UNPACKING THE BOXES FROM THE DATA WAREHOUSE

Connie Boynton, Department of National Defence, Ottawa, Canada

ABSTRACT

Many organizations are faced with legacy computer systems. The mainframe's cumbersome technology inhibits the efficient and accurate flow of data to upper management and decision-makers. Management level information needs to be produced in a presentable format, an activity better suited to the PC environment. Data warehousing, the SAS® system and other Windows® applications are useful tools for transforming data stored on the mainframe into meaningful, presentable information. This paper discusses a data warehouse project in the civilian human resource environment of Canada's Department of National Defence. A faster, more efficient method of information flow was necessary since the previous method was not meeting the requirements of decision-makers. The data warehouse project involves taking operational data from the mainframe, storing it on a server and using SAS and Windows tools to provide current and comprehensive information to upper management. The project was successful in creating an efficient, accurate information flow. This paper will be useful to other organizations faced with urgent data needs and inefficient technology. It shows the value of setting up a data warehouse along with a group of people who thoroughly understand the data.

INTRODUCTION

The purpose of this paper is to highlight some of the problems and solutions related to information delivery. The system discussed in this paper deals particularly with information delivery for the civilian human resources directorate of Canada's Department of National Defence. DND is the largest single employer in Canada. An efficiently functioning employee DBMS and information delivery service are essential to meeting the information requirements of policy-makers. Timely, accurate information became especially important as DND entered a restructuring phase. It also became important for the data to be in electronic format for use in applications and to create presentations.

Data accessibility problems are inherent to mainframe technology and inefficient information flows. There are two main problem areas. The first problem is to get electronic data into a more common environment such as Windows. The second problem is the time-lag between the initial request for information and the final product such as graphs, spreadsheets or applications.

Employee data is gathered from across Canada and stored on an IBM mainframe as VSAM (Virtual Storage Access Method) files in a CICS (Customer Information Computer System) environment. The data contains personal information about each employee such as work location, age, years of service, and tenure, as well as system codes used for identification purposes. Since not all of the stored data is required for analytical purposes, a data warehouse is an effective way to separate the operational data from the decision support environment.

Regularly scheduled programs are run against the employee database to accumulate flow data on such topics as recruitment, promotions and separations. Other data, such as position appointments, method of hiring, etc., is accumulated from an on-line pay transaction system. The data has been gathered and stored regularly over many years but has not always been easily accessible or utilized to its full potential by decision and policy-makers, a scenario which has been changing in recent years.

INFORMATION FLOW: WITHOUT A DATA WAREHOUSE

A group composed of mainframe users are designated to provide human resource data within the Department of National Defence. The group performs data manipulation and queries against the mainframe files to supply hard copy reports or mainframe screens to information seekers. Reports are mailed to the party requesting information where the figures are assessed and manually input into presentation packages.

© Info-seekers must make their data requests in writing or verbally which has many disadvantages. Reports usually need elaboration or redefinition, once the info-seeker has viewed the output. This causes expensive delays in delivery of the final information product.
Figure 1: Without a data warehouse

Figure 1 depicts the flow of data which currently exists in most of the department. The collected data is stored centrally on the mainframe and manipulated by a group of mainframe users. The data is not subset or warehoused on particular topics since few people are performing queries against these files. Hard copy reports are provided to people requesting information and the data is manually entered into presentation software packages.

Standard reports are created for periodically updated screens on the mainframe. The screens display data electronically but do not allow any further manipulations by the user. The user also cannot drill-down or load the data directly into spreadsheet or graphic packages. In addition, the person requesting information is disconnected from the mainframe data management process. Once receiving the data, it needs to be deciphered and manually entered into graphic or spreadsheet applications.

Most of the department is still functioning with this process. The mainframe group undergoes much pressure to disseminate all the various types of information to clients in a timely fashion.

Information Flow: With a data warehouse

The civilian human resource department required a faster, more efficient flow of information. It needed a group who had a thorough understanding of the data and who work effectively in a team environment. An IT group was gathered together and guided carefully by the needs of senior management for the duration of the data warehousing project.

The informatics group was assembled to bridge the gap between the raw data stored on the mainframe and the information that needs to be presented to senior management for decision making purposes. The role of this middle layer, IT group, is to summarize the necessary data into the format required for analysis and presentation.

Figure 2 shows the flow of data through to the new informatics group. The data is subset and warehoused onto a LAN server. Each member of the group has access to the data and can develop automated presentations as well as explore the data.

The group analyzed the information requirements and selected the data fields to meet the need. The specified fields are so useful that this information meets 95% of upper-management's analysis needs. The other 5% of requests are made via the information flow described in Figure 1.

Prior to start-up, there was a considerable time-lag between the request for information and the return of the final product. There was also a lack of understanding of exactly what information was being obtained. Personnel people requested reports from mainframe users who were distant to the decision-makers of the human resource world. The group helped to streamline the process and to bring a better understanding of the data that was buried in the depths of the mainframe.
Building the Data Warehouse: Storage facilities

The informatics team's first step in building their data warehouse was to select meaningful fields from the mainframe files. The necessary fields were determined through consultation with senior management and group discussion. Some new fields were created to meet the specific information requirements of the directorate. These fields are created in a SAS program which is run using each month-end file before being stored.

The files are stored on a LAN server where they can be accessed by each member of the informatics group. Subsetting and storing the files separately to the operational files on the mainframe means that no querying is done to the mainframe files which could slow down the entire system.

A series of tests are performed in order to locate anomalies in the datasets and to ensure that they were downloaded properly. Once it has passed those tests, a program is run to create the necessary analysis fields, such as age bands and salary bands. A library of various formats is also on file to use with the appropriate dataset. A well documented data dictionary is maintained to provide maps and layouts and to describe transformations that fields went through to reach their current state. Organization and clarity are an important part of a successful data warehouse project.

Month-end snapshots of the data are stored on the server. Month-end data was found to be current enough to keep up with the information requirements of upper-management.

The informatics group found it to be good practice to use numbers from month-end to allow for accurate cross-checking and verification of previously derived figures.

The datasets are compressed to allow for larger storage capabilities. These datasets are suitable for compression since they are on the output side of the information system and no data-entry is required. Typical datasets start off between 20-30 megabytes and compress down to between 13-19 megabytes, decreases of close to 35%, as shown in Table 1.

Table 1

<table>
<thead>
<tr>
<th>File Sizes Before Compression</th>
<th>Average File Size After Compression</th>
<th>Average Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-30 Mb</td>
<td>13-19 Mb</td>
<td>35%</td>
</tr>
</tbody>
</table>

Unpacking the Boxes: Software Applications

The data has been subset and transformed from its original state on the mainframe and is now available to be used in an automated fashion by information delivery tools. Successful, high profile applications have been developed via the data warehousing project. The IT group is able to answer questions and provide statistics within minutes, where it used to take days and weeks to answer the same questions, depending upon the priority of the request.

Standard reports, spreadsheets and graphs have been automated for use in presentations and the decision making process. The IT group employs the SAS system for its strength in manipulating data and producing statistics. Presentation quality reports are prepared in other well established applications and tools owned by the department, such as Lotus®, Harvard Graphics®, Foxpro® and Microsoft Access®. SAS programs format the data and open, for example, LOTUS or Foxpro using DOE to produce spreadsheets and reports. The combination of software applications efficiently produces statistics in presentation format.

Data warehousing technology allowed for the rapid development of a PC-based system which dealt with the restructuring in the department. The system was easily accessible to upper-management who worked closely with the development team. Management was able to shape the system according to policies and react promptly to problems as they arose. Progress reports and statistics were constantly in demand and easily provided due to the proximity, versatility and clarity of the system. The success of the project was evident to all involved.

Conclusion

The informatics team chose the topics and fields that would produce the swiftest results with the most visible benefits to their directorate. The data warehouse was laid out and has been evolving since the project implementation. The IT group is
preparing to explore the next information area on the priority list. The data warehousing concept has proven itself to be an efficient method of disseminating information to people with questions.

Use of the data warehouse concept will expand business intelligence in two ways in the Department of National Defence. More employees will be accessing information electronically, and different types of information will become more widely used and accessible.

It is inevitable that data use will increase as people become more familiar with the user-friendly querying software now available. The scope of the data warehouse will expand the type of information that can be analyzed by interested parties. Frequent use of information in a data warehousing environment also facilitates monitoring of data integrity and validation. More users will be able to access the data and can answer questions as they arise. Acquiring information is finally expanding beyond the immediate realm of the data maintenance group and reaching the people who are asking questions.

SAS is a registered trademark of SAS Institute Inc. Lotus is a registered trademark of International Business Machines Corporation. Harvard Graphics is a registered trademark of Harvard Graphics. Foxpro and Microsoft Access are registered trademarks of Microsoft. ® indicates USA registration.