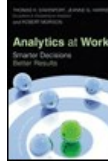# Analytics at Work: Smarter Decisions, Better Results

by Thomas H. Davenport, Jeanne G. Harris and Robert Morison
Harvard Business Press. (c) 2010. Copying Prohibited.

Chapter 2: Data – The Prerequisite for Everything Analytical

You can't be analytical without data, and you can't be really good at analytics without really good data. Now, on to the next chapter!

Well, in case you need more information, the rest of this chapter is about the data environment your organization needs to become more analytical. We'll begin by describing the key components of data management for analytics as employed by the most sophisticated or "stage 5" companies. As we suggest in the introduction to part one, not every organization needs to be at stage 5 for data, but unlike other topics in this book, data management is best addressed by considering how close your organization can come to this ideal. Next we'll discuss how to progress from stage to stage to improve your data and data management for analytics—and even if you're already pretty good, you still need to know about stage 5 data management for analytics.

Here's what you need to know about data, moving from the most fundamental issues onward: structure (what is the nature of the data you have?), uniqueness (how do you exploit data that no one else has?), integration (how do you consolidate it from various sources?), quality (how do you rely on it?), access (how do you get at it?), privacy (how do you guard it?), and governance (how do you pull it all together?). We'll take each topic in turn.

Structure

Companies basically have a choice of three methods of structuring data for analysis: "cubes," arrays, and nonnumeric. If you are tempted to stop reading and tune into ESPN or the Weather Channel, stick with us for a while—the topic is less dry than it sounds. How your data is structured matters because it affects the types of analyses you can do.

Data in transaction systems is generally stored in tables. Tables are very good for processing transactions and for making lists, but less useful for analysis. (One reason: tables rarely contain historical data—three to twelve months at most.) So when data is extracted from a database or transaction system and stored in a warehouse, it frequently is formatted into "cubes." Data cubes are collections of prepackaged multidimensional tables. For example, sales by region by quarter would yield a conventional, three-dimensional cube. However, (unlike the physical world) "data cubes" can have more than three dimensions (though more than four or five can be confusing to carbon-based life forms). Cubes are useful for reporting and "slicing and dicing" data, but less useful for analytical exploration because the variables they contain are limited to what some analyst thought should be in the cube and in the resulting report.

Data arrays consist of structured content, such as numbers in rows and columns (a spreadsheet is a specialized form of array). By storing your data in this format, you can use a particular field or variable for analysis if it is in the database. Arrays may consist of hundreds or even thousands of variables. This format allows for the most flexibility, but may be confusing to nontechnical users who don't understand the structure of the database or the locations and fields of the data within it.

Unstructured, nonnumeric data—the "last frontier" for data analysis—isn't in the formats or content types that databases normally contain. It can take a variety of forms, and companies are increasingly interested in analyzing it. You may hypothesize, for example, that the vocal tone of your customers during service calls is a good predictor of how likely they are to remain customers, so you would want to capture that attribute. Or you may analyze social media—blogs, Web pages, and Web-based ratings and comments—to understand consumer sentiments about your company. In this case, the entire Internet becomes the data warehouse (although you may want to extract and copy some of it for detailed analysis). Firms are also increasingly interested in mining text in internal databases—like warranty reports and customer complaint letters—for customer services issues, "reason fields" (for example, in denying credit), and product descriptions (for example, to reconcile multiple product hierarchies following mergers and acquisitions). There's potential value in unstructured data, but just like mining for gold, you have to sort through a lot of dirt to find what you want. For example, words that you really care about—like *fire*—can have a variety of meanings, so you have to do some semantic analysis to be sure you're getting the meanings you want.

Highly analytical (stage 5) organizations, then, engage in many different projects involving both cubes and arrays. They also tend to use, or at least experiment with, a wider variety of data—not just numbers, but data like images, Web text, and voice analyses.

Uniqueness

How can you tap into and exploit data that no one else has? Inevitably, companies that have the same data will have similar analytics. To get an analytical edge, you must have some unique data. For instance, no one else knows what your customers bought from you—and you can certainly get value from that data. But deciding what information is valuable and going out and getting proprietary data that doesn't exist in your or anybody else's organization is a different matter, and may require creating a new metric.

As Al Parisian, chief information officer and head of strategic planning for Montana State Fund, notes, "You are what you eat with regard to data . . . Just as a seriously health-oriented person must get much more engaged with what they eat, those who are serious about informing fact-based leadership must get very engaged with data."[1] We concur, and conclude that a unique strategy requires unique data. Since stage 5 organizations by definition seek an edge with their analytical capabilities, they need to seek data that other firms don't have or use.

There are several levels of unique data: one is simply to be the first company in your industry to use commercially available data. Progressive Insurance did this in 1996 when it began to use consumers' credit scores as an input to its automobile insurance underwriting. Whether you pay your bills turned out to be a surprisingly good predictor of whether you will crash your car—no one knows exactly why—but it took other firms in the industry at least four years to start using this data (and some still haven't caught on).

It was inevitable that competitors would catch up to Progressive, because anybody can buy a credit score and competitive secrets don't stay secret for long—particularly in the insurance industry, which has to publish underwriting approaches in regulatory filings. Nevertheless, Progressive kept innovating in other areas, as we describe in chapter 9, proving that even if you're using industry-standard data and your

competitors have glommed onto the idea, it's still possible to differentiate your company with that data. Capital One, for example, made extensive use of consumer credit scores for extending credit and pricing in its credit card business, but soon most of its competitors followed suit. So it started to fool around with the credit score data, determining through detailed analysis that some low-score applicants might be more likely to pay back their loans than the score would predict. By identifying some data that differentiates customers, it was able to differentiate its own services—even though the initial input was a widely employed data source.

Of course, it's easier to get proprietary advantage when the data is sourced from internal operations or customer relationships. Let's look at some examples of the latter:

- Olive Garden, an Italian restaurant chain owned by Darden Restaurants, uses data on store operations to forecast almost every aspect of its restaurants. The guest forecasting application produces forecasts for staffing and food preparation down to the individual menu item and component. Over the past two years, Darden has reduced unplanned staff hours by more than 40 percent and cut food waste by 10 percent.[2]

- The Nike+ program uses sensors in running shoes to collect data on how far and fast its customers run. The data is uploaded to the runner's iPod, and then to the Nike Web site. Through analysis of this data, Nike has learned that the most popular day for running is Sunday, that wearers of Nike+ shoes tend to work out after 5 p.m., and that many runners set new goals as part of their New Year's resolutions. Nike has also learned that after five uploads, a runner is likely to be hooked on the shoe and the program.[3]

- Best Buy was able to determine through analysis of its Reward-Zone loyalty program member data that its best customers represented only 7 percent of total customers, but were responsible for 43 percent of its sales. It then segmented its stores to focus on the needs of these customers in an extensive "customer centricity" initiative.

- In the United Kingdom, the Royal Shakespeare Company carefully examined ticket sales data that it had accumulated over seven years to grow its share of wallet of existing customers—and to identify new audiences. Using audience analytics to look at names, addresses, shows attended, and prices paid for tickets, the RSC developed a targeted marketing program that increased the number of "regulars" by more than 70 percent.[4]

- Consumer packaged goods companies often don't know their customers, but Coca-Cola has developed a relationship with (mostly young) customers through the MyCokeRewards.com Web site, which the company believes has increased its sales and allowed it to market to consumers as individuals. The site attracts almost three hundred thousand visitors a day—up 13,000 percent from 2007 to 2008.[5]

- A top ten U.S. bank with over three thousand branches found it nearly doubled the balance per customer interaction by using collaboration-based analytic technology when dealing with customers. First-year profitability per interaction increased by 75 percent after taking into account the additional value provided to the customers by the bank. Sales productivity of front-line bank staff also increased by almost 100 percent in terms of balances sold per hour.

Of course, data that was once unique and proprietary can become commoditized too. For example, every airline has a loyalty program, but these programs all offer similar benefits, and the data from them is not generally used to create and maintain strong relationships with customers. At one point these programs were great, but now they are simply a me-too capability. There is probably some potential for an airline to break out from the pack and do something distinctive with its loyalty data, but most airlines are perhaps too preoccupied with fuel costs and mergers to seize this opportunity.

Data gold mines can also potentially come from basic company operations, if the company realizes their value. For example, Cisco Systems has been maintaining the data (and increasingly voice) networks of its customers for years. Recently, the company realized that it could analyze the data on network configurations to identify which customers were mostly likely to be facing a network failure and would need to upgrade equipment. Cisco can benchmark and analyze a customer's network and all its component products across multiple dimensions—the network's configuration, its use, the position of devices in the network, and so forth. It can predict the network's stability and anticipate pending problems like "toxic combinations" of network equipment. Cisco analysts can also compare a network's likely stability to that of others in the same industry or of similar size. The ability to do such diagnoses differentiates Cisco's services and improves sales of its products.

Many more organizations will, we predict, realize that their operational data is an important asset. Delta Dental of California realized that by analyzing years of claims data, it could begin to understand patterns of behavior among insured customers and the dentists it pays: are a particular dentist's patients developing more problems than others? Are root canals more common in some areas than others? Another health insurer realized that it could identify older insured customers at risk of diabetes from inactivity, and now works to head off the disease through a program called Silver Sneakers Steps (run by Healthways, a disease management company), which uses a pedometer to measure daily steps taken.

A proprietary performance metric can also lead to improved decision making that can differentiate one company from another. Wal-Mart used the ratio of wages to sales at the store level as a new indicator of performance. Marriott created a new revenue management metric called "revenue opportunity" that relates actual revenues to optimal revenues at a particular property. Even if the metric already exists in other industries, if it's not yet used in yours, it can create some value. Harrah's, for example, imported the metric of "same store sales" from the retail industry, and was the first to employ it in the casino business. Harrah's also measured the frequency of employee smiles on the casino floor, because it determined that they were positively associated with customer satisfaction. All that remains for Harrah's, seemingly, is correlating a player's success in craps against the number of times he blows on a pair of dice.

Regardless of the source of proprietary data, any organization that wants to succeed with analytics needs to start identifying some data that it alone possesses. The next decade is going to see an explosion of attempts to analyze proprietary data. Stage 5 companies are doing it today.

[1]Interview with Al Parisian, January 2009, and e-mail correspondence.

[2]Chuck Salter, "Why America Is Addicted to Olive Garden," *Fast Company*, July 1, 2009, http://www.fastcompany.com/magazine/137/why-america- is-addicted-to-olive-garden.html.

[3]Mark McClusky, "The Nike Experiment: How the Shoe Giant Unleashed the Power of Personal Metrics," *Wired*, June 22, 2009, http://www.wired.com/medtech/health/magazine/17-07/lbnp_ nike.

[4]Accenture, "Helping the Royal Shakespeare Company Achieve High Performance Through Audience Analytics, Segmentation and Targeted Marketing," 2008, http://www.accenture.com/NR/rdonlyres/891F5AA1-A1C2- 4828-81E8-BE4EBAD7948B/0/RSCcredentialFinal.pdf.

[5]Mary Hayes Weier, "Coke's Customer-Loyalty Web Site Scores Big with Consumers," *InformationWeek*, July 21, 2008.

Integration

Data integration, which is the aggregation of data from multiple sources inside and outside an organization, is critical for organizations that want to be more analytical. Transactional systems are often "stovepiped," addressing only a particular part of the business, such as order management, human resources, or customer relationship management. Enterprise resource planning (ERP) systems, which cover broad business functionality, are a notable exception. Thanks to these, organizations have come closer to solving a lot of the basic data integration challenges that bedeviled the early years of IT management. But even with an ERP system in place, you will undoubtedly need to consolidate and integrate data from a variety of systems if you want to do analytical work.

For example, you may want to do analysis to find out whether shipping delays affect what your customers buy from you—a problem that may well require integration across multiple systems. Or you may want to merge Web data on your customers with data from the order management module in your ERP system. Perhaps you want to combine data from your organization with market share or customer satisfaction data from an external supplier. Again, most organizations can't escape the need for data integration.

Stage 5 companies define and maintain key data elements such as customer, product, and supplier identifiers throughout the organization. Hence, they avoid complaints of, "Why can't I get a list of our top one hundred customers?" or "Why do I get different answers every time I ask how many employees we have?" It takes constant vigilance to have integrated, high-quality data. Citigroup's Institutional Bank (not necessarily a stage 5 firm at analytics overall, but very strong at customer data management) established a unique identifier for corporate customers in 1974, and has been refining it ever since. It maintains a group of data analysts in Manila to continually classify, tag, clean, and refine information about customers. Even though Citi's Institutional Bank—like most business-to-business organizations—has a relatively small number of customers, it's not easy to keep track of which organizations are parents and which are subsidiaries, or of name, location, and ownership changes. And it's a real nightmare if your organization has millions of consumers as customers.

In data integration, pundits often advocate "one version of the truth." You probably know the syndrome behind this recommendation. Several different groups come to a meeting to discuss something-or-other, each camp armed with facts to support its position. Sales are up for this reason, new hires are down for that reason, and so on. Trouble is, each group's data has different numbers for revenues, profits, total employees, daily moving average of cafeteria profits, you name it. The groups then spend more time arguing about whose data is correct and less time analyzing and acting on it.

The problem is indeed debilitating and worthy of attention. But you must narrow your focus: instead of attempting the Sisyphean task of cleaning up every data object in the company, select the master (or reference) data used in decision making and analysis. Employing certain processes and technologies to manage data objects (like customer, product, etc.) that are commonly used across the organization is called master data management, or MDM. MDM gets a bad reputation partly because it is rather unglamorous work. Just like eating your vegetables, managing your master data is good for you but not always satisfying or fun.

Furthermore, companies often make MDM a lot more complicated than it needs to be. According to Wikipedia (*Encyclopedia Britannica* and *Webster's* don't weigh in on this arcane subject—as of this writing, at least), MDM "has the objective of providing processes for collecting, aggregating, matching, consolidating, quality-assuring, persisting and distributing such data throughout an organization in such a way as to ensure consistency and control in the ongoing maintenance and application use of this information."

That gerund-rich definition suggests that MDM is a big hassle, and it is when MDM morphs into an unending data purification ordeal. Still, key data items do need to be defined in a common fashion and policed so they don't vary across the organization. So just be selective and start small. Focus on a pressing problem where cleaning some limited set of financial or customer data will bring a sizable payoff. Standardize data definitions and eliminate or correct incomplete, inaccurate, and inconsistent data. Then improve the lax data management and governance processes that caused the data to become dirty. When done, move on to other important chunks of data.

Ultimately you need to set a balance between integration efforts and analytical initiatives. If you embark on an MDM project, make sure you have plenty of money, time, and executive support. There's a pretty good chance that at some point during your MDM project somebody is going to say, "What the hell is master data management, and why's it taking so long and costing so much?" It would be good to have an answer at the ready.

In summary, stage 5 organizations have some data integration in place, but the perfect, flawlessly integrated, data-managed company is largely a fantasy. Even the best organizations don't have perfect data everywhere. They focus their data integration where it really makes a difference to their performance. The business need should drive the data integration efforts, not "just in case" mass integration or a misguided search for perfection.[6] Your targets for analytical work, as discussed in chapter 5, will steer you toward integration of the data that matters most.

[6]In the business intelligence (BI) world, this is known as the "Field of Dreams" approach. But if you build it, they will not come—and anyway, you will probably suffer in the effort.

Quality

Ironically, while data quality is important in analytical decision making, data doesn't have to be quite as perfect as it does in transactional systems or in basic business intelligence reporting applications. Skilled analysts can deal with missing data, and can even estimate substantial amounts of missing data or create statistical samples of data that get around the problem.

Nevertheless, flawed or misleading data is a problem for analytics. Having integrated data is only the first step. Keep in mind that most data is originally gathered for transactional purposes, not analytical ones. And every type of transactional data can have its own specific problems. Web transaction data, for example, can be plagued with problems that may inhibit your ability to extract meaningful analytics from your Web server logs. One Web analytics expert, Judah Phillips, identified eighteen glitches, ranging from spiders and bots that crawl your site and inflate your visit counts, to untagged pages that generate uncounted page views.[7]

With a particular decision and use of analytics in mind, analysts may need to trace data problems back to the source—often to the point where the data was originally entered—to find the root cause and to fix incorrect data. Even high-quality, integrated source systems like a modern ERP can't prevent front-line employees from entering data incorrectly. You may have to undertake some detective work to identify persistent sources of poor-quality data.

Stage 5 companies don't have perfectly clean data, but they have addressed many of their glaring data quality problems. They have data of sufficient quality for analytics in areas that really matter to their decision making. If they are focused on customer analytics, they have a high-quality customer database that has very few duplicated, inactive, or dead customers (the dead don't tend to respond well even to targeted promotions). Many of the customers' addresses may be outdated, but such errors may not matter for analytical purposes. Stage 5 companies also have a well-defined, relatively painless process for improving the quality of data as needed. Moreover, they have good processes up front to capture and validate data, so there isn't much cleanup to do.

In the late 1990s, Montana State Fund, a quasi-government agency that issues workers compensation insurance, dealt with data quality and integration by borrowing tactics from data piracy prevention. The resulting data contents were often challenged, so in 2006 it began a new initiative that certifies key reports and analyses with digital water-marks. The watermarks indicate that the data within the reports—over seventeen hundred elements extracted from core applications—are the official versions and have been audited. Users can print out the reports with the watermark, but can't download the data without losing the watermark. Al Parisian, the head of IT and strategy for Montana State Fund, reports that the culture of the organization is evolving to embrace the vastly improved data platform: "I've been in meetings where people argued based on selected events and partial data. Other people there have said, 'You can't use those isolated examples. Here's a report based on all the data from that period.'" Parisian knew that the approach was working when he saw that behavior.

[7]Blog post by Judah Phillips of Web Analytics Demystified, July 19, 2008, http://judah.webanalyticsdemystified.com/2007/08/web- analytics-data-quality.html.

Access

Data must be accessible in order to be analyzed—that is, it must be separated from the transaction-oriented applications (like sales order management or general ledger) in which it was created, and located where analysts can actually find and manipulate it. Stage 5 companies provide access to data by creating a data warehouse. Many companies have proliferated warehouses and single-purpose "data marts," but since integration is critical for advanced analytics, stage 5 companies will most likely have an enterprise data warehouse (EDW)—one that cuts across multiple functions and business units—for key analytical applications to draw from.

An EDW contains all the information that you might want to analyze—both current and historical values. If you think this is vague as an "information requirements" definition, you're exactly right. Because of an EDW's all-inclusive nature, firms always have to add new data elements to their warehouses, such as external data from Nielsen or other third-party providers. Thus the warehouses usually end up being so big that they overwhelm casual users. Since the original idea of a data warehouse is to make data more accessible to nontechnical users, something of a contradiction is built into the notion of an EDW. Still, many organizations have them, and they are a more feasible avenue for analytics than working directly with transaction data.

Unlike EDWs, data marts are departmental versions of data warehouses and sometimes are created independently of IT. [8] Although department-based data marts can limit analytics by undercutting integration, they can still play a role by solving some of the size problems of EDWs. For example, if you are pretty sure that most of your financial analytics will be restricted to data in the finance data mart, you should be fine relying on that source alone.

While you're thinking about access, you may want to (or, more appropriately, have to) think about speed. If you're going to be doing a lot of analytics you may need a special "data warehouse appliance." This dedicated system of software and hardware is optimized to do rapid queries and analytics. If you need your answers to analytical questions fast, you'll probably need such an appliance.

Some organizations have concluded that they don't need to—or can't afford to—make all their data available for analysis. If you have lots of different transaction systems and data sources on customers, for example, creating an entire customer information file would be very difficult. What these organizations have done might be called the "10 percent solution." They've taken a sample—most commonly 10 percent—of data for a particular domain—usually customers—and made that accessible for analysis. Such samples of large data populations can be very satisfactory for some analyses. A large bank did this as an early step toward building a customer data warehouse, and was able to make significant progress on analytical issues such as segmentation and promotion targeting. The data sampling strategy may also be useful as a

pilot of an enterprise approach when business units are very independent and it's not clear whether managing data at the enterprise level will be feasible or effective.

[8]From an enterprise IT perspective, a data mart and an EDW serve different purposes. An EDW is truly a warehouse—a storage area where you can get bulk deliveries of very detailed data to fill the "store shelves." A mart is more like a grocery store, where items are already sitting on store shelves, packaged and ready for consumption.

Privacy

Highly analytical organizations tend to gather a lot of information about the entities they care about most—usually customers, but sometimes employees or business partners. Then they guard it with their lives. Stage 5 firms follow the Hippocratic oath of information privacy: above all, they do no harm. They have well-defined privacy policies about customer and employee information. They don't break the privacy laws of the territories or industries in which they operate (and that's not easy for a global company, because policies vary widely; Europe's laws are particularly strict). They don't lose information because of hackers or careless mistakes. They don't sell or give away information without the permission of the customer or employee. They get the information in the first place through "opt in" policies—that is, customers or employees give explicit permission for the information to be captured and used. They have clear restrictions about the frequency of customer contacts—they don't ever want to be viewed as pests by their customers. And if there is any doubt about when a particular analytical activity might cross the line of propriety, they don't cross it.

It's not just about having effective privacy policies, however. Stage 5 firms work with customer contact employees to ensure that they don't reveal sensitive information to or about customers. For example, one of Tesco's Clubcard loyalty card customers called the retailer's call center to complain about having received a coupon for condoms. Tesco often sends customers coupons based on what they have bought in the past, a fact that this customer seemed to know. "Does this mean," she asked, "that someone has bought condoms using my card?" The call center rep protected private information by coolly replying, "Sometimes we just send out coupons randomly," even though it was apparent from database records that condoms had been purchased using the customer's Clubcard number. In doing so, the service rep may also have saved a marriage, a feat to which every analytical practitioner should aspire.

Governance

Our discussions thus far may have suggested that data gets where it needs to be through supernatural forces. Alas, it is we imperfect humans who must manage data. And the term *governance* suggests that some humans are more important than others in managing it. For us, governance means all the ways that people ensure that data is useful for analysis: it is consistently defined, of sufficient quality, standardized, integrated, and accessible. Of course, one could argue that ensuring that an organization has good data is everyone's job, but that will guarantee that it's nobody's job. There are certain roles that need to be played if an organization wants to have stage 5 analytics. We'll describe the most important: executive decision makers, owners/stewards, and analytical data advocates.

Executive Decision Makers. Getting the organization aligned regarding the key data to be used in analytical projects is a job for senior managers. At a minimum, they must decide which information needs to be defined and managed in common across how much of the business. For example, if customers are to be a main analytical focus (probably the most common data domain for analytical work), senior executives must agree on a common term and meaning for "customer" throughout the organization. Most executive teams don't discuss information in this fashion, but their organization will not be able to integrate its data successfully without such deliberations.

High-level decisions about data also can only be made by senior management. Even if they are not comfortable with issues like ownership, stewardship, and relationship to strategy, they are the only ones who can deliberate about such matters. Since all data can't be perfect, only senior leaders can decide what kind of data—customer, product, supplier, zodiac sign, and so forth—is most critical to the organization's success. They will have to discuss what kinds of data assets correspond to particular strategic and analytical targets. Finally, they have to sign the checks for investments in data, so they will ultimately have to decide on major data-related programs and initiatives. If you're an IT or analytics person responsible for managing data, you need to engage your senior executives, or your lack of a relationship with upper management will come back to bite you.

Owners/Stewards. Many organizations will need to define specialized responsibilities for particular types of data—customer data, financial data, product data, and so on. Ownership is a highly loaded term that is likely to cause political difficulty and resentment; stewardship is a better term that avoids raising hackles. Stewardship entails taking responsibility for all the factors that make data useful to the business. This would typically be the job—perhaps full time, but more frequently part time—of business managers rather than IT people.

BMO (Bank of Montreal) Financial Group has adopted information stewardship to a large degree. BMO executives feel that the bank owns all of its information, but that they "need Business Information Stewards to ensure that it is managed appropriately across processes and functions."[9] BMO gives its stewards the following responsibilities:

- *Business definitions and standards.* Consistent interpretation of information and ability to integrate.

- *Information quality.* Accuracy, consistency, timeliness, validity, and completeness of information.

- *Information protection.* Appropriate controls to address security and privacy requirements.

- *Information life cycle.* Treatment of information from creation or collection through to retention or disposal.

BMO specifies particular stewardship functions at the strategic level (e.g., "Develop an information strategy and high-level 3–5 year plan"), the operational level (e.g., "Develop an information management change management strategy and program"), and the tactical level ("Develop, deliver, and maintain information operating procedures to support the information management corporate standards"). Information stewards at

BMO are business executives, and the steward-ship role is typically part time.

Analytical Data Advocates. While IT organizations are usually skilled at building data infrastructures and installing and maintaining applications that generate transaction data, they are not often oriented to helping the organization use data in reporting and analytical processes. One way to ensure that focus is to create a group that emphasizes information management and ensures that data and information can easily be accessed and analyzed. Such groups are becoming increasingly common; some call them business intelligence competency centers (BICCs).[10]

Other organizations with slightly broader objectives than facilitating business intelligence refer to their "analytical data advocate" group as information management (IM) or business information management. Two organizations that have established such groups are the health insurer Humana and the South African banking unit of Barclays, Absa Bank.

Humana's group is responsible for information management and "informatics," the term health care organizations use for analytics for patient care and disease management. One of the group's first priorities was to develop a strategy, enlisting the support of senior executives throughout the organization. Lisa Tourville, the head of the group, reports to the firm's chief financial officer, although her group addresses all sorts of information—not just financial. Lisa has a strong actuarial background, and describes her personal vision as "to be an advocate of all matters quantitative and relentlessly search to improve analytic capabilities in support of corporate decision-making efforts."[11] That's a great set of goals for the head of such a group.

Absa Bank established its IM group in 2001, and was initially focused on customer information. David Donkin, the first head of the IM group, explained the group's mission: allow information- and knowledge-based strategy formulation and decision making, and leverage information to improve business performance. These are at the heart of what it takes to make an organization more analytical.

The IM group at Absa is responsible for the data warehouses, BI tools and applications, data mining, and geographic information systems. IM also develops the bank's information strategy and architecture, which defines how the bank stores and manipulates information. Donkin has represented Absa at broader gatherings of Barclays' analytical community. The corporate IT organization manages Absa's operational applications, databases, and the IT and network architecture.

According to Donkin, when the IM group was formed, Absa's data warehouse was "not customer centric, not operationally stable, and not business directed." It stored information that no one really needed, and that few knew how to find. Today the IM group improves the relationship between IT at the back end and business decision makers on the front end. It facilitates such analytical applications as scorecards, fraud detection, risk management, and customer analytics, which drive cross-sell, up-sell, retention, customer segmentation, and lifetime value scores.[12]

Even if you have a BICC or an IM group, you can never do enough connecting between IT and rest of the business. Whether the group is separate from or part of the IT organization, you must have some data people in IT who are familiar with the typical types of analysis done in your industry and company. If they know that, they can help to ensure that the data is structured for easy access and analysis.

[9]"Getting a Handle on Our Information! Information Stewardship at BMO FG" (document furnished to author), May 12, 2006, and discussion with BMO executives, April 19, 2007. This document suggests that BICCs are responsible for two key data functions: acquisition and steward-ship. Acquisition includes activities such as data integration, data storage, testing, and maintenance. Stewardship includes the responsibility for data standards, quality, and governance.

[10]Gloria J. Miller, Dagmar Brautigam, and Stefanie V. Gerlach, *Business Intelligence Competency Centers: A Team Approach to Maximizing Competitive Advantage* (New York: John Wiley & Sons, 2006), 38.

[11]Interview with Lisa Tourville, March 6, 2008.

[12]Interview with David Donkin, October 28, 2006; Wayne Eckerson, "New Ways to Organize the BI Team," Data Warehousing Institute *Business Intelligence Journal*, March 28, 2006, http://www.tdwi.org/Publications/BIJournal/display.aspx? ID=7896.

Data Before Analytics

Getting data in order is so critical to analytics that most organizations have to undertake substantial data management efforts before they can do a lot of analysis. For example, at Albert Heijn, the largest supermarket chain in the Netherlands with over eight hundred stores, considerable data efforts were undertaken to create the ability to do analytics.

In the 1990s, Albert Heijn embarked on a program of differentiating stores along several dimensions such as assortments, replenishments, targeted customer segments, and so forth. To accomplish this objective while maintaining cost parity, Albert Heijn's managers concluded that it needed a more integrated data environment. They developed a blueprint of the company's envisioned information environment, covering and integrating data from the total value chain, resulting in an enterprise data warehouse. Previously Albert Heijn had a great deal of data, but it was spread across a number of different systems and databases. The goal was one integrated environment for all enterprise processes and transactions, using the most granular data possible and supporting the entire company. Albert Heijn embarked upon a multiyear data integration project that ultimately cost €30 million.

The resulting database was called PALLAS, after the Greek goddess of knowledge (a classical name adds class to data integration projects). It eventually drew from 75 percent of the company's transaction systems and contained ten years of online detailed data. Over three thousand employees now run reports from the data or perform analysis, and each week over sixty thousand customers use the system to see what items they recently bought. Management questions on store operations can be answered nearly in real time. Forecasts of store demand for

particular items, for example, are updated every five minutes, and stores are automatically replenished based on the forecasts.

Once PALLAS was created and the initial demand for reporting satisfied, Albert Heijn began to turn its attention to analytics. It formed a business analytics group to perform analytical projects across the organization and professionalize the area of analytics within the company. It is also using sophisticated artificial intelligence technologies rarely seen in retail. Single-purpose data marts created over the years support analysis in particular domains. The first one involved replenishment, with the goal of reducing stockouts and shrinkage of inventory. Now a variety of analytical projects are under way using PALLAS data, including projects on loyalty, assortment optimization, promotion analysis, and introduction of nonfood items to Albert Heijn stores. These projects wouldn't have been possible without the availability of integrated, high-quality data.

Data Through the Stages

We've described a number of attributes of the most sophisticated analytical competitors. What if you don't aspire to that level of analytical orientation? How can the opera houses, circus colleges, and other businesses of the world make analytics work for them? The remainder of the chapter addresses what organizations do with data at lower levels of analytical focus, and how they can move to the next step. Table 2-1 provides an overview of these transitions.

Table 2-1: Moving to the next stage: Data

From stage 1 *Analytically Impaired* to stage 2 *Localized Analytics*	From stage 2 *Localized Analytics* to stage 3 *Analytical Aspirations*	From stage 3 *Analytical Aspirations* to stage 4 *Analytical Companies*	From stage 4 *Analytical Companies* to stage 5 *Analytical Competitors*
Gain mastery over local data of importance, including building functional data marts.	Build enterprise consensus around some analytical targets and their data needs. Build some domain data warehouses (e.g., customer) and corresponding analytical expertise. Motivate and reward cross-functional data contributions and management.	Build enterprise data warehouses and integrate external data. Engage senior executives in EDW plans and management. Monitor emerging data sources.	Educate and engage senior executives in competitive potential of analytical data. Exploit unique data. Establish strong data governance, especially stewardship. Form a BICC if you don't have one yet.

From Stage 1 to Stage 2. The story from stage 1 to stage 2 is about basic data mastery, usually the prerequisite missing in stage 1 companies. If it's available at all, data is inconsistent and of poor quality. Therefore, to move to stage 2, particular functions need to create the necessary data from capable transaction systems, and make it available for analysis purposes. There are no enterprise data warehouses at this stage, but we may see the beginning of functional data marts or operational data stores.

In moving from stage 1 to stage 2, all analytical activities, including those involving data, tend to start and remain at the local level. Functional or business units marshal the necessary data and analysts to undertake analytical initiatives in their areas. There are many different targets throughout the organization, so there is little or no ability for the enterprise to focus on data hygiene and accessibility projects.

From Stage 2 to Stage 3. Now senior executives show signs of interest in analytics, encouraging analytical and data-oriented people from around the organization to start communicating and collaborating. The key here is to create some successes with data—identifying new sources, extracting some from transaction databases, buying it externally, and using it for analytical purposes. There is also the beginning of organizational consensus on a key analytical target, and some recognition that data needs to be integrated and shared.

An enterprise-level target allows the organization to begin focusing data initiatives in subject areas that support future analytical strategies. For example, if a company believes its primary future lies in customer analytics, building a customer data warehouse is the first order of business. Similarly, creating human expertise around customer data and analytics should be a priority. If the long-term vision involves something else—product data or claims data or genomic data or what have you—then the initial focus should be in that area.

In transitioning from stage 2 to stage 3, data and other resources begin to be viewed as organizational, rather than departmental, resources. Enterprise-level data strategies begin to appear. Functional or business unit managers who have built up data capabilities may be either threatened or, ideally, flattered by the adoption of their data and analysis approaches by the organization at large. Therefore, solid—not necessarily passionate—leadership at the enterprise level is necessary to get to this stage. Local owners of data need to be rewarded for giving it up, or have their wrists slapped for holding on to it.

From Stage 3 to Stage 4. In stage 3 the organization has a long-term vision of where it wants to go with analytics, but for some reason feels the goal is not achievable in the short run. The key in moving to stage 4, then, is to facilitate the organization's efforts to create tangible analytical projects at the enterprise level. From a data standpoint, the focus must be on building cross-functional data capabilities. This means that organizations need to replace function-level data marts with an enterprise-level warehouse. The data in the warehouse will come mostly from internal transactional systems, but increasingly from integrating internal and external data. To justify and pay for these activities, senior managers will have to be engaged and consulted on data issues.

We've found several organizations whose long-term vision involves using data that simply isn't practical to gather today. For example, health care institutions and pharmaceutical firms see a future of "personalized medicine" in which drugs are based on a patient's genomic and proteomic profile. Today, that information is both expensive and difficult to manage. So it's crucial to watch the development of the needed data to see when it will be available, and perhaps to institute a pilot project to analyze the data available now—so that you will know what to do with it when it finally arrives.

From Stage 4 to Stage 5. How do stage 5 companies differ from stage 4? Stage 4 organizations are generally competent at managing data and have most of the required data resources in place. They have adequate transaction systems, data warehouses, and perhaps even some nonnumerical content. However, because these firms are not yet passionate about using the data for competitive purposes, they have not optimized their data environments for analysis. They probably do not have a strong focus on unique data for their industry. They may not have

governance functions in place to mediate between business-side analysts and decision makers, and people in the IT organization who are data infrastructure providers for analytical projects.

Since the primary difference between stage 4 and stage 5 organizations is a passion for analytics, the key step in that transition is to excite senior executives about analytical possibilities. We're not suggesting the implementation of a weekly Algorithm Appreciation Day, but educational activities for an organization's leaders can emphasize the role of well-managed and differentiated data in an analytically focused strategy—with as many examples as possible of competitors and other firms that have used data for competitive advantage. A simple assessment of key data resources around the organization may help to stimulate thinking and action.

In terms of governance, it may be useful to pilot some stewardship or business-side information management functions, or to select a friendly executive as an initial steward. It may also be advisable to create a small business intelligence competency center.

Skipping Steps or Accelerating Progress

In the data space, it would certainly be possible to skip a step or two, or move quickly through a stage. If you're now in stage 1 and want to become a more analytical enterprise, it would definitely be a good idea to skip the silo-based approach of stage 2. If your executive team will support a cross-functional, enterprise-oriented approach, it makes good sense to begin building an enterprise data warehouse, for example. However, since you'll need some sense of where to focus your efforts, you need the targets and vision that typically come only with engaged executives at stage 3.

It probably wouldn't work to skip from stage 1 or 2 to stage 4 or 5. Putting the data infrastructure in place—along with the human infrastructure relative to data, analysis, and analytical strategy—takes time. Further, many of the investments necessary to move along the maturity curve for analytics may not be supported by executives until they see evidence of value from early projects. However, if you have a very supportive CEO and executive team, you can make progress quickly.

Keep in Mind . . .

- Have someone in your data management organization who understands analytics and how to create them, and who can educate others about the differences between data for analytics and data for business transactions or reporting.

- If you can't obtain all the data for a particular subject area (e.g., customers), then create a statistically valid sample with a fraction of the data. In cases when you just need some directional insight, creating a sample is a lot faster and cheaper. In most cases, though, try to obtain as much detailed data as possible.

- Create an organization whose job it is to ensure that business information is well defined, well maintained, and well used.

- Identify data stewards, generally from outside IT, for key information content domains (customer, product, employee, etc.).

- Find some data that is unique and proprietary that your organization can exploit analytically.

- Experiment with nonnumerical data—video, social media, voice, text, scent, and so on. (Okay, maybe not scent, unless you're running a company that deals in personal hygiene.)

- Don't spend all your time and resources on achieving perfection with data completeness, quality, or integration—save time for analysis!