VOL 4 (2020) NO 4

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Big Healthcare Data: Survey of Challenges and Privacy

Mohammed BinJubeir[#1], Mohd Arfian Ismail[#2], Shahreen Kasim[*], Hidra Amnur[**1], Defni[**2]

[#] *Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang (UMP), 26300, Kuantan, Pahang, Malaysia.*
E-mail: [1]moh77421143@gmail.com, [2]arfian@ump.edu.my

* *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia*
E-mail: shahreen@uthm.edu.my

** *Department of Information Technology, Politeknik Negeri Padang, Sumatera Barat, Indonesia*
E-mail: hidraamnur@gmail.com, defni@pnp.ac.id

***Abstract*— The last century witnessed a dramatic leap in the shift towards digitizing the healthcare workflow and moving to e-patients' records. Health information is consistently becoming more diverse and complex, leading to the so-called massive data. Additionally, the demand for big data analytics in healthcare organizations is increasingly growing with the aim of providing a wide range of unprecedented potentials that are considered necessary for the provision of meaningful information about big data and improve the quality of healthcare delivery. It also aims to increase the effectiveness and efficiency of healthcare organizations; provide doctors and care providers better decision-making information and help them in the early detection of diseases. It also assists in evidence-based medicine and helps to minimize healthcare cost. However, a clear contradiction exists between the privacy and security of big data and its widespread usage. In this paper, the focus is on big data with respect to its characteristics, trends, and challenges. Additionally, the risks and benefits associated with data analytics were reviewed.**

***Keywords*— big healthcare data; big data; MapReduce; Hadoop; security; privacy.**

I. Introduction

In the past few years, the broad impact of the emerging computing techniques has reinforced the generation of massive data volumes, known as "*big data*." This has led to a profound transformation of our society and has attracted the attention of several researchers in the field of information sciences. It is obvious that the huge of data generated through the emerging ubiquitous computing processes is continuously expanding. Currently, the world is experiencing an era of data deluge as evidenced by the massive volume of generated data which keeps increasing with time. Data volume has been doubling every 2 years since 2011[1][2]; data generated by the U.S. healthcare system alone has been reported to reach 150 exabytes in 2011 and is expected to soon reach the zettabyte (2^{21} gigabytes) and yottabyte (2^{24} gigabytes) scales [3]. The amount of data currently used on a daily basis is believed to be more than the entire data used by our ancestors all through their generation [4]. The term '*big data'* was invented to describe the thoughtful meaning of the massive data volume. As per several scholars, *big data* can be seen as the revolution of the digital era when considering its importance to the society [5]. One feature of these data is that they are normally complex and unstructured; a significant amount of these data comes from processes like sales records, IoT sensors, medical patient records, social media, image, and video archives. *Big data* processing using traditional data. This technological revolution has increased the interest in *big data* among both researchers, government decision makers, and technological experts. With the ever-increasing rate of Internet usage and the increasing number of connected devices to ubiquitous computing, there is a need to transform the huge amounts of generated data into different formats to extract valuable information. This will reveal the information embedded in the huge data and provide several opportunities with great unprecedented benefits in many fields [6]. Healthcare is one of the fields where the application of *big data* can bring about significant changes. It could considerably improve the quality of healthcare delivery and enhance the effectiveness and efficiency of the healthcare organization. This can be achieved by obtaining valuable insights that will help to improve patient outcomes, reduce healthcare delivery cost, avoid preventable diseases, and improve the general quality of life. However, the potentials of *big data* are yet to be realized as the mere availability of data does not translate into knowledge or clinical practice. The preservation of patients' rights and ensuring the security of their information is a difficult task[7][8][9]. Besides, the privacy of individuals may be violated by using personal information for other purposes other than what it was intended for. So, the realization of the potentials of *big data* towards medical science advancement, as well as its significance to the success of healthcare organizations demands to address data privacy and security concerns [10][11][12].

This article overviews the concept of *big data* in healthcare, starting with the definition and discussion of the features of *big data* in healthcare. Then, the capabilities of *big data* analytics and its limitations were identified. Lastly, *big data* processing capability was discussed, followed by the description of how some of these platforms work.

II. Big Data Definition

Generally, the definition of *big data* is relatively new in IT and business and rather diverse due to the rapid evolution of *big data*, and reaching a consensus is difficult. Several organizations have strived to define *big data*; for instance, Cox & Ellsworth [13] defined *big data* in 1997 as "a large volume of data produced by the digital world for visualization". However, Mckinsey [14] defined *big data* in 2011 as "a set of huge multi-source data sets with a great diversity such that it is difficult to capture, store, manage, and process them efficiently using the recent or traditional data processing techniques." Several researchers and practitioners have defined *big data* based on certain major features or dimensions: a process known as the *Vs* model. This model has played an important role in determining the description of *big data* concept in different fields like healthcare where it began to acclimatize to today's digital data era. New technologies help in capturing most of the healthcare information over a large timescale. However, such information has vastly remained underutilized despite the advent of medical electronics, and thus, wasted. The following subsection explained some important features or dimensions of *big data* in all disciplines, including *big data* in healthcare. This will help to understand both the challenges and advantages of *big data* initiatives [8].

III. V's of Big Data

Big data does not mean only a huge volume of data; it is an opportunity towards finding information on the emerging data types and content which may not have been inferred if such data is not processed. The definition of big data based on where it begins and where the targeted usage become a big data project requires a consideration of the key attributes of big data. Big data definition is commonly based on the Vs model as it can help to understand both the challenges related with big data and the advantages of big data initiatives [15].

In addition, breakthrough leaps of data have resulted in even more added challenges in the area of big data. These dimensions are proposed as candidates in identifying the challenges of big data. Each feature or dimension describes a specific property of big data and none of the features stand on its own in identifying big data from not-so-big-data. The confusing questions in the number of V's remain: What are the three, four, ten (or more) most important V's in big data and what are the widely accepted V's of big data? [2].

There are originally only 3 big data dimensions - high-volume, high-velocity, and variety (called the 3V's). These were first introduced in 2001 by Gartner analyst Doug Laney long before "big data" gained popularity [16]. As the level of data generated by enterprises keeps growing, most of these data are incomplete or poorly architected. The constantly increasing data volume and the diversity of resources and contents of big data cannot be properly depicted by the 3Vs model. This has led to the addition of more Vs to the list to describe big data and identify certain characteristics and specific dimensions of big data [17].

Therefore, the International Data Corporation (IDC) defined big data in 2011 as [18] "a new technological and architectural generation designed to economically extract value from huge data volumes by enabling high-velocity capture, discovery, and/or analysis". By this definition, big data is specified as not just characterized by the earlier mentioned 3 Vs but may be up to 4 Vs which are volume, variety, velocity, & value. This 4 Vs definition of big data is more recognized as it portrays the necessity and meaning of big data.

International Business Machines Corporation (IBM) presented veracity as the fifth V characteristic of big data. Veracity addresses the inherent data trustworthiness. Since big data is used in critical processes (e.g. for decision making), it is necessary to ensure that it can be trusted [19].

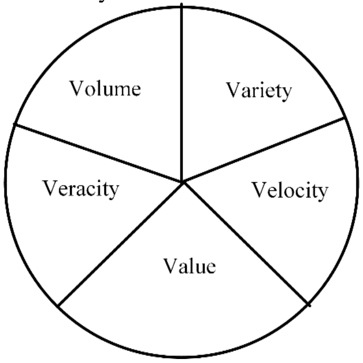

Fig 1. 5Vs of Big Data.

As it turns out, big data practitioners have a general conception that certain sets of criteria must be met in order to arrive at "big" data. Such criteria are high-volume, high-velocity, and variety (commonly called the 3Vs of big data). The 3Vs model is still used in many industries to describe big data. However, big data is not just about 3Vs but have some other features. Some have gone to add more Vs to the list, such as veracity, value, etc. To process breakthrough changes in data and for efficient business operations and profits, big data must be analyzed carefully by the organizations to reach better decisions. Currently, this technology is being used in several areas, but one of the areas where it can bring a huge change is in healthcare. In the healthcare sector, this technology can transform the healthcare system and increase their effectiveness and efficiency [20] [1][2]. [1][2]. McKinsey Global Institute believed that the US healthcare sector could make >$300 billion per year in value if it were to creatively and effectively use big data. A good portion of this value would be accrued from reducing the cost of the US healthcare system [8]. So, a discussion and understanding of the 5Vs will open doors towards finding the true value of big data in healthcare. This is depicted in Figure 1.

Volume: It is synonymous with the term "big" in big data [21]. Volume is a critical and important factor for differentiating big data from normal data. It could be said that if volume is removed from big data, such a data will not be big enough to be considered as big data, hence, will becomes a small set of data. Volume describes the amount of all types of data generated from different sources per second. According to a survey by IBM in 2012 [22], about 50 % of the 1144 participants classified datasets over one terabyte as big data. Whereas the collection and storage of big data require considerable effort and underlying investments with the rising quantity of data, the challenge of big data collection comes from trying to find innovative ways of processing information which will provide more insight and help people in making decision and automating processes.

Velocity: The description of big data can be based on its velocity/speed [23]. Velocity refers to the speed with which data are being generated and its relative accessibility, storage, and analysis. Velocity makes it possible for organizations to understand the relative growth of their big data and where they come from. Velocity aims at data analysis based on their speed of generation [16]. The Internet usage, along with the number of connected devices, has caused a constant data flow at a rate that has made it difficult to analyze such data using the traditional systems [16].

Variety refers to the different data forms, such as video, image, text, audio, & data logs which are collected from different sources. Internet usage, along with growing the number of connected devices, has led to high data generation. As such, there are several types of data, including structured, unstructured, & semi-structured data [21]. Structured data refers to data that is clustered into relational schemes with specific formats and lengths. The data consistency & configuration may allow easy dealing with these data and respond on simple queries to arrive at usable information [24]. Semi-structured data refers to a form of structured data with semantic tags which do not conform to the structure of typical relational databases (also known as self-describing structure) [21].

Value refers to the value that could be extracted from certain data and how big data analysis techniques could increase the extraction of useful information while increasing the flow of data [25]. The analysis of big data has increasingly become a hot field that several organizations are depending on it to derive vital information from big data. Meanwhile, several data can be captured in other situations, but there is no benefit from such data. It is costly to implement the required IT infrastructure for big data storage; hence, businesses must demand investment returns [25][16].

Veracity refers to the degree of confidence in the information to make a decision. Therefore, finding useful and accurate data from the "dirty data" is very important. Confidence generation is, therefore, a major challenge as the number and type of sources grow. The "dirty data" can easily result in several errors, incorrect results, and costly big data environment. Veracity is the reliability, accuracy, and context of the data source; it represents how meaningful it is to depend on such data for analysis [26].

The aforementioned 5Vs provide a different set of features for big data that differentiate big data concept from 'massive data' and 'very large data' concepts. In all fields, including healthcare, big data requires significant resources, powerful technologies and new methods to analyze, process, clean, secure and provide access to big data. These are impossible to be achieved using common or traditional data management methods [3]. Big data is attractive in the healthcare sector not just for its volume, but equally because of the data diversity and speed required for its management [27].

As such, new platforms that focus on big data storage and processing have emerged [28]. These platforms consist of several servers with a wide range of analytical platforms. Each platform excels in a specific aspect of big data analytics with a competitive advantage. In other words, there isn't one single platform that provides all the capabilities. Besides, there are common capabilities on all platforms which can be divided into data analysis capability and data processing capability. On this occasion, how to manage the core capabilities of these platforms to work efficiently and ensure the entire system running is a tremendous challenge [28][29]. The following subsections considered the common core capabilities of such platforms

IV. Data Analysis Capability

Data analysis capability is the closest component to the users in the data processing platform. It aims at keeping away the complex technical details in the bottom layer of the processing platform via abstract data access and analysis to extract useful information that can represent, interpret, or identify significant patterns (a process called *data mining* and is considered one of the subfields of computer science) [11][30][31][32][33].

The term "*data mining*" can be considered another term for knowledge discovery from data (KDD) which is an alternative expression of the objective of mining processes. The processes involved in *data mining* are patterns discovery & extraction; it also patterns identification and recognition, as well as identification of frameworks normally used in *data mining*. *Data mining* is still considered as just a basic stage during knowledge discovery in certain cases because it has various stages. The stages of knowledge discovery, as depicted in **Error! Reference source not found.**, have an iterative pattern and are presented below as interactive stages [11][30].

Stage 1: Data pre-processing: The processes involved in this stage are data selection, data cleaning (to remove irrelevant data), & data integration (to combine data from different sources).

Stage 2: Transformation of data: This involves the conversion and integration of data the data into proper mining formats.

Stage 3: Data mining: This is the use of intelligent techniques to mine sequences of the data.

Stage 4: The operations in this step are evaluation of the patterns and presentation of the extracted knowledge in an easy-to-comprehend mode.

Data mining provides a range of unprecedented promises and attractive opportunities. It is addressed as the basic to offer the required insights to increase the effectiveness and efficiency of many organizations, including healthcare organizations by obtaining meaningful information from *big data* to improve quality. However, the potentials of *big data* are yet to be realized as scholars are facing several problems when exploring *big data* sets and during knowledge extraction from such mines of information. One of such challenges is issues related to privacy and the problem of internet phishing of data. This issue is threatening the secure propagation of private patients' data over the web. It has led to the limited availability of large clinical data sets to researchers [12][34]. Thus, useful information must be extracted from large data sets to improve the quality of health care delivery and increase the effectiveness and efficiency of health care organizations. Besides the prevention of private health care information disclosure, data must be secured from falling into wrong hands or at risk. This procedure is known as privacy-preserving data mining (PPDM), a novel research area that aims at provision of guaranteed security and privacy for *big data* during [35][36][37].

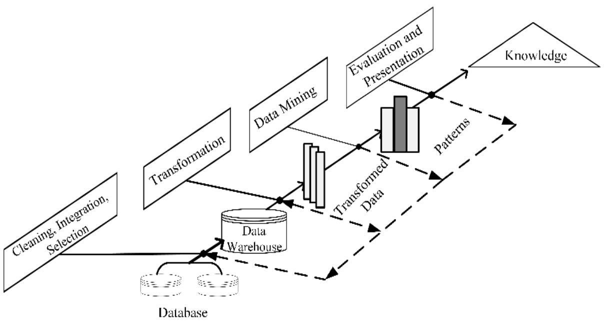

Fig 2. An outline of the KDD process

As depicted in Figure 2, the entire KDD process is comprised of several operation stages, and healthcare data can be exposed to phishing attacks in one of the stages of KDD [11]. People's privacy can be violated due to several factors, such as unauthorized access and use of personal data [11]; hence, the prevention of private healthcare information disclosure can also minimize the chances of data utilization and can cause errors or make knowledge extraction an impossible task through data mining [38].

This gap presents an opportunity for improving field of KDD and resolving privacy-related issues. This is becoming very important with advances in learning technology [11][12][39]. It is necessary that healthcare organizations safeguard and manage personal information during their propagation to various data mining servers. This can be achieved by identifying the least amount of private info required for accurate construction of data mining models[10][38]. There are various technologies towards ensuring the privacy and security of big healthcare data. Such technologies are classified as data perturbation techniques and anonymization-based techniques. The aim of the anonymization techniques is to prevent the recognition of the identity records of the owner in big data. The data providers (healthcare organizations) modify the original data (independently) using the data perturbation techniques before forwarding same to the server [40][41][42]. This process ensures that the garbled data values are used rather than the original values. It also ensures the privacy of individuals when deploying data mining techniques. Each of these approaches will be detailed in the subsequent section. Readers are referred to [35][43][44] and [45] for a complete analysis of these subjects.

A. Anonymization Technique

Big data in healthcare is comprised of a wide range of records (tuples) and each record (tuple) is considered a client that consists of various client-related specific attributes (see Table 1) [46][44]. These attributes can be categorized into Identity Attribute (IA) (identifies the records of the owner, such as the name, address, phone number); Quasi-Identifier (QI) attribute (denotes a set of attributes wherein no single attribute can provide specific identification of the person, rather, all the attributes must be combined to identify the person); and Sensitive Attributes (SA) (denotes the confidential information of the person, such as the disease type) [47][48].

Healthcare organizations may have the intention to publish partial data derived from big data sets which can support future plans in enhancing the effectiveness of the healthcare organizations without divulging the proprietorship of the sensitive data. It has been demonstrated that solely eliminating attributes (IAs) that explicitly identify users from the table is not an efficient approach [38][12]. Where the remaining data in most of these cases can be used to re-identify the person. Therefore, effective preservation of privacy in healthcare can be attained by controlling the disclosure of information which represents a set of individuals' non-explicit attributes (QI). Where, anonymization techniques can be used for this purpose prior to the data release, as they take personal data and make it anonymous or not attributable to one specific source or person by breaking the relations among attribute values [49]. Among the most favourable techniques of anonymity are the K-anonymity approach [50], L-diversity approach [51], and T-closeness approach [46].

TABLE I
MEDICAL PATIENT DATABASE

Identifier (IAs)	**Quasi-Identifier (QI)**			**Sensitive (SA)**
Name	Age	Gender	Zip code	Disease
Mike	29	Male	462350	Heart Disease
Bob	22	Male	462351	Cancer
Michel	27	Male	462352	Flu
Alice	52	Male	462350	Heart Disease
Sofia	38	Female	462350	Heart Disease

B. Data Perturbation Techniques

Perturbation is based on altering the original values of a dataset D to its anonymized version D1 by (1) swapping cells within columns [52], (2) adding noise to the data [53][54], or (3) creating synthetic data [55][56][57]. These made it difficult for an attacker to launch attribute linkage attacks for pinpointing an individual in a published dataset or to infer the exact sensitive value of an individual. It generally brings about uncertainty in published datasets and negatively affects the chances of inferring the individuals sensitive information [58]. The several proposed data perturbation techniques can be classified into dimension-based approaches and value-based approaches. The value value-based approaches, such as Uniform Perturbation approach [43] and Probability Distribution approach [59], focus on single-dimensional perturbation. On the other hand, the dimension-based approaches, such as the Random Rotation Transformation approach [60] and Random Projection approach [43][61], focus on multi-dimensional data perturbation.

Overall, both techniques attempt to protect sensitive information based on the data in use and the way in which it is used [62]. However, optimal anonymization is an "NP-Hard" problem [63][64], Owing to the recent technological developments and the nature of *big data*, the volume of generated data is daily increasing in multiple formats. As such, it is becoming extremely difficult to manage *big data* using the traditional methods [28]. Relating to healthcare, *big data* could be structured, semi-structured, or even unstructured. This has increased the complexity of *big data* processing or storage. Many methods are used to ensure data privacy [62]. Additionally, the techniques of privacy protection consider data utility when effective data mining. Besides the protection of privacy when other users utilize data, the sensitive information must be protected in a manner that an attacker will find difficult to identify the owner of the record. There is currently no existing generic solution to all the privacy issues as relates to sensitive information protection [62][65][38][66].

V. Data Processing Capability

Data processing capability refers to large data management characterized by the "3*Vs*" (high volume, high velocity, and wide variety). It provides reliable and fast data access and the ability to satisfy the demands for *big data* computing. Data processing is a wide concept that consists of procedures, policies, and technology for data collection, storage, organization, administration, governance, and delivery. Also included are data cleansing, migration, preparation and integration for reporting and analytics purposes. Most *big data* environments in today's businesses go beyond relational database platforms and traditional data warehouse which requires powerful technologies and new methods to process the vast quantities of data [28].

It is obvious that the vast quantity of data is slowly changing the way of data analysis capability, procedures and technology used. The renewed attention on *big data* analysis and processing is shaping new platforms for the combination of conventional databases with *big data* systems in logical database architecture. A part of the process is deciding the data aspects that can be discarded and those that can be analyzed to improve the current business processes or enhance competitive advantage. This requires a careful classification of data to ensure a fast analysis of the smaller data sets [10]. The next section discussed the mode of action of some of these platforms

A. The MapReduce platform

MapReduce is a platform that was introduced and used initially by Google in 2003 [67]. It is a simple and easy programming model that enables massive scalability across hundreds or even thousands of server commodity machines. It enables *big data* processing via distributed ways [68][69]. It is based on the Divide and Conquer algorithm. The basic idea of this algorithm is that it splits a large problem (large tasks) into smaller sub-problems to the extent that the subproblems are independent. The sub-problems (subtasks) are handled in parallel by different computing nodes (workers or threads) in a processor core, multiple processors in a machine, cores in a multi-core processor, or many machines in a cluster. Finally, the results are aggregated and returned to the master core [70].

MapReduce program is usually executed based on a master-slave framework. The master machine assigns tasks and controls the slave machines. The execution of a MapReduce program involves two separate jobs, namely-Map and Reduce. After dividing the data set and distributing them across the computing nodes (workers) through the master node, the Map operation is performed. This operation involves taking and converting a set of data into another data set. In the new data set, the individual data elements are disintegrated into tuples (key/value pairs). Once all Map tasks have finished, the Reduce operation is performed by taking the output from a map as an input. The results are grouped by key and redistributed so that all pairs belonging to one key are in the same node [68]. A schematic for the execution of a MapReduce program is given in **Error! Reference source not found.**.

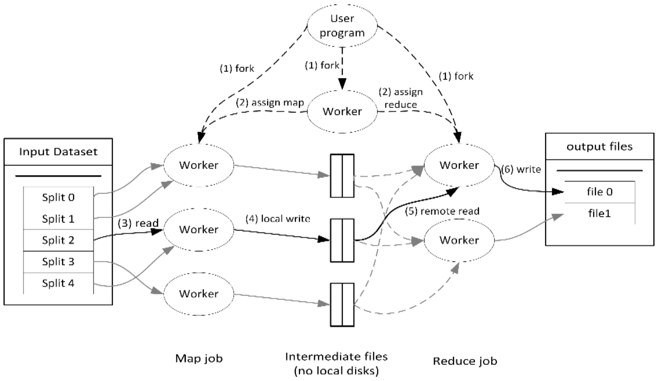

Fig 3. Execution of a MapReduce program [71]

B. Apache Hadoop platforms

Hadoop is a platform managed by the Apache Software Foundation. It allows for solving huge data problems via distributed processing. It also allows the use of simple programming models for computation across clusters of computers. Apache Hadoop framework is considered a realistic standard for data analysis in large-scale. It is one of the oldest available systems for abstracting the problems of distributed computing and fault tolerance. It reduces the entry barriers in the *big data* space [72]. The success of Hadoop later ushered in the creation of systems such as Apache Spark [73] and Apache Flink (formerly Stratosphere) [74]. These new systems offer higher levels of distributed computing. Apache Spark and Apache Flink are considered rivals; they have received much interest due to their merits and drawbacks [75].

Hadoop was initiated as a Yahoo project in 2005; it was created by Doug Cutting and Mike Cafarella [76]. The approach to dealing with an avalanche of data was inspired by papers published by Google. Hadoop has since become a top-level Apache open-source framework for reliable, scalable, and distributed computing. The core of Apache Hadoop is designed from a storage part known as Hadoop Distributed File System (HDFS) and a processing part which is implemented as a MapReduce framework on this file system to process data. It has become a core component of Hadoop [77]

ACKNOWLEDGMENT

Special appreciation to Universiti Malaysia Pahang for the sponsorship of this study by Internal Grant with Vot No. RDU190311 and Geran Antarabangsa Vot W004 from Research Management Centre (RMC), Universiti Tun Hussein Onn Malaysia (UTHM). Special thanks to the reviewers and editor who review this manuscript

REFERENCES

[1] D. S. Terzi, R. Terzi, and S. Sagiroglu, "A survey on security and privacy issues in big data," in *Internet Technology and Secured Transactions (ICITST), 2015 10th International Conference for*, 2015, pp. 202–207.

[2] R. Patgiri and A. Ahmed, "Big Data: The V," in *2016 IEEE 18th International Conference on High-Performance Computing and Communications, IEEE 14th International Conference on Smart City, and IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2016, pp. 17–24.

[3] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014.

[4] A. Woodie, "Documentary Probes the Human Face of Big Data," *Datanami Inc*, 2016. .

[5] P. Rotella, "Is Data The New Oil?," *web publication*. [Online]. Available: www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/#2042b4cd7db3. [Accessed: 25-Oct-2018].

[6] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, vol. 275, pp. 314–347, 2014.

[7] W. I. Yudhistyra, E. M. Risal, I. Raungratanaamporn, and V. Ratanavaraha, "Using Big Data Analytics for Decision Making: Analyzing Customer Behavior using Association Rule Mining in a Gold, Silver, and Precious Metal Trading Company in Indonesia," *Int. J. Data Sci.*, vol. 1, no. 2, pp. 57–71, 2020.

[8] A. Belle, R. Thiagarajan, S. M. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," *Biomed Res. Int.*, vol. 2015, 2015.

[9] M. Adibuzzaman, P. DeLaurentis, J. Hill, and B. D. Benneyworth, "Big data in healthcare--the promises, challenges and opportunities from a research perspective: A case study with a model database," in *AMIA Annual Symposium Proceedings*, 2017, vol. 2017, p. 384.

[10] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: preserving security and privacy," *J. Big Data*, vol. 5, no. 1, p. 1, 2018.

[11] L. E. I. Xu, C. Jiang, and J. Wang, "Information Security in Big Data : Privacy and Data Mining," pp. 1149–1176, 2014.

[12] S. Yu, "Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016, doi: 10.1109/ACCESS.2016.2577036.

[13] M. Cox and D. Ellsworth, "Managing big data for scientific visualization," in *ACM Siggraph*, 1997, vol. 97, pp. 21–38.

[14] J. Manyika *et al.*, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[15] G. Firican, "The 10 Vs of Big Data | Transforming Data with Intelligence." .

[16] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014, doi: 10.1109/ACCESS.2014.2332453.

[17] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review," *J. Big Data*, vol. 3, no. 1, p. 25, 2016.

[18] J. Gantz and D. Reinsel, "Extracting value from chaos," *IDC iview*, vol. 1142, no. 2011, pp. 1–12, 2011.

[19] 2014 IBM Big Data & Analytics Hub, 2014IBM Big Data & Analytics Hub, "Infographic: The Four V's of Big Data | IBM Big Data & Analytics Hub," *electronic file*, 2014. .

[20] J. M. Cavanillas, E. Curry, and W. Wahlster, *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe*. Springer, 2016.

[21] M. Moorthy, R. Baby, and S. Senthamaraiselvi, "An Analysis for Big Data and its Technologies.," *Int. J. Comput. Sci. Eng. Technol.*, vol. 4, no. 12, 2014.

[22] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: the real-world use of big data: How innovative enterprises extract value from uncertain data, Executive Report," *IBM Inst. Bus. Value Said Bus. Sch. Univ. Oxford*, 2012.

[23] P. Russom and others, "Big data analytics," *TDWI best Pract. report, fourth Quart.*, vol. 19, no. 4, pp. 1–34, 2011.

[24] J. S. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, *Big data for dummies*. John Wiley & Sons, 2013.

[25] E. Ahmed *et al.*, "The role of big data analytics in Internet of Things," *Comput. Networks*, vol. 129, pp. 459–471, 2017.

[26] J. Anuradha and others, "A brief introduction on Big Data 5Vs characteristics and Hadoop technology," *Procedia Comput. Sci.*, vol. 48, pp. 319–324, 2015.

[27] S. Frost, "Drowning in big data? reducing information technology complexities and costs for healthcare organizations." 2015.

[28] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big Data technologies: A survey," *J. King Saud Univ. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, 2018.

[29] "Big Data Platforms | Max Kanaskar's Blog." [Online]. Available: https://www.mendeley.com/catalogue/6918b847-cc6c-31fa-9702-318518c065c6/?utm_source=desktop&utm_medium=1.19.4&utm_campaign=open_catalog&userDocumentId=%7Bdb6347fb-7643-4959-9d8b-b62b999cfc2a%7D.

[30] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, 3rd ed. Elsevier, 2011.

[31] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[32] C. Clifton, "Encyclop{æ}dia britannica: definition of data mining," *Retrieved on*, vol. 9, no. 12, p. 2010, 2010.

[33] N. Thushika and S. Premaratne, "A Data Mining Approach for Parameter Optimization in Weather Prediction," *Int. J. Data Sci.*, vol. 1, no. 1, pp. 1–13, 2020.

[34] M. Siddique, M. A. Mirza, M. Ahmad, J. Chaudhry, and R. Islam, "A Survey of Big Data Security Solutions in Healthcare," in *International Conference on Security and Privacy in Communication Systems*, 2018, pp. 391–406.

[35] R. Mendes and J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2017, doi: 10.1109/ACCESS.2017.2706947.

[36] N. Bairagi, "Available Online at www.ijarcs.info A Survey on Privacy Preserving Data mining," vol. 8, no. 5, pp. 2015–2018, 2017.

[37] K. P. Rao and A. Chaudhary, "Survey on Privacy Preserving Data Mining," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 3342–3343, 2014.

[38] D. Nashik, "Novel Approaches for Privacy Preserving Data Mining in k- Anonymity Model," *J. Inf. Sci. Eng.*, vol. 78, pp. 63–78, 2016.

[39] O. Maimon and A. Browarnik, "NHECD-Nano health and environmental commented database," in *Data mining and knowledge discovery handbook*, Springer, Boston, MA, 2009, pp. 1221–1241.

[40] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2001, pp. 247–255.

[41] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, 2000, vol. 29, no. 2, pp. 439–450.

[42] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, 2005, pp. 193–204.

[43] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y.

Saygin, and Y. Theodoridis, “State-of-the-art in privacy preserving data mining,” *ACM SIGMOD Rec.*, vol. 33, no. 1, pp. 50–57, Mar. 2004, doi: 10.1145/974121.974131.

[44] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, “A comprehensive review on privacy preserving data mining,” *Springerplus*, vol. 4, no. 1, pp. 1–36, 2015, doi: 10.1186/s40064-015-1481-x.

[45] N. Zhang, “Privacy-Preserving Data Mining Systems,” *IEEE Comput. Soc.*, pp. 52–58, 2007.

[46] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, 2007, pp. 106–115.

[47] A. Sharma and N. Badal, “Literature Survey of Privacy Preserving Data Publishing (PPDP) Techniques,” *Int. J. Eng. Comput. Sci.*, vol. 6, no. 5, pp. 1–12, 2017, doi: 10.18535/ijecs/v6i4.12.

[48] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “ℓ-Diversity: privacy beyond k.” Anonymity, ICDE, 2006.

[49] T. Li, N. Li, J. Zhang, and I. Molloy, “Slicing: A New Approach for Privacy Preserving Data Publishing,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 561–574, Mar. 2012, doi: 10.1109/TKDE.2010.236.

[50] R. J. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, 2005, pp. 217–228.

[51] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” in *Data Engineering, 2006. ICDE’06. Proceedings of the 22nd International Conference on*, 2006, p. 24.

[52] S. E. Fienberg and J. McIntyre, “Data swapping: Variations on a theme by dalenius and reiss,” in *International Workshop on Privacy in Statistical Databases*, 2004, pp. 14–29.

[53] R. Brand, “Microdata Protection through Noise Addition,” in *Inference control in statistical databases*, Springer, 2002, pp. 97–116.

[54] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, “Toward privacy in public databases,” in *Theory of Cryptography Conference*, 2005, pp. 363–385.

[55] C. K. Liew, U. J. Choi, and C. J. Liew, “A data distortion by probability distribution,” *ACM Trans. Database Syst.*, vol. 10, no. 3, pp. 395–411, 1985.

[56] D. B. Rubin, “Statistical disclosure limitation,” *J. Off. Stat.*, vol. 9, no. 2, pp. 461–468, 1993.

[57] J. Domingo-Ferrer, *Inference Control in Statistical Databases*, vol. 2316. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002.

[58] R. C.-W. Wong and A. W.-C. Fu, “Privacy-preserving data publishing: An overview,” *Synth. Lect. Data Manag.*, vol. 2, no. 1, pp. 1–138, 2010, doi: https://doi.org/10.2200/S00237ED1V01Y201003DTM002.

[59] L. Yang, J. Wu, L. Peng, and F. Liu, “Privacy-Preserving Data Mining Algorithm Based on Modified Particle Swarm Optimization,” in *International Conference on Intelligent Computing*, 2014, pp. 529–541.

[60] K. Chen and L. Liu, “A random rotation perturbation approach to privacy preserving data classification,” 2005.

[61] X. Li, Z. Yan, and P. Zhang, “A review on privacy-preserving data mining,” in *2014 IEEE International Conference on Computer and Information Technology*, 2014, pp. 769–774.

[62] A. Shah and R. Gulati, “Privacy Preserving Data Mining: Techniques, Classification and Implications-A Survey,” *Int. J. Comput. Appl.*, vol. 137, no. 12, 2016.

[63] A. Meyerson and R. Williams, “On the complexity of optimal k-anonymity,” in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2004, pp. 223–228.

[64] S. Mohana and S. A. S. A. Mary, “Heuristics for privacy preserving data mining: An evaluation,” in *Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), 2017 International Conference on*, 2017, pp. 1–9.

[65] Y. Ding and K. Klein, “Model-driven application-level encryption for the privacy of e-health data,” in *Availability, Reliability, and Security, 2010. ARES’10 International Conference on*, 2010, pp. 341–346.

[66] M. Binjubeir, A. A. Ahmed, M. A. Bin Ismail, A. S. Sadiq, and M. Khurram Khan, “Comprehensive Survey on Big Data Privacy Protection,” *IEEE Access*, vol. 8, pp. 20067–20079, 2020, doi: 10.1109/ACCESS.2019.2962368.

[67] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,? in Proceedings of the 6th Conference on Operating Systems Design & Implementation,” *Berkeley, CA, USA USENIX Assoc.*, p. 10, 2004.

[68] S. Shahrivari, “Beyond batch processing: towards real-time and streaming big data,” *Computers*, vol. 3, no. 4, pp. 117–129, 2014.

[69] J. Dean and S. Ghemawat, “MapReduce: a flexible data processing tool,” *Commun. ACM*, vol. 53, no. 1, pp. 72–77, 2010.

[70] D. Garc\'\ia-Gil, S. Ram\'\irez-Gallego, S. Garc\'\ia, and F. Herrera, “A comparison on scalability for batch big data processing on Apache Spark and Apache Flink,” *Big Data Anal.*, vol. 2, no. 1, p. 1, 2017.

[71] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[72] S. M. Banaei, H. K. Moghaddam, and others, “Hadoop and its role in modern image processing,” *Open J. Mar. Sci.*, vol. 4, no. 4, pp. 239–245, 2014.

[73] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, “Spark: Cluster computing with working sets.,” *HotCloud*, vol. 10, no. 10–10, p. 95, 2010.

[74] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, “Apache flink: Stream and batch processing in a single engine,” *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.*, vol. 36, no. 4, 2015.

[75] O.-C. Marcu, A. Costan, G. Antoniu, and M. S. Pérez-Hernández, “Spark versus flink: Understanding performance in big data analytics frameworks,” in *2016 IEEE International Conference on Cluster Computing (CLUSTER)*, 2016, pp. 433–442.

[76] T. White, *Hadoop: The definitive guide*. “ O’Reilly Media, Inc.,” 2012.

[77] Wikipedia, “Hadoop, Apache.” [Online]. Available: https://en.wikipedia.org/wiki/Apache_Hadoop..