UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA

CENTRO DE INVESTIGACIÓN EN TECNOLOGÍAS DE INFORMACIÓN Y SISTEMAS

TESIS DE MAESTRÍA EN CIENCIAS COMPUTACIONALES

UTILIZACIÓN DE LA TECNOLOGÍA DATA WAREHOUSE EN INSTITUCIONES EDUCATIVAS

CASO DE ESTUDIO:

INSTITUTO HIDALGUENSE DE EDUCACIÓN MEDIA SUPERIOR Y SUPERIOR
(IHEMSyS)

TESIS DE MAESTRÍA EN CIENCIAS COMPUTACIONALES

AUTOR: Rene Cruz Guerrero.

DIRECTOR: Dr. Gustavo Núñez Esquer.

Pachuca de Soto, Hgo. Octubre de 2003.

Universidad Autónoma del Estado de Hidalgo
Centro de Investigación en Tecnologías de Información y Sistemas

Oficio No. CITIS-0628/2003

Ing. en Sistemas Comp. Rene Cruz Guerrero
P r e s e n t e ,

Por este conducto le comunico que el jurado asignado para la revisión de su trabajo de tesis titulado **"Utilización de la Tecnología Data Warehouse en Instituciones Educativas Caso de Estudio: IHEMSyS",** que para obtener el grado de Maestro en Ciencias Computacionales fue presentado por usted, ha tenido a bien, en reunión de sinodales, autorizarlo para impresión.

A continuación se integran las firmas de conformidad de los integrantes del Jurado:

PRESIDENTE: Dr. Joel Suárez Cansino.

PRIMER VOCAL: Dr. Guillermo Sánchez Díaz.

SECRETARIO: M. en C. María de los Ángeles Alonso L.

PRIMER SUPLENTE: M. en C. Félix A. Castro Espinoza.

SEGUNDO SUPLENTE: Dr. Roberto Hernández Gómez

A T E N T A M E N T E
"AMOR, ORDEN Y PROGRESO"
Pachuca, Hgo. a 25 de septiembre de 2003.

UNIVERSIDAD AUTONOMA DEL ESTADO DE HIDALGO C. I. T. I. S.

Dr. Roberto A. Hernández Gómez
Coordinador de la Maestría
en Ciencias Computacionales.

c.c.p. M. en D. Adolfo Pontigo Loyola.- Director de Control Escolar c.c.p
M. en C. Raúl García Rubio.-Director del ICBI c.c.p. Archivo / apl.

Carretera Pachuca-Tulancingo, Km. 4.5
6738,
Ciudad Universitaria C.P. 42084
Pachuca de Soto, Hidalgo, México.

Tel (01) (771) 71-72000 ext. 6735 y
fax (01) (771) 71-72109
email: citis@uaeh.reduaeh.mx

Resumen

El contenido del presente trabajo, se enfoca al uso de la tecnología Data Warehouse (DW) en instituciones educativas, considerando como caso de estudio al IHEMSyS (Instituto Hidalguense de Educación Media Superior y Superior). El proyecto surge por la necesidad de solucionar el problema de no contar con un sistema, que solucione las necesidades de tipo informativo dentro de la institución para ayudar en el soporte a la toma de decisiones. El sistema a implementar, se enfoca básicamente, a solucionar necesidades de tipo OLAP (On Line Analitic Process).

Las consultas respecto a múltiples dimensiones necesarias para la toma de decisiones, no se pueden realizar sobre las bases de datos de los sistemas operacionales con que cuenta la institución. Esto, debido a que los datos no son completamente históricos y la información no está estructurada de forma adecuada para este tipo de consultas.

Este trabajo incluye conceptos básicos que se relacionan con la tecnología DW, las metodologías existentes para la creación de un DW, así como la descripción de las diferentes etapas que intervienen en el proceso de desarrollo.

Contiene el análisis multidimensional para tres de los departamentos de la institución del caso de estudio (Académico, Recursos Financieros y Recursos Materiales), y el diseño multidimensional para el Departamento Académico (esquemas estrella, snowflake, tablas de hechos y dimensiones).

Se describe el proceso de conversión de los datos, incluyendo la definición de reglas para dicha conversión. Este proceso, se lleva a cabo con la ayuda de la herramienta Microsoft OLAP.

Se muestran ejemplos de resultados de la explotación de los datos, de los tipos de operaciones OLAP, haciendo consultas con el uso del lenguaje MDX, que es una extensión del lenguaje SQL y está incluido en las herramientas de Microsoft OLAP.

Por último, se Lace un análisis del proceso de implantación de un DW en el IHEMSyS, el cuál incluye un estudio de su pertinencia y factibilidad.

Índice

Capitulo 1

Introducción

En este capítulo se explica brevemente que es un Data Warehouse. Además, se describe en que consiste el proyecto, sus objetivos, alcances u limitantes.

1.1 INTRODUCCIÓN

En la actualidad, las tecnologías de los sistemas de información, se han utilizado principalmente para automatizar los procesos de tipo repetitivo, generándose con esto, los sistemas operacionales. Entendemos por sistemas operacionales, aquellos programas que resuelven las necesidades respecto al procesamiento de datos de alguna organización. En los sistemas operacionales, los conceptos más importantes son la actualización, el procesamiento u el tiempo de respuesta.

Sin embargo, además de las necesidades operacionales, las corporaciones también tienen necesidades informacionales. Estas tienen por objetivo obtener la información necesaria, que sirva de base para la toma de decisiones dentro de una organización. Las necesidades informacionales, se basan en gran medida en el análisis de una enorme cantidad de datos.

El dar solución a las necesidades de tipo informacional, utilizando las bases de datos de los sistemas operacionales, presenta varios problemas, debido a que para realizar consultas con alto grado de dificultad, existen diversas desventajas como: falta de visión global en la información, las bases de datos no contienen datos históricos, es decir; que especifiquen periodos de tiempo. Estas son algunas razones por las que surge la necesidad de una nueva tecnología, tal es el caso de los Data Warehouse (DW), considerada la tecnología que puede solucionar las deficiencias mencionadas.

12 CONCEPTO DE DATA WAREHOUSE

El concepto de Data Warehouse (almacén de datos), surge como una solución para obtener la información necesaria para la toma de decisiones, sin embargo, no es únicamente un almacén de datos, sino que su característica principal es la forma en como están estructurados esos datos, de modo que solucione cualquier tipo de consulta de manera eficiente u en el menor tiempo posible. A continuación, se muestran algunos conceptos de DW:

- Un DW, es un repositorio de información coleccionada desde múltiples fuentes, bajo un esquema uniforme u que usualmente reside en un solo sitio [1]. Los DW son construidos vía n proceso de limpieza, transformación, integración u carga de datos, este proceso se explica en el capítulo 2.
- Es una colección de datos orientados al sujeto, integrados, de tiempos variantes y no volátiles, que sirven de soporte para el proceso de toma de decisiones [2].
- Es un almacenamiento de información homogénea u fiable, en una estructura basada en la consulta y el tratamiento jerarquizado de la misma, en un entorno separado de los sistemas operacionales. [3].

Casi todos los conceptos coinciden, por lo que se puede resumir que un DW, es un almacén de datos que es manipulada separadamente de las bases de datos de una organización, la que se obtiene por la integración de información de diversos sistemas de aplicación; u soporta información procesada para proveer una plataforma sólida de datos históricos consolidados para su análisis.

Como se mencionó en el segundo concepto de DW, las características básicas que debe cumplir son: integrado, temático, no volátil e histórico, dichas características se explicarán en forma detallada en el Capítulo 2.

13 DATA MARTS

Un DW es una agrupación de unidades de información llamados *data marts*. Sin embargo, se considera que un *data mart* es una parte de un DW para un propósito específico (ejemplo, un *data mart* para el departamento de ventas en una empresa).

Un data martes una colección de datos, que es usada para el análisis de consultas dentro de una empresa en uno de sus departamentos o grupo de trabajo [1].

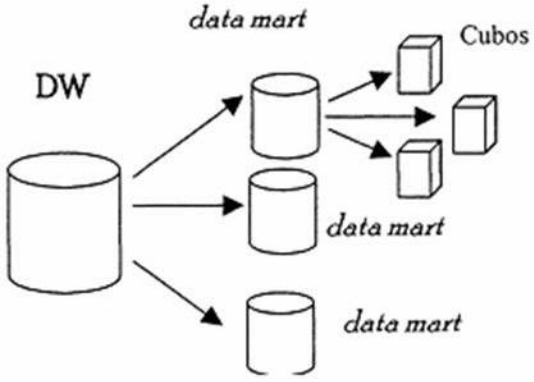


Figura 1.1 DW y Data marts

En la figura 1.1, se ilustra cómo un DW está conformado de varios *data marts*. A su vez, un *data mart* se compone de varios cubos de datos. Los *data marts* son preferidos frecuentemente por las empresas, como un primer paso para construir un DW.

1.4 LA TECNOLOGÍA DW PARA EL CASO DE ESTUDIO

En general, el buen desempeño de cualquier institución educativa depende, en gran parte, de las constantes decisiones, que se tomen a nivel directivo para corregir o mejorar los aspectos que están afectando su buen funcionamiento, principalmente en sus áreas más importantes, como por ejemplo el área académica.

Actualmente los DW se aplican en mayor porcentaje en los negocios, sin embargo, toda organización que controla grandes volúmenes de información u requiere de un soporte para la toma de decisiones, puede hacer uso de la tecnología DW.

El propósito del presente proyecto, es utilizar la tecnología DW en instituciones educativas. Respecto a la cantidad de información de los DW, su uso es aceptable cuando las instituciones manejen grandes cantidades. Por ejemplo, instituciones que están conformadas por diversos subsistemas, en donde cada subsistema está compuesto por diversos planteles educativos que pueden estar ubicados en diversas zonas o áreas geográficas, como es la situación del caso de estudio (IHEMSyS).

Por ejemplo, considere a un directivo de una institución educativa, que quiere saber el índice de aprovechamiento en los planteles educativos de determinada zona, comprendiendo el periodo de los últimos ocho semestres en el área de ciencias exactas. Esto sería bastante difícil si la información necesaria se encuentra en forma distribuida, además de que el tiempo de respuesta sería tardado.

En la actualidad, la mayoría de las instituciones educativas que requieren de un DW, no lo han implementado debido a las siguientes razones:

- Por tener una arquitectura que les exige la tecnología de *hardware* más actualizada, debido a los grandes volúmenes de información que manipula u el tiempo de respuesta requerido.
- Debido a que los beneficios de la inversión realizada al implementar el DW no se obtienen a corto plazo, únicamente las instituciones con posibilidades económicas, consideran la necesidad de su implantación.
- Para algunas instituciones, la tecnología DW, es un nuevo concepto de manipulación de datos.

1.5 OBJETIVOS

Objetivo general

Aplicar la tecnología Data Warehouse en instituciones educativas, con la finalidad de que cuenten con un sistema que resuelva sus necesidades de tipo informacional en procesamiento OLAP, en el ámbito de soporte para la toma de decisiones (caso de estudio IHEMSyS).

Objetivos particulares

- Explicar las metodologías u proceso de desarrollo de un DW.
- Hacer un análisis de la estructura organizacional del caso de estudio.
- Realizar el análisis multidimensional para los Departamentos Académico, Recursos Financieros u Recursos Materiales.
- Realizar el diseño multidimensional para el Departamento Académico.
- Definir u aplicar reglas de limpieza de datos.
- Realizar la conversión u carga de los datos del Departamento Académico
- Utilizar una herramienta para la explotación de datos, creando ejemplos de consultas.
- Determinar si es pertinente u factible implantar un DW en el IHEMSyS.

1.6 JUSTIFICACIÓN

Para poder realizar consultas que ayuden a la toma de decisiones en el IHEMSyS, no es conveniente utilizar las bases de datos que utilizan sus sistemas operacionales, debido a que la información de dichas bases de datos cambia continuamente. Una de las alternativas para poder solucionar estos problemas, es utilizar la tecnología Data Warehouse, debido a que la información de las bases de datos está separada de los sistemas operacionales.

En el proceso que se sigue para obtener la información que se requiere para la toma de decisiones, el usuario se encuentra con los siguientes problemas:

- Se tienen que realizar múltiples procesos para ejecutar consultas que involucran un alto grado de dificultad.
- No se tiene un almacenamiento histórico, pues la información de cada periodo está almacenada por separado.
- Los datos no están estructurados en un modelo multidimensional, para poder realizar consultas OLAP.

Debido a los aspectos mencionados anteriormente, el desarrollo de un Data Warehouse ayudará al IHEMSyS a obtener información que es utilizada frecuentemente por el personal directivo para la toma de decisiones.

1.7 BENEFICIOS

Con la utilización del DW se obtendrán los siguientes beneficios:

- Aprovechar el valor potencial de los recursos de información con que cuenta una institución educativa.
- Los usuarios del DW podrán acceder a una riqueza de información multidimensional, presentada coherentemente como una fuente única y confiable, disponible para ellos.
- Contar con una sola fuente de información, lo que elimina los múltiples procesos para realizar consultas, disminuyendo con esto, el tiempo de respuesta.

1.8 ALCANCES

Los alcances contemplados en el presente trabajo son los siguientes:

- En el desarrollo del DW, se pretenden cubrir todas las áreas o departamentos del IHEMSyS; sin embargo, para efecto del presente trabajo, comprenderá el análisis para los departamentos: Académico, Recursos Financieros y Recursos Materiales.
- El diseño se hará para el Departamento Académico, con el fin de implementarse como sistema piloto y posteriormente extenderse a los demás departamentos. La razón por la que se seleccionó el Departamento Académico, es por considerarse el de mayor importancia dentro del IHEMSyS.
- El sistema está destinado al personal de nivel directivo, es decir, está destinado para ser usado únicamente por el personal que dentro de la institución se considere que tiene facultad para tomar decisiones.
- El trabajo realizado llega hasta la creación del *data mart* del Departamento Académico, cargando los cubos de los indicadores obtenidos. Respecto a los datos de las tablas de hechos, se cargan los del año 2002, los datos de 4 años anteriores se simulan para efecto de analizar tendencias.
- Proponer una herramienta de explotación de datos u crear algunas consultas en dicha herramienta.

1.9 LIMITANTES

- La alimentación de los datos del depósito no se realiza en forma automática.
- La herramienta propuesta para la explotación de datos se llama MDX. Dicha herramienta permite realizar consultas por medio de sentencias más sencillas que SQL, pero no es lo suficientemente sencilla para ser utilizada por usuarios que desconocen sobre un lenguaje de consulta.

Capitulo 2

Características y Arquitectura de un DW

En este capítulo se explican de forma detallada, las características u la arquitectura de un DW. También se exponen las diferencias entre un DW u un Sistema de Procesamiento de Transacciones en Línea (OLTP).

2.1 CARACTERÍSTICAS DE UN DATA WAREHOUSE

Un DW tiene varias características que deben considerarse antes de su creación; entre las más importantes se puede mencionar las siguientes: datos integrados, temáticos, históricos y no volátiles. Cada una de estas características se explica en forma detallada a continuación:

Integrado

Un DW, es construido usualmente por la integración de datos de múltiples fuentes heterogéneas, ya sea desde bases de datos o archivos planos. Las técnicas de limpieza e integración de datos, son aplicadas para garantizar consistencia en convenciones de nombres uniformes, codificación de estructuras homologadas u atributos de medida, entreoíros [3].

Respecto a las convenciones de nombramiento, uno de los problemas que se presenta más comúnmente en el proceso de integración, es que el mismo elemento es frecuentemente referido por nombres diferentes en las diversas aplicaciones. Los datos almacenados en el DW deben integrarse en una estructura homóloga, por lo que las inconsistencias existentes en los diversos sistemas operacionales, deben ser eliminadas para que al momento de obtener los datos mediante la realización de consultas, los resultados sean confiables.

Otro de los aspectos importantes para lograr la integración de los datos, es traducir las diversas unidades de medida usadas en las diferentes bases de datos, en una medida estándar. Por ejemplo, evitar que en algunas bases de datos, la unidad de medida sean pesos y en otras dólares.

Temático

El ambiente operacional se diseña en base a funciones o actividades como: préstamos, facturación, depósitos, etc. Por ejemplo, una aplicación de facturación puede acceder a los datos sobre clientes, productos y precios. La base de datos combina estos elementos en una estructura que acomoda las necesidades de la aplicación.

En el ambiente DW, los datos se organizan en base a sujetos, tales como cliente, vendedor, producto, etc. Por ejemplo, para un fabricante, los sujetos pueden ser clientes, productos, proveedores y vendedores. Para una institución educativa pueden ser estudiantes, clases y profesores. Para un hospital pueden ser pacientes, personal médico, medicamentos, etc.

La diferencia entre la orientación a funciones en los sistemas operacionales y la orientación a sujetos en un DW, radica en el contenido de los datos a nivel de detalle [5]. En el DW, se excluye la información que no será usada para la toma de decisiones, mientras que la información utilizada por los sistemas operacionales, contiene datos requeridos para llevar a cabo sus diversas funciones o procesos.

Histórico

Los datos son almacenados para proveer información desde una perspectiva histórica. Por lo tanto, cada estructura clave en un DW contiene, implícita o explícitamente, un elemento de tiempo.

En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el presente. Por el contrario, la información de un DW sirve, entre otras cosas, para realizar análisis de tendencias por periodos de tiempo [6]. Por lo tanto, el DW se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

Como la información en el DW puede ser solicitada en cualquier momento, los datos encontrados en el depósito se llaman de tiempo variante". Los datos históricos son de poco uso en el procesamiento operacional. La información del almacén de datos, por el contrario, debe incluir los datos históricos para usarse en la evaluación de tendencias.

El tiempo variante se muestra de las siguientes formas:

- En un DW, la información representa los datos sobre un horizonte largo de tiempo (semestres, años). El horizonte de tiempo representado para el ambiente operacional, es mucho más corto (días, semanas).
- La segunda forma en la que se muestra el tiempo variante en el DW, está en la estructura de sus datos. Cada estructura clave en el DW contiene, implícita o explícitamente, un elemento de tiempo como día, semana, mes, etc. El elemento de tiempo está casi siempre concatenado al dato que lo requiere.
- La tercera forma en que se detecta el tiempo variante, es cuando la información del DW una vez almacenada correctamente, no puede ser actualizada.

No volátil

El almacén de información de un DW existe para ser leído, u no modificado. La información es por tanto permanente, por lo que la actualización del DW consiste en la incorporación de los últimos valores que tomaron las distintas variables contenidas, sin realizar ninguna modificación sobre lo que ya existía [3].

La información es útil sólo cuando es estable. Los datos de los Sistemas Operacionales cambian constantemente. Sin embargo, la perspectiva esencial para el análisis en la toma de decisiones, requiere una base de datos estable.

En los sistemas operacionales, la actualización (insertar, borrar u modificar) se hace por registro o por lotes de datos. Pero la manipulación de los datos que ocurre en el DW es mucho más simple, hay dos únicos tipos de operaciones: la carga inicial de datos u el acceso a los mismos. No hay actualización de datos (en el sentido de modificación) en el depósito, como una parte normal de procesamiento, sin embargo se pueden agregar datos nuevos correspondientes al último periodo.

2.2 COMPARACIÓN ENTRE DATA WAREHOUSE Y EL PROCESAMIENTO DE TRANSACCIONES EN LÍNEA (OLTP)

Los sistemas tradicionales de transacciones u las aplicaciones DW, utilizan conceptos diferentes, en cuanto a sus requerimientos de diseño u sus características de operación. Es de suma importancia comprender perfectamente sus diferencias, para evitar diseñar

un DW con algunas características similares a una aplicación de transacciones en línea (OLTP) m.

Es importante recordar que las aplicaciones de OLTP están organizadas para ejecutar las transacciones para las cuales fueron hechas, como por ejemplo: mover dinero entre cuentas, un cargo o abono, una devolución de inventario, etc. Por otro lado, un DW está organizado en base a conceptos (sujetos), como por ejemplo: clientes, facturas, productos, etc.

Otra diferencia radica en el tipo de usuarios, en el UW los usuarios por lo general son personal directivo, por lo que, el número de usuarios de un DW es menor al de un OLTP. Por lo tanto, es común encontrar que los sistemas transaccionales son accedidos por miles de usuarios simultáneamente, mientras que los DW sólo por cientos, debido a que estos son usados principalmente a nivel directivo [8]. A continuación, se muestra una tabla que resume las diferencias básicas entre sistemas OLTP y DW.

SISTEMA OLTP	**DW**
Predomina la actualización	Predomina la consulta
Maneja de 100 MB a GB	Maneja de 100 GB a TB
La actividad más importante es de tipo operativo (día a día)	La actividad más importante es el análisis y la decisión estratégica
Predomina el proceso puntual	Predomina el proceso masivo
Importancia del dato actual	Importancia del dato histórico
Importa el tiempo de respuesta de la Transacción	Importa el resultado de la consulta
Estructura relacional	Estructuras estrella, snowflake y constelación
Usuarios de perfiles medios o bajos	Usuarios de perfiles altos
Miles de usuarios	Cientos de usuarios
Explotación de la información relacionada parte operativa de cada aplicación	Explotación de toda la información interna y externa relacionada con cada parte del sistema

Tabla 2.1 Comparación entre DW y OLTP

2.3 ARQUITECTURA DE UN DW

La arquitectura de un DW, se puede clasificar de acuerdo al nivel de resumen de los datos que se proporcionan al usuario final, pudiendo ser de dos o tres capas. Ambas arquitecturas constan de un depósito donde se almacena la base de datos general, y

sistemas cliente que cuentan con Herramientas para explotar los ciatos del depósito. Sin embargo, la diferencia entre ambas, es que la arquitectura ele tres capas tiene en medio, entre las dos componentes anteriormente mencionadas, un grupo de *data marts* que tiene como función resumir los datos obtenidos del depósito por departamentos para proporcionarlos a los usuarios finales [9]. La arquitectura que se utilizó es la de tres capas, a continuación se explica cada una de ellas.

Arquitectura de dos capas

En la siguiente figura se muestra una arquitectura de dos capas:

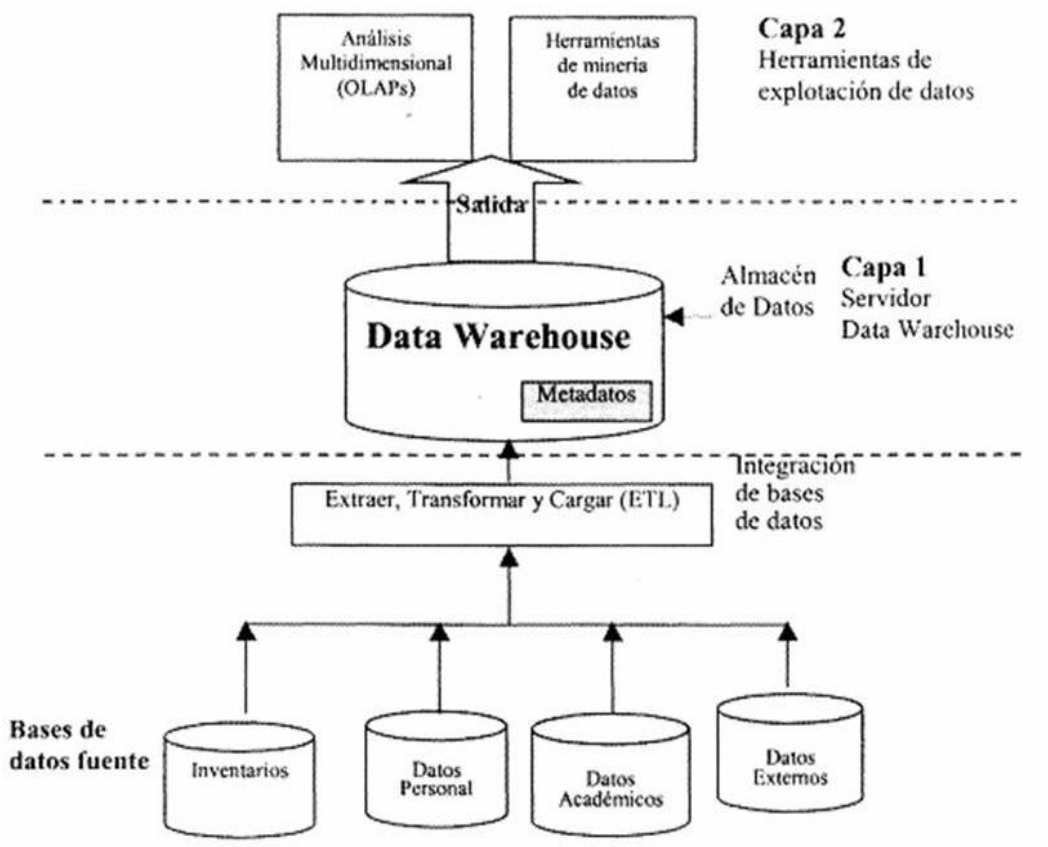


Figura 2.1 Arquitectura DW de *dos* capas

La primera capa consta de un servidor de base de datos warehouse, que es casi siempre un sistema de datos relacional Los datos almacenados en este servidor, son extraídos desde bases de datos operacionales o algunas fuentes externas, son extraídos usando interfaces de programas de aplicación conocidas como *gateways* [1]. Un *gateway* es soportado bajo un Sistema Manejador de Base de Datos y permite a los programas cliente, generar código en SQL para ser ejecutado en un servidor. Ejemplos de *gateway* incluyen ODBC (Open Database Connection) y OLE DB (Open Linking end Embedding for Databases) de Microsoft u JDBC (Java Database Connection).

La segunda capa es mi cliente, que puede contener herramientas con las que el usuario puede explotar los datos a partir del almacén warehouse, tales como herramientas OLAP o herramientas de minería de datos. Este tipo de herramientas, se explicarán más adelante en éste mismo capítulo,

Este tipo de arquitectura, se utiliza cuando el usuario final realiza consultas sobre datos no tan resumidos. Un ejemplo de donde más se utiliza este tipo de arquitectura es con la explotación de datos mediante herramientas de minería de datos. La ventaja de esta arquitectura es que se tiene acceso a todos los ciatos de la organización, sin embargo, la desventaja es que cuando se requiere consultar información específica, el proceso es más laborioso y tardado [10].

Arquitectura ele tres capas

A diferencia de la arquitectura de un DW de dos capas, esta arquitectura contiene otro nivel de resumen o sumarización de datos, que consiste en utilizar un *data mart* por departamento, como se muestra en la siguiente figura:

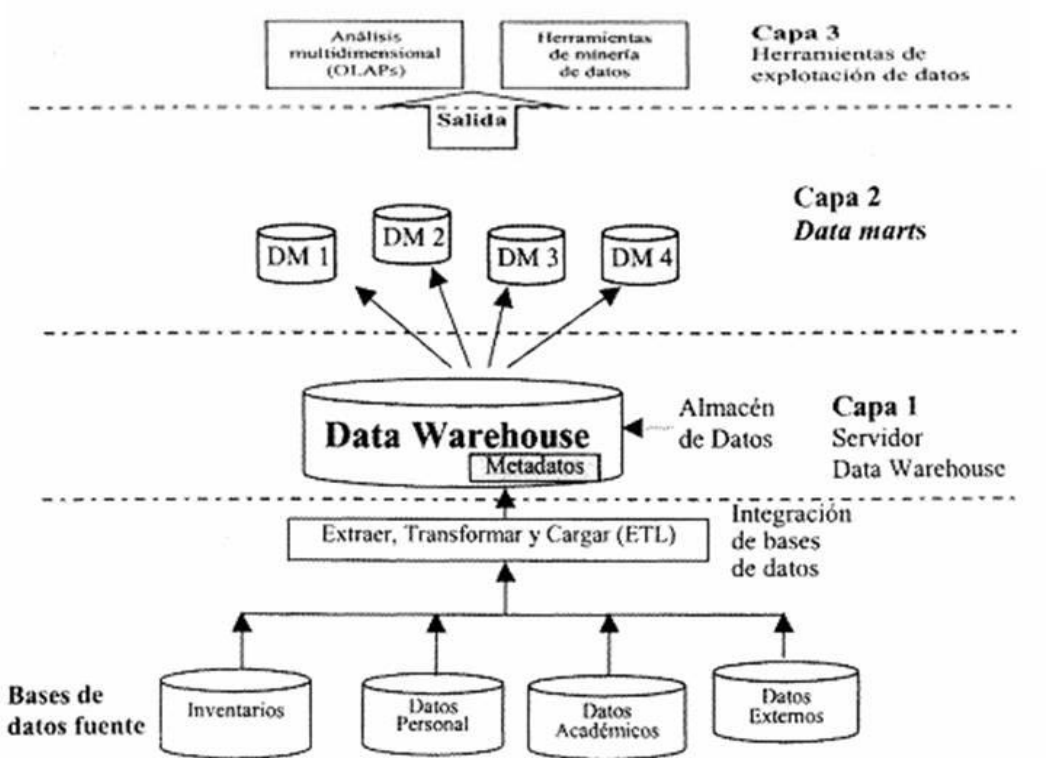


Figura 2,2 Arquitectura DW de tres capas

La segunda capa se compone por un conjunto de data mart, en el DW los datos son almacenados contó bases de datos relaciónales, sin embargo, en los data marts, los datos pueden ser almacenados como bases de datos multidimensionales,

El uso de *data marts*, ayuda al usuario a obtener datos clasificados por área, ayudando con esto a que los usuarios finales tengan acceso a información mas trabajada y que sólo es de su interés, además de que las consultas las pueden efectuar de forma mas rápida, es por eso que, ésta es la arquitectura DW más comúnmente utilizada.

La arquitectura DW a utilizar en nuestro caso de estudio, es la de tres capas, debido a que es la más adecuada cuando se requiere de realizar explotación de datos con análisis OLAP, por lo cual se creará un *data mart* departamento. A continuación, se explican los componentes y procesos de una arquitectura DW.

2.4 COMPONENTES DE UNA ARQUITECTURA DW

Como se pudo observar en la figura 2.2, los componentes principales son las bases de datos fuente, el almacén de datos, incluyendo los metadatos y las herramientas de explotación de los datos, considerando que existe un proceso para convertir los datos fuente a datos del almacén del DW. A continuación se explica cada uno de los elementos que intervienen en dicha arquitectura.

2.4.1 Bases de datos fuente

Son el conjunto de bases de datos de la organización que utilizan los sistemas operacionales, dichas bases de datos pueden ser homogéneas (contener formatos similares) o heterogéneas (contener diferentes formatos). La cantidad de trabajo a realizar y el tipo de herramientas a utilizar para la extracción de los datos, dependerá del grado de homogeneidad.

Además de las bases de datos de la organización, también puede llegar a ser necesario utilizar bases de datos externas que contienen datos que tienen alguna relación con la organización. Por ejemplo, en una institución educativa algún tipo de información perteneciente a la corporación, son los alumnos que cuentan con beca, en este caso, un tipo de información proporcionada por fuentes externas serían los datos sobre las instituciones externas que ofrecen dichas becas.

2.4.2 Metadatos

Los metadatos, son conocimientos acerca de datos, están compuestos por definiciones de los elementos de datos en el depósito, dichos metadatos son almacenados en un repositorio que maneja la herramienta [4], Las razones por las que son necesarios los metadatos en un DW, son las siguientes:

- Puede ser necesario involucrar diversos tipos de fuentes de datos, cada una de las cuales define sus datos en diferentes formas. La metadata provee una forma consistente para describir las estructuras de los datos.
- Es importante describir los cambios que son hechos a los datos cuando están siendo transformados. Algunas herramientas permiten crear metadatos durante el proceso de transformación.
- Los usuarios que utilizan los datos del DW, necesitan tener una clara explicación del manejo de los diferentes campos, hechos, niveles y dimensiones.

Una de las partes importantes de los metadatos, es realizar una descripción entendible de los datos al usuario, tales descripciones son almacenadas en la herramienta en una base de datos conocida también como repositorio de datos [11]. Casi todas las herramientas crean en forma automática la mayoría de los metadatos, sin embargo, dichas herramientas también permiten agregar metadatos en forma manual. Una de las formas más comunes de clasificar los metadatos es en técnicos y corporativos, a continuación se describen ambos.

Los metadatos técnicos, describen los datos en una forma clara y no ambigua. Esta es la clase de información que un programa de computadora necesita para procesar correctamente los datos, tales como:

- Nombre de campos, tablas y bases de datos.
- Nombre de nivel es, jerarquías.
- Nombre de dimensiones, cubos y bases de datos OLAR
- Tipos de datos.
- Longitud de campos.
- índices y relaciones.
- Llaves primarias y foráneas.

Los metadatos sobre la corporación describen datos a usuarios no técnicos, de esta forma ellos pueden entender más fácilmente la información con la que están trabajando y consultando. Los metadatos que se incluyen en este tipo son:

- Descripciones de campos, tablas y bases de datos.
- Descripciones de niveles, jerarquías, dimensiones, cubos y bases de datos OLAP.
- Descripciones de transformaciones.
- Referencias a metadatos técnicos.

Como se observa, además del tipo de metadatos manejados en un sistema operacional, un DW incluye metadatos sobre elementos OLAP, como son cubos, dimensiones, etc.
Á continuación, se muestran ejemplos de metadatos sobre cubos y dimensiones manejados por la herramienta Microsoft OLAP.

Figura 23 Ejemplo de metadatos sobre un cubo

En la figura 23, se muestra un ejemplo de metadatos sobre un cubo, los cuales almacenan información sobre cada una de las dimensiones que lo componen (ejemplo: turno, sexo, etc. los hechos que se manejan en ese cubo, las tablas origen, etc. Otro ejemplo es el de metadatos sobre la dimensión de un cubo, los cuales se muestran en la siguiente figura.

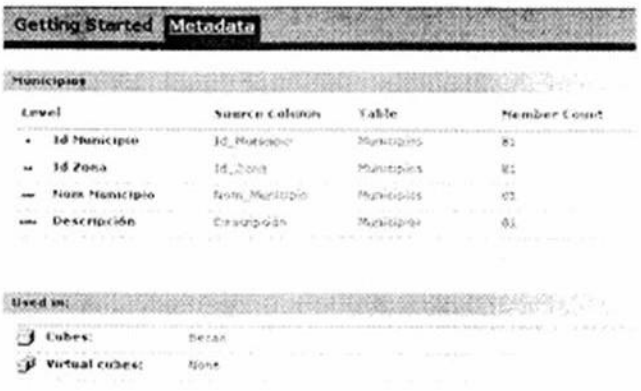


Figura 2.4 Ejemplo de metadatos sobre *una* dimensión

En esta figura se muestran metadatos sobre una dimensión, conteniendo la lista de atributos, cada uno con características como nombre de campo origen, tabla origen u número de miembros.

Como se mostró en la arquitectura de un DW, para poder transportar la información de las bases de datos operacionales al almacén del DW, es necesario realizar un proceso de conversión de datos, el cual se explica a continuación.

2.43 Proceso de conversión de datos

En la mayoría de los DW, ésta es la etapa que requiere más trabajo. El proceso de conversión, consiste en realizar los subprocesos de extracción, transformación y carga a partir de los datos que se obtienen de los sistemas operacionales de la empresa [12]. A continuación se muestran en la figura 2.D el orden de cada uno de los subprocesos:

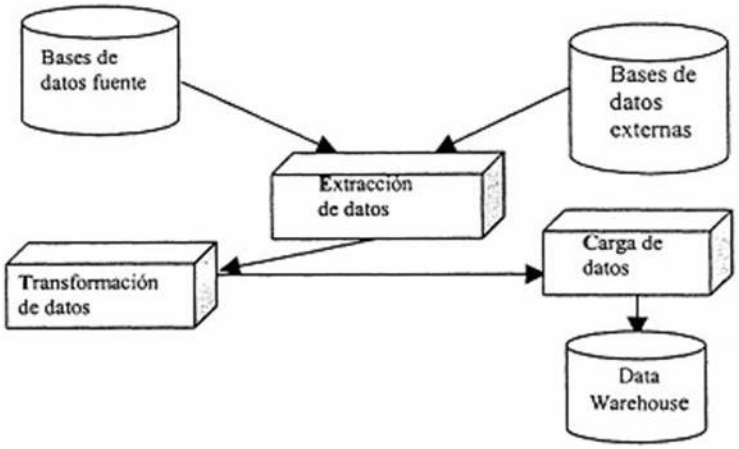


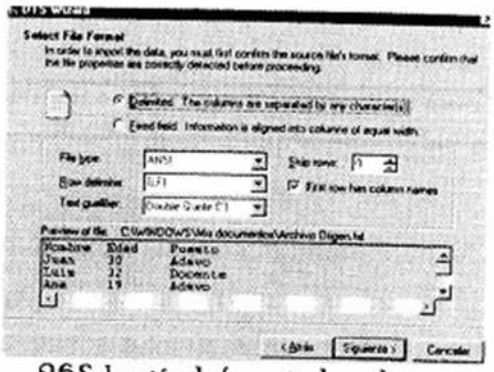
Figura 25 Proceso para convertir datos fuente

PROCESO DE EXTRACCIÓN

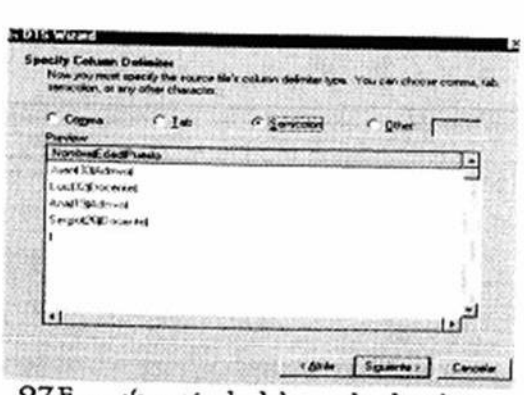
Es considerado el primero de los procesos de conversión de datos, el término extracción, se refiere al proceso de recuperar los datos requeridos desde archivos o bases de datos de los sistemas operacionales. La extracción se puede realizar con el uso de herramientas en el mercado o mediante programas desarrollados dentro de la propia corporación.

La mayoría de las herramientas de extracción, ofrecen la posibilidad de extraer datos desde diferentes formatos, incluyendo archivos con formato texto, hojas de cálculo y los formatos de los distintos tipos de manejadores de bases de datos (Oracle, Paradox, etc.). Dentro de los tipos de información fuente a obtener, el proceso de extracción a partir de bases de datos y hojas de cálculo, es más sencillo que la importación desde archivos con formato texto, sin embargo, este último tipo de extracción es raramente utilizado. El problema de importación desde archivos con formato texto, se debe a la forma en como están estructurados los datos, una de las tareas a resolver es detectar los diferentes separadores, ya sea de campos o de líneas [13].

MS OLAP, permite realizar la importación de archivos de formato texto. Para realizar este proceso, primero se debe especificar el origen de los datos, posteriormente se deben introducir algunos parámetros. En las figuras 2.6 y 2.7, se muestra el tipo de parámetros requeridos, entre los cuales se encuentra, el tipo de archivo, delimitador de filas y de columnas, etc.



2.6 Selección de formato de archivo



2.7 Especificación de delimitador de columna

En comparación con la extracción de los datos a partir de archivos de formato texto la extracción a partir de archivos de hojas de cálculo o tases de datos, se facilita debido a que los datos origen tienen una estructura definida. En la extracción con este tipo de formatos, la herramienta realiza tareas más sencillas, por ejemplo, la especificación de que campos u tablas se desean extraer.

En la figura 2.8, se muestra un ejemplo de como con la herramienta DTS de SQL se pueden extraer datos especificando la fuente u destino de los datos, u seleccionando las tablas que se desean extraer.

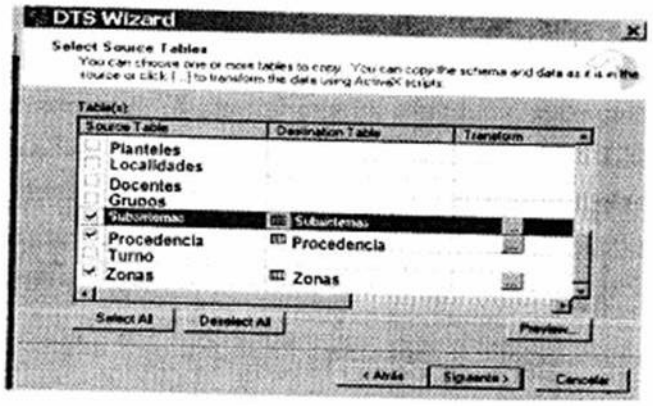


Figura 2.8 Proceso para convertir datos fuente

SQL Server, cuenta con una herramienta para la extracción de datos, pudiendo aceptar varios formatos de datos, entre los que se mencionan: Access, SQL, Excel, DBase. Paradox, Oracle, etc.

La mayoría de las herramientas existentes, también permiten documentar el proceso de extracción, ellas tienen provisiones para almacenar metadatos sobre la extracción, sin embargo, algunas de ellas son costosas, lo que provoca que algunas organizaciones prefieran tomar la decisión de escribir sus propios programas de extracción. Sin embargo esta alternativa es viable si los sistemas fuente son ambientes computacionales uniformes o heterogéneos (ejemplo: todos los datos residen en un mismo SMBD « hacen uso de algún sistema operacional).

Independientemente del tipo de datos a extraer, existen básicamente dos métodos de extracción, a saber, extracción a granel u extracción por replicación basada en cambios, a continuación se explica cada una de ellas.

Extracción a granel

Por lo general, consiste en recuperar la totalidad de los datos a partir de los sistemas fuente para ser cargados en el DW (ver figura 2.9). La recuperación de los datos se puede hacer, ya sea desde archivos de texto o desde bases de datos. Sin embargo, una característica importante de este tipo de extracción, es que no se pueden realizar transformaciones durante el proceso de recuperación de los datos [3].

La ventaja de este tipo de extracción, es la rapidez con que se efectúa el proceso, sin embargo, puede resultar más difícil realizar posteriormente los cambios necesarios debido a que se trabaja con parte de la información no necesaria. Por lo general, este tipo de extracción exige contar con conexión de red entre bases de datos fuente y bases de datos creadas. El trabajo de este tipo de extracción, se facilita cuando las bases de datos fuente son menos heterogéneas.

Replicación Basada en cambios

Sólo los datos que tienen que ser utilizados o actualizados en los sistemas de bases de datos fuente, son extraídos u cargados dentro del DW (ver figura 2.10). Este tipo de extracción le da menos importancia a la red (adecuado cuando los volúmenes de datos a ser transportados no son tan enormes), sin embargo, requiere de una selección o tablas o registros deben ser actualizados [3].

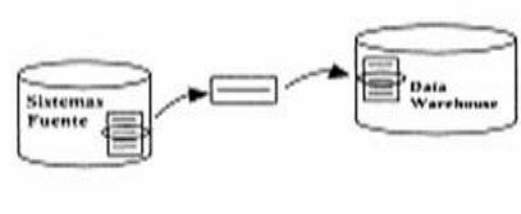


Figura 2.9 Replicación basada en cambios

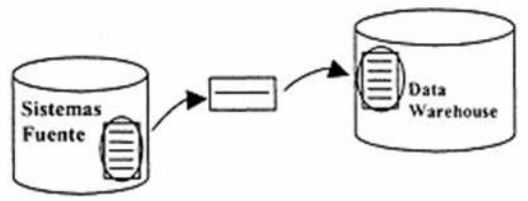


Figura 2.10 Extracciones a granel

La replicación basada en cambios, es el tipo de extracción de datos mas comúnmente utilizado, debido a que permite realizar transformaciones solo con los datos necesarios. La herramienta MS OLAP, soporta ambos tipos de extracción.

Para el caso de estudio, se utilizó el método de extracción por replicación basada en cambios, debido a que se permite trabajar sólo con los datos necesarios. En la figura 2.8 mostrada anteriormente, se observa como MS OLAP permite seleccionar los datos fuente y destino, u además permite realizar transformaciones sobre dichos datos.

Para poder seleccionar alguna ele las herramientas existentes en el mercado, es importante analizar sobre qué plataformas y tipos ele bases de datos pueden trabajar, debido a que la mayoría de las herramientas no pueden accesar a todos los tipos de fuentes de datos en todos los tipos de plataformas.

Además del tipo de herramientas mencionadas anteriormente, existen herramientas de middlware y conectividad, las que aunque son costosas, proveen acceso a los sistemas fuente en ambientes de cómputo heterogéneos, es decir, tener acceso a diferentes tipos de bases de datos, residentes en diferentes plataformas. Algunos ejemplos de middleware comercial y herramientas de conectividad son: IBM-Data Joiner, ORACLE-Transparent Gateway, SAS-SAS / Connect y SYBASE-Enterprice Connect

PROCESO DE TRANSFORMACIÓN

El proceso de transformación, permite realizar las modificaciones necesarias a los datos, acorde a las reglas o estándares que la corporación establece. La mayoría de las herramientas permiten efectuar algunas de las operaciones básicas de transformación de datos por medio de opciones que proporciona dicha herramienta, por ejemplo, la eliminación e inserción de campos y de tablas, etc.

A continuación se muestra la figura 2.11, donde se observa un ejemplo de cómo MS OLAF permite realizar operaciones básicas sobre campos como cambio de nombre, eliminación y agregación.

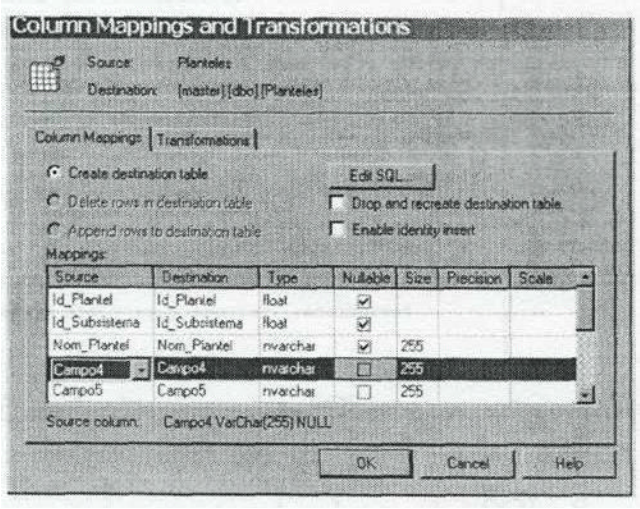


Figura 2,11 Ejemplo de operaciones de transformación

Como se observa en la figura, el usuario agrega los Id_Plantel, Id_Subsistema y Nom_Plantel, pudiendo realizar otras operaciones en la tabla destino, como crear campos o modificar el tipo o tamaño de los campos.

Otra de las formas en que las herramientas permiten realizar operaciones de transformación, es proporcionando ambientes de programación en determinado lenguaje, para que el usuario pueda introducir el código necesario u efectuar cierta operación, por ejemplo SQL.

Algunos de los ejemplos de operaciones de transformación de datos que no se consideran básicas, son la detección de llaves primarias incorrectas, las referencias incorrectas entre tablas, u la existencia de valores nulos, entre otros. A continuación, en la figura 2.12 se muestra como ejemplo, la sintaxis correcta de un conjunto de instrucciones SQL, las que pueden aplicarse para realizar algunas de las operaciones mencionadas:

OPERACIÓN	INSTRUCCIÓN SQL	SEMÁNTICA
VIOLACIÓN POR VALOR ÚNICO	SELECT*FROM<Tabla> GROUP BY <Atributo> HAVING COUNT(*) >1	<Tabla>.<Atributo>
EXISTENCIA DE VALORES NULOS	SELECT * FROM <Tabla> WHERE <Atributo> IS NULL	<Tabla>.<Atributo>
DIFERENTE DOMINIO	SELECT * FROM <Tabla> WHERE <Atributo> NOT IN <Especificación de dominio>	<Tabla>.<Atributo> NOT IN<Especificación de dominio >
VIOLACIÓN EN, LLAVE PRIMARIA	SELECT * FROM <Tabla> GROUP BY (<Atributo 1,..., atributo n>) HAVING COUNT(*) >1	< Tabla>.(<Atributo 1, ..., atributo n>)
VIOLACIÓN POR REFERENCIA	SELECT * FROM <Tabla> WHERE <Atributo> NOT IN (SELECT <Atributo en blanco> FROM <Table en blanco>.<Atributo en blanco>	<Tabla>. <Atributo> NOT IN
DIFERENTE FORMATO	SELECT APPLY(<exp regular>, <Atributo>) FROM <Tabla> WHERE APPLY(<exp regular>,<Atributo>)	TARGET APPLY(<exp regular>, <Atributo>) SOURCE APPLY(<exp regular>, <Atributo>

Figura 2.12 Ejemplos de instrucciones SQL para operaciones de transformación

La herramienta de transformación de datos de SQL Server, permite ejecutar instrucciones SQL para poder realizar transformaciones sobre las tablas de las bases de datos extraídas. En la figura 2.13 se muestra un sencillo ejemplo, donde se realiza la modificación del contenido de los datos del campo puesto de la tabla personal substituyendo la palabra'"Admvo" por "Administrativo".

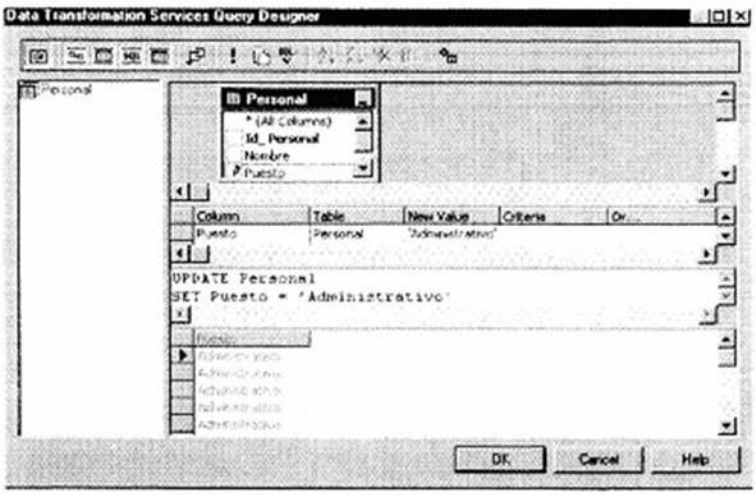
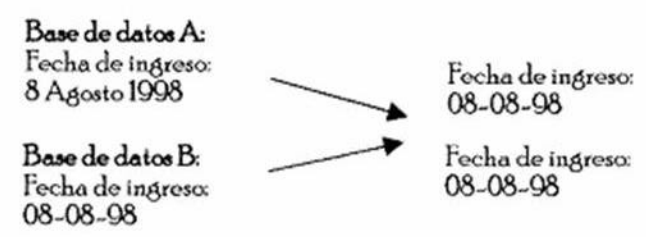


Figura 2.13 Ejemplo de transformación en ambiente de programación

Los tipos de operaciones para transformar los datos, van a depender del estado en que se encuentre la información en cada corporación, A continuación, se describen otros tipos de transformación que se requieren realizar con mayor frecuencia.

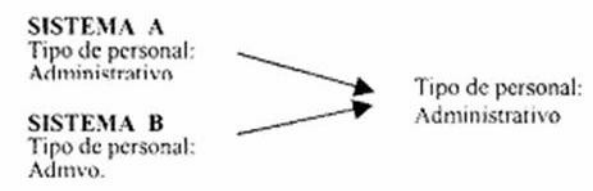
Estandarización de formatos

Los mismos datos pueden almacenarse con diferentes formatos y tipos de dato en distintas bases de datos. Estas diferencias pueden ser eliminadas durante el proceso de transformación, con la finalidad de obtener un formato o tipo de datos estándar [1]. A continuación se muestra un ejemplo de como se obtiene un formato estándar del dato fecha, después de obtenerlos de dos bases de datos fuente distintas.



Estandarización en el nombramiento de los datos

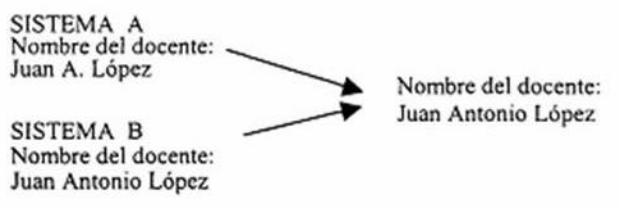
Es similar a la operación anterior, con la excepción de que en este caso se hace tratamiento sobre datos y no sobre formatos. Consiste en obtener una forma estándar de nombrar a un dato. A continuación se muestra un ejemplo sobre este tipo de operación.



Datos duplicados

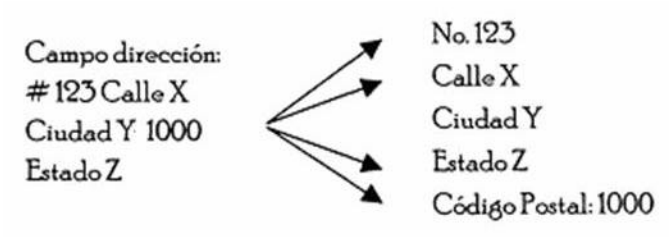
Existen situaciones en las cuales se encuentran repetidos algunos registros con distintos nombres en diferentes tablas o bases de datos. En esta situación, existe la necesidad de fusionar los registros duplicados para crear uno solo. Dependiendo de la cantidad de duplicaciones de un registro, es posible que algunos casos deban ser resueltos de forma manual y en otras aplicar alguna función para que lo resuelva.

A continuación, se muestra un ejemplo de cómo resolver la duplicación de dos registros con diferente nombre.



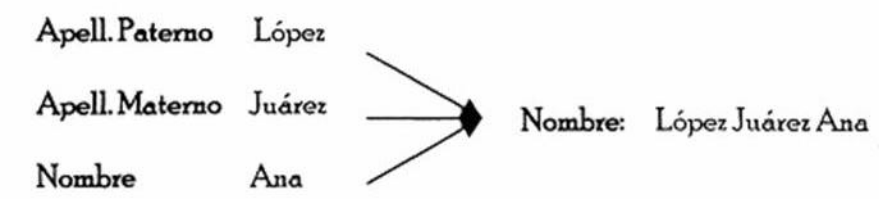
División de campos

Algunos datos en los sistemas fuente pueden necesitar ser divididos en uno o más campos. Por ejemplo, uno de los problemas de esta naturaleza mas comúnmente encontrados, es sobre las direcciones de las personas que han sido almacenadas en una sola línea de texto. Estos valores textuales deben ser divididos en algunos campos como: número y calle, colonia, ciudad, estado, etc. A continuación, se ilustra el ejemplo mencionado:



Integración de campos

Esta es la operación contraria a la anterior, es poco común, pero puede presentarse la necesidad de que a partir de dos o más campos deban ser integrados para componer uno solo. En el siguiente ejemplo, se muestra cómo a partir de tres campos se integra sólo uno. Si existen operaciones que el usuario no puede realizar fácilmente, ni con las opciones que brinda la herramienta, ni con el uso de un lenguaje, es necesario que éstas se efectúen manualmente **[12].**



CARGA DE LOS DATOS

Inserción sistemática de los datos y transformadores, en el almacenamiento físico del DW. En caso de aplicarse procesamiento OLAP, las tareas necesarias son: cargar tablas de dimensión, cargar tablas de hechos, reconstruir y regenerar índices. La mayoría de las herramientas proporcionan el ambiente para realizar todas estas operaciones.

Una vez cargada la información, los usuarios podrán realizar la explotación de los datos mediante la técnica que deseen. Microsoft GLAP, permite cargar la información correspondiente a un *data mart* medio de la creación de los cubos que lo integrarán. Para crear \\n cubo, es necesario cargar sus correspondientes tablas de hechos y dimensiones. A continuación, se muestra un ejemplo de cómo Microsoft OLAP, permite cargar las tablas de hechos 13 de dimensiones para crear cubos.

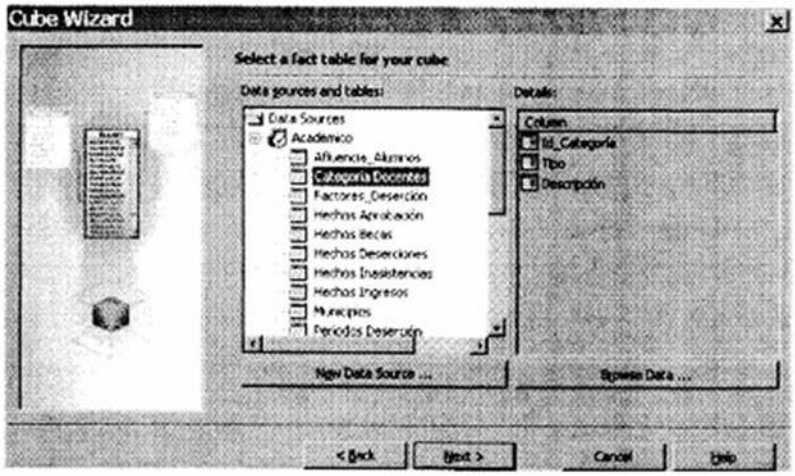


Figura 2.14 Ejemplo de carga de tablas de dimensión y de hechos con MS OLAP

2.4.4 Formas de explotan los datos

Como se observó en la figura 2,1 (arquitectura del DW), una vez que está creado el DW, los usuarios pueden explotar su información de dos formas, por medio del uso del modelado multidimensional, o usando herramientas de minería de datos, a continuación se explica cada una de ellas:

MODELO MULTIDIMENSIONAL

El modelo de datos entidad—relación, es comúnmente usado en el diseño de bases de datos relaciónales. Sin embargo, el modelo de datos comúnmente utilizado en un DW es el Modelo Multidimensional, este modelo permite representar los datos en términos ele dimensiones.

Las herramientas OLAP están basadas en un modelo de datos multidimensional, este modelo muestra los datos en una forma de cubo de datos. Un cubo de datos permite modelar datos y vistas en múltiples dimensiones [11],

A continuación, se muestra un ejemplo de un cubo de datos, el cuál Lace uso de las dimensiones: plantel, periodo u entidad.

Plantel =”Tula" Plantel =”Actopan” Plantel = “Reforma” Plantel =”Tulancingo"

	Entidad				Entidad				Entidad				Entidad			
Periodo	Alumnos	Aulas	Docentes	Grupos	Alumnos	Aulas	Docentes	Grupos	Alumnos	Aulas	Docentes	Grupos	Alumnos	Aulas	Docentes	Gps
96	990	21	50	18	820	17	45	18	900	17	45	18	800	16	38	17
97	1150	22	52	19	850	18	45	18	940	17	45	18	880	17	39	18
98	1100	12	55	18	830	18	47	18	980	18	46	19	820	18	40	17
99	1200	22	55	20	900	18	48	18	1000	18	46	19	870	18	40	18

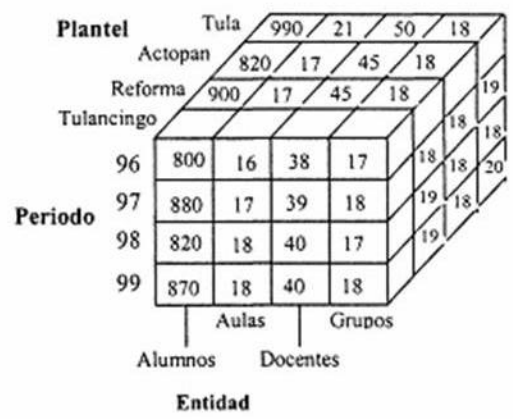


Figura 2.15 Cuto de datos 3D

Agregando la dimensión subsistema (cada uno puede contener un grupo de planteles) al cubo de datos anterior, se tiene un cubo de datos con cuatro dimensiones, como se muestra a continuación:

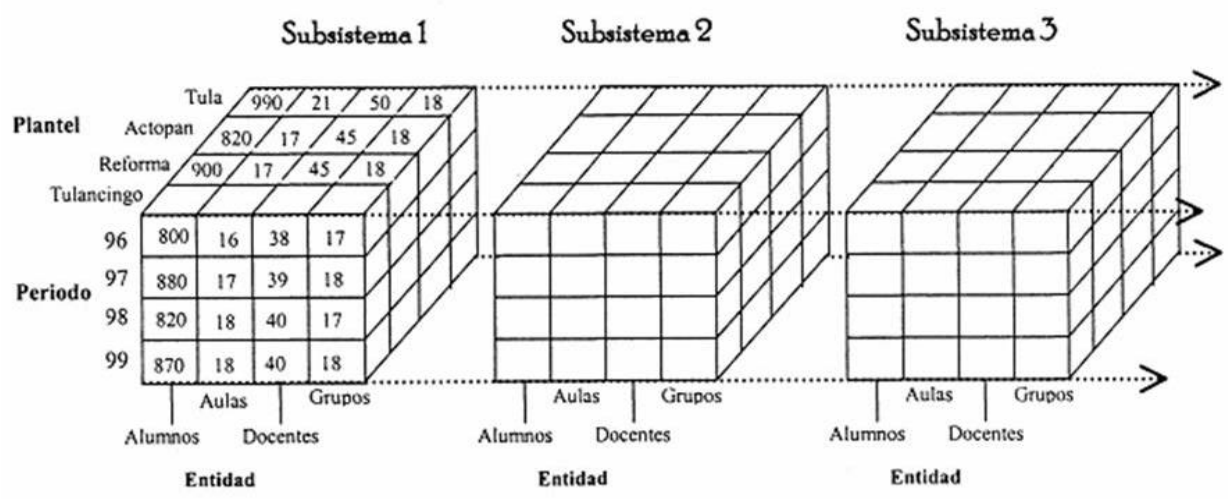


Figura 2.16 Cuto de datos 4D

OPERACIONES OLAP EN EL MODELO DE DATOS MULTIDIMENSIONAL

Las operaciones OLAP básicas que se pueden realizar en el modelo de datos multidimensional, a saber, Dice for Roll up, Drill Down y Pivot, se ilustran en la siguiente y se explican en la siguiente página.

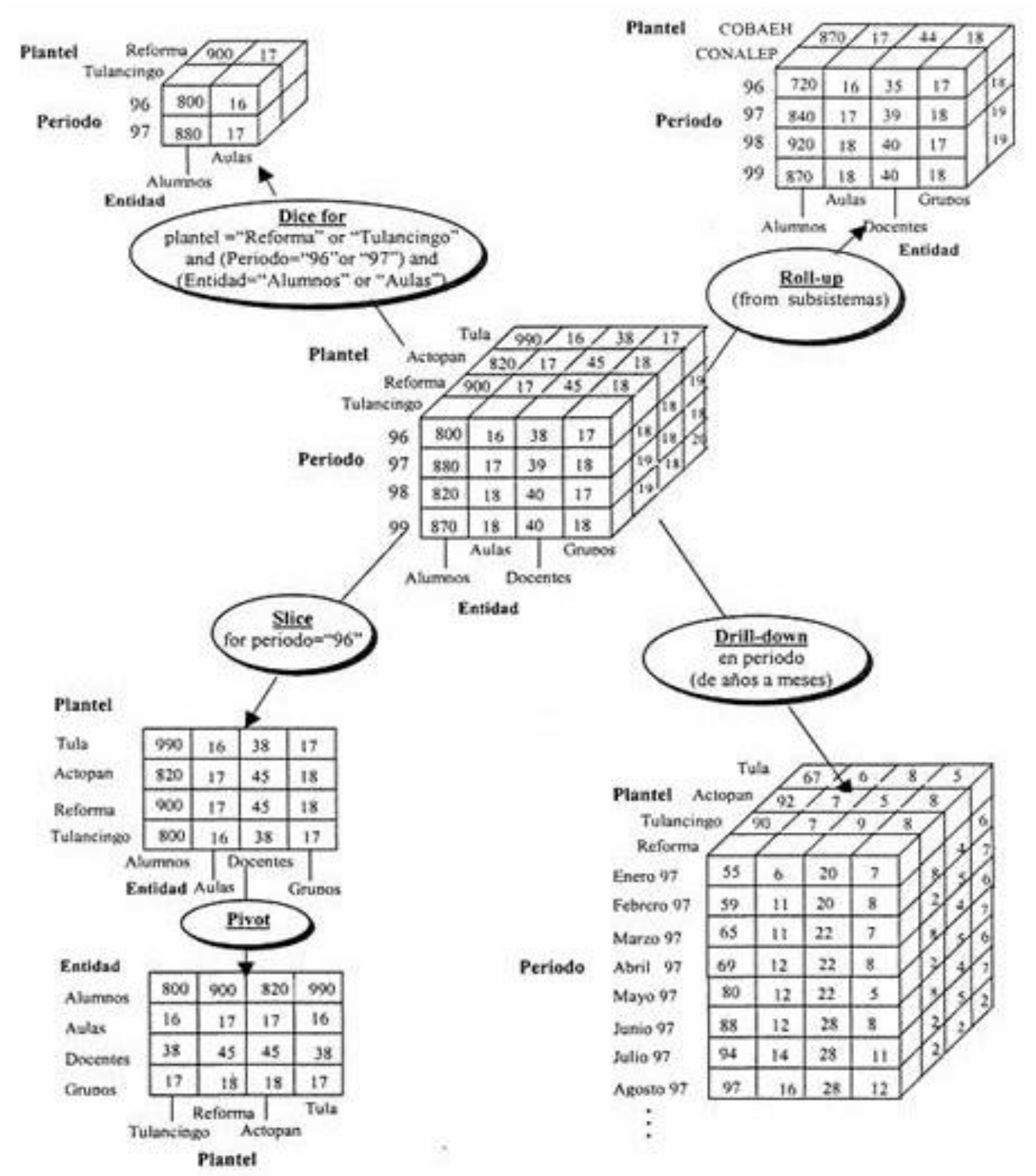


Figura 2.17 Operaciones OLAP

En el modelo de datos multidimensional, los datos son organizados en múltiples dimensiones, u a su vez cada dimensión contiene niveles de abstracción definidos por conceptos jerárquicos (por ejemplo, la dimensión periodo tiene la jerarquía año->semestre->mes). Con este tipo de organización se provee a los usuarios de flexibilidad para ver los datos desde diferentes perspectivas, las operaciones OLAP que se pueden realizar sobre unos cubos de datos son las siguientes:

Roll Up.- Esta operación (también llamada Drill Up) efectúa agregaciones sobre los cubos de datos OLAP existentes, ya sea escalando hacia arriba o por reducción de dimensiones. En la figura 2.17, se muestra el resultado de una operación roll up ejecutada sobre el cubo central por escalación hacia arriba, es decir, en lugar de obtener los datos por plantel se obtienen por subsistema (cada subsistema educativo contiene varios planteles). Cuando la operación roll up es ejecutada por reducción de dimensiones, una o más dimensiones son removidas del cubo de datos.

Drill Down.- Esta operación es la operación contraria a roll up. En esta se navega desde el menor hasta el mayor detalle, puede realizarse, ya sea por escalado hacia abajo o por adición de dimensiones. La figura 2.17 muestra el resultado de una operación drill down ejecutada en el cubo central por escalación hacia abajo, es decir, los datos son obtenidos a mayor nivel de detalle (de años a meses).

Slice and dice.- Esta operación ejecuta una selección en una dimensión del cubo de datos, obteniendo como resultado un subcubo. La figura 2.17, muestra una operación slice, donde los datos son obtenidos, tomando en cuenta el criterio de ano = 96. La operación dice define un subcubo al ejecutar una selección en dos o más dimensiones, la figura 2.17, muestra una operación dice tomando en cuenta el siguiente criterio de selección: (Plantel ="Reforma" or "Tulancingo") and (Periodo="96"or "97") and (Entidad =”Alumnos" or "Aulas").

Pivot (rotate).- Esta operación es de visualización, donde se rotan los ejes de datos en vistas en orden para proporcionar una presentación alternativa de los datos. La figura 2.17, muestra una operación pivot, donde los ejes plantel u elemento son rotados [12].

MINERÍA DE DATOS

Son procesos que se encargan de la explotación avanzada de los datos, se apoyan en Herramientas visuales de Software, estas herramientas son sofisticadas u específicas para la explotación de datos, especialmente en lo que respecta a la capacidad de detectar situaciones complejas u difíciles de detectar. Usan distintas técnicas: estadísticas, redes neuronales, de inteligencia artificial, etc. Permiten encontrar patrones de relación e incluso sugerir relaciones entre datos que no pueden ser detectados fácilmente por un analista [15].

En general, las tareas de la minería de datos pueden clasificarse en dos tipos: descriptivas y predictivas. Las tareas de minería descriptiva caracterizan las propiedades generales de los datos en la BD [29]. La minería predictiva ejecuta inferencias sobre los datos existentes en orden para realizar predicciones.

Los sistemas de minería de datos, también se pueden clasificar de acuerdo a los siguientes criterios: [1]

Tipos de bases de datos analizadas: De acuerdo al modelo de datos, se puede tener un sistema de minería de datos relacional, orientada a objetos, objeto-relacional, etc. De acuerdo a los tipos de datos, los sistemas de minería de datos pueden ser de texto, multimedia ó de páginas Web.

Tipos de conocimientos analizados: Los sistemas se clasifican según el nivel de abstracción de los conocimientos minados. Los conocimientos pueden ser generalizados (con un alto grado de abstracción), o conocimientos con múltiples niveles (considerando varios niveles de abstracción). Un sistema de minería de datos debe facilitar el descubrimiento de conocimientos con múltiples niveles de abstracción.

Tipos de técnicas utilizadas: Estas pueden ser descritas de acuerdo al a los métodos de análisis empleados. Dichas técnicas se pueden clasificar en técnicas de clasificación u de predicción. Algunas de las técnicas utilizadas para clasificación son: detección de desviaciones, agrupamientos (clustering), reglas secuénciales. Mientras que algunas de las técnicas de predicción son: Árboles de decisión, inducción neuronal, series temporales, etc., a continuación se describe un ejemplo da cada una.

Técnica clustering.- Contiene algoritmos que permiten agrupar un conjunto de objetos físicos o abstractos en clases de objetos similares. Algunos de sus métodos son: De particionamiento, jerárquicos, basados en densidad, etc.

Árboles de decisión inductivos.- Sus algoritmos manejan objetos o datos en estructuras de árbol u consisten en ir induciendo decisiones haciendo uso de recursividad del primer al último nivel del árbol (top down).

De acuerdo a las aplicaciones adaptadas: Los sistemas de minería de datos, también pueden ser clasificados de acuerdo al tipo de aplicación adaptada. Por ejemplo, pueden ser sistemas utilizados específicamente para finanzas, telecomunicaciones, etc. Por lo tanto, un sistema de minería de datos genérico, no debe enfocarse a un área de aplicación en particular.

Capítulo 3

Proceso de desarrollo del DW

3.1 INTRODUCCIÓN

En este capítulo, primero se describe que metodología se seleccionó para el desarrollo del DW, las metodologías existentes se explican a detalle en la sección 3.4. Posteriormente, en la sección 33, se explican las actividades que se llevaron a cabo en los distintos procesos que se siguieron para la creación del DW en el IHEMSyS, en esta sección, se mencionan algunos términos que se describen detalladamente en la sección 3.5, donde se explican los procesos de creación de un DW.

3.2 METODOLOGÍA UTILIZADA

Para el desarrollo de un DW, básicamente existen dos metodologías, las cuales son: Rapid Warehousing y Big Bang. Estas metodologías, se diferencian básicamente en que la primera consiste en crear los *data marts* del DW en forma evolutiva según las necesidades de la organización, y la segunda, en crear en forma paralela los *data marts* requeridos. Estas metodologías se explican más a detalle en la sección 3.4 de este capítulo.

La metodología utilizada para el caso de estudio es *Rapid Warehousing*, las razones por las que se seleccionó esta metodología, son las siguientes:

- Se requieren resultados a corto plazo, en el Departamento Académico del IHEMSyS.
- Se determinó crear un *data mart* para observar los resultados, y en base a la experiencia obtenida, continuar con la creación de los *data marts* restantes.

Como se utilizó la metodología *Rapid Warehousing*, el primer *data mart* a crear es del Departamento Académico, por considerarse el más importante en el IHEMSyS. Sin embargo, en este trabajo, se incluye también el análisis de los departamentos de Recursos Materiales y Recursos Financieros.

La razón por la cual se decidió realizar el análisis de los dos departamentos mencionados, fue porque después de entrevistar al personal directivo del IHEMSyS, consideraron que dichos departamentos seguían, en orden de importancia, al Académico.

33 PROCESO DE CREACIÓN DEL DW EN EL IHEMSyS

En esta sección, se describen las actividades que se llevaron a cabo en la creación del DW. Algunos términos nuevos se explican más adelante en la sección 3.5.

Actividades realizadas

Los procesos para el desarrollo de un DW, básicamente son: el análisis y diseño multidimensional y la conversión de los datos, dichos procesos se explican en forma detallada en la sección 3.5. Las actividades que se llevaron a cabo en la creación del DW, en cada uno de los procesos mencionados, se muestran en forma resumida en la figura 3.1.

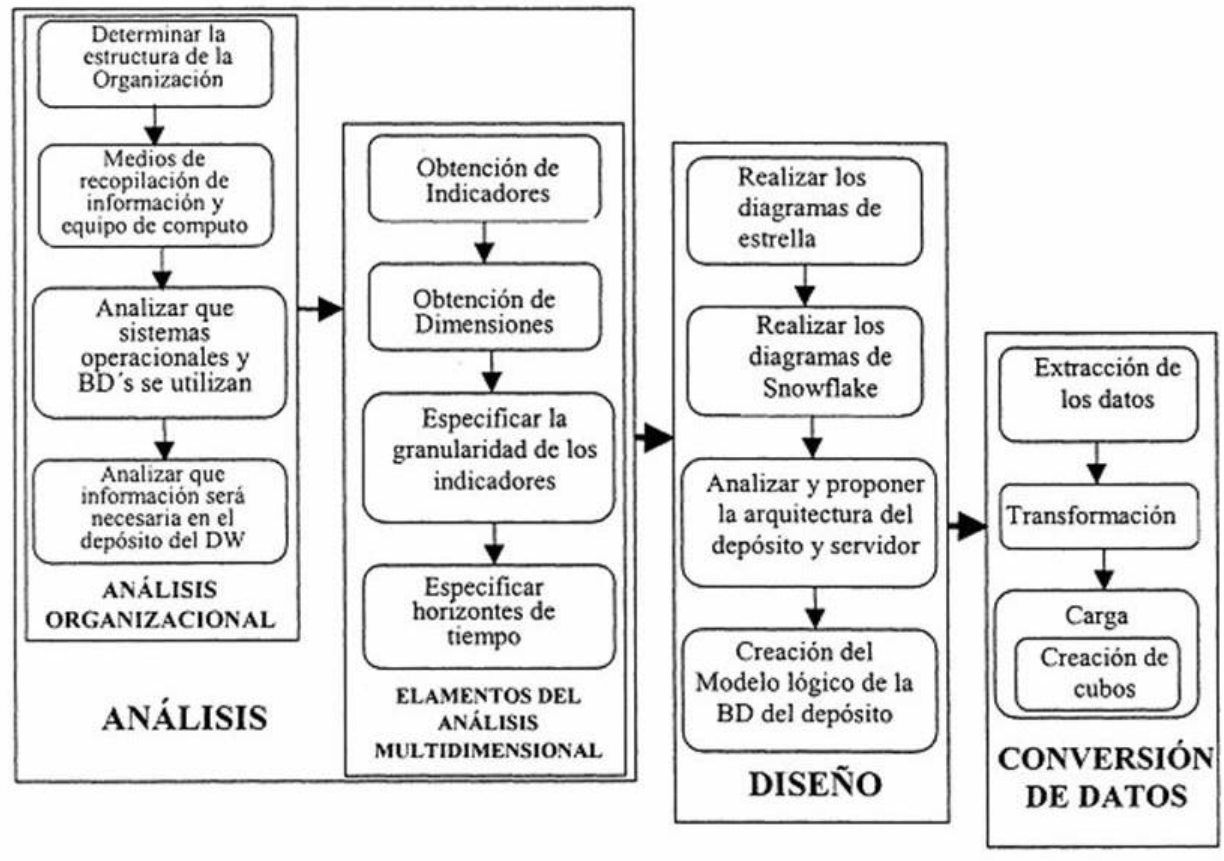


Figura 3.1 Actividades realizadas para la creación del DW en el IHEMSy S

A continuación, se explica cada una de las actividades realizadas, clasificándolas por cada proceso llevado a cabo.

ANÁLISIS

Como se muestra en la figura 3.1, el análisis se puede clasificar en el análisis de la organización y el análisis multidimensional, a continuación se explican las actividades que se realizaron en cada vino.

Análisis de la Organización

Fue la primera etapa que se realizó, u consistió en realizar las siguientes actividades:

Determinar la estructura de la organización.- Consistió en especificar que subsistemas y planteles forman parte del IHEMSyS, así como mostrar su jerarquía y la forma en que son dirigidos. También se detectó la matrícula de cada subsistema. Los resultados de esta actividad, se muestran a detalle en la sección 4.1.1del siguiente capítulo.

Medios de recopilación de información y equipo informático con que se trabaja.- En esta actividad, fue necesario detectar porque que medios se envía o intercambia la información entre los diferentes subsistemas u planteles con la Dirección General del IHEMSyS. Para esto fue necesario determinar los medios de comunicación u de almacenamiento. También se determinó el equipo de cómputo que se utiliza. Los resultados de esta actividad, se muestran en la sección 4.1.2 del siguiente capítulo.

Determinar que sistemas operacionales y bases de datos se utilizan.- Especificar que sistemas operacionales u bases de datos son utilizados en los subsistemas y planteles, así como especificar que formato manejan las diferentes bases de datos. Los resultados de ésta actividad, se muestran en la sección 4.13 del siguiente capítulo.

Analizarla información necesaria en el depósito del DW.-Tomando en cuenta las bases de datos existentes u el tipo de información requerida para el depósito del DW, en esta actividad se determinó que información es útil u como se obtiene. La finalidad es que éste tipo de información sea extraída u posteriormente almacenada en el depósito del DW. Los resultados de esta actividad, se muestran en la sección 4.1.4 del siguiente capítulo.

Análisis Multidimensional

Obtención de indicadores. Consistió en elaborar una lista de datos, respecto a los que se desea consultar la información almacenada. Esto se logró, de acuerdo a los resultados obtenidos después de aplicar un cuestionario al personal involucrado en la realización de consultas (directivos u personal técnico) u de hacer un análisis de la información manejada en las distintas bases de datos y formatos. De la misma manera se obtuvieron los elementos restantes del análisis multidimensional. Los resultados de esta actividad, se muestran en la sección 4.2.1 del capítulo 4.

Obtención de las dimensiones. Se elaboró una lista de tablas de dimensión, respecto a las cuales se analizan los indicadores obtenidos. Cada uno de dichos

indicadores se relaciona con un grupo de dimensiones. Todo esto se explica a detalle en la sección 4.2.2 del capítulo 4.

Especificar la granularidad. ~ En ésta tarea, se determinó a que nivel de detalle se desea analizar cada indicador, es decir, respecto a que tablas de dimensión se analizarán dichos indicadores. Los resultados correspondientes, se muestran en la sección 4.2.4 del capítulo 4.

Especificar horizontes de tiempo.- Se determinó respecto a que periodo se desea analizar cada uno de los indicadores obtenidos (mes, trimestre, semestre, año, etc.). Los resultados de esta actividad, se muestran en la sección 4.2.4 del capítulo 4.

DISEÑO MULTIDIMENSIONAL

Como se muestra en la figura 3.1, el proceso del diseño multidimensional, consistió en realizar las siguientes actividades:

Creación de diagramas de estrella.- En esta tarea fue necesario realizar los diagramas de los distintos indicadores, los cuales se crean en base a los resultados obtenidos en el análisis multidimensional, específicamente de acuerdo a la granularidad de los distintos indicadores. Estos diagramas, se muestran en la sección 5.1.1 del capítulo 5.

Creación de diagramas snowflake. Consistió en crear los diagramas de snowflake para los indicadores obtenidos en el análisis, es importante recordar que este tipo de diagramas son parecidos a los anteriores, a diferencia de que en estos las tablas de dimensión están normalizadas. Estos diagramas se muestran en la sección 5.1.2 del capítulo 5.

Determinar la arquitectura del depósito u del servidor. ~ Esta determinación, se realizó tomando en cuenta la cantidad de información que se va a manejar y a la forma en que se quieren concentrar los datos. El análisis multidimensional, da una idea de la cantidad de información se almacenará (magnitud del depósito), considerando el número de indicadores, de dimensiones y la periodicidad de los indicadores. Los resultados de esta actividad, se muestran en la sección 53 del capítulo 5.

Creación de un modelo lógico de la base de datos del depósito Este modelo se creó en base a los resultados obtenidos del análisis y diseño multidimensional. Es decir, tomando en cuenta las listas de indicadores y dimensiones recopiladas, se creó el modelo lógico que contiene todas las tablas de hechos necesarias y sus correspondientes dimensiones. Este modelo, servirá de base para saber que información es necesaria extraer de las bases de datos o archivos fuente. Los resultados de esta actividad, se muestran en la sección 5.4 del capítulo 5.

CONVERSIÓN DE DATOS

El proceso de conversión consistió en realizar las siguientes tareas:

Extracción de datos. Para llevar a cabo ésta tarea fue necesario auxiliarse de la herramienta Data Transformation Services (DTS) de MS OLAP. Está tarea consistió en extraer los datos a partir de bases de datos fuente y de archivos de Excel, la información sobre tablas no existentes fue necesario crearlas, por ejemplo, tablas sobre dimensión tiempo.
Transformación de datos.- Consistió en estandarizar los datos fuente obtenidos con la aplicación de una serie de reglas previamente establecidas. Para esta tarea, también se utilizó la herramienta DTS.

Carga de datos.- Esta actividad consistió en crear el data mart del departamento académico, cargando los diferentes cubos.

3.4 METODOLOGÍAS DW

Una metodología es un proceso detallado, a menudo especificado en secuencia de pasos que se deben seguir para lograr una meta (en nuestro caso, la creación de un Data Warehouse). Es importante recordar que un Data Warehouse no se puede adquirir, se tiene que construir siguiendo determinada metodología. En la actualidad, las metodologías de desarrollo de un DW están aún en proceso de maduración, en contraste con las metodologías existentes para el desarrollo de sistemas tradicionales.

Antes de describir las metodologías existentes, se describen los modelos de desarrollo que algunos autores consideran son importantes a tomar en cuenta para el uso de cierta metodología en el proceso de creación del DW.

Los desarrolladores que proponen estos modelos, consideran que la técnica a utilizar en la creación del DW, depende de hacia quién se enfoca como punto principal el desarrollo del mismo, puede ser hacia el manejo de datos, de metas o de usuarios [7]. Los modelos propuestos son: *"Data-Driven", "Goal-Dríven" y "User-Driven"*. A continuación se describe en forma general en qué consiste cada uno:

Data-Dríven

Este modelo considera que en un DW lo que se manejan son datos, a diferencia de los sistemas clásicos, en los que se manejan requerimientos; los cuales son el último aspecto a ser considerado en la toma de decisiones, considerando las necesidades de los usuarios en segundo término [18]. El modelo de datos consiste de pocas dimensiones y de grupos de hechos. La dimensión representa la estructura básica del diseño. Los hechos son basados en el tiempo y tienen poco nivel de granularidad.

Goal Driven

Este modelo considera que el proceso de desarrollo, gira en tomo a los objetivos y metas establecidas en un principio. Al contrario del modelo anterior, este contiene mas dimensiones y pocos hechos, los cuales son basados en el tiempo y tienen un bajo nivel de granularidad [18].

User Driven

Considera que el factor principal a considerar» son las necesidades de los usuarios, pues son quienes utilizarán finalmente el sistema El modelo consta de pocos hechos, los cuales tienen un nivel moderado de granularidad.

Independientemente de los modelos de desarrollo mencionados, las metodologías a seguir para el desarrollo del DW, dependen en gran parte del tamaño del DW a crear y de la prontitud con que se requiera el DW, A continuación, se hace una descripción general de las dos principales metodologías para el desarrollo de un DW, a saber la *"Big Bang" y la "Rapid Warehousing"*.

METODOLOGÍA BIG BANG

Esta metodología trata de resolver todos los problemas conocidos para crear un Data Warehouse de gran tamaño, antes de liberarlo para su evaluación y prueba [22]. El proceso de desarrollo consiste en crear en forma paralela los diversos *data marts* que componen el DW, como se muestra en la figura 3.2. Esto tiene como consecuencia, que los desabolladores requieren de toda la información que involucra los diversos departamentos desde el inicio de la construcción del DW y que el periodo para obtener resultados sea mayor.

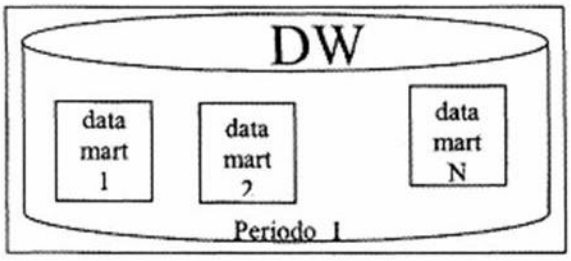


Figura 35 Metodología *Big Bang*

Las características principales de la metodología Bíg Bang son las siguientes:

- Se requiere ele más personal de desarrollo, debido a que se deben satisfacer los objetivos en cada departamento.
- Los resultados requeridos, por lo regular tardan, este tiempo depende del numero de *data marts* a desarrollar,
- Se tienen que Homologar en un principio las estructuras de datos de los distintos departamentos.

Este tipo ele metodología es menos utilizada, debido a que para la mayoría de las corporaciones la tecnología DW es algo nuevo, prefieren crear un programa piloto en uno de sus departamentos, para después de acuerdo a los resultados, decidir si se incorpora en sus departamentos restantes,

METODOLOGÍA RAPD WAREHOUSING

Esta es también conocida como metodología evolutiva o incremental y considera que la construcción e implantación de un DW es un proceso evolutivo, el cual consiste en crear rápidamente una parte de un DW con la integración de *data marts* (ver figura 33). Ésta metodología implica que cada vez que un *data mart* sea integrado, se debe operar simultáneamente en el DW (8). Así, con la integración en forma periódica de cada componente *data mart*, se integra la estructura final del DW,

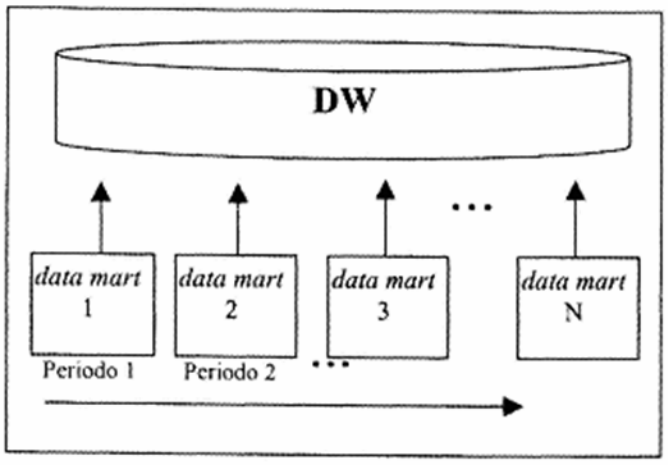


Figura 33 Metodología *Rapid Weurehousing*

Las características principales de ésta metodología son las siguientes:

- La solución de la parte del DW necesaria, requiere de poco tiempo.
- Permite adquirir experiencia en el proceso de creación. Con la implementación de los primeros data marts, se va adquiriendo experiencia para creaciones posteriores [23].
- Permite estandarizar las estructuras de los datos, respecto a los primeros data marts creados.
- Reduce la cantidad de errores en el proceso de desarrollo, debido a que involucra menos personal

Este tipo de metodología es la más usual pues requiere de que las corporaciones inviertan menos recursos que con la metodología *Big Bang*. Para aplicar ésta metodología, de ser necesario, se debe de realizar un análisis de cuál es el departamento más importante para crear en éste el primer *data mart*.

Independientemente de cualquiera de metodología que se utilice en la construcción del DW, los procesos básicos para la creación de sus *data marts*, son el modelado multidimensional u la conversión de datos.

35 PROCESOS PARA DESARROLLAR UN DW

Como se menciono en la sección 3.3, las etapas del desarrollo de un DW son el análisis multidimensional, el diseño multidimensional y la conversión de datos (ver figura 3.4). El análisis y diseño multidimensional, proporcionan los elementos necesarios para obtener el modelo lógico de una base de datos estándar, la cual se almacenara en el depósito del DW. La conversión de datos, consiste en realizar las tareas necesarias para depurar u homogeneizar los datos que se obtienen a partir de las bases de datos fuente.

Se considera que los subprocesos del modelado multidimensional son: el análisis y diseño multidimensional. El proceso de análisis multidimensional, involucra las siguientes tareas: láser un análisis organizacional, obtener los elementos necesarios para el modelado multidimensional que consiste en detectar indicadores, dimensiones, granularidad y horizonte de tiempo.

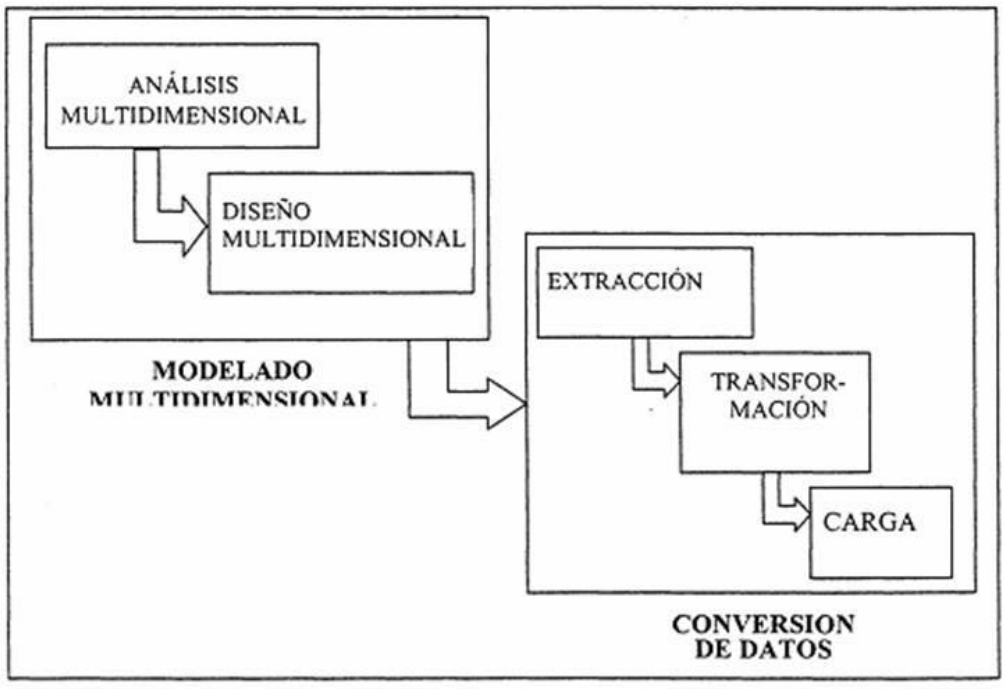


Figura 3.4 Proceso de desarrollo del DW

El diseño multidimensional, involucra las siguientes tareas: diseñar los diagramas de estrella, de snowflake, las tablas de dimensión, de hechos u especificar la arquitectura del depósito y del servidor del DW.

Los subprocesos principales de la conversión de datos son: la extracción, transformación y la carga o migración. La secuencia en el proceso de conversión de datos, primero se extraen los datos a partir de las tases de datos fuente [24]. Posteriormente se realiza la limpieza sobre las bases de datos extraídos, y por último, a partir de las base de datos limpia o depurada, se realiza la carga de los datos a los *data marts*.

A continuación se explican a detalle cada uno de los procesos de desarrollo del DW.

35.1 ANÁLISIS MULTIDIMENSIONAL

Como se mencionó en el capítulo 2, una de las principales formas de explotar los datos a partir del DW, es con el procesamiento analítico en línea, para lo cual es necesario realizar el análisis y diseño multidimensional.

Para poder realizar el análisis multidimensional, es necesario realizar las siguientes tareas:

- Determinar los indicadores.
- Determinar las dimensiones.
- Detectar las dependencias entre las dimensiones.
- Detectar las dependencias entre dimensiones e indicadores.
- Establecer la granularidad y definir horizontes de tiempo [16].

A continuación se explica qué significa y cómo se determina cada uno de los componentes anteriormente mencionados.

Determinar los indicadores

Los indicadores dentro de una organización, son aquellos datos que permiten realizar las mediciones necesarias. En general, representan resultados de los procesos realizados por los sistemas operacionales dentro de la organización; por lo general, son datos que obtienen valores numéricos.

La importancia de un indicador depende de que tan importante es la información que proporciona y la frecuencia con que se requiere su uso. Por ejemplo, en una institución educativa, un indicador importante que se consulta continuamente, es el total de deserciones [10].

Determinar las dimensiones

Se les llama dimensiones a las perspectivas de análisis de los indicadores [16], es decir, cuando se analiza un indicador (ejemplo: deserciones), generalmente se divide en dimensiones. Por ejemplo, las deserciones se pueden clasificar por subsistema y por plantel. Las dimensiones representan en general, entidades de la organización (planteles, alumnos, tiempo etc.).

Detectar la dependencia entre dimensiones

Las dimensiones pueden tener entre sí diferentes dependencias, por lo cual es importante su detección. Por ejemplo, un cliente pertenece a una determinada zona, un producto es provisto por varios proveedores, etc. Se debe considerar que las relaciones de dependencia entre las dimensiones seguramente ya existen en la bases de datos operacional, sin embargo, pueden existir variaciones en las mismas al incluirse la perspectiva histórica (adición de tiempo).

Por ejemplo, en la base de datos operacional, la dependencia entre cliente y zona puede estar definida por la relación mudaos a uno (n-1), significando que un cliente pertenece a una zona. Sin embargo, si el cliente a lo largo del tiempo puede ir variando su zona y en el análisis de los indicadores interesa diferenciar las diferentes zonas por las que él mismo fue pasando, la relación entre cliente y zona, es en realidad muchos a muchos (n-n).

Detectar la dependencia entre indicadores u dimensiones

Cuando se definen los indicadores y dimensiones, paralelamente se van identificando sus relaciones. Para cada indicador, se debe determinar por que dimensiones es analizable. Por ejemplo, el indicador deserciones depende de dimensiones tales como: tiempo, zona, subsistema, etc.

Establecer la granularidad y definir horizontes de tiempo

La granularidad de un indicador, representa el nivel de detalle por el cual será almacenado. El horizonte de tiempo de un indicador, representa el periodo que deseamos tener del mismo.

Para cada indicador se debe definir el o los diferentes horizontes de tiempo para los cuales se desea almacenar información. Para cada horizonte de tiempo definido, se debe determinar el nivel de granularidad necesario [25]. Es decir, lo que hay que determinar para cada indicador, es el subconjunto de dimensiones, mediante las cuales es analizable. Al establecer la granularidad, se debe de especificar el tipo de periodo (un mes, un semestre, un año, etc.).

35.2 DISEÑO MULTIDIMENSIONAL

Así como en los sistemas operacionales, el modelado de las bases de datos se puede hacer con el uso del modelo relacional u orientado a objetos, principalmente, el modelo de datos en un DW para procesamiento OLAP es el Modelo Multidimensional. La estructura básica de un modelo multidimensional está definida por dos elementos; a saber, tablas u esquemas [20].

Las tablas pueden ser de dos tipos: tablas de hechos, que contienen los valores, que por lo regular son resultados finales de determinado proceso u tablas de dimensiones, que contienen el detalle de los valores que se encuentran asociados a las tablas de hechos. Los esquemas están compuestos por agrupaciones de tablas. Estos pueden ser de dos tipos: esquemas estrella u esquemas snowflake [22]. A continuación se explican ambos conceptos.

Esquemas estrella

Son uno de los elementos principales del modelo multidimensional, su estructura base, está conformada por una tabla de hechos u un conjunto de tablas que la rodean radialmente (ver figura 3.5). El esquema estrella deriva su nombre debido a que su diagrama forma una estrella, con puntos radiales desde el centro. El centro de la estrella consiste de una tabla de hechos, y las puntas de la estrella son las tablas de dimensiones.

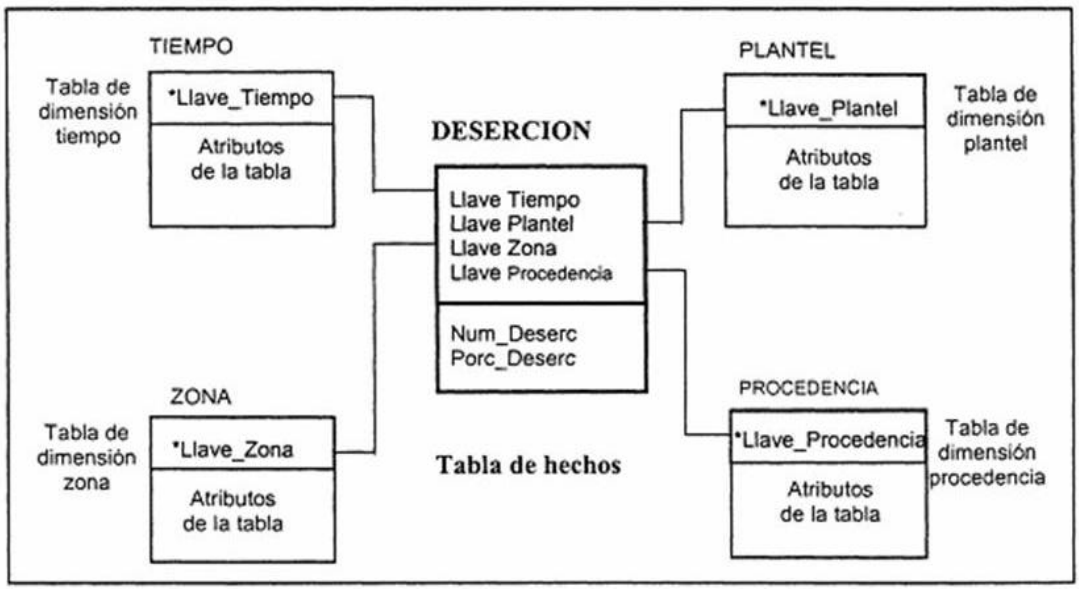


Figura 35 Esquema Estrella

En la figura 3.5, se muestra un esquema estrella que consta de cuatro tablas de dimensión no normalizadas (tiempo, plantel, zona u asignatura) u una tabla de hechos que contiene tanto las llaves principales de cada tabla de dimensión, así como, los indicadores sobre el número u porcentaje de deserciones.

Esquemas Snowflake: la diferencia de los esquemas snowflake, comparados con los esquemas estrella, es la estructura de las tablas de dimensiones (ver figura 3.6), debido a que en los esquemas snowflake están normalizadas. Cada tabla de dimensión, contiene la clave primaria y la llave foránea del nivel más cercano [16]. Por ejemplo, la dimensión plantel se divide en plantel y sistema.

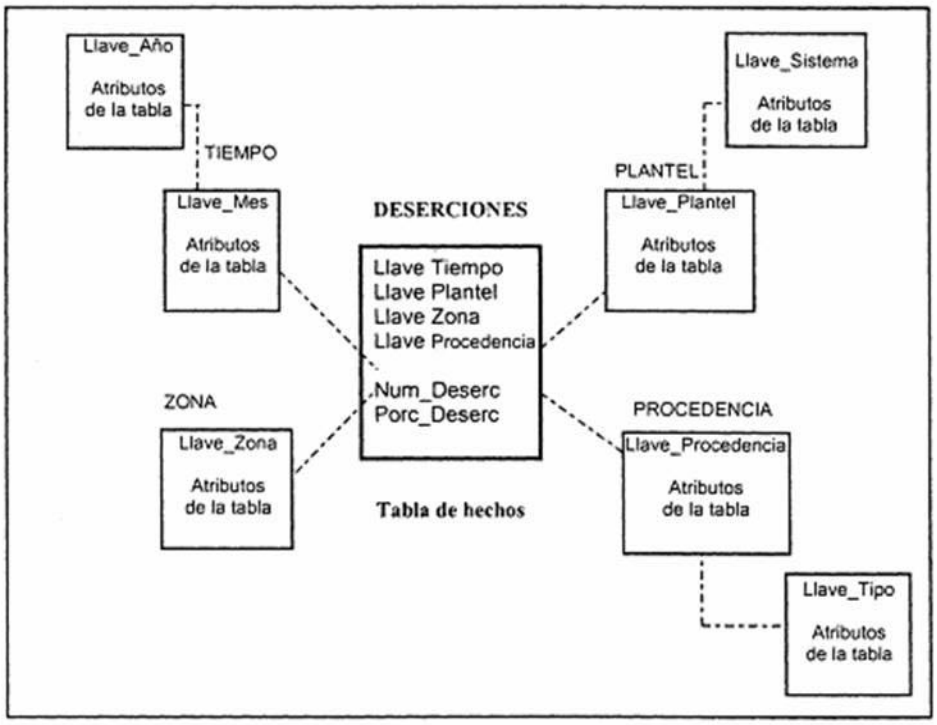


Figura 3.6 Esquema Snowflake

Tabla Fact o de hechos.

Es la tabla central en un esquema dimensional, donde se almacenan las mediciones numéricas, las cuales se hacen sobre el grano o unidad básica de la tabla. Como se mencionó anteriormente, el grano o la granularidad de la tabla queda determinada por el nivel de detalle que se almacenará en la tabla. Por ejemplo, en las deserciones en la dimensión tiempo, el grano puede ser la cantidad de alumnos que desertan en determinado semestre.

La clave de la tabla de hechos, recibe el nombre de clave compuesta o concatenada. Esta se forma de la composición (concatenación) de las llaves primarias de las tablas dimensionales a las que está unida. Así entonces, se distinguen dos tipos de columnas en una tabla de hechos: columna de hechos y de llaves. La columna de hechos es la que almacena alguna medida y la columna llave, forma parte de la clave compuesta de la tabla [21].

Tablas de dimensión

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

Una tabla de dimensión, está compuesta de una llave primaria que identifica unívocamente una fila en la tabla junto con un conjunto de atributos. Dependiendo del diseño del modelo multidimensional, puede existir una llave foránea que determina su relación con otra tabla de dimensión [10]. Para decidir si un campo de datos es un atributo o un hecho, se analiza la variación de la medida a través del tiempo: si varía continuamente seria un hecho, en caso contrario, un atributo.

Los atributos dimensionales, tienen un rol muy importante en un DW, pues son la fuente de información para las consultas. Esto significa que la base de datos será tan buena como lo sean los atributos dimensionales: mientras más descriptivos y manejables sean, mejorará la calidad del DW.

ARQUITECTURA DEL DEPÓSITO Y DEL SERVIDOR

Arquitectura del depósito

En la creación del DW, es importante considerar la estructura lógica y física de la base de datos del depósito, además de los servicios requeridos para operar y mantenerlo [30]. Esta elección determina la selección del servidor de hardware.

La plataforma física, puede centralizarse en una sola ubicación o distribuirse regionalmente. A continuación se explica cada una de estas.

a) Arquitectura central

Una forma de almacenar los datos de la organización, obtenidos de múltiples fuentes, tanto internas como externas, es consolidar la base de datos en un DW integrado. Este enfoque proporciona eficiencia, tanto en la potencia de procesamiento, como en los costos de soporte.

Un DW central (figura 3.7) está compuesto por una sola base de datos física, la cual contiene los datos de los diversos departamentos de una organización [13]. Los DW centrales se seleccionan, por lo general, cuando diversos departamentos tienen en común una buena parte de la información, así como un número grande de usuarios finales conectados a una red mediante un servidor local o a una computadora central (mainframe). Este es el tipo de arquitectura del depósito a utilizar en el

IHEMSyS, las razones se explican en el siguiente capítulo en la especificación de la arquitectura a utilizar.

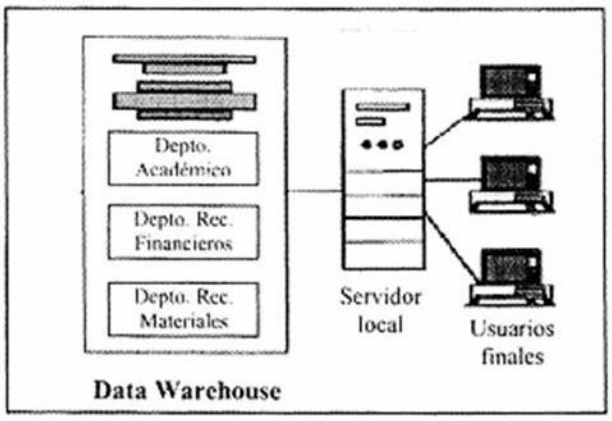


Figura. 3,7 Data Warehouse Central

En la figura 3.7, se muestra como en el almacén del DW, se encuentran concentrados los datos de los departamentos Académico, Recursos Financieros y Recursos Materiales.

b) Arquitectura distribuida

La arquitectura distribuida, consiste en dividir físicamente la información por áreas o departamentos. Los datos son consolidados lógicamente, pero se almacenan por separado, con el uso de esta arquitectura el costo de soporte es mayor [13].

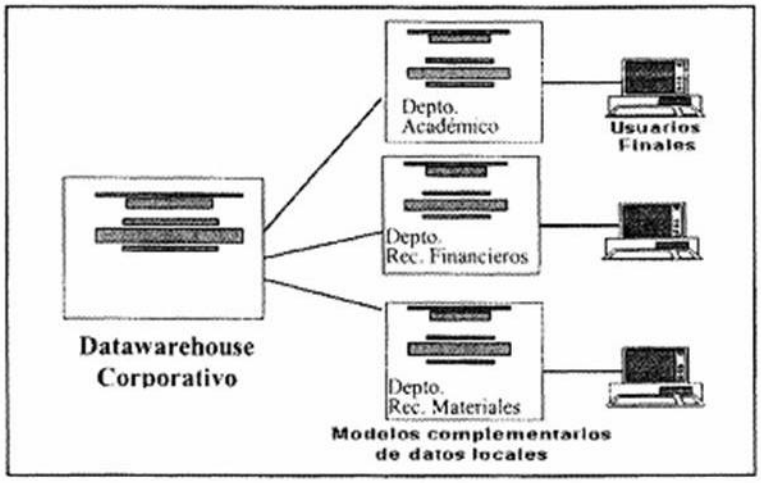


Figura 3.8 Data Warehouse Distribuido

En la figura 3,8, se muestra como la información del almacén principal, es distribuida en tres subsecciones, cada una de las cuales pertenece a un departamento, Académico, Recursos Financieros y Recursos Materiales.

Arquitectura del servidor

Después de decidir sobre una estructura de depósito distribuida o centralizada, es necesario considerar la forma en como los servidores cargarán y entregarán los datos. El tamaño de la implementación del DW y las necesidades de la organización, en cuanto a escalabilidad, influirá en la elección de la arquitectura del servidor. A continuación se describen los diferentes tipos:

a) Servidores de un solo procesador

Los servidores de un solo procesador son los mas fáciles de administrar, pero ofrecen limitada potencia de procesamiento y escalabilidad. Esta arquitectura provee seguridad en la información, debido a que las operaciones pueden cambiarse a un servidor de respaldo, si es que falla el servidor principal.

b) Multiprocesamiento simétrico

Los equipos de multiprocesamiento simétrico (Symmetric MultiProcessing - SMP), aumentan capacidad de procesamiento mediante la adición de procesadores, que comparten la memoria interna de los servidores y los dispositivos de almacenamiento secundario. En general, debe adquirirse el SMP con configuraciones mínimas (es decir, con dos procesadores) y escalar cuando sea necesario, justificando el crecimiento con las necesidades de procesamiento [26]. Este es el tipo de servidor que se propone para el IHEMSyS, lo cual se explica más a detalle en el siguiente capítulo.

c) Procesamiento en paralelo masivo

Un equipo de procesamiento en paralelo masivo (Massively Parallel Processing - MPP), conecta un conjunto de procesadores por medio de un enlace de banda ancha y de alta velocidad. Cada nodo es un servidor, completo con su propio procesador (posiblemente SMP) y memoria interna.

Para optimizar una arquitectura MPP, las aplicaciones y sistemas manejadores de bases de datos, deben ser diseñadas para ofrecer las facilidades del paralelismo. Esta arquitectura es ideal, cuando se requiere realizar búsqueda de información en grandes bases de datos [13].

En casi todos los DW, para la obtención de un modelo lógico estándar, se requiere agregar tablas que manejen datos sobre tiempo o periodo. Un ejemplo del caso de estudio en el Departamento Académico, es estandarizar los nombres de las tablas. Por ejemplo, si las bases de datos denominan de forma distinta a la tabla docentes, con nombres como catedráticos, profesores, etc., es necesario asignar un nombre único a todas.

Mas adelante, al final del capítulo 5, se muestran los cambios realizados para obtener la base de datos estándar del Departamento Académico del caso de estudio.

353 CONVERSIÓN DE DATOS

Una vez que se ka obtenido el modelo lógico que cumple con los requerimientos del modelado multidimensional realizado, el siguiente paso a realizar es la CONVERSIÓN DE DATOS. Para esta tarea se tienen que desarrollar los procedimientos de extracción y transformación sobre las bases de datos fuente, así como la carga al almacén del DW.

Los procesos de conversión, se explicaron en el capítulo 2. En el capítulo 6, se explica la forma en que se realizó la conversión de los datos en el caso de estudio, utilizando la herramienta OLAP Services.

Capitulo 4

Análisis Multidimensional

Como se mencionó en el capítulo anterior, en el proceso de desarrollo de un DW, la primera etapa es el análisis. Retomando la figura 3.1 que se mostró al inicio del capítulo anterior, a continuación se muestran las actividades que se realizaron, correspondientes a la etapa de análisis.

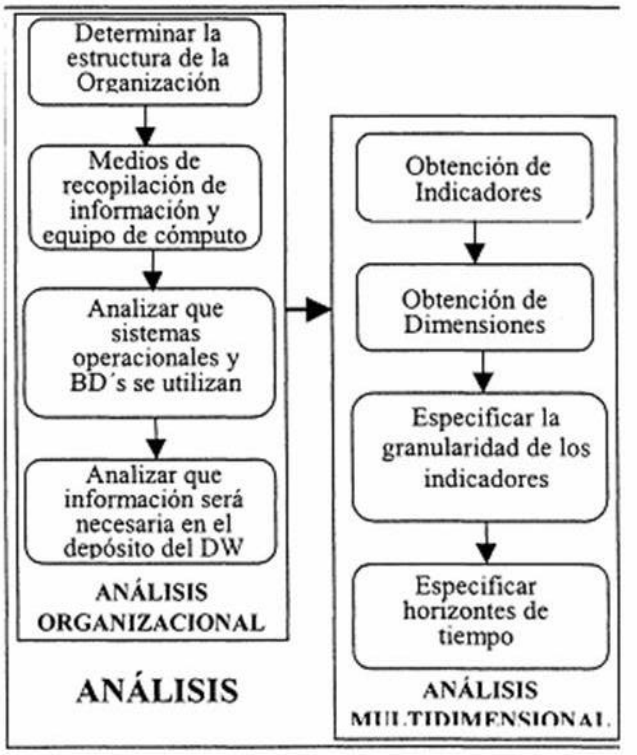


Figura 4.1 Actividades realizadas en el análisis

En la sección 3.3 del capítulo anterior, se explicó como se realizó cada una de las actividades mostradas en la figura 4.1. El procedimiento que se llevó a cato y los resultados de cada tina de las actividades se muestran en las siguientes secciones, en el orden en que se muestran en la figura 4.1.

41,1 Estructura de la Organización

El Instituto Hidalguense de Educación Media Superior y Superior (IHEMSyS) es un Organismo Publico Descentralizado del Gobierno del Estado, cuyo propósito es contribuir al desarrollo social, cien tilico y tecnológico de la entidad Esto, a través de la planeador*, coordinación u evaluación de las instituciones de educación media superior y superior» no autónomas del Estado, así como los organismos que se encargan de impartir capacitación y formación para el trabajo.

El IHEMSyS, esta conformado por dos niveles de educación, nivel medio superior y nivel superior. En la siguiente figura, se muestran los subsistemas que pertenecen a cada nivel

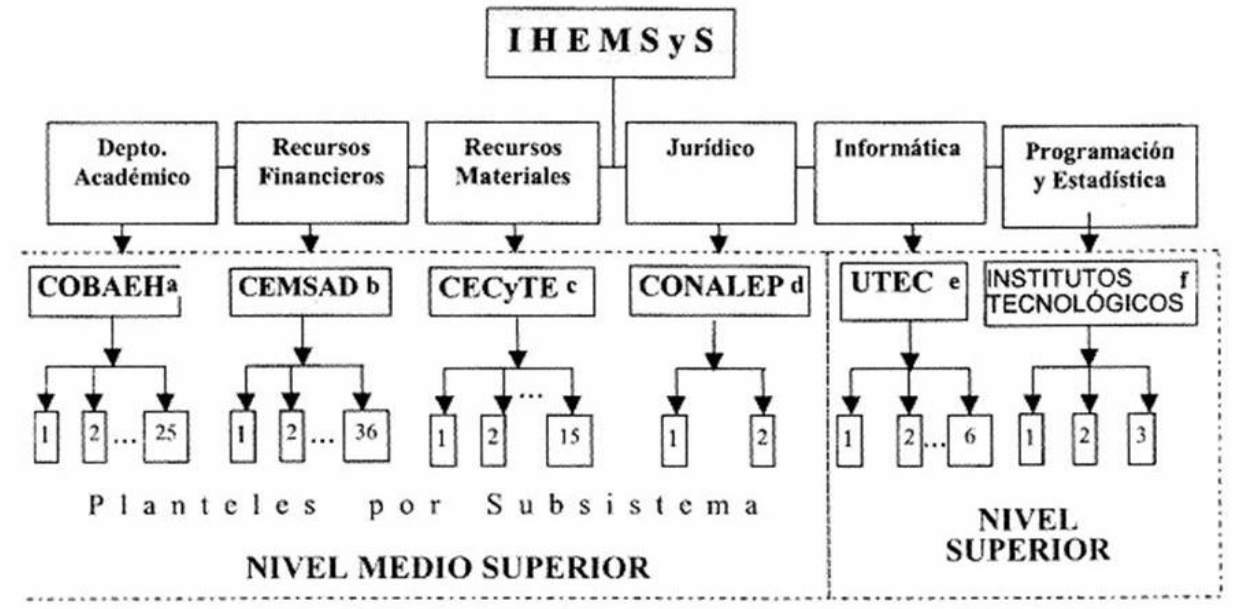


Figura 4,2 Estructura Organizacional del IHEMSyS

Los diferentes subsistemas, responden ce

a Colegio de Bachilleres del Estado de Hidalgo,
b Centro de Educación Media Superior a Distancia.
c Centro de Educación Científica u Tecnológica.
d Comisión Nacional de Educación Pública,
e Universidad Tecnológica,
f Instituto Tecnológico,

Como se muestra en la figura anterior, los subsistemas del nivel medio superior son, CONALEP, COBAEH, CEMSAD, y CECyTE Los subsistemas del nivel superior, son las Universidades Tecnológicas (UTECs) y los Institutos Tecnológicos Estatales.

En la figura, también se muestra el número de planteles por los que está constituido cada subsistema. Por ejemplo, el subsistema COBAEH está conformado por 25 planteles. Es importante mencionar que el IHEMSyS, también clasifica a los planteles de los diversos subsistemas por zonas geográficas, lo cual ayuda a detectar situaciones que se presenten por región o zona.

La matrícula de cada subsistema basta diciembre de 2002, fue la siguiente: COBAEH 14,200 alumnos, CECyTE 8,300, CONALEP 3,550, CEMSAD 9,800, UTECs 4,800 e Institutos Tecnológicos 1,700, sumando un total aproximado de 43,000 alumnos.

Actualmente, los planteles se ubican en cinco diferentes zonas, por lo que los planteles de diferentes subsistemas pueden estar localizados en una misma zona. Por ejemplo, dos planteles del subsistema CECyTE y uno de COBAEH. El IHEMSyS coordina a los diversos subsistemas a través de sus distintos departamentos, los cuales también se muestran en la figura 4.2, a saber: Académico, Recursos Financieros, Recursos Materiales, Jurídico, Programación y Estadística e Informática. La Dirección General de cada subsistema coordina a sus correspondientes planteles. Es importante mencionar que las direcciones generales de cada subsistema, se encuentran geográficamente distribuidas.

4.1.2 Medios de recopilación de información y equipo informático con que se trabaja

Una vez analizada la estructura de la organización, el siguiente paso realizado fue, analizar por que medios de almacenamiento se intercambia la información entre las dependencias que conforman el IHEMSyS, así como verificar con que equipo de cómputo se trabaja.

Se detectó que no existe una red informática que comunique los diversos subsistemas del IHEMSyS. La información la procesa y almacena cada uno de los planteles y subsistemas. Respecto a la forma de como se recopila la información, la mayoría de los planteles cuentan con el servicio de Internet, para enviar su información por correo electrónico a las oficinas centrales de su subsistema correspondiente. Sin embargo, tanto los planteles, como la Dirección General de cada subsistema y la Dirección General del IHEMSyS, cuentan con una red local que trabaja sobre la plataforma Windows NT.

Los planteles que no cuentan con servicio de Internet envían su información mediante unidad de CD ROM Dichos planteles, que por lo general son los más marginados geográficamente, de los cuales 4 pertenecen al COBAEH, 3 al CECyTE y 12 al CEMSAD.

El hecho de no contar con una red que comunique a todos los planteles y subsistemas del IHEMSyS, ocasiona algunos problemas cuando se requiere de un informe global. Por ejemplo, si se requiere saber el número de deserciones en todo el sistema, es necesario solicitar la información a cada subsistema, los cuales a su vez, lo solicita a sus

correspondientes planteles, provocando retrasos en el tiempo de respuesta. Esta situación afecta en los retrasos de los programas establecidos en los calendarios de actividades del IHEMSyS. Además, provoca que en ocasiones la toma de decisiones no se realice oportunamente.

En lo que se refiere al equipo de cómputo, en los planteles, se cuenta con el 40% de equipos *Pentium II, 40% Pentium III y 20% Pentium IV, en la Dirección General de cada subsistema el 10% de los equipos son Pentium II, e/40% Pentium III y el 50% Pentium /V. En las oficinas del IHEMSyS el 40% de las máquinas son Pentium IIIy el 60% Pentium IV.*

4.13 Sistemas operacionales y Bases de datos utilizadas

Una de las tareas de mayor importancia en el análisis, es investigar que sistemas y bases de datos están siendo utilizados actualmente por la organización, debido a que pueden manejar información necesaria para el depósito del DW.

Sistemas Operacionales utilizados

Respecto a los sistemas operacionales que se utilizan actualmente, en los planteles de los distintos subsistemas, únicamente el Departamento Académico cuenta con un sistema de control escolar y los departamentos restantes manejan su información en hojas de cálculo de Excel, respecto a la direcciones de los diferentes subsistemas, en el Departamento Académico todos cuentan con un Sistema de Control Escolar que controla la información de sus planteles y un sistema para generar reactivos de exámenes; dichos sistemas fueron creados por los mismos subsistemas.

En el Departamento de Recursos Financieros utilizan el Sistema de Contabilidad Microsip, el cual fue comprado; y por último, en el Departamento de Recursos Materiales se maneja un sistema de inventarios desarrollado por los subsistemas.

Respecto a las oficinas del IHEMSyS, en el Departamento Académico cuentan con un sistema que da seguimiento a alumnos egresados, en el Departamento de Recursos Financieros cuentan con un sistema de Contabilidad y en el de Recursos Materiales manejan la información en archivos de Excel

Bases de datos utilizadas

La información y procesos que se maneja en los diversos subsistemas en los Departamentos de Recursos Financieros y Recursos Materiales es similar. Por ejemplo, en el Departamento de Recursos Materiales los datos requeridos para controlar los inventarios de los equipos son los mismos. Sin embargo, en el Departamento Académico aunque los procesos que se efectúan son los mismos (inscripciones, cálculo de promedios, etc.), la información que manejan las bases de datos tiene algunas diferencias debido a que

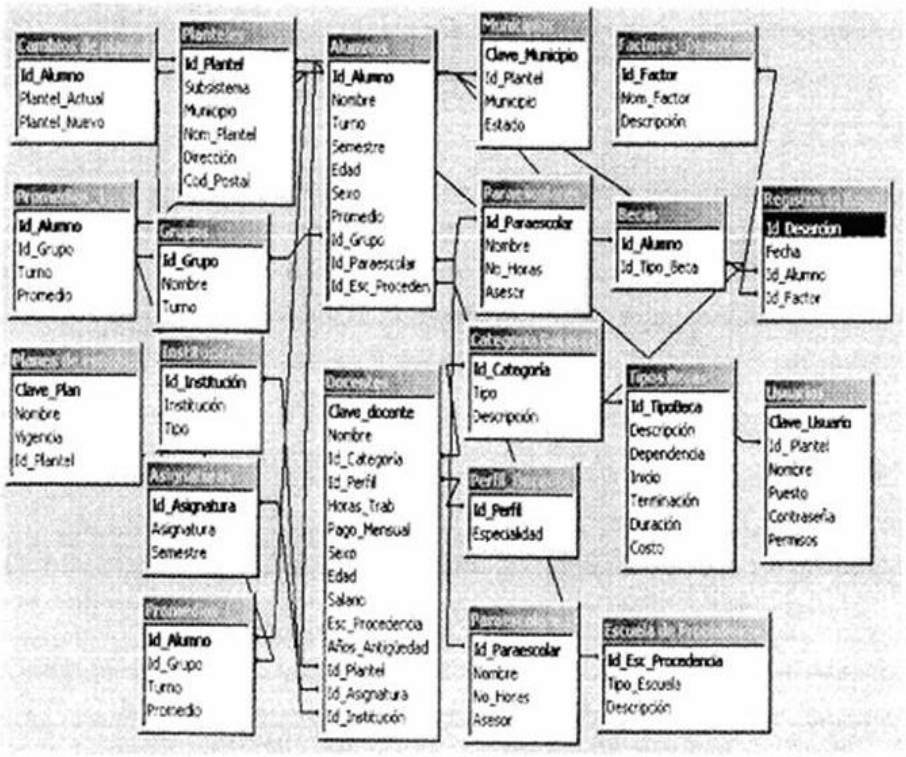
los subsistemas difieren en su modelo educativo, a continuación se explica como se encuentran las bases de datos del Departamento Académico.

Los planéeles que conforman cada subsistema, trabajan con el mismo modelo lógico de las bases de datos, pues su correspondiente Dirección General les proporciona los mismos sistemas operacionales a utilizar. Como se mencionó anteriormente, lo que varía es la información entre subsistemas.

Por ejemplo, el modelo educativo del subsistema CECyTE se enfoca a proporcionar una carrera técnica; en cambio el subsistema COBAEH proporciona un modelo educativo denominado bachillerato general Esto ocasiona que en éste Departamento, aunque las bases de datos de los distintos subsistemas manejan información sobre alumnos, docentes, grupos, etc. existen algunas diferencias.

Es importante mencionar, que tanto en todos los planteles como en la Dirección de cada subsistema y la Dirección del IHEMSyS, el manejador de base de datos que se utiliza es Access, A continuación, se muestra el modelo lógico de las bases de datos utilizadas en el Departamento Académico de cada subsistema;

Subsistema COBAEH;



***Tienta* 43 Base de datos del COBAEH**

La tase ele datos ele este subsistema difiere en el modelo lógico de las bases de datos de los otros subsistemas, como el que se muestra a continuación.

Subsistema CECyTE

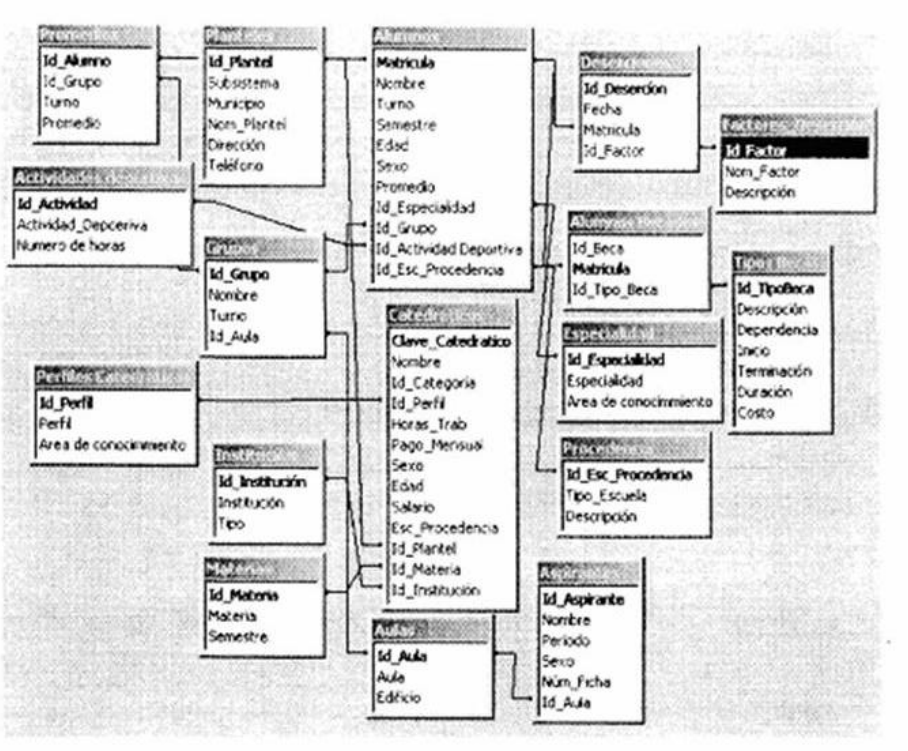


Figura 4,4 Base de datos del CECyTE

Considerando las dos bases de datos anteriores, se puede observar que, aunque en esencia ambas manejan información sobre alumnos, docentes, etc., existen varías diferencias entre ellas como son: diferente forma de nombrar una tabla o campo, también contienen tablas y/o campos distintos. A continuación se mencionan algunos ejemplos de las diferencias existentes entre las dos bases de datos anteriores:

> En la BD del subsistema CECyTE, existe la tabla especialidad, mientras que en la de COBAEH, no, Esto se debe a que en el COBAEH, su modelo educativo es bachillerato general
>
> En la BD del COBAEH, la actividad cultural o deportiva se llama paraescolar, mientras que en la tabla del CECyTE se llama actividad deportiva»
>
> En ambas tablas existe la tabla de docentes; sin embargo, en la base de datos del COBAEH se llama así, mientras que en la del CECyTE se llama catedráticos.

Las diferencias antes mencionadas, son un ejemplo, de que para utilizarse información ele las bases de datos anteriores, tendría que realizarse primero una estandarización, A

continuación se muestran las tases de datos de los subsistemas restantes, las cuales de la misma manera que las anteriores, tienen entre ellas algunas diferencias.

Subsistema CEMSAD

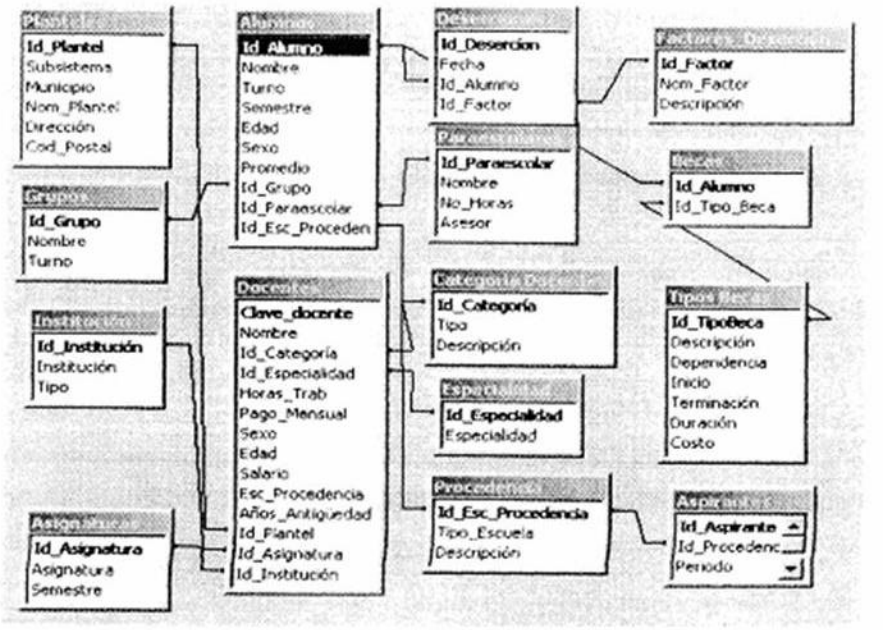


Figura *45* Base *de* datos de CEMSAD

Subsistema CONALEP

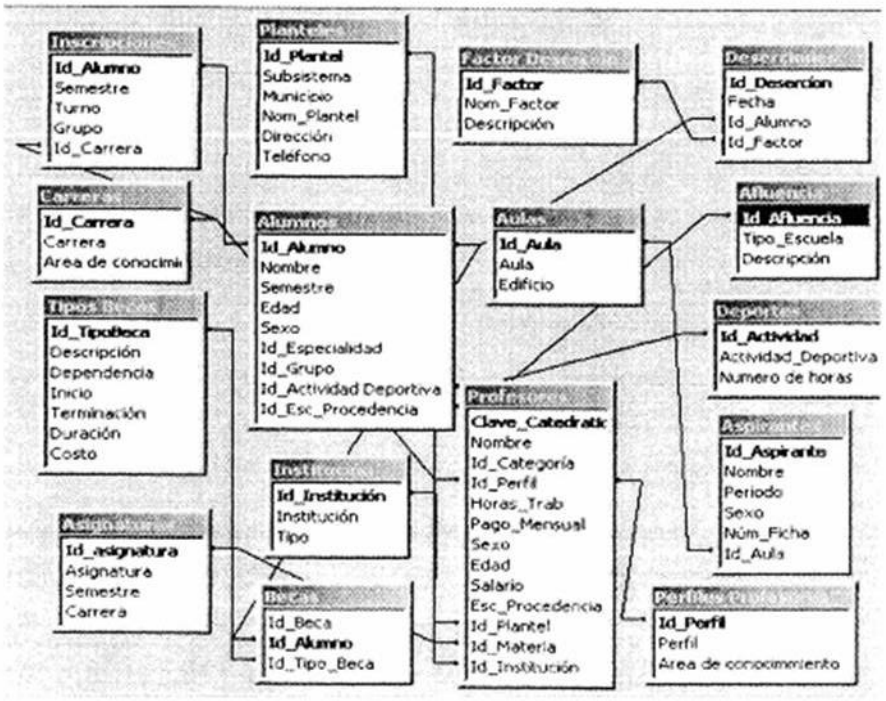


Figura 4.6 Base de datos de CONALEP

Subsistema UTEC

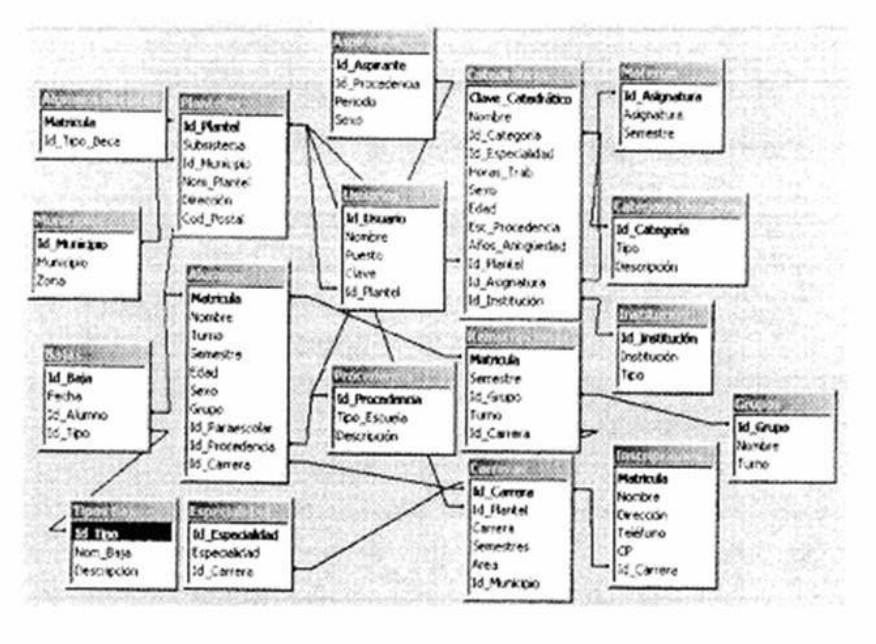


Figura 4.7 Base de datos del subsistema

Subsistemas Institutos Tecnológicos

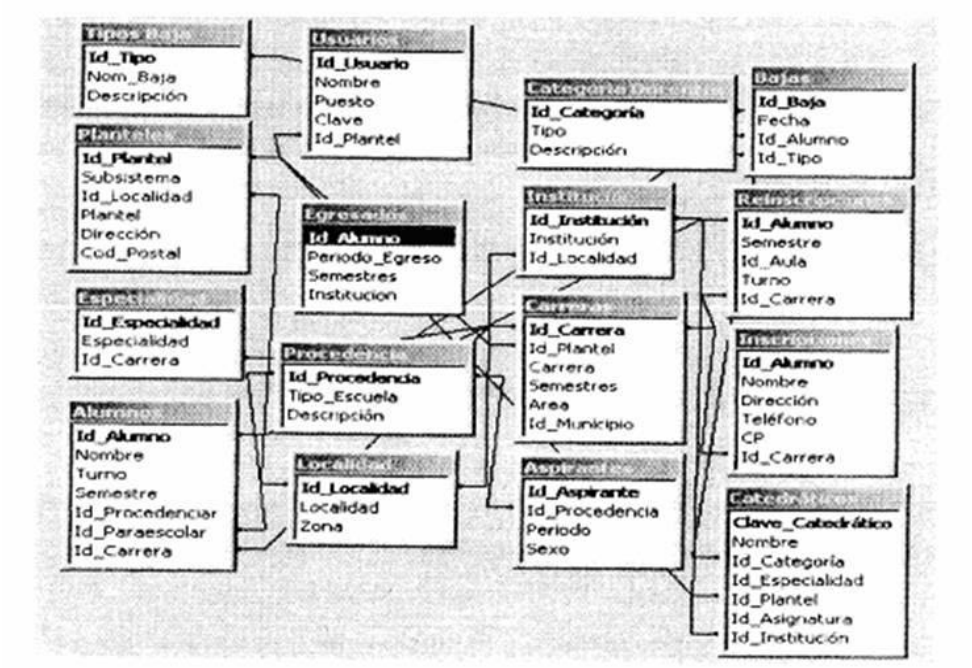


Figura 4.8 Base de datos del Subsistema Institutos Tecnológicos

Una vez analizadas estas bases de datos, se detectó que no es posible obtener de forma directa los datos finales de los diversos procesos que se llevan a cabo (por ejemplo, el total de ingresos en el plantel), que son el tipo de datos que más interesan en el depósito del DW. Sin embargo, éste tipo de información, los planteles lo manejan en formatos de Excel, los cuales se expondrán más adelante en información para las tablas de hechos.

4.14 Información requerida para el depósito del DW

Una vez analizadas las bases de datos que actualmente se manejan en los distintos subsistemas, el siguiente paso es saber que tipo de información es la que se requiere para implementar el DW en el IHEMSyS.

La información detallada que manejan las bases de datos de los sistemas operacionales antes mencionados, no son de utilidad para la implementación del DW; lo que se puede aprovechar son algunos resultados que estas proporcionan, mismas que el IHEMSyS maneja mediante reportes. Por ejemplo, en el Departamento Académico, los promedios por materia de cada alumno que se obtienen en un sistema operacional no son necesarios; lo que se requiere son datos a nivel mas general, como el promedio por grupo o plantel en determinada materia. Esto se debe a que, como se mencionó en el capítulo 1, un DW es un sistema que sirve de soporte para toma de decisiones a nivel directivo, por lo que interesan datos generales o resumidos.

Considerando otro ejemplo de información innecesaria, en el Departamento de Recursos Materiales, se requiere saber con cuantos equipos de cada tipo cuenta el plantel. En la siguiente figura, se muestra el formato que utiliza dicho departamento para almacenar los datos de los equipos. Se puede observar que para el tipo de consulta requerida, se maneja información innecesaria, pues en la consulta no se requiere saber datos sobre descripción, serie e inventario de cada equipo, lo que se necesita es saber el total de equipos por tipo, con los que cuenta el plantel.

DEPARTAMENTO DE RECURSOS MATERIALES
REPORTE DE INVENTARIOS DE EQUIPO DE COMPUTO

NOMBRE DEL SISTEMA_____
NOMBRE DEL PLANTEL:_____
CLAVE DEL PLANTEL _____ **SEMESTRE:**_____

TIPO DE EQUIPO	DESCRIPCIÓN	SERIE	INVENTARIO	OBSERVACIONES

Figura 4.9 Formato de inventarios del Departamento de Recursos Materiales

Respecto al tipo de información que se requiere para la implantación del DW, es importante recordar que el tipo de explotación de datos que se va a realizar es de consultas OLAP, para lo que es necesario un modelado multidimensional. Por lo tanto, la información se puede clasificar en: información para las tablas de dimensión; e información que se requiere en las tablas de hechos; a continuación se explica cada una de ellas.

1- Información que se requiere en las tablas de dimensión

Este tipo de información se puede clasificar en:

- Información general de las entidades que pertenecen al IHEMSyS, como datos de subsistemas, planteles, zonas, etc.

- Información de las diversas tablas dimensión, sobre las cuales se quiere analizar cada indicador Algunos de los ejemplos son: tipos de deserciones, categorías de docentes, tipos de becas, etc. Más delante de este capítulo, en el análisis multidimensional se muestra cada una de las dimensiones a utilizar.

Se detectó que la mayor parte de esta información la almacenan en archivos de Excel u la maneja la Dirección General de cada subsistema a la Dirección General del IHEMSyS. Este tipo de información se carga cuando se crea el DW u a futuro requiere de pocas agregaciones, esto se debe a que surge nueva información de este tipo con poca frecuencia. Algunos ejemplos de agregación de nuevos datos serían, los datos de los nuevos planteles, nuevos tipos de becas, nuevas categorías de docentes, etc.

A continuación se muestra un ejemplo de cómo los subsistemas tienen almacenada esta información:

RELACIÓN DE PLANTELES
SUBSISTEMA COBAEH

NOMBRE DEL PLANTEL	MUNICIPIO	NUM. DE ZONA
ACTOPAN	ACTOPAN	1
ATOTONILCO	ATOTONILCO DE TULA	1
CARDONAL	ÍXMIQUILPAN	2
CUAUTEPEC	CUAUTEPEC DE HINOJOSA	2
CHILCUAUTLA	CHILCUAUTLA	1
EMILIANO ZAPATA	TEPEAPULCO	4
FRANCISCO I. MADERO	FRANCISCO I. MADERO	3
LOS OTATES	LOS OTATES, HUEJUTLA DE REYES	1
HUICHAPAN	HUICHAPAN	2
REFORMA	MINERAL DE LAREFORMA	4
NOPALA	NOPALA DE VILLAGRAN	1
AHUATITLA	SAN FELIPE ORIZATLAN	3
SAN AGUSTÍN	SAN AGUSTÍN TLAXIACA	1
TASQUILLO	TASQUILLO	
TECOZAUTLA	TECOZAUTLA	2
TENANGO DE DORIA	TENANGO DE DORIA	3
TEPEAPULCO	TEPEAPULCO	4
TIANGUISTENGO	TIANGUISTENGO	2
TLANCHINOL	TLANCHINOL	1
TOLCAYUCA	TOLCAYUCA	1
TULA	TULA	2

Figura 4.10 Formato de relación de planteles por subsistema

Este tipo de información se recopiló de los diversos subsistemas, para homologarse posteriormente en el proceso de conversión de datos.

2.- Información que se requiere en las tablas de hechos

Datos resumidos o finales que se obtienen a partir de los procesos realizados en los sistemas operacionales que manejan los diferentes departamentos en cada plantel. Esta información, constituye los tedios sobre indicadores que van a ser almacenados haciendo uso de las tablas de dimensión.

Algunos ejemplos de este tipo de información en el Departamento Académico, son: número de alumnos becados, número de alumnos dados de baja, número de alumnos que ingresaron, número de docentes con grado de licenciatura, etc. Más adelante, en este capítulo, en el análisis multidimensional se muestra cada uno de los indicadores a utilizar.

Esta información es proporcionada por cada uno de los planteles, sin embargo, como se mencionó anteriormente, la desventaja es que dicha información no la proporcionan en forma directa sus bases de datos, sino que la tienen que almacenar en formatos de Excel en forma manual. Estos formatos son llenados, de acuerdo a la información solicitada por la Dirección General del IHEMSyS por medio de reportes.

Cuando se crea el DW, se puede almacenar tanto información reciente como la de periodos anteriores; por ejemplo de los últimos 3 años. Sin embargo, una vez cargada la información al DW, este tipo de información requiere de la agregación de nuevos datos en forma periódica, dependiendo del indicador que se trate. Por ejemplo, si se trata del número de alumnos que ingresaron, el periodo sería semestral.

A continuación se muestra el proceso que se lleva a cabo para recopilar dicha información.

Proceso para recopilar la información

En la siguiente figura, se muestra de forma general el proceso que se lleva a cabo para recopilar la información a ser utilizada en el DW.

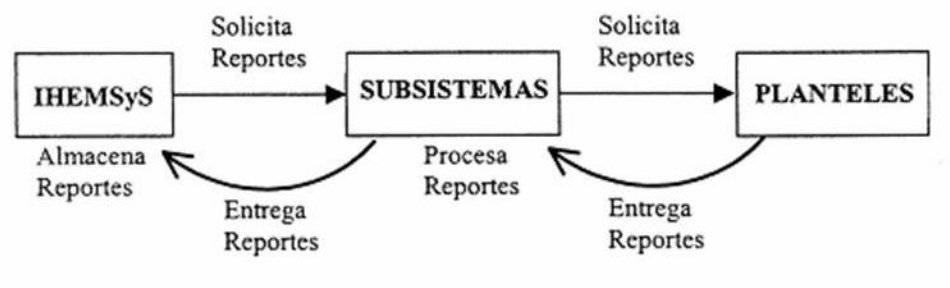


Figura 4.11 Proceso para recopilar la información

Como se muestra en la figura, el proceso inicia cuando el IHEMSyS, solicita los reportes a los diversos subsistemas en fechas que han sido establecidas anteriormente, enseguida los

subsistemas hacen la misma solicitud a cada uno de sus planteles. Los planteles llenan los reportes solicitados de acuerdo a su información obtenida a partir de los sistemas operacionales o de los formatos; posteriormente, los envían a la Dirección General de su subsistema.

A continuación se muestran los formatos para los reportes más importantes solicitados a los planteles, clasificándolos por Departamento:

Departamento Académico

NOMBRE DEL SISTEMA _____

NOMBRE DEL PLANTEL: _____

CLAVE DEL PLANTEL _____ SEMESTRE: _____

INFORMACIÓN SOBRE ALUMNOS

	TURNO		SEXO		ESCUELA DE PROCED	
DATO REQUERIDO	MATUTINO	VESPERT.	MASCULINO	FEMENINO	SECUNDARIA	TÉCNICA
ASPIRANTES. INGRESOS Y EGRESOS						
Número de aspirantes						
Número de alumnos que ingresan						
Número de alumnos que egresan						

ÍNDICE DE APROVECHAMIENTO

Número de alumnos que aprobaron todas sus materias						
Número de alumnos que reprobaron de 1 a 3 materias						
Número de alumnos que reprobaron más de 3 materias						
Número de alumnos que revalidan materias						

BECAS

	TIPO DE BACA		
DATO REQUERIDO	**TRANSPORTE**	**CONAFE**	**OPORTUNIDADES**
Número de alumnos que cuentan con beca			
Número de becas suspendidas			
Número ele becas nuevas otorgadas			

INFORMACIÓN SOBRE DOCENTES

	GRAO DE ESTUDIOS		**SEXO**		**TURNO**	
DATO REQUERIDO	**LICENCIATURA**	**MAESTRÍA**	**MASCULINO**	**FEMENINO**	**MATUTINO**	**VESPERTINO**
Número de decentes basificados						
Número de docentes contratados						
Número de decentes despedidos						

	ANTIGÜEDAD		SEXO		TURNO	
DATO REQUERIDO	1-5 Años	Más de 5 Años	MASCULINO	FEMENINO	MATUTINO	VESPERTINO
Personal Docente						
Personal Administrativo						

Figura 4.12 Formato ole reporte semestral del Departamento Acad émico

Este formato, pertenece a un reporte semestral, en el cual los planteles tienen que reportar datos sobre alumnos y docentes, de acuerdo a diversos factores. Respecto a alumnos, se solicitan ingresos, egresos, índice de aprovechamiento, número de becas, etc.

Respecto a los docentes, se solicita información sobre número de docentes por tipo de contrato y por antigüedad.

A continuación se muestra un formato, con el que se obtiene un reporte mensual sobre las deserciones ocurridas en el plantel, de acuerdo al sexo, turno y escuela de procedencia.

DEPARTAMENTO ACADÉMICO

REPORTE MENSUAL

NOMBRE DEL SISTEMA _____

NOM3RE DEL PLANTEL _____

CLAVE DEL PLANTEL _____ **MES:** _____

DESERCIONES

	TURNO		SEXO		ESCUELA DE PROCED.	
DATO REQUERIDO	**MATUTINO**	**VESPERT.**	**MASCULINO**	**FEMENINO**	**SECUNDARIA**	**TÉCNICA**
Número de alumnos Que desertaron por problemas escolares						
Número de alumnos que desertaron por problemas sociales						
Número de alumnos que desertaron por problemas económicos						

Figura 4.13 Formato ole reporte mensual del Departamento Académico

En el formato se muestra que las deserciones se obtienen de acuerdo a la causa que la originó (problemas escolares, sociales ó económicos). A continuación se muestra un formato, con el que se obtienen datos cada bimestre.

DEPARTAMENTO ACADÉMICO

REPORTE BIMESTRAL

NOMBRE DEL SISTEMA _____

NOMBRE DEL PLANTEL: _____

CLAVE DEL PLANTEL _____ BIMESTRE _____

DATOS DE ALUMNOS

	SEXO		CATEGORÍA		
DATO REQUERIDO	**MASCULINO**	**FEMENINO**	**C.EXACTAS**	**C.SOCIALES**	**LENG. Y COMUNIC.**
Porcentaje de aprobación					
Porcentaje de reprobación					

DATOS DE DOCENTES

DATO REQUERIDO	**MASCULINO**	**FEMENINO**	**C.EXACTAS**	**C. SOCIALES**	**LENG. Y COMUNIC.**
Promedio docente					
Porc Inasistencias					

Figura 4.14 Formato de reporte bimestral del Departamento Académico

Como se muestra en el formato anterior, los datos se obtienen tanto de alumnos (porcentaje de aprobación y reprobación) como de docentes (promedio docente, porcentaje de inasistencias).

Departamento de Recursos Materiales

A continuación se muestra un formato que en el Departamento de Recursos Materiales, permite obtener trimestralmente datos sobre el mobiliario, equipo de cómputo, equipo electrónico y bibliografía de cada plantel.

DEPARTAMENTO DE RECURSOS MATERIALES
REPORTE TRIM ESTRAL

NOMBRE DEL SISTEMA: _____
NOMBRE DEL PLANTEL: _____
CLAVE DEL PLANTEL: _____ **TRIMESTRE:**

DESCRIPCION DEL EQUIPO	EN EXISTENCIA	DADOS DE BAJA	EN REPARACIÓN
MOBILIARIO			
Número de butacas			
Número de escritorios			
Número de pizarrones			
Número de sillones			
Número de archiveros			
EQUIPO DE COMPUTO			
Número de computadoras			
Número de impresoras Lasser			
Número de impresoras de inyección de tinta			
Número de impresoras de matriz de puntos			
Número de Scanners			
Número de CD ROM			
Número de reguladores			
Número de concentradores			
EQUIPO ELECTRÓNICO			
Número de televisores			
Número de stereos			
Número de videos			
Número de proyectores de acetatos			
Número de proyectores de computadora			
Número de maquinas de escribir			
BIBLIOGRAFÍA			
Ciencias exactas			
Ciencias sociales			
Ciencias biologicas			
Ciencias humanisticas			
Lenguaje y comunicaciones			

Figura 4.15 Formato de *reporte* trimestral del Departamento de Recursos Materiales

Como se puede observar, en la mayoría de los reportes solicitados, la información que se pide, se refiere a los resultados finales que se obtuvieron dentro del plantel.

DEPARTAMENTO DE RECURSOS MATERIALES
REPORTE SEMESTRAL

NOMBRE DEL SISTEMA:_____
NOMBRE DEL PLANTEL: _____ **SEMESTRE:** _____
CLAVE DEL PLANTEL: _____

DESCRIPCION DEL EQUIPO	ADQUIRIDOS
MOBILIARIO	
Número de butacas	
Número de escritorios	
Número de pizarrones	
Número de sillones	
Número de archiveros	
EQUIPO DE COMPUTO	
Número de computadoras	
Número de impresoras Láser	
Número de impresoras de inyección de tinta	
Número de impresoras de matriz de puntos	
Numero de Scanner	
Número de CD ROM	
Número de reguladores	
Número de concentradores	

DESCRIPCION DEL EQUIPO	ADQUIRIDOS
EQUIPO ELECTRONICO	
Número de televisores	
Número de estéreos	
Número de videos	
Número de proyectores de acetatos	
Número de proyectores de computadora	
Número de maquinas de escribir	
BIBLIOGRAFIA	
Ciencias exactas	
Ciencias sociales	
Ciencias biológicas	
Ciencias humanísticas	
Lenguaje y comunicaciones	

Figura 4.16 Formato de reporte semestral del Departamento de Recursos Materiales

A continuación se muestran los formatos utilizados en el Departamento de Recursos Financieros.

Departamento de Recursos Financieros

DEPARTAMENTO DE RECURSOS FINANCIEROS
REPORTE SEMESTRAL

NOMBRE DEL SISTEMA. _____

NOMBRE DEL PLANTEL: _____ SEMESTRE:

CLAVE DEL PLANTEL _____

INGRESOS

TIPO DE INGRESO	TOTAL
Total de ingresos por fichas de admisión	
Total de ingresos por inscripciones	
Total de ingresos por reinscripciones	
Total de ingresos por exámenes de recuperación	
Total de ingresos por exámenes especiales	

EGRESOS

TIPO DE EGRESO	TOTAL
Gastos por cursos de actualización	
Gastos en construcciones	

Figura 4.17 Formato de reporte Semestral del Departamento de Recursos Financieros

Con este formato, se pueden obtener semestralmente datos sobre los ingresos de los planteles

DEPARTAMENTO DE RECURSOS FINANCIEROS
REPORTE MENSUAL

NOMBRE DEL SISTEMA _____

NOMBRE DEL PLANTEL: _____ MES: _____

CLAVE DEL PLANTEL: _____

INGRESOS

TIPO DE INGRESO	TOTAL
Tota de ingresos por trámites de documentación	
Total de ingresos por donativos	
Total de ingresos por reinscripciones	
Total de ingresos por actividades o eventos realizados	
Total de ingresos por cuotas de padres de familia	
Total de ingresos por permisos de venta	

EGRESOS

TIPO DE EGRESO	TOTAL
Egresos por pago de personal	
Egresos por mantenimiento del plantel	
Egresos por viáticos	
Egresos por compra de consumibles	
Egresos por concepto de publicidad	
Egresos por eventos efectuados	

Figura 4.18 Formato de reporte Mensual del Departamento de Recursos Financieros

Con este formato, se pueden obtener mensualmente datos sobre los ingresos y egresos de los planteles.

Una vez que cada subsistema recibe la información, para cada reporte hace la unión de la información de sus planteles de forma manual en Excel y los envía a la Dirección General del IHEMSyS. En la siguiente figura, se muestra un ejemplo de un formato donde el subsistema COBAEH hace la unión de la información proporcionada por los planteles que lo integran, tomando como ejemplo el primer reporte del Departamento Académico mostrado anteriormente en la figura 4.12.

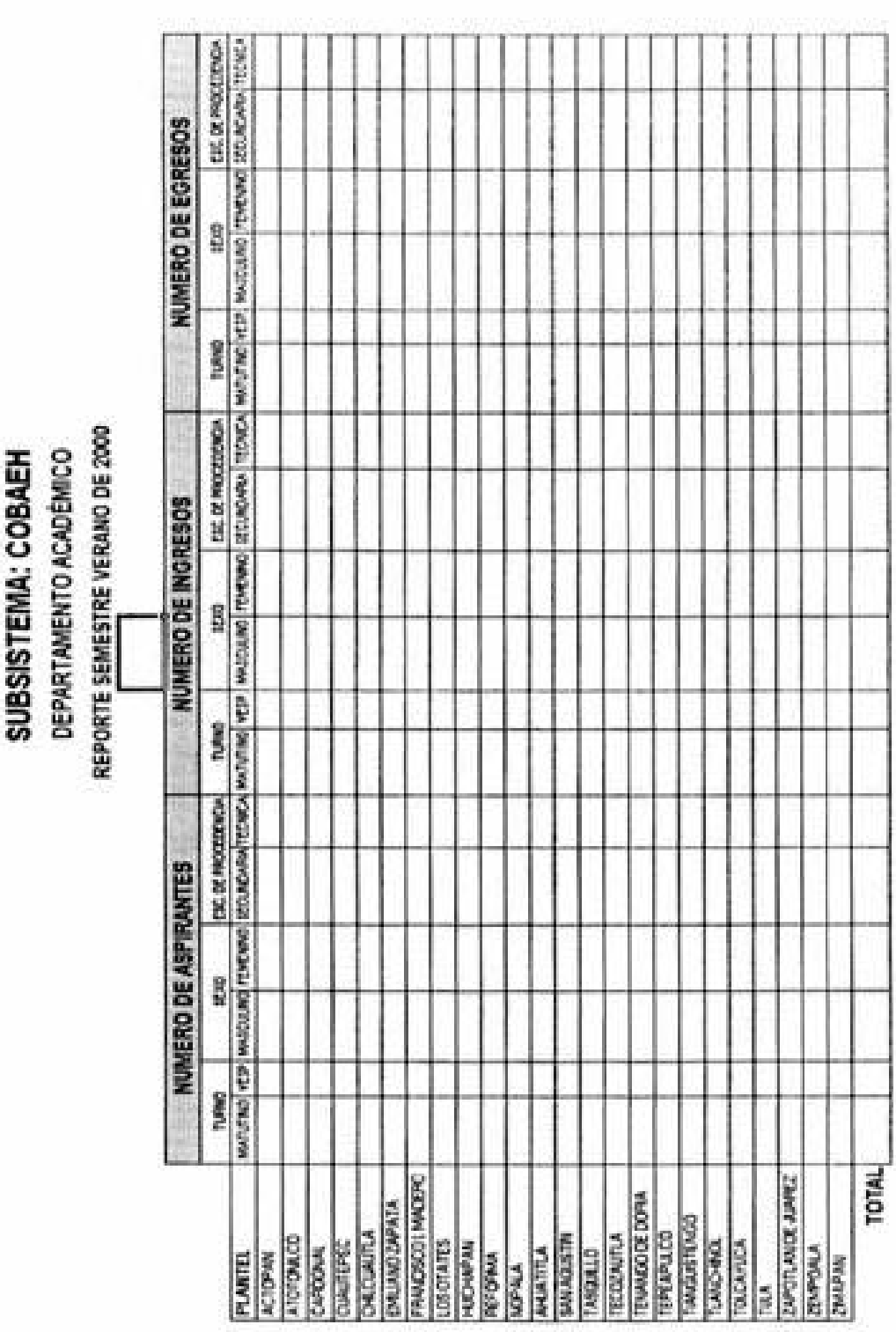


Figura 4.19 Formato de reporte semestral del Departa mento Académico

El formato sólo muestra parte de la información; sin embargo, contiene la información ele tocios los planteles. De manera similar, los subsistemas Lacen lo mismo con el resto ele los formatos, para enviarlos en archivos y en forma impresa a la Dirección del IHEMSyS.

Por último, el IHEMSyS reúne todos los reportes enviados por cada subsistema y nace uso de ellos cuando se requiere de alguna consulta, de estos reportes es de donde se obtuvo la información de tablas de hechos. Sin embargo, esta forma de trabajar, presenta varios inconvenientes como los que se muestran a continuación.

Los reportes están separados por subsistema u por periodos. Por ejemplo, en el caso de que se trate de un reporte trimestral, u que se requiera obtener información de ese reporte en todos los subsistemas en los últimos dos anos. Se tendrían que reunir los reportes de todos los subsistemas de los últimos ocho trimestres y considerando que son seis subsistemas, se tendrían que analizar 54 reportes para obtener la información.

La información concentrada en el depósito del DW soluciona este problema, debido a que la almacena en un solo lugar.

Otro inconveniente, es que en los reportes los datos no están almacenados con el mayor detalle posible. Por ejemplo, el reporte anterior permite saber cuantos alumnos del turno matutino ingresaron por plantel o por subsistema, también permite saber cuantos alumnos provenientes de escuela secundaria, ingresaron, en este caso, turno y escuela de procedencia son las dimensiones por las cuales se quiere analizar el indicador ingresos. Sin embargo, el reporte no permite realizar consultas mas detalladas o con la combinación de varias dimensiones, por ejemplo, saber cuantos alumnos de los que ingresaron en el turno matutino son del sexo masculino.

Para resolver esta situación, se propone que el IHEMSyS solicite a los subsistemas los reportes en forma más detallada, especificando las posibles combinaciones entre las diferentes dimensiones de las que requiera la información de un dato o indicador. Sin embargo, para realizar esta tarea, primero se debe saber cuales son los indicadores necesarios en cada Departamento u sobre que dimensiones se van a analizar; esto se obtiene después de realizar un análisis u diseño multidimensional, el cual se presenta mas adelante.

En la figura 4 20 se muestra la forma en como se soluciona el problema del ejemplo antes mencionado, solicitando la información de acuerdo a la combinación de las distintas dimensiones (turno, sexo y escuela de procedencia).

Por ejemplo, con dicto reporte en la columna 1 se puede obtener información sobre los alumnos que ingresaron en el turno matutino, que son del sexo masculino u que provienen de escuela secundaria.

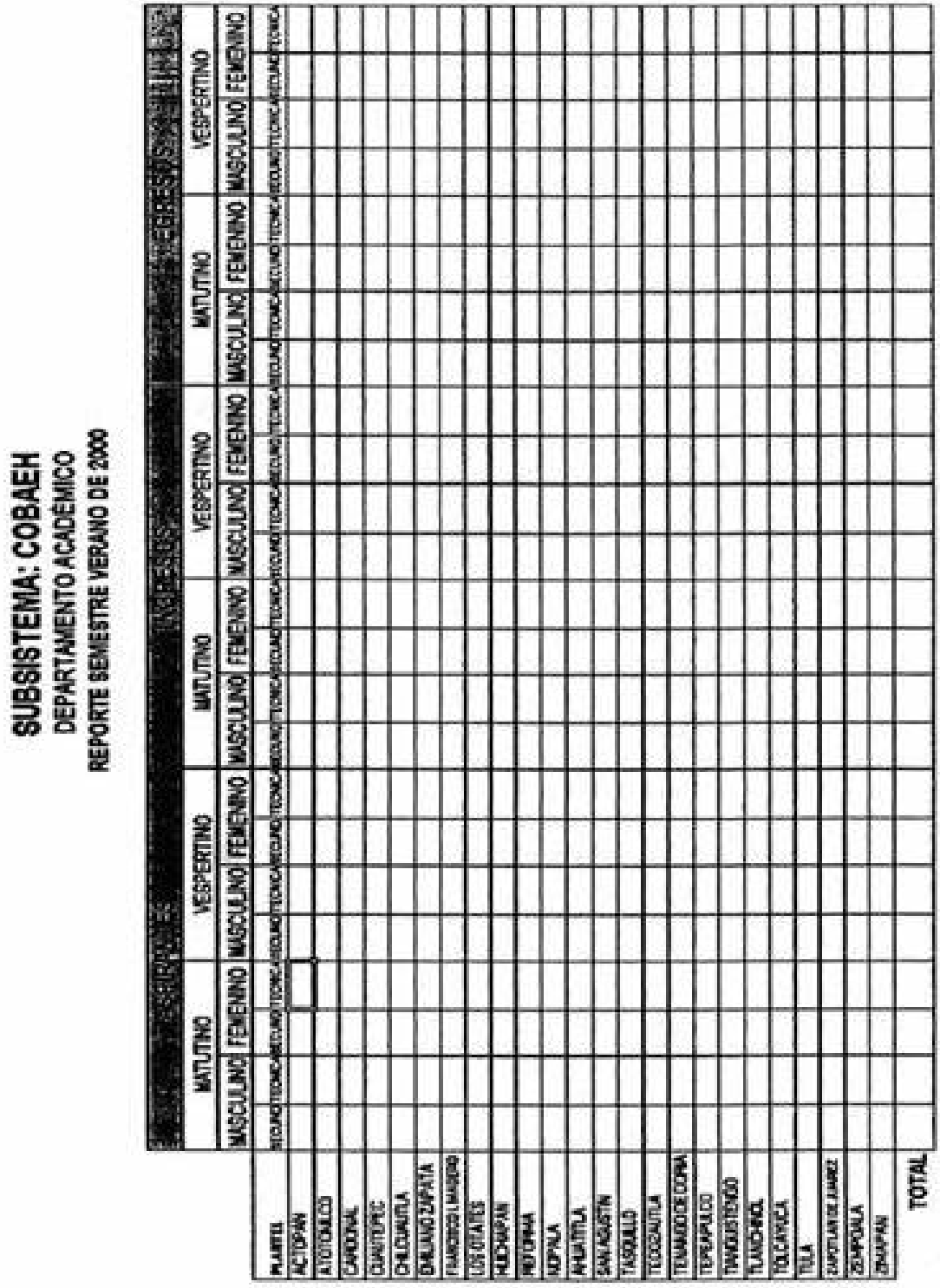


Figura 4.20 Formato de reporte semestral del Depto. Académico en forma detallada

Una vez analizadas las bases de datos fuente y el tipo de información que solicita el IHEMSyS a los diversos subsistemas, es necesario saber cuales son las necesidades de los usuarios finales del DW. En el siguiente punto se exponen los requerimientos de dichos usuarios.

Requerimientos

Antes de obtener los elementos del análisis multidimensional (indicadores, dimensiones, etc.), fue necesario detectar cuales son los requerimientos dentro de cada Departamento. Para realizar esta tarea, fue necesario obtener la información a partir de los usuarios finales; es decir, el personal que va a realizar las consultas y que esta involucrado en la tarea de realizar toma de decisiones.

A continuación se muestra el formato del cuestionario realizado al personal directivo (Director y Subdirector) de los distintos Departamentos, así como a los responsables del área de cada subsistema, pues son quienes tienen dominio de las necesidades informativas y quienes utilizarán el DW.

FORMATO PARA OBTENER REQUERIMIENTOS

Número de entrevista: _____
Nombre del encuestado: _____
Departamento: _____
Puesto: _____

Describa en forma resumida el tipo de información que continuamente requiere consultar:

Requerimientos:

1.-_____
_____
2.- _____
_____
3.- _____
_____
4.- _____

Figura 4.21 Formato para la obtención de requerimientos

Una vez entrevistado al personal de los Departamentos Académico, Recursos Financieros y Recursos Materiales del IHEMSyS, con la aplicación de la entrevista anterior, se obtuvieron de forma general los requerimientos. A continuación se muestran clasificándolos por Departamento:

Departamento Académico

Obtención de datos finales sobre procesos efectuados a alumnos en el transcurso de su estancia en un plantel, los más comunes son de tipo académico (ejemplo, promedios, inasistencias, etc.)

- Datos sobre docentes que laboran en un plantel.
- Datos del plantel.
- Datos sobre materias.

Departamento de Recursos Financieros

- Control de los gastos efectuados, según las diversas actividades del plantel.
- Control de los gastos efectuados por mantenimiento y equipo.
- Control de los ingresos obtenidos en los planteles por aspectos académicos.
- Control de los ingresos obtenidos en los planteles por aspectos no académicos.

Departamento de Recursos Materiales:

- Control de inventarios de los diferentes tipos de equipo que conforman el plantel educativo.
- Control de cambios en el equipo.
- Control de bajas de equipo y material.
- Control de equipos y material dañado.

Estos requerimientos forman parte de un análisis previo, pues se describen en forma general. Por ejemplo, uno de los requerimientos de el Departamento Académico son datos sobre docentes, pero un tipo de requerimiento más específico sería el nivel académico de cada uno. Para obtener los requerimientos en forma detallada, es necesario obtener los indicadores existentes en cada Departamento, lo que forma parte del análisis multidimensional, que se explica a continuación.

4.2 ANÁLISIS MULTIDIMENSIONAL

Como se indicó en los alcances de este trabajo, el análisis multidimensional se realizó únicamente en los Departamentos Académico, Recursos Financieros y Recursos Materiales del IHEMSyS, por ser los tres más importantes dentro de la organización.

Aunque todos los Departamentos son importantes para la organización, se considera que los tres que se tomaron en cuenta para el proceso de análisis multidimensional, son en los que se requiere realizar consultas que ayuden a la toma de decisiones con mayor prontitud. El Departamento Académico se considera de los más importantes, debido a que en el caso de estudio se está tratando con una organización educativa, cuya razón de existencia son los alumnos y los docentes.

El Departamento de Recursos Financieros tiene especial importancia, puesto que proporciona información sobre los ingresos y egresos monetarios en los diversos planteles de los subsistemas del IHEMSyS. Por último, la importancia del Departamento de Recursos Materiales se basa en que permite proporcionar información sobre el equipo disponible en los planteles.

Por otra parte, como se expuso anteriormente, las tareas principales para el análisis multidimensional, son: obtener indicadores y dimensiones; determinar la granularidad; y detectar los horizontes de tiempo. A continuación se muestran los elementos que son necesarios para realizar el análisis multidimensional.

4.21 Obtención de Indicadores

Como se mencionó en el capítulo anterior, los indicadores son datos que representan resultados de procesos que se realizan dentro de la organización, por lo general son numéricos. Por ejemplo, en el proceso de evaluación de alumnos, el dato a obtener es el promedio del alumno, sin importar el proceso que se realiza.

La forma más fácil de obtener los indicadores, es a partir del análisis de los datos manejados actualmente por la organización; sin embargo, para tener resultados mas aproximados a las necesidades, otra forma de obtenerlos es entrevistando al personal de la organización en sus distintos Departamentos. Esto, con la finalidad de saber cuales son sus necesidades informacionales, es decir, que consultas requieren hacer al almacén de datos para que les ayude en la toma de decisiones.

Posteriormente, se deben seleccionar los indicadores a partir de la información obtenida mediante la aplicación de un cuestionario o entrevista al personal directivo de cada Departamento. En realidad, al terminar la tarea de obtención de los indicadores, se combinan ambas formas de obtenerlos, debido a que después de analizar los resultados de las encuestas, es necesario agregar los indicadores que se consideren también necesarios o con posibilidad de ser requeridos en algún momento.

El análisis de los elementos del modelado multidimensional, es una de las tareas mas importantes dentro del proceso de desarrollo del DW, debido a que de esta tarea depende que una vez implementado el DW, la información almacenada sea la necesaria o correcta para satisfacer las necesidades informacionales de los usuarios.

En el caso de estudio, el proceso llevado a cabo para obtener los indicadores fue, en primer lugar realizando un cuestionario principalmente a personal directivo que tiene facultad para tomar decisiones en los Departamentos Académico, Recursos Financieros y Recursos Materiales del IHEMSyS, así como a los directivos (Director y Subdirector) de cada uno de los Departamentos mencionados de los diferentes subsistemas del IHEMSyS.

Posteriormente, a partir de los indicadores obtenidos, por medio del análisis de los reportes de información que solicita el IHEMSyS (los cuales se mostraron anteriormente) y las bases de datos que se manejan actualmente en los distintos subsistemas, se obtuvieron otros indicadores más, los que se agregaron a la lista obtenida anteriormente. Por último, nuevamente se validaron los indicadores anexados con el personal entrevistado.

A continuación se presenta el formato de la encuesta:

FORMATO DE CUESTIONARIO

DATOS LLENADOS POR EL PERSONAL CUESTIONADO

Número de encuesta _____

Nombre del subsistema _____

Nombre del plantel _____

Nombre del departamento _____

Puesto _____ Experiencia _____

¿Qué consultas realiza frecuentemente?, de preferencia indicarlas por orden de importancia

En cada consulta se debe especificar el indicador requerido y respecto a que dimensiones se desea obtener Ej. Obtener el número de alumnos de nuevo ingreso por plantel, zona, sexo n semestre.

1.-_____

2- _____

3.-_____

4.-_____

5,- _____

6.-_____

¿Aproximadamente, con que periodicidad se realiza cada una de las consultas anteriores?

Especificar en qué periodo se realiza cada una de las consultas especificadas anteriormente, relacionando los números de consulta separados con comas con su correspondiente periodo, Ej. 2,4,5.

Número de consultas	Periodicidad
_____	Diariamente
_____	Semanalmente
_____	Mensualmente
_____	Trimestralmente
_____	Semestral mente

DATOS LLENADOS POR EL ANALISTA

Indicadores involucrados

Consulta 1 _____

Consulta2 _____

Consulta3 _____

Consulta 4 _____

Consulta5 _____

Consulta6 _____

Dimensiones u horizonte de tiempo que involucra cada indicador

Indicador:	Horizonte de Tiempo	Dimensiones (Granularidad)	
_____	_____	_____	_____
		_____	_____
		_____	_____
_____	_____	_____	_____
		_____	_____
		_____	_____

Figura 4.22 Formato de encuesta

El formato de la encuesta se divide en dos partes. La primera (parte superior), contiene los datos que el encuestador va a obtener. Incluye datos del encuestado, lista de consultas que realiza y por último, la especificación de la frecuencia con que se presenta cada consulta. Es importante informarle al encuestado, que para el orden de especificación de sus consultas en el formato, debe considerar los siguientes criterios:

- La importancia que tiene cada consulta en su Departamento; es decir, que tan importantes son los resultados que la consulta va a proporcionar, para la toma de decisiones dentro del Departamento.
- La frecuencia con que se realiza la consulta.

Respecto a la especificación de la frecuencia en que sucede cada consulta, en el formato se puede observar que se solicita el número de consultas relacionándolas con cada periodo, lo que le ayudará al analista a especificar los horizontes tiempo. Por ejemplo, al periodo semestralmente, el usuario puede hacer corresponder el número de consultas que especificó anteriormente.

La segunda parte del formato de la figura 4.22 (parte inferior), contiene datos que deben ser llenados por el analista, aplicando sus conocimientos de los términos del análisis multidimensional (indicador, dimensión, etc.). A partir del contenido textual de las consultas proporcionadas por el encuestado, el encuestador va a determinar la lista de indicadores y sus correspondientes dimensiones (granularidad).

Por ejemplo, si en una de las consultas proporcionadas por el encuestado en la parte superior del formato es, ¿Cuántos alumnos se dieron de baja en todos los subsistemas en el último año? Para este caso, los datos a llenar por el encuestador en la parte inferior del formato son: como indicador es deserciones y como dimensiones son el subsistema y el tiempo.

Como se observa en el formato de la encuesta, además de permitir obtener indicadores y dimensiones, también se solicita información sobre la relación entre ambos elementos y sobre su correspondiente horizonte de tiempo, elementos que se expondrán más adelante.

Una vez aplicados los cuestionarios se analizaron los resultados para determinar qué elementos del modelo multidimensional se pueden considerar. El primer elemento a obtener fueron los indicadores, debido a que de ellos dependen los demás elementos, como son las dimensiones, horizonte de tiempo, etc.

En las siguientes tablas, se muestran los indicadores obtenidos y cuantos de los cuestionados sugirieron cada uno, los indicadores se clasifican por Departamento. Es importante recordar que el personal a quien se le aplicó el cuestionario fue al personal directivo de los diferentes Departamentos (Director y Subdirector) así como a los directivos de cada uno de los Departamentos mencionados de los diferentes subsistemas del IHEMSyS.

Departamento: Académico No. de encuestados: 27

Indicador	No. de encuestados
Deserciones	19
Ingresos	17
Egresos	15
Aspirantes	13
Promedio de ingresos	13
Revalidaciones	13
Aprobación	13
Reprobación	12
Regulares	12
Irregulares	12
Repetidores	12
Promedio Docente	15
Grado de Estudios	14
Inasistencia	13
Contratación	14
Basificados	14

Tabla 4.1 Encuestados en el Departamento Académico

El número de encuestados se debe a que fueron 3 del Departamento Académico de la Dirección General del IHEMSyS (Director, Subdirector, Asistente Técnico), 2 por subsistema (Director y Subdirector del Área Académica), y 2 personas por plantel (responsable del área académica y un profesor), tomando dos planteles por subsistema.

Departamento: Recursos Financieros No. de Encuestados: 27

Indicador	No. de encuestados
Gastos en pago a personal	19
Gastos en compra de equipo	19
Gastos en mantenimiento	17
Gastos en consumibles	15
Gastos en publicidad	13
Gastos en eventos	13
Gastos en becas	13
Gastos en cursos ole actualización	13
Ingresos de fictas ole examen	11
Ingresos de inscripciones	11
Ingresos de reinscripciones	10
Ingresos de trámites	10
Ingresos de exámenes de recuperación	10
Ingresos de exámenes especiales	10

Tabla 4.2 Encuestados en el Departamento de Recursos Financieros

Departamento: Recursos Materiales No. de Encuestados: 27

Indicador	No. de encuestados
Matrícula	20
Aulas	20
Mobiliario en uso	18
Mobiliario en reparación	14
Mobiliario dado de baja	14
Equipo de cómputo en uso	14
Equipo de cómputo en reparación	14
Equipo de cómputo dados de baja	13
Equipo electrónico en uso	13
Equipo electrónico adquirido	12
Equipo electrónico dado de baja	11
Libros en uso	11
Libros perdidos	106

Tabla 43 Encuestados en el Departamento de Recursos Materiales

Las listas de indicadores mostradas anteriormente son el resultado de las encuestas realizadas, sin embargo, es importante mencionar que la información que proporciona el personal encuestado no debe considerarse como total. Fue necesario analizar información que manejan los departamentos, para obtener los indicadores que pudieran necesitarse además de los proporcionados. Para esto, fue necesario analizar los formatos de recopilación de datos; en este capítulo se mostraron algunos de ellos, en la parte de información requerida para las tablas de hechos.

Una vez que se agregaron los demás indicadores, estos se validaron con el personal anteriormente cuestionado. A continuación se muestran las listas completas de indicadores, tanto los que ya se sabían mostrados, como los que se agregaron después de un análisis.

Departamento Académico

Nota: Los indicadores marcados con asterisco ( ), son los que se agregaron a los obtenidos en las encuestas. Esta misma notación se utiliza en las listas restantes.

Indicador	Descripción
Aspirantes	Alumnos que realizan examen para ingresar a un plantel
Ingresos	Alumnos que se inscriben en el plantel
Revalidaciones	Alumnos que ingresan por cambio de plantel, revalidando materias
Egresos	Alumnos que terminan sus estudios

Deserciones	Alumnos que son dados de taja de un plantel
Aprobación	Alumnos que acreditan una materia
Reprobación	Alumnos que no acreditan una materia
Regulares	Alumnos que en el semestre no reprobaron ninguna materia
Irregulares	Alumnos que en el semestre reprobaron de 1 a 3 materias *
Repetidores	Alumnos que en el semestre reprobaron mas de 3 materias
Becados	Cantidad de alumnos, a los que se les proporciona alguna beca
Promedio docente	Promedio de las evaluaciones que aplica el docente, englobando las asignaturas que imparte
Inasistencias	Porcentaje de inasistencia del docente
Contratación	Número de docentes que son contratados
Basificados	Número de docentes de tiempo completo *
Eventuales	Número de docentes que trabajan por Loras
Despidos	Número de docentes que se despiden en los planteles *
Antigüedad	Promedio de antigüedad por docente
Grado de estudios	Número de docentes con licenciatura, maestría o doctorado
Incapacidades	Número de incapacidades de los docentes

Tabla 4.4 Indicadores en el Departamento Académico

A continuación, se explica la importancia de los indicadores que se muestran en la lista anterior:

Todos los indicadores, permiten obtener resultados respecto a distintas dimensiones (periodo, plantel, zona, etc.), dichas dimensiones se muestran más adelante en la sección 4.2.2.

Los indicadores aspirantes, ingresos, deserciones y egresos, permite saber la cantidad de alumnos que desean incorporarse, se incorporan, truncan ó terminan sus estudios en un plantel, tomando en cuenta de que tipo de institución provienen. Los indicadores *aprobación, reprobación, regulares, irregulares y repetidores* permiten saber en forma general, la situación académica de los alumnos respecto a los resultados de sus evaluaciones.

El indicador becas, permite saber que cantidad de dinero se destina en cada uno de los planteles, en los distintos tipos de beca que se proporcionan a los alumnos. Los indicadores contratación, basificados, eventuales y despedidos, permiten saber el número de empleados con los que cuenta un plantel, de acuerdo a su tipo de contrato, así como sus bajas. El indicador antigüedad, ayuda a saber el número de anos ka laborado el personal en un plantel, lo que permite detectar si este factor influye en su rendimiento dentro de la institución.

El indicador *promedio docente*, permite saber resultados del rendimiento del los profesores de un plantel, de acuerdo a la categoría donde se ubican las materias que imparte. El indicador *revalidaciones*, permite saber la cantidad de alumnos que ingresan a un plantel para continuar sus estudios.

El indicador *grado de estudios*, ayuda a conocer la cantidad de docentes con grado de licenciatura, maestría o doctorado están laborando y poder detectar si esto afecta los resultados del desempeño docente en un plantel. Por último, el indicador *incapacidades*, permite saber el porcentaje de ausencias a sus labores por parte del personal de un plantel.

A continuación, se muestra la lista de indicadores pertenecientes al Departamento de recursos financieros

Departamento de Recursos Financieros

Indicador	Descripción Gastos
Gastos	
Gastos en pago a personal	Pagos a docentes
Gastos en Viáticos	Gastos por comisiones extralaborales
Gastos en compra de equipo	Compras de equipos nuevos para la institución
Gastos en mantenimiento	Reparaciones, pago de agua, etc.
Gastos en consumibles	Gastos en compra de material
Gastos en publicidad	Promoción para la institución
Gastos en eventos	Clausuras, concursos, etc.
Gastos en becas	Becas otorgadas a alumnos
Gastos en cursos de actualización	Cursos impartidos a personal
*Gastos en incentivos	Gratificaciones, estímulos, etc.
*Gastos en construcciones	Egresos por construcción de instalaciones
*Gastos en bibliografía	Egresos por compra de libros a bibliotecas
Ingresos	
Ingresos de fictas de examen	Pagos por derecho a examen de admisión
Ingresos de inscripciones	Pagos de alumnos de nuevo ingreso
Ingresos de reinscripciones	Pagos de alumnos que se reinscriben
Ingresos de tramites	Tramitación de documentos
Ingresos de exámenes de recuperación	Exámenes no ordinarios
Ingresos de exámenes especiales	Exámenes especiales
Ingresos de donativos	Realizados por instituciones externas
*Ingresos de eventos	Eventos realizados por la institución
*Ingresos de cuotas	Cuota de padres de familia, Permisos de Venta, etc.
*Ingresos por permisos de ventas	Permisos para realizar ventas dentro de la institución

Tabla 45 Indicadores en el Departamento de Recursos Financieros

Departamento de Recursos Materiales

Indicador	Descripción
Aulas	Número de salones de clase
Mobiliario en uso	Número de escritorios, pizarrones, butacas, etc. *
Mobiliario adquirido	Número de escritorios, pizarrones, butacas, etc.
Mobiliario en reparación	Número de escritorios, pizarrones, butacas, etc.
Mobiliario dado de baja	Número de escritorios, pizarrones, butacas, etc.
Equipo de cómputo en uso	Número de computadoras, impresoras, scanner, etc.
Equipo de cómputo en reparación	Número de computadoras, impresoras, scanner, etc.
Equipo de cómputo adquirido	Número de computadoras, impresoras, scanner, etc.
Equipo de cómputo dado de baja	Número de computadoras, impresoras, scanner, etc.
Equipo electrónico en uso	Número de cañones, proyectores, videos, televisores, etc.
Equipo electrónico en reparación	Número de cañones, proyectores, videos, televisores, etc.
Equipo electrónico adquirido	Número de cañones, proyectores, videos, televisores, etc.
Equipo electrónico dado de baja	Número de cañones, proyectores, videos, televisores, etc.
Libros en uso	Número de libros que se pueden consultar
Libros adquiridos	Número de libros que compra el plantel
Libros perdidos	Libros extraviados

Tabla 4.6 Indicadores en el Departamento de Recursos Materiales

Es importante mencionar, que una vez agregados los indicadores que se consideraron necesarios, estos se validaron nuevamente con los encuestados. Una vez realizada esta validación, los indicadores agregados fueron aceptados.

4.2.2 Dimensiones

Las dimensiones son entidades de la organización que sirven como perspectiva de análisis para los indicadores. Cada indicador involucra un grupo de dimensiones para su obtención. Las dimensiones se obtuvieron de los cuestionarios aplicados, tomando en cuenta que se quiere y que se puede analizar de cada indicador.

Por ejemplo, si en los resultados de un cuestionario, en una consulta se quiere obtener el promedio de los alumnos por plantel, por docente y sistema, las dimensiones son planteles, docentes y sistemas. El analista debe tomar en cuenta que es posible obtener información de dicho indicador respecto a otras dimensiones como son, por semestre, por turno, etc. La tarea de agregar y validar todas las dimensiones que considere necesarias, fue trabajo que también se realizó.

4.23 Dependencia entre dimensiones

Una vez obtenidas las dimensiones, es importante detectar la relación entre ellas. Una dimensión puede tener varias dependencias respecto de otras; sin embargo, lo más conveniente es especificarlas por separado. Las dependencias entre dimensiones, también se obtuvieron a partir de las encuestas realizadas, debido a que en el formato de encuesta se solicita respecto a que dimensiones se quiere consultar cada indicador.

A continuación se muestra la tabla con las dependencias existentes:

El sistema IHEMSyS tiene N subsistemas
Un subsistema tiene N zonas Una zona tiene N subsistemas
Un subsistema tiene N planteles
Un plantel pertenece sólo a un subsistema
Una localidad tiene una zona
Una zona tiene N localidades
Una zona tiene N planteles
Un plantel tiene N grupos
Un plantel tiene N docentes
Un grupo tiene N alumnos
Un alumno tiene un grupo
Un docente tiene varios grupos
Un docente pertenece a N planteles
Un plantel tiene N aulas

Tabla 4.7 Dependencia entre dimensiones

4.2.4 Granularidad y Horizonte de tiempo

Como se mencionó en el capítulo anterior, la granularidad consiste en especificar sobre que dimensiones es analizable cada indicador. A continuación se muestra una tabla p< cada Departamento, en cada una de las cuales se describe la dependencia entre cada indicador con sus correspondientes dimensiones.

Departamento Académico	
Indicador	**Dimensiones**
Deserción	Periodo, zona, localidad, subsistema, plantel, sexo, turno, escuela_ procedencia factor deserción
Ingresos, Egresos Aspirantes Revalidaciones	Periodo, zona, localidad, subsistema, plantel, sexo, escuela_ procedencia
Regulares Irregulares Repetidores	Periodo, zona, localidad, subsistema, plantel, turno y sexo
Becados	Periodo, zona, localidad, subsistema, plantel, tipo_beca, sexo, turno y dependencia.
Inasistencias	Periodo, localidad, zona, subsistema, plantel, categoría, sexo y turno
Aprobación Reprobación	Periodo, zona, localidad, subsistema, plantel, turno, sexo, semestre
Promedio docente	Periodo, zona, localidad, subsistema, plantel, sexo y categoría

Contratación Basificados Eventuales Despidos	Periodo, zona, localidad, subsistema, plantel, turno, sexo y categoría
Antigüedad	Periodo, zona, localidad, subsistema, plantel, antigüedad, turno y sexo
Grado de estudios	Periodo, zona, localidad, subsistema, plantel, turno, sexo, institución y especialidad
Incapacidades	Periodo, zona, localidad, subsistema, plantel, institución médica, turno y sexo

Tabla 4.8 Granularidad de los indicadores del departamento académico

Como se observa en la tabla 4.9, algunos indicadores pueden depender de las mismas dimensiones. Por ejemplo, los indicadores ingresos, egresos y aspirantes, dependen de las mismas dimensiones. Por lo tanto, en el diseño estos indicadores se pueden representar en un mismo diagrama.

Departamento de Recursos Financieros	
Indicador	**Dimensiones**
Gastos en pago a personal	Período, zona, localidad, subsistema, plantel, tipo personal
Gastos en compra de equipo	Periodo, zona, localidad, subsistema, plantel, tipo equipo
Gastos en mantenimiento, viáticos, Consumibles, publicidad, eventos, incentivos, construcciones	Periodo, zona, localidad, subsistema, plantel
Gastos en becas	Periodo, zona, localidad, subsistema, plantel, tipo beca
Gastos en cursos de actualización	Periodo, zona, localidad, subsistema, plantel, tipo_curso
Gastos en bibliografía	Periodo, zona, localidad, subsistema, plantel, tipo_bibliografía

Tabla 4.9 Granularidad de los indicadores del Departamento de Recursos Financieros

Departamento de Recursos Materiales	
Indicador	**Dimensiones**
Aulas	Periodo, zona, localidad, subsistema, plantel
Mobiliario en uso, adquirido, en reparación, dado de baja	Periodo, zona, localidad, subsistema, plantel, Tipo_mobiliario
Equipo de cómputo en uso, en reparación, adquirido, dado de baja	Periodo, zona, localidad, subsistema, plantel, Tipo_equipo_computo
Equipo electrónico en uso, en reparación, adquirido, dado de baja	Periodo, zona, localidad, subsistema, plantel, Tipo_equipo_eléctrico
Libros en uso, adquiridos, perdidos	Periodo, zona, localidad, subsistema, plantel

Tabla 4.10 Granularidad de los indicadores del Departamento de Recursos Materiales

Horizonte de tiempo

Esta fue le última de las tareas realizadas para la obtención de los elementos del análisis multidimensional. Es importante recordar que el horizonte de tiempo de un indicador representa el periodo por el que es necesario analizarlo. En la siguiente tabla se muestran los horizontes de tiempo correspondientes a cada indicador.

Indicador	Horizonte de tiempo
Departamento Académico	
Deserciones, Inasistencias	Mensual mente
Regulares, Irregulares, Repetidores, Aprobación, Reprobación, Promedio docente, Incapacidades	Bimestralmente
Ingresos, Egresos, Aspirantes, Revalidaciones, Becados Contratación, Basificados, Eventuales, Despidos, Antigüedad, Grado de estudios	Semestralmente.
Departamento de Recursos Financieros	
Gastos en pago a personal, Gastos en mantenimiento, viáticos, consumibles, publicidad, eventos, incentivos, construcciones Ingresos de trámites, donativos, eventos, cuotas, permisos de ventas	Mensualmente
Gastos en compra de equipo, en becas	Trimestralmente
Gastos en cursos de actualización, en bibliografía Ingresos de fichas de examen, inscripciones, reinscripciones, exámenes de recuperación, exámenes especiales	Semestralmente
Departamento de Recursos Materiales	
Aulas, mobiliario adquirido, equipo electrónico adquirido, libros adquiridos	Semestralmente
Mobiliario en uso, en reparación, dado de baja Equipo de cómputo en uso, reparación, dado de baja Equipo electrónico en uso, en reparación, dado de baja Libros en uso, perdidos	Trimestralmente

Tabla 4.11 Tabla de horizontes de tiempo

Capitulo 5

Diseño Multidimensional

Este capítulo contiene el diseño multidimensional efectuado para el Departamento Académico, en el capítulo tres se mostró en la figura 3.1, todas las actividades realizadas para crear el DW del IHEMSyS. A continuación, en la figura 5.1 se muestran las actividades que se llevaron a cabo para efectuar el diseño multidimensional.

Diagrama 5.1 Actividades realizadas en el diseño

En la sección 3.3 del capítulo 3, se explicó en forma resumida como se realizó cada una de las actividades mostradas en la figura 5.1. En las siguientes secciones, se describe el procedimiento que se llevó a cabo en cada actividad u se muestran sus resultados.

5.1 ESQUEMAS DE ESTRELLA Y SNOWFLAKE

Una vez seleccionada la Arquitectura del DW, la siguiente tarea realizada fue el diseño de los esquemas de estrella u de snowflake, para los diferentes indicadores analizados en el capítulo anterior. Es importante mencionar, que para efecto del presente proyecto, se realizó el diseño multidimensional únicamente para el departamento académico del IHEMSyS, por ser éste el más importante.

5.1.1 Esquemas estrella

Como se explicó en el capítulo anterior, un esquema estrella consta de un grupo de tablas. En la parte central se coloca la tabla de hechos que contiene la clave principal de cada una de las tablas de dimensión y los indicadores obtener. Alrededor de la tabla de hechos se colocan las tablas de dimensión, las cuales contienen su clave principal y los

campos necesarios para la propia dimensión. A continuación se muestran los diagramas ele estrella délos distintos indicadores correspondientes al Departamento Académico.

ESQUEMAS ESTRELLA

Esquema para obtener índices de deserción (por período, sexo, turno, localidad, zona, subsistema, nivel, plantel, factor_deserción y escuela Procedencia)

Este esquema permite obtener los índices de deserción, consta de una tabla de hechos y seis tablas de dimensión (Período, Factor-Deserción, Plantel, Escuela _ procedencia, Sexo y Turno). Como se muestra en la figura 6.1, no se incluyen las dimensiones localidad, tona y subsistema» puesto que se incluyen en la dimensión plantel

La tabla ele hechos está formada por una llave compuesta de las llaves foráneas de las tablas dimensionales y el indicador a obtener (numero de deserciones), Las líneas que unen cada tabla de dimensión a la tabla de hechos, indican una relación que existe entre ellas por medio del campo clave.

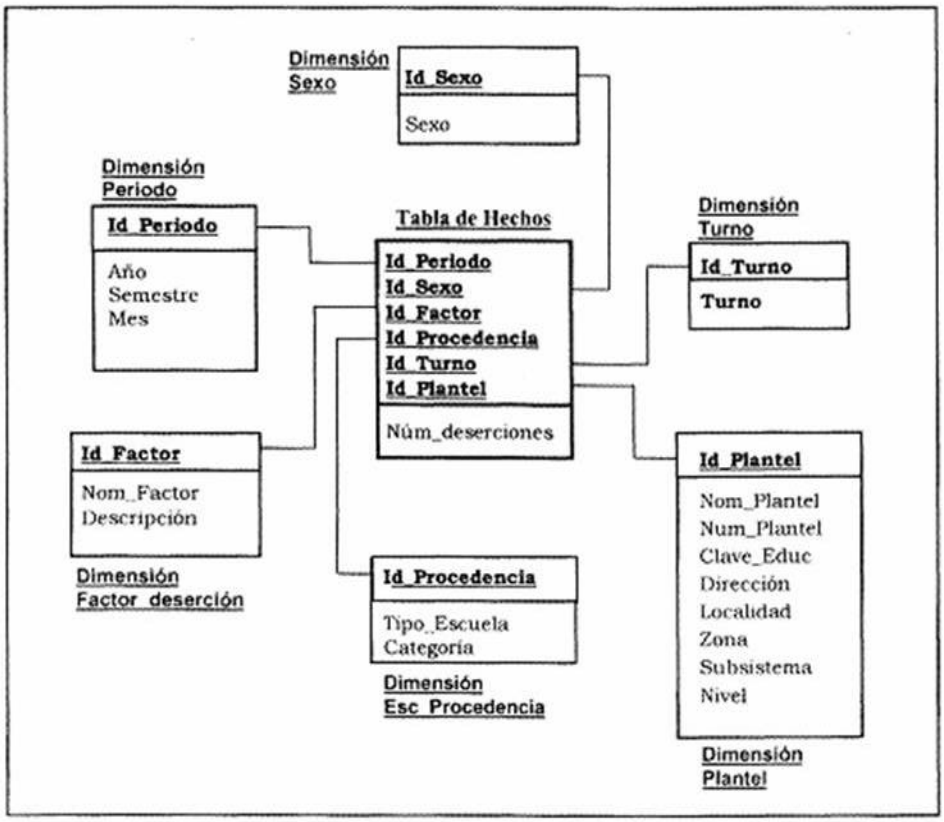


Diagrama 51 índices de deserción

La dimensión Plantel permite saber su nombre, y a qué subsistema, nivel» municipio y zona geográfica pertenece. La tabla Factor_Deserción, permite manejar información ele los motivos por los cuales un alumno deserta (ejemplo, por problemas económicos, familiares, sociales, etc. Con la tabla Esc_Procedencia, se puede tener control del tipo de escuela que proceden los alumnos desertados, normalmente a este dato también se le conoce como área de afluencia. Por último, en la tabla Periodo, a diferencia del esquema anterior, se puede observar que los intervalos de tiempo que se manejan para la obtención de estos indicadores son: mensual, trimestral y anualmente.

Esquema pauta obtener índices de aspirantes, ingresos, egresos q revalidaciones (por periodo, localidad, zona, subsistema, nivel, plantel, procedencia, sexo g turno)

En el siguiente esquema, en la tabla de techos, además de contener las llaves foráneas de las tablas dimensionales, cuenta con los campos que almacenaran los indicadores aspirantes, ingresos y egresos. El indicador aspirantes permitirá obtener el número de alumnos que solicitan ficha para examen de admisión, el indicador ingresos se refiere a los alumnos que se inscriben por primera vez en un plantel, revalidaciones son alumnos que ingresan de otras escuelas revalidando materias y el indicador egreso respecto a los que terminan sus estudios. Los tres indicadores se unen en la misma tabla, debido a que trabajan con las mismas dimensiones

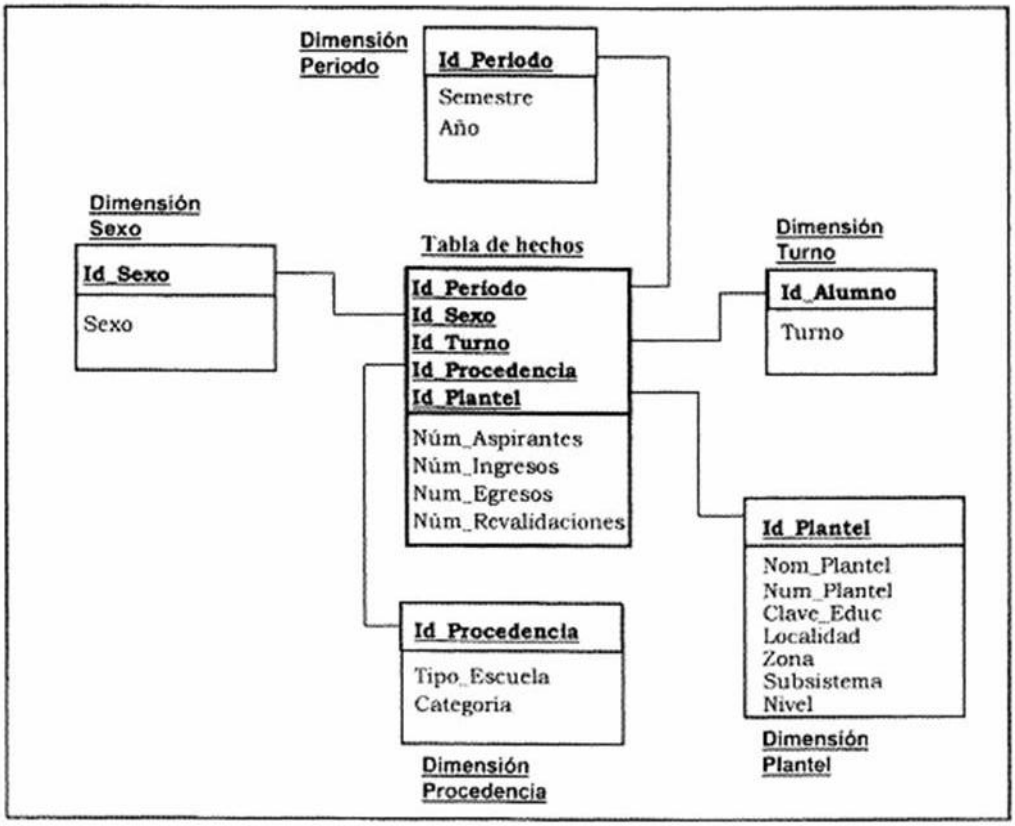


Diagrama 5.2 índices de aspirantes, ingresos y egresos

Es importante resaltar que a diferencia del esquema anterior» en la dimensión periodo, no se incluye el campo mes, esto debido a que la obtención de sus tres indicadores sólo se necesitan realizar por semestre o ano, Al igual que en el indicador anterior, para obtener índices de ingresos y egresos, se utiliza la dimensión procedencia, lo cual permitirá saber de que tipo de escuela ingresan 13 egresan mas alumnos.

Esquema estrella para obtener índices de alumnos reculares, irregulares y repetidores (por periodo, localidad, zona, subsistema, nivel, plantel, turno y sexo)

En el siguiente esquema, la tabla de hechos contiene los campos de los siguientes tres indicadores a obtener: cantidad de alumnos regulares (aquellos que durante el semestre no reprobaron ninguna materia), cantidad de alumnos irregulares (quienes durante el semestre reprobaron de una a tres materias y por último, la cantidad de repetidores (alumnos que reprobaron mas de tres materiales repiten semestre).

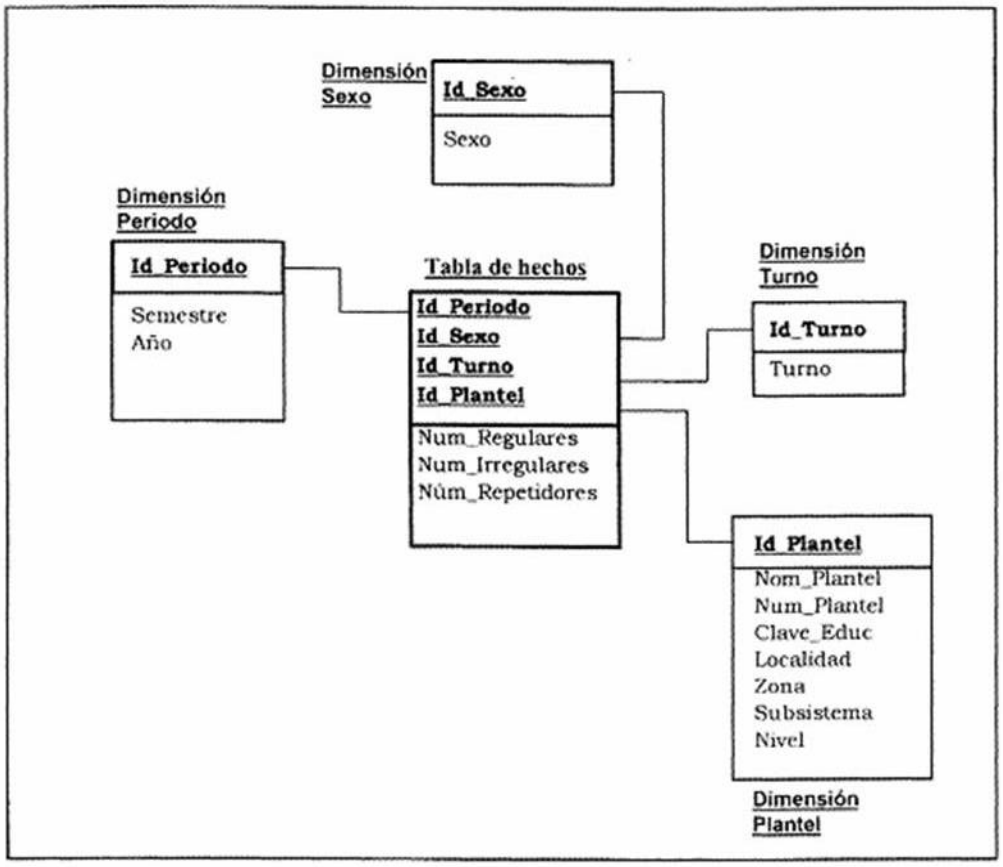


Diagrama 5.3 índices de regulares, irregulares y repetidores

Esquema estrella paira obtener índices de inasistencias
(por categoría, sexo, período, localidad, zona, subsistema, nivel, plantel y turno)

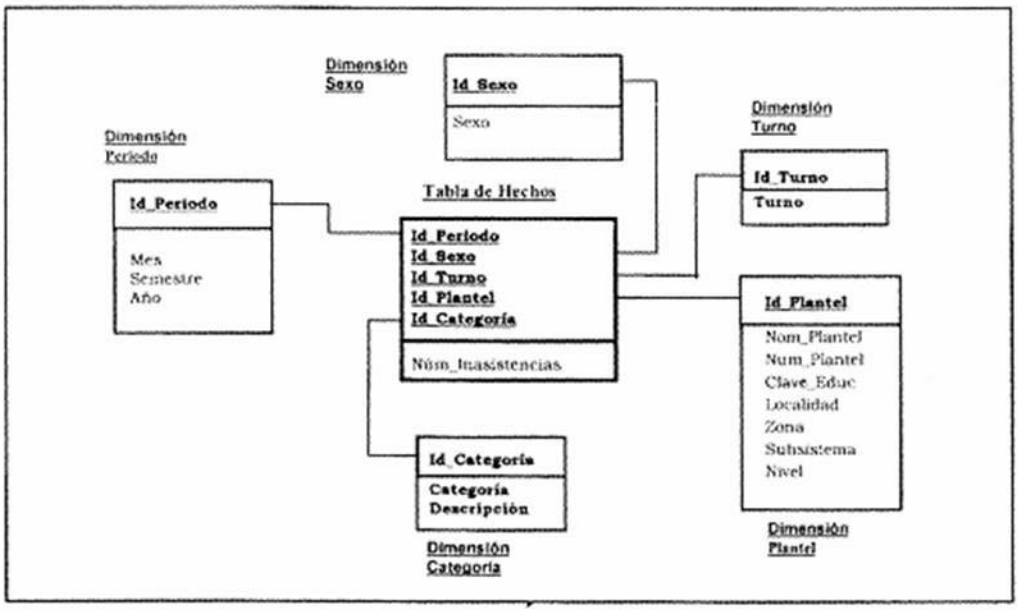


Diagrama 5.4 índices de inasistencias

Esquema estrella para obtener índices de becas
(por periodo, localidad, zona, subsistema, nivel, plantel, tipo_beca, dependencias, sexo y turno)

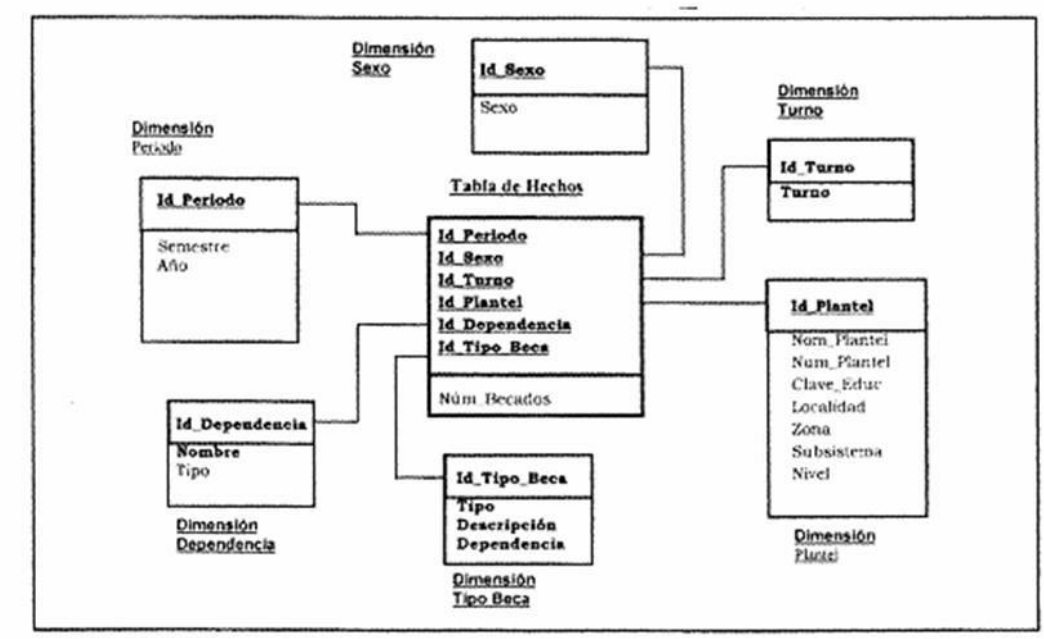


Diagrama53 índices debocas.

La dimensión tipo-beca, permite saber el tipo de apoyo otorgado al alumno (ejemplo, beca de transporte, colegiatura, alimenticia, etc. La dimensión dependencia aijada a saber que institución proporciona la teca y si dicha institución es educativa o de gobierno.

Esquema para obtener índices de personal contratado, basificado eventual y despedido (Por periodo, localidad, zona, subsistema nivel plantel categoría, sexo nocturno)

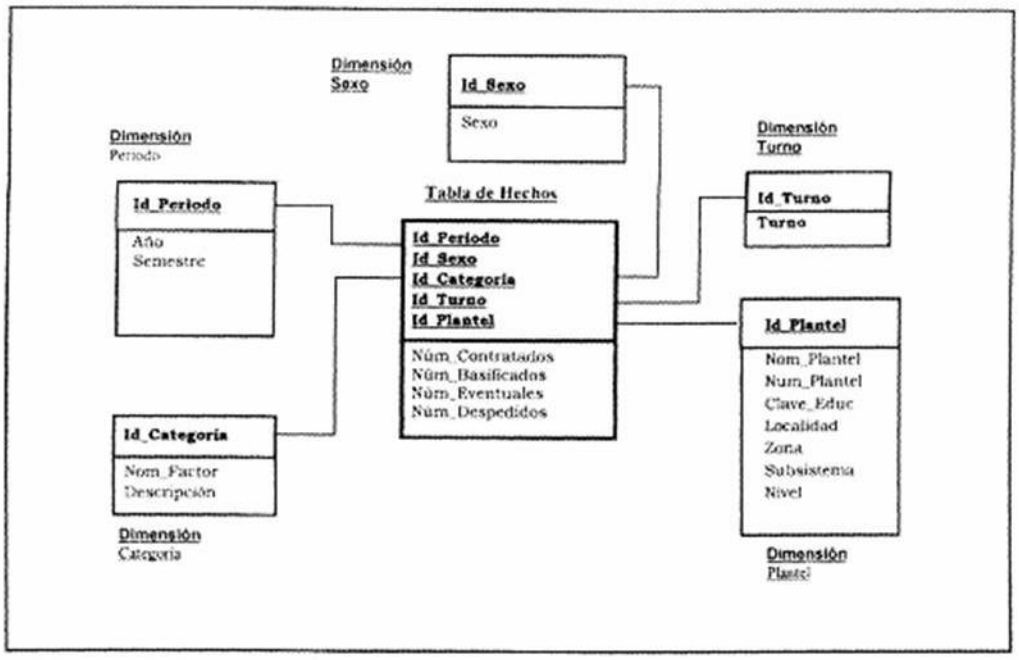


Diagrama 5.6 índices de personal contratadlo, basificado, eventual y despedido

Esquema para obtener índices de aprobación u reprobación (por periodo localidad, zona, subsistema nivel, plantel, sexo y tumo)

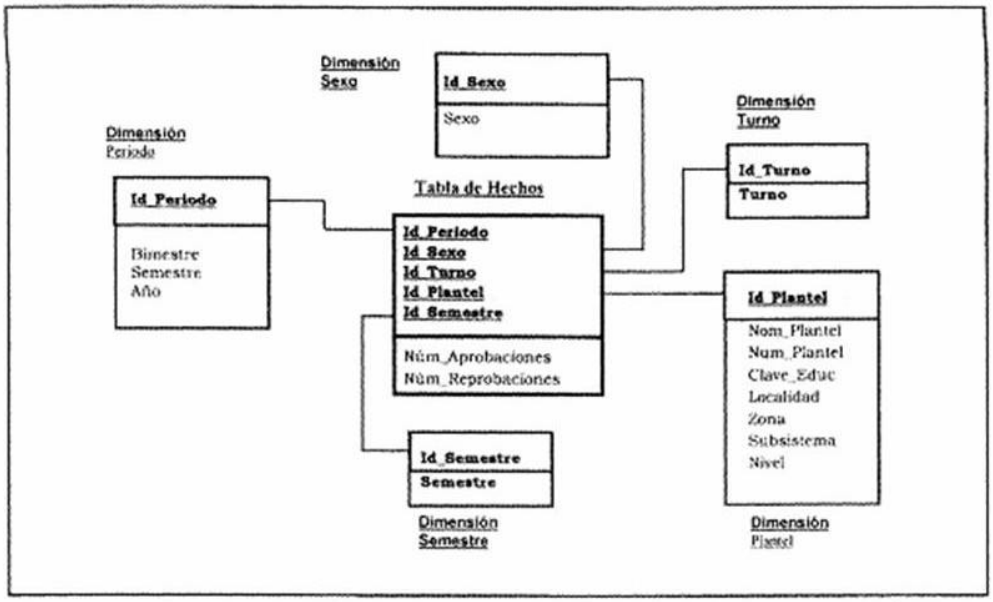


Diagrama 5.7 Índices de aprobación y reprobación

Esquema para obtener índices de promedio docente
(por periodo, localidad, nona, subsistema, nivel, plantel, categoría y sexo)

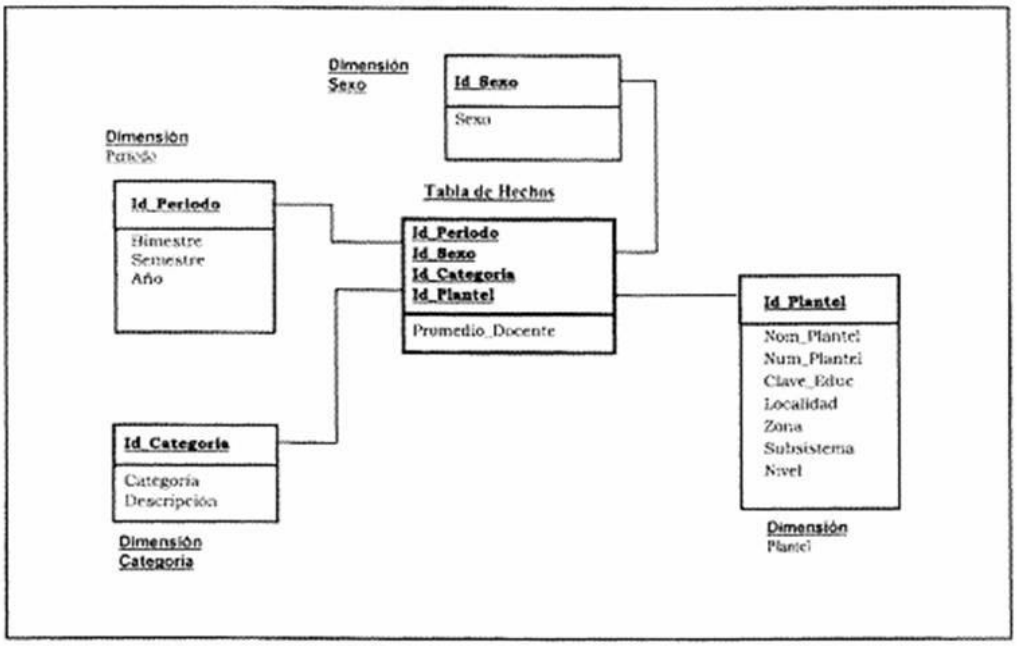
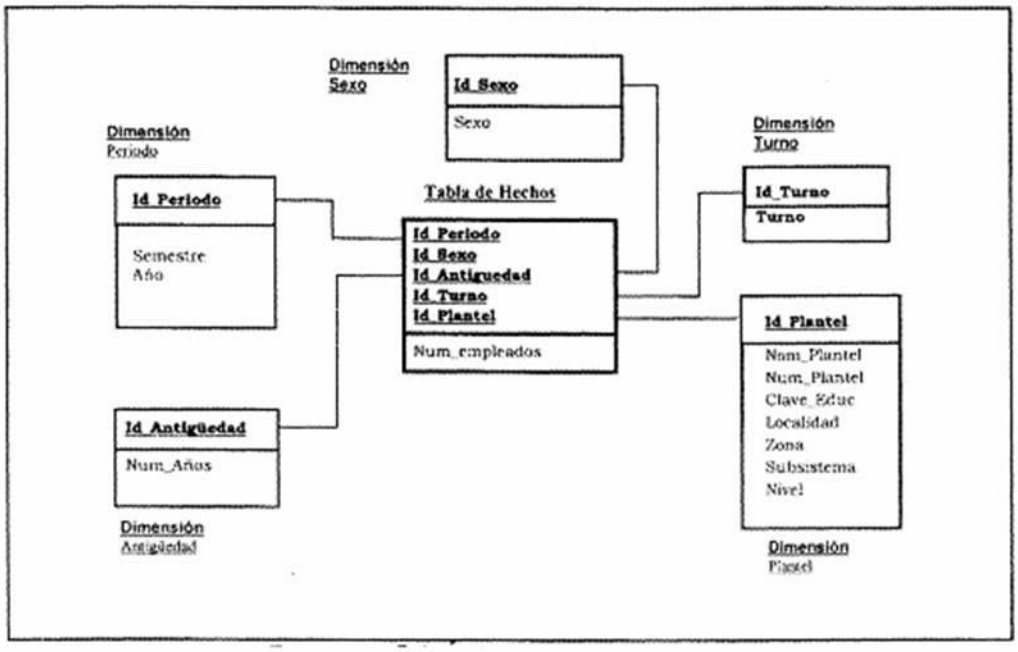


Diagrama 5.8 índices de promedio docente

Esquema para obtener índices de antigüedad
(por periodo, localidad, zona, subsistema, nivel, plantel, antigüedad, sexo q temo)



Diagrma5.9 Índices de antigüedad

Esquema para obtener índices de incapacidades
(por período, localidad, sena, subsistema, nivel, plantel, institución médica, sexo y turno)

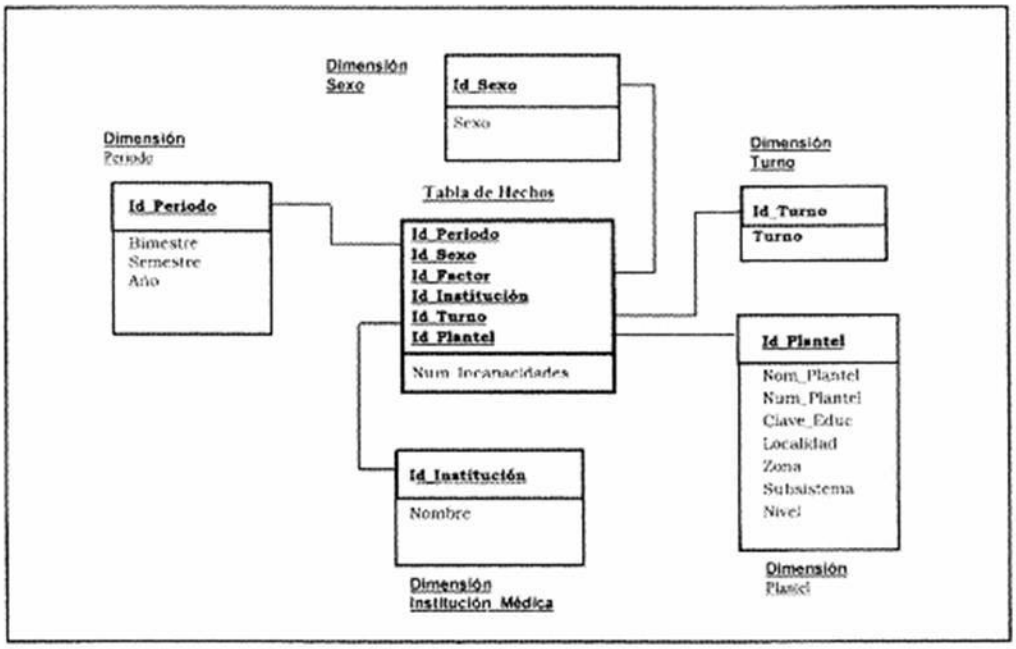


Diagrama 5.10 índices de incapacidades

Esquema para obtener índices de grado de estudios
(por periodo, localidad, zona, subsistema, nivel, plantel, especialidad, institución, sexo y tumo)

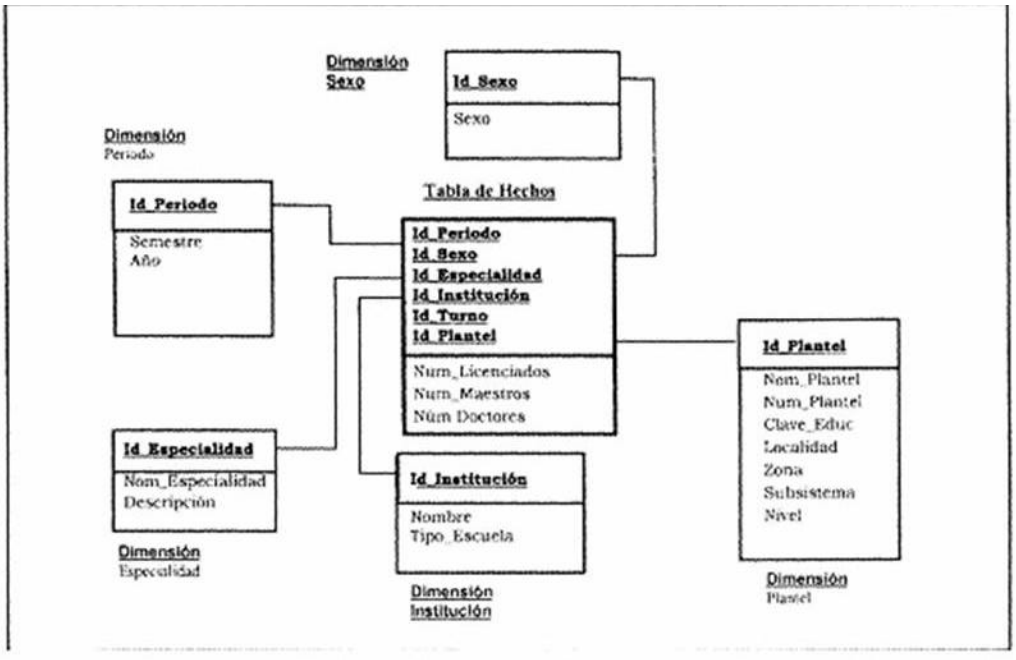


Diagrama 5.11 índices de docentes con Licenciatura ti Maestría

5.12 ESQUEMAS SNOWFLAKE

Recordemos que los esquemas snowflake, al igual que los estrella, también contienen tablas ele hedías y dimensiones, solo que sus tablas de dimensión están normalizadas. Las tablas de dimensión contienen únicamente el campo clave de la tabla y la llave foránea del nivel más cercano. Es importante mencionar, que a diferencia de los esquemas estrella, en este tipo de esquemas se omiten los campos restantes de las tablas de dimensión. A continuación se muestran los diagramas snowflake de los distintos indicadores:

Esquema snowflake para obtener índices de deserción

(por período sexo, turno, localidad, zona, subsistema, nivel, plantel, factor_deserción y escuela Procedencia)

Este esquema, consta de una tabla de hechos y tablas normalizadas en las que cada tabla dimensional contiene sólo el nivel de detalle (clave primaria en la tabla) y la llave foránea de su parentesco del nivel más cercano.

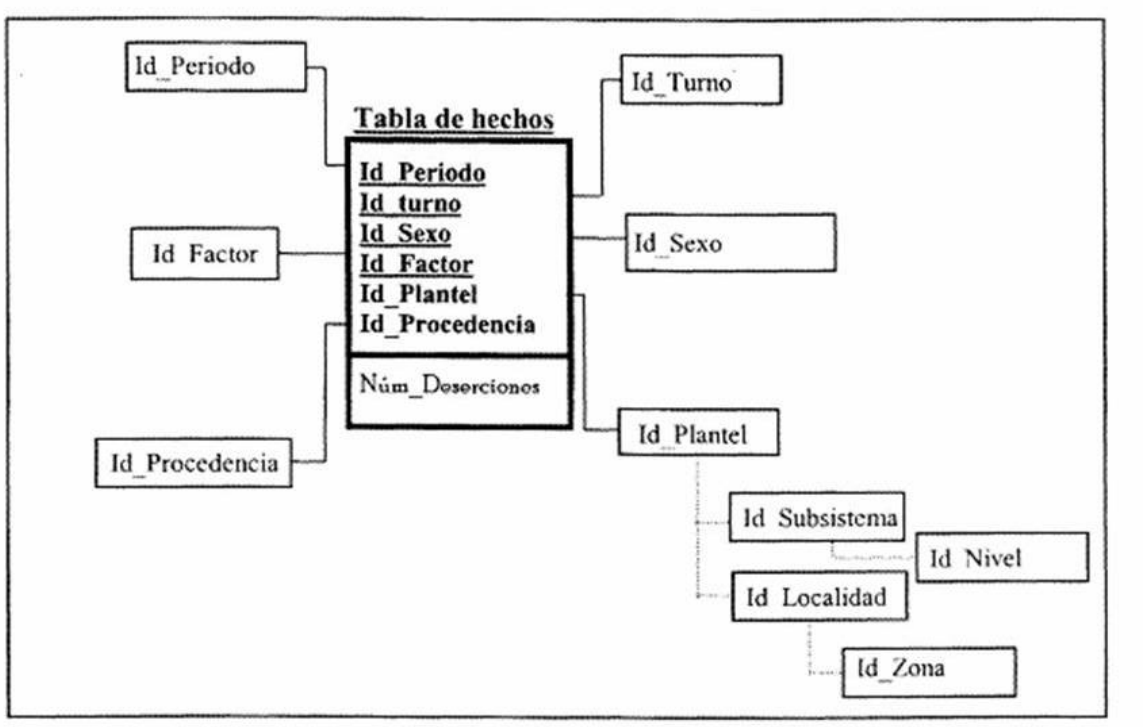


Diagrama 5.12 índices de deserción

En el presente diagrama, como se puede observar, a diferencia de su equivalente diagrama de estrella, las tablas de dimensión están normalizadas. La dimensión plantel ahora está conformada por las dimensiones localidad y subsistema, y a su vez de localidad se deriva zona y subsistema de nivel.

Esquema snowflake para obtener índices de aspirantes, ingresos y egresos (por periodo, localidad, zona, subsistema, nivel, plantel, escuela_procedencia, sexo y tumo)

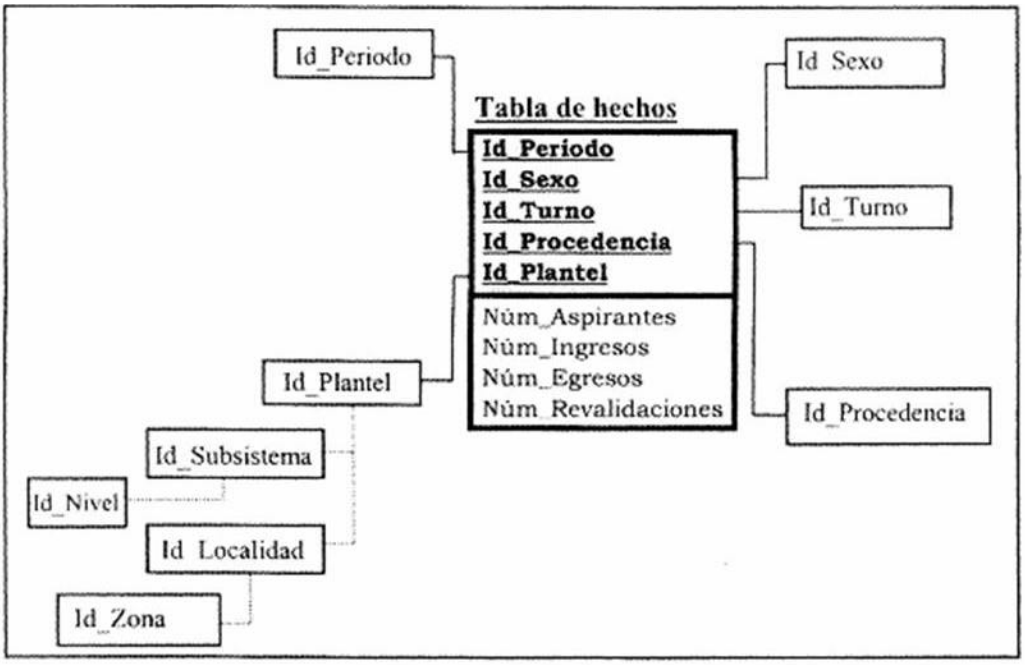


Diagrama 5.13 índices de ingresos y egresos y aspirantes

Esquema snowflake para obtener índices de alumnos regulares, irregulares y repetidores (por periodo, localidad, zona, subsistema, nivel, plantel, humo y sexo)

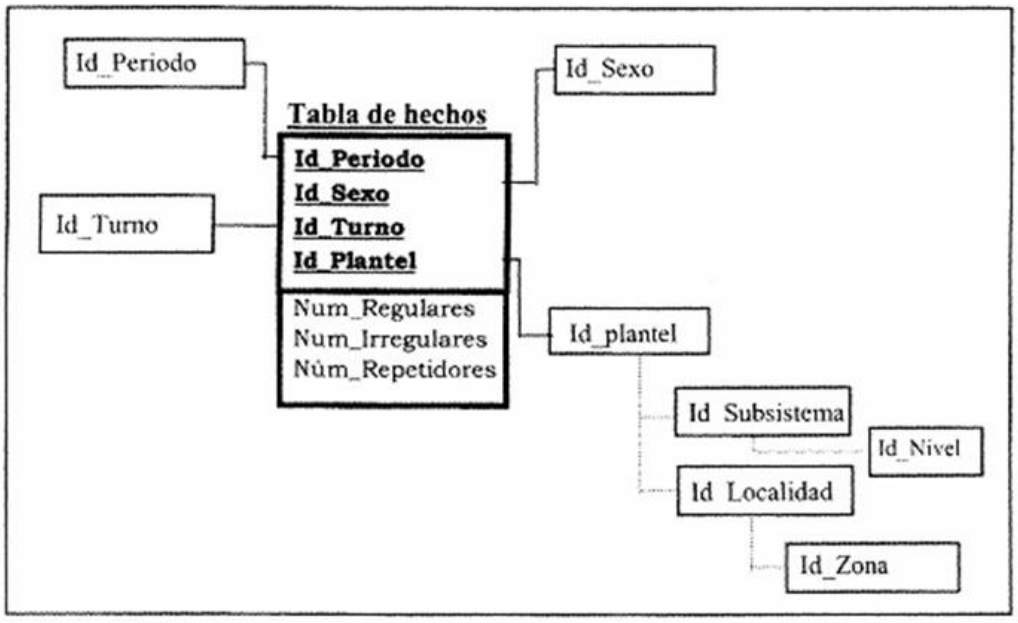


Diagrama 514 índices de alumnos regulares, irregulares y repetidores

Esquema snowflake para obtener índices de inasistencias
(por categoría, sexo, periodo, localidad, zona, subsistema, nivel, plantel y turno)

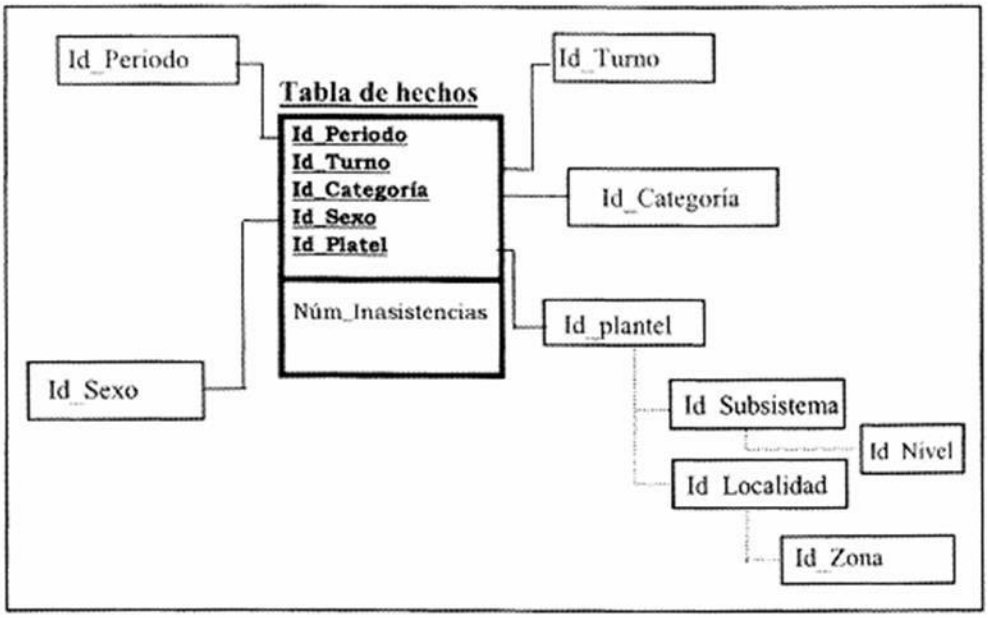


Diagrama 5.15 índices de inasistencias

Esquema snowflake para obtener índices de becas
(por periodo, localidad, zona, subsistema, nivel, plantel, tipo-beca, dependencia, sexo y turno)

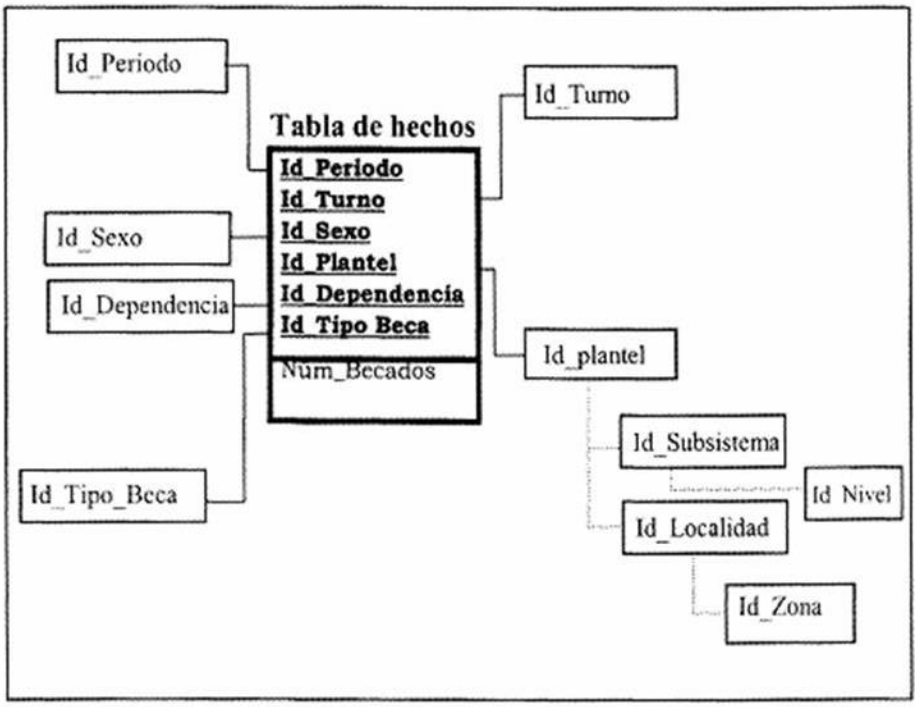


Diagrama 5.16 índices de becas

Esquema para obtener índices de personal contratado, basificado, eventual y despedido (por periodo, localidad, zona, subsistema, nivel, plantel, categoría, sexo y turno)

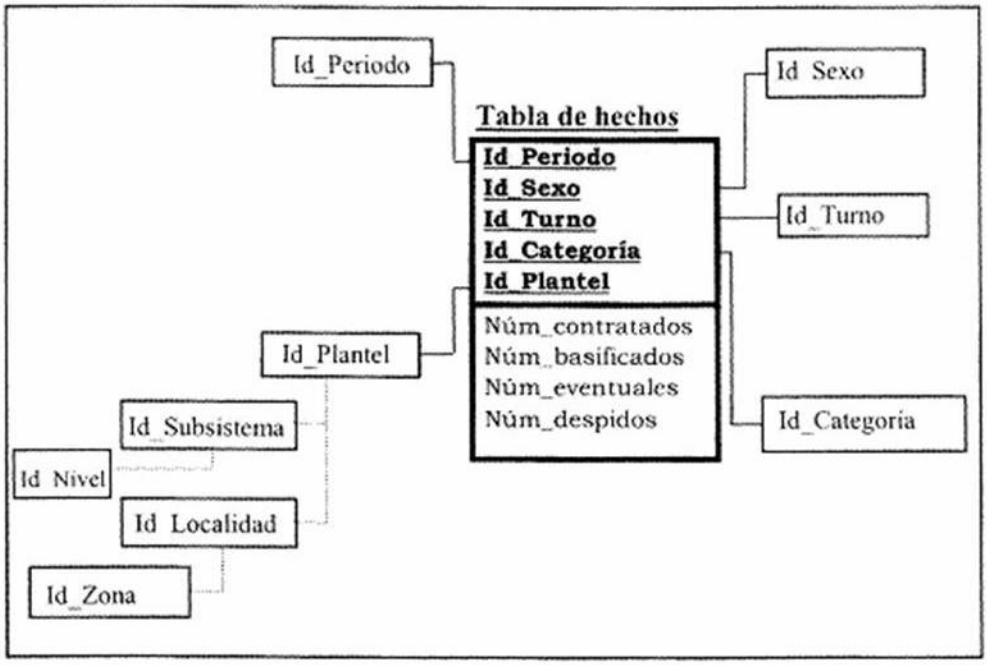


Diagrama 5.17 Índices de personal contratado, basificado, despedido y eventual

Esquema para obtener índices de aprobación y reprobación (por periodo, localidad, zona, subsistema, nivel, plantel, semestre, sexo y turno)

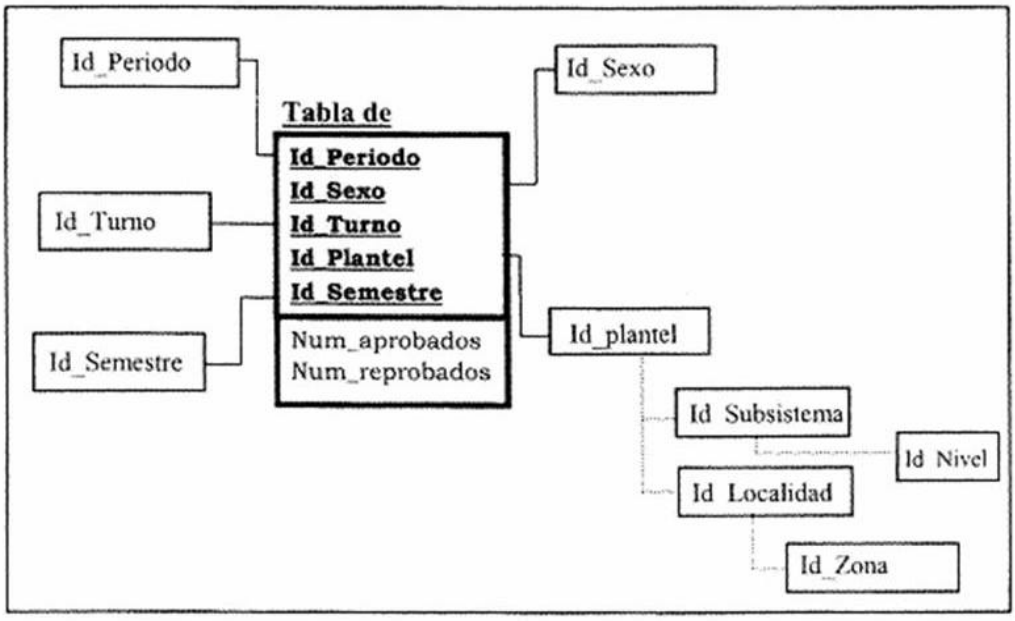


Diagrama 5.18 índices de aprobación y reprobación

Esquema para obtener índices de promedio docente
(por periodo, localidad, zona, subsistema, nivel, plantel, categoría y sexo)

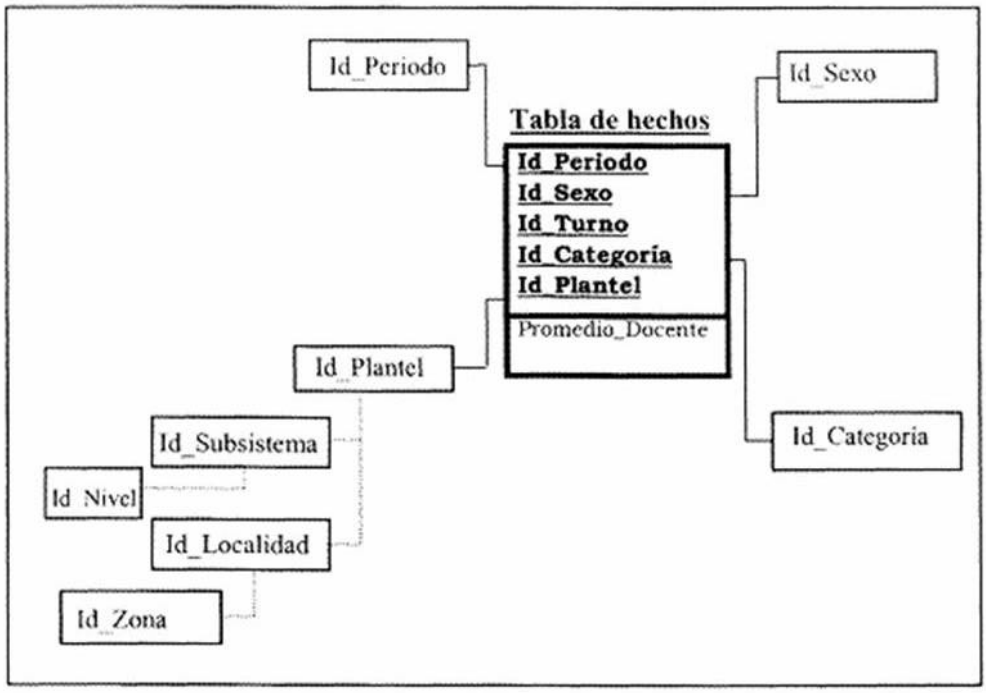


Diagrama 5.19 índices de promedio docente

Esquema para obtener índices de antigüedad
(por periodo, localidad, zona, subsistema, nivel, plantel, antigüedad, sexo g turno)

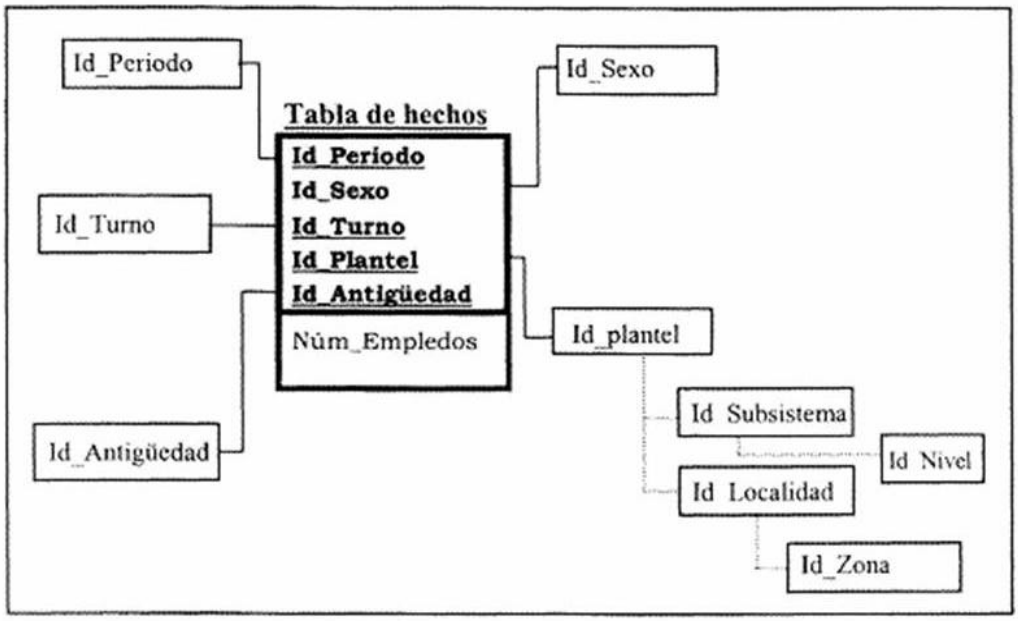


Diagrama 5.20 índices de antigüedad

Esquema para obtener índices de grado de estudios
(por periodo, localidad, tona, subsistema, nivel, plantel, especialidad, institución y sexo)

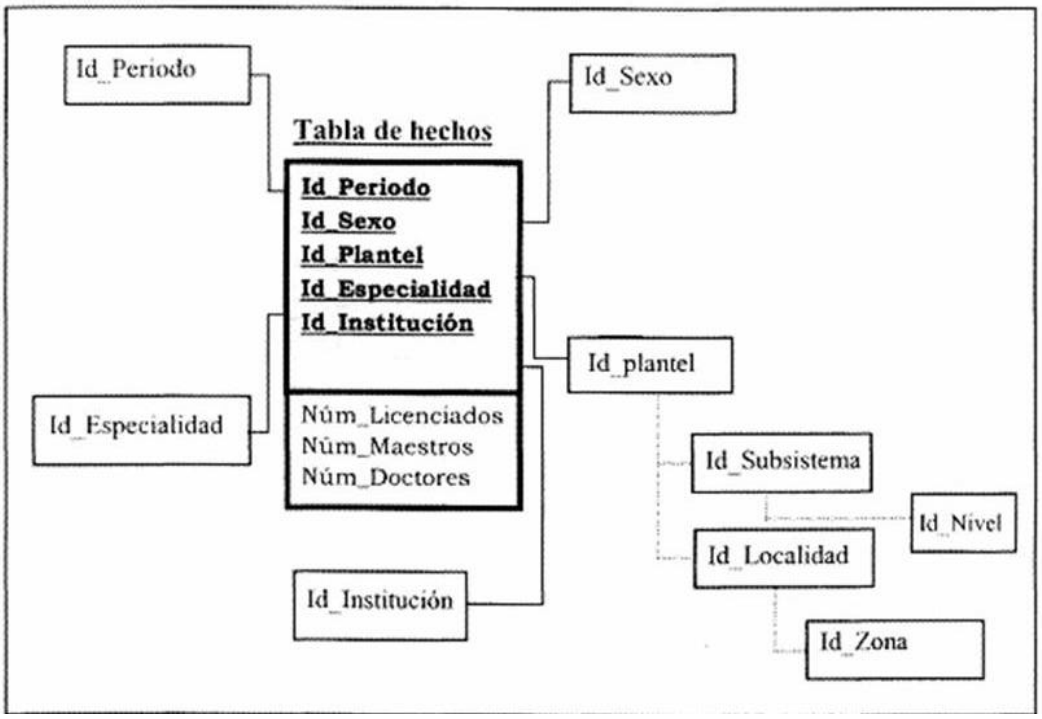


Diagrama 5.21 índices de docentes con licenciatura y Maestría

Una vez realizado los diagramas de estrella y snowflake, el siguiente paso consiste en analizar las características de los campos que se utilizan en las distintas tablas de hechos y de dimensiones,

5.2 TABLAS DE HECHOS Y DIMENSIONES

Hasta el momento, en los diagramas de dimensión se han especificado cuales son las tablas (hechos y dimensiones) que se involucran en la obtención de los diversos indicadores, el siguiente paso consiste en mostrar a detalle la información del diseño de cada una de las tablas, es decir describir las características de cáela uno de sus campos.

Las tablas de hechos contienen los campos clave de cada tabla de dimensión y los campos de los indicadores, A continuación se muestran las tablas de hechos que se utilizaron en los esquemas ele estrella y snowflake.

Tablas de hechos

Campos Clave		
Atributo	Tipo	Descripción
Id _ Período	Int	Referencia a dimensión tiempo
Id_Localidad	I*nt*	Referencia a dimensión localidad
Id _ zona	Int	Referencia a dimensión zona
Id_Subsistema	Int	Referencia a la dimensión subsistema
Id_Nivel	Int	Referencia a la dimensión nivel
Id_Plantel	Int	Referencia a dimensión plantel
Id_sexo	Int	Referencia a dimensión sexo
Id,_Turno	Int	Referencia a la dimensión turno
Id_Semestre	Int	Referencia a la dimensión semestre
Id_Escuela Procedencia	Int	Referencia a dimensión procedencia
Id_Categoría	Int	Referencia a dimensión categoría
Id_tipo Beca	*Int*	Referencia a la dimensión tipo beca
Id_Dependencia	Int	Referencia a la dimensión dependencia
Id_Antigüedad	Int	Referencia a la dimensión antigüedad
Id_Factor Deserción	Int	Referencia a dimensión factor deserción
Id_Institución	Int	Referencia a dimensión Institución
Id_Especialidad	Int	Referencia a la dimensión especialidad
Id_Institución Medica	Int	Referencia a la dimensión institución médica
Indicadores		
Atributo	**Tipo**	**Descripción**
Num_Aspirantes	Int	**Alumnos que realizan examen para ingresar a un plantel**
Núm_Ingresos	Int	Alumnos que se inscriben en el plantel
Núm_Revalidaciones	Int	Alumnos que ingresan por cambio de plantel, revalidando materias
Num_Egreso	Int	Alumnos que terminan sus estudios
Núm_Deserciones	Int	Alumnos que son dados de baja de *un* plantel
Núm_Aprobados	Int	Alumnos *que* acreditan una materia
Núm_Reprobados	Int	Alumnos que no acreditan *una* materia
Núm_Regulares	Int	Alumnos *que en* el semestre no reprobaron ninguna materia
Num_Irregulares	Int	Alumnos que en el *semestre* reprobaron de 1 a 3 materias
Num_Inasistencias	Int	Porcentaje de inasistencia
Núm_Repetidores	Int	Alumnos que en el semestre reprobaron más de 3 materias
Num_Becados	Int	Cantidad de alumnos, a los que se les proporciona alguna beca
Promedio_docente	Int	Promedio general del docente
Núm_Inasistencias	Int	Porcentaje de inasistencia del docente
Núm-Contrataciones	Int	Número de docentes que son contratados
Núm_Basificados	Int	Número de docentes de tiempo completo
Núm_Eventuales	Int	Número de docentes que trabajan por toras
Núm_Despidos	Int	Número de decentes que se despiden en los planteles
Num_Empleados	Int	Número de empleados con determinada antigüedad
Núm_Licenciados	Int	Número de docentes con título de licenciatura
Núm_Maestros	Int	Número de docentes con grado de maestría
Núm_Doctores	Int	Número de docentes con grado de doctorado
Num _ncapacidades	Int	Número de incapacidades de los docentes

Tabla 51 Tabla de hechos

Tablas de dimensión

Las tablas de dimensión contienen su campo clave y todos los campos necesarios para lograr obtener la información de los indicadores. A continuación se muestran las tablas de dimensión utilizadas en los diagramas de estrella y snowflake.

Tabla de dimensión Periodo

Atributo	Tipo	Descripción
Id_periodo	Int	Llave de la dimensión
Año	Int	Año (entero de cuatro dígitos)
Num_mes	Int	Mes(enero .. diciembre)
Bimestre	Int	Bimestre (1..6)
Trimestre	Int	Trimestre (1...4)
Semestre	Int	Trimestre (1..2)

Tabla 5.2 Dimensión Tiempo

Tabla de dimensión Zonas

Atributo	Tipo	Descripción
Id_zona	*Int*	Llave de la dimensión
Nom_zona	Char(15)	Nombre de la zona
Descripción	Char(15)	Tipo de zona

Tabla 53 Dimensión Zona

Tabla de dimensión Subsistemas

Atributo	Tipo	Descripción
Id_Subsistema	Int	Llave de la dimensión
Nom_nivel	Char(10)	Nombre del subsistema
Descripción	Char (25)	Tipo de subsistema
Id_Nivel	Int	Clave del nivel

Tabla 5,4 Dimensión Subsistema

Tabla de dimensión Nivel

Atributo	Tipo	Descripción
Id__Nível	Int	Llave de la dimensión
Nom_nivel	Char (10)	Nombre del nivel
Descripción	Char (25)	Tipo de nivel

Tabla 5.4 Dimensión Nivel

Tabla de dimensión Planteles

Atributo	Tipo	Descripción
Id _ plantel	Int	Llave de la dimensión
Nom plantel	Char (15)	Nombre del plantel
|Id_subsistema	Int	Relación cotí subsistema
Id_Localidad	Int	Relación con zona del plantel

Tabla 5.5 Dimensión Plantel

Tabla de dimensión Localidades

Atributo	Tipo	Descripción
Id_localidad	Int	Llave dé la dimensión
Nom_localidad	Char (15)	Nombre de *la* zona
Descripción	Char (15)	Tipo de zona,
Id_*zona*	*Int*	Relación con *zona* del plantel

Tabla 5.6 Dimensión Localidades

Tabla de dimensión Categoría

Atributo	Tipo	Descripción
Id_Categoría	Int	Llave de la dimensión
Descripción	Char (10)	Nombre de la categoría
Tipo	Char(lO)	Tipo ó nivel de la categoría

Tabla 5.7 Dimensión Categoría

Tabla de dimensión Factor deserción

Atributo	Tipo	Descripción
Id_factor	Int	Llave de la dimensión
Factor	Char(í5)	Nombre del factor
Descripción	Char (15)	Tipo de factor

Tabla 5.8 Dimensión factor _ deserción

Tabla de dimensión Escuela_Procedencia

Atributo	Tipo	Descripción
Id_típo escuela	Int	Llave de la dimensión
Tipo Escuela	Char (15)	Tipo do escuela
Descripción	Char (15)	

Tabla 5.9 Dimensión Escuela *de procedencia*

Tabla de dimensión Tipos_Becas

Atributo	Tipo	Descripción
Id_tipo_beca	Int	Llave de la dimensión
Descripción	Char (10)	Descripción del tipo de beca
Dependencia	Char (10)	Dependencia que la otorga

Tabla 5,10 Dimensión Tipos de beca

Tabla de dimensión Sexo

Atributo	Tipo	Descripción
Id_Sexo	Int	Llave de la dimensión
Sexo	Char (15)	Descripción del sexo

Tabla 511 Dimensión Sexo

Tabla de Dimensión turno

Atributo	Tipo	Descripción
Id _ turno	Int	Llave de la dimensión
Turno	Char (15)	Descripción del turno

Tabla 512 Dimensión turno

Tabla de dimensión Semestre

Atributo	Tipo	Descripción
Id_Semestre	*Int*	Llave de la dimensión
Semestre	Char (10)	Nombre del semestre

Tabla 5,13 Dimensión Semestre

Tabla de dimensión Institución

Atributo	Tipo	Descripción
Id _ institución	Int	Llave de la dimensión
Institución	Char (20)	Nombre de la institución
Tipo	Char (15)	Tipo de institución

Tabla 514 Dimensión Institución

Tabla de dimensión Especialidad

Atributo	Tipo	Descripción
Id _Especialidad	Int	Llave de la dimensión
Especialidad	Char (25)	Nombre de la especialidad

Tabla 515 Dimensión Especialidad

Tabla de dimensión Institución_Medica

Atributo	Tipo	Descripción
Id _ institución	Int	Llave de la dimensión
Institución	Char (20)	Nombre de la institución
Tipo	Char (13)	Tipo de institución

Tabla 516 Dimensión Institución Médica

53 ARQUITECTURA DEL DEPÓSITO Y DEL SERVIDOR

En el DW a desarrollar, el tipo ele arquitectura del depósito que se eligió es centralizado, debido a que se requiere que toda la información esté almacenada en un solo depósito, ya que en el IHEMSyS los distintos Departamentos, se encuentran en un mismo edificio.

Con este tipo de arquitectura se pueden obtener algunas ventajas, una de ellas es mejoramiento en el procesamiento de consultas, pues la velocidad de respuesta es mayor que en el caso de usar una arquitectura distribuida. Otra ventaja es el costo de soporte, debido a que se requiere de un solo medio de almacenamiento, además de que no se tienen que actualizar periódicamente varias fuentes de datos.

Respecto al tipo de servidor, se propone un servidor con multiprocesamiento simétrico, debido a que permite agregar procesadores, lo cual lo hace escalable, previendo el crecimiento del DW.

5.4 CREACIÓN DEL MODELO LÓGICO ESTÁNDAR

Debido a que el datamart a crear es el del Departamento Académico, en el presente trabajo se obtuvo sólo el modelo lógico de la base de datos de dicho Departamento. El modelo lógico, se creó en base a los resultados obtenidos en el análisis u diseño multidimensional., el cual se conformó por

Este modelo se conformó por las tablas de hechos de dimensiones obtenidas en el diseñó, y servirá de base para saber que información es necesario extraer de las bases de datos fuente.

De acuerdo a los resultados del análisis u diseño multidimensional, se detectó que existen algunas tablas en las bases de datos fuente, sin embargo, algunas son necesario de agregar. Las tablas que se pueden ocupar son las de la mayoría de las dimensiones obtenidas. Por ejemplo, planteles, docentes, tipos de becas, etc.

Sin embargo, fue necesario hacer agregación de tablas u realizar algunas modificaciones, de las cuales a continuación se describen las principales:

- Agregar algunas tablas de dimensión que no existen como: niveles, subsistemas, zonas, etc.
- Agregar las tablas de los diferentes periodos como, años, semestres, etc.
- Agregar todas las tablas que manejarán los hechos de los distintos indicadores, por ejemplo hechos de deserciones, becas, ingresos, etc. ^
- A algunas tablas existentes les faltan campos, principalmente los que se sirven de relación con las tablas faltantes. ^ Falta establecer relaciones entre diversas tablas.

Con la aplicación de todos estos cambios, se obtuvo el diseño lógico de la base de datos estándar que se muestra a continuación:

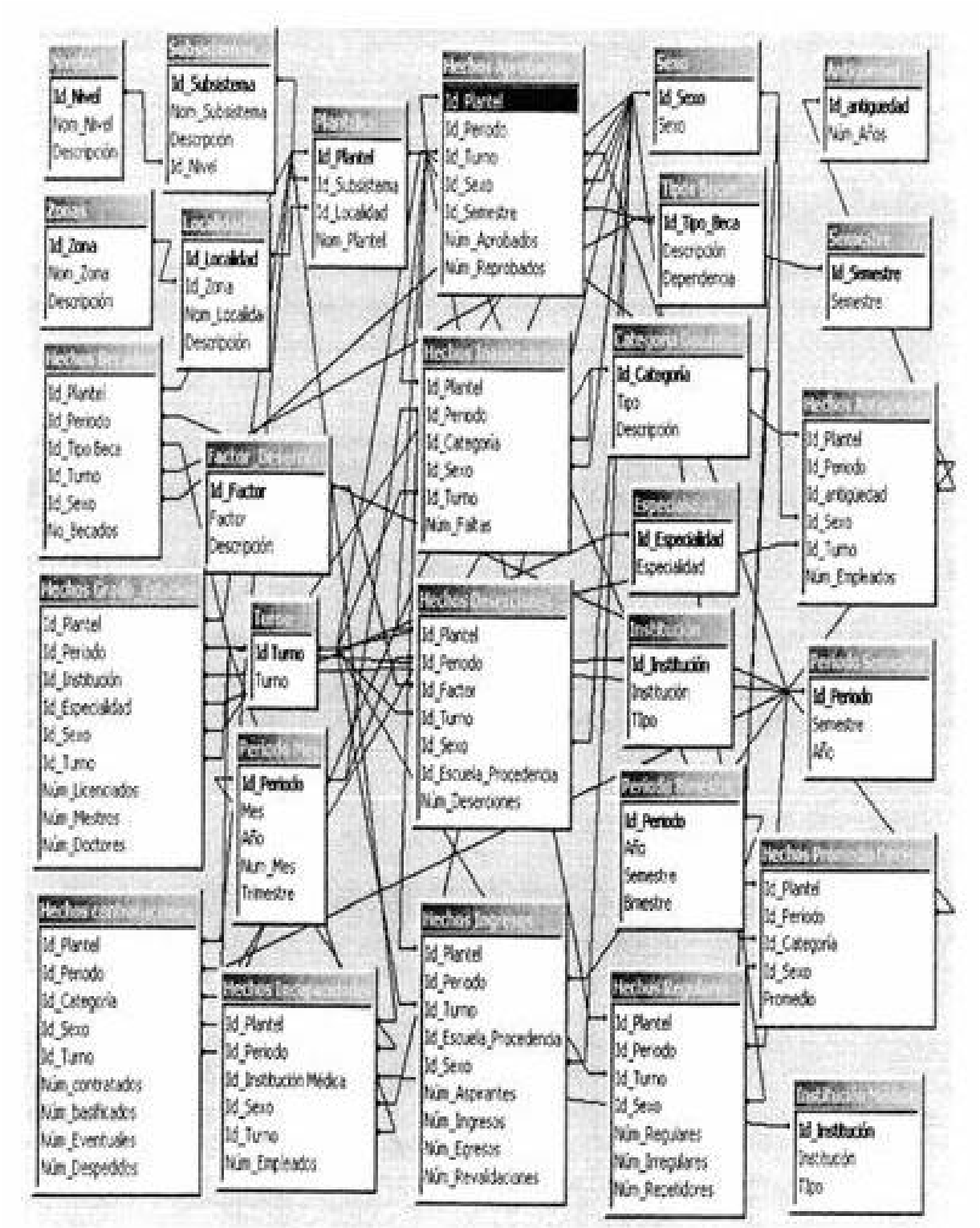


Figura 5.2 Diseño de BD estándar

Capitulo 6

Conversión de Datos

Continuando con el proceso de desarrollo, en el presente capítulo se expone la forma en que se convirtieron los datos con el uso de la herramienta Microsoft OLAP.

6.1 INTRODUCCIÓN

Para realizar el proceso de conversión de datos, en el caso de este proyecto, se utilizó la herramienta Microsoft OLAP. Las razones por las que se seleccionó la herramienta Microsoft OLAP son las siguientes:

- Tiene la capacidad de auxiliar en los diversos subprocesos de conversión.
- Esta herramienta, soporta el manejo de la cantidad de datos para el tamaño del DW del caso de estudio.
- Es una herramienta práctica, que no tiene mucha dificultad en su uso, lo que se debe a que trabaja en ambiente Windows.
- No es una herramienta muy costosa.

Para el proceso de creación de un DW, Microsoft OLAP proporciona básicamente 2 herramientas que son:

- Microsoft Data Transformation Services (DTS)
- Microsoft OLAP (OLAP Services)

DTS es una herramienta que permite realizar las tareas de extracción, transformación y carga de los datos, es decir, DTS puede ser usado para preparar datos para OLAP [23].

Por otra parte, la herramienta OLAP Services, permite construir los cubos de datos multidimensionales y almacenar información en ellos, por lo tanto, OLAP Services, ayuda a realizar el proceso de carga. Además, OLAP Services, proporciona un lenguaje de expresiones multidimensionales (MDX) para realizar consultas, dicho lenguaje se explica más a detalle en el siguiente capítulo.

6.2 CONVERSIÓN DE DATOS

A continuación, en la figura 6.1 se muestran las actividades que se realizaron en el proceso de la conversión de los datos, los cuales se llevaron a cabo utilizando tanto la herramienta DTS como OLAP Services.

Figura 6.1 Selección de la fuente de datos

En la sección 3.3 del capítulo 3, se explicó en forma resumida como se realizó cada una de las actividades mostradas en la figura 6.1. En las siguientes secciones, se describe el procedimiento que se llevó a cato en cada proceso.

6.2.1 Extracción de datos

Después de haberse realizado el modelado multidimensional y haberse creado un modelo lógico de la base de datos estándar, el siguiente proceso realizado fue la conversión de los datos. Para realizar esta tarea, lo primero que se realizó fue la extracción de los datos.

En el capítulo 2 se explicaron los tipos de extracción que permiten realizar las herramientas existentes; a saber, extracción a granel u de replicación basada en cambios. Para el caso de estudio se utilizó el segundo tipo, debido a que permite realizar transformaciones antes de copiar los datos a la base de datos destino, lo cual evitó que se copiaran datos innecesarios.

Para llevar a cato el proceso de extracción con el uso de la herramienta DTS, fue necesario especificar algunas características de las bases de datos fuente, como son: el tipo de manejador de base de datos o formato de archivo en el que se encuentran, en nuestro caso fueron Access u Excel; u también fue necesario especificar la ruta donde se localizan dichos datos, como se muestra en la figura 6.2. Los archivos origen se cargaron a la computadora donde se instaló MS OLAP.

Es importante recordar que los datos de las tablas de dimensiones, se obtuvieron de información proporcionada por los distintos subsistemas, mientras que los datos de las tablas de hechos se obtuvieron de los reportes recopilados por la dirección del IHEMSyS.

Figura 6.2 Selección de la fuente de datos

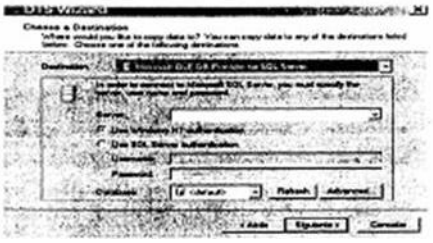


Figura 63 Selección de las bases de datos destino

Posteriormente, fue necesario especificar las mismas características para la tase de datos destino, como se muestra en la figura 63. Es importante mencionar que la herramienta DTS, permite obtener datos a partir de los siguientes tipos de manejadores de bases de datos o formatos de archivos: Access, Paradox, SQL, Oracle, DBASE, Excel y Archivos de texto. En nuestro caso, el formato de la base de datos destino fue SQL.

Una vez especificados los formatos y ubicación de las bases de datos fuente y destino, el siguiente paso consistió en realizar una conexión con la base de datos fuente, en nuestro caso para dicha conexión se utilizó la interfase ODBC (Open Data Base Connectivity). Sin embargo, la herramienta también cuenta con la interfase JDBC para cuando se requiere de utilizar ambiente de programación en el lenguaje Java.

Los datos requeridos son, el nombre del manejador de base de datos fuente y la ubicación de los mismos, como se muestra en las figuras 6.4 u 60.

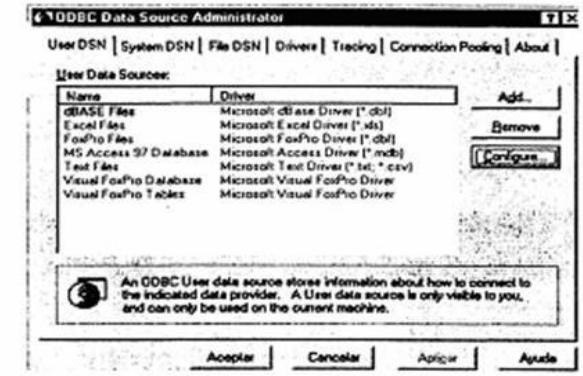


Figura 6.4 Especificación del ODBC

Figura 65 Conexión con la base de datos fuente

El siguiente paso consistió en seleccionar las tablas a extraer de las bases de datos fuente; esto, con la ayuda de un asistente con el que cuenta la herramienta DTS. En la siguiente figura se muestra un ejemplo de como se seleccionaron tablas de una base de datos fuente y se copiaron a la base de datos destino.

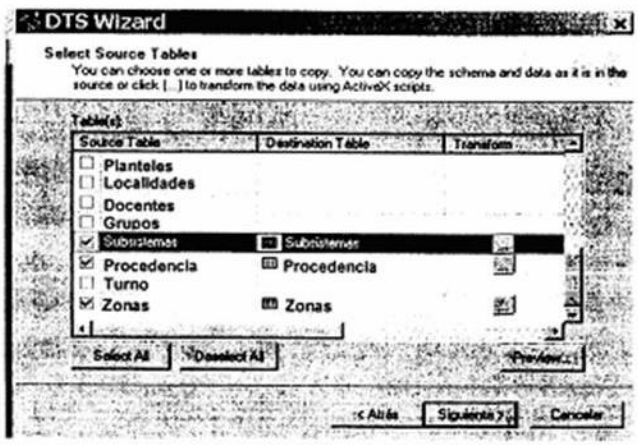


Figura 6.6 Selección de tablas de la base de datos fuente

Las tablas que se seleccionen, se van a copiaron a la tase de datos principal del DW; la cual va a proporcionar la información que utilizarán las tablas de dimensión u de hechos; las que se especificaron en el capítulo anterior. El siguiente paso, es la transformación de los datos, que se muestra a continuación.

6.2.2 Transformación de los datos

Una vez extraídas las diferentes tablas de las bases de datos, como se utilizó el método de extracción de replicación basada en cambios, la herramienta DTS permitió desplegar las tablas de la base de datos y realizar operaciones como agregar, eliminar u modificar campos. En la siguiente figura, se muestra un ejemplo del tipo de modificaciones que se pudieron realizar a los campos de una tabla.

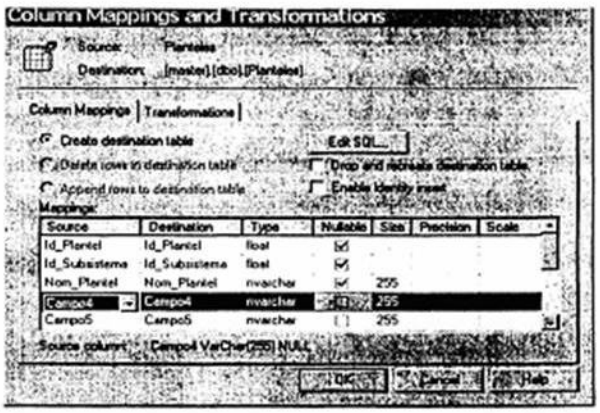


Figura 6.7 Operaciones para modificar datos

Como se muestra en la figura anterior, se realizaron operaciones como cambiar el nombre de los campos destino, el tipo de datos, el tamaño de los campos, etc. Sin embargo, otra forma que permitió la herramienta para realizar operaciones, fue con el uso del lenguaje SQL, como se explicó en el capítulo 2, en el proceso de transformación de datos.

A continuación se muestran las reglas o normas que se aplicaron en el proceso de limpieza:

- No permitir datos vacíos, en los casos donde se consideró no son necesarios.
- No se aceptan datos abreviados en los casos que no es permitido, ejemplo: Martínez -Mtz.
- Los datos de tipo numérico, deben ser únicamente números u no cadenas, ejemplo: el número de mes es correcto almacenarlo con número u no con la palabra Tres.
- Los datos de diferentes tablas que signifiquen lo mismo, deben redactarse de la misma forma. Ejemplo, el sexo no puede estar escrito en una tabla con la palabra Masculino u en otra con la letra M.
- El número de ano, en todas las fechas debe ocupar 4 dígitos.
- El formato de fecha para todos los casos es: día -mes - año

El sexo de una persona, se debe especificar con M o F (masculino, femenino) en todos los casos.

- El campo turno, en cualquier tabla se debe almacenar con las palabras matutino u vespertino, no es permitido usar M u V.
- Todos los nombres almacenados, deben tener el orden siguiente: apellido paterno, apellido materno, primer nombre u segundo nombre.
- Todos los nombres propios, deben iniciar con mayúsculas Cualquier dirección debe tener el siguiente formato: calle, número, colonia, municipio u entidad federativa.

6.23 Carga de datos

Una vez hecha la limpieza de los datos, se cuenta con la base de datos principal depurada, a partir de la cual se obtuvieron los datos para realizar el proceso de carga, por medio de la creación de los cubos de cada data mart Para esta tarea se utilizó la herramienta OLAP Services, la que a su vez cuenta con la herramienta OLAP Manager, la que permite crear tanto las bases de datos de los distintos data marts, como sus correspondientes cubos. Una base de datos OLAP es una estructura que es usada para almacenar un conjunto de cubos relacionados. Cuando los cubos de datos contienen información que se relaciona, estos deben ser almacenados en la misma base de datos [23].

En la siguiente figura se muestra la base de datos creada para el Departamento Académico del IHEMSyS:

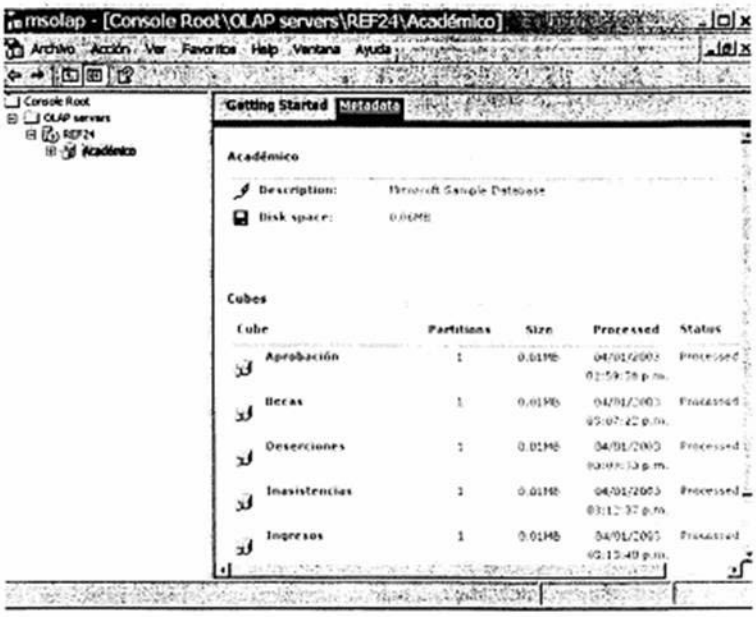


Figura 6.8 Creación de las Bases de datos del DW

Como se puede observar, las bases de datos creadas son tres: Académico, Recursos financieros u Recursos materiales. Es importante observar, que la tase de datos Académico está constituida de cinco cubos. Los cubos son la unidad fundamental para almacenar y recuperar datos en un sistema OLAP. El cubo es equivalente a una tabla en un sistema de base de datos relacional, los cubos son hechos sobre medidas y dimensiones.

Las dimensiones de un cubo son la perspectiva sobre la que los datos pueden ser vistos y analizados [23]. Esto es fácil de visualizar en dos dimensiones de un sistema relacional (en filas u columnas), pero un cubo OLAP puede tener hasta 64 dimensiones.

La herramienta OLAP Manager, proporciona asistentes para crear los cubos de datos a partir de la base de datos del DW, para lo que se tienen que seleccionar las tablas de hechos y dimensiones, como se muestra a continuación:

Figura 6.9 Extracción de tablas de hechos y dimensiones

En la figura anterior, se muestra una lista de tablas que pertenecen a la base de datos del DW, de la que se seleccionaron las tablas de hechos u de dimensión. Este proceso se repite para la creación de cada cubo.

Uno de los aspectos principales que se deben considerar en la creación de un cubo, es el tipo de almacenamiento, en la siguiente figura se muestran las tres formas de almacena* datos que permite OLAP Manager

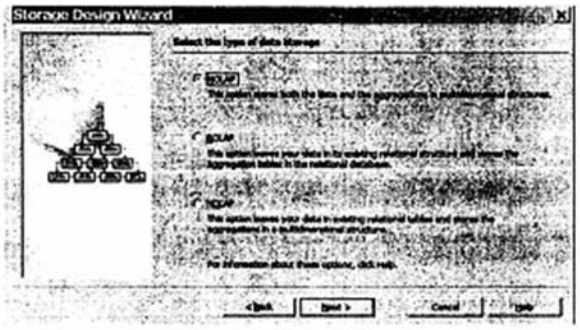


Figura 6.10 formas de almacenar dato de un cubo

Como se observa en la figura anterior, las formas de almacenar los datos de un cubo son: ROLAP, MOLAP y I1OLAP. La diferencia principal entre estos tres tipos de almacenamiento, está en como se estructuran los datos y las agregaciones de estos [26]. A continuación se describe cada uno.

MOLAP.- Usa un tipo de almacenamiento de datos que fue creado específicamente para análisis multidimensional. Los datos son copiados desde su origen y almacenados en una estructura MOLAP. La ventaja que proporciona este tipo de almacenamiento, es que el acceso a los datos es mas rápido, sin embargo, ocupa mas espacio.

ROLAP.- Usan la estructura de bases de datos relacional para almacenar las agregaciones en cubos como un conjunto de tablas. Este tipo de almacenamiento, requiere de menos espacio.

HOLAP.- Es un tipo de almacenamiento de datos que combina las características de MOLAP y ROLAP [27].

En la siguiente figura se muestran los cubos creados para el data mart del Departamento Académico.

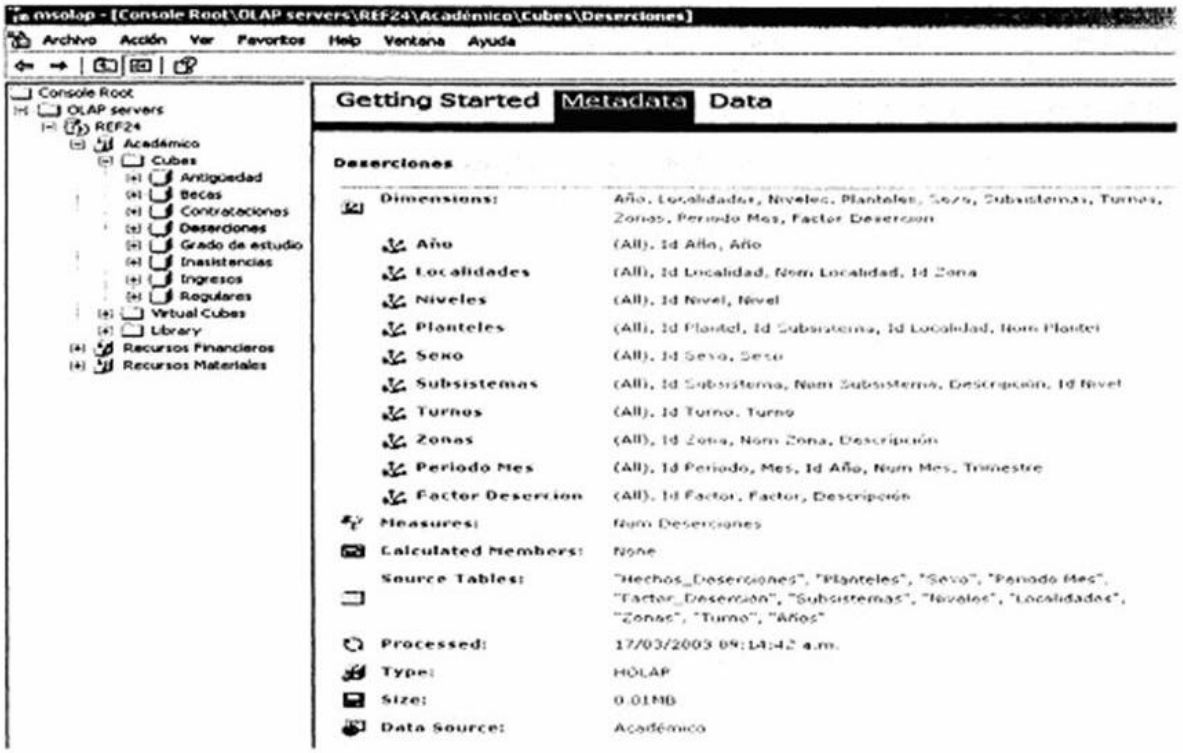


Figura 6.11 Cubos de base de datos académico

Es importante mencionar que algunos cubos como becas, deserción e inasistencias, contienen un solo indicador (por ejemplo, el cubo deserciones contiene el indicador número de deserciones). Sin embargo, otros contienen más de uno, un ejemplo es el cubo

aprobación que contiene los indicadores número de aprobados u número de reprobados, u otro ejemplo es el cubo ingresos que contiene los siguientes tres indicadores: número de aspirantes, número de ingresos u número de egresos. Un cubo puede contener varios indicadores, cuando estos trabajan con las mismas tablas de dimensiones.

En la figura 0.11, también se observan los metadatos correspondientes al cubo seleccionado, que en este caso es deserciones. Una de las grandes ventajas que proporciona la herramienta OLAP Services, es que genera automáticamente los metadatos. En la figura 0.11, se muestran metadatos sobre las dimensiones que contiene el cubo, así como sus características generales como son: tipo de estructura (MOLAP), tamaño, fuente de datos u tipo de acceso (Read Onlu).

La estructura de un cubo, está conformada por la tabla de hechos u un conjunto de tablas de dimensión, un cubo puede contener uno o más indicadores [23]. En la figura 6.12, se muestra la estructura del cubo de datos ***Deserciones***, el cual corresponde a su correspondiente esquema snowflake expuesto en el capítulo 5 en el diseño multidimensional.

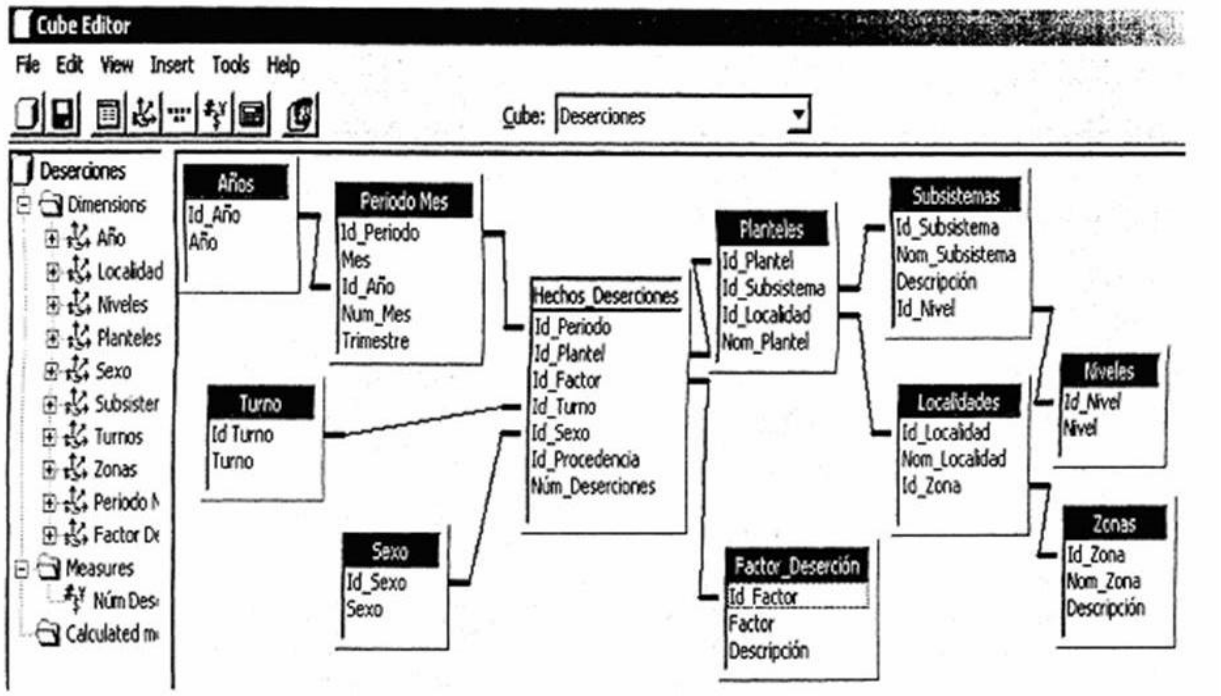
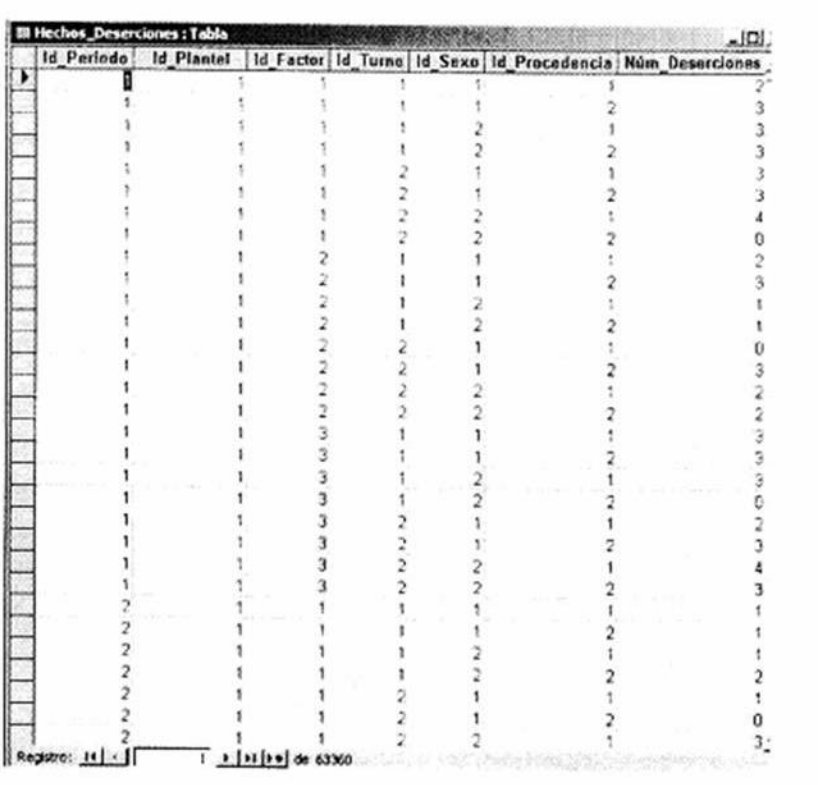


Figura 6.12 Estructura del cuto Deserciones

Como se muestra en la figura anterior, el cubo contiene una tabla de hechos que maneja el indicador deserciones y tablas de dimensión, las cuales están normalizadas. Con el fin de mostrar la estructura física de una tabla de hechos, en la siguiente figura se muestra como ejemplo la tabla correspondiente al indicador deserciones del cubo anterior.


Hechos_Deserciones : Tabla

Id_Periodo	Id_Plantel	Id_Factor	Id_Turno	Id_Sexo	Id_Procedencia	Núm_Deserciones
1	1	1	1	1	1	2
1	1	1	1	1	2	3
1	1	1	1	2	1	3
1	1	1	1	2	2	3
1	1	1	2	1	1	3
1	1	1	2	1	2	3
1	1	1	2	2	1	4
1	1	1	2	2	2	0
1	1	2	1	1	1	2
1	1	2	1	1	2	3
1	1	2	1	2	1	1
1	1	2	1	2	2	1
1	1	2	2	1	1	0
1	1	2	2	1	2	3
1	1	2	2	2	1	2
1	1	2	2	2	2	2
1	1	3	1	1	1	3
1	1	3	1	1	2	3
1	1	3	1	2	1	3
1	1	3	1	2	2	0
1	1	3	2	1	1	2
1	1	3	2	1	2	3
1	1	3	2	2	1	4
1	1	3	2	2	2	3
2	1	1	1	1	1	1
2	1	1	1	1	2	1
2	1	1	1	2	1	1
2	1	1	1	2	2	2
2	1	1	2	1	1	1
2	1	1	2	1	2	0
2	1	1	2	2	1	3

Registro: 1 de 63360

Figura 613 Tabla de hechos del cubo Deserciones

Como se puede observar, la tabla contiene un campo para almacenar los datos del indicador y cinco llaves que permiten relacionar la tabla de hechos con las cinco dimensiones con que se relaciona el indicador. La tabla almacena en su totalidad 63,360 registros, sólo que en la figura únicamente se muestra el comienzo de la tabla.

Capitulo 7

Explotación de los datos con Microsoft OLAP

7.1 INTRODUCCIÓN

La explotación de los datos a partir del depósito del DW es la base donde se verifica si el diseño ha sido realizado en forma adecuada. A continuación se describe la herramienta seleccionada para llevar a cabo esta tarea.

Se seleccionó la herramienta Microsoft OLAP para la explotación de los datos, debido a las siguientes razones:

- Permite proporcionar al usuario, vistas de datos con gran velocidad de respuesta, normalmente en menos de cinco segundos en una consulta.
- Contiene herramientas analíticas, tanto para el desarrollado como para el usuario final.
- Permite implementar los requerimientos de seguridad necesarios para compartir datos confidenciales entre los grupos de usuarios.

Las herramientas de Microsoft OLAP para la explotación de datos, proporcionan a las organizaciones los medios para acceder, ver y analizar los datos con alto desempeño y flexibilidad. Lo primero y mas importante es que Microsoft OLAP presenta los datos a los usuarios a través de un modelo de datos intuitivo y natural, por lo cual los usuarios finales pueden ver y entender más fácilmente la información del almacén de datos.

En segundo lugar, Microsoft OLAP acelera la entrega de información a los usuarios finales porque puede preparar algunos valores computados en los datos por adelantado (por ejemplo, en los casos más comunes se pueden tener datos ya agregados almacenados para cuando el usuario lo solicite) en vez de hacerlo al momento de ejecutarse una consulta. La combinación de navegación fácil y rápida, le permite a los usuarios ver y analizar información más rápida y eficientemente de lo que es posible con tecnologías de bases de datos relaciónales. Por lo tanto, esto da como resultado que el usuario se pase más tiempo analizando los datos y menos tiempo analizando las bases de datos.

Como se mencionó en el capítulo anterior, Microsoft OLAP además de contar con herramientas para la Transformación de datos (DTS), para el proceso de explotación de los datos cuenta con un lenguaje de expresiones multidimensionales (MDX), el que es usado para realizar consultar multidimensionales.

7.2 EXPLOTACIÓN DE DATOS CON MICROSOFT OLAP

Microsoft OLAP implemento un lenguaje para consultar cubos, este lenguaje llamado de expresiones multidimensionales (MDX). MDX es un lenguaje que contiene varias funciones especializadas con las cuales se pueden realizar operaciones drill down, roll up, combinación de múltiples dimensiones en un eje, unir miembros de una dimensión en una vista combinada, sumarizar datos, encontrar valores a alto o bajo nivel, ejecutar cálculos, etc. Algunas de las operaciones mencionadas, como drill down y roll up, se vieron en el capítulo 2.

Comparación entre MDX y SQL

Los servicios de Microsoft OLAP, también proveen la capacidad de ejecutar consultas en SQL (Lenguaje de consulta estructurado). La diferencia entre ambos lenguajes, está en que cuando se ejecuta una consulta construida en MDX, los servicios OLAP actúan como proveedores de este tipo de datos retornando un conjunto de datos multidimensionales. Sin embargo, cuando se ejecutan las consultas SQL, los servicios OLAP retornan un conjunto de filas como un proveedor de datos tabular [41].

Para realizar una consulta en SQL dentro de los servicios OLAP, sólo se usa parte del lenguaje que comúnmente se utiliza para consultar bases de datos relaciónales. La sintaxis básica de una sentencia SQL para servicios OLAP es la misma que una sentencia estándar SQL, a continuación se muestra la sintaxis correcta en que se puede utilizar cada una de sus cláusulas:

La cláusula SELECT, se utiliza para especificar en una lista separada con comas de los niveles u miembros de la dimensión de medidas, como se muestra en el siguiente ejemplo:

```
SELECT [Sexo]. [Sexo], [Sistema]. [Tipo_Sistema], [Medidas]. [Núm_Deserciones]
```

Es importante mencionar que en esta sentencia, solo son permitidos los miembros niveles y medidas. Esto se debe a que los niveles de los cubos son definidos con columnas desde la tabla de dimensión en el esquema estrella, de manera similar, los miembros de medidas son definidos con columnas desde la tabla de hechos, y por lo tanto, ellos también pueden ser incluidos en la lista SELECT. Sin embargo, los

miembros de las otras dimensiones no pueden ser incluidos en la lista de la cláusula select.

La cláusula FROM simplemente contiene el nombre del cuto. Cuando SQL es ejecutado en servicios OLAP, el cubo funciona como una tabla simple, desde la cual todos los datos son recuperados [23].

La cláusula WHERE, solo incluye niveles y miembros desde las dimensiones de medidas. De la misma manera, la cláusula GROUP BY, incluye solo niveles y miembros desde la dimensión de medidas.

Finalmente, es posible incluir la palabra DISTINCT en la cláusula SELECT como sigue: SELECT DISTINCT.

El uso de SQL en Servicios OLAP, tiene las siguientes restricciones con respecto al uso de la palabra reservada DISTINCT y la cláusula GROUP BY.

- Ni la palabra reservada DISTINCT ni la cláusula GROUP BY, pueden ser usadas si la lista SELECT contiene miembros.
- DISTINCT puede también causar algunos problemas con los niveles en los cuales pueden ser retornadas filas duplicadas. Este problema es probable que crezca cuando el nivel cercano o el nivel raíz contiene más de un miembro.
- Las cláusulas DISTINCT y GROUP BY, pueden retornar filas duplicadas si el servidor contiene más de un segmento.

Sin embargo, como se mencionó anteriormente, los servicios OLAP solo presentan un subconjunto de toda la sintaxis de los comandos SQL Las cláusulas FROM y GROUP BY son simples e implementadas en el nivel más básico, debido a que las uniones no son necesarias, la sintaxis de esas dos cláusulas no es compleja. por lo tanto, esto hace posible usar las cláusulas SELECT y WHERE hacia funcionalidades más avanzadas.

Mientras la lista SELECT primero consta de una lista de nombres de columnas, ésta puede también incluir la cláusula DISTINCT como se describe anteriormente, éste puede incluir identificadores de los nombres de las columnas y también puede incluir funciones agregadas (SUM, MIN, MAX, COUNT).

Comparación entre SQL y MDX

Para tener una idea más clara de la diferencia entre SQL y MDX en los servicios OLAP, es necesario comparar como estos lenguajes son aplicados a situaciones reales. Esto se realiza mediante los siguientes ejemplos:

Ejemplo de Consulta a alto nivel

El objetivo del siguiente ejemplo es mostrar la diferencia entre ambos lenguajes con una consulta que obtiene datos a bajo nivel de detalle. Es importante recordar que cuando se usa SQL como el lenguaje de consulta, no se hace empleo de dos de las mejores capacidades de servicios OLAP, que son: agregación y miembros por default Sin embargo, una consulta simple MDX usa ambas capacidades.

El primer ejemplo compara dos consultas similares. La primera es escrita en SQL y la segunda en MDX. No obstante, el hecho de que estas dos consultas sean similares en la sintaxis y contenido» la ejecución y los resultados de las consultas son diferentes.

La sentencia SQL es:

```
SELECT
        [Sexo]. [Sexo], [Sistema], [Sistema],
        [Medidas]. [Número_Deserciones]
FROM    [Deserciones]
```

El grupo de registros creados con la sentencia tiene 6,905 registros, una porción de la cual es mostrada en la siguiente figura,

Sexo.Sexo	Sistema.Plantel	Medidas.Número_deserciones
F	COBAEH	120
M	COBAEH	123
F	COBAEH	104
M	COBAEH	113
M	COBAEH	99
F	COBAEH	190
.	.	.
M	CECYTE	111
F	CECYTE	210
M	CECYTE	109
F	CECYTE	114
.	.	.
F	CONALEP	100
M	CONALEP	99
M	CONALEP	115
M	CONALEP	122
.	.	.

Figura 71 Resultado de la consulta 1 en SQL

Para mejorar los resultados con el uso de SQL, se puede realizar la agregación de datos dentro de la sentencia SQL Para esto, es necesario utilizar tanto la sentencia GROUP BY, como la función SUM() dentro de la cláusula SELECT. La siguiente sentencia SQL suma el dato Número de deserciones usando la columna sexo y la columna tipo de subsistema.

```
SELECT
        [Sistema] [Tipo_Sistema],
        [Sexo]. [Sexo]
        Sum[Número  Deserciones]
FROM [Deserciones j
GROUP BY
                [Sistema], [Tipo_Sistema],
                [Sexo]. [Tipo_Sexo]
```

El conmuto de registro resultantes desde esta consulta es mostrado en la siguiente figura:

COBAEH	F	2947
COBAEH	M	3100
CECYTE	F	1791
CECYTE	M	1900
CONALEP	M	497
CONALEP	F	450
COBAEH	M	2646
COBAEH	F	1900
CECYTE	M	1792
CECYTE	F	1900
CONALEP	M	454
CONALEP	F	526

Figura 7,2 Resultado de consulta 1 en SQL Mejorada

Como se puede observar, el conjunto de registros resultantes retorna 12 filas y estas no están agregadas completamente, lo cual no es lo óptimo. La sentencia equivalente en MDX a la consulta precedente, es la siguiente: SELECT

```
                {[Sexo]. [Sexo]. Members} ON COLUMNS,
                {[Subsistemas]. [Nom _subsistema]. Members} ON ROWS
        FROM [Deserciones]
        WHERE ([Medidas]. [Núm_deserciones])
```

El conjunto de datos generado con la sentencia MDX tiene dos columnas, tres filas y seis datos, como se muestra en la figura 73. En contraste con el grupo de registros obtenidos con SQL, los datos retornados con el uso de la sentencia MDX son más útiles por que están agregados.

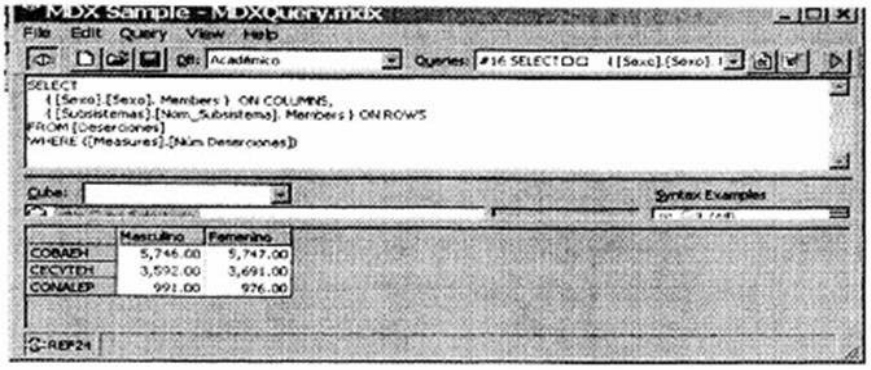


Figura 73 Resultado de consulta 1 en MDX

Cu la figura 73, el número de deserciones de cada plantel, están agregados por plantel y por sexo, En este ejemplo, se muestra que en una consulta SQL es necesario definir explícitamente la agregación, En cambio, las consultas MDX usan una agregación implícita cuando son ejecutadas.

Hasta el momento se han expuesto las similaridades y diferencias entre MDX y SQL. Lo importante es notar que, típicamente un cuto OLAP debe configurarse de tal manera que utilice sentencias MDX con alto nivel de agregación por default. En contraste SQL, retorna datos granulares con poca agregación.

73 CONSULTAS EJECUTADAS

Antes de utilizar la aplicación MDX, fue necesario realizar una conexión con el servidor OLAP. Es decir, conectar con el procesador donde está instalado MS OLAP.

A continuación, se muestra la ventana principal de la aplicación MDX, la cual se subdivide en tres partes.

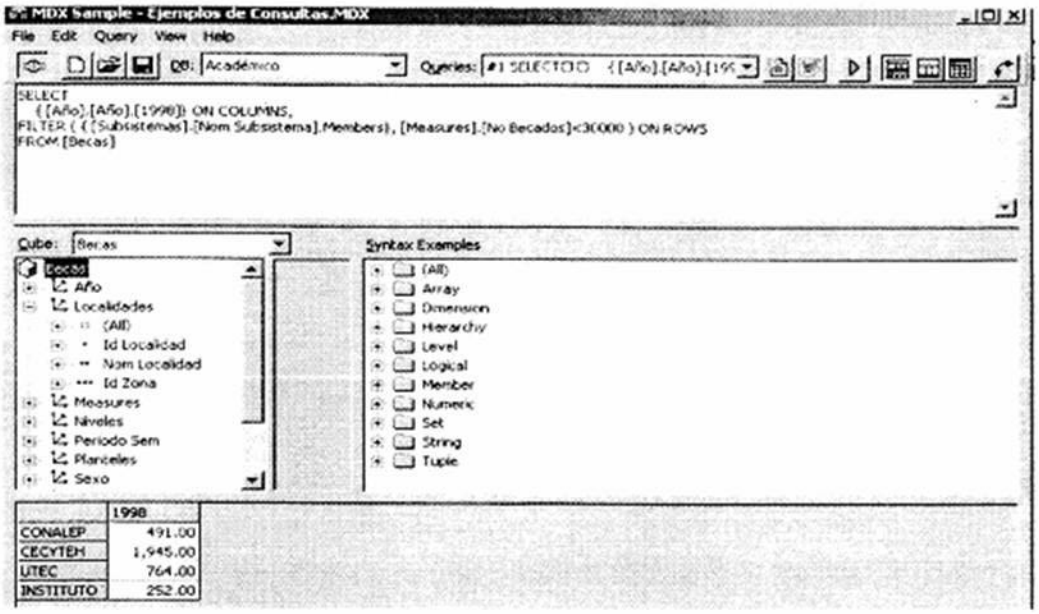


Figura 74 Pantalla principal del la aplicación MDX

La primer parte corresponde al área donde el usuario puede capturar las instrucciones de las consultas.

La segunda parte se subdivide en dos áreas:

- En la parte izquierda, el usuario puede seleccionar tanto el nombre del cubo con el que desea trabajar como la lista ele dimensiones y campos de las distintas dimensiones correspondientes al cuto seleccionado. Tanto los nombres de las dimensiones como de los campos se pueden agregar a una

consulta al seleccionarlo u dar doble clic.

- En la parte derecha, la herramienta le proporciona al usuario la sintaxis de las instrucciones y funciones básicas utilizadas en una consulta, teniendo que introducir únicamente los parámetros requeridos.

La tercera parte es el área donde se le muestran al usuario los resultados de las consultas ejecutadas.

Es importante recordar que a diferencia de SQL, una consulta en MDX se compone de pocas instrucciones. Además, la herramienta proporciona al usuario los nombres de las dimensiones, campos u tablas de hechos de cubo tratado para facilitarle el trabajo.

MDX no es una herramienta que pueda utilizarse sin saber nada del lenguaje de consultas. Sin embargo, los conocimientos que exige para su utilización son mínimos, por lo cual se considera que usuarios con una poca capacitación la pueden utilizar.

Con el fin de demostrar como se explotan los datos con MDX, a continuación se muestra una serie de consultas aplicadas sobre los cubos creados en el Departamento Académico, mostrando sus correspondientes resultados al ser ejecutadas.

Como se mencionó en el capítulo 2, las operaciones básicas OLAP son: Dice for, Roll up, Slice u Drill down. Dichas operaciones, se clasifican básicamente de acuerdo al nivel de detalle con que se quieren obtener los datos, a partir de los cubos existentes.

A continuación se muestran las consultas, clasificándolas de acuerdo al tipo de operación que realizan.

Operación Dice For

Este tipo de operación, consiste en realizar consultas donde se pueden especificar una o varias condiciones, es similar a las consultas tradicionales en SQL. Este tipo de operación es la menos requerida en el ambiente OLAP, sin embargo existen situaciones en las cuales es necesaria su utilización.

Ejemplo 1.
En este ejemplo se obtiene el número de alumnos becados bajo las siguientes condiciones:

- Que la información se obtenga sobre los planteles de Actopan" y "Zempoala'.
- Que los anos a consultar sean 1998 u 2002.

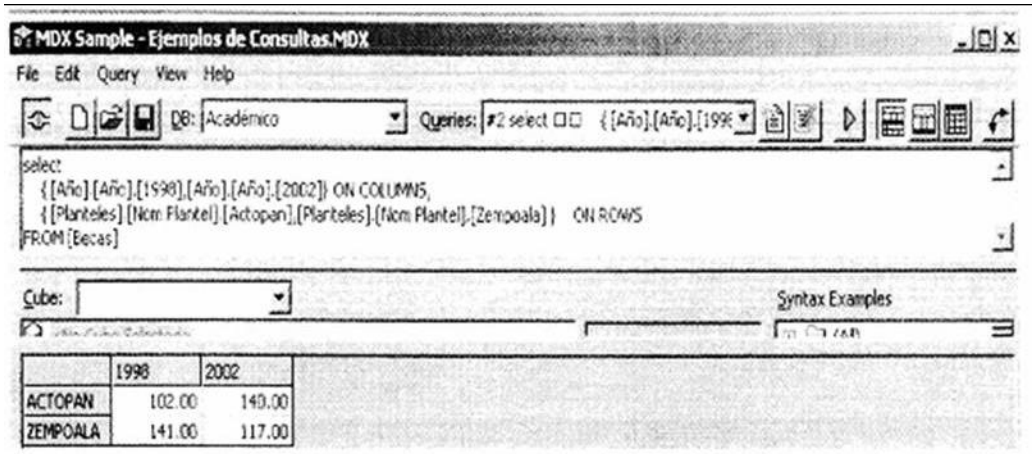


Figura 7.5 Resultado del primer ejemplo de operación *Dice lar*

La operación anterior, muestra como este tipo ele operación permite realizar cónsul t datos no secuénciales sobre determinadla dimensión. En este caso, la especificaciones dos años que no son continuos (1998 y 2000) y de dos planteles, que no se almacena* después de otro. Sí se realizara una consulta para obtener el número de alumnos becados entre los periodos ele 1998 a 2000, la operación a utilizar sería de tipo Slice, la cual se explica mas adelante.

Ejemplo 2

En este ejemplo se utiliza la operación Dice for para obtener el numero de becas bajo las siguientes condiciones;

Plantel "Actopan ".
Periodo sea el año 1999,
Semestre sea del periodo Enero-Junio (Verano).
Turno sea Matutino,
Sexo sea Masculino

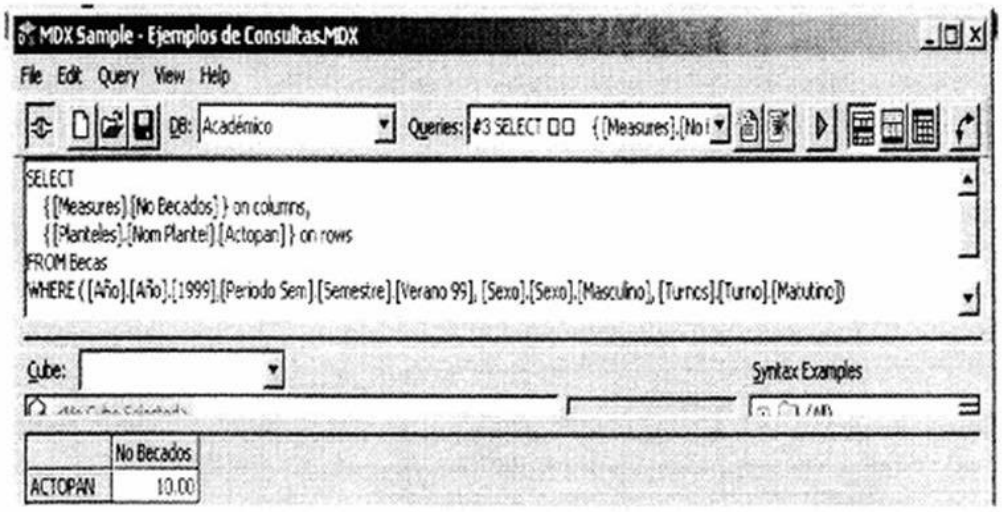


Figura 7.6 Resultado del segundo ejemplo de operaciones Dice For

Este ejemplo muestra una consulta que contiene cinco condiciones, utilizando cuatro dimensiones (plantel periodo, turno y sexo), pues ano y semestre pertenecen a la dimensión periodo.

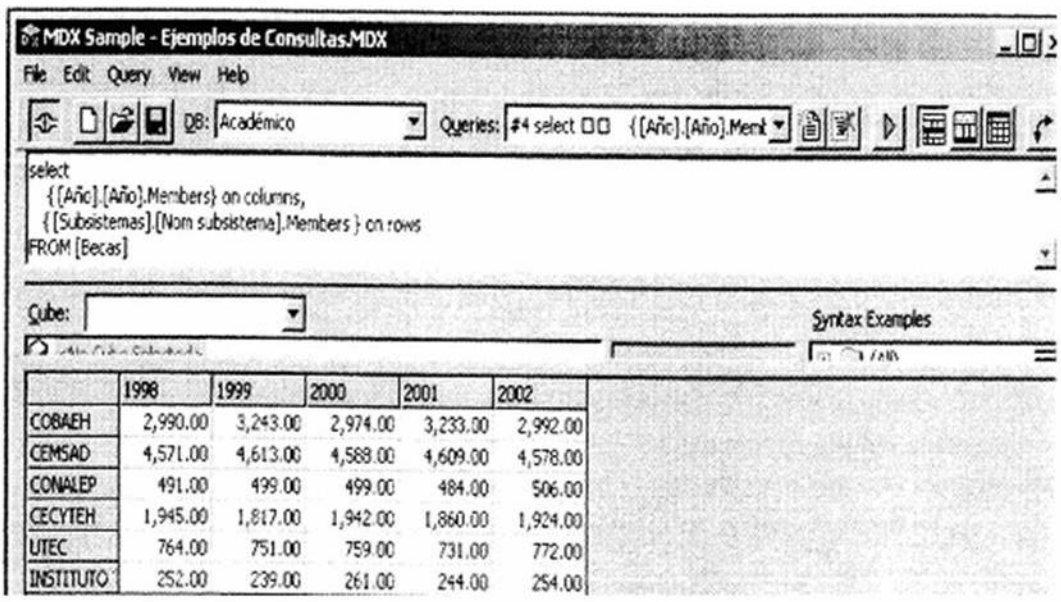
Operación RolI Up

Este tipo de operación, consiste en obtener información en forma resumida. Al aplicarla, se efectúan agregaciones sobre los datos existentes, que consiste en realizar las sumas necesarias de los datos involucrados.

Esta operación se utiliza cuando el usuario realiza consultas de indicadores sobre dimensiones de mayor jerarquía o nivel Algunos ejemplos de este tipo de dimensiones en el datamart del Departamento Académico son; subsistemas, nivel educativo, periodo de tiempo en años, etc. Es importante recordar que las operaciones OLAP se utilizan por personal directivo para la toma de decisiones donde se requieren mayormente resultados finales, por lo tanto, la operación Roll Upes la más utilizada en este ámbito.

Ejemplo 1

Aquí se utiliza la operación Roll Up, para obtener el número de deserciones en los años de 1998 a 2002 a nivel de subsistema.



	1998	1999	2000	2001	2002
COBAEH	2,990.00	3,243.00	2,974.00	3,233.00	2,992.00
CEMSAD	4,571.00	4,613.00	4,588.00	4,609.00	4,578.00
CONALEP	491.00	499.00	499.00	484.00	506.00
CECYTEH	1,945.00	1,817.00	1,942.00	1,860.00	1,924.00
UTEC	764.00	751.00	759.00	731.00	772.00
INSTITUTO	252.00	239.00	261.00	244.00	254.00

Figura 7.7 Resultado del primer ejemplo de operación *Roll Up*

Para mostrar los resultados de esta consulta, el lenguaje realiza las agregaciones o sumas necesarias. Por ejemplo, debido a que el indicador deserciones se maneja mensualmente, para mostrar el resultado del subsistema COBAEI1 en 1998, se tuvieron que sumar por

plantel los números ele deserciones de los diferentes meses de ese año y, finalmente, sumar los resultados de los diferentes planteles que conforman el COBAEH En MDX no se tienen que introducir funciones para sumar los datos, como ocurre con SQL

Ejemplo 2

En este ejemplo se obtiene el número de deserciones entre el ciño 2000 y 2002 por nivel educativo (medio superior y superior).

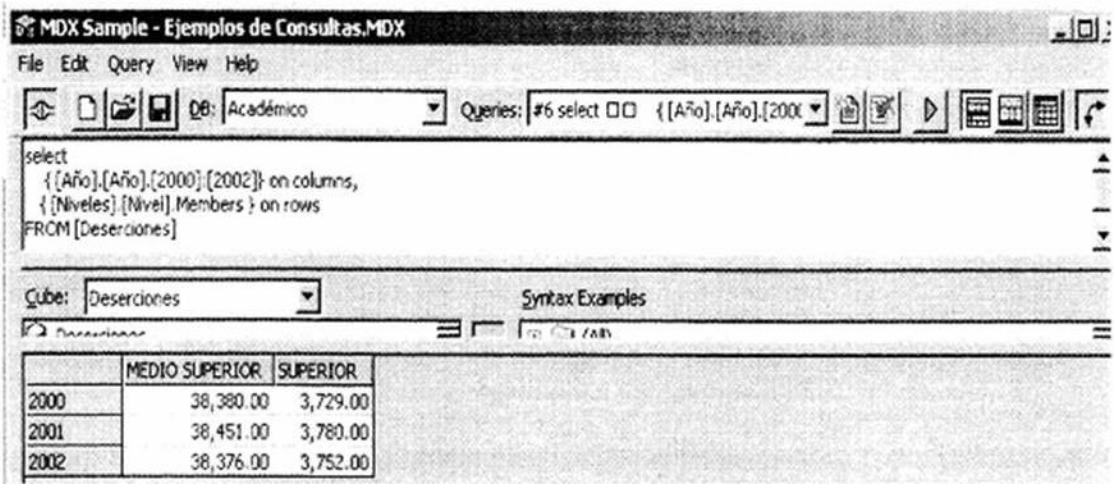


Figura 78 Resultado del segundo ejemplo de operación *Roll up*

En la figura 7,8, se muestra el resultado de las deserciones al más alto nivel, que es el nivel educativo (Medio Superior y Superior), Por lo tanto, el nivel de sumarización de datos es un nivel mayor que el del ejemplo anterior, esto debido a que el nivel educativo se clasifica por subsistemas»

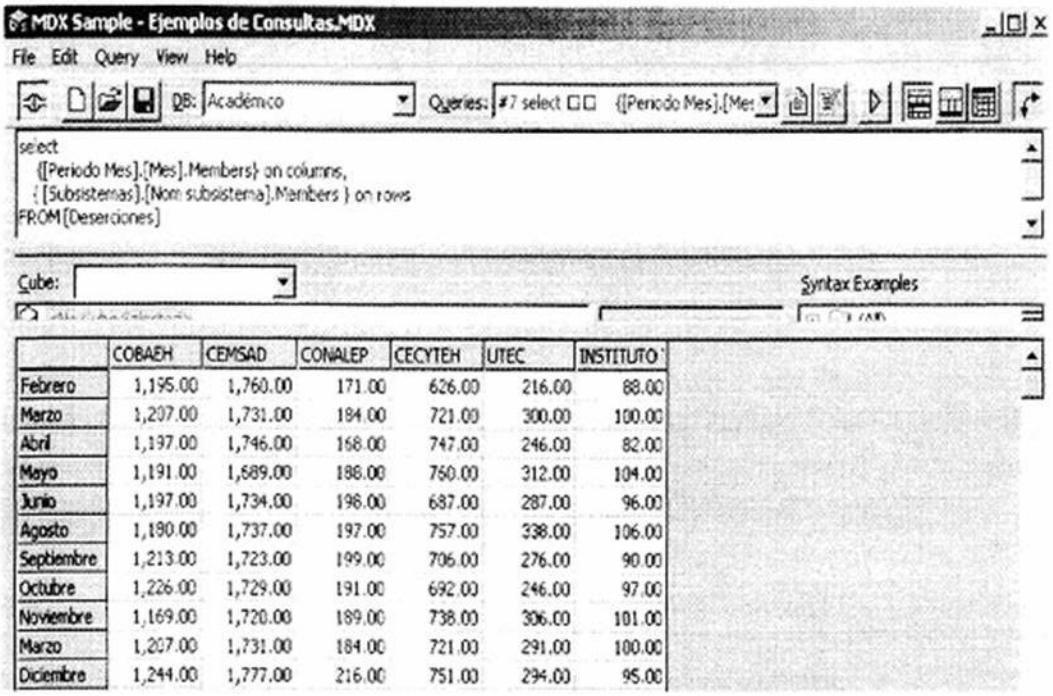
En el caso del cubo deserciones el resultado mas resumido, sería el de la suma de los datos de los dos niveles educativos en los distintos periodos. Esto daría como resultado el total de deserciones existentes en todo el IHEMSyS durante los períodos analizados.

Operación Drill Down

Este tipo de operación consiste en obtener información a alto nivel de detalle, es la operación contraria a Roll Up Se utiliza cuando se requiere consultar los indicadores respecto a dimensiones de menor jerarquía o nivel, Por ejemplo, planteles, periodo en meses o semestres, etc. En la mayoría de consultas de éste tipo de operación, los datos se muestran como están almacenados, sin necesidad de hacer sumas.

Ejemplo 1

En este ejemplo, se obtiene el número de deserciones por subsistema y por mes.



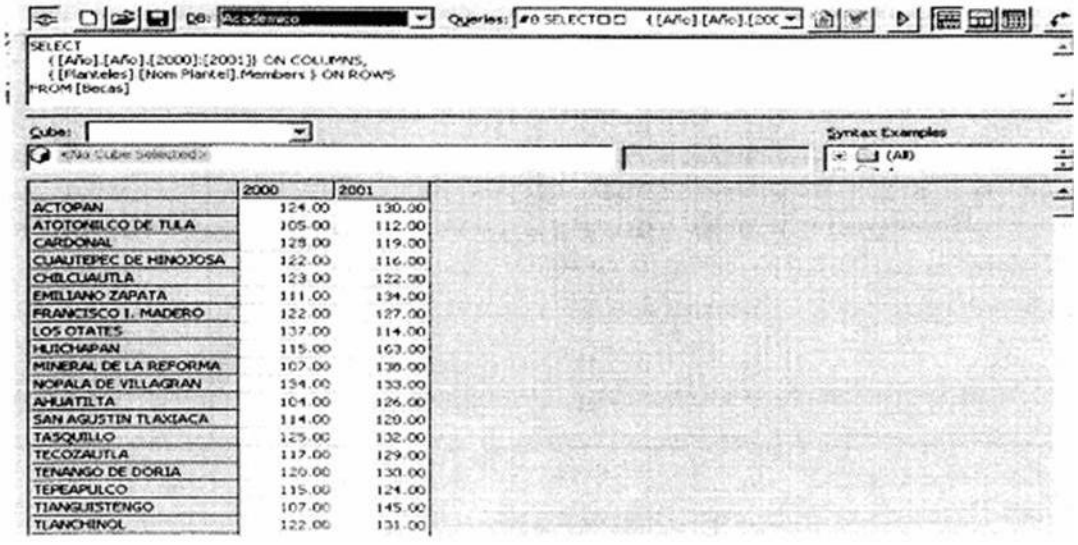
	COBAEH	CEMSAD	CONALEP	CECYTEH	UTEC	INSTITUTO
Febrero	1,195.00	1,760.00	171.00	626.00	216.00	88.00
Marzo	1,207.00	1,731.00	184.00	721.00	300.00	100.00
Abril	1,197.00	1,746.00	168.00	747.00	246.00	82.00
Mayo	1,191.00	1,689.00	186.00	760.00	312.00	104.00
Junio	1,197.00	1,734.00	196.00	687.00	287.00	96.00
Agosto	1,180.00	1,737.00	197.00	757.00	338.00	106.00
Septiembre	1,213.00	1,723.00	199.00	706.00	276.00	90.00
Octubre	1,226.00	1,729.00	191.00	692.00	246.00	97.00
Noviembre	1,169.00	1,720.00	189.00	738.00	306.00	101.00
Marzo	1,207.00	1,731.00	184.00	721.00	291.00	100.00
Diciembre	1,244.00	1,777.00	216.00	751.00	294.00	95.00

Figura 7.9 Resultado del *primer* ejemplo de operación *Drill down*

Nótese que los datos son mostrados a nivel de detalle en cuanto a la dimensión periodo En el resultado de la consulta, sólo se muestran los datos de los meses correspondientes c un solo ano.

Ejemplo 2

En este ejemplo, se obtiene el número de alumnos becados, por plantel* en los años 2000 y 200l



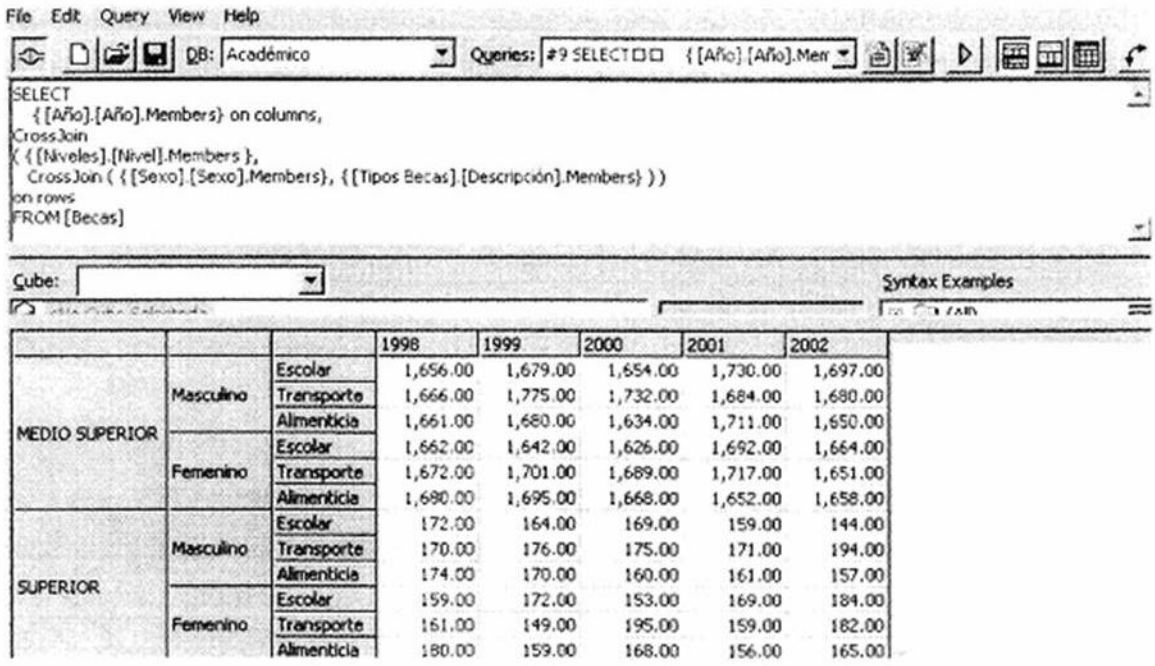
	2000	2001
ACTOPAN	124.00	130.00
ATOTONILCO DE TULA	105.00	112.00
CARDONAL	128.00	119.00
CUAUTEPEC DE HINOJOSA	122.00	116.00
CHILCUAUTLA	123.00	122.00
EMILIANO ZAPATA	111.00	134.00
FRANCISCO I. MADERO	122.00	127.00
LOS OTATES	137.00	114.00
HUICHAPAN	115.00	163.00
MINERAL DE LA REFORMA	107.00	136.00
NOPALA DE VILLAGRAN	154.00	133.00
AHUATILTA	104.00	126.00
SAN AGUSTIN TLAXIACA	114.00	129.00
TASQUILLO	125.00	132.00
TECOZAUTLA	117.00	129.00
TENANGO DE DORIA	120.00	130.00
TEPEAPULCO	115.00	124.00
TIANGUISTENGO	107.00	145.00
TLANCHINOL	122.00	131.00

Figura 7.10 Resultado del segundo ejemplo de operación *Drill down*

De manera similar al ejemplo anterior, aquí también se muestran los ciatos en forma detallada, pero en este caso respecto a la dimensión planteles En los resultados solo se muestra parte ele los datos.

Ejemplo 3

Aquí se utiliza la operación *Drill down*, para obtener el numero de alumnos becados, por nivel educativo, sexo y tipo de beca.



			1998	1999	2000	2001	2002
MEDIO SUPERIOR	Masculino	Escolar	1,656.00	1,679.00	1,654.00	1,730.00	1,697.00
		Transporte	1,666.00	1,775.00	1,732.00	1,684.00	1,680.00
		Alimenticia	1,661.00	1,680.00	1,634.00	1,711.00	1,650.00
	Femenino	Escolar	1,662.00	1,642.00	1,626.00	1,692.00	1,664.00
		Transporte	1,672.00	1,701.00	1,689.00	1,717.00	1,651.00
		Alimenticia	1,680.00	1,695.00	1,668.00	1,652.00	1,658.00
SUPERIOR	Masculino	Escolar	172.00	164.00	169.00	159.00	144.00
		Transporte	170.00	176.00	175.00	171.00	194.00
		Alimenticia	174.00	170.00	160.00	161.00	157.00
	Femenino	Escolar	159.00	172.00	153.00	169.00	184.00
		Transporte	161.00	149.00	195.00	159.00	182.00
		Alimenticia	180.00	159.00	168.00	156.00	165.00

Figura 711 Resultado del tercer ejemplo de operación *Drill down*

Como se observa en el resultado anterior, los datos son mostrados a nivel de detalle de tipo de beca (escolar, transporte o alimenticia). Sin embargo, las consultas se pueden realizar al nivel de detalle que se requiera, pudiendo involucrar todas las dimensiones que intervienen. Para esto, es importante considerar une en la consulta se deben especificar las dimensiones en orden de mayor a menor jerarquía. Por ejemplo, en la consulta anterior se debe especificar primero el nivel educativo, y después el tipo de beca.

Ejemplo 4

El lenguaje MDX, cuenta con la función DrillDownLevel ( ) que permite realizar operaciones Drill Down, mostrando los datos en forma resumida. A continuación, se muestra un ejemplo, donde se obtiene el primero de alumnos becados en el ano 2000, en el nivel medio superior y del sexo masculino.

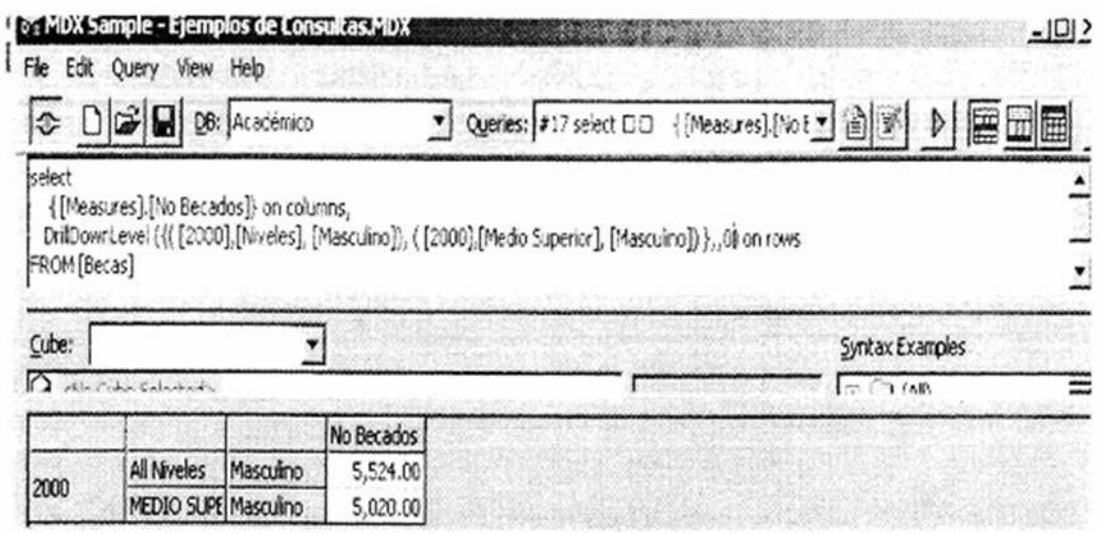


Figura 712 Resultado del cuarto ejemplo de operación *Drill down*

Como se muestra en este ejemplo, la función *DrillDownLevel* requiere como parámetro: las dimensiones sobre las cuales se desea obtener el indicador, obteniendo con resultado la sumatoria de las dimensiones involucradas.

Operaciones Slice

Este tipo de operación consiste en extraer sólo una parte del cuto multidimensional d datos. No requiere de alguna función en especial, sino que se logra al especificar que le datos se muestren sólo en determinados valores de una dimensión» Por ejemplo, Mostar datos de sólo tres subsistemas, de un total de 7,

Ejemplo 1

Aquí se obtiene el número de alumnos regulares por año, en el subsistema COBAEH

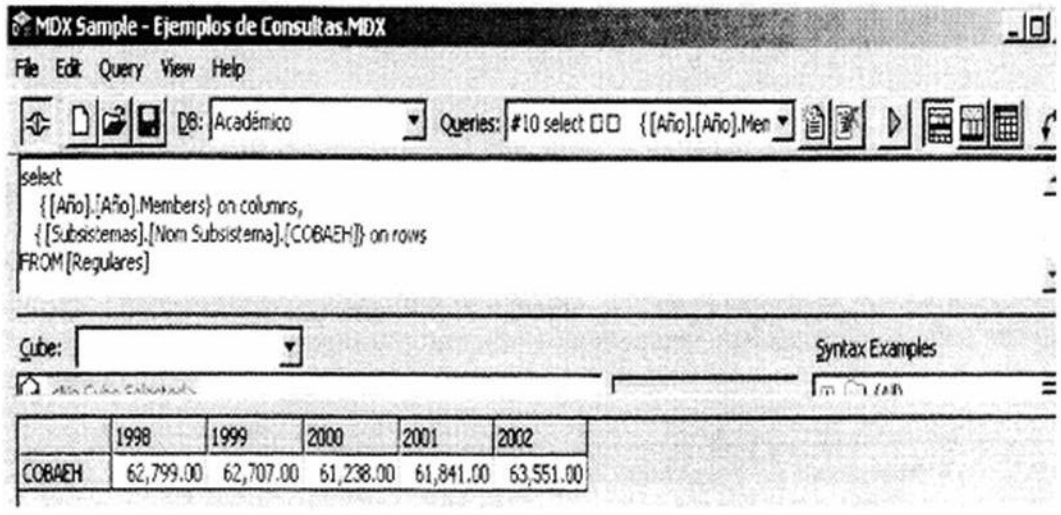
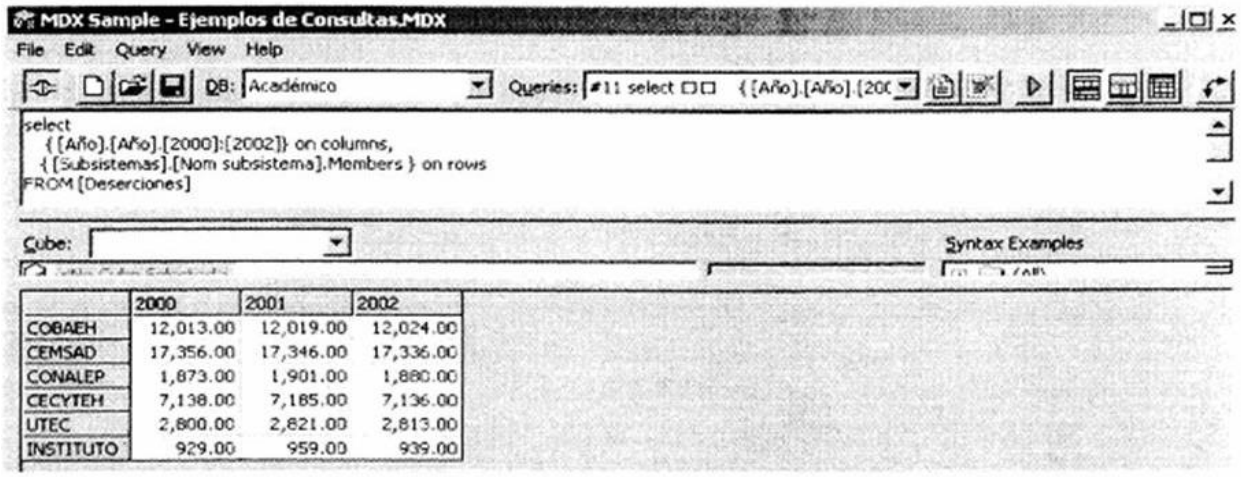


Figura 713 Resultado del primer ejemplo de operación Slice

Como se comentó anteriormente, con ésta operación se obtiene sólo parte de la información respecto a una dimensión (en este caso subsistemas), extrayendo sólo los datos del COBAEH Por ello, es que algunos autores consideran que esta operación es equivalente a extraer una rebanada de un cubo multidimensional de datos; y las rebanadas restantes serían los demás subsistemas.

Ejemplo 2

Cu este ejemplo se obtiene el número de deserciones en los distintos subsistemas en los años 2000,2001 y 2002



	2000	2001	2002
COBAEH	12,013.00	12,019.00	12,024.00
CEMSAD	17,356.00	17,346.00	17,336.00
CONALEP	1,873.00	1,901.00	1,880.00
CECYTEH	7,138.00	7,185.00	7,136.00
UTEC	2,800.00	2,821.00	2,813.00
INSTITUTO	929.00	959.00	939.00

Figura 714 Resultado del segundo ejemplo de operación Slice

Aquí también se extrae sólo una sección de los datos, debido a quien únicamente se obtiene la información correspondiente a un periodo de tres años.

Fundones adicionales

Es importante mencionar que la herramienta MDX, además de permitir realizar las operaciones anteriores, contiene una serie de funciones adicionales que ayudan a realizar algunas operaciones complementarías a las expuestas anteriormente. Algunos ejemplos de este tipo de funciones son para ordenar los datos en forma ascendente o descendente, para filtrar datos, operadores para representar condiciones, etc.

A continuación, se muestra un ejemplo de una consulta, en la cual se hace uso de la función filterí), para mostrar únicamente los datos que cumplen con cierta restricción.

Ejemplo1

Este ejemplo consiste en realizar una consulta para obtener el número de alumnos becados en el año 1998, de todos los subsistemas que hayan tenido menos de 1000 becados.

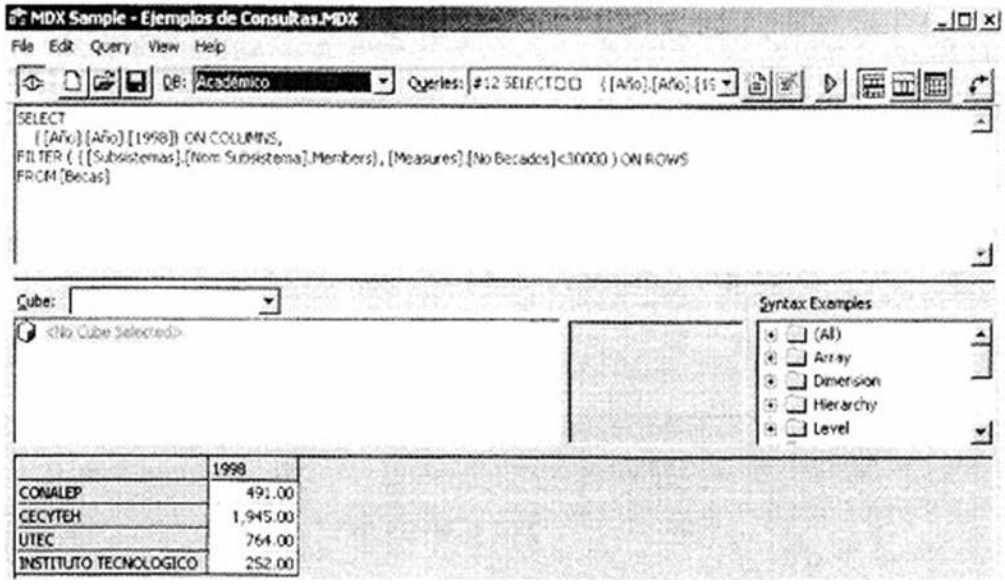


Figura 7.15 Resultado del ejemplo utilizando *la* función Filter ()

En el resultado, se muestran datos correspondientes a cuatro de los siete subsistemas, debido a que filtra únicamente aquellos subsistemas que cumplen con la condición establecida,

Ejemplo 2

A continuación se muestra un ejemplo de consulta que utiliza la función Order(), que consiste en obtener el número de alumnos becados por cada subsistema en el año 2000,

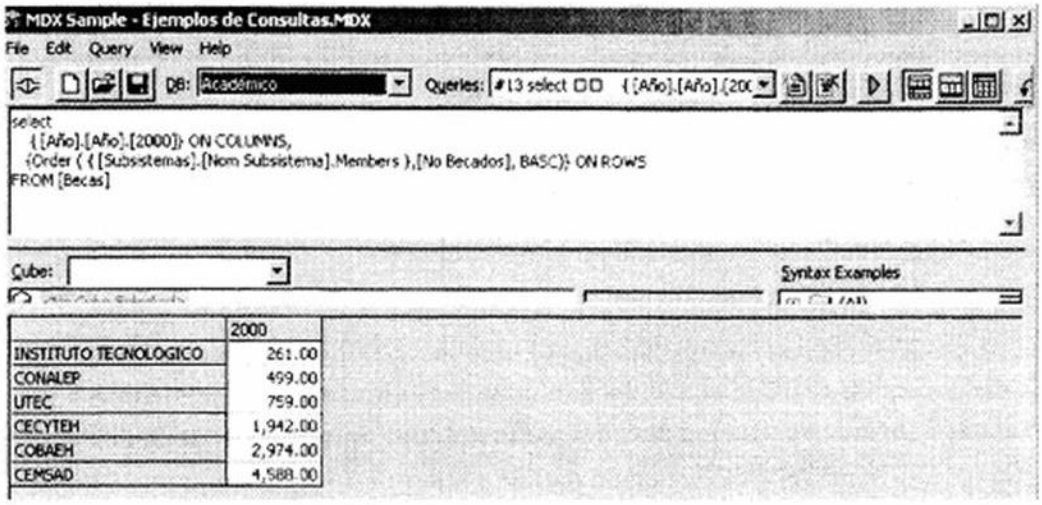


Figura 7-16 Resultado del ejemplo utilizando la función *Orden*

Los parámetros que requiere la función Order (), son la dimensión a tratar, el indicador y el tipo de orden (Ascendente o Descendente). En este caso se utilizó la ordenación ascendente.

Capitulo 8

Implantación del DW en el IHEMSyS

8.1 PERTINENCIA

Actualmente el IHEMSyS, no cuenta con un adecuado sistema que sirva de soporte para la toma de decisiones a nivel directivo. Conforme en la institución crece el número de planteles y alumnos, cada vez resulta más difícil controlar la información para realizar consultas. Actualmente el IHEMSyS, cuenta con aproximadamente 85 planteles distribuidos en sus diferentes subsistemas, sin embargo, en los últimos cinco años se han creado un promedio de cuatro planteles por año.

El IHEMSyS, además de almacenar la información actual, almacena la de semestres anteriores, lo cual permite realizar consultas donde se puedan analizar tendencias. Por lo tanto, conforme pasa el tiempo la cantidad de información que se tiene que almacenar crece, cuando se requiere realizar consultas que involucran los periodos anteriores para detectar el comportamiento de ciertas situaciones, se dificulta la operación. Por ejemplo, para realizar una consulta para determinar el nivel académico que ha tenido alguno de sus subsistemas desde su creación (por ejemplo, cinco años atrás), la dificultad aumenta con la antigüedad del mismo, por la cantidad de información que se necesita manejar.

El IHEMSyS, requiere consultar información para ayudar en la toma de decisiones a sus organismos internos como es el caso de los subsistemas o planteles. Sin embargo, también otras dependencias externas requieren que el IHEMSyS les proporcione algún tipo de información, como ejemplos pueden mencionarse el Instituto Hidalguense de Educación y el Gobierno del Estado. Por lo tanto, el uso de un sistema DW, ayuda a que dicha información se proporcione de forma correcta y oportuna.

Tanto los planteles como la dirección de cada uno de los subsistemas, requieren que el organismo que los controla, en este caso la Dirección General del IHEMSyS, cuente con un sistema adecuado que sirva de soporte a la toma de decisiones. Esto debido a que la toma de dichas decisiones, es a ellos a quienes afecta principalmente. Por lo tanto, en cuanto los

resultados de las consultas sean de mejor calidad y se realicen de forma más oportuna, mejor beneficiados se verán los subsistemas u planteles.

La forma como actualmente el y IHEMSyS maneja la información, no permite visualizar fácilmente alguna situación sobresaliente que se presente, debido a que se tienen distribuidos los reportes en varios archivos. Con un sistema DW implantado en una organización, se puede analizar fácilmente la información almacenada, pudiendo realizar comparaciones de acuerdo a distintos indicadores entre los subsistemas o planteles, para detectar situaciones importantes o sobresalientes. Por ejemplo, teniendo la información almacenada en forma global, es mas fácil visualizar y detectar que en la zona 3 del estado, el promedio en las materias de matemáticas es el más bajo.

Con el sistema de realizar consultas actualmente, se realizan los reportes y consultas que se consideran necesarias o que son solicitadas al IHEMSyS. Sin embargo, con el uso del sistema DW aún se obtienen consultas que no han sido planeadas.

Debido a estos factores, cada vez es más necesario en el IHEMSyS, la creación de un DW, debido a que es una tecnología que permite resolver las necesidades de un sistema de soporte a la toma de decisiones.

8.2 FACTIBILIDAD

En el desarrollo de un DW, se requiere de un análisis costo-beneficio para saber si es factible o no su implantación; a continuación se describen los beneficios a obtener con la implantación del DW.

8.2.1 BENEFICIOS

La implantación del DW beneficiará a la Dirección General del IHEMSyS, debido a que es quien contará con el almacén centralizado de datos. Principalmente al personal directivo de cada Departamento, que serán los usuarios finales, dado que con el DW las consultas se realizarán de forma mas sencilla u se podrán proporcionar resultados de mejor calidad.

Los subsistemas, planteles, u alumnos son los más beneficiados, pues en base a las consultas realizadas, las decisiones que se toman se aplican a estos elementos. El beneficio será aún mayor si dicha toma de decisiones es oportuna.

Por ejemplo, supóngase que con la ayuda del DW, en el Departamento Académico se detecta que en determinada zona del estado, el nivel académico en el área de ciencias exactas es bajo en comparación con las zonas restantes. A partir de dichos resultados, la Dirección General del IHEMSyS va a tomar una decisión que solucionará el problema, por ejemplo, dar cursos de capacitación en esa área al personal docente involucrado.

Como se puede ver, en este caso la decisión beneficiará directamente a los alumnos y docentes de estos planteles y en consecuencia, a los subsistemas y a la institución IHEMSyS en general.

Los beneficios de la implantación de un DW, pueden ser analizados de acuerdo a los siguientes factores:

Obtención de la información necesaria para la toma de decisiones en menos tiempo

La principal ventaja en la implantación de un DW, es que sirve de soporte para la toma de decisiones. La integración de los datos en forma estructurada en un almacén centralizado, da como ventaja el poder obtener información en menor tiempo, lo cual ayudará a realizar la toma de decisiones sin retrasos. Por ejemplo, para saber el índice de reprobación por materia en los diferentes subsistemas en los últimos tres años, puede llevar hasta un día, mientras que con el DW llevaría solo minutos.

Utilizar en otras tareas el personal de apoyó en la realización de consultas

Con la implementación del DW en el IHEMSyS, se requeriría de menos personal para obtener la información necesaria en las consultas. Actualmente, para realizar esta tarea, en cada Departamento, el personal directivo solicita el apoyo de dos técnicos para recuperar la información necesaria. Sin embargo, con el uso del DW, solo se necesitaría del personal directivo y un técnico. Por lo tanto, el personal que en la actualidad se ocupa de apoyo para recuperar la información, puede ser utilizado en otras actividades. Es decir, se ahorraría el trabajo de un técnico por Departamento, lo que traducido en costos, equivaldría a $ 5,000 pesos mensuales por Departamento.

Disminución del tiempo invertido por el personal y mejorar el análisis de los datos

Actualmente, dentro de los diferentes Departamentos del IHEMSyS, los analistas tienen que realizar varios procesos diarios para realizar tareas como: localizar, disponer y analizar datos, con el fin de proporciona* información o recomendar alguna solución a determinada problemática. Desafortunadamente, para realizar estas tareas, dichos analistas dedican mucho tiempo; pues, en ocasiones, es necesario dedicar todo un día para localizar y recuperar datos.

Por ejemplo, supóngase que el Departamento Académico del IHEMSyS requiere saber los porcentajes de deserciones en los últimos 4 años en los diversos subsistemas. Como está estructurada la información actualmente, es necesario recopilarla a partir de los informes de los diversos subsistemas, además de tener que reunir la de los periodos anteriores.

La disponibilidad de integrar y preparar datos accesibles en el DW, reduciría significativamente el tiempo que los analistas ocupan en tareas de colección de

información e incrementaría el tiempo que tienen disponible para analizar los datos que se tienen coleccionados en el DW. Esto da como ventaja disminuir el tiempo que involucra el personal para la toma de decisiones y propiciar una mejor calidad en el análisis de los datos.

8.2.2 COSTOS

Los costos de la implantación de un DW, típicamente caen dentro de las categorías que se explican a continuación:

Hardware

Se refiere a los costos asociados con lo referente al Hardware y ambiente operativo requerido para la creación de un DW. Dentro de una implantación DW, el hardware es el requerimiento que se considera más costoso, debido a que puede requerir hasta un 50 % de la inversión total.

Dependiendo del equipo con el que cuente una corporación, puede ser necesaria la adquisición de nuevo equipo o la actualización del equipo existente. Obviamente, los DW de mayor magnitud tienen costos mas elevados en la adquisición de hardware. Un DW crece en forma rápida, lo que provoca que en poco tiempo tenga que ser necesario actualizar el servidor. Una buena opción para evitar este problema es seleccionar un servidor que pueda ser escalable.

A continuación se describe el equipo de hardware requerido tanto para el servidor como para los equipos cliente, incluyendo su precio aproximado:

1 Servidor HP 9000 Enteprice (Multiprocesamiento simétrico)	$ 22,000
3 Equipos Pentium IV, 400 MB en disco duro, 128 MB en RAM	$ 15,000
Cable y conectores	
Tarjetas de red	
Concentrador de 24 puertos	

Se propone un servidor con multiprocesamiento simétrico (sus características y ventajas se expusieron en el capítulo 4, en tipos de procesadores), debido a que este tipo de senador en un principio puede trabajar con un mínimo de dos procesadores, pero conforme se requiere de mayor capacidad de procesamiento, puede escalar adicionándosele mas procesadores. Esto se hace previendo el crecimiento del DW a futuro.

Como se mencionó en el capítulo 4, la metodología a utilizar es la Rapid Warehousing, que consiste en ir creando los datamarts de uno en uno. Debido a que en el IHEMSyS, en un principio se iniciará la implantación del DW con el data mart en el Departamento

Académico, se considera que en el inicio son suficientes tres equipos cliente, en un futuro se necesitaran mas para el uso de los Departamentos restantes.

Los equipos que actualmente se utilizan para realizar las consultas, se ocupan en forma paralela para otras actividades y el uso de algunos sistemas. Sin embargo, para mejor funcionamiento en la explotación de los datos, se sugiere adquirir los equipos antes mencionados, para que sean dedicados exclusivamente para la explotación de los datos a partir del DW.

Software

Se refiere al costo de las licencias para el uso de los productos de software. Un tipo de estos productos son aquellos que ayudan a realizar procesos como la extracción, limpieza, carga y recuperación de los datos. Otro tipo de producto es la plataforma sobre la cual trabajará el software antes mencionado A continuación se mencionan los requerimientos del Software requerido, tanto para el servidor como para los equipos cliente, con su precio aproximado:

Windows NT Server 25 clientes	$ 12,000
Microsoft SQL Server 2000 edición empresarial (25 Clientes)	$ 26,250
Windows 2000 Profesional	$ 1,500

La razón por la que se propone Microsoft SQL Server edición empresarial, es porque a diferencia de la edición estándar, la edición empresarial contiene las herramientas necesarias para el proceso de desarrollo del DW. Para el proceso de Extracción, Transformación y Carga cuenta con la herramienta Data Transformation Services (DTS). Para el análisis multidimensional incluye OLAP Services y, por último, para el proceso de la explotación de datos tiene el lenguaje de consulta MDX.

El número de clientes considerados tanto en Windows NT como en SQL Server prevén el crecimiento del número de clientes en los distintos departamentos del IHEMSyS.

Costos de personal

Este elemento se refiere a los costos efectuados en el personal que se utilizará durante el desarrollo del DW. En el caso de estudio, se liará uso del personal que está laborando en el IHEMSyS, este personal actualmente se dedica a ayudar al personal directivo a recopilar la información necesaria.

Una vez implementado el DW, el personal que se requiere contratar, es únicamente una persona del área de informática con experiencia en la tecnología DW, que se dedique a dar mantenimiento al almacén de datos, el costo aproximado para este fin sería de 5 10,000 pesos mensuales, debido a que en este momento no se cuenta con una persona que domine la tecnología DW.

Servicios de consultoría

Este factor se refiere a los servicios provistos por consultores o asesores durante el proceso de creación del DW, el cual se considera lleva un tiempo aproximado de 2 a 3 meses por data mart. El uso de consultores, es más popular cuando la empresa aún esta aprendiendo sobre tecnologías y técnicas DW, por lo tanto la asesoría se requerirá solo cuando se cree el primer data mart

El IHEMSyS podrá auxiliarse en el departamento de informática, misma que deberá apoyar en el proceso de creación y asesoría. Sin embargo en el caso más extremo, si el personal de informática no logra resolver todas las necesidades, se podría requerir de la asesoría de personal externo, la cual se considera tiene un costo aproximado de $ 300 pesos por hora. De acuerdo al tiempo requerido en asesoría, se sugiere que se destine un aproximado de $ 15,000 pesos para esta actividad.

Capacitación

Una vez implementado el DW, será necesario invertir en capacitar a los usuarios en el uso del Software de los sistemas cliente, es decir, a los directivos de cada Departamento. La capacitación, principalmente se refiere enseñar a los usuarios a manejar la aplicación y el lenguaje de consulta, por medio de cursos internos en el lugar donde esté implementado el DW.

El Departamento de Informática puede capacitar al personal con un curso básico del lenguaje SQL, debido a que tiene dominio sobre este lenguaje. El lenguaje MDX, se sugiere sea impartido por personal externo, el cual tendrá una duración aproximada de
50 Loras y un costo aproximado de $ 20,000 pesos.

51se unen los costos de cada tina de las categorías antes mencionadas, se obtiene un total aproximado de $125,000 pesos. En cuanto al costo de personal contratado, como se mencionó anteriormente en el administrador del DW se invertiría un aproximado de $10,000 pesos mensuales. Considerando el presupuesto con el que cuenta el IHEMSyS disponible para la actualización de equipo y creación de sistemas que ayuden a mejorar el nivel educativo, estas cantidades están al alcance de dicho presupuesto.

Respecto a los ahorros que tendría el IHEMSyS, estos se verían reflejados principalmente en menos utilización de personal. Considerando que a mediano plazo la implementación se llevaría a cabo en tres Departamentos (Académico, Recursos Financieros y Recursos Materiales), debido a que por su importancia son quienes más lo necesitan. Como se mencionó anteriormente, el ahorro por departamento es de S 5000 pesos, por lo tanto, una vez implantado el DW, el ahorro del IHEMSyS sería de $ 15,000 pesos mensuales.

Lo anterior implica que tomando en cuenta la contratación del administrador del DW por $ 10,000 pesos mensuales, en dos anos, aproximadamente se recuperaría la inversión realizada.

Considerando los beneficios que la institución IHEMSyS va a obtener al contar con el DW que le sirva de base para la toma de decisiones, que en dicha institución es un elemento indispensable y tomando en cuenta la inversión presupuestada, se considera que es factible hacer la implementación del DW.

8.3 PROCESO DE IMPLANTACIÓN DEL DW EN EL IHEMSyS

En este punto se mencionan los requerimientos y tareas necesarias para poder realizar la implantación del DW en el IHEMSyS.

Hasta el momento, el trabajo que se ha realizado es analizar la organización, la forma en cómo manejan la información que se requiere en la creación de un DW; y la detección de sus bases de datos fuente.

Además se ha realizado el análisis multidimensional de los Departamentos más importantes (Académico, Recursos Materiales y Recursos Financieros) y el diseño multidimensional para el Departamento Académico.

Se han convertido los datos requeridos para el Departamento Académico, utilizando una base de datos estándar que se obtuvo a partir del diseño multidimensional. Se ka cargado la información, por medio de la creación de los cubos de los distintos indicadores. El trabajo realizado, hasta el momento en el Departamento Académico, puede ser aprovechado cuando se tenga el equipo necesario para la implantación del DW, con el solo hecho de copiar las bases de datos y los cubos creados, al equipo que va a funcionar como servidor.

Para iniciar la implementación del DW en la Dirección General del IHEMSyS, primero es necesario adquirir e instalar o configurar el equipo necesario. A continuación se muestra una relación de las necesidades de Hardware y Software.

Hardware

- Servidor HP 9000 Enteprice (Multiprocesamiento simétrico)
  - 3 Equipos Pentium IV, 400 MB en disco duro, 128 MB en RAM
  - Cable y conectores
  - Tarjetas de red
  - Concentrador de 24 puertos

Software

- Windows NT Server 25 clientes

Microsoft SQL Server 2000 edición empresarial (25 Clientes)
Windows 2000 Profesional

Una vez adquiridos estos requerimientos, se deben realizar las siguientes tareas:

Conexiónala red
Conectar a la red existente el equipo adquirido, el servidor con el sistema operativo Windows NT y los equipos cliente con Windows 2000. El servidor adquirido se debe conectar como un nodo a la red principal, y los equipos cliente a dicho servidor, utilizando una arquitectura cliente servidor.

Instalación del Software
Una vez instalada la red, se debe instalar el Software restante en el siguiente orden: En el servidor SQL Server y OLAP Services y en los equipos cliente SQL Cliente y el lenguaje MDX,

Una vez instalado el equipo, se pueden realizar las siguientes tareas:

Cargar la información
Como se mencionó anteriormente, en el Departamento Académico es donde se implantara el primer data mart, por lo tanto, se puede aprovechar la base de datos y cubos ya diseñados. Como primer paso, se debe copiar al servidor la base de datos del Departamento Académico.

Posteriormente, los cubos que hasta el momento se han creado, pueden copiarse al servidor, naciendo uso de la herramienta OLAP Services, específicamente de OLAP Manager.

Capacitar al personal
Para explotar los datos del DW creado, es necesario que se capacite a los usuarios finales. Como se mencionó anteriormente en el estudio de factibilidad, la capacitación se debe realizar en el lenguaje SQL y MDX El curso en SQL puede ser a nivel básico, sin embargo, el lenguaje MDX debe ser a un nivel avanzado, debido a que de esto depende con que facilidad obtengan los usuarios la información.

Respecto a los subsistemas, se sugiere que todos sus planteles cuenten con servicio de Internet, para agilizar el proceso de recuperación de datos.

CONCLUSIONES

La selección de una herramienta que se utilice para el desarrollo de un DW no es fácil, debido a que se deben considerar diversos aspectos. Uno de ellos, es verificar que procesos permite realizar (extracción, transformación, carga y explotación de los datos). Algunas herramientas, permiten realizar un solo proceso, otras la combinación de algunos de ellos y otras todos los procesos. Otros de los aspectos a considerar son: el tamaño del DW, la compatibilidad de plataformas y su costo.

También se pudo detectar, que el proceso más laborioso es el de la transformación de los datos. La dificultad de este proceso varía de una organización a otra, dependiendo de la cantidad de información que maneje y que tan heterogéneas sean las diversas bases de datos fuente. Sin embargo, en el caso de estudio disminuyó la dificultad de este proceso, debido a que existía un estándar en plataformas y manejadores de bases de datos.

Se logró realizar el análisis en los tres Departamentos más importantes del IHEMSyS (Académico, Recursos Financieros y recursos Materiales).

Se realizó el diseño multidimensional para el Departamento Académico, para lo cual, se modelaron los diagramas de estrella y snowflake, y se determinó la arquitectura del depósito y del servidor.

Se logró realizar la conversión de datos para el Departamento Académico, lo cual se realizó con la herramienta DTS de MS OLAP.

Se creó el datamart del Departamento Académico con la integración ele los cubos, esto se realizó con la herramienta OALP Manager de MS OLAP.

Se crearon ejemplos de explotación de los datos a partir del data mari creado, realizando operaciones OLAP, esto se realizó con el uso de la herramienta MDX.

Se demostró que es pertinente y factible desarrollar e implantar un DW en el II"IHEMSyS. Con ello, el IHEMSyS mejorará sus resultados en el ámbito de toma de decisiones, además de que necesitará menos personal.

TRABAJO FUTURO

Para que el DW se implante de forma total en el IHEMSyS, es necesario realizar varias tareas, a continuación se describen las más importantes:

- Cargar la información respecto a los cuatro años anteriores a 2002 en el Departamento Académico.
- Realizar el diseño multidimensional, la conversión u carga de datos para los departamentos de Recursos Materiales u Recursos Financieros.
- Realizar el análisis de los Departamentos restantes (Jurídico, Programación u Estadística e Informática), para determinar si requieren la creación de un data mart
- Para aquellos departamentos que lo requieran, realizar el análisis u el diseño multidimensional, la conversión de datos u la carga de la información.
- Crear una aplicación que sea más fácil de utilizar por el usuario que el lenguaje MDX.
- Rediseñar las bases de datos de los diversos subsistemas u planteles para que proporcionen la información necesaria en forma directa.

En lo que respecta a los subsistemas, una de las tareas que deben realizarse en los diferentes Departamentos, es ajustar los formatos de recopilación de datos como se sugiere en el capítulo 3, en el análisis de la información necesaria para el DW. Es decir, que en los reportes solicitados a los planteles, se involucren todos los indicadores u dimensiones necesarias que se obtuvieron en el análisis y diseño multidimensional.

Respecto a los planteles, se sugiere que los que aún no cuentan con servicio de Internet (se mencionaron en el capítulo 3 en el análisis de la organización), llagan su instalación. Esto, debido a que por la ubicación geográfica en que se encuentran dichos planteles, el trabajo puede retrasarse y afectar a toda la institución.

REFERENCIAS

[1] Jiawei H. and Kamber M.; Data Mining Concepts and Techniques; Morgan Kaufmann Publisher; 2001.

[2] Mukesh M.; 'Data Warehousing and Knowledge Discovery"; Ed. Springer; 1999.

[3] Humphries M. and Hawkins W.; Data Warehousing: Architecture and Implementation; Prentice Hall; 1999.

[4] Chauduri S. and Day al U.; An overview of data Warehousing and OLAJP technology"; Prentice Hall; 1997.

[5] Berson A, and Smith S.; "Data Warehousing, Datamining and OLAP"; Me Graw Hill; 1997.

[6] SaltonM.; "Introduction to Modern Information Retrieval; McGraw Hill; 1993.

[7] Gutting R; "An introduction to spatial database systems"; The VLDB Journal; 1994.

[8] Widom J.; Research Problems in data warehousing; Conf. Information and Knowledge Management, Baltimore; 1995.

[9] Inmon W.; "Building the Data Warehouse"; John Wiley & Sons; 1996.

[10] Mattison R; "Data Warehousing and Datamining"; Artech House; 1997.

[11] Sarawagi S., Agrawa lR and Migiddo R; "Exploration of OLAP data cubes"; Prentice Hall; 1999.

[12] Thomson E.; "OLAP Solutions: Building Multidimensional Information Systems; John Wiley & Sons; 1997.

[13] Kimball R; "The DW toolkit"; John Wiley & Sons; 1996

[14] Westphal C. and Blaxton T.;" Datamining Solutions: Methods and Tools for Solving Real-World Problems"; John Wiley & Sons; 1999.

[15] Fayyad M. and Smyth P.; "Advancing Knowledge Discovery and Data Mining"; Editors Cambridge; 1996.

[16] Chen M.; 'Datamining: An overview from a database perspective; IEEE Data Engineering; 1996.

[17] Ullman J.; Principles of Database and Knowledge-Base systems ; Computer Science; 1998.

[18] Poe V.; "Building a Data Warehouse for Decision Support; Prentice Hall; 1996.

[19] Kueng P.; "A Holistic Process Performance Analysis through a process Data Warehouse; American Conference of Information Su stems; 2001

[20] Han J.; " Data warehouses and Datamining; Addison; 1998.

[21] Shoshani A; "OLAP and statistical databases"; Tucson AC; 1997.

[22] Harinarayan V.; "Implementation data cubes efficiently'; Conf. SIGMOD 96, 1996.

[23] Peterson T. and Pinkelman I;" Microsoft OLAP"; Ed. SAMS; 1999.

[24] Blaxton J.; Genexus Warehouse; Editors; 1999.

[25] Rose R; "Manual de Genexus OLAP"; Artech; 1998.

[26] Inmon, W.; "Data Warehouse Performance"; John Wileu, 1999.

[27] Jarke, M.; "Fundamentals of Data Warehouses"; Springer; 2000.

[28] Harjinder S.; "Data Warehousing"; Prentice Hall, 1996.

[29] Witten H.; "Data Mining"; Morgan Kaufmann; 2000.

[30] Moeller R; "Distributed Data Warehousing using Web Technology"; 2001.