UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

F A C U L T A D D E I N G E N I E R Í A

ANÁLISIS Y DESARROLLO DE MATERIAL DIDÁCTICO DE LA ASIGNATURA "DEPÓSITO DE DATOS".

T E S I S

QUE PARA OBTENER EL GRADO DE

I N G E N I E R O E N C O M P U T A C I Ó N

P R E S E N T A N :

HORACIO HERNANDEZ ALVARADO
DAVID RAMÍREZ ARREOLA
MA. MAIYEC URRUTIA LUNA

DIRECTOR DE TESIS:
M.I. JORGE VALERIANO ASSEM

MÉXICO, D.F. MARZO, 2007.

A Dios.
Por la vida y por permitirme tener a la mayor parte de mis seres queridos a mi lado.

A mis padres.
Por el cariño brindado en el transcurso de mi vida y el apoyo durante mis estudios.

A mi esposa.
Por su apoyo y comprensión en la realización de una meta mas en mi vida.

A mi hijo.
Por brindarme su amor y ser una de las fuentes principales de motivación de mi existencia.

A la Facultad de Ingeniería.
Por compartir conmigo la luz del conocimiento y darme las herramientas necesarias para construir mi vida profesional.

A todos.
En especial a todas aquellas personas que directa o indirectamente influyeron en mi formación profesional.

Un agradecimiento especial a M. en I. Jorge Valeriano Assem. Por transmitirnos su entusiasmo y conocimiento. Además de brindarnos su invaluable apoyo para la realización de este trabajo, que se significa la culminación de una meta muy importante para nosotros.

Horacio Hernández Alvarado

Esta Tesis no hubiera sido posible sin el esfuerzo e intervención de muchas personalidades.

Agradezco y admiro todo el trabajo, apoyo y sacrificio de mis Padres. Los Amo.

A la cooperación de mis compañero de tesis.

Agradezco el aliento de cada unos de mis estimados amigos, en especial a Agustín, David, Arlette y Ruth quienes a través de su amistad y apoyo he superado obstáculos.

Y particularmente a la Facultad de Ingeniería, a la UNAM, así como a la excelencia de cada uno de su profesorado y enseñanza.

A todos y cada unos de mis hermanos que de alguna u otra forma han contribuido.

Citarlos a todos sería imposible, aunque cada uno de ellos debe verse personalmente en este agradecimiento. A todos ellos que han compartido conmigo momentos importantes, que me han ayudado a superar todo obstáculo, y mejorar cada día en mi vida, a entender que el esfuerzo, honestidad, trabajo y sacrificios, siempre al final nos tiene una recompensa, a todos ellos les dedico esta tesis.

Ma. Maiyec Urrutia Luna

A mis Padres:

Con profundo cariño y respeto y sobre todo el mas preciado tesoro que puede tener un hijo de sus padres el Amor, les agradezco el tiempo, la dedicación el esfuerzo por darme los instrumentos necesarios para enfrentar la vida diaria, les agradezco por su infinita sabiduría que día a día me acompaña y me cobija durante los momentos difíciles, solo puedo expresar estas pobres palabras pero llenas de sinceridad, los amo, y agradezco a dios el darme la oportunidad de tenerlos como guías, por sobre todo a mi querida Madre Guadalupe Arreola Benigno sus desvelos y lagrimas, sus cuidados y esa gran madera para sobrellevar a toda una familia sobre su espalda y bajo su firmeza como mujer quien siempre ha sido será y seguirá siendo la mujer que mas quiero en la vida, a mi padre David Ramírez González este gran hombre que merece por el resto de mis días mi agradecimiento y respeto por nunca dejar de apoyarme y enseñarme a no titubear al momento de confrontarme con mis problemas.

Por todo lo que son ustedes soy yo…

A mis Hermanos:

Agradezco a todos mis hermanos Fily, Mari, Elsa, Josefina, José, Carmen, Lupita, esos grandes momentos que vivimos durante nuestra niñez, y que me enseño a tener amigos, y respetar el espacio de los demás, a saber que aunque la vida sea dura siempre hasta el ultimo día de mis días estaré con ustedes y ustedes conmigo les doy las gracias por apoyarme y colocarme en un lugar especial entre ustedes, gracias hermanos y con una lagrima en mis ojos para el mejor de mis hermanos José que no se lo llevaron de nuestra vida al contrario el esta mas presente cada día y la promesa de sobresalir y la fe que nunca dejo de tener hacia mi persona le agradezco es mi promesa algún día estaremos juntos y jamás podrán separarnos gracias hermano.

A mi Universidad:

Pocos somos los privilegiados en ser parte de esta gran familia llamada UNAM , la mejor universidad del mundo, porque nunca en ningún lugar encontraremos tanta riqueza cultural, ni el apoyo académico como nuestra alma mater, y sobre todo a este gran núcleo que forma parte de ella la Facultad de Ingeniería, gracias Facultad, gracias UNAM, y sobre todo gracias México, por darme la oportunidad de estar aquí, agradezco a los profesores que forjaron mi camino gracias

A mis Amigos:

Quizá los tuve contados pero sinceros, por eso no quiero olvidar a ninguno, gracias Maiyec, Agustín, Ruth, Jesús, Julián que a pesar de conocernos por un tiempo de nuestra vida los aprecio y estimo ya que sin ustedes no hubiese podido continuar mi camino sin saber que hay alguien que también es mi familia a los que también me han a acompañado, claudio, oscar… y demás gracias.

David Ramírez Arreola.

Índice

INTRODUCCIÓN 3
CAPÍTULO 1 6
ANÁLISIS DE LA ASIGNATURA DEPÓSITO DE DATOS 6
1. 1 Antecedentes Curriculares 7
1. 2 Análisis de los temas de la asignatura 14
1.2.1 Metodología de Análisis. 14
1. 3 Sugerencias y comentarios 21
CAPÍTULO 2 22
DESARROLLO DEL MATERIAL DIDÁCTICO 22
2.1 Introducción a los Depósitos de datos 23
2.1.1 Antecedentes 23
2.1.2 Datos del mundo real 29
2.2 Planeación de los depósitos de datos 32
2.2.1 La Arquitectura de un Depósito de Datos. 32
2.2.2 Un caso de Negocio 51
2.2.3 Un enfoque para el desarrollo de un Depósito de Datos 61
2.3 Los Datos 87
2.3.1 Calidad de los datos: 87
2.3.2 Metadatos 91
2.3.3 El papel de un directorio/catálogo 92
2.3.4 Transformación de los datos 94
2.4 Diseño e implementación 105
2.4.1 Diseño físico 105
2.4.2 OLAP multidimensional y OLAP relacional 116
2.4.3 Replicación de datos en un ambiente de Depósito de datos global 122
2.4.4 VLDS y paralelismo 125
2.4.5 Implementación de un Depósito de datos en un ambiente multiservidor con tecnología paralela 143
2.4.6 El Depósito de datos virtual 167
2.4.7 Minería de datos 167
2.5 Administración de los depósitos de datos 178
2.5.1 Administración de aplicaciones 178
2.5.2 Administración de bases de datos en un ambiente heterogéneo 185
2.5.3 Seguridad en los depósitos de datos 189
2.5.4 Selección de las herramientas para el usuario final 191
2.6 Tendencias 196
CAPÍTULO 3 201
3.1 Caso Práctico 202

Anexo (A) 239
Anexo (B) 242
Conclusiones 247
Glosario 250
BIBLIOGRAFÍA 255

INTRODUCCIÓN

Hoy en día las organizaciones requieren de sistemas que les permitan disponer de la información de manera rápida y eficaz, esta disposición se torna crítica al momento de tomar las decisiones adecuadas, poder responder a los cambios de las empresas y manejar esta información de forma competitiva, pero la cantidad de información disponible en diferentes fuentes de datos heterogéneas, como Internet, librerías digitales, bases de datos antiguas, etc. de forma separada, a menudo es inconsistente y contradictoria no permite la toma de decisiones de forma inmediata y adecuada, las fuentes de datos en Internet son dinámicas y cambian constantemente por lo que es necesario recopilar la información contenida en las fuentes, así también como la incorporación a un Depósito de Datos.

Son limitadas las instituciones que presentan como materia el tema de Depósitos de datos. El objetivo de esta tesis es analizar el temario de la asignatura "Depósito de Datos" impartida en la Facultad de Ingeniería de la UNAM, con el fin de desarrollar el material de consulta que sirva como apoyo a los estudiantes que cursen la asignatura.

Para que los futuros Ingenieros tengan la capacidad y preparación adecuada para la construcción de soluciones de toma de desiciones es decir de un Deposito de Datos es necesario que se encuentren apoyados por un material didáctico que abarque los concepto que les permitan aplicar los métodos y técnicas para la implementación e incorporación de información de cualquier empresa, de forma semiautomática y poder utilizarlos como soporte a los sistemas de ayuda de decisión obteniendo dichos datos de fuentes heterogéneas. Los alumnos obtendrán un conocimiento que les ayudara a obtener un nivel competitivo en el exterior.

Es importante saber que los Depósitos de datos no se pueden comprar se tienen que construir, no es exclusivamente un almacén de datos, y dan respuestas a las demandas de alto rendimiento de datos de información de una organización. Así, los Depósitos de datos proporcionan al usuario una interfaz consolidada única para los datos, lo que hace más fáciles de escribir las consultas para ayuda a las decisiones.

Los Depósitos de datos son las arquitecturas centrales de los sistemas de información actuales, permite organizar y almacenar los datos necesarios para el procesamiento desde la perspectiva histórica a través del tiempo, La ventaja principal de este tipo de sistema es que se basan en "La estructura de información", es decir, el almacenamiento de la información homogénea y fiable, en una estructura basada en la consulta y el tratamiento jerarquizado de la misma,

y en un entorno diferente de los sistemas operacionales. También utilizan o soportan varios tipos de aplicaciones como OLAP y aplicaciones de minería de datos OLAP.

Existen diferentes niveles de datos de los Depósitos de datos, desde datos muy detallados hasta datos resumidos, otro tipo de datos son los meta datos, que contienen información acerca de los datos, una especie de directorio que contiene información de cómo son los datos almacenados en un Depósitos de datos y donde se pueden encontrar.

También es imprescindible aplicar la calidad de la información, que viene determinada por la calidad tanto del Depósito de datos como por la calidad de la presentación de los datos. De hecho, es muy importante que los datos del Depósito reflejen correctamente el mundo real, pero es también muy importante que los datos sean interpretados correctamente.

Para implantar un Depósito de datos a grandes rasgos se debe construir el modelo de datos; determinar los datos que vamos a almacenar; diseñar el modelo físico; desarrollar los programas de transformación de los datos; cargar y mantener el Depósito de datos; y construir y mantener el directorio de los meta datos.

Y con el fin de mejorar el acceso a los datos, debido a veces por el almacenamiento masivo de información, se requiere de la construcción de arquitectura en paralelo que descomponen grandes problemas en pequeños fragmentos de modo que cada fragmento del problema pueda ser ejecutado en paralelo por cada nodo, es decir, a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas, llamadas los centros comerciales ó DataMarts. El Paralelismo entre consultas es para proporcionar mayor rendimiento al dividir una única consulta compleja en varias partes y distribuir la carga de trabajo entre múltiples procesadores.

En material didáctico se ve más a detalle cada uno de los componentes y métodos que ayudarán al alumno de la asignatura Depósito de datos.

CAPÍTULO 1

ANÁLISIS DE LA ASIGNATURA DEPÓSITO DE DATOS

1. 1 Antecedentes Curriculares

La Facultad de Ingeniería es una institución de educación superior la cual imparte varias carreras de ingeniería, esta institución es generadora de profesionales capacitados para el sector público y privado.

La carrera Ingeniería en Computación fue creada en el año de 1977 y ha tenido una serie de actualizaciones con el fin de mantener vigentes los planes y programas de estudio. El avance científico que se ha logrado en las diferentes áreas de conocimiento y las innovaciones tecnológicas en telecomunicaciones, manufacturas, procesos e informática ha motivado la revisión y modificación de los planes de estudio de esta disciplina.

La penúltima modificación realizada a los planes y programas de estudio de la carrera de Ingeniería en computación fue en el año de 1996, sin embargo el área de la Informática presenta una dinámica propia del avance tecnológico y de las condiciones económicas y sociales, este avance debe ser tomado en cuenta para formar recursos humanos capacitados en las nuevas tecnologías de información, tratando con esto de satisfacer las nuevas necesidades que demandan las empresas hoy en día.

Debido a esto, la facultad de Ingeniería realizo una revisión integral del plan y programa de estudios de la carrera de Ingeniería en Computación con el fin de que esta siga manteniendo el liderazgo en la formación de profesionales en el área. Esta revisión del plan y programa de estudio fue aprobado por el consejo técnico de la facultad en el año 2005, en este plan de estudios se busca una congruencia en los contenidos y en la secuencia de las asignaturas así como mejorar la calidad académica de la facultad.

Una modificación importante al plan de estudios vigente de la carrera de ingeniería en computación, es su reducción de 10 a 9 semestres y de 448 créditos a 408. Además, mientras que el plan 96 se estructura en 3 bloques de asignaturas, el plan de estudios propuesto está organizado por asignaturas con un mínimo de seriaciones obligatorias. Por otro lado, se eliminan los cursos propedéuticos del plan de estudios vigente, al incorporarse sus contenidos relevantes a otras asignaturas básicas curriculares.

Este plan de estudios propone asignaturas las cuales abarcan temas de vanguardia de acuerdo a los intereses y vocación de los alumnos con el fin de satisfacer las expectativas de los alumnos y con esto aumentar la eficiencia terminal de los estudiantes de la carrera ingeniería en computación.

Con todos estos cambios se busca dar bases al egresado de la carrera de ingeniería en computación para poder competir en un mercado laboral donde las empresas buscan personal capacitado en las últimas tendencias de la información.

El plan de estudio 2005 agrupa varias modificaciones una de estas fue estructurar los contenidos de las asignaturas de ciencias de la ingeniería evitando la duplicidad de conocimientos de ciencias básicas.

Este plan de estudios cuida la relación contenido-tiempo de impartición de las asignaturas para que los alumnos puedan asimilar mejor los conocimientos. También se depuraron los programas de las asignaturas eliminando temas superfluos a fin de detallar y lograr mayor profundidad de los restantes.

Una modificación fundamental en este plan de estudios fue el actualizar los contenidos de las asignaturas ante el constante avance del conocimiento, este plan incorpora asignaturas con temas nuevos para que los alumnos estén actualizados, la actualización de los temas también consistió en eliminar asignaturas cuyos temas han perdido relevancia.

De acuerdo a los datos publicados y en base a la fundamentación para el cambio de plan de estudios partimos del mapa curricular del plan de estudios propuesto para el año 2005 y con una duración de nueve semestres como se muestra en la figura 1.1.

Con este plan de estudio se pretende formar egresados con conocimientos generales y específicos esto se pretende lograr con los módulos terminales, estos módulos están formados por asignaturas las cuales están diseñadas para satisfacer los requerimientos y expectativas tanto de los estudiantes como de los empleadores.

Estos módulos terminales permiten encaminar la formación del egresado hacia alguna de las áreas del campo de trabajo de la ingeniería en computación de una manera estructurada y ordenada enfocándose a las áreas de mayor demanda laboral, con lo que el estudiante adquiere conocimiento especializado el cual lo hace competitivo en un área de oportunidad acorde con sus intereses y vocación.

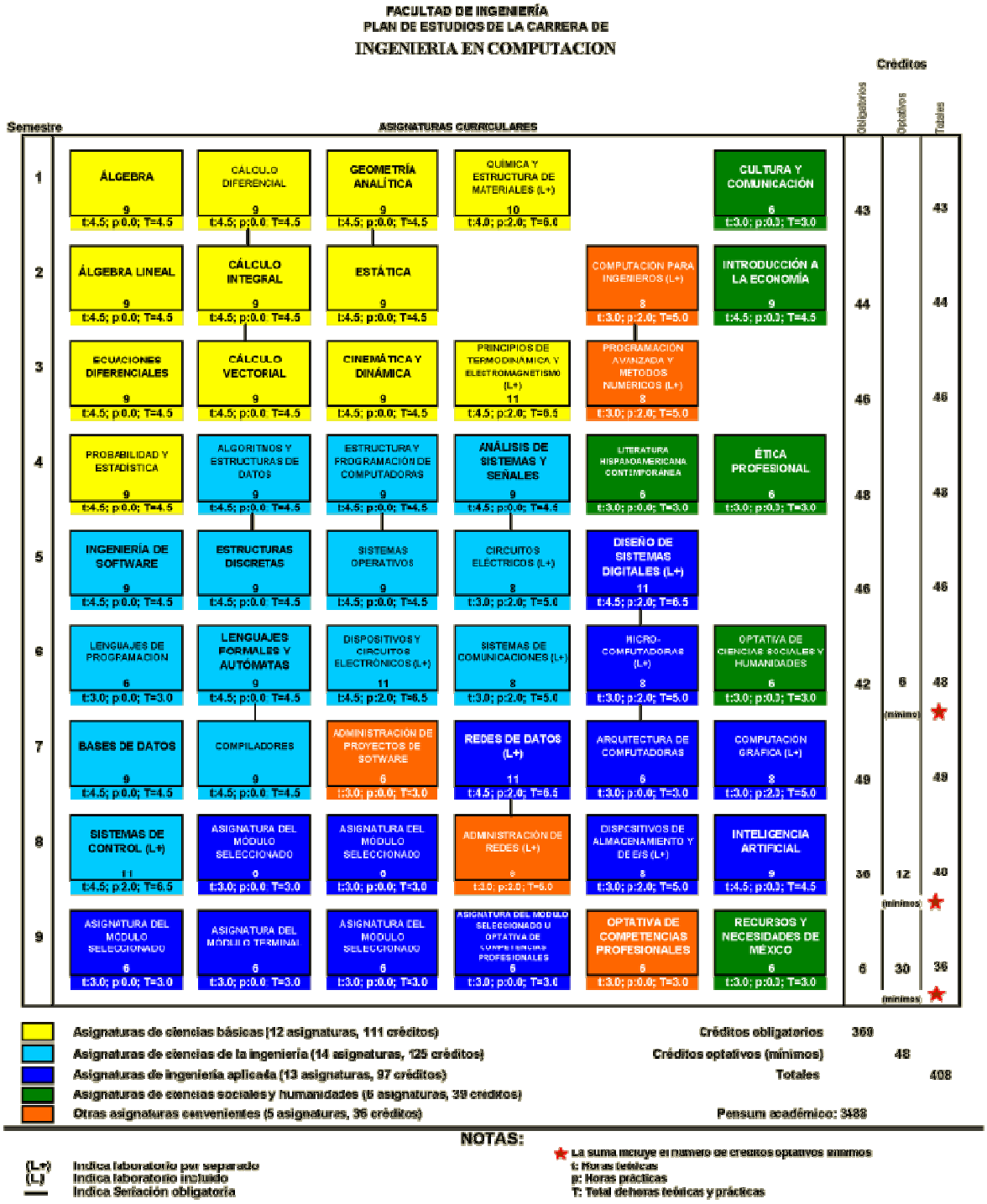

Figura 1.1
Plan de Estudios de la Carrera de Ingeniería en Computación.

El plan de estudios de la carrera de ingeniería en computación fue aprobado por el Consejo Académico del Área de las Ciencias Físico Matemáticas y de las Ingenierías el día 11 de agosto de 2005. Incluye materias optativas en todos sus módulos de salida, las cuales ayudan a la formación académica del estudiante.
Estos módulos de salida son:

- INGENIERIA DE HARDWARE
- REDES Y SEGURIDAD
- BASES DE DATOS
- INGENIERIA DE SOFTWARE
- SISTEMAS INTELIGENTES Y COMPUTACION GRAFICA
- INGENIERIA BIOMEDICA

Cada módulo de salida cuenta con materias obligatorias y optativas las cuales se muestran en el Anexo (A).

Estas asignaturas pueden ser cursadas dependiendo de los intereses propios de cada alumno.

El módulo de salida donde se encuentra la materia que estamos analizando se llama Bases de Datos. Este modulo incluye 3 asignaturas obligatorias y 5 optativas, las cuales son:

OBLIGATORIAS

- BASES DE DATOS AVANZADAS
- BASES DE DATOS DISTRIBUIDAS
- BASES DE DATOS ESPACIALES

OPTATIVAS

- DEPÓSITOS DE DATOS
- MINERÍA DE DATOS
- PROYECTO DE INVESTIGACIÓN
- SEMINARIO DE TITULACIÓN
- TEMAS SELECTOS DE BASES DE DATOS

De las cuales se toma la materia depósito de datos como caso de estudio.
La materia depósito de datos pretende dar herramientas básicas al egresado para que este sea capaz de entender conceptos de análisis, diseño, implantación y administración de información.

Los temas considerados dentro de la materia Depósito de Datos abarcan los conceptos básicos así como las tendencias de los depósitos de datos, cuya finalidad es la obtención de los conocimientos necesarios para que el estudiante

pueda proporcionar una alternativa de solución a los problemas que se le presenten en el ámbito de las tecnologías de información y se enlistan en la figura 1.2

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE INGENIERÍA

PROGRAMA DE ESTUDIO

DEPÓSITOS DE DATOS	0684	8°, 9°	06
Asignatura	Clave	Semestre	Créditos

Ingeniería Eléctrica	Ingeniería en Computación	Ingeniería en Computación
División	Departamento	Carrera en que se imparte

Asignatura:		Horas:		Total (horas):	
Obligatoria		Teóricas	3.0	Semana	3.0
Optativa	X	Prácticas	0.0	16 Semanas	48.0

Aprobado:	Fecha:
Consejo Técnico de la Facultad	25 de febrero, 17 de marzo y 16 de junio de 2005
Consejo Académico del Área de las Ciencias Físico Matemáticas y de las Ingenierías	11 de agosto de 2005

Modalidad: Curso.

Asignatura obligatoria antecedente: Ninguna.

Asignatura obligatoria consecuente: Ninguna.

Objetivo(s) del curso: El alumno cubrirá todos los aspectos de la planeación, diseño, desarrollo, implementación y administración de los depósitos de datos.

Temario

NÚM.	NOMBRE	HORAS
1.	Introducción a los depósitos de datos	2.0
2.	Planeación de los depósitos de datos	9.0
3.	Los datos	9.0
4.	Diseño e implementación	10.0
5.	Administración de los depósitos de datos	10.0
6.	Tendencias	8.0
		48.0
	Prácticas de laboratorio	0.0
	Total	48.0

DEPÓSITOS DE DATOS **(2 / 4)**

1 Introducción a los depósitos de datos

Objetivo: El alumno(a) expondrá (verbalmente y/o mediante un ensayo) la importancia de los depósitos de datos.

Contenido:

- **1.1** Introducción a los depósitos de datos
- **1.2** Datos del mundo real: El reto administrativo

2 Planeación de los depósitos de datos

Objetivo: El alumno modelará la arquitectura de un depósito de datos para un caso de negocios.

Contenido:

- **2.1** La arquitectura de un depósito de datos
- **2.2** Un caso de negocios
- **2.3** Un enfoque para el desarrollo de un depósito de datos

3 Los datos

Objetivo: El alumno modelará datos bien definidos, integrados y consistentes, explicará el papel de los metadatos en el ciclo de vida de los depósitos de datos, el papel del directorio/catalogo en la empresa, el proceso de extraer datos de aplicaciones operacionales.

Contenido:

- **3.1** Calidad en los datos
- **3.2** Metadatos
- **3.3** El papel de un directorio/catalogo
- **3.4** Transformación de los datos

4 Diseño e implementación

Objetivo: El alumno diseñara e implementará un depósito de datos

Contenido:

- **4.1** Diseño físico
- **4.2** OLAP multidimensional y OLAP relacional
- **4.3** Replicación de datos en un ambiente de depósito de datos global
- **4.4** VLDS y paralelismo
- **4.5** Implementación de un depósito de datos en un ambiente multiservidor con tecnología paralela
- **4.6** El depósito de datos virtual
- **4.7** Minería de datos

DEPÓSITOS DE DATOS **(3 / 4)**

5 Administración de los depósitos de datos

Objetivo: El alumno explicara los aspectos de administración, seguridad de los depósitos de datos.

Contenido:

- **5.1** Administración de aplicaciones
- **5.2** Administración de bases de datos en un ambiente heterogéneo
- **5.3** Seguridad en los depósitos de datos
- **5.4** Selección de las herramientas para el usuario final

6 Tendencias

Objetivo: El alumno expondrá (verbalmente y/o mediante un ensayo) las tendencias de los depósitos de datos.

Contenido:

- **6.1** Tendencias en los depósitos de datos

Bibliografía básica: **Temas para los que se recomienda:**

BISCHOFF, JOYCE; ALEXANDE, TED — **Todos**
Data warehousing practical advice from experts
New Jersey
Prentice Hall, 1997

RALPH KIMBALL, MARGY ROSS — **Todos**
The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling
2a. edición
New Jersey
Wiley, 2002

Bibliografía complementaria:

Sugerencias didácticas:

Exposición oral	X	Lecturas obligatorias	X
Exposición audiovisual		Trabajos de investigación	X
Ejercicios dentro de clase	X	Prácticas de taller o laboratorio	X
Ejercicios fuera del aula		Prácticas de campo	
Seminarios		Otras	

DEPÓSITOS DE DATOS **(4 / 4)**

Forma de evaluar:

Exámenes parciales	X	Participación en clase	X
Exámenes finales	X	Asistencias a prácticas	X
Trabajos y tareas fuera del aula	X	Otras	

Perfil profesiográfico de quienes pueden impartir la asignatura

Perfil profesiográfico: Profesional con experiencia en depósitos de datos, de preferencia con un postgrado (maestría o doctorado) en el área.

Figura 1.2
Temario de la materia Depósito de Datos

1. 2 Análisis de los temas de la asignatura

En base a lo descrito en el tema anterior los temas de la materia depósito de datos se analizarán con el objeto de proporcionar valor agregado al plan de estudios y proporcionar al estudiante información relevante al respecto.

Para el análisis de los temas utilizaremos una metodología que consiste en la comparación de los temas que integran el temario de la materia Depósito de Datos con temarios similares de otras instituciones y publicaciones.

1.2.1 Metodología de Análisis.

La metodología de análisis consiste en investigar y revisar temarios de diferentes universidades, tanto dentro como fuera del país, así como libros especializados en la materia y cursos o diplomados al respecto.

Para el desarrollo de este análisis, se tomaron como base los temarios de las instituciones siguientes:

Universidades Nacionales:

- Universidad Iberoamericana,
- Universidad Anáhuac
- Instituto Tecnológico de Puebla
- Instituto Tecnológico y de Estudios Superiores de Monterrey
- Instituto Tecnológico Autónomo de México (Sistemas para las decisiones)

Universidades Extranjeras:

- Universidad Europea de Madrid

- Universidad de Pennsylvania (OPIM 672/Decision Support Systems)
- Universidad de Calgary

Cursos y Seminarios

- National University of Singapore
    - Data Warehouse Design
- CESINE. Escuela Superior de Informática y Negocios (España)
    - Data warehouse: Integración, Tratamiento Y Presentación De La Información
- Universidad Politécnica De Madrid
    - Data Warehouse : Diseño y Utilización

Libros Especializados

- Building The Data Warehouse - William H. Inmon - Wiley
- Data Warehouse For Dummies-Alan R. Simon -Idg Books Worldwide Inc.
- Data Warehouse Design Solutions -Christopher Adamson, Michael Venerable-Wiley
- Data Warehouse Fundamentals A Comprehensive Guide For IT Professionals-Paulraj Ponniah-Wiley
- The Data Warehouse Lifecycle Toolkit : Expert Methods For Designing, Developing, And Deploying Data Warehouses-Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite-Wiley

Esquemáticamente, en la figura 1.3 se muestran los procesos realizados en la metodología de análisis.

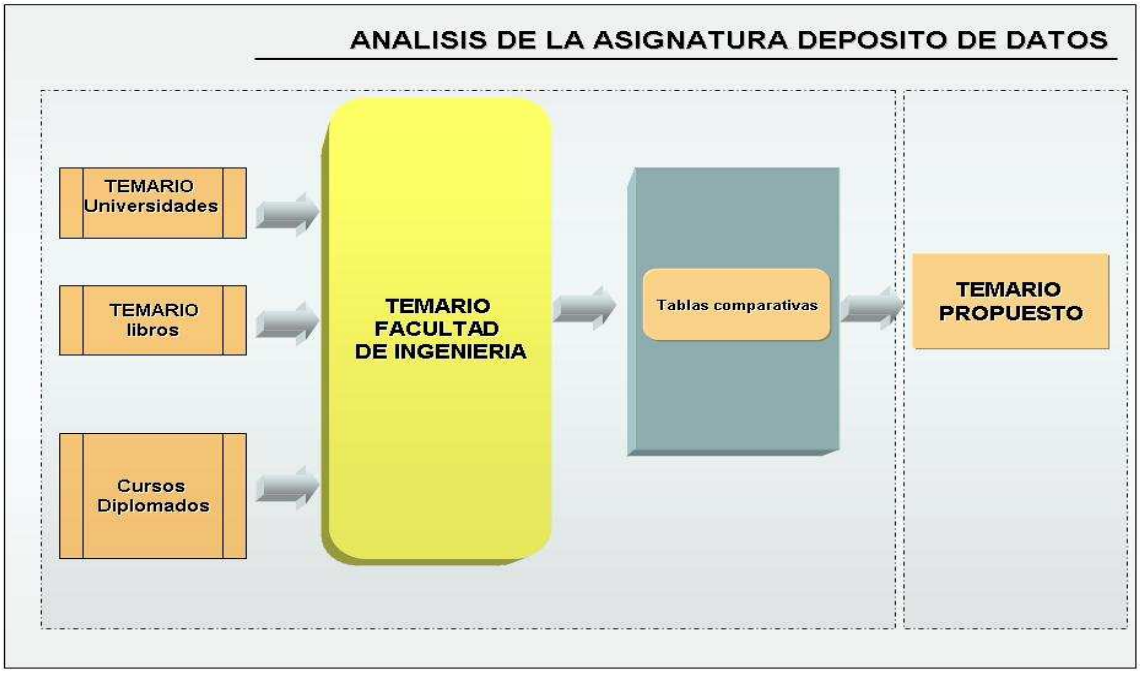

Figura 1. 1 Procesos de la Metodología de Análisis

Una vez revisada la información de los temarios descritos, se realiza la generación de tablas comparativas contra los temas de la materia depósito de datos de la Facultad de Ingeniería, con el objeto de visualizar el nivel de actualización tecnológica en el ámbito de los sistemas para la toma de decisiones entre las instituciones y publicaciones revisadas.

En las figura 1.4 se muestra el comparativo del temario de la Facultad de Ingeniería con universidades nacionales.

TEMARIOS DEPOSITOS DE DATOS	UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO	UNIVERSIDAD IBEROAMERICANA	INSTITUTO TECNOLOGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY *******	INSTITUTO TECNOLOGICO AUTONOMO DE MEXICO *******	UNIVERSIDAD ANAHUAC *******	INSTITUTO TECNOLOGICO DE PUEBLA
INTRODUCCION A LOS DEPOSITOS DE DATOS	✓	✓				✓
ARQUITECTURA DE LOS DEPOSITOS DE DATOS	✓	✓				✓
LOS DATOS	✓					
DISEÑO E IMPLEMENTACION DE UN DEPOSITO DE DATOS	✓	✓				✓
ADMINISTRACION DE LOS DEPOSITOS DE DATOS	✓					
TENDENCIAS DE LOS DEPOSITOS DE DATOS	✓					
MINERIA DE DATOS	✓					✓
PROGRAMACION DE SISTEMAS DATA WAREHOUSE		✓				
SISTEMAS INTELIGENTES		✓				

******* *Nota: En estas universidades, se imparte la materia, pero no se cuenta con información del temario*

Figura 1.4
Comparativo con Universidades Nacionales

De la figura 1.4, se observa que el temario de la UNAM tiene ventajas respecto al de la UIA, ya que contempla los temas de datos, administración y tendencias de los depósitos de datos. Así mismo, en la UIA incluyen los temas de programación de sistemas data warehouse y sistemas inteligentes, que no se tienen en la Facultad de Ingeniería. El Tecnológico de Puebla solo incluye 4 de los 7 temas que se incluyen en el temario de la UNAM.

En el desarrollo del análisis se detectó que en muchas universidades nacionales de prestigio, no se incluye la materia de depósitos de datos o alguna equivalente en sus planes de estudio.

En las figura 1.5 se muestra el comparativo del temario de la Facultad de Ingeniería con universidades extranjeras.

TEMARIOS DEPOSITOS DE DATOS	UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO	UNIVERSIDAD EUROPEA DE MADRID	UNIVERSIDAD DE PENNSYLVANIA *******	UNIVERSIDAD DE CALGARY
INTRODUCCION A LOS DEPOSITOS DE DATOS	✓	✓		✓
ARQUITECTURA DE LOS DEPOSITOS DE DATOS	✓	✓		✓
LOS DATOS	✓	✓		✓
DISEÑO E IMPLEMENTACION DE UN DEPOSITO DE DATOS	✓	✓		✓
ADMINISTRACION DE LOS DEPOSITOS DE DATOS	✓			
TENDENCIAS DE LOS DEPOSITOS DE DATOS	✓			
MINERIA DE DATOS	✓			
PROGRAMACION DE SISTEMAS DATA WAREHOUSE				
SISTEMAS INTELIGENTES				

******* *Nota: En estas universidades, se imparte la materia, pero no se cuenta con información del temario*

Figura 1.5
Comparativo con Universidades Extranjeras

De la figura 1.5, podemos interpretar que el temario de la Universidad Europea de Madrid y la Universidad de Calgary, tiene solamente los cuatro temas iniciales respecto al de la UNAM y no contempla la administración, tendencias y minería de datos.

En las figura 1.6 se muestra el comparativo del temario de la Facultad de Ingeniería con cursos.

TEMARIO DEPOSITOS DE DATOS	UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO	ESCUELA SUPERIOR DE INFORMÁTICA Y NEGOCIOS (ESPAÑA)	UNIVERSIDAD POLITECNICA DE MADRID	NATIONAL UNIVERSITY OF SINGAPORE
CURSO ➤		DATA WAREHOUSE: INTEGRACIÓN, TRATAMIENTO Y PRESENTACIÓN DE LA INFORMACIÓN	DATA WAREHOUSE : DISEÑO Y UTILIZACIÓN	DATA WAREHOUSE DESIGN
INTRODUCCION A LOS DEPOSITOS DE DATOS	✓	✓		✓
ARQUITECTURA DE LOS DEPOSITOS DE DATOS	✓	✓	✓	✓
LOS DATOS	✓		✓	✓
DISEÑO E IMPLEMENTACION DE UN DEPOSITO DE DATOS	✓	✓	✓	✓
ADMINISTRACION DE LOS DEPOSITOS DE DATOS	✓	✓		✓
TENDENCIAS DE LOS DEPOSITOS DE DATOS	✓	✓		
MINERIA DE DATOS	✓	✓		
PROGRAMACION DE SISTEMAS DATA WAREHOUSE				
SISTEMAS INTELIGENTES				

Figura 1.6
Comparativo con Cursos

Observamos que te los cursos de la figura 1.6, solo uno de ellos contempla prácticamente los mismos temas que el temario de la UNAM, mientras que los otros solo consideran algunos de ellos. La universidad de Singapur solo deja de lado los temas de tendencias y minería.

En las figura 1.7 se muestra el comparativo del temario de la Facultad de Ingeniería con publicaciones especializadas en la materia.

TEMARIO DEPOSITOS DE DATOS	UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO	BUILDING THE DATA WAREHOUSE WILLIAM H. INMON - WILEY	DATA WAREHOUSE FOR DUMMIES ALAN R. SIMON - IDG BOOKS WORLDWIDE	DATA WAREHOUSE DESIGN SOLUTIONS CHRISTOPHER ADAMSON, MICHAEL VENERABLE - WILEY	DATA WAREHOUSE FUNDAMENTALS (A COMPREHENSIVE GUIDE FOR IT PROFESSIONALS) PAULRAJ PONNIAH - WILEY	THE DATA WAREHOUSE LIFECYCLE TOOLKIT : EXPERT METHODS FOR DESIGNING, DEVELOPING, AND DEPLOYING DATA WAREHOUSES - RALPH KIMBALL, LAURA REEVES, MARGY ROSS, WARREN THORNTHWAITE - WILEY
INTRODUCCION A LOS DEPOSITOS DE DATOS	✓		✓	✓		✓
ARQUITECTURA DE LOS DEPOSITOS DE DATOS	✓	✓	✓		✓	✓
LOS DATOS	✓		✓		✓	
DISEÑO E IMPLEMENTACION DE UN DEPOSITO DE DATOS	✓	✓	✓			✓
ADMINISTRACION DE LOS DEPOSITOS DE DATOS	✓					✓
TENDENCIAS DE LOS DEPOSITOS DE DATOS	✓				✓	
MINERIA DE DATOS	✓					
PROGRAMACION DE SISTEMAS DATA WAREHOUSE						
SISTEMAS INTELIGENTES						

Figura 1.7
Comparativo con Libros Especializados

Los libros que se consideraron para este análisis, son especializados en la materia, los cuales tocan desde diferentes puntos de vista los temas incluidos en la Facultad de Ingeniería. Algunos enfatizan en el diseño y arquitectura de un depósito de datos, mientras que otros detallan de mejor forma la parte de implementación y tendencias.

Algunos se enfocan solamente en los conceptos básicos del diseño o bien desglosan de manera genérica pero muy superficial en más de la mitad de los temas de la asignatura.

Se puede deducir que el temario de la Facultad de Ingeniería de la UNAM, contempla los aspectos más importantes que se requieren para el adecuado entendimiento de los depósitos de datos, lo cual queda claramente visible en las figuras anteriores, por lo que podemos concluir que la Facultad de Ingeniería está a la vanguardia en sus planes de estudio con respecto a instituciones de educación superior dentro y fuera del país.

Así mismo denota la importancia y relevancia que tienen los depósitos de datos en diversas universidades, ya que encontramos que en una gran cantidad de universidades investigadas, no contemplan la asignatura dentro de sus planes de estudio.

1. 3 Sugerencias y comentarios

En base al análisis realizado para comparar el temario de la materia depósitos de datos de la Facultad de Ingeniería de la UNAM con respecto a otras universidades, cursos y libros, podemos sugerir lo siguiente:

- El tema de tendencias solo se contempla en el libro Data Warehouse Fundamentals de Paulraj Ponniah Editorial Wiley y en el curso Data Warehouse: integración, tratamiento y presentación de la información, de la Escuela Superior de Informática y Negocios de España, por lo que consideramos que se disminuya ese tema de 8 a 4 horas y se dediquen al tema de Diseño e Implementación.
- En el caso de que se acepte la sugerencia anterior, podríamos incluir algunas metodologías para enriquecer el capitulo de diseño e implementación, como por ejemplo la CRISP-DM o la Information Factory Framework.

CAPÍTULO 2

DESARROLLO DEL MATERIAL DIDÁCTICO

2.1 Introducción a los Depósitos de datos

2.1.1 Antecedentes

Los sistemas informáticos operacionales han sido de gran ayuda a las empresas, estos sistemas han automatizado los procesos de carácter típicamente repetitivo o administrativo.

Sin embargo un buen sistema operacional es de gran ayuda en una empresa debido a que los conceptos más importantes para estos sistemas son la actualización y el tiempo de respuesta en la información, siendo la información correcta y oportuna la que resuelve las necesidades de funcionamiento, administración y control de una organización.

Hoy en día toda organización necesita depositar mucha confianza en la toma de decisiones tanto a escala estratégica como táctica, sabemos que la competencia crece en todo momento entonces las decisiones que se deben tomar en cualquier organización deben ser acertadas. Esta idea esta presente en la mayoría de los ejecutores de decisiones pero se topan con grandes cantidades de información la cual debe ser analizada y procesada para obtener información útil, esta información es costosa debido a la gran cantidad de tiempo y recursos que se tiene que invertir y que muchas veces no están disponibles en la organización.

Por este motivo se requieren herramientas que ayuden a minimizar el tiempo para analizar grandes volúmenes de información con mayor velocidad y precisión, con estas herramientas se logra mantener la competitividad ya que las organizaciones pueden reaccionar a los cambios del mercado. De otro modo el mercado globalizado, la enorme competencia y los avances tecnológicos debilitan cualquier organización.

Una herramienta que proporciona una solución al tratamiento y manejo, de grandes volúmenes de información son los depósitos de datos.

Un depósito de datos se crea al extraer información de una o más bases de datos de aplicaciones operacionales. Los datos extraídos son transformados para eliminar inconsistencias y resumirlos si es necesario, una vez tratada es cargada al repositorio de datos.

Para obtener el máximo beneficio de los depósitos de datos se deben considerar las estrategias tecnológicas necesarias para la implementación de una arquitectura completa de un depósito de datos.

Tipos de sistemas de información en una organización

En las organizaciones existen varios tipos de sistemas los cuales cumplen funciones específicas que ayudan al adecuado funcionamiento de la organización, dentro de los más relevantes se encuentran los sistemas operacionales y los depósitos de datos.

Los sistemas operacionales como su nombre lo indica se utilizan para manejar las operaciones cotidianas de la organización como son sistemas de contabilidad, inventario, fabricación, recursos humanos entre otros. Estos sistemas son fundamentales en la organización por lo que se han revisado, mejorado y mantenido, y hoy en día están completamente integrados.

Por otra parte, hay otras funciones dentro de la empresa que tienen que ver con la planeación, previsión y administración de la información por ejemplo "planeación de marketing", "planeación de producción" y "análisis financiero", requieren, de sistemas que estén relacionados con el análisis de los datos y la toma de decisiones importantes sobre cómo debe operar la empresa ahora y en el futuro, esta funciones son cubiertas por los depósitos de datos.

Los depósitos de datos no sólo tienen un enfoque diferente al de los sistemas operacionales, sino que, por lo general, tienen un alcance diferente. Mientras las necesidades de los datos operacionales se enfocan normalmente hacia una sola área, los datos para la toma de decisiones, con frecuencia, abarcan diferentes áreas y utiliza grandes cantidades de datos operacionales relacionados.

Definición de un Depósito de Datos

Existen varias definiciones de lo que es un Depósito de Datos pero a continuación se citará la más conocida:

“El Depósito de Datos es un conjunto de datos integrados orientados a un tema cuyos datos son históricos, no transitorios y organizados para apoyar al proceso de toma de decisiones de una administración.”[1]

Las características mencionadas en la definición de depósito de datos son:

- **Orientado a un tema.**

Los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno de sistemas operacionales. Estos datos se organizan por materia para facilitar su acceso y entendimiento a los usuarios finales. Es decir en el Depósito de Datos solo se incluye información de los temas del negocio que serán utilizadas en el proceso de soporte de toma de decisiones. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del Depósito de Datos. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.

- **Integrado.**

[1] Definición de W. H. Inmon de su libro Using the Data Warehouse

Los datos almacenados en el Depósito de Datos deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas.

La información debe estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.

- **Históricos.**

El tiempo es parte fundamental de la información contenida en un Depósito de Datos, esta información representa los datos sobre un horizonte de tiempo.

La información almacenada en el Depósito de Datos sirve, entre otras cosas, para realizar análisis de tendencias.

El Depósito de Datos se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones y muestra una perspectiva histórica de dicha variable a través del tiempo.

- **No transitorios.**

La información es útil sólo cuando es estable. La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere un depósito de datos no volátil.

El almacén de información de un Depósito de Datos existe para ser leído, y no modificado. La información es por tanto permanente, es decir la actualización del Depósito de Datos al incorporar los últimos valores que tomaron las distintas variables contenidas de este depósito de datos no debe modificar la información que ya se encontraba en este almacén de datos.

En la tabla 2.1.1 se muestra un comparativo entre los sistemas operacionales y los depósitos de datos:

Sistemas operacionales	Depósito de Datos
Predomina la actualización	Predomina la consulta
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

Datos en general no agrupados	Datos agrupados
Importancia del dato actual	Importancia del dato histórico
Importante del tiempo de respuesta de la transacción instantánea	Importancia de la respuesta masiva
Estructura relacional	Visión multidimensional
Usuarios de perfiles medios o bajos	Usuarios de perfiles altos
Explotación de la información relacionada con la operativa de cada aplicación	Explotación de toda la información interna y externa relacionada con el negocio

Tabla 2.1.1
Tabla comparativa entre un sistema operacional y un depósito de datos.

Aplicación de las Depósito de Datos

Las necesidades empresariales de una tecnología de depósito de datos debe de estar bien articulada, el depósito de datos debe tener el potencial para proporcionar un servicio de tecnología de la información estratégicamente importante.

La información de la empresa es un activo y quienes trabajan en una empresa necesitan un acceso fácil y rápido a la información correcta para conocer a sus clientes, acrecentar sus ingresos e incrementar su participación en el mercado.

El desarrollo del potencial de un depósito de datos sólo será limitado por la capacidad, habilidad y creatividad de los usuarios empresariales. En la Figura 2.1.1 muestra los múltiples usos de los depósitos de datos en la industria.

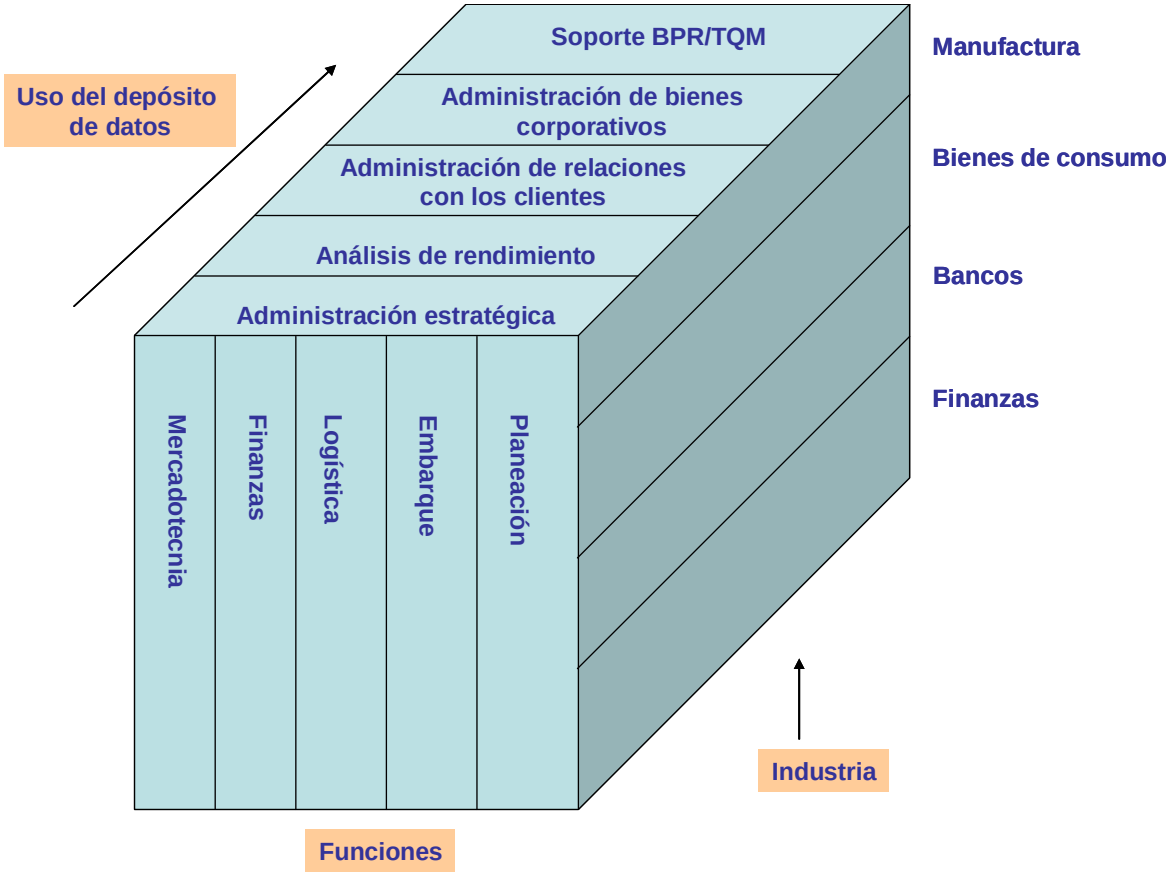

Figura 2.1.1
Aplicación de los Depósito de Datos

Los depósitos de datos se están aplicando con éxito en la manufactura, los bienes de consumo, la distribución, en el sector financiero y bancario. La tendencia es contar con depósitos de datos en las empresas para incrementar las ganancias y vencer a la competencia.

Los datos de la empresa en el depósito de datos son un activo importante. Estos datos son una historia detallada de los negocios de la empresa y su relación con sus clientes.

Procesos de un Depósito de Datos

Para comprender el concepto de Depósito de Datos, es importante considerar los procesos que lo conforman.

A continuación se describen estos procesos que son clave en la gestión de un Depósito de Datos.

Extracción
Se necesita la extracción y obtención de información de las distintas fuentes tanto internas como externas para conformar el depósito de datos.

Transformación.
En este proceso se realiza el filtrado, limpieza, depuración, homogeneización y agrupación de la información.

Carga.
En este proceso se realiza la organización y actualización de los datos y los metadatos en el repositorio de datos.

Explotación.
En este proceso se realiza la extracción y análisis de la información en los distintos niveles de agrupación, es aquí donde el usuario final puede explotar el depósito de datos y obtener información que le ayude en la toma de decisiones para el beneficio de la organización.

Desde el punto de vista del usuario, el único proceso visible es la explotación del almacén de datos, aunque el éxito del Depósito de Datos radica en los tres procesos iniciales que alimentan la información del mismo y suponen el mayor porcentaje de esfuerzo (en torno a un 80%) a la hora de desarrollar el depósito de datos.

Beneficios de utilizar un Depósito de Datos

Existen muchos beneficios al utilizar un Depósito de Datos los cuales se enuncian a continuación:

- ✓ Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.

- ✓ Facilita la aplicación de técnicas estadísticas de análisis para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor agregado para el negocio en cuanto a información.

- ✓ Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.

- ✓ Simplifica dentro de la empresa la implantación de sistemas de gestión integral de la relación con el cliente.
- ✓ Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes con rápidos retornos de la inversión.

2.1.2 Datos del mundo real

Un parámetro práctico y contundente es la realidad, en este mundo podemos medir que tan aterrizado es el analizar diseñar y construir un Depósito de Datos.

Impacto Empresarial al utilizar un Depósito de Datos

Son varios los impactos que una organización experimenta al utilizar un depósito de datos dentro de los más importantes podemos listar los siguientes:

Los Procesos de Toma de Decisiones pueden ser mejorados mediante la disponibilidad de información. Decisiones empresariales se hacen más rápidas por gente más informada.

Los procesos empresariales pueden ser optimizados. Al reducir el tiempo en la búsqueda y tratamiento de información.

Conexiones y dependencias entre procesos empresariales se vuelven más claras. Secuencias de procesos empresariales pueden ser optimizadas para ganar eficiencia y reducir costos.

Procesos y datos de los sistemas operacionales, así como los datos en el Depósito de Datos, son usados y examinados. Cuando los datos son organizados y estructurados para tener significado empresarial, la gente aprende mucho de los sistemas de información. Pueden quedar expuestos posibles defectos en aplicaciones actuales, siendo posible entonces mejorar la calidad de nuevas aplicaciones.

Las organizaciones empresariales y la gente de la cual ella se compone queda determinada por el acceso a la información. De esta manera, la gente queda mejor habilitada para entender su propio rol y responsabilidades como también los efectos de sus contribuciones; a la vez, desarrollan un mejor entendimiento y apreciación con las contribuciones de otros.

Visibilidad, accesibilidad, y conocimiento de los datos producen mayor confianza en la operación de los sistemas transaccionales.

Dadas las características de un sistema de Depósito de Datos, su aplicación puede tener varios fines, en una diversidad de organizaciones. No obstante, en términos generales, podemos decir que su aplicación más rica corresponde a entornos de empresas en los que se identifican grandes volúmenes de datos, asociados a: cantidad de clientes, variedad de productos y cantidad de transacciones.

Casos de Depósito de Datos en el mundo real

Se mencionan algunos ejemplos de aplicaciones típicas y casos puntuales en distintas organizaciones.

Comercio Minorista

Utilizan grandes sistemas de Procesamiento Paralelo Masivo para acceder a meses o años de historia transaccional tomada directamente en los puntos de venta de cientos, o miles, de sucursales. Con esta información detallada pueden efectuar en forma más precisa y eficiente actividades de compra, fijación de precios y/o manejo de inventarios.

Las promociones y ofertas de cada comercio minorista son seguidas, analizadas y corregidas. Se administran las tendencias a efectos de maximizar utilidades y reducir costos de inventario. La existencia de artículos es asignado por sucursales o regiones según ventas y tendencias. Estos sistemas con capacidad de procesar gran cantidad de datos detallados permiten implementar eficientemente prácticas de venta a consignación, en esta modalidad la cadena minorista paga al proveedor hasta que los productos son vendidos y pasados por el lector de códigos de barras (scanner) del punto de venta.

Esta información detallada permite ejercer mayor poder de negociación sobre los proveedores, dado que el comercio minorista puede llegar a saber más que el fabricante sobre sus productos: quién lo compra, dónde, cuándo, con que otros productos, etc.

Un ejemplo practico de un deposito de datos es Wal*Mart, que es considerada como un ejemplo del manejo de un depósito de datos.

El fundador de Wal*Mart, Sam Walton, escribe: "me dicen que es el depósito de datos comercial más grande del mundo. Lo que me gusta es la clase de información que puedo obtener de este al instante ¡todos esos números!, llevamos 65 semanas de historia de cada artículo que vendemos. Esto significa que puedo elegir cualquiera y decir exactamente cuantos vendimos... no en promedio, sino en cualquier región, distrito o sucursal. Es difícil que un proveedor sepa más acerca de su producto de lo que sabemos nosotros. Nos da el poder de la ventaja competitiva." [2].

Se debe considerar que las sucursales de las que se hace referencia son aproximadamente 2500 y que cada una de ellas tiene una variedad de entre 50.000 y 80.000 artículos, Toda esta información es concentrada al final del día y se realizan 20 millones de actualizaciones al repositorio de datos lo que muestra un claro ejemplo del manejo de un depósito de datos.

Manufactura de Bienes de Consumo Masivo

Las empresas de este sector necesitan hacer un manejo cada vez más ágil de la información para mantenerse competitivas en la industria. Los Depósito de Datos se utilizan para predecir la cantidad de producto que se venderá a un determinado precio y, por consiguiente, producir la cantidad adecuada para una entrega "justo a tiempo". A su vez se coordina el suministro a las grandes cadenas minoristas con inmensas cantidades de productos "en consignación", que no son pagados hasta que estos productos son vendidos al consumidor final.

[2] De acuerdo al libro “Made in america: My story” escrito por Sam Walton

Las cadenas minoristas y sus proveedores utilizan su Depósito de Datos para compartir información, permitiéndole a las empresas de manufactura conocer el nivel de existencia en los almacenes y eventualmente hacerse responsables de la reposición de inventario de la cadena minorista. Como es de esperar esto reduce fuertemente la intermediación. También se utilizan para campañas de marketing, planificación de publicidad y promociones y se coordinan las ofertas de cupones y promociones con las cadenas minoristas.

Un ejemplo interesante es el de Whirlpool. Este fabricante global de electrodomésticos con base en Benton Harbor, Michigan, utiliza su Depósito de Datos para hacer un seguimiento directo de sus casi 15 millones de clientes y de sus más de 20 millones de aparatos instalados. Las mayores aplicaciones del sistema son para marketing, ventas, mantenimiento, garantía y diseño de productos. Permite mantener stock de partes más ajustados y mejorar las condiciones de negociación con los proveedores de las mismas. Si, por ejemplo, un determinado motor se identifica como poseedor de una tasa de falla superior, Whirlpool puede utilizar la información para hacer renegociaciones de garantía con el proveedor.

Transporte de Cargas y Pasajeros

Se utilizan Depósito de Datos para almacenar y acceder a meses o años de datos de clientes y sistemas de reservas para realizar actividades de marketing, planeamiento de capacidad, monitoreo de ganancias, proyecciones y análisis de ventas y costos, programas de calidad y servicio a clientes.

Las empresas de transporte de cargas llevan datos históricos de años, de millones de cargamentos, capacidades, tiempos de entrega, costos, ventas, márgenes, equipamiento.

Las aerolíneas utilizan su Depósito de Datos para sus programas de viajeros frecuentes, para compartir información con los fabricantes de naves, para la administración del transporte de cargas, para compras y administración de inventarios, etc. Hacen un seguimiento de partes de repuesto, cumplimiento con las regulaciones aeronáuticas, desempeño de los proveedores, seguimiento de equipaje, historia de reservas, ventas y devoluciones de boletos, reservas telefónicas, desempeño de las agencias de viajes, estadísticas de vuelo, contratos de mantenimiento.

Telecomunicaciones

Estas empresas utilizan su Depósito de Datos para operar en un mercado crecientemente competitivo y global que, a su vez, atraviesa profundos cambios tecnológicos. Se almacenan datos de millones de clientes: sus circuitos, facturas mensuales, volúmenes de llamados, servicios utilizados, equipamiento vendido, configuraciones de redes, etc. así como también información de facturación, utilidades, y costos son utilizadas con propósitos de marketing, contabilidad, reportes gubernamentales, inventarios, compras y administración de redes.

2.2 Planeación de los depósitos de datos

Objetivo: El alumno modelará la arquitectura de un depósito de datos para un caso de negocios.

Se pretende que el estudiante conozca y utilice tecnologías de bases de datos para el desarrollo de aplicaciones relacionadas con el tratamiento de Información y soporte para la toma de decisiones. El alumno cubrirá todos los aspectos de la planeación, diseño, desarrollo, implementación y administración de los depósitos de datos.

2.2.1 La Arquitectura de un Depósito de Datos.

La información en un ambiente de Depósito de Datos, permite a cualquier organización hacer un uso óptimo de la misma, como parte clave para un proceso de toma de decisiones más efectivo, por lo que deben aprovechar sus recursos de información, pero deben considerarse las estrategias tecnológicas necesarias para la implementación de una arquitectura completa de Depósito de Datos.

La arquitectura de un depósito de datos se compone de:

- Origen de datos.
- Procesos de extracción, transformación y carga.
- Repositorio de datos
- Herramientas para el procesamiento analítico en línea (OLAP).

Para que un usuario final obtenga información que le sea útil es necesario contar con herramientas de análisis de información y estas a su vez requieren de un repositorio de datos, el cual contenga la información de las fuentes de datos debidamente procesada y transformada que permita el análisis de la misma, como se muestra en la figura 2.2.1.

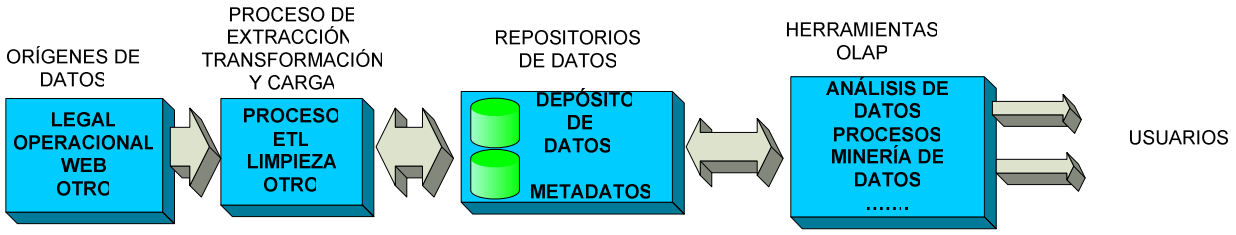

Figura 2.2.1
Arquitectura conceptual de un Depósito de Datos

Origen de datos

Se considera como orígen de datos a toda aquella información que posee una organización, y puede estar contenida en cualquier medio electrónico. Los datos administrados por los sistemas de aplicación operacionales son la principal fuente de información para el Depósito

de Datos, y estos a su vez pueden ser parte del origen de datos para otro Depósito de datos como se muestra en la figura 2.2.2.

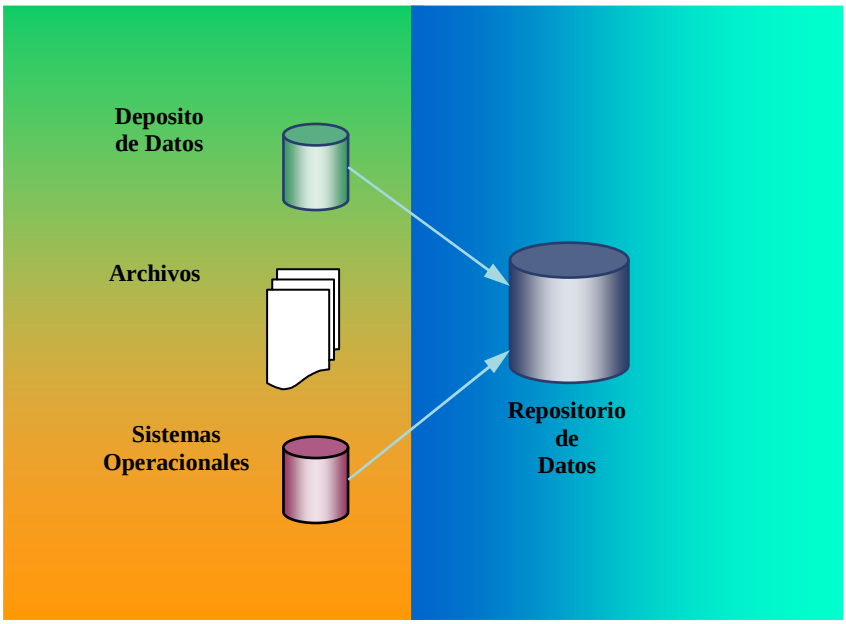

Figura 2.2.2
Origen de Datos

Procesos de extracción, transformación y carga.

El ingreso de datos en el Depósito de datos viene desde el ambiente operacional en casi todos los casos. El Depósito de datos es siempre un almacén de datos transformados y separados físicamente de la aplicación donde se encontraron los datos en el ambiente operacional, pero es necesario realizar el proceso ETL.

Los procesos de extracción, transformación y carga (ETL) involucran:

- Extracción de Información de los orígenes de datos
- Transformación de la información para adecuarla a las necesidades de la organización.
- Carga dentro del repositorio de datos

Extracción

La primera parte del proceso ETL consiste en la Extracción de los datos desde los sistemas origen. La mayoría de los proyectos de los depósitos de datos consolidan datos desde diferentes fuentes. Cada sistema puede organizar los datos de diferente forma. Los formatos más comúnmente utilizados son bases de datos relacionales y archivos planos que pueden

incluir bases de datos no relacionales u otras estructuras de datos. La extracción convierte los datos a un formato adecuado para el proceso de transformación.

Transformación

La fase de transformación aplica una serie de reglas o funciones a los datos extraídos que derivan en datos listos para ser cargados en el repositorio de datos. Algunos de los orígenes de datos requieren una pequeña manipulación de sus datos. Sin embargo en otros casos, alguna combinación de los siguientes tipos de transformaciones será requerida:

- Selección de ciertas columnas a cargar. Es decir, las columnas en nulo no se procesan.
- Traducción de valores codificados por ejemplo, si el origen de datos almacena una M para masculino y F para femenino, el depósito de datos almacena 1 para masculino y 2 para femenino.
- Codificación de valores de texto libre
- Calculo de nuevos valores
- Unión de datos de múltiples orígenes
- Resumen de múltiples registros de datos
- Transponer registros (Convertir filas en columnas o viceversa)

Carga

Es el proceso encargado de insertar los registros en el repositorio de datos, dependiendo de los requerimientos de la organización, este proceso puede ser tan extenso o tan corto como se necesite.

Los procesos de ETL pueden ser muy complejos y pueden acarrear problemas operacionales significativos debido a un diseño inapropiado del sistema ETL[3].

El rango de valores o la calidad de los datos de un sistema operacional pueden estar fuera de las expectativas de los diseñadores en el momento en que las reglas de validación y transformación son especificadas. Es recomendado la creación de perfiles de datos de un origen durante el análisis de los mismos para identificar las condiciones de los datos que serán manipulados por las reglas de transformación.

[3] Fuentes:
http://en.wikipedia.org/wiki/Extract%2C_transform%2C_load
Manual para la construcción de un Data Warehause. Publicado por el instituto Nacional de Estadística e Informática INEI/1997, Lima. Revisado y editado por el profesor J. Elliott.
DATA WAREHAUSING, La integración para la mejor toma de decisiones. Harjinder S. Gill y Prakash C. Rao. Ed. Prentice Hall Hispanoamericana, S.A. Pags. 45-73 . Que contiene un software de herramienta para los Depósitos de Datos de Indica, llamado el PLANNER (Diseñador).

La escalabilidad de un sistema ETL durante su periodo de vida necesita ser considerado durante su análisis. Esto incluye el conocer el volumen de datos que serán procesados dentro de un periodo de tiempo, Es importante entender si el volumen de datos aumenta el tiempo de proceso requerido también pudiera aumentar.

Una dificultad adicional está en asegurarse que los datos cargados sean consistentes. Debido a que los múltiples orígenes de datos pueden tener diferentes ciclos de actualización, un sistema ETL puede requerir almacenar ciertos datos hasta que todos los orígenes estén sincronizados.

Una buen ETL puede ser capaz de comunicarse con las diferentes bases de datos relacionales y leer varios formatos de archivo utilizados dentro de la organización.

La figura 2.2.3 muestra el proceso gráfico del ETL por el cual los datos pasan para ser almacenados en el repositorio de datos.

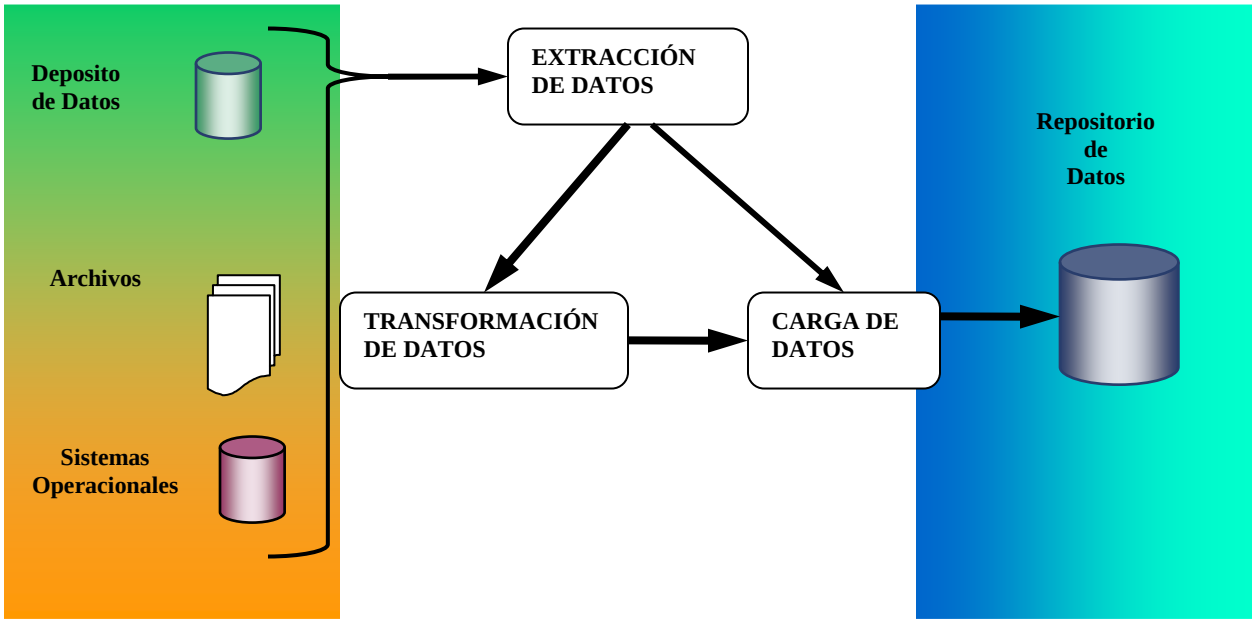

Figura 2.2.3
Proceso ETL

La transformación de datos también se encarga de las inconsistencias en el contenido de datos. Una vez que se toma la decisión sobre que reglas de transformación serán establecidas, debe crearse e incluirse las definiciones en las rutinas de transformación.

Se requieren herramientas de gestión de datos para extraer datos desde bases de datos y/o archivos operacionales, luego es necesario manipular o transformar los datos antes de cargar los resultados en el depósito de datos.

Repositorio de datos

Es el almacenamiento físico de datos de la arquitectura Depósito de Datos.
El origen de datos puede ser heterogéneo y autónomo, provee datos para el depósito de datos. Los datos de interés son capturados, almacenados y actualizados en los depósitos por el administrador, procesado de forma multidimensional, supliendo a los usuarios por analizadores. Todos los componentes actualizan e integran los depósitos de datos, sirven a los usuarios con funciones de análisis de datos y ejecución de tareas de soporte de decisión.

Después del proceso ETL para poblar el repositorio de almacenamiento del Depósito de Datos, también llamado repositorio de datos, otro paso necesario es crear el metadato, mostrado en la figura 2.2.4.

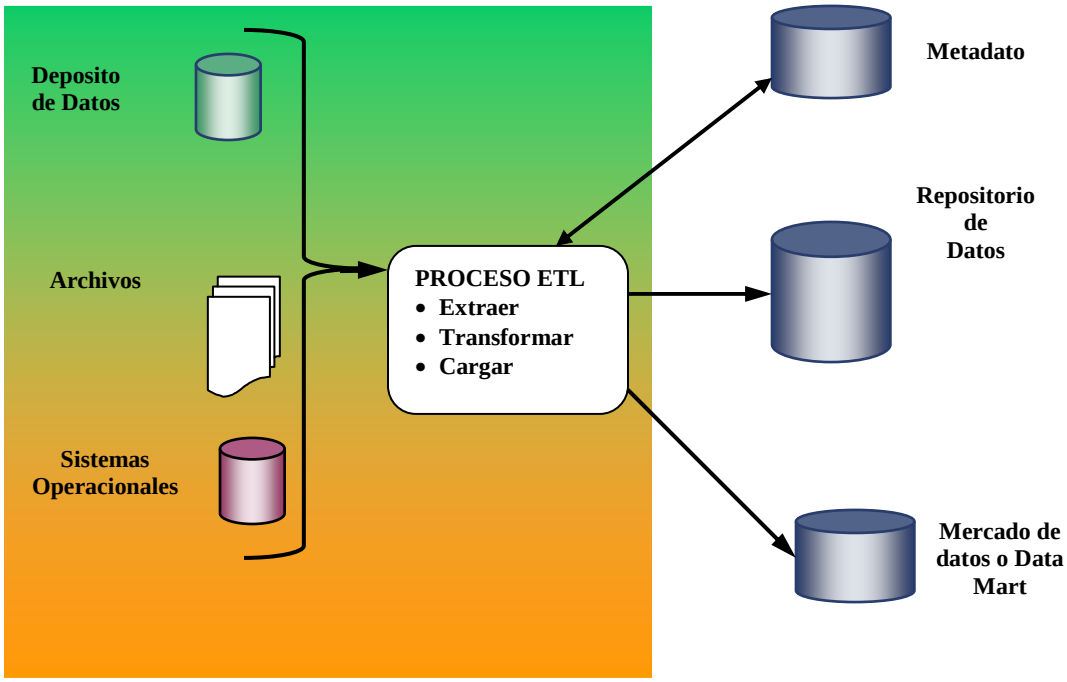

Figura 2.2.4
Repositorio de Datos

Metadato[4]

Otro aspecto de la arquitectura de los Depósito de datos es crear soporte al metadato. Metadato es la información sobre los datos que se alimenta, se transforma y existe en el Depósito de datos, es decir, datos acerca de datos que describen los contenidos del

[4] Combinación de base de datos mediante herramientas informáticas OLAP – ERP en Costos.

Revista Electrónica FCE. Universidad Católica del Uruguay

Depósito de Datos. La metadato consiste de definiciones de los elementos de datos en el depósito, sistema(s) del (os) elemento(s) fuente. Como los datos se integra y transforma antes de ser almacenada en información similar.

Los Metadato se encuentran en un repositorio separado de los depósitos de datos, crea una relación entre los usuarios, los orígenes de datos y los depósitos de datos. Las bases de metadatos almacenan información para la transformación, integración, almacenaje y uso de los depósitos de datos.

Metadato es un concepto genérico, pero cada implementación de la metadato usa técnicas y métodos específicos.

Estos métodos y técnicas son dependientes de los requerimientos de cada organización, de las capacidades existentes y de los requerimientos de interfaces de usuario. Hasta ahora no hay normas para el Metadato, por lo que debe definirse desde el punto de vista del software para el Depósito de datos seleccionado para una implementación específica.

Típicamente, el Metadato incluye los siguientes puntos:

• Las estructuras de datos que dan una visión de los datos al administrador de datos.

• Las definiciones del sistema de registro desde el cual se construye el Depósito de datos.

• Las especificaciones de transformaciones de datos que ocurren tal como la fuente de datos se replica al Depósito de datos.

• El modelo de datos del Depósito de datos (es decir, los elementos de datos y sus relaciones).

• Un registro de cuando los nuevos elementos de datos se agregan al Depósito de datos y cuando los elementos de datos antiguos se eliminan o se resumen.

• Los niveles de sumarización, el método de sumarización y las tablas de registros de su Depósito de datos.

Data Mart o Mercado de Datos

Es un subgrupo lógico del Depósito de Datos, es decir, es una base de datos o colección de bases de datos, pero con contenidos específicos, volumen de datos más limitado y un alcance histórico menor, que pueda proporcionar los datos para divulgar, análisis y soporte en una sección, una unidad, un departamento, una operación en la compañía o área de negocio de una empresa grande. Mientras que un Depósito de datos combina bases de datos a través de una empresa entera, los Data Mart son los Depósitos de datos a veces completos de los datos que son generalmente más pequeños que el Depósito de datos corporativo.

Algunos Data Mart Dependientes de los datos, son subconjuntos de Depósitos de datos más grandes.

Muchas veces los Data Mart son utilizados como estrategia para lograr un objetivo mayor, que es construir un Depósito de datos corporativo.

Datos Externos

Dependiendo de la aplicación, el alcance del depósito de datos puede extenderse por la capacidad de acceder a los datos externos. Por ejemplo, los datos accesibles por medio de servicios de computadora en línea y/o vía Internet, pueden estar disponibles a los usuarios del depósito de datos.

Construir un depósito de datos es una tarea grande. No es recomendable emprender el desarrollo del depósito de datos de la empresa como un proyecto cualquiera. Más bien, se recomienda que los requerimientos de una serie de fases se desarrollen e implementen en modelos consecutivos que permitan un proceso de implementación más gradual e iterativa.

Los datos en el Depósito de Datos no son volátiles y es un repositorio de datos de sólo lectura (en general). Sin embargo, pueden añadirse nuevos elementos sobre una base regular para que el contenido siga la evolución de los datos en la base de datos fuente, tanto en los contenidos como en el tiempo.

Uno de los desafíos de mantener un Depósito de Datos, es idear métodos para identificar datos nuevos o modificados en las bases de datos operacionales. Algunas maneras para identificar estos datos incluyen insertar fecha/tiempo en los registros de base de datos y entonces crear copias de registros actualizados y copiar información de los registros de transacción y/o base de datos diarios.

Estos elementos de datos nuevos y/o modificados son extraídos, integrados, transformados y agregados al depósito de datos en pasos periódicos programados. Como se añaden las nuevas ocurrencias de datos, los datos antiguos son eliminados.

Herramientas para el procesamiento analítico en línea (OLAP)

Proceso analítico en línea es el nombre formal para el análisis de cubos multidimensionales una forma mas intuitiva de ver la información empresarial.

Los usuarios acceden al depósito de datos por medio de herramientas de productividad basadas en GUI (Graphical User Interface - Interfase gráfica de usuario). Pueden proveerse a los usuarios del depósito de datos muchos de estos tipos de herramientas.

Las características operativas de las herramientas OLAP se dividen en tres módulos principales:

- Interfase de usuario gráfico OLAP (GUI)
- Lógica de procesamiento analítico OLAP

- Lógica de procesamiento de datos OLAP

Residen en el ambiente Cliente/servidor. Los componentes del sistema OLAP están localizados en una sola computadora, cada analista de datos debe contar con una poderosa computadora para guardar el sistema OLAP y procesar los datos localmente.
Y cada analista debe contar con su propia copia “privada” (extracto) de los datos y programas, lo cual lleva a los problemas de isla de información.

Estos pueden incluir software de consultas, generadores de reportes, procesamiento analítico en línea, herramientas data/visual mining, etc., dependiendo de los tipos de usuarios y sus requerimientos particulares. Sin embargo, una sola herramienta no satisface todos los requerimientos, por lo que es necesaria la integración de una serie de herramientas.

Las herramientas de Acceso al componente de almacenamiento físico Depósito de Datos son herramientas que proveen acceso a los datos. Se accede a ellos por servidores OLAP. Éstos se incluyen en la sección front-end.

Los servidores OLAP aceptan y procesan preguntas desde aplicaciones clientes y envía los datos requeridos en los servidores de depósitos o en las vistas de administrador. Después de las preguntas y respuestas, el servidor OLAP procesa los datos en una ruta multidimensional y presenta los datos a los usuarios. El servidor es implementado con drill-down y roll-up (operaciones OLAP) técnicamente alimentados por los datos necesarios y diferentes. La aplicación cliente es un componente contiene aplicaciones para la especificación de preguntas, análisis de datos, minería de datos, tendencias de pronóstico, y formateo de reportes. Estas aplicaciones envían preguntas a los servidores OLAP y obtienen datos multidimensionales desde el servidor OLAP, proveedor de diferentes funciones de análisis para el usuario y el manejo de los datos.

Las funciones que brindan estas herramientas son:

- Funcionan sobre un sistema de información (transaccional o almacén de datos)
- Permiten realizar agregaciones y combinaciones de los datos de maneras mucho más complejas y ambiciosas, con objetivos de análisis más estratégicos.
- Están basadas, generalmente, en sistemas o interfaces multidimensionales,
- Utilizando operadores específicos (además de los clásicos): drill, roll, pívot y slice.
- El resultado se presenta de una manera matricial o híbrida.
- Proporcionan facilidades para “manejar” y “transformar” los datos.
- Producen otros “datos” (más agregados, combinados).
- Ayudan a analizar los datos porque producen diferentes vistas de los mismos.

Operadores OLAP:

• Roll-up: Cambiar una categoría en la granularidad por una categoría menos fina. Figura 2.2.5.

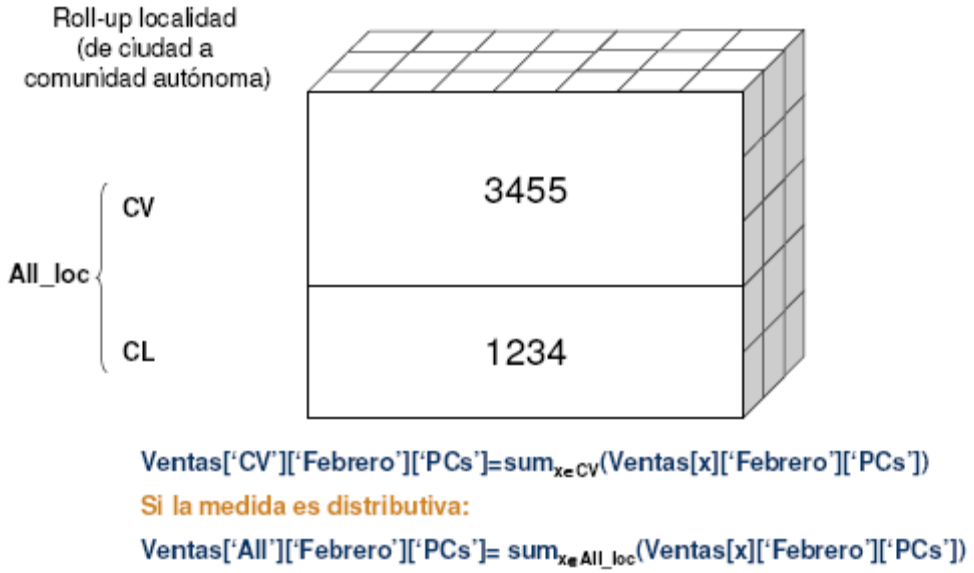

Figura 2.2.5
Roll-up

• Drill-down: inverso de Roll-Up. Aumenta el nivel de detalle. Figura 2.2.6.

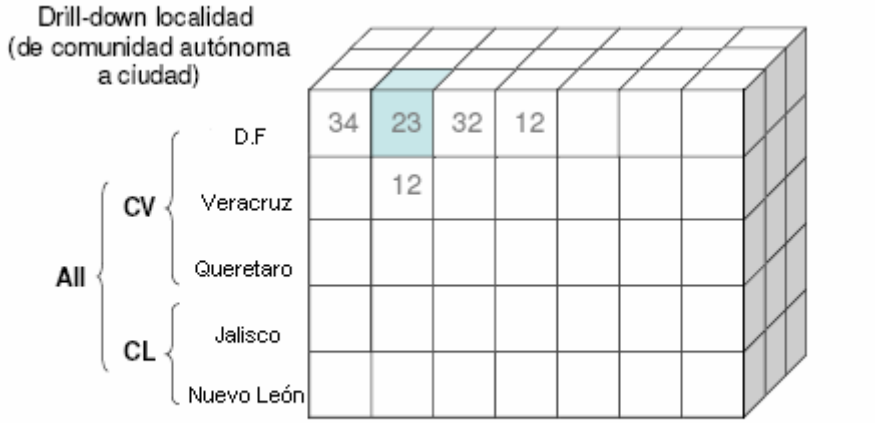

Figura 2.2.6
Drill-down

• Navegación: secuencia de roll-ups y drilldowns

• Drill-across y Roll-across: cruzar más de una tabla de hechos. (Añade dimensiones, elimina dimensiones, respectivamente). Figura 2.2.7.

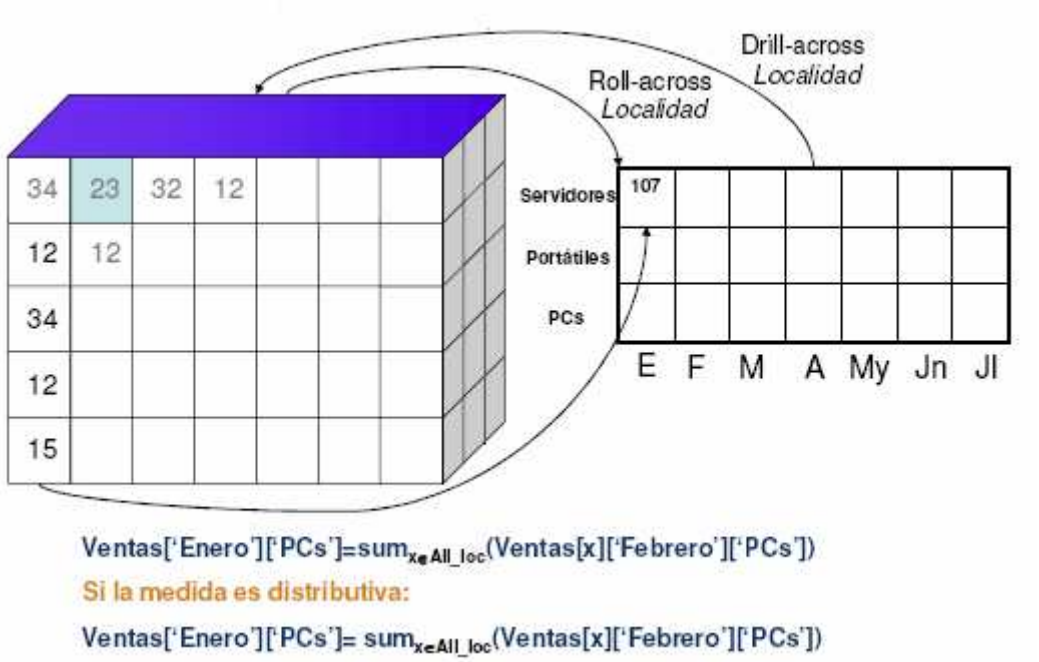

Figura 2.2.7
Drill-across

• Slice & Dice: imponer condiciones sobre las dimensiones. (Selecciones y proyecciones de valores). Figura 2.2.8.

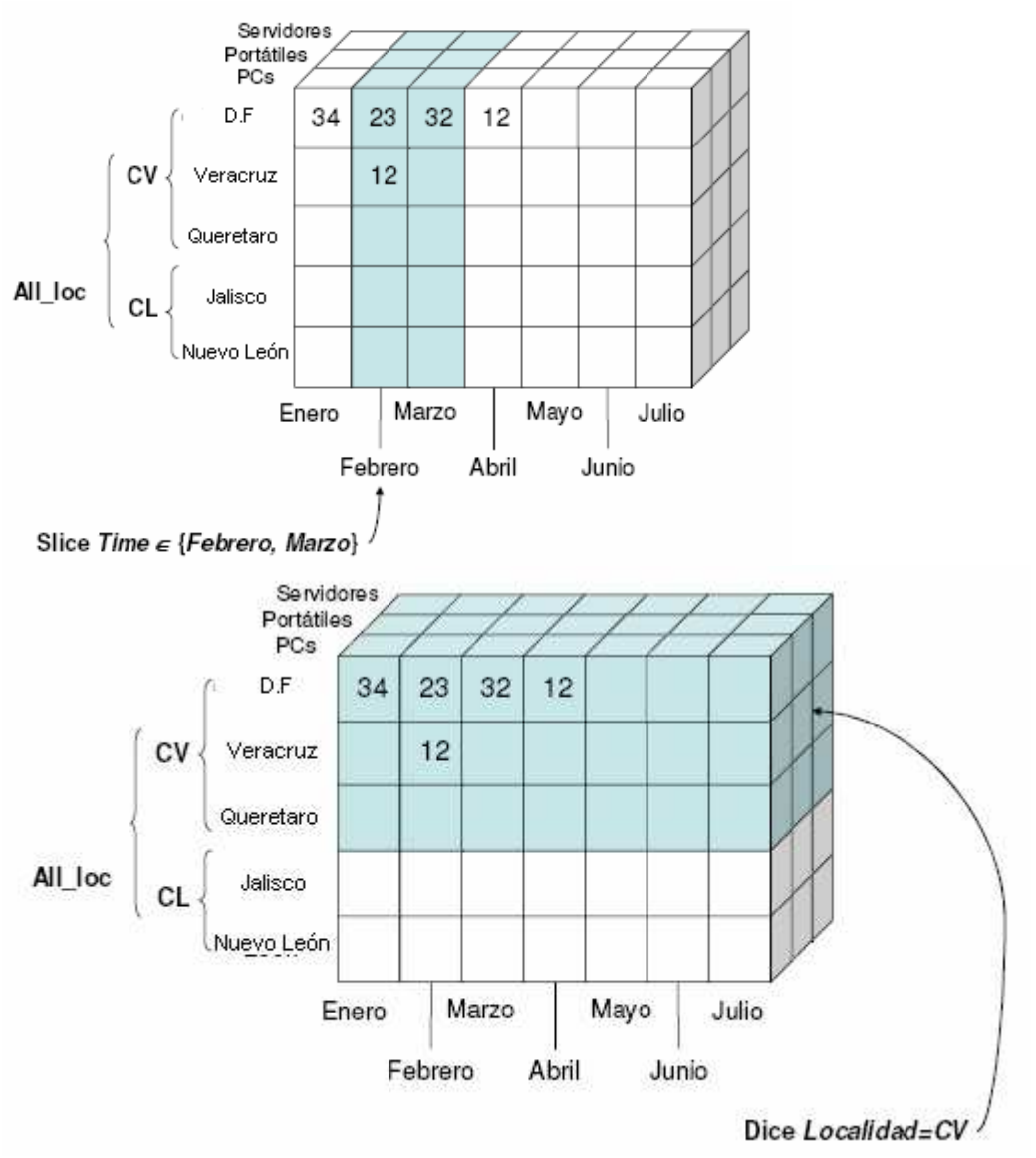

Figura 2.2.8
Slice & Dice

• Pivot: elegir atributos para la tabla de salida y cambiar la disposición de los atributos.

Existen algunas clasificaciones entre las implementaciones OLAP. La clasificación está hecha sobre la base de en qué tipo de motor son almacenados los datos:

- ROLAP es una implementación OLAP que almacena los datos en un motor relacional.
- MOLAP es una implementación OLAP que almacena los datos en una base de datos multidimensional.
- HOLAP (Hybrid OLAP) almacena algunos datos en un motor relacional y otros en una base de datos multidimensional.
- DOLAP es un OLAP orientado a equipos de escritorio (Desktop OLAP). Trae toda la información que necesita analizar desde la base de datos relacional y la guarda en el escritorio. Desde ese momento, todas las consultas y análisis son hechas contra los datos guardados en el escritorio.

Crean un ambiente más avanzado de análisis de datos que apoya la toma de decisiones, el modelado de negocios y las actividades de investigación de operaciones. Los sistemas OLAP comparten cuatro características principales.

- Utilizar técnicas multidimensionales de análisis de datos.
- Proporcionan un soporte avanzado para bases de datos.
- Proporcionan interfaces de usuario final fáciles de usar.
- Soportan la arquitectura cliente/servidor.

Técnicas multidimensionales de análisis de datos

Ésta es la característica más distintiva de las herramientas OLAP. El análisis de los datos multidimensionales se refiere al procesamiento de datos de modo que estos sean vistos como parte de una estructura multidimensional. El interés en el aspecto multidimensional del análisis de datos se deriva del hecho de que los tomadores de decisiones de negocio generalmente visualizan los datos desde una perspectiva de negocio. Es decir, como si estuvieran relacionados con otros datos de negocio.

La visualización multidimensional de los datos permite que los usuarios finales consoliden o agreguen datos a diferentes niveles: cifras de ventas totales por cliente y fecha. Por ultimo la visualización multidimencional de los datos permite que un analista de datos de negocio cambie con facilidad las perspectivas de negocio (dimensiones) de ventas por cliente a ventas por división, por región y así sucesivamente.

Las técnicas de análisis de los datos multidimensionales se amplían con las siguientes funciones:

- Funciones de presentación de datos avanzadas: gráficos tridimensionales, tablas pivote, referencias cruzadas, rotación de datos, cubos tridimensionales, etc. Tales

funciones de presentación son compatibles con hojas de cálculo de computadora de escritorio, paquetes estadísticos y paquetes de consulta y de escritura de reportes.

- Funciones avanzadas de agregación, consolidación y clasificación de datos que permiten que el analista de datos de negocio cree múltiples niveles de agregación de datos, rebanar y cortar en forma de cubos los datos, y bajar y subir los datos a través de diferentes dimensiones y niveles de agregación.

- Funciones computacionales avanzadas: variables orientadas al negocio, relaciones financieras y contables, funciones estadísticas y de pronóstico, etc. Estas funciones se proporcionan automáticamente y el usuario no tiene que redefinir sus componentes cada vez que accede a ellos.

- Funciones avanzadas de modelado de datos: soporte para escenarios “qué sucederá si”, evaluaciones variables, contribuciones variables a los resultados, programación lineal y otras herramientas de modelado.

Para proporcionar un soporte de decisiones eficientes, las herramientas OLAP deben contar con funciones de acceso a datos avanzados. Tales funciones incluyen.

- Acceso a muchas clases diferentes de DBMS, archivos simples y fuentes de datos internas y externas.
- Acceso a datos agregados del almacén de datos, así como a los datos detallados encontrados en bases de datos operativas.
- Funciones avanzadas de navegación por los datos, por ejemplo, de reducción y de agregación.
- Tiempo de respuesta a consultas rápidas y consistentes.
- La capacidad de proyectar la solicitud de los usuarios, expresadas en términos de negocio o modelo, en la fuente de datos apropiada y luego en el lenguaje de acceso a los datos apropiados (por lo general SQL). El código de consulta debe optimizarse conforme a la fuente de datos, sin importar si los datos son operativos o provienen del almacén de datos.
- Soporte para bases de datos muy grandes. El almacén de datos puede crecer con facilidad y rapidez hasta múltiples gigabytes o incluso, terabytes.

Para proporcionar una interfase perfectamente integrada, las herramientas OLAP proyectan en sus propios diccionarios de datos los elementos de datos provenientes del almacén de datos y de las bases de datos operativas. Estos metadatos posteriormente se utilizan para transformar las solicitudes de análisis de solicitudes en códigos consulta (optimizados) apropiados, los que luego se dirigen a la(s) fuente(s) de dato(s) apropiada(s).

La arquitectura Cliente/servidor proporciona un marco de referencia dentro de los cuales pueden diseñarse, desarrollarse y ejecutarse a sistemas nuevos.

El ambiente cliente /servidor permite dividir un sistema OLAP en varios componentes que definen su arquitectura. Después estos componentes pueden colocarse en la misma computadora o distribuirse en varias. Por lo tanto OLAP esta diseñado para uso fácil y flexible.

Plataforma del Depósito de Datos

En la arquitectura se requiere el sistema Cliente/Servidor y tomar en cuenta dos capas; la capa de aplicación y la capa del servidor OLAP, donde el servidor OLAP maneja modelos multidimensionales y la lógica de los depósitos de datos.

La plataforma para el Depósito de Datos es casi siempre un servidor de base de datos relacional. Cuando se manipulan volúmenes muy grandes de datos puede requerirse una configuración en bloque de servidores UNIX con multiprocesador simétrico (SMP) o un servidor con procesador paralelo masivo (MPP) especializado.

Los extractos de los datos integrada/transformada se cargan en el Depósito de Datos. Uno de los más populares RDBMSs disponibles para los Depósitos de Datos sobre la plataforma UNIX (SMP y MPP) generalmente es Teradata. La elección de la plataforma es crítica. El depósito crecerá y hay que comprender los requerimientos después de 3 o 5 años. Uno de los errores más grandes que las organizaciones cometen al seleccionar la plataforma, es que ellos presumen que el sistema (hardware y/o DBMS) escalará con los datos.

El sistema de depósito ejecuta las consultas que se pasa a los datos por el software de acceso a los datos del usuario. Aunque un usuario visualiza las consultas desde el punto de vista de un GUI, las consultas típicamente se formulan como sentencias SQL, porque SQL es un lenguaje universal y el estándar de hecho para el acceso a datos.

La figura 2.2.9 muestra más a detalle y contenidos de cada componente de la Arquitectura. El flujo de la arquitectura se explica a continuación.

La función de los componentes en la sección back-end son los datos integrantes desde los orígenes de datos y estos se mantienen en los depósitos y siempre son consistentes con los orígenes de datos, como definición de herramientas para la extracción, trasformación, transportación y carga de datos, orígenes de datos, vistas de administrador, integradores, etc. Los componentes en la sección de almacenaje de datos y metadatos, suplen los servicios de otros componentes como repositorios de metadatos, servidores de metadatos, depósito de datos y servidores de depósitos. En la sección front-end esta formado por los componentes que suplen las interfases amigables para el usuario, abundan herramientas y aplicaciones para la funcionalidad del análisis, y son capaces de soportar los múltiples modelos de datos, como servidores OLAP, aplicaciones cliente e interfases de usuario gráficas.

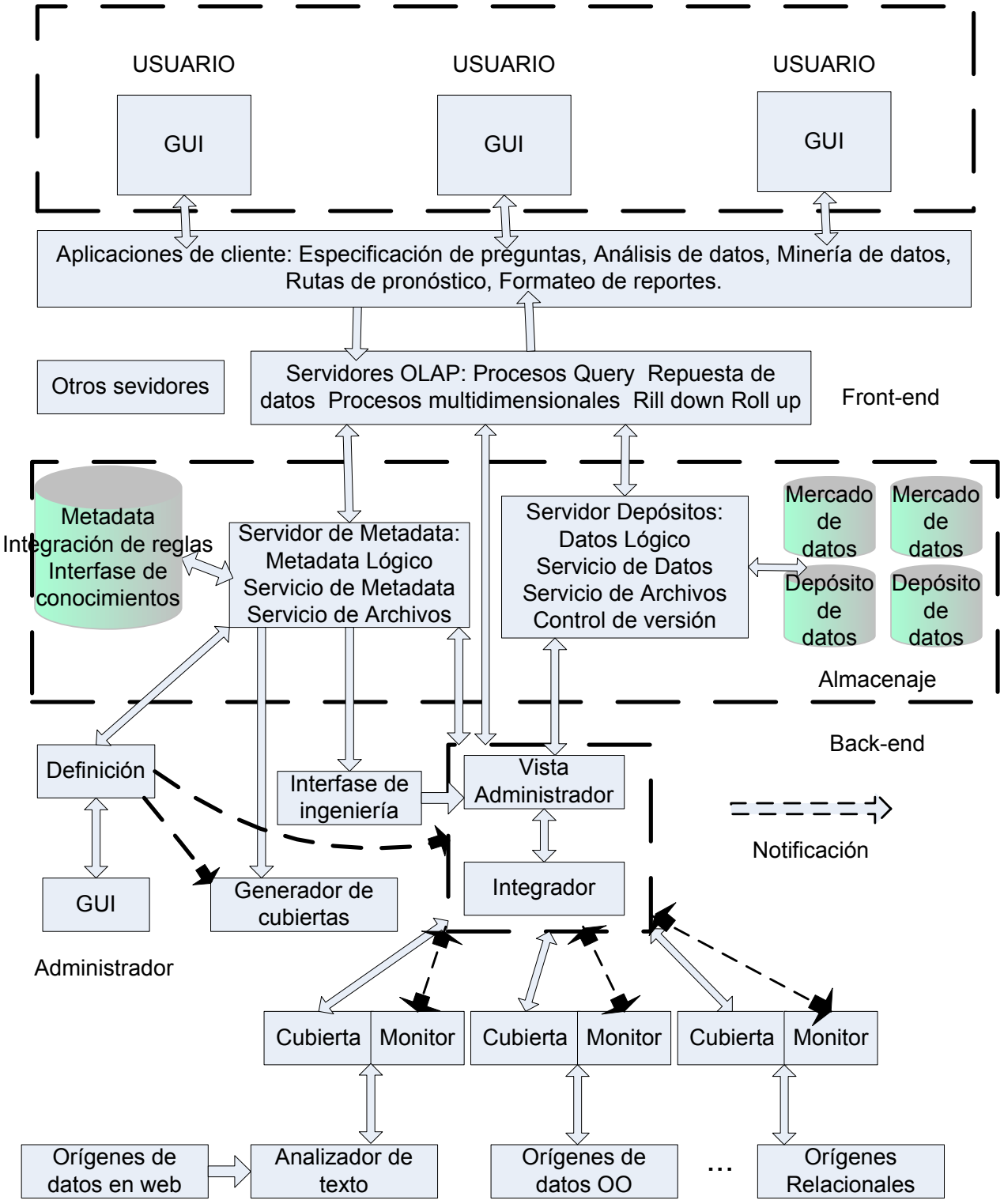

Figura 2.2.9
Arquitectura de los Depósitos de datos a detalle.[5]

[5] Fuentes: http://ieeexplore.ieee.org/Xplore/guesthome.jsp

Es usada para las siguientes tareas:

- Evaluación de las inversiones actuales.
- Análisis de los costos y beneficios.
- Administración y análisis de los riesgos.
- Evaluación de distribuidores.
- Evaluación de productos y herramientas.
- Mantenimiento y mejoramientos.
- Planeación y administración de proyectos.
- Tecnología y obsolescencia.
- Análisis y administración de técnicas.
- Simulación de proyectos.
- Arquitectura y diseño.

El proceso para llegar a la arquitectura para una implementación específica incluye los siguientes pasos:

1. Definir las distintas dimensiones arquitectónicas. La arquitectura de referencia se divide en bloques y estas a su vez se subdivide en bloques más pequeños. En éste punto, la mayor simplificación se obtiene decidiendo las estrategias generales, tales como el depósito de datos Cliente/servidor, el acceso a las microcomputadoras y a la terminal, sólo mercado de datos, o sólo un depósito de datos. El depósito de datos y el depósito operacional son los mismos o el depósito de datos contiene una información que se copia de depósito operacional.
2. Especificar las opciones disponibles a través de estas dimensiones. Las opciones pueden obtenerse del conjunto actual de inversiones e instrucciones estratégicas o pueden considerarse su compra.
3. Seleccionar una opción específica para los elementos en cada una de las dimensiones. Lo que se eligió antes tiene que ser consistente con lo que se elija después. Las primeras opciones imponen restricciones en opciones posteriores del proceso.
4. Después de tomar las decisiones, el conjunto de ellas representa una arquitectura candidata para la implementación de un depósito de datos. Esto debe examinarse detalladamente para asegurarse que las decisiones concuerden con las instrucciones y objetivos de la organización además de ser compatibles las tecnologías unas con otras.

Artículo 00797090. **AN ARCHITECTURE FOR DATA WAREHOUSE SYSTEMS, Jixue Liu Millist Vincent. School of Computer Science, The University of South Australia.**

Estilos arquitectónicos de los Depósitos de Datos

Pero no son las únicas arquitecturas. En la tabla siguiente se mencionan otros estilos arquitectónicos de los depósitos de datos[6] que ofrecen características de soporte de decisiones avanzadas, y algunas son capaces de proporcionar acceso de análisis de datos multidimensionales. Las herramientas de Depósito de datos que utilizan técnicas de análisis de datos multidimensionales se conocen como herramientas de **procesamiento de analítico en línea (OLAP)**.

Tabla 2.2.1 Estilos arquitectónicos de los depósitos de datos

ESTILOS ARQUITECTONICOS DE LOS DEPÓSITOS DE DATOS					
TIPO DE SISTEMA	**ORIGEN DE LOS SISTEMAS**	**PROCESO DE INTEGRACION Y EXTRACCION DE LOS DATOS**	**TIENDA DE DATOS DEPÓSITOS DE DATOS**	**HERRAMIENTA DE CONSULTA DEL USUARIO**	**HERRAMIENTA DE PRESENTACION DEL USUARIO**
OLTP Basado en sistemas Mainframe tradicionales	Datos Operativos	Ninguno. Reporta, lee y resume datos directamente de los datos operativos	Ninguna. Archivos temporales utilizados para propósitos de elaboración de reportes	Muy básica. Formatos de reportes predefinidos; ordenación y cálculos de totales y promedios básicos.	Muy básica Reportes predefinidos basados en menús; solo texto y números
Sistemas de información gerencial (miss)con lenguaje de tercera generación(3GL)	Datos Operativos	Extracción y agregación básica. Lee filtra y resume datos operativos en la tienda de datos intermedia	Datos ligeramente agregados en RDBMS	Muy básica. Formatos de reportes predefinidos; ordenación y cálculos de totales y promedios básicos. cierta preparación de reportes ad hoc mediante SQL	Muy básica Reportes predefinidos basados en menús; solo texto y números, mas algunas definiciones de reportes de columnas ad hoc

[6] Sistema de bases de datos. Diseño, administración e implementación. Carlos Coronel. Ed. Thomson. 2003 Pag. 627

Depósitos de datos departamental de tercera generación	Datos Operativos	Proceso de extracción integración de datos, para poblar una tienda de datos Depósito de datos; ejecutar periódicamente	Base de datos Depósitos de datos de primera generación. Por lo general RDBMS	Herramienta de consulta con algunas capacidades analíticas y reportes predefinidos	Herramientas de presentación avanzadas con capacidades de elaboración de dibujos y grafico
Almacén de datos de empresa de primera generación que utiliza un RDBMS	Datos Operativos	Herramientas avanzadas de extracción e integración de datos. Las funciones incluyen acceso a diversas fuentes de datos, transformaciones, filtros, agregaciones, clasificaciones, programación y resolución de conflictos.	Base de datos Depósitos de datos integrados al almacén de datos para soportar a toda la organización. Utiliza tecnología RDBMS optimizada para propósitos de consulta. Modelo de esquema de estrella.	herramienta de consulta con algunas capacidades analíticas y reportes predefinidos, mas un reporte de consultas mas avanzadas y funciones analíticas con extensiones	Herramientas de presentación avanzadas con capacidades de elaboración de dibujos y grafico, mas algunas herramientas de presentación multidimensional con capacidades de agregación a menor nivel
Almacén de datos de empresa de segunda generación que utiliza un RDBMS	Datos Operativos	Herramientas avanzadas de extracción e integración de datos. Las funciones incluyen acceso a diversas fuentes de datos, transformaciones, filtros, agregaciones, clasificaciones, programación y resolución de conflictos.	El almacén de datos guarda datos que utiliza tecnología de base de datos multidimensionales (MDBMS) basada en estructuras de datos conocidas como "cubos" con múltiples dimensiones	Herramienta de consulta con algunas capacidades analíticas y reportes predefinidos, mas un reporte de consultas mas avanzadas y funciones analíticas con extensiones. Utilizando una interfaz de consulta diferente para acceder al MDBS (patentada)	Herramientas de presentación avanzadas con capacidades de elaboración de dibujos y grafico, mas algunas herramientas de presentación multidimensional con capacidades de agregación a menor nivel utilizando "cubos" matrices multidimensionales limitadas por la dimensión del cubo

Una arquitectura de almacén de datos completa soporta una tienda de datos Depósitos de Datos, un filtro de integración y extracción de datos y una arquitectura de presentación especializada. Para que sea útil el almacén de datos debe contener una base de datos en una sola imagen como centro de la arquitectura de soporte de decisiones. Es decir, los datos del almacén de datos deben ajustarse a formatos y estructuras uniformes para evitar conflictos en los datos. De hecho, antes de que esta base de datos del Depósito de Datos pueda ser considerada como un Depósito de datos verdadero, deben acatar las siguientes reglas descritas a continuación.

Doce reglas que definen al Depósito de datos.

En 1994 William y Check Kelley crearon una lista de 12 reglas que definen a un Depósito de datos, los cuales definen mucho de los puntos que se han tocado con relación a los Depósitos de datos[7].

1.- El Depósito de datos y el ambiente operativo están separados.
2.- Los datos en el Depósito de datos están integrados.
3.- El Depósito de datos contienen datos históricos que abarcan un amplio horizonte de tiempo.
4.- Los datos en ele almacén de datos son capturados instantáneamente en un punto dado de tiempo.
5.- Los datos en el almacén de datos están orientados a sujeto.
6.- Los datos en el Depósito de datos principalmente son de sólo lectura con actualizaciones por lotes periódicas a partir de datos operativos. No se permiten actualizaciones en línea.
7.- El ciclo de vida del desarrollo del Depósito de datos difiere del desarrollo de sistemas clásicos. Los datos motivan el desarrollo del Depósito de datos; los procesos motivan el método clásico.
8.- El Depósito de datos contiene datos con varios niveles de detalle: datos detallados actuales, datos detallados viejos, datos ligeramente resumidos y datos altamente resumidos.
9.- El ambiente del Depósito de datos se caracteriza por transacciones de sólo lectura de conjuntos de datos muy grandes. El ambiente operativo se caracteriza por numerosas transacciones de actualizaciones de unas cuantas entidades de datos a la vez.
10.- El ambiente del Depósito de datos dispone de un sistema que rastrea fuentes, transformaciones y almacenamientos de datos.
11.- Los metadatos del Depósito de datos son un componente crítico de este ambiente. Los metadatos identifican y definen los elementos del dato; proporcionan las fuentes, transformación, integración, almacenamiento, uso, relaciones e historia de cada elemento de datos.
12.- El Depósito de datos contiene un mecanismo de retrocarga para el uso de los recursos que exigen la utilización óptima de los datos por parte de los usuarios.

Estas doce reglas capturan el ciclo de vida de los Depósitos de datos completo, desde su introducción como una entidad aparte de la base de datos operativa, hasta sus

[7] Inmon, Hill y Check Kelley, ibid

componentes, funcionalidad y procesos de administración. La generación actual de OLAP proporciona una amplia infraestructura para diseñar, desarrollar, ejecutar y utilizar Depósitos de datos para el soporte de decisiones dentro de la organización.

2.2.2 Un caso de Negocio

Durante el desarrollo del presente capitulo la definición de Depósito de Datos como una base de datos que almacena información para la toma de decisiones. Y la manera en como es construida a partir de bases de datos que registran las transacciones de los negocios de las organizaciones (bases operacionales). Y su objetivo de consolidar información y hacerla disponible para la realización de análisis de datos de tipo gerencial. Nos lleva a realizar un análisis de forma general de un caso real en donde la prioridad es el acceso interactivo e inmediato a información estratégica de un área de negocios, así mismo se puede distinguir como las operaciones preponderantes no son las transacciones, como en las bases de datos operacionales, sino consultas que involucran gran cantidad de datos y agrupaciones de los mismos. Pero cómo ayudara esto al lector, la idea principal es hacer un comparativo de la Arquitectura mostrada al inicio del capitulo.

Como resultado de este comparativo del caso de negocio se puede enunciar las siguientes características de resultados obtenidos.

- Consolidación de información de clientes.
- Consolidación de información de productos.
- Consistencia de procesos empresariales.
- Uniformidad de las interfases de clientes.
- Portafolios de análisis de productos y servicios.
- Análisis de costos, análisis de ingresos y análisis marginal.

Para cualquier organización que inicia la implementación de un Depósito de datos debe seguir la arquitectura definida para los Depósitos de datos. Por lo cual cualquiera debió haber realizado la colecta de los requisitos para el Depósito de datos y entender cómo los usuarios conducen sus negocios, que información ellos utilizan o que les gustaría hacer, al final del proceso se debe obtener lo siguiente:

1.- Un amplio entendimiento de los negocios de los usuarios.
2.- Detalles específicos sobre los datos o los elementos principales para incluir una implementación inicial de un depósito de datos
3.- Un entendimiento de uso esencial de estos datos iniciales y
4.- Un entendimiento de información común que pueden ser usadas para otras áreas de la organización.

Otra de las formas para obtener los requisitos es por medio de entrevistas utilizando preguntas –llave, cuyas respuestas permiten medir el desempeño de la información o bien en paralelo hacer un trabajo de levantamiento de disponibilidad de datos junto con el

departamento de sistema de información de la organización, para que sea posible determinar una viabilidad e indicar periodos realistas de implementación, evitando crear expectativas anticipadas a los usuarios.

Con el fin de experimentar con las técnicas y arquitectura para el modelado de un Depósito de datos fue desarrollado un caso de negocio.

Caso de Estudio: Construcción de un sistema de apoyo a la toma de decisiones para el área gerencial del Hospital de Clínicas

Se presenta la aplicación para el caso de negocio de una construcción de una base de datos para satisfacer requerimientos de la toma de decisiones (deposito de datos) para el área gerencial del Hospital de Clínicas de Uruguay. El Hospital de Clínicas se presenta como un buen exponente de la variedad de estados en que pueden encontrarse los sistemas legados existentes en una institución de su porte, y las dificultades que se presentan a la hora de reconciliarlos para obtener un depósito de datos corporativo. En este caso se analizan las etapas necesarias para la construcción de un depósito de datos, se propone el uso de un modelo conceptual multidimensional para definir las estructuras de datos que respondan a los requerimientos del área gerencial del hospital y se presenta una solución para la carga del mismo. El desarrollo de este tipo de sistema de información es relativamente nuevo en el área médica. Este caso aporta una identificación de etapas a considerar y una cuantificación de sus tiempos obtenidos en la experiencia de resolver requerimientos del área gerencial de dicho Hospital.

La metodología presentada fue utilizada para resolver los requerimientos del área de Servicios de Apoyo del Hospital de Clínicas. El ejemplo simplifica extracción para ilustrar la forma en que se aplicó la metodología y los tiempos dedicados a cada etapa.

La aplicación de la metodología limita los requerimientos asociados a la evolución de compra de un artículo en un período de tiempo dado. El resultado final será la construcción de un cubo denominado Compras.

Etapa 1. Entrevistas
Los requerimientos a satisfacer se expresan como un conjunto de consultas concretas, que en conjunto identifican el área de análisis que deberá cubrir el depósito de datos y un conjunto de cuantificadores de éxito de la gestión administrativa que se abarca.

Un ejemplo de consulta se presenta a continuación.
Para un proveedor, mostrar la evolución del porcentaje de órdenes de compra que ha cumplido respetando los plazos de entrega en un período de tiempo dado.

Un ejemplo de cuantificador de éxito es el índice de calidad de un artículo que se define como:
El cociente entre el valor de las devoluciones por baja calidad y el valor total del artículo recibido.

En el relevamiento de los sistemas legados se evidencia la inexistencia de documentación en todos, y la carencia de información almacenada electrónicamente en algunos, por lo que mediante un gran número de entrevistas con el personal de los sistemas legados se redefine el alcance de los sistemas a abarcar en el trabajo y se genera el diccionario de datos, diseño lógico, y resumen de funcionalidades para el único sistema factible de ser incluido.

Etapa 2. Análisis
Se adquirió un conocimiento mas detallado de la realidad a modelar mediante el análisis de la documentación construida en la etapa anterior y nuevas entrevistas con el personal del sistema legado.

Se construyó el Diseño Conceptual del Sistema de Recursos Materiales y control del Gasto y la tabla de correspondencias entre diseños lógico y conceptual de dicho sistema legado.

Elaboración del modelado de datos:

Con el resultado del trabajo de recolección de requisitos fue posible comprender un dominio de actividades de investigación y es específicamente.

Un tipo de modelo utilizado fue un modelo dimensional, estos modelos simplifican, facilitan la navegación de los usuarios finales, en el modelado de los indicadores de la toma de decisiones y no la relación de dependencia entre los datos.

Para la construcción del modelo dimensional se consideraron los siguientes pasos:
1.- Selección del Proceso de negocio: este paso fue realizado al mismo tiempo que inicio el caso de estudio, en el momento en que se decide que el asunto sería un producto intelectual.
2.- Detallar hechos: Una recomendación es realizar la selección a menor detalle posible. Cada registro en la tabla de hechos será un producto intelectual representado en la sociedad.
3.- Definición de las dimensiones: Observando las transcripciones de las entrevistas y las clasificaciones de los términos fue posible identificar las dimensiones para los cuales los usuarios pueden acceder a datos de productos intelectuales.

Como todo modelo dimensional, es preciso incluir las dimensiones de tiempo.

4.- Selección de los hechos: Los hechos son generalmente valores numéricos asociados en una gran tabla de hechos.

Etapa 3. Diseño Conceptual Multidimensional
Para el diseño conceptual multidimensional se utilizo CMDM (Conceptual Multidimensional Data Model). CMDM es un modelo conceptual orientado al diseño de estructuras

multidimensionales que propone estructuras de datos y un mecanismo de especificación de restricciones de integridad.

Una característica interesante de CMDM es que distingue entre dimensiones que identifican objetos de la realidad y relaciones dimensionales que representan las relaciones multidimensionales existentes entre dichos objetos.

Como ejemplos de dimensiones, citamos a Tiempo, Ofertas, Autorizaciones para Gastar (APG), Artículos, Proveedores y Transacciones de Ofertas. Como ejemplos de relaciones dimensionales, citamos RelOferta y Ordenes, las cuales se ilustran en las Figuras 2.2.10 y 2.2.11, respectivamente. La relación dimensional RelOfertas vincula todos los elementos participantes de la Oferta que un proveedor hace por la venta de un artículo en el marco de una Autorización para Gastar que emite la institución.

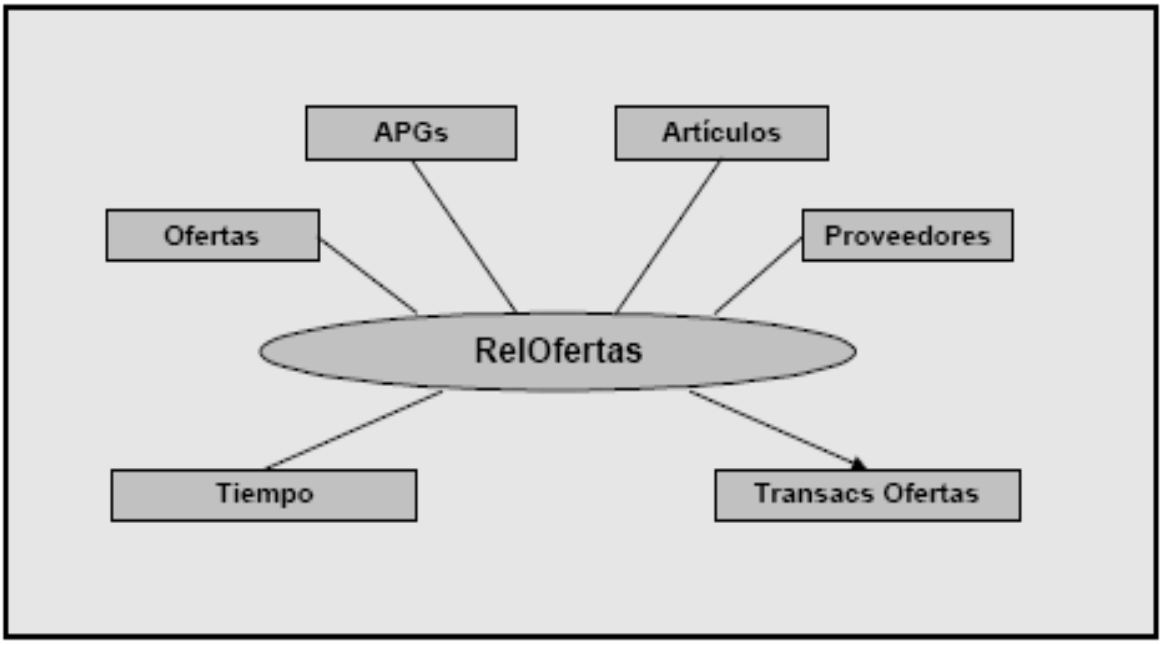

Figura 2.2.10
Relación Dimensional RealOfertas

La relación dimensional Ordenes vincula todos los elementos participantes de la Orden de Compra que la institución emite en respuesta a una Oferta que ha sido aceptada.

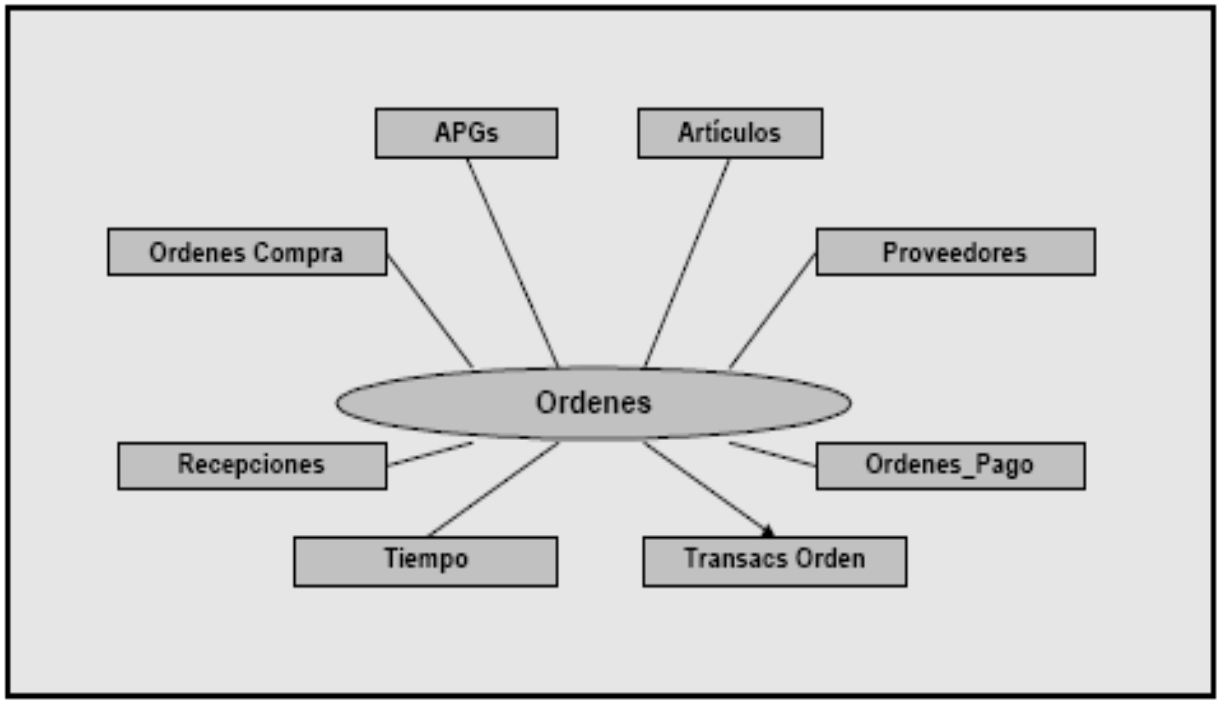

Figura 2.2.11
Relación Dimensional Ordenes

En las Figuras 2.2.10 y 2.2.11 se destaca con una flecha orientada hacia una de las dimensiones que la dimensión apuntada representa la medida de la estructura.

Se construyó además la tabla de correspondencias entre el diseño conceptual del Sistema De Recursos Materiales y Control del Gasto y el diseño conceptual del depósito de datos.

Etapa 4. Diseño Lógico Multidimensional
En esta etapa se definió un depósito de datos global con el nivel más bajo de granularidad en los datos. Se considera que los datos están al nivel más bajo de granularidad cuando no contienen resúmenes u otras operaciones de agrupación realizadas sobre los datos provenientes de los sistemas legados.

Se realizó un cálculo de tamaño relativo al momento de la finalización del proceso de carga y se proyectó el crecimiento hasta 5 años. A modo de ejemplo se presenta el volumen manejado por el Hospital.

- Tamaño al momento de la carga inicial: 300 MB.
- Tamaño proyectado en 5 años: 1,5 GB.

Se consideró el uso de tecnología relacional o multidimensional. Según consideraciones de tamaño y grado de dispersión de los datos existentes, se concluyó conveniente adoptar:

tecnología relacional para el almacenamiento de más bajo nivel de granularidad y tecnología multidimensional para los resúmenes precalculadas. En particular para el almacenamiento de los datos de más bajo nivel de granularidad mediante tecnología relacional se diseñó una tabla de hechos ilustrado en Figura 2.2.12 en el que confluyen las dos relaciones multidimensionales ilustradas en Figuras 2.2.10 y 2.2.11.

Dicha confluencia fue necesaria por la alta frecuencia prevista de operaciones entre las dos relaciones multidimensionales (conocidas como drill across) que serían necesarias para evacuar los requerimientos planteados en la Etapa 1.

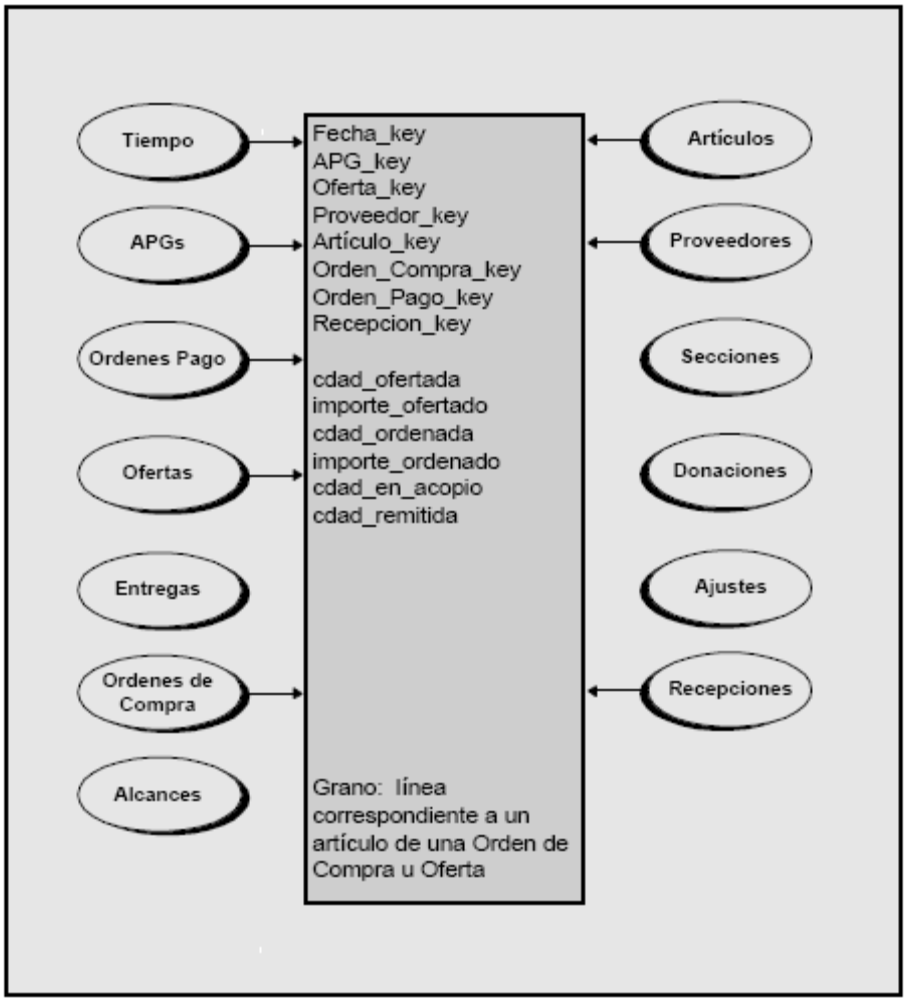

Figura 2.2.12
Tabla de Hechos Compras

El proceso de carga de dimensiones, y tablas de hechos comprende cuatro etapas. A saber,

- Extracción de datos del sistema legado.

- Almacenamiento de los datos en el área de trabajo del Depósito de datos.
- Depuración de los datos.
- Carga de las estructuras finales de las dimensiones y tablas de hechos.

La frecuencia de ejecución del proceso de refresque es diaria para optimizar el grado de frescura de los datos a usar. Dado que el tiempo completo de refresque es un lapso pequeño de tiempo de aproximadamente 20 minutos resulta perfectamente viable su realización diaria. El método de extracción de información de los sistemas legados hace uso de snapshots (volcado completo en archivos de texto). Se consideró esta opción dado que el sistema legado, por desempeño sólo mantiene datos de dos años, y que el volumen de estos datos resulta razonablemente pequeño para ser bajado en una corta ventana de tiempo, se encontró entonces justificable preferir bajar como archivo de texto todo el contenido de tablas, y realizar consultas SQL que identificarán las tuplas a considerar para refrescar el Depósito de Datos. La detección de los cambios se realiza entonces por comparación de snapshots.

La estrategia de actualización del Depósito de Datos con los cambios detectados es diferente para las tablas de hechos y para las dimensiones.

En las tablas de hechos se sobrescribe la tupla modificada.

Se siguió esta opción porque se observo que las actualizaciones que repercuten en cambios en tablas de hechos tienen su origen en la corrección de errores ocurridos en el sistema legado. Por consiguiente, se sustituyen las tuplas correspondientes para reflejar la situación existente en dicho sistema.

En las dimensiones a su vez se siguen estrategias diferentes que se describirán a continuación.

Para las dimensiones APG, Artículos, Ordenes de Compra y Proveedores, la estrategia consiste en agregar una nueva tupla a la dimensión para el mismo ítem, con la nueva versión de los atributos, de manera tal que al asociarse nuevos hechos, se vinculen a la versión más reciente de la tupla en la dimensión. Los hechos anteriores a la actualización, permanecerán vinculados a la tupla más antigua de la dimensión. Esto responde al interés manifestado por los usuarios de conservar los valores existentes previos a la actualización de las referidas dimensiones, y también actualizar al Depósito de datos con los nuevos.

Para las restantes dimensiones, la estrategia consiste en sobrescribir la tupla de la dimensión con los nuevos datos logrando así que tanto los hechos previos como posteriores al cambio resulten vinculados a la nueva descripción de dimensión. Esto se corresponde con el interés del usuario de no recordar los valores anteriores en esas dimensiones, ya sea porque se asume que los cambios se debieron a errores de digitación, o porque carezca de relevancia recordar si antes tuvieron otro valor.

Etapa 5. Implementación
Para la implementación de los procesos de carga y refresque se utilizó una herramienta ETL (Extraction, Tranformation and Loading.)

Esta categoría de herramientas está orientada a la programación de dichos procesos. Consideramos el uso de estas herramientas en lugar de realizar una codificación específica debido a dos facilidades principales que ellas ofrecen. Por un lado, la generación automática de metadatos (es decir, datos acerca de la ejecución de los procesos de carga y refresque de los datos) y por otro lado las facilidades para el mantenimiento de los programas de carga y actualización. En este trabajo se utilizó el DTS (Data Transformation Services) de Microsoft Corporation.

La Figuras 2.2.13 resumen los tiempos dedicados a cada una de las etapas en el desarrollo del Depósito de Datos teniendo en cuenta el servicio de Compras. Las horas indicadas corresponden a horas efectivas realizadas por 2 personas. Se realizan algunas consideraciones sobre las Etapas 1, 2 y 5. En la Etapa 1 se incluye la preparación previa y documentación posterior de las entrevistas así como el tiempo de análisis entre entrevistas. En la Etapa 2 se incluye entrevistas aclaratorias y de confirmación del modelo construido. Finalmente en la Etapa 5 se incluye el estudio y práctica en la herramienta usada.

Etapas			Duración
Etapa 1 Entrevistas	Entrevistas con Nivel Gerencial	20 hs / 2 sems	120 hs / 3 meses y 2 sems
	Entrevistas con Personal S. Legado	100 hs / 12 sems	
Etapa 2 Análisis	Construcción de Diseño Conceptual	20 hs / 3 sems	20 hs / 3 sems
Etapa 3 Diseño Conceptual Multidi-mensional	Estudio del modelo CMDM	10 hs / 2 días	50 hs / 5 sems y 2 días
	Construcción del modelo y tabla de correspondencias	40 hs / 5 sems	
Etapa 4 Diseño Lógico Multidi-mensional	Definición de Arquitectura	16 hs / 1 sem	44 hs / 3 sems y 2 días
	Definición del tipo de almacenamiento	10 hs / 1 sem	
	Mapeo del diseño lógico	10 hs / 1 sem	
	Diseño de los procesos de carga y refresque	10 hs / 2 días	
Etapa 5 Implemen-tación	Implementación	200 hs / 6 sems	200 hs / 1mes y 2 sems

Figura 2.2.13
Duración de las etapas.

El sistema construido se basa en el uso de tecnología conocida Depósito de datos y Olap.

Para el Hospital de Clínicas este trabajo tuvo los siguientes aportes:

- Documentación completa del sistema legado de partida.
- Documentación de diseño del sistema de Depósito de Datos e implementación de un prototipo del cubo de Compras.

- Conocimiento a nivel gerencial de la situación actual de la institución respecto de los sistemas informáticos que necesitan finalización para poder ser integrados a un Depósito de datos corporativo.
- Conocimientos a nivel gerencial de las posibilidades de apoyo que un Depósito de datos puede brindar a su gestión.

Otros ejemplos de uso y aplicación correcta del análisis del depósito de datos se enuncian a continuación en la tabla 2.2.2.

Tabla 2.2.2 Ejemplos de uso de los Depósito de datos

Solución de problemas de negocios y adición de valor con soluciones basadas en Depósitos de datos		
COMPAÑIA	**PROBLEMA**	**BENEFICIO**
Moen Inc. Fabricante de materiales y muebles de baño y cocina Fuente: Cognos Corp. Www.cognos.com	Generación de información muy limitada y laboriosa. Sólo 5 personas sabían como extraer datos con un 3GL Tiempo de respuesta inaceptable para propósito de toma de decisiones por parte de los gerentes.	Respuestas rápidas a preguntas ad hoc para la toma de decisiones. Acceso a datos para propósito de toma de decisiones. Visualización a fondo del desempeño de los productos y márgenes para clientes
Pacific Gas Transmisión co. Proveedor de Gas Natural del noroeste Fuente: Oracle Corp. Www.oracle.com	Cambios rápidos en los mercados a consecuencia de las desregularización. División de utilidades en servicios tradicionales.	los gerentes pueden analizar los datos con rapidez compañía posicionada para identificar tendencias de los mercados nuevos servicios y estructuras
Sega Corporation Sistemas de entretenimiento y juegos de video interactivos Fuente: Oracle Corp www.oracle.com	Se requiere una forma de analizar con rapidez una gran cantidad de datos. Requerida para controlar la publicidad, cupones y descuentos asociados con cambios de precios. Utilizadas para hacerlos con hojas de cálculo de Excel, lo que conduce a errores humanos.	Errores en el ingreso de datos eliminados. Estrategias de comercialización exitosas para dominar nichos de entretenimiento interactivo. Análisis de productos para identificar las mejores ofertas y mercados.
Owens and Minor, Inc. Distribuidor de material médico y quirúrgico Fuente: CFO Magazine	Perdió su cliente mas grande el cual representaba el 10% de sus ingresos anuales ($36 millones). Las	En solo cinco meses se incrementaron las ganancias por acción en 25%. Aumento el negocio, gracias a la

www.cfomagazine.com	acciones se desplomaron. Proceso engorroso para sacar información del sistema mainframe anticuado.	apertura del almacén de datos a los clientes. Los gerentes obtuvieron un rápido acceso a los datos para propósitos de toma de decisiones
La Cellular Compañía de telefonía Celular en el área de los Ángeles que se volvió parte de AT&T Wrilless en 1999. Fuente: PCWeek Online www.zdnet.com/eweek/stories/general/0,11011,392441,00.html	Se requiere un tiempo de respuesta reducido a las preguntas de negocio. Requerida para identificar que promociones están dando resultados y cuales no. Clientes a llamar- tomados de una base de datos que contiene millones de clientes- para ofrecer promociones nuevas.	El numero de suscriptores se incremento 20% gracias a ofertas apropiadas con arreglo a los clientes. Se pudo identificar que promociones eran efectivas. Los tiempos de respuesta se redujeron de 14v minutos a 1 minuto
SEDESOL Secretaria de Desarrollo Social en México	La Secretaria de Desarrollo Social cuenta con diversos programas encaminados a impulsar el desarrollo social y humano de la población. Entre ellos se pueden citar: el Programa para el Desarrollo Local (Micro regiones), Programa Opciones Productivas, Programa de Empleo Temporal, Programa de Jornaleros Agrícolas y Programa Hábitat. El cumplimiento de dichos programas requiere de una serie de mecanismos para la realización, evaluación, seguimiento y resolución de los mismos.	Sedesol cuenta con un sistema centralizado que reúne en un Depósito de Datos la información completa acerca de los distintos programas sociales y de todos sus beneficiarios a lo largo del país. En dicho Depósito de Datos, la Secretaría cuenta con datos detallados sobre el perfil socioeconómico de los hogares y el apoyo de gobierno a casi 10.5 millones de familias beneficiarias –41.9 millones de personas-.

2.2.3 Un enfoque para el desarrollo de un Depósito de Datos

El éxito de un depósito de datos comienza cuando se integran y escogen tres elementos claves, tanto de hardware como de software.

Se debe considerar configuraciones de plataformas de servidores como manejadores de base de Datos, integrar requerimientos de soporte, desempeño, eficiencia y confiabilidad.

Si la información es incorrecta, los Depósitos de Datos proporcionarán a las organizaciones problemas costosos.

Para conseguir que la implementación del depósito tenga un inicio exitoso, se necesita enfocar hacia tres bloques claves de construcción:

- Enfoque de la Arquitectura total del depósito
- Enfoque de la Arquitecturas del servidor
- Enfoque de Sistemas de Gestión de Base de Datos

Enfoque de la Arquitectura Total del Depósito

El desarrollo de los Depósitos de Datos se comienza estructurando de forma lógica y física la base de datos del depósito más los servicios requeridos para operar y mantenerlo. Esta elección conduce a la selección de otros dos ítems fundamentales: el servidor de hardware y el DBMS.

La plataforma física puede centralizarse en una sola ubicación o distribuirse regional, nacional o internacionalmente. A continuación se dan las siguientes alternativas de arquitectura:

1. Un plan para almacenar los datos de la compañía, que podría obtenerse desde fuentes múltiples internas y externas, es consolidar la base de datos en un Depósito de Datos integrado. El enfoque consolidado proporciona eficiencia tanto en la potencia de procesamiento como en los costos de soporte. (Ver Figura 2.2.14).

   En una arquitectura centralizada, una sola, el Depósito de datos integrado refleja todos los aspectos del negocio. Las bases de datos separadas son todas interrelacionadas y físicamente almacenadas en la misma plataforma.

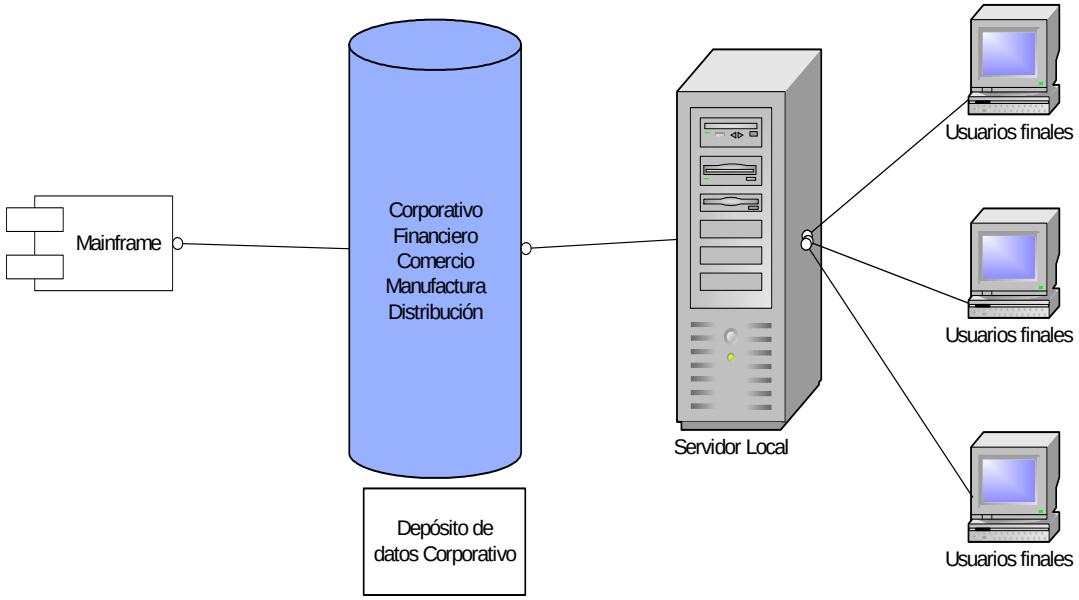

Figura 2.2. 14
Enfoque consolidado para la Arquitectura del Depósito

2. La arquitectura global distribuye información por función, con datos financieros sobre un servidor en un sitio, los datos de comercialización en otro y los datos de fabricación en un tercer lugar. (Ver Figura 2.2.15)

Los datos son consolidados lógicamente pero se almacena por separado sin las bases de datos físicas relacionadas, en los mismos sitios físicos o en diferentes.

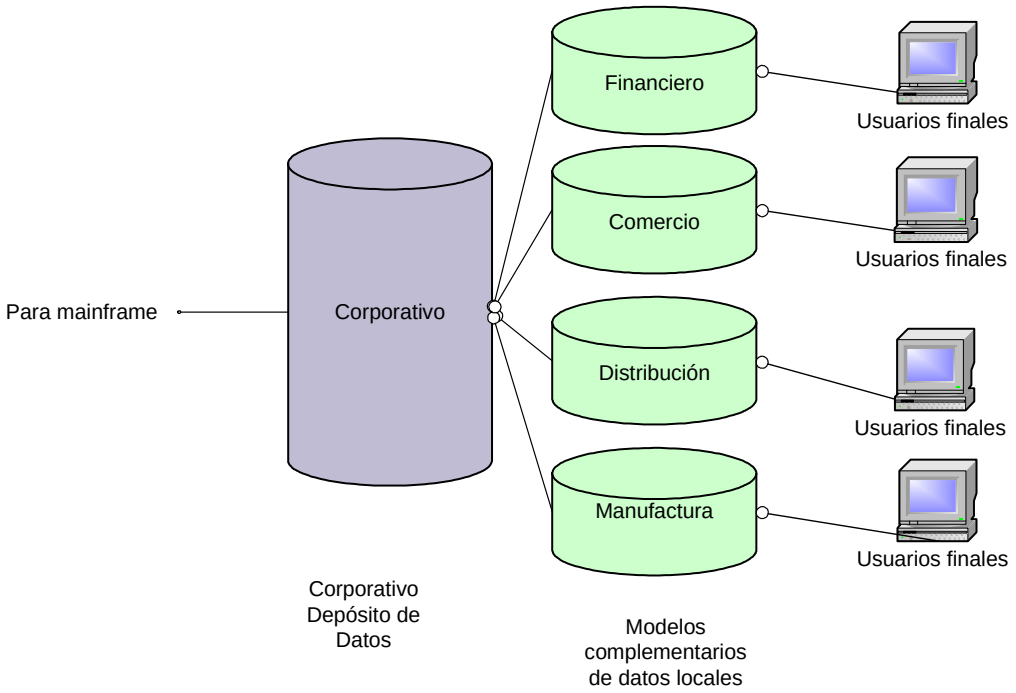

Figura 2.2. 15
Arquitectura Global

3. Una arquitectura por niveles almacena datos altamente resumidos sobre una estación de trabajo del usuario, con resúmenes más detallados en un segundo servidor y la información más detallada en un tercero.

La estación de trabajo del primer nivel maneja la mayoría de las sentencias para los datos, con pocas sentencias que pasan sucesivamente a los niveles 2 y 3 para la resolución.

Las computadoras en el primer nivel pueden optimizarse para usuarios de carga pesada y volumen bajo de datos, mientras que los servidores de los otros niveles son más adecuados para procesar los volúmenes pesados de datos, pero cargas más livianas de usuario. (Ver figura 2.2.16).

Los datos son divididos por niveles de detalle. El nivel 1 de servidores satisfacen la mayoría de los pedidos de los usuarios.

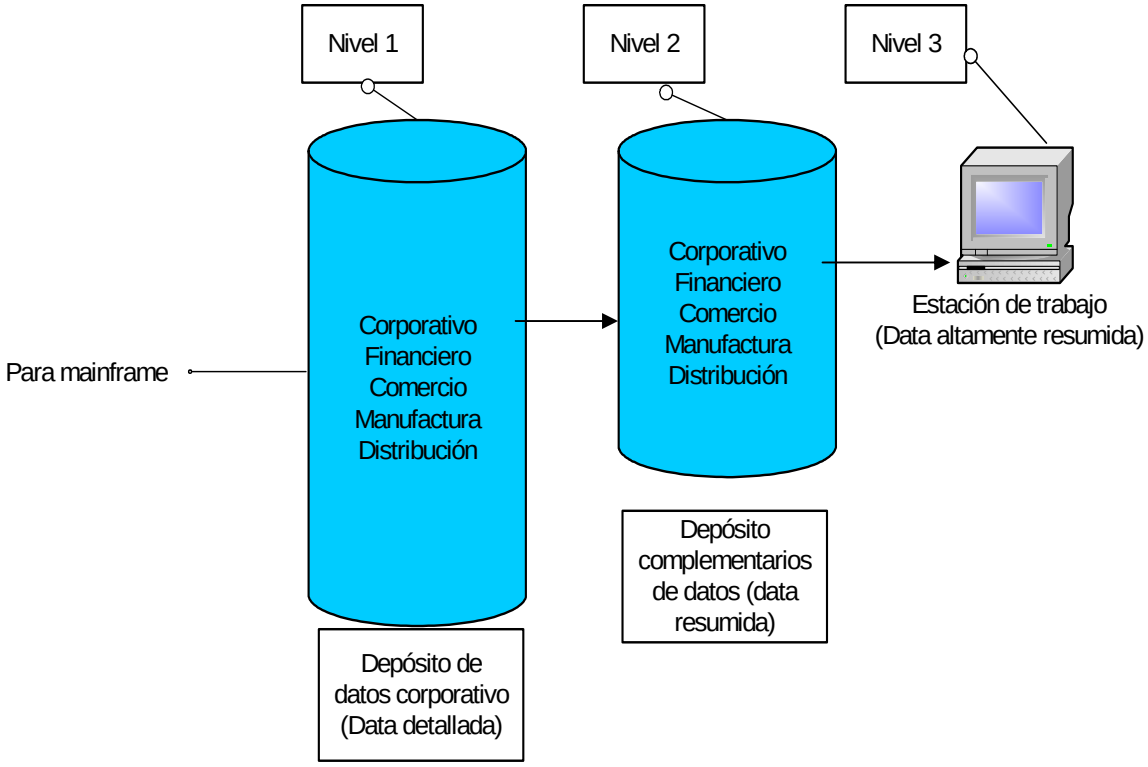

Figura 2.2 16
Arquitectura por Niveles

Enfoque de la Arquitectura del servidor

Al decidir sobre una estructura de depósito distribuida o centralizada, también se necesita considerar los servidores que retendrán y entregarán los datos. El tamaño de su implementación (y las necesidades de su empresa para escalabilidad, disponibilidad y gestión de sistemas) influirá en la elección de la arquitectura del servidor.

- Servidores de un solo procesador

Los servidores de un sólo procesador son los más fáciles de administrar, pero ofrecen limitada potencia de procesamiento y escalabilidad. Además, un servidor sólo presenta un único punto de falla, limitando la disponibilidad garantizada del depósito.

Se puede ampliar un solo servidor de redes mediante arquitecturas distribuidas que hacen uso de subproductos, tales como Ambientes de Computación Distribuida (Distributed Computing Environment - DCE) o Arquitectura Broker de Objeto Común (Common Objects Request Broker Architecture - CORBA), para distribuir el tráfico a través de servidores múltiples.

Estas arquitecturas aumentan también la disponibilidad, debido a que las operaciones pueden cambiarse al servidor de copia de seguridad si un servidor falla, pero la gestión de sistemas es más compleja.

- Multiprocesamiento simétrico

Las máquinas de multiprocesamiento simétrico (Symmetric MultiProcessing - SMP) aumentan mediante la adición de procesadores que comparten la memoria interna de los servidores y los dispositivos de almacenamiento de disco.

Se puede adquirir la mayoría de SMP en configuraciones mínimas (es decir, con dos procesadores) y levantar cuando es necesario, justificando el crecimiento con las necesidades de procesamiento. La escalabilidad de una máquina SMP alcanza su límite en el número máximo de procesadores soportados por los mecanismos de conexión (es decir, el backplane y bus compartido).

- Procesamiento en paralelo masivo

Una máquina 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 deben ser "paralelizadas" es decir, diseñadas para operar por separado, en partes paralelas.

Esta arquitectura es ideal para la búsqueda de grandes bases de datos. Sin embargo, el DBMS que se selecciona debe ser uno que ofrezca una versión paralela. Y aún entonces, se requiere un diseño y afinamiento esenciales para obtener una óptima distribución de los datos y prevenir "hot spots" o "data skew" (donde una cantidad desproporcionada del procesamiento es cambiada a un nodo de procesamiento, debido a la partición de los datos bajo su control).

- Acceso de memoria no uniforme

La dificultad de mover aplicaciones y los DBMS a agrupaciones o ambientes realmente paralelos ha conducido a nuevas y recientes arquitecturas, tales como el acceso de memoria no uniforme (Non Uniform Memory Access - NUMA).

NUMA crea una sola gran máquina SMP al conectar múltiples nodos SMP en un solo (aunque físicamente distribuida) banco de memoria y un ejemplo único de OS. NUMA facilita

el enfoque SMP para obtener los beneficios de performance de las grandes máquinas MPP (con 32 o más procesadores), mientras se mantiene las ventajas de gestión y simplicidad de un ambiente SMP estándar.

Lo más importante de todo, es que existen DBMS y aplicaciones que pueden moverse desde un solo procesador o plataforma SMP a NUMA, sin modificaciones.

Enfoque de los Sistemas de Gestión de Bases de Datos

Los depósitos de datos (conjuntamente con los sistemas de soporte de decisión [Decision Support Systems - DSS] y las aplicaciones cliente/servidor), fueron los primeros éxitos para el DBMS relacional (Relational Data Base Management Systems - RDBMS).

Mientras la gran parte de los sistemas operacionales fueron resultados de aplicaciones basadas en antiguas estructuras de datos, los depósitos y sistemas de soporte de decisiones aprovecharon el RDBMS por su flexibilidad y capacidad para efectuar consultas con un único objetivo concreto.

Los RDBMS son muy flexibles cuando se usan con una estructura de datos normalizada. En una base de datos normalizada, las estructuras de datos son no redundantes y representan las entidades básicas y las relaciones descritas por los datos (por ejemplo productos, comercio y transacción de ventas). Pero un procesamiento analítico en línea (OLAP) típico de consultas que involucra varias estructuras, requiere varias operaciones de unión para colocar los datos juntos.

La performance de los RDBMS tradicionales es mejor para consultas basadas en claves ("Encuentre cuenta de cliente #2014") que para consultas basadas en el contenido ("Encuentre a todos los clientes con un ingreso sobre $ 10,000 que hayan comprado un automóvil en los últimos seis meses").

Para el soporte de depósitos a gran escala y para mejorar el interés hacia las aplicaciones OLAP, los proveedores han añadido nuevas características al RDBMS tradicional. Estas, también llamadas características super-relacionales, incluyen el soporte para hardware de base de datos especializada, tales como la máquina de base de datos Teradata.

Los modelos super relacionales también soportan extensiones para almacenar formatos y operaciones relacionales (ofrecidas por proveedores como REDBRICK) y diagramas de indexación especializados, tales como aquellos usados por SYBASE IQ. Estas técnicas pueden mejorar el rendimiento para las recuperaciones basadas en el contenido, al juntar tablas usando índices o mediante el uso de listas de índice totalmente invertidos.

Muchas de las herramientas de acceso a los depósitos de datos explotan la naturaleza multidimensional del depósito de datos. Por ejemplo, los analistas de marketing necesitan buscar en los volúmenes de ventas por producto, por mercado, por período de tiempo, por promociones y niveles anunciados y por combinaciones de estos diferentes aspectos.

La estructura de los datos en una base de datos relacional tradicional, facilita consultas y análisis a lo largo de dimensiones diferentes que han llegado a ser comunes. Estos esquemas podrían usar tablas múltiples e indicadores para simular una estructura multidimensional. Algunos productos DBMS, tales como ESSBASE y GENTIUM, implementan técnicas de almacenamiento y operadores que soportan estructuras de datos multidimensionales.

Mientras las bases de datos multidimensionales (MultiDimensional Databases - MDDBs) ayudan directamente a manipular los objetos de datos multidimensionales (por ejemplo, la rotación fácil de los datos para verlos entre dimensiones diferentes, o las operaciones de drill down que sucesivamente exponen los niveles de datos más detallados), se debe identificar estas dimensiones cuando se construya la estructura de la base de datos. Así, agregar una nueva dimensión o cambiar las vistas deseadas, puede ser engorroso y costoso. Algunos MDDBS requieren un recargue completo de la base de datos cuando ocurre una reestructuración.

Elección de Herramientas[8]

1.- Seleccione un conjunto de herramientas que soporte la fuente de datos original. Sin ese soporte se debería optar por la solución OLAP relacional debido a que provee una arquitectura abierta.

2.- Después de haber seleccionado el soporte de su fuente de datos determine cuantos análisis se requieren realmente:

- Si sólo es requerido saber “cuantos” y “cuanto”, será suficiente una herramienta básica de consultas y reportes.
- Si es requerido un análisis más avanzado que explique las causas y los efectos de las ocurrencias y de las tendencias, se sugiere una solución OLAP.
- Para las herramientas de minería de datos se requiere expertos en análisis y se necesitan para pronósticos, clasificación y creación del modelo.

3.- Para el mejor desempeño se puede optar por una o varias soluciones. Los usuarios deben comprender los requerimientos de tecnología, desarrollar soluciones que reúnan esos requerimientos, manteniendo y mejorando el sistema.

En la siguiente tabla 2.2.3 se muestran los parámetros que se debe tomar en cuenta para la selección de las herramientas adecuadas.

[8] http://200.14.84.223/apuntesudp/docs/civil_ind/(ICI2423)Sistemas_de_Informacion/(2003-06-11)_818_datawarehouse_o_bodegaje_de_datos.pdf

Tabla 2.2.3 Parámetros para la selección de las herramientas adecuadas.

Tipo de Herramienta	Resultados	Quien los utilizan
Consultas y reportes	Reportes de ventas mensuales, histórico de inventario	Requiere datos históricos, las aptitudes técnicas son limitadas
Procesamiento analítico en Línea (OLAP)	Ventas mensuales contra Cambios de precio de los competidores	Requiere de una visión estática de los datos a “slicing and dicing”
Sistema de información ejecutiva	Libros electrónicos. Centros de comando.	Requiere de información resumida o de alto nivel.
Minería de datos	Modelos predictivos	Requiere de extraer la relación y tendencias de los datos ininteligibles.

En el Anexo B se muestran las tablas de las diferentes herramientas para consulta, reporte y creación de Depósito de Datos.

La Metodología de Evaluación Kavas

La evaluación de un sistema de Depósito de datos tiene los siguientes objetivos: guiar el diseño y el desarrollo del prototipo, cuantificar los diferentes aspectos de la presentación del sistema, y verificar su aceptabilidad. La metodología de evaluación KAVAS pone mucho énfasis en el ciclo de vida de diseño y desarrollo del sistema de decisión y soporte. El objetivo es proporcionar herramientas de validación y métodos consistentes y coherentes para guiar el desarrollo y la evaluación del sistema de Depósito de datos desde la concepción del prototipo, es decir desde el diseño, a través de su desarrollo iterativo. En otras palabras, la intención es proporcionar un formato estructurado en el cual pueda tener lugar el desarrollo iterativo del prototipo. La técnica del prototipo sigue los principios básicos de filosofía KADS. La metodología está basada en un ciclo iterativo de cuatro fases. La fase de la evaluación está representada en la figura 2.2.17.

Fase 1. Primer desarrollo del prototipo del sistema de Depósito de datos o exploración preliminar.

Fase 2. Evaluación de la validez de la base de conocimientos del sistema Depósito de datos.

Fase 3. Evaluación de la funcionalidad en circunstancias controladas (test de laboratorio).

Fase 4. Evaluación del impacto en el escenario de trabajo (pruebas de consecuencias a largo plazo de aplicación del sistema Depósito de datos).

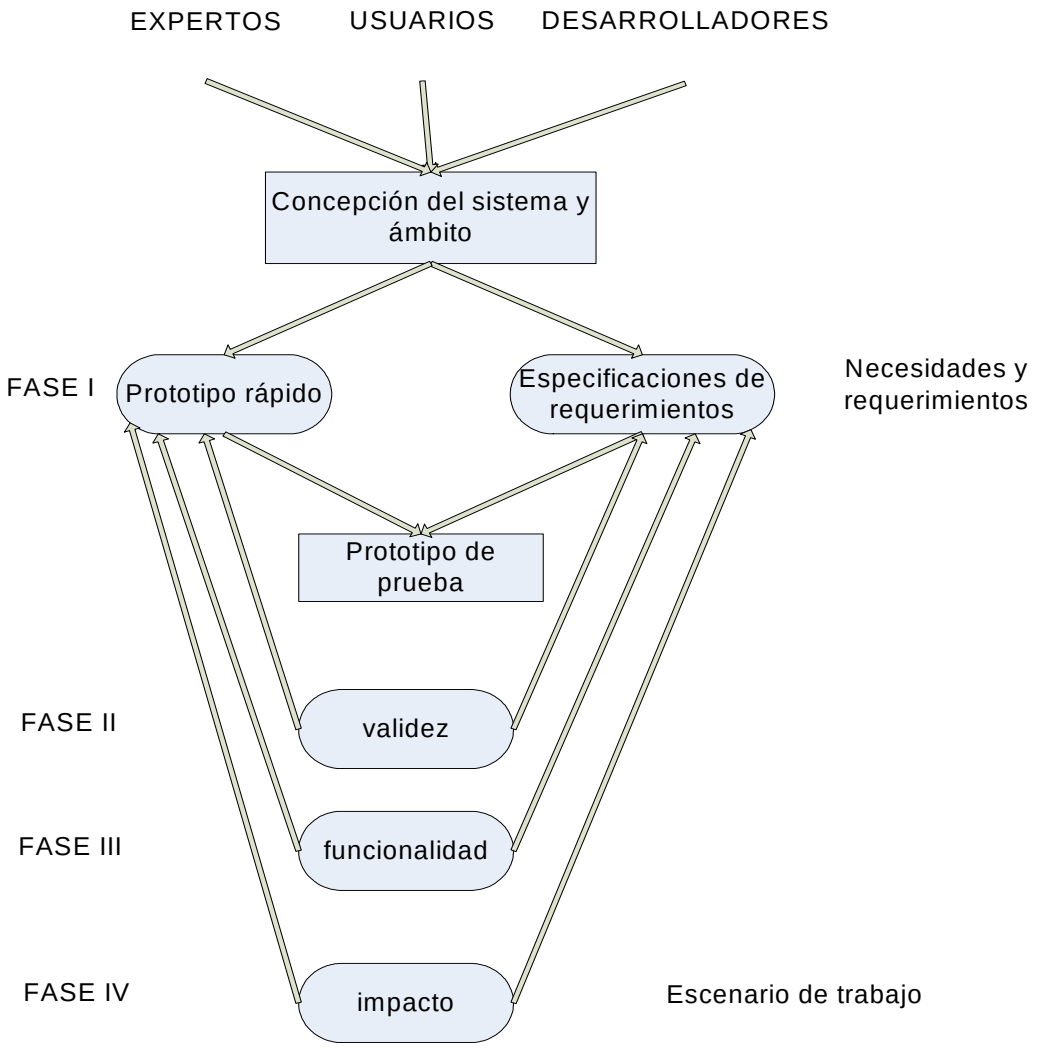

Figura 2.2.17
Metodología de evaluación KAVAS

Cada fase de desarrollo del prototipo tiene que acabar con una sesión de evaluación, durante la cual, todas las partes responsables tienen que revisar los resultados y ponerse de acuerdo sobre los aspectos mas importantes antes de pasar a la fase siguiente. Durante cada fase de evaluación las deficiencias de todas las fases anteriores deben ser examinadas de nuevo. Cada punto débil encontrado durante una fase de metodología, se necesitará una nueva iteración de la fase de desarrollo correspondiente. El proceso de

reiteración puede comenzar directamente en la fase de desarrollo anterior o en la fase preliminar de diseño, en la cual los requerimientos de los usuarios han sido previamente especificados. La primera fase es crucial en la planificación de la evaluación debido a que define metas y características:

Las necesidades para un Depósito de datos (los requerimientos de los potenciales usuarios en general). El dominio para que el Depósito de datos proporcione soporte (los límites del dominio debería definirse de manera no equivocada).

El papel de usuario utilizará el Depósito de datos, es decir, como va a influenciar a las decisiones en la salida del Depósito de datos.

Los criterios de presentación mínima, respecto a los cuales de la evaluación deben ser comprados, también tienen que ser definidos. Estos criterios corresponden a las razones potenciales de fracaso. El primer proceso de evaluación es valorar de que manera los requerimientos satisfacen las necesidades del usuario. Esta fase que corresponde a la concepción del Depósito de datos, viene seguido por la especificación de los requerimientos y de los ciclos rápidos de prototipo que llega al desarrollo del prototipo de prueba del sistema. Hay dos extremos de desarrollo de un primer prototipo:

- Un prototipo rápido con ciclos de desarrollo en un ambiente protegido de laboratorio.
- La especificación formal de la técnica de desarrollo de los requerimientos.

El desarrollo del prototipo de pruebas que se lleva a cabo a través de una combinación de especificaciones de requerimientos y ciclos rápidos de prototipo y pruebas. Un análisis detallado de la especificación de los requerimientos en más lento que el prototipo rápido pero, siendo más metódica, suele llevar aun prototipo de trabajo satisfactorio en un menor tiempo total. El prototipo rápido es particularmente útil cuando no es posible lograr una definición completa de los requerimientos del usuario sin antes desarrollar un primer prototipo. Este método puede resultar más útil en dominio donde desarrolladores de Depósito de Datos y usuarios pueden ser consistentes de las posibilidades futuras del sistema. Muchas veces estos conocimientos pueden ser adquiridos después de pruebas y errores con otros prototipos. Algunas especificaciones pueden ser definidas antes mientras que otras son definidas durante el ciclo de prototipo rápido. Antes de proceder al proceso de desarrollo a través de los estados de evaluación y desarrollo que llevan al prototipo de pruebas, deben, ser elegidos el sistema gestor de base de datos y el mecanismo de inferencia del Depósito de datos.

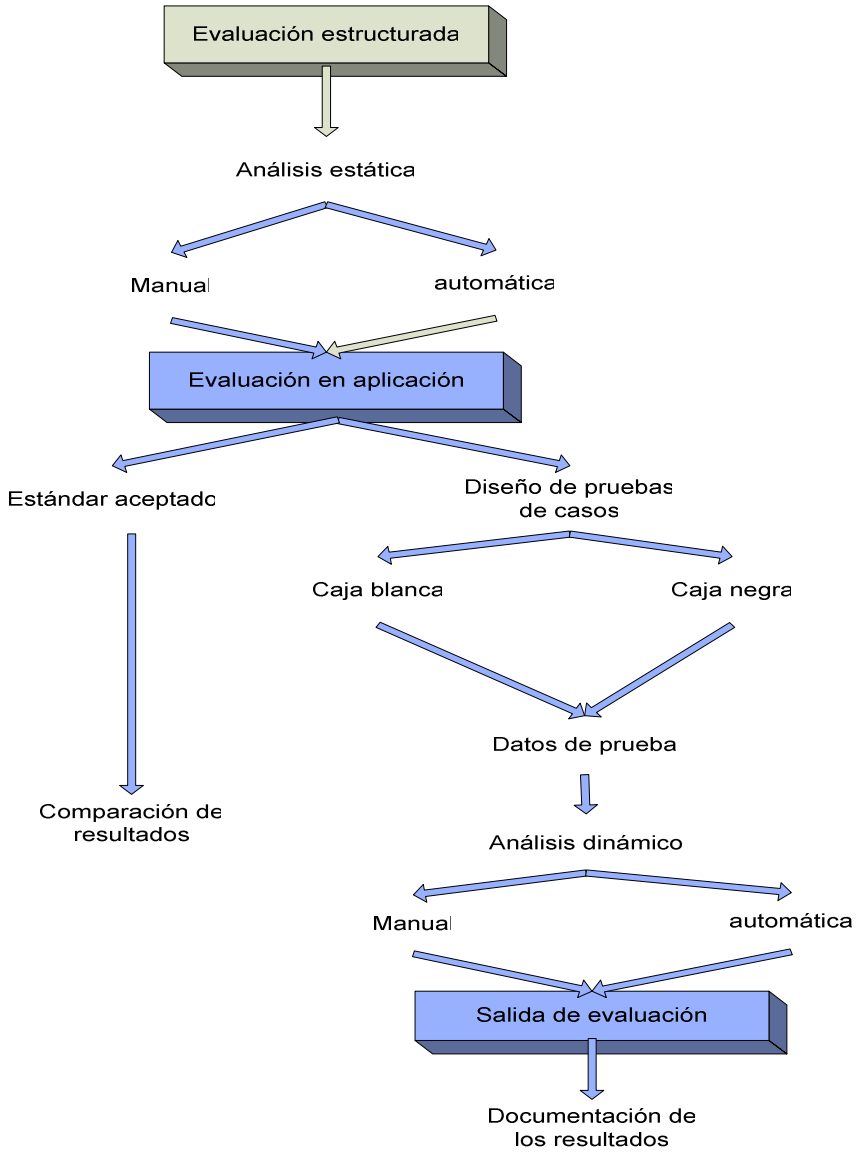

Figura 2.2.18
Evaluación de la validez de los Depósito de datos

Una vez elegido un prototipo de los depósitos de datos es apropiado avanzar a una evaluación más estructurada en el ambiente del laboratorio. El enfoque puede ser en anticipación o en predicción de las razones de fracaso de cada fase. La segunda fase de la metodología de evaluación implica examinar la validez del Depósito de datos. Eso se alcanza con una inspección de la componente base de datos, y verificando que la salida del depósito de datos sea la misma que el estándar de exactitud definido.

Para validar la corrección de la base de conocimientos del depósito de datos, puede usarse dos clases de métodos de evaluación técnica: métodos estáticos y dinámicos de pruebas. Los métodos de pruebas estáticos vienen realizados por los ingenieros de los conocimientos, cuya tarea es la verificación directa de los aspectos estructurales de la componente de base de datos. El análisis dinámico de la base de datos de los conocimientos, a través de métodos dinámicos de pruebas, implica la ejecución del depósito de datos y viene realizando después del análisis estático. El análisis dinámico de una base de datos involucra la generación de datos de pruebas que pueden desarrollarse teniendo en cuenta, o no, la estructura interna de un depósito de datos (técnica de la caja blanca y de la caja negra).

Para poder comparar los resultados, es importante definir la tasa aceptable de error. Además, se deben determinar cuáles son las respuestas aceptables, ya que puede existir no única respuesta correcta porque en cierto número de respuesta puede considerarse aceptables. Cuando ni el depósito de datos y ni los expertos pueden proporcionar respuesta perfectas, la salida del depósito de datos debe tener en consideración la tasa aceptable de error definida. Para estimar la precisión del depósito de datos es importante comparar la salida que se obtiene, como resultado del procesamiento de casos de prueba, con alguna forma de estándar comúnmente aceptado (agreed truth o gold standard).

Una tercera metodología de evaluación se refiere a la evaluación de la funcionalidad del prototipo de pruebas, e implica la presentación del depósito de datos a usuarios y expertos con el fin de proporcionar una retroalimentación. La funcionalidad de un depósito de datos no debe ser confundida con su uso o aceptabilidad, que es sólo una parte de lo que se entiende por funcionalidad. La funcionalidad se divide en interna y externa:

La funcionalidad Interna se refiere a la interacción entre usuarios y depósito de datos.

La funcionalidad externa tiene que ver con asuntos más generales de transferencia y mantenimiento del Depósito de datos.

La última fase se refiere al impacto de un depósito de datos validado y a la consecuencia a largo plazo debido a su aplicación en el contexto de trabajo.

Una Metodología para la definición de requisitos en sistemas Depósitos de Datos

Se sugiere la utilización de técnicas de ingeniería de requisitos en el soporte de una especificación de requisitos de alta calidad en sistemas de Depósitos de Datos. El uso de una metodología garantiza la correcta identificación de los requerimientos de los usuarios, un proyecto de sistema definido y la definición de los modelos dimensionales que ayuda en todas la necesidades de análisis estratégica de los usuarios finales.

Esta metodología está centrada en un modelo de fases que brinda una dirección el proceso de especificación de requisitos, en cuanto se propone un conjunto de hechos para una colecta de aspectos funcionales, no funcionales y multidimensionales que integran una

solicitud de servicio. La metodología sirve como un instrumento que hace posible que la visión dimensional y de requisitos estén siempre ajustadas a las necesidades de los usuarios.

Una de las primeras medidas de desarrollo de un sistema de software es verificar si se atienden las necesidades de los clientes. Pero muchos de los requisitos son inadecuados, inconsistentes, incompletos y ambiguos y ejercen un gran impacto en la calidad del software final. Partir de esta observación se concluye lo siguiente: *los requisitos para un dado sistema no aparecen naturalmente; al contrario requieren de ser revisado y proyectados continuamente.*

Los estudios indican que cuando son detectados apenas después de implementado el software, errores en los requisitos de software, es 20 veces mas caro la corrección que cualquier otro tipo de error. Los estudios relacionados a los requisitos de sistemas afectan a buena parte de las organizaciones que desenvuelven y usan sistemas de software. Corroborando, nuevos estudios reconocen *la Ingeniería de Requisición* como una de las fases más importantes en el proceso de la ingeniería del software.

Con la finalidad de mejorar la calidad de los requisitos, es necesario definir inicialmente que son los “requisitos”. Existen numerosas definiciones disponibles en la literatura para los términos. Según KOTONYA y SOMMERVILLE (1997), un requisito puede describirse:

(1) Una facilidad a nivel de usuario; por ejemplo, un corrector de gramática y ortografía.
(2) Una propiedad muy general de los sistemas.
(3) Una restricción especifica en el sistema.
(4) Una restricción en el desarrollo del sistema.

Una definición más simple para los requisitos es por MACAULAY (1996), según el cual requisito es “simplemente algo que los clientes necesitan”. En este caso se adapta la definición de DORFMAN y THAYER (1990), los cuales conceptualizan los requisitos de la siguiente forma:

(a) Una capacidad del software necesaria para el usuario para resolver un problema y distinguir un objetivo.
(b) Una capacidad del software que un sistema (o uno de sus componentes) debe distinguir o poseer para satisfacer un contrato, especificación, u otra documentación formalmente impuesta.

Además de los requisitos, otra definición necesaria es la de Ingeniería de Requisitos. Según la IEEE, la ingeniería de requisitos corresponde al proceso de adquisición, perfeccionamiento y verificación de las necesidades de los cliente para un sistema de software, tiene el objetivo de la especificación completa y correcta de los requisito de software (IEEE, 1984).

Una especificación de requisitos completa abarca una amplia gama de requisitos tradicionalmente, los requisitos de software son clasificados en requisitos funcionales, no funcionales y organizacionales (SOMMERVILLE, 2001; ALENCAR, 1999):

- ✓ Requisitos funcionales: son las declaraciones de las funciones que el sistema debe favorecer, cómo el sistema se comporta con entradas particulares y como el sistema se debe comportar en situaciones específicas. El termino función es usado en el sentido genérico de la operación que puede ser realizada para el sistema, sea por medio de los comandos de los usuarios, o sea por la ocurrencia de eventos internos o externos en el sistema. En algunos casos, los requisitos funcionales pueden también explícitamente definir lo que el sistema no debe hacer.

- ✓ Requisitos no funcionales: son las restricciones no funcionales ofrecidas para el sistema. Incluyen restricciones de tiempo, restricciones en proceso de desarrollo, patrones, y cualidades globales de un software, como mantenimiento, uso, desempeño, costo, etc.

- ✓ Requisitos Organizacionales: derivados directamente de procedimientos y políticas organizacionales y relacionados con los objetivos y metas de la organización.

El proceso de ingeniería de requisitos es un conjunto estructurado de las actividades que son seguidas para derivar, validar y mantener un documento de requisito. Una descripción completa del proceso debe incluir: (a) cuales actividades son ejecutadas; (b) su estructura y cronograma; (c) quienes son los responsables por cada una de las actividades;(d) sus entradas y salidas; y (e) las herramientas usadas para soportar la ingeniería de requisitos.

El proceso de ingeniería de requisitos es un proceso de proyectos con entradas y salidas, como se describe en la tabla 2.2.4.

Tabla 2.2. 4 Entradas y salidas del proceso de ingeniería de requisitos.

ENTRADAS Y SALIDAS	TIPO	DESCRIPCION
Información de los Sistemas Existentes	Entrada	Referencia de información general sobre los sistemas que será substituido o creado y de otros sistemas con el cual el sistema deberá intervenir.
Patrones corporativos	Entrada	Referencia de patrones y normas adoptadas para empresa para el desarrollo de sistemas, incluyendo métodos para el desarrollo, practicas para garantía de cualidades, etc.
Normas y regalamientos	Entrada	Normas y regalamientos externos que se aplican al sistema.
Requisitos definidos	Salida	Descripción de los requisitos levantados,

		validados y aprobados para las partes interesadas.
Especificación de los sistemas	Salida	Una especificación más detallada de los sistemas al ser desarrollados.
Modelos de los sistemas	Salida	Un conjunto de los modelos que describen el sistema a partir de diferentes perspectivas.

La metodología Hadden-Kelly

La metodología Hadden-Kelly (HADDEN y KELLY, 1997) es una de las más difundidas en la industria y licencias para grandes empresas de informática como el **Sofware AG, Reson, Va**, entre otras. Earl Hadden y Sean Nelly proponen un método evolucionado, centrado en una construcción rápida de un Depósito de Datos /Data Marts, con perfeccionamiento posterior de infraestructura tecnológica, reglas de negocio, proceso de ETL y requisitos de negocio de usuario.

El método está divido en cuatro fases:

1) Preparación, en el cual se define el "camino óptimo" (fundamentos de los negocios, ambiente técnico, disponibilidad y cualidades del dato, recursos necesario) para la construcción de un Depósito de Datos/Data Mart. Como salida, esta etapa representa un plano de acción para la organización.

2) Planeación, enfocados en objetivos, datos y áreas específicas del negocio que será parte de los Depósito de datos. Un plano de implementación describe las actividades de construcción del proyecto es preparado y sus prioridades son atribuidas de acuerdo con la importancia de los datamart para la estrategia de negocio.

3) Construcción, corresponde a la implementación de un datamart (y/o parte de los Depósito de datos) que atienden a un objetivo en particular de negocio. Incluye un mapeo de datos operacionales para el modelo del depósito, definición y generación de programas de extracción y transformación, control de calidad de datos, entre otras actividades.

4) Operación, caracterizada para conducir las actividades de respaldo y recuperación de datos, monitoreo de performance, y memoria de los procesos con base en los feddbacks obtenidos.

La metodología KIMBALL

La metodología creada por KIMBALL (1998) tiene como fundamento un *framework* conceptual que describe una secuencia de etapas de alto nivel requeridas para el proyecto,

desarrollo e implementación efectivos de un Depósito de Datos. El ciclo de vida propuso iniciarse con el planeamiento del proyecto. En esta etapa son definidos el espacio de la aplicación, los criterios de la validación y la oportunidad de negocio que justifica su implementación. KIMBALL acredita que la probabilidad de sucesos de un proyecto de Depósito de Datos es considerablemente aumentada en un entendimiento consistente de lo requisitos de los usuarios es establecido. De esta forma, el framework propio, enseguida al planeamiento, es una etapa de definición de los requisitos de negocio, cuyo objetivo principal es alcanzar el entendimiento de los requisitos de negocio que motivan la construcción de los Depósitos de Datos y traducirlos para considerarlos del proyecto. La fase subsiguiente del modelado dimensional construido, es a partir de los requisitos levantados, un modelo multidimensional que implementa las necesidades estratégicas de los usuarios y gerentes extendidos a los depósitos.

Las demás fases que integran los *framework* tratan de los aspectos relacionados con la implementación física de los modelos multidimensionales, definición de las infraestructuras para soporte de los procesos de extracción/transformación/Carga (ETL) e implementación de los Depósitos de datos. En relación con las metodologías similares, con el uso de la metodología KIMBALL se destaca por la separación entre requisitos de negocio y proyecto físico de aplicación; por la búsqueda de los requisitos multidimensionales comunes con el modelo corporativo; y por el uso de asesores de entrevista y reuniones para solicitar, analizar y negociar requisitos de sistema.

Metodología Golfarelli y Rizzi

GOLFARELLI y RIZZI (1999) propone un framework genérico para el proyecto del Depósito de datos, estructurado por seis fases:

(i) análisis del sistema de información;
(ii) especificación de requisitos;
(iii) proyecto conceptual;
(iv) perfeccionamiento y validación del esquema conceptual;
(v) proyecto lógico; y
(vi) proyecto físico.

La fase de análisis trata de un aspecto bastante peculiar sobre las aplicaciones de los Depósitos de datos, comparada con las aplicaciones convencionales: la existencia de documentación es la prioridad sobre una fuente verdadera de los datos. La documentación es analizada conjuntamente por proyectistas y equipos del sistema proveedor, para producir los esquemas iniciales de la integración. Esos esquemas engloban entradas para la fase de especificación de los requisitos, donde las necesidades de los usuarios son filtradas, produciendo como salidas una especificación preliminar de los hechos y dimensiones.

Desde el punto de vista gráfico, un esquema multidimensional es elaborado a partir de la especificación preliminar, durante la fase del proyecto conceptual, para acomodar definiciones sobre la naturaleza de las dimensiones, métricas, jerarquías y atributos pertenecientes al sistema. En la etapa de perfeccionamiento y validación, un conjunto

innovador de expresiones para la instancia de hechos permite determinase el modelo jerárquico generado que atiende todas las necesidades de las consultas e integración de los datos del sistema. El proyecto lógico recibe como entrada los esquemas dimensionales validando y produciendo un esquema multidimensional suficientemente capaz de optimizar las consultas y serán operadas sobre repositorios. La sexta y última etapa del proceso de desarrollo es la conversión de la vista lógica de los datos para un proyecto físico más adecuados es la implementación en bancos de datos, donde las consideraciones sobre los índices que serán adoptados desempeñan un papel importante.

Metodología Propuesta

La metodología propuesta está estructurada en una secuencia de fases. Cada fase define la aplicación en niveles decrecientes de abstracción, es una medida en que los requisitos del proyecto sean reunidos para formar una **línea de base (baseline)** de los requisitos, un conjunto indexado de características de los requisitos, al ser entregados en una versión específica de la aplicación (LEFFINGWELL y WIDRIG, 2000).

1) **Definición de los objetivos**: los objetivos que se establezcan para el desarrollo de un Depósito de datos, juegan un papel preponderante. Una vez establecidos los objetivos, todo proyecto debe desarrollarse de forma clara y directa. El esclarecimiento de los objetivos son los pilares básicos para el desarrollo de todo proyecto.

2) **Planeación de la Gerencia de Requisitos**: Antes de dar inicio a la tarea de solicitud de requerimientos, las reglas para un efectivo proceso de las especificaciones y gerencia de requisitos precisan ser establecidas. Esas directrices exigen utilizar metodologías y evitar que la ausencia de un patrón común que introduzca inconsistencias a los procesos de los requisitos. Las directrices desarrollan los controles de la adquisición, documentación y administración de los requisitos de sistema, y pueden ser definidas en un término de reglas de negocio, procedimientos y procesos debidamente ajustados y acordados entre las partes interesadas, con el objetivo de esclarecer los siguientes puntos:

(2.a) Papeles y responsabilidades: Un error común en los proyectos de software es la ausencia de las asignaciones claras de tareas y responsabilidades. La definición previa de los papeles deberán ser dentro de un proceso de requisitos, la asignación de las personas responsables para ejercer las funciones es decisiva para que las especificaciones de los requisitos del sistema estén bien definidas.

(2.b) Premisas de integración con fuentes de datos: Deben establecerse reglas claras para el intercambio de datos entre fuentes proveedoras y el Depósito de datos, apropiada con patrones únicos de integración. Periodicidades, prioridades de carga y mecanismos de control de calidad, entre otros aspectos, se irán construyendo una base de las directrices para el proceso de alimentación de los repositorios de los datos.

(2.c) Directrices para la Gerencia de los Requisitos: el equipo de desarrollo, conjuntamente con los demás grupos involucrados son responsables por definir, de forma clara: criterios de localización de los requisitos; restricción de los proyectos en cuanto a los pasos de entrega y puntos de control; criterios para especificación de los requisitos (tipos, atributos, reglas de numeración y procesos para la recolecta).

Los puntos antes mencionados encierran una lista, la cual puede ser ciertamente complementada con otros puntos relevantes, de acuerdo con la naturaleza de la aplicación. En cualquier caso, el plano de Gerencia de Requisitos resultante debe incorporar una relación de índices que describen (a) cómo los requisitos de los sistemas serán identificados y estructurados; (b) cuáles son los criterios para los acompañamientos de la gerencia de los requisitos a lo largo del proyecto; y (c) cuáles son los responsables por esas actividades. Es importante notar que la planeación de la gerencia de los requisitos debe coincidir los puntos generales del proyecto. Esto requiere que las versiones preliminares de los documentos de visión de los Depósitos de datos y Especificaciones de los requisitos multidimensionales serán elaboradas, para consolidar, respectivamente, aspectos fundamentales como:

- Objetivos del Proyecto.
- Puntos Multidimensionales

3) **Especificación de Requisitos:** El suceso de un proceso de ingeriría de requisitos depende de las habilidades, a partir de las declaraciones individuales de los requisitos llenas de información y confusas, verificar una especificación precisa que es comprendida por, y acordada entre todos los interesados (LOUCOPOULOS y KARAKOSTAS, 1995). Para el desarrollo de las aplicaciones de los Depósito de datos, se usa un inicio cíclico para adquisición, representación y validación de los requisitos y producción gradual de una especificación del proyecto. En este sentido un proceso iterativo es más apropiado para el soporte del flujo de actividades. Los requisitos iniciales de un dato Data mart recorre una secuencia de iteraciones según el modelo espiral, a lo largo del cual los requisitos son analizados, negociados entre las partes involucradas, documentados y validados para la garantizar la conformidad con el modelo corporativo de los Depósito de datos. El producto de cada iteración puede ser tanto un conjunto más perfeccionado de requisitos los cuales sirven de entrada para iteraciones subsecuentes, cuando una versión intermediaria (línea de base) de las especificaciones del Data mart que refleja la percepción real de los usuarios al respecto de la aplicación. El producto final de cada iteración tiende a ser cada vez mas próximo de una línea de base de requisitos acordada entre las partes, a medida que la información es cada vez mas perfeccionada es retroalimentada en el ciclo, reduciendo el esfuerzo y el tiempo de gasto con la especificación de la solución, que da una apariencia concéntrica en espiral. La figura 2.2.19 muestra un ejemplo general de un ciclo de especificaciones de requisitos en el Depósito de datos.

Figura 2.2.19
Ciclo de especificaciones de requisitos en el Depósito de datos

Tal como sucede en todo proyecto, sobre todo si involucra técnicas novedosas como son las relativas al Depósito de datos, se deben analizar todas las necesidades y hacer comprender las ventajas que este sistema puede reportar. Es en este punto donde se debe detallar los pasos a seguir, donde el usuario juega un papel preponderante.

4) **Licitación de los Requisitos**: Esta fase implementa un proceso de descubierta de requisitos para el Depósito de datos, con énfasis en aspectos multidimensionales, por medio de la comunicación con las partes interesadas. De la misa forma en que sistemas convencionales, la fase validación de requisitos en los Depósitos de datos requiere conocimientos del dominio de la aplicación de ambos usuarios e ingeniería de software. Para auxiliar estas actividades se propone las siguientes técnicas clásicas de licitación.

_ Entrevistas. Obtención de requerimientos a partir de entrevistas con el nivel gerencial de la institución. El resultado de esta actividad es el documento de requerimientos.
_Workshops. La técnica de workshop permite reunir las partes interesadas de un proyecto por un período corto, mas intensivo, de tiempo para discutir los aspectos relevantes para el desarrollo da aplicación.
_Prototipos. Utilizados dentro del proceso como ejemplos experimentales del proyecto, los prototipos demuestran a los interesados como las funcionalidades de los sistemas irán ayudando en la toma de decisiones. Los Prototipos simulan el comportamiento de los sistemas y esclarecimientos relativos al Depósito de datos.
_ Escenarios. Desenvuelven una serie de escenarios de ayuda para los ingenieros de software a aclarar y detallar los requisitos funcionales del sistema, sobre la forma de caso de uso. Las principales transacciones que componen una aplicación de Depósito de datos tienden a obedecer a un esquema de caso de uso (UML).
_ Análisis de Requisitos No Funcionales. Tratar los requisitos de calidad durante el proyecto de aplicación del Depósito de datos requiere una noción de las diferentes necesidades y

visiones de los negocios de cada parte interesada. Se identifica y define los principales requisitos no funcionales para la aplicación de los Depósitos de datos.

5) **Análisis y Negociación**: El equipo de proyecto debe asegurar que los requisitos definidos con el cliente sean coherentes con el esquema multidimensional proyectado y, principalmente, con las limitaciones de las herramientas OLAP escogidas. En contrapartida, los análisis permiten identificar puntos de ajuste en los conceptos multidimensionales adoptados antes del momento, evitando que fallas puedan comprometer la estabilidad del proyecto en fases avanzadas de su desarrollo.

Una técnica importante para dar apoyo en la ejecución de la fase son las listas de verificación (checklists) de requisitos. Para la lista de verificación definimos una lista de preguntas dirigidas a examinar cada requisito por medio de la literatura de los documentos que los registran. La lista puede ser implementada como una tabla en la cual las líneas representan índices nombrados con identificadores de los requisitos.

6) **Documentación**: Esta fase es el corazón de la metodología. El propósito aquí es proveer una documentación completa y detallada de los requisitos licitados, de manera que se englobe comprensivamente a todos los interesados.
Las actividades y documentos relacionados son:
· Obtención de requerimientos a partir de entrevistas con el nivel gerencial de la institución. El resultado de esta actividad es el documento de requerimientos.
· Relevancia de los sistemas legados, revisión de la documentación existente y entrevistas con personal de dichos sistemas. El resultado de esta actividad son los siguientes documentos:
- diccionario de datos.
- diseño lógico del sistema legado.
- resumen de funcionalidades del sistema legado.

7) **Conformidad de Requisitos**: La conformidad de los requisitos es una etapa particular de las especificaciones de Depósitos de datos. Antes de construir un datamart departamental o de un Depósito de datos de toda la organización, es imperativo que el equipo de desarrollo considere la arquitectura multidimensional propuesta dentro de una visión corporativa de todos los datos de la organización. Dentro de este principio, cuando se espera construir un Depósito de datos que sea robusto y resistente en base a la continua evolución de los requisitos, se debe adquirir una especificación del data mart en el cual las dimensiones y hechos comunes están en conformidad entre todos los datamart.

Una dimensión está en conformidad cuando tienen el mismo significado para toda tabla de hechos en el cual está ligado. La dimensión “Tiempo” es un ejemplo clásico de dimensiones en conformidad, visto que una de las funciones primarias en un Depósito de datos es una zona métrica de hechos a lo largo del tiempo. De manera similar, un hecho está en conformidad con el esquema global de los depósitos de datos, si la misma terminología para la representación de su contenido es usada a lo largo del todo los datamart. Cuando se representa en más de una tabla de hechos, esos hechos deben poseer el mismo formato,

misma regla de formación, misma ley de cálculo y definidos por el mismo contexto dimensional.

Definimos que un requisito está en conformidad si éste es común a varios (o todos los) data mart y es descrito idénticamente en cada una de las diferentes visiones de los Depósitos de datos.

Los requisitos en conformidad trazan los siguientes beneficios para la especificación de los sistemas Depósito de datos.

- *(i)* *Evita redundancia y ambigüedad entre los requisitos que permiten los Depósitos de datos como un todo.*
- *(ii)* Permite que dimensiones comunes sean relacionadas con múltiplos de hechos en el mismo espacio de bancos de datos.
- *(iii)* En conjunto con el uso orientado a escenarios, posibilita la reutilización de conocimientos previamente acordado por el espacio del proyecto, promoviendo, de esta forma, la mejor calidad.
- *(iv)* Mejora la consistencia de las interfases de usuario y de los contenidos de los datos agregados donde cualquiera que sea hace uso de los modelos comunes (en conformidad)
- *(v)* Posibilitan la operación drill-across entre data mart, la recuperación de los datos actuales a partir de las y tablas de hechos localizadas en diferentes visiones dentro de los Depósitos de datos de organización.
- *(vi)* Facilita la integración requerida entre datamarts, permitiendo que la arquitectura multidimensional trabaje como un todo único.
- *(vii)* Propicia escalabilidad y facilita la evolución del Depósito de datos.
- *(viii)* Facilita la adherencia de los patrones del proyecto y organizaciones.

8) **Validación de Requisitos**: Para validación de lo que fue definido como especificación para el Depósito de datos, se propone unir las técnicas de Revisión y prototipos como una estrategia efectiva para detectar y remover defectos en las especificaciones. Durante la reunión de revisión, la línea de base final de los data mart es representada para todos los desarrollos, y descrita en términos de sus requisitos funcionales y de calidad. El prototipo desarrollado en las herramientas OLAP auxilia el reconocimiento para usuarios de los aspectos arquitectónicos que implementan los requisitos especificados.

Cuando los problemas son identificados, la especificación es revisada por el contexto del requisito funcional, no funcional o del dominio multidimensional que dan origen a la inconsistencia. El tiempo de validación debe generar inmediatamente una lista soluciones en respuesta a cada una de los casos, y acordar entre los involucrados las soluciones propuestas. Otro de los factores importantes durante la validación de los requisitos de los Depósitos de datos es incluir especialistas de dominio que no estuvieran relacionados con la ingeniería de los requisitos del sistema.

9) **Diseño Conceptual Multidimensional**: Esta etapa se enfoca en la construcción del diseño conceptual del Depósito de datos a partir de los requerimientos identificados en y la documentación existente del sistema legado. Las actividades y documentos relacionados con esta etapa son:

· Elección de la herramienta a usar para representar el diseño conceptual del modelo de datos del Depósito de datos.
· Construcción del diseño conceptual multidimensional.
· Construcción de la tabla de correspondencias entre el diseño conceptual del sistema legado y el diseño conceptual del Depósito de datos.

10) **Diseño Lógico Multidimensional**: Esta etapa abarca la construcción del diseño lógico multidimensional que se basa fundamentalmente en el diseño conceptual multidimensional construido en la etapa anterior y los requerimientos multidimensionales identificados. Las actividades y documentos relacionados con la etapa son:

· Definición de la arquitectura de los depósitos de datos del Depósito de datos.
· Definición del tipo de almacenamiento dependiendo del uso de tecnología relacional o multidimensional.

En el caso del uso de tecnología relacional esta actividad incluye el diseño del esquema relacional del Depósito de datos.

11) **Diseño de los procesos de carga y actualización** que poblarán al Depósito de datos, que implican estrategias de extracción, limpieza, notificación de errores y administración de cambios detectados en los datos provenientes de los sistemas legados.

El proceso de carga inicial del Depósito de datos genera las instancias de las estructuras de datos del mismo con la información proveniente de los sistemas legados, depurada, integrada y eventualmente resumida, existente en el momento de ejecución del mismo.

El proceso de actualización tiene como cometido la actualización de la información del Depósito de datos con los nuevos datos generados o modificados en los sistemas legados a través del tiempo con la frecuencia adecuada. La actualización de un Depósito de datos determina el uso efectivo de los datos recolectados y resumidos desde los orígenes. La calidad de los datos provistos a quienes toman decisiones, depende de la capacidad del Depósito de datos de propagar los cambios hechos en los sistemas legados en el tiempo definido como conveniente para que la aplicación tenga datos considerados “frescos” en todo momento.

Los documentos construidos en esta etapa son los siguientes:

· Diseño lógico multidimensional.
· Tabla de correspondencias entre el diseño lógico del sistema legado y el diseño lógico del Depósito de datos.
· Diseño de procesos de carga y actualización.

12) **Implantación**: En esta última etapa se selecciona la herramienta a usar en la implantación de los procesos de carga y actualización del Depósito de datos. Además de realizarse la implantación de dichos procesos. Como resultado de la ejecución de los procesos de carga y actualización se obtienen las estructuras de datos del Depósito de datos convenientemente actualizadas y se genera un resumen de errores detectados durante el proceso de carga o actualización. El uso de este resumen de errores facilita la corrección de los datos en los sistemas legados.

Implantación de un Depósito de datos lleva implícito los siguientes pasos:

- Extracción de los datos del sistema operacional y transformación de los mismos.
- Transformación: validar, limpiar, integrar, y fechar.
- Carga de los datos validados en el Depósito de datos. Esta carga deberá ser planificada con una periodicidad que se adaptará a las necesidades de refresco detectadas durante la fase de diseño del nuevo sistema.
- Explotación del Depósito de datos mediante diversas técnicas dependiendo del tipo de aplicación que se de a los datos:

  *On-Line analytical processing (OLAP)

  *Decisión Support Systems ó Información de Gestión

  *Visualización de la información

  *Data Mining o minería de datos

13) Y por último realizar las revisiones pertinentes, lo que implica realizar **pruebas** de todo tipo.

La figura 2.2.20 representa un alto nivel de metodología incluyendo de forma general alguna de las fases descritas anteriormente.

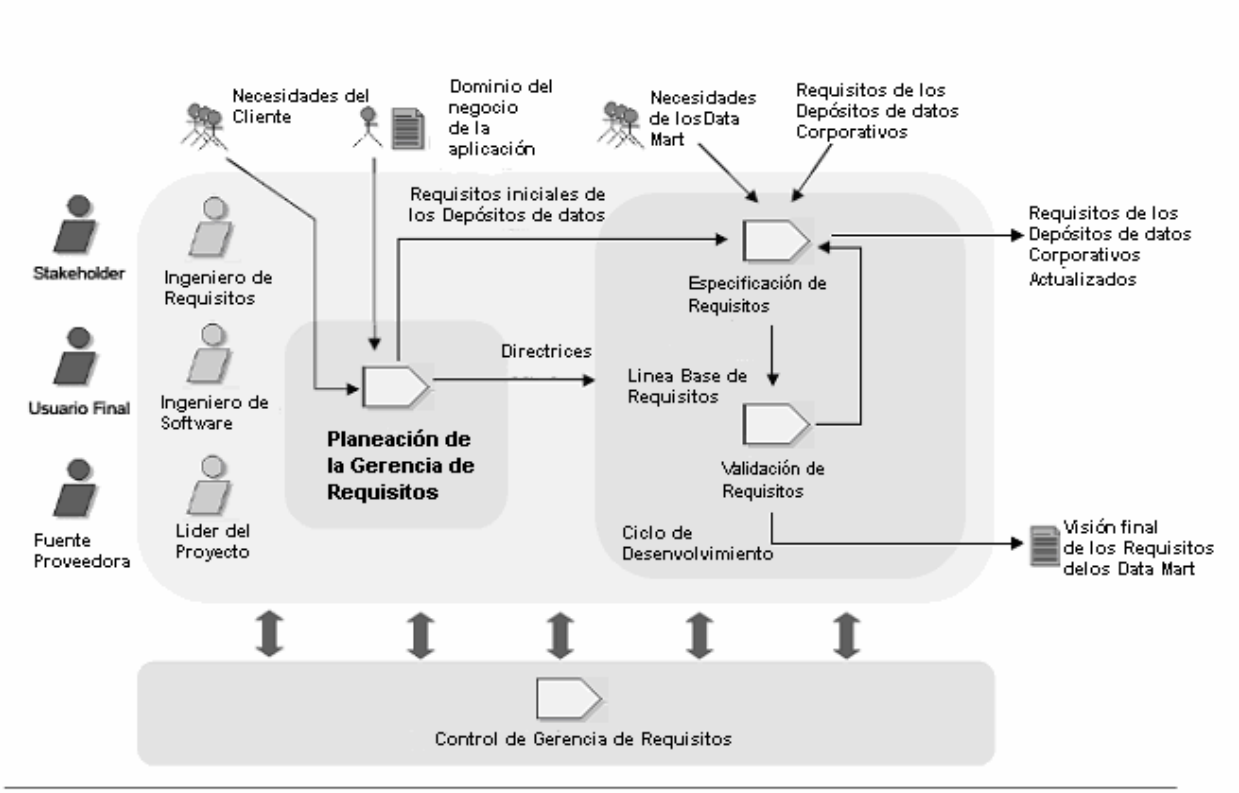

Figura 2.2.20
Descripción de un alto nivel de Metodología.

2.3 Los Datos

Objetivo: El alumno modelará datos bien definidos, integrados y consistentes, explicará el papel de los metadatos en el ciclo de vida de los depósitos de datos, el papel del directorio/catalogo en la empresa, el proceso de extraer datos de aplicaciones operacionales.

Para entender mejor los conceptos de calidad en los datos, Metadatos, el uso de directorios y catálogos damos una definición general para crear una definición que sea estándar para el estudiante, estos conceptos además se analizan a fondo en los subtemas posteriores para un mejor entendimiento de la transformación de datos.

Como es sabido para el estudiante un dato puede considerarse como el elemento mínimo que puede proporcionar una definición, es decir un elemento como 32 puede no decirnos mucho para algunos tal vez es un numero entero, pero para otros puede ser el resultado de una operación de que depende esta respuesta, depende en gran medida a la cadena de caracteres asociados a el, que pasaría si en lugar de 32 se asocia con calle 32, la respuesta a este pequeño análisis ya que no fue ninguna de las dos operaciones descritas mas bien la parte asociada fue la clave para darnos el criterio para el resultado, o lo que es aun la clave para definir la calidad de los datos es la correcta interpretación del resultado.

2.3.1 Calidad de los datos:

Las organizaciones empresariales, de educación, medicas y en general toda aquella institución que realice un manejo masivo de información, han notado la necesidad de encontrar la información adecuada para poder almacenarla, manejarla de forma rápida y precisa como un punto crítico para tomar decisiones acertadas y de esta manera atender y enfrentar a los cambios que involucran la competencia en su entorno o circulo empresarial. Las diferentes organizaciones mencionadas como se ha mencionado a lo largo de los capítulos anteriores utilizan Sistemas de Ayuda que consultan la información del Depósito de Datos o Bodegas de datos. Pero una de las necesidades básicas a cubrir es la calidad en los datos es decir; ¿Puedo obtener información pero que tan fiable veraz o cierta es?, este es uno de los cuestionamientos principales que afectan también la toma de decisiones pero de que depende el resultado?, Un buen desarrollo de análisis de nuestro sistema utilizado para proveer de información en nuestra base de datos o mejor conocido en el capitulo anterior como origen de datos nos puede llevar a un buen desarrollo para la aplicación de deposito de datos ya que de lo contrario si el diseño de nuestro sistema no tuvo una visión del crecimiento posterior y no acepta el crecimiento el desarrollo del deposito de datos si es realizable pero no confiable, ante esta situación ¿Que debemos hacer?

Para contestar a estas preguntas tomemos las siguientes consideraciones.

La calidad de los datos que a su vez en conjunto y organizada representan información viene determinada por la calidad de nuestra base de datos o almacén de datos, al inicio del capitulo describimos como se puede representar un dato a esto también se le conoce como calidad de la presentación de los datos. Para que los datos del almacén o base de datos sean un parámetro fiable de lo que ocurre en los movimientos y operaciones de las organizaciones actual, pueden también depender de la forma en como sean interpretados.

Para poder mencionar que nuestro origen de datos es de calidad en general hay que considerar aspectos como el entorno que rodea nuestro origen de datos, tomando como ejemplo una base de datos operacional, podemos mencionar los siguientes aspectos o consideraciones a seguir:

- la calidad del SGBD (Sistema Gestor de Base de Datos) relacional o multidimensional que lo soporta
- la calidad del modelo de datos (tanto conceptual, lógico como físico)
- la calidad de los propios datos contenidos en el almacén.

Dado que la calidad tiene componentes objetivas y subjetivas, es necesario catalogar los requisitos de calidad de datos de los usuarios según unas determinadas dimensiones de calidad. La mayoría de los autores intentan definir el concepto de calidad de datos y catalogar las dimensiones de calidad en función de unos determinados criterios, como pueden ser el ciclo de vida de los datos o los tipos de investigación realizadas, o simplemente la forma en la que se usan los datos. Pero todos están de acuerdo en que la calidad de datos es un concepto multidimensional que comprende distintos aspectos según las necesidades de los consumidores de datos o de los diseñadores de sistemas, y que dichas necesidades deberían introducirse desde las fases más tempranas del desarrollo a modo de requisitos de calidad de datos, con herramientas o interfaces. La tabla 2.3.1 muestra las dimensiones de calidad más importantes.

Consideraciones generales para Asegurar la calidad en los datos	
DIMENSION	**DEFINICION**
Accesibilidad	Los datos deben ser de fácil acceso y rápidamente recuperables
Manipulación	Los datos deben ser fácilmente manipulables, tanto en actualización, lectura o escritura de los mismos
Disponibilidad	La disposición de los datos debe ser en cualquier instante que sea solicitado
Interpretación	Los datos deben ser fácilmente comprensibles, en el idioma correcto y con la simbología correcta

Seguridad	El acceso a los datos debe de estar correctamente protegido, y clasificado para su consulta, actualización o eliminación.

Tabla 2.3.1 dimensiones de calidad en los datos más importantes

Tomando en cuenta la tabla anterior se hace notar que estas dimensiones son quizá las mas significativas y las mas recurrentes al analizar los datos y sobre todo para el aseguramiento de la calidad de los mismos, no obstante se hace necesaria una metodología para lograr el aseguramiento de una calidad del 100% ya que esto ayuda de gran manera a logar el objetivo principal de identificar las dimensiones de calidad que mejor describen esos requisitos, para después obtener métricas a partir de ese conjunto de dimensiones, lo que ocasionaría una manera mas ordenada de analizar los datos dentro de un deposito de datos.

La metodología que a continuación se muestra, se compone de un total de cuatro fases independientes diferenciadas. Cada una de estas fases está a su vez estar formada por una serie de o actividades. Se recomienda seguir las fases de manera consecutiva, aunque sabemos que en alguna ocasión debe de omitirse un paso de estas fases ya que no son contempladas en los objetivos de la medición. Las fases que componen esta metodología propuesta en la figura 2.3.1, son las siguientes:

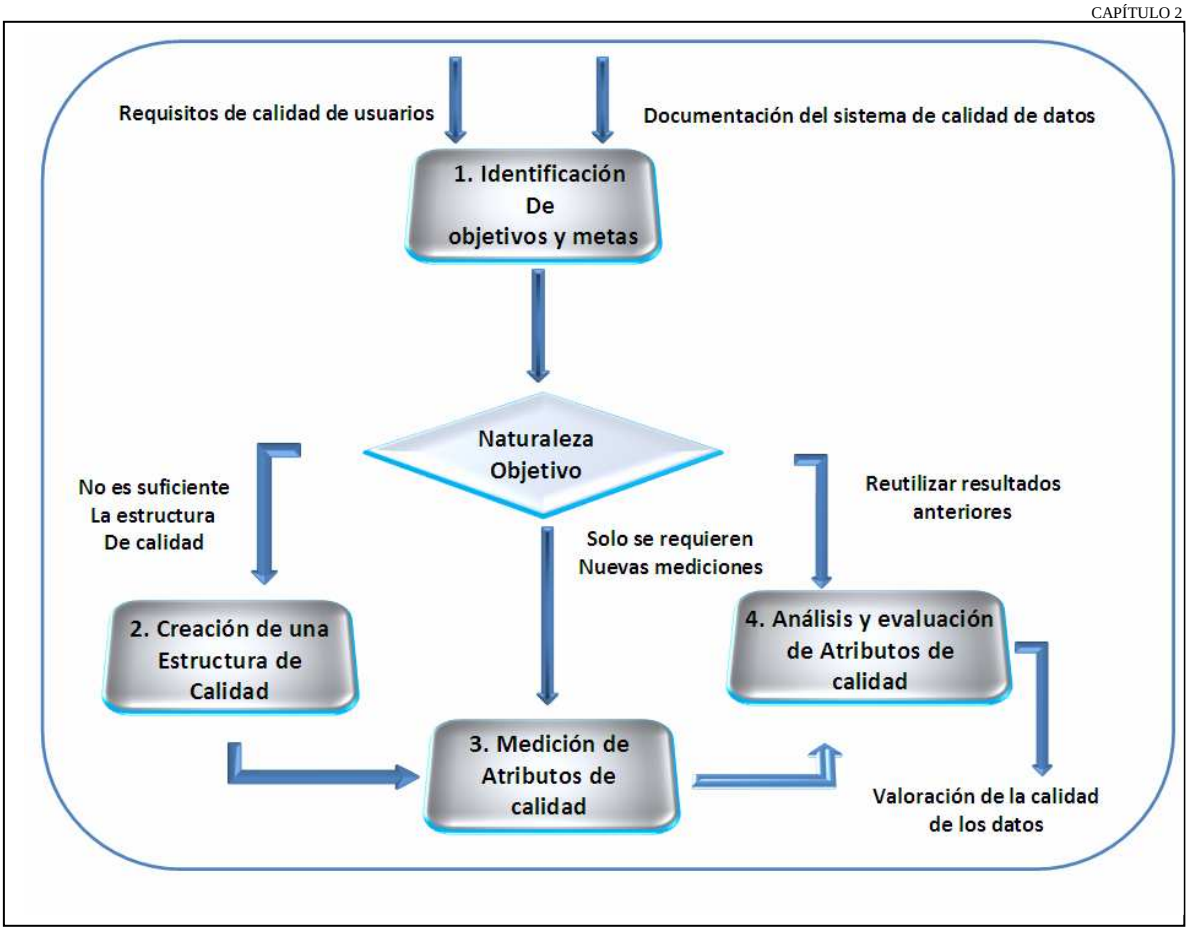

Figura 2.3. 1
Metodología para la evaluación de la Calidad

Como se identifica en la figura 2.3.1 describimos brevemente las fases involucradas en ésta metodología.

Lo primero que hay que tomar en cuenta es la identificación de los objetivos y de las medidas. En esta fase de análisis es donde a partir de los requisitos de calidad de los usuarios se obtendrían una serie de productos de trabajo como resultado de la realización de cada una de las siguientes actividades esta fase se muestra en la figura como fase 1.

- Determinar el objetivo de la medición. Se trata de determinar las razones por las que se quiere medir el nivel de calidad de datos.

- Determinar los parámetros e indicadores de calidad. A partir de los requisitos de los usuarios se identifican las dimensiones y métricas de calidad de datos más significativos para acotar el problema de calidad de datos.

- Localizar los datos a valorar. Esta actividad se divide en las siguientes subactividades.

- Definición de criterios de calidad. Se trata de establecer criterios de valoración para juzgar la bondad de un dato y de definir criterios de evaluación para determinar la bondad del conjunto de los datos.

En la fase 2 representada en la figura Creación de una estructura de calidad. Es la fase de diseño, donde el objetivo es dotar al almacén de datos de una estructura para guardar los valores que más tarde se recogerán para las medidas de calidad. Medición de los atributos de calidad. Representados en la figura como fase 3 donde una vez que el almacén de datos disponga de una estructura para guardar las medidas de las dimensiones de calidad, esta fase consiste en recoger valores para dichas medidas en las dimensiones especificadas. Puede llegar a ser necesario que para algunas dimensiones de calidad se deba conocer el valor del dato real y compararlo con el del dato almacenado. En función de la cantidad de datos y del nivel de calidad exigido puede ser necesario medir los valores de todos los datos o seleccionar por muestreo solo una parte de esa totalidad. En cualquier caso estas mediciones se guardaran en el almacén de datos.

En la fase 4 se representa el Análisis y Evaluación de los valores de los atributos de calidad. En esta fase, se someterá los valores individuales medidos en la fase anterior a los criterios de valoración para determinar el grado de bondad de un dato y según el número de datos con calidad y los criterios de evaluación establecidos se juzgarán si esos datos tienen o no el grado de calidad deseado. Si es así, se certifican los datos como válidos para la aplicación. En caso contrario se desechan como inválidos, procediendo posteriormente como mejor convenga: corrección de los datos existentes o captura de nuevos datos.

2.3.2 Metadatos

Un metadato por definición se considera como datos altamente estructurados que describen información, describen el contenido, la calidad, la condición y otras características de los datos, una definición no muy coloquial pero muy utilizada es "Información sobre información" o "datos sobre los datos" pero porque se da esta ultima definición, volvamos al capitulo I en donde se menciona que la finalidad de un deposito de datos es la obtención de datos para la toma de decisiones a un nivel gerencial de ahí la definición información sobre información, ya que el origen de los datos finales para la toma de decisiones proviene de otra fuente de datos, suena redundante pero en la practica podemos notar que la información final para esta toma de decisiones tiene su origen en bases de datos operacionales, archivos planos u otra fuente utilizada en la operación de cualquier organización que pueda proveer información para el desarrollo de un deposito de datos por lo que la definición datos sobre datos puede ser no muy redundante finalmente por lo que podemos concluir que un metadato es utilizado para describir conjuntos de datos.

Por lo tanto lo estudiado en el subtema anterior de calidad de datos tiene alta correspondencia con la información contenida en los metadatos como un conjunto para describir información esto con la finalidad de facilitar su recuperación, autentificación , evaluación y preservación o interoperabilidad, pero como se logra obtener un metadato.

Para entender mejor la definición de metadato y su utilización se muestran algunos ejemplos de metadatos:

- El resumen de un documento.
- El catálogo de una base de datos.
- Las palabras extraídas de un texto.
- Las fichas de clasificación en cualquier formato (ISBD, MARC).
- Las páginas amarillas.

Como identificamos que son ejemplos de metadatos, partiendo de la definición como un conjunto de datos, podemos decir que para el primer ejemplo resumen de un documento que proviene del análisis de n cantidad de bibliografía o de documentos para la obtención de datos específicos como el resultado de un resumen es considerado un metadato.

Contenido de un metadato

- Identificación
- Calidad de los datos:
- Organización de los datos espaciales
- Referencia espacial:
- Entidad y atributos
- Distribución:
- Referencia de los metadatos
- La estructura de los datos

2.3.3 El papel de un directorio/catálogo

A lo largo de nuestro estudio de la informática a través del tiempo se ha intentado plasmar en los SI (sistemas de información) el conocimiento que los usuarios poseían sobre su dominio de aplicación, almacenando datos relativos al mismo.

En las primeras etapas de su aparición, los SI estaban soportados básicamente por un conjunto de programas en los que se embebía la descripción de los datos, así como algunas de sus características y también, aunque en una mínima parte, sus restricciones, es decir, su semántica.

Como se ha visto estos sistemas orientados al proceso resultan, muy difíciles de mantener y demasiado complejos, planteándose la necesidad de centralizar las descripciones de los datos para conseguir sistemas más coherentes, eficientes y adaptables a los cambios.

A fin de atender estos objetivos de toma de decisiones surgen los directorios de datos cuya finalidad es la de describir las siguientes funciones de operación.

- dónde y cómo se almacenan los datos de la base,
- el modo de acceso y
- otras características físicas de los mismos

De esta forma el directorio de datos atiende las peticiones de los programas y de los procesos. Un directorio de datos contiene, en definitiva, las especificaciones necesarias para pasar de la representación externa de los datos a la representación interna de los mismos, y ha de estar siempre en un formato legible por la máquina. El objetivo principal del directorio de datos es transmitir al SGBD la información necesaria para poder acceder a los datos contenidos en la base.

Con independencia de los directorios surgen los llamados diccionarios de datos, donde se reúne la información sobre los datos almacenados (descripciones - narrativas y técnicas -, estructuras, consideraciones de seguridad, edición y usos de las aplicaciones de los mismos, etc.) y que los usuarios necesitan para comprender el significado, esto es, el aspecto lógico de los datos.

Al principio, estos diccionarios eran simplemente listas manuales elaboradas por los usuarios, o manejadas por un conjunto de programas *hechos a medida* por el administrador de la base. La complejidad de los diccionarios de datos se ha ido haciendo cada vez mayor, expandiendo el número de funciones que se le asignaron originalmente, así como el entorno operativo en el que se utilizaban. En los años 70 aparecen varios paquetes software de este tipo: DB/DC DATADICTIONARY, DATAMANAGER, DATA-DICTIONARY SYSTEM, ADR/DATADICTIONARY, LEXICON, etc.

Algunas veces estos paquetes llevan a cabo tanto las funciones de diccionario como las de directorio, denominándose entonces, diccionario/directorio de datos (DD/D).

Con la amplia difusión que experimentan los SGBD relacionales, se extiende el concepto de catálogo, que realiza funciones de directorio y diccionario. El catálogo consta de un conjunto de tablas (relaciones) que almacenan información sobre la

base de datos y a las que accede mediante el mismo lenguaje (SQL) que se utiliza para acceder a las tablas que poseen datos de las aplicaciones.

Por otro lado, la creciente complejidad de la concepción y diseño de los SI ha llevado a la construcción y comercialización de herramientas conocidas bajo el nombre genérico de CASE (Computer Aided Software/System Engineering), que contienen un diccionario llamado **enciclopedia** o **repositorio**, donde se almacena los datos generados durante el ciclo de vida de un SI: esquemas, grafos, matrices, información relativa a la gestión de proyectos, gestión de configuraciones, etc. Aunque este repositorio no suele ser activo, en algunos casos la herramienta CASE facilita instrumentos para cargar directamente las descripciones de los datos obtenidas en la etapa de diseño, en los catálogos propios de los SGBD más extendidos.

A finales de los ochenta apareció un nuevo concepto, el de diccionario de recursos de información (DRI), que pretende ser el eslabón final en la evolución de los *almacenes de datos*. El DRI constituye el depósito integrado de todos los datos sobre la organización, automatizados o no, que son utilizados para efectuar las labores de planificación, control y operación que permitan a la empresa cumplir sus objetivos.

Los DRI engloban, de algún modo, las capacidades y funciones de todos los *almacenes* de datos anteriores. Las siglas inglesas son IRD (Information Resource Dictionary), que son las utilizadas por los organismos de estandarización ANSI e ISO. En la siguiente figura se presentan de forma gráfica, los distintos elementos que hemos analizado tal como aquí se han expuesto en un intento por precisar y clarificar unos conceptos respecto a los cuales no existe en la actualidad consenso y cuya terminología, resulta, a veces, bastante confusa. De hecho, algunos expertos denominan repositorio a lo que hemos llamado DRI, dejando este último término para los estándares definidos por ISO o por ANSI.

2.3.4 Transformación de los datos

La exploración y análisis, en forma automática o semi-automática, de grandes volúmenes de información para la detección de patrones de comportamiento es lo que se denomina minería o explotación de datos, también conocido por su vocablo en inglés data mining. Se define minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Se puede definir el ciclo de vida de la explotación de datos a partir de las siguientes fases (figura3.3): [a] obtención de datos a procesar, [b] transformación de los datos para que pueda ser utilizado, [c] aplicación de la técnica de explotación de datos y [d] evaluación de los resultados obtenidos. La fase transformación de los datos, es la que insume mayor tiempo, llevando aproximadamente el 60% del esfuerzo de desarrollo. En este trabajo se propone un método de transformación de datos y se detallan las

características necesarias que debe poseer el entorno de trabajo para la automatización del mismo. En la sección a) se describe brevemente el ciclo de vida de la explotación de datos y se menciona el principal problema de la preparación de datos. En la figura 2.3.2 se detalla el ciclo de vida de la explotación de datos, y los pasos que se encuentran relacionados con la transformación de datos.

Figura 2.3. 2
Ciclo de Vida de la explotación de datos (datamining).

a) Ciclo de vida de la explotación de datos

En ciertos casos el disparador de un proceso de explotación de datos es la detección de un problema y la necesidad de corregir ese comportamiento anómalo; en otros no es necesario observar nada anormal solo se aplica el proceso de minería para detectar patrones desconocidos. De ser este último el caso aplicado, los resultados obtenidos en la explotación de datos deben ser sometidos a un proceso de validación conocido como minería de reglas de negocio, o en su forma inglesa *business rule mining*, el cual nos permitirá validar o crear una nueva regla de negocio. Las fases del ciclo de vida a seguir se describen en las siguientes subsecciones.

a.1. Obtención de datos a procesar

Este punto siempre parece mucho más sencillo de lo que realmente es, algunos de los problemas que se suelen encontrar es la falta de acceso a los datos, ya sea por razones de seguridad o por no encontrarse disponibles, es decir los datos se encuentran resguardados. Si son cuestiones de seguridad de la información, una vez superada las cuestiones burocráticas, ya estaremos en condiciones de acceder a los mismos. En caso de que los datos se encuentran resguardado el primer problema al que nos enfrentamos, es obtener el espacio suficiente para recuperar los mismos, de estar en alguna base de datos también es necesario obtener los recursos para poder acceder a la misma. Con estos pasos realizados, la próxima tarea es una primera revisión de los datos obtenidos para conocer sus características.

a.2. Transformación de los datos para que pueda ser utilizado

El primer paso para la preparación de datos es conocer el problema a resolver para lo cual se beberán Incluir como actividad preliminar la comprensión del dominio o negocio: el propósito es asegurar el entendimiento del negocio y objetivos del proyecto, o al menos hacia que objetivo queremos llegar, sin esto nos resulta imposible conocer los datos que debemos extraer. Por otra parte debemos conocer la forma en que se debe presentar la información al modelo seleccionado para la explotación de datos, con estas dos precisiones se puede comenzar a recolectar la información y trabajar con ella. Cuando se esta trabajando en explotación de datos, se están utilizando datos que representan hechos de la vida real, esos datos deben se preparados para que las herramientas de explotación puedan trabajar con ellas. La preparación de los mismos no es un proceso automático, por lo cual es necesario aplicar nuestro conocimiento para generar el conjunto de datos necesario para poder aplicar un modelo de explotación. Por lo antes dicho podemos definir como el principal objetivo de la preparación de datos (la vista minable o dataset y su descripción) es tomar información manipularla, transformarla y presentarla para que pueda ser procesada por un modelo de minería de datos. Para conocer que transformaciones debemos realizar y como la debemos presentar nos debemos hacer dos preguntas fundamentales: ¿Qué solución debemos obtener? y ¿Que técnica de explotación utilizaremos? La primera cuestión la relacionaremos con las características y cantidad de información que deberemos manipular, y la segunda cuestión, la forma en que se debe presentar la información para la explotación. Con los datos accesibles y hechos la primera revisión de los mismos los pasos comunes en la preparación de datos, se puede definir como:

Enriquecer la información: Luego de analizar la información y teniendo respuesta a las preguntas antes generadas, se plantea la posibilidad de agregar datos a los ya obtenidos, pues la información con la que se cuenta no cumple con todos los requisitos necesarios para poder generar un conjunto de datos que sea aceptado por el modelo.

Obtener casos testigos: Esto se puede convertir en un proceso muy tedioso, la obtención de estos casos testigo nos permitirán definir si el modelo al que lo vamos a aplicar es viable o no en relación al conjunto de datos que tenemos.

Determinar la estructura de los datos: Para poder entender este concepto es necesario definir el término conjunto de datos, este hace referencia a los datos que serán utilizados por el modelo de minería de datos para encontrar patrones. La estructura de datos hace referencia a la forma en que las variables se relacionan unas con otras en los conjuntos de datos. Es en esta estructura donde se buscarán relaciones y patrones de comportamiento.

Construir el modelo de entrada de datos: Se puede decir que hasta este paso en lo que nos hemos centrado es en obtener y conocer la información disponible. En este paso lo que se determinarán los procesos que se seguirán para el modelado de los datos, entre los cuales podremos nombrar:
[i] normalización
[ii] tratamiento de los valores nulo o vacíos
[iii] detección de series (las más comunes de tiempo)
[iv] reducción del ancho de los datos, es decir la cantidad de columnas y
[v] reducción de la profundidad, la cantidad de registros.

e. *Inspeccionar los datos*:
Una vez generada todas estas transformaciones, el minero de datos necesitan evaluar el resultado para poder determinar si de las transformaciones hechas al conjunto de datos lo hace viable para que el modelo elegido lo pueda procesar.

a.3. Aplicación de la técnica explotación de datos seleccionada

Luego de realizar todas las transformaciones se procede a modelar los datos en función de la técnica de minería de datos elegida para actuar sobre la vista minable obtenida anteriormente, existen diferentes técnicas a saber: de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, entre otras. Dependiendo de la técnica, se ejecutarán una o varias ejecuciones con uno o varios conjuntos de datos.

a.4. Evaluar los resultados obtenidos

Con los resultados obtenidos de las ejecuciones del modelo, se centra la atención en detectar y poder comprender el resultado de los mismos. Esta tarea no es para nada sencilla e insume gran cantidad de tiempo, esto se debe en muchos casos a la complejidad de los resultados obtenidos. De este análisis es de donde se puede

concluir, que los resultados no han sido los esperados, por varios motivos, la técnica no es la correcta para la solución del problema; otra posibilidad es que el conjunto de datos, no haya sido el adecuado, que se deba generar otro conjunto de datos, para validar los resultados obtenidos en la primera modelado; o que el modelo no se ajuste a los requerimientos de negocio, es por estas razones que el último paso sea comenzar con el ciclo nuevamente. El modelado de la explotación de datos es un proceso de aproximación cíclica, el cual se debe ir mejorando a medida que se conoce más de la información con la cual se está trabajando. Es por esto que es necesario reiniciar el ciclo has que la información obtenida satisfaga el requerimiento que la produjo.

b) Método de transformación propuesto

El Método Unificado de Transformación (MUT), es el resultado de la experiencia adquirida en el procesamiento de grandes volúmenes de información sobre distintas plataformas, desde equipos IBM 390 a redes de computadoras personales que poseen alguna de las distintas versiones de Microsoft Windows existente, pasando por el AS 400 y diversas versiones de Unix. En esta categorización se hace necesario agregar la nueva generación de aplicaciones orientadas a sistemas de planeamiento de recursos empresariales, del vocablo en inglés *enterprise resourse plainning (ERP);* esto se debe a que debido a su complejidad, el usuario puede abstraerse del sistema operativo con le cual trabaja su computadora personal y solo operar dentro del entorno que le facilita el ERP, entre estos a modo de paradigma mencionaremos SAP.

b.1. Requerimientos para la aplicación de la metodología

El encargado de la transformación de datos debe tener conocimientos básicos sobre la notación que implemente del lenguaje unificado de modelado, de su vocablo en inglés *unified modeling language,* de aquí en mas UML, específicamente se hace referencia a los casos de usos y los diagramas de secuencia. El explotador de datos conoce los formatos de los archivos con los cuales deberá trabajar, además del formato de salida obtenidos por el proceso de transformación de datos y tiene permiso a los mismos y son accesibles el conjunto de datos de entrada que deberá transformar para poder ingresar los datos al modelo de minería de datos; además se cuenta con espacio suficiente para el proceso de los datos, vale aclarar que esta metodología antepone la agilidad y velocidad de procesamiento en detrimento del espacio de almacenamiento de archivos intermedios. Esta característica del método se basa en que el espacio físico de almacenamiento hoy en día es lo mas accesible y mas barato en comparación con recursos de memoria y procesador.

b.2. Descripción de la metodología

Conociendo las dos preguntas fundamentales para el proceso de transformación, es decir, donde estoy y ha donde quiero llegar, el método recomienda la

aproximación gradual al objetivo final basado en un conjunto de pasos que se basan en: análisis de los requerimientos de transformación, modelado de las transformaciones, codificación pruebas, evaluación y nueva iteración. El principal objetivo del método propuesto no es realizar todas las transformaciones en un solo paso, sino que se realizan pequeñas modificaciones a los datos, se realizara una prueba de regresión completa de lo hecho hasta ese momento y una vez evaluada la misma de ser satisfactoria se volverá a reiniciar el ciclo con la próxima transformación a realizar. En las siguientes subsecciones se detallan los pasos mencionados anteriormente.

b.2.1. Fase de análisis de los requerimientos de transformación:

El primer paso que se deberá dar es el de recabar información acerca de que es lo que necesitamos obtener, es decir, conocer el formato de debe tener nuestro conjunto de datos para poder ser ingresado al modelo elegido para la minería de datos. En este paso se aplican las técnicas mas adecuadas que faciliten la extracción y educción (p.ej.: entrevistas), el único requisito al finalizar este paso, es poseer la especificación detallada del formato de datos para el modelo. Cabe mencionar que es posible encontrarnos ante la posibilidad que la misma persona que se encuentra encargada de las transformaciones sea la persona que ha definido el modelo de minería de datos, en tal caso solo se especificará el formato de archivo. Como resultado de la educción de requerimientos, se obtendrá la especificación del formato de archivo (ver Tabla 1) que se presenta modo de ejemplo

Con el formato de archivo de ingreso al modelo de datos ya especificado, se abren dos cursos de acción:

[a] comenzar a recabar la información necesaria para poder detectar el origen de datos para la creación del archivo solicitado ó
[b] con el formato de archivo ya especificado, volver sobre el modelo de minería de datos seleccionado, con la finalidad de detectar los requisitos del conjunto de datos para su uso, es decir, cantidad de registros necesarios para su aplicación, cantidad de conjuntos de datos necesarios para su validación o entrenamiento. En el primer curso de acción se deberán seguir los siguientes pasos:

a. *Repetir la técnica de entrevista, para detectar el origen de datos:* En esta etapa de entrevistas, lo que se observa es que la cantidad de personas involucradas es mucho mayor de lo que uno a priori puede suponer. Entre las cuestiones a tener en cuenta podemos citar:
_ Se deberá entrevistar al administrador de la base de datos, para conocer la antigüedad de los datos que se encuentran en línea en la base de datos.
_ Otra cuestión a manejar con el administrador es determinar las posibles plataformas donde se encuentran los datos, de ser todas almacenadas en Bases de Datos, cuales y que versiones.

_ Otro punto es solicitarle el diagrama de entidad - relación (DER), para conocer la estructura de las tablas y sus campos.
_ De esto surgen dos implicancias, por un lado, en función de la cantidad que se encuentran en línea, se deberá entrevistar, al encargado del resguardo de los mismos, para conocer desde hace cuanto tiempo se tienen datos resguardados y su posibilidad de acceso. La segunda implicancia es, del análisis del DER, surgirán dudas sobre el origen de los datos esto hará se conserven entrevistas con los responsables de los diversos sistemas. Aquí será necesario realizar entrevistas grupales para resolver las inconsistencias propias de todo modelo de datos.

b. *Acceso a la información:* En la medida que se detecte las fuentes de los datos, se deberán tomar todos los recaudos para poder acceder a los mismos, algunas de las cuestiones que se deberá resolver son:
_ Cuestiones referentes a la seguridad de datos, formalizar los pedidos de acceso a la información
_ Si los datos se encuentran resguardados, es necesario disponer del espacio para su recupero y calcular el tiempo que llevara esta tarea, que puede ser muy significativa

b.2.2. Fase de Modelo de las transformaciones

En esta etapa es donde se diseñan las transformaciones necesarias para que los datos tomados del origen lleguen a la estructura requerida por el modelo de minería de datos. Esto lo realizaremos utilizando Casos de Usos. Los casos de uso, del vocablo inglés *use case,* constituyen el concepto central del método OOSE de Ivan Jacobson, uno de los padres de UML. Los casos de uso representan, el medio para describir el carácter funcional de los objetos, son una representación orientada a la funcionalidad del sistema y permiten modelar las expectativas del usuario. Existen tres conceptos fundamentales en el modelado de los casos de uso: los actores que utilizan el sistema, los casos de uso y los escenarios. Los actores pueden ser de dos tipos: [a] humanos, usuarios de los programas y [b] software, programas que se comunican con nuestro sistema. Desde el punto de vista del sistema exista dos tipos de actores: [a] los actores primarios, que son los que utilizan el sistema y [b] los actores secundarios, que tienen funciones de administración y mantenimiento del mismo.

Los casos de uso representan la utilización del sistema por parte de los actores. Los casos de uso se pueden organizar desde mayor grado de abstracción hasta el detalle que se crea necesario. La representación de los casos de uso puede ser textual o gráfica. Un ejemplo de una representación textual es:

Un escenario es una serie de eventos ordenados en el tiempo, que simulan una ejecución particular del sistema, de manera general, un escenario utiliza dos tipos de conceptos: [a] objetos que normalmente forman parte del sistema y [b] eventos emitidos y recibidos por los objetos implicados en el escenario.

Los escenarios permiten experimentar las ejecuciones del sistema, por lo que resultan muy para las pruebas y el mantenimiento. El modelado de las transformaciones tendrá como actor al controlador, que es el encargado de generar los eventos, para que el flujo de los datos tenga las transformaciones necesarias. Sea el siguiente ejemplo que utilizan casos de uso, escenarios y especificación de requerimientos para el proceso de datos. El caso de uso, donde se modela un proceso de transformación de datos, consta de tres operaciones básicas: el primer paso es la validación del formato del archivo de origen, el segundo es el reemplazo de valores nulos por espacios en blanco y por último extraer de la totalidad de los datos disponibles un conjunto de datos, representativo del total. Se puede observar también que el actor de este caso de uso es el controlador de tareas quien es el encargado de invocar a todas las tareas.

En el ejemplo tratado, no es necesario profundizar los casos de usos ni el escenario de trabajo, para cerrar el mismo solo hace falta especificar el detalle de los formatos de entrada y salida de cada uno de los pasos involucrados.

b.2.3. Fase de Codificación

En este paso se codifican todos los programas que se necesiten para realizar las transformaciones necesarias para el modelo de minería de datos. El encargado de la codificación recibirá, al menos, las especificaciones de los formatos de entrada y salidas. Como el controlador de tareas es independiente del programa que debe ejecutar, se puede usar el lenguaje de programación que se desea, siempre y cuando este pueda ser soportado por la plataforma en la que se desea trabajar. Cabe mencionar que se puede entregar al codificador toda la información que el encargado de las transformaciones crea necesario. Si el lenguaje así lo permite se podría entregar los diagramas de clases necesario para la codificación, así podríamos utilizar uno de los tantos esquemas que nos facilita UML. Volviendo sobre la documentación mínima que se le debe entregar al codificador, en el cambo observaciones de la planilla con los nombres de los archivos que debe recibir y retornar el programa, es muy importante que el codificador sepa cuales son las principales características del archivo no ya de formato que las posee, sino de volumen de información pues ante distintas cantidades de información por procesar, la codificación será muy distinta. El codificador además de realizar el programa se encarga de hacer las pruebas de unidad de los programas que realiza, es pro esto que se le debe también facilitar un archivo de entrada con los datos reales, de ser posible con el volumen de información que en producción se enfrentará. Una vez que se ha finalizado con la codificación, el paso siguiente es la prueba de unidad y de regresión, por parte del encargado de la generación de la secuencia de tareas.

b.2.4. Fase de Pruebas

Esta etapa de prueba no solo se refiere a la comprobación de los programa encargados de la generación de las transformaciones, sino a la construcción del archivo que proveerá la secuencia de pasos al controlador de tareas. Con las primeras transformaciones a realizar, se carga el archivo de formatos del controlador de tareas, es necesario tener en cuenta que no es recomendable agregar varios pasos de una vez, tratando de hacer una prueba, validar la salida y agregar otro paso. Las pruebas que se realizan son:

_ *De unidad:* La finalidad de esta prueba es validar que el programa cumpla la función para la cual fue ingresado a las tareas, es necesario poder simular de la manera mas precisa posible una ejecución real.

_ La validación que se hace es en función de la documentación antes desarrollada, se toman los formatos de archivo de entrada y salida, y simplemente se evalúa si los formatos son correctos.

_ *De regresión:* En este tipo de pruebas lo que se debe realizar es la validación de los tiempos de procesamiento y recursos necesarios. Para realizar esto es necesario ejecutar la tarea completa hasta el último paso que hemos agregado, es decir hacer una corrida completa de lo que tenemos hasta este momento. Lo que se busca probar es el tiempo de procesamiento, en la sucesión de pasos ejecutados es posible detectar que el tiempo de procesamiento es inaceptable para nuestro sistema, que los recursos utilizados son demasiados, etc. De la evaluación antes descripta se pueden presentar distintas variantes:

_ *Los tiempos son aceptables:* Esta es la posibilidad mas optimista de ser así, lo que se hace es continuar con el agregado de los siguientes pasos, esto puede ser que ya se tenga la especificación de la tarea y la próxima iteración a realizar solo sea agregar un paso mas y rehacer los ciclos de prueba.

_ *Es procesamiento es demasiado extenso:* Esto hace que se deba replantear la estrategia de transformaciones a realizar, aquí se debe detectar cual es el paso que más tiempo lleva y modificarlo.

_ *Los recursos no son los óptimos:* Este tipo de alternativa se da cuando por ejemplo el espacio de almacenamiento intermedio es demasiado grande y no se dispone de más espacio en disco, esto hace que sea necesario la reformulación de la estrategia a desarrollar. Sobre los posibles caminos de acción que se puedan seguir en esta opción serán abordados en el próximo paso de la metodología propuesta

b.2.5. Fase de Evaluación:

Con toda la información de las pruebas antes realizadas el encargado de realizar las transformaciones, deberá tomar un camino de acción, como se ha dicho antes, salvo que todo halla sucedido como se esperaba, en el resto de las opciones se deberá modificar algo. La primera alternativa a seguir es una vez detectado el paso, programa, que más recursos o tiempo demora, es tratar de optimizarlo. Otra alternativas no tan costosa es, la posibilidad de ejecutar las tareas en forma paralela, esto se hace agregando un punto de bifurcación en el controlador de tareas y se hace un procesamiento en paralelo; de no poder hacer esto otro camino de acción a seguir es la posibilidad es plantear generar nuevamente el programa en un lenguaje con mejor rendimiento, a modo de ejemplo podemos citar si se ha hecho el programa en un lenguaje como Visual Basic, se lo podría pasar a C/C++, para que su ejecución sea mas optima. Otra alternativa que también podemos elegir es, a semejanza de la normalización de las bases de datos que en una primera instancia se normaliza, y para finalizar se realiza una des-normalización de las tablas para que estas posean una velocidad de acceso aceptable; se realizaran modificaciones en los programas que integran cada paso, para que se hagan mas de una transformación en un paso, como de la experiencia se ha observado que en cada paso los tiempos de acceso a disco, lectura del archivo y escritura de los mismos, es lo que mas tiempo insume, unir transformaciones puede hacer que se reduzca el tiempo de procesamiento, aunque esto va en detrimento de los reprocesos que se puedan generar, en ciertos casos es la única alternativa mejorar la performance. Esto son algunos de los caminos alternativos que se podrán seguir para la mejora del rendimiento de la tara a ejecutar, en definitiva el encargado de la realización de las transformaciones tendrá la libertad de realizar las modificaciones que desee para poder llevar a buen puerta su trabajo.

b.2.6. Fase de Nueva iteración

De lo dicho hasta el momento se deduce la necesidad de generar nuevas iteraciones con cada paso de la tarea a realizar, este proceso se repite hasta finalizar todas las transformaciones necesarias para satisfacer el Modelo de Minería de Datos. De la metodología propuesta se desprenden algunas observaciones necesarias de hacer:

_ Primero en función del método de trabajo el controlador de tareas es de suma importancia para la realización del trabajo, cuanto mas sofisticado sea el controlador y mas opciones pueda manejar, mejor será la forma que apliquemos la metodología.

_ Sobre el uso de un sistema de Monitoreo y Diagnostico para el controlador de tareas, es necesario proveerle una herramienta, en este caso un sistema experto, para que pueda manejar alternativas no contempladas por los programas hechos

para generar los pasos de las tareas, podemos citar, cuando parar ante el primer error encontrado en el archivo o después de encontrar cien registros con error, y ante esta situación que se debe hacer enviar los registros erróneos a un archivo temporal para su posterior análisis. Ante estos errores que se debe hacer, en este caso una vez decidido que se ha producido un error el cual es la política que se debe llevar a cabo. En caso de ser necesario dar avisos, entra en juego el subsistema de alarmas, que es el encargado de disparar y controlar todas las alarmas que generara el sistema y el tiempo de respuesta de los mismos. Dentro de las tareas rutinarias en el proceso de transformación llevada a cabo una tarea que es necesaria es la evaluación de las ejecuciones. Esto se trata de generar un seguimiento de los procesos cuando ya se encuentran en producción, donde ya se ha automatizado la tarea y no es controlada por ningún operador humano.

_ La experiencia dicta que todos los procesos con el tiempo se van degradando, su tiempo de respuesta empieza a ser peor, el espacio en disco utilizado aumenta, y esto puede llegar a niveles inaceptables, aunque el resultado final es el esperado. Como se habrá observado en la metodología, cuando se detalla la documentación requerida para la generación de cada paso se especificó un “Archivo de pasos”, el cual no se había tratado, este archivo es el cual nos permitirá evaluar el rendimiento de la tarea, en el mismo se contabilizaran los registros transformados , tiempo de procesamiento y demás información que se crea útil para el análisis del rendimiento en producción .Lo que se realizara se podría denominar como una minería de datos del proceso de transformación de la minería de datos. Es ahí donde la utilización de un sistema experto puede tener mucho valor agregado, pues el mismo se encargará de analizar los tiempos de proceso compararlo con el volumen de información que se ha procesado y determinar un camino de acción a seguir. Algunas situaciones que se pueden detectar con este análisis son, baja en la capacidad de procesamiento en determinados momentos del día, hay que recordar que nuestras pruebas aunque completas, no pueden simular todo el ambiente de producción donde otros procesos están corriendo en paralelo al nuestro y están compitiendo por los recursos del mismo. Aumento de la cantidad de espacio necesario para la ejecución de los procesos, esto se puede producir por la fragmentación de la información en el disco, además de la perdida de respuesta, como se puede observar cuanto antes se detecten estos problemas menos traumática será su solución.

2.4 Diseño e implementación

2.4.1 Diseño físico

El diseño físico parte del esquema lógico y da como resultado un esquema físico. Un esquema físico es una descripción de la implementación de una base de datos en memoria secundaria: las estructuras de almacenamiento y los métodos utilizados para tener un acceso eficiente a los datos.

En la estrategia de implantación de un depósito de datos debe considerarse el mejor diseño físico para el modelo de datos. El diseño físico debe estar orientado a generar buen rendimiento en el procesamiento de consultas, a diferencia del modelo lógico que está orientado al usuario y a la facilidad de consulta.

Los modelos lógicos conseguidos en la fase de análisis se convierten en modelos físicos. Se generan los diseños para programas y procesos que se requieren según la arquitectura, tanto a nivel de los datos como de aplicación. Se conoce también como diseño físico y consiste en plasmar en la práctica, los diseños lógicos de la fase de Análisis. Incluye la construcción de programas que creen y modifiquen las bases de datos, que extraigan datos de las fuentes, programas para transformación de datos tales como integración, resumen y adición, programas para la actualización de los datos, programas para búsquedas en bases de datos muy grandes.

En la figura 2.4.1 muestra en que nivel de las fases del diseño de un depósito de datos se encuentra el diseño físico y que conlleva.

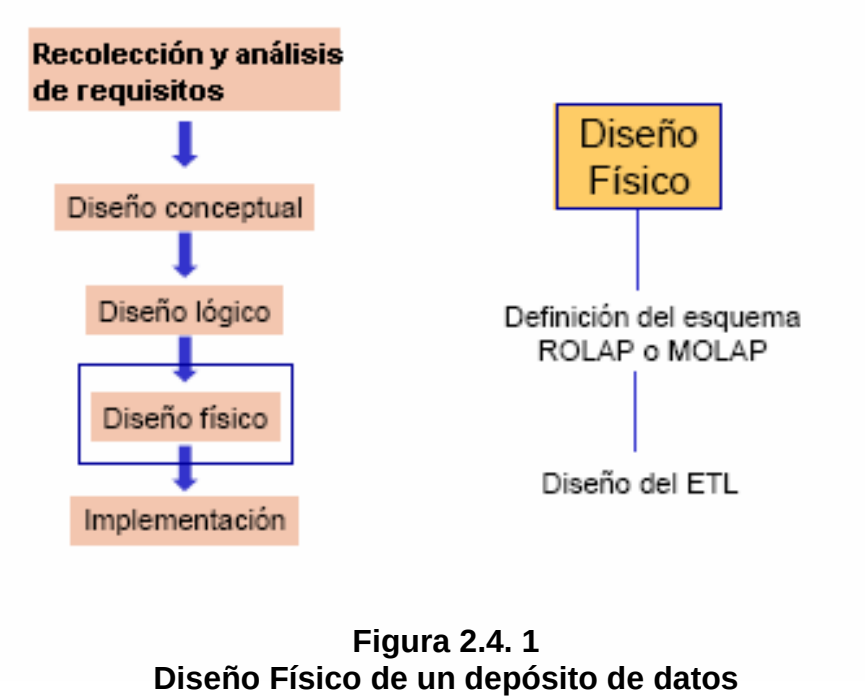

Figura 2.4. 1
Diseño Físico de un depósito de datos

Entre los objetivos fundamentales que se debe satisfacer la arquitectura de un Depósito de datos es soportar diferentes configuraciones de los datos y al mismo tiempo, brindar facilidades al usuario para la ejecución y la administración de sus tareas en ambientes complejos. Existen tres arquitecturas de datos empresariales que se definen a nivel lógico: las capas de datos de tiempo real, datos reconciliados y datos derivados. Cada una de ellas son conceptuales, pero en la medida que se avanza hacia una implementación física del Depósito de datos, se puede apreciar que no tienen una contrapartida física, que se va conformando a partir de los sistemas operacionales, del diseño del Depósito de datos Empresarial y del diseño del Depósito de datos Informacional.

El Depósito de datos Empresarial es la materialización de los datos reconciliados de forma altamente centralizada. Se diseña para ser la fuente única de todos los datos puestos a disposición de los usuarios terminales. Por su parte, Depósito de datos Informacional contiene los datos derivados y se diseña para apoyar las necesidades informacionales. Su estructura se optimiza para mejorar el rendimiento de las consultas utilizadas por los usuarios terminales.

Diseño del Depósito de datos Empresarial.

En el Depósito de datos Empresarial los datos reconciliados se ubican de forma altamente centralizada y son periódicos por naturaleza. Su diseño se basa en el modelo entidad relación genérico (que puede coincidir con el modelo entidad relación del sistema o del proyecto) y debe ser una representación lo más cercana

posible a las vistas lógicas de las aplicaciones con la estructura necesaria para almacenar la historia de la empresa.

En el Depósito de datos Empresarial se incluyen nuevas tablas normalizadas con el objetivo de incorporar la información que no está concebida en los sistemas operacionales, marcas de tiempo para almacenar explícitamente la historia y llaves sustitutas (surrogate keys) con vistas a garantizar la integridad de los datos.

Incorporar la información requerida es importante a la hora de poblar el Depósito de datos Empresarial con vistas a lograr una representación lo más completa y cercana posible a la realidad. La ausencia de esta información en los sistemas de producción, impide la carga automática del Depósito de datos Empresarial.

Usualmente para almacenar los datos periódicos se utilizan las marcas de tiempo (timestamp) que consisten en uno o varios campos definidos en el formato fecha-hora (date-time) y registran el momento en que un elemento se crea, se elimina o se modifica y que se adiciona a la llave de la tabla original. Existen varios esquemas de implementación para maximizar la eficiencia de las marcas de tiempo. El enfoque de una marca única de tiempo se basa en una marca de tiempo que identifica el momento en que un elemento cambia de estado y el enfoque de marcas dobles de tiempo se basa en dos marcas de tiempo que identifican el inicio y el final del período de validez del registro.

El uso de las llaves sustitutas puede verse como otra variación que presenta el diseño del Depósito de datos Empresarial con respecto al modelo entidad relación genérico. Una llave sustituta no es más que un campo que se agrega a una tabla y que asume el papel de llave primaria actual. Es un mecanismo necesario para eliminar la redundancia y disminuir la incoherencia de los datos, garantizando la integridad de los datos en el Depósito de datos Empresarial y acelerando las consultas que se utilizan para poblar el Depósito de datos Informacional, que en su mayoría son complejas.

Diseño del Depósito de datos Informacional

Con el fin de apoyar las necesidades informacionales de los usuarios terminales, se parte del diseño del Depósito de datos Empresarial y se sigue el flujo lógico de la definición de un Depósito de datos para llegar a la etapa final en el proceso de modelación que consiste en construir el modelo de los datos derivados que residen en el Depósito de datos Informacional.

Para diseñar el Depósito de datos Informacional se recomienda utilizar el diseño dimensional de los datos basado en el procesamiento analítico (OLAP: On Line Analytical Processing). Los productos OLAP son herramientas que permiten

construir sistemas de ayuda a la toma de decisiones con la capacidad de realizar en tiempo real sofisticados análisis multidimensionales orientados a entrecruzar información procedente de diversas fuentes de datos; posibilitando, además, navegar por la información para llegar al nivel de detalle que sea necesario.

Las estructuras de datos sobre las que se basan las herramientas OLAP pueden dividirse principalmente en tres categorías:

1. MOLAP (Multidimentional OLAP), basada en cubos de varias dimensiones para ubicar los datos, refleja con mayor exactitud la realidad modelada y es más sencilla de comprender por el usuario final que los modelos relacionales tradicionales o los ficheros planos de datos.

2. ROLAP (Relational OLAP), preserva el enfoque relacional de las bases de datos y tiene el propósito de mantener el mismo grado de eficiencia que las herramientas MOLAP.

3. HOLAP (Hybrid OLAP), híbrido entre MOLAP y ROLAP que trata de rescatar y de fusionar las virtudes de ambos modelos. Los datos primitivos que están almacenados en una base relacional y que son accedidos por métodos ROLAP conviven con los datos agregados a diferentes niveles almacenados en una base multidimensional, utilizando múltiples cubos o hipercubos en dependencia del fabricante.

Es preciso establecer comparaciones entre las principales características de ambos enfoques para determinar el modelo a utilizar en cada caso.

El modelo multidimensional utiliza las bases de datos multidimensionales basadas en arreglos n-dimensionales y admite consultas ad-hoc (consultas cuyos criterios se establecen en el momento de su formulación). Es muy rápido para conjuntos de datos pequeños o medianos, pero consume un elevado número de recursos dadas sus necesidades de almacenamiento y procesamiento.

El modelo dimensional relacional se basa en un “esquema de estrella”, que preserva las estructuras relacionales y soporta de medianas a grandes bases de datos. Admite solicitudes ad-hoc que requieran, incluso, la adición de nuevas dimensiones dinámicamente. Consume un alto grado de recursos por las necesidades de almacenamiento y se caracteriza por una velocidad buena cuando trabaja con conjuntos pequeños de datos.

La ventaja principal del modelo dimensional relacional estriba en su flexibilidad, extensibilidad y adaptación a elementos de datos inesperados y a nuevas decisiones de diseño. Existen modificaciones que se pueden realizar sobre un modelo dimensional relacional pero que no pueden aplicarse a un modelo multidimensional, como por ejemplo, adicionar hechos no previstos, adicionar dimensiones nuevas, adicionar atributos a las dimensiones existentes y adicionar nuevas tablas de hechos, entre otras. En este modelo ROLAP cualquier tabla del esquema puede transformarse dinámicamente, por lo que no es necesario reprogramar las herramientas de consulta o reporte y las aplicaciones existentes continuarán funcionando sin alterar sus resultados.

De modo general, el modelo dimensional se compone de: estructuras o elementos constituyentes, operaciones y técnicas de agregación y de desnormalización.

Entre las estructuras o elementos constituyentes del modelo dimensional relacional, las que más se destacan son los hechos, las dimensiones, los atributos y las jerarquías de atributos.

Las tablas de hechos son usualmente muy grandes en comparación con el resto de las tablas, generalmente están normalizadas y se enlazan a las dimensiones a través de llaves foráneas definidas para cada dimensión, asegurando que se mantenga una vista consistente de los datos.
Mientras que las tablas de hechos están formadas por datos detallados a nivel de transacciones, por valores numéricos para ejecutar los análisis y por identificadores que constituyen enlaces con las tablas de dimensiones, estas últimas representan categorías de información que organizan y describen los datos de las tablas de hechos, por lo que deben tener un nivel de detalle que permita un alto grado de flexibilidad para el análisis.

Las jerarquías de atributos están dadas por la dependencia o relación que exista entre los atributos, así como por el significado y la organización lógica de los mismos. Es una vía que se ofrece a los usuarios para que puedan navegar según las características de cada dimensión, propiciando una mayor flexibilidad en el análisis de los datos. Las jerarquías predeterminan una vía para orientar los análisis y cada escalón en la jerarquía proporciona un nivel de análisis.

Dentro del modelo relacional se define un conjunto de operaciones que pueden clasificarse en operaciones simples como la asignación, operaciones conceptuales que incluyen la unión, la intersección, la diferencia y el producto cartesiano; y operaciones relacionales tales como la proyección, la selección, el join y la división.

Los diseñadores del Depósito de datos deben tener en cuenta todas las tablas que se han denormalizado para evitar inconsistencias en la base de datos durante el proceso de actualización.

Además de las técnicas para construir y mantener el Depósito de datos Informacional en el enfoque relacional, existen diferentes modelos que contribuyen a la representación lógica de un esquema dimensional. El tipo de modelo que se utilice depende de los objetivos fundamentales del Depósito de datos Informacional y de las aplicaciones informacionales que se estén diseñando en particular. En el Depósito de datos Informacional pueden coexistir varios tipos de modelos que pueden integrarse o no, según las necesidades informacionales. La diferencia fundamental entre cada uno radica en la selección de las tablas a denormalizar.

Esquema de estrella: El Esquema de Estrella representa una tabla de hechos central relacionada con varias tablas de dimensiones. Es el modelo más conocido y el resto de los modelos resultan modificaciones de él.

Es una técnica de modelado de datos utilizada para proyectar datos de soporte de decisión multidimensionales en una base de datos relacional. La razón para el desarrollo del esquema en estrellas es que la técnica de modelado relacional existente, las relaciones entre entidades y la normalización, no producen una estructura de bases de datos que cumpliera bien con los requisitos de análisis de datos avanzados.

Los esquemas con estrella, figura 2.4.2, producen un modelo fácil de ejecutar para el análisis de datos multidimensionales, al mismo tiempo que conserva la estructura relacional sobre las cuales se construye las bases de datos operativas, además de que su jerarquía de dimensiones es lineal.

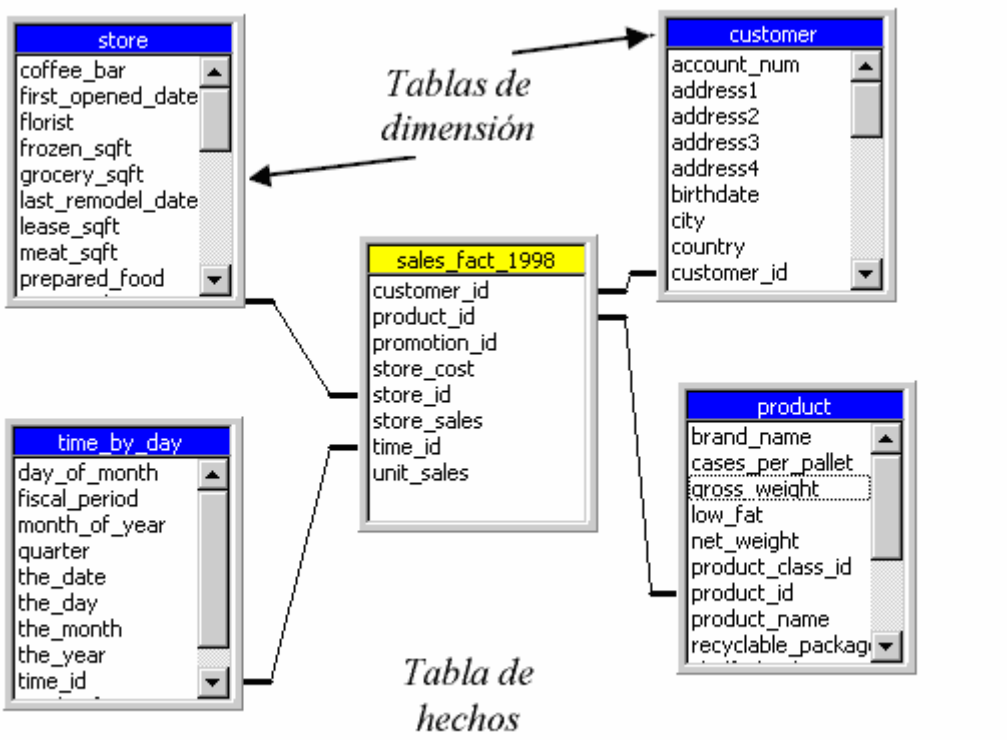

Figura 2.4. 2
Esquema de base de datos en Estrella

Esquema de Copos de Nieve: En el Esquema de Copos de Nieve, representado en la figura 2.4.3, todas las tablas deben estar normalizadas. También llamada estrella jerárquica, su jerarquía no es lineal.

Las bases de datos relacionales a menudo emplean esquemas de copo de nieve para proporcionar los mejores tiempos de respuesta posibles a las consultas complejas. Los esquemas de copos de nieve contienen una tabla de hechos central sin normalizar para el tema y numerosas tablas de dimensión para la información descriptiva sobre las dimensiones del tema. La tabla de hechos puede contener varios millones de filas. La información a la que se tiene acceso con más frecuencia se agrega previamente y se resume para mejorar aún más el rendimiento.

Si bien el esquema de copo de nieve se considera fundamentalmente una herramienta con la que el administrador de bases de datos puede aumentar el rendimiento y simplificar el diseño del Depósito de datos, también se utiliza para

representar la información del almacén de datos de forma que tenga más sentido para los usuarios finales.

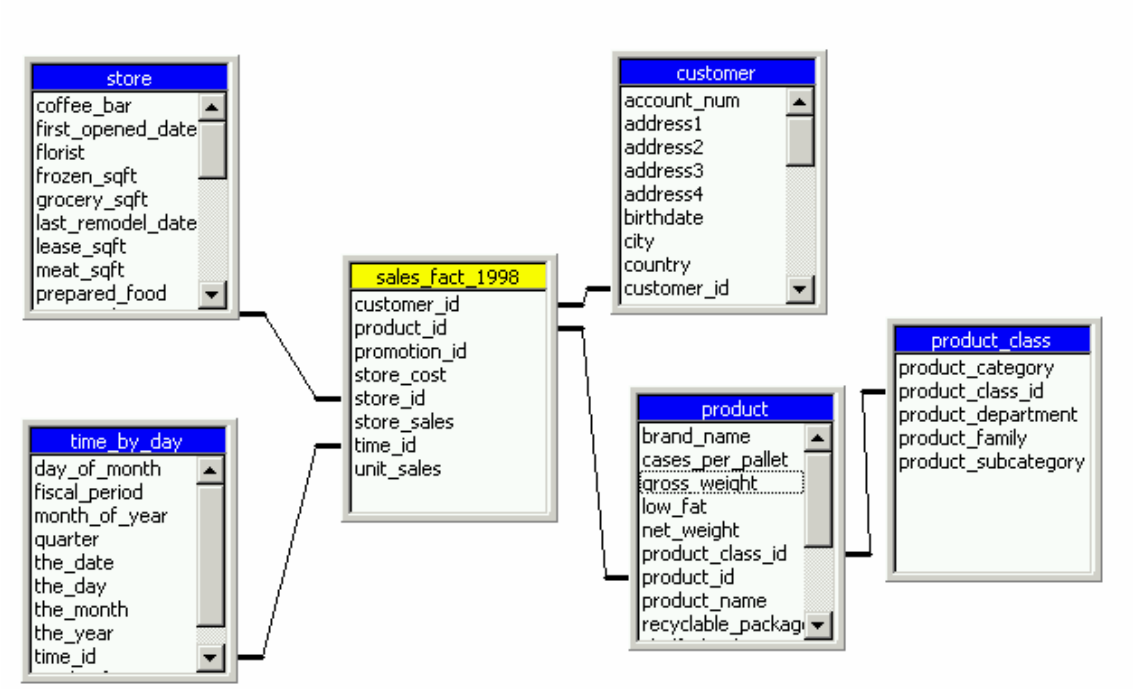

Figura 2.4. 3
Esquema de base de datos de copo de nieve

Esquema Parcial de Copos de Nieve: El Esquema Parcial de Copos de Nieve trata de integrar el esquema de copos de nieve y el de estrella, proponiendo la denormalización de algunas tablas de dimensiones de acuerdo con las consultas más frecuentes.

Esquema de Constelación: El modelo de constelación esta compuesto por una serie de modelos estrella, esta se representa en el esquema de la figura 2.4.4, donde existe una tabla de hechos principal y una serie de tablas de hecho agregadas o auxiliares las cuales pueden ser sumarizaciones de la principal no necesariamente tiene que estar relacionadas.

Con las mismas dimensiones de la principal, puede relacionarse con un subconjunto de ellas o con nuevas dimensiones de acuerdo a los requerimientos del sistema.

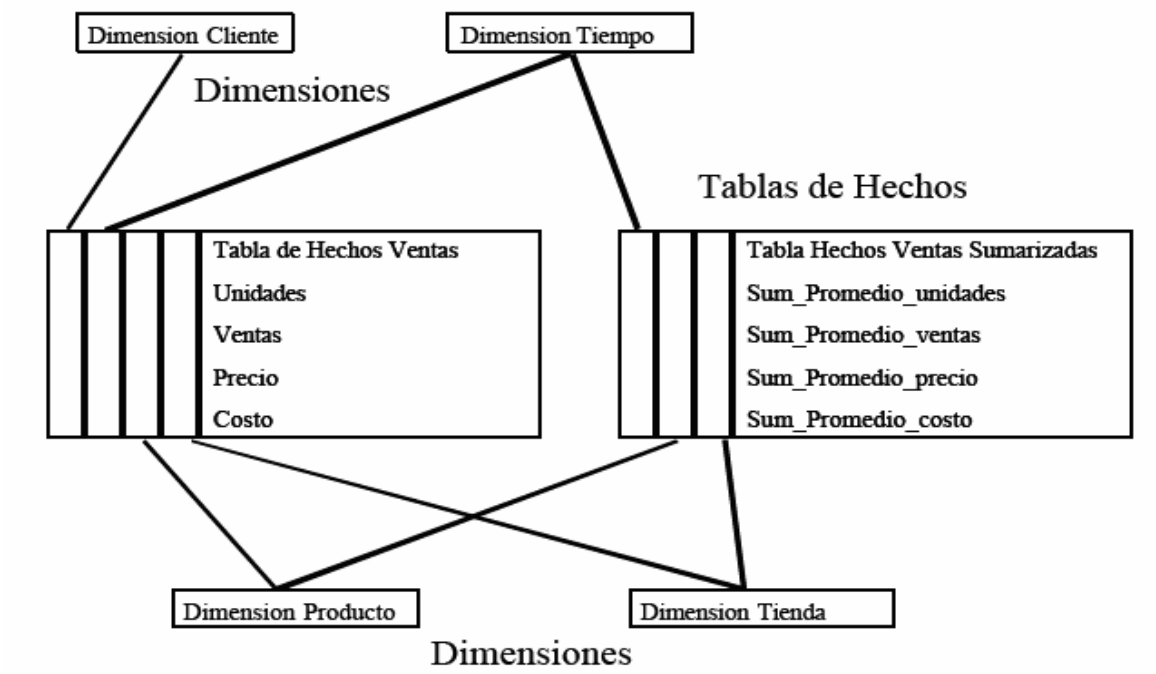

Figura 2.4. 4
Esquema de base de datos de Constelación

Se puede decir que la clave de un sistema informacional radica en que su diseño permita manipular los datos en el mismo sentido en el que se desea enfocar la investigación y el análisis.

Se muestra la correspondencia entre la arquitectura conceptual y la arquitectura física de un Depósito de datos, se detallan las características y peculiaridades para el diseño lógico y físico de un Depósito de datos, pasando por la categoría Empresarial e Informacional, respectivamente; además se mencionan las componentes a tener en cuenta en cada caso.

Modelado Multidimensional

Modelado Dimensional es una técnica para modelar bases de datos simples y entendibles al usuario final, basado en la estructura de datos denominada cubo como se muestra en la figura 2.4.5. La idea fundamental es que el usuario

visualice fácilmente la relación que existe entre las distintas componentes del modelo.

Las celdas del cubo se acceden mediante un conjunto de dimensiones que:
– Deben ser ortogonales (sin dependencias entre ellas)
– Pueden tener varios niveles de detalle (categorías)

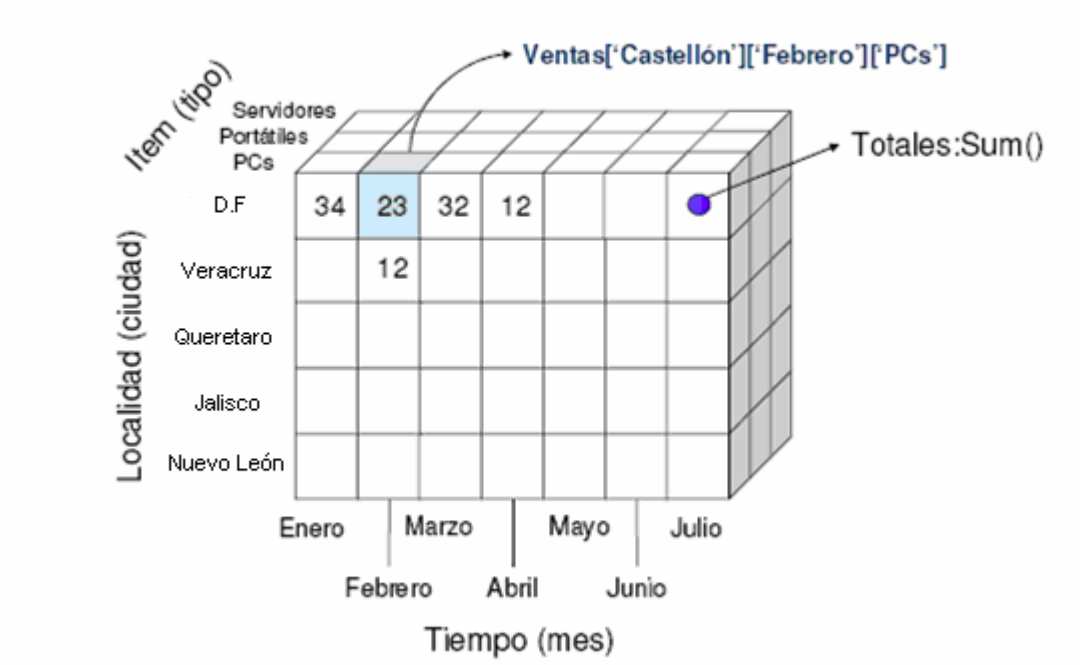

Figura 2.4. 5
Representación de Cubo dimensional

Una medida es una función numérica que puede ser evaluada en cada celda del cubo. El valor de la medida de una celda se computa agregando los datos que están contenidos en dicha celda.

Los tipos de medidas:
– Distributiva: el valor de la medida para una celda puede calcularse evaluando su función a los valores de las medidas de su sub-celdas.
Ej.: count(),sum(),min() y max()

– Algebraica: el valor de la medida para una celda puede calcularse mediante una función con M argumentos, los cuales se obtienen con funciones distributivas. Ej.: avg(), standard_deviation()

– Holistic: el resto de medidas que no son distributivas ni existe una función algebraica sobre medidas distributivas. median(), mode()

Consideremos un punto en el espacio. El espacio se define a través de sus ejes coordenados (por ejemplo X, Y, Z). Un punto cualquiera de este espacio quedará determinado por la intersección de tres valores particulares de sus ejes.

En el modelo multidimensional cada eje corresponde a una dimensión particular. Entonces la dimensionalidad de la base estará dada por la cantidad de ejes (o dimensiones) que le asociemos.

Cuando una base puede ser visualizada como un cubo de tres o más dimensiones, es más fácil para el usuario organizar la información e imaginarse en ella cortando y rebanando el cubo a través de cada una de sus dimensiones, para buscar la información deseada.

El modelo básico tiene cuatro componentes: hechos, dimensiones, atributos y jerarquía de atributo.

Los hechos son mediciones numéricas (valores) que representan un aspecto o actividad de negocio específica. Por ejemplo las cifras de ventas. Generalmente son unidades, costos, precios o ganancias; estos normalmente se guardan en una tabla de hechos que es el centro del esquema de estrella. La tabla de hechos contiene hechos vinculados por sus dimensiones. Los hechos calculados o derivados se les llama métricos para diferenciarlos de los hechos guardados. El grano o la granularidad de la tabla queda determinada por el nivel de detalle que se almacenará en la tabla. Por ejemplo, para el caso de producto, mercado y tiempo antes visto, el grano puede ser la cantidad de madera vendida ‘mensualmente’. El grano revierte las unidades atómicas en el esquema dimensional.

Las dimensiones son características calificadoras que proporcionan perspectivas adicionales de un hecho dado. Normalmente se guardan en tablas dimensiones.

Cada tabla de dimensiones contiene atributos. Con frecuencia se utilizan los atributos para buscar, filtrar o clasificar hechos. Las dimensiones proporcionan características descriptivas de los hechos mediante sus atributos.
A nivel de dimensiones es posible definir jerarquías, las cuales son grupos de atributos que siguen un orden preestablecido. Una jerarquía implica una organización de niveles dentro de una dimensión, con cada nivel representando el total agregado de los datos del nivel inferior. Las jerarquías definen cómo los datos son sumarizados desde los niveles más bajos hacia los más altos. Una dimensión típica soporta una o más jerarquías naturales. Una jerarquía puede pero no exige contener todos los valores existentes en la dimensión.

El esquema creado mediante sus dimensiones y hechos puede proporcionar los datos cuando se requiera y con el formato necesario.

Pasos básicos del Modelado Multidimensional:

1. Decidir cuáles serán los procesos de negocios a modelar, basándose en el conocimiento de éstos y de los datos disponibles. Ejemplo: Gastos realizados por cada mercado para cada ítem a nivel mensual. Productos vendidos por cada mercado según el precio en cada mes.

2. Decidir el Grano de la tabla de hechos (Fact) de cada proceso de negocio.
Ejemplo: Producto x mercado x tiempo. En este punto se debe tener especial cuidado con la magnitud de la base de datos, con la información que se tiene y con las preguntas que se quiere responder. El grano decidirá las dimensiones del Depósito de Datos. Cada dimensión debe tener el grano más pequeño que se pueda puesto que las preguntas que se realicen necesitan cortar la base en caminos precisos (aunque las preguntas no lo pidan explícitamente).

3. Decidir las dimensiones a través del grano. Las dimensiones presentes en la mayoría de los Depósitos de datos son: tiempo, mercado, producto, cliente. Un grano bien elegido determina la dimensionalidad primaria de la tabla hechos. Es posible usualmente agregar dimensiones adicionales al grano básico de la tabla de hechos, donde estas dimensiones adicionales toman un solo valor para cada combinación de las dimensiones primarias. Si se reconoce que una dimensión adicional deseada viola el grano por causar registros adicionales a los generados, entonces el grano debe ser revisado para acomodar esta dimensión adicional.

4. Elegir las mediciones del negocio para la tabla hechos. Se deben establecer los índices que quedarán determinados por la clave compuesta de la tabla hechos.

2.4.2 OLAP multidimensional y OLAP relacional

Como ya se ha mencionado anteriormente existen algunas clasificaciones entre las implementaciones OLAP. La clasificación está hecha sobre la base de en qué tipo de motor son almacenados los datos:

- ROLAP es una implementación OLAP que almacena los datos en un motor relacional.
- MOLAP es una implementación OLAP que almacena los datos en una base de datos multidimensional.
- HOLAP (Hybrid OLAP) almacena algunos datos en un motor relacional y otros en una base de datos multidimensional.
- DOLAP es un OLAP orientado a equipos de escritorio (Desktop OLAP). Trae toda la información que necesita analizar desde la base de datos relacional y la guarda en el escritorio. Desde ese momento, todas las consultas y análisis son hechas contra los datos guardados en el escritorio.

Más a detalle se define las clasificaciones MOLAP y ROLAP.

OLAP Multidimensional

Es el procesamiento analítico en línea multidimensional MOLAP por sus siglas en ingle, amplia la funcionalidad de OLAP a sistemas de administración de bases de datos multidimensionales MDBS, un MDBS utiliza técnicas patentadas especiales para guardar datos en arreglos en forma de matriz o de n-dimensiones, la premisa de MOLAP es que las bases de datos multidimensionales son mas adecuadas para manejar guardar y analizar datos multimensional4es, la mayoría de las técnicas patentadas utilizadas en MDBS, se derivan de campos de ingeniería como diseño asistido por computadora/manufactura asistida por computadora (cad/cam) y sistemas de información geográfica (GIS por sus siglas en ingles.

Conceptualmente los usuarios de MDBS visualizan los datos guardados como un cubo tridimensional conocido como cubo de datos, la ubicación de cada valor del dato en el cubo es una función de los ejes x, y y z en un espacio tridimensional los ejes x y y z, respectivamente representan las dimensiones del valor del dato. Los cubos de datos pueden crecer n numero de dimensiones con lo que llegan a ser hipercubos. Los cubos de datos se crean extrayendo datos de las bases de datos operativas o del almacén de datos. Una característica importante de los cubos de datos es que son estáticos es decir, no están sujetos a cambios y deben ser creados antes que puedan ser utilizados, en otras palabras, los cubos de datos no pueden ser creados mediante consultas ad hoc, en su lugar se consultan cubos precreados con ejes definidos por ejemplo un cubo de ventas tendrá las dimensiones de producto, ubicación y tiempo y solo podrán consultarse esas dimensiones, por consiguiente el proceso de creación del cubo de datos es critico y requiere un trabajo de diseño frontal a fondo. Este diseño de trabajo frontal se justifica muy bien por el hecho de que sabe que las bases de datos MOLAP son muchos más rápidas que su contraparte ROLAP, en particular cuando se manejan conjuntos de datos pequeños a medianos.

Para acelerar el acceso a los datos los cubos de datos normalmente se conservan en la memoria, en lo que se llama cache cúbico (un cubo es solo una ventana a un subconjunto de datos predefinido en la base de datos, un cubo de datos y una base de datos no son la misma cosa). Como MOLAP también se beneficia de la infraestructura cliente/servidor, el cache cúbico puede estar localizado en el servidor MOLAP, en el cliente MOLAP o en ambos lugares. La figura muestra la arquitectura MOLAP básica.

Existe un nuevo tipo de DBMS llamado base de datos guardada en memoria MMDB, como las MMDB atienden todas las solicitudes de consulta d4esde la memoria, funcionan 1000 veces mas rápido que las bases de datos relacionales de disco o ligadas a procesador, incluso Microsoft incluyo una en Windows 2000, una pequeña, conocida IMDB.

La capacidad de capturar el cubo de datos en la memoria permite tiempos de respuesta más rápidos, pero también hace que el MDBS sea más intensivo en cuanto a recursos, que su contraparte, además los proponentes de ROLAP argumentan que el método de cubo de datos limita la flexibilidad, la escalabilidad y facilidad de integración.

Como el cubo de datos esta predefinido con un número fijo de dimensiones, la adición de una nueva dimensión requiere que se recree todo el cubo de datos. Este proceso de recreación es una operación que lleva mucho tiempo, por consiguiente, se crean cubos de datos con mucha frecuencia, el DDBMS pierde algo de su ventaja de velocidad con respecto a la base de datos relacional, y aunque los MDBS tienen ventajas de desempeño sobre las bases de datos relacionales es el mas adecuado para conjuntos de datos pequeños a medianos. la escalabilidad esta un tanto limitada, porque el tamaño del cubo de datos se restringe para evitar tiempos de acceso largos a los datos provocados por la falta de espacio de trabajo (memoria) para el sistema operativo y los programas de aplicación. Además el MDBMS utiliza técnicas de almacenamiento de datos patentadas a las que, a su vez requieren de métodos de acceso a los datos patentados que utilizan un lenguaje de consulta multidimensional.

El análisis de datos Multidimensionales también se ve afectado por la forma en que el sistema de bases de datos maneja la raleza (RT). La raleza es una medida de densidad de los datos guardados en el cubo de datos. Esta se calcula dividiendo el numero total de valores existentes en el cubo entre el numero total de celdas del cubo. Como las dimensiones del cubo de datos están predefinidas, no todas las celdas están pobladas, algunas celdas están vacías, regresando al ejemplo de ventas, podría haber muchos productos que no se venden durante un periodo dado en un lugar dado, de hecho con frecuencia se encontrara que menos de 50 % de las celdas del cubo de datos están pobladas, en todo caso las bases de datos multidimensionales deben manejar la raleza con eficiencia para reducir la carga de procesamiento y los requerimientos de recursos.

Los proponentes relacionales también argumentan que la utilización de solucione patentadas dificultan la integración del MDBMS con otras fuentes de datos y herramientas utilizadas en la empresa. No obstante, a pesar del hecho de que se requiere de una inversión de tiempo y esfuerzo considerables para integrar la tecnología nueva y la arquitectura de sistemas de información existentes, MOLAP puede ser una buena solución para aquellas empresas en las que las bases de datos pequeñas a medianas son la norma y la velocidad del software de aplicación es critica.

El objetivo de los sistemas MOLAP es almacenar físicamente los datos en estructuras multidimensionales de forma que la representación externa y la representación interna coincidan.

Los sistemas MOLAP disponen de estructuras de almacenamiento específicas (arrays) y técnicas de compactación de datos que favorecen el rendimiento del depósito. La figura 2.4.6 representa un sistema MOLAP y sus características.

El servidor MOLAP construye y almacena datos en estructuras multidimensionales.

La herramienta de OLAP presenta estas estructuras multidimensionales.

Los Datos se encuentran:
• Arrays
• Extraídos del depósito de datos

Almacenamiento y procesos eficientes.

La complejidad de la BD se oculta a los usuarios.

El análisis se hace sobre datos agregados y métricas o indicadores precalculados.

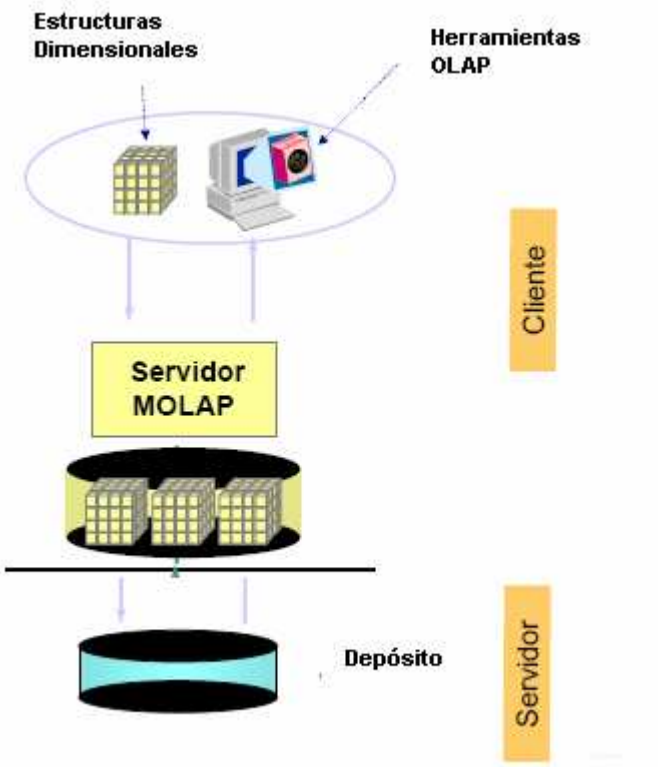

Figura 2.4. 6
Representación MOLAP

OLAP Relacional

El procesamiento analítico en línea relacional (ROLAP) proporciona funcionalidades OLAP con el uso de base de datos relacionales y herramientas de consulta relacionales utilizadas para guardar y analizar datos multidimensionales. Este método esta basado en las tecnologías relacionales existentes y representan una extensión natural de todas aquellas compañías que ya utilizan sistemas de administración de bases de datos relacional de sus

organizaciones. ROLAP agrega las siguientes extensiones a la tecnología RDBMS tradicional.

Los sistemas ROLAP se implementan sobre tecnología relacional, pero disponen de algunas facilidades para mejorar el rendimiento (índices de mapas de bits, índices de JOIN). La figura 2.4.7 representa un sistema ROLAP.

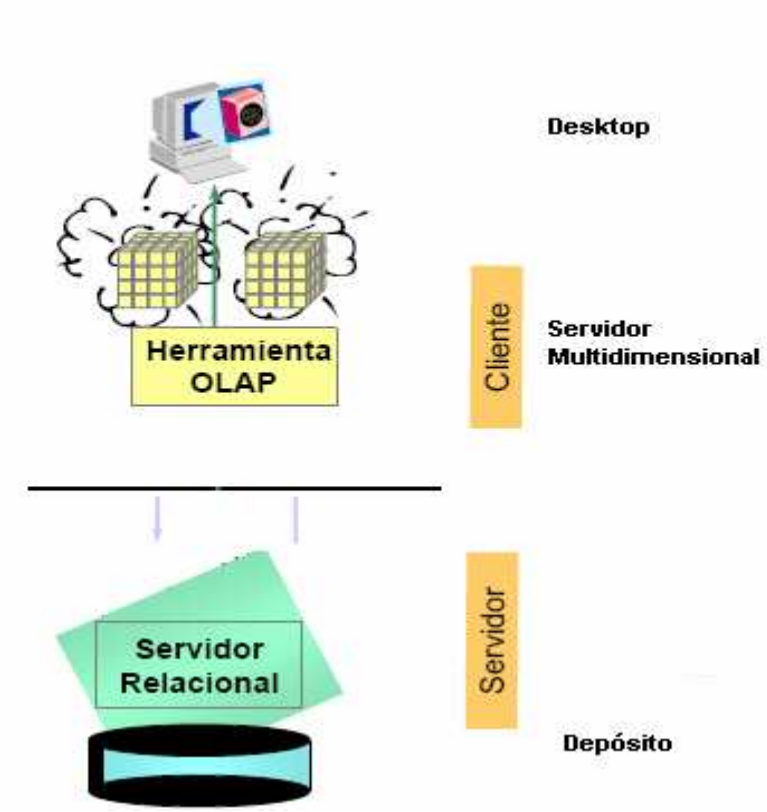

Figura 2.4. 7
Representación ROLAP

- Soporte de esquema de datos multidimensionales en el RDBMS:

  La tecnología relacional utiliza las tablas normalizadas para guardar datos. La confianza en la normalización como método de diseño de bases de datos relacionales es considerada como un obstáculo para el uso de sistemas OLAP. La normalización divide las entidades de negocio de piezas más pequeñas para producir las tablas normalizadas. Por ejemplo, los componentes de datos de ventas podrían guardarse en cuatro o cinco tablas diferentes. La razón para la utilización de tablas normalizadas es reducir las redundancias, con lo cual se eliminan las anomalías en los

datos, y facilitar su actualización. Desafortunadamente para propósitos de soporte de decisiones, es más fácil atender los datos cuando se consideran con respecto a otros. Con esta visión el ambiente de datos se ha hecho hincapié en que los datos de soporte de decisiones tiendan a ser normalizados, duplicados y pre-agregados. Estas características parece que impiden el uso de técnicas de diseño relacional estándar y RDBMS como fundamentos de los datos multidimensionales.

ROLAP utiliza una técnica de diseño especial que permite que la tecnología RDBS soporte representaciones de datos multidimensionales. Esta técnica de diseño especial se conoce como esquema en estrella, la cual se analizó en el capítulo anterior. En realidad el esquema en estrella crea casi el equivalente de un esquema de base de datos multidimensionala partir de la base de datos existente.

El esquema en estrella esta diseñado para optimizar operaciones de consulta de datos en lugar de operaciones de actualizaciones de datos. Desde luego, el cambio de fundamentos de diseño de los datos significa que las herramientas utilizadas para acceder a esos datos tendrán que cambiar. Los usuarios que conocen esas herramientas de consulta relacionadas tradicionales descubrirán que estas herramientas no funcionaran eficientemente con el nuevo esquema en estrella. No obstante, ROLAP soluciona el problema al soportar el esquema en estrella para el uso de herramientas de consultas conocidas. ROLAP proporciona funciones avanzadas de análisis de datos y mejora los datos de optimización de consulta y visualización de datos.

Los Depósitos de datos se construyen sobre un RDBMS. Los fabricantes de RDBMS ofrecen extensiones y herramientas para poder utilizar el RDBMS como un Sistema Gestor de Depósitos de Datos.

- Lenguaje de acceso a los datos y desempeño de consultas optimizados para datos multidimensionales.

Otra crítica de las base de datos relacionales es que el SQL utilizado con RDBMS no es adecuado para la realización de análisis de datos avanzados. La mayoría de las solicitudes de datos de soporte de decisión requiere el uso de consultas SQL de múltiples pasadas o múltiples sentencias de SQL anidadas. Para responder a esta crítica. ROLAP amplía el SQL de modo que pueda diferenciar entre los requerimientos de acceso a datos (basado en el esquema en estrella) y los datos operativos (tablas normalizadas). De esta manera, el sistema ROLAP es capaza de generar apropiadamente el código SQL requerido para acceder los datos en el esquema de estrella.

El desempeño de consulta también es mejor porque el optimizador de consulta se modifico para que pueda identificar los objetivos pensado del código SQL. Por ejemplo, el objetivo de la consulta es el almacén de datos, el optimizador transfiere las solicitudes al almacén de datos. No obstante, si el usuario realiza consultas de datos operativos, el optimizadote de consultas identifica esta operación y optimiza apropiadamente las solicitudes SQL antes de transferirlas al DBMS operativo.

Otra forma de uso de desempeño de consulta mejorado es el uso de técnicas de indexación avanzadas tales como índices tipografiados por bits en base de datos relacionales. Como el nombre lo dice, un índice tipografiado por bits se basa en bits 0 y 1 para representar una condición dada. Los índices tipografiados por bits son mucho mas eficientes en la manejo de grandes cantidades de datos que los índices generalmente encontrados en muchas de las bases de datos relacionales. Pero estas solo son utilizadas en las que el numero de posibles valores de un atributo son muy pequeños.

- Soporte de bases de datos muy grandes

  El soporte de base de datos muy grandes (VLDB ó VLDS Very Large Data Store en inglés) es un requerimiento obligatorio en los Depósito de datos. Por consiguiente, si se utiliza la base de datos relacional como depósito de datos, también se debe ser capaz de guardar cantidades de datos muy grandes. Los datos de soporte decisiones generalmente se cargan de modo masivo (por lotes) provenientes de los datos operativos. Por lo tanto los RDBMS deben contar con las herramientas necesarias para importar, integrar y poblar el depósito con datos operativos las operaciones por lote requieren que tanto las bases de datos origen como las base de datos destino se reserven (bloqueen).la velocidad de las operaciones de carga son importantes, sobre todo porque los sistemas de operación funcionan todos los días a toda hora.

  Dada la existencia de una arquitectura cliente/servidor abierta, ROLAP proporciona capacidades de soporte de decisión avanzadas aplicadas a toda la empresa.

2.4.3 Replicación de datos en un ambiente de Depósito de datos global

La creación de datamarts dependientes distribuidos desde un almacén de datos centralizado o incluso la duplicación del contenido de un datamart independiente, requiere la capacidad de replicar la información de forma confiable. Algunas herramientas (como SQL Server 7.0) incluyen facilidades para la distribución segura de la información desde un almacén de datos de publicación central hasta

varios datamarts de suscripción. La información puede dividirse por periodos, zonas geográficas, etc. como parte del proceso de replicación.

También proporcionan diversas tecnologías de replicación que pueden personalizarse según los requisitos específicos de la aplicación. Cada tecnología de replicación presenta diferentes beneficios y restricciones en tres importantes dimensiones[9]:

- Coherencia transaccional
- Autonomía de sitio
- Partición de datos
- Los requisitos a lo largo y ancho de estas tres dimensiones variarán entre una aplicación distribuida y la siguiente.

En la mayoría de aplicaciones de ayuda a la toma de decisiones, los datos no se actualizarán en sitios individuales. En lugar de ello, la información se preparará en un área temporal centralizada y se enviará a servidores de bases de datos distribuidos para un acceso remoto. Por este motivo, la replicación instantáneas se utiliza con frecuencia para distribuir datos.

La replicación de instantáneas toma una fotografía instantánea de los datos publicados en la base de datos en un instante concreto. En lugar de copiar las instrucciones INSERT, UPDATE y DELETE (características de la replicación transaccional) o las modificaciones de los datos (características de la replicación de mezcla), los subscriptores se actualizan mediante una actualización completa del conjunto de datos. Por tanto, la replicación de instantáneas envía todos los datos al subscriptor en lugar de enviar sólo los cambios. Si la cantidad de información que se envía es muy grande, se necesitarán recursos de red considerables para realizar la transmisión. Cuando se decida si la replicación de instantáneas es la más apropiada, habrá que considerar el tamaño del conjunto de datos en comparación con la volatilidad de los datos.

La replicación de instantáneas es el tipo de replicación más sencilla y garantiza una coherencia latente entre el editor y el subscriptor. También proporciona una gran autonomía si los subscriptores no actualizan los datos. La replicación de instantáneas es una buena solución para los subscriptores de sólo lectura que no necesitan los datos más recientes y pueden desconectarse totalmente de la red cuando no se realizan actualizaciones. Sin embargo, SQL Server proporcionará una amplia gama de posibilidades de elección para la replicación según los requisitos de aplicaciones.

[9] FUENTE:
http://msdn.microsoft.com/library/spa/default.asp?url=/library/SPA/dntaloc/html/datawarefmwk.asp
Para obtener más información sobre las capacidades de replicación en SQL Server 7.0, vea
http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnsql7/html/msdn_sqlrep.asp

Si se replica la relación r, se guardará una copia de la misma en dos o más emplazamientos. En el caso más extremo, se tendrá una réplica completa, en la que se guarda una copia en cada emplazamiento del sistema.

La réplica presenta un cierto número de ventajas e inconvenientes:

- Disponibilidad. Si falla uno de los emplazamientos que contienen la relación r, ésta aún se podrá encontrar en otro emplazamiento. Por tanto, el sistema puede seguir procesando las consultas que impliquen a r, a pesar del fallo en un emplazamiento.

- Aumento del paralelismo. En el caso de que la mayor parte de los accesos a la relación r sea de lectura de la misma, varios emplazamientos pueden procesar en paralelo las consultas que impliquen a r. Cuantas más réplicas de r haya, mayor será la posibilidad de que el dato buscado se halle en el emplazamiento en que se esté ejecutando la transacción. Por tanto, la réplica de los datos minimiza el tráfico de datos entre los emplazamientos.

- Aumento de la sobrecarga en las actualizaciones. El sistema debe asegurarse de que todas las réplicas de la relación r sean consistentes; en caso contrario, puede dar lugar a resultados erróneos. Por tanto, siempre que se actualice r, la actualización debe extenderse a todos los emplazamientos que contengan réplicas. La consecuencia es un aumento en la sobrecarga. Por ejemplo, en un sistema bancario en el que la información sobre las cuentas se replica en varios emplazamientos, hay que asegurarse de que el saldo de una cuenta concreta coincida en todos los emplazamientos.

En general, la réplica mejora el rendimiento de las operaciones de lectura y aumenta la disponibilidad de los datos para las transacciones de sólo lectura. Sin embargo, las transacciones de actualización suponen una sobrecarga mayor. Por tanto, un buen parámetro para afrontar el grado de réplica consistiría en sopesar la cantidad de consultas de lectura que se efectuarán, así como el número de consultas de escritura que se llevarán a cabo. En una red donde las consultas que se procesen sean mayoritariamente de lectura, se podría alcanzar un alto grado de réplica, no así en el caso contrario. El control de las actualizaciones concurrentes de los datos replicados por varias transacciones es más complicado que cuando se utiliza el enfoque centralizado del control de concurrencia. Se puede simplificar la administración de las réplicas de la relación r escogiendo una de ellas como copia principal de r. Por ejemplo, en un sistema bancario se puede asociar una cuenta con el emplazamiento en que se ha abierto. De manera parecida, en un sistema de reservas de unas líneas aéreas se puede asociar un vuelo con el punto en el que se origina.

2.4.4 VLDS y paralelismo

VLDS

Debido a que no existe al parecer una definición oficial o estándar para las bases de datos muy grandes (VLDB Very Large Data Base ó VLDS Very Large Data Store), se utiliza a veces para describir las bases de datos que ocupan almacenaje magnético en la gama de terabyte y que contienen mil millones de filas de la tabla. Típicamente, estos son usos de los sistemas de ayuda de decisión o del tratamiento transaccional que sirven a gran cantidad de usuarios.

La técnica del paralelismo esta involucrado a menudo en la consulta de almacenes muy grandes de datos, estos pueden utilizar un modelo muy grande de la memoria (VLM) para guardar tantos datos como sea posible en memoria física, y a menudo se requiere de la ayuda de hardware para un procesado eficiente de los datos. Grandes empresas utilizan arquitecturas paralelas para una eficiente ejecución de consultas complejas, que tienen fines analíticos dentro de sus negocios sobre grandes cantidades de datos.

La arquitectura en paralelo descompone grandes problemas en pequeños fragmentos de modo que cada fragmento del problema pueda ser ejecutado en paralelo por cada nodo, es decir, los Depósitos de datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas, llamadas los centros comerciales ó DataMarts, dependientes de los datos. Generalmente, dos ideas básicas dirigen la creación de un depósito de datos:

Integración de los datos de bases de datos distribuidas y diferentemente estructuradas, que facilita una descripción global y un análisis comprensivo en el depósito de los datos.

Separación de los datos usados en operaciones diarias de los datos usados en el depósito de datos para los propósitos de la divulgación, para la ayuda en la toma de decisiones, para el análisis y para controlar.

El ambiente de un Depósito de datos queda definido por la suma de los diferentes DataMarts integrados, no sólo a nivel físico sino también a nivel lógico.

Un DataMart es una vista lógica de los datos en bruto de sus datos provistos por el sistema de operaciones/finanzas hacia el Depósito de datos con la adición de

nuevas dimensiones o información calculada. Se les llama DataMart, porque representan un conjunto de datos relacionados con un tema en particular como Ventas, Operaciones, Recursos Humanos, etc, y están a disposición de los "clientes" a quienes les pueden interesar. Esta información puede accesarse por el Ejecutivo (Dueño) mediante "Tablas Dinámicas" de MS-Excel o programas personalizados. Las Tablas Dinámicas le permiten manipular las vistas (cruces, filtrados, organización) de la información con mucha facilidad. Los cubos de información (DataMarts) se producen con mucha rapidez. A ellos se les aplican las reglas de seguridad de acceso necesarias La información estratégica está clasificada en: Dimensiones y Variables. El análisis está basado en las dimensiones y por lo tanto es llamado: Análisis multidimensional.

Memoria muy grande (VLM) es un procesador y sistema operativo que pueden utilizar más que 4GB del RAM, que es el límite para los sistemas usando 32-BITS de direccionamiento. Las arquitecturas de VLM permiten programas de uso y bases de datos muy grandes con más que 4GB de los datos que se colocarán enteramente en memoria física, con realces grandes del funcionamiento. Algunos procesadores recientes como la DEC ALFA pueden procesar 64-BITS de datos a la vez y utilizar direcciones a la par de 32 BITS.

Un sistema de almacenes muy grande de datos (VLDS) está basado en una red rápida del ancho-área que conecta organizaciones numerosas y a individuos. El diseño de un VLDS implica problemas y las ediciones no presentes para sistemas más pequeños. Estas ediciones se centran en las áreas del nombramiento, los paradigmas y las arquitecturas de la comunicación, seguridad, y arquitectura del kernel.

Paralelismo

La arquitectura de un sistema de base de datos está influenciada en gran medida por el sistema informático en el que se ejecuta el sistema de base de datos. En la arquitectura de un sistema de base de datos se reflejan aspectos como la conexión en red, el paralelismo y la distribución.

El procesamiento paralelo dentro de una computadora permite acelerar las actividades del sistema de base de datos, proporcionando a las transacciones unas respuestas más rápidas, así como la capacidad de ejecutar más transacciones por segundo. Las consultas pueden procesarse de manera que se explote el paralelismo ofrecido por el sistema informático subyacente. La necesidad del procesamiento paralelo de consultas ha conducido al desarrollo de los sistemas de bases de datos paralelos.

El proceso paralelo divide una tarea grande en muchas tareas más pequeñas, y ejecuta las tareas más pequeñas concurrentemente en varios nodos. Consecuentemente, la tarea más grande termina más rápidamente. Paralelismo entre consultas es para proporcionar mayor rendimiento al dividir una única consulta compleja en varias partes y distribuir la carga de trabajo entre múltiples procesadores, incluidos los servidores enlazados vinculados de forma remota. El proceso paralelo divide una tarea grande en muchas tareas más pequeñas, y ejecuta las tareas más pequeñas concurrentemente en varios nodos. Consecuentemente, la tarea más grande termina más rápidamente.

Por ejemplo, en un banco con solamente una caja, todos los clientes deben formar una sola cola y la caja atenderá uno a uno de forma secuencial. Con dos cajas, la tarea puede estar partida con eficacia de modo que los clientes formen dos colas y sean atendidos dos veces de forma rápida, ó puedan formar una sola cola a proporcionar la imparcialidad. Éste es un caso en el cual el proceso paralelo es una solución eficaz.

Por el contrario, si el encargado de banco debe aprobar todas las peticiones del préstamo, el proceso paralelo no acelerará necesariamente el flujo de préstamos. No importa si muchas cajas están disponibles para procesar préstamos, todas las peticiones deben formar una sola cola para la aprobación del encargado de banco. Ninguna cantidad de proceso paralelo puede superar este embotellamiento incorporado al sistema.

La Figura 2.4.8 muestra el proceso secuencial de múltiples tareas de forma independientes.

Figura 2.4. 8
Proceso secuencial de una tarea larga

Como se observa en la figura 2.4.8, en el proceso secuencial, la tarea se ejecuta como una tarea grande. En el proceso paralelo, figura 2.4.9, la pregunta se divide en tareas más pequeñas múltiples, y cada tarea componente se ejecuta en un nodo separado. En el proceso secuencial, las tareas independientes compiten por un solo recurso. Solamente la tarea 1 se ejecuta sin tener que esperar. La tarea 2 debe esperar hasta que la tarea 1 haya terminado; la tarea 3 debe esperar hasta que las tareas 1 y 2 hayan terminado, y así sucesivamente. Por el contrario, en el proceso paralelo (por ejemplo, un servidor paralelo en un multiprocesador simétrico), más energía de la CPU es asignada a las tareas. Cada tarea independiente se ejecuta inmediatamente en su propio procesador: no hay hora de espera implicada.

La Figura 2.4.9 muestra el proceso de ejecución en paralelo.

Figura 2.4. 9
Proceso en paralelo: ejecución de tareas en paralelo

Los sistemas paralelos mejoran la velocidad de procesamiento y de E/S mediante la utilización de CPU y discos en paralelo. La fuerza que ha impulsado a los sistemas paralelos de bases de datos ha sido la demanda de aplicaciones que han de manejar bases de datos extremadamente grandes (del orden de terabytes, esto es, 1012 bytes) o que tienen que procesar un número enorme de transacciones por segundo (del orden de miles de transacciones por segundo).

La puesta en práctica eficaz del proceso paralelo implica dos desafíos:
-Tareas de estructuración de modo que ciertas tareas puedan ejecutarse al mismo tiempo (en paralelo).
-Preservar y ordenar las tareas que se deben ejecutar en serie.

Un sistema de proceso paralelo tiene las características siguientes:

- Cada procesador en un sistema puede realizar tareas concurrentemente.
- Las tareas pueden necesitar ser sincronizadas.
- Y cuenta con nodos recursos de la parte generalmente, tales como datos, discos, y otros dispositivos.

Los elementos dominantes del proceso paralelo son: Speedup y Scaleup; las metas de la sincronización del proceso paralelo: Un factor crítico del éxito Fijación Mensajería

Parámetros de rendimiento en computación paralela

Es posible medir las metas del funcionamiento del proceso paralelo en términos de las siguientes características importantes:

Velocidad de ejecución *rate (R)*. Mide *salidas (outputs)* por unidad de tiempo. Según la naturaleza de las salidas, tendremos:

- 1 que se miden con MIPS (instrucciones x =9por segundo).

- que se miden con MOPS (operaciones x =9por segundo).

$$R = \frac{instrucciones}{segundo}$$

- 10^6 que se miden con MFLOPS (op's coma flotante x =9por segundo).

Aceleración (Speedup) es el grado a el cual más hardware puede realizar la misma tarea en menos tiempo que el sistema original. Con hardware agregado, el speedup lleva a cabo la constante de la tarea y las medidas miden el tiempo de ahorros. La figura 2.4.10 demuestra cómo cada sistema paralelo del hardware realiza la mitad de la tarea original por la mitad el tiempo requerido para realizarlo en un solo sistema.

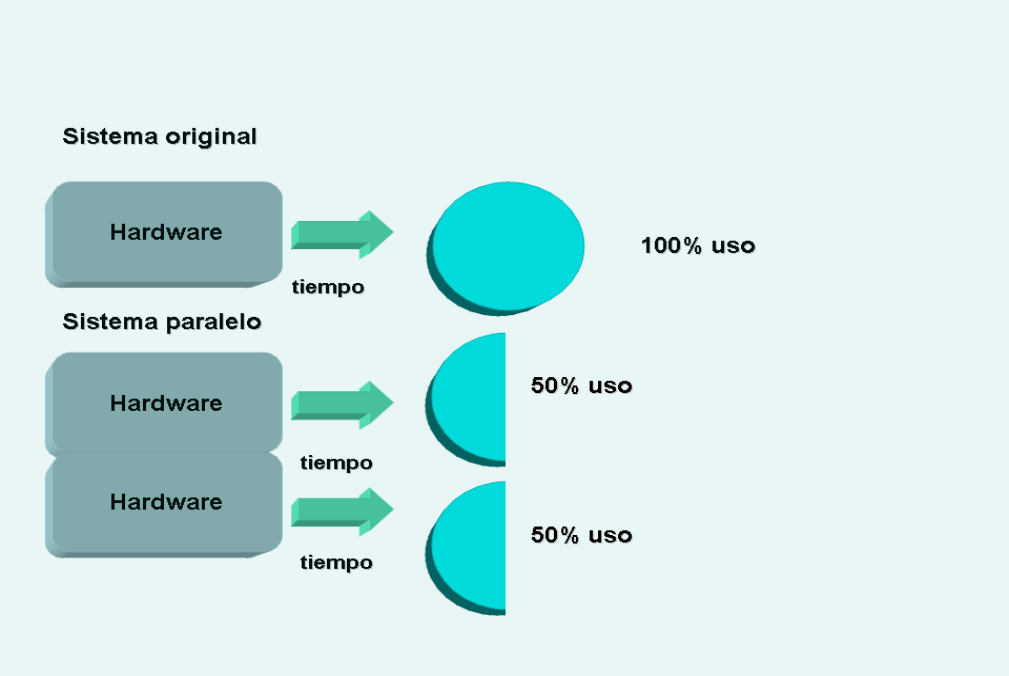

Figura 2.4. 10
Speedup

Con el speedup, los procesadores adicionales reducen tiempo de reacción de sistema. Se puede medir el speedup usando está fórmula:

$$Speedup = \frac{Time\_Original}{Time\_Parallel}$$

Dónde:

Time_Original es el tiempo transcurrido pasado por un sistema pequeño en la tarea dada.

Time_Parallel es el tiempo transcurrido pasado por un sistema más grande, paralelo en la tarea dada.

Por ejemplo, si el sistema original tomara 60 segundos para realizar una tarea, y dos sistemas paralelos tomó 30 segundos, entonces el valor del speedup sería 2.

2 = 60 / 30

Un valor de n, donde n mide el tiempo de más hardware se indica el ideal del speedup lineal: cuando más hardware puede realizar dos veces la misma tarea por la mitad del tiempo (o cuando tres veces más hardware realiza la misma tarea en una tercera parte del tiempo, y así sucesivamente).

Para la mayoría de los usos de OLTP, ningún speedup no puede esperar: solamente scaleup. El de arriba debido a la sincronización puede, de hecho, causar velocidad baja.

Scaleup es el factor m que expresa cuánto trabajo más se puede hacer en el mismo período por los tiempos del sistema en una n más grandes. Con hardware agregado, una fórmula para el scaleup lleva a cabo la constante del tiempo, y mide el tamaño creciente del trabajo que puede ser realizado.

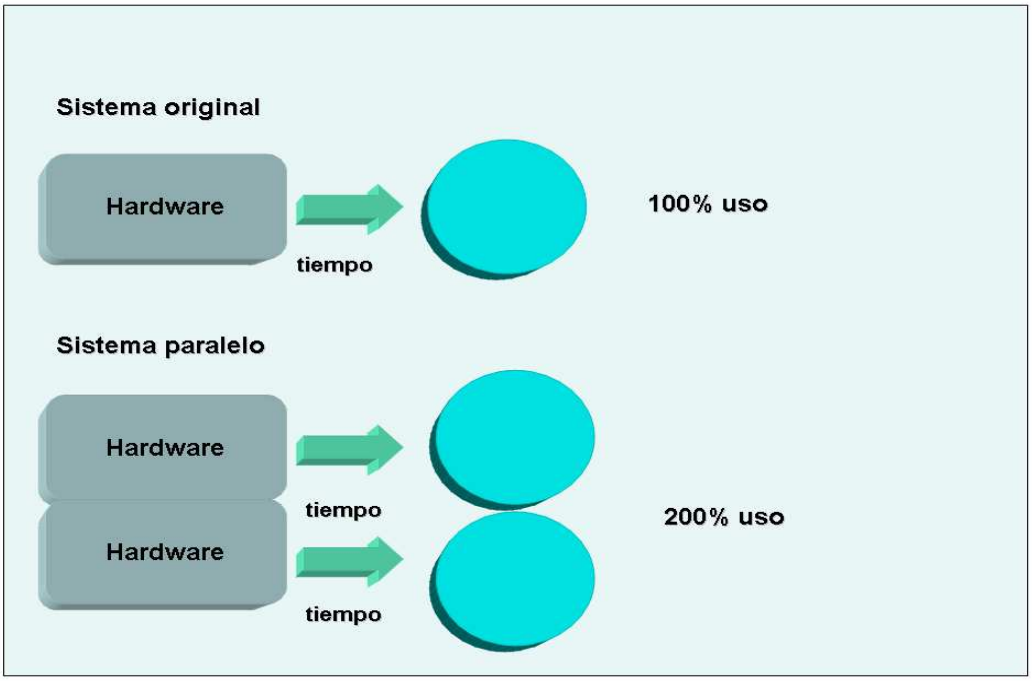

Figura 2.4. 11
Scaleup

Con el scaleup, representado en la figura 2.4.11, si los volúmenes de la transacción crecen, puedes mantener tiempo de reacción constante agregando recursos de hardware tales como CPU.

Se mede el scaleup usando la fórmula:

$$Scaleup = \frac{Volume\_Parallel}{Volume\_Original}$$

Dónde:

Volume_Original es el volumen de la transacción procesado en una cantidad de tiempo dada en un sistema pequeño.

Volume_Parallel es el volumen de la transacción procesado en una cantidad de tiempo dada en un sistema paralelo.

Por ejemplo, si el sistema original puede procesar 100 transacciones en una cantidad de tiempo dada, y el sistema paralelo puede procesar 200 transacciones en esta cantidad de tiempo, después el valor del scaleup sería igual a 2. Es decir, 200/100 = 2. Un valor de 2 indica que es ideal el scaleup lineal: cuando más hardware puede procesar dos veces el volumen de los datos en la misma cantidad de tiempo.

Eficiencia (E). Ratio entre la aceleración de una ejecución paralela y el número de procesadores.

$$p$$

Redundancia (R). Ratio entre el número de operaciones realizadas utilizando 1 procesadores en una ejecución paralela y su correspondiente ejecución serie en 1 procesador.

$$1 \leq S_p \leq p$$

Utilización (U). Ratio entre

$$E = \frac{S_p}{p} = \frac{t_1}{p\,t_p}$$

y el número de operaciones que podrían realizarse utilizando un procesador en t1 unidades de tiempo.

$$R = \frac{O_p}{O_1}$$

El proceso paralelo puede beneficiar ciertas clases de usos proporcionando:

- Rendimiento de procesamiento realzado: Scaleup
- Respuesta mejorada Tiempo: Speedup
- El tiempo de reacción mejorado puede ser alcanzado dividiendo una tarea grande en componentes más pequeños o reduciendo tiempo de la espera.

Rendimiento de procesamiento realzado: Scaleup

Si las tareas pueden funcionar independientemente una de otra, pueden ser distribuidas a diversas CPU o a los nodos y allí serán un scaleup: más procesos podrán funcionar a través de la base de datos en la misma cantidad de tiempo. Si los procesos pueden funcionar diez veces más rápidamente, entonces el sistema puede lograr diez veces más en la cantidad de tiempo original.

La característica paralela de la tarea ejecutada, por ejemplo, permite el scaleup: un sistema pude mantener el mismo tiempo de reacción si los datos ejecutados aumenta diez veces, o si más usuarios pueden ser servidos.

Respuesta mejorada Tiempo: Speedup

Los sistemas de Depósitos de datos y los querys en paralelo pueden usar un scaleup con proceso en paralelo: cada transacción corre rápidamente.

Para las aplicaciones OLTP no pueden ser alcanzado un scaleup solamente speedup. Con las aplicaciones OLTP cada proceso es independiente.

La tecnología paralela de la base de datos puede beneficiar ciertas clases de usos permitiendo:

-Un rendimiento más alto:
Con más CPU disponibles, un speedup y un scaleup más altos pueden ser logrados. La mejora en funcionamiento depende del grado de actividades de la fijación y de la sincronización del inter-nodo. El volumen de las operaciones y del contenido de la base de datos, así como el rendimiento de procesamiento y el funcionamiento del IDLM (Integrated Distributed Lock Manager), determina en última instancia la escalabilidad del sistema.

-Una disponibilidad más alta
Los nodos se aíslan entre si, así que una falta en un nodo no trae al sistema abajo. Los nodos restantes pueden recuperar del nodo faltante y continuar proporcionando datos y acceso a los usuarios. Esto significa que los datos están mucho más disponibles que si estarían con un solo nodo a la falta del nodo, y las cantidades a una disponibilidad perceptiblemente más alta de la base de datos.

-Mayor flexibilidad
Un ambiente paralelo de un servidor debe ser extremadamente flexible. Los casos se pueden asignar o desasignar cuando sea necesario. Cuando hay alta demanda para la base de datos, más casos pueden ser asignados temporalmente. Los casos se pueden desasignar y utilizar para otros propósitos una vez que ya no sean necesarios.

-Más usuarios
La tecnología paralela de la base de datos puede permitir superar límites de la memoria, permitiendo a un solo sistema servir a millares de usuarios.

La ganancia de velocidad y la ampliabilidad son dos aspectos importantes en el estudio del paralelismo. La ganancia de velocidad se refiere a la ejecución en menos tiempo de una tarea dada mediante el incremento del grado del paralelismo. La ampliabilidad se refiere al manejo de transacciones más largas mediante el incremento del grado de paralelismo. La ampliabilidad es normalmente el factor más importante para medir la eficiencia de un sistema paralelo de bases de datos. El objetivo del paralelismo en los sistemas de bases de datos suele ser asegurar que la ejecución del sistema continuará realizándose a una velocidad aceptable, incluso en el caso de que aumente el tamaño de la base de datos o el número de transacciones. El incremento de la capacidad del sistema mediante el incremento del paralelismo proporciona a una empresa un

modo de crecimiento más suave que el de reemplazar un sistema centralizado por una máquina más rápida.

Existen algunos factores que trabajan en contra de la eficiencia del paralelismo y pueden atenuar tanto la ganancia de velocidad como la ampliabilidad:

- Costes de inicio. El inicio de un único proceso lleva asociado un coste de inicio. En una operación paralela compuesta por miles de procesos, el tiempo de inicio puede llegar a ser mucho mayor que el tiempo real de procesamiento, lo que influye negativamente en la ganancia de velocidad.

- Interferencia. Como los procesos que se ejecutan en un sistema paralelo acceden con frecuencia a recursos compartidos, pueden sufrir un cierto retardo como consecuencia de la interferencia de cada nuevo proceso en la competencia con los procesos existentes por el acceso a los recursos más comunes, como el bus del sistema, los discos compartidos o incluso los bloqueos. Este fenómeno afecta tanto a la ganancia de velocidad como a la ampliabilidad.

- Sesgo. Al dividir cada tarea en un cierto número de pasos paralelos se reduce el tamaño del paso medio. Es más, el tiempo de servicio de la tarea completa vendrá determinado por el tiempo de servicio del paso más lento. Normalmente es difícil dividir una tarea en partes exactamente iguales, entonces se dice que la forma de distribución de los tamaños es sesgada. Por ejemplo, si se divide una tarea de tamaño 100 en 10 partes y la división está sesgada, puede haber algunas tareas de tamaño menor que 10 y otras de tamaño superior a 10; si el tamaño de una tarea fuera 20, la ganancia de velocidad que se obtendría al ejecutar las tareas en paralelo sólo valdría 5 en vez de lo que se esperara, 10.

Existen varios modelos de arquitecturas para las máquinas paralelas:
Memoria compartida. Todos los procesadores comparten una memoria común. normalmente a través de un bus o de una red de interconexión. El beneficio de la memoria compartida es la extremada eficiencia en cuanto a la comunicación entre procesadores.
Disco compartido. Todos los procesadores comparten un disco común.
Sin compartimiento. Los procesadores no comparten ni memoria ni disco.
Jerárquico. Es un híbrido de las anteriores.

Memoria compartida

En una arquitectura de memoria compartida, los procesadores y los discos tienen acceso a una memoria común, normalmente a través de un bus o de una red de interconexión. El beneficio de la memoria compartida es la extremada eficiencia en cuanto a la comunicación entre procesadores (cualquier procesador puede acceder a los datos de la memoria compartida sin necesidad de la intervención del software). Un procesador puede enviar mensajes a otros procesadores utilizando

escrituras en la memoria, de modo que la velocidad de envío es mucho mayor (normalmente es inferior a un microsegundo) que la que se alcanza con un mecanismo de comunicación. El inconveniente de las máquinas con memoria compartida es que la arquitectura no puede ir más allá de 32 o 64 procesadores, porque el bus o la red de interconexión se convertirían en un cuello de botella (ya que está compartido por todos los procesadores). Llega un momento en el que no sirve de nada añadir más procesadores ya que éstos emplean la mayoría de su tiempo esperando su turno para utilizar el bus y así poder acceder a la memoria. Las arquitecturas de memoria compartida suelen dotar a cada procesador de una memoria caché muy grande para evitar las referencias a la memoria compartida siempre que sea posible. No obstante, en la caché no podrán estar todos los datos y no podrá evitarse el acceso a la memoria compartida. Por estas razones, las máquinas con memoria compartida no pueden extenderse llegado un punto; las máquinas actuales con memoria compartida no pueden soportar más de 64 procesadores.

Disco compartido

En el modelo de disco compartido, todos los procesadores pueden acceder directamente a todos los discos a través de una red de interconexión, pero los procesadores tienen memorias privadas. Las arquitecturas de disco compartido ofrecen dos ventajas respecto de las de memoria compartida. Primero, el bus de la memoria deja de ser un cuello de botella, ya que cada procesador dispone de memoria propia.

Segundo, esta arquitectura ofrece una forma barata para proporcionar una cierta tolerancia ante fallos: si falla un procesador (o su memoria), los de más procesadores pueden hacerse cargo de sus tareas, ya que la base de datos reside en los discos, a los cuales tienen acceso todos los procesadores. La arquitectura de disco compartido tiene aceptación en bastantes aplicaciones; los sistemas construidos siguiendo esta arquitectura suelen denominarse agrupaciones (clusters).

El problema principal de los sistemas de discos compartidos es, de nuevo, la ampliabilidad. La interconexión con el subsistema de discos es ahora el nuevo cuello de botella. Esto es especialmente grave en situaciones en las que la base de datos realiza un gran número de acceso a los discos. Los sistemas de discos compartidos pueden soportar un mayor número de procesadores, en comparación con los sistemas de memoria compartida, pero la comunicación entre los procesadores es más lenta (hasta unos pocos milisegundos, si se carece de un hardware de propósito especial para comunicaciones), ya que se realiza a través de una red de interconexión.

Sin compartimiento

En un sistema sin compartimiento, cada nodo de la máquina consta de un procesador, memoria y uno o más discos. Los procesadores de un nodo pueden comunicarse con un procesador de otro nodo utilizando una red de interconexión de alta velocidad. Un nodo funciona como una red de interconexión de alta velocidad. Un nodo funciona como el servidor de los datos almacenados en los discos que posee. El modelo sin compartimiento salva el inconveniente de requerir que todas las operaciones de E/S vayan a través de una única red de interconexión, ya que las referencias a los discos locales son servidas por los discos locales de cada procesador; solamente van por la red las peticiones, los accesos a discos remotos y las relaciones de resultados. Es más, habitualmente, las redes de interconexión para los sistemas sin compartimiento se diseñan para ser ampliables, por lo que su capacidad de transmisión crece a medida que se añaden nuevos nodos. Como consecuencia, las arquitecturas sin compartimiento son más ampliables y pueden soportar con facilidad un gran número de procesadores. El principal inconveniente de los sistemas sin compartimiento es el coste de comunicación y de acceso a discos remotos, coste que es mayor que el que se produce en las arquitecturas de memoria o disco compartido, ya que el envío de datos provoca la intervención del software en ambos extremos.

Jerárquica

La arquitectura jerárquica combina las características de las arquitecturas de memoria compartida, de disco compartido y sin compartimiento. A alto nivel, el sistema está formado por nodos que están conectados mediante una red de interconexión y que no comparten ni memoria ni discos. Así, el nivel más alto es una arquitectura sin compartimiento. Cada nodo del sistema podría ser en realidad un sistema de memoria compartida con algunos procesadores. Alternativamente, cada nodo podría ser un sistema de disco compartido y cada uno de estos sistemas de disco compartido podría ser a su vez un sistema de memoria compartida. De esta manera, un sistema podría construirse como una jerarquía con una arquitectura de memoria compartida con pocos procesadores en la base, en lo más alto una arquitectura sin compartimiento y quizá una arquitectura de disco compartido en el medio.

Las principales diferencias entre las bases de datos paralelas sin compartimientos y las bases de datos distribuidas son las siguientes: las bases de datos distribuidas normalmente se encuentran en varios lugares geográficos distintos, se administran de forma separada y poseen una interconexión más lenta. Otra gran diferencia es que en un sistema distribuido se dan dos tipos de transacciones, las locales y las globales.

Una transacción local es aquella que accede a los datos del único emplazamiento en el cual se inició la transacción. Por otra parte, una transacción global es aquella

que o bien accede a los datos situados en un emplazamiento diferente de aquel en el que se inició la transacción, o bien accede a datos de varios emplazamientos distintos.

Paralelismo en Depósito de datos y OLAP

Los Depósitos de datos dan soporte a tecnologías OLAP (On-Line Analytical Processing), para que de forma eficiente, permitan a los analistas, managers, y ejecutivos, extraer la información necesaria para la toma de decisiones en entornos empresariales. El procesamiento de consultas de tipo OLAP es de un costo computacional muy elevado, y con un acceso masivo a disco debido al gran flujo de datos procesado. Este tipo de entornos requiere de poderosas arquitecturas paralelas con el fin de obtener un eficiente rendimiento global del sistema.

Multiprocesadores simétricos con memoria compartida (SMP), han sido ampliamente utilizados para mejorar el throughput en sistemas de bases de datos paralelos. En un sistema SMP, todos los procesadores comparten recursos de disco y memoria bajo el control de una copia del sistema operativo. Los procesadores acceden a la memoria a través de buses de alta velocidad y modernas redes de interconexión. Sin embargo, estas arquitecturas presentan problemas de escalabilidad debido a limitaciones de disco, de memoria o de contención en la red.

Cuando los Depósitos de datos han de escalar por encima del número de procesadores que puede proporcionar una arquitectura SMP, o cuando las aportaciones de un sistema de altas prestaciones son deseables, entonces, las arquitecturas cluster son la opción escogida. El uso de arquitecturas cluster se ha convertido en una solución muy común para el soporte de aplicaciones que requieren de paralelismo masivo y, en entornos Depósito de datos, se ha hecho imprescindible para alcanzar un buen rendimiento. En noviembre del 2004, cerca del 60% de los supercomputadores en la lista ’TOP500’ eran arquitecturas cluster, alcanzando un 72% en el 2005. Este tipo de arquitecturas se basan en un diseño hardware sin recursos compartidos, y están compuestas de varios nodos que se comunican entre ellos a través de una red de interconexión. En dichos sistemas, la base de datos está particionada horizontalmente entre los nodos del sistema. El particionado horizontal permite que, durante la ejecución de una consulta, se pueda realizar una distribución equitativa del trabajo. De esta forma, se alcanza un óptimo rendimiento cuando cada nodo puede trabajar de forma local con su partición de la base de datos, y con la mínima comunicación de datos posible.

Los nodos con una configuración SMP se presentan como los bloques óptimos para construir un sistema cluster (véase Figura 2.4.12). Las arquitecturas formadas por varios nodos SMP, también conocidas como arquitecturas CLUMP,

ofrecen una gran escalabilidad y efectividad, siendo comúnmente utilizadas en entornos del Depósito de datos.

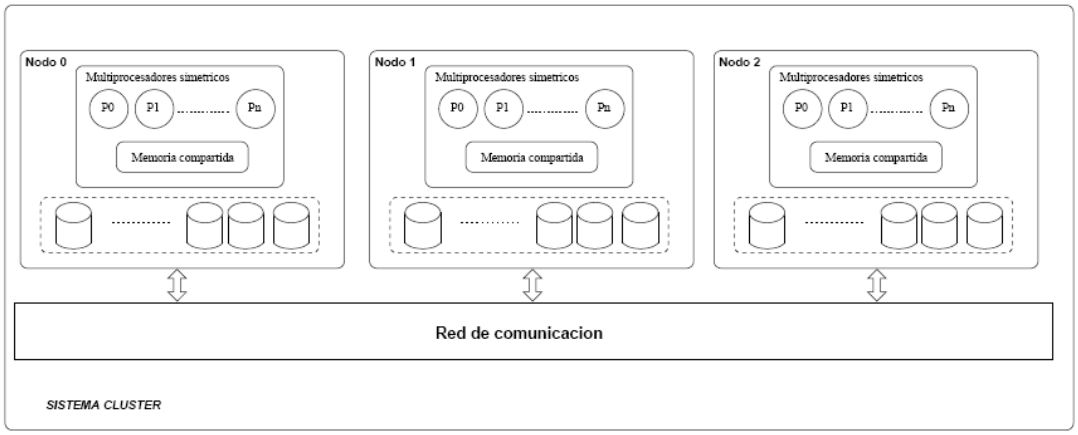

Figura 2.4. 12
Ejemplo de sistema cluster con 3 nodos con configuración SMP

Particionado de datos

Originariamente, la partición de una relación implica distribuir sus tuplas a través de los múltiples discos en una máquina sin recursos compartidos. De esta forma, la partición horizontal de los datos permite a las bases de datos paralelas explotar el ancho de banda de la E/S, leyendo y escribiendo en paralelo sobre múltiples discos. La misma filosofía se extiende a las arquitecturas cluster pero substituyendo discos por nodos.

La partición de una base de datos en arquitecturas cluster puede seguir diferentes esquemas, que son los encargados de decidir el nodo destino de cada una de las tuplas almacenadas en las tablas de la base de datos. Los esquemas de partición más utilizados son:

Particionado Round-Robin. La partición es de forma equitativa las tuplas de la base de datos entre los clusters del sistema sin tener en cuenta la naturaleza, ni los valores, de los datos almacenados. Round-Robin permite un balanceo de carga equitativo para accesos secuenciales a las tablas de la base de datos.

Particionado por rango. Particiona las tablas de la base de datos según los rangos establecidos sobre los valores de uno de sus atributos. El principal problema de este tipo de particionado es que no es equitativo, y puede provocar un claro desbalanceo de carga entre los nodos del sistema.

Particionado hash. Particiona las tablas de la base de datos a través de una función de hash que se aplica sobre los valores de uno o varios atributos de cada

tabla. El atributo/s sobre los que se aplica la función de hash se denominan clave de particionado. Por cada tupla de una tabla, la función de hash retorna un valor que especifica el nodo destino en el que se ha de almacenar. De esta forma se logra una distribución equitativa, y con conocimiento del valor de los datos por los que se particiona la base de datos.

La Figura 2.4.13 ilustra el particionado hash del esquema estrella de la Figura 2.4.12 sobre una arquitectura cluster con 4 nodos. Las claves de particionado sobre las que se aplica el particionado son o orderkey, l orderkey, p partkey, ps partkey/ps suppkey y s suppkey para las tuplas de las tablas orders, lineitem, part, partsupp y supplier respectivamente. La función de hash retorna un valor entre 0 y 3 que determina el nodo en el que se debe de almacenar cada tupla. Si la función de hash es lo suficientemente precisa y asumiendo una distribución uniforme de los datos, entonces se obtiene un buen balanceo de los datos entre los nodos del sistema.

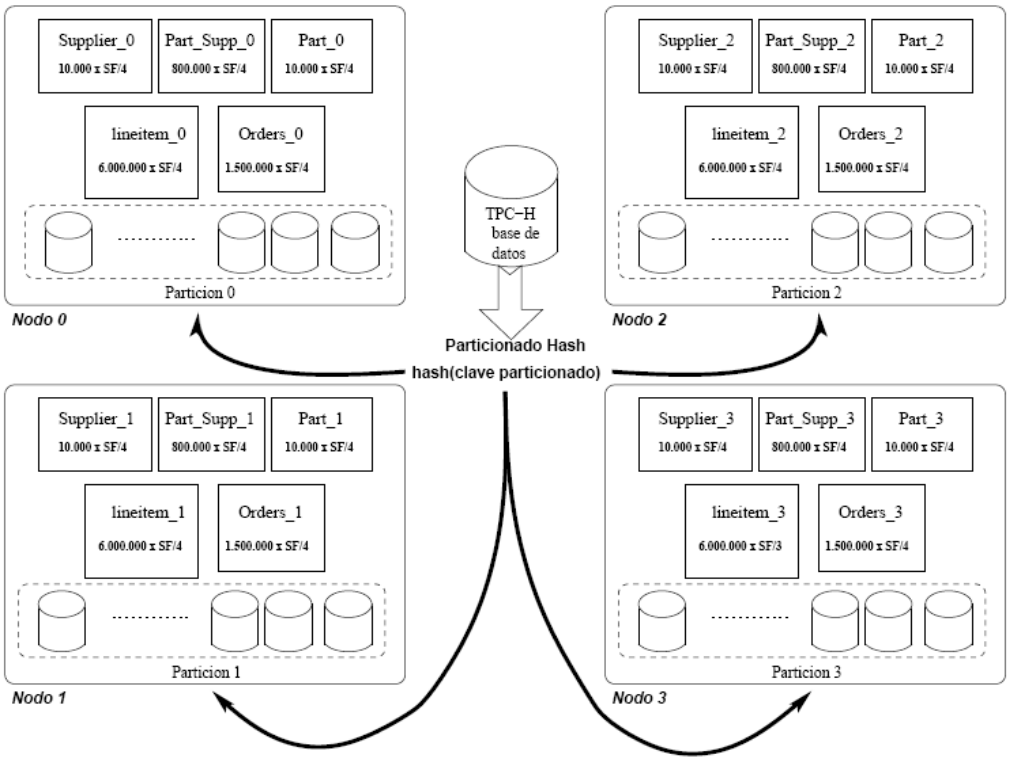

Figura 2.4. 13

TPH-H. Esquema de partición Hash

La operación de join paralela

Cuando se ejecuta una operación de join sobre arquitecturas cluster, y asumiendo un esquema de partición hash como el descrito anteriormente, entonces, se dice que el join es localizado si la clave de join es la misma que la clave de partición de las tablas involucradas en la operación. De la misma forma, se dice que una tabla está localizada si su clave de partición coincide con la clave de la operación de join. Cuando el join es localizado, se puede realizar la ejecución de forma local en cada nodo, pues, en este caso, la partición hash nos asegura que las tuplas de un nodo no harán join con las tuplas de un nodo remoto. Si la clave de partición no coincide con la clave de join, entonces, se dice que el join es no localizado, y precisa de comunicación de datos para su ejecución. El cómo se comunican los datos durante la ejecución de un join no localizado depende del estado de las tablas involucradas en la operación:

Una de las dos tablas no está localizada. En este caso, bien se reparticiona la tabla no localizada por la clave de join, o bien se realiza un broadcast de una de las dos tablas. La decisión entre una opción u otra la toma el optimizador en función de su coste. Notar que la operación de broadcast suele ser de un coste muy elevado, y en este caso, la mejor opción suele ser reparticionar la tabla no localizada.

Las dos tablas no están localizadas. Para este caso, bien se reparticionan las dos tablas por la clave de join, o bien se realiza el broadcast de una de las dos tablas. De nuevo la decisión la toma el optimizador en función del coste de cada una. En este caso, sin embargo, puede ser que el reparticionado de dos tablas resulte más costoso que realizar la operación broadcast.

En general, la paralelización de un join cuando las relaciones están localizadas es simple: el plan de ejecución se replica en todos los nodos, de forma que hay un operador de join en cada nodo ejecutándose en paralelo sobre la partición de la base de datos que le corresponde. La paralelización de un join no localizado añade más complejidad al plan de ejecución. En este caso, un nuevo operador de reparticionado es añadido para comunicar los datos de las relaciones entrantes del join no localizado.

Cada operador de reparticionado debe de llevar a cabo dos acciones diferentes: enviar datos al resto de los nodos del sistema, así como recibir datos de cada uno de estos nodos. En el primero de los casos, el operador de reparticionado actúa como el sending end, y en el segundo caso actúa como el receiving end. El operador de reparticionado tiene conocimiento del esquema de particionado llevado a cabo y de la topología de los nodos involucrados en la ejecución de la consulta. De esta forma, para cada tupla proyectada por los nodos inferiores del plan de ejecución local, el operador de reparticionado decide si la tupla tiene que ser procesada de forma local (es proyectada directamente al operador local

inmediatamente superior del plan de ejecución), o si tiene que ser procesada por uno o más nodos (es enviada a través de la red de interconexión).

2.4.5 Implementación de un Depósito de datos en un ambiente multiservidor con tecnología paralela

Para la programación paralela y los problemas que pueden surgir cuando se trata de utilizar este paradigma es necesario considerar el sistema sobre el que se programa y se lanzan los programas paralelos: ya sean multiprocesadores o, como será el caso, multicomputadores.

En el capítulo anterior se definió lo que es paralelismo. Con este concepto se indicará todos los conceptos que implica la implementación de un Depósito de datos en un ambiente multiservidor con tecnología paralela. En general en esta se tocarán puntos acerca de todos los conceptos que interesan a la hora de programar aplicaciones y hacerlo utilizando de una manera u otra el paralelismo.

Cuando se tienen demasiados usos en el funcionamiento de una máquina, se puede obtener información para mejorar el funcionamiento. Sin embargo, si el servidor de la base de datos está alcanzando sus límites de proceso puede ser que desees trasladarse a una máquina más grande o a un sistema del multinodo. Recordar que una configuración del cliente/ servidor requiere que todas las comunicaciones entre el uso del cliente y la base de datos ocurran sobre la red. Esto puede no ser apropiado donde está un volumen muy alto de tales comunicaciones requerido.

Un servidor paralelo puede consolidar varias bases de datos para simplificar tareas administrativas. Las bases de datos múltiples pueden proporcionar mayor disponibilidad que un solo caso que tiene acceso a una sola base de datos, porque una falta del caso en un sistema de la base de datos distribuida no previene el acceso a los datos en las otras bases de datos: solamente la base de datos poseída por el caso fallado es inaccesible. Un servidor paralelo, sin embargo, permite el acceso continuado a todos los datos cuando un caso falla, incluyendo los datos que fueron alcanzados por caso que funciona en el nodo fallado.

Un servidor paralelo que tiene acceso a una sola base de datos consolidada puede evitar la necesidad de las actualizaciones distribuidas, rellenos, o las cancelaciones y un bifásico más costoso permite que una transacción en cualquier nodo escriba a las tablas múltiples simultáneamente, sin importar cuales son los nodos que escriben generalmente a esas tablas.

El ambiente multiservidor con tecnología paralela es un cluster y podemos entenderlo como:

Un conjunto de máquinas unidas por una red de comunicación trabajando por un servicio conjunto. Según el servicio puede ser dar alta disponibilidad, alto rendimiento, etc...

Hay definiciones que distinguen entre cluster de máquinas SMP y clusters formados por nodos monoprocesadores. Hay arquitecturas clusters que se denominan *constelaciones* y se caracterizan por que cada nodo contiene más procesadores que el número de nodos. A pesar de todo, las constelaciones siguen siendo clusters de componentes o nodos aventajados y caros.

Otras definiciones con respecto al tema son:

Un cluster consiste en un conjunto de máquinas y un servidor de cluster dedicado, para realizar los relativamente infrecuentes accesos a los recursos de otros procesos, se accede al servidor de cluster de cada grupo del libro Operating System Concepts de Silberschatz Galvin.

Un cluster es la variación de bajo precio de un multiprocesador masivamente paralelo (miles de procesadores, memoria distribuida, red de baja latencia), con las siguientes diferencias: cada nodo es una máquina quizás sin algo del hardware (monitor, teclado, mouse, etc.), el nodo podría ser SMP. Los nodos se conectan por una red de bajo precio como Ethernet o ATM aunque en clusters comerciales se pueden usar tecnologías de red propias. El interfaz de red no está muy acoplado al bus I/O. Todos los nodos tienen disco local. Cada nodo tiene un sistema operativo UNIX con una capa de software para soportar todas las características del cluster del libro Scalable Parallel Computing de Kai Hwang y Khiwei Xu.

Es una clase de arquitectura de computador paralelo que se basa en unir máquinas independientes cooperativas integradas por medio de redes de interconexión para proveer un sistema coordinado, capaz de procesar una carga del autor Dr. Thomas Sterling.

Un cluster es un tipo particular de computadora paralela, es decir, un conjunto de computadoras que pueden trabajar de manera coordinada en la solución de un mismo problema. Aunque no existe un acuerdo general en cuanto a la definición exacta de lo que significa un cluster, y muchas computadoras diferentes pueden llegar a ser clasificadas como tales, dos de las características más aceptadas de estos equipos y quizás las que más han influido en su rápido desarrollo son que se construyen a partir de componentes que pueden encontrarse en el mercado común de cómputo, lo que en EU se conoce como “commodity of the shelf” (COTS), y que se desarrollan bajo el esquema de “hágalo usted mismo”.

No hay una un acuerdo en la definición sin embargo si se tiene un acuerdo en las características.

Para crear un cluster se necesitan al menos dos nodos. Una de las características principales de estas arquitecturas es que exista un medio de comunicación (red) donde los procesos puedan migrar para comunicarse en diferentes estaciones paralelamente. Un solo nodo no cumple este requerimiento por su condición de aislamiento para poder compartir información. Las arquitecturas con varios procesadores en placa tampoco son consideradas clusters, bien sean máquinas SMP o mainframes, debido a que el bus de comunicación no suele ser de red, sino interno.

Por esta razón se deduce la primera característica de un cluster:

1.- Un cluster consta de 2 o más nodos.

Los nodos necesitan estar conectados para llevar a cabo su objetivo.

2.- Los nodos de un cluster están conectados entre sí por al menos un canal de comunicación.

3.- Los clusters necesitan software de control especializado.

El problema también se plantea por los distintos tipos de clusters, cada uno de ellos requiere un modelado y diseño del software distinto.

Como es obvio las características del cluster son completamente dependientes del software, por lo que no se tratarán las funcionalidades del software sino el modelo general de software que compone un cluster.

El software se debe dedicar a la comunicación entre los nodos. Existen varios tipos de software que pueden conformar un cluster:

Software a nivel de aplicación.

Este tipo de software se sitúa a nivel de aplicación, se utilizan generalmente bibliotecas de carácter general que permiten la abstracción de un nodo a un sistema conjunto, permitiendo crear aplicaciones en un entorno distribuido de manera lo más abstracta posible. Este tipo de software suele generar elementos de proceso del tipo rutinas, procesos o tareas, que se ejecutan en cada nodo del cluster y se comunican entre sí a través de la red. (Parte inferior figura 2.4.14)

Software a nivel de sistema.

Este tipo de software se sitúa a nivel de sistema, suele estar implementado como parte del sistema operativo de cada nodo, o ser la totalidad de éste. (Parte superior figura 2.4.14)

Es más crítico y complejo, por otro lado suele resolver problemas de carácter más general que los anteriores y su eficiencia, por norma general, es mayor.

A pesar de esta división existen casos en los cuales se hace uso de un conjunto de piezas de software de cada tipo para conformar un sistema cluster completo. Son implementaciones híbridas donde un cluster puede tener implementado a nivel de kernel parte del sistema y otra parte estar preparada a nivel de usuario.

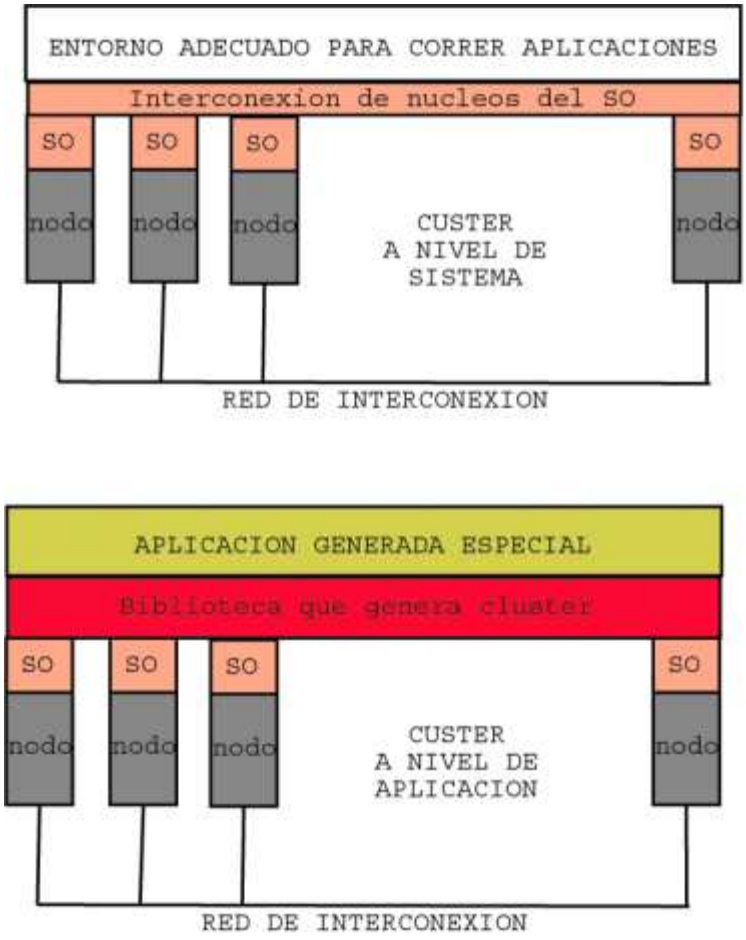

Figura 2.4. 14

Cluster a nivel de sistema y nivel de aplicación

Acoplamiento de un cluster

Dependiendo del tipo de software, el sistema puede estar más o menos acoplado.

Se entiende por acoplamiento del software a la integración que tengan todos los elementos software que existan en cada nodo. Gran parte de la integración del sistema la produce la comunicación entre los nodos, y es por esta razón por la que se define el acoplamiento; otra parte es la que implica cómo de crítico es el software y su capacidad de recuperación ante errores.

Depende también si el sistema es centralizado o distribuido. En cualquier caso, el acoplamiento del software es una medida subjetiva basada en la integración de un sistema cluster a nivel general.

Se distingue 3 tipos de acoplamiento:

- Acoplamiento fuerte
- Acoplamiento medio
- Acoplamiento débil

Acoplamiento fuerte:

El software que entra en este grupo es software cuyos elementos se interrelacionan mucho unos con otros y posibilitan la mayoría de las funcionalidades del cluster de manera altamente cooperativa. El caso de acoplamiento más fuerte que se puede dar es que solamente haya una imagen del kernel del sistema operativo, distribuida entre un conjunto de nodos que la compartirán. Por supuesto algo fundamental es poder acceder a todas las partes de este sistema operativo, estrechamente relacionadas entre sí y distribuidas entre los nodos.

Este caso es el que se considera como más acoplado, de hecho no está catalogado como cluster, sino como sistema operativo distribuido.

Otro ejemplo son los cluster SSI, en estos clusters todos los nodos ven una misma imagen del sistema, pero todos los nodos tienen su propio sistema operativo, aunque estos sistemas están estrechamente relacionados para dar la sensación a las aplicaciones que todos los nodos son idénticos y se acceda de una manera homogénea a los recursos del sistema total.

Acoplamiento medio:

Este grupo pertenece un software que no necesita un conocimiento tan exhaustivo de todos los recursos de otros nodos, pero que sigue usando el software de otros

nodos para aplicaciones de muy bajo nivel. Como ejemplos hay openMosix y Linux-HA.

Acoplamiento débil:

Generalmente se basan en aplicaciones construidas por bibliotecas preparadas para aplicaciones distribuidas. Es el caso de por ejemplo PVM, MPI o CORBA. Éstos por sí mismos no funcionan en modo alguno con las características que antes se han descrito (como Beowulf) y hay que dotarles de una estructura superior que utilice las capacidades del cluster para que éste funcione.

En general la catalogación de los clusters se hace en base a cuatro factores de diseño bastante ortogonales entre sí:

- Acoplamiento
- Control
- Homogeneidad
- Seguridad

Por otro lado está el factor de control del cluster. El parámetro de control implica el modelo de gestión que propone el cluster. Este modelo de gestión hace referencia a la manera de configurar el cluster y es dependiente del modelo de conexión o colaboración que surgen entre los nodos. Puede ser de dos tipos:

- Control centralizado: se hace uso de un nodo maestro desde el cual se puede configurar el comportamiento de todo el sistema. Este nodo es un punto crítico del sistema aunque es una ventaja para una mejor gestión del cluster.
- Control descentralizado: en un modelo distribuido donde cada nodo debe administrarse y gestionarse. También pueden ser gestionados mediante aplicaciones de más alto nivel de manera centralizada, pero la mayoría de la gestión que hace el nodo local es leer archivos de configuración de su propio nodo.

Por otro lado se ha demostrado que es posible crear sistemas de una sola imagen o heterogéneos con una implementación práctica.

Los clusters heterogéneos son más difíciles de conseguir ya que se necesitan notaciones abstractas de transferencias e interfaces especiales entre los nodos para que éstas se entiendan, por otro lado los clusters homogéneos obtienen más beneficios de estos sistemas y pueden ser implementados directamente a nivel de sistema.

Homogeneidad de un cluster:

- Homogéneos: formados por equipos de la misma arquitectura. Todos los nodos tienen una arquitectura y recursos similares, de manera que no existen muchas diferencias entre cada nodo.
- Heterogéneos: formados por nodos con distinciones que pueden estar en los siguientes puntos.
    - Tiempos de acceso distintos
    - Arquitectura distinta
    - Sistema operativo distinto
    - Rendimiento de los procesadores o recursos sobre una misma arquitectura distintos

Existen otros muchos factores de diseño que limitan el comportamiento y modelado de un cluster. La imposibilidad de llegar a clusters que paralelicen cualquier proceso se basa en que la mayoría de las aplicaciones hacen uso, en mayor o menor medida, de algoritmos secuenciales no paralelizables.

Generalmente el diseño de un cluster se realiza para solucionar problemas de tipo:

- Mejora de rendimiento
- Abaratamiento del costo
- Distribución de factores de riesgo del sistema
- Escalabilidad

El modelo de los clusters permite que la mejora de rendimiento sea evidente respecto a grandes mainframes a un precio realmente accesible. Lo que explica a su vez el segundo punto, acerca del costo de los clusters, que permite relaciones rendimiento-precio que se acercan a un margen lineal dependiendo del cluster implementado.

Por otro lado esta la distribución de riesgos. La mayoría de los usuarios tienen sus servicios, aplicaciones, bases de datos o recursos en un solo ordenador, o dependientes de un solo ordenador. Otro paso más adelante es colocar las bases de datos replicadas sobre sistemas de archivos distribuidos de manera que estos no se pierdan por que los datos son un recurso importante.

Por último está el factor de escalabilidad. Cuanto más escalable es un sistema menos costará mejorar el rendimiento, lo cual abarata el costo, y en el caso de que el cluster lo implemente distribuye más el riesgo de caída de un sistema.

Todas éstas características dan inicio a los tipos de clusters que se ven a continuación.

-Alto rendimiento (HP, high performance)

Los clusters de alto rendimiento han sido creados para compartir el recurso más valioso de un ordenador, es decir, el tiempo de proceso. Cualquier operación que necesite altos tiempos de CPU puede ser utilizada en un cluster de alto rendimiento, siempre que se encuentre un algoritmo que sea paralelizable.

-Alta disponibilidad (HA, high availability)

Los clusters de alta disponibilidad son bastante ortogonales en lo que se refieren a funcionalidad a un cluster de alto rendimiento. Los clusters de alta disponibilidad pretenden dar servicios 7/24 de cualquier tipo, son clusters donde la principal funcionalidad es estar controlando y actuando para que un servicio o varios se encuentren activos durante el máximo periodo de tiempo posible. En estos casos se puede comprobar como la monitorización de otros es parte de la colaboración entre los nodos del cluster.

-Alta confiabilidad (HR, high reliability)

Por ultimo, están los clusters de alta confiabilidad. Estos clusters tratan de aportar la máxima confiabilidad en un entorno, en la cual se necesite saber que el sistema se va a comportar de una manera determinada. Puede tratarse por ejemplo de sistemas de respuesta en tiempo real.

Este tipo de clusters son los más difíciles de implementar. No se basan solamente en conceder servicios de alta disponibilidad, sino en ofrecer un entorno de sistema altamente confiable. Esto implica muchísima sobrecarga en el sistema, son también clusters muy acoplados.

Dar a un cluster SSI capacidad de alta confiabilidad implica gastar recursos necesarios para evitar que aplicaciones caigan.

Límites en el hardware

En lo que se refiere a explotación de programas paralelos en máquinas SMP hay que tener en cuenta el modo de ejecución de las instrucciones. Actualmente la mayoría de los ordenadores comerciales son de los llamados ordenadores superescalares: lanzan varias instrucciones a la vez aprovechando una división por fases. Esta división se construye sobre el principio que cada fase requiere de una parte específica del hardware del procesador para completarse. De esta manera se aprovechan mejor todo los componentes del procesador pues se utilizan a cada ciclo.

Para lograrlo deben solucionar de la manera más apropiada posible retardos por dependencias. Generalmente esto se soluciona con renombramiento de registros (fase que puede o no hacer el compilador), lanzamiento de instrucciones fuera de orden, unidades superescalares y supervectoriales y algunas otras.

Límites del software

Como se ha visto, la ley de Amdahl pone límites en lo que se refiere al incremento del rendimiento de cualquier sistema en el que se utilicen fragmentos de código no paralelizable, es decir, de todos los sistemas que actualmente se conocen. Una vez asumida dicha limitación, no queda más opción que optimizar los programas para que el rendimiento de los mismos en el sistema sea el mejor posible. Esto a menudo implica añadir una fase más en la vida de nuestros programas.

Así pues a las fases de análisis, diseño e implementación hay que añadir una fase de paralelización, en la que se deberá tener en cuenta las necesidades del programa y del problema a resolver, así como de los medios de los que se disponen para poder hacer que el programa aumente el rendimiento a medida que aumenta la capacidad de procesamiento de sistema.

Granularidad del paralelismo

Al hablar en los apartados anteriores acerca de la granularidad del paralelismo se hacía referencia a la implicación que tiene el paralelismo en el ámbito de la programación. De este modo, se puede decir que la granularidad en el paralelismo de un sistema se encuentra en paralelizar código de procesos, rutinas, módulos o bucles a nivel de instrucción. El término granularidad se usa como el mínimo componente del sistema que puede ser preparado para ejecutarse de manera paralela. Por norma general cuanto más fuertemente acoplado (en el sentido descrito en la sección Arquitecturas) en un sistema, menor es la granularidad del paralelismo en la programación. Dependiendo del grado de granularidad del sistema se diferencia en:

1. Sistemas de granularidad fina.
    1. bucles
    2. sentencias

    En general se hace a nivel de instrucciones en ensamblador. Generalmente son explotados por sistemas hardwares muy caros con los nodos o procesadores fuertemente acoplados. Es la granularidad más pequeña y se basa prácticamente todo su funcionamiento en propiedades del hardware. El hardware puede ser suficientemente inteligente para que el programador no tenga que hacer mucho por soportar esta granularidad, por ejemplo el hardware puede aportar reordenamiento de instrucciones.

2. Sistemas de granularidad media.
    1. módulos
    2. rutinas
    3. tareas o procesos

Dentro de estos sistemas se incluyen por ejemplo el RPC, openMosix y otros, si bien estos mismos están entre el paralelismo de grano grueso y el paralelismo de grano medio. El paralelismo de grano medio en general es explotado por el programador o el compilador.

Dentro de él también se encuentran diversas librerías como pueden ser PVM o MPI. El hardware normalmente también se prepara para poder aprovechar este tipo de paralelismo, por ejemplo, los procesadores pueden disponer de instrucciones especiales para ayudar en el cambio de una tarea a otra que realiza el sistema operativo.

3. Sistemas de granularidad gruesa.
    1. trabajos o programas
    2. módulos o procesos

El paralelismo de grano grueso es el que explota el programador mediante programas que no tienen por qué requerir la utilización de ninguna librería externa, sino solamente el uso de conocimientos de programación para paralelizar un algoritmo. Se basa principalmente en cualquier tipo de medio que utilice el programador para crear un programa, que solucione un problema de manera paralela, sin tener por que hacer uso más que de su habilidad de programador y de un buen algoritmo. Son los más limitados al carecer de métodos específicos para comunicación entre nodos o procesadores, se dan en sistemas muy débilmente acoplados.

Finalmente, la estandarización de las prácticas comunes de programación paralela mediante el surgimiento del estándar MPI (Message Passing Inter-face) y la popularización de la biblioteca PVM (Parallel Virtual Machine) proporcionaron las herramientas necesarias para la construcción de programas paralelos en computadoras como Beowulf; nuevamente, el hecho de que existieran implementaciones gratuitas tanto de MPI como de PVM, representaron un factor decisivo en la popularización de los clusters.

El problema de la transparencia

Uno de los mayores problemas que existen en la creación de programas que hagan uso de paralelización (quitando los de granularidad fina que ya se ha visto que son explotados a bajo nivel por el compilador y por el hardware) es la transparencia en la programación de dichos programas. A la hora de crear un programa que resuelva un problema mediante el uso de un algoritmo que explote de alguna manera la paralelización hay que conocer el sistema de ejecución. Dependiendo del sistema elegido y teniendo en cuenta que por norma general se

paralelizan tareas, procesos, procedimientos, rutinas o componentes distribuidos que realizan este trabajo, hay dos modelos de implementación:

Modelo de programación explícita:

En el que se requiere una biblioteca de funciones especiales que se encargan tanto de realizar la comunicación como los métodos de migración y demás factores que en un principio no debe afectar al programador, el cual debe abstraerse de esta capa. Este tipo de modelo es el que utiliza RPC, MPI o PVM. Requiere un conocimiento especial de dichas bibliotecas, lo que limita la velocidad de desarrollo del software y además lo hace más costoso debido al tiempo que se debe gastar en el conocimiento de las funciones.

Modelo de programación implícita.

Es un modelo quizá más atractivo. Basa todo su funcionamiento en que el programador sepa lo mínimo del sistema para paralelizar sus procesos. Generalmente este modelo lo explotan los compiladores especiales de un sistema particular.

Por otro lado se suele depender de macros o funciones especiales que delimitan la granularidad de los procesos a migrar. Lo ideal para obtener transparencia en la programación será programar de manera completamente implícita y que al mismo tiempo el sistema implantado fuese lo menos propenso a entradas de intrusos en lo que se refiere a comportamiento con el usuario final.

Paralelización de programas

El primer paso en el desarrollo de un programa paralelizado es, como siempre, plantear el problema mediante técnicas de divide y vencerás. Deberá localizarse lo paralelizable en el problema de manera abstracta antes de pensar en la paralelización del código, es decir que existen problemas inherentemente paralelos, como pueden ser la suma de las componentes de dos vectores según sus posiciones o el tratamiento de imágenes. Estos problemas se pueden resolver de manera óptima.

Teóricamente se puede paralelizar cualquier cosa que se haya diseminado mediante técnicas de divide y vencerás, procesos, módulos rutinas, o algoritmos paralelos completamente.

Existen dos formas bien conocidas y fáciles de comprender de paralelismo:

1. El paralelismo funcional divide las aplicaciones en funciones. Se podría ver como paralelismo de código. Por ejemplo puede dividirse en: entrada, preparación del problema, solución del problema, preparación de la salida,

salida y mostrar la solución. Esto permite a todos los nodos producir una cadena. Esta aproximación es como la segmentación en funciones de un procesador.

2. El paralelismo de datos se basa en dividir los datos que se tienen que procesar. Típicamente los procesos que están usando esos datos son idénticos entre sí y lo único que hacen es dividir la cantidad de información entre los nodos y procesarla en paralelo. Esta técnica es más usada debido a que es más sencillo realizar el paralelismo.

Arquitecturas

Tanto las arquitecturas hardware como software que se han desarrollado fueron inventadas para superar los problemas de rendimiento y responden a distintos enfoques para conseguir sacar más rendimiento gracias al paralelismo. Algunas arquitecturas pretenden tener una mejor relación velocidad/precio y otras ser las máquinas más rápidas del mercado.

Las soluciones hardware fueron las primeras en aparecer, las soluciones software tuvieron que esperar hasta que el hardware diese la potencia necesaria y en el caso de los sistemas distribuidos se tuvo que esperar a que en los años 70 se desarrollaran las redes de área local.

Soluciones hardware

Las soluciones hardware han estado en el mercado de la computación desde sus inicios. Para muchas empresas la única manera de crear mejores máquinas era crear arquitecturas paralelas a partir de las que ya poseían.
Como el número de estos tipos de máquinas es elevado, existen numerosas y diversas soluciones. Quizás la división más conocida y básica sea la taxonomía de Flint. Flint dividió las soluciones hardware en cuatro categorías:

SISD: un flujo de instrucciones único trabaja sobre un flujo de datos único, a esta categoría pertenecen las CPUs simples y las superescalares.
SIMD: un flujo de instrucciones único trabaja sobre un flujo de datos múltiple, en esta categoría tenemos los computadores matriciales.
MISD: un flujo de instrucciones múltiple trabaja sobre un flujo de datos único, resultado teórico de la clasificación, el modelo que más se acerca son los computadores sistólicos.
MIMD: un flujo de instrucciones múltiple trabaja sobre un flujo de datos múltiple, estos son los multiprocesadores y multicomputadores.

Soluciones software

Una vez que el hardware obtuvo la potencia suficiente se empezó a explotar el paralelismo a nivel de software. Con la llegada de los multicomputadores, se plantearon nuevos retos, pero esta vez en vez de ser superados por los diseñadores de sistemas operativos, se torno como un problema de las aplicaciones. Quizás esto fue así porque cuando aparecieron las primeras redes no eran algo barato ni tan común, por eso se consideraba tan puntual el sistema operativo Unix al soportarlas. Tampoco se disponía de hardware barato y potente. Por lo tanto aunque el núcleo daba funcionalidades básicas (creación de procesos, comunicación, manejo de red) todo el problema se dejaba al espacio de usuario. Afortunadamente esta aproximación está cambiando y ya existen varios sistemas de ficheros enfocados a tener un sistema de ficheros único en todo el cluster.

Proceso paralelo para SMPs y MPPs

Las arquitecturas del proceso paralelo pueden apoyar: hardware del proceso cluster y masivo paralelo (MPP), en el cual cada nodo tiene su propia memoria. Y también los sistemas de la memoria conocidos como hardware simétrico del multiprocesamiento (SMP), en el cual los procesadores múltiples utilizan un recurso de la memoria

Las máquinas cluster y de MPP tienen memorias múltiples, con cada CPU típicamente teniendo su propia memoria. Tales sistemas prometen ventajas significativas del precio del funcionamiento usando componentes de la memoria y del bus de la materia para eliminar embotellamientos de la memoria.

Sistemas de gerencia de base de datos que apoyan solamente un tipo de límite del hardware la portabilidad de usos, el potencial de emigrar los usos a los nuevos sistemas del hardware, y el escalabilidad de usos. El servidor paralelo de Oracle (OPS) explota clusters y sistemas de MPP, y no tiene ninguna limitación. Oracle sin la opción paralela del servidor explota las solas máquinas de la CPU o de SMP.

Proceso paralelo para las operaciones integradas

El software paralelo de la base de datos debe desplegar con eficacia la energía de proceso del sistema de manejar usos diversos: usos en línea del tratamiento transaccional (OLTP), usos del sistema de ayuda de decisión (Depósitos de datos), así como una carga de trabajo mezclada de OLTP y de los Depósito de datos. Los usos de OLTP son caracterizados por las transacciones cortas que tienen uso bajo de la CPU y de la entrada-salida. Los usos de los Depósitos de datos son caracterizados por transacciones largas, con alto uso de la CPU y de la entrada-salida.

El software paralelo de la base de datos se especializa generalmente a menudo para servir como procesadores QUERY. Puesto que se diseñan para servir una sola función, sin embargo, los servidores especializados no proporcionan una fundación común para las operaciones integradas. Éstos incluyen ayuda de decisión en línea, datos que almacenan, OLTP, operaciones distribuidas, y altos sistemas de la disponibilidad. Los servidores especializados se han utilizado con éxito.

El software paralelo versátil de la base de datos debe ofrecer precio/funcionamiento excelentes en el hardware de los sistemas abiertos, y se diseñe para servir en una variedad amplia de necesidades que contempla la empresa. Las características tales como reserva en línea, réplica de los datos, portabilidad, interoperabilidad, y ayuda para una variedad amplia de herramientas del cliente pueden permitir un servidor paralelo a la integración del uso de ayuda, a las operaciones distribuidas, y a las cargas de trabajo mezcladas.

Arquitectura de hardware paralela

El servidor paralelo de la base de datos puede utilizar las varias arquitecturas de la máquina que permiten el proceso paralelo.

- Sistemas compartidos de la memoria
- Sistemas de disco compartidos
- Sistemas no compartidos
- Sistemas combinados discos no compartidos/compartidos

Perfiles del uso

Los usos para sistema de Depósitos de datos tienden para realizarse en SMPs, clusters, y sistemas masivos paralelos. Para ello se requiere seleccionar las herramientas correctas.

-Hardware y software de sistema operativo requerido: Cada proveedor de hardware pone el proceso en ejecución paralelo a su propia manera.

-Interconexión de alta velocidad: La interconexión puede ser Ethernet, FDDI, u otro método de la interconexión. Si la interconexión primaria falla, una interconexión de reserva está generalmente disponible. La interconexión de reserva asegurará alta disponibilidad, y previene un solo punto de la falta.

- Disco global accesible o subsistema compartido del disco: Todos los nodos en un débilmente acoplado o sistema paralelo tienen acceso masivo simultáneo a los discos compartidos. Estos subsistemas compartidos del disco se ponen en ejecución vía un SCSI compartido lo más a menudo posible SCSI (común en UNIX) conectado con una granja del disco. En plataformas de algún MPP, tales como SP de IBM, los discos se asocian a los nodos y una capa compartida virtual del software del disco permite el acceso global a todos los nodos.

Sistemas compartidos de la memoria

Un sistema compartido de la memoria representado en la figura 2.4.15, tienen las características siguientes:
-Las CPU múltiples comparten memoria.
-Cada CPU tiene de total acceso a toda la memoria compartida a través de un bus común.
-La comunicación entre los nodos ocurre vía memoria compartida.
-El funcionamiento es limitado por el ancho de banda del bus de la memoria.

Las máquinas simétricas del multiprocesador (SMP) son a menudo nodos en un cluster. Los nodos múltiples de SMP se pueden utilizar con el servidor paralelo, donde la memoria se comparte entre las CPU múltiples, y son accesibles por todas las CPU a través de un bus de la memoria. Éstos incluyen en varios componentes del sistema tales como el ancho de banda de la memoria, ancho de banda de la comunicación de la CPU, la memoria disponible en el sistema, ancho de banda de la entrada-salida, y el ancho de banda del bus común.

Las ventajas del proceso paralelo de los sistemas compartidos de la memoria son éstas:

- ✓ El acceso de memoria es más barato que la comunicación del inter-nodo.
- ✓ Esto significa que la sincronización interna es más rápida que usando al encargado de la cerradura.
- ✓ Los sistemas compartidos de la memoria son más fáciles de administrar que un cluster.

Una desventaja de los sistemas compartidos de la memoria para el proceso paralelo es como sigue:

- La escalabilidad es limitada por ancho de banda y estado latente del bus, y por memoria disponible.

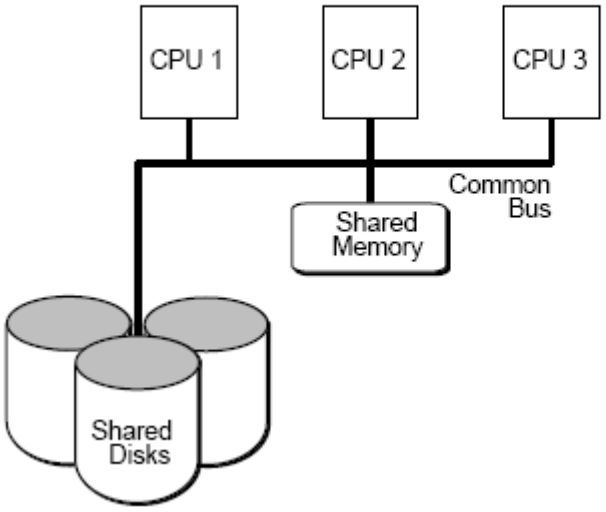

Figura 2.4. 15

Sistema compartido de la memoria

Sistemas de disco compartidos

Los sistemas de disco compartidos, mostrado en la figura 2.4.16, están típicamente débilmente acoplados. Tienen las características siguientes:

Cada nodo consiste en unas o más CPU y memoria asociada.

La memoria no se comparte entre los nodos.

La comunicación ocurre sobre un bus de alta velocidad común.

Cada nodo tiene acceso a los mismos discos y a otros recursos.

Un nodo puede ser un SMP si el hardware lo apoya.

La anchura de banda del bus de alta velocidad limita el número de los nodos (escalabilidad) del sistema.

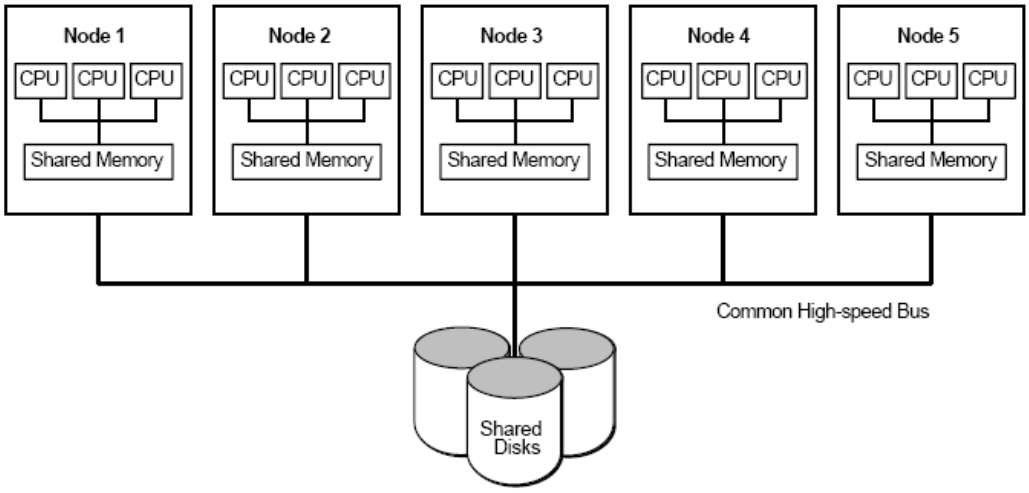

Figura 2.4. 16

Sistema de discos compartidos

Las ventajas del proceso paralelo de los sistemas de disco compartidos son las siguientes:

Los sistemas de disco compartidos permiten alta disponibilidad.

Todos los datos son accesibles en un nodo.

Estos sistemas tienen el concepto de una base de datos, que es un excedente de la ventaja de los sistemas no compartió.

Los sistemas de disco compartidos prevén crecimiento incremental.

Las desventajas del proceso paralelo de los sistemas de disco compartidos son:

Se requiere la sincronización del inter-nodo, implica gastos indirectos de IDLM y mayor dependencia en la interconexión de alta velocidad.

Si la carga de trabajo no se reparte bien, puede haber alta sincronización de arriba.

Hay gastos indirectos del sistema operativo para funcionar software compartido del disco.

Sistemas no compartidos

Los sistemas no compartido, figura 2.4.17, están típicamente débilmente acoplados. En un sistema no compartido los CPU solamente están conectados con un disco dado. Si una tabla o una base de datos están situadas en ese disco, el acceso depende enteramente de la CPU que lo posee. El sistema no compartido se puede representar como sigue:

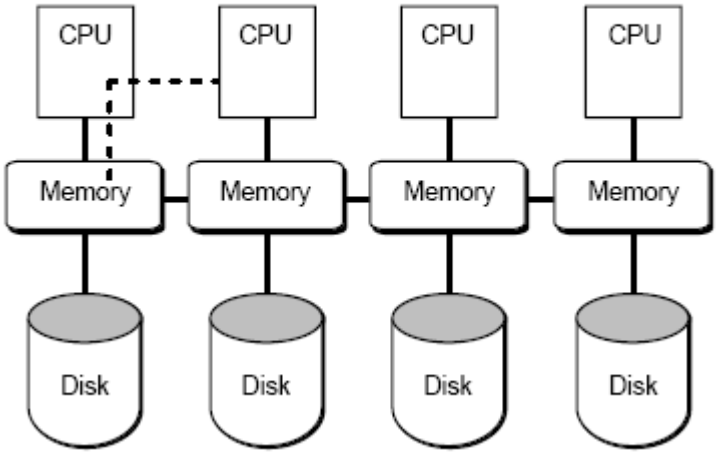

Figura 2.4. 17

Sistemas no compartidos

Los sistemas no compartidos se refieren al acceso a los discos, para no tener acceso a la memoria. El servidor puede tener acceso a los discos en sistemas compartidos mientras el sistema operativo proporcione el acceso de disco transparente, pero este acceso es costoso en términos de estado latente.

Sistemas masivos paralelos

Los sistemas masivo paralelo (MPP) tienen las características siguientes:

Solamente de algunos nodos, hasta millares de nodos se apoyan.

El costo por procesador puede ser extremadamente bajo porque cada nodo es un procesador barato.

Cada nodo ha asociado memoria no-compartida.

Cada nodo puede tener sus propios dispositivos, pero en caso de falta otros nodos pueden tener acceso a los dispositivos del nodo fallado.

Los nodos se organizan en una rejilla, un acoplamiento, o un arreglo del hypercube.

Un sistema masivo paralelo puede tener como varios miles de nodos. Un MPP tiene acceso a una cantidad enorme de memoria verdadera para todas las operaciones de la base de datos (tales como clases o el depósito del almacenador intermediario), puesto que cada nodo tiene su propia memoria asociada. Evitar la entrada-salida del disco, esta ventaja será significativa en preguntas y clases duraderas. Esto no es posible para 32 bits de máquinas que tengan 2 GB de direccionamiento límite; la cantidad total de memoria en un sistema de MPP puede manar esté sobre 2 GB. Como con los sistemas débilmente acoplados, consistencia del depósito en MPPs deben todavía ser mantenidos a través de todos los nodos en el sistema. Así, los gastos indirectos para la gerencia del depósito todavía están presentes.

Base de datos paralela

Una variedad de arquitecturas de hardware permite que las computadoras múltiples compartan el acceso a los datos, al software, o a los dispositivos periféricos. Una base de datos paralela es diseñada para aprovecharse de tales arquitecturas funcionando los casos múltiples que “compartir” un solo base de datos física. En usos apropiados, un servidor paralelo puede permitir el acceso a una sola base de datos de los usuarios en las máquinas múltiples, con funcionamiento creciente.

Un servidor paralelo procesa transacciones en paralelo manteniendo una corriente de transacciones usando las CPU múltiples en diversos nodos, donde cada CPU procesa un entero de transacción. Usando lenguaje de manipulación de datos paralela se puede tener una transacción que es realizada por nodos múltiples. Esto es un acercamiento eficiente porque muchos usos consisten en el relleno en línea y las transacciones de la actualización que tienden para tener datos cortos tienen acceso a requisitos. Además de balancear la carga de trabajo entre las CPU, la base de datos paralela prevé el acceso concurrente a los datos y protege la integridad de datos.

La tecnología paralela de la base de datos puede beneficiar ciertas clases de usos permitiendo:

- Un rendimiento más alto.

- Una disponibilidad más alta
- Mayor flexibilidad
- Más usuarios

Un rendimiento más alto

Con más CPU disponibles para un uso, un speedup y un scaleup más altos pueden ser logrados. La mejora en funcionamiento depende del grado de actividades de la fijación y de la sincronización del inter-nodo. Cada operación de la cerradura es procesado intensivamente; puede haber muchos de estado latente. El volumen de las operaciones de la cerradura y de la contención de la base de datos, así como el rendimiento de procesamiento y el funcionamiento del IDLM, determina en última instancia la escalabilidad del sistema.

Una disponibilidad más alta

Los nodos se aíslan de uno a uno, así que una falta en un nodo no trae el sistema entero abajo. Los nodos restantes pueden recuperar el nodo fallado y continuar proporcionando datos y acceso a los usuarios. Esto significa que los datos están mucho más disponibles que estarían con un solo nodo sobre todo si falta el nodo, y las cantidades a una disponibilidad perceptiblemente más alta de la base de datos.

Mayor flexibilidad

Un ambiente paralelo del servidor debe ser extremadamente flexible. Los casos se pueden asignar o desasignar cuanto sea necesario. Cuando hay alta demanda para la base de datos, más casos pueden ser asignados temporalmente. Los casos se pueden desasignar y utilizar para otros propósitos una vez que ya no sean necesarios.

Más usuarios

La tecnología paralela de la base de datos puede permitir superar límites de la memoria, permitiendo a un solo sistema servir a millares de usuarios.

Para el funcionamiento óptimo, se debe configurar el sistema según los requisitos particulares del uso y recursos disponibles, entonces se debe diseñar e implementar la base de datos y usos para hacer el mejor uso de la configuración. Considerar también la migración del hardware o del software existente al nuevo sistema o a los sistemas futuros. La puesta en práctica acertada del proceso

paralelo y de la base de datos paralela requiere escalabilidad óptima en cuatro niveles:

- Escalabilidad del hardware
- Escalabilidad del sistema operativo
- Escalabilidad del sistema de gerencia de base de datos
- Escalabilidad del uso

Escalabilidad del hardware:

Cada sistema debe tener algunos medios de conectar las CPU, si es un bus de alta velocidad o una conexión de poca velocidad de Ethernet. El ancho de banda y el estado latente de la interconexión determinan la escalabilidad del hardware.

Escalabilidad del sistema operativo:

El escalabilidad del hadware también depende de la escalabilidad del sistema operativo. Esta sección explica cómo analizar este factor. La escalabilidad del software puede ser una edición importante si un nodo es un sistema compartido de la memoria (es decir, un sistema en el cual las CPU múltiples conectan con una sola memoria del multiprocesador simétrico). Los métodos de sincronización en el sistema operativo pueden determinar la escalabilidad del sistema. El multiprocesador asimétrico, por ejemplo, solamente usa una sola CPU puede manejar interrupciones de la entrada-salida. Considerar un sistema en el cual el usuario múltiple procese toda la necesidad de solicitar un recurso del sistema operativo, en la figura 2.4.18 se muestra una comparación de los multiprocesos Asimétricos y Simétricos:

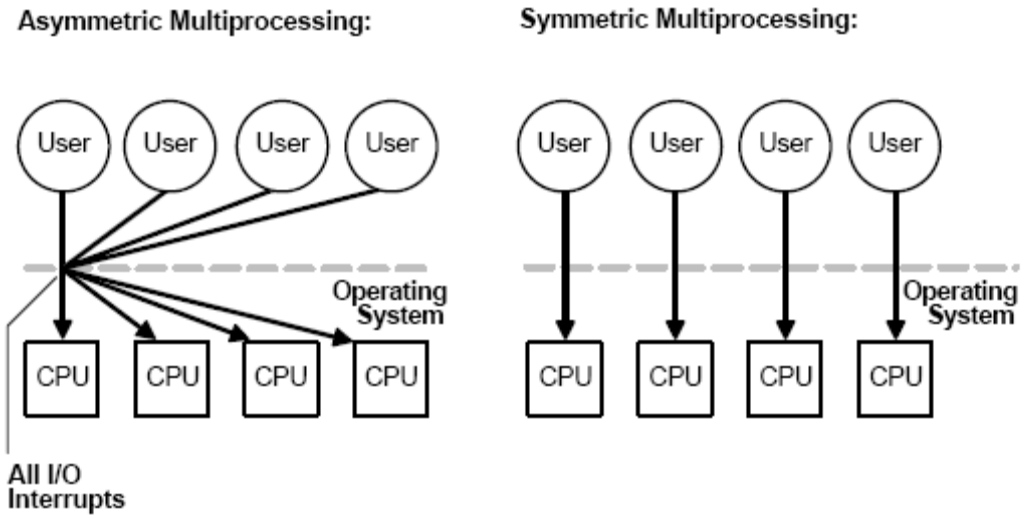

Figura 2.4. 18

Multiproceso asimétrico contra multiproceso simétrico

Aquí, la escalabilidad potencial del hardware se pierde porque el sistema operativo puede procesar solamente una petición del recurso a la vez. Cada vez que una petición incorpora el sistema operativo, una operación se sostiene para excluir las otras. En multiprocesamiento simétrico, por el contrario, no hay tal embotellamiento.

Escalabilidad del sistema de gerencia de base de datos:

Una distinción importante en arquitecturas paralelas del servidor es la comparación entre paralelismo interno contra paralelismo externo; esto tiene un efecto fuerte en la escalabilidad. La diferencia dominante es si el sistema de gerencia orientado a objetos de base de datos (ORDBMS) hace paralelismo la pregunta, o un proceso externo hace paralelismo la pregunta.

La afinidad del disco puede mejorar el funcionamiento asegurándose que los nodos tienen acceso principalmente al local.

Un mecanismo eficiente de la sincronización permite un speedup y un scaleup mejores.

Escalabilidad del uso:

El diseño del uso es dominante al aprovecharse de la escalabilidad de los otros elementos del sistema. Los usos se deben diseñar específicamente para ser

escalables. No importa cómo es escalable el hardware, el software, y la base de datos pueden ser, una tabla con solamente una fila que cada nodo esté al día sincronizada en un datablock. Considerar el proceso de generar un número de serie único:

```
UPDATE ORDER_NUM
SET NEXT_ORDER_NUM = NEXT_ORDER_NUM + 1;
COMMIT;
```

Cada nodo que necesita poner al día este número de serie tendrá que esperar para tener acceso a la misma fila de esta tabla: la situación es intrínsecamente no escalable. Un acercamiento mejor sería utilizar secuencias para mejorar escalabilidad:

```
INSERT INTO ORDERS VALUES
(order_sequence.nextval, ... )
```

Los clientes deben ser conectados con las máquinas del servidor de una manera escalable, esto significa que la red debe también ser escalable.

Estrategias en la implementación

Deben plantearse lo siguiente:

1° Definir el mejor diseño físico para el modelo de datos. El diseño físico debe estar orientado a generar buen rendimiento en el procesamiento de consultas, a diferencia del modelo lógico que está orientado al usuario y a la facilidad de consulta.

2° Definir los procesos de extracción, filtro, transformación de información y carga de datos que se deben implementar para poblar ese modelo de datos.

3° Definir los procesos de administración de la información que permanece en el Depósito de datos.

4° Definir las formas de consultas a la información del Depósito de datos que se le proporcionará al usuario. Para esto, debe considerarse la necesidad de resolver un problema y la potencia de consulta.

5° Completar el modelo de consulta base, relativo al área seleccionada.

6° Implementar los procesos estratégicos del área de trabajo, es decir, implementar herramientas especializadas de scoring, herramientas especializadas para inducción de conocimiento (Data Mining), etc.

7° Completar las áreas de interés, en forma similar a lo descrito anteriormente.

Además de:

- Estudio completo del problema.

En el cual se evalúe las maneras de atacar el problema, se dividan las distintas secciones de éste y se tenga en cuenta las implicaciones. Se corresponde con la parte de análisis del programa

- Estudio de la paralelización del problema.

En la cual se busca la ortogonalización del problema para que este pueda ser fácilmente paralelizable. Se evalúan todas las posibles vías de paralelización y se eligen las que más se adecuen al sistema concreto sobre el que se vaya a ejecutar.

- Estudio del sistema.

Tanto a nivel práctico como teórico. Cada sistema debe ser explotado de una manera diferente.

- Estudio de la paralelización del código.

Este apartado depende completamente de los dos anteriores, es decir, sólo se puede hacer correctamente en el caso que se hayan cumplido los dos apartados anteriores, corresponde con la fase de diseño del programa. Es en esta parte del desarrollo en la que muchas veces se deben cambiar algoritmos adecuados por otros que lo son menos para que sean más fácilmente paralelizables en un sistema concreto y se obtenga mejor rendimiento del mismo.

- Pruebas del sistema.

Muchos programas desarrollados mediante el paradigma de programación paralela hacen uso de un modelo de programación poco ortodoxa, que los hace difíciles de probar y de depurar al mismo tiempo. Otra consecuencia de la programación poco ortodoxa es que en muchas ocasiones el programa no se comporta de la manera o con la eficiencia con la que habíamos supuesto contaría en un principio.

2.4.6 El Depósito de datos virtual

Antes de desarrollar un Depósito de datos, es crítico el desarrollo de una estrategia equilibrada que sea apropiada para sus necesidades y sus usuarios.

Existe un número de estrategias mediante las cuales las organizaciones pueden conseguir sus Depósitos de datos.

Parte de esta estrategia es establecida un ambiente Depósito de datos Virtual, el cual puede ser creado por:

1.- Instalación de un conjunto de facilidades para el acceso a datos, directorios de datos y gestión de proceso.
2.- Entrenamiento de usuarios finales.
3.- Basados en el uso actual, crear un Depósito de datos físico para soportar los pedidos de la alta frecuencia.

Una estrategia de Depósito de datos virtual o “Point to Point”, significa que los usuarios finales pueden acceder a bases de datos operacionales directamente, usando cualquier herramienta que posibilite “la red de acceso de datos”.

Este enfoque provee flexibilidad así como también la cantidad mínima de datos redundantes que deben cargarse y mantenerse. Además, se pueden colocar las cargas de consulta no planificadas más grandes, sobre sistemas operacionales.

El almacenamiento virtual es, frecuentemente, una estrategia inicial, en organizaciones donde hay una amplia necesidad de conseguir los datos operacionales, desde una clase relativamente grande de usuarios finales y donde la frecuencia probable de pedidos es baja.

Los depósitos virtuales de datos proveen un punto de partida para las organizaciones determinadas que usuarios finales están buscando realmente.

En los Depósitos de datos virtuales el modelo describe, la ausencia del control autoritario, centralizado sobre las fuentes de datos físicas subyacentes es en desacuerdo con la meta de la interoperabilidad lógica.

2.4.7 Minería de datos

La Minería de Datos es el proceso a través del cual se descubre conocimiento interesante en forma de patrones, asociaciones, cambios, anomalías y estructuras significantes de grandes cantidades de datos almacenados en Bases de Datos, Bodegas de Datos u otros Repositorios de información. Es la extracción de información oculta y predecible de grandes bases de datos.

Las técnicas de la Minería de Datos son el resultado de un largo proceso de investigación y desarrollo de productos orientados al almacenamiento, extracción análisis de datos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. La Minería de Datos está soportada por las siguientes tecnologías:

- Soportes de almacenamiento masivo de datos
- Potentes computadoras con multiprocesadores
- Depósitos de datos
- Algoritmos de Minería de Datos.

El sistema Minería de Datos es una tecnología de soporte para usuario final cuyo objetivo es extraer conocimiento útil y utilizable a partir de la información contenida en las bases de datos de las empresas.

Las herramientas de Minería de Datos sirven para predecir tendencias y comportamientos, de esta manera permiten a las organizaciones tomar decisiones proactivas para adaptarse rápidamente a los cambios del mercado obteniendo así ventajas herramientas.

Una vez que las herramientas de Minería de Datos fueron implementadas en computadoras cliente servidor de alto performance o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a preguntas y presentar los resultados en formas de tablas, con gráficos, reportes, texto, hipertexto, etc. El origen de la información que utilizan los algoritmos de la Minería de Datos, por lo general, son datos históricos que se encuentran almacenados en un Depósito de datos. El partir de un Depósito de datos simplifica la etapa previa a la etapa de reparación de los datos ya que se construye en base a la integración de fuentes de datos múltiples y heterogéneas Bases de Datos relacionales, ficheros planos y registros de transacciones en línea.

La Minería de Datos no puede ser experimental. En muchas circunstancias, no es posible reproducir las condiciones que generaron los datos (especialmente si son datos del pasado, y una variable es el tiempo).

El Depósito de datos dota a las organizaciones de memoria, y el Depósito de datos de inteligencia. La mejor forma de aplicar las técnicas de la Minería de Datos es que éstas se encuentren totalmente integradas con el Depósito de datos así como también con herramientas flexibles e interactivas para el análisis de negocios. Varias herramientas de la Minería de Datos actualmente operan fuera del Depósito de datos, requiriendo pasos extra para extraer, importar y analizar los datos. Además la integración con el Depósito de datos permite que ni bien los cambios originados en las bases de datos operacionales son replicados al

Depósito de datos pueden ser analizados directamente y monitoreados mediante las técnicas de Minería de Datos.

La Minería de datos es el proceso de descubrir patrones de información interesante y potencialmente útiles, inmersos en una gran base de datos en la que se interactúa constantemente. Existen investigadores que ven la Minería de Datos como un paso fundamental en el proceso de descubrimiento del conocimiento el cual consiste de una secuencia iterativa de los siguientes pasos:

Limpieza De Datos: En este paso se manejan los problemas de los datos, erróneos, perdidos o irrelevantes.

Integración de datos: Se busca integrar en un único repositorio los datos provenientes de diferentes y múltiples fuentes.

Selección De Datos: Los datos relevantes la tarea del análisis son recuperados de la base de datos.

Transformación De Datos: Los datos son transformados o consolidados en formas apropiadas para la minería a través de la ejecución de operaciones de resúmenes o agregaciones.

Minería De Datos: Es un proceso esencial donde métodos inteligentes son aplicados para extraer patrones de datos.

Evaluación De Patrones: El cual permite identificar con precisión el conocimiento representado en patrones (comportamientos) basado en algunas medidas de interés.

El servidor de la Minería de Datos debe estar integrado con el Depósito de datos y el servidor OLAP para insertar el análisis de negocios directamente en esta infraestructura. Un avanzado, metadato centrado en procesos define los objetivos de la Minería de Datos para resultados específicos tales como manejos de campañas promociónales, optimización de promociones, etc. A medida que el Depósito de datos crece con nuevas decisiones y resultados, la organización puede aplicar la Minería de Datos para decisiones.

Este diseño representa una transferencia fundamental desde los sistemas de soporte de decisión convencionales. Más que simplemente proveer datos a los usuarios finales a través de software de consultas y reportes, el servidor de la Minería de Datos aplica los modelos de negocios del usuario directamente al Depósito de datos y devuelve un análisis proactivo de la información más relevante.

Estos resultados mejoran los metadatos en el Servidor OLAP proveyendo un estrato de metadatos que representa una vista fraccionada de los datos.

Generadores de reportes, visualizadores y otras herramientas de análisis pueden ser aplicadas para planificar futuras acciones y confirmar el impacto de esos planes.

KDD

KDD se define como “la extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos”. Hay una distinción clara entre el proceso de extracción de datos y el descubrimiento del conocimiento. Bajo sus convenciones, el proceso de descubrimiento del conocimiento toma los resultados tal como vienen de los datos (proceso de extraer tendencias o modelos de los datos) cuidadosamente y con precisión los transforma en información útil y entendible. KDD puede usarse como un medio de recuperación de información, de la misma manera que los agentes inteligentes realizan la recuperación de información en el Web. Nuevos modelos o tendencias en los datos podrán descubrirse usando estas técnicas. KDD también puede usarse como una base para las interfaces inteligentes del mañana, agregando un componente del descubrimiento del conocimiento a una máquina de bases de datos o integrando KDD con las hojas de cálculo y visualizaciones.

Al Descubrimiento de Conocimiento de Bases de Datos (KDD) a veces también se le conoce como minería de datos.

Muchos autores se refieren al proceso de minería de datos como el de la aplicación de un algoritmo para extraer patrones de datos y a KDD al proceso completo (pre-procesamiento, minería, post-procesamiento).

El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos.

Las metas de KDD son:

- procesar automáticamente grandes cantidades de datos crudos,
- identificar los patrones más significativos y relevantes, y
- presentarlos como conocimiento apropiado para satisfacer las metas del usuario.

KDD típicamente combina métodos automatizados con la interacción humana para asegurar resultados exactos, útiles, y entendibles.

Algoritmos de Minería de Datos

Los algoritmos de Minería de datos se clasifican en dos grandes categorías:

- supervisados o de predicción y
- no supervisados o de descubrimiento del conocimiento

Los algoritmos supervisados o de predicción predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases:

Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución de predicción, en ese caso hay que recurrir a los métodos no supervisados o del descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas.

Las herramientas de Minería de datos exploran gran cantidad de datos dentro de una BD grande, y mediante su análisis predicen posibles tendencias o comportamientos futuros dentro de una empresa, permitiendo al experto tomar decisiones en los negocios de una forma rápida y utilizando un conocimiento que de otra forma no habría encontrado. Mediante la utilización de estas herramientas se pueden generar nuevas oportunidades de negocio. Algunas posibilidades que ofrecen estas herramientas son:

- Predicción automatizada de tendencias y comportamientos.
- Descubrimiento automatizado de modelos desconocidos.
- Descubrimiento de anomalías y acciones fraudulentas por parte de clientes.

Este producto esta fuertemente relacionado con análisis estadísticos, el objetivo de generar hipótesis potenciales de interés que son posteriormente verificadas.

Estructura de Minería de datos

La estructura de minería de datos es una estructura que define el dominio de datos a partir del cual se generan los modelos de minería de datos. Una única estructura de minería de datos puede contener varios modelos de minería de datos que comparten el mismo dominio.

La unidad de creación de estructura de minería de datos son las columnas de la estructura de la minería de datos, que describen los datos que contienen el origen de datos. Estas columnas contienen información como el tipo de datos, el tipo de contenido y el modo en que se distribuyen los datos.

La estructura también puede contener tablas anidadas, esta representa una relación entre uno a varios entre la entidad de un escenario y sus atributos relacionados.

Técnicas de Minería de Datos

La minería de datos y los agentes inteligentes extraen significados y nuevos conocimientos de vastas cantidades de información, se basan en algoritmos, emparejamiento de patrones, patrones heurísticos de reconocimiento de reglas, redes neuronales e inteligencia artificial

Se utilizan las técnicas**:**

Análisis Preliminar de datos usando Query tools: el primer paso en un proyecto de Minería de datos sería siempre un análisis de los datos usando query tools, aplicando una consulta SQL a un conjunto de datos, para rescatar algunos aspectos visibles antes de aplicar las técnicas. La gran mayoría de la información (un 80 %) puede obtenerse con SQL. El 20 % restante, más importante, la información oculta requiere técnicas avanzadas.

Este primer análisis en SQL es para saber cual es la distribución de los valores posibles de los atributos. Recién después podemos ver la performance del algoritmo correspondiente.

Técnicas de Visualización: estas son buenas para ubicar patrones en un conjunto de datos y puede ser usado al comienzo de un proceso de data mining para tomar un feeling de la calidad del conjunto de datos.
Árbol de Decisión: son estructuras en forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Para poder predecir el comportamiento de un cliente es necesario poder contar con una clasificación previa esto implica una predicción de que un cliente pertenece a cierto grupo de clientes. La complejidad es de n (Log n).

Métodos específicos de árboles de decisión incluyen:
- CART Árboles de clasificación y regresión: técnica usada para la clasificación de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un

cierto resultado. Segmenta un conjunto de datos creando 2 divisiones. Requiere menos preparación de datos que CHAID.
- CHAID Detección de interacción automática de Chi cuadrado: técnica similar a la anterior, pero segmenta un conjunto de datos utilizando tests de chi cuadrado para crear múltiples divisiones.

Reglas de Asociación: establece asociaciones en base a los perfiles de los clientes sobre los cuales se está realizando el data mining. Las reglas de Asociación están siempre definidas sobre atributos binarios. No es muy complicado generar reglas en grandes bases de datos. El problema es que tal algoritmo eventualmente puede dar información que no es relevante. Data Mining envuelve modelos para determinar patterns a partir de los datos observados. Los modelos juegan un rol de conocimiento inferido. Diciendo cuando el conocimiento representa conocimiento útil o no, esto es parte del proceso de extracción de conocimiento en bases de datos (Knowledge Discovery in Databases-KDD).
Algoritmos Genéticos: son técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en los conceptos de evolución.

Redes Bayesianas: buscan determinar relaciones causales que expliquen un fenómeno en base a los datos contenidos en una base de datos. Se han usado principalmente para realizar predicción.

Procesamiento Analítico en Línea (OLAP): estas herramientas ofrecen un mayor poder para revisar, graficar y visualizar información multidimensional, en características temporales, espaciales o propias. Se valen de lenguajes menos restringidos y estructurados como lo es SQL. Requieren todavía de una alta participación de un usuario humano, pues son interactivas y requieren la guía del experto.

Redes neuronales artificiales: son modelos predecibles, no lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.

Método del vecino más cercano: una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases de k registro/s más similar/es a él en un conjunto de datos históricos. Algunas veces se llama la técnica del vecino k-más cercano.

Regla de inducción: la extracción de reglas if-then de datos basados en significado estadístico.

Modelado de la Minería de datos

La técnica usada para realizar estas hazañas en la Minería de datos se llama Modelado.

El Modelado es simplemente el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta.

Lo que ocurre en las computadoras, no es muy diferente de la manera en que la gente construye modelos. Las computadoras son cargadas con mucha información acerca de una variedad de situaciones donde una respuesta es conocida y luego el software de la Minería de datos en la computadora debe correr a través de los datos y distinguir las características de los datos que llevarán al modelo. Una vez que el modelo se construyó, puede ser usado en situaciones similares donde no se conoce la respuesta.

La minería de datos, es un proceso que invierte la dinámica del método científico, dado que se generan hipótesis a partir de los datos colectados.

Las técnicas de Minería de datos combinan la tecnología de bases de datos y “Depósitos de datos”, con técnicas de aprendizaje automático y de estadística.

La estadística es una herramienta poderosa, y es un elemento crucial en el análisis de datos. Sin embargo, a veces enfrentamos problemas muy serios en la interpretación de sus resultados, dado que no recordamos que estos resultados se aplican a grupos y no a individuos. Estos peligros se ven amplificados en el uso de software de Minería de Datos.

La Minería de Datos es una herramienta de exploración y no explicativa. Es decir, explora los datos para sugerir hipótesis. Es incorrecto aceptar dichas hipótesis como explicaciones o relaciones causa-efecto. Es necesario coleccionar nuevos datos y validar las hipótesis generadas ante los nuevos datos, y después descartar aquellas que no son confirmadas por los nuevos datos.

Las Bases de Datos proporcionan la infraestructura necesaria para almacenar, recuperar y manipular datos. La construcción y mantenimiento de un Depósito de datos, a pesar de que esta es una Base de Datos, su modo de operar es muy distinto, para soportar transacciones y la actividad de negocio en línea, además hace viable la revisión y el análisis de su información para el apoyo a las decisiones ejecutivas. Típicamente, el Depósito de datos almacena y resume información sobre transacciones cotidianas a lo largo del tiempo. Puede que contenga información que ya no es posible reproducir del sistema para la operación cotidiana, es información arcaica pero útil por su crónica histórica del funcionar. Las consultas a los depósitos no son tan sistemáticas como las transacciones y usualmente demandan más recursos de cómputo. Resulta incluso conveniente separar los equipos y sistemas de la operación cotidiana de transacciones en línea de los Depósitos de datos.

Principales tareas de la Minería de Datos

Las tareas de la minería de datos pueden ser clasificadas en: Minería de Datos Descriptiva y de Predicción. Pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas por consultas en un sistema tradicional de soporte operacional. La potencialidad de estas herramientas reside en la capacidad de explorar las bases de datos en busca de patrones ocultos, encontrando información predecible que para un experto sería casi imposible debido al gran volumen de información.

Un sistema de minería de datos puede efectuar una o más de las siguientes tareas de minería de datos:

Descripción (Caracterización) de Clases: Presenta un resumen conciso de una colección de datos y sus diferencias. Las comparaciones entre dos o más colecciones de datos son llamadas también Discriminaciones de clases.

Asociación: Es el descubrimiento de relaciones de asociación o correlación entre un conjunto de ítems. Frecuentemente son expresadas en forma de reglas (atributo - valor).

Clasificación: La clasificación analiza un conjunto de datos preparados y construye un modelo para cada clase basado en las características de los datos.

Predicción: Permite predecir los posibles valores de algunos datos perdidos o el valor de la distribución de ciertos atributos en un conjunto de objetos. Implica encontrar un conjunto de atributos relevantes a los atributos de interés y predecir el valor de la distribución basado en un conjunto de datos similares a los objetos seleccionados.

Clustering: El análisis de clustering es para identificar clusters que involucran los datos. Un cluster es una colección de objetos de datos que tienen similaridades.

Análisis De Series De Tiempo: El análisis de series de tiempo permite analizar grandes conjuntos de series de datos de tiempo para encontrar ciertas regularidades y características interesantes, incluyendo búsqueda de secuencias similares o subsecuentes, patrones secuenciales, periodicidad, tendencias y desviaciones.

Etapas principales del proceso de la minería de datos

1. Determinación de los objetivos: delimitar los objetivos que el cliente desea bajo la orientación del especialista en la minería de datos.

2. Pre-procesamiento de los datos: se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de la minería de datos.

3. Determinación del modelo: se comienza realizando un análisis estadístico de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.

4. Análisis de los resultados: verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por el análisis estadístico y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

Respecto a los modelos inteligentes, se ha comprobado que en ellos se utilizan principalmente árboles y reglas de decisión, reglas de asociación, redes neuronales, redes Bayesianas, conjuntos aproximados (rough sets), algoritmos de agrupación (clustering), máquinas de soporte vectorial, algoritmos genéticos y lógica difusa.

A pesar que la minería de datos es muy reciente, involucra un trabajo interdisciplinario en el cual intervienen áreas como: sistemas de bases de datos, bodegas de datos, estadística, máquinas de aprendizaje, visualización de datos, recuperación de información, redes neuronales, reconocimiento de patrones, análisis de datos espaciales, bases de datos de imágenes, procesamiento de señales, teoría de grafos, programación lógica inductiva y computación de alto rendimiento.

Existen una gran diferencia entre el análisis de datos tradicional y la minería de datos. En el análisis tradicional el análisis de los datos es conducido por la “suposición” (assumption-driven), es decir, la hipótesis se forma y se valida contra los datos, en cambio en la minería de datos el análisis se conduce por el “descubrimiento” (discovery-driven), en el sentido que los patrones son extractados automáticamente de los datos (grandes búsquedas).

Se puede afirmar que con la maduración de la tecnología de BD se logra almacenar, administrar y procesar grandes volúmenes de datos en diferentes repositorios que van desde las Bases de Datos tradicionales hasta las más modernas Bodegas de Datos modeladas de forma multidimensional. Este avance generó lo que hoy se conoce como “Data Explosion Problem” o Problema de Explosión de Datos, que consiste en enormes océanos de datos sin valor agregado. En la actualidad son muchas las instituciones (gubernamentales,

de servicios, sector productivo, etc.) que se están ahogando sin encontrar islas de información donde pueda residir algún conocimiento.

2.5 Administración de los depósitos de datos

2.5.1 Administración de aplicaciones

La administración de un depósito de datos es una tarea compleja que comprende la actualización de los datos del Depósito de datos, la sincronización de las fuentes de datos, la planeación de recuperación de desastres, la administración del control de acceso y seguridad, la administración del crecimiento de los datos y la administración del desempeño de base de datos.

Capa de administración de datos

Las tareas de extraer, cargar, actualizar reforzar la seguridad, archivar y restaurar el depósito de datos a partir de archivos son sustentados por la capa de administración de datos.

Desde la perspectiva del depósito de datos, los elementos de interés particular en la capa de administración de datos son los siguientes:

- Extraer los datos apropiados a fin de seleccionar las fuentes de información para el refinamiento posterior, la reingeniería y la incorporación en el depósito de datos.

- Seguir y llenar las solicitudes de datos nuevos por parte de fuentes de datos nuevas o actuales.

- Implantar un procedimiento donde se validen los datos de las fuentes de datos operacionales una vez que estos sean fidedignos, actualizar el depósito de datos.

En esta capa de administración de datos se incorporan políticas estándar como son procedimientos, programas, operaciones de seguridad, autorizaciones de acceso, archivar, restaurar y purgar los datos. Un aspecto importante es cuando un depósito de datos es potencialmente grande. El tamaño del depósito de datos afecta la administración de la compactación de datos y de índices múltiples. Esta capa también administra aspectos de procesamiento paralelo de consultas y el uso de procesadores paralelos para el acceso y recuperación de datos.

El gran reto de administrar bases de datos muy grandes con su complejidad en las áreas de índices múltiples, agrupación de datos, claves compuestas y las versiones de los datos se solventa y administra en esta capa así como también los modelos, esquemas de datos Lógicos, físicos del mercado de datos junto con el glosario técnico y empresarial.

La capa de administración de meta datos es responsable de controlar lo siguiente:

- Las definiciones estándar de los datos tanto técnicas y empresariales del depósito de datos.
- Los meta datos capturados y creados en los bloques de refinamiento y reingeniería
- Los meta datos acerca de granularidad, segmentación, áreas tema, adición y condensación.
- Los meta datos que describen las consultas y reportes predefinidos y diseñados.
- Los meta datos describen índices y perfiles que mejoran el rendimiento en el acceso y recuperación de datos.
- Los meta datos describen reglas para preparar y programar el ciclo de actualización y duplicación.

El componente de administración de sistemas ofrece la capacidad de que el constructor de sistemas y el usuario empresarial invoquen, manejen y determinen las herramientas de las aplicaciones.

El componente de administración del flujo de trabajo sustenta el proceso de integración y administración para coordinar de manera ordenada y coordinada la ejecución de herramientas, aplicaciones y actividades para extraer, resumir, actualizar, agregar y duplicar los datos en el almacén de datos.

La administración del flujo de trabajo automatiza muchas tareas requeridas para el mantenimiento y actualización del depósito de datos. Una adecuada administración permite tener reportes predefinidos y resultados de consultas que incrementan la eficiencia y la productividad de los constructores del sistema y los usuarios empresariales.

La administración de un depósito de datos inicia en el momento en que se despliega este. Se exponen los siguientes aspectos de la administración de un depósito de datos.

- Actualización y duplicación de datos.
- Recuperación de desastres.

- Controles de acceso y seguridad.
- Administración del crecimiento de datos.
- Administración del desempeño de base de datos.
- Mejora y ampliación del depósito de datos.

Actualización y duplicación de datos.

Los servicios de duplicación también permiten modificar la administración de los datos copiados y sólo permiten que los cambios se propaguen de la fuente hacia el objetivo. La duplicación es una forma muy útil de liberar datos de un sistema a otro en forma continua, al tiempo que los cambios se mantienen bajo control. Por lo regular, el servidor de duplicación es una tercera maquina que contiene una base de datos y administra el proceso de duplicación entre la fuente y los objetivos.

La duplicación es un método conveniente para mantener el almacén de datos sincronizado con las fuentes. Es también un método conveniente para mantener los mercados de datos sincronizados con los almacenes de datos, el mismo servidor de duplicaciones distribuye datos duplicados a más de un objetivo. Por lo tanto, el servidor de aplicaciones funciona como un canal de distribución de datos duplicados hacia múltiples mercados de datos.

Para tener una adecuada administración de los procesos de duplicación se debe tener un eficiente calendario de duplicación de datos. Crear una agenda es una forma importante de manejar conflictos de recursos. El tener un calendario de la duplicación de los datos se logra minimizar la necesidad incrementar el ancho de banda de la red, se programa esta tarea en horas no pico para efectuar las transferencias de duplicación, transferir datos a través de varias zonas de horarios y el uso de activadores por tiempo así como el uso de eventos para controlar los procesos de duplicación.

Recuperación de desastres

Después de que el depósito de datos es funcional, uno de los retos administrativos consiste en formular un plan de recuperación de desastres. Una vez que el depósito de datos empieza a funcionar se incrementan las presiones de la administración y de los usuarios para tener acceso continuo. Si no se tiene definido un adecuado plan de recuperación de desastres, las repercusiones para el personal de despliegue son significativas.

Ensayar una metodología para la recuperación de desastres debe formar parte de las actividades de desarrollo previas al funcionamiento del depósito de datos.

Administración del crecimiento de los datos.

El depósito de datos tiene el potencial de almacenar grandes volúmenes de información, esto se debe a que el depósito de datos maneja información histórica, además de los datos operacionales actuales. Al transcurrir el tiempo, los datos actuales de hoy se convierten en los datos históricos de mañana y se deben almacenar y manejar exactamente como los datos históricos de hace un año por ejemplo.

Existen varias razones por las que un depósito de datos tiende almacenar grandes cantidades de información dentro de las más comunes podemos mencionar las siguientes:

- El depósito de datos contiene datos históricos. Un depósito de datos contiene típicamente de cinco a diez años de datos en su estado final de madurez.

- Los datos se almacenan a nivel detalle en el depósito de datos. Este nivel de detalle se requiere para conciliar varios datos de distintos departamentos cada vez que hay conflictos en la interpretación de los datos.

- Poca o nula planeación de los resultados de las consultas hechas al depósito de datos. Los usuarios finales “crean” cantidades adicionales de datos que desean conservar (futuras consultas, respaldo de su toma de decisiones)

Estas son las principales razones por la que un depósito de datos tiende a creer de forma desmesurada, por lo que es importante tener una política de administración para contener y manejar el crecimiento de los datos, dentro de las técnicas de administración que deben tenerse en cuenta lo siguiente:

- Utilizar una adecuada técnica para resumir los datos que se cargan al depósito de datos de forma optima. Al pasar de una información muy detallada a una muy resumida, disminuye significativamente la magnitud de del almacenamiento requerido de información. Sin embargo es importante utilizar un técnica optima para resumir dicha información ya que esta no debe perder la capacidad profundizar al detalle. Todos los datos deben almacenarse y estar disponibles sin importar el grado de resumen.

El administrador del sistema deberá evaluar que tan necesario es que los usuarios tengan el nivel mas bajo de granularidad en la información, por lo general los usuarios finales pueden manejar sus tareas con menos datos de detalle que los que demandan.

- Limitar el almacenamiento de datos históricos en el depósito de datos. Las características empresariales cambian de modo significativo a través de los años; estas características podrían ser cíclicas y repetirse así mismas en intervalos de tiempo. Limitar el almacenamiento de información histórica al último ciclo empresarial pudiera ser más productivo que derivar análisis sobre datos muy antiguos los cuales pudieran tener un escaso valor.

- Limitar el ámbito de los datos que deben manejarse identificando eventos empresariales que han alternado las circunstancias en las que obtuvieron los datos cargados al depósito de datos. Un ejemplo es cuando se fusionan dos corporaciones, el beneficio de sus datos históricos individuales pudiera no ser el mismo al combinarlos en un solo depósito de datos.

- Eliminar datos de detalle que no son utilizados por los usuarios finales. Un patrón informático interesante en un depósito de datos es que a través del tiempo los datos van aumentando, el porcentaje real de datos utilizados para el procesamiento de consultas se reduce. Es decir aunque el depósito de datos contenga mas información se utilizan lo mismo o menos datos al realizar consultas de información. Una forma de controlar que estos datos no aumenten demasiado es tarea del administrador el cual debe identificar y eliminar los datos a los que las consultas no hagan referencia en absoluto.

Es importante tener en cuenta estas técnicas para evitar en lo posible sobre cargar al depósito de datos con información superflua que no aporte ningún valor a las consulta realizadas por los usuarios finales.

Sin embargo los almacenes de datos son por naturaleza receptores de un gran volumen de información estos pueden almacenar información desde unos cuantos gigabytes hasta varios terabytes. La inversión en los costos de almacenamiento en el servidor donde se implantara el depósito de datos influirá en el costo global de instrumentar el depósito de datos es por esto que se necesita tener una adecuada administración y control del crecimiento del depósito de datos.

Administración del desempeño del almacén de datos

El depósito de datos es una aplicación exigente del almacén de datos, por lo tanto el desempeño de las consultas lo determina en gran medida la organización física del almacén de datos. Un adecuado diseño del almacén asegura un desempeño

optimo del depósito de datos, es por esto que a continuación se describen algunas actividades las cuales mejoran el desempeño del almacén de datos, es importante tener en cuenta estos puntos para tener bases y poder seleccionar el software adecuado que reúna la mayoría de los puntos mencionados que a continuación se describen:

Ejecución de consultas en paralelo.

Este método consiste en mejorar la forma en que se hace una consulta al Depósito de datos. Una consulta se divide en componentes y todos los componentes que puedan ejecutarse de manera simultánea se ejecutan en paralelo por medio de procesos concurrentes, por lo tanto el desempeño de la consulta es grande y efectivo. El administrador deberá tener en cuenta este tipo de consultas para garantizar un adecuado tiempo de respuesta al usuario final, este tipo de consultas deberá ser una característica que busque el administrador en el software seleccionado para implantar un depósito de datos en su organización.

Segmentación inteligente de tablas.

En este método, el administrador del sistema es responsable de segmentar las tablas con eficiencia en varios discos y de mover los segmentos de datos y tablas no involucrados. Otra ventaja de realizar este tipo de segmentación es que también se tiene la capacidad de ejecutar respaldos y restauraciones a nivel de segmento. Esto permite el respaldo / restauración de partes de una tabla, incrementando así la disponibilidad y confiabilidad de otras partes que no se respaldan o restauran. Esto es muy importante en el ambiente de depósito de datos, en donde el tamaño de las tablas alcanza varios cientos de gigabytes para las tablas que almacenan hechos.

Métodos avanzados para la generación de índices.

Los fabricantes de depósitos de datos están ofreciendo esquemas avanzados de generación de índices, estos esquemas son más eficientes y mejoran mucho el desempeño para aplicaciones relacionadas con el depósito de datos. Uno de los esquemas, denominado tecnología de generación de índices de mapa de bits. Mejora en forma significativa el tiempo de respuesta sobre los métodos tradicionales de generación de índices, al reducir en gran medida la cantidad de operaciones de lecturas a los datos. Además de permitir que más usuarios tengan acceso simultáneo al depósito de datos.

La generación de índices utilizando el método de mapa de bits se diseñó para mejorar en forma significativa el tiempo de respuesta en consultas complejas. El principio fundamental para generar índices utilizando el método de mapa de bits es que casi todas las operaciones sobre registro del almacén de datos se realizan

sobre índices sin reclasificar. Al efectuar operaciones principalmente sobre índices, se reduce en gran medida la cantidad de lecturas al almacén de datos. Los índices de mapa de bits consumen bastante almacenamiento, pero aceleran la ejecución de consultas complejas. Como en todos los sistemas de generación de índices, construir un índice de mapa de bits hace más lenta la actualización de los datos. En un depósito de datos, esta merma aprecia con frecuencia, sólo durante la carga del almacén de datos y durante las operaciones de actualización.

Además de esta técnica existen otras, sin embargo los fabricantes ofrecen una combinación de los diversos métodos de generación de índices. Estas combinaciones se basan en el uso de árboles binarios, mapas de bits y lista de identificadores de registros. Obtener lo mejor de los índices para el almacén de datos requiere utilizar la forma apropiada de generación de índices para el tipo de tabla adecuado. Para obtener un buen desempeño de un depósito de datos se requieren diseñadores experimentados que comprendan lo que hay en el depósito de datos y cómo se usa. Por lo tanto, el diseño físico del depósito de datos es un aspecto clave para un buen desempeño de la aplicación.

Resumiendo, las funciones necesarias de administración para fines de preparación inicial, configuración y operación continua incluyen los siguientes puntos:

- Definición del modelado analítico dimensional. El modelado inicial de datos en donde es muy importante elegir las dimensiones correctas y su granularidad, prever cómo se accederán los datos y seleccionar los filtros apropiados para cargar los datos desde el depósito de datos.

- Creación y mantenimiento del depósito de meta datos.

- Control de acceso y privilegios con base en el uso. En este punto es necesario concentrarse en que desean hacer los usuarios empresariales y quien puede obtener acceso al modelo analítico y sus datos.

- Carga del modelo analítico desde el depósito de datos o el mercado de datos. Transferencias periódicas y actualizaciones en bloque, debido a que las actualizaciones en incrementos son un reto y casi imposibles mientras que el almacén de datos está es uso.

- Adición, resúmenes y precálculo durante el proceso de carga.

- Reorganización de almacén de datos para mejorar el desempeño, cambiar el modelo dimensional o actualizar los datos.

- Distribución de los datos al cliente para análisis adicional y local.

- Administración de todas las partes del sistema, incluyendo el software del depósito de datos.

- Capacitación a los usuarios finales en una tecnología diferente y uso de nuevas habilidades.

2.5.2 Administración de bases de datos en un ambiente heterogéneo

Las Tecnologías de la Información (IT) han cambiado sustancialmente la forma de hacer negocios de las empresas. En un entorno donde la competitividad, la globalización, la consolidación de industrias, ciclos de vida más cortos de los productos así como la saturación de mercados hacen que la información tenga un papel preponderante.

La información referente a mercados, competidores, clientes, incluso la relativa a los indicadores de rendimiento de la propia compañía, se ha convertido en un recurso clave. El problema radica en que las empresas disponen de una gran cantidad de datos, pero muy poca información.

Esto se debe muy comúnmente a lagunas de información, así como carencia de arquitectura, gestión, responsabilidad, posesión de los datos, deficiencia en calidad, contenido, poca accesibilidad, pobre fiabilidad de la información, múltiples y diversas aplicaciones operacionales.

Por lo que gran parte del producto generado por tecnologías de información, no es información, sino solo datos brutos. Ya que estos fueron generados por sistemas que fueron ideados para recogerlos, pero no para analizarlos.

Los datos adquieren la categoría de información cuando disponen de una estructura inteligente. A su vez, esta información se convertirá en conocimiento si se le añade las ideas, intuición, capacidad del analista, es decir, conocimiento tácito. La información sería el conocimiento explícito, es decir algo susceptible de ser transmitido, pero solo la información no será capaz de aumentar y mejorar la base de conocimiento de una compañía.

Es la inclusión del conocimiento tácito, la que promueve el ciclo virtuoso de la transformación de datos en información, información en conocimiento, y finalmente, conocimientos en acciones / decisiones mejor informadas y más afines a la realidad de la compañía.

Es este cúmulo de información que el administrador de la aplicación debe tener en cuenta para llevar un adecuado control de extracción de información de bases de datos operacionales, archivos de texto, hojas de calculo, Internet así como fuentes

de datos adquiridos, tales como Dun & Bradstreet, Standard & Poor's, Moody's y A.C. Nielsen.

El administrador del depósito de datos debe tener un control adecuado de las diferentes y nuevas fuentes de datos dentro de la organización.

Agregar nuevas fuentes de datos que requieran ampliar el modelo del almacén de datos es una tarea compleja. De inicio se debe efectuar el análisis dimensional del modelo de la fuente de datos para determinar si se necesita definir nuevas tablas de hechos o ampliaciones a las ya existentes. Dependiendo de esto se debe analizar si se tiene que definir nuevas tablas de dimensión para los datos. También se debe realizar un análisis para determinar el complemento de los nuevos datos con los existentes en el depósito de datos. El análisis producirá los requerimientos de refinamiento y reingeniería para ajustar los nuevos datos dentro del depósito.

Fuentes de datos adquiridos nuevas o adicionales.

Las fuentes de datos adquiridos son fuentes comerciales de información que se compran. Un ejemplo clásico es una base de datos de una compañía de encuestas de consumo principal que describe los perfiles de clientes prospectos, sus hábitos, patrones de compra y perfiles de ingresos. Las fuentes de datos adquiridos proporcionan mucha de la información del medio necesaria para las actividades de comercialización. También proporcionan información sobre actividades reguladoras, comparaciones competitivas, encuestas de productos, encuestas de consumidores y un enorme rango de listas de correspondencia.

La incorporación de una fuente de dados adquiridos es mucho más difícil que acomodar aplicaciones operacionales hechas en casa. Esto se debe a que las fuentes adquiridas proporcionan los datos en un formato fijo, colocando la carga del procesamiento de datos en el consumidor de los mismos. Las técnicas de extracción de datos antes expuestas son más aplicables a datos residentes en base de datos. Las fuentes adquiridas por lo general proporcionan grandes archivos con estructuras fijas. Es por esto que se recomienda cargar esta información de manera local en una base de datos antes de procesarse para su extracción y para su carga en el depósito de datos.

Incorporación de información no electrónica.

No existe ninguna restricción de que el depósito de datos deba manejar sólo la información disponible en las bases de datos operacionales. Muchas organizaciones han intentado usar documentos en papel para la recuperación de información histórica. Extraer y carga información no electrónica es una tarea compleja que comprende lo siguiente:

- Conversión de información no electrónica a electrónica. Esto implica transcribir los documentos en papel a medios magnéticos. Por lo regular esta información se convierte mediante un digitalizador en una imagen electrónica del documento.

- Conversión de una imagen electrónica en información con base en texto. La imagen electrónica no es susceptible procesamiento de datos y debe convertirse en texto. Para este fin se usa generalmente un software que realiza el reconocimiento óptico de caracteres (OCR por las siglas de Optical Carácter Recognition). La precisión de reconocimiento varía de pobre a excelente, dependiendo de factores como la calidad de imagen, la selección de fuentes en el documento inicial y el grado de contraste del papel y el texto.

- Creación de meta datos que describan y organicen la información dentro del documento. Una vez cargados los metadatos en el almacén de datos, se utilizan para organizar la información dentro de éste.

Otros aspectos que se deben tener en cuenta en un ambiente heterogéneo son los siguientes:

- Tipos de plataformas que se manejan y donde residen las bases de datos operacionales. (Windows, OS/2, UNIX, LINUX)

- Tipo y cantidad de Sistemas de administración de base de datos (DBMS) Que maneja la organización.

- Tipos de interfaz con el depósito de datos: nativa, DBMS, ODBC u otras interfaces abiertas.

- Interfaz amigable disponible. Tratar de desarrollar en lo posible aplicaciones personalizadas, divulgar resultados y exportar resultados a aplicaciones de escritorio y personalizadas.

- Manejo de un lenguaje que pueda explotar la mayoría de las bases de datos operacionales de la organización.

- Integración del ambiente de procesamiento informático. El administrador del depósito de datos debe de integrar de forma adecuada el modelo de datos del depósito de datos, el sistema de administración de base de datos, y la cultura o función de la organización.

- Manejo de acceso de múltiples bases de datos y plataformas heterogéneas para permitir un total acceso al depósito de datos y mercado de datos de la empresa.

Los usuarios también forman parte del ambiente heterogéneo en el que esta implantado un deposito de datos estos usuarios tienen diferentes niveles de confianza y experiencia con la tecnología. Esto se muestra en la figura 2.5.1.

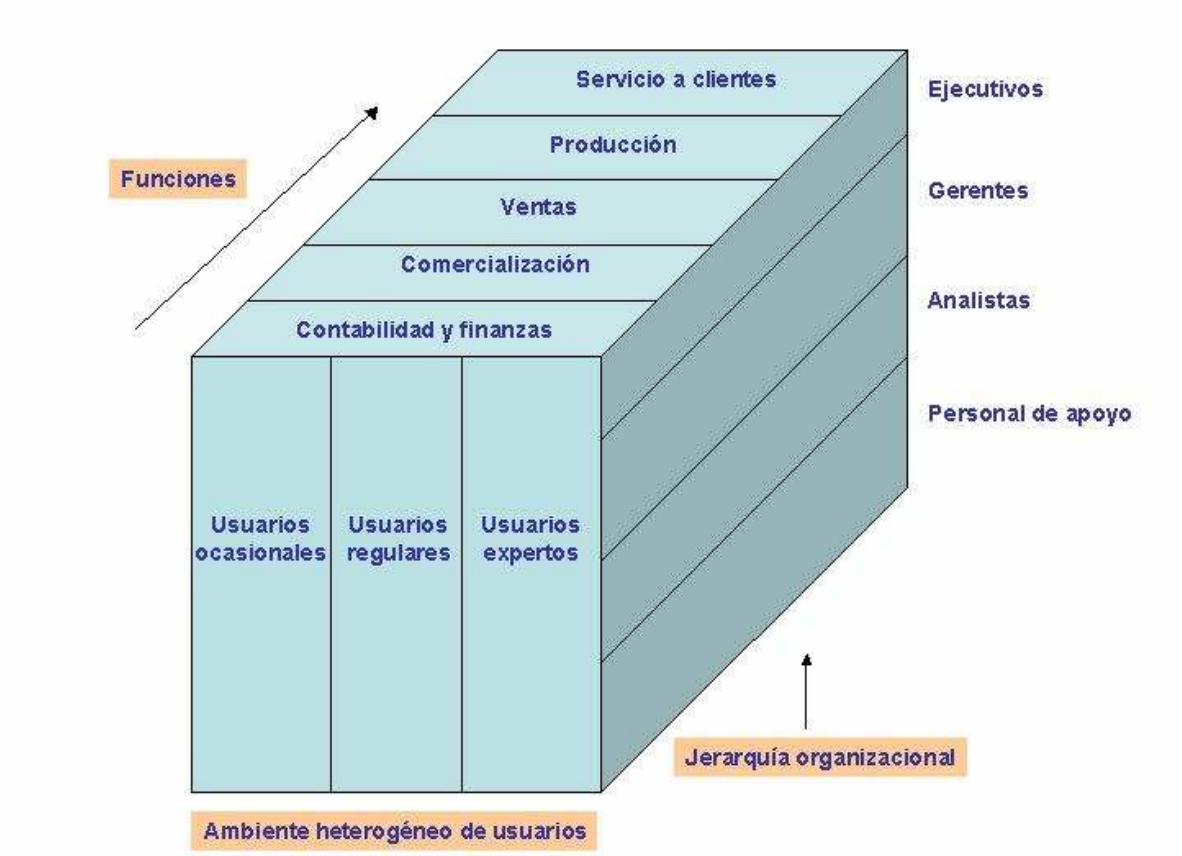

Figura 2.5. 1
Ambiente heterogéneo de Usuarios

Desde la perspectiva de un depósito de datos, los usuarios forman parte del ambiente en el que funciona un deposito de datos, debido a que estos usuarios utilizan el deposito de datos de acuerdo a sus funciones y necesidades, estos se dividen en categorías de acuerdo a su jerarquía, su función o por su nivel de competencia en cómputo en la organización. Los principales usuarios desde el punto de vista organizacional que utilizan un depósito de datos son: el director

general y al director operativo; ejecutivos de primer nivel como el director financiero y el contralor; gerentes de mandos medios; analistas empresariales y de tecnología de información y personal administrativo y de apoyo.

Los principales departamentos que utilizan el depósito de datos son: Contabilidad y finanzas, comercialización y ventas, producción e ingeniería, servicios de apoyo al cliente y administración.

Se puede dividir en tres grandes grupos a los usuarios, siendo estos usuarios ocasionales, regulares y expertos.

Todos estos usuarios explotaran la información del depósito de datos de acuerdo a sus necesidades y actividades dentro de la empresa por lo que deberán capacitarse para acceder, recuperar e interpretar la información recuperada del depósito de datos. Los usuarios deberán tener la capacidad de convertir los datos en hechos y conocimientos y utilizar dicho conocimiento para tomar decisiones o plantear recomendaciones y alternativas.

2.5.3 Seguridad en los depósitos de datos

Es muy importante la seguridad así como la administración del acceso al depósito de datos. La información de un depósito de datos es valiosa para le empresa pero también para la competencia. El control del acceso al depósito de datos se complica por varios factores:

- Un depósito de datos esta construido principalmente como un conjunto abierto de datos de la empresa. Este ayuda en la toma de decisiones y lo pueden utilizar analistas y personal operacional para mejorar sus actividades y derivar una ventaja estratégica competitiva. La incorporación de controles de seguridad va en contra de la necesidad de que el depósito de datos sea un sistema abierto.

- Los usuarios consultan información dentro del depósito de datos a diferentes niveles de resumen. El usuario puede comenzar con datos muy resumidos, pero puede tener la necesitar información cada vez mas detallada. En cambio otros usuarios solo necesiten consultar información a un solo nivel de resumen, estas diferencias en cuanto a consultas de información hacen difícil limitar a cada tipo de usuario a nivel de tabla e hilera de datos.

- La naturaleza de las herramientas de acceso y consulta de la información en un depósito de datos es principalmente la de navegar en niveles de información resumida y detallada. La utilización de intrincados controles de

seguridad puede impedir dicha navegación e impedir que los usuarios exploten la información del depósito de datos.

Estos puntos hacen difícil la seguridad en un depósito de datos sin embargo debe de implementarse una política de seguridad donde se deben imponer restricciones a las capacidades de obtener información a detalle así como el control de acceso a tablas especificas de datos resumidos e información de detalles operacionales.

Aunque un depósito de datos no maneja datos operacionales de misión crítica, la naturaleza de la amenaza a la seguridad no es al daño a los datos sino la información que se pueda obtener del detalle de estas transacciones, debido a que con este tipo de información se puede obtener secretos de estrategias corporativas.

Por todos estos motivos la seguridad del almacén de datos tiene estos principales propósitos:

- Bloquear a usuarios que no están autorizados a consultar la información.
- Controlar el acceso a porciones del almacén de datos por cada usuario.
- Restringir por usuario o por grupo de usuarios el acceso a un subconjunto del almacén de datos.

A continuación trataremos las consideraciones a contemplar en cuanto a seguridad de accesos y seguridad de datos (backup), puesto que si bien la seguridad de accesos (al nivel de datos y de aplicación) debe ser tratada de la misma manera que en los sistemas operacionales, los procedimientos de copias de seguridad merecen un especial tratamiento.

Tal y como ocurre en los sistemas operacionales, un sistema de depósito de datos debe poder realizar procedimientos de recuperación de la información desde cualquier momento en el que los datos estaban validados. Un depósito de datos, debe poder contar con procedimientos de recuperación, que permitan restaurar los datos ante cualquier situación de catástrofe.

No obstante, es preciso tener en cuenta otras consideraciones, por ejemplo dependiendo del tamaño de un mercado de datos, se puede elegir no realizar un backup, sino realizar un refresco especial desde los datos operacionales, dependiendo de la periodicidad estándar de carga.

En cuanto a la seguridad de acceso, es preciso el implantar niveles de acceso a la información, realizando un plan completo de seguridad que contemple:

- Acceso a recursos de la red (local o intranet)
- Asignación de usuarios a grupos con diferentes perfiles de seguridad
- Asignación de niveles de autorización de aplicación a grupos de usuarios
- Bloqueo de usuarios o grupos de usuarios por día, fecha, ubicación, tipo de reporte o consulta específicos.
- Seguridad a nivel de almacén de datos, mediante los procedimientos provistos por el mismo almacén.
- Manejar permisos para las restricciones de uso de recursos, tales la capacidad de crear tablas temporales y consultas ad-hoc
- Restringir la utilización de recursos informáticos a los perfiles de usuario, para evitar que estos acaparen grandes cantidades de recursos. (limitar el espacio para la generación de tablas temporales)

2.5.4 Selección de las herramientas para el usuario final

El administrador del depósito de datos realiza dos actividades importantes que son el servicio a usuarios empresariales y el servicio de administración técnica. La meta del servicio a usuarios empresariales consiste en maximizar la productividad de estos usuarios y ayudarles a extraer el máximo valor del depósito de datos. En la parte de administración técnica, el administrador del depósito de datos se encarga de proteger los activos de datos de la organización y mantener en operación el depósito de datos.

Servicios a usuarios empresariales.

El administrador del depósito de datos debe ofrecer los siguientes servicios a usuarios empresariales:

- Crear un conjunto almacenado de consultas y reportes con acceso a través de FAQ, FAR y FAD (esto mejora la reutilización y la estandarización las cuales provocan productividad y rentabilidad)
- Personalizar el ambiente de ejecución con iconos de acceso, botones, barras de herramientas y menús para facilitar el uso a los ejecutivos, gerentes y analistas empresariales.

- Incorporar opciones a las consultas y reportes predefinidos, de modo que los usuarios empresariales puedan incluir variables y filtros durante la ejecución.

- Mejorar el desempeño mediante diferentes técnicas de generación de índices. (Técnica de mapa de bits, índices de patrón e índices de esquema estrella)

- Mantener la seguridad de los datos con un acceso responsable.

- Conservar un perfil de seguridad, actualizado y sincronizado con otras políticas de seguridad de la organización.

Servicios de administración técnica.

El administrador del depósito de datos debe, como mínimo, proporcionar los siguientes servicios de administración técnica a los usuarios del depósito de datos:

- Ofrecer un depósito de meta datos de procesamiento informático, opciones para examinar y navegar, ampliando el depósito técnico de Tecnología de Información de la organización, adquiriendo uno nuevo o bien adaptando uno a la medida.
- Administrar las funciones de acceso, recuperación y examen de los meta datos.

- Mantener los datos del procesamiento informático sincronizados con los meta datos del depósito de datos. Si los meta datos del procesamiento informático están almacenados en la estación de trabajo del usuario, debe entonces sincronizar los meta datos de la estación de trabajo del usuario con los meta datos del depósito de datos.

- Manejar consultas de escape mediante diversos límites durante le ejecución como: tiempo de ejecución, ámbito de la consulta y cantidad de registros recuperados.

- Mejorar el desempeño mediante la vigilancia de la velocidad de consulta y de la velocidad del reporte, debido a que para funciones conjuntas de consulta y reporte podría haber una amplia disparidad en sus desempeños.

- Optimizar el desempeño con la programación de reportes por lotes.

- Controlar el desempeño dividiendo la carga de trabajo entre el cliente, el servidor de aplicaciones y el servidor de base de datos del depósito de datos.

- Proporcionar controles de acceso por el usuario o grupo; por tabla de base de datos, a nivel de columna o hilera (Como en el acceso a la información de nómina sólo a las porciones adecuadas. Reutilizar, si los hay, los perfiles existentes de seguridad)

Desempeño

El desempeño es un factor de éxito fundamental para la satisfacción del usuario. Dentro de las acciones para mejorar el desempeño están las siguientes:

- Controlar y administrar dónde se realiza el proceso, en la estación de trabajo del cliente, en el servidor de aplicaciones o en el servidor del almacén de datos. (El administrador necesita la capacidad de distribuir el procesamiento para obtener el máximo desempeño)

- Un rango de técnicas de índices.

- Gran apoyo a consultas.

- No reexpedir ni ejecutar consultas sólo para reformatear un reporte o analizar desde un punto de vista diferente.

- Recuperación de resultados de consulta sólo cuando sea necesario.

- Procesamiento interno o por lotes, así como programación de solicitudes para aprovechar las horas de menor uso y, por tanto, minimizar las cargas de desempeño durante las horas pico.

- Extracción de subconjuntos de datos con sus meta datos para un depósito local / personal a fin de accederlos y analizarlos después, con esto se logra reducir los costos de red e incrementar la confiabilidad.

- Un rango de controladores de consultas. Ámbito accesible del depósito de datos, duración, cantidad de registros recuperados y cancelación de una consulta en proceso.

Teniendo en cuenta los puntos mencionados anteriormente se debe escoger una herramienta que cumpla con la mayoría de las características citadas. Dentro del sistema operativo Windows NT. Microsoft SQL Server es un sistema de

administración de bases de datos escalable y de alto rendimiento, diseñado específicamente para la computación distribuida de cliente / servidor en Windows NT. Hay varias razones por las cuales los clientes han escogido a SQL Server para la implementación de sistemas de soporte a la toma de decisiones en el depósito de datos. Entre éstas se encuentran:

- Compatibilidad con muchos otros componentes y herramientas de software utilizadas en la instalación de los depósitos de datos.
- Integración con Internet.
- Opciones de conectividad.
- Capacidades para la administración del sistema.
- La plataforma de mejor rendimiento en el sistema operativo Windows NT.
- Replicación de datos incorporada.
- Habilidad para incorporar fácilmente los datos obtenidos en las aplicaciones de oficina del escritorio, entre las que se encuentran Microsoft Word, Microsoft Excel, Microsoft PowerPoint®, y el correo electrónico.
- La continúa inversión de Microsoft en nuevas capacidades y productos de software.

Microsoft ofrece la plataforma de preferencia en muchas instalaciones de sistemas de depósito de datos, y se encuentra invirtiendo en capacidades aún más novedosas en este importante mercado, como respuesta a las sugerencias de los clientes.

En cuanto a la tecnología de Internet e intranet promete a los usuarios dar un acceso barato a los datos de los Depósito de datos y Mercado de Datos, a través de los Web Browsers.

Los productos, generalmente consisten en código situado entre los servidores Web y los productos OLAP. Los vendedores líderes, están empezando a incluir Java y/o ActiveX en sus productos, en comparación a las limitadas funcionalidades HTML de los productos iniciales.

Aunque los browsers no contienen toda la capacidad de acceso disponible en las aplicaciones Cliente−Servidor, Internet es una buena solución para dar acceso a los datos, cuando los usuarios son muchos o están geográficamente dispersos, especialmente, los usuarios que no necesitan una interacción muy sofisticada.

Los vendedores están desarrollando el concepto de Mercado de Datos Virtuales para satisfacer la necesidad de los usuarios de acceder a muchos Mercados de Datos, sin necesidad de excesivas replicaciones entre ellos. Los Mercados de Datos Virtuales son vistas de varios Mercados de Datos Físicos, o del Depósito de datos corporativo, brindadas a grupos específicos de usuarios.

Otros vendedores, como Sagent Mercado de Datos Solution, de Sagent Technology Inc., proveen los conceptos de Vista Básica y Meta Vistas. Una Vista Básica es una representación gráfica de una base de datos que incluye tablas, columnas y relaciones. Una vez que una Vista Básica es creada, múltiples Meta Vistas se pueden derivar de ella. Una Meta Vista es una representación lógica de partes, de una o más Vistas Básicas. Inicialmente las tablas son desplegadas como categorías, y los campos como partes. Se pueden renombrar o remover categorías o partes de una Meta Vista. Esos cambios no afectan a las Vistas Básicas que la soportan. La Meta Vistas permite usar una única Vista Básica para presentar diferentes partes de la información a diferentes grupos de usuarios.

Esta es una de varias tecnologías las cuales la mayoría están todavía en desarrollo.

2.6 Tendencias

El gran crecimiento de las bases de datos y el aumento de las capacidades de almacenamiento de información, han hecho que todo tipo de organizaciones puedan

disponer de una gran cantidad y variedad de datos relativos a su actividad diaria. En muchas de estas organizaciones se han dado cuenta del potencial que tiene esta información para el apoyo a la gestión. Su estudio permite ver la evolución y desarrollo de las organizaciones, y por lo tanto, trazar una línea de tendencia que muestre por dónde pueden moverse en un futuro.

Así, el estudio de los datos y la información almacenada en las bases de datos ofrece una visión perspectiva (qué se está haciendo y cómo se está haciendo) y prospectiva (cómo puede evolucionar la organización en un futuro a corto-medio plazo) de la organización, y es por ello por lo que tiene una función de apoyo a la toma de decisiones.

Las nuevas tecnologías permitirán el aprovechamiento de economías de escala para hacer uso de grandes depósitos de datos, que además de almacenar información, brindarán una serie de servicios como copias de respaldo y disponibilidad permanente de los datos, estos beneficios son costosos para una empresa si esta no cuenta con un depósito de datos.

Calidad de la información.

Las tendencias que se han detectado en los últimos años obligan a reflexionar sobre la interrelación existente entre el proceso de un depósito de datos y la calidad del conocimiento de la organización y la actividad orientada a desarrollar una inteligencia del negocio. En los últimos meses es notoria una tendencia por la cual las empresas comprenden con mayor claridad su necesidad de generar y compartir conocimiento, almacenarlo eficientemente y generar capacidades y procesos de inteligencia de negocios, orientados a mejorar la convergencia entre ese conocimiento y las necesidades estratégicas, tácticas y operativas, permitiendo anticipar los diferentes escenarios futuros del negocio.

Por otra parte, se está generando una fuerte conciencia de que la solución a los grandes y crecientes volúmenes de datos que se requiere conservar en un depósito de datos no se encuentra en mayores capacidades de almacenamiento, sino en refinadas técnicas de monitoreo de los entornos de depósito de datos y novedosos métodos de almacenamiento inteligente, con tecnologías *near-line* ("casi en línea"). Por lo que el futuro de un deposito de datos no esta en su capacidad del almacenamiento si no en la calidad de la información que este puede ofrecer.

Consultas a un Depósito de datos a través de Internet.

Por último, es imposible dejar de mencionar otra tendencia significativa. El crecimiento del comercio electrónico está basado en gran medida en las posibilidades que otorga el depósito de datos. Aun cuando no sea la cara visible de una página *web*, los millones de transacciones y visitas se almacenan en bases de datos, y no pueden ser analizadas sin utilizar un proceso de depósito de datos.

El éxito de algunos sitios se basa, sin lugar a dudas, en la explotación del conocimiento de los consumidores y su comportamiento, obtenido a partir del uso de la información que se almacena sistemáticamente en un Depósito de datos, oculto detrás de Internet.

Existe una visión orientada a que algún día, cada vez más cercano, todo el procesamiento se haga detrás de Internet, lo que generará la posibilidad de acceder a infinitos proveedores de servicios. Dentro de estos servicios estará el poder consultar un d*epósito de datos*. Cada empresa o institución académica, sin requerir de grandes inversiones de *hardware* y a través de un browser cualquiera, podrá almacenar, recuperar o analizar su propio conocimiento.

Bibliotecas digitales.

Una tendencia que también esta tomando auge es el de las bibliotecas digitales debido a que estas tienen características útiles para los tomadores de decisiones y complementan de manera integral los resultados que ofrece un deposito de datos.

Las bibliotecas digitales se dividen en tres clases:

- Biblioteca Digital Autónoma (BDA). Es la biblioteca clásica normal implementada de manera completamente automatizada. La BDA es simplemente una biblioteca cuyos fondos son digitales digitalizados. La BDA es independiente - el material está localizado y centralizado. De hecho, es un ejemplo automatizado de la biblioteca clásica con las ventajas de la automatización.

- Biblioteca Digital Federada (BDF). Es una federación de varias BDAs independientes en la red, organizadas en torno a un tema común y unidas en la red. Una BDF consta de varias BDAs que forman una biblioteca en red con una interfaz de usuario transparente. Las distintas BDAs son heterogéneas y están conectadas vía comunicación en red. El mayor desafío en la construcción y mantenimiento de una BDF es la interoperabilidad (puesto que los distintos depósitos utilizan diferentes normas y formatos de meta datos).

- Biblioteca Digital Recolectada (BDR). Es una biblioteca virtual que proporciona acceso resumido a materiales relacionados dispersos en la red. Una BDR solo maneja meta datos con apuntadores a los fondos que están a un solo clic de distancia en el ciberespacio. El material alojado en las bibliotecas está recolectado (convertido en sumarios) de acuerdo con la definición de un Especialista de la Información (EI). Sin embargo, una BDR tiene las características de una biblioteca digital normal, está sutilmente estructurada y centrada en una materia. Tiene numerosos servicios bibliotecarios y un alto control de calidad mantenido por el responsable de anotar los objetos de la biblioteca.

Bioinformática.

La bioinformática es una ciencia que esta conformada por varias ciencias de la vida así como también de la informática, por lo que proporciona herramientas y recursos necesarios para favorecer la investigación biomédica. La bioinformática como campo interdisciplinario, comprende la investigación y el desarrollo de sistemas útiles para entender el flujo de información entre los genes y las estructuras moleculares así como su función bioquímica, conducta biológica y su influencia en las enfermedades y la salud.

Dentro de los principales proyectos que están dando un gran auge a la bioinformática están los siguientes:

- El Proyecto de estudio del genoma humano. El enorme volumen de datos que genera este tipo de proyectos necesita de una herramienta como lo es un depósito de datos el cual esta diseñado para soportar tal magnitud de información.

- Biochips. Los nuevos enfoques experimentales, basados en biochips, que permiten obtener datos genéticos a gran velocidad, bien de genomas individuales (mutaciones, polimorfismos) o de enfoques celulares (expresión génica).

- Internet. Cada día mas accesible lo que permite el acceso universal a las bases de datos de información biológica

La magnitud de la información que genera las investigaciones realizadas sobre el genoma humano es tal que supera la información generada por otras investigaciones en otras disciplinas científicas. La vida es la forma más compleja de organización de la materia que se conoce. Por lo que para la investigación biológica, concretamente la obtención y análisis de las secuencias de nucleótidos de los genomas conocidos. Se necesitan equipos potentes tanto en procesamiento como en almacenaje de información.

Ante tal situación, uno de los retos de la bioinformática es el desarrollo de métodos que permitan integrar los datos genómicos de secuencia, expresión, estructura e interacciones. La integración de esta información servirá para explicar el comportamiento global de la célula viva, minimizando la intervención humana. Dicha integración, sin embargo, no puede producirse sin considerar el conocimiento acumulado durante decenas de años, producto de la investigación de miles de científicos.

La bioinformática se ocupa de la utilización y almacenamiento de grandes cantidades de información biológica, es decir, trata del uso de las computadoras para el análisis de la información biológica, entendida esta como la adquisición y consulta de datos, los análisis de correlación, la extracción y el procesamiento de la información. Todas estas actividades

son naturales y propias de los depósitos de datos por lo que en un futuro cercano las depósitos de datos serán muy utilizados por la bioinformática.

La bioinformática es un área del espacio que representa la biología molecular computacional, que incluye la aplicación de las computadoras y de las ciencias de la información en áreas como la geonómica, el mapeo, la secuencia y determinación de las secuencias y estructuras por métodos clásicos. Las metas fundamentales de la bioinformática son la predicción de la estructura tridimensional de las proteínas a partir de su secuencia, la predicción de las funciones biológicas y biofísicas a partir de la secuencia o la estructura, así como simular el metabolismo y otros procesos biológicos basados en esas funciones.

Muchos de los métodos de la computación y de las ciencias de la información sirven para estos fines, incluyendo las teorías de la información, la estadística, la teoría de los gráficos, los algoritmos, la inteligencia artificial, los métodos estocásticos, la simulación y la lógica.

Actualmente uno de los cuellos de botella de los ensayos con tecnologías basadas en biochips se encuentra en la carencia de herramientas bioinformáticas adecuadas para el análisis y gestión de los datos, debido a los enormes volúmenes de datos que ellos generan. Por lo que esta ciencia debe hacer énfasis en la necesidad de emplear técnicas de minería de datos, almacenes de información así como la mejor forma de obtener conocimientos a partir de los resultados experimentales.

En la actualidad el reto es la construcción de bases de datos las puedan soportar grandes volúmenes de información así como el establecimiento de una arquitectura que permita la realización de búsquedas inteligentes, la comunicación con otras bases de datos y la unión con herramientas de análisis y de minería de datos, específicas, que permitan responder a problemas biológicos concretos.

Los científicos, encargados de la construcción de estas bases de datos, deben disponer de conocimientos previos que permitan determinar cuáles problemas científicos concretos necesitan una solución y cuál o cuáles métodos son los mejores para resolverlos. En base ha esto los científicos han empezado analizar la opción de implementar depósitos de datos ya que estas aplicaciones ofrecen herramientas para gestionar información genética en paralelo. Para ello se emplean nuevas tecnologías de extracción de conocimientos, minería de datos y visualización. En base a estos métodos de obtención de información se aplican técnicas de descubrimiento de conocimientos a problemas biológicos como análisis de datos del genoma y el proteoma.

En estos momentos, la mayoría de los proyectos que se desarrollan en el mundo en materia de genómica y proteómica, demandan la aplicación de técnicas de la minería de datos para poder determinar qué es realmente importante dentro del enorme volumen de información que se genera diariamente en el mundo. Considérese que el número total de letras (pares de bases químicas) del ADN humano ha resultado ser de 3.120 millones. El Proyecto Genoma Humano aseguró que, a los 10 años de su creación, ha terminado un primer borrador de la secuencia y completado el 85 % del ensamblaje. De los 3.120 millones de datos que componen el «libro de la vida», los científicos han encontrado que el 99,8 % son idénticos para todas las personas.

Los depósitos de datos así como la minería de datos es utilizada en el sector comercial principalmente pero la bioinformática empieza a tener en cuenta este tipo de herramientas para soporta las investigaciones en la rama biológica ya que con estas puede enfrentar la avalancha de datos que producen las investigaciones genómicas y proteómicas.

CAPÍTULO 3

CASO PRÁCTICO DE UN DEPÓSITO DE DATOS

3.1 Caso Práctico

En este capítulo se presenta el desarrollo de un caso práctico con el objetivo de aplicar los conocimientos adquiridos durante el desarrollo de los capítulos anteriores, con la finalidad de lograr crear un pequeño depósito de datos que permita visualizar el proceso básico para la creación del mismo al lector y así mismo proporcionar una herramienta básica para la explotación de la información a partir de un software comercial y de fácil acceso al usuario final.

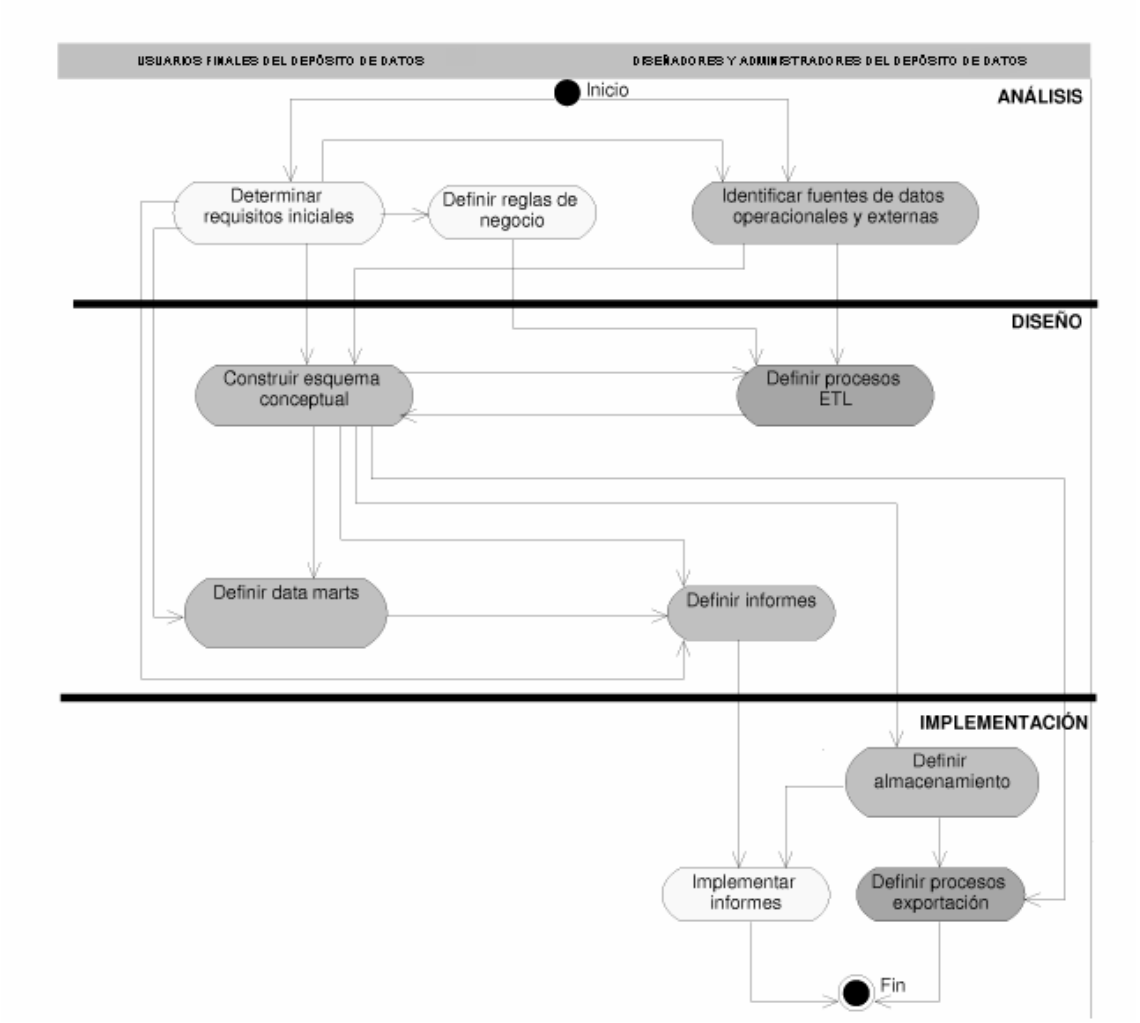

Figura 3.1. 1
Representación del proceso del caso práctico

- Comenzamos determinando lo requisitos iniciales: definiendo los alcances de los depósitos de datos mediante entrevistas con los que serían los usuarios finales, los

expertos en el análisis de la información; se revisan informes ya existentes y se recopilan los requisitos de los usuarios.

- Posteriormente se definen las reglas de negocio; definir las diversas reglas implica definir la construcción del depósito de datos, por ejemplo la definición de medidas derivadas como “beneficios netos” o “porcentajes de devolución de producto”.

- Se definen las fuentes de datos, tanto operacionales como externas (datos económico, censos de población, etc.) alimentarán el almacén de datos. Para ello se tiene en cuenta las necesidades expresadas por los usuarios finales.

Las fuentes de datos utilizadas en este caso práctico se obtuvieron de una base de datos operacional, esta base de datos es utilizada actualmente en una dependencia de gobierno dedicada a la comercialización de productos al publico, estos datos obtenidos es el resultado de la operación de ventas y compras, traspasos y movimientos en sus almacenes, franquicias y demás a través de los últimos años, en particular nos enfocaremos a un solo cuestionamiento las ventas realizadas así como de sus compras a los proveedores.

La necesidad de proyectar la información y extrapolar resultados para una toma de decisiones firme y segura sobre la adquisición, consignación y venta de los artículos adquiridos por la dependencia, esta necesidad de proyectar la información histórica de los años anteriores al 2006 para tomar la decisión correcta asignada a cada producto, proveedores, clientes es decir, la tarea principal es la decisión de qué presupuesto se tendrá para el año 2008, fue realmente correcto el manejo de las operaciones de compra y ventas realizadas en la dependencia. Para ello se consideran las tablas de la base de datos operacional del respaldo de todo el año 2006 usando principalmente el contenido de las tablas que a continuación se mencionan.

Proveedores
Productos
Entradas por compras
Salidas por ventas
Sucursales
Clasificaciones

- Para construir el esquema conceptual existen dos estrategias: top-down que define el depósito de datos según los requisitos de los usuarios finales o bottom-up que define el depósito de datos en base a los datos disponibles en la fuente de datos, ésta última fue seleccionada para el caso de práctico.

- La definición de los data mart, en este caso no será realizado. Sin embargo este caso práctico puede definirse como data mart debido a su pequeña consulta.

Con base a esta información se construyo el modelo dimensional. Debido a que el modelo de datos utiliza herramientas como usando ASP y SQL como fuente de datos operacional, entonces tiene que ser llevado a una representación relacional. El esquema multidimensional con tabla de hechos y dimensiones.

Se propuso el siguiente esquema con las jerarquías de cada dimensión para el conjunto de datos para el reporte de ventas.

Proveedor	Producto	Paises	Ventas
IdProveedor	Clave	Pais	Salidas
Estado	Concepto	NombrePais	Entradas
GpoProv1	Gpoprod1		Fecha

Está información es llevada al Depósito de datos y con base en ella se creo uno de los modelos dimensionales.

- Se define el proceso ETL como un mapeo entre el origen de datos y el depósito de datos; las reglas de negocio se aplican para calcular atributos derivados, definir transformaciones de atributos, etc. Esta actividad y la anterior definen un ciclo, ya que al crear los procesos ETL se puede detectar algún fallo en el depósito de datos, puede ser que un atributo de una dimensión no exista en el origen de datos, por lo que sería necesario modificar el depósito de datos.

Como principio se crearon todos los procesos y en especial el proceso ETL, este proceso tuvo que estar acompañado de una administración responsable, robusta y en un ambiente seguro para evitar corrupción o violación en los datos, ya que son la materia prima y más valiosa para el depósito de datos, el ambiente de seguridad y administración debe existir del lado de alimentación del depósito como del lado de usuarios y accesos a la misma.

- Y Finalmente se definirá los informes o consultas iniciales, que será mostradas mediante los cubos.

Se requiere de herramientas que permitan la extracción de información y la construcción de reportes a partir de los datos en el depósito, estas herramientas deben ser fáciles de usar, intuitivas, deben permitir que los usuarios, naveguen libremente de lo general a lo particular en los datos, así como permitir documentar los mismos y de fácil capacitación para uso masivo pero recordemos siempre en un ámbito de seguridad y administración estable; se debe establecer previo a la manipulación las dimensiones de los alcances de los requerimientos de los usuarios, ya sea para adquirir las herramientas que cumplan cabalmente con los requisitos, construirlas, o simplemente personalizar las adquiridas o las existentes. Para que en la consulta al depósito de datos se visualicen, manipulen y analicen los datos en un ambiente topológico, y en esquena cliente / servidor, donde el servidor central mantiene los datos, la administración y la seguridad, el cliente inicia una transacción con él, realiza una petición de los datos, y una vez que el servidor haya autorizado la transacción y respondido la petición, el cliente los analiza, interpreta y procesa.

En el proyecto uno de los esquemas más importantes es el de acceso al depósito, ya que la finalidad del Depósito de datos no es simplemente almacenar todos los datos del negocio por área, sino la esencia es servir de base para la inferencia y construcción de información que sustente la toma de decisiones, en la mayoría de textos tradicionales de bases de datos el termino dato es permutable y equivalente al término información, pero la diferencia es sencilla, pongamos dos ejemplos que permiten definir y diseminar claramente los dos términos, primero se realiza una consulta al repositorio de datos ¿Cuánto vendí la semana pasada y en donde?; y la segunda ¿Qué combinación de productos se vendieron mejor la semana pasada y como están segmentados a lo largo de la ciudad?, la diferencia es clara en el primer caso es una simple consulta a un dato en el depósito, y la segunda extrae, información de la misma. Combinando en línea distintas variables que determinan el comportamiento del negocio.

Las herramientas OLAP son implementadas en arquitecturas cliente servidor y su finalidad es brindar rápidas respuestas a múltiples consultas.

Se crea un caso práctico de forma sencilla con la posibilidad de utilizar herramientas al alcance del alumno.

Debido a que la base está en MySQL, entonces tiene que ser llevado a una representación relacional. El esquema multidimensional con su tabla de hechos y dimensiones se muestra la siguiente figura 3.1.2 Representándose en un esquema de Copo de Nieve.

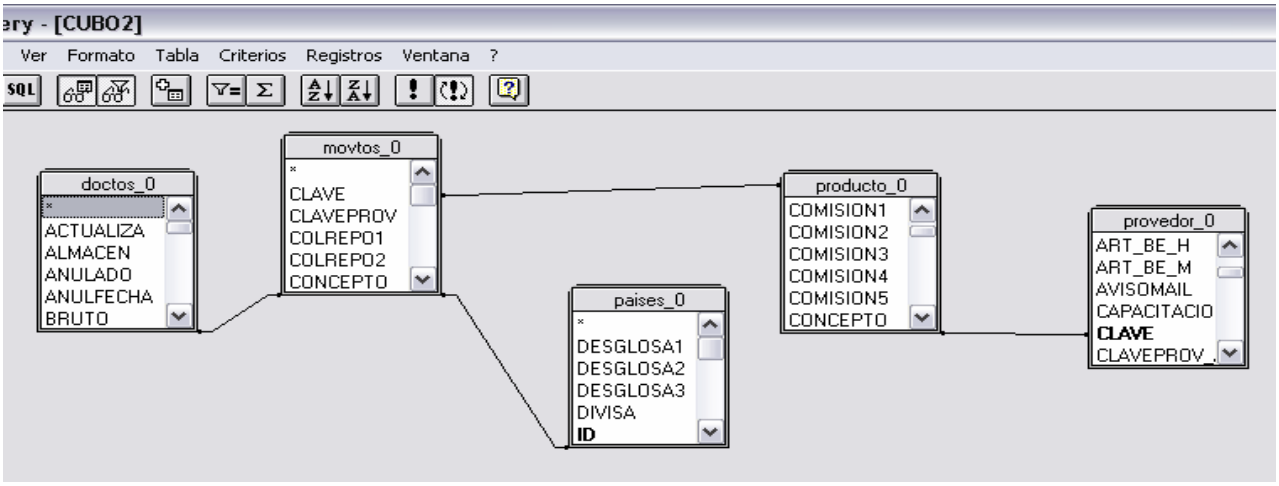

Figura 3.1. 2
Representación del depósito de datos en esquema Copo de Nieve

El programa que define el esquema multidimensional para la proyección de las ventas se encuentra en MySQL se codifico en SQL. Adicionalmente se construyeron un conjunto de índices en cada tabla para ofreces un mejor desempeño.

La arquitectura para análisis de información se construyo un programa que lleva a cabo el proceso de ETL basado en las definiciones de la proyección. Para ello se programa una aplicación en ASP que nos ayudará a la extracción, transformación y carga de los datos desde la base de datos operacional a la base que fungirá como depósito de datos.

El proceso es el siguiente, el administrador debe ejecutar la aplicación que extrae los datos, al mismo tiempo que se verifica para evitar duplicidad, además de insertar sólo los datos requeridos y eliminando la información inútil para este caso.

Debemos tomar en cuenta que tipo de manejador de base de datos estamos utilizando ya que las conexión no son las mismas, para este caso se muestra el código de el archivo .inc que nos permitirá la conexión a la base operacional de la dependencia de gobierno, así también la conexión al deposito.

Conexión1.inc

```
<%
  SET OCONN = Server.CreateObject("ADODB.Connection")
   SET RS=Server.CreateObject("ADODB.Recordset")
   OCONN.MODE = 3

   conn_string = "Driver={MySQL ODBC 3.51
Driver};Server=127.0.0.1;Port=3306;Database=baseoperacional;Uid=root;Pwd=maiyec"
   oconn.Open(conn_string)
   strSistemaManejadorBaseDatos = "MYSQL"
%>
```

Conexión2.inc

```
<%
   SET Conn2 = Server.CreateObject("ADODB.Connection")
   SET RS=Server.CreateObject("ADODB.Recordset")
   Conn2.MODE = 3

   conn2_string = "Driver={MySQL ODBC 3.51
Driver};Server=127.0.0.1;Port=3306;Database=deposito;Uid=root;Pwd=maiyec"
   Conn2.Open(conn2_string)
   'strSistemaManejadorBaseDatos = "MYSQL" conn_string = "Driver={MySQL ODBC 3.51
Driver};Server=127.0.0.1;Port=3306;Database=deposito;Uid=root;Pwd=pass"
   oconn.Open(conn_string)
%>
```

El siguiente programa llama a los archivos .inc, para la conexión a la base operación para su extracción, posteriormente su transformación y finalmente para la carga de la base que será el depósito.

Procesoetl.asp

```
<html>
<head>
<title>COPIAR REGISTROS DE MYSQL A ACCESS</title>
<style type="text/css">
<!--
body {
    font-family: Arial, Helvetica, sans-serif; font-size: x-small
}
-->
</style>
</head>
<body bgcolor="#FFFFFF" text="#000000">
   <!--#include file="CONEXION2.inc"-->
   <!--#include file="CONEXION1.inc"-->

<%
Server.ScriptTimeout=50000

'---------------------------------------------------------------------
'------------COMIENZA EL APARTADO DE FUNCIONES-------------------------
'---------------------------------------------------------------------

   '////////////////////////////////////////////////////////////////////////////////////////////////////////////
  'funcion que obtiene todas las columnas y sus tipos y los devuelve a un arreglo
   Private Function ObtenerColumnas(Conexion, NTabla, Tipo)
      dim ArrayTmp(9000,6)
      if Tipo = "MDB" then
```

```
      set RSTmp = Conexion.OpenSchema(4)
      i=0
      while RSTmp.eof = false
        if ucase(RSTmp.fields("TABLE_NAME")) = ucase(NTabla) then
          ArrayTmp(i,0)=RSTmp.fields("COLUMN_NAME") 'field
          ArrayTmp(i,1)=RSTmp.fields("DATA_TYPE") 'type
    '      RESPONSE.WRITE ArrayTmp(i,0)&"<BR>"
         if ucase(RSTmp.fields("IS_NULLABLE")) = "VERDADERO" then
            Nulleable = "NULL"
          else
            Nulleable = "NOT NULL"
          end if
                        '   if ucase(RSTmp.fields("Is_PRIMARYKEY")) = "VERDADERO" THEN RESPONSE.Write("SI ES
LLAVE")
                                  'IF RSTmp.fields("IsAutoIncrement") THEN RESPONSE.Write("SI ES AUTONUMERICO")

          ArrayTmp(i,2)=Nulleable 'null
          ArrayTmp(i,3)="" 'key
          ArrayTmp(i,4)="" 'extra
          ArrayTmp(i,5)=RSTmp.fields("CHARACTER_MAXIMUM_LENGTH") 'numero maximo de caracteres
          i=i+1
        end if
        RSTmp.movenext
      wend
      set RSTmp = nothing
    end if
    if Tipo = "MYSQL" then
      set RSTmp = Conexion.Execute("describe " & NTabla)
      i=0
      while RSTmp.eof = false
        ArrayTmp(i,0)=RSTmp.fields("FIELD") 'field
        ArrayTmp(i,1)=RSTmp.fields("TYPE") 'type
        ArrayTmp(i,2)=RSTmp.fields("NULL") 'null
        if ucase(RSTmp.fields("NULL")) = "NO" then
          Nulleable = "NOT NULL"
        else
          Nulleable = "NULL"
        end if
        ArrayTmp(i,2)=Nulleable 'null
        ArrayTmp(i,3)=RSTmp.fields("Key") 'key
        ArrayTmp(i,4)=RSTmp.fields("Extra") 'extra
        MAXIMO =""
        FOR a=1 TO LEN(ArrayTmp(i,1))
          sim=mid(ArrayTmp(i,1),a,1)
          if sim="0" or sim="1" or sim="2" or sim="3" or sim="4" or sim="5" or sim="6" or sim="7" or sim="8" or sim="9" then
            MAXIMO=MAXIMO&sim
          end if
        next
        ArrayTmp(i,5)=MAXIMO 'numero maximo de caracteres
        i=i+1
        RSTmp.movenext
      wend
    end if
    ObtenerColumnas = ArrayTmp
  End Function

  '*************************************************************
  'vamos a comprobar las columnas
  'funcion que comprueba las mismas columnas, se requiere la conexion, la tabla y el manejador por cada tabla
  'devulve tambien un arreglo bidimensional con las columnas faltantes
  Private Function ChecarColumnas(Conex1,Tab1,Manej1,Conex2,Tab2,Manej2,strALMACEN, CAMPOCOND)
    dim ColumnasNoEncontradas(1000,5)
    C1 = ObtenerColumnas(Conex1,Tab1,Manej1)
    C2 = ObtenerColumnas(Conex2,Tab2,Manej2)

    strINSSQL=""
```

```
    strSQL = " SELECT * FROM " & Tab1 &CAMPOCOND
    RESPONSE.WRITE strSQL&"<BR>"
    RESPONSE.FLUSH
    SET RS = Conex1.EXECUTE(strSQL)
    IF NOT RS.EOF  THEN
      DO WHILE NOT RS.EOF
         strINSSQL = " INSERT INTO "& Tab2 & " ("
         strCAMPOS =""
         strVALOR =""

         i=0
         while C1(i,0) <> ""
         '  IF C1(i,3)="PRI" AND C1(i,4)="auto_increment" then
         '  else
              DATO =  RS.FIELDS(C1(i,0))
              strCadena =UCASE(C1(i,1))
              if Manej1="MYSQL" then
                FOR a=1 TO LEN(C1(i,1))
                  sim=mid(C1(i,1),a,1)
                  if sim="(" then
                    strCadena=UCASE(mid(C1(i,1),1,a-1))
                  end if
                next
              end if

              if strCadena="DOUBLE" or strCadena="INT" or strCadena="SMALLINT" then
                        DATO = DATO
                        IF ISNULL(DATO) OR DATO="" THEN DATO=0
              else
                 'Esa sentencia producirá un error de "error de sintaxis, falta operador...".
                 'Para resolverlo debemos sustituir la comilla simple por un par de comillas, así:
                 'sDestinatario = "B's Beverages"
                 ' reemplazamos las comillas simples por un par
                 ' de comillas

                        DATO = DATO
                        if trim(DATO)<>"" then
                            DATO = Replace(DATO, "'", "''")
                        end if
                        DATO = "'"&DATO&"'"
              end if

              strCAMPOS = strCAMPOS & C1(i,0) & ","
              strVALOR = strVALOR &  DATO & ","
          'end if
          i=i+1
         wend
            strCAMPOS = LEFT(strCAMPOS,LEN(strCAMPOS)-1)
            strVALOR  = LEFT(strVALOR,LEN(strVALOR)-1)
             RESPONSE.WRITE strVALOR&"<BR>"
            strQUERY = strINSSQL& strCAMPOS  &")VALUES("&strVALOR&")"
            RESPONSE.WRITE strQUERY&"<BR>"
            RESPONSE.FLUSH
            Conex2.EXECUTE ( strQUERY )

        RS.MOVENEXT
      LOOP
    END IF
       RESPONSE.WRITE "<BR>"
    ChecarColumnas = ColumnasNoEncontradas
  End Function


  '////////////////////////////////////////////////////////////////
  'function que "traduce" a un mismo tipo y largo los campos
  Private Function TransType(Tipo,Maximo,BD)
    Regreso = ""
```

```
' response.write tipo & "</br>"
 RESPONSE.FLUSH

 if BD="MDB" then
  if Tipo = 130 then
     if BD="MYSQL" then
        if Maximo >= 1073741823 then
           Regreso = "LONGTEXT"
        else
           Regreso = "VARCHAR(" & Maximo & ")"
        end if
     end if
     if BD="MDB" then
        if Maximo >= 1073741823 then
           Regreso = "MEMO"
        else
           Regreso = "TEXT(" & Maximo & ")"
        end if
     end if
  end if
  if Tipo = 2 then
     if BD = "MDB" then
        Regreso = "SHORT"
     end if
     if BD = "MYSQL" then
        Regreso = "SMALLINT"
     end if
  end if
  if Tipo = 3 then
     Regreso = "INTEGER"
  end if
  if Tipo = 4 then
     Regreso = "INTEGER"
  end if
  if Tipo = 5 then
     Regreso = "DOUBLE"
  end if
  if Tipo = 11 then
     if BD="MDB" then
        Regreso = "BIT"
     end if
     if BD="MYSQL" then
        Regreso = "TINYINT"
     end if
  end if
  if Tipo = 135 then
     if BD="MYSQL" then
          Regreso = "DATETIME"
     end if
     if BD="MDB" then
          Regreso = "DATETIME"
     end if
  end if
 END IF

 'SI ES MYSQL
 if BD="MYSQL" then
              strCadena = UCASE(TRIM(Tipo))
    FOR a=1 TO LEN(Tipo)
       sim=mid(Tipo,a,1)
       if sim="(" then
          strCadena=UCASE(mid(Tipo,1,a-1))
       end if
    next
    IF strCadena="VARCHAR" THEN
            Regreso="TEXT(" & Maximo & ")"
```

```
      END IF
      IF strCadena="LONGTEXT" THEN
            Regreso="MEMO"
      END IF
      IF strCadena="DOUBLE" THEN
            Regreso="DOUBLE"
      END IF
      IF strCadena="INTEGER" THEN
            Regreso="INTEGER"
      END IF
      IF strCadena="INT" THEN
            Regreso="INTEGER"
      END IF
      IF strCadena="TINYINT" THEN
            Regreso="BIT"
      END IF
      IF strCadena="DATETIME" THEN
            Regreso="DATETIME"
      END IF
      IF strCadena="TIME" THEN
            Regreso="DATETIME"
      END IF
      IF strCadena="SMALLINT" THEN
            Regreso="SMALLINT"
      END IF
     end if


    TransType = Regreso
  End Function

'-------------------------------------------------------------------
'----------INICIA ----------------------------------------------------
%>
<form name="form1" method="post" action=" Procesoetl.ASP">

<%

      BD2= arch
      EXISTEBASE=1
      dim fs
      IF EXISTEBASE=1 THEN

        response.write "<strong>POR FAVOR ESPERE ..... </strong><BR>"


        'strALMACEN = REQUEST.FORM("PAIS")
        CAMPOCOND  = ",,,,"
        strTablas  = "PRODUCTO,PROVEDOR,PAISES,MOVTOS,DOCTOS"
        arrTABLAS  = Split(strTablas,",",-1,1)
        arrCAMPOCOND = Split(CAMPOCOND,",",-1,1)

        FOR it=0 to Ubound(arrTABLAS)
          RESPONSE.WRITE ("ESPERE, COPIANDO TABLA "&UCASE(arrTABLAS(it))&"<BR>")
          strDELEsql ="DELETE FROM "&arrTABLAS(it)
          Conn2.EXECUTE ( strDELEsql )
          IF TRIM(arrCAMPOCOND(it))<>"" THEN
             arrCAMPOCOND(it)=" WHERE " & arrCAMPOCOND(it)&"="&strALMACEN&""
          END IF
          SColF                                                                                                    =
ChecarColumnas(OCONN,arrTABLAS(it),"MYSQL",Conn2,arrTABLAS(it),"MYSQL",strALMACEN,arrCAMPOCOND(it))
          RESPONSE.WRITE ("TABLA "&UCASE(arrTABLAS(it))&" COPIADA <BR><BR>")
          Response.flush
        next
```

```
        'Conn2.EXECUTE ( strInicializasql )



        response.write "<strong>PROCESO TERMINADA</strong><br>"



       '--CIERRA LA BVASE PRINCIPAL
       Conn2.CLOSE
       OCONN.CLOSE
       SET RS = NOTHING
       SET OCONN = NOTHING
       SET Conn2 = NOTHING
     ELSE
       response.write "<strong>Base de datos no localizada.....</strong><br>"
       response.write "<strong>Verifique que la base de datos esté previamente generada por el paso 1 y el paso 2.</strong><br>"
     END IF

     'RESPONSE.WRITE("<A HREF=GENERABASE.ASP>.....REGRESAR.....</A><BR><BR>")

%>

</body>
</html>
```

Ya integrada la información en el deposito de datos, el siguiente paso es generar las consultas para obtener los cubos.

El servidor OLAP que se utilizo para el prototipo no es un servidor OLAP como tal, es una herramienta de fácil adquisición y de forma muy sencilla de utilizar, esta herramienta es de Microsoft y se llama Excel. Por medio de la cual se generará las consultas y creación de los cubos. Es importante advertir al alumno que la herramienta OLAP, para este caso, no es una herramienta que soporte demasiada información por lo cual se generarán cubos muy resumidos y de poca información.

Para el uso de esta herramienta es importante que se realice una conexión ODBC.

La forma de realizarse es la siguiente. Busque en Panel de control el ícono Herramientas Administrativas.

Figura 3.1. 3
Icono Herramientas Administrativas

En el ícono orígenes de datos ODBC:

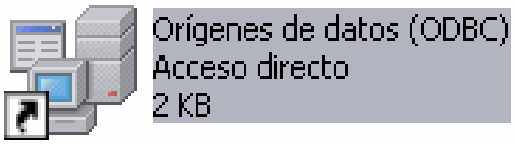

Figura 3.1. 4
Icono Herramientas Administrativas

Localice el controlador correspondiente a la base de datos utilizada.

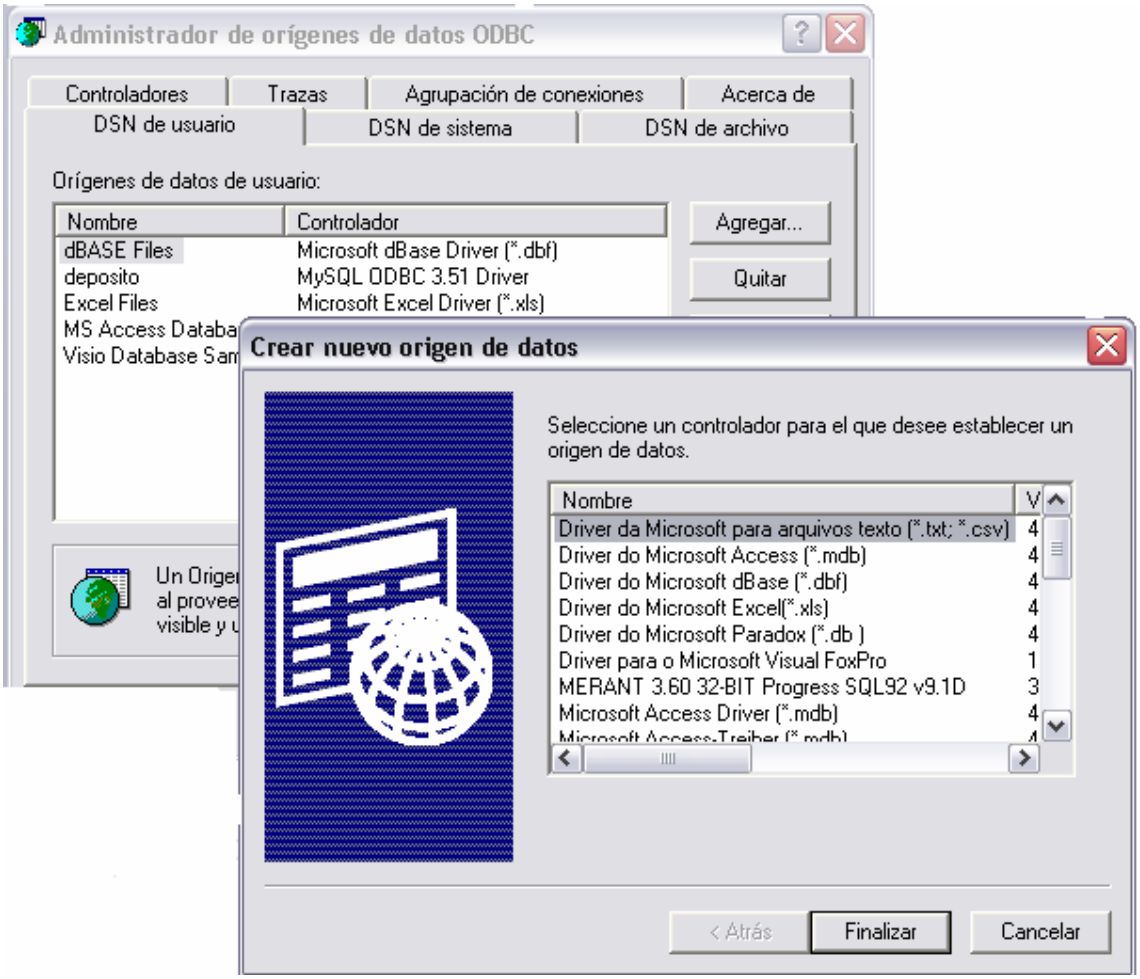

Figura 3.1.21

Localización del controlador para el manejador de base de datos

Ya localizado el Driver realizamos los pasos solicitados para la conexión por ODBC a la base de datos que funge como depósito de datos.

Ya obtenida la conexión, abrimos una hoja de Excel y en el menú Datos seleccionamos Obtener Datos Externos , y Nueva Consulta de Base de Datos.

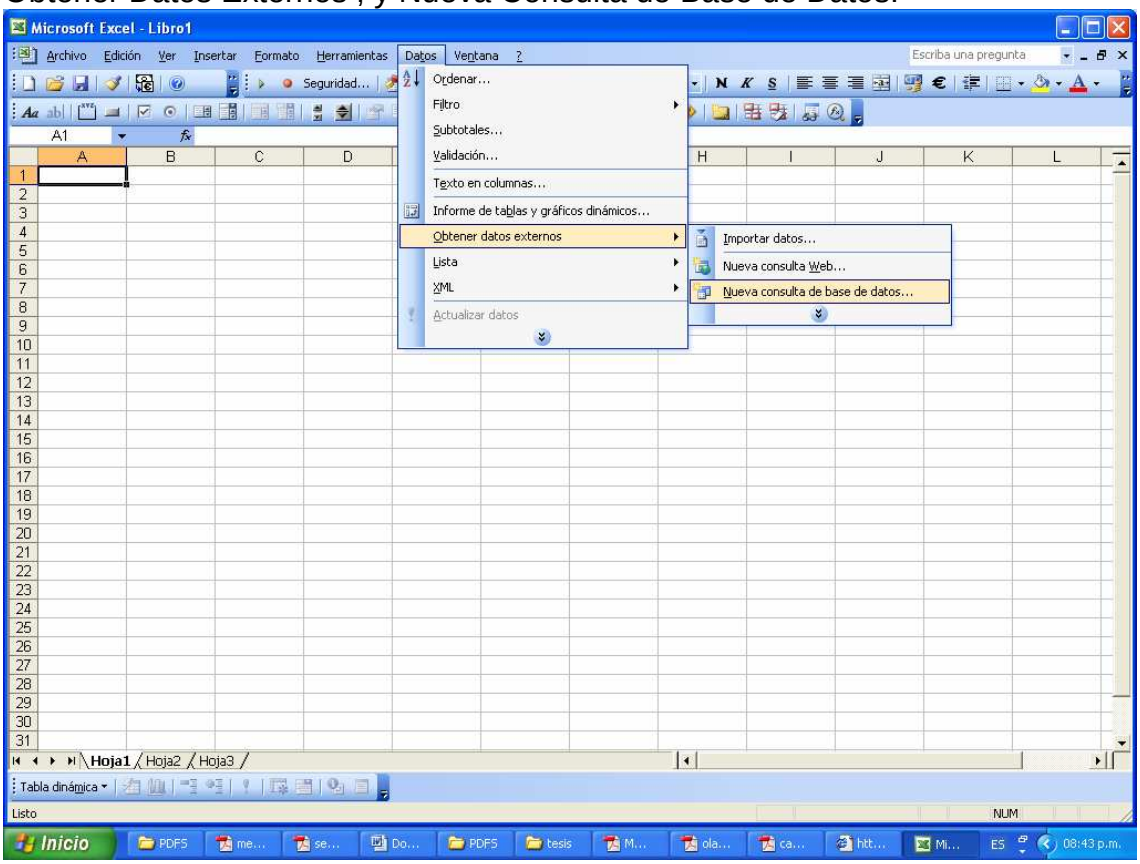

Figura 3.1. 5
Menú Datos

La siguiente ventana nos permitirá consultar cubos ya generados o bien generar nuevos cubos de consulta a la base Depósito.

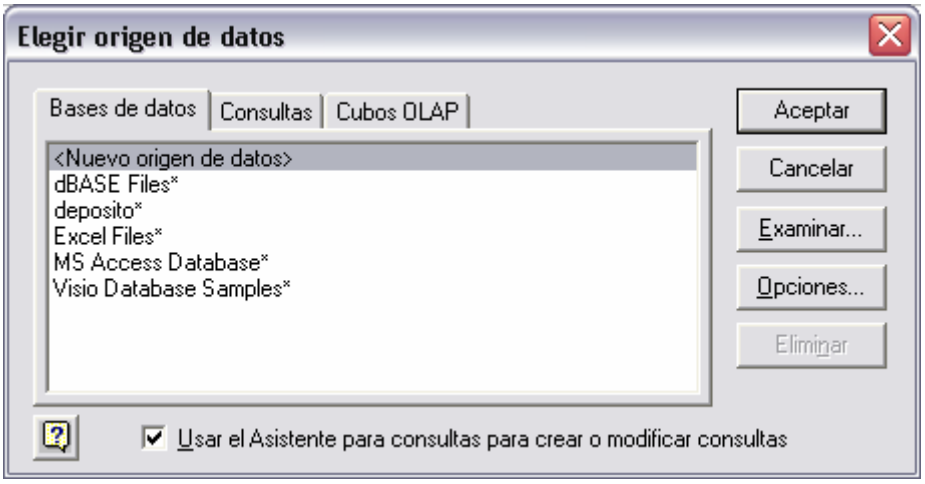

Figura 3.1. 6
Creación o Consulta de cubos

Localizada la base Depósitos, seguimos el asistente para seleccionar las tablas o campos que nos permitirá generar las columnas de nuestro cubo.

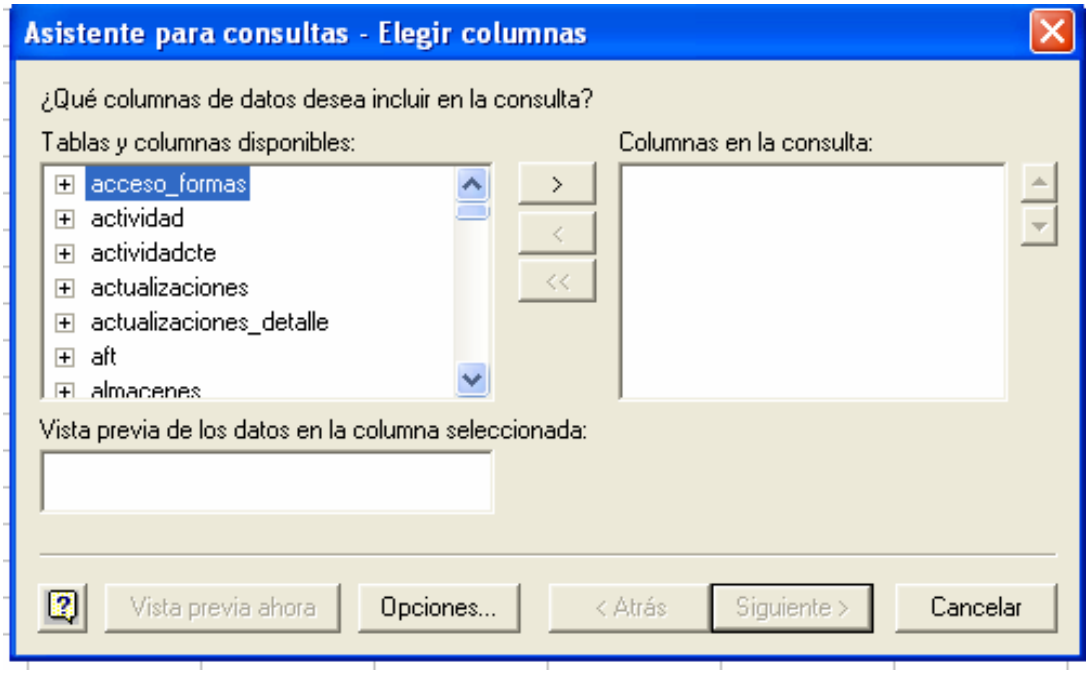

Figura 3.1. 7
Asistente para Creación de cubos

Siga el asistente, intente realizar primero la consulta con una sola tabla asignando filtros de consulta, para evitar que se filtre demasiada información o información inútil, posteriormente modifique sus consulta agregando mas tablas hasta formar cubo deseado. Procure no realizar una consulta donde implique demasiada información, recuerde que es una herramienta que no soporta demasiada información podría saturar la memoria requerida para la consulta.

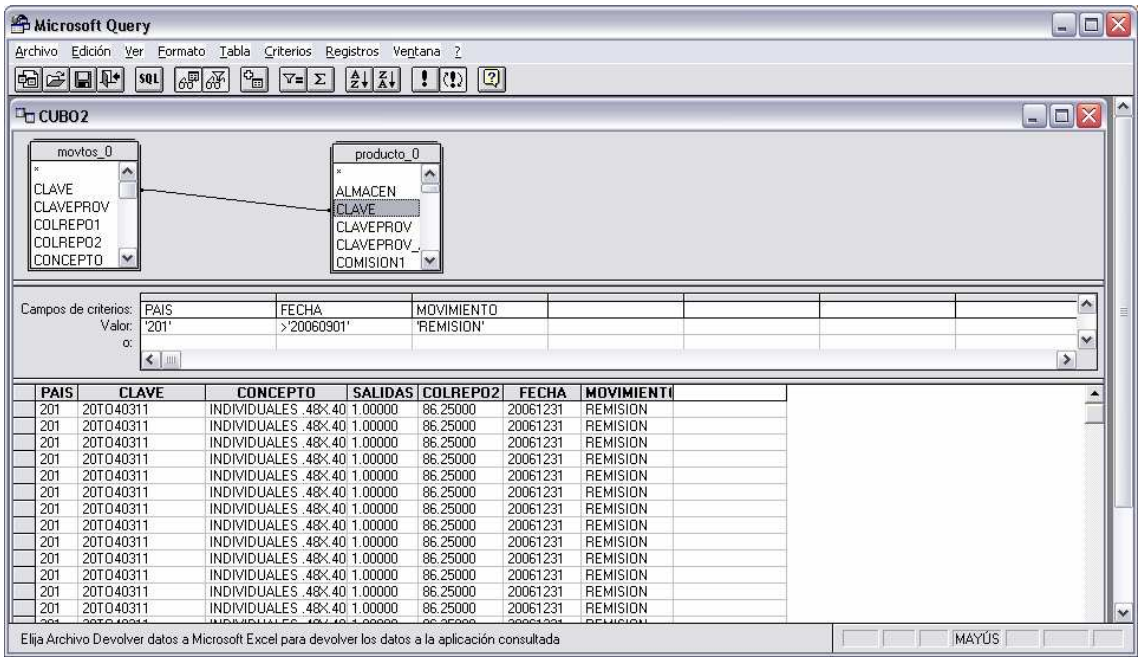

Figura 3.1. 8
Agregando mas tablas para la creación del cubo de consulta

Finalmente las consultas creadas mostrará el mismo principio para un deposito de datos de mayor tamaño, pero es necesario utilizar herramientas de igual potencia.

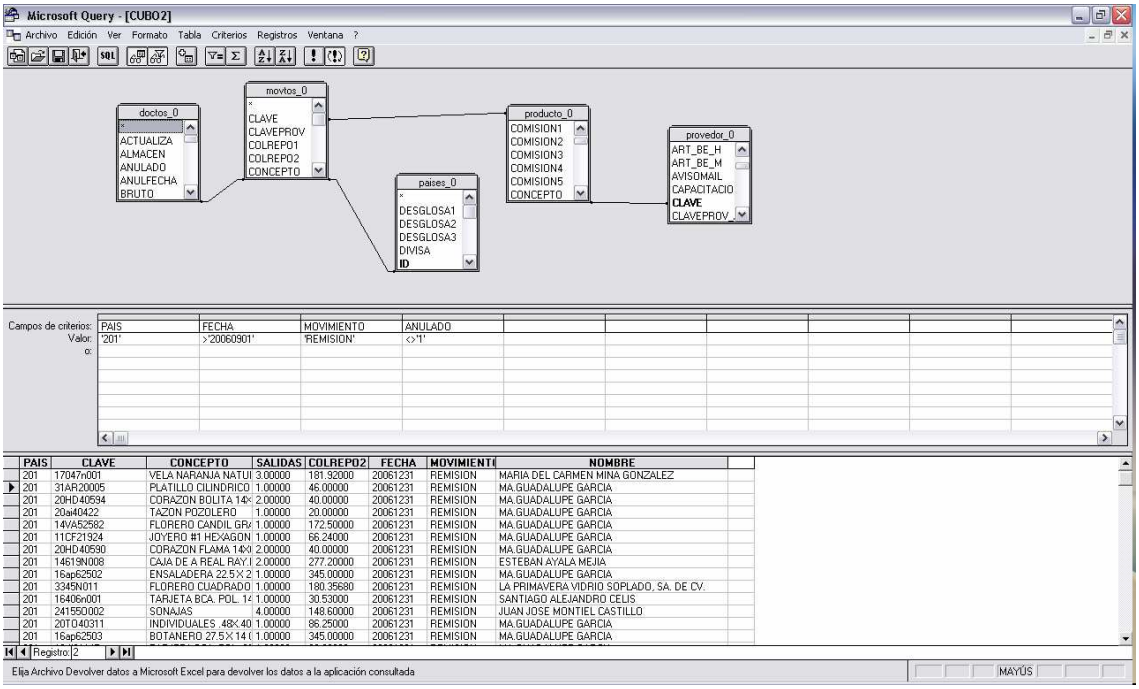

Figura 3.1. 9
Ejemplo de consulta

Finalmente podrá generar las consultas solicitadas. Arrastrando los campos requeridos ya sea como parte de la columna o parte de los costados. Analizando las salidas por venta, que son las Remisiones y las Facturas por cada columna que nos indica las claves de los productos, observando que se muestra del país 201, que nos indica la sucursal de esta forma podemos comenzar a buscar en las columnas que claves de producto son las que más se han vendido.

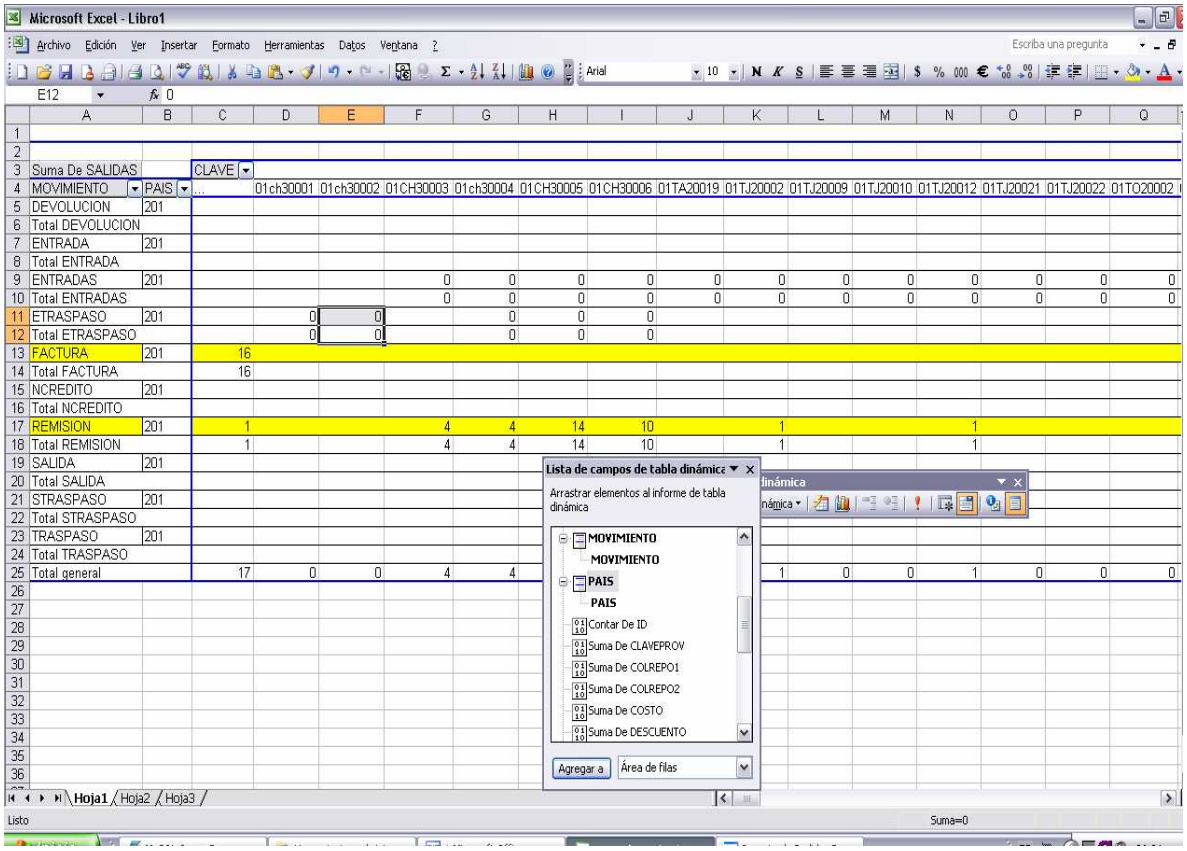

Figura 3.1. 10
Generación del Cubo de consulta por Movimiento, País y Artículo

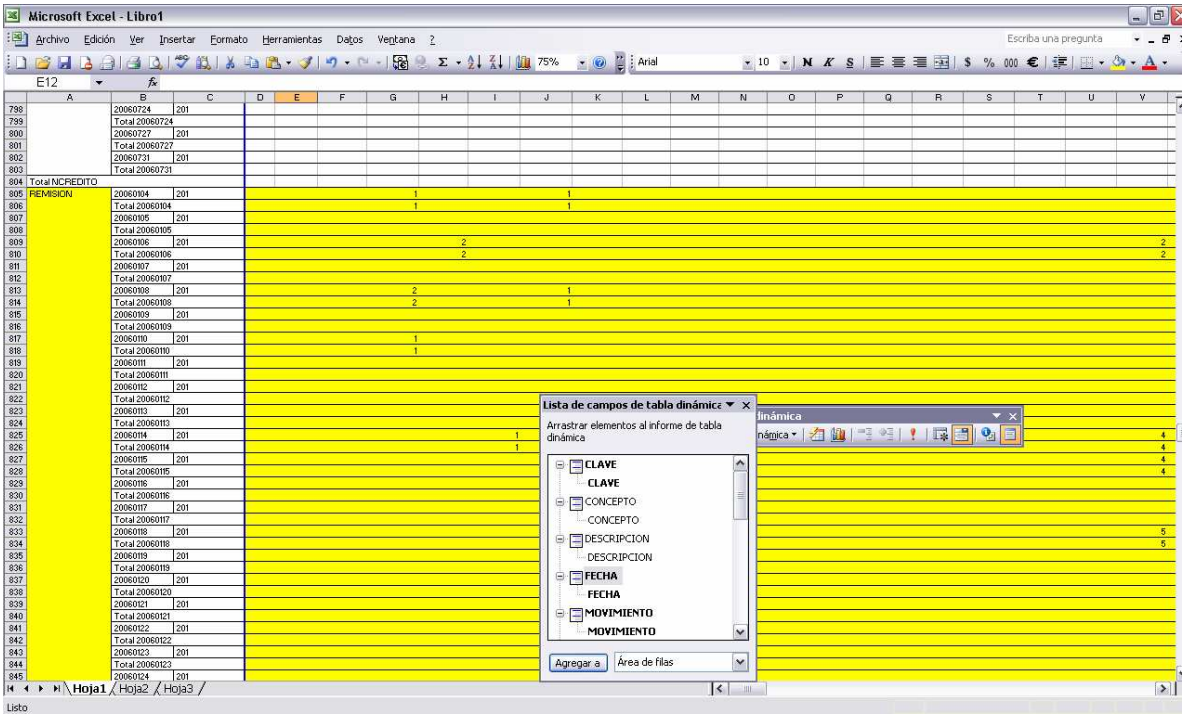

Figura 3.1. 11
Generación del Cubo de consulta por Movimiento, País, Articulo y Fecha.

Existen otras herramientas que nos facilitan la consulta y generación de cubos, especialmente para aquellos depósitos de gran información, algunas de estas herramientas son SQL Server, HamigoV01c9, etc..

Usando l Herramienta HamigoV019C9.

Se muestra una herramienta comercial de tipo ROLAP (Relational On Line Analytical Processing). Se eligió este software debido a las siguientes características:

- Cualquier base de datos relacional de la organización puede ser utilizada como un almacén de datos y el motor ROLAP proporciona la funcionalidad analítica.

- A nivel de aplicación, es el motor de la herramienta la que ejecuta las consultas multidimensionales de la información.

- Los usuarios finales ejecutan sus análisis multidimensionales, a través del motor ROLAP, que transforma dinámicamente sus peticiones a consultas que se ejecutan en las bases de datos relaciónales. Los resultados se relacionan mediante tablas cruzadas y conjuntos multidimensionales para devolver los resultados a los usuarios.

- La arquitectura ROLAP es capaz de usar datos precalculados si estos están disponibles, o de generar dinámicamente los resultados desde los datos de las bases operacionales si es preciso.

- Los resultados de las consultas realizadas por los usuarios finales a través de la herramienta ROLAP pueden ser trabajados usando sus herramientas favoritas propias con el fin de facilitar la interpretación de los resultados.

Esta herramienta es de evaluación y tiene como principal propósito mostrar la funcionalidad y ventajas de las herramientas ROLAP. Este software esta desarrollado en base al 4GL de Progress ®

Esta herramienta necesita los siguientes recursos como mínimo para poder operar:

Requerimientos

- Hardware Básico — Servidor: Pentium IV, 2.4 Ghz, 80 GB
PC´s: Pentium III en adelante
- Sistema Operativo DB — Windows, Unix, Linux
- Sistema Operativo Aplicación — Windows 98 en adelante.

A continuación se dará una breve explicación de esta herramienta y su funcionamiento, con el fin de dar una visión general del software y una introducción de la forma en la que se pueden generar consultas para la explotación y análisis de la información la cual es fundamental para la toma de decisiones.

Esta herramienta esta dividida en dos grandes módulos:

Modulo de Administración de Usuarios.

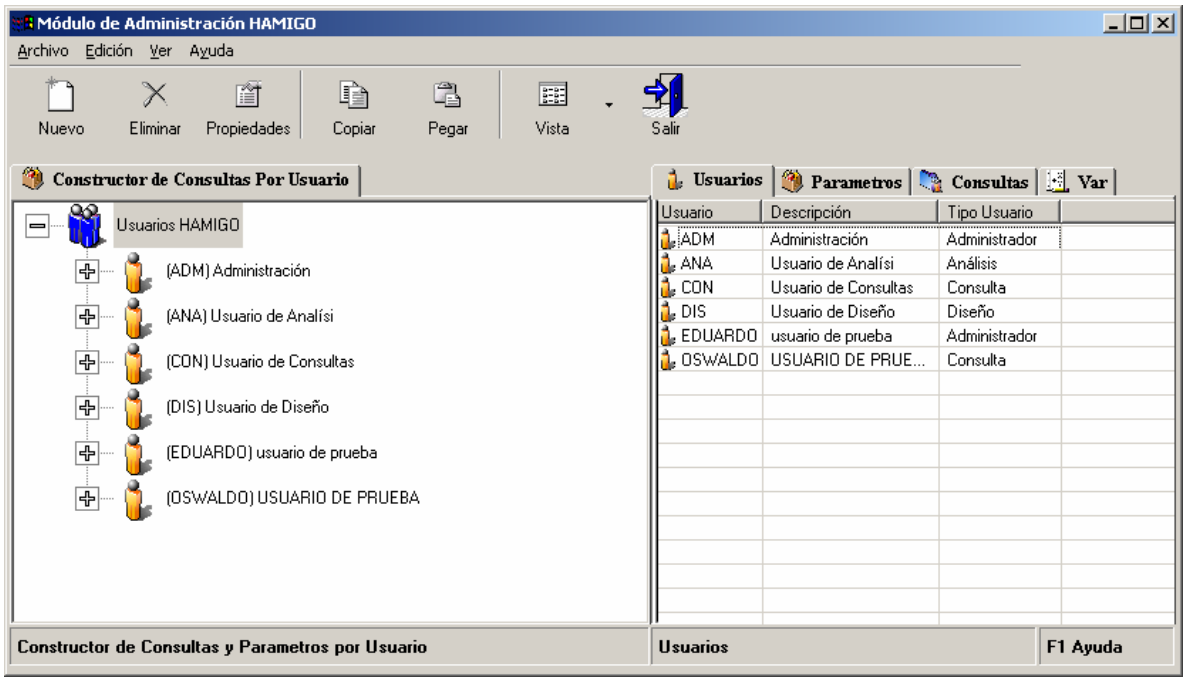

Figura 3.1. 12
Módulo de Administración de usuarios

Los usuarios forman parte del ambiente en el que funciona un deposito de datos, debido a que estos usuarios utilizan el deposito de datos de acuerdo a sus funciones y necesidades, estos se dividen en categorías de acuerdo a su jerarquía, su función o por su nivel de competencia en cómputo en la organización. Y es justamente en este modulo donde el administrador de esta herramienta debe analizar a los usuarios de la organización y establecer en que categoría debe ser incluido cada uno de estos.

La figura 3.1 muestra las diferentes categorías de usuario que se pueden registrar en esta herramienta.

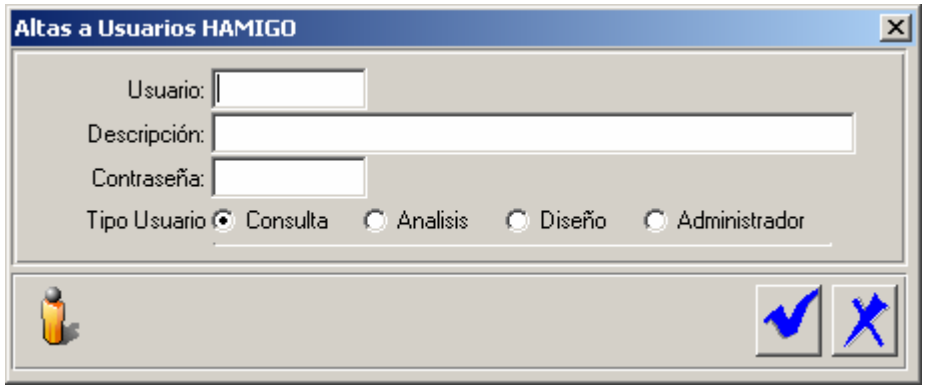

Figura 3.1. 13
Tipos de usuario registrados en el sistema

Dentro de la categoría de usuarios existen los siguientes perfiles:

- Usuario Administrador. Este usuario tiene todos los privilegios de la herramienta, es decir, podrá generar consultas, visualizar información, realizar mantenimiento a usuarios y a las consultas de información. Dentro de estos usuarios se encuentran por lo general el administrador de la herramienta, y personal asignado a generar cubos de información dependiendo de las necesidades de directores generales, ejecutivos de primer nivel como el director financiero, contable y fiscal.

- Usuario de Diseño. Este usuario esta habilitado para generar las consultas y visualizar la información, Sin embargo una vez generada la consulta este tipo de usuario no puede modificarla. Esto es con el fin de evitar que el usuario modifique de manera voluntario o involuntaria la consulta que otros usuarios también utilizan. Este tipo de usuario es asignado a personal que utiliza información en común con otros usuarios como por ejemplo usuarios que se dedican al telemarketing y la venta de servicios.

- Usuario de Análisis. Este usuario al igual que el usuario de diseño podrá generar consultas y visualizar la información la diferencia con el usuario de diseño es que podrá modificar los criterios de dichas consultas de acuerdo a sus necesidades, las cuales pueden cambiar dependiendo de la necesidad de incluir información en el cubo de consultas. Dentro de estos usuarios se encuentran por lo general personal operativo de las áreas administrativas, contables, fiscales, comercialización y ventas, producción e ingeniería.

- Usuario de Consulta. Este usuario solo podrá visualizar la consulta generada por el usuario de diseño. Dentro de estos usuarios se encuentran por lo general personal administrativo y de apoyo.

Es importante comentar que esta herramienta tiene una interfaz grafica amigable al usuario por lo que es muy intuitiva en su manejo y en la generación de cubos de información, sin embargo se recomienda que en un inicio se tenga un equipo de trabajo el cual se encargue de la recolección de las necesidades de información de los tomadores de decisiones.

Este grupo tendrá que obtener esta información a través de entrevistas, cuestionarios, informes y reportes los cuales servirán como base para el análisis y el diseño de la consulta, teniendo estos puntos definidos es sencillo el elaborar el cubo de consulta con estos datos.

Módulo de generación de consultas.

En este modulo es donde se generan los cubos de información de pendiendo del análisis y diseño previo de la consulta. Este módulo es extenso ya que es aquí donde se define el origen de los datos y como se van a presentar dependiendo de las necesidades de visualización de la información.

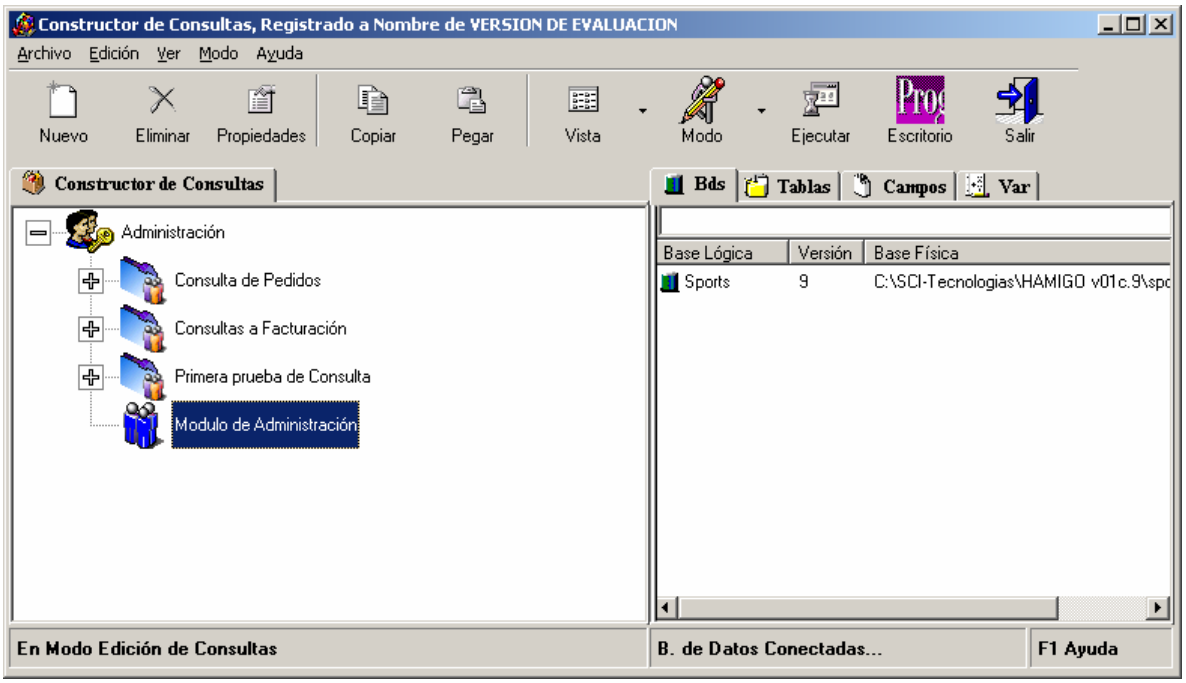

Figura 3.1. 14
Módulo de generación de consultas

A continuación se desarrollara un ejemplo de cómo generar una consulta y servirá para mostrar en forma práctica este modulo. Como se menciono en un inicio se recomienda tener

una persona o un grupo de personas dependiendo de la demanda para obtener información estratégica la cual será fundamental para tomar una decisión acertada en el ámbito laboral. Esta persona o grupo de personas debe de seguir un proceso el cual le ayude a recopilar información y obtener todos los detalles posibles del tipo de datos, vistas y hechos para que la consulta a generar este apegada a las necesidades de los usuarios ejecutores de decisiones.

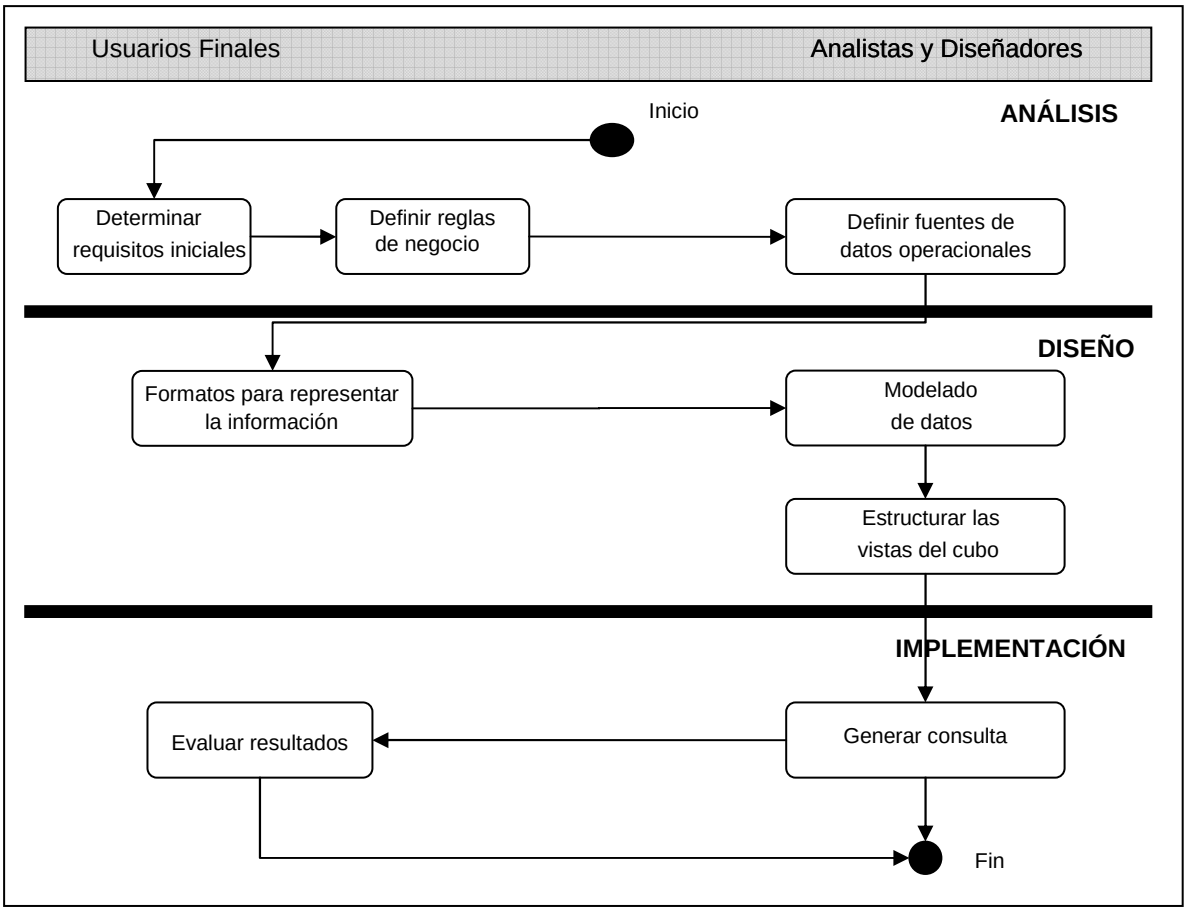

Figura 3.1. 15
Proceso para la generación de consultas

A continuación se describirá brevemente el proceso de la figura 3.1.15 después se presentaran los casos prácticos en base a este proceso los cuales también se implementaran en la herramienta HAMIGO. Estos procesos son desarrollados principalmente por el analista y el diseñador de consultas los cuales trabajan de forma estrecha con los usuarios involucrados.

- Determinar requisitos iniciales. El analista del almacén de datos debe recolectar toda la información disponible, desde reportes, informes así como entrevistas a los usuarios involucrados. El analista organiza toda esta información y la agrupa dependiendo de las necesidades de información que requiere el usuario.

- Definir reglas del negocio. El analista con ayuda del usuario deberá recopilar y asentar las reglas de negocio para complementar la información de la consulta a generar. Las reglas del negocio son muy importantes para que la consulta tenga sentido para el usuario que necesita información para una acertada toma de decisiones.

- Definir fuentes de datos operacionales. De acuerdo a las necesidades de los usuarios y los datos obtenidos por el analista, este y el diseñador organizaran y seleccionaran las bases de datos operacionales así como sus tablas para poder obtener la información necesaria que el usuario solicito.

- Formatos para representar la información. El analista con la ayuda del usuario, deberán definir uno o varios formatos los cuales muestren de forma clara la información que se solicito. Estos formatos pueden a futuro ser reportes y / o hojas de control.

- Modelado de datos. Para poder estructurar toda la información recolectada hasta este momento se va ha utilizar el modelo dimensional. Las dimensiones en este tipo de modelo son el criterio de análisis de los datos. Las medidas son valores o indicadores a analizar. Este modelo representa indicadores para estructurar cubos de información.

- Estructurar vistas del cubo de información. Una vez identificadas las dimensiones y medidas de las consultas a generar, el analista debe estructurar la forma de representar estas consultas el la herramienta HAMIGO.

- Generar consulta. Es en esta parte donde se aplica la estructuración de las vistas del cubo y es base a esto se genera la consulta.

- Evaluar resultados. Una vez que el diseñador termine de generar la consulta y realice las pruebas técnicas básicas de su funcionamiento. El usuario final deberá realizar las pruebas correspondientes a detalle para determinar si el resultado de dicha consulta satisface sus necesidades de información y ésta es de calidad y útil en la toma de dediciones.

A continuación en base a este proceso se mostraran los ejemplos de la generación de una consulta con la herramienta AMIGO.

El origen de los datos para los ejemplos que se van a mostrar en este capitulo son de una base de datos relacional, esta base de datos es de una empresa cuyo principal giro es la venta de artículos deportivos. Es importante aclarar que esta base solo tiene información del año 2006 y ciertas tablas las cuales son origen de los ejemplos presentados en este capitulo.

Determinar requisitos iniciales.

Se realizaron varias reuniones, la primera reunión fue con usuarios de mandos gerenciales estos usuarios solicitaron información variada principalmente sobre la cantidad de pedidos realizados, artículos solicitados, fecha de entrega prometida, clientes frecuentes así como sus limites de crédito.

Se realizaron cuatro reuniones más, las dos subsecuentes con personal operativo de las áreas involucradas, estos usuarios proporcionaron múltiples reportes los cuales proporcionaban parte de la información solicitada, también hojas de cálculo donde concentraban los datos de dichos reportes así como graficas de los datos concentrados.

Definir reglas del negocio.

En base a las reuniones, entrevistas con los usuarios, reportes y hojas de cálculo se organizo la información en diferentes rubros, estos grupos de información fueron presentados en la última reunión al personal operativo y gerencial de la organización. El analista con ayuda de los usuarios finales y en base a las reglas de negocio de la organización definieron una serie de preguntas las cuales se presentan a continuación:

a) Necesito saber la cantidad de los artículos que fueron pedidos el año pasado para asegurar que en este año no falten en mi inventario pero tampoco saturen el mismo.
b) Necesito saber en que mes del año los artículos tienen un menor volumen de venta para ofrecer en ese mes, un mayor descuento en ese producto e iniciar o fortalecer una campaña de publicidad.

c) Necesito saber cual es el cliente que realiza mayor cantidad de pedidos para aumentar su línea de crédito

Definir fuentes de datos operacionales.

En base a las preguntas anteriores, se procede a identificar la base de datos relacional así como sus tablas y campos para generar la consulta solicitada por los usuarios finales. En

este caso la base de datos relacional **Sports** es la que tiene la información necesaria para poder obtener respuestas a las consultas planteadas por el usuario final.

Formatos para representar la información.

Los usuarios finales solicitaron varios tipos de reportes donde piden agrupar la información solicitada, este tipo de reportes no son problema debido a que la herramienta ROLAP con la que se va ha generar las consultas tiene varias formas de representar los resultados de las mismas, estos resultados pueden ser mostrados en diferentes tipos de graficas, reportes así como también la salida a diferentes aplicaciones de Windows.

Modelado de datos.

A continuación se identifican las dimensiones y medidas de las necesidades que fueron planteadas por los usuarios y que el analista debe identificar cuando se defina las tablas que van ha ser el origen de las consultas.

Estas dimensiones son las que dictarán la organización de las caras del cubo de las consultas a generar, las dimensiones representan la información solicitada por los ejecutivos de la organización, el análisis de las preguntas planteadas por el usuario final nos da como resultado las siguientes dimensiones en la información:

- Cliente. Esta dimensión proporciona información general de los clientes así como su línea de crédito, sus descuentos y su forma de pago.

- Articulo. Esta dimensión proporciona datos generales de los artículos, cuantos han sido solicitados, su precio y la cantidad de existencias de estos en el almacén al momento de realizar un pedido.

- Vendedor. Esta dimensión proporciona información de los datos generales del vendedor así como la región geográfica en el que éste esta ubicado.

La relación dimensional de este modelo es:

- Pedidos. Esta relación dimensional agrupa la cantidad de artículos que fueron solicitados en un pedido en un periodo de tiempo así como los clientes que los solicitaron.

La figura 3.4 muestra de manera grafica la relación dimensional de las consultas solicitadas por los usuarios finales.

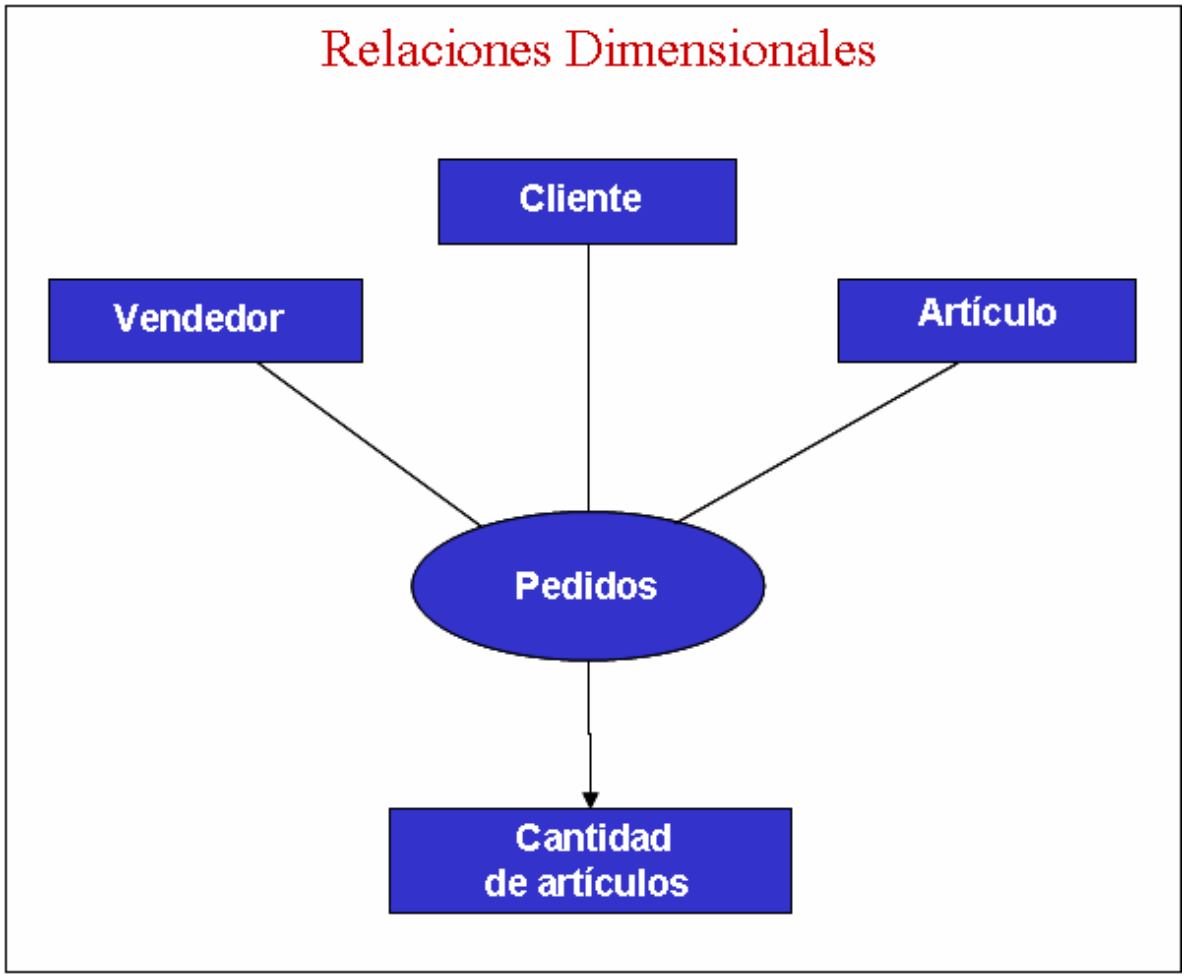

Figura 3.1. 16
Relación dimensional de la consulta

Estructurar vistas del cubo de información.

Una vez identificadas las entidades y medidas de las consultas planteadas en un inicio por los usuarios finales, se procede a desglosar las dimensiones y medidas ya que estas serán las caras del cubo de información las cuales podrán rotar, expandir, contraer y seccionar la información solicitada por los usuarios, la cual ayudara a tener información especifica en diferentes vistas, reportes y graficas.

Toda esta información ayudara al usuario final a tomar una decisión acertada en un corto tiempo.

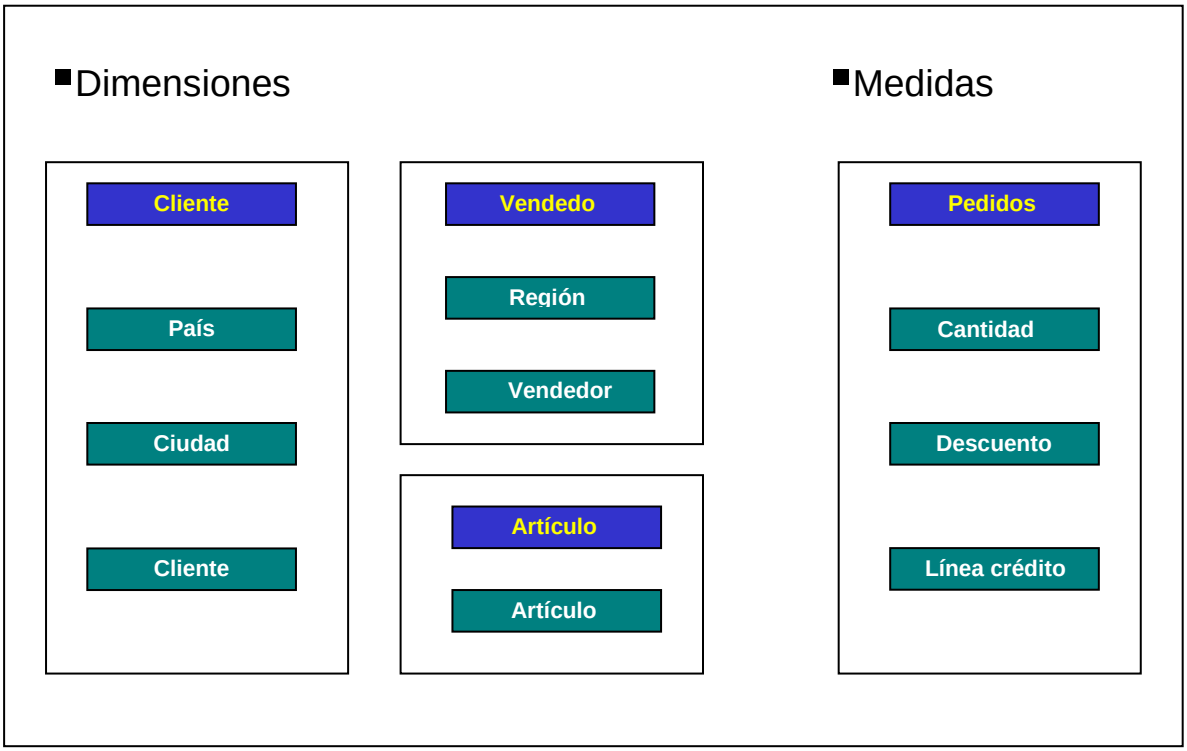

Figura 3.1. 17
Dimensiones y medidas del cubo de información

Generar consulta.

Una vez que se han realizado los pasos anteriores el analista y el diseñador deben de crear la consulta en la herramienta ROLAP. Esto se describe a continuación:

1.- De inicio se debe crear una consulta, esto se hace en el modulo de constructor de consultas colocando el puntero del ratón en el icono de usuario administrador dando clic derecho. Aparece un menú en el cual hay que seleccionar **Nueva Consulta.**

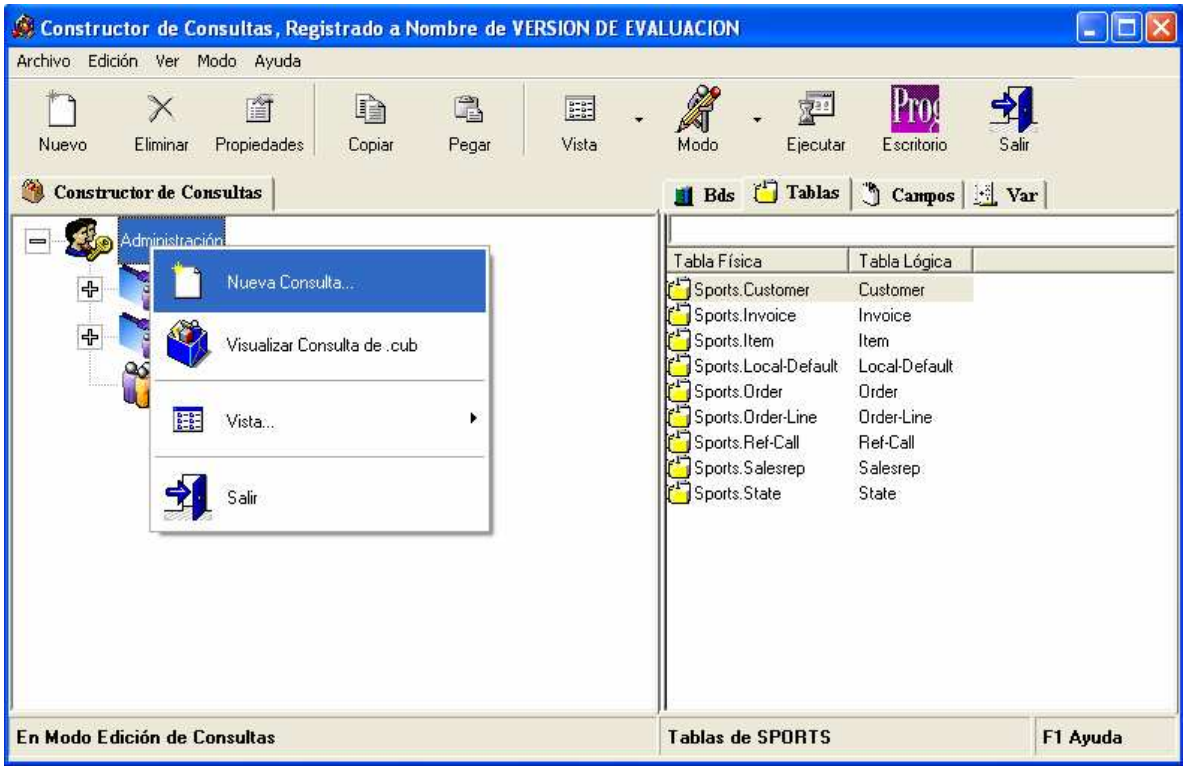

Figura 3.1. 18
Forma de generar una nueva consulta

Proporcionar un identificador y un nombre que describa la consulta que se esta generando.

2.- El siguiente paso es definir el origen de los datos para esta consulta, en este punto se procede a indicar cual es la base de datos relacional de donde se va a obtener la información. Para esto se selecciona la primera pestaña de la carpeta de lado derecho se coloca el puntero del ratón en la sección de esa pestaña y se da un clic derecho. Aparece un menú en el cual hay que seleccionar **Conectar Base de datos.**

En esta sección se puede conectar varias bases de datos relacionales dependiendo de la complejidad de la consulta, para este ejemplo solo utilizaremos la base de datos **Sports**.

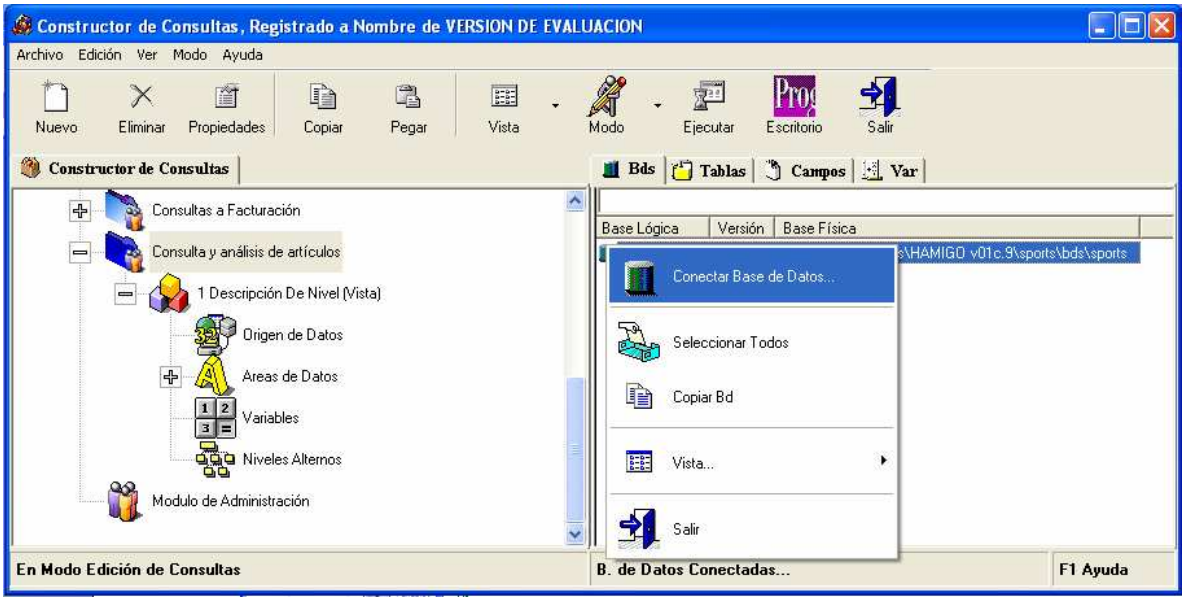

Figura 3.1. 19
Forma de Conectar una base de datos relacional

3.- Una vez conectada la base de datos solo hay que seleccionar el icono que representa la base de datos relaciona con el ratón y “arrastrar” hasta el icono **Origen de Datos** de lado izquierdo.

4.- Ahora se tiene que definir en base a las dimensiones y medidas de los datos las caras del cubo que se requiere generar. Para esto se selecciona la segunda pestaña de la carpeta de lado derecho **Tablas**, en esta pestaña aparecen todas las tablas de la base de datos relacional.

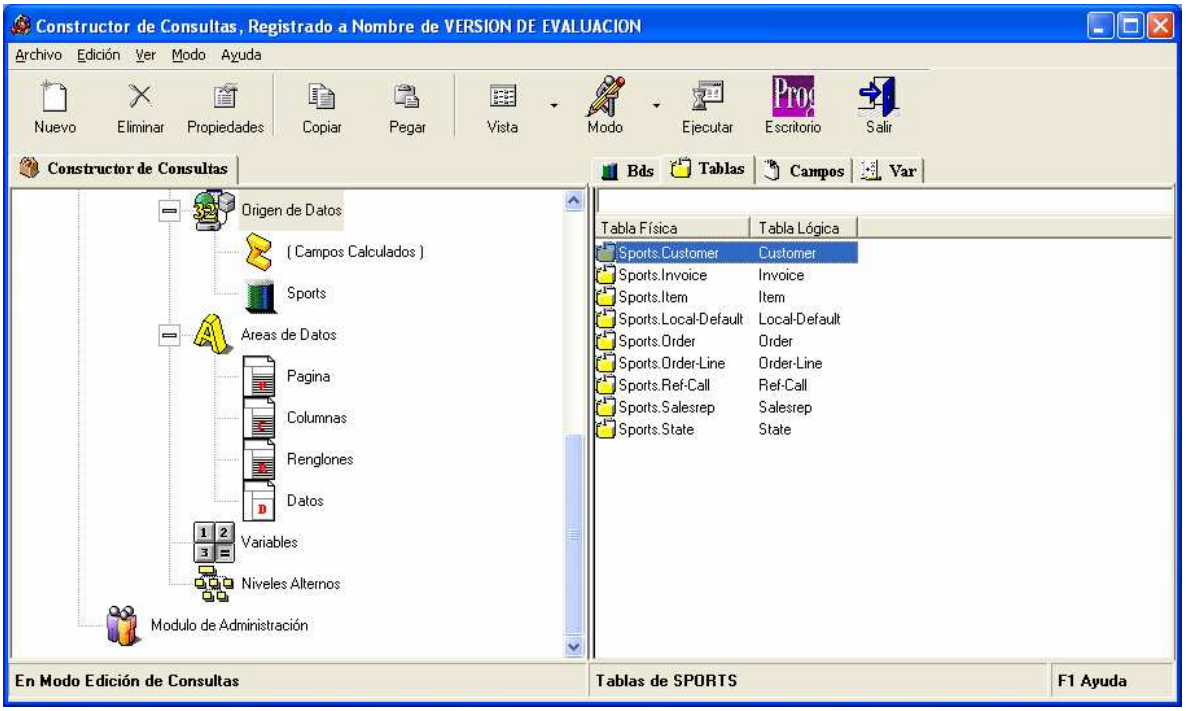

Figura 3.1. 20
Tablas contenidas en la base de datos relacional

Se debe seleccionar la tabla que tenga los campos requeridos para la consulta, en el ejemplo que estamos manejando debemos seleccionar la tabla de Clientes (customer), Artículos (Item), Pedidos (Order) y su detalle de Pedidos (Order-Line).

Para indicar que estas tablas van a formar parte de la consulta se deben “arrastrar” al icono de base de datos (Sports) que se encuentra de lado izquierdo.

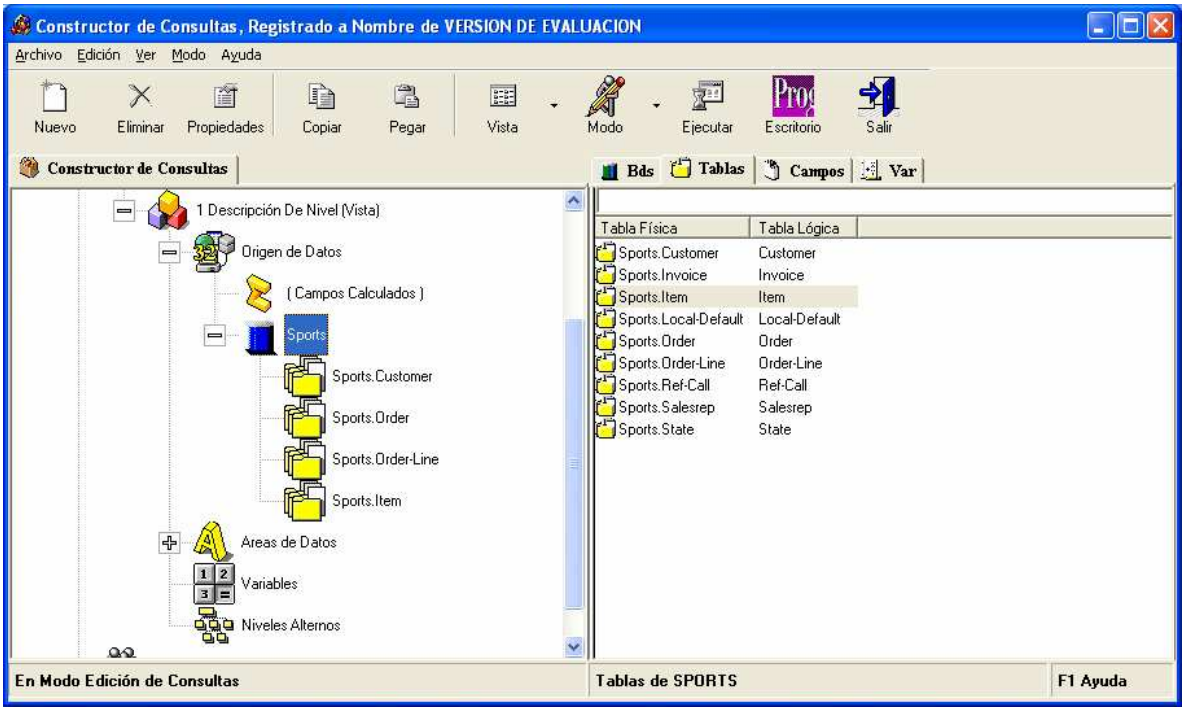

Figura 3.1. 21
Tablas necesarias para generar la consulta

5.- Agrupar los campos de las tablas de acuerdo a la consulta en el **Área de Datos**. De lado izquierdo. Para esto se selecciona la tercera pestaña de la carpeta de lado derecho **Campos**, en esta pestaña se muestran los campos de la tabla seleccionada. Por ejemplo si la tabla seleccionada en la pestaña de **Tabla**, es la de Clientes (customer) los datos que se muestran en la tercera pestaña serán los campos de la tabla de clientes.

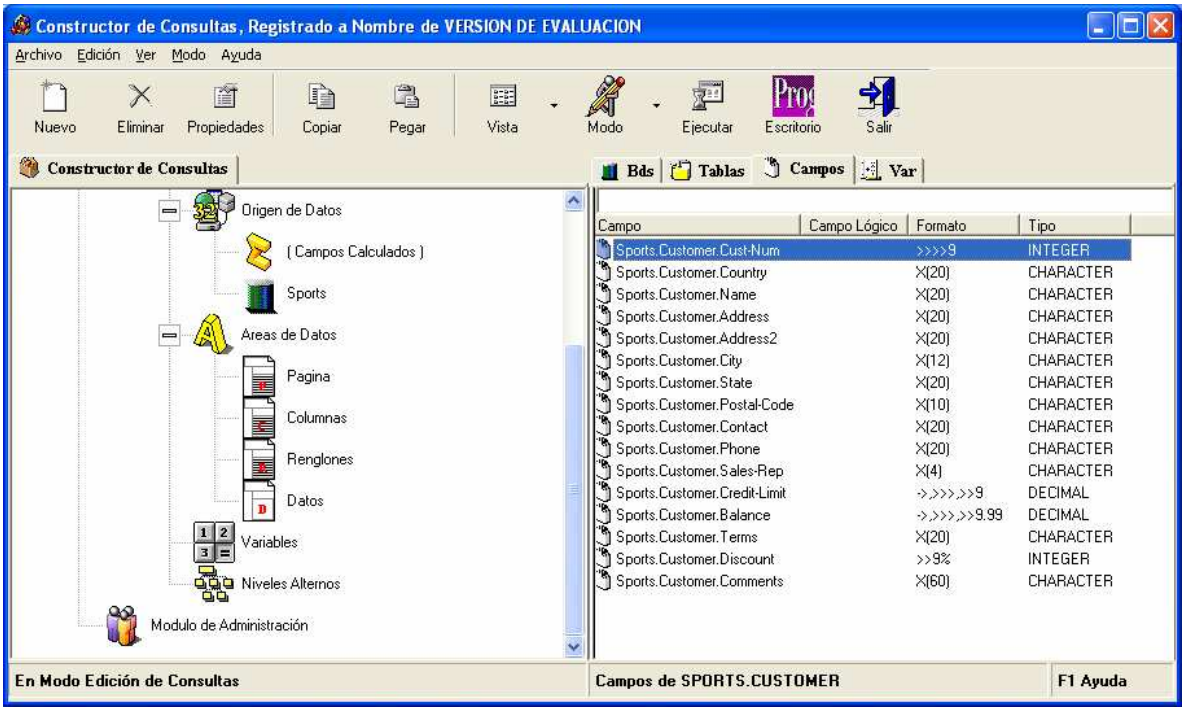

Figura 3.1. 22
Campos de la tabla Clientes

Estos campos son los que se deben agrupar en el **Área de Datos** de lado Izquierdo. La forma de agrupar es la misma que se ha estado utilizando y consiste en seleccionar el campo indicado y “arrastrar” este campo hacia los iconos de Página, Columnas, Renglones y Datos. Los iconos de Página Columna y Renglones representan las dimensiones del cubo de información y el icono de Datos representa las medidas del cubo.

Se “arrastran” los campos de acuerdo al análisis que anteriormente se hizo de las dimensiones y medidas de la consulta a generar. Por lo que estos campos se agrupan en cada Área de Datos que le corresponde.

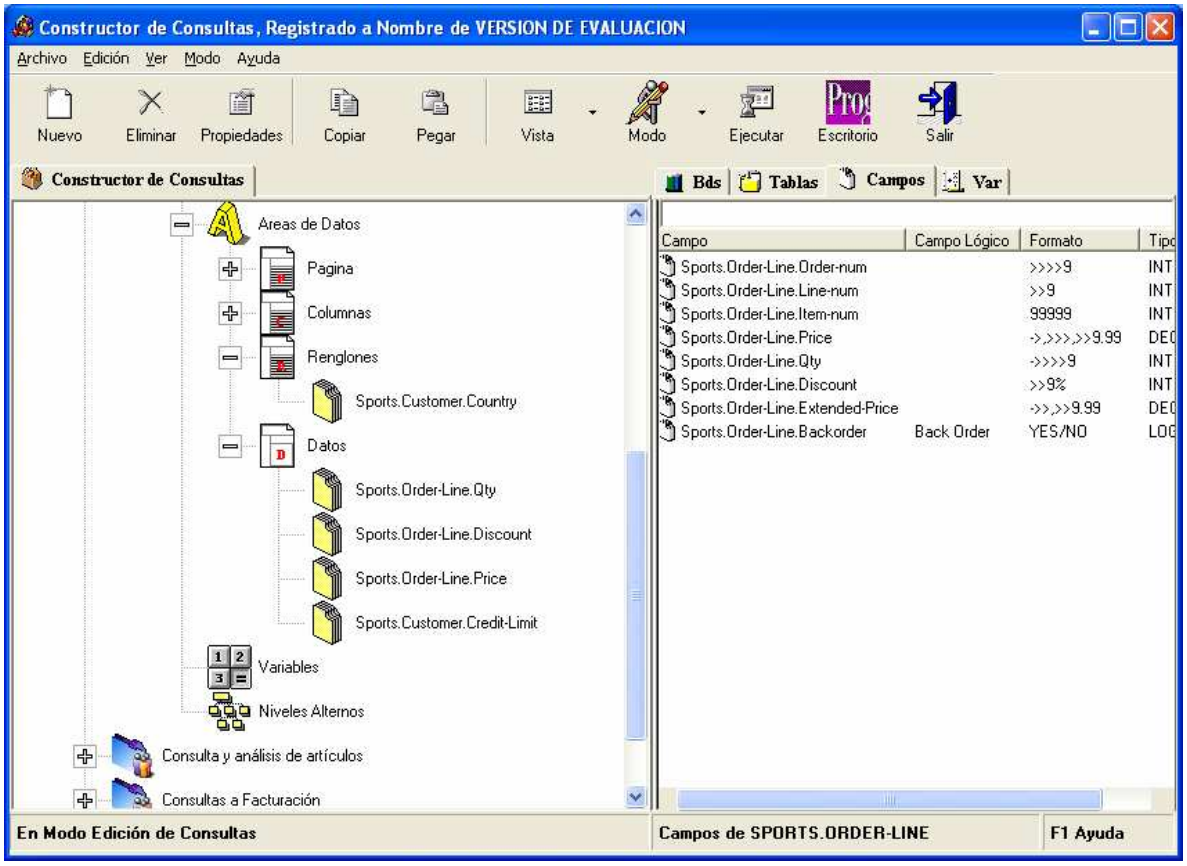

Figura 3.1. 23
Campos de las tablas en su correspondiente Área de Datos

6.- Una vez que las caras del cubo de información se encuentran definidas se termina con la construcción de la consulta y se procede a ejecutar la consulta. Para ejecutar la consulta que construimos se debe posicionar en el icono de la consulta a ejecutar dar clic derecho y seleccionar **Ejecutar Consulta**.

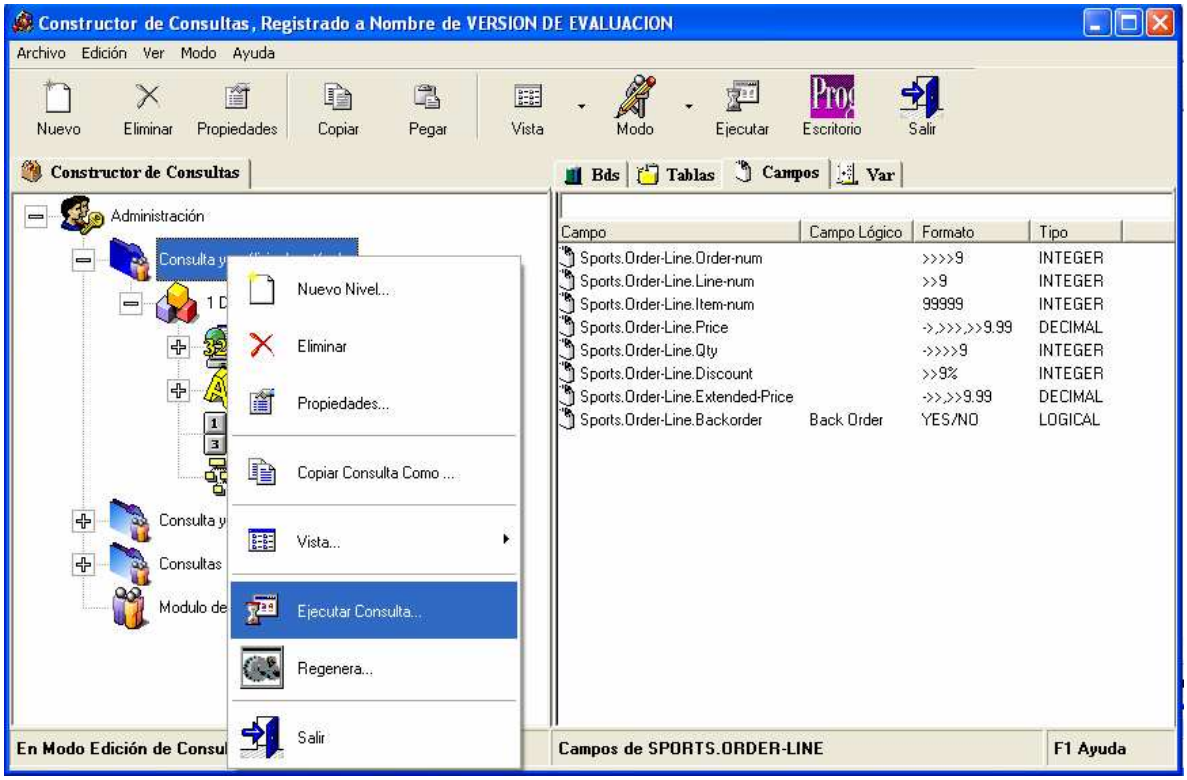

Figura 3.1. 24
Ejecutar consulta

La siguiente figura muestra el resultado de la consulta, y es en esta pantalla donde se pueden rotar las caras del cubo para ver el resultado de la consulta en diferentes perspectivas, además de que esta herramienta ofrece la posibilidad de graficar este resultado. Este resultado de la consulta se puede exportar a Excel o alguna otra herramienta de Microsoft.

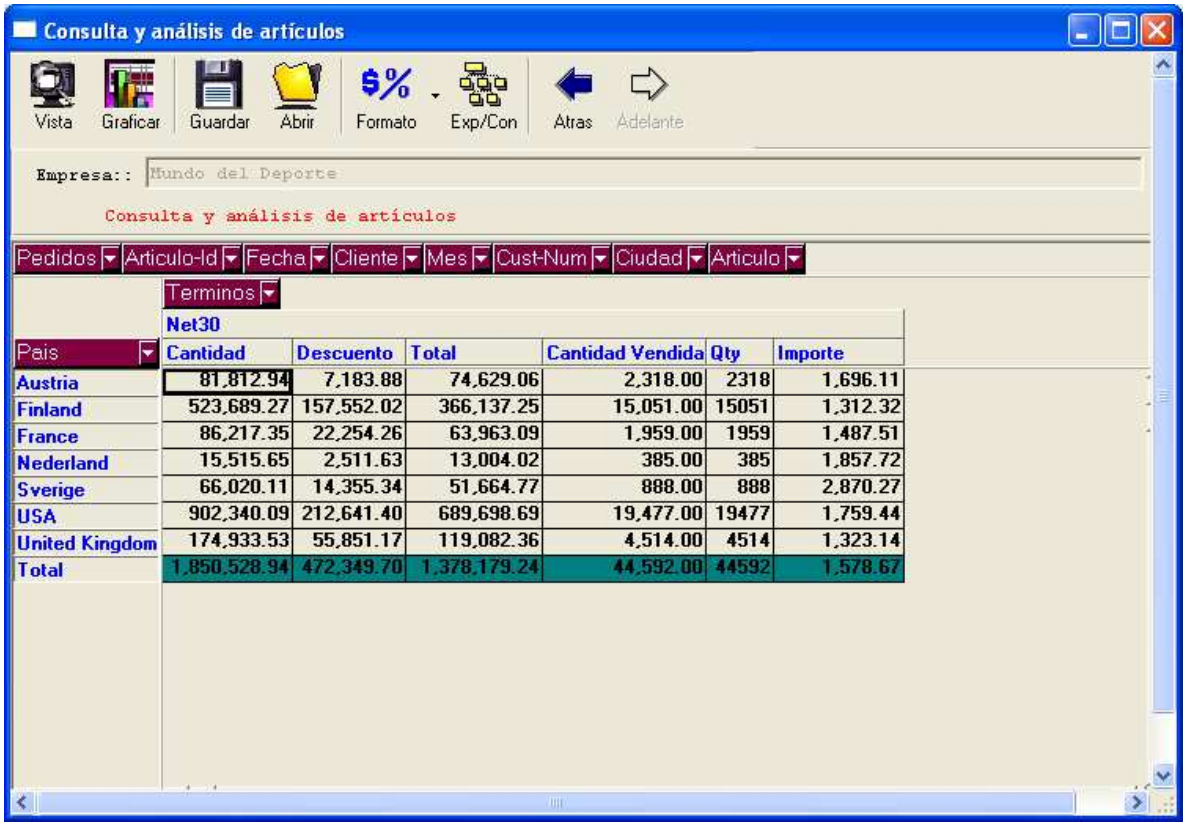

Figura 3.1. 25
Resultado de la consulta generada

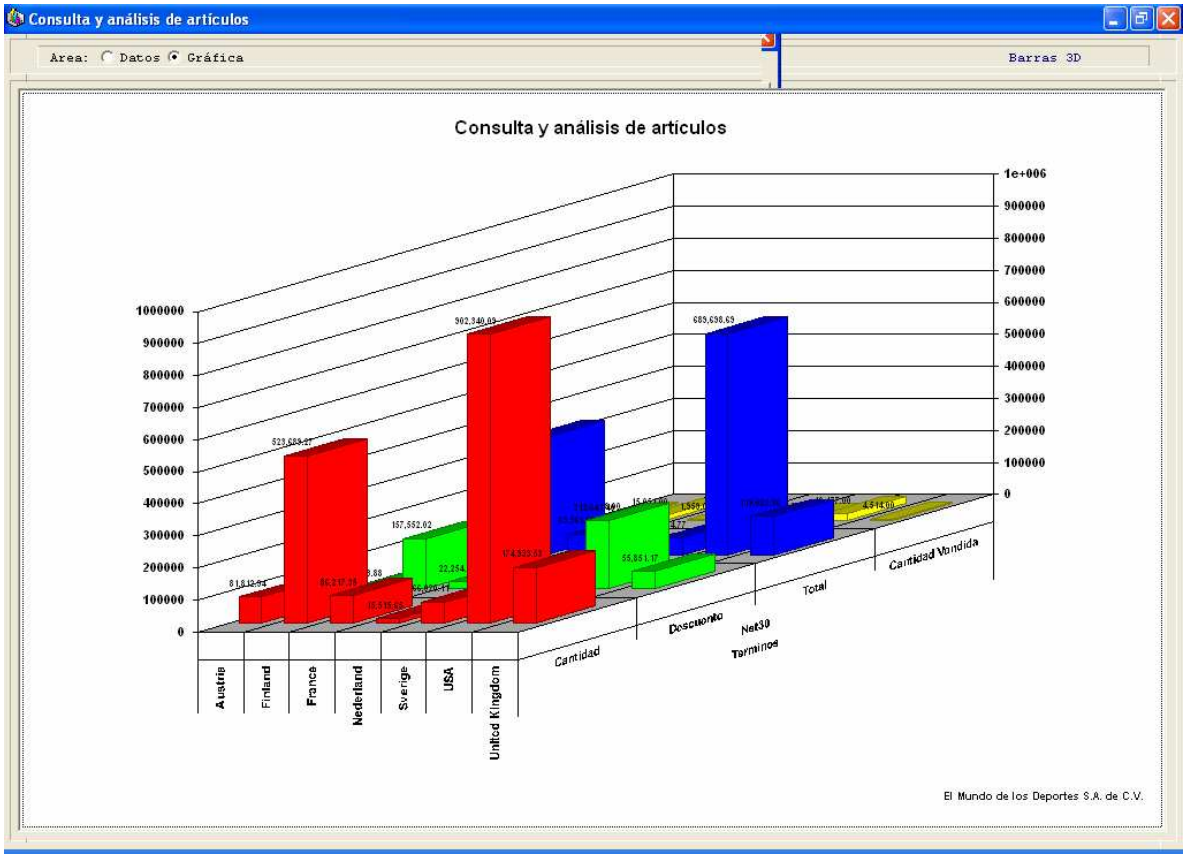

Figura 3.1. 26
Resultado en forma gráfica de la consulta generada

Existen varios beneficios al utilizar una herramienta ROLAP la cual puede ser una muy buena opción en un inicio para usuarios pioneros en los depósitos de datos, con esta herramienta se puede tener una visión de que es lo que se necesita y si esta cumple con las necesidades de los usuarios que necesitan información confiable que les brinde un apoyo para tomar una decisión oportuna en un momento crítico.

Anexo (A)

INGENIERÍA DE HARDWARE

OBLIGATORIAS
DISENO DE INTERFACES PARA COMPUTADORAS (6)
COMPUTO MOVIL (6)
SISTEMAS EMBEBIDOS (6)
OPTATIVAS
ROBOTICA (L) (10)
ROBOTS MOVILES Y AGENTES INTELIGENTES (6
PROCESAMIENTO DIGITAL DE SENALES (9)
SISTEMAS DIFUSOS (6)
INSTRUMENTACION VIRTUAL (L+)(8)
CONTROL AUTOMATICO INDUSTRIAL (L+) (8)
FISICA MODERNA (L) (6)
TEMAS SELECTOS DE INGENIERIA DE HARDWARE (6)
SEMINARIO DE TITULACION (6)*
PROYECTO DE INVESTIGACION (6)**

REDES Y SEGURIDAD
OBLIGATORIAS
CRIPTOGRAFIA (6)
SEGURIDAD INFORMATICA I (6)
SEGURIDAD INFORMATICA II (6)
ARQUITECTURAS CLIENTE/SERVIDOR (6)
OPTATIVAS
DESARROLLO DE SOFTWARE SEGURO (6)
ANALISIS Y DISENO DE REDES DE DATOS (6)
REDES INALAMBRICAS AVANZADAS (6)
TEMAS SELECTOS DE NORMALIZACION (6)
COMPRESION DE DATOS (6)
CODIFICACION DE AUDIO Y VIDEO (6)
TEMAS SELECTOS DE REDES Y SEGURIDAD (6)
SEMINARIO DE TITULACION (6)*
PROYECTO DE INVESTIGACIÓN (6)**

BASES DE DATOS
OBLIGATORIAS
BASES DE DATOS AVANZADAS (6)
BASES DE DATOS DISTRIBUIDAS (6)
BASES DE DATOS ESPACIALES (6)
OPTATIVAS
DEPOSITOS DE DATOS (6)
MINERIA DE DATOS (6)
TEMAS SELECTOS DE BASES DE DATOS (6)
SEMINARIO DE TITULACION (6)*
PROYECTO DE INVESTIGACIÓN (6)**

INGENIERIA DE SOFTWARE

OBLIGATORIAS

NEGOCIOS ELECTRONICOS (6)
VERIFICACION Y VALIDACION DE SOFTWARE (6)
COMPUTO DE ALTO DESEMPENO (6)
COMPUTO MOVIL (6)

OPTATIVAS

SISTEMAS EN TIEMPO REAL (6)
DESARROLLO DE SOFTWARE SEGURO (6)
ARQUITECTURAS CLIENTE/SERVIDOR (6)
MINERIA DE DATOS (6)
APRENDIZAJE (6)
DISENO DE INTERFACES, MULTIMEDIA Y REALIDAD VIRTUAL (6)
FISICA MODERNA (L) (6)
TEMAS SELECTOS DE INGENIERIA DE SOFTWARE (6)
SEMINARIO DE TITULACION (6)*
PROYECTO DE INVESTIGACION (6)**

SISTEMAS INTELIGENTES Y COMPUTACIÓN GRÁFICA

SISTEMAS INTELIGENTES

OBLIGATORIAS

SISTEMAS EXPERTOS (6)
ROBOTS MOVILES Y AGENTES INTELIGENTES (6
APRENDIZAJE (6)
RECONOCIMIENTO DE PATRONES (6)
TEMAS SELECTOS DE SISTEMAS INTELIGENTES (6

OPTATIVAS

SEMINARIO DE TITULACION (6)*
PROYECTO DE INVESTIGACION (6)**

TECNOLOGIA DEL LENGUAJE

OBLIGATORIAS

PROCESAMIENTO DEL LENGUAJE NATURAL (6)
PROCESAMIENTO DIGITAL DE VOZ (6)
PROCESAMIENTO DE CORPUS TEXTUALES Y ORALES (6

OPTATIVAS

RECONOCIMIENTO DE PATRONES (6)
ANALISIS Y PROCESAMIENTO INTELIGENTE DE TEXTOS (6
APRENDIZAJE (6)
TEMAS SELECTOS DE TECNOLOGIAS DEL LENGUAJE (6
SEMINARIO DE TITULACION (6)*
PROYECTO DE INVESTIGACION (6)**

COMPUTACION GRAFICA

OBLIGATORIAS

COMPUTACION GRAFICA AVANZADA (6)
PROCESAMIENTO DIGITAL DE IMAGENES (6)
DISENO DE INTERFACES, MULTIMEDIA Y REALIDAD VIRTUAL (6)
DISENO ASISTIDO POR COMPUTADORA (6)
TEMAS SELECTOS DE GRAFICACIÓN (6)

OPTATIVAS

SEMINARIO DE TITULACIÓN (6)*
PROYECTO DE INVESTIGACIÓN (6)**

INGENIERÍA BIOMÉDICA

OBLIGATORIAS

CIRCUITOS INTEGRADOS ANALOGICOS (L+) (11)
INTRODUCCION A LA FISIOLOGIA (L+) (8)
FUNDAMENTOS DE INSTRUMENTACION BIOMEDICA (L+) (8)
PROCESAMIENTO DIGITAL DE IMAGENES MEDICAS: Imagenologia (L+) (8)

OPTATIVAS

APLICACIONES DE OPTOELECTRONICA EN MEDICINA (L+) (8)
AUDIOMETRIA (6)
TELESALUD (6)
SISTEMAS Y EQUIPOS BIOMEDICOS ELECTRONICOS (6)
TEMAS SELECTOS DE INGENIERIA BIOMEDICA (8)
TRANSDUCTORES BIOMEDICOS (6)
INTRODUCCION A LA BIOFISICA (6)
SEMINARIO DE TITULACION (6)*
PROYECTO DE INVESTIGACION (6)*

OPTATIVAS DE COMPETENCIAS PROFESIONALES

CREATIVIDAD (6)
RELACIONES LABORALES Y ORGANIZACIONALES (6)
CONTABILIDAD FINANCIERA Y COSTOS (6)
SISTEMAS DE PLANEACION (8)
INTRODUCCION AL ANALISIS ECONOMICO EMPRESARIAL (6)
ADMINISTRACION DE CENTROS DE TECNOLOGIA DE INFORMACION (8)
DESARROLLO EMPRESARIAL (6)
COSTOS Y EVALUACION DE PROYECTOS (6)
CALIDAD (6)

Fuente: **http://www.ingenieria.unam.mx/revplanes/planes2006/_mapas_curriculares/computacion.pdf**

Anexo (B)

Herramientas de consulta y reporte:

PRODUCTO	EMPRESA DISTRIBUIDORA
Access	Microsoft
Access+	Sonetics
Actuate Reporting System	Actuate Software Corporation
AMIS Information Server	Hoskyns Group plc
Application System	IBM
Approach	Lotus Corporation
ARPEGGIO	Wall Data Inc.
APTuser	International Software Group
AS/Access for Microsoft Access	Martin Spencer & Associates
ASK Joe	Information Management Services
aXcess/400	Glenbrook Software
BrioQuery	Brio Technology
Business Objects	Business Objects, Inc.
Clear: Access	Sterling Software
Crystal Reports, Crystal Info	Seagate Software
d.b. Express	Computer Concepts Corp.
Databoard, Dataread	SLP Infoware
DataDirect Explorer	Intersolv
DataSite	NetScheme Solutions, Inc.
DB Publisher	Xense Technology Inc.
DbPower	Db-Tech Inc.
Decision Analyzer	Decisión Technology
DECquery, DECdecision	Touch Technologies, Inc.
Discoverer, Discoverer/2000	Oracle Corporation
DS Server, DS Modeler	Interweave
EasyReporter	Speedware Corporation
Eclipse Query/Report	Cornut Informatique
ELF	ELF Software
English Wizard	English Wizard
EnQuiry	Progress Software
Esperant	Speedware
FOCUS Six	Information Builders, Inc.
4S-Report	Four Seasons Software, Inc
Freequery	Dimension Software Systems
Front & Center for Reporting, Nomad	Thomson Software Products
GQL	Andyne
HarborLight	Harbor Software
HP Information Access	Hewlett-Packard
if…	Leep Technology, Inc.
Impress, SqlBuddy	Objective Technologies, Inc.
Impromptu	Cognos Corporation
InfoAssistant	Asymetrix

InfoMaker	Powersoft Corporation
InfoQuery	Platinum Technology, Inc.
InfoReports	Platinum Technology, Inc.
InformEnt Warehouse Desktop	Fiserv
Internet DataSpot	DTL Data Technologies Ltd.
inSight	Williams & Partner
Interactive Query	New Generation software
IQ/Objects, IQ/SmartServer	IQ Software Corporation
Iridon Panorama	The Great Elk Company Limited
Kinetix	Hilco Technologies
LANSA/Client	LANSA USA
MARKIS/400	AS Software
Nirvana	Synergy Technologies
OR-REPORTER II	Output Reporting, Inc.
Oracle Reports, Browser	Oracle Corporation
Paradox	Borland
Platinum Report Facility	Platinum Technology, Inc
ProBit	System Builder
Productivity Series Reports	michaels, ross & cole
QBE Vision	Sysdeco
QMF	IBM
QueryObject	Cross/Z International, Inc.
Quest	Centura Software Corporation
R&R Report Writer Report Writer	Concentric Data Systems Raima
Reportoire	Synergistic Systems, Inc.
Reports	Nine to Five software Co.
ReporTool	Zen Software
ReportSmith	Borland
Rocket Shuttle	Rocket Software, Inc.
Safari ReportWriter	Interactive Software Systems
Sagent Data Mart Solution	Sagent Technology, Inc.
SAS System	SAS Institute
Second Wind	Anju Technologies
Select!	Attachmate
SEQUEL	Advanced Systems Concepts
Snow Report Writer	Snow International Corporation
Spectrum Writer	Pacific Systems Group
SQLPRO Agent	Beacon Ware, Inc.
SQR Workbench	MITI
Star Tracker	Leep Technology, Inc.
Strategy	ShowCase Corporation
The Reporter	Sea Change Systems, Inc
Unique XTRA	Unique AS
URSA InfoSuite	Decision Support Inc.
ViewPoint	Informix

ViewPoint	Soliton Associates
Viper	Brann Software
VisPro/Reports	Hock Ware
Visual Cyberquery	Cyberscience Corporation
Visual Dbase	Borland
Visual Express	Computer Associates International
Visual FoxPro	Microsoft Corporation
Visual Net	CNet Svenska AB
Visualizer Query, Charts	IBM
Voyant	Brossco Systems
WebBiz	Cybercom Partners
WebSeQueL	InfoSpace Inc.
WinQL	Data Access Corporation
Xentis	GrayMatter Software Corporation

Herramientas de base de datos multidimensional/OLAP

PRODUCTO	EMPRESA DISTRIBUIDORA	TIPO
Acuity ES	Acuity Management Systems Ltd.	MDDB
Acumate ES	Kenan Systems Corporation	MDDB
Advance For Windows	Lighten, Inc.	MDDB
AMIS OLAP Server	Hoskyns Group plc	MDDB
BrioQuery	Brio Technology	MDDB
Business Objects	Business Objects, Inc.	Relacional
Commander OLAP, Decision, Prism	Comshare Inc.	MDDB
Control	KCI Computing	Relacional
CrossTarget	Dimensional Insight	MDDB
Cube-It	FICS Group	MDDB
Dataman	SLP Infoware	MDDB
DataTracker	Silvon Software, Inc.	Relacional
DecisionSuite	Information Advantage, Inc.	Relacional
Delta Solutions	MIS AG	MDDB
Demon for Windows	Data Command Limited	MDDB
DSS Agent	MicroStrategy	Relacional
DynamicCube.OCX	Data Dynamics, Ltd.	Relacional
EKS/Empower	Metapraxis, Inc.	MDDB
Essbase Analysis Server	Arbor Software Corporation	MDDB
Essbase/400	ShowCase Corporation	MDDB
Express Server, Objects	Oracle	MDDB
Fiscal	Lingo Computer Design, Inc.	Relacional
Fusion	Information Builders, Inc.	MDDB
FYI Planner	Think Systems	MDDB
Gentia	Planning Sciences	MDDB
Helm	Codeworks	MDDB
Holos	Holistic Systems	MDDB
Hyperion OLAP	Hyperion Software	MDDB
InfoBeacon	Platinum technology, Inc.	Relacional
Informer	Reportech	MDDB/Relacional

Intelligent Decision Server	IBM	Relacional
IQ/Vision	IQ Software Corporation	Relacional
Khalix	Longview Solutions, Inc.	Relacional
Lightship	Pilot Software, Inc.	MDDB
Matryx	Stone, Timber, River	MDDB
MDDB Server	SAS	Relacional
Media	Speedware Corporation	MDDB
Metacube	Informix	Relacional
MIKSolution	MIK	MDDB
MIT/400	SAMAC, Inc	MDDB
MSM	Micronetics Design Corporation	MDDB
Muse	OCCAM Research Corp.	MDDB
OLAP Office	Graphitti Software GmbH	MDDB
OpenOLAP	Inphase Software Limited	Relacional
Pablo	Andyne	MDDB/Relacional
ParaScope	DataVista	Relacional
PowerPlay	Cognos Corporation	MDDB/Relacional
StarTrieve	SelectStar	Relacional
The Ant Colony	Geppetto's Workshop LLC	Relacional
TM/1	Applix	MDDB
Toto	Ambit Research Ltd.	MDDB
Track for OLAP	Track Business Solutions	MDDB
Visualizer Plans for OS/2	IBM	MDDB

Bases de datos usadas para los Depósitos de datos.

PRODUCTO	EMPRESA DISTRIBUIDORA
Adabas D	Software AG
Advanced Pick	Pick Systems
DB2	IBM
Fast-Count DBMS	MegaPlex Software
HOPS	HOPS International
Microsoft SQL Server	Microsoft
Model 204	Computer Corporation of America
NonStop SQL	Tandem
Nucleus Server	Sand Technology Systems
OnLine Dynamic Server, Extended Parallel Server	Informix
OpenIngres	Computer Associates
Oracle Server	Oracle
Rdb	Oracle

Rdb	Oracle
Red Brick Warehouse	Red Brick Systems
SAS System	SAS
Sybase IQ	Sybase
Sybase SQL Server, SQL Server MPP	Sybase
SymfoWARE	Fujitsu
Teradata DBS	NCR
THOR	Hitachi
Time Machine	Data Management Technologies, Inc.
Titanium	Micro Data Base Systems, Inc.
Unidata	Unidata, Inc.
UniVerse	VMARK
Vision	Innovative Systems Techniques, Inc.
WX9000	White Cross Systems Inc.
XDB Server	XDB Systems, Inc.

Conclusiones

HORACIO HERNANDEZ ALVARADO

La materia depósito de datos es una materia optativa dentro del nuevo plan de estudios de la carrera de ingeniería en computación en la Facultad de Ingeniería de La UNAM y por tanto de inicio un poco “experimental” en lo que se refiere a su temario y horas propuestas para desarrollar este en clase.

Esto nos motivo ha realizar un análisis de los temas de esta materia y desarrollar cada uno de sus capítulos, con el fin de proporcionar a los alumnos que cursen esta asignatura un materia de apoyo que complemente la bibliografía de la misma. La presente tesis engloba el análisis y desarrollo de los capítulos de la asignatura deposito de datos por lo que se concluye que el objetivo de esta tesis fue cumplido.

Esta asignatura tiene capítulos muy extensos y fue uno de los principales problemas con los que nos enfrentamos, otro problema fue el caso contrario, es decir capítulos en los cuales existía muy pocas o nulas fuentes de información, debido a que los depósitos de datos son una nueva tecnología que empieza a ser, cada vez mas demandada por profesionales que necesitan información integrada y global de su organización.

Estos problemas en la obtención de información nos dieron la pauta para saber el estado actual de los depósitos de datos, es decir nos muestra que elementos o características de los depósitos de datos son poco utilizadas o implementadas en este tipo de tecnología.

Las principales aportaciones que se dan en esta tesis es la reunión de todo el material que se investigo de los temas propuestos en los capítulos de la asignatura Depósito de Datos, otra aportación es el caso práctico, el cual pretende mostrar la forma en la que funciona un Depósito de Datos utilizando como base un sistema relacional. Este tipo de herramienta genera cubos de información y es en este caso práctico donde se muestra al alumno la forma básica de analizar, diseñar, estructurar y generar cubos de información.

Una propuesta de nuestra parte es la de generar apuntes para la asignatura en base al presente trabajo. Esta información en un inicio deberá de orientar al estudiante en los temas de esta asignatura, esta información básica pretende motivar al alumno a obtener mas información actualizada de estos capítulos los cuales con toda seguridad cambiaran de acuerdo al acelerado flujo de información que existe en una organización así como la actualización en la tecnología de diseño y generación de almacenes de datos y por supuesto las nuevas necesidades de los usuarios para obtener información en tiempo y forma de este tipo de aplicaciones.

MA. MAIYEC URRUTIA LUNA

La Asignatura de Depósito de datos permitirá a los alumnos de la facultad de Ingeniería de la UNAM competir en el mundo real, sin embargo es conveniente hacer hincapié que el tiempo proporcionado para cada tema no es suficiente por lo cual es necesario que el alumno busque las respuestas y soluciones de forma autodidacta. Un material didáctico es una alternativa para la asignatura de Depósito de Datos, proporciona al alumno apoyo sobre conceptos actuales hasta el momento referentes al tema Depósito de datos.

Un material didáctico, además de incluir los conceptos será aun de más ayuda si éste incluye en sus páginas casos prácticos, los cuales sean de fácil creación y adquisición del software. “La practica hace al maestro” y no esta por de más que el alumno comprender de que forma puede implementar un Depósito de datos aun siendo éste pequeño, sin embargo, el implementar un depósito de datos pequeño no será el mismo caso que un depósito real, ya que todo depósito de datos construido es diferente por su información contenida y las necesidades de cada área o personal, pero el análisis, la explosión de información, selección de metodología, construcción, implementación, etc. le serán más fáciles para un caso real.

Referente al tiempo asignado a cada tema, el tema Tendencias en el capitulo 6 consideramos que no se asigno correctamente, este debe ser de menos tiempo, talvez de 2 horas y el resto asignarlo a los temas de mayor extensión como es a los temas Planeación de los Depósitos de datos y Diseño e Implementación de los Depósitos de datos

DAVID RAMÍREZ ARREOLA

Una parte fundamental de mi vida se ha centrado en el desarrollo de esta tesis la cual tiene como fin presentar una herramienta adicional al gran trabajo y labor que presentan los profesores de la Universidad Nacional Autónoma de México en su tarea diario de enseñar y preparar el camino de los profesionistas mexicanos , Herramienta que debe ser utilizada en el estudio de la materia Deposito de Datos al presentar un enfoque claro sobre la situación actual del manejo de datos, y su aplicación en la industria , este material se enfoca al apoyo con ejemplos claros y prácticos para el lector desde el inicio tomando en cuenta el análisis por el cual la Facultad de Ingeniería considero dar cátedra a esta materia, mientras el lector se adentra en los capítulos se encuentra con temas que son referencia en la actualidad para la migración, proceso de transformación de datos, y finalmente la conceptualización de que, para, y el porque es necesario considerar un Almacén de datos en cualquier desarrollo empresarial, o sobre algún desarrollo ya implementado, esto permitirá al lector (Alumno) tener una visión mas clara de cómo proyectar la explosión de datos en un futuro para agilizar la toma de desiciones, estrategias y rumbo a seguir por cualquier organización empresarial que pretenda extender su producto, comercio, distribución etc.

Tomamos como una conclusión general la necesidad de integrar esta materia en el currículo básico de la materia base de datos, no como una materia especialización sino como una materia base ya que esto contribuye de gran manera al desarrollo individual y colectivo de la plantilla del alumnado de la facultad de ingeniería en la carrera de ingeniería en computación, además se recomienda dar mas horas al desarrollo del tema Minería de datos, en general el estudio de la materia Deposito de Datos es una muy buena opción como materia optativa integrada en el mapa curricular.

Glosario

Ácido desoxirribonucleico (*deoxyribonucleic acid*); uno de los dos tipos fundamentales de ácidos nucleicos que se encuentran en los organismos vivos. El ácido desoxirribonucleico ó **ADN** es generalmente bicatenario, formando una estructura tridimensional helicoidal que se conoce como doble hélice. La pentosa en los residuos de nucleótidos que lo forman es la 2'-desoxiribosa, a la cual se unen indistintamente las bases nitrogenadas adenina (A), guanina (G), citosina (C) y timina (T). El ADN es la molécula portadora de la información genética.

ADN (*DNA*); v. Ácido desoxirribonucleico.
Backplane; Un backplane es un componente de las computadoras con arquitectura NUMA (Non-uniform memory access) que facilita la interconexión de los módulos o celdas que componen la computadora.

Biochip; conjunto de numerosas moléculas de ADN clonado inmovilizadas formando una estructura ordenada y compacta de pequeñas gotas (inferiores a 1 microlitro) en una matriz sólida (generalmente un portaobjetos de cristal). Se utiliza para analizar patrones de expresión génica, detectar marcadores, o secuenciar nucleótidos. La ventaja principal de estos dispositivos es el grado al que puede automatizarse el proceso de genotipado.

Bioinformática (*bioinformatics*); de manera simplificada, la aplicación de las tecnologías de la computación y la información al manejo y análisis de información de origen biológico.

BPR; Business Process Reengineering. Reingeniería del proceso empresarial

Cluster; Un clúster (o unidad de asignación según la terminología de Microsoft) es un conjunto contiguo de sectores que componen la unidad más pequeña de almacenamiento de un disco. Los archivos se almacenan en uno o varios clústeres, dependiendo de su tamaño. Sin embargo, si el archivo es más pequeño que un clúster, éste ocupa el clúster completo.

Cliente/servidor; Modelo lógico de una forma de proceso cooperativo, independiente de plataformas hardware y sistemas operativos. El concepto se refiere más a una filosofía que a un conjunto determinado de productos. Generalmente, el modelo se refiere a un puesto de trabajo o cliente que accede mediante una combinación de hardware y software a los recursos situados en un ordenador denominado servidor.

Data Mart o Mercado de Datos; es una base de datos o colección de bases de datos, pero con contenidos específicos.

DBMS; DataBase Management System. Sistema Gestor de Bases de Datos (SGBD). Son un tipo de software muy específico, dedicado a servir de interfaz entre las bases de datos y las aplicaciones que la utilizan.

Desnormalización; Una base de datos normalizada impide las dependencias funcionales de los datos para que el proceso de actualización de la base de datos sea fácil y eficiente. Sin embargo, la realización de consultas en la base de datos puede requerir la combinación de varias tablas para unir la información. A medida que el número de tablas combinadas crece, el tiempo de ejecución de la consulta aumenta considerablemente. Por este motivo, el uso de una base de datos normalizada no es siempre la mejor alternativa. Una base de datos con la medida justa de desnormalización reduce el número de tablas que deben combinarse sin dificultar en exceso el proceso de actualización. Suele ser la solución más acertada.

DOLAP es un OLAP orientado a equipos de escritorio (Desktop OLAP). Trae toda la información que necesita analizar desde la base de datos relacional y la guarda en el escritorio. Desde ese momento, todas las consultas y análisis son hechas contra los datos guardados en el escritorio.

Drill Down: Exponer progresivamente más detalle (dentro de un reporte o consulta), mediante selecciones de índices sucesivamente.

Drill-Up: Es el efecto contrario a drill -down. Significa ver menos nivel de detalle, sobre la jerarquía significa generalizar o sumarizar, es decir, subir en el árbol jerárquico.

DSS : Sistema de Soporte de Decisiones. Sistema de aplicaciones automatizadas que asiste a la organización en la toma de decisiones mediante un análisis estratégico de la información histórica. Decision Support System. Sistema automatizado de aplicación que ayudan a la organización a tomar decisiones relacionadas con el negocio.

DBMS; Los Sistemas Gestores de Bases de Datos son un tipo de software muy específico, dedicado a servir de interfaz entre las bases de datos y las aplicaciones que la utilizan. En los textos que tratan este tema, o temas relacionados, se mencionan los términos SGBD y DBMS, siendo ambos equivalentes, y acrónimos, respectivamente, de Sistema Gestor de Bases de Datos y DataBase Management System, su expresión inglesa.

ETL; proceso de extracción, transformación y carga.

F.A.Q.; Frequently Asked Questions. Preguntas frecuentes. Relación de preguntas comunes sobre un tema determinado, con sus respuestas.

F.A.R.; Frequently Asked Report. Reportes frecuentes. Relación de preguntas comunes sobre un tema determinado.

Genoma (*genome*); el total de la información genética de un organismo o entidad biológica, dado en la forma de secuencias de ADN (excepto en algunos virus cuyo genoma está constituido por ARN). El genoma puede considerarse como estático cuando se compara con el proteoma.

Genómica (*genomics*), rama de la biología que se dedica al estudio del genoma, incluyendo los métodos y técnicas específicas que se usan con este objetivo.

Genotipo (*genotype*); el total de la información genética de un organismo. A veces se intercambia con el término genoma.

Granularidad; Granularidad de sistema se usa como el mínimo componente del sistema que puede ser preparado para ejecutarse de manera paralela. La granularidad del hardware es la instrucción máquina (o en lenguaje ensamblador). Esta unidad es la más fina que puede tratar. Termino utilizado para referirse al nivel de detalle de la información. A mayor nivel de granularidad, es menor el nivel de detalle de la información.

GUI: suelen llamarse WIMP. Este acrónimo en inglés se refiere a "windows, icons, menus, pointing device" (ventanas, íconos, menús, dispositivos de señalamiento) y en general, todos estos elementos se encontrarían en una Interfase Gráfica de Usuario.

Heterogéneo; No es uniforme.

Homogéneo; Que es de la misma clase, de la misma especie, de estructura totalmente igual; que posee las mismas funciones parciales o las mismas partes
HOLAP (Hybrid OLAP) almacena algunos datos en un motor relacional y otros en una base de datos multidimensional.

Kernel; el kernel (también conocido como núcleo) es la parte fundamental de un sistema operativo. Es el software responsable de facilitar a los distintos programas acceso seguro al hardware de la computadora. Como hay muchos programas y el acceso al hardware es limitado, el núcleo también se encarga de decidir qué programa podrá hacer uso de un dispositivo de hardware y durante cuánto tiempo, lo que se conoce como multiplexado.

Metadata ; (Metadato) es la información sobre los datos que se alimenta.

MOLAP es una implementación OLAP que almacena los datos en una base de datos multidimensional.

Origen de datos; toda aquella información que posee una organización, y puede estar contenida en cualquier medio electrónico, también es llamado Fuentes de datos.

MPP; procesamiento en paralelo masivo.

Normalización; La normalización o estandarización es la redacción y aprobación de normas. Normalizar las bases de datos significa procesarlas para minimizar la redundancia y los posibles errores de inserción, eliminación y actualización. Hay tres clases de normalización, primera, segunda y tercera forma normal.

OLAP; procesamiento analítico en línea.

Oracle; es un sistema de administración de base de datos (o RDBMS por el acrónimo en inglés de Relational Data Base Management System), fabricado por Oracle Corporation. Servidor para bases de datos. Un servidor Oracle posee una base de datos Oracle y la instancia Oracle y soporta acceso por SQL y por lenguajes de programación. También posee un lenguaje de procedimientos llamado PL/SQL. Las bases de datos Oracle poseen dos estructuras primarias, la estructura física (datos almacenados) y la lógica (representación abstracta de los datos).

Ortogonalidad; es una propiedad de las CPU. Se dice que un conjunto de instrucciones es ortogonal cuando se puede utilizar cualquier modo de direccionamiento en cualquier instrucción. La búsqueda de la ortogonalidad hace que el diseño de la CPU sea más complejo pero aporta una mayor facilidad de programación.

Paralelismo; facultad de ejecutar varios programas o instancias del mismo, en dos o más procesadores a la vez.

Proteoma (*proteome*); el conjunto de proteínas que se están expresando en un momento dado, en una célula, tejido u organismo. El proteoma puede considerarse como dinámico si se compara con el genoma, pues varía con el tiempo y/o con diferentes estados fisiológicos o patológicos específicos.

RDBMS; Es un Sistema Administrador de Bases de Datos Relacionales. RDBMS viene del acrónimo en ingles Relational Data Base Manager System.
ROLAP es una implementación OLAP que almacena los datos en un motor relacional.
SMP: es el acrónimo de symetric multi-processing, multiproceso simétrico. Se trata de un tipo de arquitectura de ordenadores que se encuentra en las computadoras personales y servidores de baja gama.

Sumarización: Actividad de incremento de la granularidad de la información en una base de datos. La sumarización reduce el nivel de detalle, y es muy útil para presentar los datos para apoyar al proceso de Toma de Decisiones.

SMP: Mmultiprocesamiento simétrico.

SQL; El Lenguaje de Consulta Estructurado (Structured Query Language) es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones sobre las mismas. Aúna características del álgebra y el cálculo relacional permitiendo lanzar consultas con el fin de recuperar información de interés de una base de datos, de una forma sencilla.

SQL Server; Microsoft SQL Server es un programa informático de gestión y administración de bases de datos relacionales basada en el lenguaje SQL, que incluye también un potente entorno gráfico de administración, que permite el uso de comandos DDL y DML gráficamente.

SYBASE IQ; es la forma inteligente de llevar acabo análisis empresarial de datos de alto rendimiento. Es un servidor analítico optimizado, está diseñado específicamente parta entregar resultados dramáticamente más rápido en ambientes críticos de inteligencia de negocio, depósitos de datos y generación de reportes, sobre una variedad de plataformas y sistemas operativos.

Servidor; una computadora que realiza algunas tareas en beneficio de otras aplicaciones llamadas clientes. Algunos servicios habituales son los servicios de archivos, que permiten a los usuarios almacenar y acceder a los archivos de un ordenador y los servicios de aplicaciones, que realizan tareas en beneficio directo del usuario final.

TQM; total Quality Management. Siglas correspondientes en inglés al concepto de Gestión de Calidad Total, entendido como un proceso de implantación de la calidad en el que está implicada toda la organización.

BIBLIOGRAFÍA

SILBERSCHATZ ABRAHAM, HENRY F. KORTH, SUDARSHAN
FUNDAMENTOS DE BASES DE DATOS
CUARTA EDICION
PAG 537- 568. ED MC GRAW HILL 2002

REVISADO Y EDITADO POR EL PROFESOR J. ELLIOTT.
MANUAL PARA LA CONSTRUCCIÓN DE UN DATA WAREHAUSE. PUBLICADO POR EL INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMÁTICA INEI/1997, LIMA.

PETER ROB, CARLOS CORONEL
SISTEMAS DE BASES DE DATOS
DISEÑO IMPLEMENTACION Y ADMINISTRACION
21, 613, 622 ED. TOMSON 2003

HARJINDER S. GILL, PRAKASH C. RAO
DATA WAREHAUSING,
LA INTEGRACIÓN DE INFORMACIÓN PARA LA MEJOR TOMA DE DECISIONES.
PRENTICE HALL HISPANOAMERICANA, S.A.

GILL Y RAO
LA INTEGRACIÓN DE INFORMACIÓN PARA LA MEJOR TOMA DE DECISIONES DATA WAREHOUSING.
NEW JERSEY
PRENTICE HALL

JILL DYCHÉ
TRANSFORMANDO DATOS EN INFORMACIÓN CON DATA WAREHOUSING.
E-DATA
NEW JERSEY
PRENTICE HALL

JORGE HUMBERTO JARAMILLO MONSALVE
JOHN ALEXANDER LÓPEZ NOREÑA
MINERIA DE DATOS
DATAMINING

UNIVERSIDAD DE ANTIOQUIA
ESPECIALIZACIÓN EN GESTIÓN DE SISTEMAS Y BASES DE DATOS
MEDELLÍN, MARZO 31 DE 2000

COMBINACIÓN DE BASE DE DATOS MEDIANTE HERRAMIENTAS INFORMÁTICAS OLAP – ERP EN COSTOS.
REVISTA ELECTRÓNICA FCE. UNIVERSIDAD CATÓLICA DEL URUGUAY

REVISTA COLOMBIANA DE COMPUTACION
PAOLO ROSSO
UNIVERSIDAD POLITECNICA DE VALENCIA, ESPAÑA
VOLUMEN 2, NUMERO 1
PAG. 63-73

REVISTA I+D COMPUTACION
VOL.2, NO. 2, JULIO 2003
MANUEL SERRANO, ISMAEL CABALLERO, ET AL.

MORGAN KAUFMAN
DATA MINING. ADDISON-WESLEY
PYLE. D. (1999). DATA PREPARATION FOR DATA MINING.
ADRIAANS, P (1996).

INMON, HILL Y CHECK KELLEY,
THE TWELVE RULES OF DATA WAREHAUSE FOR A CLIENTE/SERVER WORLD
DATA MANAGEMENT REVIEW, 4(5),
MAYO DE 1994, PAGS. 6A 16

WEB METADATA: A MATTER OF SEMANTIC, IEEE INTERNET COMPUTING, LASSILA O.,
JULY/AUGUST 1998, 2, 4, 30-37.

JOSEP AGUILAR SABORIT
TESIS: TÉCNICAS PARA LA MEJORA DEL JOIN PARALELO Y DEL PROCESAMIENTO DE SECUENCIAS TEMPORALES DE DATOS.
BARCELONA, ESPAÑA, MAYOP 2006

MTTHEW JOSEPH SOTTILE
A MEASUREMENT AND SIMULATION METHOLOGY FOR PARALLEL COMPUTING PERFORMANCE STUDIES
UNIVERSIDAD DE OREGON, 1999

ORACLE8I PARALLEL SERVER CONCEPTS AND ADMINISTRATION RELEASE 8.1.5
A67778-01
FEBRUARY 1999
COPYRIGHT © 1999 ORACLE CORPORATION. ALL RIGHTS RESERVED.
PRIMARY AUTHOR: MARK BAUER.
PRIMARY CONTRIBUTORS: WILSON CHAN, ANDREW HOLDSWORTH, ANJO KOLK, RITA MORAN, GRAHAM WOOD, AND MICHAEL ZOLL.

AUTHOR: KRISTY BROWDER EDWARDS, GEORGE LUMPKIN
SECURITY AND THE DATA WAREHOUSE
AN ORACLE WHITE PAPER
APRIL 2005
CONTRIBUTING AUTHORS: MARY ANN DAVIDSON, PAUL NEEDHAM, JOHN HEIMANN

Ligas en Internet

http://www.sqlmax.com/dataw1.asp
http://www.ingenieria.unam.mx/revplanes/planes2006/_mapas_curriculares/computacion.pdf
http://msdn.microsoft.com/library/spa/default.asp?url=/library/SPA/dntaloc/html/datawarefmwk.asp
http://ieeexplore.ieee.org
http://200.14.84.223/apuntesudp/docs/civil_ind/(ICI2423)Sistemas_de_Informacion/(2003-06-11)_818_datawarehouse_o_bodegaje_de_datos.pdf
http://msdn.microsoft.com/library/spa/default.asp?url=/library/SPA/dntaloc/html/datawarefmwk.asp
http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnsql7/html/msdn_sqlrep.asp
http://www.ifla.org/IV/ifla66/papers/029-142s.htm
http://www.csee.umbc.edu/help/oracle8/server.815/a67439.pdf