Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2001 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2001

An Assessment of Company Data Warehousing Practices

Mark Hwang
Central Michigan University

James Cappel
Central Michigan University

AN ASSESSMENT OF COMPANY DATA WAREHOUSING PRACTICES

Mark Hwang
Central Michigan University
mark.hwang@cmich.edu

James J. Cappel
Central Michigan University
james.cappel@cmich.edu

Abstract

A survey of Fortune 1000 companies was conducted to assess current data warehousing (DW) practices. The survey focused on a range of issues that involve the development, use, maintenance, management, and assessment of data warehousing. For most measures, the findings of this study compare closely to those of a study by Watson et al. (2001). Similar findings were obtained for the age of data warehouses, their uses, benefits, and the use of multiple architectures for their implementation. Respondents in both studies also gave their DWs an overall favorable rating. This study produced somewhat different findings than Watson et al. and other sources for the median and average costs of DWs. In addition, this study considered a number of issues that were not addressed in the Watson et al. study such as: DW development time, DW size, and DW maintenance issues such as refresh strategies, data quality approaches, and the amount of IS staff needed for maintenance. The findings of this study, along with those of the Watson et al. study, provide important insight about DW issues and challenges for companies who are considering beginning or expanding their DW operations.

Introduction and Background

Data warehousing (DW) may be defined as "the establishment and maintenance of a large data storage facility containing data on all (or at least many) aspects of the enterprise" (Martin et al. 1999). A data warehouse is established by extracting data from source systems such as operational systems, cleaning and transforming the data, and loading it in the warehouse where it is then made available to decision makers (Watson et al. 2001). In recent years, data warehousing has grown in importance as DW tools have emerged and organizations have sought better ways to integrate their data to improve management decision making. Today, most large organizations have either built a data warehouse or are seriously thinking about developing one (Watson and Haley 1998).

Despite the growing popularity of data warehousing, there have been relatively few attempts to study companies' practices with respect to establishing and maintaining data warehouses. One exception to this was a recent study by Watson et al. (2001) that surveyed the data warehousing practices in member organizations of The Data Warehousing Institute.

The present study considers a somewhat broader sample of organizations: Fortune 1000 companies. These companies might be expected to have significant data warehousing operations. Larger organizations have a greater need to integrate diverse data sources. They also have more resources to establish DW operations, which are reported to be significant. According to Gagnon (2000), the average cost of developing a data warehouse is $2.2 million. Thus, the present study serves an important complement to the Watson et al. study by providing an improved understanding of company data warehousing practices, which are at a relatively early stage of development and still evolving.

The authors were commissioned to undertake this study by a major health care provider in the Midwest. The sponsoring organization was interested in learning more about industry experiences and best practices with data warehousing, as they looked to expand their own DW operations. The authors investigated some of the same factors as the Watson et al. study, while also considering other issues, particularly with regard to systems maintenance.

Methodology

The purpose of this study was to assess the data warehousing practices of Fortune 1000 companies. A mailing list of these companies was obtained from *Fortune*. To this list, the names of Chief Information Officers or senior manager with responsibility for data warehousing operations were added (where they could be located) in an effort to increase the response rate.

An initial survey was pilot tested by data warehousing professionals in the sponsoring company and other organizations. The survey was revised based on the input from these professionals. The final survey was mailed to the Fortune 1000 companies in September 2000, and a follow-up (reminder) mailing was sent in October 2000. Almost fifty responses were received. Some responding companies indicated that they did not have a data warehouse, they did not intend to build one, or they were currently in the process of building a data warehouse. A few companies responded that they do not participate in surveys due to time constraints or their unwillingness to reveal proprietary information. In the end, a total of 27 usable responses were obtained.

Sample Characteristics

The participants represented a variety of different industries including: beverages, building materials, chemicals, computer and data services, cosmetics, financial, outsourcing services, electronics, health care, gaming, insurance, petroleum, publishing, transportation, retailing and wholesale, utilities, and waste management. The number of employees in the participating companies ranged from 2,500 to more than 88,000; the average organization size was 21,216 employees. The average annual revenue of participating companies was $4.7 billion, with a minimum of $1.2 billion and a maximum of $21.6 billion.

Results

The results of the survey are presented in terms of four general categories: DW development, scope, and architecture; DW use; DW maintenance issues; and DW assessment.

DW Development, Scope, and Architecture. The average DW development time was reported to be 19.5 months. The longest development time was eight years, while the shortest was just one and a half months. It should be noted that a variety of different data warehousing architectures were considered in this study. The study included both data warehouses and data marts, which are smaller in scale and typically hold data that pertains to a subset of the company's business rather than being enterprise-wide like data warehouses (Stair and Reynolds 2001). This would appear to account for the large variability in development time.

The most expensive data warehouse took $30 million to develop while the least costly was $130,000. The median cost was $2 million. This figure was somewhat higher than the median cost of $1.5 million found by Watson et al. (2001). The mean cost of a DW found in our study of $5 million is also considerably higher than the average cost of $2.2 million reported by Gagnon (2000). The average cost of DWs differed by the type of architecture used. The respective averages were: $8.3 million for a data warehouse only, $5.8 million for a data mart only, $4.5 million for a data warehouse and independent data mart, $3.3 million for a data warehouse and dependent data mart.

The oldest data warehouse had been in existence for eight years while the youngest was only four months old. The average age of data warehouses among responding companies was 2.5 years, which compares closely to that age of 2.6 years found by Watson et al. As these results suggest, the data warehousing efforts at some companies are rather well established while in other firms they are still getting started.

The data warehouses also varied widely in size. Among respondents, the largest data warehouse was 4.5 terabytes whereas the smallest was 9 gigabytes. The average size was 400 gigabytes. Of course, the size of data warehouses can get much larger than this; Walmart's massive data warehouse is more than 200 terabytes. In the current survey, companies were also asked about the projected growth of their data warehouse in terms of data volume and the number of users. The average projected annual growth was reported to be 50% in data volume and 35% in the number of users.

The architecture of the data warehousing operations also varied widely among participants. As shown in Figure 1, the most common architecture used by companies (39%) was data warehouses with dependent data marts. Twenty three percent of companies had a data warehouse with independent data marts, and another 23% had data marts only. An additional 15% had a data warehouse only. These findings are consistent with the Watson et al. (2001) study, which also observed a substantial variation in DW architectures.

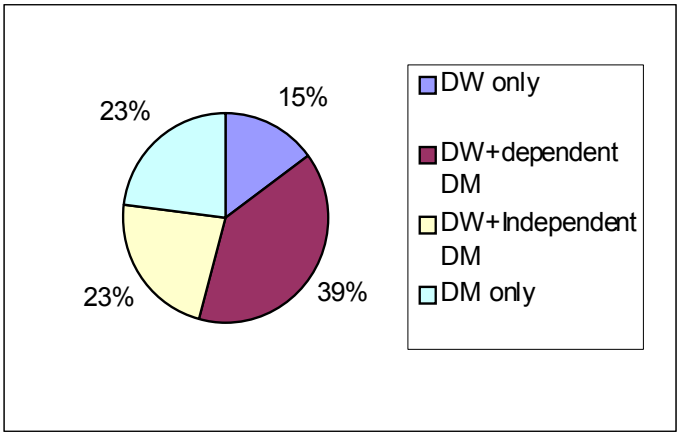

Figure 1. DW Architecture

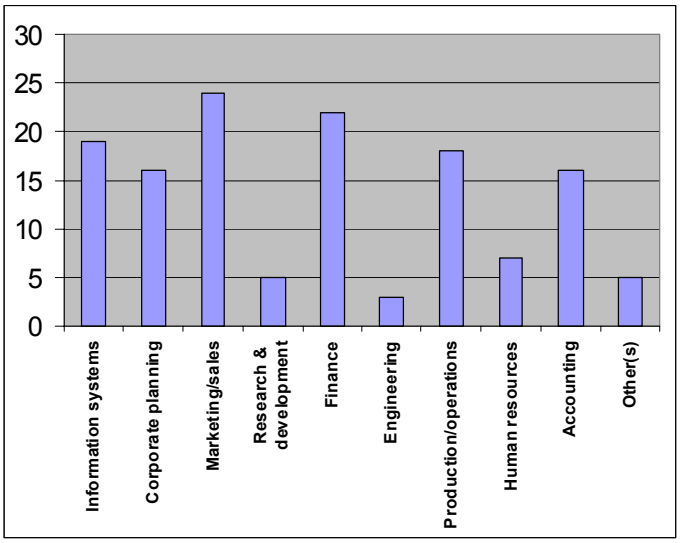

Figure 2. DW Use by Functional Areas

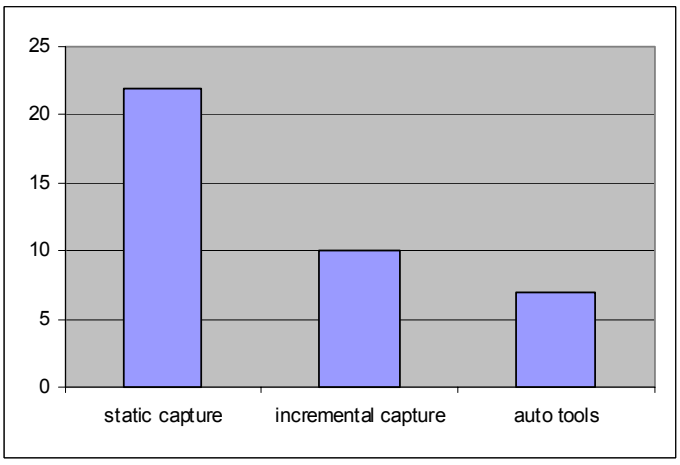

Figure 3. DW Refresh Strategies

DW Use. Data warehouses were found to be used by many functional areas within the organization. As shown in Figure 2, the leading functional users of data warehouses were: marketing/sales, finance, information systems, production/ operations, and accounting. These same functions finished as the top five functional users of data warehouses in the Watson et al. (2001) study. The largest DW system in the current study had 4,000 users while the smallest had only four users. The average number of users was 443. Of course, not all users access the data warehouse at the same time. The largest number of concurrent users was estimated to be 500, while the smallest was four; the average was 65.

DW Maintenance Issues. One important consideration is how many information systems professionals it takes to maintain the data warehouse. The results indicated that this number varied widely. The maximum number of employees involved with DW maintenance was 30, while the least was 1.5; on average, DW maintenance required 7 IS professionals.

A second DW maintenance issue is what strategies companies use to refresh the data in their warehouses. There are generally two types of DW refresh strategies: static update and incremental update (Bokun and Taglienti 1998). Static update

is a time-based approach where a snapshot of the source data is taken at fixed time intervals. The snapshots are then used to either reload or append the DW. Incremental update is a change-based approach where every change (insert, update, or delete) to the source data is captured and applied to the DW. Static capture is cheaper to implement since it can be scheduled in a way to minimize interferences with the operational systems that feed into the DW. On the other hand, changes that occur between capture cycles are lost and therefore are not available for analysis in the DW. Incremental or change-based update ensures that every change to the source data is captured and applied to the DW. It allows richer analysis than does static capture but is also more expensive to implement as it creates additional work load on the operational systems.

Figure 3 shows the number of companies that use different refresh strategies. As evident from these numbers, some companies use more than one type of strategy. The most commonly employed strategy is static capture. Most companies that used this strategy reported doing it on a daily basis. The second most common refresh strategy is incremental capture. This method may involve using an application, a trigger, a transaction, or a combination of these for the incremental capture. Most commonly, the incremental capture is done using applications. Finally, a fewer number of companies use automatic tools to refresh data, such as Maestro Scheduler or SQL server DTS.

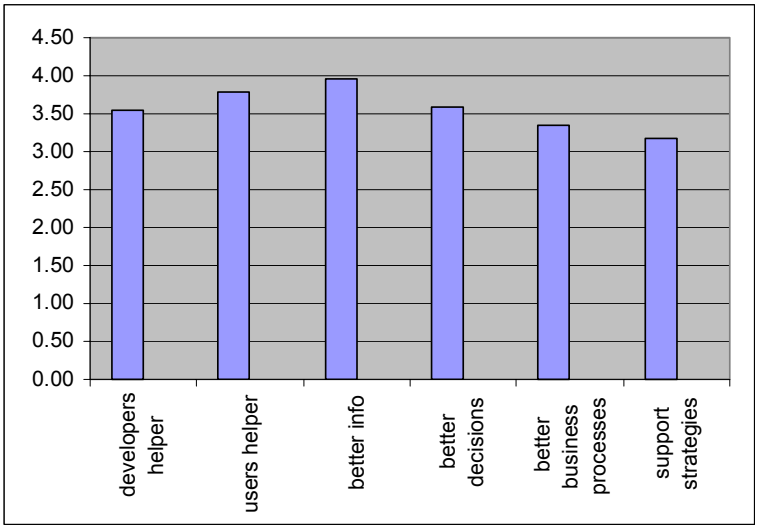

Figure 4. DW Benefits

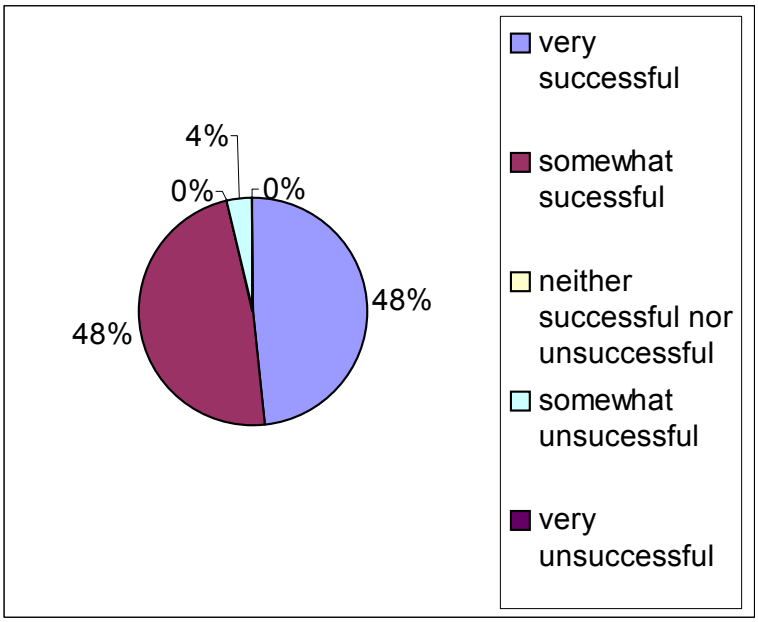

Figure 5. Overall Evaluation of DW

Another DW maintenance challenge is how to handle "dirty" data, i.e. inaccurate or inconsistent data. When "dirty" data were found, 27% of respondents reported that they always rebuilt the data warehouse, 27% said they usually rebuilt it, and 19% said they sometimes rebuilt it. A fewer number of respondents reported they seldom or never rebuilt the data warehouse (15% and 12%, respectively). Thus, for most companies, data quality is an important issue in a data warehouse. In fact, data cleansing is reported to be one of the most demanding, time consuming aspects of data warehousing (Atre 1998).

DW Benefits. Respondents were asked to assign a number from 1 (for minimum benefit) to 5 (for maximum benefit) to six major types of DW benefits that were identified by Watson et al. (2001). The average rating for these benefits is shown in Figure 4. As indicated in this figure, respondents rated their data warehouses as beneficial in all six categories; all six categories had an average rating of more than 3.0 on the five-point scale. The most highly rated benefit was better information with a mean rating of nearly 4.0, while the lowest perceived benefits were improving business processes and supporting the accomplishment of strategic objectives. This same pattern of results was observed by Watson et al. (2001).

Respondents were also asked to give an overall evaluation of their data warehouses. Consistent with the high rating on benefits, the overall ratings were quite positive (see Figure 5). Approximately 48% of respondents rated their data warehouse as very successful while an additional 48% said it was somewhat successful. Similarly, the Watson et al. (2001) study found generally positive results on this measure. In their study, only 15% of the companies described their data warehousing as a "runaway success" but an additional 67% reported it as being moderately successful.

Participants were asked to cite a number of factors that contributed to the success of their data warehouses. The most frequently mentioned factors (in descending order) were clearly defined business needs/benefits, good data quality, adequate IS staff and consultants, and strong business sponsorship. Other respondents reported factors that impeded the success of their data warehouse operations. The biggest culprit was the lack of clearly defined business needs/benefits.

Conclusions

In a number of companies, data warehousing is still at a relatively early stage of development. Consequently, many organizations are interested in learning more about the state-of-the-art DW practices of other enterprises. This study, as well as the study by Watson et al. (2001), provides important insight about these issues.

It should be noted that the sample in the current study was relatively small, so the results may not be representative of companies' DW practices in general. Still, they do provide a detailed look at the DW practices of large organizations, and where similar measures were taken, they confirm a number of the findings of the Watson et al. (2001) study.

As new DW tools continue to emerge and organizations move further ahead on the experience curve with respect to data warehouses, there is a need for ongoing research to identify company best practices. Surveys such as the one presented here need to be undertaken periodically to ensure that companies are using data warehousing technology for maximum effectiveness.

References

Atre, Shaku. "Rules for Data Cleansing," *Computerworld*, March 9, 1998, pp. 69-72.

Bokun, Michele and Taglienti, Carmen. "Incremental Data Warehouse Updates." *DM Review*, May 1998, Available at www.dmreview.com, Accessed May 7, 2001.

Gagnon, Gabrielle. "Data Warehousing: An Overview," *PC Magazine*, March 19, 1999, pp. 245-246.

Stair, Ralph M. and Reynolds, George W. *Principles of Information Systems*. 5th ed., Boston, MA: Course Technology, 2001.

Martin, E. W., Brown, C., DeHayes, D., Hoffer, J. and Perkins, W. *Managing InformationTechnology: What Mangers Need to Know*. 3rd ed., Upper Saddle River, NJ: Prentice Hall, 1999.

Watson, H., Annino, D., Wixom, B., Avery, K. and Rutherford, M. "Current Practices in Data Warehousing," *Information Systems Management,* (18.1), Winter 2001, pp. 47-55.

Watson, H., and Haley, B. "Managerial Considerations," *Communications of the ACM*, (41.9), September 1998, pp. 32-37.