Scientific Research Publishing

Journal of Information Security, 2020, 11, 304-328
https://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2)

Abou_el_ela Abdou Hussein

Computer Science Department, Modern Academy-Maddi, ARE, Maddi, Egypt
Email: abo_el_ela_2004@yahoo.com

How to cite this paper: Hussein, A.A. (2020) Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). *Journal of Information Security*, **11**, 304-328.
https://doi.org/10.4236/jis.2020.114019

Received: September 27, 2020
Accepted: October 27, 2020
Published: October 30, 2020

Abstract

The amount of data that is traveling across the internet today, including very large and complex set of raw facts that are not only large, but also, complex, noisy, heterogeneous, and longitudinal data as well. Companies, institutions, healthcare system, mobile application capturing devices and sensors, traffic management, banking, retail, education etc., use piles of data which are further used for creating reports in order to ensure continuity regarding the services that they have to offer. Recently, Big data is one of the most important topics in IT industry. Managing Big data needs new techniques because traditional security and privacy mechanisms are inadequate and unable to manage complex distributed computing for different types of data. New types of data have different and new challenges also. A lot of researches treat with big data challenges starting from Doug Laney's landmark paper, during the previous two decades; the big challenge is how to operate a huge volume of data that has to be securely delivered through the internet and reach its destination intact. The present paper highlights important concepts of Fifty-six Big Data V's characteristics. This paper also highlights the security and privacy Challenges that Big Data faces and solving this problem by proposed technological solutions that help us avoiding these challenging problems.

Keywords

Big Data, Big Data V's Characteristics, Security, Privacy, Challenges, Technological Solutions

1. Introduction

Big data term is referring to data that is so large and complex and also exceeds the processing capability of traditional data management systems and software

techniques. The term big data refers to the enormous amounts of digital information companies and governments collect about us and our surrounding people. Using large scale cloud infrastructures, with different software platforms, spread across large networks of computers, increases the attack surface of the entire system. Data becomes big data when individual data ceases to be relevant and the analyses derived from a large group of it are valuable. With new big data analyzing technologies, thoughts can be derived to empower better decision making for many critical development areas such as economic productivity, health care, energy, and natural disaster prediction. The first time the word Big Data appeared was in the year 2000 in a paper by Diebold. The era of Big Data has bought with it a plethora of opportunities for promotion of economic growth, enhancement of education system, the advancement of science, improvement of health care, and more ways of social interaction and entertainment. But as big data has its advantages it has flip side as its challenges. Security and privacy are great issues in big data due to its huge volume, high velocity, big variety represented in large scale cloud infrastructure, various data formats and sources, cloud migration, etc. The various security and privacy challenges related to big data and cloud computing and the reasons why they crop up are explained later in detail. Our paper is organized as follows. Section 2 provides a literature review of the Fifty-Six V's Characteristics of big data. In Section 3, we comprehensively describe big data privacy and security challenges. In Section 4, we propose four strategies to overcome big data security challenges. Finally, we introduce conclusion and future work in Section 5.

2. Fifty-Six V's Characteristics of Big Data [1]

Many researchers have a lot of attention for studied big data characteristics starting from 3 V's characteristics (Volume, Velocity, and Variety) which leads to add some more V's to the characterization of big data. Other authors used the term pillars, or dimensions instead of big data V's characteristics [2]. By the time the three V's with "Veracity", "Value", "Variability" reached fifty-six V's characteristics, two of them were added by the author as explained in [1]. Several V's are awarded in **Figure 1**. Different declarations for each "V" characteristic (dimension) are presented as follows as explained in **Table 1**:

Table 1. Characteristics of big data: The 56 V's.

V's Chars	Explanation
1. **Volume**	Many companies have already amount of archived "Ocean of data" in the form data or of information that can come from every possible sensor, logs, hundreds of hours of YouTube uploaded videos, billions of gigabytes from global mobile traffic [1].
2. **Variety**	Big Data is represented by different formats and varied types of data between structured, semi-structured, multi-structured and mostly unstructured data as well that came from many types of data resources, so it is heterogeneous in both size and type, consequently cannot be put together into a relational database [1] [3].

Continued

3. Velocity	Could be defined as the speed of data traveling from one side to another or moves around and the speed of processing it with high rate of receiving data and information [1].
4. Veracity	We need clear and definite answer for a very important question, does data comes from a reliable source [1].
5. Validity	How quality consistence, preciseness, reasonableness and correctness the data for its intentional use [1].
6. Value	Unless turning the enormous amount of data in big data into value, it could be useless and unusable [1].
7. Variability	Variability in big data's circumstances means variability in the data, which required to be found by deviation and aberration detection methods leading for any relevant analytics to occur [1] [4].
8. Venue	Big data is distinguished by its distributed heterogeneous data from various platforms, from numerous owners' systems, with different formatting and access needs, private or popular [1] [5].
9. Vocabulary	All metadata shapes like data models, schema, semantics, ontologies, taxonomies, and other contents that describe the data's structure, syntax, content, and origin [1] [5].
10. Vagueness	The meaning of found data is often very unclear, not only has how much data been available but also how much it is not obscure [1] [5].
11. Vulnerability	This means that no system is perfect, which means it's probable there is a way for its hardware or software to be agreement, successively meaning that any associated data can be tacked or manipulated [1] [4].
12. Volatility	What time does remain data valid and should be stored. How old does data need to be before it is considered irrelevant [1] [4].
13. Visualization	Refers to the application of more recent visualization techniques to explain the relationships between data and can display real-time changes and more illustrative graphics, thus going beyond pie, bar and other charts [1] [4].
14. Viscosity	It is occasionally used to express the delay, latency or lost time in the data relative to the phenomenon being described [1] [6].
15. Virality	Measures the rate at which data can propagate through a network [1] [6].
16. Virtual	Enterprises and other groups can benefit from big data virtualization because it authorizes them to use all the data assets they gather to accomplish various goals and objectives [1].
17. Valences	It is a measure indicating how dense the data is [1].
18. Viability	Viability could be seen as carefully choosing those attributes in the data that are most likely to forecast outcomes that matter most to organizations [1].
19. Virility	With Big Data it means that it creates itself. The more Big Data you have, the more Big Data gets strength and forceful [1] [7] [8] [9] [10].
20. Vendible	The very existence of client's for Big Data shows crucially that it is appreciable—this is evident from the communication of some known means of trading with subscribers data [1] [7] [11] [12].
21. Vanity	Vain of data means that it is glad with the effect it produces on other individuals, [1] [7] [11] [12].
22. Voracity	Big Data is potentially so insatiable that it may achieve the influence, manage and the possibility to consume itself [1] [7] [11] [12].

Continued

23. Visible	Not only pertinent information should exist, but also should be evident to the intended person at the proper time [1] [13].
24. Visual	We currently live in a world of seeing, watching, and exchanging photos and videos, whether they are personal or product pictures or weather photos through the Internet [1] [7] [13].
25. Vitality	Vitality of the data is an important perception that is vital and is included in the concept of value [1] [7] [13].
26. Vincularity	It implies in its exact meaning connectivity or linkage. This idea is very pertinent in today's interconnected world through the internet [1] [7] [14].
27. Verification	The process of initiate the fact, precision, or validity of data [1] [7] [12].
28. Valor	The specific data that has the possibility to produce value and guiding how this can be accomplished [1] [7] [15].
29. Verbosity	Understanding how to quickly separate the meaning you keep about from its repetition is important for efficiency of processing [1] [7] [13].
30. Versatility	Versatility of data shows to what extent the data is useful, in different scenarios [1] [7] [13].
31. Veritable	Data being in fact the thing named and not false, unreal, or imaginary [1].
32. Violable	Violable data capable of being or likely to be violated [1].
33. Varnish	Interaction of end-users with our work matters, and polish counts [1].
34. Vogue	Artificial intelligence are become? [1] [16].
35. Vault	Importance of data security [1] [16].
36. Voodoo	Deliver results with real-world impact [1] [16].
37. Veil	Examine latent variables from behind the curtain [1] [16].
38. Vulpine	Data leads to a new technology [1] [16].
39. Verdict	People affected by model's decision [1] [16].
40. Vet	Vetting the assumptions with evidence [1] [16].
41. Vane	Unclear direction of decision-making [1] [16].
42. Vanilla	Simple methods if tackled with care can provide value [1] [16].
43. Victual	Big Data fuel of data science [1] [16].
44. Vantage	Privileged view of complex systems [1] [16].
45. Varmint	As data gets bigger, so do software bugs [1] [16].
46. Vivify	Ability of data science to cope with every real-life aspect [1] [16].
47. Vastness	Bigness of Big Data [1] [16].
48. Voice	Ability to speak with knowledge [1] [16].
49. Vaticination	Ability to forecast [1] [16].
50. Veer	Change direction according to customer need [1] [16].
51. Voyage	Increasing knowledge [1] [16].
52. Varifocal	It is about trees and forest [1] [16].
53. Version control	You are using it right? [1] [16].
54. Vexed	Potential of data science to handle complicated problems [1] [16].
55. Vibrant	Provision of insight by data science [1] [16].
56. Vogue	Artificial intelligence will become? [1] [16].

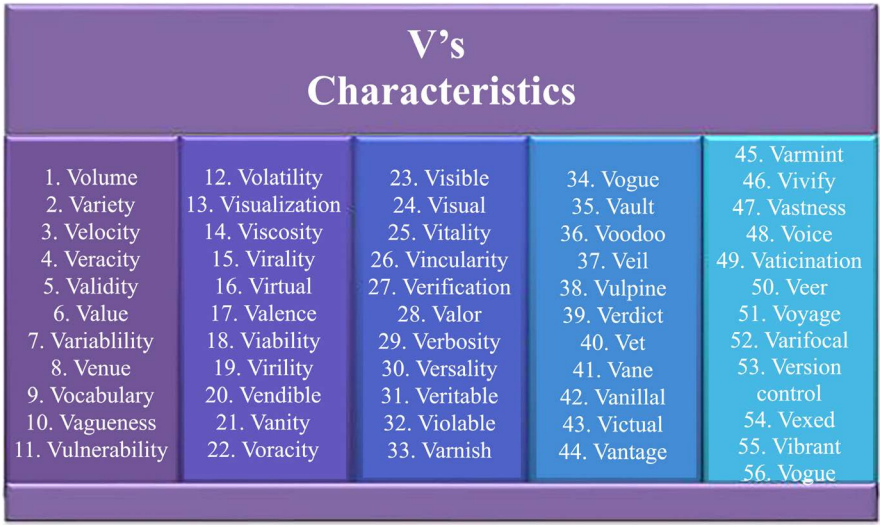

Figure 1. Fifty-six big data V's characteristics.

3. Big Data Privacy and Security Challenges

There are several studies that have addressed various threats to big data privacy and security from more than one concept or view. One view of these challenges divides challenges into categories; some of them come back to a function of the characteristics of BD, by its existing analysis methods and models, and some, through the limitations of current data processing system. All are introduced in Big Data Challenges view No. 1 [16] [17]. Another view seeks appropriate protection and privacy need to be enforced throughout the big data lifecycle phases, introduced in Big Data challenges view No. 2 [18].

3.1. Big Data Challenges View No. 1

Existing studies surrounding BD challenges have cared attention to the difficulties of understanding the notion of Big Data [16] [19], decision-making of what data are generated and collected [20], issues of privacy [17] and ethical considerations relevant to mining such data [21]. Tole [3] asserts that building a viable solution for large and multifaceted data is a challenge that businesses are constantly learning and then implementing new approaches. For example, one of the biggest problems regarding BD is the infrastructure's high costs and the actress in its most important component, namely, Hardware equipment that is very expensive even with the availability of cloud computing technologies [22]. In addition, to arrange and sort data, so that valuable information can be constructed, human analysis is often required. While the computing technologies required to facilitate these data are keeping pace, the need of the human expertise and talents to benefit from BD, that are not always available and this proves to be another big challenge. As reported by Akerkar [23] and Zicari [24], the broad challenges of BD can be seen through three main classifications, based on the data life cycle: data, process and management challenges as explained in Figure 2 [16]:

Figure 2. Conceptual classifications of big data challenges.

1) Data Challenges: Represents the group of the challenges related to the characteristics of the data itself. They have identified key challenges in this phase that are mapped to the prominent V's of big data as (variety, velocity, variety, variability, volume, value, visualization, venue, vulnerability (Poor quality data), veracity (Pressure from the top), virtual (Lack of support), volatility, valence, validity).

2) Process Challenges: are related to series of how techniques: how to capture data, how to merge data, how to modify data, how to choose the right model for analysis and how to provide the results. With the big rate of data explosion, storage systems of organizations and enterprises are confronting major challenges from mountains of data, and the ever increasing of generated data [25]. Value can be generated from large data set. For example [26], Facebook increases its ad revenue by mining its users' personal preferences and creating profiles, showcasing advertisers which products they are most interested in. Google also uses data from different applications as Google search, YouTube, and Gmail accounts to profile users' manners and habits. Despite the tremendous benefits that can be acquired in large data set, big data requests for storage and processing poses a major challenge. The total size of data that have been generated by the end of 2015 is estimated at 7.9 zettabytes (ZB), which almost five times as many as 2020, that is expected to reach 35 ZB. In this phase of big data life cycle, the key challenges of this phase are as follows:

3) Management Challenges: Cover for example privacy, security, governance and ethical aspects [16]. Data Censorship can be challenging since it includes everything from security and privacy to meeting compliance standards and the ethical use of data. With big data, management problems expand even bigger because the data shape is unstructured and unpredictable.

3.2. Data Challenges View No. 2

Web makes it easier to collect and share knowledge as well data in raw shape.

Big Data is about how these data can be stored, processed, and comprehended with an aim of using it in expecting the action in future with a reasonable accuracy and allowable time delay. Here, the security and privacy challenges of big data in each phase of the big data three phase's lifecycle are evaluated as explained in **Figure 3** [18]:

1) Big Data Acquisition,

2) Big Data Storage and Management, and

3) Big Data Analytics.

1) Big data acquisition

Different data processing constructions for big data have been suggested to address different properties of big data [27]. Data acquisition could be defined as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution. The acquisition of big data is most commonly governed by five of the V's: volume, variability, velocity, variety, and value. Most data acquisition scenarios assume high-volume, high-velocity, high-variety, high-variability but low-value data, this makes it important to have adaptive and time-efficient collection, filtering, and cleaning algorithms that only guarantee that the high-value fragments of the data are actually processed by the data-warehouse analysis. Key challenges in this phase that are mapped to the prominent V's of big data as follows [16] [18]:

a) Volume: Many companies have already amount of archived "Ocean of data" in the form data or of information that can come from different resources, this data can be mined giving benefit information that help decision making in all life fields [1].

b) Variety: of data, traditional security techniques that adopt encryption methods are not suited for all types of sources from which big data acquisition takes place.

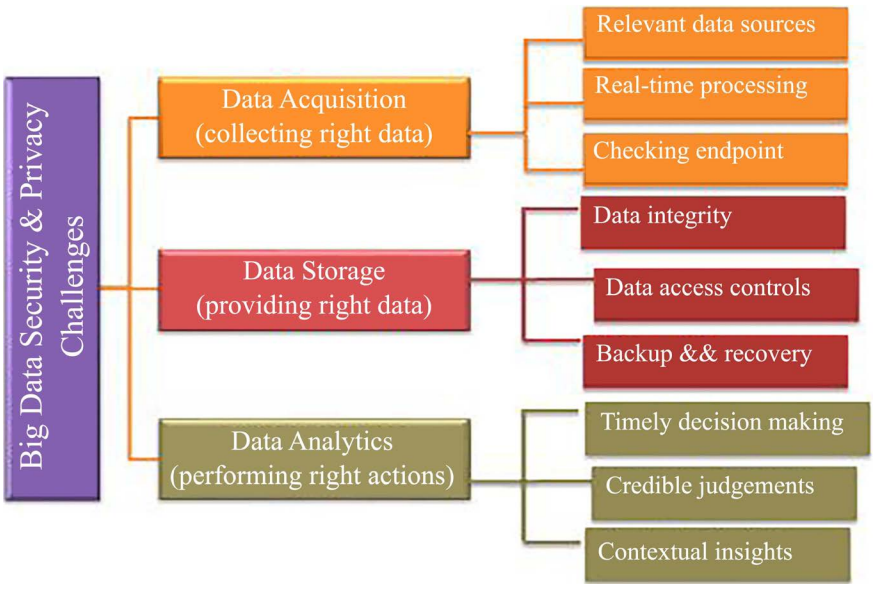

Figure 3. Security and privacy challenges of big data three phase's lifecycle.

c) Velocity: of acquisition of data is so high that it becomes difficult to monitor the traffic in real-time as data gets streamed into the storage [18].

d) Variability: Dimension of big data there can be inconsistencies in data formats, speed and types, such as in Web clickstream data leading to security vulnerabilities [1].

e) Value: Unless turning the enormous amount of data in big data into value, it could be useless and unusable, With so much data available, it's difficult to find and access the statistics most in need. Also, because employees are confused from a huge volume of data, they may not fully analyze data or only focus on the measures that are easiest procedures to collect it instead of the ones that add value. [1].

Overall, due to the above issues, big data when acquired could possibly become the carrier of advanced persistent threat (APT) [28]. APT thrives in such situations where there is a variety of data sources and non-standard data formats for data streams coming as an ongoing process in social networks [18]. When an APT code is hidden in big data, it becomes difficult to be detected in real-time. Hackers could attack the data source, destination, and all the connectivity by capitalizing on their vulnerabilities, which could result in an enlarged attack by launching a botnet. Therefore, it is important to enforce data security and privacy policies within a real-time processing environment of big data during the data acquisition stage itself. It is essential to connect the right endpoints of a network for the data flow along with sophisticated authentication and privacy policies for big data.

2) Big data storage

With the big rate of data explosion, storage systems of organizations Companies are facing challenges from massive amounts of data, and the ever-increasing number of data generated [25]. Value can be generated from large data set. For example [18], Facebook increases its ad revenue by mining its users' personal preferences and creating profiles, showcasing advertisers which products they are most interested in. Google also uses data from different applications as Google search, YouTube, and Gmail accounts to profile users' manners and habits. Despite the tremendous benefits that can be acquired in large data set, big data requests for storage and processing poses a major challenge. The total size of data that have been generated by the end of 2015 is estimated at 7.9 zettabytes (ZB), which almost five times as many as 2020, that is expected to reach 35 ZB. In this phase of big data life cycle, the key challenges of this phase are as follows:

a) Venue: With the growing massiveness of big data, the Volume dimension affects the server infrastructure of an organization. Traditional data warehouses may not suffice, and alternate storage systems such as distributed, cloud, and other outsourced big data servers are required to be employed to cope with the volume as well as the increasing Velocity of big data storage [18].

b) Volatility: Structured, unstructured, and semi-structured data are getting accumulated in the big data storage with high Volatility from various channels, including online transactions of sales, customer feedback, social media messages,

emails, marketing information, and various other logs that are both directly and indirectly associated with the business operations of an organization [18].

c) Valence: Big data is also shared among multiple related departments for their day-to-day transactions and functional operations. Hence, the big data connectivity among different data centers that are in-house, cloud-based, or outsourced can make the big data quite dense, impacting on the Valence dimension of big data [18].

d) Validity: How quality consistency, preciseness, reasonableness and correctness the data for its intentional use. Validity in data gathering means that your detections accurately represent the incident you are declared to measure. According to Forbes, estimated data scientist's spent 60 percent of their time in cleansing their data before being able to do any analysis.

Due to the above, the integrity of the data could be affected when multiparty operations take place on the same data storage in huge amounts and in increasing real-time speed [27]. Traditional encryption and security measures to maintain data integrity may not help as multiple mechanisms could be in place. Such a disparate environment could encourage sniffers to reach the servers by exploiting their security policy differences and vulnerabilities. Any misuse of data could lead to privacy leakage. These factors increase the risk of information theft and user privacy infringement.

Traditional access control methods are mainly classified as mandatory, discretionary or role-based, and none of them can be effectively applied to big data storage due to the diversity of users and their permission rights in such a highly dynamic environment [18]. Hence, new trustworthy data access controls must be established, adhering to appropriate security and privacy protection schemes and policies [29]. Good practices for backup and recovery must be followed in dealing with historic data that require be archiving or destroying at every stage of the big data life cycle.

3) Big data analytics

Big Data Analytics is the key to data value creation and that is why it is important to focus on that feature of analytics [18]. Data Analytics requires implementation of an algorithmic or mechanical process to derive insights through several data sets to look for significant correlations between each other. It is used in several fields to allow making better decisions. The focus of Data Analytics lies in deduction, which is the process of acquires conclusions that are based on what the researcher already knows. Successful implementation of big data analytics requires a combination of skills, people and processes that can work in perfect synchronization with each other. We have identified key challenges in this phase that are mapped to the prominent V's of big data as follows:

a) Volume (The amount of collected data): With today's data-driven organizations and using of big data, risk managers and other employees are often confused with the amount of data that is collected [30].

b) Value (Collecting meaningful data): With so much data available, it's dif-

ficult to find and access the statistics most in need. When employees are confused, they may not fully analyze data or only focus on the measures that are easiest procedures to collect it instead of the ones that add value.

c) Visualization (Visual representation of data): To be understood and impactful, data often needs to be visually presented in graphs or charts using visualization techniques and tools. While these tools are unbelievably useful, it's difficult to build them manually. Taking the time to drag information from multiple sources and put it into a reporting tool is disturbing and time-consuming.

d) Variety (Data collected from different sources): Different pieces of data are often collected from different systems. Users may not always recognize this, leading to incomplete or inaccurate analysis. Manually collecting data is time-consuming and can limit insights to what is easily viewed.

e) Venue (Unattainable data): Operating data into one centralized system has little effect if it is not easily accessible to the people that need it. An organization's data should be accessible by both Decision-makers and risk managers for insights on what is happening at any given moment, even if they are working off-site. Get information should be the easiest process of data analytics.

f) Vulnerability (Poor quality data): Nothing is more harmful to data analytics than inaccurate data. Without correct input, output will be unreliable. The main reason of inaccurate data is manual errors resulting during data entry which lead to large negative consequences particularly if the analysis is used to influence decisions. Another concern is asymmetrical data: when information in one system does not consider the changes made in another system, leaving it outdated. System integrations ensure that a change in one domain is instantly considered across the board.

g) Veracity (Pressure from the top): As risk management becomes more popular in organizations, Chief Financial Officers (CFOs) and other executives demand more results from risk managers. With a comprehensive analysis system, risk managers can exceed expectations and any analysis required.

h) Virtual (Lack of support): Data analytics can't be effective without organizational support, both from the top and lower-level employees. Risk managers will be powerless in many activities if chiefs don't give them the ability to act. Other employees play a key role as well: if they do not submit data for analysis process or their systems are unavailable to the risk manager, it will be hard to extrapolate any actionable information. Enterprises and other groups can benefit from big data virtualization because it authorizes them to use all the data assets they gather to accomplish various goals and objectives.

i) Variability (Confusion or anxiety): Variability in big data's circumstances means variability in the data, which is required to be found by deviation and aberration detection methods leading for any relevant analytics to occur. Also users may feel confused or anxious about switching from traditional data analysis methods, even if they understand the benefits of automation. Includes dynamic, progress, time series, spatiotemporal data, periodic & seasonal, interval of re-

porting, and any other type of dynamic (non-static) behavior in your data sources, customers, etc. Nobody likes change, especially when they are comfortable and familiar with the way things are done.

4. Proposed Big Data Security and Privacy Strategies

Four main technologies are proposed to comprehensively cope with the 56 V's during the three phases of big data lifecycle as in [18], namely data acquisition, data storage and data analytics. These are:

1) Data Provenance Technology,
2) Data Encryption and Access Control Technology,
3) Data Mining Technology, and
4) Block Chain Technology.

But with some adjustments to cope with the large and increased number of V's characteristics that reached 56 V's not only 11 V's as in [18], and also with chosen promising proposed method for each of the four techniques which explained as follows:

4.1. Data Provenance Technology

To adapt data provenance technology for addressing the security and privacy challenges in the data acquisition phase of big data. In traditional computing systems, data provenance method was used to determine the source of the data in data warehouse by adopting the labeled technique. With big data, the data acquisition involves diverse data sources from the Internet, cloud, social, and IoT networks. While big sensing data streams come with novel encryption schemes, attacks are possible right from the data acquisition phase [18] [31]. Hence, metadata about these data sources such as the data origin, the process used for dissemination and any intermediate calculations could be recorded in order to facilitate mining of the information at the time of data streaming itself. Hence, the first strategy their proposed technique is to adapt data provenance technology for effectively using data analytics techniques for detecting anomalies in the data acquisition phase of big data. However, collecting provenance metadata must adhere to privacy compliance. Another important issue is that it could become complex with application tools generating growingly large provenance graphs for establishing metadata dependencies. Data analytics of such graphs could be computationally intensive, and algorithms are being developed to detect anomalies using proximity graphs [18] [28] [32]. For instance, within a data provenance technology, the social network analysis component could adopt anomaly detection model based on a social score. Apart from balancing the trade-offs between privacy and computational complexity, it is important to monitor the data provenance technology itself as it could be attacked and requires security protection from hackers. For instance, in a real-life example of identifying the owner of the provenance documents, provenance graphs with chains of derivations of the vocabulary that contain the provenance information

could have variations based on the provenance producer's individual style [18]. In this paper, a promising algorithm by Zirije Hasani and Samedin Krrabaj [18] [33] is chosen. In this work, proposed algorithm used makes an enhancement of forecasting models and gives a short description about the algorithms and then they are categorized by type of prediction as: predictive and non-predictive algorithms [33]. They implement the Genetic Algorithm (GA) to periodically optimize Holt-Winters (HW) and Taylor's Double Holt-Winters (TDHW) smoothing parameters (used to predict the normal behavior of the periodic streams, and to detect anomalies when the deviations of observed and predicted values exceeded some predefined measures) in addition to the two sliding windows parameters that improve Hyndman's MASE measure of deviation, and value of the threshold parameter that defines no anomaly confidence interval. They also propose a new optimization function based on the input training datasets with the annotated anomaly intervals, in order to detect the right anomalies and minimize the number of false ones. The proposed method is evaluated on the known anomaly detection benchmarks NUMENTA and Yahoo datasets with annotated anomalies and real log data generated by the National education information system. However, this work brings several benefits [34]:

- Review and classification of existing literature for anomaly detection [33];
- From all the reviewed literature for anomaly detection, they assessed methods and algorithms for anomaly detection in data streams (time series) which are proper and capable to respond to the challenges that massive data streams and real-time detection have [33];
- It proposes an enhancement of the additive Holt-Winters (HW) and Taylor's Double Holt-Winters (TDHW) forecasting models that answer the stated challenges. The algorithm is implemented as a positive feedback optimization with a periodic adaptation of the algorithm parameters [33];
- Starting with ideas of numerous papers [35] [36] [37], it uses the GA optimization process, to optimize α, β, γ, ω, the HW and TDHW smoothing parameters, where they added optimization of the three new parameters k, n and δ, [33];
- Improvement is made in the new definition of the optimization function based on the input training datasets with the annotated anomaly intervals, enhanced Hyndman's MASE [38] definition where k and n define the two sliding windows intervals, and δ is the threshold parameter [33];
- The positive feedback learning process is achieved if the anomalies detected in the next time frame, by the proposed detection engine based on the computed optimal parameters from the annotated anomalies of previous one, are verified/acknowledged by human and reused for parameter optimization [33];
- The results of the experiments performed on the sets of synthetic and real data periodic streams show that proposed HW algorithm, with GA optimized parameters and with improved MASE, outperforms the other algorithms,

[33].

The data used for experiments are known as anomaly detection benchmarks NUMENTA [39] and Yahoo [40] datasets with annotated anomalies and our real log data from the Macedonian national education system e-dnevnik [33].

Based on the experimental evaluation of the detection rate and precision, performed on sets of synthetic and real data periodic streams, can be concluded that proposed HW with GA optimized parameters (α, β, γ, δ, k, n) and with improved MASE outperforms the other algorithms. This can't be concluded for the TDHW with GA optimization. Due to the HW iterative procedures, detection time is appropriate for the real-time anomaly detection. Optimization with GA that is also rather fast, with rather a small number of iterations (about 25 - 30 iterations are needed to achieve all tagged anomalies recognition in the training sets), can be done in batch mode on training sets, as also re-optimization with verified newly detected anomalies [33].

4.2. Data Encryption and Access Control Technology

To adapt advanced encryption techniques and access control schemes in big data storage systems [18]. Contemporary schemes such as homomorphic, attribute-based, and image encryption are being explored to ensure that sensitive private data is secured in cloud and other big data storage and service platforms [41] [42]. Even though homomorphic encryption allows some operations on encrypted data without decrypting it, the computing efficiency and scalability of homomorphic encryption schemes need improvement in order to be able to handle big data. On the other hand, the attribute-based encryption technique is regarded more appropriate for end-to-end security of private data in the cloud environment since the decryption of the encrypted data is possible only if a set of attributes of the user's secret key matches with the attributes of the encrypted data [42]. One of the major challenges of this scheme is the implementation of revocation since each attribute may belong to different multiple set of users [43] [44]. Anonymising data with a hidden key field could be useful for privacy protection. However, using data analytics such as correlation of data from multiple sources, an attacker would be able to identify the anonymized data. Hence, in addition to having good cryptographic techniques to ensure privacy and integrity of active big data storage, proof of data storage needs to be continuously ensured. Another important aspect is to provide proof of the archived data storage in order to verify that the files are not deleted or modified by attackers. Hadoop has become a promising platform to reliably process and store big data. It provides flexible and low cost services to huge data through Hadoop Distributed File System (HDFS) storage [45]. Unfortunately, absence of any inherent security mechanism in Hadoop increases the possibility of malicious attacks on the data processed or stored through Hadoop. In this scenario, securing the data stored in HDFS becomes a challenging task. Hence, researchers and practitioners have intensified their efforts in working on mechanisms that would

protect user's information collated in HDFS. This has led to the development of numerous encryption-decryption algorithms but their performance decreases as the file size increases. We chose a methodology to solve the issue of data security in Hadoop storage as in [45]. The authors of this methodology have integrated Attribute Based Encryption with the honey encryption on Hadoop, *i.e.*, Attribute Based Honey Encryption (ABHE). This approach works on files that are encoded inside the HDFS and decoded inside the Mapper. In addition, the authors of this methodology have evaluated the proposed ABHE algorithm by performing encryption-decryption on different sizes of files and have compared the same with existing ones including AES and AES with OTP algorithms. The ABHE algorithm shows considerable improvement in performance during the encryption-decryption of files. The main contributions of this methodology are:

- To carry out the in-depth study of big data processor, *i.e.*, Hadoop and to assess its strength and weakness in terms of security and privacy;
- To propose an ABHE, a secure and efficient algorithm executed on single and two Data Nodes in Hadoop. Also, it ensures the full security against all side channel attacks, brute force attack and collusion attack;
- To conduct experiments on test data to prove the efficacy of our proposed approach ABHE vs. other secure approaches *i.e.*, AES and AES-OTP;
- The performance of proposed ABHE has been calculated in terms of File size, Encryption Time, Decryption Time, Throughput and Power Consumption;
- The result shows that ABHE improves the execution time (total time taken for encryption and decryption of data) without affecting the size of original file.
- The proposed methodology has been able to successfully solve the weaknesses present in the security approaches available for big data security. The significance of the proposed work is as follows:
- The proposed encryption technique which uses the concept of Attributes Based Honey Encryption (ABHE) may help to secure sensitive information stored at HDFS in insecure environment such as the internet and cloud storages.
- The proposed technique provides both HDFS and Map Reduce computation in the Hadoop as well as cloud environment to secure and preserve the integrity of data during execution or at rest. Therefore, they have directed their efforts in securing the data transfer and computation paradigm in Hadoop environment by using chipper text policy attributes based honey encryption and Honey encryption for secret share of tuple of data and sent them to the cloud in a secure manner.
- The chipper text policy attributes based encryption makes the application secure and has a high performance when compared with the rest of the encryption techniques. Also, it provides the secure data transfer to all cloud applications.
- In the proposed algorithm, they have assured the data security by using sim-

plified chipper text policy attribute based encryption with Honey encryption which is difficult to decrypt by any unauthorized access.

- The user authorization access is based on the user define policy which reflects the overall organizational structure and also, depends upon a set of attributes within the system.
- With the proposed algorithm, the security of data is not only dependent on the secrecy of encryption algorithm but also on the security of the key. This provides dual layer security for the data.

This proves that the proposed technique is fast enough to secure the data without adding delay. Also, the proposed ABHE algorithm has a higher throughput which proves its applicability on big data. It provides a feasible solution for secure communication between one Data Node to other Data Node. The proposed encryption technique does not increase the file size therefore it saves the memory and bandwidth, and hence reduces traffic in a network. Also, it has an ability to encrypt structured as well as unstructured data under a single platform. Only HDFS client can encrypt or decrypt data with accurate attributes and password. The proposed technique provides a dual layer security for all Data Node as data is not confined to a specific device and clients can access the system and data from anywhere. This encryption approach may be reckoned as a premise for visualizing and designing even more robust approaches to ensure optimum security of big data.

4.3. Data Mining Technology

To adapt data mining techniques within big data analytics to intelligently perform behavior mining of access controls, authentication and incident logs [18]. Data mining technologies are on the rise to identify vulnerabilities and risks in big data and to predict any threats as a prevention technique from any possible malicious attack [46] [47] [48]. Role mining algorithms automatically extract and optimize the roles that can be automatically generated based on the user's access records for efficiently providing personalized data services for mass users. However, in big data environment, it is important to ensure the dynamic changes and the quality of data pertaining to the roles assigned to users and roles related to the permissions-set that simplify rights management. In big data environment, it may not be possible to accurately specify the data which users can access. In such a context, adopting risk-adaptive access controls using statistical methods and information theory would be applicable. However, defining and quantifying the risks are quite difficult. Hence, authentication based on behavior characteristics of users could be adopted, but the big data system needs to be trained with the training dataset as a continuous process. Incident logs pertaining to the Intranets, Internet, social, and IoT networks as well as email servers could be analyzed to detect abnormal behavior or anomaly patterns using appropriate data mining techniques [18] [32] [49]. While traditional threat analysis cannot cope with big data, by using behavior mining of metadata of various re-

source pools related to big data, anomalies can be analyzed to predict the threats, such as an APT attack. In behavior mining, trend analysis is performed, and pattern proximity is measured to define the relation between datasets. A distance function is usually used to measure the pattern proximity [28]. The distance function defines the proximity between two datasets based on their attributes. It is obvious that a group of datasets which has the minimum distance value between them belong to the same cluster. The most popular general distance function $d_{i,j}$ between two datasets x_i and x_j with p attributes is the Minkowski distance metric in the normed vector space of order m, and is used to calculate the pattern proximity as follows:

$$d_{ij} = \sqrt[m]{\sum_{k=1}^{p} \left(x_{ik} - x_{ik}\right)^m}$$

when $m = 2$, the Minkowski distance is the commonly used Euclidean distance metric as follows:

$$dist\left(x_i, x_j\right) = \left\|x_i - x_j\right\| = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{ik}\right)^2}$$

The Euclidean function works well when the datasets exhibit compact or isolated clusters and is suitable for patterns with multiple dimensions [18]. Big data security can be enhanced by studying the pattern proximity to predict threats by training with the similarity metrics of distances between anomaly datasets and normal datasets based on server/network logs, historical data of incidents and social media data. However, threat detection schemes require scalability and interoperability for the big data environment.

The parallel framework offered by MapReduce for proximity mining is a good fit for implementing data mining technologies as it can perform efficient data-intensive computations and machine learning in providing high scalability and support with large heterogeneous data sources. MapReduce uses two functions called Map and Reduce that process list of pairs <key, value>. The Map function inputs a list of keys and associated values and produces a list of intermediate <key, value> pairs. As shown in **Figure 4**, the feature attributes determined to perform proximity mining between datasets are used for representing the <key, value> pairs in the MapReduce framework. Next, grouping and shuffling of intermediate pairs are performed in proximity mining for determining Matched Pairs and Matched Classes by adapting the distance measures for similarity metrics. Finally, the Reduce function performs the merge operation for the Fusion of Classes on all intermediate pairs to arrive at the final result. The MapReduce framework can be implemented using a Hadoop Distributed File System (HDFS) as it provides parallel and scalable architecture for a large-scale data storage based on clusters in the cloud [18] [50]. However, Hadoop systems are also being attacked as data leakages in the data mappers could take place either intentionally or unintentionally in a cloud cluster.

For instance, in real-life scenarios, a common method to protect the servers

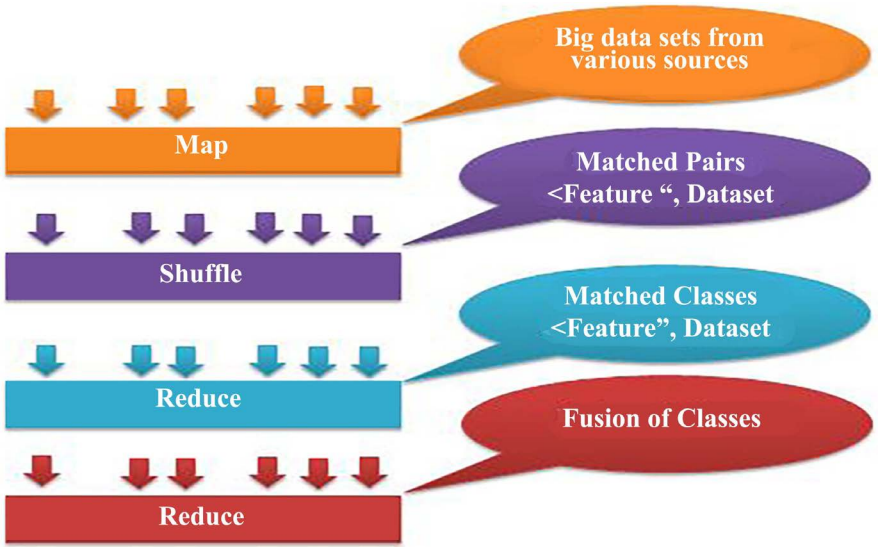

Figure 4. MapReduce framework for proximity mining of big datasets.

from malicious attacks is to record server logs that are analyzed for anomalies. In such contexts, a MapReduce model for distance-based anomaly analysis can be deployed to find k nearest neighbors for a data point and to use its total distance to the k nearest neighbors as the anomaly score in a data mining algorithm [51]. In order to perform the anomaly detection task, the MapReduce functionality can be divided into two jobs as given below:

- A MapReduce job to find pair wise distance between all data points.
- A MapReduce job to find the k nearest neighbors of a data object and to find the weight of the object with respect to the k nearest neighbors.

Data objects can be partitioned by hashing the object ID and all possible combination of hash values for every two object types can be considered [18]. The Euclidean distance computation between objects for each hash value pair can be distributed among the reducers with the mapper output key as a function of the two hash values. However, configurable number of hash buckets should be chosen appropriately to distribute the load uniformly across the reducers by using Hadoop's default reducer partitioner [18]. The parallel processing feature of Hadoop speeds up the processing time resulting in a real-time data mining of large datasets of log files efficiently for detecting anomalous events in the server. In this paper we chose proposed method for anomaly detection in log files, based on data mining techniques for dynamic rule creation [52]. To support parallel processing, this method employs Apache Hadoop framework, providing distributed storage and distributed processing of data. Outcomes of its testing show potential to discover new types of breaches and plausible error rates below 10%. Also, rule generation and anomaly detection speeds are competitive to other used algorithms, such as FP-Growth and Apriori. Such approach generates rules dynamically from certain patterns in sample files and is able to learn new types of attacks, while minimizing the need of human actions. Its implementation utilizes Apache Hadoop framework for distributed storage and distributed

processing of data, allowing computations to run in parallel on several nodes and therefore, speeding up the whole process. Compared to this method Java implementation, single-node cluster Hadoop implementation performs more than ten times faster. Another optimization was done in transformation of data into binary form, making it more efficient to analyze particular transactions.

4.4. Blockchain Technology

To adapt distributed trusted system based on blockchain for big data security and privacy protection [18]. Blockchain technology has demonstrated to be the new mode of trusted interaction and exchange of information by eliminating intermediate parties and supporting direct communication between two parties in a network through replication of information and validation processes [53]. A blockchain is a shared ledger technology ensuring appropriate visibility for any participant in the business network to see the system of record (ledger). In a blockchain system, all transactions are secure, authenticated and verifiable as all parties agree to a network verified transaction and is suitable for applications that need trust with properties such as consensus, immutability and provenance [54] [55]. Overall, it can be well-suited for big data security for various organizations. A blockchain represents a decentralized peer-to-peer network with a historical archive of decisions and actions undertaken. Blocks retain data via a list of transactions and are chained together through each block containing the hash of the previous block's header, thus forming the blockchain with inherent encryption process. **Figure 5** shows a generalized blockchain in operation with blocks storing the proof of transactions, timestamp, and other information on a user activity. Blockchains are maintained through the consensus of a set of nodes running blockchain software, called mining nodes.

In big data context, each data item or record in a database is a block containing transaction details including the transaction date and a link to the previous block. The integrity of data is maintained in blockchain technology. This is because corrupted data cannot enter into the blockchain as checks are carried out continuously in search of patterns in real-time by the various computers on the network. Also, blockchain allows sharing of data more wisely as contracted by the users, thereby preventing cybercrime and data leakage. Blockchain data could also provide valuable insights into the behaviors, trends and can be used for predictive analytics.

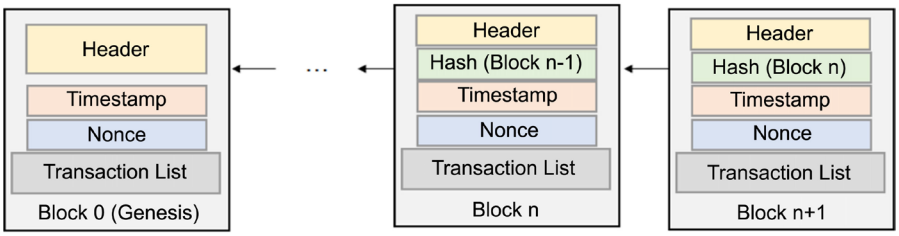

Figure 5. Generic blockchain in operation.

In a typical business environment, there is partial trust within a company or between companies, and IBM's Hyperledger Fabric blockchain could be adopted. A recent study used such a blockchain technology to implement access control of big data by combining two existing access control paradigms: 1) Identity-Based Access Control (IBAC), and 2) Role-Based Access Control (RBAC) [56]. The Blockchain Identity-Based Access Control Business Network (BIBAC-BN) uses the Hyperledger composer that consists of a model resource with definitions of a person participant, data asset with access control of five operations implemented [18]:

1) Request access,
2) Grant access,
3) Revoke access,
4) Verify access, and
5) View asset.

For the BRBAC-BN model, the resource file contains definitions of the persons and all organization participants so that access privileges are granted to a subset of the big data based on their roles. This way, users' roles enable them to access an asset only if it is enabled on the blockchain for the assigned specific roles. Overall, the three phases of big data life cycle, namely big data acquisition, big data storage, and big data analytics can have the additional layer of security when implemented on a blockchain network. However, some of the concerns associated with the blockchain technology are needed to be considered as listed below [18]:

- Irreversibility: encrypted data may be unrecoverable when the private key of a user is lost.
- Adaptability challenges: organizations need to adapt the technology for integrating it in their existing supply chain systems, which may require a big change management and learning curve.
- Current limitations: there are high operational costs associated with running blockchain technology as it requires expert developers, substantial computing power, and revamping resources to cater to its storage limitations.
- Risks and threats—while the blockchain technology greatly addresses the security challenges of big data, it is not threat proof. If the attackers are able to penetrate into majority of the network, then there is a risk of losing the entire database. In this paper blockchain implementations using [57] possibly classified into three categories [58] [59] that vary from each other by the different permission levels that different categories of participants are assigned to.
- Public blockchains are accessible to all participants, anywhere in the world. Anyone can join or leave the network at any time, record a transaction, take part in the validation of the blocks or obtain a copy of them, without any previous control.
- Permissioned blockchains have rules that set out who can take part in the validation process or even register transactions. They can, depending on the

case, be accessible to all or be restricted.

- Private blockchains are controlled by a unique actor who alone oversees participation and validation.

Due to the Bitcoin and similar cryptocurrencies, the first classification is the most-known, as these digital currencies tend to operate in public Blockchains, where any participant in the network can see all transactions already made and update the ledger with new ones. This is also the riskiest type of Blockchain, according to [60]. Permissioned Blockchains allow any user to see the history of transactions, but only selected members can update it. Because it contains more restrictive rules about who can participate, observe and validate transactions, this model is emerging in industry sectors, being used for the exchange of tangible and intangible assets between enterprises. Finally, according to some experts [3], the parameters of the private Blockchains do not respect the traditional properties of Blockchains, such as decentralization and shared validation. In any case, private Blockchains do not raise specific issues regarding their compliance with the EU GDPR. traditional distributed databases can be considered.

5. Conclusion and Future Work

We introduced Big Data challenges from more than one view. One view of these challenges divides challenges into categories; some of them come back to a function of the characteristics of BD, by its existing analysis methods and models, and some, through the limitations of current data processing system. Another view seeks appropriate protection and privacy needs to be enforced throughout the big data lifecycle phases. To avoid these challenges, we propose the use of four main technologies to comprehensively cope with the 56 V during the three phases of big data lifecycle, namely data acquisition, data storage and data analytics. The first technology we propose to use is in Data Provenance Technology. Here, we chose a promising algorithm by Zirije Hasani and Samedin Krrabaj [33] that is evaluated on the known anomaly detection benchmarks NUMENTA and Yahoo datasets with annotated anomalies and real log data generated by the National education information system. Based on the experimental evaluation of the detection rate and precision, performed on sets of synthetic and real data periodic streams, we can conclude that using proposed HW with GA optimized parameters (α, β, γ, δ, k, n) and with improved MASE outperforms the other algorithms. The second technology we propose to use is in Data Encryption and Access Control Technology, here to adapt advanced encryption techniques and access control schemes in big data storage systems. We chose Attribute Based Honey Encryption (ABHE) methodology to solve the issue of data security in Hadoop storage as in [45]. It shows considerable improvement in performance during the encryption-decryption of files. This approach works on files that are encoded inside the HDFS and decoded inside the Mapper. This proves that it is fast enough to secure the data without adding delay. Also, it has a higher throughput which proves its applicability on big data. The proposed ABHE algo-

rithm has evaluated by performing encryption-decryption on different sizes of files and has compared the same with existing ones including AES and AES with OTP algorithms that show considerable improvement in performance during the encryption-decryption of files. It provides a dual layer security for all Data Node as data is not confined to a specific device and clients can access the system and data from anywhere. The third technology we propose to use is in Data Mining Technology; here we intended to adapt data mining techniques within big data analytics to intelligently perform behavior mining of access controls, authentication and incident logs [18]. We chose proposed method for anomaly detection in log files, based on data mining techniques for dynamic rule creation [52]. To support parallel processing, this method employs Apache Hadoop framework, providing distributed storage and distributed processing of data. Outcomes of its testing show potential to discover new types of breaches and plausible error rates below 10%. Compared to this method Java implementation, single-node cluster Hadoop implementation performs more than ten times faster. The fourth technology we propose to use it in Blockchain Technology; here we intended to adapt distributed trusted system based on blockchain for big data security and privacy protection. Blockchain implementations are classified using [57] into three categories that vary from each other by the different permission levels named as, Public Blockchains, Permissioned Blockchains, and Private Blockchains. This classification gives everyone different privileges according to its need. This paper (BD2) is considered complete in our series of big data papers. We started our series with paper titled (BD1) [1], which introduces excellent details about Old and New Big Data V's Characteristics and its applications. Future work, which is already in progress, is "Using Hadoop Technology to Overcome Big Data Trials (BD3)".

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Hussien, A.A. (2020) How Many Old and New Big Data V's Characteristics, Processing Technology, and Applications (BD1). *International Journal of Application or Innovation in Engineering & Management*, **9**, 15-27. http://www.ijaiem.org/

[2] Priyadarshy, S. (2015) The 7 Pillars of Big Data. *Petrolium Review*, 14 January.

[3] Trifu, M.R. and Ivan, M.L. (2014) Big Data: Present and Future. *Database Systems Journal*, **1**, No. 1.

[4] Firican, G. (2017) The 10 V'S BIG DATA. Work Paper, 8 February.

[5] Borne, K. (2014) Top 10 Big Data Challenges—A Serious Look at 10 Big Data V's. *Blog Post*, 11 April.

[6] Vorhies, W. (2014) How Many V'S in Big Data. View Blog Work Paper, 31 October. http://www.aimspress.com/journal/Math

[7] Dhamodharavadhani, S. and Rajasekaran, G. (2018) Unlock Different V's of Big Data for Analytics. *International Journal of Computer Sciences and Engineering*, **6**, Special Issue-4.

[8] Dr. Darrin (2016). https://educationalresearchtechniques.wordpress.com/2016/05/02/characteristics-of-big-data/

[9] Gartner (2013). http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definiti-consists-of-three-parts-not-to-be-confused-with-three-vs/#5040d1cc3bf6

[10] https://hbr.org/2012/10/making-advanced-analytics-work-for-you/ar/1,2012

[11] Gewirtz, D. (2016). http://www.zdnet.com/article/volume-velocity-and-varietyunderstanding-the-three-vs-of-big-data/

[12] GoodStratTweet. (2015). http://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077?Page_number=1

[13] Laney, D. (2012) http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/

[14] Cartledge, C. (2016) How Many vs Are There in Big Data? Working Paper, 18 February.

[15] Borne, D. (2014). https://www.mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs

[16] Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017) Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Business Research*, **70**, 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

[17] Boyd, D. and Crawford, K. (2012) Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society*, **15**, 662-679. https://doi.org/10.1080/1369118X.2012.678878

[18] Venkatraman, S. and Venkatraman, R. (2019) Big Data Security Challenges and Strategies. *AIMS Mathematics*, **4**, 860-879. https://doi.org/10.3934/math.2019.3.860

[19] Hargittai, E. (2013) Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, **659**, 63-76. https://doi.org/10.1177/0002716215570866

[20] Mayer, K., *et al.* (2009) Computational Social Science. Schlüsselwerke der Netzwerkforschung ook.

[21] Wang, Y.X. and Wiebe, V.J. (2014) Big Data Analytics on the Characteristic Equilibrium of Collective Opinions in Social Networks. *International Journal of Cognitive informatics and Natural Intelligence*, **8**, Article No.: 3. https://doi.org/10.4018/IJCINI.2014070103

[22] Akerkar, R. (2014) Big Data Computing. Taylor & Francis Group, CRC Press, New York. https://doi.org/10.1201/b16014

[23] Zicari, R.V. (2014) Big Data: Challenges and Opportunities. *Big Data Computing*, 103-128.

[24] Agrawal, R. and Nyamful, C. (2016) Challenges of Big Data Storage and Management. *Global Journal of Information Technology*, **6**, 1-10. http://sproc.org/ojs/index.php/gjit https://doi.org/10.18844/gjit.v6i1.383

[25] Tarekegn, G.B. and Munaye, Y.Y. (2016) Big Data: Security Issues, Challenges and Future Scope. *International Journal of Computer Engineering & Technology*, **7**, 12-24.

[26] Cavanillas, J.M., Curry, E. and Wahlster, W. (2017) New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe. Library of Congress Control Number: 2015951834. https://doi.org/10.1007/978-3-319-21569-3

[27] Kumar, N., Vasilakos, A.V. and Rodrigues, J.J.P. (2017) A Multi-Tenant Cloud-Based DC Nano Grid for Self-Sustained Smart Buildings in Smart Cities. *IEEE Communications Magazine*, **55**, 14-21. https://doi.org/10.1109/MCOM.2017.1600228CM

[28] Yao, Z., Mark, P. and Rabbat, M. (2012) Anomaly Detection Using Proximity Graph and PageRank Algorithm. *IEEE Transactions on Information Forensics and Security*, **7**, 1288-1300. https://doi.org/10.1109/TIFS.2012.2191963

[29] Yan, Z., Ding, W., Niemi, V. and Vasilakos, A.V. (2016) Two Schemes of Privacy-Preserving Trust Evaluation. *Future Generation Computer Systems*, **62**, 175-189. https://doi.org/10.1016/j.future.2015.11.006

[30] Rebecca Webb (2018) 12 Challenges of Data Analytics and How to Fix Them. Risk management Blog-Clearrisk. https://www.clearrisk.com/risk-management-blog/challenges-of-data-analytics

[31] Puthal, D., Nepal, S., Ranjan, R. and Chen, J.J. (2017) A Dynamic Prime Number Based Efficient Security Mechanism for Big Sensing Data Streams. *Journal of Computer and System Sciences*, **83**, 22-42. https://doi.org/10.1016/j.jcss.2016.02.005

[32] Akoglu, L., Tong, H.H. and Koutra, D. (2015) Graph Based Anomaly Detection and Description: A Survey. *Data Mining and Knowledge Discovery*, **29**, 626-688. https://doi.org/10.1007/s10618-014-0365-y

[33] Hasani, Z. and Krrabaj, S. (2019) Survey and Proposal of an Adaptive Anomaly Detection Algorithm for Periodic Data Streams. *Journal of Computer and Communications*, **7**, 33-55. https://doi.org/10.4236/jcc.2019.78004

[34] Hasani, Z., Jakimovski, B., Velinov, G. and Kon-Popovska, M. (2018) An Adaptive Anomaly Detection Algorithm for Periodic Real Time Data Streams. In: *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, Berlin, 385-397. https://doi.org/10.1007/978-3-030-03493-1_41

[35] Shahin, A.A. (2016) Using Multiple Seasonal Holt-Winters Exponential Smoothing to Predict Cloud Resource Provisioning. *International Journal of Advanced Computer Science and Applications*, **7**, 91-96. https://doi.org/10.14569/IJACSA.2016.071113

[36] Cortez, P., Rocha, M. and Neves, J. (2001) Genetic and Evolutionary Algorithms for Time Series Fore-Casting. In: Monostori, L., Váncza, J. and Ali, M., Eds., *Engineering of Intelligent Systems*, *Lecture Notes in Computer Science*, Vol. 2070, Springer, Berlin, Heidelberg, 393-402. https://doi.org/10.1007/3-540-45517-5_44

[37] de Assis, M.V.O., Carvalho, L.F., Rodrigues, J.J.P.C. and Proença, M.L. (2013) Holt-Winters Statistical Forecasting and ACO Metaheuristic for Traffic Characterization. *IEEE International Conference on Communications*, Budapest, 9-13 June 2013, 2524-2528. https://doi.org/10.1109/ICC.2013.6654913

[38] Scrucca, L. (2013) GA: A Package for Genetic Algorithms. *Journal of Statistical Software*, **53**, 1-37. https://doi.org/10.18637/jss.v053.i04

[39] NUMENTA, Anomaly Benchmark with Labeled Anomalies. https://github.com/numenta/NAB/tree/master/data/artificialWithAnomaly

[40] Yahoo: S5-dA, Anomaly Detection Dataset, Version 1.0(16M). https://webscope.sandbox.yahoo.com/catalog.php?datatype=s%5c&did=70

[41] Zhou, G.M., Zhang, D.X., Liu, Y.J., *et al.* (2015) A Novel Image Encryption Algorithm Based on Chaos and Line Map. *Neurocomputing*, **169**, 150-157. https://doi.org/10.1016/j.neucom.2014.11.095

[42] Wang, Z.W., Cao, C., Yang, N.H. and Chang, V. (2017) ABE with Improved Auxiliary Input for Big Data Security. *Journal of Computer and System Sciences*, **89**, 41-50. https://doi.org/10.1016/j.jcss.2016.12.006

[43] Kshetri, N. (2014) The Emerging Role of Big Data in Key Development Issues: Opportunities, Challenges, and Concerns. *Big Data & Society*, **1**, 1-20. https://doi.org/10.1177/2053951714564227

[44] Hsu, C., Zeng, B. and Zhang, M. (2014) A Novel Group Key Transfers for Big Data Security. *Applied Mathematics and Computation*, **249**, 436-443. https://doi.org/10.1016/j.amc.2014.10.051

[45] Kapil, G., Agrawal, A., Attaallah, A., Algarni, A., Kumar, R. and Khan, R.A. (2020) Attribute Based Honey Encryption Algorithm for Securing Big Data: Hadoop Distributed File System Perspective. *PeerJ Computer Science*, **6**, e259. https://doi.org/10.7717/peerj-cs.259

[46] Wu, X.D., Zhu, X.Q., Wu, G.-Q. and Ding, W. (2014) Data Mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 97-107. https://doi.org/10.1109/TKDE.2013.109

[47] Xiao, H., Biggio, B., Brown, G., *et al.* (2015) Is Feature Selection Secure against Training Data Poisoning? *Proceedings of the* 32*nd International Conference on Machine Learning*, Lille, 1689-1698.

[48] Fuchs, G., Stange, H., Hecker, D., *et al.* (2015) Constructing Semantic Interpretation of Routine and Anomalous Mobility Behaviors from Big Data. *SIGSPATIAL Special*, **7**, 27-34. https://doi.org/10.1145/2782759.2782765

[49] Miller, B.A., Beard, M.S. and Bliss, N.T. (2011) Eigen space Analysis for Threat Detection in Social Networks. *Proceedings of the* 14*th International Conference on Information Fusion*, Chicago, 5-8 July 2011, 1-7.

[50] Hota, S. (2018) Big Data Analysis on YouTube Using Hadoop And Mapreduce. *International Journal of Computer Engineering in Research Trends*, **5**, 98-104.

[51] Remya, G. and Mohan, A. (2015) Distributed Computing Based Methods for Anomaly Analysis in Large Datasets. *International Journal of Advanced Research in Computer and Communication Engineering*, **4**, 427-430.

[52] Breier, J. and Branišová, J. (2015) Anomaly Detection from Log Files Using Data Mining Techniques. *Lecture Notes in Electrical Engineering*, **339**, 449-457. https://www.researchgate.net/publication/282923954 https://doi.org/10.1007/978-3-662-46578-3_53

[53] Restuccia, F., Salvatore d'oro, Kanhere, S.S. and Melodia, T. (2018) Blockchain for the Internet of Things: Present and Future. *IEEE Internet of Things Journal*, **1**, 1-8.

[54] Christidis, K. and Devetsiokiotis, M. (2016) Blockchains and Smart Contracts for the Internet of Things. *IEEE Access*, **4**, 2292-2303. https://doi.org/10.1109/ACCESS.2016.2566339

[55] Yaga, D., Mell, P., Roby, N. and Scarfone, K. (2018) Blockchain Technology Overview. National Institute of Standards and Technology, U.S. Department of Commerce, 1-27. https://doi.org/10.6028/NIST.IR.8202

[56] Uchibeke, U.U., Schneider, K.A., Kassani, S.H. and Deters, R. (2018) Blockchain

Access Control Ecosystem for Big Data Security. 2018 *IEEE International Conference on Internet of Things* (*iThings*) *and IEEE Green Computing and Communications* (*GreenCom*) *and IEEE Cyber*, *Physical and Social Computing* (*CPSCom*) *and IEEE Smart Data* (*SmartData*), Halifax, 30 July-3 August 2018, 1373-1378.

[57] Ramos, L.F.M. and Silva, J.M.C. (2019) Privacy and Data Protection Concerns Regarding the Use of Blockchains in Smart Cities. *ICEGOV*2019, Melbourne, 3-5 April 2019, 342-347.

[58] Maull, R., Godsiff, P., Mulligan, C., Brown, A. and Kewell, B. (2017) Distributed Ledger Technology: Applications and Implications. *Strategic Change*, **26**, No. 5. https://doi.org/10.1002/jsc.2148

[59] Tasca, P. and Tessone, C.J. (2017) Taxonomy of Blockchain Technologies. Principles of Identification and Classification. https://ssrn.com/abstract=2977811

[60] CNIL (2018) Solutions for a Responsible Use of the Blockchain in the Context of Personal Data. Technical Report. Commission Nationale Informatique & Libertés. https://www.cnil.fr/sites/default/files/atoms/files/blockchain.pdf