A Framework for Digital Forensic Investigation of Big Data

Jie Song
Department of Computer Science and Technology
Sichuan Police College
Luzhou, China
e-mail: wdsaqwe@gmail.com

Jin Li*
Department of Computer Science and Technology
Sichuan Police College
Luzhou, China
e-mail: 329743416@qq.com

***Abstract*—With the development of internet technology, the era of big data with data explosion has come. As a result, the types and the amount of data that need to be processed in cybercrime investigations are increasing. It is a great challenge to find the clues or evidences of crime from the massive data. The emergence of big data technology has brought new directions for digital forensics investigations. It is of practical significance to integrate big data technology and digital forensics to crack down on cybercrime. This paper proposes a big data forensics framework which describes the process and related technologies. In addition, the challenges to realize this framework have been taken into consideration.**

Keywords-big data; digital forensic; chain of custody; standardization

I. INTRODUCTION

Due to the development of technology and the changes of people's life-style, the Internet has become an integral part of people's lives. Social media, mobile payment, public transportation, medical, education, entertainment and so on, all aspects of people's lives are closely related to the Internet. This phenomenon will become more and more obvious under the trend of rapid development of IoT technology and popularization of 5G. According to a Cisco report [1], the number of Internet users in the world will reach 51% of the population in 2018 and 66% in 2023. According to the 44th Statistical Report on the Development of the Internet in China[2] issued by CNNIC, by June 2019, the number of Chinese Internet users has reached 854 million, accounting for 61.4% of the population. The amount of data generated by such a large number of Internet users is bound to be astonishing. According to a report issued by IDC [3], the data generated globally in 2018 was 33ZB (about 36.3 trillion GB), with an average of about 10 billion GB per day. People have really entered the era of big data.

On the other hand, a variety of Internet-based crimes are emerging. The proportion of cybercrimes has increased year by year, posing a great threat to economic and social development and the safety of people's lives and property. A report by RiskIQ[4] pointed out that in 2018, the worldwide losses caused by cybercrime reached $1.5 trillion, with an average loss of $2.9 million per minute, and the loss will increase year by year. Therefore, preventing and combating cybercrime is a common and important topic all over the world. The coming of big data era brings us challenges and opportunities. Using big data technology to conduct digital forensics investigation, prevent and crack down on crimes, improve the level of social governance is a subject worthy of study.

This paper analyzes the process of big data digital forensics, proposes a framework for digital forensic investigation using big data technology, and introduces the workflow of the proposed framework and the challenges of it.

II. BIG DATA AND DIGITAL FORENSICS

A. Big Data

The main characteristics of big data can be defined as 3Vs, which are Volume, Variety and Velocity.

- *Volume:* The size of data generated by human, corporations or sensors is very huge. And the data size reaches petabytes and zettabytes [9].
- *Variety:* Various media will generate a variety forms of data, some are text, some are images, some are multimedia and some are logs or mails. Therefore, these data may be structured or semi-structed or unstructured.
- *Velocity:* On the one hand, data generation is very fast, the data is continuously generated from a variety of data sources. On the other hand, big data analysis needs a high processing speed to ensure timely and effective.

Besides, other characteristics of big data such as Veracity, Value [10], Variability, Viscosity, Virality, Ambiguity [11], Complexity [12] and so on have been proposed. Due to the characteristics of big data, we need a technology with the following properties to process it: first, it can store a huge size data; second, it can process these data quickly; third, it can analyze these data and get useful content from it.

Using traditional processing methods, such as processing with a powerful mainframe computer, will result in very high costs. In order to solve this problem, Google proposed their distributed file system Google File System and distributed computing model MapReduce. So that a big task can be divided into several smaller tasks which are deployed on some cheaper hardware. This greatly increases scalability and reduces the costs. Inspired by this, Doug Cutting developed a distributed processing software framework,

* Corresponding author.
This work was supported by the Cybercrime Forensics Innovation Team of Sichuan Provincial Department of Education (No. 18td0039).

978-1-7281-9741-8/20/$31.00 ©2020 IEEE

Hadoop, which contains two key components: the distributed file system HDFS and the parallel data processing engine Hadoop MapReduce. After years of development, Hadoop has grown and has a complete ecosystem with numerous open-source tools for the highly scalable distributed computing field. Some commonly used components are: Yarn, HBase, Pig, Hive, Sqoop, Flume, Mahout, Spark, Impala, Oozie, ZooKeeper, etc. What's more, there are many other big data technologies and platforms in the world. This paper will not repeat them, just take Hadoop as an example.

B. Digital Forensics

In order to crack down cybercrime, investigators need to find traces of crime in various forms of digital data, such as log files, e-mails, spreadsheets, web histories, deleted files, mobile phones, monitoring and so on. The digital data was eventually provided to the judge in court as valid evidence to bring the criminals to justice. The subject of digital forensics emerged in this context. After years of development, digital forensics has a set of steps in compliance with regulations, which can be briefly summarized as follows:

- *Collection and Acquisition.* It refers to the collection, acquisition and archiving of potential digital evidence data such as logs, data files, memory dump, caches, network traffic, social media, web pages, IoT, in compliance with relevant laws, regulations and technical specifications. The technologies involved in this process mainly include read-only, clone, verification and timestamp technologies, which ensure the authenticity and integrity of digital evidence data. In addition to passively collecting and archiving the digital data of incidents that have occurred, it can also be proactive. Active collection technology includes process monitoring, network sniffing, intrusion detection, boundary detection, honey trapping technology, etc.
- *Preservation.* The collected data need to be preserved in accordance with relevant specifications. Furthermore, various data generated during digital forensics also need to be preserved and protected. Only by ensuring the security of the preserved data can the authenticity and integrity of digital evidence be guaranteed.
- *Examination and Analysis.* Examine the storage medium of raw data or the archived data through data recovery, cracking, searching, simulation, correlation, statistics, comparison or other methods to further discover clues or evidence related to the case. The objects of examine can be removable disk, CD, hard disk, image, file, or memory dump, network packets etc. Common examination and analysis technologies include: data recovery, keyword search, registry analysis, database analysis, log analysis, password cracking, steganalysis, timeline analysis, correlation analysis, memory analysis, malicious code analysis, network traffic analysis, data filtering, and tracking back [13]. Moreover, techniques such as pattern matching, machine learning, data mining, and artificial intelligence can also be used for the examination and analysis of digital data.
- *Presentation.* The results of examination and analysis are visually presented in an accessible form. And the data and reports are presented to the courts as evidence. The technologies involved in this step include data visualization, natural language processing, and human-machine interface, etc.

C. Big Data Digital Forensics

People generate more and more digital data every day, and people have more and more Internet devices. According to a report from Cisco [14], the per capita Internet devices in China will be about 2.8 in 2017 and 4.4 in 2022. In actual cases, one person often owns several or more than a dozen network devices, and even a dozen or even hundreds of devices need to be analyzed in one case. The increasing volume of data, the increasing number of devices, and the increasing complexity of analysis bring great challenges to the investigators. The traditional methods of digital forensics analysis can no longer meet the current needs, and a new direction is urgently needed to break this situation. Big data technology is the breaker.

According to the previous analysis, we found that the problems of digital forensics are consistent with the characteristics of big data. For example, the data needed for forensics is large in volume, diverse in structure and complex in relation. Therefore, it is easy to think of using big data technology to solve the current dilemma of digital forensics. First, the use of big data technology for digital forensic investigations can solve the tasks that traditional forensics takes a long time to complete, greatly reducing the labor cost. Second, it can mine out clues that are difficult to be found by traditional forensic methods from a large amount of data from different types of data sources. Third, it can also make active forewarning of crime and store the potential evidence data to provide strong support for preventing and combating cybercrime.

III. The Framework for Digital Forensic Investigation of Big Data

We propose a framework for big data forensics technology, as shown in Fig. 1. The framework describes various technologies that may be involved in the process of big data digital forensic. We divided the framework into three parts, namely digital forensics technology, intermediate technology and big data technology.

A. Digital Forensics Technology

The digital forensics technology includes the technologies involved in the above steps of digital forensics from data collection to data representation. What needs to be highlighted are as follows:

- *Collection and Acquisition process in big data forensic.* The data which collecting and acquisition from different data sources, needs to be standardized and converted into structured, semi-structured, and unstructured data for subsequent processing.

- *Preservation process in big data forensic*. Because of the huge volume of data, data reduction technology is also needed. By using data reduction technologies such as lossless compression and data deduplication, the cost of data preservation can be reduced. It should be noted that the validity of data must also be considered.
- *Examination process in big data forensic*. The data examination process of big data forensics is no longer the same as traditional digital forensics. Big data forensics uses artificial intelligence technology and distributed big data forensics tools to automatically examine the data files that have been standardized and stored in the data warehouse. Meanwhile, the results of examination are analyzed by artificial intelligence and reported by presentation module. If the results can't meet the expectation of investigator, the analysis will be repeated to obtain the desired results based on user's instructions.

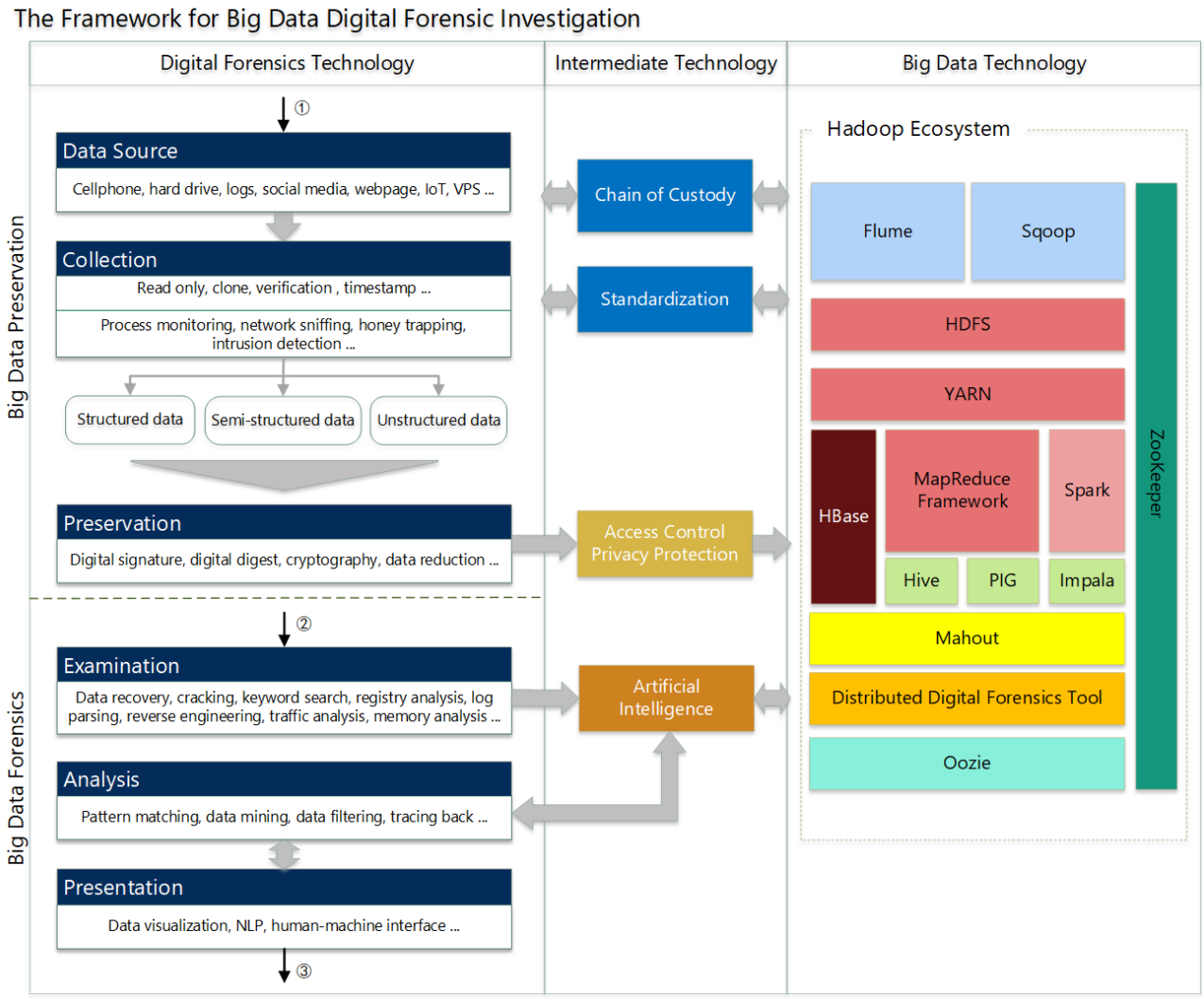

Figure 1. Big data digital forensics framwork.

B. *Intermediate Technology*

We define intermediate technologies as: A technology that bridges forensics technology with big data technology to ensure the legitimacy, feasibility, security and accuracy of the process of big data forensics.

- *Chain of Custody*. It is in charge of supervising and documenting various operations or status throughout the forensic investigation process. For example, the data collection process, the data reduction operation, the status of each distributed storage node, the process of automatic analysis, etc. The purpose of CoC is to ensure the safety and traceability of the process of big data forensics, the authenticity of data, the legitimacy of evidence, and the repeatability of results.
- *Standardization*. The purpose of standardization is to solve the problem of heterogeneous data from multi-

source. In the process of big data forensics, the tools, software, technologies and so on that can be used for the same purpose are not unique. So the results returned may also be very different. Therefore, the use of standardized data, standardized processing procedures and standardized technical solutions will ensure the reliability of the forensic investigation results.

- *Access Control and Privacy Protection.* They are designed to data security, prevent illegal access and tampering from outside, and prevent the leakage of collected and preserved data.
- *Artificial Intelligence.* On the one hand, it enables machines to master traditional digital forensics processes using various forensic tools and technologies for forensic investigations. For example, under the guidance of expert systems, use the corresponding distributed big data forensic analysis tools to conduct investigations and forensics. On the other hand, the results of examination are analyzed by the artificial intelligence system. Using pattern matching, data mining, data filtering and other technologies to analyze the examination results, and output the analysis results to the user. Meanwhile, user's feedback of the analysis results will be fed back to the artificial intelligence knowledge base for system training to improve the accuracy of artificial intelligence system.

C. Big Data Technology

Big data technology indicates the technology that processes the huge volume of data that needs to be investigated. In our framework, we take Hadoop ecosystem technology as an example. The use of other big data technologies or tool components does not affect the structure of the framework. Two points need to be explained:

- *CoC and Standardization.* The whole operation process of big data system, such as data storage, analysis and examination, needs to be supervised and audited by the chain of custody. In addition, the system structure, work flow, components, and the way of system deployment need to be standardized. In this way, the legitimacy and reliability of the big data digital forensics process can be guaranteed.
- *Distributed Big Data Forensics Tool.* Traditional digital forensics examination and analysis tools, such as file retriever, registry viewer, log parser and so on, run in a standalone computer environment. However, in big data forensics investigation, they are running on a distributed system, so it is necessary to deploy distributed forensics analysis tools to the big data system in order to quickly examine a huge volume of data in the data warehouse.

IV. Workflow of Big Data Forensics Framework

We divide the process of big data digital forensics into two steps: big data preservation and big data forensics.

A. Big Data Preservation

It mainly corresponds to the two processes of data collection and data preservation in the process of traditional digital forensics. Due to the number of devices and the amount of data involved in crimes are increasing, the collection and preservation of data need to be carried out through big data technology. At the same time, the data needs to be standardized, and the big data preservation process needs to be supervised by chain of custody.

B. Big Data Forensics

It mainly corresponds to the three processes of examination, analysis and presentation in the process of traditional forensic. The differences between big data forensics and traditional digital forensics are as follows: First, the data source of big data forensics is no longer the separate raw data, but the standardized data stored in the distributed data warehouse of big data system. Second, due to the huge volume of data, the examination and analysis process is no longer conducted by investigators manually analyzing each evidence file, but by artificial intelligence technology. This can improve work efficiency.

C. The Workflow

Now, we will demonstrate the workflow of big data forensics according to the big data forensics framework.

The arrow labeled ① in Fig. 1 indicates the beginning of the workflow. At the same time as the start of big data forensics, the chain of custody was also started to supervise all aspects of the process. Next, data are collected from different data sources. The collected data can be online data, offline data, passive forensic data or active forensic data. Then, the collected data are standardized according to certain standards and converted into structured, semi-structured, or unstructured data which can be accepted by the big data system for post-processing. Besides, data reduction can be performed to minimize the data stored in the data warehouse on the premise of ensuring the authenticity, integrity and availability of the data. Meanwhile, these data are subject to access control and privacy protection to prevent data tampering or leakage.

The arrow labeled ② in Fig. 1 indicates the beginning of the forensic analysis process. After investigators set forensic preferences, the AI system (or expert system) will automatically examine the data files stored in the data warehouse using distributed big data forensics tools. After the AI system obtains the examination results, it will further mine and analyze the examination results, and output the interested information to the presentation module. The presentation module can use data visualization and other technologies to intuitively display the analysis results and generate reports. If the results do not meet the expectations of investigator, the parameters can be adjusted through the human-machine interface until the desired result is finally obtained. What's more, the process of parameter adjustment and the feedback of user on the analysis results will be fed back to the AI system for training. In the end, exit when the flow turns to arrow ③.

V. CHALLENGES

In order to realize this big data digital forensics framework, we also need to address the following challenges.

A. Data Source

Challenges from data sources are as follows: Processing methods for heterogeneous data from various types of data sources. Privacy and security of data that may contain sensitive information, such as the IoT and social networks. Data reduction methods to reduce the cost of data preservation and processing.

B. Standardizing

Standards are needed to the big data forensics process. Processing standards for heterogeneous data and various data formats. Standardization of big data digital forensics processes, such as the standardization of technologies and processes involved in the collection, preservation, examination, analysis and presentation steps.

C. Chain of Custody

There is a big conflict in big data digital forensics, that is, the conflict between legitimacy and feasibility. To ensure the integrity of the data to ensure the legitimacy, it will often cause a significant increase in costs and processing difficulties. However, data reduction to reduce costs, it must be considered whether the data still has the legitimacy as evidence. Therefore, in addition to standardization, there should also be a set of supervision mechanisms, namely chains of custody, to supervise every link in the big data digital forensics process.

D. Distributed Forensic Analysis Tools

The traditional forensics tools and algorithms need to be improved to adapt to the distributed big data system. For example, the most commonly used "keyword search" tools in traditional forensic, if the algorithm and tools are not improved, the search speed will be very slow due to the explosion of data.

E. AI System (Expert System)

A powerful and accurate artificial intelligence system is essential for quickly and efficiently finding criminal clues or evidence in massive data and reducing the workload of investigators.

VI. CONCLUSION

Under the challenge of cybercrime in the new situation, it is of great significance to study the big data forensics investigation technology to prevent and combat cybercrime. This paper briefly introduces the process and technology of big data digital forensics, proposes a big data digital forensics framework, and analyzes some problems of the framework that need to be solved.

It is hoped that more and more people will participate in the research of big data forensics, promote the development of technology, promote the establishment of standard, and prepare for the prevention and combat of cybercrime in the era of big data.

REFERENCES

[1] Cisco, Cisco Annual Internet Report (2018–2023) White Paper [Online], https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] CNNIC, The 44th Statistical Report on Internet Development in China[Online],http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201908/t20190830_70800.htm

[3] Seagate, The Digitization of the World[Online], https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[4] RiskIQ, The Evil Internet Minute: The Cost of Cybercrime[Online], https://www.riskiq.com/videos/evil-internet-minute-cost-cybercrime/

[5] Mohammed H, Clarke N, Li F. An automated approach for digital forensic analysis of heterogeneous big data[J]. 2016.

[6] Kishore N, Saxena S, Raina P. Big data as a challenge and opportunity in digital forensic investigation[C]//2017 2nd International Conference on Telecommunication and Networks (TEL-NET). IEEE, 2017: 1-5.

[7] Bulgakova E, Bulgakov V, Trushchenkov I, et al. Big data in investigating and preventing crimes[M]//Big Data-driven World: Legislation Issues and Control Technologies. Springer, Cham, 2019: 61-69.

[8] Quick D, Choo K K R. Big forensic data reduction: digital forensic images and electronic evidence[J]. Cluster Computing, 2016, 19(2): 723-740.

[9] Zikopoulos P, Eaton C. Understanding big data: Analytics for enterprise class hadoop and streaming data[M]. McGraw-Hill Osborne Media, 2011.

[10] Lomotey R K, Deters R. Towards knowledge discovery in big data[C]//2014 IEEE 8th International Symposium on Service Oriented System Engineering. IEEE, 2014: 181-191.

[11] Krishnan K. Data warehousing in the age of big data[M]. Newnes, 2013.

[12] Katal A, Wazid M, Goudar R H. Big data: issues, challenges, tools and good practices[C]//2013 Sixth international conference on contemporary computing (IC3). IEEE, 2013: 404-409.

[13] Mai Yonghao, Zou Jinpei, Xu Rongsheng. Computer Forensics and Forensic Evaluation (Second Edition) [M]. Tsinghua University Press, 2014.

[14] Cisco, VNI Complete Forecast Highlights[Online], https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/China_Device_Growth_Traffic_Profiles.pdf

[15] Abdrabo M, Elmogy M, Eltaweel G, et al. Enhancing big data value using knowledge discovery techniques[J]. IJ Information Technology and Computer Science, 2016, 8: 1-12.

[16] Prayudi Y, Ashari A, Priyambodo T K. Digital evidence cabinets: A proposed framework for handling digital chain of custody[J]. International Journal of Computer Applications, 2014, 107(9).

[17] Nguyen, Trung, Hatua, Amartya, Sung, Andrew. Building a Learning Machine Classifier with Inadequate Data for Crime Prediction[J]. Journal of Advances in Information Technology, 2017, 141-147.

[18] Rao, Srinivas, Prasad, M., Reddy, Thammi. An Efficient Keyword Based Search of Big Data Using Map Reduce[J]. Journal of Advances in Information Technology, 2017, 159-164.