Building a Classroom Data Warehouse

Daniel D. Flies
Division of Science and Mathematics
University of Minnesota, Morris
fliesd@mrs.umn.edu

Dian Rae Lopez
Division of Science and Mathematics
University of Minnesota, Morris
lopezdr@umn.edu

Abstract

A small data warehouse was designed for use by undergraduate students in a data-mining course. Our design included the selection of appropriate hardware, software, and real world data plus the building of the interface between the warehouse and the software used for mining its contents. Guides were also created helping students to design and implement their own projects using the warehouse. Flexibility and ease of use were paramount concerns for the warehouse.

Introduction

The field of data warehousing has become a powerful tool for businesses in the information age. The idea of storing large amounts of data in a uniform way to fulfill the purposes of many departments within a business can be a formidable task. This task is so great that many companies have spent millions of dollars launching a data warehouse only to see it fail, and then they must start again. Our project was to create a very small data warehouse for a group of fellow students to use in a new data-mining course. This was a daunting task, and one that was an immense learning experience. The resulting data warehouse allowed students to test their knowledge of data mining techniques against real world data, not data that was pre-packaged for the classroom.

The design of the warehouse included the selection of appropriate hardware and software, the selection of real world data, the building of the interface between the warehouse and the software used for mining its contents, and the creation of helpful guides so that students could design their own projects using the warehouse. Flexibility and ease of use were paramount concerns for the warehouse. The warehouse needed to be flexible so that students' needs for new data could be met quickly. The data warehouse was developed so that selected data from many different databases could be brought together for data mining purposes. Students at UMM used warehoused data from the U.S. Forest Service for their projects. The US Forest Service has multiple database styles implemented in many states throughout the country. Our warehouse was made to merge this data into a consistent, compatible format that the students could mine easily and efficiently. Students in the data-mining course used warehoused data from Minnesota, Michigan, and Missouri to successfully confirm ideas from forestry professionals that in the past had relied on intuition, experience, and 'common sense'. The data warehouse was an invaluable real world experience for both its developers and its users in the data-mining course.

What is a Data Warehouse?

A data warehouse is a large system that organizes data so that the most information possible can be extracted from it. In other words "...it is more that just data, it is the processes involved in getting that data from source to table, and in getting the data from table to analysts [1]." Data warehouses can and are being used in all facets of business, government and education. Many companies have large amounts of data lying around and collecting dust. A data warehouse organizes and makes available that data to analysts so that they can make better decisions. A data warehouse can take many points of data from many different systems and give them a bond that will let the information work together. A data warehouse has benefits that other storage systems cannot give. Other types of storage such as databases and "mining applications" store the data that will best fit their application, while a warehouse tries to raise the bar and make data its main focus. Other benefits of this system over other systems are integrated data, stabilized

data values, and ad hoc retrieval. Some of these benefits take a database to the next level of data storage. A data warehouse will allow for comprehensive queries of the information sorted. Having one consistent source of information for users ensures that, when a query is submitted, the results returned are as close to complete as possible. A data warehouse also enables researchers and other users to obtain high-quality data to use when making decisions.

A data warehouse in the truest sense contains software, hardware, and the manpower needed to run such a system. Many books are being published about the importance of managing large amounts of data. Some of the data warehouse models and partitions as described in Kimball's book *The Data Warehouse Lifecycle Toolkit* include: Data Staging Area, Presentation Servers, Dimensional Models, On-Line Analytic Processing (OLAP), Relational OLAP, Multidimensional OLAP, End User Applications, and more[3]. With all of the terms associated with a data warehouse, there are still basic concepts that all books and software packages describe: Load Manager(s), Query Managers(s), User Interface, and a Data Storage Area.

Each of the four areas of a basic data warehouse are necessary to the function of a warehouse. The managers are the "work horses" of the warehouse, they handle the interaction between the user and data storage area. These buffers cloak the storage of the data in the warehouse. The Load Manager takes in the appropriate data with the correct data points and finds the optimal place to store and index the data for future searches. One optional area not mentioned is a Load Interface that interacts with the Load Manager to make the loading process easier to use. The User Interface is a vital part of the data warehouse construction and its implementation is critical to the success of the system. A user interface must make the system intuitive so that users are able to learn and make qualified decisions. The Query Manager takes control of asking the data storage for the necessary data to fulfill its request from the user interface. The interaction between these systems is the heart of a data warehouse. Other enhancements have been shown to be useful to both users and administrators but the necessity of the above four parts is what makes a basic data warehouse.

Data warehouses are gaining wide usage in various industries throughout the world. Many of the companies that try to convert their data storage to a data warehouse fail. The average business has a 50-50 chance of making it work. With the cost at an estimated one million dollars, that risk is high.

One example is Ryder; a company built on its nation-wide network of facilities. Ryder rents vehicles thousands of times a day, each with drivers, destinations, and origins. This creates a collection of information that must inform Ryder locations across North America. Ryder's first attempt to redesign their data storage solution failed. Their first attempt was a complete overhaul of the current multi-database system. It could build the consistency that they needed but, even though it failed, the importance of a data warehouse solution couldn't be overlooked. So they tried again. This time a scaled approach was used with success. Ryder's first approach was on too large of a scale, so they decided to take smaller steps on their second try. They started the new conversion

by converting a few databases into a data-mart. A data-mart is a smaller part of a data warehouse. Usually a data-mart is designed to be scalable; companies can invest in one data-mart then build another one and connect the two data-marts together to create a lager system of information. Data-marts are also usually built in various geographical locations, thereby taking a company like Ryder and its information and distributing it across the country for quicker response times for the various facilities. This can be scaled up to create an inter-connected collection of data-marts to make one large data warehouse.

Our Implementation

Data warehouse theory has many aspects that require numerous resources including time, money, and people. Our implementation was very limited compared to the ideal situation in which to construct a data warehouse. We first needed to determine the needs of the students in the class that this project was centered around. We then assessed the current capabilities and began designing a system that would be easy for students to use and for administrators to maintain. Current commercial solutions were considered as an option, but it soon became clear that without major funding our project would have to be limited to just our discipline resources. Thus our options became limited for hardware and software.

Salvaging parts from several machines throughout the various labs at the university was needed to create the data warehouse machine. Ranking our needs, we began to look for a machine that could store a large amount of data, the bigger the better, the faster access times the better. Since our lab server was being replaced, we took a 30 GB, 7200RPM hard drive from it. Since we would be processing large amount of data into our storage format, we needed a good, but not necessarily the fastest, machine. Limiting it to 400 MHz seemed reasonable. Our new warehouse architecture was a Pentium 2 400MHz processor, 128MB of SDRAM, and a 30 GB hard drive for our data storage. One factor that was not explored heavily, because of the lack of availability, was RAM; 128MB worked satisfactorily and no problems occurred with our limited processing and small warehouse. In hindsight, we discovered that the combination worked well, but was a bit unnecessary. Our data never exceeded 100MB, while containing a quarter of a million entries of data instances. This was plenty for our class to begin “mining.”

Having solved the hardware issue, a larger problem existed: Where could one find real data to store? One of the objectives of the warehouse and the class was to apply the data-mining techniques to real data. We didn’t want to use “pre-packaged” data and situations that had pre-specified results and objectives. The test of skills and knowledge would come with real data. While many solutions to the “data problem” existed, finding a match with interest and research impact was challenging. Databases from previous state elections were available to candidates, but not accessible to the general public. Data from a local agricultural research laboratory was available, but seemed to lack the interest of the average student. One vast set of data collected by the U.S. Forest Service included various descriptions of trees and their attributes. The data sets included data collected

from nearly every county in 30 states. While the data contained large amounts of data, our warehouse could only store a small amount of this data. A data warehouse is supposed to store data in the hundreds of gigabytes, while our created warehouse was struggling to reach hundreds of megabytes.

To process the data from raw text file into storage format first required a storage device (database) and a method to get from text files to the device. Our device for storing and organizing was a MySQL database. The MySQL database had some very important features that led us to choose it. The MySQL database was relatively easy to set up and use. Other options available, such as the Oracle Database library of software, were unable to be used in our laboratory setup. MySQL would also not hinder a machine in our lab from also being used as a workstation. Most importantly, MySQL is free. The UMM student lab manager set up the MySQL server on our built-from-scavenged-parts machine. Now that the architecture was in place, we needed to bring the data into our warehouse structure.

The Data Entry Process

The process of planning how to store data into the warehouse is an important step when building a warehouse. The first step in the process was to visually analyze the raw data from the US Forest Service and determine which were the necessary and relevant pieces of data for the course. The data was originally stored numerically and could be deciphered using a database manual that gave the English descriptions of each row's attributes. The data had over 30 attributes that may have needed decoding, but our class determined that only 13 attributes had to be changed from the numerical to the verbose mode. This process is normalization. Normalizing the data was an important step that helped the end users to understand the data and its implications for analysis. Also included in our normalization was the correcting of misspellings, checking abnormal data points, and appropriately clearing and zeroing missing data values.

The normalization of the data into its format was a simple process if done prudently. The process was time consuming, but correctly configuring the software side of the database would save the users and administrators time in the long run. The process of normalization was the first step for getting the data into the warehouse. This step included constructing the database format so that the data's characteristics would not be hindered by the database's table fields. To prevent confusion with field names and data value characteristics, they were kept as close to the original Forest Service guidelines as possible. To make changes to the quarter of a million entries in the database, a small program was written to change the appropriate numerical data into the correct English equivalent. Using java and several hash tables the data conversion encountered only minor problems that were easily corrected. MySQL provided tools so that the file of data could be entered into the database smoothly. After the data was entered, the data warehouse was near completion.

To aid the students though the new warehouse, several guides were created. Our guides include aids for MySQL, an introduction to SQL queries, and links to Forest Service data manuals. Several handouts and presentations given to the class were necessary to ease the students into the new concept of a data warehouse. All of the manuals were published on the world-wide-web so that they could be accessible to everyone when needed. Guides taking the students through the step by step process to access the information available to them were vital to the success of the data warehouse.

Results

The results of the data warehouse system developed at UMM were mixed. Through a semester of research and hard work, a successful course in Data Mining was implemented where students made numerous discoveries from the data in the warehouse. The process of learning, designing, implementing, and now sharing the progress of our efforts have been immensely rewarding.

The success of the data warehouse can been seen in the works of the many students that created the opportunity and the need for such a data warehouse. Students confirmed ideas that US Forest Service officials believed to be true through years of field experience and research. These results created an excitement for the project. The storage system and hardware selection withstood the usage. The versatility of the warehouse to accept data of different styles was proved successful even with the many updates and additions to the warehouse needed by the student projects. The organization and preparation of data made the data points easy to understand and interpret.

The data warehouse implementation flaws became abundantly clear from the immediate testing from fellow students. A thoroughly designed user interface, arguably the most important piece behind the warehouse besides the storage of data, was not implemented because of lack of time. While the lack of a robust user interface was a disappointment for the creators, the students were able to access and use the data. The creation of an administrator's toolkit is also an avenue that time did not permit for our exploration. Areas left for improvement make this field an exciting prospect to continue research on.

Conclusions

The data warehouse was a success! The completion of semester-long projects by the class was proof that the system was a great help for others to learn from. The educational opportunities have been immensely rewarding. Researching a new field in computer science and the opportunity to implement new ideas for data storage and retrieval was very exciting to the authors. The opportunity for an undergraduate student to experience this type of learning is one that has been difficult, but well worth the effort.

References

[1] Anahory, S., & Murray, D. (1997). *Data Warehousing in the Real World*. Harlow, England: Addison-Wesley.

[2] Pyle, D. (1999). *Data Preparations for Data Mining*. San Francisco: Morgan Kaufmann.

[3] Kimball, R., et al. (1998). *The Data Warehouse LifeCycle Toolkit*. New York: John Wiley & Sons, Inc.

[4] Witten, I. & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann.

[5] Witten, I., Moffat, A., & Bell, T. (1999). *Managing Gigabytes*. San Francisco: Morgan Kaufmann.