

Un Almacén de Datos para el Análisis de los Resultados Académicos de la U. P. Valencia

Blesa P., Hernández, J.
Depto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Valencia, Spain
{pblesa, jorallo}@dsic.upv.es

1. Introducción y objetivos

En nuestra universidad existen datos académicos sobre soporte informático desde hace más de 10 años. Con el fin de dar respuesta a determinados indicadores que podrían ser usados para la financiación de la universidad, ésta decidió la creación de un almacén de datos. Al mismo tiempo, un grupo de profesores de bases de datos que estábamos interesados en esta tecnología solicitamos financiación en nuestra universidad, y nos fue concedida, para un proyecto cuyo objetivo era poner a disposición de los responsables académicos una aplicación que apoye el proceso de evaluación del sistema de enseñanza de nuestra universidad. A continuación expondremos brevemente la estructura de los datos disponibles, el tipo de informes que se están realizando, y los recursos software y de equipo usados.

2. Estructura de los datos

El sistema de información transaccional de la Universidad Politécnica de Valencia consta de una base de datos relacional, centralizada e histórica, que contiene información de diferentes ámbitos de la Universidad: titulaciones, centros, departamentos, profesorado, resultados académicos, gestión, actividades, personal, etc. Esta base de datos se encuentra funcionando actualmente en un servidor Oracle 8i. De todos estos ámbitos nos interesan fundamentalmente aquellos directamente relacionados con los resultados académicos. En esta base de datos, dicha información se encuentra repartida en cerca de una cincuentena de tablas, denominadas tablas de matrícula, que incluyen los datos de los alumnos y todos sus datos asociados, las titulaciones, las asignaturas, los grupos, los centros, los resultados (notas) de cada alumno en cada convocatoria, etc. En la figura 1 se muestra una imagen visual de estas tablas y sus relaciones (claves ajenas). Los círculos discontinuos designan tablas o grupos de tablas.

Previamente a la construcción de un almacén de datos, una parte importante de la información original es irrelevante, con lo que es necesario realizar una importante y determinante fase de limpieza y selección de datos.

El almacén de datos se ha construido sobre otro servidor independiente. Después de la carga inicial, que realiza una importante criba de datos, se realizan actualizaciones semanales de la base de datos transaccional, con el objetivo de que el almacén de datos esté siempre actualizado.

Sobre la información filtrada en el almacén de datos se puede observar que la estructura de la base de datos resultante se centra alrededor de una tabla denominada "MAT_NOTES", que incluye el dni del alumno, la titulación, el curso académico, la

asignatura, la convocatoria y el hecho: la nota. Esta tabla servirá como tabla fundamental de hechos (“un alumno obtiene una nota en una asignatura de una titulación en una convocatoria”) para la construcción de un almacén de datos.

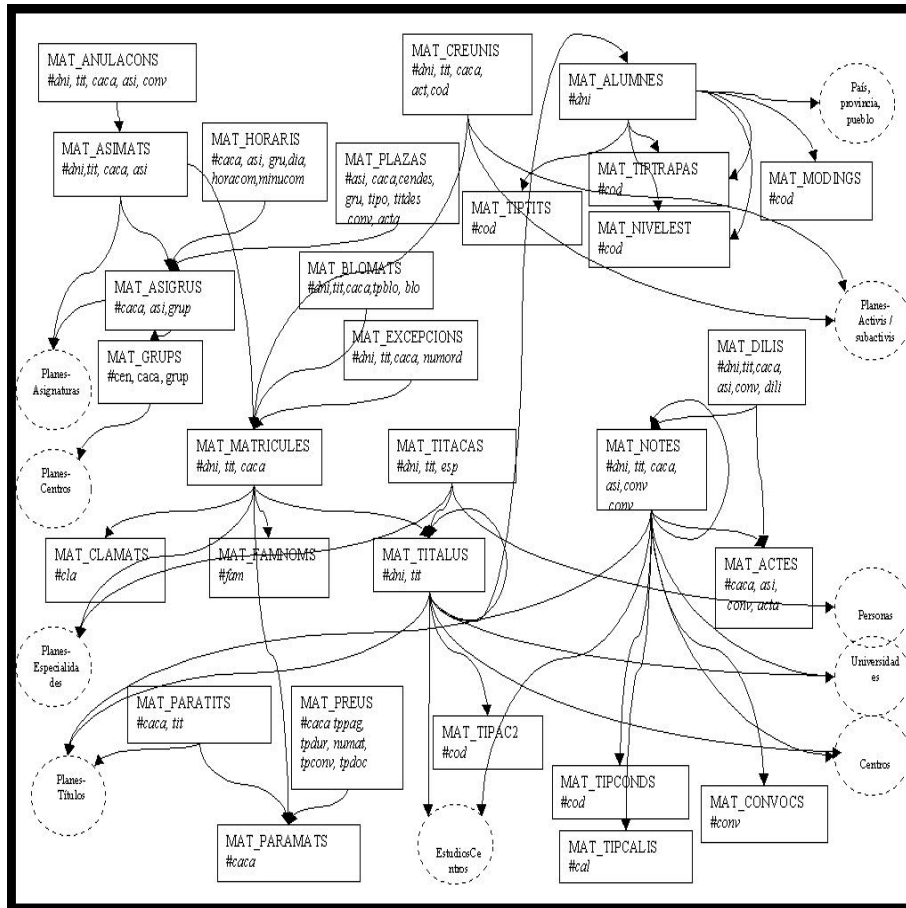


Fig. 1. Tablas de “matrícula” de la UPV.

A partir de estos hechos, se han establecido las siguientes dimensiones:

- Tiempo: Desglosable en cursos académicos y cuatrimestres.
- Grupos de alumnos: Desglosable en cohortes, grupos y alumnos.
- Estudios: Desglosables en titulaciones, centros y asignaturas.

Cada una de estas dimensiones tienen distintos niveles y caminos de agregación, dando lugar a una estructura típica de copo de nieve.

Es sobre esta base de datos Oracle y sobre estas dimensiones sobre las que se construye finalmente el diseño de almacenes de datos.

3. Arquitectura del sistema de Almacén de Datos

Como se acaba de comentar, el repositorio de los datos se obtiene del sistema operacional de alumnado de nuestra universidad. El sistema usado, MicroStrategy, se

basa en la creación de informes, cada uno de los cuales consta de una plantilla y de un filtro. En una plantilla se describe la estructura de datos buscados, y consta de varios atributos, y de al menos un indicador, que se obtiene a partir de un hecho cuantificable. Los filtros corresponden a restricciones sobre valores de atributos.

MicroStrategy dispone de un "explorador" que permite recorrer el "metadata", con el fin de seleccionar los atributos e indicadores con los que se construyen las plantillas, y los valores que se seleccionan para los filtros.

Como servidor del almacén de datos se dispone de un equipo Fujitsu GPF7000 con 4 procesadores GP64 (SPARC) de 64 bits a 275MHz, memoria RAM 4Gbytes y sistema operativo Solaris 7 (SunOS-5.7). Los discos de aplicación tienen 52 Gbytes en "mirror" (104 Gbytes). Esta máquina gestiona el almacén de datos con Oracle 8i. Por encima está montado el software de MicroStrategy

Para la explotación se disponen de tres PCs, trabajando con Windows NT, con Microstrategy Desktop, nivel 7, conectados con el almacén de datos, y que ofrece la ventana para el acceso al almacén.

4. Análisis e informes a realizar

El objetivo inicial del proyecto era proporcionar información en tiempo real frecuentemente solicitada por los responsables académicos de la universidad. Estos perfiles de consultas se han ido obteniendo a partir de las consultas disponibles en sistemas anteriores y los requisitos especiales de los indicadores de financiación de la universidad.

En general, podemos mencionar, entre otros, los siguientes modelos de consultas:

- Estudio de rendimiento, basándose en fórmulas de rendimiento académico.
- Estudio de abandonos.
- Mapas de progreso.
- Estudios comparativos de asignaturas.

Aunque los objetivos son más amplios, se ha decidido empezar por un estudio fácil de interpretar, comparando los resultados académicos de varias titulaciones, para una determinada promoción, y centrándonos en los resultados del primer curso.

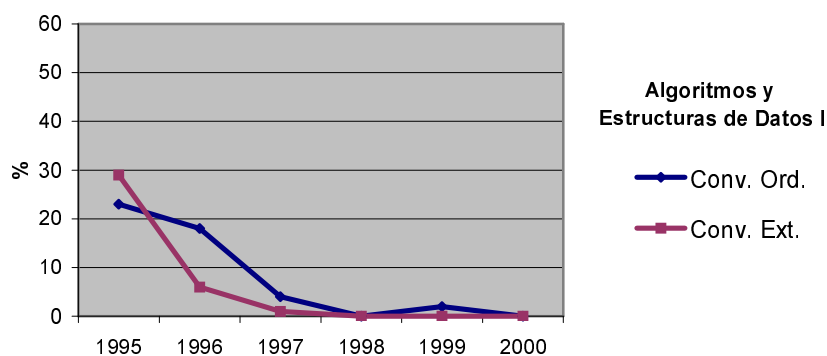


Fig. 2. Porcentaje de aprobados respecto a la cohorte ¹de 1995.

¹ Una cohorte es el conjunto de alumnos matriculados por primera vez un curso en una titulación.

Para cada titulación y cada asignatura (sólo tenemos en cuenta troncales y obligatorias) la gráfica es de la forma indicada en la figura 2.

Además, como resumen de cada cuatrimestre se realiza una gráfica donde se acumulan los resultados, teniendo en cuenta el número de créditos de cada asignatura, lo que nos da una clara idea del rendimiento del cuatrimestre, y al final de cada titulación se incluye el resumen del curso completo. Comparando dos de estas gráficas de diferentes titulaciones entre si es muy fácil discernir la evolución de ese primer curso en ambas titulaciones.

5. Conclusiones y Trabajos Futuros

Las características específicas de la información académica facilitan la creación de almacenes de datos sobre ella. Generalmente se trata de información histórica y donde es relativamente fácil reconocer los hechos (resultados académicos) y las dimensiones. Los informes obtenidos pueden ser de gran utilidad para la toma de decisiones orientadas a la mejora del servicio ofrecido.

Con el objetivo de potenciar las posibilidades y la utilidad del almacén de datos, está previsto iniciar los siguientes trabajos:

- 1- Rediseñar el almacén de datos: el diseño actual del almacén de datos está especialmente orientado para obtener indicadores bajo determinadas especificaciones de Escuelas/Facultades y Departamentos. Se ha planificado hacer en el curso 2001/2002 un estudio de requerimientos orientados al estudio del rendimiento escolar.
- 2- Minería de datos: una segunda fase de explotación consistiría en la aplicación de un paquete de minería de datos sobre el almacén de datos diseñado anteriormente. En estos momentos se contempla el uso de SPSS Clementine, conectado al almacén de datos a través de ODBC. Los modelos que puede ser interesante generar con esta herramienta podrían ser: Modelos descriptivos: (asociaciones entre atributos) o Modelos predictivos (modelo de estancia en la universidad por titulación según características del alumno, modelos de abandono).

Bibliografía Básica

1. Inmon, W. H. "Building the Data Warehouse (2nd edition)". John Wiley & Sons, 1996
2. Han, J.; Kamber, M. "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.
3. Kimball, R.; Reeves L.; Roos M.; Thornthwaite, W. "The Data Warehouse Lifecycle Toolkit". John Wiley & Sons, 1998
4. Mattison, R. "Data Warehousing: Strategies, Technologies and Techniques" McGraw-Hill, 1996