First International Conference on Construction in the 21st Century (CITC2002)
"Challenges and Opportunities in Management and Technology"
25-26 April, 2002, Miami, Florida, USA

DATA WAREHOUSING IN CONSTRUCTION: FROM CONCEPTION TO APPLICATION

Irtishad Ahmad
Associate Professor, Department of Civil and Envionmental Engineering
Florida International University, Miami, Florida, USA

Salman Azhar
Doctoral Candidate, Department of Civil and Envionmental Engineering
Florida International University, Miami, Florida, USA

ABSTRACT
Data warehousing has emerged as an effective mechanism for converting data into useful information. It is an improved approach to integrate data from multiple, often very large, distributed, heterogenous databases and other information sources. This paper examines the possibility of using data warehousing technique in the construction industry to integrate various functional and operational databases which are usually scattered across multiple, dispersed and fragmented departments, units or project sites. The concept of data warehousing is explained with the help of several examples and an insight is provided to the reference architecture of a generic data warehouse. At the end, an example on the design of a data warehouse for building developers is provided.

KEYWORDS
Data Warehouse, Information Management, Information Systems, Databases, Knowledge-based Systems

1. INTRODUCTION

Data warehousing concept is mainly used to develop decision support systems (DSS). According to Inmon W.H. (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. Many organizations are committing considerable human, technical, and financial resources to building and using data warehouses. The primary purpose of these efforts 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 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. It obviously involves technical, organizational, and financial considerations (Nunoo and Ahmad, 1999).

Data warehousing is one of the most important recent developments in the filed of information system (IS). Conceptually, a data warehouse is a database that collects and stores data from multiple remote and heterogeneous information sources. The data warehouse approach presents some advantages over the traditional on-demand approach (Theodoratos and Sellis, 1999):

- The queries can be answered locally without accessing the original information sources. Thus, high query performance can be obtained for complex aggregation queries that are needed for in-depth analysis, decision support and data mining – a way of extracting relevant data from a vast database.

- On-line Analytical Processing (OLAP) is decoupled (separated) as much as possible from On-line Transaction Processing (OLTP). Thus making information accessible to decision makers avoiding interference of OLAP with local processing at the operational sources.

Data warehousing technologies have already been successfully deployed in many industries. For example, in manufacturing for order shipment and customer support, in retail 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 outcomes analysis (Adriaans and Zantinge, 1996).

The concept of data warehousing is discussed in this paper with special reference to the construction organizations. A review of the differences between operational database and that of the data warehouse, general architecture of the data warehouse, online analytic processing (OLAP) used to analyze data extracts from the data warehouse, as opposed to online transaction processing (OLTP) are presented. Finally, an example is presented to illustrate the design of a data warehouse for a building developer.

2. DATA WAREHOUSING CONCEPT

Traditional view of database management is based on the need for data to support transaction processing. Data warehousing evolved as an answer to the need to support analytic processing for better decision-making. 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.

In general terms a data warehouse is a database created by combining data from multiple databases for purposes of analysis. A data warehouse is generally populated with data from two sources. The most frequent source is the periodic migration of data from Online Transaction Processing (OLTP) systems. The second source is from externally purchased databases (such as lists of incomes and demographic information, and in the context of construction, material pricing data) that can be linked to internal data. Thus, for a construction project management, a data warehouse can enable decision-makers, such as the project managers, to utilize project and market information for meeting client needs. A 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 it themselves (Corey *et al*. 1998).

The data warehouse should be maintained separately from an organization's transaction processing databases to reduce the impact that queries have on operational systems and 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 is that the data warehouse supports on-line analytic processing (OLAP) which enables users to leverage the information stored in databases for sophisticated decision support analysis.

Since the data warehouse is designed specially for decision support queries, only data that is needed for decision support should be extracted from the operational database and stored in the data warehouse. Decision-makers generate a breed of questions unlike those geared to transaction processing. Such queries originate from the needs to analyze and process data in order to draw conclusions. These questions are usually complex and typically cover dimensions not relevant to OLTP systems.

To summarize, 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.

2.1 Data Warehouse versus Operational Database

Operational system design and creating databases to serve operational purposes differ significantly from the goals of data warehouse design. The difference is explained in Table 1.

Table 1: Data Warehouse Vs. Operational Database

Data Warehouse	Operational Database
• Uses historical data • Data is updated in planned periodic cycles • Normalization of data into too many tables in not necessary.	• Uses current data to support day-to-day transactions • Data is updated frequently and input in real time • Normalization is necessary

2.2 Types of Data in the Data Warehouse

The data warehouse holds different flavors of data. The following represents a common sampling of the types of data contained in the data warehouse (Corey *et al.* 1998).

- *Fact data* contains the physical information that describes a factual event, e.g. contract price, material quantities, labor hours etc.
- *Dimension data* contains information about the facts used to analyze the transactions, e.g. geographic location, time, project participants.
- *Transaction downloads* from operational systems that are time-stamped to form historical record.
- *Metadata* which 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 control, 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.

The fact data and dimensional data form the foundations of a data warehouse. Figure 1 explains the relation between fact data and dimensional data.

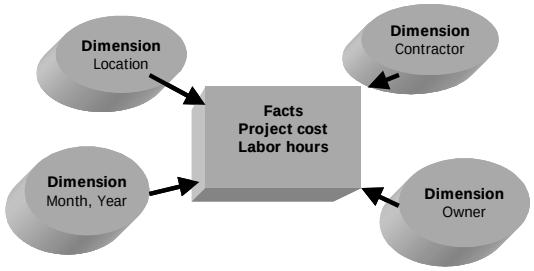

Figure 1: Relation between Fact Data and Dimension Data

2.3 Process Flow within a Data Warehouse

The process flow within a data warehouses performs the following operations, which are also highlighted in Fig. 2.

- Extract and load the data from multiple databases and sources into the data warehouse.

- Clean and transform data into a form that can cope with large data volumes, and provide good query performance. It further back-up and archive the data.
- Manage queries, and direct them to the appropriate data sources.

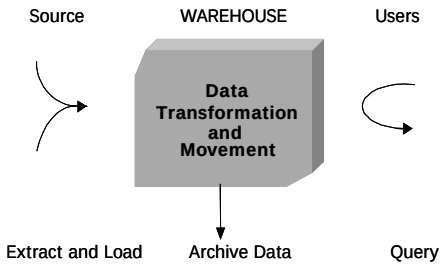

Figure 2: Process Flow within a Data Warehouse (adopted from Anahory et al, 1997)

Extract and load data
All operations necessary to support the extract and load process are performed by the "Load Manager" (LM). LM may be a combination of *off-the-shelf* tools and *shell scripts*. It extracts and loads data performing simple transformations before and during load and extracts file structure into temporary data store and then to warehouse structure, e.g. loading of payroll data.

Transform data and backup
Data transformation and backup is done by the "Warehouse Manager" (WM), which performs all the operations necessary to support the warehouse management process such as transformation and management of data, backup and archives of data warehouse, and transformation of data from temporary data store through schemas to summary tables. For example, transforms payroll data into productivity data.

Query management
Query Manager (QM) performs all the operations necessary to support the query management process. Its main functions are to: direct queries to the appropriate tables, and schedule the execution of user queries. For example, QM provides answers to questions such as, "at which location carpenters were more productive during a given time-period?"

3. GENERIC DATA WAREHOUSE ARCHITECTURE

The generic data warehouse architecture is illustrated in Fig. 3. It includes tools for extracting data from multiple operational databases and external sources; for cleaning, transforming and integrating this data; for loading data into the data warehouse; and for periodically refreshing the warehouse to reflect updates at the sources and to purge data from the warehouse, perhaps onto slower archival storage. The data is 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. The entire data warehouse then forms an integrated system that can serve the decision-maker reporting and analysis requirements of the user community (Inmon, 1992b).

Figure 3: A Generic Data Warehouse Architecture

It can be seen from the figure that data sources include existing organizational files in combination with external and operational databases. These data must be loaded into a data warehouse for online analytic processing in order to produce various reports, support queries, perform analysis and extract information.

4. DATA WAREHOUSE DESIGN

The data warehouse design essentially consists of four steps, which are as follows:

1. Identifying facts and dimensions
2. Designing fact tables
3. Designing dimension tables
4. Designing database schemas

4.1 Identifying Facts and Dimensions

The facts are analyzed by, or though different dimensions. For example, actual costs and budgeted costs are facts and can be analyzed by such dimensions as projects, line items, trades, and time period.

4.2 Designing Fact Tables

Fact tables contain the quantitative or factual data. Once the facts are identified, they can be considered as different *Entities*. An entity is a class of persons, places, objects, events, or concepts for which we need to capture and store data. Then the next step is to identify the *attributes* associated with each Entity. An attribute is a descriptive property or characteristic of an entity. This is explained in Figure 4.

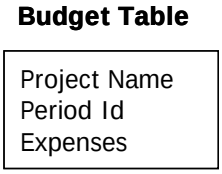

Figure 4: Example of a Fact Table

4.3 Designing Dimension Tables

Dimension tables, sometimes called minor tables hold descriptive data that reflects the dimensions of a business. The dimension tables are designed in the same way as fact tables by identifying the entities and their attributes. Once the fact and dimension tables are designed, the association between them can be identified as shown in Figure 5.

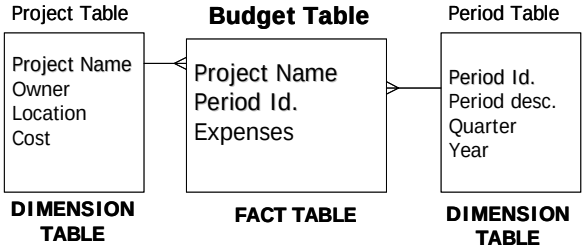

Figure 5: Example of Association between Fact Table and Dimension Table (Star Schema)

4.4 Designing Database Schemas

The schema is a database design containing the logic and showing relationships between all existing tables. There are three main types of database schemas: the *Star Schema*, the *Snowflake Schema* and the *Starflake schema.*

Star schema
Poe et al. (1998) defined Star schema as a simple structure with relatively few tables and well-defined join paths. This database design, in contrast to the normalized structure used for transaction processing databases, provides fast query response time and a simple schema.

Star schemas are physical database structures that store the factual data in the "center" surrounded by the reference (dimension) data. It can be very effective to treat fact data as primarily read-only data, and reference data as data that will change over a period of time (Anahory et al., 1997). The Star schema uses denormalization to provide fast response times, allow database optimizers to work with simpler database design in order to yield better execution plans. A star schema is denormalized in the sense that dimension tables are not broken down into normalized tables thus providing familiar end-user views. Star schemas exploit the fact that the content of factual transactions is unlikely to change, regardless of how it is analyzed. The schema shown in Fig. 5 is a simple Star schema.

Snowflake schema
The snowflake schema is a variation of the star structure, in which all dimensional information is stored in third normal form, while keeping fact table structures the same (Poe et al., 1998). This implies dividing the dimension tables into more tables, thus avoiding non-key attributes to be dependent on each other. Figure 6 shows an example of a Snowflake schema.

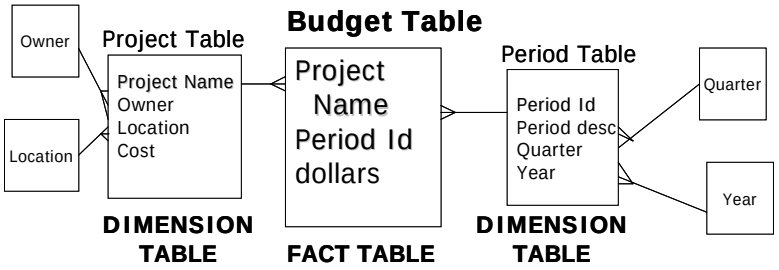

Figure 6: Snowflake Schema

Starflake schema
In decision support data warehouses, the most appropriate database schemas are combinations of denormalized Star and normalized Snowflake schemas, referred to as a *Starflake* schema, which is illustrated in Fig. 7.

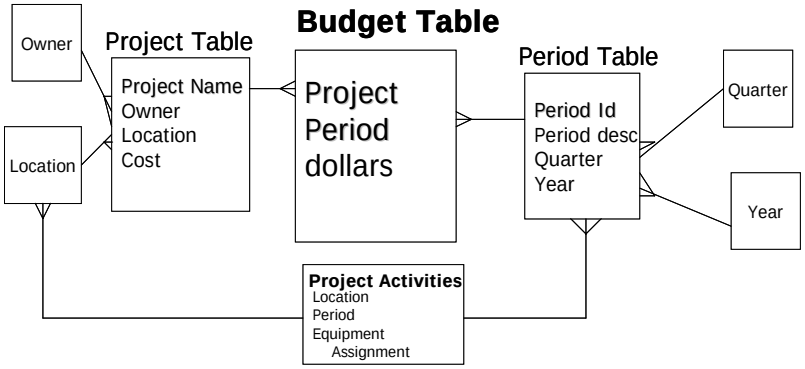

Figure 7: Starflake Schema

5. DATA WAREHOUSING IN CONSTRUCTION

In a typical construction management information system the database is designed for transactions. The transactional database is normalized to maintain referential integrity whenever a record is inserted, deleted, or updated. As a result they are not suitable for supporting online analytic processing for strategic decision-making. In

order to take advantage of the data warehousing concept construction organizations need to consider redesigning their database architectures.

Data warehousing has a tremendous potential in construction. Data Warehousing will facilitate a greater degree of coordination and will promote greater interaction and responsiveness in the process of construction. Data Warehousing will reduce the need for information processing and will eliminate unnecessary paperwork and bureaucracy. Trust and confidence among the participants will be emphasized. Partnering type arrangements will be effectively used with the support of data warehousing. Decisions will be made where they need to be made. Data warehousing will increase the capacity of information processing. With data warehousing more information can be processed in a shorter period of time. Thus decisions will be made quickly, on a timely basis, and with appropriate information.

6. EXAMPLE OF DATA WAREHOUSE DESIGN FOR A CONSTRUCTION PROJECT

This section illustrates an example of the design of a data warehouse for building developers to compare the alternative sites, building types, design and construction schedules and possible types of buyers. This example is adopted from Lukauskis (1999). Such a data warehouse will enable them to develop better cash flow projections.

6.1 Extraction and Loading Process

The data extraction includes "Land Parcel for Sale" data and location-specific spatial data. The data can be inputted in the data warehouse using pre-selected software tools and/or by using programming languages.

6.2 Transformation and Backup Process

The data can be organized using a Star (fact-dimension), or a Snowflake or a Starflake schema as shown in Figure 8. The best scheme may be selected after performing an economic and feasibility survey.

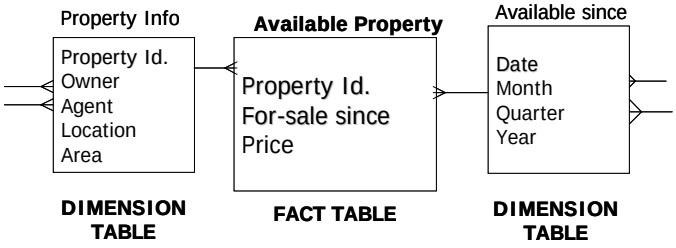

Figure 8: Star Schema for the Proposed Data Warehouse

6.3 Query Processing

The query processing must be able to perform spatial analysis on the geocoded sites (e.g. sites within 2 miles from hospitals/colleges, sites within water and sewer service areas, sites within $25k-$50k average income areas) and provide a list of all available sites with a certain targeted return on investment.

6.4 Data Warehouse Architecture

Figure 9 illustrates the architecture of the proposed data warehouse.

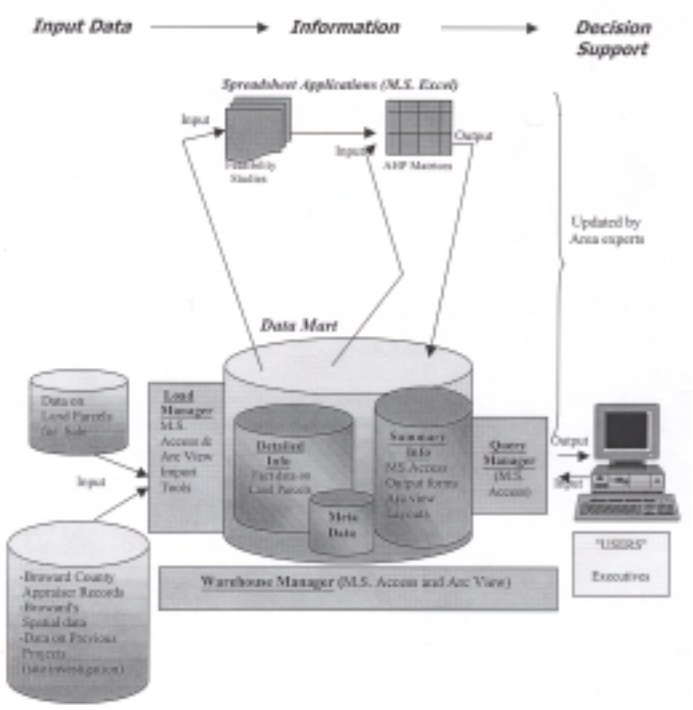

Figure 9: Reference Architecture of the Proposed Data Warehouse

7. CONCLUDING REMARKS

Data warehousing supports online analytic processing for strategic decision-making. With successful data warehouse implementation in construction organizations, managers will be less dependent on IT professionals, and will be using data more effectively. Data warehousing can enable construction companies to consolidate information from diverse operational systems into one source for consistent and reliable information. However, Data warehouse developmental efforts must be supported by the organization. Decision to invest in data warehousing must be made carefully - it is not always the most cost-effective option, it may be advisable first to build summarized reporting structure using operational data before making investing in developing a full-blown database. These structures can eventually be ported to the data warehouse of the future. Most often benefits are long-term and far less tangible than the costs.

8. REFERENCES

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

Ahmad, I. and Nunoo, C. (1999). "Data warehousing in the construction industry: Organizing and processing data for decision-making". *8th International Conference on Durability of Building Material and Components*, Institute for Research in Construction, Vancouver, British Columbia, May-June 1999.

Anahory, S., and Murray, D. (1997). *Data Warehousing in the Real World: A Practical Guide for Building Decision Support Systems*, Addison Wesley, Massachusetts.

Corey, M., Abbey, M.; and Abramsom, I. (1998). *Oracle 8 Data Warehousing-A practical Guide to Successful Data Warehouse Analysis.* ORACLE Press.

Inmon W.H. (1992a). *Building the Data Warehouse.* John Wiley.

Inmon W.H. (1992b). *Rdb/VMS: Developing the Data Warehouse*. John Wiley.

Lukauskis, P. (1999). *A Decision Support System Using Data Warehousing and GIS to Assist Builders/Developers in Site Selection*, M.S. Thesis, Florida International University, Miami, Florida, USA.

Poe, V.; Klaver, P.; and Brobst, S. (1998). *Building a Data Warehouse for Decision Support*, Prentice Hall, Upper Saddle River, New Jersey.

Theodoratos, D., and Sellis, T. (1999). "Design data warehouse". *Data and Knowledge Engineering*, Vol. 31, pp. 279-301.