DATA WAREHOUSE

A data Warehouse (*DW*) refers to a storehouse of business information gathered from production databases and multiple other sources. It consolidates and stores historical data primarily from operational databases integrated along common business dimensions. A DW is different from a database in many ways. A database is a part of a DW. Additionally, a DW uses external, historical, and operational data. A database uses operational data. Superior data analysis tools that provide quick response to queries are parts of a DW. A database usually does not come with these kinds of tools. A DW stresses the need for a single source of consistent corporate-level data. A database may provide only operational-level data. Easy-to-use applications help in accessing the data. Data warehousing, the term first coined by Inmon, has been defined as the process whereby organizations extract value from their informational assets by using special stores called data Warehouses or DWs (1, 2). While DWs took a backseat to enterprise resource planning (ERP) during the late 1990's, they have become more in vogue lately as it became apparent that the ERP systems were not well suited for analytic processing (3). In fact, a 2005 Forrester research report indicated that 30% of all information processing organizations were pursuing an active DW strategy (either operating or designing one) (1).

DATA WAREHOUSE USE

DW has several uses. First, DW produces versatile reports and graphs from consolidated data from a range of transaction systems. Thus, provisions for ad hoc requests, regular and more information-rich reliable reports, and exception reporting or alerts, add value to a firm. Second, DW supports dimensional analysis, which is a simplified way of looking at data summaries across a number of dimensions or attributes of data. DW greatly aids in answering the why's instead of what's. Third, DW is the enabler of a new technology called *data mining* which recognizes patterns in data and thus helps to predict future behavior based on the current characteristics of data. Finally, present DWs do more than mere forecasting and decision support. Large DWs with close ties to transactional systems provide an integrated business intelligence (BI) platform for firms (1).

DATA WAREHOUSE TYPES

Many types of DW can exist. A DW can be classified in many different ways: on the basis of its size and level [DW (large) versus DM (small)], underlying database support [nonvirtual (separate) versus virtual (same)], or underlying networking support [web based versus non-web based] and so on. A firm-level or enterprise DW is usually very large and provides all relevant information about a number of subjects. Some firms design *data marts* (*DM*s) or subsets of the large DW that better suit the needs of specific end-user groups. DMs themselves are DWs that focus on particular needs and are typically easier to build. DMs require less money and design time [typically thousands of dollar budget for months to finish compared with a traditional DW's millions of dollars for years of time to complete (3)]. Sometimes firms design DMs first to gain the benefits from an earlier stage. However, these DMs must be designed so as to integrate into a coherent firm-level DW. Such design is not an easy task to accomplish. DWs can also be *virtual*, that is, with no separate physical Warehouse; the data resides only in the production systems.

The evolving of relational on-line transaction processing tools (*OLTP*) may obviate the need to maintain a DW separate from the core relational database management system (*RDBMS*). The DWs can also be classified as *web-based*, because they use web browsers for consistent cross-platform interfaces rather than *LAN*-based, which use client programs for each group/PC.

DATA WAREHOUSE COMPONENTS

A DW has the following components (3):

1. Data modeling tools
2. Meta data repository, which describe the DW
3. Data transport tools, which access the source data
4. Data extraction/scrubbing (cleaning of data)/ normalization tools that transform the source data
5. Core database or the Warehouse data store that provides rapid access to data
6. Middleware connectivity tools for managing the DW environment
7. End-user data access tools for retrieving, formatting, and analyzing the data

The source data are accessed by the data transport tools from an operational environment consisting of legacy (existing mainframe-based systems), OLTP, and outside sources. The data are transformed (by extraction, scrubbing, and normalization tools) and are stored in a relational database form. The DW/DBMS repository is usually a server. Top such repositories in year 2006 are NCR Teradata, IBM system p5, z,i,x, HP Integrity, HP 9000, HP Proliant, Sun Fire U/SPARC IV+, Sun Fire x86, Unisys ES7000, Bull NovaScale, Netezza, to name a few (4). Sometimes the term ETL (extract, transport and load) is used to describe these tools. There are many ETL tools available. Some of the leading tools are Oracle Warehouse Builder, SAS ETL Studio, Business Objects-Data Integrator and Informatica Powercenter. Refer to (5) for details. Metadata repository is also part of DW (containing summaries of data). Sometimes DW data are transferred to several DMs, which are subsets of this DW. DMs can be at the departmental level of a firm. Then the DW or the DM data are transferred to end-user data access tools by middleware connectivity software. Finally, the end-user data are analyzed by data mining, on-line analytical processing (*OLAP*) tool, and other applications. Refer to (6) for a comprehensive analysis of OLAP tools.

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 2007 John Wiley & Sons, Inc.

DATA WAREHOUSE PRODUCTS

DWs are typically suitable for environments where the data are dirty or massive and require heavy analysis in real time. DWs have been extensively used in retail, health care, high tech and banking/financial applications. Since its inception, many DW products have been announced. IBM, SAS, Oracle, SAP are examples of some DW vendors that sell these products. In recent times open source platforms for DW as provided by IBM, Oracle or Hewlett-Packard are proving to be more attractive.

DW has been gaining popularity in recent years. In 1982, the Proquest ABI/Inform, a CD-ROM database produced by UMI hardly mentioned DWs. The number of article abstracts that mentioned DWs increased to more than 600 in 1996 (7). A number of experts and firms have advocated substituting DWs for RDBMS in the recent past.

MAJOR FACTORS IN FAVOR OF DATA WAREHOUSE ADOPTION AND USAGE

The recent interest in DWs has resulted from the development in distributed processing/networking technology and the need for improved decision making tools.

Other factors like lowering the cost of overall information access, improving customer responsiveness, identifying hidden business opportunities, conducting precise marketing, and mass customization have been cited that boosted the interest in DWs (8).

Several studies reported real benefits from DW implementation during early 1990s (9–14). Direct profit increase by cost reduction has been reported. After implementing a DW solution, Britain's Woolworth Co. experienced a 35% increase in profit in one year. CNA Insurance Co. created a DW for analyzing single-sourced customer and historical data. This allowed it to save money by eliminating the time and cost of checking and reconciling separate sources of data. Tora Co., a lawn mower and snow-blower firm, cut 20% off its billing cycle and thus reduced operational costs by adopting a DW solution.

Improved productivity in terms of better response time and better performance was observed in many cases. For example, by implementing a DW, Butt Grocery Co., a grocery-store chain, improved the speed of DW query significantly. Reno-air increased its weekly and monthly reports to daily ones. HCIA Inc., a health-care information industry, obtained a 90% faster response time in answering customer queries. Bank of America found that a large DW solution provides access and reduces elapsed time of a query from 2 hours to 3 minutes. An insurance brokerage firm, Sedgewick James, found that data downloading time is reduced from 2 days to a few seconds by using DW. Cole Taylor Bank reduced work hours from 5 hours a day to 40 minutes per week.

Improved business opportunities were also observed. Subaru implemented a DW system to improve its system quality by eliminating some of the cumbersome methods used in the old mainframe-based system. More complicated market-planning queries are one of the first benefits from the system. Victoria's Secret stores found from a trial set of queries that its system of allocating merchandise to its 678 shops, based on a mathematical storage average, could be improved to provide more profit.

To improve the chances of success in the implementation of a DW, Solomon lists 11 factors that need to be considered. Probably the most fundamental of them is the necessity, prior to the design of a DW, of obtaining an agreement on service level and data refresh requirements (26, p 27).

The following are other factors that have a positive impact on the user community:

1. Creation of unified and improved data. Data and information in an enterprise are usually heterogeneous and are generally randomly disseminated via spreadsheets, PC databases, XML, message queues, and functional/departmental applications, thereby making it difficult to comprehend the overall data. For other companies traditional legacy systems (for example, general ledger) can no longer supply the business with all the required reporting and analysis.
2. Creation of new values from existing resources
3. Maximization of the value of an investment—companies have invested heavily in DBMS. Firms want to maximize the value of that investment.
4. Ease of access and use: the advantage of one system access; a single access hides complexities and inconsistencies of production systems; access several different systems to obtain information on a particular subject.
5. Performance Improvements. Faster daily updates and low cost of processors with better power as well as innovations in data base design are favoring a DW design.
6. Handling of growing data size. It is common to find the volume of enterprise data in tera bytes. Wal-mart, for example, is approaching petabyte mark at the end of 2006. Many firms are doubling their data size every 12-18 months. DWs are needed in such situations to improve operating efficiency and reduce costs. In recent times a few vendors like IBM, Oracle, NCR and Netezza occupy the high-end of DW products, handling such massive data.

The importance of these factors was further confirmed in a survey of European DW firms. The Druid survey found data availability, new market opportunities, better understanding of customer base, improved competitiveness and a single and consistent version of data among the most cited gains from using DW (8).

The process of adopting DW by firms is next discussed. The important factors of DW adoption, as discussed previously, can therefore be categorized, as causal influences. Mostly nonintegrated products from a few leading vendors are used as necessary elements for DW adoption. Some of the important outcomes of the DW adoption process are increased productivity, better strategy, new values, better market exploitation, and better organizational structures for adopting firms.

These perceived benefits led several firms to implement DW projects.

Additionally, it has been observed that many present users think that data quality and end-user support are key factors that affect DW usage (15). Research has identified that using a full DW resulted in a significantly better decision performance that a partial DW design (16). Also it has been observed that non-financial performance is better with a DW solution (17).

DRAWBACKS IN DATA WAREHOUSE ADOPTION

The statistics cited to date together with the level of coverage given by the media may not be entirely correct because the extent of actual DW adoption (percentage of companies who have actually implemented DW projects) is not known. DW also has its drawbacks. A number of factors, economical, technical and organizational, may prevent firms from implementing DW projects. The typical high cost, effort (collecting data from often nonintegrated sources) and lengthy developmental time are quite prohibitive, especially to firms with a low information systems (*IS*) budget. Initial support for funding and effort can also prove difficult because this usually is an enterprise wide effort. The advantages of having a DW over an existing OLTP do not always convince upper management. Management does not always appreciate a new way of accessing data because employees need to acquire new knowledge for such data accessing, abandoning the traditional and known way (for example, using the Excel spreadsheet). The Druid report mentions several main inhibitors of DW implementation, like intangibility of benefits and a lack of corporate perspective, user understanding and sponsorship (8). Finally security is emerging as an important issue for Web-based DWs.

Others cite problems like changing from a transaction processing to a DW mindset, 80% user time spent in extracting, cleaning, and loading data, hidden systems problems feeding the DW, the possibility of more and not fewer written reports, disk space wastage due to overhead, etc. (18). One also needs to have an appropriate combination of technical capabilities (database, LAN/WAN, a server, related tools) that require compatibility of various information technology (*IT*) products. Managing large databases of typically 100 gigabytes is not an easy task. Often simple tasks, such as periodic updates and backups, become difficult. The database design and maintenance itself is a challenging task because of changing requirements. A few other studies discuss DWs in terms of *limited* benefits only. Finally, maintenance of a DW is not an easy task—it is costly as well as a complex issue that needs to be addressed adequately before an implementation.

MAJOR SHARE OF IMPLEMENTATION: DM OR DW?

For economic reasons, many firms may decide to follow a bottom-up approach in such Warehouse design by starting with a DM solution and proceeding gradually toward an enterprise wide DW solution. Others may opt for a complete DW solution from the beginning, using a top-down design approach. The DM solution has recently drawn the attention of some experts. Oracle also offers packaged DM solutions. In a recent report, Forrester Research Inc., warned the IS community against going for big, centrally managed DWs. Instead they advocate adopting a DM solution (19). Inmon, on the other hand, thinks that the long-term effect of trying to build DMs without the DW is disastrous (20).

To answer this question, archival data were collected and analyzed. The data were collected from abstracts of articles from all IS magazines, starting from 1985, as contained in the CD-ROM based database ABI/INFORM produced by UMI (7). The search command used was "Data Warehouse OR Data Mart." The abstracts were manually checked for DW implementation stories. The sample thus obtained is adequate for the results to be generalizable. An implementation abstract might contain information on the implementation date, vendor product, type of implementation, client firm name and client firm type. The article abstracts were surveyed for an eight-year period starting from November 1988 and ending in April 1997. DW became commercially available in the mid-1980s 1, p.14). The selected data set consists of 174 abstracts. From this, 1997 data were separated out for testing the forecasting ability of the best models. This reduced the total number of implementations reported in such abstracts and available for main analysis to 154. Only articles that reviewed DW project implementation were counted. Each such abstract was coded for date of appearance, client firm name, geographical place, type of industry to which the client firm belonged, type of solution (DM/DW), and type of vendor product. This kind of data gathering is fairly common in IS research. The reader may refer to Ref. 17 for such examples.

The data set revealed that some implementation articles did not differentiate between DW and DM solutions. An overwhelming number of articles refer only to DW solutions whereas a handful of articles refer to implementing the DM solution (a total number of 15 out of 174). Some earlier articles may not have been able to differentiate between a DW and a DM because of their similarities. This could also be largely due to the fact that DMs are only now being deployed, with the advent of recent products. The earliest reference to a DM implementation dates back only to November 1994. Because only a few articles refer to the DM as a DW implementation, the majority of the implementations can be categorized as DW solutions to the problem. However, some think that DW and DM solutions are inherently different. The volume of data, level of data integration, and greater flexibility in use of data, easy expansion, data reconciliation and amount of historical data may separate DW from DM solutions.

A FEW KEY MAJOR PLAYERS?

The market of many IT products or paradigms is sometimes captured by a few major players. In contrast, a few IT segments exist (particularly true for new IT products/paradigms) that have not been monopolized by a handful of firms (18). It is interesting to compare the adoption situation of DW at an earlier stage with other IT technologies. Because DW is an emerging market one may expect more than a few vendors as key market players.

According to an early stage listing, the vendors in this field numbered more than 80 (22). However, 66% of the market was controlled by 6% to 8% of the DW vendor firms. A firm with a market share in excess of 41% dominates a market (leader), whereas a firm with a market share of 26% is considered a player in that market (23). Thus this set of firms provided leadership in the market, even in early stages. There is no indication yet of any individual leader or player in this segment of the IT market, and so the market can be called unstable. The majority of the implementations reported were being shared by IBM, Oracle, Sybase, SAS, NCR, and Red Brick (was an independent unit during the time of study; later on merged with the IBM). These vendors cover DW products in construction (extract, cleansing tools), operation (storage and management tools), and access segments (information catalog, query tools, data mining, reports).

INTEGRATED PRODUCTS OR A SET OF PRODUCTS?

Because data warehousing implementation deals with a set of technologies (*WAN*/*LAN*, RDBMS, application tools, large servers, Internet/Web), product integration is a serious matter. Many products in the DW construction, operation, and access areas have appeared. SAS, Oracle, Datallegro, Netezza, IBM, NCR have all reportedly offered end-to-end DW solutions. These solutions claim to include everything from server software to development and access tools to professional consulting services. User firms may prefer to buy an integrated solution or may do the integration themselves. Inconsistency, nonscalability, noncustomizability, and lack of support have been cited as common reasons for users to be careful in selecting such integrated solutions.

The vendors released the integrated products recently. So in earlier stages, the implementation scenario was clearly dominated by a heterogeneous set of products from many vendors. Typically, a firm selected a database and then started buying OLAP and other tools and the hardware from different vendors that can help the firm run the Warehouse efficiently.

WHICH INDUSTRY SEGMENT IS PROMINENT?

In the private sector, it has been claimed that a few industry segments, such as banking and retail, have the lion's share of DW implementations. Compared with these sectors, other sectors have lagged. This is true for many IT products. For example, Electronic Data Interchange (*EDI*) was adopted earlier in the transport industry. Networks like BITNET were adopted earlier in academia.

Figure 1 shows the analysis by industry based on the data. The banking, finance, insurance, computer/communications, and retail industries are the major implementers of DW. These constitute about 11 industrial segments of 81 Standard and Poor's listed industries at the same level. Some reasons for DW adoption by these industry segments have been stated earlier. Large banks seek a full-fledged DW implementation, whereas small banks (mortgage and community banks) are also trying to enter the market, often, with a DM solution. Some experts think that insurance companies traditionally have been slow to occupy the front end of a technology curve. However, rapid changes in both the competitive and regulatory environments have compelled even smaller companies to look closely at what DWs can do for them in terms of performance analysis and reporting, product development potential, and the ability to anticipate customer needs. It is expected that computer and communication firms will adopt recent IT products like giant DWs, arguably at a more rapid pace than other segments of the industry.

STAGES OF GROWTH AND GROWTH FORECASTING

It may be mentioned that DW growth cycle of a firm has three major stages—initiation (initial stage), growth (expansion stage) and maturity (fully integrated and operational). At the initiation stage, for example, users are mostly analysts, ordinary users lack experience, applications are simple and the use is mostly operational and tactical. In maturity stage, users throughout the organization, suppliers and customers with necessary computer skills use the DW, applications are more sophisticated and well-integrated and use is additionally strategic. At this stage also, benefits are realized (24).

Historically also, DWs can be thought of as passed through several stages of growth. From databases to DW, followed by real time and integrated solutions. Better and newer models of DW are emerging. However, growth forecasting can vary at different stages of historical growth, due to a number of issues.

In order to study the historical growth of DWs in earlier years, the forecast model of DW through the year 1997–1998 was developed. Standard forecast and diffusion models can be used for this purpose (25). The standard extrapolation methods used were moving averages and exponential smoothing and the diffusion model was the mixed influence model. The models were used to analyze the ex ante 1997 data. Table 1 contains the actual values of DW cumulative data for 1997. The prediction was found to be quite accurate. The maximum forecast error estimates resulting from these models are 11 to 14%. DW growth is predicted to be robust in the future. Thus mixed influence model can be used to predict DW growth more reliably in future, assuming the nature of growth in later stages does not differ significantly from the earlier years.

FUTURE TREND

The development trends of DW seem to be many sided: a movement toward open source data bases, increased web orientation, very large DWs consisting of terabytes, faster response generation by using parallel machines (some vendors, notably NCR's Teradata DBMS, have created truly parallel systems while others, Microsoft and Oracle for example, use a parallel query/conventional processing approach (28)), intelligent pattern matching with neural networks, optical-disk based cheaper and superior storage capacity, better integration of all its components; and incorporating DWs into a firm's overall service-oriented archi-

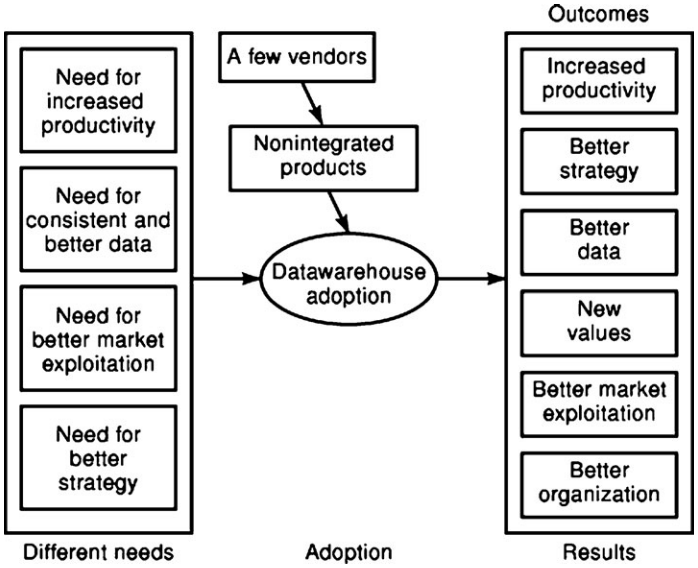

Figure 1. Share of major industry segments involved in data Warehouse implementation.

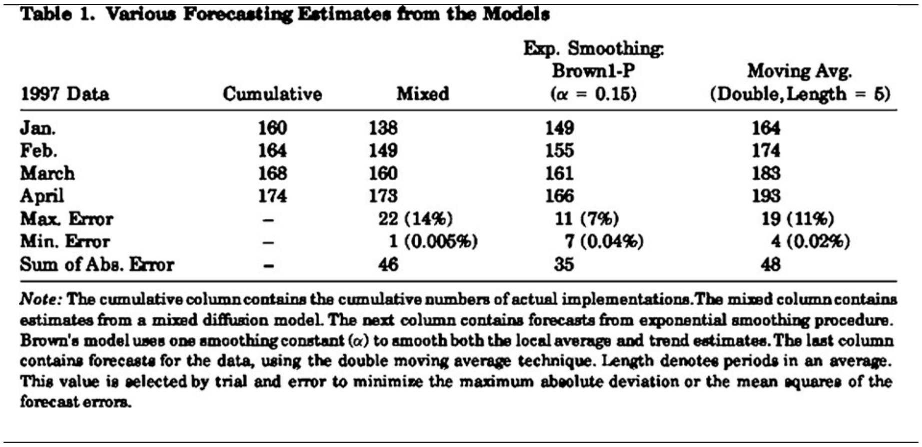

Table 1. Various Forecasting Estimates from the Models

1997 Data	Cumulative	Mixed	Exp. Smoothing: Brown1-P (α = 0.15)	Moving Avg. (Double, Length = 5)
Jan.	160	138	149	164
Feb.	164	149	155	174
March	168	160	161	183
April	174	173	166	193
Max. Error	–	22 (14%)	11 (7%)	19 (11%)
Min. Error	–	1 (0.005%)	7 (0.04%)	4 (0.02%)
Sum of Abs. Error	–	46	35	48

Note: The cumulative column contains the cumulative numbers of actual implementations.The mixed column contains estimates from a mixed diffusion model. The next column contains forecasts from exponential smoothing procedure. Brown's model uses one smoothing constant (α) to smooth both the local average and trend estimates. The last column contains forecasts for the data, using the double moving average technique. Length denotes periods in an average. This value is selected by trial and error to minimize the maximum absolute deviation or the mean squares of the forecast errors.

tecture to create an integrated BI. A move toward rapid deployment (less than 100 days) and lower total project costs can also be seen. The future application fields are environments where a single source of consistent data, quick response to management queries, data timeliness, and better access to data by the work force are needed. In such situations, a firm can gain strategic advantage by adoption of DW. Some of these application areas have been outlined above such as retail-centric ones. Apart from traditional business areas, newer areas such as bioinformatics are emerging which can widen the application base of DWs.

BIBLIOGRAPHY

1. L. Agosta The Data Strategy Advisor: A Time of Growth for Data Warehousing, *DM Review Magazine*, November 2004.
2. R. Barquin H. Edelstein (eds.) *Planning and Designing the DW*, Upper Saddle River, NJ: Prentice-Hall, 1996.
3. E. Turban E. McLean J. Wetherbee *IT for Management*, New York: Wiley, 1998.
4. D. Feinberg, J. Hardcastle, A. Butler and P. Dawson, Magic Quadrant for Data Warehouse DBMS Servers, *Gartner Research Report*, 2006.
5. ETL Tool Survey, 2004, 2005 and 2006,. http://www.etltool.com/etltoolsranking.htm
6. The OLAP Report. http://www.olapreport.com/ProductsIndex.htm
7. ABI/Inform data base, 1998.
8. The Druid Report, *Data Warehouse Network*, Ireland, 1997.
9. K. Laudon J. Laudon *Management Information Systems*, Upper Saddle River, NJ: Prentice-Hall, 1996.
10. A. Radding The 16,000% ROI, *Software Mag.*,**16**(7): S15–S18, 1996.
11. A. Reinbach Data Warehouse solution tackles competitive issue, *Bank Syst. Technol.*,**33**(10): 30, 1996.
12. P. Karon Subaru aims to take the guesswork out of consumer trends, *InfoWorld*, **18**(4): 63, 1996.
13. M. Goldberg J. Vijayan Data Warehouse gains, *Computerworld*, **30**(15): 1–16, 1996.
14. T. Studt Scientific data miners make use of all the tools available, *R&D*, **39**(5): 62C–62D, 1997.
15. S. G. Hong, P. Katerattanakul, S. K. Hong and Q. Cao Usage and perceived impact of dof data Warehouses: A study in Korean financial companies, *International Journal of Information Technology and Decision Making*, **5**(2): 297-315 June 2006.

16. Y. T. Park, An empirical investigation of the effects of data warehousing on decision per performance, *Information and Management*, **43**(1): 51–61 January 2006.
17. S. M. Lee, P. Katerattanakul, S. Hong, Impact of data warehousing on organizational performance of retailing firms, *International Journal of Information Technology and Decision Making***3**(1): 61–79 March 2004.
18. L. Greenfield Don't let data warehousing gotchas getcha, *Datamation*, **21**(11): 76–77, 1996.
19. A. E. Forrester Good news, bad news on software front, *Computing Canada*, 38–39, May 24, 1995.
20. W. Inmon Does your data mart vendor care about your architecture? *Datamation*, **43**(3): 105–107, 1997.
21. P. Todd J. D. McKeen R. B. Gallupe The Evolution of IS job skills: A content analysis of IS job advertisements from 1970 to 1990, *MIS Quarterly*, **19**(1): March, 1995.
22. C. Darling How to integrate data Warehouse, *Datamation*, **42**(10): 40–51, May 15, 1996.
23. T. Lewis Computer business or monopoly, *IEEE Computer*, **29**(1): 10–13, 1996.
24. H. Watson, T. Ariyachandra and R. Matyska Data warehousing stages of growth, *Information Systems Management*, 42–50, Summer 2001.
25. J. Martino Technological Forecasting for Decision Making, New York: North Holland, 1983.
26. D. Sammon, F. Adam, Towards a model of organizational prerequisites for enterprise-wide systems integration: Examining ERP and data warehousing, *Journal of Enterprise Information Management*, Bradford,**18**(4): 458–471, 2005.
27. M. Solomon, Ensuring a Successful Data Warehouse Initiative, *Information Systems Management*, Boston,**22**(1): 26–37, Winter 2005.
28. A. Sen and A. Sinha, A Comparison of Data Warehousing Methodologies, *Communications of the ACM*, **48**(3): 79–84, March 2005.

KALLOL K. BAGCHI
ROBERT CERVENY
PEETER KIRS
University of Texas, El Paso, TX
Florida Atlantic University,
Boca Raton, FL