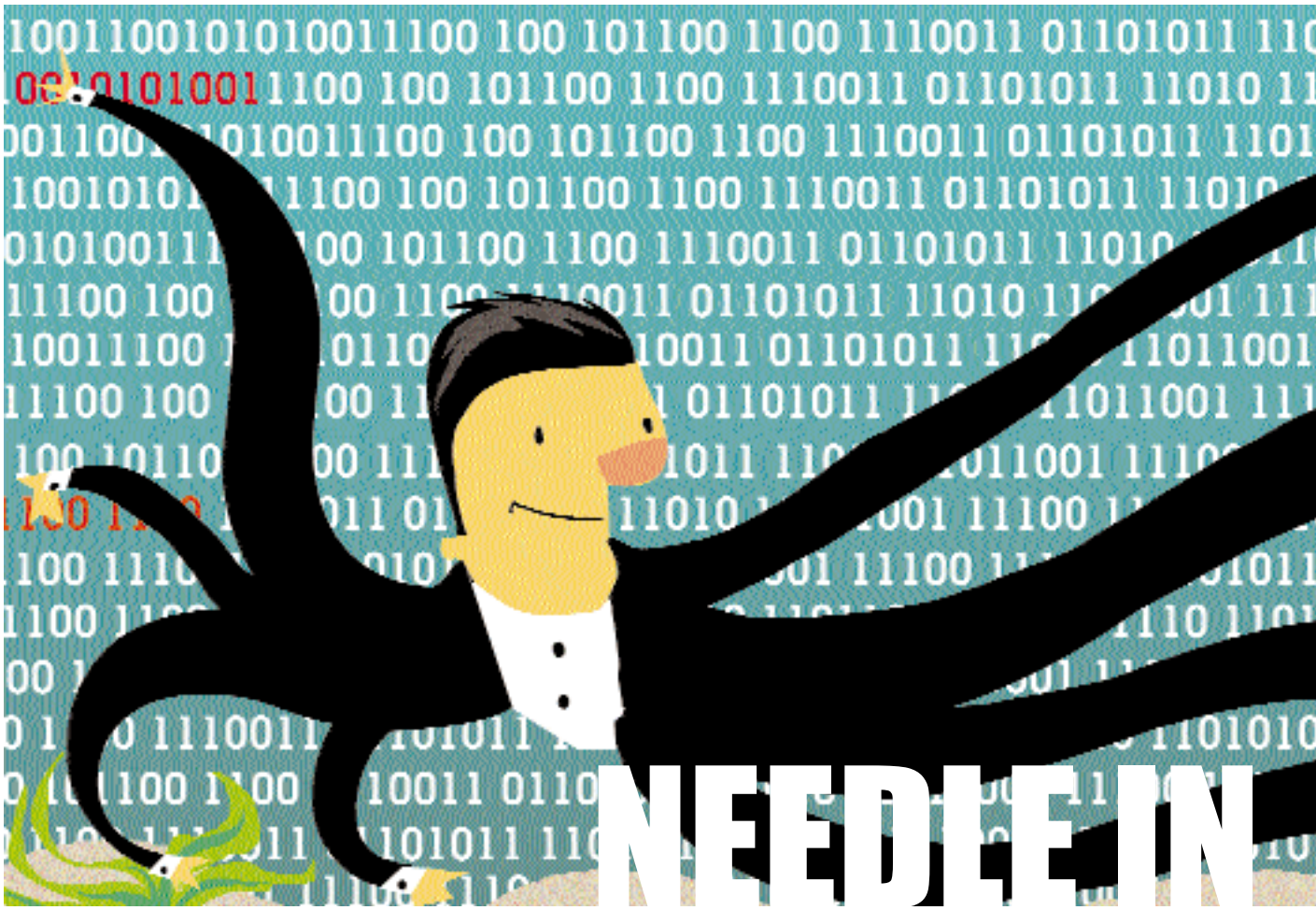

Popularly hailed as the solution to management information overload, data warehousing has become one of the most used—and abused—terms in the IT industry today

Twenty years ago, virtually all business system development was done on mainframe computers using languages such as the now obsolete COBOL. The 1980s saw the arrival of new minicomputer platforms such as IBM's AS/400 and the VAX/Digital. The late eighties and early nineties made UNIX a popular server platform with the introduction of client/server architecture.

Despite these variations in platforms, architectures, tools and technologies, a remarkably large number of business applications–by some estimates, over 70 percent–continue to run in the mainframe environment of the 1970s. These systems, generically called legacy systems, are the largest source of data.

In the days of legacy applications, integration of data and information was only a dream. Each application had its own idea of who a customer was, what a product was, and what an order implied. No two applications necessarily agreed on anything, and a corporate perspective of information did not exist. In addition, legacy applications looked at and captured only very current data. Historical data did not exist in any organised manner, and summary data was but a very small part of the operational environment.

Data warehouses got around these inadequacies by integrating data, historical data, and providing detailed as well as summary data. Data warehouses, intended to be large-scale collection/storage areas for legacy data, were based on analogies with real-life warehouses. They enabled the distribution of data to 'retail stores' or 'data marts' which were tailored for access by decision support users.

A miracle cure?

Somewhere along the way this analogy and architectural vision was lost, often manipulated by suppliers of decision support software tools. The architectural vision was frequently replaced by studies of how to design decision support databases. Suddenly the data warehouse had become a panacea for the decision support headache, and suppliers jostled for position in the burgeoning data warehousing marketplace.

In the last ten years, two factors have combined to help data warehouses proliferate. The first is a recognition of the

Illustration: SEAN D SOUZA

THE HAYSTACK

benefits of on-line analytical processing (OLAP) beyond the traditional areas of marketing and finance. Organisations found that the insights buried in the masses of data they routinely collect on their customers, products, operations, and business activities contribute to cutting operating costs and increasing revenues, not to mention making it easier to arrive at strategic decisions.

Secondly, the growth of client-server computing has spawned server hardware and software that is more sophisticated than ever. Today's servers are more powerful than yesterday's mainframes and offer technologically superior memory architectures, high-speed processors, and massive storage capacities. From this hardware/software renaissance emerges the huge data warehouses that we currently see in client-server environments.

X marks the spot

You know it is in there somewhere. Submerged deep within gigabytes of data lies the key information about an important customer trend that you wish to unearth. All you have to do is extract the proverbial needle in the haystack. Easier said than done?

Not really. Specialised tools let you capture the relevant data quickly and view it across many different data dimensions. The tool should not stop at merely accessing data—it should also enable you to meaningfully analyse the data; thereby transforming raw data into useful business information.

Business intelligence tools, the main point of contact between your warehouse application and the people who use it, sit on top of the data warehouse and provide this service. The simplest of these tools are basic querying and reporting products, which provide graphical front ends to SQL generators (or, more accurately, database access-call generators). The querying tool allows you to use point-and-click menus and buttons to specify data elements, conditions, grouping criteria, and other attributes of an information request. The query tool then generates a

WHO LEADS THE PACK?

Oracle Corporation is the leading company in the data warehousing industry, with nearly a quarter of the market share. The Redwood Shores, California, company is followed by SAS Institute Inc., Arbor Software Corp., SPSS Inc. and MicroStrategy Inc. One of the reasons for Oracle s success is that it capitalises on its large installed base of database users to sell its Express OLAP server, a luxury that vendors such as Arbor and MicroStrategy do not enjoy. The largest component of the data warehousing market was for

"A Data Warehouse is a repository of integrated information, available for queries and analysis. Data and information are extracted from heterogeneous sources as they are generated. This makes it much easier and more efficient to run queries over data that originally came from different sources."

—Stanford University

database call, extracts the relevant data, performs additional calculation and data manipulation if necessary, and presents the results in a clear format.

Digging for gold

Let us assume your data warehouse and its architectural components are ready. The next part involves exploiting the warehouse to full potential. Data mining is the next logical step in completing the circle of effective decision support. With data mining, you can discover important business patterns, examine relationships between obscure and otherwise unnoticed variables, and measure long-term trends. Data mining might be possible without building a data warehouse but the fact that data mining can be carried out does not mean that it can be done well. According to Bill Inmon, co-founder of Prism Solutions, Sunnyvale, California, (popularly called the 'father' of data warehousing) "a data warehouse provides an effective structure for data mining".

Bigger is not always better

The size of data warehouse databases has been increasing tremendously. Today, many companies are implementing warehouses in the terabyte (1000 gigabytes) range, and there seem to be few who question the boom or ask if all that data is really needed, justified or if it will even be accessed at all.

How much data, then, should be stored? If it costs a user the same to keep five years of data as it does to keep two years, the user will ask for five. However, when additional data is charged for, users will more carefully weigh the benefits and not automatically ask for additional years.

The frequency of keeping historical data

GLOSSARY

Client-Server: A distributed technology approach where the processing is divided by function. The server performs shared functions managing communications, providing database services, etc. The client performs individual user functions providing customised interfaces, performing screen to screen navigation, offering help functions etc.

Cooperative Processing: A style of computer application processing in which the presentation, business logic, and data management are split among two or more software services that operate on one or more computers. Individual software programs (services) perform specific functions that are invoked by means of parameterised messages exchanged between them.

Data Management: Controlling, protecting, and facilitating access to data in order to provide information consumers with timely access to the data they need. These functions are provided by a database management system.

Data Mining: A technique using software tools geared for the user who typically does not know exactly what he or she is searching for, but is looking for particular patterns or trends. Data mining is the process of sifting through large amounts of data to produce data content relationships. This is also known as data surfing or Knowledge Discovery (a term popularised by Gartner Group).

Data Modelling: A method used to define and analyse data requirements needed to support the business functions of an enterprise. These data requirements are recorded as a conceptual data model with associated data definitions. Data modelling defines the relationships between data elements and structures.

DRDA: Distributed Relational Database Architecture. A database access standard defined by IBM.

Enterprise Resource Planning (ERP): ERP systems consist of software programs which tie together all of an enterprise s various functions such as finance, manufacturing, sales and human resources. This software also provides for the analysis of the data from these areas to plan production, forecast sales and analyse quality. To maximise the value of the information stored in ERP systems, it is necessary to extend ERP architectures to include more advanced reporting, analytical and decision support capabilities best accomplished through the application of data warehousing tools and techniques.

ODBC: Open Database Connectivity. A standard for database access co-opted by Microsoft from the SQL Access Group consortium.

OLAP: On-Line Analytical Processing. These are applications that seek to verify complex hypotheses. An example of an OLAP query might be Compare the costs of shipping to customers in Europe to those in South America .

OLTP: On-Line Transaction Processing. OLTP describes the requirements for a system that is used in an operational environment.

SQL: Structured Query Language. A

DATA WAREHOUSE CONSTRUCTION TIPS

- It is difficult to strike gold with your first effort. You may need a few revisions before the set-up performs to expectations.
- Carefully consider your data. What formats and specific data are needed to support your application?
- Clean up your data before using it in the warehouse.
- Build a prototype mini-data warehouse as a learning experience and then revise strategies as necessary.
- Plan on more users taking advantage of a successful data warehouse than you initially expected.

is also determined by the user. It may be necessary (or cost effective) to keep daily or weekly data, but less frequent capture may be adequate. Keeping data weekly or monthly rather than daily or weekly can significantly reduce data storage requirements.

The GIGO factor

Knowing how Murphy's Law works overtime in the computer industry, one cannot be too careful about the quality of input that goes into a database. Of course, mistakes will happen—ranging from straightforward typos and minor inconsistencies (M.G. Rd vs M.G. Road, for instance) to inadvertent duplication when the same data gets listed more than once. This 'dirty' data can destroy the credibility of the entire warehouse set-up.

It is possible to 'clean' such data. However, if the amount of data is huge and the faults numerous, this can be... well, a dirty job. It may be better in such situations to rely on proprietary tools.

Even if you decide not to program the data-cleansing functions yourself or hire a consultant to do a custom job, you may not need to buy a tool specifically for the task. Your data warehouse management software might do enough cleaning and validation to meet your needs.

A boost to your business

As your warehouse grows, and the data it contains becomes more accessible, outsiders might also realise the value of the data and wish to access it. By linking your data warehouse to other systems within and outside the organisation, you can share information with little or no custom development.

E-mail, Web servers, and intranet/Internet connections can deliver inventory levels to your suppliers or order status to your business partners.

As data warehouses continue to grow in sophistication and reliability, the data housed within an enterprise will become more organised, more networked and generally more accessible to the people. This will naturally result in better business decisions, more business opportunities, and a more enlightened work force—and less time wasted on hunting for the elusive needle in the haystack!

HARIKRISHNAN MENON