SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Subsecretaría de Educación Superior
Dirección General de Educación Superior Tecnológica
Instituto Tecnológico de la Laguna

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

LA CREACIÓN DE ALMACENES DE DATOS Y LA INTELIGENCIA DE NEGOCIOS

POR

Eduardo Cuan Durón

TESIS

PRESENTADA COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE MAESTRO EN SISTEMAS COMPUTACIONALES

DIRECTOR DE TESIS
M.C. JOSE D. RUÍZ AYALA

CO-DIRECTORA DE TESIS
DRA. ELISA URQUIZO BARRAZA

ISSN: 0188-9060

RIITEC: (01)-TMSC-2013

Torreón, Coahuila. México,
Marzo 2013

Blvd. Revolución y Cda. Cuauhtémoc s/n Col. Centro C.P. 27000
Torreón, Coah. Tel. 7053113, correo: webmaster@lalaguna.edu.mx
www.lalaguna.edu.mx

Torreón, Coah., 12 / Marzo /2013

DR. JOSE LUIS MEZA MEDINA
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
PRESENTE

Por medio de la presente, hacemos de su conocimiento que después de haber sometido a revisión el trabajo de tesis titulado:

" La Creación de Almacenes de Datos y la Inteligencia de Negocios"

Desarrollado por el C. **EDUARDO CUAN DURÓN**, y habiendo cumplido con todas las correcciones que se le indicaron, estamos de acuerdo que se le conceda la autorización de la fecha de examen de grado para que proceda a la impresión de la misma.

ATENTAMENTE

M.C. JOSÉ D. RUIZ AYALA
PRESIDENTE

DR. ENRIQUE CUAN DURÓN
VOCAL

DRA. ELISA URQUIZO BARRAZA
SECRETARIO

DR. DIEGO URIBE AGUNDIS
VOCAL SUPLENTE

Dependencia: **DEPI**
Oficio: **DEPI/189b/2013**
Asunto: **Autorización de impresión**

Torreón, Coah., 12/Marzo/2013

C. EDUARDO CUAN DURÓN
CANDIDATO AL GRADO DE MAESTRO EN SISTEMAS COMPUTACIONALES
PRESENTE

Después de haber sometido a revisión su trabajo de tesis titulado:

"La Creación de Almacenes de Datos y la Inteligencia de Negocios"

Habiendo cumplido con todas las indicaciones que el jurado revisor de tesis hizo, se le comunica que se le concede la autorización con número de registro **RIITEC: 01-TMSC-2013**, para que proceda a la impresión del mismo.

ATENTAMENTE

SECRETARÍA DE EDUCACIÓN PÚBLICA
INSTITUTO TECNOLÓGICO de la Laguna
División de Estudios de Posgrado e Investigación

DR. JOSÉ LUIS MEZA MEDINA
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

ÍNDICE

ÍNDICE DE FIGURAS

Capítulo 5 – Diseño e Implementación de Modelos de Minería de Datos

SINOPSIS

En el primer capítulo se explican de forma compenciada los sistemas OLPT (On-Line Transactional Processing / Procesamiento transaccional en línea) y OLAP (On-Line Analytical Processing / Procesamiento analítico en línea), con la finalidad de proveer conceptos elementales, inherentemente asociados a la inteligencia de negocios, que clarifiquen el panorama correspondiente a esta investigación, y asimismo, faciliten la comprensión de los capítulos posteriores. Se Incluyen algunos antecedentes históricos y el análisis de sus atributos principales, funcionalidades y aplicaciones, subrayando las más ostensibles diferencias entre estas dos clases de sistemas.

En el segundo capítulo, se proporcionan conceptos y esquemas cuya finalidad es explicar la metodología estándar aplicada en el modelamiento multidimensional. Son tratados con minucia los aspectos más significativos acerca de la orientación analítica y el procesamiento de los datos relacionados con la tecnología Datawarehousing.

En el tercer capítulo se aborda la documentación de un ejemplo proporcionado por Microsoft a través de un tutorial de su página web. Los objetivos de este capítulo se focalizan en la instalación de la base de datos *AdventureWorks 2008*, la creación de un nuevo proyecto, la definición de un origen de datos y la creación y visualización de un informe desde el *ReportViewer* basado en los datos de la base precitada.

En el cuarto capítulo se aborda la documentación textual y gráfica de algunas lecciones pertenecientes al tutorial *SQL Server Analysis Services*, proporcionado por Microsoft a través de su página web., Con el fin de destacar los aspectos fundamentales concernientes a la creación, implementación y operaciones básicas de un cubo OLAP.

La finalidad del quinto capítulo, es ejemplificar, a través de la puesta en práctica de las lecciones de un tutorial de Microsoft, acerca de la minería de

datos, el uso de los procedimientos más significativos, con fines de iniciación, de este campo de las ciencias de la computación. Como puede deducirse, la minería de datos desempeña un rol principal dentro de la inteligencia de negocios, pues su facultad predictiva la convierte en un instrumento de asesoría muy valioso para la toma de decisiones empresariales.

RESUMEN

El propósito de este proyecto está orientado hacia quienes tienen la intención o la necesidad de iniciarse en el desarrollo de aplicaciones de Inteligencia de Negocios, principalmente, aunque no privativamente, a quienes integran las áreas de desarrollo de software dentro de las organizaciones empresariales.

Para conseguir tal propósito, por principio de cuentas, fue necesaria la determinación de los puntos específicos a tratar, que a la consideración de asesor y del tesista, se juzgaron fundamentales para el desarrollo de una exposición elemental de la inteligencia de negocios, clara y progresivamente organizada en función de su complejidad.

Los dos primeros capítulos constituyen el sustento teórico del proyecto. En ellos fueron considerados los conceptos, las definiciones y las explicaciones funcionales de los sistemas OLTP y OLAP, así como los aspectos que conforman el eje principal del modelo multidimensional.

Los capítulos restantes fueron desarrollados a manera de una guía de ejercicios prácticos que persiguió tres objetivos primordiales, el primero de ellos fue la familiarización con el uso de algunas herramientas que permiten el acceso a los datos y la disposición de los mismos. El segundo, la conversión de esos mismos datos en información posible de ser visualizada a través de pantallas o reportes, y finalmente, la asimilación de esa información como conocimiento, utilizado como base para la realización de cálculos predictivos capaces de orientar acertadamente durante la toma de decisiones empresariales.

ABSTRACT

The purpose of this project is geared towards those who intend or need to start developing Business intelligence applications, mainly but not exclusively to those who make up the software development areas within business organizations.

In order to achieve this purpose, was first necessary to determine the topics of discussion, that were judged by the advisor and thesis student as fundamental to the development of an exhibition elementary business intelligence clear and gradually depending on its complexity.

The first two chapters provide the theoretical basis of the project. They were considered the concepts, definitions and functional explanations of OLTP and OLAP systems, as well as the aspects that make up the backbone of the multidimensional model.

The following chapters were developed as a guide to practical exercises with three primary objectives, the first objective was familiarization with the use of some tools that allow access and disposal of data. The second objective was the conversion of data into information that was possible to view it on screens or reports. And finally the third objective was the assimilation of such information, resulting in knowledge used as the basis for performing predictive calculations, which are able to accurately guide for making business decisions.

INTRODUCCIÓN

Definición del problema

Es más común de lo que se piensa que dentro de las PYMES y organizaciones en general se presente el problema de la inconsistencia en sus sistemas de información. El manejo de un mismo tipo de información a cargo de distintos responsables y procesada desde diversas aplicaciones, a falta de un Sistema de Información Integrado (ERP Enhanced Resource Planning), tiene como consecuencias el retraso en la entrega de la información (pues comúnmente, antes de presentarla, hay que realizar algún proceso especial que la reúna y valide) y la pérdida de fiabilidad de la misma (debido a que en muchas ocasiones no es posible conciliarla completamente), por ende, los directivos desestiman dicha información como para considerarla la piedra angular sobre la que estribarán sus decisiones.

Este problema, aunado a la falta de tiempo, de estrategias y de herramientas adecuadas para analizar la información, son justamente la clase de problemas que la inteligencia de negocios resuelve, pero para su implementación, es preciso contar antes con un consolidado Sistema de Información Integrado (ERP), además de transformar la noción limitada que comúnmente suelen tener los directivos, a través de la presentación de un equilibrado análisis costo-beneficio, concerniente a la adquisición de las herramientas y tecnologías que soportan a los sistemas BI (Business Intelligence).

Justificación

El objeto de esta investigación es poner de relieve la importancia fundamental que tiene la aplicación del conjunto de estrategias y herramientas BI (Business Intelligence) para la toma de decisiones en los negocios. Para ello es menester modificar el concepto que comúnmente tienen los directivos de las organizaciones empresariales acerca de los

Antecedentes

Una sugerencia de mi asesor de proyecto despertó mi interés por este tema. Luego de leer las primeras páginas del libro: "Business Inteligence: Competir con información" de Josep Lluis Cano, me resolví a elegir este tema, con el consentimiento y apoyo de mi asesor, para dar inicio a mi investigación.

Mi experiencia personal dentro de las empresas me ha permitido comprobar en la práctica la veracidad implícita en la paradoja mencionada en el precitado libro: "La sobreabundancia de información y la falta de tiempo para su análisis".

El desarrollo de aplicaciones de software a la medida o la adquisición de software estándar, han sido bien acogidos por las organizaciones empresariales en general. No es punto de discusión la cantidad de beneficios que les han redituado en materia administrativa y operativa. No obstante, las aplicaciones comunes no suelen ir más allá de una ingente acumulación de datos que dan constancia de las acciones producidas día con día, pero que, debido a la estructura y distribución de su almacenamiento, resulta imposible analizarla en su conjunto con la finalidad de ilustrar cabalmente el panorama del negocio.

Hoy en día, las decisiones tomadas por los directivos, encaminadas a la calidad y competitividad empresarial, deben basarse, en buena medida, en los reportes de información generados por sus sistemas. Pero esto se complica hasta la imposibilidad si no se dispone de estrategias y herramientas como las ofrecidas por la inteligencia de negocios (Business Intelligence) que simplifican todos los procesos, desde la carga de datos hasta su análisis, dado su mayor número de funcionalidades, mayor rendimiento, granularidad en distintos niveles de detalle y agregación, visión multidimensional y facilidad de uso en general.

De ahí la importancia de estudiar lo más ampliamente posible este campo de la gestión empresarial, con la finalidad de destacar entre las PYMES y organizaciones en general, la enorme cuantía de contar con aplicaciones y tecnologías de esta clase.

sistemas de información. La inteligencia de negocios no se trata de una abundante acumulación de datos dispersos. Gracias a ella es posible la cohesión de dichos datos, su depuración y la conversión de los mismos, primero, en información fiable y posteriormente en conocimiento. El pronto y simple acceso a este conocimiento permite hallar su utilidad inmediata, así como constituir un sólido soporte para el desarrollo de futuros análisis y garantizar la acertada toma de decisiones. Se pretende que los beneficiarios de esta investigación sean aquéllos quienes integran las áreas de desarrollo de software dentro de las organizaciones empresariales. Que puedan servirse libremente de la información que compondrá este trabajo académico y hallen en él valor y utilidad para el desarrollo de sus aplicaciones particulares.

Objetivo

El objetivo de esta investigación es la documentación teórica, así como la inclusión de una serie de ejercicios prácticos que hagan posible la comprensión básica de los conceptos fundamentales de la inteligencia de negocios. Para el cumplimiento de este objetivo, se cuenta con el debido apoyo bibliográfico, de artículos y de referencias a sitios de internet, así como la puesta en práctica de algunos tutoriales proporcionados por Microsoft, relacionados a este tema.

Viabilidad

Viabilidad Técnica

La viabilidad técnica de este proyecto es total. El Instituto Tecnológico de la Laguna cuenta con el software necesario para el desarrollo del mismo (Visual Studio 2010 y SQL Server 2008). Asimismo, el material bibliográfico principal es asequible a través del instituto.

Viabilidad Económica

La viabilidad económica del proyecto es total. El Instituto Tecnológico de la Laguna cuenta con las licencias del software a utilizar, además de ofrecer el uso gratuito de equipo de cómputo y de material bibliográfico.

Viabilidad de Operación

El desarrollo de esta investigación es garantizado, dado que el Instituto Tecnológico de la Laguna cuenta con las instalaciones, el equipo, el material bibliográfico y el software necesario para el mismo.

Viabilidad de Programación

El tiempo para el desarrollo de este proyecto es determinado por la duración del estudio de la maestría (dos años).

Límites

Límites contextuales

El objetivo de este proyecto se orienta principalmente hacia aquéllos que integran las áreas de desarrollo de software dentro de las organizaciones empresariales y que tienen el propósito o la necesidad de iniciarse en el desarrollo de aplicaciones de Inteligencia de Negocios (Business Intelligence).

Límites conceptuales

Los conceptos más comunes a utilizar durante el proyecto son:

- Business Intelligence. También podrá ser referido como inteligencia de negocios, inteligencia empresarial, o simplemente BI. Se trata de un conjunto de herramientas y tecnologías que utilizan la inteligencia orientada a la

optimización de la toma de decisiones de carácter empresarial.

- OLAP. Las aplicaciones OLAP (On-Line Analytical Processing / Procesamiento analítico en linea) son la herramienta principal de los sistemas especializados en la toma de decisiones.
- MOLAP. (Procesamiento multidimensional OLAP).
- Data Warehouse (Almacén de datos).
- DataMart (versión especial de un Data Warehouse).
- Sistemas de Soporte a la Decisión (DSS).
- Sistemas de Información Ejecutiva (EIS).
- Cubos de datos o Cubos OLAP.
- Minería de Datos.

Límites Espaciales

El proyecto se desarrollará dentro de las instalaciones del Instituto Tecnológico de la Laguna.

Límites Temporales

El tiempo para el desarrollo de este proyecto es determinado por la duración del estudio de la maestría (dos años).

Hipótesis

La realización de una investigación que incluya el desarrollo de algunas aplicaciones prácticas de Inteligencia de Negocios (Business Intelligence) es suficiente para demostrar los alcances que esta tecnología ofrece en materia de análisis de negocios y toma de decisiones.

Fundamentos

La Inteligencia de Negocios (Business Intelligence) puede definirse como un conjunto de metodologías, aplicaciones y tecnologías que permiten conjuntar, depurar y transformar los datos de los sistemas transaccionales en información, para su uso inmediato o para su posterior análisis y conversión en conocimiento, constituyéndose así en el sustento fundamental para la toma de decisiones relacionadas a los negocios.

Los sistemas y componentes de Inteligencia de Negocios (Business Intelligence) permiten realizar consultas de alto rendimiento, aprovechando el almacenamiento desnormalizado de los datos dentro del Almacén de Datos (Data Warehouse), a diferencia de los sistemas transaccionales en los cuales los datos suelen encontrarse normalizados para apoyar operaciones continuas de inserción, modificación y borrado.

La Inteligencia de Negocios (Business Intelligence) proporciona información privilegiada para dar solución a los problemas típicos del negocio: entrada a nuevos mercados, optimización del control financiero, reducción de costos, planificación de la producción, análisis de perfiles de clientes, rentabilidad de un producto en particular, etc...

Los principales productos de la Inteligencia de Negocios (Business Intelligence) son:

- Cuadros de Mando Integrales (CMI)
- Sistemas de Soporte a la Decisión (DSS)
- Sistemas de Información Ejecutiva (EIS)

Los principales componentes para el almacenamiento de datos en la inteligencia de Negocios (Business Intelligence) son:

- Datamart
- Data Warehouse

Método

En este proyecto se desarrollaron los siguientes puntos:

- Documentación teórica de sistemas OLTP.
- Documentación teórica de sistemas OLAP.
- Documentación teórica del concepto y características de un Data Warehouse.
- Documentación teórica del Modelo Multidimensional.
- Puesta en práctica de un tutorial de Microsoft que ilustra el acceso a los datos de Adventure Works y los muestra en un reporte.
- Puesta en práctica de un tutorial de Microsoft en el cual se hace el modelamiento de un cubo OLAP y se accede a los datos del mismo
- Puesta en práctica de un tutorial de Microsoft en el cual se diseña una estructura de minería de datos, seleccionando tres tipos de modelos algoritmicos para el análisis de los datos, y, finalmente, realizando un cálculo predictivo de clientes potenciales a considerar para la emisión de correos publicitarios.

Equipo y Software a utilizar

- Computadora de escritorio
- Laptop
- Impresora
- SQL Server 2008
- Visual Estudio 2010
- Adventure Works RS4
- Office 2010

CAPÍTULO 1

Sistemas OLTP y OLAP

Introducción

En este capitulo se explican de forma compendiada los sistemas OLPT (On-Line Transactional Processing / Procesamiento transaccional en linea) y OLAP (On-Line Analytical Processing / Procesamiento analítico en linea), con la finalidad de proveer conceptos elementales, inherentemente asociados a la inteligencia de negocios, que clarifiquen el panorama correspondiente a esta investigación, y asimismo faciliten la comprensión de los capitulos posteriores. Se Incluyen algunos antecedentes históricos y el análisis de sus atributos principales, funcionalidades y aplicaciones, subrayando las más ostensibles diferencias entre estas dos clases de sistemas.

1.1. Antecedentes Históricos.

1.1.1 Sistemas de Soporte a la Decisión

Los Sistemas de Soporte a la Decisión (DSS por sus siglas en inglés) son una clase especifica de sistemas de información que tienen como finalidad asistir en las actividades correspondientes a la toma de decisiones dentro de una organización.

Un DSS diseñado apropiadamente es un sistema de software interactivo, orientado a ayudar al personal clave dentro de la organización a obtener información útil a partir de datos fuente. Esta información obtenida permite identificar problemas dentro de la organización y proporciona una herramienta robusta para poder resolverlos, efectuando una correcta toma de decisiones.

1.1.2 Síntesis Histórica

El proceso de toma de decisiones dentro de las organizaciones ha tenido significativos cambios en las últimas décadas, desde la aparición del concepto de Sistemas de Soporte a la Decisión (DSS, por sus siglas en

inglés) durante los años 60's, exclusivamente en el ámbito académico. Este concepto continuó su desarrollo hasta los años 80's donde ya se involucró el diseño de los sistemas de información para ejecutivos (EIS, por sus siglas en inglés) así como los sistemas de soporte a la decisión organizacionales (ODSS, por sus siglas en inglés). En la década de los 90's, las tecnologías de procesamiento analítico en línea (OLAP) y Data Warehousing son ya componentes fundamentales de los DSS, además de algunas técnicas basadas en aplicaciones web.

1.1.3 Utilidad de los DSS

En la actualidad el uso de los DSS se ha extendido debido a su capacidad de analizar grandes volúmenes de datos y a la forma clara y sintética de presentar la información.

La información que típicamente puede recopilar y mostrar una aplicación de soporte a la decisión incluye todos los datos almacenados en la empresa, que van desde sistemas heredados, hasta bases de datos relacionales, Data Warehouse o cubos de datos.

La información es presentada en esquemas gráficos, de tal manera que sean de fácil comprensión aun para los usuarios que no están muy familiarizados con sistemas computacionales. Una de las funciones más importantes de este tipo de sistemas es la proyección a futuro del comportamiento de algunos factores de negocio, esta proyección está basada en un modelo de sistema que ha sido diseñado de acuerdo con los expertos del área en que se está haciendo la aplicación.

1.1.4 Evolución de los DSS

La tecnología de la información se halla inmersa en cambios continuos y frecuentes, y los DSS no son la excepción. A pesar de su excelente funcionalidad, como ocurre con todos los sistemas, siempre hay optimizaciones por realizar, errores por eliminar o por resolver. Uno de los principales problemas que presentan en la actualidad los DSS está en su

diseño, ya que estos requieren de una considerable experiencia sobre cuestiones estadísticas y de un complejo análisis humano para optimizar sus tiempos de operación. Ahora una nueva etapa de sistemas de información que soportan la toma de decisiones está siendo ya desarrollada y aplicada en algunas áreas de negocios y es conocida como Toma de Decisiones Automatizada (en una traducción literaria de "Automated Decision Making"), considerada como la evolución de los Sistemas de Soporte a la Decisión y de la Inteligencia Artificial (AI) tomando las mejores características de ambos para crear el nuevo concepto.

Aunque aún se halla en desarrollo, esta tecnología se aplica ya en áreas como la prescripción médica, servicios de viajes, control del transporte y especialmente en el área financiera dentro de empresas bancarias y de seguros. Algunas de las razones para llevar a cabo este proceso evolutivo es que ahora será más fácil que antes crear y administrar las aplicaciones. Los sistemas de decisión automatizada tendrán la capacidad de detectar datos en línea, aplicar lógica o conocimiento codificado y tomar decisiones, todo esto con una mínima participación del elemento humano.

De este modo, muchas decisiones serán tomadas en forma automática por los sistemas, y aunque pareciera que sólo son buenas noticias para las empresas, ahora el personal gerencial tendrá que afrontar este nuevo reto de la tecnología adaptándose a ella y compartiendo la responsabilidad de ciertos tipos de decisiones. Igualmente las diferentes economías, principalmente en los países desarrollados, tendrán que resolver algunos aspectos derivados de la pérdida de empleos como consecuencia de la entrada de tecnologías como ésta [1].

1.2 Sistemas OLTP

Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico que puede involucrar operaciones de inserción, modificación y borrado de datos

(Gestión transaccional). El proceso transaccional es típico de las bases de datos operacionales. Su tipo de arquitectura es la de cliente-servidor (Figura 1.1). La tecnología OLTP se utiliza en innumerables aplicaciones, como en banca electrónica, procesamiento de pedidos, comercio electrónico, supermercados o industria.

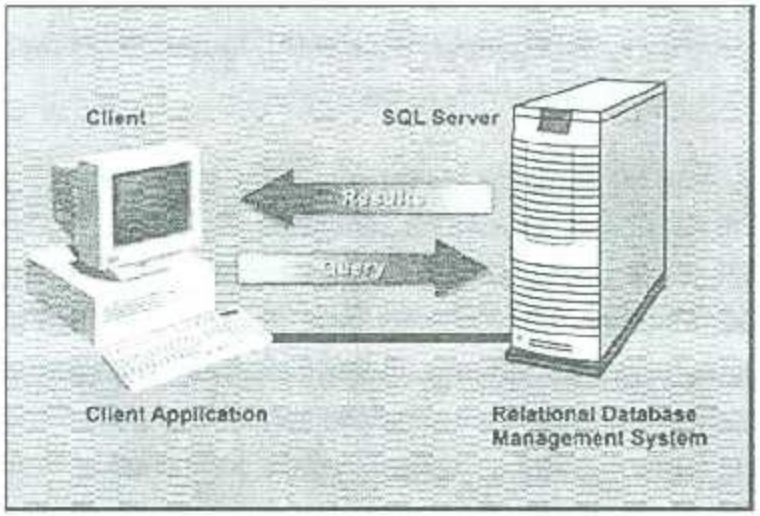


Figura 1.1 – Arquitectura de sistemas OLTP

(Figura obtenida de: http://uncontroldeuplicsa.blogspot.mx/2008/09/sistemas-oltp-epr-y-crm.html)

1.2.1 Gestión Transaccional

Es una metodología de proceso de información en tiempo real, en virtud de la cual tiene lugar una actualización fiable de la base de datos con cada transacción, que garantiza un alto grado de integridad en los datos, la eficiencia de cada transacción y la fiabilidad del sistema.

Entre las características que se le exige a un sistema OLTP están: Ante una transacción abortada, debe anularse cualquier modificación que se haya introducido con anterioridad a la interrupción (Atomicidad). Si una transacción resulta abortada, deber restituirse el anterior estado válido de los datos (Consistencia).

Los efectos de una transacción no deben ser observables por ninguna otra hasta que la transacción originaria haya concluido (Independencia).

Una vez validada una transacción, las modificaciones introducidas en los datos compartidos sobrevivirán a posibles fallos futuros en el sistema (Durabilidad).

1.2.2 Los Sistemas de Planificación de Recursos Empresariales

Los sistemas de planificación de recursos de la empresa (en inglés ERP, enterprise resource planning) son sistemas de gestión de información que integran y automatizan muchas de las prácticas de negocio asociadas con los aspectos operativos o productivos de una empresa.

Figura 1.2 – Modelo de un sistema ERP

(Figura obtenida de: http://uncontroldeupiicsa.blogspot.mx/2008/09/sistemas-oltp-erp-y-crm.html)

Se caracterizan por estar compuestos por diferentes partes integradas en una única aplicación. Estas partes son de diferente uso, por ejemplo: producción, ventas, compras, logística, contabilidad, gestión de proyectos, inventarios y control de almacenes, pedidos, nóminas, etc. (Figura 1.2)

Las características que distinguen a un ERP de cualquier otro software empresarial, es que deben de ser sistemas integrales, con modularidad y adaptables.

Integrales, porque permiten controlar los diferentes procesos de la compañía entendiendo que todos los departamentos de una empresa se relacionan entre sí, es decir, que el resultado de un proceso es punto de inicio del siguiente. Si la empresa no usa un ERP, necesitará tener varios programas que controlen todos los procesos mencionados, con la desventaja de que al no estar integrados, la información se duplica, crece el margen de contaminación en la información (sobre todo por errores de captura) y se crea un escenario favorable para malversaciones. Con un ERP, el operador simplemente captura el pedido y el sistema se encarga de todo lo demás, por lo que la información no se manipula y se encuentra protegida.

Modulares, porque su funcionalidad se encuentra dividida en módulos, los cuales pueden instalarse de acuerdo con los requerimientos del cliente. Ejemplo: ventas, materiales, finanzas, control de almacén, recursos humanos, etc.

Adaptables. Porque son desarrollados para adaptarse a la filosofía de cada empresa, así como a sus metodologías de administración y operación. Esto se logra por medio de la configuración o parametrización de los procesos de acuerdo con las salidas que se necesiten de cada uno. Los ERP más avanzados suelen incorporar herramientas de programación de 4ª Generación para el desarrollo rápido de nuevos procesos. La parametrización es el valor añadido fundamental que debe contar cualquier ERP para adaptarlo a las necesidades concretas de cada empresa. Otras características destacables de los sistemas ERP son:

- Base de datos centralizada.
- Los componentes del ERP interactúan entre sí consolidando todas las operaciones.
- En un sistema ERP los datos se ingresan sólo una vez y deben ser consistentes, completos y comunes.
- Las empresas que lo implanten suelen tener que modificar alguno de sus procesos para alinearlos con los del sistema ERP. Este proceso se conoce como: Reingeniería de Procesos. Aunque no siempre es necesario.
- Aunque el ERP pueda tener menús modulares configurables según los roles de cada usuario, es un todo. Esto significa: es un único programa con acceso a una base de datos centralizada.
- Las soluciones ERP en ocasiones son complejas y difíciles de implantar debido a que necesitan un desarrollo personalizado para cada empresa partiendo de la parametrización inicial de la aplicación que es común. Las personalizaciones y desarrollos particulares para cada empresa requieren de un gran esfuerzo en tiempo, y por consiguiente en dinero, para modelar todos los procesos de negocio de la vida real en la aplicación.
- Las metodologías de implantación de los ERP en la empresa no siempre son todo lo simples que se desearía, dado que entran en juego múltiples facetas. No hay recetas mágicas ni guiones explícitos para implantaciones exitosas; solamente trabajo bien realizado, una correcta metodología y aspectos que deben cuidarse antes y durante el proceso de implantación, e inclusive cuando el sistema entra en función.

1.2.3 Los Sistemas para la Gestión de la Relación con el Cliente

Los sistemas para la gestión de la relación con el cliente (en inglés CMR, Customer Relationship Management) son una estrategia de negocio desarrollada para lograr y gestionar mejores relaciones con los clientes. Para implementarla es preciso que la empresa adopte una cultura, estrategia y liderazgo dirigidos al cliente. (Figura 1.3).

Figura 1.3 – Modelo de un sistema CMR

(Figura obtenida de: http://uncontroldesplicsa.blogspot.mx/2008/09/sistemas-crp-erp-y-crm.html)

El uso de los sistemas CRM debe comenzar con una modificación estratégica que promueva cambios en la organización y en los procesos, y más tarde, en los sistemas de información de la empresa.

Puede atribuirse que estos sistemas han tenido buena acogida en una parte del mercado empresarial debido a que los sistemas ERP no han conseguido crear y mantener todas las ventajas competitivas que las empresas que los adoptaron esperaban de ellos. Además, el vasto acceso a la información que posibilita el Internet, pone a disposición de los clientes la oportunidad de elegir libremente la empresa a contratar para dar atención a sus problemas y necesidades.

La estrecha relación con los clientes que ofrecen estos sistemas, permite a las empresas obtener ciertas ventajas sobre sus competidores.

Algunas herramientas ofrecidas por los sistemas CRM son:

- **Herramientas de Marketing**: para realizar prospecciones del mercado y adquirir nuevos clientes gracias a la abundancia de datos y la gestión de campañas, enfatizando las relaciones duraderas con el cliente en lugar de la venta rápida.
- **Herramientas de Ventas:** Para cerrar negocios con procesos eficientes utilizando generadores de propuestas, configuradores, herramientas de gestión del conocimiento, agendas de contactos y ayudas para hacer previsiones.
- **Herramientas de comercio electrónico:** Para lograr un proceso de venta sencillo, rápido, eficaz y con el menor coste para la empresa.
- **Herramientas de Servicio:** Para gestionar el servicio post-venta y la atención al cliente con aplicaciones de call-center u opciones de auto-servicio en una página web. Y decimos "gestionar", no abandonar al cliente en una página FAQ inadecuada.

CRM es una estrategia de negocio para crear y mantener relaciones rentables y duraderas con los clientes. Los sistemas CRM que triunfan empiezan con una filosofía de negocio que cambia los procesos internos de la empresa para alinearlos con las necesidades de los clientes. Sólo entonces puede usarse la tecnología CRM como una herramienta fundamental que posibilita los procesos necesarios para convertir la estrategia en resultados de negocio [2].

1.3 Sistemas OLAP

Los sistemas de Procesamiento analítico en línea (en inglés OLAP, On-Line Analytical Processing) son bases de datos corporativas orientadas al análisis de la información que proveen de funciones de consulta especializada y de apoyo a la toma de decisiones. Esta clase de análisis suele implicar,

generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos...etc.

La razón de usar OLAP para realizar esta clase de consultas es su velocidad de respuesta, ya que éstas mismas, realizadas en sistemas gestores de bases de datos (SGBD) relacionales, se vuelven complejas e implican el cruzamiento de datos entre varias tablas, lo que se traduce en un rendimiento pobre.

Los datos se estructuran según las áreas del negocio, y los formatos de los datos se hallan integrados de manera uniforme en toda la organización. El historial de los datos es a largo plazo, regularmente en un rango comprendido entre dos y cinco años.

En los sistemas OLAP la información ya no es almacenada en tablas, sino en estructuras multidimensionales (Cubos OLAP). Los cubos OLAP contienen los datos resumidos procedentes de los distintos sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga. Un cubo OLAP Se compone de hechos numéricos llamados medidas que se clasifican por dimensiones. El cubo de metadatos es típicamente creado a partir de un esquema en estrella o copo de nieve, esquema de las tablas en una base de datos relacional. Las medidas se obtienen de los registros de una tabla de hechos y las dimensiones se derivan de la dimensión de los cuadros.

Figura 1.4 – Cubo OLAP, análisis por "Clientes"

(Figura obtenida de: Cano Josep Lluis, "Business Inteligence: Competir con Información")

En el cubo se muestran las unidades vendidas de cada uno de los libros, para los distintos clientes y en los distintos años (Figura 1.4). Este es el concepto de multidimensionalidad. Se dispone de las unidades vendidas de cada uno de los libros para cada uno de los clientes y en cada uno de los años: el contenido de un cubo individual son las ventas de un libro a un cliente en un año. Los contenidos de cada uno de los cubos individuales del cubo recogen lo que es conocido como "hechos" (en el ejemplo las unidades vendidas). En la actualidad, las soluciones OLAP permiten que cada una de los cubos individuales pueda contener más de un hecho. Las herramientas OLAP permiten "rotar" (en inglés "*slicing*") los cubos, es decir, cambiar el orden de las distintas dimensiones: En vez de analizar por "Clientes", como en el caso anterior, ahora el análisis se hará por "Libros", suponiendo que los usuarios que desean consultar son distintos y tienen distintas necesidades. Para esto, es necesario cambiar la dimensión "Clientes" por "Libros". (Figura 1.5).

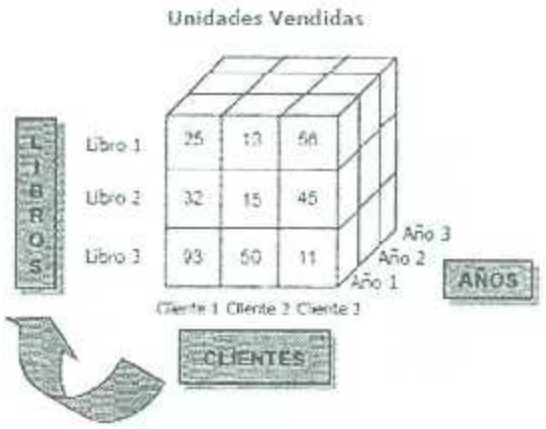


Figura 1.5 – Cubo OLAP, análisis por "Libros"

(Figura obtenida de: Carlo Joseo Lluis, "Business Intelligence: Competir con Información")

También se puede seleccionar (en inglés "*dicing*") sólo algunas de las celdas, por ejemplo: ¿Cuáles son las ventas al cliente 2, de los libros 1 y 2, en el año 1? (Figura 1.6).

Figura 1.6 – Cubo OLAP, "Dicing"

(Figura obtenida de: Carlo Joseo Lluis, "Business Intelligence: Competir con Información")

O lo que puede ser de interés es el total de libros, máximo nivel de agregación (en inglés "*roll-up*") (Figura 1.7).

Figura 1.7 – Cubo OLAP, "Roll - Up".

(Figura obtenida de: Cano Joseo Luis, "Business Intelligence: Competir con Información")

En el supuesto que se tengan libros de dos materias distintas: El libro 1 y el libro 2 son de la materia A y el libro 3 de la materia B. Partiendo del cubo anterior de las ventas agregadas, se incrementa el nivel de detalle (en inglés "*drill - down*") a través de la jerarquía "materias". En ese caso se obtendría (Figura 1.8):

Figura 1.8 – Cubo OLAP, "Drill - Down"

(Figura obtenida de: Cano Joseo Luis, "Business Intelligence: Competir con Información")

1.3.1 Otros Tipos de Herramientas OLAP

ROLAP: Relational OLAP

Las capacidades OLAP acceden directamente a la base de datos relacional. Se accede por tanto a una base de datos relacional (RDBMS). Accede habitualmente sobre un modelo "estrella". La principal ventaja es que no tiene limitaciones en cuanto al tamaño, pero es más lento que el MOLAP, aunque algunos productos comerciales nos permiten cargar cubos virtuales para acelerar los tiempos de acceso.

MOLAP: Multimensional OLAP

La implementación OLAP accede directamente sobre una base de datos multidimensional (MDDB). La ventaja principal de esta alternativa es que es muy rápida en los tiempos de respuesta y la principal desventaja es que, si queremos cambiar las dimensiones, debemos cargar de nuevo el cubo.

HOLAP: Hybrid OLAP

Accede a los datos de alto nivel en una base de datos multidimensional y a los atómicos directamente sobre la base de datos relacional. En esencia utiliza las ventajas del ROLAP y del MOLAP.

DOLAP: Desktop OLAP

Se crea un cubo con las dimensiones que le interesan al usuario, se carga en memoria en su ordenador, trabaja y, cuando acaba, se elimina de la memoria. La ventaja es que el usuario sólo recibe los hechos y las dimensiones de su interés y los analiza en forma local. Algunas de las herramientas del mercado permiten programar actividades, como por ejemplo

ejecutar consultas, publicar en web, lanzar alertas a través de la red, mediante correo electrónico o sobre agendas personales (PDA).

1.3.2 Formas de Acceso de las Herramientas OLAP

Cliente/Servidor

Cuando se tienen las instalaciones locales en los ordenadores de los usuarios.

Acceso Web: Cliente, cliente ligero, o sólo con el navegador.

En este tipo de acceso el navegador comunica con un servidor web, el cual habla con la aplicación del servidor, que es la que conecta con el *datawarehouse*. En el caso de acceder con el navegador sin ningún tipo de cliente o con cliente ligero (por ejemplo JAVA), normalmente se descargan pequeñas aplicaciones para aumentar la funcionalidad.

Principales herramientas de Business Intelligence:

- Generadores de informes: Utilizadas por desarrolladores profesionales para crear informes estándar para grupos, departamentos o la organización.
- Herramientas de usuario final de consultas e informes: Empleadas por usuarios finales para crear informes para ellos mismos o para otros; no requieren programación.
- Herramientas OLAP: Permiten a los usuarios finales tratar la información de forma multidimensional para explorarla desde distintas perspectivas y periodos de tiempo.
- Herramientas de Dashboard y Scorecard (Cuadros de mando): Permiten a los usuarios finales ver información crítica para el rendimiento con un simple vistazo utilizando iconos gráficos y con la

posibilidad de ver más detalle para analizar información detallada e informes, si lo desean.

- Herramientas de planificación, modelización y consolidación: Permite a los analistas y a los usuarios finales crear planes de negocio y simulaciones con la información de Business Intelligence. Pueden ser para elaborar la planificación, los presupuestos, las previsiones. Estas herramientas proveen a los dashboards y los scorecards con los objetivos y los umbrales de las métricas.
- Herramientas datamining: Permiten a estadísticos o analistas de negocio crear modelos estadísticos de las actividades de los negocios. Datamining es el proceso para descubrir e interpretar patrones desconocidos en la información mediante los cuales resolver problemas de negocio. Los usos más habituales del datamining son: segmentación, venta cruzada, sendas de consumo, clasificación, previsiones, optimizaciones, etc. [3].

1.4 Conclusiones

El objeto de este capítulo consistió en describir, a través de conceptos, definiciones y ejemplos gráficos, la funcionalidad principal de los sistemas OLTP y OLAP. En el capítulo fueron explicados de forma sintética sus atributos más notables. También fueron puestas de manifiesto sus implementaciones comunes, haciendo posible, de forma implícita, la identificación de sus más sustanciales diferencias. Sobre OLTP y OLAP puede concluirse, como una mención supletoria de lo anteriormente explicado, que para el desarrollo de todo sistema de inteligencia de negocios (Business Intelligence), resulta imprescindible la implementación previa de estas dos clases de sistemas. La falta o el funcionamiento deficiente de alguno de ellos, haría inviable todo intento de desarrollo de un eficiente sistema BI.

CAPÍTULO 2

Análisis y Diseño Multidimensional

Introducción

Partiendo de algunos conceptos ya someramente referidos en el capítulo anterior, en este segundo capítulo, además de la complementación de aquéllos, se proporcionan nuevos conceptos y esquemas cuya finalidad es explicar la metodología estándar aplicada en el modelamiento multidimensional.

El modelamiento multidimensional es una técnica para modelar bases de datos simples y entendibles al usuario final. La idea fundamental es que éste visualice fácilmente la relación existente entre los distintos componentes del modelo.

Aquí el espacio se define a través de ejes coordenados (por ejemplo X, Y, Z). Un punto cualquiera de este espacio quedará determinado por la intersección de tres valores particulares de sus ejes. En el modelo multidimensional cada eje corresponde a una dimensión particular. Entonces la dimensionalidad de la base de datos está en función de la cantidad de ejes (o dimensiones) que se le asocien. Cuando una base puede ser visualizada como un cubo de tres o más dimensiones, es más fácil para el usuario organizar la información e imaginarse en ella cortando y rebanando el cubo a través de cada una de sus dimensiones, para buscar la información deseada.

A lo largo de este capítulo se tratarán con minucia los aspectos más significativos acerca de la orientación analítica y el procesamiento de los datos en el modelamiento multidimensional, perteneciente a la tecnología Datawarehousing.

2.1 Introducción al Concepto Data Warehousing

Data Warehousing es el centro de la arquitectura para los sistemas de información. Soporta el procesamiento informático al proveer una plataforma sólida, a partir de los datos históricos para hacer el análisis. Facilita la integración de sistemas de aplicación no integrados. Organiza y almacena los datos que se necesitan para el procesamiento analítico de los mismos sobre una amplia perspectiva de tiempo.

Un Data Warehouse o Depósito de Datos es una colección de datos orientado a temas, integrado, de tiempo variante, no volátil, que se usa para el soporte del proceso de toma de decisiones gerenciales.

Se puede caracterizar un Data Warehouse haciendo un contraste de cómo los datos de un negocio almacenados en él, difieren de los datos operacionales utilizados por las aplicaciones de producción. (Figura 2.1):

Base de Datos Operacional	Data Warehouse
Datos Operacionales	Datos del negocio para Información
Orientado a la aplicación	Orientado al sujeto
Actual	Actual + histórico
Detallada	Detallada + más resumida
Cambia continuamente	Estable

Figura 2.1 – Datos de un Data Warehouse contrastados.
(Figura obtenida de http://fcces.unicauca.edu.co/old/datawarehouse.htm)

El ingreso de datos en el Data Warehouse proviene desde el ambiente operacional en casi todos los casos. El Data Warehouse es siempre un

almacén de datos transformados y separados físicamente de la aplicación donde se encontraron los datos en el ambiente operacional [4].

Las fuentes de información pueden incluir bases de datos relacionales, bases de conocimiento y documentos en distintos formatos. Los wrappers (encapsuladores) se encargan de extraer los datos de las distintas fuentes y transmitirlos al Data Warehouse. Los monitores están en contacto directo con las fuentes de datos para detectar los cambios que se puedan producir en ellas. El integrador es el responsable de filtrar, resumir y unificar la información proveniente de las distintas fuentes [5]. (Figura 2.2).

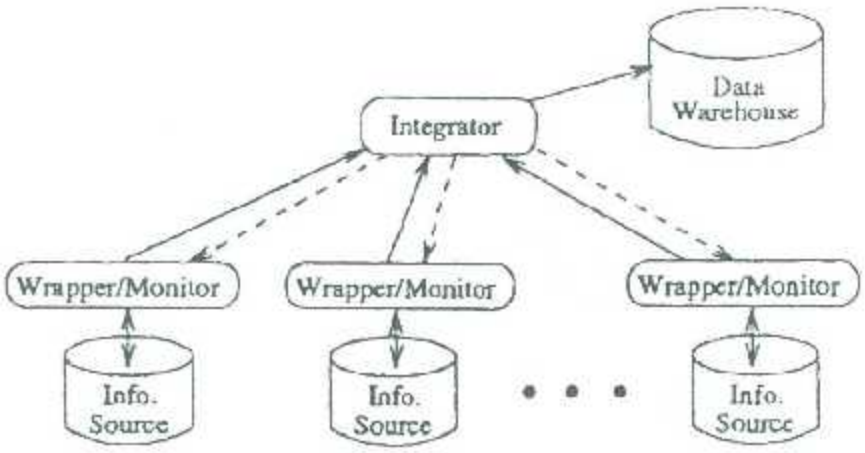


Figura 2.2 – Data Warehousing

(Figura obtenida de: http://elvex.ugr.es/idbis/db/docs/intro/1%20Modelo%20multidimensional.pdf)

2.1.1 Características de un Data Warehouse

Como se mencionó en el punto anterior, las características principales de un Data Warehouse son:

- Orientado al tema
- Integrado

- De tiempo variante
- No volátil.

2.1.1.1 Orientado a Temas

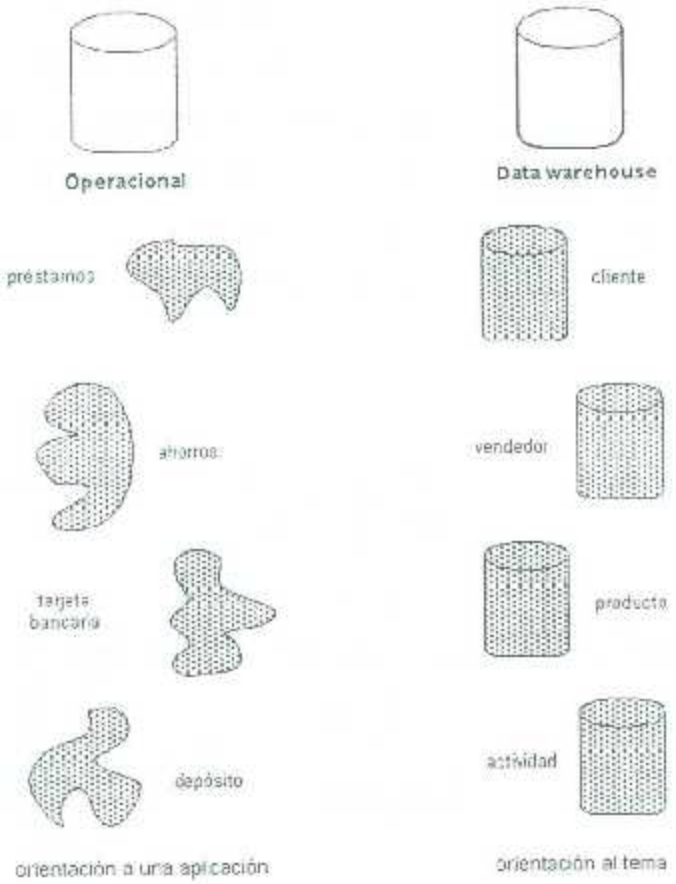


Figura 2.3 – Contraste entre los dos tipos de orientaciones.
(Figura obtenida de: http://www.ongei.gob.pe/publica/metodologias/Lib5084/131.HTM)

Una primera caracteristica del Data Warehouse es que la información se clasifica en base a los aspectos que son de interés para la empresa. Siendo así, los datos tomados están en contraste con los clásicos procesos orientados a las aplicaciones. (Figura 2.3)

El ambiente operacional se diseña alrededor de las aplicaciones y funciones tales como préstamos, ahorros, tarjeta bancaria y depósitos para una institución financiera. Por ejemplo, una aplicación de ingreso de órdenes puede acceder a los datos sobre clientes, productos y cuentas. La base de datos combina estos elementos en una estructura cue acomoda las necesidades de la aplicación.

En el ambiente Data Warehousing se organiza alrededor de sujetos tales como cliente, vendedor, producto y actividad. Por ejemplo, para un fabricante, éstos pueden ser clientes, productos, proveedores y vendedores. Para una universidad pueden ser estudiantes, clases y profesores. Para un hospital pueden ser pacientes, personal médico, medicamentos, etc.

La alineación alrededor de las áreas de los temas afecta el diseño y la implementación de los datos encontrados en el Data Warehouse. Las principales áreas de los temas influyen en la parte más importante de la estructura clave.

Las aplicaciones están relacionadas con el diseño de la base de datos y del proceso. En Datawarehousing se enfoca el modelamiento de datos y el diseño de la base de datos. El diseño del proceso (en su forma clásica) no es separado de este ambiente.

Las diferencias entre la orientación de procesos y funciones de las aplicaciones y la orientación a temas, radican en el contenido de los datos a nivel detallado. En el Data Warehouse se excluye la información que no será usada por el proceso de sistemas de soporte de decisiones, mientras que la información de las orientadas a las aplicaciones, contiene datos para satisfacer de inmediato los requerimientos funcionales y de proceso, que pueden ser usados o no por el analista de soporte de decisiones.

Otra diferencia importante está en la interrelación de la información. Los datos operacionales mantienen una relación continua entre dos o más tablas basadas en una regla comercial que está vigente. Las del Data Warehouse miden un espectro de tiempo y las relaciones encontradas en el Data Warehouse son muchas. Muchas de las reglas comerciales (y sus correspondientes relaciones de datos) se representan en el Data Warehouse, entre dos o más tablas.

2.1.1.2 Integración

El aspecto más importante del ambiente Datawarehousing es que la información encontrada al interior está siempre integrada. La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples y otros. (Figura 2.4)

Convenciones de Nombramiento.- El mismo elemento es frecuentemente referido por nombres diferentes en las diversas aplicaciones. El proceso de transformación asegura que se use preferentemente el nombre de usuario.

Medida de atributos. Los diseñadores de aplicaciones miden las unidades de medida de las tuberías en una variedad de formas. Un diseñador almacena los datos de tuberías en centímetros, otros en pulgadas, otros en millones de pies cúbicos por segundo y otros en yardas.

Al dar medidas a los atributos, la transformación traduce las diversas unidades de medida usadas en las diferentes bases de datos para transformarlas en una medida estándar común.

Cualquiera que sea la fuente, cuando la información de la tubería llegue al Data Warehouse necesitará ser medida de la misma manera.

Codificación. Los diseñadores de aplicaciones codifican el campo Género en varias formas. Un diseñador representa Género como una "M" y una "F", otros como un "1" y un "0", otros como una "X" y una "Y" e inclusive, como "masculino" y "femenino".

No importa mucho cómo el Género llega al Data Warehouse. Probablemente "M" y "F" sean tan buenas como cualquier otra representación. Lo importante es que sea de cualquier fuente de donde venga, el Género debe llegar al Data Warehouse en un estado integrado uniforme.

Por lo tanto, cuando el Género se carga en el Data Warehouse desde una aplicación, donde ha sido representado en formato "M" y "F", los datos deben convertirse al formato del Data Warehouse.

Fuentes Múltiples.- El mismo elemento puede derivarse desde fuentes múltiples. En este caso, el proceso de transformación debe asegurar que la fuente apropiada sea usada, documentada y movida al depósito.

Los puntos de integración afectan casi todos los aspectos de diseño: las características físicas de los datos, la disyuntiva de tener más de una de fuente de datos, el problema de estándares de denominación inconsistentes, formatos de fecha inconsistentes y otros.

Cualquiera que sea la forma del diseño, el resultado es el mismo - la información necesita ser almacenada en el Data Warehouse en un modelo globalmente aceptable y singular, aun cuando los sistemas operacionales subyacentes almacenen los datos de manera diferente.

Cuando el analista de sistema de soporte de decisiones observe el data Warehouse, su enfoque deberá estar en el uso de los datos que se encuentre en el depósito, antes que preguntarse sobre la confiabilidad o consistencia de los datos.

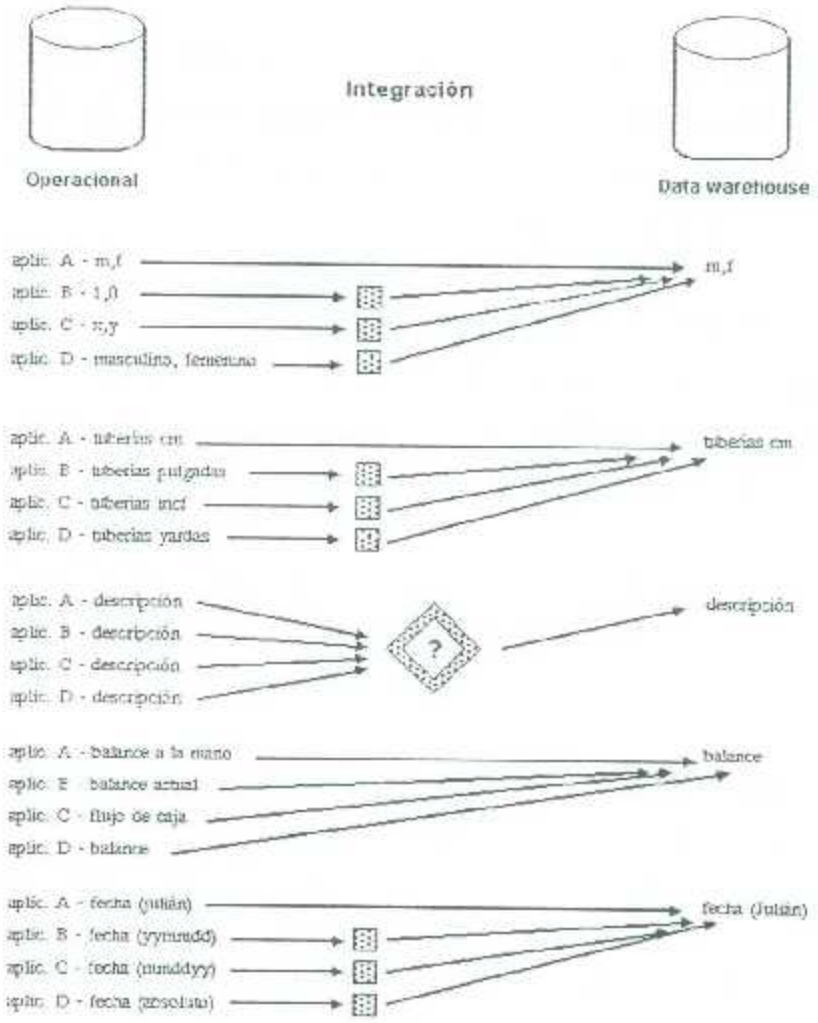


Figura 2.4 – Contraste entre los dos tipos de integraciones.
(Figura obtenida de: http://www.angel.gcb.pe/publica/metodologias/Lib5084/132.HTM)

2.1.1.3 De Tiempo Variante

Toda la información del Data Warehouse es requerida en algún momento. Esta característica básica de los datos en un depósito, es muy diferente de la información encontrada en el ambiente operacional. En éstos, la información se requiere al momento de acceder.

Como la información en el Data Warehouse es solicitada en cualquier momento (es decir, no "ahora mismo"), los datos encontrados en el depósito se llaman de "tiempo variante".

Los datos históricos son de poco uso en el procesamiento operacional. La información del depósito por el contraste, debe incluir los datos históricos para usarse en la identificación y evaluación de tendencias. (Figura 2.5).

Figura 2.5 – Contraste entre los dos horizontes de tiempo.
(Figura obtenida de: http://www.cngel.gob.pe/publica/metodologias/Jbb084/133.HTM)

El tiempo variante se muestra de varias maneras:

1.- La más simple es que la información representa los datos sobre un horizonte largo de tiempo - desde cinco a diez años. El horizonte de tiempo representado para el ambiente operacional es mucho más corto - desde valores actuales hasta sesenta a noventa días.

Las aplicaciones que tienen un buen rendimiento y están disponibles para el procesamiento de transacciones, deben llevar una cantidad mínima de datos si tienen cualquier grado de flexibilidad. Por ello, las aplicaciones

operacionales tienen un corto horizonte de tiempo, debido al diseño de aplicaciones rigidas.

2.- La segunda manera en la que se muestra el tiempo variante en el Data Warehouse está en la estructura clave. Cada estructura clave en el Data Warehouse contiene, implícita o explícitamente, un elemento de tiempo como día, semana, mes, etc.

El elemento de tiempo está casi siempre al pie de la clave concatenaca, encontrada en el Data Warehouse. En ocasiones, el elemento de tiempo existirá implícitamente, como el caso en que un archivo completo se duplica al final del mes, o al cuarto.

3° La tercera manera en que aparece el tiempo variante es cuando la información del Data Warehouse, una vez registrada correctamente, no puede ser actualizada. La información del Data Warehouse es, para todos los propósitos prácticos, una serie larga de "snapshots" (vistas instantáneas).

Por supuesto, si los snapshots de los datos se han tomado incorrectamente, entonces pueden ser cambiados. Asumiendo que los snapshots se han tomado adecuadamente, ellos no son alterados una vez hechos. En algunos casos puede ser no ético, e incluso ilegal, alterar los snapshots en el Data Warehouse. Los datos operacionales, siendo requeridos a partir del momento de acceso, pueden actualizarse de acuerdo a la necesidad.

2.1.1.4 No Volátil

La información es útil sólo cuando es estable. Los datos operacionales cambian sobre una base momento a momento. La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere una base de datos estable.

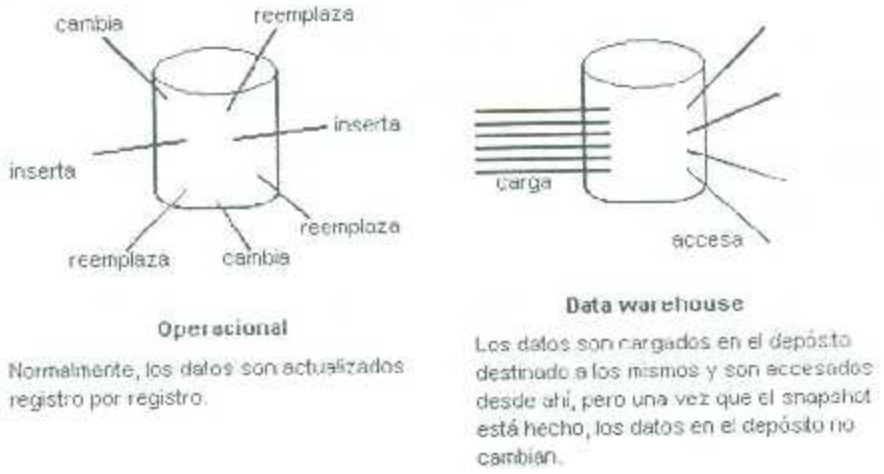


Figura 2.6 – Contraste entre las dos clases de actualizaciones de datos.
(Figura obtenida de: http://www.inge.gob.pe/publica/metodologias/L-b5084/134.HTM)

En la Figura 2.6 se muestra que la actualización (insertar, borrar y modificar), se hace regularmente en el ambiente operacional sobre una base de registro por registro. Pero la manipulación básica de los datos que ocurre en el Data Warehouse es mucho más simple. Hay dos únicos tipos de operaciones: la carga inicial de datos y el acceso a los mismos. No hay actualización de datos (en el sentido general de actualización) en el depósito, como una parte normal de procesamiento.

Hay algunas consecuencias muy importantes de esta diferencia básica, entre el procesamiento operacional y del Data Warehouse. En el nivel de diseño, la necesidad de ser precavido para actualizar las anomalías no es un factor en el Data Warehouse, ya que no se hace la actualización de datos. Esto significa que en el nivel físico de diseño, se pueden tomar libertades para optimizar el acceso a los datos, particularmente al usar la normalización y denormalización física.

Otra consecuencia de la simplicidad de la operación del Data Warehouse está en la tecnología subyacente, utilizada para correr los datos en el depósito. Teniendo que soportar la actualización de registro por registro en modo on-line (como es frecuente en el caso del procesamiento operacional) requiere que la tecnología tenga un fundamento muy complejo debajo de una fachada de simplicidad.

La tecnología permite realizar backup y recuperación, transacciones e integridad de los datos y la detección y solución al estancamiento que es más complejo. En el Data Warehouse no es necesario el procesamiento.

La fuente de casi toda la información del Data Warehouse es el ambiente operacional. A simple vista, se puede pensar que hay redundancia masiva de datos entre los dos ambientes. Desde luego, la primera impresión de muchas personas se centra en la gran redundancia de datos, entre el ambiente operacional y el ambiente de Data Warehouse. Dicho razonamiento es superficial y demuestra una carencia de entendimiento con respecto a qué ocurre en el Data Warehouse. De hecho, hay una mínima redundancia de datos entre ambos ambientes.

Se debe considerar lo siguiente:

Los datos se filtran cuando pasan desde el ambiente operacional al de depósito. Existen muchos datos que nunca salen del ambiente operacional, sólo aquéllos que realmente se necesitan ingresarán al ambiente de Data Warehouse.

El horizonte de tiempo de los datos es muy diferente de un ambiente al otro. La información en el ambiente operacional es más reciente con respecto a la del Data Warehouse. Desde la perspectiva de los horizontes de tiempo

únicos hay poca superposición entre los ambientes: operacional y de Data Warehouse.

El Data Warehouse contiene un resumen de la información que no se encuentra en el ambiente operacional. Los datos experimentan una transformación fundamental cuando pasa al Data Warehouse. La mayor parte de los datos se alteran significativamente al ser seleccionados y movidos al Data Warehouse. Dicho de otra manera, la mayoría de los datos se alteran física y radicalmente cuando se mueven al depósito. No son los mismos datos que residen en el ambiente operacional desde el punto de vista de integración.

En vista de estos factores, la redundancia de datos entre los dos ambientes es una ocurrencia rara, que resulta en menos de 1% [6].

2.2 El Modelo Multidimensional

Como ya había sido esbozado en el punto correspondiente a los sistemas OLAP del capitulo anterior, los datos dentro de un Data Warehouse se modelan en estructuras multidimensionales (cubos de datos) cuyas operaciones más comunes son:

- Roll up (incremento en el nivel de agregación de los datos).
- Drill down (incremento en el nivel de detalle, opuesto a roll up).
- Slice (reducción de la dimensionalidad de los datos mediante selección).
- Dice (reducción de la dimensionalidad de los datos mediante proyección).
- Pivotaje o rotación (reorientación de la visión multidimensional de los datos) [5]

2.2.1 Características

En general, la estructura básica de un Data Warehouse para el Modelo Multidimensional está definida por dos elementos: tablas y esquemas, que a continuación serán descritos.

2.2.1.1 Tabla Fact o de Hechos

Es la tabla central en un esquema dimensional. Es en ella donde se almacenan las mediciones numéricas del negocio. Estas medidas se hacen sobre el grano, o unidad básica de la tabla. El grano o la granularidad de la tabla queda determinada por el nivel de detalle que se almacenará en la tabla. Por ejemplo, para el caso de producto, mercado y tiempo antes visto, el grano puede ser la cantidad de madera vendida 'mensualmente'. El grano revierte las unidades atómicas en el esquema dimensional.

Cada medida es tomada de la intersección de las dimensiones que la definen. Idealmente está compuesta por valores numéricos, continuamente evaluados y aditivos. La razón de estas características es que así se facilita que los miles de registros que involucran una consulta sean comprimidos en unas pocas líneas en un set de respuesta.

La clave de la tabla fact recibe el nombre de clave compuesta o concatenada debido a que se forma de la composición (o concatenación) de las llaves primarias de las tablas dimensionales a las que está unida. Así entonces, se distinguen dos tipos de columnas en una tabla fact: columnas fact y columnas key. Donde la columna fact es la que almacena alguna medida de negocio y una columna key forma parte de la clave compuesta de la tabla.

2.2.1.2 Tablas Lock-Up o Dimensionales

Estas tablas son las que se conectan y alimentan a la tabla fact. Una tabla Lock-Up almacena un conjunto de valores que están relacionados a una dimensión particular. Las tablas Lock-Up no contienen hechos, en su lugar los valores son los elementos que determinan la estructura de las dimensiones. Así entonces, en ellas existe el detalle de los valores de la dimensión respectiva.

Una tabla Lock-Up está compuesta de una primary key que identifica unívocamente una fila en la tabla junto con un conjunto de atributos, y dependiendo del diseño del modelo multidimensional puede existir una foreign key que determina su relación con otra tabla Lock-Up.

Para decidir si un campo de datos es un atributo o un hecho se analiza la variación de la medida a través del tiempo. Si varía continuamente implicaría tomarlo como un hecho, caso contrario será un atributo. Los atributos dimensionales son un rol determinante en un Data Warehouse. Ellos son la fuente de todas las necesidades que debieran cubrirse. Esto significa que la base de datos será tan buena como lo sean los atributos dimensionales, mientras más descriptivos, manejables y de buena calidad, mejor será el Data Warehouse.

2.2.1.3 Esquema Estrella

En general, el modelo multidimensional también se conoce con el nombre de esquema estrella, pues su estructura base es similar: una tabla central y un conjunto de tablas que la atienden radialmente. (Figura 2.7).

El esquema estrella deriva su nombre del hecho que su diagrama forma una estrella, con puntos radiales desde el centro. El centro de la estrella consiste de una o más tablas fact, y las puntas de la estrella son las tablas Lock-up. Este modelo entonces, resulta ser asimétrico, pues hay una tabla dominante en el centro con varias conexiones a las otras tablas. Las tablas Lock-up tienen Sólo la conexión a la tabla fact y ninguna más.

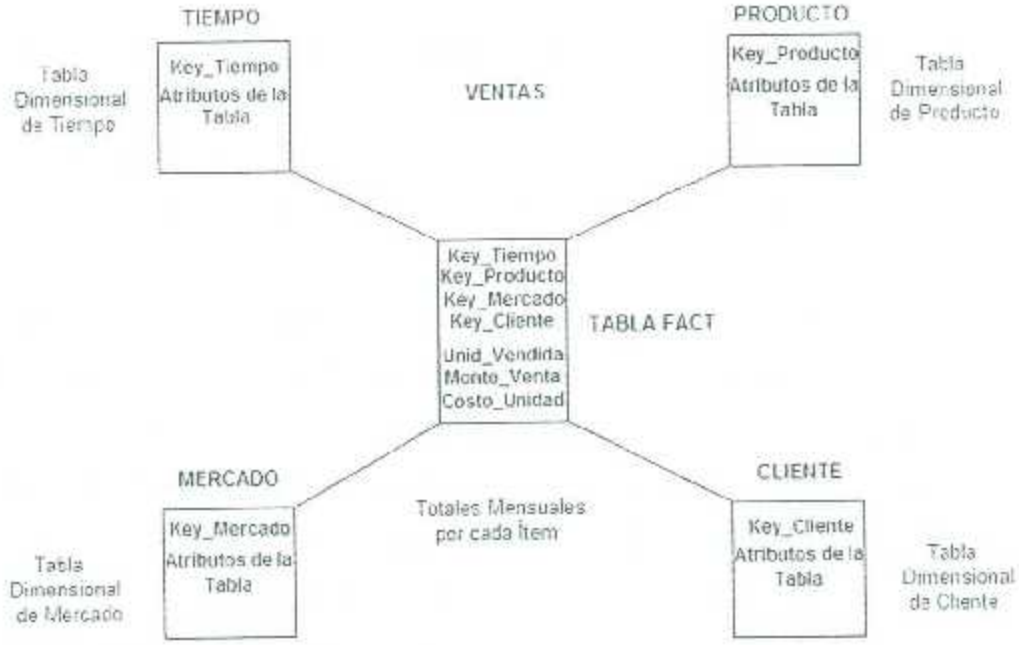


Figura 2.7 – Ilustración de un esquema estrella para una BD con dimensiones de Tiempo, Producto Mercado y Cliente.

(Figura obtenida de: http://www.inf.udec.cl/~revista/ediciones/edicion4/modmult.PDF)

2.2.1.4 Esquema Snowflake (Copo de nieve)

La diferencia del esquema copo de nieve comparado con el esquema estrella, está en la estructura de las tablas Lock-up: las tablas Lock-up en el esquema copo de nieve están normalizadas. Cada tabla Lock-up contiene

sólo el nivel que es clave primaria en la tabla y la foreign key de su parentesco del nivel más cercano del diagrama. Figura (2.8).

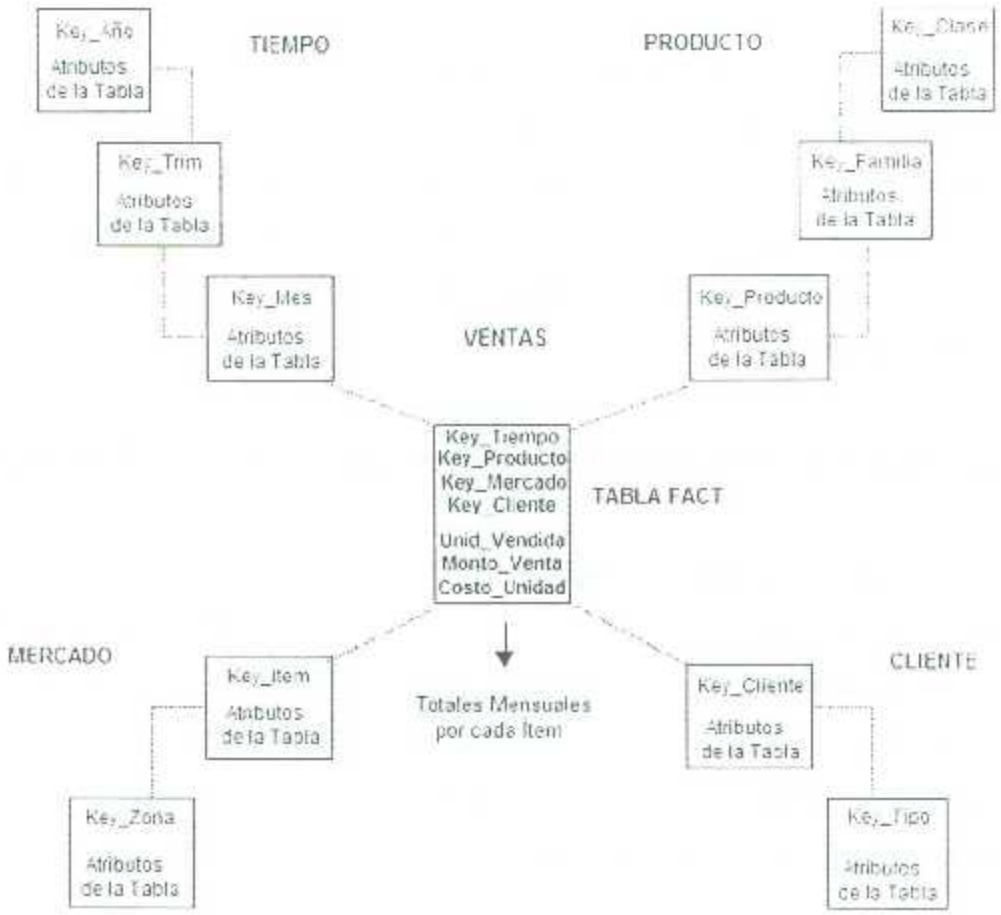


Figura 2.8 – Ilustración de un esquema Copo de nieve para el mismo ejemplo de la Fig. 2.7.
(Figura obtenica de: http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF)

2.2.2 Profundizaciones del Diseño

En este punto se definirán algunos conceptos fundamentales de diseño para el modelo multidimensional. Antes de pasar a los puntos posteriores es

conveniente comenzar explicando que las dimensiones son perspectivas o entidades respecto a las cuales una organización quiere mantener sus datos organizados (p.ej. tiempo, localización, clientes, proveedores etc.).

2.2.2.1 La Dimensión Tiempo

Virtualmente se garantiza que cada Data Warehouse tendrá una tabla dimensional de tiempo, debido a la perspectiva de almacenamiento histórica de la información. Usualmente es la primera dimensión en definirse, con el objeto de establecer un orden, ya que la inserción de datos en la base de datos multidimensional se hace por intervalos de tiempo, lo cual asegura un orden implícito.

2.2.2.2 Dimensiones que varian lentamente en el Tiempo

Son aquellas dimensiones que se mantienen "casi" constantes en el tiempo y que pueden preservar la estructura dimensional independiente del tiempo, con sólo agregados menores relativos para capturar la naturaleza cambiante del tiempo. Cuando se encuentra una de estas dimensiones se está haciendo una de las tres elecciones fundamentales siguientes. Cada elección resulta en un diferente grado de seguimiento sobre el tiempo:

Tipo 1: Sobrescribir el antiguo valor en el registro dimensional y por lo tanto perder la capacidad de dar seguimiento a la historia.

Tipo 2: Crear un registro dimensional adicional (con una nueva llave) que permita registrar el cambio presentado por el valor del atributo. De esta forma permanecerían en la base tanto el antiguo como el nuevo valor del registro con lo cual es posible segmentar la historia de la ocurrencia.

Tipo 3: Crear un campo "actual" nuevo en el registro dimensional original el cual almacene el valor del nuevo atributo, manteniendo el atributo original también. Cada vez que haya un nuevo cambio en el atributo, se modifica el campo "actual" solamente. No se mantiene un registro histórico de los cambios intermedios.

2.2.2.3 Niveles

Un nivel es un indicador de posición particular dentro de una dimensión; cada nivel sobre el nivel base representa la sumarización total de los datos desde el nivel inferior. Para un mejor entendimiento, se expondrá el siguiente ejemplo: considérese una dimensión Tiempo con tres niveles: Mes, Semestre, Año. El nivel Mes representa el nivel base, el nivel Semestre representa la sumarización de los totales por Mes y el nivel Año representa la sumarización de los totales para los Semestres. Agregar niveles de sumarización otorga flexibilidad adicional a usuarios finales de aplicaciones EIS/ DSS para analizar los datos.

2.2.2.4 Jerarquías

A nivel de dimensiones es posible definir jerarquías, las cuales son grupos de atributos que siguen un orden preestablecido. Una jerarquía implica una organización de niveles dentro de una dimensión, con cada nivel representando el total agregado de los datos del nivel inferior. Las jerarquías definen cómo los datos son sumarizados desde los niveles más bajos hacia los más altos. Una dimensión típica soporta una o más jerarquías naturales. Una jerarquía puede pero no exige contener todos los valores existentes en la dimensión. Se debe evitar caer en la tentación de convertir en tablas dimensionales separadas cada una de las relaciones muchos-a-uno presentes en las jerarquías. Esta descomposición es irrelevante en el

planeamiento del espacio ocupado en disco y sólo dificulta el entendimiento de la estructura para el usuario final, además de afectar el desempeño de la navegación [7]. Figura (2.9).

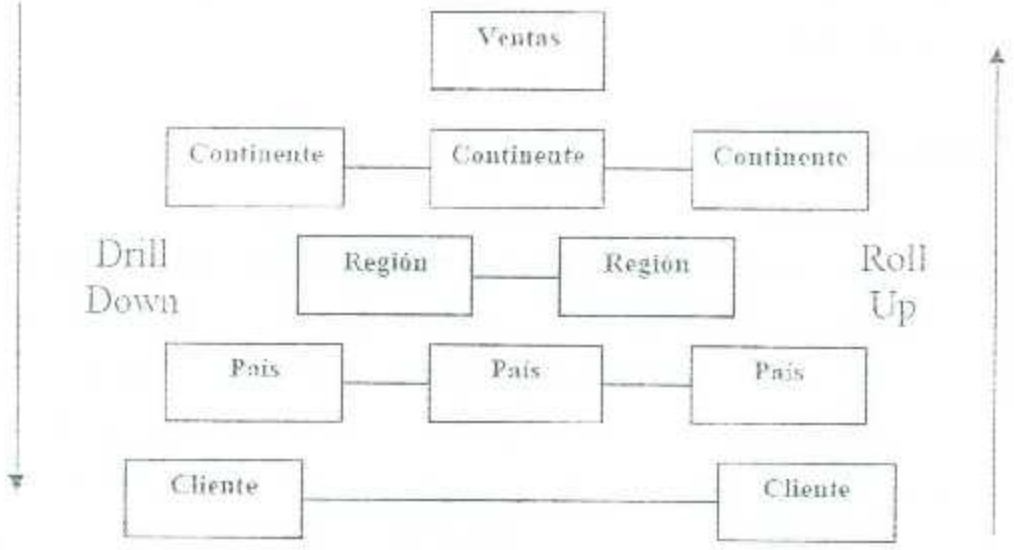


Figura 2.9 – Ilustración de una organización por jerarquías
(Figura obtenida de: http://elvex.ugr.es/idbis/db/docs/intro/F%20Modelo%20multidimensional.pdf)

2.3 Conclusiones

En el presente capitulo se trataron los aspectos elementales del análisis y diseño multidimensional, focalizando la atención en la estructura y funcionamiento del Data Warehouse.

El Data Warehouse ha conseguido desde hace tiempo atraer la atención de las grandes corporaciones porque provee un ambiente para que éstas optimicen el uso de la información administrada desde diversas aplicaciones operacionales. Como se sabe, un Data Warehouse es una colección de datos en la cual se halla integrada la información general de la empresa y que es empleada como un soporte durante el proceso de toma de decisiones gerenciales.

Pero, como es de suponer, el Data Warehouse no es ninguna herramienta que produzca por sí sola, a partir de la nada, resultados de capital importancia para las empresas. Es deber de los administradores empresariales y de los analistas, el acceso y la recuperación de los datos contenidos en el Data Warehouse, con la finalidad de convertirlos en información y en hechos.

Sobre estos hechos descansa la base de conocimientos que permite determinar el estado actual de una empresa, así como el trazo de futuros rumbos exitosos. Para esto es preciso eludir la tendencia natural al acopio desmesurado de datos, que conforman montañas de información, en muchos casos innecesaria. Lo esencial no es el conocimiento de todos los hechos asociados al negocio, sino la óptima elección de aquéllos cuyo valor sea crucial y determinante para el cumplimiento de los propósitos planteados.

CAPÍTULO 3

SQL Server 2008

y

AdventureWorks

Introducción

En este capítulo se aborda la documentación de un ejemplo proporcionado por Microsoft a través de un tutorial de su página web.

Los objetivos de este capítulo se focalizan en la instalación de la base de datos *AdventureWorks 2008*, la creación de un nuevo proyecto, la definición de un origen de datos y la creación y visualización de un informe desde el *ReportViewer* basado en los datos de la base precitada. Con esto se pretende mostrar, a groso modo, la funcionalidad de las dos aplicaciones que dan título a este capítulo de la tesis.

Para el desarrollo de este capítulo es necesaria la instalación previa de alguna de las versiones de *Microsoft SQL Server* y de *Visual Studio*, para este caso en particular, se utilizarán las versiones 2008 y 2010 respectivamente.

En virtud de que el desarrollo de este capítulo supone el despliegue de una serie de acciones técnicas, la parte medular del tema en cuestión ha sido documentada con la amplitud debida, mediante la inclusión de explicaciones textuales y gráficas, prescindiendo de instalaciones e instrucciones accesorias, con la finalidad de moderar la extensión del capítulo y asimismo no obrar en detrimento de su claridad y comprensibilidad.

3.1 Instalación de la Base de Datos AdventureWorks 2008

La base de datos está disponible para su descarga en *Codeplex*, cuya dirección web es la siguiente: http://msftdbprodsamples.codeplex.com/

Una vez en la página, dar un clic sobre la pestaña *Downloads* como se muestra en la siguiente figura. Figura (3.1).

Figura 3.1 – Pestaña para descargas

A continuación se elige la base de datos de AdventureWorks, para su descarga, según la versión deseada, como se muestra en la siguiente figura. Figura (3.2).

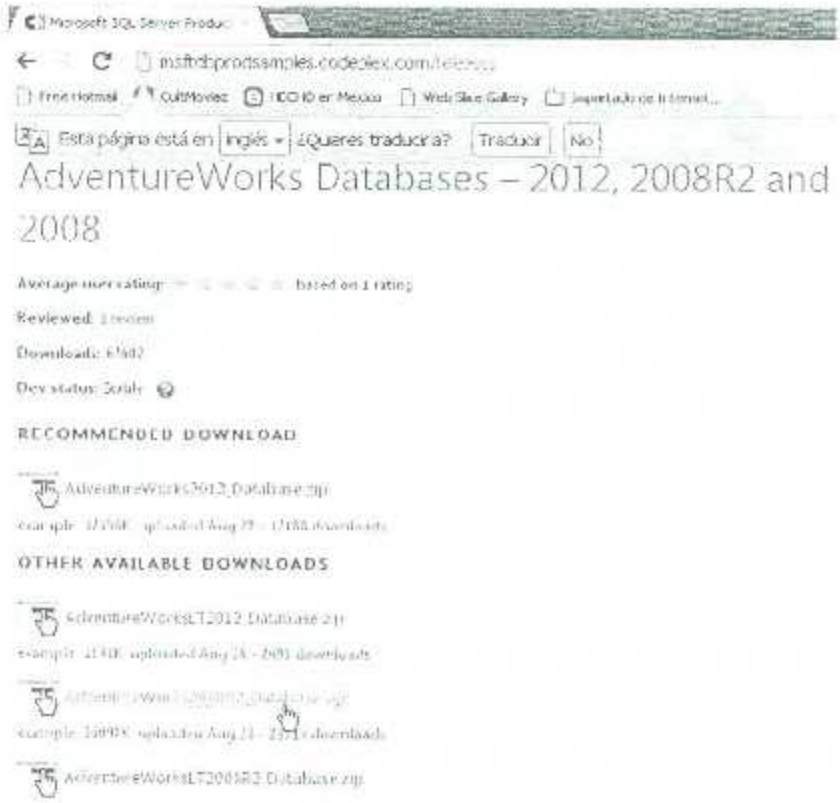


Figura 3.2 – Seleccionar la base de datos a descargar

Por último, se localiza el archivo descargado para proceder a la extracción de la base de datos [8]. Figuras (3.3) y (3.4)

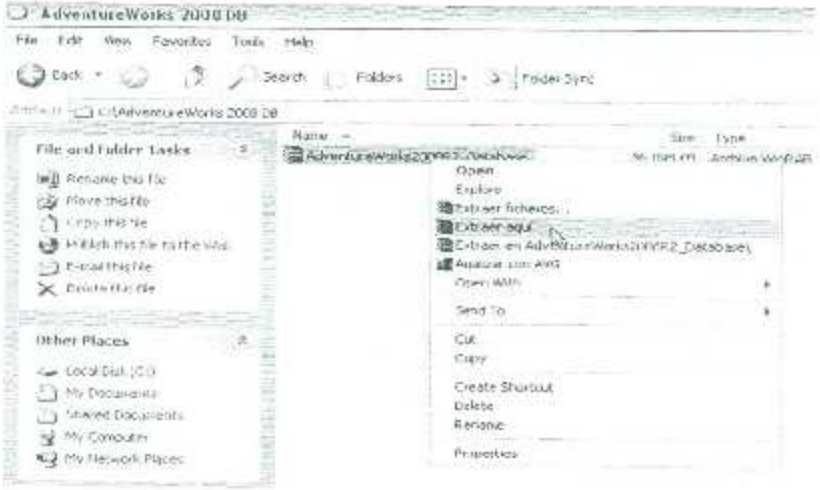


Figura 3.3 – Extracción de la base de datos

Figura 3.4 – Base de datos extraída.

3.2 Crear un Informe desde el ReportViewer

Para crear un informe hay que abrir el *Visual Studio* e iniciar un nuevo proyecto. Para este caso se elegirá dentro de las plantillas instaladas: *Visual Basic*. Dentro de ésta opción se seleccionará: *Aplicación de Windows Forms*. Después de nombrar la forma y elegir el directorio en que se desea ubicar el proyecto, el formulario *Form1* será visible.

El siguiente paso es agregar un nuevo elemento de tipo *DataSet* al formulario. Se elige: *Agregar nuevo elemento*, del menú: *Proyecto*. Posteriormente se elige el elemento DataSet, se especifica un nombre para el conjunto de datos (parte inferior de la ventana) o se mantiene el propuesto por el programa, y finalmente se da un clic sobre el botón agregar. Figura (3.5).

Esto agrega al proyecto un archivo *XSD* y abre el diseñador de *DataSet*.

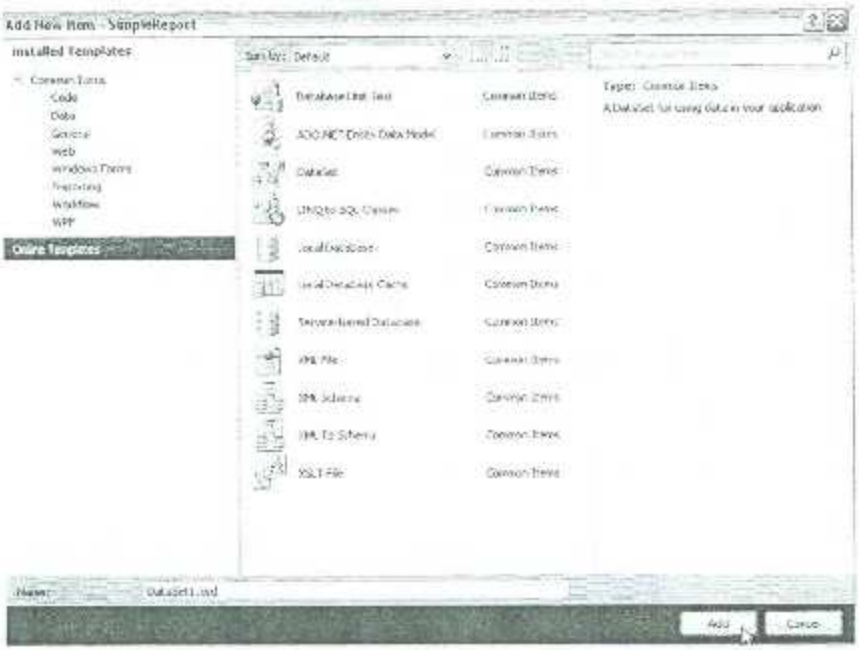


Figura 3.5 – Agregar un DataSet al formulario

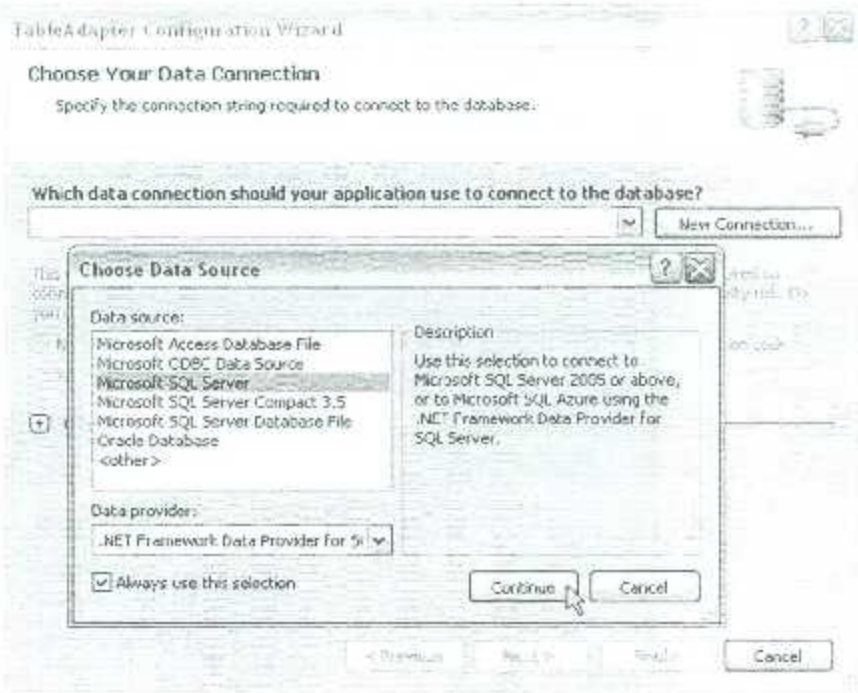


Figura 3.6 – Seleccicnar un crigen de datos en el diseñacor DataSet.

Una vez en el diseñador del *DataSet*, debe arrastrarse hasta él, desde el cuadro de herramientas, el control: *TableAdapter*, lo que dará inicio al asistente para la configuración de dicho control. Dentro del asistente hay que dar un clic sobre el botón: *nueva conexión*, y posteriormente seleccionar de lista mostrada el origen de datos: *Microsoft SQL Server*. Finalmente, debe darse un clic sobre el botón: *continuar*. Figura (3.6).

En el cuadro de diálogo: *Agregar conexión*, seleccionar el servidor y la base de datos y utilizar la opción predeterminada: *Usar Autenticación Windows* en: *Conexión con el servidor*. Figura (3.7).

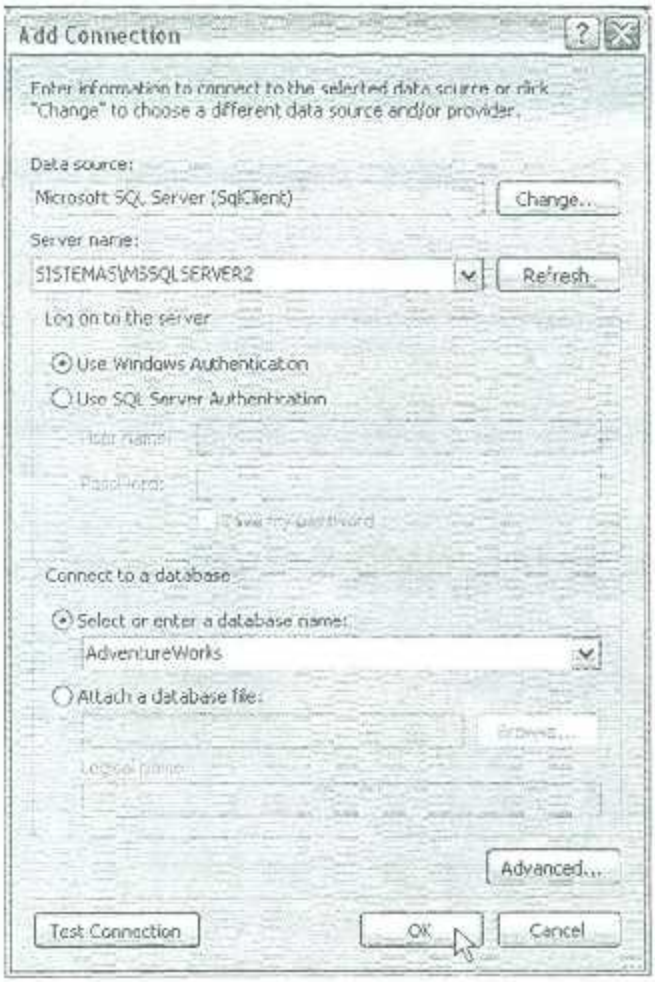


Figura 3.7 – Seleccionar el servidor, la base de datos y el tipo de conexión al servidor.

De regreso en el asistente para la configuración del *TableAdapter*, debe darse un clic sobre el botón: *Siguiente*. Si se especificó *Usar autenticación de SQL Server*, debe elegirse entre incluir los datos confidenciales en la cadena o establecer la información en el código de la aplicación.

En la página: *Guardar cadena de conexión en el archivo de configuración. de la aplicación*, debe escribirse el nombre de la cadena de conexión o aceptar el valor predeterminado *AdventureWorks2008ConnectionString*. Luego debe darse un clic en *Siguiente*. Figura (3.8).

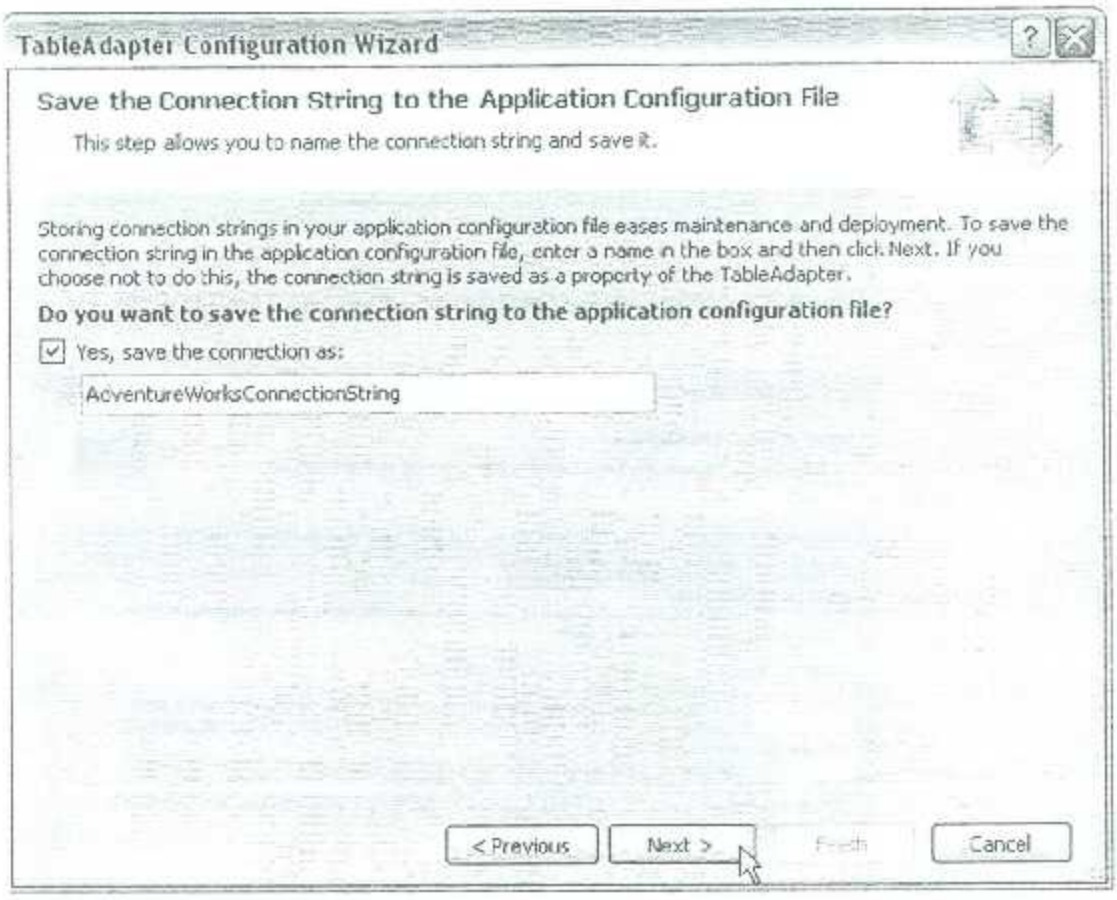


Figura 3.8 – Guardar la cadena de conexión en el archivo de configuración.

Posteriormente debe elegirse un tipo de comando, aquí se debe seleccionar: *Usar instrucciones SQL*: luego dar un clic en *Siguiente*. Figura (3.9).

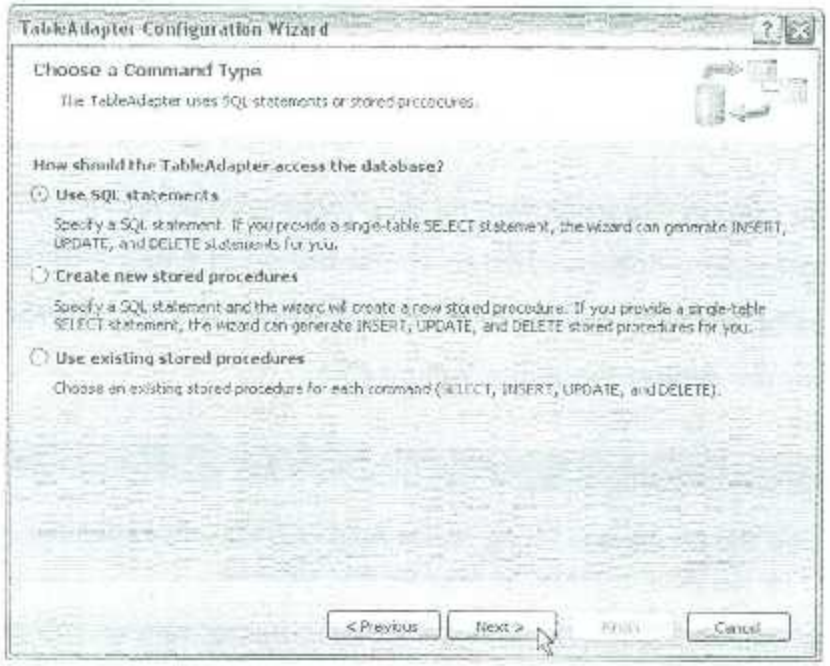


Figura 3.9 – Seleccionar el tipo de comando: Usar instrucciones SQL.

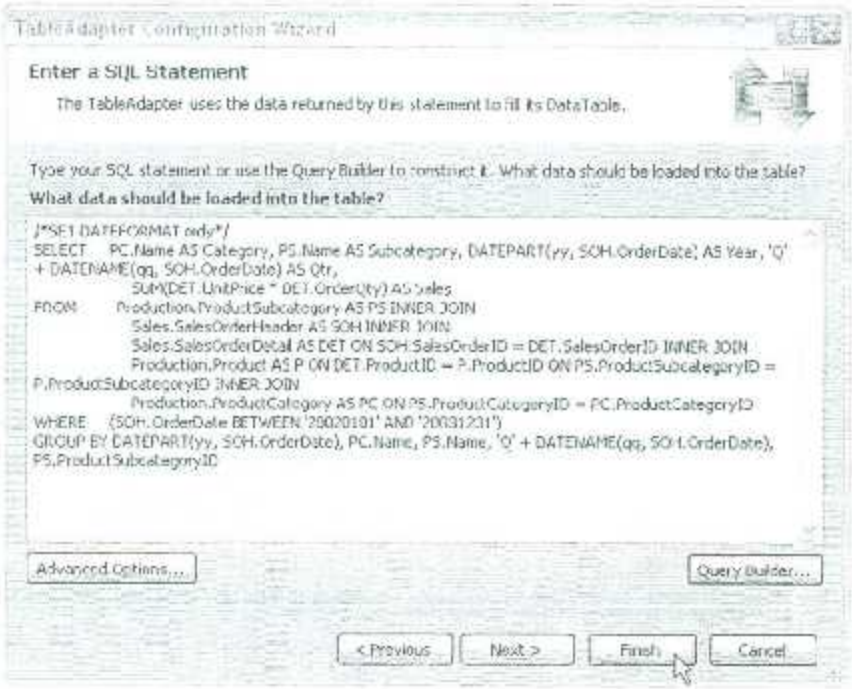


Figura 3.10 – Consulta Transact-SQL para recuperar los datos de ventas.

Después hay que agregar una consulta *Transact-SQL* para recuperar los datos de ventas de la base de datos. Antes de dar un clic en el botón: *Finalizar*, es recomendable dar un clic en el botón: *Generador de consultas*, para crear y validar la consulta. Figura (3.10). Finalmente debe guardarse el archivo *DataSet1*.

El siguiente paso es agregar un nuevo elemento de tipo *Report Wizard*. Se elige: *Agregar nuevo elemento*, del menú: *Proyecto*. Posteriormente se elige el elemento *Report Wizard*, se especifica un nombre para el conjunto de datos (parte inferior de la ventana) o se mantiene el propuesto por el programa, y finalmente se da un clic sobre el botón agregar. Detrás del cuadro de diálogo se abre una superficie de diseño gráfico. Figura (3.11).

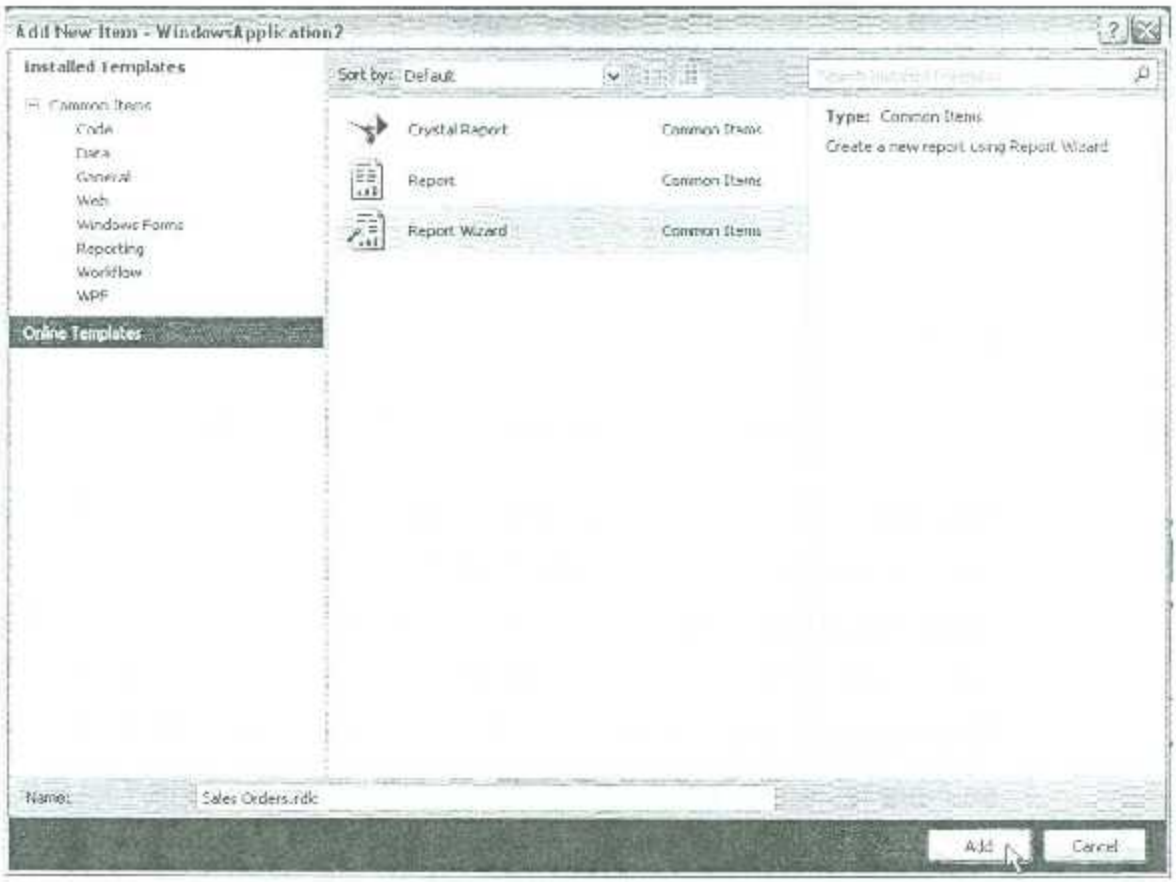


Figura 3.11 – Agregar un Report Wizard al formulario.

En la página *Propiedades del conjunto de datos*, en la lista desplegable de *Origen de datos*, se selecciona el *DataSet* creado previamente. El cuadro *Conjuntos de datos disponibles* se actualiza automáticamente con el *DataTable* creado. Luego hacer clic en *Siguiente*. Figura (3.12).

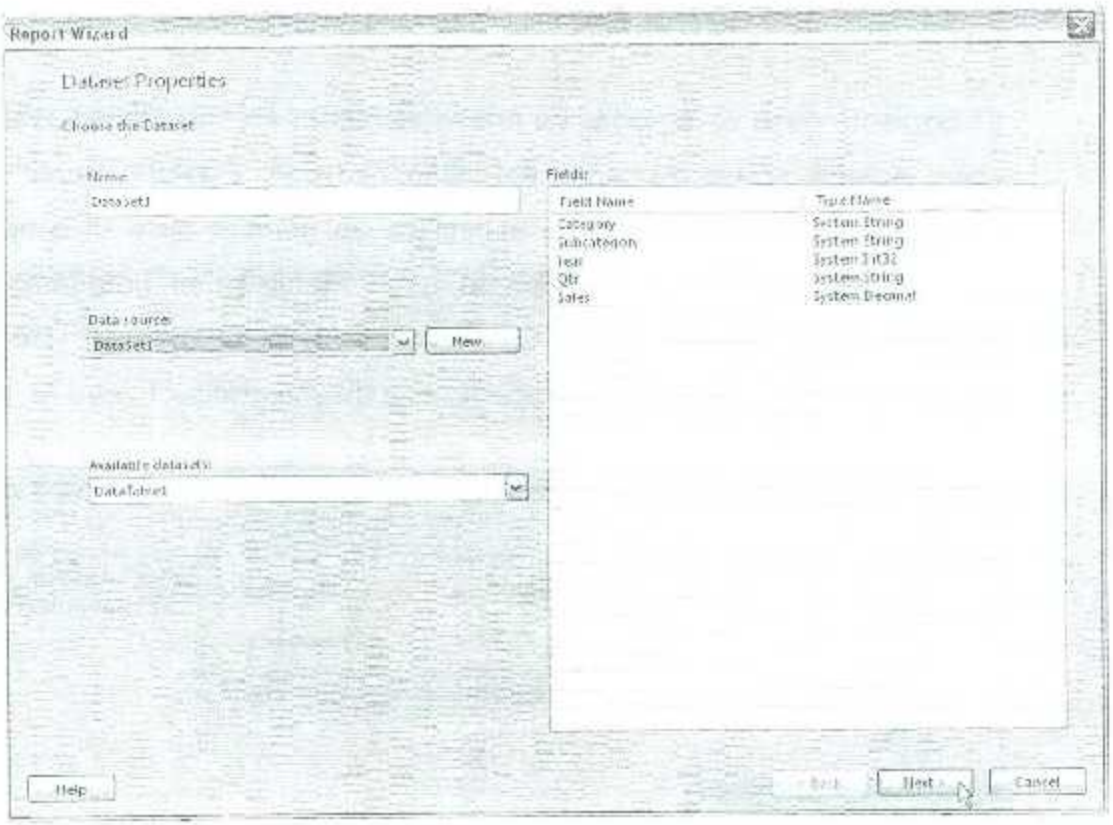


Figura 3.12 – Selección del DataSet creado anteriormente.

Para este ejemplo, en la página: *Organizar Campos*, deben arrastrarse desde el cuadro: *Campos disponibles*, los campos: *Category* y *Subcategory*, hasta el cuadro: *Grupo de Filas*. Los campos: *Year* y *Qtr*, hasta el cuadro: *Grupo de Columnas*, y el campo: *Sale* hasta el cuadro: *Valores*. Posteriormente dar clic en *Siguiente* dos veces y, a continuación, hacer clic en *Finalizar*, de esta forma se habrá creado el informe. Figura (3.13).

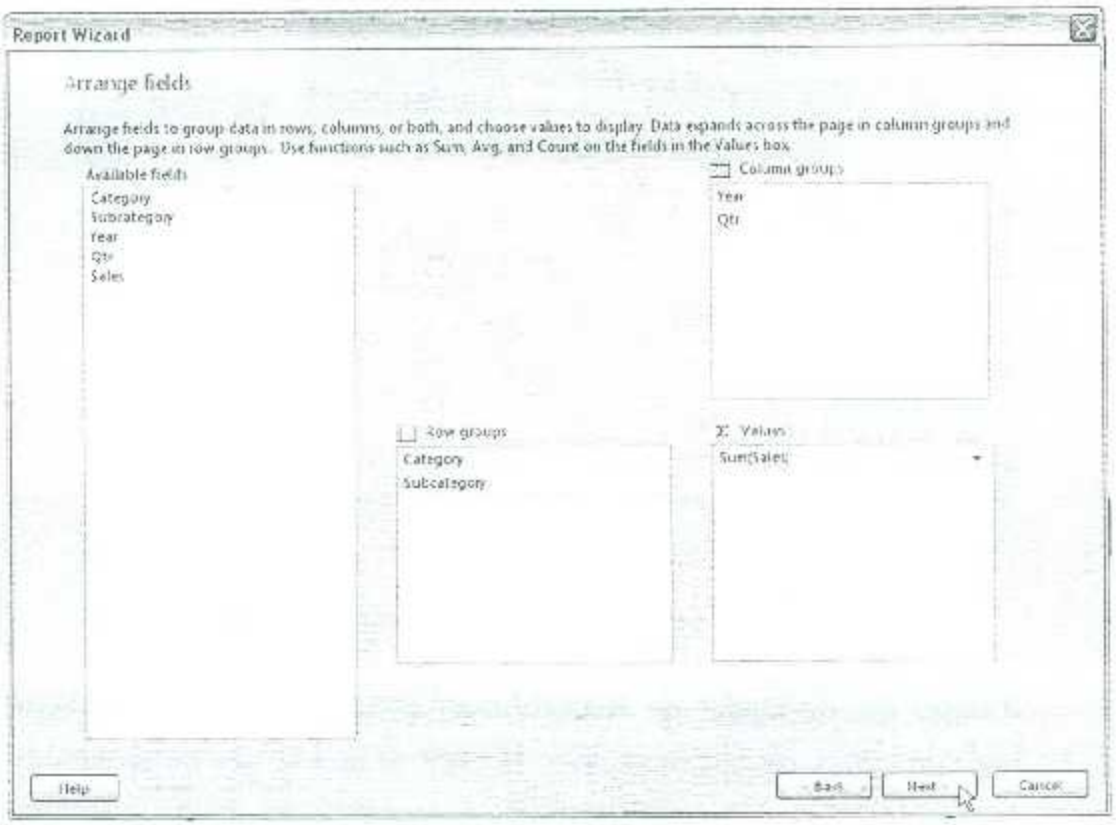


Figura 3.13 – Organización de los campos del DataSet en el Report Wizard.

Para agregar el control *ReportViewer* al formulario, debe hacerse clic en *Form1.vb*, dentro del explorador de soluciones. Luego, en el menú: *Ver*, seleccionar: *Diseñador*. Desde la sección: *Informes del cuadro de herramientas*, debe arrastrarse el control *ReportViewer* hasta el formulario. Después debe abrirse el recuadro de etiquetas inteligentes del control *ReportViewer1* haciendo clic en el glifo de la etiqueta inteligente de la esquina superior derecha. Luego hay que hacer clic en la lista desplegable, elegir informe y seleccionar: *SimpleReport.SalesOrders.rdlc*. En el recuadro de etiquetas inteligentes, elegir: *Acoplar* en contenedor principal. De esta forma el control habrá sido agregado al formulario. Figura (3.14).

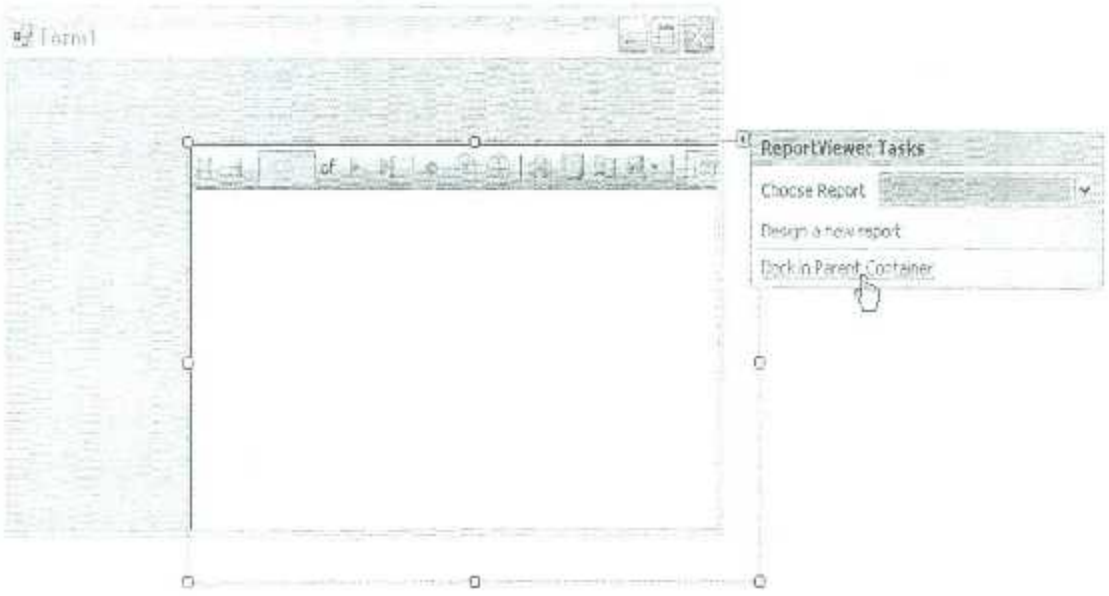


Figura 3.14 – Agregar el control ReportViewer al formulario.

Dentro del diseñador de ReportViewer puede ser un poco desbastado el reporte antes de ser mostrado. Desde el cuadro de herramientas puede arrastrarse un cuadro de texto hasta la superficie de diseño y posteriormente escribir en él un título para el reporte. Figura (3.15).

Figura 3.15 – Agregar cuacro de texto para el titulo del reporte

Para establecer el formato a los campos numéricos debe darse clic con el botón secundario sobre la celda que contiene el valor que se desea ajustar y luego dar un clic sobre la opción del menú: *Propiedades de cuadro de texto*. Posteriormente ajustar las opciones de formato deseadas. Figuras (3.16) y (3.17).

Figura 3.16 – Entrar a la configuración de formato de un valor numérico.

Figura 3.17 – Ajustar el formato del valor numérico.

Para agregar un gráfico que ilustre información significativa del reporte, desde el cuadro de herramientas puede arrastrarse un componente gráfico hasta la superficie de diseño. Figura (3.18).

Figura 3.18 – Agregar un componente gráfico al reporte.

Una vez instalado el gráfico, desde el conjunto de datos del reporte se arrastran los datos seleccionados hasta el gráfico insertado, colocándolos en las áreas de categoría, campos o series, según se deseé. Figura (3.19).

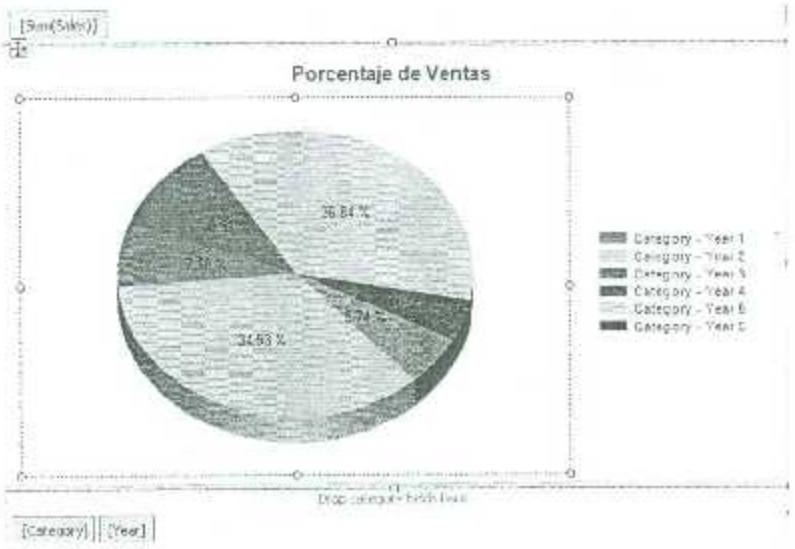


Figura 3.19 – Arrastrar los datos deseados al gráfico.

Para este ejemplo, se agregaron los campos: *Category* y *Year* en el área de categorías y el campo: *Sales* en el área de datos. De esta forma, el gráfico mostrará el porcentaje de ventas anuales según la categoría de los productos.

Finalmente, presionando la tecla F5 se compila la aplicación y el informe puede verse en el formulario [9]. Figura (3.20).

Ventas de la Compañía

Categoría	Subcategoría	Año Total	Año Total	Total
Accessories	Total	$93,796.84	$595,014.24	$688,811.08
Bikes	Total	$26,664,534.04	$35,199,346.24	$61,863,880.28
Clothing	Total	$489,820.19	$1,024,473.88	$1,514,294.08
Components	Total	$3,611,041.24	$5,489,740.88	$9,100,782.12
Total		$30,859,192.31	$42,308,575.23	$73,167,767.53

Porcentaje de Ventas

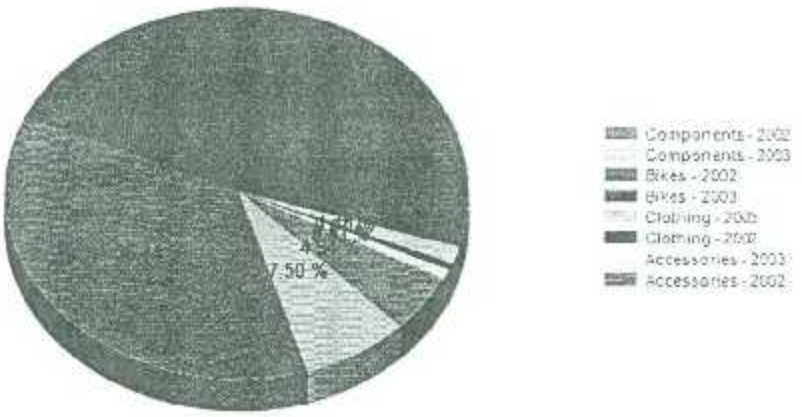


Figura 3.20 – Reporte de ventas obtenido

3.3 Conclusiones

Como se mencionó en la introducción a este capítulo, la pretensión del mismo era mostrar, a través de la realización y explicación puntual de un sencillo tutorial de Microsoft, una parte de las funcionalidades del SQL Server y del AdventureWorks.

La documentación de este capítulo, como también fue referido en la introducción, consistió en explicar textual y gráficamente los pasos esenciales para acceder a la información de ejemplo, relacionar algunos de sus datos más valiosos a través de una consulta transaccional y finalmente mostrarlos, de forma simple y clara, mediante un visor de informes.

Este breve capítulo debe ser entendido como una mera introducción a los capítulos posteriores, en los cuales se profundizará más, desde la parte práctica, en los aspectos técnicos que cimentan la inteligencia de negocios, mismos que se han ido tratando de forma teórica a lo largo de los capítulos iniciales.

CAPÍTULO 4

Construyendo un Cubo OLAP

Introducción

Con el fin de destacar los aspectos fundamentales concernientes a la creación, implementación y operaciones básicas de un cubo OLAP, este capítulo aborda la documentación textual y gráfica de algunas lecciones pertenecientes al tutorial *SQL Server Analysis Services*, proporcionado por Microsoft a través de su página web

Para tal propósito han sido obviadas algunas explicaciones técnicas elementales, así como algunas lecciones preliminares del tutorial y otras más cuya extensión temática sobrexcede los límites generales considerados para este capítulo.

Como ya se mencionó, además de la creación e implementación de un cubo OLAP, se tratan de forma sintetizada las funcionalidades principales del mismo, entre las cuales pueden mencionarse: La definición de una dimensión, agregación de atributos, modificaciones a medidas, agregación de una jerarquía, agregación de un cálculo con nombre, definición de una relación de atributos, y finalmente, la visualización de cierta información de ejemplo desde el examinador.

El desarrollo de este proyecto ha sido realizado desde el *Business Intelligence Development Studio*, incluido en *SQL Server 2008*, y la parte correspondiente a los datos fue tomada de una empresa ficticia de nombre: *Adventure Works Cycles*, incluida en las bases de datos de ejemplo de AdventureWorks.

4.1 Definir e Implementar un Cubo OLAP.

Después de efectuar algunos pasos previos como la creación de un proyecto de *Analysis Services*, de nombre: *Tutorial de Analysis Services*, la definición del origen de datos: *AdventureWorksDW2008*, la definición de una vista de origen de datos, seleccionando las tablas: *DimCustomer, DimDate, DimGeography, DimProduct y FactInternetSales*, y la modificación de los nombres predeterminados de éstas, omitiendo sus prefijos, para hacerlos más descriptivos, se procederá a explicar la definición e implementación de un cubo OLAP.

4.1.1 Definir una Dimensión.

Para definir una dimensión debe ejecutarse el *Asistente para dimensiones* desde el *Explorador de soluciones* del proyecto.

En Seleccionar un método de creación debe seleccionarse la opción *Usar una tabla existente.*

En *especificar información de Origen,* deben seleccionarse *AdventureWorksDW2008* y *Date* como la vista del origen de datos y la tabla principal respectivamente.

En *Seleccionar los atributos de la dimensión*, deben seleccionarse las casillas correspondientes a los siguientes atributos: *Date Key, Full Date Alternat, Key, English Mont Name, Calendar Quarter, Calendar Year, Calendar Semester,* y modificar, a partir del segundo, sus tipos de atributo predefinidos por: Fecha, mes, cuatrimestre, año y semestre, respectivamente. Figura (4.1).

Figura 4.1 Seleccionar los atributos de la dimensión.

En la finalización del *Asistente para dimensiones* se muestra la dimensión *Date* y la relación de los atributos que fueron agregados. Figura (4.2).

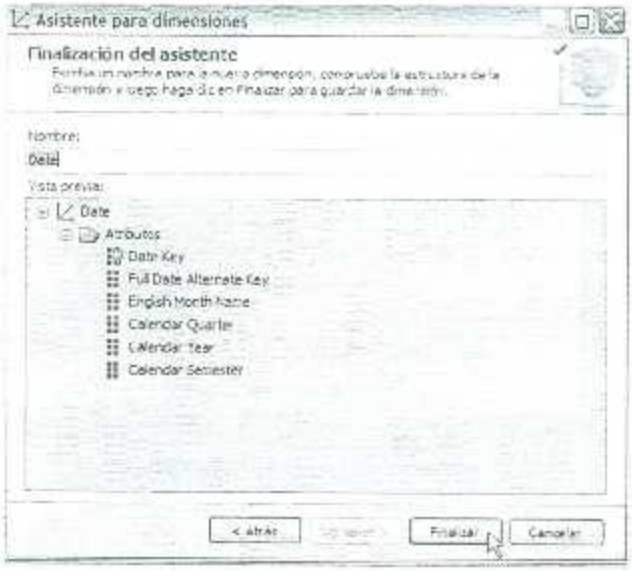


Figura 4.2 – Finalización del asistente para dimensiones.

4.1.2 Definir un Cubo

Para definir un cubo debe ejecutarse el *Asistente para cubos* desde el *Explorador de soluciones* del proyecto.

En Seleccionar un método de creación debe seleccionarse la opción *Usar una tabla existente*.

En *Seleccionar tablas de grupo de medida*, debe seleccionarse *AdventureWorksDW2008* como la vista del origen de datos. Luego dar un clic sobre el botón Sugerir. Después de esto, el asistente sugerirá *InternetSales* como tabla de grupos de medida. Las tablas de grupos de medida, también denominadas tablas de hechos, contienen las medidas que pueden ser de interés, como el número de unidades vendidas, por citar un ejemplo. Figura (4.3).

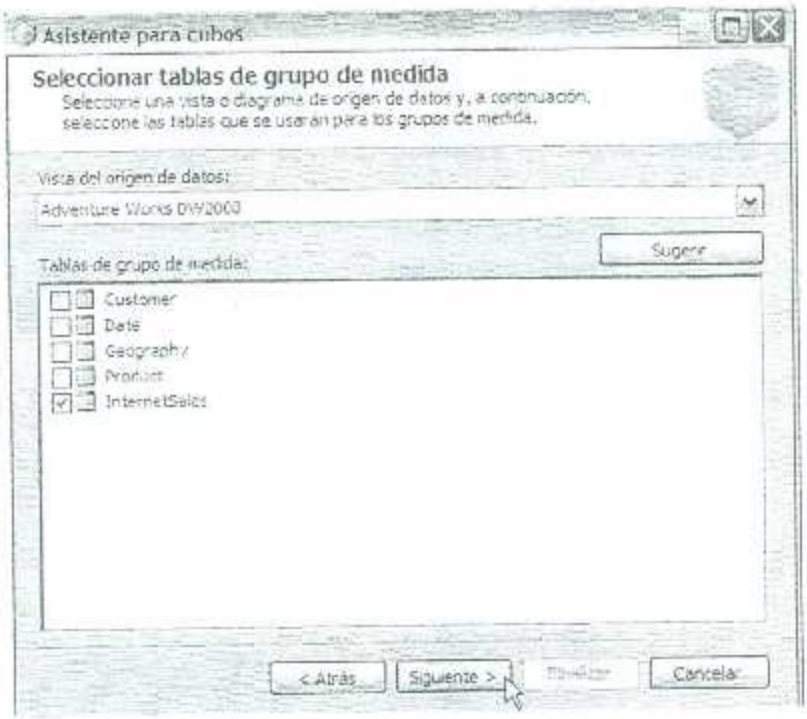


Figura 4.3 – Seleccionar tablas de grupo de medida.

En *Seleccionar medidas*, desactivar las casillas correspondientes a estas medidas: *Promotion Key*, *Currency Key*, *Sales Territory Key*, *Revision Number*.

En *Seleccionar dimensiones existentes* se activa la casilla de *Date*. Luego en *Seleccionar nuevas dimensiones* se activan las casillas de *Customer*, *Product* y *Geography* y se desactiva *Internet Sales*.

Al finalizar el asistente se modifica el nombre del cubo por *Tutorial de Analysis Services*. Figura (4.4)

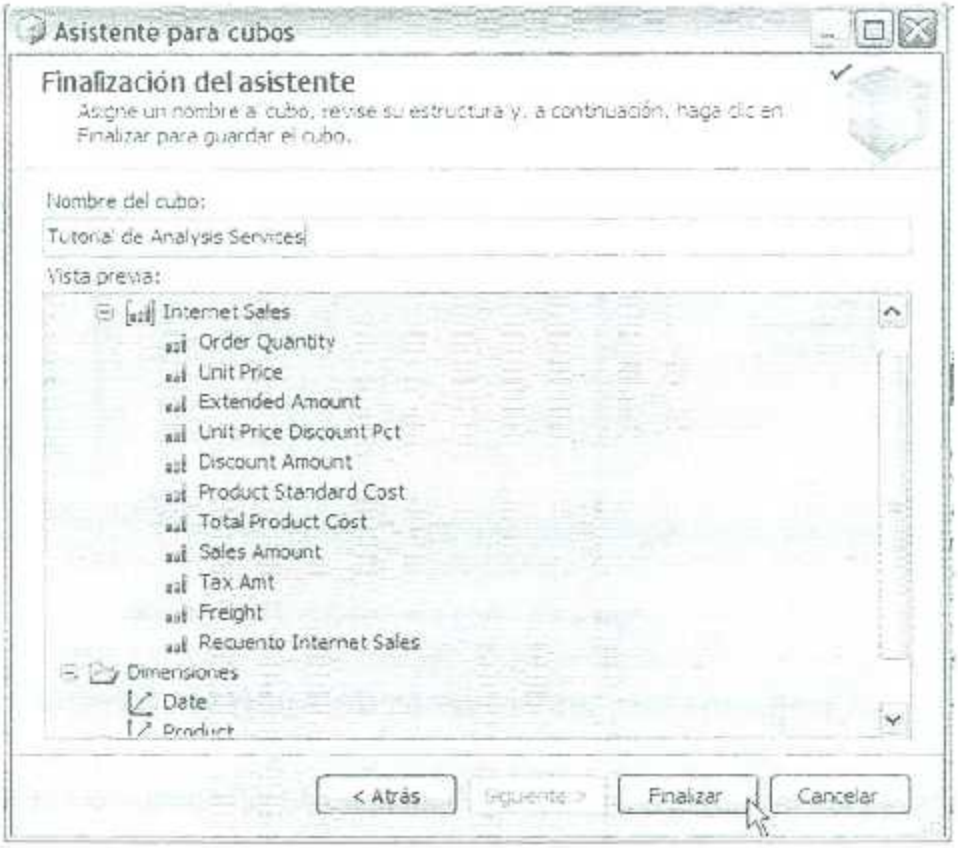


Figura 4.4 – Finalización del asistente para cubos.

4.1.3 Agregar Atributos a Dimensiones

En este punto se agregan algunos atributos de las tablas *Customer*, *Geography* y *Product* a las dimensiones *Customer* y *Product*, respectivamente. Para hacer esto, deben arrastrarse los atributos especificados desde el panel *Vista de origen de datos* hasta el panel *Atributos*. Figura (4.5)

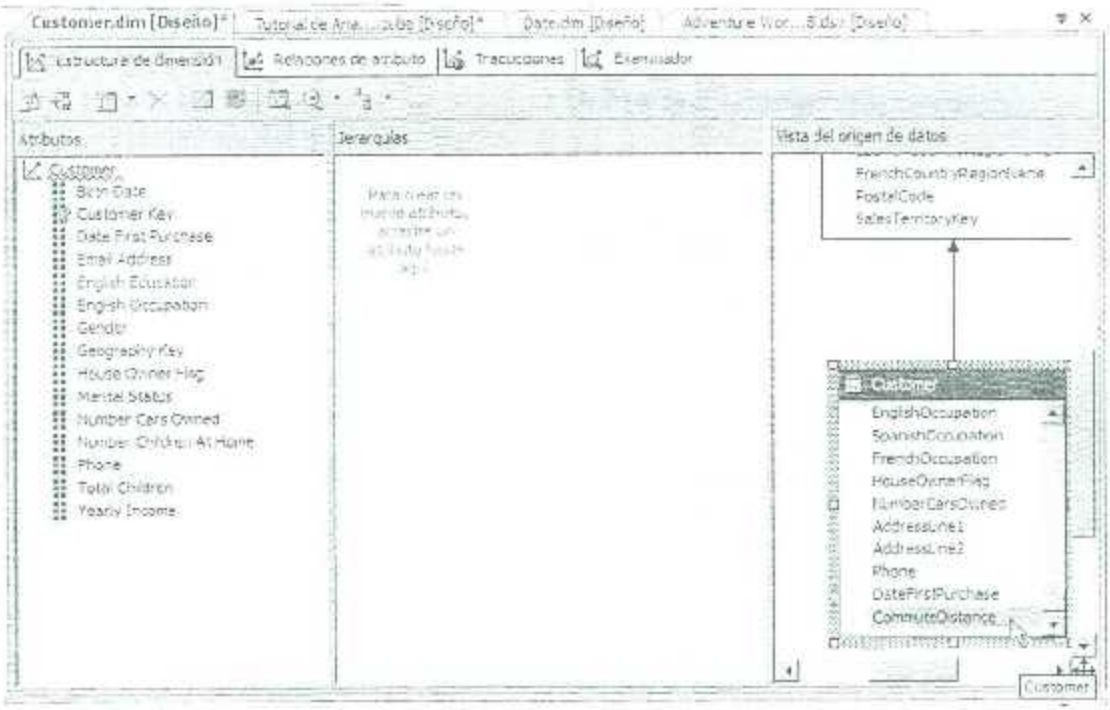


Figura 4.5 – Agregar atributos a dimensiones.

4.1.4 Implementar un Proyecto de Analysis Services

Para ver los datos de dimensión y de cubo de los objetos del cubo *Tutorial de Analysis Services* del proyecto *Tutorial de Analysis Services*, se debe implementar el proyecto en una instancia determinada de *Analysis Services* y luego procesar el cubo y sus dimensiones. Al implementar un proyecto de *Analysis Services* se crean y definen objetos en una instancia de *Analysis*

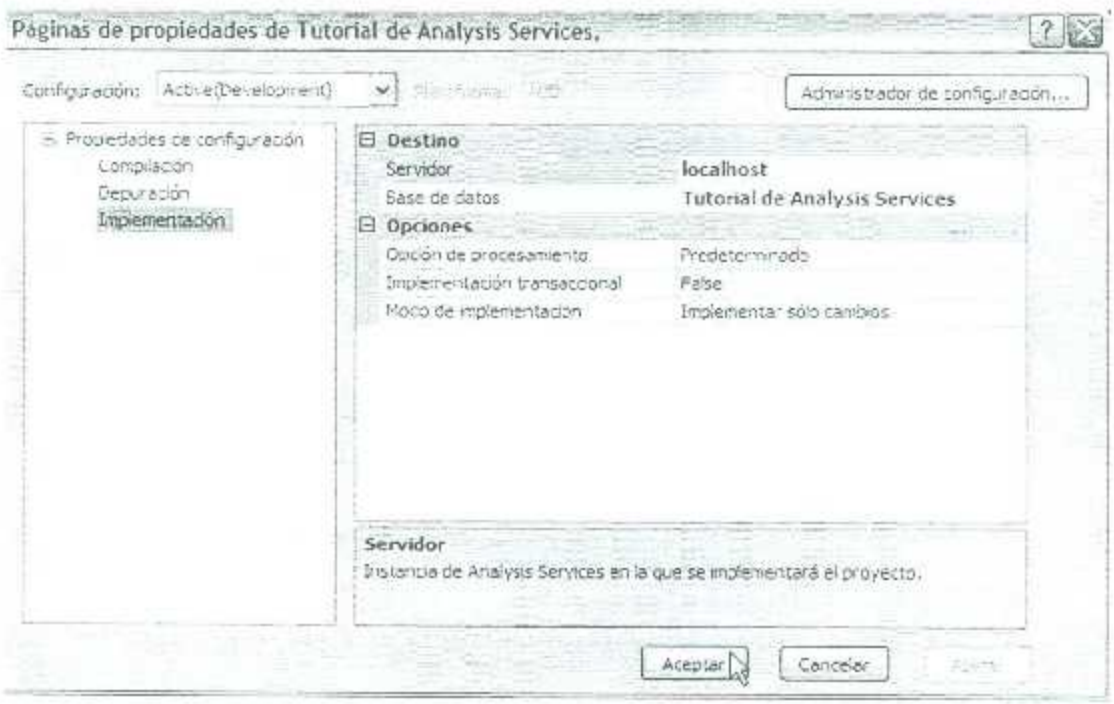


Figura 4.6 – Implementar un proyecto de Analysis Services.

En el *Explorador de soluciones* debe hacerse un clic con el botón secundario en el proyecto *Tutorial de Analysis Services*, y a continuación un clic sobre la opción del menú *Implementar*. Figura (4.7)

Figura 4.7 – Implementar un proyecto de Analysis Services.

Al finalizar la implementación, una ventana de resultados será mostrada. Figura (4.8)

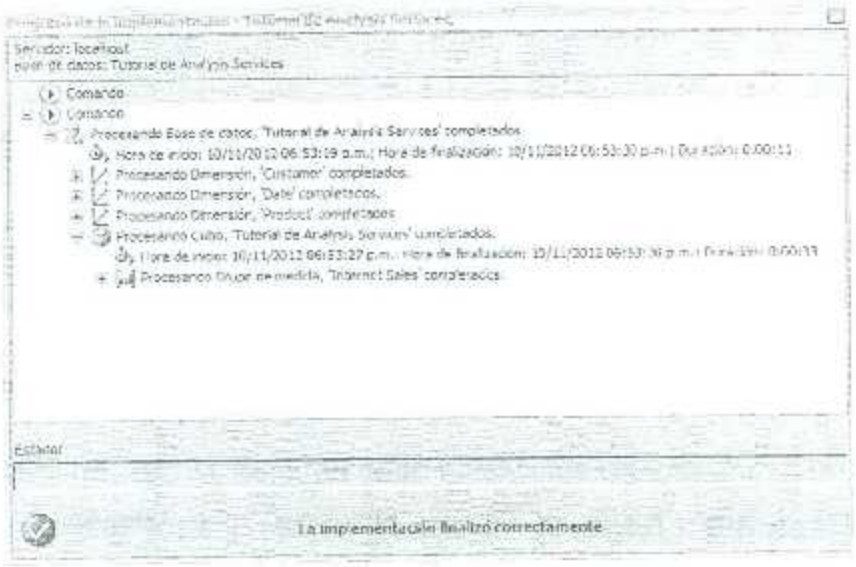


Figura 4.8 – Finalización de la implementación.

Finalmente, una vez implementado el proyecto, los datos de éste pueden verse dando doble clic en el cubo *Tutorial de Analysis Services* en el nodo *Cubos* del *Explorador de soluciones*. Se debe seleccionar la ficha *Examinador* y hacer un clic sobre el icono *Volver a conectar* en la barra de herramientas del diseñador. En el lado derecho de la ficha *Examinador* hay dos paneles: el superior es el panel *Filtro* y el inferior es el panel *Datos*. Figura (4.9)

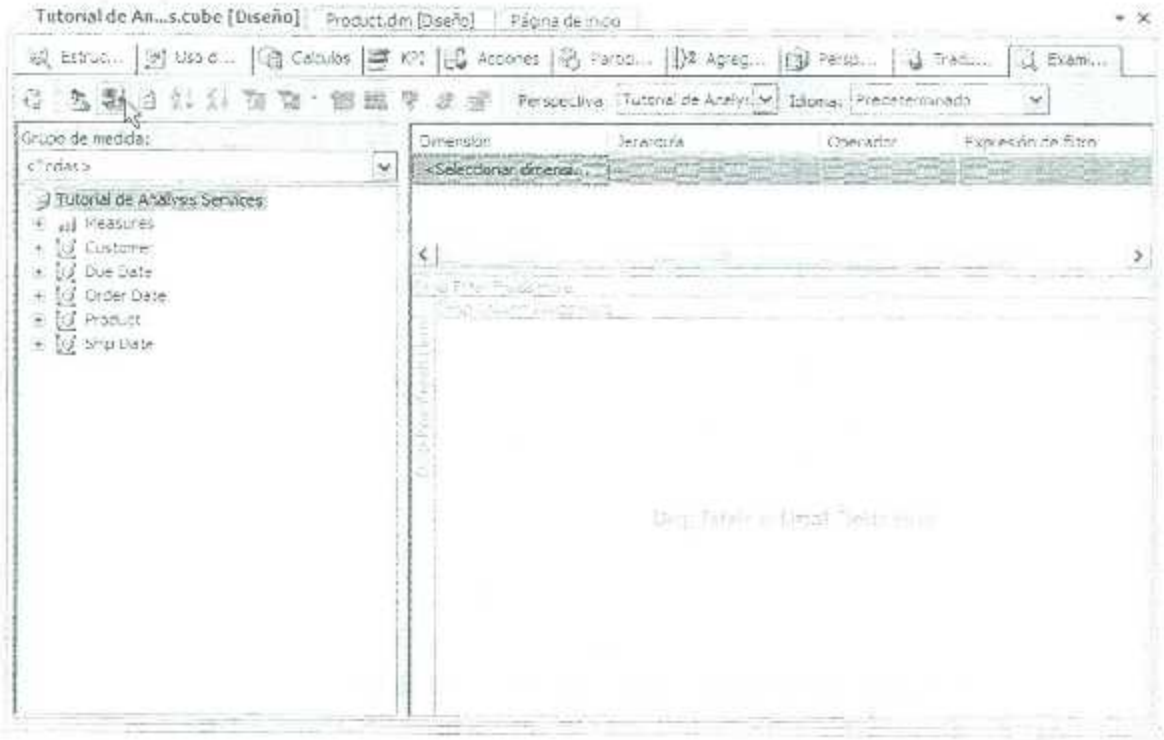


Figura 4.9 – Examinación del cubo

4.1.5 Agregar una Jerarquía

Luego de realizar algunas modificaciones a ciertas medidas para cambiar algunos tipos de formato y ajustar algún nombre, y a ciertas dimensiones del cubo para cambiar algunos nombres de atributos, según las indicaciones del tutorial, se procederá a explicar cómo se agrega una jerarquía.

Las jerarquías se crean en la ficha *Estructura de Dimensión*, arrastrando el atributo deseado desde el panel de *Atributos* hasta el panel *Jerarquías*. Pueden seguirse agregando atributos a niveles inferiores de la jerarquía siguiendo el mismo procedimiento de arrastrado. Figura (4.10)

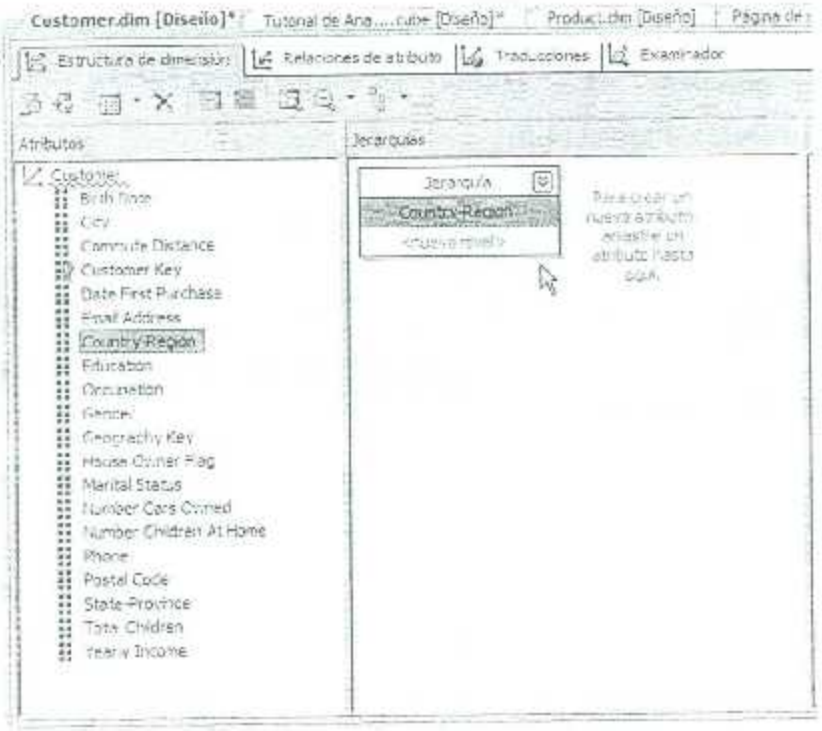


Figura 4.10 – Creación de la jerarquía Country-Region.

4.1.6 Agregar un Cálculo con Nombre

Un cálculo con nombre es una expresión SQL representada como columna calculada en una tabla de la vista de origen de datos. Los cálculos con nombre permiten ampliar el esquema relacional de las tablas existentes de la vista del origen de datos sin modificar la tabla en el origen de datos subyacente.

Para agregar un cálculo con nombre debe abrirse la vista del origen de datos *Adventure Works DW 2008* desde el *Explorador de Soluciones*.

Luego, en el panel *Tablas* hacer clic con el botón secundario en *Customer* y luego en *Nuevo cálculo con nombre*. En el cuadro de diálogo *Crear cálculo con nombre*, debe escribirse *FullName* en el cuadro *Nombre de columna* y, a continuación, escribir una instrucción *CASE* en el cuadro *Expresión*. Figura (4.11).

Figura 4.11 – Crear cálculo con nombre.

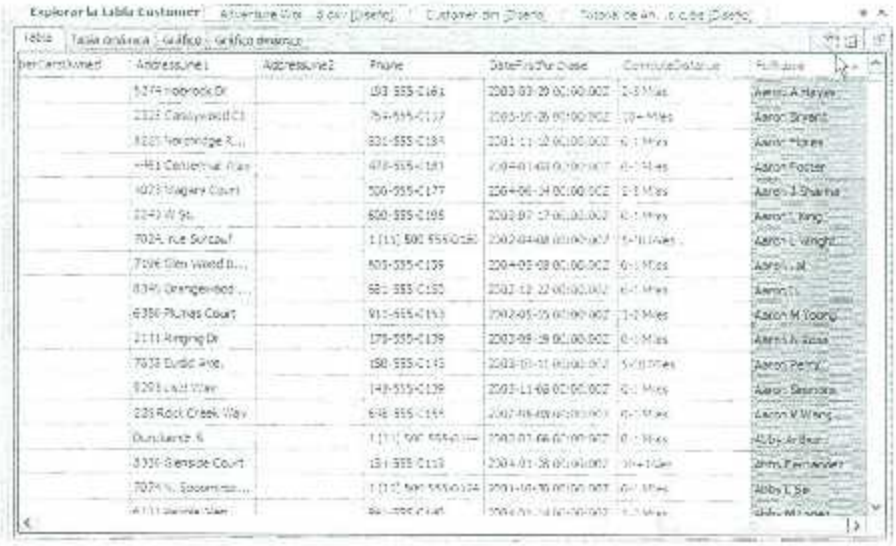


Figura 4.12 – Concatenación de columnas en FullName.

La instrucción *CASE* concatena las columnas: *FirstName, MiddleName* y *LastName*, en una única columna que se utilizará en la dimensión *Customer* como nombre mostrado para el atributo *Customer.* Figura (4.12).

Por último, el nuevo atributo *FullName* de *Customer* debe arrastrarse hasta un nuevo nivel en el panel de *Jerarquías*.

4.1.7 Definir una Relación de Atributo

Previamente realizadas las definiciones de carpetas para mostrar, así como las columnas llave, tal como lo propone el tutorial, se definirán algunas relaciones de atributo. Una relación de atributo permite acelerar el procesamiento de las dimensiones, las particiones y las consultas.

Para definir una relación de atributo debe hacerse click en la ficha *Relaciones de atributo* en el *Diseñador de dimensiones* para la dimensión *Customer*. Luego, sobre el diagrama, hacer un clic con el botón secundario en el atributo City y seleccionar *Nueva relación de atributo*. Figura (4.13).

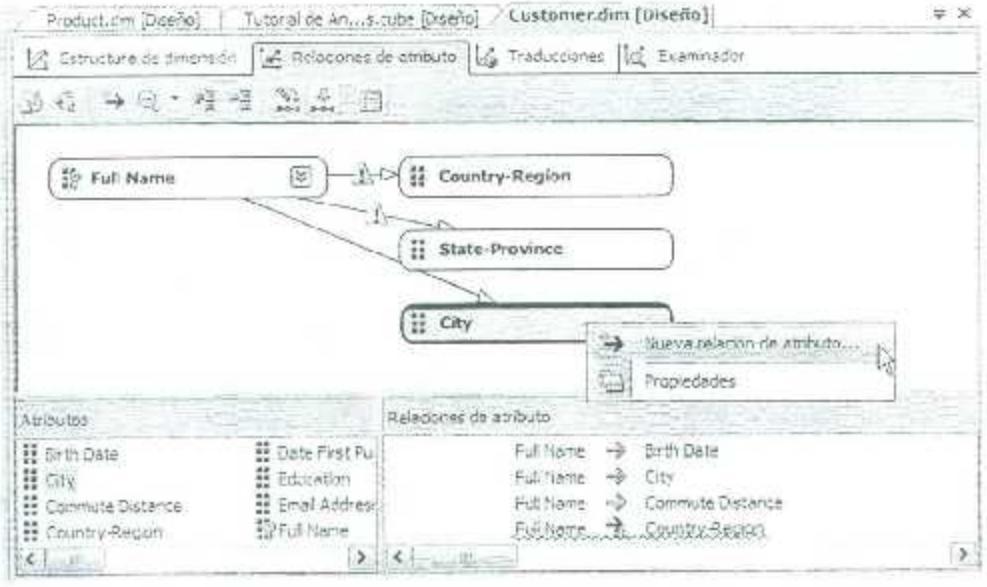


Figura 4.13 – Definición de una nueva relación de atributo.

En el cuadro de diálogo *Crear relación de atributo*, para el atributo de origen *City*, establecer *State-Province*, como atributo relacionado y en la lista tipo de relación, establecer *Rígida*, éste tipo de relación se establece cuando el nexo existente entre ciertos miembros no es susceptible a cambios en el curso del tiempo. Figura (4.14).

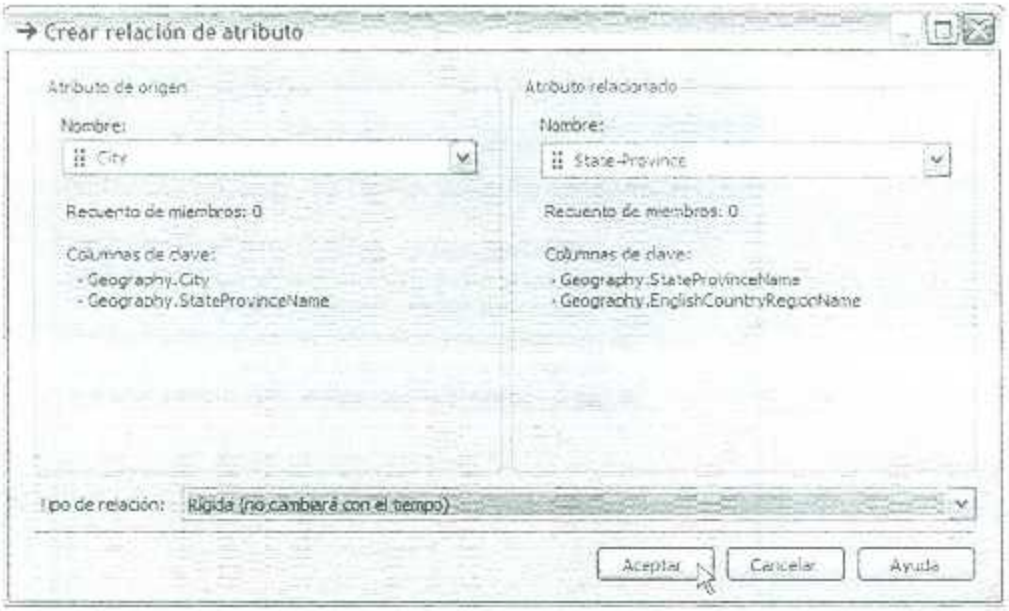


Figura 4.14 – Definición de un atributo relacionado y el tipo de relación.

Finalmente, el tutorial propone establecer otro tipo de relación *Rígida* entre el atributo de origen *State-Province* y el atributo relacionado *Country-Region*.

4.1.8 Implementación de Cambios

Una vez modificados los atributos y las jerarquías, deben implementarse dichas modificaciones y procesarse de nuevo los objetos relacionados antes de ver los cambios. En el menú *Generar* de *BI Development Studio*, debe darse un clic en *Implementar Analysis Services Tutorial* Figura (4.15). Al terminar la implementación se desplegará un mensaje informando el resultado de la misma. Figura (4.16).

Figura 4.15 – Ejecución de la implementación.

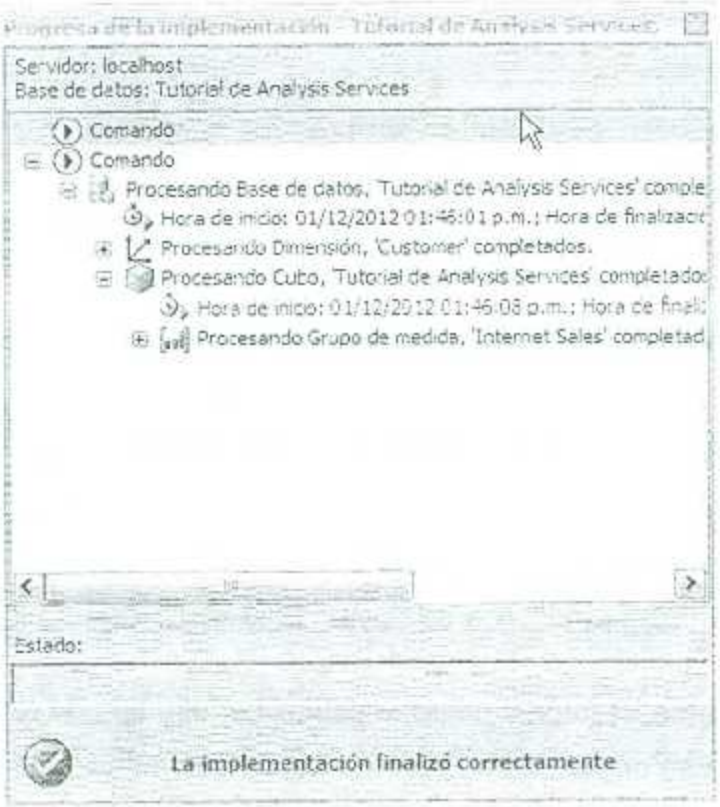


Figura 4.16 – Resultado de la implementación.

En la pestaña *Examinador* del *Diseñador de dimensiones*, pueden observarse algunos cambios en la jerarquía *Customer Geography* de la dimensión *Customer*. Al expandirse el árbol muestra una lista de países, con algunas de sus ciudades y los nombres de los clientes en cada una de éstas. Figura (4.17).

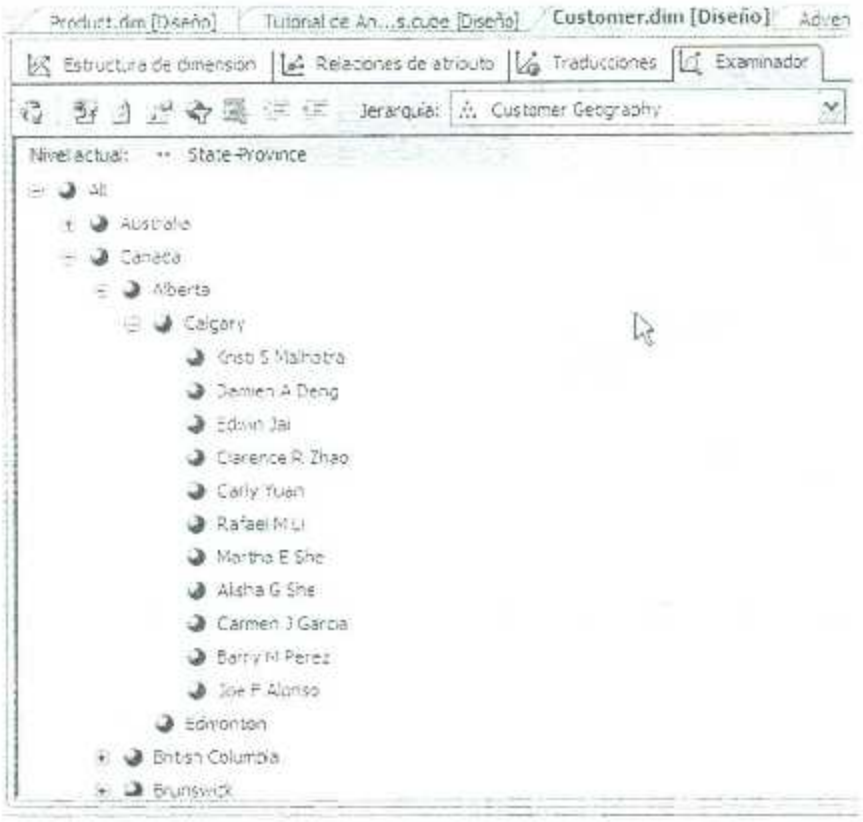
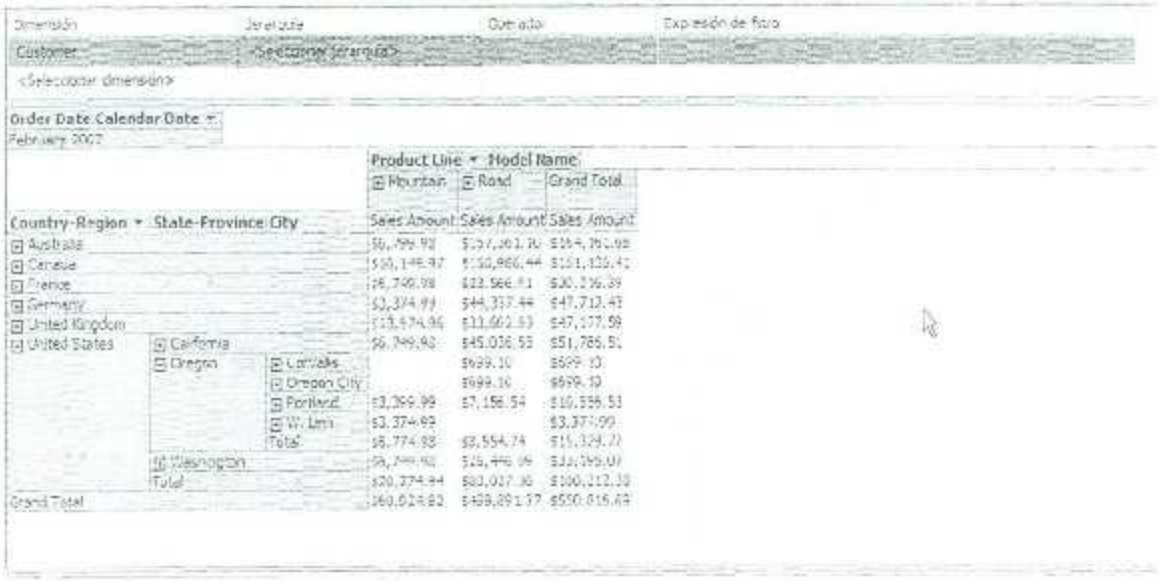


Figura 4.17 – Lista de clientes de la jerarquía Customer Geography del atributo Customer.

Realizando todas las modificaciones propuestas por el tutorial, relacionadas con las dimensiones *Product* y *Date*, agregando jerarquías, cálculos con nombre, relaciones de atributo y sus respectivas implementaciones a los cambios, tal como se explicó en los puntos anteriores para la dimensión *Customer*, es posible examinar el cubo implementado a través de la pestaña Examinador.

Arrastrando las medidas y las jerarquías propuestas en el tutorial, desde el panel metadatos hasta las diferentes áreas del panel de datos, es posible mostrar las ventas por línea de productos, para cada región, realizadas a través de Internet, en el mes de Febrero del 2012 [10] Figura (4 18)



Dimensión | Jerarquía | Operador | Expresión de filtro

Customer | <Seleccionar jerarquía>

<Seleccionar dimensión>

Order Date.Calendar Date

February 2002

Country-Region	State-Province	City	Product Line: Mountain — Sales Amount	Product Line: Road — Sales Amount	Grand Total — Sales Amount
Australia			$6,799.98	$157,361.70	$164,161.68
Canada			$10,149.97	$150,866.44	$161,016.41
France			$8,749.98	$21,566.41	$30,316.39
Germany			$2,374.99	$44,337.44	$47,712.43
United Kingdom			$13,574.96	$33,602.63	$47,177.59
United States	California		$6,749.98	$45,036.53	$51,786.51
	Oregon	Corvallis		$699.10	$699.10
		Oregon City		$699.10	$699.10
		Portland	$3,399.99	$7,156.54	$10,556.53
		W. Linn	$3,374.99		$3,374.99
		Total	$6,774.98	$8,554.74	$15,329.72
	Washington		$6,749.98	$26,446.06	$33,196.04
	Total		$20,274.94	$80,037.30	$100,312.24
Grand Total			$60,624.83	$409,891.57	$550,816.43

Figura 4.18 – Ventas por productos, para cada región, realizadas a través de Internet.

4.2 Conclusiones

En este capítulo se dio seguimiento, de forma textual y gráfica, a algunas lecciones deliberadamente seleccionadas de un tutorial de Microsoft.

El propósito, como fue anticipado en la introducción al capítulo, fue la puesta en práctica de la creación e implementación de un cubo OLAP. Como es de suponer, implícitamente fueron tratadas las funcionalidades principales para la definición y el manejo de los datos del cubo, finalizando con la visualización de la información de ejemplo desde el examinador de *Business Intelligence Development Studio*.

Con esta práctica se pretendió sintetizar, de la forma más clara y simple posible, la esquematización de las propiedades más importantes de una base de datos multidimensional.

La amplia extensión del tutorial motivó una condensación selectiva de las lecciones, según el orden de importancia, previamente consensuado entre el asesor de esta tesis y el tesista, en función de los objetivos planteados desde el comienzo de esta tesis.

CAPÍTULO 5

Diseño e Implementación de Modelos de Minería de Datos

Introducción

La finalidad del último capítulo de esta tesis, es ejemplificar, a través de la puesta en práctica de las lecciones de un tutorial de Microsoft, acerca de la minería de datos, el uso de los procedimientos más significativos, con fines de iniciación, de este campo de las ciencias de la computación.

Como un breve antecedente a este tema, se considera menester conceptualizar que el objetivo general del proceso de la minería de datos consiste en extraer información de un conjunto de datos y transformarla, para su uso posterior, en una estructura comprensible.

La minería de datos recurre a varias herramientas de análisis de datos para encontrar relaciones entre los mismos o patrones de comportamiento que hagan posible la realización de predicciones.

Como puede deducirse, la minería de datos desempeña un rol principal dentro de la inteligencia de negocios, pues su facultad predictiva la convierte en un instrumento de asesoría muy valioso para la toma de decisiones empresariales.

Con la finalidad de evitar una extensión innecesaria del capítulo, han sido obviados algunos aspectos técnicos preliminares que pueden deducirse de los capítulos anteriores.

Al igual que en el capítulo anterior, para el desarrollo de este proyecto se recurrió al *Business Intelligence Development Studio* y a la empresa ficticia: *Adventure Works Cycles*, incluida en las bases de datos de ejemplo de AdventureWorks.

5.1 Generar una Estructura de Distribución de Correo Directo

Después de preparar la base de datos de *Analysis Services*, creando para ello un proyecto de nombre: *ASDataMining2008*, definiendo un origen de datos a *AdventureWorks DW2008* y realizando una vista a éste mismo de nombre: *Targeted Mailing*, seleccionando las tablas: *ProspectiveBuyer* (tabla de compradores probables de una bicicleta) y *vTargetMail* (vista de datos históricos sobre los compradores de una bicicleta en el pasado), se procederá a realizar una lección en la cual se partirá de la suposición que el departamento de marketing de *Adventure Works Cycles* desea aumentar las ventas dirigiendo una campaña de correo directo a clientes específicos. La base de datos de la empresa, *AdventureWorks DW2008*, contiene una lista de clientes antiguos y otra de clientes nuevos potenciales. Mediante el análisis de los atributos de compradores anteriores de bicicletas, la empresa espera detectar los patrones que posteriormente se aplicarán a los clientes potenciales. La empresa pretende utilizar los patrones detectados para predecir qué clientes potenciales tienen más probabilidades de comprar una bicicleta de *Adventure Works Cycles*.

5.1.1 Crear el Modelo de Distribución de Correo Directo

Para crear una estructura, en el *Explorador de Soluciones*, debe darse un clic con el botón secundario en *Estructuras de Minería de Datos* y seleccionar: *Nueva Estructura de Minería de datos*, para iniciar el Asistente de instalación. Posteriormente, comprobar que la opción: *A partir de una base de datos relacional o un almacén de datos* esté seleccionada y, a continuación, hacer clic en *Siguiente*. En la página *Crear la estructura de minería de datos*, seleccionar *Árboles de decisión de Microsoft* y luego dar un clic en *Siguiente*. Figura (5.1)

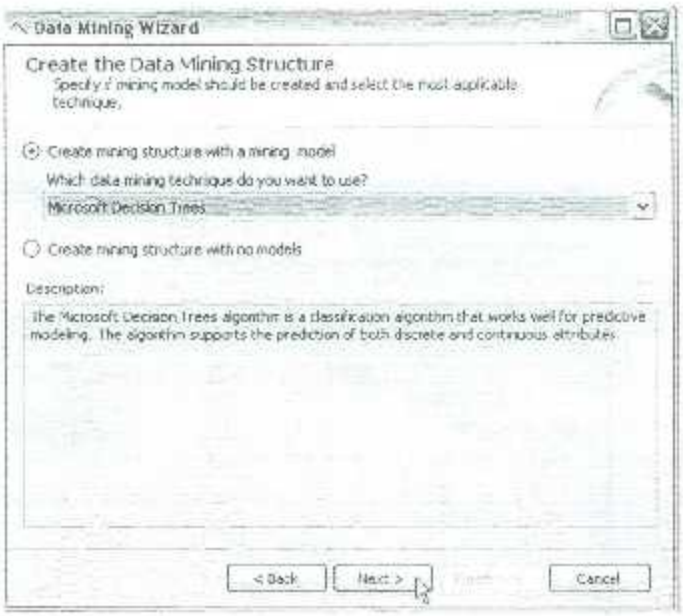


Figura 5.1 – Seleccionar la técnica de Minería de Datos.

En la página *Seleccionar vista del origen de datos*, en el panel *Vistas del origen de datos disponibles*, seleccionar ***Targeted Mailing***., luego dar un clic en *Siguiente*. Figura (5.2)

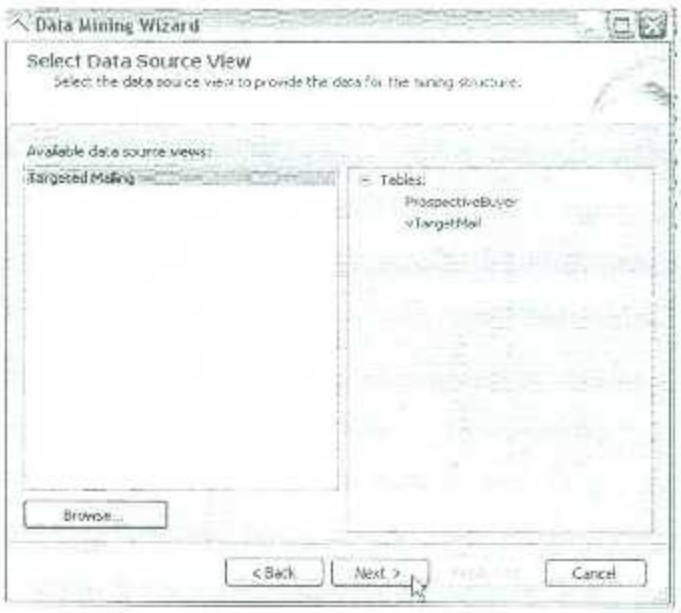


Figura 5.2 – Seleccionar la vista del origen de datos.

En la página *Especificar tipos de tablas*, activar la casilla de la columna *Caso*, correspondiente a *vTargetMail* para usarla como tabla de casos y, a continuación, dar un clic en *Siguiente*. Figura (5.3)

Figura 5.3 – Especificar tipos de tablas

En la página *Especificar los datos de aprendizaje* debe activarse la casilla de la columna *De predicción* en la fila *BikeBuyer*, la casilla de la columna *Key* en la fila *CustomerKey*, y las casillas de la columna *Entrada* en las filas siguientes: *Age, CommuteDistance,EnglishEducation, EnglishOccupation, Gender,GeographyKey,HouseOwnerFlag, MaritalStatus, NumberCarsOwned, NumberChildrenAtHome, Region, TotalChildren, YearlyIncome*.

En la columna izquierda de la página, deben activarse las columnas siguientes:*AddressLine1, AddressLine2,DateFirstPurchase, EmailAddress, FirstName, LastName*. Estas columnas se agregarán a la estructura pero no se incluirán en el modelo. Sin embargo, una vez generado el modelo, estarán disponibles para la obtención de detalles y las pruebas. Figura (5.4)

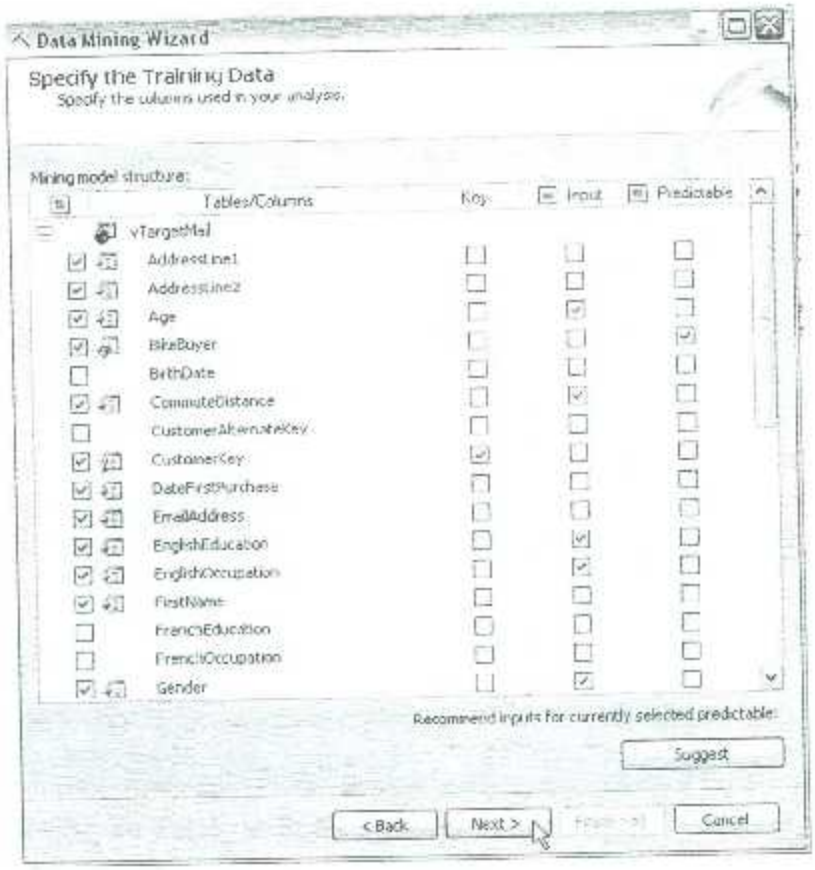


Figura 5.4 – Especificar los datos de aprendizaje.

5.1.2 Especificar el Tipo de Datos y el Tipo de Contenido

En la página *Especificar el contenido y el tipo de datos de las columnas*, hacer clic en *Detectar* para ejecutar un algoritmo que determine los tipos de contenido y los datos predeterminados de cada columna. Normalmente, el asistente detectará números y asignará un tipo de datos numérico adecuado, pero hay varias situaciones en las que se podría desear tratar un número como texto. Tal es el caso de *GeographyKey* que se debería tratar como texto, porque no sería apropiado realizar operaciones matemáticas en este identificador. Luego dar un clic en *Siguiente*. Figura (5.5)

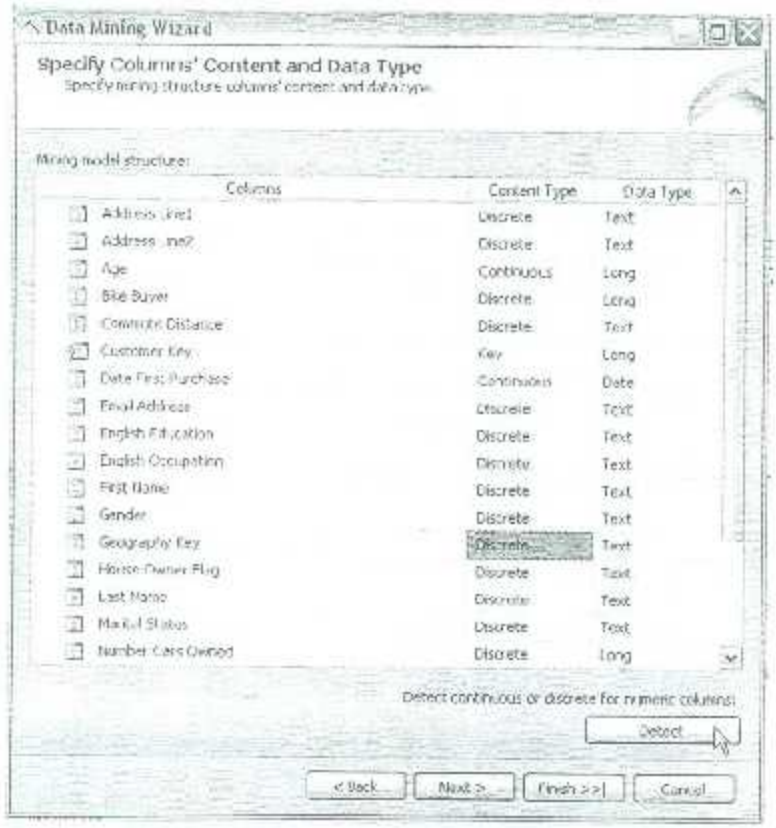


Figura 5.5 – Especificar el tipo de datos y de contenido

5.1.3 Especificar un Conjunto de Datos de Pruebas para la Estructura

Al separar los datos en conjuntos de entrenamiento y de pruebas cuando se crea una estructura de minería de datos, es posible evaluar inmediatamente la exactitud de los modelos de minería de datos que se crean después.

Para especificar el conjunto de pruebas, en *Porcentaje de datos para pruebas*, dejar el valor predeterminado 30, y En *Número máximo de casos en el conjunto de datos de prueba*, escribir 1000. Luego dar un clic en *Siguiente*. Figura (5.6)

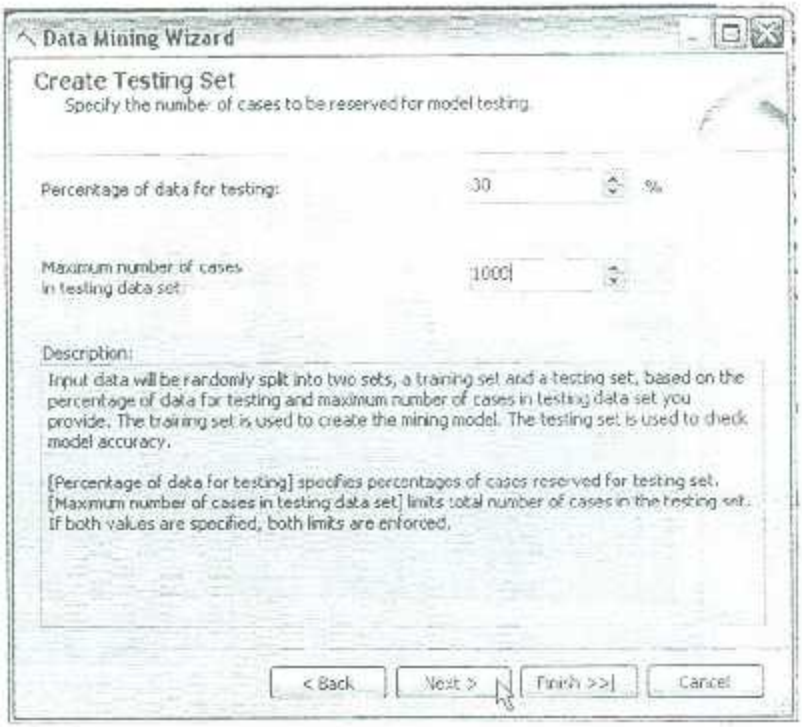


Figura 5.6 – Especificar el conjunto de pruebas

5.1.4 Denominar el Modelo y la Estructura, y Especificar la Obtención de Detalles

En la página *Finalización del asistente*, en *Nombre de la estructura de minería de datos*, escribir: *Targeted Mailing*. En *Nombre del modelo de minería de datos*, escribir: *TM_Decision_Tree*. Activar la casilla: *Permitir obtención de detalles*.

Puede observarse que sólo son mostradas las columnas seleccionadas como: *Clave, Entrada o De predicción*. Las otras columnas seleccionadas (por ejemplo, AddressLine1) no se utilizan para generar el modelo, pero estarán disponibles en la estructura subyacente y se pueden consultar una vez procesado e implementado el modelo.

Después dar un clic en *Finalizar*. Figura (5.7)

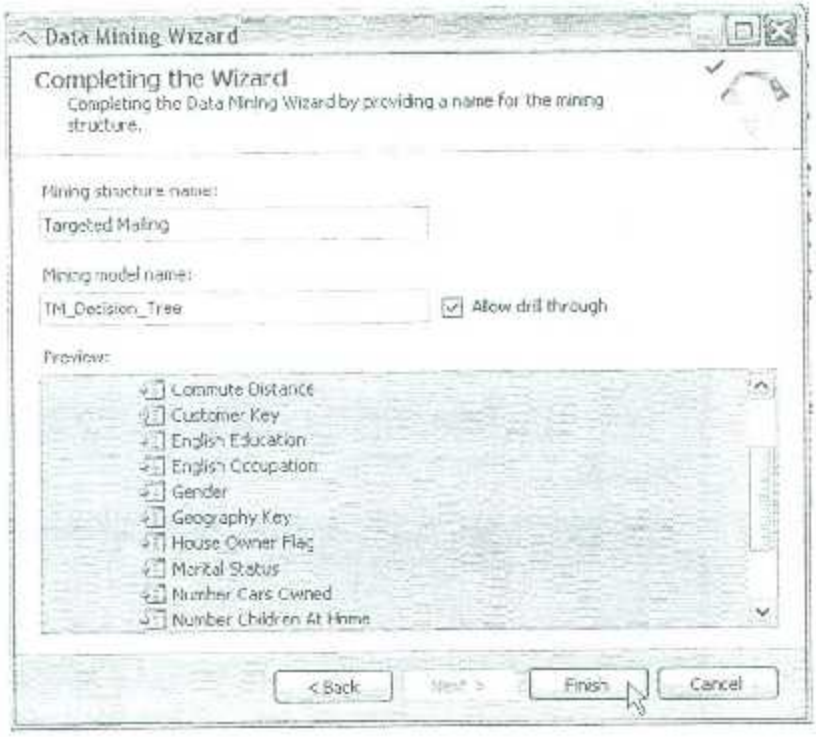


Figura 5.7 – Denominar el modelo y la estructura

5.2 Agregar y Procesar los Modelos

Aquí será creado un conjunto de modelos de minería de datos que sugerirá los clientes que tienen una mayor probabilidad de serlo entre una lista de clientes potenciales.

5.2.1 Agregar Modelos Nuevos a la Estructura de Correo de Destino

El objetivo de este punto es definir dos modelos adicionales mediante la ficha *Modelos de minería de datos* del Diseñador de minería de datos. Para crear los modelos, se usarán el algoritmo Naive Bayes y el algoritmo de clústeres de Microsoft. Estos dos algoritmos se han seleccionado debido a su capacidad de predecir un valor discreto (por ejemplo, la compra de una bicicleta).

Dar un clic sobre la pestaña *Modelos de minería de datos* del Diseñador de minería de datos en Business Intelligence Development Studio. Después hacer clic con el botón secundario en la columna *Estructura* y seleccionar *Nuevo modelo de minería de datos*. En el cuadro de diálogo *Nuevo modelo de minería de datos*, en *Nombre del modelo*, escribir TM_Clustering. En *Nombre del algoritmo*, seleccionar *Agrupación en clústeres de Microsoft*. Dar un clic en *Aceptar*. Figura (5.8).

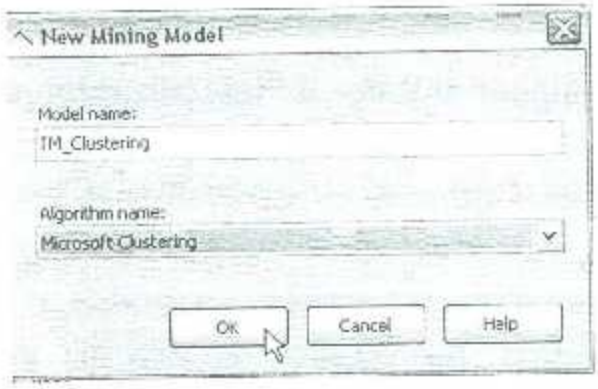


Figura 5.8 – Agregar un nuevo modelo de clústeres

Para crear un modelo de minería de datos Naive Bayes, como en el paso anterior, vuelve a crearse un nuevo modelo de minería de datos. En nombre del modelo escribir *TM_NaiveBayes* y en el nombre del algoritmo seleccionar *Naive Bayes de Microsoft*. Figura (5.9).

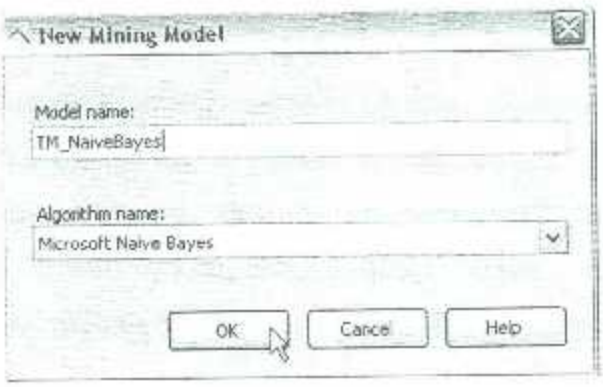


Figura 5.9 – Agregar un nuevo modelo Naive Bayes

5.2.2 Procesar los Modelos de la Estructura de Distribución de Correo Directo

Para examinar o trabajar con los modelos creados, antes debe implementarse el proyecto y procesar la estructura y los modelos de minería de datos. En la implementación se envía el proyecto a un servidor y se crean en éste los objetos del proyecto. El procesamiento es el paso, o la serie de pasos, que rellena los objetos de Analysis Services con datos de orígenes de datos relacionales.

5.2.3 Establecer el Valor de Inicialización de Exclusión

Al implementar un proyecto y procesar la estructura y los modelos, a las filas individuales de la estructura de datos se les asigna de forma aleatoria el conjunto de pruebas y entrenamiento a partir de un valor de inicialización del número aleatorio. Normalmente, el valor de inicialización del número aleatorio se calcula con los atributos de la estructura de datos. Teniendo en cuenta los fines de este tutorial, para asegurarse de que los resultados son los mismos que los descritos aquí, arbitrariamente será asignado un *valor de inicialización de exclusión* fijo de 12. El valor de inicialización de exclusión se utiliza para inicializar el muestreo aleatorio y asegurarse de que los datos se dividen aproximadamente de la misma manera para todas las estructuras de minería de datos y sus modelos.

Para establecer el valor de inicialización de exclusión, debe darse un clic en la pestaña *Estructuras de minería de datos* o *Modelos de minería de datos* en el Diseñador de minería de datos de Business Intelligence Development Studio. Después asegurarse que en el panel de propiedades *CacheMode* esté establecido en *KeepTrainingCases* y escribir *12* en *HoldoutSeed*. Figura (5.10).

Figura 5.10 – Fijar el valor de inicialización de exclusión

5.2.4 Implementar y Procesar los Modelos

En el Diseñador de minería de datos, puede procesar una estructura de minería de datos, un modelo de minería de datos específico que esté asociado a una estructura de minería de datos, o bien procesar la estructura y todos los modelos que estén asociados a esa estructura. Para este caso se procesarán la estructura y todos los modelos al mismo tiempo.

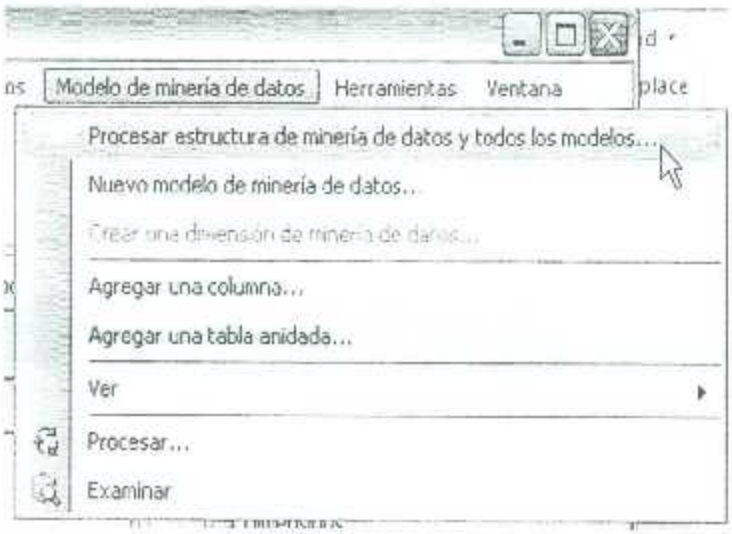


Figura 5.11 – Fijar el valor de inicialización de exclusión

Para implementar el proyecto y procesar todos los modelos de minería de datos, seleccionar *Procesar estructura de minería de datos y todos los modelos*, en el menú *Modelo de minería de datos*. Figura (5.11).

Posiblemente la implementación despliegue un error no mencionado en el tutorial. Figura (5.12).

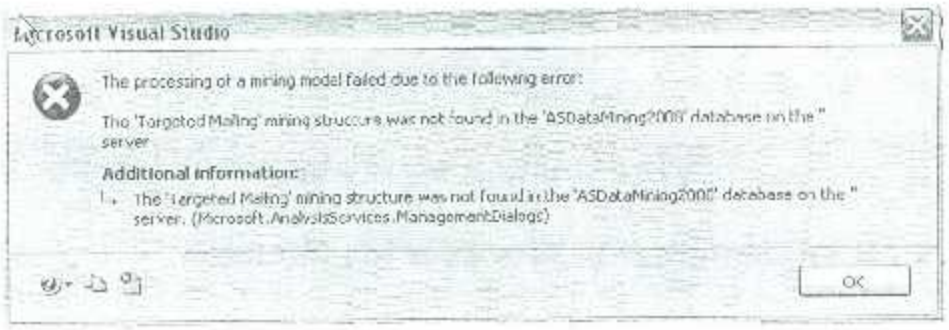


Figura 5.12 – Error en la implementación.

La solución a este problema consiste en abrir el menú *Ver*, seleccionar la opción *Código*, localizar las líneas referidas a: *ddl100_100*, y reubicarlas en otra sección del código como se muestra a continuación. Figura (5.13).

Figura 5.13 – Reubicación de líneas de código.

Una vez realizado este cambio, se repite el paso de implementación del proyecto.

En el cuadro de diálogo *Procesar estructura de minería de datos: Targeted Mailing*, hacer clic en *Ejecutar*. Figura (5.14).

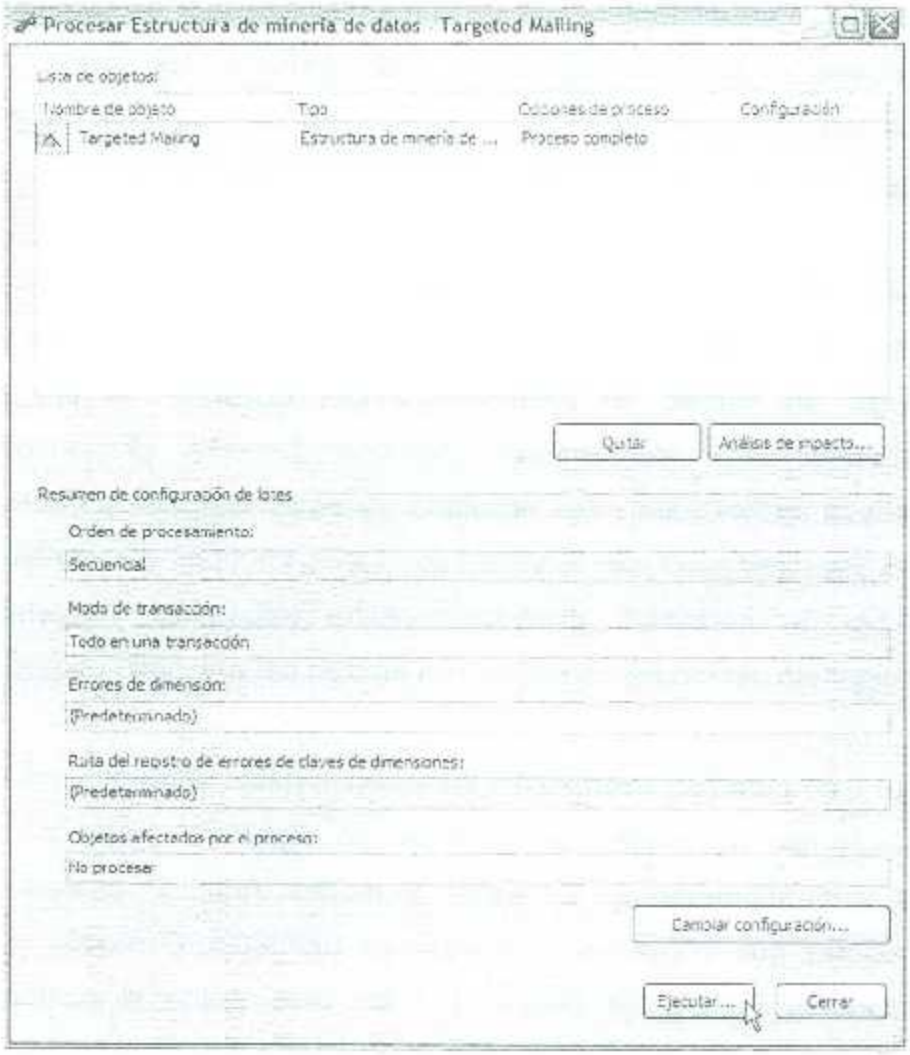


Figura 5.14 – Ejecutar el proceso de estructura de minería de datos.

Hacer clic en *Cerrar* en el cuadro de diálogo *Progreso del proceso* cuando el procesamiento de los modelos se haya completado y clic en *Cerrar* en el cuadro de diálogo *Procesando estructura de minería de datos - <estructura>*.

5.3 Explorar los Modelos de Correo Directo

Una vez procesados los modelos en el proyecto, pueden examinarse en *Business Intelligence Development Studio* para buscar tendencias interesantes. Como los resultados de los modelos de minería de datos son complejos y pueden resultar difíciles de comprender sin formato, examinar los datos visualmente suele ser la manera más fácil de entender las reglas y relaciones que los algoritmos descubren en los datos. Asimismo, la exploración servirá para entender el comportamiento del modelo y detectar qué modelo se comporta mejor antes de su implementación.

Cada modelo creado se muestra en la ficha *Visor de modelos de minería de datos* en el Diseñador de minería de datos. Cada algoritmo utilizado para crear un modelo en Analysis Services devuelve un tipo de resultado diferente. Por consiguiente, Analysis Services proporciona un visor independiente para cada algoritmo. Analysis Services proporciona también un visor genérico que funciona con todos los tipos de modelo. El Visor de árbol de contenido genérico muestra información detallada sobre el contenido del modelo, que varía en función del algoritmo utilizado.

En este punto se examinarán los mismos datos utilizando los tres modelos. Cada tipo de modelo se basa en un algoritmo diferente y proporciona visiones diferentes de los datos. El modelo Árbol de decisión le indica los factores que influyen en la compra de bicicletas. El modelo Agrupación en clústeres agrupa los clientes por atributos, como el comportamiento de compra de bicicletas y otros atributos seleccionados. El modelo Naive Bayes permite examinar las relaciones entre los diferentes atributos. Por último, el Visor de árbol de contenido genérico muestra la estructura del modelo y proporciona datos más detallados como las fórmulas, los patrones extraídos y un recuento de casos en un clúster o árbol determinado.

5.3.1 Explorar el Modelo de Árbol de Decisión

Dado que el modelo de correo directo incluido en este proyecto de tutorial contiene un único atributo de predicción, *Bike Buyer*, sólo se puede ver un árbol. Si hubiera más árboles, podría utilizarse el cuadro *Árbol* para elegir uno diferente.

Para explorar el modelo en la pestaña *Árbol de Decisión*, debe seleccionarse la pestaña *Visor de modelo de minería de datos* en *Diseñador de minería de datos*. Una vez dentro de esta pestaña, se recomienda utilizar los botones de lupa para ajustar el tamaño de la presentación del árbol. Asimismo, el número de niveles mostrados puede ajustarse utilizando el control deslizante *Mostrar nivel* o la lista *Expansión predeterminada*.

Al examinar el modelo *TM_Decision_Tree* en el Visor de árbol de decisión, puede observarse que la edad ocupa el nodo principal en el árbol, lo cual significa que es el factor más importante para la predicción de compra de bicicletas, según este ejemplo.

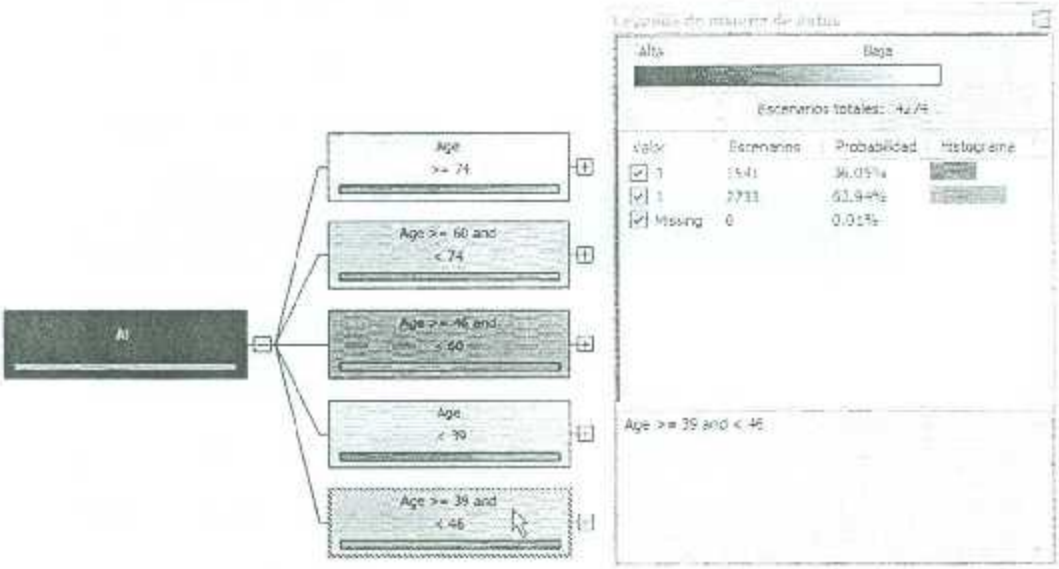


Figura 5.15 – Ejecutar el proceso de estructura de minería de datos

Al hacer clic con el botón secundario sobre un nodo y elegir del menú *Mostrar leyenda*, una ventana desplegará información relacionada al nodo, la cual indica el número de escenarios totales (la suma de los casos en los cuales se han comprado bicicletas, los casos en los cuales no se han comprado bicicletas y los casos con valores faltantes para *Bike Buyer*), el número de casos de personas que no han comprado bicicletas (representados por el número cero y el color azul) , el número de personas que sí han comprado bicicletas (representados por el número uno y el color rosa) y el número de casos, de haberlos, con valores faltantes para la predicción *Bike Buyer*. Figura (5.15).

Puede observarse a simple vista, para el nivel edad, que los compradores con una edad comprendida entre los 39 y 46 años, tienen más probabilidades de comprar una bicicleta. Puesto que el color rosa predomina sobre el azul, aunque debe aclararse que el número de escenarios para otros rangos de edades es mayor al de éste. Figura (5.16).

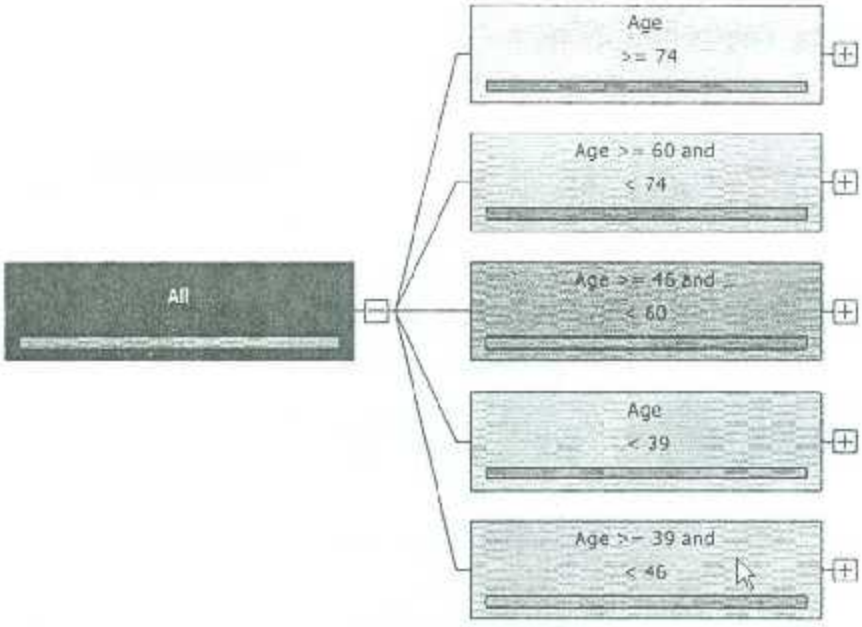


Figura 5.16 – Nodo con mayor probabilidad de compra de bicicletas.

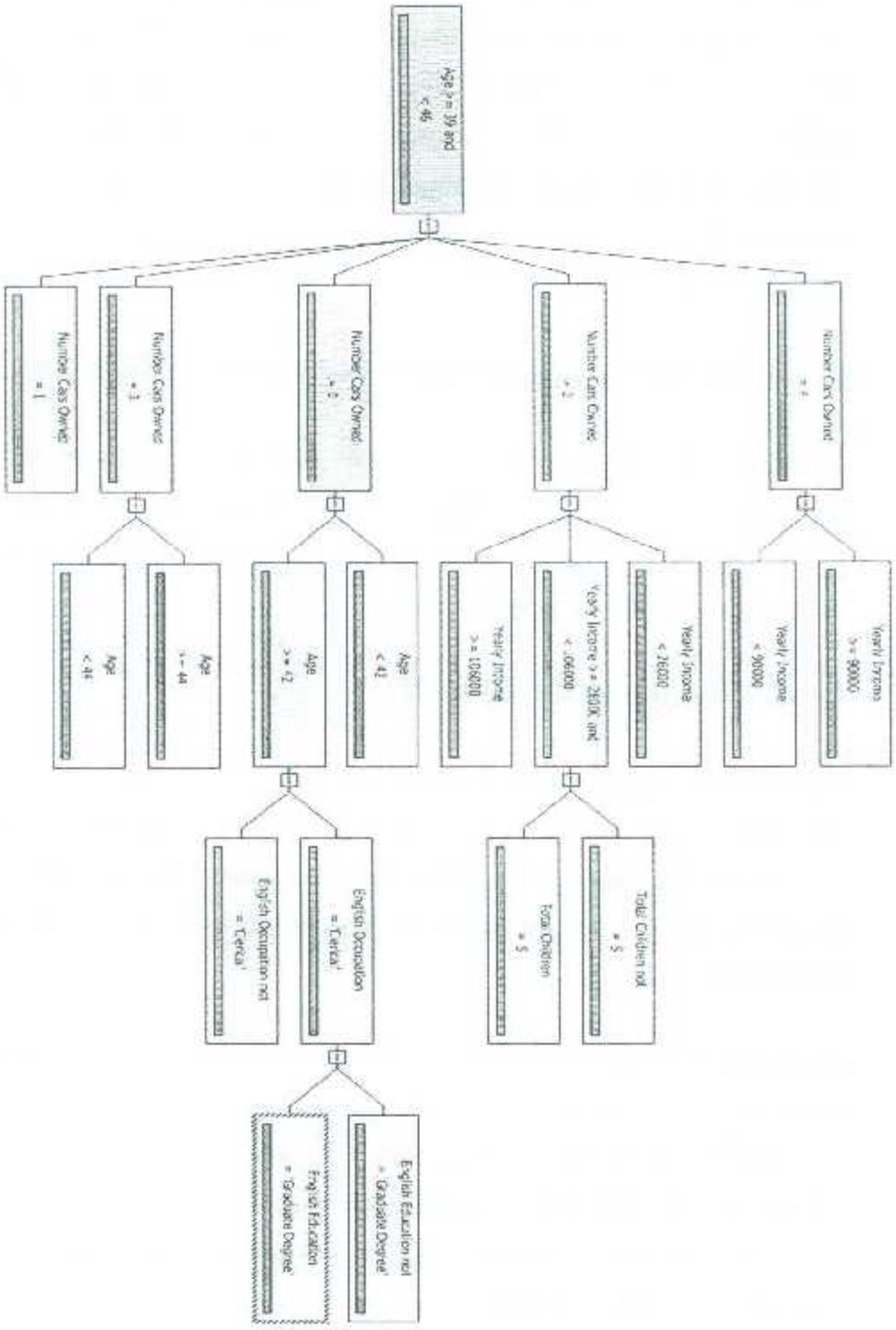


Figura 5.17 – Seguimiento hasta el último nivel del nodo Edad >=39 y <46

Dando un clic en el nodo de *Edad >= 39 y < 46*. Puede observarse que es más probable que los clientes que no cuentan con un automóvil propio o aquellos que cuentan con uno solo, compren una bicicleta. Siguiendo el orden sucesivo de los niveles del nodo de clientes que no cuentan con un automóvil propio, puede observarse que la mayor probabilidad de comprar una bicicleta la representan aquellos clientes cuya edad está comprendida entre los 42 y 45 años, cuya profesión es clérigo y cuentan con un título de posgrado. Figura (5.17).

5.3.2 Explorar el Modelo de Agrupación en Clústeres

El algoritmo de agrupación en clústeres de Microsoft agrupa los casos en clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones.

5.3.2.1 Diagrama del Clúster

La pestaña *Diagrama del clúster* muestra todos los clústeres de un modelo de minería de datos. Las líneas entre los clústeres representan la "proximidad" y aparecen sombreadas en función de la similitud entre los clústeres. El color de cada clúster representa la frecuencia de la variable y el estado del clúster.

Para explorar el modelo en la pestaña *Diagrama del clúster* seleccionar la pestaña *Visor de modelos de minería de datos* para cambiar al modelo TM_Clustering. Luego, en la lista Visor, seleccionar *Visor de clústeres de Microsoft*. En el cuadro *Variable de sombreado*, seleccionar *Bike Buyer*. La variable predeterminada es *Población*, pero puede cambiarse a cualquier atributo del modelo con el fin de determinar los clústeres que contienen miembros con los atributos deseados. Al seleccionar *1* en el

cuadro *Estado* se podrán explorar aquellos casos donde se compró una bicicleta. La leyenda *Densidad* describe la densidad del par de estados del atributo que se selecciona en Variable de sombreado y Estado. En este ejemplo se indica que el clúster con el sombreado más oscuro tiene el porcentaje superior de compradores de bicicleta. Figura (5.18).

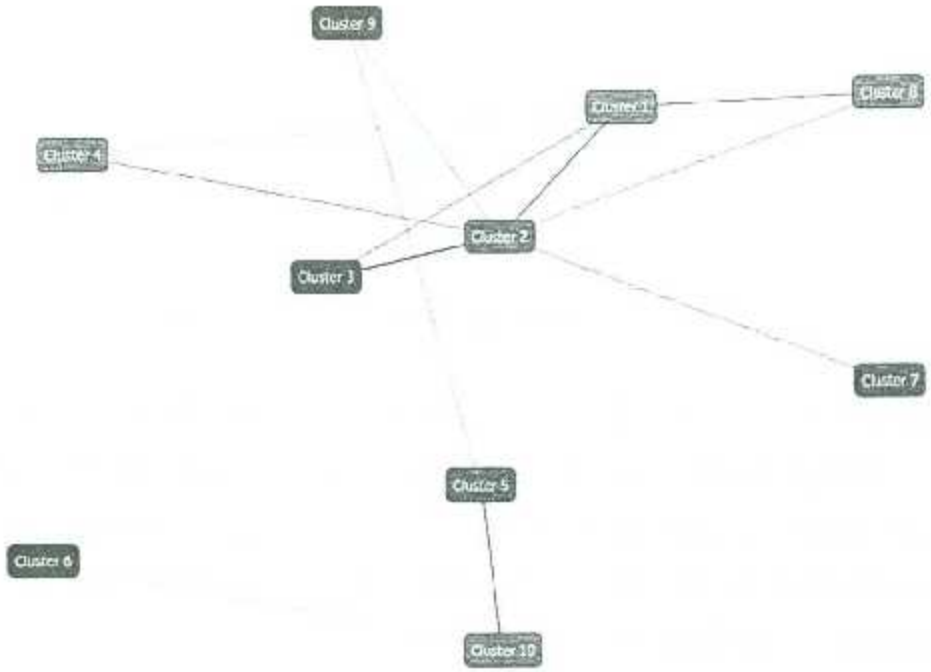


Figura 5.18 – Densidad de compradores según la variable de sombreado BiKe Buyer

Al pausar el mouse sobre el clúster con el sombreado más oscuro, una información sobre herramientas mostrará el porcentaje de casos que tienen el atributo, *Bike Buyer* = 1. Luego debe elegirse el clúster con mayor densidad, hacer clic con el botón secundario en él, seleccionar *Cambiar nombre de clúster* y luego escribir *Bike Buyers High*, para una identificación posterior. Luego dar clic en Aceptar. Figura (5.19).

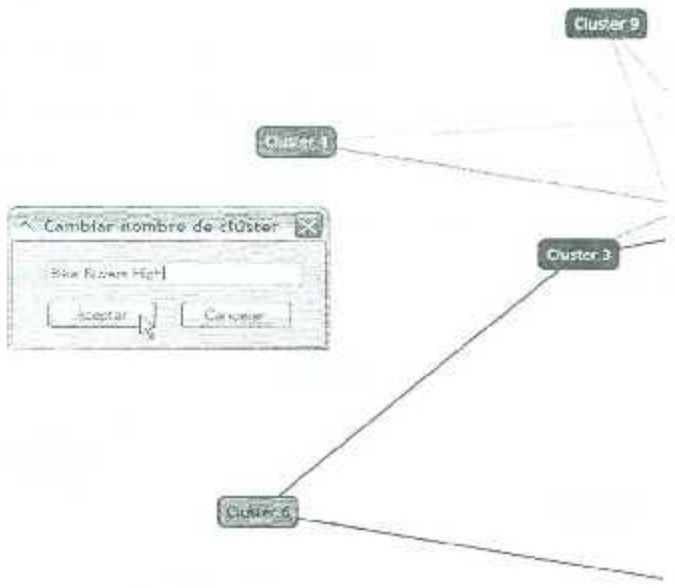


Figura 5.19 – Cambiar nombre de clúster.

Luego debe repetirse el paso anterior, eligiendo ahora el clúster con menor densidad, cambiando su nombre por: *Bike Buyers Low*. Al usar el control deslizante situado en la parte izquierda de la red puede saberse, por la oscuridad de las líneas, la intensidad de las relaciones entre todos los clústeres del diagrama.

5.3.2.2 Perfiles del Clúster

La pestaña *Perfiles del clúster* proporciona una vista global del modelo *TM_Clustering*. Contiene una columna para cada clúster del modelo. La primera columna enumera los atributos asociados a un clúster como mínimo. El resto del visor contiene la distribución de estados de un atributo por cada clúster. La distribución de una variable discreta se muestra como una barra coloreada y el número máximo de barras aparece en la lista *Barras de histograma*. Los atributos continuos se muestran con un diagrama de rombo, que representa la desviación media y estándar en cada clúster.

Para explorar el modelo en la ficha Perfiles del clúster debe establecerse las *Barras de Histograma* en *5*. Luego, deben seleccionarse las columnas *Bike Buyers High* y Bike Buyers Low y arrastrarlas hacia la derecha de la columna Población. Hacer clic en la columna *Bike Buyers High*. Hacer doble clic en la celda *Age*, en la columna *Bike Buyers High*. Hacer clic con el botón secundario en la columna *Bike Buyers Low* y seleccionar *Ocultar columna*. La columna *Variables* está organizada por orden de importancia para ese clúster. Al desplazarse por la columna y revisar las características del clúster Bike Buyer High, es muy probable que en todas ellas, la característica común sea que la distancia al trabajo sea corta. Figura (5.20).

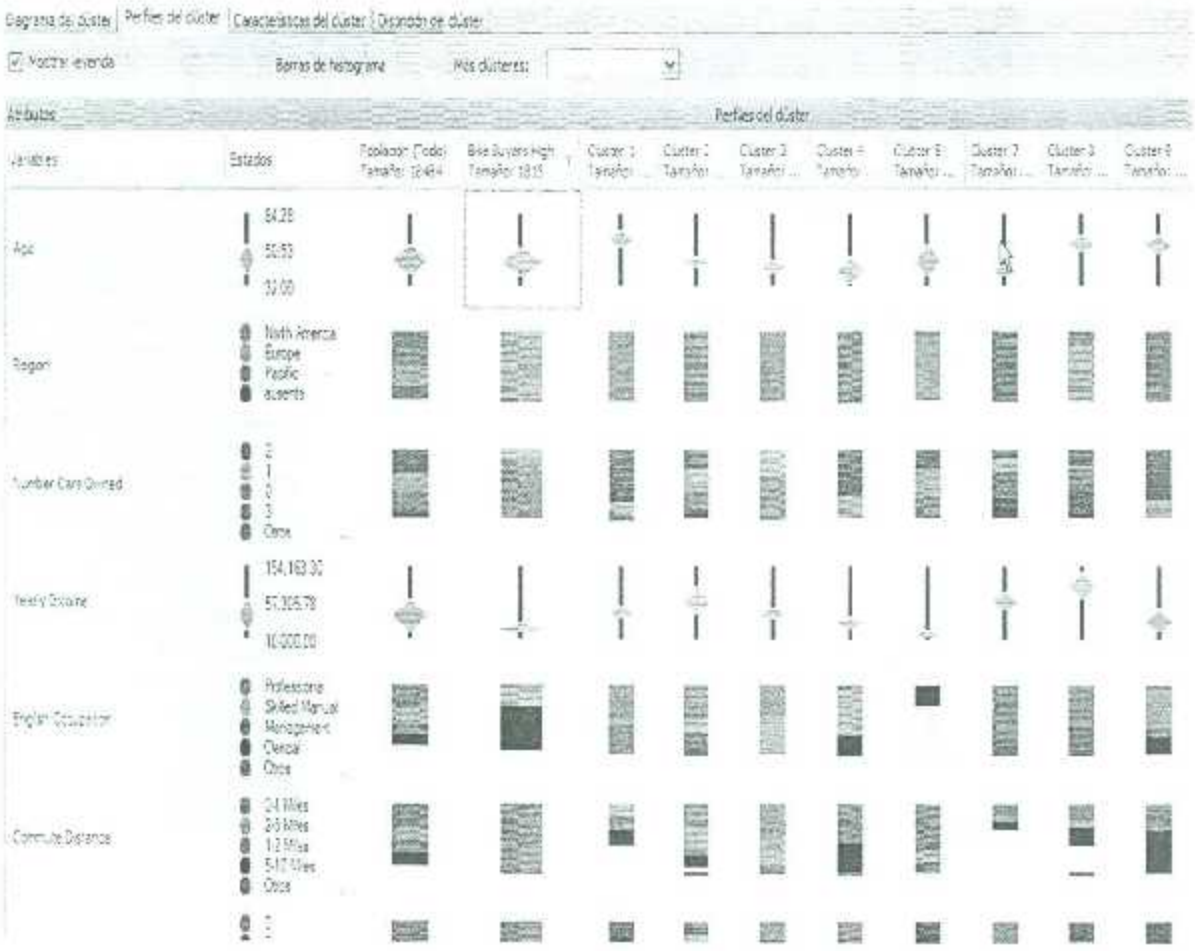


Figura 5.20 – Perfiles del clúster.

5.3.2.3 Características del Clúster

La pestaña *Características del clúster* permite examinar con más detalle las características que forman un clúster. En lugar de comparar las características de todos los clústeres (como en la pestaña Perfiles del clúster), se puede explorar un clúster a la vez. Por ejemplo, si se selecciona *Bike Buyers High* en la lista *Clúster*, pueden verse las características de los clientes en este clúster. Aunque la presentación es diferente del visor Perfiles del clúster, los resultados son los mismos. Figura (5.21).

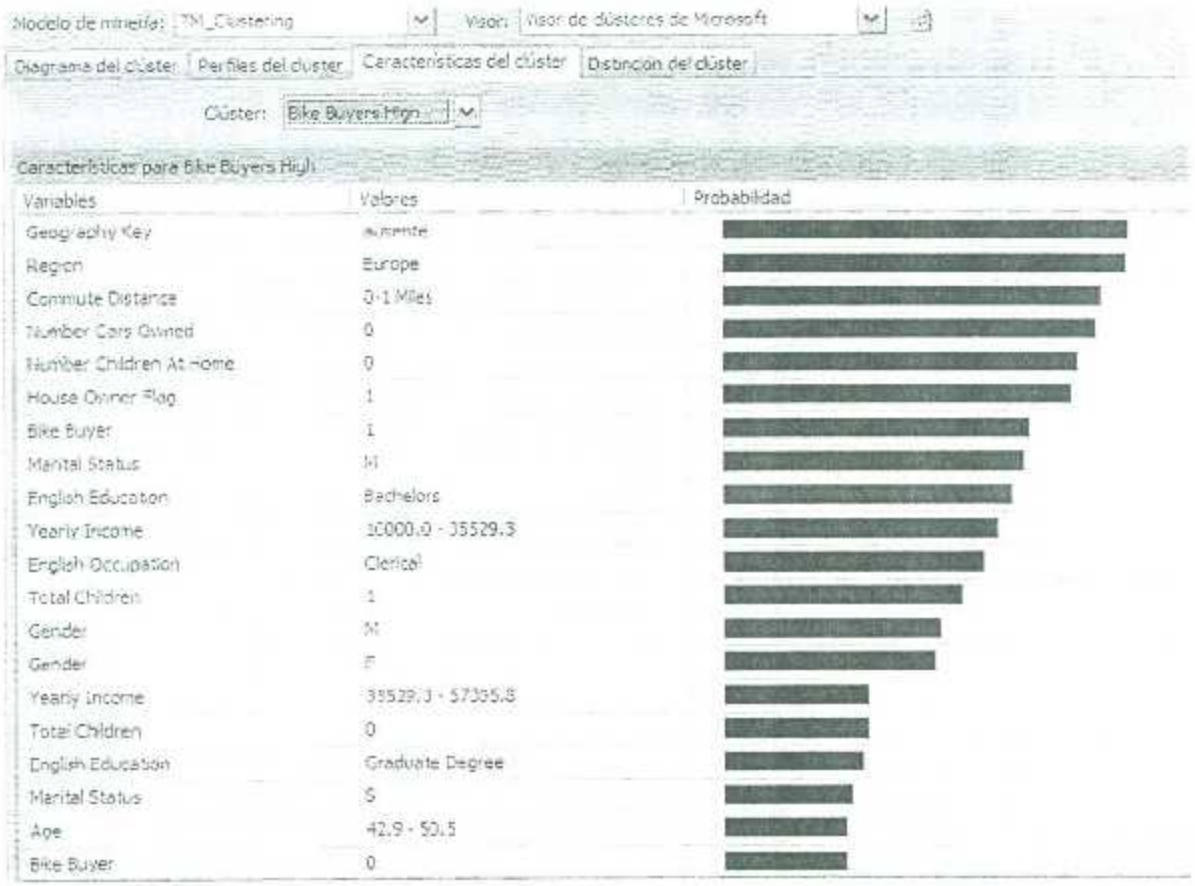


Figura 5.21 – Características del clúster Bike Buyer High.

5.3.2.4 Distinción del Clúster

La pestaña *Distinción del clúster* permite explorar las características que diferencian a un clúster de otro. Después de seleccionar dos clústeres, uno de la lista *Clúster 1* y otro de la lista *Clúster 2*, el visor calcula las diferencias existentes entre los clústeres y muestra una lista de los atributos que más distinguen a los clústeres.

Para explorar el modelo en la pestaña *Distinción del clúster*, en el cuadro *Clúster 1*, seleccionar *Bike Buyers High*. En el cuadro *Clúster 2*, seleccionar *Bike Buyers Low*. *Finalmente*, hacer clic en *Variables* para ordenar alfabéticamente.

Algunas de las diferencias sustanciales entre clientes de los clústeres *Bike Buyers Low* y *Bike Buyers High* son la edad, la posesión de un vehículo, el número de hijos y la región. Figura (5.22).

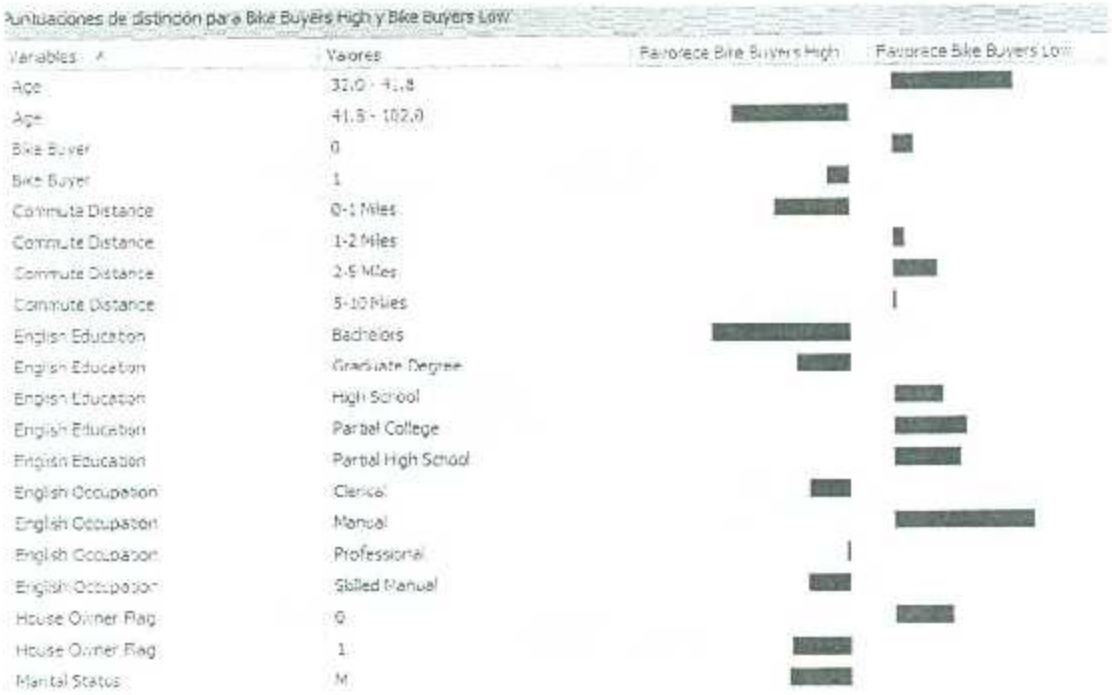


Figura 5.22 – Características que diferencian a un clúster de otro

5.3.3 Explorar el Modelo Naive Bayes

El algoritmo Naive Bayes de Microsoft proporciona varios métodos para mostrar la interacción entre los atributos de entrada y la compra de bicicletas.

5.3.3.1 Red de Dependencias

La pestaña *Red de dependencias* funciona igual que la ficha del mismo nombre del Visor de árboles de Microsoft. Cada nodo del visor representa un atributo y las líneas entre los nodos representan relaciones. En el visor, puede ver todos los atributos que afectan al estado del atributo de predicción, Bike Buyer.

Para explorar el modelo en la pestaña Red de dependencias, utilizar la lista *Modelo de minería de datos* de la parte superior de la pestaña *Visor de modelos de minería de datos* para cambiar al modelo *TM_NaiveBayes*. En la lista *Visor* cambiar a *Visor Bayes naive de Microsoft*. Al hacer clic en el nodo *Bike Buyer* pueden identificarse sus dependencias (El sombreado rosa indica que todos los atributos influyen en la compra de bicicletas). Figura (5.23)

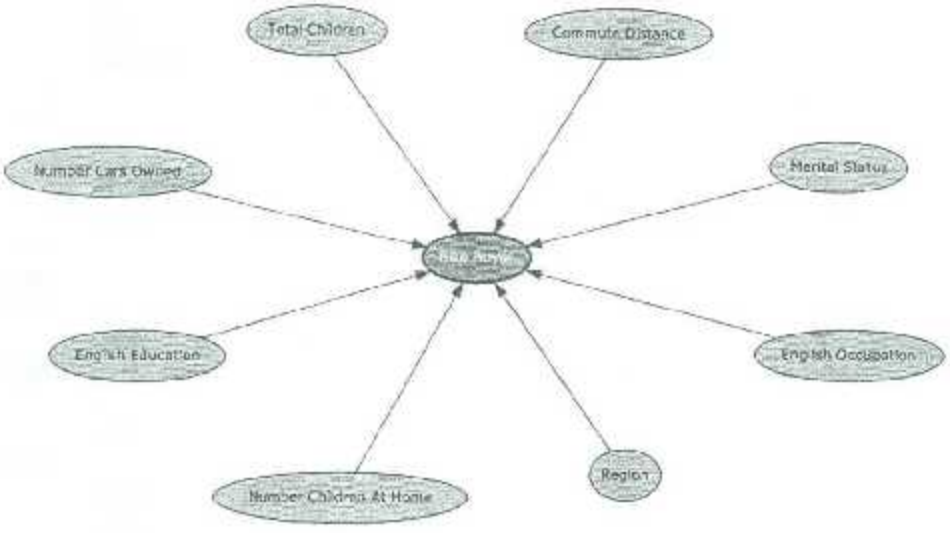


Figura 5.23 – Explorando el modelo en red de dependencias.

5.3.3.2 Perfiles del Atributo

La pestaña *Perfiles del atributo* describe la forma en que los diferentes estados de los atributos de entrada afectan al resultado del atributo de predicción.

Para explorar el modelo en la pestaña Perfiles del atributo, en el cuadro *De predicción,* comprobar que se ha seleccionado *Bike Buyer* En el cuadro de barras Histograma, seleccionar 5. En la columna Atributos, buscar *Number Cars Owned*. Puede Observarse las diferencias en los histogramas de los compradores de bicicletas (la columna con la etiqueta 1) y los no compradores (la columna con la etiqueta 0). Quienes no tienen automóvil o tienen uno solo, tienen mucha más probabilidad de comprar una bicicleta. Al hacer doble clic en la celda *Number Cars Owned* en la columna de comprador de bicicletas (la columna con la etiqueta 1) y posteriormente un clic con el botón secundario para seleccionar la opción del menú: *Mostrar leyenda,* es posible ver con mayor claridad los detalles. Figura (5.24).

Figura 5.24 – Explorando los perfiles del atributo Number cars owned.

5.3.3.3 Características del Atributo

Mediante la pestaña *Características del atributo*, se puede seleccionar un atributo y un valor para ver la frecuencia con la que aparecen los valores de otros atributos en el caso de los valores seleccionados.

Para explorar el modelo en la pestaña *Características del atributo*, en la lista *Atributo*, comprobar que *Bike Buyer* está seleccionado y que el valor sea 1. En el visor se observará que los clientes que no tienen ningún hijo conviviendo con ellos, una distancia corta al trabajo y que viven en la región de Norteamérica tienen más probabilidad de comprar una bicicleta. Figura (5.25).

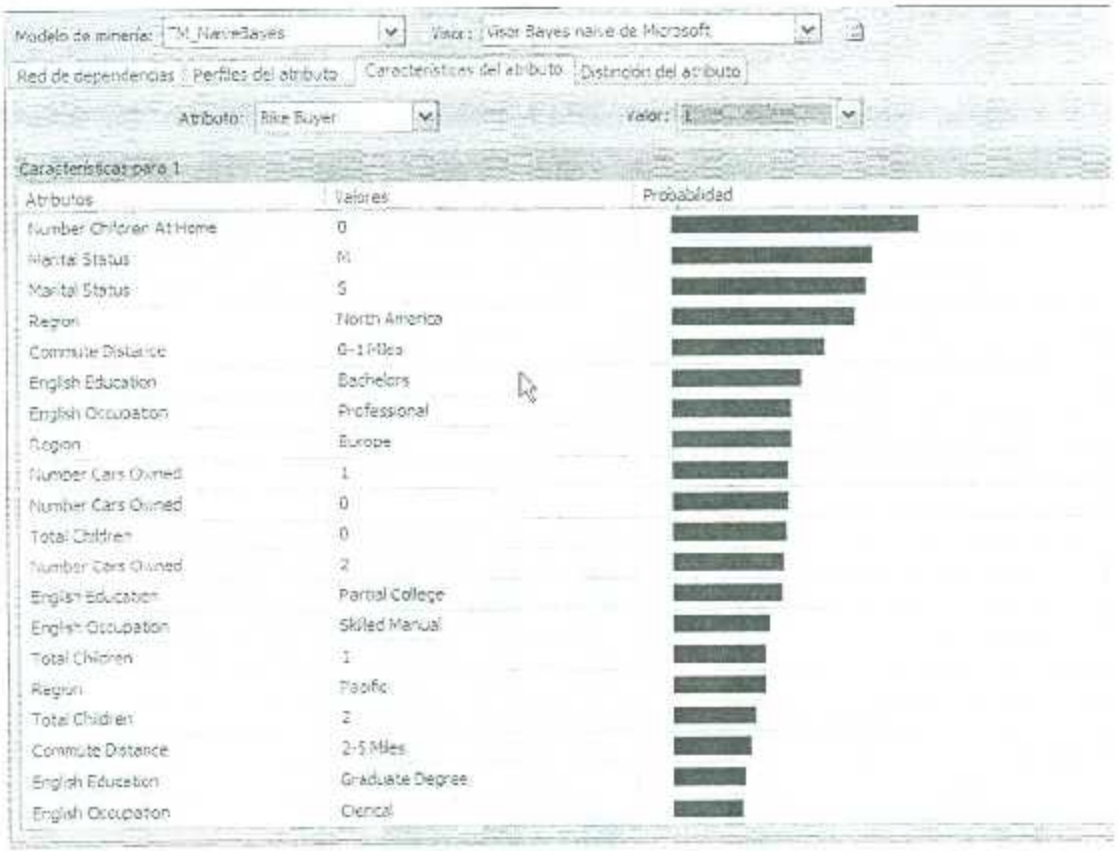


Figura 5.25 – Visor de las características del atributo

5.3.3.4 Distinción del Atributo

La pestaña *Distinción del atributo* permite examinar las relaciones entre dos valores discretos de la compra de una bicicleta y otros valores del atributo.

Dado que el modelo *TM_NaiveBayes* sólo tiene dos estados, 1 y 0, no se tiene que hacer ningún cambio en el visor. En el visor, podrá verse que las personas que no tienen un automóvil tienden a comprar bicicletas y las personas que tienen dos no suelen comprarlas. Figura (5.26)

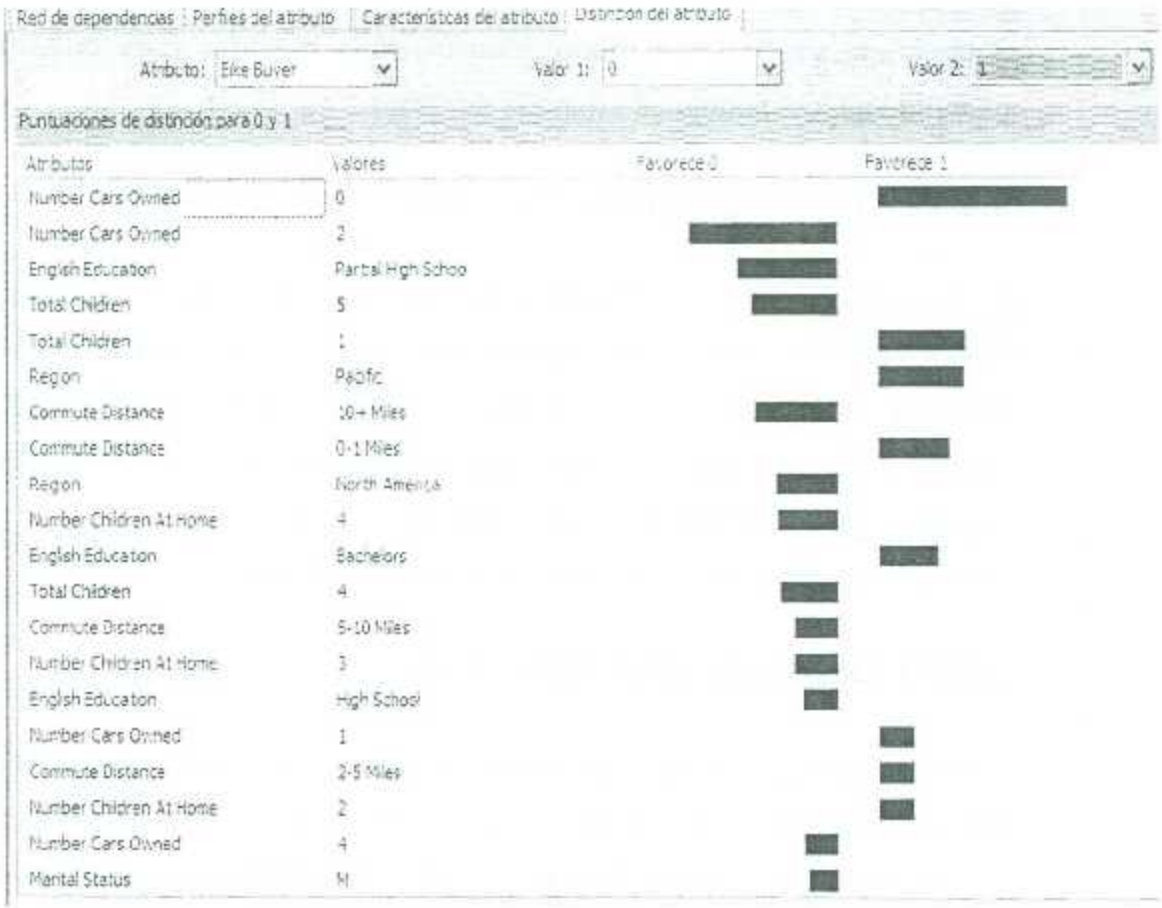


Figura 5.26 – Características que diferencian los valores de los atributos.

5.4 Prueba de los Modelos

Una vez procesados los modelos, se procederá a probar los mismos con el conjunto de pruebas. Dado que los datos del conjunto de pruebas ya contienen valores conocidos para la compra de bicicletas, es fácil determinar si las predicciones del modelo son correctas. El departamento de marketing de Adventure Works Cycles usará el modelo que mejor se comporte para identificar a los clientes para su campaña de distribución de correo directo.

En este punto se probarán primero los modelos realizando predicciones con el conjunto de pruebas. Luego, se probarán en un subconjunto filtrado de los datos. Analysis Services proporciona diversos métodos para determinar la exactitud de los modelos de minería de datos.

5.4.1 Probar la Exactitud con Gráficos de Elevación

En la pestaña *Gráfico de precisión de minería de datos* del *Diseñador de minería de datos*, se puede calcular la precisión de las predicciones de los modelos y comparar los resultados de estos mismos. Este método de comparación se conoce como *gráfico de elevación*. Normalmente, la exactitud de la predicción de un modelo de minería de datos se cuantifica mediante la elevación o la exactitud de la clasificación.

5.4.1.1 Elegir los Datos de Entrada

El primer paso a la hora de probar la precisión de los modelos de minería de datos consiste en seleccionar el origen de datos que se utilizará para realizar las pruebas. Se probará la exactitud de los modelos con un conjunto de datos específico.

Para seleccionar el conjunto de datos, debe posicionarse en la pestaña *Gráfico de precisión de minería de datos* del *Diseñador de minería de datos*

de Business Intelligence Development Studio y seleccionar la pestaña *Selección de entrada*. Posteriormente, en el cuadro de grupo *Seleccionar un conjunto de datos para usarlo en un gráfico de precisión*, seleccionar *Especificar otro conjunto de datos* para probar los modelos. Enseguida, dar un clic sobre el botón de examinar (...) que aparece a un lado de esta opción. Figura (5.27).

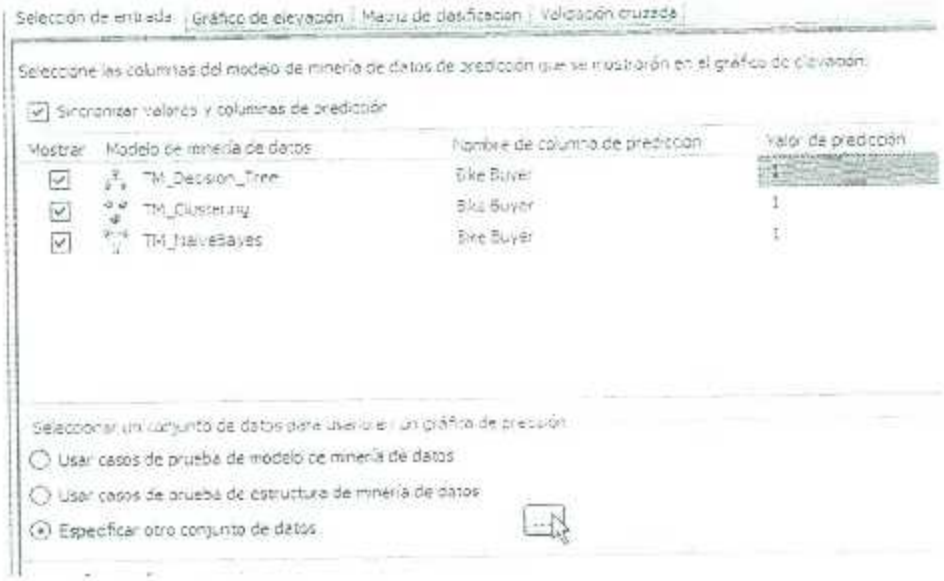


Figura 5.27 – Elección de los datos de entrada.

En el cuadro de diálogo Seleccionar estructura de mineria de datos, seleccionar la estructura que contiene los modelos con los que se desea trabajar, y, después, dar un clic sobre el botón aceptar. Figura (5.28).

Figura 5.28 – Seleccionar Estructura

En Seleccionar tablas de entrada, hacer clic en *Seleccionar tabla de casos*, para abrir el cuadro de diálogo *Seleccionar Tabla*. Una vez abierto este cuadro, selcccionar la tabla de casos previamente definida (vTargetMail) y dar un clic en el botón Aceptar. Luego, se cierra la ventana anterior. Figura (5.29).

Figura 5.29 – Seleccionar tabla de casos

Para Mostrar la elevación de los modelos, en la pestaña *Selección de entrada* del *Diseñador de minería de datos*, en *Seleccione las columnas del modelo de minería de datos de predicción que se mostrarán en el gráfico de elevación*, activar la casilla correspondiente a *Sincronizar valores y columnas de predicción*. En la columna *Nombre de columna de predicción*, comprobar que *Bike Buyer* esté seleccionado para cada modelo. Luego, en la columna *Mostrar*, seleccionar cada uno de los modelos. En la columna *Valor de predicción*, seleccionar *1*. El mismo valor se rellena automáticamente para cada modelo que tiene la misma columna de predicción.

Posteriormente, seleccionar la pestaña *Gráfico de elevación*, para mostrar el gráfico de mejora.

Los resultados se trazarán en un gráfico. El gráfico de elevación traza un modelo de suposición aleatorio así como un modelo ideal. Los modelos de minería de datos previamente creados se situarán entre estos dos extremos, entre una suposición aleatoria y una predicción perfecta.

Cualquier mejora en la suposición aleatoria se considera una *elevación*. Figura (5.30)

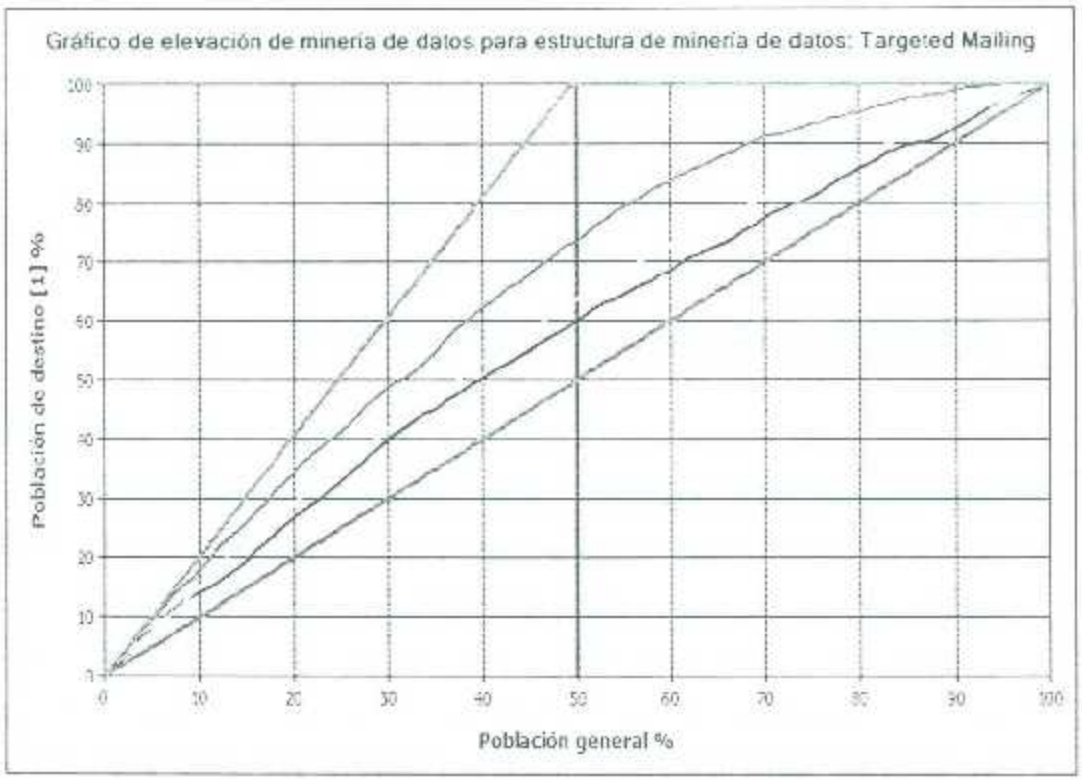


Figura 5.30 – Gráfico de elevación

En la leyenda del gráfico de elevación, se podrá apreciar más claramente que el modelo *TM_Decision_Tree* proporciona la mayor elevación,

superando tanto al modelo de clústeres como al de Naive Bayes. Figura (5.31).

Figura 5.31 – Leyenda correspondiente al gráfico de elevación.

5.5 Crear y Trabajar con Predicciones

En este punto se creará una consulta para predecir qué clientes tienen más probabilidad de comprar una bicicleta. También se recuperará la *probabilidad* de que la predicción sea correcta para poder decidir si se debe presentar o no la recomendación al departamento de marketing.

Cuando se hayan identificado a los clientes con una probabilidad alta de comprar una bicicleta, se obtendrán detalles de los casos del modelo de minería de datos para recuperar los nombres y la información de contacto correspondiente.

5.5.1 Crear una Consulta de Predicción

El primer paso para crear una consulta de predicción consiste en seleccionar un modelo de minería de datos y una tabla de entrada.

En la ficha *Predicción de modelo de minería de datos* del Diseñador de minería de datos, en el cuadro *Modelo de minería de datos*, hacer clic en *Seleccionar modelo*. Figura (5.32).

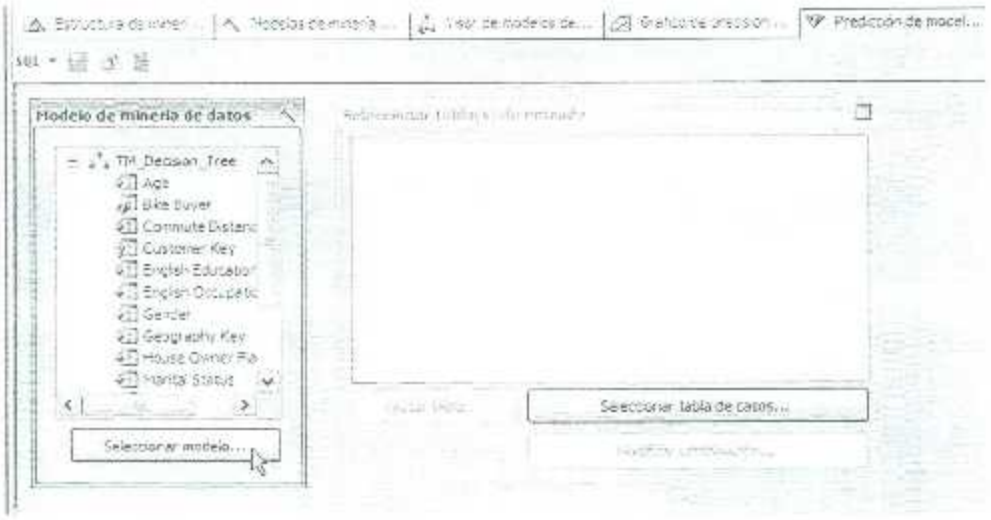


Figura 5.32 – Seleccionar modelo de minería de datos.

Figura 5.33 – Elegir modelo de minería de datos.

En el cuadro de diálogo *Seleccionar modelo de minería de datos*, navegar por el árbol hasta la estructura *Targeted Mailing*, expandirla y seleccionar *TM_Decision_Tree* y, a continuación, hacer clic en *Aceptar*. Figura (5.33).

En el cuadro *Seleccionar tabla(s) de entrada*, hacer clic en *Seleccionar tabla de casos*. En el cuadro de diálogo *Seleccionar tabla*, en la lista *Origen de datos*, seleccionar *Adventure Works DW2008*. En la lista *Nombre de tabla o vista*, seleccionar la tabla *ProspectiveBuyer (dbo)* y, a continuación, hacer clic en *Aceptar*. Figura (5.34).

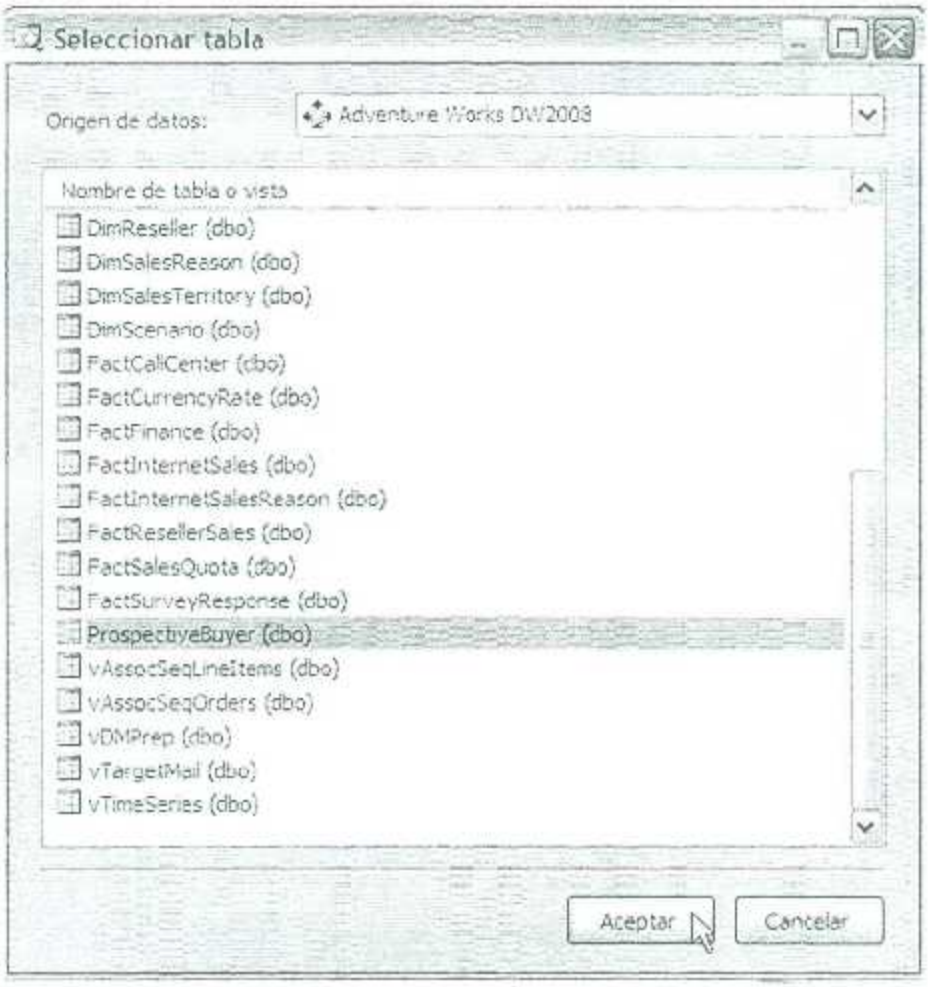


Figura 5.34 – Seleccionar tabla de casos.

5.5.1.1 Asignar las columnas

Hacer clic con el botón secundario en las líneas que conectan la ventana *Modelo de minería de datos* a la ventana *Seleccionar tabla de entrada* y seleccionar *Modificar conexiones*. Figura (5.35).

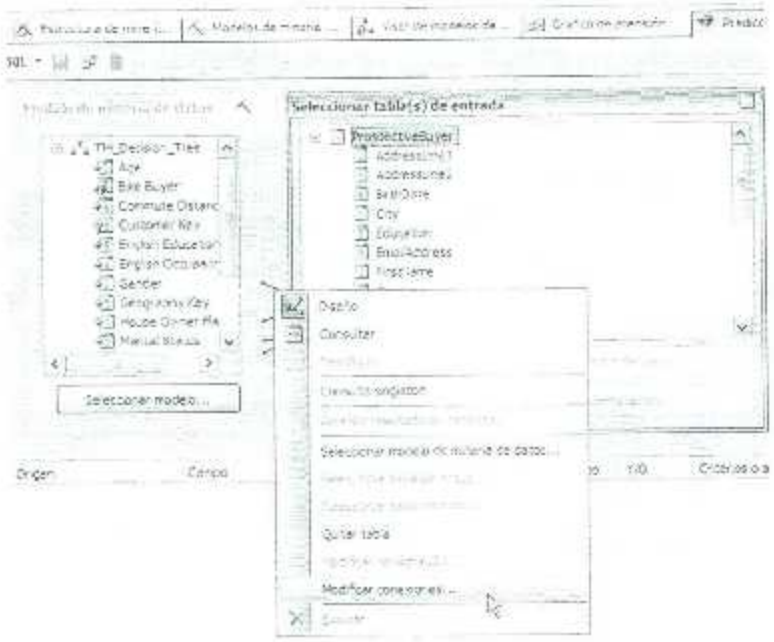


Figura 5.35 – Modificar Conexiones.

Figura 5.36 – Modificar asignación

En *Columna de la tabla*, hacer clic en la celda *Bike Buyer* y seleccionar *ProspectiveBuyer.Unknown* en el cuadro desplegable. Hacer clic en *Aceptar*. Figura (5.36).

En el *Explorador de soluciones*, hacer clic con el botón secundario en la vista del origen de datos *Targeted Mailing* y seleccionar *Diseñador de vistas*. Hacer clic con el botón secundario en el título de tabla *ProspectiveBuyer* y seleccionar *Nuevo cálculo con nombre*. Figura (5.37).

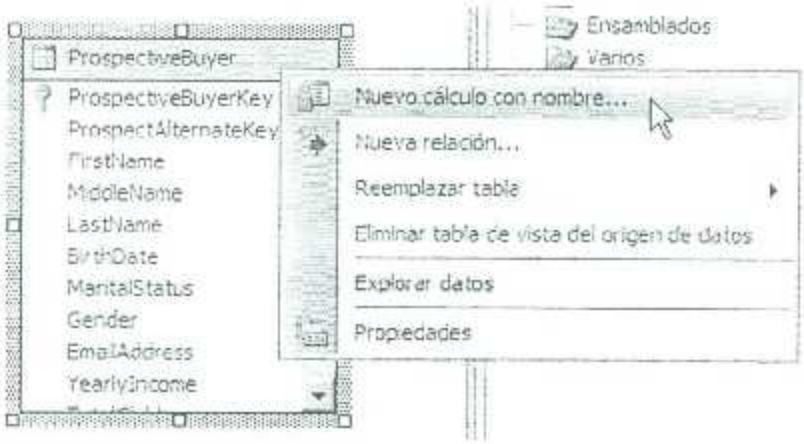


Figura 5.37 – Crear nuevo cálculo con nombre

Figura 5.38 – Definir nuevo cálculo con nombre

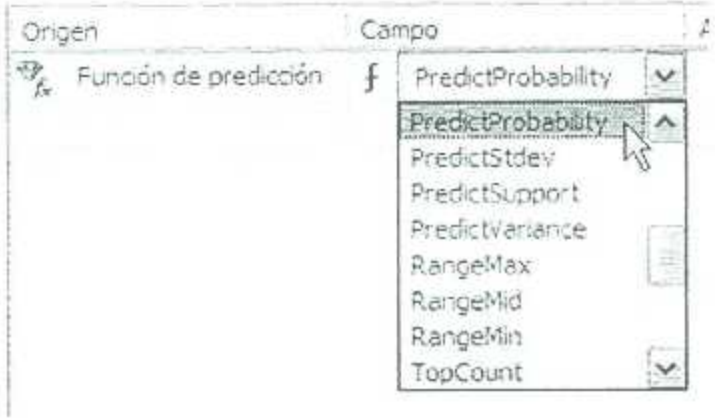
En el cuadro Nombre de columna, escribir *calcAge*. En el cuadro *Expresión*, escribir *DATEDIFF(YYYY,[BirthDate],getdate())* y hacer clic en *Aceptar*. Figura (5.38).

En el Diseñador de minería de datos, seleccionar la ficha *Predicción de modelo de minería de datos* y volver a abrir la ventana *Modificar conexiones*. En *Columna de la tabla*, hacer clic en la celda *Antigüedad* y seleccionar *ProspectiveBuyer.calcAge* en el cuadro desplegable. Hacer clic en *Aceptar* Figura (5.39).

Figura 5.39 – Modificar asignación

5.5.1.2 Diseñar la Consulta de Predicción

El primer botón de la barra de herramientas de la pestaña *Predicción de modelo de minería de datos* es el botón *Cambiar a vista de diseño de consulta / Cambiar a vista de resultado / Cambiar a vista de consulta*. Hacer clic en la flecha abajo en este botón y seleccionar *Diseño*. En la cuadrícula de la ficha *Predicción de modelo de minería de datos*, hacer clic en la celda



de la primera fila vacía de la columna *Origen* y, a continuación, seleccionar *Función de predicción*. En la fila *Función de predicción*, de la columna *Campo*, seleccione *PredictProbability*. Figura (5.40).

Figura 5.40 – Diseñar consulta de Predicción

En la ventana *Modelo de minería de datos* anterior, seleccionar y arrastrar *[Bike Buyer]* hasta la celda *Criterios o argumento*. Cuando es colocado, *[TM_Decision_Tree].[Bike Buyer]* aparece en la celda *Criterios o argumento*. Figura (5.41).

Figura 5.41 – Definición de criterio

Hacer clic en la siguiente fila vacía de la columna *Origen* y, a continuación, seleccionar *TM_Decision_Tree*. En la fila *TM_Decision_Tree*, en la columna *Campo*, seleccionar *Bike Buyer*. En la fila *TM_Decision_Tree*, en la columna *Criterios o argumento*, escribir = 1. Hacer clic en la siguiente fila vacía de la columna *Origen* y, a continuación, seleccionar *ProspectiveBuyer*. En la fila *ProspectiveBuyer*, en la columna *Campo*, seleccionar *ProspectiveBuyerKey*. Por último, agregar cinco filas más a la

cuadrícula. Para cada fila, seleccionar *ProspectiveBuyer* como *Origen* y, a continuación, agregar las columnas siguientes en las celdas *Campo*:

- calcAge
- LastName
- FirstName
- AddressLine1
- AddressLine2

Finalmente, ejecutar la consulta y examinar los resultados. Figura (5.42).

Origen	Campo	Alias	Mostrar	Grupo	Y/O	Criterios o argumento
Función de predicción	PredictProbability		☑			[TM_Decision_Tree].[Bike Buyer]
TM_Decision_Tree	Bike Buyer		☑			= 1
ProspectiveBuyer	ProspectiveBuyerKey		☑			
ProspectiveBuyer	calcAge		☑			
ProspectiveBuyer	LastName		☑			
ProspectiveBuyer	FirstName		☑			
ProspectiveBuyer	AddressLine1		☑			
ProspectiveBuyer	AddressLine2		☑			

Figura 5.42 – Diseñar consulta de Predicción

Para ejecutar la consulta y ver los resultados, en la pestaña *Predicción de modelo de minería de datos*, seleccionar el botón *Resultado*. Figura (5.43)

Figura 5.43 – Ejecutar la consulta y ver resultado.

Una vez que la consulta se ejecute y se muestren los resultados, puede notarse que la pestaña *Predicción de modelo de minería de datos* muestra información de contacto para los clientes potenciales que son probables compradores de bicicletas. La columna *Expresión* indica la probabilidad de que la predicción sea correcta. Pueden utilizarse estos resultados para determinar a qué clientes potenciales debe dirigirse el correo como parte de su campaña de publicidad. Hacer clic en el botón *Guardar* para guardar los resultados. Figura (5.44).

Expression	Bike Buyer	ProspectiveBuyerKey	talcAge	LastName	FirstName	AddressLine1	AddressLine2
0.5151521051472176	1	6	39	Bell	Angel	840 Charlotte Ave.	
0.6333151144305887	1	7	43	Bennett	Anna	312 Via Del Verdes	
0.6262256343962	1	9	47	Bhat	Arturo	7640 Isabel Dr.	
0.5401521051472176	1	11	40	Bryant	Abigail	2639 Anchor Court	
0.6262256343962	1	15	54	Carter	Adam	4082 Rosim Road	
0.6799622716801	1	17	41	Castro	Alberto	2696 Santa Rita Dr.	
0.6799622716801	1	18	45	Chandra	Alice	5935 Isabel	
0.5380335403811509	1	19	47	Chandra	April	2939 Banner Court	
0.5401521051472176	1	23	45	Clark	Alexander	3243 Lantor Ave	
0.6262256343962	1	29	46	Cooper	Alexandra	1624 Carlisle Way	
0.6262256343962	1	30	54	Davis	Alexis	6000 Wilson Richardson Ct.	
0.5123634099448295	1	31	58	Davis	Abigail	70 Nuvi Plaza	
0.6333181144305887	1	32	40	Deng	Alice	8958 Carleton Street	
0.6799622716801	1	35	43	Engineer	Andre	North 93270 Newport Highway	
0.6799622716801	1	37	44	Fernandez	Andre	1292 Springlake Drive	
0.6262256343962	1	40	52	Foster	Adam	4634 Fernsilver Lane	
0.6262256343962	1	47	53	Gonzalez	Alexandra	9078 Cole Verde	
0.6799622716801	1	48	45	Gray	Abigail	78025 E. Merrit Isl. Cswy.	
0.6333181144305887	1	50	43	Griffin	Alyssa	6916 Azores	
0.5557295992287	1	53	40	Hall	Alex	9115 Arthur Rd	
0.6261921936396291	1	55	41	Hall	Abigail	8036 Summit View Dr.	
0.6262256343962	1	57	53	He	Alejandro	Pyramid Mall	
0.6262256343962	1	67	50	Jenkins	Ana	1769 Buchanan Ct	Unit G 202

Figura 5.44 – Resultados de predicción

5.5.2 Usar la Obtención de Detalles en Datos de Estructura

Suponiendo que Adventure Works Cycles está enviando un formulario a los clientes potenciales de entre 39 y 46 años de edad como parte de su campaña de publicidad, y el departamento de marketing ha decidido que les gustaría enviar también el formulario a los clientes que compraron bicicletas

de Adventure Works Cycles hace más de cinco años. En este punto, se identificarán los clientes que anteriormente compraron bicicletas y se recuperará su información de contacto. Esta información no está incluida en el modelo, pero se incluye en la estructura. Para recuperar la información de contacto, primero hay que asegurarse que la obtención de detalles está habilitada para la estructura y, a continuación, podrá utilizarse para revelar los nombres y direcciones de los clientes que anteriormente compraron bicicletas

Para habilitar la obtención de detalles, en *Business Intelligence Development Studio*, en la ficha *Modelos de minería de datos* del *Diseñador de minería de datos*, hacer clic con el botón secundario en el modelo *TM_Decision_Tree* y seleccionar *Propiedades*. En la ventana *Propiedades*, hacer clic en *AllowDrillThrough* y seleccionar *True*. Figura (5.45)

Figura 5.45 – Habilitar la obtención de detalles

En la pestaña *Modelos de minería de datos*, hacer clic con el botón secundario en el modelo y seleccione *Procesar modelo*. Esto abrirá las ventanas para ejecutar el proceso. Una vez terminado, cerrar las ventanas.

Para ver los datos de obtención de detalles de un modelo de minería de datos, en *el Diseñador de minería de datos*, hacer clic en la pestaña *Visor de modelo de minería de datos*. Luego, seleccionar el modelo *TM_Decision_Tree* en la *lista Modelo de minería de datos*. Cambiar el valor de la lista *Fondo* por *1*. Seleccionar el visor de árboles de Microsoft en la lista *Visor* y hacer clic con el botón secundario en el nodo *Age >= 39 y <46*. Seleccionar después *Obtener detalles*, después seleccionar *Sólo Columnas de modelo* para abrir la ventana *Obtener detalles*. Figura (5.46).

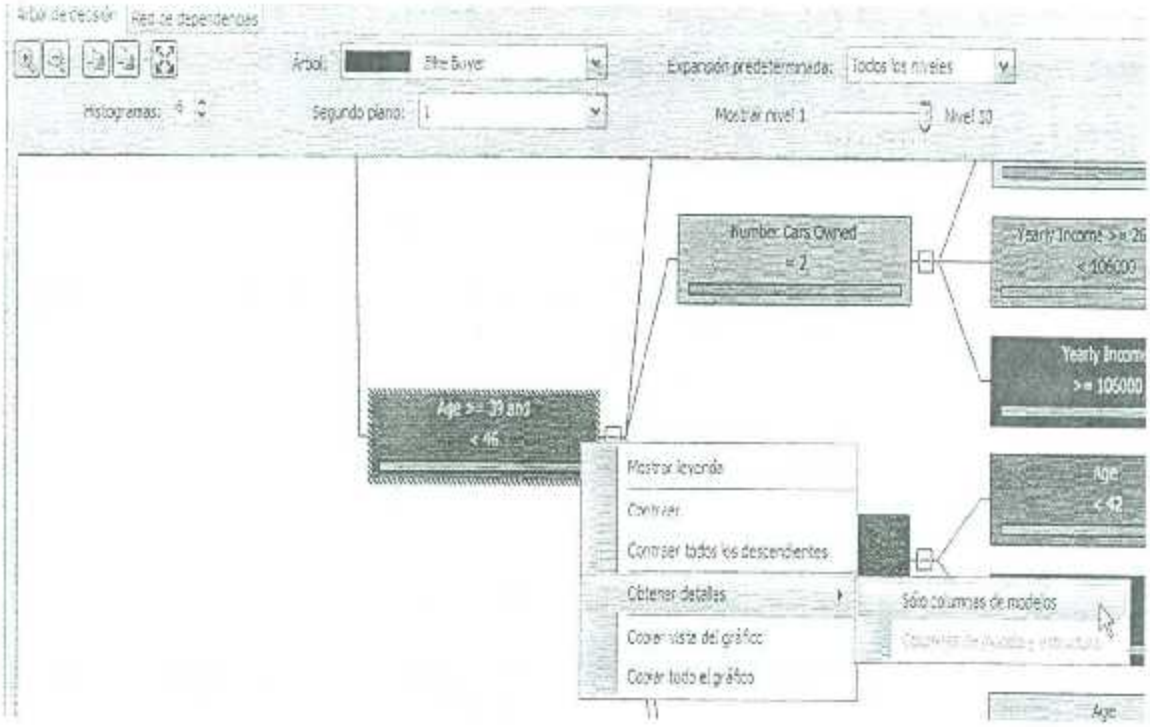
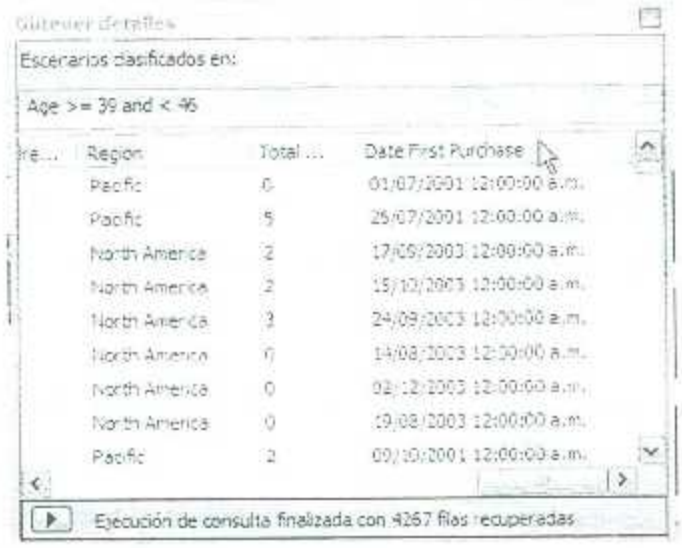


Figura 5.46 – Visualizar los datos en obtención de detalles

Desplazarse luego a la columna *Structure.Date First Purchase* para ver la fecha de compra de las bicicletas anteriores. Figura (5.47).



. Figura 5.47 – Revisar las fechas de compra de bicicletas

Si se desea copiar los datos en el Portapapeles, hacer clic con el botón secundario en cualquier fila de la tabla y seleccionar *Copiar todo* [11]. Figura (5.48).

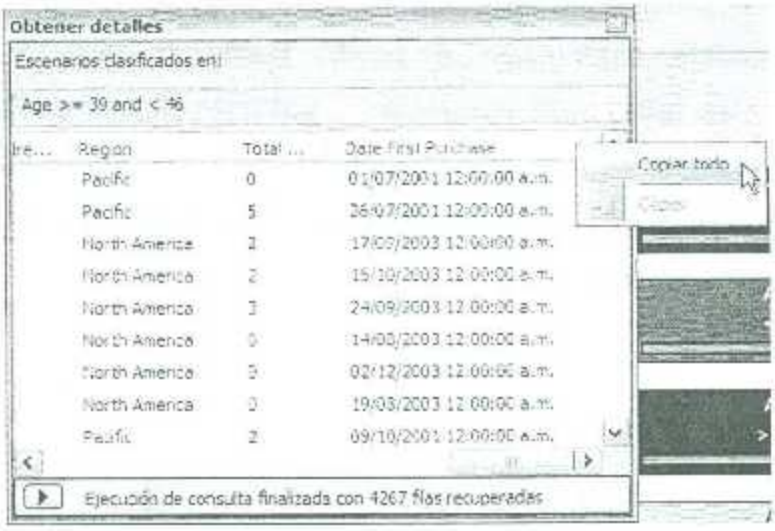


Figura 5.48 – Copiar todos los datos al portapapeles.

5.6 Conclusiones

En este capitulo se puso en práctica un tutorial básico de minería de datos de Microsoft., cuyo objetivo fue el repaso gradual de cada una de las lecciones propuestas, desde la definición de un origen de datos hasta la realización de un cálculo predictivo, para la empresa ficticia *Adventure Works Cycles*.

Algunos modelos como: los árboles de decisión, la agrupación en clústeres y el modelo de Naive Bayes, fueron utilizados para analizar los datos, y tras ser evaluada su exactitud a través del gráfico de elevación, se eligió el más preciso (árbol de decisión) para realizar un cálculo predictivo que permitiera la identificación de los clientes potenciales, con la finalidad de remitirles correos publicitarios.

Como se mencionó en la introducción a este capítulo, la minería de datos cumple una función de altísima importancia en la inteligencia de negocios. Una vez definida una estructura de minería de datos e implementada junto a sus modelos algorítmicos, es posible identificar, entre un vasto universo de datos, las relaciones, tendencias y patrones de comportamiento existentes entre los mismos, con la finalidad de obtener la información más valiosa para el negocio y contribuir con ella a la orientación acertada al momento de tomar decisiones.

CONCLUSIONES GENERALES

Con la redacción de estos cinco capítulos se da por concluida la tesis: *La Creación de Almacenes de Datos y la Inteligencia de Negocios*. Como se puntualizó en las primeras páginas, el objetivo de esta tesis fue subrayar la importancia de esta tecnología a través de la documentación teórica y práctica de sus conceptos fundamentales.

Debido a que la información relacionada a este tema es abundante y sería inviable pretender abarcar todos sus aspectos, antes de iniciar el desarrollo de esta tesis, se procedió a la determinación de los puntos específicos a tratar, que en las opiniones consensuadas entre el asesor y el tesista, se juzgaron primordiales para garantizar una exposición clara, progresiva y elemental.

Posiblemente el tema de la inteligencia de negocios, y todo lo que tecnológicamente entraña, puede considerarse poco novedoso. Se ha escrito extensamente sobre él y en internet abunda la información; sin embargo, contrario a lo que pudiera pensarse, el desconocimiento y desinterés empresarial a este respecto, por lo menos en esta región del país, se da casi en esa misma proporción. Se trata sin duda de una tecnología tan eficaz como desaprovechada, cuyo valor es universal y su vigencia no sufre limitaciones en función de las tendencias de la informática ni de la tecnología en general, puesto que para las empresas jamás perderá importancia la marcha óptima de sus negocios.

El curso seguido en esta tesis, como también fue referido anteriormente, se orienta, aunque no exclusivamente, hacia aquéllos quienes integran las áreas de desarrollo de software dentro de las organizaciones empresariales, y cuyos conocimientos rudimentarios del tema los colocan en el nivel del principiante.

La pretensión de este trabajo académico es que puedan servirse libremente de la información que lo compone y que hallen en él valor y utilidad para el desarrollo de aplicaciones particulares.

LINEAS DE INVESTIGACIÓN ABIERTA

- Desarrollo de ejercicios a un nivel más avanzado.
- Utilización de datos reales.
- Acceso a los datos y presentación de los mismos desde otras aplicaciones (Microsoft Office, Microsoft SharePoint Server, etc).
- Utilización de otros algoritmos para minería de datos (Laplace, matrices de confusión, red neural, etc.)

BIBLIOGRAFÍA

[1] Amaya Guzmán Jorge Guillermo, "Evolución de los sistemas de soporte a la decisión, sitio oficial: http://www.gestiopolis.com/canales6/ger/sistemas-soporte-para-la-toma-de-decisiones.htm

[2] Morales Ruiz Juan Antonio, "Introducción a la informática", sitio oficial: http://uncontroldeuplicsa.blogspot.mx/2008/09/sistemas-oltp-epr-y-crm.html

[3] Cano Josep Lluis, "Business Intelligence: Competir con Información", Esade, pp. 125-133.

[4] Cuéllar M. Guillermo, "Data Warehouse, aspectos técnicos, características, usos, beneficios, componentes, herramientas OLAP", Sitio Oficial: http://fccea.unicauca.edu.co/old/datawarehouse.htm

[5] Berzal Fernando, "El modelo multidimensional Data Warehousing", sitio oficial: http://elvex.ugr.es/idbis/db/docs/intro/F%20Modelo%20multidimensional.pdf

[6] Murillo Alfaro Felix, "Manual de construcción de un Data Warehouse", sitio oficial: http://www.onge.cob.pe/publica/metodologias/Lib5084/134.HTM

[7] Wolff Gloria Carmen, "Modelamiento Multidimensional", sitio oficial: http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF

[8] "Instalar la base de datos AdventureWorks", sitio oficial: http://msdn.microsoft.com/es-es/Library/aa992075(v=vs.100).aspx

[9] "Crear un informe de ReportViewer", sitio oficial: http://msdn.microsoft.com/es-es/Library/ms252073(v=vs.100).aspx

[10] "Tutorial de SQL Server Analysis Services", sitio oficial: http://technet.microsoft.com/es-es/library/ms170208(v=sql.100).aspx

[11] "Tutorial básico de mineria de datos", sitio oficial: http://technet.microsoft.com/cs-cs/library/ms167167(v=sql.100).aspx

GLOSARIO

Bussines Intelligence o BI: Inteligencia de Negocios.

Call Center: Centro de atención de llamadas telefónicas.

CMR: Sistema para la gestión de la relación con el cliente. Por sus siglas en inglés (Customer Relationship Management).

Cuadro de Mando: Herramienta de control empresarial que permite establecer y monitorizar los objetivos de una empresa y de sus diferentes áreas o unidades. Ayuda a una compañía a expresar los objetivos e iniciativas necesarias para cumplir con su estrategia, mostrando de forma continuada cuándo la empresa y los empleados alcanzan los resultados definidos en su plan estratégico.

Cubo de Datos o Cubo OLAP: Es una base de datos multidimensional, en la cual el almacenamiento físico de los datos se realiza en un vector multidimensional. Los cubos OLAP se pueden considerar como una ampliación de las dos dimensiones de una hoja de cálculo.

Dashboard: Herramienta de cuadros de mando.

Data Mining: Minería de Datos.

Data Warehouse: Depósito de Datos. Es una colección de datos orientado a temas, integrado, de tiempo variante, no volátil, que se usa para el soporte del proceso de toma de decisiones gerenciales

DataMart: Es una versión especial de almacén de datos (data warehouse). Son subconjuntos de datos con el propósito de ayudar a que un área especifica dentro del negocio pueda tomar mejores decisiones.

Denormalización: Sinónimo de Desnormalización.

Desnormalización: Proceso que a través de la redundancia controlada de datos de las tablas, pretende la optimización del esquema relacional existente entre las mismas.

Dice: Operación que permite reducir la dimensionalidad de los datos mediante proyección en los cubos OLAP.

Dimensión: En la construcción de cubos OLAP, son tablas que contienen atributos (o campos) utilizados para restringir y agrupar los datos

Tabla: Las tablas representan el formato más adecuado para organizar múltiples datos que deben aparecer relacionados. Las tablas constan de casillas de entradas de datos, denominadas celdas; cada una de ellas viene referenciada por una columna y la fila en la que se encuentra.

Wrapper: Encapsulador. Parte del Data Warehousing encargado de la extracción de datos de fuentes diversas, así como de la transmisión de los mismos al Data Warehouse.