StatSoft®

data analysis • data mining • quality improvement • web-based analytics

Business White Paper

The Big Data Revolution
And How to Extract Value from Big Data

Dr. Thomas Hill

Headquarters: StatSoft, Inc. • 2300 E 14 St • Tulsa, OK 74104 • USA • +1 (918) 749-1119 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America Ltda.
Bulgaria: StatSoft Bulgaria Ltd.
Chile: StatSoft South America Ltda.
China: StatSoft China
Czech Rep.: StatSoft Czech Rep. s.r.o.
Egypt: StatSoft Middle East
France: StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Netherlands: StatSoft Benelux
Norway: StatSoft Norway AS
Poland: StatSoft Polska Sp. z o.o.
Portugal: StatSoft Iberica Lda
Russia: StatSoft Russia
S. Africa: StatSoft S. Africa (Pty) Ltd.
Spain: StatSoft Iberica Lda
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
United Kingdom: StatSoft Ltd.
USA: StatSoft, Inc.

The Big Data Revolution
And How to Extract Value from Big Data

Overview

"Big data" is the buzzword that is currently dominating professional conferences around data science, predictive modeling, data mining, and CRM, to name only a few of the domains that have become electrified by the prospect of incorporating qualitatively larger data sizes and more voluminous high velocity data streams into business or other organizational processes. As is usually the case when new technologies begin to transform industries, the technologies also introduce new terminology and, indeed, new ways of "thinking about" or conceptualizing reality and approaches to solve problems or improve processes.

For example, while only a few years ago it was only possible and conceivable to "segment" customers into groups most likely to purchase specific items or services, it is now possible and common to build models for each customer in real time as s/he peruses the internet searching for a specific household item or electronic gadget: instantly, that interest can be analyzed and translated into relevant display advertisements and offers to specific prospects providing a degree of customization that was inconceivable only a few years ago. As technologies to record the physical location of cell phones and their owners has matured, it seems that it won't be long now until the vision depicted in the 2002 sci-fi thriller *Minority Report*, where display advertisements in malls are tailored to the specific individuals passing by, will become reality.

At the same time, there are domains and situations where, inevitably, the excitement about new technologies around big data will give way to great disappointment. Sometimes sparse data describing precisely a critical piece of reality (critical for a business' success) is much more valuable than big data describing non-critical pieces of that reality.

The purpose of this paper is to clarify and reflect on some of the exciting new opportunities around big data, and illustrate how StatSoft's *STATISTICA* analytic platform(s) can help leverage big data to optimize a process, solve problems, or gain "big insights."

How Big is Big Data

Of course, the right answer must be "it depends." In fact, in practice and in many public discussions around big data, the term is used to describe extremely large data, in the multiple gigabyte or terabyte range.

As a practical matter, such data can easily be stored and managed in "traditional" databases and with standard hardware (database servers). *STATISTICA* software is multithreaded in all critical data access (reading), transformation, and predictive modeling (and scoring) algorithms, so such (indeed very large) data sets can easily be analyzed without requiring specialized, new tools.

Large Data, Huge Data

For perspective, some of the large international banks StatSoft is working with manage somewhere around 9 to 12 million accounts. Multiplied by perhaps 1,000 parameters or characteristics (variables) collected and organized into a data warehouse for risk and other predictive modeling activities, such a file is still "only" around 100 gigabytes; certainly not a small data repository, but not of a size exceeding the capabilities of standard database technologies. Nor will such data sizes challenge *STATISTICA's* capabilities even on models/older hardware.

In practice, a large number of applications driving decision making in health care, banking and financial services, insurance, manufacturing, etc., are working with relatively well organized databases of customer data, machine data, etc. In many cases, the sizes of those databases, and the speed at which they must be analyzed to support mission critical business activities, can be challenging. However, the *STATISTICA* batch analysis and scoring solutions (*STATISTICA Enterprise*) and real-time solutions (*STATISTICA Live Score*), and the analytic tools for creating and managing models (*STATISTICA Data Miner, Decisioning* solutions), easily scale up to large numbers of servers with multiple processing cores. In practice, that means that the speed at which analytic support (e.g., predictions regarding credit risk, fraud probability, reliability of manufactured parts, etc.) can be made available to support time-critical decisions can almost always be achieved with off-the-shelf *STATISTICA* tools.

Large Data, Big Data

Large data sets: 1,000 megabytes = 1 gigabyte to hundreds of gigabytes

Huge data sets: 1,000 gigabytes = 1 terabyte to multiple terabytes

Big Data: Multiple terabytes to hundreds of terabytes

Extremely Big Data: 1,000 to 10,000 terabytes = 1 to 10 petabytes

From Huge Data to Big Data

Typically, the discussion of big data centers around data repositories (and the analyses based on such repositories) that are much larger than just a few terabytes. Specifically, some data repositories may grow to thousands of terabytes, i.e., to the petabyte range (1,000 terabytes = 1 petabyte). Beyond petabytes, data storage can be measured in exabytes; for example, the manufacturing sector worldwide in 2010 is estimated to have stored a total of 2 exabytes of new information (Manyika et al., 2011).

There are applications where data accumulate and are collected at a very high rate. For example, in automated or process manufacturing applications, such as power generation, continuous data streams are generated for sometimes tens of thousands of parameters every minute or every second. Likewise, over the past few years, so-called "smart-grid" technology is being implemented, allowing public utilities to measure electricity consumption at individual

households minute by minute or second by second. For such applications, when the data must be stored for years, extremely large big data will accumulate (Hopkins and Evelson, 2011).

There are increasing numbers of applications across commercial and government sectors where both the amount of data and the speed or velocity at which it accumulates result in data storage and data analysis requirements that can create hundreds of terabytes or petabytes of data. Modern technology has made it possible to track individuals and their behavior in many ways, e.g., as we review the Internet, purchase items from Internet retailers or large chain stores such as Walmart (according to Wikipedia, Walmart is estimated to manage a data repository of more than 2 petabytes), or move around with our cellphones turned on and leave a trail of where we have been and where we are going. The various methods of communication, from simple phone calls to shared information via social network sites such as Facebook (30 billion pieces of information shared every month, according to Wikipedia), or movie sharing sites such as YouTube (Youtube claims that it uploads 24 hours of movies every minute; see Wikipedia), generate massive amounts of new data daily. Likewise, modern health care technologies generate massive amounts of data related to the delivery of health care (images, films, real-time monitoring) as well as to the reimbursement and payment of health care providers.

Technical Challenges with Big Data

There are fundamentally three types of challenges with big data:

1. The storage and management of big data in the hundreds of terabyte or petabyte range, which exceeds what can (easily) be stored and managed through traditional relational databases
2. The management of unstructured data (which usually makes up the majority of all data in big-data scenarios), i.e., how to organize text, movies, pictures, etc.
3. The analysis of big data, both for simple reporting and advanced predictive modeling, as well as deployment

Storage of Big Data

Big data is usually stored and organized in distributed file systems. While there are multiple implementations and implementation details, in the most general terms, information is stored on one of multiple (sometimes thousands of) hard drives and standard off-the-shelf computers. An index or map keeps track of where (on which computer/drive) a specific piece of information is stored. Actually, for failover redundancy and robustness, each piece of information is usually stored multiple times, e.g., as triplets.

So, for example, suppose you collected individual transactions in a large retail chain store. Details of each transaction would be stored in triplets on different servers and hard drives, with a master table or map keeping track where exactly the respective transaction details are stored.

By using off-the-shelf standard hardware and open-source software for managing this distributed file system (such as Hadoop), reliable data repositories on the petabyte scale can be implemented relatively easily, and such storage systems are quickly becoming commonplace. Technical and implementation details of such systems can be found by searching the web for distributed file systems or Hadoop.

Unstructured Information

The majority of information collected into distributed file systems consists of unstructured information, such as text, pictures, or movie clips. This has advantages and disadvantages. The advantage is that the ability to store big data allows businesses or governmental agencies to store "all data" without worrying too much which part of the data are relevant and diagnostic for later analytic and decision support activities. The disadvantage is that in such cases a lot of subsequent processing on these very large amounts is required to extract useful information. While some of those operations may be simple (e.g., deriving simple counts, etc.), others will require more complex algorithms that have to be specifically designed to operate efficiently on the distributed file system (see also the next point).

Unstructured data and information. The general challenge here is, as it was when relational databases became popular, that just because one stores a lot of data, it is only the diagnostic information that can be extracted from that data that makes it useful. Case in point: a senior executive at a manufacturer once told a StatSoft implementation team about to deploy the *STATISTICA Enterprise* platform that he had "spent a fortune on IT and data storage, but is still not making money," because no thought had been given on how best to leverage the data to improve the core business activities. In more general terms, while the amount of data may grow exponentially, the ability to extract information and act on this information is limited and will asymptotically reach a limit (regardless of how much data is stored).

This is a critical point and consideration that will be further discussed below: The methods and procedures for extracting and updating models, and for automating decisions and decisioning processes, must be designed along with the data storage systems to ensure that such systems are useful and beneficial to the enterprise.

Analyzing Big Data

This is indeed the biggest challenge posed by big and often unstructured data: how to analyze it in a useful way. In fact, much less is written about this than about the storage solutions and technologies that make big data management possible. There are a number of issues to consider.

Map-Reduce

As a general approach, when analyzing hundreds of terabytes of data, or petabytes of data, it is not feasible to extract the data to another location for analysis (e.g., to the *STATISTICA Enterprise Analysis Server*). The process of moving data across wires to a separate server or

servers (for parallel processing) would take too long and require too much bandwidth. Instead, the analytic computations must be performed physically close to where the data are stored. It is easier to bring the analytics to the data than the data to the analytics.

Map-reduce algorithms, i.e., analytic algorithms designed according to this pattern, do exactly that. A central component of the algorithm will map sub-computations to different locations in the distributed file system and combine the results (the reduce-step) that are computed at the individual nodes of the file system. In short, to compute a count, the algorithm would compute sub-totals within each node and in parallel in the distributed file system, and report back to the map component the subtotals, which are then added up.

There is a wealth of information available on the Internet how various computations can be performed using this map-reduce pattern, including predictive modeling.

Simple Statistics, Business Intelligence (BI)

For simple BI reporting, many open-source solutions exist to enable the efficient computation of totals, averages, proportions, and so on using map-reduce. Thus, it is fairly easy to get accurate counts and other simple statistics for reporting purposes.

Predictive Modeling, Advanced Statistics

At first it may seem that it can be more complicated to build predictive models against a distributed file system; in practice, however, that is not really the case for a number of reasons.

Data preparation. Recall that much of the data in big data distributed file systems is often unstructured (e.g., text). In fact, it is hard to think of applications where actual measurements or numbers are collected resulting in petabytes of data. For example, StatSoft has conducted extensive and very successful projects involving very large datasets describing the minute-by-minute operations of large power plants, with the goal to improve efficiency and reduce emissions (Electric Power Research Institute, 2009). While such datasets can become very large, like most continuous process data, the information contained in them is of much lower dimension than the original data. For example, while data are collected second by second or minute by minute, specific damper settings, air flows, and furnace and gas temperatures remain mostly stable and invariant over large time intervals. Put another way, the data recorded second by second are mostly replications of the same information. Therefore, “smart” data aggregation performed upfront (and at the place where the data are stored) is required and easily done, yielding data sets for modeling and optimization that contain all relevant information about the dynamic changes affecting the efficiency and emissions of the plant.

Sentiment analysis and data pre-processing. This example illustrates how large datasets often contain quite obviously much lower dimensional information. Data collected by electricity meters in a smart-grid will have similar characteristics, as will sentiments expressed by the same person regarding the same topic, or perhaps by a group of people expressed for a wider range of topics. For example, StatSoft has been involved in projects involving the text mining of tweets related to such topics as satisfaction with specific airlines and their services. The

perhaps not surprising insight is that while a large number of relevant tweets can be extracted hourly and daily, the complexity and dimensionality of the sentiments expressed in them is rather straightforward (and of low dimension). Most tweets are complaints and brief single-sentence recounts of "bad experiences." In addition, the number and "strength" of those sentiments is relatively stable over time and across specific areas of complaints (e.g., lost luggage, bad food, cancelled flights).

Thus, in this case, basic compression of actual tweets to sentiment scores using text mining methods (as, for example, implemented in *STATISTICA Text Miner*; see also Miner, Elder, Fast, Hill, Delen, Nisbet, 2012) will yield much smaller data sets that can then be more easily aligned with existing structured data (actual ticketing, or frequent flyer-based info) to provide much better insight into the stratification of specific groups of customers and their complaints. Numerous tools exist to perform such data aggregation (e.g., sentiment scoring) in distributed file systems, and so this analytic workflow is easily accomplished.

Building Models

There are situations where the task is to build accurate models quickly against big data stored in a distributed file system. Actually, a more useful application is likely to build large numbers of models for smaller segments of data in a distributed file system – but more on that later.

In fact there are map-reduce implementations for various common data mining/predictive modeling algorithms suitable for massively parallel processing of data in distributed file systems (and these can be supported through StatSoft's *STATISTICA* platform). But then, just because you have crunched through a lot more data, is the final prediction model actually any more accurate?

Probability Sampling

In probability sampling, every observation in the population from which the sample is drawn has a known probability of being selected into the sample; when that probability is the same for every observation in the population, the sample is an equal probability sample or EPSEM sample (equal probability of selection method; see Kish, 1965, for details).

As put in a recent report by Forrester: "*Two plus two equals 3.9" is* [often] *good enough* (Hopkins & Evelson, 2011). The statistical and mathematical truth is that a linear regression model involving, for example, 10 predictors based on a correctly drawn probability sample of 100,000 observations will be as accurate as a model built from 100 million observations. Despite claims and, often, hype to the contrary put forth by some vendors in the big-data-analytics space that "all data must be processed," the truth is that the accuracy of a model is a

function of the quality of the sample (each observation in the population must have a known probability of selection) and its size relative to complexity of the model. The size of the population does not matter.

It is for that reason that, for example, national samples of voting preferences involving only a few thousand sampled likely voters can yield remarkably accurate predictions of actual voting outcomes.

Map-reduce sampling, data compression, data selection. What does this mean for big data analytics? There are very efficient (map-reduce) sampling algorithms available for distributed file systems that can provide an excellent solution for making big data available for straightforward and effective predictive modeling, to quickly derive insights from the investment in the storage infrastructure. For many applications and use cases, this is a very good path to choose, e.g., to deploy the *STATISTICA Enterprise* and *Data Mining* platform as the analytic tool on top of data interfaces to the distributed file system (the big data) to perform the data preparation/aggregation and/or probability sampling using map-reduce algorithms (driven by the *Enterprise* platform).

In addition to data aggregation and sampling, such a system can also perform the necessary detailed data selection (e.g., based on microsegmentation of specific customer groups and clusters) to provide the data to the *STATISTICA* analytics platform to build accurate models for specific segments (e.g., financial service offers for high-value households).

Integration of *STATISTICA* with open-source tools. A unique strength of the *STATISTICA Enterprise* and *Data Mining* platform is that it was specifically designed from the ground up as an enterprise computing platform, using industry standard scripting and data interfaces. That means that not only current StatSoft cutting edge tools, but also emerging open-source tools for data management and data preparation, as well as specialized analytics using map-reduce implementations, can easily be integrated into the platform and managed through the platform as just another analytic node in managed analytic workflows. For example, the open-source R platform is commonly used to implement highly specialized statistical computations and procedures, and the *STATISTICA* platform has supported the R platform through simple integration of R scripts into analytic workflows for many years.

Big data analytics and use cases are emerging and changing very rapidly. It is critical that the analytics platform around the distributed file system can easily leverage new methods for data preparation and aggregation, sampling, and stratification to help get the most value from the distributed file system investment.

Map-reduce implementations of specialized procedures. In addition to easy integration with open-source and other tools and platforms, it is equally important that the analytic platform of choice provides flexibility to customize analytic workflows to suit the specific analytic goals based on the distributed file system and big data. Use cases and best practices for big data analytics are emerging and evolving, and are not yet standardized in the way that analytic approaches and “traditional” predictive analytics have become well documented. That may,

however, change rather quickly, as all major vendors of database and BI tools (Microsoft, Oracle, Teradata, and others) provide interfaces and tools to access and process the data efficiently.

Either way, the *STATISTICA Enterprise* platform will provide the means for you to build customized implementations of specific analytic approaches based on data in distributed file systems, but will also support out-of-the-box interfaces and tools provided through standardized interfaces by major vendors. Probably the latter will be the most efficient and "normal" way to bring analytics to big data (through the *STATISTICA* platform).

Other Issues and Considerations for Implementation

To summarize the discussion so far, big data refers typically to mostly unstructured pieces of information stored in a distributed file system where individual pieces of data are stored across hundreds or thousands of hard drives and standard server hardware. The sizes of those distributed data repositories can easily exceed multiple petabytes (thousands of terabytes) in size given current technologies. In order to perform basic data preparation, cleaning, and extraction of such data efficiently, the respective low-level analyses are best performed at the site (the specific server) where data are stored (to reduce the data to summary statistics or aggregation), and then aggregated and mapped to produce summary data or statistics (the map-reduce framework).

Taking Advantage of Big Data by Building Large Numbers of Models

As discussed earlier in this paper, the real value of big data in distributed file systems is not to compute global (predictive) models based on all data instead of valid samples of data; the results and precision of models will be the same either way.

What makes more sense is to leverage the wealth of data and the tools to segment and cluster them efficiently to build large numbers of models for smaller clusters of data. For example, it can be expected that upsell models based on broad segmentation (20-30 year olds) will yield less accurate (valuable) results than large numbers of models built on more granular segmentations (20 to 21-year-old college students living in dorm rooms and who are business majors).

Thus, one way of taking advantage of big data is to leverage the available information by building large numbers of models for large numbers of segments, and use those models to score (predict) cases using the most appropriate model. Taken to the extreme, each individual "person," for example, in a large data repository of customers may be modeled with her/his own model to predict future purchases.

This means that the analytics platform supporting the data repository must be able to manage hundreds or even thousands of models, and have the ability to recalibrate those models if/when required. The *STATISTICA Decisioning* platform enables these critical capabilities, and

StatSoft has considerable experience with automated model building and calibration to support such systems.

Deployment of Models for Real Time Scoring

One of the core components of the *STATISTICA Enterprise* platform is *STATISTICA Livescore®,* which provides the ability to score new data in real time to support a service-architecture. In this environment, external programs can call models managed (and version controlled) through *STATISTICA Enterprise* to score new data either passed directly through the remote system call, or by passing an ID where to find a specific case or group of cases (observations) to score.

With respect to big data and distributed file systems, there is little difference between the scoring of data stored in relational databases vs. distributed file systems. The main challenge to achieve acceptable performance in this case is mostly on the data management and preparation side, which can be done using available map-reduce implementations for data preparation and extraction or by considering various other architectural choices. These may include separate analytic/scoring data warehouses based on relational databases and supported through map-reduce ETL routines, or even one of the emerging technologies based on RAM clouds, where the distributed file system itself is stored in very fast RAM "disks" to achieve much faster data access times. There are also a number of commercial solutions available, such as Netezza, Oracle Extreme, SAP Hanna, etc., which all essentially were built to accomplish the same thing: to access very large data repositories very quickly. Of course, *STATISTICA* can integrate with all of these solutions.

Criticism of Big Data Strategies, Implementation Strategies

StatSoft has been in business for more than 28 years and is dedicated to deliver best practice analytics for rapid ROI to its customers. StatSoft's *only* focus is the delivery of analytic solutions, not to sell hardware or dedicated storage solutions. Over the years, we have experienced various new technologies going through the regular cycle of early excitement and success by early adopters, and then maturing into standard solutions and workflows to optimize ROI. Along the way, inevitably, there may be certain disillusionment as the early overstated promises of new technologies do not materialize.

Here are some things to consider:

Big Data Does Not Necessarily Equate to Big Insights, Improvements

Suppose you had access to and stored the prices (last trade) of all publicly traded stocks on all US markets second by second, thus generating big data. Alternatively, also suppose you have a pipeline to targeted critical pieces of information regarding certain companies' performance indicators and earnings. Which is more valuable for driving a profitable trading strategy? Probably the latter.

Just by storing massive amounts of data describing some easily observed reality does not necessarily translate into better and more profitable insight regarding that reality. This is

equally true, regardless if you analyze stocks, twitter feeds, health care data, or CRM data, or are monitoring complex machinery for predictive maintenance.

For example, a reliable list of potential customers for home furnishings, along with basic demographics and net-worth info, can be much more valuable for a vendor of furniture than a massive data repository of click streams across various online furniture sites. When monitoring performance of power plants, we have learned [and demonstrated, see Electric Power Research Institute (EPRI), 2009] that paying attention to certain pieces of information and the changes that occur in (combinations of) some parameters is more diagnostic of subsequent performance and emissions than monitoring thousands of parameter data streams second by second.

As is the case in any project to optimize organizational or business performance, it is important to start with the questions of "What would ideal results look like," "How can I measure success (know when I am done, and know that I won)," and "What information is most likely diagnostic and useful for bringing about the ideal results." Answers to those questions may well lead to the implementation of a big data repository and analytics; in many cases, though, they may not.

Velocity of Data and Actionable Time Intervals

The other issue to consider is data velocity or the speed at which data updates. The key question here is about the "actionable time interval." Put another way, it may be so that you can build models in a manufacturing environment that predict impending problems one second forward based on continuously collected data streams for thousands of parameters. However, if it requires an engineer two hours to drill down and *do something* about it, such a system may be pointless.

Likewise, for a vendor of home furniture, it would be more important to get an "alert" a month or two *before* a home purchase occurs, instead of real-time information after it has occurred and when the prospective customer is perusing various internet sites to look for furniture. The early warning would enable a vendor to make numerous contacts with the prospect *before* the purchasing process starts, to present special offers and perhaps compel the prospect to visit a brick-and-mortar store to build personal relations with the vendor's designers. Again, a real-time platform of clickstreams may not be a useful data repository to drive traffic and build a loyal customer base.

Generally, the approach instead should be to start with a careful consideration of final use cases and strategies, on how to be successful. From that, the actionable time frames ("how much warning is required") will become obvious, and from that plan and set of requirements, an obvious optimal data collection and storage/warehousing and analysis strategy will follow.

Summary

The purpose of this paper is to provide a brief overview of the specific challenges posed by big data, i.e., data repositories in the terabyte to multiple petabyte (and beyond) range, and the technologies and approaches to overcome those challenges to derive value from big data.

Distributed file system technologies deployed on off-the-shelf server and storage hardware have made the creation and maintenance of such repositories practically and economically feasible. In those systems, instead of storing data on a single file system, data are stored and indexed across multiple (even thousands) of hard drives and servers, with redundant indexing to *map* where specific pieces of information can be found. Hadoop is one of the best known systems that uses this approach.

To process data in a distributed file system, it is necessary to move low-level computations such as counting, basic data preparation and aggregation, etc., to the actual physical location in the distributed file system where the data reside, rather than moving the data to the analytic engine. A *map* portion of the respective computational algorithms then keeps track of the local results and accumulates those (reduced) results; this approach and pattern to implement computational algorithms is known as *Map-Reduce*.

In practice, the value of big data rarely lies in the computation of statistical results over *all* data; in fact, there are statistical reasons why such computations will not yield more accurate results in most cases. Instead, the value of big data typically from a data mining and predictive modeling perspective more often is found in our ability to "micro-segment" the available information into smaller buckets, and to build large numbers of models specifically for smaller groups of observations. Other general considerations regarding the value of big data are also discussed in this paper.

From an implementation perspective, the analytic platform to mine big data must be able to leverage the emerging technologies for map-reduce algorithms, which often are available in user-supported and public domain projects. The *STATISTICA Enterprise and Decisioning* platform provides all capabilities to leverage big data, and to manage thousands of models applied against such data.

References

Eaton, C., Deroos, D., Deutsch, T., Lapis, G., Zikopoulos, P. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming data.* McGraw-Hill.

Economist Intelligence Unit Limited (2011). *Big data: Harnessing a game-changing asset.* The Economist.

Electric Power Research Institute (EPRI) / StatSoft Project 44771: *Statistical Use of Existing DCS Data for Process Optimization*; Palo Alto, 2009 (Principal Investigator: Thomas Hill, StatSoft Inc.; see also http://my.epri.com/portal/server.pt?Abstract_id=000000000001016494).

Hopkins, B., & Evelson, B. (2011). *Expand your digital horizon with Big Data.* Forrester Research Inc.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Manyika, J., Chi, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute.

Miner, G., Elder, D., Fast, A., Hill, T., Delen, D., Nisbet, R. (2012). *Practical text mining and statistical analysis for non-structured text data applications.* Academic Press.

Glossary

Big Data

Typically, the discussion of big data in the context of predictive modeling and data mining pertains to data repositories (and the analyses based on such repositories) that are larger than a few terabytes (1 terabyte = 1,000 gigabytes; 1 gigabyte = 1,000 megabytes). Some data repositories may grow to thousands of terabytes, i.e., to the petabyte range (1,000 terabytes = 1 petabyte). Beyond petabytes, data storage can be measured in exabytes; for example, the manufacturing sector worldwide in 2010 is estimated to have stored a total of 2 exabytes of new information (Manyika et al., 2011).

Distributed File System

Big data (multiple terabytes, petabytes) can be stored and organized in distributed file systems. While there are multiple implementations and implementation details, in the most general terms, information is stored on one of multiple (sometimes thousands of) hard drives and standard off-the-shelf computers; an index or map keeps track of where (on which computer/drive) a specific piece of information is stored. Actually, for failover redundancy and robustness, each piece of information is usually stored multiple times, e.g., as triplets.

So, for example, suppose you collected individual transactions in a large retail chain store. Details of each transaction would be stored in triplets on different servers and hard drives, with a master table or map keeping track of where exactly the respective transaction details can be retrieved.

By using off-the-shelf standard hardware and open-source software for managing this distributed file system (such as Hadoop), reliable data repositories on the petabyte scale can relatively easily be achieved, and such storage systems are quickly becoming commonplace.

Exabyte

1 exabyte is 1,000 petabytes, or 1,000 * 1,000 terabytes.

Hadoop

A distributed file system for storing and managing data repositories in the multiple terabytes to low petabyte range.

Map-Reduce

As a general approach, when analyzing hundreds of terabytes of data, or petabytes of data, it is not feasible to extract the data to another location for analysis; the process of moving data across wires to a separate server or servers (for parallel processing) would take too long and require too much bandwidth. Instead, the analytic computations must be performed physically close to where the data are stored. It is easier to bring the analytics to the data than the data to the analytics.

Map-reduce algorithms, i.e., data processing algorithms designed according to this pattern, do exactly that. A central component of the algorithm will map sub-computations to different locations in the distributed file system and combine the results (the reduce-step) that are computed at the individual nodes of the file system. In short, to compute a count, the algorithm would compute sub-totals within each node and in parallel in the distributed file system, and report back to the map component the subtotals, which are then added up.

Petabyte

1 petabyte = 1,000 terabytes.

Terabyte

1 terabyte = 1,000 gigabytes. Current distributed file system technology such as Hadoop allows for the storage and management of multiple terabytes of data in a single repository.