Journal of Engineering Science and Technology
Vol. 11, No. 12 (2016) 1793 - 1807

IDENTIFYING AND ANALYZING THE TRANSIENT AND PERMANENT BARRIERS FOR BIG DATA

SARFRAZ NAWAZ BROHI[1], MERVAT ADIB BAMIAH[2], MUHAMMAD NAWAZ BROHI[3]

[1]School of Computing and Information Technology, Taylor's University, Lakeside Campus, Selangor, Malaysia
[2]College of Computer Sciences and Information Systems, Prince Sultan University, Riyadh, Kingdom of Saudi Arabia
[3]Department of Computer Science SZABIST, Dubai, United Arab Emirates.
*Corresponding Author: sarfraznawaz.brohi@taylors.edu.my

Abstract

Auspiciously, big data analytics had made it possible to generate value from immense amounts of raw data. Organizations are able to seek incredible insights which assist them in effective decision making and providing quality of service by establishing innovative strategies to recognize, examine and address the customers' preferences. However, organizations are reluctant to adopt big data solutions due to several barriers such as data storage and transfer, scalability, data quality, data complexity, timeliness, security, privacy, trust, data ownership, and transparency. Despite the discussion on big data opportunities, in this paper, we present the findings of our in-depth review process that was focused on identifying as well as analyzing the transient and permanent barriers for adopting big data. Although, the transient barriers for big data can be eliminated in the near future with the advent of innovative technical contributions, however, it is challenging to eliminate the permanent barriers enduringly, though their impact could be recurrently reduced with the efficient and effective use of technology, standards, policies, and procedures.

Keywords: Big data, Analytics, Transient and permanent barriers.

1. Introduction

Big data refers to large volumes of structured, semi-structured and unstructured data that are generated from numerous sources at an alarming velocity, volume, and variety [1]. It is rapidly emerging as a critically significant driver for business

success across the sectors. Big data is not the most favored topic of the organizations mainly because of its voluminous characteristic; instead, it is due to its quality of being further analyzed to discover insights that lead to better decisions and strategic transformations.

Every day 2.5 quintillion bytes of heterogeneous data (pictures, audios, videos, tweets, click streams, sensors data, and transaction records, etc.) are generated at tremendous velocity i.e. New York stock exchange captures 1 TB of trade data during each trade session [2]. Facebook captures 500+ TB of data with 2.5 billion pieces of contents shared every day. Each day 400 million tweets are sent by 200 million active users [3].

With the use of big data tools, healthcare sector can track the emergence of disease outbreaks via social media. They can also analyze the big data to develop insightful, cost-effective, and effective treatments. Governments can track the energy usage levels to predict outage and to sustain efficient energy consumption plans. Banks can track the web clicks, transaction records, bankers' notes, and voice recordings from call centers to identify the customers' spending patterns, preferences, as well as budget limitations to offer attractive services [4]. IBM believes big data has the potential to place communications services providers in a prime position to win the battle for customers and create new revenue streams [5].

For time-sensitive processes such as payment card fraud detection and multi-channel instant marketing, big data analytics techniques are applied to analyze data in real-time to identify the desired patterns. Big data also reveals its essence in the education sector for quality improvement by examining the students' performance and behavioral data captured from sources such as social media, student-professor meeting notes, and surveys [6].

2. Barriers for Big Data

Although, it is obvious from the stated scenarios that big data has brought up significant opportunities for the organizations to improve their business process. However, it faces several technical challenges. Eventually, with the growth of innovative IT solutions major technical challenges are expected to be solved in the near future but our findings represent that the challenges such as security, privacy, trust, data ownership, and transparency will remain permanent barriers for big data because technology alone is not sufficient to overcome them as there is need for the effective standards, policies and procedures.

Consequently, the core aim of this study was to identify and critically analyze the transient and permanent barriers for big data as shown in Fig. 1. This research provides insight for practitioners and researchers in industries as well as academia in their mission to produce empirical contributions by considering the current challenges and future direction of big data research. The rest of article is organized as follows. Section 3 describes the transient barriers for big data. Permanent barriers of big data are described in Section 4. Transience and permanence analysis is discussed in Section 5, and Section 6 concludes the research by emphasizing on its future direction.

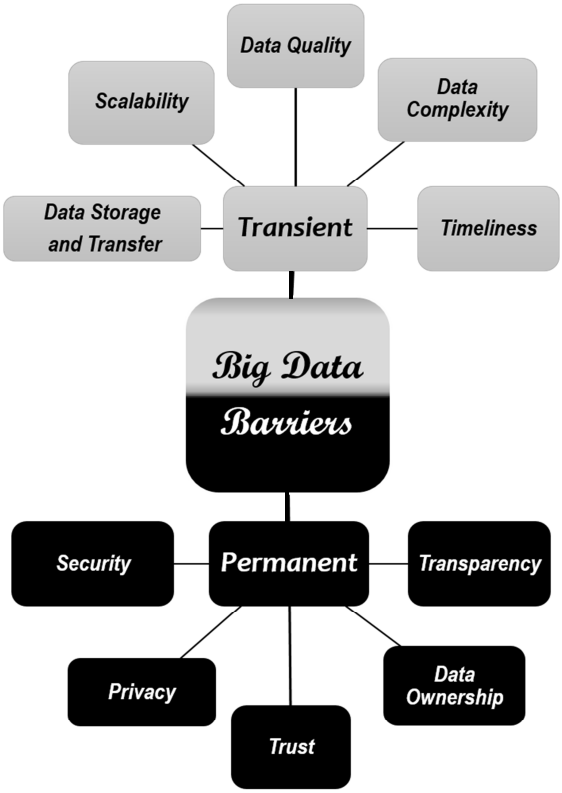

Fig. 1. Big data barriers.

3. Transient Barriers for Big Data

The challenges such as data storage and transfer, scalability, data quality, data complexity, and timeliness are severe barriers to adopt big data solutions. The transient barriers for big data are described in the following sub-sections and further analyzed in Section 5.

3.1. Data storage and transfer

Despite the use of high capacity storage mediums, it is challenging to store big data due to its alarming volume, velocity, and variety. Each minute voluminous data is being created on various platforms using numerous devices. Especially due to data generated on social media platforms, existing mediums are not capable

enough to store big data efficiently. In order to take advantage of big data, real-time analysis and reporting must be provided in tandem with the massive capacity required to store and process data. The sensitivity of data is also a major issue which arises the need for a private high-performance computing infrastructure as the best way to accommodate security requirements. Unfortunately, as datasets grow, legacy software, hardware, and transmission techniques can no longer meet these demands rapidly.

Furthermore, transfer of big data from a storage unit to a processing point is another emerging challenge. The majority of unstructured data are not generated at the place where they can be analyzed instead they are created at distributed points, henceforth in order to perform data analytics, organization need to move voluminous structured and unstructured data to an accessible storage such as Hadoop Distributed File System (HDFS) or Amazon S3 [7]. For instance, if a video has been recorded from several remote locations, it must be transferred to the shared processing Hadoop cluster to perform an efficient analysis. The high definition videos consist of several GBs of data requires highly advanced technology and communication protocols for efficient transfer [7, 8]. Data storage and transfer of complex data types remain the core barriers for the organizations to obtain marvelous benefits from big data analytics.

3.2. Scalability

Big data surpasses the typical computing, storage, and processing capabilities of the traditional databases and data analysis tools as well as techniques. There is an emergent need for the computationally powerful and efficient techniques that can be applied to analyze and filter meaningful information from large-scale data. With the emergence of big data, organizations have more data than the computing resources and processing technologies. The unmanageable voluminous data represent an instant challenge to the traditional computing infrastructures and requires scalable storage and a distributed strategy for storage, retrieval and analysis tasks. Indeed, this large amount is the essential characteristic of big data, therefore, organizations seek better techniques to develop scalable computing environments. Many organizations like Yahoo, Google, and Facebook already possess big data and they are relishing its benefits. However, considering the data growth rate at 30 to 60 percent per year as estimated by the analysts, organizations must develop enduring tactics to address the challenges of managing tasks that analyze exponentially mounting data sets with foreseeable linear costs. Although, certain organizations acquire on-premises Hadoop platforms which rely on commodity servers but they incur scalability problems and storage limitations hence they may fail to handle unexpected processing requirements when analytics and storage demands of the organization are going to increase. Consequently, it is vital for the organizations considering a big data initiative to acquire a high-performance computing platform that has ability to scale up and down on-demand. With the rapid increase in the size of heterogeneous datasets, the management and analysis of gigantic data remain the core barriers of the organizations seeking benefits of big data analytics. Usually, such issues are addressed by enhancing the processing capabilities but in big data scenarios volume of data increases at a faster rate than the processing speed [9].

3.3. Data quality

In the revolutionizing era of big data, data are generated, gathered and analyzed at an exceptional scale. Data quality is of utmost priority when it comes to effective and result-oriented analysis. The results of data analysis are vital for decision making and they can be affected by the compromised quality of data. Erroneous data costs an estimated amount of 600 billion dollars to US businesses annually. Due to huge volume, alarming velocity, and heterogeneous variety, data quality is imperfect. Organizations usually discover error rate of 1 − 5% in data, and for certain organizations, it is exceeding 30%. In some data warehousing projects, improving the data quality by the cleansing process consumes 30 − 80% of the development time and budget [10]. Data quality management has become an important area of research with the emergence of big data.

Generally, volume, velocity, and variety are considered as the core characteristics of big data. However, in order to extract value, the importance of the fourth 'V' of big data i.e. veracity is gradually being visible. The term veracity directly reflects the data irregularities and quality problems. User entry errors, duplications, and corruptions affect the value of data. Without sufficient data quality management, even minor errors may result into process inefficiency, revenue loss, and failure to comply with industry as well as government regulations [10]. Acquiring complex structured big data from various sources and effectively integrating them are the challenging tasks. When the volume of data is small, it can be checked by a manual search, and using Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT). Nevertheless, these practices are inefficient in processing peta and exa byte level data volume. The amount of global data created and copied reached 1.8 ZB in 2011. It is difficult to collect, clean, integrate, and finally obtain the required high-quality data within a practical time-frame. Since the quantity of unstructured data in big datasets is extremely high, it will take excessive time to transform unstructured data into structured types and further process the data [11]. Indeed, data quality management remains a major challenge for the current data processing techniques.

3.4. Data complexity

Big data is mainly of three types, structured, semi-structured and unstructured. Structure data retains related formats and user defined lengths. These data are generated either by the users or automated data generators without user interaction. Usually, query languages are used to process the structure data. Structure data are well-managed and stored in relational databases in a way that records are easily searchable using simple search algorithms whereas unstructured data are completely the opposite in terms of generation, storage, management, and retrieval process. The lack of manageability, format and structure makes it complex to process in an efficient manner. Nowadays, organizations are using Hadoop to process the unstructured data efficiently via clustering approach. Likewise, semi-structured data such as XML do not essentially possess a well-defined size or type [12].

Approximately 2.5 quintillion bytes of data are created from unstructured data sources such as social media, sensors, posts and digital photos each day. Undoubtedly, unstructured data is mounting rapidly. Searching as well as analyzing

unstructured data is harder than the structured data. It is highly valuable and organizations are discovering methods to extract information that can be translated into connections and patterns. Insufficient information generated from unstructured data could result into overlooked opportunities. Unstructured data are best analyzed with leading-edge analytics tools. Using these tools, organizations can formulate solutions to reduce fraud, prevent crime, ferret out waste, and detect acts of terror. Due to the diversity, velocity and size of data flowing through databases, it is extremely hard to discover patterns that lead to momentous conclusions [13].

3.5. Timeliness

In big data scenarios, timeliness is one of the most critical and significant factors. Organizations may obtain outdated and invalid information if they are not able to collect the required data in a timely manner. Processing the outdated data will result in misleading and useless conclusions which may further lead to serious mistakes in the decision-making process [11]. Considering the current research and development in the domain of big data, real-time processing, and analysis systems are still at the implementation or improvement stage and there is an emergent need for a system which can process large data sets faster. The issue of timeliness is triggered due to the large data sets which require unexpected durations to be analyzed. There are various circumstances in which the analysis results are required in real-time to take immediate actions. For instance, payment card industries would like to detect, block and flag the fraudulent transactions prior to their completion. Apparently, it will not be feasible to analyze the full purchase history of the users' in real-time, hence, it would be better to generate partial results in advance so that small amount of incremental computation with new data can be used to obtain quick decisions. In big data analytics scenarios, scanning the complete dataset to identify the required elements is impractical. Usually, index structures are created beforehand to find the desired elements efficiently. Since each index structure is designed to support only some criteria, in big data scenarios, there will be a need to formulate new index structures for unexpected data types. For example, consider a traffic management system which stores information regarding numerous vehicles and hotspots on highways and local streets. The system may be required to identify the congested locations along the routes selected by the user and recommend suitable alternatives. The successful process of this task requires an evaluation of multi-dimensional proximity queries working with the paths of moveable objects. New index structures will be required to support such queries. Designing such structures becomes particularly challenging when the data volume is increasing rapidly and the query processing has to be strictly efficient [14].

4. Permanent Barriers for Big Data

Big data concerns arise not only from the vast amounts of data generated and streamed but also from the manageability aspects since there are concerns where to store the voluminous data i.e. on or off-premises. Furthermore, who will access the data? How it will be analyzed, shared, and used? How large-scale data will be migrated from one cloud to another? These challenges give rise to permanent barriers for big data such as security, privacy, trust, data ownership, and

transparency. The permanent barriers for big data are described in the following sub-sections and further analyzed in Section 5.

4.1. Security

Big data service and product providers are required to secure their infrastructure and the way they provide services. The main objective of big data security is to maintain its confidentiality, integrity, and availability. Big data extracted from various sources such as Internet of Things (IoT) and stored on the cloud, has various concerns regarding physical and logical security. Cloud-based storage has to be secured because consumers' data are in danger of insider threats, cyber-crimes, and malicious access. Current security mechanisms are designed to secure the static limited amount of data and they cannot handle voluminous dynamic data [15]. Some big data security concerns identified by Open Web Application Security Project (OWASP) are shown in Fig. 2 [16] which allows the attackers to access the data, weak services, control updates and websites to perform malicious activities.

Fig. 2. Big data security concerns.

The traditional security mechanisms are insufficient to provide a reliable response to big data security challenges. Cloud Security Alliance (CSA) has proposed the following solutions for big data security [17].

- Securing transaction logs and data against unauthorized access.
- Ensuring availability 24/7.
- Securing computational and processing elements in cloud and distributed frameworks.
- Securing data breakdown and data sanitization capabilities.
- Ensuring endpoints security against any malicious activity.
- Providing real-time data security checks.

- Ensuring communication security.
- Ensuring the verification and validation of access control methods.
- Ensuring infrastructure security and enabling real-time protection during the initial collection of data.

4.2. Privacy

Preserving big data privacy ensures that data is protected and it is not revealed to any unauthorized party at any point in its lifecycle [18]. Big data collected from IoT devices, sensors and support systems including the personal data of the users, are not secured effectively. Failing to protect personal data causes privacy concerns. Moreover, people share personal information through their mobile devices to social networks and cloud services. The shared data are illegally copied, maliciously distributed, access and freely shared without considering the copyrights of the protected digital contents. These types of activities result in consumers' privacy breach as well as losing data and content control. Additionally, big data privacy concerns also include government authorities spying on people via social media networks. CSA has identified major big data privacy challenges including scalable privacy preserving data analysis and data mining, cryptographic solutions, granular access control, data management, integrity and reactive security [17]. Some big data privacy concerns are depicted in Fig. 3 [19].

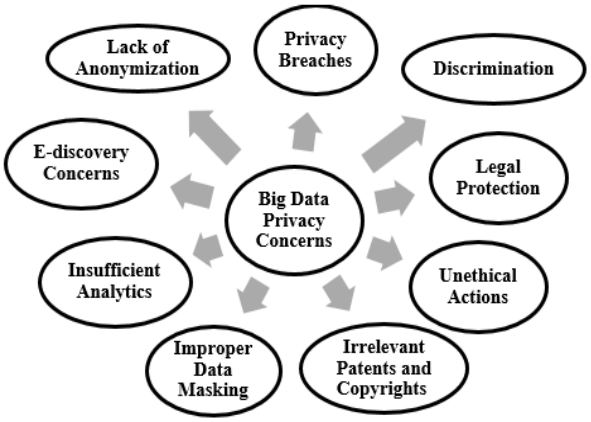

Fig. 3. Big data privacy concerns.

4.3. Trust

The increased amount of big data being generated, collected and processed, must be trusted throughout its entire lifecycle starting from trusting the source and methods of collection. Furthermore, there are trust concerns regarding the way data is collected, organized, accessed, and analyzed [20]. When considering the clients' point of view regarding trust on big data, whether it is for saving information, socialization, buying, selling or conducting any other transactions, big data suffers from authenticity and credibility issues because it can be collected

from commercial organizations which are profit-oriented business platforms, and they do not check the authenticity as well as validity of the collected data. It is challenging to trust data generation platforms such as Twitter and Facebook because they have many zombie accounts which are run by the machines. Trust can be based on policy or reputation that can be achieved through data collection, preparation, computation, storage, and communication.

4.4. Data ownership

Nowadays organizations acquire data from various resources including the external datasets. This raises ownership concerns to control the data from each perspective (access, share, store, and sell etc.) when integrating it into their existing transactions. Some of the organizations' big data sources are mostly unknown. Data ownership could be abstruse in big data scenarios, for instance, when data are generated on the social media networks, it is complex to determine whether the owners of a data items are the users whose information are recorded in the data items, or the parties who created the data items by collecting information from the users. Consumers are no longer owners of the data when it comes to big data scenarios. This results in ambiguity to data ownership and rights. Therefore, legal solutions are required to be enforced in big data scenarios for fair handling of data and its ownership [21].

4.5. Transparency

Disclosing the data logic assists the decision makers to verify whether the conclusions drawn by the organizations processing the data are accurate and fair. Organizations are expected to disclose the logic involved in big data analytics if there is a direct or indirect effect on the individuals. They have to do so proactively, without individuals having to actively take steps to seek disclosure. Personal data processed are now tracked online by smart devices and IoT. Transparency of the data collection point is required. While the complexity of data processing increases, organizations often claim secrecy over how data is processed on the grounds of commercial confidentiality. In order for consumers to efficiently exercise control over their data and provide meaningful consent, there must be transparency on who is accessing and managing their data and where does it reside. As for the case of any secondary use of the data, controllers have to notify consumers what is likely to happen to their data and they need to obtain their consent whenever required. The information imbalance between the organizations who hold the data and the consumers whose data they process increase with the deployment of big data applications [22].

5. Transience and Permanence Analysis

Oxford dictionary defines the word transience as the "state or fact of lasting only for a short time", whereas permanence is its antonym which is defined as the "state or quality of lasting indefinitely" [23]. Furthermore, Merriam-Webster defines permanence as the "quality or state of being permanent" [24]. In this research, we have analyzed the big data barriers by categorizing them as transient and permanent. The barriers such as data storage and transfer, scalability, data

quality, data complexity, and timeliness are considered as transient. Researchers have produced various contributions to overcome these problems but the proposed solutions require further improvements as well as enhancements. For example, Weisheng et al. [25] proposed a Generalized Multi-Protocol Label Switching (GMPLS) controlled network with the combination of circuit and storage to form a network for transferring big data files dynamically. Se-young et al. [26] presented a comparative performance and fairness study of three open-source big data transfer protocols using a 10 GB/s high-speed link between New Zealand and Sweden. Their results indicate that faster file systems and larger TCP socket buffers in both the operating system and application are useful in improving data transfer rates. Minoru [27] proposed a split file model that can represent big data efficiently in a low throughput storage. In the proposed model, a large file is divided into many small parts which are stored in a directory. They developed commands to support the use of split files in a transparent manner. Using these commands, replicated data is naturally excluded and effective shallow copying is supported. Pengfei [28] developed a two-level storage system to support high-performance and scalable data-intensive computing on high performance computing infrastructure using Hadoop by integrating an in-memory file system with a parallel file system. They implemented a prototype of the two-level storage system by integrating the in-memory file system Tachyon-0.6.0 with the parallel file system OrangeFS-2.9.0.

Das and Mohan [29] addressed the issue of big data complexity by proposing an approach to store and fetch unstructured data in an efficient manner. They extracted unstructured data from public tweets and parsed the XML data to store it in a NOSQL database such as HBASE. Pooja and Kulvinder [30] proposed an algorithm for the clustering problem of big data using a combination of the genetic and the K-Means algorithm. The main idea behind their algorithm was to combine the advantage of genetic and K-means algorithm to process large amounts of data. Zuhair et al. [31] stated that data cleansing approaches have usually focused on detecting and fixing errors with slight attention to the scaling of big datasets. This presents a serious impediment since data cleansing often involves costly computations such as enumerating pairs of tuples, handling inequality joins, and dealing with user-defined functions. They introduced BigDansing, a big data cleansing system to tackle efficiency, scalability, and ease-of-use issues in the data cleansing process. The system can run on the most common general purpose data processing platforms, ranging from DBMSs to MapReduce-like frameworks. BigDansing takes these rules into a series of transformations that enable distributed computations and several optimizations such as shared scans and specialized join operators. Sushovan et al. [32] provided a method for correcting individual attribute values in a structured database using a Bayesian generative model and a statistical error model learned from the noisy database directly. The proposed method avoids the necessity for a domain expert or clean master and it efficiently performs consistent querying over a complex database. In order to address the emerging issue of big data scalability, Cisco Systems introduced the Unified Computing System (UCS) in 2009 as a next-generation server and networking platform closely linked to partner storage platforms. The initial target was to reduce the cost and complexity of deploying and managing fast-growing and increasingly critical pools of virtualized applications. The UCS architecture was built with the large-scale service provider and scale-out applications in mind. The Cisco's Common Platform Architecture

based on UCS is emerging as an effective platform for enabling scalable big data and analytics environments. Companies can host both enterprise applications and big data applications under the same UCS domains, allowing data movement into Hadoop, NoSQL, or relational information management technology easily while eliminating the need for costly technology silos. Cisco UCS provides the programmable framework for its partners to provide a set of advanced provisioning and automation functions that make it possible to redeploy compute and network assets in minutes, not days or hours [33].

Sameer et al. [34] mentioned that timeliness is more important than perfect accuracy. Achieving small, bounded response times for queries on large volumes of data remains a challenge because of limited disk bandwidths, inability to fit many of these datasets in memory, network communication overhead during large data shuffles, and straggling processes. For instance, just scanning and processing a few terabytes of data spread across hundreds of machines may take several minutes. This is often accompanied by unpredictable delays due to stragglers or network congestion during large data shuffles. They introduced BlinkDB, a massively parallel approximate query processing framework optimized for interactive answers on large volumes of data. Aggregation queries in BlinkDB can be annotated with either error or maximum execution time constraints. Zhigao et al. [35], build a Real-time Big Data Processing (RTDP) architecture based on the cloud computing technology. They proposed a multi-level storage model and the LMA-based application deployment method to meet the real-time and heterogeneity requirements of RTDP system. The proposed idea is based on structuring the data, uploading it to the cloud server and MapReduce it combined with powerful computing capabilities of the cloud architecture.

The review of the related work is evident that researchers have addressed the transient barriers of big data but the proposed contributions require improvements and enhancements hence they are not ready for industrial adoption and general purpose use. Although, the transient problems are serious barriers to adopting big data solutions, however existing research indicates that there is a great possibility of them to be eliminated in the near future by producing empirical research and innovative developments in the field of big data. The challenges such as security, privacy, trust, data ownership, and transparency are considered as permanent barriers for big data. These problems will stay permanently but their impacts can be reduced periodically by producing effective contributions such as [36-40]. The permanent barriers cannot be solved only by the effective use of technology because it involves people, policies, and procedures. For instance, the problems of security, privacy, trust, data ownership, and transparency cannot be solved only by producing innovative IT solution but there is an additional need to enforce policies and procedures for acquiring the best practices so that the involved entities can perform as per the expectations to achieve a trustable big data computing environment. In permanent barriers, there is a great role of human being involved that cannot be resolved just by acquiring technical solutions. In big data scenarios, trust and cooperation of the organizations as well as the individuals play an integral role. There must be the use of laws which define the policies especially related to use of big data, and people must adhere to the well-defined policies to formulate a trusted big data computing environment. The strict process of accountability will produce a great impact in big data scenarios. There are several elements to formulate an effective accountability strategy; some of

them include auditing and compliance certification, data protection and privacy by design, impact assessment, lawyers, and data protection experts. The involved entities and processes can be used to ensure that big data is utilized responsibly.

By applying as well as practicing the suggested approaches, the impact of permanent barriers can be greatly reduced and big data could be a reality for every sector to relish its benefits such as effective data analytics which can improve the business quality, increase profit as well as productivity. However, these barriers cannot be permanently eliminated. These barriers of big data are similar to computer and network security concerns. It is not possible to establish a computing infrastructure that is completely free from threats and attacks, but organizations are securing their network and computing infrastructure with the use of latest technologies and well-defined policies to reduce the impact of threats and attacks in order to obtain a secure computing platform. Although, such infrastructure will stay secure, but its security cannot be guaranteed permanently, henceforth organizations where confidentiality of data is a vital act, they continuously monitor their computing and network infrastructures to safeguard them against the potential threats. Likewise, in big data scenarios, there must be continuous monitoring of the formulated solutions which also involve people, policies, procedures, standards and technologies to reduce the impact of permanent barriers. Furthermore, there must be trusted collaboration regarding transparency in big data environments for all the involved organization to feel comfortable in participating and sharing data with each other whenever required for analytical purposes. Data protection laws and acts must be defined especially for big data environments in order to guarantee safe and secure conduct.

6. Conclusion and Future Work

Big data analytics has revolutionized the world of information technology. Big data is not only used by the organizations to seek better insights for improving the quality of their service and profit, but it can be also used to achieve a variety of targets where success is dependent on the smart analysis. For example, Barak Obama successfully used big data analytics during his election campaign to seek the desired insights. As revealed by Dr. DJ Patel, the chief data scientist of the United States Office of Science and Technology, that big data can be used to benefit the citizens of America in several ways. Undoubtedly, big data analytics has numerous benefits but unfortunately there are several barriers to adopting and apply big data solutions such as security, privacy, trust, data ownership, and transparency. In order to make the big data analytics a reality for every organization, there is a need to formulate innovative solutions to overcome problems such as data storage and transfer, data quality, data complexity, timeliness, and scalability to just name a few. Although the transient barriers of big data can be eliminated by the effective use of innovative technology. However, the permanent barriers remain the core challenge but that does not imply that big data will never be effective and successful because the impact of such barriers can be reduced by formulating new policies, procedures, and data protection laws define explicitly for the big data environments. Despite the influential technical solutions, with the collaboration of all the involved entities and by enforcing as well as following the defined policies, we can devise trusted big data computing infrastructures that benefit the organizations as well as the individuals in several aspects of business and daily life.

References

1. Gandomi, A.; and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144.
2. Ahmed, K.B.; Bouhorma, M.; and Ahmed, M.B. (2014). Age of big data and smart cities: privacy trade-off. *International Journal of Engineering Trends and Technology*, 16(6), 298-304.
3. Madeline, G. (2013). Strategies for implementing big data analytics. Richard linowes, Kogod School of Business, American University.
4. Davenport, T.H.; and Dyche, J. (2015). *Big data in big companies*. International Institute for Analytics.
5. Fox, B.; Dam, R.; and Shockley, R. (2013). *Analytics: real-world use of big data in telecommunications*. IBM Global Business Services Business Analytics and Optimization. Produced in the United States of America.
6. Schmarzo, B. (2014). What universities can learn from big data – higher education analytics. Retrieved March 5, 2016, from https://infocus.emc.com/william_schmarzo/what-universities-can-learn-from-big-data-higher-education - analytics/.
7. EYGM, (2014). Big data: changing the way businesses compete and operate. Retrieved March 17, 2016, from http://www.ey.com/Publication/vwLUAssets/EY_-_Big_data:_changing_the_way_businesses_operate/%24FILE/EY-Insights-on-GRC-Big-data.pdf.
8. Kaisler, R.; Armour, F.; Espinosa, J.A.; and Money, W. (2013). Big data: issues and challenges moving forward. *Proceedings of 46th Hawaii International Conference on System Sciences*. Hawaii, USA, 995-1004.
9. Gorton, I. (2013). Addressing the software engineering challenges of big data. Retrieved April 4, 2016, from https://insights.sei.cmu.edu/sei_blog/2013/10/addressing-the-software-engineering-challenges-of-big-data.html.
10. Saha, B.; and Srivastava, D. (2014). Data quality: the other face of big data. *Proceedings of IEEE 30th International Conference on Data Engineering*. Chicago, USA, 1294-1297.
11. Cai, L.; and Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14(2), 1-10.
12. EDPS, (2015). Meeting the challenges of big data. Retrieved April 20, 2016, from https://secure.edps.europa.eu/EDPSWEB/edps/site/mySite/BigDataQA.
13. Digital Reasoning, (2014). Unstructured data: A big deal in big data. Retrieved April 23, 2016, from http://www.digitalreasoning.com/resources/Holistic-Analytics.pdf.
14. Jaseena, K.U.; and Julie, M.D. (2014). Issues, challenges, and solutions: big data mining. *Computer Science & Information Technology*. 131-140.
15. Yadav, N. (2016). Top ten big data security and privacy challenges. Infosecurity Magazine. Retrieved April 29, 2016, from http://www.infosecurity-magazine.com/opinions/big-data-security-privacy/.
16. Russell, B. (2015). Data security threats to the internet of things.. Parks Associates. Retrieved May 5, 2016, from https://www.parksassociates.com/blog/article/data-security-threats-to-the-internet-of-things.

17. CSA, (2013). Expanded top ten security and privacy challenges. Retrieved May 5, 2016, from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf.
18. Moura, J.; and Serrao, C. (2016). Security and privacy issues of big data. Retrieved May 13, 2016, from https://arxiv.org/ftp/arxiv/papers/1601/1601.06206.pdf.
19. Herold, R. (2016). 10 Big data analytics privacy problems. Retrieved May 11, 2016, from https://www.secureworldexpo.com/industry-news/10-big-data-analytics-privacy-problems.
20. Acquisto, D. (2015). Privacy by design in big data. an overview of privacy enhancing technologies in the era of big data analytics. *European Union Agency for Network and Information Security*.
21. Jim, D. (2015). Who owns big data?. Retrieved May 7, 2016, from http://blogs.sas.com/content/datamanagement/author/jimharris/.
22. Khan, N.; Yaqoob,1; Hashem, I.; Inayat, Z; Ali, W.; Alam,M.; Shiraz,M,; and Gani,A. (2014). Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal*.
23. Oxford dictionaries, (2016). Transient and Permanent. Retrieved May 22, 2016, from http://www.oxforddictionaries.com/.
24. Merriam-Webster, (2016). Transient and Permanent. Retrieved May 22, 2016, from http://www.merriam-webster.com/.
25. Hu, W.; Sun, W.; Jin, Y.; Guo, W.; and Xiao, S. (2013). An efficient transportation architecture for big data movement. *Proceedings of International conference on information, communications and signal processing*. Tainan, Taiwan, 1-5.
26. Yu, S.; Nevil, B.; and Aniket, M. (2013). Comparative Performance Analysis of High-speed Transfer Protocols for Big Data. *Proceedings of 38th Annual IEEE Conference on Local Computer Networks*. Sydney, Australia, 292-295.
27. Minoru, U. (2013). Split file model for big data in low throughput storage. *Proceedings of Seventh International Conference on Complex, Intelligent, and Software Intensive Systems*. Taichung, Taiwan, 250-256.
28. Pengfei, X.; Walter, B.L.; Pradip, K.S.; Rong, G.; and Feng, L. (in press). Accelerating big data analytics on HPC clusters using two-level storage. *Parallel Computing*.
29. Das, T.K.; and Mohan, P.K. (2013). BIG Data Analytics: A framework for unstructured data analysis. *International Journal of Engineering and Technology*, 5(1), 153-156.
30. Pooja, B.; and Singh, K. (2016). Big Data Mining: Analysis of genetic k-means algorithm for big data clustering. *International Journal of Advanced Research in Computer Science and Software Engineering*, 6(7), 223-228.
31. Zuhair, K.; Ihab, I.F.; Alekh, J.; and Samuel, M. (2015). BigDansing: A system for big data cleansing. *Proceedings of International Conference on Management of Data*. Victoria, Australia, 1215-1230.
32. Sushovan, D.; Yuheng, H.; Chen, Y.; and Subbarao, K. (2014). BayesWipe: A multimodal system for data cleaning and consistent query answering on

structured big data. *Proceedings of International Conference on Big Data*. Washington, USA, 15-24.

33. Richard, V.L.; and Dan, V. (2014). Building a datacenter infrastructure to support your big data plans. *Cisco in collaboration with Intel*. Retrieved May 10, 2016, from http://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/big-data/245209January.pdf.
34. Agrawal, S.; Iyer, A.P.; Panda, A.; Madden, S.; Mozafari, B.; and Stoica, I. (2012). Blink and it's done: Interactive queries on very large data. *Proceedings of the VLDB Endowment*, 5(2), 1902-1905.
35. Zhigao, Z.; Ping, W.; Jing, L.; and Shengli, S. (2015). Real-time big data processing framework: challenges and solutions. *An International Journal of Applied Mathematics and Information Sciences*. 9(6), 3169-3190.
36. Achana, R.A.; Ravindra, S.H.; and Manjunath, T.N. (2015). A novel data security framework using E-MOD for big data. *Proceedings of IEEE International conference on electrical and computer engineering*. Dhaka, Bangladesh, 546-551.
37. Changxiao, Z.; and Jinhua, L. (2015). Novel group key transfer protocol for big data security. *Proceedings of IEEE Advanced Information Technology, Electronic and Automation Control Conference*. Chongqing, China, 161-165.
38. Abid, M.; Iynkaran, N.; Xiang, Y.; Guang, H.; and Song, G. (2016). Protection of big data privacy. *IEEE Access: Translations and Content Mining*, 4(1), 1821-1834.
39. Chunming, G.; and Iwane, N. (2015). A social network model for big data privacy preserving and accountability assurance. *Proceedings of 12th Annual Consumer Communications and Networking Conference*. Las, Vegas, 19-22.
40. Johannes, S.; Christian, R.; Sabri, H.; and Pernul, G. (2014). Trust and Big Data: A roadmap for research. *Proceedings of 25th International Workshop on Database and Expert Systems Applications*. Munich, Germany, 278-282.