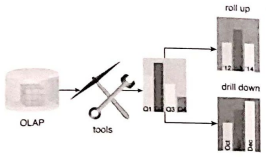
Figure 1.6
Diagnostic analytics can result in data that is suitable for performing drill-down and roll-up analysis.



Predictive Analytics

Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future. With predictive analytics, information is enhanced with meaning to generate knowledge that conveys how that information is related. The strength and magnitude of the associations form the basis of models that are used to generate future predictions based upon past events. It is important to understand that the models used for predictive analytics have implicit dependencies on the conditions under which the past events occurred. If these underlying conditions change, then the models that make predictions need to be updated.

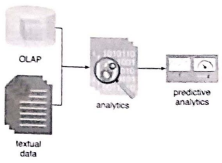
Questions are usually formulated using a what-if rationale, such as the following:

- What are the chances that a customer will default on a loan if they have missed a monthly payment?
- What will be the patient survival rate if Drug B is administered instead of Drug A?
- If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?

Predictive analytics try to predict the outcomes of events, and predictions are made based on patterns, trends and exceptions found in historical and current data. This can lead to the identification of both risks and opportunities.

This kind of analytics involves the use of large datasets comprised of internal and external data and various data analysis techniques. It provides greater value and requires a more advanced skillset than both descriptive and diagnostic analytics. The tools used generally abstract underlying statistical intricacies by providing user-friendly front-end interfaces, as shown in Figure 1.7.

Figure 1.7
Predictive analytics tools can provide user-friendly front-end interfaces.



Prescriptive Analytics

Prescriptive analytics build upon the results of predictive analytics by prescribing actions that should be taken. The focus is not only on which prescribed option is best to follow, but why. In other words, prescriptive analytics provide results that can be reasoned about because they embed elements of situational understanding. Thus, this kind of analytics can be used to gain an advantage or mitigate a risk.

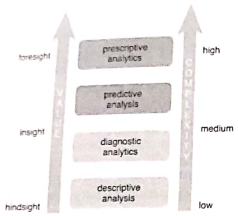
Sample questions may include:

- Among three drugs, which one provides the best results?
- When is the best time to trade a particular stock?

Prescriptive analytics provide more value than any other type of analytics and correspondingly require the most advanced skillset, as well as specialized software and tools. Various outcomes are calculated, and the best course of action for each outcome is suggested. The approach shifts from explanatory to advisory and can include the simulation of various scenarios.

This sort of analytics incorporates internal data with external data. Internal data might include current and historical sales data, customer information, product data and business rules. External data may include social media data, weather forecasts and government-produced demographic data. Prescriptive analytics involve the use of business rules and large amounts of internal and external data to simulate outcomes and prescribe the best course of action, as shown in Figure 1.8.

Figure 1.4
Value and complexity increase from descriptive to prescriptive analytics



Descriptive Analytics

Descriptive analytics are carried out to answer questions about events that have already occurred. This form of analytics contextualizes data to generate information.

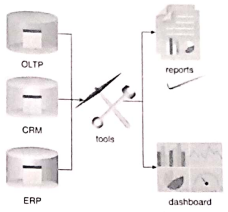
Sample questions can include:

- What was the sales volume over the past 12 months?
- What is the number of support calls received as categorized by severity and geographic location?
- What is the monthly commission earned by each sales agent?

It is estimated that 80% of generated analytics results are descriptive in nature. Value-wise, descriptive analytics provide the least worth and require a relatively basic skillset.

Descriptive analytics are often carried out via ad-hoc reporting or dashboards, as shown in Figure 1.5. The reports are generally static in nature and display historical data that is presented in the form of data grids or charts. Queries are executed on operational data stores from within an enterprise, for example a Customer Relationship Management system (CRM) or Enterprise Resource Planning (ERP) system.

Figure 1.5
The operational systems, pictured left, are queried via descriptive analytics tools to generate reports or dashboards, pictured right.



Diagnostic Analytics

Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event. The goal of this type of analytics is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.

Such questions include:

- Why were Q2 sales less than Q1 sales?
- Why have there been more support calls originating from the Eastern region than from the Western region?
- Why was there an increase in patient re-admission rates over the past three months?

Diagnostic analytics provide more value than descriptive analytics but require a more advanced skillset. Diagnostic analytics usually require collecting data from multiple sources and storing it in a structure that lends itself to performing drill-down and roll-up analysis, as shown in Figure 1.6. Diagnostic analytics results are viewed via interactive visualization tools that enable users to identify trends and patterns. The executed queries are more complex compared to those of descriptive analytics and are performed on multi-dimensional data held in analytic processing systems.

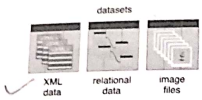


Figure 1.1
Datasets can be found in many different formats.

Data Analysis

Data analysis is the process of examining data to find facts, relationships, patterns, insights and/or trends. The overall goal of data analysis is to support better decision-making. A simple data analysis example is the analysis of ice cream sales data in order to determine how the number of ice cream cones sold is related to the daily temperature. The results of such an analysis would support decisions related to how much ice cream a store should order in relation to weather forecast information. Carrying out data analysis helps establish patterns and relationships among the data being analyzed. Figure 1.2 shows the symbol used to represent data analysis.

Figure 1.2
The symbol used to represent data analysis.

Data Analytics

Data analytics is a broader term that encompasses data analysis. Data analytics is a discipline that includes the management of the complete data lifecycle, which encompasses collecting, cleansing, organizing, storing, analyzing and governing data. The term includes the development of analysis methods, scientific techniques and automated tools. In Big Data environments, data analytics has developed methods that allow data analysis to occur through the use of highly scalable distributed technologies and frameworks that are capable of analyzing large volumes of data from different sources. Figure 1.3 shows the symbol used to represent analytics.

The Big Data analytics lifecycle generally involves identifying, procuring, preparing and analyzing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data and performing large-scale searches.

Figure 1.3
The symbol used to represent data analytics.

Different kinds of organizations use data analytics tools and techniques in different ways. Take, for example, these three sectors:

- In business-oriented environments, data analytics results can lower operational costs and facilitate strategic decision-making.
- In the scientific domain, data analytics can help identify the cause of a phenomenon to improve the accuracy of predictions.
- In service-based environments like public sector organizations, data analytics can help strengthen the focus on delivering high-quality services by driving down costs.

Data analytics enable data-driven decision-making with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone. There are four general categories of analytics that are distinguished by the results they produce:

- descriptive analytics
- diagnostic analytics
- predictive analytics
- prescriptive analytics

The different analytics types leverage different techniques and analysis algorithms. This implies that there may be varying data, storage and processing requirements to facilitate the delivery of multiple types of analytic results. Figure 1.4 depicts the reality that the generation of high value analytic results increases the complexity and cost of the analytic environment.

Data solution can be used by enterprise applications directly or can be fed into a data warehouse to enrich existing data there. The results obtained through the processing of Big Data can lead to a wide range of insights and benefits, such as:

- operational optimization
- actionable intelligence
- identification of new markets
- accurate predictions
- fault and fraud detection
- more detailed records
- improved decision-making
- scientific discoveries

Evidently, the applications and potential benefits of Big Data are broad. However, there are numerous issues that need to be considered when adopting Big Data analytics approaches. These issues need to be understood and weighed against anticipated benefits so that informed decisions and plans can be produced. These topics are discussed separately in Part II.

Concepts and Terminology

As a starting point, several fundamental concepts and terms need to be defined and understood.

Datasets

Collections or groups of related data are generally referred to as datasets. Each group or dataset member (datum) shares the same set of attributes or properties as others in the same dataset. Some examples of datasets are:

- tweets stored in a flat file
- a collection of image files in a directory
- an extract of rows from a database table stored in a CSV formatted file
- historical weather observations that are stored as XML files

Figure 1.1 shows three datasets based on three different data formats.

Time Series Plot

The analysts want to find out whether or not the fraudulent claims are time-dependent. They are particularly interested in finding out if there are any particular time periods in which the number of fraudulent claims increases. A time series of fraudulent claims for the past five years is generated based on the number of fraudulent claims that were calculated each week. A visual analysis of the time series plot reveals a seasonal trend that shows that the number of fraudulent claims goes up just before a holiday and toward the end of summer. These results suggest that either customers make false claims in order to have money for the holiday period or they upgrade their electronics and other goods after a holiday by reporting damage or theft. A few short-term irregular variations are also found, which, upon closer inspection, are discovered to be linked with catastrophes like floods and storms. The long-term trend suggests that the number of fraudulent claims is likely to increase in the future.

Clustering

Although all of the fraudulent claims are different, the analysts are interested in finding out if any similarities exist between fraudulent claims. A clustering technique is applied that groups different fraudulent claims based on a number of attributes, such as customer age, policy age, gender, number of previous claims and frequency of claim.

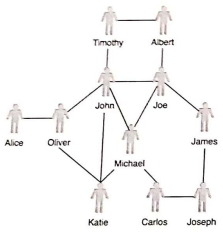
Classification

During the *utilization of analysis results* stage, the classification analysis technique is used to develop a model that can differentiate between a legitimate claim and a fraudulent claim. For this, the model is first trained using a dataset of historic claims, in which each claim is labeled as either *legitimate* or *fraudulent*. Once trained, the model is brought online, where newly-submitted, unlabeled claims are classified as fraudulent or legitimate.

Consider the social network graph in Figure 8.18 for a simple example of social network analysis:

- John has many friends, whereas Alice only has one friend.
- The results of a social network analysis reveal that Alice will most likely befriend John and Katie, since they have a common friend named Oliver.

Figure 8.18
An example of a social network graph.



Sample questions may include:

- *How can I identify influencers within a large group of users?*
- *Are two individuals related to each other via a long chain of ancestry?*
- *How can I identify interaction patterns among a very large number of protein-to-protein interactions?*

Spatial Data Mapping

Spatial or geospatial data is commonly used to identify the geographic location of individual entities that can then be mapped. Spatial data analysis is focused on analyzing location-based data in order to find different geographic relationships and patterns between entities.

Spatial data is manipulated through a Geographic Information System (GIS) that plots spatial data on a map generally using its longitude and latitude coordinates. The GIS

provides tooling that enables interactive exploration of the spatial data, for example measuring the distance between two points, or defining a region around a point as a circle with a defined distance-based radius. With the ever-increasing availability of location-based data, such as sensor and social media data, spatial data can be analyzed to gain location insights.

For example, as part of a corporate expansion, more ice cream stores are planned to open. There is a requirement that no two stores can be within a distance of 5 kilometers of each other to prevent the stores from competing with each other. Spatial data is used to plot existing store locations and to then identify optimal locations for new stores at least 5 kilometers away from existing stores.

Applications of spatial data analysis include operations and logistic optimization, environmental sciences and infrastructure planning. Data used as input for spatial data analysis can either contain exact locations, such as longitude and latitude, or the information required to calculate locations, such as zip codes or IP addresses.

Furthermore, spatial data analysis can be used to determine the number of entities that fall within a certain radius of another entity. For example, a supermarket is using spatial analysis for targeted marketing, as shown in Figure 8.19. Locations are extracted from the users' social media messages, and personalized offers are delivered in realtime based on the proximity of the user to the store.

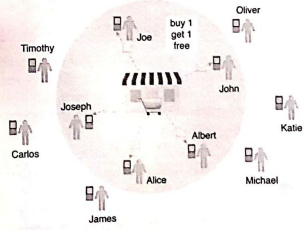


Figure 8.19
Spatial data analysis can be used for targeted marketing.

Sample questions can include:

- *How can I visually identify any patterns related to carbon emissions across a large number of cities around the world?*
- *How can I see if there are any patterns of different types of cancers in relation to different ethnicities?*
- *How can I analyze soccer players according to their strengths and weaknesses?*

Time Series Plots

Time series plots allow the analysis of data that is recorded over periodic intervals of time. This type of analysis makes use of time series, which is a time-ordered collection of values recorded over regular time intervals. An example is a time series that contains sales figures that are recorded at the end of each month.

Time series analysis helps to uncover patterns within data that are time-dependent. Once identified, the pattern can be extrapolated for future predictions. For example, to identify seasonal sales patterns, monthly ice cream sales figures are plotted as a time series, which further helps to forecast sales figures for the next season.

Time series analyses are usually used for forecasting by identifying long-term trends, seasonal periodic patterns and irregular short-term variations in the dataset. Unlike other types of analyses, time series analysis always includes time as a comparison variable, and the data collected is always time-dependent.

A time series plot is generally expressed using a line chart, with time plotted on the x-axis and the recorded data value plotted on the y-axis, as shown in Figure 8.17.

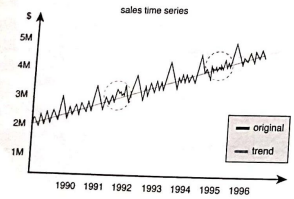


Figure 8.17
A line chart depicts a sales time [illegible]

The time series presented in Figure 8.17 spans seven years. The evenly spaced peaks toward the end of each year show seasonal periodic patterns, for example Christmas sales. The dotted red circles represent short-term irregular variations. The blue line shows an upward trend, indicating an increase in sales.

Sample questions can include:

- *How much yield should the farmer expect based on historical yield data?*
- *What is the expected increase in population in the next 5 years?*
- *Is the current decrease in sales a one-off occurrence or does it occur regularly?*

Network Graphs

Within the context of visual analysis, a network graph depicts an interconnected collection of entities. An entity can be a person, a group, or some other business domain object such as a product. Entities may be connected with one another directly or indirectly. Some connections may only be one-way, so that traversal in the reverse direction is not possible.

Network analysis is a technique that focuses on analyzing relationships between entities within the network. It involves plotting entities as nodes and connections as edges between nodes. There are specialized variations of network analysis, including:

- route optimization
- social network analysis
- spread prediction, such as the spread of a contagious disease

The following is a simple example based on ice cream sales for the application of network analysis for route optimization.

Some ice cream store managers are complaining about the time it takes for delivery trucks to drive between the central warehouse and stores in remote areas. On hotter days, ice cream delivered from the central warehouse to the remote stores melts and cannot be sold. Network analysis is used to find the shortest routes between the central warehouse and the remote stores in order to minimize the durations of deliveries.

Visual Analysis

Visual analysis is a form of data analysis that involves the graphic representation of data to enable or enhance its visual perception. Based on the premise that humans can understand and draw conclusions from graphics more quickly than from text, visual analysis acts as a discovery tool in the field of Big Data.

The objective is to use graphic representations to develop a deeper understanding of the data being analyzed. Specifically, it helps identify and highlight hidden patterns, correlations and anomalies. Visual analysis is also directly related to exploratory data analysis as it encourages the formulation of questions from different angles.

This section describes the following types of visual analysis:

- Heat Maps
- Time Series Plots
- Network Graphs
- Spatial Data Mapping

Heat Maps

Heat maps are an effective visual analysis technique for expressing patterns, data compositions via part-whole relations and geographic distributions of data. They also facilitate the identification of areas of interest and the discovery of extreme (high/low) values within a dataset.

For example, in order to identify the top- and worst-selling regions for ice cream sales, the ice cream sales data is plotted using a heat map. Green is used to highlight the best performing regions, while red is used to highlight worst performing regions.

The heat map itself is a visual, color-coded representation of data values. Each value is given a color according to its type or the range that it falls under. For example, a heat map may assign the values of 0–3 to the color red, 4–6 to amber and 7–10 to green.

A heat map can be in the form of a chart or a map. A chart represents a matrix of values in which each cell is color-coded according to the value, as shown in Figure 8.15. It can also represent hierarchical values by using color-coded nested rectangles.

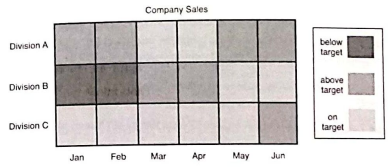


Figure 8.15
This chart heat map depicts the sales of three divisions within a company over a period of six months.

In Figure 8.16, a map represents a geographic measure by which different regions are color-coded or shaded according to a certain theme. Instead of coloring or shading the whole region, the map may be superimposed by a layer made up of collections of colored/shaded points relating to various regions, or colored/shaded shapes representing various regions.

Figure 8.16
A heat map of the US sales figures from 2013.

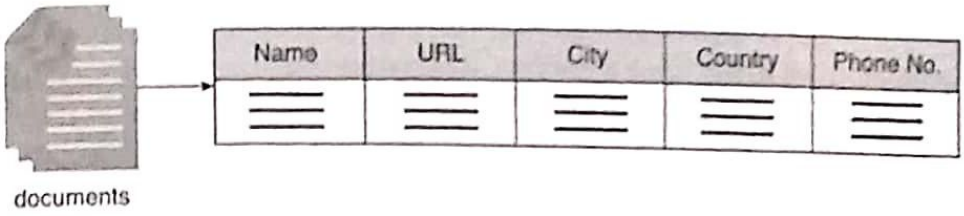


Figure 8.14
Entities are extracted from text files using semantic rules and structured so that they can be searched.

Sample questions can include:

- *How can I categorize Web sites based on the content of their Web pages?*
- *How can I find the books that contain content that is relevant to the topic that I am studying?*
- *How can I identify contracts that contain confidential company information?*

Sentiment Analysis

Sentiment analysis is a specialized form of text analysis that focuses on determining the bias or emotions of individuals. This form of analysis determines the attitude of the author of the text by analyzing the text within the context of the natural language. Sentiment analysis not only provides information about how individuals feel, but also the intensity of their feeling. This information can then be integrated into the decision-making process. Common applications for sentiment analysis include identifying customer satisfaction or dissatisfaction early, gauging product success or failure, and spotting new trends.

For example, an ice cream company would like to learn about which of its ice cream flavors are most liked by children. Sales data alone does not provide this information because the children that consume the ice cream are not necessarily the purchasers of the ice cream. Sentiment analysis is applied to archived customer feedback left on the ice cream company's Web site to extract information specifically regarding children's preferences for certain ice cream flavors over other flavors.

Sample questions can include:

- *How can customer reactions to the new packaging of the product be gauged?*
- *Which contestant is a likely winner of a singing contest?*
- *Can customer churn be measured by social media comments?*

A common medium by which filtering is implemented is via the use of a recommender system. Collaborative filtering is an item filtering technique based on the collaboration, or merging, of a user's past behavior with the behaviors of others. A target user's past behavior, including their likes, ratings, purchase history and more, is collaborated with the behavior of similar users. Based on the similarity of the users' behavior, items are filtered for the target user.

Collaborative filtering is solely based on the similarity between users' behavior. It requires a large amount of user behavior data in order to accurately filter items. It is an example of the application of the law of large numbers.

Content-based filtering is an item filtering technique focused on the similarity between users and items. A user profile is created based on that user's past behavior, for example, their likes, ratings and purchase history. The similarities identified between the user profile and the attributes of various items lead to items being filtered for the user. Contrary to collaborative filtering, content-based filtering is solely dedicated to individual user preferences and does not require data about other users.

A recommender system predicts user preferences and generates suggestions for the user accordingly. Suggestions commonly pertain to recommending items, such as movies, books, Web pages and people. A recommender system typically uses either collaborative filtering or content-based filtering to generate suggestions. It may also be based on a hybrid of both collaborative filtering and content-based filtering to fine-tune the accuracy and effectiveness of generated suggestions.

For example, in order to realize cross-selling opportunities, the bank builds a recommender system that uses content-based filtering. Based on matches found between financial products purchased by customers and the properties of similar financial products, the recommender system automates suggestions for potential financial products that customers may also be interested in.

Sample questions can include:

- *How can only the news articles that a user is interested in be displayed?*
- *Which holiday destinations can be recommended based on the travel history of a vacationer?*
- *Which other new users can be suggested as friends based on the current profile of a person?*

Semantic Analysis

A fragment of text or speech data can carry different meanings in different contexts, whereas a complete sentence may retain its meaning, even if structured in different ways. In order for the machines to extract valuable information, text and speech data needs to be understood by the machines in the same way as humans do. Semantic analysis represents practices for extracting meaningful information from textual and speech data.

This section describes the following types of semantic analysis:

- Natural Language Processing
- Text Analytics
- Sentiment Analysis

Natural Language Processing

Natural language processing is a computer's ability to comprehend human speech and text as naturally understood by humans. This allows computers to perform a variety of useful tasks, such as full-text searches.

For example, in order to increase the quality of customer care, the ice cream company employs natural language processing to transcribe customer calls into textual data that are then mined for commonly recurring reasons of customer dissatisfaction.

Instead of hard-coding the required learning rules, either supervised or unsupervised machine learning is applied to develop the computer's understanding of the natural language. In general, the more learning data the computer has, the more correctly it can decipher human text and speech.

Natural language processing includes both text and speech recognition. For speech recognition, the system attempts to comprehend the speech and then performs an action, such as transcribing text.

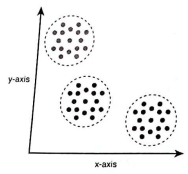
Sample questions can include:

- *How can an automated phone exchange system that can recognize the correct department extension as dictated verbally by the caller be developed?*
- *How can grammatical mistakes be automatically identified?*
- *How can a system that can correctly understand different accents of English language be designed?*

Clustering is generally used in data mining to get an understanding of the properties of a given dataset. After developing this understanding, classification can be used to make better predictions about similar but new or unseen data.

Clustering can be applied to the categorization of unknown documents and to personalized marketing campaigns by grouping together customers with similar behavior. A scatter graph provides a visual representation of clusters in Figure 8.12.

Figure 8.12
A scatter graph summarizes the results of clustering.



For example, a bank wants to introduce its existing customers to a range of new financial products based on the customer profiles it has on record. The analysts categorize customers into multiple groups using clustering. Each group is then introduced to one or more financial products most suitable to the characteristics of the overall profile of the group.

Sample questions can include:

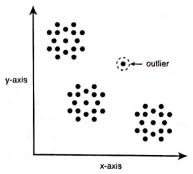
- *How many different species of trees exist based on the similarity between trees?*
- *How many groups of customers exist based upon similar purchase history?*
- *What are the different groups of viruses based on their characteristics?*

Outlier Detection

Outlier detection is the process of finding data that is significantly different from or inconsistent with the rest of the data within a given dataset. This machine learning technique is used to identify anomalies, abnormalities and deviations that can be advantageous, such as opportunities, or unfavorable, such as risks.

Outlier detection is closely related to the concept of classification and clustering, although its algorithms focus on finding abnormal values. It can be based on either supervised or unsupervised learning. Applications for outlier detection include fraud detection, medical diagnosis, network data analysis and sensor data analysis. A scatter graph visually highlights data points that are outliers, as shown in Figure 8.13.

Figure 8.13
A scatter graph highlights an outlier.



For example, in order to find out whether or not a transaction is likely to be fraudulent, the bank's IT team builds a system employing an outlier detection technique that is based on supervised learning. A set of known fraudulent transactions is first fed into the outlier detection algorithm. After training the system, unknown transactions are then fed into the outlier detection algorithm to predict if they are fraudulent or not.

Sample questions can include:

- *Is an athlete using performance enhancing drugs?*
- *Are there any wrongly identified fruits and vegetables in the training dataset used for a classification task?*
- *Is there a particular strain of virus that does not respond to medication?*

Filtering

Filtering is the automated process of finding relevant items from a pool of items. Items can be filtered either based on a user's own behavior or by matching the behavior of multiple users. Filtering is generally applied via the following two approaches:

- collaborative filtering
- content-based filtering

Regression, on the other hand, is applicable to variables that have previously been identified as dependent and independent variables and implies that there is a degree of causation between the variables. The causation may be direct or indirect.

Within Big Data, correlation can first be applied to discover if a relationship exists. Regression can then be applied to further explore the relationship and predict the values of the dependent variable, based on the known values of the independent variable.

Machine Learning

Humans are good at spotting patterns and relationships within data. Unfortunately, we cannot process large amounts of data very quickly. Machines, on the other hand, are very adept at processing large amounts of data quickly, but only if they know how.

If human knowledge can be combined with the processing speed of machines, machines will be able to process large amounts of data without requiring much human intervention. This is the basic concept of machine learning.

In this section, machine learning and its relationship to data mining are explored through coverage of the following types of machine learning techniques:

- Classification
- Clustering
- Outlier Detection
- Filtering

Classification (Supervised Machine Learning)

Classification is a supervised learning technique by which data is classified into relevant, previously learned categories. It consists of two steps:

1. The system is fed training data that is already categorized or labeled, so that it can develop an understanding of the different categories.
2. The system is fed unknown but similar data for classification and based on the understanding it developed from the training data, the algorithm will classify the unlabeled data.

A common application of this technique is for the filtering of email spam. Note that classification can be performed for two or more categories. In a simplified classification process, the machine is fed labeled data during training that builds its understanding of the classification, as shown in Figure 8.11. The machine is then fed unlabeled data, which it classifies itself.

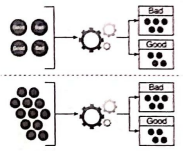


Figure 8.11
Machine learning can be used to automatically classify datasets.

For example, a bank wants to find out which of its customers is likely to default on loan payments. Based on historic data, a training dataset is compiled that contains labeled examples of customers that have or have not previously defaulted. This training data is fed to a classification algorithm that is used to develop an understanding of "good" and "bad" customers. Finally, new untagged customer data is fed in order to find out whether a given customer belongs to the defaulting category.

Sample questions can include:

- *Should an applicant's credit card application be accepted or rejected based on other accepted or rejected applications?*
- *Is a tomato a fruit or a vegetable based on the known examples of fruit and vegetables?*
- *Do the medical test results for the patient indicate a risk for a heart attack?*

Clustering (Unsupervised Machine Learning)

Clustering is an unsupervised learning technique by which data is divided into different groups so that the data in each group has similar properties. There is no prior learning of categories required. Instead, categories are implicitly generated based on the data groupings. How the data is grouped depends on the type of algorithm used. Each algorithm uses a different technique to identify clusters.

For example, managers believe that ice cream stores need to stock more ice cream for hot days, but don't know how much extra to stock. To determine if a relationship actually exists between temperature and ice cream sales, the analysts first apply correlation to the number of ice creams sold and the recorded temperature readings. A value of +0.75 suggests that there exists a strong relationship between the two. This relationship indicates that as temperature increases, more ice creams are sold.

Further sample questions addressed by correlation can include:

- *Does distance from the sea affect the temperature of a city?*
- *Do students who perform well at elementary school perform equally well at high school?*
- *To what extent is obesity linked with overeating?*

Regression

The analysis technique of regression explores how a dependent variable is related to an independent variable within a dataset. As a sample scenario, regression could help determine the type of relationship that exists between temperature, the independent variable, and crop yield, the dependent variable.

Applying this technique helps determine how the value of the dependent variable changes in relation to changes in the value of the independent variable. When the independent variable increases, for example, does the dependent variable also increase? If yes, is the increase in a linear or non-linear proportion?

For example, in order to determine how much extra stock each ice cream store needs to have, the analysts apply regression by feeding in the values of temperature readings. These values are based on the weather forecast as an independent variable and the number of ice creams sold as the dependent variable. What the analysts discover is that 15% of additional stock is required for every 5-degree increase in temperature.

More than one independent variable can be tested at the same time. However, in such cases, only one independent variable may change, while others are kept constant. Regression can help enable a better understanding of what a phenomenon is and why it occurred. It can also be used to make predictions about the values of the dependent variable.

Linear regression represents a constant rate of change, as shown in Figure 8.9.

Non-linear regression represents a variable rate of change, as shown in Figure 8.10.

Figure 8.9
Linear regression

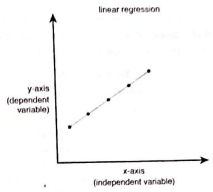
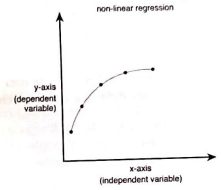


Figure 8.10
Non-linear regression



Sample questions can include:

- *What will be the temperature of a city that is 250 miles away from the sea?*
- *What will be the grades of a student studying at a high school based on their primary school grades?*
- *What are the chances that a person will be obese based on the amount of their food intake?*

Regression and correlation have a number of important differences. Correlation does not imply causation. The change in the value of one variable may not be responsible for the change in the value of the second variable, although both may change at the same rate. This can occur due to an unknown third variable, known as the confounding factor. Correlation assumes that both variables are independent.

Figure 8.5
Two different email versions are sent out simultaneously as part of a marketing campaign to see which version brings in more prospective customers.

Correlation

Correlation is an analysis technique used to determine whether two variables are related to each other. If they are found to be related, the next step is to determine what their relationship is. For example, the value of Variable A increases whenever the value of Variable B increases. We may be further interested in discovering how closely Variables A and B are related, which means we may also want to analyze the extent to which Variable B increases in relation to Variable A's increase.

The use of correlation helps to develop an understanding of a dataset and find relationships that can assist in explaining a phenomenon. Correlation is therefore commonly used for data mining where the identification of relationships between variables in a dataset leads to the discovery of patterns and anomalies. This can reveal the nature of the dataset or the cause of a phenomenon.

When two variables are considered to be correlated they are aligned based on a linear relationship. This means that when one variable changes, the other variable also changes proportionally and constantly.

Correlation is expressed as a decimal number between –1 to +1, which is known as the correlation coefficient. The degree of relationship changes from being strong to weak when moving from –1 to 0 or +1 to 0.

Figure 8.6 shows a correlation of +1, which suggests that there is a strong positive relationship between the two variables.

Figure 8.7 shows a correlation of 0, which suggests that there is no relationship at all between the two variables.

In Figure 8.8, a slope of –1 suggests that there is a strong negative relationship between the two variables.

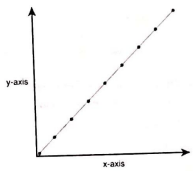


Figure 8.6
When one variable increases, the other also increases and vice versa.

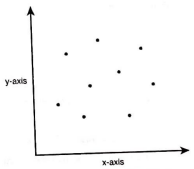


Figure 8.7
When one variable increases, the other may stay the same, or increase or decrease arbitrarily.

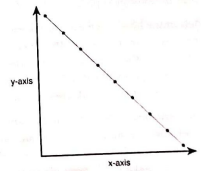


Figure 8.8
When one variable increases, the other decreases and vice versa.

Qualitative Analysis

Qualitative analysis is a data analysis technique that focuses on describing various data qualities using words. It involves analyzing a smaller sample in greater depth compared to quantitative data analysis. These analysis results cannot be generalized to an entire dataset due to the small sample size. They also cannot be measured numerically or used for numerical comparisons. For example, an analysis of ice cream sales may reveal that May's sales figures were not as high as June's. The analysis results state only that the figures were "not as high as," and do not provide a numerical difference. The output of qualitative analysis is a description of the relationship using words as shown in Figure 8.3.

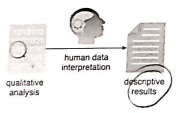


Figure 8.3
Qualitative results are descriptive in nature and not generalizable to the entire dataset.

Data Mining

Data mining, also known as data discovery, is a specialized form of data analysis that targets large datasets. In relation to Big Data analysis, data mining generally refers to automated, software-based techniques that sift through massive datasets to identify patterns and trends.

Figure 8.4
The symbol used to represent data mining.

Specifically, it involves extracting hidden or unknown patterns in the data with the intention of identifying previously unknown patterns. Data mining forms the basis for predictive analytics and business intelligence (BI). The symbol used to represent data mining is shown in Figure 8.4.

Statistical Analysis

Statistical analysis uses statistical methods based on mathematical formulas as a means for analyzing data. Statistical analysis is most often quantitative, but can also be qualitative. This type of analysis is commonly used to describe datasets via summarization, such as providing the mean, median, or mode of statistics associated with the dataset. It can also be used to infer patterns and relationships within the dataset, such as regression and correlation.

This section describes the following types of statistical analysis:

- A/B Testing
- Correlation
- Regression

A/B Testing

A/B testing, also known as split or bucket testing, compares two versions of an element to determine which version is superior based on a pre-defined metric. The element can be a range of things. For example, it can be content, such as a Web page, or an offer for a product or service, such as deals on electronic items. The current version of the element is called the control version, whereas the modified version is called the treatment. Both versions are subjected to an experiment simultaneously. The observations are recorded to determine which version is more successful.

Although A/B testing can be implemented in almost any domain, it is most often used in marketing. Generally, the objective is to gauge human behavior with the goal of increasing sales. For example, in order to determine the best possible layout for an ice cream ad on Company A's Web site, two different versions of the ad are used. Version A is an existing ad (the control) while Version B has had its layout slightly altered (the treatment). Both versions are then simultaneously shown to different users:

- Version A to Group A
- Version B to Group B

The analysis of the results reveals that Version B of the ad resulted in more sales as compared to Version A.

In other areas such as the scientific domains, the objective may simply be to observe which version works better in order to improve a process or product. Figure 8.5 provides an example of A/B testing on two different email versions sent simultaneously.

Sample questions can include:

- *Is the new version of a drug better than the old one?*
- *Do customers respond better to advertisements delivered by email or postal mail?*
- *Is the newly designed homepage of the Web site generating more user traffic?*

Big Data analysis blends traditional statistical data analysis approaches with computational ones. Statistical sampling from a population is ideal when the entire dataset is available, and this condition is typical of traditional batch processing scenarios. However, Big Data can shift batch processing to realtime processing due to the need to make sense of streaming data. With streaming data, the dataset accumulates over time, and the data is time-ordered. Streaming data places an emphasis on timely processing, for analytic results have a shelf-life. Whether it is the recognition of an upsell opportunity that presents itself due to the current context of a customer, or the detection of anomalous conditions in an industrial setting that require intervention to protect equipment or ensure product quality, time is of the essence, and freshness of the analytic result is essential.

In any fast moving field like Big Data, there are always opportunities for innovation. An example of this is the question of how to best blend statistical and computational approaches for a given analytical problem. Statistical techniques are commonly preferred for exploratory data analysis, after which computational techniques that leverage the insight gleaned from the statistical study of a dataset can be applied. The shift from batch to realtime presents other challenges as realtime techniques need to leverage computationally-efficient algorithms.

In 2003, William Agresti recognized the shift toward computational approaches and argued for the creation of a new computational discipline named Discovery Informatics. Agresti's view of this field was one that embraced composition. In other words, he believed that discovery informatics was a synthesis of the following fields: pattern recognition (data mining), artificial intelligence (machine learning), document and text processing (semantic processing), database management and information storage and retrieval. Agresti's insight into the importance and breadth of computational approaches to data analysis was forward-thinking at the time, and his perspective on the matter has only been reinforced by the passage of time and the emergence of data science as a discipline.

One challenge concerns the best way of balancing the accuracy of an analytic result against the run-time of the algorithm. In many cases, an approximation may be sufficient and affordable. From a storage perspective, multi-tiered storage solutions which leverage RAM, solid-state drives and hard-disk drives will provide near-term flexibility and realtime analytic capability with long-term, cost-effective persistent storage. In the long run, an organization will operate its Big Data analysis engine at two speeds: processing streaming data as it arrives and performing batch analysis of this data as it accumulates to look for patterns and trends. (The symbol used to represent data analysis is shown in Figure 8.1.)

Figure 8.1
The symbol used to represent data analysis.

This chapter begins with descriptions of the following basic types of data analysis:

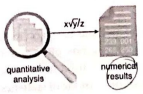
- quantitative analysis
- qualitative analysis
- data mining
- statistical analysis
- machine learning
- semantic analysis
- visual analysis

Quantitative Analysis

Quantitative analysis is a data analysis technique that focuses on quantifying the patterns and correlations found in the data. Based on statistical practices, this technique involves analyzing a large number of observations from a dataset. Since the sample size is large, the results can be applied in a generalized manner to the entire dataset. Figure 8.2 depicts the fact that quantitative analysis produces numerical results.

Quantitative analysis results are absolute in nature and can therefore be used for numerical comparisons. For example, a quantitative analysis of ice cream sales may discover that a 5 degree increase in temperature increases ice cream sales by 15%.

Figure 8.2
The output of quantitative analysis is numerical in nature.



CASE STUDY EXAMPLE

ETI's committee of senior managers investigated the company's deteriorating financial position and realized that many of the corporation's current problems could have been detected earlier. If the management at the tactical level had greater awareness, they could have proactively taken action to avoid some of the losses. This lack of early warning was due to the fact that ETI failed to sense that marketplace dynamics had changed. New competitors using advanced technologies to process claims and set premiums had disrupted the market and taken a share of ETI's business. At the same time, the company's lack of sophisticated fraud detection has been exploited by unscrupulous customers and perhaps even organized crime.

The senior management team reports their findings to the executive management team. Subsequently, in light of the previous strategic goals that were established, a new set of transformation and innovation corporate priorities are established. These initiatives will be used to direct and guide corporate resources to solutions that will enhance ETI's ability to increase profits.

Considering transformation, business process management disciplines will be adopted to document, analyze and improve the processing of claims. These business process models will then be consumed by a Business Process Management System (BPMS), which is essentially a process automation framework, to ensure consistent and auditable process execution. This will help ETI demonstrate regulatory compliance. An additional benefit of using a BPMS is that the traceability of claims processed by the system includes information about which employees have processed which claim. Although it has not been confirmed, there is a suspicion that some portion of the fraudulent claims being processed may be traceable to employees that are subverting internal manual controls driven by corporate policy. In other words, not only will the BPMS enhance the ability to meet external regulatory compliance, it will also enforce standard operating procedures and work practices within ETI.

Hyper-Connected Communities and Devices

The broadening coverage of the Internet and the proliferation of cellular and Wi-Fi networks has enabled more people and their devices to be continuously active in virtual communities. Coupled with the proliferation of Internet connected sensors, the underpinnings of the Internet of Things (IoT), a vast collection of smart Internet-connected devices, is being formed. As shown in Figure 2.6, this in turn has resulted in a massive increase in the number of available data streams. While some streams are public, other streams are channeled directly to corporations for analysis. As an example, the performance-based management contracts associated with heavy equipment used in the mining industry incentivize the optimal performance of preventive and predictive maintenance in an effort to reduce the need and avoid the downtime associated with unplanned corrective maintenance. This requires detailed analysis of sensor readings emitted by the equipment for the early detection of issues that can be resolved via the proactive scheduling of maintenance activities.

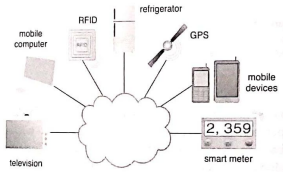


Figure 2.6
Hyper-connected communities and devices include television, mobile computing, RFIDs, refrigerators, GPS devices, mobile devices and smart meters.

Cloud Computing

Cloud computing advancements have led to the creation of environments that are capable of providing highly scalable, on-demand IT resources that can be leased via pay-as-you-go models. Businesses have the opportunity to leverage the infrastructure, storage and processing capabilities provided by these environments in order to build-out scalable Big Data solutions that can carry out large-scale processing tasks. Although traditionally thought of as off-premise environments typically depicted with a cloud

symbol, businesses are also leveraging cloud management software to create on premise clouds to more effectively utilize their existing infrastructure via virtualization. In either case, the ability of a cloud to dynamically scale based upon load allows for the creation of resilient analytic environments that maximize efficient utilization of ICT resources.

Figure 2.7 displays an example of how a cloud environment can be leveraged for its scaling capabilities to perform Big Data processing tasks. The fact that off-premise cloud-based IT resources can be leased dramatically reduces the required up-front investment of Big Data projects.

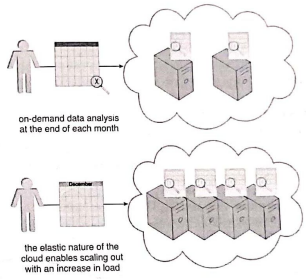


Figure 2.7
The cloud can be used to complete on-demand data analysis at the end of each month or enable the scaling out of systems with an increase in load.

It makes sense for enterprises already using cloud computing to reuse the cloud for their Big Data initiatives because:

- personnel already possesses the required cloud computing skills
- the input data already exists in the cloud

Digitization

For many businesses, digital mediums have replaced physical mediums as the de facto communications and delivery mechanism. The use of digital artifacts saves both time and cost as distribution is supported by the vast pre-existing infrastructure of the Internet. As consumers connect to a business through their interaction with these digital substitutes, it leads to an opportunity to collect further "secondary" data; for example, requesting a customer to provide feedback, complete a survey, or simply providing a hook to display a relevant advertisement and tracking its click-through rate. Collecting secondary data can be important for businesses because mining this data can allow for customized marketing, automated recommendations and the development of optimized product features. Figure 2.4 provides a visual representation of examples of digitization.

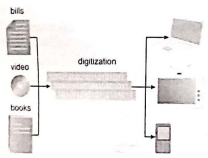


Figure 2.4
Examples of digitization include online banking, on-demand television and streaming video.

Affordable Technology and Commodity Hardware

Technology capable of storing and processing large quantities of diverse data has become increasingly affordable. Additionally, Big Data solutions often leverage open-source software that executes on commodity hardware, further reducing costs. The combination of commodity hardware and open source software has virtually eliminated the advantage that large enterprises used to hold by being able to outspend their smaller competitors due to the larger size of their IT budgets. Technology no longer delivers competitive advantage. Instead, it simply becomes the platform upon which the business executes. From a business standpoint, utilization of affordable technology and commodity hardware to generate analytic results that can further optimize the execution of its business processes is the path to competitive advantage.

The use of commodity hardware makes the adoption of Big Data solutions accessible to businesses without large capital investments. Figure 2.5 provides an example of the price decline associated with data storage prices over the past 20 years.

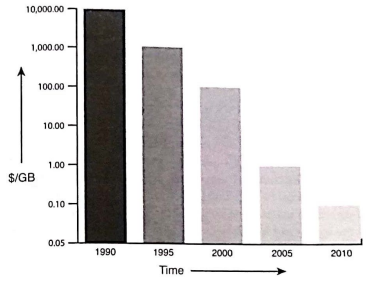


Figure 2.5
Data storage prices have dropped dramatically from more than $10,000 to less than $0.10 per GB over the decades.

Social Media

The emergence of social media has empowered customers to provide feedback in near-realtime via open and public mediums. This shift has forced businesses to consider customer feedback on their service and product offerings in their strategic planning. As a result, businesses are storing increasing amounts of data on customer interactions within their customer relationship management systems (CRM) and from harvesting customer reviews, complaints and praise from social media sites. This information feeds Big Data analysis algorithms that surface the voice of the customer in an attempt to provide better levels of service, increase sales, enable targeted marketing and even create new products and services. Businesses have realized that branding activity is no longer completely managed by internal marketing activities. Instead, product brands and corporate reputation are co-created by the company and its customers. For this reason, businesses are increasingly interested in incorporating publicly available datasets from social media and other external data sources.

Business Process Management

Businesses deliver value to customers and other stakeholders via the execution of their business processes. A business process is a description of how work is performed in an organization. It describes all work-related activities and their relationships, aligned with the organizational actors and resources responsible for conducting them. The relationships between activities may be temporal; for example, activity A is executed before activity B. The relationships can also describe whether the execution of activities is conditional, based upon the outputs or conditions generated by other activities or by sensing events generated outside of the business process itself.

Business process management applies process excellence techniques to improve corporate execution. Business Process Management Systems (BPMS) provide software developers a model driven platform that is becoming the Business Application Development Environment (BADE) of choice. A business application needs to: mediate between humans and other technology-hosted resources, execute in alignment with corporate policies and ensure the fair distribution of work to employees. As a BADE, models of a business process are joined with: models of organizational roles and structure, business entities and their relationships, business rules and the user-interface. The development environment integrates these models together to create a business application that manages screenflow and workflow and provides workload management. This is accomplished in an execution environment that enforces corporate policy and security and provides state management for long-running business processes. The state of an individual process, or all processes, can be interrogated via Business Activity Monitoring (BAM) and visualized.

When BPM is combined with BPMSs that are intelligent, processes can be executed in a goal-driven manner. Goals are connected to process fragments that are dynamically chosen and assembled at run-time in alignment with the evaluation of the goals. When the combination of Big Data analytic results and goal-driven behavior are used together, process execution can become adaptive to the marketplace and responsive to environmental conditions. As a simple example, a customer contact process has process fragments that enable communication with customers via a voice call, email, text message and traditional postal mail. In the beginning, the choice of these contact methods is unweighted, and they are chosen at random. However, behind-the-scenes analysis is being done to measure the effectiveness of the contact method via statistical analysis of customer responsiveness.

The results of this analysis are tied to a goal responsible for selecting the contact method, and when a clear preference is determined, the weighting is changed to favor the contact method that achieves the best response. A more detailed analysis could leverage customer clustering, which would assign individual customers to groups where one of the cluster dimensions is the contact method. In this case, customers can be contacted with even greater refinement, which provides a pathway to one-to-one targeted marketing.

Information and Communications Technology

This section examines the following ICT developments that have accelerated the pace of Big Data adoption in businesses:

- data analytics and data science
- digitization
- affordable technology and commodity hardware
- social media
- hyper-connected communities and devices
- cloud computing

Data Analytics and Data Science

Enterprises are collecting, procuring, storing, curating and processing increasing quantities of data. This is occurring in an effort to find new insights that can drive more efficient and effective operations, provide management the ability to steer the business proactively and allow the C-suite to better formulate and assess their strategic initiatives. Ultimately, enterprises are looking for new ways to gain a competitive edge. Thus the need for techniques and technologies that can extract meaningful information and insights has increased. Computational approaches, statistical techniques and data warehousing have advanced to the point where they have merged, each bringing their specific techniques and tools that allow the performance of Big Data analysis. The maturity of these fields of practice inspired and enabled much of the core functionality expected from contemporary Big Data solutions, environments and platforms.

information. At the managerial level, this information can be examined through the lens of corporate performance to answer questions regarding *how* the business is performing. In other words, give meaning to the information. This information may be further enriched to answer questions regarding *why* the business is performing at the level it is. When armed with this knowledge, the strategic layer can provide further insight to help answer questions of which strategy needs to change or be adopted in order to correct or enhance the performance.

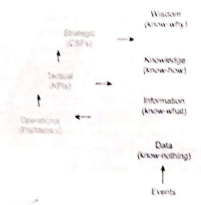


Figure 2.2
The DIKW pyramid illustrates alignment with Strategic, Tactical and Operational corporate levels.

As with any layered system, the layers do not all change at the same speed. In the case of a business enterprise, the strategic layer is the slowest moving layer, and the operational layer is the fastest moving layer. The slower moving layers provide stability and direction to the faster moving layers. In traditional organizational hierarchies, the management layer is responsible for directing the operational layer in alignment with the strategy created by the executive team. Because of this variation in regard to speed of change, it is possible to envision the three layers as being responsible for strategy execution, business execution and process execution respectively. Each of these layers relies upon different metrics and measures, presented through different visualization and reporting functions. For example, the strategy layer may rely upon balanced scorecards,

the management layer upon an interactive visualization of KPIs and corporate performance and the operational layer on visualizations of executing business processes and their statuses.

Figure 2.3, a variant of a diagram produced by Joe Gollner in his blog post "The Anatomy of Knowledge," shows how an organization can relate and align its organizational layers by creating a virtuous cycle via a feedback loop. On the right side of the figure, the strategic layer drives response via the application of judgment by making decisions regarding corporate strategy, policy, goals and objectives that are communicated as constraints to the tactical layer. The tactical layer in turn leverages this knowledge to generate priorities and actions that conform to corporate direction. These actions adjust the execution of business at the operational layer. This in turn should generate measureable change in the experience of internal stakeholders and external customers as they deliver and consume business services. This change, or result, should surface and be visible in the data in the form of changed PIs that are then aggregated into KPIs. Recall that KPIs are metrics that can be associated with critical success factors that inform the executive team as to whether or not their strategies are working. Over time, the strategic and management layers injection of judgment and action into the loop will serve to refine the delivery of business services.

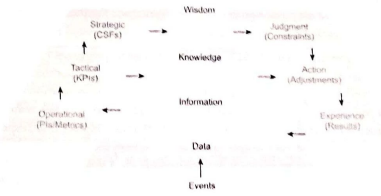


Figure 2.3
The creation of a virtuous cycle to align an organization across layers via a feedback loop

brings additional context to their internal data allows a corporation to move up the analytic value chain from hindsight to insight with greater ease. With appropriate tooling, which often supports sophisticated simulation capabilities, a company can develop analytic results that provide foresight. In this case, the tooling assists in bridging the gap between knowledge and wisdom as well as provides advisory analytic results. This is the power of Big Data—enriching corporate perspective beyond introspection, from which a business can only infer information about marketplace sentiment, to sensing the marketplace itself.

The transition from hindsight to foresight can be understood through the lens of the DIKW pyramid depicted in Figure 2.1. Note that in this figure, at the top of the triangle, wisdom is shown as an outline to indicate that it exists but is not typically generated via ICT systems. Instead, knowledge workers provide the insight and experience to frame the available knowledge so that it can be integrated to form wisdom. Wisdom generation by technological means quickly devolves into a philosophical discussion that is not within the scope of this book. Within business environments, technology is used to support knowledge management, and personnel are responsible for applying their competency and wisdom to act accordingly.

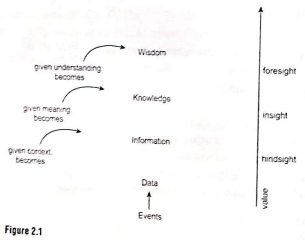


Figure 2.1
The DIKW pyramid shows how data can be enriched with context to create information, information can be supplied with meaning to create knowledge and knowledge can be integrated to form wisdom.

Business Architecture

Within the past decade, there has been a realization that too often a corporation's enterprise architecture is simply a myopic view of its technology architecture. In an effort to wrest power from the stronghold of IT, business architecture has emerged as a complementary discipline. In the future, the goal is that enterprise architecture will present a balanced view between business and technology architectures. Business architecture provides a means of blueprinting or concretely expressing the design of the business. A business architecture helps an organization align its strategic vision with its underlying execution, whether they be technical resources or human capital. Thus, a business architecture includes linkages from abstract concepts like business mission, vision, strategy and goals to more concrete ones like business services, organizational structure, key performance indicators and application services.

These linkages are important because they provide guidance as to how to align the business and its information technology. It is an accepted view that a business operates as a layered system—the top layer is the strategic layer occupied by C-level executives and advisory groups; the middle layer is the tactical or managerial layer that seeks to steer the organization in alignment with the strategy; and the bottom layer is the operations layer where a business executes its core processes and delivers value to its customers. These three layers often exhibit a degree of independence from one another, but each layer's goals and objectives are influenced by and often defined by the layer above, in other words top-down. From a monitoring perspective, communication flows upstream, or bottom-up via the collection of metrics. Business activity monitoring at the operations layer generates Performance Indicators (PIs) and metrics, for both services and processes. They are aggregated to create Key Performance Indicators (KPIs) used at the tactical layer. These KPIs can be aligned with Critical Success Factors (CSFs) at the strategic layer, which in turn help measure progress being made toward the achievement of strategic goals and objectives.

Big Data has ties to business architecture at each of the organizational layers, as depicted in Figure 2.2. Big Data enhances value as it provides additional context through the integration of external perspectives to help convert data into information and provide meaning to generate knowledge from information. For instance, at the operational level, metrics are generated that simply report on *what* is happening in the business. In essence, we are converting data through business concepts and context to generate

In many organizations it is now acceptable for a business to be architected in much the same way as its technology. This shift in perspective is reflected in the expanding domain of enterprise architecture, which used to be closely aligned with technology architecture but now includes business architecture as well. Although businesses still view themselves from a mechanistic system's point of view, with command and control being passed from executives to managers to front-line employees, feedback loops based upon linked and aligned measurements are providing greater insight into the effectiveness of management decision-making.

This cycle from decision to action to measurement and assessment of results creates opportunities for businesses to optimize their operations continuously. In fact, the mechanistic management view is being supplanted by one that is more organic and that drives the business based upon its ability to convert data into knowledge and insight. One problem with this perspective is that, traditionally, businesses were driven almost exclusively by internal data held in their information systems. However, companies are learning that this is not sufficient in order to execute their business models in a marketplace that more resembles an ecological system. As such, organizations need to consume data from the outside to sense directly the factors that influence their profitability. The use of such external data most often results in "Big Data" datasets.

This chapter explores the business motivations and drivers behind the adoption of Big Data solutions and technologies. The adoption of Big Data represents the confluence of several forces to include: marketplace dynamics, an appreciation and formalism of Business Architecture (BA), the realization that a business' ability to deliver value is directly tied to Business Process Management (BPM), innovation in Information and Communications Technology (ICT) and finally the Internet of Everything (IoE). Each of these topics will be explored in turn.

Marketplace Dynamics

There has been a fundamental shift in the way businesses view themselves and the marketplace. In the past 15 years, two large stock market corrections have taken place—the first was the dot-com bubble burst in 2000, and the second was the global recession that began in 2008. In each case, businesses entrenched and worked to improve

their efficiency and effectiveness to stabilize their profitability by reducing costs. This of course is normal. When customers are scarce, cost-cutting often ensues to maintain the corporate bottom line. In this environment, companies conduct transformation projects to improve their corporate processes to achieve savings.

As the global economies began to emerge from recession, companies began to focus outward, looking to find new customers and keep existing customers from defecting to marketplace competitors. This was accomplished by offering new products and services and delivering increased value propositions to customers. It is a very different market cycle to the one that focuses on cost-cutting, for it is not about transformation but instead innovation. Innovation brings hope to a company that it will find new ways to achieve a competitive advantage in the marketplace and a consequent increase in top line revenue.

The global economy can experience periods of uncertainty due to various factors. We generally accept that the economies of the major developed countries in the world are now inextricably intertwined; in other words, they form a system of systems. Likewise, the world's businesses are shifting their perspective about their identity and independence as they recognize that they are also intertwined in intricate product and service networks.

For this reason, companies need to expand their Business Intelligence activities beyond retrospective reflection on internal information extracted from their corporate information systems. They need to open themselves to external data sources as a means of sensing the marketplace and their position within it. Recognizing that external data

> Davenport and Prusak have provided generally-accepted working definitions of data, information and knowledge in their book *Working Knowledge*. According to Davenport and Prusak, "[d]ata is a set of discrete, objective facts about events." In a business sense, these events are activities that occur within an organization's business processes and information systems—they represent the generation, modification and completion of work associated with business entities, for example, orders, shipments, notifications and customer address updates. These events are a reflection of real-world activity that is represented within the relational data stores of corporate information systems. Davenport and Prusak further define information as "data that makes a difference." It is data that has been contextualized to provide communication; it delivers a message and informs the receiver—whether it be a human or system. Information is then enriched via experience and insight in the generation of knowledge. The authors state that "[k]nowledge is a fluid mix of framed experience, values, contextual information and expert insight that provides a framework for evaluating and incorporating new experiences and information."

tions and decide on the premium amount. The claim adjusters deal with investigating claims made against a policy and arrive at a settlement amount for the policyholder.

Some of the key departments within ETI include the underwriting, claims settlement, customer care, legal, marketing, human resource, accounts and IT departments. Both prospective and existing customers generally contact ETI's customer care department via telephone, although contact via email and social media has increased exponentially over the past few years.

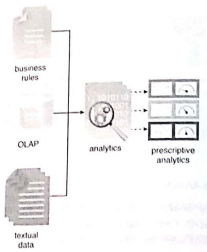
ETI strives to distinguish itself by providing competitive policies and premium customer service that does not end once a policy has been sold. Its management believes that doing so helps to achieve increased levels of customer acquisition and retention. ETI relies heavily on its actuaries to create insurance plans that reflect the needs of its customers.

Technical Infrastructure and Automation Environment

ETI's IT environment consists of a combination of client-server and mainframe platforms that support the execution of a number of systems, including policy quotation, policy administration, claims management, risk assessment, document management, billing, enterprise resource planning (ERP) and customer relationship management (CRM).

The policy quotation system is used to create new insurance plans and to provide quotes to prospective customers. It is integrated with the website and customer care portal to provide website visitors and customer care agents the ability to obtain insurance quotes. The policy administration system handles all aspects of policy lifecycle management,

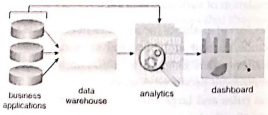
Figure 1.8
Prescriptive analytics involves the use of business rules and internal and/or external data to perform an in-depth analysis.



Business Intelligence (BI)

BI enables an organization to gain insight into the performance of an enterprise by analyzing data generated by its business processes and information systems. The results of the analysis can be used by management to steer the business in an effort to correct detected issues or otherwise enhance organizational performance. BI applies analytics to large amounts of data across the enterprise, which has typically been consolidated into an enterprise data warehouse to run analytical queries. As shown in Figure 1.9, the output of BI can be surfaced to a dashboard that allows managers to access and analyze the results and potentially refine the analytic queries to further explore the data.

Figure 1.9
BI can be used to improve business applications, consolidate data in data warehouses and analyze queries via a dashboard.

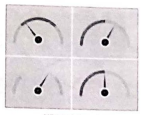


Key Performance Indicators (KPI)

A KPI is a metric that can be used to gauge success within a particular business context. KPIs are linked with an enterprise's overall strategic goals and objectives. They

are often used to identify business performance problems and demonstrate regulatory compliance. KPIs therefore act as quantifiable reference points for measuring a specific aspect of a business' overall performance. KPIs are often displayed via a KPI dashboard, as shown in Figure 1.10. The dashboard consolidates the display of multiple KPIs and compares the actual measurements with threshold values that define the acceptable value range of the KPI.

Figure 1.10
A KPI dashboard acts as a central reference point for gauging business performance.



Big Data Characteristics

For a dataset to be considered Big Data, it must possess one or more characteristics that require accommodation in the solution design and architecture of the analytic environment. Most of these data characteristics were initially identified by Doug Laney in early 2001 when he published an article describing the impact of the volume, velocity and variety of e-commerce data on enterprise data warehouses. To this list, veracity has been added to account for the lower signal-to-noise ratio of unstructured data as compared to structured data sources. Ultimately, the goal is to conduct analysis of the data in such a manner that high-quality results are delivered in a timely manner, which provides optimal value to the enterprise.

This section explores the five Big Data characteristics that can be used to help differentiate data categorized as "Big" from other forms of data. The five Big Data traits shown in Figure 1.11 are commonly referred to as the Five Vs:

- volume
- velocity
- variety
- veracity
- value

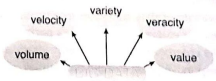
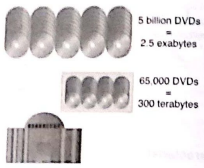


Figure 1.11
The Five Vs of Big Data.

Volume

The anticipated volume of data that is processed by Big Data solutions is substantial and ever-growing. High data volumes impose distinct data storage and processing demands, as well as additional data preparation, curation and management processes. Figure 1.12 provides a visual representation of the large volume of data being created daily by organizations and users world-wide.

Figure 1.12
Organizations and users world-wide create over 2.5 EBs of data a day. As a point of comparison, the Library of Congress currently holds more than 300 TBs of data.



Typical data sources that are responsible for generating high data volumes can include:

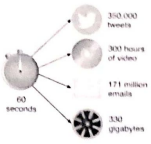
- online transactions, such as point-of-sale and banking
- scientific and research experiments, such as the Large Hadron Collider and Atacama Large Millimeter/Submillimeter Array telescope
- sensors, such as GPS sensors, RFIDs, smart meters and telematics
- social media, such as Facebook and Twitter

Velocity

In Big Data environments, data can arrive at fast speeds, and enormous datasets can accumulate within very short periods of time. From an enterprise's point of view, the velocity of data translates into the amount of time it takes for the data to be processed once it enters the enterprise's perimeter. Coping with the fast inflow of data requires the enterprise to design highly elastic and available data processing solutions and corresponding data storage capabilities.

Depending on the data source, velocity may not always be high. For example, MRI scan images are not generated as frequently as log entries from a high-traffic webserver. As illustrated in Figure 1.13, data velocity is put into perspective when considering that the following data volume can easily be generated in a given minute: 350,000 tweets, 300 hours of video footage uploaded to YouTube, 171 million emails and 330 GBs of sensor data from a jet engine.

Figure 1.13
Examples of high-velocity Big Data datasets produced every minute include tweets, video, emails and GBs generated from a jet engine.



Variety

Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions. Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage. Figure 1.14 provides a visual representation of data variety, which includes structured data in the form of financial transactions, semi-structured data in the form of emails and unstructured data in the form of images.

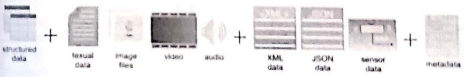


Figure 1.14
Examples of high-variety Big Data datasets include structured, textual, image, video, audio, XML, JSON, sensor data and metadata.

Veracity

Veracity refers to the quality or fidelity of data. Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise. In relation to veracity, data can be part of the signal or noise of a dataset. Noise is data that cannot be converted into information and thus has no value, whereas signals have value and lead to meaningful information. Data with a high signal-to-noise ratio has more veracity than data with a lower ratio. Data that is acquired in a controlled manner, for example via online customer registrations, usually contains less noise than data acquired via uncontrolled sources, such as blog postings. Thus the signal-to-noise ratio of data is dependent upon the source of the data and its type.

Value

Value is defined as the usefulness of data for an enterprise. The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business. Value is also dependent on how long data processing takes because analytics results have a shelf-life; for example, a 20 minute delayed stock quote has little to no value for making a trade compared to a quote that is 20 milliseconds old. As demonstrated, value and time are inversely related. The longer it takes for data to be turned into meaningful information, the less value it has for a business. Stale results inhibit the quality and speed of informed decision-making. Figure 1.15 provides two illustrations of how value is impacted by the veracity of data and the timeliness of generated analytic results.

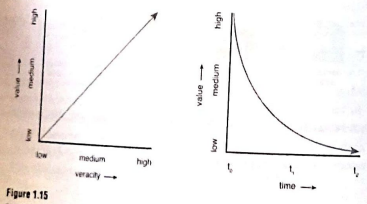


Figure 1.15
Data that has high veracity and can be analyzed quickly has more value to a business.

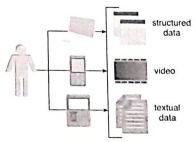
Apart from veracity and time, value is also impacted by the following lifecycle-related concerns:

- How well has the data been stored?
- Were valuable attributes of the data removed during data cleansing?
- Are the right types of questions being asked during data analysis?
- Are the results of the analysis being accurately communicated to the appropriate decision-makers?

Different Types of Data

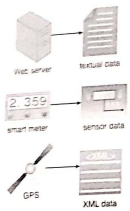
The data processed by Big Data solutions can be human-generated or machine-generated, although it is ultimately the responsibility of machines to generate the analytic results. Human-generated data is the result of human interaction with systems, such as online services and digital devices. Figure 1.16 shows examples of human-generated data.

Figure 1.16
Examples of human-generated data include social media, blog posts, emails, photo sharing and messaging.



Machine-generated data is generated by software programs and hardware devices in response to real-world events. For example, a log file captures an authorization decision made by a security service, and a point-of-sale system generates a transaction against inventory to reflect items purchased by a customer. From a hardware perspective, an example of machine-generated data would be information conveyed from the numerous sensors in a cellphone that may be reporting information, including position and cell tower signal strength. Figure 1.17 provides a visual representation of different types of machine-generated data.

Figure 1.17
Examples of machine-generated data include web logs, sensor data, telemetry data, smart meter data and appliance usage data.



As demonstrated, human-generated and machine-generated data can come from a variety of sources and be represented in various formats or types. This section examines the variety of data types that are processed by Big Data solutions. The primary types of data are:

- structured data
- unstructured data
- semi-structured data

These data types refer to the internal organization of data and are sometimes called data formats. Apart from these three fundamental data types, another important type of data in Big Data environments is metadata. Each will be explored in turn.

Structured Data

Structured data conforms to a data model or schema and is often stored in tabular form. It is used to capture relationships between different entities and is therefore most often stored in a relational database. Structured data is frequently generated by enterprise applications and information systems like ERP and CRM systems. Due to the abundance of tools and databases that natively support structured data, it rarely requires special consideration in regards to processing or storage. Examples of this type of data include banking transactions, invoices, and customer records. Figure 1.18 shows the symbol used to represent structured data.

Figure 1.18
The symbol used to represent structured data stored in a tabular form.

Unstructured Data

Data that does not conform to a data model or data schema is known as unstructured data. It is estimated that unstructured data makes up 80% of the data within any given enterprise. Unstructured data has a faster growth rate than structured data. Figure 1.19 illustrates some common types of unstructured data. This form of data is either textual or binary and often conveyed via files that are self-contained and non-relational. A text file may contain the contents of various tweets or blog postings. Binary files are often media files that contain image, audio or video data. Technically, both text and binary files have a structure defined by the file format itself, but this aspect is disregarded, and the notion of being unstructured is in relation to the format of the data contained in the file itself.

Figure 1.19
Video, image and audio files are all types of unstructured data.

Special purpose logic is usually required to process and store unstructured data. For example, to play a video file, it is essential that the correct codec (coder-decoder) is available. Unstructured data cannot be directly processed or queried using SQL. If it is required to be stored within a relational database, it is stored in a table as a Binary Large Object (BLOB). Alternatively, a Not-only SQL (NoSQL) database is a non-relational database that can be used to store unstructured data alongside structured data.

Semi-structured Data

Semi-structured data has a defined level of structure and consistency, but is not relational in nature. Instead, semi-structured data is hierarchical or graph-based. This kind of data is commonly stored in files that contain text. For instance, Figure 1.20 shows that XML and JSON files are common forms of semi-structured data. Due to the textual nature of this data and its conformance to some level of structure, it is more easily processed than unstructured data.

Examples of common sources of semi-structured data include electronic data interchange (EDI) files, spreadsheets, RSS feeds and sensor data. Semi-structured data often has special pre-processing and storage requirements, especially if the underlying