UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES
A C A T L A N

SEMINARIO TALLER EXTRACURRICULAR.

"METODOLOGIA PARA EL DISEÑO LOGICO DE UNA BASE DE DATOS DIMENSIONAL DE UN DATA WAREHOUSE"

QUE PARA OBTENER EL TITULO DE LICENCIADO EN MATEMATICAS APLICADAS Y COMPUTACION

P R E S E N T A :

ERIKA ORTIZ NAVARRO

ASESOR:
ING. RUBÉN ROMERO RUIZ

NAUCALPAN, EDO DE MEXICO Abril ,2004

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES
A C A T L A N

"METODOLOGIA PARA EL DISEÑO LOGICO DE UNA BASE DE DATOS DIMENSIONAL DE UN DATA WAREHOUSE"

SEMINARIO TALLER EXTRACURRICULAR.

QUE PARA OBTENER EL TITULO DE:
LICENCIADO EN MATEMATICAS APLICADAS Y COMPUTACION

PRESENTA

ERIKA ORTIZ NAVARRO

ASESOR: ING. RUBÉN ROMERO RUIZ

Fecha: Abril ,2004

DEDICATORIAS

A mi papá porque siempre me ha apoyado y a mi mamá que a pesar de todo siempre ha estado al pie del cañón y pendiente de nosotras.

A mis hermanas Marcela y Daniela por estar siempre conmigo y apoyarme.

A Miguel porque me ha aguantado y apoyado todos estos meses del seminario y de mucho trabajo.

A mis maestros porque gracias a su paciencia, dedicación y tiempo ha sido posible terminar esta etapa.

Y a mis amigos de la carrera Ru, Juan, Moncho, Eréndira, Blanca, Alfredo, Paco, Edith, Julián, Javier, Lázaro, Oscar, Nigie y Krusty por su amistad y porque, aunque cada quien ha tomado su camino, siempre van a quedar los recuerdos de las clases, las fiestas y las reuniones en la bardita.

ERIKA

INDICE

INTRODUCCIÓN

En la actualidad los egresados de la carrera de MAC salen al mercado laboral a competir con miles de estudiantes de otras universidades, tecnológicos, etc. por lo tanto, requieren contar con los conocimientos generales de las tecnologías aplicadas actualmente, tanto en el sector público como en el privado, con la finalidad de tener mayores oportunidades. Sin embargo, muchas de estas tecnologías no se encuentran al alcance de los estudiantes debido al costo de las mismas, a la poca información en las bibliotecas, al alto costo de libros sobre la materia o al alto grado de complejidad que manejan las fuentes existentes.

Una de las tecnologías que ha tomado mayor auge en los últimos años es la conocida como data warehouse o almacén de datos. Esta tecnología actualmente es utilizada, cada vez con mayor frecuencia tanto en la iniciativa privada como en las instituciones públicas, para el soporte a la toma de decisiones. El uso oportuno de esta información puede ayudar a una empresa a conocer el comportamiento del mercado para tomar la decisión correcta en el momento oportuno con la finalidad de ser más competitivas.

Debido a las características especiales que debe presentar un data warehouse el diseño de la base de datos que almacena esta información puede ser creada utilizando diferentes técnicas. Una de estas técnicas es el diseño dimensional que consiste en la representación de las dimensiones, los hechos y las relaciones existentes en uno o más procesos de negocio.

Para el diseño de este tipo de esquemas existen metodologías que el diseñador puede seguir para elaborar las preguntas correctas que lo orienten durante este

proceso. Por esta razón se presenta una metodología que puede ser utilizada como una guía en el diseño de bases de datos dimensionales. Esta metodología abarca el diseño conceptual, lógico y físico de la base de datos. Sin embargo, en este trabajo solo serán tratados los pasos de la metodología relacionados al diseño lógico.

En el primer capítulo se presentan los fundamentos de las bases de datos para que el estudiante adquiera o repase los conocimientos mínimos necesarios para comprender la información de los siguientes capítulos.

En el segundo capítulo, es presentada la definición de data warehouse, los elementos que lo componen, una comparación del almacén de datos con los sistemas transaccionales, la definición de un modelo dimensional y los elementos que lo componen. Esta información es importante para comprender la importancia y la función que tiene un data warehouse en una organización y el porque la necesidad de una base de datos con las características del modelo dimensional.

Por último, en el tercer capítulo, se presenta la metodología para el diseño lógico de una base de datos dimensional y la aplicación de la teoría presentada en los capítulos anteriores en un caso práctico sobre el diseño lógico de una base de datos dimensional para la consulta del tráfico de llamadas de una empresa de telefonía móvil.

El presente trabajo pretende aportar material de apoyo para la materia de bases de datos impartida a los estudiantes de la carrera de MAC o para cualquier persona que quiera diseñar una base de datos de este tipo y que cuente con conocimientos básicos sobre la materia.

CAPÍTULO I
FUNDAMENTOS DE BASES DE DATOS

En este capítulo se presenta una breve introducción a las bases de datos, su definición, características, elementos que las componen, el modelo relacional, el modelo de base de datos mas popular y utilizado actualmente, y las fases a seguir en el proceso de diseño de una base de datos relacional. Este capítulo tiene como objetivo repasar los conceptos fundamentales de bases de datos para comprender mejor los conceptos presentados en los siguientes capítulos.

1.1 Definición de Base de Datos

En la actualidad, las bases de datos desempeñan un papel primordial en casi todas las áreas, la ingeniería, la medicina, la educación, por mencionar sólo algunas de ellas. La mayoría de nosotros hemos interactuado con ellas directa o indirectamente; por ejemplo, son utilizadas cuando se lleva a cabo una reservación de un boleto de avión en alguna aerolínea, cuando se abre una cuenta en el banco o cuando se realiza el pago del teléfono celular.

Pero, ¿qué es una base de datos? "Una **base de datos** es un conjunto de datos relacionados entre sí. Por datos entendemos hechos conocidos que pueden registrarse y que tienen un significado implícito" (Elmasri y Navathe,1997).

Esto significa que una base de datos es un conjunto de datos que representan una serie de hechos que requieren ser almacenados y pueden proporcionar cierta información útil a la organización; por ejemplo, si una persona decide hacer la contratación de un plan de telefonía móvil, el cliente es dado de alta en la base de

datos de la empresa para llevar un registro de sus datos personales y la contratación realizada para facturar las llamadas generadas y cobrar los servicios que le proporciona.

Para la construcción, almacenamiento, administración y explotación de una Base de Datos es utilizado un **Sistema de Gestión de Bases de Datos** (SGBD). Un SGBD es: "Un sistema de software de propósito general que facilita al usuario el proceso de definir, construir y manipular las bases de datos" (Elmasri y Navathe, 1997). Es decir, se trata de un conjunto de programas para la administración y manipulación de la información contenida en la base de datos, este puede ser un producto existente en el mercado o un conjunto de programas creados para este fin (*figura 1.1*).

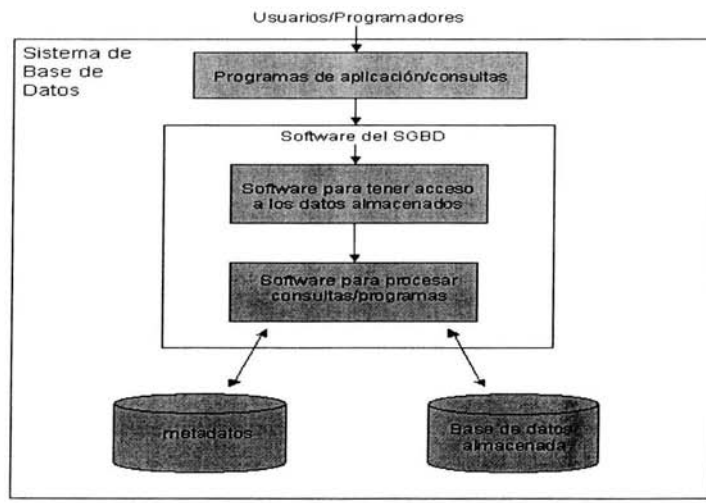

Figura 1.1 Entorno simplificado de un sistema de base de datos (Elmasri y Navathe, 1997)

1.2 ¿Por qué utilizar una base de datos?

Antes de ser utilizadas las bases de datos en las empresas y organizaciones la información era almacenada y manipulada por medio de archivos. Cada usuario definía y creaba los archivos requeridos para una aplicación específica; por ejemplo, un usuario de servicios escolares almacenaba la información sobre los estudiantes y sus calificaciones en archivos que ellos mismos actualizaban o en los casos mas automatizados mediante aplicaciones creadas para una tarea y un archivo en específico. Sin embargo, el manejo de archivos aislados presentó diversos problemas como el acceso de un solo usuario, duplicidad de la información, inconsistencias, problemas para garantizar la seguridad del sistema, etc., por citar sólo algunos de ellos.

El surgimiento de las bases de datos solucionaron estos problemas y proporcionaron otras ventajas como la compactación de la información, rapidez en la consulta de los datos, simplificación de tareas e información actualizada en el momento de ser consultada.

Todas estas ventajas se deben a las características principales que las bases de datos presentan y que son mencionadas a continuación:

- Naturaleza autodescriptiva de los sistemas de bases de datos.
- Abstracción de los datos.
- Manejo de múltiples vistas de los datos
- Compartimiento de datos y procesamiento de transacciones multiusuario.

1.2.1 Naturaleza autodescriptiva

Esta característica es fundamental ya que el sistema no sólo contiene la base de datos, sino también una definición o descripción completa de la misma. Esta definición es almacenada en un **"catálogo"** del sistema, que contiene información como la estructura, el tipo y formato de almacenamiento de cada elemento que la conforma y diversas restricciones que se aplican a los datos. A la información almacenada en este "catálogo" se le denomina **metadatos** definidos comúnmente como los datos de los datos (*figura 1.1*).

1.2.2 Abstracción de los datos.

En el procesamiento de archivos, los programas de acceso son creados dependiendo de la estructura de los datos que estos contienen, por lo mismo cualquier modificación a la estructura del archivo implicaba cambios en la aplicación que lo utilizaba. Sin embargo, en el esquema de bases de datos sólo es necesario efectuar las modificaciones necesarias en los metadatos y las aplicaciones que consultan esta información pueden continuar operando de la misma manera sin necesidad de ser modificadas.

El SGBD proporciona a los usuarios una representación conceptual de los datos que oculta muchos de los detalles de cómo son almacenados. Esta representación conceptual es presentada en los llamados **modelos de datos** *(figura 1.2)* que son un tipo de abstracción de los mismos. Más adelante se tratará con mas detalle el concepto de modelo de datos.

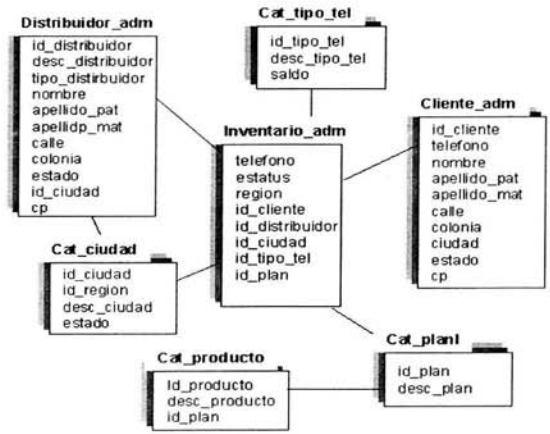

Figura 1.2 Ejemplo de un modelo de datos

1.2.3 Manejo de múltiples vistas de los datos

Una base de datos puede tener muchos usuarios, cada uno de ellos puede requerir una perspectiva o vista distinta de los mismos datos. Una vista puede estar compuesta por datos contenidos en la base de datos o presentar datos "virtuales" que se deriven de la información almacenada explícitamente en ella.

Por ejemplo, en la *figura 1.3,* la tabla *calificación* (a) almacena los datos de las materias cursadas por los alumnos, esta tabla contiene el número de cuenta, la clave de la carrera cursada por el estudiante, la clave de la materia y la calificación. Otra tabla, *cat_materia* (b), es un catálogo con las materias que existen en todas las carreras, contiene la clave de la materia, la clave de la carrera en la cual es impartida, el número de créditos de cada materia y una descripción de la misma.

(a)Tabla: Calificacion

No_cuenta	Clv_carrera	Clv_Materia	Calificacion
09553459-6	042	0213	8
09423125-5	041	0362	7
09553459-6	042	0124	9
03526458-3	043	0241	9
09553459-6	042	0163	7

(b)Tabla: Cat_materia

Clv_materia	Clv_carrera	No_creditos	Descripcion
0124	042	6	BASES DE DATOS
0213	042	8	ÁLGEBRA LINEAL I
0163	042	4	ECONOMIA

Reporte del usuario

No. Cuenta: 09553459-6

Materia	Calificación	Créditos
BASES DE DATOS	9	6
ÁLGEBRA LINEAL I	8	8

Figura 1.3 Vista de la información requerida por un usuario (Elmasri y Navathe, 1997)

Suponiendo que un usuario desea consultar las calificaciones de un alumno y sus créditos, es posible que la información requerida no se encuentre almacenada de esta manera, sin embargo, los datos pueden ser consultados por medio de uniones entre dos o más tablas de la base de datos. Incluso si el usuario necesita obtener información como el promedio de las calificaciones del alumno, el número de materias aprobadas o reprobadas, el número de alumnos que cursan una materia, etc., a pesar de que estos datos no se encuentran almacenados explícitamente en las tablas pueden ser calculados con los datos existentes en las mismas.

1.2.4 Acceso y procesamiento multiusuario

El acceso simultáneo a la base de datos de varios usuarios es indispensable para que los datos de múltiples aplicaciones se integren y mantengan la consistencia de los mismos. Esta característica es de gran utilidad ya que permite el control de la concurrencia de las transacciones para asegurar la actualización controlada de los

datos, cuando varios usuarios intentan modificar la misma información se garantiza que sea modificada con datos actuales y evitar errores.

1.2.5 Modelo de una base de datos

Una de las características más importantes de las bases de datos, ya mencionada, es la posibilidad de abstracción de los datos al ocultar al usuario los detalles de almacenamiento que no necesita conocer para que pueda comprender mejor la organización de la información. Los modelos de datos son el principal instrumento para ofrecer dicha abstracción. Un modelo de base de datos es definido de la siguiente manera:

"Un **modelo de datos** es un conjunto de conceptos que describen la estructura de una base de datos. Con el concepto de estructura de una base de datos nos referimos a los tipos de datos, los vínculos y las restricciones que deben cumplirse para esos datos" (Elmasri y Navathe,1997).

En esta definición el modelo de datos no es más que una representación de la estructura de una base de datos. La finalidad del modelo es mostrar al usuario los datos almacenados, la relación que existe entre ellos y las restricciones que deberán ser tomadas en cuenta para su almacenamiento.

Existen diferentes tipos de modelos presentados a continuación:

- El modelo relacional de base de datos.
- El modelo de base de datos orientado a objetos.
- El modelo deductivo de base de datos.
- El modelo jerárquico de base de datos.
- El modelo de base de datos en red.

La *figura 1.4* muestra una tabla (James L. Jonson, 1997) con las características de los cinco modelos de bases de datos. Cada modelo utiliza un estilo particular de lenguaje, algunos identifican sus elementos por su valor, otros en base al registro.

Actualmente el **modelo relacional** es el más popular desde su aparición en 1969 y 1970. Desde entonces este modelo ha acaparado el mercado y hasta la fecha continúa con el liderazgo en la preferencia de los usuarios.

Modelo	Organización de elementos de Datos	Identidad
Relacional	Los identificadores para filas de una tabla se insertan como valores de atributo en otra tabla.	Con base en valor
Orientado a Objetos	Los objetos relacionados, de contención lógica, se encuentran dentro de un objeto determinado por examen recursivo de atributos de un objeto que son los objetos mismos.	Con base en registro
Deductivo	Reglas que permiten generar a petición, hechos relacionalos.	Con base en valor
Jerárquico	Proximidad lógica en un árbol linealizado.	Con base en registro
Red	Cadenas que se intersectan	Con base en registro

Figura 1.4 Características de los cinco modelos de bases de datos (James L. Jonson, 1997)

Los modelos orientados a objetos y deductivo son tecnologías post-relacionales, esto quiere decir que son modelos que surgieron después del modelo relacional, y los modelos jerárquicos y de redes representan tecnologías pre-relacionales o tecnologías existentes antes de la llegada del modelo relacional.

1.3 El modelo relacional

El modelo relacional es sin lugar a dudas el fundamento de la Tecnología Moderna de Base de Datos. En 1969 el Dr. Edgar Frank Codd comenzó a trabajar en él. En 1969 publicó el primero de una serie de artículos que cambiarían la forma de ver de las bases de datos. Desde entonces otras personas han hechos contribuciones importantes a este tipo de modelo, sin embargo, ninguna tan relevante como el trabajo presentado por el Dr. Codd.

Término relacional formal	Equivalente informal
Relación	Tabla
Tupla	Fila o registro
Cardinalidad	Número de filas
Atributo	Columna o campo
Grado	Número de columnas
Dominio	Conjunto de valores válidos
Llave primaria	Identificador único

Figura 1.5 Terminología estructural (James L. Johnson, 1997)

Como se mencionó anteriormente un modelo es un conjunto de conceptos que describen la estructura de una base de datos. En el modelo relacional los conceptos estructurales más importantes son: relación, tupla, cardinalidad, atributo, grado, dominio y llave primaria (*figura 1.5)*.

Una definición de las estructuras mencionadas es la siguiente: "La **relación** es el elemento básico del modelo relacional, y es posible representarlo por medio de una tabla. En ella se distinguen un conjunto de columnas, denominadas **atributos**, que representan sus propiedades y están caracterizadas por un nombre; y un conjunto de filas llamadas **tuplas**, que son las concurrencias de la relación. El número de filas de una relación se denomina **cardinalidad**, mientras que el número de columnas es el **grado**. Existen también **dominios** de los cuales los atributos toman sus valores. Y por último, la **llave** primaria es el identificador único de una tupla o registro" (James L. Jonson, 1997). Una relación puede ser representada en forma de tabla (*figura 1.6*), aunque tiene una serie de elementos característicos que la distinguen de una tabla común:

- No puede haber filas duplicadas, es decir, todas las tuplas o renglones tienen que ser distintas.
- El orden de las filas es irrelevante.
- La tabla es plana, es decir, en el cruce de una fila y una columna sólo puede haber un valor.

Las características antes mencionadas son restricciones o reglas que se deben cumplir en el modelo relacional. Una relación siempre tiene un nombre, y se debe distinguir una cabecera o los atributos que la definen y un cuerpo formado por un conjunto de tuplas o renglones que varían con el tiempo.

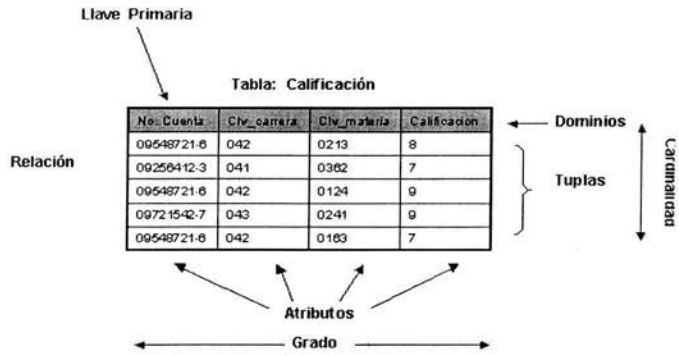

Figura 1.6 Tabla de un modelo relacional (James L. Jonson, 1997)

1.3.1 Dominio y atributo

El dominio se define de la siguiente manera: "Un **dominio** D es un conjunto finito de valores homogéneos y atómicos, V_1, V_2, ..., V_n caracterizados por un nombre" (Miguel y Mario Piattini, 1993).

Esto quiere decir que es un grupo de valores *homogéneos* (todos son del mismo tipo) y *atómicos* o indivisibles, ya que si se dividieran perderían su significado *(figura 1.7)*. Todo dominio debe tener un tipo de dato y un nombre para referirse a éste. El nombre de un dominio puede ser definido dependiendo de la información a la cual se refiera. Un ejemplo de un dominio es la edad de las personas que laboran en una empresa; este dominio puede variar entre 18 y 65 años y puede ser definido como un entero.

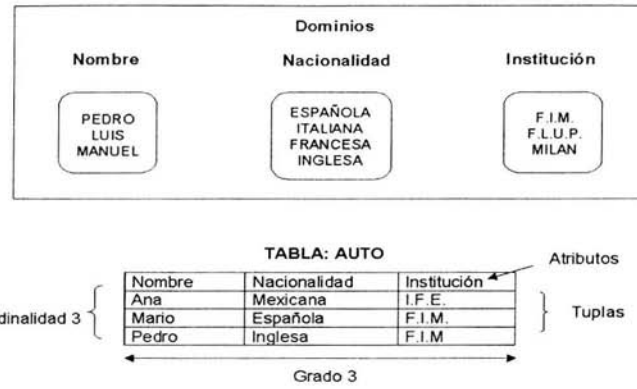

Figura 1.7 Representación de la tabla AUTOR (Miguel y Mario Piattini, 1993)

Un atributo es definido de la siguiente manera: "Un **atributo** A es el papel que tiene un determinado dominio D en una relación; se dice que D es el dominio de A y se denota como dom(A)" (Miguel y Mario Piattini, 1993). Por ejemplo, existe un atributo *nacionalidad* en la tabla *AUTOR (figura 1.7)*, definido sobre el dominio de nacionalidad (mexicana, inglés, africano, etc.) que indica los posibles valores que puede tomar este atributo.

1.3.2 Relación

La relación, en este modelo, es una tabla de dos dimensiones que contiene datos heterogéneos, los renglones que la componen no necesariamente se encuentran ordenados, y se utiliza en el modelo relacional como sinónimo de tabla.

La abstracción de los datos contenidos en una tabla es llamado extensión, muestra los atributos, sus nombres y los valores que contienen como una matriz (*figura 1.8*). Y la intención de una relación es el nombre de la relación y los nombres de los atributos listados ente paréntesis (*figura 1.8*).

INTENCIÓN DE UNA RELACIÓN

AUTOR(NOMBRE:NOMBRES, NACIONALIDAD:NACIONALIDADES, INSTITUCIÓN:INSTITUCIONES)

EXTENSIÓN DE UNA RELACIÓN

TABLA: AUTOR

NOMBRE	NACIONALIDAD	INSTITUCION
DATE	NORTEAMERICADA	RELATIONAL
DE MIGUEL	ESPAÑOLA	F.I.M.
SALTOR	ESPAÑOLA	F.I.

Figura 1.8 Intensión y extensión de una relación (Miguel y Mario Piattini, 1993)

1.3.3 Llave

Una llave o clave es definida de la siguiente manera: "La **llave** de una relación es un conjunto no vacío de atributos que identifican de manera única cada tupla" (Migue y Mario Piattini, 1993).

En un modelo relacional siempre existirá una llave o clave candidata debido a que no existen dos tuplas repetidas Una relación puede tener más de una clave o llave, entre las cuales se debe distinguir:

- **Llave o clave primaria:** es la clave que el usuario elige, por consideraciones ajenas al modelo relacional, para identificar las tuplas de la relación. Por ejemplo, en la *figura 1.9*, se observa que la llave primaria se encuentra formada por el número de cuenta del alumno (asignada a cada estudiante), la clave de la carrera que cursa, la clave de la materia y el período que cursa. Un estudiante no puede cursar la materia de una carrera dos veces en el mismo período, por tal motivo este hecho sólo puede ser registrado una vez en la base de datos.

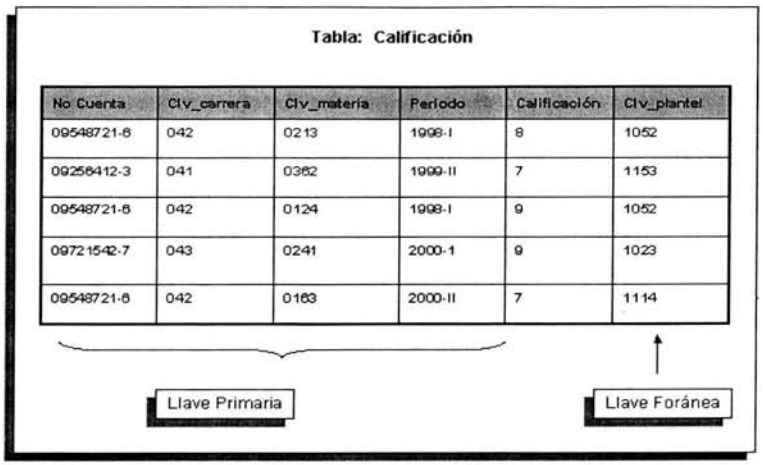

Figura 1.9 Ejemplo de llaves o claves de una relación

- **Llave o clave foránea:** es el conjunto de atributos cuyos valores coinciden con la llave primaria de otra tabla. La llave primaria y llave foránea correspondiente deberán de estar definidas sobre los mismos dominios. Esto quiere decir que el valor de la llave foránea toma los mismos valores que la llave primaria de otra tabla, en el caso del ejemplo presentado en la *figura 1.9,* existe otra tabla cuya llave primaria presenta los mismos valores que el campo clv_plantel el cual es la llave foránea de la tabla calificación.

1.3.4 Restricciones

El modelo relacional, al igual que otros modelos, presenta **restricciones**, es decir, estructuras u ocurrencias no permitidas. En el modelo relacional existen dos tipos de restricciones:

- **Restricciones inherentes al modelo:** una serie de características propias de una relación que deben ser cumplidas obligatoriamente por lo cual se

convierten en restricciones inherentes. Por ejemplo, en el modelo relacional no deben existir tuplas iguales, el orden de las tuplas no es significativo y cada atributo debe obtener un valor único del domino.

- **Restricciones indicadas por el usuario:** es un conjunto de reglas indicadas por el usuario y definidas sobre atributos, tuplas o dominios, que deben ser verificadas por los correspondientes objetos de la base de datos. Por ejemplo, en el campo donde se almacena la edad de los alumnos no es posible almacenar valores negativos o cero. El usuario se puede asegurar de cumplir con esta regla creando una restricción en la base de datos que impida la inserción de estos valores no permitidos según las reglas establecidas por el usuario.

1.3.5 Esquema

En cualquier modelo de datos es importante distinguir entre la descripción de la base de datos y la base de datos. La descripción se conoce como **esquema de la base de datos**. Este esquema es especificado durante el diseño y no debe ser modificado con frecuencia si la base de datos está bien diseñada.

En la mayoría de los modelos de datos son utilizadas ciertas convenciones para representar los esquemas en forma de diagramas. La representación de un esquema se denomina **diagrama del esquema**.

El diagrama esquemático de una base de datos (*figura 1.10*) presenta la estructura de los registros pero no los ejemplares reales de los registros. A cada uno de los objetos del esquema, como ESTUDIANTE o MATERIA, se le llama **elemento del esquema**.

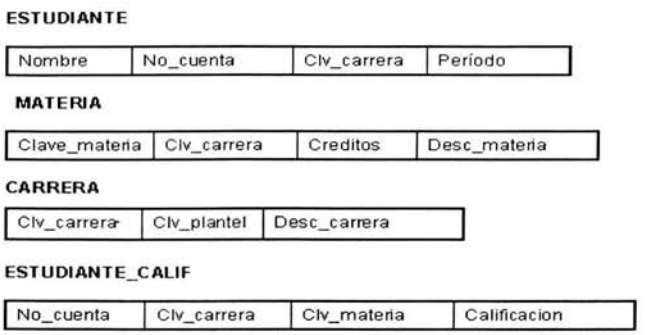

Figura 1.10 Diagrama de esquema para una base de datos (Elmasri y Navathe, 1997)

Los diagramas de esquema sólo ilustran algunos aspectos de la base de datos, como los nombres de los registros y algunas clases de restricciones. Sin embargo, no se especifican otros aspectos como tipo de dato de cada elemento de información ni las relaciones entre las diferentes tablas.

1.4 Diseño de una base de datos relacional

Una vez repasados los conceptos básicos de bases de datos y del modelo relacional son presentados los pasos a seguir en el diseño de una base de datos relacional los cuales se componen de 4 fases (Elmasri y Navathe, 1997) *(figura 1.11)* enumerados a continuación:

1. Recolección y análisis de requerimientos.
2. Diseño conceptual
3. Diseño lógico
4. Diseño físico

1.4.1 Recolección y análisis de requerimientos

El primer paso es la recolección y análisis de requerimientos *(figura 1.11)*, durante esta fase los diseñadores entrevistan a los usuarios que utilizan o utilizarán la base de datos para la comprensión y documentación de los requerimientos. En paralelo, es conveniente especificar los requerimientos funcionales conocidos de la aplicación, es decir, las operaciones definidas por el usuario (o transacciones) que serán aplicadas a la base de datos, incluyendo la obtención y actualización de los datos. Para esta tarea se acostumbra utilizar técnicas específicas como los diagramas de datos.

1.4.2 Diseño conceptual

Una vez recabados y analizados todos los requerimientos, el siguiente paso *(figura 1.11)* es crear un **esquema conceptual** independientemente de cualquier consideración física. Al construir el esquema, los diseñadores descubren el significado de los datos de la empresa: encuentran entidades, atributos y relaciones.

El objetivo de esta fase es comprender la perspectiva que cada usuario tiene de los datos, la naturaleza de los mismos y su uso a través de las áreas de aplicación. El esquema conceptual se puede utilizar para que el diseñador transmita a la empresa lo que ha entendido sobre la información que ésta maneja. Para ello, ambas partes deben estar familiarizadas con la notación utilizada en el esquema. La más popular es la notación del **modelo entidad-relación**.

El esquema conceptual se construye utilizando la información que se encuentra en la especificación de los requisitos de usuario. El diseño conceptual es completamente independiente de los aspectos de implementación, como puede ser el SGBD que se vaya a usar, los programas de aplicación, los lenguajes de programación, el hardware disponible o cualquier otra consideración física. Durante todo el proceso de

desarrollo del esquema conceptual éste se prueba y se valida con los requisitos de los usuarios. El esquema conceptual es una fuente de información para el diseño lógico de la base de datos.

1.4.3 Diseño lógico

El diseño lógico es el proceso de construir un esquema de la información que utiliza la empresa, basándose en un modelo de base de datos específico, independiente del SGBD concreto que se vaya a utilizar y de cualquier otra consideración física.

En esta etapa, se transforma el esquema conceptual en un esquema lógico que utilizará las estructuras de datos del modelo de base de datos en el que se basa el SGBD que se vaya a utilizar, como puede ser el modelo relacional, el modelo de red, el modelo jerárquico o el modelo orientado a objetos. Conforme se va desarrollando el esquema lógico, éste se va probando y validando con los requisitos de usuario.

La **normalización** es una técnica que se utiliza para comprobar la validez de los esquemas lógicos basados en el modelo relacional, ya que asegura que las relaciones (tablas) obtenidas no tienen datos redundantes.

El esquema lógico es una fuente de información para el diseño físico. Además, juega un papel importante durante la etapa de mantenimiento del sistema, ya que permite que los futuros cambios que se realicen sobre los programas de aplicación o sobre los datos, se representen correctamente en la base de datos.

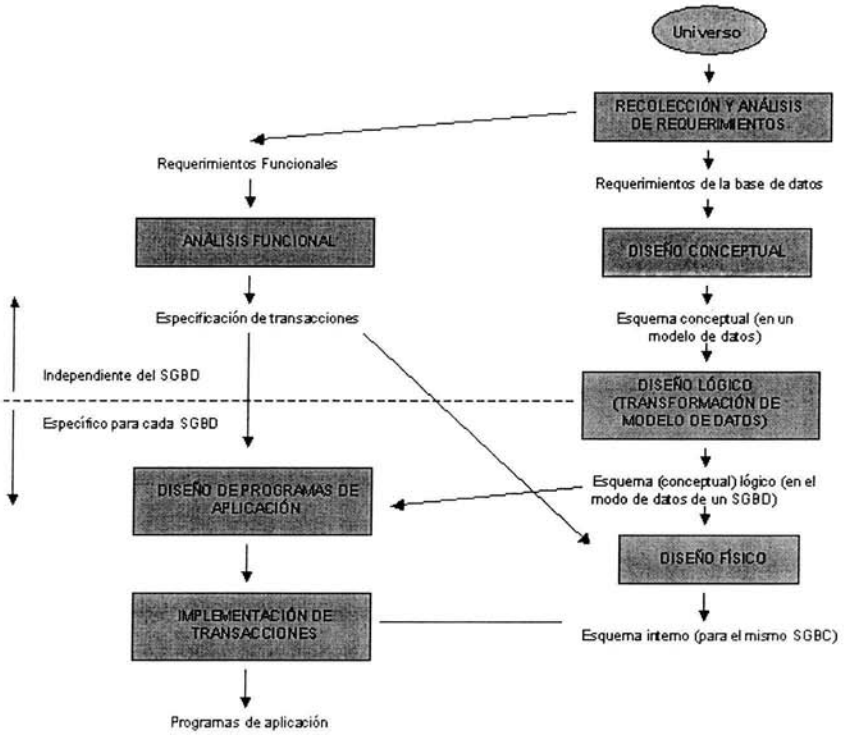

Figura 1.11 Fases del diseño de bases de datos (simplificado) (Elmasri y Navathe, 1997)

Tanto el diseño conceptual, como el diseño lógico, son procesos iterativos, tienen un punto de inicio y se van refinando continuamente. Ambos se deben ver como un proceso de aprendizaje en el que el diseñador va comprendiendo el funcionamiento de la empresa y el significado de los datos que maneja.

El diseño conceptual y el diseño lógico son etapas clave para conseguir un sistema que funcione correctamente. Si el esquema no es una representación fiel de la empresa, será difícil, sino imposible, definir todas las vistas de usuario (esquemas externos), o mantener la integridad de la base de datos. También puede ser difícil definir la implementación física.

Además, hay que tener en cuenta que la capacidad de ajustarse a futuros cambios es un sello que identifica a los buenos diseños de bases de datos. Por todo esto, es fundamental dedicar el tiempo y las energías necesarias para producir el mejor esquema posible.

1.4.4 Diseño físico

El diseño físico es el proceso de producir la descripción de la implementación de la base de datos en memoria secundaria: estructuras de almacenamiento y métodos de acceso que garanticen un acceso eficiente a los datos.

Para llevar a cabo esta etapa, se debe haber decidido cual es el SGBD que se va a utilizar, ya que el esquema físico se adapta a él. Entre el diseño físico y el diseño lógico hay una retroalimentación, debido a que algunas de las decisiones que se tomen durante el diseño físico para mejorar el performance, pueden afectar a la estructura del esquema lógico.

En general, el propósito del diseño físico es describir cómo se va a implementar físicamente el esquema lógico obtenido en la fase anterior. Concretamente, en el modelo relacional, esto consiste en:

- Obtener un conjunto de relaciones (tablas) y las restricciones que se deben cumplir sobre ellas.
- Determinar las estructuras de almacenamiento y los métodos de acceso que se van a utilizar para conseguir unas prestaciones óptimas.
- Diseñar el modelo de seguridad del sistema.

De esta manera se mostraron los conceptos fundamentales de bases de datos, sus características, tipos de modelos y se dio una breve explicación sobre las fases que comprenden el proceso de diseño de una base de datos relacional, con la finalidad de tener presentes los conceptos fundamentales de bases de datos para una mejor comprensión de los capítulos posteriores.

CAPÍTULO II
INTRODUCCIÓN AL DATA WAREHOUSE

En este capítulo es presentada una introducción al data warehouse, su definición, características, los elementos que lo componen, la definición de un sistema transaccional, las diferencias entre un data warehouse y los sistemas operacionales y el modelo dimensional como una técnica de diseño.

2.1 ¿Qué es un data warehouse?

Durante las últimas dos décadas se observó un crecimiento acelerado en la popularidad y aplicación de las computadoras para el control y organización de las empresas. La rápida evolución de la tecnología ha agilizado los procesos de las mismas y ha disminuido los costos e incrementado la competitividad de las empresas. En los últimos años se ha dado énfasis en mejorar el registro de los datos del negocio, actualmente cada transacción y cada dato es registrado en las bases de datos, en el pasado quedaron otros métodos de almacenamiento como el de archivos mencionados en el capítulo anterior.

Actualmente, la información juega un papel fundamental en el mundo de los negocios. Muchas veces el éxito de una empresa depende del uso oportuno o no de la misma. Los constantes cambios en el mercado hace necesario el acceso rápido de información que permita a las organizaciones tomar decisiones oportunas con la finalidad de ser competitivas.

Poco a poco el proceso de toma de decisiones también ha adoptado el uso de las computadoras como una herramienta para lograr organizar y analizar la información

requerida en el soporte y análisis de decisiones a problemas de estrategia y tácticas de negocio.

Sin embargo, no se ha puesto mucha atención en transformar ese registro de datos en información con un significado que pueda ser accesada fácilmente por los ejecutivos con el fin de tomar decisiones más efectivas. Aunque la mayoría de las empresas poseen sistemas que permiten el manejo de los datos de manera eficaz y eficiente, éstos únicamente se encargan de crear, almacenar y proveer de **datos** sobre ventas, compras, finanzas, etc., pero en el contexto de negocio, lo que las empresas necesitan actualmente no son sólo datos sino **información**. Esta información es proporcionada por los llamados Sistemas de Soporte a la Toma de Decisiones.

En un principio lo que el tomador de decisiones hizo fue tratar de satisfacer la necesidad de información tomando los datos de los sistemas operacionales existentes. Sin embargo, esta tarea no es sencilla debido a que estos sistemas no fueron diseñados pensando en una integración para su análisis y generalmente no contienen información histórica suficiente para cubrir las necesidades del tomador de decisiones.

Un sistema de data warehouse puede tomar el significado de los datos y usando procesos de análisis intensos, ofrece un panorama de las condiciones de mercado antes de que éstas ocurran. Por esta razón surgió el data warehouse o almacén de datos como una solución a la necesidad de información histórica y como un depósito central que permite la elaboración de resúmenes provenientes de los sistemas de producción convirtiéndose en una tecnología común y fundamental. En otras palabras, es un sistema diseñado especialmente para satisfacer las necesidades de los tomadores de decisiones.

2.1.1 Definición de data warehouse

Después de una breve explicación de lo que es un almacén de datos se presenta la definición de un data warehouse. Existen distintos conceptos de lo que es un almacén de datos. Cada organización puede tener uno diferente dependiendo del uso que se le dé a éste. Alguna de estas empresas puede señalar que se trata de un depósito de datos aislado totalmente de los sistemas operativos y de producción existentes que llenan por completo los diferentes requerimientos de acceso y reporte de datos; o también existe la idea de que es un proceso contínuo que mezcla los datos de varias fuentes heterogéneas, incluyendo datos históricos y adquiridos para soportar la constante necesidad de consultas, reportes analíticos y soporte de toma de decisiones.

De acuerdo con W. H. Inmon, considerado padre del data warehouse, un almacén de datos es: "un conjunto de datos integrados orientados a una materia, que varían con el tiempo y que son transitorios, los cuales soportan el proceso de toma de decisiones de una administración" (Inmon, 1996)

Otra definición de data warehouse es la que presenta Ralph Kimpball: "Un data warehouse es una copia de la información transaccional estructurada para llevar a cabo consultas" (Kimpball, 1996).

2.1.2 Características de un data warehouse

De la definición presentada por Inmon son identificadas cuatro características fundamentales de un data warehouse:

1. Orientado a una materia.
2. Integración de datos.
3. No volatilidad de los datos.
4. Variación con el tiempo.

2.1.2.1 Orientado a una materia

El data warehouse organiza y orienta los datos desde la perspectiva del usuario final. Los sistemas clásicos, generalmente, organizan los datos desde la perspectiva de la aplicación con el fin de recuperar y actualizar datos de manera eficiente.

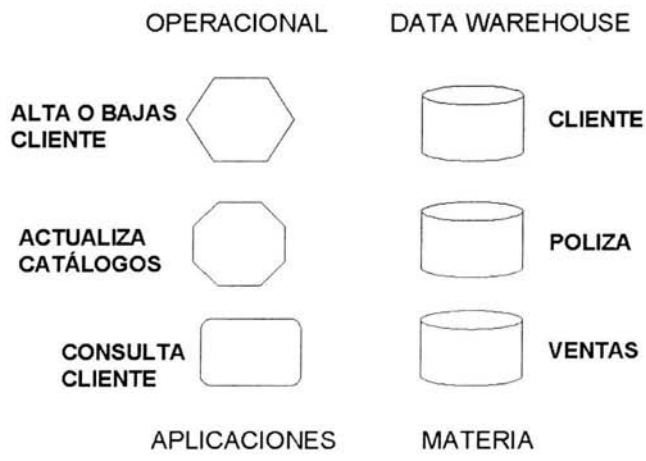

Figura 2.1 Enfoque de los sistemas OLTP y data warehouse en cuanto a la materia (Inmon, 1996)

Por ejemplo, en una compañía de seguros el sistema de ventas puede manejar información para dar de alta clientes, llevar un control de éstos y la información de los seguros que tiene contratados, etc. Sin embargo, la forma en la que están organizados los datos no es la adecuada para formular preguntas de tipo empresarial como *¿Cuál fue el tipo de póliza más vendida y la menos vendida entre el primero y el segundo bimestre del año?* o *¿Cuáles son los motivos de quejas más comunes de los clientes?*. Como es posible notar el enfoque que se da a la información y la aplicación de la misma es muy diferente, un data warehouse se enfoca a proporcionar información sobre una materia en específico, en esta caso las ventas y

los productos, el sistema de ventas se dedica a actualizar y llevar un control de los datos de manera eficiente. (*Figura 2.1*)

2.1.2.2 Integración de datos

De todos los aspectos del data warehouse este es uno de los más importantes.

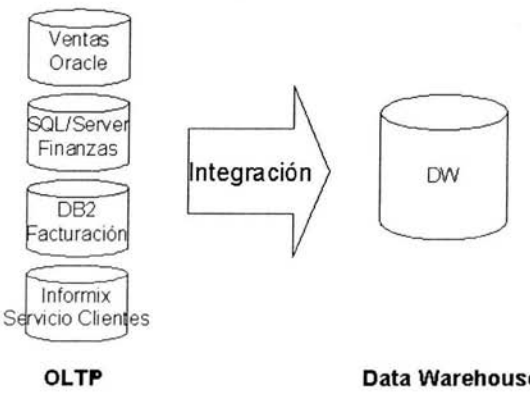

Figura 2.2 Integración de las bases de datos transaccionales de una organización en el data warehouse

Debido a que las organizaciones han administrado sus operaciones utilizando diversas aplicaciones y múltiples bases de datos (*figura 2.2*), antes de poner a disposición de los analistas y tomadores de decisiones la información recopilada de los diferentes sistemas, ésta debe ser integrada y la consistencia de su información validada.

Es posible que la información proveniente de los distintos sistemas tengan algunas diferencias como el manejo de distintos tipos de datos o de unidades (minutos, metros, etc.). Antes de cargar la información a la base de datos del data warehouse es necesario llevar a cabo su integración para eliminar las diferencias y poderla cargar.

Sist. Operacional	Valor	Integración	Valor en Data Warehouse
App. A App. B App. C App. D	m,f 1,0 x,y male,felame	código	m,f
App. A App. B App. C App. D	cm pulgadas yardas metros	medidas	cm
App. A App. B App. C App. D	descripción1 descripción2 descripción3 descripción4	múltiples valores	descripción1
App. A App. B App. C App. D	llave char(10) llave decimal llave char(12) llave integer(10)	conflictos con llaves	char(12)

Figura 2.3 Integración de los datos para cargarla al data warehouse (Inmon, 1996)

Por ejemplo, en la *figura 2.3* se observa que en la base de datos del sistema A, en una organización, las unidades manejadas en cierto campo es de centímetros (cm) mientras que en el sistema B las unidades son las pulgadas (") y en el sistema D las unidades son los metros (m). Los tres sistemas manejan el mismo concepto, sin embargo las unidades que maneja cada sistema son diferentes. En la base de datos del data warehouse este campo sólo puede contener los valores con un solo tipo de unidades por lo tanto, es necesaria la integración de la información con el fin de poder cargarla a la base de datos y poder hacer consultas como si provinieran de un solo sistema origen, para ésto las unidades de los diferentes sistemas son convertidas al que maneja el almacén de datos.

La importancia de la integración de datos no radica únicamente en el hecho de almacenar la información en un depósito de datos centralizado, la habilidad de establecer y entender la correlación entre actividades de diferentes grupos

organizacionales dentro de la empresa es una de las ventajas que ofrece la implantación de un almacén de datos.

2.1.2.3 No volatilidad de los datos

La tercera característica de un data warehouse es la no volatilidad de los datos. En los sistemas operacionales (*figura 2.4*) las entidades pasan por muchos cambios en sus atributos. A lo largo de todo el día, la base de datos recibe pequeñas transacciones que pueden ser de consulta, modficación o inserción de datos y son procesadas una por una.

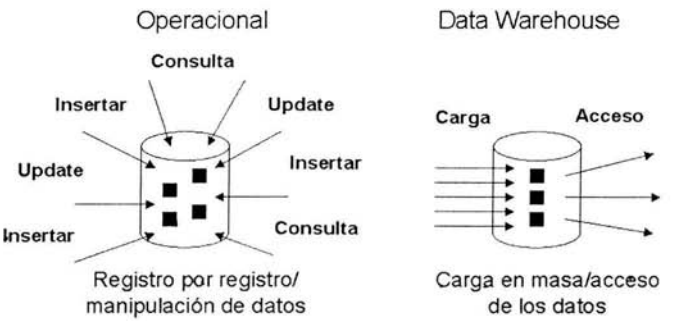

Figura 2.4 No volatilidad de los datos (Inmon, 1996)

Por ejemplo, el estatus de una orden de compra puede cambiar varias veces antes de poder ser liberada en el sistema. Durante el tiempo que la información se encuentre en la base de datos es posible que sea actualizada una o más veces.

En un almacén de datos es común que este se encuentre compuesto por una serie de "fotografías" de los datos en un momento en específico. Estas "fotografías" son almacenadas o cargadas en el data warehouse y pueden estar compuestas por miles o millones de registros. Son utilizadas para analizar y comparar unas con otras y

obtener información sobre el comportamiento de los datos en un rango de tiempo determinado. Es importante resaltar que la información cargada al almacén de datos, es modificada en raras ocasiones. Es muy difícil, sino es que imposible, mantener los datos de manera dinámica en un data warehouse.

2.1.2.4 Variación con el tiempo

La última característica es la no variación de la información en el tiempo. En los sistemas transaccionales generalmente se almacena información que ha sido generada entre un período de tiempo de 60 a 90 días aproximadamente. Cuando la información deja de estar activa, ésta es respaldada y eliminada de la base de datos.

Una de las razones por las cuales la información no se puede mantener demasiado tiempo en un sistema en línea, es el alto precio que se debe pagar en cuanto a performance para mantener estos datos productivos. Si tomamos en cuenta que el performance es una de las prioridades de este tipo de sistemas, es posible llegar a la conclusión de que no vale la pena sacrificar el performance por información que probablemente ya no es necesaria en la operación del mismo.

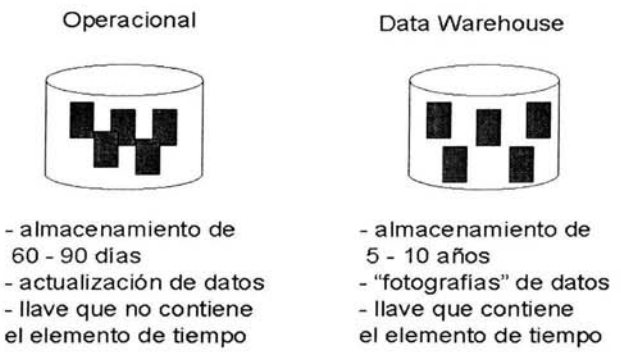

Figura 2.5 Variación con el tiempo (Inmon, 1996)

Sin embargo, un data warehouse debe contener datos que abarquen un período de tiempo de entre 5 a 10 años para cumplir con el requerimiento de los usuarios de información histórica es por ésto que la importancia del factor tiempo es muy importante (*figura 2.5*). En comparación con los sistemas operacionales, el costo de mantener información durante 5 o 6 años en un data warehouse es mínima. La caída de los precios del hardware ha ayudado a la expansión y éxito de los proyectos para la implantación de un data warehouse.

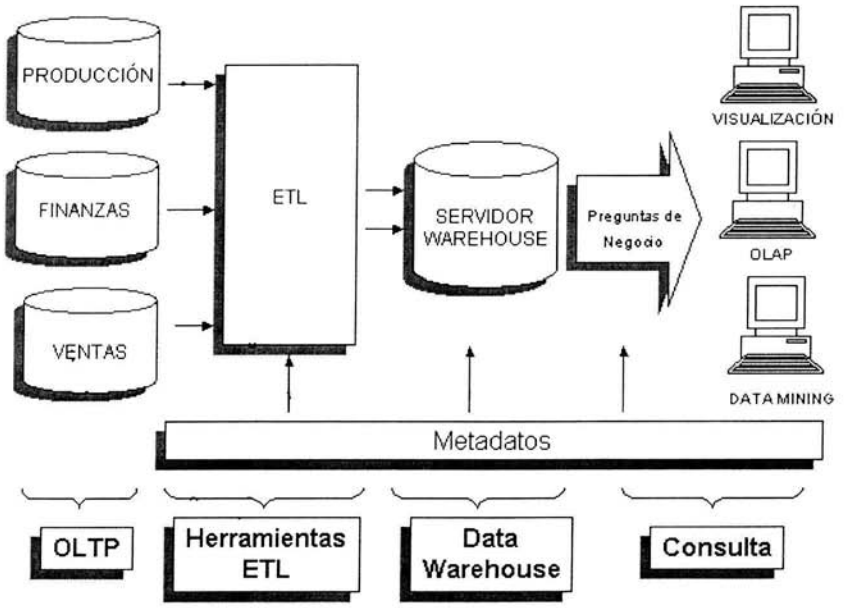

Figura 2.6 Elementos básicos en un data warehouse (Informix, 1999)

2.2 Elementos básicos de un data warehouse

Un data warehouse está compuesto básicamente por los elementos (*figura 2.6*), encargados de llevar a cabo el proceso que comienza en el momento de extracción

de la información de los sistemas transaccionales hasta el momento de consulta y análisis de la información por parte de los tomadores de decisiones.

Un data warehouse (*figura 2.6*) se encuentra compuesto de los siguientes elementos:

- Datos.
- Metadatos.
- Proceso de extracción, transformación y carga (Extraction, Transformatíon Loading).
- Data Warehouse
- Herramientas de consulta.

2.2.1 Datos

Los datos utilizados en el data warehouse provienen de los sistemas transacciones u operacionales, de archivos planos, información impresa en papel o de información contenida en otros almacenes de datos.

El camino que recorre esta información comienza desde la creación de una transacción, sigue con el procesamiento de la misma por el sistema OLTP y en el caso de ser de utilidad para el data warehouse es extraído, transformado si fuera necesario y cargado a la base de datos para ser utilizado como información en el análisis para la toma de decisiones.

2.2.2 Metadatos

El manejo de la infomación de la empresa es tan importante como la información misma. Los metadatos son parte integral de un data warehouse. Sin éstos, la información es un simple repositorio de datos sin significado alguno. Mientras un simple número *5.32* no tiene significado alguno, *5.32%* significa un poco más y

5.32% de auditorio tiene un significado mas preciso para el usuario. Los metadatos son los datos acerca de los datos y son definidos de la siguiente manera: "Los **metadatos** actúan como un directorio de información, mostrando la localización de los datos de tal manera que el acceso en el data warehouse sea sencillo" (SCN Education, 2001).

Los metadatos proveen a los administradores y usuarios de la descripción de los datos u objetos de información que pueden accesar. Existen dos tipos de metadatos:

1. **Metadatos técnicos.** Utilizados por los administradores y herramientas de software y proveen de la descripción técnica de los datos y operaciones.
2. **Metadatos de negocio.** Utilizados por los analistas de negocio y usuarios finales y proveen de la descripción del negocio.

Ambos son importantes en la construcción, mantenimiento y uso del data warehouse. La información que típicamente se almacena en los metadatos son:

- La estructura de los datos en el procesamiento transaccional.
- La estructura de los datos en el data warehouse.
- La información sobre la extracción de los datos.
- Los datos que alimentan al almacén de datos.
- La información para el proceso de transformación de los datos.

2.2.3 Proceso de extracción, transformación y carga

También conocido como proceso ETL (Extraction Transformation Loading) consiste en la extracción de los datos de las fuentes de origen, la transformación de la información en caso de ser necesario y su carga al data warehouse.

2.2.3.1 Extracción de los datos

Para elaborar un plan de extracción, es necesario que el área de data warehouse llegue a un acuerdo con el área dueña de la información para determinar la información y la forma adecuada de obtener los datos requeridos en base a las políticas de seguridad, los métodos de extracción y la frecuencia en la que se requieren los datos con el fin de garantizar la seguridad y calidad de los mismos. Por ejemplo, se debe determinar si es posible accesar directamente a la base de datos transaccional y extraer la información mediante herramientas ETL o si la información será preparada y enviada por el área que la posee cada determinado tiempo en archivos planos que presenten ciertas características determinadas por el área de data warehouse.

2.2.3.2 Transformación de los datos

La transformación de los datos, es el proceso donde la información es filtrada, combinada, transformada y preparada para convertirse en un recurso utilizable en el data warehouse. Este proceso puede abarcar la validación de los datos, estandarización de formatos o unidades, etc. Esta fase en el proceso de data warehousing es de gran importancia ya que puede ser determinante en el aseguramiento de la **calidad de los datos.** Garantizar la confiabilidad de la información es uno de los objetivos primordiales de un almacén de datos.

2.2.3.3 Carga de los datos

Durante el proceso de carga, la información no podrá ser consultada por el usuario dado que todo o parte de la base de datos del data warehouse es puesto fuera de línea. La hora y frecuencia de la carga es programada según sea requerido. En ocasiones este proceso es ejecutado en la noche cuando la información no es consultada por el usuario. El proceso de carga debe ser muy rápido y soportar la

carga de miles o millones de registros a la base de datos y prever los posible problemas que pudieran surgir durante el proceso.

2.2.3.4 Servidor warehouse

Se trata del servidor de base de datos que utiliza la información procesada en la fase de transformación para almacenarla. Cuando un servidor es elegido se deben de tomar en cuenta los siguientes puntos:

- Identificar las ventajas y desventajas del producto.
- Evaluar los requerimientos de performance y escalabilidad.
- El tipo de técnica utilizada para la organización de los datos.

En el servidor, la información se almacena y organiza para ser accesada por los usuarios y elaborar reportes y análisis para la toma de decisiones.

2.2.4 Herramientas de soporte a la toma de decisiones

Es el término utilizado para referirse a las aplicaciones y herramientas del data warehouse que se emplean para recuperar, manipular y analizar los datos. Algunos ejemplos de herramientas de consulta son: herramientas basadas en la tecnología OLAP, el data mining (minería de datos) o reporteadores desarrollados por los analistas para la explotación de la base de datos, etc.

2.2.4.1 Tecnología OLAP

Las herramientas de Procesamiento Analítico en Línea (OLAP, On-Line Analitycal Processing) proporcionan la capacidad de manipulación y análisis de la información almacenada en el data warehouse. OLAP es la tecnología que permite a los usuarios el acceso rápido, fácil e interactivo de los datos organizados de manera

multidimensional y permite a los analistas, gerentes y ejecutivos tener acceso a una gran variedad de vistas de la información que ha sido transformada para reflejar la dimensión real de la empresa.

La tecnología OLAP sigue una lista de 12 reglas presentadas por el Dr. Codd en 1992; entre las que se encuentran el hecho de que presenta una visión multidimensional, transparencia para el usuario en el acceso de los datos y facilidad de acceso de los datos.

Existen diferentes tipos de herramientas OLAP: MOLAP (Multidimensional OLAP) y ROLAP (Relational OLAP). La capacidad de análisis y la interfaz de estas dos arquitecturas son la misma, pero difieren en el almacenamiento físico de los datos. MOLAP almacena los datos en una base de datos multidimensional y ROLAP en una base de datos relacional.

2.2.4.2 Minería de datos

Minería de datos es una tecnología que aplica sofisticados y complejos algoritmos para el análisis de datos y expone información interesante para el análisis en la toma de decisiones. Mientras que OLAP organiza los datos para la exploración de la información mediante el análisis, el data mining lleva a cabo el análisis de los datos y provee de resultados a los tomadores de decisiones.

2.3 OLTP vs Data warehouse

Los sistemas de data warehouse presentan diferencias importantes si son comparados con los sistemas operacionales o transaccionales (On-line transaccional processing).

2.3.1 OLTP (On-line transaction process)

Un sistema operacional o transaccional se define como "la aplicación utilizada para correr los negocios día a día usando datos en tiempo real" (Devlin, 1997). Estos sistemas son utilizados en la operación de la empresa y constituyen el motor que mantiene funcionando a la organización. Ninguna operación debe impedir que se realice el proceso en un sistema OLTP dado que estos sistemas no pueden parar, si se detuvieran, la organización resultaría afectada.

Estos sistemas son las principales fuentes de datos de un data warehouse ya que en ellos se registran diariamente las transacciones de una empresa. Algunos ejemplos de un sistema operacional o transaccional es el sistema de facturación o el sistema de atención a clientes.

2.3.2 Diferencias entre un OLTP y un Data Warehouse

Ralph Kimpball (1996) menciona algunas diferencias entre estos dos sistemas:

- En la concurrencia.
- En el tipo de transacción.
- En el tipo de usuario.
- En el uso del elemento tiempo.
- En el modelo de datos utilizado.

2.3.3 Diferencias en la concurrencia

Concurrencia quiere decir que dos usuarios que consultan al mismo tiempo la misma información obtienen los mismos resultados.

En un OLTP la concurrencia es microscópica, esto quiere decir que un gran número de pequeñas transacciones se deben procesar correctamente una a una. En la

figura 2.7 es posible observar que todas las transacciones enviadas al sistema son procesadas sin perder ninguna.

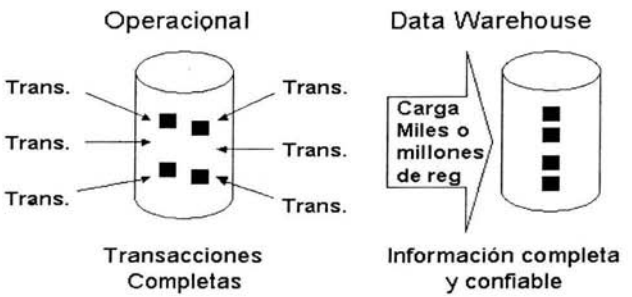

Figura 2.7 Concurrencia de los datos (Inmon, 1996)

En el data warehouse la concurrencia se mide de manera global, ya que la información cargada en la base de datos está formada por datos completos y confiables. Por lo tanto, no se presta importancia a una transacción individual sino a la **calidad** global de la información.

2.3.4 Diferencias en el tipo de transacción

Un sistema de transacción en línea generalmente procesa miles, o incluso, millones de transacciones por día (*figura 2.7*). Cada transacción enviada a la base de datos constituye una pequeña pieza de información. En el data warehouse se lleva a cabo una sola "transacción" al día compuesta por miles o millones de registros, a esta operación se le conoce con el nombre de **carga de datos productivos**. Si el proceso de carga fuera interrumpido por alguna razón, no se carga ningún dato que

compone la "fotografía" y es necesario volver a enviar la carga completa, esto garantiza que toda la información sea cargada correctamente.

2.3.5 Diferencias en el tipo de usuario

Los usuarios de un sistema OLTP son la maquinaria que hace caminar a la organización dado que actualizan la información existente, crean datos nuevos, toman órdenes, dan de alta clientes, etc.; generalmente repiten una operación una y otra vez, por lo cual un buen performance es la prioridad en un sistema transaccional. Las consultas elaboradas en un sistema transaccional casi siempre requieren de un número pequeño de registros para obtener la respuesta deseada.

Usuario de un OLTP	Usuario de un data warehouse
Información detallada	Acumulados, información refinada
Información actualizada	Información histórica
Actualización de los datos	No debe existir actualización de los datos
Información orientada a la aplicación	Información orientada al análisis
Importancia de respuesta de la transacción	Importancia de la respuesta masiva
Manejo de pequeñas cantidades de datos	Manejo de grandes cantidades de datos
Manejo de transacciones	Manejo de consultas y análisis

Figura 2.8 Necesidades de los usuarios de los sistemas OLTP y data warehouse (Inmon,1996)

Los usuarios de un data warehouse utilizan los datos que generan los sistemas transaccionales. Los reportes generados por un data warehouse comprimen miles o millones de datos en una pequeña respuesta. Los ejecutivos y el personal operativo consideran los datos agregados y condensados con mayor valor ya que proporcionan mayor información que la consulta por separado de cada una de las transacciones.

En el data warehouse el performance es importante pero no determinante. Las consultas a una tabla deben tener una respuesta casi instantánea y las consultas a más de una tabla deben tardar un par de segundos o minutos. La actividad principal de un data warehouse es la elaboración de reportes.

En la *figura 2.8* se presentan las diferencias entre las necesidades de un usuario de un sistema transaccional y un usuario de un data warehouse.

2.3.6 Diferencias en el uso del elemento tiempo

Los sistemas transaccionales y de data warehouse tratan el elemento del tiempo de manera muy distinta. Un buen sistema de data warehouse es un conjunto de "fotografías" instantáneas de momentos específicos de la organización. En un data warehouse preguntas como: *¿Qué tanto aumentaron o disminuyeron las ventas en el último mes? ¿Cuántos productos de cierto tipo fueron vendidos el último año?* son contestadas frecuentemente y en ocasiones requieren de información histórica que los sistemas transaccionales no pueden proporcionar debido a que almacenan información de un período de tiempo limitado y en cualquier momento puede ser actualizada.

2.3.7 Diferencias en el modelo de datos

La última diferencia y la más importante entre los sistemas OLTP y data warehouse es la organización de los datos. Para comprender porqué los datos están

organizados de diferente manera, es necesario tomar en cuenta el performance de los sistemas transaccionales.

En gran parte el buen performance que presenta una base de datos de este tipo se debe al modelo entidad-relación aplicado en el diseño. Este tipo de modelado busca eliminar todas las redundancias existentes en los datos. Si no existen redundancias las transacciones que modifican, agregan o borran los datos únicamente necesitan llevar a cabo estas operaciones en un solo lugar. A ésto se debe la rapidez en el procesamiento de transacciones que aportó el modelo relacional.

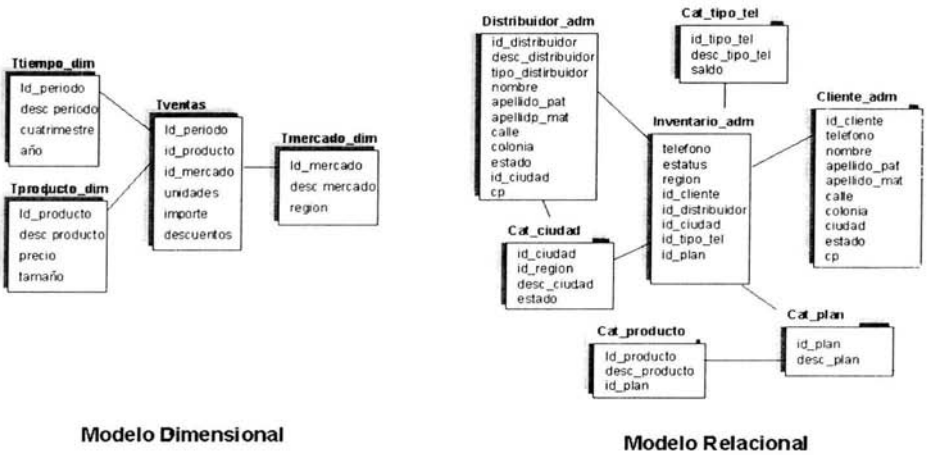

Figura 2.9 El modelo dimensional y relacional

Uno de los modelos utilizados para el data warehousing es el modelo dimensional o esquema estrella (*figura 2.9)*, donde las tablas pueden formar una especie de estrella. A diferencia del modelo relacional (*figura 2.9*), el modelo dimensional es asimétrico y permite una consulta rápida de la información y una visión de los datos acorde al contexto de negocio.

OLTP	Data Warehouse
Predominio de la actualización	Predominio de la consulta
Actividad de tipo operativo	Actividad de análisis y decisión estratégica
Proceso puntual	Proceso masivo
Datos en general desgregados	Datos en distintos niveles de detalle y agregación
Dato actual	Dato histórico
Los datos entran rápidamente	Los datos salen rápidamente
Importancia de respuesta de la transacción	Importancia de la respuesta masiva
Modelo relacional	Modelo dimensional
Usuarios de perfiles medios o bajos	Usuarios de perfiles altos
Explotación de información relacionada con la operación	Explotación de toda la información interna y externa relacionada con el negocio

Figura 2.10 Diferencias entre los sistemas OLTP y Data Warehouse

En la *figura 2.10* se presenta un resumen de las diferencias encontradas entre los sistemas transaccionales en línea y los sistemas de data warehouse.

2.4 Modelo dimensional

Las empresas de hoy en día operan en una economía global con competidores globales, necesitan buscar mercados en donde sus productos y servicios tengan ventajas competitivas y sean diferentes. Un requerimiento fundamental es buscar nuevas oportunidades de mercados y crear programas de comercialización detallados. Para lograr ésto, se utiliza el análisis multidimensional para el soporte a la toma de decisiones.

Los datos empresariales se encuentran relacionados y regularmente son jerárquicos: por ejemplo, los datos de ventas, los datos de inventario y los pronósticos de presupuesto están interrelacionados, dependen unos de otros. Tanto para la eficiencia operativa como para la planeación a futuro, se debe elaborar un análisis de estos datos. Esta necesidad empresarial se aborda mediante el procesamiento analítico.

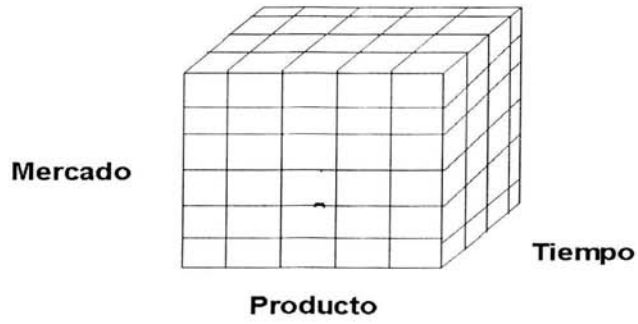

Figura 2.11 Dimensiones de un negocio (Informix, 2000)

Para facilitar un análisis complejo, el procesamiento analítico o análisis multidimensional presenta una visión empresarial sencilla de los datos. El proceso de negocio es visualizado como un cubo cuyos resultados se pueden pivotear o girar para cambiar los ejes y la perspectiva de los mismos (*figura 2.11*).

Por ejemplo, suponiendo que se tiene un negocio de ventas de diferentes productos y éste cuenta con sucursales o mercados en diferentes lugares de la ciudad, se requiere evaluar el comportamiento del negocio en un tiempo determinado. Este negocio es fácil imaginarlo como un cubo de datos que contiene **dimensiones** de tiempo, productos y mercados (*figura 2.11)*. Las distintas intersecciones entre las líneas del cubo contiene las **métricas** del negocio, que corresponden a una combinación en particular de producto, mercado y tiempo. Es decir, el número de productos de cierto tipo vendidos en un mercado determinado en un momento en el tiempo (días, meses, bimestres, etc.).

Para obtener una base de datos que permita este tipo de análisis es utilizado el modelo dimensional que es una técnica de modelado de base de datos. Ralph Kimpball define el modelo dimensional de la siguiente manera: "Es el nuevo nombre de una vieja técnica para hacer bases de datos simples y entendibles. Cuando una base de datos puede ser visualizada como un **cubo** de tres, cuatro o incluso más dimensiones, las personas pueden imaginar mover estos cubos a lo largo de cada una de sus dimensiones" (Kimpball, 1996).

Este modelo, es muy **sencillo** y **fácil de comprender** por el usuario. Además, permite elaborar preguntas de negocio sencillas pero importantes para la planeación estratégica de una organización. *¿Qué cantidad de producto X fue vendida en el mercado durante los últimos 6 meses?* o *¿Qué producto fue el más vendido y cuál fue el menos vendido en los mercados que se encuentran en la parte norte y sur de la ciudad?*.

La variedad de preguntas puede ser enorme, desde diferentes puntos de vista del negocio y proporciona información muy importante. La sencillez de este modelo es su principal cualidad y ventaja sobre otro tipo de modelos como el modelo relacional.

Otro nombre con el cual es conocido el modelo dimensional es el **esquema estrella** (*figura 2.12*). Los diseñadores de bases de datos utilizan este nombre porque el diagrama para este modelo parece formar una estrella con una tabla en el centro y un conjunto de tablas a su alrededor relacionadas a esta. La tabla central es la única tabla en el esquema con múltiples uniones conectándola a las tablas restantes. Esta tabla central es llamada **tabla de hechos** (contiene las métricas) y las tablas que la rodean **tablas de dimensiones** (contiene las dimensiones del negocio). Las tablas de dimensión tienen una sola unión con la tabla de hechos.

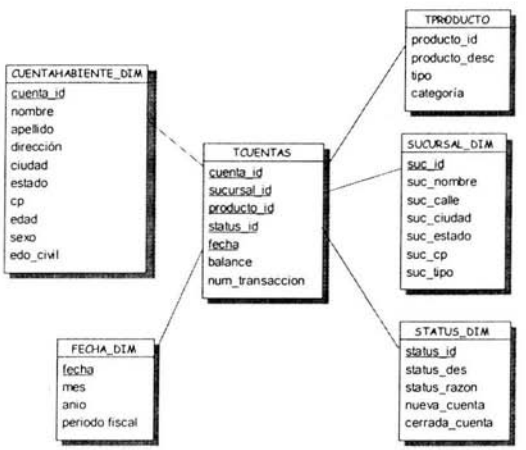

Figura 2.12 Un típico modelo dimensional (Informix, 2000)

2.4.1 Tabla de Hechos

La tabla de hechos almacena las métricas de un negocio, cada una de estas métricas son las intersecciones de las dimensiones (*figura 2.13)* y la llave de esta tabla representa el nivel más bajo de cada tabla de dimensión. Las métricas son cuantitativas o son hechos que el sujeto llevó a cabo, generalmente son numéricas y responden a las preguntas **¿cuántos?** o **¿qué tantos?**. Por ejemplo, algunas

métricas son el número de productos vendidos o los productos registrados en el inventario.

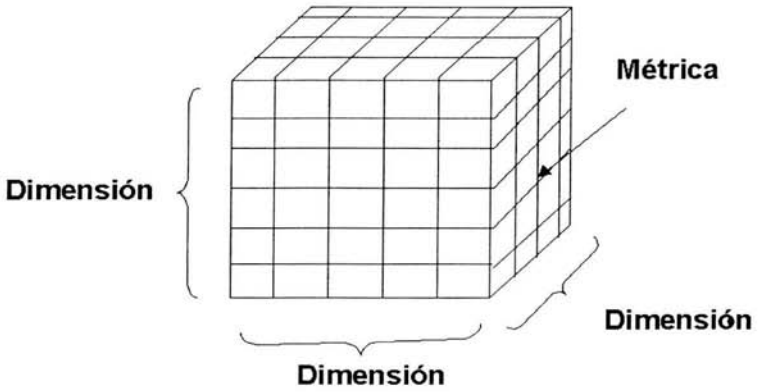

Figura 2.13 Principales elementos del modelo dimensional: dimensiones y métricas o hechos (Informix, 2000)

Como fue comentado anteriormente, la llave de la tabla de hechos está compuesto por los campos que representan el nivel más bajo de detalle en las tablas de dimensiones, a ésto también se le conoce como **granularidad**.

La granularidad puede ser una transacción individual, una "fotografía" diaria, una "fotografía" mensual, etc. El nivel de detalle o granularidad de la tabla de hechos es determinante ya que dependiendo de ésta se define el tamaño y crecimiento de la tabla de hechos.

2.4.2 Dimensiones del modelo

Una dimensión representa un solo conjunto de objetos o eventos en el mundo real. Las dimensiones son las categorías que hacen de las métricas de la tabla de hechos un dato con significado porque contestan el **¿qué?**, **¿cuándo?** y **¿dónde?** de la

pregunta elaborada por el tomador de decisiones. Cada dimensión identificada para el modelo de datos queda implementado como una tabla de dimensiones.

En las preguntas: ¿Cuántos *productos* fueron los más vendidos el último *año*? O ¿Cuál fue la ganancia por *vendedor*? Las dimensiones se identifican con letras itálicas, incluso es posible apreciar que en cada pregunta puede estar implicada una o más dimensiones.

2.4.2.1 Tablas de Dimensión

Una tabla de dimensión es una tabla que almacena la descripción de las dimensiones del negocio. Una tabla de dimensiones contiene **elementos de dimensión** y **atributos de dimensión**, que se encuentran en lo que se conoce como **jerarquías**. El nivel de jerarquía se refiere a los atributos de una tabla que tienen una relación de uno a muchos, por ejemplo, en la *figura 2.14* se muestra una relación de uno a muchos, una región puede contener varios estados y un estado a su vez puede tener distintas ciudades, cada uno de estos atributos se encuentran en un nivel de jerarquía. El nivel más bajo de detalle requerido para el análisis de datos determina el nivel más bajo de jerarquía.

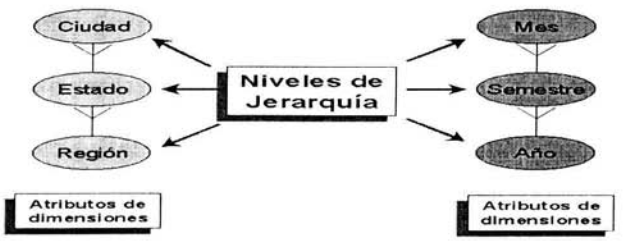

Figura 2.14 Jerarquías en una dimensión (Informix, 2000)

2.4.2.2 Elementos de dimensiones

Cada dimensión puede definir múltiples elementos de dimensión para diferentes niveles de sumarización. Debido al aspecto jerárquico de las dimensiones, los usuarios pueden elaborar consultas que accesan los datos en un nivel alto o resumido (drill up) o desde un nivel bajo o de detalle (drill down). Por ejemplo, la jerarquía de la *figura 2.14* está formado por *región*, *estado* y *ciudad*. Una ciudad está contenida en un estado, el cual a su vez está contenido en una región. Por ejemplo, si el usuario quisiera consultar todas las regiones podría ver esta información filtrando por el atributo región y si quisiera consultar con mayor detalle lo podría hacer por estado.

Los elementos de dimensión son generalmente almacenados en la base de datos como códigos numéricos o pequeñas cadenas de caracteres que faciliten uniones con otras tablas. Cada elemento de dimensión puede definir múltiples atributos de dimensión, del mismo modo que las dimensiones pueden definir múltiples elementos de dimensiones.

2.4.2.3 Atributos de dimensión

Un atributo de dimensión (*figura 2.15*) es una columna en una tabla de dimensiones donde cada atributo describe un nivel de sumarización o elemento de dimensión. Los elementos de dimensión definen las relaciones jerárquicas en una tabla de dimensiones. Los atributos describen elementos de dimensión en términos que son familiares para los usuarios.

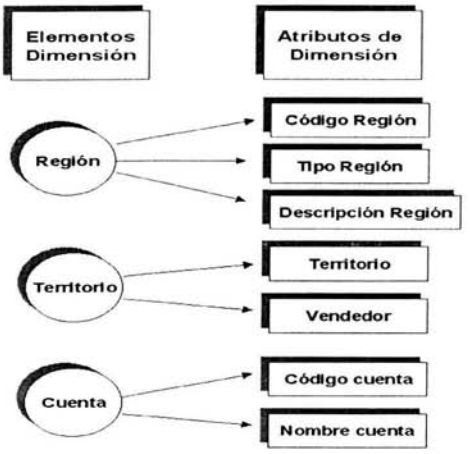

Figura 2.15 Ejemplo de los atributos de dimensión (Informix, 2000)

De esta manera se concluye la introducción al data warehouse. En este capítulo se dio una definición de lo que es un data warehouse, sus características, los elementos que lo componen y el porqué la necesidad de crear sistemas de este tipo. Se dio una explicación breve del modelo dimensional y sus elementos. La comprensión de los términos relacionados con este tema es fundamental para la correcta comprensión del siguiente capítulo.

CAPÍTULO III
DISEÑO LÓGICO DE UNA BASE DE DATOS DIMENSIONAL

En este capítulo se presentan los elementos, las fases de diseño y una metodología para el diseño lógico de una base de datos para un data warehouse. Así como un caso práctico donde se aplica la metodología presentada para modelar el esquema estrella del sistema de tráfico de llamadas de una empresa de telefonía móvil.

3.1 Diseño de una base de datos dimensional

Como en todo proceso de construcción de un sistema, en un proyecto de creación de un data warehouse se sigue un ciclo de vida que pasa por un proceso de planeación, definición de requerimientos, análisis, diseño, desarrollo, pruebas e implantación del sistema.

Sin embargo, el diseño de un data warehouse difiere al de un sistema tradicional. La diferencia radica en que el **diseño de un data warehouse** se basa en los **conceptos de negocio**, mientras que el **diseño de un sistema transaccional** gira alrededor de los **procesos operacionales**. Un sistema de data warehouse requiere de ciertas características, como consultas rápidas de los datos o almacenamiento de información histórica, las cuales deben ser tomadas en cuenta por el diseñador.

3.2 Elementos del diseño de un data warehouse

Antes de comenzar con el proceso de diseño, es importante comprender los elementos que intervienen e influyen durante este proceso. Estos elementos son: la **arquitectura**, cuyas metas deben de ser claras antes de comenzar con el proceso

de diseño; el **usuario**, para el cual será creado el data warehouse; y la **información**, cuyas características deben ser conocidas antes de comenzar el proyecto.

3.2.1 Arquitectura

Un data warehouse, tiene como meta principal proporcionar información para el soporte a la toma de decisiones, por lo tanto debe ser diseñado para satisfacer los siguientes requerimientos:

- Proporcionar experiencia de los usuarios .
- Funcionar sin interferir con los sistemas transaccionales.
- Proveer de un repositorio central de información consistente.
- Responder a consultas complejas rápidamente.
- Proveer de una variedad de poderosas herramientas para el análisis.

3.2.2 Usuarios

El éxito de un data warehouse puede ser medido en base a la aceptación que tenga por parte de los usuarios que lo utilizan. El buen diseño radica en comprender al usuario y sus necesidades. Los usuarios de este tipo de sistemas necesitan responder a preguntas como *¿cuántas reservaciones se han llevado a cabo en el último año?, ¿cuál es el porcentaje de incremento o decremento de clientes del semestre actual comparado con el del último semestre?*, etc. Las diferentes vistas de negocio que el usuario requiere consultar pueden variar según la información que necesite.

3.2.3 Consulta de información

Es importante que la información consultada en el almacén de datos sea oportuna y confiable en cuanto a la calidad de los datos y reflejen la realidad lo más fielmente posible. Las consultas se pueden efectuar a través de herramientas que faciliten el acceso a la información de manera más eficiente o permitan elaborar análisis complejos según las necesidades de los usuarios. Generalmente son utilizadas herramientas especiales de análisis aunque también pueden ser creadas aplicaciones que generen reportes requeridos frecuentemente.

3.3 Fases del diseño de un data warehouse

Una vez mencionados los elementos que intervienen en el diseño de un data warehouse, a continuación se listan las tres fases que componen este proceso:

- Diseño Conceptual
- Diseño Lógico
- Diseño Físico

En cada una de estas fases se siguen una serie de pasos que permiten un diseño adecuado para lograr las metas definidas. Sin embargo, hay que tomar en cuenta que la construcción del data warehouse es un proceso cíclico que requiere revisar varias veces el diseño antes de obtener el definitivo.

En cada fase es importante tomar en cuenta los diferentes elementos que influyen en el proceso de diseño presentados anteriormente y que varían en las diferentes fases. Estos elementos son los usuarios, las consultas y la arquitectura (*figura 3.1*).

Otro de los elementos que influyen en el proceso de diseño y que no ha sido mencionado, son las **limitaciones** a considerar como podrían ser el presupuesto

disponible para el proyecto, el nivel de conocimiento de los participantes o el plan de trabajo establecido.

	Conceptual	Lógico	Físico
Usuarios	Alta	Media	Baja
Consultas	Alta	Media	Baja
Arquitectura	Baja	Alta	Alta

Figura 3.1 Elementos que influyen en las fases de diseño de un data watehouse y su importancia

3.3.1 Diseño Conceptual

La primera fase del proceso de diseño es el diseño conceptual. En esta fase se hace la **recolección de información** acerca del proyecto. Son elaboradas una serie de entrevistas con los usuarios finales las cuales introducen a los diseñadores a las necesidades y expectativas de los usuarios. En esta parte del diseño deben ser contestadas algunas preguntas como: *¿quiénes son los usuarios que utilizarán el sistema?*, *¿qué habilidades poseen estos usuarios?*, *¿cuáles son los lugares, personas y cosas que intervienen en el proceso de negocio?*, *¿existen fuentes que proporcionan la información que se está solicitando?*, *¿qué software debe ser utilizado o cuáles son las posibles limitaciones de este?* ,etc.

Estas entrevistas y preguntas permiten al diseñador elaborar una lista de requerimientos de los usuarios, de recursos materiales y humanos necesarios para el proyecto. Se definen los recursos de datos y estructuras así como la frecuencia en la cual será utilizada la información.

Esta fase es de gran importancia ya que son identificadas las necesidades del usuario, son definidos los objetivos del proyecto, sus alcances y permite al diseñador conocer el nivel de información que almacenará el futuro data warehouse, ajustarlo y

corregirlo de acuerdo a los requerimientos del usuario. Es importante mencionar que el diseñador también debe tener cuidado de que el usuario no se haga expectativas que no serán cubiertas por el almacén de datos, deben quedar claras las limitaciones y alcances del proyecto. La importancia del usuario y la forma en que se va a consultar la información que se requiere es primordial y ayuda a definir el objetivo del proyecto (*figura 3.1*).

3.3.2 Diseño Lógico

Durante esta fase, el modelo conceptual obtenido se presenta en un esquema de base de datos. En esta fase un equipo de diseñadores conformado por analistas, administradores de bases de datos, diseñadores de aplicaciones, personal de soporte, etc. forman un equipo de diseño que se dedica a identificar los procesos de negocio a modelar, elegir las tablas de hechos y su nivel de detalle, identificar las tablas de dimensiones, etc. Se obtiene el esquema estrella que representa el proceso de negocio, la importancia del usuario y la consulta de información tienen una prioridad media y la arquitectura toma mayor importancia (*figura 3.1*).

3.3.3 Diseño Físico

Durante esta fase un grupo de diseño, debe tomar decisiones sobre el almacenamiento físico de la información. Se determina el lugar físico de los datos y los índices en los filesystems. Se crean las tablas, se elaboran los procesos de extracción y transformación de los datos, se cargan las tablas, etc. Es elegida una estrategia de indexado apropiada al manejador de la base de datos y son aplicados los principios de segmentación y fragmentación de datos; se evalúan los requerimientos para el almacenamiento de datos e índices y se elabora una estrategia para el respaldo de la información. La importancia de la arquitectura en esta fase toma fuerza y la del usuario y la consulta de información pasa a segundo plano (*figura 3.1*).

3.4 Metodología a seguir para el diseño de un data warehouse

La construcción de un data warehouse es el proceso que empata las necesidades de los usuarios con la información disponible. El diseñador se debe apoyar en una metodología para elaborar las preguntas correctas que le permitan elegir un diseño adecuado a las necesidades de la empresa, de lo contrario, corre el riesgo de obtener una base de datos que no cubra todas las necesidades de los usuarios.

Ralph Kimpball (1996) presenta una metodología para el diseño de una base de datos dimensional de un data warehouse conformada por una serie de pasos a seguir para obtener un esquema adecuado. Esta metodología es de tipo arriba-bajo porque comienza con la identificación de los procesos macros de la compañía y termina con la definición de la frecuencia con la que serán extraídos los datos, es decir, va de lo general a lo particular.

Cabe mencionar que esta metodología **es una guía para el diseñador y no una receta de cocina** que deberá ser seguida al pie de la letra. El diseño de un data warehouse puede variar según el proceso de negocio a modelar. El diseñador deberá usar su creatividad, ingenio y experiencia, en el caso de tenerla, para lograr un buen modelo.

Esta metodología abarca el diseño conceptual, lógico y físico de un data warehouse, sin embargo, este trabajo únicamente está enfocado al diseño lógico, por lo tanto solo serán tratados los cinco puntos que conforman esta fase del diseño.

3.4.1 Proceso de diseño de un data warehouse

La metodología presentada por Ralph Kimpball (*figura 3.2*) comienza con una serie de entrevistas con los usuarios finales (diseño conceptual). Después de haber recolectado la información necesaria se procede a identificar el proceso de negocio a modelar y se definen los hechos del modelo. Antes de diseñar cualquier tabla de

hechos se debe tomar la decisión de cual será el nivel más bajo de detalle o granularidad para posteriormente proceder a identificar las dimensiones del modelo. La elección de las dimensiones son la llave del diseño ya que permitirán consultar la información de la tabla de hechos desde diferentes perspectivas y tan detallada o resumida como se requiera.

Fase del diseño	Pasos de la metodología
Diseño Conceptual	1. Recolección de información y elaboración de los requerimientos del usuario
Diseño Lógico	2. Elección del proceso a modelar
	3. Definición del nivel de detalle o granularidad de cada tabla de hechos
	4. Identificación de las dimensiones de cada tabla de hechos
	5. Definición de las tablas de hechos, incluyendo las tablas de hechos precalculadas.
	6. Definición de los atributos de las dimensiones con una descripción completa y la terminología apropiada.
Diseño Físico	7. Obtener dimensiones cambiantes
	8. Definición de agregados, dimensiones heterogéneas, minidimensiones, consultas y otras decisiones sobre el almacenamiento físico
	9. Definición de la duración histórica de la base de datos
	10. Definición de la frecuencia con la cual será extraída y cargada la información en el data warehouse.

Figura 3.2 Metodología para el diseño lógico de un data warehouse (Kimpball, 1996)

Una vez que las dimensiones se han creado, se procede a elegir las métricas que contendrá la tabla de hechos y a diseñarla a detalle. Finalmente las tablas de dimensiones son completadas agregando sus atributos y descripciones completas.

Hasta este punto la metodología de Ralph Kimpball considera que el diseño lógico ha sido completado y comienza el diseño físico que incluye la obtención de dimensiones

cambiantes, tablas de agregados, dimensiones heterogéneas, minidimensiones, y formas de consulta. Finalmente, el diseñador puede planear la duración de la base de datos y definir la frecuencia con la que será extraída y cargada la información al data warehouse.

3.4.1.1 Elección del proceso a modelar

Una vez recolectada la información, el primer paso en el diseño lógico es identificar el proceso de negocio a modelar. Un **proceso de negocio** se define como "el proceso primario operacional de una organización" (Kimpball, 1996). El proceso de negocio identifica lo que el usuario hace con los datos, de donde vienen y cómo se transforman para crear información con un significado. La información puede provenir de muchas fuentes como las bases de datos de finanzas, de ventas, de distribución, de otro data warehouse, etc. Algunos ejemplos de proceso de negocio son: ventas, distribución, inventario o facturación.

Para elegir el proceso a modelar deberán ser tomados en cuenta los siguientes puntos:

- Llevar a cabo un análisis cuidadoso de la información obtenida en el diseño conceptual.
- Entender cada proceso de negocio y los datos necesarios para su soporte.
- Determinar la información que el usuario requiere consultar.
- Conocer los reportes obtenidos actualmente y la información disponible.

Esta fase del diseño es importante ya que la identificación del proceso de negocio influirá en el esquema obtenido en el diseño y ayudará a identificar fácilmente los hechos. Para este tipo de sistemas ya existen un conjunto de esquemas para los procesos de negocio más comunes en los que se puede apoyar el diseñador.

Por ejemplo, un esquema clásico del proceso de ventas *(figura 3.3)* está conformado por una tabla de hechos llamada *ventas* y tres tablas de dimensiones, *fecha_dim, producto_dim y mercado_dim.* El usuario podrá hacer consultas de las ventas por región, fecha o mercado, incluso puede variar sus consultas por categoría de producto, o hacer comparativos de ventas por mes.

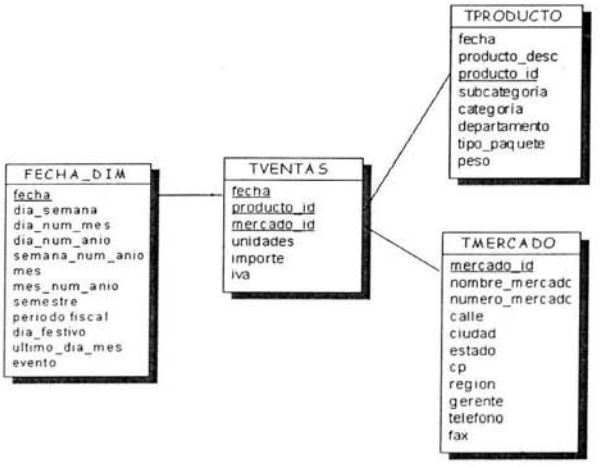

Figura 3.3 Esquema clásico del modelo de un proceso de negocio de ventas

Otro esquema clásico es el financiero (*figura 3.4*), en éste existe una tabla con las métricas *cuentas* y sus respectivas dimensiones *cuentahabientes, sucursal, fecha, producto y estatus.* Con este esquema es consultada la información sobre las cuentas de los clientes de un banco por cuenthabiente, producto o estatus de la cuenta.

3.4.1.2 Definición de la granularidad o del nivel de detalle de cada tabla de hechos

El siguiente paso podría parecer un detalle técnico, sin embargo, es uno de los puntos más importantes para obtener un esquema que cumpla con las

características adecuadas. La granularidad se refiere al nivel de detalle de la información requerida y tiene una relación directa con las actividades de resumen y adición realizadas sobre los datos en las consultas. A menor granularidad, mayor cantidad de detalle. Para incrementar la granularidad (y su utilidad para quienes toman las decisiones), los datos operacionales deben resumirse y acumularse. Por lo regular, entre mayor sea la granularidad, mayor será la cantidad de procesamiento requerido para convertir y resumir los datos operacionales. Al mismo tiempo los datos con alta granularidad requieren de menos volumen de almacenamiento y también se pueden consultar con rapidez y conveniencia. Sin embargo, un data warehouse casi siempre requiere información detallada hasta donde sea posible, no porque sea necesario consultar el detalle, sino porque las consultas muchas veces requieren de información precisa para hacerlas mas flexibles.

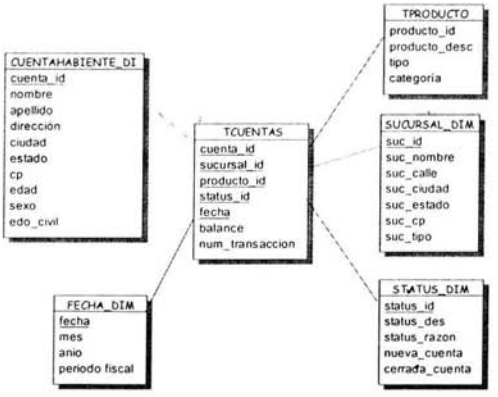

Figura 3.4 Esquema clásico del modelo de un proceso de negocio financiero

Para la elección de la granularidad se debe decidir lo que un registro en el nivel más bajo de la tabla de hechos debe contener o representar. Los componentes que crean la granularidad de una tabla de hechos corresponden directamente a las dimensiones del modelo de datos. Por lo tanto, cuando se define la granularidad de la tabla de hechos, también se identifican las dimensiones del modelo. Los detalles

deben ser consistentes para que exista una buena relación entre tabla de hechos y dimensiones.

La granularidad de una tabla de hechos también es importante en cuanto a que influye en el espacio de almacenamiento requerido para la base de datos. Por ejemplo, si se consideran las siguientes granularidades:

- Producto por día por región.
- Producto por mes por región.

El tamaño de la base de datos que tiene la granularidad de *producto por día por región* podría ser mucho más grande que la base de datos con una granularidad de *producto por mes por región,* porque la base de datos contiene registros para cada transacción hecha cada día en comparación con un resumen de las transacciones de todo el mes. La determinación de la granularidad es de gran importancia, si es demasiado pequeña se puede obtener como resultado una base de datos enorme que se salga del control de las manos del administrador, si sucede lo opuesto, el resultado podría ser información poco detallada para las consultas requeridas.

3.4.1.3 Identificación y conformación de las dimensiones

Una vez que la granularidad de la tabla de hechos ha sido elegida y las dimensiones primarias se identifican, cada llave que conforma la granularidad de la tabla de hechos corresponde a una dimensión. Por ejemplo, si la granularidad de la tabla de hechos fue identificada como *cliente por producto por ciudad por día*, entonces las dimensiones primarias o indispensables pueden ser las dimensiones *cliente, producto, geografía y tiempo* (*figura 3.5*).

Figura 3.5 Dimensiones obtenidas de la granularidad de la tabla de hechos

Como ya había sido mencionado, la tabla de dimensiones contienen las características de los datos almacenados en la tabla de hechos, cada renglón de la tabla es único y son utilizados como filtros para consultar las métricas.

En esta parte del diseño ya se pueden identificar las jerarquías y los elementos de cada dimensión.

Los elementos y las jerarquías deben ser elegidas tomando en cuenta los requerimientos de información identificados en el diseño conceptual, y previendo las consultas que el usuario podría necesitar en un futuro. Los elementos y jerarquías deben ser elegidas con la finalidad de que el data warehouse pueda ser capaz de filtrar los datos hacia arriba (drilling up) o hacia abajo (drilling down). Esto quiere decir que debe permitir consultar la información tan detallada o resumida como el usuario lo requiera. Por ejemplo, el usuario debe tener la facilidad de hacer una consulta de los productos vendidos en cierta región y de hacer la misma consulta pero por ciudad o por mercado.

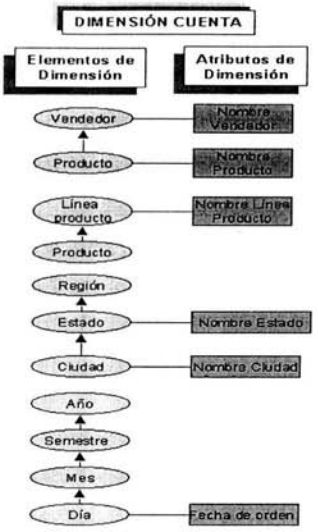

Figura 3.6 Elementos, jerarquías y atributos de una tabla de dimensión

Algunas de las dimensiones que con frecuencia se identifican en un data warehouse son las de tiempo, producto y geografía por mencionar algunas.

3.4.1.4 Definición de la tabla de hechos incluyendo las tablas de hechos precalculadas

Una vez identificado el proceso a modelar y las dimensiones, el cuarto paso en el diseño es hacer un cuidadoso análisis para determinar cuales son las métricas que debe contener la tabla de hechos.

Las métricas son accesadas a través de las dimensiones y de preferencia estas deben estar conformadas por valores numéricos que faciliten el concentrado de la información en pequeñas respuestas, sin embargo, no siempre presenta este tipo de valores. La información contenida en las tablas de hechos se clasifican en tres tipos según los datos que conforman el dominio de las métricas:

1. **Aditivas:** Aquellas tablas de hechos donde las métricas presentan valores numéricos que pueden ser sumados a través de todas las tablas de dimensiones, por ejemplo, el número de productos vendidos presentado en un modelo de ventas (*figura 3.3)*.

2. **Semiaditivas:** Aquellas tablas de hechos donde sólo algunas de las métricas disponibles pueden ser sumarizadas y otras no, por ejemplo el balance mensual del cliente de un banco presentado en un modelo financiero (*figura 3.4*). Este tipo de información no es obtenido mediante una suma, se debe obtener el promedio del balance del cliente en un mes.

3. **No aditivas:** Ninguno de los hechos pueden ser sumarizados. Este caso se puede presentar cuando una tabla contiene información sobre eventos. Por ejemplo, en un modelo de inventario se puede almacenar el estatus del producto (disponible, reservado o vendido), estos valores no pueden ser sumarizados, sin embargo, proporciona información sobre el estado del producto en diferentes momentos.

Las métricas de los procesos de negocio de una empresa son representadas en uno o más grupos de tablas de hechos. Estas comparten las dimensiones que tengan en común representando la realidad lo más fielmente posible.

Como ya se había sido mencionado, el diseño de una base de datos es un proceso cíclico y el diseñador puede detectar algunos ajustes necesarios de la granularidad o número de dimensiones.

Otro elemento que se debe tomar en cuenta en esta fase del diseño, es la creación de tablas de hechos precalculadas. En ocasiones, la tabla de hechos alcanza un tamaño considerable o la consulta requiere de cálculos que pudieran afectar el

performance de la base de datos. En estos casos, una alternativa que se presenta es la creación de tablas precalculadas donde se almacenan acumulados de las tablas de hechos que permiten agilizar la consulta de la información.

En cuanto a la normalización de los datos, en el modelo relacional la información pasa por un proceso de eliminación de redundancias con la finalidad de optimizar el performance y asegurar la integridad de los datos. En el modelo dimensional la normalización de las tablas es evitada debido a las siguientes razones:

- El modelo dimensional es la representación de la correlación existente entre las dimensiones y los hechos del negocio, por lo tanto el modelo debe ser lo más apegada a la realidad y plasmar la relación compleja que existe entre los elementos que conforman al modelo y que presentan una normalización natural.

- La existencia de redundancia en los datos ayuda a la consulta rápida y flexible de la información, una normalización impediría la consulta desde diferentes perspectivas y la rapidez de la información sería perjudicada.

- Las tablas de hechos y dimensiones no ocupan mucho espacio. En las tablas de hechos el tipo de dato de las llaves foráneas están formadas por enteros pequeños, y los datos de las métricas están compuestos por datos numéricos que ocupan una pequeña cantidad de bytes. En comparación con las tablas de hechos, las tablas de dimensiones son mucho más pequeñas y no tiene caso considerar la normalización de estas tablas ya que el espacio utilizado es mínimo.

Hasta el momento, han sido identificadas las tablas de hechos y dimensiones, sin embargo, no ha sido mencionada la forma en la que están relacionadas.

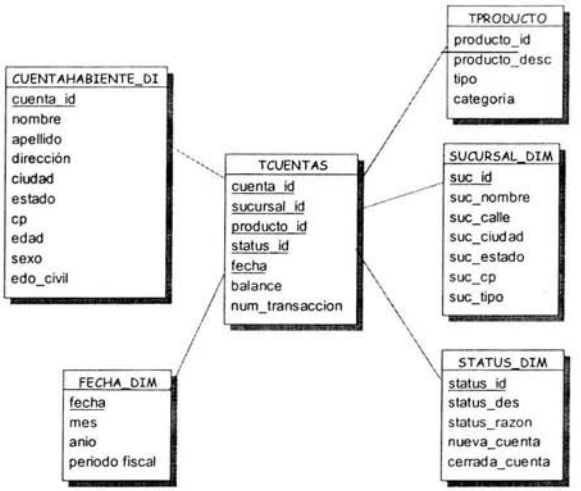

Figura 3.7 Relaciones entre la tabla de hechos y las de dimensiones

Cada tabla de dimensión presenta una llave primaria que hace único a los registros almacenados en la tabla. La llave primaria de la tabla de hechos está compuesta por una serie de llaves foráneas que corresponden a las llaves primarias de cada una de las dimensiones con las que tiene relación. Por ejemplo, en la tabla de dimensión *fecha_dim* (*figura 3.7)* la llave primaria corresponde a una de las llaves foráneas de la tabla de hechos *cuentas*. Cada llave foránea de la tabla de hechos debe tener su contraparte en las tablas de dimensiones.

3.4.1.5 Definición de los atributos de las dimensiones con una descripción completa y la terminología apropiada

Después de completar la tabla de hechos, se definen los atributos para cada tabla de dimensión. Por medio de las dimensiones, como ya había sido mencionado, el usuario obtiene la información que necesita. Por lo tanto, la asignación de una descripción completa y apropiada de los atributos es de gran importancia ya que

dependiendo de esta el usuario encontrará en el data warehouse una herramienta sencilla o difícil de utilizar, hay que recordar que el éxito de un data warehouse, en gran parte, depende de la aceptación y utilidad que represente para los usuarios.

El número de atributos definidos en cada dimensión, de preferencia, se debe de mantener al mínimo. Las tablas de dimensiones con una gran cantidad de atributos se pueden encontrar con un problema de demasiadas líneas y un pobre performance.

Para una mejor comprensión de la metodología expuesta se presenta un caso práctico sobre el modelado lógico de una base de datos dimensional. A continuación es presentada la situación de una empresa de telefonía móvil que requería consultar información histórica sobre el tráfico de llamadas para el soporte a la toma de decisiones.

3.5 Caso Práctico: modelado de una base de datos dimensional para la consulta del tráfico de llamadas mediante la metodología presentada

En una empresa de telefonía móvil todas las llamadas que cruzan tráfico por la red utilizada para dar servicio a sus clientes son almacenadas en un repositorio central llamado Collector. En este repositorio se encuentran las llamadas en forma de registros que contienen toda la información relacionada a esta (el teléfono que la originó, teléfono que la recibió, la hora y fecha en la que fue generada, si utilizó algún servicio adicional de la empresa, si el teléfono destino fue marcado con algún prefijo especial, la central que recibió y atendió esa llamada, si fue exitosa o si fue interrumpida por alguna causa anormal, etc.). A estos registros se les llama CDR (Call Detail Register) y son la materia prima que mantiene funcionando a la organización.

Todas las áreas de la empresa trabajan directa o indirectamente con estos registros, por ejemplo, el área de Facturación Postpago utiliza la información contenida en el

CDR para tasarla y cobrarle al cliente, el área de Activaciones utiliza estos registros para saber cuantos teléfonos fueron activados y en que condiciones, el área de Soporte a Usuarios analiza la información para resolver problemas reportados por los usuarios, el área de Finanzas los utiliza para conocer el monto del pago y cobro a otros operadores (Telmex, IUSACELL, etc.) con los que la empresa tiene convenios de interconexión, larga distancia, el que llama paga, etc.

Sin embargo, el acceso de esta información solo es permitido a algunas áreas que requieren trabajar directamente con el CDR como el área de Facturación Postpago, el área de Soporte a Usuarios, el área de Sistemas de Interconexión y Larga Distancia, etc. Los encargados de enviar esta información es el área de administración de CDR's quienes distribuyen los datos según los requerimientos de cada una de estas áreas.

Como se mencionó anteriormente, una de las áreas que requieren trabajar directamente con el CDR es el área de Sistemas de Interconexión y Larga Distancia. Esta área se encarga de obtener la información concerniente a las llamadas que presentan interconexión local y de larga distancia con otros operadores. Por interconexión local se conoce el intercambio de tráfico local entre la empresa y los diferentes operadores, a este tipo de llamadas se les aplica un cobro y pago en base a una tarifa establecida en los convenios celebrados con los operadores. De la misma manera las llamadas que presentaron larga distancia son cobradas y pagadas a otros operadores con la tarifa estipulada en los acuerdos.

Esto quiere decir que cuando la empresa recibe llamadas clasificadas como de interconexión local se le debe cobrar al operador el uso de la red en base a una tarifa establecida, de la misma manera, cuando algún celular de la empresa realiza una llamada a otro operador utilizando interconexión local se debe pagar según lo estipulado en los convenios. Lo mismo sucede con el cobro y pago a otros operadores por larga distancia.

A diferencia de otras áreas que también reciben los CDR's, el área de Sistemas de Interconexión y Larga Distancia requiere recibir todas las llamadas que cruzan tráfico por la red para identificar aquellas que presentaron interconexión local y larga distancia, tasarlas y obtener el monto que se debe cobrar a cada operador según los convenios celebrados. Además, se tiene que validar o conciliar que las llamadas cobradas por los operadores existan. Esta información es entregada al área de Finanzas que realiza los pagos y cobros.

Para llevar a cabo esta tarea es necesario saber algunos datos como los operadores a los que pertenecen el teléfono que generó y el que recibió la llamada, la duración, si la llamada proviene de otro operador o fue generada por un cliente de la empresa, el lugar donde fue generada y recibida, etc. Sin embargo, esta información no es recibida de esta manera en el CDR el cual únicamente presenta datos técnicos que es necesario descifrar para obtener toda esta información.

Para descifrar los datos contenidos en el CDR, el área de Interconexión y Larga Distancia desarrolló e implementó el Sistema de Interconexión (*figura 3.8*) el cual obtiene toda esta información e identifica las llamadas locales de las de larga distancia. Este sistema toma todas las llamadas recibidas diariamente y las procesa en siete pasos:

Paso 1: Formatea las llamadas recibidas preparándolas para el procesamiento posterior.

Paso 2: Da formato a la fecha y hora de llamada y en base a la marcación del teléfono origen y el teléfono destino clasifica las llamadas de la siguiente manera:

- Llamadas de larga distancia (Internacional, Mundial o Nacional)
- Llamadas normales.

- Llamadas con marcaciones especiales (consulta de saldo, consulta de buzón, consulta de la hora, etc.)
- Llamadas que impliquen uno o más teléfonos roamers.
- Llamadas que ocupen un servicio adicional de la empresa (llamada tripartita, llamada en espera, transferencia de llamada, etc.)
- Llamadas locales.
- Llamadas de ocho dígitos (actualmente se manejan diez dígitos en los números)
- Llamadas con 01900 (números de concursos) o 01800 (llamadas de no cobro).

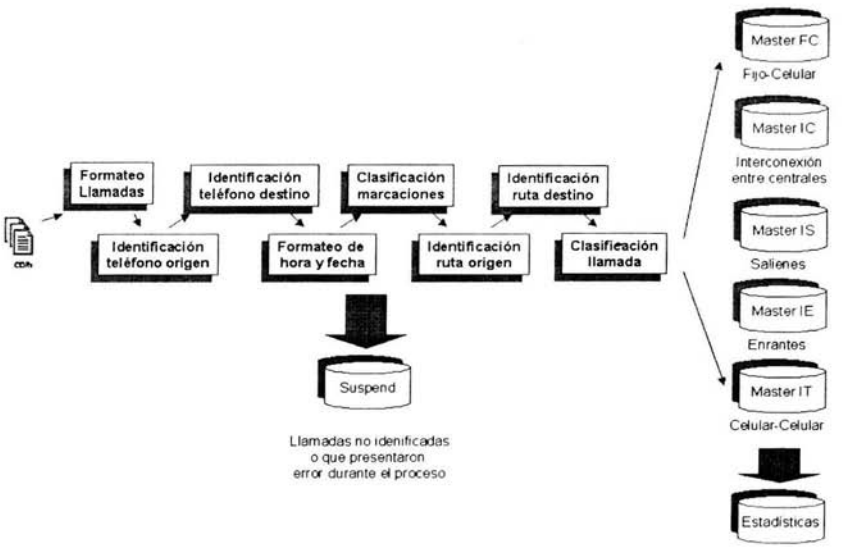

Figura 3.8 Diagrama del procesamiento de llamadas en el sistema de interconexión

Paso 3: Identifica el operador, el plan, el área de servicio local y la región a la cual pertenece el número que originó la llamada en base al plan nacional de numeración proporcionado por la COFETEL.

Paso 4: Identifica el operador, el plan, el área de servicio local y la región a la cual pertenece el número que recibió la llamada en base al plan nacional de numeración proporcionado por la COFETEL. Una vez identificados los operadores, tanto del teléfono que originó como el que recibió la llamada, ya es posible identificar la dirección de la misma y clasificarla de la siguiente manera:

1. Llamadas Entrantes: llamadas recibidas de otro operador (por ejemplo, TELMEX). Estas llamadas son cobradas al operador identificado.
2. Llamadas Salientes: llamadas generadas por un celular de la empresa hacia otro operador. Estas llamadas son pagadas a otros operadores y conciliadas por la empresa.
3. Llamadas Celular a Celular: llamadas generadas entre dos teléfonos que pertenecen a la empresa.

Paso 5. Es identificada la ruta entrante, la región donde se encuentra la ruta y el tipo de ruta. Esta identificación se ejecuta ya que proporciona información del lugar geográfico donde fue generada la llamada y por donde entró físicamente a la red.

Paso 6. Es identificada la ruta saliente, la región donde se encuentra la ruta y el tipo de ruta. Esta identificación se ejecuta ya que proporciona información del lugar donde fue entregada la llamada y por donde salió físicamente de la red, pudo haber salido por medios propios de la empresa o por una red distinta.

Paso 7. Por último, tomando en cuenta una serie de validaciones las llamadas son distribuidas en diferentes tablas utilizadas para la consulta de la información, en estas tablas se almacena la información de los últimos tres meses. Por mes se están recibiendo aproximadamente 900 millones de

llamadas que cruzan tráfico por la red. Las llamadas son almacenadas en llamadas Fijo-Celular (cierta modalidad de teléfonos que renta la empresa), llamadas entrantes, llamadas salientes, llamadas celular-celular, llamadas de interconexión entre centrales y el suspend (tabla donde se almacenan las llamadas que pudieron ser identificadas o que presentaron errores en el registro).

3.5.1 Definición del problema

Debido a que el sistema de interconexión recibe todas las llamadas que cruzan tráfico por la red, las clasifica e identifica; algunas áreas como Ingeniería y sobre todo Finanzas, comenzaron a solicitar información sobre el tráfico de llamadas. Muchas veces esta información era utilizada para la toma de decisiones y requería de la elaboración de reportes; por ejemplo, comparativos entre varios meses o incluso semestres del comportamiento del tráfico de llamadas. Algunos requerimientos de información solicitaban el número de llamadas de larga distancia clasificadas en internacionales, mundiales y nacionales por mes en el último semestre o el número de llamadas entrantes, salientes y de celular a celular por mes en el último año para observar el comportamiento del tráfico y por lo tanto del cliente.

Poco a poco estos reportes ya no fueron únicamente comparativos, también comenzaron a solicitarse reportes más complejos como la duración y número de llamadas salientes donde fue utilizado algún servicio adicional o el número de llamadas entrantes y salientes generadas por operador clasificadas en llamadas locales y de larga distancia.

La solicitud de esta información comenzó a ser un problema para el área de Interconexión y Larga Distancia ya que implicaba el desarrollo de un nuevo reporte y recursos dedicados a su generación, el área no podía proporcionar esta información con la rapidez que el usuario requería, debido a la gran cantidad de información

almacenada en la base de datos, no se tenía toda la información histórica requerida y las tareas principales del área (la identificación de las llamadas con interconexión y larga distancia y la conciliación de estas para el pago y cobro a operadores) comenzaron a tener retrasos debido a que no se contaban con los recursos materiales y humanos suficientes para atender todas estas tareas.

3.5.2 Diseño conceptual del caso práctico

Debido a que la información requerida es utilizada para la toma de decisiones, el departamento de data warehouse llevó acabo el análisis del problema para encontrar una solución al requerimiento de información sobre el tráfico de llamadas en las diferentes áreas de la empresa, principalmente Finanzas.

Después de una serie de entrevistas con las áreas involucradas (el área de Finanzas, el área de Sistemas de Interconexión y Larga Distancia, el área de Ingeniería y la Dirección de Informática), y después de revisar las fuentes que ya proporcionan dicha información, se llevó a cabo el análisis del problema (*figura 3.9*) donde fueron identificadas las áreas involucradas, los aspectos controlables y no controlables del problema así como las limitaciones existentes y los posibles resultados de las soluciones que pudieran darse a este problema.

También fue elaborado un árbol del problema (*figura 3.10*) donde se pueden apreciar las conclusiones a las que llegó el área de data warehouse, identificando el problema, las causas y sus efectos.

En resumen, el área de data warehouse encontró que el usuario requiere analizar y generar reportes sobre el tráfico de llamadas de manera eficiente, con suficiente información histórica para generar reportes de tres años atrás a la fecha y contar con ésta cuando lo requiera, si es posible, poder generar él mismo estos reportes y tener la flexibilidad de cambiar las preguntas de negocio que requiera; no necesita

visualizar todo el detalle de las llamadas, sólo los datos que pudieran ser útiles para comprender el comportamiento del tráfico. La información que el usuario requiere es la duración de las llamadas y el número de llamadas generadas por fecha, dirección de la llamada, servicios utilizados, tipo de tráfico, operador origen, operador destino, etc.

QUIEN EXPERIMENTA EL PROBLEMA Y QUIENES TOMAN DECISIONES	ASPECTOS DEL PROBLEMA QUE SE PUEDEN CONTROLAR	ASPECTOS DE LA SITUACIÓN QUE SE ESCAPAN DEL CONTROL DEL QUE TOMA LA DECISIÓN	LIMITACIONES	POSIBLES RESULTADOS
Área de Interconexión y Larga Distancia. Tomadores de decisiones de las áreas de Informática, Finanzas e Ingeniería.	Organización del personal para un mayor rendimiento con la finalidad de generar los reportes solicitados con mayor rapidez y cumplir con las tareas de conciliación y facturación. Requerimiento de recursos humanos, materiales y financieros necesarios para generar esta información. Monitoreo del Sistema de Interconexión para revisar la recepción de todas las llamadas del collector y analizar aquellas que no fueron identificadas asegurando la calidad de los datos.	Manejo de grandes volúmenes de información (900 millones de registros al mes). La calidad de la información depende del correcto envío por parte del área de Administración del Collector y del correcto registro en la red de las llamadas que cruzan tráfico. Cada vez hay mas necesidad, por parte de las diferentes áreas de la empresa, de información sobre el tráfico de llamadas. Cambios en la preguntas de negocio por parte de los tomadores de decisiones según las necesidades de información para la planeación estratégica. Existen otras tareas que presentan prioridad en el área de Interconexión y Larga Distancia debido a la importancia que representan para los ingresos y egresos de la empresa (conciliación y facturación a otros operadores de interconexión local y de larga distancia).	Recursos humanos, materiales y financieros limitados que ocasionan retraso en la generación de reportes para toma de decisiones. Información histórica limitada (tres meses de información en base de datos). La calidad de la información reportada depende de la total recepción de las llamadas, del correcto registro de llamadas en la red y del conocimiento de los analistas sobre casos de tráfico para la correcta identificación y clasificación de las llamadas.	El área de Interconexión y Larga Distancia continúa generando reportes cada determinado tiempo ocupando mayores recursos aunque limitados por información histórica. El área de data warehouse construye una base de datos con información histórica, alimentándola de la información generada por el Sistema de Interconexión, para la generación de información proporcionada de manera eficaz, eficiente y en el momento en el que el tomador de decisiones la necesite. Esta base de datos también permitiría.

Figura 3.9 Análisis del problema

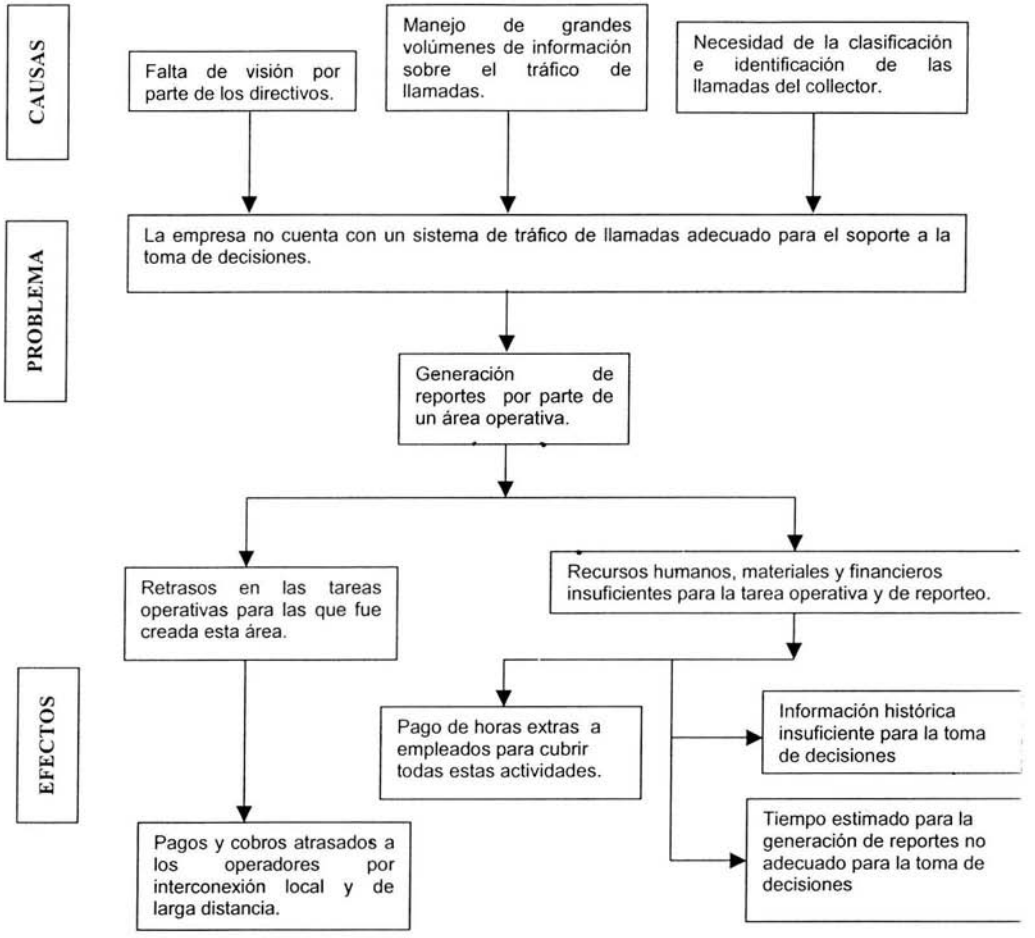

Figura 3.10 Árbol del problema

Con la información obtenida del análisis fueron identificados los principales requerimientos que a continuación se presentan.

3.5.2.1 Requerimientos obligatorios

Aquellos que el sistema debe cubrir para considerar que cumple con las funciones para las cuales fue creado.

- Proporcionar información oportuna y confiable sobre el tráfico de llamadas para una buena toma de decisiones.
- Proporcionar información histórica de por lo menos tres años.
- Presentar la información organizada en un modelo de datos flexible y adecuado para el análisis de información desde el punto de vista del negocio.
- Esquema de base de datos que el usuario pueda comprender fácilmente para consultarla él mismo.
- Elaboración de un plan de administración y mantenimiento de la base de datos.
- Elaboración de un plan de seguridad de la información para garantizar la calidad y confidencialidad de los datos.
- Elaboración de un plan de extracción y carga de información a la base de datos. (recepción diaria de la información enviada por el área de Interconexión y Larga Distancia en archivos planos, proceso ETL para la carga de la información al data warehouse, etc.)
- Manejador de base de datos adecuado para la creación y administración de un data warehouse.
- Contar con herramientas de explotación de la base de datos (aplicaciones y herramientas de consulta)
- Capacitación de los usuarios sobre el uso de las herramientas de explotación de la base de datos y del uso de la misma para obtener la información que necesita.

3.5.2.2 Requerimientos deseados

Son aquellos que no es necesario que sean cubiertos por el sistema para garantizar que cumpla con su función.

- Manual de usuario de la base de datos.
- Manual técnico de la base de datos.

3.5.2.3 Requerimientos funcionales

Los requerimientos funcionales son aquellos servicios que el sistema será capaz de realizar, se refieren a las entradas y salidas del sistema.

- Generación de reportes sobre la información del tráfico de llamadas en el Collector para el soporte a la toma de decisiones.
- Análsis de los datos desde diferentes perspectivas del negocio.
- Información histórica
- Administración de los usuarios para el control del acceso a la información de la base de datos.
- Información confiable y consultas rápidas de la información.

3.5.2.4 Requerimientos no funcionales

Los requerimientos no funcionales son aquellos referentes a las características del sistema que podrían representar una limitación para el sistema (mantenimiento, seguridad, rendimiento, etc.)

- Redbrick 6.0 como manejador de base de datos.
- BrioQuery 6.5 como herramienta de consulta.
- PC con Windows 9X o 2000.
- Conexión de la aplicación por medio de ODBC.

- Data Stage como herramienta ETL (Extraction Transformation Loading Tool).
- Servidor con sistema operativo UNIX.
- Información fuente enviada diariamente por el área de Interconexión y Larga Distancia en archivos planos.

3.5.3 Diseño lógico del caso práctico

Después del diseño conceptual, el área de data warehouse procedió a elaborar el diseño lógico tomando como guía los pasos especificados en la metodología presentada:

1) Elección del proceso a modelar en el caso práctico.

Después del diseño conceptual donde se hizo un análisis del problema, se hicieron entrevistas con las áreas involucradas y los requerimientos fueron definidos; el área de data warehouse concluyó que el proceso a modelar sería **el tráfico de llamadas** y la información sería obtenida del Sistema de Interconexión que es el principal proveedor actual de esta información.

Los diseñadores tuvieron que realizar un análisis del Sistema de Interconexión y la forma en la que son identificadas y clasificadas las llamadas. Aunque esto parece sencillo el conocimiento de este proceso es complicado y requiere del conocimiento de 273 casos de tráfico. Es necesario remarcar, que el conocimiento del proceso de negocio a modelar es fundamental para obtener un esquema de base de datos que cubra las necesidades de información para las cuales es creada.

2) Definición del nivel de granularidad o detalle en el caso práctico

Una vez definido y estudiado el proceso a modelar fue necesario especificar la granularidad. En el diseño conceptual del proceso de tráfico de llamadas se

encontró que los usuarios no requieren consultar todo el detalle de las llamadas ya que éste se puede obtener del Sistema de Interconexión y solo es consultada durante la conciliación y la facturación. Sin embargo, sí requieren consultar información sobre la dirección, los operadores de los teléfonos que originan y reciben las llamadas, la central por donde fue recibida y generada, región de la llamada, el tipo de tráfico y los servicios utilizados.

Hay que tener cuidado en el momento de definir la granularidad, aunque un data warehouse requiere de información detallada que permita flexibilidad en las consultas, esta puede ser un arma de doble filo. Si se elige poca granularidad (información muy detallada) la cantidad de información podría ser demasiada y ocasionaría problemas de espacio en la base de datos o en filesystem (si la información es recibida en archivos como en este caso, se recomienda guardar o respaldar estos archivos como medida de seguridad en caso de perder la información de alguna de las tablas). Para asegurar que la cantidad de información obtenida con la granularidad determinada es manejable, es recomendable hacer un cálculo de la información que será almacenada durante los próximos tres años. Además, hay que revisar si existen otros procesos que ya se encuentren productivos y que pudieran verse afectados al tener que compartir demasiados recursos con el nuevo proceso.

Por otro lado, si es elegida alta granularidad (poco detalle) se corre el riesgo de que las consultas sean poco flexibles, que no sean lo suficientemente rápidas y no cumplan con las expectativas de los usuarios. Es muy importante que el usuario encuentre útil la información contenida en el data warehouse, de lo contrario comenzará a buscarla en otras fuentes. Se han observado casos en los que el usuario no encuentra utilidad en los datos almacenados en el data warehouse (por consultas lentas, información inconsistente, por ser muy difícil de utilizar o comprender), estos sistemas con el tiempo son eliminados e incluso sustituidos por otros que sí cumplan los requerimientos del usuario.

Para determinar si la granularidad propuesta por el data warehouse era la mas adecuada, en el proceso de tráfico de llamadas se hicieron los siguientes cálculos:

Dimensión	Cálculo de registros
Tiempo	365 días X 3 = 1,095 días en 3 años.
Central Entrante	59 centrales que atienden llamadas diariamente repartidas en nueve regiones.
Central Saliente	59 centrales que atienden llamadas diariamente repartidas en nueve regiones.
Dirección	3 direcciones (entrante, saliente, celular-celular).
Operador origen	15 operadores (Telmex, IUSACELL, Pegaso, etc.).
Operador destino	15 operadores (Telmex, IUSACELL, Pegaso, etc.).
Tipo Tráfico de Llamadas	Una llamada puede ser clasificada en uno de los 6 tipos de tráfico identificados (larga distancia internacional, larga distancia nacional, larga distancia mundial, llamada local, roamer internacional, roamer nacional).
Tipo de Servicio	Una llamada puede ser clasificada en un solo tipo de servicio (normal, conferencia tripartita, transferencia de llamada, consulta a buzón, marcaciones especiales).

Figura 3.11 Posibles valores que presenta cada dimensión

Si la granularidad se define como llamada por tiempo por central entrante, por central saliente por dirección por operador origen por operador destino por tipo de tráfico y por servicio (*figura 3.11*) y se hace el cálculo del espacio que ocupará la tabla de hechos en un período de tres años, el número de registros que puede presentar la tabla de hechos en tres años con la granularidad definida sería:

Número de registros máximo que puede contener la tabla de hechos: 1,095 X 59 X 59 X 3 X 15 X 15 X 6 X 6 = 92,624,188,500 de registros.
Número de campos que forman la llave: 8.

Número de campos que forman las métricas: 3 (si los hechos están formados por número de llamadas, duración y duración send).

Tamaño estimado de la tabla de hechos para 3 años:
92 billones de registros X 11 campos X 4 bytes = 4 Tb

Si se comparan los 4 Tb ocupados en tres años en el data warehouse con los 1.5 Tb que ocupa por tres meses en base de datos el Sistema de Interconexión, el espacio que va a ocupar la tabla de hechos durante tres años es menor y si además de esto se divide la información por año, se tendrían tablas de 1.25 Tb por año. Además de presentar información histórica las consultas serán mas rápidas y la información no presentará datos innecesarios para el usuario pero tampoco suprime datos que pudiera necesitar para el análisis del comportamiento del tráfico de llamadas.

En cuanto a las tablas de dimensión el espacio que requieren es tan pequeño que no representan preocupación para el diseñador. En el cálculo del número de registros de la tabla de hechos ya fue presentado el número aproximado de registros que podría contener cada dimensión (*figura 3.11*). El espacio que ocupará cada tabla de dimensiones es muy pequeño, la tabla más grande sería la tabla de la dimensión de tiempo con 1,095 registros. Durante los próximos tres años los registros de las tablas de dimensiones podrían aumentar, aún así, el tamaño de las tablas no representa un problema ya que el número de operadores podría crecer uno o dos registros por año o los servicios ofrecidos por la empresa se podrían incrementar en tres o cuatro registros y lo mismo sucede con el crecimiento de las dimensiones restantes.

3) Identificación de las dimensiones de cada tabla de hechos para el caso práctico

Una vez definida la granularidad es más sencillo identificar las dimensiones. Como había sido explicado anteriormente cada uno de los elementos que forman la granularidad representan una dimensión. Como en este caso la granularidad está formada por fecha, central entrante, central saliente, dirección, operador origen, operador destino, tipo de tráfico y servicio serán creadas una dimensión para cada uno de estos elementos (*figura 3.12 y 3.13*).

Figura 3.12 Dimensiones de la tabla de hechos tráfico de llamadas

En este paso también son definidos los elementos y jerarquías de las dimensiones tomando en cuenta las consultas que se requieren y procurando que el modelo sea lo más flexible posible para el usuario. Una manera de identificar los elementos y jerarquías de las dimensiones es revisando los reportes y la información generada por las áreas usuarias. En ocasiones, durante las entrevistas no es posible obtener toda la información que se requiere, ya sea por que el usuario no considere importante mencionar algunos detalles o porque no quiere dar demasiada información. El diseñador puede apoyarse en los reportes y documentos generados por las áreas si se le permite tener acceso a ellos.

Por ejemplo, para obtener los elementos y jerarquías de las dimensiones del tráfico de llamadas fueron revisados los requerimientos de información levantados al área de sistemas de interconexión en los últimos meses. En uno de ellos era solicitado por el área de Finanzas la información sobre el número de llamadas de larga distancia mundial, internacional y nacional; se requería el número de llamadas entrantes y salientes que presentaran estos tipos de tráfico de llamadas por mes durante los primeros seis meses del año. En otro requerimiento era solicitado un comparativo de los meses de marzo, abril y mayo de las llamadas que consultaban a buzón con una marcación especial. Otro requerimiento, presentado por Ingeniería, requería el número de llamadas por central y por día de la semana.

Nombre dimensión	Descripción dimensión
Fecha_dim	Fecha en la cual es realizada la llamada.
Central_entrante_dim	Lugar geográfico donde fue registrada la generación de la llamada.
Central_saliente_dim	Lugar geográfico donde fue registrada la recepción de la llamada.
Dirección_dim	Dirección de la llamada.
Operador_origen_dim	Operador al cual pertenece el teléfono que generó la llamada.
Operador_destino_dim	Operador al cual pertenece el teléfono que recibió la llamada.
Tráfico_dim	Tipo de tráfico de la llamada.
Servicio_dim	Servicio utilizado en la llamada.

Figura 3.13 Dimensiones de la tabla de hechos tráfico de llamadas

Es necesario remarcar que la definición de los elementos y jerarquías de las dimensiones es de gran importancia ya que de estos dependerá que el modelo obtenido sea lo suficientemente flexible poder elaborar diferentes preguntas de negocio. Además, la correcta conformación de una tabla de dimensión influye para que el usuario consulte apropiadamente las métricas en la tabla de hechos.

En la gráfica de la dimensión de tiempo (*figura 3.14*), con el nombre *fecha_dim*, se pueden observar las jerarquías que permiten generar reportes por día, mes, cuatrimestre, semestre o año, el usuario puede consultar la información tan detallada

o resumida como lo requiera. Además las consultas se pueden realizar por día hábil o no hábil, por día de la semana (en ocasiones el tráfico de llamadas varía dependiendo del día de la semana en la que fue generada) o día festivo.

Figura 3.14 Jerarquías y elementos de la dimensión fecha_dim

En la gráfica de la dimensión de central entrante y de la dimensión de central saliente (*figura 3.15*), con los nombres de *central_entrante_dim* y *central_saliente_dim* respectivamente, las consultas de la información se pueden llevar a cabo por región, por central, estado, colonia, ciudad o calle donde se encuentra localizada la central que capta la llamada por una radiobase o que la entrega a un celular o a otra red de telefonía según el escenario de tráfico que se presente. Esta información es de gran ayuda ya que se puede saber si una central está atendiendo más tráfico de lo recomendable o si está ociosa, incluso de esta manera se pueden localizar los lugares donde hay mayor o menor tráfico de llamadas y por lo tanto donde los clientes están ocupando más el servicio y donde lo ocupan menos. Por ejemplo, el tráfico generado en la Ciudad de México no es el mismo al generado en La Paz.

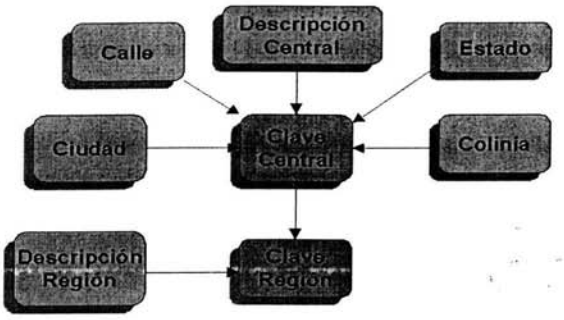

Figura 3.15 Jerarquías y elementos de las dimensiones Central_entrante_dim y Central_saliente_dim

Figura 3.16 Jerarquías y elementos de la dimensión Dirección_dim

La dimensión de la dirección (*figura 3.16*), con nombre *dirección_dim*, permite consultar las métricas en base a la dirección de las llamadas. Si la llamada fue recibida de otro operador, es una llamada entrante, si es enviada a otro operador entonces es saliente y si es generada de un celular a otro celular de la empresa es una llamada celular-celular. Esta información es importante conocerla, sobre todo porque permite conocer el tráfico que es intercambiado con los operadores y conocer el monto a pagar y cobrar por interconexión, por larga distancia o por la modalidad "el que llama paga".

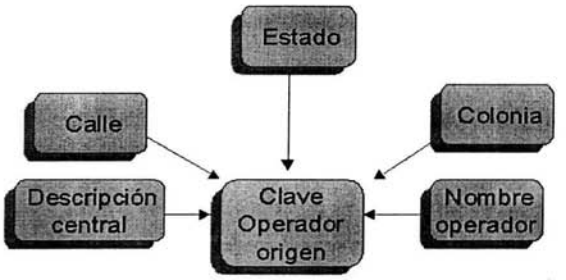

Figura 3.17 Jerarquías y elementos de la dimensión Operador_orig_dim

Figura 3.18 Jerarquías y elementos de la dimensión Operador_dest_dim

En las dimensiones de operador origen y operador destino (*figura 3.17 y 3.18*), conocidas con los nombres *operador_origen_dim* y *operador_destino_dim* respectivamente, las consultas se pueden hacer por el operador que generó la llamada y por el que la recibió, es decir, se puede saber cuantas llamadas ha generado la empresa a la competencia y viceversa, esto también es de gran ayuda sobre todo en el cobro y pago de interconexión y larga distancia o en la modalidad "el que llama paga".

En la dimensión de tráfico (*figura 3.19*), conocida como trafico_dim, existen dos elementos de dimensión, las consultas se pueden hacer en base al tráfico que presentó la llamada, es decir, si la llamada fue local, de larga distancia o roamer y

también se puede conocer con más detalle si el tipo de tráfico fue de larga distancia mundial, nacional o internacional o si la llamada fue de un roamer internacional o nacional. Esto también es importante para el pago y cobro de interconexión local y de larga distancia al operador correspondiente o para conocer si los clientes están utilizando el servicio de roamer.

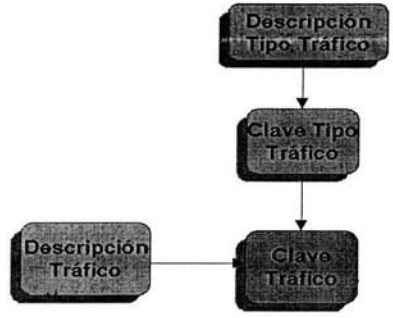

Figura 3.19 Jerarquías y elementos de la dimensión Táfico_dim

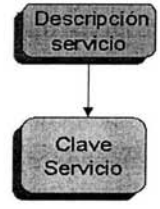

Figura 3.20 Jerarquías y elementos de la dimensión Servicio_dim

Y por último, por medio de la dimensión de servicio (*figura 3.20*), conocida como servicio_dim, se pueden saber el número de llamadas que se hicieron utilizando algún tipo de servicio, por ejemplo, consultas a buzón, conferencia tripartitas, transferencia de llamadas a otros celulares o incluso a un teléfono fijo o algún otro servicio que en ese momento ofrezca la empresa.

Si la información contenida en los hechos se consulta a través de las dimensiones se puede obtener la información deseada a través de una o más dimensiones para contestar la preguntas de negocio. Algunas de estas preguntas podrían ser *¿cuántas llamadas de larga distancia nacional son atendidas por medios propios de la empresa tomando en cuenta que se tiene una conseción de larga distancia y cuántas se hicieron por medios de otro operador?*, tomando en cuenta que el costo del servicio de buzón es de un peso *¿cuál es la ganancia que se está percibiendo por este servicio?, ¿cuántas llamadas locales van a ser cobradas a cierto operador o tienen que ser pagadas por la empresa? o ¿cuáles son las ciudades donde se presenta mayor y menor tráfico de llamadas?.*

4) Definición de la tabla de hechos incluyendo las tablas de hechos precalculadas

Después de definir la granularidad e identificar las dimensiones se procedió a definir la tabla de hechos a detalle. Según el análisis elaborado en el diseño conceptual, el usuario solo necesita saber el número de llamadas, la duración y la duración send. ¿Por qué sólo estos datos? El número de llamadas, la duración y la duración send son datos importantes que pueden proporcionar mucha información como el aproximado de la cantidad de dinero pagado o cobrado a los operadores, el o los servicios que generan ganancias o pérdidas a la empresa, los lugares donde los usuarios utilizan más el servicio, los lugares donde es menos utilizado, es posible saber si los usuarios están usando el celular para generar llamadas o si solo lo están utilizando para recibirlas (como fue detectado con los teléfonos que utilizan la modalidad de prepago), se pueden identificar las centrales que tienen que atender un exceso de tráfico y por lo tanto presentan más fallas o aquellas que son poco utilizadas, el número de llamadas que son generadas con larga distancia, el número de llamadas generadas por teléfonos que no se encuentran en su región (roamers), el monto a pagar por interconexión por operador, etc. Toda esta variedad de información permite al tomador de decisiones conocer el comportamiento del cliente y hacer pronósticos sobre el mercado.

Por lo tanto, la tabla de hechos del proceso de tráfico de llamadas presenta como llave la fecha de llamada, clave de la central entrante, clave de la central saliente, clave de la dirección, clave del operador origen, clave del operador destino, clave del tipo de tráfico y clave del servicio. Y como métricas el número de llamadas, duración en segundos y duración en segundos send como se muestra en la *figura 3.21*. Por cada combinación de las dimensiones que conforman la llave de la tabla de hechos se pueden hacer las diferentes consultas que el tomadores de decisiones requiere.

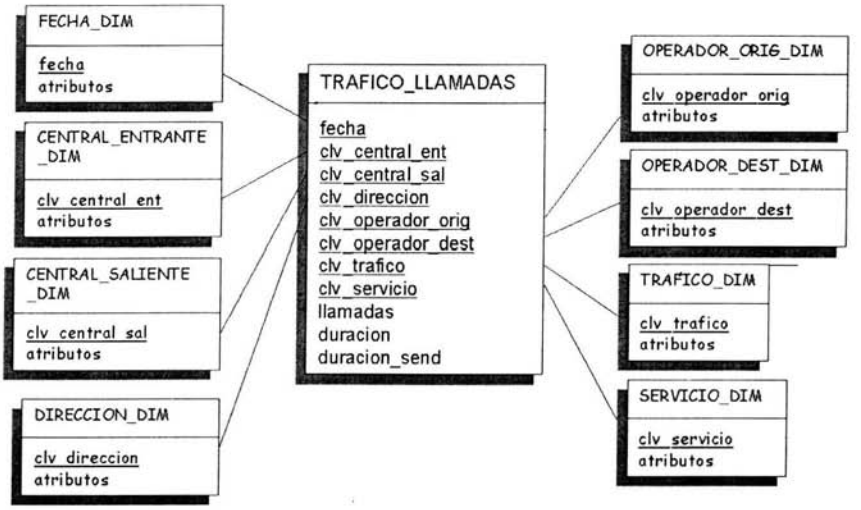

Figura 3.21 Tabla de hechos de tráfico de llamadas

A continuación, es presentada la descripción de los campos que conforman la tabla de hechos y las dimensiones a través de las cuales se puede llevar a cabo una consulta.

Hechos	Dimensiones
Tráfico_llamadas	Fecha_dim
	Central_entrante_dim
	Central_saliente_dim
	Dirección_dim
	Operador_orig_dim
	Operador_dest_dim
	Tripo_trafico_dim
	Servicio_dim

Figura 3.22 Dimensiones que tienen relación con la tabla de hechos de tráfico de llamadas

Nombre columna	Descripción	Ejemplo de valores
Fecha	Fecha en la cual se llevó a cabo la llamada	DD/MM/YYYY
Clv_central_ent	Clave de la central donde fue recibida la llamada	10
Clv_central_dest	Clave de la central donde fue entregada la llamada	10
Clv_dirección	Clave de la dirección de la llamada	EN
Clv_operador_orig	Clave del operador del teléfono que generó la llamada	2
Clv_operador_dest	Clave del operador del teléfono que recibió la llamada	6
Clv_tipo_tráfico	Clave del tráfico de la llamada	RI, LI, NO
Clv_servicio	Clave del servicio utilizado en la llamada	BZ, CS, TL
Duración	Duración en segundos de la duración de la llamadas	1421
Duración_send	Duración en segundos a partir de dar send	1623
No_llamadas	Número de llamadas	52

Figura 3.23 Descripción de las columnas que contiene la tabla de tráfico de llamadas

5) Definición de los atributos de las dimensiones con una descripción completa y la terminología apropiada.

Una vez definida la granularidad, las tablas de hechos y las dimensiones, se procede a definir a detalle las tablas de dimensiones con sus atributos y la descripción de cada uno de ellos. La descripción de los atributos debe ser clara y en términos que el usuario comprenda para facilitar las consulta de la información. La terminología elegida debe ser la que el usuario maneje en el proceso del negocio. Toda esta información fue obtenida en la fase del diseño conceptual; sin embargo, no hay que

el olvidar que el diseño de la base de datos puede ser revisado varias veces antes de ser concluido.

A continuación son presentadas las dimensiones con los atributos y la descripción de cada uno de ellos.

En la tabla de dimensión *fecha_dim*, la consulta se puede hacer especificando la fecha de la llamada o el número del día en el mes, en el año, si es día festivo o hábil, el mes, cuatrimestre, semestre o año de las llamadas que se requieren consultar en la tabla de hechos (figura 3.24).

Nombre atributo	Descripción	Ejemplo de valores
Fecha	Fecha en la que fue generada la llamada	DD/MM/YYYY
No_día_cal_mes	Número del día en el mes	1 a 28,30 ò 31 en el mes
No_día_cal_anio	Número del día en el año	Día 85 en el año
Día_semana	Día de la semana	Lunes, Miércoles, Viernes, etc.
Día_festivo	Bandera que indica si se trata de un día festivo	1= día festivo; 0=día no festivo
Día_hábil	Bandera que indica si se trata de un día hábil	1= día hábil; 0=fin de semana
Mes	Número del mes en el año	1 al 12
Cuatrimestre	Cuatrimestre que contiene la fecha	1 al 4
Semestre	Semestre que contiene la fecha	1 al 2
Anio	Año	Ejemplo: 2003

Figura 3.24 Atributos y descripción de la dimensión fecha_dim

La dimensión *central_entrante_dim y central_saliente_dim*, es la tabla donde se encuentran los datos geográficos de la llamada. Esta puede ser utilizada para especificar en la consulta la región o central donde se generó y se recibió la llamada (*figura 3.25*).

Nombre atributo	Descripción	Ejemplo de valores
Clv_central_ent	Clave de la central por donde entró la llamada	10
Clv_central_sal	Clave de la central por donde salió la llamada	20
Desc_central	Descripción de la central	Xochimilco
Clv_región	Clave de la región	R09
Desc_región	Descripción de la región	México
Calle	Calle donde se encuentra ubicada la central	Bosques
Colonia	Colonia donde se encuentra ubicada la central	San Ángel
Ciudad	Ciudad donde se encuentra ubicada la central	Ciudad de México
Estado	Estado donde se encuentra ubicada la central	Distrito Federal

Figura 3.25 Atributos y descripción de las dimensiones central_entrante_dim y central_saliente_dim

Con la dimensión *direccion_dim*, el usuario puede especificar en la consulta si requiere conocer la información concerniente a las llamadas entrantes, salientes o de celular-celular (*figura 3.26*).

Nombre atributo	Descripción	Ejemplo de valores
Clv_dirección	Clave de la dirección	EN, SA, IC, IT
Desc_dirección	Descripción de la dirección	Entrante, saliente, interconexión entre centrales, celular a celular

Figura 3.26 Atributos y descripción de la dimensión dirección_dim

La dimensión *operador_origen_dim*, permite hacer consultas del tráfico de llamadas generado por operador, esta es una de las dimensiones más importantes ya que la información puede ser consultada para conocer el cobro y pago a otros operadores lo cual representa parte importante de los ingresos y egresos de la empresa (*figura 3.27*).

Nombre atributo	Descripción	Ejemplo de valores
Clv_operador_orig	Clave del operador origen	4
Nombre_operador	Descripción del operador origen	IUSACELL PCS, S.A de C.V.
Calle	Calle donde se encuentra ubicado el operador origen	Margaritas
Colonia	Colonia donde se encuentra ubicado el operador origen	Santa Fe
Ciudad	Ciudad donde se encuentra ubicado el operador origen	Ciudad de México
Estado	Estado donde se encuentra ubicado el operador origen	Distrito Federal

Figura 3.27 Atributos y descripción de la dimensión operador_origen_dim

Lo mismo sucede con la dimensión *operador_destino_dim* (*figura 3.28*).

Nombre atributo	Descripción	Ejemplo de valores
Clv_operador_dest	Clave del operador destino	4
Nombre_operador	Descripción del operador destino	IUSACELL PCS, S.A de C.V.
Calle	Calle donde se encuentra ubicado el operador destino	Margaritas
Colonia	Colonia donde se encuentra ubicado el operador destino	Santa Fe
Ciudad	Ciudad donde se encuentra ubicado el operador destino	Ciudad de México
Estado	Estado donde se encuentra ubicado el operador destino	Distrito Federal

Figura 3.28 Atributos y descripción de la dimensión operador_destino_dim

Con la dimensión *trafico_dim*, el usuario puede consultar las llamadas clasificadas en larga distancia, llamadas locales, roamers o si necesita consultar mayor detalle por el tipo de tráfico, larga distancia mundial, nacional, etc. (*figura 3.29*)

Nombre atributo	Descripción	Ejemplo de valores
Clv_tráfico	Clave del tráfico	RM, LD,NO.
Desc_tráfico	Descripción del tráfico	Roamer, larga distancia, normal.
Clv_tipo_tráfico	Clave del tipo de tráfico	RI, RN, LN, LI, LM, NO..
Desc_tipo_tráfico	Descripción del tipo de tráfico	Roamer nacional, roamer internacional, larga distancia nacional, larga distancia internacional, larga distancia mundial, llamada normal.

Figura 3.29 Atributos y descripción de la dimensión Trafico_dim

Por último, la dimensión *servicio_dim*, permite hacer consultas por el servicio presentado en las llamadas. Si en la llamada fue utilizado el servicio de buzón, de consulta de saldo, de transferencia de llamada, etc. (*figura 3.30*)

Nombre atributo	Descripción	Ejemplo de valores
Clv_servicio	Clave del servicio utilizado	BZ, CS, TL, CT, NM
Desc_servicio	Descripción del servicio utilizado	Buzón, consulta de saldo, transferencia de llamada, conferencia tripartita.

Figura 3.30 Atributos y descripción de la dimensión servicio_dim

3.5.4 Diagrama final del modelo lógico de la base de datos dimensional del tráfico de llamadas de una empresa de telefonía móvil

Hasta este punto finaliza el modelado lógico de la base de datos dimensional del tráfico de llamadas utilizando la metodología presentada. En la *figura 3.31* es expuesto el esquema estrella obtenido para la consulta de información sobre el tráfico de llamadas para el soporte a la toma de decisiones de una empresa de telefonía móvil.

Utilizando la metodología presentada, el modelo obtenido fue un sencillo esquema con una tabla de hechos y siete dimensiones con las cuales el usuario podrá consultar las llamadas según el tipo de llamada, el lugar geográfico donde fue originada y recibida, los operadores que intervienen, la dirección, etc. Como fue

presentado en el caso práctico, esta metodología aportó una guía para obtener el esquema de base de datos de tráfico adecuado a las necesidades de los usuarios y evitar que el diseñador se confundiera con toda la información recopilada en el diseño conceptual y se centrara en los puntos más importantes.

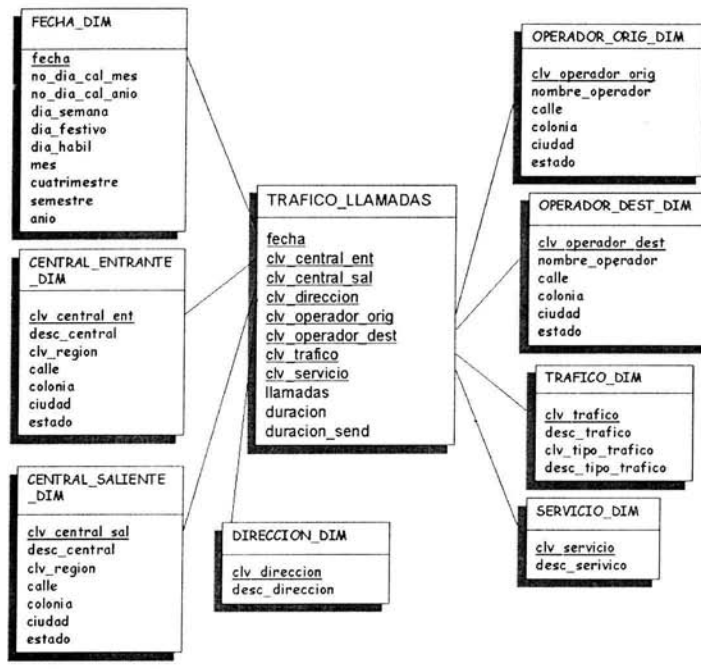

Figura 3.31 Esquema estrella de la base de datos dimensional del sistema de interconexión

De esta manera se concluye la presentación de la metodología y su aplicación al caso práctico.

3.6 Puntos a considerar en la aplicación de la metodología

Por último, a continuación se mencionan algunos puntos que se deben de tomar en cuenta en la aplicación de esta metodología para lograr una base de datos que cubra las necesidades de los tomadores de decisiones:

1. Es muy importante tener conocimiento del proceso de negocio para poder diseñar esquemas que sean útiles y cumplan con las necesidades de información de los tomadores de decisiones.

2. Es importante tener presente que este tipo de base de datos no funcionan como una base de datos relacional. Se han dado casos donde los diseñadores, debido al poco conocimiento que tienen sobre este tipo de sistemas o por evitar hacer un análisis profundo, quieren manejar las bases de datos dimensionales como una base de datos relacional, lo cual es un error y únicamente se tendrá como resultado una base de datos que no es relacional ni dimensional y que no presenta las ventajas de ninguno de los dos modelos. Este problema ha sido observado en las empresas que buscan solucionar sus problemas momentáneamente y no de raíz.

3. En ocasiones no se da la importancia debida al diseño de la base de datos por diversas razones (limitaciones en el presupuesto, falta del conocimiento necesario de los analistas, poco tiempo asignado al proyecto, etc.); sin embargo, se ha observado que muchos de los problemas que presentan este tipo de sistemas (consultas lentas, información no confiable, expectativas de los usuarios pobremente cubiertas, sistemas demasiado complejos) pudieron haberse evitado si se hubiera hecho un buen diseño desde un principio.

4. Es muy importante verificar la calidad de los datos que provienen de los sistemas fuente. Aunque la información haya sido validada en un principio, el analista no debe confiarse. Muchas veces los sistemas fuente hacen cambios que pudieran afectar la generación de la información que va a data warehouse. Aunque se tenga un esquema de base de datos perfecto, si la información que es cargada a las tablas no es confiable de nada sirve el trabajo realizado durante el proceso de diseño.

5. Es imponte la capacitación de los usuarios para que conozcan el sistema y aprendan a utilizarlo de manera que les sea útil y que obtengan todos los beneficios que un sistema de este tipo puede ofrecer.

CONCLUSIONES

El presente trabajo fue presentado con la finalidad de proporcionar material de apoyo para los estudiantes de la carrera de MAC en relación al tema de base de datos. Son expuestos los fundamentos de bases de datos y una introducción sobre el almacén de datos (data warehouse) con la finalidad de que el estudiante, o cualquier persona interesada en el tema, comprenda la metodología expuesta para el diseño lógico de una base de datos dimensional utilizada en el almacenamiento y consulta de información para el soporte a la toma de decisiones.

Además, es presentado un caso práctico donde es aplicada la metodología expuesta y donde se aprecia la utilidad de ésta así como la funcionalidad de la técnica de diseño utilizada (dimensional) y los beneficios que proporciona un almacén de datos a una organización como puede ser una empresa de telefonía móvil.

Por último, es importante remarcar, la importancia que en los últimos años ha tomado la rápida y efectiva toma de decisiones en la planeación estratégica de una organización. Ahora no es suficiente la productividad y buen funcionamiento de una empresa, también se debe tener conocimiento del comportamiento del cliente y del mercado en el cambiante mundo de los negocios y la competencia con otras empresas. El data warehouse o almacén de datos es una herramienta que poco a poco ha tomado mayor importancia en las organizaciones por la información y los análisis que los tomadores de decisiones extraen del sistema.

GLOSARIO

Metodología

Aplicación de un método.

Método

Modo de decir o hacer una cosa con orden y según ciertos principios.

Base de datos

Conjunto de datos relacionados entre si.

Hechos

Acontecimiento o suceso.

Sistema de Gestión de Base de Datos (SGBD)

Sistema de software de propósito general que facilita al usuario el proceso de definir, constuir y manipular las bases de datos.

Metadatos

Datos de los datos contenidos en una base de datos. Los metadatos describen como, cuando y por quien determinada información fue recolectada y como se encuentra almacenada y formateada. Los metadatos son esenciales para comprender la información almacenada en un data warehouse.

Vista

Una vista es una forma particular de ver una base de datos.

Transacción

Se le llama transacción al procesamiento que envía como respuesta un sistema a la solicitud de un usuario.

Modelo de datos

Conjunto de conceptos que describen la estructura de una base de datos.

Estructura de base de datos

Tipos de datos, vínculos y restricciones que deben cumplirse para esos datos.

Modelo Relacional

Conjunto de principios abstractos para el manejo de base de datos propuesto por el Dr. Edgar F. Codd en 1970.

Relación

Elemento básico del modelo relacional representado por una tabla.

Atributo

Propiedad o columna de una tabla.

Tupla

Concurrencia o fila de una relación.

Cardinalidad

Número de filas de una tabla.

Grado

Número de columnas de una tabla.

Dominio

Conjunto finito de valores homogéneos y atómicos.

Llave primaria

Clave que identifica cada tupla de una relación.

Llave foránea

Conjunto de atributos de una tabla cuyos valores coinciden con la llave primaria de otra tabla.

Restricción de base de datos

Estructuras u ocurrencias no permitidas en la base de datos.

Esquema de base de datos

Descripción de una base de datos.

Diseño de base de datos

Proceso a seguir para obtener una base de datos que cubra las necesidades del usuario para las cuales fue creada.

Normalización

Técnica utilizada para comprobar la validez de los esquemas lógicos de una base de datos basados en el modelo relacional.

Datos

Pequeña pieza de información. Puede existir en una gran variedad de formas, como números, texto, bits, etc.

Sistemas de soporte a la toma de decisiones

Conjunto de herramientas de tecnología de la información que permite a los tomadores de decisiones interactuar y resolver problemas para la toma de decisiones.

Data warehouse

Conjunto de datos integrados orientados a una materia, que varían con el tiempo y que son transitorios, los cuales soportan el proceso de toma de decisiones de una administración.

Performance

Resultado.

Herramientas de soporte a la toma de decisiones

Aplicaciones y herramientas utilizadas para recuperar, manipular y analizar los datos.

OLAP

Herramientas de proceso analítico en línea. Proporcionan la capacidad de manipulación y análisis de la información almacenada en el data warehouse.

Minería de datos

Tecnología que aplica sofisticados y complejos algoritmos para el análisis de datos y expone información interesante para el análisis en la toma de decisiones.

Sistema operacional o transaccional

Aplicación utilizada para correr los negocios día a día usando datos en tiempo real.

Modelo dimensional

Técnica para hacer bases de datos simples y entendibles. Cuando una base de datos puede ser visualizada como un cubo de tres, cuatro o más dimensiones.

Métrica

Relativo a las medidas.

Dimensión

Conjunto de objetos o eventos en el mundo real.

Esquema estrella

Diagrama que esquematiza una base de datos dimensional y que contiene tablas de hechos y tablas de dimensiones.

Tabla de hechos

Tabla que contiene las métricas del negocio.

Tabla de dimensión

Tabla que contiene una dimensión del negocio.

Jerarquía

Atributos de una tabla que tienen una relación de uno a muchos.

Elementos de dimensión

Un elemento de dimensión corresponde a un nivel de sumarización en una dimensión. Definen las relaciones jerárquicas en una tabla de dimensiones.

Atributos de dimensión

Describen elementos de dimensión en términos que son familiares para los usuarios.

Proceso de negocio

Proceso primario operacional de una organización.

BIBLIOGRAFÍA

Bases de Datos: Modelos, Lenguajes, Diseño
Johnson, James L.
Ed. Oxford University Press
México, 1997

Building The Data Warehouse
Inmon, W.H.
Ed. John Wiley & Sons, USA 1993

Concepción y Diseño de Bases de Datos del Modelo E/R al Modelo Relacional
Miguel/Mario Piattini
Ed. Addison-Wesley Iberoamérica
Madrid 1993

Data Warehousing
Ed. SCN Education B.V.
1a. edición,
Germany, 2001

Diseño de Bases de Datos Relacionales
María Mercedes Marqués Andrés
Mayo, 2003
http://nuvol.uji.es/~mmarques/f47/apun/node1.html

Informix Guide for Designing Databases and Data Warehouses
Informix Software, Inc.
Ed. Prentice Hall, USA 2000

Introducción a los Sistemas de Bases de Datos
Date, C. J.
Ed. Pearson Educación
México, 2001

La integración de Información para la Mejor Toma de Decisiones: Data Warehousing
Harjinder S. Grill, Prakash C. Rao
Ed. Prentice Hall, México 1996

Redbrick Warehouse V. 5.0 Warehouse Administrator's. Guide for UNIX Platforms
Redbrick Systems Inc.
June 1997

Sistemas de Bases de Datos, Conceptos Fundamentales
Elmasri, Rames; Navathe, Shamkant B.
Ed. Addison-Wesley Iberoamericana
E.U.A, 1997

The Data Warehouse Toolkit, Practical Techniques for Building Dimensional Data Warehouse
Ralph Kimpball
Ed. John Wiley & Sons Inc. USA 1996

Nine Decisions in the Design of a Data Warehouse.
Ralph Kimpball
http://www.dbmsmag.com/
Marzo 2003

Practical Guide to Implementing Data warehouse
http://www.essentialstrategies.com/publications/datawarehouse/relmult.htm
David C. Hay
Marzo 2003