

DATA WAREHOUSING: Investigación y Sistematización de Conceptos

—

HEFESTO: Metodología propia para la Construcción de un Data Warehouse

Ing. Bernabeu, Ricardo Dario

Córdoba, Argentina – Miércoles 07 de Noviembre de 2007

Copyright ©2007 Ing. Bernabeu, Ricardo Dario. Se otorga permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.2 o cualquier otra versión posterior publicada por la Free Software Foundation; requiriendo permanecer invariable el nombre de la metodología (HEFESTO), en cuanto al diseño de su logotipo, debe mantenerse el estilo medieval para su confección y letra "O" representada por el símbolo de radioactividad (). Una copia de la licencia está incluida en la sección titulada Licencia de Documentación Libre de GNU.

Fecha	Versión	Autor	Detalle del cambio
Miércoles 07 de Noviembre de 2007	0.1	Ing. Bernabeu, Ricardo Dario	Versión Inicial

*...si supiese qué es lo que estoy haciendo,
no lo llamaría INVESTIGACIÓN...*

Albert Einstein

Índice general

I DATA WAREHOUSING: Investigación y Sistematización de Conceptos	1
RESUMEN	3
1. BUSINESS INTELLIGENCE	5
1.1. Introducción	5
1.2. Definición	6
1.3. Proceso de BI	6
1.4. Beneficios	7
2. DATA WAREHOUSE	9
2.1. Introducción	9
2.2. Definición	9
2.3. Características	10
2.3.1. Orientada al negocio	10
2.3.2. Integrada	11
2.3.3. Variante en el tiempo	12
2.3.4. No volátil	12
2.4. Cualidades	13
2.5. Ventajas	14
2.6. Desventajas	15
2.7. Redundancia	15
2.8. Estructura	16
2.9. Flujo de Datos	17
3. ARQUITECTURA DEL DATA WAREHOUSE	19
3.1. Introducción	19
3.2. OLTP	20
3.3. Load Manager	21
3.3.1. Extracción	21
3.3.2. Transformación	22
3.3.2.1. Codificación	22
3.3.2.2. Medida de atributos	23
3.3.2.3. Convenciones de nombramiento	23
3.3.2.4. Fuentes múltiples	24
3.3.2.5. Limpieza de datos	24
3.3.3. Carga	25
3.3.4. Proceso ETL	25
3.3.5. Tareas del ETL	25
3.4. Data Warehouse Manager	26
3.4.1. Base de datos multidimensional	27
3.4.2. Tablas de Dimensiones	28

3.4.2.1. Dimensión Tiempo	29
3.4.2.2. Jerarquías	29
3.4.2.3. Relación	30
3.4.2.4. Granularidad	30
3.4.3. Tablas de Hechos	31
3.4.4. Esquema en Estrella	32
3.4.5. Esquema Copo de Nieve	34
3.4.6. Esquema Constelación	35
3.4.7. OLTP vs DW	36
3.4.8. ROLAP	36
3.4.9. MOLAP	37
3.4.10. HOLAP	37
3.4.11. Cubo Multidimensional	38
3.4.12. Metadatos	40
3.4.12.1. Mapping	41
3.5. Query Manager	42
3.5.1. Drill-down	43
3.5.2. Drill-up	45
3.5.3. Drill-across	46
3.5.4. Roll-across	47
3.5.5. Pivot	48
3.5.6. Page	49
3.6. Herramientas de Consulta y Análisis	51
3.6.1. Reportes y Consultas	52
3.6.2. OLAP	52
3.6.3. Data Mining	53
3.6.3.1. Redes Neuronales	54
3.6.3.2. Sistemas Expertos	55
3.6.3.3. Programación Genética	55
3.6.3.4. Árboles de Decisión	55
3.6.3.5. Detección de Desviación	56
3.6.4. EIS	56
3.7. Usuarios	56
4. OTROS CONCEPTOS A TENER EN CUENTA	59
4.1. Sistema de Misión Crítica	59
4.2. Data Mart	59
4.3. SGBD	61
II HEFESTO: Metodología propia para la Construcción de un Data Warehouse	63
RESUMEN	65
5. METODOLOGÍA HEFESTO	67
5.1. Introducción	67
5.2. Descripción	68
5.3. Características	69
5.4. Empresa analizada	69
5.5. Pasos y aplicación metodológica	70
5.5.1. PASO 1) ANÁLISIS DE REQUERIMIENTOS	70
5.5.1.1. a) Identificar preguntas	70
5.5.1.2. b) Identificar indicadores y perspectivas de análisis	71

5.5.1.3. c) Modelo Conceptual	72
5.5.2. PASO 2) ANÁLISIS DE LOS OLTP	73
5.5.2.1. a) Establecer correspondencias con los requerimientos	73
5.5.2.2. b) Seleccionar los campos que integrarán cada perspectiva. Nivel de granularidad	75
5.5.3. PASO 3) ELABORACIÓN DEL MODELO LÓGICO DE LA ESTRUCTURA DEL DW	78
5.5.3.1. a) Diseñar tablas de dimensiones	79
5.5.3.2. b) Diseñar tablas de hechos	80
5.5.3.3. c) Realizar uniones	83
5.5.3.4. d) Determinar jerarquías	83
5.5.4. PASO 4) PROCESOS ETL, LIMPIEZA DE DATOS Y SENTENCIAS SQL	86
6. OTRAS CONSIDERACIONES A TENER EN CUENTA	91
6.1. Tamaño del DW	91
6.2. Tiempo de construcción	92
6.3. Implementación	92
6.4. Performance	92
6.5. Mantenimiento	93
6.6. Impactos	93
6.7. DM como sub proyectos	94
6.8. Teoría de grafos	94
6.9. Elección de columnas	95
Apéndice A	97
A. Descripción de la empresa	97
A.1. Identificación de la empresa	97
A.2. Objetivos	97
A.3. Políticas	97
A.4. Estrategias	98
A.5. Organigrama	98
A.6. Datos del entorno específico	98
A.7. Relación de las metas de la organización con las del DW	99
A.8. Procesos	99
Apéndice B	101
B. Licencia de Documentación Libre de GNU	101
B.1. Preámbulo	101
B.2. Aplicabilidad y definiciones	102
B.3. Copia literal	103
B.4. Copiado en cantidad	104
B.5. Modificaciones	104
B.6. Combinación de documentos	106
B.7. Colecciones de documentos	106
B.8. Agregación con trabajos independientes	106
B.9. Traducción	107
B.10Terminación	107
B.11Revisiones futuras de esta licencia	107
B.12Adenda: cómo usar esta Licencia en sus documentos	108

Bibliografía	109
Índice de figuras	111
Índice alfabético	115

Parte I

DATA WAREHOUSING:

Investigación y

Sistematización de Conceptos

RESUMEN

En esta primera parte de la publicación, se sistematizarán todos los conceptos inherentes al Data Warehouse, haciendo referencia a cada uno de ellos en forma ordenada, en un marco conceptual claro, en el que se desplegarán sus características y cualidades, y teniendo siempre en cuenta su relación o interrelación con los demás componentes del ambiente.

Inicialmente, se definirá el concepto de Business Intelligence y sus respectivas características. Seguidamente, se introducirá al Data Warehouse y se expondrán sus aspectos más relevantes y significativos. Luego, se precisarán y detallarán todos los componentes que intervienen en su arquitectura, de manera organizada e intuitiva, atendiendo su interrelación. Finalmente, se describirán algunos conceptos que deben tenerse en cuenta.

El principal objetivo de esta investigación, es ayudar a comprender el complejo ambiente de Data Warehouse, sus respectivos componentes y la interrelación entre los mismos, así como también cuales son sus ventajas, desventajas y características propias. Es por ello, que se hará énfasis en la sistematización de todos los conceptos de la estructura del Data Warehouse, debido a que la documentación existente se enfoca en tratar temas independientes sin tener en cuenta su vinculación y referencias a otros componentes del mismo.

Por último, se desea publicar el proyecto a través de la Licencia de Documentación Libre de GNU (GFDL – GNU Free Documentation License), para permitir y proteger su libre difusión, distribución, modificación y utilización, en pos de su futura evolución y actualización.

Capítulo 1

BUSINESS INTELLIGENCE

1.1. Introducción

Actualmente, en las actividades diarias de cualquier organización, se generan datos como producto secundario, que son el resultado de todas las transacciones que se realizan. Es muy común, que los mismos se almacenen y administren a través de sistemas transaccionales en bases de datos relacionales.

Pero, la idea central de esta publicación, es que estos dejen de solo ser simples datos, para convertirse en información que enriquezca las decisiones de los ejecutivos.

Las organizaciones desean explotar y maximizar el valor de su información para lograr tener una mayor ventaja competitiva. Además, es un factor muy importante para las mismas, con el fin de incrementar sus ganancias, enfocarse en retener a sus clientes actuales, como así también conseguir nuevos. Cabe destacar, que comprender el comportamiento de los clientes es de vital importancia, ya que generalmente, el 20 % de estos siempre generan un 80 % del total de las ganancias.

Debido a lo expuesto anteriormente, sería ideal que las organizaciones tuvieran la posibilidad de segmentar y/o clasificar a sus clientes de alguna manera, en este caso, por su rentabilidad, para poder actuar en base a ello, confeccionando una estrategia que permita cumplir con este objetivo.

La Inteligencia de Negocios (Business Intelligence - BI), permite realizar este tipo de segmentación, además, está orientada a encontrar información que no solo se encargue de responder a preguntas de lo que está sucediendo o ya sucedió, sino también, posibilita la construcción de modelos, mediante los cuales se podrán predecir eventos futuros.

Cuando se nombra el término inteligencia, se refiere a la aplicación combinada de información, habilidad, experiencia y razonamientos, para resolver un problema de negocios.

Como ejemplo, se mencionarán algunas preguntas típicas:

- ¿Cuáles son los mejores proveedores?
- ¿Cuáles son los clientes más rentables? ¿Cómo hacer para retenerlos?
- ¿Cuál es el margen de cada línea de producto?
- ¿Cuáles fueron los productos más vendidos en el mes pasado, en la zona norte?

1.2. Definición

Se puede describir BI, como un concepto que integra por un lado el almacenamiento y por el otro el procesamiento de grandes cantidades de datos, con el principal objetivo de transformarlos en conocimiento y en decisiones en tiempo real, a través de un sencillo análisis y exploración.

La definición antes expuesta puede representarse a través de la siguiente fórmula:

$$\text{Datos} + \text{Análisis} = \text{Conocimiento}$$

Existe una definición muy popular acerca de BI, que dice: "Inteligencia de Negocios es el proceso de convertir datos en conocimiento y el conocimiento en acción, para la toma de decisiones".

BI hace hincapié en los procesos de recolectar y utilizar efectivamente la información, con el fin de mejorar la operación de un negocio, brindando a sus usuarios, el acceso a la información clave que necesitan para llevar a cabo sus tareas habituales y más precisamente, para poder tomar decisiones oportunas basadas en datos correctos y certeros.

Al contar con la información exacta y en tiempo real, es posible, aparte de lo ya mencionado, identificar y corregir situaciones antes de que se conviertan en problemas y en potenciales pérdidas de control de la empresa, pudiendo conseguir nuevas oportunidades o readaptarse frente a la ocurrencia de sucesos inesperados. Por ejemplo, este hecho podría darse en los casos en que la respuesta de los clientes no conforma las expectativas que le dieron origen.

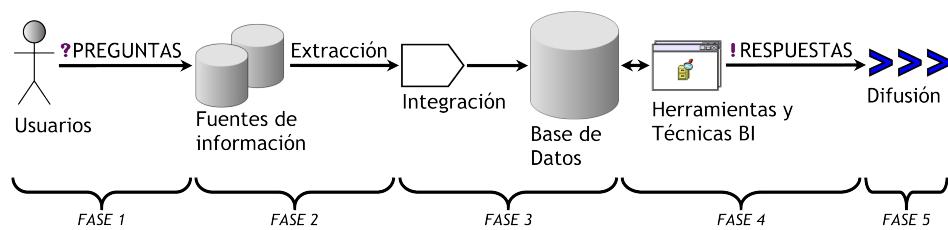
Cuanto más relevante y útil sea la inteligencia que posea una organización sobre un negocio, sus clientes, proveedores, socios, operaciones, etc, mayor será su ventaja competitiva y se podrán tomar mejores decisiones. Esto se debe simplemente a que, por ejemplo, cuanto más se conoce a los clientes, se logra satisfacer sus necesidades de mejor manera y por supuesto, anticipar sus necesidades.

La Inteligencia de Negocios tiene sus raíces en los Sistemas de Información Ejecutiva¹ (Executive Information Systems – EIS) y en los Sistemas para la Toma de Decisiones (Decision Support Systems – DSS), pero ha evolucionado y se ha transformado en todo un conjunto de tecnologías capaces de satisfacer a una gran gama de usuarios junto a sus necesidades específicas en cuanto al análisis de información.

1.3. Proceso de BI

A fin de comprender cómo es que una organización puede crear inteligencia de sus datos, para, como ya se ha mencionado, proveer a los usuarios finales oportuna y acertadamente acceso a esta información, se describirá a continuación el proceso de BI. El mismo está dividido en cinco fases, las cuales serán explicadas teniendo como referencia el siguiente gráfico, que sintetiza todo el proceso:

¹Ver sección 3.6.4, en la página 56.

**Figura 1.1: Fases del proceso BI.**

- Fase 1 – Dirigir y Planear. En esta fase inicial es donde se deberán recolectar los requerimientos de información específicos de los diferentes usuarios, así como entender sus diversas necesidades, para que luego en conjunto con ellos se generen efectivamente las preguntas que les ayudarán a alcanzar sus objetivos.
- Fase 2 – Recolección de Información. Es aquí en donde se realiza el proceso de extraer desde las diferentes fuentes de información de la empresa, tanto internas como externas, los datos que serán necesarios para encontrar las respuestas a las preguntas planteadas en el paso anterior.
- Fase 3 – Procesamiento de Datos. En esta fase es donde se integran y cargan los datos en crudo en un formato utilizable para el análisis. Esta actividad puede realizarse mediante la creación de una nueva base de datos, agregando datos a una base de datos ya existente o bien consolidando la información.
- Fase 4 – Análisis y Producción. Ahora, se procederá a trabajar sobre los datos extraídos e integrados, utilizando herramientas y técnicas propias de la tecnología BI, para crear inteligencia. Como resultado final de esta fase se obtendrán las respuestas a las preguntas, mediante la creación de reportes, indicadores, etc.
- Fase 5 – Difusión. Finalmente se les entregará a los usuarios que lo requieran las herramientas necesarias, que les permitirán explorar los datos de manera veloz y sencilla.

1.4. Beneficios

Entre los beneficios más importantes que BI proporciona a las organizaciones, vale la pena destacar los siguientes:

- Reduce el tiempo mínimo que se requiere para recoger toda la información relevante del negocio, ya que la misma se encontrará integrada en una fuente única de fácil acceso.
- Automatiza la asimilación de la información, debido a que la extracción y carga de los datos necesarios se realizará a través de procesos predefinidos.
- Proporciona herramientas de análisis para establecer comparaciones y tomar decisiones.
- Cierra el círculo que hace pasar de la decisión a la acción.
- Permite a los usuarios no depender de reportes o informes programados, porque los mismos serán generados de manera dinámica.

- Posibilita la formulación y respuesta de preguntas que son claves para el desempeño de la empresa.
- Permite acceder y analizar directamente los indicadores de éxito.
- Se pueden identificar cuáles son los factores que inciden en el buen o mal funcionamiento de la empresa.
- Se podrán detectar situaciones fuera de lo normal.
- Se encontrarán y/o descubrirán cuáles son los factores que maximizarán el beneficio.
- Permitirá predecir el comportamiento futuro con un alto porcentaje de certeza, basado en el entendimiento del pasado.
- El usuario podrá consultar y analizar los datos de manera sencilla.

Capítulo 2

DATA WAREHOUSE

2.1. Introducción

Debido a que para llevar a cabo BI, es necesario gestionar datos guardados en diversos formatos, fuentes y tipos, para luego depurarlos e integrarlos, además de almacenarlos en un solo destino, depósito o base de datos que permita su posterior análisis y exploración, es imperativo y de vital importancia contar con una herramienta que satisfaga todas estas necesidades.

Esta herramienta es el Data Warehouse (DW), que básicamente se encarga de consolidar, integrar y centralizar los datos que la empresa genera en todos los ámbitos de una actividad de negocios (Compras, Ventas, Producción, etc), para luego ser almacenados mediante una estructura que permite el acceso y exploración de la información requerida con buena performance, facilitando posteriormente, una amplia gama de posibilidad de análisis multivariados, que permitirá la toma de decisiones estratégicas y tácticas.

2.2. Definición

El DW posibilita la extracción de datos de sistemas operacionales y fuentes externas, permite la integración y homogenización de los datos de toda la empresa, provee información que ha sido transformada y sumarizada, para que ayude en el proceso de toma de decisiones estratégicas y tácticas.

El DW, convertirá entonces los datos operacionales de la empresa en una herramienta competitiva, debido a que pondrá a disposición de los usuarios indicados la información pertinente, correcta e integrada, en el momento que se necesita.

Una de las definiciones más famosas sobre DW, es la de W. H. Inmon, quien define: “Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia”.

Debido a que W. H. Inmon, es reconocido mundialmente como el padre del DW, la explicación de las características más sobresalientes de esta herramienta se basó en su definición.

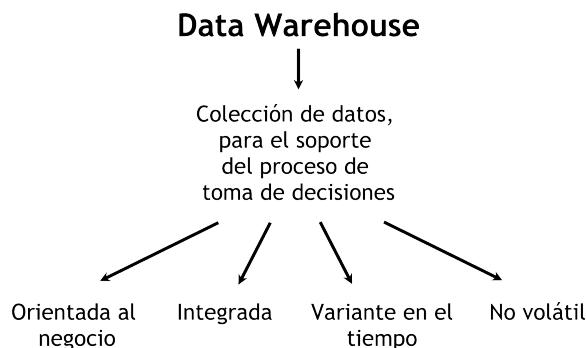


Figura 2.1: Data Warehouse, características.

Cabe aclarar que los términos almacén de datos y depósito de datos, son análogos a DW, y se utilizarán de aquí en adelante para referirse al mismo.

2.3. Características

2.3.1. Orientada al negocio

La primera característica del DW, es que la información se clasifica en base a los aspectos que son de interés para la empresa. Esta clasificación afecta el diseño y la implementación de los datos encontrados en el almacén de datos, debido a que la estructura del mismo difiere considerablemente a la de los clásicos procesos operacionales orientados a las aplicaciones.

A continuación, y con el fin de obtener una mejor comprensión de las diferencias existentes entre estos dos tipos de orientación, se realizará un análisis comparativo:

- Con respecto al nivel de detalle de los datos, el DW excluye la información que no será utilizada exclusivamente en el proceso de toma de decisiones; mientras que en los procesos orientados a las aplicaciones, se incluyen todos aquellos datos que son necesarios para satisfacer de manera inmediata los requerimientos funcionales de la actividad que soporten. Por ejemplo los datos comunes referidos al cliente, como su dirección de correo electrónico, fax, teléfono, D.N.I., código postal, etc, que son tan importantes de almacenar en cualquier sistema operacional, no son tenidos en cuenta en el depósito de datos por carecer de valor para la toma de decisiones, pero sí lo serán aquellos que indiquen el tipo de cliente, su clasificación, ubicación geográfica, sexo, edad, etc.
- En lo que concierne a la interacción de la información, los datos operacionales mantienen una relación continua entre dos o más tablas, basadas en alguna regla comercial vigente; en cambio las relaciones encontradas en los datos residentes del DW son muchas, debido a que por lo general cada tabla del mismo estará conformada por la integración de varias tablas u otras fuentes del ambiente operacional, cada una con sus propias reglas de negocio inherentes.

El origen de este contraste es totalmente lógico, ya que el ambiente operacional se diseña alrededor de las aplicaciones u programas que necesita la organización para llevar a cabo sus actividades diarias y funciones específicas. Por ejemplo, una aplicación de una

institución financiera manejará: préstamos, ahorros, tarjetas bancarias, cuentas, depósitos, etc. De esta manera, la base de datos combinará estos elementos en una estructura que se adapte a sus necesidades.

En contraposición, por ejemplo, para un fabricante el ambiente DW se organizará alrededor de entidades de alto nivel tales como: clientes, productos, rubros, proveedores, vendedores, zonas, etc. Que son precisamente aquellos sujetos mediante los cuales se desea analizar la información. Esto se debe a que el depósito de datos se diseña para realizar consultas e investigaciones sobre las actividades de la organización y no para soportar los procesos que se realizan en ella.

En síntesis, la ventaja de contar con procesos orientados a la aplicación, esta fundamentada en la alta accesibilidad de los datos, lo que implica un elevado desempeño y velocidad en la ejecución de consultas, ya que las mismas están predeterminadas; mientras que en el DW para satisfacer esta ventaja se requiere que la información este desnormalizada, es decir, con redundancia¹, duplicidad de los datos y que la misma esté dimensionada, para evitar tener que recorrer toda la base de datos cuando se necesite realizar algún análisis determinado, sino que simplemente la consulta sea enfocada por vectores y variables que permitan localizar los datos de manera rápida y eficaz, para poder de esta manera satisfacer una alta demanda de complejos exámenes en un mínimo tiempo de respuesta.

2.3.2. Integrada

La integración implica que todos los datos de diversas fuentes que son producidos por distintos departamentos, secciones y aplicaciones, tanto internos como externos, deben ser consolidados en una instancia antes de ser agregados al DW. A este proceso se lo conoce como Extracción, Transformación y Carga de Datos² (Extraction, Transformation and Load - ETL).

La integración de datos, resuelve diferentes variados tipos de problemas relacionados con las convenciones de nombres, unidades de medidas, codificaciones, fuentes múltiples, etc, cada uno de los cuales será correctamente detallado y ejemplificado más adelante.

Esto se debe a que a través de los años los diseñadores y programadores no se han basado en ningún estándar para definir nombres de variables, tipos de datos, etc, ya sea por carecer de ellos o por no creer que sean necesarios. Por lo cual, cada uno por su parte ha dejado en cada aplicación, módulo, tabla, etc, su propio estilo personalizado, confluendo de esta manera en la creación de modelos muy inconsistentes e incompatibles entre sí.

Los puntos de integración afectan casi todos los aspectos de diseño, y cualquiera sea su forma, el resultado es el mismo, ya que la información será almacenada en el DW en un modelo globalmente aceptable y singular, aún cuando los sistemas operacionales y demás fuentes almacenen los datos de maneras disímiles, para que de esta manera el usuario final este enfocado en la utilización de los datos del depósito y no deba cuestionarse sobre la confiabilidad o solidez de los mismos.

¹Ver sección 2.7, en la página 15.

²Ver sección 3.3, en la página 21.

2.3.3. Variante en el tiempo

Debido al gran volumen de información que se manejará en el DW, cuando se le realiza una consulta, los resultados deseados demorarán en originarse. Este espacio de tiempo que se produce desde la búsqueda de datos hasta su consecución es del todo normal en este ambiente y es, precisamente por ello, que la información que se encuentra dentro del depósito de datos se denomina de tiempo variable.

Esta característica básica, es muy diferente de la información encontrada en el ambiente operacional, en el cual, los datos se requieren en el momento de acceder, es decir, que se espera que los valores procurados se obtengan a partir del momento mismo de acceso.

Además, toda la información en el DW posee su propio sello de tiempo:

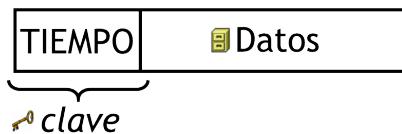


Figura 2.2: Data Warehouse, variante en el tiempo.

Esto contribuye a una de las principales ventajas del almacén de datos: los datos son almacenados junto a sus respectivos históricos. Esta cualidad que no se encuentra en fuentes de datos operacionales, garantiza poder desarrollar análisis de la dinámica de la información, pues ella es procesada como una serie de instantáneas, cada una representando un periodo de tiempo. Es decir, que gracias al sello de tiempo se podrá tener acceso a diferentes versiones de la misma información.

Es importante tener en cuenta la granularidad³ de los datos, así como también la intensidad de cambio natural del comportamiento de los fenómenos del negocio, para evitar crecimientos incontrolables y desbordamientos de la base de datos.

El intervalo de tiempo y periodicidad de los datos debe definirse de acuerdo a la necesidad y requisitos de los usuarios.

Es elemental aclarar, que el almacenamiento de datos históricos, es lo que permite al DW desarrollar pronósticos y análisis de tendencias y patrones, a partir de una base estadística de información, ya que las instantáneas son actualizadas de acuerdo con las actividades del negocio.

2.3.4. No volátil

La información es útil para el análisis y la toma de decisiones solo cuando es estable. Los datos operacionales varían momento a momento, en cambio, los datos una vez que entran en el DW no cambian.

La actualización, o sea, insertar, eliminar y modificar, se hace de forma muy habitual en el ambiente operacional sobre una base, registro por registro, en cambio en el depó-

³Ver sección 3.4.2.4, en la página 30.

sito de datos la manipulación básica de los datos es mucho más simple, debido a que solo existen dos tipos de operaciones: la carga de datos y el acceso a los mismos.

Por esta razón es que en el DW no se requieren mecanismos de control de la concurrencia y recuperación.

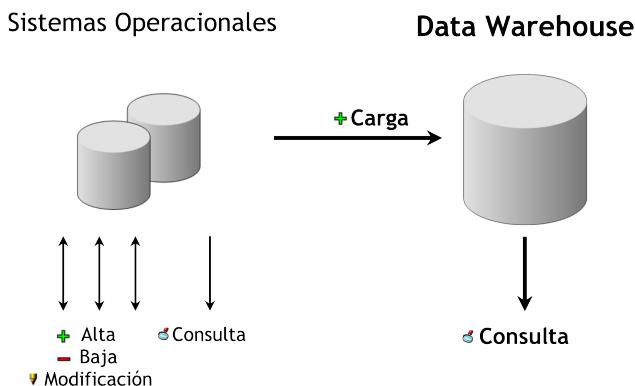


Figura 2.3: Data Warehouse, no volátil.

2.4. Cualidades

Una de las primeras cualidades que se puede mencionar del DW, es que maneja un gran volumen de datos, debido a que consolida en su estructura la información recolectada durante años, proveniente de diversas fuentes, en un solo lugar centralizado. Es por esta razón que el depósito puede ser soportado y mantenido sobre diversos medios de almacenamiento.

Además, como ya se ha mencionado, el almacén de datos presenta la información sumarizada y agregada desde múltiples versiones, y maneja información histórica.

Organiza y almacena los datos que se necesitan para el procesamiento analítico e informático, con el propósito de responder a preguntas de negocios y brindarles a los usuarios finales una interface amigable, comprensible y fácil de utilizar, para que los mismos puedan tomar decisiones sobre los datos sin tener que poseer demasiados conocimientos informáticos. El DW permite un acceso más directo, es decir, la información gira en torno al negocio, y es por ello que también los usuarios pueden sentirse cómodos al explorar los datos y encontrar relaciones complejas entre los mismos.

El DW no es solo datos, sino un conjunto de herramientas para consultar, analizar y presentar información, que permiten obtener o realizar análisis, reporting, extracción y explotación de los datos, con alta performance, para transformar dichos datos en información valiosa para la organización.

Con respecto a las tecnologías empleadas, en un almacén de datos se pueden encontrar las siguientes:

- Arquitectura cliente/servidor.
- Técnicas avanzadas para replicar, refrescar y actualizar datos.

- Software front-end, para acceso y análisis de datos.
- Herramientas para extraer, transformar y cargar datos en el depósito, desde múltiples fuentes muy heterogéneas.
- Sistema de Gestión de Base de Datos⁴ (SGBD).

Cabe destacar, que todas las cualidades expuestas anteriormente, son imposibles de saldar en un típico ambiente operacional, y esto es una de las razones de ser del DW.

2.5. Ventajas

A continuación se enumerarán algunas de las ventajas más sobresalientes que trae aparejada la implementación de un DW y que ejemplifican de mejor modo sus características y cualidades:

- Transforma datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- Integra y consolida diferentes fuentes de datos y departamentos empresariales, que anteriormente formaban islas, en una única plataforma sólida y centralizada.
- Provee la capacidad de analizar y explotar las diferentes áreas de trabajo y de realizar un análisis inmediato de las mismas.
- Permite reaccionar rápidamente a los cambios del mercado.
- Aumenta la competitividad en el mercado.
- Elimina la producción y el procesamiento de datos que no son utilizados ni necesarios, producto de aplicaciones mal diseñadas o ya no utilizadas.
- Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible. Información que los usuarios necesitan, en el momento adecuado y en el formato apropiado.
- Logra un impacto positivo sobre los procesos empresariales. Cuando los usuarios tienen acceso a una mejor calidad de información, la empresa puede lograr por sí misma: aprovechar el enorme valor potencial de sus recursos de información y transformarlo en valor verdadero; eliminar los retardos de los procesos empresariales que resultan de información incorrecta, inconsistente y/o inexistente; integrar y optimizar procesos a través del uso compartido e integrado de las fuentes de información; permitir al usuario adquirir mayor confianza acerca de sus propias decisiones y de las del resto, y lograr así, un mayor entendimiento de los impactos ocasionados.
- Aumento de la competitividad de los encargados de tomar decisiones.
- Los usuarios pueden acceder directamente a la información en línea, lo que contribuye a su capacidad para operar con mayor efectividad en las tareas rutinarias o no. Además, pueden tener a su disposición una gran cantidad de valiosa información multidimensional, presentada coherentemente como fuente única, confiable y disponible en sus estaciones de trabajo. Así mismo, los usuarios tienen la facilidad de contar con herramientas que les son familiares para manipular y evaluar la información obtenida en el DW, tales como: hojas de cálculo, procesadores de texto, software de análisis de datos, software de análisis estadístico, reportes, etc.
- Permite la toma de decisiones estratégicas y tácticas.

⁴Ver sección 4.3, en la página 61.

2.6. Desventajas

A continuación se consignarán algunas de las desventajas encontradas en la implementación de un DW:

- Requiere una gran inversión, debido a que su correcta construcción no es tarea sencilla y consume muchos recursos, además, su misma implementación implica desde la adquisición de herramientas de consulta y análisis, hasta la capacitación de los usuarios.
- Existe resistencia al cambio por parte de los usuarios.
- Los beneficios del almacén de datos son apreciados en el mediano y largo plazo. Este punto deriva del anterior, y básicamente se refiere a que no todos los usuarios confiarán en el DW en una primera instancia, pero sí lo harán una vez que comprueben su efectividad y ventajas. Además, su correcta utilización surge de la propia experiencia.
- Si se incluyen datos propios y confidenciales de clientes, proveedores, etc, el depósito de datos atentará contra la privacidad de los mismos, ya que cualquier usuario podrá tener acceso a ellos.
- Infravaloración de los recursos necesarios para la captura, carga y almacenamiento de los datos.
- Infravaloración del esfuerzo necesario para su diseño y creación.
- Incremento continuo de los requerimientos del usuario.

2.7. Redundancia

Debido a que el DW recibe información histórica de diferentes fuentes, sencillamente se podría suponer que existe una repetición de datos masiva entre el ambiente DW y el operacional. Por supuesto, este razonamiento es superficial y erróneo, de hecho, hay una mínima redundancia de datos entre ambos ambientes.

Para entender claramente lo antes expuesto, se debe considerar lo siguiente:

- Los datos del ambiente operacional se filtran antes de pertenecer al DW. Existen muchos datos que nunca ingresarán, ya que no conforman información necesaria o suficientemente relevante para la toma de decisiones.
- El horizonte de tiempo es muy diferente entre los dos ambientes.
- El almacén de datos contiene un resumen de la información que no se encuentra en el ambiente operacional.
- Los datos experimentan una considerable transformación, antes de ser cargados al DW. La mayor parte de los datos se alteran significativamente al ser seleccionados, consolidados y movidos al depósito.

En vista de estos factores, se puede afirmar que, la redundancia encontrada al cotejar los datos de ambos ambientes es mínima, ya que generalmente resulta en un porcentaje menor del 1 %.

2.8. Estructura

Los DW estructuran los datos de manera muy particular y existen diferentes niveles de esquematización y detalle que los delimitan.

En la siguiente figura se puede apreciar mejor su respectiva estructura.

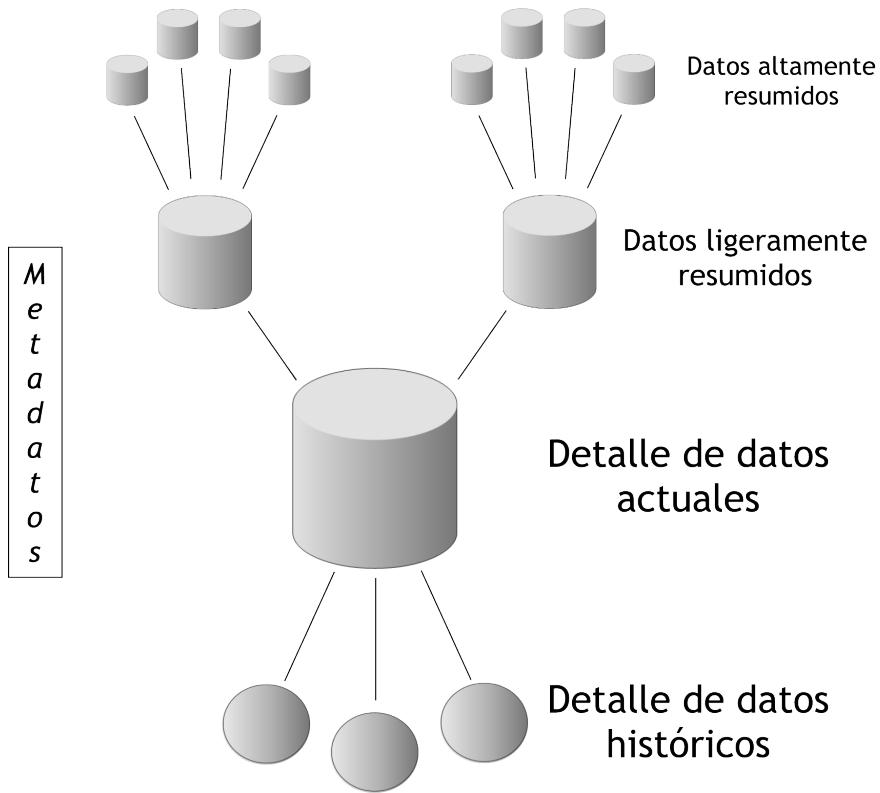


Figura 2.4: Data Warehouse, estructura.

Como se puede observar, los almacenes de datos están compuestos por diversos tipos de datos, que se organizan y dividen de acuerdo al nivel de detalle que posean.

A continuación se explicarán cada uno de estos tipos de datos:

- Detalle de datos actuales: son aquellos que reflejan las ocurrencias más recientes. Generalmente se almacenan en disco, aunque su administración sea costosa y compleja, con el fin de conseguir que el acceso a la información sea sencillo y veloz, ya que son bastante voluminosos. Su gran tamaño se debe a que los datos residentes poseen el más bajo nivel de granularidad, o sea, se almacenan a nivel de detalle. Por ejemplo, aquí es donde se guardaría el detalle de una venta realizada en tal fecha.
- Detalle de datos históricos: representan aquellos datos antiguos, que no son frecuentemente consultados. También se almacenan a nivel de detalle, normalmente sobre alguna forma de almacenamiento externa, ya que son muy pesados y en adición a esto, no son requeridos con mucha periodicidad. Este tipo de datos son

consistentes con los de Detalle de datos actuales. Por ejemplo, en este nivel, al igual que en el anterior, se encontraría el detalle de una venta realizada en tal fecha, pero con la particularidad de que el día en que se registró la venta debe ser lo suficientemente antigua, para que se considere como histórica.

- Datos ligeramente resumidos: son los que provienen desde un bajo nivel de detalle y sumarizan o agrupan los datos bajo algún criterio o condición de análisis. Habitualmente son almacenados en disco. Por ejemplo, en este caso se almacenaría la sumarización del detalle de las ventas realizadas en cada mes.
- Datos altamente resumidos: son aquellos que compactan aún más a los datos ligeramente resumidos. Se guardan en disco y son muy fáciles de acceder. Por ejemplo, aquí se encontraría la sumarización de las ventas realizadas en cada año.
- Metadatos⁵: representan la información acerca de los datos. De muchas maneras se sitúa en una dimensión diferente al de otros datos del DW, ya que su contenido no es tomado directamente desde el ambiente operacional.

2.9. Flujo de Datos

El DW posee un flujo de datos estándar y generalizado, el cual puede apreciarse mejor en la siguiente figura.

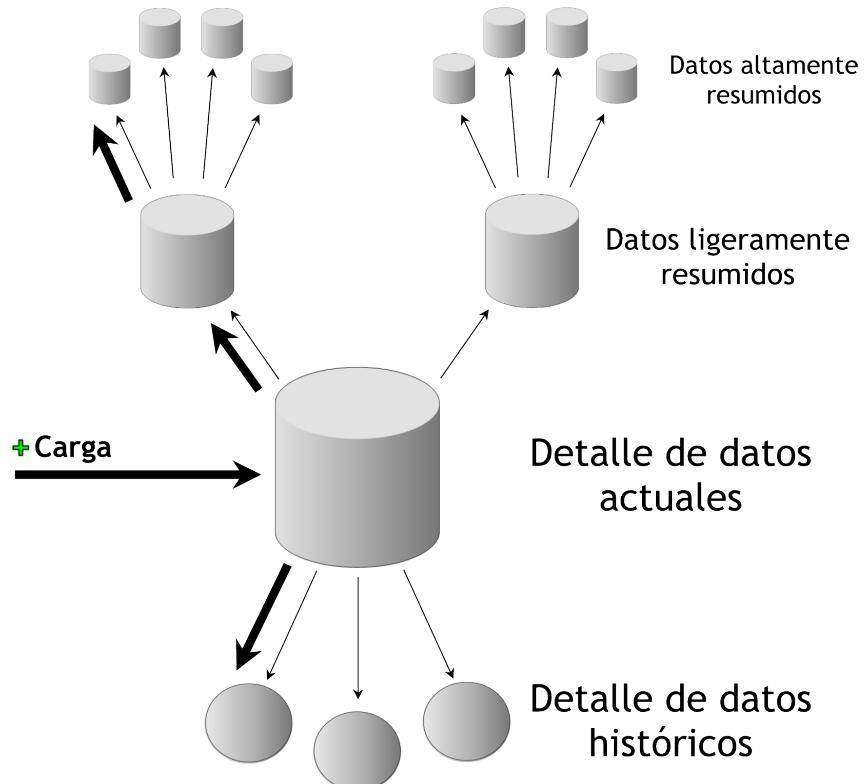


Figura 2.5: Data Warehouse, flujo de datos.

⁵Ver sección 3.4.12, en la página 40.

Cuando la información ingresa al depósito de datos se almacena a nivel de Detalle de datos actuales. Los datos permanecerán allí hasta que ocurra alguno de los tres eventos siguientes:

- Sean borrados del depósito de datos.
- Sean resumidos, ya sea a nivel de Datos ligeramente resumidos o a nivel de Datos altamente resumidos.
- Sean archivados a nivel de Detalle de datos históricos.

Capítulo 3

ARQUITECTURA DEL DATA WAREHOUSE

3.1. Introducción

En este punto y teniendo en cuenta que ya se han detallado claramente las características generales del DW, se definirán y describirán todos los componentes que intervienen en su arquitectura o ambiente.

A través del siguiente gráfico se explicitará la estructura del depósito de datos:

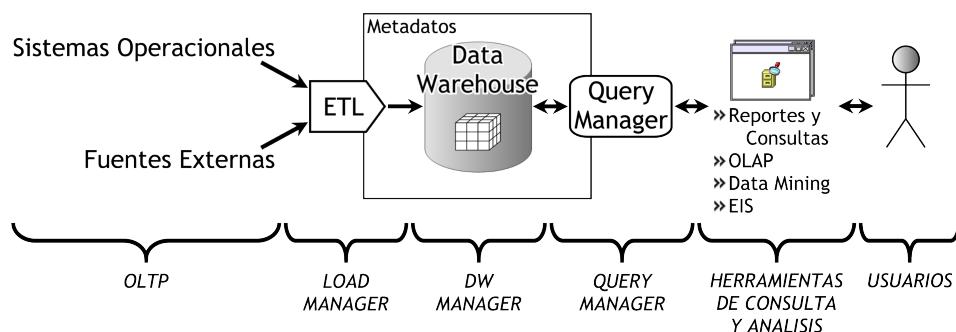


Figura 3.1: Data Warehouse, arquitectura.

Tal y como se puede apreciar, el ambiente del depósito de datos esta formado por diversos elementos que interactúan entre sí y que cumplen una función específica dentro del sistema. Por ello es que al abordar la exposición de cada elemento se lo hará en forma ordenada y teniendo en cuenta su relación con las demás partes.

Básicamente, la forma de operar del esquema superior se resume de la siguiente manera:

- Los datos son extraídos desde aplicaciones, bases de datos, archivos, etc. Esta información generalmente reside en diferentes tipos de sistemas y arquitecturas y tienen formatos muy variados.

- Los datos son integrados, transformados y limpiados, para luego ser cargados en el DW.
- La información del DW se estructura en cubos multidimensionales, los cuales preparan esta información para responder a consultas dinámicas con una buena performance. Los cubos pueden ser explotados tanto por un usuario no experto como por usuarios avanzados.
- Los usuarios acceden al DW utilizando diversas herramientas de consulta, exploración, análisis, reportes, etc.

A continuación se detallará cada uno de los componentes de la arquitectura del DW, teniendo como referencia siempre el gráfico antes expuesto, pero resaltando el tema que se tratará.

3.2. OLTP

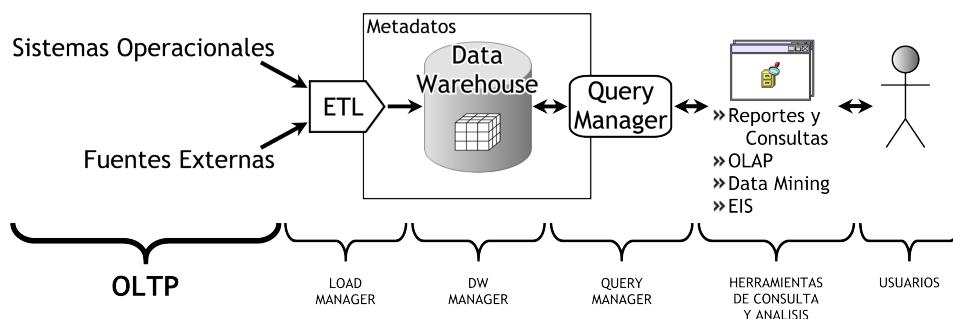


Figura 3.2: OLTP.

OLTP (On Line Transaction Processing), representa toda aquella información transaccional que genera la empresa en su accionar diario, además, de las fuentes externas con las que puede llegar a disponer.

Como ya se ha mencionado, estas fuentes de información, son de características muy disímiles entre sí, en formato, procedencia, función, etc.

Entre los OLTP más habituales que pueden existir en cualquier organización se encuentran:

- Archivos de textos.
- Hipertextos.
- Hojas de cálculos.
- Informes semanales, mensuales, anuales, etc.
- Bases de datos transaccionales.

3.3. Load Manager

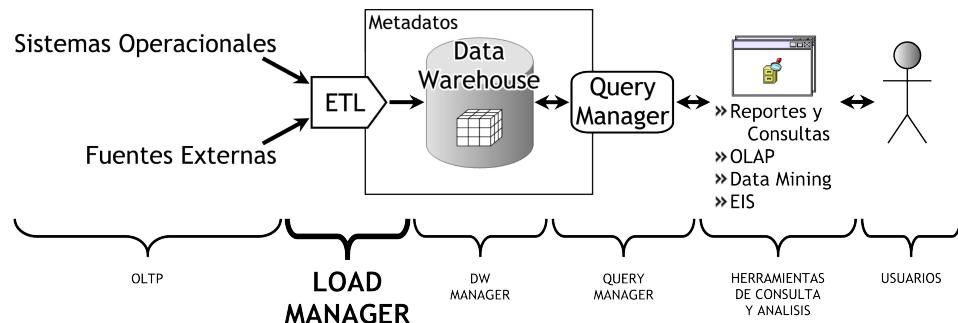


Figura 3.3: Load Manager.

Para poder extraer los datos desde los OLTP, para luego manipularlos, integrarlos y transformarlos, para posteriormente cargar los resultados obtenidos en el DW, es necesario contar con algún sistema que se encargue de ello. Precisamente los ETL (Extracción, Transformación y Carga) son los que cumplirán con tal fin.

Tal y como sus siglas lo indican, los ETL, extraen datos de las diversas fuentes que se requieran, los transforman para resolver posibles problemas de inconsistencias entre los mismos y finalmente, después de haberlos depurado se procede a su carga en el depósito de datos.

En síntesis, las funciones específicas de los ETL son tres:

- Extracción.
- Transformación.
- Carga.

3.3.1. Extracción

Es aquí, en donde, basándose en las necesidades y requisitos del usuario, se exploran las diversas fuentes OLTP que se tengan a disposición, y se extrae la información que se considere relevante al caso.

Si los datos operacionales residen en un SGBD Relacional, el proceso de extracción se puede reducir a, por ejemplo, consultas en SQL o rutinas programadas. En cambio, si se encuentran en un sistema propietario o fuentes externas, ya sean textuales, hipertextuales, hojas de cálculos, etc, la obtención de los mismos puede ser un tanto más difícil, debido a que, por ejemplo, se tendrán que realizar cambios de formato y/o volcado de información a partir de alguna herramienta específica.

Una vez que los datos son seleccionados y extraídos, se guardan en un almacenamiento intermedio, lo cual permite, entre otras ventajas:

- Manipular los datos sin interrumpir ni paralizar los OLTP, ni tampoco el DW.

- Almacenar y gestionar los metadatos que se generarán en los procesos ETL.
- Facilitar la integración de las diversas fuentes, internas y externas.

3.3.2. Transformación

Esta función es la encargada de convertir aquellos datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el DW. Estas acciones se llevan a cabo, debido a que pueden existir diferentes fuentes de información, y es vital conciliar un formato y forma única, definiendo estándares, para que todos los datos que ingresarán al DW estén integrados.

Los casos más comunes en los que se deberá realizar integración, son los siguientes:

- Codificación.
- Medida de atributos.
- Convenciones de nombramiento.
- Fuentes múltiples.

Además de lo antes mencionado, esta función se encarga de realizar el proceso de Limpieza de Datos (Data Cleansing).

3.3.2.1. Codificación

Una inconsistencia muy típica que se encuentra al intentar integrar varias fuentes de datos, es la de tener más de una sola forma de codificar un atributo en común. Por ejemplo, en el campo “género”, algunos diseñadores completan su valor con “0” y “1”, otros con “F” y “M”, otros con “Femenino” y “Masculino”, etc. Lo que se debe realizar en estos casos, es seleccionar o recodificar estos atributos, para que cuando la información llegue al DW, esté integrada de manera uniforme.

En la siguiente figura, se puede apreciar que de varias formas de codificar se escoge una, entonces cuando surge una codificación diferente a la seleccionada, se procede a su transformación.

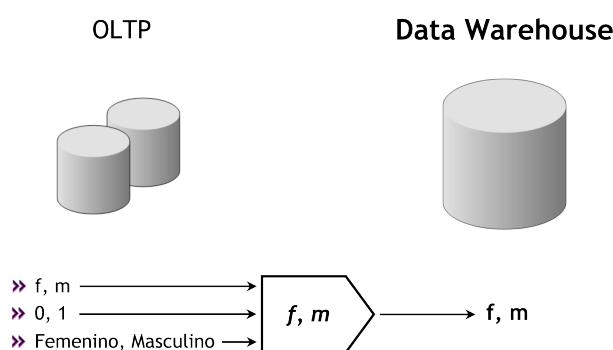


Figura 3.4: Transformación: codificación.

3.3.2.2. Medida de atributos

Los tipos de unidades de medidas utilizados para representar los atributos de una entidad, varían considerablemente entre sí, a través de los diferentes OLTP. Por ejemplo, al registrar la longitud de un producto determinado, de acuerdo a la aplicación que se emplee para tal fin, las unidades de medidas pueden ser explicitadas en centímetros, metros, pulgadas, etc.

En esta ocasión, se deberán estandarizar las unidades de medidas de los atributos, para que todas las fuentes de datos expresen sus valores de igual manera. Los algoritmos que resuelven estas inconsistencias son generalmente los más complejos.

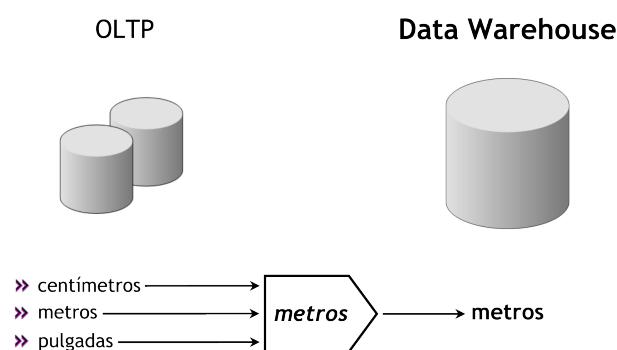


Figura 3.5: Transformación: medida de atributos.

3.3.2.3. Convenciones de nombramiento

Usualmente, un mismo atributo es nombrado de diversas maneras en los diferentes OLTP. Por ejemplo al referirse al nombre del proveedor, puede hacerse como "nombre", "razón_social", "proveedor", etc. Aquí, se debe utilizar la convención de nombramiento que para el usuario sea más comprensible.

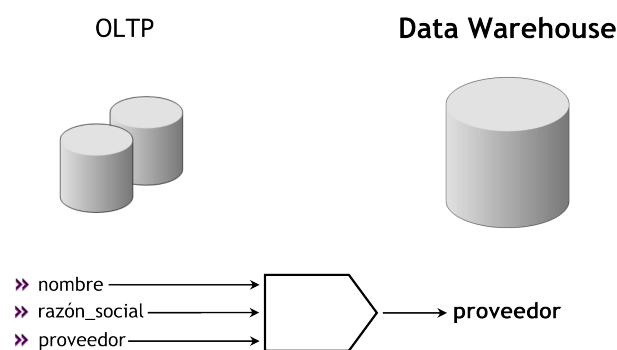


Figura 3.6: Transformación: convenciones de nombramiento.

3.3.2.4. Fuentes múltiples

Un mismo elemento puede derivarse desde varias fuentes. En este caso, se debe elegir aquella fuente que se considere más fiable y apropiada.

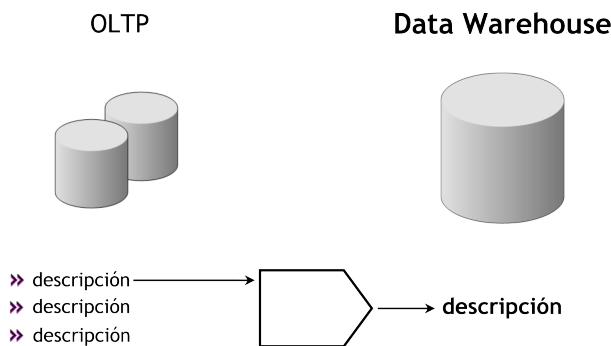


Figura 3.7: Transformación: fuentes múltiples.

3.3.2.5. Limpieza de datos

Su objetivo principal es el de realizar distintos tipos de acciones contra el mayor número de datos erróneos, inconsistentes e irrelevantes.

- Las acciones más típicas que se pueden llevar a cabo al encontrarse con Datos Anómalos (Outliers) son:
 - Ignorarlos.
 - Eliminar la columna.
 - Filtrar la columna.
 - Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
 - Reemplazar el valor.
 - Discretizar los valores de las columnas. Por ejemplo de 1 a 2, poner "bajo"; de 3 a 7, "óptimo"; de 8 a 10, "alto". Para que los outliers caigan en "bajo" o en "alto" sin mayores problemas.
- Las acciones que suelen efectuarse contra Datos Faltantes (Missing Values) son:
 - Ignorarlos.
 - Eliminar la columna.
 - Filtrar la columna.
 - Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
 - Reemplazar el valor.
 - Esperar hasta que los datos faltantes estén disponibles.

Un punto muy importante que se debe tener en cuenta al elegir alguna acción, es el de identificar el por qué de la anomalía, para luego actuar en consecuencia, con el fin de evitar que se repitan, agregándole de esta manera más valor a los datos de la organización. Se puede dar que en algunos casos, los valores faltantes sean inexistentes, ya que por ejemplo, un nuevo asociado o cliente, no poseerá consumo medio del último año.

3.3.3. Carga

Este proceso es el responsable de cargar la estructura de datos del DW con:

- Aquellos datos que han sido transformados y que residen en el almacenamiento intermedio.
- Aquellos datos de los OLTP que tienen correspondencia directa con el depósito de datos.

Se debe tener en cuenta, que los datos antes de moverse al almacén de datos, deben ser analizados con el propósito de asegurar su calidad, ya que este es un factor clave, que no debe dejarse de lado.

3.3.4. Proceso ETL

A continuación, se explicará en síntesis el accionar del proceso ETL, y cuál es la relación existente entre sus diversas funciones. En la siguiente figura se puede apreciar mejor lo antes descrito.

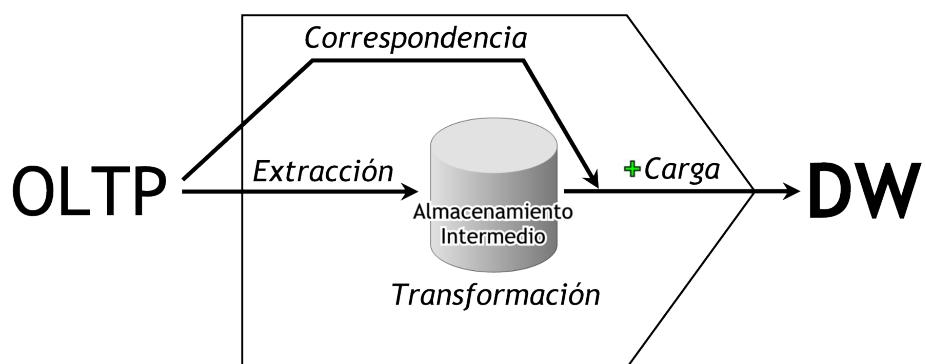


Figura 3.8: Proceso ETL.

Los pasos que se siguen son:

- Se extraen los datos relevantes desde los OLTP.
- Estos datos se depositan en un almacenamiento intermedio.
- Se integran y transforman los datos, para evitar inconsistencias.
- Finalmente los datos depurados son cargados desde el almacenamiento intermedio hasta el DW. Si existiesen correspondencias directas entre datos de los OLTP y el DW, se procede también a su respectiva carga.

3.3.5. Tareas del ETL

Los ETL, son los encargados de realizar dos tareas bien definidas:

- Carga Inicial (Initial Load).
- Actualización, mantenimiento o refresco periódico (siempre teniendo en cuenta un intervalo de tiempo predefinido para tal operación).

La carga inicial, se refiere precisamente a la primera carga de datos que se le realizará al DW. Por lo general esta tarea consume un tiempo bastante considerable, ya que se deben insertar registros que han sido generados aproximadamente, y en casos ideales, durante más de cinco años.

Los mantenimientos periódicos mueven pequeños volúmenes de datos, y su frecuencia está dada en función del gránulo del DW y los requerimientos del usuario. El objetivo de esta tarea es añadir al depósito aquellos datos nuevos que se fueron generando desde el último refresco.

Antes de realizar una nueva actualización, es necesario identificar si se han producido cambios en las fuentes originales de los datos recogidos, desde la fecha del último mantenimiento, a fin de no atentar contra la consistencia del DW.

Para efectuar esta operación, se pueden realizar las siguientes acciones:

- Cotejar las instancias de los OLTP involucrados.
- Utilizar disparadores en los OLTP.
- Recurrir a Marcas de Tiempo (Time Stamp), en los registros de los OLTP.
- Hacer uso de técnicas mixtas.

Si este control consume demasiado tiempo y esfuerzo, o simplemente no puede llevarse a cabo por algún motivo en particular, existe la posibilidad de cargar el DW desde cero, este proceso se denomina Carga Total (Full Load).

3.4. Data Warehouse Manager

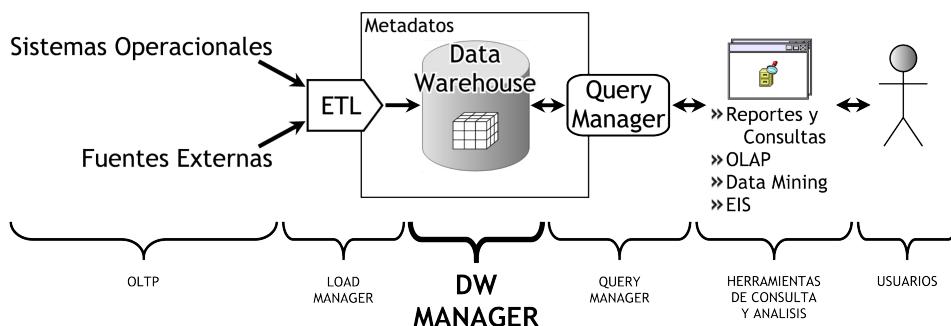


Figura 3.9: Data Warehouse Manager.

El DW Manager presenta las siguientes características y funciones:

- Transforma e integra los datos fuentes y del almacenamiento intermedio en un modelo adecuado para la toma de decisiones.
- Gestiona el depósito de datos y lo organiza en torno a una base de datos multi-dimensional¹, que tal y como lo indica su nombre almacena los datos en diversas

¹Ver sección 3.4.1, en la página 27.

dimensiones, que conforman un cubo multidimensional², en donde el cruce de los valores de los atributos de cada dimensión³ a lo largo de las abscisas, determinan un hecho específico. Los cálculos que se aplican sobre las dimensiones son matriciales, los cuales se procesan dando como resultado reportes tabulares.

- Permite realizar todas las funciones de definición y manipulación del depósito de datos, para poder soportar todos los procesos de gestión del mismo.
- Es el encargado de ejecutar y definir las políticas de particionamiento. El objetivo de realizar esto, es conseguir una mayor eficiencia y performance en las consultas al no tener que manejar todo el grueso de los datos. Esta política debe aplicarse sobre la tabla de hechos⁴ que, como se explicará más adelante, es en la que se almacena toda la información que será analizada.
- Realiza copias de resguardo incrementales o totales de los datos del DW.
- Se constituye típicamente al combinar un SGBD con aplicaciones y software propietarios.
- Posee un repositorio de datos propio.
- Gestiona y mantiene metadatos.

3.4.1. Base de datos multidimensional

Las bases de datos multidimensionales, proveen una estructura que permite tener acceso flexible a los datos, para explorar y analizar sus relaciones, y resultados consecuentes. Estas se pueden visualizar como un cubo multidimensional, en donde las variables asociadas existen a lo largo de varios ejes o dimensiones, y la intersección de las mismas representa la medida, indicador o el hecho que se está evaluando.

En la siguiente representación matricial se puede ver más claramente lo que se acaba de decir.

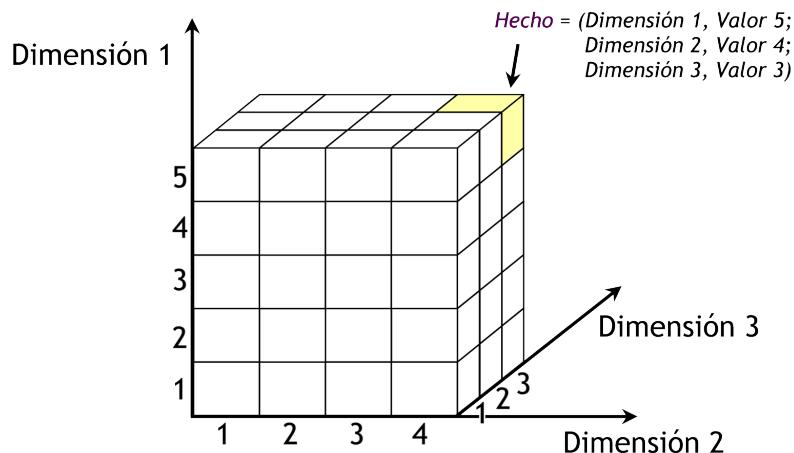


Figura 3.10: Cubo multidimensional.

²Ver sección 3.4.11, en la página 38.

³Ver sección 3.4.2, en la página 28.

⁴Ver sección 3.4.3, en la página 31.

En el cubo de la figura anterior, existen tres dimensiones, “Dimensión 1”, “Dimensión 2” y “Dimensión 3”, cada una con sus respectivos valores asociados. También, se ha seleccionado un hecho al azar para demostrar su correspondencia con los valores de las dimensiones. En este caso, la medida a la que se hace referencia, representa el cruce del Valor “5” de “Dimensión 1”, con el Valor “4” de “Dimensión 2” y con el Valor “3” de “Dimensión 3”. Se puede observar que el resultado del análisis está dado por los cruces matriciales de acuerdo a los valores de las dimensiones seleccionadas.

Las bases de datos multidimensionales implican tres variantes posibles de modelamiento, que permiten realizar consultas de soporte de decisión:

- Esquema en estrella⁵ (Star Scheme).
- Esquema copo de nieve⁶ (Snowflake Scheme).
- Esquema constelación⁷ o copo de estrellas (Starflake Scheme).

Los mencionados esquemas pueden ser implementados de diversas maneras, que, independientemente al tipo de arquitectura, requieren que toda la estructura de datos este desnormalizada o semi desnormalizada, para evitar desarrollar uniones (Join) complejas para acceder a la información, con el fin de agilizar la ejecución de consultas. Los diferentes tipos de implementación son los siguientes:

- Relacional – ROLAP⁸.
- Multidimensional – MOLAP⁹.
- Híbrido – HOLAP¹⁰.

3.4.2. Tablas de Dimensiones

Las tablas de dimensiones definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio.

Representan los ejes del cubo, y los aspectos de interés, mediante los cuales el usuario podrá filtrar y manipular la información almacenada en la tabla de hechos.

En la siguiente figura se pueden apreciar algunos ejemplos:

GEOGRAFIA	PRODUCTOS	CLIENTES	FECHAS
id_Geografía País Provincia Ciudad Barrio	id_Producto Rubro Tipo NombreProducto	id_Cliente NombreCliente	id_Fecha Año Trimestre Mes Día

Figura 3.11: Tablas de Dimensiones.

⁵Ver sección 3.4.4, en la página 32.

⁶Ver sección 3.4.5, en la página 34.

⁷Ver sección 3.4.6, en la página 35.

⁸Ver sección 3.4.8, en la página 36.

⁹Ver sección 3.4.9, en la página 37.

¹⁰Ver sección 3.4.10, en la página 37.

Como se puede observar, cada tabla posee un identificador único y al menos un atributo que describe los criterios de análisis relevantes de la organización, estos son por lo general de tipo texto. Usualmente la cantidad de tablas de dimensiones, aplicadas a un tema de interés en particular, varían entre tres y quince.

Así mismo, dentro de estas tablas pueden existir jerarquías¹¹ de datos, además, de acuerdo a las dimensiones del negocio, estará dada la granularidad que adoptará el modelo.

Los datos dentro de estas tablas, que proveen información del negocio o que describen alguna de sus características, son llamados datos de referencia. Entonces, se puede afirmar que una tabla de dimensión posee una clave primaria y uno o más datos de referencia.

3.4.2.1. Dimensión Tiempo

En un DW, la dimensión Tiempo es obligatoria, y la definición de granularidad y jerarquía de la misma depende de la dinámica del negocio que se esté analizando, toda la información dentro de la bodega, como ya se ha explicado, posee su propio sello de tiempo que determina la ocurrencia y ubicación con elementos en iguales condiciones, representando de esta manera diferentes versiones de una misma situación.

Es importante tener en cuenta que el tiempo no es solo una secuencia cronológica representada de forma numérica, sino que posee fechas especiales que inciden notablemente en las actividades de la organización. Esto se debe a que los usuarios podrán por ejemplo analizar las ventas realizadas teniendo en cuenta el día de la semana en que se produjeron, quincena, mes, trimestre, semestre, año, etc.

Existen muchas maneras de diseñar esta tabla, y en adición a ello no es una tarea sencilla de llevar a cabo. Por estas razones se considera una buena práctica evaluar con cuidado la temporalidad de los datos, la forma en que trabaja la organización, los resultados que se esperan obtener del almacén de datos relacionados con una unidad de tiempo y la flexibilidad que se desea obtener de dicha tabla.

Si bien, el lenguaje SQL ofrece funciones del tipo DATE, en la dimensión Tiempo, se modelan y presentan atributos temporales que no pueden calcularse en SQL, lo cual le añade una ventaja más.

3.4.2.2. Jerarquías

Una jerarquía representa una relación¹² lógica entre dos o más atributos dentro de una misma dimensión.

Las jerarquías poseen las siguientes características:

- Pueden existir varias en una misma dimensión.
- Están compuestas por dos o más niveles.
- Se tiene una relación “1-n” entre atributos consecutivos de un nivel superior y uno inferior.

¹¹Ver sección 3.4.2.2, en la página 29.

¹²Ver sección 3.4.2.3, en la página 30.

La principal ventaja de manejar jerarquías, reside en poder analizar los datos desde su nivel más general al más detallado y viceversa.

La siguiente figura muestra un pequeño ejemplo de lo recién expuesto:

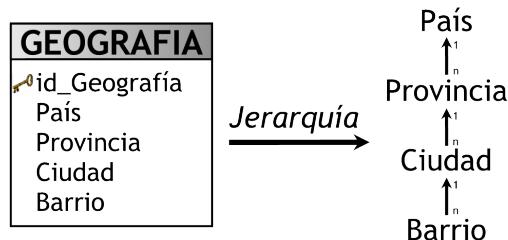


Figura 3.12: Jerarquía de "GEOGRAFIA".

En la figura anterior, se puede apreciar claramente como se construye una jerarquía. Se coloca el atributo más general en la cabecera y se comienza a disgregar los niveles hacia abajo, dependiendo siempre del rango del dato. En este caso, la figura se explica como sigue:

- Un barrio pertenece solo a una ciudad. Una ciudad puede poseer uno o más barrios.
- Una ciudad pertenece solo a una provincia. Una provincia puede poseer una o más ciudades.
- Una provincia pertenece solo a un país. Un país puede poseer una o más provincias.

3.4.2.3. Relación

Una relación representa la forma en que dos atributos interactúan dentro de una jerarquía. Existen básicamente dos tipos de relaciones:

- Explícitas: son las más comunes y se pueden modelar a partir de atributos directos y están en línea continua de una jerarquía, por ejemplo, la figura anterior posee este tipo de relación, en donde un país posee una o más provincias y una provincia pertenece solo a un país.
- Implícitas: son las que ocurren en la vida real, pero su relación no es de vista directa, por ejemplo, un país tiene uno o más ríos, pero un río pertenece a uno o más países.

3.4.2.4. Granularidad

La granularidad representa el nivel de detalle al que se desea almacenar la información sobre el negocio que se esté analizando. Por ejemplo, los datos referentes a ventas o compras realizadas por la empresa, pueden registrarse día a día, en cambio, los datos pertinentes a pagos de sueldos o cuotas de socios, podrán almacenarse a nivel de mes.

Mientras mayor sea el nivel de detalle de los datos, se tendrán mayores posibilidades de análisis, ya que los mismos podrán ser resumidos o sumarizados. Es decir, los datos que posean granularidad fina (nivel de detalle) podrán ser resumidos hasta obtener una granularidad media o gruesa. No sucede lo mismo en sentido contrario, ya que

por ejemplo, los datos almacenados con granularidad media podrán resumirse, pero no tendrán la facultad de ser analizados a nivel de detalle. O sea, si la granularidad con que se guardan los registros es a nivel de día, estos datos podrán sumarizarse por semana, mes, semestre y año, en cambio, si estos registros se almacenan a nivel de mes, podrán sumarizarse por semestre y año, pero no lo podrán hacer por día y semana.

3.4.3. Tablas de Hechos

Las tablas de hechos contienen los hechos, medidas o indicadores que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones.

Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones.

Los datos presentes en las tablas de hechos constituyen el volumen de la bodega, y pueden estar compuestos por millones de registros dependiendo de su granularidad y de los intervalos de tiempo de los mismos. Los más importantes son los de tipo numérico.

El registro del hecho posee una clave primaria que está compuesta por las claves primarias de las tablas de dimensiones relacionadas a este.

En la siguiente imagen se puede apreciar un ejemplo de lo antes mencionado:

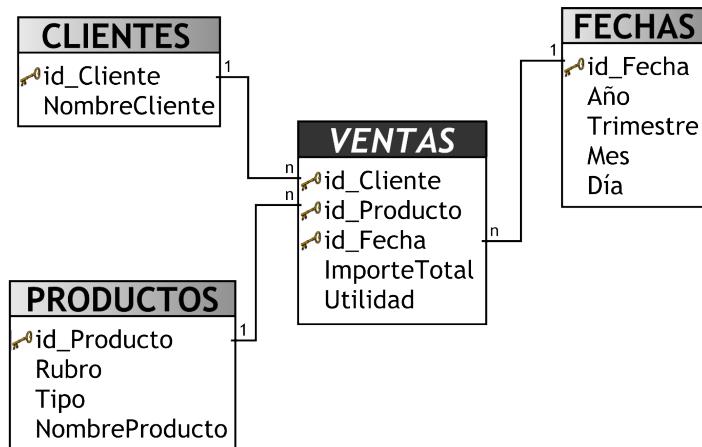


Figura 3.13: Tablas de Hechos.

Como se muestra en la figura anterior, la tabla de hechos “VENTAS” se ubica en el centro, e irradiando de ella se encuentran las tablas de dimensiones “CLIENTES”, “PRODUCTOS” y “FECHAS”, que están conectadas mediante sus claves primarias. Es por ello precisamente que la clave primaria de la tabla de hechos es la combinación de las claves primarias de sus dimensiones. Las medidas en este caso son “ImporteTotal” y “Utilidad”.

A continuación, se entrará más en detalle sobre la definición de un hecho, también llamado dato agregado:

- Los hechos son todas aquellas summarizaciones o acumulaciones preestablecidas que residen en una tabla de hechos para agilizar las consultas y permitir que los da-

tos puedan ser accedidos por las diferentes dimensiones, y desde luego, explorados por ellas. Las sumarizaciones no están referidas solo a sumas, sino que también a promedios, mínimos, máximos, totales por sector, porcentajes, fórmulas predefinidas, etc, dependiendo de los requerimientos de información del negocio.

Para ejemplificar este nuevo concepto de hechos, se enumerarán algunos que son muy típicos y fáciles de comprender:

- ImporteTotal = SUM(precioProducto * cantidadVendida)
- Rentabilidad = SUM(utilidad / PN)
- PromedioGeneral = AVG(notasFinales)
- CantidadVentas = COUNT(idVenta)

3.4.4. Esquema en Estrella

El esquema en estrella, consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves. En la siguiente figura se puede apreciar un esquema en estrella estándar:

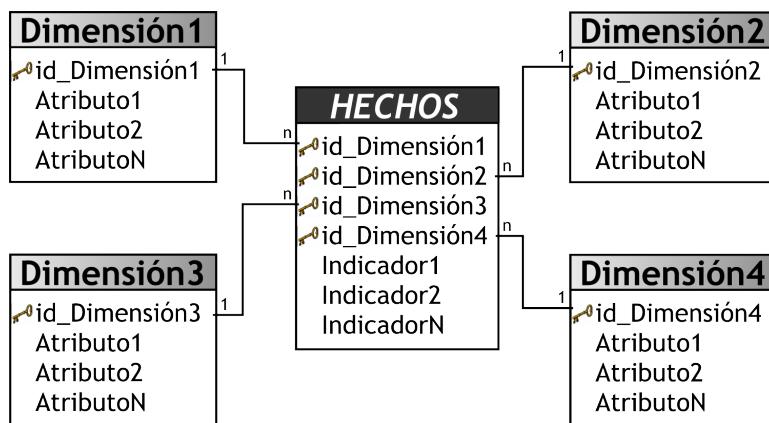
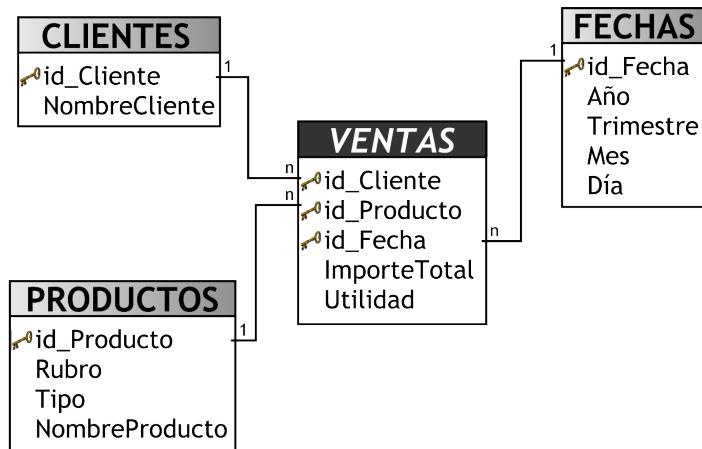
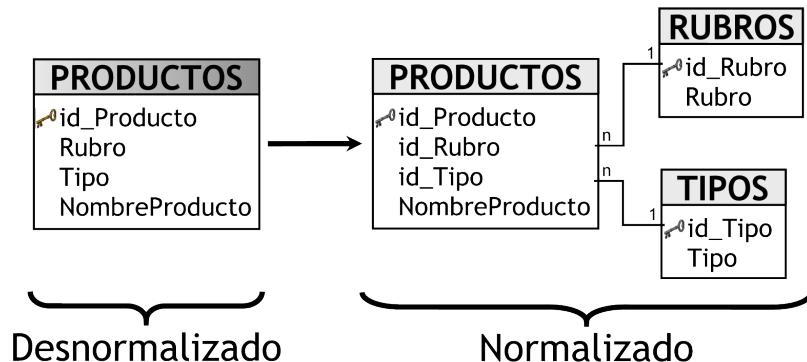


Figura 3.14: Esquema en Estrella.

El modelo ejemplificado cuando se abordó el tema de las tablas de hechos, es un esquema en estrella, por lo cual se lo volverá a mencionar para explicar sus cualidades.

**Figura 3.15: Esquema en Estrella, ejemplo.**

Este modelo debe estar totalmente desnormalizado, es decir que no puede presentarse en tercera forma normal (3ra FN), es por ello que por ejemplo, la dimensión "PRODUCTOS" contiene los atributos "Rubro", "Tipo" y "NombreProducto". Si se normaliza esta tabla, se obtendrá el siguiente resultado:

**Figura 3.16: Desnormalización.**

Cuando se normaliza, se pretende eliminar la redundancia, la repetición de datos y que las claves sean independientes de las columnas, pero en este tipo de modelos se requiere no evitar precisamente esto.

Las ventajas que trae aparejada la desnormalización, son las de obviar uniones (Join) entre las tablas cuando se realizan consultas, procurando así un mejor tiempo de respuesta y una mayor sencillez con respecto a su utilización. El punto en contra, es que se genera un cierto grado de redundancia, pero el ahorro de espacio no es significativo.

El esquema en estrella es el más simple de interpretar y optimiza los tiempos de respuesta ante las consultas de los usuarios. Este modelo es soportado por casi todas las herramientas de consulta y análisis, y los metadatos son fáciles de documentar y man-

tener, sin embargo es el menos robusto para la carga y es el más lento de construir.

A continuación se destacarán algunas características de este modelo, que ayudarán a comprender mejor el por qué de sus ventajas:

- Posee los mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.
- Existe paralelismo entre su diseño y la forma en que los usuarios visualizan y manipulan los datos.
- Simplifica el análisis.
- Facilita la interacción con herramientas de consulta y análisis.

3.4.5. Esquema Copo de Nieve

Este esquema representa una extensión del modelo en estrella cuando las dimensiones se organizan en jerarquías de dimensiones.

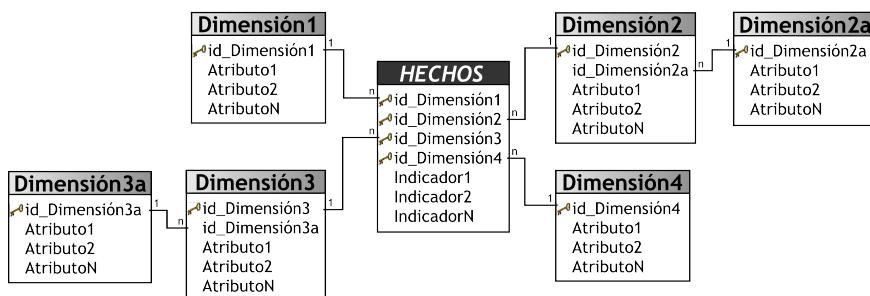


Figura 3.17: Esquema Copo de Nieve.

Como se puede apreciar en la figura anterior, existe una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quienes a su vez pueden estar relacionadas o no con una o más tablas de dimensiones.

Este modelo es más cercano a un modelo de entidad relación, que al modelo en estrella, debido a que sus tablas de dimensiones están normalizadas.

Una de los motivos principales de utilizar este tipo de modelo, es la posibilidad de segregar los datos de las dimensiones y proveer un esquema que sustente los requerimientos de diseño. Otra razón es que es muy flexible y puede implementarse después de que se haya desarrollado un esquema en estrella.

Se pueden definir las siguientes características de este tipo de modelo:

- Posee mayor complejidad en su estructura.
- Hace una mejor utilización del espacio.
- Es muy útil en tablas de dimensiones de muchas tuplas.

- Las dimensiones están normalizadas, por lo que requiere menos esfuerzo de diseño.
- Puede desarrollar clases de jerarquías fuera de las dimensiones, que permiten realizar análisis de lo general a lo detallado y viceversa.

A pesar de todas las características y ventajas que trae aparejada la implementación del esquema copo de nieve, existen dos grandes inconvenientes de ello:

- Si se poseen múltiples dimensiones, cada una de ellas con varias jerarquías, se creará un número de dimensiones bastante considerable, que pueden llegar al punto de ser inmanejables.
- Al existir muchas uniones y relaciones entre tablas, el desempeño puede verse reducido.

3.4.6. Esquema Constelación

Este modelo está compuesto por una serie de esquemas en estrella, y tal como se puede apreciar en la siguiente figura, está formado por una tabla de hechos principal ("HECHOS_A") y por una o más tablas de hechos auxiliares ("HECHOS_B"), las cuales pueden ser summarizaciones de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones.

No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, ya que, las tablas de hechos auxiliares pueden vincularse con solo algunas de las tablas de dimensiones asignadas a la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones.

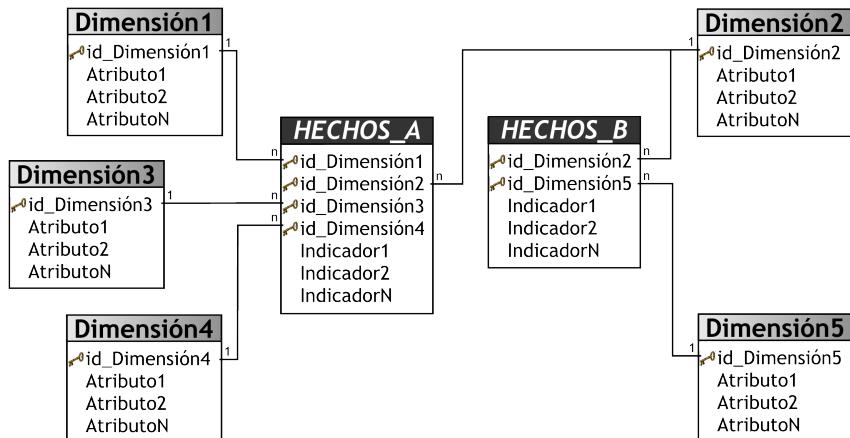


Figura 3.18: Esquema Constelación.

Su diseño y cualidades son muy similares a las del esquema en estrella, pero posee una serie de diferencias con el mismo, que son precisamente las que lo destacan y caracterizan. Entre ellas se pueden mencionar:

- Permite tener más de una tabla de hechos, por lo cual se podrán analizar más aspectos claves del negocio con un mínimo esfuerzo adicional de diseño.

- Contribuye a la reutilización de dimensiones, ya que una misma dimensión puede utilizarse para varias tablas de hechos.
- No es soportado por todas las herramientas de consulta y análisis.

3.4.7. OLTP vs DW

Debido a que, ya se han explicado y caracterizado los distintos tipos de esquemas del DW, se procederá a exponer las razones de su utilización, como así también las causas de por qué no se emplean simplemente las estructuras de las bases de datos tradicionales:

- Los OLTP son diseñados para soportar el procesamiento de información diaria de las empresas, y el énfasis recae en maximizar la capacidad transaccional de sus datos. Su estructura es altamente normalizada, para brindar mayor eficiencia a sistemas con muchas transacciones que acceden a un pequeño número de registros y están fuertemente condicionadas por los procesos operacionales que deben soportar, para la óptima actualización de sus datos. Esta estructura es ideal para llevar a cabo el proceso transaccional diario, brindar consultas sobre los datos cargados y tomar decisiones diarias, en cambio los esquemas de DW están diseñados para poder llevar a cabo procesos de consulta y análisis para luego tomar decisiones estratégicas de alto nivel.

A continuación se presentará una tabla comparativa entre los dos ambientes, que resume sus principales diferencias:

	OLTP	Data Warehouse
Objetivo	Soportar actividades transaccionales diarias.	Consultar y analizar información estratégica y táctica.
Tipo de datos	Operacionales.	Para la toma de decisiones.
Modelo de datos	Normalizado.	Desnormalizado.
Consulta	SQL.	SQL más extensiones.
Datos consultados	Actuales.	Actuales e históricos.
Horizonte de tiempo	60 - 90 días.	5 - 10 años.
Tipos de consultas	Repetitivas, predefinidas	No previsibles, dinámicas
Nivel de almacenamiento	Nivel de detalle.	Nivel de detalle y diferentes niveles de sumarización.
Acciones disponibles	Alta, baja, modificación y consulta.	Carga y consulta.
Número de transacciones	Elevado	Medio o bajo
Tamaño	Pequeño - Mediano.	Grande.
Tiempo de respuesta	Pequeño (segundos - minutos).	Variable (minutos - horas).
Orientación	Orientado a las aplicaciones.	Orientado al negocio.
Sello de tiempo	La clave puede o no tener un elemento de tiempo.	La clave tiene un elemento de tiempo.
Estructura	Generalmente estable.	Generalmente varía de acuerdo a su propia evolución y utilización.

Figura 3.19: OLTP vs Data Warehouse.

3.4.8. ROLAP

Este tipo de organización física se implementa sobre tecnología relacional, pero disponen de algunas facilidades para mejorar el rendimiento.

Es decir, ROLAP (Relational On Line Analytic Processing) cuenta con todos los beneficios de una SGBD Relacional a los cuales se les provee extensiones y herramientas para poder utilizarlo como un Sistema Gestor de DW.

Entre las características más importantes y sobresalientes de ROLAP, se encuentran las siguientes:

- Almacena la información en una base de datos relacional.
- Posee tres capas lógicas: de almacenamiento, de análisis y de presentación.
- Utiliza índices de mapas de bits.
- Utiliza índices de Join.
- Posee técnicas de particionamiento de datos.
- Posee optimizadores de consultas.
- Cuenta con extensiones del SQL (drill-up, drill-down, etc).

Como se aclaró anteriormente, el almacén de datos se organiza a través de una base de datos multidimensional, sin embargo, puede ser soportado por un SGBD Relacional. Para lograr esto se utilizan los diferentes esquemas, en estrella, copo de nieve y constelación, los cuales transformarán el modelo multidimensional y permitirán que pueda ser gestionado por un SGDB Relacional, ya que solo se almacenarán tablas.

3.4.9. MOLAP

El objetivo de los sistemas MOLAP (Multidimensional On Line Analytic Processing) es almacenar físicamente los datos en estructuras multidimensionales de manera que la representación externa y la interna coincidan.

Para ello, se dispone de estructuras de almacenamiento específicas (Arrays) y técnicas de compactación de datos que favorecen el rendimiento del depósito de datos.

Las principales características de MOLAP son:

- Posee tecnología optimizada para consultas y análisis, basada en el modelo multidimensional.
- Cuenta con un motor especializado.
- Provee herramientas limitadas y propietarias.
- No es adecuada para muchas dimensiones.
- Construye y almacena datos en estructuras multidimensionales.

Las herramientas de consulta y exploración OLAP¹³, presentan estas estructuras multidimensionales.

3.4.10. HOLAP

HOLAP (Hybrid On Line Analytic Processing) constituye un sistema híbrido entre MOLAP y ROLAP, que combina estas dos implementaciones para almacenar algunos datos en un motor relacional y otros en una base de datos multidimensional.

¹³Ver sección 3.6.2, en la página 52.

3.4.11. Cubo Multidimensional

Un cubo multidimensional o hipercubo, representa o convierte los datos planos que se encuentran en filas y columnas, en una matriz de N dimensiones.

En los ejes de la matriz se encuentran los criterios o dimensiones, mediante los cuales se analizará el negocio y en los cruces o intersecciones, residen los valores o hechos que se desean consultar.

Para comprender más respecto al concepto de cubo multidimensional, primero se expondrá un cubo de dos dimensiones:

PRODUCTOS	P1	40	25	60
	P2	21	55	45
	P3	13	32	43
	M1 M2 M3	MARCAS		

Figura 3.20: Cubo multidimensional de dos dimensiones.

Este cubo posee dos ejes o dimensiones, “PRODUCTOS” y “MARCAS”. La intersección de los ejes representa la cantidad de unidades vendidas de cada producto con su respectiva marca. Con el objetivo de que se entienda mejor la relación existente entre el cubo y un modelo de DW, se presentará a continuación el esquema en estrella que sería la base del cubo en cuestión.



Figura 3.21: Esquema en Estrella de dos dimensiones.

Aquí, se pueden apreciar nuevamente las dos dimensiones, “MARCAS” y “PRODUCTOS”, que condicionan el hecho “Venta”, que está dado por la siguiente fórmula:

- $\text{Venta} = \text{SUM}(\text{unidadesVendidas})$

El campo “unidadesVendidas” representa las cantidades totales que se han vendido de cada producto y marca. Por ejemplo, de acuerdo al cubo que se explicitó al inicio:

- Las unidades vendidas del producto “P1” de la marca “M1” son 40.

- Las unidades vendidas del producto “P1” de la marca “M2” son 25.
- Las unidades vendidas del producto “P1” de la marca “M3” son 60.

Ahora, al esquema en estrella que ya se ha presentado se le agregará una nueva dimensión, “CLIENTES”:

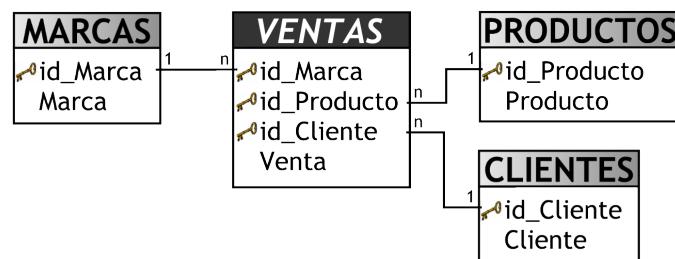


Figura 3.22: Esquema en Estrella de tres dimensiones.

Entonces, el nuevo cubo quedará expresado de la siguiente manera:

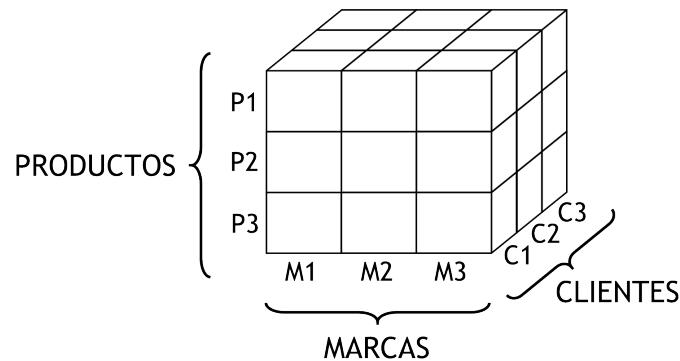
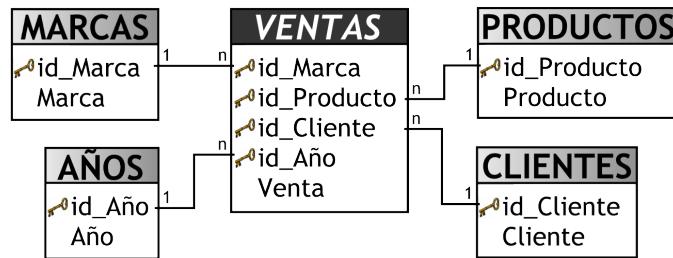


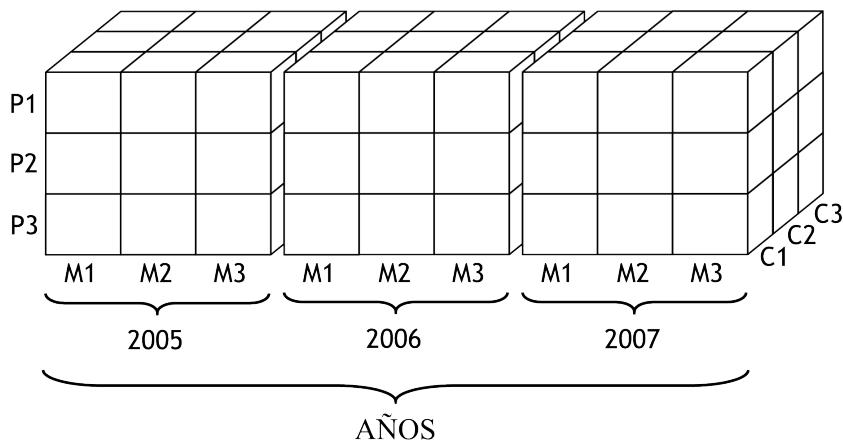
Figura 3.23: Cubo multidimensional de tres dimensiones.

En este caso los valores del hecho “Venta” están dados de acuerdo a la cantidad de unidades vendidas de cada producto, de cada marca, a cada cliente.

Para finalizar, se añadirá una cuarta dimensión, la dimensión de tiempo “AÑOS”:

**Figura 3.24: Esquema en Estrella de cuatro dimensiones.**

De esta manera, el cubo resultante será:

**Figura 3.25: Cubo multidimensional de cuatro dimensiones.**

Los hechos en este momento, estarán condicionados por la cantidad de unidades vendidas de cada producto, de cada marca, a cada cliente, en cada año.

Esta última imagen, demuestra claramente los conceptos expuestos de la tabla de dimensión tiempo, donde se decía que pueden existir diferentes versiones de la situación del negocio. Se puede apreciar que cada periodo de tiempo es representado por un cubo diferente.

Cabe destacar, que los ejemplos seleccionados, son solo con fines didácticos, ya que como se ha dicho, en todo esquema es requisito fundamental agregar la dimensión tiempo, tal como se hizo cuando se agregó la cuarta dimensión, por ende, de implementarse los modelos anteriores serían erróneos.

3.4.12. Metadatos

Los metadatos son datos que describen o dan información de otros datos, que en este caso, existen en el DW. Brindan información de localización, estructura y significado de los datos, básicamente mapean los mismos.

El concepto de metadatos es análogo al uso de índices para localizar objetos en lugar de datos.

Su creación es una novedad que aporta el DW como sistema de análisis de información. Los metadatos constituyen un archivo al que se lo considera como un diccionario de estructuras de datos, cuyo objetivo es asistir en los procesos de consulta a la base de datos, que los usuarios utilizarán posteriormente para analizar y explotar la información.

Las funciones que cumplen los metadatos en el almacén de datos son muy importantes y significativas, algunas de ellas son:

- Facilitan el flujo de trabajo, convirtiendo datos automáticamente de un formato a otro.
- Contienen un directorio para facilitar la búsqueda y descripción de los contenidos del DW, tales como: bases de datos, tablas, nombres de atributos, summarizaciones, acumulaciones, reglas de negocios, estructuras y modelos de datos, relaciones de integridad, etc.
- Poseen un guía para el mapping¹⁴, de cómo se transforman e integran los datos fuentes operacionales y externos al ambiente del depósito de datos.
- Almacenan las referencias de los algoritmos utilizados para la esquematización entre el detalle de datos actuales, con los datos ligeramente resumidos y éstos con los datos altamente resumidos, etc.
- Contienen las definiciones del sistema de registro desde el cual se construye el DW.

Dentro del ambiente DW, existen diferentes tipos de Metadatos:

- Los metadatos de los procesos ETL, referidos a las diversas fuentes utilizadas, reglas de extracción, transformación, limpieza, depuración y carga de los datos al depósito.
- Los metadatos operacionales, que son los que básicamente almacenan todos los contenidos del DW, para que este pueda desempeñar sus tareas.
- Los metadatos de consulta, que contienen las reglas para analizar y explotar la información del almacén, tales como drill-up y drill-down. Son estos metadatos los que las herramientas de análisis y consulta del usuario emplearán para realizar documentaciones y para navegar por los datos.

3.4.12.1. Mapping

El término mapping, se refiere a relacionar un conjunto de objetos, tal como actualmente están almacenados en memoria o en disco, con otros objetos. Por ejemplo: una estructura de base de datos lógica, se proyecta sobre la base de datos física.

¹⁴Ver sección 3.4.12.1, en la página 41.

3.5. Query Manager

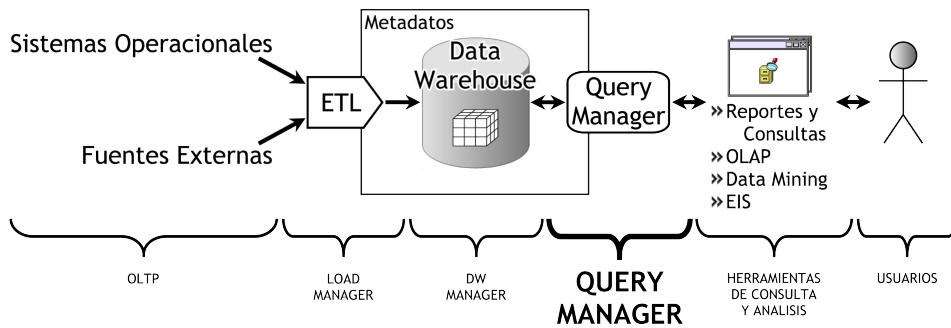


Figura 3.26: Query Manager.

Este componente realiza las operaciones necesarias para soportar los procesos de gestión y ejecución de consultas relacionales, tales como Join y agregaciones, y de consultas propias del análisis de datos, como drill-up y drill-down.

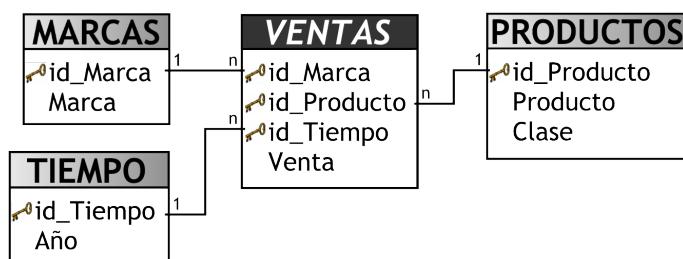
Query Manager recibe las consultas del usuario, las aplica a las tablas correspondientes y devuelve los resultados obtenidos.

Cabe aclarar que una consulta a un DW, generalmente consiste en la obtención de medidas o indicadores a partir de los datos de una tabla de hechos, restringidas por las propiedades o condiciones de las dimensiones seleccionadas.

Las operaciones que se pueden realizar sobre modelos multidimensionales y que son las que verdaderamente les permitirán a los usuarios explorar y desmenuzar los datos en busca de respuestas, son:

- Drill-down.
- Drill-up.
- Drill-across.
- Roll-across.
- Pivot.
- Page.

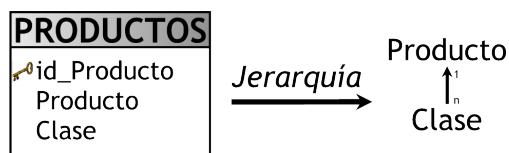
A continuación, se explicará cada una de ellas y se ejemplificará su utilización, para lo cual se utilizará como guía el siguiente esquema en estrella.

**Figura 3.27: Esquema en Estrella.**

El mismo posee tres dimensiones, “MARCAS”, “PRODUCTOS” y “TIEMPO”; y una tabla de hechos central, “VENTAS”, en la cual el hecho “Venta” representa las unidades vendidas de un producto en particular, de una marca específica en un año dado, su fórmula esta dada por:

- Venta = SUM(unidadesVendidas)

Además, se definió una jerarquía sobre la dimensión “PRODUCTOS”:

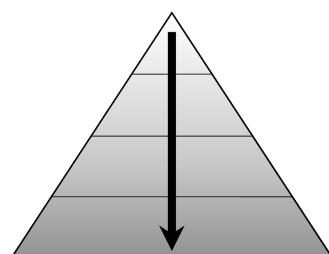
**Figura 3.28: Jerarquía de “PRODUCTOS”.**

Para simplificar los ejemplos que se presentarán, se supondrá que solo se ha cargado en el DW información del año “2007”.

3.5.1. Drill-down

Permite apreciar los datos en un mayor detalle, bajando por la jerarquía de una dimensión. Esto brinda la posibilidad de introducir un nuevo nivel o criterio de agregación en el análisis, disregando los grupos actuales.

Drill-down es ir de lo general a lo específico. Gráficamente:

**Figura 3.29: Drill-down.**

Para explicar esta operación se utilizará la siguiente representación tabular:

TIEMPO - Año	PRODUCTOS - Producto	MARCAS - Marca	VENTAS - Venta
2007	Producto1	M1	40
2007	Producto1	M2	52
2007	Producto1	M3	25
2007	Producto2	M1	39
2007	Producto2	M2	65
2007	Producto2	M3	48

Figura 3.30: Tabla antes de aplicar Drill-down.

Como puede apreciarse, el nombre de la cabecera de la tabla posee el formato “TABLA – Atributo”, y en el cuerpo de la misma se encuentran los valores correspondientes. Se ha resaltado la primera columna, ya que es la que se analizará más en detalle.

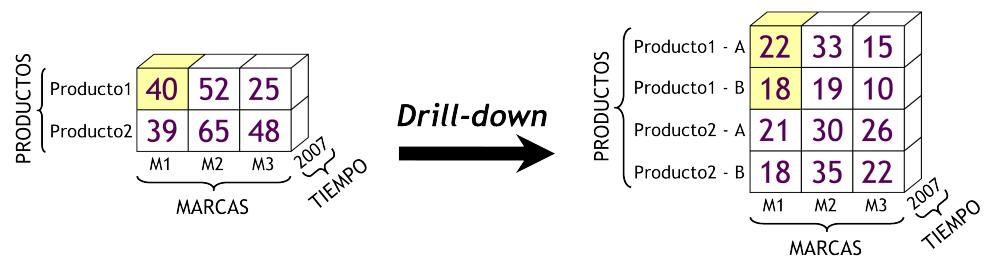
En este caso, se realizará drill-down sobre la jerarquía de la dimensión “PRODUCTOS”, entonces:

TIEMPO - Año	PRODUCTOS - Producto	PRODUCTOS - Clase	MARCAS - Marca	VENTAS - Venta
2007	Producto1	A	M1	22
2007	Producto1	B	M1	18
2007	Producto1	A	M2	33
2007	Producto1	B	M2	19
2007	Producto1	A	M3	15
2007	Producto1	B	M3	10
2007	Producto2	A	M1	21
2007	Producto2	B	M1	18
2007	Producto2	A	M2	30
2007	Producto2	B	M2	35
2007	Producto2	A	M3	26
2007	Producto2	B	M3	22

Figura 3.31: Tabla después de aplicar Drill-down.

Tal y como puede apreciarse en los ítems resaltados de la tabla, se agregó un nuevo nivel de detalle (“PRODUCTOS – Clase”) a la lista inicial, y el valor “40” que pertenecía a las unidades vendidas del “Producto1”, de la marca “M1”, en el año “2007”, se dividió en dos filas. Esto se debe a que ahora se tendrá en cuenta la “Clase” de producto para realizar las summarizaciones sobre “Venta”.

La siguiente imagen muestra este mismo proceso pero, representado matricialmente:

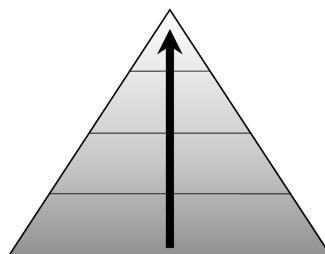
**Figura 3.32: Drill-down, representación matricial.**

De aquí en más se utilizará esta forma para explicar cada operación.

3.5.2. Drill-up

Permite apreciar los datos en menor nivel de detalle, subiendo por la jerarquía de una dimensión. Esto brinda la posibilidad de quitar un nivel o criterio de agregación en el análisis, agregando los grupos actuales.

Drill-up es ir de lo específico a lo general. Gráficamente:

**Figura 3.33: Drill-up.**

Se tomará como referencia la tabla anterior:

TIEMPO - Año	PRODUCTOS - Producto	PRODUCTOS - Clase	MARCAS - Marca	VENTAS - Venta
2007	Producto1	A	M1	22
2007	Producto1	B	M1	18
2007	Producto1	A	M2	33
2007	Producto1	B	M2	19
2007	Producto1	A	M3	15
2007	Producto1	B	M3	10
2007	Producto2	A	M1	21
2007	Producto2	B	M1	18
2007	Producto2	A	M2	30
2007	Producto2	B	M2	35
2007	Producto2	A	M3	26
2007	Producto2	B	M3	22

Figura 3.34: Tabla antes de aplicar Drill-up.

Se aplicará drill-up sobre la jerarquía de la dimensión “PRODUCTOS”, entonces:

TIEMPO - Año	PRODUCTOS - Producto	MARCAS - Marca	VENTAS - Venta
2007	Producto1	M1	40
2007	Producto1	M2	52
2007	Producto1	M3	25
2007	Producto2	M1	39
2007	Producto2	M2	65
2007	Producto2	M3	48

Figura 3.35: Tabla después de aplicar Drill-up.

Como se puede observar en la lista resultante, en la fila resaltada se sumarizaron los valores “22” y “18” de la tabla inicial, debido a que al eliminar el criterio “Clase”, las ventas se agruparon o sumarizaron de acuerdo a “Producto”, “Marca” y “Año”.

La siguiente imagen muestra este mismo proceso pero, representado matricialmente:

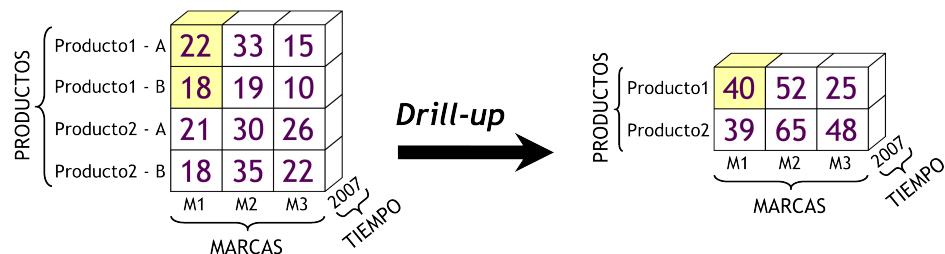


Figura 3.36: Drill-up, representación matricial.

3.5.3. Drill-across

Es muy similar al funcionamiento de drill-down, con la diferencia de que drill-across no se realiza sobre jerarquías de una dimensión, sino que su forma de ir de lo general a lo específico es agregar como nuevo criterio de análisis una nueva dimensión.

Se partirá de la siguiente tabla:

TIEMPO - Año	PRODUCTOS - Producto	VENTAS - Venta
2007	Producto1	117
2007	Producto2	152

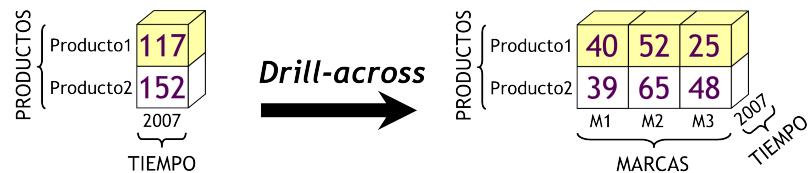
Figura 3.37: Tabla antes de aplicar Drill-across.

Ahora, se aplicará drill-across, al agregar la dimensión “MARCA”, entonces:

TIEMPO - Año	PRODUCTOS - Producto	MARCAS - Marca	VENTAS - Venta
2007	Producto1	M1	40
2007	Producto1	M2	52
2007	Producto1	M3	25
2007	Producto2	M1	39
2007	Producto2	M2	65
2007	Producto2	M3	48

Figura 3.38: Tabla después de aplicar Drill-across.

La siguiente imagen muestra este mismo proceso pero, representado matricialmente:

**Figura 3.39: Drill-across, representación matricial.**

3.5.4. Roll-across

Es muy similar al funcionamiento de drill-up, con la diferencia de que roll-across no se hace sobre jerarquías de una dimensión, sino que su forma de ir de lo específico a lo general es quitar un criterio de análisis eliminando de la consulta una dimensión.

Se tomará como base la representación tabular anterior:

TIEMPO - Año	PRODUCTOS - Producto	MARCAS - Marca	VENTAS - Venta
2007	Producto1	M1	40
2007	Producto1	M2	52
2007	Producto1	M3	25
2007	Producto2	M1	39
2007	Producto2	M2	65
2007	Producto2	M3	48

Figura 3.40: Tabla antes de aplicar Roll-across.

Se aplicará roll-across sobre la dimensión "MARCAS", eliminándola de la tabla, entonces:

TIEMPO - Año	PRODUCTOS - Producto	VENTAS - Venta
2007	Producto1	117
2007	Producto2	152

Figura 3.41: Tabla después de aplicar Roll-across.

La siguiente imagen muestra este mismo proceso pero, representado matricialmente:

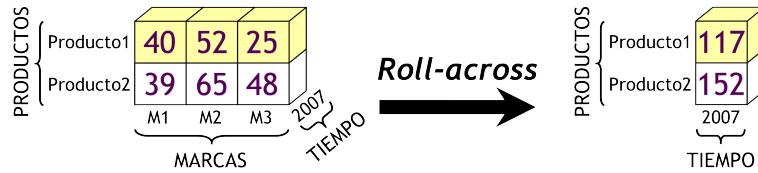


Figura 3.42: Roll-across, representación matricial.

3.5.5. Pivot

Permite seleccionar el orden de visualización de las dimensiones, con el objetivo de analizar la información desde diferentes perspectivas.

Se tomará como referencia, para explicar esta operación, la siguiente tabla:

TIEMPO - Año	PRODUCTOS - Producto	MARCAS - Marca	VENTAS - Venta
2007	Producto1	M1	40
2007	Producto1	M2	52
2007	Producto1	M3	25
2007	Producto2	M1	39
2007	Producto2	M2	65
2007	Producto2	M3	48

Figura 3.43: Tabla antes de aplicar Pivot.

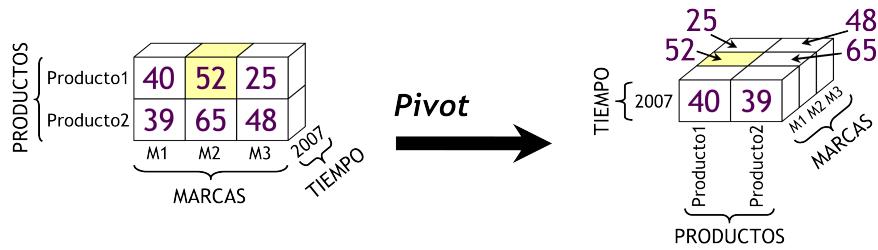
Como puede apreciarse, el orden de las dimensiones es: "Año", "Producto" y "Marca". Ahora, se hará pivot, reorientando la vista multidimensional:

MARCAS - Marca	TIEMPO - Año	PRODUCTOS - Producto	VENTAS - Venta
M1	2007	Producto1	40
M1	2007	Producto2	39
M2	2007	Producto1	52
M2	2007	Producto2	65
M3	2007	Producto1	25
M3	2007	Producto2	48

Figura 3.44: Tabla después de aplicar Pivot.

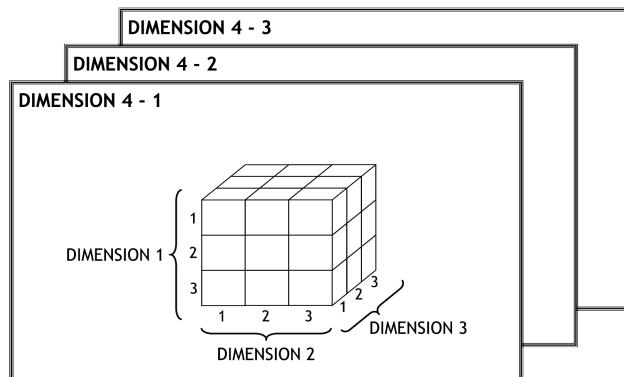
El nuevo orden de las dimensiones es: "Marca", "Año" y "Producto".

La siguiente imagen muestra este mismo proceso pero, representado matricialmente:

**Figura 3.45: Pivot, representación matricial.**

3.5.6. Page

Presenta el cubo dividido en secciones, a través de los valores de una dimensión, como si se tratase de páginas de un libro. Gráficamente:

**Figura 3.46: Page.**

Se tomará como referencia, para explicar esta operación, la siguiente tabla:

TIEMPO - Año	PRODUCTOS - Producto	MARCAS - Marca	VENTAS - Venta
2007	Producto1	M1	40
2007	Producto1	M2	52
2007	Producto1	M3	25
2007	Producto2	M1	39
2007	Producto2	M2	65
2007	Producto2	M3	48

Figura 3.47: Tabla antes de aplicar Page.

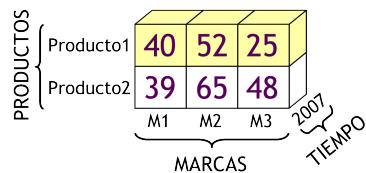
Se realizará Page sobre “PRODUCTOS - Producto”, entonces se obtendrán las siguientes páginas:

- Página Nro 1:

Producto1		
TIEMPO - Año	MARCAS - Marca	VENTAS - Venta
2007	M1	40
2007	M2	52
2007	M3	25

Figura 3.48: Página Nro 1, representación tabular.

Matricialmente:

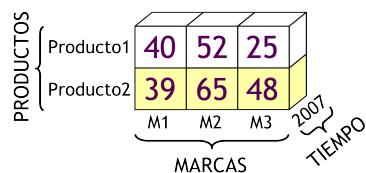
**Figura 3.49: Página Nro 1, representación matricial.**

- Página Nro 2:

Producto2		
TIEMPO - Año	MARCAS - Marca	VENTAS - Venta
2007	M1	39
2007	M2	65
2007	M3	48

Figura 3.50: Página Nro 2, representación tabular.

Matricialmente:

**Figura 3.51: Página Nro 2, representación matricial.**

3.6. Herramientas de Consulta y Análisis

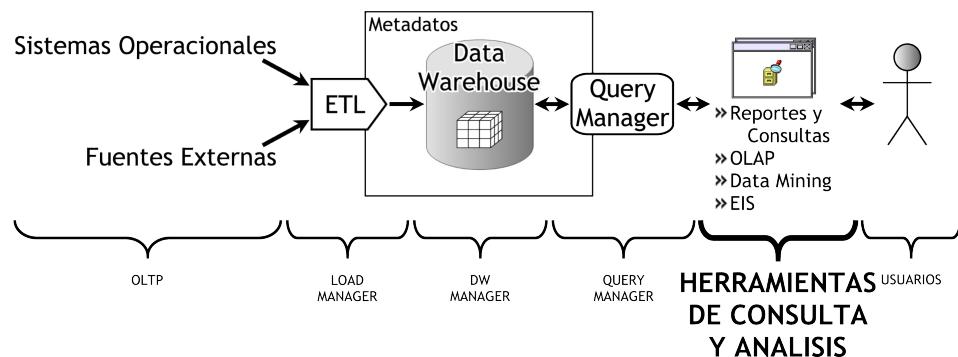


Figura 3.52: Herramientas de Consulta y Análisis.

Las herramientas de consulta y análisis son sistemas que permiten al usuario realizar la exploración de datos del DW. Básicamente constituyen el nexo entre el depósito de datos y los usuarios.

A través de una amigable interfaz gráfica y una serie de simples pasos, el usuario genera consultas que son enviadas desde la herramienta de consulta y análisis al Query Manager, este a su vez realiza la extracción de información al DW Manager y devuelve los resultados obtenidos a la herramienta que se los solicitó. Luego, estos resultados son expuestos ante el usuario en formatos que le son familiares.

Este proceso se puede comprender mejor al observar la siguiente figura:

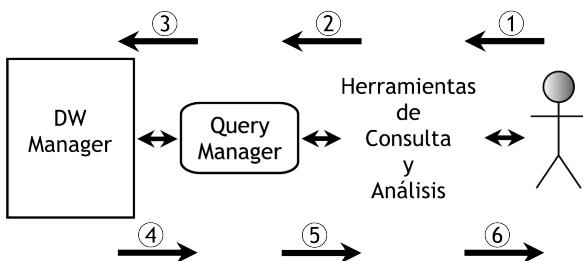


Figura 3.53: Proceso de Consulta y Análisis

El mismo, se lleva a cabo a través de seis pasos sucesivos:

1. El usuario selecciona o establece que datos desea obtener del DW, mediante las interfaces de la herramienta que utilice.
2. La herramienta recibe el pedido del usuario, construye la consulta y la envía al Query Manager.
3. El Query Manager ejecuta la consulta en las tablas del DW.
4. El Query Manager obtiene los resultados de la consulta.

5. El Query Manager envía los datos a las herramientas de consulta y análisis.
6. Las herramientas presentan al usuario la información requerida.

Una de las principales ventajas de utilizar estas herramientas, es que los usuarios no se tienen que preocupar por conocer cual es la estructura de los datos, solo se deben enfocar en el análisis.

Existen diferentes tipos de herramientas de consulta y análisis, y de acuerdo a la necesidad, tipos de usuarios y requerimientos del negocio, se deberán seleccionar las más propicias al caso. Entre ellas se destacan las siguientes:

- Reportes y Consultas
- OLAP
- Data Mining
- EIS

3.6.1. Reportes y Consultas

Se han desarrollado varias herramientas para la producción de consultas y reportes, que ofrecen a los usuarios, a través de pantallas gráficas intuitivas, la posibilidad de generar informes avanzados y detallados del área de interés del negocio que se este analizando. El usuario solo debe seguir una serie de simples pasos, como por ejemplo seleccionar opciones de un menú, presionar tal o cual botón para especificar los elementos de datos, sus condiciones, criterios de agrupación y demás atributos que se consideren significativos.

3.6.2. OLAP

El procesamiento analítico en línea OLAP (On Line Analytic Processing), es la componente más poderosa de los DW, ya que es el motor de consultas especializado de la bodega.

Las herramientas OLAP, son una tecnología de software para análisis en línea, administración y ejecución de consultas, que permiten inferir información del comportamiento del negocio.

Su principal objetivo es el de brindar rápidas respuestas a complejas preguntas, para interpretar la situación del negocio y tomar decisiones. Cabe destacar que lo que es realmente interesante en OLAP, no es la ejecución de simples consultas tradicionales, sino la posibilidad de utilizar operadores tales como drill-up, drill-down, etc, para explotar profundamente la información.

Además, a través de este tipo de herramientas, se puede analizar el negocio desde diferentes escenarios históricos, y proyectar como se ha venido comportando y evolucionando en un ambiente multidimensional, o sea, mediante la combinación de diferentes perspectivas, temas de interés o dimensiones. Esto permite deducir tendencias, por medio del descubrimiento de relaciones entre las perspectivas que a simple vista no se podrían encontrar sencillamente.

Las herramientas OLAP requieren que los datos estén organizados dentro del depósito en forma multidimensional, por lo cual es que utilizan los cubos multidimensionales.

Además de las características ya descritas, se pueden enumerar las siguientes:

- Permite recolectar y organizar la información analítica necesaria para los usuarios y disponer de ella en diversos formatos, tales como tablas, gráficos, reportes, etc.
- Soporta análisis complejos de grandes volúmenes de datos.
- Complementa las actividades de otras herramientas que requieran procesamiento analítico en línea.
- Presenta al usuario una visión multidimensional de los datos (matricial) para cada tema de interés del negocio.
- Es transparente al tipo de tecnología que soporta el DW, ya sea ROLAP, MOLAP o HOLAP.
- Permite definir de forma flexible las dimensiones que se quieren analizar, sus restricciones, jerarquías y combinaciones.
- No tiene limitaciones con respecto al número máximo de dimensiones permitidas.
- Permite a los usuarios, analizar la información basándose en más criterios que un análisis de forma tradicional.
- Al contar con muestras grandes, se pueden explorar mejor los datos en busca de respuestas.
- Permiten realizar agregaciones y combinaciones de los datos de maneras complejas y específicas, con el fin de realizar análisis más estratégicos.

3.6.3. Data Mining

Esta herramienta constituye una poderosa tecnología con un gran potencial que ayuda y brinda soporte a los usuarios, con el fin de permitirles analizar y extraer conocimientos ocultos y predecibles a partir de los datos almacenados en un DW o en un OLTP. Claro que es deseable que la fuente de información sea un DW, por todas las ventajas que aporta.

La integración con el depósito de datos facilita que las decisiones operacionales sean implementadas directamente y monitoreadas. A medida que el DW crece con nuevos requerimientos y resultados, la organización puede minar las mejores prácticas y aplicarlas en futuras decisiones.

Implementar Data Mining permitirá analizar factores de influencia en determinados procesos, predecir o estimar variables o comportamientos futuros, segmentar o agrupar ítems similares, además de obtener secuencias de eventos que provocan comportamientos específicos.

Una de las principales ventajas del Data Mining es que, como recién se ha hecho mención, permite inferir comportamientos, modelos, relaciones y estimaciones de los datos, para poder desarrollar predicciones de los mismos, sin la necesidad de tener que contar con algún patrón o regla preestablecida o conocida de antemano, permitiendo tomar decisiones proactivas y basadas en un conocimiento acabado de la información.

Además brinda la posibilidad de dar respuesta a preguntas complicadas sobre los temas de interés, como por ejemplo ¿Qué está pasando?, ¿Por qué? y ¿Qué pasaría si?, estos cuestionamientos aplicados a una empresa podrían ser: ¿Cuál de los productos de tal marca y clase serán más vendidos en la zona norte en el próximo semestre? y ¿por qué? Además se podrán ver los resultados en forma de reportes tabulares, con gráficos, textos, etc.

Entonces, se puede definir Data Mining como una técnica para descubrir patrones y relaciones entre abundantes cantidades de datos, que a simple vista o que mediante otros tipos de análisis, no se pueden deducir, ya que tradicionalmente consumiría demasiado tiempo o estaría fuera de las expectativas.

Los sistemas Data Mining se desarrollan bajo lenguajes de última generación basados en la Inteligencia Artificial y utilizan métodos matemáticos tales como:

- Redes Neuronales.
- Sistemas Expertos.
- Programación Genética.
- Árboles de Decisión.

Soporta además, sofisticadas operaciones de análisis como los sistemas Scoring y aplicaciones de Detección de Desviación.

Es muy importante tener en cuenta que en las herramientas OLAP y en los reportes y consultas, el análisis parte de una pregunta o hipótesis generada por el usuario, en cambio Data Mining permite generar estas hipótesis.

Generalmente las herramientas de Data Mining se integran con plataformas de hardware y software existentes, como DW, para incrementar el valor de las fuentes de datos establecidas y para que puedan ser integradas con nuevos productos y sistemas en línea, como OLAP. En adición a esto, hacer minería de datos sobre un depósito de datos permite entre otras ventajas contar con los beneficios de los procesos ETL y de las técnicas de limpieza de datos, tan necesarios en este tipo de análisis.

3.6.3.1. Redes Neuronales

Se utilizan para construir modelos predictivos no lineales que aprenden a través de entrenamiento y que semejan la estructura de una red neuronal biológica.

Una red neuronal es un modelo computacional con un conjunto de propiedades específicas, como la habilidad de adaptarse o aprender, generalizar u organizar la información, todo ello basado en un procesamiento eminentemente paralelo.

Por ejemplo, las redes neuronales pueden emplearse para:

- Resolver problemas en dominios complejos con variables continuas y categóricas.
- Modelizar relaciones no lineales.
- Clasificar y predecir resultados.

3.6.3.2. Sistemas Expertos

Un sistema experto, puede definirse como un sistema informático (hardware y software) que simula a los expertos humanos en un área de especialización dada.

La principal ventaja de estos sistemas es que un usuario con poca experiencia puede resolver problemas que requieren el conocimiento de un experto en el tema.

Por ejemplo, los sistemas expertos pueden utilizarse para:

- Realizar transacciones bancarias a través de cajeros automáticos.
- Controlar y regular el flujo de tráfico en las calles y en los ferrocarriles, mediante la operación automática de semáforos.
- Resolver complicados problemas de planificación en los cuales intervienen muchas variables.
- Descubrir relaciones entre diversos conjuntos de variables.

3.6.3.3. Programación Genética

El principal objetivo de la programación genética es lograr que las computadoras aprendan a resolver problemas sin ser explícitamente programadas para ello, generando de esta manera soluciones a partir de la inducción de los programas. El verdadero valor de esta inducción está fundamentado en que todos los problemas se pueden expresar como un programa de computadora.

Por ejemplo, la programación genética se utiliza para:

- Resolver problemas, para los cuales es difícil y no natural tratar de especificar o restringir con anticipación el tamaño y forma de una solución eventual.
- Analizar sistemas que actúan sobre condiciones inestables en ambientes cambiantes.
- Generar de manera automática programas que solucionen problemas planteados.

3.6.3.4. Árboles de Decisión

Son estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos, las cuales explican el comportamiento de una variable con relación a otras, y pueden traducirse fácilmente en reglas de negocio.

Son utilizados con finalidad predictiva y de clasificación.

Por ejemplo, los árboles de decisión pueden emplearse para:

- Optimizar respuestas de campañas.
- Identificar clientes potenciales.
- Realizar evaluación de riesgos.

3.6.3.5. Detección de Desviación

Analiza una serie de datos similares, y cuando encuentra un elemento que no coincide con el resto lo considera una desviación.

Usualmente para la detección de la desviación en base de datos grandes se utiliza la información explícita externa a los datos, así como las limitaciones de integridad o modelos predefinidos. En un método lineal, al contrario, se enfoca el problema desde el interior de los datos, empleando la redundancia implícita de los mismos.

Por ejemplo, la detección de desviación puede utilizarse para:

- Descubrir excepciones a modelos establecidos.
- Delimitar grupos que cumplan con condiciones preestablecidas.

3.6.4. EIS

EIS (Executive Information System) proporciona medios sencillos para consultar, analizar y acceder a la información de estado del negocio. Además, pone a disposición facilidades para que el usuario pueda conseguir los datos buscados rápidamente, empleando el menor tiempo posible para comprender el uso de la herramienta.

Usualmente, EIS se utiliza para analizar las métricas e indicadores de performance y desempeño del negocio o área de interés, a través de la presentación de vistas con datos simplificados, altamente consolidados, mayormente estáticos y preferentemente gráficos.

El concepto principal de esta herramienta, se basa en el simple hecho de que los ejecutivos no poseen tiempo, ni las habilidades necesarias para analizar grandes cantidades de datos.

Al igual que OLAP y Data Mining, los EIS, se pueden aplicar independientemente de la plataforma DW. Pero tener como base un depósito de datos para implementar esta herramienta, conlleva todas las ventajas implícitas del mismo.

3.7. Usuarios

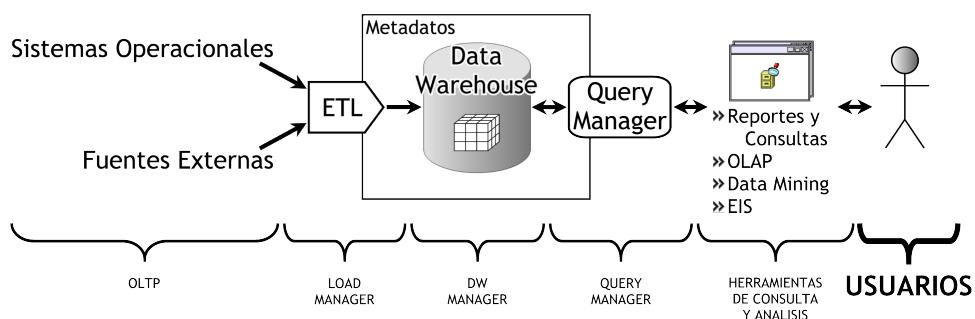


Figura 3.54: Usuarios.

Los usuarios que posee el DW son aquellos que se encargan de tomar decisiones y de planificar las actividades del negocio, es por ello que se hace tanto énfasis en la integración, limpieza de datos, etc, para poder conseguir que la información posea toda la calidad posible.

Es a través de las herramientas de consulta y análisis, que los usuarios exploran los datos en busca de respuestas para poder tomar decisiones proactivas.

Para comprender mejor a los usuarios del almacén de datos, se hará referencia a las diferencias que estos tienen con respecto a los del OLTP:

- Los usuarios que acceden al DW concurrentemente son pocos, en cambio los que acceden a los OLTP en un tiempo determinado son muchos más, pueden ser cientos o incluso miles. Esto se debe principalmente al tipo de información que contiene cada fuente.
- Los usuarios del DW generan por lo general consultas complejas, no predecibles y no anticipadas. Usualmente, cuando se encuentra una respuesta a una consulta se formulan nuevas preguntas más detalladas y así sucesivamente. Es decir, primero se analiza la información a nivel de datos actual para averiguar el “qué”, luego, para obtener mayor detalle y examinar el “cómo”, se trabajan con los datos ligeramente resumidos, derivados de la consulta anterior, y desde allí se puede explorar los datos altamente resumidos. Teniendo en cuenta siempre la posibilidad de utilizar el detalle de datos histórico. Al contrario, los usuarios de los OLTP solo manejan consultas predefinidas.
- Los usuarios del DW, generan consultas sobre una gran cantidad de registros, en cambio los del OLTP lo hacen sobre un pequeño grupo. Esto se debe a que como ya se ha mencionado, el depósito contiene información histórica e integra varias fuentes de datos.
- Las consultas de los usuarios del DW no tienen tiempos de respuesta críticos, aunque sí se espera que se produzcan en el mismo día en que fueron realizadas. Mientras mayor sea el tamaño del depósito y mientras más compleja sea la consulta, mayores serán los tiempos de respuestas. En cambio, las respuestas de las consultas en un OLTP son y deben ser inmediatas.
- En los OLTP, los usuarios típicamente realizan actualizaciones, tales como agregar, modificar, eliminar y consultar algún registro. En cambio en un DW, la única operación que pueden realizar es la de consulta.

Las mencionadas diferencias entre estos dos tipos de usuarios se pueden apreciar mejor en la siguiente tabla comparativa:

	Usuarios de OLTP	Usuarios de Data Warehouse
Acceso concurrente	Muchos.	Pocos.
Tipo de consultas	Predefinidas.	Complejas, no predecibles y no anticipadas.
Registros consultados	Pocos.	Muchos.
Tiempo de respuesta	Crítico.	No crítico.
Acciones permitidas	Agregar, modificar, eliminar y consultar.	Consultar.

Figura 3.55: Usuarios de OLTP vs Usuarios de DW.

Capítulo 4

OTROS CONCEPTOS A TENER EN CUENTA

4.1. Sistema de Misión Crítica

Los usuarios siempre poseen una cierta resistencia al cambio cada vez que se les presenta una nueva herramienta o software, es por ello que al principio no todos confiarán en el DW, y por ende no lo utilizarán. Pero a medida que pasa el tiempo y los usuarios pueden comprobar por sí mismos su buen funcionamiento, se adaptan, aprenden a usarlo y disuelven sus dudas e incertidumbres, tanto el número de usuarios como su utilización se incrementará de manera considerable.

Además, a medida que las empresas confían y emplean más el almacén de datos, y están más pendientes de la disponibilidad de información que él contiene, como así también en su acceso, este se torna fundamental para la misión del negocio o área que apoya, convirtiéndose paulatinamente en un Sistema de Misión Crítica. Llegando al punto en que, un error en el mismo puede provocar una falla en las actividades del negocio.

En resumen, conforme la empresa comienza a utilizar cada vez más los datos del DW, y desde luego se fían de su buen funcionamiento y desempeño para producir de forma sencilla, rápidas consultas, los usuarios comenzarán a dejar para último momento la generación de la información necesaria. Por este motivo, es de suma importancia que el DW posea una buena performance, seguridad y consistencia, y que todas las aplicaciones o herramientas que lo manipulen estén a disposición en todo momento.

4.2. Data Mart

En adición al DW principal pueden existir varios Data Mart (DM) departamentales. Un DM es la implementación de un DW con alcance restringido a un área funcional, problema en particular, departamento, tema o grupo de necesidades.

Muchos depósitos de datos comienzan siendo Data Mart, para, entre otros motivos, minimizar riesgos y producir una primera entrega en tiempos razonables. Pero, una vez que estos se han implementado exitosamente, su ámbito se irá ampliando paulatinamente.

De acuerdo a las operaciones que se deseen o requieran desarrollar, los DM pueden adoptar las siguientes arquitecturas:

- Top-Down: primero se define el DW y luego se desarrollan, construyen y cargan los DM a partir del mismo. En la siguiente figura se encuentra detallada esta arquitectura:

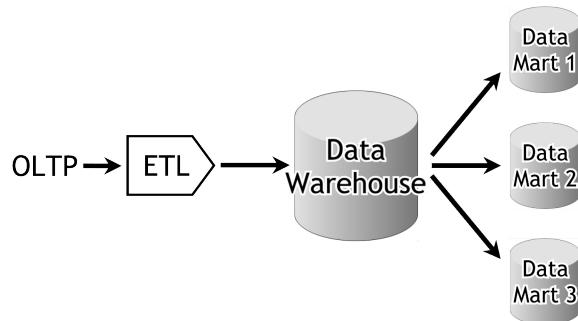


Figura 4.1: Top-Down.

Como se puede apreciar, el DW es cargado a través de procesos ETL y luego este alimenta a los diferentes DM, cada uno de los cuales recibirá los datos que correspondan al tema o departamento que traten. Esta forma de implementación cuenta con la ventaja de no tener que incurrir en complicadas sincronizaciones de hechos, pero requiere una gran inversión y una gran cantidad de tiempo de construcción.

- Bottom-Up: en esta arquitectura, se definen previamente los DM y luego se integran en un DW centralizado. La siguiente figura presenta esta implementación:

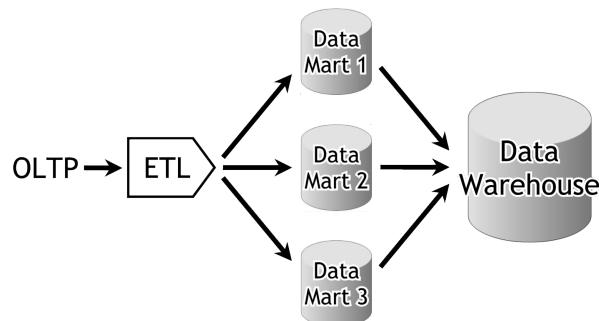


Figura 4.2: Bottom-Up

Los DM se cargan a través de procesos ETL, los cuales suministrarán la información adecuada a cada uno de ellos. En muchas ocasiones, los DM son implementados sin que exista el DW, ya que tienen sus mismas características pero con la particularidad de que están enfocados en un tema específico. Luego de que hayan sido creados y cargados todos los DM, se procederá a su integración con el depósito. La ventaja que trae aparejada este modelo es que cada DM se crea y pone en funcionamiento en un corto lapso de tiempo y se puede tener una pequeña solución a un costo no tan elevado. Luego que todos los DM estén puestos en marcha, se puede decidir si construir el DW o no. El mayor inconveniente está dado en tener

que sincronizar los hechos al momento de la consolidación en el depósito.

Dentro de las ventajas de aplicar un Data Mart a un negocio, se han seleccionado las siguientes:

- Son simples de implementar.
- Conllevan poco tiempo de construcción y puesta en marcha.
- Permiten manejar información confidencial.
- Reflejan rápidamente sus beneficios y cualidades.
- Reducen la demanda del depósito de datos.

4.3. SGBD

Los SGBD (Sistema de Gestión de Base de Datos) son un tipo de software muy específico, dedicados a servir de interfaz entre la base de datos, el usuario y las aplicaciones que lo utilizan. Se compone de lenguajes de definición, manipulación, consulta y seguridad de datos.

El propósito general de los SGBD es el de manejar de manera clara, sencilla y ordenada un conjunto de datos.

Existen diferentes objetivos que deben cumplir los SGBD, de los cuales se han enumerado los siguientes:

- Hacer transparente al usuario los detalles del almacenamiento físico de los datos, mediante varios niveles de abstracción de la información.
- Permitir la realización de cambios a la estructura de la base de datos, sin tener que modificar la aplicación que la emplea.
- Proveer al usuario la seguridad de que sus datos no podrán ser accedidos, ni manipulados por quien no tenga permiso para ello. Debido a esto, debe poseer un complejo sistema que maneje grupos, usuarios y permisos para las diferentes actividades que se pueden realizar dentro del mismo.
- Mantener la integridad de los datos.
- Proporcionar una manera eficiente de realizar copias de seguridad de la información almacenada en ellos, y permitir a partir de estas copias restaurar los datos.
- Controlar el acceso concurrente de los usuarios.
- Facilitar el manejo de grandes volúmenes de información.

Parte II

HEFESTO: Metodología propia para la Construcción de un Data Warehouse

RESUMEN

En esta segunda parte de la publicación, se propondrá una metodología propia para la construcción de un Data Warehouse, que partirá de la recolección de requerimientos y necesidades de información del usuario, y concluirá en la confección de un esquema lógico y sus respectivos procesos de extracción, transformación y carga de datos. Además, se exemplificará cada etapa de la metodología a través de su aplicación a una empresa real, que servirá de guía para que se puedan visualizar los resultados que se esperan de cada paso y para clarificar los conceptos enunciados.

Primero, se describirán los aspectos más sobresalientes de la metodología y luego se explicará cada paso con su respectiva aplicación. Finalmente, se expondrán algunas consideraciones que deben tenerse en cuenta al momento de construir e implementar un Data Warehouse.

El principal objetivo es facilitar el arduo trabajo que significa construir un Data Warehouse desde cero, aportando información que permitirá aumentar la performance del mismo. En adición a ello, esta nueva metodología estará orientada a evitar el tedio que provoca el tener que seguir pasos sin terminar de comprender el por qué de los mismos.

Capítulo 5

METODOLOGÍA HEFESTO

5.1. Introducción

En esta sección se presentará la metodología HEFESTO, que permitirá la construcción de Data Warehouse de forma sencilla, ordenada e intuitiva. Su nombre fue inspirado en el dios griego de la construcción y el fuego, y su logotipo es el siguiente:

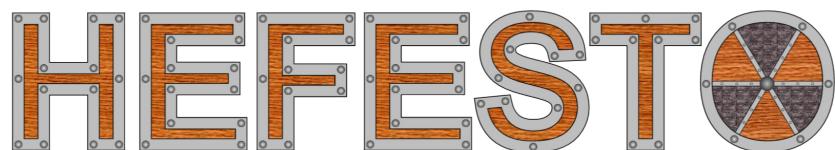


Figura 5.1: Metodología HEFESTO, logotipo.

HEFESTO es una metodología propia, cuya propuesta está fundamentada en una muy amplia investigación, comparación de metodologías existentes y experiencias propias en procesos de confección de almacenes de datos.

La idea principal, es comprender cada paso que se realizará, para no caer en el tedio de tener que seguir un método al pie de la letra sin saber exactamente qué se está haciendo, ni por qué.

La construcción e implementación de un DW puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, con la salvedad de que para algunas fases en particular, las acciones que se han de realizar serán muy diferentes. Lo que se debe tener muy en cuenta, es no entrar en la utilización de metodologías que requieran fases extensas de reunión de requerimientos y análisis, fases de desarrollo monolítico que conlleve demasiado tiempo y fases de despliegue muy largas. Lo que se busca, es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del DW y motivar a los usuarios.

La metodología HEFESTO, puede ser embebida en cualquier ciclo de vida que cumpla con la condición antes declarada.

Con el fin de que se llegue a una total comprensión de cada paso o etapa, se acompañará con la implementación en una empresa real, para demostrar los resultados que se deben obtener y exemplificar cada concepto.

5.2. Descripción

La metodología HEFESTO puede resumirse a través del siguiente gráfico:



Figura 5.2: Metodología HEFESTO, pasos.

Como se puede apreciar, se comienza recolectando las necesidades de información de los usuarios y se obtienen las preguntas claves del negocio. Luego, se deben identificar los indicadores resultantes de los interrogativos y sus respectivas perspectivas de análisis, mediante las cuales se construirá el modelo conceptual de datos del DW.

Después, se analizarán los OLTP para señalar las correspondencias con los datos fuentes y seleccionar los campos de estudio de cada perspectiva.

Una vez hecho esto, se pasará a la construcción del modelo lógico del depósito, explicando las jerarquías que intervendrán.

Por último, se definirán los procesos de carga, transformación, extracción y limpieza de los datos fuente.

5.3. Características

Esta metodología cuenta con las siguientes características:

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- Se basa en los requerimientos del usuario, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- Reduce la resistencia al cambio, ya que involucra al usuario final en cada etapa para que tome decisiones respecto al comportamiento y funciones del DW.
- Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- Es independiente de las herramientas que se utilicen para su implementación.
- Es independiente de las estructuras físicas que contengan el DW y de su respectiva distribución.
- Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente.
- Se aplica tanto para DM como para DW.

5.4. Empresa analizada

Antes de comenzar con el primer paso, es menester describir las características principales de la empresa a la cual se le aplicará la metodología HEFESTO, así se podrá tener como base un ámbito predefinido y se comprenderá mejor cada decisión que se tome con respecto a la implementación y diseño del DW.

Además, este análisis ayudará a conocer el funcionamiento y accionar de la empresa, lo que permitirá examinar e interpretar de forma óptima las necesidades de información de la misma, como así también apoyará a una mejor construcción y adaptación del depósito de datos.

La descripción de la empresa se encuentra en el Apéndice A (página 97).

5.5. Pasos y aplicación metodológica

5.5.1. PASO 1) ANÁLISIS DE REQUERIMIENTOS

a) Identificar preguntas

El primer paso comienza con el acopio de las necesidades de información, el cual puede llevarse a cabo a través de muy variadas y diferentes técnicas, cada una de las cuales poseen características inherentes y específicas, como por ejemplo entrevistas, cuestionarios, observaciones, etc.

El análisis de los requerimientos de los diferentes usuarios, es el punto de partida de esta metodología, ya que ellos son los que deben, en cierto modo, guiar la investigación hacia un desarrollo que refleje claramente lo que se espera del depósito de datos, en relación a sus funciones y cualidades.

El objetivo principal de esta fase, es la de obtener e identificar las necesidades de información clave de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la empresa, y que facilitará una eficaz y eficiente toma de decisiones.

Debe tenerse en cuenta que dicha información, es la que proveerá el soporte para desarrollar los pasos sucesivos, por lo cual, es muy importante que se preste especial atención al relevar los datos.

Una forma de asegurarse de que se ha realizado un buen análisis, es que el resultado del mismo debe hacer explícitos los objetivos estratégicos planteados por la empresa que se está estudiando.

Otra forma de encaminar el relevamiento, es enfocar las necesidades de información en los procesos principales que desarrolle la empresa en cuestión.

La idea central es, que se formulen preguntas complejas sobre el negocio, que incluyan variables de análisis que se consideren relevantes, ya que son estas las que permitirán estudiar la información desde diferentes perspectivas.

Un punto importante que debe tenerse muy en cuenta, es que la información debe estar soportada de alguna manera por algún OLTP, ya que de otra forma, no se podrá elaborar el DW.

Caso práctico:

Se indagó a los usuarios en busca de sus necesidades de información, pero las mismas abarcaban casi todas las actividades de la empresa, por lo cual se les pidió que escogieran el proceso que considerasen más importante en las actividades diarias de la misma y que estuviese soportado de alguna manera por algún OLTP. El proceso elegido fue el de Ventas.

A continuación, se procedió a identificar que era lo que les interesaba conocer acerca de este proceso y cuáles eran las variables o perspectivas que debían tenerse en cuenta para poder tomar decisiones basadas en ello.

Se les preguntó cuáles eran según ellos, los indicadores que representan de mejor modo el proceso de Ventas y qué sería exactamente lo que se desea analizar del mismo. La respuesta obtenida, fue que se deben tener en cuenta y consultar datos sobre la cantidad de unidades vendidas y el

monto total de ventas.

Luego se les preguntó cuáles serían las variables o perspectivas desde las cuales se consultarán dichos indicadores. Para simplificar esta tarea se les presentó una serie de ejemplos concretos de otros casos similares.

El resultado obtenido fue el siguiente:

- Se desea conocer cuántas unidades de cada producto fueron vendidas a sus clientes en un periodo determinado. O en otras palabras: "Unidades vendidas de cada producto a cada cliente en un tiempo determinado".
- Se desea conocer cuál fue el monto total de ventas de productos a cada cliente en un periodo determinado. O en otras palabras: "Monto total de ventas de cada producto a cada cliente en un tiempo determinado".

Debido a que la dimensión Tiempo es un elemento fundamental en el DW, se hizo hincapié en él. Además, se puso mucho énfasis en dejar en claro a los usuarios, a través de ejemplos prácticos, que es este componente el que permitirá tener varias versiones de los datos a fin de realizar un correcto análisis posterior.

Como se puede apreciar, las necesidades de información expuestas están acorde a los objetivos y estrategias de la empresa, ya que es precisamente esta información requerida la que proveerá un ámbito para la toma de decisiones, que en este caso permitirá analizar el comportamiento de los clientes a los que se pretende satisfacer ampliamente, para así lograr obtener una ventaja competitiva y maximizar las ganancias.

5.5.1.2. b) Identificar indicadores y perspectivas de análisis

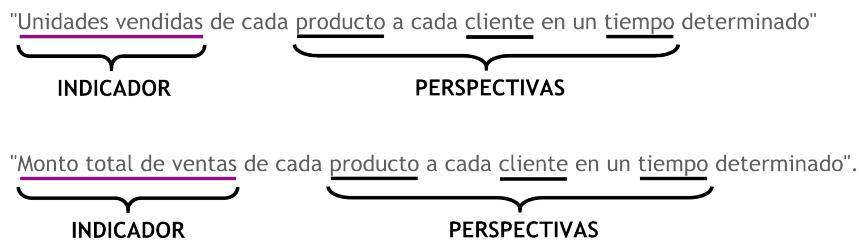
Una vez que se han establecido las preguntas claves, se debe proceder a su descomposición para descubrir los indicadores que se utilizarán y las perspectivas de análisis que intervendrán.

Para ello, se debe tener en cuenta que los indicadores, para que sean realmente efectivos son, en general, valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, etc.

En cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, etc. Cabe destacar, que el Tiempo es muy comúnmente una perspectiva.

Caso práctico:

A continuación, se analizarán las preguntas obtenidas en el paso anterior y se detallarán cuáles son sus respectivos indicadores y perspectivas.

**Figura 5.3: Caso práctico, indicadores y perspectivas.**

En síntesis, los indicadores son:

- Unidades vendidas.
- Monto total de ventas.

Y las perspectivas de análisis son:

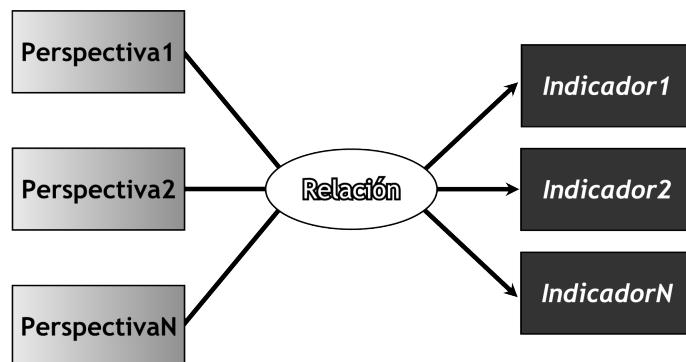
- Clientes.
- Productos.
- Tiempo.

5.5.1.3. c) Modelo Conceptual

En esta etapa, se construirá un modelo conceptual¹ a partir de los indicadores y perspectivas obtenidas en el paso anterior.

A través de este modelo, se podrá observar con claridad cuales son los alcances del proyecto, para luego poder trabajar sobre ellos, además al poseer un alto nivel de definición de los datos, permite que pueda ser presentado ante los usuarios y explicado con facilidad.

La representación gráfica del modelo conceptual es la siguiente:

**Figura 5.4: Modelo Conceptual.**

¹Modelo Conceptual: descripción de alto nivel de la estructura de la base de datos, en la cual la información es representada a través de objetos, relaciones y atributos.

A la izquierda se colocan las perspectivas seleccionadas, que serán unidas a un óvalo central que representa y lleva el nombre de la relación que existe entre ellas. La relación, constituye el proceso o área de estudio elegida. De dicha relación y entrelazadas con flechas, se desprenden los indicadores o medidas, estos se ubican a la derecha del esquema.

Como puede apreciarse en la figura anterior, el modelo conceptual permite de un solo vistazo y sin poseer demasiados conocimientos previos, comprender cuáles serán los resultados que se obtendrán, cuáles serán las variables que se utilizarán para analizarlos y cuál es la relación que existe entre ellos.

Caso práctico:

El modelo conceptual resultante de los datos que se han recolectado, es el siguiente:

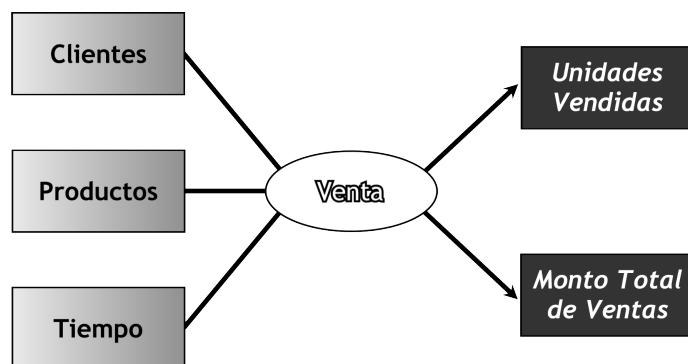


Figura 5.5: Caso práctico, Modelo Conceptual.

Como puede observarse, la relación mediante la cual se unen las diferentes perspectivas, para obtener como resultado los indicadores requeridos por los usuarios, es precisamente "Venta".

5.5.2. PASO 2) ANÁLISIS DE LOS OLTP

5.5.2.1. a) Establecer correspondencias con los requerimientos

El objetivo de este análisis, es el de examinar los OLTP disponibles que contengan la información requerida, como así también sus características, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos.

En el caso de los indicadores, deben explicitarse como se calcularán, y más aún si son fórmulas u operaciones complejas.

La idea es, que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

Caso práctico:

En el OLTP de la empresa analizada, el proceso de venta está representado por el diagrama de entidad relación² de la siguiente figura.

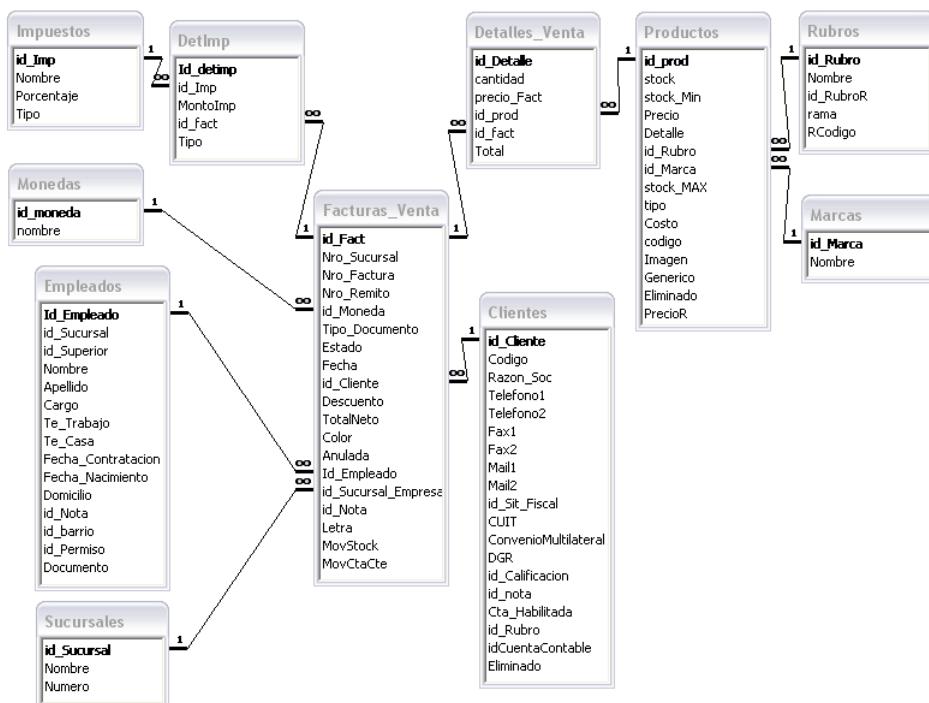


Figura 5.6: Caso práctico, Diagrama de Entidad Relación.

Los indicadores se calcularán de la siguiente manera:

- “Unidades Vendidas” representa las unidades que se han vendido de un producto en particular.
- “Monto Total de Ventas” representa en monto total que se ha vendido de cada producto, y se obtiene al multiplicar la cantidad de unidades vendidas, por su respectivo precio.

A continuación, se expondrá la correspondencia entre los dos modelos:

²Diagrama de Entidad Relación: representa la información a través de entidades, relaciones, cardinalidades, claves, atributos y jerarquías de generalización.

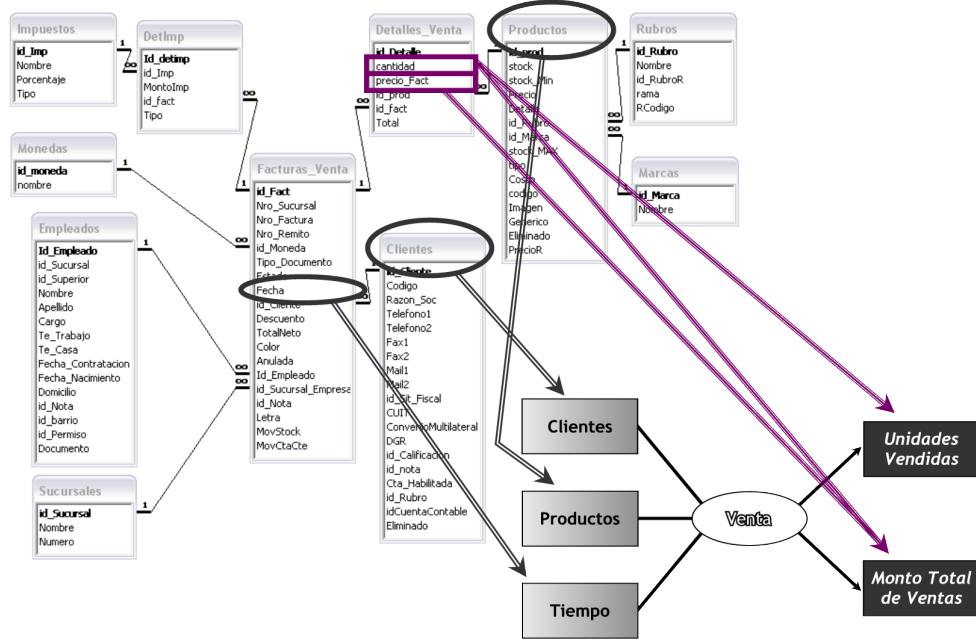


Figura 5.7: Caso práctico, correspondencia.

Las relaciones identificadas fueron las siguientes:

- La tabla "Productos" se relaciona con la perspectiva "Productos".
- La tabla "Clientes" con la perspectiva "Clientes".
- El campo "fecha" de la tabla "Facturas_Venta" con la perspectiva "Tiempo" (debido a que es la fecha principal en el proceso de venta).
- El campo "cantidad" de la tabla "Detalles_Venta" con el indicador "Unidades Vendidas".
- El campo "cantidad" de la tabla "Detalles_Venta" multiplicado por el campo "precio_Fact" de la misma tabla, con el indicador "Monto Total de Ventas".

5.5.2.2. b) Seleccionar los campos que integrarán cada perspectiva. Nivel de granularidad

Una vez que se han establecido las relaciones con los OLTP, se examinarán y seleccionarán los campos que contendrá cada perspectiva, ya que será a través de estos por los que se manipularán y filtrarán los indicadores.

Para ello, basándose en las correspondencias establecidas en el paso anterior, se debe presentar al usuario los datos de análisis disponibles para cada perspectiva. Es muy importante conocer en detalle qué significa cada campo y/o valor de los datos encontrados en los OLTP, por lo cual, es conveniente investigar su sentido, ya sea a través de diccionarios de datos, reuniones con los encargados del sistema, análisis de los datos propiamente dichos, etc.

Luego de exponer frente al usuario los datos existentes, explicando su significado, valores posibles y características, este debe decidir cuáles son los que considera relevantes para consultar los indicadores y cuáles no.

Con respecto a la perspectiva “Tiempo”, es muy importante definir el ámbito mediante el cual se agruparán o sumarizarán los datos. Este punto es fundamental precisarlo con claridad, debido a que, determinará la granularidad de la información encontrada en el DW. Sus campos posibles pueden ser: día de la semana, quincena, mes, trimestres, semestre, año, etc.

Finalmente, y con el fin de graficar los resultados obtenidos, se ampliará el modelo conceptual expuesto anteriormente, colocando bajo cada perspectiva los campos o atributos elegidos.

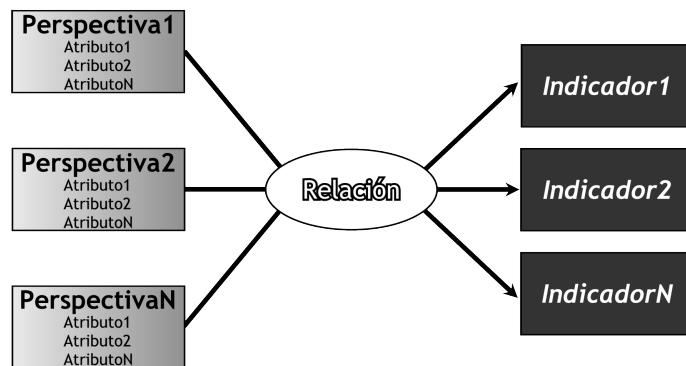


Figura 5.8: Modelo Conceptual con atributos.

Caso práctico:

De acuerdo a las correspondencias establecidas, se analizaron los campos residentes en cada tabla a la que se hacia referencia, a través de dos métodos diferentes. Primero se examinó la base de datos para intuir los significados de cada campo, y luego se consultó con el encargado del sistema sobre algunos aspectos de los cuales no se comprendía su sentido.

De todas formas, y como puede apreciarse en el diagrama de entidad relación antes expuesto, los nombres de los campos son bastante explícitos y se deducen con facilidad, pero aún así fue necesario investigarlos para evitar cualquier tipo de inconvenientes.

- Con respecto a la perspectiva ”Clientes”, los datos disponibles son los siguientes:
 - id_Cliente: es la clave primaria de la tabla ”Clientes”, y representa únicamente a un cliente en particular.
 - Codigo: representa el código del cliente, este campo es calculado de acuerdo a una combinación de las iniciales del nombre del cliente, el grupo al que pertenece y un número incremental.
 - Razon_Soc: nombre o razón social del cliente.
 - Telefono1: número de teléfono del cliente.
 - Telefono2: segundo número telefónico del cliente.
 - Fax1: número de fax del cliente.
 - Fax2: segundo número de fax del cliente.
 - Mail1: dirección de correo electrónico del cliente.
 - Mail2: segunda dirección de correo del cliente.
 - id_Sit_Fiscal: representa a través de una clave foránea el tipo de situación fiscal que posee el cliente. Por ejemplo: Consumidor Final, Exento, Responsable No Inscripto, Responsable Inscripto.

- CUIT: número de C.U.I.T. del cliente.
 - ConvenioMultilateral: indica si el cliente posee o no convenio multilateral.
 - DGR: número de D.G.R. del cliente.
 - id_Clasificación: representa a través de una clave foránea la clasificación del cliente. Por ejemplo: Muy Bueno, Bueno, Regular, Malo, Muy Malo.
 - id_nota: representa a través de una clave foránea una observación realizada acerca del cliente.
 - Cta_Habilitada: indica si el cliente posee su cuenta habilitada.
 - id_Rubro: representa a través de una clave foránea el grupo al que pertenece el cliente. Por ejemplo: Bancos, Construcción, Educación Privada, Educación Pública, Particulares.
 - idCuentaContable: representa la cuenta contable asociada al cliente, la cual se utilizará para imputar los movimientos contables que este genere.
 - Eliminado: indica si el cliente fue eliminado o no. Si fue eliminado, no figura en las listas de clientes actuales.
- En la perspectiva "Productos", los datos que se pueden utilizar son los siguientes:
- id_prod: es la clave primaria de la tabla "Productos", y representa unívocamente a un producto en particular.
 - stock: stock actual del producto.
 - stock_min: stock mínimo del producto, se utiliza para dar alerta si el stock actual está cerca del mismo, al ras o si ya lo superó.
 - Precio: precio de venta del producto.
 - Detalle: nombre o descripción del producto.
 - id_Rubro: representa a través de una clave foránea el rubro al que pertenece el producto.
 - id_Marca: representa a través de una clave foránea la marca a la que pertenece el producto.
 - stock_MAX: stock máximo del producto. Al igual que "stock_min", se utiliza para dar alertas del nivel de stock actual.
 - tipo: clasificación del producto. Por ejemplo: Producto, Servicio, Compuesto.
 - Costo: precio de costo del producto.
 - codigo: representa el código del producto, este campo es calculado de acuerdo a una combinación de las iniciales del nombre del producto, el rubro al que pertenece y un número incremental.
 - Imagen: ruta de acceso a una imagen o dibujo mediante la cual se quiera representar al producto. Este campo no es utilizado actualmente.
 - Generico: indica si el producto es genérico o no.
 - Eliminado: indica si el producto fue eliminado o no. Si fue eliminado, no figura en las listas de productos actuales.
 - PrecioR: precio de lista del producto.
- Con respecto a la perspectiva "Tiempo", que es la que determinará la granularidad del depósito de datos, los atributos más típicos que pueden emplearse son los siguientes:
- Año.
 - Semestre.
 - Cuatrimestre.
 - Trimestre.
 - Número de mes.
 - Nombre del mes.
 - Quincena.
 - Semana.
 - Número de día.

- Nombre del día.

Una vez que se recolectó toda la información pertinente y se consultó con los usuarios cuales eran los datos que consideraban de interés para analizar los indicadores ya expuestos, los resultados obtenidos fueron los siguientes:

- En la perspectiva "Clientes" sólo se tendrá en cuenta el nombre del cliente, o sea, el campo "Razon_Soc" de la tabla "Clientes".
- En la perspectiva "Productos", se utilizarán los campos que hacen referencia al nombre del producto ("detalle" de la tabla "Productos") y a la marca a la que pertenecen ("Nombre" de la tabla "Marcas", obtenido a través de la unión con la tabla "Productos").
- En la perspectiva "Tiempo", se seleccionaron los campos "Mes" (referido al nombre del mes), "Trimestre" y "Año".

Teniendo esto en cuenta, se completará el diseño del diagrama conceptual:

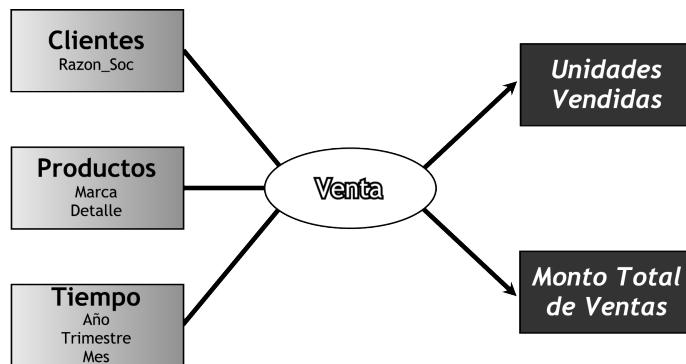


Figura 5.9: Caso práctico, Modelo Conceptual con atributos.

5.5.3. PASO 3) ELABORACIÓN DEL MODELO LÓGICO DE LA ESTRUCTURA DEL DW

A continuación, se confeccionará el modelo lógico³ de la estructura del DW, teniendo como base el modelo conceptual que ya ha sido creado.

Se debe seleccionar cuál será el tipo de esquema que se utilizará para contener la estructura del depósito de datos, que se adapte mejor a los requerimientos y necesidades del usuario. Es muy importante definir objetivamente si se empleará un esquema en estrella, constelación o copo de nieve, ya que esta decisión afectará considerablemente la elaboración del modelo lógico.

Caso práctico:

El esquema que se utilizará será en estrella, debido a sus características, ventajas y diferencias con los otros esquemas.

³Modelo Lógico: representación de una estructura de datos, que puede procesarse y almacenarse en algún SGBD.

5.5.3.1. a) Diseñar tablas de dimensiones

Este paso, se aplicará por igual a todos los tipos de esquemas lógicos.

Lo primero que se hará será crear las dimensiones del mismo, para ello se tomará cada perspectiva con sus atributos relacionados y se les realizará el siguiente proceso:

- Se elegirá un nombre que identifique la dimensión.
- Se añadirá un campo que represente su clave principal.
- Se redefinirán los nombres de los atributos si es que no son lo bastante explicativos.

Gráficamente:

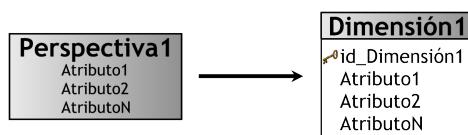


Figura 5.10: Diseño de tablas de dimensiones.

Caso práctico:

A continuación, se diseñaran las tablas de dimensiones.

- Perspectiva “Clientes”:

- La nueva dimensión tendrá el nombre “CLIENTE”.
- Se le agregará una clave principal con el nombre “idCliente”.
- Se modificará el nombre del atributo “Razon_Soc” por “Cliente”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

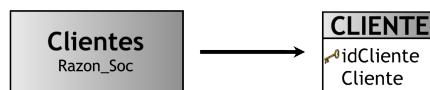
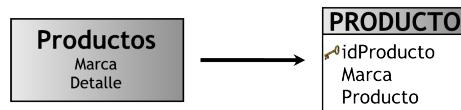


Figura 5.11: Caso práctico, dimensión “CLIENTE”.

- Perspectiva “Productos”:

- La nueva dimensión tendrá el nombre “PRODUCTO”.
- Se le agregará una clave principal con el nombre “idProducto”.
- El nombre del atributo “Marca” no será cambiado.
- Se modificará el nombre del atributo “Detalle” por “Producto”.

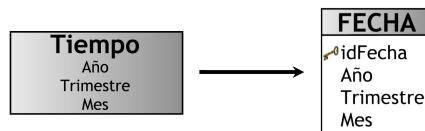
Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

**Figura 5.12: Caso práctico, dimensión "PRODUCTO".**

- Perspectiva “Tiempo”:

- La nueva dimensión tendrá el nombre “FECHA”.
- Se le agregará una clave principal con el nombre “idFecha”.
- El nombre los atributos no serán modificados.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

**Figura 5.13: Caso práctico, dimensión "FECHA".**

5.5.3.2. b) Diseñar tablas de hechos

En este paso, se definirán las tablas de hechos, que son las que contendrán los indicadores de estudio.

- Para los esquemas en estrella y copo de nieve, se realizará lo siguiente:

- Al igual que las dimensiones, se le deberá asignar un nombre a la tabla de hechos que en este caso represente la información analizada, área de investigación, negocio enfocado, etc.
- Se definirá su clave primaria, que se compone de la combinación de las claves primarias de cada dimensión que se utilizará para generar las consultas.
- Se renombrarán los hechos o indicadores si es que no llegasen a ser lo suficientemente explícitos.

Gráficamente:

**Figura 5.14: Tabla de hechos.**

- Para los esquemas constelación se realizará lo siguiente:

- Las tablas de hechos se deben confeccionar teniendo en cuenta el análisis de las preguntas realizadas por el usuario en pasos anteriores y sus respectivos indicadores y dimensiones.
- Cada tabla de hechos debe poseer un nombre que la identifique, contener sus indicadores correspondientes y su clave debe estar formada por la combinación de las claves de las dimensiones que intervendrán.

Antes de continuar, vale la pena recordar que las perspectivas fueron convertidas en dimensiones en el paso anterior, razón por la cual, las preguntas realizadas por el usuario son examinadas a través de indicadores y dimensiones.

- Al diseñar las tablas de hechos, se deberá tener en cuenta:
 - Caso 1: Si en dos o más preguntas figuran los mismos indicadores pero con diferentes dimensiones de análisis, existirán tantas tablas de hechos como preguntas cumplan esta condición. Por ejemplo:

"Analizar el Indicador1 por Dimensión1 y por Dimensión2".
 "Analizar el Indicador1 por Dimensión2 y por Dimensión3".

Figura 5.15: Caso 1, preguntas.

Entonces se obtendrá:

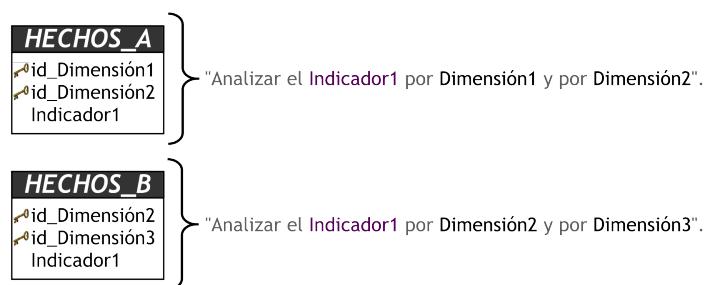


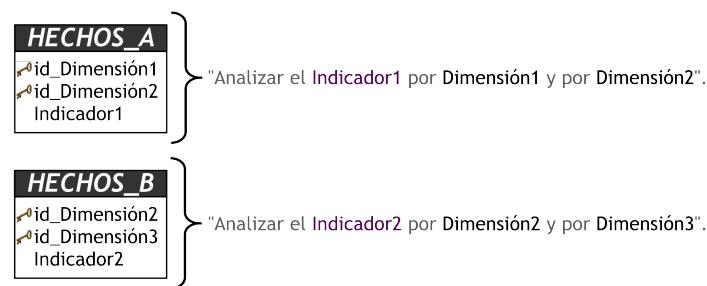
Figura 5.16: Caso 1, diseño de tablas de hechos.

- Caso 2: Si en dos o más preguntas figuran diferentes indicadores con diferentes dimensiones de análisis, existirán tantas tablas de hechos como preguntas cumplan esta condición. Por ejemplo:

"Analizar el Indicador1 por Dimensión1 y por Dimensión2".
 "Analizar el Indicador2 por Dimensión2 y por Dimensión3".

Figura 5.17: Caso 2, preguntas.

Entonces se obtendrá:

**Figura 5.18: Caso 2, diseño de tablas de hechos.**

- Caso 3: Si el conjunto de preguntas cumplen con las condiciones de los dos puntos anteriores se deberán unificar aquellos interrogantes que posean diferentes indicadores pero iguales dimensiones de análisis, para luego reanudar el estudio de las preguntas. Por ejemplo:

"Analizar el Indicador1 por Dimensión1 y por Dimensión2".
 "Analizar el Indicador2 por Dimensión1 y por Dimensión2".

Figura 5.19: Caso 3, preguntas.

Se unificarán en:

"Analizar el Indicador1 y el Indicador2 por Dimensión1 y por Dimensión2".

Figura 5.20: Caso 3, unificación.

Caso práctico:

A continuación, se confeccionará la tabla de hechos:

- La tabla de hechos tendrá el nombre “VENTAS”.
- Su clave principal será la combinación de las claves principales de las dimensiones antes definidas: “idCliente”, “idProducto” e “idFecha”.
- Los indicadores serán renombrados, “Unidades Vendidas” por “Cantidad” y “Monto Total de Ventas” por “MontoTotal”.

En el gráfico siguiente se puede apreciar mejor este paso:

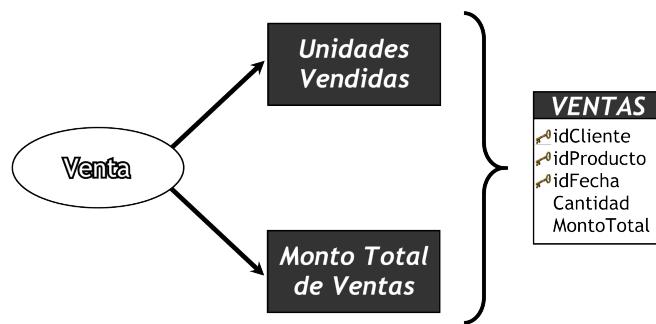


Figura 5.21: Caso práctico, diseño de la tabla de hechos.

5.5.3.3. c) Realizar uniones

Para los tres tipos de esquemas, se realizarán las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos.

Caso práctico:

Se realizarán las uniones pertinentes, de acuerdo corresponda:

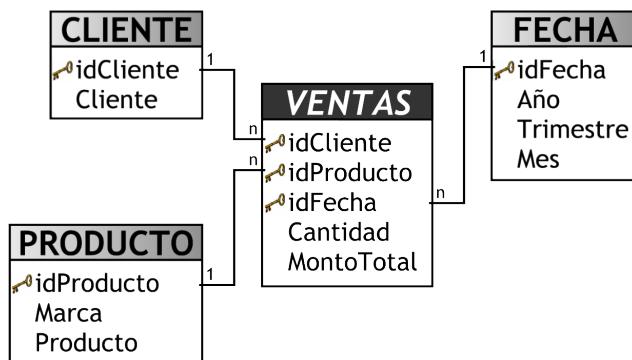
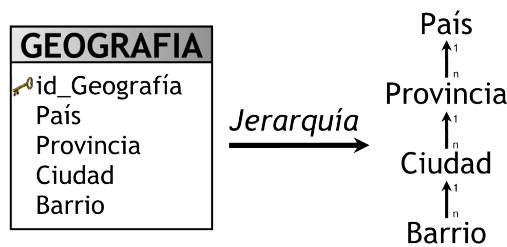


Figura 5.22: Caso práctico, uniones.

5.5.3.4. d) Determinar jerarquías

- Para los esquemas en estrella y constelación, se deberán especificar las jerarquías que existirán dentro de cada tabla de dimensión, teniendo siempre presente cual es el objetivo de las mismas. Para representar las jerarquías en el modelo lógico, se deberán colocar los atributos pertenecientes a las jerarquías en sus respectivas tablas, en orden descendente y acompañado con un número ordinal encerrado entre corchetes.

Por ejemplo, si se posee la siguiente jerarquía:

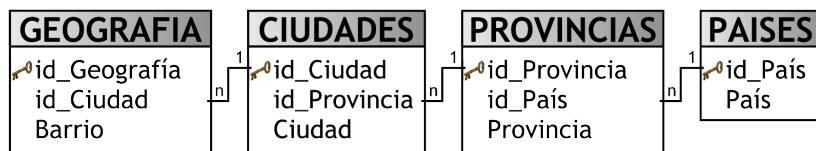
**Figura 5.23: Jerarquía de "GEOGRAFIA".**

Entonces, para representar esta jerarquía, la tabla “GEOGRAFIA” debe quedar como sigue:

GEOGRAFIA	
id_Geografía	
País	[1]
Provincia	[2]
Ciudad	[3]
Barrio	[4]

Figura 5.24: Jerarquía de "GEOGRAFIA", representación.

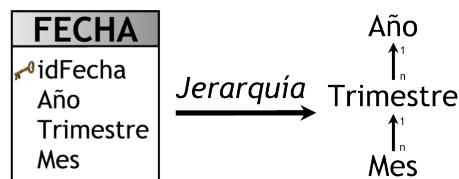
- Para los esquemas copo de nieve, cuando existan jerarquías dentro de una dimensión, esta tabla deberá ser normalizada. Por ejemplo, si se toma como referencia la dimensión “GEOGRAFIA” de la imagen anterior y su respectiva jerarquía, entonces, al normalizar esta tabla se obtendrá:

**Figura 5.25: Normalización de la jerarquía de "GEOGRAFIA".**

Caso práctico:

A continuación, se definirán las jerarquías existentes en el esquema, las cuales fueron acordadas con los usuarios, tras haberles explicado sus objetivos y utilidades.

En la dimensión “FECHA” se seleccionó la siguiente:

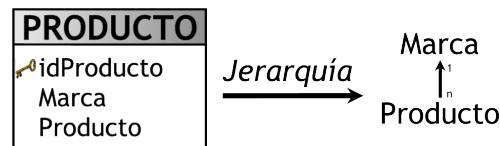
**Figura 5.26: Caso práctico, jerarquía de "FECHA"**

Entonces, la tabla “FECHA” queda como sigue:

FECHA	
idFecha	
Año	[1]
Trimestre	[2]
Mes	[3]

Figura 5.27: Caso práctico, representación de la jerarquía de "FECHA"

En la dimensión “PRODUCTO” se seleccionó la siguiente jerarquía:

**Figura 5.28: Caso práctico, jerarquía de "PRODUCTO"**

Entonces, la tabla “PRODUCTO” queda como sigue:

PRODUCTO	
idProducto	
Marca	[1]
Producto	[2]

Figura 5.29: Caso práctico, representación de la jerarquía de "FECHA"

En la figura que se presentará a continuación, se puede apreciar el esquema lógico del DW resultante, tras haber definido sus jerarquías correspondientes.

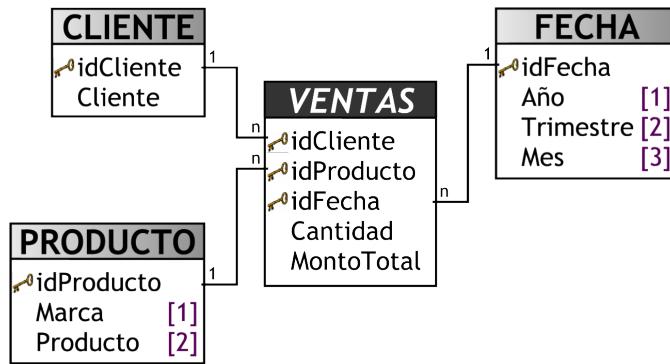


Figura 5.30: Caso práctico, Modelo Lógico.

5.5.4. PASO 4) PROCESOS ETL, LIMPIEZA DE DATOS Y SENTENCIAS SQL

Una vez construido el modelo lógico, se deberá proceder a probarlo con datos, a través de procesos ETL.

Para realizar la compleja actividad de extraer datos de diferentes fuentes, para luego integrarlos, filtrarlos y depurarlos, existen varios software que facilitan estas tareas, por lo cual este paso se centrará solo en la generación de las sentencias SQL que contendrán los datos que serán de interés.

Antes de realizar la carga de datos, es conveniente efectuar una limpieza de los mismos, para evitar valores faltantes y anómalos.

Al generar los ETL, se debe tener en cuenta cual es la información que se desea almacenar en el depósito de datos, para ello se pueden establecer condiciones adicionales y restricciones. Estas condiciones deben ser analizadas y llevadas a cabo con mucha prudencia para evitar pérdidas de datos importantes.

En la cláusula ORDER BY de las sentencias SQL, que se efectuarán para cargar cada tabla, deben figurar los atributos, medidas y claves en orden de aparición de sus respectivas tablas. Al realizar esta acción, se logrará aportar mayor eficiencia cuando se realizan búsquedas de datos.

Cuando se trabaja con un esquema constelación, hay que tener presente que varias dimensiones serán compartidas con diferentes tablas de hechos, ya que puede darse el caso de que algunas restricciones aplicadas sobre una tabla de dimensión en particular para analizar los indicadores de una tabla de hechos, se puedan contraponer con otras restricciones o condiciones de análisis de otros indicadores de otras tablas de hechos.

Primero se cargarán los datos de las dimensiones y luego los de las tablas de hechos, teniendo en cuenta siempre, la correcta correspondencia entre cada elemento y las summarizaciones que se requieran. En el caso en que se esté utilizando un esquema copo de nieve, cada vez que existan jerarquías de dimensiones, se comenzarán cargando las tablas de dimensiones del nivel más general al más detallado.

Cuando se haya cargado en su totalidad el DW, se deben establecer sus políticas de actualización o refresco de datos.

Caso práctico:

A continuación, se generarán las sentencias SQL para cargar las diferentes dimensiones y la tabla de hechos.

■ Dimensión “CLIENTE”:

Se tomará como fuente de entrada la tabla “Clientes” del OLTP mencionado anteriormente.

Se consultó con los usuarios y se averiguó que deseaban tener en cuenta solo aquellos clientes que no estén eliminados y que tengan su cuenta habilitada.

Es importante destacar que aunque existían numerosos movimientos de clientes que en la actualidad no poseen su cuenta habilitada o que figuran como eliminados, se decidió no incluirlos debido a que el énfasis está puesto en analizar los indicadores a través de aquellos clientes que no cuentan con estas condiciones.

Los clientes eliminados son referenciados mediante el campo “Eliminado”, en el cual un valor “1” indica que este fue eliminado, y un valor “0” que aún permanece vigente. Cuando se examinaron los registros de la tabla, para muchos clientes no había ningún valor asignado para este campo, lo cual, según comunicó el encargado del sistema, se debía a que este se agregó poco después de haberse creado la base de datos inicial, razón por la cual existían valores faltantes. Además, comentó que en el sistema, si un cliente posee en el campo “Eliminado” un valor “0” o un valor faltante, es considerado como vigente.

Con respecto a la cuenta habilitada, el campo del OLTP que le hace mención es “Cta_Habilitada”, y un valor “0” indica que no está habilitada y un valor “1” que sí.

Seguidamente, se generará la sentencia SQL, sobre el OLTP “Clientes”, con los datos requeridos para cargar esta dimensión:

```
SELECT
    Clientes.id_Cliente AS idCliente,
    Clientes.Razon_Soc AS Cliente
FROM
    Clientes
WHERE
    (Clientes.Eliminado <> 1)
AND (Clientes.Cta_Habilitada <> 0)
ORDER BY
    Clientes.id_Cliente,
    Clientes.Razon_Soc
```

Figura 5.31: Caso práctico, sentencia SQL de “CLIENTES”.

■ Dimensión “PRODUCTO”:

Las fuentes que se utilizarán, son las tablas “Productos” y “Marcas”.

En este caso, aunque existían productos eliminados, el usuario decidió que está condición no fuese tomada en cuenta, ya que habían movimientos que hacían referencia a productos con este estado.

Es necesario realizar una unión entre la tabla “Productos” y “Marcas”, por lo cual se debió asegurar que ningún producto hiciera mención a alguna marca que no existiese, y se tomaron medidas contra su futura aparición.

El SQL es el siguiente:

```
SELECT
    Productos.id_prod AS idProducto,
    Marcas.Nombre AS Marca,
    Productos.Detalle AS Producto
FROM
    Productos LEFT OUTER JOIN
    Marcas ON Productos.id_Marca = Marcas.id_Marca
ORDER BY
    Productos.id_prod,
    Marcas.Nombre,
    Productos.Detalle
```

Figura 5.32: Caso práctico, sentencia SQL de “PRODUCTOS”.

■ Dimensión “FECHA”:

Para generar esta dimensión, infaltable en todo DW, existen varias herramientas y utilidades de software que proporcionan diversas opciones para su confección. Pero, si no se cuenta con ninguna, se puede realizar a mano o mediante algún programa, llenando los datos en un archivo, tabla, hoja de cálculo, etc, y luego exportándolos a donde se requiera.

Lo que se hizo, fue realizar un pequeño programa para cargar en un archivo plano las fechas desde el año del primer movimiento de la empresa, en este caso el año 2000, hasta la fecha actual.

A continuación, se puede apreciar una muestra de los datos generados:

```
"idFecha";"Año";"Trimestre";"Mes"
20000101;2000;"1er Tri";"Enero"
20000102;2000;"1er Tri";"Enero"
20000103;2000;"1er Tri";"Enero"
20000104;2000;"1er Tri";"Enero"
20000105;2000;"1er Tri";"Enero"
20000106;2000;"1er Tri";"Enero"
20000107;2000;"1er Tri";"Enero"
20000108;2000;"1er Tri";"Enero"
20000109;2000;"1er Tri";"Enero"
20000110;2000;"1er Tri";"Enero"
```

Figura 5.33: Caso práctico, datos de "FECHA".

Como se puede observar, la primera fila representa los nombres de las columnas, las cuales están separadas entre sí, y para establecer delimitadores, por “;”, y sus nombres figuran entre comillas dobles. De la segunda fila en adelante se encuentran todos los datos de la dimensión. Los campos que son del tipo texto están encerrados entre comillas dobles, los que son numéricos no.

La clave principal es un campo numérico representado por el formato “yyyymmdd”. La misma también puede calcularse mediante la siguiente fórmula:

$$((\text{YEAR}(\text{fecha}) * 10000) + (\text{MONTH}(\text{fecha}) * 100) + (\text{DAY}(\text{fecha})))$$

Figura 5.34: Caso práctico, fórmula "yyyymmdd".

- Tabla de hechos “VENTAS”:

Para la confección de la tabla de hechos, se tuvieron que tomar como fuente las tablas “Facturas_Ventas” y “Detalles_Venta”. Al igual que en las tablas de dimensiones, se recolectaron las condiciones que deben cumplir los datos para considerarse de interés, y en este caso, se trabajará solamente con aquellas facturas que no hayan sido anuladas.

Se investigó al respecto, y se llegó a la conclusión de que el campo que da dicha información en “Anulada” de la tabla “Facturas_Ventas” y si el mismo posee el valor “1” significa que efectivamente fue anulada.

Otro punto importante a tener en cuenta es que la fecha se debe convertir al formato numérico “yyyymmdd”.

Aquí también se realizarán las sumarizaciones correspondientes de los hechos, para ello se utilizará la cláusula GROUP BY para agrupar todos los registros a través de las claves primarias de esta tabla.

La sentencia SQL resultante fue la siguiente:

```

SELECT
    Facturas_Venta.id_Cliente AS idCliente,
    Detalles_Venta.id_prod AS idProducto,
    ((YEAR(Facturas_Venta.Fecha) * 10000) + (MONTH(Facturas_Venta.Fecha) * 100) +
     (DAY(Facturas_Venta.Fecha))) AS idFecha,
    SUM(Detalles_Venta.cantidad) AS Cantidad,
    SUM(Detalles_Venta.cantidad * Detalles_Venta.precio_Fact) AS MontoTotal
FROM
    Facturas_Venta INNER JOIN
    Detalles_Venta ON Facturas_Venta.id_Fact = Detalles_Venta.id_fact
WHERE
    (Facturas_Venta.Anulada <> 1)
GROUP BY
    Facturas_Venta.id_Cliente,
    Detalles_Venta.id_prod,
    Facturas_Venta.Fecha
ORDER BY
    Facturas_Venta.id_Cliente,
    Detalles_Venta.id_prod,
    idFecha,
    Cantidad,
    MontoTotal

```

Figura 5.35: Caso práctico, sentencia SQL de "VENTAS".

Con respecto a las actualizaciones del depósito de datos, también existen diversas herramientas DW, que proveen muchas facilidades, por lo cual no se entrará en detalle en su utilización, pero sí se establecerá por escrito las políticas que se han convenido con los usuarios:

- La información se refrescará cada semana.
- Los datos de las dimensiones “PRODUCTO” y “CLIENTE” serán cargados totalmente cada vez.
- Los datos de la dimensión “FECHA” se cargarán de manera incremental teniendo en cuenta la fecha de la última actualización.
- Los datos de la tabla de hechos que corresponden al último año a partir de la fecha actual, serán reemplazados cada vez.
- Estas acciones se realizarán durante un periodo de prueba, para analizar cuál es la manera más eficiente de generar las actualizaciones, basadas en el estudio de los cambios que se producen en los OLTP y que afectan al contenido del DW.

Capítulo 6

OTRAS CONSIDERACIONES A TENER EN CUENTA

6.1. Tamaño del DW

Dependiendo del negocio, el volumen de datos y el alcance del proyecto, el tamaño del DW puede variar considerablemente, por lo cual, es una buena práctica tener esto en cuenta al momento de diseñar el depósito y al determinar los recursos físicos, los tiempos de desarrollo y los respectivos costos inherentes.

De acuerdo al tamaño del depósito de datos, se lo puede clasificar como:

- Personal: si su tamaño es menor a 1 Gigabyte.

DW < 1 GB

- Pequeño: si su tamaño es mayor a 1 Gigabyte y menor a 50 Gigabyte.

1 GB < DW < 50 GB

- Mediano: si su tamaño es mayor a 50 Gigabyte y menor a 100 Gigabyte.

50 GB < DW < 100 GB

- Grande: si su tamaño es mayor a 100 Gigabyte y menor a 1 Terabyte.

100 GB < DW < 1 TB

- Muy grande: si su tamaño es mayor a 1 Terabyte.

DW > 1 TB

6.2. Tiempo de construcción

Diversos autores resaltan la importancia del factor tiempo en la construcción de un DW, por lo cual se ha considerado interesante exponer dos frases seleccionadas al respecto:

- “El 70 % del tiempo total dedicado al proyecto se insume en definir el problema y en preparar la tabla de datos”.
- “Estime el tiempo necesario, multiplíquelo por dos y agregue una semana de resguardo”.
- “Regla 90 – 90”: el primer 90 % de la construcción de un sistema absorbe el 90 % del tiempo y esfuerzo asignados; el último 10 % se lleva el otro 90 % del tiempo y esfuerzo asignado.

6.3. Implementación

Las implementaciones de los depósitos de datos varían entre sí de forma considerable, teniendo en cuenta las herramientas de software que se empleen, los modelos que se utilicen, recursos disponibles, SGBD que lo soporten, herramientas de análisis y consulta, entre otros.

6.4. Performance

Cuando se diseñan los ETL, es muy importante que los mismos sean lo más eficientes posible, ya que una vez que se tenga un gran volumen de datos, el espacio en disco se volverá fundamental y los tiempos incurridos en el procesamiento y acceso a la información serán esenciales, y más aún si el DW es considerado o tomado como un sistema de misión crítica.

De aquí surge una bifurcación, debido a que para mejorar la performance del depósito pueden realizarse dos acciones diferentes:

- Darle mayor interés a velocidad de acceso y procesamiento de los datos.
- Optimizar el espacio en disco, reduciendo su volumen.

Estas opciones por lo general son opuestas y no pueden satisfacerse en conjunto, ya que usualmente de acuerdo a la que se le de mayor prioridad afectará el rendimiento de manera inversa a la de menor.

Dependiendo de lo que se considere más importante priorizar se deben tomar medidas acorde a la situación, por ejemplo si se desea tener performance en lo referido a espacio en disco, se podrían llevar a cabo algunas de las siguientes prácticas:

- Crear procedimientos que compriman los datos que no se utilicen frecuentemente y que los descompriman en el momento en que se requieran.
- Prestar especial atención a los tipos de datos utilizados, por ejemplo, para valores enteros pequeños conviene utilizar tinyint o smallint en lugar de int, con el fin de no asignar tamaños de datos mayores a los necesarios. Esto toma vital importancia cuando se aplica en las claves primarias, debido a que formarán parte de la tabla de hechos que es la que contiene el volumen del almacén de datos.

- Definir siempre que sea posible claves numéricas, ya que ocupan mucho menos espacio que las de cadena de caracteres.
- Realizar particiones al DW, para poder dividir su volumen entre varias unidades de almacenamiento.

En cambio, si se tiene preferencia por la velocidad de acceso y procesamiento, se pueden seguir algunas de las prácticas que se expondrán a continuación:

- En los procesos ETL, se puede contar con un manejador de errores que registre en un archivo el detalle de las fallas que ocurran. Este archivo permitirá a quien lo acceda identificar cual fue específicamente el error que ha sucedido, evitando de esta manera demoras innecesarias en detectar y probar posibles causas.
- Realizar procesos de actualización incrementales en aquellas dimensiones que no sean demasiado variables en el tiempo, con el fin de no utilizar Full Load cada vez que se requiera refrescar el depósito de datos. Por ejemplo, esto se puede aplicar en dimensiones que contengan nombres de rubros, productos, marcas, etc, ya que en general, la única modificación que sufren es la de inserción de nuevos registros.
- Realizar particiones al cubo multidimensional, para solo trabajar con los datos seleccionados. Esto puede llevarse a cabo al particionar, por ejemplo, por el año actual y el anterior, de esta manera solamente se trabajará con estos dos años, y se evitarán tiempos innecesarios en cargar datos que no se utilizarán.
- Definir siempre que sea posible claves numéricas, ya que permiten realizar uniones mucho más veloces y eficientes.
- Utilizar técnicas de multiprocesamiento, con el objetivo de agilizar la obtención de resultados, a través de la realización de procesos en forma concurrente.

6.5. Mantenimiento

Un punto muy importante es mantener en correcto funcionamiento al DW, ya que a medida que pase el tiempo, este tenderá a crecer significativamente, y surgirán cambios, tanto en los requerimientos como en las fuentes de datos.

6.6. Impactos

Al construir un DW, es fundamental que los usuarios del mismo participen activamente durante todo su desarrollo, debido a que son ellos los que conocen en profundidad su negocio y saben cuáles son los resultados que se desean obtener. Además, es precisamente en base a la utilización que se le de, que el depósito de datos madurará y se adaptará a las situaciones cambiantes por las que atraviese la empresa. Los usuarios, al trabajar junto a los desarrolladores y analistas podrán comprender más en profundidad sus propios sistemas operacionales, con todo lo que esto implica.

Con la implementación del DW, los procesos de toma de decisiones serán optimizados, al obtener información correcta al instante en que se necesita, evitando perdidas de tiempo y anomalías en los datos. Al contar con esta información, los usuarios tendrán más confianza en las decisiones que tomarán y en adición a ello, poseerán una base sustentable para justificarlas.

Usualmente, los DW integrarán fuentes de datos de diversas áreas y sectores de la empresa, esto tendrá como beneficio contar con una sola fuente de información centralizada y común para todos los usuarios. Esto posibilitará que en las diferentes áreas se compartan los mismos datos, lo cual conducirá a un mayor entendimiento, comunicación, confianza y cooperación entre las mismas.

El depósito de datos introducirá nuevos conceptos tecnológicos y de inteligencia de negocios, lo cual requerirá que se aprendan nuevas técnicas, herramientas, métodos, destrezas, formas de trabajar, etc.

6.7. DM como sub proyectos

Al diseñar e implementar DM como partes de un proyecto DW, se debe tener en cuenta que el análisis que se efectuará, los modelos que intervendrán y el alcance, deben ser globales, con el fin de determinar, por ejemplo, dimensiones comunes entre las diferentes áreas de trabajo. Esto evitará que se realicen tareas repetidas, ahorrando tiempos y enfocándose en la consolidación, unificación y centralización de la información de los diferentes sectores.

6.8. Teoría de grafos

Para evaluar la validez de la estructura lógica del depósito de datos, puede emplearse la teoría de grafos, la cual afirma que su estructura será correcta sí y solo si está conformada únicamente por trayectorias acíclicas.

Si se encuentran trayectorias cíclicas, deberán ser transformadas para que las consultas al DW sean válidas y confiables.

Una trayectoria acíclica, es aquella que sólo tiene una forma de recorrido (en un solo sentido). Por ejemplo, en la siguiente figura se puede apreciar que existe una sola manera de recorrer las dimensiones.

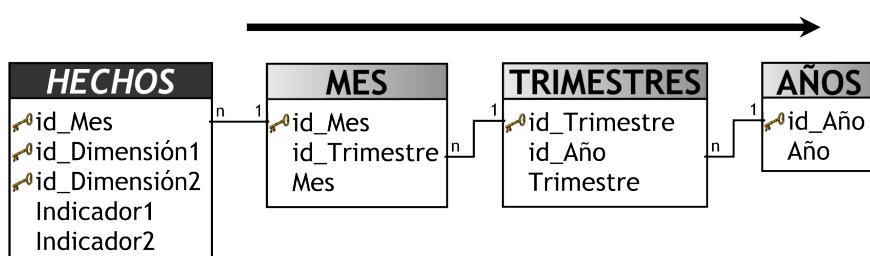
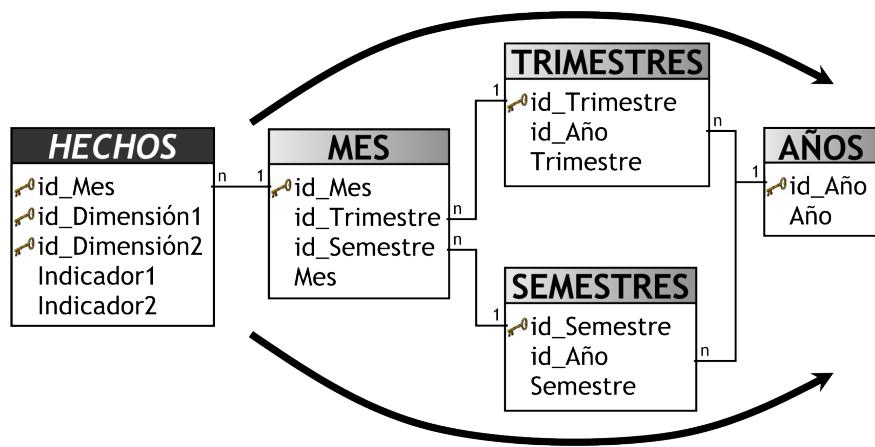


Figura 6.1: Trayectoria acíclica.

Una trayectoria cíclica, es aquella que se puede recorrer en dos o más secuencias diferentes. Por ejemplo, en la siguiente imagen se pueden distinguir dos sentidos por los cuales recorrer las dimensiones.

**Figura 6.2: Trayectoria cíclica.**

6.9. Elección de columnas

Cuando se seleccionan los atributos que integrarán el DW, se debe tener en cuenta lo siguiente:

- Se deben descartar aquellos campos cuyos valores tengan muy poca variabilidad.
- Se deben descartar los campos que tengan valores diferentes para cada objeto, por ejemplo el número de D.N.I. cuando se analizan personas.
- En los casos en que no existan jerarquías dentro de alguna dimensión, en la cual la cantidad de registros que posee la misma son demasiados, es conveniente, conjuntamente con el usuario, definirlas. Pero, si llegase a suceder que no se encontrase ningún criterio por el cual jerarquizar los atributos, es una buena práctica crear jerarquías propias. El objetivo de llevar a cabo esta acción, es la de poder dividir los registros en grupos, propiciando de esta manera una exploración más amena y controlable. Para ejemplificar este punto, se utilizará como referencia la tabla de dimensión de la siguiente figura. La misma no posee ninguna jerarquía definida y la cantidad de registros con que cuenta son cientos:

**Figura 6.3: Tabla de dimensión "PRODUCTO".**

Entonces, lo que se realizará será crear una nueva jerarquía a partir de los campos disponibles:

- Se añadirá a la tabla un nuevo campo ("Letra"), el mismo estará formado por la primera letra del atributo "Producto" que lo acompaña. Por ejemplo, si el valor de "Producto" es "Lapicera", "Letra" será "L"; si es "Cartuchera" será "C", etc.

- Se definirá la nueva jerarquía:

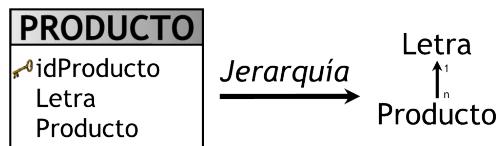


Figura 6.4: Jerarquía de "PRODUCTO".

Además, se pueden aplicar algunas de las acciones que se expondrán a continuación sobre los valores de los campos que se incluirán en el depósito de datos:

- Factorizar: se utiliza para descomponer un valor en dos o más componentes. Por ejemplo, el atributo “código” perteneciente a un producto está formado por tres identificadores separados por guiones medios, que representan su rubro, marca y tipo (“idRubro-idMarca-idTipo”), entonces este campo puede factorizarse y separarse en tres valores independientes (“idRubro”, “idMarca” e “idTipo”).
- Estandarizar: se utiliza para ajustar valores a un tipo de formato o norma preestablecida. Por ejemplo, se puede emplear este método cuando se desea que todos los campos del tipo texto sean convertidos a mayúscula.
- Codificar: es utilizado para representar valores a través de las reglas de un código preestablecido. Por ejemplo, en el campo “género” se pueden codificar sus valores, “M” y “F”, para transformarlos en “Masculino” y “Femenino” respectivamente.
- Discretizar: es empleado para convertir un conjunto continuo de valores en uno discreto. Por ejemplo, cuando se especificaron los tamaños del DW se realizó esta operación.

Apéndice A

Descripción de la empresa

A.1. Identificación de la empresa

La empresa analizada, desarrolla las actividades comerciales de mayorista y minorista de artículos de limpieza, en un ambiente geográfico de alcance nacional. De acuerdo a su volumen de operaciones, se la puede considerar de tamaño mediano.

Con respecto a su clasificación, es una sociedad de responsabilidad limitada con fines de lucro.

Su estructura está formalizada y posee características de una organización funcional.

A.2. Objetivos

Su objetivo principal es el de maximizar sus ganancias. Pero también, se puede adicionar el objetivo de expandirse a un nuevo nivel de mercado, con el fin de conseguir una mayor cantidad de clientes y posicionarse competitivamente por sobre sus rivales.

Otra meta que persigue, pero que aún no está definida como tal, es la de incursionar en otros rubros para lograr diversificarse.

A.3. Políticas

La empresa posee escasos grandes clientes con un gran poder adquisitivo, y son precisamente estos, los que adquieren el volumen de los productos que se comercializan. Debido a ello, la política que se utiliza para cubrir los objetivos antes mencionados, es la de satisfacer ampliamente las necesidades de sus clientes, brindándoles confianza y promoviendo un ambiente familiar entre los mismos. Esta acción se realiza con el fin de mantener los clientes actuales y para que nuevos se interesen en su forma de operar.

Existe otra política que es implícita, por lo cual, no está definida tan estrictamente como la anterior, y es la de mejorar continuamente, con el objetivo de sosegar las exigencias y cambios en el mercado en el que actúa y para conseguir una mejor posición respecto a sus competidores.

A.4. Estrategias

Dentro de las estrategias existentes, se han destacado dos por considerarse más significativas, ellas son:

- Expandir el ámbito geográfico, creando varias sucursales en puntos estratégicos del país.
 - Añadir nuevos rubros a su actividad de comercialización.

A.5. Organigrama

A continuación, se expondrá un organigrama que fue confeccionado a partir de los datos suministrados en la empresa, ya que no existía ninguno previamente predefinido.

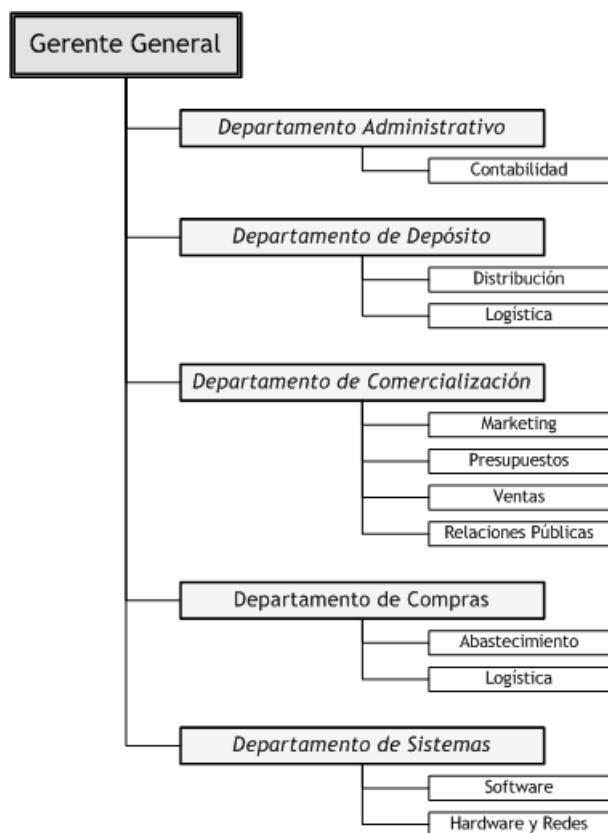


Figura A.1: Organigrama.

A.6. Datos del entorno específico

Los clientes con que cuenta son bastante variados y cubren un amplio margen. Los mismos son tanto provinciales, como nacionales, con diferentes tipos de poder adquisitivo.

Con respecto a sus proveedores, la empresa posee en algunos rubros diversas opciones de las cuales puede elegir y comparar, pero en otros solo cuenta con pocas alternativas.

Además, tiene como rivales a nivel de mayoreo, varios competidores importantes y ya consolidados en el mercado, pero, a nivel minorista aventaja por su tamaño y volumen de actividades a sus principales competidores.

A.7. Relación de las metas de la organización con las del DW

El DW coincide con las metas de la empresa, ya que esta necesita mejorar su eficiencia en la toma de decisiones y contar con información detallada a tal fin. Esto es vital, ya que es muy importante para procurar una mayor ventaja competitiva conocer cuáles son los factores que inciden directamente sobre su rentabilidad, como así también, analizar su relación con otros factores y sus respectivos por qué.

El DW aportará un gran valor a la empresa, entre las principales ventajas e inconvenientes que solucionará se pueden mencionar los siguientes:

- Permitirá a los usuarios tener una visión general del negocio.
- Transformará datos operativos en información analítica, enfocada a la toma de decisiones.
- Se podrán generar reportes dinámicos, ya que actualmente son estáticos y no ofrecen ninguna facilidad de análisis.
- Soportará la estrategia de la empresa.
- Aportará a la mejora continua de la estructura de la empresa.

A.8. Procesos

Los principales procesos que se llevan a cabo son los siguientes:

- Venta:
 - Minorista: es la que se le realiza a los clientes particulares que se acercan hasta la empresa para adquirir los productos que requieren.
 - Mayorista: es la que se le efectúa a los grandes clientes, ya sea por medio de comunicaciones telefónicas, o a través de visitas o reuniones.
 - Al realizarse una venta, el departamento de Depósito se encarga de controlar el stock, realizar encargos de mercadería en caso de no cubrir lo solicitado, armar el pedido y enviarlo por medio de transporte propio o de terceros al destino correspondiente.
- Compra:
 - El departamento de Compras, al recibir del departamento de Depósito las necesidades de mercadería, realiza una comparación de los productos ofrecidos por sus diferentes proveedores en cuestión de precio, calidad y confianza. Posteriormente, se efectúa el pedido correspondiente.

Apéndice B

Licencia de Documentación Libre de GNU

Versión 1.2, Noviembre 2002

This is an unofficial translation of the GNU Free Documentation License into Spanish. It was not published by the Free Software Foundation, and does not legally state the distribution terms for documentation that uses the GNU FDL – only the original English text of the GNU FDL does that. However, we hope that this translation will help Spanish speakers understand the GNU FDL better.

Ésta es una traducción no oficial de la GNU Free Document License a Español (Castellano). No ha sido publicada por la Free Software Foundation y no establece legalmente los términos de distribución para trabajos que usen la GFDL (sólo el texto de la versión original en Inglés de la GFDL lo hace). Sin embargo, esperamos que esta traducción ayude los hispanohablantes a entender mejor la GFDL. La versión original de la GFDL esta disponible en la Free Software Foundation.

Esta traducción está basada en una de la versión 1.1 de Igor Támara y Pablo Reyes. Sin embargo la responsabilidad de su interpretación es de Joaquín Seoane.

Copyright ©2000, 2001, 2002 Free Software Foundation, Inc. 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA. Se permite la copia y distribución de copias literales de este documento de licencia, pero no se permiten cambios¹.

B.1. Preámbulo

El propósito de esta Licencia es permitir que un manual, libro de texto, u otro documento escrito sea *libre* en el sentido de libertad: asegurar a todo el mundo la libertad efectiva de copiarlo y redistribuirlo, con o sin modificaciones, de manera comercial o no. En segundo término, esta Licencia proporciona al autor y al editor² una manera de

¹Ésta es la traducción del Copyright de la Licencia, no es el Copyright de esta traducción no autorizada.

²La licencia original dice *publisher*, que es, estrictamente, quien publica, diferente de *editor*, que es más bien quien prepara un texto para publicar. En castellano *editor* se usa para ambas cosas.

obtener reconocimiento por su trabajo, sin que se le considere responsable de las modificaciones realizadas por otros.

Esta Licencia es de tipo *copyleft*, lo que significa que los trabajos derivados del documento deben a su vez ser libres en el mismo sentido. Complementa la Licencia Pública General de GNU, que es una licencia tipo copyleft diseñada para el software libre.

Hemos diseñado esta Licencia para usarla en manuales de software libre, ya que el software libre necesita documentación libre: un programa libre debe venir con manuales que ofrezcan la mismas libertades que el software. Pero esta licencia no se limita a manuales de software; puede usarse para cualquier texto, sin tener en cuenta su temática o si se publica como libro impreso o no. Recomendamos esta licencia principalmente para trabajos cuyo fin sea instructivo o de referencia.

B.2. Aplicabilidad y definiciones

Esta Licencia se aplica a cualquier manual u otro trabajo, en cualquier soporte, que contenga una nota del propietario de los derechos de autor que indique que puede ser distribuido bajo los términos de esta Licencia. Tal nota garantiza en cualquier lugar del mundo, sin pago de derechos y sin límite de tiempo, el uso de dicho trabajo según las condiciones aquí estipuladas. En adelante la palabra *Documento* se referirá a cualquiera de dichos manuales o trabajos. Cualquier persona es un licenciatario y será referido como *Usted*. Usted acepta la licencia si copia, modifica o distribuye el trabajo de cualquier modo que requiera permiso según la ley de propiedad intelectual.

Una *Versión Modificada* del Documento significa cualquier trabajo que contenga el Documento o una porción del mismo, ya sea una copia literal o con modificaciones y/o traducciones a otro idioma.

Una *Sección Secundaria* es un apéndice con título o una sección preliminar del Documento que trata exclusivamente de la relación entre los autores o editores y el tema general del Documento (o temas relacionados) pero que no contiene nada que entre directamente en dicho tema general (por ejemplo, si el Documento es en parte un texto de matemáticas, una Sección Secundaria puede no explicar nada de matemáticas). La relación puede ser una conexión histórica con el tema o temas relacionados, o una opinión legal, comercial, filosófica, ética o política acerca de ellos.

Las *Secciones Invariantes* son ciertas Secciones Secundarias cuyos títulos son designados como Secciones Invariantes en la nota que indica que el documento es liberado bajo esta Licencia. Si una sección no entra en la definición de Secundaria, no puede designarse como Invariante. El documento puede no tener Secciones Invariantes. Si el Documento no identifica las Secciones Invariantes, es que no las tiene.

Los *Textos de Cubierta* son ciertos pasajes cortos de texto que se listan como Textos de Cubierta Delantera o Textos de Cubierta Trasera en la nota que indica que el documento es liberado bajo esta Licencia. Un Texto de Cubierta Delantera puede tener como mucho 5 palabras, y uno de Cubierta Trasera puede tener hasta 25 palabras.

Una copia *Transparente* del Documento, significa una copia para lectura en máquina, representada en un formato cuya especificación está disponible al público en general, apto para que los contenidos puedan ser vistos y editados directamente con editores de texto genéricos o (para imágenes compuestas por puntos) con programas genéricos

de manipulación de imágenes o (para dibujos) con algún editor de dibujos ampliamente disponible, y que sea adecuado como entrada para formateadores de texto o para su traducción automática a formatos adecuados para formateadores de texto. Una copia hecha en un formato definido como Transparente, pero cuyo marcaje o ausencia de él haya sido diseñado para impedir o dificultar modificaciones posteriores por parte de los lectores no es Transparente. Un formato de imagen no es Transparente si se usa para una cantidad de texto sustancial. Una copia que no es *Transparente* se denomina *Opaca*.

Como ejemplos de formatos adecuados para copias Transparentes están ASCII puro sin marcaje, formato de entrada de Texinfo, formato de entrada de L^AT_EX, SGML o XML usando una DTD disponible públicamente, y HTML, PostScript o PDF simples, que sigan los estándares y diseñados para que los modifiquen personas. Ejemplos de formatos de imagen transparentes son PNG, XCF y JPG. Los formatos Opacos incluyen formatos propietarios que pueden ser leídos y editados únicamente en procesadores de palabras propietarios, SGML o XML para los cuáles las DTD y/o herramientas de procesamiento no estén ampliamente disponibles, y HTML, PostScript o PDF generados por algunos procesadores de palabras sólo como salida.

La *Portada* significa, en un libro impreso, la página de título, más las páginas siguientes que sean necesarias para mantener legiblemente el material que esta Licencia requiere en la portada. Para trabajos en formatos que no tienen página de portada como tal, *Portada* significa el texto cercano a la aparición más prominente del título del trabajo, precediendo el comienzo del cuerpo del texto.

Una sección *Titulada XYZ* significa una parte del Documento cuyo título es precisamente XYZ o contiene XYZ entre paréntesis, a continuación de texto que traduce XYZ a otro idioma (aquí XYZ se refiere a nombres de sección específicos mencionados más abajo, como *Agradecimientos*, *Dedicatorias*, *Aprobaciones* o *Historia*. *Conservar el Título* de tal sección cuando se modifica el Documento significa que permanece una sección *Titulada XYZ* según esta definición³.

El Documento puede incluir Limitaciones de Garantía cercanas a la nota donde se declara que al Documento se le aplica esta Licencia. Se considera que estas Limitaciones de Garantía están incluidas, por referencia, en la Licencia, pero sólo en cuanto a limitaciones de garantía: cualquier otra implicación que estas Limitaciones de Garantía puedan tener es nula y no tiene efecto en el significado de esta Licencia.

B.3. Copia literal

Usted puede copiar y distribuir el Documento en cualquier soporte, sea en forma comercial o no, siempre y cuando esta Licencia, las notas de copyright y la nota que indica que esta Licencia se aplica al Documento se reproduzcan en todas las copias y que usted no añada ninguna otra condición a las expuestas en esta Licencia. Usted no puede usar medidas técnicas para obstruir o controlar la lectura o copia posterior de las copias que usted haga o distribuya. Sin embargo, usted puede aceptar compensación a cambio de las copias. Si distribuye un número suficientemente grande de copias también deberá seguir las condiciones de la sección 3.

Usted también puede prestar copias, bajo las mismas condiciones establecidas anteriormente, y puede exhibir copias públicamente.

³En sentido estricto esta licencia parece exigir que los títulos sean exactamente *Acknowledgements*, *Dedications*, *Endorsements* e *History*, en inglés.

B.4. Copiado en cantidad

Si publica copias impresas del Documento (o copias en soportes que tengan normalmente cubiertas impresas) que sobrepasen las 100, y la nota de licencia del Documento exige Textos de Cubierta, debe incluir las copias con cubiertas que lleven en forma clara y legible todos esos Textos de Cubierta: Textos de Cubierta Delantera en la cubierta delantera y Textos de Cubierta Trasera en la cubierta trasera. Ambas cubiertas deben identificarlo a Usted clara y legiblemente como editor de tales copias. La cubierta debe mostrar el título completo con todas las palabras igualmente prominentes y visibles. Además puede añadir otro material en las cubiertas. Las copias con cambios limitados a las cubiertas, siempre que conserven el título del Documento y satisfagan estas condiciones, pueden considerarse como copias literales.

Si los textos requeridos para la cubierta son muy voluminosos para que ajusten legiblemente, debe colocar los primeros (tantos como sea razonable colocar) en la verdadera cubierta y situar el resto en páginas adyacentes.

Si Usted publica o distribuye copias Opacas del Documento cuya cantidad excede las 100, debe incluir una copia Transparente, que pueda ser leída por una máquina, con cada copia Opaca, o bien mostrar, en cada copia Opaca, una dirección de red donde cualquier usuario de la misma tenga acceso por medio de protocolos públicos y estandarizados a una copia Transparente del Documento completa, sin material adicional. Si usted hace uso de la última opción, deberá tomar las medidas necesarias, cuando comience la distribución de las copias Opacas en cantidad, para asegurar que esta copia Transparente permanecerá accesible en el sitio establecido por lo menos un año después de la última vez que distribuya una copia Opaca de esa edición al público (directamente o a través de sus agentes o distribuidores).

Se solicita, aunque no es requisito, que se ponga en contacto con los autores del Documento antes de redistribuir gran número de copias, para darles la oportunidad de que le proporcionen una versión actualizada del Documento.

B.5. Modificaciones

Puede copiar y distribuir una Versión Modificada del Documento bajo las condiciones de las secciones 2 y 3 anteriores, siempre que usted libere la Versión Modificada bajo esta misma Licencia, con la Versión Modificada haciendo el rol del Documento, por lo tanto dando licencia de distribución y modificación de la Versión Modificada a quienquiera posea una copia de la misma. Además, debe hacer lo siguiente en la Versión Modificada:

- A) Usar en la Portada (y en las cubiertas, si hay alguna) un título distinto al del Documento y de sus versiones anteriores (que deberían, si hay alguna, estar listadas en la sección de Historia del Documento). Puede usar el mismo título de versiones anteriores al original siempre y cuando quien las publicó originalmente otorgue permiso.
- B) Listar en la Portada, como autores, una o más personas o entidades responsables de la autoría de las modificaciones de la Versión Modificada, junto con por lo menos cinco de los autores principales del Documento (todos sus autores principales, si hay menos de cinco), a menos que le eximan de tal requisito.
- C) Mostrar en la Portada como editor el nombre del editor de la Versión Modificada.
- D) Conservar todas las notas de copyright del Documento.

- E) Añadir una nota de copyright apropiada a sus modificaciones, adyacente a las otras notas de copyright.
- F) Incluir, inmediatamente después de las notas de copyright, una nota de licencia dando el permiso para usar la Versión Modificada bajo los términos de esta Licencia, como se muestra en la Adenda al final de este documento.
- G) Conservar en esa nota de licencia el listado completo de las Secciones Invariantes y de los Textos de Cubierta que sean requeridos en la nota de Licencia del Documento original.
- H) Incluir una copia sin modificación de esta Licencia.
- I) Conservar la sección Titulada *Historia*, conservar su Título y añadirle un elemento que declare al menos el título, el año, los nuevos autores y el editor de la Versión Modificada, tal como figuran en la Portada. Si no hay una sección Titulada *Historia* en el Documento, crear una estableciendo el título, el año, los autores y el editor del Documento, tal como figuran en su Portada, añadiendo además un elemento describiendo la Versión Modificada, como se estableció en la oración anterior.
- J) Conservar la dirección en red, si la hay, dada en el Documento para el acceso público a una copia Transparente del mismo, así como las otras direcciones de red dadas en el Documento para versiones anteriores en las que estuviese basado. Pueden ubicarse en la sección *Historia*. Se puede omitir la ubicación en red de un trabajo que haya sido publicado por lo menos cuatro años antes que el Documento mismo, o si el editor original de dicha versión da permiso.
- K) En cualquier sección Titulada *Agradecimientos o Dedicatorias*, Conservar el Título de la sección y conservar en ella toda la sustancia y el tono de los agradecimientos y/o dedicatorias incluidas por cada contribuyente.
- L) Conservar todas las Secciones Invariantes del Documento, sin alterar su texto ni sus títulos. Números de sección o el equivalente no son considerados parte de los títulos de la sección.
- M) Borrar cualquier sección titulada *Aprobaciones*. Tales secciones no pueden estar incluidas en las Versiones Modificadas.
- N) No cambiar el título de ninguna sección existente a *Aprobaciones* ni a uno que entre en conflicto con el de alguna Sección Invariante.
- O) Conservar todas las Limitaciones de Garantía.

Si la Versión Modificada incluye secciones o apéndices nuevos que califiquen como Secciones Secundarias y contienen material no copiado del Documento, puede opcionalmente designar algunas o todas esas secciones como invariantes. Para hacerlo, añada sus títulos a la lista de Secciones Invariantes en la nota de licencia de la Versión Modificada. Tales títulos deben ser distintos de cualquier otro título de sección.

Puede añadir una sección titulada *Aprobaciones*, siempre que contenga únicamente aprobaciones de su Versión Modificada por otras fuentes –por ejemplo, observaciones de peritos o que el texto ha sido aprobado por una organización como la definición oficial de un estándar.

Puede añadir un pasaje de hasta cinco palabras como Texto de Cubierta Delantera y un pasaje de hasta 25 palabras como Texto de Cubierta Trasera en la Versión Modificada. Una entidad solo puede añadir (o hacer que se añada) un pasaje al Texto de Cubierta Delantera y uno al de Cubierta Trasera. Si el Documento ya incluye un textos de cubiertas

añadidos previamente por usted o por la misma entidad que usted representa, usted no puede añadir otro; pero puede reemplazar el anterior, con permiso explícito del editor que agregó el texto anterior.

Con esta Licencia ni los autores ni los editores del Documento dan permiso para usar sus nombres para publicidad ni para asegurar o implicar aprobación de cualquier Versión Modificada.

B.6. Combinación de documentos

Usted puede combinar el Documento con otros documentos liberados bajo esta Licencia, bajo los términos definidos en la sección 4 anterior para versiones modificadas, siempre que incluya en la combinación todas las Secciones Invariantes de todos los documentos originales, sin modificar, listadas todas como Secciones Invariantes del trabajo combinado en su nota de licencia. Así mismo debe incluir la Limitación de Garantía.

El trabajo combinado necesita contener solamente una copia de esta Licencia, y puede reemplazar varias Secciones Invariantes idénticas por una sola copia. Si hay varias Secciones Invariantes con el mismo nombre pero con contenidos diferentes, haga el título de cada una de estas secciones único añadiéndole al final del mismo, entre paréntesis, el nombre del autor o editor original de esa sección, si es conocido, o si no, un número único. Haga el mismo ajuste a los títulos de sección en la lista de Secciones Invariantes de la nota de licencia del trabajo combinado.

En la combinación, debe combinar cualquier sección Titulada *Historia* de los documentos originales, formando una sección Titulada *Historia*; de la misma forma combine cualquier sección Titulada *Agradecimientos*, y cualquier sección Titulada *Dedicatorias*. Debe borrar todas las secciones tituladas *Aprobaciones*.

B.7. Colecciones de documentos

Puede hacer una colección que conste del Documento y de otros documentos liberados bajo esta Licencia, y reemplazar las copias individuales de esta Licencia en todos los documentos por una sola copia que esté incluida en la colección, siempre que siga las reglas de esta Licencia para cada copia literal de cada uno de los documentos en cualquiera de los demás aspectos.

Puede extraer un solo documento de una de tales colecciones y distribuirlo individualmente bajo esta Licencia, siempre que inserte una copia de esta Licencia en el documento extraído, y siga esta Licencia en todos los demás aspectos relativos a la copia literal de dicho documento.

B.8. Agregación con trabajos independientes

Una recopilación que conste del Documento o sus derivados y de otros documentos o trabajos separados e independientes, en cualquier soporte de almacenamiento o distribución, se denomina un *agregado* si el copyright resultante de la compilación no se usa para limitar los derechos de los usuarios de la misma más allá de lo que los de los trabajos individuales permiten. Cuando el Documento se incluye en un agregado, esta Licencia no se aplica a otros trabajos del agregado que no sean en sí mismos derivados

del Documento.

Si el requisito de la sección 3 sobre el Texto de Cubierta es aplicable a estas copias del Documento y el Documento es menor que la mitad del agregado entero, los Textos de Cubierta del Documento pueden colocarse en cubiertas que enmarquen solamente el Documento dentro del agregado, o el equivalente electrónico de las cubiertas si el documento está en forma electrónica. En caso contrario deben aparecer en cubiertas impresas enmarcando todo el agregado.

B.9. Traducción

La Traducción es considerada como un tipo de modificación, por lo que usted puede distribuir traducciones del Documento bajo los términos de la sección 4. El reemplazo las Secciones Invariantes con traducciones requiere permiso especial de los dueños de derecho de autor, pero usted puede añadir traducciones de algunas o todas las Secciones Invariantes a las versiones originales de las mismas. Puede incluir una traducción de esta Licencia, de todas las notas de licencia del documento, así como de las Limitaciones de Garantía, siempre que incluya también la versión en Inglés de esta Licencia y las versiones originales de las notas de licencia y Limitaciones de Garantía. En caso de desacuerdo entre la traducción y la versión original en Inglés de esta Licencia, la nota de licencia o la limitación de garantía, la versión original en Inglés prevalecerá.

Si una sección del Documento está Titulada *Agradecimientos, Dedicatorias o Historia* el requisito (sección 4) de Conservar su Título (Sección 1) requerirá, típicamente, cambiar su título.

B.10. Terminación

Usted no puede copiar, modificar, sublicenciar o distribuir el Documento salvo por lo permitido expresamente por esta Licencia. Cualquier otro intento de copia, modificación, sublicenciamiento o distribución del Documento es nulo, y dará por terminados automáticamente sus derechos bajo esa Licencia. Sin embargo, los terceros que hayan recibido copias, o derechos, de usted bajo esta Licencia no verán terminadas sus licencias, siempre que permanezcan en total conformidad con ella.

B.11. Revisiones futuras de esta licencia

De vez en cuando la Free Software Foundation puede publicar versiones nuevas y revisadas de la Licencia de Documentación Libre GNU. Tales versiones nuevas serán similares en espíritu a la presente versión, pero pueden diferir en detalles para solucionar nuevos problemas o intereses. Vea <http://www.gnu.org/copyleft/>.

Cada versión de la Licencia tiene un número de versión que la distingue. Si el Documento especifica que se aplica una versión numerada en particular de esta licencia o *cualquier versión posterior*, usted tiene la opción de seguir los términos y condiciones de la versión especificada o cualquiera posterior que haya sido publicada (no como borrador) por la Free Software Foundation. Si el Documento no especifica un número de versión de esta Licencia, puede escoger cualquier versión que haya sido publicada (no como borrador) por la Free Software Foundation.

B.12. Adenda: cómo usar esta Licencia en sus documentos

Para usar esta licencia en un documento que usted haya escrito, incluya una copia de la Licencia en el documento y ponga el siguiente copyright y nota de licencia justo después de la página de título:

Copyright ©AÑO SU NOMBRE. Se otorga permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.2 o cualquier otra versión posterior publicada por la Free Software Foundation; sin Secciones Invariantes ni Textos de Cubierta Delantera ni Textos de Cubierta Trasera. Una copia de la licencia está incluida en la sección titulada *GNU Free Documentation License*.

Si tiene Secciones Invariantes, Textos de Cubierta Delantera y Textos de Cubierta Trasera, reemplace la frase *sin ... Trasera* por esto:

- siendo las Secciones Invariantes LISTE SUS TÍTULOS, siendo los Textos de Cubierta Delantera LISTAR, y siendo sus Textos de Cubierta Trasera LISTAR.

Si tiene Secciones Invariantes sin Textos de Cubierta o cualquier otra combinación de los tres, mezcle ambas alternativas para adaptarse a la situación.

Si su documento contiene ejemplos de código de programa no triviales, recomendamos liberar estos ejemplos en paralelo bajo la licencia de software libre que usted elija, como la Licencia Pública General de GNU (*GNU General Public License*), para permitir su uso en software libre.

Bibliografía

- [1] Laboratorios, prácticos, apuntes y bibliografía de la materia MOTORES DE BASE DE DATOS – Ing. Mauricio Rizzi, Ing. Mariano García Mattío – Instituto Universitario Aeronáutico (IUA) - Año 2006.
- [2] Laboratorios, prácticos, apuntes y bibliografía de la materia SOPORTE PARA LA TOMA DE DECISIONES – Ing. María Alejandra Boggio – Instituto Universitario Aeronáutico (IUA) – Año 2006.
- [3] Laboratorios, prácticos, apuntes y bibliografía del curso SISTEMAS AVANZADOS DE BASE DE DATOS CON SOPORTE PARA LA TOMA DE DECISIONES – Ing. Mauricio Rizzi – Universidad Católica de Córdoba (UCC) – Año 2006.
- [4] WIKIPEDIA.org y links relacionados.
- [5] ESTRATEGIA COMPETITIVA, Técnicas para el Análisis de los Sectores Industriales y de la Competencia – Michael E. Porter – Año 2000 – Vigésima séptima reimpresión.
- [6] EL NUEVO DIRECTIVO RACIONAL, Análisis de problemas y toma de decisiones – Charles H. Kepner, Benjamin B. Tregoe – Ed. McGraw-Hill – Año 1992
- [7] Ingeniería del Software, Un enfoque práctico – Roger S. Pressman. MacGraw-Hill – Año 2001 – 5ta Edición.

Índice de figuras

1.1. Fases del proceso BI.	7
2.1. Data Warehouse, características.	10
2.2. Data Warehouse, variante en el tiempo.	12
2.3. Data Warehouse, no volátil.	13
2.4. Data Warehouse, estructura.	16
2.5. Data Warehouse, flujo de datos.	17
3.1. Data Warehouse, arquitectura.	19
3.2. OLTP.	20
3.3. Load Manager.	21
3.4. Transformación: codificación.	22
3.5. Transformación: medida de atributos.	23
3.6. Transformación: convenciones de nombramiento.	23
3.7. Transformación: fuentes múltiples.	24
3.8. Proceso ETL.	25
3.9. Data Warehouse Manager.	26
3.10Cubo multidimensional.	27
3.11Tablas de Dimensiones.	28
3.12Jerarquía de "GEOGRAFIA".	30
3.13Tablas de Hechos.	31
3.14Esquema en Estrella.	32
3.15Esquema en Estrella, ejemplo.	33
3.16Desnormalización.	33
3.17Esquema Copo de Nieve.	34
3.18Esquema Constelación.	35
3.19OLTP vs Data Warehouse.	36
3.20Cubo multidimensional de dos dimensiones.	38
3.21Esquema en Estrella de dos dimensiones.	38
3.22Esquema en Estrella de tres dimensiones.	39
3.23Cubo multidimensional de tres dimensiones.	39
3.24Esquema en Estrella de cuatro dimensiones.	40
3.25Cubo multidimensional de cuatro dimensiones.	40
3.26Query Manager.	42
3.27Esquema en Estrella.	43
3.28Jerarquía de "PRODUCTOS".	43
3.29Drill-down.	43
3.30Tabla antes de aplicar Drill-down.	44
3.31Tabla después de aplicar Drill-down.	44
3.32Drill-down, representación matricial.	45
3.33Drill-up.	45
3.34Tabla antes de aplicar Drill-up.	45

3.35.Tabla después de aplicar Drill-up.	46
3.36Drill-up, representación matricial.	46
3.37.Tabla antes de aplicar Drill-across.	46
3.38.Tabla después de aplicar Drill-across.	47
3.39Drill-across, representación matricial.	47
3.40.Tabla antes de aplicar Roll-across.	47
3.41.Tabla después de aplicar Roll-across.	47
3.42Roll-across, representación matricial.	48
3.43.Tabla antes de aplicar Pivot.	48
3.44.Tabla después de aplicar Pivot.	48
3.45Pivot, representación matricial.	49
3.46Page.	49
3.47.Tabla antes de aplicar Page.	49
3.48Página Nro 1, representación tabular.	50
3.49Página Nro 1, representación matricial.	50
3.50Página Nro 2, representación tabular.	50
3.51Página Nro 2, representación matricial.	50
3.52Herramientas de Consulta y Análisis.	51
3.53Proceso de Consulta y Análisis	51
3.54Usuarios.	56
3.55Usuarios de OLTP vs Usuarios de DW.	57
 4.1. Top-Down.	60
4.2. Bottom-Up	60
 5.1. Metodología HEFESTO, logotipo.	67
5.2. Metodología HEFESTO, pasos.	68
5.3. Caso práctico, indicadores y perspectivas.	72
5.4. Modelo Conceptual.	72
5.5. Caso práctico, Modelo Conceptual.	73
5.6. Caso práctico, Diagrama de Entidad Relación.	74
5.7. Caso práctico, correspondencia.	75
5.8. Modelo Conceptual con atributos.	76
5.9. Caso práctico, Modelo Conceptual con atributos.	78
5.10Diseño de tablas de dimensiones.	79
5.11Caso práctico, dimensión "CLIENTE".	79
5.12Caso práctico, dimensión "PRODUCTO".	80
5.13Caso práctico, dimensión "FECHA".	80
5.14.Tabla de hechos.	80
5.15Caso 1, preguntas.	81
5.16Caso 1, diseño de tablas de hechos.	81
5.17Caso 2, preguntas.	81
5.18Caso 2, diseño de tablas de hechos.	82
5.19Caso 3, preguntas.	82
5.20Caso 3, unificación.	82
5.21Caso práctico, diseño de la tabla de hechos.	83
5.22Caso práctico, uniones.	83
5.23Jerarquía de "GEOGRAFIA".	84
5.24Jerarquía de "GEOGRAFIA", representación.	84
5.25Normalización de la jerarquía de "GEOGRAFIA".	84
5.26Caso práctico, jerarquía de "FECHA"	85
5.27Caso práctico, representación de la jerarquía de "FECHA"	85
5.28Caso práctico, jerarquía de "PRODUCTO"	85
5.29Caso práctico, representación de la jerarquía de "FECHA"	85

5.30 Caso práctico, Modelo Lógico.	86
5.31 Caso práctico, sentencia SQL de "CLIENTES".	87
5.32 Caso práctico, sentencia SQL de "PRODUCTOS".	88
5.33 Caso práctico, datos de "FECHA".	89
5.34 Caso práctico, fórmula "yyyymmdd".	89
5.35 Caso práctico, sentencia SQL de "VENTAS".	90
6.1. Trayectoria acíclica.	94
6.2. Trayectoria cíclica.	95
6.3. Tabla de dimensión "PRODUCTO".	95
6.4. Jerarquía de "PRODUCTO".	96
A.1. Organigrama.	98

Índice alfabético

- Análisis de requerimientos, 70
Base de datos multidimensional, 27
Bottom-Up, 60
Business Intelligence, 5

Carga, 25
Carga Inicial (Initial Load), 25
Carga Total (Full Load), 26
Codificación, 22
Codificar, 96
Consultas, 52
Convenciones de nombramiento, 23
Correspondencias, 73
Cubo Multidimensional, 38

Data Mart, 59, 94
Data Mining, 53
Data Warehouse, 9
Data Warehouse Manager, 26
Dato agregado, 31
Datos altamente resumidos, 17
Datos Anómalos (Outliers), 24
Datos de referencia, 29
Datos Faltantes (Missing Values), 24
Datos ligeramente resumidos, 17
Desnormalización, 33
Detalle de datos actuales, 16
Detalle de datos históricos, 16
Detección de Desviación, 56
Dimensión Tiempo, 29
Discretizar, 24, 96
Drill-across, 46
Drill-down, 43
Drill-up, 45

EIS, 56
Esquema Constelación, 35
Esquema Copo de Nieve, 34
Esquema en Estrella, 32
Estandarizar, 96
ETL, 21, 86, 92
Extracción, 21

Factorizar, 96
Fuentes múltiples, 24
Granularidad, 30
Hechos, 31
HEFESTO, 67
Herramientas de Consulta y Análisis, 51
HOLAP, 37

Indicadores, 71
Integrada, 11
Jerarquías, 29, 83

Limpieza de Datos (Data Cleansing), 24, 86
Load Manager, 21

Mapping, 41
Medida de atributos, 23
Metadatos, 17, 40
Modelo Conceptual, 72
Modelo Conceptual con atributos, 75
Modelo Lógico, 78
MOLAP, 37

No volátil, 12
Normalización, 33

OLAP, 52
OLTP, 20
Orientada al negocio, 10

Page, 49
Perspectivas, 71
Pivot, 48
Programación Genética, 55

Query Manager, 42

Redes Neuronales, 54
Redundancia, 15
Relación, 30
Reportes, 52
ROLAP, 36

Roll-across, 47
Sello de tiempo, 12
Sentencias SQL, 86
SGBD, 61
Sistema de Misión Crítica, 59, 92
Sistemas Expertos, 55

Tablas de Dimensiones, 28, 79
Tablas de Hechos, 31, 80
Top-Down, 60
Transformación, 22

Uniones, 83
Usuarios, 56

Variante en el tiempo, 12
Árboles de Decisión, 55