Construcción y explotación de un data warehouse para el análisis de información sobre el tránsito rodado de vehículos

Rubén Sánchez Ginés
Grado Ingeniería Informática

Carles Llorach Rius

6 de Enero de 2014

Esta obra está sujeta a una licencia de
Reconocimiento-NoComercial-SinObraDerivada 3.0
España de Creative Commons

Título del trabajo:	*Construcción y explotación de un data warehouse para el análisis de información sobre el tránsito rodado de vehículos*
Nombre del autor:	*Rubén Sánchez Ginés*
Nombre del consultor:	*Carles Llorach Rius*
Fecha de entrega:	*01/2014*
Área del Trabajo Final:	*Data Warehouse*
Titulación:	*Grado en ingeniería informática*

Resumen del Trabajo:

Los almacenes de datos se han convertido en una herramienta indispensable en las grandes organizaciones para la mejora en la toma de decisiones, lo que las convierte en imprescindibles en un desarrollo de inteligencia de negocios de una empresa es que permiten aunar toda la información de la empresa que se encuentra diseminada en diferentes formatos y estructuras de datos, permitiendo obtener una imagen global de toda la organización. Además permiten mantener un histórico de los datos almacenados que de otra manera no sería posible.

En las siguientes páginas documentaré todo el proceso de creación de un almacén de datos, el análisis y tratamiento de las fuentes de datos, la creación de bases de datos relacionales tanto para el proceso de ETL, Extracción Transformación y Carga, como para la creación del cubo OLAP, la generación de índices y constraints para una mayor eficiencia, la creación de tablas de dimensiones y de hechos para la generación del cubo OLAP y la generación de informes para el análisis de los datos.

El área de la inteligencia de negocios así como en nuestro caso más particular los almacenes de datos permiten al estudiante de una manera práctica demostrar los conocimientos adquiridos durante los estudios de grado de ingeniería informática. Este hecho sumado a la necesidad actual y futura de las empresas en este terreno me hizo decidirme en la elección del proyecto actual.

Abstract:

Data warehouse have become an indispensable tool in large organizations to improve making decisions. What makes them essential in developing business intelligence of a company is that allow to unify all the information of the company that is scattered in different formats and data structures , allowing to obtain a global image of all the organization. Furthermore, it allows to maintain an historical data warehouse that couldn't be possible otherwise.

In the following pages I will document the whole process of creating a data warehouse: analysis and treatment of data sources; creating relational databases both for the ETL process (Extraction Transformation and Loading) and for the creation of the OLAP cube; generation of indexes and constraints for a better efficiency; the creation of dimension tables and facts for the generation of the OLAP cube and the generation or reports for data analysis.

The area of business intelligence and, in our particularly case, data warehouse allow students to demonstrate the knowledge acquired during the computer engineering grade. This fact, added to the current and future needs of businesses in this area, made me decide on the choice of the current project.

Palabras clave:

Inteligencia empresarial, FIC, Almacén de datos, ETL, OLAP, ROLAP, MOLAP

A Sara, por los días que no pude estar contigo.

Índice de contenidos

1. **Introducción** 10
1.1. Contexto y justificación del proyecto 10
1.2. Objetivos del proyecto 12
1.3. Enfoque y método seguido 12
1.4. Planificación del proyecto 14
1.5. Productos obtenidos 16
1.6. Descripción del resto de capítulos de la memoria 16

2. **Análisis** 16
2.1. Análisis de requerimientos 16
2.2. Casos de uso 20
2.3. Diseño conceptual 21

3. **Diseño** 28
3.1. Diagrama de arquitectura del software 28
3.2. Diagrama de arquitectura del hardware 29
3.3. Diseño de la base de datos y diagrama del modelo físico 29
3.4. Diseño y descripción de los informes creados 33

4. **Conclusiones** 44

5. **Líneas de evolución futuras** 45

6. **Glosario** 46

7. **Bibliografía** 47
7.1. Apuntes 47
7.2. Web 47

8. **Anexo** 48
A. Análisis de riesgos 48
B. Estimación de tiempo de desarrollo 49
C. Viabilidad 49
D. Análisis de tecnologías 50
E. Análisis de los datos de origen y calidad de la información 51
F. Fuentes extras de información 55
G. Modificaciones en las fuentes de información 55
H. Proceso ETL 55
I. Cubo OLAP 70

Índice de ilustraciones

Ilustración 1. Cronograma de actividades 14
Ilustración 2. Diagrama de Gantt 15
Ilustración 3. Casos de uso del usuario administrador 20
Ilustración 4. Casos de uso del usuario analista 20
Ilustración 5. Casos de uso del usuario consultor 20
Ilustración 6. Relación de métricas y dimensiones 25
Ilustración 7. Esquema relacional 26
Ilustración 8. Esquema multidimensional 27
Ilustración 9. Diagrama de arquitectura del software 28
Ilustración 10. Diagrama de arquitectura del hardware 29
Ilustración 11. Modelo entidad relación 32
Ilustración 12. Informe total vehículos 34
Ilustración 13. Informe total conductores 35
Ilustración 14. Informe porcentaje de vehículos respecto a población 36
Ilustración 15. Informe densidad de población 37
Ilustración 16. Informe número de vehículos por radar 38
Ilustración 17. Informe porcentaje de conductores por radar 39
Ilustración 18. Informe indicador de conductores vs habitantes 40
Ilustración 19. Informe indicador de radares vs vehículos 41
Ilustración 20. Informe ratio de vehículos por conductor 42
Ilustración 21. Informe densidad de vehículos por Km^2 43
Ilustración 22. Informe radares por tipo de vía 44
Ilustración 23. Tabla tipo_vias 57
Ilustración 24. Transformación municipios 58
Ilustración 25. Ejemplo reemplazo nombres de municipio 59
Ilustración 26. Transformación provincia-comarcas 60
Ilustración 27. Transformación habitantes 60
Ilustración 28. Normalización de los campos anuales 61
Ilustración 29. Normalización porcentajes por sexo 62
Ilustración 30. Transformación vías 62
Ilustración 31. Transformación radares 63

Ilustración 32. Transformación autorizaciones 64
Ilustración 33. Creación del campo año 64
Ilustración 34. Normalización de tipo-sexo 64
Ilustración 35. Partir campos id_autorizacio e id_sexo 65
Ilustración 36. Transformación update licencias 2007 65
Ilustración 37. Transformación vehículos 66
Ilustración 38. Normalización de filas tipo-año 67
Ilustración 39. Transformación de zonas 68
Ilustración 40. Transformación de hechos 68
Ilustración 41. Trabajo que ejecuta todas las transformaciones 69
Ilustración 42. Explorador de proyectos 70
Ilustración 43. Cubo OLAP 71
Ilustración 44. Jerarquías de agregación 71

Índice de tablas

Tabla 1. Objetivos a cumplir y fechas de entrega 13
Tabla 2. Descripción de los campos de la tabla tiempo 30
Tabla 3. Descripción de los campos de la tabla vehículos 30
Tabla 4. Descripción de los campos de la tabla vías 30
Tabla 5. Descripción de los campos de la tabla sexo 30
Tabla 6. Descripción de los campos de la tabla autorizaciones 30
Tabla 7. Descripción de los campos de la tabla zonas 31
Tabla 8. Descripción de los campos de la tabla hechos 31
Tabla 9. Descripción de incidencias y planes de contingencia 48
Tabla 10. Tabla de estudio de viabilidad 49
Tabla 11. Tabla comparativa de tecnologías MOLAP vs ROLAP 51

1 Introducción

1.1 Contexto y justificación del proyecto

La Fundació d'Estudis per a la Conducció RESponsable (FECRES) quiere profundizar en la evolución del tráfico rodado de vehículos y analizar las posibles correlaciones entre medios de transporte, perfiles de conductor y algunas variables de seguridad vial.

Actualmente FECRES no dispone de software que le permita analizar estos datos, para ello es necesario el desarrollo de un sistema BI ya que los datos de los que dispone FECRES se encuentran en diferentes fuentes, el motivo es que no se disponen de datos propios si no que depende de los datos de los organismos IDESCAT, la DGT y el Servei Català de Trànsit.

Descripción del proyecto

La solución a la problemática de FECRES es la creación de una factoría de la información corporativa o FIC.

La FIC es un conjunto de componentes que interactúan para ayudar a gestionar los flujos de datos de los sistemas operacionales de la empresa. Tiene como objetivo transformar la información proporcionada por las fuentes de datos para su posterior análisis.

Componentes de la FIC

Fuentes de información

Las fuentes de información son las encargadas de dar soporte al proceso de negocio. Están formadas por un conjunto de aplicaciones independientes entre sí, que fueron creados a medida que surgía la necesidad, pueden estar en un mismo entorno o pertenecer a entornos diferentes.

Data Warehouse

Partiendo de las fuentes de información se crea el DW, este es independiente a las fuentes de información y su uso está reservado a la consulta.

Existen tres tipos de DW.

1. **Almacén operacional de datos:** este almacén de datos contiene información actual de las fuentes de información, se utiliza para realizar consultas que requieren información actualizada. A diferencia de los otros dos tipos que veremos no mantiene un histórico y su volumen de datos es menor.

2. **Almacén de datos corporativo:** Aglutina toda la información relevante a las fuentes de información. La información no se actualiza si no que mantiene un histórico. Es tipo de almacén de datos no suele usarse para consultas ya que su volumen de información es muy grande. Si existe un almacén operacional de datos se obtienen los datos directamente de él y no de las fuentes de datos.

3. **Almacén de datos departamental:** Surge de dividir en subconjuntos más pequeños el almacén de datos corporativo, con la información imprescindible de un área determinada del negocio, y son estos tipos de almacén de datos los que son consultados por los usuarios.

Componentes de integración i transformación.

También conocidos como ETL, son los encargados de extraer, transformar y cargar los datos para su posterior inserción en los almacenes de datos.

Para su desarrollo se usan herramientas middleware, que usan software encargado de la conectividad sobre diferentes plataformas.

El proceso tiene dos fases, la primera se trata de la obtención de la imagen inicial y la segunda es la obtención de las actualizaciones, en la primera fase se obtiene la información hasta el momento actual, esto implica que no se dispondrá de datos históricos, la segunda fase es la encargada de obtener los nuevos datos que se van a generar.

Metadatos

Son los datos que ofrecen información sobre los datos transferidos y almacenados.

Herramientas OLAP

Son las herramientas encargadas de generar las agregaciones de datos para su uso posterior, históricamente se caracterizan por la generación de tablas multidimensionales de información, aunque actualmente existen diferentes maneras de modelar la información.

Generación de informes

Es la información extraída y formateada para poder ser consultada y analizada.

1.2 Objetivos del proyecto

Objetivos generales

El objetivo general del TFG data warehouse es adquirir experiencia en el análisis, diseño, construcción y explotación de un data warehouse, además de los objetivos del área escogida, en este caso data warehouse, el TFG tiene como objetivo que el alumno ponga en práctica todos los conocimientos adquiridos durante el curso y aprenda a documentar y realizar una memoria del proyecto.

Objetivos específicos

- Creación de herramientas ETL para el tratamiento e importación de datos desde diferentes fuentes.
- Creación de un data warehouse.
- Creación de sistemas OLAP .
- Generación de Informes.
- Aprendizaje de software orientado a la creación de la FIC.
- Aportar una solución eficiente a una problemática planteada.
- Generar una documentación profesional de la solución aportada.
- Cumplir un plazo prefijado de entrega.

1.3 Enfoque y Método seguido

FECRES no dispone de un software con el que estuviera dando solución a la problemática referida, es por ello que no se parte de una base a analizar y mejorar, será necesario el desarrollo de un producto nuevo.

Se ha seguido el método SDLC, ciclo vital del desarrollo de aplicaciones, en cascada, es decir que cada fase del proyecto es continuación de la fase anterior.

Este sistema de trabajo permite avanzar de forma segura una vez se han cumplido los requisitos.

Las fases de cada etapa son:

1. **Plan de trabajo**: Elaboración del documento donde se establece la planificación de las diferentes tareas a realizar durante el proyecto y donde se hace un análisis de los datos disponibles y de la solución para determinar el alcance del proyecto.

2. **Análisis de requisitos y diseño conceptual**: Elaboración de un documento con el análisis más detallado de los requerimientos basados en el análisis preliminar, diseño del modelo dimensional que dará soporte a las necesidades de los usuarios.
3. **Implementación**: Construcción del almacén de datos, configuración de las herramientas ETL y OLAP, construcción de los informes y análisis de estos con la información obtenida.

4. **Memoria y presentación visual:** Recopilación de los documentos de las fases anteriores y conclusiones finales, creación de una presentación virtual.

Objetivos a cumplir y fechas de entrega

Objetivo	Inicio	Entrega
Tarea 1 - Plan de trabajo	19/09/2013	01/10/2013
Tarea 2 - Análisis y diseño	02/10/2013	05/11/2013
Tarea 3 - Implementación	06/11/2013	18/12/2013
Tarea 4 - Entrega final y defensa	19/12/2013	06/01/2014

Tabla 1 – Objetivos a cumplir y fechas de entrega

1.4 Planificación del proyecto

Propuesta de actividades i cronograma

	Nombre de tarea	Comienzo	Fin	Predecesoras
1	**TFG Data Whare House**	**jue 19/09/13**	**lun 06/01/14**	
2	**Tarea 1 - Plan de trabajo**	**jue 19/09/13**	**mar 01/10/13**	
3	Busqueda de información	jue 19/09/13	jue 26/09/13	
4	Análisis inicial del proyecto	jue 19/09/13	sáb 21/09/13	
5	Planificación	dom 22/09/13	vie 27/09/13	4
6	Documentación	dom 29/09/13	dom 29/09/13	5
7	Revisión y entrega	lun 30/09/13	mar 01/10/13	6
8	**Tarea 2 - Analisis y diseño**	**mié 02/10/13**	**mar 05/11/13**	**2**
9	Familiarización con el Entorno virtual	mié 02/10/13	mié 02/10/13	
10	Familiarización con el software	jue 03/10/13	mar 08/10/13	9
11	Requerimientos funcionales/no funcionales	dom 06/10/13	lun 07/10/13	
12	Modelo conceptual	mar 08/10/13	vie 11/10/13	11
13	Diseño base de datos	sáb 12/10/13	sáb 19/10/13	12
14	Analisis modelo multidemensional	dom 20/10/13	vie 25/10/13	13
15	Calidad de datos	sáb 26/10/13	dom 27/10/13	14
16	Proceso ETL	lun 28/10/13	jue 31/10/13	15
17	Documentación	vie 01/11/13	sáb 02/11/13	16
18	Revisión y entrega	dom 03/11/13	mar 05/11/13	17
19	**Tarea 3 - Implementación**	**mié 06/11/13**	**mié 18/12/13**	**8**
20	Creación Herramientas ETL	mié 06/11/13	sáb 16/11/13	
21	Creación del almacen de datos	dom 17/11/13	dom 24/11/13	20
22	Creación Herramientas Olap	lun 25/11/13	sáb 30/11/13	21
23	Creación Informes	dom 01/12/13	vie 06/12/13	22
24	Interpretación de datos	sáb 07/12/13	mar 10/12/13	23
25	Documentación	mié 11/12/13	dom 15/12/13	24
26	Revisión y entrega	lun 16/12/13	mié 18/12/13	25
27	**Tarea 4 - Entrega final y defensa**	**jue 19/12/13**	**lun 06/01/14**	**19**
28	Documentación memoria	jue 19/12/13	dom 22/12/13	
29	Guión presentación	lun 23/12/13	mié 25/12/13	28
30	Documentación gráfica	jue 26/12/13	vie 27/12/13	29
31	Familiarización software video	sáb 28/12/13	sáb 28/12/13	30
32	Gravación sonido presentación	dom 29/12/13	mar 31/12/13	31
33	Montaje presentación	mié 01/01/14	sáb 04/01/14	32
34	Revisión y entrega	dom 05/01/14	lun 06/01/14	33

Ilustración 1 - Cronograma de actividades

Diagrama de Gantt

	Nombre de tarea	Comienzo	Fin
1	**TFG Data Whare House**	**jue 19/09/13**	**lun 06/01/14**
2	**Tarea 1 - Plan de trabajo**	**jue 19/09/13**	**mar 01/10/13**
3	Busqueda de información	jue 19/09/13	jue 26/09/13
4	Análisis inicial del proyecto	jue 19/09/13	sáb 21/09/13
5	Planificación	dom 22/09/13	vie 27/09/13
6	Documentación	dom 29/09/13	dom 29/09/13
7	Revisión y entrega	lun 30/09/13	mar 01/10/13
8	**Tarea 2 - Analisis y diseño**	**mié 02/10/13**	**mar 05/11/13**
9	Familiarización con el Entorno virtual	mié 02/10/13	mié 02/10/13
10	Familiarización con el software	jue 03/10/13	mar 08/10/13
11	Requerimientos funcionales/no funcionales	dom 06/10/13	lun 07/10/13
12	Modelo conceptual	mar 08/10/13	vie 11/10/13
13	Diseño base de datos	sáb 12/10/13	sáb 19/10/13
14	Analisis modelo multidemensional	dom 20/10/13	vie 25/10/13
15	Calidad de datos	sáb 26/10/13	dom 27/10/13
16	Proceso ETL	lun 28/10/13	jue 31/10/13
17	Documentación	vie 01/11/13	sáb 02/11/13
18	Revisión y entrega	dom 03/11/13	mar 05/11/13
19	**Tarea 3 - Implementación**	**mié 06/11/13**	**mié 18/12/13**
20	Creación Herramientas ETL	mié 06/11/13	sáb 16/11/13
21	Creación del almacen de datos	dom 17/11/13	dom 24/11/13
22	Creación Herramientas Olap	lun 25/11/13	sáb 30/11/13
23	Creación Informes	dom 01/12/13	vie 06/12/13
24	Interpretación de datos	sáb 07/12/13	mar 10/12/13
25	Documentación	mié 11/12/13	dom 15/12/13
26	Revisión y entrega	lun 16/12/13	mié 18/12/13
27	**Tarea 4 - Entrega final y defensa**	**jue 19/12/13**	**lun 06/01/14**
28	Documentación memoria	jue 19/12/13	dom 22/12/13
29	Guión presentación	lun 23/12/13	mié 25/12/13
30	Documentación gráfica	jue 26/12/13	vie 27/12/13
31	Familiarización software video	sáb 28/12/13	sáb 28/12/13
32	Gravación sonido presentación	dom 29/12/13	mar 31/12/13
33	Montaje presentación	mié 01/01/14	sáb 04/01/14
34	Revisión y entrega	dom 05/01/14	lun 06/01/14

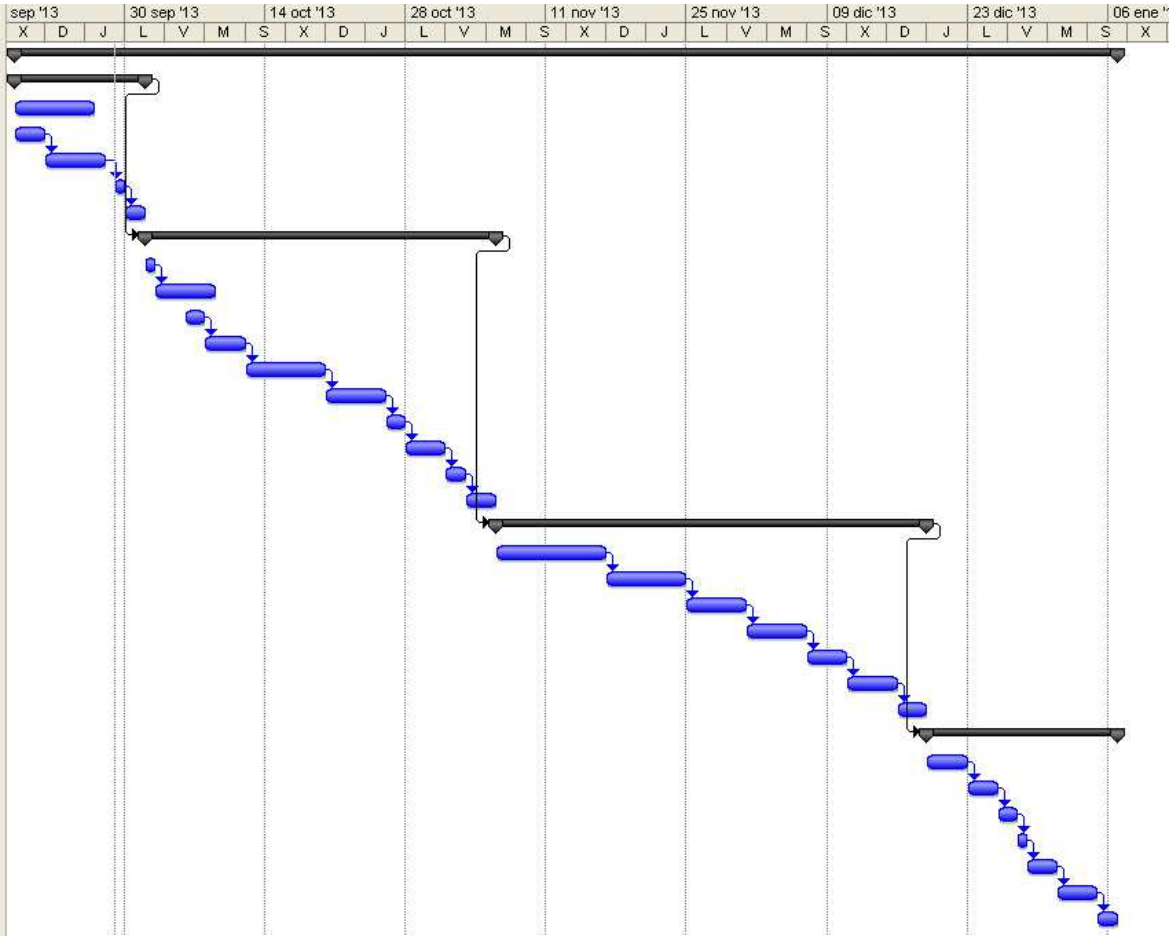

Ilustración 2 - Diagrama de Gantt del proceso de creación del proyecto

1.5 Productos obtenidos

Los productos obtenidos en el desarrollo del proyecto son:

Base de datos relacional: Se han creado dos bases de datos relacionales en MySQL, una se usará durante el proceso de ETL como tabla auxiliar de carga, la otra será la que albergará la información del cubo OLAP.

Proceso ETL: Desarrollado en Pentaho y almacenado en C:\TFG\ETL encontramos los archivos que componen el proceso de ETL.

Cubo OLAP: Creado en Microsoft Analysis Services.

Informes: Creados en Microsoft Analysis Services.

Memoria: Documento que expone todo el proceso de desarrollo del proyecto.

Presentación virtual: Video de presentación del proyecto.

1.6 Descripción del resto de capítulos de la memoria.

El resto de capítulos de la memoria se compone de:

Análisis: En este capítulo se analizan los requisitos del proyecto y la solución a nivel conceptual.

Diseño: En este apartado se mostrará tanto el diagrama del software como el de hardware, el diseño de la base de datos y de los informes.

Conclusiones: Se redactaran las conclusiones obtenidas durante el desarrollo del proyecto.

Líneas de evolución futura: Se describirá una propuesta de mejora para el proyecto.

Glosario: Definición de términos propios del área de almacén de datos.

Bibliografía: Enumeración de los recursos bibliográficos utilizados durante el desarrollo del proyecto.

Anexos: Ampliación de la información presentada en la memoria.

2 Análisis

2.1 Análisis de requerimientos

En este apartado se especificaran los requerimientos necesarios del sistema, estos están divididos en funcionales y no funcionales.

Para poder dar solución a la petición de FECRES, el aplicativo deberá ser capaz de generar y mantener los siguientes informes:

- Total de vehículos.
- Total de conductores.
- % de vehículos respecto a población.
- Densidad de población y densidad de tráfico.
- 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 de territorio.

Estos informes deberán poder consultarse de forma agregada por comarca/provincia/comunidad autónoma, tipo de vehículo, tipo de permiso, tipo de vía y año.

Requerimientos Funcionales:

Los requerimientos funcionales describen el funcionamiento del sistema y el tratamiento de errores.

Funcionalidades de Interfaz

La interfaz grafica de usuario o GUI usará un sistema de ventanas.

El usuario accederá al aplicativo mediante usuario y contraseña.

Las modificaciones y eliminaciones de datos se realizaran mediante formularios.

La presentación de datos se realizará en forma de árbol de categorías, lo que permitirá una representación organizada de la información y facilitará al usuario la comprensión de la estructura de datos.

En los informes se permitirán realizar búsquedas y usar filtros.

Funcionalidades de usuario:

Los usuarios estarán clasificados en perfiles, según el perfil al que pertenezca el usuario tendrá más o menos permisos de acceso a las diferentes herramientas. Los usuarios se clasifican en:

- Consultor, podrá consultar y exportar los informes.
- Analista, podrá modificar los datos definidos, generar, modificar y eliminar informes.
- Administrador, con acceso a todo el aplicativo además de poder crear, editar y eliminar usuarios.

Funcionalidades de datos:

Importación de datos de diferentes formatos.

El aplicativo deberá ser capaz de importar datos de diferentes formatos mediante las herramientas ETL.

Transformación, normalización y carga de datos.

Los datos deberán ser transformados y normalizados en el proceso ETL antes de su carga en la base de datos.

Limpieza de datos erróneos o parciales.

Los datos erróneos o parciales deberán ser tratados o eliminados, evitando así almacenar datos incorrectos o poco fiables en la base de datos, además se generará un informe de errores con los errores producidos durante la carga.

Modificación y eliminación de datos.

Los datos cargados en base de datos podrán ser modificados y eliminados.

Funcionalidades de la estructura de datos

Creación de una base de datos relacional donde cada dimensión estará representada por una tabla y todas ellas estarán relacionadas con la tabla de hechos.

Creación de un cubo multidimensional usando tecnología MOLAP.

Funcionalidades de consulta de información

Creación, edición y eliminación de consultas.

Creación, edición y eliminación de informes.

Requisitos no funcionales:

Los requerimientos no funcionales definen las características del sistema en seguridad, rendimiento, calidad etc.

Seguridad:

Para acceder a la información del almacén de datos será necesaria la autentificación del usuario, además se controlará mediante el uso de perfiles las funcionalidades a las que tendrá acceso cada usuario.

Calidad:

En el proceso ETL se eliminarán los datos erróneos y trataran los datos parciales que puedan alterar la calidad de los datos.

Se tendrán en cuenta los cambios de legislación y se ampliaran las fuentes de información para cubrir los vacios de las fuentes proporcionadas.

Escalabilidad:

El diseño de la base de datos y el modelo en forma de estrella permite de manera fácil y rápida adaptarnos a nuevas necesidades que puedan surgir en un futuro, se ha tenido en cuenta la posibilidad de ampliar los datos a nivel de comunidades autónomas.

Facilidad de uso:

El sistema basado en una interfaz de ventanas será diseñado para un manejo sencillo e intuitivo, y será orientado a un usuario de conocimientos en informática de nivel medio.

Rendimiento:

Se sistema ha sido optimizado para un rendimiento alto, se han creado índices para los niveles de agregación para mejorar las agregaciones de datos, en caso de que no fuera suficiente y el rendimiento se viera a afectado se procedería a la creación de cubos por nivel de agregación.

2.2 Casos de uso

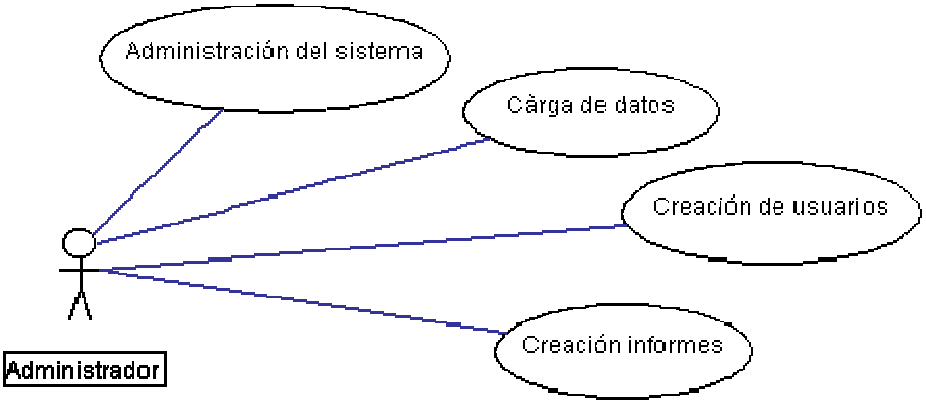

Ilustración 3 - Casos de uso del usuario administrador

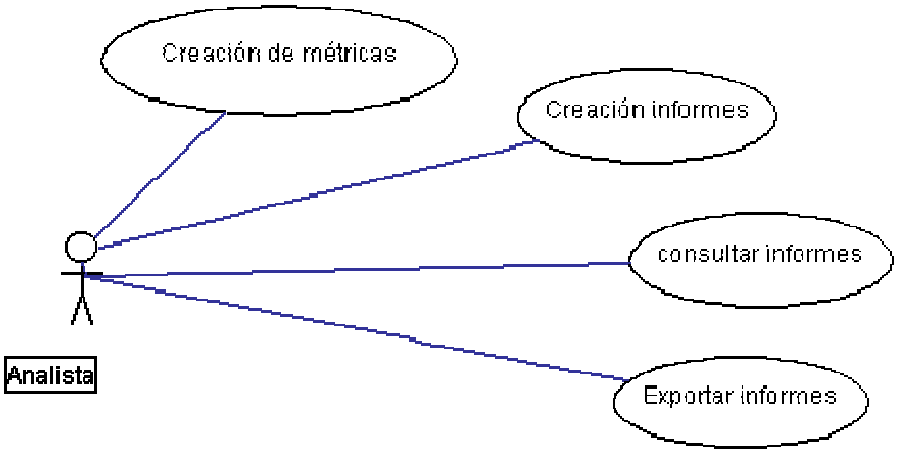

Ilustración 4 - Casos de uso del usuario analista

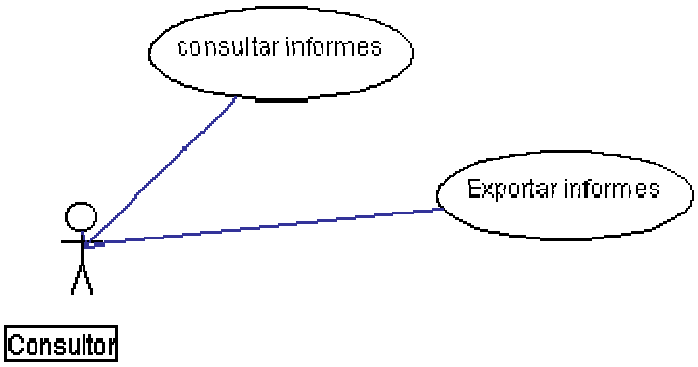

Ilustración 5 - Casos de uso del usuario consultor

2.3 Diseño conceptual

Analizando la estructura de datos he decidido crear un diseño en forma de estrella, considero que es una estructura sencilla y flexible que proporcionará un buen equilibrio entre rendimiento y escalabilidad.

Búsqueda de Hechos

Los hechos contienen los datos que se necesitan analizar, en nuestro caso los hechos que necesitaríamos serían:

- **El número de vehículos:** Este hecho nos permitirá analizar el número de vehículos por tipo de vehículo, tipo de permiso, zona y año.

- **El número de conductores:** Este hecho nos permitirá analizar el número de conductores por tipo de permiso, sexo, zona y año.

- **% de vehículos respecto a la población**: Este hecho nos permitirá analizar el % de vehículos de un tipo y tipo de permiso respecto a la población por sexo en una zona y año.

- **Densidad de población:** Este hecho nos permitirá analizar la densidad de población por sexo para una zona y año.

- **Densidad de transito:** Este hecho nos permitirá analizar la densidad de vehículos por km^2 por tipo de vehículo, tipo de permiso, para una zona y año

- **Número de vehículos por radar:** Este hecho nos permitirá analizar el número de vehículos de un tipo y tipo de permiso que hay por cada radar en una vía, zona y año

- **% de conductores por radar:** Este hecho nos permitirá analizar el número de conductores por tipo de permiso y sexo que hay por cada radar en una vía, zona y año.

- **Indicador de conductores por radar:** Este hecho nos permitirá analizar el indicador de conductores por sexo, con un tipo de autorización por radar de un tipo de vía en una zona y año.

- **Indicador de radares por vehículo:** Este hecho nos permitirá analizar el indicador de radares de una vía por vehículo de un tipo y tipo de permiso, en una zona y año.

- **Ratio de vehículos por conductor:** Este hecho nos permitirá analizar el ratio de vehículos de un tipo por conductor de un sexo para un tipo de permiso una

zona y un año.

- **Cantidad de vehículos / superficie del territorio:** Este hecho nos permitirá analizar la cantidad de vehículos de un tipo y tipo de permiso respecto a la superficie de una zona en un año.

Analizando los hechos necesarios se puede observar que muchos de ellos son operaciones matemáticas realizadas sobre unos datos primarios, estos datos primarios serán las medidas.

Solo las medidas formarán parte de la tabla de hechos, esto evitará que la tabla de hechos se haya de modificar con cada nuevo dato derivado que se quiera analizar, el resto de datos necesarios se pueden generar a través de operaciones por lo que serán calculados en tiempo de ejecución.

Granularidad

La granularidad representa el nivel de detalle que existe en las dimensiones y los hechos, es importante elegir correctamente la granularidad de los datos que almacenaremos, si elegimos un nivel de detalle demasiado bajo estaremos perdiendo información, por otro lado no podemos escoger un nivel de detalle muy alto si carecemos de la información necesaria, en algunos casos aunque dispongamos de la información necesaria podría ser contraproducente elegir un nivel de detalle demasiado elevado, ya que esto afectaría exponencialmente al tamaño de la base de datos.

Para el nivel de detalle de tiempo estamos limitados a nivel de año, ya que no disponemos de información más detallada como podría ser a nivel de trimestres, meses, semanas etc.

En nuestro caso cada entrada en la tabla de hechos contendrá información sobre una vía, un tipo de autorización, un tipo de vehículo en un municipio para las personas de un sexo en un año.

Dimensiones

Las dimensiones representan puntos de vista desde los cuales se pueden analizar los datos.

Como solución a las necesidades de FECRES he definido seis dimensiones, cada una de ellas tendrá un identificador único y numérico además de los atributos que expondré más adelante.

Tiempo: Momento en el que se analizan los datos.

Descriptores

- Año: año que se analiza.
- Bisiesto: Si ese año es bisiesto o no.
- Elecciones municipales: Si ese año ha tenido elecciones municipales.
- Elecciones autonómicas: Si ese año ha tenido elecciones autonómicas.
- Elecciones Estatales: Si ese año ha tenido elecciones estatales.
- Elecciones Europeas. Si ese año ha tenido elecciones europeas.

Zonas: Lugar geográfico del que analizamos los datos.

Jerarquías de agregación

- Comarca.
- Provincia.
- Comunidad autónoma.

Descriptores

- Municipio.
- Extensión: Extensión del municipio en kilómetros cuadrados.
- INE: Código INE del municipio.

Autorizaciones: Tipos de autorizaciones que permiten la conducción de vehículos.

Descriptores

- Tipo de autorización.

Vías: Vías en las que se encuentran situados los radares.

Categorización:

- por tipo de vía.

Descriptores

- Nombre de la vía.

Sexo: Sexo al cual pertenecen los habitantes o conductores.

Descriptores

- Sexo.

Vehículos: Tipo de vehículos.

Descriptores

- Tipo de vehículo.

Métricas

Las métricas son atributos numéricos que se usarán para el análisis y representaran el valor de las celdas de nuestro cubo.

Número de vehículos: Será la suma de vehículos.

Población: Será la suma de habitantes.

Radares: Será la suma de radares.

Extensión: Será la suma de extensiones de municipios.

Conductores: Al no disponer de datos de conductores debemos hallar una manera de calcularlos, teniendo en cuenta que el número de permisos/licencias no refleja fielmente el número de conductores ya que un conductor puede tener más de un permiso/licencia y que tampoco el número de vehículos puede considerarse cómo el número de conductores ya que un conductor puede tener ninguno, uno o más vehículos y que más de un conductor puede usar un mismo vehículo.

Por este motivo la mejor manera de encontrar el número más exacto de conductores sería cogiendo el valor mínimo entre autorizaciones y vehículos y así evitando en la medida de lo posible contabilizar un conductor varias veces.

Habría que tener en cuenta que el impuesto de matriculación al ser fijado por los ayuntamientos provoca que en algunos municipios contiguos haya una diferencia considerable de tasas, por lo que muchos vecinos deciden matricular su coche en pueblos cercanos, para evitar este desfase en los datos incluiremos el número de habitantes de un municipio en la formula, quedando de esta manera Min (Autorizaciones, Vehículos, Habitantes).

Relación de métricas y dimensiones

En el siguiente gráfico vemos la relación de las dimensiones con las métricas, como se puede observar no todas las dimensiones serán determinantes para el valor de una métrica dentro de una celda, por ejemplo la extensión de un territorio dependerá únicamente de la granularidad de la dimensión zonas independientemente del valor o nivel de gránulo del resto de dimensiones, se puede ver claramente que la extensión de un municipio no se ve afectada por el año que se consulte o por el tipo de vía seleccionado.

Cuando el nivel de gránulo varia en una dimensión que no afecta a un valor, sobre este valor no se realiza ninguna operación de agregación, a estas métricas se las conoce como semiaditivas, ya que se comportan de forma aditiva para unas dimensiones y de manera no aditiva para otras.

Se puede observar que todas las métricas son del tipo semiaditiva ya que hay como mínimo una dimensión relacionada y cómo mínimo una no relacionada.

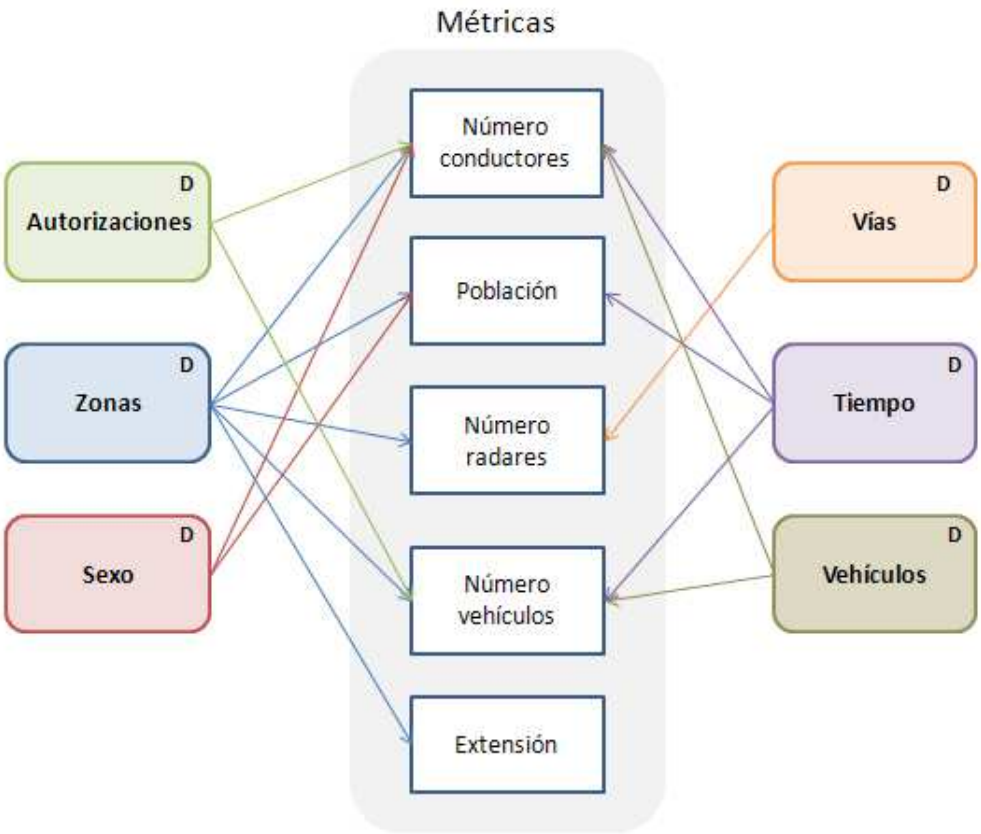

Ilustración 6 –Relación de métricas y dimensiones

Esquema relacional

La siguiente imagen ilustra el esquema relacional que tendrá nuestra estrella, en el se puede ver la relación de las dimensiones con la tabla de hechos.

Ilustración 7 –Esquema relacional

Esquema multidimensional

Una vez hecho el análisis nuestro esquema multidimensional.

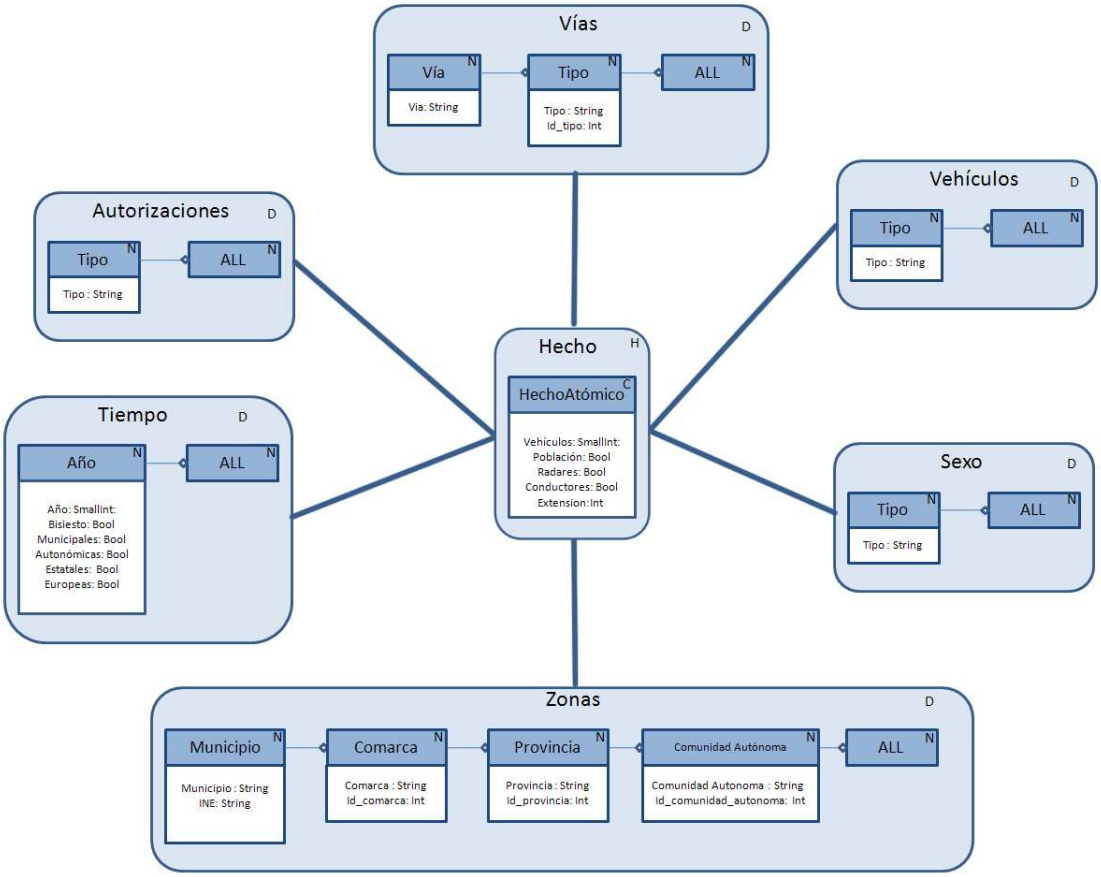

Ilustración 8 -Esquema multidimensional

3 Diseño

3.1 Diagrama de Arquitectura del Software

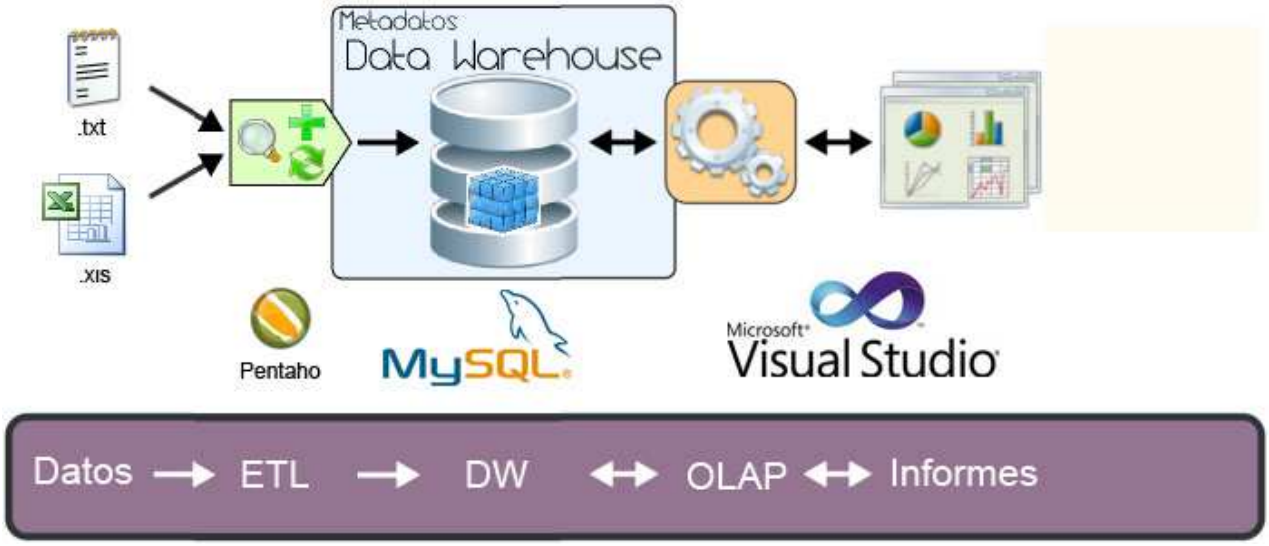

Ilustración 9 -Diagrama de arquitectura del software

Los datos han sido proporcionados en formato de texto y en archivos de Excel, se ha utilizado la herramienta Spoon de Pentaho para el proceso ETL, insertando los datos en una base de datos MySQL para la creación del cubo OLAP y los informes se ha usado Microsoft Analysis Services.

3.2 Diagrama de Arquitectura del Hardware

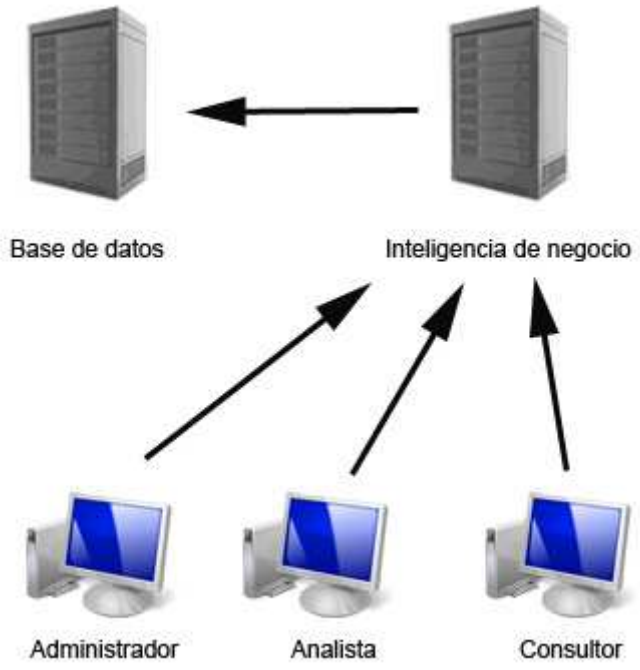

Ilustración 10 - Diagrama de arquitectura del hardware

El proceso de Inteligencia de negocio ha sido alojado en un servidor que se comunicará con el servidor de base de datos para proporcionar información a los distintos usuarios, las flechas simbolizan la red de comunicaciones y el sentido de la comunicación.

3.3 Diseño de la base de datos y diagrama del modelo Físico.

El diseño de la base de datos se ha creado a partir del diseño multidimensional a pesar de ello presenta una diferencia en la distribución de los datos, la medida extensión que pertenecía a la tabla de hechos ha sido desplazada a la dimensión zonas con el fin de optimizar el espacio ocupado.

Tablas de la base de datos

Todas las tablas son indexadas por defecto por el campo clave, además de esta indexación existen restricciones de tipo unique e indexaciones para agilizar las consultas y mantener una coherencia en los datos.

Tiempo			
	campos	Tipo	Espacio en bytes
	id	SmallInt	2
	año	SmallInt	2
	bisiesto	Bool	1
	elecciones municipales	Bool	1
	elecciones autonómicas	Bool	1
	elecciones estatales	Bool	1
	elecciones europeas	Bool	1

Tabla 2 – Descripción de los campos de la tabla tiempo

Vehiculos			
	campos	Tipo	Espacio en bytes
	id	SmallInt	2
	tipo	Varchar(20)	21

Tabla 3 – Descripción de los campos de la tabla vehículos

Vias			
	campos	Tipo	Espacio en bytes
	id	SmallInt	2
	via	Varchar(15)	16
	tipo	Varchar(20)	21
	id_tipo	SmallInt	2

Tabla 4 – Descripción de los campos de la tabla vías

El campo vía tiene un índice de tipo unique que evita que haya dos vías con el mismo nombre, además el campo id_tipo ha sido indexado para mejorar el rendimiento al agrupar por tipo de vía.

Sexo			
	campos	Tipo	Espacio en bytes
	id	SmallInt	2
	tipo	Varchar(10)	11

Tabla 5 – Descripción de los campos de la tabla sexo

Autorizaciones			
	campos	Tipo	Espacio en bytes
	id	SmallInt	2
	tipo	Varchar(15)	16

Tabla 6 – Descripción de los campos de la tabla autorizaciones

Zonas			
	campos	Tipo	Espacio en bytes
	id	SmallInt	2
	municipio	Varchar (50)	51
	INE	Cha5(5)	5
	comarca	Varchar (50)	51
	id_comarca	SmallInt	2
	provincia	Varchar(50)	51
	id_provincia	SmallInt	2
	comunidad_autonoma	Varchar(50)	51
	id_comunidad_autonoma	SmallInt	2
	extension	SmallInt	2

Tabla 7 – Descripción de los campos de la tabla zonas

El campo municipio tiene un índice de tipo unique para evitar duplicados, y se han creado índices para los campos id_comarca, id_provincia e id_comunidad autónoma para mejorar el rendimiento al realizar agrupaciones.

Hechos			
	campos	Tipo	Espacio en bytes
	id	Bigint	12
	id_autorizacion	SmallInt	2
	id_via	SmallInt	2
	id_sexo	SmallInt	2
	id_tiempo	SmallInt	2
	id_vehiculo	SmallInt	2
	id_zona	SmallInt	2
	vehiculos	Integer	4
	poblacion	Integer	4
	radares	Integer	4
	conductores	Integer	4

Tabla 8 – Descripción de los campos de la tabla hechos

Se ha creado un índice de tipo unique para el conjunto de campos id_autorización, id_via, id_sexo, id_tiempo, id_vehiculo e id_zona, para evitar duplicados que no deberían existir.

Modelo entidad relación

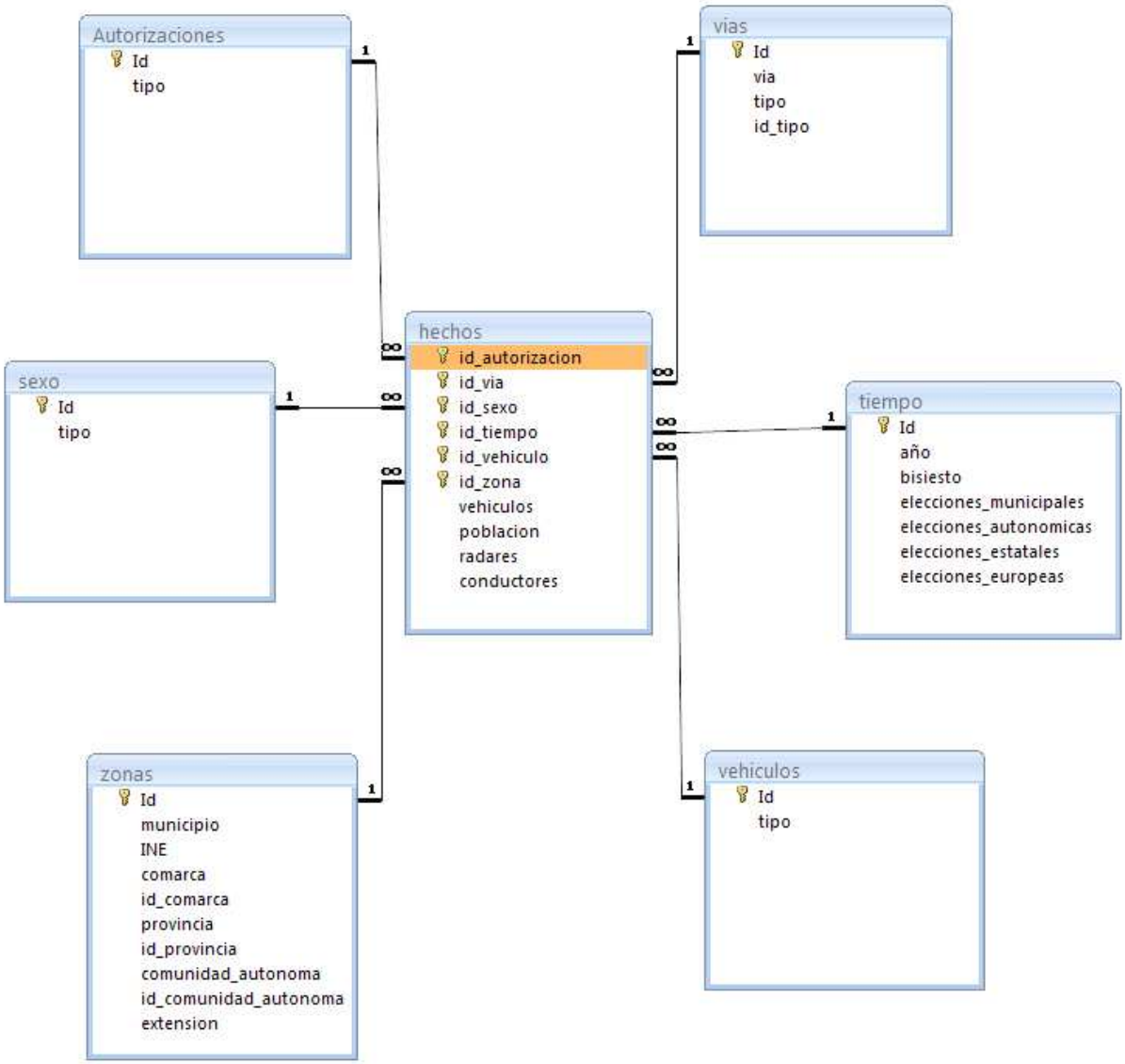

Ilustración 11 - Modelo entidad relación

3.4 Diseño y descripción de los informes creados

Se han diseñado los informes de manera que se pueda consultar la información en tiempo real de manera agregada por año, provincia, comarca, tipo de vehículo, tipo de permiso de conducir y sexo, esta estructura de consulta resulta muy ágil e intuitiva de usar.

La información se presenta inicialmente de manera anual y se puede ir desglosando por provincia, comarca, tipo de vehículo etc.

En algunos informes se han omitido los niveles de agregación que no representaban ninguna información.

Total Vehículos

Catalunya

Año	Provincia	Comarca	Permiso	Vehículo	Total vehiculos
⊞ 2007					4689267
⊞ 2008					4920415
⊟ 2009					4973967
	⊟ Barcelona				3578765
		⊞ Alt penedes			77238
		⊞ Anoia			83115
		⊞ Bages			132381
		⊟ Baix Llobregat			522391
			⊞ Licencia		4
			⊟ Permiso		522387
				Automovil	355430
				Camion o Furgoneta	71441
				Otros	95516
		⊞ Barcelonès			1343324
		⊞ Berguedà (Barcelona)			28228
		⊞ Garraf			99296
		⊞ Maresme			302113
		⊞ Osona (Barcelona)			110047
		⊞ Selva (Barcelona)			1254
		⊞ Vallès Occidental			594295
		⊞ Vallès Oriental			285083

Ilustración 12 - Informe total vehículos

En el informe total vehículos encontramos la información del número de vehículos totales, agregado por año, provincia, comarca, tipo de permiso de conducir y tipo de vehículo, viendo los datos del Baix Llobregat observamos que el número de licencias es extremadamente reducido y que la mayoría de vehículos son automóviles.

Total Conductores

Catalunya

Año	Provincia	Comarca	Permiso	Sexo	Conductores
⊞ 2007					3944503
⊞ 2008					3964262
⊟ 2009					3979683
	⊞ Barcelona				2943773
	⊟ Girona				405855
		⊞ Alt Empordá			62889
		⊞ Baix Empordà			75367
		⊞ Cerdanya (Girona)			6327
		⊞ Garrotxa			32517
		⊟ Gironès			103793
			⊞ Licencia		3
			⊟ Permiso		103790
				Hombre	59380
				Mujer	44410
		⊞ Osona (Girona)			586
		⊞ Pla de l'Estany			18125
		⊞ Ripollès			14207
		⊞ Selva (Girona)			92044
	⊞ Lleida				208235
	⊞ Tarragona				421820
⊞ 2010					4009332
⊞ 2011					4034820

Ilustración 13 - Informe total conductores

En el informe Total conductores podemos ver el número total de conductores, agregado por año, provincia, comarca, tipo de permiso y sexo.

Podemos ver como el número de conductores ha ido creciendo año tras año, y que la mayoría pertenecen a la provincia de Barcelona, en el caso de la comarca del Girones vemos que hay una mayor cantidad de hombres que tienen permiso de conducción y que la cifra de licencias es muy baja.

Porcentaje de vehículos respecto a población

Catalunya

Año	Provincia	Comarca	Vehículo	% vehículos
⊞ 2007				66.78%
⊞ 2008				68.60%
⊞ 2009				68.31%
⊞ 2010				67.85%
⊟ 2011				67.82%
	⊟ Barcelona			65.10%
		⊞ Alt penedes		77.02%
		⊞ Anoia		75.02%
		⊞ Bages		73.40%
		⊞ Baix Llobregat		65.10%
		⊟ Barcelonès		58.98%
			⊞ Automovil	36.44%
			⊞ Camion o Furgoneta	6.52%
			⊞ Licencia	0.00%
			⊞ Otros	16.02%
		⊞ Berguedà (Barcelona)		80.63%
		⊞ Garraf		68.45%
		⊞ Maresme		70.33%
		⊞ Osona (Barcelona)		77.32%
		⊞ Selva (Barcelona)		91.05%
		⊞ Vallès Occidental		67.19%
		⊞ Vallès Oriental		72.51%

Ilustración 14 - Informe porcentaje de vehículos respecto a población

En este informe podemos observar el porcentaje de vehículos respecto a la población total, la información ha sido agregada por año, provincia, comarca y tipo de vehículo.

Según los datos observados se hace evidente que el Barcelonès con casi un 59% de vehículos por habitante tiene las cifras más bajas de toda la provincia, esto puede ser debido a que dispone de unas infraestructuras de transporte público mejores que el resto de comarcas de la provincia.

Densidad de Población

Catalunya

Año	Provincia	Comarca	Sexo	Habitantes/Km2
⊞ 2007				438.24
⊞ 2008				447.61
⊞ 2009				454.44
⊞ 2010				456.71
⊞ 2011				458.04
⊟ 2012				459.99
	⊞ Barcelona			1115.59
	⊞ Girona			237.44
	⊞ Lleida			92.79
	⊟ Tarragona			192.06
		⊞ Alt Camp		169.28
		⊞ Baix Camp		417.44
		⊟ Baix Ebre		105.26
			Hombre	53.24
			Mujer	52.02
		⊞ Baix penedès		378.64
		⊞ Conca de Barberà		56.59
		⊞ Montsià		119.36
		⊞ Priorat		41.25
		⊞ Ribera d'Ebre		32.72
		⊞ Tarragonès		917.04
		⊞ Terra Alta		20.65

Ilustración 15 - Informe densidad de población

Este informe muestra el número de habitantes por kilometro cuadrado de manera agregada por año, provincia, comarca y sexo, viendo los datos de 2012 se aprecia como la provincia de Barcelona es la que mayor densidad posee y Lleida la que menos, también podemos ver esta densidad por comarca siendo el Tarragonès la comarca de Tarragona con mayor densidad, muy por encima de la media de la provincia.

En la comarca de Baix Ebre observamos como el número de hombres por kilometro cuadrado es ligeramente mayor que el de mujeres.

Numero de vehículos por radar

Catalunya

Año	Provincia	Comarca	Permiso	Vehículo	densidad
⊞ 2007					16888.85
⊞ 2008					17675.76
⊞ 2009					17813.82
⊟ 2010					17733.32
	⊞ Barcelona				29074.20
	⊞ Girona				7199.46
	⊞ Lleida				5108.07
	⊟ Tarragona				10250.04
		⊞ Alt Camp			2316.00
		⊞ Baix Camp			12099.36
		⊞ Baix Ebre			5688.88
		⊞ Baix penedès			11354.25
		⊞ Conca de Barberà			5451.00
		⊟ Montsià			5894.80
			⊞ Licencia		3.20
			⊟ Permiso		7858.67
				Automovil	3526.40
				Camion o Furgoneta	1426.80
				Otros	940.80
		⊞ Ribera d'Ebre			1723.33
		⊞ Tarragonès			27108.22
⊞ 2011					17741.67

Ilustración 16 - Informe número de vehículos por radar

El informe de número de vehículos por radar nos muestra el número de vehículos que existen por cada radar, la información está agregada por año, provincia, comarca, tipo de permiso y tipo de vehículo.

Analizando los datos vemos que para el año 2010 la provincia Barcelona es la que mas vehículos tiene por cada radar, viendo los datos de Montisià (Tarragona) vemos que el número de automóviles respecto al número de radares es mas de 3500 veces.

% de conductores por radar

Catalunya

Año	Provincia	Comarca	Permiso	Sexo	% conductores por radar
⊞ 2007					2029761.41%
⊞ 2008					2039928.99%
⊞ 2009					2047864.32%
⊟ 2010					2063121.10%
	⊞ Barcelona				3505230.77%
	⊞ Girona				1108016.22%
	⊞ Lleida				568329.73%
	⊟ Tarragona				1192085.58%
		⊟ Alt Camp			730400.00%
			⊞ Licencia		166.67%
			⊞ Permiso		730233.33%
		⊟ Baix Camp			1242768.00%
			⊟ Licencia		192.00%
				Hombre	180.00%
				Mujer	12.00%
			⊟ Permiso		1242576.00%
				Hombre	730704.00%
				Mujer	511872.00%
		⊟ Baix Ebre			609437.50%
			⊞ Licencia		200.00%
			⊞ Permiso		609237.50%
		⊟ Baix penedès			1318725.00%
			⊞ Licencia		250.00%
			⊞ Permiso		1318475.00%
		⊞ Conca de			1008400.00%

Ilustración 17 - Informe porcentaje de conductores por radar

El siguiente informe nos muestra de forma agregada por año, provincia, comarca, permiso y sexo el porcentaje de conductores por radar, podemos apreciar que el número de conductores respecto al de radares aumenta cada año, siendo la provincia de Barcelona donde hay más conductores por cada radar.

Indicador conductores vs habitantes

Catalunya

Año	Provincia	Comarca	Permiso	Sexo	Habitantes por conductor
⊞ 2007					1.78
⊟ 2008					1.81
	⊟ Barcelona				1.83
		⊞ Alt penedes			1.67
		⊞ Anoia			1.67
		⊞ Bages			1.70
		⊞ Baix Llobregat			1.80
		⊟ Barcelonès			1.99
			⊟ Licencia		124198.78
				Hombre	132718.75
				Mujer	1173828.00
			⊟ Permiso		1.00
				Hombre	1.56
				Mujer	2.68
		⊞ Berguedà (Barcelona)			1.68
		⊞ Garraf			1.82
		⊞ Maresme			1.73
		⊞ Osona (Barcelona)			1.63
		⊞ Selva (Barcelona)			1.74
		⊞ Vallès Occidental			1.73
		⊞ Vallès Oriental			1.69

Ilustración 18 - Informe indicador conductores vs habitantes

En este informe vemos el número de habitantes por cada conductor, este tipo de indicador nos permite evaluar el número de habitantes que tienes permisos de conducción, es significativo que en el caso del barcelonés haya 2,68 mujeres por cada mujer con permiso y solo 1.56 hombres por cada hombre con permiso.

Indicador radares vs vehículos

Catalunya

Año	Provincia	Comarca	Permiso	Vehiculo	Vehículos por radar
⊞ 2007					30986
⊞ 2008					32514
⊟ 2009					32868
	⊞ Barcelona				46178
	⊞ Girona				26021
	⊞ Lleida				14937
	⊟ Tarragona				16371
		⊟ Alt Camp			14277
			⊞ Licencia		5
			⊟ Permiso		14275
				Automovil	9135
				Camion o Furgoneta	2610
				Otros	2530
		⊟ Baix Camp			16170
			⊞ Licencia		8
			⊟ Permiso		16168
				Automovil	10224
				Camion o Furgoneta	2786
				Otros	3157
		⊞ Baix Ebre			7870
		⊞ Baix penedès			16488
		⊞ Conca de Barberà			13771
		⊞ Montsià			10367

Ilustración 19 - informe indicador radares vs vehículos

Este informe nos muestra el número de vehículos por radar que existe, la información se encuentra agregada por año, provincia, comarca, tipo de permiso y tipo de vehículo.

Ratio vehiculos por conductor

Catalunya

Año	Provincia	Comarca	Permiso	Vehiculo	Vehiculos por cada conductor
⊞ 2007					1.59
⊞ 2008					1.66
⊞ 2009					1.67
⊟ 2010					1.65
	⊞ Barcelona				1.61
	⊞ Girona				1.86
	⊞ Lleida				1.71
	⊟ Tarragona				1.73
		⊞ Alt Camp			1.75
		⊞ Baix Camp			1.73
		⊟ Baix Ebre			1.72
			⊞ Licencia		4.00
			⊟ Permiso		1.72
				Automovil	0.96
				Camion o Furgoneta	0.39
				Otros	0.38
		⊞ Baix penedès			1.69
		⊞ Conca de Barberà			1.84
		⊞ Montsià			1.74
		⊞ Priorat			1.93
		⊞ Ribera d'Ebre			1.92

Ilustración 20 - informe ratio de vehículos por conductor

Este informe nos muestra el número de vehículos que existen por cada conductor, la información se puede consultar de manera agregada por año, provincia, comarca, tipo de permiso, tipo de vehículo.

Densidad vehículos Km2

Catalunya

Año	Provincia	Comarca	Tipos de Vehiculo	Vehiculos por Km2
⊞ 2007				292.64
⊞ 2008				307.07
⊟ 2009				310.41
	⊟ Barcelona			723.86
		⊞ Alt penedes		140.69
		⊞ Anoia		230.24
		⊞ Bages		158.73
		⊞ Baix Llobregat		1074.88
		⊞ Barcelonès		9328.64
		⊞ Berguedà (Barcelona)		103.02
		⊞ Garraf		536.74
		⊞ Maresme		766.78
		⊞ Osona (Barcelona)		221.42
		⊟ Selva (Barcelona)		39.19
			Automovil	22.22
			Camion o Furgoneta	8.53
			Licencia	0.00
			Otros	8.44
		⊞ Vallès Occidental		1088.45
		⊞ Vallès Oriental		444.05
	⊞ Girona			192.49

Ilustración 21 - Informe densidad de vehículos por Km2

En el siguiente informe se puede ver el número de vehículos por cada kilometro cuadrado de territorio, esta información está agregada por año, provincia, comarca, tipo de vehículo.

Analizando los datos se ve que la comarca del barcelonés presenta un gran número de vehículos por cada kilometro cuadrado, son datos esperados ya que también es una comarca con una gran densidad de población.

Radares por tipo de via

Provincia	Tipo de vía	Radares
⊟ Barcelona		99
	Autonómicas	62
	Estatal	21
	Provincial	16
⊟ Girona		37
	Autonómicas	20
	Estatal	13
	Provincial	4
⊟ Lleida		38
	Autonómicas	15
	Estatal	20
	Provincial	3
⊟ Tarragona		41
	Autonómicas	12
	Estatal	22
	Provincial	7

Ilustración 22 -Informe de radares por tipo de vía

En el siguiente informe se puede ver una clasificación de los radares por provincia y tipo de vía, la información de tipo de vía se ha obtenido a través del nombre de las vías.

4 Conclusiones

El desarrollo del proyecto ha supuesto una inversión de tiempo mayor de la esperada, las previsiones de tiempo realizadas en la primera entrega se han quedado cortas en muchos puntos, sobretodo en el aprendizaje del software, aun así ha sido gratificante todo este proceso por los conocimientos adquiridos.

Durante la creación del proyecto ha quedado patente la necesidad de un análisis inicial, que permite ver de antemano los problemas y sus posibles soluciones así como definir las estrategias y plazos de entrega, también se ha mostrado las ventajas de trabajar en un proceso con un ciclo de desarrollo en cascada, ya que ha permitido conseguir los objetivos fijados

5 Líneas de Evolución futuras

- **Automatización ETL:** Aunque la ETL ha sido preparada para servir tanto para la carga inicial como para siguientes cargas, no se ha desarrollado ningún proceso que active estas siguientes cargas cuando hayan nuevos datos.

- **Mejorar los informes:** Los informes creados permiten consultar los datos de manera dinámica, pero resulta complicado hacer comparaciones de datos por años, se podrían agregar filtros dinámicos para permitir mostrar información de una sola comarca. También sería interesante agregar gráficos a los resultados parciales y mostrar % acompañando a los valores totales para ver que porcentaje representa esa comarca sobre el valor de una provincia.

- **Uso de roles:** a pesar que estaban en el planteamiento inicial no se ha podido usar los perfiles de usuario para el acceso restringido al aplicativo.

- **Más datos sobre los que trabajar:** Se podría ampliar la información que se dispone para realizar nuevos estudios.
    - **Franjas de edades población/conductores:** Si se tuviera la población y los conductores segmentados en franjas de edades se podría analizar el envejecimiento de la población/conductores año tras año.

    - **Tipo de motor:** Sería interesante saber el número de coches de gasolina, diesel, híbridos y eléctricos, se podría estudiar tendencias de mercado y ver la evolución anual.

    - **Edad del vehículo:** Podría aportar información sobre el envejecimiento del parque de vehículos.

    - **Accidentes:** si se tuviera información de accidentes, podría asociarse con las franjas de edades de los conductores, la edad del vehículo y el tipo de motor para sacar informes de seguridad.

6 Glosario

Inteligencia empresarial: Del inglés *Business Intelligence* o BI, es el conjunto de estrategias que permiten la toma de decisiones teniendo en cuenta los datos de una organización.

Factoría de la información corporativa: También denominada con sus siglas, FIC, es un conjunto de componentes que interactúan para ayudar a gestionar los flujos de datos de los sistemas operacionales de la empresa.

Almacén de datos: Del inglés *Data Warehouse*, es una colección de datos que integra toda la información referente a una empresa de manera no volátil.

ETL: Acrónimo en inglés de *Extract, Transform and Load,* es el proceso que consiste en extraer la información de las fuentes de datos, transformarla y cargarla en una base de datos.

SDLC: Acrónimo en inglés de *System Development Life Cycle*, define el ciclo de vida del desarrollo del proyecto.

OLAP: Acrónimo en inglés de *On-Line Analytical Processing*, solución que permite agilizar las consultas de grandes cantidades de datos.

ROLAP: Es un tipo de implementación OLAP que almacena los datos en un motor relacional.

MOLAP: Es un tipo de implementación OLAP que almacena los datos en una estructura multidimensional.

HOLAP: Es una tecnología OLAP híbrida mezcla de ROLAP Y MOLAP.

7 Bibliografía

7.1 Apuntes

Rius, Àngels; Serra, Montse; Abelló, Alberto; Samos, José; Vidal, Josep; Curto, Josep Almacén de datos y modelos multidimensionales, segunda edición, FUOC, Barcelona, 2012

Gironés, Jordi; Conesa, Jordi; Masip, David Data mining, primera edición, FUOC, Barcelona, 2013

7.2 WEB

Definición de requisitos funcionales y no funcionales:

(09/2013) http://synergix.wordpress.com/2008/07/07/requisito-funcional-y-no-funcional/

Definiciones y teoría de Tecnologías OLAP:

(09/2013) http://es.wikipedia.org/wiki/ROLAP

(09/2013) http://es.wikipedia.org/wiki/MOLAP

(09/2013) http://es.wikipedia.org/wiki/HOLAP

(10/2013) http://www.bi.dev42.es/2011/02/23/olap-molap-rolap/

Información sobre data warehouse

(09/2013) http://www.businessintelligence.info/serie-dwh/tablas-agregadas.html

Diseño de estrella vs copo de nieve:

(11/2013)http://informationmanagement.wordpress.com/2007/11/19/diseno-de-un-data-warehouse-estrella-y-copo-de-nieve/

Diseño de Lógico de bases de datos:

(10/2013) http://www.slideshare.net/argenis85/diseo-lgico-de-base-de-datos

Definición granularidad

(10/2013) http://es.wikipedia.org/wiki/Granularidad

8 Anexos

A. Análisis de riesgos

En la siguiente tabla se muestra un análisis de incidencias y como se solucionarían.

Incidencia	Plan de contingencia
Perdida de una a tres jornadas por imprevistos, obligaciones o enfermedades	El tiempo será recuperado en las siguientes jornadas haciendo un esfuerzo para disponer de más horas.
Pérdida de más de tres jornadas	El tiempo se recuperará cogiendo días de vacaciones guardados para tal efecto.
Horas extras en el trabajo de manera continuada	Se deberá pasar más carga de trabajo al fin de semana, viendo modificada la disponibilidad horaria.
Fallo en servidor virtual	Se realizaran faenas de documentación o investigación que no dependan de la conectividad con el servidor
Fallo en el equipo habitual	Se hará uso de un segundo equipo.
Perdida de datos o documentación en el equipo habitual.	Se recuperará una versión anterior almacenada en la nube.
Perdida de datos o documentación en el equipo virtual.	Se recuperará una versión anterior almacenada en local.

Tabla 9 – Descripción de incidencias y planes de contingencia

B. Estimación de tiempo disponible

El tiempo estimado que dispongo es de unas 5 horas entre semana y 11 el fin de semana, por lo que he distribuido la mayor carga de trabajo durante el fin de semana.

- De lunes a viernes 1h, de 22:30 a 23:30
- Sábados 5h, de 15:00 a 20:00
- Domingos 6h, de 11:00 a 13:00 y de 16:00 a 20:00

Los días festivos de este semestre y la posible disponibilidad.

- 1 de noviembre 5h, de 15:00 a 20:00
- 6 de diciembre 5h, de 15:00 a 20:00
- 25 de diciembre 0h
- 26 de diciembre 0h
- 1 de enero 0h
- 6 de enero 6h, de 11:00 a 13:00 y de 16:00 a 20:00

C. Viabilidad:

El modelo planteado en forma de estrella requerirá como máximo un poco más de un Terabyte de espacio en memoria por lo que la viabilidad está asegurada.

Ha que tener en cuenta que el cálculo de bytes por registro se realiza sobre el máximo posible y que este no se llegará a alcanzar ya que la base de datos gestiona el espacio en caso que los datos no ocupen el espacio asignado, como por ejemplo en los tipos de datos Varchar.

Tablas	Campos	Bytes por registro	Registros	Total Bytes	Total kB	Total MB
Tiempo	7	9	6	54	0,05	0,00
Zonas	10	219	947	207393	202,53	0,20
Vías	4	41	217	8897	8,69	0,01
Vehículos	2	23	8	184	0,18	0,00
Autorizaciones	2	18	2	36	0,04	0,00
Sexo	2	13	2	26	0,03	0,00
Hechos	10	28	39.455.808	1.104.762.624,00	1.078.869,75	1.053,58
					Total:	**1.053,79**

Tabla 10 – Tabla de estudio de viabilidad

Otro factor que afectará al tamaño es el número de registros, el número representado es el máximo posible, si el número de registros de una tabla de dimensiones es menor al máximo el número de registros de hechos disminuirá proporcionalmente, se puede realizar un cálculo sencillo haciendo la resta del valor máximo con el valor real y multiplicando el resultado con el valor obtenido de multiplicar el número de registros de las demás dimensiones.

Otro factor a tener en cuenta es la evolución de estos datos en el tiempo, si bien en algunas dimensiones los valores máximos de registros no es muy probable que varíen (zonas, vehículos, autorizaciones o sexo) en otros como en tiempo es seguro que aumentaran, con cada nuevo año estaríamos aumentando como máximo en 6.575.968 de nuevos registros la tabla hechos lo que supondría un aumento máximo de unos 175 MB anuales.

D. Análisis de las tecnologías a utilizar

He considerado adecuado trabajar con la herramienta Pentaho Data Integration ya que es el software con el que estoy más familiarizado para el desarrollo del proceso ETL, esto me facilitará avanzar más rápidamente y dedicar más tiempo a otros aspectos de la práctica.

En la elección del método OLAP a utilizar lo he tenido más complicado, ya que aunque poseo experiencia en bases de datos relacionales y eso me facilitaría el desarrollo usando la tecnología ROLAP creo que es una buena oportunidad para iniciarme en un modelo de datos multidimensional.

A fin de ayudarme en la decisión he realizado una tabla comparativa con las principales características de las tecnologías MOLAP y ROLAP.

Además de estos dos sistemas OLAP existe el HOLAP que es un hibrido de ambos, combinando sus principales atributos lo que le permite almacenar las agregaciones en una estructura multidimensional cómo MOLAP, mientras que las consultas sobre los datos de origen se realizan cómo ROLAP, mejorando con ello el rendimiento y permitiendo una mayor escalabilidad.

Es por estas razones que me he decidido a utilizar la tecnología HOLAP.

Para su desarrollo usaré Microsoft SQL Server Analysis Services, ya que permite trabajar con los tres métodos OLAP descritos, lo cual es interesante para poder realizar comparativas de eficiencia y facilitaría el paso de un sistema a otro si fuera necesario.

Tabla comparativa MOLAP vs ROLAP

	MOLAP	ROLAP
Rendimiento	Su rendimiento es superior debido a la indexación multidimensional y la memoria cache.	Rendimiento lento, depende de tablas agregadas para mejorarlo, pero no es práctico realizar toda la combinatoria de casos.
Tamaño	Usa técnicas de compresión que le permiten ocupar un menor espacio, pero introduce redundancia de datos.	El uso de tablas agregadas y la falta de compresión hacen que tenga un mayor tamaño.
Funcionalidad	Permiten operaciones específicas sobre el cubo para potenciar su análisis.	Están limitadas por las funcionalidades del lenguaje SQL, aunque van surgiendo cada vez más instrucciones especificas como CUBE y ROLLUP.
Escalabilidad	Los cubos deben generarse con cada nueva información, ralentización con modelos con dimensiones muy altas.	Es más escalable y soporta mejor grandes volúmenes de datos con alta cardinalidad.

Tabla 11 – Tabla comparativa de las tecnologías MOLAP vs ROLAP

E. Análisis de los datos de origen y calidad de la información

Datos de conductores

A pesar de la denominación de los archivos como datos de conductores, la DGT ha proporcionado datos de los permisos/licencia agrupando por municipio y sexo desde el año 2007 hasta 2011, en total 5 archivos de texto donde la información ha sido organizada en 5 columnas separadas por tabulaciones, en la primera columna podemos encontrar los datos de los municipios precedido por el nombre de la provincia, las siguientes dos columnas están formadas por el número de permisos pertenecientes a mujeres y a hombres respectivamente, de igual manera se organizan las dos últimas columnas con la información de licencias.

Calidad de la información.

Analizando los datos de los cinco archivos vemos que hay varios casos a tener en cuenta.

Diferentes formatos de datos: Los datos del archivo de 2011 tienen un formato diferente al del resto de años. Los campos no están separados por tabulaciones si no que por punto y coma, pero por error algunas de estas se encuentran entre comillas lo que evita que puedan ser interpretadas las diferentes columnas.

Solución: Si se eliminan todas las comillas del archivo este puede ser interpretado correctamente y su información no se verá.

Nuevos municipios: El municipio de la Canonja fue constituido como nuevo municipio en Abril de 2010, por lo que no se disponen de datos anteriores, este municipio formaba parte del municipio de Tarragona.

Solución: La falta de datos no se limita solo a los permisos/licencias, es por esto que este municipio no se tendrá en consideración.

Cambio de legislación en ciclomotores: A partir del año 2008 las licencias de ciclomotor dejan de considerarse como tales y pasan a ser permisos.

Solución: Pasaremos el número de licencias de ciclomotor del año 2007 a permisos, como no disponemos de datos fiables para saber el número de licencias de ciclomotor las calcularemos restando al número de licencias de 2007 el número de licencias de 2008, como resultado obtendremos el número de licencias que pertenece a ciclomotores y lo sumaremos a permisos, dejando el valor de licencias de 2007 idéntico al de 2008.

Municipios sin especificar: Dentro del conjunto de datos hayamos para cada provincia un campo de “municipio sin especificar” que aporta información sobre uno o varios municipios sin que podamos determinar cuáles.

Solución: Esta información no será tratada ya que no se puede relacionar con ningún municipio.

Ausencia de datos anuales: En los archivos proporcionados por la DGT no se encuentra la estadística de permisos y licencias de 2012, esta estadística no se puede obtener de su web ya que no está disponible.

Solución: Analizando el actual panorama económico sería osado calcular los datos de 2012 basándonos en los datos que tenemos de los años más recientes, es por esto que los datos serán considerados como NULL.

Datos de Municipios:

Los datos de los municipios nos vienen dados por el archivo Dades_municipis.xls, que se encuentra en formato Excel, está formado por nueve columnas, la primera corresponde a los municipios, seguida por el código INE de estos, después encontramos 6 columnas más con la población por años desde el 2012 al 2007, por último se hallan los datos de extensión calculados en Km^2.

Calidad de la información.

Nuevos municipios: El municipio de la Canonja fue constituido como nuevo municipio en Abril de 2010, por lo que no se disponen de datos anteriores, este municipio formaba parte del municipio de Tarragona.

Solución: La falta de datos no se limita solo a los datos de habitantes, es por esto que este municipio no se tendrá en consideración.

Falta de datos por sexos: No se disponen de datos por sexos.

Solución: Estos datos se conseguirán de una fuente externa que los proporciones, y serán mostrados cómo porcentaje.

Falta de municipios: Solo se dispone de información de 466 municipios del total de 947 que pertenecen a Cataluña.

Solución: Los datos del resto de municipios será almacenado como NULL para no perder información referente al resto de contenidos que si disponen información de estos municipios.

Municipio que pertenece a una provincia diferente que su comarca:
Algunos municipios de la Cerdanya pertenecen a Lleida:

- Bellver de Cerdanya
- Lles de Cerdanya
- Montellà i Martinet
- Prats i Sampsor,
- Prullans
- Riu de Cerdanya.

Y otros a Girona:

- Alp
- Bolvir
- Das
- Fontanals de Cerdanya
- Ger
- Guils de Cerdanya
- Isòvol
- Llívia
- Meranges
- Puigcerdà
- Urús

Solución: Para no crear confusión a efectos prácticos se diferenciara la Cerdanya perteneciente de Lleida con la perteneciente a Girona.

Datos de vehículos

Los datos de Vehículos nos vienen dados por el archivo de Excel Dades_vehicles.xls, está formado por 29 columnas, la primera de ellas muestra el nombre de las comarcas y seguido de el código INE. Las cinco columnas siguiente muestran información sobre el número de vehículos a motor entre los años 2011 y 2007, posteriormente encontramos seis columnas más con el número de automóviles desde el año 2012 hasta el año 2007 seguidos por otras tantas columnas con la información de los camiones y furgonetas para el mismo periodo de tiempo, la estructura se repite con los datos de otros vehículos a motor, por ultimo encontramos cuatro columnas con los datos de motocicletas, autobuses, tractores industriales y resto de vehículos.

Calidad de la información

Datos no disponibles por municipio: El municipio de la Canonja solo dispone información para el año 2012

Solución: La falta de datos en todos los ámbitos hace que no lo tengamos en consideración.

Error en datos Otros vehículos de motor 2012: Por error los datos que se muestran en otros vehículos de motor 2012 son los mismos que en el año 2011.

Solución: Se calcularan los datos de otros vehículos de motor 2012 con la

suma de las columnas Motocicletas, autobuses, tractores industriales, Resto de vehículos a motor, ya que pertenecen al periodo 2012.

Datos no disponibles en vehículos de motor 2012: no se dispone de la columna de vehículos de motor 2012.

Solución: No supone ningún problema ya que es un dato que no se usará directamente si no que se calculará con la suma de camiones y furgonetas, automóviles y otros vehículos a motor.

Falta de municipios: Solo se dispone de información de 466 municipios del total de 947 que pertenecen a Cataluña.

Solución: Los datos del resto de municipios será almacenados como NULL para no perder información referente al resto de contenidos que si disponen información de estos municipios.

Datos de Radares

Los datos de radares son proporcionados a través del archivo de texto Radars_SCT.txt , en este archivo encontraremos los datos organizados en cuatro columnas, la primera de ellas nos da el nombre de la vía donde está situado el radar, el resto de columnas nos dan información del municipio la comarca y la provincia respectivamente.

Calidad de la información

Número limitado de comarcas: En los datos de radares podemos encontrar la relación entre municipio y comarca, el resto de archivos no dispone de esta información por lo que solo podremos relacionar los municipios que dispongan de radares.

Solución: Los municipios que no se encuentren entre los proporcionados por radares tendrán un valor NULL para la comarca.

F. Fuentes de Información extras

Además de las fuentes de información proporcionadas por IDESCAT, la DGT, y el Servei Català de Trànsit ha sido necesario incorporar dos fuentes nuevas, una con la información sobre la distribución de la población por sexos, la otra con información sobre los municipios comarcas y provincias, estos datos han sido extraídos de la web de IDESCAT con el objetivo de proporcionar información extra sobre la distribución de la población por sexos en los municipios, y la relación de los municipios con comarcas y provincias.

G. Modificaciones en las fuentes de información

Se ha analizado la necesidad de hacer modificaciones en los archivos de entrada, procurando tratar la mayor parte de anomalías desde el programa Spoon.

Ha sido necesario modificar el archivo Dades_conductors 2011.txt ya que la información estaba distribuida con un formato diferente al resto de archivos proporcionados por la DGT. Se han substituido los caracteres punto y coma (;) por tabulaciones, y se han suprimido las comillas dobles (“).

En el archivo Dades_vehicles.xls, se ha calculado el valor para vehículos de motor del año 2012 con los valores de Motocicletas, Autobuses, Tractores industriales y Resto de vehículos de motor, ya que la información proporcionada por IDESCAT era errónea respecto al año 2012 para el valor de vehículos de motor.

H. Proceso ETL

El proceso ETL se ha dividido principalmente en dos partes, por un lado la carga de datos auxiliares y por otro la carga real de datos, es decir los que formaran el cubo OLAP.

Carga auxiliar

Para la carga auxiliar de datos se ha generado una base de datos con el nombre etl, esta base de datos contiene la información que se ha ido generando mediante las transformaciones, todas sus tablas a excepción de tipo_vias han sido rellenadas con la información extraída de las fuentes de datos.

La información de la tabla tipo_vias ha sido generada para poder relacionar las vías con el tipo al que pertenecen, y así poderlas clasificar en estatales, autonómicas y provinciales.

id_tipo	compare_case	tipo
1	N	Estatal
1	A	Estatal
1	AP	Estatal
2	C	Autonómicas
3	B	Provincial
3	BP	Provincial
3	BV	Provincial
3	GI	Provincial
3	GIP	Provincial
3	GIV	Provincial
3	L	Provincial
3	LP	Provincial
3	LV	Provincial
3	T	Provincial
3	TP	Provincial
3	TV	Provincial

Ilustración 23. Tabla tipo_vías.

Carga Real

Para la carga Real de datos se ha generado una base de datos con la estructura presentada en las anteriores entregas, las tablas autorizaciones, sexo, tiempo y vehículos han sido rellenadas manualmente, mientras que las tablas vías, zonas y hechos han sido rellenadas con los datos generados en el proceso ETL.

Transformaciones

En este apartado explicaré las transformaciones que intervienen en el proceso ETL, expondré que datos cargan, que tablas atacan y los pasos más relevantes.

En la mayoría de transformaciones encontraremos que antes de un paso de unión o agrupación se ordenan los campos, estos pasos son obligatorios, incluso aunque se recuperen datos de base de datos ordenados por un campo, ya que la ordenación realizada por SQL y la herramienta de ordenación de Spoon no son equivalentes y esto podría provocar inconsistencia en los datos.

Muchos de los pasos finales utilizados en las transformaciones son de inserción/modificación, en todos ellos han sido definidos los campos que se compararan para saber si ese registro ya existe y hay que modificar los datos, o en caso contrario se han de insertar. Estos campos cumplirán la función de campos unique.

Municipios

La transformación municipios es la encarga de rellenar la tabla auxiliar municipios, en esta tabla almacenaremos los municipios, su código INE, su extensión, la relación con la tabla comarcas y con la tabla provincias y el nombre que tienen los municipios en las diferentes fuentes de información para poder hacer una relación de manera sencilla con estas fuentes en los siguientes pasos.

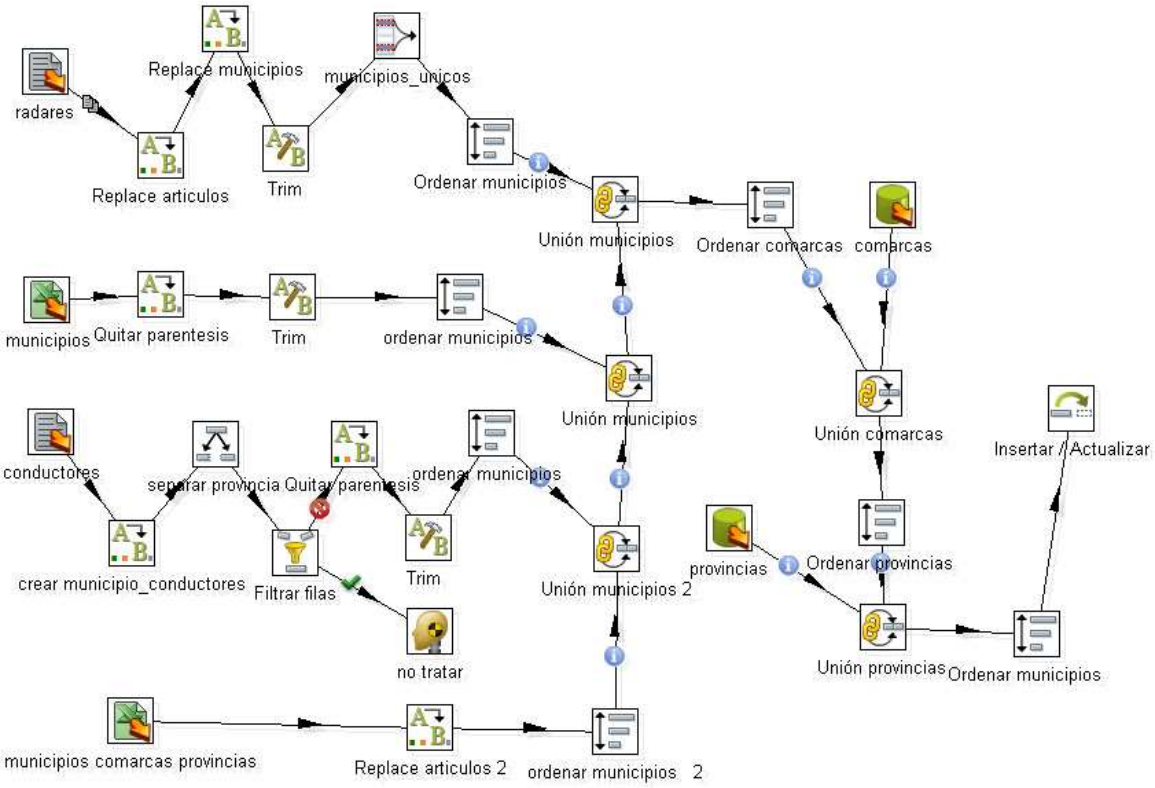

Ilustración 24. Transformación Municipios.

Cargaremos las fuentes de radars_SCT.txt, dades_municipis.xls, Dades_conductors 2007.txt. y municipios-comarcas-provincias.xls

Municipios-comarcas-provincia.xls nos ofrecen información sobre la relación entre municipios, comarcas y provincias, los datos de municipios.txt nos dan información sobre la extensión y el código INE , el resto de fuentes los cargamos para almacenar la relación con los municipios para el resto de cargas.

En las tres fuentes de datos proporcionadas por FERCRES duplicaremos el campo que las relaciona con municipios, este nuevo campo recibirá el nombre de municipio_X_join, siendo la X la primera letra de la fuente de datos, por ejemplo en radares quedaría municipio_r_join.

Al campo denominado municipio_X_join se le aplicaran las transformaciones necesarias para que se pueda unir al resto de fuentes, por ejemplo, eliminación de artículos, eliminación de texto entre paréntesis, eliminación de espacios al principio y al final e incluso modificar un nombre de municipio.

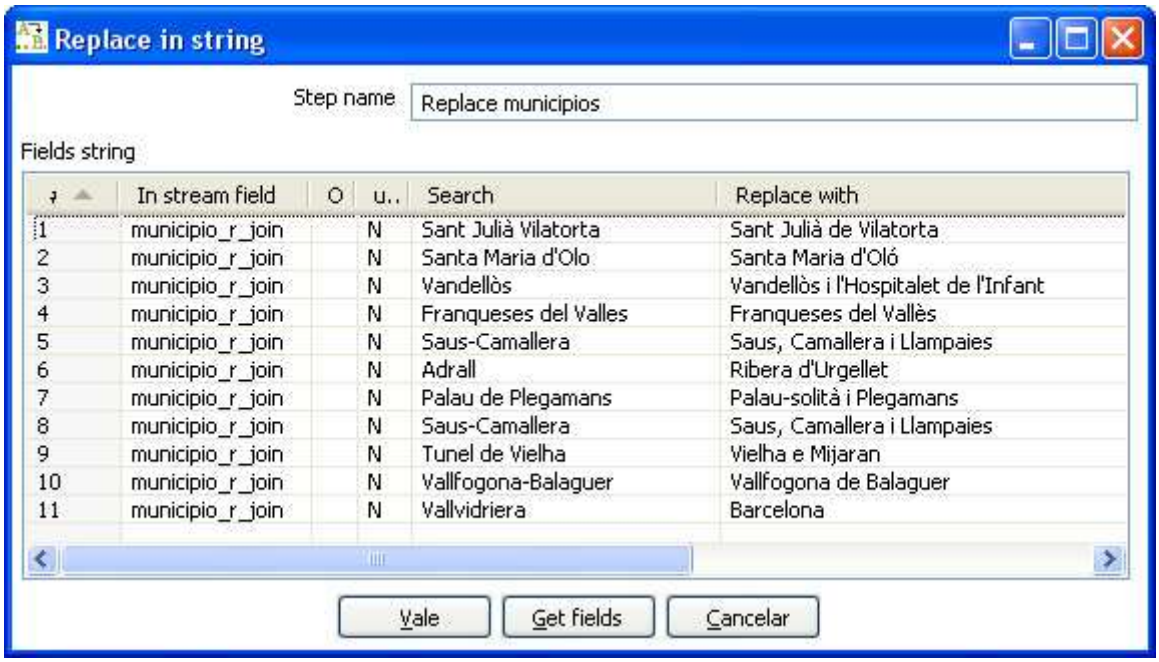

Ilustración 25. Ejemplo reemplazo nombres de municipio.

Los datos de conductores contienen filas de municipios sin especificar, estas filas serán rechazadas usando la herramienta filtrar filas y enviando los datos a una salida sin tratar.

Una vez se han unido los datos de las tres fuentes a través de los municipios y mediante uniones del tipo FULL OUTER, agregaremos los identificadores de comarcas y provincias mediante la lectura de sus tablas y con sendas uniones. Usaremos los datos que provienen de radares para la unión con comarcas y los que provienen de conductores para la unión con provincias, ya que en conductores todos los municipios tienen asociado una provincia.

De este proceso hay que destacar el uso de las uniones de tipo FULL OUTER, que nos permite generar un registro aunque no haya habido unión por parte de alguno de los pasos, esto es importante ya que no se perderán municipios si no están disponibles en alguna fuente, pero para estos municipios sin ambas uniones no se dispondrá de la información que proporcionaba la fuente que falta.

Provincia-comarcas

En esta transformación se cargaran los datos de la fuente de datos Radars_SCT.txt ya que contiene información sobre las comarcas y provincias, se agrupan los datos para que no haya repeticiones y se insertan/modifican en las tablas auxiliares de comarcas y provincias.

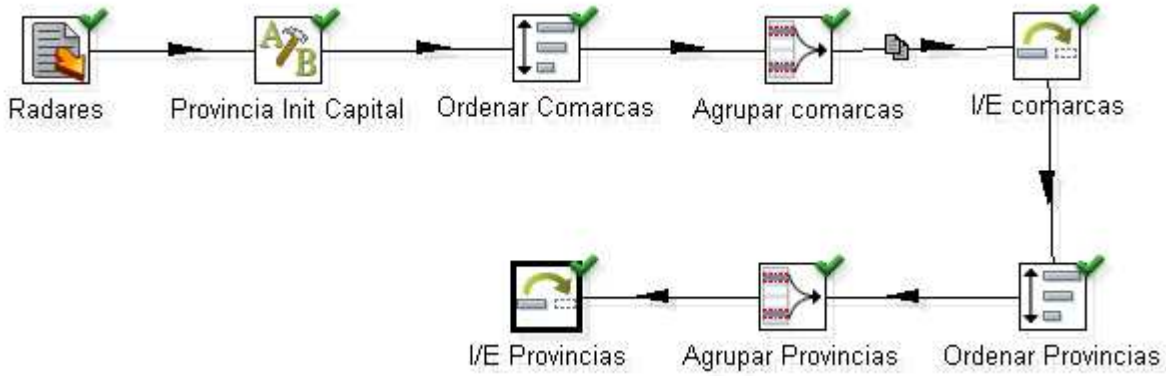

Ilustración 26. Transformación Provincia-comarcas

Habitantes

En esta transformación se cargan los datos de municipios y de porcentaje de sexos para generar la tabla auxiliar población, esta tabla estará relacionada con municipios mediante el id_municipio y con sexos mediante el id_sexo y contendrá el número de habitantes de cada sexo por año y municipio.

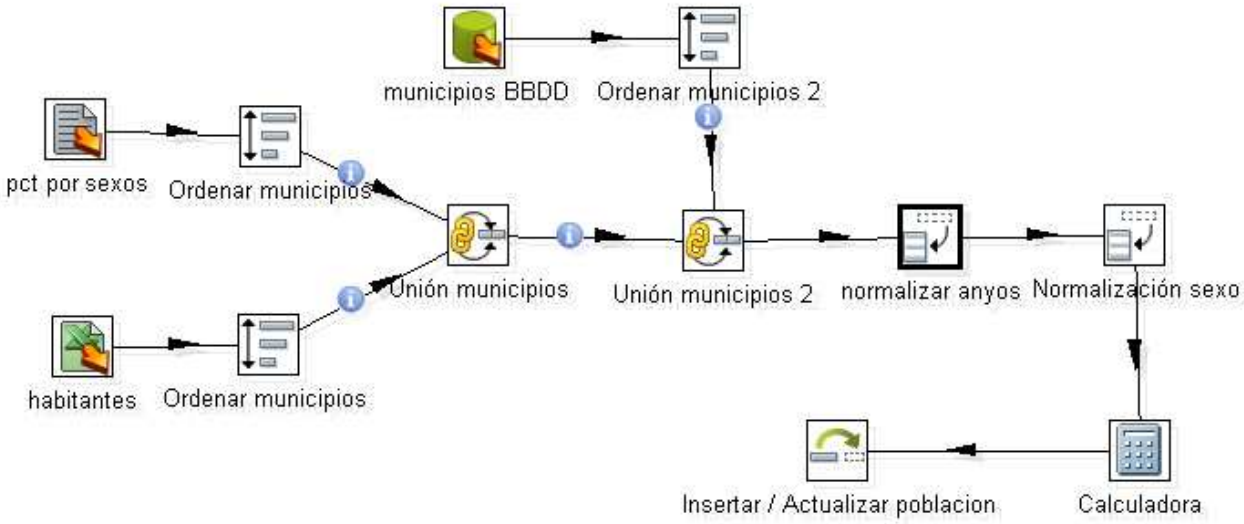

Ilustración 27. Transformación Habitantes

Cargaremos las fuentes de datos y una vez ordenadas las uniremos por nombre de municipio, no ha sido necesario ningún tratamiento ya que los nombres de municipio coincidían en todos los casos. Esta unión se realizará con el método INNER para evitar que haya porcentajes sin número de habitantes o número de habitantes sin porcentajes, una vez unidos los datos se procederá a realizar otra unión para conseguir el id_municipio asociado a los municipios, para ello cargaremos el id de municipios y el campo municipio_municipios, que había sido almacenado anteriormente para poder relacionar la tabla municipios con el archivo dades_municipis. Esta última unión se realizará con el método RIGHT OUTER para relacionar solo los municipios de los cuales tenemos información sobre su población.

Una vez obtenida la relación con la tabla municipios se realizará una normalización de los campos con valores anuales.

Ilustración 28. Normalización de los campos anuales

La normalización consiste en generar a partir de un conjunto de campos nuevos registros que contaran con dos campos nuevos, uno con el valor de los campos anteriores, y el otro con la categorización que se le otorgara a cada campo antiguo.

En el caso se ha generado a partir de los campos “población xxxx” donde xxxx es el año, dos campos nuevos, el campo población con el valor que contenían los antiguos campos, y el campo anyo con el valor del año al que pertenecen.

Para poder calcular la población por sexos necesitaríamos duplicar los registros actuales uno por cada sexo, es por ello que realizamos una nueva normalización,

esta vez sobre el campo porcentaje de hombres/mujeres, generando un nuevo campo llamado sexo. Y separando los porcentajes de población por sexos.

Ilustración 29. Normalización porcentajes por sexo

Una vez tenemos los registros duplicados calculamos el número de habitantes por sexo y lo insertamos/modificamos en la tabla población.

Vías

En la transformación vías, cargaremos datos de la tabla auxiliar tipo_vias y los datos del archivo radares_SCT.txt.

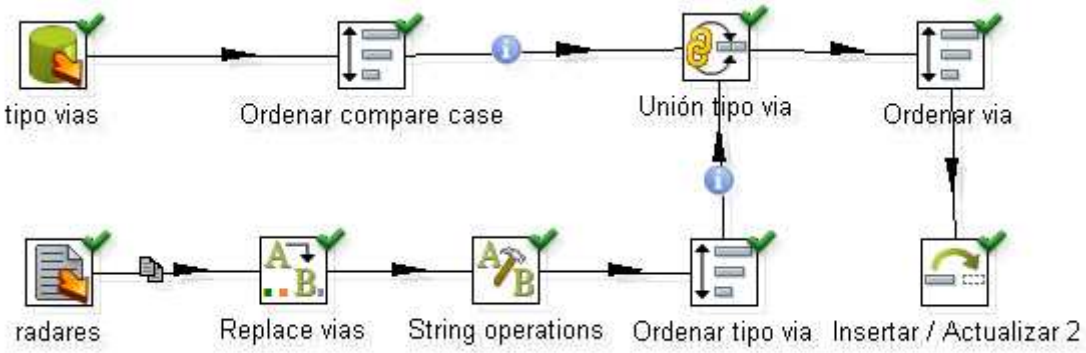

Ilustración 30. Transformación Vías

La tabla auxiliar tipo_vias contiene los tipos de vías, estatal, autonómicas y provincial, con sus respectivos identificadores y un campo llamado compare_case que servirá para hacer las comparaciones con el nombre de las vías y detectar a que tipo pertenecen.

Una vez cargados los datos de radares realizaremos una modificación con la herramienta remplace string sobre los datos de vías, creando un campo nuevo llamado tipo_via que contendrá solo las iníciales de las vías para poderlo comparar con el campo compare_case de la tabla tipo_vias, con esta misma herramienta eliminaremos los textos nord y sud de vías, también eliminaremos los espacios en blanco sobrantes del campo vías.

Una vez ordenados los campos tanto de la tabla como del archivo procederemos a unir los dos flujos comparando los tipos de vías del archivo con el campo compare_case, esta unión nos permitirá obtener el campo id_tipo y tipo_via para insertarlo en la tabla final vías.

Radares

La transformación radares es la encargada de contabilizar el número de radares que hay por cada vía en un municipio, estos datos los almacenará en la tabla auxiliar radares, donde guardará los identificadores de municipio y vías y el número de radares asociados.

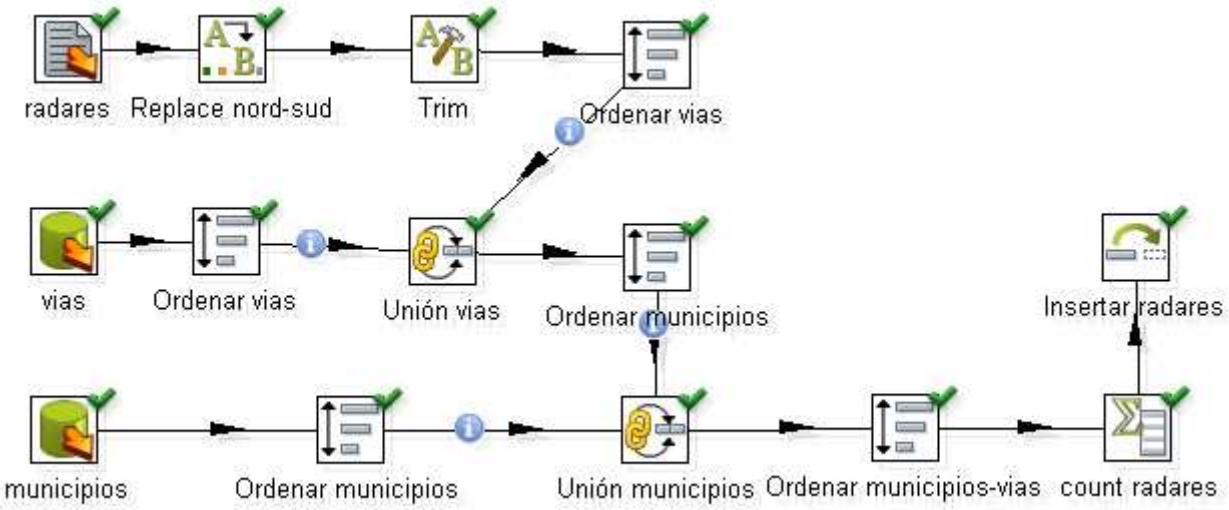

Ilustración 31. Transformación Radares.

Lo primero que se realizará será la unión de los radares con las vías creadas para obtener su identificador, una vez realizado este paso se unirán a los municipios usando para ello la clave almacenada en municipios municipio_radares. Una vez asociados los radares con las vías y los municipios usaremos la herramienta count para contar el número de radares por vía y municipio y se insertarán/modificaran los datos de la tabla radares.

Autorizaciones

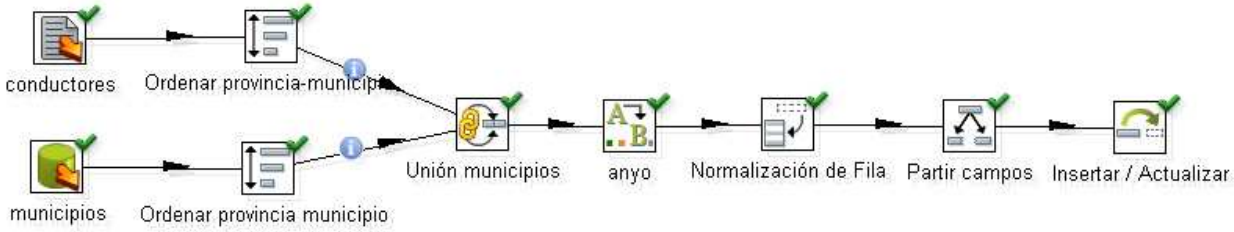

Ilustración 32. Transformación autorizaciones

Para generar el número de autorizaciones asociadas a un municipio y sexo cargaremos la fuente de datos de conductores, que se encuentra en 5 archivos, y la tabla de municipios, haremos la relación con la clave almacenada en municipios y con el nombre del municipio de conductores, una vez hecho esto extraeremos el año del nombre de los archivos.

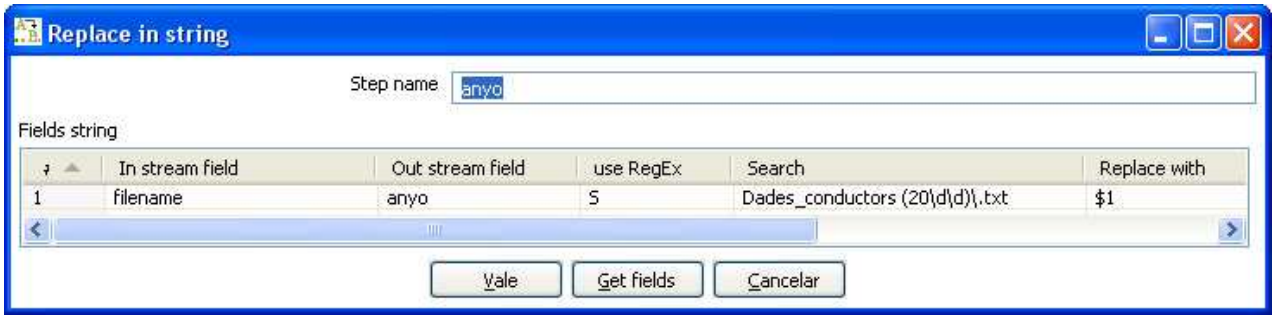

Ilustración 33. Creación del campo año.

Una vez obtenidos los años, usaremos la herramienta normalizar filas y partir campos para crear los campos de id_autorizacion e id_sexo.

Ilustración 34. Normalización de tipo-sexo

Ilustración 35. Partir campos id_autorizacion y id_sexo

Por último insertaremos/modificaremos los datos en la tabla auxiliar num_autorizaciones.

Update licencias 2007

Debido a un cambio de legislación los ciclomotores dejaron de necesitar licencia para su conducción a partir del año 2008, por este motivo los datos de licencias del año 2007 debían ser modificados. El proceso que se ha seguido ha sido recuperar las licencias de 2008 y sobrescribir este valor al de licencias de 2007, con las licencias sobrantes de 2007 se las pasó a permisos.

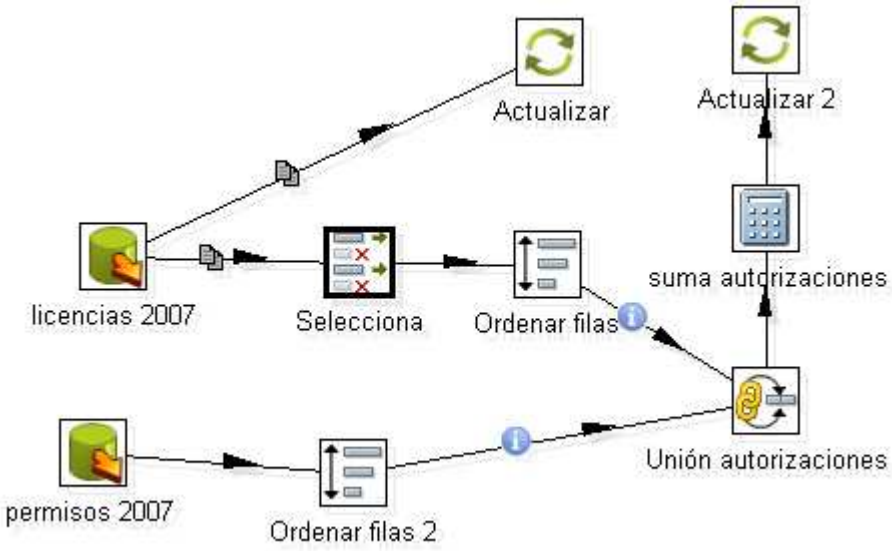

Ilustración 36. Transformación update licencias 2007

El paso licencias 2007 contiene una sentencia SQL que carga los datos de licencias de 2008 y el nuevo valor que habrá que sumarle a permisos de 2007.

Query:

```
select na_2007.anyo,na_2007.id_autorizacion, na_2007.id_municipio, na_2007.id_sexo, (na_2007.num_autorizaciones - na_2008.num_autorizaciones) as sum_permisos, na_2008.num_autorizaciones as autorizaciones_2008 from num_autorizaciones na_2007
JOIN num_autorizaciones AS na_2008 ON na_2008.id_municipio = na_2007.id_municipio AND na_2008.id_autorizacion = na_2007.id_autorizacion AND na_2008.id_sexo = na_2007.id_sexo
WHERE na_2007.anyo =2007 AND na_2008.anyo = 2008 AND na_2007.id_autorizacion =2;
```

Una vez cargados estos datos se modificaran los valores de licencias de 2007 en el paso actualizar, y se unirán los valores de permisos con los permisos ya existentes de 2007 cargados en el paso permisos 2007, una vez unificados los registros por año, municipio y sexo en el paso unión autorizaciones se realiza la suma de los permisos y se modifican los valores en la tabla num_autorizaciones.

Vehículos

Para poder relacionar vehículos y autorizaciones he tenido que crear una categoría nueva de vehículos llamada licencia, en esta categoría encontraremos los vehículos que requieren licencia, para saber el número de vehículos que pertenecen a esta categoría usaré el número de licencias, para reflejar este cambio en los datos que se disponen de vehículos restaremos este valor al actual otros.

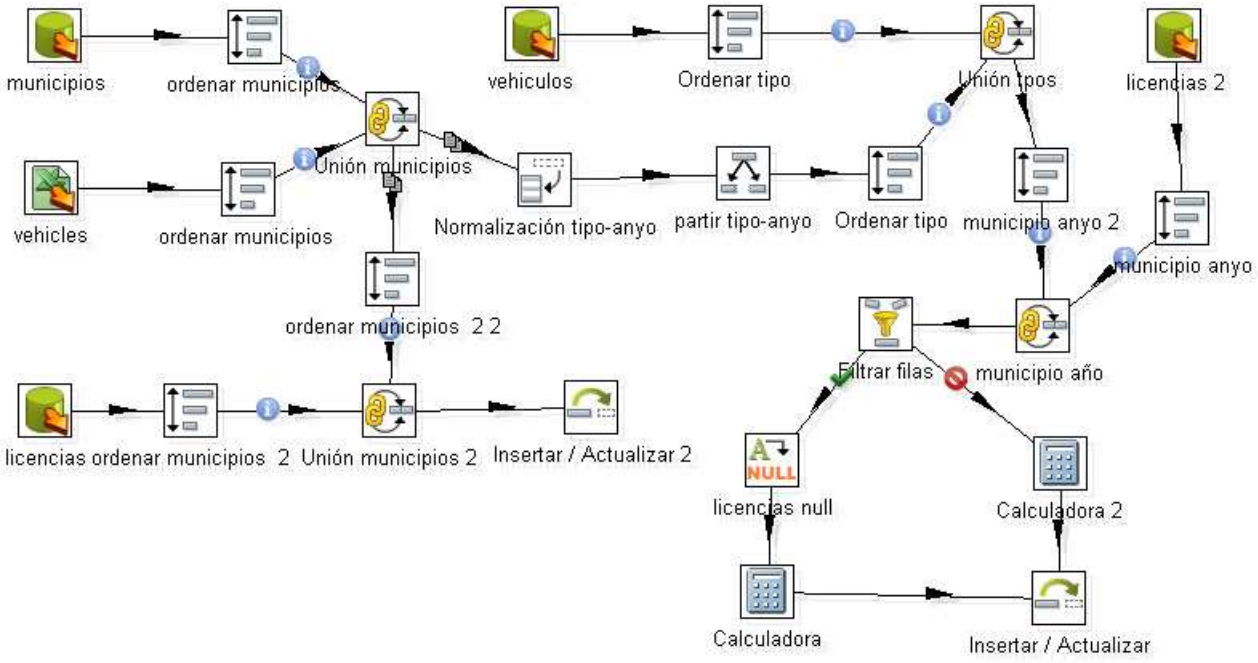

Ilustración 37. Transformación vehículos

La transformación vehículos consta de dos partes, en la primera intervienen la tabla auxiliar municipios el archivo de vehículos y la tabla auxiliar de num_autorizaciones, después de unir los vehículos con los municipios para obtener el identificador de municipios, unimos el resultado a los datos de licencias por años para generar los datos de los vehículos de tipo licencia por cada municipio y año.

La segunda parte de la transformación es la encargada de insertar el resto de tipos de vehículos.

Se ha continuado a partir de la unión de municipios y vehículos y se han normalizado los datos para obtener el tipo de vehículo y el año.

#	Nombre campo	Tipo
1	Automóviles 2007	Automovil-2007
2	Automóviles 2008	Automovil-2008
3	Automóviles 2009	Automovil-2009
4	Automóviles 2010	Automovil-2010
5	Automóviles 2011	Automovil-2011
6	Automóviles 2012	Automovil-2012
7	Camiones y furgonetas 2007	Camion o Furgoneta-2007
8	Camiones y furgonetas 2008	Camion o Furgoneta-2008
9	Camiones y furgonetas 2009	Camion o Furgoneta-2009
10	Camiones y furgonetas 2010	Camion o Furgoneta-2010
11	Camiones y furgonetas 2011	Camion o Furgoneta-2011
12	Camiones y furgonetas 2012	Camion o Furgoneta-2012
13	Otros vehículos de motor 2007	Otros-2007
14	Otros vehículos de motor 2008	Otros-2008
15	Otros vehículos de motor 2009	Otros-2009
16	Otros vehículos de motor 2010	Otros-2010
17	Otros vehículos de motor 2011	Otros-2011
18	Otros vehículos de motor 2012	Otros-2012

Ilustración 38. Normalización de filas tipo-año

Una vez obtenido el tipo-año se ha partido el campo generando el campo año y el campo tipo, este último campo se ha unido con la tabla vehículos para obtener el id_tipo, después se ha unido con la tabla auxiliar autorizaciones para añadir el campo de licencias y así poder restarlo a los registros con el tipo otros, para hacer esta resta solo a los registros correspondientes se ha usado el paso filtrar filas que se encargaba de comparar los tipos y distribuir el flujo. Una vez realizadas las operaciones se ha insertado/actualizado los datos en la tabla num_vehiculos.

Zonas

Esta transformación es la encargada de generar la tabla final zonas, lee los datos de las tablas municipios, comarcas y provincias y los almacena en zonas.

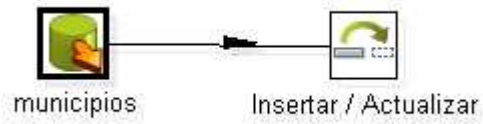

Ilustración 39. Transformación zonas.

Query:

```
SELECT municipios.id, municipios.municipio, municipios.INE,
municipios.id_comarca, municipios.id_provincia, municipios.extension,
comarcas.comarca,provincias.provincia FROM municipios
LEFT JOIN comarcas on comarcas.id = municipios.id_comarca
LEFT JOIN provincias on provincias.id = municipios.id_provincia
```

Generar hechos

Esta es la última transformación en ejecutarse.

Mediante una sentencia SQL unimos las tablas auxiliares para generar todos los hechos, una vez generados los insertamos en la tabla final.

Ilustración 40. Transformación hechos

Query:

```
SELECT municipios.id AS id_zona, coalesce(poblacion.id_sexo,3) id_sexo,
poblacion.id_tiempo, coalesce(num_autorizaciones.id_autorizacion,3)
id_autorizacion, coalesce(radares.id_via, 67) id_via,
coalesce(num_vehiculos.id_vehiculo,5) id_vehiculo, poblacion.habitantes AS
poblacion, radares.radares, IF(num_autorizaciones.num_autorizaciones <
aux_vehiculos.total_vehiculos, num_autorizaciones.num_autorizaciones,
aux_vehiculos.total_vehiculos)  AS conductores, num_vehiculos.num_vehiculos AS
vehiculos
FROM municipios
LEFT JOIN poblacion ON poblacion.id_municipio = municipios.id
LEFT JOIN radares ON radares.id_municipio = municipios.id
LEFT JOIN num_autorizaciones ON num_autorizaciones.id_municipio = municipios.id
AND num_autorizaciones.id_sexo = poblacion.id_sexo
AND num_autorizaciones.anyo = poblacion.id_tiempo
LEFT JOIN num_vehiculos ON num_vehiculos.id_municipio = municipios.id
AND num_vehiculos.anyo = poblacion.id_tiempo
AND num_vehiculos.id_autorizacion = num_autorizaciones.id_autorizacion
LEFT JOIN (select id_municipio,id_autorizacion, anyo, sum(num_vehiculos) AS
total_vehiculos FROM num_vehiculos GROUP BY  id_autorizacion,id_municipio,  anyo)
AS  aux_vehiculos ON aux_vehiculos.id_municipio = municipios.id
AND aux_vehiculos.anyo = poblacion.id_tiempo
AND aux_vehiculos.id_autorizacion = num_autorizaciones.id_autorizacion
```

Execute ETL

Execute ETL es el trabajo encargado de ejecutar de manera ordenada todo el proceso de ETL, las transformaciones han sido organizadas para que se cumplan los prerrequisitos necesarios antes de que se lancen.

Ilustración 31 - Trabajo que ejecuta todas las transformaciones

1. Cubo OLAP

Con la herramienta Microsoft Analysis Services se ha generado la estructura del cubo OLAP.

Se ha definido la base de datos tfg alojada en MySQL como fuente de datos, se han generado las dimensiones con las tablas de vías, vehículos, tiempo, sexo, autorizaciones y zonas, y se ha generado el cubo con las métricas de vehículos, población, radares y conductores pertenecientes a la tabla hechos y la métrica extensión perteneciente a la tabla zonas.

Se han definido tres tipos de roles para limitar el acceso a los usuarios según el perfil de estos.

Ilustración 42. Explorador de proyecto.

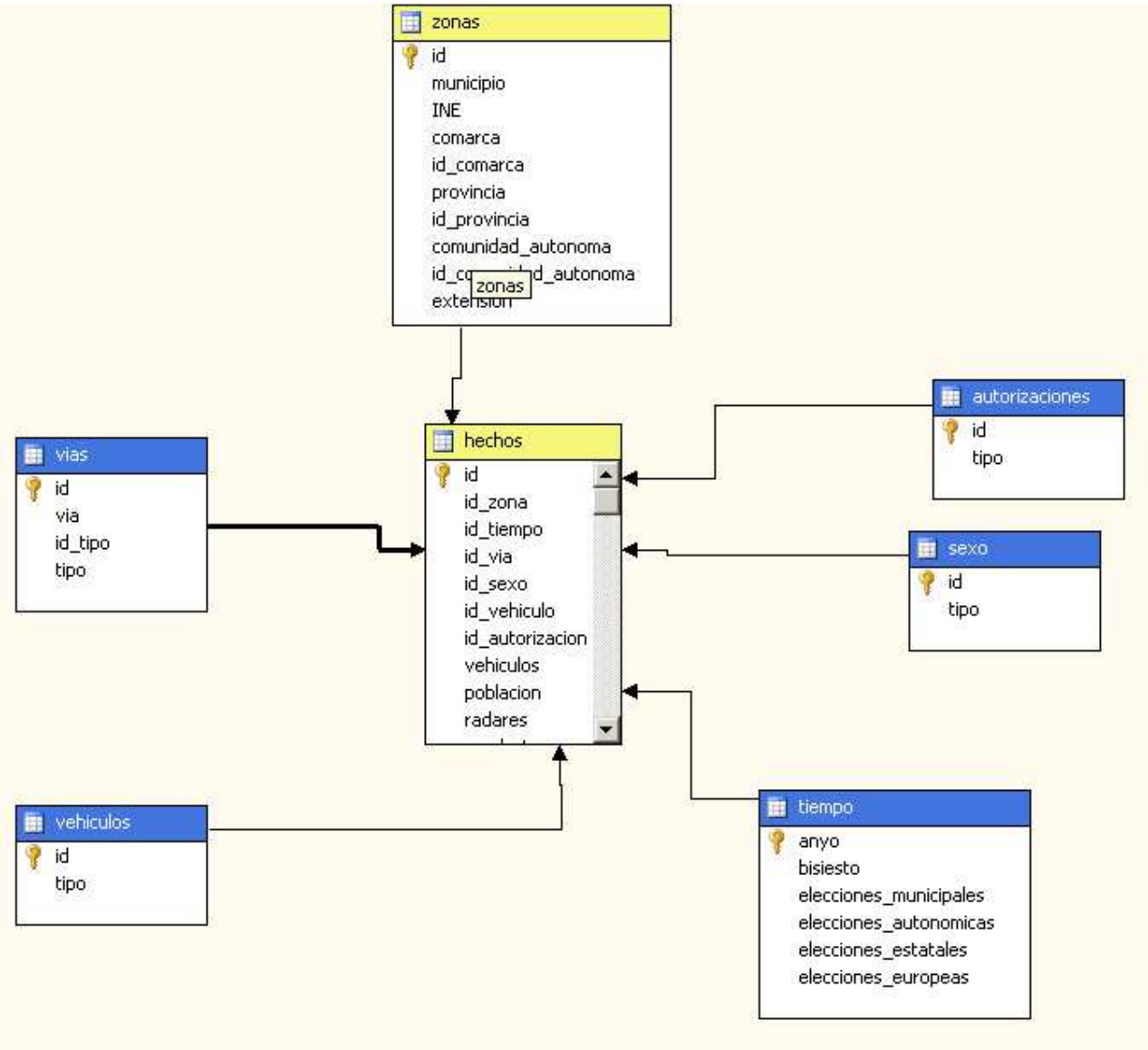

Ilustración 43 - Cubo OLAP

En las dimensiones de vías y zonas se han definido las jerarquías de agregación.

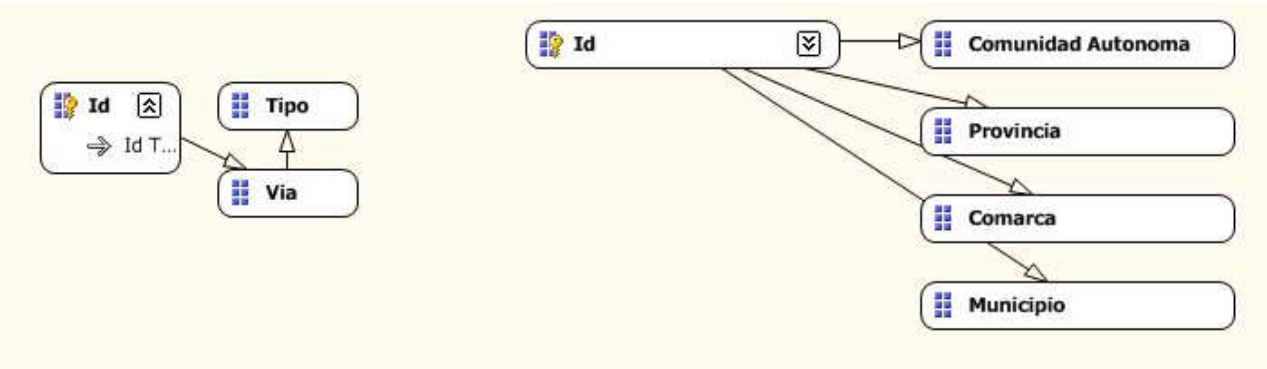

Ilustración 44 - Jerarquías de agregación de vías y municipios respectivamente