DATAWAREHOUSE CON GEOLOCALIZACIÓN Y CLUSTERING

Carolina Zambrano
Departamento de Informática
y Ciencias de la Computación,
Universidad de Atacama
Chile
carolina.zambrano@uda.cl

Darío Rojas
Departamento de Informática
y Ciencias de la Computación,
Universidad de Atacama
Chile
dario.rojas@uda.cl

Marcela Varas
Departamento de Informática
y Ciencias de la Computación
Universidad de Concepción
Chile
mvaras@udec.cl

ABSTRACT

Every day organizations have more information because their systems produce a large amount of daily operations that are usually stored in transactional databases. In order to analyze this historical information, an interesting alternative is to implement a Data Warehouse. On the other hand, the Data Warehouse by themselves do not support geographic analysis, and clustering techniques. However, in the ETL stage can apply techniques of geo location and machine learning techniques to classify and group information to improve the quality of historical analysis. This paper proposes a Data Warehouse architecture to perform an analysis of students' academic performance which considers geo referenced and classified information. The results show the viability of the architecture of data warehouse for analysis of academic performance with clustering and application of geo location for geographical dimensions and improve the analysis.

RESUMEN

Cada día las organizaciones tienen más información porque sus sistemas producen una gran cantidad de operaciones diarias que se almacenan generalmente en bases de datos transaccionales. Con el fin de analizar esta información histórica, una alternativa interesante es implementar un Data Warehouse. Por otro lado, los Data Warehouse por sí mismos no soportan análisis geográfico, ni técnicas de clustering. Sin embargo, en la etapa ETL se pueden aplicar técnicas de geo localización y técnicas de machine learning para clasificar y agrupar información con el fin de mejorar la calidad del análisis histórico. En este trabajo se propone una arquitectura de Data Warehouse con el fin de realizar un análisis del desempeño académico de los estudiantes que considera información geo referenciada y clasificada mediante clustering. Los resultados muestran la viabilidad de la arquitectura de Data Warehouse para el análisis de rendimiento académico con aplicación de clustering y geo localización para obtener dimensiones geográficas y mejorar el análisis.

KEYWORDS

Data Warehouse, Geo localización, Clustering.

INTRODUCCIÓN

Actualmente las organizaciones cuentan con mucha información que puede ser aprovechada usándola en el proceso de toma de decisiones. El ámbito educacional no se encuentra exento de datos, por lo cual es interesante aplicar análisis en este contexto. Para analizar esta información se cuenta con diversas técnicas como las que se describen a continuación.

Una de las acciones más utilizadas en las instituciones educacionales para dar valor a la información y dar apoyo a la toma de decisiones, es la confección de reportes. La confección de los reportes es una acción exploratoria, es decir, se hacen ciertos cruces de datos y dependiendo de los resultados, se van analizando otros criterios hasta que se llega a un punto en el cual los resultados son satisfactorios para tomar decisiones sobre la organización.

El apoyo a la toma de decisiones puede ser realizado mediante sistemas especialmente diseñados para ello como son los DSS [18] (Decision Support Systems), los cuales pueden generar informes parametrizables en forma periódica, rápida y fácil, como los presentados en [14].

Otro método comúnmente utilizado, es la creación de reportes mediante la manipulación directa de bases de datos transaccionales a través del lenguaje SQL (Structured Query Language), lo cual tiene el inconveniente de requerir una persona experta en la utilización de SQL. Además el desarrollo de reportes puede tomar un tiempo considerable debido a que las bases de datos transaccionales no están diseñadas específicamente para el análisis. Otro método muy utilizado trata sobre el uso de planillas de cálculo y datos tabulados, sin embargo, este método a pesar de necesitar menos conocimientos técnicos sufre de la imposibilidad de manejar eficientemente grandes cantidades de datos directamente, como también sufren de la dificultad de poder realizar el cruzamiento de datos en forma sencilla desde distintas fuentes de datos.

Por otro lado, los Data Warehouse (DW), son repositorios de datos electrónicos especialmente diseñados para la generación de reportes y análisis de datos [10] [20]. Las características distintivas de los DW respecto a los sistemas descritos anteriormente es que son flexibles, integran todos los aspectos organizacionales de interés, pueden manejar grandes volúmenes de datos eficientemente, permiten la creación y cálculo de indicadores de gestión. Además, los DW se diseñan con el objetivo de ser eficientes en los requerimientos de análisis para niveles estratégicos en las organizaciones, por lo que toman en cuenta los objetivos estratégicos de la organización directamente [12]. En el mismo contexto, los DW permiten analizar de forma eficiente la información histórica de una organización, y de esta forma visualizar tendencias de comportamiento de los indicadores de gestión en el tiempo. Por lo cual en una organización educacional serían muy útiles para analizar por ejemplo las tendencias académicas de los estudiantes.

Por otro lado las técnicas de machine learning [13] tales como la de clustering pueden ayudar agrupar información. En este contexto se han aplicado técnicas de machine learning como las redes neuronales [7] [9] a datos educacionales en [4] [16] [17], sin embargo estas técnicas no se han aplicado en conjunto a un DW como se propone en este trabajo donde se aplica clustering para ayudar a obtener la información de una dimensión de zona geográfica clasificada o agrupada.

En este trabajo se ha implementado un DW en base a información obtenida de un sistema de base de datos no relacional (basado en archivos o también llamado sistema heredado). El DW se ha diseñado con el objetivo de analizar el comportamiento de aprobación y avance en una malla curricular con datos reales de los currículos de los estudiantes de la carrera de Ingeniería Civil en Computación e Informática de la Universidad de Atacama.

El artículo tiene la siguiente estructura: primero se presenta un apartado de Metodología que presenta la metodología de trabajo que explica la arquitectura del DW diseñado e implementado. Posteriormente se presenta un apartado de Análisis y Resultados que incluye los principales resultados del análisis ROLAP obteniendo las tendencias de comportamiento. Finalmente se presenta la Conclusión y Trabajos Futuros que incluyen comentarios sobre los resultados, y posibles trabajos futuros.

METODOLOGÍA

Diseño e Implementación del Data Warehouse

Un DW está compuesto de elementos básicos, entre los que podemos encontrar las dimensiones de análisis, las medidas también conocidas como indicadores de gestión y los hechos que representan los datos reales. En este contexto, los DW se diseñan para poder calcular y analizar un conjunto de indicadores de gestión. Con este enfoque, los "indicadores de gestión" dirigirán el diseño, y se convertirán en las "medidas", y las "variables/criterios" a analizar se convertirán en las "dimensiones" de un modelo multidimensional [10] [20]. Cada celda o hecho, contiene uno o más indicadores de gestión, como por ejemplo podría ser la cantidad de estudiantes por asignatura y región, promedio de notas, etc. Otro concepto en el ámbito de los DW es el de Data Mart que representan pequeños DW centrados en un tema o un área de negocio específico dentro de una organización [1].

La tecnología que permite una acción exploratoria de los datos del DW es OLAP [5] (Online Analytical Processing), que no sólo permite flexibilidad en cuanto a la navegación a través del modelo multidimensional de la información, sino que también es flexible en la definición de los reportes y aplicaciones que se construyen a partir de ella. Además, las herramientas OLAP definen claramente operadores especiales de refinamiento o manipulación de consultas que pueden ser comprendidas mucho más fácilmente que las sentencias SQL y que además son eficientes, ya que se realizan sobre datos y resúmenes pre-computados.

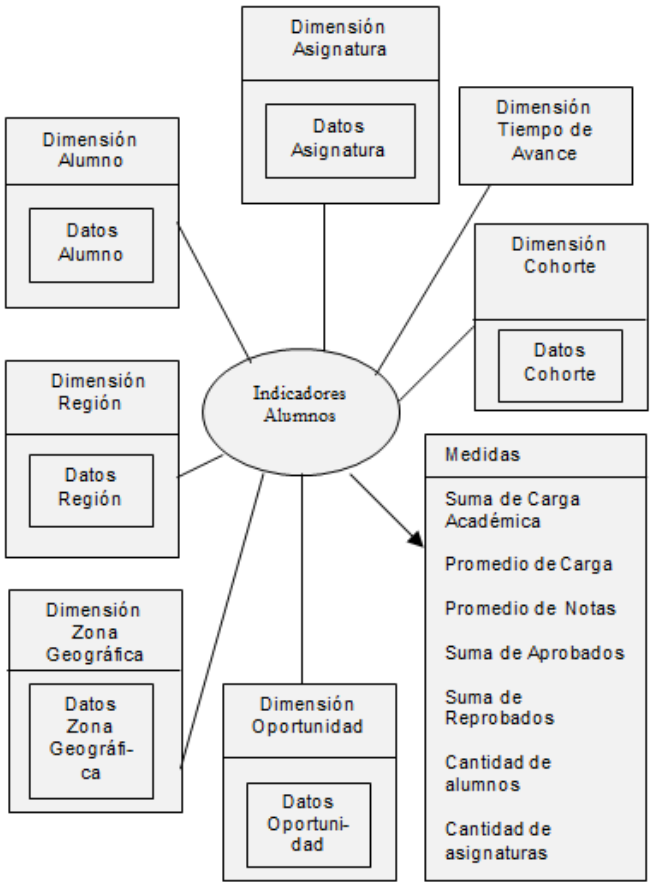

Figura 1. Esquema Conceptual del DW (usando modelo conceptual CMDM [3] para especificar el diseño del DW implementado para el análisis de indicadores de estudiantes.

Un sistema de DW puede ser implementado bajo enfoque Molap (MultidimensionalOlap), Rolap (RelacionalOlap) o mediante el hibrido Holap (permite tanto Molap como Rolap)

[5]. En este trabajo se utilizó enfoque Rolap. Independiente del enfoque los principales procesos que se llevan a cabo en el desarrollo de un DW son los siguientes:

Proceso de Modelamiento Conceptual: El modelo conceptual es independiente de la tecnología y es primordial para especificar los requerimientos de análisis y disponibilidad de información. A nivel de modelos conceptuales de DW no existe consenso en la comunidad de investigadores sobre cuál es el modelo aceptado como estándar para la representación de un DW, sin embargo, hay varias propuestas algunas de ellas se presentan en [3] [6] [8] [19]. Durante el proceso de modelamiento conceptual se genera el esquema conceptual del DW. En este trabajo se utilizó el modelo conceptual CMDM [3] debido a la sencillez de su notación y porque su objetivo es justamente la especificación conceptual de un DW.

Proceso de Modelado Lógico e implementación Física: El modelo lógico, específica formalmente el esquema multidimensional, sus restricciones y capacidades. Por otro lado el esquema lógico es implementado directamente en un motor de base de datos transformándose en tablas físicas. En el caso de los DW esquemas de diseño lógico son el esquema estrella y el esquema copo de nieve [2]. En la etapa de implementación física se crean las tablas de dimensión y tabla de hecho dependiendo del tipo de esquema estrella o copo de nieve.

Proceso de Carga de Datos ETL: El proceso ETL (Extraction, Transformation, Load) es el encargado de extraer los datos de las bases de datos originales, transformarlos y cargarlos en el DW. La Figura 2 muestra un esquema del proceso ETL que se llevó a cabo durante el desarrollo de este trabajo. Es importante indicar que el proceso de clustering y geo localización se llevó a cabo durante esta etapa de tal forma que se pueda aprovechar la información clasificada de las zonas geográficas donde se ubican los domicilios de los estudiantes, que luego se cargan en una dimensión llamada zona geográfica de la arquitectura propuesta. Esto permitirá analizar si el lugar donde vive un estudiante tiene alguna influencia en su rendimiento académico.

Proceso de Análisis Rolap: Permite la acción exploratoria a través de las operaciones definidas en OLAP para el análisis y creación de reportes bajo modelo relacional.

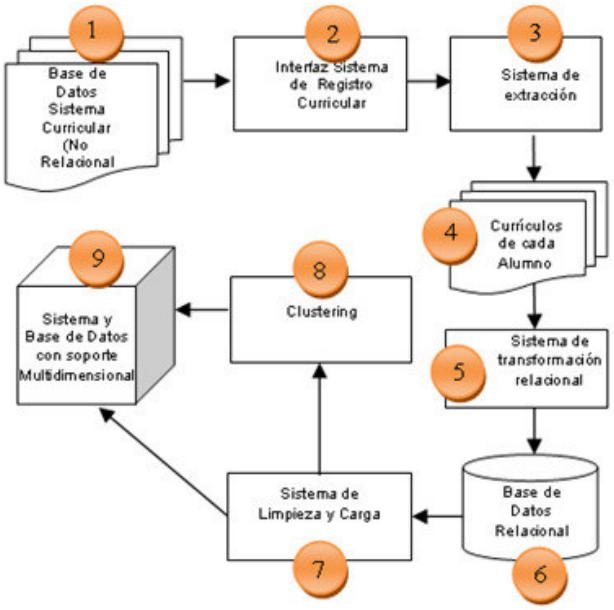

Figura 2. Esquema ETL simplificado para la carga del DW.

En la implementación del DW la primera etapa consistió en diseñar el esquema conceptual para el análisis como se muestra en la Figura 1. Este esquema de modelo conceptual, posee siete dimensiones de análisis:

1) Alumno: con los datos personales de los estudiantes y su estado.
2) Asignatura: con los datos de las asignaturas impartidas y condiciones de entrada a la universidad (PSU).
3) Región: con las regiones y ciudades de donde provienen los estudiantes.
4) Oportunidad: Representa los datos sobre las oportunidades posibles de cursar las asignaturas.
5) Tiempo de Avance: Tiempo de permanencia de un alumno en la carrera, en base a los semestres.
6) Zona Geográfica: Representa la zona geográfica donde se ubica el alumno.
7) Cohorte: Cohorte a la que pertenecen los estudiantes.

Por otro lado, los indicadores multidimensionales son implementados a través de las Medidas tales como cantidad de estudiantes, suma de aprobados, etc. según se aprecia en la Figura 1. En este contexto, cabe notar que el esquema lógico no es presentado por efectos de simplicidad y extensión.

El proceso ETL simplificado se presenta en la Figura 2, el cual consiste, en extraer los datos desde la base de datos del sistema de información curricular de estudiantes de la universidad (1), el cual no está soportado por un motor relacional y funciona a través de archivos (sistemas heredados). Este sistema sólo es accesible mediante una interfaz de usuario a través de la red mediante una aplicación de consola heredada del lenguaje COBOL (2). Para extraer esta información se simulo el proceso manual de extracción mediante una aplicación especialmente diseñada para ello (3), luego de lo cual se extrajo el currículo de cada alumno en

formato de texto (4). Estos archivos de texto, son transformados mediante la utilización de un software diseñado a medida (5) y cargados en una base de datos relacional (6), tras lo cual son transformados nuevamente por otra aplicación implementada a través de procedimientos almacenados (7) que los carga en el DW (9). Por otro lado, los datos sobre ubicación geográfica de los estudiantes son transformados a coordenadas de latitud y longitud mediante un software de geo localización especialmente diseñado, utilizando como base de datos Google Maps (ver Figura 3a), para luego realizar un proceso de clustering (8) mediante el algoritmo k-means [11]. Posteriormente se procede a etiquetar la ubicación en cuatro zonas como se puede apreciar en la Figura 3b. Luego, estos datos son cargados igualmente al DW como parte de la dimensión Zona Geográfica.

Figura 3a. Sistema de geo localización automática utilizando Google Maps como base de datos. (Disponible en http://frodo.diicc.uda.cl/demogeoloc/)

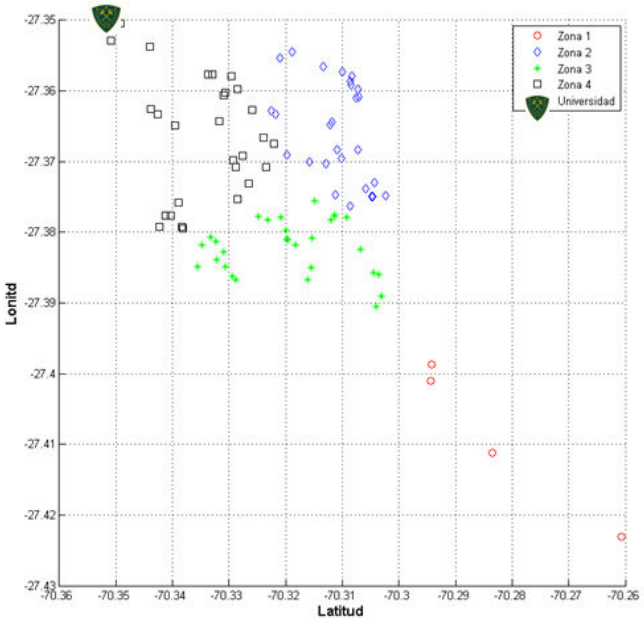

Figura 3b. Resultado de la etiquetación de la ubicación geográfica de los estudiantes mediante clustering.

ANALISIS Y RESULTADOS

En este apartado se analiza el comportamiento de ciertos indicadores en el tiempo a través de la arquitectura de DW propuesta e implementada.

Análisis usando la arquitectura del DW

La plataforma utilizada para el análisis Rolap, fue Pentaho Business Intelligence [15], en su versión open source que cubre las necesidades de Análisis de los Datos y de Reportes, siendo una de sus características su funcionalidad y simplicidad en la implementación.

El objetivo de los análisis que se presentan a continuación es demostrar la versatilidad de los resultados de las operaciones mediante Rolap, debido a que todos los reportes presentados en este trabajo fueron generados en poco tiempo (en relación al diseño e implementación del DW), lo que indica claramente la capacidad de la plataforma DW – Rolap para consultar y analizar datos dispuestos multidimensionalmente desde distintos puntos de vistas, sin un diseño pre-establecido del sistema, sino más bien, sólo del modelo de datos y análisis previo que permite llegar a una arquitectura de diseño de DW robusta para el análisis.

A continuación se presentan los gráficos de algunos reportes de análisis obtenidos de la propuesta de DW.

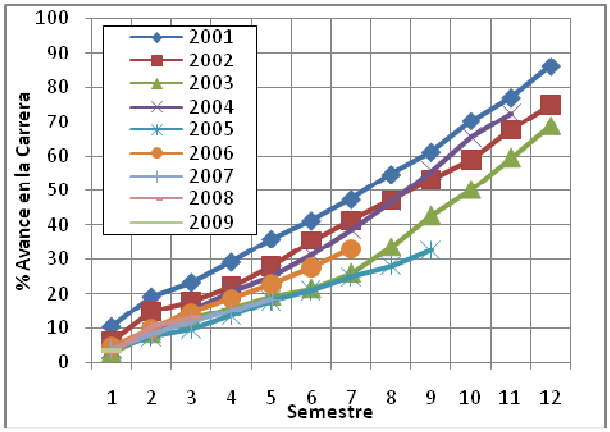

Figura 4: Porcentaje de Asignaturas Aprobadas Acumuladas (PAAA) de la carrera por semestre de permanencia para cada cohorte.

En el gráfico de la Figura 4, se muestra la tendencia del Porcentaje de Aprobación de Asignaturas Acumuladas (PAAA), por semestre de permanencia para las distintas cohortes a partir del año 2001. Como se puede apreciar, en el semestre 12 de permanencia los estudiantes de la cohorte 2001 presentan en promedio, un 85% de los ramos de la carrera aprobado, siendo el mejor desempeño según las cohortes analizadas. Por otro lado, se puede ver que las cohortes 2002 y 2004 y 2006 se escapan al comportamiento común de las cohortes 2003, 2005, 2007, las cuales tienen un PAAA en el tiempo bastante más bajo. Cabe notar que cohortes más nuevas no poseen más información, debido a que aún no había datos para los semestres posteriores, sin embargo la tendencia inicial de las curvas permiten predecir a simple vista su comportamiento futuro. Cabe destacar que esta predicción es sólo por análisis de la curva del gráfico.

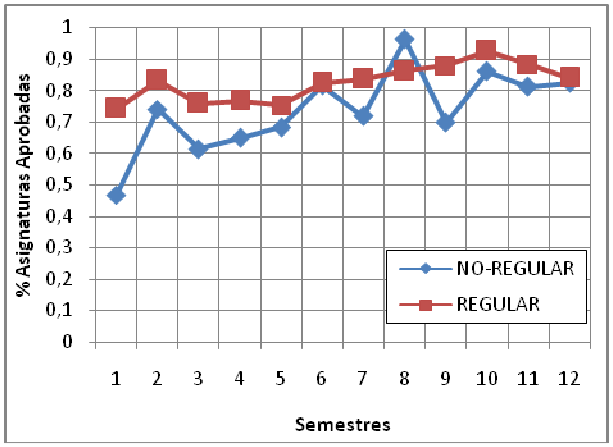

Figura 5: Porcentaje de Aprobación por Semestre (PAS) en asignaturas de la carrera por semestre de permanencia para estudiantes regulares y no-regulares.

En el gráfico de la Figura 5, se muestra el Porcentaje de Aprobación de Asignaturas por Semestre (PAS), de los estudiantes regulares de la carrera y los en situación no-regular. Los estudiantes no-regulares son aquellos estudiantes eliminados o que no renovaron matriculas o que se encuentran en cualquier otra situación que les quite la condición de alumno regular. Como se aprecia, la aprobación de los estudiantes regulares, es siempre superior que los estudiantes no-regulares, salvo para el semestre 8, el cual presenta una inferioridad respecto a los no regulares. Esto último es debido a que en el semestre 8, hay muy pocos estudiantes en condición no-regular y por lo tanto pocas asignaturas reprobadas en relación a las aprobadas por parte de estos estudiantes.

En el gráfico de la Figura 6, se puede apreciar que el PAAA por zona geográfica es muy similar, excepto para la Zona 1, lo que puede estar justificado por la distancia geográfica de estos estudiantes con respecto a la universidad la cual se encuentra marcada por su escudo. Este análisis sólo se pudo obtener producto de la incorporación del clustering y geo localización.

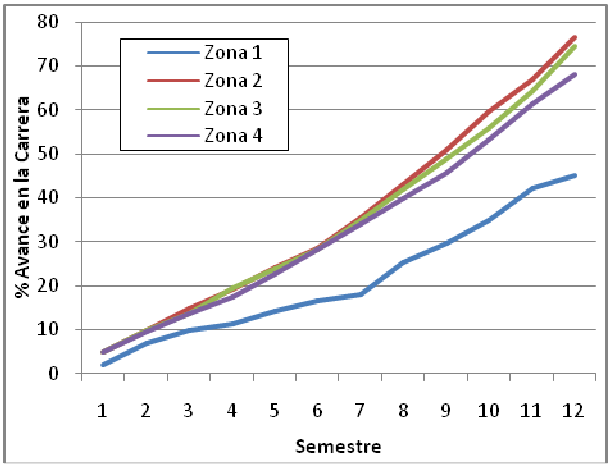

Figura 6: Porcentaje de Asignaturas Aprobadas Acumulada (PAAA) de la carrera por semestre de permanencia para cada Zona geográfica.

En el gráfico de la Figura 7, se puede apreciar, que las cohortes tienen en general un comportamiento irregular del PAS. Por ejemplo, la cohorte 2002, tiene un comportamiento además de inferior en porcentaje a las otras cohortes en cada semestre, su variabilidad en el tiempo también es mayor. Este comportamiento puede ser explicado porque los estudiantes en un semestre determinado, en su mayoría reprueban un ramo y luego al segundo semestre tienen una menor carga y aprueban regularmente los ramos reprobados con anterioridad. Luego, un estudiante, nuevamente se encuentra con nuevos ramos los cuales reprueba provocando el comportamiento de subidas y bajadas en el indicador PAS. En este gráfico sólo se muestran las cohortes 2001 a 2004 debido a que las otras cohortes aún no han cursado todos los semestres a analizar. Además, este gráfico es revelador, desde el punto de vista del comportamiento para este indicador está determinado por la cantidad de asignaturas aprobadas y la cantidad de asignaturas cursadas en un semestre, por lo que se presume la posibilidad de predecir por lo menos estos datos del próximo semestre dado el historial de un alumno.

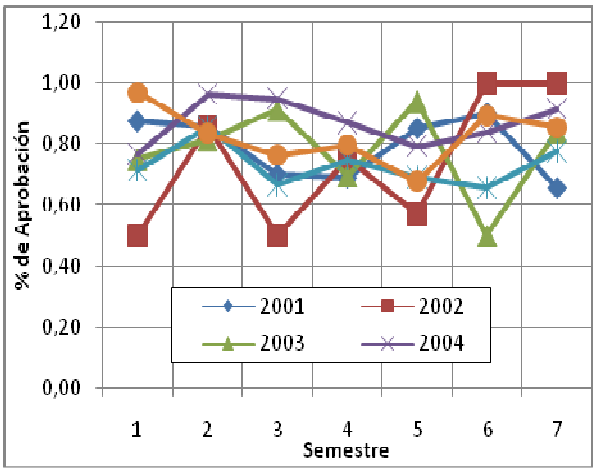

Figura 7: Porcentaje de Aprobación por Semestre (PAS) en asignaturas de la carrera (por semestre de permanencia para las cohortes de la 2001 a las 2004).

CONCLUSIONES Y TRABAJO FUTURO

Se ha realizado la implementación de un Data Warehouse para el análisis de rendimiento de los datos académicos de los estudiantes de Ingeniería Civil en Computación e Informática de la Universidad de Atacama. La principal ventaja en la utilización de un DW, radica en la posibilidad de cruzar distintas dimensiones de análisis de forma simple y rápida, con tal de realizar un análisis exploratorio de los datos para la creación de reportes. Se puede destacar que el proceso de extracción, transformación y carga (ETL), es el que más tiempo y recursos demando, debido principalmente a que la información debe ser cruzada desde distintas fuentes. Además, los sistemas operacionales no están diseñados para analizar datos y la heterogeneidad de las plataformas donde se encuentra la información, añade una mayor dificultad que obliga a la creación de aplicaciones y sistemas específicos que permitan aprovechar los datos históricos. Es preciso agregar, que la utilización de un modelo conceptual multidimensional para generar el esquema conceptual del DW, se convierte en una gran herramienta que, independiente de las plataformas, permite acotar el dominio de análisis y dar claridad al proceso posterior de ETL.

Para finalizar respecto a la implementación de DW, podemos indicar que el análisis mediante Rolap es eficiente y permite realizar operaciones en el cubo en tiempo real para poder navegar por los datos desde distintas perspectivas de una manera sencilla e intuitiva. Además el clustering y la geo localización permiten mejorar el análisis respecto de una arquitectura tradicional de DW que no tiene estas características.

El análisis mediante DW permite a la institución tomar medidas remediales para poder analizar, modificar y validar los indicadores de gestión o quizás para generar nuevas estrategias que le permitan mejorar y/o optimizar su proceso de gestión, pues el conocimiento se extrae de sus mismas bases de datos, dando valor a la información de gestión que se registra pero que quizás no siempre se tiene en cuenta.

Como trabajos futuros se trabaja para agregar dimensiones sociales, dimensiones económicas y dimensiones con datos de encuestas de perfiles biosicosociales de los estudiantes con los que ya se cuenta, para generar una nueva arquitectura de DW.

AGRADECIMENTOS

Al laboratorio de investigación Midal, del DIICC – UDA. Este trabajo fue parcialmente financiado por la Dirección de Investigación de la Universidad de Atacama, Chile, Proyecto 221219 “Data Warehouse Para Análisis Con Jerarquías Difusas”.

REFERENCIAS

[1] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, S. Paraboschi. “Designing Data Marts for Data Warehouses”. ACM Transactions on Software Engineering and Methodology. Vol. 10, Issue 4, pp. 452-483. October, 2001.

[2] L. Cabibbo, R. Torlone. “A Logical Approach to Multidimensional Databases”. Lecture Notes in Computer Science. Vol. 1377. 1998.

[3] F. Carpani. “CMDM: Un Modelo Conceptual para la Especificación de Bases de Datos Multidimensionales”, Tesis para optar al grado de Maestría. Universidad de la República. Uruguay.2000.URL:http://www.fing.edu.uy/inco /pedeciba/bibliote/tesis/tesis-carpani.pdf.

[4] Z. Cataldi, F. Salgueiro, F. Lage. “Predicción del rendimiento de los estudiantes y diagnóstico usando redes neuronales.” XIII Jornadas de Enseñanza Universitaria de la Informática, España, 2006.

[5] S. Chaudhuri, U. Dayal. “An Overview of Data Warehousing and OLAP Technology” SIGMOD Record, 26(1):65.74, 1997.

[6] M. Golfarelli, D. Maio, S. Rizzi. “Conceptual Design of Data Warehouses from E/R Schemes”.Proceedings of the Thirty-First Hawaii International Conference on System Sciences. 1998.

[7] S. Haykin, “Neural Networks a Comprehensive Foundation”, Macmillan College Publishing, Inc., USA, Book ISBN number 9780023527616. 1994.

[8] B. Hüsemann, J. Lechtenbörger, G. Vossen. “Conceptual Data Warehouse Design”. DMDW’00. Sweden. 2000.

[9] P. Isasi, I. Galván, “Redes de Neuronas Artificiales Un enfoque Práctico”, Pearson. Book ISBN number 8420540250. 2004.

[10] R. Kimball, M. Ross, R. Merz. “The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling” . John Wiley & Sons, Book ISBN number 0471200247. 2002.

[11] U. Maulik, S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices”, IEEE transaction on pattern analysis and machine intelligence. Vol. 24. pp. 1650-1655. 2002.

[12] J.N. Mazón, J. Trujillo, M. Serrano and M. Piattini. "Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling". In Proceedings of the 1 Workshop on Requirements Engineering for Business Need and IT Alignment. Paris, France. September, 2005.

[13] T. Mitchell, "Machine Learning", McGraw-Hill, USA. Book ISBN number 0070428077. 1997.

[14] G. Olguín. Sistema de Monitoreo y Análisis del Comportamiento Académico del Alumnado, XXIII Congreso Chileno de Educación en Ingeniería, Concepción, Chile, 2009.

[15] Pentaho. "Pentaho Business Intelligence", URL: http://www.pentaho.com.

[16] M. A. Pinninghoff, P. Salcedo, R. Contreras, "Neural Networks to Predict Schooling Failure/Sucess", Lecture Notes Computer Science. Vol. 4528. 2007.

[17] M. A. Pinninghoff., M. Herrera, R. Contreras, P. Salcedo, "Predicción de rendimiento académico mediante redes neuronales", VI Congreso Chileno de Educación Superior en Computación. Jornadas Chilenas de Computación, Arica, Chile, 2004.

[18] G. Salvendy "Decision Support Systems".Chapter 4: Handbook of Industrial Engineering: Technology and Operations Management, John Wiley & Sons Book ISBN number 0471330574. 2001.

[19] C. Sapia, M. Blaschka, G. Höfling, B. Dinter. "Extending the E/R Model for the Multidimensional Paradigm". DWDM'98. Singapur, pp. 105-116. 1998.

[20] C. Todman, Designing a Data Warehouse, Prentice Hall, Book ISBN number 9780130897121. 2001.