacenan estos datos han sido registrados con formato UTF8. Además, antes

de ejecutan cualquier script “Oracle SQL* Loader” es necesario ejecutar la siguiente sentencia desde la línea de comando:

```
set NLS_LANG=SPANISH_SPAIN.UTF8
```

Para poder ejecutar un script “Oracle SQL* Loader” desde la línea de comando hay que emplear la siguiente sentencia:

```
sqlldr USERID=EstudiantDW/DW2013 CONTROL=nom_fichero_control.ctl LOG=.\log\nom_fichero_control.log
```

Los script “SQLPlus” deben ejecutarse en este orden carga_TD_INF_GEOGRAFICA_desde_WT.sql, y después, carga_TH_EVOLUCION_TRAFICO_desde_WT.sql. Ya que el primero termina de poblar la tabla TD_INF_GEOGRAFICA y el segundo script obtiene datos de esta para poblar TH_EVOLUCION_TRAFICO.

Para poder ejecutar un script “SQLPlus” desde la línea de comando hay que emplear la siguiente sentencia:

```
exit | sqlplus EstudiantDW/DW2013 @carga_TD_INF_GEOGRAFICA_desde_WT.sql
```

Para facilitar la ejecución del proceso ETL se ha creado un fichero ETL.bat que realiza todos los pasos en el orden adecuado. Los comandos contenidos en este fichero ETL.bat son las siguientes:

```
set NLS_LANG=SPANISH_SPAIN.UTF8

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_tabla_TD_GENERO.ctl
LOG=.\log\carga_tabla_TD_GENERO.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_tabla_TD_VEHICULO.ctl
LOG=.\log\carga_tabla_TD_VEHICULO.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_tabla_TD_PERMISO.ctl
LOG=.\log\carga_tabla_TD_PERMISO.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_tabla_TD_TIEMPO.ctl
LOG=.\log\carga_tabla_TD_TIEMPO.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_fichero_Dades_municipis.ctl
LOG=.\log\carga_fichero_Dades_municipis.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_fichero_Dades_vehicles.ctl
LOG=.\log\carga_fichero_Dades_vehicles.log
```

```
sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_ficheros_Dades_conductors_2011.ctl
LOG=.\log\carga_ficheros_Dades_conductors_2011.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_ficheros_Dades_conductors_2010.ctl
LOG=.\log\carga_ficheros_Dades_conductors_2010.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_ficheros_Dades_conductors_2009.ctl
LOG=.\log\carga_ficheros_Dades_conductors_2009.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_ficheros_Dades_conductors_2008.ctl
LOG=.\log\carga_ficheros_Dades_conductors_2008.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_ficheros_Dades_conductors_2007.ctl
LOG=.\log\carga_ficheros_Dades_conductors_2007.log

sqlldr USERID=EstudiantDW/DW2013 CONTROL=carga_fichero_Radars_SCT.ctl
LOG=.\log\carga_fichero_Radars_SCT.log

exit | sqlplus EstudiantDW/DW2013 @carga_TD_INF_GEOGRAFICA_desde_WT.sql

exit | sqlplus EstudiantDW/DW2013 @carga_TH_EVOLUCION_TRAFICO_desde_WT.sql
```

5.2.3.1 Automatización de la ejecución.

La automatización de la ejecución del proceso ETL según se modificaran los datos origen no es posible, ya que a los distintos ficheros que contienen estos datos se les ha dado un tratamiento, que si bien es liviano también es necesario. En cualquier caso, como se ha comentado en la sección de ejecución de los scripts, existe el fichero ETL.bat que facilita la ejecución del proceso.

5.2.4 Ficheros con el resultado de la ejecución; Ficheros log y bad.

Como resultado de los distintos scripts del proceso ETL se obtiene un conjunto de ficheros log y bad. A continuación se detalla la información contenida en ellos.

- Scripts “Oracle SQL* Loader”. En este caso, los ficheros son generados por la propia herramienta, por lo que no es posible establecer otro formato.
    - Ficheros log. En estos ficheros se incluye un resumen del número de registros cargados correctamente, así como, del número de registros que han producido algún error. Para este último caso, se indica el número del registro y el error encontrado, con lo que se facilita su detección en el fichero de datos de entrada. Durante la fase de análisis se comprobó que en algunos ficheros de entrada existía el valor “n.d.”, en lugar de un valor numérico. Esta situación sería

detectada en la ejecución de estos scripts y, por tanto, reportada en el fichero log.
    - Ficheros bad. En estos ficheros se guardan los registros que han producido algún error.

  Podría decirse que los ficheros log y bad trabajan conjuntamente. En los ficheros log se guardan el número de los registros de entrada que producen algún error, y en los ficheros bad se guardan estos registros.
- Scripts "SQLPlus". Estos scripts generan un fichero log con el mismo nombre que el script (sustituyendo las extensiones sql por log). Como su cometido es obtener datos de las tablas de trabajo y guardar estos datos en tablas del almacén de datos, la información registrada en los ficheros log está separada por secciones en donde se marca el inicio y el fin del tratamiento que se da a los registros de cada tabla de trabajo, así como el número de registros tratados y el número de registros con algún error. Concretando por script, la información que se registra en el log es la siguiente:
    - carga_TD_INF_GEOGRAFICA_desde_WT.sql.
        - Municipios registrados en Dades_vehicles, que no se encuentra en Dades_municipis.
        - Registros del fichero Dades_conductors cuyo municipio es el literal "Municipio sin especificar".
        - Municipios registrados en Dades_conductors, que no se encuentra en Dades_municipis.
        - Municipios registrados en Radars_SCT, que no se encuentra en Dades_municipis.
    - carga_TH_EVOLUCION_TRAFICO_desde_WT.sql
        - Número de registros insertados en la tabla TH_EVOLUCION_TRAFICO.

Para cada registro de la tabla TH_EVOLUCION_TRAFICO en el que se ha producido un error al actualizarlo, se indica el municipio correspondiente a los datos con los que se pretendía actualizar la tabla, así como, dichos datos.

5.3 Implementación de los distintos informes.

La implementación de los informes ha sido realizada empleando la herramienta Visual Studio 2010, si bien una vez terminados se realiza un despliegue de los mismos en Microsoft Reporter Server, y podrán ser ejecutadas con un navegador web por medio de la URL de SQL Server Reporting Services. El diseño de distintos informes se verá en el siguiente apartado.

6. INFORMES DEL SISTEMA

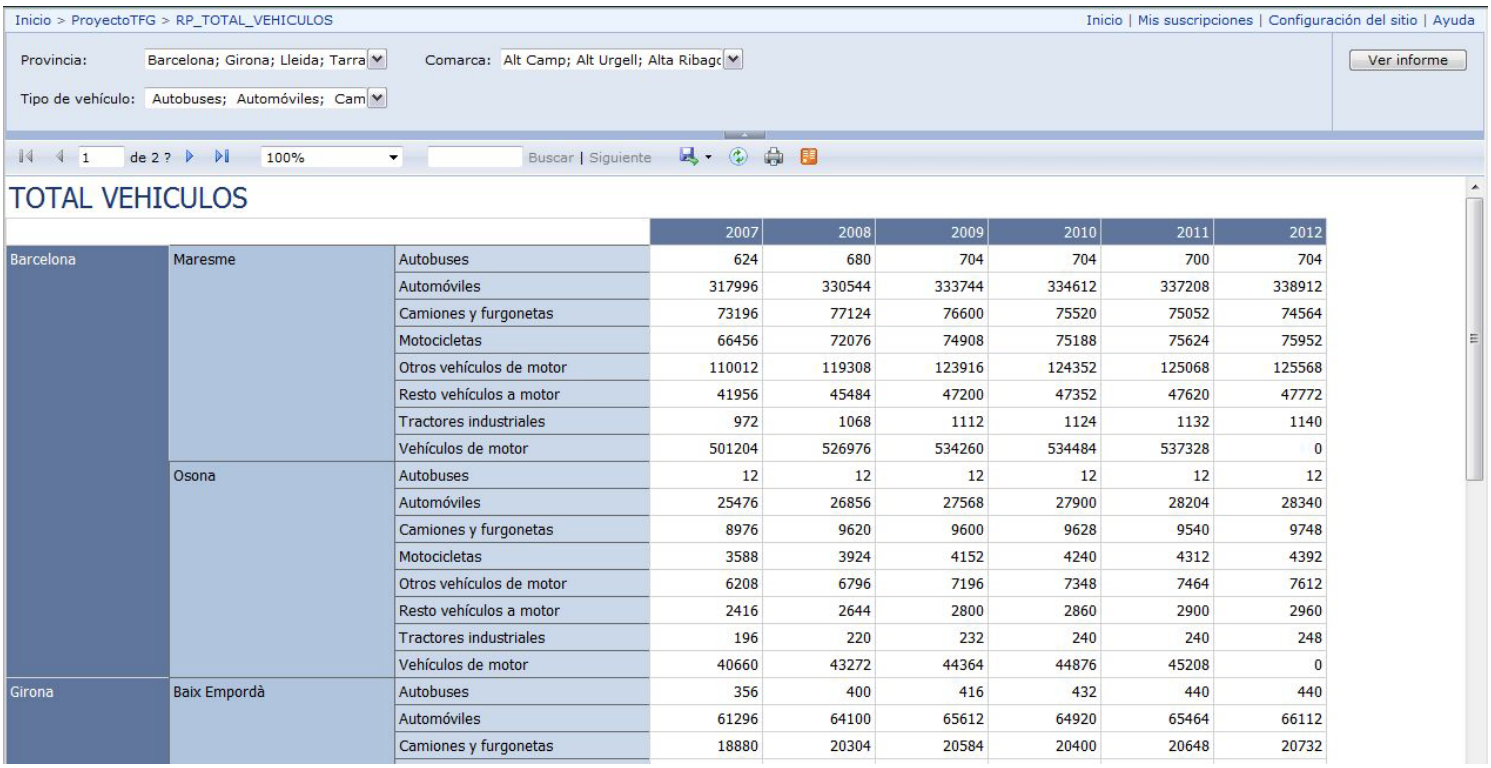
Si bien cada informe tiene su propio formato, en todos ellos la temporalidad de los datos es a nivel de año y todos tienen un conjunto de parámetros que permiten filtrar la información. Este conjunto de parámetros depende de la información mostrada por cada informe, no obstante, todos los parámetros permiten seleccionar uno o varios de sus valores. En todos los informes se incluyen los parámetros provincia y comarca. Estos parámetros están sincronizados de manera que primero hay que seleccionar una o varias provincias, de forma que, en el parámetro comarca sólo se podrán seleccionar comarcas correspondientes a las provincias previamente seleccionadas.

A continuación se muestran capturas de pantalla con ejemplos de la ejecución de cada informe.

6.1 Total de vehículos

Este informe muestra el número total de vehículos en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca y tipo de vehículo. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



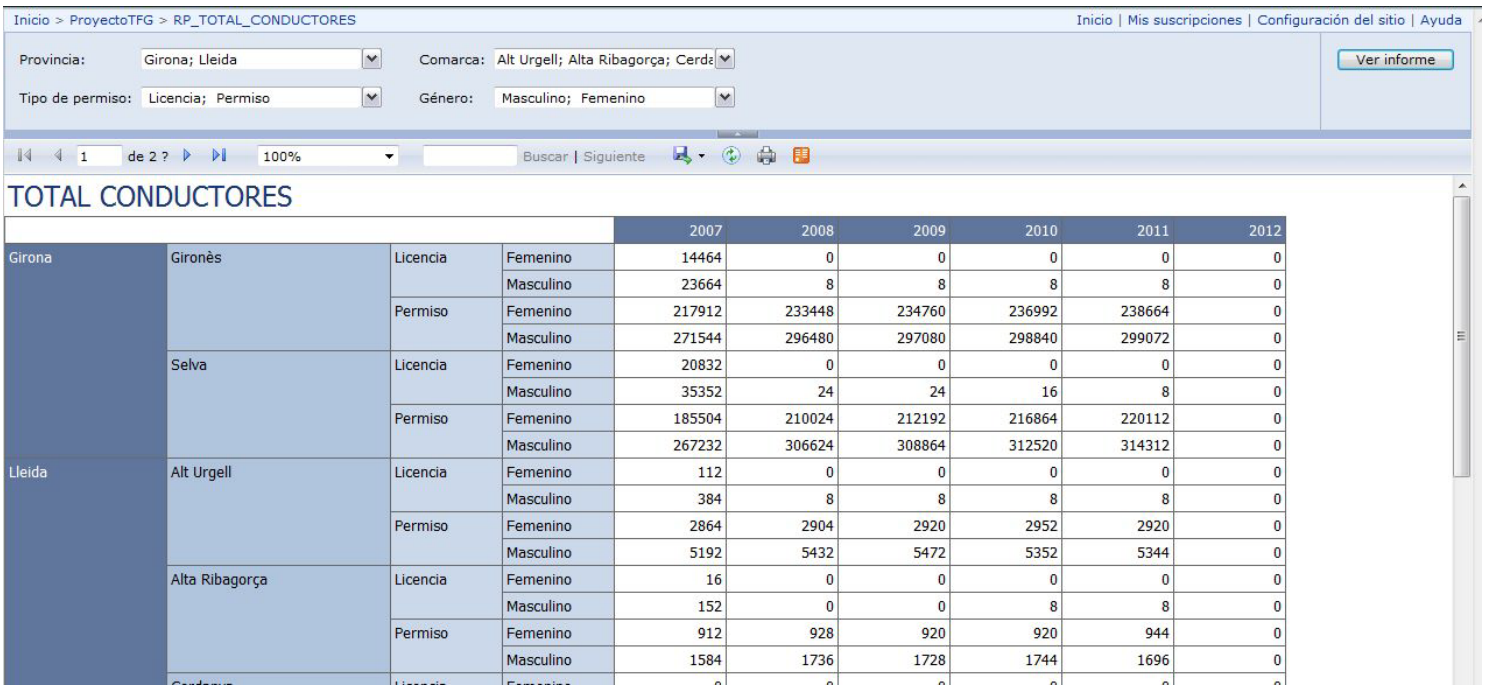
TOTAL VEHICULOS

			2007	2008	2009	2010	2011	2012
Barcelona	Maresme	Autobuses	624	680	704	704	700	704
		Automóviles	317996	330544	333744	334612	337208	338912
		Camiones y furgonetas	73196	77124	76600	75520	75052	74564
		Motocicletas	66456	72076	74908	75188	75624	75952
		Otros vehículos de motor	110012	119308	123916	124352	125068	125568
		Resto vehículos a motor	41956	45484	47200	47352	47620	47772
		Tractores industriales	972	1068	1112	1124	1132	1140
		Vehículos de motor	501204	526976	534260	534484	537328	0
	Osona	Autobuses	12	12	12	12	12	12
		Automóviles	25476	26856	27568	27900	28204	28340
		Camiones y furgonetas	8976	9620	9600	9628	9540	9748
		Motocicletas	3588	3924	4152	4240	4312	4392
		Otros vehículos de motor	6208	6796	7196	7348	7464	7612
		Resto vehículos a motor	2416	2644	2800	2860	2900	2960
		Tractores industriales	196	220	232	240	240	248
		Vehículos de motor	40660	43272	44364	44876	45208	0
Girona	Baix Empordà	Autobuses	356	400	416	432	440	440
		Automóviles	61296	64100	65612	64920	65464	66112
		Camiones y furgonetas	18880	20304	20584	20400	20648	20732

6.2 Total de conductores

Este informe muestra el número total de conductores en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca, tipo de permiso y género. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



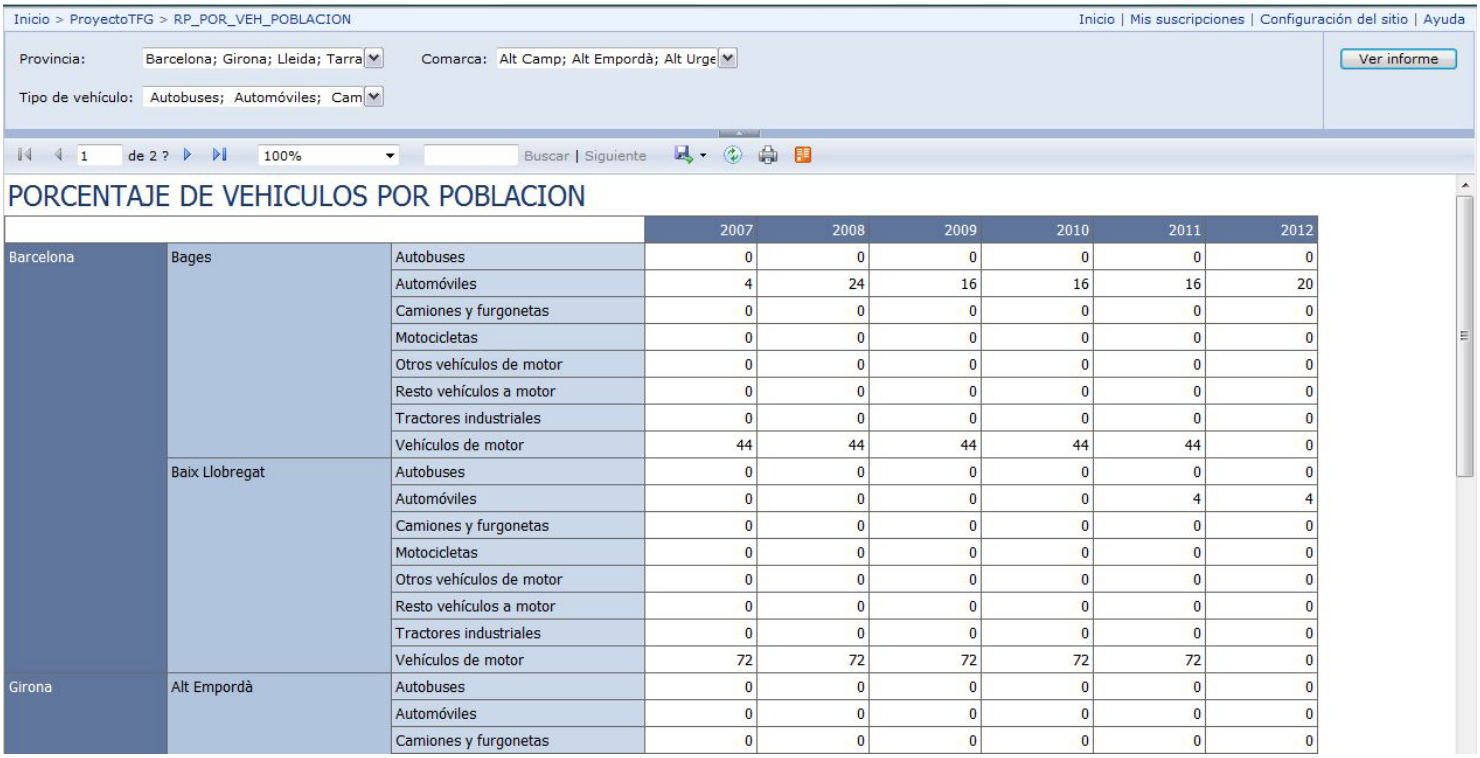
TOTAL CONDUCTORES

				2007	2008	2009	2010	2011	2012
Girona	Gironès	Licencia	Femenino	14464	0	0	0	0	0
			Masculino	23664	8	8	8	8	0
		Permiso	Femenino	217912	233448	234760	236992	238664	0
			Masculino	271544	296480	297080	298840	299072	0
	Selva	Licencia	Femenino	20832	0	0	0	0	0
			Masculino	35352	24	24	16	8	0
		Permiso	Femenino	185504	210024	212192	216864	220112	0
			Masculino	267232	306624	308864	312520	314312	0
Lleida	Alt Urgell	Licencia	Femenino	112	0	0	0	0	0
			Masculino	384	8	8	8	8	0
		Permiso	Femenino	2864	2904	2920	2952	2920	0
			Masculino	5192	5432	5472	5352	5344	0
	Alta Ribagorça	Licencia	Femenino	16	0	0	0	0	0
			Masculino	152	0	0	8	8	0
		Permiso	Femenino	912	928	920	920	944	0
			Masculino	1584	1736	1728	1744	1696	0

6.3 Porcentaje de vehículos respecto a la población

Este informe muestra el porcentaje de vehículos respecto a la población en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca y tipo de vehículo. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



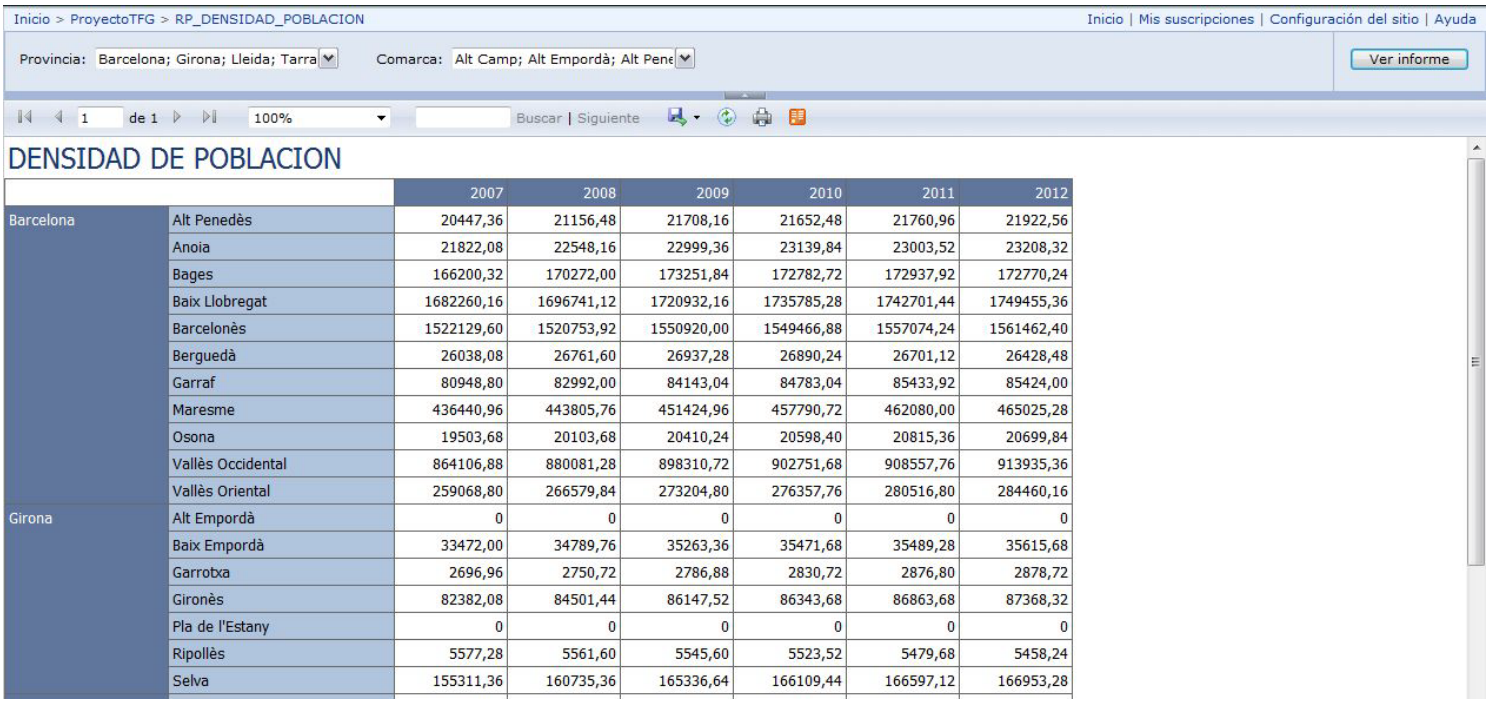
PORCENTAJE DE VEHICULOS POR POBLACION

			2007	2008	2009	2010	2011	2012
Barcelona	Bages	Autobuses	0	0	0	0	0	0
		Automóviles	4	24	16	16	16	20
		Camiones y furgonetas	0	0	0	0	0	0
		Motocicletas	0	0	0	0	0	0
		Otros vehículos de motor	0	0	0	0	0	0
		Resto vehículos a motor	0	0	0	0	0	0
		Tractores industriales	0	0	0	0	0	0
		Vehículos de motor	44	44	44	44	44	0
	Baix Llobregat	Autobuses	0	0	0	0	0	0
		Automóviles	0	0	0	0	4	4
		Camiones y furgonetas	0	0	0	0	0	0
		Motocicletas	0	0	0	0	0	0
		Otros vehículos de motor	0	0	0	0	0	0
		Resto vehículos a motor	0	0	0	0	0	0
		Tractores industriales	0	0	0	0	0	0
		Vehículos de motor	72	72	72	72	72	0
Girona	Alt Empordà	Autobuses	0	0	0	0	0	0
		Automóviles	0	0	0	0	0	0
		Camiones y furgonetas	0	0	0	0	0	0

6.4 Densidad de población (habitantes/km2)

Este informe muestra la densidad de población en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia y comarca. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



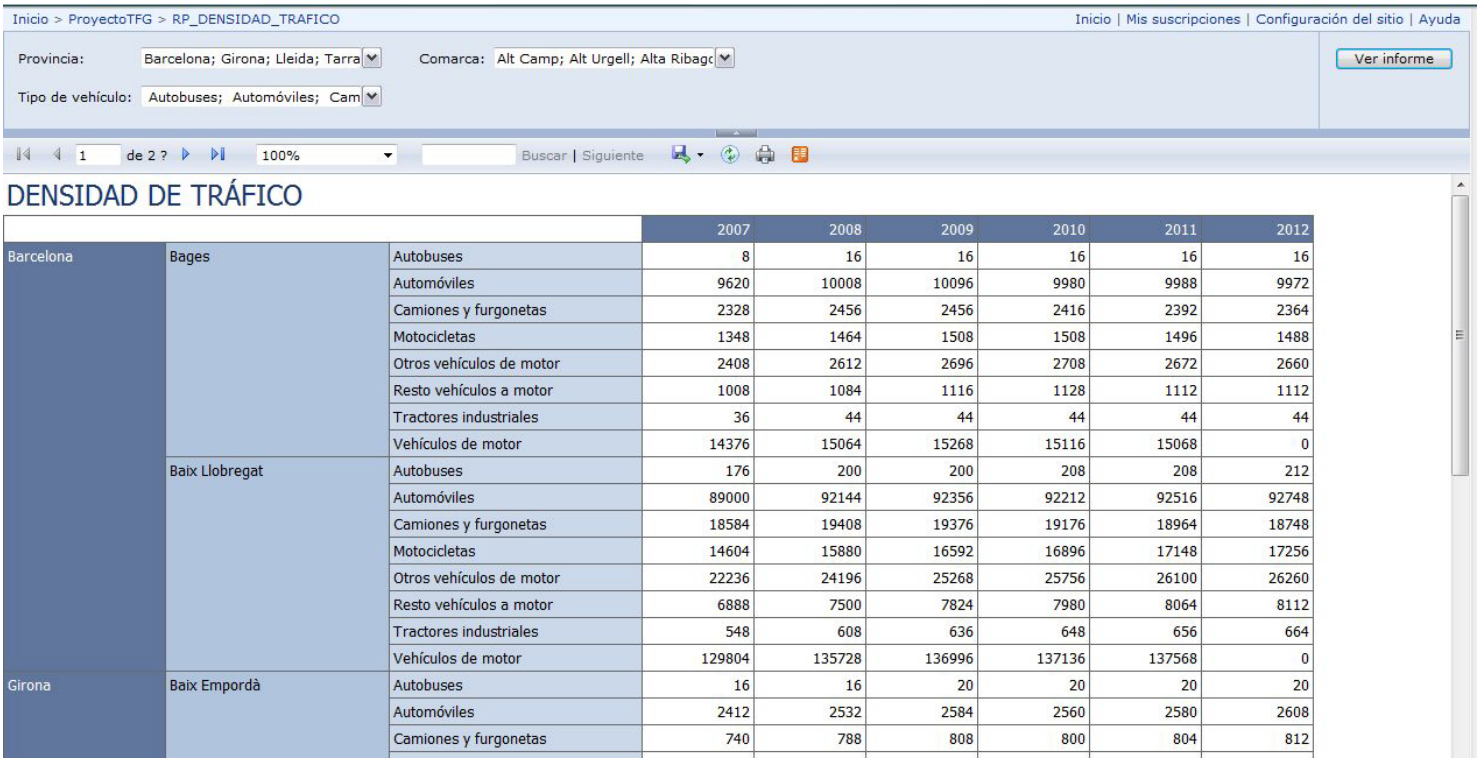
DENSIDAD DE POBLACION

		2007	2008	2009	2010	2011	2012
Barcelona	Alt Penedès	20447,36	21156,48	21708,16	21652,48	21760,96	21922,56
	Anoia	21822,08	22548,16	22999,36	23139,84	23003,52	23208,32
	Bages	166200,32	170272,00	173251,84	172782,72	172937,92	172770,24
	Baix Llobregat	1682260,16	1696741,12	1720932,16	1735785,28	1742701,44	1749455,36
	Barcelonès	1522129,60	1520753,92	1550920,00	1549466,88	1557074,24	1561462,40
	Berguedà	26038,08	26761,60	26937,28	26890,24	26701,12	26428,48
	Garraf	80948,80	82992,00	84143,04	84783,04	85433,92	85424,00
	Maresme	436440,96	443805,76	451424,96	457790,72	462080,00	465025,28
	Osona	19503,68	20103,68	20410,24	20598,40	20815,36	20699,84
	Vallès Occidental	864106,88	880081,28	898310,72	902751,68	908557,76	913935,36
	Vallès Oriental	259068,80	266579,84	273204,80	276357,76	280516,80	284460,16
Girona	Alt Empordà	0	0	0	0	0	0
	Baix Empordà	33472,00	34789,76	35263,36	35471,68	35489,28	35615,68
	Garrotxa	2696,96	2750,72	2786,88	2830,72	2876,80	2878,72
	Gironès	82382,08	84501,44	86147,52	86343,68	86863,68	87368,32
	Pla de l'Estany	0	0	0	0	0	0
	Ripollès	5577,28	5561,60	5545,60	5523,52	5479,68	5458,24
	Selva	155311,36	160735,36	165336,64	166109,44	166597,12	166953,28

6.5 Densidad de tráfico (vehículos/km2)

Este informe muestra la densidad de tráfico en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca y tipo de vehículo. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



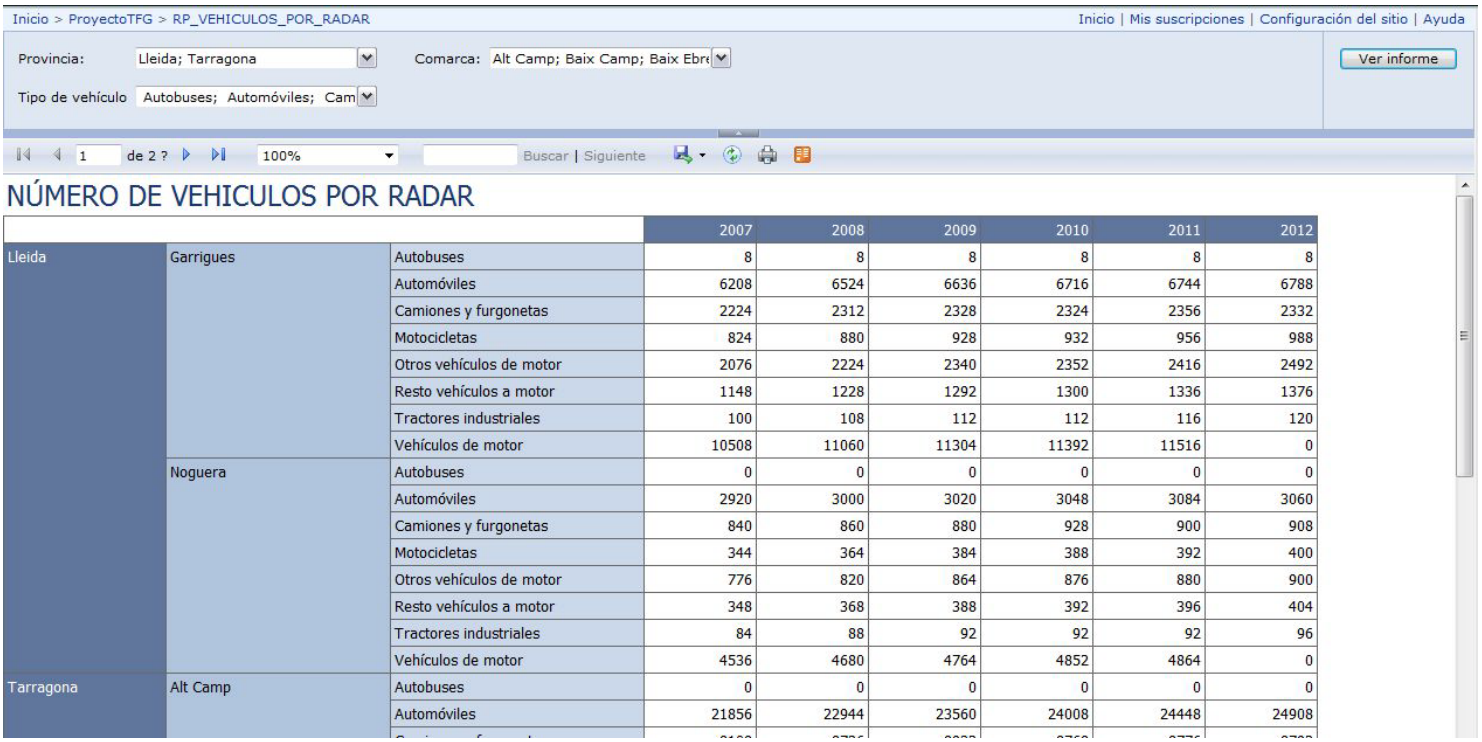
DENSIDAD DE TRÁFICO

			2007	2008	2009	2010	2011	2012
Barcelona	Bages	Autobuses	8	16	16	16	16	16
		Automóviles	9620	10008	10096	9980	9988	9972
		Camiones y furgonetas	2328	2456	2456	2416	2392	2364
		Motocicletas	1348	1464	1508	1508	1496	1488
		Otros vehículos de motor	2408	2612	2696	2708	2672	2660
		Resto vehículos a motor	1008	1084	1116	1128	1112	1112
		Tractores industriales	36	44	44	44	44	44
		Vehículos de motor	14376	15064	15268	15116	15068	0
	Baix Llobregat	Autobuses	176	200	200	208	208	212
		Automóviles	89000	92144	92356	92212	92516	92748
		Camiones y furgonetas	18584	19408	19376	19176	18964	18748
		Motocicletas	14604	15880	16592	16896	17148	17256
		Otros vehículos de motor	22236	24196	25268	25756	26100	26260
		Resto vehículos a motor	6888	7500	7824	7980	8064	8112
		Tractores industriales	548	608	636	648	656	664
		Vehículos de motor	129804	135728	136996	137136	137568	0
Girona	Baix Empordà	Autobuses	16	16	20	20	20	20
		Automóviles	2412	2532	2584	2560	2580	2608
		Camiones y furgonetas	740	788	808	800	804	812

6.6 Número de vehículos respecto al número de radares

Este informe muestra la cantidad de vehículos por radar en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca y tipo de vehículo. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



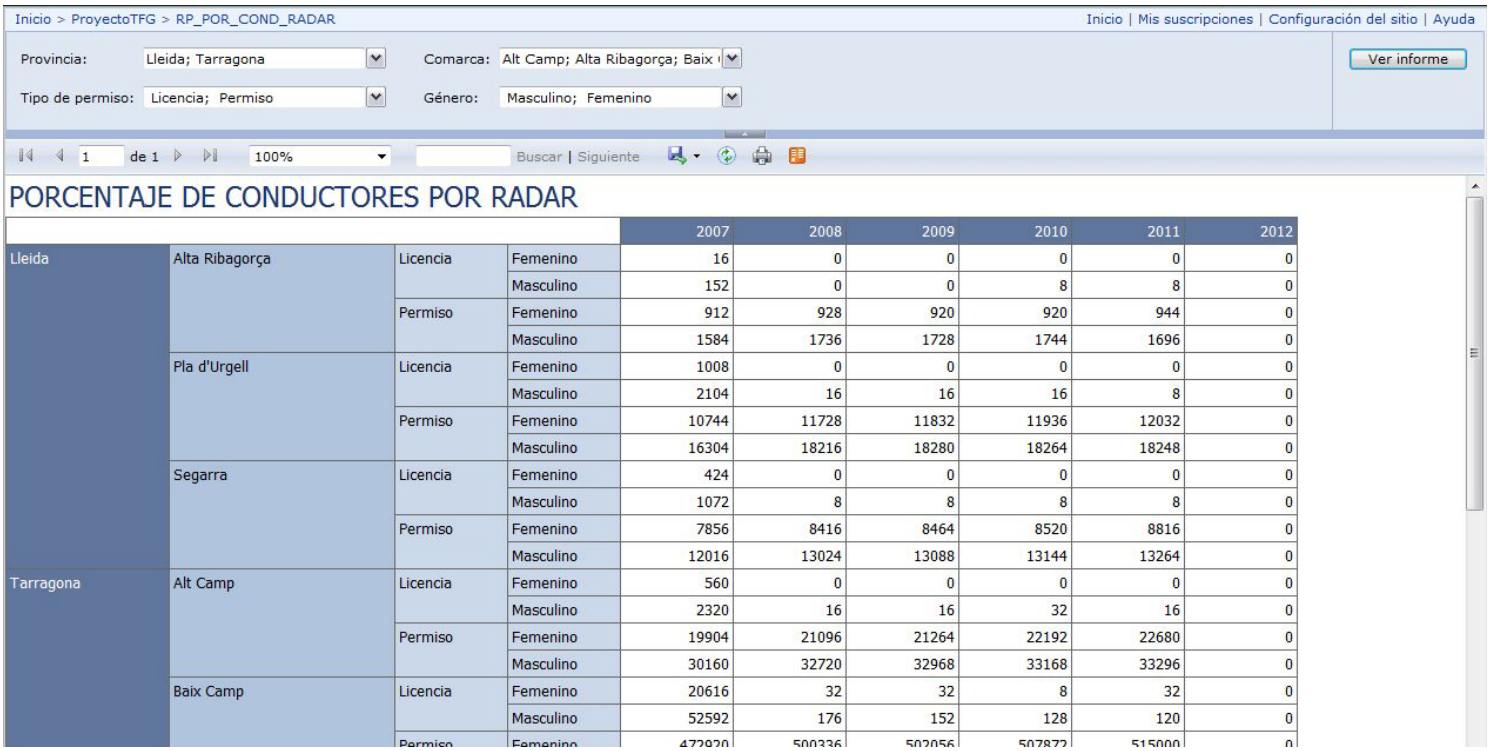
NÚMERO DE VEHICULOS POR RADAR

			2007	2008	2009	2010	2011	2012
Lleida	Garrigues	Autobuses	8	8	8	8	8	8
		Automóviles	6208	6524	6636	6716	6744	6788
		Camiones y furgonetas	2224	2312	2328	2324	2356	2332
		Motocicletas	824	880	928	932	956	988
		Otros vehículos de motor	2076	2224	2340	2352	2416	2492
		Resto vehículos a motor	1148	1228	1292	1300	1336	1376
		Tractores industriales	100	108	112	112	116	120
		Vehículos de motor	10508	11060	11304	11392	11516	0
	Noguera	Autobuses	0	0	0	0	0	0
		Automóviles	2920	3000	3020	3048	3084	3060
		Camiones y furgonetas	840	860	880	928	900	908
		Motocicletas	344	364	384	388	392	400
		Otros vehículos de motor	776	820	864	876	880	900
		Resto vehículos a motor	348	368	388	392	396	404
		Tractores industriales	84	88	92	92	92	96
		Vehículos de motor	4536	4680	4764	4852	4864	0
Tarragona	Alt Camp	Autobuses	0	0	0	0	0	0
		Automóviles	21856	22944	23560	24008	24448	24908

6.7 Porcentaje de conductores por radar

Este informe muestra el porcentaje de conductores por radar en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca tipo de permiso y género. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



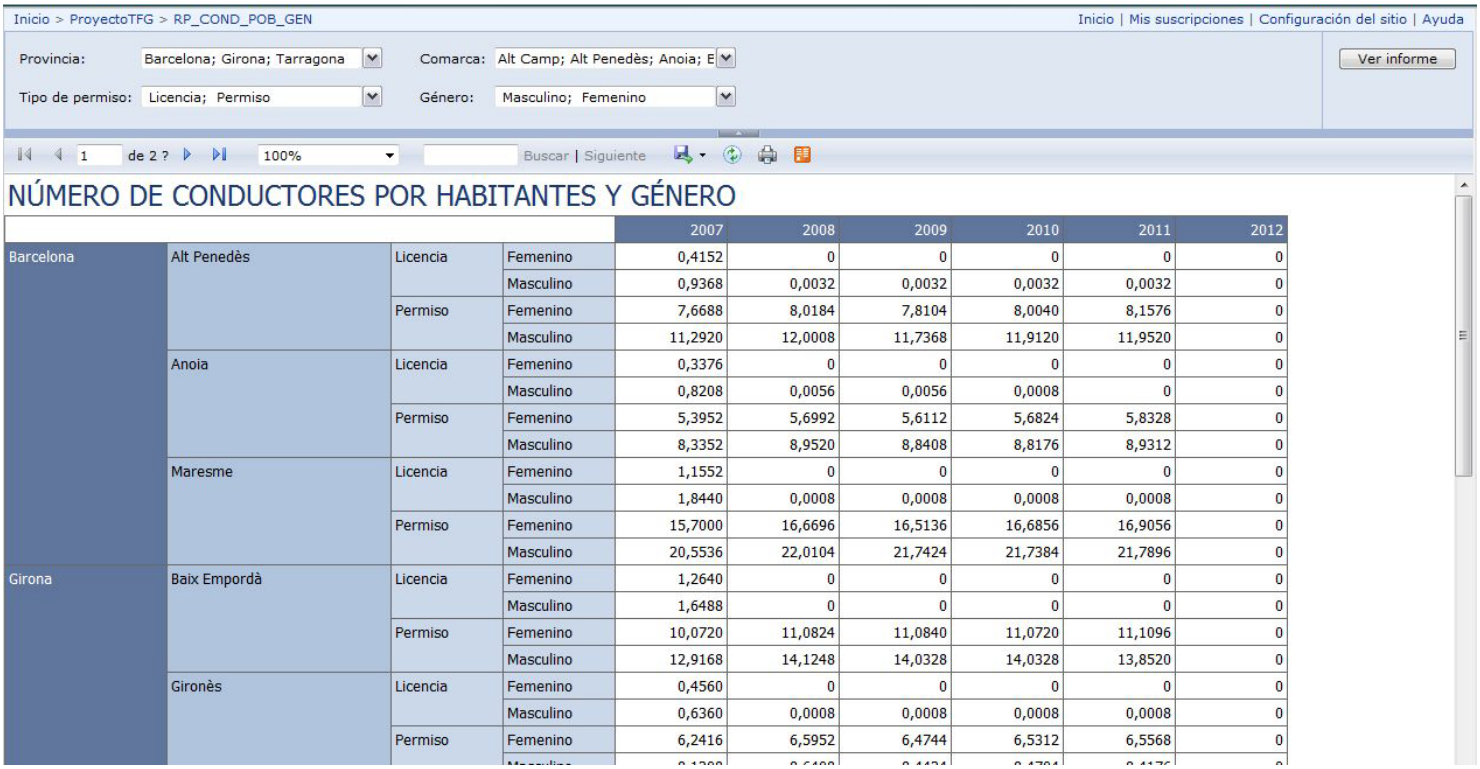
PORCENTAJE DE CONDUCTORES POR RADAR

				2007	2008	2009	2010	2011	2012
Lleida	Alta Ribagorça	Licencia	Femenino	16	0	0	0	0	0
			Masculino	152	0	0	8	8	0
		Permiso	Femenino	912	928	920	920	944	0
			Masculino	1584	1736	1728	1744	1696	0
	Pla d'Urgell	Licencia	Femenino	1008	0	0	0	0	0
			Masculino	2104	16	16	16	8	0
		Permiso	Femenino	10744	11728	11832	11936	12032	0
			Masculino	16304	18216	18280	18264	18248	0
	Segarra	Licencia	Femenino	424	0	0	0	0	0
			Masculino	1072	8	8	8	8	0
		Permiso	Femenino	7856	8416	8464	8520	8816	0
			Masculino	12016	13024	13088	13144	13264	0
Tarragona	Alt Camp	Licencia	Femenino	560	0	0	0	0	0
			Masculino	2320	16	16	32	16	0
		Permiso	Femenino	19904	21096	21264	22192	22680	0
			Masculino	30160	32720	32968	33168	33296	0
	Baix Camp	Licencia	Femenino	20616	32	32	8	32	0
			Masculino	52592	176	152	128	120	0
		Permiso	Femenino	472920	500336	502056	507872	515000	0

6.8 Indicador de conductores vs habitantes por género

Este informe muestra el porcentaje de conductores por habitantes según su género en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca tipo de permiso y género. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



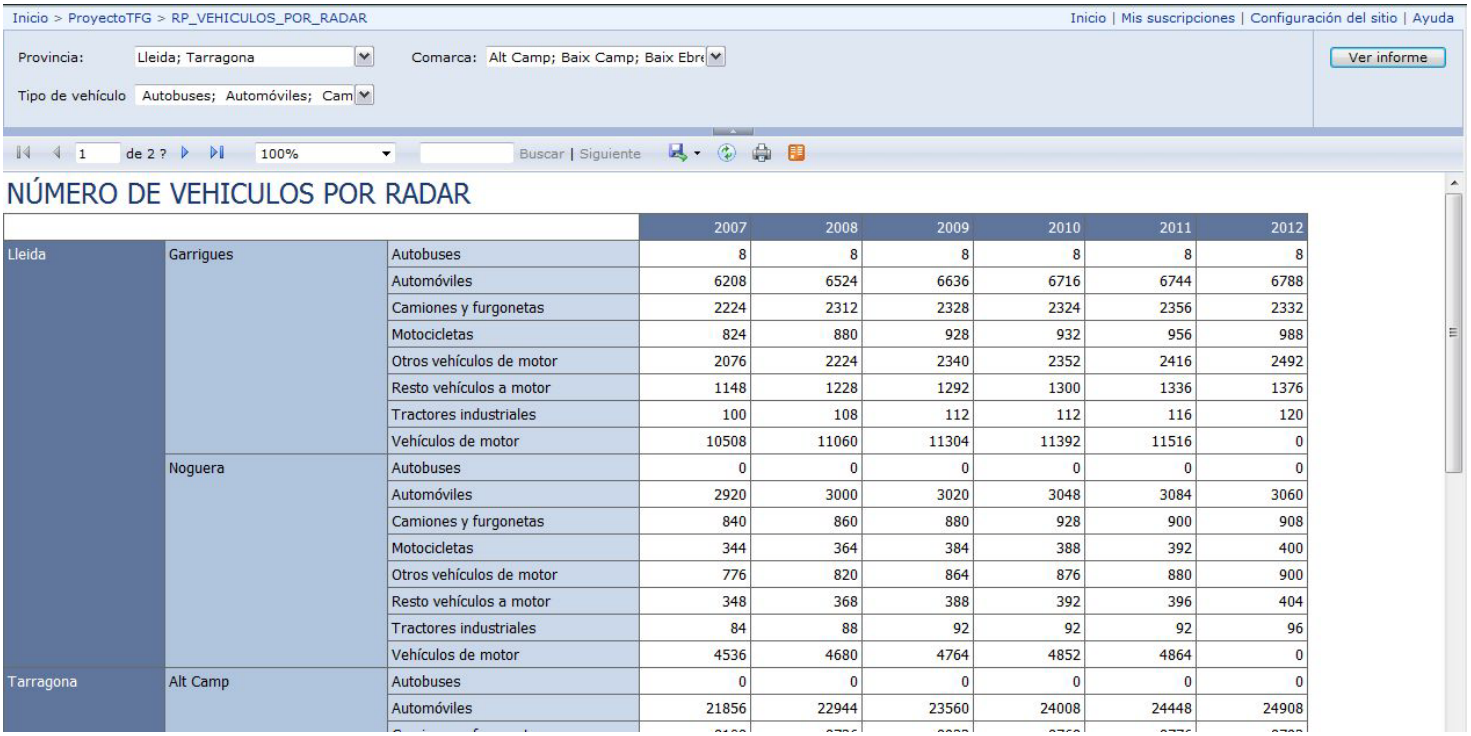
NÚMERO DE CONDUCTORES POR HABITANTES Y GÉNERO

				2007	2008	2009	2010	2011	2012
Barcelona	Alt Penedès	Licencia	Femenino	0,4152	0	0	0	0	0
			Masculino	0,9368	0,0032	0,0032	0,0032	0,0032	0
		Permiso	Femenino	7,6688	8,0184	7,8104	8,0040	8,1576	0
			Masculino	11,2920	12,0008	11,7368	11,9120	11,9520	0
	Anoia	Licencia	Femenino	0,3376	0	0	0	0	0
			Masculino	0,8208	0,0056	0,0056	0,0008	0	0
		Permiso	Femenino	5,3952	5,6992	5,6112	5,6824	5,8328	0
			Masculino	8,3352	8,9520	8,8408	8,8176	8,9312	0
	Maresme	Licencia	Femenino	1,1552	0	0	0	0	0
			Masculino	1,8440	0,0008	0,0008	0,0008	0,0008	0
		Permiso	Femenino	15,7000	16,6696	16,5136	16,6856	16,9056	0
			Masculino	20,5536	22,0104	21,7424	21,7384	21,7896	0
Girona	Baix Empordà	Licencia	Femenino	1,2640	0	0	0	0	0
			Masculino	1,6488	0	0	0	0	0
		Permiso	Femenino	10,0720	11,0824	11,0840	11,0720	11,1096	0
			Masculino	12,9168	14,1248	14,0328	14,0328	13,8520	0
	Gironès	Licencia	Femenino	0,4560	0	0	0	0	0
			Masculino	0,6360	0,0008	0,0008	0,0008	0,0008	0
		Permiso	Femenino	6,2416	6,5952	6,4744	6,5312	6,5568	0

6.9 Indicador de radares vs vehículos

Este informe muestra la cantidad de radares por vehículo en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca y tipo de vehículo. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



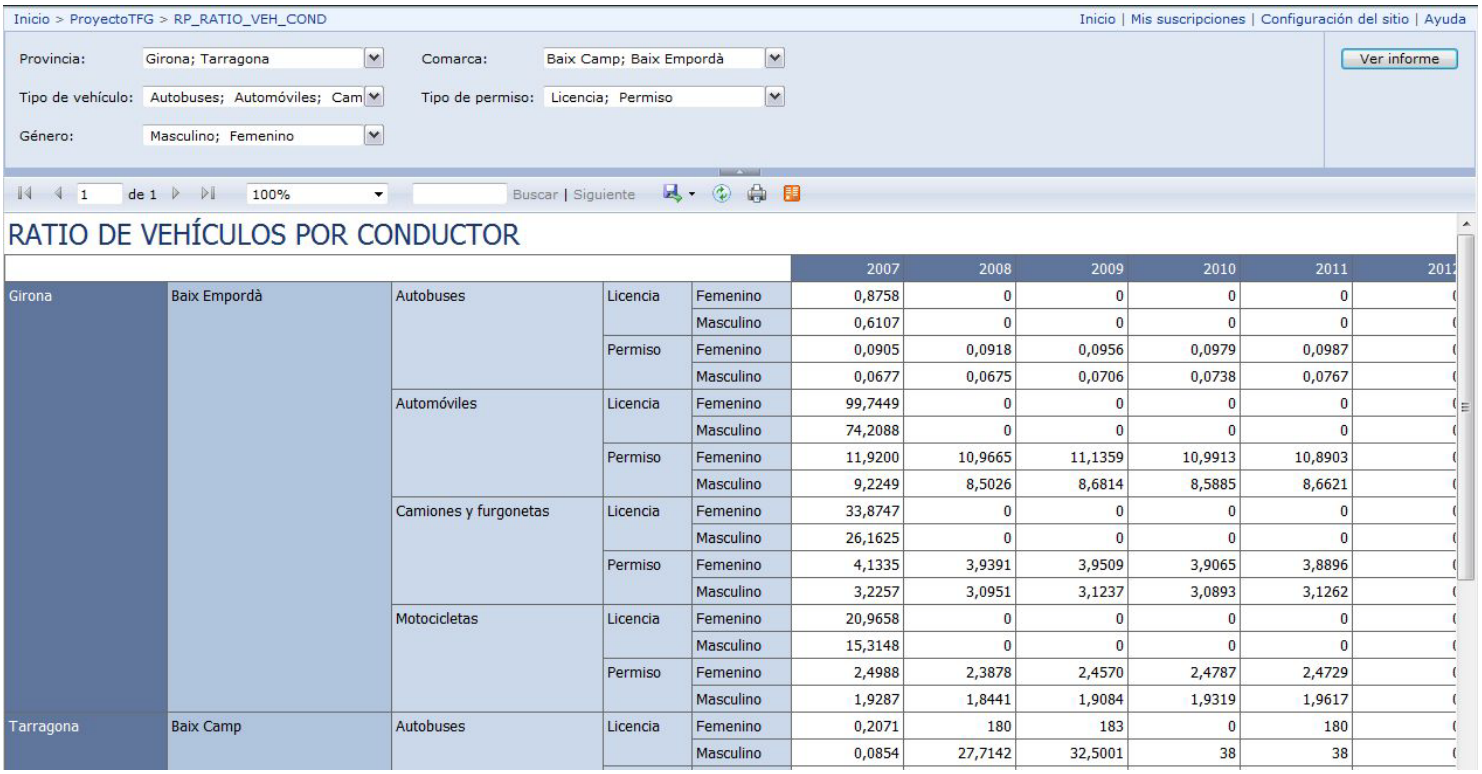
NÚMERO DE VEHICULOS POR RADAR

			2007	2008	2009	2010	2011	2012
Lleida	Garrigues	Autobuses	8	8	8	8	8	8
		Automóviles	6208	6524	6636	6716	6744	6788
		Camiones y furgonetas	2224	2312	2328	2324	2356	2332
		Motocicletas	824	880	928	932	956	988
		Otros vehículos de motor	2076	2224	2340	2352	2416	2492
		Resto vehículos a motor	1148	1228	1292	1300	1336	1376
		Tractores industriales	100	108	112	112	116	120
		Vehículos de motor	10508	11060	11304	11392	11516	0
	Noguera	Autobuses	0	0	0	0	0	0
		Automóviles	2920	3000	3020	3048	3084	3060
		Camiones y furgonetas	840	860	880	928	900	908
		Motocicletas	344	364	384	388	392	400
		Otros vehículos de motor	776	820	864	876	880	900
		Resto vehículos a motor	348	368	388	392	396	404
		Tractores industriales	84	88	92	92	92	96
		Vehículos de motor	4536	4680	4764	4852	4864	0
Tarragona	Alt Camp	Autobuses	0	0	0	0	0	0
		Automóviles	21856	22944	23560	24008	24448	24908

6.10 Ratio de vehículos por conductor

Este informe muestra el ratio de vehículos por conductor en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca tipo de permiso y género. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



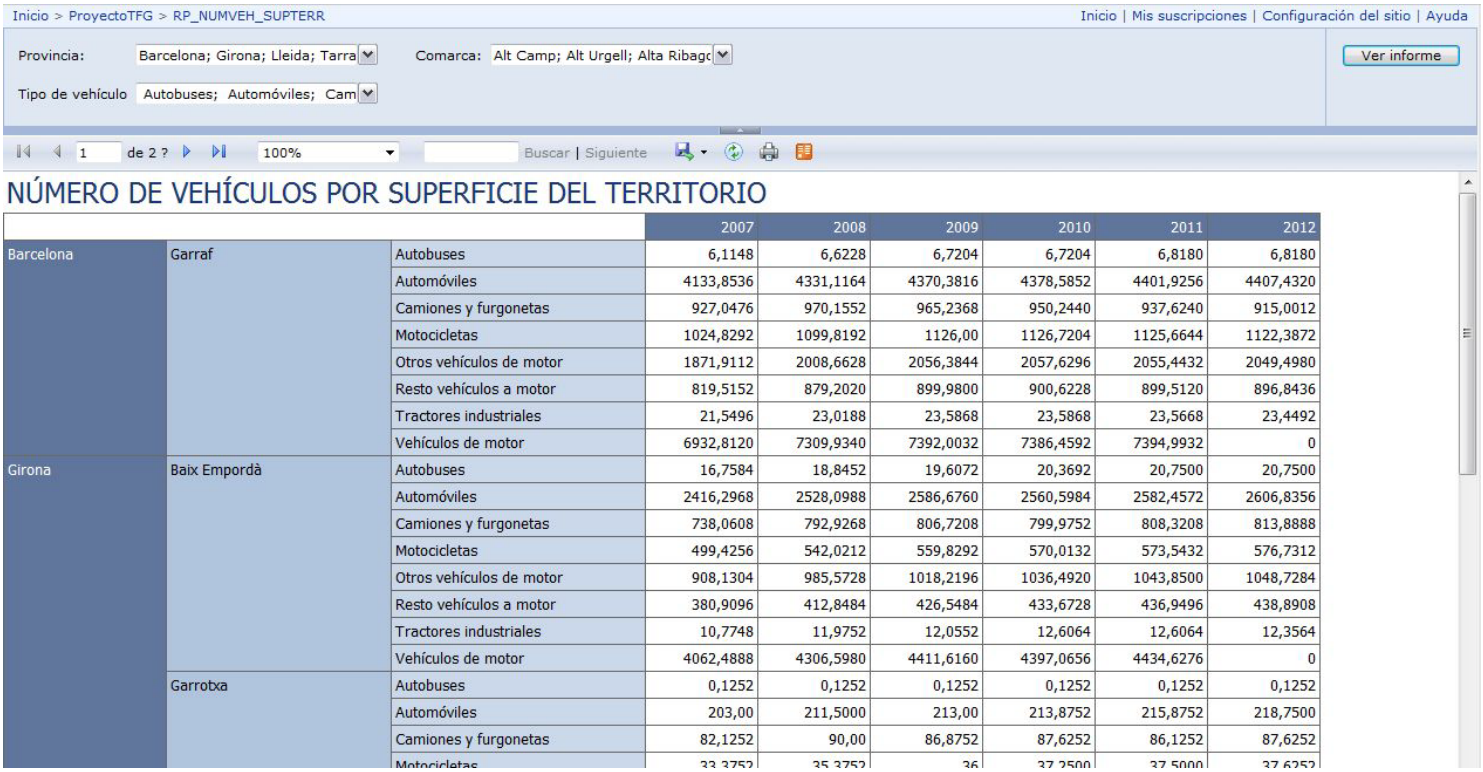
RATIO DE VEHÍCULOS POR CONDUCTOR

					2007	2008	2009	2010	2011	2012
Girona	Baix Empordà	Autobuses	Licencia	Femenino	0,8758	0	0	0	0	0
				Masculino	0,6107	0	0	0	0	0
			Permiso	Femenino	0,0905	0,0918	0,0956	0,0979	0,0987	0
				Masculino	0,0677	0,0675	0,0706	0,0738	0,0767	0
		Automóviles	Licencia	Femenino	99,7449	0	0	0	0	0
				Masculino	74,2088	0	0	0	0	0
			Permiso	Femenino	11,9200	10,9665	11,1359	10,9913	10,8903	0
				Masculino	9,2249	8,5026	8,6814	8,5885	8,6621	0
		Camiones y furgonetas	Licencia	Femenino	33,8747	0	0	0	0	0
				Masculino	26,1625	0	0	0	0	0
			Permiso	Femenino	4,1335	3,9391	3,9509	3,9065	3,8896	0
				Masculino	3,2257	3,0951	3,1237	3,0893	3,1262	0
		Motocicletas	Licencia	Femenino	20,9658	0	0	0	0	0
				Masculino	15,3148	0	0	0	0	0
			Permiso	Femenino	2,4988	2,3878	2,4570	2,4787	2,4729	0
				Masculino	1,9287	1,8441	1,9084	1,9319	1,9617	0
Tarragona	Baix Camp	Autobuses	Licencia	Femenino	0,2071	180	183	0	180	0
				Masculino	0,0854	27,7142	32,5001	38	38	0

6.11 Cantidad de vehículos por superficie del territorio

Este informe muestra la cantidad de vehículos por superficie del territorio en los años de 2007 a 2012. Estos años se muestran como columnas del informe mientras que las filas están agrupadas por provincia, comarca y tipo de vehículo. Los datos de las filas se corresponden con los parámetros disponibles.

Un ejemplo de su ejecución sería el siguiente:



NÚMERO DE VEHÍCULOS POR SUPERFICIE DEL TERRITORIO

			2007	2008	2009	2010	2011	2012
Barcelona	Garraf	Autobuses	6,1148	6,6228	6,7204	6,7204	6,8180	6,8180
		Automóviles	4133,8536	4331,1164	4370,3816	4378,5852	4401,9256	4407,4320
		Camiones y furgonetas	927,0476	970,1552	965,2368	950,2440	937,6240	915,0012
		Motocicletas	1024,8292	1099,8192	1126,00	1126,7204	1125,6644	1122,3872
		Otros vehículos de motor	1871,9112	2008,6628	2056,3844	2057,6296	2055,4432	2049,4980
		Resto vehículos a motor	819,5152	879,2020	899,9800	900,6228	899,5120	896,8436
		Tractores industriales	21,5496	23,0188	23,5868	23,5868	23,5668	23,4492
		Vehículos de motor	6932,8120	7309,9340	7392,0032	7386,4592	7394,9932	0
Girona	Baix Empordà	Autobuses	16,7584	18,8452	19,6072	20,3692	20,7500	20,7500
		Automóviles	2416,2968	2528,0988	2586,6760	2560,5984	2582,4572	2606,8356
		Camiones y furgonetas	738,0608	792,9268	806,7208	799,9752	808,3208	813,8888
		Motocicletas	499,4256	542,0212	559,8292	570,0132	573,5432	576,7312
		Otros vehículos de motor	908,1304	985,5728	1018,2196	1036,4920	1043,8500	1048,7284
		Resto vehículos a motor	380,9096	412,8484	426,5484	433,6728	436,9496	438,8908
		Tractores industriales	10,7748	11,9752	12,0552	12,6064	12,6064	12,3564
		Vehículos de motor	4062,4888	4306,5980	4411,6160	4397,0656	4434,6276	0
	Garrotxa	Autobuses	0,1252	0,1252	0,1252	0,1252	0,1252	0,1252
		Automóviles	203,00	211,5000	213,00	213,8752	215,8752	218,7500
		Camiones y furgonetas	82,1252	90,00	86,8752	87,6252	86,1252	87,6252
		Motocicletas	33,3752	35,3752	36	37,2500	37,5000	37,6252

7. CONCLUSIONES

Una vez concluida la realización del proyecto podemos comparar los planteamientos inicialmente establecidos con el producto final y con ello obtener las siguientes conclusiones:

- Se ha conseguido alcanzar los distintos hitos parciales establecidos en la planificación inicial, además, la calidad de los distintos productos obtenidos en estos hitos ha sido la adecuada. De lo cual podemos deducir que el reparto del esfuerzo dedicado a la realización de todos los hitos en conjunto ha sido equilibrado.

- A la vista del producto final y los objetivos inicialmente establecidos podemos concluir que se han conseguido estos objetivos por las siguientes razones:
    - Para la realización de este proyecto, el cual se corresponde con el TFG, se han empleado las destrezas de planificación, análisis, diseño e implementación aprendidas durante la realización de los estudios que ahora se concluyen.
    - El proyecto se centra en el área del conocimiento del Data Warehousing, por lo que ha sido necesario adquirir conocimientos en esta área temática. Situación habitual en el mundo informático a nivel empresarial.
    - En cuanto a los objetivos propios del proyecto, el sistema implementa un almacén de datos y el de conjunto de informes necesario.
- Además de los objetivos del proyecto, en la fase de análisis se determinaron un conjunto de requerimientos funcionales y no funcionales que el sistema debía tener en cuenta. Ya que el producto final contempla todos los requerimientos de ambos tipos, podemos concluir que tiene la calidad requerida y que satisface las necesidades del cliente FECRES.
- Afortunadamente no podemos indicar si el plan de riesgos definido es el adecuado, puesto que no se ha producido ninguno. No obstante, si podemos concluir que es muy importante la creación detallada de un plan de riesgos al principio de la realización de un proyecto, porque de esta forma quedarán cubiertos todos (o casi todos) los posibles inconvenientes que puedan surgir y, en caso que se produzca alguno, se sabrá cómo actuar de forma que el proyecto no sufra retrasos.
- La metodología empleada como referencia para el análisis y diseño del almacén de datos (HEFESTO), ha sido la apropiada.
- A la vista las distintas tareas realizadas y el tiempo empleado, podemos concluir que en este tipo de aplicaciones tanto los datos de entrada como el propio almacén de datos tienen una gran importancia. Por tanto, es básico realizar un análisis exhaustivo de ambos, así como, un buen diseño en el caso del modelo físico del almacén de datos.
- Desde un punto de vista personal, he aumentado mi experiencia la realización de todas las fases que conlleva la construcción de un sistema informático. Además, he adquirido conocimientos y destrezas en el área del Data Warehousing, así como, en las herramientas que se emplean en este entorno de conocimiento. Igualmente, he adquirido conocimientos en el proceso ETL y como llevarlo a cabo. Este proceso, si bien es empleado en el área del Data Warehousing, también puede serlo en el desarrollo de aplicaciones convencionales. Por todo ello, puedo concluir que la realización del TFG ha enriquecido mis capacidades profesionales.

8. LÍNEAS DE EVOLUCIÓN FUTURAS

Finalizado el proyecto y, ya que hemos concluido que su calidad es la adecuada, cabe plantearse posibles líneas de evolución que permitan en el futuro mejorar las capacidades del sistema actual, así como, añadir nuevas funcionalidades o requerimientos no funcionales. En este sentido, se han detectados las siguientes mejoras:

- Aumentar las capacidades del proceso ETL de forma que tenga en cuenta datos corregidos. El proceso ETL detecta errores de datos, por ejemplo valores "n.d." en campos numéricos, así como circunstancias atípicas, por ejemplo municipios de los que se tiene información sobre sus vehículos pero no sobre su población y extensión. Se podría mejorar el proceso ETL para que una vez que se disponga de la información corregida se pueda añadir al almacén de datos. De esta forma se mejoraría la precisión de los informes.
- Añadir gráficos a los informes. Puesto que la intención de FECRES es comprobar la evolución del tráfico la inclusión de gráficos facilitaría la comprensión visual de esta evolución.
- Aumentar el conjunto de datos origen incluyendo los datos de otras comunidades autónomas, e incluso extenderlo al ámbito nacional.
- Establecer en el sistema la capacidad del multiidioma, de manera que, al acceder al sistema el usuario podría escoger el idioma en el que aparecerían los distintos literales. En este sentido hay que decir que, puesto que para la creación y ejecución de los informes se emplean ciertas herramientas, habría que asegurarse que estas herramientas permiten cumplir con este requerimiento.
- Añadir nuevos informes a partir de los datos incluidos en el almacén, por ejemplo, número de radares por vía y número de radares por extensión (km2).
- Emplear técnicas de minería de datos que nos permitan inferir un posible comportamiento, por ejemplo, el posible aumento de tráfico en una determinada demarcación geográfica. Este aumento estaría determinado por la evolución del número de conductores y la evolución del número de vehículos en dicha demarcación geográfica.
- Podrían añadirse nuevos datos al sistema, como por ejemplo el número de multas por exceso de velocidad expendidas en un año en una determinada vía. Con ello, se podría comprobar si un aumento en el número de radares de una vía conlleva un aumento en este tipo de multas o, por el contrario, una disminución por una conducción más correcta.

9. BIBLIOGRAFÍA

Para la realización de este trabajo se ha consultado la siguiente bibliografía:

- Wikipedia
  www.wikipedia.org/
- Artículo sobre Data Warehousing
  http://www.1keydata.com/datawarehousing/datawarehouse.html
- Adictos al trabajo, portal con artículos y tutoriales informáticos
  http://www.adictosaltrabajo.com/
- HEFESTO: Metodología propia para la Construcción de un Data Warehouse
- Dataprix, portal con documentación y foros sobre Tecnologías de la Información y la Comunicación
  http://www.dataprix.com
- http://www.kimballgroup.com/.
- Sinnexus, empresa de servicios Bussiness Intelligence
  http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx
- Plan docente de la asignatura TFG – Data Warehouse
- http://docs.oracle.com/
- http://www.orafaq.com/wiki/SQL*Loader_FAQ
- http://msdn.microsoft.com/

10. ANEXOS

10.1 Software empleado

Para la realización del TFG, si tenemos en cuenta la parte técnica y la correspondiente a la documentación, se empleó el siguiente software:

- Microsoft Office (Word, Excel y PowerPoint)
- Adobe Reader
- Microsoft Project 2010
- Microsoft Visual Studio
- Open ModelSphere
- yED
- Windows Vista
- Máquina Virtual de la UOC alojada en Amazon Web Services, con el siguiente software:

- Windows Server 2008 32bits
- Oracle 11 XE 32bits
- Microsoft SQL Analysis Services 2012 32bits