CAPÍTULO II

Inteligencia Empresarial

El objetivo de este capítulo es explicar con detalle con detalle cada una de las tecnologías de la inteligencia empresarial que se mencionaron en el capítulo I y cómo encajan dentro del proyecto desarrollado. El capítulo se encuentra organizado de la siguiente manera: la sección 2.1 corresponde a los conceptos de *datawarehousing*, la sección 2.2 la parte de OLAP, la sección 2.3 los conceptos de *data mining*, en la sección 2.4 se explica lo que es *Customer Relationship Management* y *Market Basket Análisis*, en la sección 2.5 se muestra el estándar para proyectos de *data mining*, en la sección 2.6 se describen las áreas aplicaciones de apoyo a la toma de decisiones y por último, en la sección 2.7 las conclusiones del capítulo.

2.1 Datawarehousing

W. H. Inmon quien es considerado padre del *datawarehouse*, lo define como una colección de datos diseñada para apoyar la toma de decisiones. Los *datawarehouses* integran datos que han sido seleccionados y organizados de manera histórica, sobre los que se realizan análisis que ayuden a justificar las decisiones estratégicas tomadas en las empresas [Inmon, 2005] [Imhoff, 2003] [Kimball, 2002].

Las características principales de un *datawarehouse* son las siguientes:

- "El *datawarehouse* está orientado a un contexto, organiza y orienta los datos desde la perspectiva del usuario final" [Harjinder, 1996].

- "***Administra grandes cantidades de información***: la mayoría de los *datawarehouses* contienen información histórica que se retira con frecuencia de las bases de datos operacionales" [Inmon, 2005].
- ***Comprende múltiples versiones de esquemas de datos***: debido a que el *datawarehouse* tiene que guardar información histórica, y como ésta ha sido manejada en distintos momentos por diferentes versiones de esquemas, debe poder administrar información originada diferentes bases de datos [Kimball, 2002].
- "***Condensa y agrega información***: Con frecuencia, es muy alto el nivel de detalle de la información guardada. Un *datawarehouse* condensa y agrega la información para presentarla en forma comprensible a las personas" [Harjinder, 1996].

Según Ralph Kimball (et.al. 2002), un *datawarehouse* se compone de los siguientes elementos:

- ***Fuentes de datos de sistemas operacionales***: se refieren a las bases de datos operacionales que contienen información recopilada de las aplicaciones operacionales, que pueden venir en diversos esquemas tales como modelos relacionales, no relaciones o basados en archivos, y pueden ser tanto internas como externas a la organización.
- ***Área mediación de datos***: en esta área se realiza un conjunto de procesos conocidos como de extracción, transformación y carga *(ETL* por sus siglas en inglés*)*. Esta área comprende todo aquello que se encuentra entre la fuente de datos operacionales y el área de presentación de datos.

- ***Área de presentación de datos***: es el lugar en donde los datos son organizados, almacenados y están disponibles para las consultas, reportes y resúmenes que realicen los usuarios.
- ***Herramientas de acceso a los datos***: todas las herramientas de acceso a los datos realizarán consultas a partir del área de presentación de datos. Una herramienta de acceso puede consistir desde una simple consulta, hasta minería de datos y aplicación de modelos.

Para mostrar de manera gráfica los aspectos comunes a todos los *datawarehouses* mencionados anteriormente, se propone la figura 2.1, que representa en sí la arquitectura general de un *datawarehouse*.

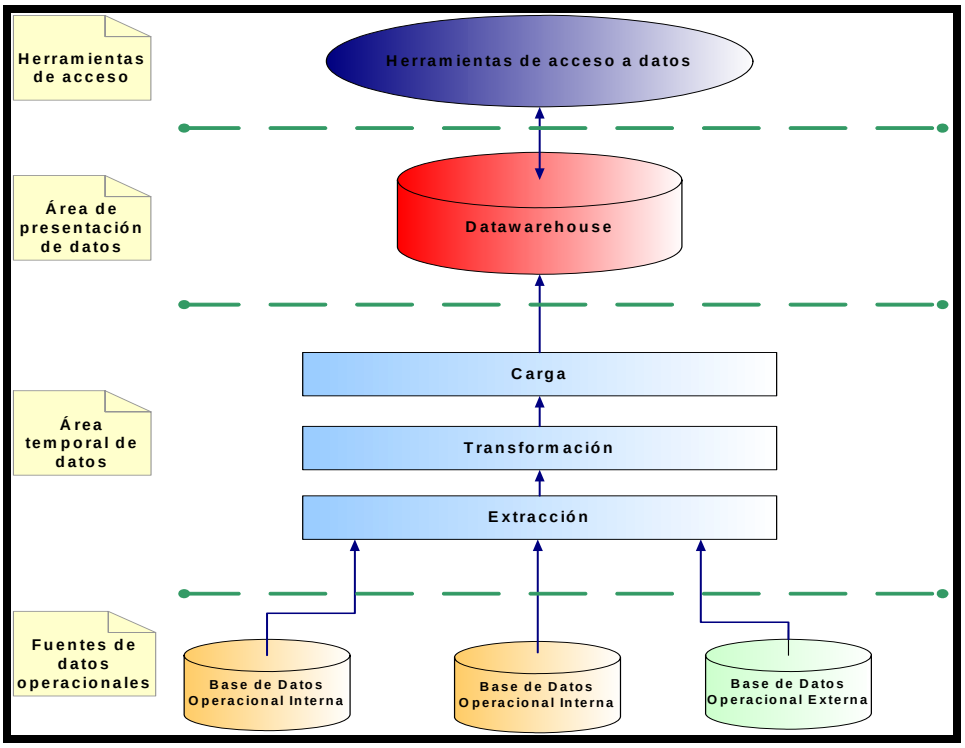

Figura 2.1 Arquitectura general de un datawarehouse

A pesar de que se puede pensar que un *datawarehouse* es semejante a un OLTP, la verdad es que existen diferencias muy significativas entre ellos como podemos ver en la tabla 2.1 [Gardner, 1998] [Imhoff, 2003] [Inmon, 2005] [Kimball, 2002].

	Base de datos transaccional	**Datawarehouse**
Propósito	Operaciones diarias. Soporte a las aplicaciones.	Recuperación de información, informes, análisis y minería de datos.
Tipo de datos	Datos de funcionamiento de la organización.	Datos útiles para el análisis, la sumarización, etc.
Características de los datos	Datos de funcionamiento, cambiantes, internos, incompletos,…	Datos históricos, datos internos y externos, datos descriptivos…
Modelo de datos	Datos normalizados.	Datos en estrella, copo de nieve, parcialmente denormalizados, multidimensionales…
Número y tipo de usuarios	Cientos/miles: aplicaciones, operarios, administrador de la base de datos.	Decenas: directores, ejecutivos, analistas.
Acceso	SQL. Lectura y escritura	SQL y herramientas propias (slice & dice, drill, roll, pívot,etc.) Lectura.

Tabla 2.1 Diferencias entre una base de datos transaccional y un datawarehouse [Hernández, 2004]

Una vez comprendida la definición de un *datawarehouse*, se puede comenzar a diseñarlo e implementarlo. Para construir un *datawarehouse* existen 2 tipos de modelado de datos, [Imhoff, 2003] [Hernández, 2004] [Ponniah, 2007] los cuales son:

- ***Modelo relacional***: propuesto por Inmon (2005). Podemos encontrar más información del modelo relacional en las siguientes referencias [Imhoff, 2003] [Gardner, 1998] [Harjinder, 1996].
- ***Modelo multidimensional***: propuesto por Kimball (et.al. 2002). Se puede encontrar más información al respecto en las siguientes referencias [Imhoff, 2003] [Ponniah, 2007] [Harjinder, 1996].

Los 2 tipos de modelado se explicarán más detalladamente en las secciones siguientes.

2.1.1 Modelo Relacional

El modelo relacional se puede usar para implementar un *datawarehouse* aplicando, sobre el modelo de datos de la empresa, un proceso de transformación de datos de 8 pasos, en este caso particular, de las PyMEs. [Inmon, 2005] [Imhoff, 2003] Los 8 pasos a seguir son los siguientes:

1. ***Selección de los datos de interés.*** El modelo de datos de la empresa es una de las entradas en este paso, además existen otras como el alcance del proyecto, reportes, prototipos, consultas y requerimientos de información. Se debe tener cuidado en la selección de lo datos para no sobresaturar la información que se almacene en el *datawarehouse*.
2. ***Añadir la dimensión de tiempo a las llaves***. Debido a que el modelo del *datawarehouse* representa la información a lo largo del tiempo, se debe agregar el tiempo o fecha a la llave de cada entidad de interés.
3. ***Añadir datos derivados***. Los datos derivados se obtienen como resultado de aplicar operaciones matemáticas a otros datos. Es necesario incluirlos en el *datawarehouse* por razones de optimización y para asegurar la consistencia de los datos.
4. ***Determinar el nivel de granularidad.*** El nivel de detalle que se almacenará en el *datawarehouse* puede variar dependiendo la perspectiva del negocio, técnica o del proyecto. Antes de determinar el nivel de granularidad, es necesario considerar varios factores tales como las necesidades del negocio, de los procesos de *data mining*, el costo de almacenamiento y el desempeño.
5. ***Sumarizar los datos.*** Se utiliza con fines de rendimiento en la entrega de los resultados. Comúnmente se sumarizan los datos por periodos de tiempo.

6. ***Mezclar entidades.*** Combinar 2 o más entidades en una sola, es decir, se denormalizan los datos, para esto deben tener una llave en común.
7. ***Crear arreglos.*** Se utiliza poco, pero si es necesario, puede mejorar considerablemente la población de los *datawarehouses*.
8. ***Separar los datos.*** Se separan tablas con base en su estabilidad y uso.

Estos 8 pasos pueden dividirse en 2 categorías. Los primeros 4 pertenecen a la etapa de creación del *datawarehouse* y los 4 siguientes pasos sirven para mejorar el rendimiento y optimizar el tiempo de respuesta [Inmon, 2005]. Sobre el modelo relacional se pueden aplicar las técnicas de reporteo general y permite almacenar el histórico de la información.

2.1.2 Modelo multidimensional

"El modelo multidimensional, permite tener datos organizados en torno a hechos, que tienen unos atributos o medidas, que pueden verse con mayor o menor detalle según ciertas dimensiones" [Kimball,2002] Se puede encontrar más información al respecto en las siguiente referencias [Imhoff, 2003] [Ponniah, 2007]. Los conceptos importantes que se manejan dentro del modelo multidimensional son:

- ***Hecho***: corresponde a la actividad de la empresa que se desea representar, por ejemplo, las ventas de un supermercado.
- ***Medidas:*** son el conjunto de indicadores del hecho que se escogió para representar. Generalmente responden a la pregunta *¿Cuánto?* Retomando el ejemplo anterior, las medidas para el hecho de las ventas podrían ser: ¿Cuántos productos se vendieron?, ¿Cuánto fue el total de la venta en pesos? ¿Cuánto costaron esos productos vendidos?

- ***Dimensiones:*** son las que van a caracterizar al hecho y responden a las preguntas *¿Dónde? ¿Cuándo? ¿Qué?* Siguiendo con ejemplo, las dimensiones para el hecho de las ventas podrían ser: la fecha de la venta, la hora, o el lugar.
- ***Granularidad***: corresponde al nivel de detalle que será almacenado en las dimensiones. Por ejemplo: para la dimensión de tiempo podemos tener: año, semestre, trimestre, mes, semana, día, hora.

Una vez comprendidos estos conceptos, podemos definir los elementos de los que se compone el modelo multidimensional para su implementación: las tablas y los esquemas [Kimball, 2002] [Silverston, 2001].

Existen dos tipos de tablas, que se muestran gráficamente en la figura 2.2:

- ***Tabla hecho***: en ella se almacenan las medidas y las claves de las tablas de dimensión u otras medidas derivadas, conocidas como funciones de agregación.
- ***Tabla de dimensión***: contiene los datos descriptivos de cada dimensión, también conocidos como atributos de la dimensión.

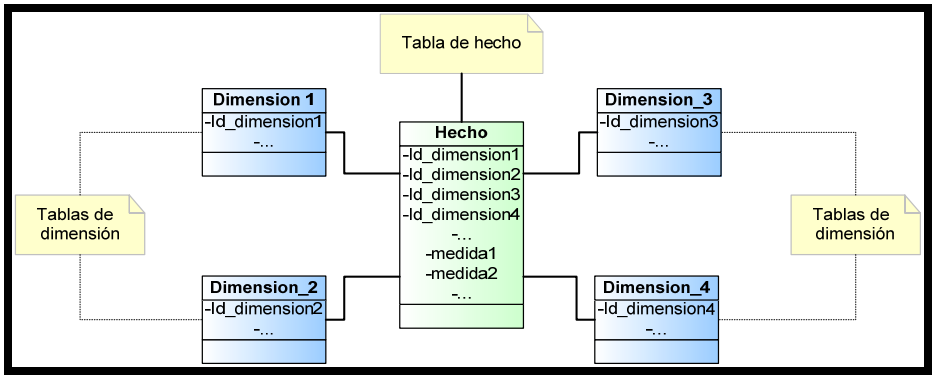

Figura 2.2 Ejemplo de tablas hecho y dimensión

Los esquemas son colecciones de tablas y pueden ser de dos tipos:

- ***Esquema Estrella***: el centro de la estrella consiste en una tabla hecho y las puntas de la estrella son las tablas de dimensión, las cuales tienen una sola conexión a la tabla hecho, a través de su llave primaria, que debe ser un número único y de tipo entero, y no hay caminos alternativos en las dimensiones, es decir no existen jerarquías, como vemos en la figura 2.3. Este esquema es una representación sencilla de los datos que agiliza el tiempo de respuesta en las consultas multidimensionales, sin embargo, dificulta el proceso de actualización de datos, porque usualmente está denormalizado, lo que provoca duplicación de datos o sustitución de llaves por valores de registros y esto ocupa más espacio de almacenamiento [Kimball, 2002].

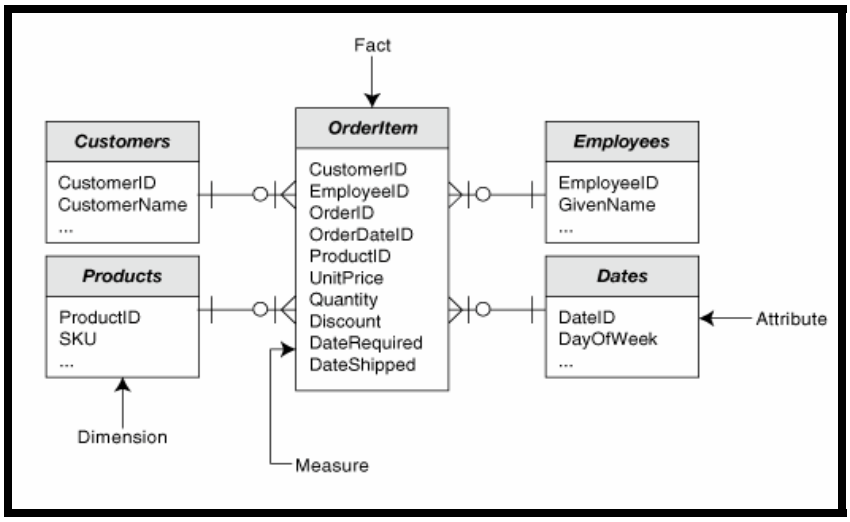

Figura 2.3 Ejemplo de esquema estrella [Riordan, 2005]

- ***Esquema Copo de Nieve***: La diferencia con el esquema estrella es que sí existen caminos alternativos en las dimensiones, es decir jerarquías en las dimensiones. Una de las ventajas de este esquema es que facilita la actualización de los datos del

datawarehouse y ahorra espacio de almacenamiento. Sin embargo, es una representación de datos más compleja que disminuye el tiempo de respuesta en las consultas multidimensionales [Kimball, 2002]. La figura 2.4 es un ejemplo del esquema copo de nieve.

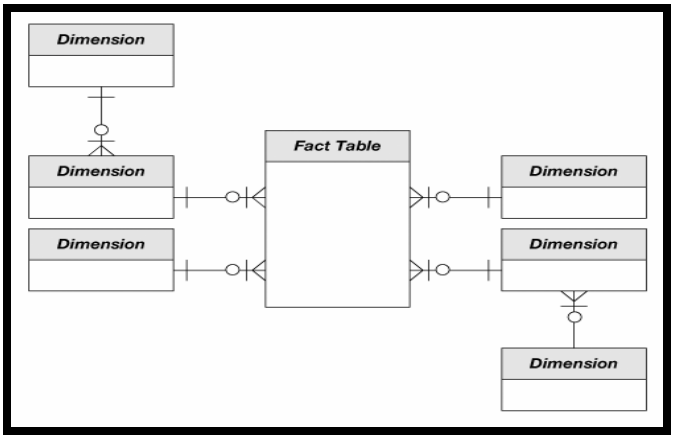

Figura 2.4 Ejemplo de esquema copo de nieve [Riordan, 2005]

- ***Constelación***: es un conjunto de esquemas estrella o copos de nieve que comparten dimensiones, como podemos ver en la figura 2.5.

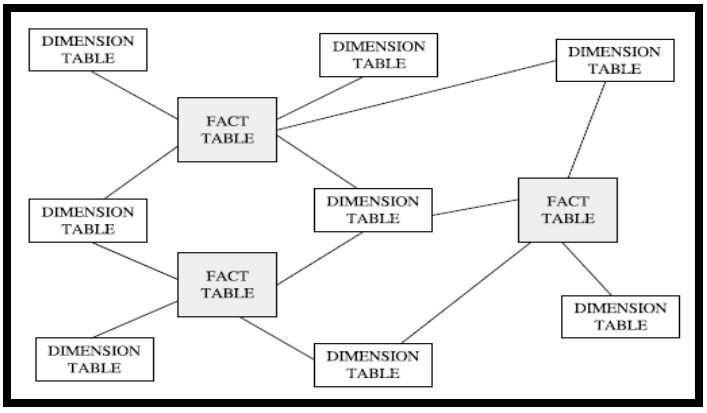

Figura 2.5 Familia de esquemas estrellas: constelación [Ponniah, 2007]

Existen 4 pasos importantes que se deben considerar para la elaboración de un modelo multidimensional:

1. Selección del proceso a modelar.
2. Seleccionar el hecho central y el gránulo máximo que se va a necesitar sobre él.
3. Identificar las dimensiones que caracterizarán el dominio.
4. Determinar y refinar las medidas y atributos a almacenar sobre el proceso.

El modelo multidimensional también permite almacenar el histórico de la información, una ventaja importante es que los esquemas estrella ayudan a enforcarse en hechos particulares y permiten crecer tanto como se quiera. La figura 2.6 presenta gráficamente las diferencias entre ambos modelos.

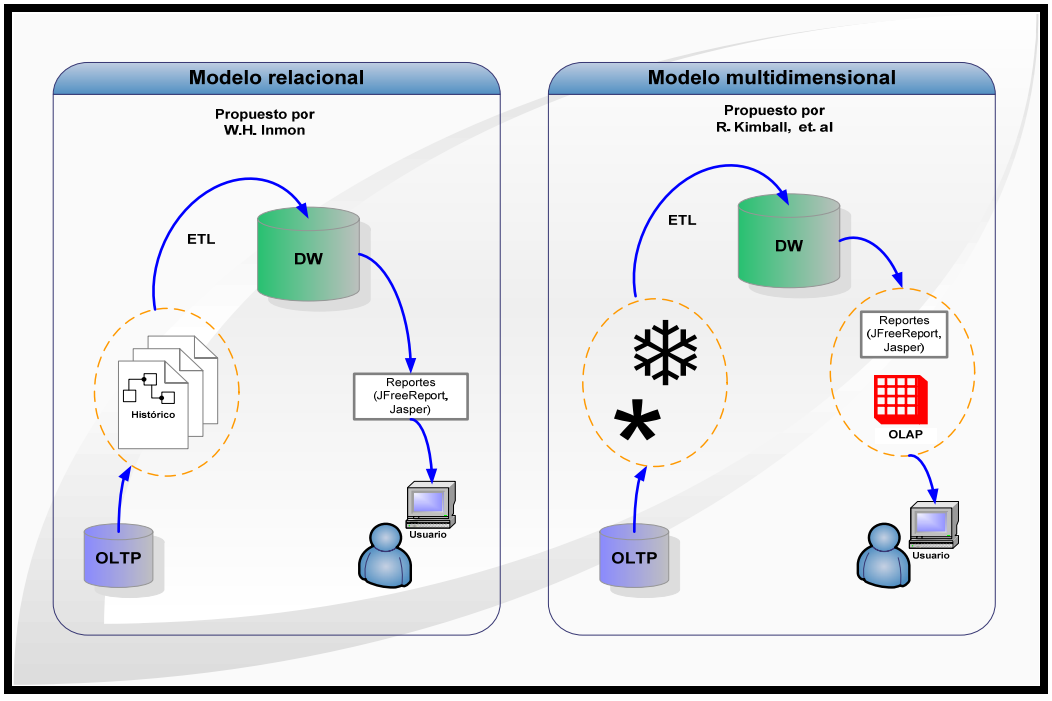

Figura 2.6 Diferencia entre los modelos relacional y multidimensional

A diferencia del modelo relacional, sobre el modelo multidimensional se pueden construir cubos de OLAP, que serán explicados en la siguiente sección, además de las técnicas de reporteo general.

2.1.3 Administración del datawarehouse

Una de las razones principales por las que se construye el *datawarehouse* separado de la base de datos operacional es para conseguir que se realice el análisis de datos de una manera eficiente.

La carga y el mantenimiento de un *datawarehouse* es uno de los aspectos más delicados y que más esfuerzo requiere. Como se mencionó anteriormente el proceso ETL, es el encargado de realizar estas tareas.

La extracción, transformación y carga, comprende las siguientes tareas [Harjinder, 1996]:

- ***Lectura de datos transaccionales***: se trata de obtener, mediante consultas SQL, la información que se necesita de la base de datos transaccional. La primera carga de datos, suele ser la más difícil ya que los datos pueden encontrarse en distintos formatos.
- ***Creación de claves***: es recomendable hacer una distinción entre las claves de las bases de datos transaccionales y las del *datawarehouse* para evitar confusiones.
- ***Integración de datos***: consiste en la unión de datos de distintas fuentes, detectar cuándo representan los mismos objetos y generar las referencias y restricciones adecuadas para conectar la información y proporcionar la integridad referencial.

- ***Limpieza y transformación de datos***: en esta tarea se trata de evitar datos redundantes, inconsistentes, estandarizar medidas, formatos, fechas, tratar valores nulos, etc.
- ***Creación y mantenimiento de metadatos***: para que todo el proceso de ETL pueda funcionar, es necesario crear y mantener metadatos sobre el propio proceso de ETL, los pasos realizados y por realizar.
- ***Identificación de cambios***: se puede realizar de distintas maneras: una carga total cada vez que haya un cambio, comparaciones entre instancias, marcas de tiempo o técnicas mixtas.
- ***Planificación de la carga y mantenimiento***: consiste en definir las fases de carga y el orden de las migraciones para evitar violar las restricciones de integridad. El objetivo es poder hacer la carga sin saturar la base de datos transaccional, así como el mantenimiento sin paralizar el almacén de datos.
- ***Indización***: se deben crear índices en las claves y atributos del *datawarehouse* que se consideren relevantes.

Una vez que se ha implementado de manera exitosa el *datawarehouse* se puede proceder a la aplicación de técnicas que exploten y manipulen la información almacenada, las cuales se verán en las siguientes secciones.

2.2 OLAP

Como se mencionó en el capítulo I, *On Line Analytical Processing* (OLAP) permite analizar grandes cantidades de datos a través del modelo multidimensional, explicado en la sección 2.1.

Esta representación permite mostrar los datos al usuario final de una manera más sencilla y tiene la flexibilidad necesaria para cambiar las perspectivas de visión de la información. OLAP permite realizar análisis históricos complejos con amplia manipulación de los datos [Imhoff, 2003] [Moss, 2003] [Ponniah, 2007].

El análisis de la información se realiza mediante cubos, que son colecciones de dimensiones y medidas, alrededor un hecho particular, sobre los cuales se aplican distintos operadores para dar los resultados a las consultas que se ejecuten. En la figura 2.7 se puede observar que un cubo se compone de ejes, representados por las dimensiones y celdas que son las medidas que se quieren analizar.

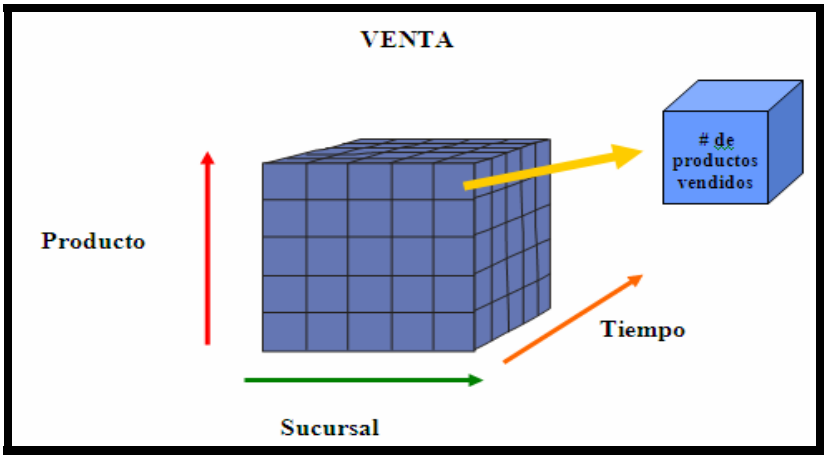

Figura 2.7 Cubo de OLAP

En la figura 2.8 podemos observar un ejemplo de los reportes que podemos obtener con los cubos de OLAP.

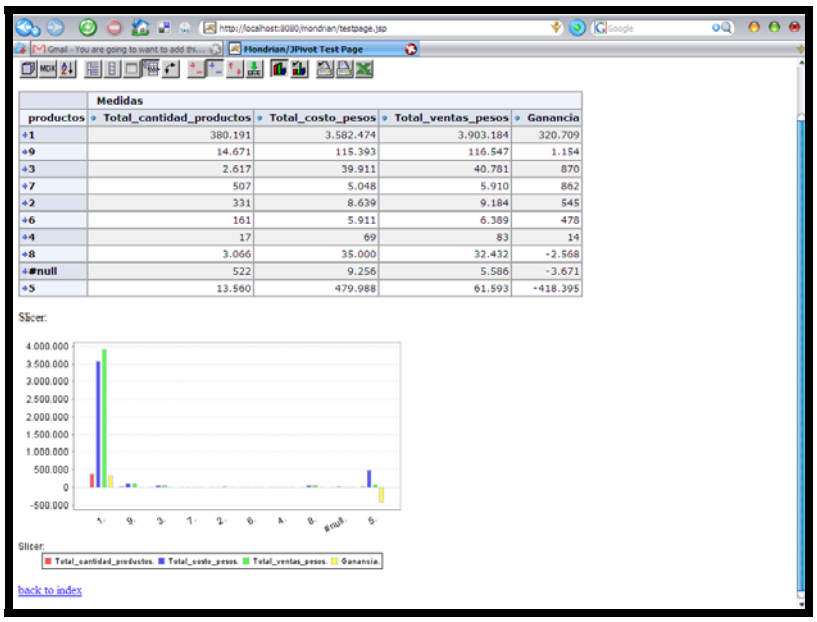

Figura 2.8 Ejemplo de reporte de OLAP

Existen 2 técnicas de almacenamiento/implementación de cubos de OLAP, como se muestra en la figura 2.9, que son:

- ***ROLAP***: "físicamente, el *datawarehouse* se construye sobre una base de datos relacional" [Harjinder, 1996]. Una ventaja de este tipo de esquema es que se pueden utilizar los sistemas de administración de bases de datos relacionales, que son muy populares para el OLTP, y sus herramientas asociadas, además de que el costo necesario para la implementación es mucho menor.
- ***MOLAP***: "físicamente, el *datawarehouse* se construye sobre estructuras basadas en matrices multidimensionales" [Harjinder, 1996]. Las ventajas de este esquema son la

especialización y la correspondencia entre el nivel lógico y el nivel físico. Por esto generalmente MOLAP es más eficiente, debido a que es una solución ad-hoc.

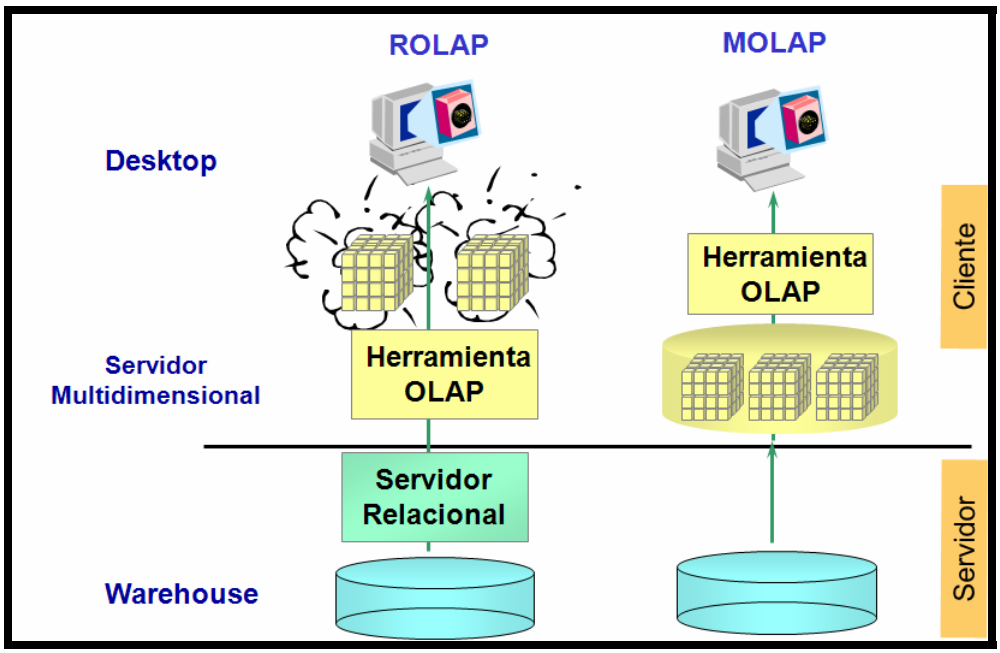

Figura 2.9 Técnicas de almacenamiento de cubos ROLAP y MOLAP [Hernández, 2004]

La diferencia entre ambas técnicas radica en la implementación física y no en la manera en que las herramientas muestren los resultados de las consultas, muchos autores lo manejan de esta manera [Hernández, 2004].

Entre las características principales de OLAP se encuentran las siguientes:

- Presenta una visión multidimensional lógica de los datos en el *datawarehouse*. La visión es independiente de cómo se almacenan los datos.
- Incluye siempre la consulta interactiva y el análisis de los datos. Por lo regular la interacción es de varias pasadas, lo cual comprende la profundización en niveles cada vez más detallados o el ascenso a niveles superiores de resumen y adición.

- Ofrece opciones de modelado analítico, incluyendo un motor de cálculo para obtener proporciones, desviaciones, etc., que comprende mediciones de datos numéricos a través de muchas dimensiones.
- Crea resúmenes, adiciones, jerarquías y cuestiona todos los niveles de adición y resumen en cada intersección de las dimensiones.
- Maneja modelos funcionales de pronóstico, análisis de tendencias y análisis estadísticos.
- Recupera y exhibe datos tabulares en dos o tres dimensiones, cuadros y gráficas. Esto permite analizar los datos desde diferentes perspectivas.

Para explotar la información almacenada en el *datawarehouse* una vez definido el cubo, se pueden aplicar distintos operadores que se muestran en la tabla 2.2. [Spofford, 2006] [Stackowiak, 2007].

Operador	Significado
Drill	**Ofrece mayor nivel de detalle y menos agregación.**
Roll	**Lo contrario a Drill, se tiene mayor agregación y menor nivel de detalle.**
Slice & Dice	**Se proyectan datos de áreas específicas por selección, no por agregación.**
Pívot	**Se reorientan las dimensiones, es decir, las columnas ocupan el lugar de las filas y las filas de las columnas.**
Drill-down/Roll-up	**Se aumentan o disminuyen agregaciones dentro de una consulta ya predefinida.**
Drill-across/Roll-across	**Se obtienen agregaciones en otras dimensiones que no hayan sido comprendidas inicialmente en la consulta, o se desaparecen dimensiones.**

Tabla 2.2 Operadores de OLAP

En la figura 2.10 podemos ver algunos ejemplos de la aplicación de los operadores de OLAP.

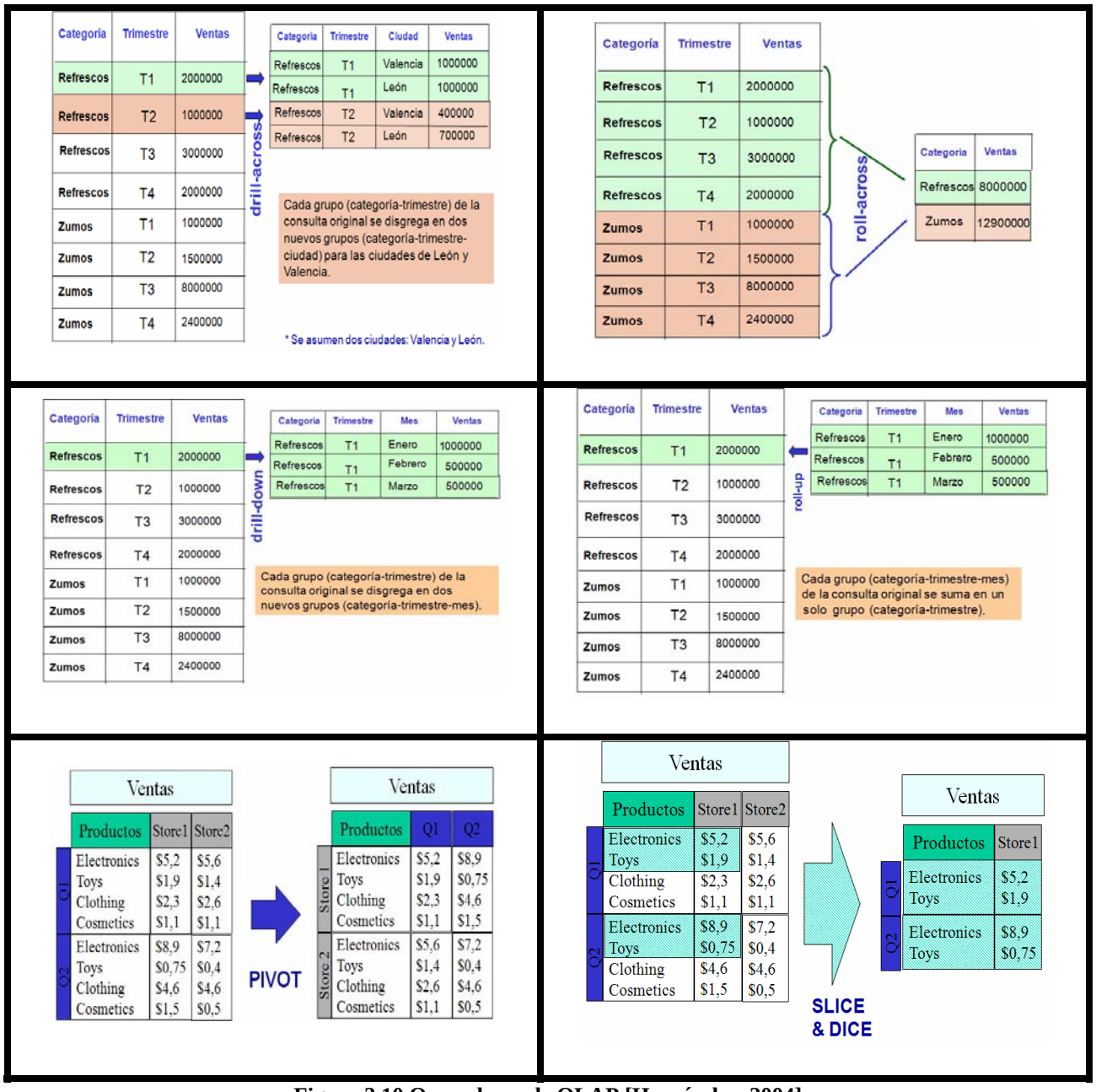

Figura 2.10 Operadores de OLAP [Hernández, 2004]

Una vez que hemos visto la definición de OLAP y los operadores de manipulación de datos que existen, podemos pasar a la realización de consultas, las cuales se definen en leguaje MDX que se explicará con detalle en la siguiente sección.

2.2.1 MDX

MDX significa *multi-dimensional expressions*. MDX fue introducido por Microsoft con *Microsoft SQL Server OLAP Services* alrededor de 1998. Más recientemente MDX apareció como parte de *XML for Analysis API*. Microsoft lo propuso como estándar y su adopción por los desarrolladores de aplicación y otros proveedores de servicios OLAP sigue creciendo [Spofford, 2006].

Un ejemplo sencillo de consultas en lenguaje MDX luce como se muestra en la figura 2.11 y cuyo resultado se muestra en la figura 2.12.

```
select    {[Measures].[Cantidad    (unidades)],[Measures].[Costo    Promedio],
[Measures].[Costo_unitario],[Measures].[Precio (pesos)]} ON COLUMNS,
NON EMPTY {[productos].[productos.Departamento].Members}
ON ROWS from [ventas]
```

Figura 2.11 Consulta en MDX

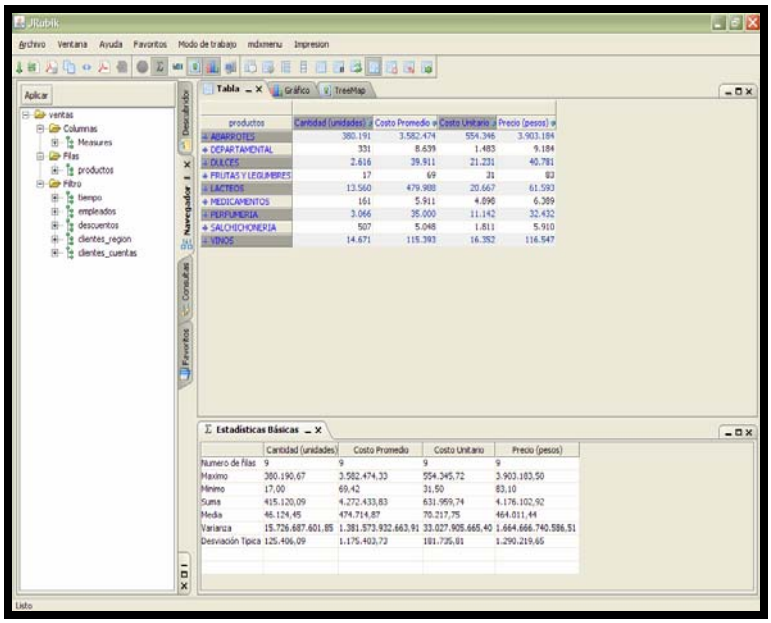

Figura 2.12 Resultado de la consulta MDX

Tiene cierto parecido con SQL [ISO/IEC, 2003] [Silberchatz, 2006] sin embargo su estructura presenta ciertas diferencias. SQL es un lenguaje diseñado para bases relacionales y

transaccionales. Para realizar consultas multidimensionales, se necesita de un nivel de abstracción superior que simplifique las consultas desde el punto de vista multidimensional, el cual finalmente se traduzca a SQL.

En conclusión, el objetivo de OLAP es ayudar al usuario final a entender lo que está sucediendo en la empresa, ya que permite mostrar un análisis concentrado de los datos de la empresa. Esto ayuda a los administradores a mantenerse informados de la situación actual de la organización. Otro tipo de análisis es la técnica de *data mining*, que también utiliza los datos almacenados en el *datawarehouse* y que veremos con más detalle en la siguiente sección.

2.3 Data Mining

Como se mencionó en el capítulo I, *data mining* es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos [Wu, 2004].

Para que este proceso sea efectivo debería ser automático o semi-automático y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras. Por lo tanto, el objetivo de *data mining* es descubrir patrones válidos, novedosos, interesantes y comprensibles, que reporten algún beneficio a la organización [Fayyad, 2000] [Fayyad, 2001]

Dos conceptos importantes en *data mining* son las tareas y los métodos. En las secciones siguientes se explican con más detalle los tipos de tareas y métodos que existen para resolverlas.

2.3.1 Tareas

Una tarea es un tipo de problema de *data mining*. Por ejemplo: clasificar piezas en defectuosas, no defectuosas, defectuosas reparables y defectuosas no reparables es una tarea. Esta tarea se podría resolver mediante árboles de decisión o redes neuronales, entre otros métodos, éstos, son métodos o técnicas que permiten resolver las tareas [Hernández, 2004] [Soukup, 2002].

Existen dos grandes grupos en los que se pueden dividir las tareas, como se pueden ver en la tabla 2.3 [Berry, 2004] [Larose, 2005] [Larose, 2007].

Tipo de Tareas	Descripción	Ejemplos
Predictivas	Se trata de problemas en los que hay que predecir uno o más valores para uno o más ejemplos	Clasificación.
		Clasificación suave.
		Estimación de probabilidad de clasificación.
		Categorización.
		Preferencias o priorización.
		Regresión
Descriptivas	Los ejemplos se presentan como un conjunto de datos sin ordenar ni etiquetar de ninguna manera. Por lo tanto, el objetivo, no es predecir nuevos datos sino describir los existentes	Clustering.
		Correlación y factorizaciones
		Reglas de asociación.
		Dependencias funcionales.
		Detección de valores e instancias anómalas.

Tabla 2.3 Tareas de data mining

De todas estas tareas, las que se abordan en la tesis son tareas descriptivas, específicamente reglas de asociación y clustering. Las tareas de clustering no se implementan físicamente en la tesis, pero se retoman posteriormente en las conclusiones. En seguida se explicarán con más detalle estas 2 tareas.

2.3.1.1 Reglas de Asociación

"Las reglas de asociación son una manera muy popular de expresar patrones de datos" [Hilderman, 1998] [Witten, et.al, 2000]. Estos patrones pueden servir para conocer el comportamiento general de un problema, y de esta manera, tener más información que pueda apoyar en la toma de decisiones.

Las reglas de asociación surgieron inicialmente para afrontar el análisis de las canastas de compra en los comercios (MBA). Una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertos estados en una base de datos. Una típica regla de asociación sería la que se muestra en la figura 2.13

SI gansito "Marinela" Y leche "Lala" ENTONCES galletas "Marías"

Figura 2.13 Ejemplo de una regla de asociación

La parte izquierda de la regla "**SI** gansito Marinela **Y** leche Lala" se conoce como *premisa*, mientras que la parte derecha "**ENTONCES** galletas Marías" se conoce como *consecuencia* o *consecuente*. Esto significa que si se compró un gansito "Marinela" y una leche "Lala", probablemente se compren unas galletas "Marías".

Dada una regla de asociación se suele trabajar con 2 medidas para conocer y evaluar la calidad de la regla: soporte y confianza.

"El soporte, también conocido como cobertura, de una regla se define como el número de instancias que la regla predice correctamente. Por otro lado, la confianza, también conocida como precisión, mide el porcentaje de veces que la regla se cumple cuando se puede aplicar"

[Hilderman, 1998]. Los algoritmos de *data mining* trabajan en la búsqueda de reglas que cumplan con unos requisitos mínimos en estas medidas.

Por ejemplo como vemos en la figura 2.14 tenemos 4 reglas de asociación en donde solo 3 de ellas se cumplen para: ***Si*** *aceite* **y** *harina* ***entonces*** *frijoles,* por lo tanto tenemos que el soporte de la regla es de 3, es decir, el número de reglas que se encuentra esa pareja en todas las reglas y tiene una confianza de ¾ o 0.75 es decir, el número de veces que se cumple la regla sobre el número de veces que aparece esa pareja.

SI aceite Y harina ENTONCES frijoles | **SI aceite Y harina ENTONCES frijoles |** **SI aceite Y harina ENTONCES frijoles |** **SI aceite Y harina ENTONCES galletas**	**Soporte=3 Confianza =$^{3}/_{4}$ o 0.75**

Figura 2.14 Ejemplos de reglas de asociación con soporte y confianza

Es importante resaltar que al cambiar el orden de las partes de una regla tenemos reglas diferentes, ya que las medidas de soporte y confianza varían, es decir que se tienen distintas posibilidades de que la regla se cumpla si sus partes se encuentran en distinto orden. Para el ejemplo de la regla "**SI** gansito Marinela **Y** leche Lala **ENTONCES** galletas Marías" tenemos un soporte de 0.4 y una confianza de 0.90 al cambiar la regla a "**SI** galletas Marías **ENTONCES** gansito Marinela **Y** leche Lala" tendremos un soporte de 0.6 y una confianza de 0.95.

2.3.1.2 Clustering

"El clustering es una de las tareas más frecuentes en *data mining*. Se trata de encontrar grupos entre un conjunto de individuos. El concepto de distancia puede jugar un papel crucial, que individuos similares deberían ir a para al mismo grupo" [Witten, et.al, 2000].

Para evaluar los resultados de este tipo de tarea se suele utilizar la distancia entre los grupos. Cuanto mayor sea la distancia entre los grupos, significa que se ha efectuado una mejor separación, y por tanto el modelo de agrupamiento se considera mejor, por ejemplo: la agrupación de páginas de Internet en grupos, conocidos como directorios, que realizan Google o Yahoo!.

2.3.2 Métodos/Técnicas

Cada una de las tareas mencionadas en la sección 2.3.1, requiere de técnicas, métodos o algoritmos, para resolverlas. Entre las técnicas más importantes se encuentran:

- Técnicas algebraicas y estadísticas.
- Técnicas bayesianas.
- Técnicas basadas en conteo de frecuencias y tablas de contingencia.
- Técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas.
- Técnicas relacionales declarativas y estructurales.
- Técnicas basadas en redes neuronales artificiales.
- Técnicas basadas en núcleo y máquinas de soporte vectorial.
- Técnicas estadísticas y difusas.
- Técnicas basadas en casos, en densidad o distancia.

Como se puede observar, existen muchas técnicas que pueden aplicarse a la solución de distintas tareas, en ocasiones una misma tarea puede ser resuelta por diferentes técnicas sin embargo, algunas de ellas entregan resultados más pertinentes que otras.

Para resolver las tareas de reglas de asociación y clustering, descritas anteriormente, hemos escogido los siguientes métodos.

2.3.2.1 Método Apriori

Este algoritmo resuelve tareas de reglas de asociación. Su funcionamiento se basa en la búsqueda de los conjuntos de productos con determinado soporte. Para ello, en primer lugar se construyen simplemente los conjuntos formados por un solo producto que superan el soporte mínimo. Este conjunto de conjuntos se utiliza para construir el conjunto de conjuntos de 2 productos, y así sucesivamente hasta que se llegue a un tamaño en el cual no existan conjuntos de productos con el soporte y confianza requerida [Larose, 2005] [Pérez, et.al., 2006] [Witten, et.al., 2000].

2.3.2.2 K-medias

El algoritmo K medias o *K-means* se trata de un método de clustering por vecindad, en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar, sin etiquetar.

Es el método más popular para resolver tareas de clustering. La idea de K medias es situar los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares.

El método tiene una fase de entrenamiento, que puede ser lenta, dependiendo del número de puntos a clasificar y de la dimensión del problema. Pero una vez terminado el entrenamiento, la clasificación de nuevos datos es muy rápida, gracias a que la comparación de distancias se realiza

sólo con los prototipos. Normalmente se utiliza la distancia euclidiana. Un ejemplo de éste algoritmo lo podemos ver en la figura 2.15 en donde se realizan grupos de clustering [Berry, et.al., 2004]. Para más información respecto al algoritmo se pueden consultar las siguientes referencias [Larose, 2005] [Larose, 2007] [Witten, et.al., 2000].

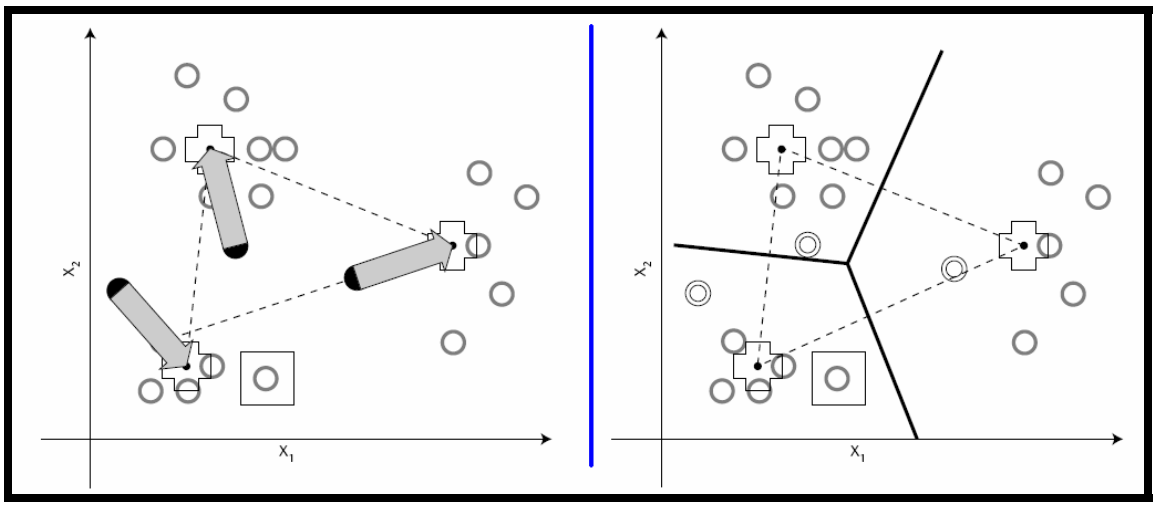

Figura 2.15 Ejemplo del algoritmo de clustering K-medias [Berry, et.al., 2004]

En las siguientes secciones se describirán más detalladamente los conceptos de negocios que están involucrados en la tesis.

2.4 Customer Relationship Management

Customer Relationship Management (CRM) es una estrategia de mercado cuyo objetivo principal es establecer relaciones duraderas con los clientes, como se menciona en el capítulo I. Las empresas requieren entender a cada cliente de manera individual y utilizar ese conocimiento para hacer negocios con ellos más fácilmente que sus competidores.

Construir negocios alrededor de las relaciones con el cliente es un cambio revolucionario para muchas compañías. Se necesita mucho más que *data mining* para cambiar una empresa enfocada en el producto a una empresa enfocada en el cliente. Por ejemplo: un resultado de *data mining* sugiere que ofrecer a un cliente en particular el producto A en vez del producto B, será ignorado si la comisión del administrador depende del número de productos B vendidos en un trimestre y no el número de productos A (incluso si el producto A es más rentable) [Quevedo et.al., 2006] [Berry, et.al, 2004].

Sin embargo las tecnologías de *datawarehousing*, OLAP y *data mining*, son herramientas que pueden ayudar a las organizaciones a concentrarse en sus clientes en vez de sus productos. Recordemos que estas tecnologías son procesos y metodologías que se deben adoptar para obtener los beneficios deseados para el negocio [Cunningham, 2004].

La empresa debe ser capaz de aprender de lo que ha sucedido en el pasado. Deben existir sistemas de procesamiento de transacciones que capturen las interacciones con el cliente, *datawarehouses* para almacenar la información del comportamiento histórico de los clientes, tecnologías de *data mining* y OLAP para planificar acciones futuras a realizar, todo esto para elaborar un estrategia para CRM que se pueda poner en práctica [Gondar, 2003].

No basta lograr un cierto nivel de satisfacción del cliente, es necesario obtener su fidelidad. Monitorear y capturar la satisfacción del cliente no revela a las empresas aquello que quieren saber acerca de sus clientes.

Si la calidad percibida por el cliente es mayor a la calidad esperada, entonces obtenemos la satisfacción del cliente y por consiguiente su fidelidad. Esto quiere decir que si damos al cliente más de lo que espera, estará satisfecho y se convertirá en un cliente fiel a la empresa.

Existen diversas técnicas que CRM emplea para fortalecer las relaciones con los clientes. Algunas de las más populares son [Quevedo et.al., 2006]:

- ***Target Marketing*:** Mediante esta técnica se puede utilizar una lista de potenciales/actuales clientes para enviarles publicidad dirigida.
- ***Risk analysis. Credit Scoring*:** Reducir la posibilidad de otorgar préstamos a personas potencialmente insolventes.
- ***Market Basket Analysis***: Determina grupos de productos que tiendan a presentarse juntos en una transacción o compra.
- ***Cluster Analysis*:** Segmentación de mercado.

Específicamente en esta tesis se ha elegido la técnica de *Market Basket Analysis* o MBA que se explicará con más detalle en la siguiente sección.

2.4.1 Market Basket Analysis

Market Basket Analysis (MBA) analiza las combinaciones de las compras realizadas por los clientes y el número de veces que se repiten, a través de esto se obtienen reglas de asociación, que explican la probabilidad de compra simultánea de productos diferentes. Un ejemplo de la técnica de MBA se muestra en la tabla 2.4 y en la figura 2.16 las reglas de asociación obtenidas [Hernández, 2004].

	Vino "El cabezón"	Gaseosa "Chispa"	Vino "Tío Paco"	Horchata "Xufer"	Bizcochos "Goloso"	Galletas "Trigo"	Chocolate "La vaca"
T1	1	1	0	0	0	1	0
T2	0	1	1	0	0	0	0
T3	0	0	0	1	1	1	0
T4	1	1	0	1	1	1	1
T5	0	0	0	0	0	1	0
T6	1	0	0	0	0	1	1
T7	0	1	1	1	1	0	0
T8	0	0	0	1	1	1	1
T9	1	1	0	0	1	0	1
T10	0	1	0	0	1	0	0

Tabla 2.4 Canasta de compras [Hernández, 2004]

1. Si bizcochos "Goloso" Y horchata "Xufer" ENTONCES galletas "Trigo"
2. Si bizcochos "Goloso" Y galletas "Trigo" ENTONCES horchata "Xufer"
3. Si galletas "Trigo" Y horchata "Xufer" ENTONCES bizcochos "Goloso"
4. Si galletas "Trigo" ENTONCES horchata "Xufer" Y bizcochos "Goloso"
5. Si bizcochos "Goloso" ENTONCES horchata "Xufer" Y galletas "Trigo"
6. Si horchata "Xufer" ENTONCES bizcochos "Goloso" Y galletas "Trigo"

Figura 2.16 Reglas de asociación obtenidas de la canasta de compras

MBA es una de las técnicas más importantes para CRM y marketing, sobre todo para el área de las ofertas o promociones. Se aplica con el fin de que aquellos productos que están fuertemente asociados no sean colocados en una promoción de manera simultánea. De esta manera, al colocar en promoción un producto determinado, se obtiene además, el efecto de aumentar las ventas del producto asociado [Girija, 2006] [Kohavi, 2004].

Los resultados de aplicar esta técnica se pueden utilizar para:

- Determinar la organización física del negocio.

- Diseñar estrategias de marketing del negocio en cuanto a: promociones, displays, publicidad, paquetes de productos.

El ámbito de aplicación de MBA es principalmente el de los supermercados. Sin embargo, también se puede utilizar en tarjetas de crédito, servicios de telecomunicaciones, servicios bancarios, aseguradoras, servicios médicos, entre otros [Kohavi, 2004].

En la siguiente sección veremos el estándar que se sigue para la realización de proyectos de *data mining*.

2.5 Estándar CRISP-DM

El estándar CRISP-DM versión 1.0 (Cross Industry Standard Process for *Data mining*) [CRISP-DM, 2006], es un modelo que se aplica a los proyectos de *data mining*, y aunque el proyecto desarrollado en esta tesis es de inteligencia empresarial, se decidió aplicar el estándar ya que de hecho en un proyecto de *data mining* se involucra la integración de los datos y existe una gran semejanza entre los 2 tipos de proyectos.

La figura 2.17 muestra el modelo del proceso, que sigue un proyecto de *data mining*, en este caso de inteligencia empresarial, así como las fases y tareas que se deben seguir.

Como se observa el ciclo de vida del proyecto consiste en 6 fases que no necesariamente son secuenciales. Esto dependerá de los datos que se tengan y de los resultados que se quieran obtener con dicho proyecto. Las flechas indican las dependencias más importantes entre las fases y el círculo exterior simboliza el ciclo del proyecto, el cual vuelve a iniciar aún después de haber

terminado el proyecto para obtener nuevos datos y resultados interesantes con la información obtenida del ciclo anterior.

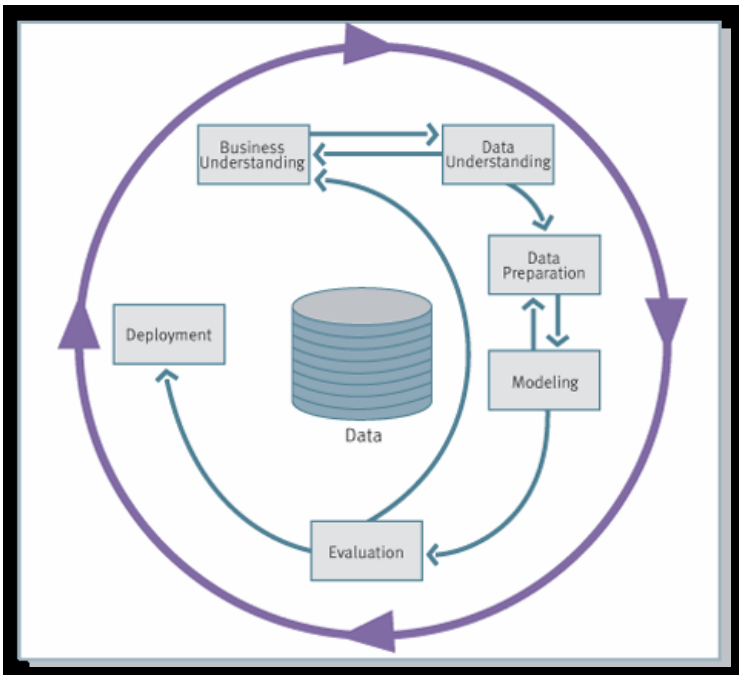

Figura 2.17 Ciclo de vida del proyecto [CRISP-DM, 2006]

En seguida se describirán cada una de las fases de las que se compone el ciclo:

1. ***Comprensión del negocio***. Esta fase inicial se enfoca en la compresión de los objetivos del proyecto y los requerimientos desde la perspectiva del negocio. Esta información se convierte en conocimiento para la definición del problema y el diseño del plan para alcanzar los objetivos.
2. ***Compresión de los datos***. Esta fase comienza con una colección de datos y sobre la cual se realizan actividades para familiarizarse con ellos, para identificar los

problemas de calidad y detectar subconjuntos interesantes para formar hipótesis sobre la información escondida.

3. ***Preparación de los datos***. Esta fase involucra todas las actividades para construir el conjunto final de datos a partir del conjunto inicial. Estas actividades se llevan a cabo varias veces y no en un orden predefinido. Entre estas tareas se encuentran la selección de atributos así como la transformación y limpieza de los datos para las herramientas de modelado.
4. ***Modelado***. En esta fase se seleccionan y aplican varias técnicas de modelado o algoritmos y sus parámetros son calibrados para obtener los mejores resultados. Algunos algoritmos tienen requerimientos específicos para el formato de los datos, en cuyo caso se debe regresar a la fase de preparación de datos y realizar las tareas necesarias para obtener dicho formato, las veces que sean necesarias.
5. ***Evaluación***. Antes de seguir con el despliegue final del modelo, es importante evaluar el modelo propuesto, revisar los pasos ejecutados para construirlo y estar seguros de que se han alcanzado los objetivos iniciales. También es importante verificar si existe algún punto del negocio que no se haya considerado antes.
6. ***Despliegue.*** La creación del modelo generalmente no es el final del proyecto. El conocimiento obtenido del modelo de datos necesita organizarse y presentarse de una manera para que el usuario lo pueda utilizar. En muchos casos será el usuario y no el analista el que desempeñe estos pasos. Sin embargo es importante que el analista explique al usuario las acciones necesarias para utilizar el o los modelos creados.

2.6 Aplicación a la toma de decisiones

El área de apoyo a la toma de decisiones constituye un área multidisciplinar cuyo objetivo es la introducción de métodos y/o herramientas que ayuden a las personas en la toma de decisiones clave.

La fase de toma de decisiones usualmente se refiere al proceso necesario para realizar la selección de una opción o alternativa. Este proceso incluye: conocer el problema, recoger información sobre el problema, identificar alternativas, anticipar consecuencias de posibles decisiones, realizar la selección utilizando juicios lógicos y coherentes basados en la información disponible.

Entonces se puede definir el área del apoyo a la toma de decisiones en 2 partes: la primera que concierne a la toma de decisiones por parte del personal involucrado y la segunda que corresponde al estudio de técnicas que asistan a las personas a mejorar las decisiones tomadas [Sprague, 1996]. Entre esas técnicas podemos ubicar a: *datawarehousing*, OLAP y *data mining*, que nos ayudan a formar un sistema completo de inteligencia empresarial.

Existen diversas áreas de aplicación en las que ya se han incorporado estas técnicas para apoyar la toma de decisiones, las más importantes se muestran en la tabla 2.5:

Área de aplicación	Ejemplos
Aplicaciones financieras	Obtención de patrones de uso fraudulento de tarjetas de crédito, determinación del gasto en tarjeta de crédito por grupos, cálculo de correlaciones entre indicadores financieros, análisis de riesgo en créditos.
Análisis de mercado, distribución y comercio	Análisis de la canasta básica de mercado, evaluación de campañas publicitarias, análisis de la fidelidad de los clientes, estimación de inventarios, costos y ventas.
Seguros y salud privada	Determinación de clientes potencialmente caros, identificación de patrones de comportamiento para clientes con riesgo, identificación de comportamiento fraudulento,

	predicción de clientes que podrían ampliar su póliza.
Educación	Selección o captación de estudiantes, detección de abandonos y fracasos, estimación de tiempo de estancia en la institución.
Procesos industriales	Extracción de modelos sobre comportamiento de compuestos, detección de piezas con defectos, predicción de fallos y accidentes, estimación de composiciones óptimas en mezclas, extracción de modelos de costos, extracción de modelos de producción.
Medicina, biología, bioingeniería y otras ciencias	Diagnóstico de enfermedades, detección de pacientes con riesgo de sufrir una enfermedad concreta, recomendación priorizada de fármacos para una misma enfermedad, predecir si un compuesto químico causa cáncer, clasificación de cuerpos celestes, predicción del recorrido y distribución de inundaciones, modelos de calidad de aguas.
Telecomunicaciones	Establecimiento de patrones de llamadas, modelos de carga en redes, detección de fraude.

Tabla 2.5 Áreas de aplicación de apoyo a la toma de decisiones [Hernández, 2004]

2.7 Discusión final

En este capítulo se han mencionado los conceptos esenciales en los que se basa la inteligencia empresarial: *datawarehousing*, OLAP y *data mining*. Así como los conceptos de negocios que están relacionados con la problemática de la tesis: CRM y MBA. Estos conceptos ayudarán y facilitarán la comprensión de los siguientes capítulos, en donde se explican algunas herramientas, el análisis, diseño e implementación del prototipo.

En el siguiente capítulo se describirán algunas de las herramientas más populares que dan soporte a las tecnologías de la inteligencia empresarial.