Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

Data Warehouses in Public Schools: Lessons Learned in an On-Going Implementation

Hemant K. Jain
University of Wisconsin - Milwaukee, jain@uwm.edu

Derek L. Nazareth
University of Wisconsin - Milwaukee, derek@uwm.edu

Robert Nelson
Milwaukee Public Schools, nelsonrw@mail.milwaukee.k12.wi.us

Data Warehouses in Public Schools:
Lessons Learned in an On-going Implementation

Hemant K. Jain, University of Wisconsin-Milwaukee, jain@uwm.edu
Derek L. Nazareth, University of Wisconsin-Milwaukee, derek@uwm.edu
Robert W. Nelson, Milwaukee Public Schools, nelsonrw@mail.milwaukee.k12.wi.us

Abstract

Over the last several years data warehouses and data marts have been increasingly used in commercial business enterprises to support decision making. However, this technology has not been as widely adopted in not-for-profit enterprises and public education in particular. Public education is coming under increased scrutiny as grass-roots support for spending is on the wane and accountability expectations for resource utilization are increasing. Data warehousing can effectively support data-driven decision making and represents a logical area for investment. This paper describes a partnership between a major metropolitan university and a large urban school district to deploy a comprehensive student-centered data warehouse to support educators and administrators in their day-to-day operations. It addresses managerial, technical, and political issues faced in the implementation.

Introduction

Interest in data warehousing in the corporate setting in the last decade has exhibited spiraling growth since the early formulation of the ideas by Inmon (Inmon 1992; Inmon et. al. 1996). With closer integration of information systems and the need to improve performance at all levels of operation, data warehouses are an attractive prospect in that they offer all levels of knowledge workers the opportunities to improve their day-to-day operation and decision making. Data-driven decision making allows knowledge workers to utilize their resources in a more effective manner.

While there are many instances of successful data warehouse implementations in the corporate setting, the degree of adoption in not-for-profit organizations is less pronounced. In the public education sector, the extent of data warehousing use lags further behind. Typically, these organizations function within a bureaucratic structure that is constantly buffeted by political changes. Many of these organizations employ an IT infrastructure that dates back several years and lack the appropriate tools to assist in the day-to-day decision making that would improve their effectiveness at the knowledge worker level.

In an era where more innovative approaches to education are being deployed, such as charter schools, school choice programs, experimental scheduling and assessment strategies, the need to properly judge the effectiveness of these initiatives assumes greater importance. This paper describes the efforts to create a data warehouse to help address these questions at Milwaukee Public Schools. It describes the initial championing for the technology, data modeling sessions, data warehouse design, product selection and purchase, as well as operational hurdles regarding data extraction and cleaning in an effort to provide high quality data to the users. The warehouse was implemented in a phased manner, and scale-up problems are also touched upon. A wealth of insights have been gained by all stakeholders, including end-users, champions, sponsors, developers, and consultants. Of some interest is the way in which the data warehouse is being used or projected to be used, given that many of these uses were not originally planned or sold to the end-users.

Decision Making at the School District

Milwaukee Public Schools is a large urban school district with over 105,000 students, in over 150 schools. In recent years, the school district has invested a sizeable amount of resources in technology to assist in improving instructional effectiveness. While some of this technology is directed at hands-on student use, other components are aimed at improving the day-to-day operation of the schools and the central administrative unit. The strategic plan for technology for the district (MPS 2000) includes technology initiatives like distance learning, wireless communications, internet access, and multimedia, in an effort to create a connected community of learners. The vision for the data warehouse is that it will provide a basis for longitudinal analysis and data mining to facilitate operational decision making in the school district.

Decision making at the public school system occurs at several levels. Budgetary and planning decisions tend to be performed at the Central Services location. These decisions tend to focus on appropriate channels for spending in an effort to meet student education requirements. Administrative decisions for each school –

staffing, course offering, etc. – are made by the school administrative leadership personnel, e.g. principals, vice-principals, etc. Other stakeholders are involved in day-to-day operational decisions. For example, a teacher may decide to adopt lesson plans that help a specific underachieving segment of the student population acquire the appropriate knowledge and concepts in a more effective manner.

Some form of support is available for most of the "important" decisions. However, the support for low-level operational decisions is spotty at best, and is frequently based on individual knowledge worker intuition. Formal mechanisms to validate the effectiveness of these decisions are lacking. Thus, the effect of adopting a new lesson plan for underachieving students may be hard to assess unless the teacher specifically tracks prior and subsequent test scores for that student population. Clearly, a data warehouse with historical data about student testing would greatly assist in the evaluation of this decision.

It is important to recognize that improvements in decision making attributable to the warehouse are expected to be at the micro level. This is due to the drill-down capabilities afforded by OLAP and query tools, thereby going beyond the reporting from traditional application systems – which generally provide high-level aggregate snapshots. It is hoped that with widespread use of the warehouse within the district, these micro improvements can have a macro effect.

Data Warehouse Development Methodology

The data warehouse was intended to assist all levels of users in day-to-day decision making in the organization – including low-level decisions that make a difference in the education of a single student. For example, individual teachers could examine specific responses of individual students on standardized tests to determine the deficiencies in student knowledge, and prescribe appropriate remedial tutoring for the student. Given the large scope of the warehouse, and the concomitant costs associated with its deployment, it was necessary to ensure buy-in from all stakeholders, not just the managerial ranks. The Director of Technology served as the champion for the data warehouse, creating a core team involving stakeholders from several areas (including district and school representation), and at several levels (including executives, supervisors, and operations personnel).

It was important to ensure buy-in from the core team, particularly since they would also double as the source of information for data in the warehouse. Two all-day orientation seminars were conducted to educate the core team and potential users about the benefits of data-driven decision making, and the need for a data warehouse to support it. Another all-day session with the IT staff was devoted to understanding the organization of the current student databases, spanning content, frequency of update, usage patterns, and the perceived quality and currency of the data.

This data formed the basis for directed data modeling sessions with various stakeholders. In total, the researchers interviewed about 40 individuals representing a dozen subject areas of interest. The sessions typically ranged from 1 to 4 hours, and were performed over a two month span. Data models for each subject area were shared with the users as part of the model validation process. Two iterations of updates allowed the researchers and users to be fairly confident about the data that needed to be incorporated into the warehouse. Individual data models were then merged into an enterprise data model – representing about 120 entities, and requiring a large poster for eventual display.

Logical and physical database designs were compiled from the enterprise data model. Some denormalization of the data was performed to improve performance, as well as to facilitate understanding by users.

Since the data should be accessible and manipulated by users from their desktop, it was deemed that a client/server implementation would be preferable to a mainframe solution. An RFP for the hardware/ software platform was assembled. This process was a little slower than anticipated given the lack of DBA expertise in the organization. Also of some concern was the need to provide access from different client platforms – since the standard administrative desktop was an Intel-based machine, while most teachers used Macintosh computers. Despite considerable interest from several potential vendors at a vendor conference, only 2 proposals were received. A detailed analysis of both proposals by the researchers allowed the core technical group to eventually select Oracle running on a HP platform under Unix as the preferred solution.

Installation and initial management of the hardware and software was performed by the vendor. Over a period of several months, several IT staff acquired the expertise to serve as the DBA for the data warehouse. On a parallel front, several different query tools were examined as possible options for users. Eventually, the group standardized on Brio as the query tool of choice. This was based on its friendly user interface, the ability to easily create views incorporating the user's vocabulary, and the ability to create some data cubes for further analysis. It was installed on some users' desktops, and some ad-hoc training was provided to get them started in using the data warehouse.

Data Warehouse Implementation

The implementation of the data warehouse posed some interesting challenges. While it was desirable that referential integrity be maintained, it quickly became apparent that the strategy adopted for loading the data could preclude this. The warehouse stored a combination of historical and current data. Student attendance and transcript data was largely historical, and expected to grow over time. Student demographic data, on the other hand, was largely current and needed a different update strategy. Given the relative volatility of the data – in a population of over 105,000 students there were roughly 1,000 school changes per day – it was decided that reloading entire tables would be far more efficient than updating specific rows. This strategy meant that referential integrity would have to be sacrificed in some cases. After careful review of the situation, it was decided that the warehouse would not use referential integrity, but would rely on the data extraction and scrubbing processes instead to maintain quality.

Data for the warehouse came from several sources. Some of it resided in the existing student database implemented in IMS (a hierarchical DBMS) on an IBM mainframe. Other segments of data, e.g. assignment of students to schools, were obtained from existing mainframe applications. Some other segments were provided by external sources, e.g. student testing data. The population of the warehouse was expected to be a daunting task. Programmers familiar with the various application systems were designated to create programs and procedures to extract and clean the data, placing the results in flat files for loading into the warehouse.

The sheer size and scope of the data warehouse precluded implementation in a single release. Aside from the hardware resource limitations, the task of collecting all the relevant data, developing data extraction and scrubbing programs, and loading the warehouse, dictated that a phased implementation be adopted. Prioritization of the areas for inclusion in the initial release followed standard organizational procedure – an assessment of the criticality of each area with some political undertones. Eventually, it was determined that the initial release of the data warehouse would contain data relating to accountability measures – an area that was of concern to the school board, and one that would benefit from the availability of a flexible query system.

Implementation Experiences and Issues

While the researchers were prepared for some challenges during the implementation, the problems faced during implementation proved to be an interesting set. First, the volatility of data was much more than originally anticipated. Student mobility in the district was exceptionally high, and the original decision to load some of the tables on a weekly cycle needed to be revised to a daily cycle. The data extraction and scrubbing programs also proved to be an eye-opener. While the district acknowledged that there could be some data quality problems in its current student database, the effort required to fix these proved to be far more than originally anticipated – due to underestimates of the error rates. Each error had to be examined for remedial action – a process that took far more effort than initially budgeted for. An unintended consequence was that the district was able to now address a significant portion of the data quality issues in the current student database.

Data quality problems relating to external data proved to be a greater challenge. The data on student testing, which was provided on tape by an external vendor contain all sorts of errors – scores for fictitious students, missing scores for some students, testing dates not consistent with the testing process, etc. As before, the data was corrected manually, to the extent possible. The result was that data in the warehouse was relatively free from error. However, this caused some problems in generating the accountability measures. Measures generated via the traditional process by the research department and those generated from the warehouse were relatively close – differing in a few percentage points. Nonetheless, this difference caused the warehouse to be viewed with some distrust, particularly since these measures were used to evaluate the performance of administrators and those in leadership roles. While it was clear that the data sets used to generate the measures accounted for the differences, initial acceptance of the measures from the data warehouse was slow, despite acknowledgement of the higher data quality. Over a period of time, users grew more accepting of the results, and currently the accountability measures on student testing are generated from the warehouse. The query flexibility afforded by the warehouse has allowed decision makers to focus on specific goals and targets, e.g. improving performance for Hispanic students at the middle school level.

The creation of the data warehouse sparked interest from a wide variety of external researchers. The district has ties to education departments at several universities, as well as private research agencies. The availability of high quality current data on students in public education with flexible access was appealing. However, issues like confidentiality of student data, and research intent, need to be worked out. At present, the district is in the process of formulating policies to address these open issues.

The continued growth of the warehouse required significant additional resources. The inclusion of additional subject areas required the creation of new extraction programs, and sustained data loading of the current subject areas. The district was also seeking to replace its current mainframe systems with a more user-friendly client/server system. With over 150 schools in the district, this process consumed a great deal of resources, and further deployment of the data warehouse was scaled back. The second subject area to be added in was student attendance. Data for this segment came from four different sources – representing the different systems for elementary, middle school, high school, and alternative school systems. However, the subject area was added in considerably later than originally intended, due to resource deployment in other projects. Use of this data for decision making has not kept pace with expectations, though.

Lessons Learned

The foremost lesson from the experience is that a large and complex endeavor like a data warehouse requires the sustained efforts of a visible champion. The literature is very clear about this (Bischoff and Alexander 1997), and it was borne out in this case. Without the active involvement of the Director of Technology, this project would have not gotten off the ground.

User involvement played a critical role in this project. The sustained interaction with various user groups during the data modeling processes, and the periodic involvement of the core team helped keep the project on track during the first few phases and helped build user support for the project.

Resources available to the project proved to be a critical factor. In the early phases, sufficient personnel were assigned to the project, and the project progressed on a fast track. During the implementation phase, as the school district pursued other initiatives, these personnel were not available to the data warehouse project in a sustained manner. This indicates the need for dedicated personnel resources for implementation, with possible additions as the scope of the warehouse grows.

Data quality issues were also a big factor in the implementation of the warehouse. While it is expected that there will be some quality problems in the core business applications, the data warehouse brings them to the forefront due to its wide access. Unfortunately, this has the untended consequence of undermining confidence in the data warehouse, since the quality problems are readily apparent, and are frequently attributed to the warehouse.

Political factors can also be an issue of some concern. While the researchers expected this, they did not anticipate all possible political ramifications. The warehouse was sometimes employed as a basis for evaluating performance of other departments. Perceived loss of ownership and control over data was manifested as a general reluctance to populate the warehouse with current and error-free data.

If we had to do this again, some things would clearly be done differently. A commitment on personnel dedicated to the project would be a prerequisite. More energies would be devoted to user education – particularly in terms of the sources of data, their respective currency, the mechanisms to ensure quality, and the effects of erroneous data. Additional effort is needed to maintain the high profile of the data warehouse, and release cycles for future phases need to be shortened, or the initial enthusiasm for the project will wear off quickly. The successful applications of the warehouse also need to be publicized to a greater degree.

Conclusions

The implementation of a data warehouse is fraught with difficulties. Doing so at a public institution brought about some additional challenges. While the nature and scope of these problems may have been different, ultimately the success of the warehouse boiled down to the ability to sustain interest in its capabilities and deliver a product that met user expectations. The lack of sustained resource deployment for the data warehouse has led to delayed releases of various subject areas. While there are isolated pockets of users, the lack of more widespread adoption belies the initial interest and expectations for the data warehouse.

References

Bischoff, J. and Alexander, T. *Data Warehouse: Practical Advice from the Experts*, Prentice Hall, 1997.

Inmon, W.H. *Building a Data Warehouse*, QED Technical Publishing Group, 1992.

Inmon, W.H.,Welch, J.D. and Glassey, K.L. *Managing the Data Warehouse*, Wiley, 1996

Milwaukee Public Schools, “Technology Strategic Plan – April 2000”, Department of Technology, http://ftp.milwaukee.k12.wi.us/departments/technology/mps_tsp.pdf, (Current 5/1/2000)