Paper 113-26

Data Warehousing at Baylor University: The Second Year

Phil Rhodes, Baylor University, Waco, TX
Sue Herring, Baylor University, Waco, TX

ABSTRACT

The office of Institutional Reseach and Testing at Baylor University began a data warehousing project in the spring of 1999. First year activities included data modeling and software selection and customization, along with the creation of several data marts. The second year of the project has resulted in a large increase in the number of users, along with the number of data marts available to users. This paper presents a brief overview of the data warehousing process using SAS® software, as well as warehouse design considerations specific to the academic environment. Additionally, important lessons learned throughout the process will be presented in the form of a 'Top Ten' list.

INTRODUCTION

The office of Institutional Research and Testing (IRT) at Baylor University is charged with providing timely and accurate information for decision support to the University administration. This information ranges from student enrollment and retention to faculty workload and compensation, as well as financial data to support tutition and fees rates setting and the University's five year budget planning process. A data warehousing project was begun in the spring of 1999 in order to consolidate this information and to provide a single point of access to the data. The responsibility for designing and implementing the warehouse was given to IRT, due to the broad-based knowledge IRT had of the University's data. Using The SAS System®, a data warehouse was designed, built, and delivered via a web interface accessible by Univesity administrators. During this process, some important new lessons were learned, and some old lessons reinforced.

THE DATA WAREHOUSE

Academic institutions often have an advantage over many businesses when it comes to building a data warehouse. Many times these institutions have years of data extracts, frozen on the twelfth class day of each semester. These frozen datasets are used for official reporting purposes. They can also form the core of a new data warehouse.

Baylor University's formal data warehousing project began early in the spring of 1999. The staff of IRT began to research data warehousing methods and best practices, as well as contacting possible users and supporters of the data warehouse. Data modeling was begun, starting with the student data that was the data our office knew best. At the same time, a committee was formed to evaluate data warehousing software. The SAS System was chosen because it was the only end-to-end solution available, as well as being the tool already used by IRT. This enabled the use of existing programs and datasets to build the warehouse, avoiding a costly and time-consuming conversion process.

Once the software was chosen and installed, the Graduate School was chosen as the first subject area in the warehouse. This choice was made for two reasons: first, the Dean of the Graduate School was new to the position, and was requesting a large number of ad hoc reports from IRT. Second, the President of the University had recently made the decision to strengthen the University's graduate programs in an effort to move up the Carnegie classification scale. IRT staff members began meeting with the Dean to determine what type of questions he needed answered, as well as to find out what information could be provided now. As questions were identified by the dean, data to answer those questions was located and loaded into the warehouse. The SAS/Warehouse Administrator® product was used to load the data and manage the large amount of metadata generated by this phase of the project.

The second year of the project has seen an expansion in the number of users of the warehouse, and the deployment of additional data marts. The first data mart added after the Graduate School was to serve the needs of the office of Admission Services. This office includes the recruiting arm of the university, and it was here that the data mart was focused. The recruiters can now identify students with a high potential to enroll before they visit a high school and set up private meetings with these students. Additionally, the director of recruiting can now track progress toward recruiting goals at the click of a button, a process that used to take several weeks.

The second data mart added focused on undergraduate students. This allowed Deans and program directors to evaluate programs over time with regard to quality of entering students, retention, and graduation rates. Additionally, work was begun with the School of Business, the School of Education, and the College of Arts and Sciences to identify other questions that were not answered by the current version of the data mart.

The next phase of the project will be to incorporate Human Resources data into the warehouse. This will involve the use of the current interface as well as the SAS HR Vision® product for some users.

LESSONS LEARNED

During the first two years Baylor University's data warehousing project, many lessons were learned - some old, some new. Here, in reverse order, are ten of the most imporant.

LESSON 10: TOOLS ARE JUST TOOLS

There are many tools available for data warehousing today, from administrative tools to presentation tools. From the perspective of a data warehouse, perhaps the most important tool is the user interface. This is the 'face' of the warehouse to the world, and is what the user will see every time they use the system. If the interface is poorly thought out, it won't matter how elegantly the warehouse has been designed - the users most likely will not use the system. The tools are only important in so far as the help the users get the information that they need.

The vendor sales representatives gave beautiful presentations and many talked about using their software "right out of the box". It is possible that this could be done, but in 99.99% of all cases, the interface software will have to be modified before it is used (see Lesson 5). This requires that whatever software is chosen must be flexible enough to fit the user's needs, not the other way around. The SAS/Intrnet® package, and in particular the SAS/MDDB® Report Viewer, fit this need perfectly. From 'canned' static reports to data-driven, user-customized analysis, it was possible to tailor the interface to the user's needs.

LESSON 9: IT TAKES LONGER THAN YOU THINK

The primary sponsor of the first phase of the data warehouse (as well as the first 'test' case) was the Graduate School. The Dean was a researcher and was able to formulate some specific questions that he needed answers to. However, he had been told in the past that the data was not available to answer those questions. After some investigation, IRT determined that the data was indeed available to answer many of the questions the Dean had posed, it just wasn't easily accessible.

A timetable was agreed upon, including a tentative date for the warehouse to be operation. Unfortunately, the first three or four intermediate deadlines were not met, although progress was made. Several problems were due to delays in software delivery. Several new SAS modules, including SAS/Intrnet, SAS/MDDB Server, and the SAS/Warehouse Administrator, were late in arriving. After installation, there were additional issues with configuration and learning to use the software. There were also data problems. The more the data sources were examined, the more problems we found, even in 'clean' archived data (see Lesson 8). This all took time.

Additionally, working with the users took longer than expected. Even though the Graduate Dean was eager and willing to do whatever it took to bring the project to fruition, it was difficult to develop a clear understanding of what he wanted and where to locate the data to get what he wanted (see Lesson 7). A case in point was stipends and tuition remission for graduate students. After discussions with the Dean, we extracted, cleansed, and summarized the data on a student level. When he was given access to the data, he informed us that he really only needed it on a departmental level. The effort was not wasted, however, as he later decided that he needed the data on a student basis as well.

LESSON 8: CLEAN DATA IS STILL DIRTY

Data quality has been a primary concern in IRT for many years. As we create extract files each semester, many pages of error reports are generated and sent to the appropriate offices for correction before the final extract is frozen. However, these error reports were not all-inclusive. They were limited to the relationships that IRT knew of, or could deduce from the data themselves. When data was loaded into the warehouse, it became evident that the data was not as clean as had been thought.

Data that were clean for a specific term did not appear so clean when viewed across multiple terms or years. Inconsistencies appeared, including changes in data usage or values. Formatting changes in past terms made data comparison difficult. Some of these problems were well-known by IRT staff and accommodations were made when longitudinal studies were conducted. However, the datasets themselves had not been corrected. The alterations were temporary inside specific programs or projects. Finding these alternations and making the data consistent and accurate for the warehouse was a large part of loading data.

One of the larger issues that arose had to do with organizational changes. An academic department might move from the College of Arts & Sciences to Institutes and Special Studies, or a particular academic program might change from one department to another. In order to give academic department chairs access to the correct data, all of these changes had to be tracked down and verified. Along with this, consistency with previously-published information had to be maintained. If the warehouse was not consistent with other data published by IRT, users would not trust the data.

LESSON 7: COMMUNICATION 101 IS REQUIRED

One of the reasons that delays were experienced in the project was that communicating with end users was difficult. In essence, end users and data warehouse developers (or any software developer, for that matter) do not speak the same language. As the developers, IRT staff spoke in terms of data elements, DBD elements, and processes and procedures. The managers with whom we worked spoke in more global terms. They know generally what they need and want, but they do not always know what data is required to get the answers. Having staff who understand the data and also the language of the managers is a key to the success of any data warehousing project.

It was also important that the concept of data warehousing be understood by the prospective users. The primary data in the warehouse was historical, not real-time, and it was to be used for evaluation and trend studies. It was never intended to be used for real-time ad hoc reporting. When talking to some potential users, however, it was very difficult to communicate the differences between the two. Some universities have built 'data warehouses' that are daily copies of their transactional systems to be used for reporting purposes. This was not the design path followed at Baylor, and it was imperative that the users understood this to avoid unrealistic expectations.

LESSON 6: DOCUMENTATION IS ESSENTIAL

'Document, document, document' is a lesson that is easily forgotten, especially on a data warehousing project with so many other tasks to be accomplished. However, it tends to be put off or thought of as unimportant. Due to past experience, our office management is insistent on cross-training and multiple people being able to perform specific tasks. This is only wise but when learning new software and working with users and writing new code and developing new web pages, it is a tendency to just do it rather than to document how someone else can do it also.

In addition to documenting processes and procedures, the SAS/Warehouse Administrator tool was used to document the metadata - where individual pieces of data came from, the formats, conversions, etc. Data dictionaries were written for users to explain uses and limitations of individual data elements. This provided security so that maintenance could occur even if the original staff were not present. Because of the magnitude of the project, it also allowed the staff themselves to remember what had been done.

LESSON 5: INTUITIVENESS IS WORTH THE EFFORT

During one of the initial user interface demonstrations for the Graduate School, the Dean stated that he didn't like the interface. His exact words were "It is not intuitive enough." This phrase has been repeated many times since that day, and has become a yardstick by which we measure any new tool or report that is added to the warehouse interface. There were several reasons 'intuitiveness' is important:

Warehouse users are, for the most part, infrequent users. Managers have a question and go to the warehouse for the answer. It may occur once a month or once a semester, but it more than likely will not be daily. Infrequent users need something that is simple to use. If it requires a lot of thought or training or lots of things to remember, then managers will not use the warehouse. They will call IRT as they have done in the past. No one is being served.

If the interface requires a great deal of training to use effectively, this requires both time and commitment from the users. Additionally, someone must be responsible for preparing the traing tools, conducting the training sessions, and serving as the 'help desk'. If the interface is developed so that training is unnecessary, then the effort saved on training can go into improving the warehouse.

Finally, the Dean also stated that if the tools are hard to use, he would not use them. Most users would not make the effort to move around a complicated, convoluted web site if there were another choice. Intuitiveness and ease of use are more than worth the effort put into them.

LESSON 4: ANSWERS BEGET MORE QUESTIONS

When managers finally have tools in their hands that allow them to examine data and search for relationships, more questions will surface (see Lesson 1). During the modeling process, no one can identify every question that needs to be answered and thus, every piece of data that needs to be available. The magnitude of the project just continues to grow. This is not a bad thing, even though it requires requires more work on the part of the implementation team. While software engineers may cringe at this ‘feature creep’, it is a strong measure of the success of the warehouse. As users explore data, additional questions come to light that they had not thought of before. This only happens if they are actually using the warehouse. A data warehouse thatnever moves beyond its initial questions is a data warehouse that is not not being used to its full potential.

LESSON 3: SUCCESS IS CONTAGIOUS

A data warehouse is most successful if the users themselves do the selling. While IRT staff made many presentations to potential users, a large number of users called us asking for access after speaking with other users. Additionally, when supervisors use the warehouse and discover something about a department, they often confront that department head with the findings and request an explanation. Shortly thereafter, the department chair calls to get access to the warehouse. They want see the information used to evaluate their program.

The reverse of this lesson is also true: success is contagious, but so is failure. Difficulty using the interface can sink a data warehousing project quickly, as can inconsistent data. Both of these problems are quickly shared with other users and can bias them against the warehouse, even if it could benefit them.

LESSON 2: COMMUNICATION 201 IS HIGHLY RECOMMENDED

Communication 101 was required, and Communication 201 is highly recommended. One cannot say that a project will fail without Communication 201, but success will come much easier with it. Communication 201 involves staying in touch with users and sponsors even after datamarts and reports are created and are being used. Continual feedback and refinement of data and tools is necessary for a warehouse to continue to grow. The job doesn’t stop with the deployment of a data mart (see Lesson 1).

One way to keep communication lines open is via email. An email listserver allows quick dissemination of information to users, including updates and the availability of new data. Any of the users can use the listserv to pose questions or share success stories. This has proven to be an excellent way to keep the warehouse in front of the infrequent user until habit takes over and they begin to think first of going to the warehouse.

LESSON 1: IT AIN’T OVER ‘TIL IT’S OVER

…and it’s never over. It has been said many times that a data warehouse is a journey and not a destination. Between maintenance, cleansing data, updating software and documentation, and bringing in new users, the data warehouse is never finished. This is a major difference from the typical software development project. Starting a data warehousing project with the idea that it will be completely finished is a mistake, and may prevent the warehouse from growing and becoming completely successful.

CONTACT INFORMATION

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

Phil Rhodes
Baylor University
PO Box 97032
Waco, TX 76798
(254) 710-8860
Fax: (254) 710-2062
Email: Phillip_Rhodes@baylor.edu

Sue Herring
Baylor University
PO Box 97032
Waco, TX 76798
(254) 710-8836
Fax: (254) 710-2062
Email: Sue_Herring@baylor.edu