Doctoral Program in Informatics

Data Warehousing Systems

Proposal for a Course (2010-2011)

MAP-i – Joint Doctoral Program in Informatics
University of Minho, University of Porto, and University of Aveiro

Orlando Belo
obelo@di.uminho.pt

Department of Informatics
School of Engineering
University of Minho

Gabriel Davis
gtd@fe.up.pt

Department of Informatics Engineering
Faculty of Engineering
University of Porto

Maribel Santos
maribel@dsi.uminho.pt

Department of Information Systems
School of Engineering
University of Minho

Keywords: Decision Support Systems, Data Warehousing Systems, ETL, OLAP, Dimensional Modeling, and Operational Systems.

>>> Context

During his existence, companies based on their own experience have learn that the separation of operational systems requisites and functionalities from decision support ones it was the better policy and practice to follow in the improvement of their decision making abilities. Thus, it is not a big surprise to see that some of them have implemented effective decision support systems developing, step by step, a corporative data warehouse complemented, when necessary, with on-line analytical processing and sophisticated mechanisms of reporting. The implementation of a data warehousing system provides an efficient mean to store high level quality information, organizing it accordingly decision making agents' perspectives, and offering, as well, advanced resources to explore it dynamically.

>>> Objectives

This course was especially designed to present, discuss and deal with data warehousing systems, providing students with knowledge and skills to plan, design, implement, manage, and explore such systems for real-world application, and complement its functionalities through the integration of On-Line Analytical Processing (OLAP) infra-structures and services. All technological and scientific topics approached here are explored based on the implementation of data warehousing systems and its consequent exploitation through conventional means of database querying and reporting, or through on-line analytical processing mechanisms.

>>> Prerequisites

It is expect that students have basic knowledge about real-world database systems design, implementation, and administration.

>>> Learning Outcomes

Upon successful completion of this course, students should be able to:

> Understanding the mission and goals of a data warehousing system inside an organization, and characterize clearly the process how to implement them and justified the necessary investments.
> Know how to design a data warehousing system from scratch to its deployment, and consequent evaluation of its future evolution.
> Apply effectively dimensional modelling techniques projecting adequate structures to support a (useful) data warehouse, following de more realistic decision-making perspectives.
> Design and manage a data warehousing project, giving particular attention to system's end-users, database structures and services, and populating infrastructures and processes.
> Conceive and implement OLAP applications, knowing to explore effectively a multi dimensional database with appropriate querying languages.

>>> Program Contents

> Introduction to Decision Support Systems.
> Decision-making life cycle and decision-making systems implementation.
> From Operational to analytical systems.
> Data warehousing systems, architectures and services.
> The project of a data warehousing system.
    - Business case and justification.
    - Project planning, management, and risk evaluation.
    - Collecting and analysing requirements.
    - Technical architecture design and specification.
    - Dimensional modelling.
    - Databases physical design.
    - ETL systems design and development.
    - Business applications identification, specification and implementation.
    - System deployment and maintenance.
> Conceptual modelling for data warehouses.
    - Basics and first steps.
    - ER vs Dimensional modelling.
    - Decision matrix and data mart characterization.
    - Schemas and variations for data marts.
    - Fact tables and dimension tables design.
    - Natural and surrogate keys.
    - Slowly changing dimensions.
    - Snowflaking, outriggers, bridge tables, and others.
    - Transaction, periodic snapshot and accumulating fact tables.
> Extracting, transforming and loading data into a data warehouse.
    - Analysing data sources and data profiling.
    - Business rules definition for transformation processes.
    - Cleaning, conforming, and transforming data.
    - Data transformers and data flow controllers.
    - Populating dimension tables and appliance of updating policies.
    - Populating fact tables.
    - Monitoring and controlling ETL tasks and error handling.

- Managing and optimizing performance of a data warehousing system.
- Data webhouses and clickstream processing.
- Distributed data warehouses.
- Security models for ETL processes and storage structures.
- OLAP – Online Analytical processing.
    - Multidimensional structures.
    - Storage options – ROLAP, MOLAP and HOLAP.
    - Cube processing and optimization.
    - Multidimensional querying languages.
    - Applications.
- Tools and applications.

>>> Format

The course will be organized around formal lectures (60%) and practical demonstrations in laboratory (40%) of data warehousing systems applications. It is also planned a seminar period for presentation of real data warehousing systems scenarios presented by some of our industrial partners.

>>> Student Evaluation

The final evaluation of course's students will be based in a single component: a report about a data warehousing system design project.

>>> Lecturing Team

- **Orlando Belo.** Associate Professor at the Department of Informatics, School of Engineering, University of Minho.

  Orlando Belo is associate professor at the Department of Informatics of Minho University, and a researcher at CCTC in the areas of Data Warehousing Systems, OLAP, and Data Mining. His main research topics are related with data warehouse design, ETL services, and distributed multidimensional structures processing. Currently, he is the coordinator of the Decision Support Systems curricular unit of the first cycle level in Informatics, and of the Informatics Engineering licenciatura. During the last years he maintained several R&D projects with industrial companies and academic partners, in particular related to the implementation of business intelligent platforms and data mining applications in areas such as fraud detection, specific data warehousing systems (raw material transportation and recycling material), data preservation (with Gabriel David). For additional information, please visit http://www.di.uminho-pt/~omb.

- **Gabriel David.** Associate Professor Informatics Engineering Department, Engineering Faculty of the University of Porto.

  Gabriel David is currently Associate Professor at the Informatics Engineering Department, Engineering Faculty of the University of Porto (FEUP), where he integrates FEUP Executive Board and the Scientific Committees of the Information Science Bachelor and Master Programs. He leads the development team of SIGARRA, the U.PORTO Academic Information System. He has been a Researcher at INESC since 1985. His main research interests are in Information Systems, Databases, and Information Management. He has been the leader of the project MetaMedia (funded by Portuguese FCT) on multimedia archives and is currently working (as well as Orlando Belo) on the project DBPreserve(Portuguese FCT) on preservation of databases.

> **Maribel Santos.** Assistant Professor at the Department of Information Systems, School of Engineering, University of Minho.

Maribel Yasmina Santos is an Assistant Professor in the Department of Information Systems at the University of Minho in Portugal. She has a degree in Informatics and Systems Engineering from the University of Minho (1991), a MSc in Informatics and a Ph.D. in Information Systems and Technologies, both from University of Minho (1996 and 2001, respectively). Her research interests include business intelligence, data mining, geographic information systems, spatial reasoning and space models.

>>> Some Publications of the Lecturing Team

> Catalin Calistru, Cristina Ribeiro, Gabriel David, "Multidimensional descriptor indexing: Exploring the BitMatrix Multidimensional descriptor indexing: Exploring the BitMatrix". Lecture Notes in Computer Science Series, nº 4071, pp.401-410, 2006

> Lopes CT, David G, "Higher Education Web Information System Usage Analysis with a Data Webhouse" in ICCSA 2006 Intl. Conf. on Computational Science and Its Applications, Lecture Notes in Computer Science 3983: 78-87, 2006.

> Lourenço, A., Belo, O., Applying Clickstream Data Mining to Real-Time Web Crawler Detection and Containment Using ClickTips Platform, Advances in Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, pp 351-358, Springer, 2007.

> Loureiro, J., Belo, O., Swarm Intelligence in Cube Selection and Allocation for Multi-node OLAP Systems, in Advances and Innovations in Systems, Computing Sciences and Software Engineering, Elleithy, K., (ed), Springer, 2007.

> Alves, R., Belo, O., Ribeiro, J., Mining Top-K Multidimensional Gradients, 9th International Conference DaWaK 2007, Data Warehousing and Knowledge Discovery, LMCS 4654, pp 375-384, Regensburg, Germany, September, 2007.

> Loureiro, J., Belo, O., A Metamorphosis Algorithm for the Optimization of a Multi-node OLAP System, Progress in Artificial Intelligence, 13th Portuguese Conference on Artificial Intelligence, EPIA'2007, LNAI 4874, pp 383-394, Guimarães, Portugal, Dezembro, 2007.

> Moreira, Adriano e Maribel Yasmina Santos, "Concave Hull: A k-nearest neighbours approach for the calculation of the concave hull of a set of points", Proceedings of the 2nd International Conference on Computer Graphics Theory and Applications (GRAPP'2007), Barcelona, Spain, 8-11 March, 2007, pp. 61-68 (ISBN 978-972-8865-71-9).

> Sérgio Nunes, Cristina Ribeiro, Gabriel David, Use of Temporal Expressions in Web Queries in ECIR'08, pp.580-584, 2008.

> Gabriel David, Data Warehouses in the Path from Databases to Archives in International Workshop on Database Preservation, 2007-03-23, National e-Science Centre, Edinburgh, Scotland, 2007 (http://homepages.inf.ed.ac.uk/hmueller/presdb07/papers/PRESDB_GTD.pdf, seen on 2008-01-31)

> Santos, Maribel Yasmina, e Luís Amaral, "Mining Geo-referenced Databases: a way to improve decision-making", In James Pick (Ed.), GIS in Business, Idea Group Publishing, 2005, pp. 113-149, ISBN 1-59140-400-2 (URI: http://hdl.handle.net/1822/1353 <http://hdl.handle.net/1822/1353>).

> Maribel Yasmina Santos e Adriano Moreira, "Decision Trees in the Identification of Space Models", The Mediterranean Journal of Computers and Networks, Special Issue on Mobile and Ubiquitous Systems, 3 (1), January, 2007, ISSN: 1744-2397, pp. 15-23 .

> Santos, Maribel Yasmina, e Luís Alfredo Amaral, "Geo-Spatial Data Mining in the Analysis of a Demographic Database", Soft Computing – A Fusion of Foundations, Methodologies and Applications, Special Issue on Soft Computing for Spatial Data Analysis, May, 9(5), 2005, pp. 374-384 (ISSN 1432-7643 – Paper, 1433-7479).

>>> References

> The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaninng, Conforming and Delivering Data, Ralph Kimball, Joe Caserta, Wiley, September, 2004. ISBN-13: 978-0764567575.
> The Data Warehouse Lifecycle Toolkit, Ralph Kimball, Margy Ross, Warren Thornthwaite, Joy Mundy, Bob Becker, Wiley, 2nd ed, January, 2008. ISBN-13: 978-0470149775.
> The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, Ralph Kimball, Margy Ross, Wiley; 2nd ed, April 26, 2002. ISBN-13: 978-0471200246.
> Building the Data Warehouse, W. H. Inmon, Wiley, 4th ed, October, 2005. ISBN-13: 978-0764599446.
> Mastering Data Warehouse Design: Relational and Dimensional Techniques, Claudia Imhoff, Nicholas Galemmo, Jonathan G. Geiger, Wiley, August, 2003. ISBN-13: 978-0471324218.
> Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance, Christopher Adamson, Wiley, July, 2006. ISBN-13: 978-0471777090.
> Data Webhouse Toolkit : Building the Web-Enabled Data Warehouse, Kimball, R., Merz, R., John Wiley, 2000.
> Database Systems - A Practical Approach to Design, Implementation, and Management, Connolly, T., Begg, C., III Edição, Addison-Wesley, 2001.

>>> Other Resources

> P. Vassiliadis, A. Simitsis, M. Terrovitis, and S. Skiadopoulos. Blueprints for ETL workflows. In Proceedings of the 24th International Conference on Conseptual Modeling (ER'05), volume 3716 of LNCS, pages 385--400. Springer, 2005.
> A. Simitsis, P. Vassiliadis, M. Terrovitis, and S. Skiadopoulos. Graph-Based Modeling of ETL Activities with Multi-Level Transformations and Updates. In Proceedings of the 7th Int'l Conference on Data Warehousing and Knowledge Discovery (DaWaK'05), volume 2589 of LNCS, pages 43--52. Springer, 2005.
> R. Kimball. The 38 Subsystems of ETL. www.intelligententerprise.com. December 4, 2004.
> M. Ross, R. Kimball. Slowly Changing Dimensions Are Not Always as Easy as 1, 2, 3. www.intelligententerprise.com. March 1, 2005.
> R. Kimball. Pipelining Your Surrogates - A good surrogate key system is worth the work. www.dbmsmag.com. June, 1998.

>>> Tools

> Microsoft SQL Server 2008.
> Microsoft Integration Services 2008.
> Microsoft Analysis Services 2008.
> Microsoft Excel 2008.

>>> Site

> http://www.di.uminho.pt/~omb/mapidws (to be developed).