laSalle

UNIVERSITAT RAMON LLULL

Escola Tècnica Superior d'Enginyeria La Salle

Treball Final de Màster

Màster Universitari en Enginyeria Informàtica i la seva gestió

Desarrollo de un sistema de información para el análisis energético: modelos de información, técnicas y ámbitos de aplicación

Alumne
Fàtima Galan Armell

Professor Ponent
Leandro Madrazo

ACTA DE L'EXAMEN DEL TREBALL FI DE CARRERA

Reunit el Tribunal qualificador en el dia de la data, l'alumne

D. Fàtima Galan Armell

va exposar el seu Treball de Fi de Carrera, el qual va tractar sobre el tema següent:

Desarrollo de un sistema de información para el análisis energético: modelos de información, técnicas y ámbitos de aplicación

Acabada l'exposició i contestades per part de l'alumne les objeccions formulades pels Srs. membres del tribunal, aquest valorà l'esmentat Treball amb la qualificació de

Barcelona,

VOCAL DEL TRIBUNAL VOCAL DEL TRIBUNAL

PRESIDENT DEL TRIBUNAL

Abstract

Este trabajo recoge los resultados obtenidos en la fase de estudios previos del proyecto REPENER financiado por el Plan Nacional de I+D+i 2009-2012. El objetivo de este proyecto es diseñar e implementar un prototipo de repositorio digital online capaz de almacenar, acceder y analizar datos sobre el comportamiento energético de los edificios con el fin de optimizar su eficiencia energética.

REPENER requiere un sistema de información capaz de recoger y organizar los datos energéticos para aplicar sobre ellos herramientas de extracción de conocimiento y así facilitar a diferentes tipos de usuarios la toma de decisiones encaminadas a mejorar la eficiencia energética de los edificios. Para ello, es determinante disponer de un sistema de información capaz de organizar y simplificar los datos energéticos para luego poder aplicar de manera efectiva las técnicas y herramientas de minería de datos.

El trabajo aborda los fundamentos teóricos que en los que se basan los métodos de diseño y almacenamiento de la información, su publicación en la red y las herramientas de minería de datos más destacadas. El trabajo concluye con una demostración de la aplicabilidad de estas herramientas y del valor añadido que aportan a la información contenida en el sistema.

Resumen

El trabajo desarrollado para el Proyecto Final de Máster “Desarrollo de un sistema de información para el análisis energético: modelos de información, técnicas y ámbitos de aplicación” forma parte de la fase de estudios previos del proyecto de investigación REPENER, financiado por el Plan Nacional de I+D+i 2009-2012. El objetivo de este proyecto es diseñar e implementar un prototipo de repositorio de información energética que permita desarrollar una metodología para adquirir y analizar datos energéticos y extraer conclusiones que permitan a distintos tipos de usuarios tomar acciones encaminadas a mejorar la eficiencia energética de los edificios, tanto en los edificios existentes como en los nuevos proyectos.

En este trabajo se abordan los fundamentos teóricos en los que se basa el diseño de un sistema de información energética, y la aplicación de las herramientas de minería de datos para extraer conocimiento a partir de él.

El sistema de información que se plantea desarrollar en el proyecto comprende tres tipos de procesos:

1. Recogida de datos o diseño de la información.
2. Pre procesado de los datos
3. Análisis de la información.
4. Extracción de las reglas que permitan visualizar conclusiones aplicables a la eficiencia energética.

El objetivo final del estudio es la extracción de conocimiento a partir del sistema de información con el fin de mejorar la eficiencia energética, tanto de los edificios existentes como de nuevos proyectos.

En la sección práctica de este estudio se aplican herramientas de minería de datos a un caso de estudio consistente en una base de datos de consumos energéticos facilitados por una empresa suministradora.

Índice

Índice

Índice de Figuras

1. Introducción

El objetivo del proyecto REPENER REPENER, financiado por el Plan Nacional de I+D+i 2009-2012, es diseñar e implementar un prototipo de repositorio digital online capaz de almacenar, acceder y analizar datos sobre el comportamiento energético de los edificios con el fin de optimizar la eficiencia energética.

El repositorio de información energética permitirá desarrollar una metodología para adquirir y analizar diferentes tipos de datos, estáticos (valores técnicos, geometría) y dinámicos (consumos, patrones de ocupación, clima). A partir de la combinación de los datos reales y los simulados se podrán realizar los análisis comparativos y derivar conclusiones a partir de los resultados.

La información recogida en el repositorio permitirá ajustar el comportamiento simulado con el real mejorando así la eficiencia energética a medio plazo. A corto plazo, los análisis sobre la información energética pueden ser útiles para la rehabilitación y el mantenimiento de edificios existentes. A largo plazo, la información recogida en el repositorio puede ser utilizada para evidenciar las mejoras del parque de viviendas a lo largo de los años.

Mediante la elaboración de este trabajo se evalúan los procesos más importantes para la creación de un sistema de información como el que el citado proyecto requiere.

El trabajo se estructura en cuatro secciones.

En primer lugar, en la sección Estado del Arte se abordan los conceptos básicos sobre la transformación de los datos en información, y de ésta en conocimiento, tal como se lleva a cabo en el Business Intelligence. En esta sección también se trata la cuestión de la infraestructura necesaria para alojar la información en la red, como el Cloud Computing. A lo largo de esta sección se definen los aspectos más importantes, se buscan las plataformas existentes relacionadas y finalmente se evalúan las diferentes tecnologías y sus alternativas para el futuro desarrollo del proyecto REPENER.

La segunda sección está dedicada al Descubrimiento de Conocimiento (KDD, Knowledge Discovery in Databases) Con el fin de distinguir y caracterizar los conceptos más relevantes, el estudio de esta sección se ha dividido en cuatro bloques:

1. Recogida de datos o diseño de la información.
2. Pre procesado de los datos
3. Análisis de la información.
4. Extracción de reglas y conclusiones

En esta sección se analizan las metas y las tendencias actuales en el desarrollo de procesos de minería de datos y se evalúan la idoneidad de los procesos disponibles para su aplicación al análisis de un sistema de información energética.

Finalmente, la tercera sección está dedicada a recoger los resultados de la aplicación práctica de las herramientas de minería de datos a un caso de estudio. Se ha llevado a cabo una comparación de algunas de las herramientas, utilizando los datos facilitados por Leako, empresa vasca que ha elaborado un sistema de monitorización y recogida de los valores de los consumos de agua y calefacción de un conjunto de viviendas del País Vasco.

A los datos proporcionados por Leako se les han aplicado los algoritmos más representativos de la minería de datos, entre ellos : algoritmos de clasificación de predicción y de clusterización, con el fin de valorar los distintos resultados, Los resultados obtenidos con la aplicación de las herramientas de minería de datos proporcionan un valor adicional a la información almacenada en la base de datos que hay que saber interpretar para poder sacar las conclusiones pertinentes que contribuyan a la mejora de la eficiencia energética de los edificios estudiados.

2. Estado del Arte

En el presente trabajo se han evaluado todas aquellas tecnologías que deben ser tenidas en cuenta para el desarrollo de un repositorio de información energética como el que se plantea en el proyecto REPENER. En esta sección se identifican y analizan algunas de las tecnologías necesarias, las cuales se organizan en cuatro apartados:

Procesos: introducción al concepto de Business Intelligence. El Business Intelligence, desde el punto de vista más empresarial, representa el conjunto de procesos que se aplicará a los datos con el fin de obtener conocimiento útil para la empresa.

Datos: introducción a los conceptos de ontologías y OpenData. Ambos conceptos tratan sobre la estructuración de la información en la red y son en la actualidad las propuestas tecnológicas más innovadoras que se están llevando a cabo.

Infraestructura: El debate "On vs Off Premises". Off Premises representa el término más conocido como Cloud Computing. Se estudian los conceptos básicos del Cloud Computing, sus principales arquitecturas y los beneficios que aporta.

Proyectos sobre eficiencia energética: Estudio de casos prácticos existentes sobre el uso de herramientas de Minería de Datos para mejorar la eficiencia energética en viviendas y edificios de oficinas.

2.1 Procesos

El objetivo práctico de este estudio es disponer de un sistema de información que permita extraer conclusiones sobre la información sobre consumos energéticos en viviendas acumulada en una base de datos. Para ello, se comienza estudiando los procesos de Business Intelligence, en el ámbito de las empresas, su auge en los últimos tiempos y, finalmente, su potencial para apoyar

2.1.1 El Business Intelligence en las empresas

El Business Intelligence o BI representa la habilidad para transformar los datos en información, y la información en conocimiento. De este modo, se define el Business Intelligence como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar los datos en información estructurada para su análisis y conversión en conocimiento. (Sinnexus, 2010)

El objetivo del BI es permitir la toma de decisiones a través de los datos almacenados. Serán necesarios un conjunto de procesos que lleven de la propia ordenación de los datos en sistemas integrados, pasando por el tratamiento de dichos datos para la extracción del conocimiento asociado a estos hasta la toma de decisiones a partir de las conclusiones extraídas.

Desde principios de 1990, las aplicaciones de BI han evolucionado en diversas direcciones, en respuesta al crecimiento exponencial de la información. Desde informes operacionales generados por mainframes, modelación estadística de campañas publicitarias, ambientes OLAP multidimensionales para analistas así como dashboards y scorecards para ejecutivos. Las compañías empiezan a demandar más métodos para analizar y realizar informes a partir de datos. (Anandarajan Murugan, 2004)

Tal y como indican los expertos como Sondergaard (Taylor, 2011): "En los próximos 20 años el principal cambio se dará a partir de la información. La información será el petróleo del siglo 21". Los últimos informes desarrollados por las compañías que evalúan las tendencias tecnológicas sostienen que la recopilación y uso de Business Intelligence son fundamentales para el impulso del crecimiento. (Pettey, 2010)

Las aplicaciones de los procesos de BI se pueden encontrar en diferentes plataformas especializadas en la extracción de conocimiento o en herramientas integradas en plataformas que engloban procesos de gestión de información como, por ejemplo, un Data Warehouse

En la actualidad, las aplicaciones BI ya son imprescindibles en la toma de decisiones para una gran variedad de organizaciones de todo el mundo. El BI hace que las empresas sean más competitivas dado que les permite visualizar interacciones entre diferentes ámbitos de su negocio en un formato fácilmente interpretable, mediante dashboards diseñados de un modo eficaz y preciso. De este modo los gerentes de las empresas pueden tener acceso a la información crítica de su negocio y compartir dicha información con socios, clientes e incluso empleados. Esta información permite mejorar el rendimiento operativo, encontrar soluciones para hacer frente a problemas y , en el caso de los directivos, verificar las consecuencias de sus decisiones.

En definitiva, el BI permite que las empresas sean más inteligentes, mejores y más rápidas en diferentes ámbitos de decisión como son la reducción de pérdidas, el uso eficiente de recursos humanos, y las mejoras en ventas y marketing.

2.1.2 Análisis de tendencias tecnológicas

Gartner es una organización dedicada a proporcionar informes sobre el análisis de las últimas tendencias que facilita consejos sobre Tecnologías de la Información para profesionales y empresas de tecnología. En los últimos informes y notas de prensa publicadas por esta organización, esta define el Business Intelligence como una de las cuatro grandes tendencias que van a cambiar la tecnología y la economía en los próximos 10 años. (Pettey, 2010)

Se prevé que para el año 2012, el 20 por ciento de los procesos orientados al cliente tengan como objetivo facilitar el conocimiento, adaptable y montado justo a tiempo para satisfacer las demandas y preferencias de cada cliente, con la asistencia de las tecnologías BPM(Business Process Management).

Hoy en día la capacidad de cambiar de forma proactiva los procesos no es más que un paso intermedio para la mejora de estos. La evolución creará procesos que se adapten automáticamente a las referencias del usuario, a la demanda de los consumidores, a la

capacidad de predicción, de tendencias, análisis de la competencia y las relaciones sociales. (Pettey, Goasduff, 2010) basándose en la detección de patrones.

2.1.3 Business Intelligence aplicado a proyectos energéticos

Los procesos de Business Intelligence no sólo son aplicables al análisis de organizaciones y empresas, sino también a otros campos como el medio ambiente, la energía o la sostenibilidad. La aplicación de los procesos BI a estos ámbitos puede contribuir a la mejora de la eficiencia energética, y a la optimización de la utilización de los recursos energéticos disponibles.

Además del ámbito empresarial, esto son algunos de los ámbitos sobre los que se aplican procesos de Business Intelligence son:

- Energía
- Tráfico o transporte
- Medio ambiente
- Agricultura
- Ámbitos horizontales

Energía

Por lo que se refiere al sector energético, el que guarda más relación con el proyecto, las aplicaciones del BI pueden tener cabida en: el sector eléctrico y el petrolífero.

En el caso del sector eléctrico, desde hace algunos años se han planteado las SmartGrids (redes eléctricas que utilizan la tecnología digital como alternativas para adaptar el consumo y la demanda a las necesidades del consumidor - como alternativa a as redes eléctricas existentes, consideradas anticuadas, frágiles e ineficientes. En la era digital mediante el uso de una red como la de Internet se obtiene información del estado del flujo energético a través de la red en tiempo real. Para adaptar en tiempo real demanda y consumo, son necesario procesos de BI.

Algunos de los beneficios para los distintos agentes implicados en este sector son:

- Para las empresas distribuidoras:
  - ✓ Reducción de pérdidas de energía: la compañía puede gestionar su energía de manera autónoma, identificando y controlando el gasto de la misma.
  - ✓ Eficiencia: realizando sofisticados análisis de los patrones de consumo, identificación de oportunidades que posibiliten la reducción del consumo.
  - ✓ Optimización de la infraestructura de red.
    - Incluyendo ante caso de accidentes, desastres medioambientales o atentados.
  - ✓ Permiten ofrecer un mejor servicio al cliente, con más ventajas comerciales (nuevas tarifas, pago por uso, etc.).

- Para los usuarios:
  - ✓ Pago por uso: al no ser necesaria una lectura manual, se eliminan los recibos estimados y los consumidores sólo pagan por lo que consumen.
  - ✓ Tarifas flexibles: las empresas gestionan diversas tarifas para optimizar el consumo de la energía.
  - ✓ Gestión en remoto del suministro de energía: no será necesario una intervención local para activar, terminar o incrementar el suministro.

En el caso del sector petrolífero es también necesario que se tomen medidas para la reducción de su consumo, dado que cada vez las reservas mundiales de crudo son menores y esto obliga a buscar energías alternativas. Este proceso de cambio, por ejemplo, obliga a las empresas de venta de gasolina a redefinir y reformular los diferentes operativos existentes como las estaciones de servicio.

Transporte

El principal leitmotiv en este área es la de obtener información para construir políticas que hagan un mundo más sostenible.

El tráfico genera contaminación pero para la sociedad representa una calidad de vida a la que no está dispuesta a renunciar.

Se desarrollan iniciativas que permitan reducir estos excesos mediante iniciativas como el coche eléctrico, el control del coche rodado, el control aéreo, la previsión meteorológica.

Medio ambiente

Además de la energía eléctrica existen otros ámbitos que aportan datos que permiten valorar y mejorar la eficiencia energética tales como el control y las mediciones meteorológicas.

En el caso del control de las aguas, conocer datos del flujo se suministro, por ejemplo permite identificar las pérdidas en la red de distribución. Este hecho representa una ganancia doble, por un lado se ahorra agua, y por el otro se ahorra parte de la energía necesaria para la distribución del agua.

2.2 Datos

El principal objetivo de la plataforma a desarrollar será la de extracción de conocimiento por parte de la información con el fin de sacar conclusiones hacia la eficiencia energética. Pero para tal objetivo son necesarios datos con los que trabajar.

En esta sección se introducen los conceptos de ontologías y OpenData. Entre ellos, a priori, no representan el mismo concepto pero comparten el mismo propósito, la estructuración de la información entendida como un conjunto de datos que una vez ordenados adquieren significado.

2.2.1 Ontologías

Una ontología es un modelo de representación del conocimiento o como indica Gruber (Gruber, 1993) "Una ontología se entiende como una artefacto de diseño formulada para propósitos específicos y evaluada a través de un criterio de diseño específico". Más sucintamente, según Gruber "Una ontología es una especificación explícita y formal de una conceptualización".

Las ontologías permiten a los usuarios entender el dominio mediante una vista conceptual de los elementos que la forman y permiten relacionar el poder expresivo y la complejidad del razonamiento.

Una ontología resultará la representación de una taxonomía relacional de conceptos y conjunto de axiomas, a través de los cuales se podrá inferir nuevo conocimiento.

Para el modelado de ontologías se encuentran históricamente tres clasificaciones:

- los sistemas basados en redes semánticas: representación gráfica a través de la declaración de conceptos y sus relaciones. Los conceptos representan nodos y las relaciones las flechas que los unifican.
- los sistemas basados en marcos o frames: cada marco (frame) representa un concepto, y se le añaden categorías (slots) que pueden tener especificaciones (fillers). De este modo un marco que representa el concepto "vino" puede tener un slot "color", que puede obtener los valores "blanco", "rosado" o "tinto". (Lassila, 2001)
- los sistemas basados en lógicas descriptivas: son una familia de formalismos basados en la lógica para la representación del conocimiento. Provee teorías y sistemas para expresar información estructurada, permitir su acceso, y poder razonar de forma semánticamente precisa.

Las definiciones y conceptos introducidos hablan de modelo de información pero no tienen en cuenta en la aclaración realizada por Borst (Borst, 1997) en la cual clarifica que "Una ontología es una especificación formal de una conceptualización compartida". Lo que en este caso Borst indica dos conceptos clave. En primer lugar Borst sostiene que la ontología representa conocimiento compartido y en segundo lugar sugiere que una ontología debe ser especificada usando un lenguaje formal para que pueda ser procesado por personas y por ordenadores.

Esta ampliación del concepto de ontología que abarca el conocimiento compartido conduce hacia la Web Semántica y a la estructuración de la información en la Web.

La Web Semántica

La Web Semántica representa el concepto según el cual la Web no es el mecanismo de localización perfecto dado que al gran volumen de información que este maneja no puede solucionar la problemática de la recuperación de la información de la Web. Como indican Peis (Peis, Herrera-Viedma, & Hassan, 2003) la Web Semántica aspira en última instancia a que las máquinas sean capaces de "comprender" el significado de los hiperdocumentos, con el fin de analizar su viabilidad y las ventajas que implicaría en la Recuperación de la Información en la Web.

Para tal objetivo, a continuación, se describirá la infraestructura de tecnologías y lenguajes necesaria, la cual se esquematiza en 4 niveles:

1. Modelo básico para establecer las propiedades sobre los recursos o asertos. Para el que se emplea RDF.
2. Modelo para definir las relaciones entre los recursos, a través de clases y objetos. Para el que se emplea RDF Schema.
3. Capa lógica que permite realizar las consultas e inferir el conocimiento. Para el que se emplean las ontologías. Este es el apartado sobre el que se centra nuestro interés.
4. Capa de seguridad que asignen fiabilidad a determinados recursos. Para el que se emplea la Firma Digital.

Por lo que respecta a las ontologías y la Red Semántica, existen dos lenguajes estándar que se utilizan para estructurar y modelar la información semánticamente: RDF y OWL

RDF y RDF Schema

RDF es un lenguaje que fue desarrollado para describir los recursos en la Web. Es un lenguaje de etiquetado, creado mediante sintaxis XML, que define un modelo de datos para describir recursos usando identificadores Web (Uniform Resource Identifiers, URI). Para tal efecto se introduce el concepto de tripletas “sujeto-predicado-objeto”.

RDF Schema es un vocabulario RDF que permite describir recursos mediante un mecanismo para definir clases, objetos y propiedades, relaciones entre clases y propiedades y restricciones.

Con RDFS se pueden describir jerarquías de clases sobre las que realizar las consultas.

OWL

OWL o Web Ontology Language es un lenguaje de representación de ontologías, se cimenta en RDFS. Mediante este se podrá definir formalmente el significado de la terminología usada en los documentos Web.

2.2.2 OpenData

El OpenData es una iniciativa llevada a cabo por el gobierno y la administración pública mediante la apertura a información pública fomentando la transparencia del gobierno electrónico.

El OpenData proporciona la accesibilidad de datos a través de la red de una manera pública.

De este modo los gobiernos de diferentes países han llevado a cabo proyectos para fomentar este concepto. Esta interoperabilidad no deja de formar parte de las iniciativas llevadas a cabo por los gobiernos con el fin de acercar la administración a los ciudadanos.

En el proceso de búsqueda se encontraron dos fuentes que pueden ser de gran importancia:

1) La Agencia Estatal de Meteorologíao o AEMET

La Agencia es la encargada del desarrollo, implantación, y prestación de los servicios meteorológicos de competencia del Estado y el apoyo al ejercicio de otras políticas públicas y actividades privadas, contribuyendo a la seguridad de personas y bienes, y al bienestar y desarrollo sostenible de la sociedad española. (AEMet, 2010)

A partir del 30 de Noviembre del 2010 AEMET facilita el acceso libre a sus datos meteorológicos.

A través de un entorno web (Figura 1) permiten el acceso, visualización y descarga (a través de un servidor ftp) de diversos parámetros relacionados con la información meteorológica. Mediante esta conexión la plataforma permite el acceso a la consulta de los valores proporcionados por la red de centrales que tiene distribuidas por todo el territorio español.

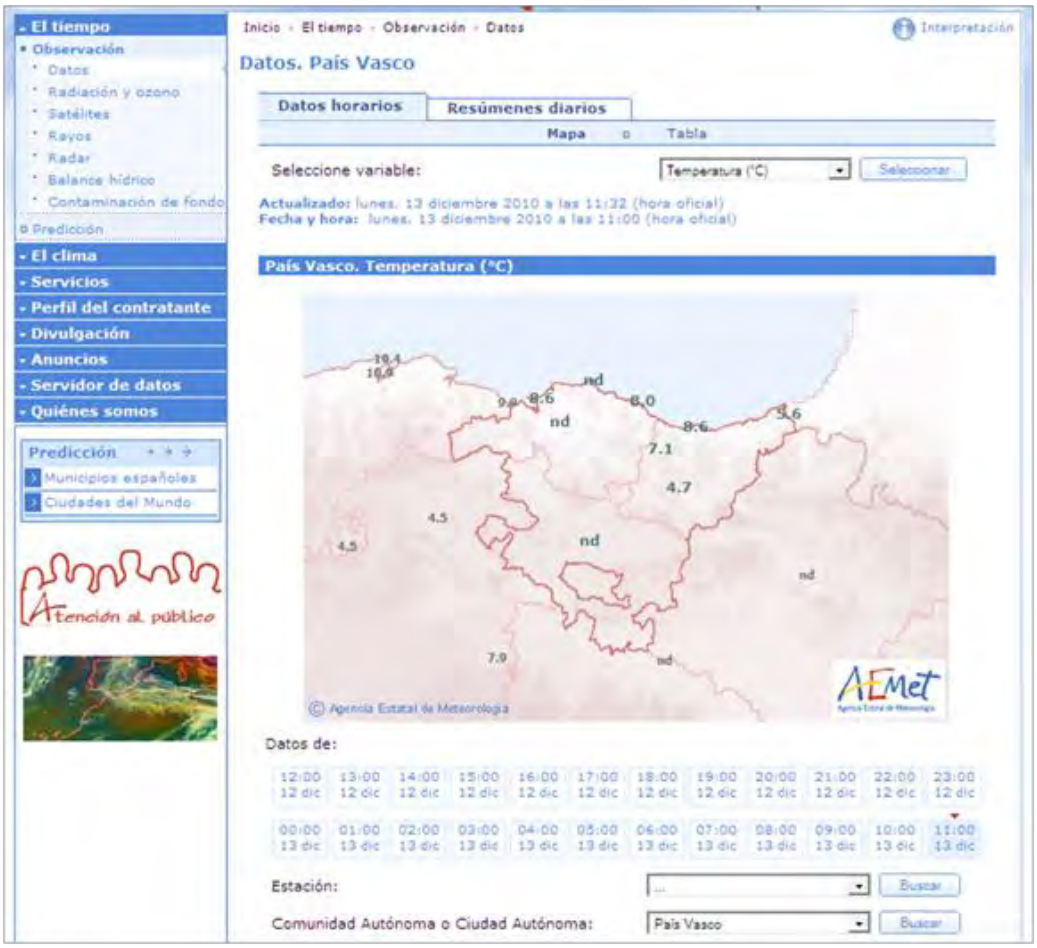


Figura 1: Plataforma de visualización de información de AEMet

El sistema entre otras actividades recoge datos diez minutales de las 250 estaciones de observación de las redes de AEMET.

Parámetros que se pueden encontrar:
Velocidad media del viento
Dirección media del viento

Velocidad máxima del viento
Dirección de la velocidad máxima del viento
Temperatura del aire
Humedad relativa
Temperatura del punto de rocío
Presión
Precipitación
Presión reducida al nivel del mar
Capa nieve
Temperatura máxima en 10 min
Temperatura mínima en 10 minutos
Temperatura máxima en 1 hora
Temperatura mínima en 1 hora
Reducción de la presión a altura del geopotencial ...
Reducción de la presión a altura del geopotencial ...
Reducción de la presión a altura del geopotencial ...
Hora y minuto de la temperatura máxima en 1 hora
Hora y minuto de la temperatura mínima en 1 hora
Visibilidad
Tiempo presente
Insolación
Radiación global
Temperatura suelo
Temperatura subsuelo 5cm
Precipitación acumulada líquida
Precipitación acumulada sólida
Recorrido del viento

Figura 2: Listado de parámetros proporcionados por AEMet

En caso que realizar la descarga de la información a través del servidor FTP, la información se proporciona mediante archivos de tipo *.xls (Microsoft Office Excel). La información se puede seleccionar por fecha o por identificador de la estación.

2) Open Data Euskadi

Tal y como indican en la propia web, Open Data Euskadi es una iniciativa enmarcada dentro de la política del Gobierno de Euskadi. Mediante este proyecto el Gobierno Vasco expone sus datos públicos con el fin de que terceros puedan crear servicios derivados de los mismos. (Open Data Euskadi, 2010)

En este caso el catálogo de datos proporcionado es muy amplio y engloba áreas tan diversas como:

- Actividades económicas
- Administración Pública
- Asuntos Sociales
- Euskera

- Educación
- Medio Ambiente
- Meteorología
- Ocio y Turismo
- Salud
- Seguridad e Interior
- Transporte y movilidad
- Trabajo y Empleo
- Urbanismo y territorio
- Vivienda

En este caso la plataforma web (Figura 3) es solo para la búsqueda de la información y descarga de los archivos que sean de interés.

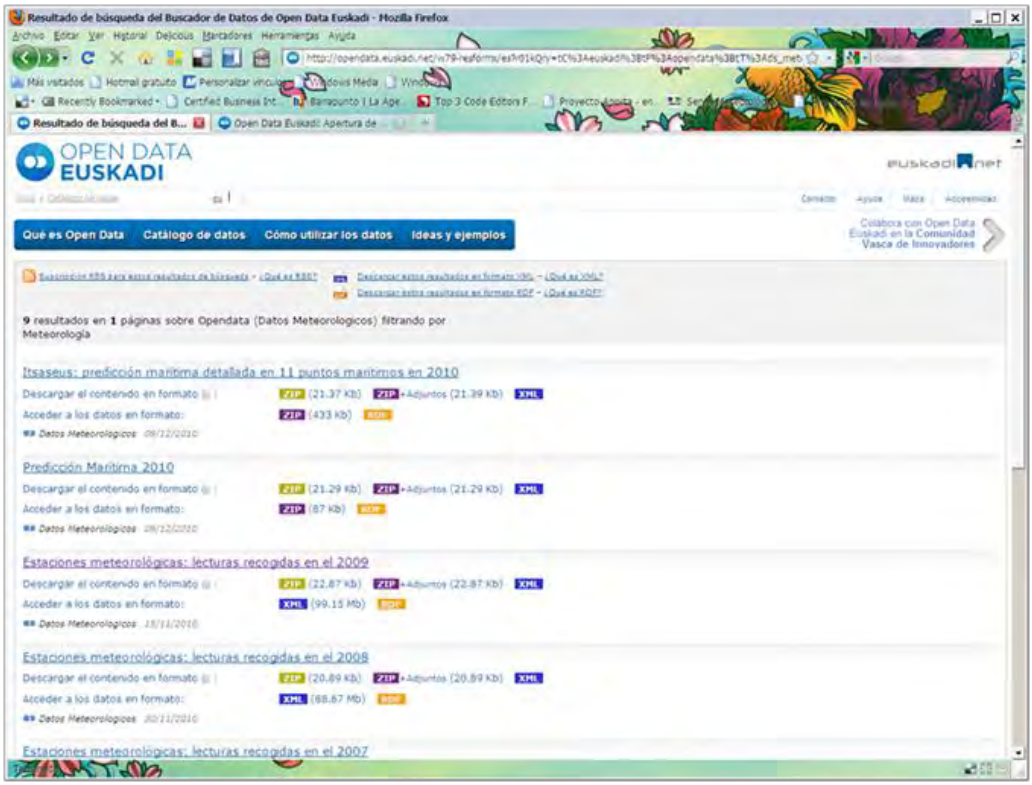


Figura 3: Interfaz Web Open Data Euskadi

En este caso el sistema ofrece distintos formatos de descarga de la información. Entre estos figuran: XML , CSV, WMS u otros.

En el caso que influye al proyecto se han podido encontrar archivos sobre las lecturas recogidas por las estaciones meteorológicas desde el 2003 y las predicciones meteorológicas del 2010. La información se encuentra estructurada en documentos del tipo XML en los que podremos ver la totalidad de la información recogida tal y como se ve en la Figura 4.

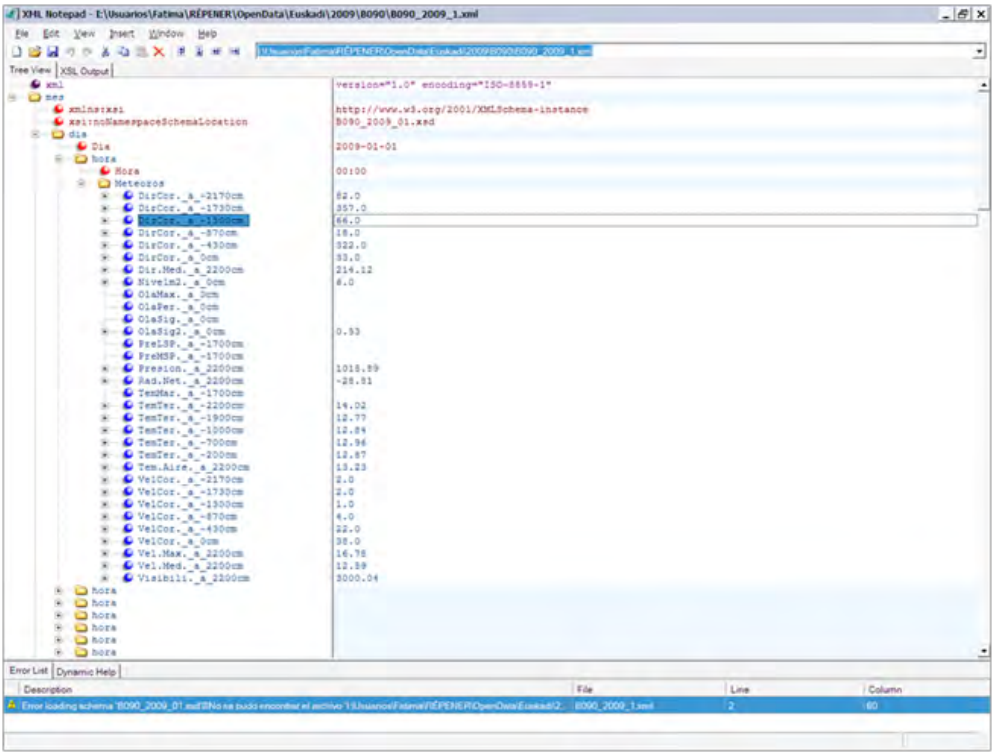


Figura 4: Documento XML proporcionado por Open Data Euskadi

2.3 Infraestructura

Mediante este proyecto se quiere desarrollar un sistema de almacenamiento de información que se localizará en Internet para que los usuarios la puedan consultar y puedan operar con ella en línea.

Para que este objetivo sea posible es necesario considerar los mecanismos y tecnologías con las que poder vincular la información con Internet. En este caso se plantean preguntas como:

- ¿Dónde se debe ubicar el sistema? ¿En un servidor privado permitiendo el acceso desde la red? o ¿Directamente en la red?
- ¿Dónde se ubicarán los datos y los procesos? ¿Los datos en la red y los procesos en local? o ¿Todo en el mismo lugar?

Son muchos los escenarios posibles que responden a estas preguntas. Por ello, a continuación, se evalúa cual puede ser la mejor infraestructura para este caso: On vs Off Premises.

2.3.1 On vs Off Premises

1) On Premises

El término On Premises hace referencia al modelo clásico donde los datos y el software instalado funcionan en servidores privados (dentro de una organización).

Desde que en 2005 nació la posibilidad de la ejecución del software de forma remota ha sido el modelo más usado. Aunque es un modelo que se considera anticuada en las condiciones actuales, este seguirá siendo el preferido en ciertos sectores económicos.

2) Off Premises

Off premises software hace referencia al Cloud Computing. El Cloud Computing representa el concepto mediante el cual los datos y/o los procesos se encuentran localizados en alguna granja de servidores en Internet. Esta solución está ganando fuerza, dado que representa un entorno para las tecnologías de la información escalable y que permite a las organizaciones desvincularse de la carga que supone la gestión de su infraestructura y del software. Además facilita el acceso a la información de una forma más sencilla incrementando la colaboración en una plataforma compartida.

Cloud Computing es un concepto que engloba diferentes tipos de soluciones que pueden proporcionar a sus clientes. Cloud Computing describe el uso de servicios, aplicaciones, información y infraestructuras proporcionadas en la red, con capacidad de computación y almacenamiento. Es por ello que se van a diferenciar los diferentes modelos de servicios que se proporcionan.

Tal y como se nos indica en la guía elaborada por la (Cloud Security Alliance, 2009) y se visualiza en la Figura 5, los tipos de virtualización se clasifican:

- en un primer nivel se encuentra la “Infrastructure as a Service” (IaaS) la cual representa el modelo en el que un operador web proporciona servidores virtuales a través de una dirección IP única y con un espacio de almacenamiento. Mediante esta arquitectura el cliente controla sus servidores y paga por el servicio que usa.
- en segundo nivel se encuentra la “Platform as a Service” (PaaS) la cual representa el modelo en el que se proporciona software ubicado en los servidores del proveedor, por ejemplo, un Windows o las aplicaciones de Google.
- finalmente el “Software as a Service” (SaaS) el cual representa el modelo en el que el proveedor da permisos al cliente para el uso de su software el cual se encuentra trabajando en la infraestructura de la cloud y será accesible para diferentes clientes web.

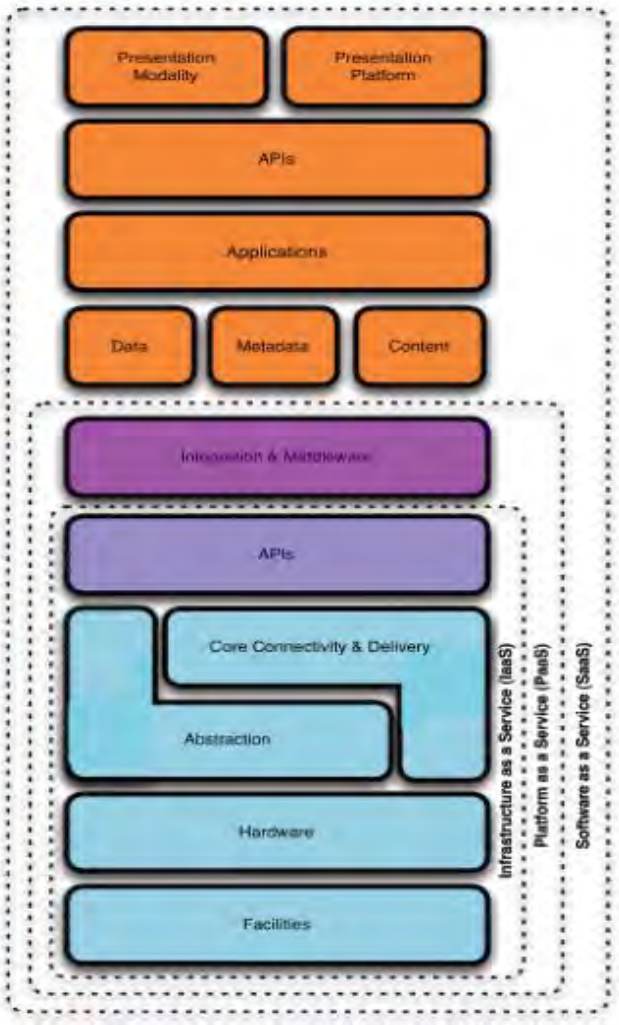


Figura 5: Esquema Cloud Reference Model (Cloud Security Alliance, 2009)

Modelos de explotación según los usuarios

En la cloud se ubican procesos y información pero la cuestión más importante que se plantea es cuál el control sobre quien accede y como a la información. Existen cuatro modelos de explotación en función de quien tiene los derechos de acceso a cada tipo de cloud:

- **Cloud pública:** es la solución más abierta donde una organización vende servicios en la cloud para un público general.
- **Cloud privada:** es la solución en la que la infraestructura cloud es propiedad de una organización y por ende solo ésta podrá operar en ella.
- **Cloud híbrida:** es la solución en la que la infraestructura cloud está formada por 2 o más infraestructuras cloud (públicas, privadas o de comunidad). Cada una mantiene su unidad como entidad pero se mantienen unidas por la estandarización de la tecnología que les permite el intercambio de datos y portabilidad de aplicaciones.
- **Cloud de comunidad:** es la solución en la que la infraestructura cloud es compartida por distintas organizaciones y compartible dado sus preocupaciones compartidas en requisitos de seguridad, política y consideraciones de cumplimiento.

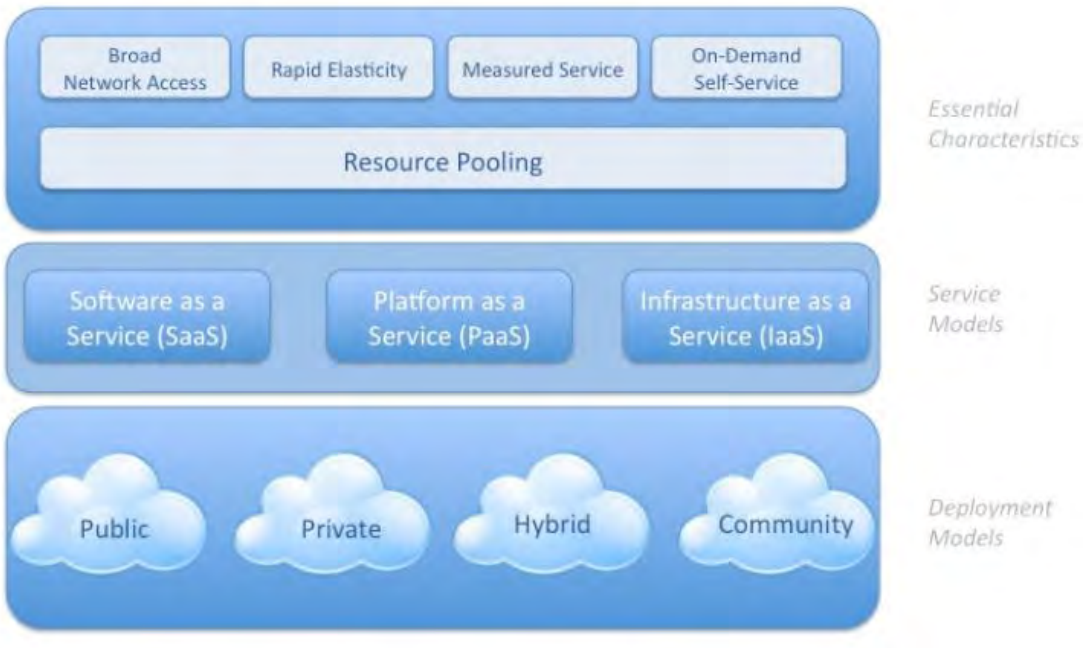


Figura 6: Esquema Cloud Global (Cloud Security Alliance, 2009)

Cada modelo de explotación cuenta con diferentes modelos de instancias: la cloud interna y la externa. La interna reside dentro del perímetro de seguridad de la red y la externa reside fuera de dicho perímetro.

Principales beneficios

Como ya se ha comentado esta plataforma cuenta cada vez con más adeptos. Está solución presenta un gran número de beneficios entre los que cabe destacar:

Reducción de costes: dado que las organizaciones ya no deben hacerse cargo de los costes hardware y software. Este hecho es muy significativo sobre todo desde el punto de vista de la escalabilidad o a la hora de tener que desarrollar una nueva infraestructura.

Rápido de crear: el servidor cloud proporciona la aplicación web para la creación de la infraestructura y en un momento la tienes creada.

Disponibilidad: el proveedor cloud tiene el último hardware, software y ancho de banda para las necesidades de su negocio. El proveedor del servicio proporciona la infraestructura, la plataforma y los servidores los cuales no se saturaran en caso de sobrecarga. La disponibilidad de los servicios estará garantizada por el proveedor.

Escalabilidad: como ya se ha comentado, esta solución ofrece mayor flexibilidad a la necesidad de incrementar infraestructura o servicios.

Eficiencia: gracias al ahorro representativo de esta nueva arquitectura, las organizaciones pueden invertir dicho dinero en otras áreas de la empresa.

Resistencia: dado que el proveedor garantiza proporcionar sus servicios incluso en caso de desastre natural.

Ejemplos de Soluciones Cloud comerciales

La arquitectura del Cloud Computing requiere una gran infraestructura de servidores ubicados en alguna granja, capaz de responder a la necesidad y potencia de procesamiento de información que los clientes requieren. Existen numerosas empresas que ofrecen un buen abanico de servicios entre las que destacan: Amazon, Google y Microsoft.

Amazon

Era una compañía que destacaba por proporcionar capacidad de computación a gran escala a sus empleados y consumidores a través de su web de ventas Amazon. Ofrecer capacidad de computación en bruto a través de Internet fue su paso natural. Sólo debía aprovechar su propia experiencia e infraestructura de centros de datos masivos para llegar a ser uno de los primeros proveedores de nubes más importantes.

Es uno de los innovadores en el sector del "Web-based computing" ofreciendo acceso a servidores virtuales y almacenamiento en la red. Se le puede considerar el padre del cloud computing desde que en 2006 lanzó su solución "Elastic Compute Cloud".

Los servicios que ofrecen son Amazon Web Services, el Elastic Compute Cloudy el Simple Storage Service.

Google

Es la empresa poseedora del buscador más usado en todo el mundo. Nadie puede concebir un Internet sin Google. Mientras Google proporciona su mejor buscador avanza desarrollando aplicaciones "Software as a Service". y su AppEngine resulta una buena alternativa en el mercado del platform as a service.

En 2007 Google hace el salto creando Google Apps, una suite que ofrece a sus clientes una plataforma de herramientas totalmente gratuita.

Los servicios que ofrecen son Google Apps, un conjunto de herramientas online de productividad como email, calendario, editor de texto, etc.

Microsoft

Es el desarrollador del sistema operativo más importante y ahora quiere poder proporcionarlo a través de la cloud. Windows ya ofrece un conjunto de servicios para el negocio a través de la web como Exchange, SharePoint, Office Communications Server, etc. A través de Azure Microsoft quiere proporcionar una herramienta que simplifique a sus clientes tener que mover sus aplicaciones a operadores externos.

Los servicios que ofrecen son Azure, plataforma "Windows as a Service", la cual es una plataforma de servicios que consiste en un sistema operativo y servicios para desarrolladores que se puede utilizar para construir y mejorar las aplicaciones alojadas en Web.

2.4 Proyectos sobre eficiencia energética

Las condiciones de vida y los patrones de consumo de la sociedad, son un ámbito que siempre despierta interés. Éste puede ser concebido por diferentes sectores entre los que se podrían citar: el de las compañías de suministros o el de los arquitectos.

2.4.1 Sistemas de gestión por monitorización de datos meteorológicos

Los avances sobre servicios de obtención de información real van principalmente enfocados a la reducción de costes, tanto en energía como en dinero. Esto aporta beneficios a todos los actores, es decir, los propietarios, los gestores del edificio y las propias empresas suministradoras. Un ejemplo ya citado anteriormente es el caso de las SmartGrid.

Cada vez los edificios están más automatizados y son más inteligentes y eso a su vez aumenta las expectativas y los niveles de exigencia de sistemas capaces por parte de los actores. Los sistemas de monitorización de edificios deben ser cada vez más eficientes para reducir y realizar un buen mantenimiento del edificio y a la vez mejorar el confort de los ocupantes de este.

Disponer de datos meteorológicos en tiempo real proporciona un control eficiente de la temperatura. Por ejemplo, un día soleado representa una entrada importante de energía solar a través de las ventanas, por lo que, en función de la orientación del edificio, hay sectores que requieren de menos potencia en la calefacción y viceversa. Con este razonamiento se pueden realizar muchas políticas de control en un edificio, desde la temperatura, la iluminación o el ahorro de agua mediante regular el uso de esta.

Un ejemplo de ello es el sistema Kepware (Kepware's WeatherBug for Automation, 2010). Kepware proporciona datos meteorológicos en tiempo real para el control de edificios y viviendas. El sistema se basa en la información recogida por las más de 8000 estaciones meteorológicas que poseen en América del Norte, la información es recogida por sus servidores y lista para ser tratada. El sistema provee de datos actuales y pronósticos relacionados con variables como la luz, la humedad, la velocidad del viento, niveles de lluvia, temperaturas y dirección del viento, entre las más importantes.

Los datos climatológicos son muy importantes también en el entorno de la edificación para poder elaborar los análisis energéticos de los edificios y el diseño de los equipamientos.

Otro ejemplo de repositorios útiles sobre datos energéticos son las bases de datos ODYSEE y MURE (Energy Efficiency Indicators in Europe, 2010). ODYSEE es un proyecto que involucra distintas agencias de la comunidad Europea. Esta contiene los valores de los consumos eléctricos finales y subsectoriales y a su vez indicadores de eficiencia energética y CO2. De este modo, lo que se quiere es comparar y evaluar las medidas y costumbres relativas a eficiencia energética de los diferentes países miembros y así detectar potenciales mejoras. MURE es la base de datos que contiene las medidas políticas relacionadas con la energía.

2.4.2 Sistemas gestión aplicando herramientas de Minería de Datos

Por lo que se refiere a los beneficios obtenidos mediante el uso de herramientas de Minería de Datos son numerosos los trabajos encontrados de ejemplo de uso de estas técnicas.

Los procesos aplicados en Minería de Datos incluyen redes neuronales, modelos en árbol, modelos lógicos y otros modelos estadísticos como series temporales, razonamiento basado en memoria y componentes principales.

Técnicas de regresión lineal y no lineal sirvieron para obtener modelos de regresión y ecuaciones energéticas para la predicción del uso de la electricidad anual de un rascacielos de oficinas con aire acondicionado de Hong Kong. (Lam, Hui, & Chan, 1997)

Mediante el uso de Redes Neuronales de Regresión General (GRNN) se optimizó el almacenamiento térmico del sistema HVAC (Heating, Ventilating, and Air Conditioning) en edificios públicos y de oficinas. Los resultados determinaron que una red neuronal bien diseñada podía ser una herramienta muy potente para optimizar el almacenamiento de energía térmica en edificios y funcionar sin problemas con registros de temperatura exterior. (Ben-Nakhi & Mahmoud, 2003)

Mediante análisis de regresión se realizaron procesos de benchmarking en los que se desarrollaron relaciones entre patrones de gran consumo energético y factores de uso que los justifiquen, por ejemplo por franjas horarias. Se tuvieron en cuenta variables tales como la edad del edificio, la ocupación y el tipo de sistema energético. (Chung, 2004)

Mediante algoritmos SOM se han identificado perfiles de uso de energía. Las herramientas de clusterización se utilizaron como metodología para identificar perfiles de consumidores eléctricos dentro de un área comercial o geográfica mediante el estudio de los patrones de sus consumos históricos. También se han encontrado aplicaciones de algoritmos SOM y algoritmos avanzados basados en SOM para desarrollar estrategias avanzadas en el control de HVAC. Estos son capaces de gestionar grandes volúmenes de información y extraer perfiles y patrones de datos de ocupación y temperatura que permitan adquirir, al sistema de control, de consciencia sobre el comportamiento y las tendencias de los usuarios y así ser capaz de anticipar sus necesidades y preferencias. (Cantos, Iglesias, & Vidal, 2009)

3. Descubrimiento de conocimiento en Bases de Datos (KDD)

Knowledge Discovery in Databases (KDD) hace referencia al proceso de varias fases que se debe realizar para la obtención de conocimiento de los datos almacenados en las bases de datos.

La sociedad se encuentra en la era de la información. El crecimiento constante de esta, desencadena procesos de análisis que permiten la extracción del conocimiento que en ella se contiene.

La información que se almacena es histórica y representan situaciones que se han producido en el pasado. Tal y como enseña Hernandez et al (Hernandez, Juan, Minaya, & Monserrat, 2004) la información histórica es útil para predecir la información futura. El mismo autor indica que la predicción del futuro no representa algo muy complejo dado que la mayoría de los actos cotidianos se basan en predicciones y experiencias pasadas y en su extrapolación futura.

Pero la información no es sinónimo de conocimiento. El conocimiento nace del diseño y desarrollo de herramientas que permiten el análisis del flujo de la información con el fin de potenciarla proporcionándole un valor añadido y diferencial.

El almacenamiento de datos es uno de los procesos y ha pasado a ser una de las rutinas básicas que deben desarrollar los sistemas de información de las organizaciones. ¿Pero con almacenar esta información hay suficiente? El siguiente paso debe ser el Descubrimiento de Conocimiento que estos almacenan (Knowledge Discovery in Databases, KDD).

Sin embargo, a su vez tal y como indica Fayyad et al (Fayyad, Piatetsky-Shiapiro, Smyth, & Uthurusamy, 1996) la gran cantidad de datos que almacenan las organizaciones hace imposible la utilización de métodos manuales para su análisis, por ello serán necesarias técnicas que ayuden al hombre.

El Descubrimiento de Conocimiento en Bases de Datos es un proceso de exploración que conlleva la aplicación de procedimientos algorítmicos para la manipulación de datos, construcción de modelos desde los datos y la manipulación de estos.

Algunos autores consideran que la minería de datos y KDD representan un mismo concepto. Sin embargo tal y como lo definieron Fayyad et al (Fayyad, Piatetsky-Shiapiro, Smyth, & Uthurusamy, 1996)y Chapman (Chapman, y otros, 2000)hay que distinguir tres procesos: el pre procesado de los datos, la aplicación de algoritmos de inducción y post-proceso de modelos.

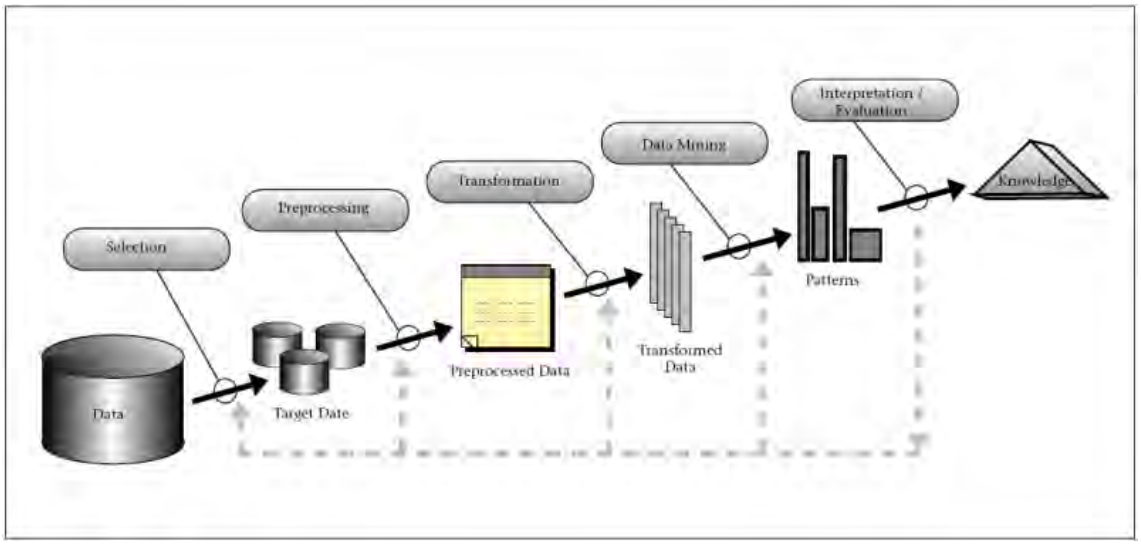


Figura 7: Proceso de Descubrimiento de Conocimiento según Fayyad (Fayyad, Piatetsky-Shiapiro, Smyth, & Uthurusamy, 1996)

Pre procesado de los datos:

La información en las bases de datos se encuentra desestructurada y para su utilización es necesario realizar un análisis exhaustivo.

En todas las fuentes evaluadas se concluye que el proceso fundamental es el de la selección de los datos. Todo el proceso se verá frustrado si el conjunto de datos con el que se va a trabajar no es el adecuado para el tipo de análisis que se quiere desarrollar.

Algoritmos de Minería de Datos:

La minería de datos se denomina al proyecto de investigación que se centraliza en un subconjunto de estados de los procesos de KDD. Tal y como define, una vez más, (Fayyad, Piatetsky-Shiapiro, Smyth, & Uthurusamy, 1996) la minería de datos es una disciplina que combina técnicas de aprendizaje-máquina, reconocimiento de patrones, estadística, bases de datos y visualización para extraer conceptos y patrones de interés.

De todos modos también haremos hincapié en la definición proporcionada por (Mena, 1999) la cual la describe como el proceso iterativo de extracción de patrones escondidos en grandes bases de datos usando técnicas estadísticas y de Inteligencia Artificial.

Mediante esta definición enfatizamos las raíces de la Minería de Datos, las cuales son la estadística y la Inteligencia Artificial en especial el Machine Learning. Este concepto será ampliado en el apartado 3.2.

Post-proceso de modelos:

Finalmente en esta etapa se deberán realizar las fases de:

1. Interpretación, transformación y representación de los patrones obtenidos.
2. Difusión y uso del nuevo conocimiento.

Es la fase de conclusiones y en la que se realizan las acciones de interpretación, representación y difusión y uso del nuevo conocimiento realimentando la plataforma.

Si se globalizan los procesos especificados por Fayyad se puede elaborar el diagrama especificado en la Figura 8.

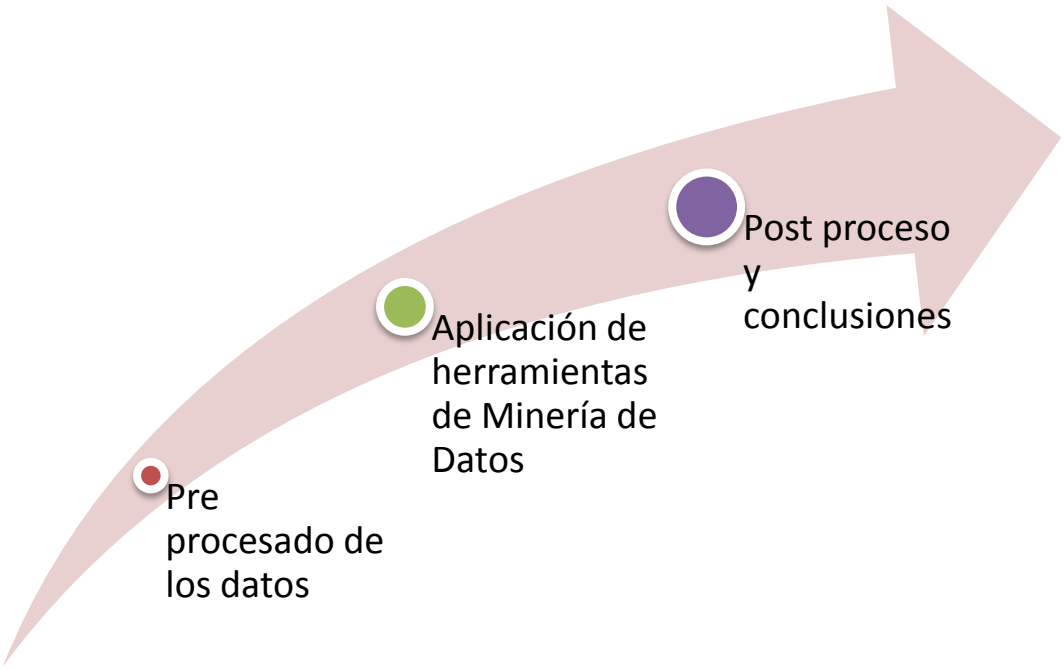


Figura 8: Diagrama procesos para el descubrimiento de conocimiento

3.1 Diseño de la información

En esta sección se detallan los principales requerimientos que se deben tener en cuenta a la hora de diseñar un sistema de información como el que se plantea.

Se ha realizado un estudio exhaustivo para identificar los errores más comunes en el planteamiento de un sistema como este, que tienen su origen en un diseño inapropiado de los datos y de las tablas que los contienen.

La información es la base del repositorio de datos energéticos que se plantea en este proyecto. Para su funcionamiento y eficiencia se deberán tener en cuenta un conjunto de normas para la correcta modelización de los datos, intentando evitar los fallos que se pueden derivar de un mal diseño de los datos.

3.1.1 Consideraciones en el diseño de un sistema de información

El diseño de un sistema donde la base de funcionamiento son los datos deberá tener en cuenta factores para el diseño que impliquen desde la calidad de estos, los costes de acceso y consulta hasta los propios parámetros básicos en el diseño de cualquier aplicación. Es por ello que se

van a describir un conjunto de los más importantes y que consideraciones se deben tener en cada uno de los casos. (Cacciapaglia, 2008)

1) Calidad de los datos

En el sistema almacenará datos históricos y se encargará de recoger aquellos que reciba en tiempo real. Es por ello que se deberá tener en cuenta la calidad de los datos que vaya a recibir, teniendo en cuenta los errores y teniendo que corregirlos a la vez que se reciban.

2) Coste de latencia

Considerando la latencia como el retardo que es producido por la demora en la propagación y transmisión de paquetes dentro de una red. La latencia representa un elemento clave para evaluar en el diseño del sistema.

Evaluación en tres momentos distintos:

- Latencia a nivel de datos: tiempo necesario para su recepción, adecuación y preparación para los procesos de análisis.
- Latencia de análisis: intervalo de acceso y aplicar los procesos para el análisis de los datos.
- Latencia de decisión: tiempo necesario para la extracción de las lógicas en función de los resultados de los procesos de análisis aplicados.

3) Adquisición de los datos

El sistema debe organizar y controlar todo el conjunto de información que contendrá y recibirá, es por ellos que este deberá registrar todas aquellas transacciones de datos que se realicen en el sistema. Para ello este debe ser flexible y tendrá que definir los niveles de calidad a los que los datos deben estar sujetos.

Se deberán normalizar los datos, para ello se darán tres metodologías usadas en aplicaciones BI:

- Fecha de consolidación: fecha en la que se insertaron los datos de las diferentes fuentes al sistema. Dicha información ha pasado previamente procesos de limpieza (eliminación de duplicados o incoherencias) y estructurado.
- Fecha de federación: en un sistema de BI existen datos almacenados y datos creados, por ese motivo se deberá hacer esta distinción. Mediante la fecha de federación se representaran dichos datos almacenados en las fuentes que los generaron.
- Fecha de propagación: se utiliza para copiar algunos casos o cambios en los datos. La propagación puede ser síncrona o asíncrona. La propagación síncrona representa aquella mediante la que se crean copias en el mismo momento en la que se produce el cambio y la asíncrona representa en la que se guardan los cambios en un pre-definido período de tiempo, por ejemplo por la noche.

4) Arquitectura escalable

En el caso de herramientas BI la escalabilidad implica se ve afectada a muchos niveles dado que se deben tener en cuenta diferentes aspectos en las diferentes dimensiones de escalabilidad de un proyecto que implica hardware y software.

Se debe controlar:

- Disponer de un hardware potente para poder responder a una alta demanda del sistema (mayor número de usuarios).
- Disponer de una buena estructura en el caso que se deban añadir más servidores.

5) Plataformas y bases de datos

En el momento de elaboración de una nueva aplicación una de las primeras cuestiones que se deben tratar es la plataforma sobre la cual se quiere trabajar. Esta cuestión se debe principalmente a los costes que en licencias pueden representar el desarrollo bajo una plataforma u otra.

Al hacer la búsqueda sobre las diferentes soluciones se ha podido comprobar que al tratarse de herramientas en muchas ocasiones muy enfocadas para que las empresas puedan extraer la lógica de su negocio, las empresas que ofrecen este tipo de software se han preocupado en poder ofrecer sus sistemas para las distintas plataformas existentes. Su ímpetu por estar a la última y poder ofrecer los mejores servicios a sus clientes les lleva incluso a la adaptación de nuevas tecnologías como el Ipad.

“Con MicroStrategy y la nueva Apple iPad, los usuarios móviles pueden acceder a la información que precisen, donde y cuando lo necesiten.” (Microstrategy, 2010)

6) Flexibilidad en el diseño de los procesos

Este tipo de sistemas donde los datos son, no solo un elemento de consulta de información si no, un elemento de trabajo para la extracción de información serán muchos los procesos que operaran con estos datos. Las lógicas que se quieran extraer de estos son muchas y cambiantes al largo del tiempo. Es por ello que será muy importante que los procesos mediante los cuales se realizará la extracción de conocimiento sean flexibles.

Factores de calidad

Como ya se ha citado anteriormente son muchos los factores que influyen en el diseño y desarrollo de aplicaciones como ésta. No existe, a priori, ninguna ISO(International Organization for Standardization) pero se va a estudiar los principales factores de calidad que se podrían aplicar a un data warehouse basándose en la ISO 9126 del standard [ISO97]. (Vassiliadis, Quix, Vassiliou, & Jarke, 2008). La ISO 9126 es la encargada de definir cuáles son los principales parámetros para la evaluación del software. (Figura 9)

Dimensiones de calidad	**Factores de calidad**
Funcionalidad	
Idoneidad	Requerimientos software
Exactitud	Integridad, precisión y consistencia de los datos
Interoperabilidad	Módulos que interactuaran en el sistema.
Seguridad	Módulos que evitaran el acceso no permitido, accidental o deliberado, al sistema.
Cumplimiento de normas	Módulos que no se ajusten a los estándares o regulaciones.
Fiabilidad	
Madurez	Frecuencia de errores por fallos en el software.
Recuperabilidad	Ocasiones en las que el software no es capaz de recuperarse o recuperar los datos en caso de fallo. Tiempo y esfuerzo en restablecerse.
Tolerancia a fallos	Ocasiones en las que el software no es capaz de mantener el nivel de funcionamiento.
Usabilidad	
Aprendizaje	Porcentaje de aceptación por parte de los usuarios.
Comprensión	Porcentaje de aceptación por parte de los usuarios.
Operatividad	Predicción del tiempo por operación.
Atractivo	Porcentaje de satisfacción por parte de los usuarios.
Eficiencia	
Comportamiento en el tiempo	Tiempo de respuesta, procesadoy ratios de salida
Comportamiento de recursos	Volumen de información que el sistema extrae o carga. Tamaño máximo de datos que el sistema puede suportar.
Mantenimiento	
Estabilidad	Número de rutas lógicas en un módulo, el control de flujo de intersección, tamaño, etc.
Facilidad de análisis	Porcentaje de análisis, índice de legibilidad.
Facilidad de cambio	Número de rutas lógicas en un módulo, el control de flujo de intersección, tamaño, etc.

Facilidad de pruebas	

Figura 9: Dimensiones de calidad (ISO, 2010)

3.1.2 Consideraciones en el modelado de la información

En esta sección se van a explicar algunos de los conceptos más básicos en el ámbito de las bases de datos, entenderlos y saberlos aplicar serán la base a la hora de desarrollar un gran almacén de datos como el que se plantea en este proyecto.

1) Diseño del Modelo dimensional:

En el proceso de creación del modelo se deben analizar los datos en función de un proceso de negocio para:

- identificar la granularidad de las tablas de hechos
- identificar las dimensiones y atributos asociados
- y evaluar los hechos numéricos.

Un modelo dimensional contiene los mismos datos y relaciones que un modelo normalizado en la 3FN, pero estructurado de manera diferente.

Este permite mejorar el entendimiento y desempeño de consultas al DW

Las construcciones primarias son: las tablas de dimensiones y las tablas de hechos.

Tablas de dimensiones

Las tablas de dimensiones contienen la descripción de atributos y características relacionadas con valores tangibles y específicos, tales como clientes, productos, representantes de ventas.

Los atributos de dimensión son usados para limitar o agrupar. Las relaciones jerárquicas N:1 son normalizadas en tablas de dimensión simples.

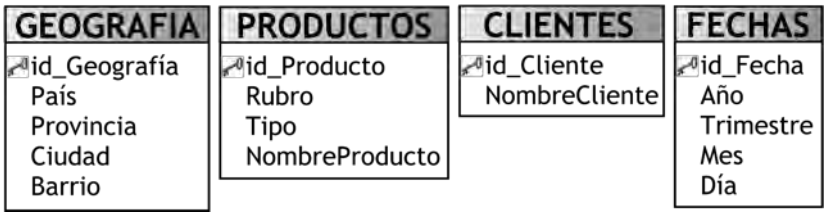


Figura 10: Visualización del diseño mediante Tablas de dimensiones (Dataprix, 2010)

Se pueden distinguir varios tipos de dimensiones:

- **Dimensiones Normales**: aquellas que agrupan diferentes atributos relacionados por el ámbito al que se refieren.
- **Dimensiones Causales**: aquellas que incluye atributos que pueden causar cambios en los procesos de negocio.

- **Dimensiones Heterogéneas**: dimensiones que agrupan conjuntos heterogéneos de atributos, que no están relacionados entre sí.
- **Dimensiones Roll-Up**: son dimensiones subconjunto de otra, necesarias para el caso en que se tratan tablas de hechos con diferente granularidad.
- **Dimensiones Junk**: dimensiones que agrupan indicadores de baja cardinalidad como pueden ser flags o indicadores.
- **Dimensiones Role-playing**: cuando una misma dimensión interviene en una tabla de hechos varias veces (por ejemplo, la fecha en una tabla de hechos donde se registran varias fechas referidas a conceptos diferentes), es necesario reutilizar la misma dimensión, pues no tiene sentido crear tantas dimensiones como usos se hagan de ella. Para ello se definen las dimensiones Role-playing. Se pueden crear vistas sobre la tabla de la dimensión completa que permiten utilizarla varias veces o jugar con los alias de tabla. La misma dimensión juega un rol diferente según el sitio donde se utiliza.
- **Dimensiones Degeneradas**: dimensiones que no tienen ningún atributo y por tanto, no tienen una tabla especifica de dimensión. Solo se incluye para ellas un identificador en la tabla de hechos, que identifica completamente a la dimensión (por ejemplo, un pedido de ventas). Interesará tener identificada la transacción (para realizar DataMining, por ejemplo), pero los datos interesantes de este elemento se encontraran repartidos en las diferentes dimensiones (cliente, producto, etc.).
- **Mini dimensiones o Dimensiones Outrigger**: conjunto de atributos de una dimensión que se extraen la tabla de dimensión principal pues se suelen analizar de forma diferente. El típico ejemplo son los datos socio demográficos asociados a un cliente (que se utilizan, por ejemplo, para el DataMining).

Cada una de las dimensiones tiene una clave que identifica cada uno de los registros que la conforman. Para definir esta clave se introduce el concepto de las Claves Subrogadas, en inglés Surrogated Keys, que son identificadores que se permitirán optimizar las consultas SQL evitando las limitaciones de las claves existentes.

Las Claves Subrogadas

Las claves subrogadas son aquellas que se definen artificialmente, son de tipo numérico secuencial, no tienen relación directa con ningún dato y no poseen ningún significado en especial. Són las claves existentes en los OLTP (**Procesamiento de Transacciones En Línea** (**O**n**L**ine **T**ransaction **P**rocessing) son un tipo de sistemas que facilitan y administran aplicaciones transaccionales, para la entrada de datos, recuperación y procesamiento de transacciones.

Las claves Subrogadas definen una serie de ventajas más:

- Ocupan menos espacio y son más performantes que las tradicionales claves naturales.
- Son de tipo numérico entero (auto numérico o secuencial).
- Simplifican la construcción y mantenimiento de índices.
- Un Data Warehouse, concepto que se analiza en la sección 3.1.2, no dependerá de la codificación interna del OLTP.
- Si se modifica el valor de una clave en el OLTP, el Data Warehouse lo tomará como un nuevo elemento, permitiendo de esta manera, almacenar diferentes versiones del mismo dato.
- Permiten la correcta aplicación de técnicas SCD (Dimensiones lentamente cambiantes).

Una clave subrogada debe ser el único campo que sea clave principal de cada tabla de dimensión. Una forma de implementación sería, a través de la utilización de herramientas ETL, mantener una tabla que contenga la clave primaria de la tabla del OLTP y la clave subrogada correspondiente a la dimensión del Data Warehouse.

Tablas de hechos

Los hechos son los indicadores de negocio que dan sentido al análisis de las dimensiones.

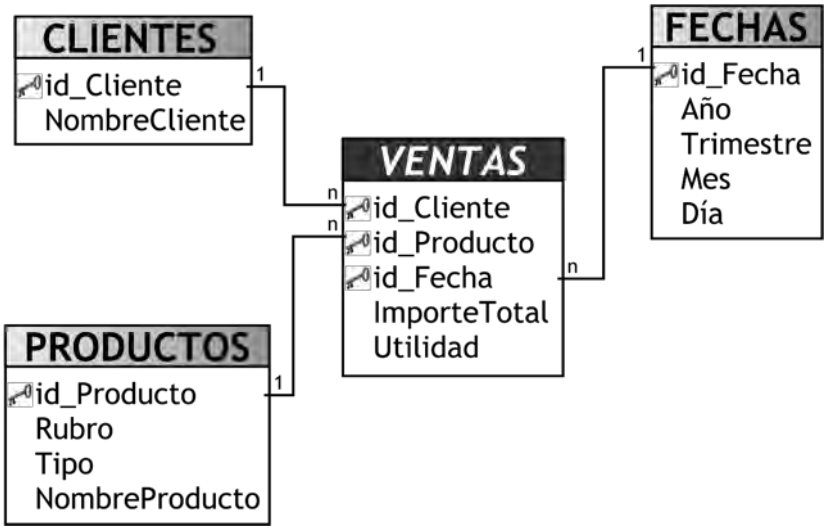


Figura 11: Visualización del diseño mediante Tablas de hechos. (Dataprix, 2010)

Principales características:

- Contienen métricas derivadas de un proceso de negocio o un evento.
- El MD debe ser estructurado alrededor de un proceso del negocio
- Se diseñan vistas similares y consistentes de los datos para toda la organización.
- La granularidad de la tabla de hechos, debe ser el más atómico posible
- Esto permite mayor flexibilidad y extensibilidad.

Tipos de tablas de hechos:

- **Transaction Fact Tables:** representan eventos que suceden en un determinado espacio-tiempo. Se caracterizan por permitir analizar los datos con el máximo detalle. Reflejan las transacciones relacionadas con nuestros procesos de negocio (ventas, compras, inventario, contabilidad, etc).
- **Factless Fact Tables:** Son tablas que no tienen medidas y representan la ocurrencia de un evento determinado. Por ejemplo, la asistencia a un curso puede ser una tabla de hechos sin metricas asociadas.
- **Periodic Snapshot Fact Tables:** Son tablas de hecho usadas para recoger información de forma periódica a intervalos de tiempo regulares sobre un hecho. Permiten tomar una foto de la situación en un momento determinado (por ejemplo al final del dia, de una semana o de un mes). Un ejemplo puede ser la foto del stock de materiales al final de cada día.
- **Accumulating Snapshot Fact Table:** representan el ciclo de vida completo de una actividad o proceso, que tiene un principio y final. Suelen representar valores acumulados.

- **Consolidated Fact Tables:** tablas de hechos construidas como la acumulación, en un nivel de granuralidad o detalle diferente, de las tablas de hechos de transacciones.

Se pueden distinguir diferentes **tipos de medidas o indicadores**, basadas en el tipo de información que recopilan así como su funcionalidad asociada:

- **Métricas**: valores que recogen el proceso de una actividad o los resultados de la misma. Esto medidas proceden del resultado de la actividad de negocio.
  - Métricas de realización de actividad (*leading*): miden la realización de un actividad. Por ejemplo, la participación de una persona en un evento.
  - Métricas de resultado de una actividad (*lagging*): recogen los resultados de una actividad. Por ejemplo, la cantidad de unidades vendidas.
- **Indicadores clave**: por este concepto se entienede por valores correspondientes que hay que alcanzar, y que suponen el grado de asunción de los objetivos. Estas medidas proporcionan información sobre el rendimiento de una actividad o sobre la consecución de una meta.
  - **Key Performance Indicator** (KPI): Indicadores clave de rendimiento. Más allá de la eficacia, se definen unos valores que se explican en qué rango óptimo de rendimiento se deberían situar al alcanzar los objetivos. Son métricas del proceso.
  - **Key Goal Indicator** (KGI): Indicadores de metas. Aquí se pueden incluir por ejemplo, el objetivo de rentabilidad del proceso de negocio de ventas.

Según si se des normalizan las tablas de dimensiones o no, resultará un esquema de estrella (star) o copo de nieve (snowflaked). Kimball recomienda utilizar siempre la des normalización total, pero está claro que hay situaciones en las que no queda más remedio que pasarse al esquema copo de nieve (aunque solo sea para alguna dimensión).

2) Modelos de datos

Esquema de estrella (Star)

El esquema en estrella es la representación más importante del modelo dimensional. Este estará formado por hechos y dimensiones.

Un hecho**:** es todo objeto de análisis. Este se representa en forma de tabla de hechos o fact table.

Las dimensiones: analizan los hechos. Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales.

Los hechos tienen columnas de datos denominadas métricas y las dimensiones tienen columnas que representan los niveles de jerarquías.

Representa un diseño lógico relacional de base de datos que resulta en que las tablas de hechos representan la Tercera Forma Normal (3FN) y las dimensiones representan la Segunda Forma Normal (2FN).

Este esquema es ideal por su simplicidad y velocidad para ser usado en análisis multidimensionales (OLAP, Data Marts, EIS, etc.). Permite acceder tanto a datos agregados como de detalle.

El diseño de esquemas en estrella permite implementar la funcionalidad de una base de datos multidimensional utilizando una clásica base de datos relacional.

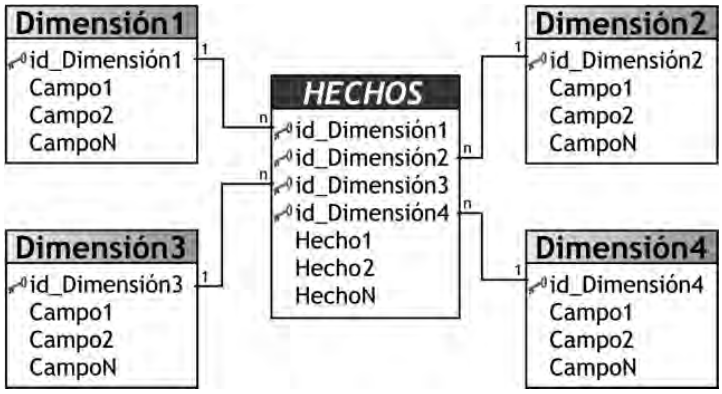


Figura 12: Esquema en estrella (Dataprix, 2010)

Esquema de copo de nieve (Snowflaked)

Es una variación del esquema de estrella. Es un esquema más complejo que el esquema de estrella porque las tablas que describen las dimensiones están normalizadas. Se da cuando alguna de las dimensiones se implementa con más de una tabla de datos. La finalidad es normalizar las tablas y así reducir el espacio de almacenamiento al eliminar la redundancia de datos; pero tiene la contrapartida de generar peores rendimientos al tener que crear más tablas de dimensiones y más relaciones entre las tablas (*JOINS*) lo que tiene un impacto directo sobre el rendimiento. Está orientado a facilitar el mantenimiento de dimensiones.

Las tablas de dimensiones en este modelo representan relaciones normalizadas (3NF) y forman parte de un modelo relacional de base de datos.

Finalmente este diseño estará muy relacionado con los cubos OLAP.

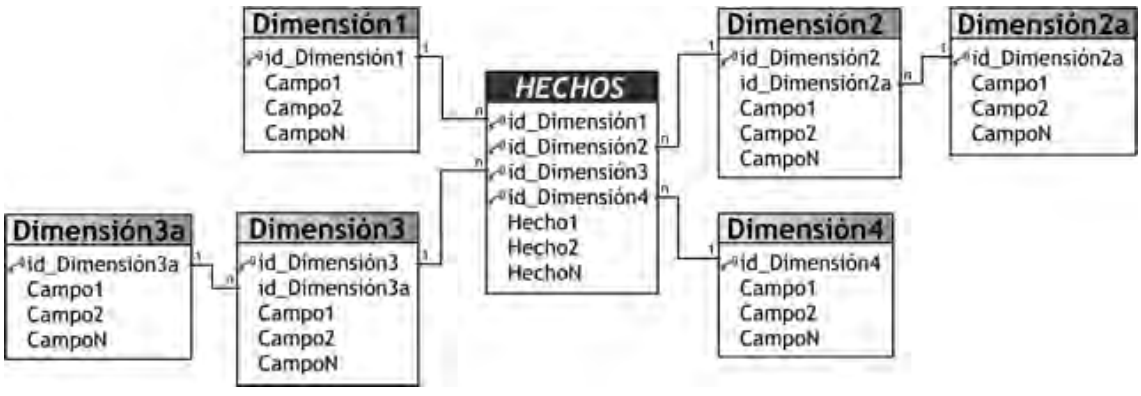


Figura 13: Esquema de copo de nieve (Dataprix, 2010)

Desventajas:

1) Las tablas de hecho ocupan el noventa por ciento del almacenamiento, (el beneficio es poco).
2) Normalizar las tablas de dimensión pueda deteriorar la ejecución de un DWH.
3) Para extraer datos de las tablas en esquema de copo de nieve, a veces hay que vincular muchas tablas en las sentencias SQL que puede llegar a ser muy complejo y difícil para mantener.

Ventajas:

1) Es apropiado si se presenta alguna de las siguientes condiciones:
   - Una dimensión es esparcida
   - Una dimensión tiene una lista muy larga de atributos
2) En la práctica, muchos DWH normalizarán algunas dimensiones y otros no (usan una combinación de copo de nieve y de estrella) (Malinowski & Zimányi, 2008)

3.1.3 Consideraciones en el modelo físico

Hasta este momento se han analizado aquellas configuraciones físicas más óptimas para el modelado de la información, optimizando el diseño para una buena gestión de un gran repositorio de datos.

Sin embargo, REPENER es un sistema de información online. En la siguiente sección se analizan las dos principales arquitecturas para el almacenamiento teniendo en cuenta el factor online del proyecto.

1) Repositorios

Un repositorio es un almacén el cual permitirá almacenar diferentes tipos de datos. Pero un repositorio no es solo la tecnología que gestiona los datos, un repositorio al encontrarse en la red, es un sistema digital de almacenamiento social. (Bradley, 2006)

Los repositorios pueden estar en internet o no, en un medio extraíble como un CD, en un disco duro, etc. Pero es este principio social mediante internet el que le ha dotado de mayor sentido.

Pero la principal limitación que se presenta con un sistema como este es el grado de diseño que se debe realizar sobre los datos. La granularidad, el grado de complejidad de los datos y el tiempo de respuesta al acceso a dicha información también serán un factor decisivo en el diseño del sistema. La granularidad es el concepto mediante el cual se indica el nivel atómico de los datos.

Pero lo que más interesa del concepto repositorio es el hecho de que se localizan en internet y como ya se ha comentado anteriormente, REPENER es “Sistema de almacenamiento en internet”. Este tipo de estructura tiene nombre propio y es Open Repositories.

Un OpenRepositori es un repositorio localizado en internet más o menos público que almacena y mantiene archivos digitales. Su cualidad más destacada es el hecho de al localizar la

información directamente en internet, las entradas a los datos y archivos almacenados en esta pueden ser referenciados y etiquetados desde cualquier página web.

Es por ello que algunos repositorios implementan y almacenan tipos de datos algo más complejos con el fin de sacar el máximo partido a la información que en éste almacenan y dotarlos de sostenibilidad. Dentro de este enfoque se va a citar el caso práctico de Fedora.

Fedora

Es un repositorio online diseñado para el almacenamiento de cualquier tipo de dato digital.

Fedora define un modelo genérico de objeto digital que se puede utilizar para tratar las características esenciales para la mayoría de tipos de datos digitales, incluyendo documentos, imágenes, libros electrónicos, multimedia objetos de aprendizaje, bases de datos, metadatos y muchos otros. Este modelo de objeto digital es componente fundamental de la arquitectura Modelo de contenido proporcionado por Fedora (Fedora, 2010)

Con esta arquitectura lo que se pretende es:

- Definir patrones reutilizables como los modelos de contenido con el fin de reducir el esfuerzo de crear o capturar, ingerir, almacenar, administrar, conservar, transformar, y acceder a contenido digital.
- Definir una representación de la información y procesado de la arquitectura. El contenido digital no se define por su formato o tecnología, y también pueden incorporar funciones como parte de su naturaleza. Por ejemplo, al acceder a una imagen digital poder obtener su resolución y calidad de color. Esas características que deben estar presentes para ofrecer las características esenciales de los contenidos digitales. Estas mismas características también facilitaran el intercambio de los contenidos.

Soluciones OpenSource

DSpace: Código Abierto

DSpace es un software de código abierto que provee herramientas para la administración de colecciones digitales, y comúnmente es usada como solución de repositorio institucional. Soporta una gran variedad de datos, incluyendo libros, tesis, fotografías, video, datos de investigación y otras formas de contenido. Los datos son organizados como ítems que pertenecen a una colección; cada colección pertenece a una comunidad.

DSpace es una aplicación cliente/servidor que se gestiona vía web, es decir, que la mayor parte de procesos pueden llevarse a cabo con un navegador estándar como Internet Explorer, Firefox u Opera.

Fedora

Fedora es una arquitectura modular que consigue mejor interoperabilidad gracias a su integración de datos, interfaces, y los mecanismos/módulos claramente definidos. Fedora es la arquitectura subyacente de un repositorio digital y un gestor de activos digitales en el que se pueden almacenar muchos tipos de datos digitales, repositorios institucionales, además de la posibilidad de la creación de bibliotecas digitales.

Sus características y ventajas más relevantes son el hecho de ser una arquitectura modular y fácilmente adaptable.

DuraSPace

Nace de la alianza entre alianza Fedora Commons and DSpace Foundation.

Es una triada como resultado de unir dos organizaciones sin fines de lucro creada para mantener sus repositorios de código abierto. Las dos organizaciones combinan sus fuerzas para realizar una misión común y para ampliar su oferta en informática en la nube y académico / científico "ciberinfraestructura" para las universidades, bibliotecas e instituciones de investigación, archivos, museos, organizaciones no gubernamentales, y más.

Solución Comercial: Open Repocitory

Es una solución de almacenamiento para construir y mantener repositorios DSpace.

Es una solución de BioMed Central, que construye y mantiene depósitos en nombre de las organizaciones. Este servicio cumple OAI (Open Archives Iniciative), se basa en el último software de DSpace y ofrece una solución profesional, fiable,

Usando el software de código abierto DSpace, Open UAB gestiona y mantiene todo el contenido dentro de los repositorios del cliente garantizándoles una mayor visibilidad para las organizaciones y permitiendo a los administradores de tiempo y libertad para concentrarse en su información.

2) Data Warehouse

El concepto de Data Warehouse surge tras las dificultades de los sistemas tradicionales en satisfacer las necesidades informacionales. Este término acuñado por Bill Inmon, se traduce literalmente como Almacén de Datos aunque sus funciones re-definen un Data-Warehouse como una herramienta diseñada específicamente para entregar resultados más rápidos en soluciones esenciales de inteligencia empresarial, de generación de informes y de almacenamiento de datos, en cualquier hardware y sistema operativo estándar.

Es una base de datos orientada al análisis y que será el corazón de todo proyecto de Business Intelligence. Esta base de datos deberá soportar todos los tipos de herramientas de análisis que se vayan a utilizar. Este permitirá aplicar análisis de datos, para extraer, transformar y cargar datos, así como gestionarlos.

Un Data Warehouse representa una estructura de almacenamiento mucho más compleja que un repositorio. Representa un sistema que almacenará y recibirá un alto grado de información diaria y a la que se aplican técnicas de tratamiento para poder extraer conocimiento de la información almacenada. Es por ello que uno de los factores más importantes a tener en cuenta será como se formatee la información y su tiempo de respuesta para que su acceso y tratamiento no resulte una carga para el sistema.

El objetivo más importante será obtener la información necesaria que sirva de base para la toma de decisiones tanto a escala estratégica como táctica. También será vital importancia la visión histórica de todas las variables analizadas, y el análisis de los datos del entorno.

A. Principios básicos sobre los datos

El elemento principal de este tipo de sistemas es la estructura de la información. Este concepto representa un sistema de información homogénea y fiable, en una estructura basada en la consulta y el tratamiento jerarquizado de la misma, y en un entorno diferenciado de los sistemas operacionales.

Según definió Bill Inmon, el Data Warehouse se caracteriza por ser:

Integrado: los datos almacenados en el Data Warehouse deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.

Temático: sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del Data Warehouse. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.

Histórico: el tiempo es parte implícita de la información contenida en un Data Warehouse. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el Data Warehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el Data Warehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

No volátil: el almacén de información de un Data Warehouse existe para ser leído, y no modificado. La información es por tanto permanente, significando la actualización del Data Warehouse la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía. (Malinowski & Zimányi, 2008)

Pero todavía se deben tener en cuenta otros aspectos sobre los datos que se van a almacenar, introduciendo el concepto de metadatos. Los metadatos son datos relativos a otros datos, este concepto permite mantener información de la procedencia de la información, la periodicidad de refresco, su fiabilidad, forma de cálculo, etc., de los datos que se mantienen en el sistema. Serán estos metadatos los que permitirán simplificar y automatizar la obtención de la información.

Es por ello que los objetivos que los metadatos deberán cumplir son aquellos que ayuden al usuario final y a los administradores del sistema a introducir y trabajar con los datos almacenados:

- Perfil usuario: ayudándole a acceder al sistema, indicando que información hay y qué significado tiene. Ayudar a construir consultas, informes y análisis.
- Perfil del administrador: mediante herramientas de gestión de información histórica, administración del sistema, elaboración de programas de extracción de la información, especificación de las interfaces para la realimentación a los sistemas con los resultados obtenidos, etc.

El papel de los metadatos es uno de los más complejos dado que si no se entienden los datos estos caerían en una especie de saco roto del cual no se podrían extraer ningún tipo de conclusión. Es por ello que se dedicará un apartado más extenso en una sección especializada del documento (ver apartado 1.3 sección 3).

B. Procesos básicos

Para comprender el concepto de Data Warehouse, es importante considerar los procesos básicos que lo forman. Tal y como se ve en la Figura 14 distinguimos:

Extracción: obtención de información de las distintas fuentes tanto internas como externas.
Elaboración: filtrado, limpieza, depuración, homogeneización y agrupación de la información.
Carga: organización y actualización de los datos y los metadatos en la base de datos.
Explotación: extracción y análisis de la información en los distintos niveles de agrupación.

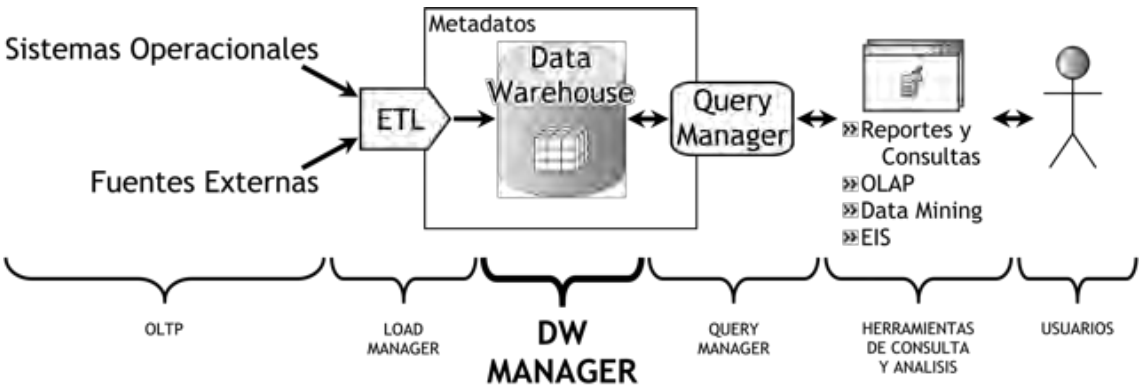


Figura 14: Esquema general procesos en un Data Warehouse (Dataprix, 2010)

Desde un punto de visto totalmente global el proceso más importante radica en la explotación de la información pero el éxito del sistema dependerá directamente de los tres procesos iniciales dado que son los que alimentan al sistema y suponen todo el esfuerzo a la hora del desarrollo del sistema.

C. Conclusiones de explotación

Hasta el momento se ha hablado de cómo deben ser los datos y los procesos que trabajan en el sistema pero todavía no se ha comentado del tipo de herramientas que se darán y cuál será su objetivo.

Pero como ya se ha comentado un Data Warehouse será una base de almacenamiento pero que permitirá responder mediante herramientas a cuatro cuestiones muy básicas sobre los datos gracias a todos los parámetros ya comentados.

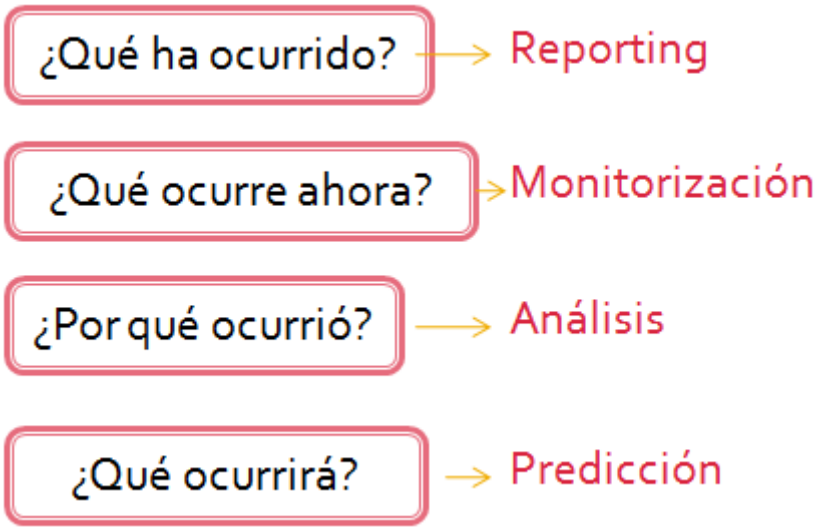


Figura 15: Esquema cuestiones que responde un Data Warehouse

Mediante las cuatro principales cuestiones que se observan en la Figura 15 se resumen los beneficios que un Data Warehouse puede aportar:

- Proporciona herramientas para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén, obteniendo un valor añadido para el negocio de dicha información.
- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes.

D. Factores críticos

Pero una de las principales variables que más preocupa en el desarrollo de un sistema que integra desarrollo de herramientas de BI es que el usuario final sea capaz de comprender las herramientas que se le proporcionen y su funcionamiento, es por ello que se van a describir los principales factores críticos de éxito en común que ya fueron mencionados por Lokken en el 2001:

- Proveer acceso a datos adecuados. Organizar los datos para que el usuario sepa localizarlos.
- Incrementar la habilidad de los usuarios para entender los resultados. Saturar a las personas de números en estos días crea más problemas que los que resuelven. Diez años atrás el problema era obtener los datos, pero hoy en día tiene que ver más con el manejo de ellos.
- Incrementar el entendimiento de los negocios por parte de los usuarios. Conocer que es lo que los datos dicen es algo bueno, pero en la actualidad es necesario saber qué hacer con ellos. Este conocimiento es difícil de construir dentro de una pieza de software.
- Ayudar a comunicar los hallazgos y tomar acciones. Es raro que un individuo pueda ejecutar cualquier cosa significativa dentro de una organización sin involucrar a otros.

E. Modelos de patrones arquitectónicos según arquitectura de los datos

Existen diversas metodologías o enfoques para la construcción de un Data Warehouse. Las más importantes son las que definieron Ralph Kimball y Will Inmon.

Para entender los enfoques se va a definir el concepto de DataMart.

DataMart: es un repositorio de datos específico y construida para soportar una línea de negocio simple. Es un subconjunto de los datos del Data Warehouse con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica. Al igual que en un data warehouse, los datos están estructurados en modelos de estrella o copo de nieve y un data mart puede ser dependiente o independiente de un data warehouse.

Esta estrategia es particularmente apropiada cuando el Data Warehouse central crece muy rápidamente y los distintos departamentos requieren sólo una pequeña porción de los datos contenidos en él. La creación de estos Data Marts requiere algo más que una simple réplica de los datos: se necesitarán tanto la segmentación como algunos métodos adicionales de consolidación.

Metodologías para desarrollar un Data Warehouse:

	Top-Down	**Bottom-Up**
Profesional	Bill Inmon	Rodolfo Kimball
Metodología y Arquitectura		
Énfasis	DWH	DataMarts
Diseño	Modelo normalizado basado en la empresa, mediante bases de datos departamentales.	El modelo dimensional de DataMarts, usa esquema de estrella.

Arquitectura	Compuesto de varios niveles de áreas de interés y DataMarts dependientes.	Área de interés y DataMarts. Los DataMarts representan un proceso de negocio único. La coherencia de la información se dá mediante el bus de datos y las dimensiones.
Complejidad	Más complejo	Más sencillo
Diseño físico	Más complejo	Más sencillo
Modelización de los datos		
Orientación de la información	Por temas o por datos	Por procesos
Data set	DWH datos a nivel atómico; DataMarts datos sumarizados	Contiene datos atómicos y sumarizados
Accesibilidad del usuario final	Baja	Alta
Filosofía		
Principales usuarios	Profesionales IT (Information Technology)	Usuarios finales
Objetivo	Base de datos adecuada basada en métodos y tecnologías.	Solución que facilite a los usuarios finales la consulta de los datos y les permita obtener respuestas razonables.

Figura 16: Esquema comparativo tipos de implementación de un Data Warehouse (Fuente: propia)

En 1990 Bill Inmon presenta la primera arquitectura: Top-Down. Inmon ve la necesidad de transferir la información de los diferentes sistemas de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis. Los datos son extraídos de los sistemas operacionales por los procesos ETL y cargados en las áreas de stage, donde son validados y consolidados en el DW corporativo, donde ademas existen los llamados metadatos que documentan de una forma clara y precisa el contenido del DW. Una vez realizado este

proceso, los procesos de refresco de los Data Mart departamentales obtienen la información de él, y con las consiguientes transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos, refrescando su contenido.

Resumiendo:

- El DWH usa modelo de datos de toda la empresa
- El DWH es un depósito de Data Marts
- Más tiempo para implementar.
- Fracasos por falta de paciencia y de compromiso

En 1996 Ralph Kimball presenta una nueva arquitectura: Bottom-Up. El Data Warehouse es un conglomerado de Data Marts, representan una copia de los datos transaccionales estructurados de una forma especial para el análisis, de acuerdo al Modelo Dimensional, que incluye las dimensiones de análisis y de sus atributos, su organización jerárquica, incluso los diferentes factores que se quieren analizar. Por un lado existen tablas para representar las dimensiones y por otro lado las tablas para los hechos (las facts tables). Los diferentes Data Marts están conectados y estructurados entre sí por un bus structure (que permite que los usuarios puedan realizar querys sobre los diferentes DataMarts, dado que el bus contiene los elementos en común que los comunican).

Inicia con un Data Mart, luego otros Data Marts.

- El flujo de datos:
    - Fuente: un Data Mart
    - La unión de los Data Mart formarán el DWH
- Rápido de implementar, por etapas
- Necesita asegurar:
    - La consistencia de los metadatos.
    - Estar seguro que cada cosa es llamado por su nombre.
- Hacer que la información sea de fácil acceso
  (Oporto, 2006)

Si se enfoca esta comparativa hacia las necesidades según los datos que el sistema debe controlar, se extrae la tabla siguiente:

	Inmon	**Kimball**
Naturaleza de las decisiones	Tácticas	Estratégicas
Requerimientos de la información	Integración de toda la organización.	Áreas individuales de negocio.

Escalabilidad	La evolución de las necesidades es crítico.	Alta necesidad de adaptar nuevos datos.
Persistencia de los datos	El grado de cambio de las fuentes es alto.	La fuente de los sistemas es estable.
Trabajadores cualificados	Grandes grupos de especialistas.	Pequeños grupos generalizados
Tiempo de desarrollo	Los requisitos retrasan la puesta en marcha.	Bueno cuando se necesita un almacén de datos urgente.
Coste de desarrollo	Alto coste de la puesta en marcha con menores costes en los proyectos posteriores..	Reducción de costes en la puesta en marcha. Cada proyecto posterior representa el mismo coste.

Figura 17: Esquema comparativo tipos dedesarrollo de un Data Warehouse basandose en la información (Breslin)

Solución OpenSource: Pentaho

Pentaho Business Intelligence Suite es una plataforma de inteligencia de negocios open-source comercial. Desarrollado en Java, es multiplataforma y soporta múltiples bases de datos relacionales. Esta plataforma está disponible en dos versiones: Community Edition y Enterprise Edition, esta última se puede descargar como trial y se puede utilizar por un tiempo de 30 días.

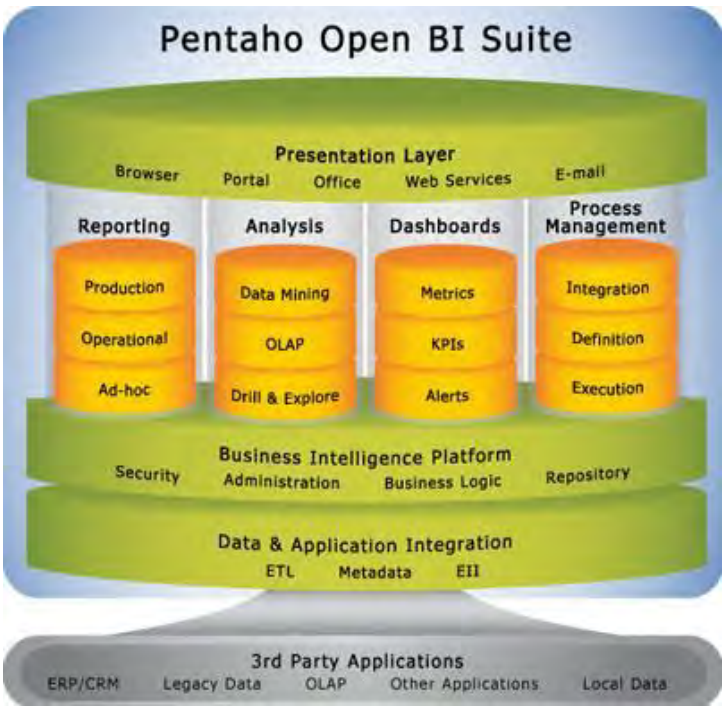


Figura 18: Esquema Suite de Pentaho (Pentaho Community, 2010)

Componentes de la plataforma

La suite de BI de Pentaho incluye las siguientes herramientas, las cuales también pueden ser usadas por separado:

Kettle (Integración de Datos / ETL): Es una herramienta de integración de datos, su rol es similar al de Integration Services de SQL Server, permite extraer datos desde fuentes externas, transformarlos y cargarlos en un destino. Este proceso es dirigido por un sistema de configuración mediante metadatos.

Mondrian (Análisis / OLAP): Es un servidor OLAP escrito en el lenguaje de programación JAVA. Su configuración de almacenamiento es ROLAP, el modelo multidimensional se ubica en una base de datos relacional, las consultas multidimensionales se traducen en consultas SQL que se envían al motor relacional configurado con antelación. Mediante conectores JDBC el servidor Mondrian accede a la base de datos relacional, permitiendo independencia respecto de éste, por ejemplo, se pueden utilizar como servidor relacional Oracle, PostgreSQL, MySQL y Microsoft SQL Server.

Pentaho Reporting (Reportes): Es un conjunto de aplicaciones y servidores que permite crear, generar y distribuir reportes a los usuarios. Permite utilizar distintas fuentes de datos: relacional, OLAP (Mondrian), metadatos de Pentaho y XML para generar los reportes. Los reportes soportan elementos visuales como encabezado, pie de página, imágenes, gráficos, entre otros.

Weka (Data Mining): Son un conjunto de herramientas para Data Mining. Posee algoritmos de clasificación, regresión, asociación y clustering que permiten realizar análisis predictivo.

Pentaho BI Platform: Incluye todo lo necesario para dar solución a problemas en BI. Esta plataforma posee un motor de Workflows que puede ser integrado a los procesos de negocios, servicios unificados de identificación, seguridad, registros, auditoría y Web Services. También integra reportes (dado por Pentaho Reporting), análisis OLAP (dado por Mondrian) y Dashboards. Cuando se utiliza este servidor no es necesario descargar los otros, puesto que vienen todos integrados en un solo sistema.

Solución Comercial: Microstrategy

“Desde 1989, MicroStrategy ha ayudado a las empresas a transformar sus datos en valiosa información de negocio. Nuestra plataforma de Business Intelligence, MicroStrategy 8™, da soluciones a todas sus consultas de negocio, reporting, necesidades de análisis avanzado y distribución de información vía web, wireless y voz. Con una gran cantidad de clientes satisfechos MicroStrategy ha demostrado ser la mejor y más completa solución de Business Intelligence.” (Microstrategy, 2010)

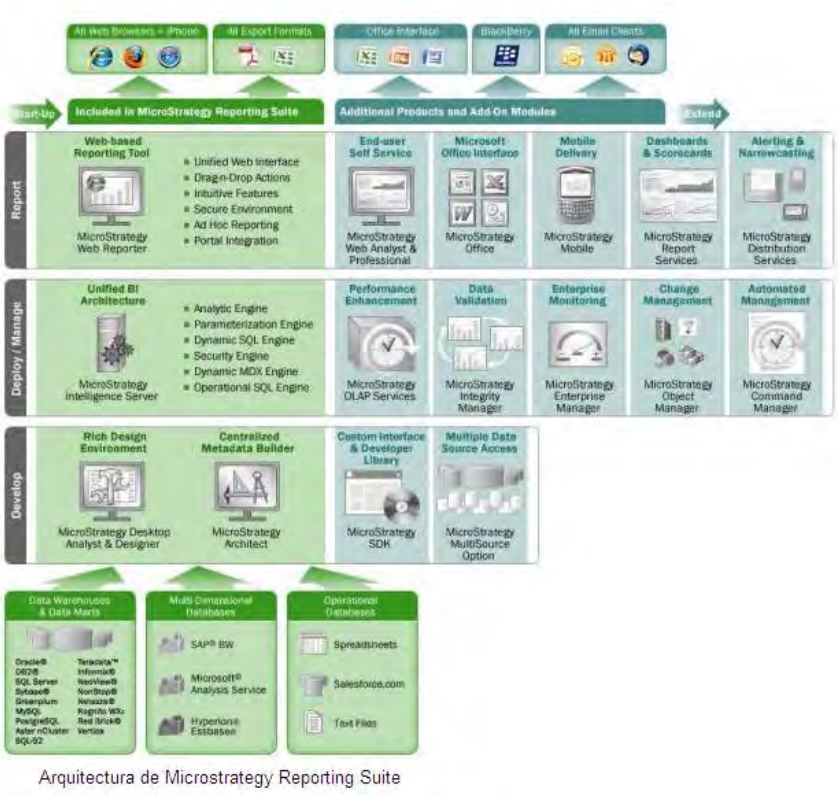


Figura 19: Esquema Suite de Microstrategy (Microstrategy, 2010)

Componentes de la plataforma

Los elementos más importantes que forman la herramienta de Microstrategy tal y como se vé en la Figura 19, son los siguientes:

MicroStrategy metadata: repositorio que almacena las definción de los objetos de MicroStrategyy la información sobre el data warehouse.
MicroStrategy Intelligence Server: servidor analitico optimizado para el reporting empresarial y para el análisis Olap.
MicroStrategy Desktop: aplicación en entorno windows que proporciona un completo abanico de funciones analíticas diseñadas para facilitar el desarrollo de informes.
MicroStrategy Web and Web Universal: interfaz de usuario altamente interactivo para la ejecución de informes y análisis.

MicroStrategy project: lugar donde se definen y almacenan los objetos del esquema y la información que se necesitan para trabajar con el sistema de reporting y análisis.
MicroStrategy Architect: herramienta para el diseño de los proyectos, que permite definir de forma gráfica todos los componentes requeridos para el proyecto desde una interfaz centralizada.

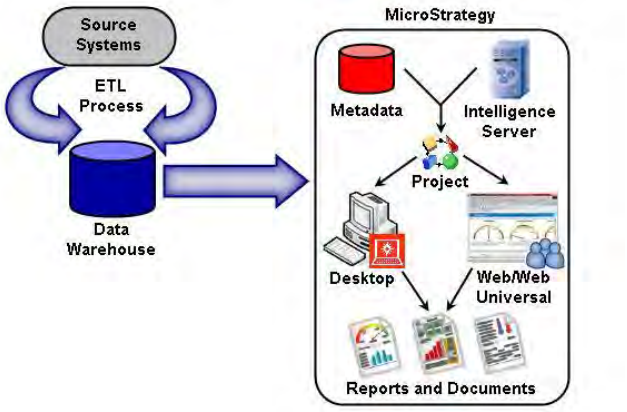


Figura 20: Esquema de los procesos involucrados en Microstrategy (Microstrategy, 2010)

Se ha podido evaluar cómo trabaja Microstrategy realizando diferentes pruebas de análisis sobre los datos. De este modo se ha podido ver como este proporciona a los usuarios las diferentes técnicas de Data Mining proporcionadas por el sistema. A la vez que al igual que enseñan en la Figura 21 se han seguido los diferentes procesos para la obtención final del conocimiento.

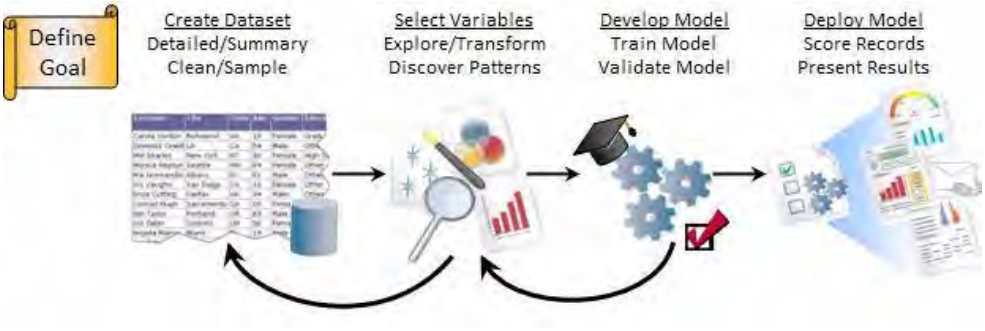


Figura 21: Esquema Mining Process de Microstrategy

Conclusiones de la sección

A lo largo de todo el proceso de estudio a través de documentos de investigación y casos de uso de sistemas de información como el que se quiere desarrollar siempre existe un factor común entre todos, la lista de los 10 errores más comunes. Es por ello que teniendo en cuenta esta lista se han tratado de analizar los factores más relevantes.
De este modo mediante esta sección se ha querido dar respuesta a tres de los componentes más importantes a la hora de desarrollar un sistema de información:

- **Consideraciones genéricas en el diseño de un sistema de información.**
- **Estudio a fondo de conceptos básicos en el desarrollo de una base de datos. Este ha sido de interés para el entendimiento de algunos conceptos y consejos introducidos en el siguiente apartado. Tales como que la estructura en estrella facilita la tarea a los usuarios exploradores encargados de encontrar nuevos patrones significativos mediante la minería.**
- **Presentación de conceptos tales como el significado y diferencias entre un Repositorio y un Data Warehouse y plataformas existentes.**

El desarrollo de un sistema de información como el que se plantea perdería toda su potencia en el caso que se implementara mediante un Repositorio. La correcta

implementación será mediante la creación de Data Marts que formaran el Data Warehouse final.

Destacar que en la elaboración de las pruebas prácticas se realizaran mediante la creación de Data Marts.

Con la información analizada en esta sección se podrán especificar las tres primeras etapas del proceso de descubrimiento de conocimiento (KDD)

1. **Determinación de las fuentes de información que pueden ser útiles.**
2. **Diseñar el esquema de un almacén de datos (Data Warehouse) que consiga unificar de manera operativa la información recogida.**

Implantación del almacén de datos que permita la "navegación" y visualización de los datos, para enfatizar los datos a ser estudiados.

3.2 Pre proceso de los datos

En el diagrama proporcionado por Fayyad en la Figura 7 se precisa que para un buen análisis de la información es necesario haber hecho un análisis, limpieza y selección de la información.

Este proceso nace con el nombre de Transformación de los datos o ETL.

Los procesos ETL tienen dos objetivos muy diferenciados que son:

- Integradores de la información de diferentes fuentes de datos: Los procesos ETL permiten extraer datos de diversas fuentes de datos. Corresponde al 70% del riesgo y esfuerzo de un proyecto de DWH. Los procesos ETL son muy necesarios para la integración de la información en un solo sistema como es el Data Warehouse. Mediante este proceso se verificará la calidad de la información y los errores que esta pueda contener.
- Especificadores de la información con la que se quiere trabajar, separándola en Bases de Datos diferenciadas o Data Marts.

Ahora que ya conocemos lo que son los Data Warehouse y los Data Marts se puede hacer la explicación pertinente sobre las funciones que los procesos ETL proporcionan a los sistemas de información.

EL proceso de ejecución de un proceso ETL:

3) Transforma los datos:
    - Desde: optimizarlo para las transacciones
    - Para: optimizarlo para el análisis y la presentación de informes.
4) Sincroniza los datos procedentes de diferentes bases de datos.
5) Limpia los datos para eliminar los errores.
6) Carga de datos en una nueva Base de Datos o Data Mart.

En función de cuando y como se aplique el proceso ETL se construirán arquitecturas diferentes en el proceso de desarrollo del Sistema de Información.

3.2.1 Diseño lineal con Data Marts independientes

Este diseño es el más sencillo dado que no se llega nunca a integrar toda la información y finalmente no permite relacionar la información entre las diferentes Bases de Datos. Los datos se trataban individualmente y el desarrollo de las soluciones se tratará por silos funcionales. Las dimensiones no se ajustan.

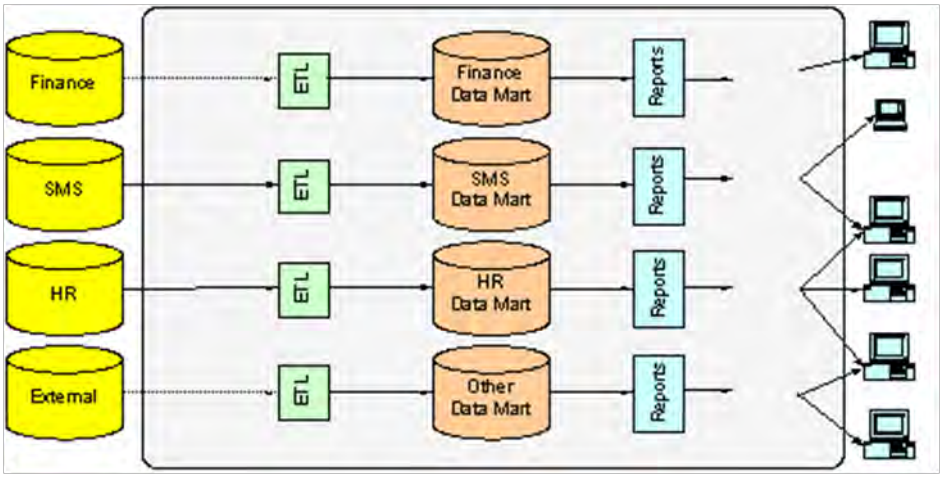


Figura 22:Esquema de un modelo de Diseño lineal

3.2.2 Diseño Top-Down (Pull)

Mediante este diseño se refina la información contenida en el sistema conforme se avanza. Como se visualiza en la figura hay un primer proceso ETL, encargado de la limpieza y puesta a punto de los datos, que deriva en el Data Warehouse, a continuación y mediante un segundo proceso ETL más complejo se construirá y constituirán los diferentes Data Marts para el análisis especializado de lo que ya se puede denominar como información.

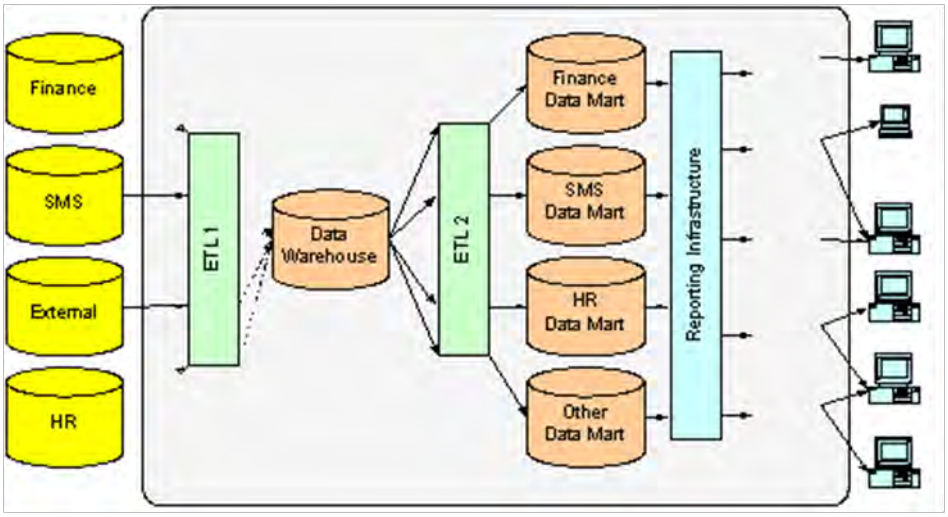


Figura 23: Esquema de un modelo de Diseño Top Down

3.2.3 Diseño Bottom-Up (Push)

Este diseño es aquel en el que las partes individuales se diseñan con detalle y luego se enlazan para formar componentes más grandes que a su vez se enlazan para formar el sistema completo. En este diseño se aplica un proceso previo ELT no demasiado complejo y a continuación se aplica el segundo proceso el cual aplicará las políticas realmente complejas. De este modo el bloque central de almacenamiento servirá meramente de elemento de sincronización de la información.

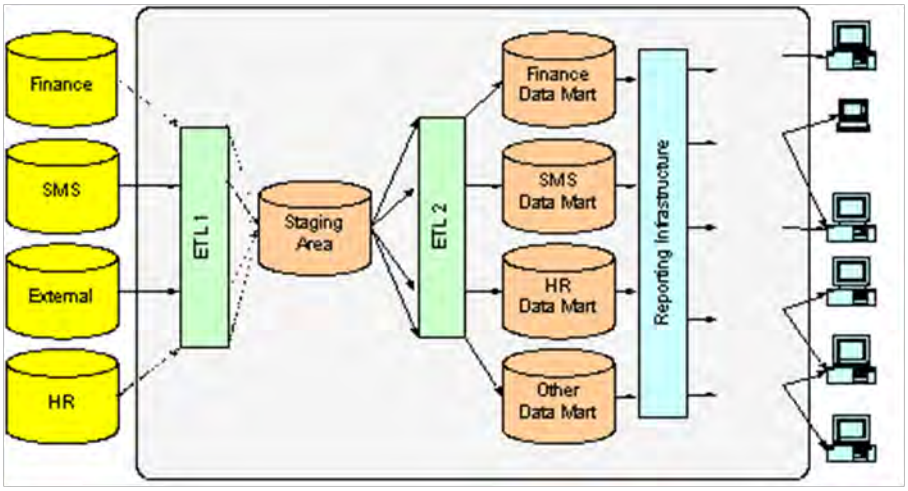


Figura 24: Esquema de un modelo de Diseño Bottom-Up

3.2.4 Diseño Central hibrido (Push an Pull)

Este diseño busca superar las limitaciones de las arquitecturas anteriores. Se realiza un primer proceso de extracción, no demasiado complejo, que permita eliminar la información menos importante o redundante. Una vez realizada y organizado el nuevo sistema aplica un nuevo proceso ETL más complejo y que dará resultado al conjunto de datos que finalmente se organizará dentro del Data Warehouse.

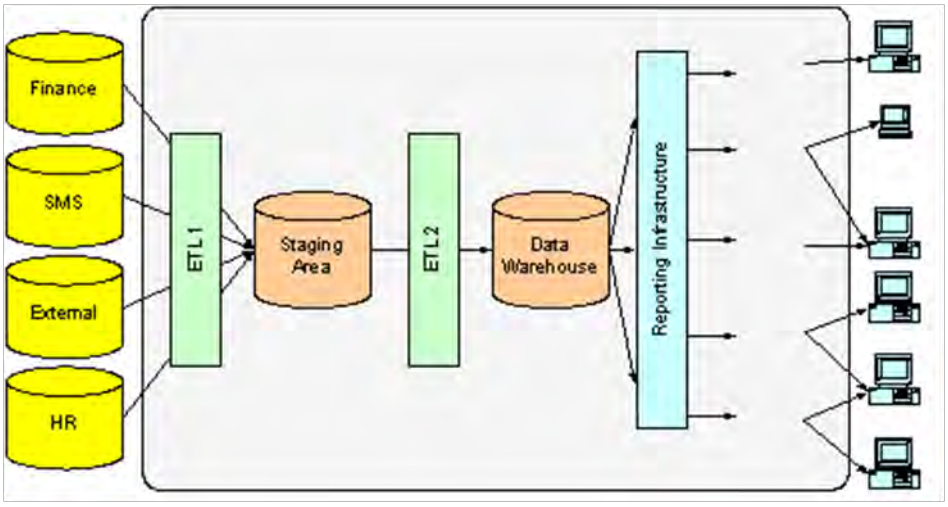


Figura 25: Esquema de un modelo de Diseño Central Híbrido

3.2.5 Diseño Federado

Es una arquitectura que intenta aprovechar los DataMarts ya existentes. Realizando un primer procesado ELT que se almacenará dentro de los DataMarts y a continuación se aplicará el proceso ETL que finalmente organizará la información dentro del DataWarehouse centralizado.

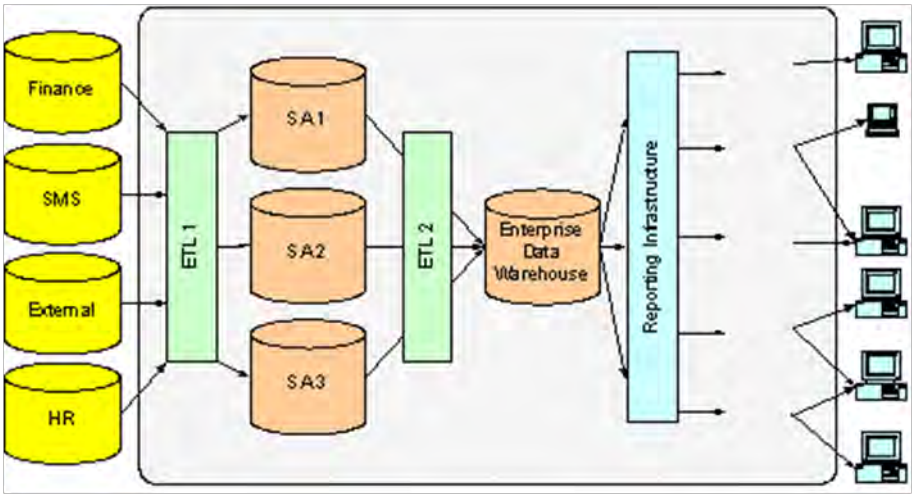


Figura 26: Esquema de un modelo de Diseño Federado

Conclusiones de la sección

Aunque parece que esta información se debía integrar junto a la definición de un Data Warehouse se ha mantenido fuera porqué representa procesos de limpieza de los datos una vez ya definida la estructuración de estos en el sistema.
Los procesos ETL son parte fundamental en el proceso de desarrollo de un sistema de información como el que se quiere desarrollar.

3.3 Minería de Datos

La Minería de Datos, en inglés el Data Mining es el proceso de análisis sobre las bases de datos con el fin de encontrar relaciones inesperadas que serán de valor para los propietarios de la base de datos. (Hand, 1998)

3.3.1 Origen

Pero tal y como introducíamos anteriormente mediante la cita de (Mena, 1999) el cual describe la minería de datos como: proceso iterativo de extracción de patrones escondidos en grandes bases de datos usando técnicas estadísticas y de Inteligencia Artificial, más exactamente Machine Learning.

Esta breve explicación es importante dado que en general existe una ambigüedad que provoca la difusión de conceptos generalizados y en algunos casos erróneos. Pero la realidad es que ambas disciplinas han evolucionado por separado dando lugar a nomenclaturas distintas con el mismo propósito, el de extraer información de los datos y expresarla en las decisiones que se tomaran.

A continuación se analizan las distintas definiciones de ambos conceptos:

El Machine Learning tal y como nos indica (Mena, 1999) es una rama de la Inteligencia Artificial que se encarga del diseño y aplicabilidad de algoritmos de aprendizaje (learning algorithms).

La estadística es una rama de las matemáticas aplicadas que considera tres disciplinas: el estudio de poblaciones, la variabilidad (para la modelización de fenómenos) y métodos de reducción de la información contenida en los datos.

Si nos fijamos en ambas definiciones se puede concluir que la estadística es la disciplina más cercana a la minería de datos pero la gran diferencia parte de la observación de la información. Las técnicas estadísticas requieren del usuario que proponga un modelo y sean estas las que lo parametricen. Sin embargo, tal y como indica (Aluja, 2001) la Inteligencia Artificial se ha centrado en el desarrollo de soluciones algorítmicas.

3.3.2 Objetivos

Como indica Bernstein et al (Bernstein, Provost, & Hill, 2005) los especialistas en Minería de Datos no están necesariamente familiarizados con todo el abanico de componentes y procesos posibles.

Mediante herramientas de Minería de Datos y datos históricos es posible tratar, en principio, cualquier tipo de problema. El principal problema es que el conocimiento parte implícitamente del mismo analista, sabiendo identificar cual es el patrón de información que se quiere extraer.

La visualización de los datos de una manera gráfica ayuda al usuario a examinar los volúmenes de datos.

A continuación se propone una clasificación de los tipos de objetivos o tipologías de patrones que se podrían considerar para el proyecto que nos ocupa. De este modo a posteriori ayudará a filtrar entre las herramientas a aplicar:

Asociaciones: la búsqueda de sucesos que se relacionen a través de una transacción. El objetivo de *las* reglas de asociación *es* encontrar asociaciones o correlaciones entre los elementos u objetos de la base de datos. Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente sea relativamente alta.

Dependencias: la búsqueda representativa de hechos comunes entre valores que permitan establecer que uno o más atributos determinan el valor de un tercero. En estos casos el gran problema parte del hecho que suelen existir muchas dependencias que no resulten interesantes.

Clasificación: representación del comportamiento de un hecho determinado por el conjunto de valores que lo determinan. Se puede ver como el esclarecimietno de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases ya conocidas.

Segmentación: o clustering representa la detección de grupos o racimos diferenciados del resto.

Detección de ciclos temporales o tendencias: detección de patrones que sean repetitivos, predecir valores de una variable continua a partir de la evolución de otra variable continua o de la misma en un pasado.

Predicción: representa uno de los problemas a abordar más claros. Por ejemplo, ante la pregunta ¿Lloverá mañana? el proceso a aplicar variará en función de la variable que buscamos. Si la el resultado es una variable continua (por ejemplo un conjunto de datos representados por un curva) se usaran técnicas de regresión, sin embargo si el resultado es un valor categórico (por ejemplo un si o no) entonces se usaran técnicas de clasificación.

3.3.3 Técnicas de Minería de Datos

Como se ha comentado anteriormente no existe una sola técnica para solucionar un mismo problema. La respuesta a cuál será el mejor método para resolver un problema se verá claramente influenciado por la experiencia del analista humano.

Aún así todos los algoritmos deben cumplir las siguientes características:

- Robusto: para evitar problemas de ruido, incertidumbre y descripciones superficiales.
- Fiables
- Escalables y eficientes: dado que su desarrollo implica dos costes, el de construir el modelo y el de clasificar los ejemplos.
- Explicativo
- Determinista: el algoritmo deberá adaptarse en caso que los datos cambien.

Hemos considerado necesario hacer un pequeño estudio de todos los posibles procesos a aplicar a fin de conocer las herramientas.

A continuación se van a presentar algunas de las técnicas más destacadas que se emplean en el análisis de la información. La clasificación se ha realizado en función de los datos de entrada y basándose en la organización descrita en el apartado anterior.

Por lo que se refiere a la información de salida, existen unas técnicas más visuales que otras. Este hecho es importante dado que las técnicas de visualización ayudan a comprender mejor y más rápidamente los resultados obtenidos. Cabe destacar la gran capacidad humana de extraer patrones a partir de imágenes. Este hecho resultará muy representativo dado que incluso una visualización previa de la estructuración de la información ayudará a entenderlos y así determinar qué tipo de técnica aplicar.

Las técnicas menos visuales resultaran ser las más apropiadas para variables continuas y con conocimiento más claro sobre el resultado a obtener. Pero sin embargo su poca inteligibilidad representa un problema a la hora de interpretar los resultados o entender porqué no se ha obtenido conocimiento.

A continuación se presenta una primera conceptualización con que se abordará con más detalle la sección práctica.

Se distinguirán dos clasificaciones: aprendizaje supervisado y no supervisado

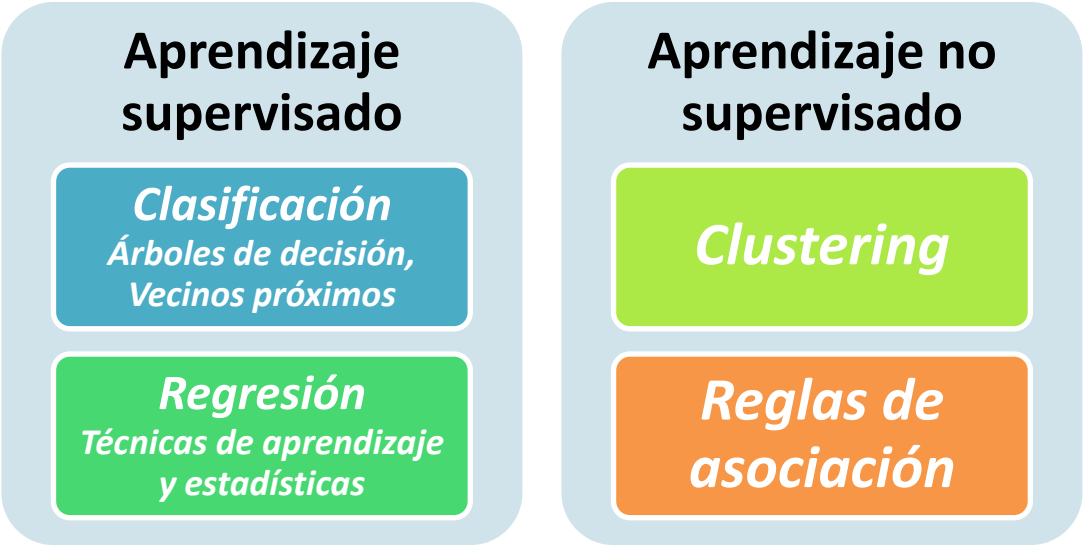


Figura 27: Clasificación de herramientas de Minería de Datos

1) Aprendizaje supervisado

Representa aquellos algoritmos en los que el algoritmo relaciona cada entrada con la salida deseada, es decir, la salida es el aprendizaje obtenido a partir de la información de entrada y los parámetros que determinan los objetivos a obtener. Este es el caso de los algoritmos de clasificación y regresión.

Clasificación

Los algoritmos de clasificación tratan de construir un modelo que les permita poder predecir una clase a partir de un conjunto de valores atributos. Por este motivo el algoritmo trata de dividir la información según uno de sus atributos sucesivamente (por valor o intervalos de valores).

Árboles de decisión: se aplican fundamentalmente para clasificación y segmentación. Permiten obtener de forma visual las reglas de decisión bajo las cuales se determinan algunos de los parámetros. A través de la construcción del árbol se podrán extraer reglas que a simple vista no son tan triviales o incluso cerciorarse de cómo otros parámetros no influyen en la solución final.

Vecinos más próximos y razonamiento por casos: se aplican para casos de clasificación y segmentación basándose en medidas de distancias o similitud de grupos.

Regresión

Los algoritmos de regresión representaran modelos de predicción.

Técnicas de aprendizaje: se aplican para casos de clasificación y segmentación mediante herramientas de redes neuronales, lógica difusa y algoritmos genéticos.

Técnicas estadísticas: se aplican para confirmar asociaciones, dependencias y análisis de segmentaciones. Son las técnicas de regresión lineal (y no lineal).

2) Aprendizaje no supervisado

Representa aquellos algoritmos capaces de extraer información relevante de la información sin la intervención de ningún factor, se auto-organizan. La única información que participa en el proceso es la de entrada en el sistema. Es el caso de los algoritmos de clustering y reglas de asociación.

Clustering

Permiten buscar elementos afines dentro de un conjunto de datos. Sus claves permiten:

- Definir la función de similitud de los datos.
- Agrupar los objetos similares en clusters.

Se definen diferentes tipos de clustering: particional, jerárquico o difuso.

Reglas de asociación

Permiten detectar las asociaciones comunes entre elementos identificando subconjuntos posibles de sus atributos, identificando las relaciones entre las transacciones y generando las reglas que los justifican.

3.3.4 Herramientas existentes

Todos los procesos descritos anteriormente son programables de manera manual, pero ya existen plataformas que los proporcionan, de este modo no es necesario reinventar la rueda.

Las dos herramientas más populares son Weka y RapidMiner. Ambas con OpenSource y desarrolladas en Java.

1) Weka

Es un entorno creado por desarrolladores de la Universidad de Waikato en Nueva Zelanda para el desarrollo de herramientas de aprendizaje automático y minería de datos.

Este proporciona una interfaz de usuario que ofrece el acceso a las herramientas de visualización, a los algoritmos para el análisis de los datos y a los de modelado predictivo.

Contiene una extensa colección de técnicas para el pre procesamiento de datos y modelado, tales como clustering, clasificación, regresión.

La alimentación de la información se hace mayoritariamente a través de archivos planos en formato propio ARFF, también permite la conexión con bases de datos SQL.

2) RapidMiner

Es una completa solución OpenSource que proporciona un entorno para el desarrollo y implementación de herramientas de minería de datos. Esta herramienta posee una buena interfaz gráfica de usuario y simplifica en gran medida algunas de las tareas más complejas.

Ésta plataforma parte de la también conocida YALE (Yet Another Learning Environment) y que en su día también resulto de gran importancia.

RapidMiner es un entorno multiplataforma desarrollado en Java y cuenta con una potente interfaz gráfica la cual permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de bloques. Cada bloque representa cada uno de los operadores, el encadenamiento de estos representa la construcción del proceso de extracción de conocimiento desde la lectura de estos hasta su análisis final. Dispone de más de 400 operadores de minería de datos.

Son muy numerosos los operadores que proporciona y las herramientas que orientan en el proceso de trabajo. Además, a través de la instalación de una extensión, permite la integración de otras plataformas de gran importancia, como es la integración de los algoritmos en Weka o la integración de un módulo para R. R es la plataforma usada por los expertos estadísticos.

Conclusiones de la sección

Para el desarrollo de la parte práctica se escogerá la herramienta RapidMiner ya que proporciona beneficios a nivel de la carga de datos y visualización de estos e integra las soluciones de Weka.

En esta etapa se deberán realizar las fases de:

- **Selección, limpieza y transformación de los datos que se van a analizar. La selección representa fusión y criba de filas y atributos.**
- **Selección del método de minería de datos y aplicación. Este proceso incluye la selección del algoritmo a aplicar y transformar los datos al formato requerido por el algoritmo.**

4. Comparación práctica de técnicas

4.1 Leako

Leako es una empresa sita en el País Vasco que se dedica a la venta de sistemas energéticos centralizados. Son sistemas que se implantan a nivel de edificio y que, entre otros servicios, ofrecen la gestión del control de este a través de una interfaz web. (Leako, 2011)

Esta empresa ha proporcionado una base de datos con los valores de los consumos recogidos por el sistema de monitorización implantado en algunos edificios de viviendas y de oficinas.

Cada edificio dispone de una estación central la cual se encarga del suministro de energía térmica para la calefacción y el agua caliente. Cada apartamento tiene una subcentral a través de la cual se proporciona el servicio a cada cliente.

La base de datos de Leako proporciona la información sobre los consumos realizados por más de 700 viviendas del País Vasco sobre:

- consumo de agua (litros),
- consumos térmicos para el agua caliente
- consumos térmicos de la calefacción
- temperatura interior

Todos los edificios se encuentran en alguna de las cuatro ciudades especificadas en el mapa de la Figura 28. Estas ciudades son: Bilbao, Donosti, Llodio o Amorebieta.

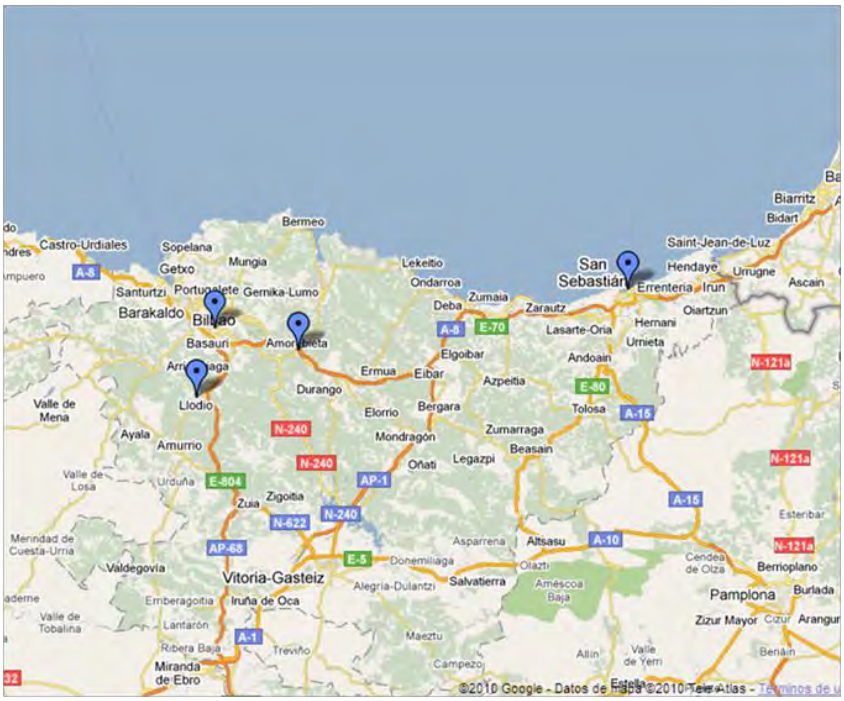


Figura 28: Distribución de los edificios monitorizados por Leako

Información proporcionada por Leako

Como se ha podido aprender en las fases anteriores de este documento, en el desarrollo de un proceso de KDD, el primer proceso a desarrollar es la de integración de la información. Esta ha sido una de las tareas más laboriosas. Leako facilita la información a través de un conjunto de archivos del tipo *.xls (Microsoft Office Excel). Los datos almacenados en ellos son distintos en función del archivo al cual se accede. Así hay archivos que contienen la información para una fecha concreta y otros con información según año, día y hora.

La primera tarea consistió en analizar todos los tipos de archivos y entender los datos y variables que estos contenían.

A continuación se detalla el contenido de los archivos facilitados por Leako:

En primer lugar encontramos la información referente al edificio (o como lo denomina Leako Central).

Información del edificio:	
Ncentral:26	Número de referencia del edificio
Maestro Elorza	Dirección de localización del edificio
Llodio (Plz Estacion)	Pueblo donde se ubica el edificio
c\ Maestro Elorza 1 - Llodio	Dirección completa sobre la localización del edificio

Figura 29: Información sobre el edificio proporcionada por Leako

A continuación se puede acceder a la información de los consumos de dos maneras:

1. A partir de una fecha determinada todas las lecturas realizadas a todos los apartamentos (o como lo denomina Leako Subcentral) para un edificio determinado tal y como se ve en la Figura 30.

Información por fecha de lectura		
NºSub	1	Identificador del edificio
Portal	1	Identificador del portal
Piso	7	Identificador del piso
Mano		Identificador de la puerta
CodigoReferencia		Código de referencia
NumCuenta		Número de cuenta
Detalle de los valores térmicos		
KwhTermTotal	1174,5	KWh consumidos en el apartamento desde que se puso el contador.
Fecha	04/09/2007	Fecha de la lectura
KwhTermParcial	87,6	Kwh consumidos desde la última lectura
(€)Base	39,06	Precio a pagar por los KWh consumidos

(€)Iva	0	Impuestos
Detalle de los valores sobre el agua		
LitrosTotal	54113	Litros de agua consumidos desde que se puso el contador
Fecha	04/09/2007	Fecha de la lectura
LitrosParcial	7548	Litros consumidos desde la última lectura
(€)Base	5,84	Precio a pagar por los litros consumidos
(€)Iva	0	Impuestos
(€)TotalFacturar	44,9	Valor en euros del total a pagar en la facture para ambos consumos.
NºRecibo	0	Identificador de la factura

Figura 30: Información de los consumos realizados para una fecha determinada proporcionada por Leako

2. A partir del identificador de un apartamento se puede acceder a todas las lecturas realizadas a este. En este segundo caso la información es muy detallada dado que proporciona los valores de consumos por hora como se ve en la Figura 31.

Información por Apartamento o Subcentral		**La información se recoge hora por hora**
Anio	2001	Año
Mes	{1…12}	Mes
Dia	{1..31}	Día
Hora	{0..24}	Hora
KwhCalefaccion	0	Kwh consumidos para la calefacción
KwhACS	0	Kwh conusmidos para calendar el agua
LitrosAgua	0	Litros de agua consumidos
Temperatura	166	Temperatura

Figura 31: Información de todos los consumos realizados por un apartamento hora por hora proporcionado por Leako

El principal problema que se planteó al tener la información a partir de las lecturas realizadas es que existe una gran diversidad de fechas provocando dos tipos de problemas:

1. En ciertos casos para una misma central las fechas para los consumos térmicos y los de agua no coinciden.
2. Hay centrales con más mediciones que otras o con fechas de lecturas muy distantes, si se tiene en cuenta que las lecturas parciales miden los consumos de un mes para el otro.

Estos problemas dificultan el desarrollo del sistema de información porque la inconsistencia de los datos puede corromper la totalidad de la información. Por este motivo fue necesario analizar en profundidad las fechas de las distintas lecturas para así llegar a un buen diseño de los Data Marts que hiciese posible la realización de las pruebas prácticas.

4.2 Pruebas realizadas

A continuación se presentan las pruebas realizadas con algunos de los procesos descritos anteriormente mediante la información proporcionada por Leako y utilizando la herramienta RapidMiner.

Para llevar a cabo el proceso de extracción de conocimiento se van a seguir las tres etapas especificadas y descritas anteriormente en la introducción de la sección 3.

4.2.1 Pre procesado de los datos

Hasta el momento conocemos la información y los valores proporcionados por Leako, pero para el desarrollo de un sistema más generalizado se deben tener en cuenta todos los valores que podrían ser necesarios desde el punto de vista de un arquitecto.

Por este motivo mediante la colaboración de un arquitecto especializado en proyectos de eficiencia energética se evaluó perfiles de usuario y se creó una taxonomía que incluye y describe parámetros necesarios.

taxonomy		data
project data		owner design team building type building life cycle phase ...
environment	*climate*	external air temperatura (time profile) external air relative humidity (time profile) horizontal solar radiation (time profile) ...
	surroundings	height on the see a level distance from the urban density ...
building	*building construction*	usefull floor area ... envelope transmittance ...
	building systems	HVAC type ...
operation		internal heat gains internal set point temperature ...
performance	*energy performance*	heating demand cooling demand hot water demand ...
	indoor space performance	internal air temperature (time profile) internal air relative humidity (time profile) ...
	cost	enegry cost by gas ...

Figura 32: Taxonomía de información relevante para le eficiencia energética

Tal y como se muestra en la Figura 33 la taxonomía se divide en 5 secciones:

Project Data: incluye los parámetros de carácter arquitectónico.

Environment: incluye los valores que hacen referencia a parámetros que describen el exterior del edificio. Tales como la temperatura exterior, altura respecto al nivel del mar, densidad urbana, etc.

Building: incluye los datos propios de las características técnicas del edificio.

Operation: incluye todos aquellos parámetros que intervendrán en el control del edificio, tales como patrones ocupacionales y de consumo como temperaturas óptimas (considerando regímenes de invierno y verano). Estos valores normalmente se extraen de los estándares proporcionados por el Ministerio de cada país.

Performance: en esta se incluyen finalmente los valores de los consumos realizados por el edificio y/o vivienda, precios, etc.

Mediante este esquema se complementó la información de Leako con los nuevos parámetros que debían considerarse desde el punto de vista de un arquitecto.

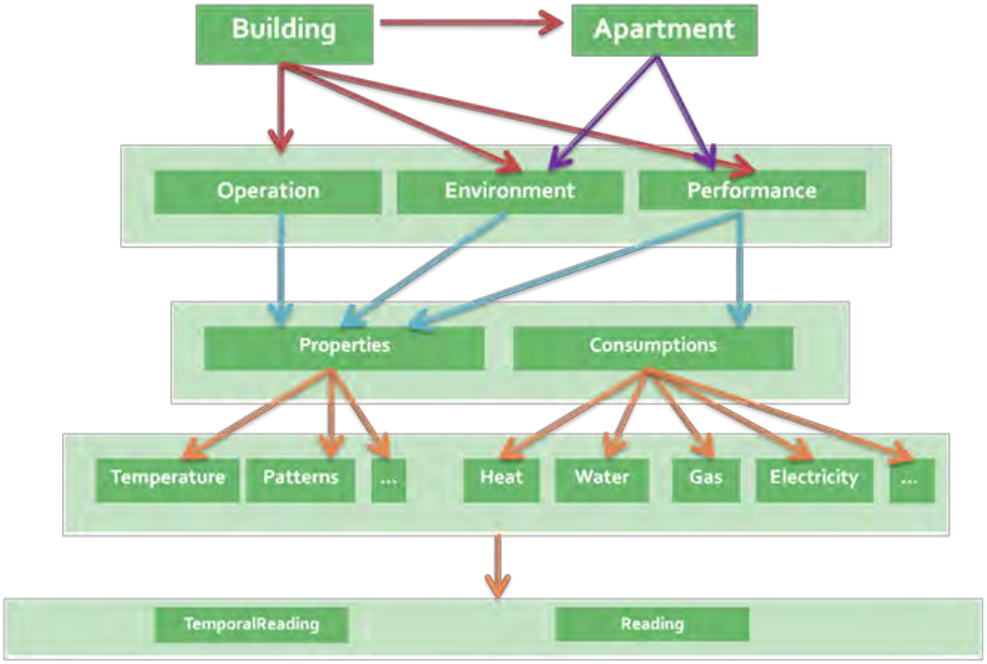


Figura 33: Diagrama de valores proporcionados por Leako

Esta nueva clasificación de la información se representa mediante un semidiagrama UML.(Figura 34)

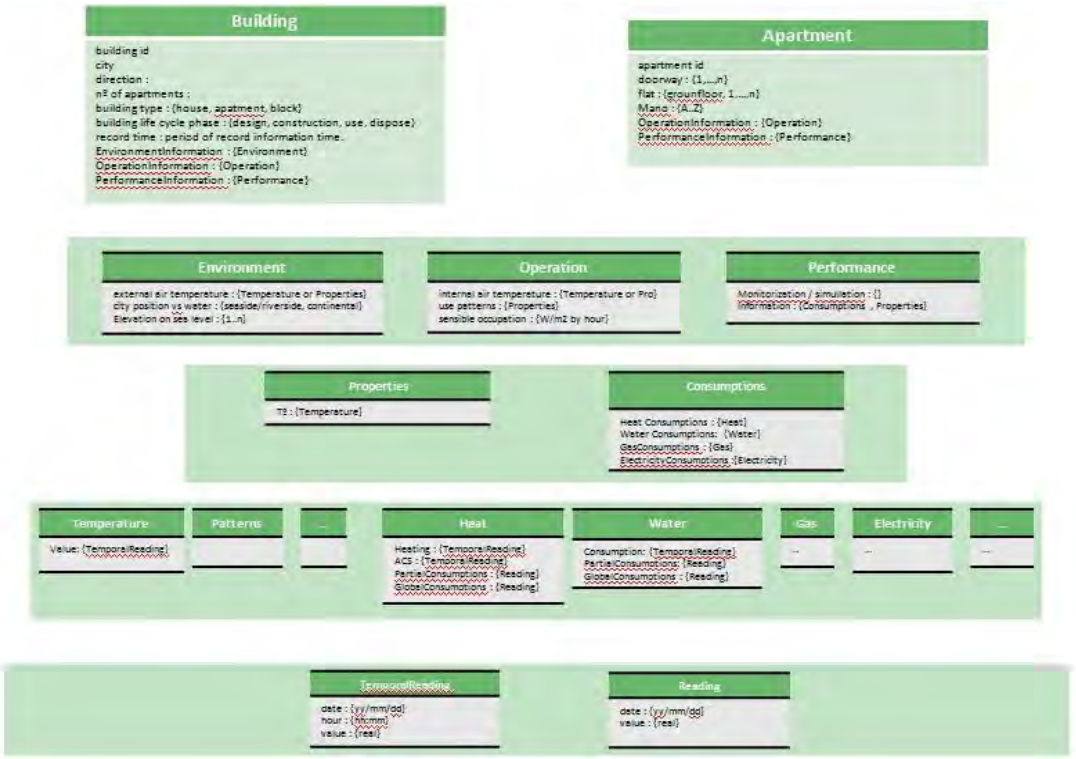


Figura 34: Primer diagrama UML cruzando datos Leako con concepto BIM

4.2.2 Aplicación de herramientas de Minería de Datos

En esta sección se realiza la demostración práctica de los algoritmos de minería de datos estudiados en la sección 3.3.3.

Para la demostración práctica se han usado los mismos datos con el fin averiguar el potencial de los distintos algoritmos y extraer las conclusiones que se derivan de su aplicación para cada caso.

1. Primer caso de aplicación

En este primer caso se uso se va a tratar de realizar un análisis de la información pero a un nivel muy visual y sin el uso de ninguna herramienta de minería de datos.

Tal y como se comentaba en la sección 3.3.3 de la memoria, la capacidad humana para reconocer la información a través de imágenes es una herramienta muy útil.

Para iniciar el ejercicio se crea un Data Mart en el cual almacena:

- consumos térmicos parciales
- consumos térmicos totales
- realizados por la Central con identificador 40
- para 3 fechas: 24/2/2006, el 24/3/2006 y el 23/4/2006

Como visualizamos en la configuración de la información contenida por el Data Mart.

ExampleSet (147 examples, 0 special attributes, 6 regular attributes)

Role	Name	Type	Statistics	Range	Missings
regular	Ncentral	nominal	mode = 40 (147), least = 40 (147)	40 (147)	0
regular	NºSub	nominal	mode = 1 (3), least = 1 (3)	1 (3), 2 (3), 3 (3), 4 (	0
regular	Piso	nominal	mode = 1 (27), least = 6 (12)	6 (12), 5 (27), 4 (27)	0
regular	KwhTermParcial	real	avg = 194.952 +/- 220.388	[0.000 ; 1476.300]	0
regular	KwhTermTotal	real	avg = 10240.751 +/- 8068.438	[0.000 ; 32922.100]	0
regular	Fecha	nominal	mode = 2/24/06 (49), least = 2/24/06 (49)	2/24/06 (49), 3/24/0	0

Figura 35: Configuración del Data Mart

La totalidad de los parámetros se configuran como valores nominales a excepción de los consumos que se defininen como valores reales.

A la hora de aplicar algoritmos de minería de datos la configuración de los parámetros puede diferir en función del algoritmo y de los resultados a obtener. Por este motivo se ha adoptado un enfoque generalista con el fin de evitar tener más de un Data Mart con la misma información para las distintas configuraciones de los parámetros. Posterioriormente mediante los módulos proporcionados por RapidMiner se podrán modificar las configuraciones.

A continuación se muestran algunos ejemplos:

ExampleSet (252 examples, 0 special attributes, 6 regular attributes)

Row No.	Ncentral	NºSub	Piso	KwhTermParcial	KwhTermTotal	Fecha
40	40	40	2	123.100	1780	2/24/06
41	40	41	3	97.200	5996.900	2/24/06
42	40	42	3	127.900	3868.900	2/24/06
43	40	43	3	424.200	4023.200	2/24/06
44	40	44	4	391.500	11534.300	2/24/06
45	40	45	4	122.700	5621.300	2/24/06
46	40	46	4	82.600	12207.100	2/24/06
47	40	47	5	122.700	4651.400	2/24/06
48	40	48	5	0	4.700	2/24/06
49	40	49	5	2.200	121.300	2/24/06
50	40	1	6	176.200	15591.200	3/24/06
51	40	2	5	408.100	12855	3/24/06
52	40	3	5	167.900	18807.600	3/24/06
53	40	4	4	50.600	7790.700	3/24/06
54	40	5	4	62.100	6095.800	3/24/06
55	40	6	3	80.800	5025.100	3/24/06
56	40	7	3	71.600	10980	3/24/06
57	40	8	2	47.400	5225.900	3/24/06
58	40	9	2	90.600	7253.500	3/24/06

Figura 36: Visualización de información contenida en el Data Mart

Como se puede observar en la siguiente imagen (Figura 39) los consumos térmicos realizados por esta central en el mes de febrero fueron superiores a los consumos realizados en los meses de marzo o abril.

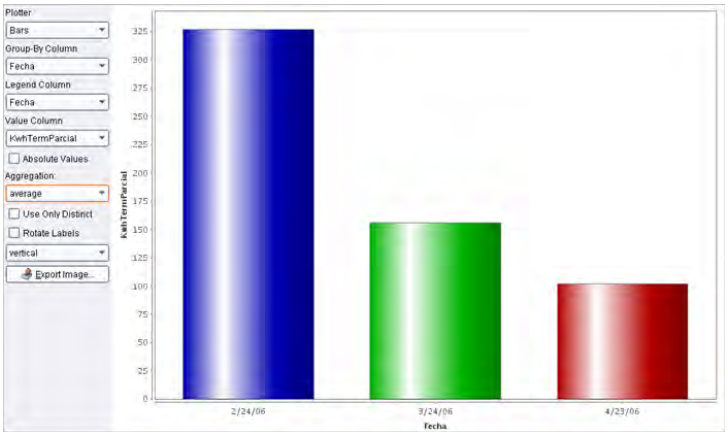


Figura 37: Gráfico de observación de los consumos para unas fechas determinadas.

RapidMiner proporciona diferentes herramientas que permiten la visualización de la información. Por ejemplo, en la Figura 27 se muestran los mismos consumos pero cada círculo representa una de las lecturas.

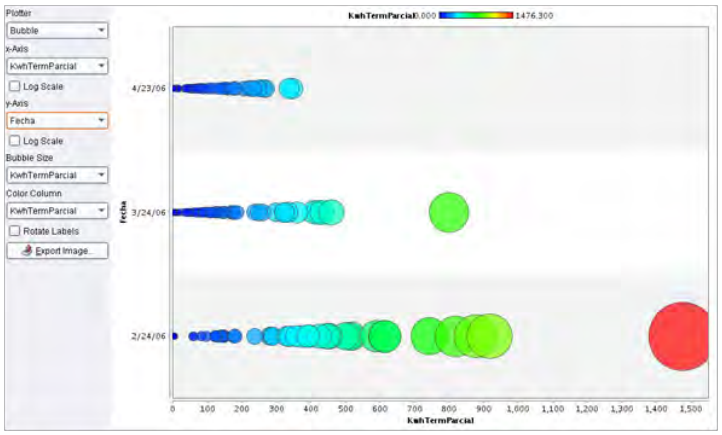


Figura 38: Visualización de los consumos para fechas determinadas.

Conclusiones del caso de uso:

Mediante el primer caso se ha podido hacer la primera toma de contacto con los datos y cómo se puede trabajar con ellos. La plataforma RapidMiner es de gran ayuda para cada uno de los procesos pero uno de los trabajos más importantes realizados es el de adaptación y aprendizaje de las diferentes herramientas que esta proporciona y cómo usarlas, dado que para los siguientes casos los parámetros de configuración se van complicando.

2. Segundo caso de aplicación

En el caso anterior se ha hecho uso de herramientas de visualización de los datos contenidos en el Data Mart, sin embargo, en esta segunda sección se usaran herramientas visuales siguiendo el procedimiento estadístico.

Como ya introducíamos en la sección 3.3.1 de la teoría, la estadística trata de construir un modelo a partir de la observación de los datos. Para ello hace uso de herramientas como Series o la Transformada de Fourier que le permiten ver de una manera gráfica como los datos se ven influenciados por la aplicación de estos algoritmos.

A continuación se presenta el potencial que ofrecen herramientas estadísticas como las Series para evaluar posibles resultados de los datos a tratar. Para ello se usará el módulo Moving Average, el cual crea un nuevo parámetro resultado del cálculo del promedio de los valores de la variable que se determine.

Los datos que se analizarán serán los consumos térmicos parciales.

Como se muestra en la Figura 41, en primer lugar construimos el modelo mediante los módulos y el escritorio de trabajo proporcionado por RapidMiner.

Figura 39: Modelo de análisis

A continuación se podrán representar los valores reales de los consumos realizados (línea azul) y los valores calculados mediante el módulo que ha calculado un promedio de todos estos (línea roja).

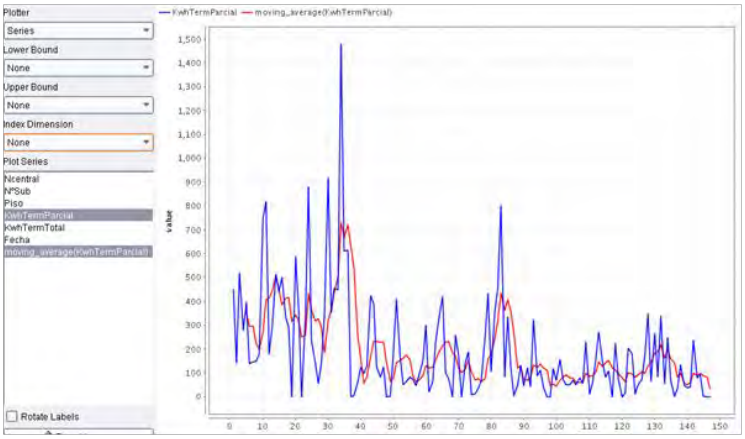


Figura 40: Análisis mediante Series

Conclusiones del caso de uso:
Mediante el segundo caso se han explorado los beneficios que la estadística proporciona mediante herramientas de visualización que ayudaran, a posteriori, a los expertos en el desarrollo de hipótesis específicas.

3. Tercer caso de aplicación

Tras las primeras tomas de contacto con la aplicación, se aplican algunos de los procesos de minería de datos. Dado que se ha iniciado los casos de estudio partiendo de ejemplos basados en las herramientas de visualización se continuará mediante la clasificación de la información mediante un árbol de decisión. Recordemos que los árboles de decisión son herramientas del tipo supervisado.

El algoritmo de construcción de un árbol de decisión implementa el siguiente bucle:

1. Verificar el criterio de parada del proceso en el nodo.
2. Definir la lista de todas las particiones posibles en el nodo.
3. Seleccionar la partición óptima
4. Generar la partición seleccionada.

Exiten varios algoritmos para la construcción de árboles de decisión. La herramienta RapidMiner ofrece los suyos propios y los desarrollados por Weka, como es el caso del W-J48. Solo de Weka te ofrecen 15 algoritmos de árboles de decisión distintos.En este caso se ha usado el proporcionado por RapidMiner dado que el resultado era el mismo que el de Weka.

Para este ejemplo se han desarrollado todas las fases del proceso de KDD:

1 Pre procesado

Creación del Data Mart: se ha generado un nuevo Data Mart el cual contiene:

- consumos térmicos parciales
- consumos térmicos totales
- realizados por dos Centrales: la 40 y la 26
- las mediciones de la central 40 son de las fechas: 24/2/2006, el 24/3/2006 y el 23/4/2006
- las mediciones de la central 26 son de las fechas: 2/1/2007, el 2/5/2007 y el 4/9/2007

Detalle del Data Mart

ExampleSet (252 examples, 0 special attributes, 7 regular attributes)

Role	Name	Type	Statistics	Range	Missings
regular	Ncentral	nominal	mode = 40 (147), least = 26 (105)	40 (147), 26 (105)	0
regular	N°Sub	nominal	mode = 1 (6), least = 36 (3)	1 (6), 2 (6), 3 (6), 4 (6), 5 (6), 6 (6),	0
regular	Piso	nominal	mode = 5 (42), least = 7 (15)	6 (27), 5 (42), 4 (42), 3 (42), 2 (42)	0
regular	KwhTermParcial	real	avg = 360.537 +/- 487.532	[0.000 ; 3120.600]	0
regular	KwhTermTotal	real	avg = 14749.461 +/- 12367.944	[0.000 ; 57644.700]	0
regular	Fecha	nominal	mode = 2/24/06 (49), least = 5/2/07 (35)	2/24/06 (49), 3/24/06 (49), 4/23/06	0
regular	TMedia	nominal	mode = 16 (49), least = 21 (35)	16 (49), 18 (49), 20 (49), 21 (35),	0

Figura 41: Data Mart Centrales 26 y 40

2 Aplicación de herramientas

Elaboración del proceso mediante los módulos en el escritorio de RapidMiner.

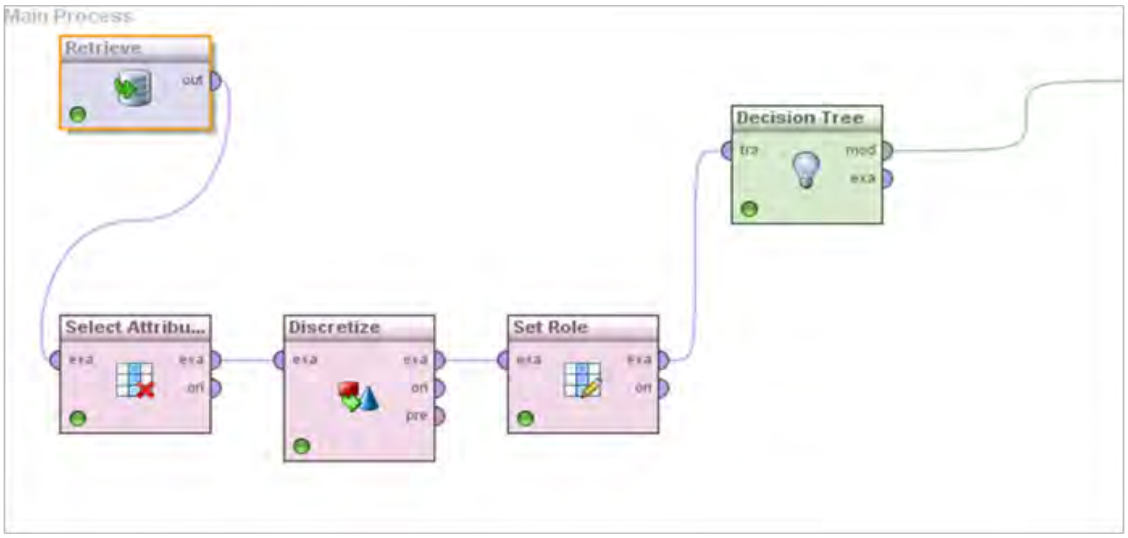


Figura 42: Diagrama de módulos para el análisis mediante árboles de decisión.

A continuación se explica la funcionalidad de cada uno de los módulos:

El primer módulo representa el Data Mart ya especificado.

El segundo módulo lleva a cabo la función de selección de atributos o filtraje de parámetros con el fin de no incluir en el análisis aquellos parámetros innecesarios y que pueden llevar a confusión. En este caso se han excluido en el análisis estos valores:

- el número de la subcentral: es un identificador que puede llevar a la confusión al haber incluido los consumos de dos centrales dado que habrá valores repetidos y que el algoritmo puede asociar.
- los consumos térmicos totales: a pesar de parecer que representen información muy importante, los valores realmente importantes serán los consumos parciales.

EL tercer módulo discretiza los valores de los consumos parciales. El valor de los consumos se ha definido como un valor real. Los árboles de decisión solo trabajan con valores nominales. Mediante el módulo de discretización se agrupan los valores numéricos en un número determinado de grupos (parámetros especificado por el usuario) y se convierten los rangos en valores nominales. Se puede discretizar por frecuencia, por tamaño, etc. Este módulo es muy importante para el desarrollo del proceso. Se ha usado un discretizador por frecuencia para 3 rangos de valores. Esta discretización se realiza mediante intervalos de igual frecuencia, es decir, se seleccionan los umbrales de los intervalos de tal manera que cada uno contenga la misma cantidad de valores numéricos.

En cuarto lugar mediante el módulo SetRole se define la variable a evaluar indicando mediante el atributo label que es un atributo especial, en este caso el consumo parcial.

Finalmente se le pasa la información al algoritmo que desarrollará el árbol de decisión.

3 Post proceso y toma de decisiones

El resultado obtenido ha sido el siguiente:

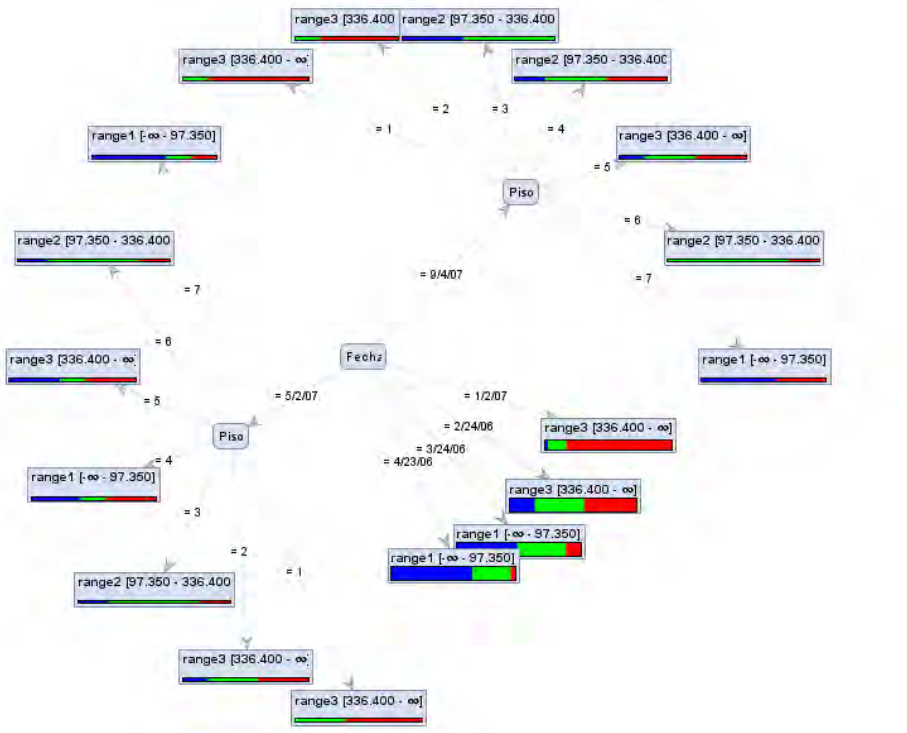


Figura 43: Árbol de decisión resultante

Conclusiones del caso de uso:

Como se visualiza en el diagrama de la Figura 45, los valores de los consumos recogidos por la estación 26 entre el 2/5/2007 y el 4/9/2007 han estado claramente determinados por la altura a la que se encuentra el piso. Sin embargo los consumos para las otras fechas se mantienen en el mismo rango de valores. Cabe destacar el hecho de que el sistema haya obviado la variable que indicaba el valor de la temperatura media exterior en esas fechas.

4. Cuarto caso de aplicación

Los algoritmos de predicción son una herramienta potente que permiten realizar otro tipo de análisis por ejemplo, completar un proceso con la estimación de datos que no se poseen.

El problema planteado surge al tener un Data Mart con unos consumos térmicos cuya fecha de adquisición es desconocida. Para poder hacer una estimación de la fecha en la que se realizaron los consumos se lleva a cabo una comparación con otro Data Mart que contiene los consumos de otra central para tres fechas.

Este mismo proceso de predicción se podría aplicar en diferentes direcciones, es decir, relacionando los consumos para el mismo año o para diferentes años para una misma central o entre diferentes centrales. En este caso comparamos entre dos centrales para distintos años.

Inicio del proceso de KDD

1 Pre procesado

Los Data Marts proporcionados son los ya anteriormente descritos los cuales contienen:

Data Mart ENTRENAMIENTO	Data Mart LEARNER
realizados por la Central 26 consumos térmicos parciales consumos térmicos totales para las fechas: 2/1/2007, el 2/5/2007 y el 4/9/2007. Si nos fijamos las tres fechas que se han seleccionado distan bastante en tiempo para que el proceso pueda escoger claramente entre una de ellas.	realizados por la Central 40 consumos térmicos parciales consumos térmicos totales

Si antes de iniciar el proceso se reflexiona sobre el resultado esperado se determinaría que dado que el entrenamiento posee consumos de los meses de febrero, mayo y septiembre y los que queremos clasificar son de febrero, marzo y abril, lo más acorde sería que el sistema estimara que los consumos son del mes de febrero.

2 Aplicación de herramientas

La Figura 33 describe el proceso completo:

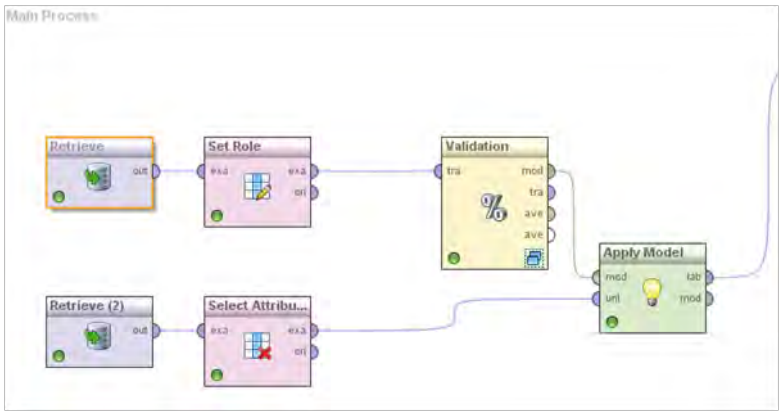


Figura 44: Proceso de predicción de conocimiento

A continuación se explica la funcionalidad de cada uno de los módulos:

El primer módulo representa el Data Mart de entrenamiento, el que contiene los datos de la Central 26.

El segundo módulo mediante el Set Role indica que el valor especial será la fecha.

El tercer bloque es el que realiza el entrenamiento. Mediante el módulo de Validación del tipo "Split Validation" divide los ejemplos en dos conjuntos, uno de prueba y otro de modelo y de este modo produce un Vector Performance que representará el entrenamiento. Pero hay que

indicarle la acción que se pretende desarrollar. Por ello internamente se debe configurar de la siguiente manera(Figura 47):

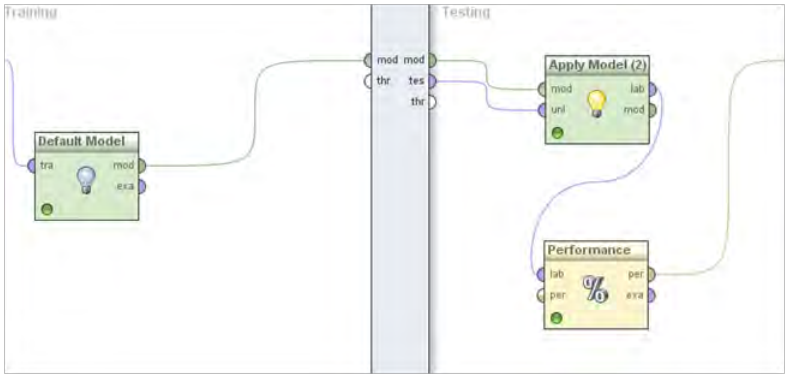


Figura 45: Proceso especificado dentro del módulo de Validación

El módulo Default model hace referencia a un módulo del tipo “Lazy Modeling”. Este módulo es el encargado de aprender y crear el modelo que permitirá predecir a través de un conjunto de valores.

A continuación se lleva a cabo una fase de testeo de la información mediante los siguientes módulos:

Módulo “Apply Model”: este modelo es el que se aplicará a un conjunto de datos de los proporcionados a la entrada.

Módulo “Performance”: es un operador evaluador que se usa para tareas de clasificación. Este operador espera un conjunto de ejemplos de prueba como entrada que contienen el atributo label y el rol prediction. Sobre estos dos atributos se calcula el Vector Performance.

Retomando el proceso principal a segundo nivel se encuentra el Data Mart con los valores de los consumos de la Central 40. Mediante el módulo “SelectAttributes” solo cogemos los valores del identificador de la central y sus consumos parciales para evitar que el sistema se distraiga.

Finalmente mediante el bloque “ApplyModel” se le proporciona el entrenamiento y los valores de la Central 40.

3 Post proceso y toma de decisiones

El resultado se muestra en la Figura 35:

ExampleSet (147 examples, 4 special attributes, 2 regular attributes)

Role	Name	Type	Statistics	Range	Missin...
confidence_5/2/07	confidence(5/2/07)	real	avg = 0.333 +/- ?	[0.333 ; 0.333]	0
confidence_1/2/07	confidence(1/2/07)	real	avg = 0.333 +/- ?	[0.333 ; 0.333]	0
confidence_9/4/07	confidence(9/4/07)	real	avg = 0.333 +/- ?	[0.333 ; 0.333]	0
prediction	prediction(Fecha)	nominal	mode = 5/2/07 (147), least = 1/2/07 (0)	5/2/07 (147), 1/2/07 (0), 9/4/07 (0)	0
regular	Ncentral	nominal	mode = 40 (147), least = 40 (147)	40 (147)	0
regular	KwhTermParcial	real	avg = 194.952 +/- 220.388	[0.000 ; 1476.300]	0

Figura 46: Resultado de la predicción

El proceso determina que los valores proporcionados por la Central 40 respecto al entrenamiento de la Central 27 son consumos realizados el mes de mayo.

Este resultado no coincide con la reflexión realizada al inicio del caso. Si se visualizan los datos se observa el motivo del resultado del sistema predictor.

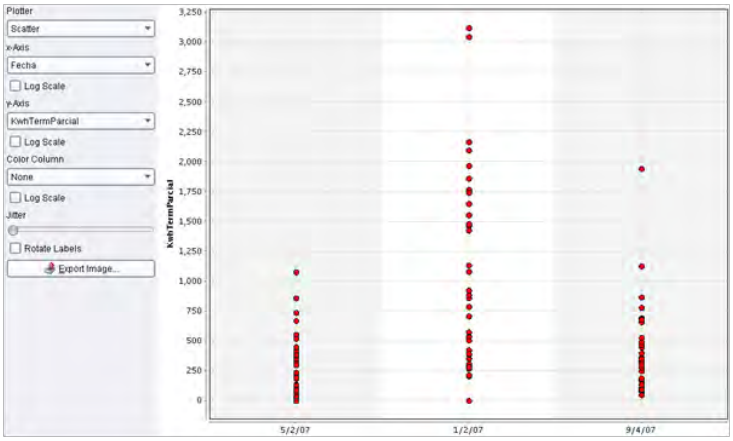


Figura 47: Consumos Central 26 año 2007

Los consumos realizados en la Central 26 en el año 2007 fueron en general bastante altos. Para el mes de febrero hay consumos entre 1000 y 2000 KWh y para los meses de mayo y septiembre aunque llegan a los 1000KWh (Figura 49).

Sin embargo si se observan los consumos realizados por la Central 40 para los tres meses consecutivos se observa que los consumos fueron más estables y en general no sobrepasaron los 500 KWh a excepción del mes de febrero. (Figura 50).

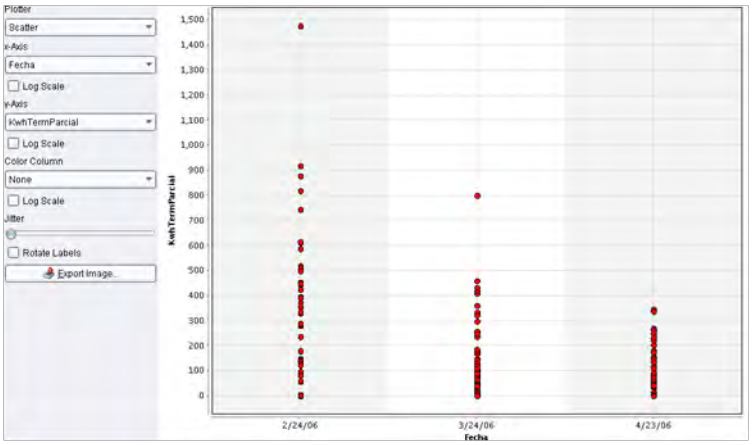


Figura 48: Consumos Central 40 año 2006

Se puedo concluir, por tanto, que el proceso ha determinado correctamente que los consumos se corresponden con los realizados en el mes de mayo. Cabe destacar que se trata del mismo tipo de viviendas en ambos casos.

Conclusiones del caso de uso:

Mediante el proceso de predicción se ha podido observar que el año 2007 en general fue más frio y provocó consumos más elevados respecto el 2006.

5. Quinto caso de aplicación

Los algoritmos de clustering son técnicas que dividen la información en función de la proximidad entre dos individuos y a partir de ahí, buscan los grupos de individuos más parecidos entre sí.

1 Pre procesado

Para el proceso de clustering se ha creado un nuevo Data Mart el cual contiene la información de las dos Centrales, la 26 y la 40.

2 Aplicación de herramientas

La Figura 51 muestra el proceso mediante los módulos en el escritorio de RapidMiner.

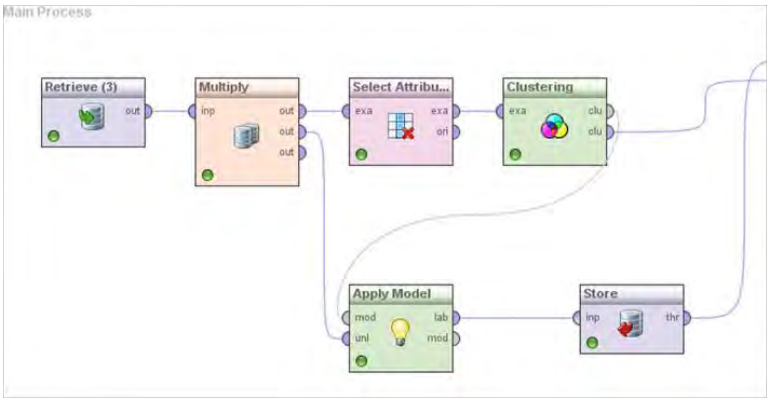


Figura 49: Proceso para la aplicación de herramientas de clustering

A continuación se explica la funcionalidad de cada uno de los módulos:

El primer módulo representa el Data Mart que contiene los datos para las Centrales 26 y 40.

El segundo módulo es una herramienta de RapidMiner que permite multiplicar la base de datos de entrada para los dos procesos en paralelo.

El tercer bloque es el de “SelectAttributes” mediante el que se seleccionaran los consumos parciales para el proceso de clusterización.

El cuarto bloque es el módulo de clusterización al cual se le configuran el número de clusters en los que se quieres segmentar la información. En este caso se ha configurado que la salida sea de 4 clusters.

Como se puede ver en las Figuras 52 y 53 estas son las salidas del proceso de clusterización.

ExampleSet (252 examples, 2 special attributes, 1 regular attribute)

Row No.	id	cluster	KwhTermP...
1	1	cluster_0	449.800
2	2	cluster_1	144.100
3	3	cluster_0	519
4	4	cluster_0	280.700
5	5	cluster_0	394.500
6	6	cluster_1	139.500
7	7	cluster_1	145.900
8	8	cluster_1	147.800
9	9	cluster_1	177.400
10	10	cluster_3	743.400
11	11	cluster_3	818.400
12	12	cluster_1	179.700
13	13	cluster_0	286
14	14	cluster_0	511.700
15	15	cluster_0	443.400
16	16	cluster_0	498.700
17	17	cluster_0	335.100
18	18	cluster_0	288.900
19	19	cluster_1	0
20	20	cluster_0	586.600
21	21	cluster_0	372.800
22	22	cluster_1	0
23	23	cluster_0	329.700
24	24	cluster_3	877.100
25	25	cluster_0	236.500
26	26	cluster_1	143.200
27	27	cluster_1	57.600

Figura 50: Resultado proceso de clustering

Con el análisis gráfico se identifica que ha configurado el cluster 1 con los consumos inferiores. Si se ordenan los clusters por niveles de consumos se obtiene:

Cluster 1 < Cluster 0 < Cluster 2 < Cluster 3

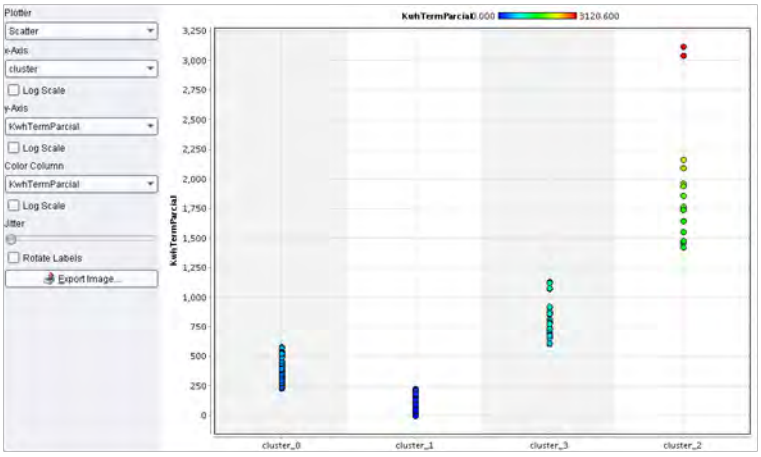


Figura 51: Resultados de la distribución entre los consumos y los clusters

Pero con esta clasificación de la información no se pueden visualizar como se relacionan el resto de los parámetros con la clasificación realizada por el clustering. Por este motivo se necesita el bloque final “ApplyModel”.

Tal y como se ve en la Figura 54 ahora se podrá ver la relación entre la totalidad de los datos de entrada y la clasificación realizada a través del proceso de clusterización.

Row No.	label	Ncentral	N°Sub	Piso	KwhTermParcial	KwhTermT...	Fecha
1	cluster_0	40	1	6	449.800	15415	2/24/06
2	cluster_1	40	2	5	144.100	12446.900	2/24/06
3	cluster_0	40	3	5	519	18639.700	2/24/06
4	cluster_0	40	4	4	280.700	7740.100	2/24/06
5	cluster_0	40	5	4	394.500	6033.700	2/24/06
6	cluster_1	40	6	3	139.500	4944.300	2/24/06
7	cluster_1	40	7	3	145.900	10908.400	2/24/06
8	cluster_1	40	8	2	147.800	5178.500	2/24/06
9	cluster_1	40	9	2	177.400	7162.900	2/24/06
10	cluster_3	40	10	1	743.400	23246	2/24/06
11	cluster_3	40	11	1	818.400	20327.700	2/24/06
12	cluster_1	40	12	1	179.700	1935.800	2/24/06
13	cluster_0	40	13	1	286	9619.500	2/24/06
14	cluster_0	40	14	2	511.700	19618.200	2/24/06
15	cluster_0	40	15	2	443.400	16449	2/24/06
16	cluster_0	40	16	3	498.700	12809.900	2/24/06
17	cluster_0	40	17	3	335.100	11157.700	2/24/06
18	cluster_0	40	18	4	288.900	10594	2/24/06
19	cluster_1	40	19	4	0	78.900	2/24/06
20	cluster_0	40	20	5	586.600	19087.600	2/24/06
21	cluster_0	40	21	5	372.800	9450.100	2/24/06
22	cluster_1	40	22	6	0	0	2/24/06
23	cluster_0	40	23	6	329.700	15248	2/24/06
24	cluster_3	40	24	6	877.100	32533.400	2/24/06
25	cluster_0	40	25	5	236.500	7806.400	2/24/06
26	cluster_1	40	26	5	143.200	1820.500	2/24/06
27	cluster_1	40	27	4	57.600	1179.300	2/24/06
28	cluster_1	40	28	4	139.400	4688.300	2/24/06

Figura 52: Resultado del proceso de clustering realimentado

Para finalizar se guardaran los resultados como un nuevo Data Mart mediante el módulo “Store”. De esta manera se realimenta el sistema con los conocimientos obtenidos.

3 Post proceso y toma de decisiones

Para entender los resultados obtenidos se usaran una vez más las herramientas de visualización. En el gráfico que muestra la Figura 55 se muestra la relación global entre las cuatro variables más importantes.

Se visualizan seis barras, las cuales representan los consumos parciales para las seis fechas incluidas en el DataMart. Las tres primeras son las de la central 26 y las restantes las de la 40. Mediante esta visualización permite ver la relación entre los clusters indicados por los colores y las fechas para los consumos de ambas centrales para la totalidad de los consumos parciales proporcionados..

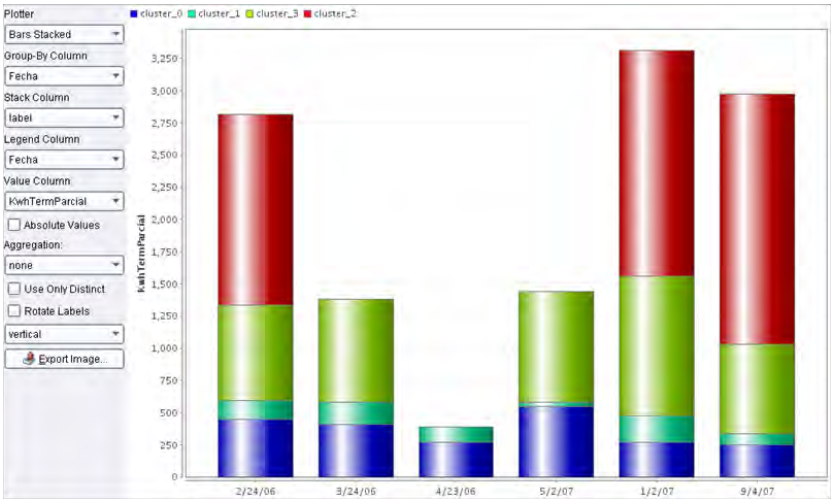


Figura 53: Visualización de los resultados en los que se implican los consumos, las fechas y los clusters

Conclusiones del caso de uso:

Las herramientas de clusterización son las más utilizadas de todo el conjunto de algoritmos que se han presentado, por la simplicidad a la hora de visualizar los resultados obtenidos y permiten encontrar la similitud dentro de un conjunto de datos.

6. Sexto caso de aplicación

Las reglas de asociación son el clásico ejemplo de la aplicabilidad de las herramientas de minería de datos sobre hechos de la vida cotidiana. El ejemplo más clásico de la aplicación de un algoritmo de reglas de asociación surgió analizando los carritos de la compra de los clientes de un gran almacén. Mediante este análisis se determinó que los clientes que compraban pañales también solían comprar cervezas. A posteriori también detectaron que las ventas se realizaban principalmente los viernes y sus clientes eran varones de entre 25 y 35 años. Tal y como se visualiza en la Figura 57 los algoritmos de reglas de asociación se simbolizan con un carrito de la compra.

1 Pre procesado

Para este caso de aplicación se ha decidido introducir un nuevo Data Mart. En el cual se ha intentado recoger los consumos de todas las subcentrales para un periodo de lectura muy cercano El nuevo Data Mart contiene:

- consumos térmicos parciales
- consumos térmicos totales
- euros pagados de los consumos parciales
- ciudad donde se encuentra la Central
- la relación Central, fecha de facturación y ciudad es la que se muestra en la Figura 56

NCetral	Fecha facturación	Ciudad
20	25/06/2007	Bilbao
25	28/06/2007	Donosti

26	2/7/2007	Llodio
31	22/08/2007	Amorebieta
33	3/05/2007	Amorebieta
37	29/06/2007	Bilbao
40	22/08/2007	Llodio
46	25/06/2007	Bilbao

Figura 54: Datos incluidos en el Data Mart

2 Aplicación de herramientas

La Figura 57 muestra el proceso mediante los módulos en el escritorio de RapidMiner

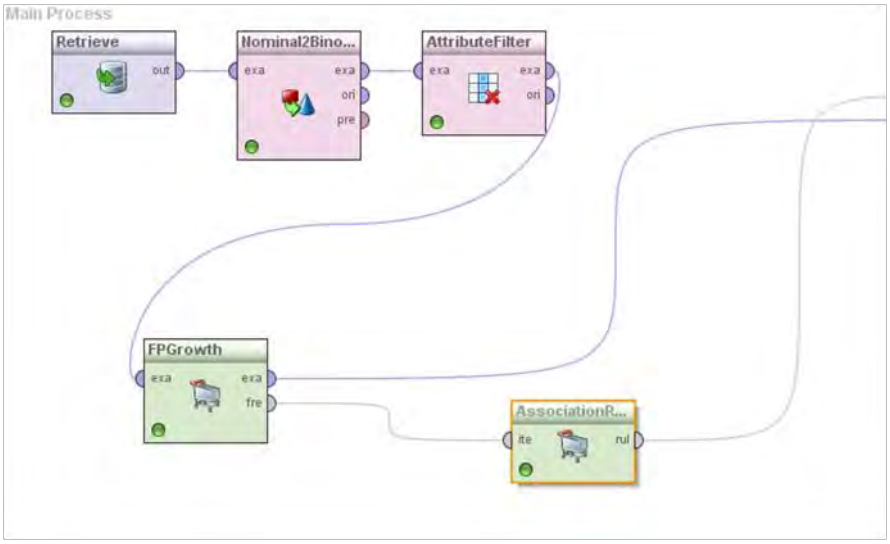


Figura 55: Proceso para la búsqueda de reglas de asociación

El primer módulo contiene el Data Mart.

El segundo módulo sirve para convertir los valores definidos como nominales en binomiales dado que será necesario para su tratamiento por el módulo de reglas de asociación.

El tercer módulo es el de selección de atributos el cual ha permitido filtrar aquellos parámetros que no se quiere que intervengan en este proceso.

El cuarto módulo es el "AttributeFilter", mediante este módulo se le indicaran los parámetros que se quiere que intervengan en las relaciones.

Y finalmente se encuentran el módulo "FPGrowth" el cual será el que aprenda y calcule todos las combinaciones más frecuetes y por último el módulo "Create Asociation Rules" el cual generará las reglas a partir del entrenamiento proporcionado por el módulo anterior.

El resultado final es el observado en la Figura 58 en la cual muestra la relación entre los atributos

Conjunction Type:

And

Conclusions:

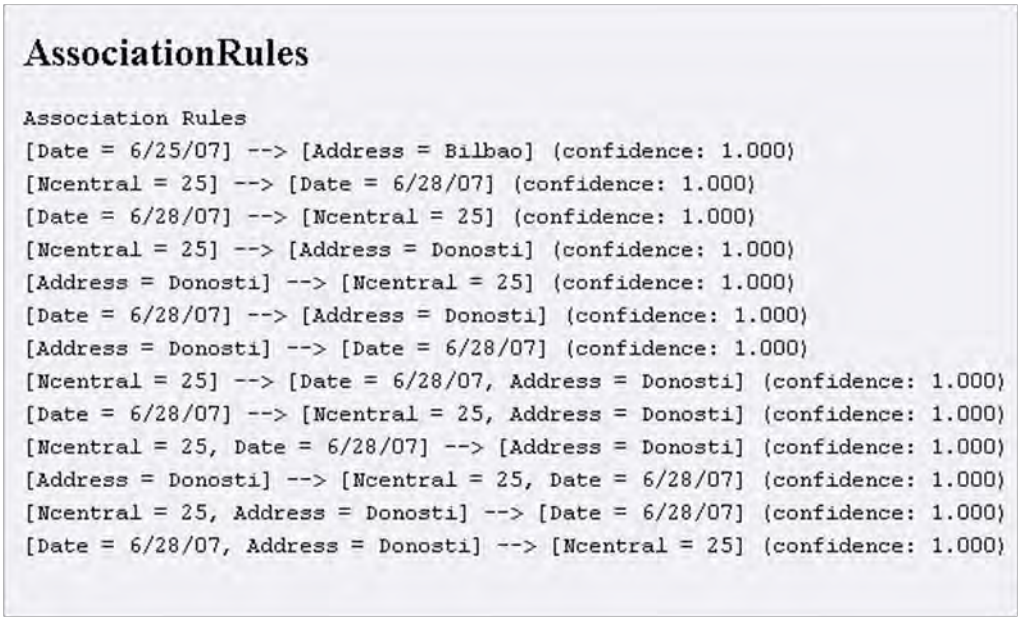
Address = Bilbao
Ncentral = 25
Date = 6/28/07
Address = Donosti

No.	Premises	Conclusion	Support	Confid.	LaPl.	Gain	p-s	Lift	Conv.
1	Date = 6/25/07	Address = Bilbao	0.230	1	1	-0.23(	0.139	2.523	∞
2	Ncentral = 25	Date = 6/28/07	0.305	1	1	-0.30(	0.212	3.281	∞
3	Date = 6/28/07	Ncentral = 25	0.305	1	1	-0.30(	0.212	3.281	∞
4	Ncentral = 25	Address = Donosti	0.305	1	1	-0.30(	0.212	3.281	∞
5	Address = Donosti	Ncentral = 25	0.305	1	1	-0.30(	0.212	3.281	∞
6	Date = 6/28/07	Address = Donosti	0.305	1	1	-0.30(	0.212	3.281	∞
7	Address = Donosti	Date = 6/28/07	0.305	1	1	-0.30(	0.212	3.281	∞
8	Ncentral = 25	Date = 6/28/07, Address = Donosti	0.305	1	1	-0.30(	0.212	3.281	∞
9	Date = 6/28/07	Ncentral = 25, Address = Donosti	0.305	1	1	-0.30(	0.212	3.281	∞
10	Ncentral = 25, Date = 6/28/07	Address = Donosti	0.305	1	1	-0.30(	0.212	3.281	∞
11	Address = Donosti	Ncentral = 25, Date = 6/28/07	0.305	1	1	-0.30(	0.212	3.281	∞
12	Ncentral = 25, Address = Donosti	Date = 6/28/07	0.305	1	1	-0.30(	0.212	3.281	∞
13	Date = 6/28/07, Address = Donosti	Ncentral = 25	0.305	1	1	-0.30(	0.212	3.281	∞

Figura 56: Resultado obtenido de las Reglas de asociación

3 Post proceso y toma de decisiones

Analizando las reglas de asociación proporcionadas por el proceso (Figura 59) se puede comprobar la veracidad de las reglas proporcionadas.



AssociationRules

```
Association Rules
[Date = 6/25/07] --> [Address = Bilbao] (confidence: 1.000)
[Ncentral = 25] --> [Date = 6/28/07] (confidence: 1.000)
[Date = 6/28/07] --> [Ncentral = 25] (confidence: 1.000)
[Ncentral = 25] --> [Address = Donosti] (confidence: 1.000)
[Address = Donosti] --> [Ncentral = 25] (confidence: 1.000)
[Date = 6/28/07] --> [Address = Donosti] (confidence: 1.000)
[Address = Donosti] --> [Date = 6/28/07] (confidence: 1.000)
[Ncentral = 25] --> [Date = 6/28/07, Address = Donosti] (confidence: 1.000)
[Date = 6/28/07] --> [Ncentral = 25, Address = Donosti] (confidence: 1.000)
[Ncentral = 25, Date = 6/28/07] --> [Address = Donosti] (confidence: 1.000)
[Address = Donosti] --> [Ncentral = 25, Date = 6/28/07] (confidence: 1.000)
[Ncentral = 25, Address = Donosti] --> [Date = 6/28/07] (confidence: 1.000)
[Date = 6/28/07, Address = Donosti] --> [Ncentral = 25] (confidence: 1.000)
```

Figura 57: Reglas de asociación extraídas

Inicialmente el proceso a encontrado que la fecha de facturación 25/06/07 es relativa a una de las centrales sita en Bilbao, pero si buscamos más información sobre dicha central el proceso ya no muestra más información, sin embargo, si que detecta que para la central con identificador 25 le consta una fecha de facturación del 28/6/07 y que finalmente pertenece a una central sita en Donosti.

Ambas afirmaciones son correctas, pero por qué el proceso no concluye nada más del resto de centrales? La respuesta es que Donosti es la única ciudad que solo posee una central. En el resto de casos existe más de una central para cada una de las ciudades. Es entonces cuando el proceso detecta que para una ciudad tiene más de un identificador de Central asociado y más de una fecha de facturación, por eso de centra en la regla que si podrá asegurar.

Conclusiones del caso de uso:
Una vez más los algoritmos de reglas de asociación han demostrado su eficacia en la búsqueda de patrones de comportamiento en los datos de entrada.

7. Séptimo caso de aplicación

Para finalizar es necesario demostrar como mediante RapidMiner también se pueden utilizar los algoritmos de la plataforma Weka.

1 Pre procesado

Para este ejercicio se ha escogido hacer uso de conocimiento adquirido mediante los casos de aplicación haciendo uso del Data Mart generado al final del proceso del quinto caso de uso, Sección 4.2.7.

En este Data Mart se encuentran los valores de los consumos térmicos para las dos Centrales además de una nueva variable que indica el cluster al que pertenecen.

2 Aplicación de herramientas

En la Figura 56 se muestra el proceso completo. Este caso de aplicación se ha querido simplificar al máximo para poder mostrar los resultados del algoritmo W-J48 de Weka.

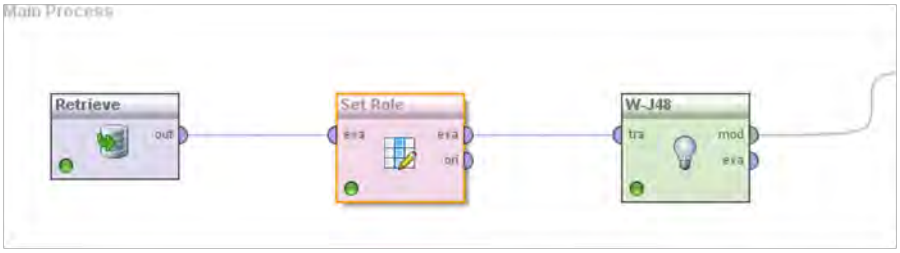


Figura 58: Proceso para el desarrollo de un árbol de decisión proporcionado por Weka

El proceso está formado por tres cloques:

El primero es el Data Mart ya descrito.

El segundo es del tipo “SetRole” y con el que se configurará que la variable especial a tratar sea la del resultado de la clusterización.

EL tercero es el módulo que ejecuta el algoritmo del tipo árbol de decisión desarrollado por Weka.

El resultado proporcionado en forma de reglas es el siguiente: (Figura 57)

W-J48

```
J48 pruned tree
------------------

KwhTermParcial <= 229.7: cluster_1 (138.0)
KwhTermParcial > 229.7
|   KwhTermParcial <= 586.6: cluster_0 (73.0)
|   KwhTermParcial > 586.6
|   |   KwhTermParcial <= 1134.3: cluster_3 (25.0)
|   |   KwhTermParcial > 1134.3: cluster_2 (16.0)

Number of Leaves  :     4

Size of the tree :      7
```

Figura 59: Reglas de inducción desarrolladas por RapidMiner mediante el algoritmo W-J48

3 Post proceso y toma de decisiones

El árbol de decisión final elaborado por el algoritmo Weka muestra las relaciones entre los valores de los consumos parciales con su segmentación en un cluster u otro.

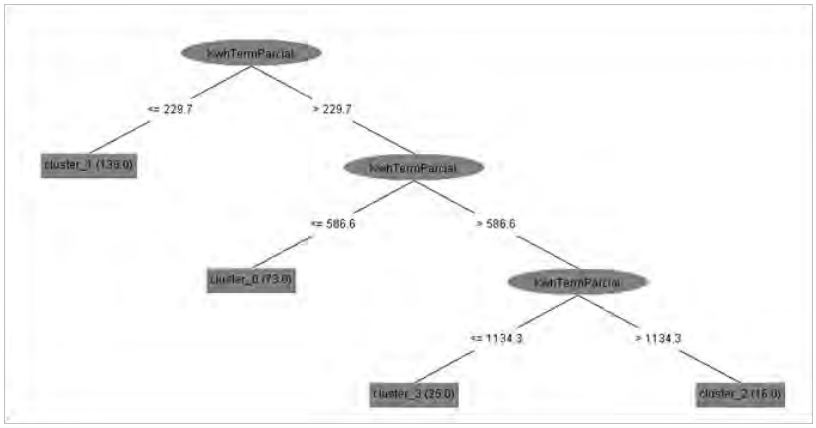


Figura 60: Árbol de decisión resultante de aplicar el algoritmo W-J48 de Weka

Conclusiones del caso de uso:

Como muestra el gráfico resultante los módulos que provienen de Weka no cuentan con la misma interfaz de colores que a los algoritmos proporcionados por RapidMiner. Para el ejemplo planteado tanto Weka como RapidMiner ofrecerán soluciones de árboles de decisión validas.
Sin embargo, los algoritmos desarrollados por Weka son mas valorados en el mundo de la minería de datos. Por ello algunos expertos se inclinan a trabajar con ellos.
El hecho que RapidMiner haya integrado estos algoritmos le proporciona una gran potencia de uso. Además, cabe destacar la facilidad que representa trabajar con RapidMiner dada la sencillez con la que permite cargar datos mediante la creación de Data Marts.

Conclusiones de la sección

El proceso de KDD tal y como se ha demostrado finaliza mediante las fases:

- **Evaluación, interpretación, transformación y representación de los patrones extraídos y interpretar los resultados.**
- **Difusión y uso del nuevo conocimiento. Incorporando el conocimiento descubierto al sistema.**

Se puede concluir de los diferentes ejemplos analizados que la herramienta RapidMiner es de gran potencia y simplicidad. Mediante RapidMiner se pueden combinar los operadores y acompañar al usuario en su proceso de construcción.

Finalmente se puede afirmar que mediante la aplicación de estas herramientas en cada uno de los casos planteados, las técnicas de minería de datos han permitido analizar la información y extraer de ella un conocimiento útil para los evaluadores y consultores energéticos.

8. Conclusiones

El trabajo aquí expuesto forma parte de la fase de estudios previos del proyecto REPENER. Durante esta fase, ha sido necesario llevar a cabo un estudio del arte sobre los sistemas de información y los procedimientos de extracción de conocimiento disponibles que puedan ser de utilizados para crear un repositorio de información energética.

En este sentido, el primer objetivo de este trabajo ha sido llevar a cabo este estudio del arte. A lo largo del estudio se han identificado algunos de los conceptos básicos sobre sistemas de información, como Business Intelligence y Knowledge Discovery Databases.

El estudio del KDD ha permitido diferenciar cuatro procesos básicos: obtención, preprocesado, análisis y extracción de conclusiones.

En el transcurso del proyecto se han analizado y evaluado las diferentes alternativas para la obtención y modelización los datos como paso previo necesario para el desarrollo de un sistema de información energética. Asimismo, se han estudiado las arquitecturas a desarrollar, como es el caso del Data Warehouse.

A continuación se han analizado las técnicas para la toma de decisiones que permiten aprender de los datos del pasado y predecir a partir de ellos situaciones futuras. Para ello se han evaluado un conjunto de algortimos de minería de datos como son árboles de decisión, reglas de asociación o clustering.

Para finalizar, se han aplicado estas técnicas de minería de datos a un caso de estudio utilizando el software Rapidminer. La aplicación prácticca se ha llevado a cabo a partir del análisis de datos reales demostrando la capacidad de extraer un conocimiento a partir de ellos que podría realimentar el sistema alimentándolo con nuevas relaciones entre los datos.

Se puede concluir, por tanto, que el análisis de los datos energéticos con las técnicas de minería de datos podría dar lugar un conocimiento que, utilizado adecuadamente, podría contribuir a mejorar la eficiencia energética de los edificios analizados. . La visualización de las relaciones entre los distintos tipos de datos (consumo, comporamiento, características de los edificios) permitiría a los futuros usuarios de la plataforma REPENER (arquitectos, usuarios, gestores energéticos); proyectar edificios más eficaces energéticamente y mejorar el mantenimiento de los existentes, con la adopción de estrategias más eficientes para minimizar el gasto energético.

a. Líneas de futuro

La principal línea de futuro es el propio desarrollo del proyecto REPENER. La siguiente fase es la de las especificaciones. En esta fase se deberán definir y diseñar los métodos y herramientas explicados en este trabajo, tales como el diseño de la arquitectura del

sistema de información y los protocoles de acceso o configuración de las herramientas de minería de datos.

Existe gran cantidad de proyectos de investigación tanto en ámbito nacional y europeo que permiten considerar otras alternativas que amplíen los resultados de este estudio. A continuación se presentan dos alternativas:

1. La primera es el desarrollo del proyecto basándose en la Red Semántica. Esta alternativa se ha considerado en el Estado del Arte en la Sección 2.2.1. La información se mantendría contenida en el Data Warehouse pero se realizaría el proceso de KDD mediante ontologías. Los resultados se integrarían a la red ontológica y a la vez de guardarían en el Data Warehouse. El desarrollo de una plataforma como la que se describe permitiría estructurar la información basándose en la Red Semántica y la integración de distintas técnicas de minería de datos.

2. La segunda es potenciar la comunicación de la futura plataforma REPENER con otros sistemas de extracción de conocimiento para sí poder relacionar el comportamiento de los materiales o el estudio del comportamiento climático con otros datos energéticos..

b. Tiempo invertido

El proyecto se ha realizado durante el periodo Enero 2010 a Enero 2011.

El cálculo de las horas invertidas en el desarrollo de este proyecto se ha realizado teniendo en cuenta mi dedicación semanal que es de 20 horas.

Las tareas realizadas se dividen en tres:

Estudios: en esta fase se engloban los estudios realizados para la contextualización del proyecto, tecnologías existentes y las nuevas tendencias a aplicar. Representa algo más del 50% del trabajo desarrollado.

Aprendizaje práctico: en esta fase se engloban los procesos de aprendizaje del funcionamiento de las herramientas con las que trabaja REPENER. Tales como la base de datos de Leako, los algoritmos de minería de datos, el uso de diferentes plataformas en el desarrollo de herramientas para la extracción de conocimiento como RapidMiner y la realización de distintas pruebas

Memoria: coste temporal necesitado para la elaboración de esta memoria.

	Tiempo	Porcentaje
Estudios	480 horas	67%
Aprendizaje	320 horas	38%
Memoria	50 horas	10%
Total	850 horas	100%

Figura 61: Coste temporal del proyecto

Agradecer la colaboración del Prof. Dr. German Nemirovskij, profesor de ciencias de la computación en negocios en la universidad Hochschule Albstadt-Sigmaringen, por sus consejos en el proceso de elaboración del proyecto.

También agradecer la colaboración de Félix Iglesias profesor y investigador por sus aportes en el ámbito de la ingeniería relacionada con la eficiencia energética.

9. Bibliografía

Adeva, A. (20 de Septiembre de 2010). *computing.es*. Recuperado el 3 de Enero de 2011, de http://pruebas.computing.es/Noticias/201009200029/COMUNICACIONES-Renfe-colaborara-con-nueve-universidades-y-fundaciones-para-impulsar-la-innovacion.aspx.

AEMet. (30 de Noviembre de 2010). *AEMet*. Obtenido de http://www.aemet.es/es/quienes_somos/que_es.

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial.

Anandarajan Murugan, A. A. (2004). Business Inteligence techniques.

Anandarajan, M., Anandarajan, A., & Srinivasan, C. (2004). *Business Inteligence Techniques.* Springer.

Ben-Nakhi, A. E., & Mahmoud, M. A. (2003). Cooling load prediction for buildings using general regression neural networks.

Bernstein, A., Provost, F., & Hill, S. (2005). Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification.

Borst, W. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse. thesis.*

Bradley, K. (2006). Digital Sustinability and Digital Repositories.

Breslin, M. (s.f.). Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models.

Cacciapaglia, A. (2008). Comparativa de Suitesde Business Intelligence.

Cantos, S., Iglesias, F., & Vidal, J. (2009). Comparison of standard and case-based user profiles in building's energy Performance Simulation.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., C., C. S., y otros. (2000). *Step-by-step data mining guide.* USA: SPSS.

Chung, W. (2004). Benchmarking the energy efficiency of commercial building.

Cloud Security Alliance. (2009). *Security Guidance for Critical Areas of Focus in CLoud Computing v2.1.*

Dataprix. (22 de Abril de 2010). *Dataprix*. Obtenido de http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager#x1-510003.4.5.2.

Energy Efficiency Indicators in Europe. (2010). *http://www.odyssee-indicators.org/*.

Energy future COALITION. (20 de Diciembre de 2010). *Energy future COALITION*. Obtenido de http://www.energyfuturecoalition.org/files/webfmuploads/EFC_Report/EFCReport.pdf.

Fayyad, U., Piatetsky-Shiapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* Merlo Park, California: AAAi Press.

Fedora. (6 de Julio de 2010). *Fedora*. Obtenido de https://wiki.duraspace.org/display/FCR30/Fedora+Repository+3.3+Documentation.

Gruber. (1993). *Toward principles for the design of ontologies used for knowledge sharing.*

Hand, D. (1998). Data Mining: Stadistics and more. *The american Statiscian* .

Hernandez, J., Juan, M., Minaya, N., & Monserrat, C. (2004). Extracción y visualización de Conocimiento de Bases de Datos médicas.

Infovis. (s.f.). *Infovis*. Obtenido de http://www.infovis-wiki.net/index.php?title=Knowledge_Discovery_in_Databases_(KDD).

ISO. (10 de Abril de 2010). *Organización for Standarization*. Obtenido de http://www.iso.org/iso/home.html.

Kepware's WeatherBug for Automation. (2010). *http://www.kepware.com/WeatherBug/*.

Lam, J., Hui, S., & Chan, A. (1997). Regression analysis of high-rise fully air-conditioned office buildings.

Lassila, M. (2001). *The Role of Frame-Based Representation on the Semantic Web.*

Leako. (2 de Enero de 2011). *Leako*. Obtenido de http://www.leako.com/index.php.

Malinowski, E., & Zimányi, E. (2008). Advanced Data Warehouse Design.

Mena, J. (1999). Data Mining your website.

Microstrategy. (10 de Mayo de 2010). *Microstrategy*. Obtenido de http://www.microstrategy.es/Software/businessintelligence/.

Open Data Euskadi. (10 de Diciembre de 2010). *Open Data Euskadi*. Obtenido de http://opendata.euskadi.net/w79-opendata/es/contenidos/informacion/que_es_opendata/es_que_es/que_es_opendata.html.

Oporto, S. (2006). Metodologías para el Data Warehousing.

Pandey, A., Pandey, A., Tandon, A., Maurya, B., & Kushwaha, U. (2001). *Cloud Computing: Exploring the scope.*

Peis, E., Herrera-Viedma, E., & Hassan, Y. a. (2003). *Ontologías, metadatos y agentes: recuperación "semántica" de la información.*

Pentaho Community. (10 de Mayo de 2010). *Pentaho Community*. Obtenido de http://community.pentaho.com/.

Pettey. (4 de Febrero de 2010). *Gartner Announces Business Intelligence Summit 2010*. Obtenido de http://www.gartner.com/it/page.jsp?id=1295220.

Pettey, Goasduff. (13 de Enero de 2010). *Gartner Reveals Five Business Process Management Predictions for 2010 and Beyond*. Obtenido de http://www.gartner.com/it/page.jsp?id=1278415.

Sinnexus. (5 de Diciembre de 2010). *Sinnexus*. Obtenido de http://www.sinnexus.com/business_intelligence/.

Taylor. (10 de Enero de 2011). *Finantial Times*. Obtenido de http://www.ft.com/cms/s/0/0c941764-db04-11df-a870-00144feabdc0.html#axzz1Aizra4Yr.

Vassiliadis, P., Quix, C., Vassiliou, Y., & Jarke, M. (2008). Data warehouse process management.