Paper #3003
Report Mining - A New Information Delivery Paradigm

Gordon Croston
Genesee Software, Inc.
7977 South Wabash Court
Englewood, CO 80112
303/850-9128

It's Groundhog Day For The I.S. Department

In the movie *Groundhog Day*, an overburdened television weatherman finds himself trapped in a strange time warp. The calendar is stuck on February 2nd, Groundhog Day, and he is condemned to repeat it, over and over, until he gets it right. The only way he can break the cycle and escape from his "day-in-hell" is to change his whole outlook on life.

Many IS managers face a similar (but real-world) problem. In their role as the keepers of corporate information, they are caught in a seemingly endless cycle of reporting. For years, they have been cranking out reports full of valuable information, but users are never truly satisfied. Users want a *different cut* at the data -- a new sort, a new roll-up, a new way to get the information they need.

Each request results in the creation of a new report which provides temporary relief until another new cut is desired. Even the smallest change can mean one more "day-in-hell" for the IS department.

The situation is especially frustrating because IS managers *know* how to break the cycle. In theory, the answer is simple: Give users tools to access the data *directly*, so they can perform their own queries and construct their own reports. In a 1994 Gartner Group survey of IS professionals, more than 80% said that providing such tools is a strategic priority within their organizations. Everyone wants to empower users to access information. *How* to do it is the tough part.

The Evolution of Information Delivery

In this white paper, we'll look at three common approaches to delivering corporate information, and introduce a hybrid approach called *report mining* which borrows some of the best features from each.

Hardcopy Reports: The Common Currency Of Corporate Information

Far and away the most popular delivery vehicle for corporate information is the hardcopy report. In 1995, data center print production was expected to surpass 600 billion pages per year, a 500% increase since 1984.

The sheer volume of print production is astonishing. But even more astonishing is the growth rate. In this age of client/server technology, one might expect that hardcopy reports would be

quickly losing favor. In fact, reports are gaining momentum! Arguably, reports are the *fastest-growing* information delivery technology. It's scary.

Fortunately, most reports are actually quite useful. After all, they embody a tremendous investment of hard work. Through many generations of revision, the reports have evolved to become rich in valuable information.

A good report is not merely a data dump. Rather, it is a carefully-crafted document which turns raw data into information. Typically, the data is drawn from multiple tables and joined at the time of reporting. It is sorted, summarized and formatted in order that it may bring value to it's recipients. On rare occasions, a report can be a masterpiece.

More often, however, reports contain too much information. Everything that anybody might possibly need is included. One size fits all. And all that information is *frozen* on the page. There's just no easy way to get at it.

How many hours have been wasted in your organization because people must dig through mountains of printouts to find the answers they need? And how many more hours have been wasted because data must often be rekeyed into PCs for further analysis or presentation? Chances are, the lost hours are costing you real money on your bottom line.

For all their shortcomings, hardcopy reports serve a vital need. Day in and day out, they are the currency of corporate information. Direct access technologies hold promise to one day supplant them, but, so far, progress has been slow and expensive.

Direct Access Technologies: First Attempts At Data Mining

Going directly into operational databases to retrieve information is a tantalizing concept. Vast deposits of data sit waiting to be tapped. If PC users could be equipped with data mining tools, an information bonanza would soon result -- such were the predictions just a few years ago.

Powerful data mining tools have indeed been invented, but no bonanza has followed. Why? The problem, essentially, is "low grade ore." All the data sitting in operational databases is structured to support transaction processing, *not queries and analysis*.

Wayne Eckerson, an authority on data access technology, cites important differences between operational data and informational data.

Operational Data	Informational Data
Atomic	Summarized
Application Oriented	Subject-Oriented
Current	Historical
Dynamic	Static

According to Eckerson, these differences make it very difficult to tap operational databases directly. Another problem is that operational data tends to be spread across multiple physical databases, on different platforms, in different formats.

"Every system we have seems to deal with name, address, and telephone number information in a different way," says Randy Sprague, information systems project leader at the San Francisco Newspaper Agency, as quoted in a recent InfoWorld article.

Without consistency in naming and format conventions, data mining across multiple operational databases is nearly impossible. Executive information systems, business intelligence systems, query systems and other data mining tools depend on higher grade ore to work effectively.

A third problem is performance. As Eckerson reports, “Companies often don’t want to risk letting a user issue a complex, data-intensive query that brings their production systems -- and business -- to a stand still.”

“Because of the inherent differences between operational and informational systems,” says Eckerson, “many companies choose to build separate databases to house decision support data.” These separate information-oriented databases are called data warehouses.

Welcome To The Warehouse

In simple terms, a data warehouse is a staging area for corporate data, designed to support information access without impacting production systems. Raw data is extracted from operational databases at pre-determined intervals and repackaged so that it is more accessible. Data from multiple databases are frequently combined.

Many experts believe that data warehousing is the correct approach to bridging the gap between operational and informational systems. Specialized data warehouse vendors have emerged, including Trinzic, Prism, Redbrick, ETI and Carleton, and mainstream database vendors such as Oracle, CA/Ingress and IBM have added warehousing facilities to their standard database systems.

Some organizations have already embarked on data warehousing projects. A recent article in Forbes magazine describes several such projects, undertaken by companies who are determined to unlock the "hidden jewels" in their corporate databases. One example: Managers at Chevron needed information to help them manage petroleum inventories held in storage tanks throughout their nearly 100 U.S. refineries and marketing terminals.

Gathering that information was not easy. According to Jim Seid, a systems analyst for Chevron’s technology group, the data “was scattered throughout a number of different databases across different computers all over the continent.” The data had to be cleansed, formatted and transferred electronically to an SQL-based data warehouse, an effort described in Forbes as “a Herculean task."

In another example, a similar effort at Monsanto was found to be “more complex than we had imagined.” Although the staging-and-cleansing process should have been automated, this was found to be “impossible.”

Both Chevron and Monsanto have benefited greatly from their warehousing efforts, but costs have been substantial. (In Chevron’s case, $750,000.) Building and maintaining a data warehouse is obviously a big job.

Building a Data Warehouse

Bill Inmon, who is credited for laying much of the conceptual groundwork for data warehousing, points out that “the warehouse is only a small part of a larger infrastructure.

Surrounding the warehouse is a complex and sophisticated set of hardware, software and procedures."

The surrounding infrastructure serves as the interface between legacy systems and the warehouse. "As a rule," says Inmon, "this interface -- where integration and transformation of data is done -- is particularly complex." He cites the following functions as necessary constituents of the interface process:

- Extract Data
- Convert Data
- Map Data
- Reformat Data
- Recalculate Data
- Restructure Keys of Data
- Summarize Data

All these functions must be performed as data is moved into the warehouse and made ready to support queries, reporting and analysis.

If developing the interface seems like a daunting task, consider that it is just one step in the entire implementation process. One expert lists 33 separate steps. A partial list gives a sense of the problem:

Warehouse Implementation Steps:

- **Identify requirements.** *(What information does the user need?)*
- **Identify potential sources for the information.**
- **Develop the warehouse project team.**
- **Develop warehouse conceptual, logical, and physical models.**
- **Identify and document sources of derived data.**
- **Develop systems to move data into the warehouse.** *(The interface.)*
- **Develop, test, transform and load warehouse programs.**
- **Connect/load meta data to dictionary.**
- **Develop, install and test cataloging software.**
- **Load meta data and dictionary to catalog.**
- **Provide user training, tolls, data and processes.**

Many of these steps involve *meta data*, which define the names, relationships and operations that apply to the data elements. (e.g., parts count = sum(part 1 .. part n)). Meta data tells the software how the data should be processed. Overall, the meta data must embrace the business rules which govern the enterprise.

For organizations who complete the implementation process, the payoff can be enormous. With a data warehouse in place, users of corporate information are empowered to perform their own queries and analyses. The cost for IS? Sweat, toil and infinite patience. Imagine if users could achieve substantially the same result using existing reports (report files) as a *proxy* for the data warehouse!

The Warehouse Alternative: Report Mining

The idea is simple -- instead of mining data buried in a central database, users mine data buried in reports. Any report used in your organization can be brought on line and mined by

PC users. Gartner Group analyst Howard Dresner explains the value of report mining: "Existing reports already contain meta data, data, and business rules, which have endured the test of time," says Dresner. "I believe this message will be welcomed by many CIOs and IS managers who question the value and wisdom of data warehousing projects." Report mining tools can help IS rediscover the value of their existing mainframe report investments.

Report mining tools let users access and manipulate data buried in computer-generated reports. They read report files downloaded from any mainframe, midrange or client/server system. (These files are commonly known as print files, spool files, TXT files, formatted ASCII files, PRN files and SDF files.)

Report Mining tools exploit standard report generation procedures: Whenever a report is generated, a report file is produced which contains all the characters and control codes that are sent to the printer to produce the actual printout. Report mining reads that same report file, but instead of producing a *hardcopy* printout, it displays a *softcopy* of the report on screen, with live data the user can access and manipulate.

Using report files as a source for data offers many advantages:

Data is instantly available. Every report produced by the IS organization represents a ready-made database that report mining tools can exploit. No restructuring is necessary. ("The real value of report mining tools," says Dresner, "comes from their ability to promote access and analysis of report information directly -- *without an intermediary transformation process.*")

Compatibility is achieved across computing environments. Because the computer industry has adopted standard conventions for sending characters to printers, report mining tools can read report files generated in virtually any computing environment. The tools don't care how, when or where the reports were created.

Users are immediately productive. Users adapt easily to report mining tools because they already rely on reports to get information. No training is required.

Data security is maintained. Since report mining tools read report files and not the central database, production data stays secure and out-of-reach.

Existing reports are leveraged. The IS organization has put tremendous effort into building existing reports and reporting systems. Report mining tools leverage that investment.

IS resources are conserved. This benefit takes several forms. Since report mining tools run on the desktop, host systems are not impacted. And fewer demands are placed on IS to produce custom reports because users can create new reports with data extracted from existing reports. Most importantly, report mining takes pressure off the IS organization to deploy warehouse technology before it is ready. Report mining tools work today, are easy to use, and provide immediate value.

Real-World Experience

If report mining tools are here today, who's using them? What real-world problems do they solve? What kind of payback can be expected?

At Lehman Brothers, report mining is used primarily for ad hoc reporting and analysis. Director of Systems Operations Ron Queler says, “Because more of our users access report data through report mining, our custom programming requests have dropped to an all time low.”

At Aroostook Medical Center, report mining is used to prepare special audit reports and management reports with data pulled from AS/400 report files. “Report mining is the precise toll needed to bridge the gap between the information expectations of the PC user and the information realities of the Management Information Department,” says Benton W. Cash, the hospital’s controller. “It saved us thousands and thousands of dollars by filling the niche that our data processing department just couldn’t fulfill.”

A similar story is told by Clark R. Abrahams, Vice President, Manager of Portfolio Quality Information at NationsBank. “Report mining has eliminated the programming middleman, and has put the power in the hands of the end-user of the information. We have a critical need to capture data from both the mainframe reports and client/server based reports in a spreadsheet environment,” says Abrahams. Once in the spreadsheet, the data can be further analyzed, graphed and reported.

“Report mining is the universal data adapter plug which makes this possible,” says Abrahams. “Thanks to it, we have cut turnaround time for management reporting by two days and saved hours of error-prone re-keying and subsequent re-validation of the data.”

At Digital Equipment Corporation, Robert Yorke uses report mining to compute contract accruals. Prior to adopting it, contract accruals were calculated manually from computer printouts at the end of every quarter, a monotonous task which required the efforts of four people working four days to complete. “Report mining cut that to two people working part-time for two days,” says Yorke. “The original reports (500+ pages) are summarized down to just one page. It’s wonderful to see the results coming out right to the penny.”

Digital also uses report mining to produce custom contract summaries for very large customers. “Contract administrators send contracts by e-mail and we produce summaries according to customer specifications such as by customer sites. This has greatly reduced the billing cycle for both our contract administrators and for our customers in reconciling their invoices. It is a tremendous help for everyone involved.”

Deploying Report Mining Solutions

In some organizations, report mining tools are used purely on an ad hoc basis, with report files downloaded from the host as needed. In others, report files are deliberately placed on a network server and accessed throughout the day by multiple users. The report files are organized into directories based on report type and run date. Access rights are established by the network administrator.

Report mining tools are especially valuable to organizations which have adopted mainframe-based report distribution systems and/or COM-replacement systems. (COM is an acronym for Computer Output Microfiche.) These systems, from vendors such as Mobius, Legent, Computer Associates, 4th Dimension and RSD, help organizations manage the distribution and archival storage or reports. Some organizations have terabytes of reports held in magnetic and optical storage.

Joel Wecksell, a Gartner Group analyst who follows the market for report distribution systems, says, “Tools that enable users to manipulate reports delivered by RDS products add significant value to the process.” Users can extract, analyze and redeploy data from any report held in the RDS archive.

Making Use of the E-Mail Engine

A great way to deploy reports is to deliver them electronically, using the same vehicle that electronically delivers mail - the corporate e-mail system. Basically what happens is the report file becomes a file attachment within an e-mail message. That e-mail message is sent by the e-mail engine to the mailbox of the report recipient. The recipient then executes the report mining software against the attached file. Some Windows based e-mail software will even have the report mining tools icon in place of the attached file. The report recipient simply double clicks on the icon and views and mines the report immediately.

The easiest implementation of this strategy is when the machine where reports are produced (probably a mainframe computer) also houses an e-mail system that is at least part of the corporate e-mail network. Other parts of the corporate e-mail network (LAN based e-mail, Internet e-mail, MCI Mail, etc.) are usually linked via gateways to this mainframe. With this e-mail network structure an e-mail message, with the attached ASCII spool file, is sent to the mainframe e-mail system. The message is pushed to the proper e-mailbox, anywhere in the corporate e-mail network, by the mainframe e-mail engine. Tools are available to automate the process of message creation. For example, a message and file attachment can be automatically created and mailed to a predefined destination every time the spool file appears in the spooler.

What if the computer that creates the spool file is not part of the corporate e-mail network. Somehow that spool file needs to get to the e-mail network. This can be done in one of two ways. The e-mail network could be expanded to include the mainframe where the report resides. This would take e-mail software on the mainframe along with a gateway attached to the network e-mail system. The report could also be moved to a machine that is part of the e-mail network. From there proper tools could be used to pass the mail message and file attachment to the e-mail system residing on that machine.

“With a report mining tool,” says Wecksell, “users can mine existing reports, finding nuggets of value that enable them to make better, faster business decisions. These reports are of great value and represent a significant investment to the enterprise.” Report mining tools enable organizations to leverage this investment, providing users with a new tool for information empowerment, while retaining the management, control and security of legacy systems, something that comes at a high cost in emerging client/server computing paradigms.

Report Mining and the IS Manager

It’s not often that service organizations get the opportunity to hand their internal customers a ready-made solution to a difficult problem -- a solution that is virtually free when compared to other alternatives. Report mining can make heroes out of IS managers.

Stable, long-term, consistent, and evolutionary are words most IS managers would like to apply to their IS strategies. Report mining fits this model well. It requires no reengineering and complements whatever comprehensive information delivery systems are ultimately put

in place. It can be part of a strategic data access solution, or simply a tactical measure to satisfy data-hungry users who want immediate results.