Universidad de Costa Rica
Facultad de Ingeniería
Escuela de Ciencias de la Computación e Informática

Carta al Estudiante

CI-2355 – Almacenes de datos y OLAP
I Semestre 2012
Lunes, 5:00 pm
Aula # 301
Profesor: Diego Villalba
Casillero # 19

INTRODUCCIÓN

Las organizaciones de hoy se enfrentan a retos relacionados con la creciente competencia y rápidos cambios, a los cuales deben adaptarse con el fin de lograr sus metas operacionales. Para poder tomar las decisiones acertadas con respecto a la dirección de desarrollo de la organización es necesario apoyarlas en los datos. Sin embargo, las bases de datos tradicionales, llamadas también operacionales o transaccionales, no cumplen los requisitos para el análisis de datos.

Los almacenes de datos (bodegas de datos o data warehouses en inglés) por medio del modelo multidimensional ofrecen una mejor comprensión de los datos para fines de análisis y permiten un mejor rendimiento para las consultas complejas necesarias para el soporte de la toma de decisiones. Estos grandes volúmenes de datos pueden ser analizados usando sistemas de procesamiento analítico en línea (OLAP- on-line analytical processing) que permiten la manipulación dinámica y agregación automática de los datos. Además, debido a que los almacenes de datos integran datos desde diferentes sistemas internos o externos a la organización, es necesario desarrollar los procesos de limpieza y transformación de los datos antes de cargarlos al almacén.

DESTINATARIOS

El curso está dirigido fundamentalmente a estudiantes de Pregrado y Licenciatura en Ciencias de la Computación e Informática.

OBJETIVO GENERAL

En el presente curso el estudiante conocerá las nuevas tendencias relacionadas con el diseño, implementación y mantenimiento de los almacenes de datos y se familiarizará con los procesos necesarios para su población y actualizaciones. Además, el estudiante aprenderá los conceptos relacionados con bases de datos multidimensionales sobre las cuales operan las herramientas del procesamiento analítico en línea.

OBJETIVOS ESPECÍFICOS

En el presente curso el estudiante:

- Entenderá la diferencia entre las bases de datos convencionales y almacenes de datos.

- Aprenderá a modelar los almacenes de datos a nivel conceptual usando la versión extendida del modelo entidad relación, en nivel lógico por medio de esquemas de estrella y copo de nieve como también en el nivel físico.
- Conocerá los diferentes tipos de jerarquías que existen en la vida real y que deben ser representados en los almacenes de datos.
- Será capaz de identificar los problemas relacionados con el proceso de extracción, transformación y carga (ETL – extraction, transformation, and loading) de los datos desde sistemas operacionales a los almacenes de datos y usar las herramientas que facilitan este proceso.
- Aprenderá conceptos de bases de datos multidimensionales usados para construir cubos de datos en sistemas OLAP e incorporarlos a las herramientas de cliente.
- Conocerá diferentes metodologías de desarrollo de almacenes de datos.
- Se introducirá a las nuevas tendencias en el desarrollo de almacenes de datos y OLAP.

CONTENIDO GENERAL DEL CURSO

1. Introducción
2. Conceptos básicos sobre los almacenes de datos
    a. Definición de almacenes de datos.
    b. Diferencia entre un sistema operacional o transaccional y un sistema informativo.
    c. Breve reseña histórica sobre el desarrollo de almacenes de datos.
    d. Metas, beneficios y obstáculos de la implementación del almacén de datos.
    e. La arquitectura de los almacenes de datos: sus componentes y variaciones.
    f. El ciclo de vida del almacén de datos.
3. Modelo multidimensional de datos
    a. Visión abstracta de cubo y operaciones de drill-down, roll-up, slice-and-dice.
    b. Representación por medio de tablas relacionales: esquema de estrella y copo de nieve.
    c. Conceptos de diseño: tabla de hechos, medidas, dimensiones, jerarquías y granularidad.
    d. Introducción al modelo conceptual multidimensional.
    e. Uso del modelo conceptual para la representación de los elementos del almacén de datos de supermercado.
    f. Clasificación de diferentes tipos de medidas.
4. Mapeo del modelo conceptual multidimensional al modelo lógico relacional
    a. Reglas generales.
    b. Refinamiento de las reglas para diferentes tipos de jerarquías.
    c. La importancia de llaves tipo “surrogate”.
5. Diseño físico de los almacenes de datos
    a. Características que diferencian bases de datos operacionales y almacenes de datos con respecto al almacenamiento de datos y consultas.
    b. Métodos de almacenamiento presentes en herramientas DBMS actuales.
    c. Diferentes tipos de índices.
    d. Vistas materializadas.
    e. Fragmentación y ejecución paralela.
    f. Mediciones de rendimiento (Benchmarking).
6. Procesos de extracción, transformación y carga (ETL).

a. Etapas en el proceso de ETL.
b. Identificación de campos fuentes.
c. Aspecto de validación y calidad de datos.
d. Transformaciones de datos con el objetivo de su integración.
e. Área intermedia de almacenamiento (staging area).
f. Diferentes tipos de actualizaciones de almacenes de datos.
g. Metadatos y su administración.

7. Diferentes métodos para el diseño de almacenes de datos
    a. Las fases de diseño.
    b. Diferentes métodos para la fase de especificación de requerimientos y modelaje conceptual.
8. Procesamiento analítico en línea (OLAP)
    a. Limitaciones de análisis de los datos presentes en hojas electrónicas y SQL.
    b. Breve reseña histórica sobre el desarrollo de sistemas OLAP.
    c. Concepto de cubo, medidas, dimensiones, atributos, jerarquías, funciones y agregados presentes en las herramientas OLAP.
    d. Diferentes formas de almacenamiento físico en servidores OLAP.
    e. Cubos de datos, su computación y problema de explosión de datos.
    f. Métodos para diseño de cubos OLAP.
    g. Procesamiento y re-procesamiento de cubos.
    h. Visualización de cubos OLAP.
    i. Herramientas clientes para consultas dinámicos sobre cubos OLAP.
    j. Ejemplo de implementación de diferentes elementos multidimensionales en un servidor OLAP de software libre o comercial.
9. Conceptos avanzados de diseño multidimensional:
    a. Diferentes tipos de jerarquías
    b. Esquema de constelación y la operación de drill-across.
    c. Dimensiones que cambian valores de sus atributos.
    d. Dimensiones grandes.
    e. Roles de las dimensiones.
    f. Dimensiones degeneradas (dimensiones representadas como hechos).
    g. Dimensiones multi-valuadas.
    h. Relaciones factuales sin medidas.
    i. Ejemplos de aplicación de modelaje multidimensional en diferentes áreas de desarrollo humano: esquema de transacción, de snapshot, de entrega de productos, entre otos.
10. Nuevas tendencias en almacenes de datos y procesamiento analítico en línea.

METODOLOGÍA

Las clases teóricas serán complementadas con evaluaciones y tareas teórico-prácticas semanales de las lecturas y demás material. Para fortalecer el aprendizaje se realizarán las prácticas guiadas, donde se enseña los conceptos teóricos aplicados a casos prácticos. Para verificar la adquisición del conocimiento, se realizarán tareas que cubren diferentes etapas de aprendizaje.

Partiremos del interés de cada grupo de estudiantes (máximo 2) para diseñar un almacén de datos por grupo. Este proyecto del curso pretende ser un elemento de trabajo de unión y aplicación de los conceptos aprendidos en el curso a un caso real de mediana complejidad. El proyecto debe tener adecuada documentación y los resultados deben ser presentados al profesor en forma de una defensa preparada adecuadamente.

Además, cada estudiante desarrollará un tema de actualidad, en colaboración con el profesor, y culminará con una presentación de un artículo y su exposición en clase.

EVALUACIÓN

- Evaluaciones semanales (25%). El estudiante desarrollará pequeñas tareas y examenes cortos semanales relacionados al material de cada semana de clases para demostrar la comprensión de cada tema. En otras palabras la evaluación se distribuirá a lo largo de todo el período de clases.
- Examen (25%).
- Investigación de un tema actual (25% - 15% por el artículo, 10% por la presentación) Trabajo individual.
- Proyecto (25%). Trabajo en grupos de 1 ó 2 estudiantes.

BIBLIOGRAFÍA

1. E. Malinowski y E. Zimányi. “Designing Conventional, Spatial, and Temporal Data Warehouses: From Conventional to Spatial and Temporal Applications”. Springer, 2008.

2. R. Wrembel y Ch. Koncilia (eds.). “Data Warehouses and OLAP”. IRM Press, 2007.

3. M. Rafanelli (ed.). “Multidimensional Databases: Problems and Solutions”. Idea Group Publishing, 2003.

4. M. Jarke, M. Lenzerini, Y.Vassiluiou y P. Vassiliadis. “Fundamentals of Data Warehouse”. Springer, 2003.

5. R. Kimbal, M. Ross y R. Merz. “The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling”. John Wiley & Sons, 2002.

6. R. Kimball, M. Ross, W. Thornthwaite, J. Mundy y B. Becker. “The Data Warehouse Lifecycle Toolkit. John Wiley & Sons, 2008.

7. W. Inmon. “Building the Data Warehouse”. John Wiley & Sons, 2002.

8. C. Imhoff, N. Galemmo y J. Geiger. “Mastering Data Warehouse Design”. John Wiley & Sons, 2003.

9. T. Lachev. “Applied Microsoft Analysis Services 2005”. Prologica Press, 2005.

10. E. Thomsen. “OLAP Solutions. Building Multidimensional Information Systems”. John Wiley & Sons, 2002.

11. B. Larson. “Delivering Business Intelligence with Microsoft SQl Server 2005”. McGraww-Hill, 2006.

12. B. Knight, A. Mitchell, D. Green y otros. “Professional SQL Server 2005 Integration Services”.

Wiley Publishing, Inc., 2006.

13. Diferentes artículos de revistas y conferencias.

14. Documentación en línea de diferentes productos comerciales.