International Journal of Engineering & Technology, 7 (4.35) (2018) 866-869

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Digital Forensics Challenges in Big Data Environment: a Review

Gopinath Muruti[1]*, Siti Hawa Mokhtar[1], Fiza Abdul Rahim[1,2], Zul-Azri Ibrahim[1,2], Abbas M. Al-Ghaili[2]

[1]College of Computer Science and Information Technology, Universiti Tenaga Nasional,Malaysia
[2]Institute of Informatics and Computing in Energy (IICE), Universiti Tenaga Nasional, Malaysia
**Corresponding author E-mail: gopinathmuruti@gmail.com*

Abstract

Forensics is a vital point for law enforcement, civil litigators, and different experts who manage complex advanced examinations. Digital forensics has assumed a noteworthy part in a portion of the biggest criminal and civil investigations. However, the ascent in the prevalence of big data as a better approach for unraveling the challenges exhibited by huge, complex data due to the progression of innovations such as the Internet, Internet of Things (IoT), and Cloud Computing. These challenges have contributed to data deluge and forensics tool limitations in the digital forensics investigation. In this paper, a number of challenges faced by the digital forensics investigator in a big data environment are discussed. The identified challenges could significantly contribute to a more efficient digital forensics process in the big data environment.

Keywords: *Big Data; Big Data Forensics; Digital Forensics; Forensics Investigation*

1. Introduction

As we enter the period of technological advancements, a monstrous amount of information via our daily routines and tasks is generated. For instance, our world is undergoing complete datafication, everything from human activities or otherwise in our daily life are being digitalized. We are increasingly leaving digital records from our conversations, emails in corporate systems, social media up-dates and also phone conversations every day. Furthermore, the advancement and progression of technological innovations such as the Internet, Cloud Computing, and Internet of Thing (IoT) have also helped to fuel the explosion of huge complex digital data [1].

Big data was introduced as a new method of overcoming the problems presented by this huge complex digital data. The increase in usage of big data solutions, such as Hadoop, have gained popularity across a wide number of organizations which forced digital forensics investigators to adapt to these solutions. Remarkably, a large number of organizations have implemented big data solutions in their operations over the past decade. These systems house critical information that can provide vital information to digital forensics investigations.

The arena of digital forensics has gotten a growing extent of attention over the past years as traces and digital evidence bring into being on variety of devices has become more treasured during an investigation. Notwithstanding this reality, new hurdles confronting digital forensics are rising, requiring a need to reevaluate some of the conducts in digital forensics that have been conducted as of now.

The increase in computing ubiquity, low-cost digital storage, and the progression in the number of the IoT have contributed to this massive rise in the volume of digital data. As indicated by an article by Gartner in 2017, there will be 8.4 billion associated IoT gadgets in use by 2017 and is relied upon to achieve 20.4 billion by 2020, which will make new difficulties for all parts of data administration [2]. Additionally, the volume of data produced by multiple service providers are getting larger. As an example, Sprout Social stated that there are 147,000 photos uploaded every 60 seconds to the Facebook platform alone [2]. Big data grants a real prospect in recognizing significant bits of knowledge from data. However, the huge increment in the volume of accessible digital data also played a key part in the rise in the volume of digital evidence. As per to a yearly report by the Federal Bureau of Investigation (FBI), the usual case size of the digital forensics case is increasing at thirty five percent annually in the United States [3].

2. Definition of Key Terms

In this section, a brief definition of digital forensics, challenges in digital forensics phases and big data is given.

2.1. Digital Forensics

In the course recent forty years, the arena of digital forensics has fully-fledged from being an inconsequential portion of criminal investigations to a vital part of many inquiries encompassing of digital data. Digital forensics is a subdivision of forensics science covering the retrieval and examination of material found in digital devices, often in relative to computer crime. Digital forensics utilizes systematically derived and proven techniques toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence with the determination of aiding or advancing the rebuilding of events observed to be to be criminal in nature, or aiding to antedate illegal activities shown to be disrupting to prearranged operations [4].

Since 2001, several scholars have suggested multiple frameworks and process models for digital forensics inquires. Based on an

overall summary of these models, digital forensics encompasses six main phases, which are preparation, preservation, collection, examination, analysis, and presentation [5].

2.2. Challenges in Digital Forensics Phases

This section discusses the challenges by concentrating on the stages of the digital forensics.

i. Preparation: This stage relates to all of the works that need to be done prior to the actual investigation and official collection of data [6]. Among the tasks to be performed are getting the required consent from appropriate authority preparing and setting up of the gears to be used in the investigation [7]. However the key challenge confronted in the preparation stage orbits around the volume and variety of data on which evidence might be found during investigation process[7]. Guaranteeing that the essential standards guidelines and techniques to keep an eye on during an inquiry are in order may be equitably simple preparing the examiner but equipping the right gears for all circumstances is still a challenge and ensuring that the examiners gears are ready to manage every operating system, file format, and protocol is nearly an impossible mission.

ii. Collection: In general, collection stage is where all relevant data are captured, stored and be made available for the next stage [8]. This stage faces the major hurdle as they are pretentious by the rising volume, variety and variability of data. In the facet of big data, gathering all potential evidence in an inquiry is frequently impossible.

iii. Preservation: Safeguarding that the reliability and validity of evidence is conserved during an inquiry also presents the hurdle of knowing how to handle different devices [8]. In spite of the fact that the methods and standard practices to guarantee that information is saved may not differ, having the proper instrument for a media or gadget can still be a hurdle. Likewise, with the expanding volume of information, the period engaged with preservation turns out to be extensively greater prompting higher response times amid an examination.

iv. Examination: The task of examining data containing huge volumes and a varied variety of data during an investigation is a perplexing task [9]. Though computational power is growing conferring to the Moore's law, volume of data is snowballing at a substantially speedier rate prompting to buildups in analysis and scrutiny of multiple inquiries [10]. Digital forensics tools must have the capability to digest current, emerging and old devices and data formats, and this is a foremost hurdle in an investigation.

v. Analysis: The requirement for examination of volumes of information in a brief timeframe is becoming more vital [11] and new algorithms or techniques are necessary to address this issue. The wide assortment of data that will be stumble upon in an inquiry also stances a hurdle during the scrutiny and investigation stages as digital forensics gears are insufficient in the amount of file extensions and gadgets that they can handle.

vi. Presentation: The point of this stage is to circulate the findings and deductions from an inquiry in a method that is comprehensible to an onlookers or the court of law. The major hurdle in this can be detecting the most suitable way to define the processes and techniques and processes utilized in the inspection and scrutiny of such huge volumes of data. Rationalization of efficiency and precision of such processes and techniques may also be essential in order for the evidence to be admissible.

2.3. Big Data

Generally, big data can be defined with three features or three Vs: Volume, Variety, and Velocity [12]. Volume denotes the amount of data, Variety denotes the number of types of data, and Velocity denotes to the speed of data processing. Big data generally includes numerous datasets which are complex to digest using standard software tools, predominantly within an acceptable period of time. Datasets are unceasingly increasing in size from a few terabytes to petabytes [13].

Big data has become a favorable business application and is rapidly gaining popularity as a segment of the IT industry. The potential of big data is expected to impact all sectors, ranging from the energy sector to the retail sector and has generated significant interest in the field of data analytics and data mining [14]. Data analytics and mining of big data offer acumens to businesses, it allows to perceive inaccuracies and deception which significantly mitigates against losses. Additionally, it also permits businesses to grow more operative strategies towards other industry players in a reduced amount of time, offering deep insight into consumer trends and sales while improving their profits and customer service [15]. In digital forensics, data analytics and mining of big data may help investigators to recover potential evidence to assist an investigation.

Although big data can be considered as a valuable commodity in data analytics, it does present serious challenges to digital forensics investigators. Big data poses a logistical issue to companies who wish to use big data in their operations. They are required to modify their whole approach as the data flow into their operations becomes constant rather than periodic. Next, analysis of big data requires the capability to conduct sophisticated analyses but however many presently used data tools are not able to handle real-time analysis.

3. Review Method

In the review process, Systematic Literature Review (SLR) was conducted and reported based on the guideline by [16] and [17] for research in an information system. SLR provides a complete, exhaustive summary of current evidence prove significant to a research question.

The guideline [16] pinpoints four main stages of review which are; preparation, selection, extraction, and execution. The guideline was chosen as it provides a clear guidance on how to conduct and report a systematic review.

In the preparation stage, the aim and the research question of this review are defined. In the subsequent stage, literature search is performed to select articles for review purpose. A quality assessment for each article is then performed in stage three [18]. In the final stage, the findings are discussed and the review is reported.

3.1. Research Questions

To line up with the aim of this study, the research question for this SLR is:

"What are the digital forensics challenges in big data environment?"

3.2. Primary Search

The main search procedure included 3 online databases which are IEEE-Xplore Digital Library, Springer Link, and Science Direct. The selection of databases depended on the accessibility of full-text articles subscribed by the Universiti Tenaga Nasional's library. The consideration criteria involved articles circulated in English language, availability in full text, and published between the year 2012 and the year of 2017, which deal with the digital forensics challenges in the big data environment.

3.3. Search Strategy

The initial search strings are Digital Forensics, Forensics, Big Data, and Challenges. The search string is then made by using Boolean "and" and Boolean "or" operators to permit synonyms and word class variants of each keyword. The resulting search string is ("Big Data") AND ("Big Data Challenges" OR "Big Data Forensics" OR "Big Data Forensics Challenges")

3.4. Study Selection

The study selection was ordered in three stages:

i. Primary Search: The quest for publications from three online databases. This stage was piloted by using the search string.
ii. Exploration of title, abstract and keywords of identified articles and selection based on suitability criteria.
iii. The selection only considered full-text articles subscribed by the Universiti Tenaga Nasional's library.

3.5. Data Collection

In order to assist the data collection procedure and appraise the reliability of the review, a quality checklist was utilized in order to collect evidence relevant to the research question. In the process of designing the quality checklist, a portion of the questions itemized in the previous literature were reutilized [19], [20], [21], [22], [17], and [23]. The quality checklist contained five general questions as shown in Table 1 to measure the quality of the selected studies. The checklist also used to identify the suitability aspects of studies for further review.

Table 1: Quality Checklist

No	Answer	Yes / No
SQ1	Are the aims and objectives of the research clearly stated?	Yes/No
SQ2	Is the research design clearly stated and suitable for the aims and objectives of the research?	Yes/No
SQ3	Does the researcher(s) provide(s) a clear account of the process by which their findings were produced?	Yes/No
SQ4	Does the researcher(s) display(s) enough data to support their interpretations and conclusions?	Yes/No
SQ5	Is the method of analysis suitable and sufficiently explicated?	Yes/No

4. Findings and Discussion

The initial search stage delivered a total of 3995 studies by utilizing the search string. The titles, abstracts, and keywords of the studies were filtered and only 356 articles were selected. Finally, after an in-depth consideration, a total of twenty articles were counted in for the review as tabulated in Appendix. Discussions of each challenge are explained in detail in the following subsections.

4.1. Data Deluge

Digital forensics investigators have been encountering an exponential increase in the size of forensics data collection as highlighted in S1, S14, and S17 in relation to an investigation [24]. Due to the explosion of big data which has increased the turnaround time for digital forensics investigations, it might affect the efficiency of the investigations even to the major digital forensics laboratories.

S2, S5, and S15 have also highlighted the exponential growth of data is driven by the growth in storage technologies and the drop in storage price. Additionally, the ever increasing size of device storage which has rapidly increased over the years has also resulted in data deluge in big data affecting the digital forensics analysis of an investigation. Similarly, the same issues are found in S10, S18, and S20.

Based on a review of S4, S8, and S19, identification of residual evidence in increasingly massive datasets which can hide small amounts of information is like pinpointing a needle in a haystack for digital forensics investigators [25]. The small amounts of information in the massive datasets can be relevant to an investigation[26]. Meanwhile, S3, S5, S7, and S12 have highlighted the growth and explosion of IoT devices and their integration with the cloud environment. In which it will contribute to creating a huge burst of data in the environment posing a challenge to the digital forensics investigators. The investigators might unable to seize the whole device in the data center unlike the traditional method of seizing mobile devices or desktop computers [27]. Furthermore, cloud data storage services are universally accessible to devices and users, and criminal users might utilize the big data environment that exist within the service to hide potential evidence among the huge volume of data [28].

4.2. Limitation of Current Forensics Tools

S1, S13, and S20 pinpointed that current forensics tools are not efficient in analyzing and storing unstructured dimension of data stored in huge datasets. Current forensics tools are designed only to utilize relational databases which support structured data instead of unstructured data. Furthermore, from the same articles, it was found current forensics tools offer less automation and often experience input-output bottlenecks and demand more processing power in analyzing huge datasets comprising of unstructured data.

Identically in S2 and S5, digital forensics investigators are faced with a barrier due to the limitations of file formats and devices that are supported by current forensics tools. Digital forensics investigations are hindered by the cost involved in developing a forensics tool to support a variety of operating systems, file formats, and media types. Additionally, most of the existing forensics tools are designed for personal owned computing system and are not built to appropriately address today's big data paradigm as mentioned in S12 and S18.

Due to the nature of cloud-enabled big data storage solutions also poses a challenge to the digital forensics investigator. This is due to the lack and limitation of current forensics tools in analyzing and handling the huge volume of data stored in this cloud-enabled big data storage as mentioned in S3, S8, and S11. And for any digital evidence to be utilized in court, examiners must follow a proper set of procedures when gathering and dissecting information from computer systems [29]. Hence the interface between the investigator and the data is facilitated by the tools but due to the differences in terms of the technologies, investigators will have to engage different methods and tools depending on the category of data involved. The first challenge for forensic software tools is the linking of the structures of the data, protocols, and designs in which the data resides. The second challenge is then to be able to identify precisely patterns that may occur within the big data.

4.3. Diversity of Devices

According to S1, S2, S12, and S15, mobile devices, computing devices and IoT equipment's uses wide range of operating systems, communications specifications and file arrangements, which in turn surges the complexity of data that are being generated diversity big data sources. Furthermore, scrutinizing many gadgets also adds up to the correlation and uniformity problem.

The ability to interconnect and generating correlations various devices using multiple file formats, operating systems and media types generate volumes of data of which poses a challenge to the digital forensics investigator in a big data environment in terms of the complexity of data exploitation as discussed in S17, S18, and S19.

Additionally, with the rising amount of contrasting devices, there is a need for digital forensic analyst to collect particular targeted data, not necessarily a device specific data, as the applicable data may be stored on an IoT device, mobile gadgets, personal computer, or have been transmitted to the cloud storage [30]. There is a developing need to gather and analyze data from an extensive extent of gadgets, with an emphasis on the data that empowers the individuals who are associated in a digital forensic investigation to settle on a choice with as much relevant information that is profit capable, in a timely manner [30].

4.4. Dimension of Data

Big data can originate from various sources such as a web page, network logs, social media posts, emails, documents, and can include various data from IoT sensors which then can be categorized into three types of data dimension; structured, semi-structured and unstructured dimensions. Unstructured data tends to not have a generalized format which makes a forensics analysis of such data a troublesome errand as specified in S5.

S1, S6, and S9 have highlighted the variety of data dimension stored within a big data environment pose a challenge to traditional relational database management system (RDBMS) used in current forensics tools. Hence making current forensics tools not efficient in analyzing or storing complex data since RDBMS were not intended to support unstructured data and semi-structured data [31].

Additionally, S17 have stated that crucial evidence is most likely missing when statistical or random data reduction methods are applied to selecting data containing evidence from large volumes of structured and unstructured data. The data reduction method varies in structured and unstructured data as usually perceived in big datasets, the data can range from databases, compressed files, image files, text files, and video files.

In handling a variety of dimensions in big data processing in digital forensics, models such as MapReduce [32] can be applied. Alternatively, NoSQL based database systems can be integrated into the next generation of digital forensics tools which has the capability to manage and analyze big data.

4.5. Presentation of Big Data Forensics Evidence in the Eyes of Law

The purpose of the presentation stage in a forensics investigation is to distribute the findings and conclusions from an investigation in a manner that is understandable to the members of the courtroom and the audience. S2, S5, and S13 highlighted the emergence of big data and a wide variety of data that often needed to be presented during various stages of the investigation.

The large volume of data in big data makes it a difficult task for the investigators to present the whole data without missing out on potentially important information which might be relevant to the investigation. Indeed, S5 mentioned the difficulties the judge or the court audience might face in digesting the big data technical jargons and digital forensics terms. The judge or the court audience might have elementary knowledge of using personal computers but the workings involved in filtering and analyzing big data can be very mind boggling for them to understand.

From the reviews, five issues showed up as the main digital forensics challenges in a big data environment as tabulated in Table 2, data deluge in big data and limitation of current forensics tools were found to be the most frequently discussed topic in existing literature.

Table 2: List of Digital Forensics Challenges in Big Data

Digital Forensics Challenges In Big Data	Study ID
Data Deluge	S1, S2, S3, S4, S5, S7, S8, S10, S12, S14, S15, S17, S18, S19, S20
Limitation of Current Forensics Tools	S1, S2, S3, S5, S8, S11, S12, S13, S16, S18
Diversity of Devices	S1, S2, S12, S15, S17, S18, S19
Dimension of Data in Big Data	S1, S5, S6 S9, S17
Presentation of Big Data Forensics Evidence In The Eyes of Law	S2, S5, S13

5. Conclusion

In this review paper, five digital forensics challenges in big data environment have been identified. The identified challenges are; data deluge, limitation of current forensics tools, diversity of devices, dimension of data in big data, and presentation of big data forensics evidence in the eyes of law.

The growth of digital data has resulted in data deluge due to growth in storage technologies and the drop in storage price has affected the efficiency of digital forensics investigations. Pinpointing and identifying evidence among the data deluge is like finding a needle in a haystack for the digital forensics investigators. Furthermore, current forensics tools are found inadequately efficient in analyzing and storing unstructured dimension of data stored in big data environment and have compatibility issues with multiple file formats and diversity of devices that exist in a big data environment. Digital forensics investigators are also faced with multiple dimensions of data when performing an investigation in a big data environment. The final challenge highlighted in this review is on the difficulties in presenting big data forensics evidence in the eyes of the law due to the technicality involved. Given the current situation, the future developments of digital forensics tools may focus on addressing specific challenges identified in this review.

Acknowledgment

The authors would like to thank the Ministry of Higher Education Malaysia (MoHE) and Universiti Tenaga Nasional (UNITEN) for funding this study under the Fundamental Research Grant Scheme (FRGS) of Grant No. FRGS/1/2017/ICT03/UNITEN/03/1.