BODEGAS DE DATOS Y OLAP EN UNICAUCA VIRTUAL

MARTHA MENDOZA
Ingeniera de Sistemas, Magíster en Informática
Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones
Universidad del Cauca
mmendoza@unicauca.edu.co

CARLOS COBOS
Ingeniero de Sistemas, Magíster en Informática
Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones
Universidad del Cauca
ccobos@unicauca.edu.co

JAIME MUÑOZ
Estudiante de Ingeniería de Sistemas Décimo Semestre
Programa de Ingeniería de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones
Universidad del Cauca
jaimem@unicauca.edu.co

LISANDRO ACOSTA
Estudiante de Ingeniería de Sistemas Décimo Semestre
Programa de Ingeniería de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones
Universidad del Cauca
lisandroam@unicauca.edu.co

LUIS CARLOS GOMEZ FLOREZ
Ingeniero de Sistemas, Magíster en Informática
Escuela de Sistemas, Facultad de Ingenierías Físico Mecánicas
Universidad Industrial de Santander
lcgomezf@uis.edu.co

RESUMEN

Fecha Recepción: 7 de marzo de 2006
Fecha Aceptación: 13 de junio de 2006

El análisis de información se ha convertido en uno de los procesos más importantes de la mayoría de empresas que manejan gran cantidad de clientes. Ellas, han encontrado en la información generada por sus propios sistemas, conocimiento valioso que les puede ayudar a mejorar sus procesos internos, mejorar la relación con los clientes, ser más competitivos, minimizar costos e incrementar utilidades entre muchas otras cosas. Sin embargo, no solamente en las empresas grandes, con procesos industriales y gran cantidad de clientes es útil el análisis de información, existen casos exitosos del uso de tecnologías de análisis en áreas como el transporte, la medicina, la biología, la química, la genética y por supuesto la educación. Es así como en la Universidad del Cauca, apoyados por el Grupo de I+D en Tecnologías de la información y Colciencias, se ha iniciado un proyecto en el cual se pretende implementar un Sistema de Soporte a la toma de Decisiones (DSS), que permita a directivos y profesores de Unicauca Virtual (Universidad Virtual del Cauca), utilizar la potencialidad de las tecnologías Data Warehouse y OLAP en el análisis de información generadas en la interacción del estudiante y el sistema gestor de aprendizaje (LMS) de Unicauca Virtual. Este artículo presenta el proceso de desarrollo de este proyecto haciendo énfasis en el modelado y la construcción de la bodega de datos en el motor de base de datos Oracle 10g Release 2.

PALABRAS CLAVE: Modelado dimensional, bodega de datos, herramienta OLAP, ciclo de vida, esquema estrella

ABSTRACT

The analysis of information has become one of most important processes in major enterprises who count with a high number of customers. These Enterprises, have found in information generated by its own systems, valuable knowledge that might help to improve inner process, customers relationship, competitiveness, minimize costs and increase revenue, etc. However, not only major enterprises with industrial process and great quantity of customer, find useful the analysis of information; there are successful cases in use of this kind of technology in other areas like transport, health, biology, chemistry, genetic, and education. This is why in the University of Cauca, sponsored by The Information Technology Group and Colciencias, has begun a project that want to implement a Decision Support System (DSS), that let directives and teachers of Unicauca Virtual, use the data warehouse potentiality and OLAP technologies to analyze information generated between students and the Learning Management System (LMS) of Unicauca Virtual. This article, present the development process of this project emphasizing in modeling and construction of the data warehouse in the Oracle 10g Release2 database engine.

KEYWORDS: *Dimensional Modeling, Data Warehouse, Online Analytical Processing (OLAP), Life Cycle, Star Schema.*

INTRODUCCIÓN

Muchas empresas utilizan sistemas de información que soportan toda su operación diaria, es decir, todas las transacciones que se pueden generar en una jornada normal de trabajo. Estos sistemas permiten gestionar datos, hacer los procesos del negocio más fácil y se caracterizan por recolectar gran cantidad de información.

La búsqueda de formas para aprovechar la información almacenada en las bases de datos transaccionales, ha llevado a diferentes expertos a plantear nuevas metodologías y estándares centrados en el análisis de información, conocidos en el ambiente tecnológico como Bodegas de Datos (Data Warehouse), Procesamiento Analítico en Línea (OLAP) y Minería de Datos (Data Mining). Todos estos terminan convirtiéndose en sistemas informáticos muy robustos denominados Sistemas de Soporte para la toma de Decisiones (DSS), que son muy utilizados por empresas en distintas áreas como: ventas, transporte, seguridad, sector financiero, medicina, educación, entre otras.

Unicauca Virtual siendo un ambiente de aprendizaje virtual, planeó incluir en su segunda fase un servicio en el cual se utilicen estas tecnologías como soporte al análisis de las diferentes actividades que realizan los estudiantes dentro del ambiente y que ayuden a tomar decisiones importantes que pueden afectar el desempeño del estudiante dentro del sistema.

El proyecto del cual trata este articulo, pretende en términos mas específicos, en primera instancia, diseñar y construir una bodega de datos que permita almacenar información relacionada con las actividades académicas que desarrolla el estudiante en Unicauca Virtual y que será utilizada por el sistema de soporte a decisiones.

En segunda instancia construir un prototipo de una herramienta OLAP que permita obtener y procesar información contenida en la bodega de datos, para apoyar la toma de decisiones estratégicas de los docentes y directivos de Unicauca Virtual. Este prototipo tendrá las siguientes funcionalidades:

* Gestionar un conjunto de consultas analíticas (se refiere a un conjunto de informes o reportes que presentan información en un formato determinado, por ejemplo: graficas, estadísticas y/o, tablas) predefinidas que cumpla con las necesidades de información requerida para apoyar la toma de decisiones estratégica de los usuarios.

* Gestionar nuevas consultas analíticas definidas por los usuarios.

Este proyecto se enmarca dentro del macroproyecto Unicauca Virtual planteado por el Grupo de I+D en Tecnologías de la información (GTI) de la Universidad del Cauca, que es cofinanciado por Colciencias.

CONTEXTO TEÓRICO

Para la construcción de un DSS, se deben tener claro conceptos como: Bodegas de datos, Metodología de Desarrollo de la Bodega de Datos, y OLAP. Las definiciones de estos conceptos se presentan a continuación.

SISTEMAS DE SOPORTE PARA LA TOMA DE DECISIONES (DSS)

Se puede pensar en los sistemas de soporte para la toma de decisiones como sistemas que ayudan en el análisis de

información de negocios. Su propósito es ayudar a la administración para que "marque tendencias, señale problemas y tome.... decisiones inteligentes" [1].

Sin embargo, no se pueden vincular los DSS solamente al mundo de los negocios ya que existen diferentes escenarios que involucran también la toma de decisiones y en los cuales resultaría muy útil analizar información como en la genética o en la educación, como es el caso de Unicauca Virtual.

BODEGAS DE DATOS

No se puede hablar de DSS si no se incluyen las bodegas de datos, las cuales han sido concebidas como repositorios de gran cantidad de datos que se encuentran organizados para optimizar la realización de consultas y la obtención rápida información. Una definición formal, dada por W.H.Inmon, expresa que "una bodega de datos es una colección de datos integrados, orientados a temas, que dan soporte a las funcionalidades del DSS, donde cada unidad de dato es relevante en algún momento en el tiempo" [2].

Por lo general, las bodegas de datos se constituyen por unidades menores de estructuras de datos o subconjuntos lógicos denominados data marts, como se puede apreciar en la **Figura 1a**. Es decir, un data mart es una pequeña bodega de datos, que al unirse con otros data marts conforman la gran bodega de datos.

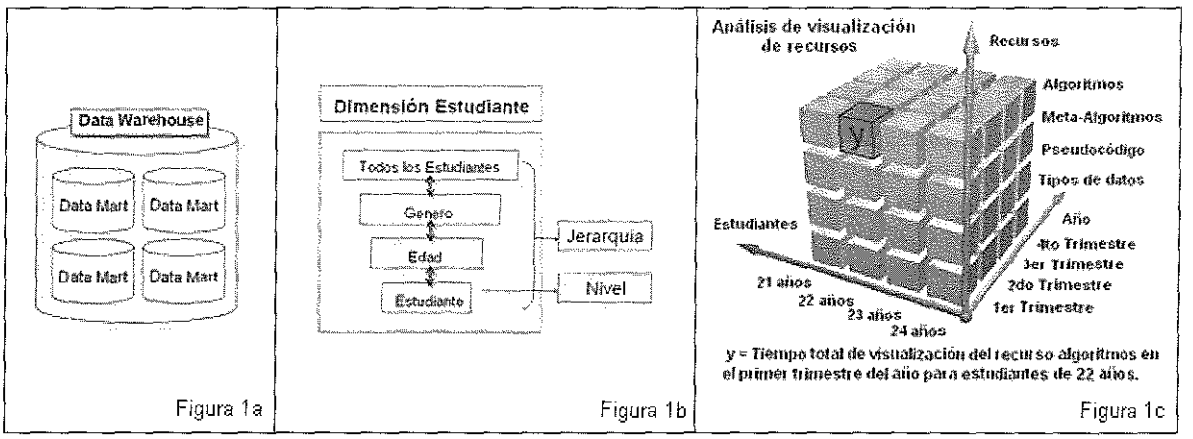

Figura 1a Representación de la formación de una bodega de datos con unidades más pequeñas llamadas "Data Marts". Figura 1b. Representación de una jerarquía dentro de una dimensión. Figura 1c. Representación de un cubo con sus dimensiones como ejes y las medidas.

dimensionales cuyos principales componentes son: tablas de hechos, medidas o hechos, dimensiones, jerarquías, y niveles.

Una tabla de hechos es una gran tabla dentro del data mart que almacena medidas de negocio. La tabla de hechos representa datos, usualmente numéricos y aditivos, que pueden ser analizados y examinados (medidas). Una tabla de hechos usualmente posee dos tipos de columnas: aquellas que contienen hechos y otra que tiene las llaves foráneas hacia otras tablas [3].

Las dimensiones son estructuras de datos, a menudo compuesta de una o más jerarquías. Sus atributos, ayudan a describir el valor de la(s) medida(s) en la tabla de hechos; son normalmente valores textuales y junto a las tablas de hechos, permiten responder a preguntas del negocio [3].

Las jerarquías se componen de diferentes niveles de detalle, cada valor tiene una relación padre-nivel hasta el nivel mas bajo de la jerarquía (ver Figura 1b).

Todos estos componentes conforman lo que comúnmente se denomina cubo o hyper-cubo (si tiene más de tres dimensiones), donde cada dimensión es un eje del mismo (ver **Figura 1c**).

Procesamiento Analítico en Línea (Olap)

El termino OLAP fue acuñado por el Dr. E. F. Codd para describir una tecnología que forma un puente entre la computación personal y el manejo empresarial de datos. Los modelos relacionales fallaban al suministrar capacidades analíticas para gerentes, analistas y ejecutivos, Codd reconoció la necesidad de un modelo multidimensional adicional que permitiera análisis más rápidos y de mayor capacidad de las crecientes bases de datos relacionales [6].

Formalmente se puede pensar en OLAP como la actividad general de consultar y presentar datos numéricos y textuales de una bodega de datos [3], o como el proceso interactivo de crear, mantener, analizar y realizar informes sobre datos, añadiendo que los datos en cuestión son percibidos y manejados como si estuvieran almacenados en un arreglo multidimensional [1].

En el mercado existen diferentes implementaciones OLAP, las más conocidas son:

* **ROLAP (Relational OLAP):** Puede ser definida como un conjunto de aplicaciones e interfases que le dan a las bases de datos relacionales una tratamiento multidimensional [3], es decir OLAP sobre una base de datos relacional. Algunas características de ROLAP son [3] [7]:

* **Puede manejar gran cantidad de datos:** La limitación del tamaño de datos en la tecnología ROLAP depende de los límites de la cantidad de datos que la base de datos pueda manejar.

* **Puede apoyarse en funcionalidades inherentes de la base de datos relacional:** A menudo, las bases de datos relacionales ya vienen con un contenedor de funcionalidades. Las tecnologías ROLAP, desde que se ubicaron sobre las bases de datos relacionales, pueden apoyarse en estas funciones.

* **El desempeño puede ser bajo:** Porque cada reporte ROLAP es esencialmente una consulta SQL (o múltiples consultas SQL) en la base de datos relacional, la duración de la consulta puede ser larga si el tamaño de la base de datos es grande.

* **Limitado para funcionalidades SQL:** La tecnología ROLAP trabaja principalmente con la generación de sentencias SQL para consultar la base de datos relacional y las sentencias SQL no cubren con todas las necesidades de análisis, por consiguiente es necesario que la base de datos cuente con extensiones especiales del lenguaje SQL para el manejo de OLAP relacional..

* **MOLAP (Multidimensional OLAP):** Suministra capacidad de análisis de datos almacenados en un sistema de base de datos multidimensional (MDBS) [6]. Algunas características son [3] [7]:

* **Mayor desempeño:** Los cubos MOLAP son construidos para realizar una rápida recuperación de datos.

* **Requieren inversión adicional:** La tecnología es usualmente propietaria y aun no es muy común en las organizaciones. Por consiguiente, adoptar la tecnología MOLAP, involucra necesidades adicionales de inversión en recursos humanos y tecnológicos.

Metodología de Desarrollo de la Bodega de Datos

Durante la etapa de investigación se encontraron varios autores quienes proponen metodologías para el desarrollo de una bodega de datos, el más reconocido es R. Kimball [3], quien brinda conceptos necesarios para llevar acabo el análisis y diseño de la bodega de datos planteada para Unicauca Virtual. El método presentado por Kimball (ver Ciclo de vida de la bodega de datos) además de estar extensamente expuesto en su bibliografía, se apoya en gran variedad de ejemplos que facilitan la comprensión de los conceptos relacionados con diseño de bodegas de datos.

Existen otras metodologías que aunque no son tan conocidas podrían convertirse en buenas alternativas a la metodología propuesta por R. Kimball, como por ejemplo la propuesta por Sergio Lujan Mora y Juan Trujillo denominada "Un método global basado en UML para el

diseño de Almacenes de Datos" [4] pero que aun carecen de reconocimiento practico o se encuentran en fase de desarrollo.

Ciclo de Vida de la Bodega de Datos

La metodología de Kimball, presenta un marco de trabajo ilustrado en la **Figura 2**, en la cual se muestran las diferentes etapas durante todo el proceso de creación de la bodega.

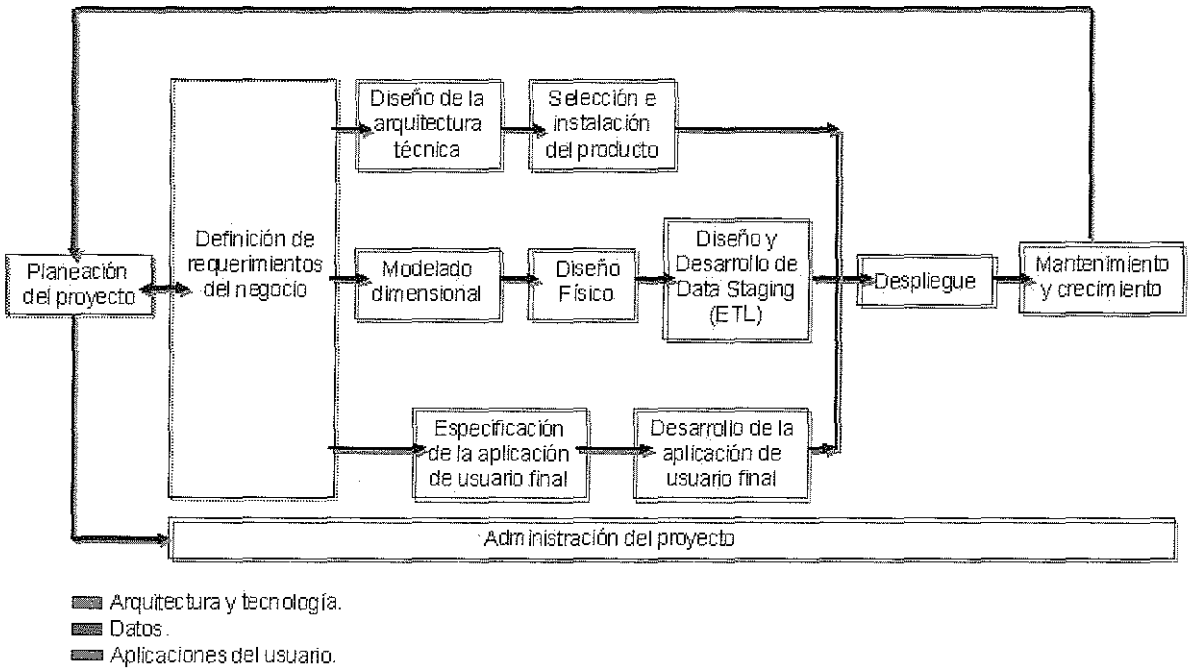

Figura 2. Ciclo de vida para la construcción de una bodega de datos según Ralph Kimball.

La fase de planeación del proyecto, pretende establecer la definición y el alcance del proyecto de la bodega de datos, incluyendo la valoración y justificación del negocio. La fase de definición de requerimientos del negocio es donde se establece la base relacionada con la tecnología, los datos y las aplicaciones del usuario.

La ruta de mayor importancia es la relacionada con los datos, en la cual se realiza el modelado dimensional, partiendo de los requerimientos obtenidos y de las necesidades de análisis de los usuarios; el diseño físico, el cual se enfoca en definir las estructuras físicas necesarias para soportar el modelado dimensional; y la etapa ETL (Extract-Transform-Load Data Staging) en la cual se diseña y desarrollan procesos para extraer, transformar y cargar datos.

Además se deben considerar otras rutas como la de tecnología en la cual se establece la arquitectura y visión general del marco de trabajo y la ruta de aplicación en la cual se definen un conjunto de aplicaciones de usuario final para consulta de datos.

Por último, a lo largo de todo el ciclo de vida se debe seguir una administración general del proyecto la cual asegura que todas las actividades del ciclo de vida se alcancen y se sincronicen.

Asimismo, este autor basado en su experiencia en la construcción de bodegas de datos, plantea una serie de soluciones para tener en cuenta en algunos casos especiales que se

presentan en el momento de realizar el modelado dimensional. Estas soluciones ayudan a garantizar una buena representación de los requerimientos del negocio señalando conceptos relacionados con el diseño de los datos, la arquitectura y la implementación. Estos conceptos se exponen detalladamente por el autor en dos de sus libros [3][5].

MODELADO DIMENSIONAL: El modelado dimensional es una técnica de diseño lógico que busca presentar los datos en un marco de trabajo estándar que es intuitivo y permite acceso de alto desempeño. Es inherentemente dimensional y se adhiere a una disciplina que usa el modelo relacional con restricciones de consideración. Cada modelo dimensional esta compuesto de una tabla con una llave múltiple llamada tabla de hechos y un conjunto de tablas llamadas tablas dimensión. Cada tabla dimensión esta compuesta por una llave simple que corresponde exactamente a uno de los componentes de la llave múltiple en la tabla de hechos. Esta estructura característica, similar a una "estrella" es a menudo llamada "Esquema Estrella" [3] (ver **Figura 3**).

Bodegas De Datos Y Olap En Oracle

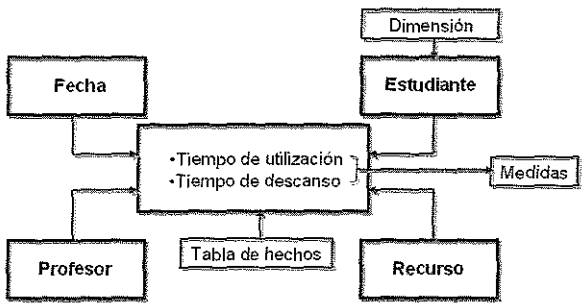

Figura 3. Esquema estrella.

La bodega de datos del presente proyecto se implementará en el motor de base de datos relacional Oracle; que desde la versión 9i extiende su funcionalidad para permitir el manejo de bases de datos multidimencionales de manera integrada. Como ya se menciono, existen dos formas de crear bodegas de datos: sobre una base de datos relacional (ROLAP) y sobre una base de datos multidimensional (MOLAP).

ROLAP

Si se desea implementar una bodega de datos relacional en Oracle, se cuenta con las siguientes opciones:

* **Oracle Warehouse Builder (OWB) [8].** Suministra un completo entorno de trabajo para realizar un proyecto desde la definición misma de las tablas relacionales hasta la creación de los objetos lógicos, como por ejemplo dimensiones, jerarquías, niveles, cubos, medidas, etc. Además, presenta un entorno grafico que permite realizar el proceso de extracción, transformación y carga de datos desde el sistema transaccional al esquema de la bodega de datos. Una desventaja del OWB es que carece de la creación directa de los metadatos CWM (Common Warehouse Metamodel) para los objetos.

* **Oracle Enterprise Manager (OEM)** en las versiones de escritorio y Web. Permite crear los objetos lógicos con sus correspondientes metadatos CWM [9], es decir, quedan visibles para que aplicaciones de terceros que soporten el estándar puedan trabajar con los objetos definidos. La consola de escritorio necesita un componente adicional (parche) para que queden completamente habilitadas las opciones de bodegas de datos. Estos componentes no son necesarios si se usa el OEM versión Web donde no se presentan restricciones de ningún tipo y permite crear todos los objetos lógicos necesarios.

MOLAP

Si se desea implementar una bodega de datos multidimensional en Oracle, se tienen las siguientes opciones:

* **Analytic Workspace Manager (AWM) [10].** Permite crear la definición de los objetos multidimensionales y además realizar la carga de datos. Los objetos creados desde el AWM quedan con sus correspondientes metadatos CWM. Usando el AWM, la creación de los objetos dimensionales (dimensiones, jerarquías, niveles, cubos, medidas, etc.) se realiza de una manera muy cómoda ayudado por los asistentes. Sin embargo, si se encuentran problemas en el momento de hacer el mapeo de los datos, se presentan dificultades para corregirlos debido a que no es fácil identificar el motivo de los errores. Los objetos que se crean en la "vista de diseño" (AWM presenta dos opciones de implementación, una basada en la vista de diseño y otra basada en la vista de objeto) [10] del AWM cumplen con la forma estándar para espacios de trabajo analíticos, los cuales, tienen una gran cantidad de extensiones que no están documentadas para el uso público, dificultando así la comprensión de la definición de los objetos.

* **OLAP DML [12].** Lenguaje de manipulación de objetos multidimensionales de Oracle que extiende las capacidades de análisis de lenguajes de consulta como el SQL. La implementación directa en lenguaje OLAP

DML no es sencilla, teniendo en cuenta que es un lenguaje relativamente nuevo, extenso y la bibliografía que se encuentra al respecto es bastante genérica. Lo que se recomienda según expertos [11] en el manejo de bodegas de datos en Oracle, es usar AWM para la definición de los objetos más genéricos como dimensiones, jerarquías, etc. y el lenguaje OLAP DML para definir objetos más complejos y específicos como cálculos o algún tipo de operación de pronósticos.

¿ROLAP O MOLAP?

Para muchos administradores de bases de datos, las tablas relacionales y el SQL representan un ambiente de desarrollo familiar. Por otro lado, los espacios de trabajo analíticos requieren transformación de datos y aprender nuevos conceptos. ¿Entonces porque utilizar espacios de trabajo analítico? La respuesta es: las grandes capacidades de análisis y el desempeño en tiempo de ejecución representan en muchos casos el mejor soporte para la toma de decisiones. Sin embargo, los esquemas relacionales son la mejor decisión en algunas situaciones, por ejemplo, si el almacén de datos soporta la creación de reportes rutinarios con carácter exploratorio, y ese uso aparenta ser estable.

Los espacios de trabajo analíticos son una buena opción para los siguientes requerimientos:

* Consultas ad-hoc o "no genéricas" de todas las áreas de datos.
* Cálculos avanzados como modelos, pronósticos, radios de crecimiento y tendencias.
* Escenarios "que pase si".
* Cálculos en series de tiempo como retardos, elementos principales, promedios de movimiento.
* Cálculos complejos en tiempo de ejecución.
* Alto desempeño para calcular resúmenes de datos.

Los esquemas relacionales se pueden preferir si se tienen los siguientes requerimientos:

* Patrones de consulta predecibles y reportes preparados.
* Cálculos no avanzados.
* Cálculos complejos en tiempo de ejecución infrecuentes.
* Medidas con un gran número de dimensiones.
* Dimensiones con pocos niveles de agregación.

MODELOS DIMENSIONALES PARA UNICAUCA VIRTUAL

Para identificar los temas de negocio (data marts) de este proyecto, se utilizó la técnica de priorización de requerimientos del negocio[3], estos temas son:

* *Comportamiento de los estudiantes,* con la cual se pretende ofrecer información de las actividades que el estudiante realiza dentro del LMS.
* *Estilos de aprendizaje,* que junto con el tema de evaluación y el tema de comportamiento de los estudiantes, pretende señalar la relación que existe entre los estudiantes y los estilos de aprendizaje asociados a ellos.
* *Evaluación,* este tema permitirá que los usuarios de la bodega de datos, obtengan información relacionada con las evaluaciones que involucraban al estudiante y los temas del contenido que cursa.

Inicialmente se han enfocado todos los esfuerzos en el tema de comportamiento de los estudiantes, donde se han identificado una serie de indicadores que hablan de la interacción del estudiante con el LMS. Algunos de ellos son:

* *Tiempo de utilización:* permite medir el tiempo utilizado en visualizar un determinado recurso.
* *Tiempo de descanso:* permite medir el tiempo que se toma como descanso cuando se encuentra visualizando un recurso (Objeto de Contenido Compartible-SCO ó Medio-ASSET).
* *Indicador de visualización:* indica que el estudiante visualizo un recurso en un momento dado, sirve para calcular el número de veces que se ha visualizado un recurso.
* *Tiempo total:* es el tiempo total medido en la interacción del estudiante con el recurso (tiempo de utilización más tiempo de descanso).

En este momento se ha definido un esquema estrella para el tema de comportamiento de los estudiantes, que incluye las siguientes dimensiones (ver **Figura 4**):

* *Estudiante:* esta dimensión permite mantener información descriptiva del estudiante que forma parte de Unicauca Virtual.
* *Profesor:* esta dimensión permite mantener información descriptiva del profesor dentro de Unicauca Virtual.
* *Recurso:* en esta dimensión se mantienen datos

descriptivos de los recursos (SCO ó ASSET). Los recursos son las unidades de información dentro de Unicauca Virtual con los cuales el estudiante interactúa directamente para llevar a cabo su proceso de aprendizaje.

* *Actividad:* esta dimensión mantiene información que describe ampliamente cada una de las actividades o temas que se encuentran en las estructuras de contenidos y que el estudiante tiene la posibilidad de revisar.
* *Curso:* conserva información descriptiva de las asignaturas y grupos que se forman para dictar las clases dentro de Unicauca Virtual.
* *Fecha:* guarda datos descriptivos de la fecha de calendario de los hechos ocurridos dentro del ambiente. Esta dimensión permite obtener información de aspectos importantes ocurridos a través del tiempo.
* *Periodo:* esta dimensión permite mantener datos descriptivos de los diferentes periodos académicos que se presentan dentro del ambiente de aprendizaje virtual.
* *Localidad:* esta subdimensión almacena datos descriptivos sobre las distintas localidades que se manejan para los usuarios del ambiente virtual.

Con este esquema estrella se pretende que los docentes y directivos puedan responder preguntas como:

* Tiempo de descanso que se toma un estudiante en la revisión de un recurso o un conjunto de recursos.
* El número de recursos visualizados por un estudiante durante un periodo de tiempo dado (semana, mes, trimestre, semestre, año).
* Número de recursos que visualiza un estudiante en un periodo académico.
* Tiempo que utilizan los estudiantes de un determinado curso para revisar un recurso.
* Tiempo que pierden los estudiantes de un curso determinado revisando un recurso.
* Tiempo empleado por los estudiantes para revisar una actividad determinada.

Implementación de los Modelos Dimensionales en Oracle

Como parte de la exploración tecnológica realizada en el proyecto, se probó la implementación del esquema estrella de la **Figura 4**, tanto en un ambiente multidimensional como en uno relacional. La realización de estas pruebas ayudo a identificar que era necesario una base relacional sobre la cual soportar la estructura de los objetos multidimencionales, es decir, un esquema relacional distinto al OLTP donde se encuentren los datos "limpios" y estructurados obtenidos después de realizar el proceso ETL.

Después de realizar las pruebas y tomando como base las recomendaciones presentadas por Oracle en el momento de decidir que tipo de implementación (ROLAP o MOLAP) es recomendable para la construcción del sistema [10], se optó por implementar el modelo dimensional visualización recurso enteramente relacional. Los puntos a favor de esta decisión son los siguientes:

* El proyecto de bodega de datos para Unicauca Virtual, representa una etapa inicial de lo que puede ser en un futuro un sistema de soporte para la toma de decisiones completo, que incluya todas las necesidades de análisis que se puedan presentar una vez el ambiente virtual se encuentre en fase de producción. Por lo tanto, es recomendable realizar una implementación relacional que soporte las necesidades de análisis que hasta el momento se presentan, las cuales, no se encuentran definidas completamente y carecen de complejidad.
* Por el estado actual del macroproyecto Unicauca Virtual, el modelo dimensional obtenido hasta el momento y los que se obtendrán al terminar el proyecto, están sujetos a modificaciones, producto de las nuevas necesidades de análisis de los usuarios finales. Estas modificaciones permitirán extender el significado de los modelos y la cantidad de información disponible para análisis.

El resultado después de la exploración tecnológica fue el esquema dimensional visualización recurso (ver **Figura 4**) implementado completamente en Oracle Enterprise Manager Web de forma relacional con sus correspondientes metadatos. La implementación involucra la creación de cada una de las dimensiones, jerarquías, cubos y medidas.

Prototipo De Herramienta Olap Para Consultar Datos

En el mercado existen diferentes herramientas OLAP, que permiten consultar la información de la Bodega de Datos. Oracle, por ejemplo, "Oracle Business Intelligence Spreadsheet Add-In", la cual se adiciona a Microsoft Excel y permite a los usuarios finales desplegar y navegar a través de los datos Oracle OLAP utilizando una hoja de cálculo de Excel.

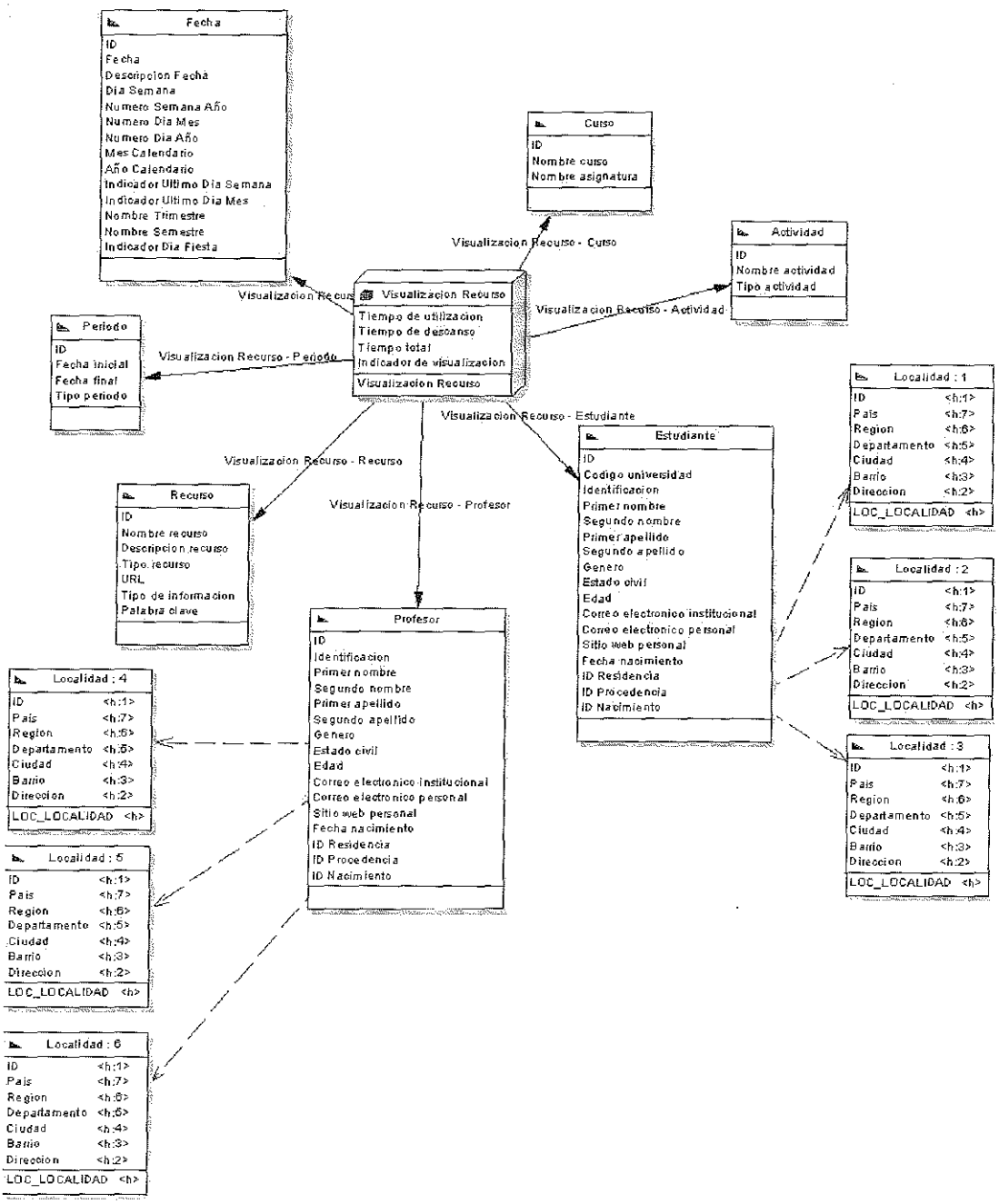

Figura 4. Modelo dimensional obtenido para Unicauca Virtual.

Otras casas de software como "CNS International" o "Cognos", líderes en inteligencia de negocios, ofrecen soluciones OLAP muy completas, "Data Warehouse Explorer" por ejemplo, para acceder a datos basados en sistemas Microsoft, o "Power Play" para otras bases de datos como Oracle, Microsoft, SAP BW e IBM. Aunque existen muchas soluciones, la mayoría de ellas involucran elevados costos adicionales, no estimados para la elaboración de este proyecto. El mercado también ofrece otras alternativas enmarcadas dentro del software libre como PALO [13], Mondrian [14] y JPALO [15] que a pesar de no representar incrementos en los costos del proyecto no permiten conectividad con el motor de base de datos ORACLE. Dado lo anterior el grupo del proyecto decidió iniciar el desarrollo de una aplicación prototipo de herramienta OLAP, que se encuentre integrada a los demás

módulos de la Fase 2 de Unicauca Virtual y que permite consultar los datos que se encuentran dentro de los esquemas dimensionales implementados en Oracle 10g Release 2.

Aunque Oracle ofrece una Interfase de Programación de Aplicaciones Java (API) para OLAP, es decir un conjunto de clases que son útiles para quienes quieren desarrollar aplicaciones que ejecutan procesamiento analítico en línea, la plataforma de desarrollo elegida por el proyecto es "Microsoft Visual Studio .NET", utilizando el proveedor de datos Oracle (ODP para .NET) para tener acceso a los datos. Todo esto debido a que el proyecto se encuentra enmarcado en un macroproyecto, el cual ha elegido este ambiente de desarrollo.

El prototipo se encuentra en su primera fase de construcción, en la que se hizo necesario utilizar algunas vistas del motor Oracle, para consultar y desplegar los metadatos de información de los esquemas estrella implementados (ver Figura 5). Aquí se puede notar las carpetas de medidas, cuyo objetivo es agrupar medidas relacionadas, los cubos y los elementos que lo forman (dimensiones y medidas).

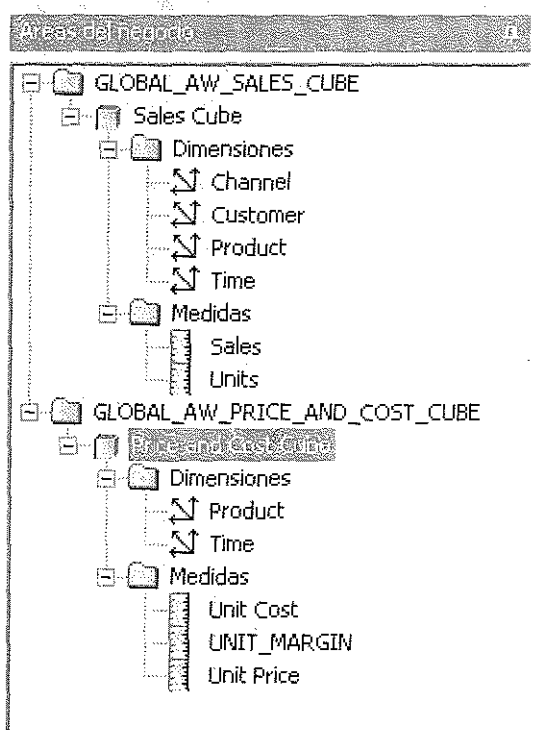

Figura 5. Metadatos de los modelos dimensionales implementados a través del prototipo OLAP.

En la segunda fase se implementará el modulo de consulta de datos sobre una grilla, que permita realizar el "Drill Down" y "Roll Up", es decir bajar y subir a través de los diferentes niveles en las jerarquías de la dimensiones incluidas para el análisis. Para esto es necesario, construir un modulo que forme de acuerdo a los diferentes parámetros enviados la sentencia de consulta sobre el esquema implementado.

Para la fase final, se debe incluir un modulo que permita visualizar los datos de forma grafica para brindar un mayor entendimiento de los mismos.

Problemas Encontrados Durante la Realización del Proyecto

Desde las etapas iniciales se han presentado diferentes inconvenientes que se describirán a continuación:

* Las herramientas de Oracle disponibles para implementar una bodega de datos, no soportan la especificación de algunos casos especiales que surgen en el modelado dimensional. En el proyecto, el hecho de no poder representar casos como relaciones de muchos a muchos entre dimensiones y tablas de hechos y personalización de hechos y dimensiones, obligó a la creación de nuevos modelos dimensionales reduciendo la posibilidad de obtener información más significativa para los usuarios de la bodega.

* Un aspecto importante a tener en cuenta al momento de construir una Bodega de Datos, es contar con la base de datos que soporta el proceso operacional del negocio. En el caso de Unicauca Virtual, solamente una parte del sistema OLTP se ha definido hasta el momento, lo cual limita la definición de nuevos temas y modelos dimensionales útiles para los usuarios.

* Oracle 10g Release 2, permite la implementación de las Bodegas de Datos de dos formas: Relacional y Multidimensional. Aunque se ha hecho un gran esfuerzo por brindar una solución fuerte que soporte la inteligencia de negocios, la etapa de transición desde la versión 9i a 10g produjo una serie de problemas:

* ODP NET (proveedor de datos de Oracle para Visual Studio .NET) para Oracle 10g R1 no funcionaba correctamente. Después de instalado no se podía conectar a la base de datos desde Visual Studio .NET utilizando este componente. Este error solo se corrigió hasta la realización por parte de Oracle de la versión R2 de su motor de base de datos con su correspondiente ODP NET.

* El contenido de alguna documentación que se encuentra en el sitio oficial de Oracle sobre el tema, difiere entre versiones de la base de datos, haciendo que algunos de los ejemplos que se encuentran no puedan ser implementados a través de las herramientas. Además, esta documentación presenta el contenido de forma muy general limitando el entendimiento de algunos de los casos especiales que se presentan durante el desarrollo de una Bodega de Datos como por ejemplo el manejo de subdimensiones, relaciones de muchos a muchos entre la tabla de hechos y las dimensiones, dimensiones y tablas de hechos personalizados y tablas puente para el manejo de jerarquías.

Conclusiones y Trabajo Futuro

La exploración tecnológica es una excelente alternativa cuando se inicia un proyecto relacionado con bodegas de datos y OLAP y de gran utilidad cuando no se cuenta con el escenario ideal para su desarrollo. La exploración tecnológica no solo sirve para conocer a fondo la tecnología en la que se implementará la bodega de datos y se desarrollarán las consultas OLAP, sino que también facilita el entendimiento de los conceptos teóricos generales relacionados con la tecnología.

Guiar el modelado dimensional por medio de una metodología planteada para este tipo de modelado es muy importante, sobretodo para el diseño dimensional, sin embargo, se debe tener en cuenta que la tecnología de bodegas de datos que se va a utilizar para construirla, permita implementar los casos que se modelan en el diseño, o de otra forma redefinir el modelado de acuerdo a las limitaciones de la tecnología usada.

La experiencia obtenida al desarrollar el proyecto, mostró que cuando se esta iniciando el desarrollo de este tipo de proyectos es importante, empezar a modelar los temas del negocio de los que se tiene mayor información. Además una vez seleccionado los hechos o medidas, es aconsejable iniciar con un modelado básico, sin pretender de una vez vislumbrar e incluir todos los casos posibles que se obtienen de un modelado dimensional (hechos conformados, relaciones muchos a muchos, jerarquías, subdimensiones, minidimensiones, roles, etc.). Estos se pueden incluir en un próximo "prototipo" que rápidamente puede convertirse en un modelo dimensional completo.
Dentro del trabajo futuro, se puede decir que es necesario identificar e implementar otros modelos dimensionales que permitan conocer más sobre el comportamiento del estudiante y su experiencia dentro del ambiente virtual. Además, el prototipo de herramienta OLAP debe incrementar sus funcionalidades facilitando la interacción del usuario con los datos contenidos en la bodega de datos. Algunas de estas funcionalidades están relacionadas con la realización dinámica de consultas, la presentación matricial y grafica de los datos, la gestión de consultas realizadas, el filtrado de datos y el cambio de jerarquías en las dimensiones.

AGRADECIMIENTOS

A COLCIENCIAS por el apoyo entregado a través de su proyecto de I+D en ETI titulado "Unicauca Virtual Fase II" con código 1103-14-14897.

BIBLIOGRAFÍA

[1] C. J. Date, Introducción a los Sistemas de Bases de Datos, Apoyo para la toma de decisiones, 21 (México: Pearson Education, 2001, 694 - 724).

[2] W. H. Inmon. Building the Data Warehouse. (United States of America: Wiley Computer Publishing, 1996).

[3] R. Kimball, L. Reeves, M. Ross, W. Thornthwaite, The Data Warehouse Lifecycle Toolkit. (United States of America: Wiley Computer Publishing, 1998).

[4] S. L. Mora, J. Trujillo. Un método global basado en UML para el diseño de Almacenes de Datos. Actas de las VIII Jornadas de Ingeniería de Software y Bases de Datos. 2003. 539- 549.

[5] R. Kimball, M. Ross. The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling. Wiley Computer Publishing. 2002.

[6] G. Dodge, T. Gorman, Essential Oracle8i Data Warehousing, Analytical Processing in the Oracle Data Warehouse, 13 (United States of America: Wiley Computer Publishing, 2000, 809 - 812).

[7] 1keydata, (Visitado 2005, Marzo 03). Información relacionada con conceptos de OLAP, ROLAP y MOLAP. [Documento WWW]. URL http://www.1keydata.com/datawarehousing/datawarehouse.html.

[8] S. Alison, K. Nayar, M. Bird, J. Stein, Oracle Warehouse Builder User's Guide 10g Release 1 (10.1), 2003, 2 - 28.

[9] D. T. Chang, (Visitado 2004, Noviembre 10). Common Warehouse Metamodel specification. [Documento WWW]. URL http://www.omg.org/cwm.

[10] Oracle Corporation. Oracle OLAP Application Developer Guide 10g Release 1 (10.1.0.4), 2005, 1 - 22.

[11] H. H. Eriksen, (Visitada 2005, Octubre 09). Oracle OLAP Forum. [Documento WWW]. URL http://forums.oracle.com/forums/thread.jspa?messageID=1085783&#1085783

[12] Oracle Corporation. Oracle OLAP DML Reference 10g Release 2 (10.2), 2005, 5 - 21.

[13] Jedox. (Visitada 2006, Marzo 21). PALO - Open Source MOLAP. [Documento WWW]. URL http://www.jedox.com/show.php?index=content/Open%20Source%20-%20Palo.07.

[14] Pentaho. (Visitada 2006, Marzo 21). Mondrian OLAP. [Documento WWW]. URL http://mondrian.sourceforge.net.

[15] Tensegrity Software. (Visitada 2006, Marzo 21). PALO Eclipse Client. [Documento WWW]. URL http://www.tensegrity-software.com/jpalo.html.