The Big Data Revolution

The amount of data generated by people, Internet-connected devices and companies is growing at an exponential rate. Financial institutions, companies and health service providers generate large quantities of data through their interactions with suppliers, patients, customers and employees. Beyond those interactions, large volumes of data are created through Internet searches, social networks, GPS systems and stock market transactions. This widespread production of data has resulted in the “data revolution” or the Age of Big Data.

The term “Big Data” is used to describe a universe of very large sets of data composed of a variety of elements. This gives way to a new generation of information technology designed to make available the increased processing speeds necessary to analyze and extract value from large sets of data, employing – of course – specialized materials and software. The phenomenon of Big Data not only refers to the explosion in the volume of data produced, which was made possible by the development of information storage and dissemination capacities on all sorts of platforms, but the term also refers to a second phenomenon, which involves newfound data processing capabilities.

In general terms, the concept of Big Data describes the current state of affairs in the world, in which there is a constant question of how to manage lumps of data in a better way, and how to make sense of the massive volume of data produced daily.
Data sources are multiplying: smartphones, tablets, social networks, web services and so on. Once these intelligent objects are connected to the Internet, they can feed data into enormous databases and communicate with other objects and humans [PRI 02]. This data must be processed and developed in order to become “intelligent” or “smart”. Intelligence, which can be brought out by using analysis techniques, can provide essential information that top management will require in order to determine strategies, boost operational performance and manage risks.

To this end, “data scientists” must pool their strengths in order to face the challenges of analyzing and processing large pools of data, gaining clarity and precision. Data scientists must make data “speak” by using statistical techniques and specialized software designed to organize, synthesize and translate the information that companies need to facilitate their individual decision-making processes.

1.1. Understanding the Big Data universe

The IT craze that has swept through our society has reached a new level of maturity. When we analyze this tendency, we cannot help being overwhelmed by the transformations that it has produced across all sectors. This massive wave developed very quickly and has resulted in new applications. Information and communication technologies (ICTs) and the advent of the Internet have triggered an explosion in the flow of information (Big Data). The world has become digital, and technological advances have multiplied points of access to data.

But, what exactly is Big Data? The concept really took off with the publication of three important reports from the McKinsey Institute:

- Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch [BUG 10];
- Are You Ready for the Era of “Big Data”? [BRO 11];
- Big Data: The Next Frontier for Innovation, Competition and Productivity [MAN 11].
“Big Data” describes: “a series of data, types of data, and tools to respond quickly to the growing amount of data that companies process throughout the world." The amount of data gathered, stored and processed by a wide range of companies has increased exponentially. This has partially benefited from an explosion in the amount of data resulting from web transactions, social media and bots.

The growth of available data in terms of quantity, diversity, access speed and value has been enormous, giving way to the “four Vs”: “Volume”, “Variety”, “Velocity” and “Value”, that are used to define the term Big Data:

– Volume: the advent of the Internet, with the wave of transformations in social media it has produced; data from device sensors; and an explosion of e-commerce all mean that industries are inundated with data that can be extremely valuable. All these new devices produce more and more data, and in turns, enrich the volume of existing data;

– Variety: with the rise of Internet and Wi-Fi networks, smartphones, connected objects and social networks, more and more diverse data is produced. This data comes from different sources and varies in nature (SMSs, Tweets, social networks, messaging platforms, etc.);

– Velocity: the speed at which data is produced, made available, and interpreted in real-time. The possibility of processing data in real-time represents a field of particular interest, since it allows companies to obtain results like personalized advertisements on websites, considering our purchase history, etc.;

– Value: the objective of companies is to benefit from data, especially by making sense out of it.

The challenges of Big Data are related to the volume of data, its variety, the speed at which it is processed, and its value. Some

2 In this sense, Viktor Mayer-Schönberger has underscored the difficulty in quantifying the value of data in his book, Big data, A Revolution That Will Transform How We Live, Work and Think.
scholars add another three “Vs”, namely3: “Variability”, “Veracity”, and “Visualization”.

The first V refers to data whose meaning evolves constantly. The second qualifies the result of the data’s use, since even though there is a general consensus about the potential value of Big Data, data has almost no value at all if it is not accurate. This, particularly, is the case for programs that involve automatic decision-making, or for data feeding into unmonitored machine learning algorithms. The last V, which touches on one of the greatest challenges of Big Data, has to do with the way in which the results of data processing (information) are presented in order to ensure superior clarity.

The expression “Big Data” represents a market in and of itself. Gilles Grapinet, deputy CEO of Atos notes that “with Big Data, organizations’ data has become a strategic asset. A giant source of unexpected resources has been discovered.” This enormous quantity of data is a valuable asset in our information society.

Big Data, therefore, represents a large discipline that is not limited to the technological aspect of things. During recent years, the concept has sparked growing interest from actors in the information management systems sector. The concept of the “four Vs” or even that of the “seven Vs” opens up new avenues for consideration and research, but they do not provide a clear definition of the phenomenon. The sum of these “Vs” gives way to new perspectives for new product creation through improved risk management and enhanced client targeting. Actions aimed at anticipating and reducing subscription suspensions or at making customers more loyal can also be envisioned.

The increase in the volume of data, processing speed and data diversity all present new challenges to companies and affect their decision-making processes. And yet, companies that produce, manage and analyze vast sets of data on a daily basis now commonly use terms such as terabyte, petabyte, exabyte and zettabyte.

### Table 1.1. Data units of measurement

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 byte</td>
<td>8 bits</td>
<td>1</td>
</tr>
<tr>
<td>1 kilobyte</td>
<td>KB</td>
<td>$10^3$ bytes</td>
</tr>
<tr>
<td>1 megabyte</td>
<td>MB</td>
<td>$10^6$ bytes</td>
</tr>
<tr>
<td>1 gigabyte</td>
<td>GB</td>
<td>$10^9$ bytes</td>
</tr>
<tr>
<td>1 terabyte</td>
<td>TB</td>
<td>$10^{12}$ bytes</td>
</tr>
<tr>
<td>1 petabyte</td>
<td>PB</td>
<td>$10^{15}$ bytes</td>
</tr>
<tr>
<td>1 exabyte</td>
<td>EB</td>
<td>$10^{18}$ bytes</td>
</tr>
<tr>
<td>1 zettabyte</td>
<td>ZB</td>
<td>$10^{21}$ bytes</td>
</tr>
<tr>
<td>1 yottabyte</td>
<td>YB</td>
<td>$10^{24}$ bytes</td>
</tr>
</tbody>
</table>

In 1999, Wal-Mart (one of the US’s most important retail chains) had a database of 1,000 terabytes (that is, 1,000,000 gigabytes of data). By 2012, this quantity had grown to over 2.5 petabytes (2.5 million gigabytes) of data.

**Box 1.1. Wal-Mart’s database**

The Big Data phenomenon has rendered classical data processing methods antiquated, and now stands as an opportunity in the business world, especially for companies that know how to use it.

Amazon.com, a leader in product suggestions processed through Big Data, analyzes purchase histories, product listings, and commentaries posted on its site.

**Box 1.2. Amazon and Big Data**
There are several methods with which a company can create value from its data assets. Data can be used to improve understanding of customers’ needs and adapt products accordingly. Companies can use data to monitor and control performance of key functions on their website, identify factors that contribute to the gaps observed and discover necessary corrective measures, or find new ways of optimizing the existing management systems.

Some companies combine data to predict customers’ behavior and thus take the necessary measures. Several other uses allow companies to better navigate their environment.

To show how to use a large volume of data for practical purposes, we will take the example of Wal-Mart, which was able to identify a strong correlation between purchases of baby diapers and beer by analyzing sales receipts. This real-time correlation system operating on a constant flow of data allowed Wal-Mart to increase its sales by simply placing diapers and beer closer together on their shelves.

**Example 1.1. A sales receipts analysis by Wal-Mart**

Another illustrative example is Amazon’s automatic book suggestion service for customers, which is based on the books that the clients have already purchased or consulted. Amazon develops a profile for each of its customers based on his or her previous searches and purchases, and then compares it to the content of other available books, which allows the company to suggest books according to clients’ tastes.

**Example 1.2. Book suggestions for Amazon customers**
As another example, the company Nike provides its clients with a complete “ecosystem” focused on health and well-being. The data sources used (see, for example, Nike+ and FuelBand) provide a personalized service for clients and allow them to monitor their physical activity. In exchange, the company gains access to real-time intimate data on its clients’ habits and uses, which it can, in turns, use to improve its product offerings.

**Example 1.3. An ecosystem provided by Nike**

Big Data has, therefore, transformed companies in all sectors, as well as their operations and ways of acting. We can also add dynamic real-time analyses, thanks to the speed at which techniques and software make it possible to obtain results. Analyses bring to light the changes in client behavior and reveal new needs. They also make it possible to predict needs that do not even exist yet, which enables strategic decision-making.

Big Data allows companies to measure different aspects of daily life and to find correlations between these different measures, all with the aim of finding relations that companies themselves might never have imagined. It opens up the possibility of examining a market composed of millions of clients and to see them not as a vague mass, but rather as individuals with specific tastes and values. It enables companies to have a statistical base for identifying tendencies through data analysis tools.

The rise of Big Data reflects the growing awareness of the “power” behind data, and of the need to enhance gathering, exploitation and sharing processes within companies. As it enables more efficient decision-making procedures, gaining access to a large volume of information and to the tools necessary to process it can allow companies to attain a better strategic position. Data thus becomes companies’ new strategic asset, no matter their sector.
1.2. What changes have occurred in data analysis?

Companies have always needed to analyze data in order to have a precise understanding of their situation and to predict their future business moves.

Data analysis, when it is not preceded by the word “Big”, refers to the development and sharing of useful and effective models. For the most part, it uses a variety of methods from different research fields, like statistics, data mining, visual analysis, etc. It caters to a wide range of applications, including data summarization, classification, prediction, correlation, etc.

In the 1970s and 1980s, computers could process information, but they were too large and too costly. Only large firms could hope to analyze data with them. Edgar F. Codd and Hubert Tardieu were the first to work on data organization by designing database management systems (DBMSs), in particular relational databases. Data processing and analysis, in the present day, are brought together under the notion of “Business Intelligence”, due especially to computers’ increased processing capabilities.

A fundamental requirement for successful data analysis is to have access to semantically rich data that links together pertinent information elements for objective analysis. However, the situation has changed with Big Data because data now comes from several sources of very different kinds and in different forms (structured, unstructured). This leads us to say that new data processing tools are now necessary, as are methods capable of combining thousands of datasets.

In the Big Data universe, companies seek to unlock the potential of data in order to generate value. They are also impatient to find new ways to process that data and make more intelligent decisions, which will result in better client service, improved process efficiency and better strategic results.

In the literature, the concept of Big Data is defined in terms of the theory of the “four Vs” or of the “seven Vs”. The exponential speed at
which data is generated, as well as the multiplicity of sources that generate it in different formats (digital, text, images, etc.), are characteristic of this phenomenon:

“Big Data refers to volume, variety, and velocity of data – structured or unstructured – that is transmitted across networks in transformation processes and across storage devices until it becomes knowledge that is useful for companies”, (Gartner Research Firm).

The following image shows one perspective on the massiveness of data and on its growing evolution through different interconnected technologies. This volume of data is available today due to storage capacities that have increased while their cost has correspondingly diminished.

This large collection of data is often created in real-time and its quick processing provides knowledge to managers that was previously inaccessible. At the same time, it allows them to optimize their decision-making processes. Data is, therefore, transformed into a plan of action to be put into place, into decisions to be taken and into new markets to explore.

---

4 American consultant and IT research specialist.
There are essentially three types of challenges surrounding Big Data:

– massive data storage management, in the order of hundreds of terabytes or of petabytes, which go beyond the current limits of classic relational databases in terms of data storage and management;

– unstructured data management (which often constitutes the largest portion of data in Big Data scenarios). In other words, how to organize text, videos, images, etc.;

– analysis of this massive data, both for reporting and for advanced predictive modeling, but also for its deployment.

In current usage, the term “Big Data” does not refer exclusively to vast sets of data. It also involves data analysis and value extraction operating on large volumes of data. The expression “Big Data” thus refers to the technologies, processes and techniques that enable organizations to create, manipulate and manage data on a large scale [HOP 11], as well as to extract new knowledge in order to create new economic value.

The large volume of data collected, stored and disseminated through different processing technologies is currently transforming priorities and developing new analysis tools, which are in line with changes in companies’ operations and which will transform the business landscape. At the same time, new analytic techniques make it possible to examine the datasets. Processing them will play a crucial role, and will allow companies to gain a competitive advantage.

The process has to do with gathering data, cleaning it up, and organizing it in different databases. Next, the data is compared in order to be organized and uploaded to software capable of analyzing it (to find correlations within it). The data from the different sources is then combined to be transformed into a technical diagram. In other words, data warehouses, created for the analysis, are examined in order to present the results in different forms (visualization).
The present day software tools make it possible to process and assimilate this massive volume of data quite quickly. Understanding the technological dimension of things is nevertheless fundamental because it makes it possible to understand its limits and potentialities, as well as to identify the most relevant actions to take. With exponential increase in the volume of data, companies attempt to use available analysis tools to find out how to extract value from their gathered data.

A study carried out by [MAC 12] showed that companies that adopted advanced data analysis tools attain more productivity and better profit margins than their competitors. In fact, technical competence in data processing is now a genuine strategic asset for companies’ competitive differentiation [BUG 11].

Thanks to new Big Data methods and tools, it has become possible to work on large volumes of data. The result is an advantage stemming from the possibility of bringing to light correlations in new data. Interpreting this large volume of data is the greatest challenge facing Big Data, since information resulting from it can be the basis for new knowledge that brings about development opportunities.

Technology now offers a perspective on data as structured and therefore, static. Technological limits, in terms of performance and storage, reduce the scope of possible analysis to sets of explicit data. Most solutions provide the couple: “storage and processing”. It is worth noting that growth in the volume of data has been accompanied by a reduction in the price of storage.

Currently, one of the innovations that make it possible to share and store large volumes of data is “Cloud Computing”. The “Cloud” allows access to shared computing resources through an on-demand telecommunication network or self-service modules. The cloud transforms storage infrastructure and computing power into services through the intermediary of companies that possess servers and rent out their capacities. This approach makes it possible to share costs and to provide greater data storage and processing flexibility for users.
1.3. From Big Data to Smart Data: making data warehouses intelligent

Data has always held strategic value, but the scale of the data available and processing capacities today have resulted in a new category of assets. We find ourselves at the beginning of a long journey where, with the right principles and guidelines, we will be able to gather, measure and analyze more and more data to make better decisions, individually or collectively.

The massive flow of data, or “Big Data”, which is generated by the Internet, social media, cloud computing, etc. is developing very quickly. This prompts companies to rethink their strategies and go beyond the difficulties involved in processing large volumes of data. It will soon become possible to organize and transform data into
information, which will, in turn, be transformed into knowledge useful for cognitive or intellectual operations. However, attaining the complete potential of data depends on the way in which it is presented. It must be used and reused in different ways, without its value being diminished. This requires making data available in the right form and at the right time to any party interested in exploiting and adding value to it.

“Data Is the New Oil” [ROT 12]. It is the indispensable raw material of one of the new century’s most important activities: data intelligence. However, it is important to be prudent in our predictions because a lot of data is not yet “the right data”. There is, therefore, an underlying difficulty behind Big Data, since more data is not necessarily better data. It is possible to obtain better results by making better use of available data.

Twitter has become a popular source for Big Data exploitation, but working with Twitter data represents a huge methodological challenge that is seldom taken up.

**Box 1.3. A methodological challenge: Twitter**

When researchers encounter a set of data, they need to understand not only the limits of the available set of data, but also the limits of the questions that it can respond to, as well as the range of possible appropriate interpretations.

However, it is imperative for those combinations to be made rigorously and with methodological transparency. This leads us to say that it is not so much a question of size, but rather of what can be done with a given set of data. After all, the objective of gathering and analyzing that data is not simply to attain knowledge, but also to act. In the field of marketing, for example, Big Data processing must employ the necessary tools to determine the appropriate type and level of action it recommended from attracting and retaining each client at the lowest possible cost and for negotiating the continuous relationship with an optimal profit level.
The McKinsey Global Institute’s July 2013 report estimates that Big Data applications could result in a $30 million increase in retail sales due to productivity gains, increased transparency, and more sophisticated targeted marketing, advertisements and sales.

Box 1.4. McKinsey Global Institute’s 2013 report

But how can we obtain or make progress with such benefits by using Big data? How can companies bring together and combine data from disparate sources to achieve projected gains? What role can the data analysis play in what amounts to an IT challenge? What changes are required in order for data analysis to become a more practical discipline? These questions refer to some of Big Data’s greatest challenges, and they represent the difficulties that make it a “Big Challenge”.

The greatest objective for Big Data involves intelligent database management aimed at identifying and extracting pertinent information allowing companies or users to establish strategies that actually address identified needs. Intelligent data makes it possible to go from raw (structured or unstructured) data coming from internal or external sources to strategic information.

The ultimate goal is not only to collect, combine or process all data, but also to increase its value and efficiency. This means that we must evolve from “Big” data to “Smart” data, since the effectiveness of companies’ strategies now depends on the quality of data.

Data quality refers to its adequacy for its envisioned use in operations, processes, decision-making and planning. Data quality, in this regard, also has an impact on product lifecycle analysis. Data quality is highly important because it represents a key source of value for companies.

5 In its flexibility approaches, companies seek operational efficiency. Exploiting high-quality data makes it possible to optimize participation and interaction of different staff members in the decision-making process. In its analysis and presentation of data, companies must ensure quality because its impact is important. The general idea is to manage data as a company asset, just like products, employees and clients.
Data quality is important for monitoring and evaluating progress towards objectives. It is all the more important when it relates to reliable and accurate information gathered through company data management systems. Having access to accurate information makes it possible to:

– demonstrate responsibility and good governance;

– provide decision-makers with the information necessary to plan, allocate resources and elaborate strategies;

– monitor progress towards the attainment of established goals and objectives.

Indeed, companies must not rely on the size of their data – it is not useful unless it is applied in an intelligent manner. Therefore, the volume of data is of little importance, since internal data must be combined with external data in order for a company to obtain the most out of its data. What is truly necessary are excellent analytic skills, a capacity to understand and manipulate large sets of data, and the capacity to interpret and apply the results.

The challenge is to consider the data’s use, rather than its quantity. This could become the most profitable way of extracting the value of data from the massive sources available. The evolution from “Big Data” to “Smart Data” represents a new awareness of the importance of data processing. It is at this level that the figure of the “data scientist” appears. They are well-trained in computer science, mathematics and statistics, which they combine with good knowledge of the business world. They must be able to analyze a phenomenon from all possible angles in order to draw profit from the company’s data assets.

In this way, the term “variety” involves several different issues. First of all, data – especially in an industrial environment – can be presented in several different ways, such as texts, functions, curves, images and graphs, or a combination of these elements. On the other hand, this data shows great variety, which often reflects the complexity of the studied phenomenon. It is, therefore, important to
be open about the structure of the observations’ content in order to draw the right conclusions from it.

Processing large volumes of data by enlisting people from the “new” profession of data scientist is a major focus for companies that have placed themselves at the center of the flood of data requiring specialized processing. The development of information technology and computation tools makes storage of large databases possible, as well as processing and analysis of very large sets of data.

More recently, improvements in software and their interfaces, both for statisticians and non-specialized users, have made it much simpler to apply these methods. This evolution, as well as the popularization of new algorithmic techniques (neural networks) and graphing tools, have led to the development and commercialization of software that brings together a subset of statistical and algorithmic methods, and which is known as “data mining”.

In this regard, data mining refers to the search for pertinent information that may be helpful for decision-making and planning. It employs statistical machine learning techniques that can handle the specificity of large to very large volumes of data.

However, in the Big Data universe, the main objective is still related to a much more important “V”, namely value-extraction. “The consensus today is to place the data scientist at the intersection of three fields of expertise: computer science, statistics, and mathematics, “Business Knowledge” [ABI 13]. In relation to new data structures, the statistician-turned-data scientist revisits basic notions to focus on tools and methods that can lead to useful applications compatible with new information systems.

1.4. High-quality information extraction and the emergence of a new profession: data scientists

The massive amount of data currently produced in real-time requires analysis, processing and exploration. In order to face the Big Data challenge, companies must have the capacity to make their
constant flow of data “speak”. This has resulted in the emergence of employment prospects in new careers that attract increasingly more attention. We are of course talking about “data scientists” and “data analysts”.

Data scientists examine large volumes of data from which they, in turn, identify the main tendencies in order to help companies make better decisions. Data scientists analyze data from different sources and examine information from all possible angles; it is therefore useful to understand the importance of their contributions to companies.

For Simon Rogers, “a data expert is above all capable of bringing data and analysis together, and to make his or her work accessible. Theirs is a translation job: by translating data into a form that can be understood by more people, data experts improve the world a little bit”.

Moreover, John Foreman, as chief data scientist at MailChimp, confirms, “If by data scientist you mean someone who can create a data summary or aggregate, or model a task specifically assigned in advance, it’s not surprising that the job can be paid at $30 an hour”.

On that same note, DJ Patil, data expert for LinkedIn, the professional social network, explains that “the role of data scientists requires striking a balance between technical data skills and a capacity to tell that data’s story”. This is a perspective shared by Hilary Mason, science director at Bitly, who describes the ideal candidate for this atypical job: “A data scientist is a rare hybrid between a developer, a statistician, and a careful analyst of data and human behavior”.

Data scientists therefore have a digital-age job that is as related to finance as it is to banking, insurance, marketing and human resources. The job of data scientist is a profession born out of data science. It is a new discipline that brings together elements from different fields including mathematics, statistics, computer science and data science.

6 The Twitter data journalist.
visualization and modeling. In fact, data science extracts knowledge both from companies’ internal and external data.

Since 2010, demand for this new career profile has increased dramatically, as is shown in Figure 1.2.

![Figure 1.2. The importance of data scientists. Source: http://www.indeed.com/jobtrends/Data-scientist.html](image)

The increase in data scientist demand is fed by the success of companies like Google, Facebook, LinkedIn and Amazon, which have invested in data science only in order to use their databases in a creative manner. Once the data is organized, a company can focus on understanding its meaning and implications instead of wasting time managing it.

The need to analyze and use enormous amounts of data more efficiently drives companies towards data science in the hope of unlocking the power of Big Data. A data scientist must have a general grasp of business and be capable of analyzing data in order to extract knowledge by using computer science applications.

---

Data scientists are not only highly-trained computer scientists, they are also innovative thinkers capable of gleaning new perspectives on general tendencies in the data available. A data analyst analyzes data from a variety of sources and examines its different tabs in order to attain a general understanding of the phenomenon it describes and enable a company to develop competitive improvements.

A data scientist’s interpretations enable top management personnel to take advantage of relevant information and thereby obtain excellent results. Google’s main economist, Hal Varian, confirms: “The most popular job in the next ten years will be that of statistician: being able to take data, understand it, process it, extract information from it, and visualize and communicate it”.

A study by the McKinsey Global Institute estimates that by 2018, there will be a deficit of 140,000 to 190,000 people with analytic skills in the US, as well as 1.5 million managers capable of using Big Data analytics in order to make enhanced decisions. The recruitment firm Robert Half published a list of 6 golden jobs, including data scientist, for 2014 and 2015.

### The six “golden jobs” according to Robert Half 2014-2015:
- Financial Comptroller (Corporate Finance)
- General Accountant (Accountancy)
- Compliance Officer (Banking)
- Data Scientist (Insurance)
- Compensation and Benefits Manager (Human Resources)
- Technical Director (IT)

**Box 1.5. The six jobs of the future according to R. Half**

Data science occupies the central position in companies’ priorities. As an example, Yahoo devoted a significant amount of resources to data science. After witnessing Google’s use of “MapReduce” to analyze enormous quantities of data, companies realized they had similar needs. The result was “Hadoop”, which today is one of the most important tools in a data scientist’s toolbox.
“Hadoop” is a specific kind of open source database architecture, which makes it possible to process all kinds of data in mass (including unstructured data) based on a non-relational mode (data is not processed according to its internal relations). Hadoop constitutes a sort of library from which multi-thread computations can be carried out.

“MapReduce” is a massive multi-thread programming model designed to process very large volumes of data. MapReduce is a Google Corporation product. Programs that adopt this model are automatically threaded together and run on computer clusters.

Example 1.5. Two examples of open source data

The best data scientists will also understand the latest business trends and be able to compare a company’s data to industry or competitor indicators in order to correctly diagnose the situation and obtain useful information. An “analyst” or data scientist employs different available methods to interpret data in relation to a decision-making context. Their role requires a variety of skills, including:

– technical and statistics training;
– general technology and IT-savviness;
– familiarity with the field in which the analyzed data will be applied.

Data has always played an important role and contributed to companies’ operational and strategic planning. However, it is necessary for each company to define the settings that help it understand the enormous heaps of available data the most. But in the Big Data universe, it is also important to open up new research and analysis perspectives, which requires highly skilled personnel.
capable of exploiting and processing sets of data and adding value to them.

1.5. Conclusion

Big Data is gradually becoming an inevitable concept that revolutionizes the way in which many companies do business, since the scope of their business goes beyond the boundaries of their specific sectors and belongs to a globalized world. Smartphones, tablets, clouds, GPS, the Web, etc. – these are the new tools of a trade whose goal is to refine a certain raw material: data. Data, the new strategic asset, will without a doubt, influence global strategies throughout. As a consequence, the data processed by technological platforms has become fundamental in overhauling decision-making processes.

It becomes necessary to clean up, analyze and compare (structured or unstructured) data produced by companies, governments and social networks to develop new uses. Big Data is the general term used to describe the exponential increase in the volume of data, which has been accompanied, of course, by growth in the capacity to transfer, store and analyze that data. However, storing large amounts of data is one thing; processing it is another.

Big Data does not amount to processing more data; it is rather a question of developing it. Analytic capacities capable of developing new applications and uses of the massive amount of data available are crucial. This is the work of data scientists who possess both the technical skills required for data analysis and the capacity to understand the strategic stakes involved in the analysis.

But data quality is of utmost importance: well-prepared, well-classified and integrated data that allows companies to benefit completely from its analysis. Without this preparatory phase, analytic processing will not produce performance data, which companies need in order to become more competitive. The challenges of Big Data are therefore related to the volume of data, its variety, the velocity in which it is processed and its value.