Benefits, Challenges and Tools of Big Data Management

Fernando L. F. Almeida

Faculty of Engineering of the University of Porto, Portugal
almd@fe.up.pt

Abstract: Big Data is one of the most predominant field of knowledge and research that has generated high repercussion in the process of digital transformation of organizations in recent years. The Big Data's main goal is to improve work processes through analysis and interpretation of large amounts of data. Knowing how Big Data works, its benefits, challenges and tools, are essential elements for business success. Our study performs a systematic review on Big Data field adopting a mind map approach, which allows us to easily and visually identify its main elements and dependencies. The findings identified and mapped a total of 12 main branches of benefits, challenges and tools, and also a total of 52 sub branches in each of the main areas of the model.

Key words: Big Data, data management, data analysis, massive data, business intelligence, cloud computing, Hadoop, Map reduce

1.   Introduction

Big Data is the term that describes the large amount of data, which can be structured and unstructured, that affects business. Big Data requires specialized tools for their treatment in order to generate important results that, in smaller volumes, would hardly be achieved. The focus should not be exclusively in dealing with high quantity of data, but on the possibility that these data offer in the intention of creating information and knowledge that can make companies and public entities more competitive, which will let them to offer better services for consumers and citizens.

Big data has been defined in terms of the five v's (Russom, 2017, Chen 2012, Abdullah 2015): volume, velocity, variety, veracity and value. The volume is the quantity of data that can be stored and managed; the velocity is the speed of calculation needed to query the data relative to the rate of change of the data; the variety measures the number of different formats of data (e.g, text, audio, video, etc.); the veracity refers to the messiness or the trustworthiness of the data; and the value is the importance given by companies/entities to access this data.

The process of working with a large volume of data offers numerous advantages as it allows us to be more rigorous in the decisions. However, and in parallel, it is necessary that its increase in terms of volume is also accompanied by an improvement in data quality. Several challenges emerge in various dimensions. From the technological point of view, it is important that technologies evolve and be able to handle with huge amount of data in a fast and reliable way. On the business side, it becomes necessary for companies to be able to filter the information that is really important from what is not. Finally, on the social side, it is important to ensure security and confidentiality in the handling of information.

This study aims to identify and synthesize the main key benefits, challenges and tools in Big Data environments. For that, we initially perform a revision of literature in the field of Big Data. After that, we present the adopted methodology, we draw the mind maps that visually represent the most relevant benefits, challenges and tools of Big Data, and we discuss the main results. Finally, we draw the conclusions of our work and we give insights about future research directions that we expect to get with the evolution of Big Data.

2.   Literature Review

Big Data has been a very emerging area of study and research. The main literature works have been published in the last 5 years. There are some literature reviews in the field that focus on aspects such as data quality (Abdullah 2015, Tsai 2015), data storage (Acharjya, 2014), data integration [Chen, 2014], data analysis [Chen, 2014, Hammad, 2015, Padgavankar 2014], knowledge discovery [Hammad, 2015, Padgavankar 2014, Tsai, 2015), scalability (Phule, 2013), and visualization (Acharjya, 2014). These studies generically analyze the various concepts and stages of Big Data management, focusing in greater detail on some of its dimensions.

The benefits of using big data techniques are quite wide. Two main groups of benefits emerge (Mohan, 2016): (i) cost savings; and (ii) competitive advantage. In terms of cost savings Big Data tools allow businesses to store massive volumes of data at a much cheaper price tag than a traditional database. Furthermore, Big Data offers competitive advantage for businesses by offering them the possibility of exploring new business opportunities. In fact, new products, services, and even business models can emerge from analysis of Big Data (Manyika, 2011). Executives also state the top five areas that have benefited from the adoption of Big Data are (Porres, 2017): (i) increasing insights into consumer behavior; (ii) increasing sales; (iii) increasing sign-ups and registrations; (iv) increasing Return on Investment (ROI); (v) increasing customer satisfaction; and (vi) increasing sales leads.

The potentialities of Big Data depend on the sectors of activity where it is used. Five examples can be given (Ularu, 2012): (i) in information technology in order to improve the security; (ii) in customer service in call centers to enhance customer satisfaction; (iii) in retail by the use of social media to understand customer preferences; (iv) in banks to detect fraud in online transactions; and (v) in the financial market to analyze and classify risk assessment. Significant advances have been registered in science by the adoption of Big Data, particularly in astronomy, biology and bioinformatics (Oguntimilehin, 2014). Additionally, Big Data can be used in education. Students and teachers can afford to shape a modern and dynamic education system. The list of benefits includes (Drigas, 2014): (i) improved instruction; (ii) matching students to programs; (iii) matching students to employment; (iv) transparent education financing; and (v) efficient system administration. Finally, there are also studies that use case studies to demonstrate the benefits of Big Data. In the field of marketing, it was reported benefits in cross-sell and up-sell, reduction in churn, increased customer experience and better customer service between all channels of the company (Moorthy, 2015). In the government field, it was mentioned the high potential of Big Data adoption, through the integration of data from multiple applications and also with the evolution of the Internet of Things (IoT) technology (Moorthy, 2015).

The adoption of Big Data in companies, regardless of their size, has been one of the biggest implementation challenges today. There are still many companies that are struggling to enter into this new world of information and others are already enjoying the technology, but still in a limited and restricted way. The integration, manipulation, quality and governance of Big Data emerge as key points that should be considered when building a Big Data management solution (Kaur, 2017) Four groups of challenges in the Big Data analysis were identified (Acharjya, 2014): (i) data storage and analysis; (ii) knowledge discovery and computational complexities; (iii) scalability and visualization of data; and (iv) information security. Other challenges were also identified by other authors, such as: heterogeneity and incompleteness (Lawal, 2016, Satyanarayana 2015); high-dimensional data (Najafabadi, 2015); large-scale models (Najafabadi, 2015); failure handling (Lawal, 2016, Satyanarayana 2015); energy management; and human resources and manpower (Satyanarayana 2015).

The challenges of Big Data are not only at technical levels. A key challenge that emerges is to provide appropriate data processing solutions for effective and efficient integration of data and process management and appropriate analysis tools (Wulff, 2016). In fact, every department within a company will need to make some adjustments. Big Data needs to be an integrated part of the business process, instead of a distinct function performed only by well qualified and trained specialists (Almeida, 2013). This vision is also confirmed by statistics, which indicate that until 2020 around 84% of businesses believe that data will be an integral part of forming business strategy and also 77% believe that data management will be driven by multiple stakeholders in their organization, rather than by a single data specialist (Carmody, 2016)

The emergence of new technologies and the adhesion to Big Data by more and more organizations, causes the appearance of emerging challenges in this field. At this level, study (Acharjya, 2015) identifies four areas of research issues in the Big Data analysis: (i) IoT for Big Data analytics; (ii) cloud computing for Big Data analysis; (iii) bio-inspired computing for Big Data analysis; and (iv) quantum computing for Big Data analysis. Clearly within these themes, the area of cloud computing has been the most explored. Cloud computing offers groups of servers, storages and various networking resources that can be exploited by Big Data analytics. Therefore, cloud computing appears as an efficient way to increase productivity while reducing cost to process huge amount of data (Assunção, 2016, Gupta, 2015, Yang, 2017).

One of the essential aspects relates to the process of choosing a Big Data tool. The study (Singh, 2014) establishes six relevant criteria: (i) scalability; (ii) data I/O performance; (iii) fault tolerance; (iv) real-time processing; (v) data size supported; and (vi) iterative task support. Big Data tools can be categorized into two groups (Prasad, 2016): (i) computing tools; and (ii) storage tools. In the former

group we may find frameworks as Hadoop Map Reduce, Cloudera Impala, IBM Netezza or Apache Giraph. In the latter group we have applications such as HBase, Apache Hive, Cassandra or Neo4j. Hadoop is one of the most adopted tools at this level and we may find several studies that apply this framework in the exploitation and distribution of large data volumes (Jeysudha, 2017, Sharmila , 2016, Gade, 2016). This framework is characterized by two fundamental components (Bhosale, 2014): HDFS architecture and MapReduce architecture. Associated with Big Data tools there are some other technologies that have an important role. Among them, it is relevant to highlight the JSON and Machine-to-Machine technologies (Trifu, 2014, Zan, 2015).

Big Data has been used in a large set of fields. The top 10 industry verticals are (Sivakumar, 2015): (i) banking and securities; (ii) communications, media & entertainment; (iii) healthcare providers; (iv) education; (v) manufacturing and natural resources; (vi) government; (vii) insurance; (viii) retail and wholesale trade; (ix) transportation; and (x) energy and utilities. Several research studies may be found in these domains, particularly in the healthcare industry (Raghupathi, 2014, Bains, 2016). Finally, the popularity of social media has also attracted a strong interest from the scientific and business communities. Studies on how to explore data from various social networks and the semantic analysis of this information have arisen in recent years (Olshannikova, 2017).

3.   Methodology

A systematic review approach was adopted. This methodology allows us to synthesize existing work and identify the most relevant studies in a comprehensive manner. The adopted strategy intends to completeness our research in the domain of Big Data by looking to its benefits, challenges and tools. The main advantages of systematic literature reviews are (Kitchenham, 2009): (i) avoid that the results of the literature are biased; (ii) systematic reviews provide evidence that the phenomenon is robust and transferable when studies give consistent results; and (iii) in the case of quantitative studies, it is possible to combine data using meta-analytic techniques.

We followed a systematic review approach composed of five steps (Khan, 2003)
  o Step 1: Framing questions for a review - the problems to be addressed by the review should be clearly and unambiguously identified;
  o Step 2: Identifying relevant work - the search for studies should be extensive. Multiple resources should be used;
  o Step 3: Assessing the quality of studies - select studies should be subjected to a more refined quality assessment. The studies should help us in assessing the strength of inferences and making recommendations for future work;
  o Step 4: Summarizing the evidences - data synthesis consists on identifying the evidences that can be confirmed by several studies;
  o Step 5: Interpreting the findings - the issues highlighted in each of the four steps above should be met. Exploration of heterogeneity should help determine whether the overall summary can be trusted, and, if not, the effects observed in high-quality studies should be used for generating inferences.

In the context of our study, we synthesize the adopted five steps in the Table 1.

Table 1: Description of the steps of the adopted methodology

Methodology steps	Description
Step 1: Framing questions for a review	Identification of the most relevant benefits, challenges and tools offered by Big Data.
Step 2: Identifying relevant work	The most relevant issues published in the last four years and indexed in Web of Science, Scopus or Google Scholar were identified and analyzed.
Step 3: Assessing the quality of studies	Each study was deeply analyzed in order to identify relevant benefits, challenges and tools of Big Data. Studies that don't add significant value or that don't focus on the three identified themes (see Step 1) were not considered.
Step 4: Summarizing the evidences	The evidences were summarized using a mind map. The mind map is composed by major branches and sub branches. Two levels of depth were used.
Step 5: Interpreting the findings	The findings were analyzed and debated considering the consistency of the studies and their impact in scientific, business and social terms.

We start by presenting the mind map of the benefits offered by Big Data in Fig. 1. We do not intend to list exhaustively all the elements in each of the analysis dimensions, but only those that deserve greater relevance and acceptance of the scientific community.

The benefits were classified into three groups: (i) technology; (ii) financial; and (iii) competitive advantage. A total of 17 sub branches was also identified.
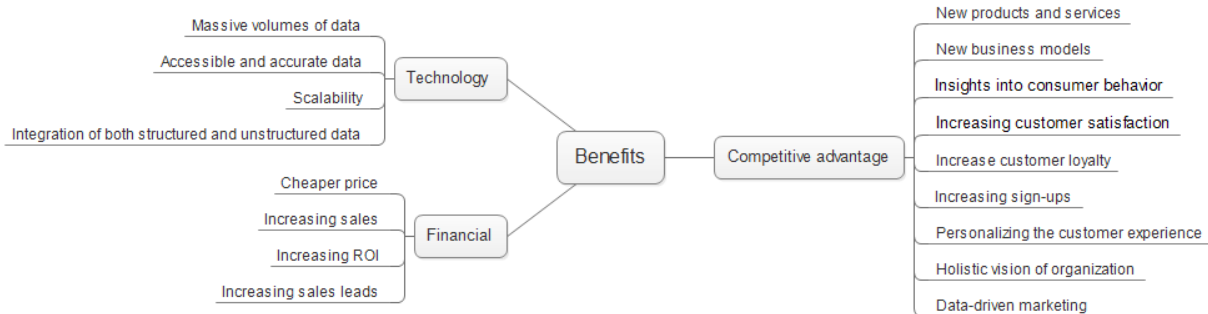
Fig. 1: Mind map representation of Big Data benefits

The challenges posed by the adoption of Big Challenges are depicted in Fig. 2. The challenges were grouped into six domains: (i) data storage and analysis; (ii) scalability and data visualization; (iii) knowledge discovery; (iv) information security; (v) human resources and manpower; and (vi) appearance of new technologies. A total of 20 sub branches was also identified.
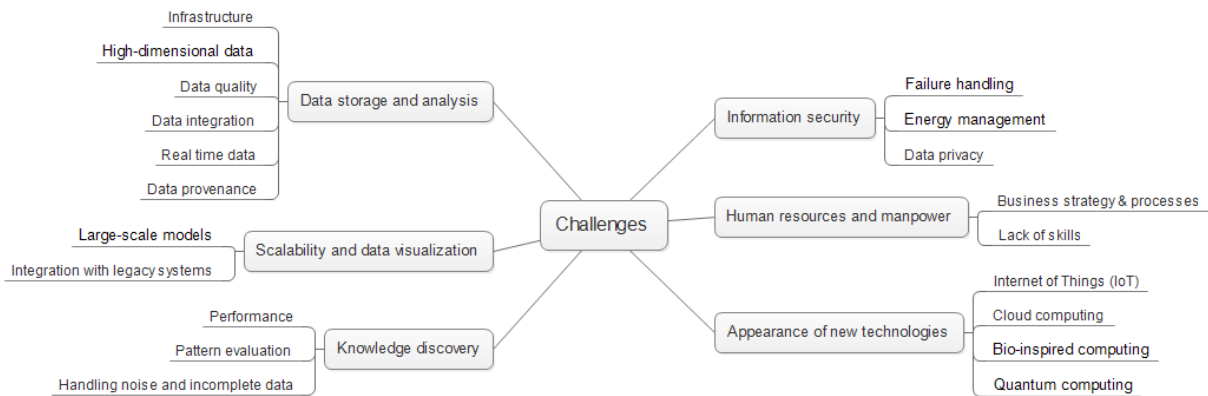
Fig. 2: Mind map representation of Big Data challenges

Finally, the Big Data tools are presented in Fig. 3. They were categorized into three groups: (i) computing tools; (ii) storage tools; and (iii) support technologies. A total of 15 sub branches was also identified.
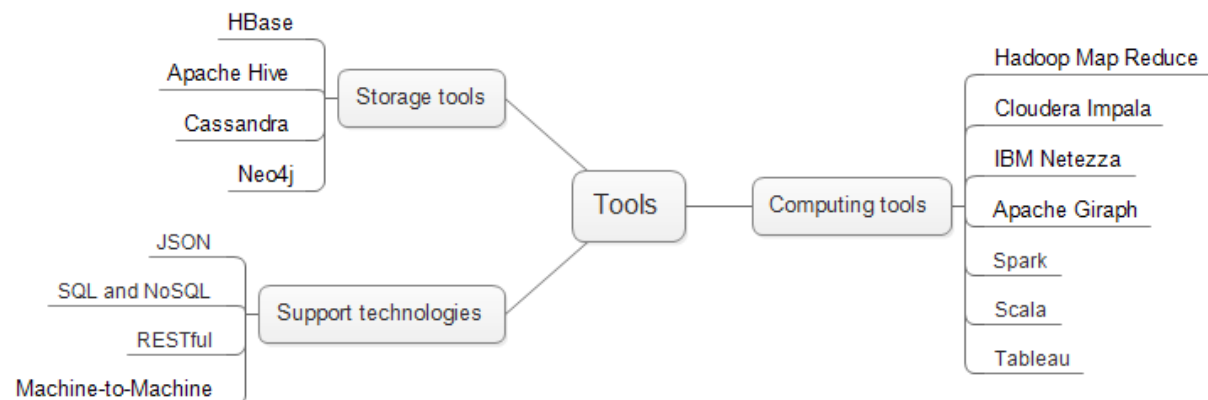
Fig. 3: Mind map representation of Big Data tools

4. Results and Discussion

4.1. Big Data Benefits

At the technological level, we can find benefits by the dealing of massive volumes of data, accessible and accurate data, scalability, and integration of both structured and unstructured of data. In fact, an essential characteristic of Big Data is to deal with huge amount of structured and unstructured data. The computing models and high volume data processing have led storage systems to evolve with high performance, high efficiency and scalability solutions. Therefore, there are storage alternatives based on horizontal scalability or scale-out. This type of scalability is most appropriate in situations where it is difficult to predict changes in storage needs rather than acquiring anticipated storage to meet short-term demand, capacity can be increased when necessary. By adding nodes that work in parallel, performance, capacity and throughput can be increased, since each node has its own storage and processing capacity. In this way, the space and computing power are increased simultaneously.

The financial benefits offered by Big Data are one of the most obvious advantages. Large volumes of storage space are available at cheaper prices. The companies can process more date for the same price, which will increase their offer in the market. Therefore, they potentially can increase the total amount of sales, sales leads and ROI.

Finally, a large panoply of competitive advantages can be reached by companies. A total of nine benefits was identified in this segment: new products and services, new business models, insights into consumer behavior, increasing customer satisfaction, increase customer loyalty, increasing sign-ups, personalization the customer experience, holistic vision of the organization and data-driven marketing. In fact, some of them are correlated. For instance, when we personalize the customer experience, we expect to have a more insight into consumer behavior and better customer satisfaction, which will potentially increase the customer loyalty. The last two identified benefits deserve a better analysis.

Big Data contributes to a holistic vision of the organization. In fact, traditional organizational models cause that information is kept in silos, both at departmental (e.g., sales division, marketing division, HR department) and technological (e.g., ERP, CRM, email marketing systems, social analytics). The use of Big Data ensures that all information is integrated and explored, and can be used later fairly by all departments of the company.

The data-driven marketing, as the name suggests, is the marketing oriented by the availability of data. The strategic marketing decisions are based on information about the customer. Data-driven marketing enables more assertive and fully measurable actions. Because marketing is customer-centric, it is possible to identify the customer profile and send them a communication policy with the right approach at the right time. Data-driven marketing is the basis of machine learning and predictive marketing, which is supported on the premise of a high quantity and quality of data. These are two fundamental elements for the adoption of Big Data.

4.2. Big Data Challenges

Data storage and analysis is still one of the main concerns in Big Data environments. At this level, we may find challenges in terms of hardware infrastructure, high-dimensional data, data quality, data integration, real-time data and data provenance. The last two issues deserve a careful analysis due to their impact in all Big Data process. Currently business decisions have to be made in real time. Data collection devices are increasingly heterogeneous, from mobile devices, sensors and social media that produce massive amounts of changing data. To get the highest possible value from these data, companies need to process data and make decisions much faster. On the other side, the data provenance refers to the metadata used to identify information sources and transformation processes. This information is extremely relevant particularly at the granularity level. A lower level of granularity implies a higher cost of collection and storage. However, a higher degree of granularity may lead to a small variety of queries that can be answered and, thus, condition the potential of Big Data.

The scalability and data visualization is another challenge. Large scale scalability models that promote horizontal scalability of processing and storage over a variety of different infrastructures are required. For its part, integration with legacy systems aims to combat the isolation effect of a system and monetize old equipments, analytical processing algorithms and increase data availability. One of the emerging areas is the use of legacy systems in hybrid clouds (Shein, 2017).

The discovery of knowledge is one of the Big Data's goals. In this field, we can find challenges in terms of performance, patter evaluation, and handling noise and incomplete data. Several knowledge

discovery algorithms are well known and tested in the field of data mining. The question that arises when using them in Big Data environments is whether these same algorithms offer acceptable performance on Big Data environment. The algorithms that offer best conditions are those that can be more easily distributed and take advantage of cloud computing.

The risks associated with information security are typically considered as one of the main challenges of cloud computing. The biggest security flaws are the lack of authentication mechanisms, as well as the lack of use of secure channels for access to information, such as the use of encryption. These situations lead to vulnerabilities in the exclusion, inclusion or modification of information. In addition to these aspects, also appear security issues related to the use of the network, namely in terms of the implementation of mechanisms of failure handling and energy management. Consequently, one of the challenges is to ensure a high rate of system operation.

Another challenge is about people. Big Data cannot be seen exclusively as a new technology, but must be integrated as a transversal process between the various divisions and departments of an organization. In addition, the company needs to have qualified human resources, emerging the role of the data scientist in organizations. A data scientist is a professional with multidisciplinary skills in areas such as data engineering, scientific method, math, statistics, advanced computing, visualization, hacker mindset and domain expertise. This new job is currently considered one of the activities with greatest growth potential in information technologies in the 21st century (Tharwat, 2017).

Finally, the appearance of new technologies brings new challenges. Cloud computing is one technology that is always associated with Big Data due to its potentialities to increase the processing capacity with low costs. Other technologies have also begun to emerge in the market, but whose effects on Big Data are more uncertain, like the Internet of Things (IoT), bio-inspired computing or quantum computing.

4.3. Big Data Tools

Computing tools for Big Data are fundamental tools that can be used to process Big Data at different levels. They typically provide a fast engine for Big Data analysis, integrating different processing techniques, such as machine learning and graph processing. One of the most popular tools is the Hadoop MapReduce that is a software framework for distributed processing of large data sets on computer clusters of commodity hardware. Other tools allow to fulfill other goals. For instance, Scala is an object-oriented language that is particularly suitable for pattern matching; Apache Giraph is an extension of the Hadoop's MapReduce framework to perform graph processing on Big Data; and Tableau is a business intelligence tool that can be used to create reports, charts, graphs and dashboards.

Storage tools in Big Data have a double purpose. They offer an infrastructure on which is possible to run analytics tools and, simultaneously, a place to store and query Big Data. The most relevant variables in choosing a Big Data storage tool include the existing environment, current storage platform, growth expectations, size and type of files, database and application mix (Robb, 2017).

Finally, it isn't possible to process Big Data without the contribution of other languages and technologies. Machine-to-machine offers great possibilities in BIg Data environments and refers to direct communication between devices using any communications channel, including wired and wireless. JSON is a universal format that is very convenient for exchanging information between applications through various protocols. RESTful is an API that allows the communication between a web-based client and server that employs representational state transfer (REST). Finally, SQL and NoSQL offer mechanisms for storage and retrieval data. NoSQL databases are suitable for data that is constantly evolving or frequently changing (Palovská, 2015) hey are increasingly used in big data and real-time web applications, due to their simplicity of design and scalability.

5.   Conclusions

In the current organizational context, it is no longer enough to have technological tools that collect, consolidate, analyze and ingrate data into the decision process. It is also necessary to have technological resources that can boost them, perceive the business and know how to extract information relevant to the development of the organization. Big Data analytics are currently a tool that helps organizations harness their data and use it to identify new opportunities.

The systematic review approach allowed us to identify and synthesize the main benefits, challenges and tools in Big Data. The benefits offered by Big Data were grouped into three domains: (i) technology; (ii) financial; and (iii) competitive advantage. On the other side, the challenges found include: (i) data storage and analysis; (ii) scalability and data visualization; (iii) knowledge discovery; (iv) information security; (v) human resources and manpower; and (vi) appearance of new technologies. Finally, the big data tools were categorized into three groups: (i) computing tools; (ii) storage tools; and (iii) support technologies. In addition, we have mapped a total of 52 sub branches for all the three main categories.

As future work, it should be noted that this systematic review work must necessarily be revised in the face of technological developments that will arise in the Big Data field. In fact, Big Data has been an area where the emergence of new technologies has been constant, turning existing models quickly obsolete. In light of this, new technological solutions with greater speed, processing and storage capacity are created, which brings with them new benefits, but also new technological, business and organizational challenges.

6. References

Abdullah, N., Ismail, S., Sophiayati, S., & Sam, S., 2015: Data Quality in Big Data: A Review. *International Journal of Advances in Soft Computing and its Applications (IJASCA)* 7(3), pp. 16-27

Acharjya, D. & Ahmed, K., 2016: A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. *International Journal of Advanced Computer Science and Applications (IJACSA)* 7(2), pp. 511-518

Almeida, F. & Calistru, C., 2013: The main challenges and issues of big data management. *International Journal of Research Studies in Computing* 2(1), pp. 11-20

Assunção, M., Calheiros, R., Bianchi, S., Netto, M., & Buyya, R., 2015: Big Data computing and clouds: trends and future directions. *Journal of Parallel and Distributed Computing* 79-80, pp. 3-15

Bains, J., 2016: Big Data Analytics in Healthcare - Its Benefits, Phases and Challenges. *International Journal of Advanced Research in Computer Science and Software Engineering* 6(4), pp. 430-435

Bhosale, H. & Gadekar, D., 2014: A review paper on Big Data and Hadoop. *International Journal of Scientific Research Publications* 4(10), pp. 1-7

Carmody, B., 2016: *Biggest problem with Big Data Management in 2016*. Available at: https://www.inc.com/bill-carmody/biggest-problem-with-big-data-management-in-2016.html [Accessed 26 June 2017]

Chen, H., Chiang, R., & Storey, V., 2012: Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36(4), pp. 1165-1188

Chen, M., Mao, S., & Liu, Y., 2014: Big Data: A Survey. *Mobile Networks and Applications* 19, 171-209.

Drigas, A. & Leliopoulos, P., 2014: The Use of Big Data in Education. *International Journal of Computer Science Issues* 11(5), pp. 58-63

Gade, S., Pathan, A., Tomar, S., & Razdan, S., 2016: Big Data on Cloud using Hadoop. *Imperial Journal of Interdisciplinary Research* 2(7), pp. 255-257

Gupta, A., Mehrotra, A., & Khan, P., 2015: Challenges of Cloud Computing & Big Data Analytics. In: *2nd International Conference on Computing for Sustainable Global Development*, 2015, New Delhi, India

Hammad, K., Fakharaldien, M., Zain, J., & Majid, M., 2015: Big Data Analysis and Storage. In: *International Conference on Operations Excellence and Service Engineering*, September 10-11. 2015, Orlando, Florida, USA

Jeysudha, A., Muthukutty, L., Krishnan, A., & Shivadekar, S., 2017: Real Time Video Copy Detection using Hadoop. *International Journal of Computer Applications* 162(9), pp. 42-45

Kaur, P. & Monga, A., 2016: Managing Big Data: A Step towards Huge Data Security. *International Journal of Wireless and Microwave Technologies* 2, pp. 10-20

Khan, K., Kunz, R., Kleijnen, J., & Antes, G., 2003: Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine* 96(3), pp. 118-121

Kitchenham, B., 2007: *Guidelines for performing systematic literature reviews in software engineering*. EBSE Technical Report, EBSE-2007-01, Keele University and University of Durhan

Lawal, Z., Zakari, R., Shuaibu, M., & Bala, A., 2016. A Review: Issues and Challenges in Big Data from Analytic and Storage Perspectives. *International Journal of Engineering and Computer Science* 5(3), pp. 15947-15961

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A., 2011. *Big Data: The next frontier for innovation, competition, and productivity*. Available at: http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation [Accessed 23 June 2017]

Mohan, A., 2016: Big Data Analytics: Recent Achievements and New Challenges. *International Journal of Computer Applications Technology and Research* 5(7), pp. 460-464

Moorthy, J., Lahiri, R., Sanyal, D., & Nanath, K., 2015: Big Data: Prospects and Challenges. *The Journal of Decision Makers* 40(1), pp. 74-96

Najafabadi, M., Villanustre, F., Khoshgoftaar, T., Seliya, N., Wald, R., & Muharemagic, E., 2015: Deep learning applications and challenges in Big Data analytics. *Journal of Big Data* (2), pp. 1-21

Oguntimilehin, A. & Ademola, E., 2014: A Review of Big Data Management, Benefits and Challenges. *Journal of Emerging Trends in Computing and Information Sciences* 5(6), pp. 434-438

Olshannikova, E., Olsson, T., Huhtamäki, J., Kärkkäinen, H., 2017: Conceptualizing Big Social Data. *Journal of Big Data* 4(3), pp. 1-19.

Padgavankar, M. & Gupta, D., 2014: Big Data Storage and Challenges. *International Journal of Computer Science and Information Technologies* 5(2), pp. 2218-2223

Palovská, H., 2015: What can NoSQL serve an enterprise. *Journal of Systems Integration* 6(3), pp. 44-49

Phule, R. & Ingle, M., 2013: A Survey on Scalable Big Data Analytics Platform. *International Journal of Science and Research (IJRS)* 4(5), pp. 1164-1169

Porres, E., 2013: *The Big Potential of Big Data*. Forbes Insights. Available at: https://images.forbes.com/forbesinsights/StudyPDFs/RocketFuel_BigData_REPORT.pdf [Accessed 23 June 2017]

Prasad, B. & Agarwal, S., 2016: Comparative study of Big Data computing and storage tools: a review. *International Journal of Database Theory and Application* 9(1), pp. 45-66

Raghupathi, W. & Raghupathi, V., 2014: Big Data analytics in healthcare: promise and potential. *Health Information Science and Systems* 2(3), pp. 1-10

Robb, D., 2016: *Top Ten Big Data Storage Tools*. InfoStor. Available at: http://www.infostor.com/backup-and_recovery/top-ten-big-data-storage-tools.html [Accessed on 29 June 2017].

Russom, P., 2011: *TDWI Best Practices Report*. Available at: https://tdwi.org/research/2011/09/~/media/TDWI/TDWI/Research/BPR/2011/TDWI_BPReport_Q411_Big_Data_Analytics_Web/TDWI_BPReport_Q411_Big%20Data_ExecSummary.ashx [Accessed 18 June 2017]

Satyanarayana, L., 2015: A Survey on Challenges and Advantages in Big Data. *International Journal of Computer Science and Technology* 6(2), pp. 115-119

Sharmila, K. & Manickam, S., 2016: Diagnosing Diabetic Dataset using Hadoop and K-means Clustering Techniques. *Indian Journal of Science and Technology* 9(40), pp. 1-5

Shein, E., 2016: *Integrating legacy systems, hybrid clouds: channel partner tips*. Available at: http://searchcloudprovider.techtarget.com/feature/Integration-of-legacy-systems-hybrid-clouds-Channel-partner-tips [Accessed 28 June 2017]

Singh, D. & Reddy, C., 2014: A survey on platforms for Big Data analytics. *Journal of Big Data* 1(8), pp. 1-20

Sivakumar, S., 2015: *How Top 10 Industries Use Big Data Applications*. Data Science Association. Available at: http://www.datascienceassn.org/content/how-top-10-industries-use-big-data-applications [Accessed 26 June 2017]

Tharwat, M., 2017: *Data scientist: 21st century sexiest job for free*. John Snow Labs. Available at: http://www.johnsnowlabs.com/dataops-blog/data-scientist-21st-century-sexiest-job-for-free/ [Accessed 29 June 2017].

Trifu, M. & Ivan, M., 2014: Big Data: present and future. *Database Systems Journal* 5(1), pp. 32-41

Tsai, C., Lai, C., Chao, H., & Vasilakos, A., 2015: Big Data Analytics: A Survey. *Journal of Big Data* 2(21), pp. 1-32

Ularu, E., Puican, F., Apostu, A., & Velicanu, M., 2012: Perspectives on Big Data and Big Data Analytics. *Database Systems Journal* 3(4), pp. 3-14

Wulff, A. & Wunck, C., 2016: Integration of Business Process Management and Big Data Technologies. In: *International Conference on Industrial Engineering and Operations Management*, March 8-10, 2016, Kuala Lumpur, Malaysia

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F., 2017: Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth* 10(1), pp. 13-53

Zan, M. & Yanfei, L., 2015: Research of Big Data based on the views of technology and application. *American Journal of Industrial and Business Management* 5, pp.192-197

JEL Classification: C88, M15