

ABC embraces data mart

"This new Data Mart system enables us to generate meaningful information much quicker and allows us to make more effective decisions throughout the organization."

-- Marketing and financial analyst, ABC Corporation.

Let us begin with simple math. Imagine you are a provider of long distance telephone service to a customer base of 300,000 customers. Assuming you bill them on a monthly basis, you would store 300,000 multiplied by 12, or 3.6 million billing entries in a year. That would be 7.2 million per year if you track usage by two methods of calling, say direct dial and calling card. And 36 million if you keep five years of history. If you want to analyze this information to identify calling patterns, you may have to summarize the bill amount and average number of minutes by method, by time of day of call (peak, evening, night), by type of call (intra-LATA, interstate, intrastate, international), and by year and month, which might take days to produce any results at all.

This is precisely what ABC Corporation (ABC) experienced in the pre-data mart era. ABC Corporation is a socially responsible organization and a leading provider of Long distance and associated products such as calling cards, pager and Internet . What is fascinating about ABC is that it does not produce any of its products. So the telephone lines are leased from Sprint , calling cards from MCI , and credit cards co-branded with Fleet Bank. In addition, the billing and collection functions are outsourced to a Data Center provider in Pennsylvania, and Customer service to some other agencies. ABC limits itself to the remaining core functions of planning, marketing and finance. As you may have guessed it, the basic need of such a user group is access to operational and external data suitable for marketing and financial analysis.

The Problem

Unfortunately, the kind of queries the ABC wanted to run for decision support simply overwhelmed the operational database. This is not surprising since the operational database was designed and tuned to support just one thing: day-to-day operations. And it did that well. But it failed to support the analysis queries that summarized, sorted and merged several large tables. Some queries ran overnight and over the weekend. Often some users had to be requested not to run their queries so that other queries could go through.

But there were other problems too! The billing vendor to whom the IT operations were outsourced was not willing to host the volumes of historical data that ABC required for decision support, but no longer required by the application systems. The Customer service levels were going down because the DSS queries squeezed all the computing resources. Then there were other usability problems with the operational database, since it did not follow standard naming and coding conventions, and stored dates as numeric fields in a variety of formats that could not be easily queried.

The Solution

Since the operational data was in Unisys environment, ABC found Unisys's Data Mart Foundation Kit the perfect solution for their problem. The kit includes Unisys Clearpath HMP server that now hosts its proprietary operating system as well as Windows/NT environment, connected by a high speed bus. It also includes Attachmate's DATABridge middleware that provides database connectivity between Unisys DMS-II and MS SQLServer in Windows/NT environment. MS SQLServer stores over 60 GB of the user data. Sagent Data Mart Solution provided the scaleable 3-tier delivery vehicle to access data for On-line Analytical processing (OLAP).

It was then left to us to bring these components together into a coherent architecture, and to design and develop the data mart application.. This required installation of components, integration, tuning, building of metadata, and development of custom software to refresh the data mart on a daily basis.

Performance Challenge

Four measures were taken to improve the performance of ad-hoc queries. Firstly, the RAID drives were striped to provide efficient retrieval of large volumes of data. Second, Sagent’s caching mechanism improved performance for repeat queries, since the data is retrieved from cache instead of extracting and summarizing the same data over and over for multiple users. The third measure taken was identifying and building index for typical access paths. Last, and perhaps the most significant measure was restructuring of the OLTP data model to a dimensional model called “star-schema”. This includes pre-defined joins, de-normalizations and other physical optimizations to improve the performance. The star-schema model revolves around facts and dimensions. Facts are variables or measures that can be aggregated and summarized, such as discount and bill amounts. Dimensions are the business perspectives from which the fact can be analyzed, such as customer, product and time. The primary key of the fact table comprises of the foreign keys from all associated dimension tables. Since the fact table is pre-joined this way, it is easy to frame and execute queries that look at the fact table from multiple dimensions.

The sales example here shows how the Product Sales fact can be analyzed by account, product, marketing channel, and time dimensions such as year and month. This kind of analysis helps ABC in product cross selling, and monitoring growth and attrition.

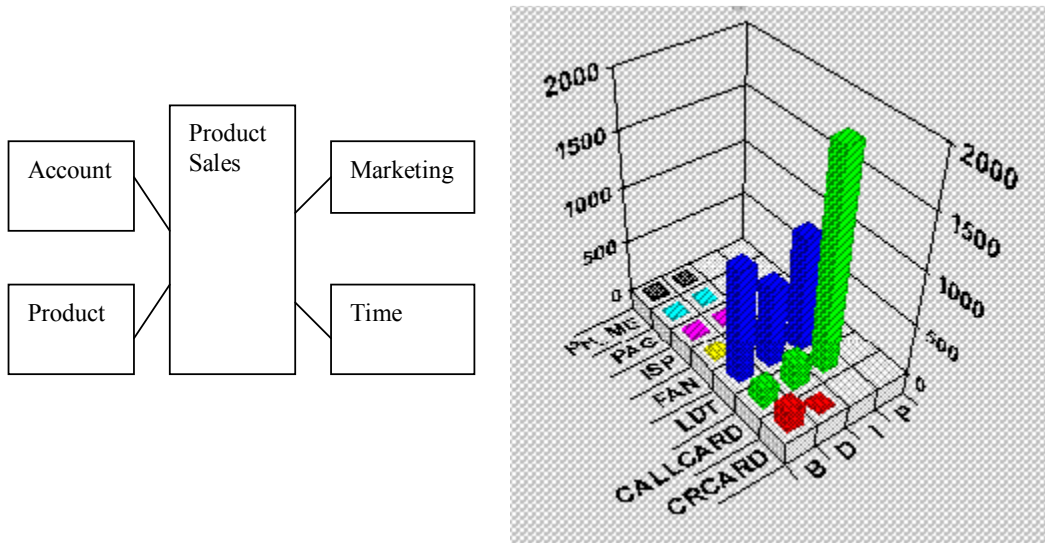


Figure 1: Product sales star-schema and its multi-dimensional view

Refreshing the data mart without affecting the performance of operational database was another challenge. The data mart includes 50+ tables with some with over 10 million rows. These tables need to be refreshed on a daily basis from the operational database. The refreshing process was so designed that a process keeps browsing the DMS-II audit files for modified data, instead of the DMS-II database. These changes are immediately extracted and copied “as is” into a buffer area. Overnight the data mart update process cleans, transforms, and moves the data from buffer area into the data mart.

New Capabilites

The most widely used new feature is the time-series analysis capability. This feature helps users analyze the snapshots of operational data stored in the data mart in sequence over a period of time to determine the emerging trends and patterns, or to investigate anomalies as of a particular date.

Users also like the non-volatile environment the data mart provides. Since the data mart stores operational data as “snapshots”, the results are consistent across users and when users drill down and roll-up the data at different times.

The Sagent metaview makes it easy to build ad-hoc queries without the need to know the underlying relationships. The Sagent Analysis interface provides OLAP features such as pivoting, slicing/dicing, drill-downs and roll-ups. It presents the relational data from SQLServer in a multi-dimensional form, as tables and charts. The interface to Excel and Crystal Reports makes it handy to leverage the existing skills in ABC.

Lessons learned

Building this data mart that users really like and use was not without surprises. The version of the middleware we started with fell short of the performance expectations, but the vendor quickly responded with an improved version. And, the data entry system that ABC used called ‘Smart system’ was not very smart and accepted invalid data. In some instances, the dates actually contained account number, perhaps because date and account number fields are placed adjacently in the GUI. We also made the mistakes that everybody makes in data warehousing, i.e. underestimating the disk space and memory requirements. Both had to be doubled immediately after implementation. The backup drive had to be upgraded as well, with a high speed one that can catch up with the increasing volumes of data.

Future direction

The next steps include extending the data mart to include additional information. ABC may soon be entering the “green energy” and local telephone markets, that need to be brought into data mart. And perhaps other data marts for Customer service and Customer value analysis.