ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

A Survey on Digital Forensics to Address Big Data Challenges

Radha Dhekane[1]

Graduate Student, Dept. of Computer Science, California State University, Sacramento, California[1]

ABSTRACT: Digital Forensics is a procedure used to handle the difficulties of investigating and handling big data during unlawful time, safeguarding the forensic principles to be presented in the court of law as digital evidence, transmitted from or stored on PCs. One of the fundamental difficulties in digital forensics is the increasing volume of information that should be examined. The current tools and infrastructures cannot meet the expected response time when we investigate on a big dataset. Forensics specialists will face challenges while identifying necessary pieces of evidence from a big dataset, and gathering and analysing the evidence. This paper briefly talks about the challenges and necessity of digital forensics in the face of the current trends and requirements of different crime investigations. A digital forensics analysis structure that puts into consideration the existing techniques as well as the current challenges is proposed. The aim of this structure is to reassess the various stages of the digital forensics examination process and introduce into each stage the required techniques to enhance better collection, preservation, analysis and presentation in the face of big data and other challenges facing digital forensics.

KEYWORDS: Digital Forensics; digital forensics analysis structure; digital forensics examination process; big data.

I. INTRODUCTION

The field of digital forensics has received an increasing amount of attention in the previous years as digital proof found on different devices has become more and more valuable during examinations. Digital forensics handles challenges of analysing and handling enormous data. The low price of digital storage, the increasing ubiquity of computing, and the growth of the type and number of the Internet of Things (IoT) drive this massive increment in the amount of digital data. The developing challenges can be attributed to technological advances, the capacity to interconnect different devices capable of generating huge volumes of data, the need to gather and investigate data found on data (both structured and un-structured) as well as the need to conduct forensic analysis on information stored on cloud. All of these factors emphasize a relationship between the field of data science and digital forensics and point out the need to analyse "big data" in digital forensics.

The more information is accessible, the harder it is to spot fraudulent activity and malicious users behind those activities. Large number of IoT devices produces huge datasets of activity logs. Hence, using traditional log correlation and visualization techniques, we cannot distinguish malicious activities and users from a big dataset of logs.

Data comes in a wide variety of formats and types, all of which are generated through the user and device relationship. In 2010, IBM [1] mentioned that 80 percent of data generated is unstructured. This is a concern to the digital forensic community because existing forensic tools cannot efficiently analyse or store complex data. They only employ relational databases, which are not designed to support unstructured data but structured data.

The proposed framework contributes to existing framework by presenting more efficient ways of preserving, analyzing and presenting information during an investigation despite the present challenges faced by digital forensics. Section III briefly discusses about digital forensics and the current existing process models. Section IV emphasizes on the concept of big data. Section V explores the issues faced in the different stages of the digital forensics process from the perspective of big data. Section VI defines the proposed process model. Section VII discusses the structure and points out some of the pros and cons of the model. Section VIII gives the conclusion and future work.

DOI: 10.15680/IJIRCCE.2018. 0612004

II. RELATED WORK

Digital gadgets can extend from basic sensors to complex devices, which can potentially be commandeered for criminal operations. The data on these devices change according to the device, and can be unstructured, structured or semi-structured. These devices exhibit a number of technical and legal issues which include proprietary file and operating systems, communication protocols, encryption, and rapid introduction of new devices [2]. Identifying, collecting, analysing and presenting the data from different digital devices (e.g. in a smart home, Amazon Echo, and other devices connected to Amazon Echo, such as mobile devices, smart TV etc.) can be challenging and difficult, particularly when data from a variety of devices is combined. Also when data from different sources are combined, their relevance to an investigation may turn out to be more evident.

Digital devices have also been utilized to commit crime, for example using devices for Distributed Denial of Service (DDOS) attacks, controlling cloud-based CCTV units and accessing Internet connected printers [3]. It is reported that the Lizard Stressor malware accesses connected digital devices to dispatch DDOS attacks against telecommunication companies and government agencies. The malware is effective because it uses of devices which often run embedded Linux based operating systems that have no bandwidth limitations, have minimal security, and often have default passwords shared across multiple devices. Gartner have anticipated that by 2020 more than one fourth of cyber-attacks will involve disparate digital devices [4].

The growing number of devices and storage potential is also adding to a rise in digital forensic evidence, which is overpowering many digital forensic labs [5]. The concerns with respect to the increase in data volume relate to gathering and preservation of increasing volumes of data, timely analysis of the increasing volume of data, and storage costs. Data reduction is one method to address collection, preservation, analysis, and storage concerns, utilizing a technique of Data Reduction by Selective Imaging (DRbSI) [6]. When given information has been gathered and put away, there is a need to undertake timely analysis of growing volume of disparate data, including the non-structured data generated by devices. This results in a need to be able to undertake quick processing and analysis of an increasing volume of disparate device and case information.

III. DIGITAL FORENSICS

Digital forensics is defined by Palmer as “the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations” [7].

From a general review of these models, digital forensics comprises of the following processes:

- Identification:

The initial step of a digital forensics investigation is identification, where an investigator identifies the issues that are vital to prosecute litigation and recognize the evidence related with those incidents. This stage identifies potential sources of relevant information in devices and location of data.

- Preservation:

This stage involves the isolation and securing of the state of the digital and physical evidence at the crime scene. Essential actions are taken to maintain the integrity and credibility of evidence during the investigation. This is achieved by capturing visual pictures of the scene and documenting all relevant information about the evidence and how it was obtained.

- Collection:

This stage involves collection of digital information that may be relevant to the investigation. Collection may include removing the electronic devices from the crime scene and then imaging, duplicating or printing out their content.

- Analysis:

Based on the relevant information gathered, analysis is performed to determine the significance of the data and make conclusions. The outputs of this examination are data objects found in the data. Analysis aims to draw conclusions dependent on the evidence found.

- Reporting:

This stage provides an outline and explanation of the findings and conclusions drawn, during the examination and analysis stage. The purpose of this phase is to disseminate the findings and conclusions from an investigation in a way that is understandable to audience and the court. Other competent legal forensic examiners should also be able to duplicate and reproduce similar results.

IV. **BIG DATA**

Big Data is a progressing concept, with its definition going through changes with time, context and rapid technological advances. The intriguing fact is that the data considered to be vast ten years ago is almost insignificant when compared to the current volume of data. It is not just about the size of the data, it incorporates different dimensions of data, best understood by the 5 Vs, viz. volume, velocity, variety, veracity and value.

- Volume:

Volume refers to the huge amount of data that is generated. This involves scalability in terms of information storage as well as the requirement for a distributed approach in processing information. In the age of big data, organizations manage terabytes and petabytes of data. For instance, Walmart handles more than a million client transactions each hour and imports those into databases valued to contain more than 3 petabytes of data [8], AT &T has a database which is 312 terabytes in size [9].

- Velocity:

Velocity is defined as the rate at which data is generated or moved around. Velocities differ depending on the source of information, which may generate data in real time or by batch or stream processes. Velocity does not only refer to the speed of the incoming data but also the speed of data flow within a system. Velocity of big data infers that data acquisition and examination need to be conducted expeditiously so as to maximize the value of the data.

- Variety:

Big data can be categorized in three general types: structured, semi-structured, and unstructured. Data stored in a regular relational database is treated as structured data. Unstructured data do not have a fixed format or arrangement and hence, it is harder to conduct forensics analysis on such data. Semi-structured data may not have specific fixed fields but can contain tags to speed up the data analysing process. With the rise of unstructured and semi-structured data, there is a requirement for better storage and analysis of data.

- Veracity:

Veracity refers to correctness and accuracy of data. Veracity involves information quality, information governance, and metadata management, along with considerations for legal and privacy issues. With different forms of big data, quality and precision are less controllable and big volumes often cause the loss of quality or precision.

- Value:

Value refers to the ability of transforming collected information into a value. Big Data is concerned on extracting value from enormous stored data. The availability of big data can uncover previously obscure insights from data and this leads to increasing its value. It is important to ensure that the data generated is accurate and leads to measurable improvements.

V. Big Data Challenges in Digital Forensics

Big data forensics can be considered as a special branch of digital forensics where the identification, preservation, collection, analysis, and reporting processes deal with a huge-scale dataset of possible evidence to build up the facts about a crime.

This section describes few of the challenges facing digital forensics while dealing with big data. These described by focusing on the stages of the digital forensics process given above:

- Identification:

The major challenge faced in the identification stage involves the volume and variety of data, as well as diversity of devices on which evidence can be found. Although ensuring that the necessary standards, procedures to follow during an investigation are in place; training the investigator and having the right tools for each situation is still considered as a challenge.

- Preservation:

Ensuring that the authenticity and integrity of evidence is preserved all through an investigation also presents the challenge of knowing how to handle various gadgets. Although the techniques to guarantee that information is preserved may not change, having the appropriate tool for that particular gadget may still be a challenge. Also, with the increasing volume of information, the time involved in preservation becomes significantly higher prompting to larger response times during an investigation.

- Collection:

The collection stage of the forensics process face the biggest test as they are influenced a lot by the growing volume, variety and variability of data. Although there has been a significant decrease in the cost of storage devices over the years, storing huge volumes of data in an uncompressed manner still has a huge cost involved. On the other hand, if the data is compressed, the price involved in the collection is reduced however this implies that the data is not easily available for a forensic analysis.

- Analysis:

Analysing data during an investigation involves big volumes and an extensive variety of data. Many of the techniques used in traditional digital forensic investigation such as string matching, pattern search and text mining are not reasonable for the present problem space of digital forensics mainly because the techniques do not scale and the accessible computing power is underutilized [10]. The need to analyse volumes of data in a short span of time is becoming more important and new techniques/algorithms are required to address this issue.

- Reporting:

The primary challenge in reporting phase is identifying the most appropriate way to describe the techniques and processes used in the examination and analysis of such substantial volumes of information. Justification of validity and correctness of such techniques may also be required for evidence to be valid and acceptable.

VI. PROPOSED DIGITAL FORENSIC PROCESS MODEL

Considering the challenges identified above, it is important to reconsider each stage of the digital forensic process model to ensure that the issues are dealt with. Specialists have stressed the issue of increasing data volume in digital investigations. The new model also emphasizes on various techniques that can be connected at different phases of an investigation.

The aim of this model is to present a thorough view on the digital forensic process with focus on steps and techniques that may be applied in each phase of the process. Figure 1 presents the proposed framework.

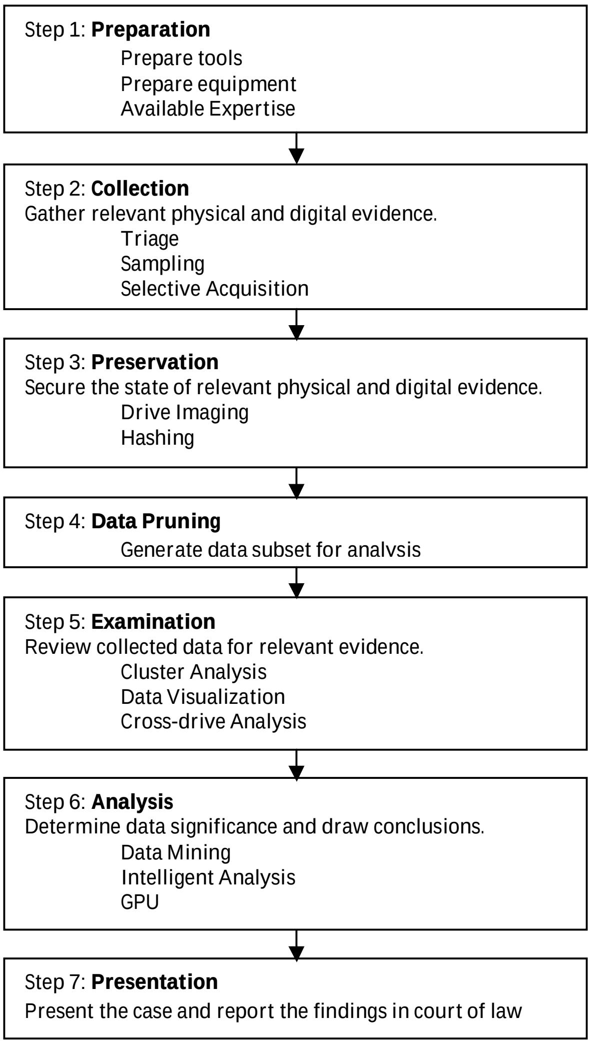

Fig.1. Proposed Digital Forensics Framework

- Step1-Preparation:

The first phase of the framework is the preparation phase and it highlights on ensuring that the tools, equipment and specialists required for the investigation are available. A methodology that has been recommended before beginning an investigation is the utilization of a matrix which lays out the possible scenario of the crime scene, digital evidence and the suspect, and specifies the required skills for the specialists on the team. Understanding the aim of the investigation and expertise will assist in preparing for additional expertise when required.

- Step 2-Collection:

The second step of the framework involves the collection of potential evidence for the investigation. Techniques such as triage, sampling selective and intelligent acquisition have been suggested to constrain the amount of data that has to be collected.

Triage is a process in which potential sources of evidence are positioned in the order of priority so that forensic tasks are ordered properly during an investigation. Storage media and devices are prioritized based on the relevant data stored in them which are retrievable within a short period of time. Another approach that can be used to handle huge volumes of data during collection is sampling of the media so that only the data in specific sections have to be collected. Using statistical techniques, it is possible to gain information about the content of the media and reduce the volume of data collected. Selective acquisition involves the collection of logically selected file types of data based on the specific investigation.

Triage, sampling and selective and intelligent acquisition can be connected in the collection of cloud data to reduce collection time, storage space and time required for investigation.

- Step 3-Preservation:

Preservation of collected data is similar to the strategies and techniques involved in existing frameworks. Drive imaging and hashing can be done in order to prevent modification or destruction of the original evidence. Imaging a drive is a forensic technique in which an analyst creates a bit-for-bit duplicate of a drive. This forensic image of the digital media helps retain evidence for the investigation. Hashing is used to generate digital fingerprints of data files, in order to prove the equivalence of the original and the duplicate data. All of the analysis is conducted on "forensic clone" so as to preserve original data.

- Step 4- Data Pruning:

Many of the techniques that have been proposed for reducing data volume in an investigation are often applied as initial phase of data collection. In investigations that need to obtain a full forensic image, the data pruning step can be used to reduce the quantity of data that has to be analyzed.

The pruning step ensures that the evidence is preserved and is rendered usable by filtering out data part with greatest potential for investigation, using forensic tools. The advantage of the pruning step is that an investigator can reassess the full forensic image in the event that trace evidence is missed in the pruning process.

- Step 5 - Examination:

This step of the framework involves searching collected evidence or a part of evidence to locate relevant data. The need here is for smart algorithms that work better than techniques currently used in investigation and which reduces retrieval overhead time so that examination can be completed fast, despite increase in data volumes. Some of the techniques that have been proposed to achieve this objective include cluster analysis, data visualization, outlier analysis and cross-drive analysis. Cluster analysis involves utilization of neural networks for clustering search results so that the number of hits is reduced [11]. Data visualization helps an investigator to visually interpret data and aids the process of anomaly detection [12]. Cross-drive analysis [13] includes the use of statistical techniques for the correlation of data on multiple disk images. In multiple drives, the techniques can be applied to identify which drive has the most relevant information for examination.

- Step 6- Analysis:

The analysis stage of the digital forensics process model is possibly the stage most hampered by the growing variety, variability and data volume. In order to handle these challenges, techniques such as data mining, intelligent forensics, use of graphical processing unit (GPU) or other ways of improving available processing power during an analysis have been proposed.

Data and text mining techniques can be applied to digital forensics to improve the response time in an analysis as well as reduce associated cost. Since this field is well developed, techniques applied to digital forensic investigations can be better presented in a court during a case.

Intelligent analysis involves the use of tools and techniques such as artificial intelligence, computational modeling and social network analysis in digital investigations with the aim of reducing the amount of time involved in analysis [14].

The use of graphical processing unit (GPU) to enhance the performance of digital forensic tools is another way in which the amount of time required for data analysis can be reduced [15].

- Step 7- Presentation:

The presentation step includes the communication of findings during forensic investigation. Reports include the summary, list of evidence examined, tools used for analysis, findings and conclusion. Detailed information on the techniques and algorithms employed in the above steps will be of benefit during presentation.

VII. DISCUSSION OF PROPOSED FRAMEWORK

The digital forensic framework defined in this paper goes a step further by mapping different methods that have been proposed for use during a digital forensic investigation. To a great extent, the principle steps identified in the framework are based on an amalgamation of most of the current framework. The purpose of the steps identified in the framework is to generalize it as much as possible.

Additionally, the framework provides a general mapping of the process in the big data environment. It also gives an overview of the processes and techniques that can be applied when handling huge data volumes and is intended to complement existing models.

It is important to note that the techniques deployed at any sub-step of the framework will be dependent on the unique situation in an examination. Although we have highlighted different techniques that may be used each step, the applicability of each in a particular situation is a decision that has to be made by the investigator. Depending on the investigation, it is possible that there could be other applicable techniques at each step of the framework.

VIII. CONCLUSION AND FUTURE WORK

The need to re-evaluate the procedures and techniques used during a digital forensic examination is vital because of the present challenges facing digital forensics, particularly with respect to expanding data volumes in an investigation. In this paper we have defined a framework that is planned to complement existing models by describing methods and techniques that can be applied at various stages of an investigation. The paper has described the notion of big data and its attributes. The challenges faced at different stages of a digital forensic investigation with regards to handling big data have also been discussed.

Although we have discussed the techniques mentioned in the framework, research into their application and ramifications of some of the procedures is still required even though their potential in an investigation are clearly visible.

REFERENCES

1."IBM", The enterprise answer formanaging unstructured data, [online] Available: https://www-304.ibm.com/events/idr/idrevents/detail.action?meid=6320.
2. The 2016 Internet Organized Crime Threat Assessment (IOCTA), The Hague, Netherlands, 2016.
3. Q. Do, B. Martini, K.-K. R. Choo, "A data exfiltration and remote exploitation attack on consumer 3D printers", IEEE Trans. Inf. Forensics Security, vol. 11, no. 10, pp. 2174-2186, Oct. 2016.
4. W. Ashford, Lizard Stresser IoT botnet Launches 400 Gbps DDoS Attack, Oct. 2016, [online] Available: http://www.computerweekly.com/news/450299445/LizardStresser-IoT-botnet-launches-400Gbps-DDoS-attack.
5. H. Parsonage, Computer Forensics Case Assessment and Triage—Some Ideas for Discussion, Aug. 2013, [online] Available: http://computerforensics.parsonage.co.uk/triage/triage.htm.
6. D. Quick, K.-K. R. Choo, "Big forensic data reduction: Digital forensic images and electronic evidence", Cluster Comput., vol. 19, no. 2, pp. 723-740, 2016.
7. G. Palmer, "A Road Map for Digital Forensic Research," Utica, New York, 2001.
8. "SAS", Big data meets big data analytics, [online] Available: http://www.sas.com/resources/whitepaper/wp_46345.pdf.
9. B. Davis, "How Much Data We Create Daily", 2013, [online] Available: http://goo.gl/aOImFT.
10. S. L. Garfinkel, "Digital forensics research The next 10 years", Digital Investigation, vol. 7, no. Supplement, pp. S64-S73, 2010
11. N. L. Beebe, J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results", Digital Investigation, vol. 4, pp. 49-54, 2007.
12. S. Teelink, R. F. Erbacher, "Improving the Computer Forensic Analysis Process Through Visualization", Communications of the ACM, vol. 49, no. 2, pp. 71-75, 2006.
13. S. L. Garfinkel, "Forensic feature extraction and cross-drive analysis", Digital Investigation, vol. 3, pp. 71-81, 2006.
14. B. W. P. Hoelz, C. G. Ralha, R. Geeverghese, "Artificial Intelligence Applied to Computer Forensics", Proceedings of the 2009 ACM Symposium on Applied Computing, 2009.
15. L. Marziale, G. G. Richard, V. Roussev, "Massive threading: Using GPU s to increase the performance of digital forensics tools" in Digital Investigation, vol. 4, pp. 73-81, 2007.