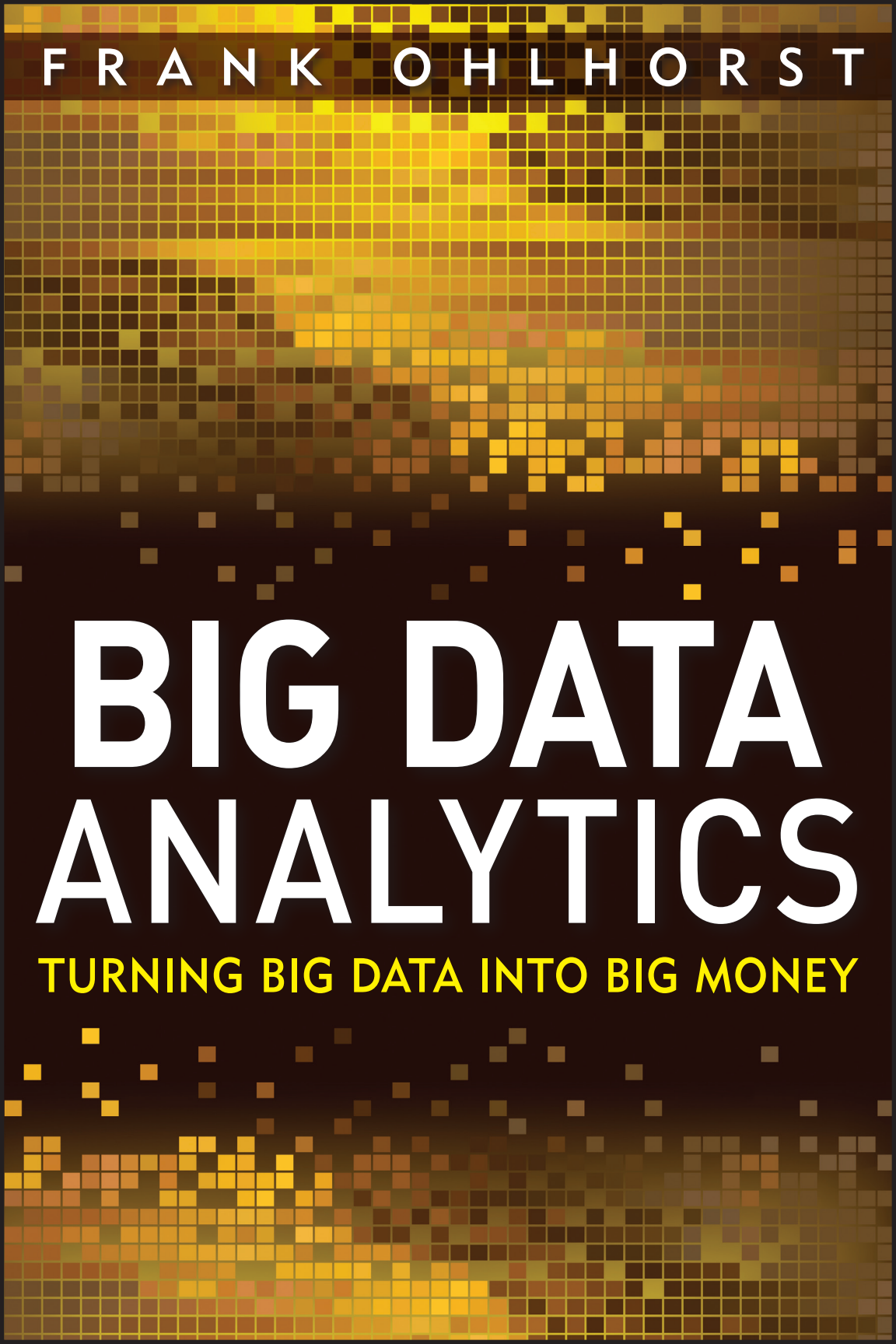

FRANK OHLHORST
BIG DATA ANALYTICS
TURNING BIG DATA INTO BIG MONEY

Contents

CHAPTER 1

What Is Big Data?

What exactly is *Big Data*? At first glance, the term seems rather vague, referring to something that is large and full of information. That description does indeed fit the bill, yet it provides no information on what Big Data really is.

Big Data is often described as extremely large data sets that have grown beyond the ability to manage and analyze them with traditional data processing tools. Searching the Web for clues reveals an almost universal definition, shared by the majority of those promoting the ideology of Big Data, that can be condensed into something like this: *Big Data* defines a situation in which data sets have grown to such enormous sizes that conventional information technologies can no longer effectively handle either the size of the data set or the scale and growth of the data set. In other words, the data set has grown so large that it is difficult to manage and even harder to garner value out of it. The primary difficulties are the acquisition, storage, searching, sharing, analytics, and visualization of data.

There is much more to be said about what Big Data actually is. The concept has evolved to include not only the size of the data set but also the processes involved in leveraging the data. Big Data has even become synonymous with other business concepts, such as business intelligence, analytics, and data mining.

Paradoxically, Big Data is not that new. Although massive data sets have been created in just the last two years, Big Data has its roots in the scientific and medical communities, where the complex analysis of

massive amounts of data has been done for drug development, physics modeling, and other forms of research, all of which involve large data sets. Yet it is these very roots of the concept that have changed what Big Data has come to be.

THE ARRIVAL OF ANALYTICS

As analytics and research were applied to large data sets, scientists came to the conclusion that more is better—in this case, more data, more analysis, and more results. Researchers started to incorporate related data sets, unstructured data, archival data, and real-time data into the process, which in turn gave birth to what we now call Big Data.

In the business world, Big Data is all about opportunity. According to IBM, every day we create 2.5 quintillion (2.5×10^{18}) bytes of data, so much that 90 percent of the data in the world today has been created in the last two years. These data come from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and cell phone GPS signals, to name just a few. That is the catalyst for Big Data, along with the more important fact that all of these data have intrinsic value that can be extrapolated using analytics, algorithms, and other techniques.

Big Data has already proved its importance and value in several areas. Organizations such as the National Oceanic and Atmospheric Administration (NOAA), the National Aeronautics and Space Administration (NASA), several pharmaceutical companies, and numerous energy companies have amassed huge amounts of data and now leverage Big Data technologies on a daily basis to extract value from them.

NOAA uses Big Data approaches to aid in climate, ecosystem, weather, and commercial research, while NASA uses Big Data for aeronautical and other research. Pharmaceutical companies and energy companies have leveraged Big Data for more tangible results, such as drug testing and geophysical analysis. The *New York Times* has used Big Data tools for text analysis and Web mining, while the Walt Disney Company uses them to correlate and understand customer behavior in all of its stores, theme parks, and Web properties.

Big Data plays another role in today's businesses: Large organizations increasingly face the need to maintain massive amounts of structured and unstructured data—from transaction information in data warehouses to employee tweets, from supplier records to regulatory filings—to comply with government regulations. That need has been driven even more by recent court cases that have encouraged companies to keep large quantities of documents, e-mail messages, and other electronic communications, such as instant messaging and Internet provider telephony, that may be required for e-discovery if they face litigation.

WHERE IS THE VALUE?

Extracting value is much more easily said than done. Big Data is full of challenges, ranging from the technical to the conceptual to the operational, any of which can derail the ability to discover value and leverage what Big Data is all about.

Perhaps it is best to think of Big Data in multidimensional terms, in which four dimensions relate to the primary aspects of Big Data. These dimensions can be defined as follows:

1. **Volume.** Big Data comes in one size: large. Enterprises are awash with data, easily amassing terabytes and even petabytes of information.
2. **Variety.** Big Data extends beyond structured data to include unstructured data of all varieties: text, audio, video, click streams, log files, and more.
3. **Veracity.** The massive amounts of data collected for Big Data purposes can lead to statistical errors and misinterpretation of the collected information. Purity of the information is critical for value.
4. **Velocity.** Often time sensitive, Big Data must be used as it is streaming into the enterprise in order to maximize its value to the business, but it must also still be available from the archival sources as well.

These 4Vs of Big Data lay out the path to analytics, with each having intrinsic value in the process of discovering value.

Nevertheless, the complexity of Big Data does not end with just four dimensions. There are other factors at work as well: the processes that Big Data drives. These processes are a conglomeration of technologies and analytics that are used to define the value of data sources, which translates to actionable elements that move businesses forward.

Many of those technologies or concepts are not new but have come to fall under the umbrella of Big Data. Best defined as analysis categories, these technologies and concepts include the following:

- **Traditional business intelligence (BI).** This consists of a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data. BI delivers actionable information, which helps enterprise users make better business decisions using fact-based support systems. BI works by using an in-depth analysis of detailed business data, provided by databases, application data, and other tangible data sources. In some circles, BI can provide historical, current, and predictive views of business operations.
- **Data mining.** This is a process in which data are analyzed from different perspectives and then turned into summary data that are deemed useful. Data mining is normally used with data at rest or with archival data. Data mining techniques focus on modeling and knowledge discovery for predictive, rather than purely descriptive, purposes—an ideal process for uncovering new patterns from large data sets.
- **Statistical applications.** These look at data using algorithms based on statistical principles and normally concentrate on data sets related to polls, census, and other static data sets. Statistical applications ideally deliver sample observations that can be used to study populated data sets for the purpose of estimating, testing, and predictive analysis. Empirical data, such as surveys and experimental reporting, are the primary sources for analyzable information.
- **Predictive analysis.** This is a subset of statistical applications in which data sets are examined to come up with predictions, based on trends and information gleaned from databases. Predictive analysis tends to be big in the financial and scientific

worlds, where trending tends to drive predictions, once external elements are added to the data set. One of the main goals of predictive analysis is to identify the risks and opportunities for business process, markets, and manufacturing.

- **Data modeling.** This is a conceptual application of analytics in which multiple "what-if" scenarios can be applied via algorithms to multiple data sets. Ideally, the modeled information changes based on the information made available to the algorithms, which then provide insight to the effects of the change on the data sets. Data modeling works hand in hand with data visualization, in which uncovering information can help with a particular business endeavor.

The preceding analysis categories constitute only a portion of where Big Data is headed and why it has intrinsic value to business. That value is driven by the never-ending quest for a competitive advantage, encouraging organizations to turn to large repositories of corporate and external data to uncover trends, statistics, and other actionable information to help them decide on their next move. This has helped the concept of Big Data to gain popularity with technologists and executives alike, along with its associated tools, platforms, and analytics.

MORE TO BIG DATA THAN MEETS THE EYE

The volume and overall size of the data set is only one portion of the Big Data equation. There is a growing consensus that both semi-structured and unstructured data sources contain business-critical information and must therefore be made accessible for both BI and operational needs. It is also clear that the amount of relevant unstructured business data is not only growing but will continue to grow for the foreseeable future.

Data can be classified under several categories: structured data, semistructured data, and unstructured data. Structured data are normally found in traditional databases (SQL or others) where data are organized into tables based on defined business rules. Structured data usually prove to be the easiest type of data to work with, simply

because the data are defined and indexed, making access and filtering easier.

Unstructured data, in contrast, normally have no BI behind them. Unstructured data are not organized into tables and cannot be natively used by applications or interpreted by a database. A good example of unstructured data would be a collection of binary image files.

Semistructured data fall between unstructured and structured data. Semistructured data do not have a formal structure like a database with tables and relationships. However, unlike unstructured data, semistructured data have tags or other markers to separate the elements and provide a hierarchy of records and fields, which define the data.

DEALING WITH THE NUANCES OF BIG DATA

Dealing with different types of data is converging, thanks to utilities and applications that can process the data sets using standard XML formats and industry-specific XML data standards (e.g., ACORD in insurance, HL7 in health care). These XML technologies are expanding the types of data that can be handled by Big Data analytics and integration tools, yet the transformation capabilities of these processes are still being strained by the complexity and volume of the data, leading to a mismatch between the existing transformation capabilities and the emerging needs. This is opening the door for a new type of universal data transformation product that will allow transformations to be defined for all classes of data (structured, semistructured, and unstructured), without writing code, and able to be deployed to any software application or platform architecture.

Both the definition of Big Data and the execution of the related analytics are still in a state of flux; the tools, technologies, and procedures continue to evolve. Yet this situation does not mean that those who seek value from large data sets should wait. Big Data is far too important to business processes to take a wait-and-see approach.

The real trick with Big Data is to find the best way to deal with the varied data sources and still meet the objectives of the analytical process. This takes a savvy approach that integrates hardware, software, and procedures into a manageable process that delivers results within an acceptable time frame—and it all starts with the data.

Storage is the critical element for Big Data. The data have to be stored somewhere, readily accessible and protected. This has proved to be an expensive challenge for many organizations, since network-based storage, such as SANS and NAS, can be very expensive to purchase and manage.

Storage has evolved to become one of the more pedestrian elements in the typical data center—after all, storage technologies have matured and have started to approach commodity status. Nevertheless, today's enterprises are faced with evolving needs that can put the strain on storage technologies. A case in point is the push for Big Data analytics, a concept that brings BI capabilities to large data sets.

The Big Data analytics process demands capabilities that are usually beyond the typical storage paradigms. Traditional storage technologies, such as SANS, NAS, and others, cannot natively deal with the terabytes and petabytes of unstructured information presented by Big Data. Success with Big Data analytics demands something more: a new way to deal with large volumes of data, a new storage platform ideology.

AN OPEN SOURCE BRINGS FORTH TOOLS

Enter Hadoop, an open source project that offers a platform to work with Big Data. Although Hadoop has been around for some time, more and more businesses are just now starting to leverage its capabilities. The Hadoop platform is designed to solve problems caused by massive amounts of data, especially data that contain a mixture of complex structured and unstructured data, which does not lend itself well to being placed in tables. Hadoop works well in situations that require the support of analytics that are deep and computationally extensive, like clustering and targeting.

For the decision maker seeking to leverage Big Data, Hadoop solves the most common problem associated with Big Data: storing and accessing large amounts of data in an efficient fashion.

The intrinsic design of Hadoop allows it to run as a platform that is able to work on a large number of machines that don't share any memory or disks. With that in mind, it becomes easy to see how Hadoop offers additional value: Network managers can simply buy a

whole bunch of commodity servers, slap them in a rack, and run the Hadoop software on each one.

Hadoop also helps to remove much of the management overhead associated with large data sets. Operationally, as an organization's data are being loaded into a Hadoop platform, the software breaks down the data into manageable pieces and then automatically spreads them to different servers. The distributed nature of the data means there is no one place to go to access the data; Hadoop keeps track of where the data reside, and it protects the data by creating multiple copy stores. Resiliency is enhanced, because if a server goes offline or fails, the data can be automatically replicated from a known good copy.

The Hadoop paradigm goes several steps further in working with data. Take, for example, the limitations associated with a traditional centralized database system, which may consist of a large disk drive connected to a server class system and featuring multiple processors. In that scenario, analytics is limited by the performance of the disk and, ultimately, the number of processors that can be bought to bear.

With a Hadoop cluster, every server in the cluster can participate in the processing of the data by utilizing Hadoop's ability to spread the work and the data across the cluster. In other words, an indexing job works by sending code to each of the servers in the cluster, and each server then operates on its own little piece of the data. The results are then delivered back as a unified whole. With Hadoop, the process is referred to as MapReduce, in which the code or processes are mapped to all the servers and the results are reduced to a single set.

This process is what makes Hadoop so good at dealing with large amounts of data: Hadoop spreads out the data and can handle complex computational questions by harnessing all of the available cluster processors to work in parallel.

CAUTION: OBSTACLES AHEAD

Nevertheless, venturing into the world of Hadoop is not a plug-and-play experience; there are certain prerequisites, hardware requirements, and configuration chores that must be met to ensure success. The first step

consists of understanding and defining the analytics process. Most chief information officers are familiar with business analytics (BA) or BI processes and can relate to the most common process layer used: the extract, transform, and load (ETL) layer and the critical role it plays when building BA or BI solutions. Big Data analytics requires that organizations choose the data to analyze, consolidate them, and then apply aggregation methods before the data can be subjected to the ETL process. This has to occur with large volumes of data, which can be structured, unstructured, or from multiple sources, such as social networks, data logs, web sites, mobile devices, and sensors.

Hadoop accomplishes that by incorporating pragmatic processes and considerations, such as a fault-tolerant clustered architecture, the ability to move computing power closer to the data, parallel and/or batch processing of large data sets, and an open ecosystem that supports enterprise architecture layers from data storage to analytics processes.

Not all enterprises require what Big Data analytics has to offer; those that do must consider Hadoop's ability to meet the challenge. However, Hadoop cannot accomplish everything on its own. Enterprises will need to consider what additional Hadoop components are needed to build a Hadoop project.

For example, a starter set of Hadoop components may consist of the following: HDFS and HBase for data management, MapReduce and OOZIE as a processing framework, Pig and Hive as development frameworks for developer productivity, and open source Pentaho for BI. A pilot project does not require massive amounts of hardware. The hardware requirements can be as simple as a pair of servers with multiple cores, 24 or more gigabytes of RAM, and a dozen or so hard disk drives of 2 terabytes each. This should prove sufficient to get a pilot project off the ground.

Data managers should be forewarned that the effective management and implementation of Hadoop requires some expertise and experience, and if that expertise is not readily available, information technology management should consider partnering with a service provider that can offer full support for the Hadoop project. Such expertise proves especially important for security; Hadoop, HDFS, and

HBase offer very little in the form of integrated security. In other words, the data still need to be protected from compromise or theft.

All things considered, an in-house Hadoop project makes the most sense for a pilot test of Big Data analytics capabilities. After the pilot, a plethora of commercial and/or hosted solutions are available to those who want to tread further into the realm of Big Data analytics.

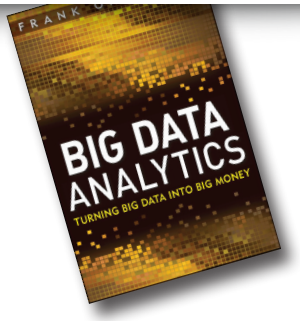
About the Author

Frank J. Ohlhorst is an award-winning technology journalist, professional speaker, and IT business consultant with over 25 years of experience in the technology arena. Frank has written for several leading technology publications, including *ComputerWorld*, *TechTarget*, *CRN*, *Network Computing*, *PCWorld*, *ExtremeTech*, and *Tom's Hardware*. Frank has contributed to business publications, including *Entrepreneur* and *BNET*, and to multiple technology books. He has written several white papers, case studies, reviewers' guides, and channel guides for leading technology vendors.

ANALYTICS

Know what's hot.

The topic of analytics is on fire right now. With SAS,® you can discover innovative ways to increase profits, reduce risk, predict trends and turn data assets into new business opportunities.



Visit **go.sas.com/bigmoney** to buy *Big Data Analytics: Turning Big Data into Big Money.* This recommended book discusses how to make Big Data a key component in your organization's growth strategy.

sas®
THE POWER TO KNOW®