IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP5205

This paper appears in *Managing Modern Organizations Through Information Technology*, Proceedings of the 2005 Information Resources Management Association International Conference, edited by Mehdi Khosrow-Pour. Copyright 2005, Idea Group Inc.

Data Warehouse Projects: Increasing Awareness for Assumptions and Complexity from the Perspectives of Business and IT

Alexander Dreiling
European Research Ctr for Info. Sys., University of Münster, Leonardo-Campus 3, 48149 Münster, Germany, alexander.dreiling@ercis.de

Marta Indulska
UQ Business School, University of Queensland, Ipswich QLD 4305, Australia, m.indulska@business.uq.edu.au

ABSTRACT

Data warehouse research quickly gained momentum in the mid 1990s after conceptual foundations have been laid and rapid advancements in database technology made the management of large amounts of data possible. Since that time, many technical solutions for various aspects of data warehouse development projects have been proposed. However, to this day, this area of research appears to be firmly nested within technical departments, with the role of business often being neglected or relegated to simple assumptions. In this paper, we argue that there is a need for broadening the scope in order to include both the business and IT aspects in these long running projects. This more holistic approach should potentially increase the chances of project success. Assuming such an approach, we briefly outline the stages required in data warehouse projects and the necessity for communication at the different stages. We argue that continuous absence of communication between business and IT significantly decreases the chances for data warehousing success by providing examples and motivation for mandatory alignment processes at certain vital stages in such projects.

INTRODUCTION

Recent research in understanding the impacts of aligning business and Information Systems (IS) strategies suggests that closing the gap between business and IS leads to improved business performance, and thus needs to be pursued continually (Chan and Huff, 1997, Sabherwal and Chan, 2001). Two of the critical factors influencing alignment success are *communication between business and IT executives* and *shared domain knowledge* (Reich and Benbasat, 2000).

Yet, currently, data warehouse (DWH) research is mainly nested within Information Technology (IT) departments. The role of business in data warehousing projects seems to be typically of a lesser concern. The complex interdependencies that arise from the area of conflict between business and IT seem to be of an even lesser concern still. The divide and conquer nature of contemporary DWH research leads to certain problematic phenomena, such as the IT team requiring specific input in a predefined format that might not be deliverable by the business, or the business demanding solutions that might not be deliverable by IT (either because they are not possible, e.g. NP-complete such as the traveling salesman problem or maximum clique problem, or because they are not feasible, e.g. the required data becomes too expensive).

Typically, IT projects—we use this term here in the absence of other commonly accepted terms even though we argue throughout the paper that DWH projects are not solely IT projects—can be divided into several phases. For the purpose of this paper we use a five phase model that includes phases of *project definition*, *requirements engineering*, *design*, *implementation*, and *maintenance.* These terms and underlying concepts, which are common in system development, are slightly adapted here in order to cater for the specifics of DWH projects. Figure 1 depicts this model and some decomposed terms within the several stages. The intention of the arrows is to depict that there is a necessity for constant communication between business and IT in order to successfully conduct DWH projects.

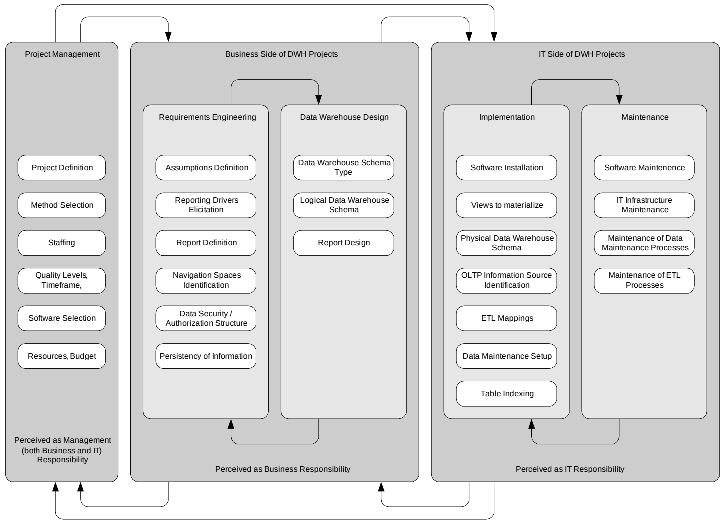
Using this model, this paper aims at highlighting some of the complexities within DWH projects and thus tries to motivate the necessity of enlarging the scope of DWH research in order to make it a truly interdisciplinary research area. In order to do so, we sketch the role of project management in the next section. Subsequently, we will elaborate on requirements engineering and DWH design and motivate why these tasks should be mainly within the hands of the business with significant input from IT. The subsequent section examines the IT side of DWH projects and motivates the necessity for constant business input. We will then discuss the interdependencies between business and IT and conclude with a short summary and outline of future work.

PROJECT MANAGEMENT

The management of DWH projects does not significantly differ from other major information systems implementation or customization projects. The main purpose of project management is to achieve project success. This is everything but trivial. Ongoing research on escalation and failure of information systems development and implementation projects indicates that we are just beginning to understand the reasons for escalation and failure (Keil et al., 2000, Keil and Robey, 2001, Schmidt et al., 2001, Smith et al., 2001). In DeLone and McLean's IS success model (DeLone and McLean, 2003, DeLone and McLean, 1992) and Seddon's extension to this model (Seddon, 1997), *information quality* and *system quality* are influencing IS success. Wixom and Watson's model for DWH success (Wixom and Watson, 2001) includes the constructs *data quality* and *system quality* to explain success. Given the importance of these factors, it is obvious that if IT departments are to provide computer-based information systems that support the business adequately, the quality of the deliverable information and systems depend on the project as a whole rather than on single isolated steps.

Basic decisions during project management include the staffing of the project, the definition of quality levels, the setting of a time frame, creating a budget, or allocating resources to the project. Software, as well as methods for the implementation or configuration of a data warehouse,

Figure 1. Decomposition of DWH Projects



needs to be selected, and project goals need to be defined. From these basic tasks it becomes apparent that project management cannot be clearly classified as a business or an IT responsibility. Apart from the necessity of including both business and IT experts in order gain adequate estimations of the complexity and cost of the steps within the course of a DWH project, proper project management also resolves conflict which, in the case of a DWH project, can mainly arise from the fact that business and IT people typically do not speak the same language. Hence, project management must not be the sole responsibility of business or IT. Rather, an interdisciplinary team needs to be managed.

"BUSINESS SIDE" OF DATA WAREHOUSE PROJECTS

Two major tasks are typically perceived as a business responsibility in DWH projects: *requirements engineering* and *data warehouse design*. Even though requirements engineering is clearly a business responsibility, a significant contribution of IT is necessary in order to perform this task appropriately. DWH design is at the borderline between business and IT but we argue below that it should be mainly within the hands of business. However, it is apparent that DWH design requires significantly more contribution from IT.

Requirements Engineering

Data warehouses serve a business purpose and thus, their specification needs to be business-driven. When it comes to the business purpose, a commonly used assumption within IT is that information needs to be provided in order to make a reasoned decision. Decisions are typically seen as the choice of one course of action from a set of alternatives after having evaluated their estimated consequences (e.g., Koontz and Cyril, 1968, March and Simon, 1958). Since this process always deals with uncertainties, and thus in itself maybe consistent but meaningless towards actual real-world problems, business as a science provides some alternatives. Postmodern Organization Theory questions concepts such as *management* or *decision*, and, in its most radical form, even denies the existence of these concepts (e.g., Boje et al., 1996). This must not be ignored as, particularly in large or public organizations, the level of bureaucracy can lead to power structures that differ from organizational structures. Management decisions become irrelevant if they are not appropriately communicated or if the organization resists them. Hence, making a decision, and providing information in order to make a decision, is not necessarily the bottleneck. Organizational Learning (Argyris and Schön, 1978) provides an alternative fundamental concept to decision making, however, this is typically not considered in DWH research even if DWH design may be significantly different with altered—and possibly more meaningful—assumptions.

After clearly articulating the underlying assumptions of a reporting system as the main deliverable of a DWH project, there is a need to define *reporting drivers* such as *critical success factors* (Rockart, 1979), *managerial tasks*, or *managerial goals and objectives* (for use in context of data warehouses Becker et al., 2003, for a general discussion on goal setting Locke and Latham, 1990). These reporting drivers will fundamentally determine the content and structure of the DWH. If there has been an agreement as to what will drive the managerial reporting system, *report definition* needs to be based on elicited reporting drivers. Prospected users and managers, being the target group of the DWH, must be involved at this stage since *user participation* and *management support* are factors influencing DWH success (Wixom and Watson, 2001). Finally, with the notion of OLAP, a managerial report is dynamic rather than static. An initial view provided to a manager is only the entry point for a range of managerial analyses. In order to avoid information overflow, and in order to protect information from abuse, it might become necessary to reduce the scope of the object of examination for a manager/user. Hence, *managerial navigation spaces* through enterprise information must be defined (Becker et al., 2003). This task is further detailed within the definition of authorization structures. As DWH projects potentially integrate a broad variety of information sourced from across the whole organization, and, as some countries have very strict data security laws, authorization structures need to be thoroughly defined. Finally, during the definition of *information persistency* it needs to be decided from a business perspective for how long information will be made available, and at what level of detail.

Data Warehouse Design

The major step in DWH design can be seen in defining the logical DWH schema (Lechtenbörger, 2001). Accordance with the requirements elicited in the previous step needs to be ensured. Several possibilities, such as creating separate data marts for separate managerial functions or organizational units, arise during this phase. We treat this step mainly as a business responsibility because the decision as to what information needs to be accessed, and by whom should not be decided within IT.

After the logical definition of the DWH schema the defined reports from the previous step have to be designed. This step becomes necessary for DWH environments that store data at a very detailed level. If, for example, only *sales quantities* and *sales prices* of single transactions are stored, then the figure *total sales of a period* needs to be calculated. Report design enables the definition of reports in a more operational way than in the previous step where the definition might have led to the general statement that a sales report is necessary.

Interactions within Business

The previous two subsections only very superficially sketch the necessary steps carried out during data requirements engineering and DWH design. However, it becomes immediately apparent that there are complex interdependencies within these steps. This is especially true due to the necessity of alignment of the sub-steps within the phases, which thus require continuous adjustment. It is furthermore unreasonable to assume that there is a strict order that implies that requirements engineering is always carried out before DWH design. Rather, these phases need to be executed in parallel or as an iterative process (as the reader will see, this applies to the entire DWH project).

Examples of interactions within business start with *assumptions* and *reporting driver elicitation*: How do reporting drivers need to be elicited based on the assumptions? In order to assist decision making it would be necessary to enquire about decision situations. In order to assist learning it would be necessary to confront managers with consequences of their actions. If the enquiries during reporting driver elicitation do not lead to the desired results, assumptions might need to be adjusted. Another example concerns *report definition* and *report design*: whereas the former defines the necessity, the latter precisely specifies the information, leaving no room for ambiguity. For example, *turnover* can be defined in several ways. It can be defined as the aggregated sales price within a period of invoiced items, or of paid invoices, or, if the payment does not match the invoiced amount, it can be defined as the aggregated amount of incoming money resulting from the payment of the invoices. If ambiguities cannot be resolved, or if their resolution leads to discrepancies within the defined reports, redefinition of reports might become necessary. A third example concerns *report design* and *navigation spaces identification*: re-specification may be necessary if the report design phase leads to reports that contain information which is structurally not included in a navigation space. For example, turnover analysis along the ‘sales representative’ dimension is required, but this dimension is not included in the navigation space.

The presented examples suggest that DWH design might not be a linear process through several phases within the tasks that are mainly business responsibilities. If, and only if this is not the case, how can DWH research that considers only parts of the entire process possibly guarantee the applicability of its deliverables? Since constant steps back and forth can become unavoidable in certain situations, does it make sense to separate requirements engineering and DWH design in a divide and conquer fashion? We argue here that both phases need to be treated holistically, considering the manifold interdependencies within the single steps outlined here.

THE “IT SIDE” OF DATA WAREHOUSE PROJECTS

The IT component of data warehouse projects can be defined in terms of two main tasks; the *implementation* and the subsequent *maintenance* of the data warehouse. While these tasks are mainly the responsibility of the IT team, they also require the involvement of managers, users, and project stakeholders in order to increase the chances of project success. The implementation and maintenance stages are therefore perceived to be driven by IT in consultation with business.

Data Warehouse Implementation

The first step in the implementation of the data warehouse is the setup of the system infrastructure that was decided upon at the project management stage. This step involves the installation of any required hardware, networks, software, and database management systems. Once the installations are carried out and tested, the IT team is able to undertake the core data warehouse implementation tasks.

As depicted in the model proposed in Figure 1, the IT team designs the data warehouse based on the information obtained from the earlier stages of requirements engineering and data warehouse design. One of the critical factors in the implementation of the data warehouse is often the selection of the views that will be materialized in the data warehouse, and the selection of those that will remain virtual (i.e. will require computation on the fly). The logical data warehouse schema that is passed on to the IT team describes the data and the relationships between the data that will be stored in the data warehouse. However, it is important to identify any pre-computed aggregations (views) that should exist in order to improve the response time of the system (Indulska, 2000). The IT team is responsible for taking requirements like query and update frequencies, and producing an optimal mix of virtual and materialized views for the data warehouse. Such selection of views is not an easy task, it is a fine balance that can potentially significantly affect the performance and the cost of the data warehouse. Since some open-period data will require updates at predetermined intervals, it is important to consider how often these updates will occur versus how often the data will actually be queried in the data warehouse. This area has been an active research topic since the birth of data warehouses and various IT solutions have been proposed (Baralis et al., 1997, Han et al., 1998, Indulska, 2000, Theodoratos et al., 2001, Theodoratos and Sellis, 2000, Yang et al., 1997).

Once these data structures are agreed upon, the next step is the creation of the physical data warehouse schema and the identification of OLTP sources. At the same time it is imperative that meta data be defined as well. The meta-data includes the descriptions and definitions of the various data items in the data warehouse. It also includes information about the original sources of the data, any relevant business rules, as well as the documented extent of extraction and transformation that the data has undergone.

The process of data extraction, transformation and loading (ETL) can begin once a data map has been defined, which defines the sources of the data (i.e. specific fields in the underlying databases), their corresponding target fields in the data warehouse, and any necessary data modification. The extraction stage of the ETL process is responsible for the selection of data from the underlying sources and making that data available to the transformation component of ETL. Due to the possibility of heterogeneous data sources, the extraction stage may consist of a number of different extraction mechanisms for the different underlying sources. The transformation stage of the ETL process receives as input the extracted data and performs the required data manipulation in order to prepare the data for loading into the data warehouse. This is one of the critical steps in ensuring data quality and integrity in data warehouse projects. The final stage of ETL, loading, has the simple yet potentially time consuming task of loading each cleansed and transformed data record into its corresponding data structure in the data warehouse.

After the ETL process has been finalized and the data warehouse has been tested, it is ready to be used. There are however a number of additional steps that can be perfomed depending on the needs of the organization. Depending on the expected query patterns, index structures may be set up in order to improve the response time of the data warehouse. Additionally, if the data warehouse is not a purely historical one, a data maintenance strategy has to be operationalised.

The maintenance of the data warehouse in the presence of underlying data source updates has been well researched and different strategies exist, from incremental updates to self-maintaining views (Colby et al., 1997, Labio et al., 1999, Yang and Widom, 2000, Zhuge et al., 1995). The critical issue with any data warehouse which requires data updates is the issue of minimizing the maintenance window during which the data warehouse, or parts of it, are unusable (Labio et al., 1999).

Data Warehouse Maintenance

Once the data warehouse has been implemented it requires constant monitoring much like any other system. The core issues taken into consideration at this stage are the maintenance of software, hardware and data maintenance processes. Maintenance of software and hardware includes regular updates to systems or software, updates to the ETL process in case of changes of the underlying data sources or structures, as well as hardware upgrades in case of failures or increased processing demands. Modifications to the data maintenance processes may also be required if the data sources or the data warehouse have undergone any changes. Such modifications would also be required if additional data requirements were to be identified and implemented.

Interactions within IT

While the previous two sub-sections briefly outlined the core elements of the IT aspect of data warehouse projects, it is clear that there is a degree of freedom in terms of the order in which these tasks have to be performed. For example, the ETL process can begin before the data structures have been implemented in the data warehouse, if a data staging area is being used for data transformation. Likewise, the selection of views to be materialized can take place before or during the stages of system infrastructure setup.

As with the business aspect of data warehouse projects discussed in earlier sections, there are also some interactions between the two stages of the IT related activities. For example, the update of data maintenance processes might occur due to a change in the data map defined prior to carrying out ETL. Hardware maintenance of the other hand may be impacted by any significant increases to data storage needs which may result due to changes in the selection of materialized views. Additionally, new software or new versions of software might reduce some of the restrictions that were put in place at the implementation stage of the project.

INTERACTIONS BETWEEN BUSINESS AND IT

Data warehouse projects are long running complex projects which, to be successful, require thorough communication between the users, the planning team and the technical development team. These projects typically do not follow the systems development lifecycle in the same manner that information systems projects do. While the data warehouse project is an iterative one, it is rare that all requirements are specified before the implementation stage begins. In fact, data warehouse projects are often built incrementally, starting with the fundamental data requirements first and then adding more functionality in later stages. However, whichever method is used for DWH design and implementation, it requires constant interaction between the business and the IT aspects of the project. The requirement here is not just for interaction, but for the efficient sharing of requirements at the required level of granularity for a specific stage of the project.

Clearly, the initial stage of project planning is heavily dependant on the expertise of both business analysts as well as the IT team. Project Management is an ongoing activity in terms of resolving conflict, and controlling progress. *Distancing* (Argyris and Schön, 1978) between business and IT must be avoided in order to facilitate a learning process throughout a DWH project. This is the only way to effectively build up a shared domain knowledge between business and IT and effectively solve upcoming problems within the project. Without the involvement of both teams, there is a lack of alignment between the needs of the organization and the data warehouse project. This necessity for communication is evident at many stages of data warehouse projects. For example, such communication is vital at the project planning stages in order to correctly budget for the necessary resources. Also, there is a need for the communication of potential query patterns and query frequencies. While the potential query patterns mostly impact the logical data warehouse schema development, they also impact view selection, which in turn impacts the performance of the data warehouse, as well as indexing and data maintenance strategies. Furthermore, as business requirements, and thus query patterns, change over time, there needs to be communication between the two teams to make sure that the new requirements are reflected in the data warehouse.

CONCLUSIONS AND FURTHER RESEARCH

This paper presented a brief outline of the tasks and division of labour involved in data warehousing projects, from the initiation stage right through to the maintenance stage. Each identified stage was briefly described, followed by examples of the required interactions at the specified stages.

In general, current literature on data warehouse design makes some assumptions which are not always realistic or aligned with business needs. For example, many bodies of work approach the problem of view selection in data warehouses with very naive assumptions about user requirements. Other assumptions can be made if the developments in business science are considered more carefully. Taking them into account can potentially lead to different and more relevant solutions within data warehouse research.

Additionally, these specific solutions need to be applicable within a complex area of conflict between business and IT. Conflict can occur because business people sometimes demand (much like customers from suppliers) what cannot be feasibly delivered. On the other hand, IT people sometimes demand input from business that cannot be feasibly delivered, or, they deliver technical solutions that do not solve the identified business problems. Therefore, we argue that a broader scope is necessary in order to grant DWH research more relevance. Approaching DWH projects in a 'divide and conquer' manner may result in lack of relevance to the business needs (through a lack of communication of user requirements at the required granularity of detail).

We envision future research in this area to include a more comprehensive high-level guide to data warehouse projects, and, the development of a modeling language, or an integrated set of modeling languages, that assists an entire DWH project holistically. Such a modeling language should model the requirements at different levels of granularity so that it can be used at different stages of the DWH project and by different team members (be it business or IT). For example, such a modeling language should be able to capture not just the high-level data requirements, but also information about query patterns, frequencies, and relative importance of data items.

REFERENCES

Argyris, C. and Schön, D. A. (1978) Organizational Learning: A Theory of Action Perspective, Addison-Wesley Publishing Company, Reading, MA

Baralis, E., Paraboschi, S. and Teniente, E. (1997) Materialized View Selection in a Multidimensional Database, In Proceedings of the Proc. 23rd International Conference on Very Large Databases, (Eds, Jarke, M., Carey, M., Dittrich, K. R., Lockovsky, F., Loucopoulos, P. and Jeusfeld, M. A.), Morgan Kaufmann Publishers, pp. 156-65.

Becker, J., Dreiling, A., Holten, R. and Ribbert, M. (2003) Specifying Information Systems for Business Process Integration - a Management Perspective, Information Systems and e-Business Management, 1(3), 231-263.

Boje, D. M., Gephart Jr., R. and Thatchenkery, T. J. (Eds.) (1996) Postmodern Management and Organization Theory, SAGE Publications, Thousand Oaks, CA

Chan, Y. E. and Huff, S. L. (1997) Business Strategic Orientation, Information Systems Strategic Orientation, and Strategic Alignment, Information Systems Research, 8(2), 125-150.

Colby, L., Kawaguchi, A., Lieuwen, D., Mumick, I. and Ross, K. A. (1997) Supporting Multiple View Maintenance Policies, In Proceedings of the Proc. ACM SIGMOD Conference on Management of Data, Tucson, Arizona, pp. 405-416.

DeLone, W. H. and McLean, E. R. (1992) Information Systems Success: The Quest for the Dependent Variable, Information Systems Research, 3(1), 60-95.

DeLone, W. H. and McLean, E. R. (2003) The Delone and Mclean Model of Information Systems Success: A Ten-Year Update, Journal of Management Information Systems, 19(4), 9.

Han, J., Stefanovic, N. and Koperski, K. (1998) Selective Materialization: An Efficient Method for Spatial Data Cube Construction, In Proceedings of the Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, Australia.

Indulska, M. (2000) Shared Result Identification for Materialized View Selection, In Proceedings of the Proc. 11th Australasian Database Conference, Canberra, (Ed, Orlowska, M. E.), IEEE Computer Society, pp. 49-56.

Keil, M., Mann, J. and Rai, A. (2000) Why Software Projects Escalate: An Empirical Analysis and Test of Four Theoretical Models, MIS Quarterly, 24(4), 631-664.

Keil, M. and Robey, D. (2001) Blowing the Whistle on Troubled Software Projects, Communications of the ACM, 44(4), 87-93.

Koontz, H. and Cyril, O. D. (1968) Principles of Management: An Analysis of Managerial Functions, McGraw-Hill Book Company, edition 4, New York

Labio, W. J., Yerneni, R. and Garcia Molina, H. (1999) Shrinking the Warehouse Update Window, SIGMOD Record, 28(2), 383-94.

Lechtenbörger, J. (2001) Data Warehouse Schema Design, Dissertation Thesis, University of Münster, Dept. of Mathematics and Computer Science, Münster

Locke, E. A. and Latham, G. P. (1990) A Theory of Goal Setting and Task Performance, Prentice-Hall, Englewood Cliffs, NJ

March, J. G. and Simon, H. A. (1958) Organizations, John Wiley & Sons, New York

Reich, B. H. and Benbasat, I. (2000) Factors That Influence the Social Dimension of Alignment between Business and Information Technology Objectives, MIS Quarterly, 24(1), 81-113.

Rockart, J. F. (1979) Chief Executives Define Their Own Data Needs, Harvard Business Review, 30(Mar-Apr), 81-93.

Sabherwal, R. and Chan, Y. E. (2001) Alignment between Business and Is Strategies: A Study of Prospectors, Analyzers, and Defenders, Information Systems Research, 12(1), 11-33.

Schmidt, R., Lyytinen, K., Keil, M. and Cule, P. (2001) Identifying Software Project Risks: An International Delphi Study, Journal of Management Information Systems, 17(4), 5-36.

Seddon, P. B. (1997) A Respecification and Extension of the Delone and Mclean Model of Is Success, Information Systems Research, 8(3), 240-253.

Smith, H. J., Keil, M. and Depledge, G. (2001) Keeping Mum as the Project Goes Under: Toward an Explanatory Model, Journal of Management Information Systems, 18(2), 189-228.

Theodoratos, D., Ligoudistianos, S. and Sellis, T. (2001) View Selection for Designing the Global Data Warehouse, Data & Knowledge Engineering, 39(3), 219-40.

Theodoratos, D. and Sellis, T. (2000) Incremental Design of a Data Warehouse, Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies, 15(1), 7-27.

Wixom, B. H. and Watson, H. J. (2001) An Empirical Investigation of the Factors Affecting Data Warehousing Success, MIS Quarterly, 25(1), 17-41.

Yang, J., Karlapalem, K. and Li, Q. (1997) Algorithms for Materialized View Design in Data Warehousing Environment, In Proceedings of the Proc. 23rd International Conference on Very Large Databases, (Eds, Jarke, M., Carey, M., Dittrich, K. R., Lockovsky, F., Loucopoulos, P. and Jeusfeld, M. A.), Morgan Kaufmann Publishers, pp. 136-45.

Yang, J. and Widom, J. (2000) Temporal View Self-Maintenance, In Lecture Notes in Computer Science, Vol. 1777, (Eds, Zaniolo, C., Lockemann, P. C., Scholl, M. H. and Grust, T.) Springer Verlag, pp. 395-412.

Zhuge, Y., Garcia Molina, H., Hammer, J. and Widom, J. (1995) View Maintenance in a Warehousing Environment, SIGMOD Record, 24(2), 316-27.