DBTech EXT Virtual Lab Workshop: Business Intelligence*

D.A. Dervos[1], G. Evangelidis[2], L. Karamitopoulos[1], L. Aunimo[3], J. F. Aldana Montes[4], J.J. Farfán-Leiva[4], K. Siakas[1], H. Valkama[3]

[1]Information Technology Dept., Alexander Technology Educational Institute, Thessaloniki, Greece, Tel. +30.2310791295, email: {dad, lkaramit, siaka}@it.teithe.gr,
[2]Applied Informatics Dept., University of Macedonia, Thessaloniki, Greece, Tel. +30.2310891844, email: gevan@uom.gr,
[3]Haaga-Helia University of Applied Sciences, Helsinki, Finland, Tel. +35.8922965525, email: {lili.aunimo, hannu.valkama}@haaga-helia.fi,
[4]Departamento de Lenguajes y Ciencias de la Computación, Málaga, Spain, Tel. +34.952132813, email: jfam@lcc.uma.es, jjfl@satd.uma.es

Database Technology Extended (DBTech EXT) is the continuation of the Database Technology Professional (DBTech Pro) project. Both projects comprise actions of the DBTechNet consortium [1]. DBTech EXT involves eleven inner-circle (core) partners from five EU member states: seven Higher Education (HE) academic institutions, one company, and three VET centers. Academic institutions, organizantions, companies, and VET centers are invited to join-in as outer-circle members and act as project deliverables recipients. At project initiation time (January 2009), its outer-circle membership involved eight companies, and four academic institutions. In this paper, a typical DBTech EXT VLW (*Business Intelligence*) is outlined.

1. Introduction

Database technology comprises the underlying framework of today's information systems and has a major impact to the way organizations and individuals administer, organize, and process data and information today. In response to this reality and need, database technology forms a core area of study in undergraduate and postgraduate academic curricula relating to computer science, information technology, and information systems, as well as in professional knowledge and skills certification programs [3,4]. Moreover, database technology, much like a boiling pot, embraces all new ICT developments, including a wide spectrum of techniques and methodologies in relation with data and information administration and processing. Long gone is the time when the RDBMS consisted of just its relational data processing engine. Today, algorithms and data are unified by integrating programming and relational query languages. Extensible object-relational DBMS systems support non-procedural relational operators that manipulate complex object sets. Data warehouses organize data in multi-dimensional structures, and online analytical processing is baked in most database environments. In addition, the latter support a framework of data mining and machine learning algorithms, text, temporal, and spatial data access methods, web-DBMS integration APIs, XML, XQuery, .NET / JavaEE, SOA, etc. [5].

* *This report is the result of collaborative work undertaken along the lines of the DBTechNet Consortium. All the co-authors participate in DBTech EXT, a project partially funded by the EU LLP Transversal Programme (Project Number: 143371-LLP-1-2008-1-FI-KA3-KA3MP)*

DBTechNet was established in 1997 as an initiative to focus on the need for (a) identifying the differences in the way database technology topics are addressed in academic curricula across the European member states, and (b) the academics who teach database courses to join forces in order to cope with the diversity of technologies and specializations that co-exist in the modern DBMS application design and development environment.

One of the key areas of activity in the DBTech EXT project is the development and the support of virtual laboratory workshops (VLWs). The latter are offered over the internet to selected audiences of university students, VET trainees, and IT professionals who wish to update their knowledge and skills with the recent developments and new trends in database technologies. Figure 1 outlines the structure of a typical DBTech EXT VLW [6].

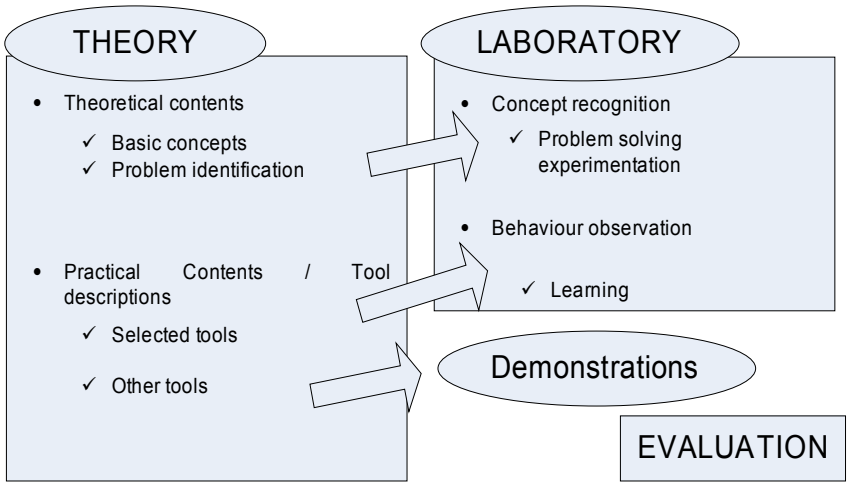

Figure 1: DBTech EXT VLW structure

The content of the typical VLW includes lecture slides, digital video lectures and software 'how-to' presentations, self-training exercises with model answers, hands-on laboratory exercises, assessment exercises, etc. The trainees who participate in the DBTech EXT VLWs are the project's inner- and outer- circle partners, alike. In the course of their distance-learning type of hands-on laboratory training, VLW participants receive guidance and assistance from local trainers as well as from a network of DBTechNet experts located at disparate geographic locations.

2. The Business Intelligence VLW (BI-VLW)

The educational and training content of the BI-VLW may be seen to consist of two logical units: (a) the data preparation and organization phase, including online analytical data processing (Data Warehousing and OLAP), and (b) the information extraction/discovery phase, utilizing data mining algorithms and techniques (Knowledge Discovery from Databases, KDD).

2.1 Data Warehousing and OLAP

The first part of the BI-VLW begins by introducing the participant to the data extraction, transformation, and loading (ETL) process. Next, the Data Warehousing architectures are considered, together with advanced indexing implementations that facilitate OLAP operations. In the hands-on laboratory part, the participant is to utilize both open source (GNU licenced), and commercial software to construct the data warehouse (DW), populate it with data, and conduct OLAP analytics on the latter, utilizing both Java- as well as web-based client GUIs.

Data preparation and ETL operations are to be conducted utilizing both the MS SQL Server® (commercial, [7]), and MySQL® (GNU licence, [9]) software. Data warehouse construction and OLAP will be conducted utilizing the MS Analysis Services® (commercial, [8]), as well as PALO® (Open Source, [10]), and Mondrian Pentaho® (Open Source, [11]) software.

2.2 Knowledge Discovery from Databases (KDD)

Having introduced the participant to the need of pre-processing and organizing the data in order to proceed and extract information from them, the concepts and the representations involved are revisited. Data and metadata are differentiated from information [12] which emerges in the form of a number of representations in the output of the knowledge discovery / data mining process [13]. Next, a number of data mining algorithms are considered and realized in the form of self-training exercises, the model answers provided. The latter focus on affinity analysis (association rules mining), supervised learning (decision tree classification), and clustering (K-Means, BIRCH, demographic).

In the hands-on laboratory session, the participant is introduced to the use of the IBM DB2 DWE Intelligent Miner® (commercial, [14]), and the WEKA® (GNU license, [15]) software. Use is made of the IBM “Mining your business in retail” tutorial [16] which utilizes SQL code and the Easy Mining Procedures for the IBM Intelligent Miner® for hands-on training on the following:

Association Rules Mining

- Data exploration phase
- Data preparation phase (SQL views)
- Association rules model, building of
- Evaluation phase: the rules (SQL View), and IM Visualization
- Deployment phase: products recommendation

Clustering

- Data preparation phase (SQL views)
- Model building-1: demographic clustering
- Evaluation phase-1: IM Visualization
- Model building-2: improved clustering model
- Evaluation phase-2: results interpretation, cluster analysis
- Deployment phase: Improved products recommendations

BI-VLW participants are required tackle a number of hands-on laboratory exercises and file their answers in the form of a laboratory report which they submit (electronically, from a distance) to their instructors. They make use of portal usage log data made available with the kind permission of the Hellenic Academic Libraries Link (HEAL-link, [17]). Through HEAL-link, the academic communities in Greece and Cyprus can search and retrieve the electronic full text versions of research articles published with some 9,000+ scientific journals. They also make use of a dataset relating to accident of the TITANIC ocean liner, originating from [18]. More specifically, they make use of the IBM Intelligent Miner software to conduct association rules mining on the HEAL-link data, and they are required to use the WEKA to conduct clustering analysis on the TITANIC data.

3. The UMA VDI service

As it is mentioned in sections 1 and 2, (a) the BI-VLW utilizes a wide range of relevant software, part of which is commercial and accessible only via academic licensing [19, 20]. To make it possible for the industrial and VET center participants to have access to the complete range of software resources relating to the BI virtual workshop, use is made of the University of Malaga (UMA) virtual desktop infrastructure and service (VDI, [2,6]).

VDI provides remote access to a pool of virtual machines through a connection broker. This allows:

- Access control to the virtual machines by means of directory attributes (LDAP, Active Directory).
- Dynamic supply of virtual machines, according to their purpose, thanks to the concept of Linked PC. Linked PC supports multiple and customizable instances of a virtual machine using “templates”.
- Automated management of the linked PC life cycle (creation, supply, destruction), optimizing the VDI infrastructure management.
- Assignment of independent virtual machines from a VDI pool of resources to users.
- Access to virtual machines with different operating systems (Linux, Windows 2000, 2003, XP, Vista).
- Linked PC assignment can be persistent or not persistent.

In order to access a virtual machine from a client, a local area network (LAN) or a private or public wide area network (WAN) is needed. Clients (PCs, Laptops, thin clients) need a browser which supports the ActiveX technology (IE), or the java applet technology (IE, Mozilla Firefox, Safari, etc).

The broker supports different kinds of connections: Remote Desktop Protocol (RPD), Remote Desktop Protocol with Re-compressor (RDP Rec), Virtual Network Computing (VNC) and Virtual Network Computing with Recompressor (VNC Rec).

Six (6) virtual machines can be setup and run concurrently in the VDI infrastructure. The number of concurrently running virtual machines depends on the number of processors and the number of processor cores. Six virtual machines per core is a good estimate. For example, one infrastructure with two VMware servers, each with two quad-core processors and 32 GB of RAM may operate 96 concurrent virtual machines, each with 64 GB of RAM.

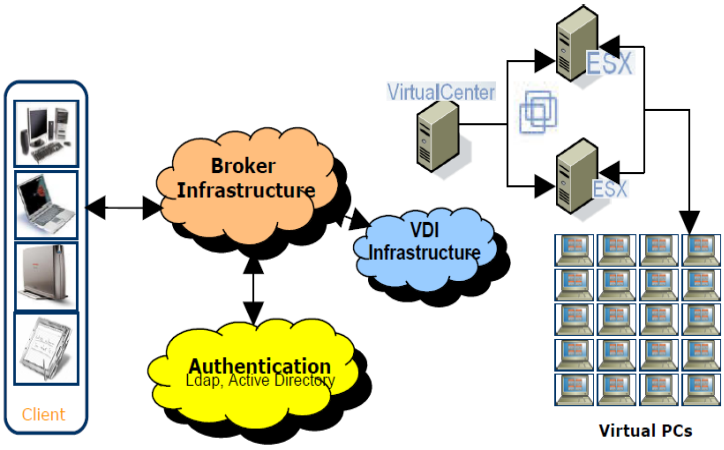

Figure 2: UMA Virtual Desktop Infrastructure Architecture

The UMA VDI infrastructure architecture is presented graphically in Figure 2, and involves:

- VMware infrastructure virtualization servers (Virtual center plus n ESX VMware servers),
- A Connection broker,
- Re-compressors, and
- Additional servers: DNS servers, DHCP servers, databases, authentication servers (LDAP, ActiveDirectory).

4. Pedagogical Evaluation

The DBTech EXT projects includes a workpackage on quality assurance and continuous quality management, process evaluation, and assessment of the pedagogical methodologies applied. The plan for quality management incorporates the continuous assessment of the whole project, including flow of events, working group outcomes, meetings, communication media, and administration. Regarding the pedagogical quality, the demands of prior learning recognition, credit transfer and industry needs are taken into consideration. The learning outcomes for knowledge, skills and targeted competences are created by using pedagogic models, such as the European Qualification Framework (EQF, [21]).

For the DBTech EXT workshops in particular, a *Pre-teaching plan* template aims to help planning groups and instructors to focus on relevant topics in the planning of workshops. Next, a *Peer review* template serves as a tool for collecting comments and observations, aiming to achieve coherence in the releveant information. An on-line *Questionnaire form* collects the views of workshop participants on the arrangements, materials and content, processes and the outcomes of learning. In addition structured interviews including open ended questions will be carried out with selected learners and instructors in order to record additional instances of useful, unbiased views. The aim is to compile a SWOT analysis to identify strengths (what was good in the workshop), weaknesses (what did not work or was not suitable), opportunities (appropriateness to market related requirements in setting up analogous project activities in the future), and threats (what can go wrong in the workshop) [22].

5. Conclusion

The DBTech EXT project activities include the development of virtual laboratopry workshops (VLWs) offered over the internet for training I.T. professionals on topics relating to new trends in database technology. In this paper, a typical DBTech EXT VLW (*Business Intelligence*) is presented. The organization of its educational and training (ET) content, the trainer and trainee activities involved, the proprietary and open-source software used, the self-study and assessment laboratory exercises, as well as the quality evaluation procedure applied are outlined. The utilization of proprietary software possessed a challenge to the DBTechNet community: virtual workshop participants include trainees from I.T. companies and VET centers who are not entitled to use proprietary software made available via academic licencing (i.e. for educational use only). In this respect, use is made of the University of Malaga PCVirtual/VDI internet based service [2].

References

1. The DBTechNet website: http://www.dbtechnet.org. Accessed: July 26, 2009.
2. The University of Malaga *PCVirtual/VDI* service: https://pcvirtual.cv.uma.es. Accessed: July 26, 2009.
3. Curricula Recommendations, Association for Computing Machinery, http://www.acm.org/education/curricula-recommendations. Accessed: July 26, 2009.
4. European Certification of Informatics Professionals (EUCIP), http://www.eucip.com/. Accessed: July 26, 2009.
5. J. Gray, The Revolution in Database Architecture, Extended abstract of keynote talk at ACM SIGMOD 2004, Paris, France, June, 2004, Also MSR-TR-2004-31, March 2004, http://research.microsoft.com/pubs/64551/tr-2004-31.pdf. Retrieved: July 26, 2009.
6. I. Navas-Delgado, M. del Mar Roldán-García, J. J. Farfán-Leiva, J. F. Aldana-Montes, M. Laiho, F. Laux, D. A. Dervos, DBTechNet: Innovation in Education and Training on Information Management Topics, Proceedings, 5th International Conference on Education, pp. 63-67, Bilbao, Spain, June 2009.
7. Microsoft SQL Server 2005 website http://www.microsoft.com/Sqlserver/2005/en/us/. Accessed: July 29, 2009.
8. Microsoft SQL Server 2005 Analysis Services website, http://www.microsoft.com/Sqlserver/2005/en/us/analysis-services.aspx. Accessed: July 29, 2009.
9. MySQL website, http://www.mysql.com/. Accessed: July 29, 2009.
10. PALO website, http://www.jedox.com/en/home/overview.html. Accessed: July 29, 2009.
11. Mondrian Pentaho website, http://mondrian.pentaho.org/. Accessed: July 29, 2009.
12. Dervos D., Coleman A., "A Common Sense Approach to Defining Data, Information, and Metadata", Proceedings of the Ninth International ISKO Conference, Ed. G. Budin, C. Swertz, K. Mitgutsch, Advances in Knowledge Organization, Vol. 10, Ergon Verlag, pp. 51-58, Vienna, Austria, July 2006, http://dlist.sir.arizona.edu/1516/. Retrieved: July 29, 2009.
13. I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann 2005, http://www.cs.waikato.ac.nz/~ml/weka/book.html. Accessed: July 29, 2009.
14. IBM DB2 DWE Intelligent Miner for Data, http://www-01.ibm.com/software/data/iminer/. Accessed: July 29, 2009.
15. WEKA 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/. Accessed: July 29, 2009.
16. IBM tutorial: Mining your business in retail with IBM DB2 Intelligent Miner, http://www.ibm.com/developerworks/edu/dm-dw-dm-retail_tutorial-i.html. Accessed: July 29, 2009.
17. HEAL-link portal, http://www.heal-link.gr/journals/en/. Accessed: July 29, 2009.
18. R.J. MacG. Dawson, The "Unusual Episode" Data Revisited, Journal of Statistics, v.3, n.3, 1995, http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html, Retrieved: July 29, 2009.
19. IBM academic initiative program, http://www.ibm.com/developerworks/university/academicinitiative/. Accessed: July 29, 2009.
20. Microsoft MSDN Academic Alliance program, http://msdn.microsoft.com/en-us/academic/. Accessed: July 29, 2009.
21. The European Qualifications Framework for Lifelong Learning, http://ec.europa.eu/education/policies/educ/eqf/eqf08_en.pdf. Retrieced on July 18, 2009.
22. Hill, T. & R. Westbrook (1997). "SWOT Analysis: It's Time for a Product Recall". *Long Range Planning* v.30, n.3, pp. 46–52, 1997.