2

Information Creation through Analytics

Stephan Kudyba

CONTENTS

The primary initiative in leveraging the value of data resources lies in the realm of analytics. This term, however, encompasses a wide variety of methodologies that can provide descriptive, comparative, and predictive information for the end user. This chapter will provide a brief background and description of some noteworthy analytic approaches as applied to more historical, structured data and include references to big data issues

along the way. The area of big data and analytics will be addressed in greater detail in the real time and continuous analysis section at the end of this chapter and in Chapter 3.

Analytic methods can range from simple reports, tables, and graphics to more statistically based endeavors to quantitative-based methods. We provided some analytic approaches according to some commonly referred to categories below. Regardless of the techniques deployed, the end result of an analytic endeavor is to extract/generate information to provide a resource to enhance the decision-making process.

1. Spreadsheet applications (also facilitated by vendor software packages)
   a. Data/variable calculations, sorting, formatting, organizing
   b. Distribution analysis and statistics (max, min, average, median, percentages, etc.)
   c. Correlation calculation between variables
   d. Linear and goal programming (optimization)
   e. Pivot tables (an intro to online analytic processing (OLAP) and business intelligence)
2. Business intelligence
   a. Query and report creating
   b. Online analytic processing
   c. Dashboards
3. Multivariate analysis (also part of business intelligence)
   a. Regression (hypothesis approach)
   b. Data mining applications (data-driven information creation)
      – Neural networks
      – Clustering
      – Segmentation classification
      – Real-time mining
4. Analysis of unstructured data
   a. Text mining
5. Six Sigma
6. Visualization

The type of analytic approach is generally dictated by the objective of what the user of the analysis requires, and where the objective and overall initiative needs to be clearly defined to achieve the most effective and informative results. This problem definition process generally involves the selection of a performance metric and identification of variables that

impact that metric. Once the scope of the analytic endeavor (problem definition) has been established, then corresponding data resources must be managed (variables selected at a particular level of detail) and analysis can begin. The steps to conducting a problem definition for analytics will be addressed in detail in Chapter 5. The remainder of this chapter will provide an overview of some of the analytic methods mentioned above.

INTRODUCTION TO THE CONCEPT OF ANALYTICS

One of the initial stages of any analytic endeavor is the incorporation of an investigative study of a data resource. In other words, before a report is generated or quantitative modeling is conducted, an analyst needs to better understand what's in a data file. This investigative process involves conducting a distribution analysis of various data variables, perhaps calculating maximum, minimum, and variance metrics such as standard deviations. This provides a descriptive character of what the data variables are comprised of and renders additional analysis more robust, as it identifies the presence of such issues as data bias or skew, outliers, and even errors in data resources.

BUSINESS INTELLIGENCE

Reports

The focus of this book involves the utilization of business intelligence applications (e.g., OLAP, dashboards, mining) to extract actionable information from all types of data to enhance the decision-making process. One of the most basic levels of this approach is the creation of business reports that incorporate sequel-related queries of data resources to extract variables that describe a business scenario. The introduction of big data involves additional requirements to this process; namely, when devising the parameters of the report to be created, the decision maker now must consider new variables that impact that conceptual report. The volume of data that must be processed must also be considered, and finally, the currency of the report (e.g., how often a report must be updated to provide

adequate information for the decision maker). However, as simple as the process of generating a report may be, creating one that provides essential information to those that receive it may be a quite complex task.

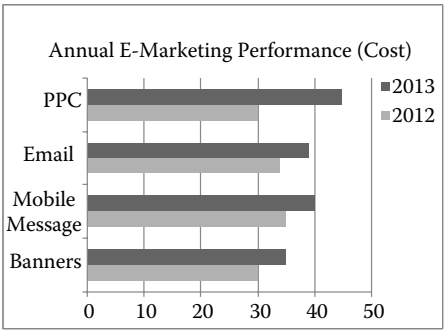
Consider a request by an Internet marketing department to produce an analytic report that depicts the performance of various Internet marketing tactics that drive traffic to a company's landing page. Although this initiative appears to be straightforward and simplistic in nature, one must consider all the variables that comprise the area to be analyzed, along with the needs of the user of the report.

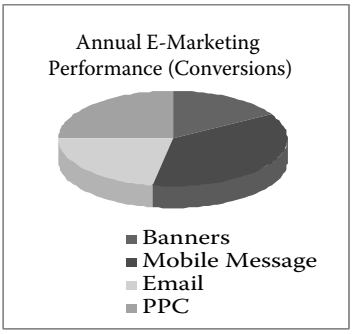
Some dimensions and variables that could be included in this analysis would involve:

Time	**Performance Metric**	**Marketing Source**	**Source Traffic Location**
Hour	Clicks	Paid, Generic search	Town
Day	Bounce	Mobile device	County
Month	Conversions	Banners	State
	Cost	Referral site	

Platforms such as Google Analytics provide robust functionality to accomplish extensive report generation in the e-commerce spectrum. When conducting customized analytics (tailored analytics to a specific company's activities) data experts and analysts must apply due diligence to acquire that information that provides a strategic advantage in the marketplace. This involves the storage, processing, management, and ultimate analysis of data resources that describe a particular process.

Well-designed reports that incorporate all the pertinent and available variables that describe a business activity can be an important source of information to decision makers (see Figure 2.1). However, the limitation of information creation at the report level is that the user often scans a report, assimilates the information, and quickly thinks of alternative business scenarios that are essential to providing more robust information regarding a process or activity. The report is limited to its current level of data aggregation and variables depicted. The next step to analysis or business intelligence involves the application of OLAP, which gives users the flexibility to view and analyze multiple scenarios of a business process. Before we describe the application of OLAP functionality that leverages large data resources and addresses currency of data, consider the more simplistic spreadsheet application of Pivot Tables.





Tactic	2012 (cost)	2013 (cost)	Change	Tactic	2013 Conv
PPC	30,000	44,000	46%	PPC	22%
Email	34,000	39,000	15%	Email	20%
Mobile	35,000	40,000	14%	Mobile	42%
Banner	30,000	34,000	13%	Banner	16%

FIGURE 2.1
Annual report of e-marketing cost and conversions.

Pivot Tables

A simplistic version of OLAP that many users can quickly relate to includes the use of pivot tables in a spreadsheet environment. Pivot tables leverage data in a flat, spreadsheet file to present alternative scenarios that describe a business activity. Through basic spreadsheet functionality, users can quickly generate a table view of relevant variables at a particular level of aggregation. For example, a spreadsheet of data that describes a software company's sales activities can include numerous rows according to corresponding variables. Hypothetical data recording national sales activities of branches across the country is illustrated in Table 2.1.

With a simple pivot function, Table 2.2 could be calculated with ease.

Dynamic Reporting through OLAP

Pivot tables are similar to OLAP in that they provide a multidimensional view of an activity. Enterprise OLAP provides greater scale to the analytic process, as it provides the platform to address multiple levels of aggregation of data resources, can depict updated views as source data is updated, and can process extremely large volumes of data. With this flexibility OLAP can help decision makers investigate information addressing multiple descriptive scenarios regarding an operation's activity, therefore

TABLE 2.1

Hypothetical Data Recording National Sales Activities

Salesperson	Product Category	City/Area	Customer Industry	Units	Sales
KDE	ETL	NY	Finance	90	$45,000
SEF	Reporting	NY	Insurance	80	$24,000
CHT	Analytics	Boston	Finance	10	$20,000
HHT	Database	Phili	Retail	55	$41,250
GGN	Database	Atlanta	Manufact	65	$48,750
THT	ETL	DC	Retail	18	$9,000
TTW	ETL	Phili	Retail	42	$21,000
AHY	Analytics	Chicago	Retail	30	$60,000
FDO	Reporting	San Fran	Manufact	39	$11,700
JJT	Reporting	Chicago	Finance	42	$12,600
GHI	ETL	NY	Transport	32	$16,000
BDE	Analytics	DC	Transport	71	$142,000
PEC	Reporting	NY	Finance	26	$57,045
LYJ	Database	Chicago	Insurance	52	$39,000
KIP	Analytics	San Fran	Insurance	75	$150,000
OBN	Database	NY	Retail	53	$39,750
ERB	Database	San Fran	Manufact	93	$69,750
SEN	Reporting	LA	Retail	17	$5,100
JJR	ETL	NY	Retail	96	$48,000
WNS	ETL	Phili	Manufact	32	$16,000
DHK	Reporting	Boston	Finance	26	$7,800
TRN	Reporting	Boston	Transport	30	$9,000
RGH	Database	Phili	Retail	54	$40,500
MMR	Database	Atlanta	Retail	46	$34,500
SJP	ETL	Atlanta	GPU	80	$40,000

enhancing the knowledge generation process and overall ability to generate effective strategic conclusions. The diversity of information views involves various dimensions of time, performance metrics, and descriptive variables.

General Cube Inputs

Time	Descriptive Variables	Performance Metrics
Daily	Demographics	Sales
Weekly	Behavioral	Response rate
Monthly	Strategic	Operational
Quarterly	Process related	Units

TABLE 2.2

Sales by Product Category by City

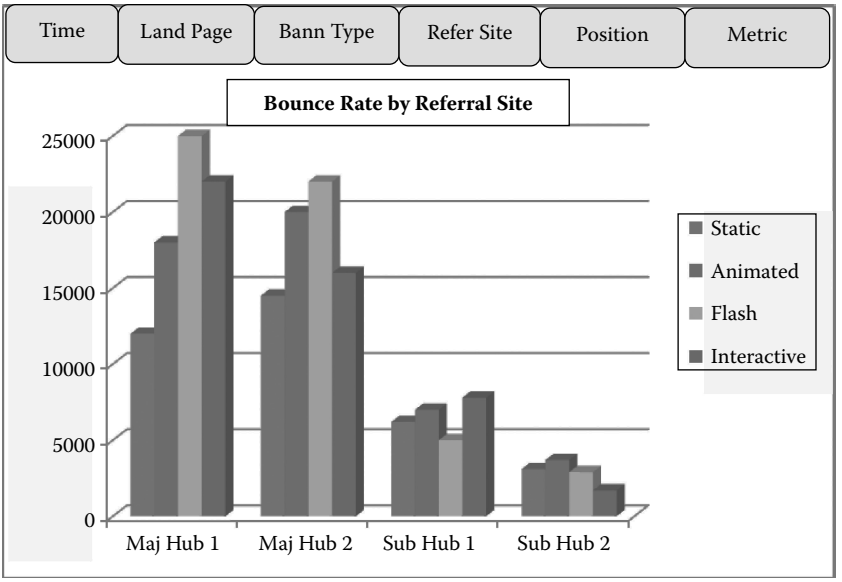
ETL (Extract Transfer and Load)	
New York	$61,000
DC	$9,000
Philadelphia	$37,000
Atlanta	$40,000
Total	$195,000
Reporting	
New York	$81,045
San Francisco	$11,700
Chicago	$12,600
Boston	$16,800
Los Angeles	$5,100
Total	$127,245

These inputs must be organized to provide information (variables at levels of detail) that describes a business scenario in order to facilitate decision support for the end user. Consider the graphical view of a cube in Figure 2.2.

Figure 2.2 depicts an illustration of an OLAP cube that facilitates analytics of banner Internet marketing tactics. The cube presents a multidimensional view of the variables that describe the activities involved in banner advertising. The platform gives the analyst the ability to query data variables from different levels of detail and in different combinations, through both numeric data and visualization. The tabs at the top of the graphic depict the variables that are available to be analyzed. The scenario depicted illustrates the bounce rate (number of bounces) according to different types of referral sites where the various banner styles (static, animated, flash, and interactive) are displayed.

Users have the ability to change variable views from different perspectives, including:

Time (hourly, daily, quarterly)
Landing page (social media, home, custom landing design)
Banner type (static, animated, etc.)
Referral site (main hub, MSN, Yahoo; subhub, complementary site)
Position (banner position, top, middle, bottom)



	Maj Hub 1	Maj Hub 2	Sub Hub 1	Sub Hub 2	Total
Static	12000	14500	6200	3100	35800
Animated	18000	20000	7000	3700	48700
Flash	25000	22000	5000	2900	54900
Interactive	22000	16000	7800	1700	47500
Total	77000	72500	2000	11400	186900

FIGURE 2.2
(See color insert.) Banner performance multidimensional cube.

These perspectives can be analyzed according to predefined metrics, including bounce, views, click-throughs, and conversions. By navigating the different dimensions of the cube, the analyst can quickly identify strengths and weaknesses in different banner advertising initiatives. OLAP enhances the decision makers' ability to more fully understand the attributes that comprise the activities of banner advertising.

So what about big data, you say? Remember, big data entails not only volume of data but also the new variables (sources of data). Both these factors are considered when conducting analytics. In other words, a conceptual model must be generated that best describes the attributes of a desired process (entity to be better understood), and then data corresponding to those variables must be applied to that analytic framework. Big data adds complexity to the generation of the conceptual model as it introduces new

descriptive variables that may not have been available or incorporated in the traditional structure of the particular process. The value of big data follows the basic concepts just mentioned; however, it can provide even greater value to the user by providing more robust models that provide greater descriptions and understanding of what affects process performance. In the banner ad scenario above, perhaps the new variable that must be added to provide more insightful information to decision makers regarding the effectiveness of their e-commerce advertising is the source of where traffic is coming from regarding the technological platform. In other words, is traffic coming from mobile devices, laptops, or tablets? When considering big volumes and velocities of data in an OLAP environment, methods such as parallel processing and map reduction of data resources must be considered. This topic will be addressed in greater detail in Chapter 3.

OLAP provides a robust source of business intelligence to decision makers, as it can leverage data resources including big data volumes and provides a platform that offers a flexible, accurate, and user-friendly mechanism to quickly understand what has happened and what is happening to a business process. The multidimensional framework will give users the power to view multiple scenarios of a given process, such as the following:

- What is the bounce rate if I utilize a specific type of landing page?
- Where are my highest conversion rates coming from?
- Is there seasonality according to day of the week or month of the year for my traffic?

The key to a valuable OLAP cube involves the combination of a few factors. One of these relates to the concept mentioned earlier, namely, that a cube must effectively describe a business scenario. The conceptual model that is used to build the cube must include noteworthy variables (relevant) with an appropriate detailed format that give users true business intelligence. The next major factor is filling the cube with accurate, current, and consistent data. Deficiencies in either of these areas can quickly render the analytic method useless for decision making.

Analytics at a Glance through Dashboards

In today's ultra-fast, ultra-competitive information-based economy, it seems that the more senior a manager you may be, the less time that is available for investigation and drilling around multidimensional cubes.

Often the level of analytics is filtered down to a few insightful reports, ongoing insights absorbed in the marketplace, and the access to real-time dashboards that display key performance indicators relevant to a particular process. These dashboards are designed to provide decision makers with a feedback mechanism as to how an organization is performing. The key elements of dashboards are the delineation of relevant key performance indicators (KPIs) to a particular process, timeliness of their readings (currency of information), and finally, a user-friendly visual that provides the decision maker with a clear way of determining whether a process is operating successfully or not. The more traditional visual platform resembles that of an odometer in an automobile, where color schemes of performance reflect that of traffic lights (e.g., green, all is well; yellow, caution; and red, something is wrong and needs to be investigated). However, dashboard technology is quickly evolving where styles can include combinations of a variety of visuals (bar, line, pie charts) according to designated scales and are being utilized by decision makers at all levels in an organization.

The key to the effectiveness of a dashboard design involves its connection to the process at hand and use for decision making. Displays must be simple to understand and interpret. Just as a simple graphic display must adhere to design conventions (e.g., coherent color scheme, axis labeling, scale), so too must dashboard design, which adds complexity to the process as it combines various visual elements. The true key to a successful dashboard is evident by its effectiveness in providing timely, easy-to-understand decision support of a corresponding process. Dashboards that are too busy (include too many visuals), that are difficult to interpret, can quickly become omitted from an analyst's arsenal of decision support information.

Consider the dashboard example in Figure 2.3. The various graphic displays are clearly delineated from one another (separate sections) and are clearly labeled. Also, the design includes different visual displays, so the information presentation does not appear to overlap or include a blended view. Finally, complementary but distinctly different key performance indicators give the decision maker a well-rounded view of a human capital management application in this case.

Robust BI and Drill-Down behind Dashboard Views

Dashboards provide an instantaneous mechanism to analyze the performance status of a process. Organizations with extensive analytic

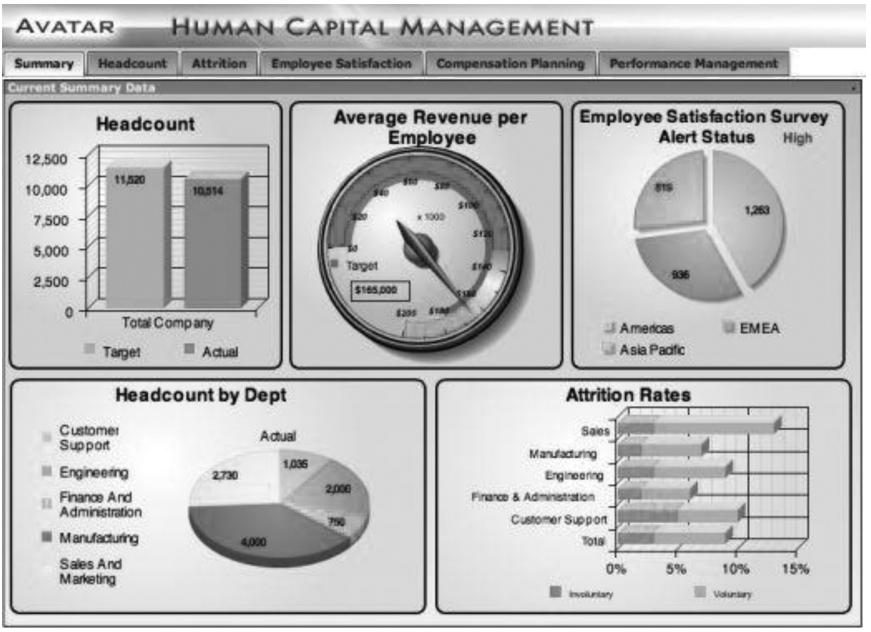


FIGURE 2.3
(See color insert.) Clearly designed employee analytic dashboard. (From http://www.dashboards-for-business.com/dashboards-templates/business-intelligence/business-intelligence-executive-dashboard; Domo, Inc., http://www.domo.com.)

capabilities through business intelligence applications can have OLAP cubes that can be quickly drilled into from a dashboard KPI that provides descriptive analytics of underlying variables that underpin the KPI. A prime example of an e-commerce-based KPI is the bounce rate on a landing page for an organization, especially when a new marketing initiative has been launched. Perhaps an organization has initiated an Internet marketing campaign with banners listed on various complementary referral sites. A red signal indicating a higher than acceptable bounce rate would provide decision makers with a timely analytic alert mechanism to investigate the source of the problem. A real-time cube or report could quickly depict which referral site may be the greatest source of misdirected traffic.

Not all dashboard displays need to be real time, where a simple refresh of data on an interim basis provides decision makers with an accurate indication of whether a process's performance is adequate. However, the big data era involving high velocity of streaming data resources often requires a real-time dashboard visual of a given process to provide users with a quick view of variable impacts on KPIs.

DATA MINING AND THE VALUE OF DATA

As we've illustrated in the business intelligence section (e.g., reporting, OLAP, dashboards), a primary approach to generating value from data resources is to manage it into useful information assets (e.g., building conceptual models and viewing data according to level of details according to variables that describe a process). The next step in the valuation process is to generate a higher level of knowledge through the information created from data. Data mining involves the application of quantitative methods (equations and algorithms), along with forms of statistical testing that process data resources, which can identify reliable patterns, trends, and associations among variables that describe a particular process. Techniques such as segmentation classification, neural networks, logistic regression, and clustering, to name a few, incorporate the use of algorithms and code or mathematical equations to extract actionable information from data resources. Chapter 4 provides detailed descriptions and applications of major mining methods.

Why Things Are Happening

Data mining can provide decision makers with two major sources of valuable information. The first refers to descriptive information, or the identification of why things may be occurring in a business process. This is done through the identification of recurring patterns between variables. Cross-sectional graphic displays can add significant information to decision makers to illustrate patterns between variables. Figure 2.4 provides a simple graphical view that illustrates an ad spend vs. dollar revenue elasticity curve as identified in the mining process. The figure depicts that a recurring pattern exists between the two variables, and that a direct relationship is prominent, where an increase in ad spend yields an increase in product revenue. Many non-mining-centric analysts would quickly raise the point that this information is not noteworthy, given the natural relationship between the two variables (e.g., the more spent on advertising, the more sales that are generated); however, this criticism is quickly dispelled when posing the question: If ad spend is increased by 5% from $200,000, what is the expected increase in revenue? That question is difficult to answer without the use of mining.

Mining methods can yield insightful patterns as to demographic and behavioral attributes of consumer response to marketing initiatives, the

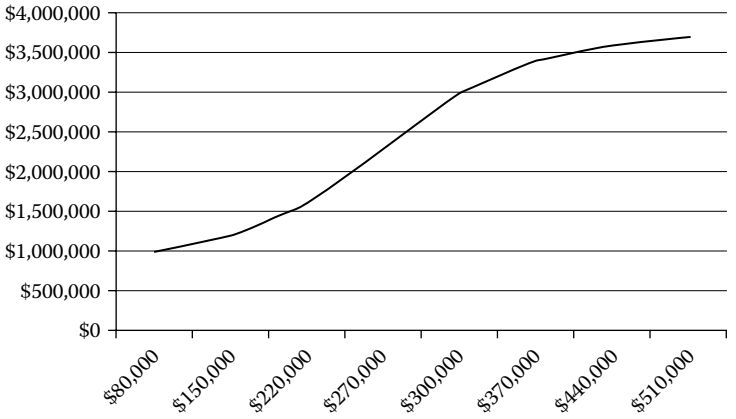


FIGURE 2.4
Advertising spend vs. revenue curve.

impacts of process components on performance metrics, and many more. Below are a few prominent applications where mining is often utilized:

- Consumer propensities
- Marketing and advertising effectiveness
- E-commerce initiatives
- Fraud detection
- Worker and team performance
- Pricing policies
- Process-related applications (throughput, workflow, traffic analysis)
- Healthcare-related areas (outcomes measurement, treatment effectiveness)
- Risk assessment

What Is Likely to Happen

The other main source of information where mining provides value to decision makers is in the deployment of mining results. The patterns that have been identified are often embedded in an equation or algorithmic function, often referred to as the model, which can be used to perform a "what if" analysis or estimate future expected results based on inputs. In other words, if I market my product to a particular market segment defined by demographics, what is my expected response rate? Or, is a particular activity (e.g., credit card use) likely to be fraudulent? If the analysis

is based on a time series approach, mining models can provide forecasts for product sales. The analyst in this case needs to make assumptions as to future input values.

Real-Time Mining and Big Data

The evolution of the big data era has increased the utilization of the concept of real-time or streaming mining approaches. More traditional streaming mining involves the creation of models through analyzing a data sample or historical data of a given process. The resulting model then becomes a function that can be used to process streaming or real-time incoming data, and corresponding actionable outputs are generated in real time as well. Streaming mining addresses the big data concept of velocity and volume of data and is incorporated in processes where timely results are needed to improve strategies. Streaming mining applications are commonly applied in

- Website traffic analysis for real-time online marketing
- Fraud detection for online transactions
- Financial market risk and trading

The real time and continuous analytic section described later in this chapter along with Chapters 8 and 9 provide more detailed descriptions and applications in this area.

Some big data sources (e.g., sensor and satellite producing entities) with extreme velocity and volume sometimes render the ability to extract a sample that represents the entire data source as difficult, to say the least. In these instances, the ability to create optimized quantitative models to process this streaming data is limited. Techniques such as multisampling (Rajaraman and Ullman, 2011) and the implementation of self-optimizing quantitative techniques that learn as data is encountered have evolved to address this issue.

Analysis of Unstructured Data and Combining Structured and Unstructured Sources

Up to this point, this chapter has dealt with analytics of structured data. The big data era, however, largely involves the incorporation of unstructured data resources that need to be analyzed in order to identify actionable information that enhances strategic initiatives. Text mining addresses the

analytics of textual data (words, phrases, messages, emails, etc.). At a high level of description, text analytics seeks to create structure from unstructured sources. It does this by processing various unstructured forms and classifies them into particular categories. Processing is generally based in mathematics or linguistics.

In the realm of the vastly growing utilization of electronic communication, which includes texting, tweeting, leaving content on social media, emailing, etc., one can quickly see the possible value that exists in deploying analytic techniques to extract information that describes responses to marketing initiatives and product and service offerings, reactions to news and events, and general consumer behavior and sentiment.

An example involving the analysis of both structured and unstructured data for informative decision support is evident when examining patients' electronic health records (EHR) to better understand treatment outcomes and patient diagnosis.

More structured physiological data (e.g., blood sugar levels) can be combined with unstructured data (e.g., physician comments on treatment) to better understand a patient's status. Analytic techniques such as semantic mining can be applied in this situation to extract actionable information. The concept of mining unstructured data will be addressed in great detail in Chapters 4, 10, and 11.

SIX SIGMA ANALYTICS

Still many other analytic approaches exist outside the realm of BI applications. More intensive user-generated analytics include Six Sigma-based initiatives. The core of Six Sigma is a philosophy and focus for reducing variability in process operations. It involves process definition and the incorporation of an array of statistical analytic methods to measure the performance of various attributes (Pande and Neuman, 2000). Classic Six Sigma is underpinned by the DMAIC methodology, which is an acronym for the following:

Define: Process attributes and project objectives.
Measure: Identify relevant data variables and measure performance of the process.
Analyze: Identification of sources of unacceptable variances.
Improve: Initiate strategic tactics to address causes of variance.

Control: Establish metrics to measure performance for ongoing feedback and take appropriate actions to address shortcomings of process.

The initial three steps to the methodology clearly depict classic analytics as they involve the definition of the problem objective and corresponding use of statistics and techniques to analyze the performance of the process. In the big data era, new sources of descriptive variables and volumes can enhance the application of Six Sigma across processes and industries. Consider the recent evolution of the healthcare industry that has involved an aggressive adoption of information technologies to underpin the vast processes that exist in a healthcare provider's operations in treating patients.

Activity time stamps are commonplace for many processes in healthcare organizations that simply record when an activity of a subprocess begins and ends. This data is available at the patient level of detail. This seemingly trivial data element yields great significance in its facilitation of conducting analytics. Consider the activity of a patient checking in to an emergency room.

The entire process of checking in to an ER to being diagnosed is comprised of various subcomponents (Table 2.3). Variability or breakdowns in throughput in any of these subcomponents can adversely increase waiting times for patients, which can result in poor customer satisfaction ratings and the subpar outcome of the patient's well-being. A DMAIC scenario is provided to illustrate the analytic initiative.

In the ER scenario provided, the process has been defined (e.g., tracking the time to patient disposition from checking in to an ER), and the [D]/define step for DMAIC has been addressed. The next step is to create data variables that describe the various subcomponents of the process and measure corresponding performance rates.

- Patient checks in to ER
- Patient is moved to triage, where nurse is assigned and patient is moved to bed
- Nurse collects patient information (medical history)
- Medical service exam (MSE) is performed by a physician and tests are ordered
- Test results are received and patient disposition (admitted to hospital or sent home) is conducted

Time stamps of corresponding subcomponents to the process are generated and stored, where duration of each of the subcomponents must be

TABLE 2.3

Subcomponents of ER Throughput

Activity	Time	Duration	% Change	Alert
Check in at ER	2:00 a.m.			
Move to Triage	2:20 a.m.	20 min.	5%	
Information Collection	2:28 a.m.	8 min.	10%	
MSE by Physician	2:42 a.m.	14 min.	12%	
Disposition of Patient	3:15 a.m.	33 min.	85%	XXX

measured. In this case, the healthcare service provider has a historic perspective of measuring the process and has calculated the previous quarter's average duration for all the subcomponents. The next step is to analyze current performance (average for current month) to identify any significant changes to the baseline. Table 2.3 depicts a significant change (variance) in the duration of finishing the MSE to receiving lab results for patient disposition. Statistical techniques considering variance measurement are incorporated at the analytic stage to determine the level of significance, and therefore a need to implement the improve (I) stage. Here the analyst drills down into details of the process of ordering, conducting tests, and receiving results and communicating them back to the ER. At this stage, another, more detailed DMAIC study can be conducted to determine the factors that cause a high time duration from ordering to receiving test results to occur. Once this is accomplished, the decision maker can then formulate a strategic plan to address bottlenecks affecting the process (e.g., add radiology staff, adjust technology platform that communicates information in the test ordering process, implement an activity scheduling system). Once strategic initiatives have been implemented, the final step, control (C), follows to monitor effectiveness of the strategic endeavor and overall performance of the process (Kudyba and Radar, 2010).

The combination of available data (e.g., simple but voluminous sources of activity time stamps) in conjunction with a project and process definition and analytics enhances efficiency and organizational outcomes.

AN OFTEN OVERLOOKED SECTOR OF ANALYTICS (POWER OF THE SIMPLE GRAPHIC)

Although many think of analytics as crunching numbers through an array of techniques and interpreting metrics to support decision making,

analytics are greatly enhanced by the incorporation of an often taken for granted application of visual displays. Just think of having to analyze tables and columns of pure numbers when reviewing analytic reports. The process can quickly become mundane and even painful. In the host of analytic applications we described above and for numerous additional analytic methods, there is a common denominator to a successful endeavor, and that is the use of graphics to disseminate information. A simple view of a well-designed graphic can provide the decision maker with a clear presentation of extensive analytic results in a comprehendible manner.

In order to successfully leverage graphics, a few key points need to be considered. Before you become intrigued with robust colors and images that quickly draw you to generate dramatic conclusions about a particular process, take a step back and increase your understanding of what the information is actually portraying. In other words:

1. Analyze the titles and legends.
2. Take notice of the scale of the axis.
3. Understand the graphic/chart method used.

When you fully understand the variables that are depicted in the graphic, what the type of graphic focuses on, and the scale of the axis, only then can the analyst begin to generate effective interpretations. In the following section, a variety of graphical styles are listed with some simple descriptions of when they should be used. Keep in mind that when considering graphics in a big data era, the most significant elements are real-time graphics that provide analysts with a streaming view of processes. The real-time streaming visualization of data actually becomes a dashboard that analysts can monitor to observe variances in KPIs in relation to some event.

Graphic Types

Figure 2.5 illustrates the classic pie chart that depicts how a whole unit is divided among some subcomponents (pieces of an established pie). Market share is a prime example for pie charts, where share can be delineated by product lines, regions, industry competitors, etc. Pie charts have limitations when considering negative values.

Despite the seemingly simplistic bar chart depicted in Figure 2.6, the visual actually incorporates a number of important elements in the realm of analytics. The graphic depicts a comparative view of a multicomponent

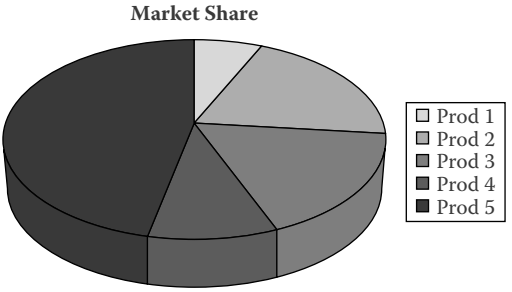


FIGURE 2.5
Pie chart depicting market share.

process (call centers in this case) in a time series setting (quarterly views). With a quick glance, the analyst can make inferences regarding relative performance (customer satisfaction) of three different call centers over time. Bar charts are more appropriate in depicting quantities or amounts of select variables.

Bar charts are also often used to illustrate variable distributions (percentages of ranges or categories of a given variable). Figure 2.7 depicts a categorical age variable and the amount of data that exists in selected ranges. This gives analysts a better understanding of the dimensions of a given data variable, and in this case enables them to determine if there is any age skew or bias (high percentage of one age range relative to the population).

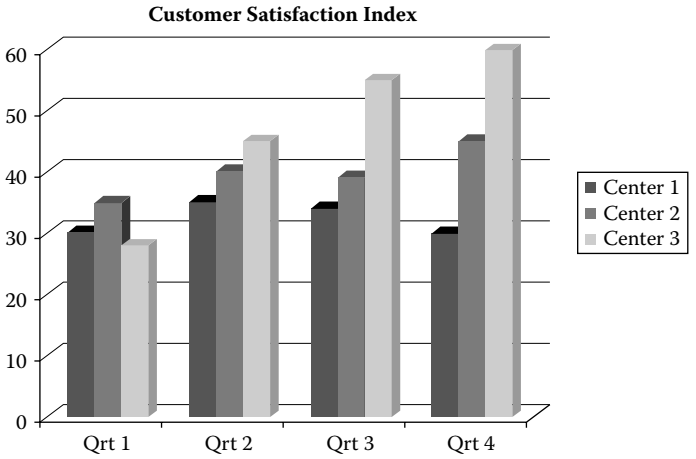


FIGURE 2.6
Bar chart (comparative view of multi-component process).

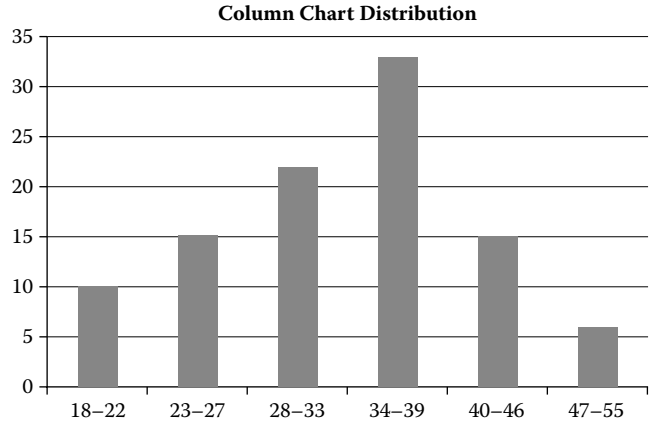


FIGURE 2.7
Age distribution chart.

In conducting market research, a variable distribution view enables the researcher to determine if a target market is included in a data resource.

Variable distribution analysis can often include visuals via line graphs that are useful in illustrating scenarios involving continuous variables. Figure 2.8 illustrates the continuous data variable of mall foot traffic for a given day according to retailers.

Time series line charts provide users with a visual of potential seasonality in processes. Figure 2.9 depicts the classic holiday effect in retail as is seen in the repetitive bump in sales in Q4.

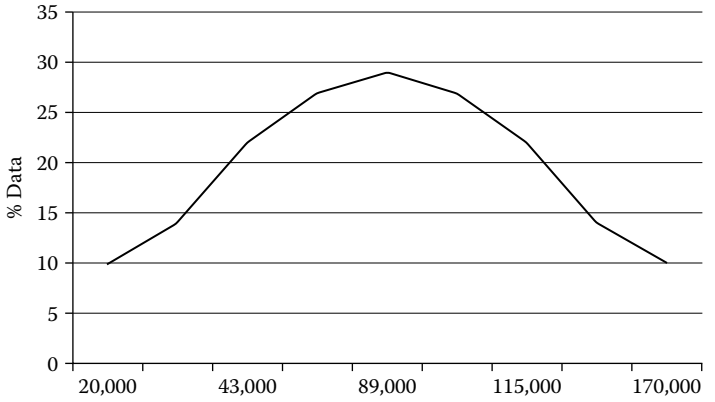


FIGURE 2.8
Line chart of continuous variable distribution of mall traffic.

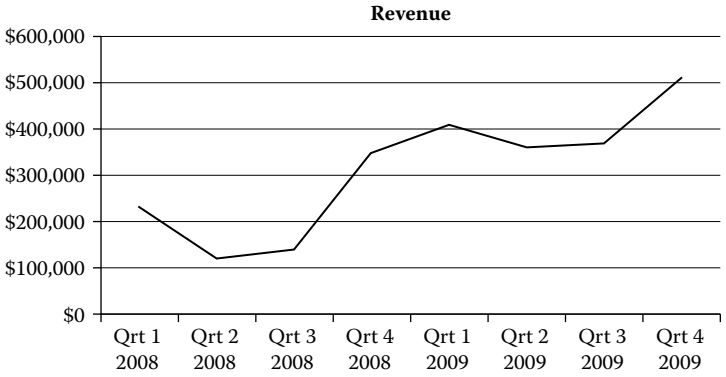


FIGURE 2.9
Time series line charts for seasonality.

Another type of chart involves the scatter plot that is commonly used to illustrate correlations between variables, where simple plots of individual data points are depicted. Figure 2.10 depicts data points depicting correlations between employee performance and training received.

A rather insightful chart style is the bubble chart. The bubble graphic enables analysts to depict three-dimensional scenarios in a coherent fashion by incorporating bubble size to illustrate variable attributes. Figure 2.11 depicts the multi-dimensional scenario of organizational team performance according to workload and team size.

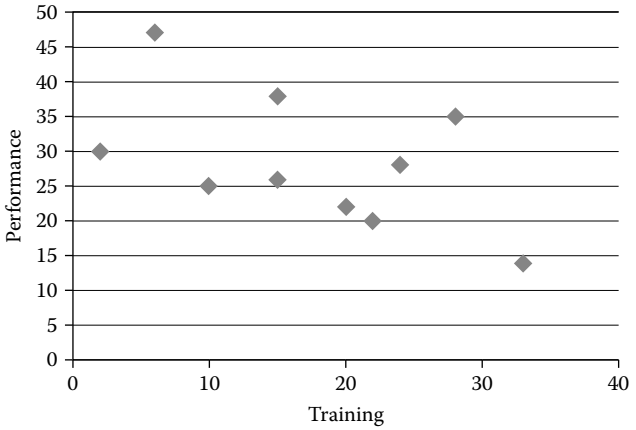


FIGURE 2.10
Scatterplot for correlations.

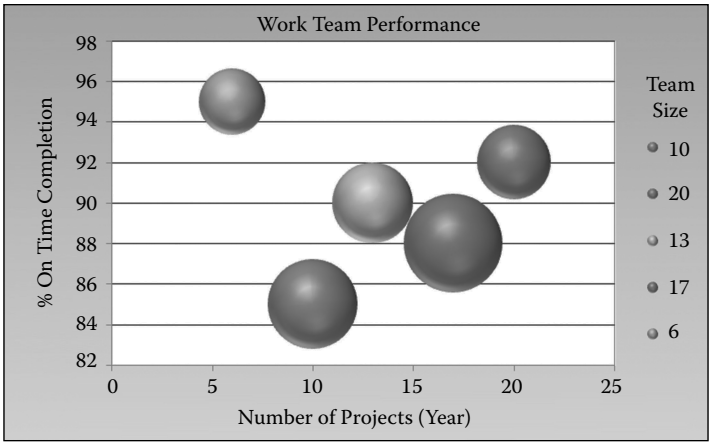


FIGURE 2.11
(See color insert.) Bubble chart depicting workforce team performance.

Yet another graphic style that has increased in importance over the evolution of the big data era is the use of maps. Map visuals are generally utilized when an analysis involving location is emphasized; however, location can also refer to a process location. Applications such as traffic analysis or population analytics are common examples. Traffic can refer to website activities, vehicular, consumer, or some type of designated activity. See Chapter 7 for more on heat maps for the web.

In a simple web traffic visual, a map can illustrate cross sections of time and area of a web page that are receiving high user traffic. This can provide strategists with actionable information to more effectively apply online marketing tactics (e.g., display banners in hot spots on a particular page at a particular time).

Civil engineering can leverage heat maps by incorporating GPS data to investigate hot areas of traffic incidents (congestion, accidents) and optimize new designs to alleviate existing trouble areas and in designing new roadways.

Figure 2.12 provides a standard heat map where "hot colors" depict more intense activity. In this case, the hotter areas depict areas where job vacancies are difficult to fill.

Map visuals are particularly applicable in the big data era, when real-time, high-velocity analytics and voluminous sources are involved. Applications that leverage big data include geovisualization that involves the analysis of geographic specific flows of data and bioinformatics and sensor output in the healthcare spectrum. For example, the healthcare industry is increasingly

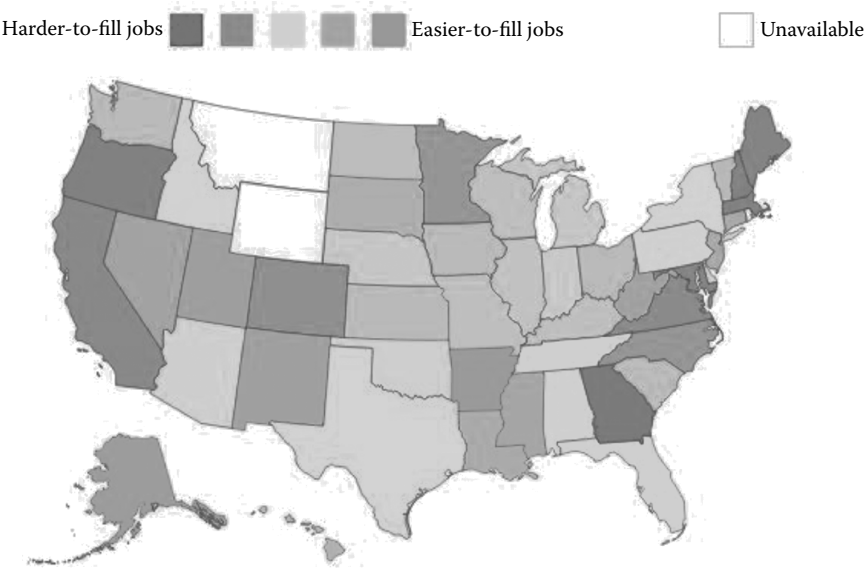


FIGURE 2.12
(See color insert.) Heat map that illustrates areas of hard-to-fill job vacancies. (From Wanted Analytics, http://www.wantedanalytics.com.)

utilizing streaming sensor data generated by various treatment and diagnostic technologies. For diagnosis (MRI), these data describe the characteristics of a patient. Visual displays of this source are essential to extract information on trouble areas for patients. Chapter 12 provides more detailed information corresponding to this concept. As big data sources emerge, the application of heat maps should become a common visual technique for providing analysts with a mechanism to enhance strategic initiatives.

This concludes our section on visual analytics. The following section provides a more technical description of data management and analytic concepts in a high volume, big data environment. Event stream processing (ESP) can be utilized for analyzing streaming big data in motion, where this technique is often utilized as a front end for historical analytics as well.

REAL-TIME AND CONTINUOUS ANALYTICS*

Complex event processing (CEP) refers to systems that process or analyze events closer to the creation of those events, prior to storage. Some

* This section contributed by Jerry Baulier of SAS, Inc.

definitions refer to CEP events from the "event cloud" because it's not always defined where events will be published into CEP for processing and whether they will be ordered or not. This is very different than traditional analytics, or historical analytics, where data is first stored persistently (usually in a database) before it is analyzed or processed. Complex event processing systems reduce latency times of analyzing data, usually referred to as events, by analyzing or processing the data before it is stored. In fact, sometimes this form of analytics is done as a front end to historical analytics in order to aggregate (or reduce) raw data before storing it persistently for further analytics. This combination of complex event processing and front-ending historical analytics (as depicted in Figure 2.13) can be a very powerful, multistage approach to analyzing data for actionable intelligence and consequently acting on intelligence sooner than otherwise possible, which can sometimes create new business opportunities than otherwise possible.

The term *complex* in complex event processing sometimes creates confusion relative to what it refers to. Complex event processing systems analyze events that are published into them and create derived (or synthetic) events that represent some transformation of the events that were published into them. These derived events usually represent an aggregated view over time of many input events, and hence are referred to as complex events. While the name *complex event processing* is relatively new, the science of systems analyzing events in motion (before they are stored) has been around for some time. This science of analyzing events in motion consisted largely of two types of processing models. The first type was more rule based, both

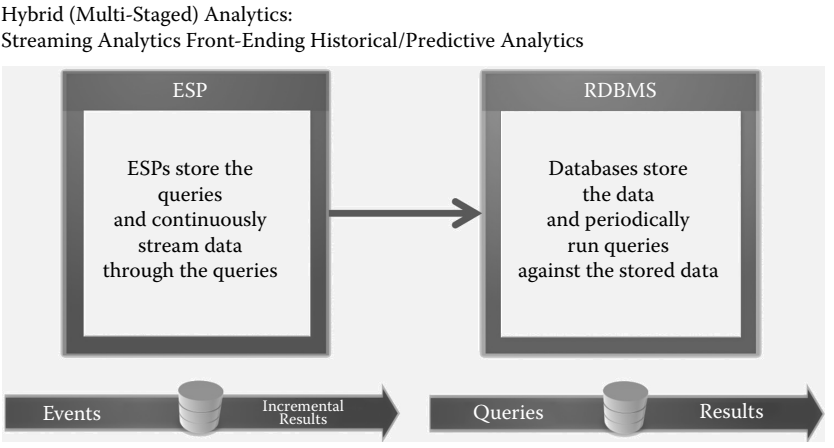


FIGURE 2.13
Event processing and front-ending historical analysis for faster decision making.

inference and event, condition, action (ECA). The second type of systems were based more on the concept of continuous queries, were mostly relational in nature, and they analyzed (or processed) events that were published into them as event streams. These systems tended to focus on processing high volumes of events in very low latencies, therefore creating new opportunities around leveraging actionable intelligence in very small time frames from the event occurrences. Systems that utilize continuous queries to analyze events in motion are referred to as event stream processing (ESP), even though many of these systems have product names that include complex event processing. So event stream processing can be viewed as a type of complex event processing system that utilizes continuous query models to analyze event streams in motion. Having made this initial distinction between rules-based and continuous query-based CEPs, there has been a blending of these approaches over time, and at least one rather prominent rules-based engine has focused on low-latency applications such as algorithmic trading in capital markets. The remainder of this section will focus on continuous query-based CEPs, which we will refer to as ESP.

As mentioned above, ESP systems are usually architected to handle large volumes of events with very low latencies, which is why a lot of these systems focused on capital markets, where front office trading, risk, and position management systems needed to make decisions in milliseconds, or more likely today, microseconds. These systems process financial market data from real-time financial feeds (such as trades, quotes, and orders) where the volumes could reach millions of events per second. ESP systems are being applied across many markets and applications, including personalized marketing, trading systems, operational predictive asset management, fraud prevention, and even cyber security.

The continuous annual growth of data, much of this being computer generated, has spawned the evolution of new methods to be able to analyze all of this data, and ESP helps fill that need by analyzing large volumes of raw data (or event streams) and looking for actionable intelligence. This actionable intelligence is usually aggregated events, and hence a significant level of data reduction by ESP systems can be done before it makes its way to the more traditional historical analytics. This form of multistage analytics helps keep up with the growth of data and enables actionable intelligence to be obtained in lower latencies, hence enabling opportunity. ESP systems will sometimes find patterns or other aggregations of events that are directly actionable, and other times they find aggregations of events that need to be analyzed further before they become actionable.

This multistage analytics approach is also very useful given that the type of analytics done by ESP is quite different than traditional statistical-based analytics. Historical analytics typically works on a static data set at rest and applies complex algorithms and searches that can be very statistical in nature. Event stream processing models, however, are much more additive and data transformational in nature. ESP models are often referred to as continuous queries because they essentially query data in motion, where the resultant set of the queries is continuously being updated. For this type of processing, it is important that the queries are additive in nature. In other words, when a new event comes in, these queries typically update the retained state of the queries without having to reread all the events already processed.

These continuous queries can be modeled as directed graphs where input event streams are published into the top of the graph, and each subsequent node in the graph performs transformations on the events it receives. The nodes of the graph can be thought of as windows with operators and retention policies. So each node performs a transformation on each event it gets, producing zero, one, or more resultant events that are retained in the window for a defined period of time or volume, and are passed on to the sibling nodes connected to that window. The nodes come in various forms, including relational, procedural, and rules. Relational nodes include all the primitives of Structured Query Language (SQL), such as joins, aggregates, projections, and filters. Most ESP systems also support pattern matching, which is a rule-based window that allows one to specify event patterns of a temporal nature. For example:

Event-A **Followed-by** (within 2 minutes) *Event-B*
Followed-by (within 1 minute) **Not** *Event-C*

where we are looking for an event pattern such that *Event-A* occurs, and that is followed within 2 minutes by *Event-B* occurring, which in turn is not followed by the occurrence of *Event-C* for the next minute.

Procedural nodes, or windows, typically allow one to write event stream handlers using some procedural language (like C++) to process the events from the event streams that are input into that node or window. So one can model continuous queries as one or more directed graphs of data transformation nodes that work in parallel to produce continuous results in very low latencies.

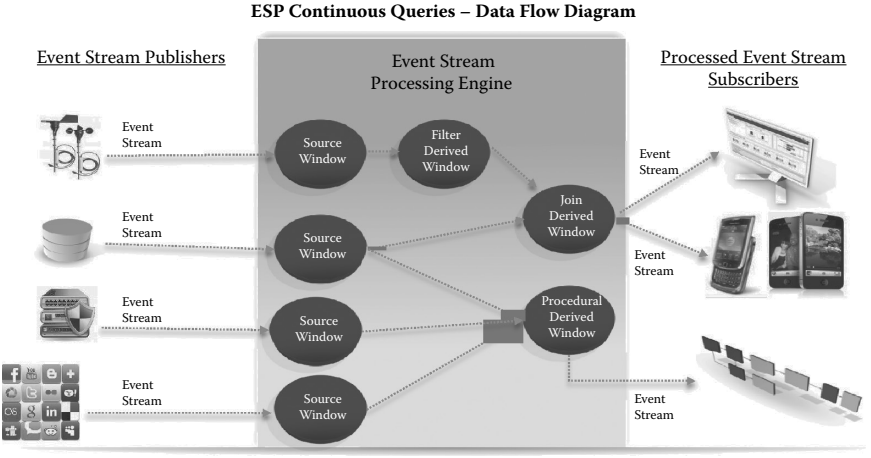


FIGURE 2.14
ESP continuous query.

Figure 2.14 is an example of a very simplistic ESP continuous query where event streams are published into the continuous query via source windows. Once each event is absorbed into the source window, it makes its way to any connected windows or subscribers. Every window type that is not a source window is referred to as a derived (or synthetic) window because they have an associated operator that transforms each event coming into it to zero, one, or more resultant events. In our simple example, we show window types filter, join, and procedural. The terminology (such as windows) is not standardized across ESP systems, but the basic concepts around continuous queries and their node types hold true. ESP continuous queries are data transformation models that continuously query event streams as they flow through the ESP where each window or operator in the flow could retain an event state that represents some period of recent events. One can model very powerful continuous queries that can reduce data into something immediately actionable or at least more relevant for further historical or predictive analytics.

VALUE OF DATA AND ANALYTICS

We began this chapter by stressing the importance of analytics as an essential component to deriving value from data, where the era of big data

adds intensity to the concept, as it adds new dimensions to the equation. Regardless of the source of data, its value is not realized unless it provides some resource to a strategic endeavor. Rarely does a decision maker reference a data source without first formulating a reason to do so. Once the conceptual need is defined, only then can data provide value.

The conceptual need involves the quest to better understand a process with the goal of enhancing its efficiency, productivity, or profitability. Simply analyzing random data and coming up with associations between variables may actually generate negative returns since the analytic process requires time and resources, and the result may not add meaningful information.

Consider the growing data resource in the area of sports. More variables (variety) and real-time downloads of various athletic activities at corresponding events (velocity and volume) may seemingly provide great value to understanding various attributes of different athletes and sports teams. However, in order to truly generate value for decision making, a conceptual model must be created. Consider the quest to better understand what leads a team to achieve a winning record. An analysis of corresponding data could yield the following result: a basketball team is more likely to win a game the more 3-point shots they make.

At first glance, this may seem to be very valuable information, but the revelation proves limited at best when looking to make a strategic decision. What does a coach do in leveraging this associative pattern—encourage players to take more shots from the 3-point zone? Does he change practice to intensify skills for increasing 3-point percentages for players? And if so, what happens to the team's performance from the 2-point zone, and does a reduction in 2-point conversions decrease the likelihood of winning a game despite an increase in 3-point shots? In the case at hand, does the variable of number of 3-point shots really add descriptive value to what leads to a team's success. Perhaps more appropriate variables that can provide strategic action could entail:

Team practice data (frequency, drills, duration)
Player descriptions (height, speed, position, age)
Type of offensive and defensive tactics

Identifying patterns among these types of variables empowers a coach (decision maker/strategist) to implement strategic initiatives that impact a performance metric or defined objective—winning.

Efficiency, Productivity, and Profitability

The concept of value also extends to three often cited benchmarks in the realm of commerce: efficiency, productivity, and profitably. One should note that although the three terms appear synonymous, there are noteworthy differences among them, so when seeking to succeed in strategic endeavors, decision makers must clearly understand the entire initiative from the perspective of these three concepts.

Analytics of all types naturally address the quest for enhancing efficiencies of corresponding processes. Enhancing efficiency naturally leads to cost reduction for the defined process; however, simply increasing efficiency for a particular activity does not necessarily imply an increase in productivity and profitability at the organizational level. Consider a marketing department for a small retailer that depends on more traditional mail order initiatives to generate sales. The department could consistently achieve increased efficiency in the process of creating printed marketing materials, generating addresses, and mailing literature to the market. These efficiencies could be achieved by implementing new printing technologies, data-based endeavors, etc. However productivity as measured by response rate or increased product sales may not necessarily increase. Perhaps traditional mail is no longer the most effective marketing medium for the type of product given the evolution of e-marketing tactics and the adoption of smartphones and electronic communication by consumers, or perhaps the target market has changed its behavior and a different segment is actually more appropriate for the product. What may actually transpire for this endeavor is an efficient process that yields decreased productivity for the organization (deploying resources and achieving decreased returns).

Just as analytics were utilized to better understand what drives wasteful activities for the mail order marketing initiative, so too should they be utilized for such endeavors as better understanding overall marketing effectiveness and target marketing. Simply put, strategic endeavors must incorporate a bigger picture than simple processes, which provides a segue to the third concept of value, which involves profitability. Profitability must be included in any endeavor where productive and efficient strategies must make sense on a cost and revenue basis.

Investment in Internet advertising has grown dramatically over the past years, partially because of the cost-effectiveness of some tactics; however, the recent growth of this sector has also resulted in increased costs, a variable that needs to be monitored. A core tactic to e-marketing is search

engine optimization for websites, an initiative that is ongoing and incurs costs of continuous management (verbiage of meta-tags, frequency of key phrases, reciprocal links). These costs must be considered in the overall spectrum of e-commerce initiatives. So when leveraging big data feeds involving site traffic relative to page layouts and cross-selling tactics, costs of an entire space need to be managed to understand profitability.

This introduces an important issue in the big data era. Big data is not free and involves technological infrastructure (servers, software) for data management along with labor resources (analytic and technical minds) to leverage this complex resource. Organizations must therefore fully understand how big data resources may impact their operations. Just because new volumes and velocities of data exist doesn't imply that the resource will be a value to every organization. An important concept to keep in mind to estimate this value is answering the question: Will new data resources help an organization better understand its process performance or marketplace?

With new resources come new ideas that leverage them. The evolving big data era in conjunction with new information technologies has introduced an opportunity for organizations to create new markets. Through analytics, big data is transformed into information that provides value by enhancing decision-making capabilities through knowledge generation. Organizations can better understand those processes essential to their operations. The concept of the creation and dissemination of information goes beyond value to organizations as it extends to individuals. Insightful information derived from big data generated from wireless sensor devices can be made available to consumers that may provide beneficial value to them. Consider a wireless device (body sensor) manufacturer that extracts information from users that may enable them to estimate and offer optimized fitness programs on a personalized basis, thus augmenting the value of the product device to existing consumers and adding value to new prospects. As the big data, analytic, and communication era continues to evolve, innovative initiatives that involve information creation in areas previously untapped can prove to be a bold new market.

The big data era requires the analyst to more intensely exhaust data resources that may provide descriptive elements of a given process as new varieties of data variables evolve. Analysts must also consider whether a process must be monitored in a real-time environment (velocity/volume) in order to uncover strategic insights in the information creation and decision-making process. There is little doubt that the job of analyst has

become more important and also more complex. The remainder of this book should provide insights to help enlighten individuals as to some of the important concepts that are involved in this space.

REFERENCES

Baulier, J. (contrib.). Real Time and Continuous Analytics. Cary, NC: SAS.

Kudyba, S., and Hoptroff, R. *Data Mining and Business Intelligence: A Guide to Productivity.* Hershey, Pennsylvania: IDEA Group Publishing, 2001.

Kudyba, S., and Radar, R. Enhancing Data Resources and Business Intelligence in Healthcare. In *Healthcare Informatics: Increasing Efficiency and Productivity.* Boca Raton, Florida: Taylor & Francis, 2010.

Pande, P., and Neuman, R. *The Six Sigma Way: How GE, Motorola, and Other Top Companies Are Honing Their Performance*. New York: McGraw Hill, 2000.

Rajaraman, A., and Ullman, J. Mining Data Streams. In *Mining of Massive Data Sets*. Cambridge: Cambridge University Press, 2011, pp. 129–132.