Data warehousing with IBM cloud db2 warehouse

Abstract

Nowadays, data warehouse tools and technologies cannot handle the load and analytic process of data into meaningful information for top management. Big data technology should be implemented to extend the existing data warehouse solutions. Universities already collect vast amounts of data so the academic data of university has been growing significantly and become a big academic data. These datasets are rich and growing. University's top-level management needs tools to produce information from the records. The generated information is expected to support the decision-making process of top-level management. This paper explores how big data technology could be implemented with data warehouse to support decision making process. In this framework, we propose Hadoop as big data analytic tools to be implemented for data ingestion/staging. The paper concludes by outlining future directions relating to the development and implementation of an institutional project on Big Data.

1. Introduction

Higher educations are working in a more and more complex and competitive environment. They have to compete with other institutions to answer to national and global economic, political and social changes. Moreover, different stakeholders are expecting higher education institutions to produce right solution in a timely manner to these demands. To overcome this condition, higher education needs to produce the right decisions required for dealing with these rapid changes by analyzing vast data sources that have been generated. Most of higher education institution invest enormous resources in information technology to implement data warehouse system .

The development of data warehouse is a way to extract the important information from the scattered data in some information systems into a centralized integrated storage and support the need for data history. This integrated data can be utilized for information delivery activities that can be reviewed from various dimensions and can be set the level of detail.

Further utilization of the information contained in the data warehouse is the activity of data analysis using certain techniques and methods. There are several algorithm for knowledge data discovery, like classifying, clustering and mining . The data contained in the data warehouse can used as input for the application system for example like a dashboard. With the existence of this dashboard is expected to be a solution for the learning process to monitor the academic condition and then could take the right decision. However, organizations are recognizing that traditional data warehouse technologies are dying to meet new business requirements, especially around streaming data, real-time analytics, large volumes of unstructured and complex data sets.

To solve this problem, this paper aims to design and implement a modern data warehouse for academic information system to support decision making process. The designed system accommodates Hadoop platform, a powerful analytical tools which is able to produce a graph that displays the student data information statistically. To support parallel and distributed processing of

large volumes of data, most solutions involve Hadoop technology. Hadoop is capable to perform analysis of large heterogeneous datasets at unprecedented speeds

As a result, top management will have a dashboard to monitor the existing condition of the academic atmosphere of university. The reporting dashboard itself will cover operational, strategic and analytical dashboard. The operational dashboards will tell us what is happening now, while strategic dashboards will track key performance indicators in academic process. Moreover, analytical dashboards will process data to identify trends.

The main contributions of this paper are as follows:

(1) the designed system enables the communication among different platform and datasets, including smart phones, web, and desktop application whether it is structured, semistructured and unstructured data.

2) The system provides solution to the top level management in order to know the academic condition in their university.

3) The proposed system could be implemented to other university who need a decision support system for big data.

The remaining part of this paper is organized as follows. Section 2 presents the background and the related work. Section 3 presents the design of the system and Section 4 present the testing of the proposed system. Finally, the conclusions are drawn in Section 5.

2. Traditional data warehouse and modern data warehouse

This section describes about traditional data warehouse and modern data warehouse. The differences between them are also discussed.

Data warehouse is the combination of concepts and technologies that facilitate organizations to manage and maintain historical data obtained from operational and transactional applications . It helps knowledge workers (executives, managers, analysts) to make quicker and more informed decisions. Data warehouse is a new paradigm in strategic decision making environment. Data warehouse is not a product but an environment in which users can find strategic information . Data warehouse is a place to store information that is devoted to help make decisions . The Data warehouse contains a collection of logical data separate from the operational database and is a summary. Data warehouse allows the integration of various types of data from a variety of applications or systems. This ensures a one-door access mechanism for management to obtain information and analyze it for decision making. Data warehouse has several characteristics : subject-oriented, integrated data, nonvolatile, time-variant, and not normalized.

Data warehouse used data modeling technique called dimensional modeling technique. Dimensional modeling is a call-based model that supports high-level query access. Star Schema is a form of dimensional modeling scheme that contains a fact table at its center and dimensional tables. Fact table contains descriptive attribute that is used for query and foreign key process to connect to dimension table. Decision analysis attributes consist of performance measures, operational metrics, aggregate sizes, and all other metrics needed to analyze organizational performance. Fact table shows what is

supported by data warehouse for decision analysis. The dimension table contains attributes that describe the entered data in the fact table.

Extract, Transform, and Load (ETL) is a data integration process that extracts data from outside sources, transforms the data according to business needs, and stores it into data warehouse . The data used in the ETL process can come from a variety of sources including enterprise resource planning (ERP) applications, flat files, and spreadsheets.

Data warehouse support decision support system. Decision Support Systems (DSS) is a computer-based system that helps decision makers use the data and models available to solve problems . DSS functions combine the resources of each individual with the ability of the computer to improve the quality of the decision. DSS requires data coming from various sources to solve the problem. Every problem needs to be solved and every opportunity and strategy requires data. Data is the first component of the DSS architecture. The data relate to a state that can be simulated using a model that is the second component of the DSS architecture. Some systems also have knowledge which is the third component of the DSS architecture. The fourth user interacts with the system through a user interface which is the fifth component in the DSS architecture. In building the DSS, it is necessary to plan a mature system accompanied by the preparation and incorporation of components well.

Data warehouse is widely implemented, including in the education industry. It is possible to implement data warehouse for typical university information system . Academic data warehouse supports the decisional and analytical activities regarding the three major components in the university context: didactics, research, and management . Data warehouse has important role in educational data analysis.

With the arriving of big data, traditional data warehouse cannot handle large amount of data . In the past, educational data has been gathered mainly through academic information system and traditional assessments. However, it is increasingly being gathered through online educational systems, educational games, simulations and social media now. Huge workload, concurrent users and data volumes require optimization of both logical and physical design. Therefore, data processing must be in parallel. Moreover, traditional data warehouse cannot extract unstructured data that has varying data structure into information. Traditional data warehouse was design with the purpose of integrating structured data from transactional sources that is supported by OLAP-based analysis. It is the opportunity for big data technology to solve the problem. The integration between big data technology such as Hadoop and data warehouse is very important. To support parallel and distributed processing of large volumes of data, most solutions involve Hadoop technology. Hadoop is capable to perform analysis of large heterogeneous datasets at unprecedented speeds.

The Table 1 summarizes the characteristics of traditional data warehouse and modern data warehouse, from the several point of views like the purpose, data sources, scope, architecture, technology, and end-user.

3. System design

ETL is the main process in traditional data warehouse technology which cannot handle unstructured data. In this system, we need a flexible ETL process which can handle several data quality

issues, as for instance duplicated data, inconsistency data, and garbage data. The proposed system can be seen in Fig

1. In the system, there is a combination between Hadoop and RDBMS. Hadoop can enhances RDBMS as data ingestion/staging tool, but also as data management and data presentation platform.

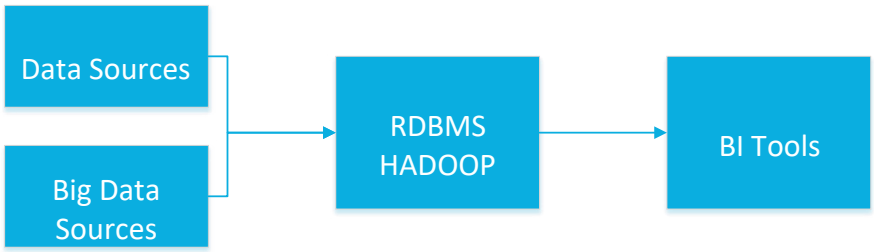
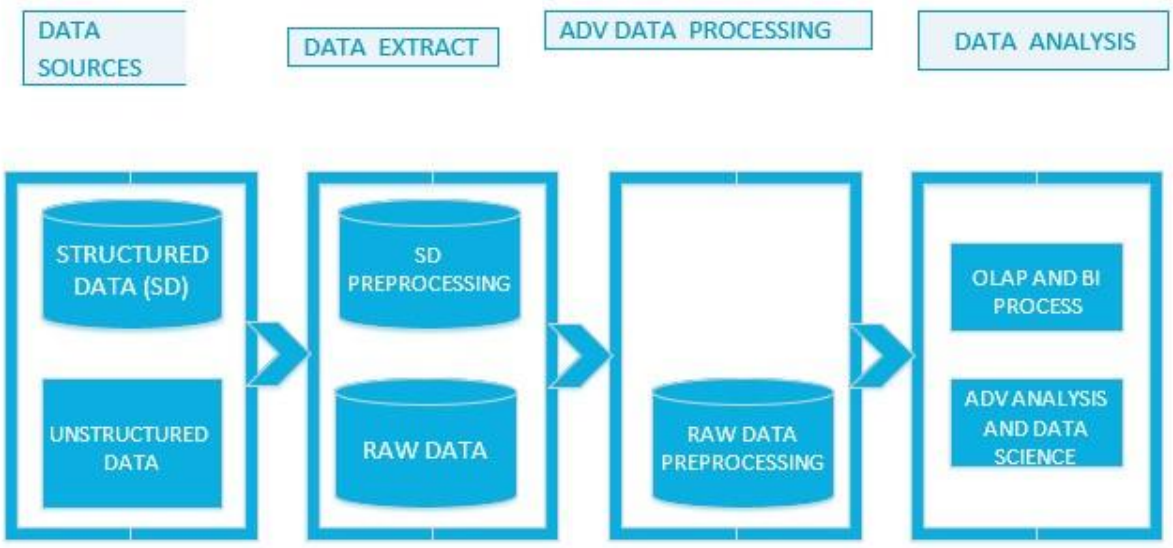


Fig. 1. The proposed system.



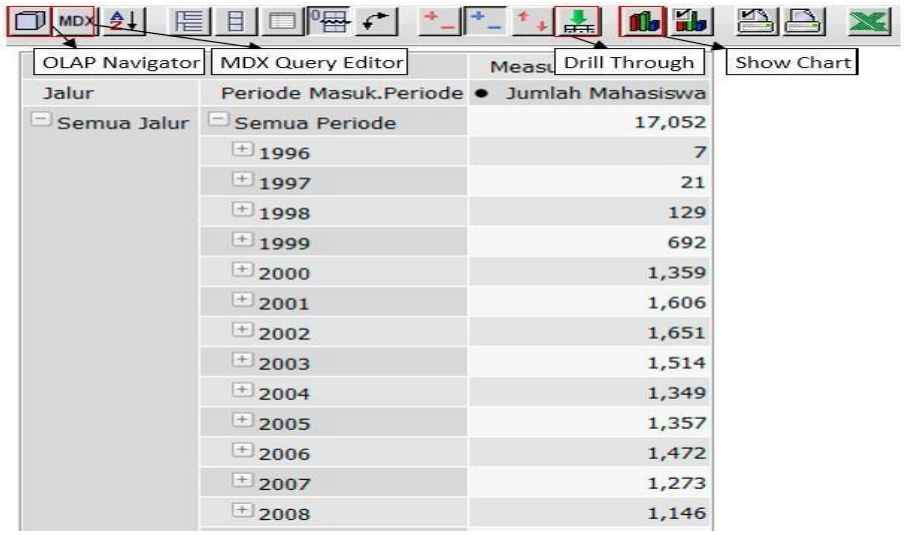
The architecture of system.

4. Implementation and testing

In this section will be discussed about the implementation of the system in accordance with the analysis and system design. The structured data comes from PostgreSQL databases, while unstructured data comes from social media such as Facebook, twitter and LinkedIn.

shows the analysis page. In this application, users could create new analysis so the report can be customised as they need. In every analysis, it is possible to produce some graphs or charts to support

the generated report. Some advanced users need OLAP Navigator and MDX Query Editor to create powerful report.



Jalur	Periode Masuk.Periode	Jumlah Mahasiswa
Semua Jalur	Semua Periode	17,052
	1996	7
	1997	21
	1998	129
	1999	692
	2000	1,359
	2001	1,606
	2002	1,651
	2003	1,514
	2004	1,349
	2005	1,357
	2006	1,472
	2007	1,273
	2008	1,146

Analysis page of DSS application.

The sample chart can be seen in User can customised the type of the chart, so the generated report will be more meaningful for the reader.

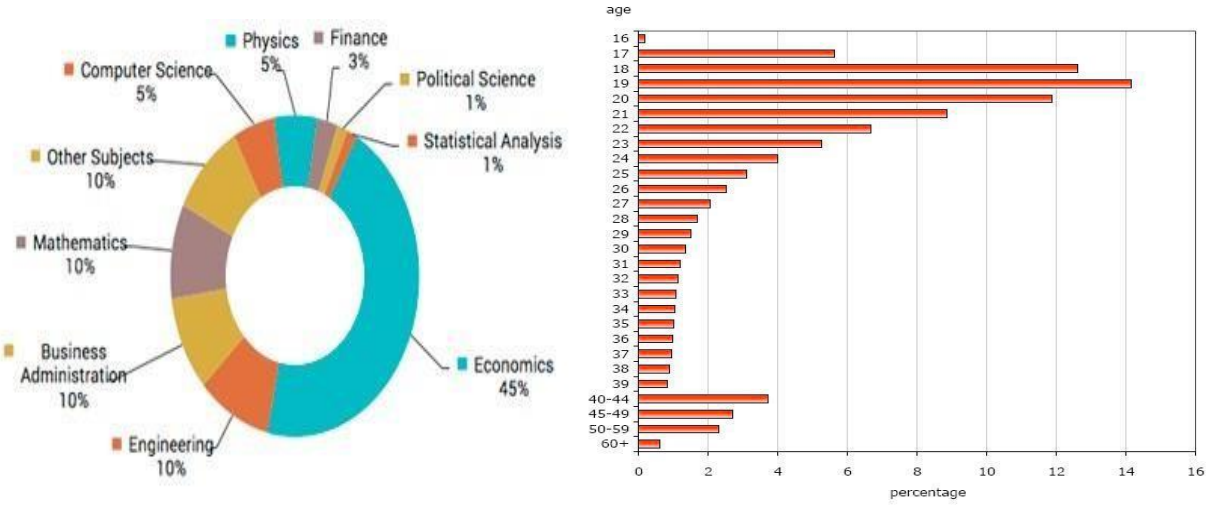
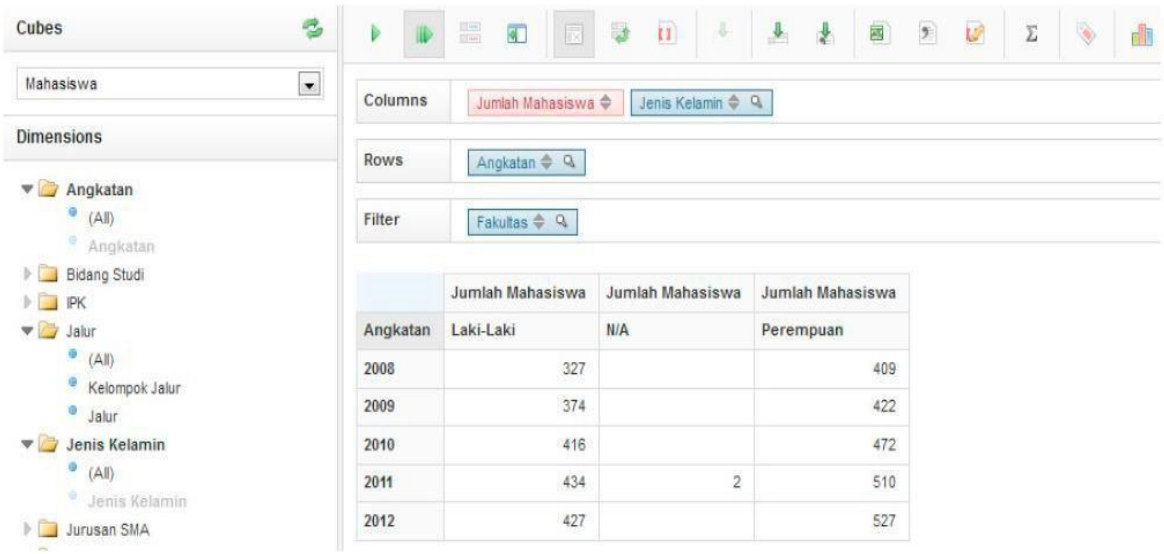


Chart page of DSS application.

The more advanced interface for analytical feature can be seen in Fig. 5. Advance users could drag and drop item dimensions in the left panel, then put into the column, row or filter in the right panel to produce the insightful report.



	Jumlah Mahasiswa	Jumlah Mahasiswa	Jumlah Mahasiswa
Angkatan	Laki-Laki	N/A	Perempuan
2008	327		409
2009	374		422
2010	416		472
2011	434	2	510
2012	427		527

Advanced analysis of DSS academic

Questionnaires were distributed among thirty university staffs which cover from top management, middle management and bottom management. Rector and vice rectors are categorized as top management. In the middle management, it contains dean, vice dean and their staffs. Head of departments are grouped into bottom management. The assessed indicators include application interface, graphic customization features, ease of use of the application, ability to meet the user needs, and overall application. Detailed assessment of application usage can be seen in Table 2.

Description of the rating scale:

- Value 1: Very bad
- Value 2: Bad
- Value 3: Enough
- Value 4: Good
- Value 5: Very good

From the questionnaire responses, 77% of users has good user experience. Mostly, the respondents said the graphics customization feature is good. According to the users, 87% of application is easy to use. 93% of users said the applications meet the business requirements. Overall, 97% of survey respondents said that the application is good.

5.INTEROPERABILITY AND METADATA

The heterogeneity in conceptual and logical models proposed for DWs, together with the wide variety of tools and software products available on the market, has lead to a broad diversity in metadata modeling. In practice, tools with dissimilar metadata are integrated by building complex metadata bridges, but some information is lost when translating from one form of metadata to another. Thus, there is a need for a standard definition of metadata in order to better support DW interoperability and integration, which is particularly relevant in the recurrent case of mergers and acquisitions.

Two industry standards developed by multi-vendor organizations have arisen in this context: the Open Information Model (OIM) by the Meta Data Coalition (MDC) and the Common Warehouse

Metamodel (CWM) by the OMG (see for a comparison of the two competing specifications). In 2000, MDC joined OMG for developing the CWM as a standard metadata model. The CWM is a platformindependent metamodel definition for interchanging DW specifications between different platforms and tools. It is based on the standards UML, XMI, and MOF, and basically provides a set of metamodels that are comprehensive enough to model an entire DW including data sources, ETL, multidimensional cubes, relational implementations, and so on. These metamodels are meant to be generic, external representations of shared metadata and to provide a framework for data exchange. Unfortunately, their expressivity is not sufficient to capture all the complex semantics represented by conceptual models, so they hardly can be used for effective integration of different DWs.

An alternative approach in this direction is described in [8], where a notion of dimension compatibility based on information consistency is proposed, aimed at cross-querying over autonomous, federated data marts. We believe that another interesting possibility for integration would be to use domain ontologies in order to establish semantic mappings between different data marts.

6.DESIGN FOR NEW ARCHITECTURES AND APPLICATIONS

Advanced architectures for business intelligence are emerging to support new kinds of applications, possibly involving new and more complex data types. The modeling and design techniques devised so far are mainly targeted towards traditional business applications, and aimed at managing simple alphanumerical data. Thus, it appears inevitable that more general, broader techniques will have to be devised. In this section we discuss the impact of some of the new applications and architectures on modeling and design; other related topics, that we do not address here due to space constraints, are active DWs and DWs for the life sciences.

6.1Spatial data warehousing

Spatial DWs are characterized by a strong emphasis on spatial data, coming in the form of spatial dimensions or spatial measures. Several works, like, show the advantages of using Geographic Information Systems (GIS) characteristics in the analysis of multidimensional data in specific domains. Other works, like , implemented more general systems mixing GIS and OLAP.

While all existing conceptual models support basic modeling of a spatial dimension (most business DWs include a geographic hierarchy built on customers), location data are usually represented in an alphanumeric format. Conversely, picking a more expressive and intuitive representation for these data would reveal patterns that are difficult to discover otherwise.

Preliminary approaches to conceptual modeling for spatial DWs are proposed in , where multidimensional models are extended with spatial dimensions, spatial hierarchies, and spatial measures. Also topological relationships and operators such as *intersect* and *inside* as well as userdefined aggregate functions are included to augment the expressivity of these models. From the point of view of logical modeling, the main issue raised by spatial warehousing is how to seamlessly integrate the classical ROLAP and MOLAP solutions (the star schema) with the

specialized data structures used in GISs while preserving high-level performance. In this line, investigates the definition of mappings between the geographical dimension of an OLAP tool and a GIS. Finally, as concerns design methods, adequate solutions for properly moving from conceptual to logical schemata in presence of spatial information must be devised.

6.2Web warehousing

Web warehouses are DWs that collect Web data. The characteristics of the Web raise new difficulties, mainly due to the semi-structured nature of data, to the lack of control over the sources, and to the frequency of changes on them.

The main challenges in this field are how to integrate heterogeneous web sources and how to automate the process of conceptual design when some or most data sources reside on the Web. Some attempts have been made in this direction, mainly aimed at building a conceptual schema from XML data . In other approaches, like , the design of the Web warehouse is driven by frequent user queries and by data quality. Importantly, the development of the Semantic Web opens new exciting possibilities since knowledge is represented according to formal ontologies capable of expressing semantic relationships, which will allow more powerful methods for conceptual design and for data integration to be devised.

6.3Real-time data warehousing and BPM

As DW systems provide an integrated view of an enterprise, they represent an ideal starting point to build a platform for business process monitoring (BPM). However, performing BPM on top of a DW has a deep impact on design and modeling, since BPM requires extended architectures that may include components not present on standard DW architectures and may be fed by nonstandard types of data (such as data streams). In particular, the fact that BPM implies real-time requirements leads to rethinking ETL components, making the ETL design techniques devised so far questionable. In addition, achieving satisfactory performance for continuous monitoring queries will require more sophisticated logical models for storing data cubes. Arising design issues are summarized in [26]:

- *Right-time design*. While strict real-time will not actually be needed for most applications, data processing must take place in so-called right-time, meaning that information must be ready and complete not later than required by the decision-making process. Thus, a relevant problem for the designer is to understand what is the right-time for the specific business domain.

- *KPI and rule design*. BPM architectures typically include dashboards for viewing key performance indicators (KPIs) and inference engines for managing business rules aimed at giving the decision maker an accurate and timely picture of the business. Hence, suitable techniques for modeling and designing KPIs and business rules, capable of establishing a conceptual connection with the related business goals and of coping with quickly changing requirements, will be necessary.

- *Process design*. In BPM a leading role is played by processes. Hence, BPM design also requires to understand business processes and their relationships in order to find out the relevant KPIs and rules, and to determine where the data to compute them can be found.

6.4Distributed data warehousing

As in distributed databases, in distributed data warehousing a new phase needs to be added to the design method: the one for designing the distribution, from both the architectural and the physical points of view. During architectural design, general decisions will be taken about which distribution paradigm (P2P, federation, grid) better suits the requirements, how to deploy the DW on the infrastructure, which communication protocols to use, etc. For example, makes the case for a P2P infrastructure for warehousing XML resources, whereas reports how DW systems can be deployed on a grid. On the other hand, the physical point of view mainly addresses how to fragment the DW and how to allocate fragments on the different sites in order to maximize local references to data and to take advantage of the intrinsic parallelism arising from distribution, thus optimizing the overall performance. Though some approaches to fragmentation of DWs have been tempted , they are mainly aimed at exploiting local parallelism or at designing ad hoc view fragments for a given workload.

Indeed, distribution is particularly useful in contexts where new data marts are often added, typically because of company mergers or acquisitions. In this case, the most relevant issue is related to integration of heterogeneous data marts as already mentioned in Section 5.

7.CONCLUSION

In this paper we have discussed open issues related to modeling and design of DWs. It is apparent that, though these topics have been investigated for about a decade, several important challenges still arise. Furthermore, ad hoc techniques are required for dealing with the emerging applications of data warehousing and with advanced architectures for business intelligence. Besides, the need for real-time data processing raises original issues that were not addressed within traditional periodically-refreshed DWs. Thus, overall, we believe that research on DW modeling and design is far from being dead, partly because more sophisticated techniques are needed for solving known problems, partly because of the new problems raised during the adaptation of DWs to the peculiar requirements of today's business.

We would like to thank the anonymous referees for their careful reading and constructive comments, which helped to improve the presentation. In addition, we would like to warmly thank all the friends who participated in the Dagstuhl Seminar for sharing their ideas with us: Alex

Buchmann, Karen Davis, Matteo Golfarelli, Joachim Hammer, Matthias Jarke, Manfred Jeusfeld,

Mirek Riedewald, Nick Roussopoulos, Markus Schneider, Timos Sellis, Alkis Simitsis, Dimitri

Theodoratos, A Min Tjoa, and Panos Vassiliadis. This paper is based upon our section "Design and Modeling" of an unpublished draft co-authored by all Dagstuhl participants. Our work has been partially supported by the Spanish Research Program PRONTIC and FEDER under project TIN200505406, by the Valencia Government (Spain)under DADASMECA, and by the CastillaLa Manc.