DATA WAREHOUSING IN THE CONSTRUCTION INDUSTRY: ORGANIZING AND PROCESSING DATA FOR DECISION-MAKING

Data warehousing in construction

I. AHMAD and C. NUNOO
Department of Civil and Environmental Engineering, Florida International University, Miami, Florida, USA

Durability of Building Materials and Components 8. (1999) *Edited by M.A. Lacasse and D.J. Vanier*. Institute for Research in Construction, Ottawa ON, K1A 0R6, Canada, pp. 2395-2406.

Abstract

Construction organizations are critically dependent on data. But data must be available in suitable forms for use. Timely access to useful and meaningful information can enable construction companies gain competitive edge, increase client satisfaction, expand market share and enhance profitability. Vast amounts of construction operational data are scattered across multiple, dispersed and fragmented departments, units or project sites. In this paper, we present data warehousing as an emerging database management technology that can provide the resource for decision-making. We point out the difference between an operational database - used for transaction processing; and a data warehouse - intended to be used for analytic processing in management decision-making in the context of construction organizations.

Keywords: Database management, data warehousing, construction organizations, decision-making

1 Introduction

Effective management of construction projects and, for that matter, construction organizations is critically dependent on the effectiveness of the decision-making process used by the organizations. Most often, the process is informal and unstructured and sometimes arbitrary. Although, ideally, decisions should be based on data and information, in practice they seldom are. Conventionally, data are viewed as a resource that supports transaction processing in an organization. Transaction processing involves day-to-day operations of an organization. In a

typical construction contracting company, payroll processing, time card processing, materials management, estimating, scheduling, and routine activities like these are examples of transaction processing. The concept of normalization was developed to facilitate organization, maintenance and utilization of data for supporting transaction processing in organizations.

While normalized data are essential for supporting day-to-day transaction processing, they are not as useful to upper level decision-makers. Data should be organized in a different way in order to satisfy the need of the decision-makers. The recently introduced concept of data warehousing may provide a solution to the need.

2 Data warehousing fundamentals

Construction organizations generate and use vast amount of operational data that are distributed across various functional databases. For the most part, however, the fundamental value of these scattered data remains uncovered and unutilized. Decision-makers often have to wait days or weeks for responses from IT (information technology) or MIS (management information system) personnel that handle requested database queries in order to tap into such data. Such long waiting periods can cause irreversible damages to the reputation of a construction organization. Accessing the appropriate information from these databases at the time when it is needed, and in the appropriate format, is not an easy task. Data warehousing is presented in this paper as a concept that is meant for gaining timely access to these data in order to facilitate effective decision-making.

According to Inmon (1992a), data warehousing is a collection of decision-support technologies, aimed at enabling the "knowledge worker" (executive, manager, and analyst) to make better and faster decisions. The primary purpose of a data warehouse is to provide easy access to specially-prepared data that can be used with decision-support applications, such as management reporting, queries, decision-support systems, and executive information systems. Data warehousing is broad in scope, including:

- Extracting data from transactional systems, operational systems and other external sources;
- Cleansing, scrubbing, and preparing data for decision-support;
- Maintaining data in appropriate data stores;
- Accessing and analyzing data using a variety of end user tools; and
- Mining data for significant relationships.

Obviously, data warehousing involves technical, organizational, and financial considerations. Data warehousing technologies have already been successfully deployed in many industries. According to Adriaans and Zantinge (1996), data warehousing is being effectively-used in manufacturing for order shipment and customer support, in retailing for user profiling and inventory management, in financial services for claims analysis, risk analysis, credit card analysis, and fraud

detection, in transportation for fleet management, in telecommunications for call analysis and fraud detection, in utilities for power usage analysis, and in healthcare for outcome analysis

In this paper, the potential role of data warehousing in strategic decision-making with special reference to construction organizations is discussed. A review of the differences between operational database and that of the data warehouse, the general architecture of the data warehouse, and online analytic processing (OLAP) used to analyze data extracts from the data warehouse, as opposed to online transaction processing (OLTP), are presented. Finally, benefits that could potentially be derived from data warehousing implementation within a construction organization are presented.

3 Understanding data warehousing

In general terms a data warehouse is a database created by combining data from multiple databases for the purposes of analysis. A data warehouse is used solely for reporting purposes unlike the traditional data capture or online-transaction processing (OLTP) systems such as general ledger, accounts payable, financial management, order entry, or equipment inventory. A data warehouse is populated with data from two sources. The most frequent source is the periodic migration of data from OLTP systems. The second source is from externally purchased databases such as list of incomes and demographic information that can be linked to internal data. The warehouse gives management the ability to access and analyze information about its business. It allows users to utilize business and market information for directing corporate objectives. The data warehouse collects all of the data into one system, organizes the data so it is consistent and easy to read, keeps "old" data for historical analysis, and makes access to and use of data easy so that users can do the analysis themselves (Corey *et al*. 1998).

From the above general discussion, data warehousing can be defined as a *subject-oriented, integrated, non-volatile, time-variant* collection of data for supporting the process of management decision-making (Inmon 1992c). Understanding these key characteristics of data warehouse is crucial for the understanding of how it can be effectively used to facilitate efficient and effective decision-making

3.1 Subject-oriented data

A data warehouse is established around all the existing applications of the operational data of an organization and it gives information about a particular subject about an organization’s on-going operations (e.g., equipment, supplier, orders, shipments) instead of around applications (e.g., general ledger or payroll). Since the data warehouse is designed specifically for decision-support while an operational database contain information for day-to-day use, not all the information in the operational database is useful for a data warehouse.

3.2 Integrated data

In an operational data environment many types of information used in a variety of application may be found and different names for the same entity may exist. The data warehouse reconciles the different data representations in various operational databases used by the organization. To create a useful subject area, the source data must be integrated by modifying it to comply with common coding rules. The data warehouse gathers data into the warehouse from all these variety of sources and merges them into coherent whole. In this way they reflect the business information of the organization.

3.3 Time-variant data

Most business analyses are, in fact, analyses of trends. Trend analysis requires access to historical data. The data warehouse stores data that can be identified with particular time periods.

This implies that there must always be a connection between the information in the data warehouse and the time it was entered. The data warehouse therefore maintains historical data (typically three to five years of data) to enable the knowledge worker to do trend analysis over time. Operational systems on the other hand, are designed for immediate response and therefore data is often purged within a few months of its capture.

3.4 Non-volatile data

Data in the data warehouse are never updated, but only used for queries. It contains stable data, which are normally read-only environments that are enriched only on a nightly, or weekend basis. In this way such data can only be loaded from the operational database of the organization since that is the only database that can be updated, changed or deleted. Because the data are non-volatile in the data warehouse, management gets a consistent picture of the state of affairs of an organization’s business.

4 Decision-support versus transaction processing

The data warehouse is maintained separately from an organization's OLTP databases to reduce the impact that queries have on operational systems and to safeguard operational data from being changed or lost. It also allows database administrators to combine fields from different systems to create new, subject-oriented data that “end users” can access directly using powerful graphical query and reporting tools. Another reason for separating the data warehouse from an organization's OLTP databases are that the data warehouse supports on-line analytical processing (OLAP) which enables users to leverage the information stored in databases for sophisticated decision-support analysis. The functional and performance requirements of these systems are quite different from those of the OLTP applications traditionally supported by the operational databases (Zhuge *et al.* 1995). The operational database is clearly in the operational environment with data entering into it from unintegrated environment applications. Once the data have been

successfully populated into a well-designed data warehouse they can then be made available to all the management decision-support systems in the organization, such as management information systems (MIS), executive information system (EIS) and decision-support systems (DSS). With such major differences in the two environments, technology for supporting them should also be different (Adriaans and Zantinge 1996).

Since the data warehouse is designed specially for decision-support queries, only that data needed for decision-support are extracted from the operational database and stored in the data warehouse. Traditional OLTP systems, designed to handle day-to-day mission critical data and to have access to information through simple query operations, are very good at putting data into the databases quickly, safely and efficiently but not good at delivering meaningful analysis. Retrieval of facts for a typical *ad hoc* report takes too long to specify and convey to an IMS (information management systems) organization, with even a longer wait as a result. Operational systems, therefore, can not serve as repositories for facts and historical data for business analysis. In the world of transaction processing systems, more demands are made of information system than questions asked of the data. The occasional queries are typically limited to locating a particular record in the information system and preparing it for updates, or performing simple aggregations.

Decision-makers, on the other hand, generate a breed of questions unlike those geared to transaction processing. Such queries originate from the business need to analyze and process data in order to draw conclusions. These questions are usually complex and typically cover dimensions not relevant to OLTP systems, such as time period, product families or region of the world (Adriaans and Zantinge 1996). As a data warehouse is only applicable to a decision-support system environment, it is of little use to a firm that has a lack of integration in its operational environment. Although the transactional data provide the backbone for operational processing, they are fragmented since they were previously built separately from each other. An attempt can be made to integrate operational applications by producing data and process models that show the complete information flow requirements of the organization.

5 Types of data in the data warehouse

One can think of the data warehouse as a collection of historical transactions and summarization. It may also be thought of as a giant spreadsheet that sits on a big computer. The data warehouse holds different flavors of data. The following represents a common sampling of the type of data contained in the data warehouse (Corey *et al*. 1998).

- Transaction downloads from operational systems that are time-stamped to form historical records.
- Dimensional support data (client, suppliers, materials, equipment, human resources time, cost, techniques that worked etc.).

- Table to supports the joining of dimensional data and numeric facts relating to this data.
- Summarization of transactions (e.g., daily production by departments). These are in actual facts preemptive queries; the data are aggregated when it is added to the data warehouse rather than when a user requests for it.
- Miscellaneous coding data.
- Metadata that represents data about the data. This category might include sources of the warehouse data, replication rules, rollup categories and rules, availability of summarization, security and controls, purge criteria, logical and physical data mapping.
- Event data sourced from outside services, such as demographic information collated into the geographic areas in which a company operates.

6 Operational database versus data warehouse design

Operational system design and creating databases to serve operational purposes differ significantly from the goals of data warehouse design as we discussed in earlier sections. When creating a database for an online operational system in a relational model, the concern is with quick response time and efficient data storage and therefore the data model created must be in the third or higher normal form (Atre 1980). The process of normalization or arranging data in "normal" forms involves creating "relations" or tables in order to make the tasks of data maintenance (such as adding, deleting, updating and retrieving) less complex and more systematic. As data are moved from operational systems into the warehouse, they go through a process known as systematic denormalization that violates all the rules relational database architectures apply when modeling most systems. Systematic denormalization is carried out to enhance the performance of the warehouse by reducing *join* (an SQL operator) operations that are resource intensive. Compared to single table statements, *join* operations have the following characteristics (Corey *et al*.1998):

- Consume significantly more CPU;
- Require more temporary work space (on disk and in memory) for sorting;
- Require temporary tables for holding of intermediate results;
- Perform I/O since at least one I/O is required per table in the query

Ironically, when implemented, a data warehouse will assume expansive sizes, does not necessitate dynamic updates, and is used primarily for analytical rather than repetitive processes. Hence, normalization is not only futile but is in fact counterproductive towards factors like performance and ease of use when it comes to decision-support applications. The numerous *join* processes generated between normally distributed tables causes slow and cryptic query systems (SQL or Structured Query Language-based) with non-instinctive application logic for data warehouse users. The "Nested Relational Model" is presented as a possible solution to this problem by limiting database operations within predefined modules (Inmon 1992b).

7 Data architecture perspective of data warehousing

In this section, we will look at the components that constitute the data warehouse, a data architecture perspective and information flow process in a typical data-warehousing environment. The data warehouse database is a combination of many different components, including the following (Corey *et al*. 1998):

- Operational data source
- The staging area
- The data warehouse
- The subject data marts
- OLAP Server(s)
- Reporting tools
- Metadata repository
- Monitoring and administration of the warehouse.

A typical data warehouse architecture is shown in Fig. 1.

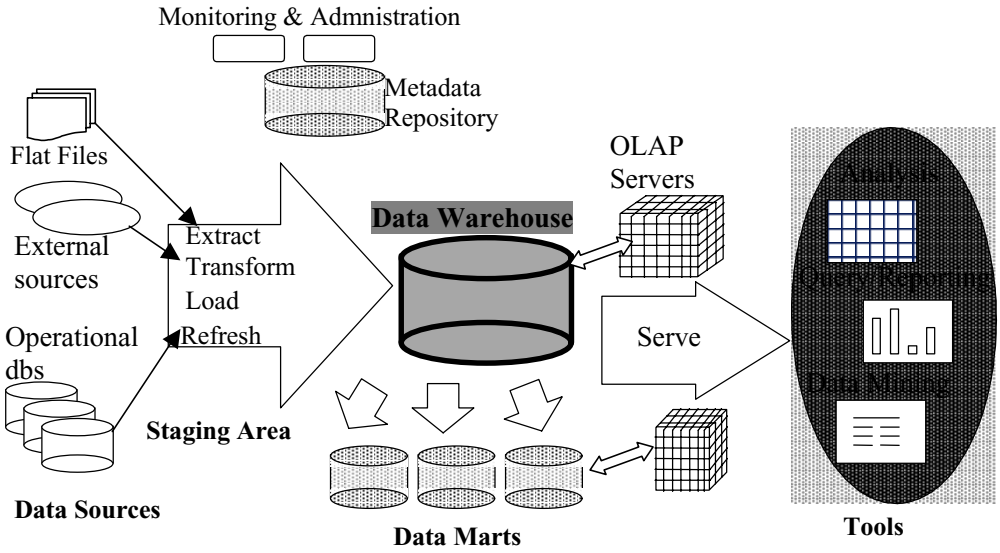

Fig. 1: A typical data warehousing environment architecture (Chaudhuri and Dayal 1996)

Thus an effective data warehouse should include tools for:

- extracting data from multiple operational databases and external sources;
- cleaning, transforming and integrating this data;
- loading data into the data warehouse; and
- periodically refreshing the warehouse to reflect updates at the sources and to purge data from the warehouse, perhaps onto slower archival storage.

The data are initially extracted from the source such as operational data or flat files, through the staging area, and then loaded into a data warehouse using various third party loaders such as "SQL* Loader" and data warehouse loading tools. The warehouse is then used to populate the various process-oriented data marts and OLAP servers, as illustrated in Fig. 1. The entire data warehouse then forms an integrated system that can serve the decision-maker reporting and analysis requirements of the user community (Chaudhuri and Dayal 1996).

The staging area is a set of database tables that will be used to receive data from the operational data source. This accelerates the manipulation of the data in the relational database in the operational environment and the loading into the data warehouse. The data structure in the data warehouse simplifies the enterprise's data while still retaining a non-process oriented database structure. In addition to the main warehouse, there may be several departmental data marts. Data in the warehouse and data marts are stored and managed by one or more warehouse servers (OLAP servers), which present multidimensional views of data to a variety of front end tools, namely, query tools, report writers, analysis tools, and data mining tools. Finally, there is a repository for storing and managing Metadata, and tools for monitoring and administering the warehousing system. (Inmon 1992a; Chaudhuri and Dayal 1997)

8 Data warehousing and construction information management

Normally the operational database is the central storage area of all the data needed to serve each business function of the entire organization. Today, many construction organizations are moving toward the new technology offered by client/server architecture to centralize their database at one location and decentralize their decision-making initiatives. In a typical construction management information system the database is essentially normalized to shrink the overall database size, thereby easing updates, and establishing program/data independence. As a result they are not able to support online analytic processing for strategic decision-making.

Incorporation of *historical, projected* and *derived* data is an important process in the implementation of data warehousing solutions in any construction organization. Historical data are most often compiled from existing operational data; projected information usually comes from spread sheets or external data feeds and derived data is computed by OLAP (online analytic processing) server's calculation engine. By integrating all three types of data, construction professionals will be able to support an operationally oriented decision-making process that looks forward as opposed to a system that reports only what has happened in the past. Support for historical data alone limits an operational system to only *what-happened* questions. Support for projected and derived data enables a system to additionally support *what-if* and *what-next* analyses. The ability of an OLAP server to deliver historical, projected and derived data all side-by-side creates forward-looking applications that provide significant value to end users that are otherwise impossible to deploy.

In a typical construction organization, the data will be gathered from vast amount of valuable information contained in operational sources such as accounting payroll, cost estimates, company and project finance, material inventory, equipment

inventory, human resource data, and contract data. These data will then be transformed into a single, integrated, subject-oriented database. Project engineers, architects, suppliers, construction managers and all personnel involved in a construction venture, can then query and analyze the information in the warehouse to support business decisions in ways that were not possible before. A typical construction data warehouse will contain a large volume of historical data that will support more strategic decision-support needs such as long-term trend analysis in project cost, project duration, tender quotations, and experience derived from past projects. Building managers, for example, can compare alternative sites, building types, design and construction schedules and possible types of tenants or buyers in a local real estate market and be able to project expenditure and income streams to enable them to develop a more precise discount cash flow.

9 Overall benefits of implementing data warehousing in a construction organization

The overall benefit of data warehousing in any construction organization, in the view of the authors, will be to serve as a distinct centralized repository for online transaction processing systems in the organization. This data may contain extracts of vital business data from a variety of corporate databases, which can be analyzed and used as a strategic, competitive weapon. Successful data warehouse implementation in a construction organization will increase the productivity of construction managers, project developers, and the organization as whole. The inherent flexibility of online analytic processing (OLAP) systems means business users of OLAP applications can become more self-sufficient (Roussopoulos 1995). Managers are no longer going to depend on information technologists to make schema changes to create *joins*. Perhaps more importantly, data warehouse capabilities will enable managers to model problems that would be impossible to model using less flexible systems with lengthy and inconsistent response times. More control and timely access to strategic information both equal more effective decision-making.

Considering the fact that the construction industry is rapidly becoming a global industry, it is expected that many construction organizations are going to commit their effort to achieve tremendous benefits by implementing data warehousing technology in their organizations. The reasons for this are not far fetched from the numerous advantages inherent in data warehousing technology discussed in this paper. These benefits can more than compensate for the investment made in the implementation of a workable data warehouse.

Construction organizations will be able to compete overcoming barriers of time and distance and learn from the past, adjust to the present, and position for the future. For optimally utilizing the concept of data warehousing, however, fundamental differences between conventional databases and data warehouse such as current vs. historical data, large volume vs. very large volume data, and mission critical vs. decision-support applications must be understood. Rather than immediately adopting Relational Database Management Systems (RDBMS) for transaction-based, as well as information-based applications; it is recommended that DBMS (Database

Management System) are selected for the former whereas DWMS (Data Warehouse Management System) are chosen for the latter (The Data Warehouse 1998).

10 Organizational issues

Data warehouse developmental efforts, including attempts for improvements and changes, mandates considerations of how data warehousing will be supported in the construction organization. There will be the need for specific decisions regarding who will maintain the database, networks, third party OLAP tools and programming as well as any other decentralized applications. Because the effectiveness of data warehouse depends on the willingness of users to share data, this coordination can be difficult. Traditionally, this type of responsibility rests with the data processing or information system (IS) entity within the organization. For success of data warehousing in a construction organization the administration function must be separated into a new unit that has authority across the entire organization. In any of this organizational structure, the enterprise nature of data warehousing requires that stakeholders be integrated into the functional system. A typical approach currently used in most organizations practicing data warehousing is to create a data warehouse coordinating group represented by all the various units of the organization (Kimball and Strehlo 1995).

Despite the convincing potential of data warehousing in construction information management, the decision to invest in this direction must be done with care since data warehousing is not always the most cost-effective or even the best solution for an organization. It may sometimes be advisable to start by building summarized reporting structure in the OLTP database. These structures can eventually be ported to the warehouse. Prior to any system development effort, it is also important to undertake careful financial analysis in terms of cost and benefits to determine the viability of the proposed investment in data warehousing. The costs are straightforward to estimate in a data warehousing environment as in any online transaction processing environment, however benefits are far less tangible because in most cases the exact target audience of the system is rarely known. In general, the decision to implement a data warehouse should be based on the following questions (Corey *et al*. 1998):

- Does it give us competitive advantage?
- Does it improve the bottom line?
- Will it deliver on all its promises?
- Will it deliver on time?
- What will be the risk if it is not implemented?
- What will be the risk if it is implemented?
- Will it deliver on budget?

It must be emphasized that while it is important to discuss the benefits of a planned warehouse, it is usually impossible to quantify these benefits in dollar terms. In many organizations that are currently practicing data warehousing technology, the decision to construct the warehouse is frequently an act of faith.

11 Conclusion

Most construction organizations have accumulated vast amounts of data over the years through the normal daily transaction processing activities. These databases have, in most cases, not been designed to store historical data or respond to queries except to support all the applications used in the organization. Data warehouse represents another type of database, solely designed to support strategic decision that can be set up in any construction organization. Data warehousing technology can enable construction companies to consolidate information from diverse operational systems into one source for consistent and reliable information. This will give construction project managers the opportunity to bring wisdom and insight into their decision-making process.

The premise of data warehousing, as presented in this paper, is to make large amount of information available to large community of end users in an organization. Presumably, the more users that are able to access data warehousing applications, the more value will be provided to the organization by the data warehouse. This implies that to maximize success, the online analytic processing (OLAP) server in the data warehouse must be made accessible to a wide variety of end-user tools (OLAP 1998). As a data warehouse stores sensitive business information in a centralized location, the importance of data security and user management cannot be overemphasized.

While many commercial products and services exist, there are still several interesting avenues for research in order to adopt the most appropriate technique and customization to suite the needs of individual construction organizations. By providing the ability to model real business problems and more efficient use of people resources, data warehousing can provide the means for construction organizations to respond more quickly to market demands. Market responsiveness in turn often will yield improved revenue and profitability. A good data warehouse should provide the right data to the right people at the right time (DISC 1997) – which is the main purpose of any organizational information system.

12 References

Adriaans, P. and Zantinge, D. (1996) *Data Mining*. Addison-Wesley Longman Limited, Reading, Massachusetts.

Atre, S. (1980) *Data Base: Structured Techniques for Design, Performance*, and Management. John Wiley, New York.

Chaudhuri S. and Dayal, U. (1997) *An Overview of Data Warehousing and OLAP Technology.* Technical Report MSR-TR-97-14 Microsoft Research, Advanced Technology Division Microsoft Corporation: Redmond, WA.

Corey M., Abbey M. and Abramson I. (1998) *Oracle 8 Data Warehousing-A practical Guide to Successful Data Warehouse Analysis.* ORACLE Press Edition. Berkeley, California: Osborne/MacGraw-Hill.

DISC (1997) Dynamic Information Systems Corporation (DISC) White Paper, *Bringing Performance to Your Data Warehouse.* (http://sun2.dic.com/dwhper.html).

Inmon W.H. (1992a) *Building the Data Warehouse.* John Wiley and Sons, New York, N.Y.

Inmon W.H. (1992b) *Rdb/VMS: Developing the Data Warehouse*. John Wiley and Sons, New York, N.Y.

Inmon W.H. (1992c) *Using the Data Warehouse*. John Wiley and Sons, New York, NY.

Kimball R. and Strehlo (1995) Why decision support fails and how to fix it. reprinted in *Sigmod Record*, 24(3).

Kimball, R. (1996) *The Data Warehouse Toolkit*. John Wiley and Sons, New York, NY.

OLAP (1998) OLAP in Data Warehousing: *An Arbor Software White Paper,* [online] Available: www.Arborsoft.com/cgi-bin/rbox/wh…?t=711604194818&f=WHOLAPDW&p=olapdw.html. (6/15/98).

Rezgui, Y., G. Cooper and P. Brandon, (1998) Information Management In a Collaborative Multiactor Environment: The COMMIT Approach. *Journal of Computing in Civil Engineering: 136-142.*

Roussopoulos, N., Chungmin, M. C. and Kelley, S. (1995) The ADMS Project: Views "R" Us. *IEEE Data Eng. Bulletin*, 18(2), June.

The Data Warehouse: Achieving Better Decisions Faster, (1998) *Database and Network Journal*: An International Journal of Databases and Network Practice, 28(3) p. 3-6.

Veshosky, D. (1998) Managing Innovation in Engineering and Construction Firms. *Journal of management in engineering*, January/February, 58-66.

Zhuge, Y., Garcia-Molina, H. Hammer, J. and Widom, J. (1995) View Maintenance in a Warehousing Environment. *Proceedings of the ACM SIGMOD International Conference on Management of Data, San Jose, California, June.*