Big Data: Concepts, Challenges and Applications

Imen Chebbi(✉), Wadii Boulila, and Imed Riadh Farah

Laboratoire RIADI, Ecole Nationale des Sciences de L'Informatique, Manouba, Tunisia
ichebbi88@gmail.com, wadii.boulila@riadi.rnu.tn, riadh.farah@ensi.rnu.tn

Abstract. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.Big Data it's varied, it's growing, it's moving fast, and it's very much in need of smart management In this paper, we first review a literature survey of big data, second review the related technologies, such as could computinq and Hadoop, We then focus on the five phases of the value chain of big data, i.e.,data sources, data collection, data management, data analysis and data visualization. And we finally examine the several representative applications of big data.

Keywords: Big data · Cloud computing · Hadoop · Big data analytic · Mapreduce

1 Introduction

Today the term big data [1] [2] draws a lot of attention, but behind the hype there's a simple story. For decades, companies have been making business decisions based on transactional data stored in relational databases. Beyond that critical data, however, is a potential treasure trove of non-traditional, less structured data: weblogs, social media, email, sensors, and photographs that can be mined for useful information. Decreases in the cost of both storage and compute power have made it feasible to collect this data which would have been thrown away only a few years ago. As a result, more and more companies are looking to include non-traditional yet potentially very valuable data with their traditional enterprise data in their business intelligence analysis. Every day, we create 2.5 quintillion bytes of data so much that 90By now, the term of big data is mainly used to describe enormous datasets, a new fact of business life, one that requires having strategies in place for managing large volumes of both structured and unstructured data. This data comes from everywhere: sensors and videos, purchase transaction record, and cell phone GPS signals to name a few [3]. Compared with traditional datasets, big data [5] typically includes masses of unstructured data that need more real-time analysis. In addition, big data also brings about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and also incurs new challenge, e.g, how to effectively organize and process such datasets (e.g. datasets

M. Núñez et al. (Eds.): ICCCI 2015, Part II, LNCS 9330, pp. 638–647, 2015.
DOI: 10.1007/978-3-319-24306-1_62

of satellite images) The rest of this paper is organized as follows. In Section 2, we present the definition of big data and its four features. Then, in Section 3 we introduce the related technologies. Section 4 focuses on the big data value chain (which is composed of five phases).Section 5 examine the several representative applications of big data. A brief conclusion with recommendations for future studies is presented in Section 6.

2 The era of Big Data

In this section, we first present a list of popular definitions of big data, followed by a definition of the 4V.

2.1 What Is Big Data

There is no universal definition of what constitutes "Big Data". In fact, several definitions for big data are found in the literature:

- IDC [4] Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.
- McKinsey Global Institute [6] "Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data–i.e., we don't define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).
- Gartner [7] Big Data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.
- Oracle [8] big data is the data characterized by 4 key attributes: volume, variety, velocity and value.
- IBM [9] big data is the data characterized by 3 attributes: volume, variety and velocity. Big data is an abstract concept. Apart from masses of data, it also has some other features, which determine the difference between itself and "massive data" or "very big data."

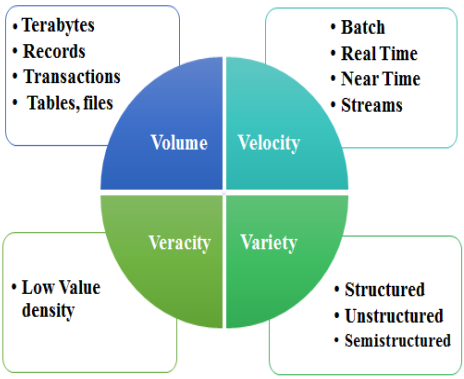

Fig. 1. The 4Vs features of Big Data

2.2 Features of Big Data

Big data is characterized by 4Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily [9]. IBM [4] offers a good and simple overview for the four critical features of big data :

- Big data is always large in volume. It actually doesn't have to be a certain number of petabytes to qualify. If your store of old data and new incoming data has gotten so large that you are having difficulty handling it, that's big data.
- Velocity or speed refers to how fast the data is coming in, but also to how fast you need to be able to analyze and utilize it. If we have one or more business processes that require real-time data analysis, we have a velocity challenge. Solving this issue might mean expanding our private cloud using a hybrid model that allows bursting for additional compute power as-needed for data analysis.
- Variety points to the number of sources or incoming vectors leading to your databases. That might be embedded sensor data, phone conversations, documents, video uploads or feeds, social media, and much more. Variety in data means variety in databases.
- Veracity is probably the toughest nut to crack. If we can't trust the data itself, the source of the data, or the processes you are using to identify which data points are important, we have a veracity problem. One of the biggest problems with big data is the tendency for errors to snowball. User entry errors, redundancy and corruption all affect the value of data.

2.3 Big Data Challenges

As big data wends its inextricable way into the enterprise, information technology (IT)practitioners and business sponsors alike will bump up against a number of challenges that must be addressed before any big data program can be successful.Five of those challenges are [10] [11]:

- Uncertainty of the Data Management : There are many competing technologies, and within each technical area there are numerous rivals. Our first challenge is making the best choices while not introducing additional unknowns and risk to big data adoption.
- Transparency: Making data accessible to relevant stakeholders in a timely manner.
- Experimentation: Enabling experimentation to discover needs, expose variability, and improve performance. As more transactional data is stored in digital form, organizations can collect more accurate and detailed performance data.
- Decision Support: Replacing/supporting human decision making with automated algorithms which can improve decision making, minimize risks, and uncover valuable insights that would otherwise remain hidden.
- Innovation: Big Data enables companies to create new products and services, enhance existing ones, and invent or refine business models.

3 Related Technologies

In order to gain a deep understanding of big data, this section will introduce several fundamental technologies that are closely related to big data, including cloud computing, MapReduce and Hadoop. Big data analytics is often associated with cloud computing because the analysis of large data sets in real-time requires a platform like Hadoop to store large data sets across a distributed cluster and MapReduce to coordinate, combine and process data from multiple sources.

3.1 Cloud Computing for Big Data

Cloud computing [12] is closely related to big data. Big Data is all about extracting value out of "Variety, Velocity and Volume" from the information assets available, while Cloud focuses on On-Demand, Elastic, Scalable, Pay-Per use Self Service models .Big Data need large on-demand compute power and distributed storage to crunch the 3V data problem and Cloud seamlessly provides this elastic on-demand compute required for the same. With the "Apache Hadoop", the de-facto standard for Big Data processing, the big data processing has been more batch oriented in the current state. The burst workload nature of the Big Data Computing Infrastructure makes it a true case for the Cloud. Big data is the object of the computation-intensive operation and stresses the storage capacity of a cloud system. The main objective of cloud computing [9] is to use huge computing and storage resources under concentrated management, so as to provide

big data applications with fine-grained computing capacity. The development of cloud computing provides solutions for the storage and processing of big data. On the other hand, the emergence of big data also accelerates the development of cloud computing. The distributed storage technology based on cloud computing can effectively manage big data; the parallel computing capacity by virtue of cloud computing can improve the efficiency of acquisition and analyzing big data.

3.2 Hadoop/MapReduce for Big Data

Presently, Hadoop [13] is widely used in big data applications in the industry, e.g., spam filtering, network searching, clickstream analysis, and social recommendation. In addition, considerable academic research is now based on Hadoop. Some representative cases are given below. As declared in June 2012, Yahoo runs Hadoop in 42,000 servers at four data centers to support its products and services, e.g., searching and spam filtering, etc. At present, the biggest Hadoop cluster has 4,000 nodes, but the number of nodes will be increased to 10,000 with the release of Hadoop 2.0. In the same month, Facebook announced that their Hadoop cluster can process 100 PB data, which grew by 0.5 PB per day as in November 2012. In addition, many companies provide Hadoop commercial execution and/or support, including Cloudera, IBM, MapReduce, EMC, and Oracle. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

4 Big Data Architecture

In this section we will focus on the value chain of big data [14] [9], which can be generally divided into five phases: data sources, data collection, data management, data analysis and data visualization.

4.1 Data Sources

Data sources is the first step of big data pipeline.Success for a big data strategy lies in recognizing the different types of big data sources [3], using the proper mining technologies to find the treasure within each type, and then integrating and presenting those new insights appropriately according to your unique goals, to enable the organization to make more effective steering decisions. The different sources of information includes [15] network usage, sensors, connected devices, mobile devices, worldwideweb applications and social networks.

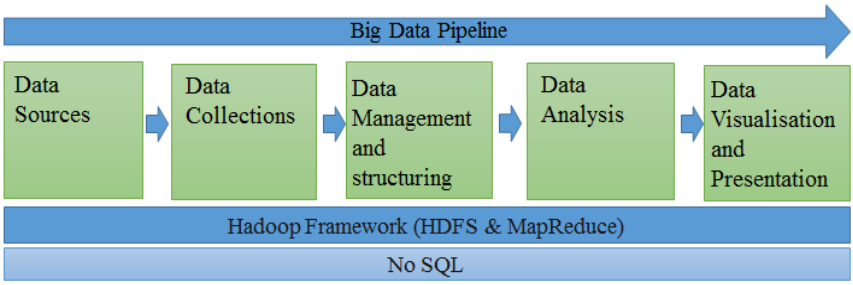

Fig. 2. The Big Data Value Chain

4.2 Data Collection

The second phase of the big data system is big data collection. Data collection refers to the process of retrieving raw data from real-world objects. The process needs to be well designed.Otherwise,inaccurate data collection would impact the subsequent data analysis procedure and ultimately lead to invalid results. At the same time, data collection methods not only depend on the physics characteristics of data sources, but also the objectives of data analysis. As a result, there are many kinds of data collection methods. In the subsection, we will focus on two common methods for big data collection.

- Log File: Log files [16], one of the most widely deployed data collection methods, are generated by data source systems to record activities in a specified file format for subsequent analysis. Log files are useful in almost all the applications running on digital devices. There are three main types of web server log file formats available to capture the activities of users on a website: Common Log File Format (NCSA), Extended Log Format (W3C), and IIS Log Format (Microsoft). All three log file formats are in the ASCII text format. Alternatively, databases can be utilized instead of text files to store log information to improve the querying efficiency of massive log repositories.Otherexamples of log file-based data collection include stock ticks in financial applications, performance measurement in network monitoring, and traffic management.
- Sensors [17] are used commonly to measure a physical quantity and convert it into a readable digital signal for processing. Sensor types include acoustic, sound, vibration, automotive, chemical, electric current, weather, pressure, thermal, and proximity. Through wired or wireless networks,this information can be transferred to a data collection point. Wired sensor networks leverage wired networks to connect a collection of sensors and transmit the collected information.

4.3 Data Management and Structuring

The data source developer makes deliberate organizational choices about the datas syntax, structure, and semantics and makes that information available

either from schemata or from a metadata repository [15] [18]. Either mechanism can provide the basis for tracking the shared semantics needed to organize the data before integrating it. Metadata repositories are commercially available, and numerous generic metamodels exist, many of which rely on Extensible Markup Language Metadata Interchange (XMI). However, because of XMIs generality, each tool provides customized extensions, which can lead to vendor lock, problems sharing schemata among participants, and other tool-interoperability issues. Analysts often skip formal data organization because theyre more focused on their own data needs than on considering how to share data. However, sharing knowledge about internal data organization can enable more seamless integration with data providers environments (upstream) and data consumers environments (downstream).

4.4 Data Analysis

The most important stage of the big data value chain is data analysis [18]. Data sources are then ready for analysis, which includes maintaining the provenance between the input and results and maintaining metadata so that another analyst can recreate those results and strengthen their validity. Popular data analysis techniques, such as MapReduce, enable the creation of a programming model and associated implementations for processing and generating large datasets. Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software [19] and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop [13] and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems.

4.5 Data Visualization

Visualization [20] involves presenting analytic results to decision makers as a static report or an interactive application that supports the exploration and refinement of results. The goal is to provide key stakeholders with meaningful information in a format that they can readily consume to make critical decisions. Industries, such as media and training, have a wealth of data visualization techniques, which others could adopt. Virtual and augmented realities, for example, enhance the user experience and make it easier to grasp information thats elusive in two-dimensional media. Although this technology has promising implications,

virtual and augmented reality systems continue to be viewed as only suitable for training, education, and other highly customized uses. Big data environments help organizations capture, manage, process and analyze large amounts of data from both new data formats, as well as traditional formats, in real time. When managing the data lifecycle of big data, organizations should consider the volume, velocity and complexity of big data.

5 Big Data Applications

In this section we present a diverse and a representative set of applications dealing with big data.

5.1 Social Media Applications

IBM estimates that 2.5 quintillion bytes of new data are created every day. To put this into perspective, social media alone generates more information in a short period of time than existed in the entire world just several generations ago. Popular sites [21] like Facebook, Instagram, Foursquare, Twitter, and Pinterest create massive quantities of data that if translated properly by large-scale applications would be any brands golden ticket into the minds of its consumers. Unfortunately, the data produced by social media is not only enormous its also unstructured. The task of capturing, processing, and managing this data is unquestionably beyond human scale. In fact, its beyond the scale of most common software. Because of this, a glass wall exists between marketers and the data they can see it, but they cant harness it. Its easy to see how Big Data fits into the picture. The Big Data industry deals in sets of data that range from a few dozen terrabytes to many hundreds of petabytes. A slew of Big Data applications have been created specifically to make sense of social media data.

5.2 Enterprise Applications

Theres no doubt Big Data is changing how companies are able to process and use data.The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) [19] programs. That could include Web server logs and Internet clickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and For rester Research Inc. also consider transactions and other structured data to be valid components of big data analytics applications [22].

5.3 Science Applications

The sciences have always been big consumers of Big Data, with the added challenge of analyzing disparate data from multiple instruments and sources. For example, astronomy involves continuously collecting and analyzing huge amounts of image data, from increasingly sophisticated telescopes. The next big science astronomy project is the Large Synoptic Survey Telescope (LSST). The telescope will ultimately collect about 55 Petabytes of raw data. There is a streaming pipeline where software looks for patterns in the images (e.g., stars and other celestial objects) and then looks for the same object in different telescope images to obtain trajectory information. All of the data will be stored, and astronomers want to reprocess the raw imagery with different algorithms, as there is no universal image-processing algorithm that pleases all astronomers. New Big Data technologies can provide the capacity to store, process, analyze, visualize and share large amounts of image data, as well as remote sensing data from satellites [23]. Astronomy looks out to the sky from telescopes; remote sensing looks in towards the earth from satellites. The two are roughly mirror images of each other, presenting an opportunity for new approaches to analyzing both kinds of data simultaneously.

6 Conclusion

The Big data concept has progressively become the next evolutionary phase in batch processing, storing, manipulation and relations visualization in vast number of records. The era of big data is upon us, bringing with it an urgent need for advanced data acquisition, management, and analysis mechanisms. In this paper, we have presented the concept of big data, challenges, related technologies, applications and highlighted the big data value chain, which covers the entire big data lifecycle. The big data value chain consists of five phases: data sources, data collection, data management, data analysis and data visualization. We regard Big Data as an emerging trend in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at realtime. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

References

1. Tsuchiya, S., Sakamoto, Y., Tsuchimoto, Y., Lee, V.: Big data processing in cloud environments. FUJITSU Science and Technology J. **48**(2), 159–168 (2012)
2. Big Data: Science in the Petabyte Era: Community Cleverness Required. Nature **455**(7209) (2008)
3. What is Big Data. IBM, New York (2013). http://www-01.ibm.com/software/data/bigdata/

4. Gantz, J., Reinsel, D.: The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. In: Proc. IDC iView, IDC Anal. Future (2012)
5. Mayer-Schonberger, V., Cukier, K.: Big data: a revolution that will transform how we live, work, and think. Eamon Dolan/Houghton Mifflin Harcourt (2013)
6. Manyikaetal, J.: Bigdata: The Next Frontier for Innovation, Competition, and Productivity, p. 1137. McKinsey Global Institute, San Francisco (2011)
7. Gartner Group, Inc. http://www.gartner.com/it-glossary/big-data/
8. Oracle Big Data. https://www.oracle.com/bigdata/index.html
9. Chen, M., Mao, S., Liu, Y.: Big Data: a survey. MONET **19**(2), 171–209 (2014). Spring science business Media, New York
10. Xindong, W., Zhu, X., Gong-Qing, W., Ding, W.: Data Mining with Big Data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
11. Kong, W., Wu, Q., Li, L., Qiao, F.: Intelligent data analysis and its challenges in big data environment. In: 2014 IEEE International Conference on System Science and Engineering (ICSSE), pp. 108–113 (2014)
12. Gupta, R., Gupta, H., Mohania, M.: Cloud computing and big data analytics: what is new from databases perspective? In: Srinivasa, S., Bhatnagar, V. (eds.) BDA 2012. LNCS, vol. 7678, pp. 42–61. Springer, Heidelberg (2012)
13. The Apache Software Foundation (2014). http://hadoop.apache.org/
14. Han, H., Wen, Y., Chua, T.-S., Li, X.: Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access **2**, 652–687 (2014)
15. Evans, D., Hutley, R.: The explosion of data. white paper (2010)
16. Wahab, M.H.A., Mohd, M.N.H., Hanafi, H.F., Mohsin, M.F.M.: Data pre-processing on web server logs for generalized association rules mining algorithm. World Acad. Sci. Eng. Technol. **48**, 36 (2008)
17. Chandramohan, V., Christensen, K.: A first look at wired sensor networks for video surveillance systems. In: Proceedings LCN 2002, 27th annual IEEE conference on local computer networks, pp. 728–729. IEEE
18. Bhatt, C.A., Kankanhalli, M.S.: Multimedia data mining: state of the art and challenges. Multimedia Tools Appl. **51**(1), 35–76 (2011). [184] G. Blackett (2013)
19. Richardson, J., Schlegel, K., Hostmann, B., McMurchy, N.: Magic quadrant for business intelligence platforms (2012). [Online]
20. Friedman, V.: Data visualization and infographics (2008). http://www.smashingmagazine.com
21. Li, Y., Chen, W., Wang, Y., Zhang, Z.-L.: Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. ACM, pp. 657–666 (2013)
22. Agrawal, D., Bernstein, P., Bertino, E.: Challenges andopportunities withbigdata. Acommunity white paper (2012)
23. Ma, Y., Wu H., Wang, L.: Remote sensing big data computing: challenges and opportunities. Future Generation Computer Systems (2014)