Construcción y explotación de un almacén de datos para el análisis de ventas de una cadena de supermercados (Cadena GLDP)

Memoria del proyecto

Fecha entrega: 07 / 01 / 2013
Titulación: ITIG
Alumno: Ignacio Fernández Sánchez
Consultor: Carles Llorach Rius
Correo-e: igfersan@gmail.com

CONTROL DE CAMBIOS

EDICIÓN	APARTADO CAMBIO	DESCRIPCION	FECHA CAMBIO
1.00	Documento Original	Primera versión del documento.	30-12-2012
2.00	Reestructuración de puntos	Ante el nuevo modelo de memoria a entregar recibido, se cambia la estructura y contenido del documento	03-01-2012

1. Resumen y palabras claves

En este capítulo se describe brevemente el alcance del proyecto y las palabras claves que lo definen:

1.1. Resumen

El presente documento, contiene la memoria del Trabajo Final de Carrera (TFC) del área de almacenes de datos, perteneciente a la titulación de Ingeniería Técnica en Informática de Gestión (ITIG) de la Universidad Oberta de Catalunya (UOC), en el cual se pretende recoger los aspectos más relevantes del trabajo realizado sintetizando en la medida de lo posible la documentación aportada a lo largo del proyecto.

El objetivo principal de este TFC es la construcción de la base de datos de un Data Warehouse (DW) y la utilización de las técnicas y herramientas existentes en la actualidad para llevarlo a cabo. La implementación de dicho modelo debe de estar orientado a un almacén de datos físico ROLAP, el cual presentará los elementos propios de dicho modelo (dimensiones, atributos y hechos) y las características típicas que lo definen (desnormalización de tablas, inclusión de información agregada, historificación de la información, etc).

Para adquirir estas capacidades se pretende desarrollar un caso práctico que permita la aplicación de los conceptos estudiados en un escenario concreto. Este escenario es aportado mediante el documento de propuesta, o enunciado, en el que se explica la problemática planteada y los objetivos que se pretenden cumplir con la construcción del almacén de datos. Además junto con el enunciado se reciben los datos necesarios se servirán para poblar dicho almacén, los cuales serán necesario tratar y manipular, para adaptarlos al modelo multidimensional que dotará al producto de las capacidades de análisis requeridas.

Finalmente, junto con el almacén de datos y el desarrollo de la ETL que permita la carga de datos recibidos, se espera que el producto desarrollado incluya un conjunto de informes iniciales que permitan de manera rápida y sencilla analizar y visualizar los principales indicadores de negocio demandados en el enunciado.

1.2. Palabras Clave

Business Intelligence ,Data Warehouse, almacén de datos, OLAP, cubo OLAP, dimensión, modelo dimensional, tabla de dimensiones, tabla de hechos, ETL, Pentaho BI Platform and Server, Pentaho Reporting, Pentaho Data Integration, MySQL,

Índice de contenidos

Índice de figuras

Índice de tablas

2. Introducción

2.1. Justificación del proyecto (idoneidad)

En este capítulo se explica los motivos por los cuales nace la necesidad de desarrollar el proyecto y una breve descripción del mismo.

2.1.1 ¿Por qué el proyecto?

El proyecto surge ante una necesidad del Grupo Líder en Distribución de Proximidad (GLDP), que ante la actual situación de crisis y la constante disminución de las ventas, decide invertir en una solución BI que le proporcione conocimiento del comportamiento de su negocio, a través del análisis de sus ventas, y le permita de esta manera, estar preparado para tomar decisiones adecuadas y mejorar la tendencia negativa que sufre en los últimos años.

El problema con el que se encuentra el grupo es que, debido a la gestión distribuida de las compras de cada establecimiento y debido a la consolidación trimestral de ventas, no tiene acceso a todos los datos que necesita para poner en marcha las campañas de marketing oportunas que deben conducirlo a un incremento de ventas en los establecimientos y periodos más flojos en ventas.

2.1.2 Descripción del proyecto

La solución a implementar consistirá en un sistema de carga automatizada de información heterogénea proporcionada por el propio cliente, la cual quedará integrada en un repositorio único y centralizado, que permitirá la publicación de informes complejos, accesibles para diferentes usuarios, y le dotará de capacidades para la generación propia de nuevos informes de una manera fácil e intuitiva.

El sistema se soportará sobre productos de licenciamiento open-source, líderes en el mercado para este tipo de sistemas, los cuales quedarán integrados mediante desarrollos específicos destinados a optimizar los distintos procesos definidos, incrementando en gran medida el grado de automatismo de la solución final y permitiendo un notable grado de control sobre los procesos con mínimo coste de mantenimiento.

2.2. Objetivos del proyecto

En este capítulo, se describen los objetivos del proyecto atendiendo a diferentes puntos de vista como son: los objetivos de la asignatura, y los propios del Trabajo fin de carrera.

2.2.1 Objetivos de la asignatura

El objetivo principal del proyecto, desde el punto de vista de la asignatura, es adquirir experiencia en el diseño, construcción y explotación de un almacén de datos a partir de la información disponible en una base de datos transaccional.

Al alumno le será requerido un análisis de técnicas existentes para proyectar la base de datos de un DW. El diseño deberá estar orientado a un almacén de datos físico ROLAP, creando las dimensiones, atributos y hechos necesarios. Han de considerarse factores del tipo: desnormalización de tablas, inclusión de información agregada, historificación de la información, etc.

La descripción del trabajo a realizar es la siguiente:

- **Plan de Trabajo:** Lo primero que se pide al estudiante es la creación de un plan de trabajo donde se deberá indicar con un cierto nivel de detalle las tareas que se deberán realizar, así como un análisis de riesgos y un diagrama de Gantt.
- **Análisis y Diseño**: En este segundo apartado del TFC, se deberá de realizar el análisis de la problemática propuesta en el enunciado y se deberá de realizar el diseño de la solución generando el modelo multidimensional de datos, así como toda la información necesaria para poder conseguir los objetivos solicitados.
- **Implementación:** En esta fase se deberá crear la BDD tal y como se ha diseñado en el apartado anterior, y también se deberá cargar los datos que se disponen. En este proceso de carga deberá haber un proceso de transformación y adecuación a lo que se solicita. Una vez cargada la BDD se deberá generar los informes necesarios para poder resolver lo que se solicita. Estos informes se han de generar con la herramienta indicada por el consultor.
- **Memoria y presentación virtual:** Al final del semestre, y una vez entregadas las fases anteriores, se deberá crear la memoria del TFC donde se explicará el trabajo realizado (habitualmente será una suma y adecuación de los entregables anteriores). También se deberá crear una presentación donde se explique el trabajo realizado. Esta presentación será la que permitirá la defensa el trabajo delante del tribunal.

2.2.2 Objetivos del TFC

En este apartado se van a diferenciar entre los objetivos generales que deben de poseer todos los proyectos enmarcados dentro del mundo de las tecnologías BI, y los objetivos específicos requeridos en la construcción del almacén de datos que nos compete según el enunciado.

2.2.2.1. Generales

Como objetivos generales nos encontramos los siguientes:

- **Procesos de carga:** Elaborar un conjunto de procesos que se encargarán de manipular información heterogénea proveniente de diferentes orígenes, transformarla y enriquecerla, para por último pasar a cargarla en un modelo de datos propio del aplicativo.
- **Almacén de datos**: Elaborar un almacén de datos que contenga un modelo multidimensional que integre la información de los diferentes orígenes, y que sirva como repositorio de información único sobre el cual los usuarios finales realicen el análisis de los datos.

- **Sistema de *reporting***: Se dotará al sistema de una interfaz definida exclusivamente en términos de negocio que permitirá a los usuarios explotar el almacén de datos del aplicativo, sin requerir unos conocimientos técnicos específicos por parte de los mismos.

2.2.2.2. Específicos

El objetivo específico del TFC es dotar a al Grupo Líder en Distribución de Proximidad (GLDP) de un sistema BI que le permita conocer mejor el comportamiento de sus ventas para cambiar la evolución negativa de los últimos años. Para se debe implementar un almacén de datos, mediante el cual se pueda obtener, como mínimo, la siguiente información:

- Total de ventas y margen neto del grupo
- Importe medio de compra por socio y establecimiento
- % de ventas respecto el total del grupo (por establecimiento y por demarcación)
- Precios máximos y mínimos por tipo de establecimiento y características de producto.
- Ranking de establecimientos por número de ventas y volumen total
- "*Top ten" de productos
- % de ventas de "marcas blancas" por habitantes
- Categorización de Clientes A/B/C
- Análisis de compra por impulso
- Distribución semanal y estacionalidad de ventas

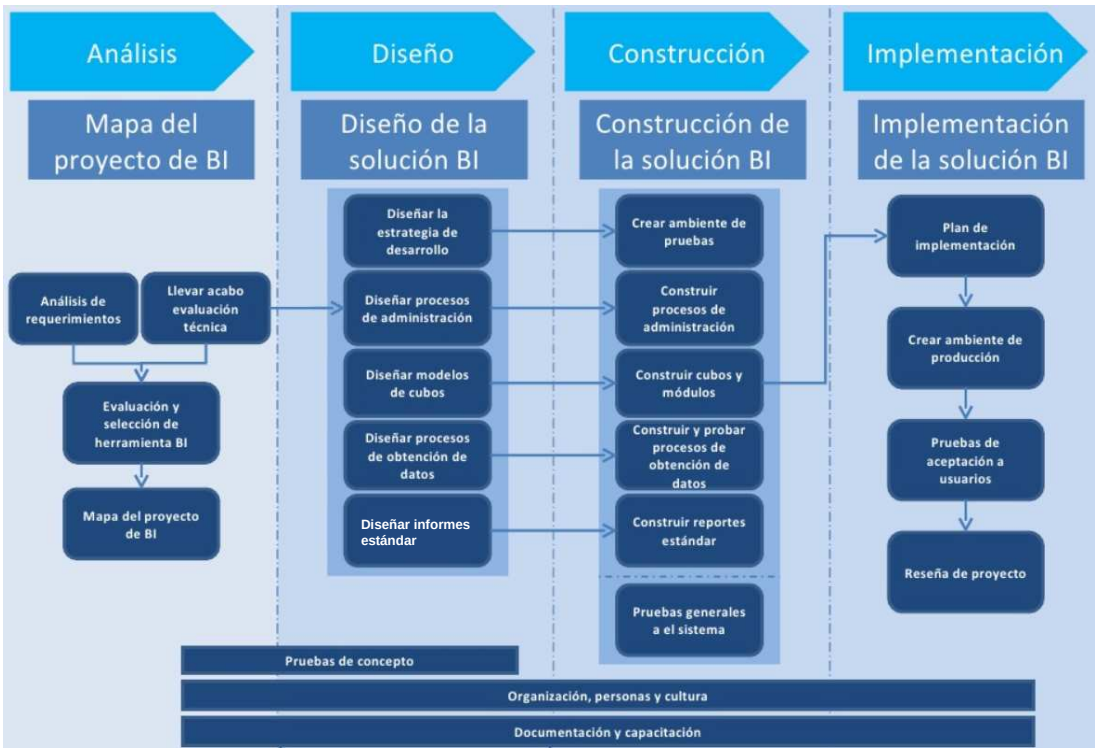
La información se proporcionará dentro de una temporalidad a nivel de mes y año.
Se podrá consultar de forma agregada por demarcación territorial, tipo de establecimiento y familia de productos.

Adicionalmente, se proporcionará un conjunto predefinido de informes que muestren la información solicitada y cualquier otra que pueda ser útil para GLDP.

2.3. Enfoque y método seguido

En general, la planificación propuesta para las entregas de las PECs (comentado en el apartado **2.2.1 Objetivos de la asignatura)** junto con los contenidos mínimos de cada una de ellas, han hecho que se siga la metodología estándar que rigen las implementaciones de sistemas BI, en lo que se refiere principalmente a las fases de Análisis, Diseño y Construcción.

Figura 1: Metodología proyectos BI



La figura anterior muestra cada una de las actividades que se presentan las fases principales mencionadas en el párrafo anterior, la cuales pueden variar dependiendo del autor que se consulten. Las que han tenido más peso en el proyecto, y que han sido llevadas a cabo son las siguientes:

- **Análisis:**
  - **Análisis de requerimientos:** Actividad de recopilar los requisitos del cliente, basados en necesidades de negocio y que requiere una adaptación a conceptos más técnicos
  - **Mapa del proyecto:** Planificación y esquematización general del proyecto que sirve para hacerse una idea del alcance que tendrá el producto a construir
  - **Evaluación y selección de herramientas BI:** Es importantísimo tener una idea clara de las herramientas BI a utilizar, ya que aunque existen multitud en el mercado y con características semejantes, la solución técnica final puede variar entre unas y otras.
  - **Evaluación técnica:** Donde se verá si **es factible** con la información que se le proporciona al sistema **dar la solución requerida**. Además, se tendrán en cuenta detalles como el hardware y elementos técnicos dependientes del entorno de implantación del sistema.
- **Diseño:**
  - **Diseño de modelos de cubos:** Consistente en el diseño del modelo multidimensional, que permitirá la construcción de cubos OLAP que satisfagan las necesidades de información del usuario. También hay que tener en cuenta cómo será su implementación física, ya que puede depender según el sistema gestor de BBDD que se elija.
  - **Diseño de procesos de obtención de datos:** Una vez definido el modelo de datos, hay que definir cómo serán los procesos de transformación y carga (ETL) que lo poblarán.
  - **Diseño de informes estándar:** Consiste en tener una idea clara de cuáles van a ser el conjunto inicial de informes que cumplen las principales necesidades de información del usuario. Este conjunto de informes servirá como punto de partida, para que trabajando conjuntamente con el usuario, el sistema evolucione y aporte así más valor a su negocio.

- **Construcción:**
  - **Construcción cubos y módulos:** Consistente en transformar al modelo físico, el diseño de la fase anterior y su especificación técnica para que los motores OLAP puedan utilizarlos.
  - **Construcción y pruebas de procesos de obtención de datos:** Preparación de la ETL. Este punto incluye además las pruebas que verifican su correcto funcionamiento.
  - **Construcción de informes estándar:** Una vez que los datos están en el modelo, es posible generar los informes estándar definidos en el apartado de diseño.
  - **Pruebas generales al sistema:** Si bien las pruebas deben realizarse desde y en cada uno de los componentes que forman el sistema, una vez que se disponen de los informes estándar se entra en una etapa de pruebas generales, en las que se evalúa el sistema como un todo para ver si está listo para entregarlo al usuario. Esta fase incluiría también tareas de **ajuste de rendimiento (*Performance Tuning*),** que influirán tanto a los procesos de carga como a las consultas de los informes.

2.4. Propuesta de actividades y cronograma

En este capítulo se detallan principalmente el conjunto de tareas a realizar durante la ejecución del proyecto, la estimación de tiempos que cada una ellas requiere y los hitos más importantes que se encuentran en el proyecto. Finalmente se pasar a mostrar el diagrama de Gantt que recoge todos los puntos anteriores.

2.4.1 Relación de actividades

A continuación se describen las actividades principales que conforman el proyecto:

- **Plan de Trabajo:** Consiste en la elaboración del documento en el que se establece un plan de trabajo y el detalle de actividades junto con un análisis a alto nivel de la solución. Esta primera actividad exige un gran trabajo de búsqueda de información; por una parte para manejar los principales conceptos manejados en el desarrollo de sistemas relacionados con los almacenes de datos y así comprender el alcance real del proyecto, y por otra sobre las herramientas de BI a utilizar en la solución aportada.
- **Análisis y Diseño**: En este apartado se llevarán **dos actividades en forma paralela**. Por una parte, el análisis y el diseño de la solución a implantar, la cual requiere la comprensión exacta de las expectativas del cliente y las posibilidades de cumplirlas, y por otra parte, una familiarización con el uso de herramientas que serán utilizadas en la siguiente fase. Esta preparación a su vez tiene dos funciones: la primera, estar preparados para la fase de implementación lo cual es muy importante dado los tiempos ajustados y el ritmo al que se plantea proyecto/asignatura, y por otra, la función de entender las capacidades de una herramienta y lo que se va a poder ofrecer con la misma, para así de esta manera, poder afinar más los requisitos y la solución final a implementar.
- **Implementación:** Esta fase consiste en la creación del almacén de información propio del sistema, la construcción de la ETL encargada de cargar los datos, y la implantación de la herramienta de análisis junto con el conjunto de informes predefinidos y los *metadatas* necesarios para poder aportar al cliente una solución que le permita generar sus propios informes.
- **Memoria y presentación virtual:** Esta fase consiste en la aglutinación del trabajo realizado en cada una de las fases anteriores, que junto a la recopilación del conocimiento adquirido durante la elaboración del proyecto, permitirán la generación de un documento de memoria y una presentación virtual del trabajo

2.4.2 Estimación de tiempo

La planificación se ha pensado teniendo en cuenta un ritmo de trabajo de unas 3 horas de media al día, pudiendo disponer del siguiente horario aproximado:

- **Lunes a viernes:** de 20:30 a 23:00
- **Sábados y Domingos:** de 10:00 a 13:00 y de 17:00 a 20:00

La siguiente tabla muestra el detalle de tareas propuestas y el tiempo dedicado a cada una de ellas en jornadas que no se corresponden con el calendario estándar de 8 horas:

Tabla 1: Detalle de tareas y duranción en jornadas

TAREA	INICIO	FIN	JORNADAS
Plan de trabajo y Análisis preliminar	19-sep-12	2-oct-12	14
Descarga de documentación del aula	19-sep-12	19-sep-12	1
Lectura de documentación	20-sep-12	20-sep-12	1
Estudio del enunciado	21-sep-12	21-sep-12	1
Búsqueda de información	22-sep-12	26-sep-12	5
Búsqueda y análisis herramientas BI	22-sep-12	23-sep-12	2
Búsqueda información conceptos DW	24-sep-12	24-sep-12	1
Búsqueda información proyectos BI	25-sep-12	26-sep-12	2
Análisis de los datos fuentes	27-sep-12	27-sep-12	1
Descarga y familiarización OpenProj	28-sep-12	28-sep-12	1
Lectura ejemplos PEC1	29-sep-12	29-sep-12	1
Preparación del documento PEC1	30-sep-12	2-oct-12	3
Análisis y diseño	3-oct-12	6-nov-12	35
Revisión análisis de requerimientos	3-oct-12	7-oct-12	5
Correlación información aportada-solicitada	8-oct-12	9-oct-12	2
Preguntas al consultor	9-oct-12	9-oct-12	0
Búsqueda de información modelos BI	10-oct-12	13-oct-12	4
Modelo conceptual	14-oct-12	17-oct-12	4
Diseño de la BD / Diagrama ER	18-oct-12	22-oct-12	5
Modelo multidimensional detallado	23-oct-12	28-oct-12	6
Diseño de cadenas y procesos	29-oct-12	3-nov-12	6
Familiarización Herramientas	9-oct-12	2-nov-12	25
Familiarización maquina y entorno linux	9-oct-12	12-oct-12	4
Familiarización PDI	13-oct-12	19-oct-12	7
Familiarización MySQL	20-oct-12	21-oct-12	2
Instalación Pentaho BI Platform	22-oct-12	24-oct-12	3
Familiarización Pentaho BI Platform	25-oct-12	2-nov-12	9
Realización documentación PEC2	4-nov-12	6-nov-12	3
Implementación	7-nov-12	19-dic-12	43
Creación base de datos	7-nov-12	9-nov-12	3
Análisis calidad del dato fuentes	10-nov-12	11-nov-12	2
Preguntas al consultor	11-nov-12	11-nov-12	0
Creación de procesos ETL	12-nov-12	21-nov-12	10
Creación de script cadenas	22-nov-12	24-nov-12	3
Creación del metadata BI	25-nov-12	29-nov-12	5
Construcción Informes y Dashboard	30-nov-12	9-dic-12	10
Análisis de la información	10-dic-12	11-dic-12	2
Reajustes y modificaciones	12-dic-12	15-dic-12	4
Realización documentación PEC3	16-dic-12	18-dic-12	3
Preparación entrega	19-dic-12	19-dic-12	1

TAREA	INICIO	FIN	JORNADAS
Memoria y Presentación Virtual	20-dic-12	7-ene-13	19
Preparación Memoria	20-dic-12	26-dic-12	7
Familiarización herramienta presentación	27-dic-12	27-dic-12	1
Elaboración presentación virtual	28-dic-12	7-ene-13	11

2.4.3 Hitos a cumplir

En este apartado se recogen los hitos principales a cumplir, los cuales son estipulados por el propio ritmo de la asignatura TFC- Almacenes de datos y la evaluación continua propuesta.

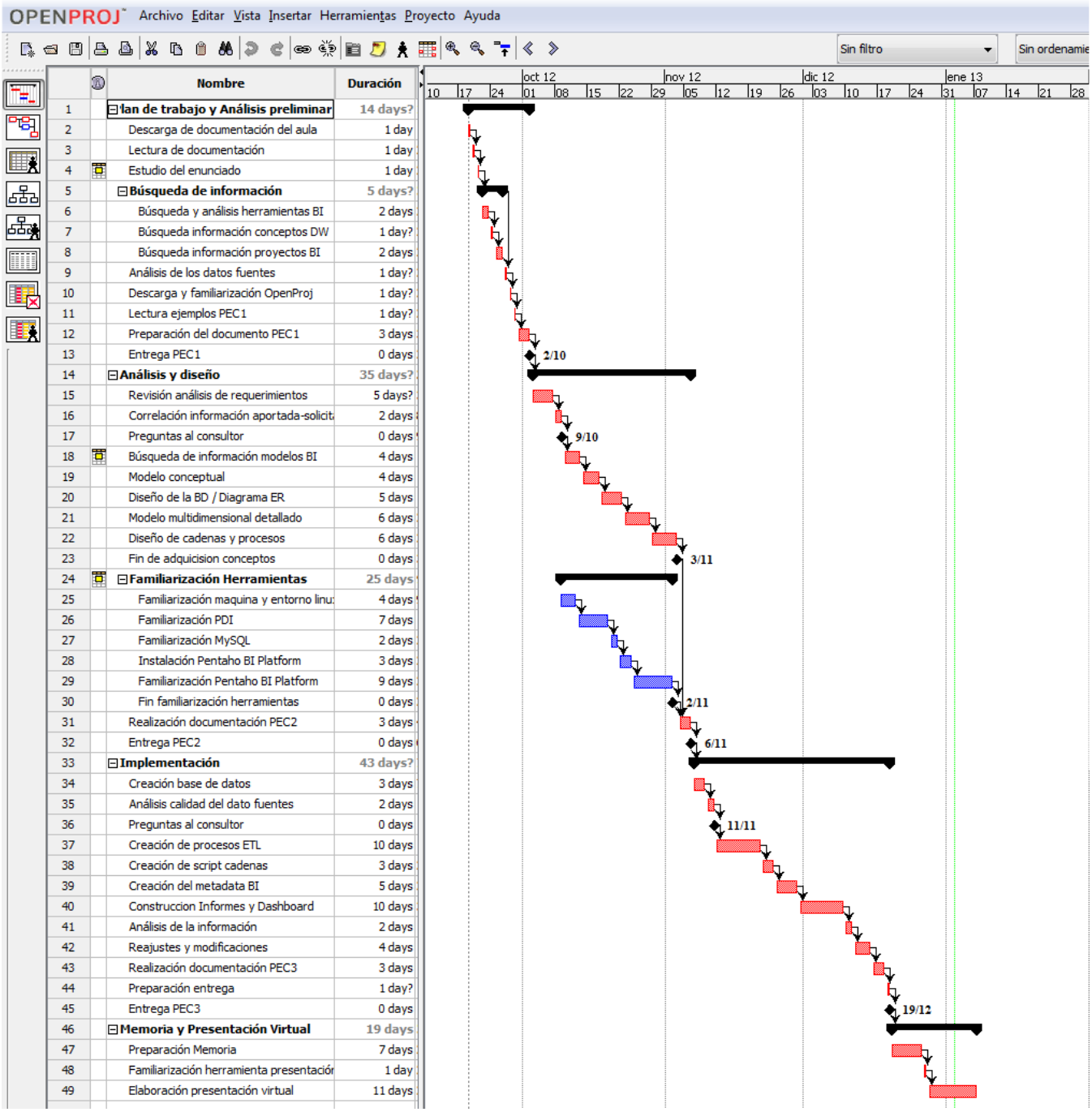
Tabla 2: Hitos a cumplir

HITO	INICIO	FIN
PEC1	19/09/2012	02/10/2012
PEC2	03/10/2012	06/11/2012
PEC3	07/11/2012	19/12/2012
Entrega Final	20/12/2012	07/01/2013

2.4.4 Diagrama de Gantt

La siguiente figura muestra el diagrama de Gantt para las fases “Análisis y diseño”, “Implementación” y “Memoria y Presentación Virtual”:

Figura 2: Diagrama de Gantt



El diagrama completo queda recogido en el siguiente archivo adjunto:

2.4.5 Planificación inicial vs planificación final

La coincidencia de los hitos intermedios que cumplir con las entregas obligatorias de las distintas PECs de la asignatura, han hecho posible que de una manera general las planificaciones inicial y final coincidieran en el tiempo.

Si bien es cierto, que en la fase de implementación, se hubieran querido abordar multitud de nuevas funcionalidades y mejoras, que fueron estimadas bajo una planificación inicial bastante optimista, y que debido al tiempo real con el que se contaba para la realización de dicha fase, han tenido que quedarse fuera de la entrega.

El modo de trabajo ha consistido en preparar un conjunto de funcionalidades e informes que cumplan sobradamente los requisitos iniciales del proyecto, e ir insertando mejoras en la medida que la planificación y las fechas de entrega lo permitían. Esto fue tenido en cuenta, aunque no tan severamente en una de las tareas reflejadas en la planificación inicial denominada “Reajustes y modificaciones” la cual contaba con la duración de 4 jornadas.

Este conjunto de funcionalidades que han quedado fuera y que se consideran importantes incluir en proyecto de esta tipología, serán consideradas como mejoras a la versión inicial, quedando su detalle reflejado en el apartado ¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia. de este mismo documento.

2.4.6 Análisis de riesgos

En este apartado se definen los posibles problemas que pueden acontecer durante la ejecución del proyecto, y un posible plan de contingencia en caso de su aparición.

Como posibles riesgos del proyecto se han planteado los siguientes:

2.4.6.1. Problemas hardware / software

Ningún proyecto queda fuera del alcance de este riesgo, en el que un fallo en algún elemento del hardware o del software involucrado, puede penalizar en las planificaciones llegando incluso a suponer su fracaso.

Como plan de contingencia, se realizarán copias de seguridad diarias de la documentación elaborada y desarrollo. Estas copias de seguridad se realizarán sobre soporte físico (disco externo o *pendrive*), o en *“la nube”* utilizando un espacio de *DropBox* (www.dropbox.com)

La penalización en el proyecto debería de ser menor cuanta más cantidad de software sea respaldado. Además, tratándose de un entorno de desarrollo virtual se intentará semanalmente realizar respaldo de los discos físicos de la máquina virtual, lo cual agilizaría en gran medida la restauración del sistema en caso de fallo grave del sistema host que la sustenta.

2.4.6.2. Enfermedades y aumento considerable actividad laboral

Estos males son ajenos a cualquier planificación ya que pueden darse en cualquier proyecto y en diferente medida.

La única manera de actuar en este caso, ya que no hay posibilidad de demora en los plazos (sobre todo en los finales), es informar cuanto antes al consultor, por ver la posibilidad de una pequeña demora si el problema no es muy grave, y realizar un sobreesfuerzo en los días que le siguen. También se cuenta con la posibilidad de disponer de días de vacaciones que han sido reservados para si llegase la necesidad utilizarlos a tal efecto.

2.4.6.3. Familiarización con las herramientas

El proyecto exige la utilización de multitud de herramientas de las cuales no se conoce de manera exacta su curva de aprendizaje. Esto puede ocasionar una demora en la parte final del proyecto (implementación), demora que podría desembocar en el fracaso del mismo.

Para minimizar riesgos, se intentará solapar actividades de investigación con las del propio análisis de la solución. Esta medida tiene dos ventajas inmediatas: por una parte la comentada anteriormente de ir ganando tiempo y estar preparado para la fase de implementación, y por otra, no comprometerse con funcionalidades que no puedan ser soportada por las herramienta o en el tiempo que exige la finalización del proyecto.

Además, como ocurría con el riesgo del apartado anterior, existe la posibilidad de disponer de días de vacaciones reservados para la realización de sobreesfuerzos, que aumentarán las horas dedicadas al proyecto para intentar recuperar la planificación del mismo.

2.5. Productos obtenidos

Este capítulo recoge el conjunto de entregables que han sido generados a lo largo de la ejecución del TFC, los cuales se dividen en los siguientes apartados:

- **Documento PEC1:** Documento con el plan de trabajo y el detalle de actividades, junto con un análisis a alto nivel de la solución.
- **Documento PEC2:** Contiene un análisis exhaustivo de los requerimientos del cliente y de la información proporcionada (plasmados como requisitos funcionales y no funcionales), el diseño del modelo de datos (conceptual y físico) y el diseño de procesos ETL
- **Documento PEC3:** Explicación de cómo se ha llevado a cabo la fase de implementación del producto, captura y explicación de los informes generados y comentarios relevantes sobre el desarrollo llevado a cabo.
- **Documento de Memoria del Proyecto** Documento de síntesis de las actividades más relevantes que se han llevado a cabo durante la ejecución del TFC.
- **Presentación virtual:** vídeo con la captura del escritorio de la máquina de trabajo, que ofrece una perspectiva general del TFC y permite al tribunal formular al alumno las preguntas que consideren oportunas.
- **Máquina Virtual:** contiene la implementación del almacén de datos requerido:
  - **BBDD:** Base de datos con el modelo multidimensional diseñado
  - **Procesos ETL:** Procesos de carga de información de cliente e información inicial que pueblan la BBDD del almacén de datos.

- **Informes predefinidos:** Conjunto de informes que contienen los indicadores de negocio más relevantes para el cliente que fueron solicitados mediante el enunciado.
- **Procesos de automatización de las cargas:** Procesos que gestionan las ejecuciones de los procesos de carga, y que permiten su automatización y control mediante la recepción de email con el resultado de los mismos.
- **Procesos de generación dinámica de informes:** Procesos que permiten la planificación y generación de informes de manera dinámica.

2.6. Contenido de los siguientes capítulos

En este apartado se describe los siguientes capítulos en los que está estructurado el documento:

- **Análisis:** Recoge las principales tareas que se llevaron a cabo en cuanto al análisis del proyecto, mostrando el diagrama de casos de uso, un análisis de la información recibida y de la calidad que ésta presenta y el modelo conceptual en el que se basa el almacén de datos.

- **Diseño:** En este capítulo se describen las principales características que definen el proyecto en cuanto al diseño del mismo, reflejando los apartados de arquitecturas HW y SW

- **Desarrollo:** Este capítulo describen los aspectos más relevantes sobre las fases de desarrollo y la solución final que se ha implementado.

- **Informes realizados:** Recoge una explicación inicial del concepto “informes predefinidos”, para a continuación pasar a mostrar una captura de pantalla y una explicación de cada uno de ellos.

- **Conclusiones:** Apartado con las principales conclusiones obtenidas tras la finalización del proyecto.

- **Líneas de evolución futuras:** Nuevas funcionalidades y mejoras que serán candidatas a ser incluidas en el sistema una vez finalizada esta primera fase de desarrollo.

3. Análisis

Este capítulo recoge las principales tareas que se llevaron a cabo en cuanto al análisis del proyecto.

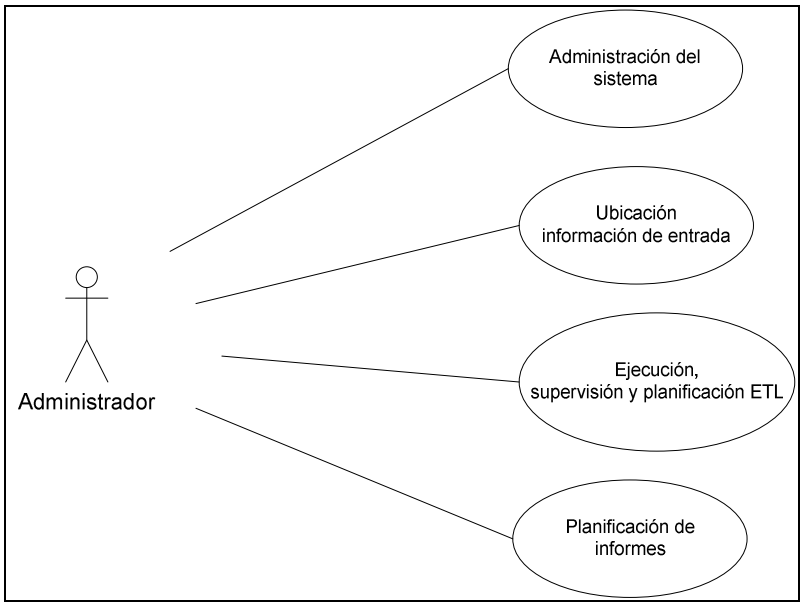
3.1. Diagrama de casos de uso

En el sistema se pueden distinguir dos actores o de perfiles diferentes, los cuales llevarán a cabo un conjunto de actividades bien definidas:

3.1.1 Perfil "Administrador"

La siguiente figura muestra el conjunto de actividades que realizará este perfil en el sistema:

Figura 3: Casos de uso "Administrador"

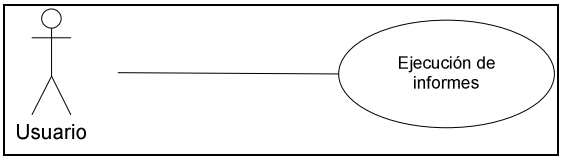


- **Administración del sistema:** Este usuario se encargará de administrar el sistema operativo, la herramienta de análisis y el servidor de base de datos, creando los usuarios y adjudicando los permisos que se requieran
- **Ubicación de información:** Ubicará la información de entrada al sistema en los directorios definidos a tal efecto para sea procesada y consolidada en el almacén de datos por la ETL.
- **Ejecución, supervisión y planificación ETL:** El usuario será el encargado de ejecutar de manera manual o planificada (con las herramientas que se aportan en la solución) los procesos ETL. También supervisará el correcto funcionamiento y será candidato a ser el receptor de los correos con el resultado de la ejecución del sistema.
- **Planificación de:** Configurará los destinatarios de la distribución de informes y la ejecutará, como ocurría en el caso anterior, de manera manual o planificada

3.1.2 Perfil “Usuario”

La siguiente figura muestra el conjunto de actividades que realizará este perfil en el sistema:

Figura 4: Casos de uso “Usuario”



- **Ejecución de informes:** En primera instancia este usuario sólo tendrá como caso de uso la ejecución de los informes predefinidos utilizando los parámetros que considere oportuno para la ejecución de los mismos. Es previsible que en futuras entregas los usuarios puedan elaborarse sus propios informes.

3.2. Análisis inicial de información

En este capítulo se recoge el análisis inicial de la información de entrada al sistema, la información requerida por el cliente y la correspondencia entre la información proporcionada y requerida con el objetivo de advertir riesgos en el compromiso de datos que no puedan suministrarse.

3.2.1 Análisis de la información de entrada al sistema

El sistema será nutrido por cuatro grupos de ficheros diferentes, los cuales proporcionaran principalmente información sobre su actividad y sobre información externa relacionada con la distribución de la población en Cataluña.

Los ficheros recibidos deben de cumplir tanto la estructura que se recoge a continuación, como una nomenclatura fija que permita al sistema su tratamiento automático.

Dado que el nombre de los ficheros del cliente será utilizado en las transformaciones al proporcionar información complementaria sobre los mismos, será imprescindible que el cliente nos proporciones siempre los mismos nombres de ficheros y con el mismo formato.

Un incumplimiento de los requisitos de interfaz de entrada implicaría que el sistema no procesara correctamente la información.

A continuación se procederá a realizar un análisis detallado de la información que se presenta en cada uno de los ficheros que serán tratados por el sistema.

3.2.1.1. Catálogo de productos

Se trata de un archivo CSV por cada uno de los años en los que se registran productos en el sistema origen, con todos los productos ofrecidos en alguno de sus puntos de venta. El fichero incluirá una cabecera con la descripción de los nombres de los campos

Cada uno de los registros del fichero CSV contiene la siguiente información:

Tabla 3: Productos

CAMPO	DESCRIPCIÓN	TIPO
Id_producte	Código del producto	CHAR
Nom	Nombre del producto	CHAR
Preu_venda	Precio de venta del producto	NUMBER
Proveïdor	Proveedor del producto	CHAR
Tipologia	Clasificación del producto	CHAR
IVA	Impuesto aplicado al producto	NUMBER
Preu_cost	Precio de coste del producto	NUMBER
Codi_barres	Código de barras del producto	NUMBER
Venda_impuls	Indicador de venta por impulso	BOOLEANS

El formato del nombre de los archivos será el siguiente:

- **Ficheros de productos:** Productes XXXX.csv
  - XXXX**:** Año al que pertenecen los registros de productos del fichero.

3.2.1.2. Establecimientos

Son proporcionados mediante un archivo Excel con la ficha descriptiva de cada centro comercial. Todos los establecimientos serán indicados en la primera pestaña de la hoja Excel proporcionada. Además, todos los establecimientos serán descritos con el mismo número de campos y en el mismo orden.

El fichero Excel de establecimientos contiene la siguiente información para cada establecimiento:

Tabla 4: Establecimientos

CAMPO	DESCRIPCIÓN	TIPO
Establiment	Nombre del establecimiento	CHAR
Superfície	Metros cuadrados del establecimiento	CHAR
Dia del soci	Día del socio del establecimiento	CHAR
Cost fix mensual	Coste del mantenimiento del establecimiento	NUMBER
Tipologia	Clasificación del establecimiento	CHAR

3.2.1.3. Ventas

Un archivo de Microsoft Access por cada establecimiento, con la relación de ventas y el detalle de las mismas, distribuida en dos tablas “Venda” y “Detall_venda” respectivamente.

En cada fichero Microsoft Access, nos encontramos con dos tablas que contienen la siguiente información:

Tabla 5: Venta

CAMPO	DESCRIPCIÓN	TIPO
Id_venda	Identificador de la venta en el establecimiento	NUMBER
Data	Fecha de la venta	DATE (DD/MM/YYYY)
Hora	Hora de la venta	DATE (HH24:MI:SS)
Forma_pagament	Forma de pago de la venta	CHAR
Id_cliente	Cliente que realizó la venta	NUMBER
Descompte	Descuento sobre la venta	NUMBER

Tabla 6: Detalle Venta

CAMPO	DESCRIPCIÓN	TIPO
Id_venda	Identificador de la venta en el establecimiento	NUMBER
Id_línea	Identificador de la línea de la venta	NUMBER
Unitats	Unidades vendidas	NUMBER
Id_producte	Identificador del producto vendido	NUMBER

El formato del nombre de los archivos será el siguiente:

- **Ficheros de ventas:** XXXXXXXXX.accdb

XXXXXXXXX**:** Nombre del establecimiento. Siempre el mismo nombre para un determinado establecimiento.

3.2.1.4. Información sobre número de habitantes

La información será suministrada por IDESCAT, y queda por determinar el mecanismo y formato mediante el cual será suministrada. Este podrá ser vía API o un procedimiento manual que incluya la descarga y posterior importación al modelo de datos del sistema.

La información a recuperar de IDESCAT es la siguiente:

Tabla 7: Información sobre número de habitantes

CAMPO	DESCRIPCIÓN	TIPO
Año	Fecha a 1 de enero del año objeto a informar	NUMBER
Demarcación	Demarcación sobre la que se adjunta información	CHAR
Habitantes	Nº de habitantes por demarcación	NUMBER

3.2.2 Información requerida por el cliente

Como se ha comentado anteriormente, la información mínima a proporcionar por el sistema requerida por el cliente es la siguiente:

- **Total de ventas y margen neto del grupo**
- **Importe medio de compra por socio y establecimiento**
- **% de ventas respecto el total del grupo (por establecimiento y por demarcación)**
- **Precios máximos y mínimos por tipo de establecimiento y características de producto.**
- **Ranking de establecimientos por número de ventas y volumen total**
- **"*Top ten" de productos**
- **% de ventas de "marcas blancas" por habitantes**
- **Categorización de Clientes A/B/C**
- **Análisis de compra por impulso**
- **Distribución semanal y estacionalidad de ventas**

3.2.3 Correspondencia de conceptos

En este apartado se intenta hacer una aproximación de la correspondencia de información que demanda el Cliente con la que es aportada al sistema, para estudiar la viabilidad de su posterior ofrecimiento por el sistema.

A continuación, se presenta una tabla donde se muestra una columna para cada origen de la información.

Tabla 8: Correspondencia de información

INFORMACIÓN SOLICITADA	UBICACIÓN	ORIGEN DE LA INFORMACIÓN
Nº Ventas	Ficheros MDBs	Nº distinto de campo Id_venda
Volumen de ventas	Ficheros MDBs, Excel producto	Sumatorio del precio del producto identificado por id_producte en la venta
Margen neto grupo	Ficheros MDBs, Excel producto y Establecimiento	Una vez obtenido el volumen de ventas por establecimiento se le restaría el coste del mismo
Importe medio compra	Ficheros MDBs, Excel producto	Sumatorio del precio del producto identificado por id_producte en la venta / Nº Ventas
% Ventas	Ficheros MDBs, Excel producto y Establecimiento	Nº Ventas para el criterio seleccionado / Nº Ventas para un criterio más global
% Volumen ventas	Ficheros MDBs, Excel producto y Establecimiento	Volumen de ventas para el criterio seleccionado / Volumen de ventas para un criterio más global
Precios máximos y mínimos	Ficheros MDBs, Excel producto	Máximo o mínimo del total de precio del producto identificado por id_producte en la venta
Nº Productos vendidos	Ficheros MDBs	Nº distinto de campo Id_producte

INFORMACIÓN SOLICITADA	UBICACIÓN	ORIGEN DE LA INFORMACIÓN
Margen del producto	Ficheros MDBs, Excel producto	Sumatorio del campo Preu_venta de los productos vendidos – sumatorio campo Preu_cost
Nº Ventas (por impulso)	Ficheros MDBs, Excel producto	Nº distinto de campo Id_venda para productos con campo Venda_impuls a Sí
Nº Habitantes	IDESCAT	Número de habitantes para una determinada demarcación
Socio	Ficheros MDBs	Campo id_client
Establecimiento	Ficheros MDBs, Excel establecimiento	El propio nombre del fichero MDB nos indica el establecimiento cuyo detalle se encuentra en la Excel
Tipología de Establecimiento	Ficheros MDBs, Excel establecimiento	Una vez identificado el establecimiento por el nombre, se obtiene su tipología en la Excel (campo Tipología)
Demarcación	Manual	Se enriquecerá de manera manual la tabla de establecimientos para indicar su demarcación
Producto	Excel de Productos	Se asociará a la venta mediante el campo id_producte
Tipología de Producto	Excel Productos	Una vez identificado el producto sería el campo Tipología
Proveedor	Excel Productos	Una vez identificado el producto sería el campo Proveidor
Marcas blancas	Excel Productos	Una vez identificado el producto sería cuando el campo Proveidor fuese GLDP
Tiempo	Ficheros MDBs	Campos Data y Hora de la venta que serán dimensionados

3.3. Análisis de la Calidad del dato

En este capítulo primeramente se recoge el análisis preliminar sobre la información a procesar, para a continuación proceder a comentar el tratamiento que ser realizará sobre los diferentes escenarios anómalos o de baja calidad del dato que se presenten durante las ejecuciones de los procesos ETL.

3.3.1 Análisis de la calidad de la información a procesar

En este apartado realiza un análisis inicial de la calidad de la información que va a ser tratada por los procesos de carga encargados de los mecanismos de clasificación y enriquecimiento del detalle de las ventas, verificando así la viabilidad de los mismos y previendo su futuro comportamiento. El objetivo de este análisis es estar preparado para la fase de desarrollo y tener más claro los posibles errores y anomalías que van a presentar los datos y que en ocasiones requieren de tratamientos específicos para asegurar el funcionamiento global del sistema.

A continuación se comentará por concepto los datos más relevantes descubiertos:

3.3.1.1. Productos

Las principales anomalías detectadas sobre productos son:

- **Productos sin precio de coste:** Existen productos sin precio de coste lo que ocasionará problemas sobre cálculos de margen.
- **Productos con precio de venta y precios de coste "anómalos":** Se han advertidos precios de productos muy por encima de su valor habitual, o con diferencias grandísimas entre coste y venta. Esto no ocasionará problemas en los cálculos más allá de un enturbiamiento del análisis.
- **Ventas de Productos no catalogados:** Existen registros de las ventas asociadas a códigos de productos no catalogados, o de fechas de las cuales no se tiene información. Este escenario es bastante grave en el establecimiento de Tarragona, donde aproximadamente el 33% de sus ventas no podrían asociarse a un producto catalogado en el sistema. Esta anomalía implica que no se disponga del precio de la venta lo cual afecta a multitud de indicadores a recuperar por el sistema.

3.3.1.2. Clientes

Las principales anomalías detectadas sobre clientes son:

- **Ventas sobre clientes no informados:** Existen ventas que no tienen indicado el identificador de cliente. Esto a priori no afecta los informes requeridos, pero sí puede originar que no se tengan en cuenta esos registros en determinados cálculos.

3.3.1.3. Ventas

Las principales anomalías detectadas sobre clientes son:

- **Registros del detalle de venta que no casan con el registro de la venta y viceversa:** Existen registros de detalle de ventas cuyo identificador de venta no tiene referencia en los registros de ventas. Esto implica que no pueda identificar para ese detalle, ni el cliente, ni la forma de pago, ni el descuento, y lo que es más importante la fecha y hora de la realización de esa venta. La siguiente tabla muestra por establecimiento, el número de registros de detalle, los registros en los que sí se establece la relación y una última columna con el porcentaje de registros que no casan.

Tabla 9: Correspondencia Detalle Venta / Venta

ESTABLECIMIENTO	REGISTROS DETALLE	REGISTROS CON CORRESPONDENCIA	% SIN CORRESPONDENCIA
Figueres	154.441	0	100%
Girona	151.920	150.341	1,04%
Lleida	74.345	74.254	0,12%
Olot	63.932	63.848	0,13%
Poble Sec	112.931	112.480	0,40%
Reus	83.809	83.713	0,11%

ESTABLECIMIENTO	REGISTROS DETALLE	REGISTROS CON CORRESPONDENCIA	% SIN CORRESPONDENCIA
Sants	63.932	63.848	0,13%
Tarragona	231.848	227.217	2,00%
Terrassa	169.031	123.482	26,95%
Tortosa	69.408	69.284	0,18%
Vielha	64.069	62.923	1,79%

3.3.2 Tratamiento de errores

El objetivo de este apartado es indicar cómo se comportarán los procesos de carga cuando se encuentren con los escenarios comentados en el apartado anterior y con otras anomalías que se puedan producir y que no hayan sido detectadas en esta fase de análisis.

Básicamente, y como se comenta en el detalle de la funcionalidad de los procesos, cada proceso de carga generará un fichero CSV de errores (uno por cada fichero/tabla que trate) con los registros que sufren anomalías, y un campo adicional de observaciones en el que se muestra el motivo. El objetivo es que estos ficheros se remitan al Cliente para que pueda modificar los ficheros de origen que nos proporcionan y así aumente la calidad de los datos que presenta el sistema.

Además, se puede dar la situación de que alguno de los procesos rechace el fichero íntegramente. Esta situación se informará por consola y se renombrará el fichero a *.ER. Esta situación sólo puede darse en caso de que el proceso no sea capaz de obtener del nombre del fichero la información necesaria (en el caso de productos, el año, y en el de las ventas, el establecimiento).

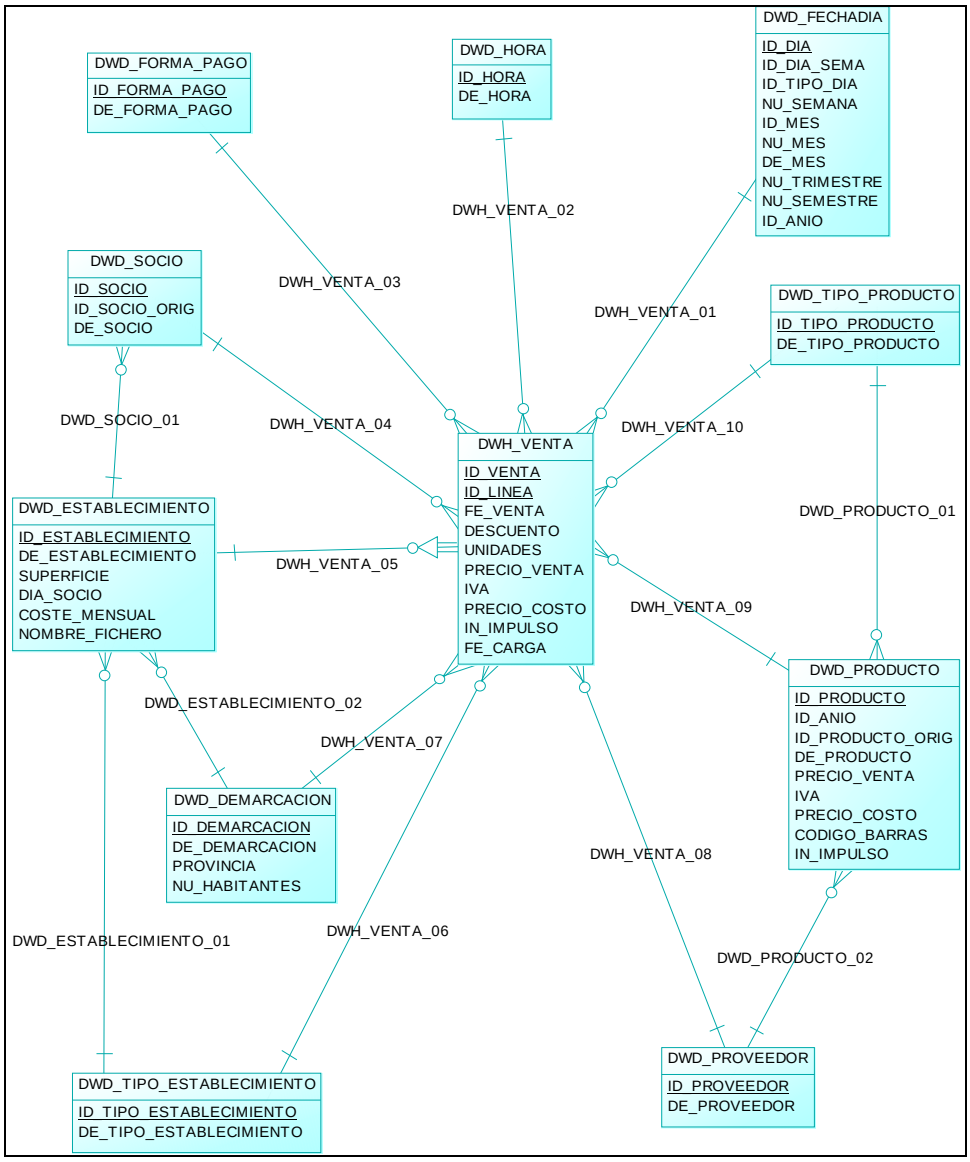
3.4. Modelo Conceptual

En este capítulo se recogen los elementos nivel conceptual a tener en cuenta en un almacén de datos de carácter multidimensional dentro del ámbito del proyecto. Para ello, primeramente se mostrará el diagrama conceptual que soportará la actividad del sistema, para a continuación pasar a comentar las diferentes consideraciones que se han tenido en cuenta para su generación.

3.4.1 Diagrama Modelo Conceptual

A continuación se muestra el modelo conceptual que recogerá la información del sistema:

Figura 5: Modelo Conceptual



3.4.2 Identificación de los Hechos

Llamamos evento o **Hecho** a una operación que se realiza en el negocio en un tiempo determinado. Son objeto de análisis para la toma de decisiones. En nuestro proyecto este concepto se instanciaría en el concepto “venta”, entendiendo como tal, cada una de las líneas de las que consta una operación de venta de un determinado establecimiento, es decir, el detalle de la venta.

3.4.3 Definición de la granularidad

Se refiere a la especificidad a la que se define un nivel de detalle en una tabla, es decir, si hablamos de una jerarquía la granularidad empieza por la parte más alta de la jerarquía, siendo la granularidad mínima, el nivel más alto de ésta.

Definir correctamente la granularidad de nuestro modelo es primordial para asegurar que somos capaces de cumplir los requisitos de información solicitados. Esto significa que si prescindimos de algún nivel de detalle y luego el cliente quiere obtener información a ese nivel, el sistema no podrá proporcionar el dato. En ocasiones, en un principio las peticiones del cliente no son al máximo nivel de detalle, pero esto hay que analizarlo muy bien, ya que una vez implementado el modelo no es posible aumentar el detalle si no se reprocesa toda la información, acción que en ocasiones no es viable dado la cantidad de información a procesar en sistemas de este tipo.

Por otro lado, no hay que olvidar que las consultas OLAP, por concepto, deben de cumplir unos requisitos mínimos de eficiencia en cuanto al tiempo de ejecución que presentan. Un nivel de granularidad muy grande (gran nivel de detalle) puede perjudicar el rendimiento de las consultas ya que obliga al sistema a manejar mucha cantidad de información para calcular/presentar los datos.

En nuestro proyecto, y dado a que la mayoría de los informes se requieren a nivel mensual, se podría pensar en consolidar la información a este nivel de detalle, teniendo en cuenta el resto de información asociada a la venta (Cliente, Forma de pago, Producto, etc…) pero prescindiendo del instante preciso del mes en el que se realiza la venta. De esta manera y como se puede observar en la tabla siguiente se conseguiría reducir el volumen de información en más de un 20%. La siguiente tabla muestra la reducción de registros que causaría almacenar la información a este nivel para los establecimientos con mayor número de ventas:

Tabla 10: Granularidad mensual

ESTABLECIMIENTO	REGISTROS DETALLE	GRANULARIDAD MENSUAL	% REDUCCIÓN
Tarragona	227.217	51.497	22,66%
Terrassa	123.482	26.197	21,22%
Girona	150.341	32.301	21,49%
Vielha	62.923	17.247	27,41%

Aunque esta reducción sería muy interesante de aplicar, información del tipo “distribución semanal y estacional de ventas” y el “Análisis de compra por impulso (se basa en el día de la semana)”, nos van a obligar a conservar el detalle a nivel de día. Puesto que en un mismo día un cliente no suele realizar varias compras (excepto el cliente genérico “0”), no nos planteamos eliminar la hora exacta de la misma, ya que no se conseguiría reducir considerablemente el número de registros, y por lo tanto nuestro nivel será el mismo que el que presentan los datos en el origen.

3.4.4 Definición de agregaciones

Relacionado con el punto anterior, nos encontramos con la posibilidad de realizar agregaciones en los datos. Las tablas de agregadas se utilizan para almacenar un resumen de los datos, es decir, se guardan los datos en con niveles de granularidad menores a los que inicialmente fueron obtenidos y/o consolidados.

La necesidad de crear tablas agregadas pueden ser, o bien para mejorar el rendimiento de consultas sobre el detalle que se demoran, o bien cuando los resúmenes de información a mostrar son complicados de calcular en tiempo de consulta.

En nuestro caso, en cuanto a número de registros, se estaría hablando de un máximo nivel de detalle con 2.500.000 millones de registros, esto unido a que las complejidades de cálculos pueden mediante implementaciones que ofrece la herramienta analítica por la que se ha optado (*Pentaho Business*

Analitycs), nos hace considerar que en esta fase incipiente de *Business Intelligence* no será necesario la elaboración de tablas agregadas para cubrir las expectativas del cliente.

3.4.5 Identificación de Dimensiones y Atributos

Una **Dimensión** es una característica de un hecho que permite su análisis posterior, en el proceso de toma de decisiones. Un claro ejemplo de dimensión en nuestro proyecto son "Cliente", "Establecimiento", "Producto". En cuanto a **Atributo,** lo podemos definir cómo la información asociada a una dimensión y que tiene como objetivo enriquecerla. Así por ejemplo en nuestro proyecto identificamos el atributo "Superficie" de la dimensión "Establecimiento".

A continuación se recogen el conjunto de dimensiones y atributos identificados. Esta información es proporcionada del siguiente modo:

- Dimensión: Descripción de la dimensión
  - Atributo: Descripción del atributo

La lista de dimensiones y atributos es la siguiente:

- Socio: Número de cliente de una determinada tienda
- Establecimiento: Establecimiento dentro de la cadena GLDP
  - Identificador: Identificador del establecimiento
  - Superficie: Superficie en metros cuadrados del establecimiento
  - Día del Socio: Día de la semana identificado como "Día del socio"
  - Coste fijo mensual: Gastos mensuales derivados del mantenimiento del establecimiento
- Tipología de Establecimiento: Clasificación en la que se enmarca un establecimiento
- Demarcación: Provincia sobre la que se analiza la información
- Producto: Producto involucrado en la venta
  - Identificador: Identificador del producto
  - Nombre del producto: Nombre descriptivo del producto
  - Código de barras: Código de barras del producto
- Tipología de Producto: Clasificación en la que se enmarca un producto
- Proveedor: Proveedor del producto.
- Marcas blancas: Clasificación de la venta en un determinado escenario
- Tiempo: Momento en el que se analiza una venta o un producto
  - Año: Año en el que se analiza una venta o un producto
  - Estación: Estación en la que se analiza una venta o un producto
  - Mes: Mes en el que se analiza una venta o un producto
  - Semana: Semana en la que se analiza una venta o un producto

Cabe destacar que determinada información puede ser objeto de duda en cuanto a ser encuadrada dentro del concepto "dimensión" o "atributo". Esta elección depende de si el objeto tiene peso

suficiente para llegar a ser dimensión, es decir, si la información va a ser utilizada como criterio de análisis o simplemente como enriquecimiento de datos. Un ejemplo de esto puede ser la dimensión "Tipología de Producto", que si bien podría ser una descripción del producto sin más, es candidata a ser definida como dimensión ya que muchos de los indicadores van a ser consultados utilizándolo como criterio de análisis.

3.4.6 Identificación de las jerarquías

Las **Jerarquías** son una agrupación de dimensiones que se relacionan con cardinalidad "uno a muchos" y por lo tanto establecen relaciones de jerárquicas entre ellas. Cabe destacar que en ocasiones una dimensión puede pertenecer a más de una jerarquía, o lo que es lo mismo que la dimensión tenga dos jerarquías distintas:

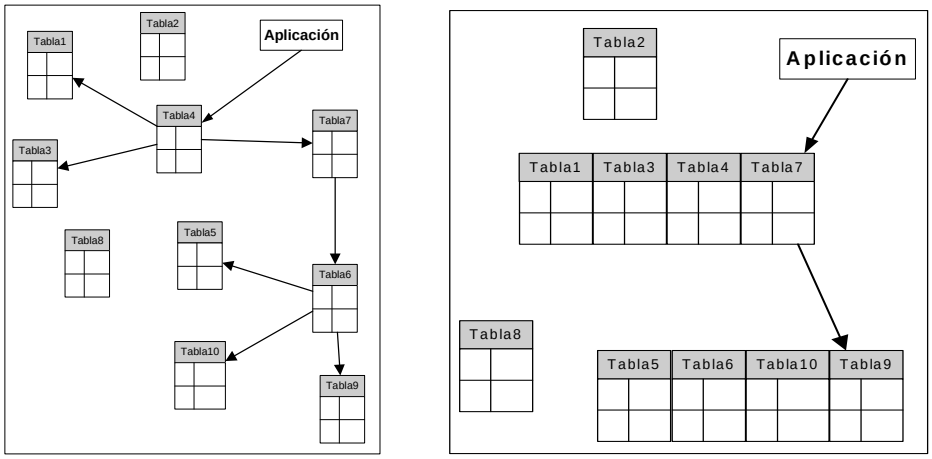
En nuestro proyecto podemos definir el siguiente conjunto de jerarquías::

- Tiempo Día: Año → Semestre → Trimestre → Mes → Día
- Producto:
  - Tipo Producto → Producto
  - Proveedor → Producto
- Establecimiento: Tipo Establecimiento → Establecimiento
- Socio: Establecimiento → Socio
- Demarcación: Provincia → Demarcación

3.4.7 Desnormalización

El proceso de desnormalización consiste en "mezclar" físicamente o desnormalizar (en la medida de lo posible y en función de la naturaleza de los datos) algunas de las tablas de dimensión para reducir el consumo de recursos utilizado para la *joins* (cruce de tablas) necesarias para llegar a sus datos y para agilizar el acceso a los mismos.

Figura 6: Ejemplo de desnormalización de tablas



La normalización / desnormalización de los datos en la tabla de hechos conllevará una serie de ventajas y de inconvenientes:

Tabla 11: Normalización / Desnormalización

NORMALIZACIÓN	DESNORMALIZACION
Los procesos que pretendan recuperar la descripción tendrán que acceder a ella mediante, al menos, un join entre tablas.	El acceso a los datos es más eficiente (no hay joins). Lo cual es imprescindible en sistemas en los que prima el tiempo de respuesta a las consultas.
Se ahorra espacio de disco, puesto que los datos normalizados suelen ser cadenas y sólo se almacenarán una vez.	Se incurre en un mayor gasto de espacio de disco, puesto que se escriben los datos sin codificar en la tabla de hechos.
Los procesos de inserción son más óptimos.	Los procesos de inserción tienen que insertar más datos al no estar éstos codificados.
Actualizar, por ejemplo, una descripción normalizada implica sólo actualizar la tabla de dimensión.	Actualizar una descripción normalizada requiere actualizar todos los registros en que se usa en la tabla de hechos.

En nuestro proyecto se realizarán principalmente las siguientes desnormalizaciones:

- **Desnormalización dimensión Tiempo:** Se mezclará toda la información de la jerarquía "Tiempo Día" en una única tabla, la del nivel mínimo de detalle "DIA". Esto evitará que para obtener información de los distintos niveles se realicen múltiples join en el momento de realizar la consulta.
- **Desnormalización de dimensiones a tabla de hechos:** Como se puede observar en el diagrama del modelo, se ha llevado información redundante en cuanto a dimensiones, a la tabla central del modelo VENTA, dimensiones que podían haberse consultado a partir de otras tablas, con las cuales ya están relacionadas. El conjunto de desnormalizaciones de este tipo es el siguiente:
  - Tipo Establecimiento
  - Tipo Producto
  - Proveedor

  En este caso, como en el apartado anterior, se permite analizar la información de manera individual, sin necesidad de realizar la join con las tablas de las que depende. Por ejemplo, es posible realizar consultas por "Tipo producto" (la cual presenta pocos valores), sin necesidad de cruzar con la tabla de "Producto" y calcular los datos a ese nivel de la jerarquía.
- **Desnormalización de medidas a tabla de hechos:** Se han introducido en la tabla de hechos, indicadores pertenecientes a alguna dimensión. En este caso estamos hablando de los campos IVA, PRECIO_COSTO, DESCUENTEO, IN_IMPULSO, los cuales podrían haber sido consultados a partir de la tabla PRODUCTO. Esta decisión permite prescindir de realizar siempre *join* con dicha tabla, la cual presenta una cardinalidad mayor, y poder realizar el análisis por otras dimensiones sin cruzar con ella.

3.4.8 Identificación de las medidas

Una **Medida** es una propiedad de un Hecho (casi siempre numérica), que es usada para su análisis. Un ejemplo de medias en nuestro proyecto serían "Nº Ventas", "Precio de la venta", etc.

El conjunto de medidas identificadas, es el siguiente:

- **Nº Ventas:** Nº transacciones realizadas en un periodo de tiempo
- **Volumen de ventas:** Suma de los importes de las ventas realizadas en un periodo de tiempo
- **Margen neto grupo:** Margen entre las ventas de productos y gastos de establecimiento
- **Importe medio de compra:** Media del valor de importe de las ventas
- **% Ventas:** Porcentaje de Nº de transacciones sobre el total
- **% Volumen de ventas:** Porcentaje de suma de importes sobre el total
- **Precios máximos y mínimos:** Indicador de precio máximo o mínimo de las ventas
- **Nº Productos vendidos:** Nº diferente de productos vendidos en un periodo de tiempo
- **Margen del Producto:** Margen entre el valor del producto y su precio de coste
- **Nº Ventas (por impulso):** Nº transacciones de productos "por impulso" realizadas en un periodo de tiempo
- **Nº Habitantes:** Número de habitantes por demarcación

Cabe destacar que un mismo indicador puede dar información diferente dependiendo del elemento de análisis por el cual se esté consultando (dimensiones). Así por ejemplo el indicador "Nº Ventas", puede servir para dar información del número de ventas por establecimiento, por demarcación, por Cliente…, o incluso por varias o todas a la vez.

3.4.9 Restricciones de integridad

Las restricciones de integridad forman parte del proceso de enriquecimiento de la información del modelo presentado. Esto significa que a partir de la información base de nuestro modelo "la venta", se tendrán que implementar mecanismos que permitan realizar búsquedas de información relacionada para poder enriquecerla y así cumplir con las restricciones que presenta el modelo. Aunque observando el modelo se perciben multitud de relaciones (SOCIO, FORMA PAGO, FECHA, etc…), la mayoría vienen dadas en el propio detalle o son obtenidas a partir de otras búsquedas. Finalmente el conjunto de restricciones que se deben de cumplir en el sistema son las siguientes:

- **Venta / Detalle Venta:** Los registros de detalle de las ventas tienen que hacer referencia a una línea de venta existente dentro del propio fichero de relación de ventas
- **Venta / Producto:** El producto identificado en el detalle de la venta, tiene que estar catalogado en la tabla de productos para la fecha en la que se produce dicha venta.
- **Venta / Establecimiento:** La identificación del establecimiento se realizará mediante el nombre del archivo suministrado por el cliente con la relación detallada de las ventas. Por ese motivo es primordial que este nombre sea siempre el mismo para que esta información se asocie de manera correcta.

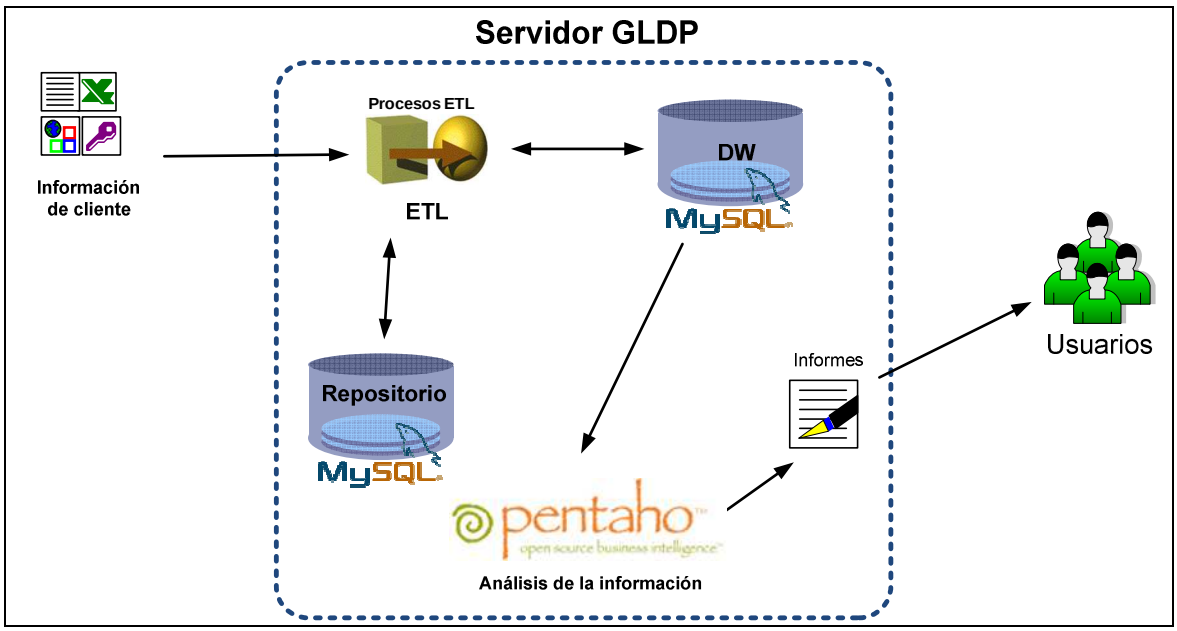
4. Diseño

En este capítulo se describen las principales características que definen el proyecto en cuanto al diseño del mismo, reflejando los apartados de arquitecturas HW y SW, diseño de modelo de datos y de procesos ETL.

4.1. Diagrama de arquitectura software

Los componentes software que definen el proyecto y su interrelación son los siguientes:

Figura 7: Arquitectura software



En el grafico anterior se distinguen los siguientes componentes:

- **Información de cliente:** Conjunto de ficheros e información externa que deberá de ser cargado en el almacén de datos.
- **Servidor GLDP:** Máquina virtual **XP** sobre ***VirtualBox*** que contiene los siguientes elementos
  - **DW:** Almacén de datos con el modelo de datos donde reside la información de cliente, implementado sobre el sistema gestor de Base de datos **MySQL 5.5**
  - **ETL:** Procesos de ***Pentaho Data Integrator y Scripts Batch***, encargado de transformar y cargar la información de cliente en el almacén DW
  - **Repositorio:** Conjunto de información o metadata manejada por la herramienta ETL de Pentaho donde reside su configuración. Implementado también sobre el mismo servidor MySQL 5.5.
  - ***Pentaho BI platform / Server:*** Plataforma ***Pentaho Business Analitycs*** que incluye herramientas que permite a través de consultas realizadas al DW, generar informes para los Usuarios
  - **Informes de usuarios:** Conjunto de informes realizados con ***Pentaho Report Designer*** que utilizan las capacidades que ofrece la plataforma *Pentaho BI.*

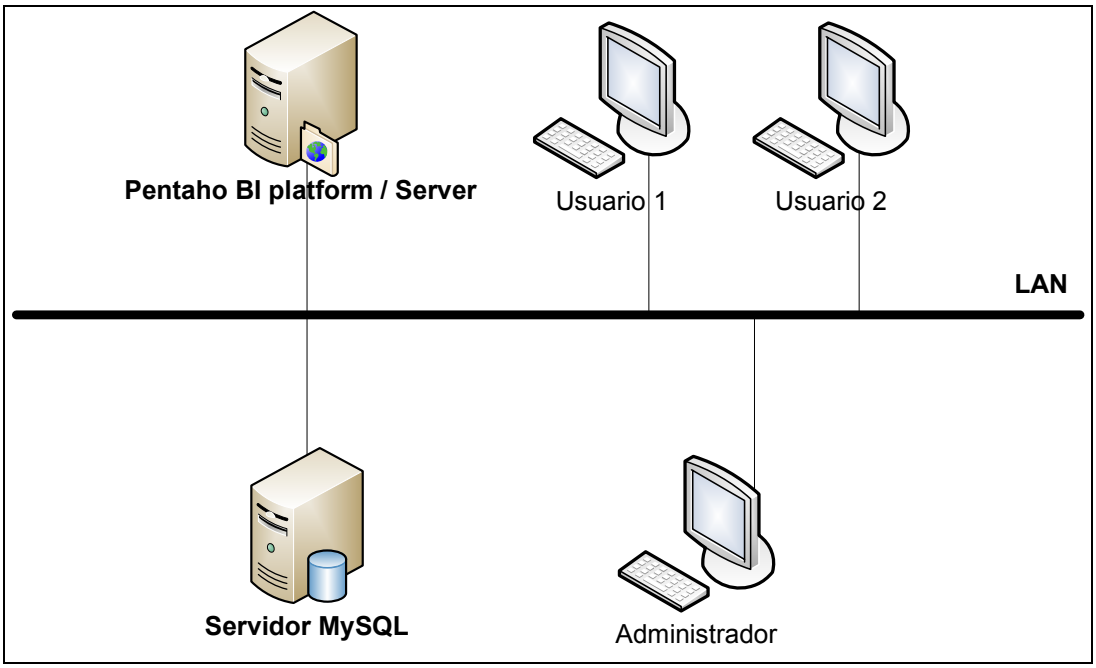
4.2. Diagrama de arquitectura hardware

El gráfico que se muestra a continuación, sugiere la exploración del sistema únicamente mediante una red local o LAN. Esto es así, debido a la propia naturaleza de los sistemas BI, que hacen que los usuarios candidatos a realizar consultas sobre el almacén de datos sean integrantes del equipo de dirección de la empresa, los cuales lo utilizan para la toma de decisiones a través del análisis de los datos que éste sistema brinda, y que suelen estar ubicados de una manera centralizada.

Además, como se ha comentado a lo largo del proyecto, el sistema dispone de un módulo de planificación y distribución de informes, lo que permite el envío de informes de manera periódica a los diferentes establecimientos que forman el grupo GLDP, dejando la actividad de análisis online y de generación de nuevos informes para efectuarse desde la propia red de la empresa.

Los componentes hardware que definen por tanto el proyecto y su interrelación son los siguientes:

Figura 8: Arquitectura hardware



- ***Pentaho BI platform / Server*:** Servidor que albergará el producto *Pentaho Business Analitycs* mencionado en el apartado anterior.
- **Servidor MySQL:** Servidor destinado para la instalación del servidor de base de datos MySQL que contendrá tanto el repositorio de la herramienta ETL como el propio almacén de datos del sistema.
- **PCs de Usuarios:** Equipos de usuarios finales que explotarán el almacén de datos mediante el acceso a la plataforma que ofrece el *Pentaho Business Analitycs*
- **PCs Administrador:** Ordenador desde el cual se elaborarán los informes que serán consumidos por el resto de usuarios, y que tendrá acceso al servidor donde corre la ETL para ubicar la información de entrada al sistema y que ésta sea procesada y consolidada en el almacén de datos.

4.3. Diseño de la base de datos

En este capítulo primeramente se recogen el conjunto de consideraciones a tener en cuenta sobre el diseño del modelo de base de datos, a continuación se muestra el diagrama del modelo físico correspondiente al modelo conceptual tratado en el capítulo anterior, y finalmente se pasan a detallar cada uno de los campos de los que consta el modelo.

4.3.1 Consideraciones sobre el modelo

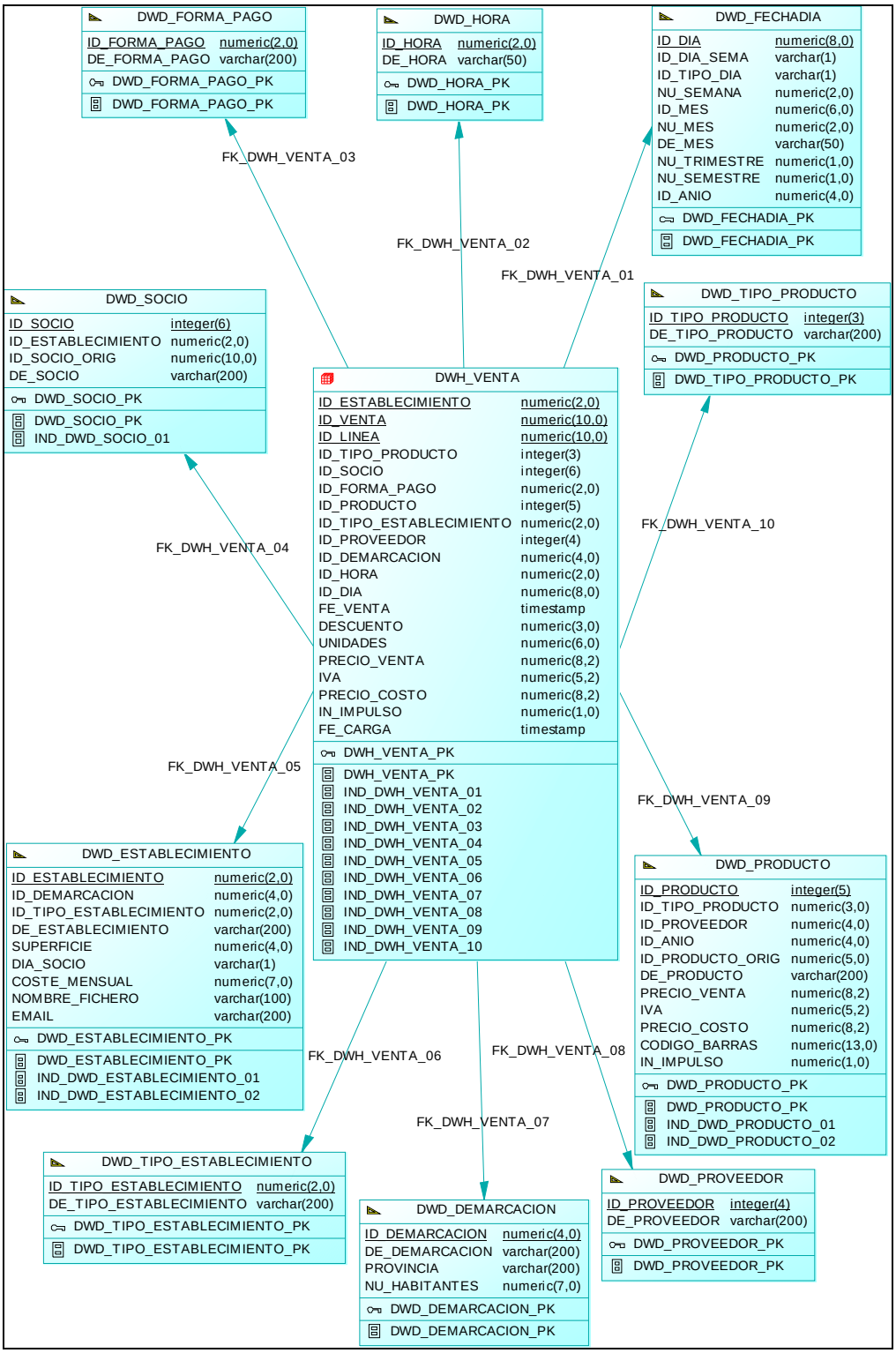
El conjunto de consideraciones a tener en cuenta que ayudan en la comprensión del modelo físico construido son las siguientes:

- Las tablas del diagrama presentan en su parte superior izquierda los iconos  y  para representar que se tratan de tablas de hechos o dimensiones respectivamente.
- Los campos subrayados representas campos que forman la clave primaria de la tabla.
- Para facilitar la comprensión e interpretación del mismo, se han eliminado las relaciones entre tablas que no pertenezcan exclusivamente al modelo en estrella núcleo del sistema.
- Las tablas incluyen en su nomenclatura los prefijos “DWH_” y “DWD_” para distinguir entre tablas de dimensión y tablas de hechos.
- Cada clave foránea establecida sobre una tabla, conlleva la generación de un índice sobre los mismos campos que son propagados a la tabla de destino de la restricción.
- Se han incluido dimensiones “Extra” DWD_FORMA_PAGO y DWD_HORA, que si a priori no se considerarían necesarias para la presentación de los informes requeridos por el cliente, en un futuro podrían dotar al sistema de capacidades de análisis interesantes para el negocio, las cuales por ejemplo responderían a las preguntas de “¿Cómo se distribuyen las ventas a lo largo del día?” ¿En qué horas se concentra la máxima actividad de venta?“ ¿En fín de semana se distribuyen más las ventas a lo largo del día? ¿Qué forma de pago es la más utilizada?” etc…
- Como se comenta en el apartado **3.4.7 Desnormalización**, se han creado nuevas dimensiones a partir de las dimensiones iniciales obtenidas de los datos recibidos por el cliente, como son DWD_TIPO_PRODUCTO, DWD_PROVEEDOR, DWD_TIPO_ESTABLECIMIENTO que han sido unidas a la tabla de hecho central DWH_VENTA para aportar mejoras de rendimiento del sistema ROLAP que se basa en el modelo.
- Se ha añadido el campo FE_CARGA en la tabla de detalle, el cual indica el momento en que se realiza la inserción del registro en BBDD para ayudar a posibles tareas administrativas relacionadas con el reproceso y borrado de información.
- Se ha añadido el campo NOMBRE_FICHERO en la dimensión DWD_ESTABLECIMIENTO, para ayudar a los procesos ETL a identificar el establecimiento a partir del nombre de fichero de detalle de ventas que se está procesando
- La clave primaria de la tabla DWD_SOCIO será autogenerada por el sistema y no utilizará los códigos que vienen en el detalle de la venta (campo ID_SODIO_ORIG), ya que éstos son compartidos por varios establecimientos, quedando identificado de manera unívoca por la dupla ID_SOCIO_ORIG, ID_ESTABLECIMIENTO. Generando su propio campo clave se consigue desvincular del establecimiento y se propaga un único campo a la tabla de detalle.

4.3.2 Diagrama Modelo Físico

A continuación se muestra el modelo conceptual que recogerá la información del sistema:

Figura 9: Modelo Físico



4.3.3 Descripción detallada de los campos del modelo

A continuación, se pasará a describir las tablas presentes en el diagrama anterior, indicando el significado de cada uno de los campos por los que están formados:

4.3.3.1. Tabla DWH_VENTA

Tabla central del modelo en estrella que contiene la información que define el hecho "Venta" y las métricas relacionada con el mismo. Cabe destacar, que la tabla está relacionada con el resto de tablas de dimensión mediante claves foráneas que propagan las claves primarias éstas dentro de la propia tabla de hechos.

El detalle de los campos que presenta es el siguiente:

- **ID_ESTABLECIMIENTO:** Clave foránea que establece la restricción de integridad con la tabla DWD_ESTABLECIMIENTO a través de su clave primaria
- **ID_VENTA:** Identificador de la venta recogido de la interfaz de origen venta, que ejerce como clave única dentro de las ventas de un mismo establecimiento.
- **ID_LINEA:** Identificador de línea dentro de la venta recogido del fichero de origen de detalle de venta.
- **ID_TIPO_PRODUCTO:** Clave foránea que establece la restricción de integridad con la tabla DWD_TIPO_PRODUCTO a través de su clave primaria
- **ID_SOCIO:** Clave foránea que establece la restricción de integridad con la tabla DWD_SOCIO a través de su clave primaria
- **ID_FORMA_PAGO:** Clave foránea que establece la restricción de integridad con la tabla DWD_FORMA_PAGO a través de su clave primaria
- **ID_PRODUCTO:** Clave foránea que establece la restricción de integridad con la tabla DWD_PRODUCTO a través de su clave primaria
- **ID_TIPO_ESTABLECIMIENTO:** Clave foránea que establece la restricción de integridad con la tabla DWD_TIPO_ESTABLECIMIENTO a través de su clave primaria
- **ID_PROVEEDOR:** Clave foránea que establece la restricción de integridad con la tabla DWD_PROVEEDOR a través de su clave primaria
- **ID_DEMARCACION:** Clave foránea que establece la restricción de integridad con la tabla DWD_DEMARCACION a través de su clave primaria
- **ID_HORA:** Clave foránea que establece la restricción de integridad con la tabla DWD_HORA a través de su clave primaria
- **ID_DIA:** Clave foránea que establece la restricción de integridad con la tabla DWD_DIA a través de su clave primaria
- **FE_VENTA:** Campo que indica el momento exacto en el que se realiza la venta con precisión "*timestamp*"
- **DESCUENTO:** Descuento aplicado a la venta
- **UNIDADES:** Unidades vendidas en la venta
- **PRECIO_VENTA:** Precio del producto que interviene en la venta

- **IVA:** Impuesto de Valor Añadido sobre el producto que interviene en la venta
- **PRECIO_COSTO:** Precio de coste del producto que interviene en la venta
- **IN_IMPULSO:** Campo que indica si el producto que interviene en la venta está catalogado como "Venta por Impulso"
- **FE_CARGA:** Fecha en la que se realizó la carga del registro en la BBDD. Útil para posible labores de mantenimiento.

4.3.3.2. Tabla DWD_FORMA_PAGO

Tabla de dimensión que contiene las diferentes modalidades de pago en lo que se produce la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_FORMA_PAGO:** Clave primaria de la tabla.
- **DE_FORMA_PAGO:** Campo con la descripción de forma de pago.

4.3.3.3. Tabla DWD_HORA

Tabla de dimensión que contiene las diferentes horas en los que se produce la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_HORA:** Clave primaria de la tabla.
- **DE_HORA:** Campo descriptivo de la hora

Añadiendo campos a esta tabla se podrían establecer horarios (horario mañana, horario tarde, etc…)

4.3.3.4. Tabla DWD_FECHADIA

Tabla de dimensión que contiene el calendario de fechas en los que se produce la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_DIA:** Clave primaria de la tabla que además representa el día en formato YYYYMMDD
- **ID_DIA_SEMA:** Identificador del día de la semana (L,M,X,J,V,S y D)
- **ID_TIPO_DIA:** Identificador del tipo de día de la semana Laborar (L) o festivo (F)
- **NU_SEMANA:** Número de la semana del año (semana 1, semana 2, …)
- **ID_MES:** Identificador del mes en formato YYYYMM
- **NU_MES:** Número del mes (1,2,3 …)
- **DE_MES:** Literal del mes (Enero, Febrero, …)
- **NU_TRIMESTRE:** Número de trimestre del año (1,2,3 …)
- **NU_SEMESTRE:** Número de semestre del año (1,2,3 …)
- **ID_AÑO:** Identificador del año en formato YYYY

4.3.3.5. Tabla DWD_TIPO_PRODUCTO

Tabla de dimensión que contiene las diferentes tipologías de productos sobre los que se genera la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_TIPO_PRODUCTO:** Clave primaria de la tabla
- **DE_TIPO_PRODUCTO:** Campo con la descripción de la tipología de producto.

4.3.3.6. Tabla DWD_PRODUCTO

Tabla de dimensión que contiene los diferentes productos sobre los que se genera la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_ PRODUCTO:** Clave primaria de la tabla
- **ID_ANIO:** Año de vigencia del producto y para el cual presenta los valores que se detallan a continuación
- **ID_PRODUCTO_ORIG:** Código de producto en la hoja de datos del origen
- **ID_TIPO_PRODUCTO:** Clave foránea que establece la restricción de integridad con la tabla DWD_TIPO_PRODUCTO a través de su clave primaria
- **ID_PROVEEDOR:** Clave foránea que establece la restricción de integridad con la tabla DWD_PROVEEDOR a través de su clave primaria
- **DE_PRODUCTO:** Descripción del producto
- **PRECIO_VENTA:** Precio por el cual se vende el producto en el año indicado
- **IVA:** Impuesto de Valor Añadido sobre el producto para ese año
- **PRECIO_COSTO:** Precio de coste para ese producto en el año indicado
- **CODIGO_BARRAS:** Código de barras que posee el producto para ese año
- **IN_IMPULSO:** Indicador de venta por impulso que informa que el producto se está promocionando como tal.

4.3.3.7. Tabla DWD_PROVEEDOR

Tabla de dimensión que contiene los diferentes proveedores de los productos sobre los que se genera la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_PROVEEDOR:** Clave primaria de la tabla
- **DE_PROVEEDOR:** Campo con la descripción del proveedor.

4.3.3.8. Tabla DWD_DEMARCACION

Tabla de dimensión que contiene las diferentes demarcaciones a la que pertenecen los establecimientos donde se genera la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_DEMARCACION:** Clave primaria de la tabla
- **DE_DEMARCACION:** Campo con la descripción de la demarcación.
- **PROVINCIA:** Provincia a la que pertenece la demarcación
- **NU_HABITANTES:** Número de habitantes que posee la demarcación

4.3.3.9. Tabla DWD_TIPO_ESTABLECIMIENTO

Tabla de dimensión que contiene los diferentes tipos de establecimientos donde se genera la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_TIPO_ESTABLECIMIENTO:** Clave primaria de la tabla
- **DE_TIPO_ESTABLECIMIENTO:** Campo con la descripción del tipo de establecimiento

4.3.3.10. Tabla DWD_ESTABLECIMIENTO

Tabla de dimensión que contiene los establecimientos donde se genera la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_ESTABLECIMIENTO:** Clave primaria de la tabla
- **ID_DEMARCACION:** Clave foránea que establece la restricción de integridad con la tabla DWD_DEMARCACION a través de su clave primaria
- **ID_TIPO_ESTABLECIMIENTO:** Clave foránea que establece la restricción de integridad con la tabla DWD_TIPO_ESTABLECIMIENTO a través de su clave primaria
- **DE_ESTABLECIMIENTO:** Campo con la descripción del tipo de establecimiento
- **SUPERFICIE:** Superficie en metros cuadrados del establecimiento
- **DIA_SOCIO:** Día de la semana que es considerado para el establecimiento como día del socio
- **COSTE_MENSUAL:** Coste de mantenimiento del establecimiento
- **NOMBRE_FICHERO:** Nombre que figurará en el fichero que contendrá las ventas de ese establecimiento
- **EMAIL:** Dirección de correo electrónico para la planificación de informes.

4.3.3.11. Tabla DWD_SOCIO

Tabla de dimensión que contiene los diferentes socios de un determinado establecimiento en el que se genera la venta.

El detalle de los campos que presenta es el siguiente:

- **ID_ SOCIO:** Clave primaria de la tabla
- **ID_ESTABLECIMIENTO:** Clave foránea que establece la restricción de integridad con la tabla DWD_ESTABLECIMIENTO a través de su clave primaria
- **ID_SOCIO_ORIG:** Código del socio en el fichero de ventas del origen
- **DE_SOCIO:** Campo para almacenar información complementaria del socio

5. Implementación

La fase de desarrollo del almacén de datos se ha dividido en cuatro partes muy diferenciadas, las cuales tenían fuerte dependencia entre ellas.

5.1. Despliegue del modelo de datos

La primera actividad abordada en la fase de implantación, ha sido la creación del modelo de datos que almacenará la información del grupo GLDP. Esto es así ya que los siguientes pasos dependen fundamentalmente de que exista dicho modelo, ya sea para la creación de los procesos de carga, que insertan en él, como para la ejecución de informes que necesitan de los datos cargados.

Para llevar a cabo esta actividad se han seguido los siguientes pasos:

- Familiarización con la herramienta "MySQL Workbench 5.2 CE" y las peculiaridades del lenguaje DDL de MySQL
- Configuración de las conexiones con la base de datos (esquema) dw.
- Unificación y creación del script con las sentencias de creación del modelo de datos
- Familiarización con la herramienta de interfaz de comandos de MySQL.
- Familiarización con el lenguaje de programación de ficheros por lotes BATCH de Microsoft Windows.
- Preparación del fichero por lotes que ejecute la creación del modelo y controle si su ejecución ha sido correcta

5.2. Construcción de la ETL

En esta fase se procedieron a generar todos los procesos de carga de información al modelo de datos creado en la etapa anterior.

La primera tarea consistía en determinar y desplegar la estructura de directorios de la aplicación, resultando la siguiente:

- **E:\etl\datos\entrada:** Carpeta donde el administrador alojará los datos a ser procesados por la aplicación
- **E:\etl\datos\inicial:** Carpeta donde se ubican los ficheros que contienen los valores iniciales de las dimensiones del modelo de datos.
- **E:\etl\datos\log:** Carpeta donde los procesos dejarán los ficheros de trazas
- **E:\etl\datos\salida:** Directorio de salida del sistema donde se alojarán los informes generados en formato *.pdf para su posterior distribución
- **E:\etl\datos\tmp:** Carpeta temporal de trabajo de la ETL candidata a ser utilizada para la generación de ficheros intermedios (de ordenación, para email, …)
- **E:\etl\mysql:** Carpeta donde residen los script de MySQL, en esta caso el script de creación del modelo de datos. También podría ser utilizado para script llamados desde la ETL como para la creación de índices, script de mantenimiento de BD, etc…
- **E:\etl\script:** Script de lanzamiento de los diferentes módulos que forman la ETL del sistema.

Lo siguiente fue crear la estructura de carpetas del repositorio de *Spoon*, las cuales albergan el conjunto de transacciones y trabajos, que forman la ETL de carga, quedando dividido por las áreas organizativas siguiente:

- **inicial:** Carpeta del repositorio que incluye los trabajos y transacciones que se encargan de la carga inicial de dimensiones.
- **incremental_dimensiones:** Carpeta que contiene los trabajos y transacciones encargados de la carga de dimensiones incrementales originadas del procesamiento de ficheros enviados por el cliente
- **hechos:** Carpeta que contiene los trabajos y transacciones encargados de la carga de ficheros que derivarán en las tablas de hechos del sistema
- **comunes (directorio */*):** Carpeta con procesos genéricos que afectan a varias áreas organizativas de la ETL:

La construcción de la ETL se puede dividir en las siguientes subfases:

5.2.1 Carga inicial

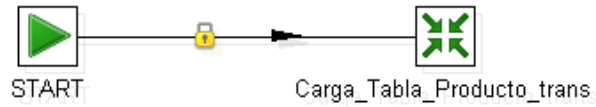
En esta fase se crearon los ficheros *.csv que contendrán los valores iniciales de las dimensiones del modelo de datos. Estos datos podrían ser los valores por defecto comunes a todas las dimensiones (-1 para valor no informado y -9 para valor no encontrado) o valores de dimensiones consideradas como estáticas las cuales no son previsibles que sufran modificaciones en sus valores en el corto plazo. El conjunto de dimensiones de carga inicial es el siguiente:

- DWD_DEMARCACION
- DWD_ESTABLECIMIENTO
- DWD_FECHADIA
- DWD_FORMA_PAGO
- DWD_HORA
- DWD_TIPO_ESTABLECIMIENTO

Una vez elaborado los ficheros anteriores se empezaron a construir los procesos ETL de carga inicial con la herramienta PDI de *Pentaho*.

Los pasos que se han seguido son los siguientes:

- Familiarización con la herramienta "*spoon*"
- Creación de las transacciones y trabajos necesarios para la carga inicial. Se trataban de transacciones y trabajos sencillos denominados "*pass through*" que son del siguiente modo:
  - Transacciones:



  - Trabajo:

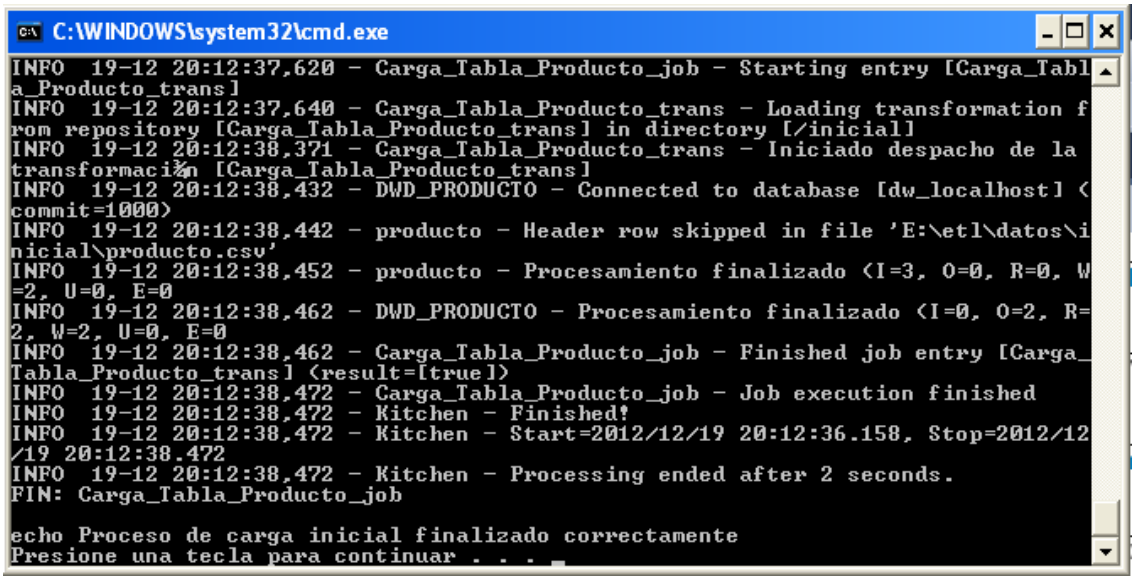
- Familiarización de los programas *"Pan" y "Kitchen"* de *Pentaho* encargado de lanzar *jobs* y transformaciones diseñadas con *Spoon*
- Preparación de los ficheros por lotes batch encargados del lanzamiento de los jobs que utilizaban los programas del punto anterior

La lista de software que forma la carga inicial ubicado en la carpeta de repositorio "/inicial" es el siguiente:

- Transacciones:
  - Carga_Tabla_Demarcacion_trans
  - Carga_Tabla_Establecimiento_trans
  - Carga_Tabla_Fechadia_trans
  - Carga_Tabla_Forma_Pago_trans
  - Carga_Tabla_Hora_trans
  - Carga_Tabla_Producto_trans
  - Carga_Tabla_Proveedor_trans
  - Carga_Tabla_Socio_trans
  - Carga_Tabla_Tipo_Establecimiento_trans
  - Carga_Tabla_Tipo_Producto_trans

- Trabajos:
  - Carga_Tabla_Demarcacion_job
  - Carga_Tabla_Establecimiento_job
  - Carga_Tabla_Fechadia_job
  - Carga_Tabla_Forma_Pago_job
  - Carga_Tabla_Hora_job
  - Carga_Tabla_Producto_job
  - Carga_Tabla_Proveedor_job
  - Carga_Tabla_Socio_job
  - Carga_Tabla_Tipo_Establecimiento_job
  - Carga_Tabla_Tipo_Producto_job

- *Script batch*:
  Todos los trabajos son lanzados mediante un único archivo *.bat que realiza tantas llamadas al ejecutable Kitchen.bat de PDI como trabajos están englobados en esta área, evaluando si su ejecución es correcta. El nombre del ejecutable es el siguiente: **Carga_Inicial.bat**. A continuación se muestra la salida por consola del ejecutable:



5.2.2 Procesos ETL de carga de dimensiones incrementales y tabla de hechos

En esta siguiente etapa se crearon el resto de procesos ETL, los cuales comprendían de una dificultad mayor. Para ello se siguió la siguiente estrategia de implementación, en la que en cada jornada de trabajo se procedía de la siguiente forma:

- **Recopilación y análisis de los resultados de las ejecuciones:** Dada la naturaleza de los procesos ETL propios de los almacenes de datos, los cuales son candidatos a requerir grandes cantidades de tiempo, se estableció desde el principio la estrategia de lanzar, en cuanto fuera posible y con la funcionalidad que se tuviera implementada en cada momento, cargas nocturnas con todo el conjunto de datos proporcionados para su carga. De esta manera se conseguía probar la mayor casuística desde el principio, y se iban recogiendo datos para la posterior fase de *"tunnig"* de procesos (mejoras en el rendimiento).
- **Desarrollo y pruebas:** Cada jornada se intentaba abordar una funcionalidad pequeña y bien definida para poder finalizarla en el día y realizar las pruebas específicas sobre cada apartado y darlo así por cerrado.
- **Ejecuciones nocturnas:** Como se ha comentado anteriormente, cuando las pruebas unitarias de un apartado se daban por concluidas, se pasaba toda la pila de ficheros disponibles por la funcionalidad terminada para encontrar posibles errores y no esperar a última hora.
- **Realización del Backup:** Cada jornada terminaba con la exportación a fichero XML de las transacciones o trabajos modificados en el día, para su posterior *backup* en "la nube" y en dispositivo de almacenamiento externo *pendrive*.

5.2.2.1. Procesos de carga de dimensiones incrementales

Los procesos de carga incremental de dimensiones, son los encargados de poblar dimensiones cuya información aumenta durante la vida del sistema, ya que no se conoce a priori todo el conjunto de datos que pueden contener.

Este tipo de dimensiones se pueden clasificar a su vez en dos tipologías diferentes: las proporcionadas por el cliente de manera explícita, y las autogeneradas a partir de los datos para cumplir integridad referencial del modelo. En nuestro proyecto, la primera tipología se corresponde con el catálogo de productos proporcionados por el cliente y que será consolidado en la tabla DWD_PRODUCTO. En cuanto a la segunda tipología, podemos encuadrar por ejemplo las tablas DWD_SOCIO, y DWD_TIPO_PRODUCTO, cuyos registros serán generados a partir del procesamiento de las hojas Excel de productos y los ficheros de ventas respectivamente. Los valores nuevos encontrados para estas tablas deben de ser insertados antes de la consolidación de la tabla de detalle para asegurar la integridad referencial del modelo.

La lista de software que forma la carga de dimensiones incrementales ubicado en la carpeta de repositorio "/incremental_dimensiones" es el siguiente:

- Transacciones:
    - Validacion_Previa_Carga_Productos_trans: Recopila la lista de ficheros de productos que se van a procesar y valida que en su nombre se indique un año válido. Este valor es pasado por parámetro posteriormente al resto de procesos involucrados en la carga de productos. Si los ficheros no cumplen las restricciones de nomenclatura son renombrados como *.ER.csv.
    - Carga_Dimensiones_Tabla_Producto_trans: Se encarga de obtener los diferentes valores de proveedores y tipo_productos de los ficheros a tratar para autogenerar las dimensiones que contienen dichos valores y asignarles un valor autoincremental.

- Carga_Tabla_Producto_Incr_trans: Carga los productos según el año indicado en la nomenclatura del fichero asignando los códigos calculados en el paso anterior. Además realiza validaciones de registros recogidas en el apartado **5.5.2 Punto unificado de validación de registros,** llevando los que no la superen a un archivo con el mismo nombre pero con extensión *.desc. También discrimina de si existe ya el producto o si ha habido alguna modificación en los datos que se están recibiendo para decidir si los inserta, modifica, o no realiza ninguna acción.

- Trabajos:
  - Carga_Productos_job: Trabajo que lanza la transacción que recopila los ficheros a procesar para posteriormente lanzar el trabajo que se indica en el punto siguiente
  - Procesa_Fichero_Productos_job: Lanza tanto la carga incremental de dimensiones como la transacción propia de carga de ficheros y llama a la transacción que renombra el fichero a OK o a ER dependiendo de si su ejecución es correcta o no.

- *Script batch*:
  El trabajo que contiene toda la lógica de control de la carga de productos es el trabajo "Carga_Productos_job" y es lanzado mediante un archivo *.bat que simplemente realiza una llamadas al ejecutable Kitchen.bat de PDI para que lance el trabajo anterior y evalúan si su ejecución es correcta. El nombre del ejecutable es el siguiente: **Carga_Productos.bat**

5.2.2.2. Procesos de carga de tabla de hechos

Los procesos de carga de tabla de hechos, son los encargados de poblar la tabla principal del almaén de datos donde se almacenan las medidas a consultar y mostrar en los informes. Estos proceso deben de estar lo suficientemente optimizados ya que manejan gran cantidad de datos y su ejecución puede llevarse varias horas de procesamiento.

La lista de software que forma la carga de la tabla de hechos ubicado en la carpeta de repositorio "/hechos" es el siguiente:

- Transacciones:
  - Validacion_Previa_Carga_Ventas_trans: Recopila la lista de ficheros de productos que se van a procesar y valida que en su nombre se indique establecimiento válido. Este valor junto con el de la demarcación y la tipología de establecimiento, es pasado por parámetro posteriormente al resto de procesos involucrados en la carga de ventas. Si los ficheros no cumplen las restricciones de nomenclatura son renombrados como *.ER.accdb.
  - Carga_Auxiliares_Venta_trans: Carga las dos tablas que llegan en los ficheros Access, en el dos tablas auxiliares ya en el modelo auxiliar de nuestro almacén de datos. Esto es así para optimizar la parte de join entre la tabla de ventas y de detalle ya que se ha demostrado empíricamente que los tiempos obtenidos utilizando este mecanismo son alrededor de 10 veces mejores
  - Carga_Dimensiones_Tabla_Venta_trans: Se encarga de obtener los diferentes valores de socios de los ficheros a tratar para autogenerar la dimensión que contienen dichos valores y asignarles un valor autoincremental
  - Carga_Tabla_Venta_Incr_trans: Descarga las ventas de las tablas auxiliares emulando un "full outer join" en MySQL (ya que de por sí el gestor no dispone de esta capacidad) para poder identificar los registros sin referencia tanto de ventas como de detalle e informar de si se produce esta casuística. Además realiza validaciones de registros recogidas en el apartado **5.5.2 Punto unificado de validación de registros,** llevando los que no la

superen a un archivo con el mismo nombre pero con extensión *.desc. Por último procede a insertar o modificar los registros de la tabla de hechos según existiesen o no en ella.

- Trabajos:
  - Carga_Ventas_job: Trabajo que lanza la transacción que recopila los ficheros a procesar para posteriormente lanzar el trabajo que se indica en el punto siguiente
  - Procesa_Fichero_Ventas_job: Lanza tanto la carga incremental de dimensiones como la transacción propia de carga de ficheros y llama a la transacción que renombra el fichero a OK o a ER dependiendo de si su ejecución es correcta o no.

- *Script batch*:
  El trabajo que contiene toda la lógica de control de la carga de ficheros de venta es el trabajo "Carga_Ventas_job" y es lanzado mediante un archivo *.bat que simplemente realiza una llamadas al ejecutable Kitchen.bat de PDI para que lance el trabajo anterior y evalúa si su ejecución es correcta. El nombre del ejecutable es el siguiente: **Carga_Ventas.bat**

5.2.2.3. Procesos generales

Por último queda un conjunto de elementos que son generales a los módulos anteriores y que por lo tanto se ubican en la carpeta raíz "/" del repositorio:

- Transacciones:
  - Renombra_Fichero_Origen_trans: Transacción que según el estado de la ejecución previa, se encarga de renombrar a OK o a ER el fichero que se le pasa por parámetro.

- Trabajos:
  - Carga_Diaria_job: Trabajo que contiene toda la lógica de control de la carga tanto de ficheros de productos como de ventas así como de realizar el envío de correo electrónico con el resumen de la ejecución acontecida. Esta funcionalidad está recogida en el apartado **5.5.5 Envío de correo con el resultado de las ejecuciones.**

- *Script batch*:
  El script **Carga_Diaria.bat** es el encargado de llamar al trabajo "Carga_Diaria_job", Por lo tanto este script es el candidato a ser planificado de manera automática para que se ejecute periódicamente ya que el trabajo que encapsula está preparado para funcionar correctamente tanto si hay ficheros para procesar como si no.

Cabe resaltar que además de un script global, como se puede deducir de la lista anterior, es posible lanzar de manera manual e independiente, las cargas de productos y de ventas mediante los respectivos script mencionados anteriormente.

5.2.3 *"Tuning"* de procesos (mejora en el rendimiento)

Gracias a la estrategia seguida indicada en el apartado anterior, ha sido posible ir optimizando paulatinamente los procesos consiguiendo mejoras significativas en su rendimiento. Este apartado será abordado en detalle en siguientes capítulos.

5.3. *Reporting* y calidad del dato

Una vez terminada la fase de la construcción de la ETL, ya estábamos en posición de abordar la elaboración de informes. Para llevar a cabo esta actividad se han seguido los siguientes pasos:

- Familiarización con la herramienta *"Report-designer" de Pentaho*
- Configuración de las conexiones y carpetas en la consola de usuario de *Pentaho*
- Configurar el sistema para permitir la exportación de informes desde la herramienta de *"reporting"*
- Preparar el primer informe o plantilla, que se utilizará como base para el resto de informes. El conjunto de información a destacar sobre esta plantilla es el siguiente:
    - Elección de una paleta de colores
    - Elaboración de una cabecera y pie de informe (con número de página) vistosa
    - Creación de un campo con la fecha de actualización del informe para la cabecera
    - Creación del logotipo de la empresa para poner en el pie:

    - Creación del efecto cebra para las tablas de los informes
- Una vez terminada la plantilla se siguió con la elaboración de todos los informes recogidos en el apartado de análisis de la primera PEC.
- Por cada informe elaborado, se procedía a revisar el significado de la información y la calidad del dato, contrastando en la medida de lo posible con los datos recibidos. Esto permitían comprobar que la información mostrada en el informe era la correcta y cuando no lo era se procedía a revisar y a subsanar el error.

Cabe destacar que como en la fase anterior, después de cada jornada, los informes eran exportados a fichero XML para su posterior *backup* en "la nube" y en dispositivo de almacenamiento externo *pen drive*.

5.4. Distribución de informes

Una vez finalizada la fase de elaboración de informes, se ha procedido a utilizar una de las nuevas funcionalidades que ofrece la versión de *Pentaho* instalada. Esta funcionalidad consiste en la preparación de una ETL que permite ejecutar informes generados con *"Report-designer"* y pasarle de manera dinámica diferentes parámetros, para que así de esta manera un mismo informe se genere varias veces y con datos diferentes. Una vez generado el fichero de salida en formato *.pdf, la ETL se encarga de enviar por correo dicho informe de manera dinámica también.

La lista de software que sirve para ejecutar de manera automática la distribución de informes se aloja en la carpeta raíz del repositorio de PDI, y está formado por los siguientes elementos:

- Transacciones:
    - Distribucion_Informes_trans: Transacción que recupera de la lista de establecimientos, los que posean en su campo EMAIL, una dirección de correo. Posteriormente establece las constantes de usuario de correo electrónico, la password, servidor SMTP, ruta e informe a distribuir, ruta de salida de la ejecución de informes. A continuación, llama al componente que permite instanciar la plantilla de informe y generarlo en formato *.pdf dependiendo del código de establecimiento del que se trate. Finalmente pasa a enviar tantos correos como ficheros se hayan generado.

- Trabajos:
  - Distribucion_Informes_job: Trabajo que llama a la transacción descrita en el punto anterior.

- *Script batch*:
  El script encargado de lanzar el trabajo "Distribucion_Informes_job" es **Distribucion_Informes.bat**, y su función es simplemente realizar una llamada al ejecutable Kitchen.bat de PDI para que lance el trabajo anterior y evaluar si su ejecución es correcta. Este es otro de los script candidatos a ser planificados de manera mensual, diaria o según se necesite realizar la distribución de informes.

Además, se ha preparado el informe plantilla para ser instanciado y distribuido de forma automática "Resumen mensual del establecimiento". El detalle de dicho informe puede encontrase en el apartado **6.4 Distribución de informes**.

5.5. Consideraciones relevantes acerca del desarrollo

A continuación, se enumeran los apartados más relevantes a comentar relacionados con la implementación del almacén de datos.

5.5.1 Dimensiones autogeneradas de manera incremental

Los procesos ETL son capaces de **poblar las dimensiones del sistema de manera automática a partir de los datos que se reciben diariamente**, asignándoles codificación incremental para reducir el espacio de almacenamiento en disco (ya que esta codificación es numérica) y para mejorar el rendimiento de las consultas. Las dimensiones en las que se ha utilizado esta técnica son las siguientes:

- DWD_SOCIO
- DWD_TIPO_PRODUCTO
- DWD_PROVEEDOR

5.5.2 Punto unificado de validación de registros

Las cargas de productos y de ventas incluyen al principio de la transacción un paso de "Valor Java Script Modificado" que incluye toda la lógica de validación y descartes:

De esta manera se consigue **descartar la información de manera temprana** para no realizar procesamientos en vano, y tener identificado el lugar donde se realiza dicha validación. Los descartes consisten en la generación de un fichero *.csv, que incluye todos los campos que se reciben del cliente, más una columna extra en la que se indican tantos incumplimientos como en los que se vea afectado el registro.

Esto permite al usuario revisar dicha información, modificarla y guardarla con un nombre válido y así el sistema lo procesará directamente en la siguiente ejecución.

Cabe destacar las siguientes validaciones:

- Productos

    - Registros sin código de producto
    - ID de producto no numérico
    - Precios de venta, de coste e IVA sin informar
    - Precios de venta, de coste e IVA con valores más largos de lo permitido o no numérico
- Ventas
    - Registro de detalle sin correspondencia con registro padre de la venta
    - Registro padre de la venta sin ningún registro de detalle relacionado
    - Fecha de la venta incorrecta

5.5.3 Evitar realizar actualizaciones en BD si no es necesario

Antes de realizar una actualización en BD, debido a la necesidad por cambio de un producto, por ejemplo, **se comprueba que alguno de los campos de dicho producto ha cambiado**. Si esto no es así, no se realiza dicha actualización, evitando actividad innecesaria en la BBDD.

5.5.4 Recuperación ante errores

El sistema de procesos ETL está diseñado para tolerar los errores en el procesamiento de los ficheros de manera que sea capaz de pasar al siguiente fichero en caso de error. De esta manera, y ya que las cargas de estos sistemas suelen tomarse mucho tiempo, conseguimos que al día siguiente se haya procesado el mayor número de información posible. **Los ficheros que generan error, son renombrados** con la extensión *.ER (y los que han sido correctos *.OK) para que no sean procesados la próxima vez que se lancen.

5.5.5 Envío de correo con el resultado de las ejecuciones

Al terminar las ejecuciones con las cargas de productos y ventas, el sistema envía un correo desde la cuenta de Gmail **"gldp.tfc.uoc.gmail.com"** (a sí misma) con el resultado de la ejecución de los procesos indicando tanto si ha ido bien la cadena como si ha ido mal y adjuntando un fichero con el resumen de los ficheros procesados y de su estado.

5.5.6 *"Tuning"* de procesos (mejora en el rendimiento)

Como se ha comentado anteriormente, se han conseguido altos grados de rendimiento en los procesos de carga, llegándose a procesar y cargar el conjunto completo de información **en menos de 30 minutos.**

Vista la posibilidad de reducir los periodos de carga de manera tan considerable, se ha preferido invertir horas de trabajo en conseguir dicha mejora, a realizar procesos específicos que se encarguen de actualizar las tablas de hechos, cuando alguna de las dimensiones se ven afectadas por cambios.

Con los tiempo obtenidos se ha considerado menos costoso realizar una recarga de toda la información disponible. Si en el futuro el volumen de datos lo requiriera, se implementarían mecanismos para llevar de la manera más eficiente este tipo de modificaciones.

Para conseguir la eficiencia comentada anteriormente se han realizado los siguientes ajustes y mejoras:

5.5.6.1. Descarte temprano de la información incorrecta

Además de lo comentado en el punto anterior en el que las validaciones de información se realizaban lo antes posible, rechazando la información incorrecta y no dejando pasar el flujo a operaciones posteriores, los ficheros que no cumplían una nomenclatura correcta y que por lo tanto no podrían aportar valor al almacén de datos, se descartaban inmediatamente y se renombraban como erróneos, realizando una única búsqueda por fichero. En este escenario se encuadran los siguientes:

- Ficheros de productos, con año inválido, campo imprescindible para poder identificar de manera unívoca el producto del que se trata para poder asignarlo a la venta
- Ficheros de ventas de establecimientos que no estén inventariados en el sistema.

5.5.6.2. Utilización de un área de precarga

Uno de los cuellos de botella que primero fue detectado durante el análisis de los datos, era el que se originaba de la actividad de realizar *"join"* con los pasos de Spoon. Este problema fue solventado, creando dos tablas auxiliares, las cuales contenían un índice sobre el campo id_venda y que se encargarían de almacenarían la información de venta y de detalle de venta por separado. A continuación en un proceso posterior, se realizaba la descarga de la información cargada en estas tablas auxiliares utilizando el mecanismo de **"join"** que ofrece MySQL. Esto permitió que los nuevos tiempos de carga fueran 10 veces más rápidos únicamente aplicando esta mejora.

5.5.6.3. Mejoras en las búsquedas de información relacionada (lookups)

Activando las diferentes opciones de cache que incluyen los pasos de búsquedas, se ha conseguido reducir el tiempo de procesamiento casi 10 veces (acumulado a la mejora anterior). La siguiente hoja Excel contiene dos pestañas (TIEMPO ORIGINAL y TIEMPO TUNING) en el que se puede ver el detalle de cómo afecta dicha optimización en el rendimiento de la transformación de carga de uno de los ficheros de ventas.

Tiempo y Mejoras_lookup.xls

5.5.7 Utilización de amplio abanico de funcionalidades en los informes

El conjunto de informes preparado ha intentado utilizar el mayor conjunto de funcionalidad posible (teniendo en cuenta el tiempo disponible). El conjunto de componentes a resaltar es el siguiente:

- Componentes de gráficos diferentes (barras, barras y líneas, de sectores),
- Etiquetas de código de barras para los productos
- Utilización de diferentes tipología de parámetros (de origen de datos tabla, origen jdbc, de formato lista, formato conjunto de opciones, etc…)
- Utilización de **"subreport"** para representar información de varias consultas en el mismo informe y sección y en posiciones diferentes

5.5.8 Distribución de informes

Según lo comentado en el apartado **5.4 Distribución de informes**, se ha preparado el informe **6.4 Distribución de informes**, el cual es instanciado y generado en formato *pdf, para cada uno de los establecimientos de la tabla DWD_ESTABLECIMEINTO que dispongan de correo electrónico en el la columna EMAIL., Finalmente, mediante la cuenta de correo gldp.tfc.uoc@gmail.com, se procede a enviar a las diferentes direcciones indicadas en la columna referenciada anteriormente.

5.5.9 Automatización procesos

El producto *Pentaho Data Integration* proporciona las utilidades *Pan y Kitchen*, las cuales ofrecen la capacidad de ejecutar transformaciones diseñadas con *Spoon* desde línea de comandos. Esta funcionalidad, junto con la proporcionada por el comando "at" del propio del sistema operativo Windows XP, permite la planificación de ejecuciones en intervalos de tiempo regulares.

En nuestro proyecto, y ya que es previsible que la explotación de estos sistemas se realice en horario de oficina, podría ser interesante planificar las cargas de manera periódica, de tal manera que cada noche se procesasen todos los ficheros que estuviesen alojados en el directorio definido como de entrada al sistema. Siguiendo este procedimiento, al día siguiente la información estaría consolidada y disponible para ser analizada mediante el aplicativo de análisis y *reporting*.

El detalle de los script objeto a planificar queda recogido en el apartado **5.2.2.3 Procesos generales**, y como se observa en dicho apartado, estos scripts informan del resultado de su ejecución (tanto si es correcta como si no lo es) mediante el envío de un correo electrónico, lo cual ayuda a la supervisión de los procesos automáticos.

Por último comentar, que tal y como se comenta en el apartado **5.4 Distribución de informes**, el sistema incluye un módulo de generación dinámica de informes. Este módulo ha sido construido con la misma herramienta que el resto de la ETL, con la herramienta PDI de Pentaho, y gracias a la planificación comentada en el punto anterior para los procesos de carga diarios, es posible dotar al sistema de un planificador de informes.

Este planificador permite que diferentes destinatarios, como pueden ser los diferentes responsables de los establecimientos o directivos de la cadena, reciban de manera automática informes que les ayuden a realizar el seguimiento de los datos más relevantes de una manera más cómoda.

6. Informes realizados

A continuación se engloban en tres apartados el conjunto de informes predefinidos a aportar junto con la entrega del aplicativo.

Este conjunto de informes es una aportación inicial, que tiene como objetivo que el cliente comprenda y experimente las capacidades y la tipología de informes que se pueden obtener con el almacén de datos implementado.

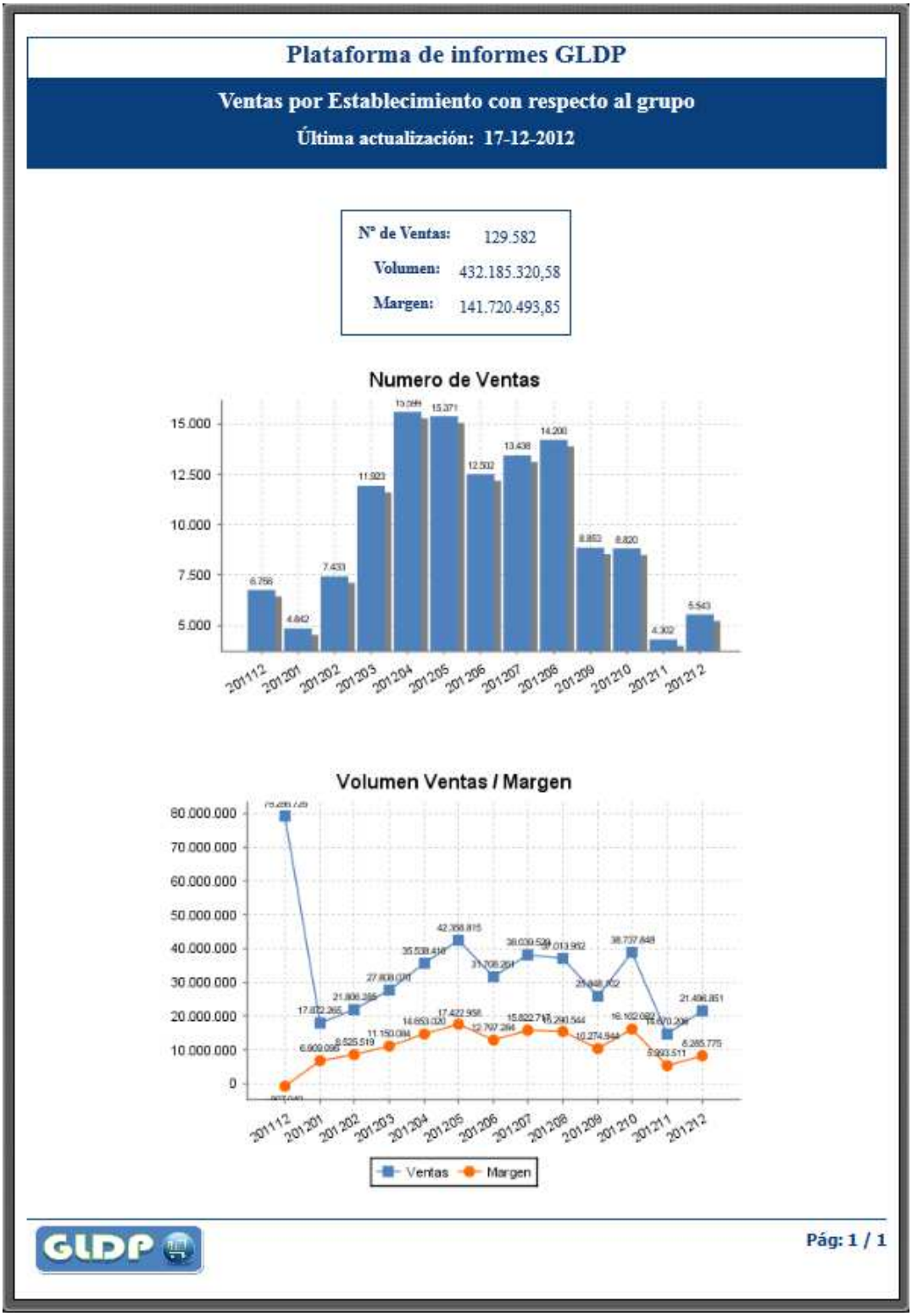
La experiencia dice que una vez que se le muestra el primer conjunto de informes, es el mismo cliente el que los evoluciona, pasándose por lo tanto a una segunda fase del proyecto, en la que si el almacén de datos está correctamente desarrollado, requerirá únicamente esfuerzo y desarrollo de la parte de informes.

Los indicadores generales que van a ser utilizados a lo largo de los informes. El motivo de explicarlo en este punto global es que su cálculo es común y siempre el mismo, diferenciando en cada uno de ellos las dimensiones o criterios de análisis para los que se muestran:

- Nº de Ventas: Representa el número diferente de ventas para el periodo analizado (distintos IDs de venta)
- Volumen: Es el importe acumulado resultante de multiplicar por cada línea de venta, el importe del precio de venta del producto por la cantidad de unidades vendidas
- Margen de productos: Este dato se consigue restando al volumen del punto anterior, el importe acumulado de multiplicar por cada línea de venta, el importe del precio de coste del producto por la cantidad de unidades vendidas.
- Margen: Para el cálculo global del margen es necesario que por cada mes, se reste también el coste originado por los gastos de los establecimientos los cuales incluyen mermas, costes de personal, consumo energético, seguridad, mantenimiento....
- Número de unidades vendidas: sumatorio del número de productos vendidos en todas las ventas involucradas en el análisis realizado.

6.1. Ventas

Figura 10: Total de ventas y margen neto del grupo



Explicación:

Se trata del informa más general preparado, y muestra el total de ventas y margen neto del grupo. Para ello utiliza los últimos 13 meses de información del sistema con el objetivo de poder comparar el mes actual con su correspondiente del año pasado.

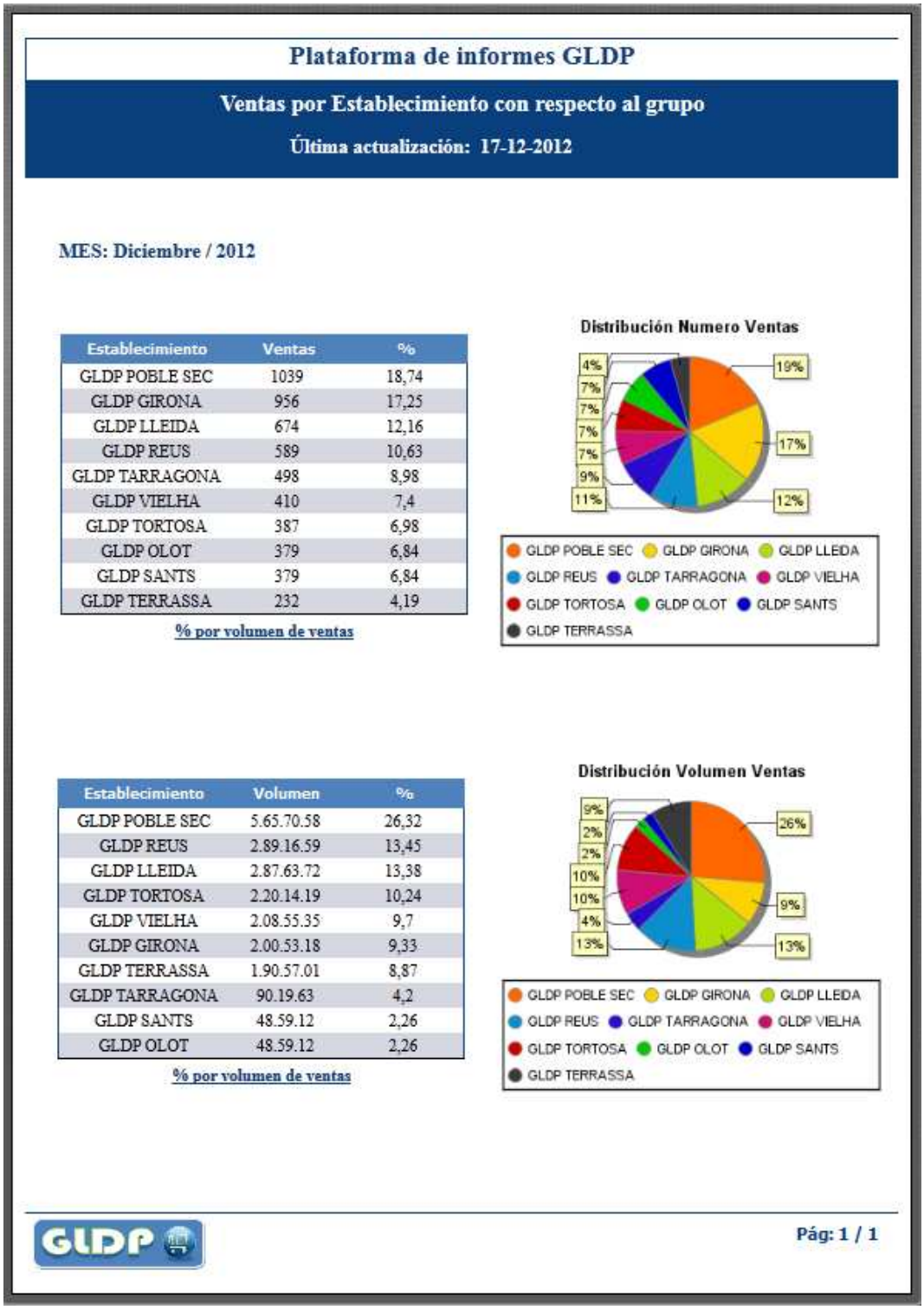
Los principales datos mostrados son los siguientes:

- Cuadro resumen de ventas con los siguientes indicadores:
  - Nº de Ventas acumulado para los 13 meses que se muestra el informe.
  - Volumen de ventas acumulado para los 13 meses que se muestra el informe.
  - Margen en las ventas acumulado para los 13 meses que se muestra el informe.
- Gráfico Número de Ventas:

  Gráfico de barras que muestra por mes, el indicador Nº de Ventas
- Gráfico Volumen Ventas / Margen:

  Gráfico de líneas y barras que muestran por mes, los indicadores Volumen y Margen según lo comentado en el apartado anterior

Figura 11: % de ventas respecto el total del grupo (por establecimiento)



Plataforma de informes GLDP

Ventas por Establecimiento con respecto al grupo

Última actualización: 17-12-2012

MES: Diciembre / 2012

Establecimiento	Ventas	%
GLDP POBLE SEC	1039	18,74
GLDP GIRONA	956	17,25
GLDP LLEIDA	674	12,16
GLDP REUS	589	10,63
GLDP TARRAGONA	498	8,98
GLDP VIELHA	410	7,4
GLDP TORTOSA	387	6,98
GLDP OLOT	379	6,84
GLDP SANTS	379	6,84
GLDP TERRASSA	232	4,19

% por volumen de ventas

Distribución Numero Ventas

Establecimiento	Volumen	%
GLDP POBLE SEC	5.65.70.58	26,32
GLDP REUS	2.89.16.59	13,45
GLDP LLEIDA	2.87.63.72	13,38
GLDP TORTOSA	2.20.14.19	10,24
GLDP VIELHA	2.08.55.35	9,7
GLDP GIRONA	2.00.53.18	9,33
GLDP TERRASSA	1.90.57.01	8,87
GLDP TARRAGONA	90.19.63	4,2
GLDP SANTS	48.59.12	2,26
GLDP OLOT	48.59.12	2,26

% por volumen de ventas

Distribución Volumen Ventas

GLDP

Pág: 1 / 1

Explicación:

El informe permite analizar las ventas de cada establecimiento con respecto el total del grupo. Para ello se muestran dos partes diferenciadas: el análisis del número de ventas y el análisis por volumen de ventas, ya que según el criterio la posición del establecimiento puede variar frente al resto.

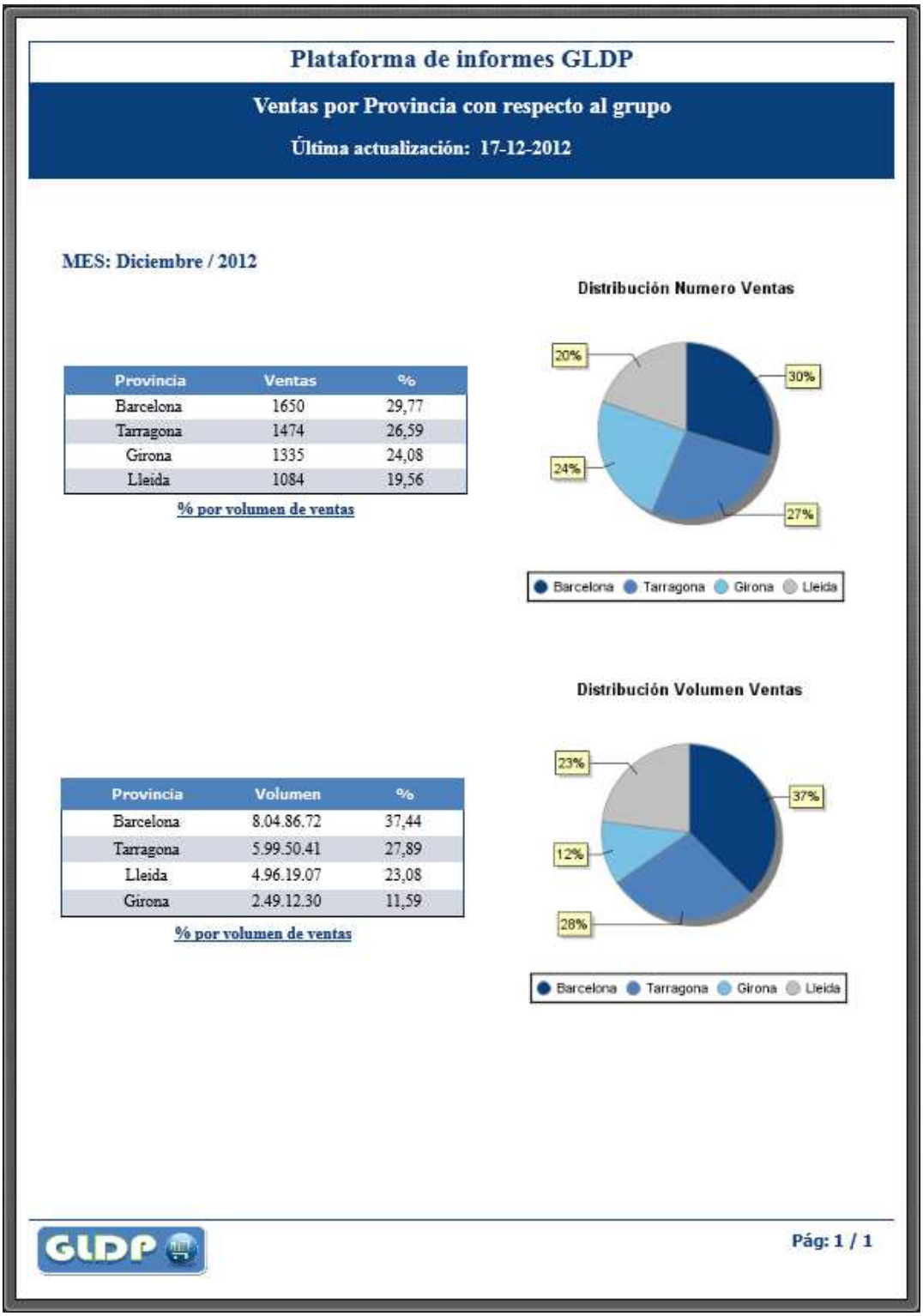
Los filtros a seleccionar en este informe son:

- Mes En formato (YYYYMM): Permite discriminar el mes para el que se quieren analizar los datos de los establecimientos. Por defecto el sistema muestra el año actual.

Los principales datos mostrados son los siguientes:

- Tabla para el análisis de el número de ventas, ordenada por dicho criterio:
  - Establecimiento: Nombre del establecimiento.
  - Ventas: Número de ventas del establecimiento
  - %: Porcentaje que representa el número de ventas del establecimiento frente al total formado por las ventas de todos los establecimientos.
- Gráfico Distribución Numero Ventas: Muestra el porcentaje de ventas de cada uno de los establecimientos utilizando un gráfico de sectores.
- Tabla para el análisis del volumen de ventas, ordenada por dicho criterio y que muestra las siguientes columnas:
  - Establecimiento: Nombre del establecimiento.
  - Volumen: Acumulado del importe de ventas por establecimiento
  - %: Porcentaje que representa volumen de ventas del establecimiento frente al total formado por todo el acumulado de ventas de todos los establecimientos.
- Gráfico Distribución Volumen Ventas: Muestra el porcentaje del volumen de ventas de cada uno de los establecimientos utilizando un gráfico de sectores.

Figura 12: % de ventas respecto el total del grupo (por demarcación)



Provincia	Ventas	%
Barcelona	1650	29,77
Tarragona	1474	26,59
Girona	1335	24,08
Lleida	1084	19,56

% por volumen de ventas

Provincia	Volumen	%
Barcelona	8.04.86.72	37,44
Tarragona	5.99.50.41	27,89
Lleida	4.96.19.07	23,08
Girona	2.49.12.30	11,59

% por volumen de ventas

Explicación:

El informe permite analizar las ventas de cada provincia con respecto el total del grupo. Para ello se muestran dos partes diferenciadas: el análisis del número de ventas y el análisis por volumen de ventas, ya que según el criterio, la posición de la provincia puede variar frente al resto.

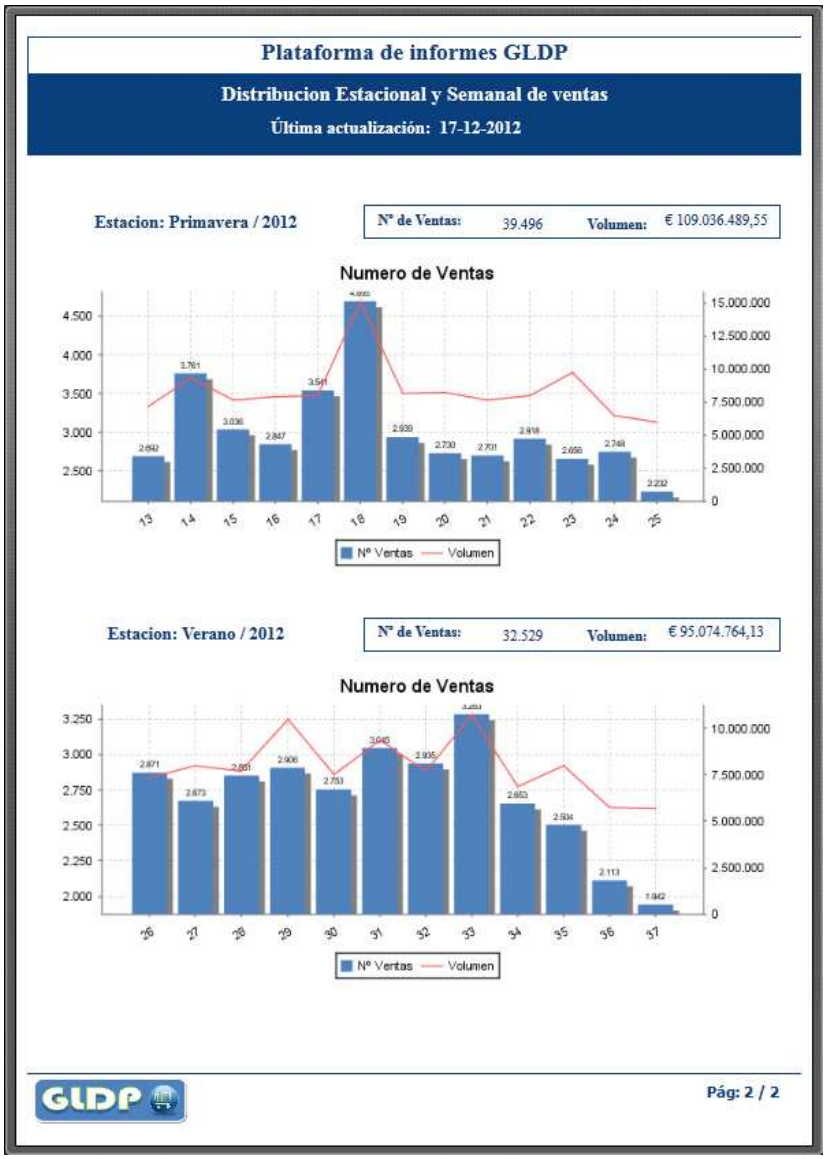
Los filtros a seleccionar en este informe son:

- Mes En formato (YYYYMM): Permite discriminar el mes para el que se quieren analizar los datos de las provincias. Por defecto el sistema muestra el mes actual.

Los principales datos mostrados son los siguientes:

- Tabla para el análisis de el número de ventas, ordenada por dicho criterio y que muestra las siguientes columnas:
    - Provincia: Nombre de la provincia.
    - Ventas: Número de ventas producidas en los establecimientos de esas provincias
    - %: Porcentaje que representa el número de ventas de la provincia frente al total formado por las ventas de todas las provincias.
- Gráfico Distribución Numero Ventas: Muestra el porcentaje de ventas de cada una de las provincias utilizando un gráfico de sectores.
- Tabla para el análisis del volumen de ventas, ordenada por dicho criterio:
    - Provincia: Nombre de la provincia.
    - Volumen: Acumulado del importe de ventas de los establecimientos de esas provincias
    - %: Porcentaje que representa volumen de ventas de la provincia frente al total formado por las ventas de todas las provincias.
- Gráfico Distribución Numero Ventas: Muestra el porcentaje del volumen de ventas de cada una de las provincias utilizando un gráfico de sectores.

Figura 13: Distribución semanal y estacionalidad de ventas



Explicación:

El informe permite analizar la distribución semanal y estacional de las ventas del grupo. Para ello se muestran los datos divididos por estación (Primavera, Verano, Otoño e Invierno) y para cada una de estas estaciones los datos de las ventas por las semanas del año incluidas en dicha estación.

Los filtros a seleccionar en este informe son:

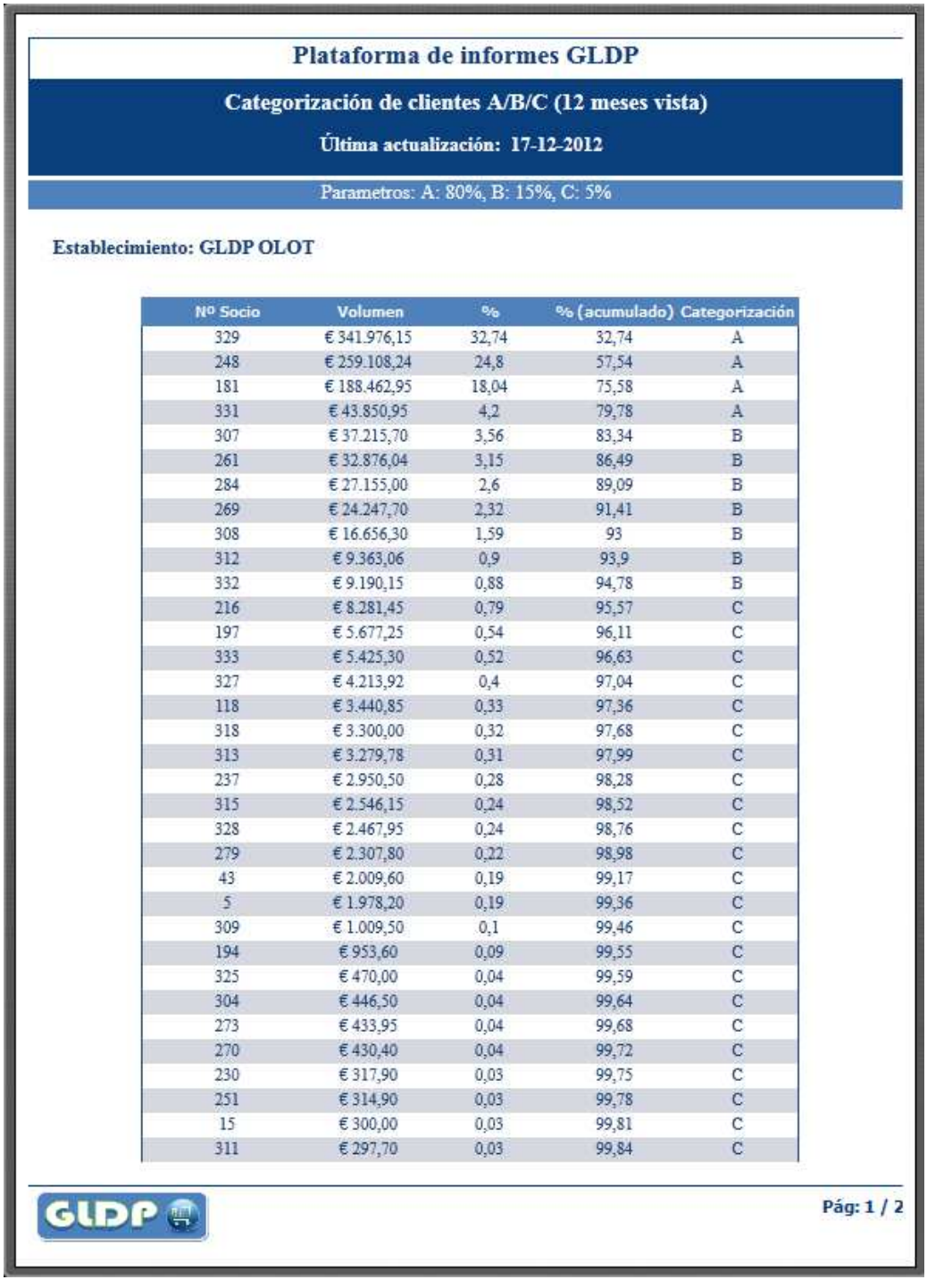
- Año. En formato (YYYY): Permite discriminar el año para el que se quieren analizar los datos. Por defecto el sistema muestra el año actual.

Los principales datos mostrados son los siguientes:

- Sección por Estación del año. Donde para cada estación del año se muestra:
  - Resumen de los indicadores de Nº de ventas y Volumen totales para dicha estación
  - Gráfico de Barra y líneas en el que para cada semana se muestra el Nº de ventas y el Volumen de ventas.

6.2. Clientes

Figura 14: Categorización de Clientes A/B/C



Plataforma de informes GLDP

Categorización de clientes A/B/C (12 meses vista)

Última actualización: 17-12-2012

Parametros: A: 80%, B: 15%, C: 5%

Establecimiento: GLDP OLOT

Nº Socio	Volumen	%	% (acumulado)	Categorización
329	€ 341.976,15	32,74	32,74	A
248	€ 259.108,24	24,8	57,54	A
181	€ 188.462,95	18,04	75,58	A
331	€ 43.850,95	4,2	79,78	A
307	€ 37.215,70	3,56	83,34	B
261	€ 32.876,04	3,15	86,49	B
284	€ 27.155,00	2,6	89,09	B
269	€ 24.247,70	2,32	91,41	B
308	€ 16.656,30	1,59	93	B
312	€ 9.363,06	0,9	93,9	B
332	€ 9.190,15	0,88	94,78	B
216	€ 8.281,45	0,79	95,57	C
197	€ 5.677,25	0,54	96,11	C
333	€ 5.425,30	0,52	96,63	C
327	€ 4.213,92	0,4	97,04	C
118	€ 3.440,85	0,33	97,36	C
318	€ 3.300,00	0,32	97,68	C
313	€ 3.279,78	0,31	97,99	C
237	€ 2.950,50	0,28	98,28	C
315	€ 2.546,15	0,24	98,52	C
328	€ 2.467,95	0,24	98,76	C
279	€ 2.307,80	0,22	98,98	C
43	€ 2.009,60	0,19	99,17	C
5	€ 1.978,20	0,19	99,36	C
309	€ 1.009,50	0,1	99,46	C
194	€ 953,60	0,09	99,55	C
325	€ 470,00	0,04	99,59	C
304	€ 446,50	0,04	99,64	C
273	€ 433,95	0,04	99,68	C
270	€ 430,40	0,04	99,72	C
230	€ 317,90	0,03	99,75	C
251	€ 314,90	0,03	99,78	C
15	€ 300,00	0,03	99,81	C
311	€ 297,70	0,03	99,84	C

GLDP

Pág: 1 / 2

Explicación:

El informe permite analizar la importancia de los clientes para el grupo realizando una categorización ABC sobre los mismos, utilizando para ello el indicador de Volumen de ventas para los últimos 12 meses. El informe ofrece la posibilidad de que los porcentajes para el estudio ABC sean introducidos por el usuario, ofreciendo por defecto los porcentajes de 80%, 15% y 5%.

Además, debido a que la gran mayoría de las ventas pertenecen a clientes “No Socios” (socio 0 o sin identificar), el informe permitirá al usuario discriminar si quiere realizar el análisis para todos los clientes, o no incluir los “No Socios” en el mismo. Esta última opción es la establecida por defecto.

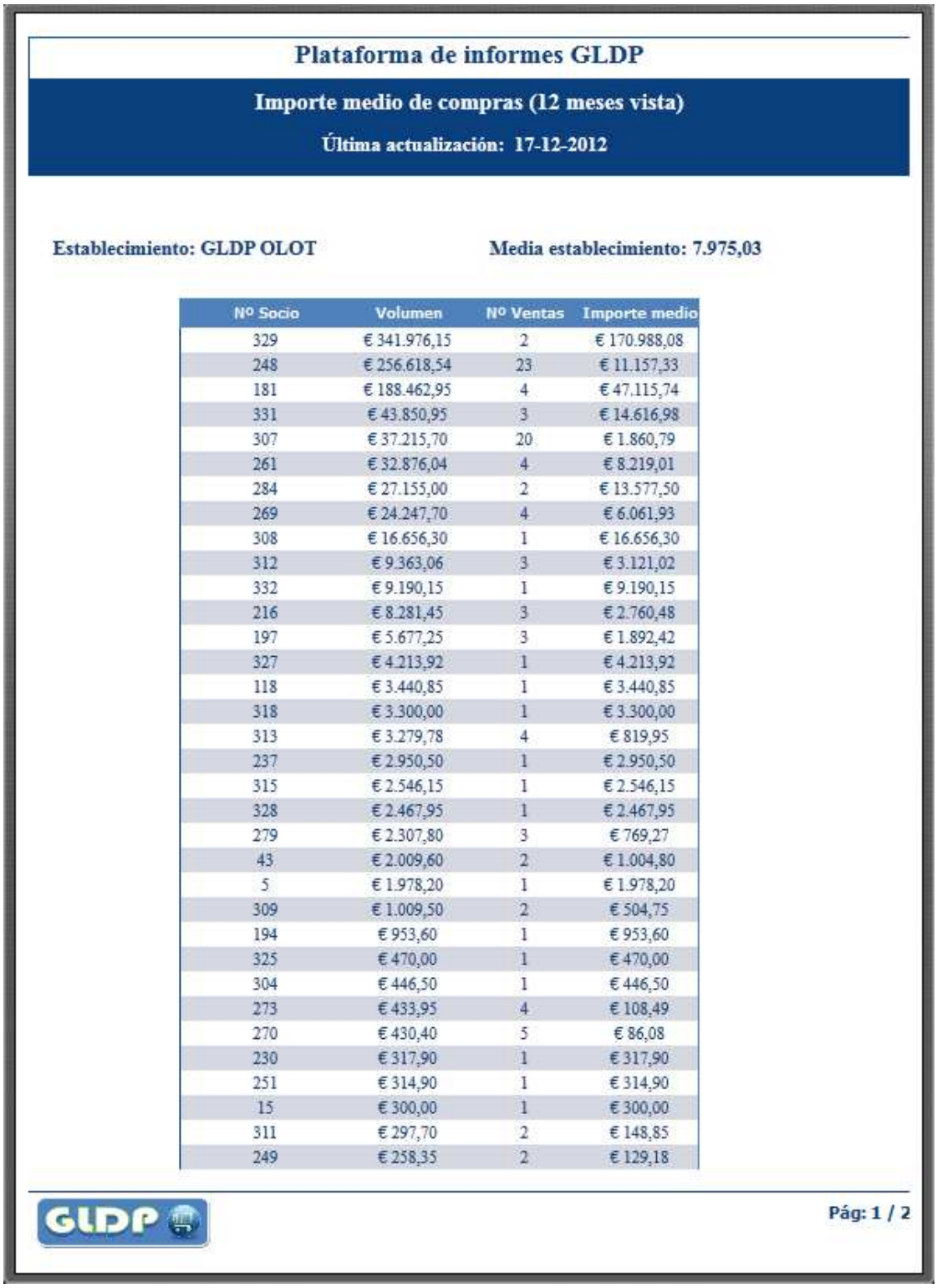
Los filtros a seleccionar en este informe son:

- Categorizar no socios (socio Nº 0): Conjunto de opciones que permite incluir/excluir en el análisis las ventas realizadas por clientes que no son socios del establecimiento.
- Porcentajes para la categorización ABC. Mediante dos cajas de texto se le permite al usuario introducir porcentajes y según estos variar el análisis ABC a realizar en función de ellos. El sistema ofrece únicamente la posibilidad de introducir los dos primeros, ya que se sobreentiende que el tercer porcentaje es la diferencia con el valor 100%.
- Establecimiento: Selecciona el establecimiento sobre el que se realiza el análisis, ya que solo socios son específicos de cada establecimiento.

Los principales datos mostrados son los siguientes:

- Título con los parámetros seleccionados para el análisis ABC.
- Sección con el nombre del Establecimiento seleccionado.
- Tabla de categorización que muestra la siguiente información:
    - Nº de Socio: Identificación del cliente para ese establecimiento.
    - Volumen: volumen de ventas realizado para ese cliente
    - %: Porcentaje de ventas de ese cliente con respecto al total de ventas del establecimiento.
    - % (acumulado): Acumulado de la columna anterior para realizar la categorización
    - Categorización: Clasifica al cliente como A, B o C dependiendo de % acumulado que tenga.

Figura 15: Importe medio de compra por socio y establecimiento



Plataforma de informes GLDP

Importe medio de compras (12 meses vista)

Última actualización: 17-12-2012

Establecimiento: GLDP OLOT

Media establecimiento: 7.975,03

Nº Socio	Volumen	Nº Ventas	Importe medio
329	€ 341.976,15	2	€ 170.988,08
248	€ 256.618,54	23	€ 11.157,33
181	€ 188.462,95	4	€ 47.115,74
331	€ 43.850,95	3	€ 14.616,98
307	€ 37.215,70	20	€ 1.860,79
261	€ 32.876,04	4	€ 8.219,01
284	€ 27.155,00	2	€ 13.577,50
269	€ 24.247,70	4	€ 6.061,93
308	€ 16.656,30	1	€ 16.656,30
312	€ 9.363,06	3	€ 3.121,02
332	€ 9.190,15	1	€ 9.190,15
216	€ 8.281,45	3	€ 2.760,48
197	€ 5.677,25	3	€ 1.892,42
327	€ 4.213,92	1	€ 4.213,92
118	€ 3.440,85	1	€ 3.440,85
318	€ 3.300,00	1	€ 3.300,00
313	€ 3.279,78	4	€ 819,95
237	€ 2.950,50	1	€ 2.950,50
315	€ 2.546,15	1	€ 2.546,15
328	€ 2.467,95	1	€ 2.467,95
279	€ 2.307,80	3	€ 769,27
43	€ 2.009,60	2	€ 1.004,80
5	€ 1.978,20	1	€ 1.978,20
309	€ 1.009,50	2	€ 504,75
194	€ 953,60	1	€ 953,60
325	€ 470,00	1	€ 470,00
304	€ 446,50	1	€ 446,50
273	€ 433,95	4	€ 108,49
270	€ 430,40	5	€ 86,08
230	€ 317,90	1	€ 317,90
251	€ 314,90	1	€ 314,90
15	€ 300,00	1	€ 300,00
311	€ 297,70	2	€ 148,85
249	€ 258,35	2	€ 129,18

GLDP

Pág: 1 / 2

Explicación:

El informe permite analizar la el comportamiento de las ventas en los diferentes establecimientos mediante la presentación del Volumen de ventas de cada socio, y de la media de dicho indicador, para los últimos 12 meses. El informe ofrece la posibilidad, debido a que la gran mayoría de las ventas pertenecen a clientes “No Socios” (socio 0 o sin identificar), de que el usuario discrimine si quiere realizar el análisis para todos los clientes, o no incluir los “No Socios” en el mismo. Esta última opción es la establecida por defecto.

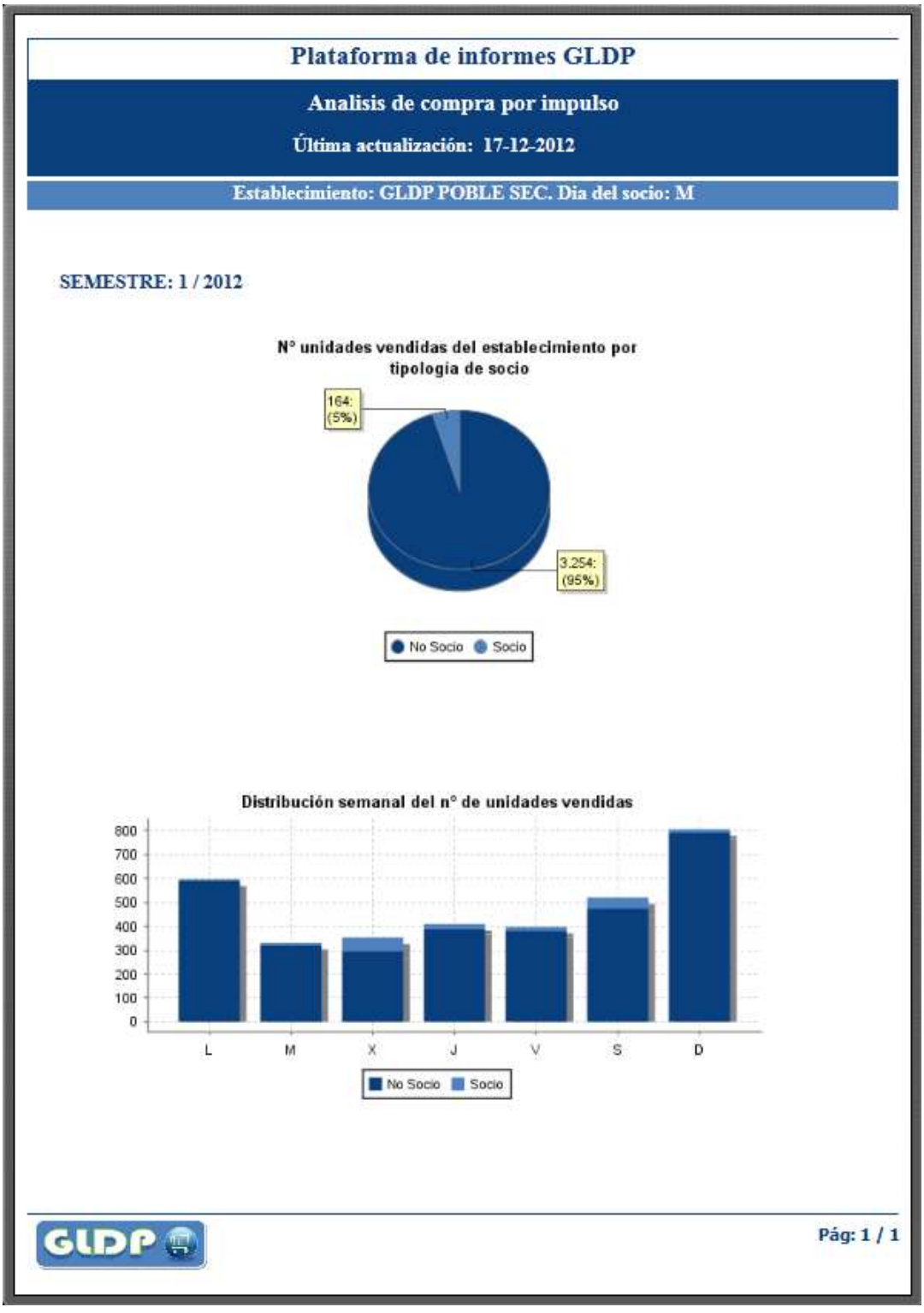
Los filtros a seleccionar en este informe son:

- Categorizar no socios (socio Nº 0): Conjunto de opciones que permite incluir/excluir en el análisis del importe medio de compras los clientes que no son socios del establecimiento.
- Establecimiento: Selecciona el establecimiento sobre el que se realiza el análisis, ya que solo socios son específicos de cada establecimiento.

Los principales datos mostrados son los siguientes:

- Sección con el nombre del Establecimiento seleccionado.
- Media establecimiento: Media formada por el importe medio de compra de todo el establecimiento
- Tabla de importes medios de compras:
    - Nº de Socio: Identificación del cliente para ese establecimiento.
    - Volumen: volumen de ventas realizado para ese cliente
    - Nº de Ventas: Número de ventas realizadas por el cliente
    - Importe medio: Importe medio calculado de la división de las columnas anteriores.

Figura 16: Análisis de compra por impulso



Explicación:

El informe permite analizar las ventas de productos señalados como “de impulso”. Lo que se pretende con este informe es comprobar si los socios son más sensibles a esta iniciativa que el resto de clientes analizando la distribución semanal de las ventas de estos productos de tal manera que exista alguna correlación entre el importe de las ventas y el hecho de contener productos "de impulso".

Para ello lo que se ofrece es un informe que analiza por semestre la actividad de este tipo de compras por establecimiento, primeramente a nivel global entre socios y no socios, para luego realizar la misma comparativa pero por día de la semana para ver si existe algún cambio en el día del socio del establecimiento.

Los filtros a seleccionar en este informe son:

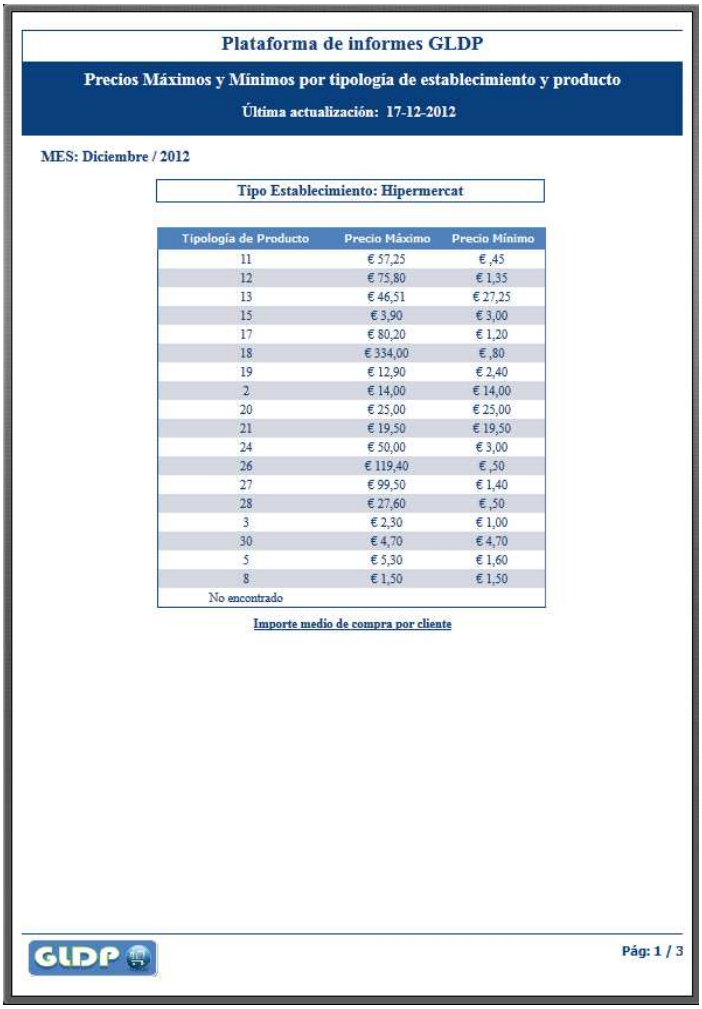
- Año: En formato (YYYY): Año para el que se quiere realizar el análisis de los datos.
- Semestre: Semestre del año del apartado anterior para el que se realizará el análisis
- Establecimiento: Selecciona el establecimiento sobre el que se realiza el análisis, ya que cada establecimiento tiene un día del socio distinto.

Los principales datos mostrados son los siguientes:

- Nombre del establecimiento seleccionado y día del socio en dicho establecimiento
- Sección con el semestre y año seleccionado.
- Grafico de unidades vendidas por tipología de socio: Indica un gráfico de sector que representa el porcentaje de ventas de productos “por impulso” de los clientes que son socios frente de los que no lo son
- Gráfico Distribución semanal del nº de unidades vendidas Gráfico de barras apiladas que representa las ventas de los clientes que son socios frente de los que no lo son para cada uno de los días de la semana.

6.3. Productos

Figura 17: Precios máximos y mínimos por tipo de establecimiento y Tipología de producto.



Plataforma de informes GLDP

Precios Máximos y Mínimos por tipología de establecimiento y producto

Última actualización: 17-12-2012

MES: Diciembre / 2012

Tipo Establecimiento: Hipermercat

Tipología de Producto	Precio Máximo	Precio Mínimo
11	€ 57,25	€ ,45
12	€ 75,80	€ 1,35
13	€ 46,51	€ 27,25
15	€ 3,90	€ 3,00
17	€ 80,20	€ 1,20
18	€ 334,00	€ ,80
19	€ 12,90	€ 2,40
2	€ 14,00	€ 14,00
20	€ 25,00	€ 25,00
21	€ 19,50	€ 19,50
24	€ 50,00	€ 3,00
26	€ 119,40	€ ,50
27	€ 99,50	€ 1,40
28	€ 27,60	€ ,50
3	€ 2,30	€ 1,00
30	€ 4,70	€ 4,70
5	€ 5,30	€ 1,60
8	€ 1,50	€ 1,50
No encontrado		

Importe medio de compra por cliente

GLDP

Pág: 1 / 3

Explicación:

El informe permite analizar los precios máximos y mínimos de las ventas por tipología de establecimiento y tipología de producto. Para ello se presenta para cada tipología de establecimiento, una tabla que recoge las tipologías de productos vendidas en el mismo, y para cada una de éstas, el importe máximo y mínimo de producto vendido.

Los filtros a seleccionar en este informe son:

- Mes: En formato (YYYYMM): Mes para el que se quiere realizar el análisis de los datos.

Los principales datos mostrados son los siguientes:

- Sección con mes y año seleccionado.
- Subsección que discrimina entre las diferentes tipologías de supermercados
- Tabla de tipología de productos con las siguientes columnas:
    - Tipología de producto: Nombre de la tipología.
    - Precio Máximo: Precio del producto vendido de mayor importe dentro de la tipología
    - Precio Mínimo: Precio del producto vendido de menor importe dentro de la tipología

Figura 18: "*Top ten" de productos

Plataforma de informes GLDP

Top Ten de productos por ventas y margen

Última actualización: 17-12-2012

MES: Diciembre / 2012

Codigo Barras	ID Producto	Descripcion	Número Unidades
	375	CHAMPÚ CABELLO SECO SENSATIONS 400 ML	3.739.818
	350	CERA LÍQUIDA AUTOBRILLANTE SUELOS 1 L.	307.185
	336	CEPILLO DENTAL BLANQUEANTE SUAVE	2.721
	1198	VINO D.O. RIOJA TINTO VIÑA RÚSTICA, 75 CL.	1.582
	873	PALILLO REDONDO PALILLERO 200/U. APROX	950
	1180	VELAS CALIENTAPLATOS, 10 UND.	898
	169	ATÚN LISTADO EN ACEITE GIRASOL, 220 G. P.N.	883
	334	CEPILLO DENTAL BLANQUEANTE DURO	571
	443	CREMA DENTAL TRIPLE ACCIÓN, 75 ML.	406
	311	CARAMELOS FRUTA ÁCIDA, 1 K.	358

Top 10 de productos por número de unidades vendidas

Codigo Barras	ID Producto	Descripcion	Margen
	375	CHAMPÚ CABELLO SECO SENSATIONS 400 ML	€ 8.900.766,84
	350	CERA LÍQUIDA AUTOBRILLANTE SUELOS 1 L.	€ 559.076,70
	455	CREMA MANOS Nº 3 ANTIMANCHAS SENSATIONS..	€ 14.630,40
	48	AGUA CON GAS, 1,5 L.	€ 13.375,17
	309	CARAMELOS FRESA NATA, 90 G.	€ 8.565,15
	336	CEPILLO DENTAL BLANQUEANTE SUAVE	€ 3.646,14
	1180	VELAS CALIENTAPLATOS, 10 UND.	€ 3.035,24
	873	PALILLO REDONDO PALILLERO 200/U. APROX	€ 2.850,00
	308	CARAMELOS DUPLEX PICA REGALIZ, 150 G.	€ 2.403,00
	466	CUBO BASURA C/PEDAL, 25 L	€ 1.995,82

Top 10 de productos por margen obtenido

GLDP

Pág: 1 / 1

Explicación:

El informe permite analizar los diez mejores productos para el grupo, utilizando para ello los criterios de unidades vendidas y el margen del producto.

Los filtros a seleccionar en este informe son:

- Mes: En formato (YYYYMM): Mes para el que se quiere realizar el análisis de los datos.

Los principales datos mostrados son los siguientes:

- Sección con mes y año seleccionado.
- Tabla de "Top Ten" de productos con mayor número de unidades vendidas, con las siguientes columnas:
    - Código de barras: Imagen con el código de barras del producto para el facilitado de su captura.
    - ID Producto: Identificador del producto según las referencias del grupo (no código interno del almacén de datos)
    - Descripción: Descripción o nombre del producto
    - Número Unidades: Número de unidades vendidas del producto para el mes seleccionado
- Tabla de "Top Ten" de productos con mayor margen obtenido en las ventas, con las siguientes columnas:
    - Código de barras: Imagen con el código de barras del producto para el facilitado de su captura.
    - ID Producto: Identificador del producto según las referencias del grupo (no código interno del almacén de datos)
    - Descripción: Descripción o nombre del producto
    - Margen: Margen de producto obtenido en las ventas del producto para el mes seleccionado

Figura 19: % de ventas de "marcas blancas" por habitantes

Plataforma de informes GLDP

Consumo marcas blancas por año demarcación

Última actualización: 17-12-2012

AÑO: 2012

Demarcación	Unidades Marcas Blancas Vendidas	Unidades Totales Vendidas	Nº Habitantes	% Por habitante	% Sobre el total
Barcelona	9370	18303483	1615448	0,58%	0,05%
Girona	5521	9438938	96722	5,71%	0,06%
Tarragona	4981	6499328	134085	3,71%	0,08%
Olot	3600	4419906	33725	10,67%	0,08%
Lleida	2174	7455281	138416	1,57%	0,03%
Reus	2001	6827042	106709	1,88%	0,03%
Tortosa	1967	6681587	34432	5,71%	0,03%
Vielha e Mijaran	1489	5752660	5601	26,58%	0,03%
Terrassa	1439	2343271	213697	0,67%	0,06%
Figueres	0	1	44765	0,00%	0,00%

% de ventas de "marcas blancas" por habitantes y sobre el total

GLDP

Pág: 1 / 1

Explicación:

El informe permite analizar el consumo de productos catalogados como “Marcas Blancas” los cuales a priori son los fabricados por el propio grupo GLDP, esto significa que el proveedor de los productos es GLDP. Para ello el informe permite al usuario seleccionar entre el conjunto de proveedores que tiene inventariado en el sistema, dando por defecto el propio proveedor GLDP y a continuación ofreciendo una lista ordenada por nombre del resto de proveedores para permitir su selección múltiple. Esto es así, para dotar al informe de la capacidad de que si un proveedor establece una alianza con el grupo pasando a ser también proveedor de productos “Marcas Blancas”, o por el motivo que sea se quiere analizar la actividad de sus ventas, con este mismo informe, se pueda realizarse dicho análisis de manera conjunta con el proveedor GLDP, o de manera independiente.

Los filtros a seleccionar en este informe son:

- Año: En formato (YYYY): Año para el que se quiere realizar el análisis de los datos.
- Lista de proveedores: Permite seleccionar el conjunto de proveedores para los cuales sus productos son considerados como “Marcas blancas” por defecto el proveedor GLDP.

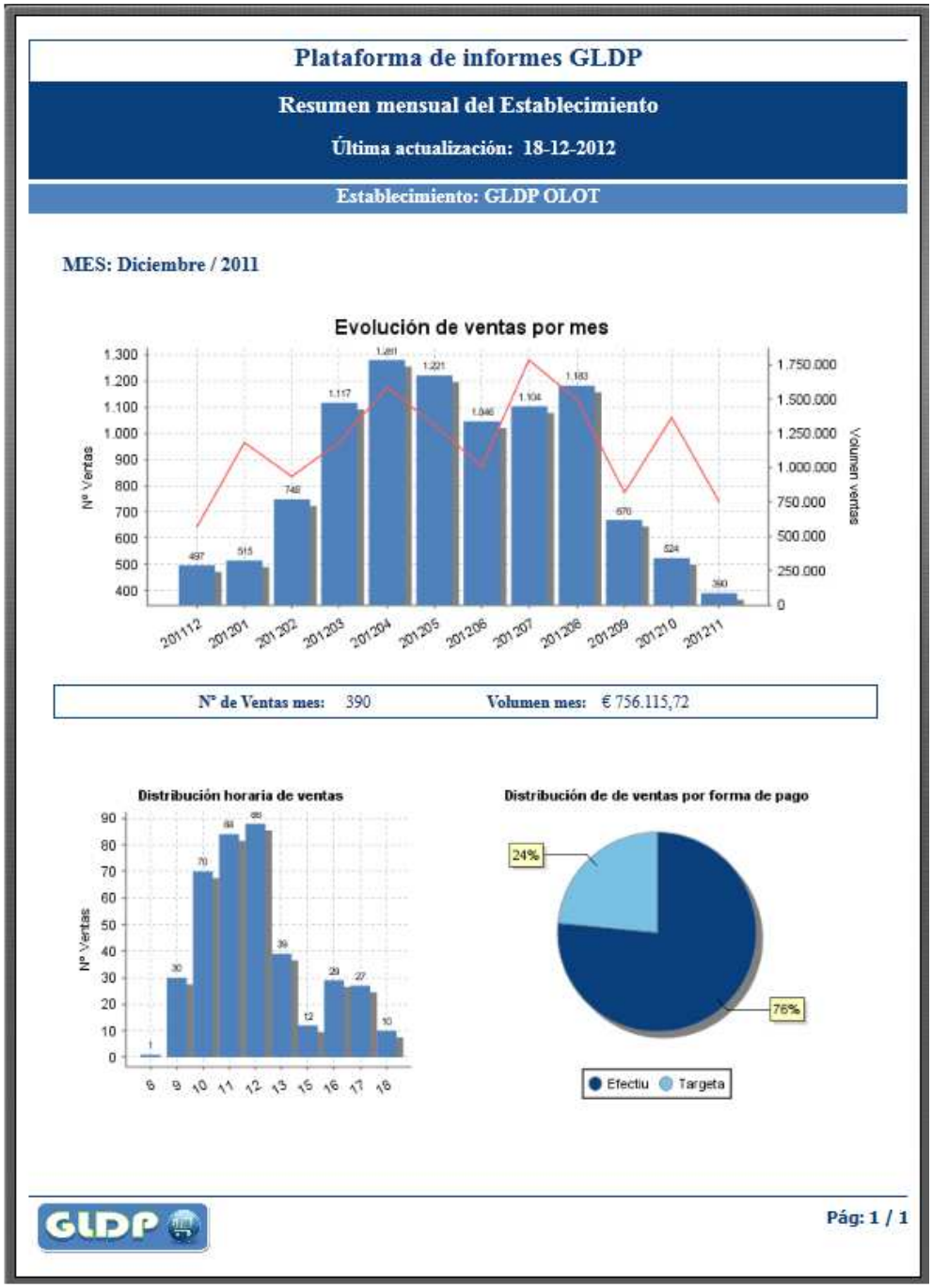
Los principales datos mostrados son los siguientes:

- Sección con el año seleccionado.
  Gráfico de barras que muestra para el año seleccionado el número de ventas por mes, dando una idea de la evolución en el consumo de estos productos.
- Tabla de ventas de “Marcas Blancas” por habitante de una determinada demarcación, con las siguientes columnas:
  - Demarcación: Demarcación para la que se analiza el consumo de “Marcas Blancas”
  - Unidades Marcas Blancas Vendidas: Número de unidades de estos productos vendidos para una demarcación en concreto en el año seleccionado
  - Unidades Totales Vendidas: Número de unidades totales (de cualquier producto) vendidas para una demarcación en concreto en el año seleccionado
  - Nº Habitantes: Número de habitantes que actualmente residen en la demarcación
  - % Por habitante: Porcentaje resultante de dividir el número de unidades vendías de productos “marcas Blancas” al año en la demarcación, por el número de habitantes
  - % Sobre el total: Porcentaje resultante de dividir el número de unidades vendías de productos “marcas Blancas” al año en la demarcación, entre el total de unidades vendidas

6.4. Distribución de informes

Este informe se ha generado para la distribución automática de informe que será explicada posteriormente en el apartado oportuno.

Figura 20: Resumen mensual del establecimiento



Explicación:

Este informe ha sido diseñado para demostrar las capacidades que ofrece el sistema en cuanto a la generación y distribución de informes, y cuyo detalle será explicado en el apartado correspondiente. En cuanto al significado del mismo no es más que un ejemplo de informe específico para un establecimiento que pretende reflejar otros datos que no han sido recogidos en el resto de informes (forma de pago, distribución horaria) además de ser para un establecimiento en concreto.

Los filtros a seleccionar en este informe son:

- ID de establecimiento: Se pide el código interno del sistema para este establecimiento. Este filtro cobrará sentido cuando se aborde el tema de la generación y distribución de informes que será explicada posteriormente.

Los principales datos mostrados son los siguientes:

- Sección con el mes y el año actual
  Gráfico de barras que muestra para los 12 meses anteriores a la fecha el número de ventas por mes del establecimiento seleccionado, dando una idea de la evolución de las mismas.
- Cuadro con resumen de datos del mes para el establecimiento seleccionado, que incluyen:
  - Nº de Ventas mes: Indica el número de ventas del mes anterior:
  - Volumen mes: Indica el volumen de ventas realizadas el mes anterior
- Gráfico de distribución horaria de ventas: Es un gráfico de barras que indica por horas el número de ventas producidas en el mes anterior, lo que ayuda a comprender los picos de carga de trabajo y por lo tanto a la elaboración de turnos del establecimiento
- Grafico de distribución de ventas por forma de pago: Se trata de un gráfico de sectores con la distribución de la forma de pago, lo cual permite saber al establecimiento si los clientes pagan más con tarjeta o en efectivo sus compras.

7. Conclusiones

Las principales conclusiones obtenidas una vez finalizado el proyecto/asignatura son las siguientes:

- Se ha conseguido sobradamente cumplir los objetivos planteados al inicio del semestre, tanto los de la asignatura, como los propios del trabajo de fin de carrera.

- El enfoque seguido para abordar el proyecto ha sido el adecuado

- La planificación ha sido la correcta y no ha habido grandes desviaciones con respecto la planteada inicialmente. Si bien es cierto que para la fase de implementación se hubieran querido abordar más funcionalidades, como la elaboración de algún cuadro de mando, que por falta de tiempo no ha podido desarrollarse.

- Un buen análisis de los datos recibidos y el cotejo entre dicha información y la que se solicita, es imprescindible para garantizar el éxito final del proyecto.

- A mi modo de ver, la parte más importante del proyecto consiste en la elaboración de un modelo de datos capaz de cubrir las necesidades de información solicitadas por el usuario. Esto es así, con independencia de lo compleja que se convierta la ETL encargada de poblarlo, ya que el tiempo y los recursos que se inviertan en esta parte, a mi juicio serán recompensados al disponer de un modelo fácil de explotar y rápido en cuanto a tiempos de respuesta se refiere. El modelo creado cumple estas características.

- Se ha realizado un gran esfuerzo, quizá demasiado, en la preparación de los procesos de carga, ya que están preparado para las cargas incrementales, el automatismo, son fácilmente monitorizables y muy eficientes en cuanto a tiempos de ejecución. Esto, junto a un completo modelo de datos, hace que las futuras mejoras se centren en su mayor parte en las herramientas de explotación de datos y no en la carga ni modelado de información. No obstante, si se hubiera dedicado menos tiempo en la ETL, se podrían haber obtenido otros resultados más tangibles para el usuario e incluso de mayor utilidad, que los que aporta una ETL optimizada.

- Se dispone de un conjunto de informes predefinidos que son un buen punto de partida para mostrar las capacidades que ofrece el sistema al usuario. Esto le permitirá comprender la utilidad y el valor que el almacén de datos proporciona a su negocio, para en una siguiente fase poder evolucionarlo según sus necesidades y teniendo una idea más clara de lo que le puede aportar.

- Finalmente, considero que el TFC propuesto en la asignatura ha sido muy interesante y de gran utilidad, ya que me ha permitido adquirir conocimientos y experiencia en proyectos relacionados con la inteligencia de negocio, los cuales pueden generar mucho valor a las empresas mediante la exploración y el análisis de sus propios datos.

8. Líneas de evolución futuras

Después de realizar las conclusiones del proyecto y analizar el estado actual en el que éste queda, consideramos que el sistema debería de evolucionar para incluir las siguientes funcionalidades:

8.1. Creación de metadatos Pentaho para *reporting ad-hoc*

La plataforma *Pentaho BI Suite* ofrece la herramienta *Pentaho Metadata Editor* que permite definir modelos de negocio a partir de almacenes de datos y distintas fuentes, para que posteriormente usuarios menos especializados puedan utilizarlo en la elaboración de sus propios informes y cuadros de mando.

El objetivo de esta herramienta es mapear la estructura física de la base de datos a un modelo lógico de negocio conocido por el usuario, de manera que el usuario se desvincula de la implementación física que hay detrás del modelo, e incluso es capaz de explotar el almacén de datos sin conocimiento de lenguajes de manipulación de datos como pueden ser las sentencias SQL.

8.2. Definir parámetros de seguridad y niveles de acceso a los datos

En estos momentos, todos los usuarios con cuenta que les permita acceder a la plataforma, tienen acceso a todos los informes que en ella están publicados. Se deberían de crear distintos perfiles de acceso con diferentes competencias en cuanto a la información que pudieran visualizar en la plataforma. Es conveniente hacer la siguiente distinción en cuanto a visibilidad se refiere:

- **Visibilidad de objetos y funcionalidades:** Este tipo de seguridad está relacionado con los diferentes informes y componentes de la estructuras de carpetas del servidor a las que un determinado usuario tiene acceso. Podrán existir determinados informes que sólo estén disponibles a un cierto perfil de usuario, o podrán existir un conjunto de funcionalidades a las que los usuarios no podrán acceder, como crear sus propios informes, etc.
- **Visibilidad del dato:** Este tipo de restricción se refiere a que aunque dos usuarios tengan acceso al mismo informe, la ejecución de éste no devuelva la misma información a ambos usuarios. Esto podría aplicarse para que responsables de un determinado establecimiento o demarcación, sólo pudieran visualizar la información de sus establecimientos. Otra utilidad sería que solamente determinados usuarios pudieran obtener información referente al volumen de ventas (dinero), siendo no obstante el número de ventas, visible para todos los usuarios de la organización.

8.3. Creación de cuadros de mandos e informes dinámicos

La entrega inicial contempla para la parte de análisis con un conjunto de informes predefinidos, que si bien abarcan un amplio abanico de tipología de informes y uso de componentes, e incluyen multitud de filtros para sus consultas, no dejan de ser informes estáticos con capacidad de análisis online limitada.

Sería recomendable la construcción de cuadros de mando empresariales dirigidos al área de dirección, que permitan de un simple vistazo ver el estado de sus KPI o indicadores claves del negocio así como funcionalidades de profundización y sintetización *(Drill down/up)* que permitan explorarlos o analizar parte del documento dependiendo de la selección realizada.

8.4. Tareas relacionadas con el *Data Quality* (Calidad del dato)

La propia naturaleza del sistema desarrollado, el cual permite analizar la información proveniente de la empresa, hace ver lo sensible que son los sistemas de información de estas empresas a las deficiencias en cuanto a calidad del dato se refiere.

En nuestro proyecto, tal y como queda reflejado en el apartado **3.3 Análisis de la Calidad del dato** la información recogida muestra multitud de “anomalías” que si bien en cierta medida han sido detectadas y controladas, existen multitud de escenarios que requieren de un tratamiento más específico y complicado para su consolidación y aceptación.

Un ejemplo de estos escenarios se manifiesta, en la existencia de multitud de proveedores contenidos en los ficheros de los productos, que aún teniendo descripciones diferentes, es muy probable que se traten del mismo proveedor y a día de hoy el sistema los está considerando como diferentes.

9. Bibliografía

- Plan docente TFC (Almacenes de datos)
- Conceptosw DW:
  http://www.monografias.com/trabajos57/data-warehouse-sql/data-warehouse-sql2.shtml
  http://es.wikipedia.org/wiki/Granularidad
  http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager#x1-450003.4.4.2
- Herramientas y desarrollo:
  OpenProj: http://www.slideshare.net/abetancur/openproj
  Pentaho: http://community.pentaho.com/
  Pentaho: http://wiki.pentaho.com/display/COM/Community+Wiki+Home
  Programación Batch:
  http://www.palomatica.info/juckar/microsoft/msdos/bat/intro.html
  http://www.taringa.net/posts/info/2071753/Curso-de-Batch-acelerado.html
  http://foro.elhacker.net/scripting/recorrer_directorios_y_generar_log_batch-t332054.0.html
  Comando At: http://www.colorconsole.de/console/es/002.htm
- Herramientas BI:
  http://www.pgconocimiento.com/Productos/AtlasSBI.html
  http://www.talend.com/index.php
  http://www.redopenbi.com/group/pentahodataintegration/forum/topics/diferencias-entre-las
  http://todobi.blogspot.com.es/2009/10/nuevo-visor-olap-para-la-version.html
- Documentación Departamentos de BI de Telefónica Soluciones
- Proyectos ejemplo de Almacenes de Datos UOC
- Resto documentación:
  Data Quality: http://es.wikipedia.org/wiki/Calidad_de_datos