Category: Applications of E-Technologies

Data Warehousing and Data Mining Lessons for EC Companies

Neerja Sethi
Nanyang Technological University, Singapore

Vijay Sethi
Nanyang Technological University, Singapore

INTRODUCTION

Internet companies are now in the second stage of evolution in which the emphasis is on building brands (Campman, 2001) and retaining customers rather than just transactions. There is also an imperative for multidimensional Web performance monitoring (Earls, 2005) and a continual fine-tuning of sites for optimal navigation, increased stickiness and transactional efficiency. Such research as the relationship between customer profiles and navigational characteristics (Garatti, Sergio, Sergio, & Broccab, 2004) and techniques for seamlessly aggregating Web data with corporate data (Wood & Ow, 2005) also testify to the importance of holistic data analysis for knowledge discovery. The technologies that are becoming critical in this fight for customer retention are data warehousing, data mining and customer relationship management.

This article presents two case studies, one on data warehousing and the other on data mining, to draw some very specific lessons about management support, organizational commitment and overall implementation of such projects. These lessons complement past recommendations that these technologies are more about organization change (Kale, 2004), about a single unified view of the business and, ultimately about building a shared data model of the enterprise.

We start with a brief overview of data warehousing and data mining. The two cases are discussed next, using a similar analytical structure to facilitate comparison among them. In the conclusion, we describe the key lessons learned from the two cases and implications for future research.

Data Warehousing and Data Mining

Data Warehousing

Data warehousing is the process of creating an integrated and summarized copy of an organization's transactional data for the purposes of data analysis and decision support. The Data Warehousing Institute (TDWI) has played an active role in highlighting and popularizing best practices in the industry. It has produced a Data Warehousing Roadmap (Barquin & Edelstein, 1997) to guide would-be implementers. Another methodology exists, called the Metis Methodology (Kelly, 1997), developed by Sean Kelley, founder of the Data Warehouse Network. This has been further refined into the Hadden-Kelly Methodology.

A key aspect of data warehousing development is rapid application development (RAD) that is best used when it is impossible to fully specify the system's requirements accurately. However, this does not mean that no part of the system is fixed. Key high-level components, such as warehouse architecture, data model, data dictionary and other logical and conceptual components, should remain reasonably stable, standard and well defined. What evolves continuously are the applications and the data in the warehouse. A useful metaphor is that of shipbuilding, where the design of the hull remains immutable, but various ships (e.g., aircraft carriers, cargo haulers) are built on the same hull. Benefits of the rapid application development approach include the ability to manage the inherent risk of a data warehousing project, the ability to prioritize resources and the continuous delivery of business functionality.

Another way to mitigate risks is to build data marts, which are essentially smaller, more focused data warehouses that cater to the needs of a single business line or function. Data marts are often marketed as quick fixes to an organization's data management problems, and many organizations implement them as part of a longer-term plan to roll out a data warehouse.

Data Mining

Once an organization has an integrated data warehouse, its members can use a variety of tools, such as online analytical processing (OLAP) applications and SQL, to query and analyze the data (Inmon, 1996). OLAP refers to the technique of performing complex, multi-dimensional analysis in an ad-hoc manner and ranges from basic

navigation and browsing (often referred to as "slice and dice") to more complex modeling and calculations. SQL analysis requires users to have a requirement or hypothesis that provides a clear and bounded focus to the data exploration.

Data mining can be done in a variety of ways, and these are often applied jointly (Chan & Lewis, 2002). Association is a method that aims to find affinities among records in a data set and is used in applications such as Market Basket Analysis. Sequential patterns are used, among other things, to detect buying patterns of individual customers. If records in a data set have been divided into various classes, classification can be used to describe the characteristics of each class. If, on the other hand, the records are not classified, clustering can be used to segment the records according to some criteria. Some of the mathematical techniques employed include rule induction, artificial neural networks, fuzzy logic and decision trees.

DESCRIPTION OF PROJECTS

Case: Data Warehousing

Company Overview

The project was implemented at the regional center of a multinational bank that employs 30,000 people globally in more than 50 countries. Its regional center was started in 1859, employs more than 2,000 people and provides the full range of Consumer Banking, Corporate and Institutional Banking and Global Markets products and services. The bank offers a complete range of Internet banking services, where users can set up personalized home pages, choose financial consultants, open online accounts, and select a variety of banking information.

Goals and Scope

The bank turned to data warehousing due to the increasingly competitive environment that had resulted in thinning profit margins. It came to the conclusion that it had to identify its most valuable customers and leverage them to use a broader range of banking services. Developing customer profiles, one of the key drivers of Internet banking as well, was important. Moreover, the bank would also need to use data mining techniques to discern trends and forecast the success of its marketing campaigns.

But customer information was scattered among various departmental databases and it was difficult to identify and profile these customers. A system called customer information management (CIM), which was meant to give each customer a unique ID, existed side-by-side with the various departmental databases. However, not every department used it and customers had different IDs in different systems; thus, CIM was reduced to mapping its IDs to the IDs in the other databases.

The IT Department of the bank decided on a data warehouse package called *Collage*. It was decided that *Collage* would only hold the customer IDs from CIM. Data from various sources would first be mapped to a customer ID in CIM and then loaded into a separate database. The data would be transformed to standardize the naming conventions and then it would be stored in the data warehouse to allow individual applications and departments to pull out whatever data was required.

Project Development and Management

The objectives of the development process were fourfold: (1) to understand the data available, (2) to prove that banking indicators—for example, credit limits—could be integrated around customers, (3) to demonstrate inconsistencies in the current data, and (4) to create an all-encompassing data warehouse with a few data marts around it. These data marts would provide the necessary subsets of information required by individuals or groups and would be located in closer proximity to the users. The deliverables included a logical data model, a data dictionary, the physical database structures in the data warehouse, a sample set of data instances for validation purposes, a validation of the database schema and integration methodology, and an action plan to address any issues that arose.

Once the project was accepted, a task force was formed and given 4 months to develop the data warehouse. The project team decided to build a prototype to test the robustness and validity of the underlying assumptions. It spent a significant amount of time evaluating various hardware and software combinations. As a result, planning itself took 4 months and it was another 2 months before data could be loaded into the warehouse. A prototype containing data for the top 30 customers was evaluated and considered successful. Full-scale development followed, and it went smoothly, according to the schedule drawn up.

Implementation

Though the system was complete, no one came forward to use it! After a period of time, the Marketing Department showed an interest in it. But while their data mart contained customer details such as reference, revenue and balances, they required customer segmental information. Modifications, thus, had to be made, and to better articu-