Decision Support with Web-Enabled Software

MARC-DAVID COHEN
SAS Institute
SAS Campus Drive
Cary, North Carolina 27513
marc.cohen@sas.com

CHARLES B. KELLY
SAS Institute
chuck.kelly@sas.com

ANDRÉS L. MEDAGLIA
SAS Institute
andres.medaglia@sas.com

Abstract

SAS develops software for building Web-based applications for data management, statistical analysis, forecasting, data mining, and operations research. Our customers have used these tools to build several kinds of Web-based decision support applications. A Web-based optimization framework is used to solve a large-scale production-and-distribution problem at United Sugars Corporation, to solve an inventory replenishment problem at Cameron & Barkley, and to find optimal portfolios of suppliers using Dun & Bradstreet data. Applications implemented at Dow Corning and the Aeronautics Division of Lockheed Martin make statistical process control and project management data available via the Web. We conclude with a discussion of Web mining and some of its challenges. Visit the companion Web sites at http://www.informs.org/interfaces/ebiz and http://www.sas.com/supplychain.

The Internet is an electronic marketplace, meeting place, and medium for communication. It has created myriad opportunities and businesses are emerging to exploit them. For over 20 years, SAS has provided software for data management, statistical analysis, forecasting, data mining, and operations research. Now, its products simplify building applications that use the Internet as a communication channel and find wide use as a tool for analyzing Web usage.

The World Wide Web has grown rapidly for several reasons including a single user interface paradigm, its distributed architecture, and the growth of open standards. Web browsers make it easy for the novice user to gain access to many diverse sources of information. As a result of this and the Internet's open structure, many people and organizations provide information for delivery and use the Internet for finding information. Moreover, because of its distributed architecture, they constantly provide fresh information, keeping data current. The OR community needs to exploit the potential of the Internet.

Solution Architecture

One approach to exploiting the Internet's potential is to distribute processing across multiple servers as a convenient and efficient way to solve large analytical problems. We have developed applications that use a multitiered architecture (Figure 1) to do this. The user uses a Web browser to drive programs on an application server. The application server collects information for building a model, for example, a linear program or a mixed integer program. Some of the information the user needs to build that model may reside in a database housed elsewhere, which may be accessed via a data server. At some point, the user may request that the model be optimized. The application server then starts a process on an optimization server, passing it the model (business logic) for solution. The optimization server may also need access to the database. When it finishes the optimization, the optimization server sends the user an e-mail message with a universal resource locator (URL) that displays the solution, or it may also publish reports with the solution to different channels of a Web portal. The application server can process this URL and serve up the solution as stored in the database.

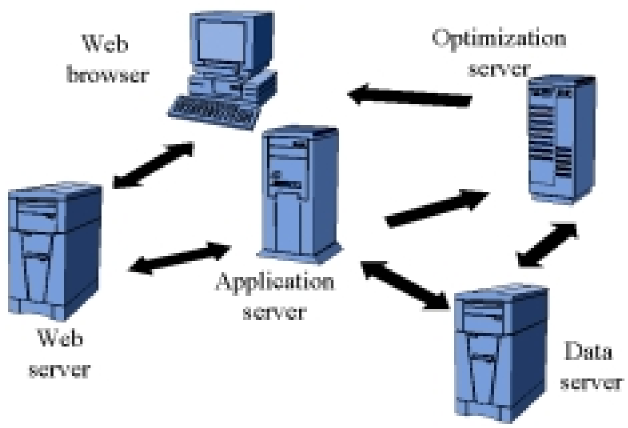


Figure 1: In this multitiered architecture for solving optimization problems, several processes communicate via both one-way and two-way links. The user builds and optimizes a model via a standard Web interface. Once the optimization problem is solved, the optimization server can e-mail the user a URL with the solution.

This architecture relies on the Web server having access to a registry of available application servers, which act as brokers that have access to registries of information on the locations of data, data servers, and optimization servers. We are building the necessary interfaces into SAS® software to provide scalable solutions that follow this model.

Data are classified as either static or dynamic. Static data, which do not change or change slowly, include product information, product specifications, and product reviews. Dynamic data, which change quickly, are often measures of process state. Examples include the number of items in inventory, item prices, the quality of a manufactured product, or how much a process departs from a schedule. The dynamic data and the process state that these data often represent are usually found in enterprise resource planning (ERP) systems, particularly in manufacturing. Because these systems are operational in design, they are not suited for tactical or strategically focused analytical solutions. Often, this limitation is remedied by building a data warehouse on top of the ERP system. These warehouses contain snapshots of process state represented in the ERP systems. They provide static platforms for developing analytical solutions and can be extensive and distributed across intranets, extranets, and the Internet (Figure 2).

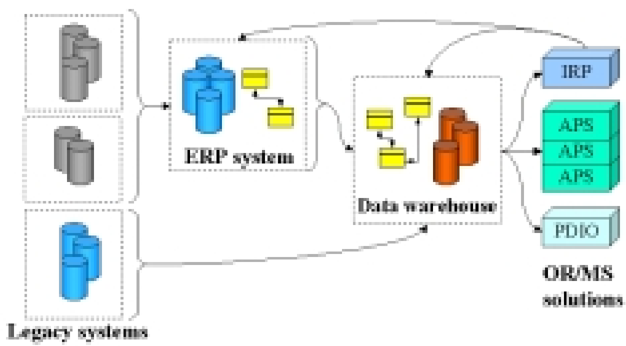


Figure 2: In this data architecture, which supports Internet-based optimization solutions, an ERP (enterprise resource planning) system and legacy data systems feed into a data warehouse. Snapshots of these transactional systems are taken periodically and uploaded into the data warehouse. Solutions (expressed by business models and solved on optimization servers) sit on top of the data in the warehouse. Output from these feed back into the warehouse and into the ERP. Shown here are APS (advanced planning and scheduling), PDIO (production distribution and inventory optimization), and IRP (inventory replenishment planning) solutions.

Optimization and Decision Support via the Web

Several organizations rely on a multitiered architecture (Figure 1) to drive Web-based applications that facilitate optimization and decision support. The United Sugars Corporation wants to determine optimal production, distribution, and inventory capacity. Cameron & Barkley Company focuses on identifying optimal inventory replenishment policies. Buyers use the Supplier Portfolio Optimizer to identify portfolios of suppliers that meet their criteria while improving purchasing leverage.

United Sugars Corporation

The United Sugars Corporation of Bloomington, Minnesota sells and distributes sugar products for its member companies. With a 25-percent share of the US sugar market and sales of over $1.0 billion, this grower-owned cooperative is one of the nation’s largest marketers of sugars to industries and consumers.

Until 1997, United Sugars produced and distributed all of its sugar products from the Red River Valley area in the upper Midwest of the United States. When the United States Sugar Corporation in southern Florida joined the cooperative in 1997, United Sugars decided to revise its marketing and distribution plans to gain access to new markets and to serve its existing markets more efficiently. Also, because United States Sugar is Florida's largest producer of cane sugar, the cooperative could supply high-quality sugar to all of its customers year-round. With this new flexibility in sugar production, United Sugars decided to take a more systematic approach to managing its supply chain.

In 1998, the United Sugars Corporation collaborated with SAS to explore methods for modeling and managing its supply chain. Following our supply-chain methodology for aligning supply and production with demand [Cohen and Kelly 1999], we designed and built a strategic model to identify the minimum cost solution for United Sugars' packaging, inventory, and distribution problem. United Sugars extracted business data from an SAP AG R/3® system and a legacy database to populate the mathematical model. The company uses the results monthly to plan operations that account for approximately 85 percent of its operational budget.

The resulting network model for planning over a 13-month horizon with approximately 80 plants (packaging, storage, distribution, and transportation transfer points), 250 sugar products and over 2,000 customers contained roughly 220,000 nodes, 1 million arcs, 3,000 non-arc variables, and 26,000 linear side constraints. United Sugars solved the optimization problem using the primal-dual predictor-corrector interior-point algorithm of the NETFLOW procedure [SAS Institute Inc. 1999b] on a PC running Windows NT (450 MHz and 1 GB of RAM) in about 2.5 hours.

The optimal solution specifies the minimum-cost schedule and indicates the packaging, distribution, and inventory assets needed to satisfy customer demand. United Sugars uses the results in its current planning cycle for budgeting all freight and warehousing expenses. Primary users include the production planning, logistics, and distribution managers. Currently, the company uses the results to make strategic decisions, but it plans to upload information to its ERP system to support operational decisions in the future.

The solution provides essential information for these areas of planning:
— production packaging,
— labor and resource requirements,
— warehousing and freight expenses,
— inventory management, and
— order distribution.

Analysis of the results has provided additional insight into several business processes. By not limiting the constraints on the capacity of some storage facilities, we used the model to identify the optimal storage levels. In some cases, this showed that a facility was being underused, while in others it suggested that the corporation should acquire additional public storage. Because the model permits flexible data input and problem generation, it is easy to perform what-if analyses with various scenarios. By making multiple problem runs and resolving many data infeasibilities, the model identified several potential problems. These included

— how to manage insufficient sugar supply,
— how to handle customer sourcing requirements,
— how to reconcile inventory inconsistencies,
— how to improve inventory turns, and
— how to incorporate *fixed* production planning.

In some cases, we modified the model formulation to incorporate additional information in the final solution, while in others, we generated problem-specific reports to provide summary results to aid decision making and enable data correction, if necessary.

United Sugars creates reports that contain detailed and summarized information. They include

— a packaging-line planning schedule,
— a packaging-line capacity-utilization report,
— distribution requirements planning flows,
— freight costs,
— inventory movement (interplant shipments), and
— warehouse utilization and costs.

In the future, United Sugars plans to create graphics-quality reports with summary information and drilldown capability to be distributed via the Web for multiple-user access (Figure 3). By using the Web browser and clicking on a particular facility or customer, users can display a related business report (Figure 4). Since reports are generated dynamically from current solution data, all information is up to date and accurate. By tailoring the set of reports available to individual groups of users, United Sugars can make information access timely and efficient for everyone.

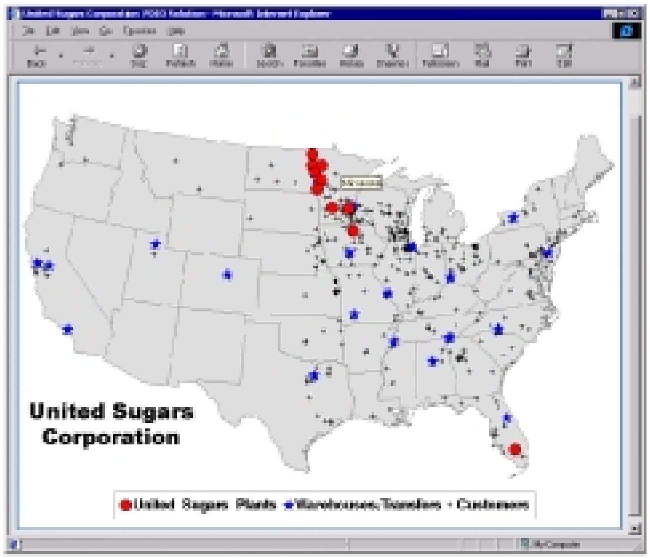
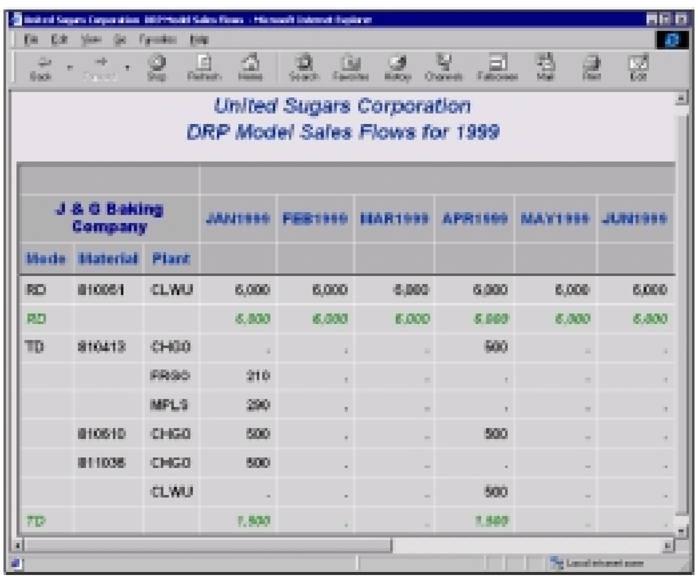


Figure 3: This Web page represents a prototype for reporting an optimal solution to the United Sugars' production, distribution, and inventory model. The US map indicates the location of the company's plants, warehouses, and customers. These are hotspots and are linked to additional information on the solution.



United Sugars Corporation
DRP Model Sales Flows for 1999

J & G Baking Company			JAN1999	FEB1999	MAR1999	APR1999	MAY1999	JUN1999
Mode	Material	Plant						
RD	810051	CLWU	6,000	6,000	6,000	6,000	6,000	6,000
RD			6,000	6,000	6,000	6,000	6,000	6,000
TD	810413	CHGO	.	.	.	500	.	.
		FRGO	210	.	.	.	.	.
		MPLS	290	.	.	.	.	.
	810510	CHGO	500	.	.	500	.	.
	811036	CHGO	500	.	.	.	.	.
		CLWU	.	.	.	500	.	.
TD			1,500	.	.	1,500	.	.

Figure 4: Selecting the customer, J & G Baking Company, on the main page (Figure 3) will produce a demand distribution report that shows the shipments from various plants and warehouses to that customer. The report indicates the mode of delivery, material delivered, and quantity by month, as calculated by the linear programming model.

Cameron & Barkley Company

Founded in 1865, the Cameron & Barkley Company (Cambar) of Charleston, South Carolina, is a distributor of industrial, electrical, and electronic supplies with facilities throughout the southeastern United States. Cambar's inventory contains nearly one-half million products. To become more profitable and competitive, Cambar wants to reduce inventory without sacrificing customer service.

In 1999, Cambar collaborated with SAS to explore methods for managing and reducing its product inventory and improving its demand forecasting. Through data analysis and exploration, we examined Cambar's product inventory and demand patterns and suggested several ordering rules. We then built a prototypical inventory-planning-and-management system that is nearing production status.

Cambar has a team of buyers charged with maintaining enough inventory to meet its strict levels of customer service. Traditionally, buyers relied on their knowledge of the market, intuition, and simple demand forecasts to determine stock levels. Because of the importance of meeting customer demand, they tended to overstock and to fail to meet the goal of four inventory turns per year established by management. More important, with their current software systems, buyers had to pay attention to all inventory items even though many were high-volume, low-cost items whose purchasing could easily be automated [Supply Works, Inc. and American Express 1997].

To automate these processes and to free up buyers' time for strategic decisions, we created the inventory replenishment planner (IRP) using the distributed architecture (Figure 1). The Web interface captures user interactions, saves the business information specified by the user, and finally builds a model of the problem to be solved. The model approximates lead-time demand and minimizes ordering and fixed costs subject to required service levels.

A user who goes to the IRP page sees a display (Figure 5) that allows them to specify service levels for three classes of inventory. The inventory is classified automatically and controlled via parameters entered in the page accessed via the Data tab. The user can specify the mean and variance of the order lead times for inventory items by class, and the type of algorithm used in calculating the inventory replenishment policies. Currently, two heuristic algorithms are available for calculating the policies [Tijms and Groenevelt 1984]. The first uses a normal or a gamma distribution approximation for the lead time demand distribution (the choice depends on the demand parameters), while the second uses the empirical distribution. An auto select option is available to let the software choose the most appropriate algorithm.

The Simulation tab is used to control a simulation of the calculated policies. The simulation samples from the user's historical demand data and drives the system using the calculated policies; the results enable the user to evaluate the effectiveness of the policies. With the Reports tab, the user produces reports on the input data, the simulation, and the policy (Figure 6).

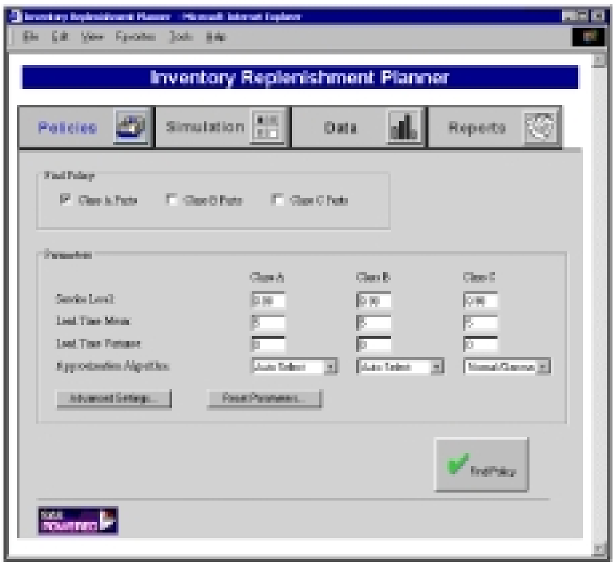


Figure 5: This Web page displays the four functional areas of the Inventory Replenishment Planner. The Policies tab enables the user to calculate replenishment policies for desired classifications of parts. The user may specify a target service level, order lead-time mean and variance, and the approximation algorithm to use.

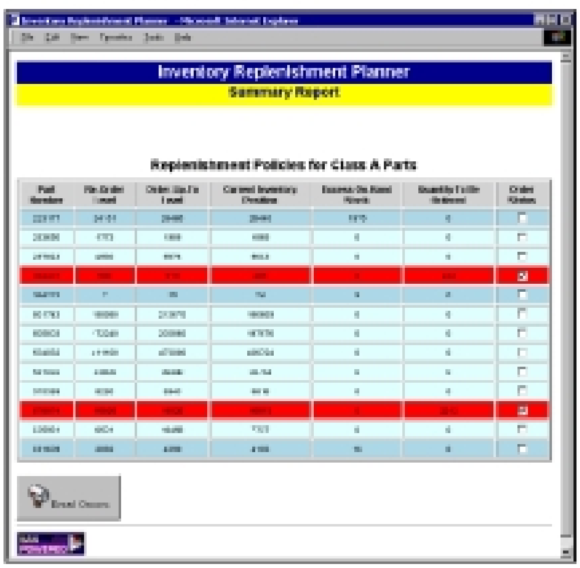


Figure 6: The Summary Report page displays the replenishment policies and the current inventory – including the re-order level, order-up-to level, inventory position, excess on-hand or to-be-ordered, and order status for each part. Parts with excess inventory are highlighted in turquoise (dark cyan), while parts that need to be ordered are highlighted in red. The Email Orders button triggers an e-mail message for those parts selected in the Order Status column.

Supplier Portfolio Optimization

In a partnership with Dun & Bradstreet, SAS developed an initiative called Supplier Relationship Management (SRM). SRM is composed of two software components, a viewer of procurement data called Procurement Vision (PV) and a procurement-planning component called Supplier Portfolio Optimizer (SPO).

Dun & Bradstreet provides a service called data rationalization to clean and normalize a company's purchasing data. This service includes making sure that all references to a given supplier are consistent, attaching a unique code to each supplier, attaching the UN/SPSC [Granada Research 1998] code to each unique commodity, and appending financial data about each supplier to the rationalized purchasing data. Dun & Bradstreet provides the data framework on which to base the SRM exploratory and analytic software.

The PV software is an exploratory tool that helps users understand the current portfolio of suppliers represented in the Dun & Bradstreet database. It helps buyers make strategic decisions about purchasing practices by identifying existing and potential suppliers and the products and services they provide. PV answers such questions as

— How much am I spending with a supplier?
— What is my risk with a specific supplier?
— How are my buying practices changing over time?

The SPO software is a planning tool that builds portfolios of suppliers that meet specific goals. It provides buyers with guidance for answering these strategic questions:

— How much should I spend with a supplier?
— How can I limit my financial risk?
— What should I buy from a given supplier?

Vendor selection and purchasing decisions are complicated because buyers must consider many elements simultaneously in making decisions [Weber, Current, and Benton 1991]. Dun & Bradstreet recommends that a company build its supplier portfolio considering such measures as the leverage in future negotiations, purchase consolidation with a select number of vendors, the demographics balance (woman- or minority-owned, industry, region, and so forth), and financial risk score. SPO uses these factors to optimize a supplier portfolio.

As input, the user of SPO specifies bounds on these and other measures to characterize the desired supplier portfolio. SPO finds the combination of vendors that meets or exceeds the desired characteristics and maximizes purchasing leverage. Leverage is defined as the percentage of a supplier's total business that the business accounts for. Thus, leverage is a measure of a firm's negotiating power with a supplier.

Since SPO accounts for demographic, financial, and other factors, the portfolio of suppliers that it identifies provides the highest leverage possible in future negotiations and also has the desired characteristics for all measures. It does this by consolidating purchases mainly from vendors for which the company is a major client while simultaneously creating a balanced and diversified portfolio. SPO uses a mixed-integer program to achieve these goals. Mixed-integer programming provides an almost perfect fit between modeling capability and the inherent features of a vendor selection and purchasing problem [Bender et al. 1985].

The SPO page (Figure 7) has four tabs: the Scenarios tab for managing solution output, the Filtering tab for preparing the data before portfolio selection is made, the Design tab for specifying the design parameters of the portfolio, and the Reports tab for reviewing the portfolio.

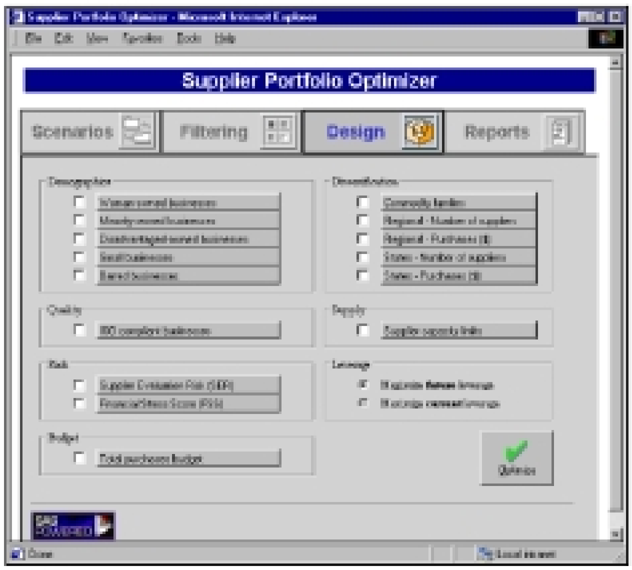


Figure 7: The Supplier Portfolio Optimizer main page provides an interface for building portfolios of suppliers. The Design tab is pictured here.

The input to SPO is the user purchasing data, rationalized and augmented by Dun & Bradstreet, and the design parameters for controlling the structure of the supplier portfolio. The design parameters capture the following kinds of rules:

— To satisfy a government regulation we should purchase at least one percent from minority-owned businesses.

— To promote quality, we should buy at least 25 percent of our supplies from ISO-compliant businesses.

— We should have 30 to 50 suppliers from our home state and buy between 15 and 20 percent of our supplies from them.

After setting the design parameters and pressing the “Optimize” button, the user will be queried for an e-mail address. The application builds the appropriate mixed-integer program (MIP), launches an optimization server, and sends this MIP to that server for solution. Using a data warehouse with 346 potential suppliers and 924 different commodities, the optimization server running under Windows NT (400 MHz and 128 MB of RAM) takes less than 5 minutes to solve problems of the order of 5,000 variables and 5,000 constraints. When the optimization server finishes, it sends an e-mail to the user. If an optimal solution to the MIP is found, a link to a Web-accessible solution report is embedded in the e-mail (Figure 8); if there are no feasible solutions, a brief note warns the user and suggests relaxing the model constraints as a possible course of action.

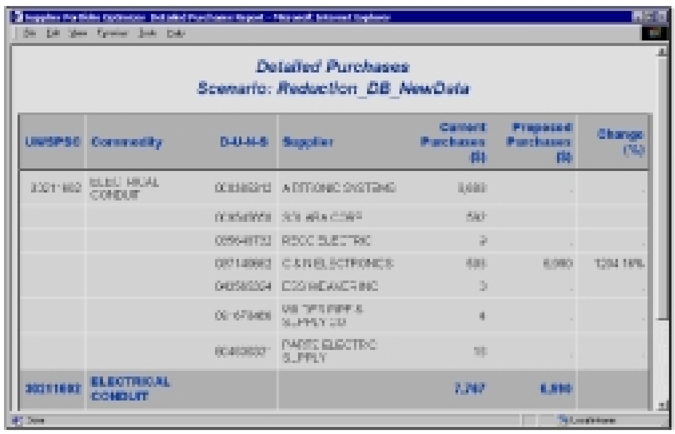


Figure 8: This Web page shows detail from an optimal supplier portfolio – the suggested purchasing for a single commodity, electrical conduit. Currently seven suppliers supply this commodity. The software recommends consolidating all purchases with C & N Electronics.

The output of the optimization process is a portfolio of suppliers that satisfies all the conditions specified by the design parameters and maximizes leverage in future negotiations. Numerous reports available summarize the optimal portfolio, giving such information as

— Detailed and aggregate purchases from each supplier,
— Detailed and aggregate purchases of each commodity,
— Purchases aggregated by region and state, and
— Interactive drill-down reports, for example, going from region to state to suppliers to commodities.

Data Visibility via the Web

Applications as illustrated in Figure 2 use summarized operational data from data warehouses to produce solutions that are also saved as data. For an analytical solution to be effective, its end-users must understand and have access to the process state represented by the input data and the solution saved in the data warehouse. In the following three examples, two companies make manufacturing production data available to a diverse user community.

Dow Corning

Dow Corning employs about 9,000 people and maintains 26 manufacturing sites to produce more than 10,000 products and specialty materials used by over 30,000 customers around the world. Its policy is to provide products and services that meet or exceed its customers' expectations through continuous improvement. Dow Corning developed an automated, Web-based application to enable staff and management to monitor product quality. The application generates statistical process charts and reports for all products. The company uses the application to monitor production sites and more than 10,000 different tests.

The application enables users to specify the particular physical or chemical property of the product that they would like to chart (say, viscosity) for any time period and production site. This generality makes it valuable to users worldwide. In addition, the application permits users to create personal profiles of product characteristics that they monitor regularly. With this feature, they do not need to reenter the same information day after day, and they can be notified automatically via e-mail when particular out-of-control situations occur. The application helps engineers focus on taking corrective and preventive actions instead of maintaining charts.

The user interface is a standard Web browser, linked to a specific URL. The first page, (Figure 9) allows the user to specify a production site and product on which information is desired. The user can search for product names containing a particular string, producing a page listing materials that meet the search criteria. Next, the user can specify parameters for a requested control chart, which the system then produces (Figure 10).

Figure 9: This Web page is the entry point for the Dow Corning SPC application for producing statistical process control charts. The user can select a plant and a search string to locate the types of materials made at that plant.

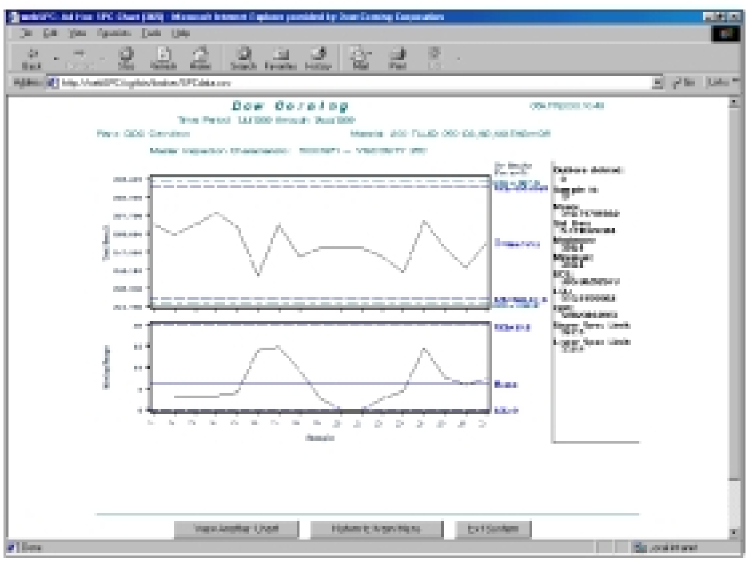


Figure 10: The Dow Corning SPC application produced this statistical process control (SPC) chart for the selected plant, material, and parameters.

Currently, about 100 quality engineers, manufacturing engineers, and statistical experts use the system. Eventually, Dow Corning expects to provide their customers direct access to data on the products they buy, reducing data reporting and management costs across the entire supply chain.

Data management is an important element of this application. Product-quality data are stored in a SAP AG R/3® system and are subsequently transferred to an Oracle® database (Figure 2). Data are extracted to a data warehouse where the application filters and processes it to produce charts on demand. As part of the filtering process, the system cleans the potentially messy data. The application contains statistical filters and logical business rules to test the data and to determine appropriate techniques for handling data with particular problems characteristic of the process industries. Also, the system alerts users to out-of-control situations before the plant makes a poor quality product, and it monitors the manufacture of virtually all Dow Corning products nightly. Soon, it will also monitor quality of all raw materials purchased for use in Dow Corning's products. The system dispatches an e-mail message containing a URL to the user if it detects preselected signals in the data. By automating data collection and the entire data-maintenance workflow, quality professionals have time to focus on the quality issues rather than data management issues. And the ability to share this information easily with customers over an Internet delivery platform enhances customer service dramatically.

Lockheed Martin Astronautics Division

The Space Launch Systems group of Lockheed Martin's Astronautics Division in Denver is responsible for designing, fabricating, and delivering two major space transportation systems: the Titan and Atlas missiles. Each program has its own managers, contracts, and customers. The two programs' managers needed a cost-effective way for over 400 middle- and upper-level decision makers to monitor performance of the production systems employed in building these missiles. These decision makers have four main questions:

— Where are the missiles in the build cycle?
— When will they be done?
— How will changes affect the schedule?
— What can be changed to improve the schedule?

Although the Astronautics Division makes a small number of missiles at a time, each missile is expensive to make, consists of about 10,000 parts (with bills of material that have about 20 levels), and has a long, complex fabrication cycle. Because of this complexity, the division does not reschedule once a production cycle has begun but focuses instead on tracking, understanding the current state, and reducing cycle-time.

The systems to address these issues meet the division's reporting needs by publishing reports to the Web [Rerecich 1998]. Much of this work was completed in 1998, and the systems continue to be integral parts of the Titan missile program. More recently, Lockheed Martin Astronautics used this structure for the Atlas program. The new projects have additional analytical content and include

— Control-point flow – identifying breakdowns in the bill-of-material structure;
— Past-due reporting – identifying work orders that are past due;
— MRP metrics – comparing the materials requirement planning (MRP) schedule to an optimal schedule;
— Work-remaining reporting; and

— Cycle-time reduction.
We discuss two of these projects: work-remaining reporting and cycle-time reduction.

Monitoring schedule status is the main focus of all the production-operations systems. In the work-remaining reporting project, the difficulties lie in accessing and transforming data so that they are consistent, scheduling the various tasks, and making the information visible. Lockheed Martin maintains schedule information in data sets with 40,000 to 250,000 records. Problems with data arise because data exist in diverse sources, have internal inconsistencies, and occur in large volume. Moreover, because some systems have over 400 users, the needs are diverse. For monitoring status, users are interested in the schedules, assembly structures, work remaining, load across factories, and resources available to support that load. They are also interested in obtaining information for various time horizons at many levels of detail.

Users access the system via a URL and can choose to monitor a project at multiple levels of detail and see the overall schedule of work remaining for each vehicle (Figure 11). Each of the bars in the chart can be a link to another Web page that shows detail on that particular vehicle. In turn, the resulting Gantt chart's bars are links to other pages that provide further detail on the selected subassembly, and so on. In this way, users can drill down to information at whatever level of detail they wish.

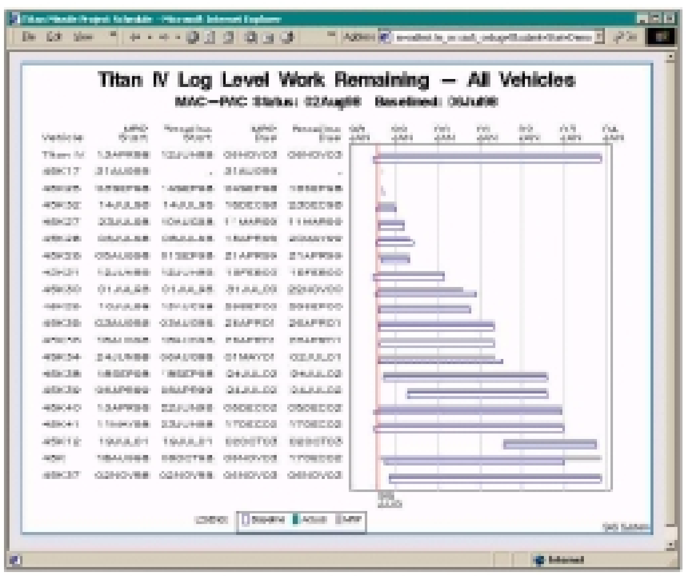


Figure 11: This Gantt chart displays the production schedules for the Titan IV missiles. Each of the schedule bars is a link to another Gantt chart that shows a high-level schedule for that particular vehicle.

In the cycle-time reduction project, Lockheed Martin sought to decrease the time taken to build Atlas launch vehicles. The heart of this project is to identify the longest paths in the activity network and to determine the activities for which to increase resources and engineering effort in order to speed up the production process. Project teams charged with reducing any paths that exceed a time span managers proclaim to be too long meet weekly and use a Web-page interface to find candidate paths and to generate reports detailing these paths. The input is the bill of materials (BOM). These BOM data describe the activity network for building an Atlas missile. The analysis determines the *n* longest paths in the schedule. Lockheed Martin built a simple and fast algorithm that identifies the path using off-the-shelf software, including the CPM procedure [SAS Institute Inc. 1999c]. They packaged this algorithm as a macro, which can be driven from a Web interface (Figure 12).

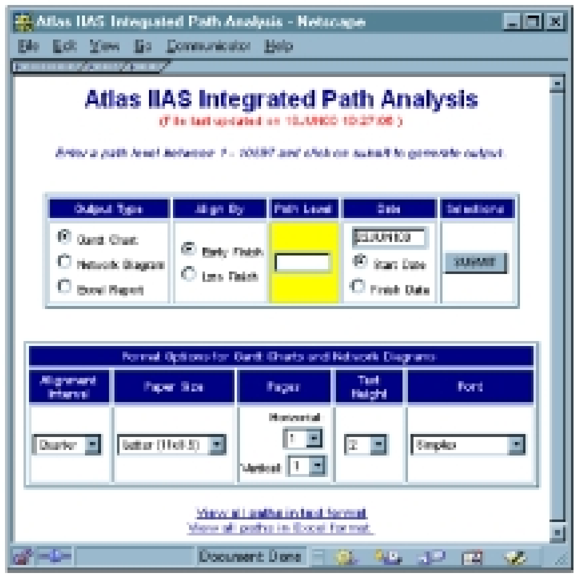


Figure 12: This Web page drives the Atlas II cycle-time reduction application. The user chooses the chart type, schedule type, path level (longest path) to find, date information, and information for the desired chart format.

These systems build Web pages programmatically via common gateway interface (CGI) [Weinman 1996] scripts that drive a server version of SAS. This configuration provides a concise representation of information and immediately captures any change in the underlying data without manual intervention.

Web Mining

Every customer-facing Web site has a fundamental and difficult challenge: how to use its logged activity data to better understand who enters the site and where, how the site is used, where and why people exit, and how the site's content and structure can be improved. Improvement usually means increasing the expected value of a visit to the site.

OR professionals have the tools with which to attack this challenge, and some are working on it, but most industry work is proprietary. This challenge is best addressed through Web mining, which can be understood in terms of a three-stage approach suggested by Parsa [2000]. The first stage, called *Web-log analysis*, summarizes raw Web-log data using basic descriptive statistics, reports, and charts. The second stage, called *click-stream analysis*, seeks to identify navigation paths through a Web site from the summarized Web-log data. The final stage is *data mining* of the summarized Web-log data augmented with data from other sources.

The challenge begins with making sense of the Web log, the raw history of user interactions with a Web site. Difficulties include enormous data volumes, incomplete information, and identification of sessions and users from the raw data. For instance, Web logs generate an essentially complete listing of user interaction at a rate of approximately 1 MB of data per 4,000-5,000 requests (or hits). A high-end server can receive 2 million hits per day and produce a daily Web log of 0.5GB. Even the preliminary step of organizing Web-log data by downloaded pages is not necessarily easy, since multiple objects (audio files, CGI calls, images, text, etc.) are downloaded to display a single Web page, with each download recorded separately in the log.

The goal of Web-log analysis is to provide accurate snapshots of the activity of the Web site by cleaning and reorganizing the Web log to extract page view information; to identify user sessions; and to extract any query keywords used by the visitors to the Web site. Reorganizing Web-log data by individual user sessions is difficult because the data are confounded by erratic user task continuity (users may switch tasks or leave their computers), multiple persons sharing a computer without ending a session, and the inconsistent enabling of cookies (see Cookie Central). The results of Web-log analysis are descriptive statistics and graphical reports useful for tuning the Web site design and for tracking and evaluating advertising campaigns and the set of Web pages visited by session and unique visitor. Two companies with successful systems for Web-log analysis are AutoTrader.com, a major auto-trade Web site, and Washingtonpost.Newsweek Interactive, a news-media and electronic publisher.

Given the Web pages visited by session and unique visitor, the next difficulty is to identify the navigation paths that users take through the Web site; for example, what are the most common paths used by the customers who buy or leave the site before buying? Another objective is to determine associations among various pages in a Web site. For example, if users who visit product A's page also frequently visit product B's page, then there is a potentially valuable cross-selling potential revealed from this association. Based on this information, one may be able to modify the site for a customer (or type of customer) to highlight the products or services most likely to be of interest. Ostdeutscher Rundfunk Brandenburg, the German public broadcaster, is currently working along these lines.

One difficulty encountered is that some of the user navigation paths may not be complete. In fact, incomplete data is to be expected unless users' computers have been fitted with special monitoring software. The main reason for this is the wide use of content caching. Caches, which are storage devices between the browser and the server, store Web content – images and HTML pages – and serve it to browsers with the aim of reducing browser response time, reducing load on content servers, and reducing Internet bandwidth consumption. However, when content is served from a cache, it is not recorded in the server log. One challenge to the OR practitioner is to develop techniques to reconstruct navigation paths that have missing links. A second challenge is to improve the Web site design so that users can find the most important content with minimal effort. Approaches are emerging to address these problems [Spiliopoulou 2000].

The greatest benefit can be realized by using data mining techniques on the summarized Web log. Ideally, the summarized Web log should be augmented with historical purchasing data, user profiles, and data obtained from third-party sources. An application server can generate score functions using data mining models offline. Then, these scores can be used in real time to generate dynamic content and guide users toward appropriate pages in the Web site. The offline models can use several standard data mining techniques [Two Crows Corporation 1999]. For example, discovering association rules of the sort mentioned above with cross-selling potential uses a technique called market basket analysis, which uses Bayes Rule to compute conditional probabilities of association. Similar opportunities exist to use classification, which can segment customers into fine-grained groups for the purpose of making customized offers or for creating a more personalized Web browsing experience. Standard implementation techniques include discriminant analysis, clustering, Kohonen networks, and decision trees, which are supported in data mining software such as Enterprise Miner [SAS Institute 1999a] and others (see KDnuggets).

Web mining presents a significant opportunity for OR professionals. Most organizations involved in Web mining are at the first stage. OR practitioners are in a good position to understand the applicable technologies, and hence to guide their organizations towards choosing and developing the best technology to address their business problems.

Conclusions

At SAS, we are working to make intranets, extranets, and the Internet more useful for delivering analytical content and exploiting the numerous opportunities for decision support in a Web environment. Opportunities range from solving tactical production, distribution, and inventory planning with a multitiered architecture; to providing visibility of product quality or process state to broad and diverse groups of users; to providing insight into web usage and ways of improving Web impact. Delivering decision support functionality over the Web improves accessibility, ensures timely information, and results in institutionalized decision support capabilities.

Acknowledgments

We thank the following people who helped us in many ways to prepare this paper: Gehan Corea, Radhika Kulkarni, and Tugrul Sanli of SAS Institute Inc.; Jamie McGee of Dow Corning Company; Roger Perala of United Sugars Corporation; Dave Jennings, Chuck Rissmiller, and Ken Robak of Lockheed Martin; and Greg Lockard and Bob McQuillan of Cameron & Barkley Company. We also thank the editors, Arthur Geoffrion and

Ramayya Krishnan, and several anonymous reviewers for their many constructive suggestions.

References

Bender, Paul S.; Brown, Richard W.; Isaac, Michael H.; and Shapiro, Jeremy F. 1985, "Improving purchasing productivity at IBM with a normative decision support system," *Interfaces*, Vol. 15, No. 3 (May-June), pp. 106-115.

Cohen, Marc-david and Kelly, Charles B. 1999, "ASAP: A methodology for strategic and tactical planning," Working paper, SAS Institute Inc., Cary, North Carolina.

Granada Research 1998, *Using the UN/SPSC: Why Coding and Classifying Products is Critical to Success in Electronic Commerce*, Granada Research, El Granada, CA.

Parsa, Ismail 2000, "Mining Web data for personalization solutions: Course notes," SAS Institute Inc., Cary, North Carolina.

Rerecich, Art 1998, "*W - O - W: Data warehouse, SAS/OR and the Web*," *Proceedings of the 23rd Annual SAS Users Group International Conference*, Nashville, Tennessee, pp.567-573.

SAS Institute Inc. 1999a, "A SAS Institute white paper: Finding the solution to data mining: A map of the features and components of SAS Enterprise Miner software," SAS Institute Inc., Cary, North Carolina.

SAS Institute Inc. 1999b, *SAS/OR User's Guide: Mathematical Programming, Version 8*, SAS Institute Inc., Cary, North Carolina.

SAS Institute Inc. 1999c, *SAS/OR User's Guide: Project Management, Version 8*, SAS Institute Inc., Cary, North Carolina.

Spiliopoulou, Myra 2000, "Web usage mining for Web site evaluation," *Communications of the ACM*, Vol. 43, No. 8 (August), pp. 127-134.

Supply Works, Inc. and American Express 1997, "Open buying on the internet," OBI Consortium, Cupertino, California.

Tijms, H. C. and Groenevelt, H. 1984, "Simple approximations for the reorder point in periodic and continuous review (s, S) inventory systems with service level constraints," *European Journal of Operational Research*, Vol. 17, pp. 175-190.

Two Crows Corporation 1999, *Introduction to Data Mining and Knowledge Discovery, Third edition*, Two Crows Corporation, Potomac, Maryland.

Weber, Charles A.; Current, John R.; and Benton, W. C. 1991, "Vendor selection criteria and methods," *European Journal of Operational Research*, Vol. 50, pp. 2-18.

Weinman, Williams 1996, *The CGI Book*, New Riders, Indianapolis, Indiana.