BUSINESS INTELLIGENCE: FROM DATA COLLECTION TO DATA MINING AND ANALYSIS

Data for EC organizations can be viewed as either transactional or analytical. Transactional data are those pieces of information that are collected in traditional transactions processing systems (TPSs), are organized mainly in a hierarchical structure, and are centrally processed. Newer systems that contain transactional data are usually Web based; in medium to large organizations, they may be part of an ERP system. These are known as operational systems, and the results of the processing are mainly summaries and reports (see Turban et al. 2005).

Today, the most successful companies are those that can respond quickly and flexibly to market changes and opportunities (i.e., they are agile). The key to this response is the effective and efficient use of data and information. EC transactions frequently must be done online in real time. This is done not only via transaction processing, but also through the supplementary activity of analytical processing, which involves analysis of accumulated data, mainly by end users. Analytical processing includes Web applications, market research, data mining, CRM activities, and decision support systems. Placing strategic information in the hands of decision makers aids productivity, empowers users to make better decisions, and improves customer service, leading to greater competitive advantage.

COLLECTING, ORGANIZING, AND STORING DATA FOR ANALYTICAL PROCESSING

Analytical processing basically can be done in two ways. One is to work directly with the operational systems (the “let’s use what we have” approach), using software tools and components known as front-end tools and middleware. This option can be optimal for companies that do not have a large number of end users running queries and conducting analyses against the operating systems. Since the mid-1990s, a wave of front-end tools that allow end users to conduct queries and report on data stored in operational databases have become available. The problem with this approach, however, is that the tools are effective only with end users who have a medium- to high-degree of knowledge about databases.

These limitations call for a second, improved option of analytical processing, which involves three concepts:

1. A business representation of data for end users
2. A user-friendly, Web-based environment that gives the customers and corporate employees query and reporting capabilities
3. A single, server-based data repository—a data warehouse (DW)—that allows centralized analysis, security, and control over the data

DATA WAREHOUSES

The purpose of a data warehouse is to establish a repository that makes operational data accessible in a form readily acceptable for analytical processing activities, such as EC applications, decision support, and other end-user applications. As part of this accessibility, detail-level operational data must be transformed into a relational form, which makes them more amenable to analytical processing. Thus, data warehousing is not a concept by itself, but is interrelated with data access, retrieval, analysis, and visualization (see Gray and Watson 1998).

The process of building and using a data warehouse is shown in Exhibit W4A.1. The organization’s data are stored in operational systems (left side of the figure). Not all data are transferred to the data warehouse, and frequently only a summary of the data is transferred in a process of extraction, transformation, and load (ETL). The data that are transferred are organized within the warehouse as a relational database so that it is easy for end users to access. Also, the data are organized by subject, such as by product, customer segment, or business partner. EC data also may be organized according to a business process, such as ordering, shipping, or available inventory. The data then can be optionally replicated in data marts (explained later). Data access is provided through Web browsers via middleware software. On the right side of the figure are various applications that may use the data.

The activities conducted during much of the process described in Exhibit W4A.1 are generally referred to as business intelligence. The major reason for the name is that these activities not only collect data warehouse (DW)
A single, server-based data repository that allows centralized analysis, security, and control over data.

business intelligence
Activities that not only collect and process data, but also make possible analysis that results in useful—intelligent—solutions to business problems.
and process data, they also enable analysis that results in useful—intelligent—solutions to business problems. The concept of business intelligence originated from executive information system (EIS) activities, but today it is used to describe online analytical processing and data mining activities as well.

Data warehouses provide for the storage of metadata, which are data about data. Metadata include software programs about data, rules for organizing data, and data summaries that are easy to index and search, especially with Web tools.

**Characteristics of Data Warehousing**

The major characteristics of data warehousing include the following:

- **Organization.** Data are organized by detailed subject (e.g., by customer, vendor, product, price level, and region) and only contain information relevant for decision support.

- **Consistency.** Data in different operational databases may be encoded differently. For example, gender data may be encoded 0 and 1 in one operational system and “m” and “f” in another. They will be coded in a consistent manner within each warehouse.

- **Time variant.** The data are kept for 5 to 10 years so that they can be used for trends, forecasting, and comparisons over time.

- **Nonvolatile.** Once entered into the warehouse, data are not updated. However, new, related data may replace or supplement old data.

- **Relational.** The data warehouse typically uses a relational structure (organized into tables of rows and columns).

**Benefits of Data Warehouses**

The major benefits of data warehouses are (1) the ability of users to reach data quickly, because data are located in one place and organized properly, and (2) the ability to reach data easily, frequently by end users themselves, using Web browsers. Another benefit is that a data warehouse provides a consolidated view of corporate data, which is better than providing many smaller (and differently formatted) views. For exam-

---

**metadata**

Data about data, including software programs about data, rules for organizing data, and data summaries.

---

EXHIBIT W4A.1  **Data Warehouse Framework and Views**

*Source: Introduction to Information Technology, 2nd Edition, by Turban, Efraim, R. Kelly Rainer, Jr., and Richard E. Potter. © 2002 by John Wiley & Sons, Inc. This material is used by permission of John Wiley & Sons, Inc.*
ple, separate production systems may track sales and coupon mailings. Combining data from these different systems may yield insights into the cost efficiency of coupon sales promotions that would not be immediately evident from the output data of either system alone. Integrated within a data warehouse, however, such information can be easily extracted.

Data warehouses allow information processing to be off-loaded from expensive operational systems onto low-cost servers (or processed by application service providers, ASPs). Once this is done, end-user tools can handle a significant number of end-user information requests. Furthermore, some operational system reporting requirements can be moved to Web-based decision support systems, thus freeing up production processing.

In addition, accessibility to data warehouse content by decision makers is provided throughout the enterprise via an intranet. Users can view, query, and analyze the data and produce reports using Web browsers. This is an extremely economical and effective method of delivering data.

The various benefits offered by data warehouses can improve business knowledge, provide competitive advantage, enhance customer service and satisfaction, facilitate decision making, and help in streamlining business processes.

**Suitability**

Data warehousing is most appropriate for organizations in which some of the following apply:

- Large amounts of data need to be accessed by end users.
- The operational data are stored in different systems.
- An information-based approach to management is in use.
- The company has a large, diverse customer base (such as in a utility company or a bank).
- The same data are represented differently in different systems.
- Data are stored in highly technical formats that are difficult to decipher.
- Extensive end-user computing is performed (many end users performing many activities).

Hundreds of successful applications have been reported (e.g., see client success stories and case studies at Web sites of vendors such as MicroStrategy, Inc., Business Objects, Cognos Corp., Information Builders, NCR Corp., Platinum Technology, Software A&G, and Pilot Software). For further discussion, see Turban et al. (2005), Gray and Watson (1998), and Inmon et al. (2000). Also visit The Data Warehouse Institute (tdwi.org).

Although data warehouses offer substantial benefits, the cost of a data warehouse can be very high, both to build and to maintain. Furthermore, it may difficult and expensive to incorporate data from obsolete legacy systems. Finally, there may be a lack of incentive among departments within a company to share data. Therefore, a careful feasibility study must be undertaken before a commitment is made to data warehousing. Alternatively, one or more data marts can be used.

**DATA MARTS**

The high cost of data warehouses confines their use mostly to large companies. An alternative used by many other firms is the creation of a lower-cost, scaled-down version of a data warehouse called a data mart. A data mart is a small warehouse designed for a strategic business unit (SBU) or department. A data mart can be fully dedicated to EC.

The advantages of data marts over data warehouses include the following:

- The cost is low (prices under $100,000 versus $1 million or more for large data warehouses).
- The lead time for implementation is significantly shorter, often less than 90 days.
- They are controlled locally rather than centrally, conferring power on the using group.
- They contain less information than the data warehouse. Hence, they have more rapid response and are more easily understood and navigated than an enterprisewide data warehouse.
- They allow an EC department to build its own decision support systems without relying on a centralized IS department.

Data marts are either replicated (dependent) or stand-alone. Replicated data marts are those in which functional subsets of the data warehouse have been replicated (copied) into smaller data marts. The reason for using replicated data marts is that sometimes it is easier to work with a small subset of the data warehouse. Each of these replicated data marts is dedicated to a certain area, as shown in Exhibit W4A.1. The replicated data mart is an addition to the data warehouse. (This is why it is also called dependent—it exists on the data warehouse.) Alternatively, a company can have one or more stand-alone data marts without having a data warehouse. Stand-alone data marts are typically used for marketing, finance, and engineering applications.
OPERATIONAL DATA STORES
An **operational data store** is a database for transaction processing systems that uses data warehouse concepts to provide clean data. That is, it brings the concepts and benefits of the data warehouse to the operational portions of the business, often at a lower cost. It is used for short-term decisions involving mission-critical applications rather than for the medium- and long-term decisions associated with the regular data warehouse. Short-term decisions often require current information. For example, when a customer sends an e-mail query to a bank, the bank will quickly need to access all of the customer’s current accounts. The operational data store can be viewed as situated between the operational data (legacy systems) and the data warehouse. A comparison between the two is provided by Gray and Watson (1998).

SUCCESSES AND FAILURES OF DATA WAREHOUSING
Since their early inception, data warehouses have produced many success stories. However, there have also been many failures. Carbone (1999) identified several types of warehouse failures:

- Warehouse did not meet the expectations of those involved
- Warehouse was completed, but went severely over budget in relation to time, money, or both
- Warehouse failed one or more times but eventually was completed
- Warehouse failed and no effort was made to revive it

Carbone identified a number of reasons for failures (which are typical for many other large information systems):

**Data Problems**
- Not enough summarization of data
- Failure to align data marts and data warehouses
- Poor data quality (e.g., omitted information)
- Incomplete user input
- Incorrectly using data marts instead of data warehouses (and vice versa)
- Insecure access to data manipulation (users should not have the ability to change any data)
- Poor upkeep of information (e.g., failure to keep information current)

**Technology Problems**
- Inappropriate architecture
- Using the warehouse only for operational, not informational, purposes
- Poor upkeep of technology
- Inappropriate format of information—a single, standard format was not used

**Other Problems**
- Training and management issues
- Vendors overselling capabilities of products
- Lack or inappropriate training and support for users
- Inexperienced/untrained/inadequate number of personnel
- Unrealistic expectations—overly optimistic time schedule or underestimation of cost
- Lack of coordination (or requiring too much coordination)
- Cultural issues were ignored
- Improperly managing multiple users with various needs
- Unclear business objectives; not knowing the information requirements
- Lack of effective project sponsorship
- Interfering corporate politics

Suggestions on how to avoid data warehouse failure are provided by Griffin (2000) at datawarehouse.com and by Ferranti (1998).

DATA ANALYSIS AND KNOWLEDGE DISCOVERY
Once the data are in the data warehouse and/or data marts, they can be accessed by end users. Users can then conduct several types of analytical activities with the data, ranging from decision support and executive support analyses to ad-hoc queries, online analytical processing (OLAP), and data mining.
AD-HOC QUERY

Ad-hoc queries allow users to request real-time information from the computer that is not available in periodic reports. Such answers are needed to expedite decision making. The system must be intelligent enough to understand what the user wants. Simple ad-hoc query systems are based on menus. More intelligent systems use SQL (structured query language) and query-by-example approaches or Web-based applications.

Web-Based Ad-Hoc Query Tools

Web-based ad-hoc query tools allow users to access, navigate, and explore relational data to make key business decisions in real time. For instance, users can gauge the success of a Web marketing campaign according to the number of Web hits received last month, last week, or even yesterday, in relation to products or services purchased. This insight helps companies better target marketing efforts and forge closer, more responsive relationships with customers. Several vendors offer such tools. For example, Cognos Corp. (see cognos.com/products/query.html) offers Web users powerful ad-hoc exploration of corporate data assets, with little or no user training needed.

Advanced query tools can be connected to intranets and extranets for B2B and CRM querying. Also, a drill-down from multidimensional analysis to DSS and other tools are available. Answers to queries can be delivered to visualization tools.

ONLINE ANALYTICAL PROCESSING

Online analytical processing (OLAP) refers to such end-user activities as DSS modeling using spreadsheets and graphics, which are done online. OLAP is an information system that enables the user to query the system, conduct an analysis, and so on, while the user is at his or her PC. The result is generated in seconds. Unlike online transaction processing (OLTP) applications, OLAP involves many data items (frequently many thousands or even millions) in complex relationships. One objective of OLAP is to analyze these relationships and look for patterns, trends, and exceptions. Another objective is to answer user queries.

A typical OLAP query might access a multigigabyte, multiyear sales database in order to find all product sales in each customer segment (female, male, young people, etc.). After reviewing the results, an analyst might further refine the query to find sales volume for each sales channel by hours of the day or by product type. As a last step, the analyst might want to perform year-to-year or quarter-to-quarter comparisons for each sales channel. This whole process must be carried out online with rapid response time.

Thus, OLAP queries are able to analyze the relationships between many types of business elements (e.g., sales, products, regions, and channels) involving aggregated data over time (e.g., sales volumes, budgeted dollars, and dollars spent, on a monthly, quarterly, or yearly basis). The ability to present data in different perspectives involving complex calculations between data elements (e.g., expected profit calculated as a function of sales revenue for each type of sales channel in a particular region) enables users to pursue an analytical thought process without being stymied by the system.

Many vendors provide ready-made analytical tools, mostly in finance, marketing, and operations (e.g., productivity analyses, profitability analyses). Such packages include built-in Web-based DSSs. For example, Cognos Finance (from cognos.com) is an enterprisewide financial application for monitoring the financial performance of a business organization. It provides a framework for completing financial processes in a timely manner: monthly and quarterly closes, the budget process, and integration of the latest actual data with user-supplied forecasts. Users also can integrate Web information for a single view of the organization.

However, although OLAP can be quite useful, it is retrospective in nature and cannot provide the automated and prospective knowledge discovery that is done by advanced data mining techniques.

KNOWLEDGE DISCOVERY

The process of extracting useful knowledge from volumes of data is known as knowledge discovery in databases (KDD), or just knowledge discovery (KD). The objective of KDD is to identify valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is useful because it is supported by three technologies that are now sufficiently mature to produce meaningful data: massive data collection, powerful multiprocessor computers, and data mining algorithms.

Formal computer-based knowledge discovery has been done since the 1960s. However, the enabling techniques have been expanded and improved over time. KDD processes have appeared under various names and have shown different characteristics. KDD tools have evolved over time. Over time, KDD has become able to answer more complex business questions. For details, see Fayyad (1996).
**DATA MINING**

**Data mining** derives its name from the similarities between searching for valuable business information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material or intelligently probing it to find exactly where the value resides. In some cases the data are consolidated in a data warehouse and data marts; in others they are kept on the Internet and intranet servers.

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing the following capabilities:

- **Automated prediction of trends and behaviors.** Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly and quickly from the data. A typical example of a predictive problem is targeted marketing. Data mining can use data on past promotional mailings to identify the targets most likely to respond favorably to future mailings. Other predictive examples include forecasting bankruptcy and other forms of default and identifying segments of a population likely to respond similarly to given events.

- **Automated discovery of previously unknown patterns.** Data mining tools identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together, such as baby diapers and beer. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying invalid (anomalous) data that may represent data entry keying errors.

When data mining tools are implemented on high-performance, parallel-processing systems, they can analyze massive databases in minutes. Often, these databases will contain several years’ worth of data. Faster processing means that users can experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

Data mining also can be conducted by nonprogrammers. The “miner” is often an end user, empowered by “data drills” and other power query tools to ask ad-hoc questions and get answers quickly, with little or no programming skill. Data mining tools can be combined with spreadsheets and other end-user software development tools, making it relatively easy to analyze and process the mined data. Data mining appears under different names, such as knowledge extraction, data dipping, data archeology, data exploration, data pattern processing, data dredging, and information harvesting. “Striking it rich” in data mining often involves finding unexpected, valuable results.

Data mining yields five types of information:

1. **Association.** Relationships between events that occur at one time (e.g., the contents of a shopping cart, such as orange juice and cough medicine)
2. **Sequences.** Relationships that exist over a period of time (e.g., repeat visits to a supermarket)
3. **Classifications.** The defining characteristics of a certain group (e.g., customers who have been lost to competitors)
4. **Clusters.** Groups of items that share a particular characteristic that was not known in advance of the data mining
5. **Forecasting.** Future values based on patterns within large sets of data (e.g., demand forecasting)

Data miners use several tools and techniques: case-based reasoning (using historical cases to recognize patterns); neural computing (a machine-learning approach by which historical data can be examined for patterns through massive parallel processing); association analysis (using a specialized set of algorithms to sort through data sets and express statistical rules among items); and intelligent agents (expert or knowledge-based software embedded in information systems).

**A Sampler of Data Mining Applications**

According to a 2000 Gartner Group report (see Linden 2002), more than half of all the Fortune 1000 companies worldwide are using data mining technology. Data mining can be very helpful, as shown by the representative examples that follow. Note that the intent of most of these examples is to identify a business opportunity in order to create a sustainable competitive advantage.

- **Retailing and sales distribution.** Predicting sales, determining correct inventory levels and distribution schedules among outlets
- **Banking.** Forecasting levels of bad loans and fraudulent credit card use, predicting credit card spending by new customers, predicting which kinds of customers will best respond to (and qualify for) new loan offers
Manufacturing and production. Predicting machinery failures, finding key factors that control optimization of manufacturing capacity

Brokerage and securities trading. Predicting when bond prices will change, forecasting the range of stock fluctuations for particular issues and the overall market, determining when to buy or sell stocks

Insurance. Forecasting claim amounts and medical coverage costs, classifying the most important elements that affect medical coverage, predicting which customers will buy new policies

Computer hardware and software. Predicting disk-drive failures, forecasting how long it will take to create new chips, predicting potential security violations

Police work. Tracking crime patterns, locations, and criminal behavior; identifying attributes to assist in solving criminal cases

Government and defense. Forecasting the cost of moving military equipment, testing strategies for potential military engagements, predicting resource consumption; improving homeland security by mining data from many sources (see Chapter 8)

Airlines. Capturing data on where customers are flying and the ultimate destination of passengers who change carriers in hub cities so that airlines can identify popular locations that they do not service, checking the feasibility of adding routes to capture lost business

Health care. Correlating demographics of patients with critical illnesses, developing better insights on symptoms and their causes, learning how to provide proper treatments

Broadcasting. Predicting the most popular programming to air during prime time, predicting how to maximize returns by interjecting advertisements

Marketing. Classifying customer demographics that can be used to predict which customers will respond to a mailing or buy a particular product

TEXT MINING

Text mining is the application of data mining to nonstructured or less-structured text files. Data mining takes advantage of the infrastructure of stored data to extract predictive information. For example, by mining a customer database, an analyst may discover that everyone who buys product A also buys products B and C, but does so 6 months later. Text mining, however, operates with less-structured information. Documents rarely have strong internal infrastructure, and when they do, it is frequently focused on document format rather than document content.

Text mining helps organizations find the “hidden” content of documents, as well as additional useful relationships. It also helps them group documents by common themes (e.g., identify all the customers of an insurance firm who have similar complaints).

Web Mining

The previous discussion of data mining refers to data that usually are stored in a data warehouse. However, to analyze a large amount of data on the Web, one needs different mining tools. Web mining is the application of data mining techniques to discover meaningful patterns, profiles, and trends from Web sites. The term Web mining is used to describe two different types of information mining. The first, Web content mining, is the process of discovering information from millions of Web documents. The second, Web usage mining, is the process of analyzing what customers are doing on the Web—that is, analyzing clickstream data.

In Web mining, the data are clickstream data, usually stored in a special clickstream data warehouse (see Sweiger et al. 2002) or in a data mart. The strategies used may be the same in both. Several companies provide tools for Web mining; for example, iOpus (iopus.com), KD Nuggets (kdnuggets.com), Megaputer (megaputer.com), and SPSS (spss.com).
REFERENCES


