CP 35
PON 69

Bases de Datos Multidimensionales con SQL Server para la Administración de la Docencia

Pedro Fletes-Gudiño, Nicandro Farías-Mendoza, Eliseo Mendoza-Luna, Iván S. González-Magaña, Rafael Espinosa-Cárdenas

Instituto Tecnológico de Colima, Av. Tecnológico, Colima, México

pfletes@itcolima.edu.mx, nfarias@itcolima.edu.mx, eliseoml@hotmail.com, ivangonz77@gmail.com, respinoza@itcolima.edu.mx

***Resumen*- Durante los últimos años las bases de datos multidimensionales se han convertido en algo común en el mundo de los negocios y científicos. La necesidad de contar con información cada vez más sistematizada y tomando en cuenta incluso información histórica, hace que esta necesidad sea cada día más necesaria e implícita en el mundo de cualquier empresa. En este trabajo se propone un modelo de base de datos multidimensional, que nos permita a través de un proceso, clasificar, filtrar información del área de docencia, particionando los registros en capas y paneles, para así, a través de instrucciones de SQL procesar información como la requiera el usuario, la intención además es abarcar la información de docencia de todas las áreas del Instituto. La meta que se pretende alcanzar a través de este modelo propuesto, es la utilización de instrucciones SQL para la consulta de la información en base de datos multidimensionales.**

Palabras Clave

Base de datos Multidimensionales, SQL, procesos, costos.

I. INTRODUCCION

En una base de datos multidimensional, la información se representa como matrices multidimensionales, cuadros de múltiples entradas o funciones de varias variables sobre conjuntos finitos. Cada una de estas matrices se denomina Cubo.

El esquema de un cubo queda determinado dando a conocer sus ejes con sus respectivas estructuras y la estructura de los datos que se presentan en cada celda de la matriz. Se asume que los datos en todas las celdas son uniformes, es decir, todas las posiciones de la matriz tienen datos con igual estructura.

Una instancia de un cubo, queda determinada por un conjunto de datos para cada eje y un conjunto de datos para la matriz ver (Fig. 1).

A los ejes se les llama Dimensiones y al dato que se presenta en la matriz, se le llama Medida. A los elementos del producto cartesiano de los ejes (dimensiones) se le llama coordenadas.

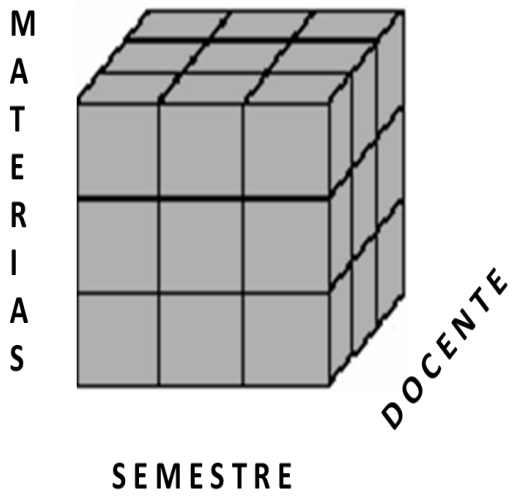


Fig.1 La estructura básica en el modelo Multidimensional: el cubo

La importancia de este proyecto, es el ahorro de tiempo y costo en la obtención de la información a través de un esquema multidimensional, en comparación con las herramientas utilizadas como las hojas de cálculo de Excel.

Trabajos relacionados

Pitarch en [1] propone un modelo conceptual para modelar jerarquías personalizadas en bases de datos multidimensionales. Golfarelli en [2] nos presenta como a través de consultas anidadas utilizando el comando Nested, se pueden optimizar estas, proponiendo consultas sencillas

ROC&C'2013 / CP 35 PONENCIA RECOMENDADA POR EL COMITE DE COMPUTACION DEL IEEE SECCION MEXICO Y PRESENTADA EN LA REUNION INTERNACIONAL DE OTOÑO, ROC&C'2013.
ACAPULCO, GRO., DEL 10 AL 14 DE NOVIEMBRE DEL 2013.

en secuencia. Un lenguaje IRAH permite aumentar las funciones tradicionales de resumen extensionales con conocimiento intencional en [3]. La propuesta de un algoritmo de construcción de elipsoides basado en (ER-Tree) en [4] nos demuestra que estos afectan significativamente la velocidad, la indexación y el rendimiento de las consultas en bases de datos grandes como de multimedia, medicina y geográficas. En [5] Liang presenta 2 algoritmos para la optimización de consultas, uno de aproximación polinómica para consultas que utilizan el método hash y otro algoritmo exponencial que ofrece un plan con la evaluación óptima de costo. En [6] es propuesto un lenguaje de consulta visual , y este permite, la visualización gráfica de datos basados en tablas. Este lenguaje compila en tanto los comandos de las consultas y los dibujos necesarios para generar la visualización, lo que nos permite diseñar sistemas que integren estrechamente análisis y visualización. Yaghmaie en [7] propone la formalización de la trayectoria del esquema relacional, que se convierte en la base para la obtención de reparaciones dimensionales, aquí se muestra que la estrella común relacional y los esquema de copo de nieve multidimensionales para bases de datos, no son la mejor opción para este proceso. Un novedoso sistema TEXplorer permite a los usuarios realizar búsquedas de palabras clave y la agregación de OLAP y de estilo exploración de los objetos de texto en un cubo construido sobre un texto multidimensional de base de datos nos presenta [8]. El modelo llamado GOLD, demuestra que su poder de expresión nos permite considerar todas las peculiaridades de modelado multidimensional es presentado en [9]. F. Carpani, presenta un modelo que permite la especificación detallada de una base de datos multidimensional. Esta especificación se construye mediante un lenguaje gráfico que permite describir las estructuras de datos y algunas restricciones de integridad, y un lenguaje de restricciones de integridad que permite dar una descripción precisa de las relaciones entre los datos en [10]. V. Esch-Bussenmakers en [11], propone un tutorial de un sistema experimental, donde el objetivo del proyecto fue evaluar las diferentes modalidades de acceso (voz, gráficas y modo táctil) para acceder y presentar ciertos tipos de información, y para ciertas estrategias de búsqueda al buscar y navegar en una base de datos musical multidimensional, utilizando un dispositivo móvil simulado.

II. MODELO

Para construir una Base de datos Multidimensional (BDM), se toma la información de las bases de datos donde están los datos de origen, además es posible hacer llegar información de fuentes de datos externas.

Para esto se sigue el proceso *Extracción* de información, *Transformación* de la misma y *Carga* en el almacén de datos (ETL) . En la Fig. 2 se puede observar este proceso.

El proceso de *Extracción* consiste en la obtención de la información de las bases de datos y de fuentes externas.

Transformación consiste en el filtrado de datos, limpieza y consolidación de los mismos.

La C*arga* significa la incorporación de los datos en el almacén de datos.

Enseguida está la capa OLAP la cual nos permite la obtención de información para proporcionar los reportes estadísticos a los usuarios.

Es aquí donde a través de las instrucciones de SQL, como por ejemplo:

- ROLLUP, agrupa datos por alguna dimensión determinada, por ej: analizar los índices de reprobación por materia, a los índices de reprobación por carrera.
- DRILL DOWN, realiza la operación inversa. Muestra la información detallada de cada agrupamiento. Por ej: analizar los índices de reprobación por carrera a los índices de reprobación por materia.
- SLICE, al seleccionar algún miembro en particular de una dimensión se forma una especie de rebanada del cubo original. Por ej: se puede tener los datos de reprobación de las materias de un docente en particular.
- DICE, selecciona varios miembros de varias dimensiones y se forma un subcubo (cubo más pequeño) o dado. Por ej: los índices de reprobación de un docente, en una materia en un año.
- PIVOT, rota el cubo para ver una cara en particular. Por ej: en vez de ver materias vemos docentes.

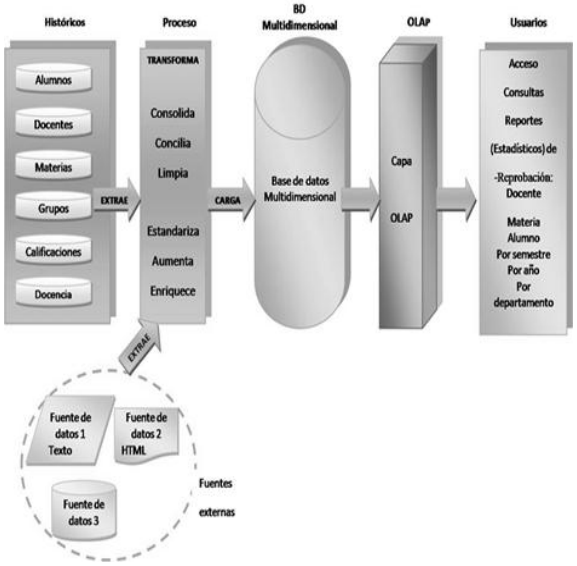


Fig. 2 Proceso de Extracción, Transformación y Carga

III. DESARROLLO

En todos los Institutos Tecnológicos a nivel nacional, existen departamentos académicos que administran las carreras que se ofrecen, en cada uno de ellos existen Jefaturas de docencia que manejan información tales como:

- Dosificaciones de las materias.
- Instrumentaciones didácticas.
- Proyectos en horas de apoyo a la docencia (Proyectos individuales).
- Docentes.
- Listas de calificaciones.
- Índices de aprobación y reprobación.

Actualmente no existe un sistema automatizado que nos permita hacer un registro de datos y posteriormente obtener información, como la que hoy se demanda.

Por lo que no es posible obtener información de control (estadísticas de aprobación, reprobación y deserción) de cada uno de los docentes (por materia, hora de la clase, semestre), de las carreras, del departamento y del Instituto.

Además tampoco es posible obtener fácilmente información de semestres anteriores.

Por lo que la propuesta es la implementación de un esquema de Base de Datos Multidimensional, que nos proporcione la información antes mencionada y con este esquema nos permitirá reducir el espacio de almacenamiento y los tiempos de acceso de la información. En la Fig. 3 se muestra el esquema multidimensional.

Con este esquema la intención es tener la información de todas las carreras que ofrece el Instituto y obtener información para:

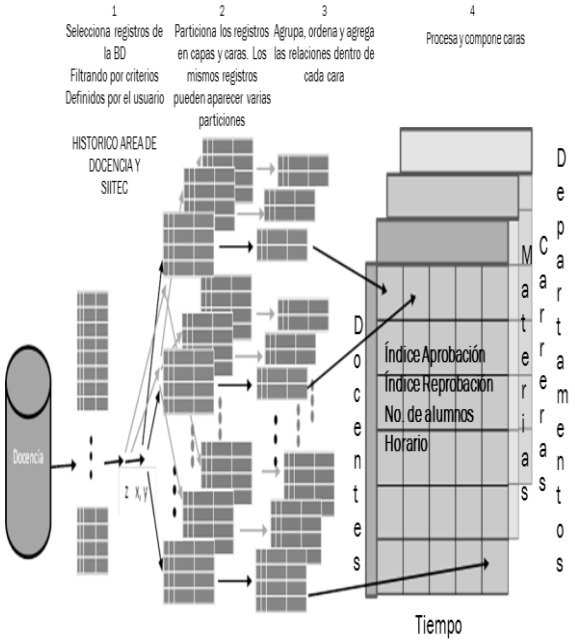


Fig. 3 Big Picture del esquema multidimensional propuesto

- Analizar índices de reprobación (por departamento, carrera, materia, semestre, docente, horario de clase).
- Prever con anticipación problemas en base a la información proporcionada (por ejemplo, horarios con mayor índice de aprobación ó reprobación, horarios inadecuados para impartir ciertas materias).
- Diseñar estrategias que nos permite elevar los índices de eficiencia terminal.

Para el desarrollo de este proyecto la actividad objeto del análisis sería: los índices de aprobación, reprobación, número de alumnos por grupo y horario de la clase. La información registrada por ejemplo sería la siguiente:

Nombre del departamento “Sistemas y Computación, Carrera “Ing. Sistemas Computacionales”, Materia “Programación básica”, Docente “Aguilar Andrade Javier”, , Semestre “ene-jun 2013”, índice de aprobación 95, , índice de reprobación 5, número de alumnos 42, hora “7:00-8:00”.

La estructura del almacén consiste de 5 dimensiones en la Fig. 4 se pueden visualizar estos.

En un esquema multidimensional se representa una actividad que es objeto de análisis (hecho) y las dimensiones que caracterizan la actividad (dimensiones). La información relevante sobre el hecho (actividad) se representa por un conjunto de indicadores (medidas o atributos de hecho), la información descriptiva de cada dimensión se representa por un conjunto de atributos (atributos de dimensión).

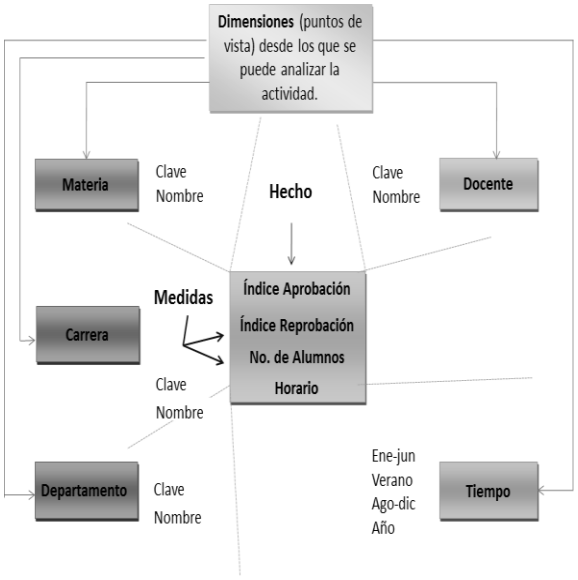


Fig. 4 Dimensiones del Almacén de datos

En este proyecto los hechos son los índices de aprobación, reprobación, número de alumnos y horario de la materia.

Una Base de Datos Multidimensional es básicamente, un Cubo OLAP, que acuña su nombre por su característica multidimensional, ampliando las posibilidades que hasta el momento ofrecen las conocidas hojas de cálculo.

Gracias a la incorporación de las bases de datos de tipo multidimensional, y el nacimiento del nuevo concepto Cubo OLAP, las herramientas de soluciones para sistemas de inteligencia de negocios han avanzado notablemente en cuanto a las prestaciones que estas aplicaciones brindan a las empresas, donde la información confiable, precisa y en el momento oportuno, son uno de los bienes más preciados.

Cada una de las dimensiones que posee la base de datos incorpora un campo determinado para un tipo de dato específico, que luego podrá ser comparado con la información contenida en el resto de dimensiones, para hacer posible la evaluación y posteriores informes de la información realmente relevante para una compañía.

Una base de datos multidimensional puede contener varios cubos o vectores que extenderán las posibilidades del sistema OLAP con el cual se trabaja.

Por ello, si bien en general los sistemas OLAP suelen estar compuestos por tres dimensiones, lo cierto es que existe la posibilidad de que el sistema OLAP albergue más de tres dimensiones mediante la utilización de estos Cubos OLAP.

Las grandes ventajas que presenta este tipo de base de datos multidimensional que incluye Cubos OLAP, es que permite obtener mayor rapidez en las consultas y en el procesamiento de la información.

Modelo de tiempo del sistema actual y el propuesto:

M = (TA, TP, T1, T2, T3, T4, T5) en donde:

TA = Tiempo del sistema actual f: $R \rightarrow R$
T1 = Tiempo de captura en Excel T1 $\in$ a R
T2 = Tiempo de análisis T2 $\in$ a R
T3 = Tiempo de impresión T3 $\in$ a R

TP = Tiempo sistema propuesto $f: R \rightarrow R$
T4 = Tiempo de Extracción, Transformación y Carga (ETL) T4 $\in$ a R
T5 = Tiempo de consulta T5 $\in$ a R

*$TA = T1 + T2 + T3$

* Este tiempo sería por cada semestre y solo obtenemos datos de índices de reprobación.

**$TP = T4 + T5 + T3$

** Este tiempo sería del sistema propuesto donde el tiempo T4 (ETL) solo sería una vez al final del semestre, por lo que el tiempo solo se reduce a T5 + T3 (Consulta e impresión), el cual es mucho menor que el tiempo TA del sistema actual. Con la ventaja además de que las consultas pueden ser por: semestre, año, índice de aprobación, índice de reprobación, no. de alumnos y horario de la materia.

La $f: R \rightarrow R$ es una función de los reales a los reales.

IV. PRESENTACIÓN DE RESULTADOS

El alcance de este proyecto se circunscribe al diseño de los cubos OLAP, para solucionar el problema planteado en la etapa del desarrollo, buscando que se cumplan las metas propuestas que es la disminución de los costos de operación, del almacenamiento y consulta de la información a través de las instrucciones de SQL. En la figura 5 se muestra un ejemplo de una consulta con la instrucción Rollup.

```
SELECT F2 'DOCENTE',F3 'MATERIA',F4 'HORA',F5
'ALUMNOS', AVG(F6) 'APROBACION'
FROM INDICES
GROUP BY F2, F3 , F4, F5,  F6 WITH ROLLUP
```

Fig. 5 Código de la instrucción Rollup

V. CONCLUSIONES

Observando los resultados mostrados previamente consideramos que se cumple el objetivo especificado para esta investigación.

La importancia de este proyecto radica en obtener información de una base de datos especificada como un esquema multidimensional, el cual es analítico y está basado tanto en información actual como histórica, que nos permite la administración eficiente de la información derivada de las tareas docentes del ITC.

Para continuar en lo futuro con esta investigación sugerimos incrementa las dimensiones de los cubos, que nos permita administrar la información de otros tecnológicos agrupados en regiones geográficas del país.

REFERENCIAS

[1]Y. Pitarch and A. Laurent, P. Poncelet *A Conceptual Model for Handling Personalized Hierarchies in Multidimensional Databases*, MEDES 2009, Lyon, France.

[2]M. Golfarelli and S. Rizzi, *Comparing Nested GPSJ Queries in Multidimensional Databases*, DOLAP 2000, VA, USA.

[3]M. Minuto and Alejandro A. Vaisman, *Efficient Intensional Redefinition of Aggregation Hierarchies in Multidimensional Databases*, DOLAP 2001, Atlanta, Georgia, USA.

[4]O. Dankoand T. Skopal,*Elliptic Indexing of Multidimensional Databases*, Australasian Database Conference 2009, Wellington, New Zealand.

[5]W. Liang, María E. Orlowska, and J.X. Yu,*Optimizing multiple dimensional queries simultaneously in multidimensional databases*, VLDB Journal 2000.

[6]C. Stolte, D. Tang, and P. Hanrahan, *Polaris: A System for Query, Analysis, and Visualization of Multidimensional Databases*, ACM, Noviembre 2006.
[7] M. Yaghmaie, L. Bertossi, and S. Ariyan**,***Repair-Oriented Relational Schemas for Multidimensional Databases*, EDBT 2012, Berlín, Germany.
[8] B. Zhao, X.Lin, B. Ding, and J. Han, *TEXplorer: Keyword-based Object Search and Exploration in Multidimensional Text Databases*, CIKM 2011, Glasgow, Scotland, UK.
[9] J. Trujillo and M. Palomar, J. Gómez, *The GOLD Definition Language (GDL): An Object Oriented Formal Specification Language For Multidimensional Databases*, SAC 2000, Como Italy.
[10] F. Carpani, *CMDM: Un Modelo Conceptual para la Especificación de Bases Multidimensionales*, Tesis de Maestría, Agosto 2000, Universidad de la República Uruguay.
[11] V. Esch-Bussenmakers, and A.H.M. Cremers, *User Walkthrough of Multimodal Access to Multidimensional Databases*, ICMI 2004, Pennsylvania, USA.

VI. CURRICULUM VITAE



Pedro Fletes Gudiño, Licenciado en Informática egresado en 1985 del Instituto Tecnológico de Colima, actualmente labora como docente en el Instituto Tecnológico de Colima, es estudiante de la Maestría en Sistemas Computacionales en el Tecnológico de Colima, su área de interés son las Bases de datos y los Sistemas Operativos.

Nicandro Farías Mendoza, estudió la Licenciatura en Física y Matemáticas en el Instituto Politécnico Nacional (IPN). En 2002 se graduó como Doctor en Ciencias en el Centro de Investigación y de Estudios Avanzados del IPN. Actualmente es profesor de tiempo completo del Instituto Tecnológico de Colima. Sus áreas de interés son los Sistemas Formales, Agentes de Software Inteligentes, Tópicos Diversos de Ingeniería en Software.

Eliseo Mendoza Luna es Ingeniero en Sistemas Computacionales, egresado del Instituto Tecnológico de Colima. Actualmente labora en el Departamento de Informática de la empresa Liconsa SA de CV y es estudiante de la Maestría en Sistemas Computacionales en el Instituto Tecnológico de Colima.

Iván Surizaday González Magaña, Ingeniero en Sistemas Computacionales egresado en el 2011 del Instituto Tecnológico de Colima. Actualmente es estudiante de la maestría en Sistemas computaciones en el Instituto Tecnológico de Colima, sus áreas de interés son Redes, Linux y Bases de datos.

Rafael Espinosa Cárdenas, Ingeniero Industrial egresado en 1981 del Instituto Tecnológico de Colima. Actualmente profesor de tiempo completo y estudiante de la Maestría en Sistemas Computacionales en el Instituto Tecnológico de Colima, sus áreas de interés son la Programación de dispositivos móviles, Métodos Numéricos y Simulación.