Paper 269-2009

Business Intelligence on the Grid: A Customer Perspective

Phil Hatfield, ISO Innovative Analytics, San Francisco, CA

ABSTRACT

As part of a continuing effort to increase its capabilities in predictive modeling and analytics, ISO Innovative Analytics is building an “advanced analytics platform.” The system centers around SAS Enterprise Miner running in a grid environment. This article discusses the process the company went through in making the software and hardware choices for the new platform and the experience of implementing the platform for a team of more than 30 predictive modelers and analysts.

BACKGROUND

ISO is a leading source of information about risk. We deliver a wide variety of products and services that help customers understand and manage their risk. From its roots as a consolidator of information and actuarial analysis for the property/casualty insurance industry, ISO has broadened its scope to include risk-assessment products for the healthcare and mortgage industries and for the risk-management function in all industries.

Advanced analytics has always played a large role in ISO's product offerings, and in recent years we have participated in the rapid development and change in analytical techniques. In 2005, the company formed a division called ISO Innovative Analytics (IIA) to develop new analytic capabilities and to enhance our products and services for customers in many fields.

Based in San Francisco, IIA is developing better ways for our customers to measure, manage, and reduce risk. The unit is a key part of ISO's commitment to produce state-of-the-art analytic products that meet the needs of our markets.

In the early days of IIA, the computing platform was fairly simple. Analysts worked on individual workstations using whatever tools they were familiar with. Base SAS and SAS/STAT were the most widely used tools, but some analysts also used other applications and programming languages for some functions.

The computing environment evolved as more analysts came on board. Eventually, we had a dozen common analytic servers accessing about 6 TB of data on an iSCSI storage area network (SAN).

As the organization grew, the analytic demands began to outstrip the infrastructure available in the San Francisco office. We decided that we needed to replicate the IIA systems in ISO's home office in Jersey City, New Jersey. The home office has a modern data center with a large support staff and a fiber-channel SAN already in place. We accomplished the move in the first half of 2008.

But although the move of IIA's computing platform had some immediate advantages, it didn't solve the underlying problem. Our system had evolved over time and was not easily scalable. So, while we were moving our platform across the country, we also started designing the follow-on system.

NEEDS

At the beginning of 2008, IIA's staff numbered about 20 people, and we expected that to grow to 35 by the end of the year. We knew that the expansion would not end in 2008. Clearly, we needed an "advanced analytic platform" that would scale to accommodate substantial growth.

Also, two-thirds of the staff are M.S.- and Ph.D.-level statisticians and mathematicians doing data mining and predictive modeling. They are highly skilled and valuable resources, and they have a low tolerance for computing tools that hamper their work. We realized that we needed a sophisticated, modern system that would increase the quality of work and the productivity of the analysts. And we knew that the availability of such a platform would aid our efforts to recruit and retain the best analysts.

Another requirement was our increasing need for standardization and documentation. The ad hoc process and myriad tools — acceptable and even encouraged in our original research-lab environment — would be unacceptable and unmanageable for a larger, more complex, more service-oriented organization.

In short, IIA needed a common platform and tool set that would meet the needs of our analysts and our customers. The fact that we were consolidating and upgrading our tools also made it feasible to look at new capabilities, such as data visualization and model management.

ANALYTIC-TOOL SELECTION

Since we were basically starting from scratch with a new system, we realized that we could allow the selected analytic-tool portfolio to drive the design of the computing environment. Whatever tool or set of tools we chose would determine the architecture of the computing platform.

We started by creating a task force to select the analytic tools. The task force included several members of IIA management and the analytic team. A systems architect from the ISO IT staff coordinated the work.

The task force made an initial scan of the analytic-tool marketplace to create a comprehensive list of possible tools. Most of our analysts were already familiar with many different tools, so it wasn't difficult to compile the list and cull it down to three prime contenders. Many of the products that *didn't* make the cut were useful for only a narrow scope of analytic problems, while others didn't work well with the large data sets we typically require.

The task force contacted the three finalists and solicited more detailed information and demonstrations. At the end of an extensive process, SAS Enterprise Miner was the clear choice.

Most of our analysts were familiar with other SAS tools. And choosing Enterprise Miner preserved most of the SAS programming and model development that we had completed over the previous three years. SAS also offered some desirable new capabilities, including data visualization (through SAS JMP) and model versioning and monitoring (through SAS Model Manager). The fact that Enterprise Miner could run in a grid computing environment was also a major positive point.

THE GRID

The initial design of IIA's system called for it to accommodate 20 concurrent users, but we knew that wouldn't be the end of it. IIA was continuing to grow, and we also envisioned the system as a corporate resource open to analysts from other ISO divisions and business units.

Informal conversations with people who had built analytic systems at other companies revealed a common story. After buying a large, expensive machine to run their analytics applications, they quickly outgrew it and had to buy another one.

A grid computing environment — applying several computers to a single problem at the same time — offers the advantage of more incremental upgrades with less expensive hardware. It's a matter of "scaling out" rather than "scaling up."

Grid computing also offers performance benefits, particularly with SAS Enterprise Miner. Unlike a SAS script — basically a linear style of program — Enterprise Miner uses a graphical data flow. The programmer can design many portions of a program to run processes in parallel. Further, many of the procedures within SAS are "grid-enabled." So a single SAS procedure can take advantage of the multiple processors available within the grid environment to spawn multiple parallel processes on the grid without any explicit coding for a multiprocessor environment.

In consultation with SAS, ISO's IT architects designed a grid system that included eight dual-processor, dual-core IBM blade servers as the initial implementation.

OPERATING SYSTEM

SAS Enterprise Miner runs on a variety of operating systems. Therefore, we decided on the operating system based on factors other than the chosen tool. Although ISO has a few Linux servers for specific applications, we are primarily a Microsoft Windows shop, and we have many more support staff for Windows servers than for Linux.

To implement a Linux system would have meant getting extra support, while a Windows system could take advantage of the depth of expertise that already exists at ISO. Furthermore, SAS Enterprise Miner can use the existing Microsoft Active Directory for authentication. Linux would have required LDAP to synchronize with ISO's Active Directory — another new capability that we would have had to build into the project. All those considerations meant that Windows would allow IIA to get better support and a shorter time to implementation.

DATA SYSTEM

The SAS Enterprise Miner server implementation uses a more structured way of accessing data than the stand-alone versions of SAS that the analysts had used previously. The server implementation can do just-in-time data pulls from a variety of sources — allowing easier centralized management and reuse of data. As our analytics group grows, as we take on more projects, and as those projects become more complex, the ability to manage data in a more structured manner is becoming ever more important to us.

So, in conjunction with implementation of the server-based SAS applications, IIA is implementing a new data-management platform. Again taking advantage of ISO's existing in-house expertise and vendor relationships, we chose Microsoft SQL Server 2008 as the DBMS. Three environments — development, testing/staging, and production — will make up the system.

We picked the three-environment model because of the unique needs of predictive analytics. As we begin work on a new model, we need to examine a large volume of data from many potentially relevant sources. As the research continues, we may eliminate individual variables or even entire data sources that we find less useful to creating an accurate model. The analysts will use the development environment for assembling analytic data sets. The rather loose structure of the development environment will let analysts easily combine and eliminate data from various data sources as the individual project requires. By contrast, the production environment will house more permanent data sets that require high availability, good documentation, and regular updates. We will reuse data from the production environment in multiple projects over a long period of time. And that data will support our implemented models.

The testing/staging environment is primarily a way station between the development and production environments. In the testing/staging environment, we will prepare data for loading into the production environment. We'll also test new data programs, complete documentation, and perform other necessary tasks.

The initial implementation calls for approximately 2 TB of storage allocated among the three environments. That will grow to 18 TB over the first few months, as IIA migrates data stored in SAS data sets and flat files from the old system to the new one.

IMPLEMENTATION

Building our advanced analytics platform was a fairly complex project for ISO — and for SAS — and the implementation required a great deal of coordination between the two organizations. Some of the most important and useful parts of the process were the meetings where we gathered user and technical requirements.

Before implementation, ISO and SAS had worked together to do a high-level design of the system. For example, based on that early effort, we knew going in that we would be implementing SAS Enterprise Miner on a grid architecture that would include eight grid servers divided between production and test systems. But the SAS system still offered many configuration options and other details that we needed to work out. The SAS project manager and system engineers held a two-day requirements-gathering meeting with several of the analysts to get a better understanding of how IIA would be using the system. That session helped to refine several parts of the architecture and optimize the implementation to IIA's unique situation.

The team then held a technical-requirements session with the ISO project team. Working together, we finalized the system architecture, determined the proper security implementation, and worked out the network architecture. The meetings were crucial to setting up the system in a way that would be appropriate for IIA's needs.

ISO ordered the new hardware, set it up, and tested it. We also built the servers. SAS then sent an engineer who worked on-site with the ISO technical team to install all the applications and test them out.

TESTING

Today, the system is nearing the end of its acceptance tests. IIA created a comprehensive plan designed to test all the individual components of the system in the way we intend to use them after implementation. The plan also tests the system's ability to handle multiple concurrent sessions of heavy analytic computing.

We learned our most important lessons during acceptance testing. First, IIA had planned on two weeks of acceptance testing. We discovered that we clearly needed more time for a system this complex. In the end, acceptance testing required two months.

Second, we learned that proper documentation is critical. We found that we must document requirements, issues, procedures, and assumptions — to make sure that the items we resolve stay resolved.

CONCLUSION

Today, the new SAS analytic system is just about ready for production. To get from requirements gathering to this point has required six months of hard work by the SAS team, the ISO technical team, and the IIA analysts. Along the way, we've learned a lot, and we've overcome some hurdles, as you'd expect with such a complex and technologically advanced platform. The results look very promising.

Was it worth it? The ultimate test will come soon, when IIA puts the platform into production and 30 statistical analysts log on for their daily work of creating predictive models. Tune in next year.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Phil Hatfield
Enterprise: ISO Innovative Analytics
Address: 388 Market St, Ste. 750
City, State ZIP: San Francisco, CA 94954
Work Phone: 415-276-4106
E-mail: phatfield@iso.com