ALMACENES DE DATOS: Análisis de ventas de una compañía.

Abel Aránega Hernández
Máster Universitario en Ingeniería Informática

Víctor Ruíz Marques

16 de junio de 2015

FICHA DEL TRABAJO FINAL

Título del trabajo:	ALMACENES DE DATOS: Análisis de ventas de una compañía.
Nombre del autor:	Abel Aránega Hernández
Nombre del consultor:	Víctor Ruíz Marques
Fecha de entrega (mm/aaaa):	06/2015
Área del Trabajo Final:	Business Intelligence
Titulación:	*Máster en Ingeniería Informática*

Resumen del trabajo (máximo 250 palabras):

Hoy en día, las organizaciones manejan un flujo de información inimaginable hace tan sólo unos años. Basta con observar la valiosa información que atesoran grandes empresas como Google o Amazon que usan con fines comerciales para generar valor a sus clientes, al mismo tiempo que buscan incrementar sus ingresos.

Otras empresas, en cambio, tienen un gran volumen de información, pero no saben cómo analizarla. Surge así la necesidad de hacer uso de herramientas de análisis de datos ofrecidas por el Business Intelligence (BI).

Este trabajo intenta estudiar y documentar todo el proceso necesario para analizar los datos de una serie de fuentes de datos dadas por la empresa y poder contestar a una serie de preguntas de negocio. Para ello se han realizado tareas de análisis, tratamiento de fuentes de datos, creación de las bases de datos tanto para el repositorio de la herramienta ETL como para la construcción del *data warehouse,* del que se alimentará el cubo OLAP, que finalmente se utilizará para la generación de informes de análisis de estos datos.

Este trabajo me ha permitido entrar en contacto con el diseño y desarrollo de un sistema de BI, el cual me ha servido para adquirir una serie de conocimientos que van desde los tipos de arquitectura de un *data warehouse*, pasando por su diseño y creación, hasta la elección y configuración de un sistema de informes basado en OLAP.

Abstract (in English, 250 words or less):

Nowadays, organizations deal with large volumes of information that, only a couple of years ago, were unthinkable. We only need to have a look at how companies such as Google or Amazon work with the information they own. These companies have managed to give value to their customers at the same time they have managed to increase their own sales and profits.

By contrast, there are companies that, with similar or smaller amounts of information, do not have the knowledge to analyze it. Business Intelligence arises then as the needed set of tools to transform this information.

The purpose of this project is to study and, at the same time, document the whole process required to analyze data provided by a certain company. The analysis of the given data should be aimed at answering a set of questions defined by the company.

This process requires:

- Analytical processing
- Data cleansing
- Specific database transformation for the ETL process
- Specific database transformation for the data warehouse
- The information obtained will be contained in an OLAP cube which will be used to generate the requested reports from the analyzed data.

This project has allowed me to learn how to design, develop and implement a BI system. The whole process has helped me to acquire a set of tools ranging from different data warehouse architectures, their design and generation, to the selection and establishment of a report system through OLAP.

Palabras clave (entre 4 y 8):

Business Intelligence, ETL, cubo OLAP, inteligencia de negocio, data warehouse

A Ana, por su paciencia y dedicación.

Índice

Lista de ilustraciones

Lista de tablas

1. Introducción

1.1. Contexto y justificación del Trabajo

La empresa TOTSALES tiene actualmente cinco delegaciones repartidas por el territorio español y necesita hacer un estudio de la información almacenada en su sistema con el fin de analizar su negocio. Este sistema está formado por diferentes módulos integrados de compras, producción, pedidos, ventas, logística, inventarios, contabilidad, nóminas, etc.

Actualmente TOTSALES no dispone de un software que le permita analizar estos datos, por lo que es necesario la creación de un sistema de *Business Intelligence* (BI) que recuperará los datos de diferentes fuentes para su posterior análisis. El módulo que se analizará será el de las ventas de la compañía.

Descripción del proyecto

Se propone la creación de una arquitectura similar a la *Corporate Information Factory* para dar solución a las necesidades propuestas por la empresa TOTSALES.

La arquitectura que se ha elegido está formada por un conjunto de componentes que darán respuestas a las necesidades del cliente mediante la tecnología *Business Intelligence*, en la que se encontrarán: un *data warehouse, procesos* ETL (*Extract, Transform and Load*) y las herramientas de análisis necesarias.

Componentes de la *Corporate Information Factory*

Fuentes de información
Existen diferentes fuentes de información en formato csv. Estos ficheros se han exportado siguiendo una estructura determinada con el fin de ser analizada para la obtención de datos y su explotación.

Herramientas de integración
Conocidos como ETL (*Extract, Transform and Load*), será el encargado de extraer la información de las fuentes, transformarlas usando métodos de *data cleansing* para mejorar la calidad de la información a tratar para su posterior carga en el *data warehouse.*

Data warehouse
El análisis de las fuentes de datos determinará el diseño del *data warehouse* que albergará los datos necesarios para dar respuesta al negocio.

En la arquitectura seguida, aparecen elementos como:

1. *Staging Area***:** es el Sistema que permanece entre las fuentes de datos y el *data warehouse* con el objetivo de facilitar la extracción de datos desde fuentes de origen diferentes, mejorar la calidad de los datos, usado como caché de datos operacionales o incluso para acceder en detalle a información no contenida en el *data warehouse.*

2. *Data warehouse Corporativo:* el cuál contiene información de todas las fuentes de información de los distintos sistemas del cliente. Este tipo de *data warehouse* no se suele utilizar para consultas al ser un volumen de información muy grande.

3. *Operational Data Store* (ODS): contendría los datos con la información actual de las fuentes de información. Se utiliza para la realización de consultas que requieren información actualizada. No contiene un histórico y su volumen de datos es menor.

4. *Data Mart:* utilizado como subconjunto de datos del almacén corporativo, con la información imprescindible de un área determinada del negocio, que finalmente serían consultados por los usuarios a través de herramientas de *reporting*.

Habiendo tenido en consideración los elementos que aparecen en esta arquitectura, se desestima la utilización tanto de ODS como de un *staging area* por la simplicidad del modelo.

Por otro lado se deberá aclarar con la empresa, la utilización de un *data warehouse* para albergar la información de las diferentes fuentes. En este sentido se podría entender que esta información al ser sólo sobre las ventas, se podría tratar de un *data mart*, pero se ha decidido la realización de un *data warehouse* por la escalabilidad que permite.

Cubos OLAP

OLAP (*Online Analytical Processing)* es una solución de Business Intelligence que permite al usuario extraer y ver información de manera sencilla desde diferentes puntos de vista. Las herramientas OLAP estructuran los datos jerárquicamente y permiten a la dirección modificar cómo se ve la información, cambiando las relaciones para poder tener una visión más detallada, ayudando a identificar fortalezas o debilidades.

Generación de informes.

Información extraída y formateada de los *data mart* o cubos OLAP, que facilitan su consulta así como su análisis.

1.2. Objetivos del Trabajo

El objetivo de este trabajo es hacer el análisis de ventas de la empresa planteada. A partir de un conjunto de ficheros con datos de ventas que simularán los datos recogidos en el sistema ERP de la empresa.

Será necesario diseñar, implementar, cargar y explotar un *data warehouse* para satisfacer las necesidades analíticas de este sistema.

Listado de los objetivos

1. Análisis de tres plataformas diferentes de BI disponibles en el mercado, así como la elección de una de ellas para la explotación de la información.

2. Diseño e implementación de un *data warehouse.*

3. Carga de los datos facilitados en los ficheros fuentes en el *data warehouse*, a través de procesos ETL.

4. Una vez hayan sido cargados, se deberá explotar la información para poder extraer ciertos indicadores clave con el fin de proporcionar a los responsables de la empresa TOTSALES una visión más analítica de su sistema así como dar respuesta a las preguntas especificadas a continuación:

    - ¿Se ha producido un decremento de las ventas en las delegaciones respecto a las ventas del año anterior?
    - ¿Qué productos y familias de productos son los más vendidos?
    - ¿A qué clientes se les factura más?
    - ¿Cuál es la distribución de las ventas según la zona del cliente?
    - ¿Cuál es la evolución de las ventas en función del tiempo?
    - ¿Qué comisiones deben liquidar a los comerciales?
    - ¿Cuál es el margen anual obtenido?

1.3. Enfoque y método seguido

La empresa TOTSALES no dispone de un software para poder contestar a las preguntas planteadas en el anterior punto, por lo que en este proyecto no se parte de una base que se puede analizar y mejorar, sino que será necesario el desarrollo de un nuevo producto.

Para la realización de este proyecto se utilizará una metodología de producción llamada *Systems Development Life Cycle* (SDLC), en castellano, ciclo de vida de desarrollo de un sistema de información, en donde cada fase del proyecto es una continuación de la fase anterior.

Esta metodología permite avanzar de forma segura cumpliendo los objetivos marcados siempre que se sigan unas fases:

1. **Planificación:** donde se identificarán los objetivos del producto y se establecerán dentro del tiempo disponible los procesos, actividades, herramientas, métodos de trabajo y recursos.

2. **Análisis de requisitos:** documento que recoge las necesidades de la empresa que definirán el modelo lógico de funcionamiento del sistema de análisis.

3. **Diseño conceptual:** a partir del análisis de requisitos y del modelo lógico, se construirá el modelo dimensional que dará soporte a las necesidades de los usuarios.

4. **Implementación:** construcción del *data warehouse.* Creación de los diferentes pasos del ETL, así como la construcción de un cubo OLAP. Seguidamente se realizará la elaboración de los informes y el análisis de los mismos.

1.4. Planificación del Trabajo

1.4.1. Propuesta de hitos

Nombre	Duración (días)	Inicio	Final
Tarea 1 – Plan de trabajo	**13**	**26/02/2015**	**10/03/2015**
Lectura documentación proyecto	5	26/02/2015	02/03/2015
Búsqueda de información	3	03/03/2015	05/03/2015
Elección sistema BI	1	06/03/2015	06/03/2015
Realización planificación	2	07/03/2015	08/03/2015
Redacción de entregables	1	09/03/2015	09/03/2015
Revisión y entrega	1	10/03/2015	10/03/2015
Tarea 2 – Análisis y diseño	**38**	**11/03/2015**	**21/04/2015**
Lectura documentación	4	11/03/2015	14/03/2015
Búsqueda y consulta de información	4	11/03/2015	14/03/2015
Elección de Base de Datos	4	11/03/2015	14/03/2015
Aprendizaje del uso de herramientas	4	15/03/2015	18/03/2015
Análisis de requisitos	7	19/03/2015	25/03/2015
Diseño conceptual del modelo de datos (DW)	7	26/03/2015	01/04/2015
Diseño del modelo lógico de datos (DW)	6	06/04/2015	11/04/2015
Diseño del modelo físico de datos (DW)	6	12/04/2015	17/04/2015
Redacción de entregables	29	19/03/2015	20/04/2015
Revisión y entrega	1	21/04/2015	21/04/2015
Tarea 3 - Implementación	**24**	**22/04/2015**	**15/05/2015**
Lectura documentación	3	22/04/2015	24/04/2015
Búsqueda y consulta de información	3	25/04/2015	27/04/2015
Instalación de ETL y herramientas de *reporting*	3	28/04/2015	30/04/2015
Aprendizaje de las herramientas instaladas	4	01/05/2015	04/05/2015
Diseño de la extracción y transformación de los datos	5	05/05/2015	09/05/2015

Pruebas	5	05/05/2015	09/05/2015
Construcción de informes requeridos	5	10/05/2015	14/05/2015
Prueba de informes	5	10/05/2015	14/05/2015
Redacción de entregables	10	05/05/2015	14/05/2015
Revisión y entrega	1	15/05/2015	15/05/2015
Tarea 4 – Entrega final y defensa	**32**	**16/05/2015**	**16/06/2015**
Instalación de SW necesario para la presentación	2	16/05/2015	17/05/2015
Realización de Memoria	10	18/05/2015	27/05/2015
Revisión de ortografía y formatos	2	28/05/2015	29/05/2015
Realización de la presentación	10	30/05/2015	08/06/2015
Revisión del proyecto	7	09/06/2015	15/06/2015
Revisión y entrega	1	16/06/2015	16/06/2015
Debate virtual	**3**	**29/06/2015**	**01/07/2015**

Tabla 1 - Cronograma de actividades

1.4.2. Diagrama de Gantt

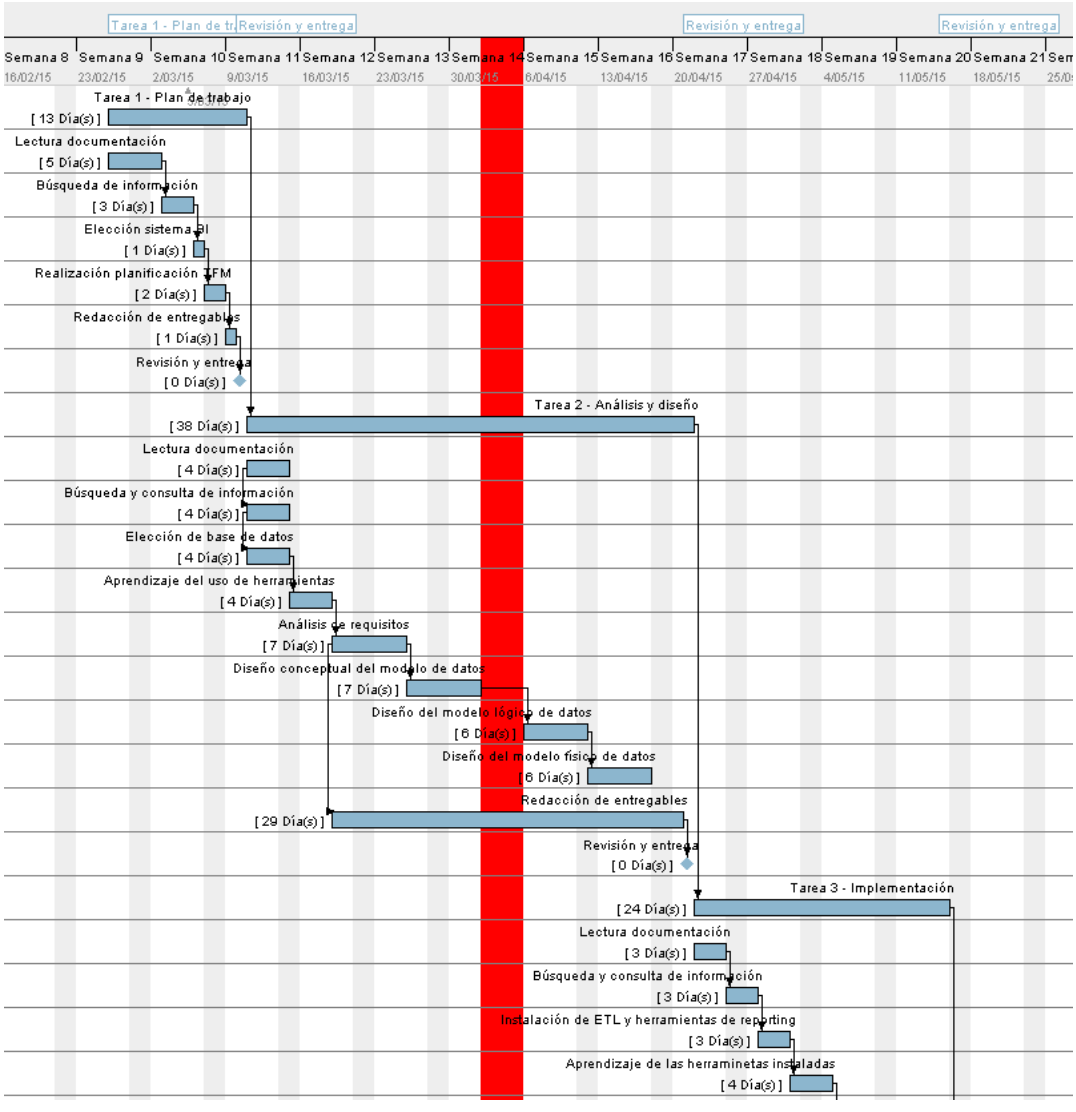

Ilustración 1 - Diagrama de Gantt (parte 1)

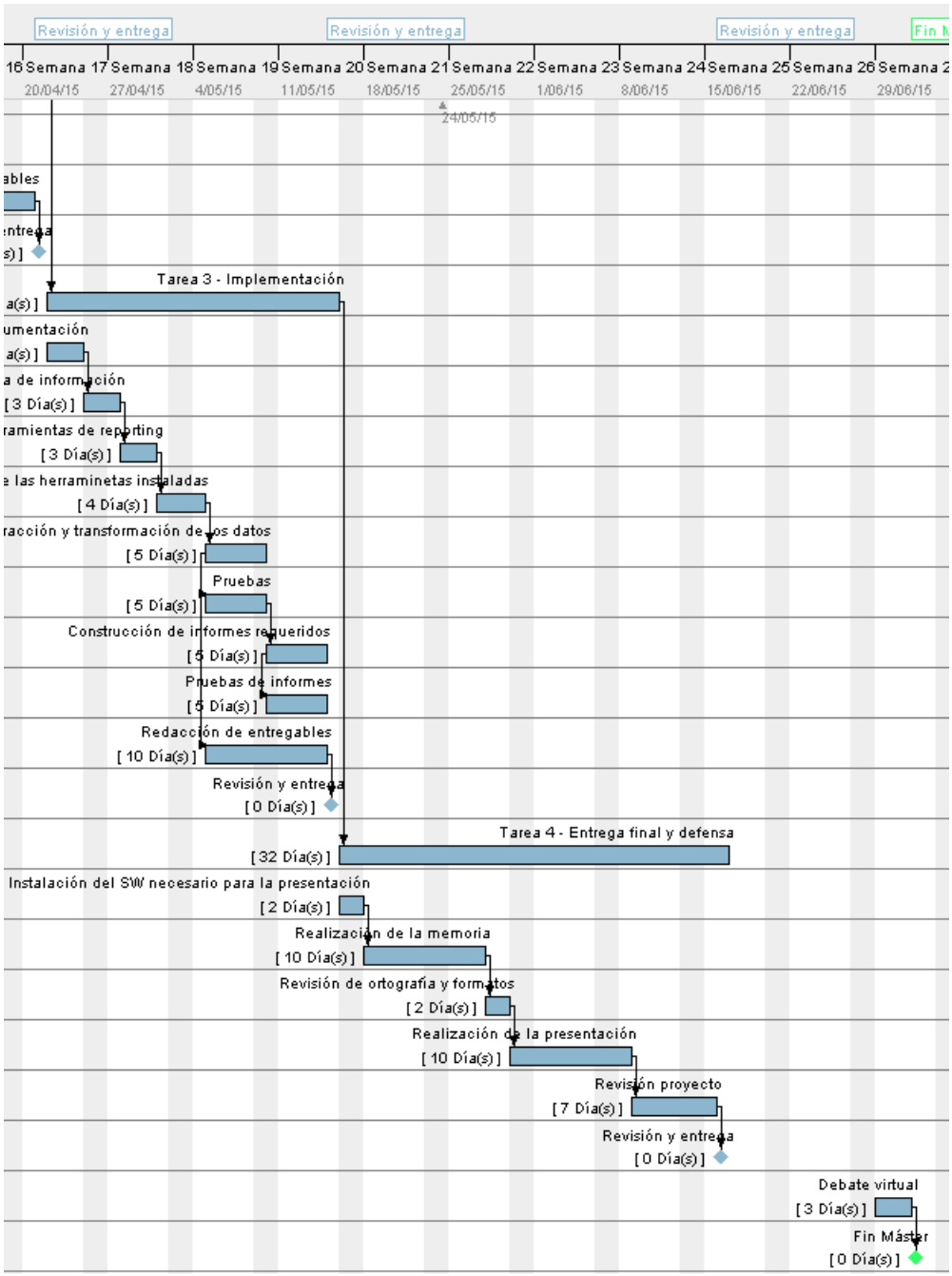

Ilustración 2 – Diagrama de Gantt (parte 2)

1.5. Breve sumario de productos obtenidos

Los productos que se obtendrán en el desarrollo del proyecto serán los siguientes:

Data warehouse: scripts para la creación del *schema* así como los scripts de creación de las tablas de dimensiones y hechos (en el mismo script). También se podrá encontrar el modelo de datos que se ha seguido utilizando la herramienta de MySQL Workbench.

Repositorio ETL: exportación del repositorio ETL de Pentaho (Kettle).

Cubo OLAP: fichero con el esquema de Mondrian en XML realizado con el Pentaho Schema Workbench.

Informes: creados con Pentaho Business Analytics, utilizando el plug-in de Saiku.

Memoria: documento que describe el proceso seguido en el desarrollo del proyecto.

Presentación virtual: vídeo con la presentación del proyecto.

Estos productos, salvo el vídeo, se encuentran en la carpeta "\AránegaHernándezAbel_TFM_2015\Productos\" del fichero comprimido entregado.

1.6. Breve descripción de los otros capítulos

Los otros capítulos de la memoria son:

Análisis y elección de las herramientas utilizadas: en este apartado se analizarán diferentes herramientas y se seleccionará aquella que de una mejor solución al proyecto garantizando al mismo tiempo el mínimo gasto.

Diseño técnico: en este apartado se mostrará la arquitectura seguida para la realización del proyecto, así como el diseño de la base de datos y de los diferentes informes.

Construcción: en este apartado se detallarán los diferentes pasos que se han seguido en el desarrollo y puesta a punto de las diferentes herramientas usadas.

Conclusiones: en donde se intentará contestar a los siguientes puntos:

- Una reflexión crítica sobre el logro de los objetivos planteados inicialmente.
- Un análisis crítico del seguimiento de la planificación y metodología a lo largo del producto.

- ❖ Las líneas de trabajo futuras que no se han podido explorar durante este proyecto y han quedado pendientes.

2. Análisis y elección de las herramientas utilizadas

2.1. Elección de las herramientas utilizadas

2.1.1. Análisis de las bases de datos a usar

Al realizar un proyecto basado en herramientas open source, se ha pensado en dos de los gestores de base de datos que cumplan el mismo requisito como son: MySQL y PostgreSQL dos de las más reconocidas en el desarrollo open source. Hay otras opciones como las versiones Express de Oracle y SQL que se han desestimado por no estar preparados para el *data warehousing*, OLAP o Analysis Services de manera gratuita.

SGDB	MySQL	PostgreSQL
Origen	Sistema de gestión de base de datos relacional y multiusuario. Desde enero de 2008 pertenece a Sun Microsystems, que, a su vez, pertenece a Oracle Corp. desde abril de 2009 desarrolla MySQL como software libre en un esquema de licenciamiento dual.	Sistema de gestión de base de datos relacionales orientado a objetos, con código fuente disponible de forma libre. Creación en 1996 utilizando un modelo cliente/servidor, que hace uso del multiproceso para garantizar la estabilidad del sistema.
Características	• Optimizado para equipos de múltiples procesadores. • Se puede utilizar como cliente/servidor o incrustado en aplicaciones. • Soporta múltiples métodos de almacenamiento de las tablas con prestaciones y diferentes rendimientos para poder optimizar el SGDB a cada caso concreto.	• Texto de largo ilimitado. • Figuras geométricas (con una variedad de funciones asociadas). • Integridad referencial. • Replicación asíncrona/síncrona. • Múltiples métodos de autenticación.
Ventajas	• Soporte para el control de las transacciones. • Replicación de bases de datos. • Bajo costo en	• Multiplataforma. • Herramientas gráficas de diseño. • Bloqueo a nivel de registro. • Integridad referencial. • Ahorros considerables en costes

	requerimientos para la elaboración de bases de datos. • Licencia GLP.	de operación. • Estabilidad y confiabilidad. • Código abierto. • Licencia BSD.
Manejo de concurrencia y bloqueos	• Concurrencia: MVCC. • Bloqueos a nivel de fila: InnoDB.	• Concurrencia: MVCC. • Bloqueos: tanto a nivel de tabla como de fila.
Permite el procesamiento distribuido de consultas	Sí	Sí
Transacciones distribuidas	• Soporte OLAP. • Soporte OLTP. • Data Warehousing. • Data Mining. • Clustering.	• Soporte OLAP Básico. • Soporte OLTP. • Data Warehousing. • Data Mining.
Seguridad	Utiliza lista de control de acceso (ACL) en todas las conexiones, consultas y operaciones.	Modelo de seguridad para acceso a objetos de base de datos por usuarios y grupos de usuarios.
Licencia	Licencia GNU/GPL.	Licencia BSD.
Coste	0 €	0 €

Tabla 2 - Comparativa MySQL vs PostgreSQL

Elección de la base de datos:

Después del análisis de ambos gestores, de los conocimientos previos tenidos y teniendo en cuenta el modelo y la carga de datos elegidos, el gestor MySQL es el que mejor se adapta a este proyecto.

Si por el contrario, la empresa TOTSALES quisiera tener un sistema de business intelligence propio, quizás sería interesante pensar en PostgreSQL por su escalabilidad.

2.1.2. Análisis de plataformas BI

Para esta comparativa se hará uso de las tecnologías open source por la gran relevancia que están adquiriendo a día de hoy.

Teniendo en cuenta estas premisas, y después de haber investigado sobre las plataformas BI open source más utilizadas en la actualidad, se revisarán tres de las más importantes:

- Pentaho
- Jasper
- BIRT

Pentaho

Líder del mercado open source, ofrece una suite en donde se puede encontrar todo lo necesario para la implementación del BI en la empresa:

- Herramienta ETL con Pentaho Data Integration o también conocido por Kettle.
- Motor OLAP basado en Mondrian.
- Sistemas de análisis predictivo de los datos con el uso de algoritmos de *data mining* a través de Weka (plataforma escrita en JAVA para el aprendizaje y la minería de datos).
- Herramienta de *reporting para poder realizar cuadros de mandos y vistas OLAPs, como puede ser el propio Business Analytics o el Reporting Services de Pentaho,* con cuadros de mando y vistas OLAP, entre otras posibilidades.

Pentaho cuenta con una versión *community* en la que podemos encontrar de forma gratuita las herramientas comentadas, además del soporte de una gran comunidad de usuarios.

Jasper

Construido basándose en el motor de informes Jasper Reports cuenta con:

- Una solución ETL de Talend.
- Un motor OLAP basado en Mondrian.
- Integración con el lenguaje R, que resulta bastante interesante para poder hacer minería de datos.
- Para cuadros de mandos y reportes se necesitará la versión *Enterprise*.

BIRT

Su comercialización se hace bajo el nombre de Actuate, pero a nivel de open source es conocido como BIRT y funciona como plug-in de Eclipse, el cual sólo permite la realización de informes y cuadros de mandos que destacan por su nivel de granularidad. En su versión *Enterprise*, incluye varias funcionalidades como son el análisis OLAP, ETL y minería de datos, entre otras.

Servicios ofrecidos por cada plataforma

La mejor manera de observar la diferencia entre los servicios ofrecidos por una u otra plataforma será categorizándolos en la siguiente tabla:

Servicios	Pentaho	Jasper	BIRT
Informes	✓	✓	✓
Servidor Web	✓	✓	Enterprise

Cuadros de mando	✓	Enterprise	✓
ETL	✓	✓	Enterprise
Análisis OLAP	✓	✓	Enterprise
Minería de datos	✓	✓	Enterprise
Dispositivos móviles	Enterprise	✓	Enterprise
Cloud	Enterprise	Enterprise	Enterprise

Tabla 3 - Comparativa de servicios ofrecidos por plataformas BI

Elección de la plataforma:

En los puntos anteriores hemos hablado de tres de las plataformas open source más utilizadas hasta el momento. Se ha elegido la suite de Pentaho por su madurez en el mercado y por su versión *community* ya que sin necesidad de asumir los costes de la versión Enterprise, hay una comunidad de desarrollo detrás bastante activa.

Cabe también destacar que en mi entorno laboral utilizamos Pentaho como herramienta ETL (Kettle), por lo que también existe un grado de conocimiento previo y familiaridad que facilitará su uso.

2.2. Análisis de requisitos

El análisis que se ha realizado a partir de los requisitos dados permite utilizar como hecho principal las **ventas**, que podrán medirse mediante las dimensiones e indicadores que se ven a continuación:

Hecho principal:

- Ventas

Este hecho podrá medirse mediante los siguientes indicadores:

- Evolución de las ventas en unidades vendidas/importe.
- Productos/familias más/menos vendidos por delegación y año.
- Clientes a los que más se les factura en el año.
- Las zonas en donde más/menos se vende.
- Comisiones de los comerciales.
- Evaluación de las ventas en el tiempo, por importe.
- Margen anual para las familias de los productos.

Y a través de las siguientes dimensiones:

- Fecha
- Zona
- Comercial
- Cliente
- Tipo de cliente

- Artículo
- Almacén
- Delegación
- Concepto
- Tipo de dato
- Forma de pago

A continuación se podrá encontrar la tabla que muestra la relación de como, tanto las dimensiones, como los indicadores citados anteriormente sirven para construir los informes necesarios.

Informe	Dimensiones	Indicadores
Evolución de las ventas en cada una de sus delegaciones, tanto en unidades vendidas como en importe.	Fecha Delegación TipoDato	Evolución de las ventas en unidades vendidas/importe.
Detectar la familia de productos y los productos con más y menos éxito de ventas.	Fecha Artículo TipoDato	Productos/familias más/menos vendidos por delegación y año.
Clientes a los que más se les factura en el año.	Fecha Cliente TipoDato	Clientes a los que más se les factura.
Las zonas donde se producen más o menos ventas según la zona del cliente.	Fecha Cliente Zona TipoDato	Las zonas en donde más/menos se vende.
Evolución de las ventas en función del tiempo.	Fecha TipoDato	Evaluación de las ventas en el tiempo, por importe.
Comisiones que deben recibir los comerciales por su trabajo.	Fecha Comercial TipoDato	Comisiones de las ventas realizadas por los comerciales.
Margen anual para las familias de productos.	Fecha Artículo TipoDato	Margen anual obtenido para la familia de los productos.

Tabla 4 – Tabla relacionando tanto las dimensiones como los indicadores con los informes a generar

2.3. Análisis de las fuentes de datos

La información dada consiste en trece ficheros en formato csv (separado por puntos y comas) que se tendrán que tratar previamente para poder trabajar con ellos. Se podría pensar en hacer uso de una *staging area*, pero teniendo en cuenta el alcance del proyecto se entiende que no es necesario, al realizar un *data cleansing* desde los procesos ETL. En el apartado de diseño técnico se podrá encontrar más información al respecto.

A continuación se muestra un breve resumen de la estructura que tienen los ficheros utilizados como fuente de información con los datos necesarios para establecer el sistema que se requiere.

Almacenes.csv: contiene la información de los almacenes que tiene la empresa, con su código de almacén y su descripción.

Articulos.csv: contiene los datos de los diferentes artículos, con su referencia, descripción y su sub-familia.

Clientes.csv: contiene la información de un maestro de cliente, donde aparece NIF/CIF, el nombre completo del cliente así como el tipo de cliente al que se hace referencia (pequeño comercio, hoteles, servicios, industria, distribuidores, mayoristas, …).

ClientesZona.csv: contiene la información de los clientes con su NIF/CIF, nombre completo y población. También se podrá encontrar la provincia y el país al que pertenece.

Comerciales.csv: contiene el NIF y nombre completo de los diferentes comerciales.

Conceptos.csv: contiene el código del concepto de facturación de servicios adicionales en la venta y su descripción.

Divisa.csv: contiene el código de la divisa utilizada por el sistema o del sistema local y su descripción.

Empresas.csv: códigos de las empresas así como su descripción.

FamiliasArticulos.csv: contiene el código de referencia de los productos, su descripción y su familia.

FormasPago.csv: contiene los códigos de las diferentes formas de pago así como su descripción.

TipoDato.csv: contiene el código del tipo de dato y su descripción, teniendo en cuenta si es un dato de presupuesto, una previsión o si es un dato real.

TipoLinea.csv: contiene el código de la línea y su descripción, hay dos tipos, las líneas y el concepto.

Ventas.csv: contiene toda la información de las ventas durante el período comentado en el enunciado del proyecto.

2.4. Modelo conceptual

Con este modelo se identificará qué tipo de procesos y vistas de negocio proporcionan las respuestas adecuadas a las preguntas citadas en el apartado 1.1. Será necesario pensar en las necesidades de la empresa a largo plazo.

Este modelo conceptual está compuesto por la tabla de hechos (el proceso de negocio) y las dimensiones de análisis.

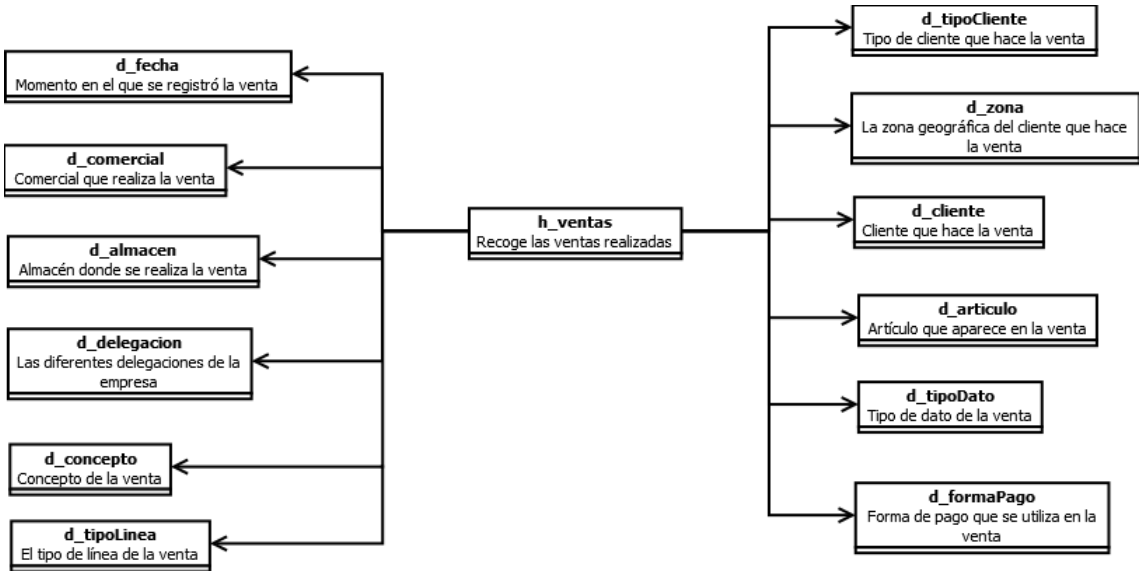

Ilustración 3 – Modelo conceptual

Como se puede observar en la ilustración, se ha decidido por un tipo de esquema en estrella, donde los datos están desnormalizados. En este modelo las dimensiones están directamente relacionadas con la tabla de hechos y la jerarquía no está implementada a través de integridad referencial entre las dimensiones.

Con este esquema se busca un mejor rendimiento al tener menos uniones entre las dimensiones que las que tendría un esquema tipo copo de nieve.

Por último, comentar que el esquema en estrella ofrece una rápida respuesta y además es la fuente ideal para las estructuras de cubo que se pretenden utilizar en este proyecto.

2.5. Modelo lógico

Una vez se ha llegado a realizar el modelo conceptual, se deberá realizar el modelo lógico del sistema en el que se identificarán las métricas de las tablas de hechos y los atributos de las dimensiones.

2.5.1. Hechos

Para el estudio que se llevará a cabo, el hecho relevante de estudio son las **ventas**.

Tabla de hecho	Claves foráneas	Métricas
h_venta	id_fecha, id_zona, id_comercial, id_cliente, id_articulo. id_delegacion, id_almacen, id_tipoCliente, id_formaPago, id_concepto, id_tipoDato, id_tipoLinea	total_unidades, total_coste, total_comisiones, total_importes

Tabla 5 – Tabla de hechos

2.5.2. Dimensiones y sus atributos

Tabla de dimensiones	Clave primaria	Jerarquía (0 … n)	Atributos
d_fecha	id_fecha	0 - anio, 1 - mes, 2 - semana, 3 - día	anio, mes, desc_mes, semana_anio, dia, desc_dia_semana
d_zona	id_zona	0 - zona, 1 - provincia, 2 - población	país, provincia, población
d_comercial	id_comercial	comercial	nif, nombre
d_delegacion	id_delegacion	delegación	cod_delegacion, desc_delegacion
d_almacen	id_almacen	almacén	cod_almacen, desc_almacen
d_cliente	id_cliente	cliente	nif, nombre_completo, zona, tipo
d_articulo	id_articulo	0 - familia, 1 – subFamilia, 2 - articulo	cod_articulo, desc_articulo, familia, desc_familia, sub_familia, desc_sub_familia
d_concepto	id_concepto	concepto	cod_concepto, desc_concepto
d_tipoDato	id_tipoDato	tipoDato	cod_tipoDato, desc_tipoDato
d_tipoCliente	id_tipoCliente	tipoCliente	cod_tipoCliente, desc_tipoCliente
d_tipoLinea	id_tipoLinea	tipoLinea	cod_tipoLinea, desc_tipoLinea
d_formaPago	id_formaPago	formaPago	cod_forma, desc_forma

Tabla 6 – Tabla de dimensiones

2.5.3. Diseño lógico

El diseño lógico estaría representado mediante la siguiente estructura:

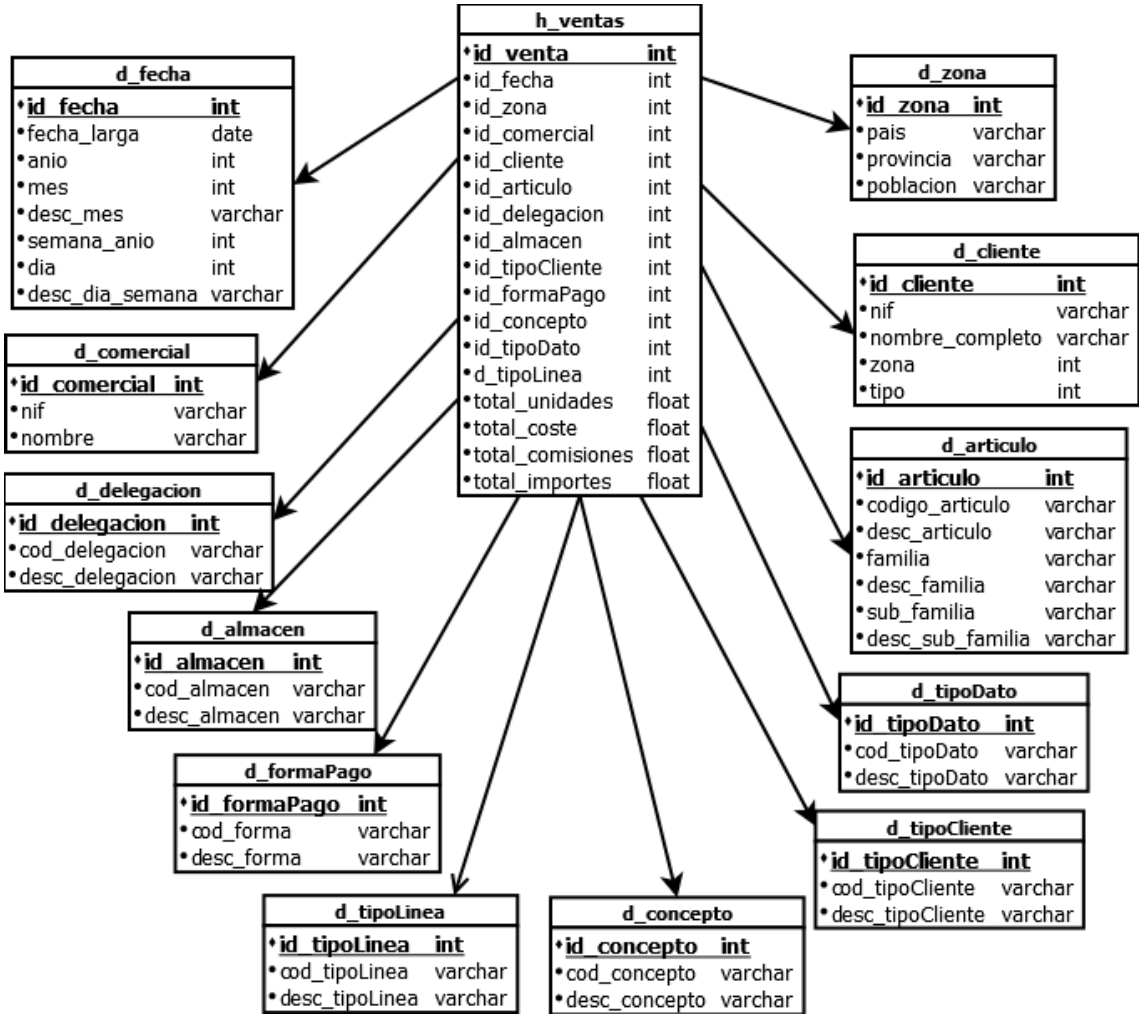

Ilustración 4 - Modelo lógico

3. Diseño técnico

3.1. Modelo físico

En un primer momento se había pensado utilizar una *staging area* donde cargar los datos de las fuentes de datos, pero se ha desestimado por la sencillez del modelo. En su lugar se realizarán tareas de *data cleansing*, que corregirán y eliminarán datos erróneos en los ficheros fuentes durante los procesos ETL.

3.1.1. Data warehouse

La definición de las tablas del modelo conceptual es la siguiente:

h_venta:

Nombre del campo	Tipo de dato
id_venta	Integer
id_fecha	Integer
id_zona	Integer
id_comercial	Integer
id_cliente	Integer
id_articulo	Integer
id_delegacion	Integer
id_almacen	Integer
id_tipoCliente	Integer
id_formaPago	Integer
id_concepto	Integer
id_tipoDato	Integer
id_tipoLinea	Integer
total_unidades	Float
total_coste	Float
total_comisiones	Float
total_importes	Float

Tabla 7 – Definición de la tabla de hechos: venta

d_articulo:

Nombre del campo	Tipo de dato
id_articulo	Integer
cod_articulo	Varchar(15)
desc_articulo	Varchar(100)
Familia	Varchar(30)
desc_familia	Varchar(45)
sub_familia	Varchar(10)
desc_sub_familia	Varchar(45)

Tabla 8 – Definición de la tabla de la dimensión: artículo

d_cliente:

Nombre del campo	Tipo de dato
id_cliente	Integer
nif	Varchar(9)
nombre_completo	Varchar(45)
Zona	Integer
Tipo	Integer

Tabla 9 – Definición de la tabla de la dimensión: cliente

d_fecha:

Nombre del campo	Tipo de dato
id_fecha	Integer
fecha_larga	Varchar(10)
anio	Integer
mes	Integer
desc_mes	Varchar(20)
semana_anio	Integer
dia	Integer
desc_dia_semana	Varchar(20)

Tabla 10 – Definición de la tabla de la dimensión: fecha

d_comercial:

Nombre del campo	Tipo de dato
id_comercial	Integer
Nif	Varchar(9)
nombre	Varchar(45)

Tabla 11 – Definición de la tabla de la dimensión: comercial

d_zona:

Nombre del campo	Tipo de dato
id_zona	Integer
pais	Varchar(20)
provincia	Varchar(30)
poblacion	Varchar(45)

Tabla 12 – Definición de la tabla de la dimensión: zona

d_almacen:

Nombre del campo	Tipo de dato
id_almacen	Integer
cod_almacen	Varchar(10)
desc_almacen	Varchar(45)

Tabla 13 – Definición de la tabla de la dimensión: almacén

d_delegacion:

Nombre del campo	Tipo de dato
id_delegacion	Integer
cod_delegacion	Varchar(10)
desc_delegacion	Varchar(45)

Tabla 14 – Definición de la tabla de la dimensión: delegación

d_tipoCliente:

Nombre del campo	Tipo de dato
id_tipoCliente	Integer
cod_tipoCliente	Varchar(20)
desc_tipoCliente	Varchar(45)

Tabla 15 – Definición de la tabla de la dimensión: tipoCliente

d_formaPago:

Nombre del campo	Tipo de dato
id_forma	Integer
cod_forma	Varchar(10)
desc_forma	Varchar(45)

Tabla 16 – Definición de la tabla de la dimensión: formaPago

d_concepto:

Nombre del campo	Tipo de dato
id_concepto	Integer
cod_concepto	Varchar(10)
desc_concepto	Varchar(45)

Tabla 17 – Definición de la tabla de la dimensión: concepto

d_tipoDato:

Nombre del campo	Tipo de dato
id_tipoDato	Integer
cod_tipoDato	Varchar(10)
desc_tipoDato	Varchar(45)

Tabla 18 – Definición de la tabla de la dimensión: tipoDato

d_tipoLinea:

Nombre del campo	Tipo de dato
id_tipoLinea	Integer
cod_tipoLinea	Varchar(10)
desc_tipoLinea	Varchar(20)

Tabla 19 - Definición de la tabla de la dimensión: tipoLinea

En el siguiente modelo se observará la implementación del modelo conceptual definido previamente:

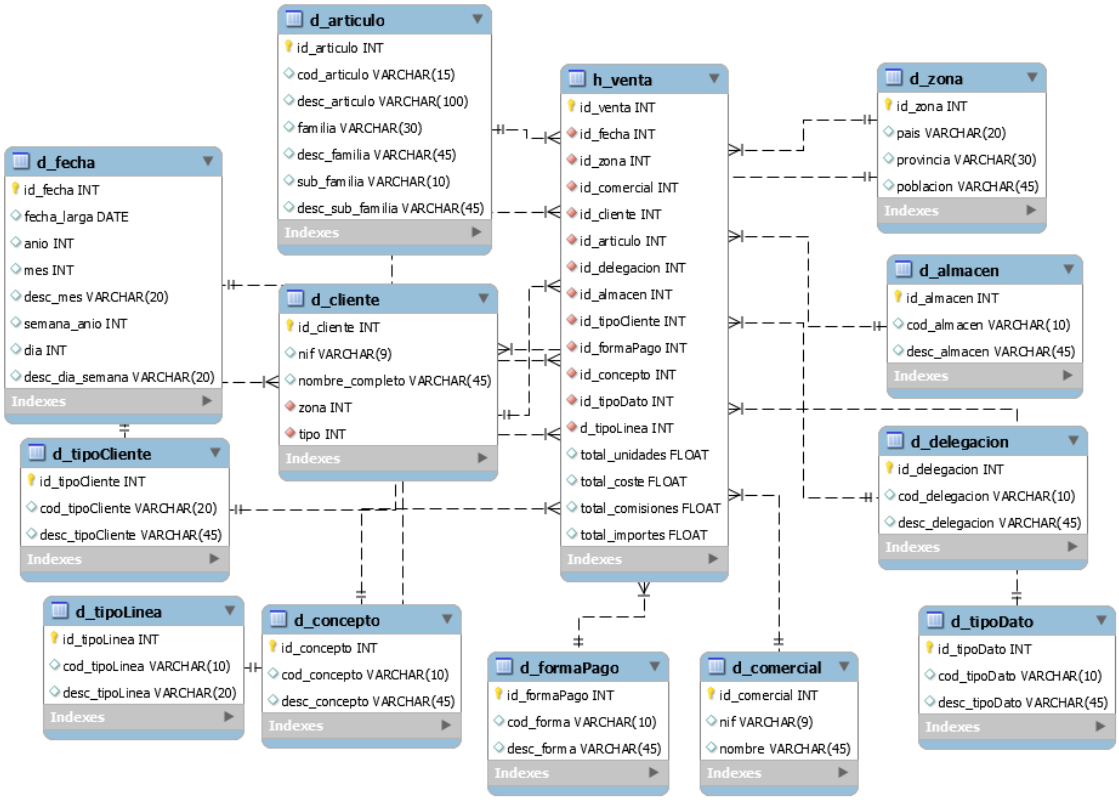

Ilustración 5 - Modelo lógico del *data warehouse*

3.2. Arquitectura

La empresa TOTSALES necesita hacer un estudio de la información almacenada en su sistema para analizar su negocio. La fuente de los datos son varios ficheros CSV que contienen datos exportados de los diferentes sistemas de que dispone la empresa.

Para dicho análisis no se contemplará la necesidad de un *staging area* ni tampoco de un espacio de datos operacional, como se ha comentado en epígrafes anteriores, ya que lo que el cliente demanda es el análisis de los datos obtenidos durante un intervalo de tiempo y de manera puntual.

Es por ello, que la arquitectura utilizada será simple, diseñando una serie de transformaciones que obtendrán los datos de las fuentes y cargará los datos en el *data warehouse.*

Una vez en el *data warehouse* se hará uso del motor Mondrian utilizado por Pentaho para cubos OLAP y para realizar el análisis necesario, que servirá para obtener los informes y cuadros de mandos (*dashboards*) requeridos por la empresa TOTSALES.

Arquitectura utilizada

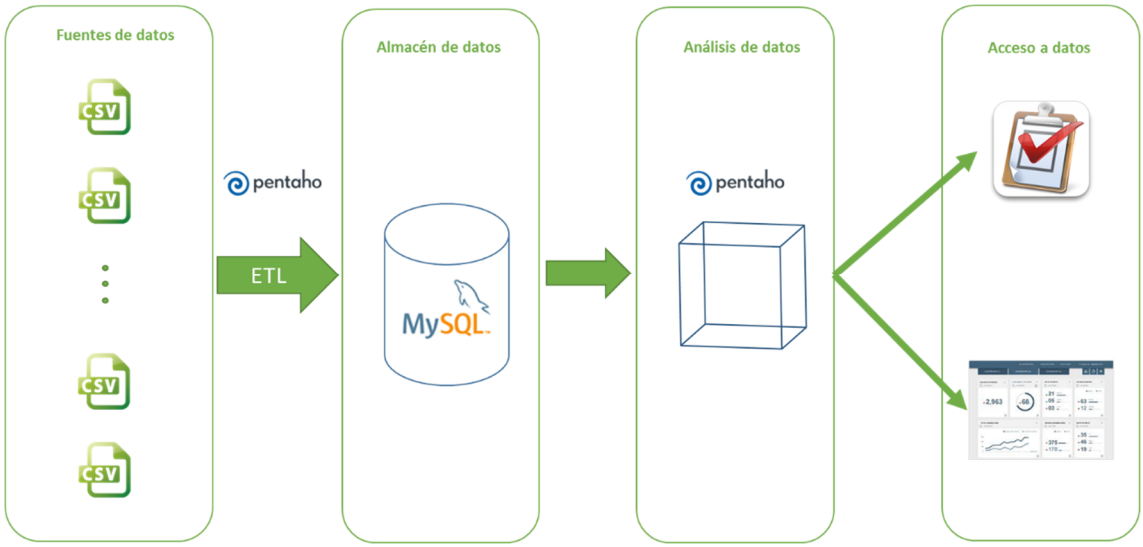

Ilustración 6 - Diseño de la arquitectura seguida

La elección de MySQL, de SGBD y de la solución de Pentaho para las tareas de ETL, así como de análisis, se puede encontrar explicadas en el epígrafe “Análisis de la tecnología a utilizar” que se encuentra en los anexos.

3.3. Carga

El proceso de carga de los datos en el *data warehouse* se realizará utilizando la herramienta de ETL, Kettle. Se hará un *job* principal que irá lanzando las diferentes transformaciones, las cuales irán realizando, tanto la lectura de los ficheros como el *data cleansing* necesario para poder procesar y analizar los datos de manera eficiente.

El *job* irá lanzando cada transformación, que irá rellenando las tablas de dimensiones del *data warehouse*. Una vez rellenas procederá a rellenar las tablas de hechos.

3.4. Cubo OLAP

La creación del cubo OLAP se realiza utilizando la herramienta Mondrian Schema Workbench de Pentaho de una manera visual. El motor Mondrian procesa peticiones MDX al esquema OLAP creado.

El fichero con el esquema obtenido (Ventas.xml) es un modelo de metadatos XML creado con una estructura específica que es la usada por el motor de Mondrian.

Este modelo XML está considerado como un cubo y utiliza la tabla de hechos h_venta, las tablas de dimensiones del *data warehouse* y las métricas

definidas. No requiere de mantenimiento físico del cubo como tal, solamente de los metadatos creados.

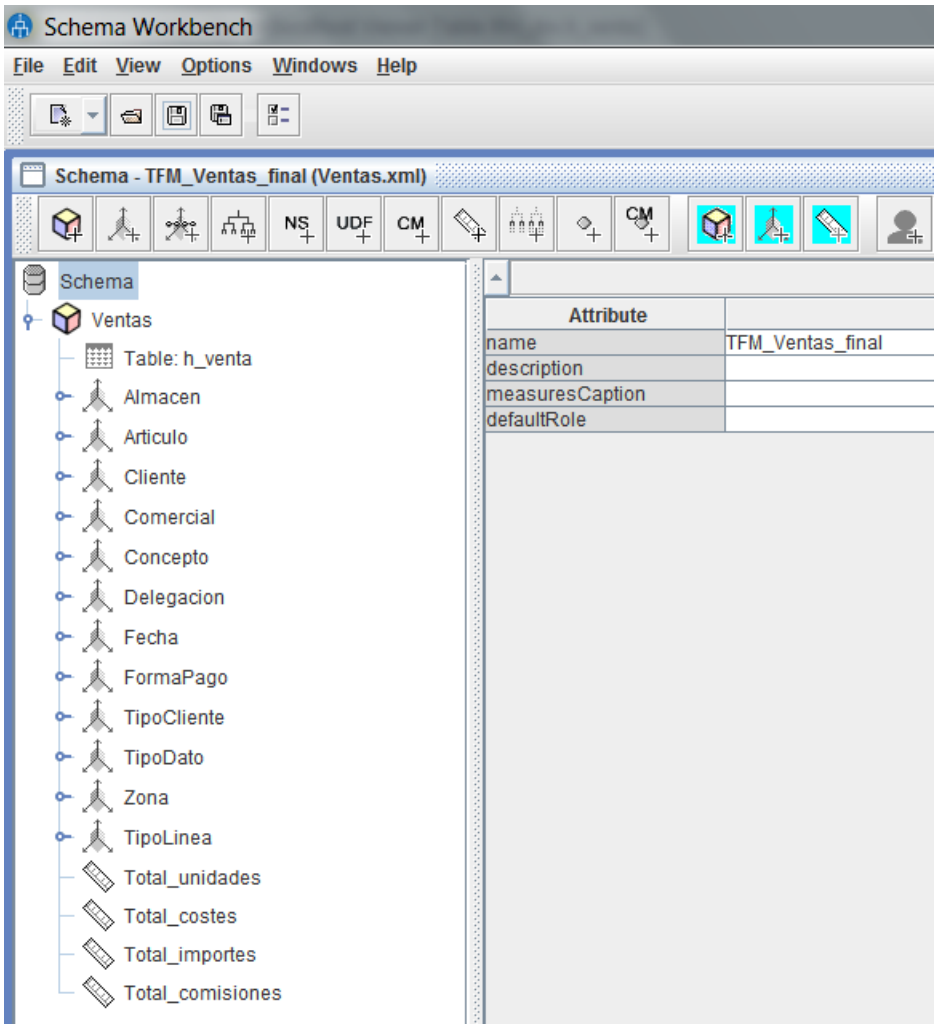

Ilustración 7 - Schema Workbench de Pentaho, creación del esquema del cubo

3.5. Informes

A continuación se muestra el diseño de cada uno de los informes realizados para contestar a las diferentes preguntas de negocio requeridas por la empresa TOTSALES.

Se hace uso de la herramienta de Saiku Analytics para el diseño de los informes. Existe la posibilidad de enviar los informes a una dirección de carpetas de la red de la empresa, o de dejarlos en un portal web para acceder desde la intranet.

1) **Evolución de las ventas en unidades vendidas/importe.**
   Aunque no está diseñado se piensa que la mejor manera de mostrar la evolución es a través de una gráfica de columnas, donde en el eje de abscisas estará el mes del año en curso y en el eje de ordenadas

el importe o las unidades vendidas. El informe tendrá la posibilidad de intercambiar la métrica, ya sea por unidades vendidas o importe

2) **Productos más/menos vendidos por delegación y año.**
   Gráfico de columnas que mostrará en el eje de abscisas los productos, permitiéndose elegir el año de análisis deseado, así como la delegación. En el eje de ordenadas se situará el número de unidades vendidas.

3) **Familias más/menos vendidas por delegación y año.**
   Al igual que en el informe anterior, el gráfico utilizado será de columnas. El eje de abscisas contendrá la descripción de las familias mientras que el de ordenadas, el total de unidades. Se podrá cambiar tanto la delegación como el año, así como la agrupación (más o menos vendidas).

4) **Clientes a los que más se les factura en el año.**
   Gráfico de columnas donde en el eje de abscisas estará el Top 10 de clientes y en el de ordenadas el total de ventas por cada cliente.

5) **Zonas en donde más/menos se vende.**
   Gráfico de columnas que contendrá en el eje de abscisas las zonas donde más/menos se vende y en el de ordenadas el total del importe. Se dará la posibilidad de dejar cambiar la relación (más/ menos venta).

6) **Comisiones de los comerciales.**
   Listado de los comerciales junto con el número de ventas, así como con su importe y el importe correspondiente a sus comisiones.

7) **Evaluación de las ventas en el tiempo, por importe.**
   Gráfico de líneas en cuyo eje de abscisas estarán las marcas temporales y en el de ordenadas el importe. Se verá la evolución según la variación de las líneas.

8) **Margen anual para las familias de los productos.**
   Listado con las diferentes familias vendidas con su importe total anual y el margen obtenido.

4. Construcción

Sistemas elegidos:

Sistema Operativo
El entorno utilizado para realizar el caso que se ha descrito está basado en un sistema operativo Windows 7.

Sistema gestor de base de datos
MySQL 5.6.23, tanto para el *data warehouse* como para el repositorio del ETL.

Data Integration (ETL)
Kettle de Pentaho con repositorio en MySQL 5.6.23.

Cubo OLAP
Pentaho Schema Workbench 3.9.0.0-213, Mondrian.

Plataforma de Business Analytics*:* Pentaho BI Server 5.3.

Sistema de reporting: Saiku Analytics.

A nivel de hardware, se dispone de un disco duro SSD de 256GB y 8GB de memoria RAM. Adecuados para tratar de dar respuesta a las preguntas de la empresa.

4.1. Base de datos

Como se ha comentado en el epígrafe anterior, tanto el esquema que contiene el repositorio de Kettle (esquema: kettle) como el esquema del *data warehouse* (esquema: tfm_dw), se encuentran en MySQL.

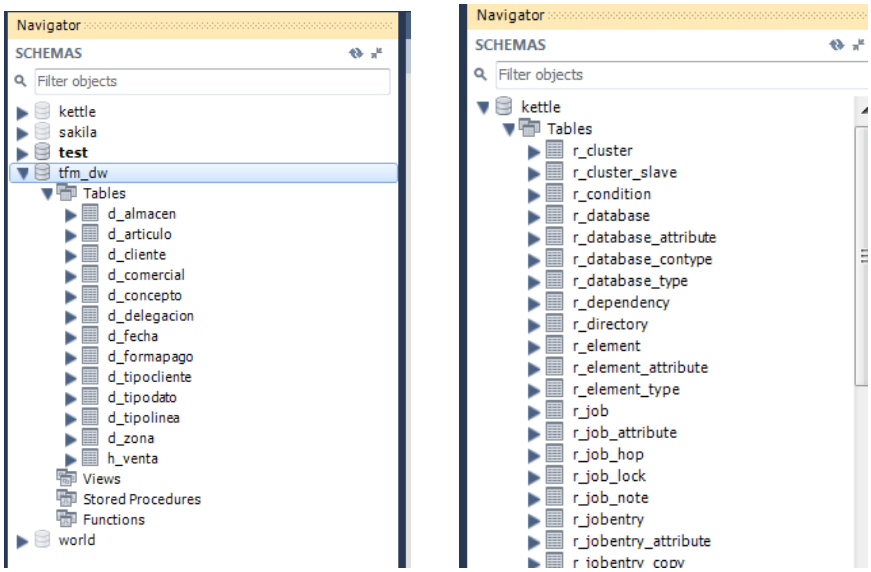

Ilustración 8 - Esquema del *data warehouse* y del repositorio de Kettle

En los siguientes sub-apartados se encontrarán los scripts para la creación tanto de las dimensiones como de la tabla de hechos del *data warehouse.*

4.1.1. Dimensiones

d_almacen

```
CREATE TABLE `d_almacen` (
  `id_almacen` int(11) NOT NULL,
  `cod_almacen` varchar(10) DEFAULT NULL,
  `desc_almacen` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_almacen`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_articulo

```
CREATE TABLE `d_articulo` (
  `id_articulo` int(11) NOT NULL,
  `cod_articulo` varchar(15) DEFAULT NULL,
  `desc_articulo` varchar(100) DEFAULT NULL,
  `familia` varchar(30) DEFAULT NULL,
  `desc_familia` varchar(45) DEFAULT NULL,
  `sub_familia` varchar(10) DEFAULT NULL,
  `desc_sub_familia` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_articulo`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_cliente

```
CREATE TABLE `d_cliente` (
  `id_cliente` int(11) NOT NULL,
  `nif` varchar(9) DEFAULT NULL,
  `nombre_completo` varchar(45) DEFAULT NULL,
  `zona` int(11) NOT NULL,
  `tipo` int(11) NOT NULL,
  PRIMARY KEY (`id_cliente`),
  KEY `fk_zona_idx` (`zona`),
  KEY `fk_tipoCliente_idx` (`tipo`),
  CONSTRAINT `fk_tipoCliente` FOREIGN KEY (`tipo`)
REFERENCES `d_tipocliente` (`id_tipoCliente`) ON
DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `fk_zona` FOREIGN KEY (`zona`) REFERENCES
`d_zona` (`id_zona`) ON DELETE NO ACTION ON UPDATE NO
ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_comercial

```
CREATE TABLE `d_comercial` (
  `id_comercial` int(11) NOT NULL,
  `nif` varchar(9) DEFAULT NULL,
```

```
  `nombre` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_comercial`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_concepto

```
CREATE TABLE `d_concepto` (
  `id_concepto` int(11) NOT NULL,
  `cod_concepto` varchar(10) DEFAULT NULL,
  `desc_concepto` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_concepto`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_delegacion

```
CREATE TABLE `d_delegacion` (
  `id_delegacion` int(11) NOT NULL,
  `cod_delegacion` varchar(10) DEFAULT NULL,
  `desc_delegacion` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_delegacion`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_fecha

```
CREATE TABLE `d_fecha` (
  `id_fecha` int(11) NOT NULL,
  `fecha_larga` varchar(10) DEFAULT NULL,
  `anio` int(11) DEFAULT NULL,
  `mes` int(11) DEFAULT NULL,
  `desc_mes` varchar(20) DEFAULT NULL,
  `semana_anio` int(11) DEFAULT NULL,
  `dia` int(11) DEFAULT NULL,
  `desc_dia_semana` varchar(20) DEFAULT NULL,
  PRIMARY KEY (`id_fecha`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_formaPago

```
CREATE TABLE `d_formapago` (
  `id_formaPago` int(11) NOT NULL,
  `cod_forma` varchar(10) DEFAULT NULL,
  `desc_forma` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_formaPago`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_tipoCliente

```
CREATE TABLE `d_tipocliente` (
  `id_tipoCliente` int(11) NOT NULL,
  `cod_tipoCliente` varchar(20) DEFAULT NULL,
  `desc_tipoCliente` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_tipoCliente`)
```

```
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_tipoDato

```
CREATE TABLE `d_tipodato` (
  `id_tipoDato` int(11) NOT NULL,
  `cod_tipoDato` varchar(10) DEFAULT NULL,
  `desc_tipoDato` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_tipoDato`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_tipoLinea

```
CREATE TABLE `d_tipolinea` (
  `id_tipoLinea` int(11) NOT NULL,
  `cod_tipoLinea` varchar(10) DEFAULT NULL,
  `desc_tipoLinea` varchar(20) DEFAULT NULL,
  PRIMARY KEY (`id_tipoLinea`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

d_zona

```
CREATE TABLE `d_zona` (
  `id_zona` int(11) NOT NULL,
  `pais` varchar(20) DEFAULT NULL,
  `provincia` varchar(30) DEFAULT NULL,
  `poblacion` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id_zona`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

4.1.2. Hechos

h_venta

```
CREATE TABLE `h_venta` (
  `id_venta` int(11) NOT NULL,
  `id_fecha` int(11) NOT NULL,
  `id_zona` int(11) NOT NULL,
  `id_comercial` int(11) NOT NULL,
  `id_cliente` int(11) NOT NULL,
  `id_articulo` int(11) NOT NULL,
  `id_delegacion` int(11) NOT NULL,
  `id_almacen` int(11) NOT NULL,
  `id_tipoCliente` int(11) NOT NULL,
  `id_formaPago` int(11) NOT NULL,
  `id_concepto` int(11) NOT NULL,
  `id_tipoDato` int(11) NOT NULL,
  `id_tipoLinea` int(11) DEFAULT NULL,
  `total_unidades` float DEFAULT NULL,
  `total_coste` float DEFAULT NULL,
```

```
  `total_comisiones` float DEFAULT NULL,
  `total_importes` float DEFAULT NULL,
  PRIMARY KEY (`id_venta`),
  KEY `fk_h_venta_d_articulo_idx` (`id_articulo`),
  KEY `fk_h_venta_d_fecha1_idx` (`id_fecha`),
  KEY `fk_h_venta_d_cliente1_idx` (`id_cliente`),
  KEY `fk_h_venta_d_comercial1_idx` (`id_comercial`),
  KEY `fk_h_venta_d_zona1_idx` (`id_zona`),
  KEY `fk_h_venta_d_almacen1_idx` (`id_almacen`),
  KEY `fk_h_venta_d_delegacion1_idx`
(`id_delegacion`),
  KEY `fk_h_venta_d_tipoCliente1_idx`
(`id_tipoCliente`),
  KEY `fk_h_venta_d_formaPago_idx` (`id_formaPago`),
  KEY `fk_h_venta_d_concepto_idx` (`id_concepto`),
  KEY `fk_h_venta_d_tipoDato_idx` (`id_tipoDato`),
  CONSTRAINT `fk_h_venta_d_almacen` FOREIGN KEY
(`id_almacen`) REFERENCES `d_almacen` (`id_almacen`)
ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `fk_h_venta_d_articulo` FOREIGN KEY
(`id_articulo`) REFERENCES `d_articulo`
(`id_articulo`) ON DELETE NO ACTION ON UPDATE NO
ACTION,
  CONSTRAINT `fk_h_venta_d_cliente` FOREIGN KEY
(`id_cliente`) REFERENCES `d_cliente` (`id_cliente`)
ON DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `fk_h_venta_d_comercial` FOREIGN KEY
(`id_comercial`) REFERENCES `d_comercial`
(`id_comercial`) ON DELETE NO ACTION ON UPDATE NO
ACTION,
  CONSTRAINT `fk_h_venta_d_concepto` FOREIGN KEY
(`id_concepto`) REFERENCES `d_concepto`
(`id_concepto`) ON DELETE NO ACTION ON UPDATE NO
ACTION,
  CONSTRAINT `fk_h_venta_d_delegacion` FOREIGN KEY
(`id_delegacion`) REFERENCES `d_delegacion`
(`id_delegacion`) ON DELETE NO ACTION ON UPDATE NO
ACTION,
  CONSTRAINT `fk_h_venta_d_fecha` FOREIGN KEY
(`id_fecha`) REFERENCES `d_fecha` (`id_fecha`) ON
DELETE NO ACTION ON UPDATE NO ACTION,
  CONSTRAINT `fk_h_venta_d_formaPago` FOREIGN KEY
(`id_formaPago`) REFERENCES `d_formapago`
(`id_formaPago`) ON DELETE NO ACTION ON UPDATE NO
ACTION,
  CONSTRAINT `fk_h_venta_d_tipoCliente` FOREIGN KEY
(`id_tipoCliente`) REFERENCES `d_tipocliente`
(`id_tipoCliente`) ON DELETE NO ACTION ON UPDATE NO
ACTION,
  CONSTRAINT `fk_h_venta_d_tipoDato` FOREIGN KEY
(`id_tipoDato`) REFERENCES `d_tipodato`
```

```
(`id_tipoDato`) ON DELETE NO ACTION ON UPDATE NO
ACTION,
  CONSTRAINT `fk_h_venta_d_zona1` FOREIGN KEY
(`id_zona`) REFERENCES `d_zona` (`id_zona`) ON DELETE
NO ACTION ON UPDATE NO ACTION
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COMMENT='    ';
```

4.2. Proceso de carga

A continuación se verá tanto el *job* inicial como las diferentes transformaciones que irán insertando los datos en las tablas, tanto de dimensiones como de hechos.

JOB principal: J_ETL_C_DW

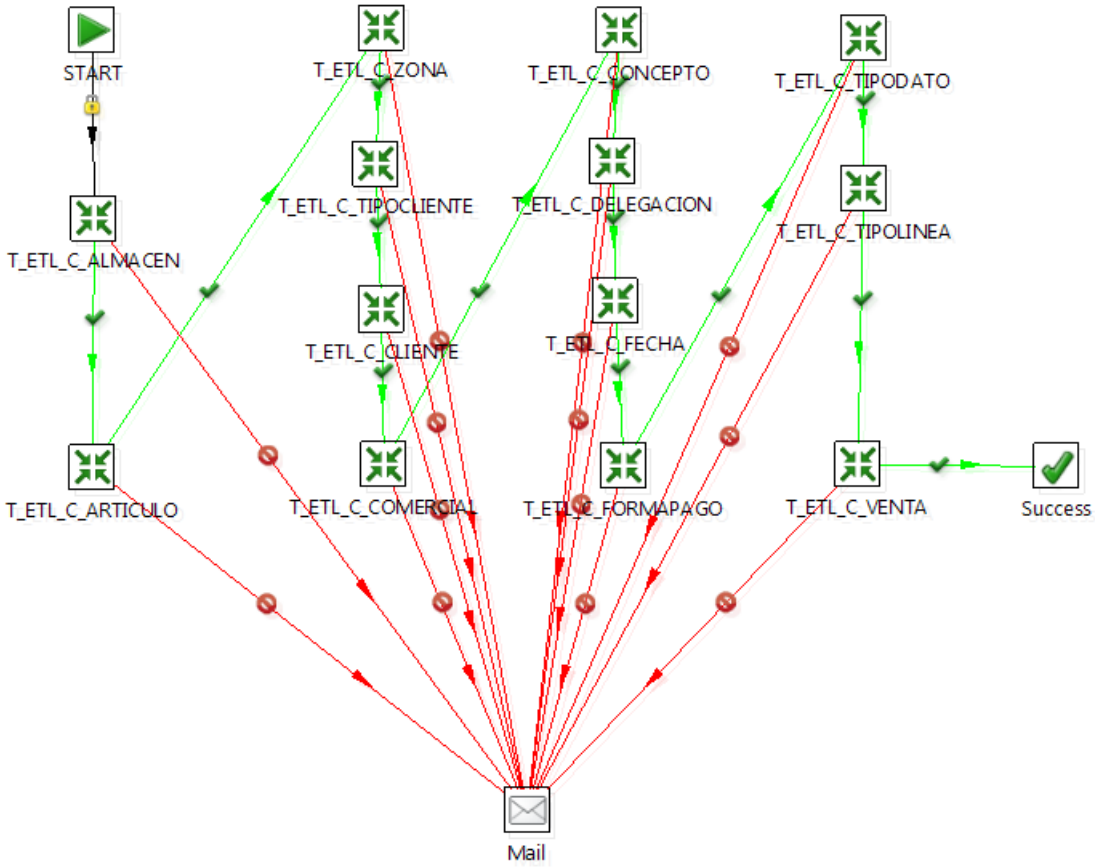

Ilustración 9 - Job principal J_ETL_C_DW

El job se encarga de lanzar las transformaciones y en caso de error manda un correo con el log y el detalle del mismo.

START

Paso que dispara el comienzo del job.

Ilustración 10 - Paso que indica el comienzo del job

T_ETL_C_ALMACEN, T_ETL_C_COMERCIAL, T_ETL_C_CONCEPTO, T_ETL_C_DELEGACION, T_ETL_C_FORMAPAGO, T_ETL_C_TIPODATO, T_ETL_TIPOLINEA.

Son transformaciones simples, en donde se abre el fichero de cada una de las fuentes de datos, y se realizan pasos de ordenación y eliminación de datos duplicados. Una vez tratados, se añade el secuencial para poderlos insertar en la tabla de dimensión correspondiente.

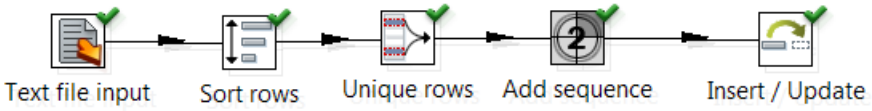

Ilustración 11 - Transformación simple

T_ETL_C_ARTICULO

Transformación más compleja debido a la estructura de los ficheros. La transformación se divide en dos partes. Una consiste en la lectura del fichero FamiliasArticulos.csv el cuál separará los registros que marcan la jerarquía de los registros que realmente interesan dentro del fichero.

La otra parte, lee el fichero de Articulos.csv que contiene el mismo número de códigos de artículos que el de FamiliasArticulos.csv, y al igual que en el primer paso, separa los registros de jerarquías de los artículos.

El paso “Merge Join” realiza una unión de los datos de los dos flujos en donde se obtendrá el código de artículo, su descripción, la sub-familia y familia a la que pertenece. Se limpiará el *stream* de la información innecesaria para aligerar el procesamiento, se añadirá la secuencia y se insertará en la tabla de dimensión correspondiente.

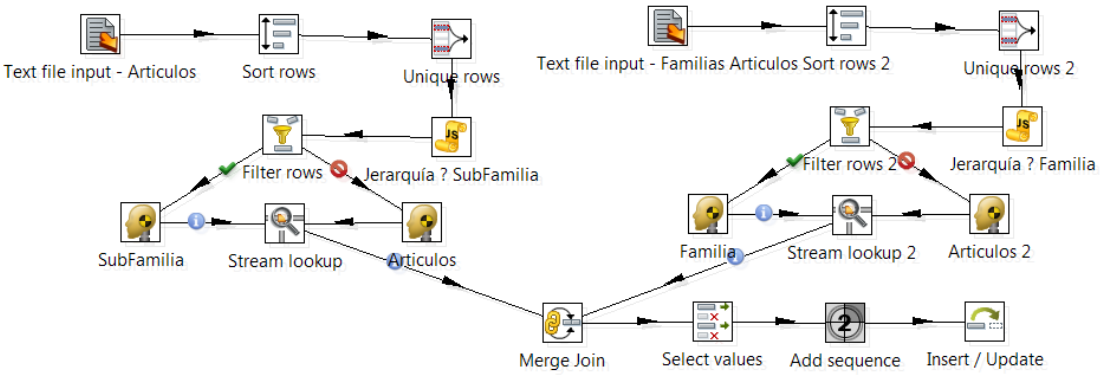

Ilustración 12 - Transformación T_ETL_C_ARTICULO

T_ETL_C_CLIENTE

Transformación que requiere obtener el código de cliente, su nombre, la zona a la que pertenece y el tipo de cliente. Hay que tener en cuenta que esta transformación se realiza después de cargar las dimensiones de zona y de tipo Cliente y no antes.

Esta transformación, al igual que la anterior se divide en dos partes. Por un lado se lee el fichero ClientesZona.csv, que contiene tanto información del cliente (nif, nombre y población), como el listado de poblaciones y zonas. Se hace necesario hacer una distinción que separe las zonas del resto de información utilizando el paso “Zonas del cliente”. A través del paso de “Asignar Zona según BD” se buscará el id correspondiente en la dimensión zona. Por otro lado, se lee el fichero Clientes.csv, en el que hay que separar los clientes, de los tipos de clientes, estos últimos situados al final del mismo. Al igual que en el primer paso se busca el id correspondiente en la dimensión tipoCliente.

La información obtenida en ambos procesos se une, se añade el secuencial y se inserta en la tabla correspondiente.

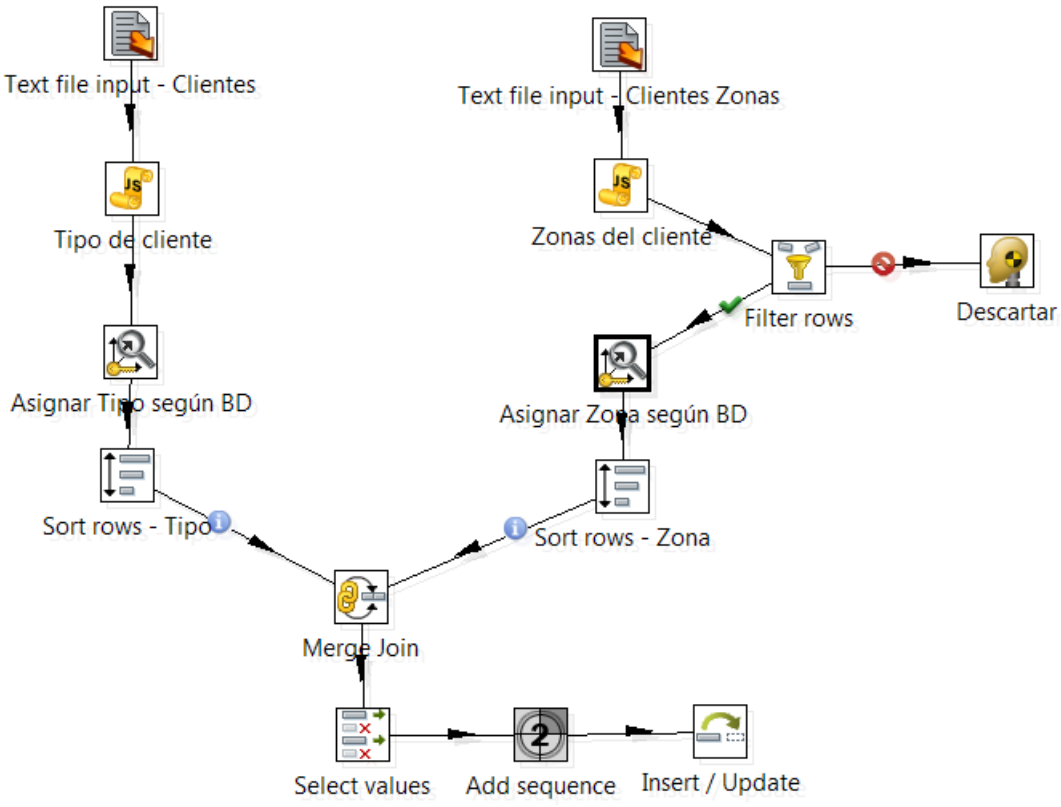

Ilustración 13 - Transformación T_ETL_CLIENTE

T_ETL_C_FECHA

Para la generación de fechas con la que se podrá rellenar la dimensión correspondiente, se inicializa una variable con la fecha 1/1/2009 y se genera un secuencial de los días que se utilizará en el paso “Calculator” para realizar las diferentes operaciones de fechas necesarias para completar la dimensión.

Se termina añadiendo el secuencial y haciendo la inserción en su dimensión

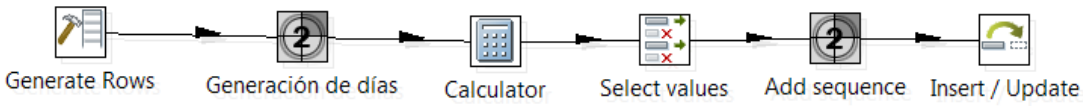

Ilustración 14 - Transformación T_ETL_C_FECHA

Step name: Calculator

s:

New field	Calculation	Field A	Field B	Field C
fechas	Date A + B Days	fechaInicial	seqDia	
dia	Day of month of date A	fechas		
diaSemana	Day of week of date A	fechas		
mes	Month of date A	fechas		
descMes	Month of date A	fechas		
año	Year of date A	fechas		
semana	Week of year of date A	fechas		

Ilustración 15 - Operaciones para calcular las fechas

T_ETL_C_TIPOCLIENTE

En esta transformación se lee el fichero y se evalúa si es un registro con información sobre el cliente o si no lo es. Una vez procesado, se separa la información que se necesita y el resto se descarta, se añade una secuencia y se inserta en la tabla de dimensiones correspondiente.

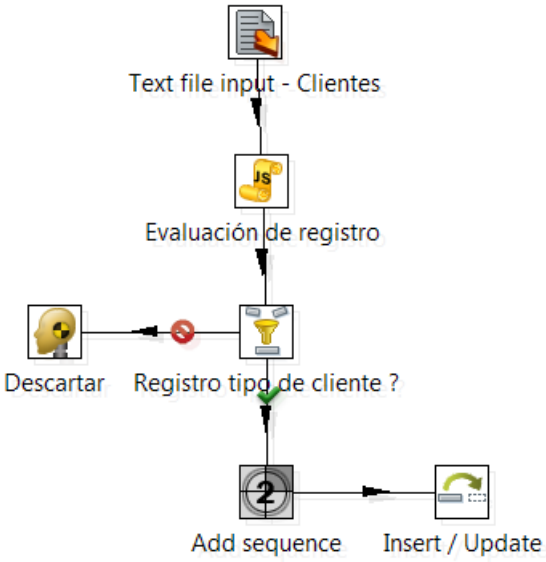

Ilustración 16 - Transformación E_ETL_C_TIPOCLIENTE

T_ETL_C_ZONA

Para poder realizar la transformación hay que separar la información en varios trozos, ya que en el mismo fichero ClientesZona.csv, viene información de los clientes junto con un maestro de poblaciones y localidades.

Por la estructura que tiene el fichero se decide leer dos veces, una para poder separar los registros relacionados con zonas de los de clientes y la otra lectura para saber si el registro tiene o no información de la provincia, para más adelante unirlo al *stream* principal e insertarlo en la dimensión correspondiente.

Se han encontrado varios problemas a la hora de realizar este paso:

1. Información de clientes con información de las zonas geográficas
2. En las zonas geográficas, la estructura empieza como población1, población2 y provincia pero luego cambia a provincia, país y jerarquía.
3. Registros en diferentes formatos (Jaén Provincia/Jaén provincia).
4. Añadirle el país que le corresponde y no hacerlo de manera fija.

Para poder solventar estos problemas, se hace uso del paso "*Fuzzy Match"* que se encarga de buscar cadenas de caracteres que identifican, usando algoritmos de detección de duplicidades, el valor que más posibilidades obtiene, calculando la similitud de dos *stream* de datos.

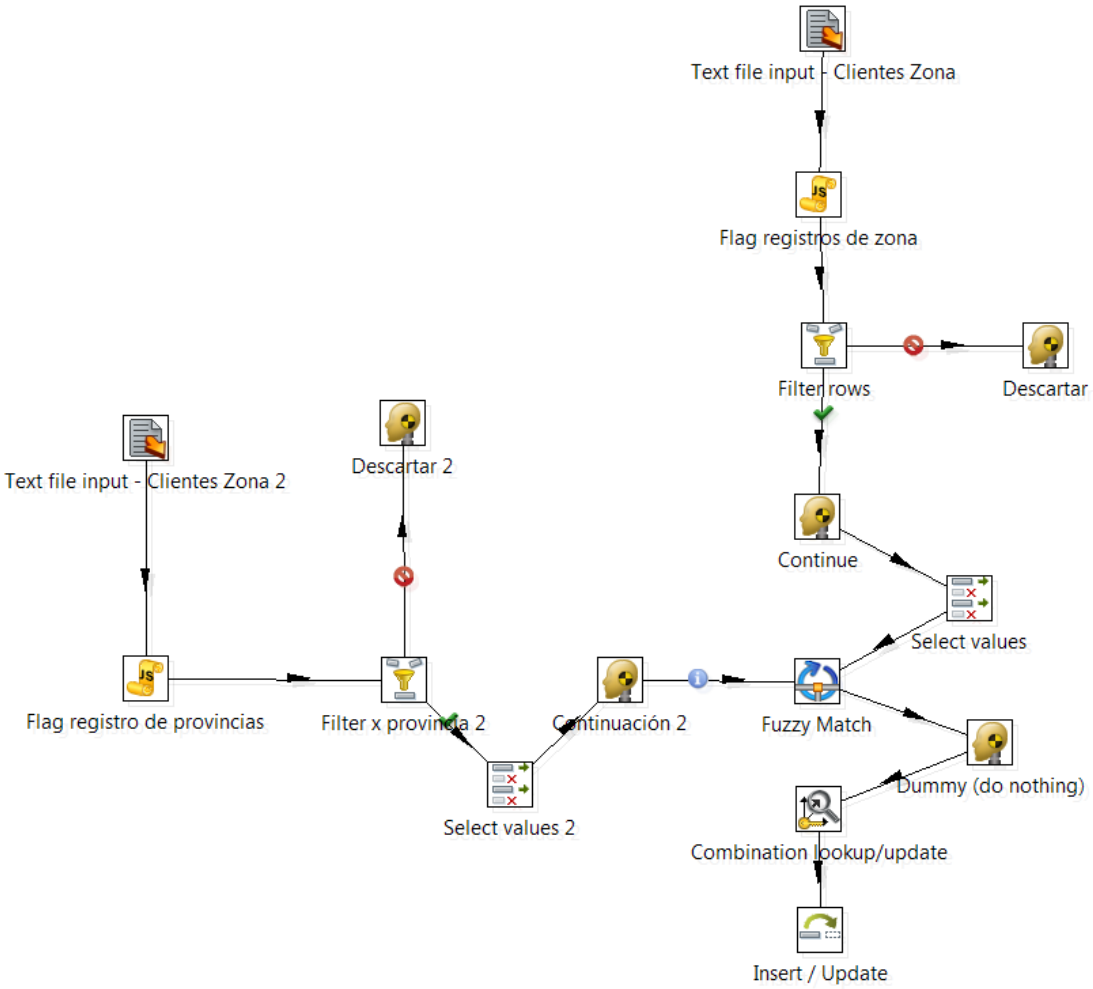

Ilustración 17 - Transformación T_ETL_C_ZONA

T_ETL_C_VENTA

Una vez se han cargado las dimensiones, se llegará a la última transformación, que cargará el hecho ventas h_venta.

Para poder cargar la información correctamente se realizan ciertas tareas de *data cleansing.* Varias columnas contienen datos erróneos y se ha decidido cambiarlos por un único valor que se pueda analizar a posteriori y evaluar consecuentemente.

Primero se lee el fichero para hacer el *data cleansing*:

Java script :

Script 1

```
//Script here

var almacenes_filtered;
var articulos_filtered;
var conceptos_filtered;
var formasPago_filtered;
var familias_filtered;

if (ALMACENES == "GEN_ID" || ALMACENES == "+" || ALMACENES == "-" || ALMACENES == "--"){
    almacenes_filtered = "-";
}else{
    almacenes_filtered = ALMACENES;
}

if (ARTICULOS == "" || ARTICULOS == "+" || ARTICULOS == "-" || ARTICULOS == "--" || ARTICULOS == "---"){
    articulos_filtered = "-";
}else{
    articulos_filtered = ARTICULOS;
}

if (CONCEPTOS == "GEN_ID" || CONCEPTOS == "-" || CONCEPTOS == "--" || CONCEPTOS == "+"){
    conceptos_filtered = "-";
}else{
    conceptos_filtered = CONCEPTOS;
}

if (FORMASPAGO == "GEN_ID" || FORMASPAGO == "-" || FORMASPAGO == "--" || FORMASPAGO == "+" ){
    formasPago_filtered = "-";
}else{
    formasPago_filtered = FORMASPAGO;
}

if (FAMILIASARTICULO == "GEN_ID" || FAMILIASARTICULO == "+" || FAMILIASARTICULO == "-" || FAMILIASARTICULO == "--" || FAMILIASAR
    familias_filtered = "-";
}else{
    familias_filtered = FAMILIASARTICULO;
}
```

Ilustración 18 - *Data cleansing* en T_ETL_C_VENTA

Se analiza cada una de las columnas problemáticas, y en donde se encuentra alguna discrepancia se cambia por “-“, o si no, se deja el dato contenido en el fichero.

Otro de los pasos complejos es el de artículos y familias de artículos que pueden tener indistintamente códigos de familia y códigos de artículos. Por lo que se hace una lectura de la dimensión y se filtran los artículos para asignarles el código de artículo que le corresponda.

El resto de pasos, al igual que el de la fecha, buscan en la dimensión correspondiente su id, se añade al *stream* y en el último paso se añade la secuencia y se hace la inserción en la base de datos.

Problemas encontrados: en la columna de Almacenes aparecen datos que no indican el almacén correspondiente i.e: GEN_ID. En la columna de Artículos no sólo aparecen códigos de artículos, sino también de familia, lo que hace que se generen incongruencias.

En el *data cleansing* que se realiza, se cambiarán estos valores por “-“, indicando que se desconoce el dato para el análisis.

Gracias a este proceso, se podrá indicar a la empresa, la cantidad de información que no es posible analizar y que se debe mejorar para futuros análisis que se realicen.

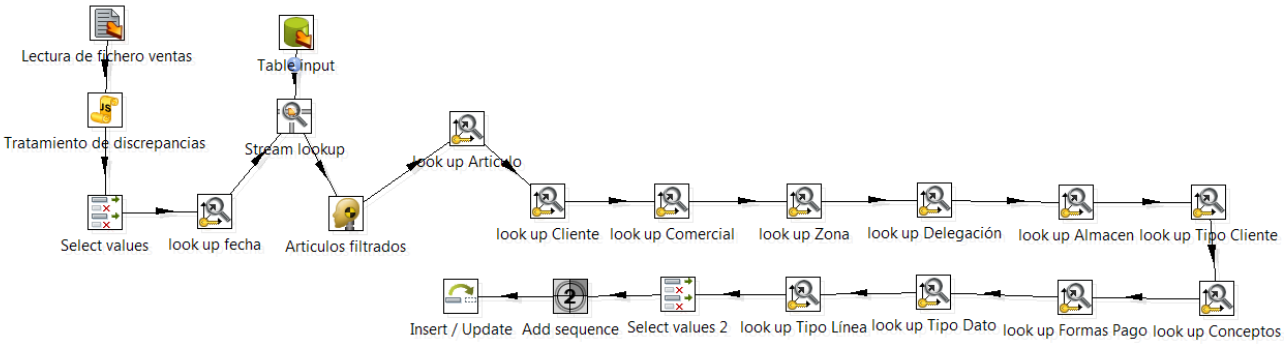

Ilustración 19 - Transformación T_ETL_C_VENTA

En las transformaciones de las dimensiones anteriormente expuestas, se ha utilizado la propiedad de “filtrado de datos”. Esta propiedad se ha usado con el fin de filtrar aquellos datos que se querían omitir. Las medidas llevadas a cabo han sido las siguientes:

- Omisión de los registros que empiezan por “+”.
- Omisión de los registros que empiezan con dos o tres guiones, i.e.: “--”, “---”.

Una vez realizado el análisis de la información, hay que explicar a la empresa la importancia de tener la información correcta en su sistema para que no genere inconsistencias y el análisis de los datos no se vea afectado por las mismas.

4.3. Cubo OLAP

Como se ha comentado en el punto 3.4, la creación del cubo OLAP se realiza usando la herramienta de Pentaho: Schema Workbench, que genera un esquema Mondrian llamado TFM_Ventas_final. Este esquema será analizado por el sistema Business Analytics de Pentaho.

Este cubo OLAP, Ventas, está formado por la tabla de hechos, las dimensiones, explicadas con anterioridad, y las jerarquías, que permitirán al usuario tener más nivel de detalle cuando consulte los informes.

Una vez se ha generado el cubo, se guarda el esquema como Ventas.xml.

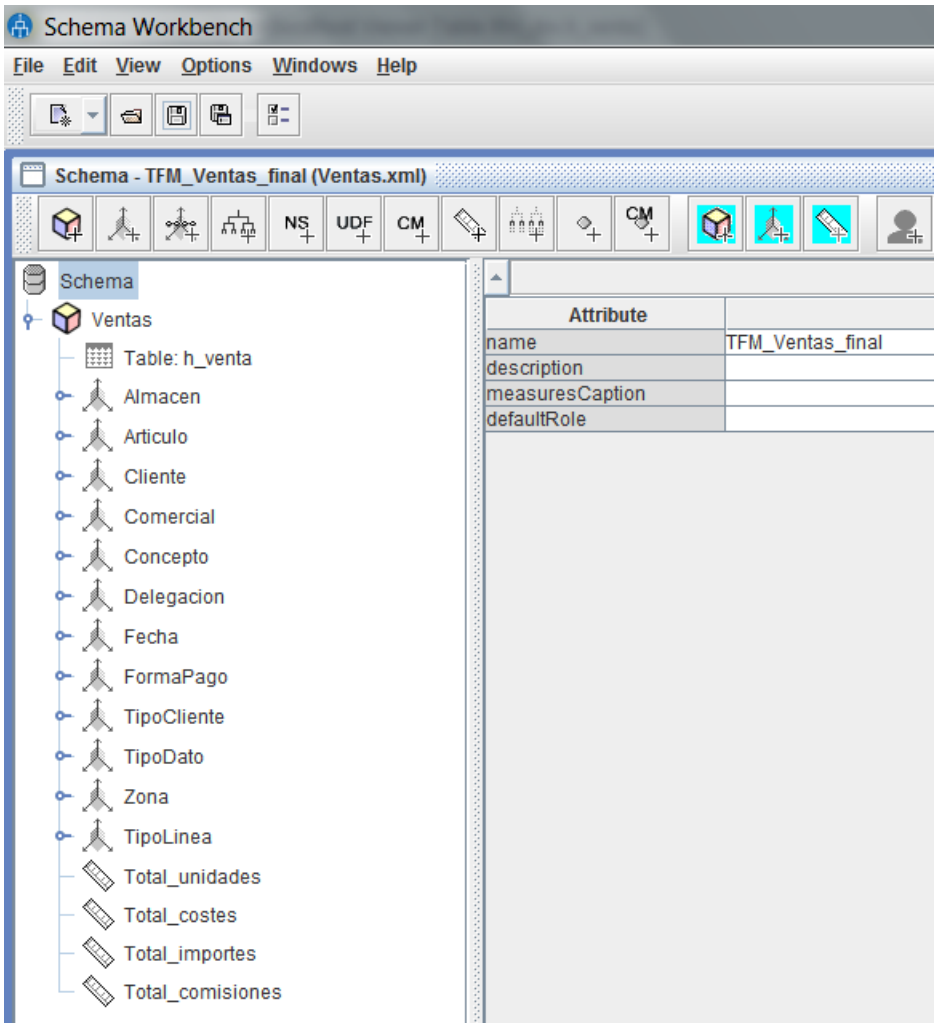

Ilustración 20 - Cubo OLAP

4.4. Pentaho Business Analytics

Gracias a esta versión *community* de Pentaho se realizarán los informes necesarios para el análisis de la empresa TOTSALES.

Para poder generar estos informes será necesario establecer un *data source* (TFM_Ventas_final) y para ello lo que se hará es importar el esquema Mondrian que se ha generado con Schema Workbench (Ventas.XML).

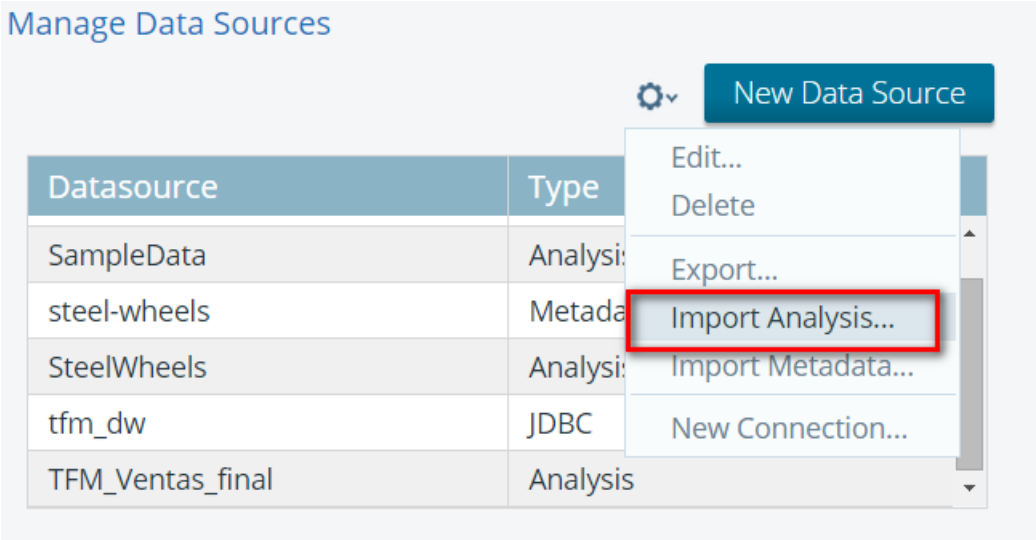

Ilustración 21 - Creación *data source* de análisis

Una vez creado el nuevo *data source*, se podrá explotar el cubo a través del plug-in de Saiku Analytics.

5. Explotación

5.1. Entrada en el sistema

A través del navegador y conectándose a la URL establecida, el usuario podrá acceder al portal de análisis donde tendrá a su disposición la conexión y el cubo a analizar. A través de las diferentes herramientas de explotación que dispone la herramienta de Pentaho, el usuario podrá analizar y diseñar sus propios reportes así como, establecer tareas de generación de informes para su publicación.

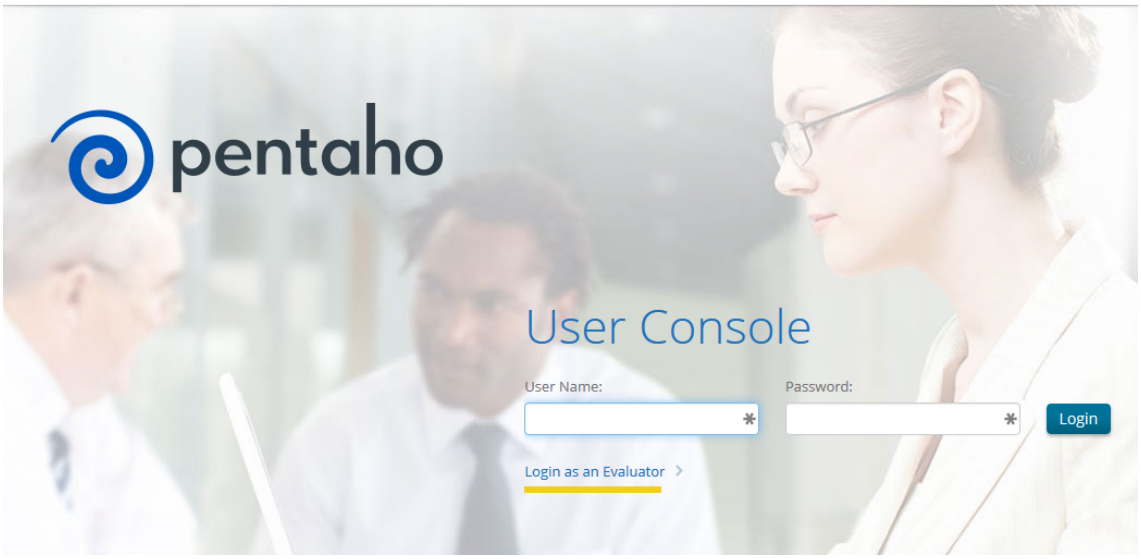

Ilustración 22 - Pantalla de log-in en el servidor BA de Pentaho ce

El usuario podrá realizar varias acciones, desde explotar el cubo, pudiendo sacar los informes necesarios para su propia explotación, hasta crear tareas para la generación de los informes con una frecuencia determinada. Pudiendo enviar estos informes vía correo electrónico o publicándolos en una dirección de red que se le indique.

5.2. Creación de informes

Para ello se utilizará la herramienta de reporting que aparece en el Marketplace de Pentaho llamada Saiku Analytics. Herramienta de análisis intuitiva y con un buen diseño.

Se hará uso del interfaz de diseño para la realización de los informes que pide la empresa TOTSALES.

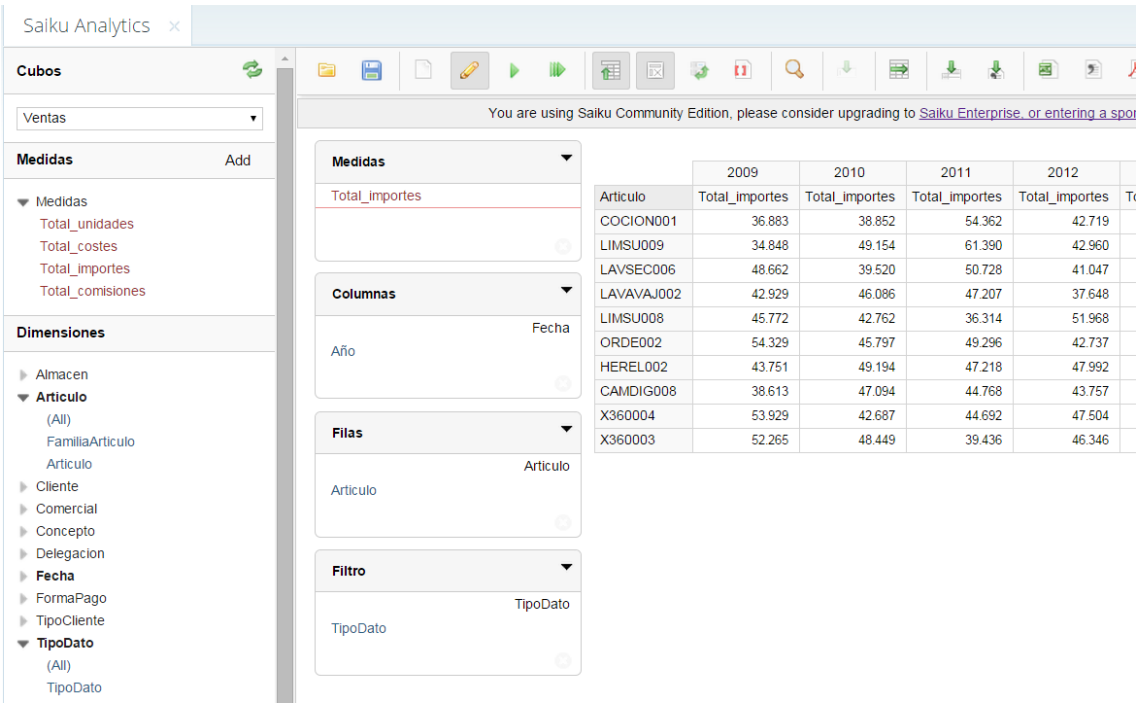

Ilustración 23 - Herramienta de diseño de informes

5.2.1. Evolución de las ventas en unidades vendidas/importe.

El análisis se hace en base al último año completo de los datos obtenidos por las fuentes (2012), comparados con el año anterior (2011).

En donde se puede observar la evolución de las ventas por delegación:

Delegación de Asturias (AIBE_001): se observa que las ventas del 2012 son más bajas que las del 2011, pero la tendencia es bastante parecida, donde habría que remarcar como los meses de más ventas del 2011 a febrero, septiembre y noviembre mientras que en el 2012 serían febrero, mayo y octubre.

Delegación de Bilbao (AIBE_002): durante el 2011 tiene bastantes altibajos, siendo julio de 2011 el peor mes, seguramente por la salida de vacaciones de la gente que coincide para los dos años, mientras que agosto es un buen mes en ventas para el 2011, pero en 2012 las ventas siguen bajando. En los dos años, septiembre es un punto de inflexión y vuelven a subir.

Delegación de Levante (AIBE_003): para la delegación de levante, las ventas del 2012 son mejores que las del año anterior, aunque el 2011 fue un año con ventas más constantes salvo septiembre-octubre, mientras que en el 2012, las ventas caen a partir de julio para volver a subir de octubre-noviembre llegando a máximos.

Delegación de Sevilla (AIBE_004): la tendencia en las ventas es parecida durante todos los meses tanto el 2011 como en el 2012, salvo los meses que van de mayo a septiembre, que en el 2012 bajan las ventas con respecto al mes anterior, para volver a subir en los meses de octubre y noviembre.

Delegación de Galicia (AIBE_005): el 2011 es un año bastante constante en las ventas, mientras que en el 2012 se encuentran altibajos, llegando a mínimos de ventas en los meses de abril y de agosto.

		Delegación Asturias		Delegación Bilbao		Delegación Levante		Delegación Sevilla		Delegación Galicia	
Año	Mes	Total_unidades	Total_importes	Total_unidades	Total_importes	Total_unidades	Total_importes	Total_unidades	Total_importes	Total_unidades	Total_importes
2011	enero	540	195.086	500	173.722	542	193.716	539	188.541	569	207.910
	febrero	593	207.684	616	211.854	576	211.686	602	223.381	522	185.577
	marzo	494	178.344	466	165.182	496	179.719	467	171.931	522	182.318
	abril	532	190.206	492	182.756	526	177.301	497	178.395	552	194.206
	mayo	529	182.062	491	173.369	526	190.302	571	212.025	556	194.332
	junio	471	171.850	492	174.889	503	178.215	543	197.329	534	191.434
	julio	526	178.861	446	151.317	482	174.400	474	165.413	495	175.460
	agosto	504	173.268	566	210.064	531	188.147	468	165.341	494	172.603
	septiembre	601	211.405	490	170.725	495	170.538	530	181.371	544	199.264
	octubre	521	191.923	546	183.833	433	152.194	570	191.294	574	195.301
	noviembre	580	210.350	584	212.232	524	181.549	565	202.323	558	191.259
	diciembre	575	195.052	547	190.452	541	184.190	490	170.813	562	192.155
2012	enero	516	175.607	500	172.879	553	191.918	533	182.842	542	189.617
	febrero	492	176.562	593	199.762	556	198.077	591	198.822	546	193.434
	marzo	556	195.436	562	193.830	560	192.934	590	202.316	654	221.684
	abril	518	182.562	494	174.658	514	186.829	493	164.730	460	157.836
	mayo	575	191.891	567	202.756	596	207.922	501	187.409	502	170.716
	junio	516	182.057	529	182.022	500	173.337	530	185.173	543	188.125
	julio	467	165.294	477	170.857	541	185.367	489	178.529	495	173.285
	agosto	494	172.915	481	167.270	467	165.427	467	154.868	394	134.364
	septiembre	514	176.063	435	149.955	402	138.709	407	133.588	433	156.169
	octubre	567	195.478	528	184.412	516	171.905	518	181.958	504	172.509
	noviembre	567	188.033	644	205.561	661	234.195	654	228.753	647	217.340
	diciembre	500	174.634	525	180.962	486	171.330	514	186.144	494	170.825
2013	enero	469	150.140	453	158.027	496	174.242	438	151.258	459	157.924

Ilustración 24 - Evolución de las ventas

Como se puede ver en la siguiente gráfica la evolución de las ventas que van del 2009 al 2012 no es significativa, sí que es verdad que tanto en las ventas por importe como las ventas por unidades, las ventas ascienden del período que va desde el 2009 al 2011, mientras que en el de 2012 bajan al mismo nivel que las del 2011.

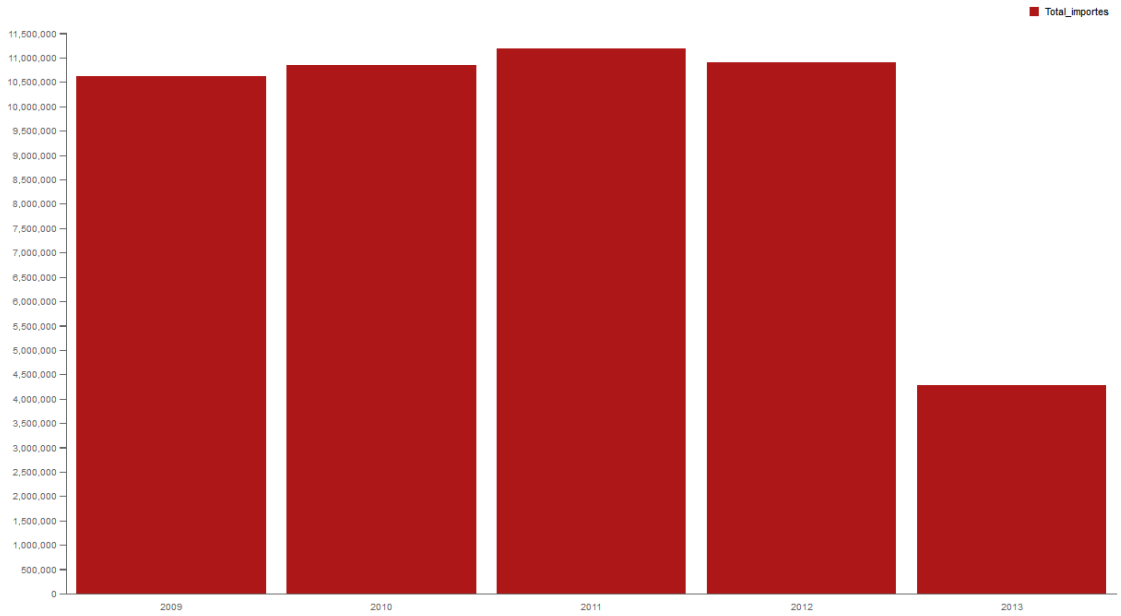

Ilustración 25 - Ventas anuales por importes

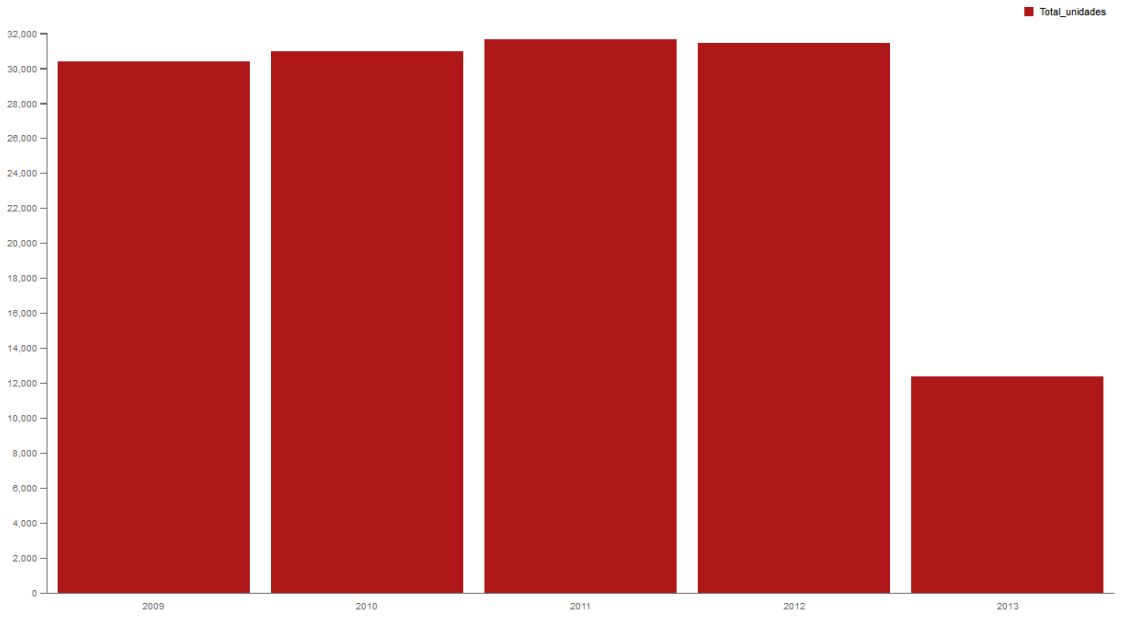

Ilustración 26 - Ventas anuales por unidades

En la siguiente gráfica se puede ver una comparativa de la evolución de los meses del 2013 en los que se tienen datos, con respecto a los mismos meses del 2012.

En donde se puede apreciar, que en enero de los dos años las ventas eran bajas, quizás debido a lo que se llama “la cuesta de enero”. La subida en las ventas del 2013 es más pronunciada que la de 2012 pero más corta, terminado de estancarse durante hasta mediados de abril que vuelve a

bajar. Quizás esta última bajada se deba a que no se tienen datos del mes completo.

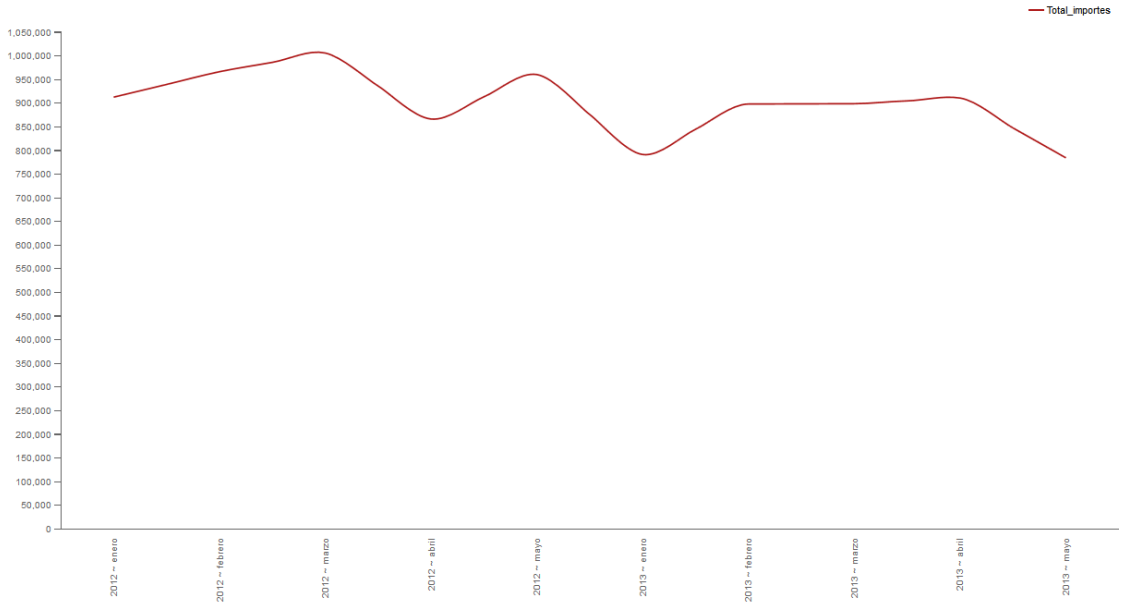

Ilustración 27 - Evolución ventas en los primeros meses 2012 y 2013

5.2.2. Artículos más/menos vendidos por delegación y año.

Como se puede ver en el informe, el ordenador Asus y la video consola XBOX con el juego MadGatz seguidos muy de cerca del microondas Tristar.

		Ordenador Asus CM6730-ESCH18	X BOX 360 MadCatz MLG Pro-Circuit Contoller XBOX	Microondas Tristar MW 2890	Lijadora Dremel 4000 1/45 (F0134000 JC)	Aspiradora de mano Electrolux ZB2906	X BOX 360 Electronic Arts XBOX360 FIFA 13	Secadora Bosch WVH2846XEE	Lavavajillas Bosch SMS40D12EU	Cámara digital Samsung ES90 Rojo PACK	Aspiradora de mano Electrolux ZB2908W
Año	Delegacion	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades
2009	Delegación Asturias	40	38	11	34	18	34	38	15	24	26
	Delegación Bilbao	49	28	18	18	7	16	24	22	17	13
	Delegación Levante	13	21	16	19	19	31	17	26	24	37
	Delegación Sevilla	28	28	32	16	29	34	28	33	32	22
	Delegación Galicia	21	31	29	37	35	28	26	27	19	31
2010	Delegación Asturias	24	25	24	16	38	39	21	24	23	20
	Delegación Bilbao	38	17	19	24	26	40	19	30	27	30
	Delegación Levante	18	18	27	24	22	24	25	18	32	35
	Delegación Sevilla	37	36	34	43	23	23	24	24	31	18
	Delegación Galicia	16	29	19	26	25	19	20	30	19	28
2011	Delegación Asturias	33	17	24	42	36	31	28	35	22	10
	Delegación Bilbao	16	20	30	27	42	18	31	23	25	23
	Delegación Levante	24	21	43	18	15	25	41	33	34	31
	Delegación Sevilla	39	31	30	27	35	23	29	19	17	25
	Delegación Galicia	24	36	27	26	35	9	21	33	35	15
2012	Delegación Asturias	23	28	33	22	22	28	26	20	19	28
	Delegación Bilbao	32	36	26	38	25	31	28	17	43	29
	Delegación Levante	31	28	25	9	21	9	28	21	15	30
	Delegación Sevilla	19	26	25	30	21	27	13	26	36	36
	Delegación Galicia	20	17	23	29	27	34	24	33	14	26
2013	Delegación Asturias	10	16	10	3	7	10	12	8	8	6
	Delegación Bilbao	9		6	10	14	13	8	8	17	14
	Delegación Levante	5	8	22	10	17	5	13	11	9	4
	Delegación Sevilla	2	14	17	10	9	8	9	10	14	10
	Delegación Galicia	8	10	6	13	2	5	9	14	3	10
	Grand Total	579	579	576	571	570	564	562	560	559	557

Ilustración 28 - Artículos más vendidos

Mientras que los menos vendidos en el mismo período son la cafetera Krups, el televisor Philips de 32” y el lavavajillas Fagor.

		Cafetera Krups KP 5001 Dolce Gusto Circolo	Televisor Philips 32PFL7606H/12	Lavavajillas Fagor LVF-13 X	Robot iRobot Roomba 555	X BOX 360 TRITTON AX 180	Carro Kärcher HT 4.520 Kit 20Mts 2.645-169.0	Lavadora y secadora Hotpoint WMG 823 B EU	Apple Nuevo iPad Wi-Fi 4G 64GB Negro	Frogorífico Indesit RAA 24 N	Televisor Philips 40PFL5527H/12
Año	Delegacion	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades
2009	Delegación Asturias	15	16	9	23	18	17	18	14	16	31
	Delegación Bilbao	13	10	31	15	29	21	19	15	23	25
	Delegación Levante	14	24	11	9	17	14	27	24	17	16
	Delegación Sevilla	14	9	12	26	18	13	15	18	25	25
	Delegación Galicia	27	24	17	20	13	7	25	8	20	24
2010	Delegación Asturias	7	5	12	30	10	19	15	18	24	25
	Delegación Bilbao	11	43	24	14	15	8	29	16	14	23
	Delegación Levante	5	30	11	15	23	37	20	18	20	14
	Delegación Sevilla	24	24	8	16	32	30	25	29	16	10
	Delegación Galicia	20	23	24	17	14	22	21	18	24	21
2011	Delegación Asturias	32	18	29	16	28	22	24	17	22	26
	Delegación Bilbao	23	16	23	29	16	23	20	16	22	20
	Delegación Levante	23	17	15	20	17	32	17	26	20	14
	Delegación Sevilla	19	23	25	17	18	20	18	18	24	18
	Delegación Galicia	14	11	19	20	27	22	18	29	23	23
2012	Delegación Asturias	20	25	21	19	12	21	18	18	22	17
	Delegación Bilbao	28	24	15	17	11	18	20	19	21	9
	Delegación Levante	26	15	30	29	17	17	21	24	15	24
	Delegación Sevilla	11	7	21	16	18	7	23	32	20	28
	Delegación Galicia	28	12	18	18	26	13	7	14	19	14
2013	Delegación Asturias	8	7	3	15	12	12	8	12	14	3
	Delegación Bilbao	4	11	3	10	4	4	5	12		8
	Delegación Levante	12	12	15	7	8	8	8	11	11	8
	Delegación Sevilla	8	9	17	8	15	10	6	5	7	8
	Delegación Galicia	11	5	14	2	10	14	8	7	4	10
	Grand Total	**417**	**420**	**427**	**428**	**428**	**431**	**435**	**438**	**443**	**444**

Ilustración 29 - Artículos menos vendidos

5.2.3. Familias más/menos vendidas por delegación y año.

En este informe se puede observar que la familia de artículos más vendidos en todas las delegaciones es la de los electrodomésticos, seguida por la de informática. Encabeza la lista de mayor número de electrodomésticos vendidos la delegación de Asturias en el año 2012 con 1758 unidades.

		Electrodomésticos	Informática	Electrónica consumo	Casa y jardín	fotografía, video y óptica	Cónsolas y videojuegos
Año	Delegacion	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades
2009	Delegación Asturias	1.727	763	576	524	427	261
	Delegación Bilbao	1.522	793	592	478	403	236
	Delegación Levante	1.610	728	551	500	486	218
	Delegación Sevilla	1.645	767	615	464	512	263
	Delegación Galicia	1.668	732	550	420	375	276
2010	Delegación Asturias	1.665	816	618	481	424	261
	Delegación Bilbao	1.722	773	639	495	379	270
	Delegación Levante	1.601	787	659	535	448	219
	Delegación Sevilla	1.685	796	515	517	427	248
	Delegación Galicia	1.638	822	580	481	441	241
2011	Delegación Asturias	1.706	918	613	481	422	267
	Delegación Bilbao	1.698	790	595	490	423	230
	Delegación Levante	1.717	697	654	528	419	235
	Delegación Sevilla	1.740	839	611	518	398	276
	Delegación Galicia	1.747	816	566	567	498	259
2012	Delegación Asturias	1.758	778	630	529	384	265
	Delegación Bilbao	1.693	786	603	547	425	244
	Delegación Levante	1.695	786	644	543	447	230
	Delegación Sevilla	1.668	864	678	472	484	249
	Delegación Galicia	1.757	779	570	461	419	279
2013	Delegación Asturias	718	293	254	182	161	109
	Delegación Bilbao	648	329	247	167	199	96
	Delegación Levante	696	287	231	226	152	95
	Delegación Sevilla	691	290	211	207	203	125
	Delegación Galicia	632	334	209	239	146	90
	Grand Total	**37.047**	**17.363**	**13.211**	**11.052**	**9.502**	**5.542**

Ilustración 30 – Familias más vendidas

Mientras que las menos vendidas son las de consolas y videojuegos y fotografía, vídeo y óptica.

		Cónsolas y videojuegos	fotografía, video y óptica	Casa y jardín	Electrónica consumo	Informática	Electrodomésticos
Año	Delegacion	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades	Total_unidades
2009	Delegación Asturias	261	427	524	576	763	1.727
	Delegación Bilbao	236	403	478	592	793	1.522
	Delegación Levante	218	486	500	551	728	1.610
	Delegación Sevilla	263	512	464	615	767	1.645
	Delegación Galicia	276	375	420	550	732	1.668
2010	Delegación Asturias	261	424	481	618	816	1.665
	Delegación Bilbao	270	379	495	639	773	1.722
	Delegación Levante	219	448	535	659	787	1.601
	Delegación Sevilla	248	427	517	515	796	1.685
	Delegación Galicia	241	441	481	580	822	1.638
2011	Delegación Asturias	267	422	481	613	918	1.706
	Delegación Bilbao	230	423	490	595	790	1.698
	Delegación Levante	235	419	528	654	697	1.717
	Delegación Sevilla	276	398	518	611	839	1.740
	Delegación Galicia	259	498	567	566	816	1.747
2012	Delegación Asturias	265	384	529	630	778	1.758
	Delegación Bilbao	244	425	547	603	786	1.693
	Delegación Levante	230	447	543	644	786	1.695
	Delegación Sevilla	249	484	472	678	864	1.668
	Delegación Galicia	279	419	461	570	779	1.757
2013	Delegación Asturias	109	161	182	254	293	718
	Delegación Bilbao	96	199	167	247	329	648
	Delegación Levante	95	152	226	231	287	696
	Delegación Sevilla	125	203	207	211	290	691
	Delegación Galicia	90	146	239	209	334	632
	Grand Total	**5.542**	**9.502**	**11.052**	**13.211**	**17.363**	**37.047**

Ilustración 31 - Familias menos vendidas

Como se puede observar en la gráfica, se ve claramente la diferencia que hay entre la familia “electrodomésticos” y el resto, llamando también la atención la igualdad de las gráficas para cada año.

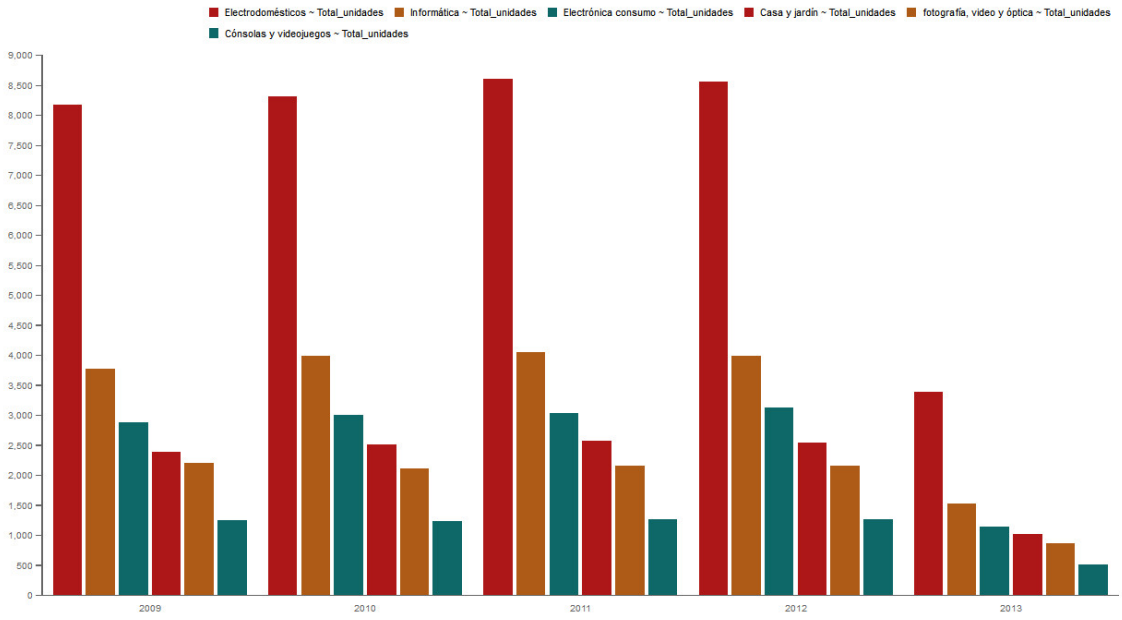

Ilustración 32 - Familias más/menos vendidas

5.2.4. Clientes a los que más se les factura.

Los diez clientes que aparecen en el listado a continuación son aquellos a los que la empresa deberá prestar más atención utilizando alguna política de fidelización y siguiendo de cerca cualquier comentario o incidencia que pudiesen tener.

	JAnual	
Cliente	Total_unidades	Total_importes
MIRA MORENO, ISABEL	2.183	772.129
RIBE PIE, JORDI	1.959	686.618
Nuria Torres Suárez	1.885	656.406
CUEVAS- CAMACHO, M.JOSE	1.903	654.454
SAEZ PELAYO, JOSEFA	1.839	645.564
GASCO CORRAL, CARLOS JAVIER	1.854	645.398
HIGUERAS RODRIGUEZ, ANGUS	1.802	627.333
ALTES BALANYA, MIREIA	1.753	613.370
ROY BELLO, LORENZO	1.787	613.186
MAYO FERNANDEZ, MANUEL	1.726	612.276

Ilustración 33 - Clientes a los que más se les factura

En la siguiente gráfica se puede ver claramente cuál es el TOP10 de clientes a los que más se factura.

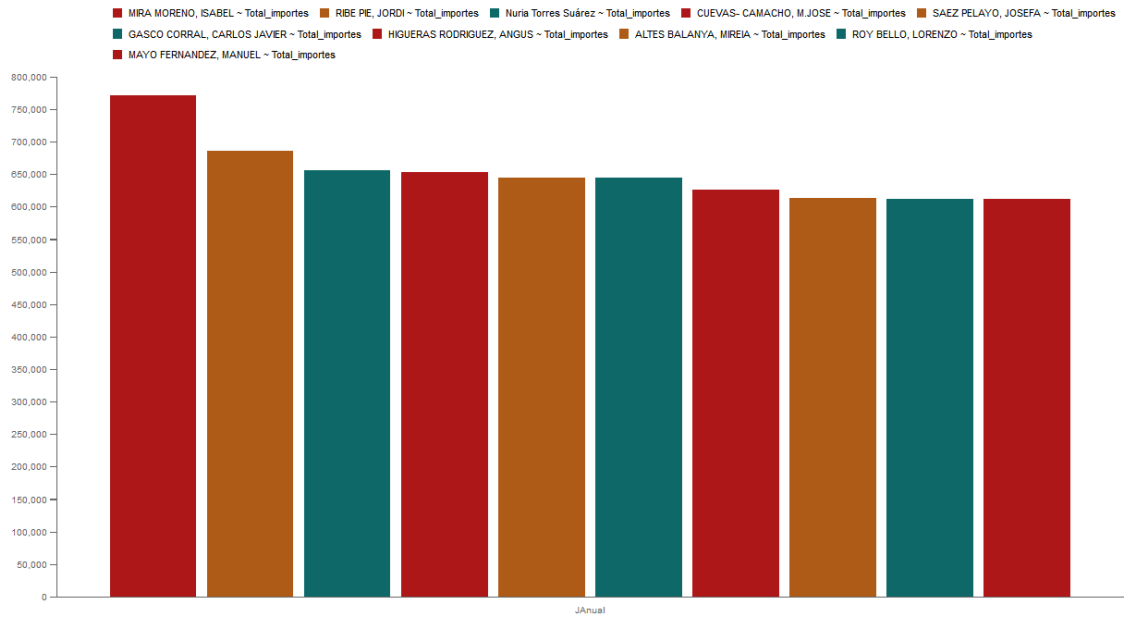

Ilustración 34 - Clientes a los que más se les factura

5.2.5. Las zonas en donde más/menos se vende.

En el listado se muestran las ventas durante los últimos años, ordenadas por total facturado acumulado durante estos años, donde se pueden apreciar las que más y menos movimientos tienen.

Se podría utilizar este listado y sacar uno con los mejores vendedores para intentar mejorar las ventas de las zonas más problemáticas. Aunque la movilidad de un comercial será siempre complicada, se podría motivar a que cambiaran de zona con incentivos según su desempeño.

En el 2013 sólo aparecen las ventas hasta mayo.

	2009	2011	2012	2013	
JProvincia	Total_importes	Total_importes	Total_importes	Total_importes	Total_importes Grand Total
Girona	656.197	676.081	673.631	275.517	**2.281.426**
Cáceres	567.999	729.512	636.872	270.520	**2.204.903**
Rioja (La)	559.131	621.985	558.646	200.717	**1.940.479**
Zaragoza	544.114	506.402	572.646	184.747	**1.807.909**
Barcelona	610.865	422.639	605.878	193.021	**1.832.403**
Burgos	436.440	462.050	358.366	231.151	**1.488.007**
Guadalajara	510.971	447.910	444.262	170.397	**1.573.540**
Valencia/València	370.669	573.934	421.481	233.263	**1.599.347**
Zamora	416.564	423.350	528.841	188.753	**1.557.508**
Valladolid	365.076	491.042	423.149	109.872	**1.389.139**
Salamanca	407.298	353.568	376.535	126.086	**1.263.487**
Tarragona	414.682	412.644	407.746	131.733	**1.366.805**
Lleida	443.821	341.774	445.487	82.432	**1.313.514**
Almería	294.134	358.336	371.159	186.867	**1.210.496**
Soria	331.949	321.178	261.860	91.344	**1.006.331**
Avila	286.615	239.523	306.027	72.611	**904.776**
Cuenca	224.069	305.746	258.104	159.836	**947.755**
Navarra	311.817	256.822	276.292	86.552	**931.483**
Palencia	180.299	250.095	245.621	157.545	**833.560**
Granada	287.351	242.355	228.254	70.905	**828.865**
Pontevedra	147.953	271.027	255.114	102.640	**776.734**
León	196.820	236.794	220.935	76.553	**731.102**
Madrid	267.125	201.774	163.719	90.982	**723.600**
Badajoz	167.692	193.026	236.718	78.096	**675.532**
Vizcaya	168.182	132.825	190.535	56.601	**548.143**
Teruel	112.813	133.681	163.485	37.128	**447.107**
Huelva	139.527	159.066	119.583	38.095	**456.271**
Toledo	95.288	156.466	118.104	62.932	**432.790**
Asturias	146.839	165.122	56.870	77.833	**446.664**
Jaén Provincia	150.798	144.646	121.688	40.772	**457.904**

Ilustración 35 - Zonas con más ventas (ordenadas de mayor a menor)

	2009	2011	2012	2013	
JProvincia	Total_importes	Total_importes	Total_importes	Total_importes	Total_importes Grand Total
Castellón/Castelló	86.371	77.484	57.205	81.514	**302.574**
Córdoba	62.671	126.107	134.855	32.250	**355.883**
Cantabria	98.899	104.598	90.672	61.905	**356.074**
Soria provincia	105.215	150.799	87.446	39.203	**382.663**
Huesca	114.852	136.319	166.415	37.362	**454.948**
Balears (Illes)	96.402	111.898	108.974	58.136	**375.410**
Sin asignar	139.216	123.835	125.067	34.622	**422.740**
Sevilla	121.420	136.006	92.200	54.114	**403.740**
Jaén Provincia	150.798	144.646	121.688	40.772	**457.904**
Asturias	146.839	165.122	56.870	77.833	**446.664**
Toledo	95.288	156.466	118.104	62.932	**432.790**
Huelva	139.527	159.066	119.583	38.095	**456.271**
Teruel	112.813	133.681	163.485	37.128	**447.107**
Vizcaya	168.182	132.825	190.535	56.601	**548.143**
Badajoz	167.692	193.026	236.718	78.096	**675.532**
Madrid	267.125	201.774	163.719	90.982	**723.600**
León	196.820	236.794	220.935	76.553	**731.102**
Pontevedra	147.953	271.027	255.114	102.640	**776.734**
Granada	287.351	242.355	228.254	70.905	**828.865**
Palencia	180.299	250.095	245.621	157.545	**833.560**
Navarra	311.817	256.822	276.292	86.552	**931.483**
Cuenca	224.069	305.746	258.104	159.836	**947.755**
Avila	286.615	239.523	306.027	72.611	**904.776**
Soria	331.949	321.178	261.860	91.344	**1.006.331**
Almería	294.134	358.336	371.159	186.867	**1.210.496**
Lleida	443.821	341.774	445.487	82.432	**1.313.514**
Tarragona	414.682	412.644	407.746	131.733	**1.366.805**
Salamanca	407.298	353.568	376.535	126.086	**1.263.487**
Valladolid	365.076	491.042	423.149	109.872	**1.389.139**

Ilustración 36 - Zonas con menos ventas (ordenadas de menor a mayor)

En la gráfica que se encuentra a continuación, se puede apreciar con una simple observación cuáles son las zonas que más venden, como Girona, y las que menos, como Castellón.

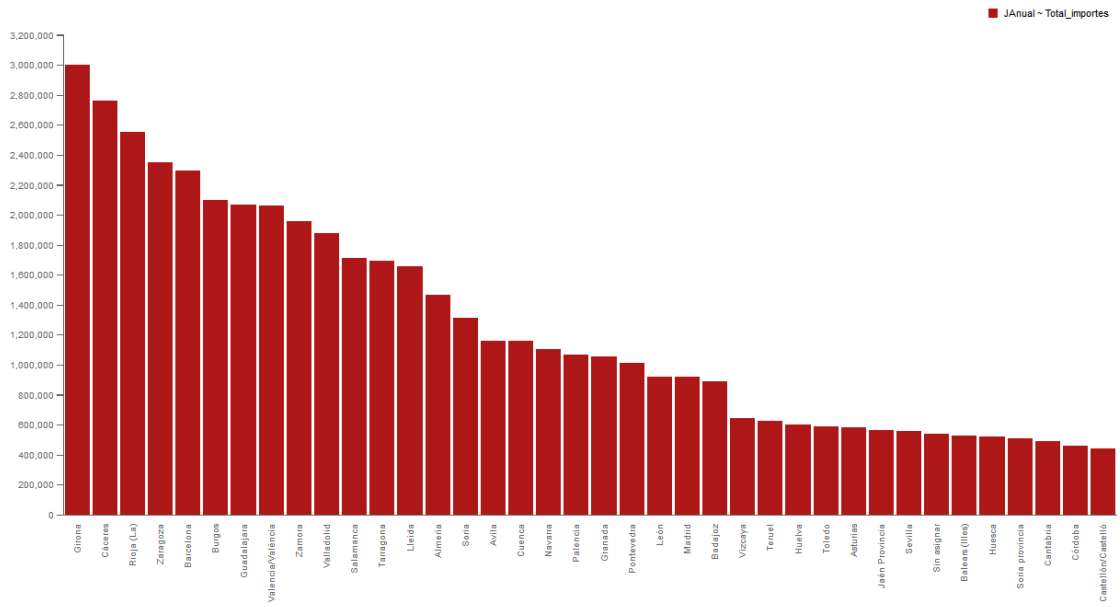

Ilustración 37 - Distribución de las ventas por zonas

5.2.6. Evolución de las ventas en el tiempo, por importe.

Este listado muestra la evolución de las ventas, tanto por unidades como por importes.

	2009		2010		2011		2012		2013	
Mes	Total_unidades	Total_importes	Total_unidades	Total_importes	Total_unidades	Total_importes	Total_unidades	Total_importes	Total_unidades	Total_importes
enero	2.394	837.773	2.572	902.181	2.690	958.975	2.644	912.863	2.315	791.591
febrero	2.702	950.620	2.562	891.148	2.909	1.040.182	2.778	966.657	2.605	898.361
marzo	2.785	971.609	2.322	822.859	2.445	877.494	2.922	1.006.200	2.621	899.007
abril	2.011	712.411	2.776	998.976	2.599	922.864	2.479	866.615	2.580	911.031
mayo	2.175	755.523	2.494	877.030	2.673	952.090	2.741	960.694	2.259	784.617
junio	2.539	912.616	2.484	862.159	2.543	913.717	2.618	910.714		
julio	2.506	864.030	2.269	793.882	2.423	845.451	2.469	873.332		
agosto	2.757	972.447	2.637	921.284	2.563	909.423	2.303	794.844		
septiembre	2.365	815.850	2.914	1.017.421	2.660	933.303	2.191	754.484		
octubre	2.875	1.008.219	2.666	929.118	2.644	914.545	2.633	906.262		
noviembre	2.580	885.951	2.543	887.635	2.811	997.713	3.173	1.073.882		
diciembre	2.721	951.095	2.788	956.106	2.715	932.662	2.519	883.895		

Ilustración 38 - Ventas en el tiempo

Gracias a esta gráfica se puede ver la variación de las ventas para todos los años y meses. La disminución de las ventas a partir de mayo puede deberse a los datos faltantes desde mayo 2013 en adelante.

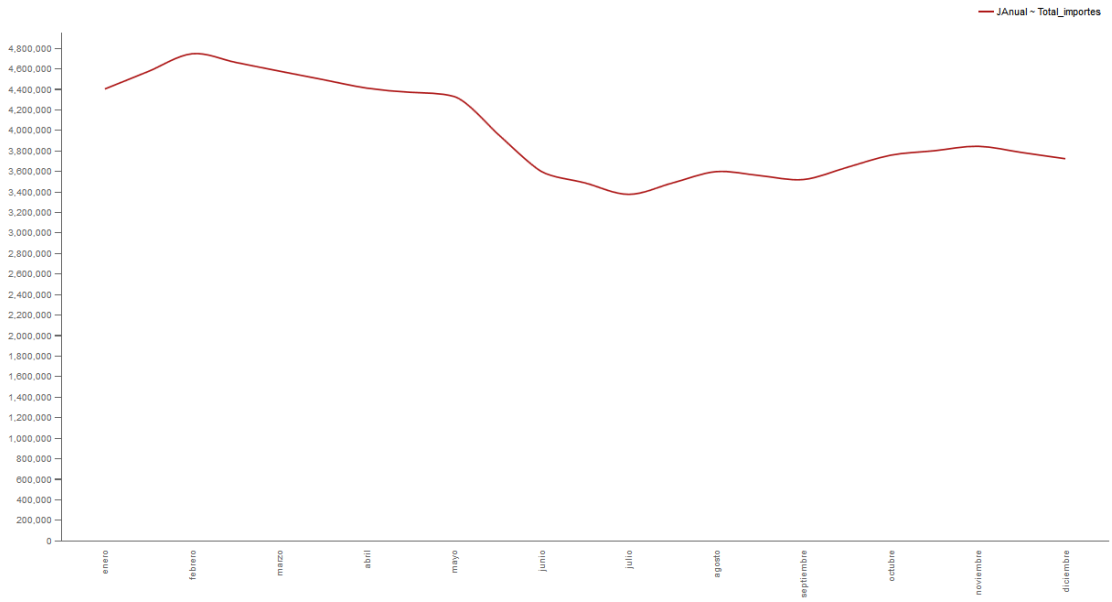

Ilustración 39 - Evolución de las ventas

5.2.7. Comisiones de los comerciales.

Listado donde se muestran las comisiones cobradas por los comerciales durante un período de tiempo determinado. Llama la atención que de los 50 comerciales que hay dados de alta en el sistema del cliente, sólo tengan registradas ventas 6 comerciales. Puesto que la empresa tiene 5 delegaciones. Se entiende que estos comerciales representan a cada una de estas delegaciones. Teniendo una de ellas, dos comerciales asignados quizás, por motivos de volumen de ventas. De la información facilitada podría entenderse que todos los comerciales de una zona se reparten las comisiones de forma equitativa indistintamente de las ventas que hiciera cada uno, de ahí que las ventas de cada zona las aglutine un único comercial.

		MANCEBO LOZANO, FRANCISCO JAVIER	GOMEZ DURAN, JOSE	DE TORRES ESCUDERO, MARIANO	DE LA CRUZ VILLARUBIA, JOSE LUIS	CABRERA CARRANZA, FELIPA	CORCHERO LENA, ROSA MARIA
Año	Mes	Total_comisiones	Total_comisiones	Total_comisiones	Total_comisiones	Total_comisiones	Total_comisiones
2009	enero	6.884	8.904	6.397	6.515	7.385	5.839
	febrero	7.678	7.314	8.809	7.280	8.225	8.281
	marzo	8.735	8.488	7.651	8.261	7.845	7.621
	abril	6.323	5.989	5.410	6.429	5.757	5.743
	mayo	5.772	5.801	7.992	5.365	5.971	6.900
	junio	8.074	8.116	7.380	7.229	6.867	7.996
	julio	7.645	6.457	8.188	7.676	6.918	6.346
	agosto	8.399	8.186	7.640	8.964	8.166	7.327
	septiembre	6.844	6.909	6.267	6.799	7.248	6.787
	octubre	7.567	7.874	7.703	8.853	9.302	9.145
	noviembre	7.490	7.276	6.397	7.818	8.042	7.321
	diciembre	7.772	6.754	8.409	8.928	7.214	8.529
2010	enero	8.017	7.689	5.939	7.979	8.212	7.323
	febrero	7.937	7.133	7.349	7.653	7.726	6.821
	marzo	7.071	7.592	7.251	6.927	6.645	5.687
	abril	7.292	8.101	8.697	9.011	8.262	8.633
	mayo	6.617	8.948	7.487	7.020	6.918	6.901
	junio	7.270	7.629	5.372	7.157	8.012	7.701
	julio	6.260	6.279	6.812	6.661	6.945	6.781
	agosto	8.268	7.328	8.232	7.769	6.698	7.814
	septiembre	8.596	9.671	7.587	7.219	8.401	9.422
	octubre	7.477	8.121	8.281	8.188	7.780	6.651
	noviembre	7.172	7.190	8.140	6.284	7.590	8.057
	diciembre	9.118	7.947	8.423	7.306	7.492	7.577
2011	enero	7.529	9.017	7.691	7.901	8.258	7.611
	febrero	8.608	9.173	9.060	8.067	8.886	8.254
	marzo	7.823	6.667	7.491	7.341	7.463	7.135
	abril	7.506	9.150	8.221	6.839	6.807	7.662
	mayo	8.195	7.347	7.950	7.759	7.602	8.800
	junio	7.748	8.419	7.474	6.622	7.712	7.753
	julio	7.250	6.798	7.893	7.106	7.171	6.088

Ilustración 40 - Comisiones de los comerciales

En la siguiente gráfica se aprecia las comisiones anuales por comercial, donde se puede ver, la poca diferencia que hay entre ellos.

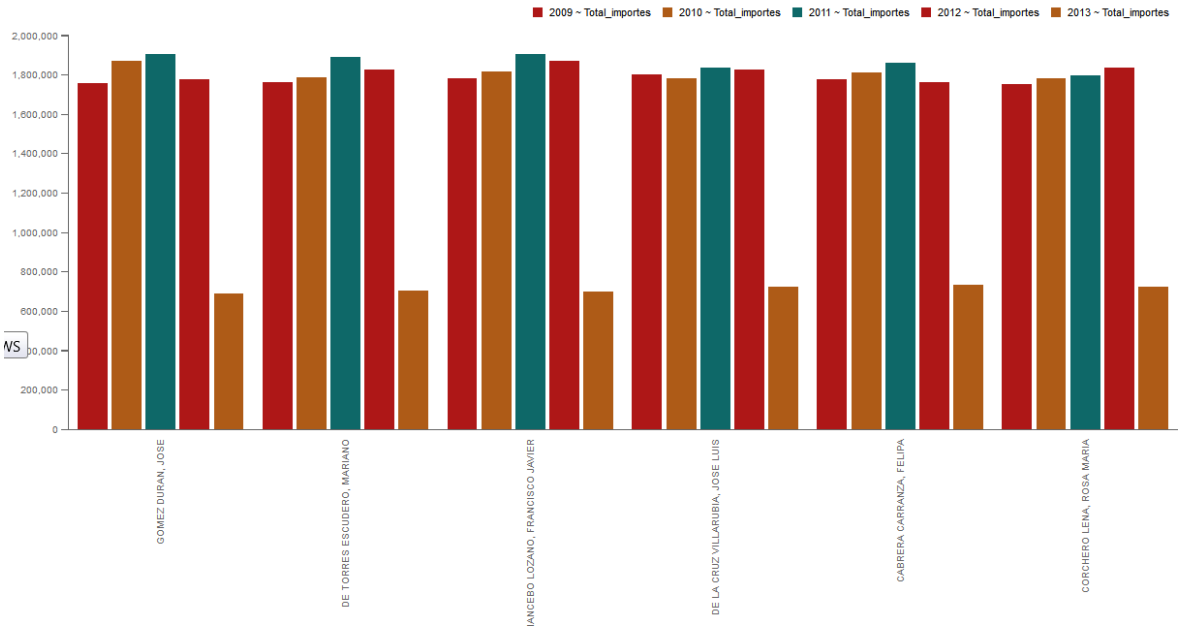

Ilustración 41 - Comisiones anuales por comercial

5.2.8. Margen anual para las familias de los productos.

Listado con el total del margen que cada familia ha tenido durante el período indicado.

	2009			2010			2011			2012		
JFamiliaArticulo	Total_importes	Total_costes	Total margen	Total_importes	Total_costes	Total margen	Total_importes	Total_costes	Total margen	Total_importes	Total_costes	Total margen
Electrodomésticos	2.864.737	1.280.095	1.584.642,00	2.873.598	1.295.607	1.577.991,00	3.049.907	1.367.634	1.682.273,00	3.002.242	1.356.024	1.646.218,00
Informática	1.316.944	595.753	721.191,00	1.419.076	632.787	786.289,00	1.470.081	670.086	799.995,00	1.374.908	615.931	758.977,00
Electrónica consumo	1.005.983	450.710	555.273,00	1.061.584	474.811	586.773,00	1.063.216	479.871	583.345,00	1.073.345	487.814	585.531,00
Casa y jardín	836.301	382.388	453.913,00	892.955	403.907	489.048,00	910.924	409.106	501.818,00	858.299	392.022	466.277,00
fotografía, video y óptica	758.573	342.139	416.434,00	742.002	335.053	406.949,00	762.565	344.802	417.763,00	741.863	332.388	409.475,00
Cónsolas y videojuegos	440.842	194.920	245.922,00	425.462	191.889	233.573,00	443.317	199.152	244.165,00	434.745	196.945	237.800,00
Grand Total	**7.223.380**	**3.246.005**	**3.977.375**	**7.414.677**	**3.334.054**	**4.080.623**	**7.700.010**	**3.470.651**	**4.229.359**	**7.485.402**	**3.381.124**	**4.104.278**

Ilustración 42 - Margen anual para las familias

En el informe anterior no se puede apreciar las diferencias en el margen por lo que requeriría un análisis más exhaustivo.
A simple vista según la gráfica que se muestra a continuación no se aprecia variación en el margen ya que parece estar entorno al 55 - 56%.

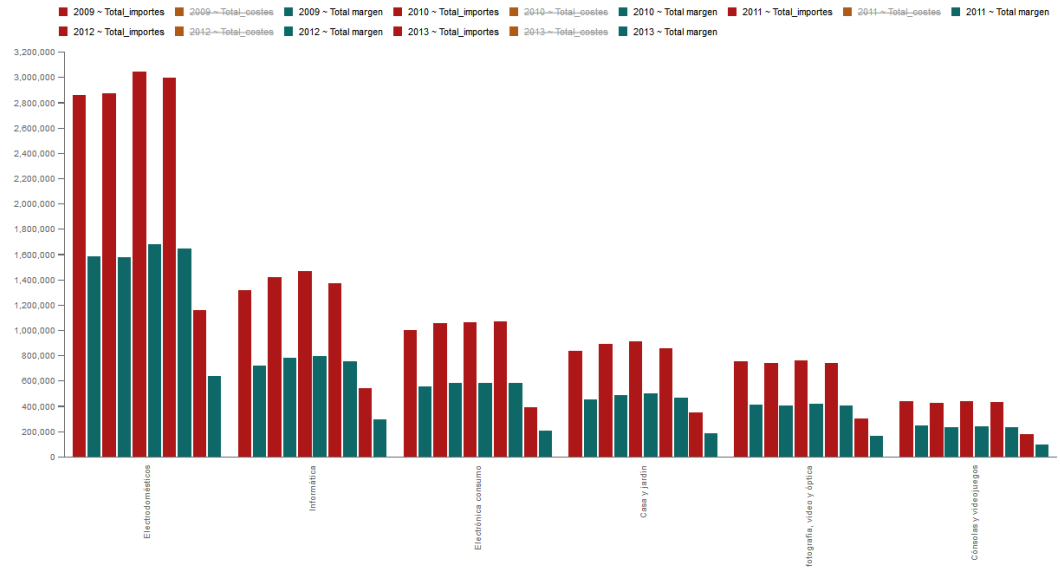

Ilustración 43 - Margen de las ventas

6. Conclusiones

La realización de este proyecto ha servido para entrar en contacto con el mundo del análisis de los datos y con la tecnología que se aplica. Descubrir las infinitas posibilidades del BI y ampliar los conocimientos en este campo son, sin duda, herramientas clave para el desarrollo profesional dentro de las tecnologías de la información.

Por otra parte, durante la realización del mismo, se han tenido que afrontar los problemas habituales cuando manejas datos desconocidos. Ha sido fundamental la parte previa de documentación y lecturas relativas a este tema.

Se considera que se han logrado los objetivos marcados por la empresa TOTSALES, ya que se ha conseguido dar respuesta de forma satisfactoria a las preguntas formuladas por el cliente en base a los datos aportados por él.

Por otro lado se recomienda a la empresa que para mejorar los análisis en un futuro se tomen medidas en lo que respecta a la coherencia y homogeneidad de sus datos, puesto que se han encontrado incongruencias, ya no sólo en el formato de los ficheros, donde se mezclaban poblaciones con información de clientes, sino principalmente en lo relativo a la información de las ventas, donde como ejemplo, se puede indicar que en el fichero de ventas.csv aparecen ventas por el código de familia en vez de por el código de artículo.

En todo lo relativo a la planificación, se han podido experimentar desviaciones en la parte de análisis de información de las fuentes de datos así como en la creación de los diferentes pasos seguidos en el ETL, debido a la complejidad de la estructura establecida en los ficheros. Para poder solventar estas complicaciones se le ha tenido que dedicar más horas de las planificadas en un primer momento. Este esfuerzo adicional ha permitido llegar a la fecha de entrega convenida.

La metodología seguida durante el proyecto ha sido la adecuada, se ha seguido con facilidad, salvo por las desviaciones encontradas que han hecho que la planificación tuviese que ser modificada de forma inmediata.

6.1. Líneas de trabajo futuro

Mejora en la consistencia de las fuentes de datos por parte del cliente:

Durante el análisis de las fuentes de datos se han descubierto ciertas incongruencias en lo que respecta a los datos de las ventas debido a un mal uso de las herramientas de ventas. Esto ha desencadenado en no poder realizar un análisis completo de los datos. Es por ello que se deberá instar a la empresa TOTSALES a mejorar el uso de sus herramientas así como a crear conciencia empresarial sobre la importancia de su buen uso para el análisis adecuado de los datos.

Nuevos informes:

Durante los procesos ETL en los que se ha cargado la información en el *data warehouse*, se han aislado las incongruencias para que no tenga impacto en los datos a analizar. Por lo que la empresa, antes de tomar cual decisión, debe tener en cuenta, estos datos para poderla cruzar con la información analizada en los informes vistos en este proyecto. Dichos datos incongruentes han sido identificados con un guion medio y la descripción "sin definir".

Instalación de sistema BI en la empresa TOTSALES:

Se le recomendará al cliente la incorporación de un sistema BI en su empresa para mejorar el análisis de sus datos, así como la toma de decisiones. Todo ello permitirá tener un control más detallado de lo que está pasando en la empresa y adelantarse a causalidades que pudieran suceder.

En el caso de que se instale la solución de Pentaho como herramienta de Business Analytics se recomienda:

- **Instalación de PostgreSQL como SGBD**: tanto para el repositorio como para el *data warehouse*, no sólo por el coste cero, sino por el rendimiento y escalabilidad que tiene en los sistemas de *data mining* y *data warehousing*.

- **Replantear arquitectura utilizada en el *data warehouse:*** el cliente necesitará plantearse sus necesidades para ver la conveniencia de la arquitectura a seguir. Por el contexto de la empresa, sería recomendable que siguiera con la planteada en este proyecto, aunque añadiendo un *staging area* junto con los diferentes *data marts* de los departamentos.

- **Automatización ETL:** creación de tareas para la ejecución del *job* que dispara la carga del *data warehouse.*

- **Configuración de servidor de correos para las notificaciones de errores:** en caso de que haya algún error en alguna de las transformaciones podrá ser notificado al administrador del sistema.

- **Uso de roles:** configurar los roles del servidor según las personas que tengan acceso al mismo, mejorando así la seguridad en el acceso de los datos.

Mejora de informes:
Se recomienda el uso de Pentaho Reporting Service para la creación de informes, ya que abre la posibilidad a la creación de informes más detallados así como la publicación automática de los mismos en el propio servidor de Pentaho BI, para el acceso del personal de la empresa.

En estos informes se podrá hacer uso de gráficas que podrán adaptar su diseño y personalizarse, tal y como se ha detallado en el epígrafe 3.5 de este documento.

7. Glosario

***Business Intelligence*:** en castellano, inteligencia empresarial. Conjunto de estrategias que permiten la toma de decisiones teniendo en cuenta los datos de una organización.

Corporate Information Factory: en castellano, factoría de la información coporativa. Consiste en una arquitectura en la que existe un *data warehouse* corporativo y unos *data marts* dependientes del mismo. El acceso a datos se realiza a los *data marts* o al ODS en caso de existir, pero nunca al propio *data warehouse.*

Data cleansing: en castellano, limpieza de datos. Proceso que permite identificar datos incompletos, incorrectos, inexactos para sustituir, modificar o eliminar estos datos incompletos evitando las posibles inconsistencias del sistema.

Data mart: versión especial del almacén de datos. Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.

Data mining: en castellano, minería de datos. Campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

Data warehouse: en castellano, almacén de datos. 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.* Proceso que consiste en extraer la información de las fuentes de datos, transformarla y cargarla en una base de datos.

Fuzzy Match: paso en una transformación del sistema ETL Kettle, el cual encuentra el igual del dato buscado siguiendo un algoritmo que detecta la duplicidad de los datos calculando la similitud de dos flujos de datos.

MDX: acrónimo en inglés de *MultiDimensional eXpressions*. Lenguaje de consulta para bases de datos multidimensionales sobre cubos OLAP.

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

ODS: acrónimo en inglés de *Operational Data Store*. Es un tipo de almacén de datos que proporciona sólo los últimos valores de los datos y no su historial.

SDLC: acrónimo en inglés de *System Development Life Cycle.* Define el ciclo de vida del desarrollo de un Proyecto.

SGBD: acrónimo de Sistema Gestor de Base de Datos. Conjunto de programas que permiten el almacenamiento, modificación y extracción de la información en una base de datos, además de proporcionar herramientas para añadir, borrar, modificar y analizar datos.

Staging area: en castellano, área de pruebas. Área intermedia de almacenamiento de datos utilizada para el procesamiento de los mismos durante procesos de extracción, transformación y carga. Esta área se encuentra entre las fuentes de los datos y su destino, que a menudo son almacenes de datos, *data marts* u otros repositorios de datos.

8. Bibliografía

8.1. Apuntes

Rodríguez, José Ramón. "Ciclo de vida de un proyecto".
La gestión de proyectos. Conceptos básicos. Barcelona: FUOC (PID_00153562).

8.2. Libros

Curto Díaz, Josep; Conesa Caralt, Jordi (2010)**.** *Introducción al Business Intelligence (1ª ed.)*. Barcelona: editorial UOC.

Kimball, Ralph; Ross, Margy (2002). *The Data Warehouse Toolkit* (2ª ed.). United States of America: John Wiley and Sons, Inc.

8.3. Web

Comparativa de las herramientas de análisis a utilizar:
(04/2015)
http://community.pentaho.com/
https://community.jaspersoft.com/
http://www.actuate.com/products/

Comparativa de SGBD:
(03/2015)
http://es.slideshare.net/jazpekcobain/cuadro-comparativ-35729496

Información sobre tipos de esquemas:
(04/2015)
http://searchbusinessintelligence.techtarget.in/answer/Star-schema-vs-snowflake-schema-Which-is-better

Información sobre MDX:
(05/2015)
http://blog.crossjoin.co.uk/2013/02/09/topcounts-with-ties-in-mdx/
http://www.mdxpert.com/Functions/FunctionList.aspx?c=All%20MDX

Información sobre OLAP:
(05/2015)
http://www.informationbuilders.com/olap-online-analytical-processing-tools

MOOC:
(01/2015) MOOC UOC - Introducción al Business Intelligence - MiriadaX