ISSN 2029-9338 (ONLINE)
ISSN 1822-6515 (CD-ROM)

FASTER AND BETTER DECISIONS IN CHANGING ENVIRONMENTS USING A HYBRID APPROACH OF DATA WAREHOUSE INTEGRATION

Marcus Preis[1], Juergen Seitz[2]

[1]*Baden-Wuerttemberg Cooperative State University, Germany, marcus.preis@web.de*
[2]*Baden-Wuerttemberg Cooperative State University, Germany, seitz@dhbw-heidenheim.de*
crossref http://dx.doi.org/10.5755/j01.em.17.2.2209

Abstract

We live in a globalized and fast changing world. The financial crisis means a lot of trouble for governments, administrations, companies and people. Sometimes the environment changes faster than the actors are able to react. So, at the moment Heraclitus' proverb 'Nothing is more constant than the change' is still up-to date, still valid and of high relevance. Companies merge with other companies; build joint ventures and virtual organizations. What has not been changed for a long time is the fact that only successful companies survive. Successful companies could be characterized as companies which are able to make better decisions faster. Therefore it is necessary to have the relevant data at the right time. This means that data for strategic decisions have to be available shortly after a merger or the establishment of a virtual organization for the whole new organization in a unified, complete and consistent form. It is not possible to build a data warehouse all from the scratch because this means a project of several years which is often still not in time and not in budget. A new way to develop an integrated data warehouse is needed. In this paper we show how a hybrid approach of data warehouse integration could look like.

Keywords: data warehouse, data warehouse integration, hybrid approach.
JEL Classification: L20, M10, M15.

Introduction

At the beginning of the 1980s, a suggestion was pulled forward to not only use the IT as a 'Data Processing System' (DTP) or a 'Management Information System' (MIS) (Hofmann & Schmidt, 2007, p. 13) as it has been taught before, but to use the IT as a weapon to gain competitive advantages compared to competitors (Wiseman & MacMillan, 1984). At the latest since then, the awareness was present that the proper and adequate employment of Information Technology could be vital in the emerging competition driven markets.

Nevertheless, there was always some kind of critical undertone, which scrutinised the value proposition of the IT. At this point two of these opposing points of view will be presented:

First, the productivity paradox which was introduced by Brynjolfsson in the year 1993 (Brynjolfsson, 1993): With the help of an empirical study, he released the hypothesis of a contradiction between the rapid growth in computational power in conjunction with increased investments in this sector and the relatively slow growth of the entire economy. Since this phenomenon was confuted (most publically by himself five years later (Brynjolfsson & Hitt, 1998), it will not be discussed in detail, but is only mentioned to clarify one specific mindset at that time.

Second, an article by Nicolas Carr, published in the year 2003 in the Harvard Business Review 'IT doesn't matter' (Carr, 2003): Therein, he depicts in the first step the undeniable importance of IT. In spite of everything, he does not see any strategic value and advantages justified by IT because according to Carr every company more or less uses the core functions of IT in the same manner. What he calls commodity factors have become 'costs of doing business that must be paid by all but provide distinction to none' (Carr 2003, p. 42). Carr draws a scenario of an ubiquitous IT with vanishing strategic advantages owed to IT degenerating to a commodity.

Both, the productivity paradox and the 'IT doesn't matter' issue indicate management's deficits with the exhaustion of the potentials of the IT (Hofmann & Schmidt, 2007, p. 16). But how are managers able to manage such a volatile and insecure business effectively without following the wrong trends and missing significant changes?

In 1985, a study by Michael Porter and Viktor Millar tried to emphasize the actual value of information. They state that in the framework of a business and its processes, information must be dedicated some kind of metaphysical role. Information is not a commodity or a process for one single incident, but rather an omnipresent constant which causes an 'information revolution' (Porter & Millar, 1985, p. 149). This paper is an attempt to manifest the IT as an integral building block within the process of creating added values. The result is their famous so-called 'value chain' (Porter & Millar, 1985, p. 150) where they divide 'a company's activities into the technologically and economically distinct activities' (Porter & Millar, 1985, p.

150) to perform its business. They further proposed nine distinct categories for assigning these value activities. But none of these categories was IT: 'Information technology is permeating the value chain at every point, transforming the way, value activities are performed and the nature of the linkages among them' (Porter & Millar, 1985, p. 151). This approach can be seen as highly subtle and ahead of the times because gradually the general perception of the IT changes and gradually managers start to think different.

Now the question arises which are the current or future trends in the field of IT. The following section will make a differenciation of these two terms in respect of the holistic situation of the IT business.

Trend or hype?

What is the difference between trend and a hype? – The factor time. While a hype is temporary matter a trend implies an ongoing paradigm.

An important and comprehensive disclosure about the happenings in the world of IT provides the 'CIO Agenda' by the Gartner research institute. Every year numerous Chief Information Officers (CIOs) are asked to participate in a survey which puts its focus on trends and future developments of enterprises. The aim of this report is to gain a better insight into businesses and to deviate trends in which ways enterprises and the economy are influenced by IT. Having a look at the CIO Agenda 2011 (McDonald & Aron, 2011) there remains no doubt that there is a huge change. The headline itself challenges CIOs to reimagine IT in order to be ready for the path breaking technologies. The most significant finding of this report is that 'over the next four years almost half of all CIOs expect to operate the majority of their applications and infrastructures via cloud technologies' (McDonald & Aron, 2011).

Therefore, a 'creative destruction' will take place, thinking about which IT processes remain unchanged, which IT processes will be slightly changed and which IT processes will be build up from scratch – reimagining so to say (McDonald & Aron, 2011). There will be a great focus on streamlining processes, getting rid of inflated, over-dimensioned solutions to be more business ready. The business impact of an IT department needs to rise urgently. The IT should do it's utmost to not being regarded as a cost pool in the balance sheet or a millstone around the neck of innovation pushers and key users. In contrast, the IT should deliver a solid base for employed systems, be an enabler and facilitator for new technologies and fertilizing changes and enhancements.

Further, business demands of the IT department to be modern and flexible and provide the state of the art of top technologies like mobility or business intelligence. Especially, in the complex domain of business intelligence and data warehousing managers need to pursuit a clear strategy and should also know about possible alternatives because business intelligence projects tend to be huge and expensive.

This paper discusses an approach of data warehouse integration, taking the fast changes of the companies' environments and the demanded reaction rate of decision makers into account.

A data warehouse

Being called 'information warehouse' at that time, the data warehouse concept was developed at the end of the 1970s as an internal project of IBM (Stahlknecht & Hasenkamp, 2005). Nowadays, the term data warehouse (DWH) is used as a synonym. The most popular and well-known definition of a data warehouse was given by the Bill Inmon in the year 1996:

'A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process' (Inmon 1996). Thus, a data warehouse is characterized by these four distinctive traits:

- **Subject-orientation**: In contrast to a transactional system, the subject-orientation deals with business cases, which means the data warehouse focuses basically on observance objects for managers like turnover or income return.
- **Integration**: The logical data definition (e. g., how the 'return on investment' is defined) and the physical data definition (e. g., number of places after decimal point) have to match.
- **Time-variation**: the information which is stored in a data warehouse usually has a reference to a point in time or a period of time and is regularly kept for several years.
- **Non-volatility**: Once committed to the data warehouse the data cannot be deleted again (persistence).

The term 'business intelligence' (BI) can also be traced back to the year 1996: 'Data analysis, reporting, and query tools can help business users wade through a sea of data to synthesize valuable

information from it – today these tools collectively fall into a category called 'Business Intelligence' (Anandarajan et al., 2004).

In conclusion, both terms have a very strong relationship but 'business intelligence' tends to cover more the frontend and reporting domain whereas 'data warehouse' is more about the underlying data collection, storage and provision. In the literature the combination of both is sometimes called 'analytical information system' (AIS) (Bauer & Günzel, 2008). For obvious reasons of difficult separation SAP for example calls its software solution 'business warehouse' to convey the holistic approach.

Data warehouse integration

Data integration is defined as the meaningful exchange of information between systems that are not originally designed to work together (Seligman *et al.*, 2002). Hence, data warehouse integration can in analogy be defined as the meaningful exchange of information between data warehouses that are not originally designed to work together. There exist two basic concepts for the integration of heterogeneous data sources: on the one hand side the physical integration approach and on the other hand the logical integration approach. Torlone differentiates 'tightly coupled' and 'loosely coupled' integration in this context (Torlone 2008; Torlone 2009).

Physical Data warehouse integration

Physical integration, also named as data migration (Kimball, 2002), is the process of creating a new system from heterogeneous source systems (Schmitt & Saake, 2005). To achieve one global schema it is necessary to migrate all the data from the original source systems to the newly built integrated information system (Berger, 2009). This solution seems remarkably simple because all diversities between the source systems have to be eliminated before the integration. Since queries can be sent to the global system query performance issues are not a matter. Furthermore, difficult drill-across queries are not of interest (Abelló et al., 2002). Hence, the major task of the physical integration is to eliminate all heterogeneities between the source systems, for example in data structures, the data model, the universe of discourse etc., a priori. This process is extremely labor intensive and error-prone (Doan & Halevy, 2005). Nevertheless, physical integration is not applicable in any case. For example in co-operations of companies or virtual organizations the partners will only be willing to share non-confidential data. Furthermore, privacy issues indicate a limit, e. g., in the health sector (Stolba et al., 2007). Kimball found out that the volume of data which is stored in a data warehouse system often causes an unmanageably high complexity of data migration projects (Kimball, 2002).

Logical Data Warehouse Integration

Logical integration describes the non-physical matching of heterogeneous data sources in order to receive one uniform representation of the underlying data. Such a relationship between the sources is modeled using mappings (Lenzerini, 2002). Since logical integration is in the spotlight of researchers for decades (for surveys, see Seligman *et al.*, 2002 and Doan & Halevy, 2005) a brief outline is provided:

The first noticeable efforts towards the logical integration of data sources were made by the extension of query languages to acquire so-called multi-database query languages (Grant *et al.*, 1993). With the formulation of the query the user is completely responsible for overcoming heterogeneity (Berger, 2009). Later, autonomous data marts are mapped; either complying with the global-as-view (GAV) or the local-as-view (LAV) approach (Lenzerini, 2002). Currently, an approach which is called both-as-view (BAV) (Boyd *et al.*, 2004) is in discussion which tries to combine the advantages of the previous two approaches. Several publications (Niemi *et al.*, 2002; Pedersen *et al.*, 2002a; Pedersen *et al.*, 2002b; Pérez *et al.*, 2006; Hümmer *et al.*, 2003; Trujillo *et al.*, 2004) aim on the integration of autonomous data marts without a global schema, using the Extensible Markup Language (XML).

Implications of the modern economy on data warehousing

In the two previous sections, two approaches of data warehouses integration were introduced. But why should managers even care about those scenarios? In the following two business cases it will be outlined in order to deliver a justification and to draw awareness for the importance of data warehouse integration.

First of all the scenario of mergers and acquisitions. 'A recent study from Thomson Reuters and Freeman Consulting Services concludes that the global market for M&A will surge 36 % in 2011 to over US$3 trillion.'(Kramer, 2011). The reason for M&As is rather simple: Instead of spending huge amounts in R&D a company aims to buy another company which is an evicted specialist in a particular sector. What comes on top is the struggle for patents and protected intellectual property. One could only marvel at the millions and billions of dollars, which are in dispute if big companies like Apple, Google or Samsung argue in court over patent infringement. Hence, the big companies also acquire other companies to receive better patent portfolios as a result. In the year 2010, there have been more than 70 acquisitions related to technology and IT companies conducted by the major industry leaders. As the largest player, Google alone bought out 25 companies in 2010. Oracle made nine deals, spending US$7.4 billion for Sun Microsystems and $1 billion for the Art Technology Group. 'The catalysts are growth, technology, cash and the economy' (Kramer, 2011). And the outlook for the future is bright. Those, who survived the crisis look at recovering stock markets and rapid technological progress.

Secondly, the environment of virtual organizations (VO) where two or more parties work together in a rather loosely-coupled framework and only contribute their core competencies (Byrne *et al.*, 1993). IT is in this constellation responsible for the speed and the quality of the information flow and exchange and the coordination of a large variety of tasks (Malone & Rockart, 1993). Arnold *et al.* (1995) present a list of tools for supporting the creation and the running of daily business activities with the aid of a VO. One system among these that is very significant for the underlying article is the so-called 'management information system' which can be understood as a reporting system with a subjacent data warehouse. The crucial point is that the applied IT systems, first and foremost the data warehouse, have to be of overall validity. Only in this way it is guaranteed that all relevant facts from all involved organisations are considered. Mertens & Faisst (1997) demand the IT systems to be flexible and adaptable like a 'plug and play' system to facilitate the response to new processes, systems or even partners. This may be realizable with easy applications which communicate straight forward over standard interfaces but not with data warehouses.

Logical or physical data warehouse integration – an either or decision?

The previous section showed the demand for effective and most important of all fast integration methods for data warehouses. If a data warehouse in its original meaning serves as a collection of data to support the decision making process valid and well-founded decisions will only be made if all relevant data is taken into consideration. Given the case that two companies merge it is not econmically reasonable to make decisions across the group with information which is related to only a certain part of the group. For this reason data warehouse integration methods and strategies have to be evaluated (Preis & Seitz, 2012). The physical data warehouse integration is, as a total migration project of systems and databases labour-intensive, complex and expensive. There are also certain restrictions like privacy (in the health sector) or confidentiality (in VO situations). Hence, the physical data warehouse integration is not applicable for certain projects or not applicable for the entire scope of the project. The logical data warehouse integration offers a fast integration of additive systems and is therefore very flexible. In the fast changing environment of the internet, it is more open to the demands of the web 2.0.

From a business point of view a hybrid approch of data warehouse integration would be desireable, combining the best of both worlds.

The hybrid approach

The issue of facing heterogeneities in the environment of data warehouses will be in the spotlight of researchers in the near future. In an era of growing data volumes caused by the information age (Web 2.0) and elaborated techniques, e. g., data mining, successful solutions for homogeneous analysis of data and reporting have to be developed. The hybrid approach is not a concept of yesterday but of tomorrow. Even though it may be technically harder to implement, the result is worth to be considered.

ISSN 2029-9338 (ONLINE)
ISSN 1822-6515 (CD-ROM)

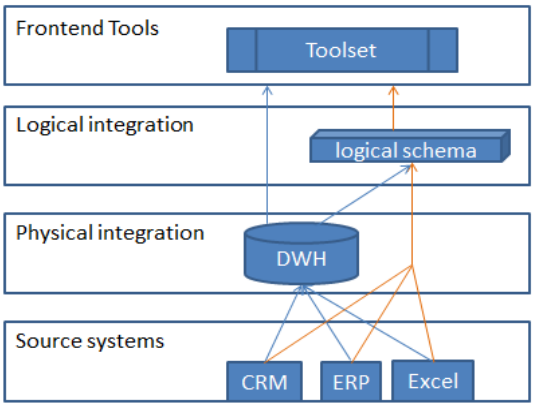

Figure 1. Possible architecture framework of the hybrid data warehouse integration

Figure 1 shows the discussed hybrid framework matched to the environment of logical and physical data warehouse integration. Forrester draws an advanced picture of the hybrid data warehouse integration, calls it the 'hourglass architecture' and states that leading companies see this solution as a best practice (Hopkins, 2011). Besides the need of additional costs and resources several potential advantages have to be mentioned:

- Meeting the demand of the business to a better extent because of the relatively easy combination of historical and real-time data.
- Simplification of the roll-out processes because the hybrid approach covers a broad spectrum of functional requirements which leads to a higher level of satisfaction with the customers and a higher approval rate.
- Consistency of the data model: Upcoming new systems do not destroy the whole system landscape and can smoothly be integrated.
- Increase in performance: Reports that used to run for several minutes or even hours can be generated in a fraction of the time needed before because of the effective interplay between OLTP and OLAP system.

This interplay is technically realized within the HyPer project which combines OLTP and OLAP main memory databases (Kemper & Neumann, 2011) and leads to future of business intelligence solutions based on the in-memory technology.

Conclusion and Outlook: The Web 2.0 as a facilitator for data warehousing

People who grew up with computers and the Internet are called 'digital natives' (Gluchowski *et al.*, 2011). This highly computer affine generation will be the office workers of tomorrow. For them, terms like 'feeds', 'mashups', 'blogs', 'social software' are common. Not only the future working force but also the managers of tomorrow will have a strong relationship to the cutting edge of innovations and the progress of key technologies in the IT sector. Therefore, neither the management, nor the employees of a company will be satisfied with a less sophisticated level of information technologies than they employ at home.

On the opposite side, business intelligence or analytical information systems rely on business processes, system architectures and frameworks that have not significantly changed within the last ten years (Gluchowski et al., 2011). For the average backend system and the core processes like the extract, transform and load (ETL) process this devolopement might be adequate. In a three tier architecture the user does not care whatever happens in the two underlying tiers as long as the presentation layer complies with his/her needs and requirements.

For managers and whole IT departements the current rapid changes in the main domains of interest (cloud computing, virtualization, mobile technologies, IT management and business intelligence (McDonald, & Aron, 2011) present both threads and opportunities. Threads in terms of not being able to mainly satisfy the customers' needs and opportunities in terms of drawing awareness to the productivity of the IT and to astonish users.

The Internet protocol (IP) traffic is doubling every two years whereas mobile traffic is doubling even every year. In 2012 the Internet will be 75 times larger than it was in 2002 (McDonald & Aron, 2011). May

it be for reasons of patents or any other cause, the Internet facilitates mergers and acquisitions. Hence the data volumens tend to be eminently huger than in the past.

Since the Internet is the all-dominant changing paradigm of data warehousing managers need to take new approaches of consolidating data into consideration. As depicted in the introduction managers need to evaluate more agile and target oriented methods for the integration of data warehouses. The underlying paper presents a suggestion how a hybrid framework of data warehouse integration can possibly look like. Especially the agility and the faster reaction time needs to be pointed out for the subjacent data layer serves as a mere foundation for the development of highly interactive, ergonomically designed front end applications.

References

1. Abelló, A., Samos, J., & Saltor, F. (2002). On relationships offering new drill-across possibilities. In: D. Theodoratos (Ed.), Proceedings of DOLAP 2002. ACM Fifth International Workshop on Data Warehousing and OLAP, 7-13.
2. Anandarajan, M., Anandarajan, A., & Srinivasan C. A. (2004). Business Intelligence Techniques. Berlin, Heidelberg et al.: Springer.
3. Arnold, O., Faisst, W., Härtling, M., & Sieber, P. (1995). Virtuelle Unternehmen alsUnternehmenstyp der Zukunft? HMD, 185, 8-23.
4. Bauer, A., & Günzel, H. (Eds.) (2008). Data Warehouse Systeme – Architektur, Entwicklung, Anwendung. Heidelberg: Dpunkt-Verlag GmbH.
5. Berger, S. (2009). FedDW: a Model-Driven Approach for Querying Federations of Autonomous Data Marts. Dissertation. Johannes-Kepler-Universität Linz.
6. Boyd, M., Kittivoravitkul, S., Lazanitis, C., McBrien, P., & Rizopoulos, N. (2004). Automed: A BAV data integration system for heterogeneous data sources. In: A. Persson & J. Stirna (Eds.), Proceedings of the 16th International Conference on Advanced Information Systems Engineering CAiSE (pp. 82-97). Berlin: Springer.
7. Brynjolfsson, E. (1993). The productivity paradox of information technology. Communications of the ACM, 36(12), 66-77.
8. Brynjolfsson, E., & Hitt, L. M. (1998). Beyond the Productivity Paradox: Computers are the Catalyst for Bigger Changes. Communications of the ACM, 41(8), 49-55.
9. Byrne, J.A., Brandt, R., & Port, O. (1993, February 8). The virtual corporation. Business Week, pp. 36-40.
10. Carr, N. (2003). IT Doesn't Matter. Harvard Business Review, 81(5), 41-49.
11. Doan, A., & Halevy, A. Y. (2005). Semantic integration research in the database community: A brief survey. AI Magazine, 26(1), 83-94.
12. Gluchowski, P., Schieder, C., & Böhringer, M. (2011). Web-2.0-inspirierte Business-Intelligence-Lösungen für die Anwender der Zukunft. HMD, 48(282), 16-25.
13. Grant, J., Litwin, W., Roussopoulos, N., & Sellis, T. K. (1993). Query languages for relational multidatabases. VLDB Journal, 2(2), 153-171.
14. Hofmann, J., & Schmidt, W. (Eds) (2007). Masterkurs IT-Management. Wiesbaden: Vieweg Verlag.
15. Hopkins, B. (2011). Data virtualization reaches the critical mass. Retrieved from http://purl.manticoretechnology.com/ImgHost/582/12917/2011/resources/white_papers/DataVirtualizationReachesCriticalMass.pdf.
16. Hümmer, W., Bauer, A., & Harde, G. (2003). XCube – XML For Data Warehosues. In: Proceedings of the 6th ACM International Workshop on Data Warehousing and OLAP. DOLAP '03. 1-8.
17. Inmon, W. H. (1996). Building the Data Warehouse. New York: Wiley Computer Publishing.
18. Kemper, A., & Neumann, T. (2011). HyPer: A hybrid OLTP & OLAP main memory database system based on virtual memory snapshots. In: Proceedings of 2011 IEEE 27th International Conference on Data Engineering (ICDE). 195-206.
19. Kimball, R. (2002). The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. New York: John Wiley & Sons.
20. Kramer, H. (2010). 2011 Will Be The Year Of M&A. Retrieved from http://www.forbes.com/2010/12/22/mergers-acquisitions-business-trend-2011-kramer-intelligent-investing.html.
21. Lenzerini, M. (2002). Data integration. A theoretical perspective. In: L. Popa (Eds.), Proceedings of PODS. 233-246.

22. Malone, T. W., & Rockart, J. F. (1993). How will information technology reshape organizations? In: S. P. Bradley, J. A. Hausman, & R. L. Nolan (Eds.), Globalization, Technology, and Competition: The Fusion of Computers and Telecommunications in the 1990s, Boston: Harvard Business School Press.
23. McDonald, M. P., & Aron, D. (2011). Reimagining IT: The 2011 CIO Agenda. Retrieved from http://www.gartner.com/id=1524714.
24. Mertens, P., & Faisst, W. (1997). Virtuelle Unternehmen, Einführung und Überblick. In: D. Hahn, & B. Taylor (Eds.), Strategische Unternehmensplanung – Strategische Unternehmensführung (pp. 953-968). Heidelberg: Physica-Verlag.
25. Niemi, T., Niinimäki, M., Nummenmaa, J., & Thanisch, P. (2002). Constructing an OLAP cube from distributed XML data. In: D. Theodoratos (Ed.), Proceedings of DOLAP 2002. ACM Fifth International Workshop on Data Warehousing and OLAP, 22-27.
26. Pedersen, D., Riis, K., & Pedersen, T. B. (2002a). Query optimization for OLAP-XML federations. In: D. Theodoratos (Ed.), Proceedings of DOLAP 2002. ACM Fifth International Workshop on Data Warehousing and OLAP, 57-64.
27. Pedersen, D., Riis, K., & Pedersen, T. B. (2002b). XML-extended OLAP querying. In: Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM'02), 195-206.
28. Pérez, J. M., Berlanga, R., Aramburu, M. J., & Pedersen, T. B. (2006). Integrating Data Warehouses with Web Data: A Survey. A DB Technical Report.
29. Porter, M., & Millar, V. (1985). How information gives you competitive advantage. Harvard Business Review, 63(4), 149-160.
30. Preis, M., & Seitz, J. (2012): A Hybrid Approach of Data Warehouse Integration Based on New Storage Technologies. In: International Journal of Advances in Computing and Management (IJACM). 1(1), 40 - 46.
31. Schmitt, I., & Saake, G. (2005). A comprehensive database schema integration method based on the theory of formal concepts. Acta Inf. 41(7-8), 475-524.
32. Seligman, L., Rosenthal, A., Lehner, P., & Smith, A. (2002). Data integration: Where does the time go? IEEE Data Engineering Bulletin.
33. Stahlknecht, P., & Hasenkamp, U. (2005). Einführung in die Wirtschaftsinformatik. Berlin: Springer.
34. Stolba, N., Tjoa, A. M., Mueck, T., & Banek, M. (2007). Federated Data Warehouse Approach to Support the National and International Interoperability of Healthcare Information Systems. In: Proceedings of the 15th European Conference on Information Systems (ECIS), 1-4.
35. Torlone, R. (2008). Two approaches to the integration of heterogeneous data warehouses. Distributed and Parallel Databases, 23(1), 69-97.
36. Torlone, R. (2009). Interoperability in Data Warehouses. In: L. Liu & T. Öszu (Eds.), Encyclopedia of Database Systems (pp. 1560-1564). Berlin: Springer.
37. Trujillo, J., Luján-Mora, S., & Song, I. (2004). Applying UML and XML for Designing and Interchanging Information for Data Warehouses and OLAP Applications. Journal of Database Management (JDM), 14(1), 41-72.
38. Wiseman, C., & MacMillan, I. C. (1984). Creating competitive weapons from information systems. The Journal of business strategy, 5(2), 42-49.