

A survey of data warehouse architectures - preliminary results

Moh'd Alsour
Wrocław University of Economics
ul. Komandorska 118/120,
53-345 Wrocław, Poland
Email: mohsqour@wp.pl

Kamal Matouk
Wrocław University of Economics
ul. Komandorska 118/120,
53-345 Wrocław, Poland
Email: kamal.matouk@ue.wroc.pl

Mieczysław L. Owoc
Wrocław University of Economics
ul. Komandorska 118/120,
53-345 Wrocław, Poland
Email: mieczyslaw.owoc@ue.wroc.pl

Abstract— In this abridged version, which is a summary of the authors' study, we present some of our findings for consideration. The principal objective of the study was to investigate empirically the architectures of data warehouse (DW), or more specifically, the types of the architectures and a number of factors, which are believed to influence their selection, were explored. A questionnaire survey, which targeted the information systems managers, was used to collect data from 150 Polish companies about the respondents' firms, the architecture they use, and the factors which influence the selection of the architecture. The findings of the study give us practical insights into DW's field in Poland.

I. INTRODUCTION

The early 1970s ushered in decision support systems (DSS), which were fundamentally different from operational or transactional systems [5]. A DSS requires the construction of a DW in order to complete its life cycle. A DW unifies the data scattered throughout an organization into a single centralized data structure with a common format [24]. For the data component, it was recognized that a separate data repository was needed to draw data from operational systems and other data sources and therefore independent data marts were developed in response to that need as the first decision support data infrastructure [5].

Data warehousing began in the 1980s as a response to the lack of information provided by the many online application systems that were being built, online applications served the needs of a limited community of users, and they were rarely integrated with each other [22]. Additionally, online applications had no appreciable amount of historical data because they jettisoned their historical data as quickly as possible in the name of high performance. Thus, corporations had lots of data and very little information [22]. Inmon claims that a DW was the first attempt at architecture that most organizations had ever encountered." Prior to data warehousing, everything had been a new application; however, it became apparent that applications were not going to get the organization where it needed to go over time. The solution was to build an architecture or at least the first fledgling steps of an architecture". Inmon argues that there is still a great deal of confusion as to what a data warehouse really is." Bill Inmon [22], [21p.31], the world-renowned expert, said that the definition for a DW was and still is today. "A source of data that is subject-oriented, integrated,

and time-variant for the purpose of management's decision processes". Inmon, who coined the term DW, said that the underlying architecture for a DW has evolved over the years, although the original definition remains the same.

Because of the need for further investigation on this topic, considering the importance of the architecture's choice and the shortage of the empirical research [3], we have every reason to investigate the success of the various architectures. In this paper, we discuss different architectures of DW and analyze their structures and features, along with the factors which influence their choice. Its sections as follows; section 2 provides general description and characters of the different architectures. In section 3, we review the available literature which pertains to the study. In section 4, we discuss the research methodology. Section 5 provides a brief account of the results, and the paper's core argument. Finally, section 6 contains the real meat of our argument and abstracts the main points from it.

II. TYPES OF DATA WAREHOUSE ARCHITECTURES

The review of the available literature on DW has identified five predominantly architectures, i.e. independent data marts, bus architecture, hub and spoke, centralized and federated [1], [2], [3], [4], [5], [6p.56], [7], [8], [9]. Although other architectures (e.g. hybrid) are mentioned in the literature, they tend to be variations on these five [2]-[4], [9]. These architectures are grouped together in Fig.1.

Fig. 1 shows the principal architectures, independent data marts (IDM), which are the first endeavors to provide a repository of decision support data, are typically independent of other data stores and serve specific and localized needs such as providing data for a particular application or business unit, do not provide "a single version of the truth" [2]-[5]. However, IDM is not a formally advocated architecture in the industry. Nevertheless, data marts exist and are used in organizations as a DW solution. They are also commonly considered in discussions and surveys of the various data warehouse architectures [5]. For the purpose of this paper and in line with the literature (e.g.,[5]), the data marts are treated as an architecture.

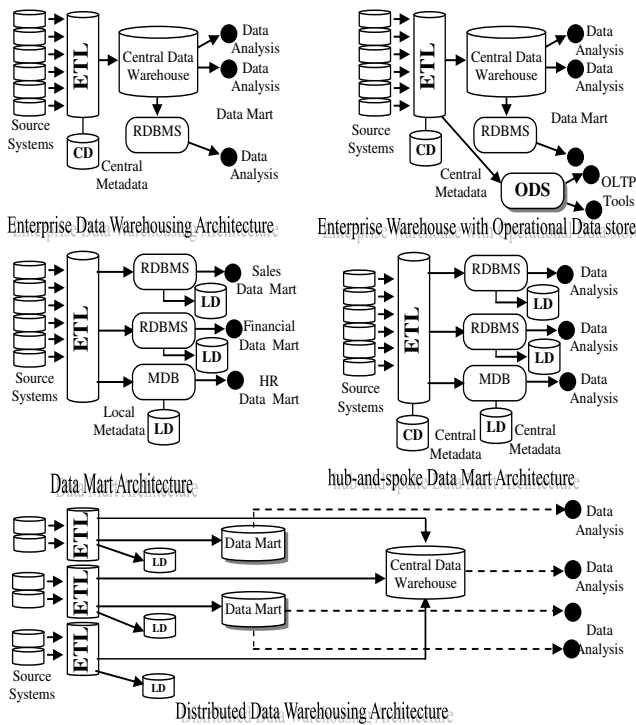


Fig.1. The common architectures of data warehouse [1]

Data mart bus architecture with linked dimensional data marts (DBA) has data marts that support various business processes, the first mart is built for a single business process using dimensions and measures that are used with other marts (i.e., conformed dimensions), additional marts are developed using these conformed dimensions, which results in logically integrated marts and an enterprise view of the data [2]-[5], [7]. Hub and Spoke Architecture is developed in an iterative manner, subject area by subject area. In this architecture, atomic level data is maintained in the warehouse in 3rd normal form [7]. Dependent data marts are created that source data from the warehouse, thus maintaining a “single version of the truth” The dependent data marts may be developed for departmental, functional area, or specialized purposes (e.g., data mining) and may have normalized, denormalized, or summarized dimensional data structures depending on user needs [2] , [4], [6p.59]. Fig. 1 clearly shows the similarity between the centralized DW architecture and the hub and spoke except that there are no dependent data marts, contains atomic level data, some summarized data, and logical dimensional views of the data[2]-[5], [6p.60],[7]. FED is recommended when there is a fragmented decision support data environment and there is a need to integrate at least some of the data. These data are either logically or physically integrated using shared keys, global metadata, distributed queries, or other methods [2]-[5], [6 p.60].

Because of the companies’ need to an enterprise-wide data repository to support a variety of analytical applications (e.g. queries, OLAP, and data mining), the DWs began to emerge and resulted in two competing architectures, i.e. enterprise

DW and data mart bus architecture [5], [13p.229], [9], [6p.55], [16p.53], [17], [19]. According to Ariyachandra and Watson [3]-[5], there has been much debate on the issue of DW architecture. Bill Inmon and Ralph Kimball, who are pioneers in the field of DW, are among the debaters. Inmon argues in favour of a hub-and-spoke architecture (e.g., the Corporate Information Factory), whereas the data mart bus architecture with conformed dimensions is approved by Kimball. Inmon supports a top-down development approach that adapts traditional relational database tools to the development needs of an EDW. From this enterprise wide data store, individual departmental databases are developed to serve most decision support needs [17].

Table 1 summarizes the most essential characteristic differences between the two ideologies.

TABLE I.

COMPARISON OF ESSENTIAL FEATURES OF INMON’S AND KIMBAL’S IDEOLOGY [6 P.56]

	Kimball’s Approach	Inmon’s Approach
1.	Everyone is allowed to fabricate their database according to their requirements and department structure. All these independent repositories can be integrated as and when required. This methodology is known as bottom up approach.	Supports a top down approach. Here no one is allowed to develop any database independently. The database for an organization should be planned and designed centrally. Every department within the organization will follow the centrally designed schema to fabricate their database.
2.	This structure is easier to build.	The structure proposed is very typical one to craft.
3.	It is a nimble approach.	Rigorous analysis and designing is required.
4.	Problematic to maintain as an enterprise resource.	Easier to maintain as an enterprise resource.
5.	Data is often redundant.	Redundancy is regulated to a great extent.
6.	Very difficult to integrate independent data marts with varying structure.	Integration of data marts is comparatively easier.
7.	This approach is flexible.	This approach is comparatively rigid.

The second approach is that of Ralph Kimball, who proposes a bottom-up approach that employs dimensional modeling [8], [17]. A data modeling approach unique to data warehousing. Rather than building a single enterprise wide database, Kimball suggests creating one database (or data mart) per major business process. Enterprise-wide cohesion is accomplished by using another Kimball innovation, a data bus standard [17]. The EDW approach does not preclude the creation of data marts. The EDW is the ideal in this approach because it provides a consistent and comprehensive view of the enterprise [13p.229]. Kimball’s data mart strategy is a “plan big, build small” approach. A data mart is a subject-oriented DW. It is a scaled-down version of a DW that focuses on the requests of a specific department, such as

marketing or sales. This model applies dimensional data modeling, which starts with tables. Kimball advocated a development methodology that entails a bottom-up approach. Which in the case of DWs means building one data mart at a time [13p.230]. According to McKenna [18], a key decision in any data warehouse implementation is what overall architectural style to employ for the warehouse. "Should we follow Bill Inmon's conventions to build a CIF, or is Ralph Kimball's bus architecture more appropriate?. Should an operational data store (ODS) be utilized for the near-real-time data?" McKenna claims that the right answer depends on many factors, including the current state of DW development, organizational readiness, and the scope of the project being undertaken.

According to Watson [23], not long ago, BI managers and professionals were struggling with architecture decisions. "Is the Inmon hub-and-spoke or the Kimball data mart bus architecture best? (Both can be successful) Should we build logical or physical data marts? (You will end up with at least a few physical ones)." Watson argues that these decisions do not seem challenging today, they were for the people developing their first DW at a time when data warehousing knowledge was less codified.

Conradie 2005 investigated and compared the views of Inmon and Kimball and found that the concept of the CIF is appealing [16 p.112]. Although he fully supported the concept of a CIF, the Kimball approach to the design of the data warehouse (simple data mart by data mart, driven by specific business needs and glued together by the Bus architecture of conformed dimensions), led him to lean towards the Kimball approach when developing the Bigger Picture BI context Model in his thesis. he claimed that the idea to accommodate the detailed transactional data requirements in a detailed data mart as part of the data warehouse (instead of a separate ODS), is a further plus point for the Kimball's approach [16 p.76].

III. LITERATURE REVIEW

Watson and Ariyachandra 2005 [2] conducted a Web-based survey of 454 to better understand the factors that influence the selection of a DW architecture and the success of the various architectures. For the study, the authors, in addition to reviewing the data warehousing literature, formed a group of 20 experts to help identify the architectures to study and the success measures to use. Bill Inmon and Ralph Kimball, who are leading experts in the field of DW, were among the participants. The results showed the predominance of the hub-and-spoke architecture (39%) followed by the bus architecture (26%), centralized (17%), IDM (12%), and federated (4%) [2]-[4]. The researchers found that the initial development time and costs for independent data marts, the bus architecture, and the centralized architecture are very similar to one another. When the bus architecture was compared to the hub and spoke/centralized architectures, no statistically significant differences were found, indicating that these architectures

are equally successful. Perhaps the most interesting study finding is how similar the bus, hub and spoke, and centralized architectures score on many of the metrics. It helps explain why these competing architectures have survived over time. The data reveals that all of the selection factors have some influence. The lowest average score is over 4.3 (for the perceived ability of the in-house IT staff), thus indicating that even the lowest rated factor is important. The most important factors (with average scores over 5.0) are information interdependence between organizational units, the strategic view of the warehouse prior to implementation, and upper management's information needs.

In 2010, Ariyachandra and Watson [5] conducted another study in order to investigate the factors which influence the selection of a particular DW architecture, or more specifically, the research questions, "What factors are most important to the architecture selection decision?" and "What factors influence the selection of a particular architecture?". The results, which are based on responses from 400 organizations and interviews with experts, suggested that various combinations of organizational factors influence data warehouse architecture selection. The overall findings suggest that different combinations of specific factors affect the likelihood of selecting one architecture over the others. Three selection factors emerged as being important in choosing between the IDM and EDW architectures. They are the perceived of the IT staff, resource constraints, and the strategic view of the warehouse. In addition, further analysis revealed that interdependence, task routine, and the level of sponsorship influence the selection of an EDW through strategic view. When interdependence is high, task routine is low, and the sponsorship level is high, organizations perceive the ability implementation of a data warehouse as a strategic initiative. The distribution of the four data warehouse architectures was: IDM 11%, data mart bus architecture (DBA) 26%, EDW 58%, and Federated architecture 5%.

A survey of 213 practitioners at the Data Warehousing Institute San Diego Conference in August 2004 by Forrester, showed 42.25% of those questioned were in favour of centralized with hub-and-spokes architecture followed by IDM without consistent design, the centralized architecture, federated and virtual data warehousing, with 15.49%, 18.78%, 17.84 % and 0.94% respectively [10]. Lou Agosta, the lead industry analyst at Forrester Research, Inc, said it was the type of the firm that influence the choice of the architecture." In the real world, firms that are highly centralized in geography and governance should pursue a centralized data warehouse architecture to reap the greatest operational efficiencies and business benefits. In contrast, those firms that are highly decentralized will prefer a distributed architecture, and those with a mixed organizational pattern should implement a federated one." Agosta says no DW architecture is right or wrong in itself." Enterprises have succeeded with all the alternative architectures surveyed - and individual cases of failure have also occurred. The data warehousing architecture will often mirror the form of the enterprise that implements it."

A survey by Mawilmada 2011 revealed that a centralised DW architecture addresses current needs and can also be upgraded to an enterprise wide warehouse architecture or federated DW architecture. "There is lack of support available for the current decision-making process from the current information systems and need a centralised data management (process) to improve decision-making". Analysis of the question regarding current decision-making issues, revealed most of the end users (75%) have selected integration of data from other data repositories as the main problem for current decision-making [8p.49]. The centralised data warehouse model improves the access to data integrated from the different units compared with the architecture of independent data marts [8p.79]. Research suggests that implementing centralised data warehouse will minimise the issues faced in the current decision-making process and also, provide many benefits such as improved access to data and improved decision-making [8p.85].

William McKnight, who architected and directed the development of several of the largest and most successful business intelligence programs in the world, claims that the mighty struggle within the ranks of IT architects and consultancies about the best architecture no longer subsists." We cannot deny the efficacy of having a centralized data store with quality, integrated, accessible, high performance and scalable data. The discussion now revolves around the EDW architecture." McKnight, the founder and president of McKnight Associates, Inc., a leading business intelligence consultancy, argued that EDW is a "big tent" architecture, and the bigger, the better; however, there still must be digestible pieces along the way [11]. Increasingly, large enterprises are recognizing the value of an EDW in their information and knowledge strategies [12]. Srikanth, a senior DW consultant for Teradata, discussed the potential benefits of an EDW. "The potential benefits include cost-effective consolidation of data for a single view of the business and creation of a powerful platform for everything from predictive analysis to near real-time strategic and tactical decision support throughout the organization."

According to Singh and Malhotra 2011 [7], IDM is more likely to be selected if resources are limited and bus architecture when there is a dire need to share data or information between departments. Whereas in Hub and Spoke/Centralised architecture tends to be selected where the data warehouse is considered to be an integral part of a strategic solution hence there is a high need for data to be made freely available between business units. While the need for information to be made freely available between business units is important in the choice of either of these architectures, it is the bus architecture that will tend to be chosen. Similarly, if the data warehouse is required quickly the bus architecture has proved popular. However, as a strategic solution, it is the hub and spoke/centralised architecture that is the more likely choice [7].

Salonen 2004 [14 p.98] compared different architectural models of DW and found clear advantages for the federated DW model with a centralized metadata repository providing common dimensions and measures for both dependent and as

independent data marts. This architectural model could be called the "architecture of all architectures" as it provides a "virtual view" of the whole enterprise. This approach is also based on the "bottom-up" implementation approach, wherein the enterprise-wide analytical solutions are built from dependent data marts with an enterprise view in mind. Another advantage of the federated DW model is the added data staging area, which enables easier implementation of analytical solutions as a staging area hides the complexities of operational data sources. Salonen recommended selection of some of the characteristics of a federated DW model combine with distributed DW architectural model as the best foundation for analytical application software development [14 p.98]. In his analysis of different DW implementation models and DW architectural models, it became clear that selection of a product development strategy requires additional dimensions for analysis, such as market segmentation and technology selections [14 p. 99].

Kimball and Ross 2002 [15p.13] argue that the bus architecture is the secret to building distributed DW systems. "Let's be real-most of us don't have the budget, time, or political power to build a fully centralized DW. When the bus architecture is used as a framework, we can allow the enterprise DW to develop in a decentralized (and far more realistic) way."

The studies by [2]-[5] suggested why there are agreements and disagreements over which architecture is best. "In some ways, the architectures have evolved over time and become more similar. For example, the hub-and spoke architecture often includes dimensional data marts, which is at the heart of the bus architecture. Even the development methodologies (e.g., top down for the hub-and-spoke and centralized architectures, and life cycle or bottom up for the bus architecture) have evolved and become more similar."

According to Sen and Sinha 2005 [1], these different definitions and concepts gave rise to an array of data warehousing methodologies and technologies. "Data warehousing methodologies are rapidly evolving but vary widely because the field of data warehousing is not very mature." They added, no architecture has achieved the status of a widely recognized standard yet. "As the industry matures, there could be a convergence of the methodologies, similar to what happened with database design methodologies. It is apparent that the core vendor-based methodologies are appropriate for those organizations that understand their business issues clearly and can create information models. Otherwise, the organizations should adopt the information modeling based methodologies. If the focus is on the infrastructure of the data warehouse such as metadata or cube design, it is advisable to use the infrastructure-based methodologies. Although the methodologies used by these companies differ in details, they all focus on the techniques of capturing and modeling user requirements in a meaningful way."

There is no one-size-fits-all strategy to data warehousing. An enterprise's data warehousing strategy can evolve from a simple data mart to a complex DW in response to several factors such as user demands [17 p. 230]. For many

enterprises a data mart is frequently a convenient first step to acquiring experience in constructing and managing a DW [17 p. 231].

According to Watson and Ariyachandra [2], organizations are not the same, and consequently, companies may differ on their architecture selection. "There isn't a single architecture that is best for all situations and companies. If it were that simple, there wouldn't be disagreements over architecture selection."

Jennings and Lewis 2011 [20], concluded that "the best solution for one enterprise is not necessarily the best for another. Although best practices and reference architectures should be considered during design, they should not be followed blindly."

A typical architecture model for data warehousing as defined by data warehousing literature is a process model in which data are extracted, transformed and loaded into a DW database [14p.61]. From the software development perspective, the traditional centralized DW model is arguably the easiest architectural model but also the most inefficient environment in end user organizations with heterogeneous data sources and hardware and software environments [14 p. 76].

Based on the literature review, the above mentioned argument, and for the manageability sake, in this study, five architectures and five factors were investigated. The actual survey questions were; what is the architecture of your decision support system (DW)? which factors led your firm to choose this particular architecture? Furthermore, the research objectives include the following items:

1. To identify the various architectures of DW in Polish firms.
2. To study the factors which influence the choice of DW architecture in Polish firms.

IV. RESEARCH METHODOLOGY

The DW implementation and success literature were reviewed to identify the architectures to study, and factors that affect the DW architecture selection decision. the investigated architectures were independent data marts, data mart bus architecture with linked dimensional data marts, hub and-spoke, centralized data warehouse (no dependent data marts), and federated. The following five factors were extracted [2],[5]:

- upper management's information needs,
- constraints on resources,
- technical issues,
- information interdependence between organizational units, and
- source of sponsorship.

After the literature review, survey data were collected from firms listed in Warsaw Stock Exchange - WSE (150) via a mail questionnaire. 60 usable responses were analyzed. A brief description of some findings is shown in the next section.

V. ANALYSIS AND RESULTIS

Fig.2 shows the distribution of the respondents across different sections of the industry, the most represented are financial services 24% followed by manufacturing 18%.

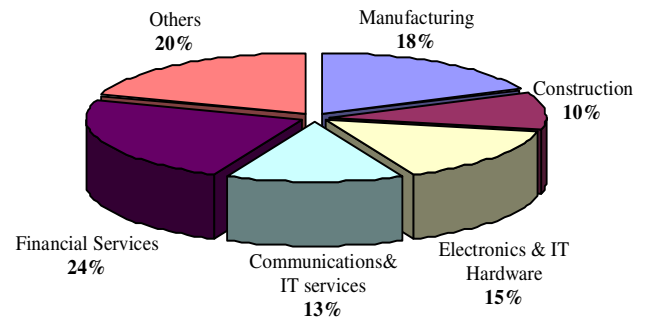


Fig. 2. The Percentage of firms by Industry

In response to the question of which architecture they have in their firms. The results are represented in Fig. 3 below. The leading architecture is hub and spoke 35% followed by data mart bus architecture 23%. EDW was placed third with 20%. 7% of the respondents reported having a federated architecture.

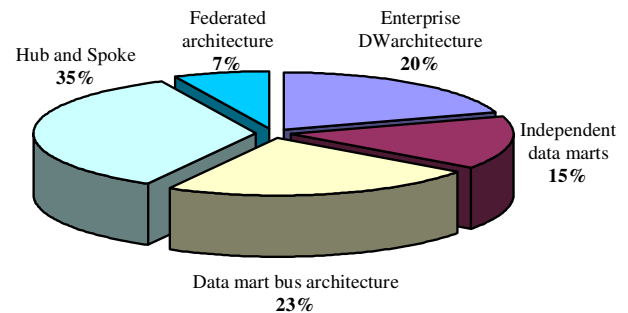


Fig. 3. The Distribution of the Architectures

The respondents asserted the importance of the five factors in choosing the architecture. However, they claimed that upper management's information needs and constraints on resources played an important part in determining their choice of the architecture. The findings of the study conformed to the most of the literature mentioned in section 3.

VI. CONCLUSION

This paper, which is part of a study under sifting and scrutinizing, focuses on the status of DW in Poland in general and the different architectures of DW, and the factors which affect their choices in particular. In this study, we concentrated on the most common architectures specially Inmon and Kimbl's ideologies.

The preliminary findings show the predominance of hub and spoke architecture in the study's sample. 35% of the

respondents use hub and spoke architecture in their firms, 23% use data mart bus architecture which is the second highest choice. The results are represented in Fig. 3 section 5.

At this stage of data analysis, we are not able to determine precisely the limitations, hence the analysis of the study's results has not completed yet. However, the sample's size. There is, in addition, the use of on method of data collection, i.e. a questionnaire. May be considered some of those limitations.

To some extent this study is one of the academic contributions. It contributes to our understanding of the DW's architecture in general and in Poland in particular, and may form a basis and motivation for future research in the important field.

REFERENCES

- [1] A. Sen, A. P. Sinha, "A Comparison of Data Warehousing Methodologies Using a common set of attributes to determine which methodology to use in a particular data warehousing project", *Communications of the ACM*, March 2005, Vol. 48, No. 3 pp. 79-84.
- [2] H. J. Watson., T. Ariyachandra, "Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures", Terry College of Business, University of Georgia, (July 2005) http://www.terry.uga.edu/~hwatson/DW_Architecture_Report.pdf accessed 1.03.2012.
- [3] T. Ariyachandra, H. J. Watson., "Which Data Warehouse Architecture Is Most Successful?", *Business Intelligence Journal* Vol. 11, No. 1 2006 pp. 4-6.
- [4] T. Ariyachandra, H. J. Watson, "Which Data Warehouse Architecture is Best?", *Communications of the ACM* October 2008, Vol. 51 No. 10 PP. 146-147.
- [5] T. Ariyachandra, H. J. Watson, "Key organizational factors in data warehouse architecture selection", *Decision Support Systems* 49 (2010) 200-212.
- [6] J. Singh, promulgation of contemporary testing techniques for effective engineering of data warehouses doctorate thesis, faculty of engineering and technology Punjabi University, India July, 2010.
- [7] S. Singh, S. Malhotra, "Data warehouse and its methods", *Journal of Global Research in Computer Science*, Volume 2, No. 5, May 2011 pp. 113-115.
- [8] Mawilmada, Pubudika Kumari (2011) Impact of a data warehouse model for improved decision-making process in healthcare. Masters by Research thesis, Queensland University of Technology October.
- [9] S. Nilakanta, K. Scheibe., A. Rai., Dimensional issues in agricultural data warehouse designs, *computers and electronics in agriculture* 60, 2008, pp. 263-278.
- [10] L. Agosta, "Hub-and- Spoke Architecture Favored", *DM Review*, Mar 2005, Vol. 15 Issue 3, p14-63.
- [11] W. McKnight, "The New Business Intelligence Architecture Discussion", *DM Review* September 2004 Vol. 14 Issue 9 pp. 96-96.
- [12] S. Srikant, "Logical Data Modeling: A Key to Successful Enterprise Data Warehouse Implementations". *DM Review*, Sep2006, Vol. 16 Issue 9, p.13-16.
- [13] E. Turban, J. E. Aronson, T. Liang., R. Sharda . Decision support and business intelligence systems. Eighth edition, Pearson Education, 2007, Inc.
- [14] P. I. Salonen, "Evaluation of a Product Platform Strategy for Analytical Application Software", Ph.D. Thesis, Helsinki School of Economics 2004.
- [15] R. Kimball, M. Ross, "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling", 2nd Edition, Wiley 2002.
- [16] P. J. Conradie, "An industrial engineering perspective of business intelligence", doctorate thesis, faculty of engineering, built environment and information technology, university of Pretoria 2005.
- [17] M. Breslin., "Data warehousing battle of giants: comparing the basics of the Kimball and Inmon models", *Business Intelligence Journal* 9 (4) (2004).
- [18] J. P. McKenna., "Moving Toward Real-Time Data Warehousing,," *business intelligence Journal* vol. 16, No. 3 pp. 14-19, 2011.
- [19] K. Corral, D. Schuff and R.D. St. Louis, "The impact of alternative diagrams on the accuracy of recall: A comparison of star-schema diagrams and entity-relationship diagrams", *Decision Support Systems* 42 (1) (2006) 450-468.
- [20] C. Jennings., L. Tristan, "The Perils of Over-Complicating Data Warehouses How Avoiding Unnecessary Complexity Can Provide Satisfaction and Savings", *Business Intelligence Journal*, vol. 16, No. (2) (2011), p. 39-43.
- [21] W.H. Inmon, "Building the Data Warehouse", Third Edition, New York: John Wiley & Sons, 2002.
- [22] W.H. Inmon., "DW 2.0 Architecture for the Next Generation of Data Warehousing", *DM Review*, Apr 2006, Vol. 16 Issue 4, p.8-25.
- [23] H. J. Watson, "This Isn't Your Mother's BI Architecture", *Business Intelligence Journal*, 2012, Vol. 17 Issue 1, p4-6.
- [24] T. R. Sahama, P. R. Croll, "A Data Warehouse Architecture for Clinical Data Warehousing", in Roddick, J. F. and Warren, J. R., Eds. *Proceedings Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007) CRPIT*, 68, pages pp. 227-232, Ballarat, Victoria.