International Journal of Hybrid Information Technology
Vol.8, No. 5 (2015), pp. 349-356
http://dx.doi.org/10.14257/ijhit.2015.8.5.37

A Survey: Data Warehouse Architecture

[1,2]Muhammad Arif, [1]Ghulam Mujtaba

*[1]Faculty of Computer Science and Information Technology, University of Malaya
50603 Kuala Lumpur, Malaysia
[2]Computer Science Department, Comsats Institute of Information and Technology
Islamabad Pakistan*

Abstract

Data Warehouse and Data mining are technologies that deliver optimallyvaluable information to ease effective decision making. This survey paper defines architecture of traditional data warehouse and ways in which data warehouse techniques are used to support academic decision making. This paper defines different data warehouse types and techniques used in educational environment to extract, transform and load data, and the ways to improve these techniques to have maximum benefit of data warehouse in educational environment. Further this paper have define different data warehouse framework for different situations.

Keywords: *Data Warehouse (DWH), Online Analytical Processing (OLAP), Decision Support System (DSS), Dimensional Modeling, Data Extraction Transformation and Loading (ETL), Academic Decision Support System, Data marts, Educational Data mining*

1. Introduction

A data warehouse is a kind of management technique that collect business data from different stations of the enterprise network, so that it can provide efficient data analysis to decision makers [1]. There are some architectural requirements which would govern development of architecture, some of them are: identifying potential users, defining security requirements, skill requirements *etc.* [2]. A general data warehouse is designed as follow:

a) Select the "*business Process*" to model, if business processes are multiple so "data warehouse model" must be followed, but if "business process" involve a single process so then "data mart" must be made [3].

b) Select the "*grain*" of "business process", where "grain" is the level at which a fact in a fact table is represented [3].

c) Select the "*dimensions*" that are applied to each fact, dimension is usually categorized as time, date, product and geography [3].

d) Select the "*measures*" that will populate each fact, they are generally in a form of numbers [3].

Data warehouses normally adopt three-tier architecture:

a) **Bottom tier**: It is a "*warehouse database server*", which is almost "relational database system". Data from different sources are mined through gateways. A gateway is maintained by underlying DBMS [3].

b) **Middle tier**: It is an "*OLAP server*", which is applied by using "relational OLAP model" [3].

c) **Top tier**: It is a "*Client*", which consist of tools for querying, reporting, analysis and data mining [3].

Data warehouse architecture is divided into two (2) portions / parts:

1.1. The Back Room: It is composed of:

i. Operational Source System: At this stage data is gathered from different sources, and passed onto data staging area [4].

ii. Data Staging Area: Here few operations are performed on data collected from various sources, operations are:

Extract: Extract data from different sources.

Transform: Transform data into useful information; perform data cleaning / cleansing and data aggregation.

Load: Load the information on Data Warehouse / Data Marts [4].

iii. Data Representation Area: It consists of “Data Warehouse and Data Marts”. These are used to store historical data in a bulk, which are later use for decision making [4].

Data from here is passed to OLAP server, which apply techniques like MOLAP and ROLAP to increase performance [4].

1.2. The Front Room: It consist:

i. Data Access Tools: It consists of tools to access data like:

Analysis / OLAP
Query / Reporting
Data Mining [4]

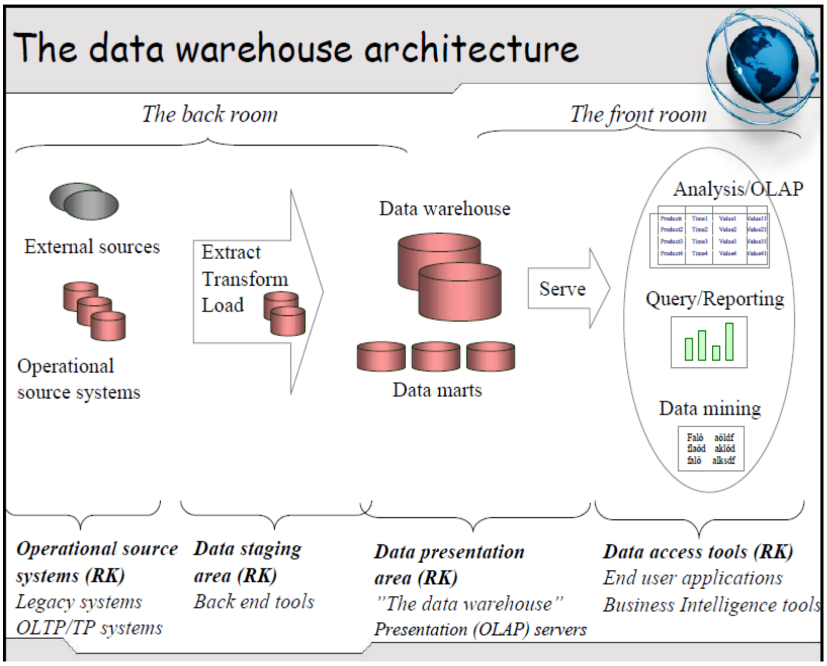


Figure 1. The Data Warehouse Architecture

2. Literature Review

Education related decision support system was required which should gather the data (like financial data and student data from different schools/universities) and on the basis of that data performs decision making(like is the annual number of students increase?, annual number of teachers increase or not ?, what kind of subjects students are preferring etc.).The system has setup a theme based,unified and integrated “data warehouse (DWH)”, by using different tools and technique that facilitates the user with data at various levels and from various views, on the basis of which decision can be made easily [5].

This paper discusses that how “education related decision support system” uses “data warehouse” techniques and OLAP to make decision. Data from all schools/universities are on a data warehouse. Designing DWH and management of metadata they have used ware house studio and for storing data they have used Sybase IQ [6], [10]. Data reporting is done by filling online forms or filling specific formatted files such as excel or text files .Theme oriented DWH is designed, where theme consist of basic facts from all schools and universities [7]. Focusing on theme model chosen is star model. Sybase warehouse is chosen as a logic model where scripts are created automatically [8].

To extract data for DWH; “educational decision-making support” deposit its data to terminal “Data Base", through Data reporting .Data is organized differently. Data must be selected, reflected and transformed.

After these operations data is moved to target [9]. Then they used “warehouse control center” tool to create meta data ,which helps the administrator and developer to find data of their interest [10].After gathering interest information they used different OLAP tools to provide different dimension of data [7] and based on these different dimensions, decisions can be made.

Data Warehouse (DWH) and Data Mining (DM)” is not only deliver critical and useful information but also gaining popularity in industries. In the light of continuously growing demand, schools are trying to prepare students with this technology. This paper describes the key components that comprise a course which would introduce both DWH and DM technologies to a graduate program of information technology [11].

Although it has been more than three decades that DWH and DM technologies are recognized but implementation began early nineties [12]. According to one of the most popular IT jobs search website [13]most of the jobs in computer science are in DWH and DM. so they have started teaching DWH and DM to graduate program of IT ,with the following 7 learning outcomes[11].

1. **Define and describe DWH and its application:**

   “DWH is subject oriented, integrated, non-volatile and time variant collection of data in support of management decision”[14] and have various application like banks/financial, education, hospitals etc.

2. **Distinguish between OLTP and OLAP:**

Online Transactional Process (OLTP) is day to day operation of an organization primary business. Such as ATM of banks, flight ticket of travel agency etc. Whereas Online Analytical Processing (OLAP) is capable of handling huge amount of integrated data to process ad-hoc queries, e.g. which books customers likely to buy together? [15]

3. **Multidimensional views and pivot tables:**

OLAP provides multidimensional views of data on the basis of time/year, geography/location product. And to exercise these views it use Microsoft Excel, Access or Analysis services of SQL server [16].

4. **Dimensional Modeling:**

It is a new data modeling technique that uses fact tables and dimension tables to form star schema.

5. **ETL process:**

It stands for Extract Transform and Load. Data is extracted from different sources, then it is aggregated and finally loaded into fact tables and dimension [18].

6. **Querying a multidimensional cube using MDX:**

It tells how MDX can be used to create OLAP query [18].

7. **Data Mining:**

It is used to discover knowledge from large Data Base [17].

This paper addresses the issue related to modeling and implementing a DWH for “Higher Education Information System (HEIS) in Croatia”. Its purpose was to deliver “data querying service” that will help in understanding, planning and operational work of HEIS. The solution provided by designing a specific DWH which interact with transactional system and copy all the data to relational DB Replica. Data integrity rules are also verified [19].

In this paper DWH for HEIS is proposed [19] this system provides quick and easy to use, internet based system for querying and analyzing data, to support "operational work and decision making". Its main advantage is speed and ability to execute some queries that cannot be done in a single query contrary to the transactional DB [19]. Its architecture is composed of transactional system and HEIS DWH. Relational DB (RDB) replica is saved on "data warehouse server machine", into which a subset of relational DB of HEIS is copied. Then data from RDB replica is "extracted, transformed and loaded" into the "copy of multidimensional DB (MDB)", where "data integrity" rules are checked, if it is correct, then data from "copy of MDB" is loaded in MDB, where MDB is storage mode for MOLAP. If rules are failed, so data is not inserted into MDB and administrator is informed to check faults. "MOLAP server" refreshes data from MDDB and user can query "MDDB or MOLAP" through browser [20]. To protect data, communication is done on a safe channel through Secure Hyper Text Transfer Protocol (HTTPs). Data is presented through internet [21]. There are 3 categories of queries that a user can follow.

1. "Pre-defined queries".
2. "Detailed ad-hoc queries"
3. "Summary ad-hoc queries" [22].

Decision making academia planning requires a lot of analysis data from various systems. This paper provides a method for accessing educational capabilities and preparing its utilization, implemented as a "decision support system" permitting simulation and assessment of various applications and situations. Solution is provided by modeling a "supply –demand relationship" between teaching resources and students [23].

This paper offers a trustworthy "decision support" to the process of balancing educational demand and its supply in universities [23]. In this scenario major components are, teaching resources and students, so a "supply demand relationship" is established between them [24]. Methodology for assessment is based on matrix (which will map errors –faculty dependencies into a university wide circular contribution matrix) [25]. Online model is constructed which will support multiuser "Decision Support System (DSS)" [26].System integrates data from different universities, where DSS help as a reporting tool for solving specific tasks by permitting users to query the data ,produce report and visualize that report to given insight [27]. It is implemented as multi-layered "client-server architecture"". Where calculations are performed at server side using PHP and clients can access it by using web browser and network connection. In this approach major challenge is "pre-processing phase", in which whole data is recognized, collected and combined into DWH.

In this paper the architecture of academia Data Warehouse (DWH) is presented which provides a centralize source for accessing information from different academia units, to quickly analyze the problem and provide solutions, supply data to develop instructions strategic plan and enabling administration to make better business decision based on historical data [28].

This paper presents architecture and design of academia DWH supporting decisional and logical tasks related to three major components i.e. didactics, research and management [28]. "A business intelligence system" in university contest has extensive information about the performance on candidate, teaching staff and didactics. This university DWH fulfils the subsequent requirements.

a. A unique system of analysis and reporting for the managerial staff of the athenaeum and for a single organizational and administrative structure such as departments or secretariats for the students.

b. A system that provides in real time data to "information external agencies" [28].

Data is extracted from six (6) different sources (ESSE3, NOGE, CIA, CSA, SAPERI, and SINBAD), then transformed and loaded into academic DWH. Where data is organized into four (4) data marts .i.e. Didactics, Finance, Research, and Human Resources), from where data is accessed by MIUR, CRUI, academic supervisory staff , organizational structures and Administrative structures [28].

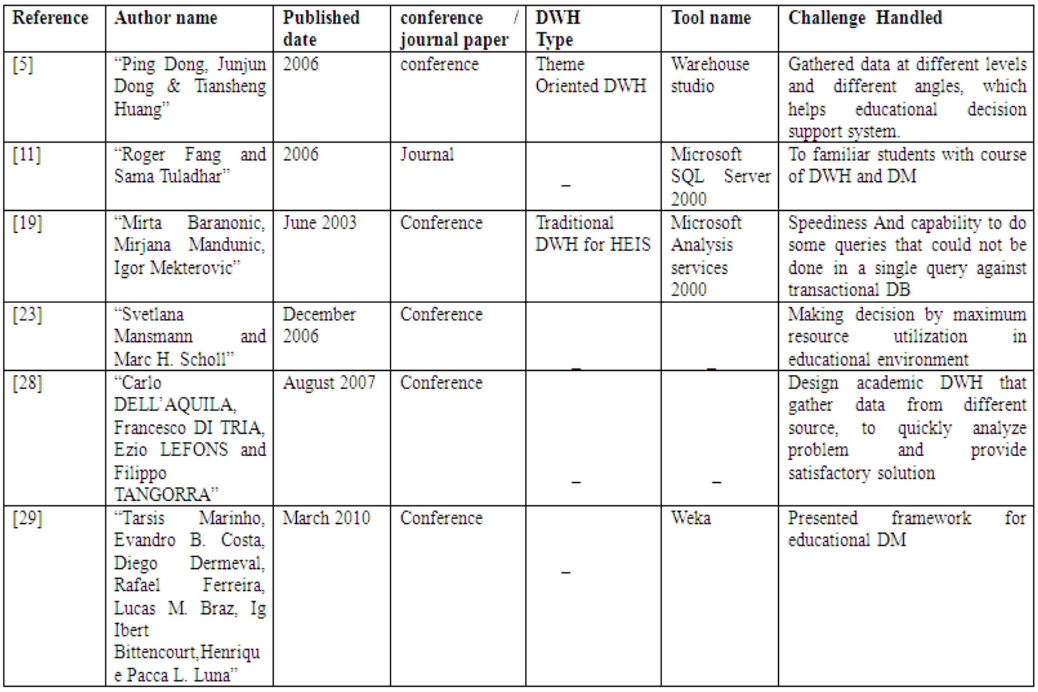
This paper present an "ontology based software framework" for providing "educational Data Mining (DM)" applications. It primarily offers elasticity in encapsulating mining techniques with semantic web services. It consists of architecture based on four layers. This framework provides benefits to both developer and teacher. [29]

This paper provides software framework for educational DM based on semantic web service and will support both developer and teacher [29]. It used SEDAM framework which offer flexibility on encapsulating mining techniques with semantic web services. Its architecture consists of four (4) layers i.e. Tools, web services, service manager and Ontology. It chooses a "tool" WEKA[30], which consists of algorithms for common DM problems like regression, classification etc. Then comes "web services" which provide generalization of necessities applied in DM tools. Through this layer educational environment can use algorithm without implementing them, which is less complex and saves time [31]. The next layer then maps web services available in subordinate layer. Providing meaning to it, this task is achieved in three (3) phases:

a. "Data preprocessing"
b. "Conversion to ARFF format"
c. "Execution of association algorithm" [32]

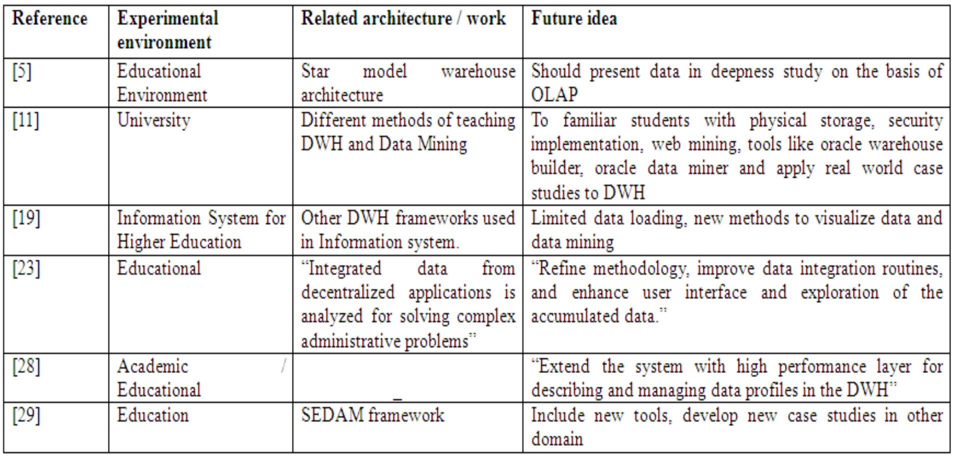
Ontology provides a formal knowledge representation. In this frame work some ontology classes are: i) Data Mining, ii) Data Mining Tasks, iii) Data Mining Techniques, iv) Parameters [30]. Further it uses "Fra W tutoring system" for guidance [30].

Table 1. Explains the Author Name, DWH Type, Tool Name, and Changed Handled by the Data Warehouse Architecture



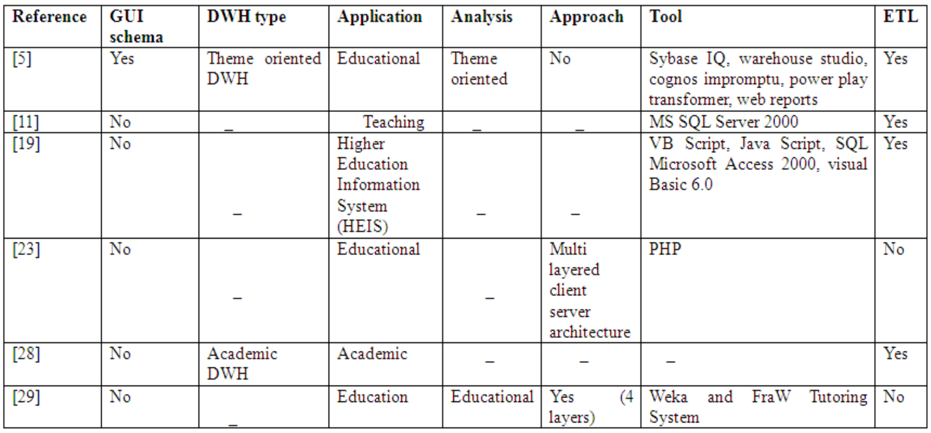
Reference	Author name	Published date	conference / journal paper	DWH Type	Tool name	Challenge Handled
[5]	"Ping Dong, Junjun Dong & Tiansheng Huang"	2006	conference	Theme Oriented DWH	Warehouse studio	Gathered data at different levels and different angles, which helps educational decision support system.
[11]	"Roger Fang and Sama Tuladhar"	2006	Journal	–	Microsoft SQL Server 2000	To familiar students with course of DWH and DM
[19]	"Mirta Baranonic, Mirjana Mandunic, Igor Mekterovic"	June 2003	Conference	Traditional DWH for HEIS	Microsoft Analysis services 2000	Speediness And capability to do some queries that could not be done in a single query against transactional DB
[23]	"Svetlana Mansmann and Marc H. Scholl"	December 2006	Conference	–	–	Making decision by maximum resource utilization in educational environment
[28]	"Carlo DELL'AQUILA, Francesco DI TRIA, Ezio LEFONS and Filippo TANGORRA"	August 2007	Conference	–	–	Design academic DWH that gather data from different source, to quickly analyze problem and provide satisfactory solution
[29]	"Tarsis Marinho, Evandro B. Costa, Diego Dermeval, Rafael Ferreira, Lucas M. Braz, Ig Ibert Bittencourt,Henrique Pacca L. Luna"	March 2010	Conference	–	Weka	Presented framework for educational DM

Table 2. Explain the Experimental Analysis, Related Architecture and Future Idea of the Data Warehouse Architecture



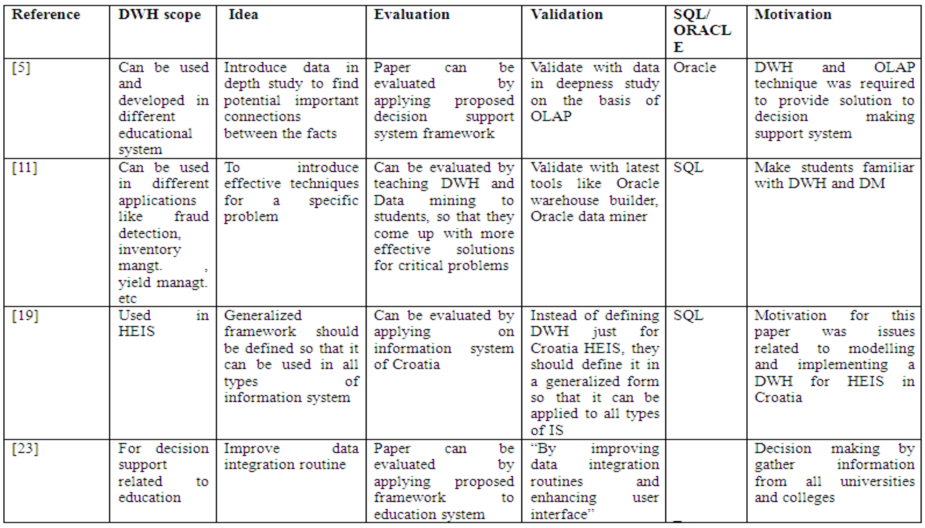
Reference	Experimental environment	Related architecture / work	Future idea
[5]	Educational Environment	Star model warehouse architecture	Should present data in deepness study on the basis of OLAP
[11]	University	Different methods of teaching DWH and Data Mining	To familiar students with physical storage, security implementation, web mining, tools like oracle warehouse builder, oracle data miner and apply real world case studies to DWH
[19]	Information System for Higher Education	Other DWH frameworks used in Information system.	Limited data loading, new methods to visualize data and data mining
[23]	Educational	"Integrated data from decentralized applications is analyzed for solving complex administrative problems"	"Refine methodology, improve data integration routines, and enhance user interface and exploration of the accumulated data."
[28]	Academic / Educational	_	"Extend the system with high performance layer for describing and managing data profiles in the DWH"
[29]	Education	SEDAM framework	Include new tools, develop new case studies in other domain

Table 3. Explain the DWH type, Application Analysis, Approach and ETL Process of the Data Warehouse Architecture



Reference	GUI schema	DWH type	Application	Analysis	Approach	Tool	ETL
[5]	Yes	Theme oriented DWH	Educational	Theme oriented	No	Sybase IQ, warehouse studio, cognos impromptu, power play transformer, web reports	Yes
[11]	No	_	Teaching	_	_	MS SQL Server 2000	Yes
[19]	No	_	Higher Education Information System (HEIS)	_	_	VB Script, Java Script, SQL Microsoft Access 2000, visual Basic 6.0	Yes
[23]	No	_	Educational	_	Multi layered client server architecture	PHP	No
[28]	No	Academic DWH	Academic	_	_	_	Yes
[29]	No	_	Education	Educational	Yes (4 layers)	Weka and FraW Tutoring System	No

Table 4. Explain the Scope, Idea, Evaluation, Validation and Motivation of the Data Warehouse Architecture



Reference	DWH scope	Idea	Evaluation	Validation	SQL/ ORACLE	Motivation
[5]	Can be used and developed in different educational system	Introduce data in depth study to find potential important connections between the facts	Paper can be evaluated by applying proposed decision support system framework	Validate with data in deepness study on the basis of OLAP	Oracle	DWH and OLAP technique was required to provide solution to decision making support system
[11]	Can be used in different applications like fraud detection, inventory mangt. , yield managt. etc	To introduce effective techniques for a specific problem	Can be evaluated by teaching DWH and Data mining to students, so that they come up with more effective solutions for critical problems	Validate with latest tools like Oracle warehouse builder, Oracle data miner	SQL	Make students familiar with DWH and DM
[19]	Used in HEIS	Generalized framework should be defined so that it can be used in all types of information system	Can be evaluated by applying on information system of Croatia	Instead of defining DWH just for Croatia HEIS, they should define it in a generalized form so that it can be applied to all types of IS	SQL	Motivation for this paper was issues related to modelling and implementing a DWH for HEIS in Croatia
[23]	For decision support related to education	Improve data integration routine	Paper can be evaluated by applying proposed framework to education system	"By improving data integration routines and enhancing user interface"	_	Decision making by gather information from all universities and colleges

3. Conclusion

This paper proposed solution for DWH, how data warehouse techniques can be used to have maximum benefit in academic environment.This survey also included the sort of data warehouse is used in some specific environment. Different tools and techniques used to build data warehouse. Technique works better in different situations and environment is listed. Data extraction, transformation and loading on a data warehouse server, improving efficiency and performance of data warehouse, limitations of a specific technique and how those limitations can be overcome. The future work of Data warehouse architecture includes, datashould beintroduced in depth study on the basis of OLAP. To familiarize students with physical storage, security implementation, web mining, tools like oracle warehouse builder, oracle data miner and apply real world case studies to DWH.It also includes Partial data loading, new ways to data visualization and data mining. Furthermore, it includes refine methodology, improves data integration routines, and enhances user interfaces and exploration of the accumulated data.Moreover, future work includes, extend the system with high performance layer for describing and managing data profiles in the DWH, new tools and develop new case studies in other domain.

References

[1] J. A. Senn, **(2000)**, “Information technology in business: principles, practices, and opportunities”, Prentice-Hall, Inc..

[2] [Online] http://edw.berkeley.edu/documents/EDW%20architecture%20package.pdf, **(2013)** April 9.

[3] [Online] http://dataminingzone.weebly.com/uploads/6/5/9/4/6594749/ch4_dw_architecture.pdf, **(2013)** April 9.

[4] [Online] http://people.dsv.su.se/~petia/is5/Lectures/F4.pdf, **(2013)** April 9.

[5] P. Dong, J. Dong and T. Huang, “Application of Data Warehouse Technique in Educational Decision Support System”, In 2006 IEEE International Conference on Service Operations and Logistics, and Informatics, **(2006)**, pp. 818-822.

[6] B. Bebel, Z. Królikowski and R. Wrembel, “On Methods’ Materialization in Object-Relational Data Warehouse”, In Advances in Information Systems, **(2002)**, pp. 425-434, Springer Berlin Heidelberg.

[7] T. Morzy, R. Wrembel and T. Koszlajda, “Hierarchical materialisation of method results in object-oriented views”, In Current Issues in Databases and Information Systems, **(2000)**, pp. 200-214, Springer Berlin Heidelberg.

[8] L. Burmester and M. Goeken, “Method for User Oriented Modelling of Data Warehouse Systems”, In ICEIS, vol. 3, **(2006)**, pp. 366-374.

[9] P. Vassiliadis, C. Quix, Y. Vassiliou and M. Jarke, “Data warehouse process management”, Information Systems, vol. 26, no. 3, **(2001)**, pp. 205-236.

[10] R. Wrembel and B. Bębel, “Metadata management in a multiversion data warehouse”, In Journal on data semantics VIII, **(2007)**, pp. 118-157, Springer Berlin Heidelberg.

[11] R. Fang and S. Tuladhar, “Teaching data warehousing and data mining in a graduate program of information technology”, Journal of Computing Sciences in Colleges, vol. 21, no. 5, **(2006)**, pp. 137-144.

[12] B. Bębel, J. Eder, C. Koncilia, T. Morzy, R. Wrembel, , “Creation and management of versions in multiversion data warehouse. InProceedings of the 2004 ACM symposium on Applied computing, **(2004)** March, pp. 717-723, ACM.

[13] www.computerjobs.com.

[14] E. Levine, “Building a data warehouse. American School Board Journal”, vol. 189, no. 11, **(2002)**, pp. 48-50.

[15] G. M. Marakas, “Modern data warehousing, mining”, and visualization: core concepts, **(2003)** pp. 100-101), Prentice Hall.

[16] J. C. Collins, “Microsoft Excel Pivot Tables”, **(2005)** June 3. http://www.microsoft.com/businesssolutions/excel_pivot_tables_collins.mspx

[17] J. MacLennan, Z. Tang and B. Crivat, “Data mining with Microsoft SQL server 2008”, John Wiley & Sons, **(2011)**.

[18] R. Kimball and M. Ross, “The data warehouse toolkit: the complete guide to dimensional modeling”, John Wiley & Sons, **(2011)**.

[19] M. Baranovic, M. Madunic and I. Mekterovic, “Data warehouse as a part of the higher education information system in Croatia”, In Information Technology Interfaces, ITI 2003. Proceedings of the 25th International Conference on, **(2003)** June, pp. 121-126, IEEE.

[20] R. G. Allan and D. R. May, “Data models for a registrar's data mart”, Journal of Data Warehousing, vol. 6, no. 3, **(2001)**, pp. 38-53.

[21] H. G. Molina, W. J. Labio, J. L. Wiener and Y. Zhuge “Distributed and Parallel Computing Issues in Data Warehousing (Invited Talk)”, In Proc. of the Tenth Annual ACM Symposium on Parallel Algorithms and Architectures, **(1998)**.

[22] J. W. Seifert, “Data mining: An overview”, National security issues, **(2004)**, pp. 201-217.

[23] S. Mansmann and M. H. Schol, “Decision support system for managing educational capacity utilization”, Education, IEEE Transactions on, vol. 50, no. 2, **(2007)**, pp. 143-150.

[24] D. Z. Deniz and I. Ersan, “Using an academic DSS for student, course and program assessment”, In Proceedings of the ICEE 2001 Conference, **(2001)** August.

[25] C. A. Casper and M. S. Henry, “Developing performance -oriented models for university resource allocation”, Research in Higher Education, vol. 42, no. 3, **(2001)**, pp. 353-376.

[26] N. K. Kwak and C. Lee, “A multicriteria decision-making approach to university resource allocations and information infrastructure planning”, European Journal of Operational Research, vol. 110, no. 2, **(1998)**, pp. 234-242.

[27] S. Vinnik and M. H. Scholl, “UNICAP: Efficient decision support for academic resource and capacity management” **(2005)**, pp. 235-246, Springer Berlin Heidelberg.

[28] C. Dell’Aquila, F. Di Tria, E. Lefons and F. Tangorra, , “An academic data warehouse”, In Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications, **(2007)** August, pp. 229-235.

[29] T. Marinho, E. B. Costa, D. Dermeval, R. Ferreira, B. L. Braz, I. I. Bittencourt and H. P. L. Luna, , “An ontology-based software framework to provide educational data mining”, In Proceedings of the 2010 ACM Symposium on Applied Computing, **(2010)** March, pp. 1433-1437, ACM.

[30] I. H. Witten and E. Frank, “Data Mining: Practical machine learning tools and techniques”, Morgan Kaufmann, **(2005)**.

[31] O. R. Zaïane and J. Luo, “Web usage mining for a better web-based learning environment”, In Proceedings of conference on advanced technology for education, **(2001)** June pp. 60-64.

[32] A. Preece and S. Decker, **(2002)**, “Guest Editors' Introduction: Intelligent Web Services”, IEEE Intelligent Systems, vol. 17, no. 1, **(2002)**, pp. 15-17.

Author

Muhammad Arif is a PhD student at Faculty of CS and IT, University of Malaya. Currently he is working on Medical image Processing. His research interests include image processing, E learning, Artificial intelligence and datamining. He joined UM as a Bright Spark Scholar in September 2013 for the period of 3 years. Before this he completed masters and bachelor degrees in Pakistan. He received his BS degree in Computer Science from University of Sargodha, Pakistan in 2011. He obtained his MS degree in Computer Science from COMSATS Islamabad 2013 Pakistan.