The Technology Guide Series™
techguide.com™

Building a Successful CRM Environment

This Guide has been sponsored by

IBM®

Visit ATG's Web Site
to read, download, and print
all the Technology Guides
in this series.

www.techguide.com

"The significant problems we face cannot be solved by the same level of thinking that created them."

Albert Einstein

Table of Contents

About the Editor...

Thomas Flanagan and Elias Safdie are Vice Presidents of Editorial Development and Principals at ATG. They are also co-founders of The Cordis Group, a consulting firm specializing in helping management use information technology to meet their business objectives. Together Thomas Flanagan and Elias Safdie have nearly fifty years of combined experience advising both vendors and system users in the strategic and tactical uses of information technology. You may contact the authors at esafdie@techguide.com or tflanagan@techguide.com or at 978-256-4974.

Introduction

The 1990s have been a dynamic period within the information systems field. One of the most exciting uses of new technologies has been in the area of Customer Relationship Management (CRM) systems. Today the availability of econometric, demographic, lifestyle and psychographic data, decision support systems, the Internet, and other customer access techniques are helping marketing and senior management make customer care a reality rather than just a vision. Companies no longer want to treat their customer base as a homogeneous collection of revenue generating units; they want to get up close and personal with each of them individually.

Demographic	Lifestyle	Purchase Behavior	Automotive
Who They Are	*How They Spend Their Money*	*How They Buy*	*You Are What You Drive*
Female/Male Age Income Marital Status Children & Ages Occupation	Frequent Flyers Gardening Arts and Crafts Mutual Funds Home PC	Shops by mail for: Home Furnishings Children's Apparel Power Tools	Make/Model Year Bought New/Used Leased Lease Expiration Date

Figure 1: Typical Demographic and Lifestyle Data

Everyone has been exposed to these changes: the mailings from banks offering special interest rates, phone companies offering incentives to switch to their services, etc. At the core of all of these is the offering company's desire to develop a "relationship" with a target audience.

Today's technology has reached a price/performance point where it is possible to acquire, consolidate, analyze, and manage the volumes of information that make this concept possible. This guide will explore the nature of CRM, why companies are moving to this strategy in such astounding numbers, why information technologists should be aware of these trends, and what the IT community can do to support these initiatives and contribute to the success of the organization.

CRM—The Basics

What is CRM?

First of all, it must be understood that at its core, CRM is more than just a set of technologies: it is a process. This fact will be of significant importance to Information Technology (IT) professionals who will be asked to support CRM with information and applications. Furthermore, it is intended to be a *repeatable* process to ensure ongoing, continually improving, and consistent results. Simply stated, CRM comprises the acquisition and deployment of knowledge about customers to enable a company to sell more of their product or service more efficiently.

CRM begins with in-depth analysis of customer behavior and attributes to achieve complete knowledge of the customers, their habits and desires, and their needs. It then applies this knowledge to the formulation of marketing campaigns, strategies, and treatment plans. However, managing the relationship also implies customer interaction. Therefore, CRM also encompasses enabling a network of "touchpoints" by which the organization can establish, cultivate, and maintain long-lasting and mutually beneficial interactions with the customer. These are the two cornerstones of CRM—the knowledge or customer information platform and the customer interaction platform.

Finally, in order to achieve continuous improvement, it is necessary to track the results of the customer interaction and use those results to refine future actions. This implies capturing the essential information exchanged through the touchpoints.

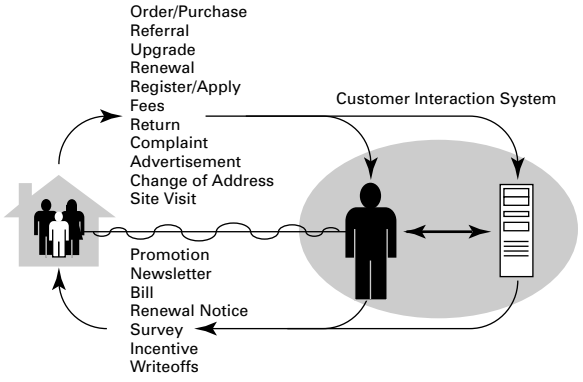

Figure 2: Customer Interaction Examples

The role of enabling IT in all three of these areas—deriving knowledge, enabling intelligent customer channels, and capturing and analyzing feedback—cannot be overemphasized. IT professionals have a tremendous opportunity to become enablers of a customer-centric business strategy and to have an impact on the organization's bottom line.

Evolution of CRM

While revolutionary in many respects, CRM is also a natural and predictable extension of how marketing and sales have evolved over the years, and in many ways, are coming to a full circle.

In the past, customers were served by the corner store and door-to-door sales forces. The corner stores were small, intimate, and provided one-on-one service to their clientele. The door-to-door salesperson was the other face of the company and the personal relationship established by the salesperson was the key to success. This model provided, through personal interactions, an intimacy and knowledge about the customer and developed customer loyalty and trust.

Mass Marketing

The age of mass marketing replaced the intimacy of a direct sales force in many organizations. Centralized large-scale production, wide-geographic distribution, and one-way communication on a grand scale created a tremendous variety of easily available, affordable goods. This put pressure on the relatively inefficient corner store and door-to-door models. Over time, the local corner store gave way to the supermarkets, malls, and megastores of today. While society has benefited from the cost efficiencies of these arrangements, something was lost in the bargain. That loss was the sense of connection customers had with the local storekeeper—personalized service.

Mass marketing was enabled through technological improvements in TV, radio, and the printed press, all of which created simple and powerful means to communicate a company's message to millions of people at once. Marketing's major goal was to push product and create brand recognition. The main measure of success for this business strategy was market share.

Target Marketing

In the mid 1980s, with the advances of technology and refinement in direct mail and telemarketing, another approach to communicate directly with the customer evolved. The use of Information System technology allowed the selection of specific ("targeted") customers via mail or telephone. Unlike mass marketing, targeted marketing had the advantage of potentially receiving a direct response from a customer. The general strategy was to unearth potential customers by canvassing large numbers. Response rates became the central metric in gauging success, with response rates of two to three percent being considered successful. Market share still remained the primary measure of business success.

Target marketing recognized the need to interact more with customers, albeit at a very superficial level, but did not go far enough. There was a lack of specific

data as it relates to responses from the targeting means resorting to "averages" for response rates, customer purchases, and other data. Nonetheless, target marketing was a significant step in the evolution to today's CRM in that it moved the relationship between producer and consumer one more step towards a personal interaction.

Obviously, mass marketing has not vanished, nor have sales forces or telemarketing efforts. In many companies, all of these techniques are still used in combination but with no cohesive overall plan between them. This leads to confusion as customers receive multiple, uncoordinated messages through separate channels.

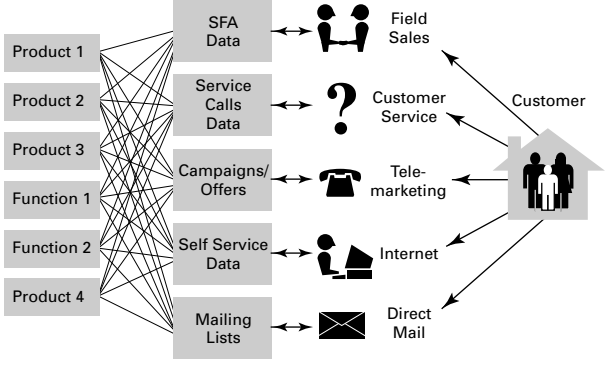

Figure 3: Islands of Information

Customer Relationship Management

CRM is the next step in the evolution, and it moves us back towards developing an intimacy with today's customers, using today's tools, and maintaining our mass production and distribution systems. It recognizes that the equation that yields trust and loyalty from a customer has two variables. The first variable is information and analysis (knowledge): one has to know what the customer wants, needs, and values. The second variable is the need for interactivity and personal contact and the way in which the customer wants to be contacted. The success of a customer-centric business strategy is measured not only by "share-of-market" but by "share-of-customer".

In the following model, the four quadrants represent approaches that combine relative measures of knowledge about a customer and interactivity with that customer.

- The knowledge scale is a measure of what is known about the customer's behavior and values. This is the informational and analytical part of equation.
- Interactivity is the measure of dialogue with a target customer, from one way communications at the low end to full interactivity at the upper end. This represents the personal contact and interaction part of the equation.

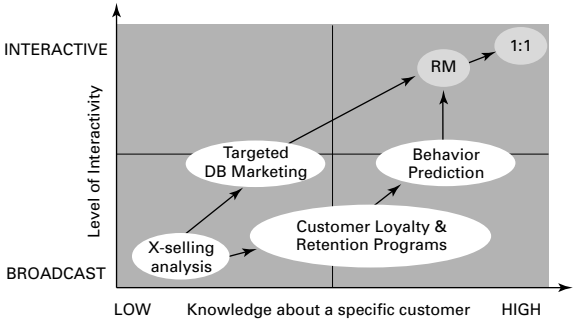

Figure 4: Evolution of Marketing Initiatives

The Customer Relationship Management Cycle

Because the process is intended to be repeatable, it is only natural that a cycle is associated with the implementation of CRM. This cycle consists of an

assessment phase, a planning phase, and an execution phase. In this cycle, assessment is made up of the knowledge acquisition portions of the process, planning comprises the creative part of the marketing process, while the execution phase maps to the customer interaction elements.

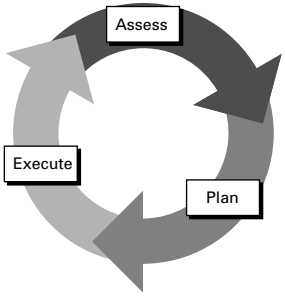

Figure 5: The CRM Cycle

Assess

The assessment phase develops a model of the behavior of target customers, using a combination of in-house data and external demographic, psychographic, and other data. Here, marketing will explore a number of questions, including:

- Who are the customers? What are their demographics and lifestyle?
- Where do they live? Here geocoding and proximity measures are applied.
- What are they worth? What is their lifetime value potential?
- What and how do they buy? What are their purchasing patterns? Is there a model of their profitability or the risk associated with doing business with them?
- How can they be reached? How have they responded to promotions in the past and through which channels do they prefer to be contacted?

This phase is arguably the most technology intensive phase. To be successful, IT will need to integrate many legacy and external data sources, usually through data warehousing technologies. The analysis of the data will require tools such as OLAP (On-Line Analytical Processing), data mining, and statistical analysis tools such as CHAID (CHI-squared Automated InDuction), CART (Classification And Regression Trees), and other complementary tools to report, analyze, and unearth hidden trends in the data. Because this is the part of the process where the target segment of the customer base is selected and customer requirements are analyzed, it is also the most critical phase of the cycle.

Plan

During this phase, marketing decides how best to approach the customers defined in the assessment stage by designing marketing campaigns and strategies. Although there are IT solutions available for campaign planning, this phase is less dependent on technology for its success. Traditionally, the planning phase is the creative part of marketing, with support provided by tools and frameworks.

Execute

The execution phase of the cycle is where an organization puts all this knowledge to work, using all of the customer touchpoints available. Effective customer interaction, which has two dimensions, is the key here. The first dimension is the execution and management of marketing campaigns and customer treatment strategies through these interaction touchpoints.

The second dimension is the tracking of responses, which also represents an important aspect of this phase: gathering data on the results of each plan which are to be used in the next assessment cycle or the next time a customer interacts with the company. From an IT

perspective, it is essential to understand the nature of the data to be gathered and how it will be used so that appropriate data standards and metadata definitions can be deployed to make the subsequent use of this data simple and meaningful. If this data is captured successfully, then the next cycle will be more productive and the repeatability of the process will be beneficial. If the organization does a poor job of defining and using the execution data, then the repeatability of the process will only lead to a rehash of the same mistakes over and over again.

Business Drivers—Why CRM? Why Now?

There are three primary reasons why CRM has taken hold as rapidly as it has: competition is fierce, the economics of customer retention are unequivocal, and technology allows organizations to do this more effectively and profitably today.

The Return on CRM Methods

There are only three ways to increase the profitability of a customer base; acquire more customers, optimize the value of existing customers, or retain the right customers longer. All of these benefits must be achieved with lower costs.

As the economic climate continues to become more competitive, the fight over customers intensifies. Of the three choices above, acquiring new customers is the most expensive. Research shows that acquiring a new customer costs 5 to 10 times more than retaining an existing one. Studies also show that loyal customers will buy more over their lifetime and are willing to pay a premium for doing business with someone they like and trust. Therefore, while organizations will clearly continue looking for new customers, once acquired, they now know that it is worth a significant investment to keep them. CRM is a way to do that.

The chart below shows relative areas of improvement by companies who made an effort to improve their target marketing and move toward a CRM approach, underlining the tremendous advantage that a well-implemented CRM strategy can bring to an organization.

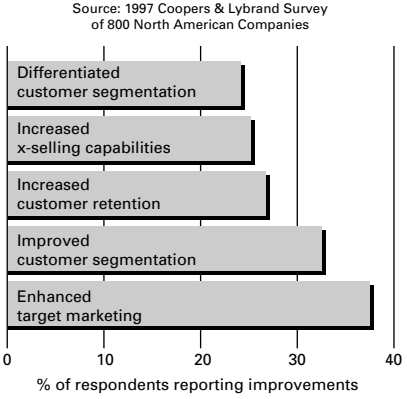

Figure 6: Benefits of a CRM Strategy

A Confluence of Factors

Technology has become a prime factor in the move to CRM. A confluence of multiple disciplines, from a data warehouse and its underpinnings in VLDB and parallel computing, to OLAP, data mining, and other complementary technologies, have enabled marketers to sift through mountains of data to extract invaluable information and knowledge about their customer base. Without these technologies, the ability to achieve a market-of-one concept would not exist. Integrating these technologies with operational front-end and back-end systems provides the necessary seamless collaboration, and the IT challenge that

comes along with it.

Furthermore, new channels like the Internet, information kiosks, sophisticated call center automation, and the associated communities they each embrace, provide organizations unprecedented opportunities to reach out to their customer base and integrate these new channels into new business processes not possible a few years ago. Over time, as these capabilities are further developed, they will provide huge economies of scale to companies that take advantage of them. Some companies are already recognizing the fact that each of these channels appeals to a different community. Therefore, the customer "channel preference" can become a significant profiling or segmentation dimension in planning for specific promotions. Likewise, each channel has different costs associated with its deployment and use, and can therefore lead to differing levels of profitability per customer for the same product. The challenge then becomes devising programs that will optimize the goals of the organization based on the distribution of customers that will buy using the overall combination of channels.

Companies have also recently become aware of the need to manage their business processes more effectively. Workflow technology, which allows enterprises to design automated processes to enhance the productivity and responsiveness of their workforce, as well as to deliver new levels of service, has also played a major role in enabling CRM initiatives and providing a higher quality of customer service.

Finally, as data mining and segmentation technology has evolved, there has been an observed trend towards insourcing of database marketing activities, rather than the more traditional outsourcing to service bureaus. As the technology for identifying target customers becomes affordable and available, individual companies are finding that they want to gain control of the process of connecting with their customers. Bringing the process in-house brings not only more control to the organization, but better results. Companies can overlay their own data to better understand customer attributes and get better results. Rather than relying on purchasing extensive lists of customer attributes (many of which they can't use and throw away), companies can intelligently select these attributes, thus saving money and achieving better tailored results.

Furthermore, as the algorithms for segmentation become better refined, companies are coming to understand the power of the segmentation tool. By bringing the tools in-house, they are able to build models that meet their own specific needs. These highly customized segmentation and propensity models have become highly treasured trade secrets, which must be assiduously guarded. This represents an opportunity and a challenge for the IT professional to work with the marketing organization in implementing and deploying these new technologies.

Critical Success Factors

An Architecture for Developing CRM

An architecture is a blueprint created to define the solution to a problem. The architecture brings together the tools and elements of the solution in such a way that they work together as a single, cohesive entity. The more complex the system, the more need for a good blueprint to help guide the development team members in their individual tasks. This is the job of an architecture.

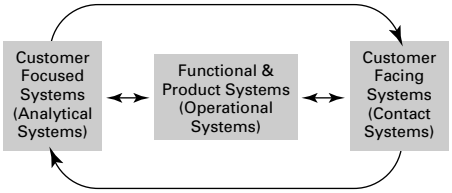

Figure 7: Key Elements of a CRM Architecture

Data warehouses represent the linchpin of the architecture for CRM initiatives. It is the warehouse that becomes the nucleus for all the data required to perform the assessment phase, and from which all of the detail information is disseminated to call centers, SFA, and campaign management components. The data warehouse infrastructure for CRM needs to be a comprehensive customer information platform that supports the marketing process.

A successful architecture is more than just the software components and the hardware boxes on which they operate. A significant portion of the architecture effort must deal with the data standards, data models, and data structures required to support the effort. Remember, in most data warehousing projects, as much as 80% of the service costs involve the data archaeology (finding the right data), data transformation and cleansing, and the creation of data models and metadata repositories, which support the work of the analysts and marketers.

Unless the work performed by the analysts and marketers is well understood, it is impossible to develop data models that adequately support them. For example, parallel database technology, coupled with parallel processing techniques, allows data to be partitioned to take advantage of the parallelism. However, if data is partitioned by month of the year, for example, and the analysis needs to be by geographical area, the partitioning may actually work against the parallelism and the system could develop hot spots, which slow down the overall response time and create frustrated and dissatisfied users. This requires that the work processes and activities also be part of the overall CRM architecture.

Finally, at the ground floor of all of this is the condition of the infrastructure of the organization. What kinds and how many processors will be required to perform the analyses? Should there be one intergalactic data warehouse, or should there be multiple data marts? What kinds of traffic will be expected on the network infrastructure, and is there sufficient bandwidth available to accommodate this load?

A typical architecture for a CRM system could include the following process components supported by a customer information platform.

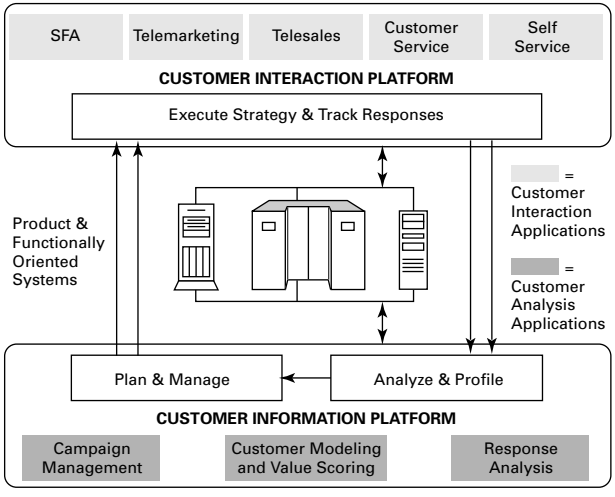

Figure 8: A Comprehensive CRM Architecture

Analysis/Profiling

The data warehouse is the technology and infrastructure heart of the architecture. All the marketing process components are enabled by the data warehouse and depend on the quality of the data warehouse for the accuracy of the derived results. The analysis and profiling activity uses a number of tools and technologies to derive knowledge about the organization's customers, as well as attribute the results of previous campaigns and customer interactions to the actions performed. This process utilizes various analytical tools, such as, OLAP, data mining, and statistical analysis. The analysis and profiling activity primarily evaluates previous customer treatment strategies and uses the analysis to modify different kinds of models that are used to indicate a particular customer, or customer segment's propensity to buy, churn, up-sell, cross-sell, etc. The results of the analysis and profiling activity are used extensively in the planning phase to construct the different promotional campaigns and treatment strategies, which are then fed to the customer interaction systems. The overall interaction workflow with customers is also modified and changed based on the results of the analysis and profiling process.

Campaign Planning and Management

Once the segmentation of the database has been completed and the analysts have finished the assesment phase, a variety of customer management strategies, ranging from customer treatment or customer value to customer offers, can be planned and initiated. One such strategy, campaign planning and management, helps tie the analysis and execution phases of the CRM cycle.

The results of each campaign must be tracked so organizations can learn from the success or failure of the promotion. After a number of campaigns have been run and results captured, those results can be analyzed, and adjustments can be made to targeting algorithms and models.

Customer Interaction

Organizations cannot truly develop a relationship with their customer base without instituting mechanisms for direct, bi-directional contact. These mechanisms are called "touchpoints" or "channels." They consist primarily of three architectural constructs:

- The sales force automation system that allows direct sales forces to promote products and gather customer feedback;
- The call center, which allows both customer-initiated direct contact and outbound telemarketing/telesales; and
- The expanding Web access component for self service.

The key to success here is the need for these touchpoints to be integrated into a cohesive whole. Customers must be able to get the same information from the Web, the sales force, or the call center. The sales force and call center must know the same facts about the customer and their history. Treatment strategies and campaigns must be consistent across these channels so as to ensure consistent relationship management. All of these elements must be connected within a "closed-loop" to make a CRM system successful.

The Role of a Methodology

As earlier noted, CRM integrates a number of different technologies, both in the data acquisition and analysis as well as in the creation and maintenance of the many touchpoints to the customer. This means

integrating many technologies, groups, and disciplines together to build a successful system. A methodology offers risk reduction in CRM projects by applying a structure and checklist for all to follow. A methodology helps avoid the all too familiar "slipping through the cracks" syndrome, which can spell disaster for a large, complex project. Furthermore, it also offers opportunities to apply best practices and adaptive techniques with appropriate feedback loops to keep the project on course.

Best Practices: Phased Implementations

Often, users and management alike are enthralled with the vision of the ultimate CRM system. In their zeal, the project implementation takes on the form of a behemoth project. These approaches simply do not work, and it is important for everyone to understand that up front.

One thing that is constant in today's environment is change. Everything changes, from technology to organizational structure, to the economy, and to market dynamics. Therefore, it is inescapable that the project be designed in a phased approach. This accomplishes three important goals. First and foremost, it makes the project manageable. Resources are easier to obtain and manage for a three to five month project, as opposed to an 18 to 24 month project and funding is easier to acquire. Secondly, in a well-structured plan, benefits can be inserted into the organization earlier, so that the company can be enjoying the fruits of phase 1 while phases 2 through 4 are being worked on. Finally, each phase gives the organization a chance to reassess how much was actually accomplished vs. what was expected to be accomplished, whether or not the original plan now needs modification to react to some change, and whether the rest of the milestones and even the ultimate end point needs to be reviewed.

Best Practices: Teaming/Dealing With the Systems Integrator

Unfortunately, in spite of its simplicity, the execution of the plan-the-work/work-the-plan concept is elusive to many IT professionals. Following are some examples of techniques that can effectively bring a system integrator into the project team.

- **Make the Integrator Part of the Team**
  The relationship with an integrator has to be one of teamwork.One effective way to make this work is to appoint two co-project leaders, one from the systems integrator and another from the organization. By developing a cross-functional team of internal and external professionals, companies can take advantage of the best practices the system integrators have to offer and apply what has been learned during previous implementation experiences to the current project.
- **Keep a Channel Open**
  Communications between the project team and the end users must be continuous and concise. This communication must also be written down clearly and mutually accepted. Every project should start off with a document which details the problem the organization is trying to solve, the scope of the project, the business objectives, and the approach or strategy which will be employed. These are a precursor to a detailed project plan, which must be accepted by both parties prior to commencing any work.

Best Practices: Managing the Change

Few things are more important in a project than defining a process for changing the scope, approach, or plan details. As part of this change management process, a log must be maintained recording any proposed changes, and whether those changes were driven by user demands or by implementation problems

encountered by the team that were not foreseen. A mechanism must be defined for determining whether this change carries with it additional fees or not. There must be a well-defined chain of command to the project and a well-defined escalation procedure to elevate problems to appropriate decision makers.

Best Practices: No Surprises

Involving end users from the beginning stages of defining the business requirements to the final production rollout is vital to avoiding any unexpected surprises in the project. It is imperative that users of the system be part of the implementation team and are involved in the decision-making process as it relates to the project. Regular team meetings ensure that all team members, and especially the end users, are updated on the progress of the project. A useful format for conducting these meetings is to have each team member report on the following three topics:

- What was accomplished last week?
- What is intended to be accomplished this week?
- What problems have come up or still exist since the last meeting?

Best Practices: Team Structure and Executive Sponsorship

Executive sponsorship is key to achieving cross-functional agreement and overcoming obstacles. It also means working with the different business units to change the way business is conducted today so as to take advantage of what the technology has to offer. The sponsors of the effort must enlist the support of top management, as there will be numerous issues which cross departmental lines and will need to be mediated or arbitrated by senior management. Experience has shown that forming a steering committee is extremely useful. The committee should be made up of senior management from both the client and the system integrator. The committee should meet no more than once a month and should be used to review deliverables, set direction, and resolve conflicts only after other lower-level negotiations have failed.

Achieving a Successful CRM Environment

Successful Planning: A Solutions Approach

The first task involved in creating a CRM environment is to build a solid and detailed plan which involves all aspects of the job: technology, process, and people.

Building all of the information required to make CRM a reality, creating the central repository, and integrating the different components is a difficult and complex undertaking. The importance of aligning and interfacing these systems with an organization's accepted business process cannot be overstated. It is an undertaking that calls for careful planning and the adoption of a complete solutions approach to its implementation. This means understanding the business processes that rely on the information housed in the system and understanding how analysts work and in what form they need various pieces of information. Another key element to successful planning may involve a formalized ROI approach. Although many marketing organizations may not follow a formalized ROI analysis, usually the CFO will not share that view, and so the financial benefits of a CRM initiative must be sold in terms relevant to the company's senior executives.

Successful Implementation: A Modular Approach

Selecting the right components for the CRM implementation is the next step in achieving the vision. This section will discuss the various components of the architecture and their salient characteristics. There are generally two classes of tools: those that assist in getting the data into the warehouse and those that assist analysts in getting information out. In either case, the fundamental purpose of these tools is to provide decision support simply. Users should be shielded from the complexities of getting at the data, processing it, and gaining insight from the analysis to develop actionable plans.

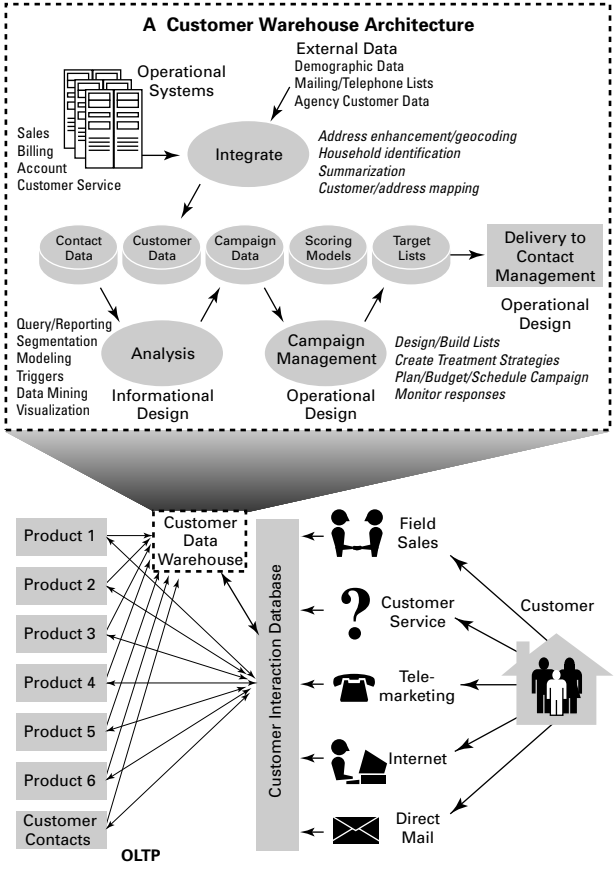

Figure 9: Functional Components of a CRM Architecture

The Data Warehouse—A Decision Driven Environment

The principal raison d'être for constructing the data warehouse is to allow marketers to make decisions on customer segmentation, matching products to customer profiles, and demographics/psychographics. These all involve multiple decisions with quick turnaround, which in turn depend on well-organized, accessible data and a set of tools to support those decisions.

The data warehouse then becomes the technology centerpiece of CRM; it is the officially recognized repository of all relevant data for anyone involved in the CRM process. It is organized in such a way as to support the decision making process of the marketing organization.

Analysts and marketers rely on the data in the warehouse to segment their target markets and decide upon promotions based on the results of the analysis. Call center agents look to the warehouse to supply them with information on the customers calling them, their purchases, their transaction history, and their history of complaints. The sales force relies on customer contact and demographic information as part of the SFA. And, all of the users rely on being able to capture results and feed them back into the warehouse to help make subsequent cycles of the process more productive and successful.

Therefore, it is imperative that the organization develop a good understanding of how it wants to implement the warehouse that drives the CRM system. There are a number of architectural constructs available, from central data warehouses to data marts to data malls and beyond. All have advantages and disadvan-

tages particular to specific implementations and any organization embarking on such a project has to understand in detail how the different pros and cons will impact their specific endeavor. However, the key to success is developing a good data model that will enable the entire customer management cycle and increase its velocity.

Data Archaeology

A fundamental premise of CRM is making decisions based on valid, well-organized and well-understood information. Therefore, the set of tools required to acquire and manage the information is of significant interest. Once the determination has been made that certain data elements are required to support the business decisions, the system of record for those data elements must be found. In many organizations, data is duplicated, either manually or automatically. Sometimes, the data is replicated and then the resulting duplicate database is separately managed, such that it grows differently from the original.

Depth and Breadth of Data

Data archaeology in a CRM environment largely consists of gathering information about customers and products from the universe of information systems. Subject areas in which information must be maintained include the following:

- *Contact Information:* Who is the customer and what is the best way to get in touch with them?
- *Household Information:* Who else are customers associated with, and is it possible to somehow take advantage of the family grouping to offer special promotions?
- *Group Information:* What kinds of groups are they associated with? For example, a popular way to offer credit cards today is via an association with university or frequent flyer groups.
- *Customer History:* What has this customer bought before, and when, and in response to what offers?
- *Promotion History:* How successful have promotions been? This will include large quantities of response data to the promotions and attribution models.
- *Product Purchases/Product Usage:* Can the kinds of behavior of this customer be classified in terms of size of purchases, kinds of products, etc.?
- *Transaction Rollup:* What can be determined by summarizing the different transactions with the customer?
- *Customer Service History:* Is this a high-maintenance customer?
- *Survey and customer response data.*
- *Demographic, psychographic, firmographic and/or credit data if available.*
- *Customer Interaction Information:* How does the customer like to interact with the company? What are the preferred channels, etc.?

Gathering, collating, integrating, matching, de-duplicating, and managing all of this information is a challenge, and success depends on a very well-defined process and best practices.

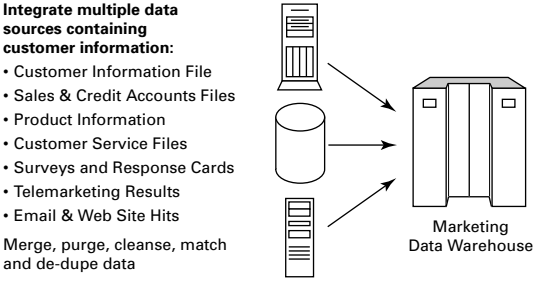

Figure 10: Data Cleansing and Transformation

Data Extraction and Cleansing

Once all the data elements have been found, the challenge is extracting them from the legacy systems, cleansing the data, and populating the database structures defined for the warehouse. Data must be cleansed because of the nearly universal lack of data standards in organizations today. The data extraction and cleansing tools transform data from operational systems on a regular basis and place them into the data warehouse according to the model defined.

Any single company's internal data resources, regardless of size, will necessarily be limited in scope regarding overall demographics and psychographics of the population as a whole. For example, analysts working in a CRM environment will want to see how the specific customer sets of that organization compare with the population as a whole.

It is important to note that, although some of the databases are external, the cleansing and transformation problem is still there. There is little likelihood that the target data model defined in response to the needs of the analysts and marketers of any one company will map exactly to the structures of the external databases. Furthermore, since many of these databases are purchased from independent organizations, it is also likely that they will suffer from lack of data standardization, and therefore it is important to understand the metadata of the external databases as clearly as those of the legacy systems.

Data Management and Storage

Successful implementation of CRM necessarily involves an appropriate structuring and management of the information. Therefore, within the data warehouse there are a number of segments or logical stores of information. A warehouse (or mart) may contain several years' worth of transactions by customer. Many retailers today will key their customer records on phone numbers or other similar identifier. For example, many retailers will always ask for a phone number at the register, and verify the customer's name and address from that number. (As a brief aside, it is interesting to note that use of a phone number may also help in householding analyses.) For many analyses, the use of detail data will slow the system down beyond tolerable limits and so in many cases, data is summarized and held in separate tables to speed up the analysis. Furthermore, in order to facilitate OLAP (see below), some data may have to be replicated in a star or snowflake schema to allow for dimensional analysis.

The sheer magnitude of the data forces many implementations to constrain the domain of the project. This challenge is depicted in the following diagram, showing the increase in database size per unit of information as the number of precalculated dimensions grows beyond seven. (Source: *The OLAP Report*).

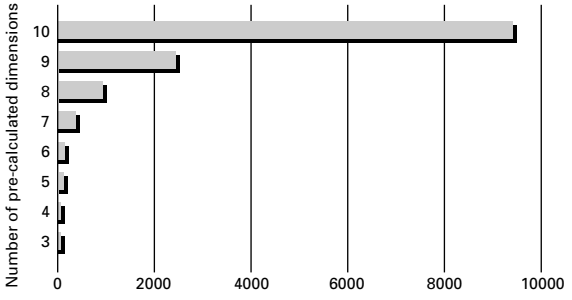

Figure 11: Database Size per Unit of Input Data

The warehouse will also keep scores from OLAP or data mining applications, campaign histories, contact history from customer interaction systems, segmentation models, and other telemarketing and direct mail data about customers.

Lists of prospective customers are also contained within the warehouse. These can be sourced from promotions or outside agencies. The handling of these may require special procedures since some legal

requirements exist regarding what type of information can be purchased and how it is to be used (e.g. banking and credit card operations).

All of this means that CRM warehouse databases can be gargantuan in size, with many organizations now requiring multi-terabyte warehouses. This not only presents problems in managing the warehouse—taking backups and so on—but it also impacts the operational aspects of refreshing the warehouse. Warehouses are typically updated on some periodic basis that matches the business needs: quarterly, monthly, weekly, or even daily in some rare instances. However, updating a terabyte data store is no mean feat, and as the size of the uploads increase, the time window during which analysts may not use the system also increases, causing frustration and delays.

Scalability and Open Technologies

Data warehousing projects usually fall into two camps, when viewed from the perspective of their success: those that grow wildly and those that die off. Particularly for companies that choose to implement data marts, initial success in one department leads to increased demand within that department and across to other departments. Part of the reason for this is that warehousing technologies offer the promise of empowering the end user to derive answers independently of the IT organization, which traditionally has been overloaded with requests for special queries. This provides a win-win situation, allowing IT to concentrate on their development backlog while simultaneously providing users the answers they need.

In any case, success feeds on itself and more and more users will pile more and more applications onto the warehouse, growing the size of the database, growing the user population, growing the number and complexity of queries as the users become more sophisticated. It is at this point that the warehouse either continues to grow or dies off.

If the warehouse was designed originally in a scalable fashion, such that it could accommodate the increases in database size, number of queries, and complexity of queries, the success rate continues to climb. If, on the other hand, the system begins to stagger and hang; if queries which originally took two minutes begin to take 20 minutes to run, users will abandon the system without hesitation. All of this dictates an open design, which can be scaled by adding incremental processing power, additional databases, and even spinning off separate data marts with relative ease.

There are other reasons for moving to industry-standard and open architectures. Many campaign management tools are designed to work against databases stored in a warehouse. This means that the warehouse must be built with a common database technology accessible by the broadest possible spectrum of products. In today's market, the only answer that makes any sense here is to use open, relational database products that are supported by not only campaign management and data warehousing tools, but also by call centers and sales force automation packages.

Data Analysis and Query Tools

Once the warehouse has been populated it can finally be exploited by a set of tools that divide it into three categories: query, data mining, and statistical analysis.

Query tools usually refer to traditional report writers, but have been extended to include On-Line Analytical Processing (OLAP). The concept of OLAP is to provide user tools to analyze a multidimensional hypercube of information, such as sales of product X across geographic regions, by quarter, by division, by model number, and by profitability. A user/analyst could then take that hypercube and "slice and dice" it, getting system reports on which products performed best by geography, by profitability, by quarter, and so on.

Some OLAP vendors roll OLAP technology in with data mining, because they claim that OLAP allows for the discovery of patterns in the data. However, there is a significant difference between OLAP and data mining. OLAP is an interactive tool by which a user drives subsequent levels of analysis to unearth the patterns. As will be pointed out, data mining is fully automated and returns results without user intervention.

Data mining and statistical analysis tools are both used to further analyze and segment customers, in terms of buying patterns and behavior, preferences, and profitability. Data mining is a relatively recent addition to the arsenal of warehousing analytical tools. These tools look for hidden relationships in the data and are very compute intensive, using pattern recognition technologies like neural networks and genetic algorithms. They could be batch jobs that run during off-hours. Analysts interact with the output, make changes, and submit the data again to the algorithms. These models can be difficult to build and require some skill in formulating the data mining requests. If not well defined and designed, data mining results can be disappointing, pointing out obvious patterns such as people who eat baby food are typically under three years old. One important dimension of a CRM system becomes the degree to which the existing staff has the skills to build these models.

Another set of tools in the arsenal is statistical analysis. In contrast to data mining tools, these techniques are used extensively to build association trees, which help to classify and segment the information to identify customer profiles. For example, a CHAID analysis of a database might conclude that a certain percentage of the population that bought product X had household income of over $50,000 per year, lived in the Northeast, and the product was primarily purchased by males between 35 and 50 years of age. If the percentages of the affinities are sufficiently compelling, a promotion might be initiated to specifically try and sell the target population Product X.

It is important to note that OLAP and statistical analysis tools are not only useful in developing the promotions and plans, but also in gauging the effectiveness of previous campaigns. Patterns relating to which promotions worked better than others, and the underlying reasons for those differences, are of tremendous value in refining future campaigns.

Conclusion

The rewards of executing an effective CRM program are largely self-evident: increased customer value, higher customer retention, increased customer recruitment, and higher profitability. The traps and pitfalls or lessons learned are equally important as they can affect the outcome and prevent a CRM from achieving its potential. Some final recommendations:

- Start with a clear vision of a customer-centric, as opposed to a product-focused approach;
- Remain focused on a disciplined and structured marketing process driven by detailed data;
- Retain business focus—don't get distracted by the technology;
- Ensure the technology infrastructure can adequately capture and track promotional history and customer responses to campaigns;
- Recognize that multiple steps may be required to move from the current state to the ideal;
- A marketing warehouse infrastructure must exist for any serious CRM initiative;
- Data mining and analytical tools add complexity

to a marketing warehouse but provide powerful behavioral and segmentation insights efficiently;

- New campaign management technology can augment previous investments in targeting and other marketing programs; and
- Reduce the risk associated by building a project team made up of business users, IT professionals, executive management, and external consultants as appropriate.

If organizations follow the steps outlined in this guide, they should be well on their way to building the IT foundation of a customer-centric business strategy.

Glossary

Accuracy—A measure of the effectiveness of a data mining model. The goal of a successful data mining project is to deliver results which are significantly better than a random mass mailing. Therefore, the measure of the success of a data mining project can be measured in the accuracy of the results: how effective was the prediction of the model.

Actionable strategies—Strategies formulated in such a way as to be able to construct action plans, consisting of concrete activities which are shown to support and further the strategy.

Ad-Hoc Query—Any query that cannot be determined prior to the moment the query is issued. A query that consists of dynamically constructed SQL, which is usually constructed by desktop-resident query tools.

Aggregates—Facts added together, or "aggregated" to form summaries of information.

Backpropagation Neural Network— Neural network structures in which connections exist between a layer and one or more previous layers. Backpropagation allows a neural network to adjust itself and "learn" from successive attempts.

Bitmapped Indexing—A family of advanced indexing algorithms that optimize RDBMS query performance by maximizing the search capability of the index per unit of memory and per CPU instruction. Properly implemented, bitmapped indices eliminate all table sans in query and join processing.

Business Drivers—Business Drivers are external forces which propel a business in certain directions. For example, the pace of technology forces businesses to modernize their technology holdings on a regular basis to remain competitive.

Business-Driven Approach—The process of identifying the data needed to support business activities, acquiring or capturing those data, and maintaining them in the data resource.

Campaign Management—The business process that is put in place to manage the lifecycle of a marketing campaign. Campaign management measures the yield or return from an effort to reach a set of customers. A marketing campaign is a set of promotions which are directed at a specific set of customers to get them to buy specific products and/or services. Managing the campaign involves coordinating the activities such as market segmentation and telemarketing with the collateral information required for each step.

Change Data Capture—The process of capturing changes made to a production data source. Change data capture is typically performed by reading the source DBMS log. It consolidates units of work, ensures data is synchronized with the original source, and reduces data volume in a data warehousing environment.

Chi Square Automatic Interaction Detection (CHAID)—A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will will have a given outcome. Segments a dataset by using chi square tests to create multi-way splits.

Classification—The process of dividing a data set into mutually exclusive groups such that the members of each group are as "close" as possible to one another and different groups are as "far" as possible to one another, where distance is measured with respect to specific variable(s) you are trying to predict. For example, a typical classification problem is to divide a database of customers into groups that are as homogeneous as possible with respect to a creditworthiness variable with values "good" and "bad."

Classification and Regression Trees (CART)—A decision tree technique used for classification of a dataset. Provides a set of classification rules that you can apply to a new (unclassified) data set to predict which records will have a given outcome. Segments a dataset by creating 2-way splits. Requires less data preparation than CHAID.

Cluster Sampling—The selection of a subset of a clustered set of data.

Control Groups—A component of an experiment in which a group other than the targeted group is utilized, usually via a randomly selected list. The purpose of the control group is to provide a baseline against which to measure the success or failure of a marketing campaign by noting the difference in response between the target group and the control group.

Corporate Data— All the databases of the company. This includes legacy systems, old and new transaction systems, general business systems, client/server databases, data warehouses, and data marts.

Critical Success Factors—Key areas of activity in which favorable results are necessary for a company to reach its goal. For example, a retail institution needs to have excellent customer service systems to be successful.

Cross Selling—The business techniques of associating like products to customer purchases. For example, if a customer buys letterhead there is an opportunity to sell envelopes in the same transaction.

Customer Acquisition—The goal of marketing programs that seek to increase the customer base by using effective and efficient identification and acquisition strategies.

Customer-Centric View—A view of business that puts the needs and desires of a customer as a principal criterion in any decision that impacts the customer.

Customer Information Systems—Systems which are geared to providing companies information about the purchasing preferences of their customers. These systems are used to identify potential customers and to retain existing customers, as well as to find out what products and services should be promoted to which segment of the customer population.

Customer Relationship Management(CRM)—The set of decisions that guide the actions of an organization in developing a positive relationship with customers. Customer buying profiles and churn analysis are two examples of decision support activities that can significantly affect the success of customer relationships.

Customer Retention—(1) The goal of marketing programs that seek to maintain high levels of customer continuance. The cost of retaining customers is significantly less than the cost associated with acquiring a customer (2) A measure of the success of customer service organizations, customer retention is ensuring your customers continue to buy from you in the future.

Daily Update Window—An allowable period of time by which a database must be refreshed so its users can perform analysis of it.

Data Accuracy—The component of data integrity that deals with how well data stored in the data resource represent the real world. It includes a definition of the current data accuracy and the adjustment in data accuracy to meet the business needs.

Data Architecture—The component of the data resource framework that contains all activities, and the products of those activities, related to the identification, naming, definition, structuring, quality, and documentation of the data resource for an organization.

Data Cleansing—The process of manipulating the data extracted from operational systems so as to make it usable by the data warehouse. When loading data from existing operational systems, it is likely that few if any of the operational systems will have data to present in a format which is compatible with the data model developed for the warehouse. For example, a product number may be held as a numeric field in one system while a second system appends an alpha suffix to the number for reporting purposes.

Data Completeness—An indication of whether or not all the data necessary to meet the current and future business information demand are available in the source systems data resource. It deals with determining the data needed to meet the business information demand and ensuring those data are captured and maintained in the data resource so they are available when needed.

Data Concurrency—The situation where the replicated data values at are synchronized with the corresponding data values at the official data source. When the data values at the official data source are updated, the replicated data values must also be updated so they are consistent with the official data source.

Data Conversion—The process of changing data from one physical environment to another. This process makes any changes necessary to move data from one electronic medium or database product to another.

Data Extract—Data which normally resides on an operational system and which is removed from that system for loading into a Data Warehouse.

Data Mapping—The process of assigning a source data element to a target data element.

Data Mart—A subset of the data resource, usually oriented to a specific purpose or major data subject, that may be distributed to support business needs. The concept of a data mart can apply to any data whether they are operational data, evaluational data, spatial data, or metadata.

Data Mining—(1) The process of utilizing the results of data exploration to adjust or enhance business strategies. It builds on the patterns, trends, and exceptions found through data exploration to support the business. It is also known as data harvesting. (2) A technique using software tools geared for the user who typically does not know exactly what he's searching for, but is looking for particular patterns or trends. Data mining is the process of sifting through large amounts of data to produce data content relationships. This is also known as data surfing.

Data Mining Methodology—A set of processes and tools that in combination provide a means of implementing a data mining application. Methodologies are primarily used to standardize the approach to implementing products, and as a byproduct, increase consistency and reduce error.

Data Modeling—A method used to define and analyze data requirements needed to support the business functions of an enterprise. These data requirements are recorded as a conceptual data model with associated data definitions. Data modeling defines the relationships between data elements and structures.

Data Repository—A logical (and sometimes physical) partitioning of data where multiple databases which apply to specific applications or sets of applications reside. For example, several databases (revenues, expenses) which support financial applications (A/R, A/P) could reside in a single financial data repository.

Data Restructuring—The process to restructure the source data to the target data during data transformation.

Data Source—A specific data site where data are stored and can be obtained. Any source of data from a specific organization, such as a data base or data file. A data source may include non-automated data, but it does not include unpublished documents containing data.

Data Store—A place where data is stored; data at rest. A generic term that includes databases and flat files.

Data Synchronization—The process of identifying active data replicates and ensuring that data concurrency is maintained. Also known as data version synchronization or data version concurrency because all replicated data values are consistent with the same version as the official data.

Data Transformation—(1) The formal process of transforming data in the data resource within a common data architecture. It includes transforming disparate data to an integrated data resource, transforming data within the integrated data resource, and transforming disparate data. It includes transforming operational, historical, and evaluational data within a common data architecture. (2) Creating "information" from data. This includes decoding production data and merging of records from multiple DBMS formats. It is also known as data scrubbing or data cleansing.

Data Visualization—The process of creating and presenting a chart from a set of data based on a set of attributes. It deals with understanding patterns, trends, and relationships in historical data, and providing visual information to the decision maker.

Data Warehouse—(1) A subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process. A repository of consistent historical data can that can be easily accessed and manipulated for decision support. (2) An implementation of an informational database used to store sharable data sourced from an operational database-of-record. It is typically a subject database that allows users to tap into a company's vast store of operational data to track and respond to business trends and facilitate forecasting and planning efforts.

Database Marketing—A term used to describe the art/science of selecting a database of a potential set of customers for a given product or need. For example, defining a target mailing list for people likely to acquire a new mutual fund product.

Decision Support—A set of software applications intended to allow users to search vast stores of information for specific reports which are critical for making management decisions.

Decision trees—A tree-shaped structure that represents a set of decisions. These decisions generate rules for the classification of a dataset.

Demographic Data—Any data that locate, identify, or describe populations and their properties or characteristics. For example, demographic data will describe the age groups of people living in certain geographies, perhaps in certain income categories. Other dimensions or characteristics of demographic data include race, religion, political preferences, spending preferences, family size, and so on.

Derived Data—(1) Data that is the result of a computational step applied to reference or event data. Derived data is the result of either relating two or more elements of a single transaction (such as an aggregation), or of relating one or more elements of a transaction to an external algorithm or rule. (2) Data that are derived from other data through a data derivation procedure, not by the measurement or observation of an object or event.

Derived Data Maintenance—The process for ensuring that active derived data are properly rederived when their contributing data characteristics values change or when new contributing data characteristics appear.

Dimension—In data analysis, dimensions are variables in a situation. For example, time, product type, region are three dimensions of a sales' situation: product types are sold over time in different regions.

Direct Marketing—A technique which brings the vendor's message directly to a market segment which has been identified as a potential buyer for the goods and services. Methods include direct mail and telemarketing.

Drill Down—A method of exploring detailed data that was used in creating a summary level of data. Drill down levels depend on the granularity of the data in the data warehouse.

Enterprise Data Warehouse—A Data Warehouse implementation in which a single warehouse serves the need of several business units simultaneously with a single data model which spans the needs of the multiple business divisions.

Executive Information Systems (EIS)—Tools programmed to provide canned reports or briefing books to top-level executives. They offer strong reporting and drill-down capabilities. Today, these tools allow ad-hoc querying against a multi-dimensional database, and most offer analytical applications along functional lines such as sales or financial analysis.

Extract—A set of data which resides normally on the operational systems which is uploaded into the data warehouse.

Functional Data Warehouse—A warehouse that draws data from nearby operational systems. Each functional warehouse serves a distinct and separate group (such as a division), functional area (such as manufacturing), geographic unit, or product marketing group.

Genetic Algorithms—An evolutionary algorithm that generates each individual from some encoded form known as a "chromosome." Chromosomes are combined or mutated to create new individuals.

"Crossover," the kind of recombination of chromosomes found in sexual reproduction in nature, is often also used in GAS. Here, an offspring's chromosome is created by joining segments chosen alternately from each of two parents' chromosomes which are of a fixed length. GAS are useful for multidimensional optimization problems in that they can encode the values for the different variable being optimized.

Geocoding—Classification of customers based on geographic location.

Geographic Information System—An organized collection of computer hardware, software, geographic data, and personnel designed to efficiently capture, store, update, manipulate, analyze, and display all forms of geographically referenced information.

Horizontal Data Aggregation—The edge-matching of data between different sets of data.

Horizontal Data Replication—The formal process of creating an exact copy of official data and placing that exact copy at one or more data sites.

Householding—A methodology of consolidating names and addresses and grouping them for business purposes. For example, if multiple members of the same family have accounts at a certain bank, the bank may grant them certain privileges based on their combined accounts. Householding allows a bank to understand that the multiple accounts are related.

Incremental Refresh—A technique which loads only data which has changed since the last load into a data warehouse or data mart.

Induction—A method of proving statements about well-ordered sets. If S is a well-ordered set with ordering "<", and we want to show that a property P holds for every element of S, it is sufficient to show that, for all s in S, (IF for all t in S, t<s=>P(t) THEN P(s)) i.e. if P holds for anything less that s then it holds for s. In this case, we say P is proved by induction.

Intelligent Agent—A software routine that waits in the background and performs an action when a specified event occurs. For example, agents could transmit a summary file on the first day of the month or monitor incoming data and alert the user when certain transactions have arrived.

Intranet—A private network inside a company or organization that uses the same kinds of software that you would find on the public Internet, but that is only for internal use. As the Internet has become more popular, many of the tools used on the Internet are being used in private networks. For example, many companies have Web servers that are available only to employees.

Inverted File Indexes—A more efficient method to access data in an ad-hoc or analysis environment. It maintains indexes to all values contained in an indexed field. Those values, in turn, can be used in any combination to identify records that contain them, without actually scanning them from disk.

Java Database Connectivity (JDBC)—A standard implemented by Sun Microsystems that allows developers to write applications in the Java language, without having to use database-specific code

Lifestyle Data—Similar to demographic filtering of databases, a number of databases are available which cluster groups of neighborhoods according to multivariate criteria, identifying homogeneous lifestyle habits in certain areas of the country. These are used by marketers to target and direct campaigns at appropriate clusters.

Lifetime Value Ranking—A technique for understanding the value of customers by ranking them according to their potential lifetime value. Lifetime Value is defined by Don Peppers and Martha Rogers

in their book *Enterprise One-on-One* as "the stream of expected future profits, net of costs, on a customer's transactions, discounted at some appropriate rate back to its current, net present value."

Lift Chart—Also called a "Gains Chart." A chart in which all observations from the scored data set are sorted from highest expected profit to lowest expected profit. Then, the observations from the scored data set are sorted from highest expected profit to lowest expected profit. Then, the observations are grouped into cumulative deciles. The positive responses for all observations within each decile are tallied within decile. If the model used to create the scored data set has no predictive power, then the distribution of positive responses will be concentrated in the highest deciles.

Logical Data Model—(1) A data model that represents the normalized design of data needed to support an information system. Data are drawn from the common data model and normalized to support the design of a specific information system. (2) Actual implementation of a conceptual module in a database. It may take multiple logical data models to implement one conceptual data model.

Market Basket Analysis—A technique employed in the retail industry where the contents of "typical" sets of purchases are analyzed to determine statistically which products are selling best, what kinds of affinities might exist between and among products, and similar associations.

Metadata—Metadata is information about the data stored in a warehouse. Examples of metadata include data element descriptions, data type descriptions, attribute/property descriptions (alpha/numeric), range/domain descriptions (allowable zip codes), and process/method descriptions. The repository environment encompasses all corporate metadata resources: database catalogs, data dictionaries, and navigation services. It insulates the data warehouse from changes in the schema of operational systems.

Model—A model is a representation of a situation designed to answer questions about that situation. In a data mining scenario, models are built to reflect how customers may react to different campaigns, and to predict (answer the question) which customers are most likely to buy certain products and/or services.

Multi-Dimensional Analysis—Informational Analysis on data which takes into account many different relationships, each of which represents a dimension. For example, a retail analysis may want to understand the relationships among sales by region, by quarter, by demographic distribution (income, education level, gender), by product. Multi-dimensional analysis will yield results for these complex relationships.

Neural Networks—Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Normalization—The process of reducing a complex data structure into its simplest, most stable structure. In general, the process entails the removal of redundant attributes, keys, and relationships from a conceptual data model.

ODBC—Open Database Connectivity. A standard for database access co-opted by Microsoft from the SQL Access Group consortium.

OLAP—On-Line Analytical Processing, originally introduced in 1994 in a paper by E. F. Codd, is a decision support counterpart to On-Line Transaction Processing. OLAP allows users to derive information and business intelligence from data warehouse systems by providing tools for querying and analyzing the information in the Warehouse. In particular, OLAP allows multidimensional views and analysis of that data for decision support processes.

OLTP—On-Line Transaction Processing.

Operational Data Store—Contains timely, current, and integrated information. The data is typically very granular. These systems are subject oriented, not application oriented, and are optimized for looking up one or two records at a time for decision making.

Operational Database—The database-of-record, consisting of system-specific reference data and event data belonging to a transaction-update system. It may also contain system control data such as indicators, flags, and counters. The operational database is the source of data for the data warehouse. It contains detailed data used to run the day-to-day operations of the business. The data continually changes as updates are made, and reflect the current value of the last transaction.

Overlays—A set of demographic filters used to screen out only those individuals who fit a target profile. An overlay typically will consist of demographic data such as income and education, as well as cluster codes.

Parallelism—The ability to perform functions in parallel.

Patterns—Associations among data in a database which recur with some frequency. A pattern may be the fact that whenever product A is purchased, 73% of the time a customer will also purchase product B; or product C is most often purchased at a particular time of year. In each of these examples, there is an association or linking of two or more facts to form a pattern.

Physical Data Model—A data model that represents the denormalized physical implementation of data that support an information system. The logical data model is denormalized to a physical data model according to specific criteria that do not compromise the logical data model but allow the database to operate efficiently in a specific operating environment.

Promotional History—Information regarding the number of times a prospect has been contacted, what was sent, and his or her response. This is used frequently in building predictive models in a target marketing situation.

Query Governor—A facility that terminates a database query when it has exceeded a predefined threshold.

Random Sampling—The selection of a set of data points from the entire data set, selected at random for purposes of analysis.

Redundant Data—The situation where the same data characteristic exists at two or more data sites. Redundant data are created, stored, and maintained independent of each other and are often unknown to the organization.

Referential Integrity—The part of data structure integrity that ensures a parent data occurrence exists for each subordinate data occurrence. A subordinate data occurrence cannot be added if there is no parent data occurrence, and a parent data occurrence cannot be deleted if subordinate data occurrences still exist.

Regression—A form of statistical analysis in which the behavior of one variable is predicted from the behavior of others.

Relational On-Line Analytic Processing (ROLAP)—OLAP based on conventional relational databases rather than specialized multi-dimensional databases.

Relationship Profitability—A way to measure the value of a relationship by measuring how profitable each relationship is.

Sampling—The use of a subset of the population under study to determine the behavior of the entire population by the behavior of a subset.

Schema—(1) A diagrammatic representation of the structure or framework of something. (2) The logical and physical definition of data elements, physical characteristics and inter-relationships.

Scoring—The process of computing outputs by applying a model to data.

Segmentation—Techniques for deriving clusters and classes by creating two-way and multi-way splits from a dataset according to specified variables.

Slice and Dice—A term used to describe a complex data analysis function provided by MDBMS tools.

Summarization Tables—These tables are created along commonly used access dimensions to speed query performance, although the redundancies increase the amount of data in the warehouse.

Target Marketing. The concept of identifying groups of people (targets) who are more likely to buy your product or service and then delivering the message directly through direct marketing techniques.

Temporal Data—Any data that represent a point in time or a time interval. They are data with a time component.

Training Data—Data that contain input and target values that are used for training a model. For example, in a data mining environment a neural network may be trained to look for certain patterns by providing the network with a data set of known matching patterns.

Validation Data—Data that are used indirectly during training for model selection, for early stopping, or for other methods intended to improve generalization.

Validation Data Set—A set of demographic and historical data against which hypotheses are tested during the building of the models. Once the model is trained, a validation dataset with known observable patterns is presented to see if the model is capable of finding the subject patterns.

Vertical Data Aggregation—The summarization of data to higher levels of generalization.

Vertical Data Modeling—The process of moving through the logical schema, tactical schema, and strategic schema. Transforming a general schema to a more detailed schema is a specialization process, and transforming a detailed schema to a more general schema is a generalization process.

NOTES

NOTES

NOTES

NOTES

This Technology Guide is one in a series of topic-focused Guides that provide a comprehensive examination of important and emerging technologies.

This series of Guides offers objective information and practical guidance on technologies related to Communications & Networking, the Internet, Computer Telephony, Document Management, Data Warehousing, and Enterprise Computing.

Built upon the extensive experience and ongoing research of our writers and editorial team, these Technology Guides assist IT professionals in making informed decisions about all aspects of technology development and strategic deployment.

techguide.com is supported by a consortium of leading technology providers. IBM has lent its support to produce this Guide.

Visit our Web Site at

www.techguide.com

to view and print this Guide, as well as all of our other Technology Guides.

This is a free service.

CP-1324 REV. 3/99

techguide.com