Doctoral Program in Informatics

Data Warehousing Systems

Proposal for a Curricular Unit (2013/2014)

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 David
gtd@fe.up.pt
Department of Informatics
Engineering Faculty of
Engineering University of Porto

Maribel Yasmina Santos
maribel@dsi.uminho.pt
Department of Information
Systems School of Engineering
University of Minho

Keywords: Decision Support Systems, Data Warehousing Systems, Dimensional Modelling, and On-Line Analytical Processing.

>>> Context

The requirements of transactions processing systems in large companies lead to the design of normalized databases. However, the requirements of decision support systems for those organizations, even when building on the same data, are very different and lead to the design of different data models. Thus, it is not a big surprise to see that some of them have implemented effective decision support systems developing, step by step, a corporate 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 means to store high volumes of quality information, organizing it according to the 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 their 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 of how to implement them and justify 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 the 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 application of updating policies. o 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.
- OLAP – Online Analytical processing.
    - Multidimensional structures.
    - Storage options – ROLAP, MOLAP and HOLAP. o 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 with Habilitation, at the Department of Informatics, School of Engineering, University of Minho.

  Orlando Belo is Associate Professor with Habilitation at the Department of Informatics, School of Engineering, University of Minho, and a researcher at ALGORITMI Centre in the areas of Business Intelligence, Data Warehousing Systems, and On-Line Analytical Processing. 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 MSc Course in Informatics Engineering. During the last years he maintained several R&D projects with industrial companies, in particular related to the implementation of business intelligent platforms and data mining applications. For additional information, visit http://www.di.uminho-pt/~omb.

- **Gabriel David.** Associate Professor at the 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 Scientific 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 until 2010. He has led 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 the project DBPreserve (Portuguese FCT) on preservation of databases.

- **Maribel Yasmina Santos.** Associate Professor with Habilitation, at the Department of Information Systems, School of Engineering, University of Minho.

  Maribel Yasmina Santos is an Associate Professor at the Department of Information Systems, University of Minho, in Portugal. Maribel received the Aggregated title

(Habilitation) in Information Systems and Technologies from the University of Minho in 2012 and a Ph.D. in Information Systems and Technologies from the University of Minho in 2001. Maribel has a degree in Informatics and Systems Engineering and a MSc in Informatics both from the University of Minho (1991 and 1996, respectively). Since 1997, her research work has integrated decision support systems with GI Science, given particular emphasis to the development of spatial decision support systems, spatial data mining algorithms, spatial databases, spatial data warehousing and spatial on-line analytical processing systems, among others. Maribel has participated in several research projects funded by national and international institutions, and has authored and co-authored more than 80 international publications, including papers in journals, book chapters and papers in international conferences. Since May 2010, Maribel is sub-director of the Department of Information Systems at the University of Minho in Portugal. Since April 2011, she integrates the AGILE (Association for Geographic Information Laboratories in Europe) council. For additional information, please visit http://www.dsi.uminho-pt/~maribel.

>>> Some Publications of the Lecturing Team

> Ademar Aguiar, Gabriel David. “Patterns for Effectively Documenting Frameworks” in “Transactions on Pattern Languages of Programming II”, Lecture Notes in Computer Science 6510, pp.79-124, Springer-Verlag Berlin Heidelberg, 2011.

> BELO, O., MONSANTO, P., LOURENÇO, A., “Signature Based Credentials - An Alternative Method for Validating Student Access in eLearning Systems”, In Proceedings of 12th European Conference on e-Learning (ECEL-2013), Sophia Antipolis, France, 30-31 October, 2013.

> BELO, O., RODRIGUES, P., BARROS; R., “Adaptive Dashboarding - Reflecting Usage Preferences in OLAP”, In Proceedings of European Conference on Data Analysis (ECDA’2013), Luxembourg, 10-12 July, 2013.

> Catalin Calistru, Cristina Ribeiro, Gabriel David, Multimedia in Cultural Heritage Manuscripts: Integrating Description, Transcription, and Image Content EURASIP Journal on Image and Video Processing, Vol.2009 nº 2009, pp.1-9, 2009.

> Sérgio Nunes, Cristina Ribeiro, Gabriel David. “Term Weighting based on Document Revision History”. Journal of the American Society for Information Science and Technology, Vol.62 nº 12, pp.2471-2478, 2011.

> Sérgio Nunes, Cristina Ribeiro, Gabriel David. "Term Frequency Dynamics in Collaborative Articles" in 10th ACM Symposium on Document Engineering (DocEng2010), Manchester, UK, 2010.

> OLIVEIRA, B., BELO, O., “Approaching ETL Conceptual Modelling and Validation Using BPMN and BPEL”, In Proceedings of The 2nd International Conference on Data Management Technologies and Applications (DATA-2013), Reykjavik, Iceland, July 29-31, 2013.

> Painho, Marco, Maribel Yasmina Santos and Hardy Pundt (Eds.), “Geospatial Thinking”, Lectures Notes in Geoinformation and Cartography, Springer-Verlag, 1st Edition, 418 p., 2010, May, ISBN: 978-3-642-12325-2 (URI: http://hdl.handle.net/1822/11596).

> Pereira, Óscar M., Rui L. Aguiar, and Maribel Yasmina Santos, “CRUD-DOM: A Model for Bridging the Gap Between the Object-Oriented and the Relational Paradigms - an Enhanced Performance Assessment Based on a Case Study”, International Journal on Advances in Software, Vol. 4, n∫ 1 & 2, 158-180, ISSN 1942-2628, pp. 158-180, 2011 (URI: http://hdl.handle.net/1822/13937).

> Arif Ur Rahman, Gabriel David, Cristina Ribeiro.“SIARD Archive Browser” in Theory and Practice of Digital Libraries, pp.496-499, 2012.

> Arif Ur Rahman, Gabriel David, Cristina Ribeiro. “Transformation Rules for Model Migration in Relational Database Preservation” in Proceedings of the 8th International Conference on Preservation of Digital Objects, pp.-, 2011.

- Arif Ur Rahman, Gabriel David, Cristina Ribeiro, “Model Migration Approach for Database Preservation” in The Role of Digital Libraries in a Time of Global Change, pp.81-90, 2010.
- RIBEIRO, D., SANFINS, A., BELO, O., “Wastewater Treatment Plant Performance Prediction with Support Vector Machines”, In Proceedings of 13th Industrial Conference on Data Mining (ICDM’ 2013), LNAI Springer, New York, USA July 16-21, 2013.
- SANTOS, J., BELO, O., “Estimating Risk Management in Software Engineering Projects”, In Proceedings of 13th Industrial Conference on Data Mining (ICDM’ 2013), LNAI Springer, New York, USA July 16-21, 2013.
- Santos, Maribel Yasmina, Joaquim Silva, João Moura-Pires and Monica Wachowicz, “Automated Traffic Route Identification through the Shared Nearest Neighbour Algorithm”, Proceedings of the 15th AGILE International Conference on Geographic Information Science, April 24-27, 2012, Avignon, France, Springer-Verlag (URI: http://hdl.handle.net/1822/19194).
- Santos, Maribel Yasmina and Adriano Moreira, “GUESS: on the prediction of mobile usersí movement in space”, in Monica Wachowicz (Ed.), Movement-aware Applications for Sustainable Mobility: Technologies and Approaches, IGI Global Publishing, 2010, pp. 87-104, ISBN: 978-1615207695 (URI: http://hdl.handle.net/1822/12998).
- Santos, Maribel Yasmina e Isabel Ramos, “Business Intelligence ñ Tecnologias de InformaÁ„o na Gest„o de Conhecimento”, 2ª Edição Actualizada e Aumentada, FCA Editora de Informática, Fevereiro, 2009, ISBN 978-972-722-516-3 (URI: http://hdl.handle.net/1822/11110).
- SILVA, D., FERNANDES, J.M., BELO, O., “Assisting Data Warehousing Populating Processes Design Through Simulation Using Coloured Petri Nets”, In Proceedings of 3rd Industrial Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH’ 2013), Reykjavik, Iceland, July 29-31, 2013.
- Silva, Ricardo, João Moura-Pires and Maribel Yasmina Santos, ìSpatial Clustering in SOLAP systems to Enhance Map Visualizationî, International Journal of Data Warehousing and Mining (SCImago Journal Rank Q1, Indexed By SCOPUS, ISI, DBLP), Vol. 8, n∫ 2, ISSN (printed): 1548-3924. ISSN (electronic): 1548-3932, 2012 (URI: http://hdl.handle.net/1822/19182) (DOI: 10.4018/jdwm.2012040102).

>>> 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 2012.
- Microsoft Integration Services 2012.
- Microsoft Analysis Services 2012.
- Microsoft Reporting Services 2012.
- Microsoft Excel 2010.

>>> Site

- http://www.di.uminho.pt/~omb/mapidws.