CAPÍTULO IV

Análisis y diseño

El objetivo de este capítulo es describir la PyME que será nuestro caso de estudio, así como el análisis y diseño del sistema PIE, las consideraciones y requerimientos tomados en cuenta para su implementación, su arquitectura, y el modelo de datos definido. El capítulo IV se encuentra organizado de la siguiente manera: en la sección 4.1 se describe la empresa que sirvió como caso de estudio, se enlistan las necesidades de tecnologías de información, y los recursos tecnológicos con los que cuenta y los que hacen falta, en la sección 4.2 se describe la arquitectura del sistema, en la sección 4.3 se describe el modelo de datos para el *datawarehouse*, en la sección 4.4 se describe las consideraciones del proceso de ETL, en la sección 4.5 se describen las consultas y reportes que se generan con la técnica de OLAP, en la sección 4.6 se describen los algoritmos, archivos y otros datos necesarios para el proceso de *data mining* y por último, las conclusiones del capítulo en la sección 4.7.

4.1 Caso de estudio

La *Abarrotera Coscomatepec*, la cual fungió como caso de estudio, se encuentra ubicada en Coscomatepec de Bravo, municipio de Veracruz en Av. Bravo y Ocampo No. 1 Col. Centro. Tiene cinco sucursales, cada una de ellas cuenta aproximadamente con quince empleados y maneja un promedio de 10,000 transacciones de ventas mensuales, en cada sucursal.

Figura 4.1 Bodega del supermercado

Retomando la tabla de clasificación de empresas que se mostró en la sección 1.1, se puede observar que debido al número de empleados y por el rango de facturación de la empresa *Abarrotera Coscomatepec* se ubica, específicamente, en la clasificación de mediana empresa del sector comercio.

Figura 4.2 Área de cajas del supermercado

Hace 9 meses la empresa adquirió un sistema de administración, llamado SAN (Sistema Administrativo de Negocios), el cual ha sido instalado en una de las sucursales, y está en proceso de instalación en las demás sucursales. El sistema SAN, que vemos en la figura 4.3, está diseñado para controlar las operaciones administrativas de los negocios. Algunas de sus principales características son:

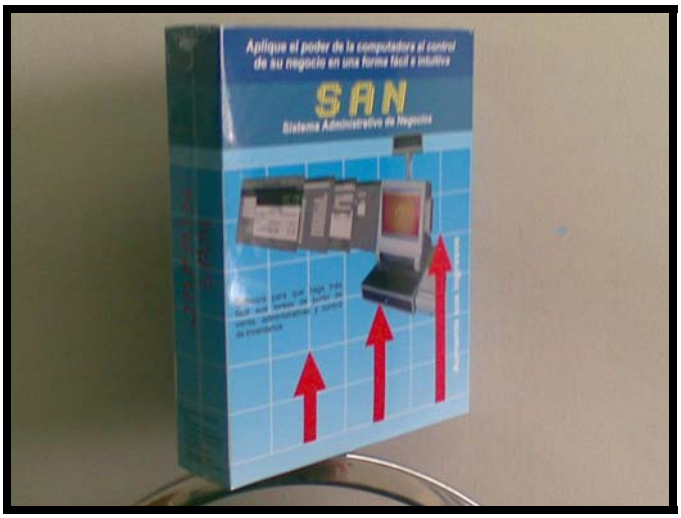

Figura 4.3 Sistema Administrativo de Negocios (SAN)

- Aplicación de control de negocios para ambiente WINDOWS® desarrollado en DELPHI®.
- Multi-Usuario (operación en red).
- Control de inventarios de múltiples.
- Módulo de punto de venta y módulo verificador de precios (opcional) que trabajan en forma integral con el sistema.

- Control de artículos con código de barras o código libre, múltiples precios, múltiples agrupaciones, múltiples empaques (manejo de código DUN), ensambles, rompimientos de precio automáticos, artículos pesables, imagen por artículo, ofertas programadas por artículo, puntos mínimos, mix & match etc.
- Control de clientes, proveedores, cuentas de cheques, cuentas de gastos.
- Facturación para ventas de mayoreo y venta al detalle por medio del punto de venta con emisión de comprobantes simplificados, compras, ajustes y transferencias entre almacenes.
- Manejo de pedidos de clientes y de órdenes de compra a proveedores.
- Consolidación de información de tiendas que operan en puntos remotos usando medios de almacenamiento o por medio de red.
- Gran variedad de reportes para el control del negocio.

Este sistema le permite al supermercado mantener el control de su negocio y tener un panorama general y detallado de la situación de la empresa, lo que ayuda a tomar decisiones justificadas con la información reciente y de determinados procesos. Sin embargo no tiene la flexibilidad para realizar análisis de la información más específicos. Por ejemplo: a lo largo del tiempo, historial de compra de un cliente específico, patrones de productos vendidos conjuntamente.

El supermercado maneja lo que se conoce como venta en ruta, este proceso consiste en la toma de pedidos por parte de los clientes mayoristas y posteriormente sale una camioneta a surtir dichos pedidos.

En ocasiones este proceso se complica porque lleva tiempo cargar la camioneta y en el orden que requieren los pedidos, ya que los productos de la bodega se encuentran dispersos. Para mejorar este proceso sería recomendable realizar un análisis para reestructurar la distribución de la bodega de manera que se facilite la carga de la camioneta.

Para poder realizar el análisis es necesario entender la estructura del negocio como se estable en la primera fase del estándar CRISP-DM. En seguida veremos el modelo de datos que maneja en el supermercado.

4.1.1 Modelo de datos de la empresa

La empresa cuenta con un catálogo de 13169 productos. Existen 2 tipos de clientes: clientes mayoristas, es decir, clientes de los cuales si se conoce su identidad y se les expide factura, y clientes generales o de menudeo de los cuales no se conoce su identidad y no se les expide factura. Existen alrededor de 625 clientes mayoristas.

En el apéndice A se puede observar el diagrama entidad-relación de los datos que maneja la empresa. Este diagrama representa un componente importante del *datawarehouse*, el esquema de las fuentes de datos operacionales que se menciona en el capítulo 2. El diagrama será utilizado para definir y seleccionar la información que se requiere almacenar el *datawarehouse*.

En las siguientes secciones veremos las necesidades de tecnologías de información que se requieren en el supermercado.

4.1.2 Necesidades de tecnologías de la información

Los sistemas de administración ofrecen cierta información que puede ayudar a la toma de decisiones en algunas ocasiones, sin embargo esta información es limitada y poco flexible.

Dado que se busca el análisis de toda la información relevante de la empresa, y el sistema de administración que maneja la empresa ofrece un análisis limitado, se requiere implementar las técnicas mencionadas en el capítulo II, para poder realizar un análisis con mayor profundidad que otorgue resultados con base en toda la información histórica de la empresa y no en análisis parciales de algunos segmentos de la información.

Se requiere implementar un *datawarehouse*, para almacenar los datos necesarios que mediante la aplicación de las técnicas de OLAP y explotación con las técnicas de *data mining*, que otorguen información adicional y relevante a la que brinda el sistema de administración.

También se requiere la posibilidad de visualización la información en diferentes perspectivas (dimensiones) para tener más argumentos que justifiquen las decisiones que sean tomadas en la empresa. Es necesaria también, la búsqueda de patrones en las ventas que permitan modelar el comportamiento de compra de los clientes, para tomar decisiones que comprueben las ventajas competitivas existentes o que proporcionen datos necesarios para generar nuevas y que ayuden a mejorar los procesos que se manejan en el supermercado.

4.1.3 Recursos tecnológicos con los que cuenta

La empresa cuenta en promedio con 4 terminales de punto de venta para el público general, 3 terminales para mayoristas y un servidor por cada sucursal.

Las características de las terminales de punto de venta, mayoristas y servidor son las siguientes:

- Sistema Operativo Windows XP
- Procesador Intel Pentium IV a 3.2Ghz
- 1GB en memoria RAM
- Disco Duro S-ATA 120 Gb

4.1.4 Recursos tecnológicos que necesita

Probablemente se requiera una nueva terminal, con características similares a las que ya se tienen, que cuente con el sistema de inteligencia empresarial para realizar el análisis de la información o designar una de las terminales que no tenga tanta actividad operacional, para realizar esta tarea.

No sería conveniente designar al servidor también como terminal para el análisis de datos ya que se pueden entorpecer las actividades operacionales respecto al tiempo de respuesta y rapidez en las transacciones.

Una vez seleccionada la máquina en la que estará operando el sistema de inteligencia empresarial, se podrá implementar todo el proceso de inteligencia empresarial.

4.1.5 Hipótesis sobre ventajas competitivas

El sistema que utiliza la empresa para control y administración de la misma (SAN), les ha ayudado a saber algunas preguntas básicas que todo sistema de este tipo proporciona. Por ejemplo: ¿Cuáles son los 10 productos más vendidos?

Mediante la observación y conocimiento del negocio, el gerente administrativo de la empresa, tiene algunas hipótesis con respecto a los productos que tienen mayor venta y ha formulado algunas estrategias que adopta como sus ventajas competitivas con respecto a su competencia.

Una de las hipótesis, es la siguiente:

> "Los 10 productos más vendidos, sirven como productos *gancho*, es decir, se venden a un precio muy accesible, para que los clientes identifiquen a la tienda, como la tienda que tiene mejores precios en esos productos y se compensa con el aumento de precio a otros productos que los clientes compran *de paso"*.

La ventaja competitiva consta en ofrecer mejores precios en los productos más vendidos. De esta manera, los clientes identifican al supermercado como uno de los que tienen mejores precios y comprarán de paso también otros productos aunque estos no tengan el mejor precio con respecto a la competencia.

Con la aplicación del sistema de inteligencia empresarial, se podrá corroborar la hipótesis que tiene el gerente como su ventaja competitiva, o por el contrario descartarla. Esto se llevará a cabo a partir del análisis de las compras de los clientes haciendo la distinción entre mayoristas y de menudeo.

4.2 Arquitectura del sistema

Dado que el objetivo principal de la tesis era desarrollar un prototipo que implemente las técnicas de inteligencia empresarial, que sea factible para una PyME se contempló que el prototipo cumpliera con los siguientes aspectos:

- Que sea un sistema que brinde una solución al área de CRM utilizando la técnica MBA.

- La integración de los datos debe realizarse de manera transparente para el usuario.

- El sistema debe realizar el procesamiento de la información en poco tiempo, es decir, que los resultados se entreguen en un tiempo considerable, directamente proporcional al tamaño de los resultados a entregar.

- Debe permitir la manipulación de la información desde diferentes perspectivas. La variedad de perspectivas le permite al usuario descubrir información más interesante que solo se encuentra con algunas combinaciones de parámetros.

- Debe ofrecer distintos tipos de reportes/consultas que muestren información relevante.

- Debe seguir el estándar de CRISP-DM, explicado en la sección 2.5.

- Los resultados que entregue deben ser comprensibles y fáciles de interpretar. Los resultados mostrados a través de gráficos facilitan al usuario la comprensión de los mismos. Por esta razón, el sistema debe poder ilustrar los resultados con gráficos que den al usuario una mejor idea de lo que está recibiendo.

- Sencillo y de fácil interacción, debido a que las PyMEs no cuentan con personal experto en el área de inteligencia empresarial, el sistema debe ser sencillo, intuitivo y fácil de utilizar para el usuario final, en este caso, el administrador o encargado de la PyME.

- Debe ser económico: como hemos mencionado anteriormente, las PyMEs cuentan con recursos limitados, tanto económicos como tecnológicos, por lo mismo el sistema debe ser accesible económicamente y debe ser capaz de trabajar con recursos tecnológicos limitados, sin que por estas dos razones se vea afectado el desempeño del mismo.

Una vez definidas las consideraciones y los requerimientos del sistema en seguida describiremos la solución que se propuso como para resolver la problemática de la inteligencia empresarial en las PyMEs.

La arquitectura de PIE está conformada por tres grandes capas, que son: integración, análisis y visualización. En la figura 4.4 se puede observar de manera gráfica la arquitectura diseñada para el sistema PIE.

En la ***capa de integración*** se extraen los datos de las bases de datos operacionales y se seleccionan los campos necesarios conforme al modelo de datos. Posteriormente los datos pasan por un proceso de ETL en donde se limpian y estandarizan, esto con el fin de eliminar inconsistencias y posibles errores que llegaran a existir. Después serán almacenados en estructuras (tablas) relacionales y de esta manera queda implementado el *datawarehouse*.

La ***capa de análisis*** comprende la aplicación de las técnicas de OLAP y los algoritmos de *data mining*.

Para OLAP, desde la capa de visualización, que se explicará posteriormente, el usuario ejecuta una consulta, la cual es formulada en lenguaje MDX. Posteriormente el motor de OLAP

se encarga de mapear las consultas en lenguaje MDX a sentencias SQL, que serán ejecutadas en la base de datos relacional donde reside el *datawarehouse*. La información resultante es regresada al motor de OLAP y éste se encarga de enviarla nuevamente a la capa de visualización.

Para *data mining*, se construyen archivos con los datos del *datawarehouse* necesarios para aplicar el algoritmo seleccionado, se eligió el algoritmo para reglas de asociación *Apriori*. El archivo contiene los datos que se deseen minar, en un formato específico, que se envía a la herramienta de *data mining* para ejecutar el algoritmo. Posteriormente, esta herramienta entregará los resultados a la capa de visualización.

Por último, la ***capa de visualización*** es la que permite mostrar al usuario final los resultados que se obtienen de la aplicación de las técnicas de OLAP y algoritmos de *data mining* de una manera que el usuario los pueda interpretar más fácilmente.

Los resultados pueden visualizarse a través de texto, tablas y gráficos. Esta variedad facilita la comprensión e interpretación de los mismos. De esta manera, el usuario puede interactuar y manipular la información de su interés para analizarla desde diferentes perspectivas. Dichas perspectivas le permitirán obtener información relevante que le ayudará a crear estrategias justificadas que traigan beneficios a la PyME.

Se definió una arquitectura de tres capas porque de esta manera cada procedimiento se encuentra bien definido e independiente de los demás.

Esta arquitectura permite cambiar herramientas en caso de que fuera necesario, aislar entradas y salidas bien definidas, buscando crear un flujo de la información, donde se puede localizar fácilmente cada etapa y buscar puntos de mejora.

Para la implementación del sistema PIE, se escogieron las herramientas: *Kettle* para la parte de integración de datos, *Mondrian* para la parte de OLAP, *Weka* para la aplicación de algoritmos de *data mining*, *JRubik* para la visualización de resultados, y el DBMS relacional MySQL para la implementación del *datawarehouse*, dichas herramientas que se explicarán con más detalle en el capítulo V.

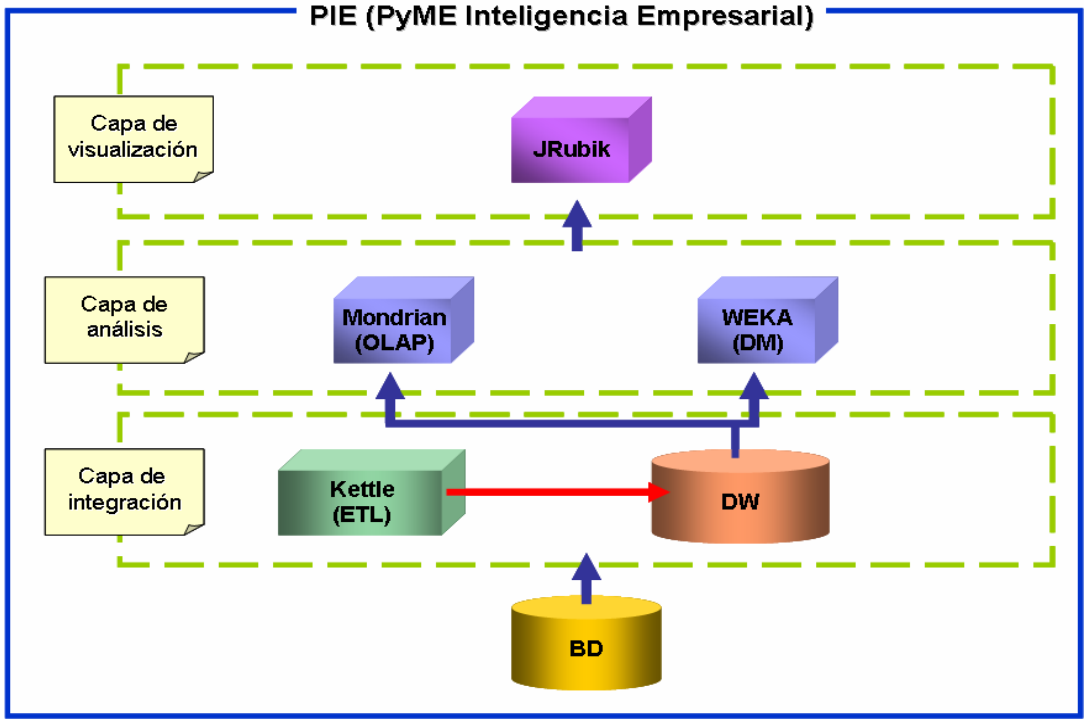

Figura 4.4 Arquitectura del sistema PIE

A pesar de que las tres herramientas forman parte de la suite de Pentaho, mencionada en el capítulo 3, se seleccionaron las 3 herramientas por separado, por las siguientes razones:

- No se usa la suite de Pentaho completa, debido a que las necesidades de las PyMEs no lo requieren y tampoco se necesita todo el soporte que brinda.

- Ya que la suite es una solución más general, maneja procedimientos, técnicas y funciones complejas que la PyME no necesita, ésta es la principal razón por la que los sistemas de inteligencia empresarial se vuelven muy complicados y difíciles de utilizar, y recordemos que una de las consideraciones del sistema es que fuera sencillo y de fácil interacción.

- Las necesidades y requerimientos de las PyMEs pueden satisfacerse con un sistema que esté a la medida de su estructura y de su economía.

- Con las herramientas no estamos atados a una marca o compañía, éstas nos permiten tener independencia para intercambiar o agregar componentes, siempre que se requiera.

Una vez que se definió la arquitectura del sistema, en las siguientes secciones se explicará cada componente más detalladamente.

4.3 Modelo de datos multidimensional

El primer componente del sistema PIE es el *datawarehouse*. Como se mencionó en el capítulo II para implementar un *datawarehouse* se requiere primero hacer el modelado de los datos que se van a almacenar.

De los 2 tipos de modelos de datos que existen, mencionados en el capítulo 2 (relacional y multidimensional), se eligió el modelo de datos multidimensional para la representación de los

datos de la PyME, porque se tiene una estructura general y homogénea de los datos, además de que nos da la ventaja de agregar información de otras fuentes de datos ya sean internas o externas a la PyME, en caso de que así se requiriera, por ejemplo: para analizar otro proceso de la empresa como las compras o el manejo de inventarios.

Otra razón por la que se escogió el modelo multidimensional, fue por su escalabilidad, es decir, se pueden agregar más hechos y dimensiones conforme se requieran en un futuro, para colocar información que se genere posteriormente y que sea importante para el apoyo a la toma de decisiones en la PyME. Los hechos agregados, pueden compartir dimensiones con los hechos existentes y formar así constelaciones, que se explicará más adelante en las conclusiones de la tesis [Laker, 2006].

A diferencia del modelo relacional, el modelo multidimensional nos permite analizar la información mediante cubos de OLAP, otra de las razones por las que preferimos este tipo de modelado, que resulta adecuado para el estudio de los datos de la PyME, además de las técnicas de reporteo general.

El modelo multidimensional también contempla el histórico de la información, el cual se va almacenando cada vez que hacen actualizaciones al *datawarehouse*. La información almacenada se encuentra agregada, lo que permite analizar grandes volúmenes de datos en espacios menores de almacenamiento. Esta característica facilita visualizar la información de manera gráfica, en reportes, mapas o gráficos, que a su vez simplifica la comprensión de los resultados que se entreguen.

Para la elaboración del modelo multidimensional de la PyME, se generó, del diagrama entidad-relación (ER) de los datos operacionales de la PyME, visto en el apéndice A, otro diagrama ER simplificado, que se muestra en la figura 4.5, el cual contiene una versión más definida de los datos necesarios que nos interesan analizar.

El diagrama muestra dos procesos importantes que se llevan a cabo en la PyME, ventas y compras, representados cada uno como una entidad. Las otras entidades que se muestran son: productos, clientes, empleados y proveedores con sus respectivas relaciones.

Este diagrama, sirvió como base para definir el modelo de datos multidimensional que fue usado para la implementación del *datawarehouse*, el cual explicaremos en seguida.

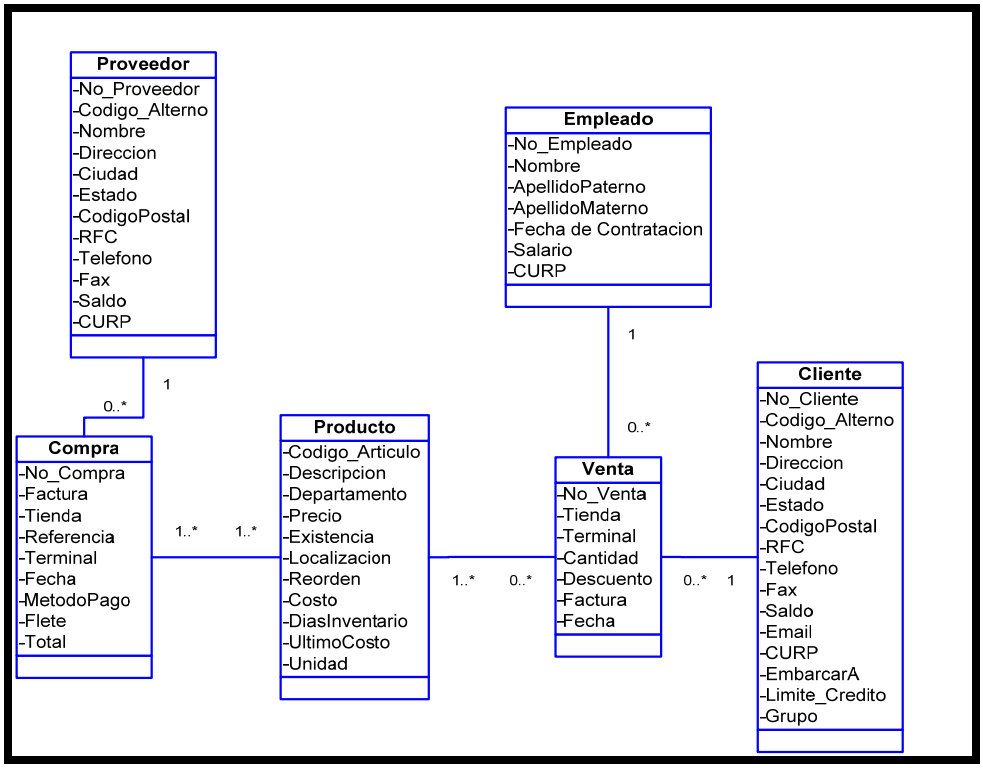

Figura 4.5 Diagrama entidad-relación simplificado de los datos de la PyME

El modelo de datos multidimensional definido para la implementación del *datawarehouse*, es un esquema estrella, con el cual analizaremos el proceso de las ventas de la PyME del caso de estudio.

Se eligió el proceso de negocios de ventas, porque el proceso de inteligencia empresarial está enfocado al área de CRM utilizando la técnica de MBA y las ventas es uno de los procesos de los que más información se puede obtener sobre los clientes. Cada transacción representa una parte del comportamiento habitual de compra de los clientes, y lo que se buscó fue modelar dicho comportamiento.

Para satisfacer las necesidades de los clientes es necesario conocer sus gustos y preferencias respecto a los productos que compran, esto con el fin de hacer clientes cautivos y apoyar la toma de decisiones que favorezcan la relación con los clientes.

En la figura 4.6 se muestra el modelo multidimensional normalizado, con cada una de las dimensiones desglosadas en jerarquías. Como podemos observar, en la parte central del diagrama se encuentra la tabla hecho *venta*, con sus respectivos atributos y llaves foráneas que representan la relación con las tablas de dimensión y las funciones de agregación o medidas que son: cantidad, precio, costo y costo unitario.

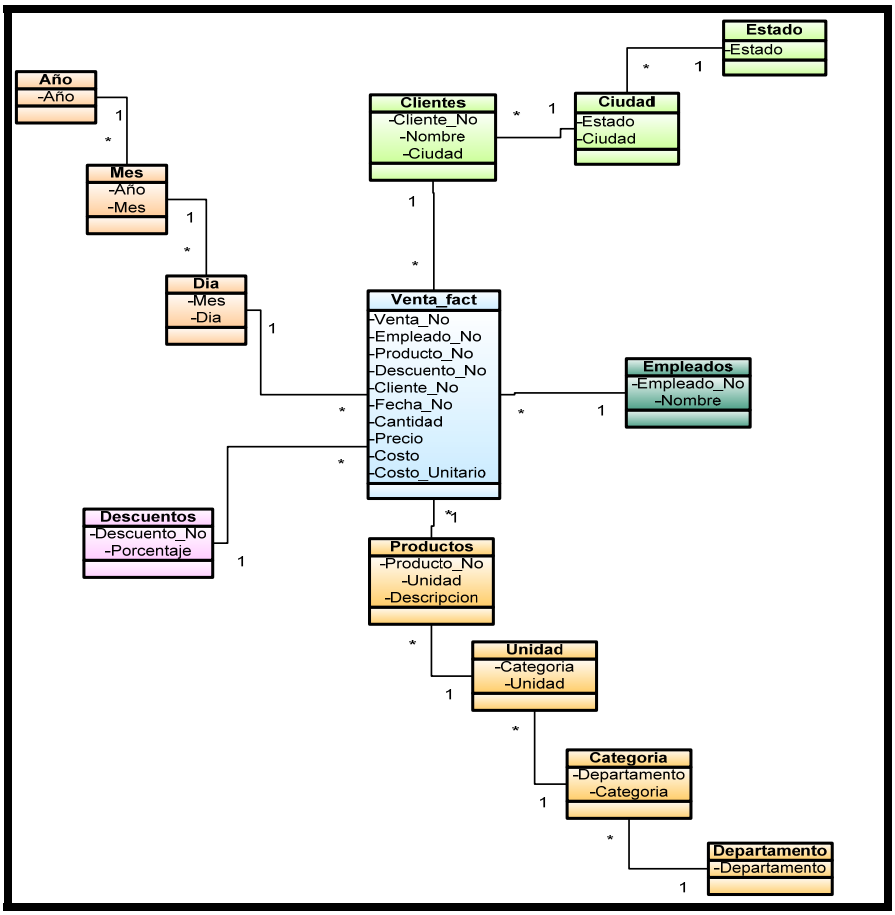

Figura 4.6 Modelo multidimensional normalizado

El modelo está representado en un esquema copo de nieve y como se mencionó en el capítulo II, existen más desventajas al tener un copo que tener un esquema estrella, ya que a pesar de que el esquema estrella ocupa mayor espacio de almacenamiento porque existe duplicación de datos, se prefiere tener un tiempo de respuesta más eficiente en las consultas multidimensionales, lo que disminuye la complejidad del modelo y por ende la de las consultas. Por ésta razón hemos definido un modelo multidimensional denormalizado, que forma una estrella más sencilla, la cual se muestra en la figura 4.7.

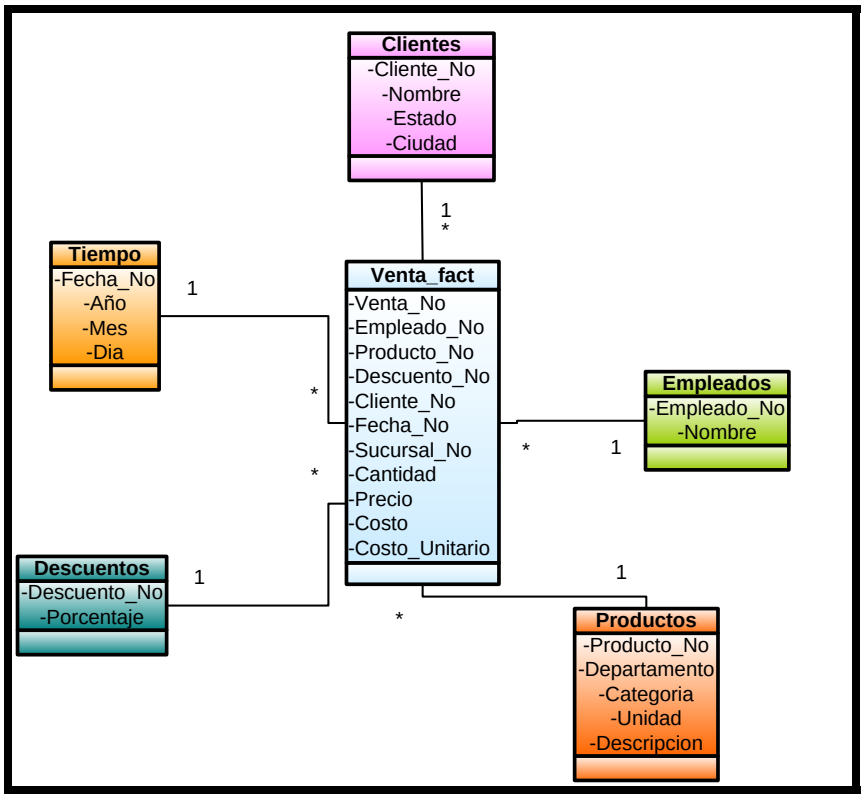

Figura 4.7 Modelo multidimensional denormalizado

Después de hacer un análisis comparativo entre todos los atributos disponibles y tener una entrevista previa con el administrador de la PyME, se obtuvieron como resultado las dimensiones definidas en el modelo, las cuales resultaron ser las más útiles para representar la información relevante en el proceso de las ventas de una PyME en general, no solamente para el caso de estudio. Los atributos de las dimensiones también fueron escogidos bajo el mismo criterio de selección.

La tabla 4.1 muestra los valores que pueden adquirir los atributos del modelo multidimensional.

Atributo		Valor
Clientes	**Cliente_No**	Cadena de 13 dígitos máximo, no admite decimales.
	Nombre	Cadena de caracteres, máximo 120, ya que podemos tener también nombres de empresas.
	Estado	Cadena de caracteres, máximo 20.
	Ciudad	Cadena de caracteres, máximo 40.
Empleados	**Empleado_No**	Cadena de 13 dígitos máximo, no admite decimales.
	Nombre	Cadena de caracteres, máximo 120, porque podemos tener múltiples nombres.
Productos	**Producto_No**	Cadena de 13 dígitos máximo, no admite decimales.
	Departamento	Cadena de caracteres, máximo 30.
	Categoría	Cadena de caracteres, máximo 30.
	Unidad	Cadena de caracteres, máximo 4.
	Descripción	Cadena de caracteres, máximo 255.
Descuentos	**Descuento_No**	Número entero.
	Porcentaje	Número flotante.
Tiempo	**Fecha_No**	Número entero.
	Año	Cadena de 4 caracteres.
	Mes	Cadena de 2 caracteres.
	Día	Cadena de 2 caracteres.
Venta_fact	**Venta_No**	Número entero.
	Cantidad	Número flotante.
	Precio	Número flotante.
	Costo_Unitario	Número flotante.
	Costo	Número flotante.

Tabla 4.1 Valores de los atributos del modelo multidimensional

Sobre el modelo multidimensional denormalizado que definimos, se construyó el cubo de análisis de OLAP y se aplicó el algoritmo de reglas de asociación de *data mining*. En las siguientes secciones veremos las consideraciones y definiciones de los procesos de ETL, OLAP y *data mining* que se emplearon para la implementación del sistema PIE.

4.4 Limpieza e integración de los datos

Los datos almacenados en las bases de datos operacionales no siempre se encuentran homogéneos y estandarizados, sobre todo cuando las bases provienen de distintas fuentes. Esto se debe a que los datos hayan sido ingresados por diferentes personas, que no se haya definido con anterioridad un estándar para la captura de los datos o a simples errores humanos.

Para realizar un buen análisis tanto de OLAP como de *data mining*, es necesario que la información almacenada en el *datawarehouse* se encuentre lo más homogénea posible. Para esto se requiere pasar los datos por un proceso que permita la integración de los datos.

Para integrar los datos de la PyME del caso de estudio tomaremos en cuenta las siguientes consideraciones:

- Definición de los nombres de las tablas hecho, dimensión y sus respectivos atributos, diferentes a los que tienen en la base de datos operacional para evitar confusiones, que serán usados para la implementación del *datawarehouse* y las transformaciones pertinentes para el transporte de los datos. Dichos nombres se establecieron como se muestran en la tabla 4.2.

Elemento	Nombre	Ejemplo
Tablas	Nombre de la tabla en plural y en minúsculas y sin acentos.	productos
Llaves	Nombre de la tabla en singular, primera letra en máyuscula, guión bajo y “No”.	Producto_No
Atributos	Primera letra en mayúscula las demás en minúscula, sin acentos.	Estado, Ciudad, Direccion
Dimensiones	Nombre de la tabla, guión bajo y “dimension”.	clientes_dimension
Hechos	Nombre de la tabla en singular, guión bajo y “fact”.	venta_fact

Tabla 4.2 Estándar de nombres definido para la implementación del datawarehouse

- Verificar las inconsistencias que existen en los datos, que pueden deberse a que por descuido se hayan eliminado registros o por pensar que ya no serían útiles en un futuro:

    - Clientes que aparecen en la tabla de ventas pero no en la tabla de clientes.

- Empleados que aparecen en la tabla de ventas pero no en la tabla de empleados.
- Productos que aparecen en la tabla de ventas, pero no en la tabla de productos.

- Debido al funcionamiento del sistema de administración, en las ventas se registran los productos tal y como son marcados por el cajero, esto con fines de optimización, por esta razón pueden existir repeticiones de los productos en cada transacción. Para solucionar esto, se requiere agrupar y sumar los productos repetidos en cada venta.

- Eliminar duplicaciones en valores que hayan sido escritos incorrectamente, o que signifiquen lo mismo pero hayan sido escritos de dos o más formas diferentes. Por ejemplo: los atributos de unidades, ciudades y estados.

Después de realizar estas transformaciones, los datos habrán quedado homogéneos, limpios y estandarizados, por lo que podremos proceder a definir las posibles consultas y reportes que el sistema PIE debe considerar.

4.5 Reportes necesarios

Para definir las consultas (reportes de OLAP) posibles que podemos aplicar sobre el modelo multidimensional, es necesario aplicar la fórmula $\boldsymbol{2^k}$ **= *número de grupos de consultas***, donde ***k*** es el número de dimensiones que tenemos en el modelo multidimensional. [ODDL, 2003] Aplicando la fórmula a nuestro modelo multidimensional tenemos como resultado 32 grupos de consultas y 6 niveles, los cuales se muestran en la figura 4.8.

Las consultas predeterminadas en el sistema PIE se encuentran entre los niveles 0 y 3, con el fin de satisfacer las preguntas como las que se muestran a continuación:

- ¿Cuál es el importe total de ventas en la fecha A del departamento B?
- ¿Cuál fue el importe total vendido del departamento A por el empleado B en la fecha C?
- ¿Cuál fue el importe total comprado del departamento A por el cliente B en la fecha C?
- ¿Cuántos productos A compró el cliente B en la fecha C?
- ¿Cuántos productos del departamento A se vendieron en la fecha B por el empleado C?
- ¿Cuántos productos del departamento A compraron los clientes de la región B?
- ¿Cuántos productos con descuento A fueron comprados por el cliente B en la fecha C?
- ¿Cuántos productos con descuento A fueron vendidos por el empleado B en la fecha C?

Se eligieron los niveles entre 0 y 3 para las consultas predeterminadas porque la visualización de datos es relativamente sencilla hasta 3 dimensiones, de 4 en adelante se complica la visualización e interpretación y se requiere de mayor interacción con el usuario, por esto lo reservamos para usuarios más experimentados.

Como vemos, el sistema PIE contempla 2 tipos de consultas, unas predeterminadas y otras donde el usuario experto tiene la posibilidad de manipular los atributos de la consulta para poder encontrar más información relevante en sus datos, o elaborar nuevas consultas incluyendo las de los niveles 4 y 5 en un área designada para un usuario más experimentado.

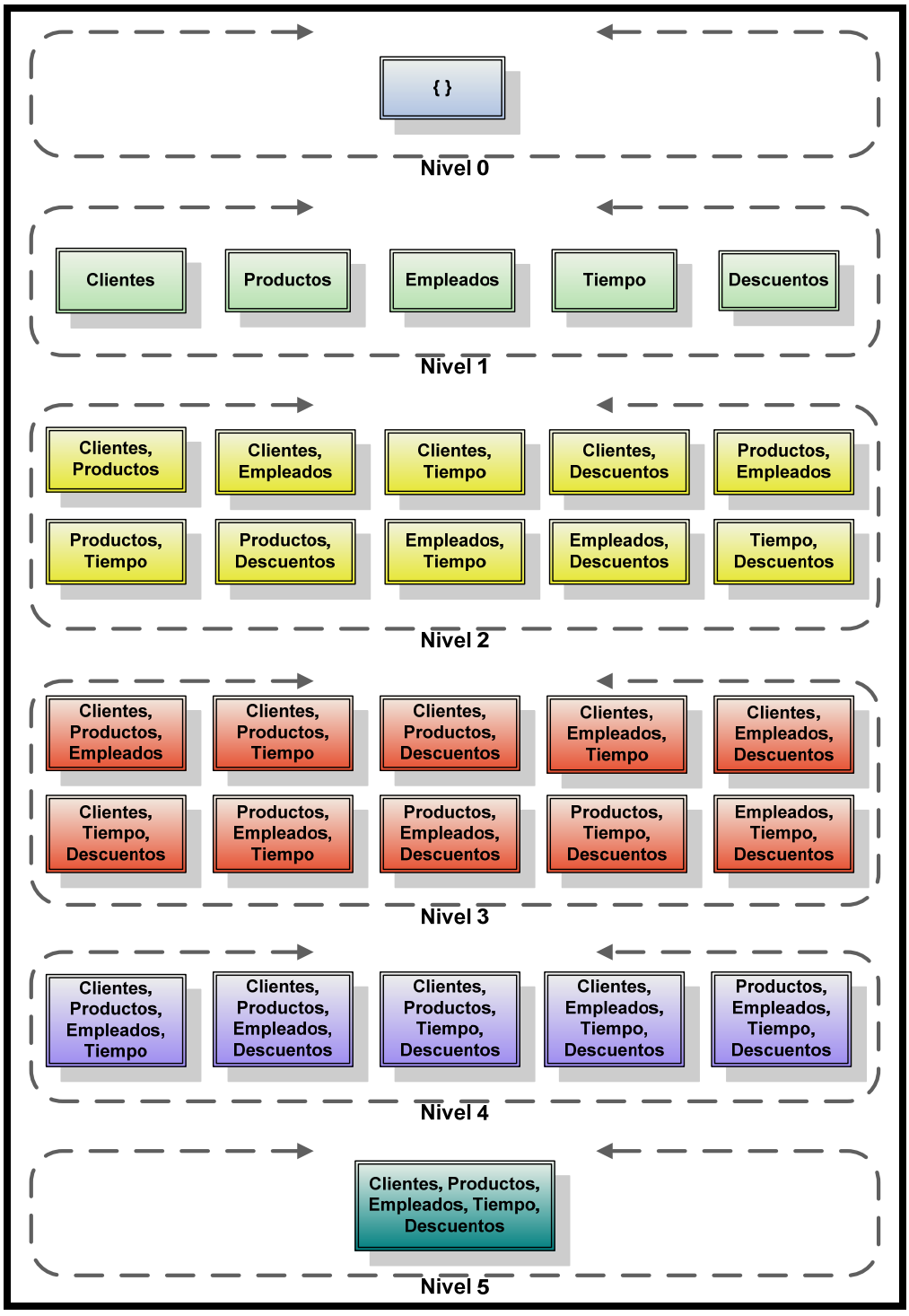

Figura 4.8 Grupos de consultas

Una vez determinadas las consultas para el procesamiento de OLAP, en la siguiente sección definiremos las consideraciones necesarias para el proceso de *data mining*.

4.6 Estructuras y parámetros para los algoritmos de data mining

Para modelar el comportamiento de compra del cliente se requiere describir los productos que suele comprar en una misma transacción o en varias con base en un historial de compras de los clientes. Este problema, desde el punto de vista de *data mining,* se considera como una tarea descriptiva ya que, como se mencionó en la sección 2.3, se requiere ***describir*** y no predecir el comportamiento de compra del cliente.

Dentro de la posibles categorías de tareas descriptivas, el problema corresponde a reglas de asociación ya que lo que buscamos son las asociaciones entre los productos que se compran juntos y esta sería la representación más adecuada, podría pensarse que también sería una tarea de clustering sin embargo, los clusters definen clases, lo que se ajustaría más a definir clases o grupos de clientes y no de productos, esto se define más adelante en el capítulo VI como trabajo a futuro.

Para la obtención de las reglas elegimos el algoritmo *Apriori*, que está basado en los conteos de frecuencias en las que dos o más sucesos se presentan conjuntamente, que es lo que buscamos para localizar los productos que se compran juntos en una misma venta y otros datos interesantes.

Existen 3 tipos de algoritmos *A priori* que son:

- *Apriori*: utilizando un soporte del 100%, el cual va disminuyendo buscando que exista el mayor porcentaje de confianza. Es decir el soporte máximo se diminuye hasta alcanzar el soporte mínimo. Tiene una visión *top-down*.

- *Predictive Apriori:* combina el soporte y la confianza en una sola medida y va aumentando el umbral del soporte. Es decir, el soporte mínimo se aumenta hasta alcanzar el soporte máximo. Tienen una visión *bottom-up.*

- *Tertius:* busca reglas de acuerdo a una medida de confirmación pero difiere de los 2 métodos anteriores, porque utiliza condiciones de OR en lugar de AND.

Se seleccionó el algoritmo *Apriori* porque el *Predictive Apriori* requiere mayor tiempo y recursos procesamiento al combinar las dos medidas de soporte y confianza y porque el *Tertius* tiene una precisión más relajada al utilizar las condiciones de OR, lo que nos entregaría reglas con múltiples productos en la premisa y por consiguiente tendríamos un análisis con menor precisión.

El algoritmo *Apriori* maneja una serie de parámetros que se deben configurar para que estemos seguros de que las reglas resultantes son confiables. Estos parámetros se configuran tomando en cuenta el tipo y cantidad de datos que estamos manejando.

Para poder determinar los parámetros correctos para obtener reglas interesantes y confiables, se realizó un análisis previo a los registros de las ventas del caso de estudio, del mes de enero 2007. De este análisis se obtuvo:

- 13169 productos diferentes.

- 625 clientes entre mayoreo (se considera mayoreo la ventas mayores o iguales a 30 productos) y menudeo.
- 13 empleados.
- 10778 transacciones de ventas del mes de enero, incluidas las de clientes de mayoreo y menudeo.
- Un promedio de 3 artículos por venta, por lo que no se pueden esperar reglas de más de 3 productos en la premisa.
- 9045 ventas del mes de enero de menudeo.
- 1733 ventas del mes de enero de mayoreo.

Los parámetros más importantes que se deben definir en el algoritmo *Apriori* son el soporte y la confianza. Recordemos que el soporte se define como el número de instancias encontradas de entre todo el conjunto de reglas, mientras que la confianza mide el porcentaje de las veces que la regla se cumple satisfactoriamente entre aquellas que tienen el soporte establecido.

Con los datos mencionados anteriormente se definieron los parámetros del algoritmo como se muestran en la tabla 4.3.

Parámetro	Valor	Significado
Delta	0.05	Iterativamente se decrementa el soporte mínimo por este parámetro.
Soporte mínimo	0.0050	Parámetro de soporte mínimo al que se llega con los decrementos.
Soporte máximo	0.5	Parámetro de soporte máximo del que se parte para empezar el decremento.
Confianza mínima	0.25	Parámetro de confianza mínima que deben tener las reglas resultantes.
Métrica	Confidence	Proporción de ejemplos cubiertos por la premisa de la regla que también son cubiertos por la consecuencia.
No. de reglas	100	Número de reglas que entregará el algoritmo.

Tabla 4.3 Parámetros del algoritmo Apriori

El soporte mínimo 0.0050 se obtuvo de dividir el porcentaje de instancias en las que queremos que se aplique la regla entre el número de datos que tenemos, 53 instancias de los datos reales cumplen con el soporte mínimo y el soporte máximo se seleccionó de 0.5 o 50%, 5389 instancias de los datos reales cumplen con el soporte máximo.

Se eligieron estas medidas, para optimizar el proceso, mejorar el tiempo de respuesta, la memoria ocupada y porque al poner un soporte del 100% significaría que los clientes han comprado todos los productos de la tienda, lo cual no es poco probable. Al poner un soporte del 50% estamos asumiendo que los clientes han comprado la mitad de los productos existentes, algo que resulta más razonable.

La confianza mínima se definió de 0.25 o 25% para obtener reglas que no se descubren a simple vista y así poder descubrir otros datos interesantes.

El número de reglas se definió en 100, ya que es el número máximo de reglas que se pueden obtener y la métrica seleccionada para clasificar las reglas fue *confidence* para poder determinar el porcentaje de calidad de las reglas.

Los parámetros mencionados anteriormente son importantes para el diseño del sistema, puesto que son adecuados para PyMEs, de otra manera por ejemplo: el poner un soporte mínimo muy alto no arrojaría ningún resultado.

Una vez definidos los parámetros del algoritmo, describiremos la estructura que se definió para los archivos con los datos a minar.

Se determinaron 4 tipos de archivos diferentes: clientes_menudeo, clientes_mayoreo, ventas_menudeo y venta_mayoreo.

- Para los archivos de ventas, la estructura es una representación binaria de los atributos como se muestra en la tabla 4.4, que contiene en las filas las ventas realizadas y en las columnas los productos existentes, este archivo nos permite descubrir los productos que se compran juntos en una misma transacción.

Venta/Producto	Aceite Maravilla	Azúcar	Cerveza Superior	Papas Sabritas	Coca-Cola	Leche Alpura	Pañales Pampers
V1	1	1	0	0	0	0	0
V2	0	0	0	1	1	0	0
V3	0	0	1	1	1	0	0
V4	0	0	1	1	0	1	1
V5	1	1	0	0	0	0	0
V6	0	0	1	1	0	0	1
V7	0	0	0	0	0	1	1
V8	1	1	0	0	1	0	0
V9	0	0	0	0	0	1	0
V10	1	1	1	1	1	0	0

Tabla 4.4 Canasta de compras con datos del caso de estudio

- Para los archivos de los clientes, la estructura también es una representación binaria como se muestra en la tabla 4.5, que contiene en las filas a los clientes y en las columnas a los productos, este tipo de archivo nos permite descubrir que artículos prefiere un cliente aunque no hayan sido comprados en la misma venta.

Cliente/Producto	Aceite Maravilla	Azúcar	Cerveza Superior	Papas Sabritas	Coca-Cola	Leche Alpura	Pañales Pampers
C1	**1**	**1**	**0**	**1**	**1**	**0**	**0**
C2	**0**	**0**	**1**	**1**	**1**	**0**	**0**
C3	**0**	**1**	**0**	**0**	**0**	**1**	**0**
C4	**1**	**1**	**1**	**1**	**0**	**0**	**0**
C5	**1**	**1**	**0**	**0**	**0**	**1**	**1**

Tabla 4.5 Compras por cliente con datos del caso de estudio

4.7 Discusión final

En este capítulo se describió el estudio que se realizó de los procesos de negocio que se llevan a cabo en el supermercado, sus necesidades tecnológicas y las implementaciones que se deben llevar a cabo para aplicar el proceso de inteligencia empresarial, así como el análisis de los datos de la PyME, y la selección de los atributos necesarios para diseñar el modelo de datos.

También hemos definido todas la consideraciones que se tomaron en cuenta para la implementación del sistema PIE en cada uno de las capas de actividades: ETL, OLAP y *data mining*. Se definió la arquitectura del sistema, el modelo de datos, los grupos de consultas de OLAP, la estructura de los archivos para *data mining* y los parámetros para el algoritmo *Apriori*.

Todas estas actividades descritas a lo largo del capítulo corresponden a las 4 primeras fases del estándar de CRISP-DM: comprensión del negocio, comprensión de los datos, preparación de los datos y modelado. Restan por cumplir 2 fases que veremos en el siguiente capítulo junto con la implementación del sistema y más detalles técnicos, así como la justificación de las herramientas seleccionadas.