



The Arab American University of Palestine AAUP
Faculty of Graduate Studies

**Operational Frameworks in Data Acquisition and Incidence
Handling for Cyber Forensics Exploiting BigData**

By:

Layth Nasser Eddin Mohammad Abuarram

Supervisor:

Dr. Islam Amro

**This thesis was submitted in partial fulfillment of the
requirements for the Master's degree in
Cybercrimes and Digital Evidence Forensics**

February 23

© Arab American University –2011. All rights reserved.

Thesis Approval
Operational Frameworks in Data Acquisition and Incidence
Handling for Cyber Forensics Exploiting BigData

By:

Layth Nasser Eddin Mohammad Abuarram

This thesis was defended successfully on 08/03/2023 and approved by:

Committee members

Signature

1. Dr. Islam Amr /Supervisor



.....


2. Dr. Ahmad Hasasneh / Internal



.....

Examiner

3. Dr. Nael Abu Halaweh /External



.....

Examiner

Declaration

I am the undersigned who submitted the thesis entitled:

**Operational Frameworks in Data Acquisition and Incidence Handling for Cyber Forensics
Exploiting BigData**

I declare that this thesis has been composed solely by myself and has not been submitted, in whole or in part, in any previous application for a degree, except where stated by reference or acknowledgment that the work presented is entirely my own.

Students name: Layth Abo Arram

Signature:



Date: Monday, January 8, 2024

Student ID: 202012754

Dedication

I am deeply thankful to God for all the blessings and opportunities He has provided me throughout my life. Without His guidance and love, I would not have achieved all that I have.

I dedicate this master's degree to my parents, who have always believed in me and supported me through every step of my educational journey. Without their love, guidance, and encouragement, I would not be where I'm today. I'm forever grateful for all that they have done for me.

I also want to thank my supervisor, doctors, and mentors, who have challenged and inspired me to think critically and strive for excellence. Their wisdom and expertise have played a significant role in my growth and development as a student and professional.

Finally, I want to dedicate this degree to my friends and colleagues, who have been invaluable sources of support, motivation, and friendship during my time. Thank you for being there for me through the highs and lows and making this experience so much more meaningful and enjoyable.

Acknowledgment

I want to express my great thanks to **God** Almighty for his generosity and grace in granting me the health and the ability to accomplish this thesis.

Fortunately, I am among the few who have had the honor of being a student of the supervisor whom the more you learned from him, the more you wanted, the teacher who taught us to act morally and honestly above all; **Dr. Islam Amr**, thank you for inspiring, supporting, guiding, and correcting me through my journey of preparing my thesis. I will never forget every piece of advice, comment, and word you gave me.

Extreme thanks and appreciation from the **committee examiners** for their efforts and time in reviewing my thesis and providing me with the needed feedback to enhance the thesis.

My heartfelt gratitude to **my parents** for their continuous support and motivation.

Abstract in English

BigData is a key component of the fourth industrial revolution and has garnered significant attention in recent years due to its ability to handle large amounts of data with speed and versatility. However, criminal investigators may face challenges when conducting digital forensics on a BigData system due to the vast volume of data stored there, the potential for various attack scenarios, and a lack of relevant research. In this thesis, we propose the Responsive BigData Digital Forensics Framework. This framework is designed to respond to forensic instances, adapt to the characteristics of the incident, and gather evidence that can be used to reconstruct digital events. It integrates the Computer Domain, Network Domain, and BigData Domain to ensure that all relevant aspects of the system are considered. The investigation process includes preserving the system, searching for and extracting digital evidence, and reconstructing digital events. Moreover, this research presents a tool that assists digital forensic investigators in decision-making and incident response processes through a stored database for a set of appropriate scenarios and tools that help speed up the investigation process.

Keywords: BigData, Digital Forensics, BigData Forensics, BigData Forensics Framework, Cybercrimes.

Table of Contents

Contents

| | |
|---|----|
| Thesis Approval | 2 |
| Declaration..... | 3 |
| Dedication | 4 |
| Acknowledgment..... | 5 |
| Abstract in English | 6 |
| Table of Contents | 7 |
| List of Tables | 11 |
| List of Figures..... | 11 |
| List of Abbreviations | 13 |
| Chapter One: Introduction and Background..... | 15 |
| 1. Overview | 15 |
| 2. Background | 15 |
| 2.1. Industry 4.0..... | 15 |
| 2.2. BigData..... | 18 |
| 2.3. Industry 4.0 and BigData | 19 |
| 3. Introduction..... | 19 |
| 3.1. BigData Introduction..... | 19 |
| 3.1.1 BigData Digital Forensics Challenges | 21 |
| 3.1.2 The Architecture of BigData | 23 |
| 3.1.3 Data Collecting and Storing Layer in BigData..... | 24 |
| 3.1.4 Data Processing | 28 |
| 3.1.5 BigData analysis and visualization | 31 |
| 3.1.6 Hadoop BigData System and HDFS..... | 33 |
| 3.1.6.1 Hadoop System Architecture | 33 |
| 3.1.6.2 HDFS Architecture | 34 |
| 3.1.6.3 Hadoop MapReduce | 39 |
| 3.1.6.4 Forensics Tools | 40 |
| 3.1.6.5 Hadoop Forensics | 41 |
| 3.1.6.6 Physical Forensics | 42 |

| | |
|--|-----------|
| 3.2. Introduction to Digital Forensics | 43 |
| 3.2.1 Computer Forensics | 48 |
| 3.2.1.1 Static Forensics | 49 |
| 3.2.1.2 Live Memory Forensics | 50 |
| 3.2.2 Network Forensics | 52 |
| 4. Research Problem | 54 |
| 5. Research Questions | 55 |
| 6. Research Aims and Objectives | 55 |
| 7. Frameworks | 56 |
| 8. Research Methodology | 58 |
| Chapter Two: Related work | 62 |
| .2 Overview | 62 |
| 3. Related Framework Papers | 62 |
| 4. Related Work Discussion | 67 |
| Chapter 3: The Proposed Frameworks | 69 |
| 1. Overview | 69 |
| 2. The proposed BigData Digital Forensics Framework (BDDFF) | 69 |
| 2.1 Digital Forensics Domains | 72 |
| 2.1.1. BigData as a forensics domain | 72 |
| 2.1.2. The 3 Domains Artifacts | 72 |
| 2.2 Readiness | 73 |
| 2.2.1. Planning: | 74 |
| .2.2.2 Technical Factors | 74 |
| .2.2.3 Organizational Factors | 75 |
| .2.2.4 Legal Factors | 75 |
| 2.3 Deployment | 77 |
| 3.5.1. Detection | 77 |
| 3.5.2. Notification | 79 |
| .3.5.3 Confirmation | 79 |
| 3.5.4. Authorization | 80 |
| 2.4 Physical Crime Scene Investigation (PCSI) | 81 |
| 2.4.1. Scene management | 81 |
| .2.4.2 Securing the scene | 81 |

| | | |
|---|--|-----|
| 2.4.3. | Physical Objects Search..... | 82 |
| 2.4.4. | Physical Objects collection | 83 |
| 2.4.5. | Physical crime reconstruction | 84 |
| 2.5 | Logical Crime Scene Investigation (LCSI) | 85 |
| 2.5.1. | Identification | 85 |
| 2.5.2. | Acquisition | 86 |
| 2.5.3. | Preservation | 87 |
| 2.5.4. | Examination..... | 87 |
| 2.5.5. | Analysis | 88 |
| 2.6 | Presentation | 89 |
| 2.6.1. | Reconstruction..... | 90 |
| 2.6.2. | Report..... | 91 |
| 2.6.3. | Dissemination | 92 |
| 2.6.4. | Return Evidence..... | 93 |
| 2.7 | Documentation | 94 |
| 2.8 | BigData Assistant Tool (BDAT)..... | 94 |
| 3. | The proposed Responsive BigData Digital Forensics Full Framework (RBDDFF)..... | 99 |
| 2.4. | Forensics Domains | 100 |
| 2.5. | Crime Type..... | 102 |
| 2.6. | RBDDFF examples..... | 103 |
| 4. | The Chapter Summary | 108 |
| Chapter 4: Case, Experiments, and Results | | 109 |
| 1. | Overview | 109 |
| 2. | The Company Attack Case..... | 109 |
| .2.1 | The company's BigData system and the physical form of the system | 109 |
| 2.2. | Logical BigData system of the company | 111 |
| 2.3. | Attack scenario from the Attacker's point of view | 113 |
| 2.4. | Attack scenario from the point of view of the Victim | 114 |
| 2.5. | Attack scenario from the Forensics side | 115 |
| 3. | Explain the scenario and the RBDDFF implementation | 116 |
| 3.1. | Select the Domains | 117 |
| 3.2. | Readiness Stage | 118 |
| 3.2.1. | Planning | 118 |

| | | |
|-----------------------------|---|-----|
| 3.2.2. | Technical Factors | 120 |
| 3.2.3. | Organizational Factors | 121 |
| 3.2.4. | Legal Factors | 122 |
| 3.3. | Deployment | 123 |
| 3.4. | Physical Crime Scene Investigation (PCSI)..... | 124 |
| 3.5. | Logical Crime Scene Investigation (LCSI) | 124 |
| 3.5.1. | During the digital forensic investigation of the Hadoop Domain | 125 |
| 3.5.1.1. | Hard Drive Imaging..... | 125 |
| 3.5.1.2. | Memory Dumping | 127 |
| 3.5.1.3. | Hashing | 128 |
| 3.5.1.4. | Hadoop Logs Analysis | 129 |
| 3.5.2. | During the digital forensic investigation of the Computer Domain | 130 |
| 3.5.3. | During the digital forensic analysis of Network Traffic | 130 |
| 3.6. | Artifacts expected to be obtained from analysis processes..... | 132 |
| 3.7. | Presentation | 133 |
| 4. | The RBDDFF of The Case: | 134 |
| 5. | Explanation of The Scenario Practically..... | 136 |
| 5.1. | Digital forensics before the process of deleting the file from the system | 136 |
| 5.2. | Digital forensics during the process of deleting the file from the system..... | 139 |
| 5.3. | Digital forensic analysis during the process of uploading the virus file to the system..... | 140 |
| 5.4. | Digital Forensics investigation during virus downloading on the victim's device | 144 |
| 5.5. | Digital forensic analysis after the process of running the virus file on the system..... | 145 |
| 6. | Results and outcomes from the case investigation | 147 |
| 7. | The Proposed Report | 148 |
| Chapter 5: Discussion | | 152 |
| 1. | Overview | 152 |
| 2. | Limitations..... | 152 |
| 3. | Conclusion: | 153 |
| 4. | Future Work..... | 154 |
| References | | 155 |
| الملخص..... | | 177 |

List of Tables

| | |
|---|-----|
| Table 1. Data processing modes | 29 |
| Table 2. Related papers comparison | 66 |
| Table 3. Forensics domains array | 102 |
| Table 4. Crimes Types matrix of possibilities | 103 |
| Table 5. Virtual machines on the system | 110 |
| Table 6. Case technical factors | 120 |
| Table 7. the proposed Alpha company attack case report..... | 149 |

List of Figures

| | |
|--|-----|
| Figure 1. Industry Revolutions [2]..... | 16 |
| Figure 2. BigData architecture [22]..... | 24 |
| Figure 3. ETL process [22]..... | 26 |
| Figure 4. Hadoop architecture..... | 34 |
| Figure 5. HDFS architecture | 36 |
| Figure 6. The task of secondary NameNode..... | 39 |
| Figure 7. MapReduce steps..... | 40 |
| Figure 8. Digital forensics evidence | 47 |
| Figure 9. Forensics Types | 47 |
| Figure 10. Investigative Process of Digital Forensics [113]..... | 47 |
| Figure 11. OSI model layers | 54 |
| Figure 12. Casey & Cohen forensics frameworks | 58 |
| Figure 13. The proposed BigData Digital Forensics Framework | 71 |
| Figure14 . Forensics Domains | 73 |
| Figure 15. Readiness..... | 76 |
| Figure 16. The workflow of the BDAT | 97 |
| Figure 17. BDATA proposed report | 98 |
| Figure 18. RBDDFF of Attacking BigData system Networking Digitally | 105 |
| Figure 19. RBDDFF of a BigData and computer attack only..... | 106 |
| Figure20 . RBDDFF of Attacking All Domains digitally..... | 107 |
| Figure 21. Physical Company System | 111 |
| Figure 22 . Company HDFS system | 113 |
| Figure 23. Attacker side..... | 114 |
| Figure 24. Victim Side..... | 114 |
| Figure 25. Scenario Forensics Side..... | 116 |
| Figure 26. Case Domains possible artifacts and files to analyze | 117 |
| Figure27 . Planning Process..... | 120 |
| Figure28 . The expected artifacts outcomes..... | 133 |
| Figure29 . The case RBDDFF | 135 |

| | |
|---|-----|
| Figure30 . listing files on HDFS | 137 |
| Figure31 . FSImage for HDFS | 137 |
| Figure 32. EditLogs for HDFS..... | 138 |
| Figure 33. Timing (Human readable) | 138 |
| Figure34 . Removing File from HDFS | 139 |
| Figure 35. Listing files on HDFS..... | 139 |
| Figure 36. The original old file deletion evidence – EditLogs file Analysis | 140 |
| Figure 37. Time (Human readable)..... | 140 |
| Figure 38 . Listing Ransomware on HDFS | 141 |
| Figure 39. Edit Logs analysis..... | 142 |
| Figure 40. Human readable conversion of the mtime | 142 |
| Figure 41. CID with the IP address of cluster 2 | 143 |
| Figure 42. CID with the IP address of cluster 3 | 143 |
| Figure 43. Source Ports, Network Traffic, and IP addresses | 143 |
| Figure 44. CID with the IP address of the masternode and the block pool ID | 143 |
| Figure 45. Put Command on. Vmem | 143 |
| Figure 46. -get command on. vmem | 144 |
| Figure 47. Virus Path | 144 |
| Figure 48. Files Encrypted..... | 145 |
| Figure 49. Encrypted files by .VMDK analysis..... | 145 |
| Figure 50. The first file before encryption | 146 |
| Figure 51. The first file after encryption..... | 146 |
| Figure 52. The second file before encryption | 146 |
| Figure 53. The second file after encryption | 146 |
| Figure 54. The found Artifacts..... | 148 |

List of Abbreviations

| Abbreviation | Definition |
|--------------|--|
| BD | BigData |
| DF | Digital Forensics |
| LCSI | Logical Crime Scene Investigation |
| PCSI | Physical Crime Scene Investigation |
| CPS | Cyber-Physical Systems |
| HCI | Human-computer interaction |
| IoT | Internet of Things |
| BDDFF | The BigData Digital Forensics Framework |
| RBDDFF | The Responsive BigData Digital Forensics Framework |
| SaaS | Software as a Service |
| PaaS | Platform as a Service |
| IaaS | Infrastructure as a Service |
| Four Vs | Volume, Velocity, Variety, and Variability |
| PB | Peta Bytes |
| ETL | Extract, Transform, and Load |
| HDFS | Hadoop Distributed File System |
| GFS | Google File System |
| PLC | Programmable Logic Control |
| ACID | Atomicity, Consistency, Isolation, and Durability |
| API | Application Programming Interface |
| JVM | Java Virtual Machine |
| PDA | Personal Digital Assistant |
| FBI | Federal Bureau of Investigation |
| OSI | The Open Systems Interconnection |
| SDN | Software Defined Network |

| | |
|------------|--|
| BDFF | Bangladesh Digital Forensics Framework |
| .VMEM | Virtual Memory |
| .VMDK | Virtual Hard Drive Disk |
| DDOS | Distributed Denial of Service |
| BDAT | BigData Assistant Tool |
| PD | Physical Domain |
| VM | Virtual Machine |
| SSH | Secure Shell |
| OIV | Hadoop Offline Image Viewer |
| OEV | Offline Edits Viewer |
| IP | Internet Protocol |
| CID | Client ID |
| TCP | Transfer Control Protocol |
| ZB | Zettabytes |
| PB | Petabytes |
| NASA | National Aeronautics and Space Administration |
| GB | Gigabytes |
| OSI | The Open Systems Interconnection |
| ADI format | Amateur Data Interchange Format |
| IT | Information Technology |
| SDNs | software-defined networks |
| BDFF | Bangladesh Digital Forensics Framework |

Chapter One: Introduction and Background

1. Overview

This chapter provides a comprehensive look at the history of the four industrial revolutions, and how big data was part of it. Then the chapter talks about an introduction in which big data is clarified, its structure, and how data is stored in it. Then the chapter deals with digital forensic investigation with an introduction about it and its types and tools.

In addition, the chapter presents the research problem that this thesis is working on to solve it, the research questions and its objectives. Finally, the types of digital forensics framework are clarified and the most important of them directly related to this thesis are allocated to work on them, and the methodology of this research.

2. Background

2.1. Industry 4.0

At the advent of the twenty-first century, the fourth industrial revolution is taking place, and the commercial world is being transformed by digitalization, this is also recognized as "Industry 4.0". In the industry 4.0 environment, connected computers, intelligent machines, and smart materials communicate with each other, engage with their surroundings, and eventually make choices with little to no human assistance [1]. Industry 4.0, originally conceptualized as the fourth industrial revolution, has evolved to refer to the complete digital transformation of the industrial and consumer markets. This transformation encompasses the digitization and "smartening" of manufacturing facilities, distribution networks, and value chain participants, resulting in the

integration of the entire value delivery process. The academic, governmental, and commercial sectors frequently discuss Industry 4.0 in terms of these digitalization efforts [1]. Industry 4.0 continues the three-revolution history of industrial growth. In industries during the first revolution, the steam engine was used in the 1780s to power the mechanical loom. The second industrial revolution, known as Industry 2.0, began in the 1870s with the introduction of mass production based on the division of labor and power by electrical energy. The third revolution of digitization was pressed upon industries in the 1960s by embedded systems and information technology. The Internet of Things, the Internet of Services, the Internet of People, and the Internet of Energy are the primary pillars of Industry 4.0, which is the fourth industrial revolution. Traditional industries are transformed into modern manufacturing. Figure 1 represents an illustration of these revolutions [2].

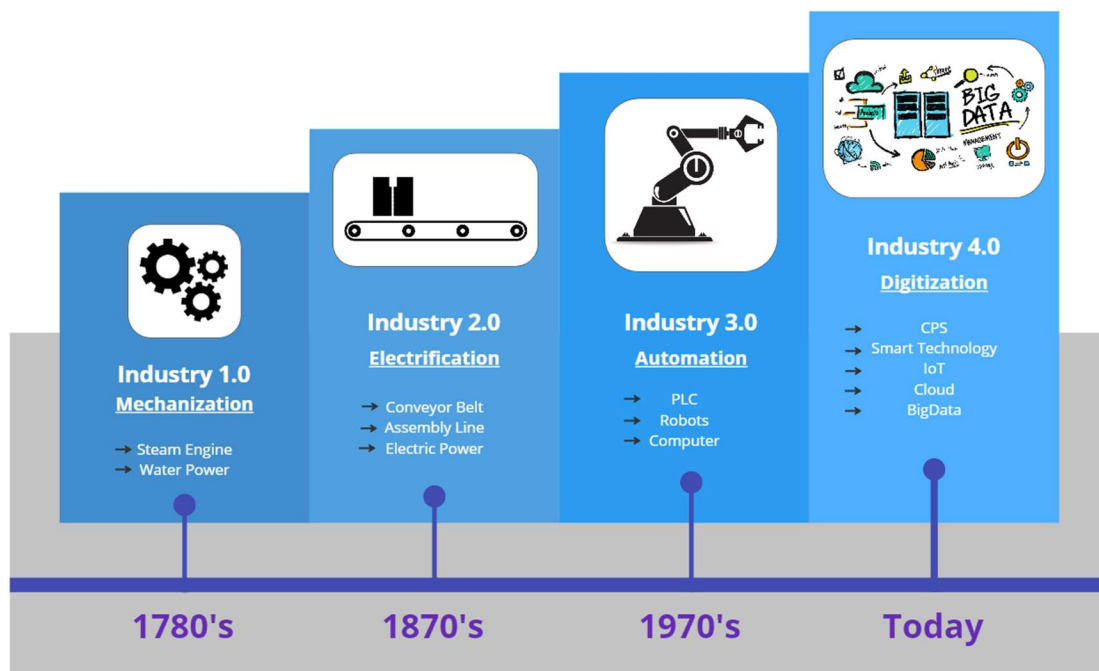


Figure 1. Industry Revolutions [2].

The objectives of Industry 4.0 are to increase operational productivity and efficiency while also automating processes to a higher extent [3]. Industry 4.0 concentrated on Cyber-Physical Systems

(CPS), intelligent manufacturing, Human-computer interaction (HCI), remote processes, the Internet of Things (IoT), cloud services, 3D printing, BigData, and other current technologies [2].

The complexity of industrial automation is steadily increasing and the data generated in manufacturing is transforming into big data. The sources of big data in Industry 4.0 include robots, sensors, actuators, wireless devices, programmable logic control (PLCs), and more. BigData refers to data that is too large to be managed and processed by traditional tools and technologies. In 2011, the world generated 1.8 ZB (Zettabytes) of data and this is expected to double every two years in the near future [4]. The widespread adoption of CPS and IoT has significantly impacted the industrial sector. Industry 4.0 refers to the integration of advanced technologies and systems. These technologies generate massive amounts of data that are challenging to process using traditional methods. BigData analytics tools offer a suitable solution for Industry 4.0 by simplifying the process of cleaning, formatting, and transforming industrial data [2].

Industry 4.0 and Big Data technologies interact and support each other. Industry 4.0 provides the data sources and infrastructure necessary for Big Data to exist and thrive, while Big Data provides the information and insights needed to drive Industry 4.0 forward. For example, Industry 4.0 generates data from IoT devices, which can then be analyzed using Big Data tools to provide insights into machine performance and productivity. Industry 4.0 and Big Data are inextricably linked, with each one supporting and advancing the other. The relationship between the two technologies will continue to evolve and shape the future of manufacturing and other industries. On the other hand, forensics plays a crucial role in ensuring the security and integrity of data generated by Industry 4.0 technologies. The application of forensics in Industry 4.0 can help in detecting and preventing cybercrimes, intellectual property theft, and other malicious activities that can compromise the data generated by these technologies.

2.2. BigData

Data has exponentially risen and increased on a vast scale during the past twenty years in many different industries in the age of industries, the use of Cyber-Physical Systems (CPS) and Internet of Things (IoT) resulted in a significant commute. Modern smart systems and technology are combined to form Industry 4.0. Smart technology creates an abundance of data that is challenging to manage using traditional tools and algorithms. BigData analytics tools are thus the best option for Industry 4.0 since they make it simple to clean, process, and convert industrial data. Volume, Velocity, Variety, and Variability are the industrial BigData [2].

BigData is characterized as datasets that are too large, too swift, and too challenging to process using relational databases or current methods. It is defined by the four variables, or the four Vs: volume, velocity, variety, and variability of data: Volume is utilized to describe the enormous amount of data that may be produced. This necessitates dispersed data processing methods and scalability in terms of data storage [5]. Velocity is the rate at which data enters the system and must be processed. Both humans and machines produce data. With more sources, the speed picks up. Velocity rises when the data production pace does, like in the case of mobile devices and sensors. Structured, semi-structured, and unstructured data are all created quickly due to variety, and they must be analyzed quickly to yield meaningful data and pertinent insights. For instance, the transactions of Walmart's worldwide client base create more than 2.5 PB of data per hour. Another excellent illustration of the rapid pace of BigData is YouTube [6], [5]. Variability is the term that refers to inconsistent changes in data volume, velocity, and variety across time. Peak data loads from data processing might occur at various periods. When it comes to storage and processing, such peak loads must be managed [5], [7].

2.3. Industry 4.0 and BigData

Before the BigData revolution, businesses were unable to effectively handle enormous data sets or keep all of their archives for extended periods. Traditional technologies are costly, have a limited storage capacity, and have strict management tools. They lack the performance, flexibility, and scalability required in a BigData environment [6].

3. Introduction

3.1. BigData Introduction

The world population is more than 8 billion, with over 5 billion people having access to the internet [8]. The rapid advancements in technology have led to a huge amount of data being generated by billions of people through the use of devices like remote sensors. And in the light of the outcomes and givens of industry 4.0 mentioned in the previous section. This data, referred to as Big Data, is characterized by its large volume, complex structure, and fast generation and processing. Unlike traditional data, Big Data often includes a lot of unstructured information which requires real-time analysis. The abundance of Big Data presents new opportunities for value discovery and a deeper understanding of underlying patterns, but also brings new challenges in organizing and managing the datasets [9], [10]. As explained in the previous section, Big Data is one of the elements of the Fourth Industrial Revolution. Every branch of science may benefit from BigData, which is a hot area of study [2]. Data Storage Market forms an essential indicator about the BigData expansion. According to Statista company statistics [11], by 2021, the data storage market's global sales had

risen to more than 42 billion USD. China produced the most income per nation, 9.9 billion US dollars, followed by the US and India, with 4.83 and 2.09 billion US dollars in sales, respectively. Except for Japan, the Statista Consumer Market Outlook predicts growth in market revenues in these nations. In 2021, the BigData analytics industry was estimated to be worth over 240 billion dollars. By 2029, it is anticipated that the market would have grown significantly and will be worth over 650 billion dollars. BigData approaches are included in many of the most widely used data processing methods, including data mining, data fusion, machine learning, artificial intelligence, social networks, and others [12].

There are several instances of Big Data. For instance, Google Maps can provide you with the least congested path to any location, Amazon gathers extensive data on the buying habits, shipping methods, and payment preferences of its millions of users, and Big Data analysis allows banks to observe and report on their operational procedures and employee behaviors.[13].

The phrase "Big Data" was first used by National Aeronautics and Space Administration (NASA) scientists in 1997 to explain the challenge of handling data sets too big to fit into a computer's RAM. The idea of Big Data is relative to the technology available at a particular moment [14]. The use of BigData in Industry 4.0 allows for the analysis of large amounts of data generated by manufacturing processes and connected devices and all kind of data gathering and dependent transactions. This can help improve efficiency, productivity, and decision-making in the manufacturing industry. The integration of Industry 4.0 technologies and BigData is enabling the development of more intelligent and connected manufacturing systems. These systems are capable of making real-time decisions and responding to changes in the environment, leading to improved efficiency, productivity, and competitiveness in the manufacturing industry.

Big data encounters a range of obstacles when it comes to the data that is assembled and examined. One of these is the complex data challenge which arises from the fact that Big Data is typically gathered from numerous sources, resulting in heavy-tailed behaviors with significant tail interdependence. Another challenge is noisy data, which refers to the presence of measurement errors, anomalies, and absent values within Big Data. Furthermore, there is a dependent data challenge found in various types of contemporary data such as financial time series, and time-course microarray data, where the samples are reliant on each other with faint signals [15].

3.1.1 BigData Digital Forensics Challenges

The field of Big Data Digital Forensics is faced with numerous difficulties and obstacles. These challenges can range from:

- a) In Preparation stage: The main difficulty in the preparation phase is handling the large amount and different types of data and the range of devices where evidence may exist. Establishing the appropriate standards, policies and procedures for the investigation is manageable but training the investigator and having the suitable tools for each scenario remains challenging. Making sure that both the investigator and their tools are equipped to deal with all applications, operating systems, mobile devices, protocols, file formats, encryption and cloud data is almost unattainable.
- b) Preservation: Preserving the evidence's authenticity and integrity throughout the investigation is a challenge, particularly with regards to knowing how to handle various devices. The methods and standard procedures for preserving data may remain the same, but having the right tool for a specific device or storage medium (e.g. a mobile phone) can still be problematic.

Additionally, the large amount of data leads to an increase in preservation time, resulting in slower response times during investigations.

- c) Evidence collection: The biggest challenge in the digital forensics process is in the collection, examination, and analysis stages due to the increasing amount, diversity, and variability of data. Despite lower storage device prices, storing large amounts of data in an uncompressed form still incurs a significant cost. Compressing the data reduces collection cost but makes it less accessible for analysis. The biggest challenge in collecting evidence comes when it is stored in the cloud, where a full forensic image of the media is not possible due to accessibility issues or large data size stored on inexpensive media. In the big data era, collecting all relevant evidence in an investigation is often not feasible.
- 4. Examination and Analysis: Investigating and scrutinizing large amounts of diverse data is a difficult task during an investigation. Despite advancements in computing technology, data volume is growing at a faster pace causing delays in examination and analysis. The traditional methods used for digital forensics, such as string search, pattern matching, and text mining, are inadequate in the current scenario due to their inability to handle the scale and underutilization of available computing resources. False positives during the examination and analysis stage contribute to longer processing times. The need to analyze large amounts of data quickly is growing and new methods or algorithms are necessary to meet this requirement. Additionally, the diversity of data in an investigation presents a challenge as digital forensic tools have limitations in the types of file formats and devices they can process. Digital forensics must be capable of handling old, current, and emerging data formats and devices, which is a significant challenge during investigations.

5. Presentation: The purpose of this phase is to communicate the findings and conclusions of an investigation in a clear manner to the audience or court. The main difficulty in this phase may be figuring out the best way to explain the examination and analysis methods used for such large amounts of data. Justifying the efficiency and accuracy of these methods may also be necessary for the evidence to be admissible [5], [16]–[19].

In this research, we work on the field of digital forensic investigation of big data, and we build a framework that helps criminal investigators in the investigation process and speeds up the process in a short time and is responsive to incidents of all kinds and targeting points.

3.1.2 The Architecture of BigData

The architecture of Big Data services can be divided into three main categories, as presented in figure 2; collecting and storage layer, processing layer, and application layer. In the layer that collects and stores data, sources for BigData services must be gathered by the appropriate hardware, and the "pre-processed" data must then be saved and processed in a distributed file system or database system. The data processing layer adopts several processing frameworks based on the various data formats. Large-scale machine learning systems that can extensively mine the potential value of data are now the key foundation for the in-depth examination of BigData. Finally, data service users are presented with results using visualization tools. BigData technology is used in a variety of sectors at the application layer. In addition, SaaS, PaaS, and IaaS-based [20] software and infrastructure are used in BigData-based cloud computing services to process BigData [21], [22].

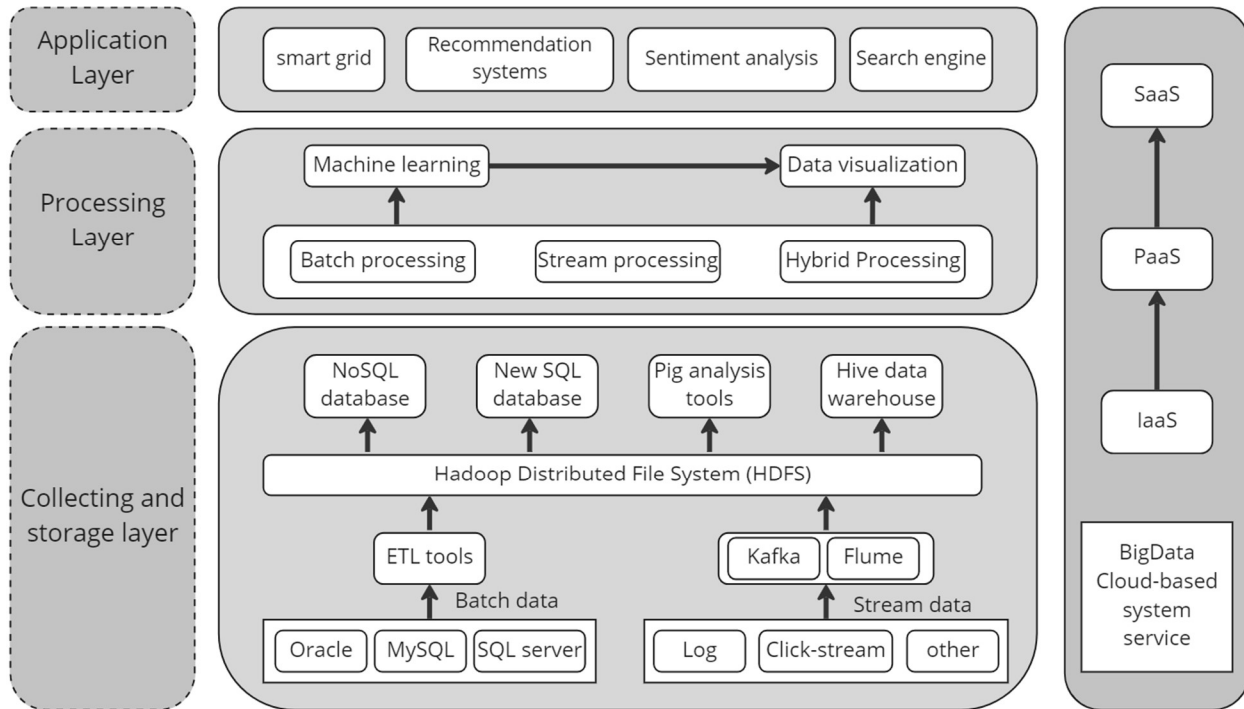


Figure 2. BigData architecture [22].

3.1.3 Data Collecting and Storing Layer in BigData

As shown in figure 2, the first layer of BigData architecture is collecting and storage layer. Static batch data and dynamic stream data are the two major types of BigData. While stream data is a continuous real-time data instance sequence, batch data is kept in a static format. The streaming data won't be saved, and many components will be discarded right away after processing [23].

Streaming data collection differs from conventional batch data collection due to the instability of stream data transfer. Extract, Transform, and Load Tools(ETL) [24] are often used to accomplish the transmission and collection of various data types for batch data from numerous data sources. Extract, Transform, and Load is the three actions that must be done in ETL. A process exists for each task. Data from diverse sources, such as ERP which refers to a category of software systems

that enable organizations to integrate and manage various business processes across different departments, SAP which is a leading enterprise software company that provides a range of business applications, and other operational systems, are consolidated during the extraction phase and placed in the data warehouse or staging area in preparation for the transformation process. The multiple subtasks that make up the ETL converting process typically include applying business rules, cleaning, filtering, separating, joining, transferring, and verifying the extracted data to prepare it for loading. Data is loaded during the loading process and stored in the data warehouse or other repositories [25] Figure 3 represents the ETL process.

The standard ETL (Extract, Transform, Load) process consists of three distinct stages. Initially, data is extracted from multiple sources, each potentially in different formats such as txt, csv, or xls. Next, transformation techniques like normalization, filtering, and sorting are applied to cleanse and prepare the extracted data. Lastly, the cleansed data is loaded into a data warehouse, where it undergoes processing and analysis to generate insights, knowledge, and wisdom [26].

Through data processing operations including connection, transformation, and cleansing, ETL eliminates corrupted or dirty data. Kettle, DataStage, Informatica, and other common ETL tools include Informatica. A collecting tool that can provide instantaneity, fault tolerance, stability, and dependability are required for the stream data that must be gathered in real-time [22]. Flume [27] is a dependable and fault-tolerant distributed stream processing system that gathers, organizes, and transmits a significant volume of log data from many sources to a centralized storage area [28]. The Hadoop environment is typically the foundation for Flume, which serves as an intermediary between data sources and recipients [29], [30]. In addition, A common open-source messaging platform called Kafka is mostly used to create real-time data pipelines and streaming applications [31]. Kafka processes data using queues to further enhance the control and processing speed of

stream data, which is capable of preventing the asynchronism in processing speed between data production and processing [32]. The other well-known platforms are Taobao's and Facebook's Scribe [33] Time Tunnel [22].

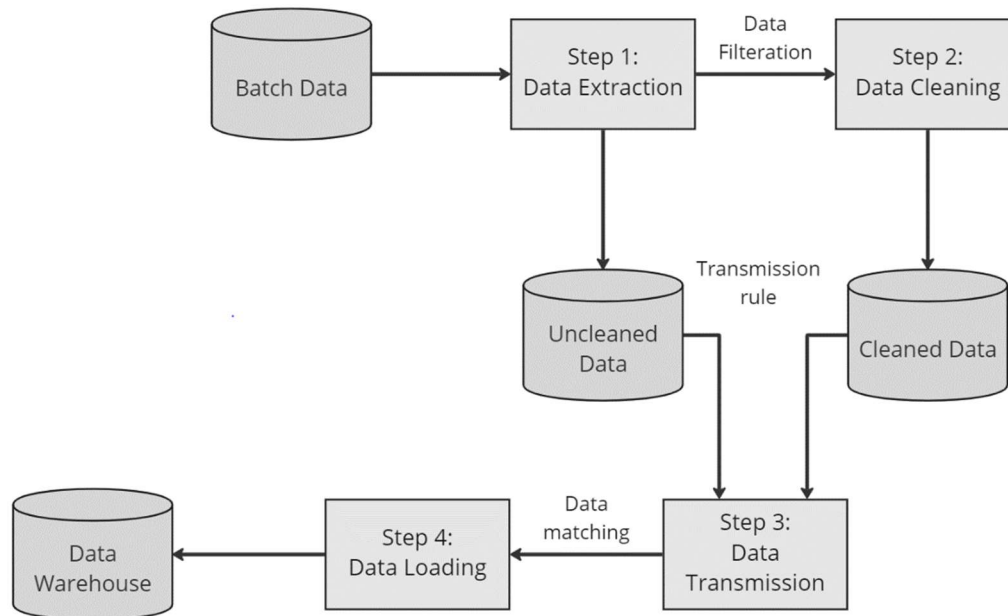


Figure 3. ETL process [22].

Relational databases and structured data management methods have been widely adopted in recent years [34]. The storage systems used are mostly distributed file systems, NoSQL, and NewSQL, which were created to manage huge data effectively, by the features of BigData [22], [35]. Google created the Google File System (GFS) in 2003 as a scalable file system for applications that required a lot of data to operate [36], [37]. It delivers excellent BigData aggregation performance and satisfies customer needs for large-capacity storage [37]. Due to its high performance and high availability, Hadoop distributed file system (HDFS), which is a component of the Apache Hadoop core project and is based on the Google File System, is frequently utilized in large-scale clusters [38]–[41]. Hadoop's main service is HDFS, a distributed file system made to run on inexpensive

hardware. It is intended to be implemented on inexpensive hardware that is extremely fault-tolerant and to offer a successful method of managing large amounts of data across clusters [41].

The increasing significance of BigData has necessitated the use of data management systems that can handle large volumes of data efficiently. Furthermore, the data generated by many applications is unstructured, making it difficult to manage. To address this issue, NoSQL data management systems were developed as a solution. These systems allow for the flexible and scalable storage and management of unstructured data. However, when dealing with very large amounts of data and the challenges of scaling, it becomes necessary to utilize distributed computing environments in order to avoid issues with latency and performance. This involves using multiple computers or servers in parallel to process the data, which can greatly enhance the speed and efficiency of the data management system. Consequently, many NoSQL data management systems utilize a distributed architecture. This type of data model involves storing data as key-value pairs, column families, graphs, or documents. This allows for the flexible and scalable storage and management of unstructured data, as well as improved performance and scalability by using multiple computers or servers in parallel to process the data. This is in contrast to traditional relational database management systems, which are designed to store structured data in a series of tables and require a fixed schema. NoSQL data management systems are particularly well-suited to handling large amounts of unstructured data and the challenges of scalability [22], [35]. NewSQL systems can be considered as a variant of traditional SQL systems that address some of their drawbacks. These systems are deployed across multiple nodes in multiple data centers, enabling distributed data management and local control. Furthermore, NewSQL systems provide Atomicity, Consistency, Isolation, and Durability (ACID) and SQL services for transactions, ensuring reliable and efficient data management [42], [43] Currently, the most commonly used NewSQL databases are Spanner

and MemSQL. These systems have been designed to address the limitations of traditional SQL systems and provide robust and scalable solutions for managing complex and high-volume data in real-time [44], [45].

3.1.4 Data Processing

Processing layer is the second layer of BigData service architecture. At the initial stage of data analysis, it is essential to clean and organize the raw data to enhance the performance of the BigData service for clients. This process, known as data preprocessing, involves cropping, integrating, and transforming the data into a usable format. There are several different approaches to data processing, including batch processing, stream processing, hybrid processing, and graph processing (see table 1). By selecting the appropriate method, researchers can effectively analyze and interpret large datasets to generate valuable insights [22]. Cleaning the data refers to the process of removing or correcting any errors, inconsistencies, or irrelevant information present in the raw data. Here are some real examples of data cleaning in the context of big data:

1. **Missing Values:** Data may have missing values, which can affect the analysis and interpretation. One approach to handle missing values is to impute them by replacing them with estimated values based on statistical techniques such as mean, median, or regression models.
2. **Outliers:** Outliers are data points that significantly deviate from the normal distribution of the data. They can distort analysis results. Outliers can be detected and removed using statistical methods like the z-score, interquartile range (IQR), or by applying domain knowledge.

3. **Inconsistent Data:** Inconsistent data occurs when there are variations in data formats, units, or values that contradict each other. For example, inconsistent date formats or conflicting measurements. Data cleaning methods involve standardizing formats, converting units, or resolving discrepancies through data transformation or rule-based validation.
4. **Duplicates:** Duplicates are identical or highly similar records present in the dataset. They can skew analysis results and introduce bias. Data deduplication techniques involve identifying duplicate records based on similarity measures and removing or merging them.
5. **Noise:** Noise refers to irrelevant or erroneous data that can hinder analysis. It can include typographical errors, sensor noise, or irrelevant data points. Techniques such as filtering, smoothing, or applying algorithms like moving averages can help reduce noise and improve data quality.
6. **Incomplete or Inaccurate Data:** Data may be incomplete or contain inaccuracies, such as inconsistent or incorrect entries. Data validation techniques, such as cross-referencing with external sources, can help identify and correct such issues [22].

Table 1. Data processing modes

| Items | Batch Processing | Stream Processing | Hybrid Processing | Graph Processing |
|-----------------------------------|---|---|--|---|
| Data Characteristics | Large-scale; High accuracy | Continuous infinite Realtime data sequence | Existing of both batch data and stream | Relevant data formed by vertices and edges |
| Processing speed | In Minutes | In Millisecond | In Millisecond | In Second |
| System Characteristics mode | Simple programming mode; Time-insensitivity; Data intensive | Serialized, low latency, event-driven triggering | Diversified workloads; High fault tolerance; Low latency | Graph data with massive nodes and edges High data correlation |

| | | | | |
|-------------------------------|---|---|---|---|
| Communication mechanism | RPC/HTTP | Realtime input stream | Memory/Disk | Distributed file systems/ databases |
| Typical processing frameworks | Mapreduce | Storm, Samza | Spark, Flink | Pregel, Giraph |
| Application scenarios | Offline analysis and processing of massive data, large-scale web information search | Pure real-life analysis; real-time scheduling, continuous calculation | Iterative machine learning; incremental calculation | Social network map; traffic road map analysis |

Because batch data is static and the amount of data is very large, data processing is typically performed using a distributed offline computing method that is capable of parallel computing. In 2004, Google designed and developed MapReduce, a software framework for distributing large data processing tasks across commodity hardware clusters. The MapReduce framework has become a widely-adopted standard for processing BigData [46], [47].

The stream data processing pattern is appropriate for handling data that requires real-time responses. As such, stream processing is necessary as a data processing framework that can achieve low latency. Apache Storm [48] is a leading real-time processing tool that guarantees to process of newly-generated information with extremely low latency. This makes it suitable for applications where quick responses are critical, such as in financial services or emergency response systems [22], [49].

Some tasks require both batch data processing and stream data processing. To address this, many data processing frameworks support both types by combining similar or related components and APIs, simplifying the process of handling different data types. This approach is known as hybrid data processing. One example of a hybrid data processing framework is Apache Spark, which is

an open-source project for fast and large-scale data processing. It is based on a batch data processing framework, but also includes capabilities for stream data processing. This allows it to handle a wide range of data processing tasks flexibly and efficiently [49]–[51].

In the context of BigData, certain types of data known as graph data are connected in the form of graphs or networks. These graphs often have a large number of vertices and edges, with some examples reaching hundreds of millions. This high degree of interconnectedness allows for complex analysis and modeling of the relationships between the data points [15].

3.1.5 BigData analysis and visualization

BigData analysis technologies, including machine learning, are utilized to extract valuable information from large data sets. This information is then used to predict and analyze future trends and patterns. Data visualization is often used to present this information to data service consumers. The growing popularity of machine learning in BigData analysis has made it an important focus in the field. One example of a novel approach to unsupervised feature learning is the research proposed in [52]. Some of the challenges and solutions for big data analysis and visualization:

1. **Volume and Scalability:** Big data requires scalable data processing frameworks like Hadoop and Spark for efficient analysis.
2. **Variety and Data Integration:** Integration techniques harmonize diverse data sources for meaningful analysis.
3. **Velocity and Real-Time Analysis:** Stream processing frameworks enable real-time analysis of high-velocity data.
4. **Veracity and Data Quality:** Data cleaning and validation techniques ensure data accuracy and reliability.

5. Complexity and Dimensionality: Dimensionality reduction and advanced visualization tools help handle complex data structures.
6. Interactivity and User Experience: Interactive visualizations and intuitive interfaces enhance exploration and interpretation.[22], [53]–[55].

BigData and cloud computing are closely related. Several major cloud services providers, such as Amazon Web Services [56], Microsoft Azure [57], and Aliyun [58], offer a range of services that support the development and integration of BigData technologies. These services include data computing, algorithm development, data service, and other technologies that are designed to meet the needs of various data development projects. By providing a complete set of BigData integration development environments on the cloud, these cloud service providers make it possible for organizations to easily and efficiently utilize BigData for their projects [59], [60].

The following cloud computing service models have been proposed as potential solutions to the aforementioned issues: software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS) [22], [61].

In the software as a service (SaaS) model, the customer is not required to install and run the application on their local computer, thereby relieving them of the burden of software maintenance and reducing the cost of software purchases through on-demand pricing [62]. The platform-as-a-service (PaaS) model of cloud computing provides services that automate the deployment and management of applications, reducing the complexity faced by application owners in managing the underlying infrastructure resources. This model allows users to develop, run, and manage applications with ease [63]. Infrastructure-as-a-service (IaaS) involves outsourcing the equipment

required to support operations, including storage, hardware, servers, and networking components [61].

Data consumers require access to valuable information after processing and analysis. To present data in a more understandable and compelling manner, data visualization often employs tables and images. By using vivid visual effects, the data is presented in a more intuitive form, improving its appeal and persuasiveness. Data visualization allows data analysts to study data more deeply from various angles and understand trends, patterns, relationships, and other information to further improve data analysis. Data visualization tools, such as RawGraphs and Google Charts for charting, Modest Maps and ColorBrewer for map creation, and Cube and Dipity for timeline creation, are widely used [22].

3.1.6 Hadoop BigData System and HDFS

3.1.6.1 Hadoop System Architecture

The Apache Hadoop software library is a distributed computing framework designed to handle large data sets and perform parallel processing across clusters of computers. It is a popular choice for researchers and professionals in academia and industry due to its ability to provide scalable storage and computing for large data sets. Hadoop is known for its efficiency and reliability, making it a widely adopted platform for data-intensive tasks. Hadoop consists of two main components, MapReduce and the Hadoop Distributed File System (HDFS). The masternode, which is responsible for coordinating and managing the slave daemons on all nodes, can also perform the role of a slave. The daemons running on the masternode are responsible for managing

the slave daemons on all nodes, which are responsible for storing and processing data [38], [41],[64].

Apache Hadoop consists of five daemons that run on their own Java Virtual Machine (JVM). These daemons are the NameNode, DataNode, Secondary NameNode, JobTracker, and TaskTracker. The NameNode and DataNode, which store data and metadata, are part of the HDFS layer. The JobTracker and TaskTracker, which track and execute jobs, are part of the MapReduce layer [65], as shown in figure 4

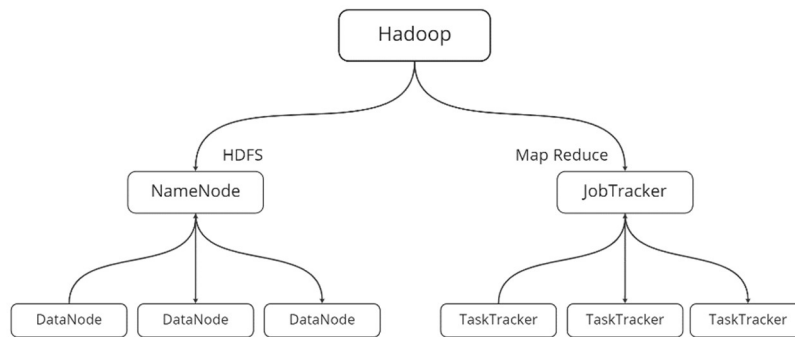


Figure 4. Hadoop architecture

3.1.6.2 HDFS Architecture

The Hadoop Distributed File System (HDFS) is a distributed file system that operates on commodity hardware and shares characteristics with other distributed file systems. However, it also exhibits notable differences. An HDFS cluster consists of a NameNode, which manages the filesystem namespace and metadata for all files, and multiple DataNodes, which store the data blocks. The key features of HDFS in Hadoop are:

1. **Distributed Storage:** HDFS distributes data across multiple machines in a cluster, allowing for parallel processing of data. It breaks down large files into smaller blocks and stores them on different data nodes in the cluster. This distributed storage model provides fault tolerance,

as data is replicated across multiple nodes to ensure data availability even in the presence of failures.

2. Replication: HDFS replicates data blocks across multiple data nodes in the cluster. By default, it maintains three replicas of each block, but the replication factor can be configured based on the desired level of fault tolerance. Replication helps in ensuring data reliability and availability.
3. NameNode and DataNode: HDFS has a master/slave architecture. The NameNode is the central metadata management component that stores information about the file system namespace, including file locations and block mappings. DataNodes are responsible for storing and serving data blocks. The NameNode keeps track of the DataNodes and coordinates data read and write operations.
4. Data Integrity: HDFS verifies data integrity by using checksums for each data block. When data is read from HDFS, the checksum is computed again and compared with the stored checksum to ensure data integrity. If a mismatch occurs, HDFS retrieves another replica of the block.
5. High Throughput: HDFS is optimized for large-scale data processing and is designed to provide high throughput. It achieves this by performing streaming reads and writes, which minimize seek time and maximize data transfer rates.
6. MapReduce Integration: HDFS is tightly integrated with the MapReduce processing model, which is a core component of Apache Hadoop. MapReduce jobs typically read input data from HDFS, process it in parallel across the cluster, and write the output back to HDFS [41], [66].

In HDFS, files are divided into one or more blocks, which are dispersed and stored on different data nodes. Each block can have multiple backups on different data nodes. When a client accesses

a file, it sends a request to the NameNode, which responds with the address of the relevant data nodes. The client can then establish a direct connection with the appropriate data nodes to perform specific read and write operations. Data nodes also regularly communicate with the NameNode to report their status and receive commands such as creating, deleting, or copying blocks. Figure 5 illustrates the HDFS architecture in more detail [64], [67].

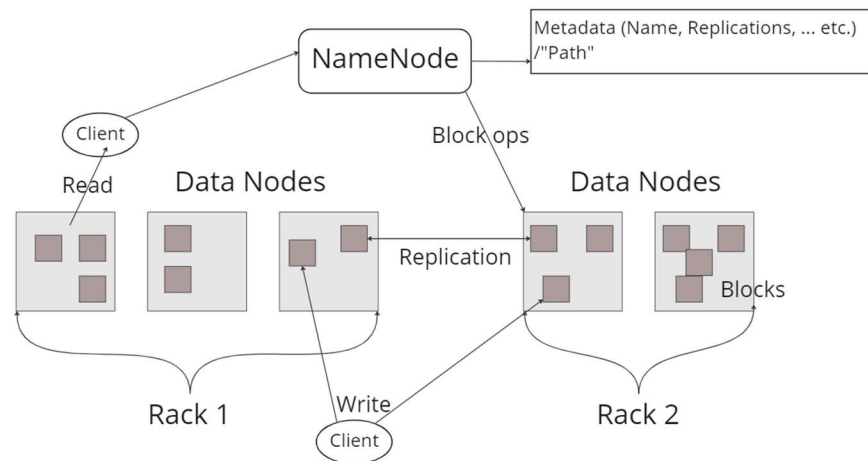


Figure 5. HDFS architecture

In HDFS, the master/slave architecture is implemented through the use of daemons. The NameNode daemon and secondary NameNode run on the master node, while the DataNode daemon runs on each slave node. The HDFS storage layer consists of three daemons: the NameNode, the DataNode, and the Secondary NameNode.

a) NameNode:

The NameNode is a daemon that runs on the master node in a Hadoop system and is responsible for managing the file system structure. It uses inodes to represent all the files and directories in the system, and stores metadata about the file system in a file called FSimage. This metadata is kept in memory to make it easier for clients to access the file system. The NameNode also controls the

slave DataNode daemons, which handle input/output tasks, and manages how files are broken down into blocks and which slave nodes should store them. It also monitors the overall health of the file system. The tasks performed by the Namenode are demanding in terms of memory and input/output [65].

The FSimage file contains all the Inodes and block information that make up the metadata of the file system. The edits file keeps track of any changes made to the FSimage file. When the NameNode starts up, it loads the FSimage file and then applies the changes in the edits file to get the current state of the file system [68]. The NameNode stores the EditLog, a file that keeps track of changes to the file system, in its local operating system file system. The file system namespace, including how blocks are associated with files and the properties of the file system, is stored in a file called the FSimage. The FSimage is also stored on the NameNode's local file system [66].

When the Namenode begins running, it retrieves the FSimage and EditLog files from the disk and combines their information in memory to create an updated version of the FSimage. This new version is then saved back to the disk, replacing the old version. The NameNode can then delete the old EditLog because its information has been incorporated into the updated FSimage. This process of creating an updated FSimage by combining the information from the FSimage and EditLog and saving it to the disk is called a checkpoint. Currently, checkpoints only happen when the NameNode starts, but there are plans to implement periodic checkpointing in the future, which would allow for regular updates to the FSimage rather than only occurring when the NameNode starts [66].

b) DataNodes:

Hadoop storage nodes are responsible for running DataNode processes, which are responsible for storing parts of the Hadoop Distributed File System (HDFS) on the local disks of the storage nodes [69].

A Hadoop cluster consists of several slave nodes, each of which runs a DataNode daemon. DataNodes are the primary storage components of HDFS and are responsible for storing data blocks and handling read and write requests for files stored in HDFS. They are controlled by the NameNode, which is the master node in the Hadoop cluster that coordinates the activities of the DataNodes. The data blocks stored in the DataNodes are replicated, as specified in the configuration, to ensure reliability and high availability. The replicated blocks are distributed across the cluster to enable fast computation [65].

When it starts up, each DataNode connects to the NameNode and performs a handshake to verify the namespace ID and software version of the DataNode. If either of these does not match the expected values, the DataNode will automatically shut down. In addition, the DataNode will send a block report to the NameNode to verify the block replicas that it is responsible for. This initial block report is sent as soon as the DataNode registers with the NameNode. The DataNode also sends a heartbeat to the NameNode every 3 seconds to confirm that it is functioning properly and that the block replicas it is hosting are available. This allows the NameNode to track the status of the DataNodes and the data they are storing [68].

c) Secondary NameNode:

The Secondary NameNode is not a backup or replacement for the primary NameNode. Its role is to periodically read the file system and log any changes, which it then applies to the FSImage file. To do this, the Secondary NameNode retrieves the FSImage and edits log files from the primary

NameNode at regular intervals, and loads them into memory. It then applies the operations from the edits log to the FSImage to create an updated version. The Secondary NameNode then sends this updated FSImage file back to the primary NameNode and updates the modified time of the FSImage file in the FStime file to track when the FSImage was last updated. The purpose of the Secondary NameNode is to help manage the workload of the primary NameNode by periodically updating the FSImage file and allowing the primary NameNode to focus on other tasks [65], [68], [70].

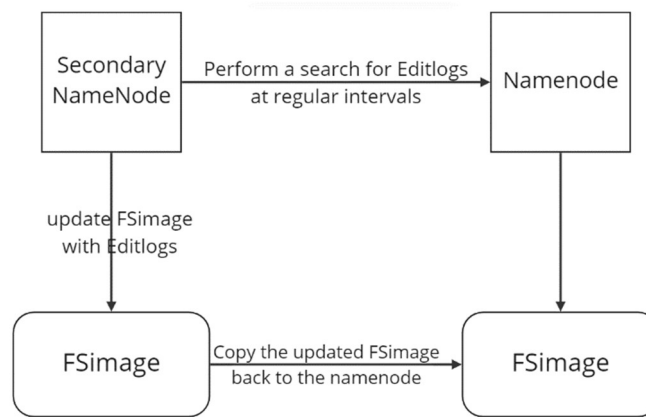


Figure 6. The task of secondary NameNode

3.1.6.3 Hadoop MapReduce

A software framework or programming paradigm called MapReduce was created to help programmers process massive volumes of data in a decentralized environment. It is implemented in many programming languages and has gained popularity for its simplicity, efficiency, and ability to handle "BigData " quickly. MapReduce consists of two main parts: a Map operation, where a simple function is used to divide the data into key/value pairs in parallel, similar to using primary keys in a relational database; and a Reduce operation, where the core processing logic is applied to produce results in a timely manner. Hadoop is a widely used, open-source

implementation of MapReduce provided by the Apache Foundation. In Hadoop, each worker node runs a TaskTracker process that executes part of each Hadoop job, while a single JobTracker coordinates the processing across all TaskTrackers [65], [69], [71]–[75].

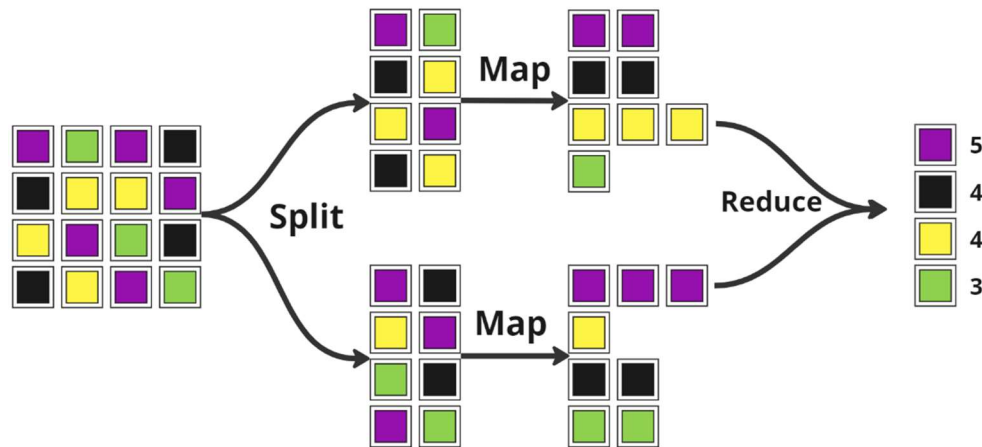


Figure 7. MapReduce steps

3.1.6.4 Forensics Tools

Collecting live memory from a system is a challenging task. It requires the collection to be performed on a live system and to have minimal impact on the system. Even a simple action such as memory de-allocation could potentially overwrite valuable data by triggering heap defragmentation. It is important to minimize the footprint of the collection process on the system [76]. FTK Imager [77], [78] can handle Amateur Data Interchange Format ADI which is a file format developed by Adobe Systems as an open standard for storing raw image data and RAW file formats RAW files store the unprocessed sensor data, which can be crucial in conducting forensic analysis. These files contain a wealth of information, including the complete pixel data, color information, exposure settings, and metadata. Belkasoft RAM capture [78] is compatible with all Windows versions and system architectures, Magnet RAM Capture [79] supports both 32 and 64-

bit systems and all Windows versions and has an easy-to-use interface, and the command-line DumpIt tool [80] is compatible with both 32-bit and 64-bit Windows machines. Wireshark [81] is an open-source tool that analyzes packets to identify potential network attacks based on trace files collected from real-world networked systems [82]–[84].

In this research we will use HashCalc for matching hashes, FTK imager for imaging virtual machine virtual hard drive and virtual memory, and FTK international for analyzing these files and extract the clues.

3.1.6.5 Hadoop Forensics

Investigating digital forensic cases in a BigData distributed file system can be difficult due to a large amount of physical data storage. The Hadoop Distributed File System (HDFS) has reportedly been targeted by cybercriminals for data theft and exfiltration. It consists of three main components: the name node, the data node, and the backup node (also known as the secondary node). The name node is the main node in the HDFS architecture and manages the data blocks and their metadata on the data nodes. The data nodes, which are referred to as slaves, store the data and handle read and write requests for the blocks of data. Each data node only stores a portion of the data and follows the instructions of the name node. The backup node periodically updates itself at designated checkpoints to be able to recover the name node in case of any failures [85]–[88].

There are still areas of the Hadoop infrastructure that have not been thoroughly explored, as many researchers tend to focus on what is easily visible rather than what is not. A common trend in Hadoop and other file system forensics is the analysis of log files. While log files contain a wealth

of forensic data, many papers in [89]–[94], [95]–[101] express skepticism about relying heavily on these files for forensic purposes.

As noted in the papers, logs are considered a fragile source of forensic information because they can be easily altered or deleted. To advance BigData forensic research and retrieve forensic artifacts from the Hadoop Distributed File System, there is a need to explore other areas of the infrastructure. This includes identifying the user responsible for actions and the time associated with those actions in the Hadoop environment [87].

A digital forensic investigation can analyze Log files and HDFS (Hadoop file system) to be powerful by linking the timeline in the Log files with the analysis of the Hadoop file system. This allows a strong Chain of custody to be written that can be relied upon in digital forensic investigations by criminal investigators and to be considered with both sides focusing on them to gather strong and sufficient evidence.

3.1.6.6 Physical Forensics

A proactive forensics module refers to immediate action that is taken to protect the evidence in case of live investigation to detect an incident/crime. It also denotes securing the evidence of an identified crime. On the other hand, to support digital forensics and legal prosecution, physical forensics can be defined as collecting and preserving physical evidence. Here are the main steps in a crime scene investigation:

Preserving the crime scene: The initial responder searches for and apprehends the suspect, and holds any witnesses. It is important to secure the location and limit access to only authorized personnel who are conducting the investigation.

crime scene survey: The investigator walks through the area to identify visible evidence and temporary evidence. They document initial observations about the who, what, where, when, and how of the situation, and form an initial hypothesis.

Documenting the crime scene: The crime scene is documented through the use of photographs, drawings, and video. Evidence should be documented clearly and collected.

Searching the crime scene: Search patterns are used to find additional evidence that was not identified during the initial examination. The theory developed earlier is used to look for specific pieces of evidence that have not yet been located, such as a USB or Hard Drive.

Reconstructing the crime scene: The events that took place at the crime scene are determined by analyzing the crime scene itself, the positions and locations of physical evidence.

3.2. Introduction to Digital Forensics

Digital forensics is a rapidly evolving field that requires constant monitoring of new technologies to stay current. The widespread use of smartphones and tablets, as well as the constant evolution of operating systems and computer file systems, necessitates a thorough analysis to determine the best methods for extracting relevant information in investigations. Techniques for forensic investigation on various technologies are also constantly being developed and updated. To stay up to date in digital forensics, it is necessary to keep track of these technological and investigative advances [102].

Digital forensics has become a crucial tool in detecting computer-related crimes, and it is widely used by law enforcement and organizations to investigate electronic media and incidents. Despite

the significant progress made by forensic tools like Encase and FTK, they still face challenges that need to be addressed. Eric Holder, the former Deputy Attorney General of the United States Subcommittee on Criminal Oversight for the Senate, categorized these challenges into three groups: technical challenges (such as encryption and live analysis), legal challenges (like jurisdictional issues and lack of international legislation), and resource challenges (including the large volume of data and the time needed for forensic analysis) [103].

The gathering, analysis, interpretation, and reporting of digital evidence have long been considered the primary tasks of digital forensics (DF) [104]. The area requires the employment of forensic software to examine digital data, much of which is stored and read as electronic signal information that is locally processed and translated. Although digital storage media, such as a conventional USB memory stick or hard drive, may withstand a thorough physical and visual examination, the data contained within can only be examined with the aid of specialized hardware and software that can decipher it and present it in a readable manner [104].

Digital Forensics has many advances such as: Digital evidence recovery aids in solving crimes and providing legal proof. Helps investigate cybercrimes and identify perpetrators, motives, and methods. Supports proactive security measures by identifying vulnerabilities. Assists in incident \ response to contain and mitigate security breaches. Provides evidence for legal proceedings, ensuring justice is served. Moreover, it faces many challenges: Keeping up with technological advancements and new devices. Overcoming encryption barriers while respecting privacy laws. Handling and analyzing large volumes of data. Dealing with anti-forensic techniques employed by perpetrators. Navigating jurisdictional and legal complexities. Preserving evidence integrity and maintaining a proper chain of custody [105], [106], [107].

Digital forensics involves the examination of data files, software, computer operations, and other electronic files or digital content stored on technology-based devices such as personal digital assistants (PDAs), digital cameras, and mobile phones. The goal of forensic science is to determine how digital evidence can be used to identify suspects and reconstruct events and to analyze and assess the state of victim machines. This analysis is often used in criminal or civil court cases to provide evidence in support of an investigation [108].

The recovery of evidence from digital devices is an aspect of digital forensic science, which is commonly described in terms of process models that represent the steps of the investigation in [109]. The procedure is divided into steps so that it can be discussed how AI methods have been used in a digital examination. These follow Interpol's acquisition, examination, analysis, and presentation rules. Several different data sources can be utilized in this general digital research process: conventional desktops, portable devices, and other embedded gadgets (such as smart home devices and other IoT devices), It is also applicable to live forensics, network forensics, and cloud forensics [109].

Legal investigations are facilitated by the analysis of digital sources of evidence in the discipline of digital forensics. In contrast, cybersecurity is a field where protecting users' privacy and the security of their digital data is a major priority [110].

Logical Forensics Investigators are responsible for extracting evidence from digital devices, such as computers, operating systems, laptops, smartphones, and other electronic devices, found at a crime scene. They use digital forensics techniques to uncover and analyze electronic data related to cybercrime and preserve it for use as legal evidence in court. Digital forensics evidence can be

obtained from a variety of sources, as shown in Figure 8. There are also several different types of forensics, as presented in Figure 9. The goal of digital forensics is to collect evidence related to cybercrime without altering or modifying it in any way [111].

Digital forensics involves the analysis of digital evidence to support or refute a hypothesis in a legal investigation. As more and more evidence is being found in digital form, investigators need to be able to access and analyze this material to properly conduct their investigations. This field is growing in importance as the amount of digital evidence increases [112].

The process of digital forensics investigation involves four main stages: preservation, collection, examination, and analysis. Preservation stage involves preventing any activity that can damage digital information during the collection process. Collection stage involves finding and collecting digital information that may be relevant to the investigation. Examination stage involves searching for evidence in a systematic manner, and the outputs of this stage include data objects found in the collected information. Analysis stage involves drawing conclusions based on the evidence found, and reporting stage involves creating a report that outlines the examination process and relevant data recovered during the investigation [113].

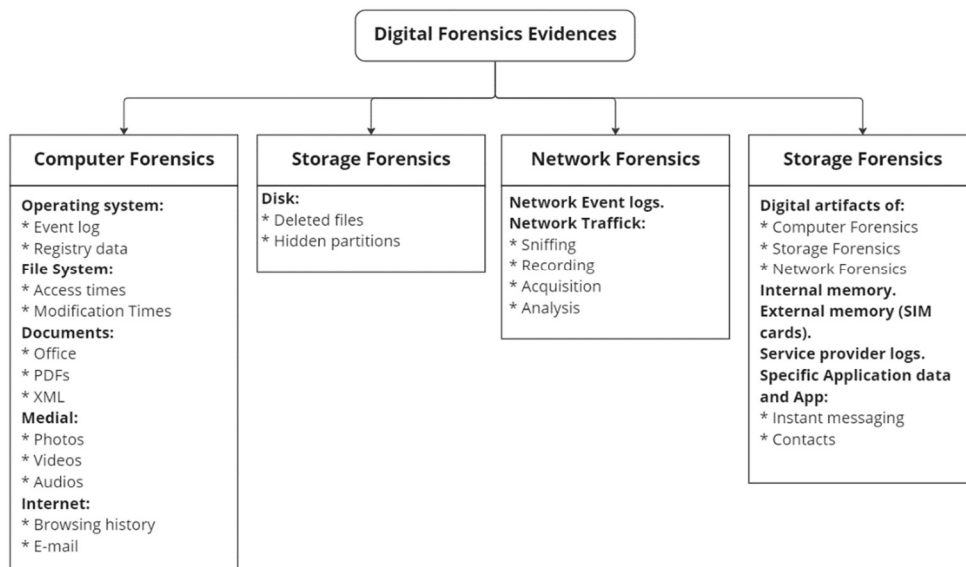


Figure 8. Digital forensics evidence

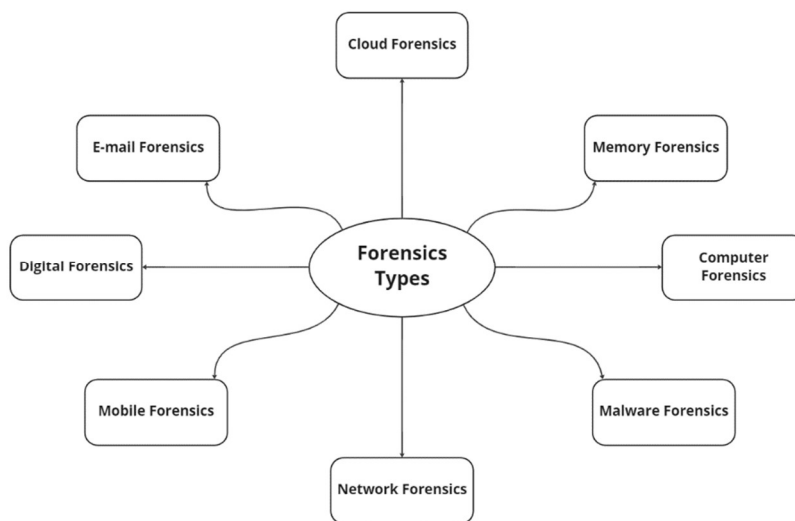


Figure 9. Forensics Types

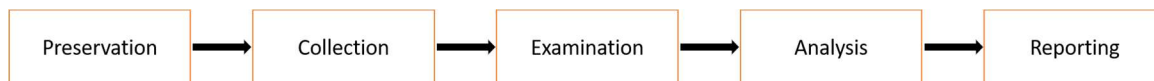


Figure 10. Investigative Process of Digital Forensics [113]

3.2.1 Computer Forensics

The FBI and other law enforcement agencies have been working on methods to analyze computer evidence since 1984. The specific techniques and procedures used in a digital forensic investigation can significantly impact the outcome of the investigation. Using inappropriate methods or skipping steps in the process may result in incomplete or missing evidence, which can lead to invalid conclusions. It is important to follow a structured and systematic process when collecting digital evidence, as evidence gathered in an unstructured manner may not be admissible in court [114], [115].

Computer forensics involves the use of specialized techniques and methods to collect and analyze digital evidence from computing devices. The goal of this type of investigation is to conduct a thorough and organized analysis to determine what happened on a device and who may be responsible. To do this, experts must carefully preserve and gather data from the device in a way that maintains the integrity of the evidence [108].

Pollitt has suggested a methodology for conducting digital evidence investigations that will produce scientifically sound and legally acceptable results. This methodology consists of four phases: Acquisition, Identification, Evaluation, and Admission. In the Acquisition phase, evidence is collected with proper legal authorization. In the Identification phase, the digital components of the evidence are identified and translated into a format that is understandable to humans. In the Evaluation phase, the relevance and legitimacy of the identified evidence to the case being investigated are determined. In the Admission phase, the collected and analyzed evidence is presented in court [114], [116], [117].

In computer forensics, the "snatch and grab" method involves shutting down a system and making a copy, or image, of the disk either on-site or in a laboratory after seizing the machine. The image is then analyzed in a controlled environment using standardized processes that are acceptable to both the prosecution and defense. This allows the analysis to be repeatable and the results to be reliable and agreed upon by both parties [118]. Traditionally, computer forensics focused on the collection and analysis of data stored on permanent storage devices such as hard drives. However, as computers have increased in storage capacity and networks have become more widespread, traditional techniques are no longer sufficient. Live forensics, which involves the analysis of data while a computer system is still running, has become more popular in the field of computer forensics to keep up with these changes [119]. Computer forensics is a type of digital forensics that involves the analysis of software, sessions, the Registry, or physical storage devices such as hard drives, RAM, or internal storage to gather evidence. This evidence is typically in the form of "artifacts," which are traces of activity or data that can be found on a computer or storage device. The goal of computer forensics is to use these artifacts to understand what has happened on a device and identify any potential wrongdoing.

3.2.1.1 Static Forensics

Static forensic analysis is a technique used to examine offline evidence, such as the contents of a hard drive when the computer is turned off. This is because the data on a computer may change when it is turned on. Static forensics involves analyzing a copy of the disk to retrieve data such as deleted files, web browsing history, network connections, and user login history. To perform static forensic analysis, investigators typically use external devices like USBs and hard drives to transfer

the data to a forensic laboratory, where it can be analyzed using various forensic techniques. The goal of static forensic analysis is to gather evidence for a legal investigation [108], [120], [121].

3.2.1.2 Live Memory Forensics

Live forensics is a technique used to extract data from a computer's volatile RAM to investigate criminal activity. Volatile RAM is a type of memory that is only present when a device is powered on and is erased when the device is powered off, making it a valuable source of information for forensic investigators [122], [123].

In today's world, it is common for computers to have large amounts of RAM to run modern operating systems efficiently. While 4 GB of RAM is a typical amount, it is not uncommon to find computers with as much as 64 GB of RAM. By creating an image of the RAM and analyzing the data stored in it, forensic investigators can gain valuable insights into the actions taken on a computer and potentially gather evidence for a criminal investigation. As such, it is important to include RAM imaging and analysis as a key step in any forensic investigation [102].

Computer RAM can be a useful source of forensic evidence because it stores important information about various computer processes, malware, chat logs, registry handles, networking events, user logs and credentials, and encrypted passwords [124]. Any activity that occurs on a computer is stored in its RAM promptly, including both simple and complex actions [102]. Memory forensics involves two main stages: acquiring a copy of the memory and analyzing it. There are various tools and techniques have been developed to assist with these phases, including FTK Imager [77], Foremost [125], Helix [126], Volatility [127], and WinHex [128], [129].

Memory forensics techniques can uncover a significant amount of evidence that would be lost if traditional forensic procedures were used. This process involves two steps: acquiring a copy of the computer's memory and analyzing it to extract evidence. The success of the analysis and the evidence produced depend on the successful acquisition of the memory. The most common method of memory acquisition is to copy the contents of physical memory to a different storage medium for further analysis [130], [131].

Physical memory is not as commonly considered in the digital forensics process as it has been in the past. Previously, digital evidence collection methods focused on file carving from static media such as hard drives, CDs, and flash memory devices. However, physical memory is different from these traditional sources of evidence. A long-term data storage space is static media, while physical memory is a short-term storage area used by the operating system to handle computing data. Overall, physical memory is more challenging to obtain substantive information and analysis from. Due to these difficulties, physical memory forensics is less commonly used in digital forensics. However, there are some situations where live forensics, or the investigation of a system while it is still running, may be necessary to gather additional information for analysis [122], [132], [133].

Memory forensics can be avoided by some investigators due to its volatility, which means that data is quickly lost and not permanently stored. Additionally, RAM has a fast overwrite rate, so even if data is recovered, it may be incomplete or damaged.

3.2.2 Network Forensics

The IT industry's emergence, along with its security concerns, has motivated the development of network forensics. Organizations are particularly concerned about the security of their networks and data due to the numerous attacks on various companies. Recently, there have been numerous attacks on various social networks, such as Google Blogger, Facebook, and Twitter. Additionally, Phishing attacks are used to get users' private information, such as bank account passwords, which can allow intruders to quickly become billionaires. These criminals must be located, and the prosecution will require proof to condemn them in court [134]–[137].

The ISO OSI model comprises seven layers, with routers and network devices usually operating in the bottom three layers and hosts operating in all seven layers. Each layer handles data in a unique way, with a Protocol Data Unit (PDU) serving as the unit of data handled by each layer. Some layers add layer-specific information in the form of a header, trailer, or both to the data. This added information contains details that help control communication between two entities, with headers added at the start of the PDU and trailers added at the end. Starting from the bottom, the seven layers are Physical layer: responsible for the transmission and reception of raw bit streams over a physical medium. Data link layer: responsible for the reliable transfer of data between network nodes, error detection, and flow control. Network layer: responsible for the delivery of packets from source to destination, including addressing, routing, and traffic control. Transport layer: responsible for end-to-end data delivery and error recovery, including segmentation and reassembly of data. Session layer: responsible for establishing, managing, and terminating sessions between applications. Presentation layer: responsible for the formatting, encryption, and compression of data for presentation to the application layer. Application layer: provides access to

network services for end-user applications and supports communication with other network applications. Each layer communicates with its counterpart on another device using protocols, and the interaction between layers is known as the protocol stack. The OSI model provides a framework for understanding how data is transmitted across a network and enables the interoperability of different network technologies and devices [138].

Online services transmit massive volumes of data in a variety of formats through communication networks, with network packets being the most prevalent. Network packets are a type of data unit that consists of groups of bits containing both data and control information. These packets are typically used to represent network traffic flow at a specific point in time, and they are the smallest unit of data that can be intercepted and logged about network traffic on packet-switched networks. They consist of control information such as source and destination IP addresses, error detection codes, and sequencing information, as well as payload, which is the intended message. Network packets are generally associated with the network layer (The Open Systems Interconnection OSI Layer 3) protocol data unit. In the OSI model, a data unit at the data link layer (OSI Layer 2) is called a frame, which is a collection of bits containing information including one or more addresses and other protocol control information. At the transport layer (OSI Layer 4), the equivalent data unit is called a segment or datagram. Network packets, when properly captured, stored, and processed, can be used in forensic investigations and potentially even as evidence in court cases against a suspect [139].



Figure 11. OSI model layers

4. Research Problem

Through the analysis of related scientific studies in Table 1 and their discussion in the previous chapter in the Related Work Discussion section, we find that no scientific study covers all the processes/operations listed in Table 1. In this study, we build a framework called the Responsive BigData Digital Forensics Framework to assist digital forensic investigators in conducting digital forensic investigations of big data and covering all the processes/operations discussed in Table 1. Additionally, it is capable of responding to different types of events that target several domains of these systems, covering digital and physical forensic investigations. Furthermore, it has a fast response time to reduce time. Moreover, the framework proposes a tool called the BigData Assistant Tool to help digital forensic investigators in conducting investigations and suggesting appropriate tools and techniques for these investigations. Additionally, it is capable of generating automatic reports ready for presentation to courts, based on standard forms, and the company has the ability to add or modify it as desired.

5. Research Questions

- Can the Framework accelerate the process of criminal investigations?
- Is the Framework capable of addressing various forms of attacks?
- Is it possible to obtain evidence from the BigData system and establish the Chain of Custody?
- Does the final report ensure admissibility in courts?

6. Research Aims and Objectives

In this study, we design the BigData Digital Forensics Framework (BDDFF) in which the steps of conducting a big data digital forensics investigation can be followed to help investigators during the investigation process. Furthermore, we develop this framework into a Responsive BigData Digital Forensics Framework (RBDDFF) to speed up decision-making and to make the criminal investigation process easier and more efficient within a fewer time. in addition, to handling events depending on scenario type, targeted areas, and incident response capabilities. moreover, we develop software that facilitates action and familiarizes criminal investigators with the necessary tools for the investigation process, and it can expand depending on new occurrences as cases are added to the database to reinforce it. finally, we make a case to test our RBDDFF to explain it more and to ensure that the investigation process follows the legal standards for acceptance in the courts.

7. Frameworks

To speed up the investigation process and get around issues like the enormous amount of storage space on digital devices, a common framework that is offered in [140] to guide the process of digital forensics is crucial. Over time, several frameworks have been developed, and each one aims to include new technology and methodologies over its predecessor. Most recent research has focused on using new techniques and technology to enhance existing frameworks from various angles, such as efficiency and accuracy; other recent research has focused on addressing brand-new issues [140].

, the results of forensic laboratory analysis, and using the scientific method [141]–[144].

➤ Forensics Frameworks

Casey 2004 Forensics Framework, is a set of guidelines for digital forensics investigations. The framework is a comprehensive, structured approach to digital forensics that covers the entire process of a digital forensics investigation, from the initial request for assistance to the presentation of findings in a court of law. Cohen Forensic Framework is a digital forensics model developed by Cohen in 2009. It is a high-level, process-oriented framework that provides guidance for the conduct of digital forensics investigations. When we go beyond the differences in terminology and level of detail, certain process models emphasize important aspects to consider during a digital investigation. The typical steps for conducting a thorough and competent digital investigation are:

- Preparation: This stage involves the identification of the scope of the investigation, the selection of appropriate tools and techniques, and the planning of the investigation.
- Acquisition: This stage involves the preservation and acquisition of digital evidence from a variety of sources, including computers, mobile devices, and other digital media.
- Examination: This stage involves the analysis of the acquired evidence to identify relevant information and extract meaningful data. This may include tasks such as data recovery, file carving, and image analysis.
- Analysis: This stage involves the interpretation of the data extracted in the examination stage and the development of a comprehensive understanding of the facts of the case. This may include tasks such as timeline analysis, email analysis, and network analysis.
- Presentation: This stage involves the preparation of the findings of the investigation in a manner that is suitable for presentation in a court of law. This may include the preparation of reports, charts, and other visual aids.

Both the Casey and Cohen frameworks provide a structured approach to digital forensics investigations and incident response. The Casey framework provides a more comprehensive and detailed approach to digital forensics, while the Cohen framework is focused specifically on incident response. Both frameworks can be useful tools for digital forensics practitioners and incident responders, but the choice of framework will depend on the specific needs of the investigation or incident response effort [145], [146].

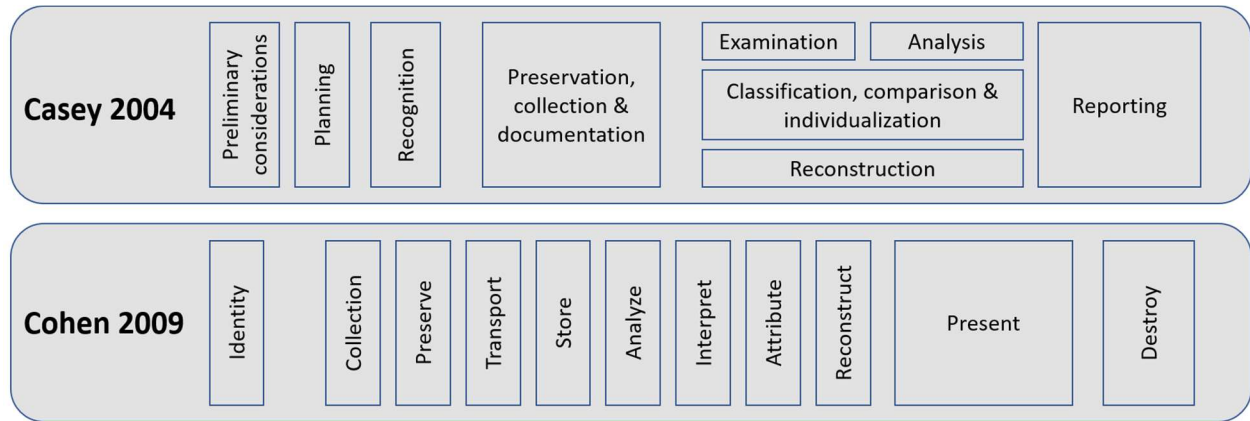


Figure 12. Casey & Cohen forensics frameworks

8. Research Methodology

Our research methodology is based on building a framework through which digital forensic investigators in the big data digital forensic investigation process help them solve the problems they encounter during the investigation processes. In addition to maintaining data integrity, ensuring its acceptance in courts, and following standards. Moreover, accident response and speed.

The research methodology includes several steps, as follows:

1. Identifying problems: This refers for identifying the problems faced by digital forensic investigators during big data forensic investigations. The problems faced by investigators were discussed in previous chapters.
2. Collection, Study, and Analysis: This involves collecting scientific research related to the problems that have been addressed in several solutions, studying these researches and then analyzing them. Additionally, identifying the strengths that each research focuses on and the weaknesses that it briefly covers or does not cover in the desired way to make the investigative

process comprehensive for all elements of the investigation. Furthermore, tracking the methodologies used in each research and the objective the researcher was trying to achieve or through the framework they developed. Also, documenting everything collected, analyzed, and concluded from each research, which will be an important reference when building the framework. The nature of the research collected is research that deals with Digital Forensics, Big Data, and Big Data Digital Forensics. Table 1 deals with the strengths and weaknesses of each research and what our framework will cover.

3. Criminal forensic analysis of the scene: According to related research, there are many scientific studies that have never addressed the physical aspect (see Table 1). We have covered the physical aspect because attacks can occur digitally and/or physically. Therefore, we have collected and identified the necessary steps that will be useful in the investigation process and linked them with the framework to be comprehensive. The physical aspect is a crucial element in the investigation process. It is possible to find many devices in the crime scene such as USBs, CDs, or others. These devices may contain important digital evidence that helps in understanding the event and building the Chain of Custody.
4. Based on the Cohen and Casey Frameworks: In this process, both the Casey framework and Cohen are studied in detail. The starting point for building the framework and defining its key elements will be based on them.
5. Digital Forensics Analysis of the BigData System: In this process, the analysis of systems that work on BigData and the study of its Filesystem, and the study of its Architecture is carried out. In addition to identifying all the points that can be used to collect Artifacts as evidence for investigation and adding them to the Framework. This system will be applied as a Case in subsequent steps.

6. Building the Framework: In this process, the Framework is built based on the accumulation of all previous steps. It relies on all the strengths and weaknesses and employs them as key elements covered by the Framework. In addition, the Framework will be developed based on the Cohen & Casey. Furthermore, the BigData system will be employed to complete the Framework under the name BigData Digital Forensics Framework (BDDFF).
7. BDAT Tool: A tool to assist digital forensic investigators under BigData, named BigData Assistant Tool, has been proposed. The forensic investigator can select similar cases to the one they are working on or may be the same based on what is stored in the database. In addition, the tool can store new cases to become stronger and evolve faster. Moreover, the tool can make reports auto generation based on the inputs entered by the forensic investigator and built on standard forms to assist in providing evidence to the judiciary.
8. After we build the BigData Digital Forensics Framework (BDFF), we start to develop it to be Responsive, that is, to be able to respond to accidents and attacks that it will be exposed to forensically, and the form of the framework can change based on the nature of the attack and what domains it targeted from the Full Framework and became The Responsive BigData Digital Forensics Framework (RBDDFF), the next sections describe it in details.
9. Finally, we perform a practical application simulation on the framework, where Hadoop system is used on a computer with 3 virtual machines for a hypothetical company called Alpha. Each machine is installed with CentOS 7 and the company does not own any protection systems other than the firewalls specific to the operating system. The company has a simple Cyber Security department. At some point, a hacker attacks the system by exploiting the privileges of the Masternode and accessing the system in some way. In addition, the hacker searches for one of the important files that are heavily loaded on a number of devices, and after

finding it, deletes it and uploads another file with the same name, extension but is Ransomware. The device from the connected devices to the system downloads this file, then runs it. After running, its files are encrypted and a ransom is requested in exchange for decryption. Investigators are looking into the case to see how it happened, where it came from, and who did it. To start the case, the team is using the RBDDFF system, as it ensures the acceptability of evidence in court, explains the investigation process, is quick to respond to incidents, and can predict the artifacts that can be collected, as well as providing a tool for auto-report generation. The investigators first identify the targeted domains, to exclude domains that have not been targeted. Then, they determine the crime type and eliminate the type that was not targeted if any. In addition, they create diagrams that explain the physical layout of the company on the ground and another diagram that explains the Big Data system architecture. Then, all the processes are applied to the framework to start the investigation, writing down the expected artifacts and where to search during the digital forensic investigation of massive data. In addition, after applying the framework, practical work is done by taking all measures from the framework and extracting evidence. After collecting digital evidence, it is entered into the BDAT for report generation.

In the next section, we will explain the BDDFF specific to this research, with a detailed explanation of all the points in it.

Chapter Two: Related work

2. Overview

In this chapter, we present the related work papers of our research. Additionally, we present the points of focus of our research and related research in a table that shows all the important points for building a framework of digital forensic investigation in the field of BigData and the key differences between the related research papers and ours.

3. Related Framework Papers

The BigData technology is a new economical service model that loads and extracts the acquired data from diverse data sources while using data as an asset. A range of specialized data processing methods, data analysis, and visualization services are available through this service architecture. The BigData service architecture is briefly defined in [22], along with the technological underpinning that handled data collection and storage, as well as features for deep BigData analysis using machine learning and outlined the technologies of BigData visualizations. After that, it highlights a few real-world uses for BigData.

One of the primary issues that digital forensics investigators in the case deal with are the increase in the amount of stored data that has to be reviewed. With the emergence of the BigData system, this issue grew more acute, prompting a reevaluation of how to conduct a digital forensics inquiry into such a large volume of data. The researchers developed a new framework based on earlier ones that are more complemented and detailed and differ from others in each stage. The purpose of the developed framework is to work on different investigations in the BigData system started,

to clarify each stage, and to propose the required techniques in each scenario. The paper [5] briefly describes the main challenges that face the digital forensics investigators in the BigData system and the needs in different scenarios.

Due to BigData's organizational structure, the study [147] describes the substantial obstacles that the Digital Forensic Investigator faces in the system. Finding the right data might be like searching for a needle in a haystack, additional difficulty is that system investigations may not be handled well by the instruments employed in digital forensics investigations. The researchers address these problems by presenting opportunities as remedies at each level of the investigative process.

Chapter [148] discusses how the limited application of BigData Digital Forensics methodologies in digital investigation and crime prevention prevents a competent analysis of the criminal information that is currently accessible. This has a detrimental impact on the efficacy of Logical crime investigation operations. It demonstrates how BigData technology may be used to analyze all available options and directions for resolving common problems and preventing Logical crime. It concludes that BigData technologies should be used in digital investigation and crime prevention in order to achieve a highly effective technique to handle challenging digital investigation jobs and prevent Logical crimes, to fulfill the criteria, it should complete the top-priority infrastructure activities, beginning with the decision-making process, processing of analytical data, organization and management of all evidence-connected data systems, and provision of necessary IT infrastructure.

According to [149] explanation, Digital Forensics Investigators must deal with a massive volume of heterogeneous growing data that has been seized more than ever, extremely varied data formats, and growth in the complexity of dispersed stored data, It sought to draw attention to the difficulties

with BigData forensics and the potential solutions for cybercrimes, concluding that one approach is to use advanced data analytics, so artificial intelligence and computer modeling must be incorporated into the investigation to help find solutions quickly and effectively.

In order to examine and identify BigData forensics difficulties, the research [150] develops a new concept of BigData forensics, It provides a theoretical framework for managing BigData forensics investigations, illustrates various cases, and suggests a few potential applications of BigData for digital investigation.

The study [151] describes the trends in BigData digital forensics, and the difficulties in acquiring BigData evidence, and suggests an application employing data science and machine learning algorithms to correctly analyze the enormous amount of data. For enterprises employing BigData for digital investigations, it is intended to establish an investigation model and calls for the use of the right technique for defining architecture and putting into place cost-effective system requirements.

The article [152] goes through the difficulties that BigData system networking theory is now facing and possible solutions. It outlines the current situation of BigData networking challenges related to processing, management, and capacity, in order to identify relevant problems and answers, it also discusses and compares the MapReduce and Hadoop model structures, research difficulties, fabric networks, and software-defined networks (SDN) that are utilized to govern today's rapidly developing digital world.

The Bangladesh Digital Forensics Framework, proposed by [144], is a novel digital forensics framework based on Logical crimes in Bangladesh (BDFF), The framework provides a detailed

and organized method for guiding digital forensics investigators from the Readiness stage to Dissemination. Additionally, it suggests a design for an expert-developed tool that includes an Assistive Expert system that uses an inference engine and a historical database to speed up and improve the efficiency of the investigative process.

The study [153] outlines a framework for BigData Forensic technology, explains several technologies that might be used during a BigData digital forensics investigation, and provides examples of each. The framework is broken down into three primary areas. The first is Digital Forensics Technology, which illustrates the forensics process from the Collection stage through Data representation. The second part, called Intermediate Technology, demonstrates how forensics and BigData system technologies may be combined. The final part, BigData Technology, demonstrates the need for analysis of the technology handling the enormous amount of data.

In order to move toward a more detailed explanation of what "BigData" is, the study [72] aims to assess and characterize its characteristics. It provides a framework for developing BigData applications, with three levels for each step and seven layers overall to break the BigData application down into modular components. To gain a competitive edge by offering management more control over data processing for the businesses, the primary objective is to efficiently manage and build large BigData applications

| Operation/ Ref no. | DF Framework | BigData Framework | BigData Forensics Framework | Forensics domains coverage | Readiness | Deployment | Physical crime scene | Logical Crime Scene Investigation | Evidence presentation | Assistant Tool | Documentati on in | Framework Scenarios implementati on | Framework Incident Handling | Framework Incident Responsivity |
|-----------------------|-----------------|----------------------|-----------------------------------|----------------------------------|-----------|------------|-------------------------|---|--------------------------|-------------------|----------------------|--|-----------------------------------|---------------------------------------|
| [22] | × | ✓ | × | × | × | × | × | × | × | × | × | × | × | × |
| [5] | ✓ | × | × | × | × | × | × | ✓ | ✓ | × | × | × | × | × |
| [147] | × | × | × | × | × | × | × | ✓ | ✓ | × | × | × | × | × |
| [148] | × | × | × | × | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × |
| [149] | × | × | × | × | × | ✓ | × | ✓ | ✓ | × | ✓ | × | × | × |
| [150] | × | × | × | × | ✓ | ✓ | × | × | ✓ | × | × | ✓ | × | × |
| [151] | × | × | × | × | × | × | × | × | ✓ | × | × | × | × | × |
| [152] | × | × | × | ✓ | × | × | × | × | × | × | × | × | × | × |
| [144] | ✓ | × | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × |
| [153] | × | × | ✓ | ✓ | × | × | × | ✓ | ✓ | ✓ | × | × | × | × |
| [72] | × | × | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × |
| Our Research | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 2. Related papers comparison

4. Related Work Discussion

Many scientific studies have been conducted on BigData, others on Digital Forensics, and still others on the relationship between BigData and Digital Forensics and the BigData Digital Forensics Framework.

Based on the Cohen & Casey Frameworks and their integration with BigData operations, a summary of the processes & stages was created in Table 1. These functions were performed on the related scientific studies mentioned in the previous section. The results were different, and the following section explains each process/operation and the studies that addressed it and those that did not: First, the scientific studies that discussed building a Digital Forensics Framework were only two in total out of 11, which are [5], [144]. Only one scientific study discussed building a BigData framework which is [22]. Secondly, only one scientific study discussed building a BigData forensics framework which is [153], indicating the lack of research in this field. In addition, several studies discussed the Forensics Domains [72], [144], [152], [153].

Several studies discussed the Readiness stage, an important phase in the investigation process, during which a plan is established on how to start the investigation, when and where, as well as determining the Organizational factors, technical factors, and finally the legal factors. The studies that discussed this are 4 studies: [72], [144], [148], [150]. Furthermore, 5 studies from the scientific research related to deployment were discussed, which were [72], [144], [148]–[150].

Some scientific studies focused on physical criminal investigations and covered this aspect and expanded upon it, which were [72], [144], [148]. Concerning digital criminal investigation of digital evidence, most of the scientific studies discussed this aspect and showed a clear interest in

it, which were [5], [72], [144], [147], [149], [153]. Evidence preservation and presentation are considered one of the most important steps in criminal investigation, during which threats are detected and verification is made of the investigation procedures, and obtaining approval and authorization to start, several important studies covered this aspect, which were [5], [72], [144], [147], [149]–[151]. In addition, researchers [144], [153] suggest a tool to assist digital forensic investigators during investigations and simplify the process for them.

Documentation is a very important step for forensic investigators as all actions and operations performed by digital forensic investigators are recorded along with the timing and date. The researchers [144], [153] talk about this process in related scientific studies. There are various types of cyberattacks that occur on digital systems and it is difficult to track them, but they can be grouped into several categories.

The attack scenarios differ according to the nature of the organization or system. Scientific research [150] discusses the possibility of different attack scenarios to be considered during investigations. Handling incidents based on their differences is a critical aspect that organizations or companies should have. There is no scientific research related to the technologies or methods to handle incidents. There is also no scientific research on how to respond to incidents in case they occur. A prompt response to incidents and handling them can greatly reduce risks. The faster the response, the more likely it is to identify the culprit in some cases. In addition, the faster the response, the less time required for the investigation process.

Chapter 3: The Proposed Frameworks

1. Overview

In this chapter, we delve to present the BigData Digital Forensics Framework (BDDFF)). This framework has been designed to address the challenges faced in digital forensics. The stages involved in the BDDFF are described in detail, highlighting the key features and benefits of this framework. Finally, the Responsive BigData Digital Forensics Framework (RBDDFF) is introduced. This framework is a recent development in the field of digital forensics and is considered a major contribution to the field. The RBDDFF is of utmost importance to criminal investigators as it provides a more effective and efficient way to conduct digital investigations. The role of the RBDDFF in the field of digital forensics is explained, along with the motivation for its development.

2. The proposed BigData Digital Forensics Framework (BDDFF)

According to our research and analysis of comparable Frameworks based on the discussion of Table 1, various crucial and missing components would significantly impact the process of gathering evidence or responding to incidents and properly handling some or all of them as explained before in Table 1 discussion section. Furthermore, they did not put it to use in practice to evaluate the effectiveness and strength of this framework and determine whether it needed to be improved. In addition to focusing primarily on BigData and addressing all of its flaws and elements

that were not addressed by other research, the proposed Digital Forensics Framework (see figure 13) also applies experience to it to assess its efficacy and strength. After discussing and showcasing our framework, the Responsive Framework with a Case Study of our work is presented in the following section.

The framework that we present consists of a set of stages, and each stage has its own procedures and steps that must be followed by digital forensic investigators to ensure a comprehensive digital criminal investigation, and to ensure the admissibility of evidence in courts.

The framework begins by identifying the Digital Forensics Domains and determining which domain the system attack occurred in. Then, the Readiness phase follows, during which a plan is developed, taking into consideration the available capabilities and specialized tools for investigating, studying, and analyzing laws to ensure the collection of evidence and presenting results in an acceptable manner in court. Additionally, in the event of an attack, the company's systems are notified, and the information security team is alerted, and permission is obtained to begin the investigation. If the attack is physical, the scene is protected, and evidence is collected in a secure and legal manner, while if it is digital, the locations where the attack occurred are determined to know what type of evidence will be legally collected and analyzed. Afterward, all evidence obtained from both physical and digital sources is collected and reconstructed to sequence events and present them in the final report to the court. All of these procedures are also collected and placed in a special database for a proposed tool that assists digital forensic investigators in expediting decision-making and facilitating future investigations.

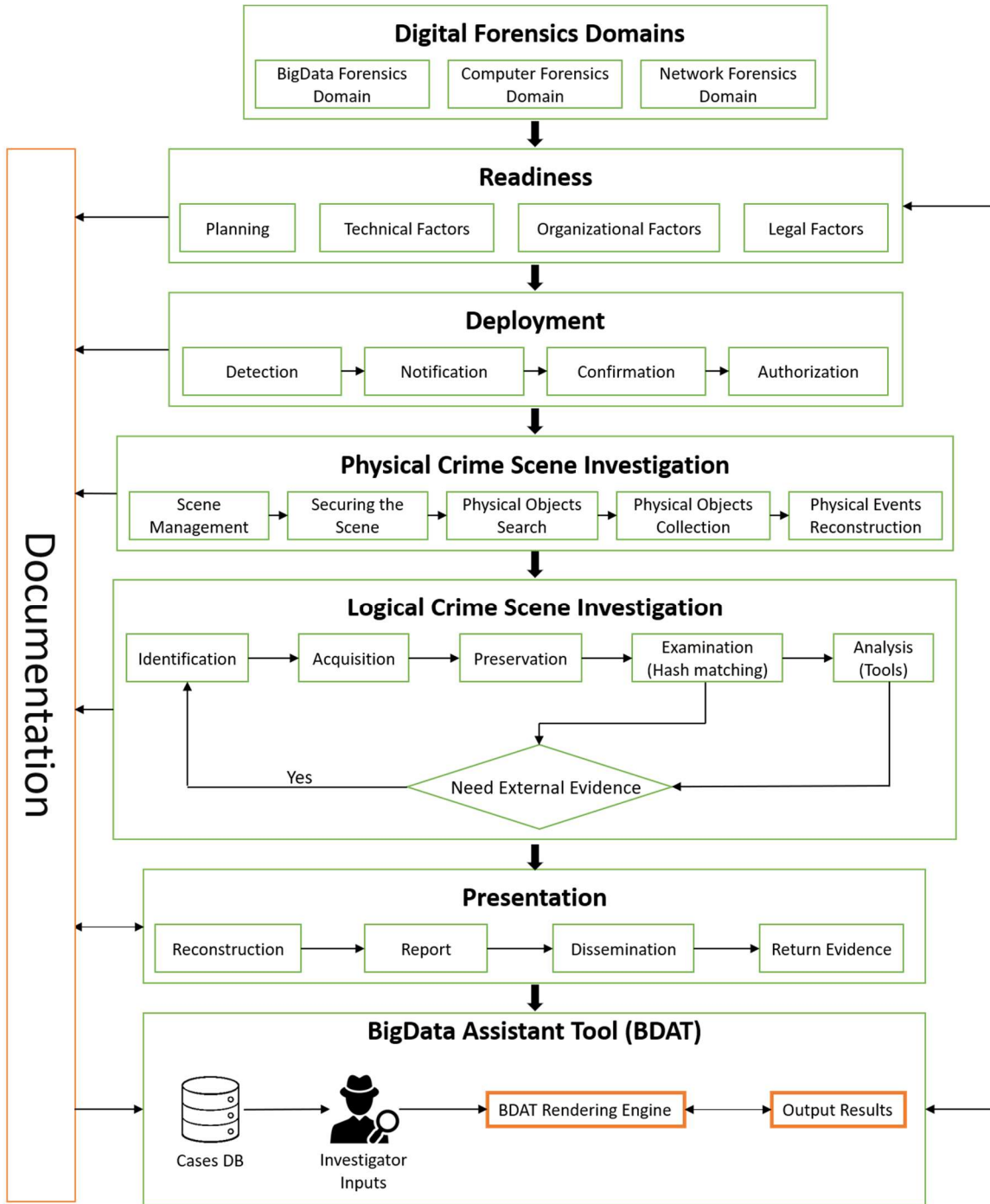


Figure 13. The proposed BigData Digital Forensics Framework

2.1 Digital Forensics Domains

depending on [70], there are 5 Digital Forensics (DF) types: Computer Forensics, Mobile Forensics, Cloud Forensics, Network Forensics, and IoT Forensics.

The nature of the problem dictates forensics domains be adapted. As computer and network domains are essential, a new BigData forensics domain is suggested. Computer and network domains constitute the incubating environment of the BigData system therefore their presence is necessary to build the BigData system on OS and to make interconnections between clusters.

2.1.1. BigData as a forensics domain

In this research, we assume that the BigData system constitutes its domain for criminal investigation within our framework. This assumption is made in order to facilitate the work of investigators and to treat the BigData system as a distinct domain, similar to the domains of digital forensics (DF). The three DF domains relevant to our framework are computer forensics, network forensics, and BigData forensics. Figure 14 illustrates the key artifacts that can be obtained from the BigData forensics domain about the BigData system.

2.1.2. The 3 Domains Artifacts

In this research, we collected all the artifacts through which evidence can be obtained from each domain on the Hadoop system, in the computer, as big data systems are installed on computers or servers, the important artifacts are the Hard Drive and Ram, and the BigData The important

Artifacts are the Log files which are FSIimage and Editlogs as well as the BigData filesystem itself, and finally the Network Domain, each of the traffic, ports, ip addresses, and cluster id.

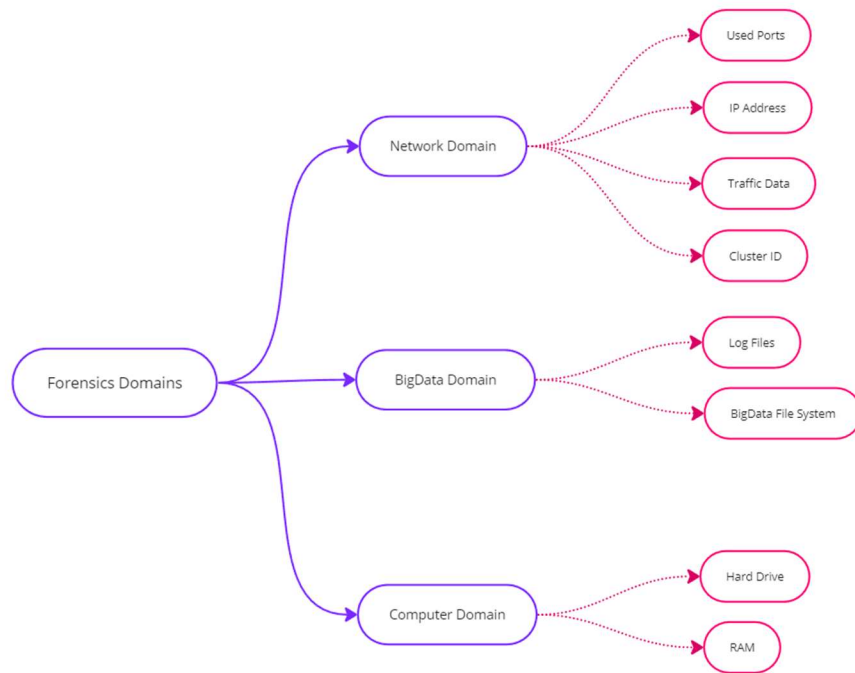


Figure14 . Forensics Domains

2.2 Readiness

After establishing the forensics domains for different forensic investigation methods of BigData systems, the framework moves on to the readiness stage. This stage is focused on preparing for and planning the investigation and includes four necessary operations: planning, addressing technical factors, addressing legal factors, and addressing organizational factors. The readiness stage is crucial for ensuring that the investigation is properly organized and has the necessary resources and support to be conducted effectively (see Figure 15).

The documentation process begins at this point and continues throughout all subsequent phases of the investigation. This process involves recording details such as the timing, location, and any processes used for each event that occurs. This documentation is important for maintaining a thorough and accurate record of the investigation and can serve as evidence if the findings of the investigation are called into question.

2.2.1. Planning:

Planning on how to begin the criminal investigation process, how to implement safeguards and measures, when to implement them, and where to implement them while taking into account the gravity of the crime and its kind, is done in this phase, whether it was physical, digital, or both. Additionally, it is essential to take the necessary precautions to secure the evidence and ensure its admissibility in court. For instance, the Cyber Security Department's office Before creating the best defensive plan for a point in the BigData system, a firm first determines who will do it and how, whether there will be losses, how to reduce them, and how proof will be acquired.

2.2.2. Technical Factors

This process covers the technological, technical, and practical measures that the team or company must consider in the criminal investigation process. it includes the infrastructure of the BigData system, its composition, the tools that must be used during the criminal investigation process, the security standards on the system, and the authorities that must be considered. Along with training

staff members on how to use the tools and handle them, is crucial. For example, the information security manager in the company or organization determines the programs that will be used during the digital forensic investigation process, from the moment evidence is collected to its analysis, and then the results are presented.

2.2.3. Organizational Factors

This process is concerned with managing resources (employees and their respective roles). This includes understanding the team, managing expertise in the organization, and distributing elements appropriately. This occurs before and during attacks on the system and pre-preparation, developing plans for protection methods and response methods. Moreover, setting company policies and principles for dealing within its scope.

2.2.4. Legal Factors

Attention must be given to the requirement to lawfully gather evidence throughout the investigative process for it to be accepted by the courts. In order to protect the integrity of the original evidence and prevent tampering with it, it must be removed, kept in particular ways, and backup copies are made. Only then will it be possible to construct and anticipate the subsequent chain of events once it has been examined. In addition, this process includes the service provider because it may help in the process of forensic investigation by analyzing network flow and IP addresses, as well as measuring the boundaries of areas whether they cover the local side or outside the country.

For example, if the original evidence was tampered with or the digital forensic investigation process was conducted during a digital forensic investigation process, likely, it will not be accepted because the evidence will be subject to change, and this may also harm the data of the institution from which the evidence was collected. Taking into account the country that will apply justice and understanding its laws, it is possible that a country may differ in the policy of accepting digital evidence from others.



Figure 15. Readiness

2.3 Deployment

3.5.1. Detection

Detection in readiness refers to the process of identifying and detecting potential threats or vulnerabilities in a system or network. In digital forensics, this phase plays a crucial role in assessing the security posture and identifying potential risks within a system or network. By conducting thorough assessments and analysis, digital forensic investigators can uncover weaknesses and potential threats that could compromise the integrity, confidentiality, or availability of data or systems.

During the threat and vulnerability identification process, several techniques and activities are typically employed, including:

1. **Risk Assessment:** Evaluating the likelihood and potential impact of identified threats to determine their level of risk. This involves considering factors such as the vulnerability of systems, potential attackers, and the potential consequences of an attack.
2. **Penetration Testing:** Simulating real-world attacks to identify vulnerabilities and weaknesses in a controlled environment. This involves attempting to exploit known vulnerabilities to gain unauthorized access or compromise systems.
3. **Security Audits:** Conducting comprehensive assessments of system configurations, network infrastructure, and security policies to identify potential security gaps or non-compliance with industry standards or best practices.

4. **Vulnerability Scanning:** Using automated tools to scan systems and networks for known vulnerabilities. This involves checking for software vulnerabilities, misconfigurations, and weak security settings.
5. **Threat Intelligence:** Staying updated on the latest security threats, emerging attack techniques, and vulnerabilities relevant to the system or network under investigation. This includes monitoring security advisories, forums, and other sources of threat information.
6. **Log Analysis:** Examining system logs, network traffic logs, and other sources of data to identify suspicious or anomalous activities that may indicate the presence of a threat or vulnerability.

By effectively identifying and detecting potential threats or vulnerabilities, digital forensic investigators can take proactive measures to mitigate risks, enhance the security posture of the system or network, and gather valuable evidence for further investigation if an incident has occurred.

This is an important aspect of cybersecurity, as it allows organizations to proactively identify and mitigate risks before they can be exploited by attackers.

One example of a digital attack is a phishing attack, where an attacker uses email or other forms of communication to trick a victim into revealing sensitive information, such as login credentials or financial information. In a phishing attack, the attacker may send an email that appears to be from a legitimate source but is a fraudulent attempt to gain access to the victim's account. To prevent falling victim to a phishing attack, it is important to be aware of the signs of a phishing attempt, such as suspicious links or requests for personal information, and to verify the identity of the sender before responding to any requests.

3.5.2. Notification

Notification refers to the process of preparing to inform relevant parties of the progress or results of an investigation. This may involve identifying the parties that need to be notified, determining the appropriate method of notification, and developing a plan for communicating the results of the investigation.

For example, consider a cybersecurity investigation into a distributed denial of service (DDoS) attack on a company's network. To be prepared for notification during this investigation, the investigator might identify the parties that need to be informed of the progress and results of the investigation, such as the company's management, employees, customers, and any relevant regulatory agencies. The investigator might also determine the appropriate method of notification, such as email, phone, or in-person communication, and develop a plan for communicating the results of the investigation in a timely and effective manner.

Ensuring notification readiness during an investigation can help to ensure that relevant parties are kept informed of the progress and results of the investigation, which can help to maintain transparency and trust in the process. This can be especially important in cases where the outcome of the investigation may have significant consequences or impacts, such as criminal or civil proceedings.

3.5.3. Confirmation

Confirmation readiness in the context of an investigation refers to the process of gathering and organizing evidence to confirm or establish the facts of a case. This may involve collecting and

analyzing various types of evidence, such as documents, physical evidence, witness statements, and other information relevant to the investigation.

For example, consider a cybersecurity investigation into a phishing attack that targeted a company's employees. To be prepared for confirmation during this investigation, the investigator might gather and analyze evidence such as email logs, network traffic logs, and other data related to the attack. The investigator might also review the company's security protocols and policies to determine whether they were followed and to identify any potential vulnerabilities that may have been exploited by the attackers.

Ensuring confirmation readiness during an investigation can help to ensure that the investigation is thorough and accurate and that the results of the investigation are reliable and supported by strong evidence. This can be especially important in cases where the outcome of the investigation may have significant consequences or impacts, such as criminal or civil proceedings.

3.5.4. Authorization

Authorization in the context of an investigation refers to the process of obtaining permission or approval to conduct the investigation. This may involve obtaining legal authorization, such as a warrant, or obtaining approval from relevant authorities or stakeholders.

Ensuring that appropriate authorization has been obtained is an important aspect of investigation readiness. Without proper authorization, an investigation may not be considered legitimate or may not be able to proceed as planned. Obtaining authorization may also involve following certain procedures and protocols, such as providing evidence of probable cause or obtaining consent from individuals whose privacy may be affected by the investigation.

It is important for investigators to carefully consider the authorization process and ensure that all necessary approvals have been obtained before beginning an investigation. This can help to ensure that the investigation is conducted fairly and legally and that the results of the investigation are considered credible and reliable.

2.4 Physical Crime Scene Investigation (PCSI)

It is also likely that the attack on the BigData system will be an attack on the ground, such as an attempt to sabotage, or cut wires, or the possibility of entering the system site and attempting to steal or copy data, or many other possibilities.

2.4.1. Scene management

There are four main parts of managing a crime scene: managing information, managing the people working on the scene, using technology, and logistics (like getting the right supplies). If any of these parts isn't handled well, it can mess up the whole investigation. Good communication is important for all these parts to work together. Different ways of organizing the investigation, called models, can be used depending on the situation, like how many people are available and what kind of crime it is. Each model has pros and cons.

2.4.2. Securing the scene

To secure a crime scene, it is important to establish a perimeter around the area, control access to the scene, document the scene thoroughly, collect and preserve evidence, and secure the scene

until it can be released. Establishing a perimeter helps to prevent unauthorized entry while controlling access ensures that only authorized personnel is allowed to enter the scene. Documenting the scene involves taking photographs and videos, making detailed notes, and drawing diagrams or maps of the area. Collecting and preserving evidence involves carefully handling and storing physical evidence, as well as digital evidence such as computers and cell phones. Finally, the crime scene should be secured until it can be released back to the owner or designated authority. This may involve sealing the area off with tape or other barriers or having law enforcement officers or other authorized personnel stationed at the scene.

2.4.3. Physical Objects Search

In a physical crime scene investigation, it is important to carefully search for and collect any physical objects that may be evidence. This process, known as "processing the scene," is typically done by specially trained forensic investigators or crime scene technicians.

Several steps are typically followed when processing a crime scene:

- Secure the scene covered in the last step.
- Conduct a preliminary assessment: The investigator will conduct a preliminary assessment of the scene to get an overall understanding of the events that took place. This may involve talking to witnesses, examining the location and layout of the scene, and looking for any obvious signs of a crime.
- Search for and collect physical evidence: The investigator will then search the scene for any physical objects that may be evidence. This may include things like USBs, Hard Drives, CDs,

and other physical trace evidence. Care must be taken to ensure that the evidence is collected and handled properly.

- Analyze and interpret the evidence: Once the physical evidence has been collected, it is sent for analysis. The results of this analysis can be used to help reconstruct the events that took place at the crime scene and may be used as evidence in court.

Overall, the process of searching for and collecting physical evidence at a crime scene is an important part of any investigation. This process must be done carefully and thoroughly, as the physical evidence collected can play a crucial role in determining what happened and who was responsible.

Document the scene: The investigator will also document the scene by taking photographs, making sketches, and creating a detailed report of the findings. This documentation is important for creating a record of the scene and the evidence that was collected.

2.4.4. Physical Objects collection

When collecting physical evidence at a crime scene, the investigator will typically follow these steps:

- Identify the evidence: The investigator will look for any physical objects that may be evidence, such as USBs, Hard Drives, CDs, and other physical trace evidence.
- Document the evidence: Before collecting the evidence, the investigator will document its location and appearance. This may involve taking photographs, making sketches, or creating a written description of the evidence.

- **Collect the evidence:** The investigator will then carefully collect the evidence using appropriate protective gear and techniques to avoid contamination. This may involve using tweezers, envelopes, or other specialized tools to handle and preserve the evidence.
- **Label and package the evidence:** The investigator will label and package the evidence using appropriate materials to ensure that it is properly protected and preserved. This may involve using airtight containers, sealed bags, or other protective packaging.
- **Transport the evidence:** The investigator will then transport the collected evidence to make images to analyze them by digital investigators. It is important to handle the evidence carefully during transportation to avoid contamination or damage.

2.4.5. Physical crime reconstruction

Crime scene reconstruction is the process of using physical evidence, witness statements, and other information to piece together what happened at a crime scene. It is often used to help investigators better understand the events leading up to, during, and after a crime, and can help to identify potential suspects or establish a timeline of events.

Several steps are typically followed when reconstructing a crime scene:

- **Interview witnesses:** The investigator will interview any witnesses who may have seen or heard something related to the crime. These witness statements can provide valuable information about what happened and can help to fill in any gaps in the investigation.
- **Re-create the crime scene:** The investigator will then use the information gathered from the physical evidence and witness statements to re-create the crime scene as it is believed to have

occurred. This may involve creating diagrams, sketches, or computer-generated models to help visualize the scene.

- Analyze the results: The investigator will then analyze the results of the crime scene reconstruction to see if it is consistent with the physical evidence and witness statements. If there are discrepancies, the investigator may need to go back and re-examine the evidence or conduct additional interviews to try to resolve them.

Overall, reconstruction is a complex process that requires a thorough understanding of forensic science and investigative techniques.

2.5 Logical Crime Scene Investigation (LCSI)

2.5.1. Identification

The process of identification in logical crime scene investigation involves collecting and analyzing various types of information about the crime scene and the evidence found there. This may include examining digital devices for evidence of tampering or malicious activity and analyzing documents for clues about the perpetrator or the crime.

It is important to properly identify and document all of the evidence found at a crime scene to establish a clear chain of custody and ensure that the evidence is properly preserved and handled. This is essential for building a strong case against the perpetrator and ensuring that the evidence is admissible in court. The criminal investigator determines what tools and programs he will use for the criminal investigation process.

2.5.2. Acquisition

Acquisition refers to digital evidence collection according to evidence type from different domains, the evidence will be either from the networks, the hard drive, or the RAM, Network Capture, and analysis.

The process of acquisition in digital crime scene investigation typically involves several steps, including:

- Identifying the types of digital evidence present at the scene and in different domains: This may include identifying the types of digital devices or systems that were involved in the crime, as well as identifying the specific locations where these devices or systems were found.
- Establishing a chain of custody: This involves creating a record of who has had access to the evidence and when to establish the authenticity and integrity of the evidence.
- Collecting the evidence: This may involve taking physical possession of the digital devices or systems involved in the crime, as well as collecting any digital documents, files, network captures, hard drives, memory dumps, and others that may contain evidence of the crime.
- Preserving the evidence: This involves taking steps to ensure that the evidence is not altered or damaged in any way, such as by making copies of the evidence or storing it in a secure location.

Acquiring digital evidence from a crime scene is a critical step in the process of investigating a digital crime, as it allows law enforcement and forensic analysts to identify and track down the perpetrator of the crime and build a strong case against them.

2.5.3. Preservation

It must be ensured that digital evidence is properly collected, handled, and stored in a way that maintains its authenticity and integrity. This is an important step in the process of investigating a digital crime, as it ensures that the evidence is admissible in court and can be used to build a strong case against the perpetrator.

Several steps can be taken to preserve digital evidence in the course of a digital crime scene investigation:

- **Making copies of the evidence:** It is important to make copies of the evidence to preserve the original evidence while still allowing it to be analyzed.
- **Storing the evidence:** The evidence should be stored in a secure location where it will not be tampered with or damaged.
- **Handling the evidence:** The evidence should be handled with care to avoid damaging or altering it in any way.

By following these steps, law enforcement and forensic analysts can ensure that the digital evidence collected from a crime scene is properly preserved and can be used effectively in the investigation and prosecution of the crime.

2.5.4. Examination

Evidence examination refers to the process of analyzing digital evidence to identify any clues or information that may be relevant to the investigation. This will involve examining forensics

domains to identify patterns of activity, trace the actions of the perpetrator, or identify other information that may be relevant to the case.

To ensure the integrity of the evidence being presented, it is necessary to verify that the copies of the evidence are identical to the original. One way to do this digitally is through the process of hash matching, which compares the unique digital hashes of the original and copies to ensure that they match.

By examining the digital evidence collected from a crime scene, law enforcement and forensic analysts can gather valuable information about the crime and the perpetrator, and use this information to build a strong case against the perpetrator.

2.5.5. Analysis

Once all the necessary preparatory steps have been completed, the investigation process can begin. This process will involve examining the evidence within the framework of a BigData system, taking into account the specific nature of the evidence. To do this effectively, it is important to use the appropriate tools and techniques for the type of evidence being examined (as we mentioned in Chapter 1). These tools may include specialized software or other technologies that are specifically designed to handle and analyze various types of evidence, such as digital documents, audio recordings, or images. By using the right tools and following a structured framework, investigators can ensure that the evidence is properly analyzed and that the results of the investigation are reliable and accurate.

2.6 Presentation

In digital crime scene investigation, evidence is collected, processed, and analyzed to understand what happened during a cyber incident and to identify the responsible parties. One important aspect of this process is the presentation of evidence, which involves organizing and presenting the collected evidence in a way that is clear, concise, and understandable.

Several considerations must be taken into account when presenting evidence in a digital crime scene investigation. One of the most important is the chain of custody, which refers to the documentation and preservation of the evidence from the time it is collected until it is presented in court. It is important to maintain the integrity of the evidence and to be able to demonstrate that it has not been tampered with or altered in any way.

Another important consideration is the format in which the evidence is presented. In many cases, digital evidence is presented in the form of computer logs, emails, and other types of digital records. It is important to ensure that these records are presented in a way that is easy to understand and that demonstrates the relevance of the evidence to the investigation.

Finally, it is important to consider the audience when presenting evidence in a digital crime scene investigation. The evidence may need to be presented to a variety of individuals, including law enforcement agents, lawyers, judges, and jurors, who may have different levels of technical expertise and knowledge of the subject matter. It is important to present the evidence in a way that is accessible and understandable to all members of the audience.

2.6.1. Reconstruction

When presenting evidence in a crime scene investigation, it is important to reconstruct the events or actions that took place to provide a complete and accurate picture of what happened. This may involve presenting the evidence in chronological order, highlighting key points or details, and using visual aids or diagrams to help illustrate the evidence and its significance.

In physical crime scene investigation, this might involve reconstructing the sequence of events based on the positions and locations of physical objects, and other types of physical evidence. In digital crime scene investigation, it might involve reconstructing the sequence of events based on BigData framework domains such as computer logs, HDFS logs, Hard drive acquisition, network traffic, and others which are shown in figure 14.

Regardless of the type of crime scene, it is important to consider the context in which the evidence was collected and to provide any necessary background information or explanations. This can help to ensure that the evidence is presented in a way that is easy to understand and that demonstrates its relevance to the investigation.

It is important to be careful and thorough when reconstructing the evidence, as any errors or inaccuracies in the presentation could potentially undermine the credibility of the evidence and the investigation as a whole.

In our framework, all collected evidence is analyzed to reconstruct the sequence of events that led to the incident under investigation. This includes considering both physical and digital attacks and integrating them into a logical and cohesive scenario. The goal of this process is to fully understand the circumstances of the incident and how it unfolded, from all stages of the event to its final form.

By thoroughly reconstructing the scenario, investigators can gain a comprehensive understanding of what occurred and determine the most appropriate course of action.

2.6.2. Report

the report is a document that summarizes the evidence that has been collected and analyzed during the investigation. The goal of an evidence report is to provide a clear and concise overview of the evidence and its relevance to the investigation.

When preparing an evidence report, it is important to be thorough and include all relevant information. This may include details about the types of evidence that were collected, the methods used to collect and analyze the evidence, and any conclusions that were drawn based on the evidence.

In the digital crime scene investigation, the evidence report may include details about digital evidence that collected from different domains in the framework. It may also include information about the chain of custody for the digital evidence, which refers to the documentation and preservation of the evidence from the time it is collected until it is presented in court.

In the physical crime scene investigation, the evidence report may include details about physical evidence such as collected Hard drives, RAMs, USBs, and other types of physical evidence. It may also include information about the methods used to collect and analyze the physical evidence, as well as any conclusions that were drawn based on the evidence.

When presenting an evidence report in a crime scene investigation, it is important to organize the information clearly and logically and to use language that is easy to understand. It is also important

to ensure that the report is complete and accurate, as any errors or omissions could potentially undermine the credibility of the evidence and the investigation as a whole.

2.6.3. Dissemination

In both digital and physical crime scene investigation, evidence dissemination refers to the process of sharing the collected evidence and the results of the investigation with relevant parties. This may include law enforcement agencies, legal professionals, and other stakeholders who are involved in the investigation or who have an interest in the outcome.

Several considerations must be taken into account when disseminating evidence in a crime scene investigation. One of the most important is the confidentiality of the evidence and the need to protect sensitive information from unauthorized disclosure. This may involve redacting or withholding certain pieces of evidence or limiting access to certain individuals.

Another important consideration is the format in which the evidence is disseminated. In many cases, digital evidence is disseminated in the form of electronic documents or files, while physical evidence may be disseminated through photographs, diagrams, or other visual aids. It is important to ensure that the evidence is presented in a way that is clear, concise, and understandable to the intended audience.

Finally, it is important to consider the purpose of the evidence dissemination and to tailor the presentation of the evidence to meet the needs of the intended audience. For example, the evidence may need to be presented in a way that is suitable for use in a legal proceeding, or it may need to be presented in a way that is designed to inform the public about the results of the investigation.

2.6.4. Return Evidence

Evidence return refers to the process of returning the collected evidence to its rightful owner once it is no longer needed for the investigation. This may include returning physical objects to their owners, deleting digital files or documents, or otherwise disposing of the evidence appropriately.

Several considerations must be taken into account when returning evidence in a crime scene investigation. One of the most important is the chain of custody, which refers to the documentation and preservation of the evidence from the time it is collected until it is returned. It is important to maintain the integrity of the evidence and to be able to demonstrate that it has not been tampered with or altered in any way.

Another important consideration is the format in which the evidence is returned. In many cases, digital evidence is returned in the form of electronic files or documents, while physical evidence may be returned in its original form or as a photograph or other type of representation. It is important to ensure that the evidence is returned in a way that is appropriate for the intended use of the evidence.

Overall, it is important to consider the legal and ethical obligations associated with the return of evidence. In some cases, there may be legal requirements or ethical considerations that dictate how the evidence should be returned or disposed of. It is important to adhere to these requirements and considerations to ensure the integrity of the investigation and to avoid any potential legal or ethical issues.

2.7 Documentation

In our framework, the process of documenting everything that occurs during the investigation is a crucial aspect of ensuring the integrity and legal acceptability of the findings. This documentation begins at the moment of initial data collection and continues through the presentation stage. It is important for verifying the procedures followed during the investigation, and can also be used to demonstrate the validity of the findings in legal proceedings, as well as to inform management and relevant authorities. Additionally, this documentation is useful for storing incidents in the BDAT program to help investigators make informed decisions in the future. It also plays a key role in the reporting stage, as it helps to provide a clear and comprehensive understanding of the investigation's findings and recommendations.

2.8 BigData Assistant Tool (BDAT)

The documentation process is a process in which all the procedures, operations, and steps during the investigation are documented. Looking at figure 13, we see that the BDAT stage is the only stage where documentation process is used as an input. In this stage, all the different digital criminal investigation operations that the investigators have carried out, from the moment of readiness to the results, are collected and stored in the database of this tool.

The BigData Assistant Tool is a tool that helps digital forensic investigators in their digital forensic investigations on massive data by selecting appropriate tools based on previous cases. It also has the ability to create a report based on standard forms. Figure 16 illustrates the working principle of this tool.

- **The tool's working principle (see figure 16):**

To make the decision of what we do want to do in it, The tool starts with two options, the first option by entering number 1, with this option it goes to the assistant in the investigation process, but if the entered number is different, it goes to writing the report.

If the user presses number 1, the operations will be as follows:

- Step 1: Defining the type of domains to attack as input to the tool, then collecting possible evidence in the form of scenarios with the preferred tools for each scenario and the possibilities of domains. Whereas, the BigData system must be present in all possibilities because it is the target system alone or with the presence of another domain from the BDDFF
- Step 2: Based on the input on the tool if it's 1 or others, scenarios are selected and displayed based on the states stored in the Database.
- Step 3: The scenario that matches or is close to the scenarios that have been presented is chosen by investigators, and if it is not shown, it will display a message that similar scenarios are not available. if there's nothing found, then this is a new case the investigators must investigate to get the clues, and then add this scenario to the tool.
- Step 4: Processes, directions for investigators, and appropriate tools depending on previous crimes for the investigation process are presented, with an explanation of each tool and its function.

On the other hand, if the user enters any other number to enter Report Designing, the steps would be as follows:

- Step 1: The user enters all the investigation inputs, including the names of the investigators, case number, etc.
- Step 2: the tool reads the inputs and move them into generating step
- Step 3: the tool start generating the inputs and match the inputs with fields.
- Step 4: the report is created
- Step 2: The report is displayed.

Figure 17 shows the complete inputs that the user enters and the report format.

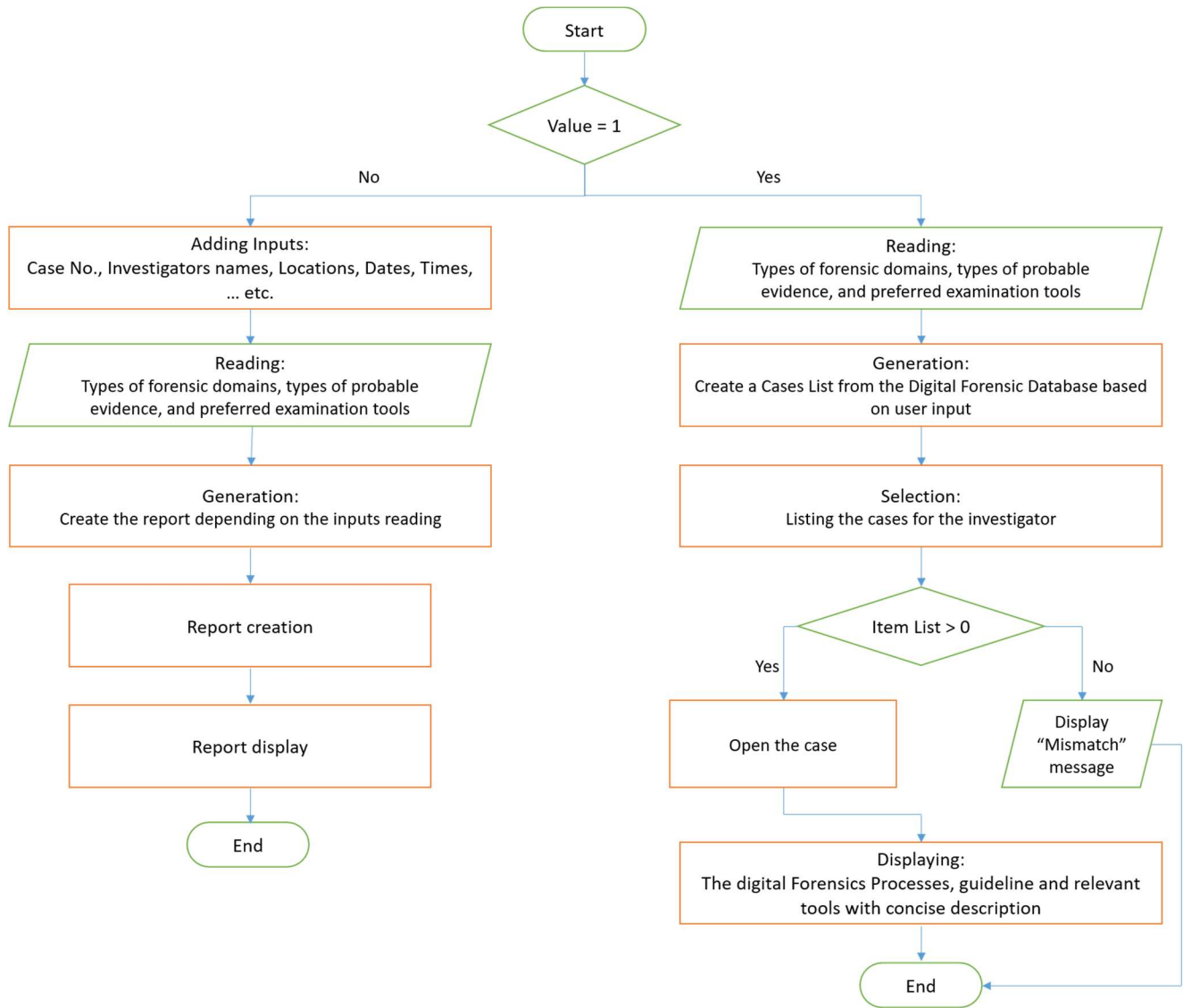


Figure 16. The workflow of the BDAT

| The proposed Digital Forensics Report | | | |
|--|---------------------|----------------------------------|--------------------|
| Case Details | | | |
| Investigator's name and Association: | | Case No.: | |
| | | Report Date: | |
| | | Report Time: | |
| Location of the Organization: | | Name of Suspect(s)/Type of Case: | |
| Devices Details | | | |
| Systems Type Serial No: | | Devices Details: | |
| | | Serial No: | |
| | | CPU: | |
| | | RAM: | |
| | | Storage: | |
| Virtual Machines Details on the organization system: | | | |
| <u>Machine Name:</u> | <u>Memory Size:</u> | <u>Storage Capacity:</u> | <u>IP Address:</u> |
| | | | |
| Framework processes | | | |
| Selected Domains: | | Type of crime: | |
| Network Domain | | Logical Crime scene only | |
| Computer Domain | | | |
| BigData Domain | | | |
| Readiness stage: | | Deployment: | |
| Logical Crime Scene Investigation: | | Other media: | |
| Findings: | | | |
| Comments: | | | |

Figure 17. BDATA proposed report

3. The proposed Responsive BigData Digital Forensics Full Framework (RBDDFF)

After designing the extended BDDFF as explained in the previous sections, work began to develop this framework more so that it would be responsive to the nature of incidents based on their type and what they target, because in some cases the speed of access to the directory is very important, and therefore dealing with the BDDFF Access to important directories may take longer, as all Processes and Stages must be checked in the framework, Therefore, it is necessary to increase the effectiveness by ignoring what is the least important (about the nature of the incident) and focusing on the relevant elements at the crucial moments. Therefore, work has been done on an advanced type of framework under the name of A Responsive BigData Digital Forensics Framework.

The idea of developing the RBDDFF depends on it being the same as the BDDFF, but some of its elements may be deleted or modified, and it all depends on the nature of the incident. In some cases, the physical crime scene in the BDDFF can be ignored if the attack is only digital, so the RBDDFF is the same as the BDDFF However, with the deletion of this section or parts thereof, the attack may be digital or physical, so physical and Logical crime scene investigations and many other scenarios are taken.

In the BDDFF, two factors are most important in determining the nature of the attack: the first is the Forensics Domains, where the nature of the domain can determine the initial goal of the nature of the attack, is it only to reduce the performance of the system, steal or delete data, or sabotage the system and other goals, and the second It is the crime type, and here the type of attack is determined, whether it is only digital or physical, or digital and physical at the same time, and

determining the crime type through which it is possible to determine how the attack should be dealt with.

2.4. Forensics Domains

Attack scenarios and digital forensic investigation depend on the investigation process of the Framework Domains because they determine the type of forensics that will be used for the BigData system in the Framework. Accordingly, several possibilities have been limited, and the nature of the attack on the system may be the same, namely:

a) Network Domain Only

So that the attack in this case may target the network of the BigData system, such as the attacker delaying the communication process between the clusters and thus transferring data at a lower speed or preventing it for some time, such as DDOS attacks.

b) Computer Domain Only

In this case, the attacker targets the operating system on which the BigData system is installed, and this case is very dangerous because if the server system is sabotaged or destroyed, it will directly affect the BigData system, such as ransom viruses. Or prevent it for some time, such as DDOS attacks

c) BigData Domain Only

The attack here targets the BigData system itself and downloads files from it, modifying or deleting them, or uploading new files to the system, such as the attacker deleting a regular file from the system and uploading another file with the same name and extension, but it is infected.

d) Network and Computer Domains

In this case, the system may target both the network and the operating system on the server or computer, as it may want to know the IP addresses of the clusters that are communicating between them at this moment, and it may want to spy on the operating system

e) Network and BigData Domains

In this case, the BigData system is accessed through the Network for specific purposes, and therefore both are checked

f) Computer and BigData Domains

The attack targets the BigData system itself with sabotage or any process, or it may target the operating system and thus affect the BigData system

These attacks can be arranged and described in a 3×3 matrix as follows including network domain.

(To note that: N means Network Domain, C is Computer Domain, and BD means BigData Domain)

Table 3. Forensics domains array

| Domains | N | C | BD |
|---------|---|---|----|
| N | 1 | 4 | 5 |
| C | 4 | 2 | 6 |
| BD | 5 | 6 | 3 |

2.5. Crime Type

From the table 2 we see that the probability of the attack being on the Network Domain alone or through it is $1/9$, the probability of the attack being on the Computer Domain alone is $1/9$, and the probability of the attack being on the BigData system only is $1/9$.

When more than one Domain is exploited, the probability of attack increases, and we see this clearly when Computer and Network Domains are exploited together, BigData and Computer Domains together, and BigData and Computer Domains, as their probabilities are equal, which is $2/9$.

The attacks on digital systems differ based on the nature of the attack they are exposed to, in the Full Framework shows both the Physical and Logical crime scenes stages and processes, but the attacks are not always like this, but in some cases, the criminal investigator analyzes the physical side And the attack did not target him, such as an attack using the network and system powers

from afar, and the Physical analysis means an important time delay during the investigation process.

In the RBDDFF, this problem was solved, and the types of crimes were classified in terms of the presence of the attacker to the following:

1. Physical Crime Attacks only
2. Logical Crime Attacks only
3. Physical and Logical Crime Attacks

And in the order and description of the crimes in a 3×3 matrix are as follows:

(To note that: P means physical, D is digital, and PD means physical and digital)

Table 4. Crimes Types matrix of possibilities

| Domains | P | D | PD |
|---------|---|---|----|
| P | 1 | 3 | 3 |
| D | 3 | 2 | 3 |
| PD | 3 | 3 | 3 |

2.6. RBDDFF examples

Figures 18-20 are RBDDFFs for hypothetical possible scenarios on the BigData system, through which it is possible to see how the BDDFF changed and became its new form in the RBDDFF based on the type and targeting of the attack, where the Domain and Crime Type were affected in this incident.

In Figure 18, we see that the criminal investigation process targeted the Network Domain or was used for the attack because the effect was exploiting Network Traffic and thus may affect the amount and speed of response between the Clusters and the speed of data transfer and that the attack also targeted the digital side only, such as DOS or DDOS attacks.

In Figure 19, we see that the criminal investigation process targeted or exploited both the Computer and BigData Domains, because the attack either affected the file system on which the BigData system is installed, or other possibilities, and that the attack targeted the digital side only.

In Figure 20, we see that the criminal investigation process targeted all domains on both the physical side and digital because the attack is likely to target the entire system and possibly complete or partial destruction or sabotage, etc. Therefore, the form of the RBDDFF is the same as the form of the BDDFF.

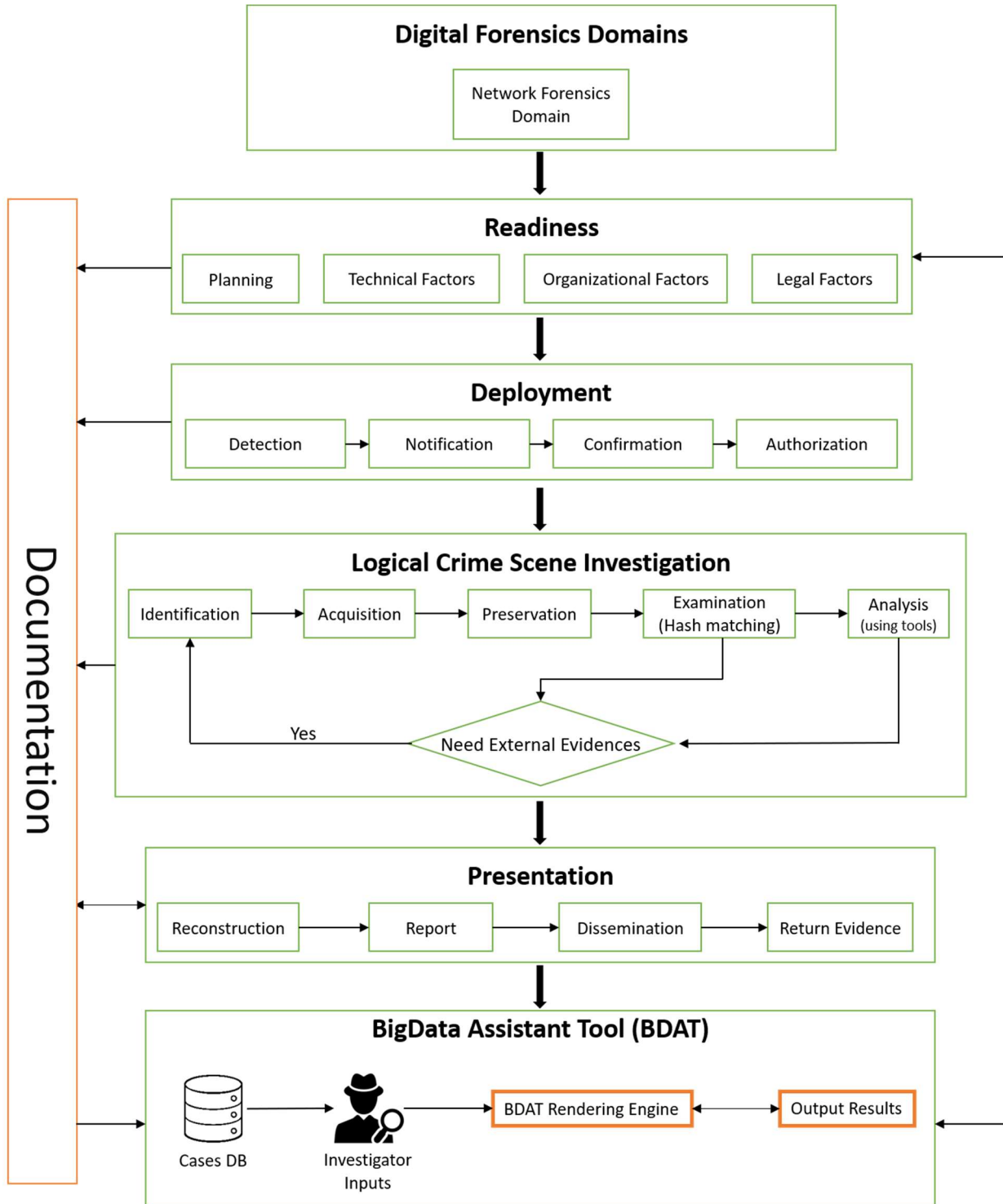


Figure 18. RBDDFF of Attacking BigData system Networking Digitally

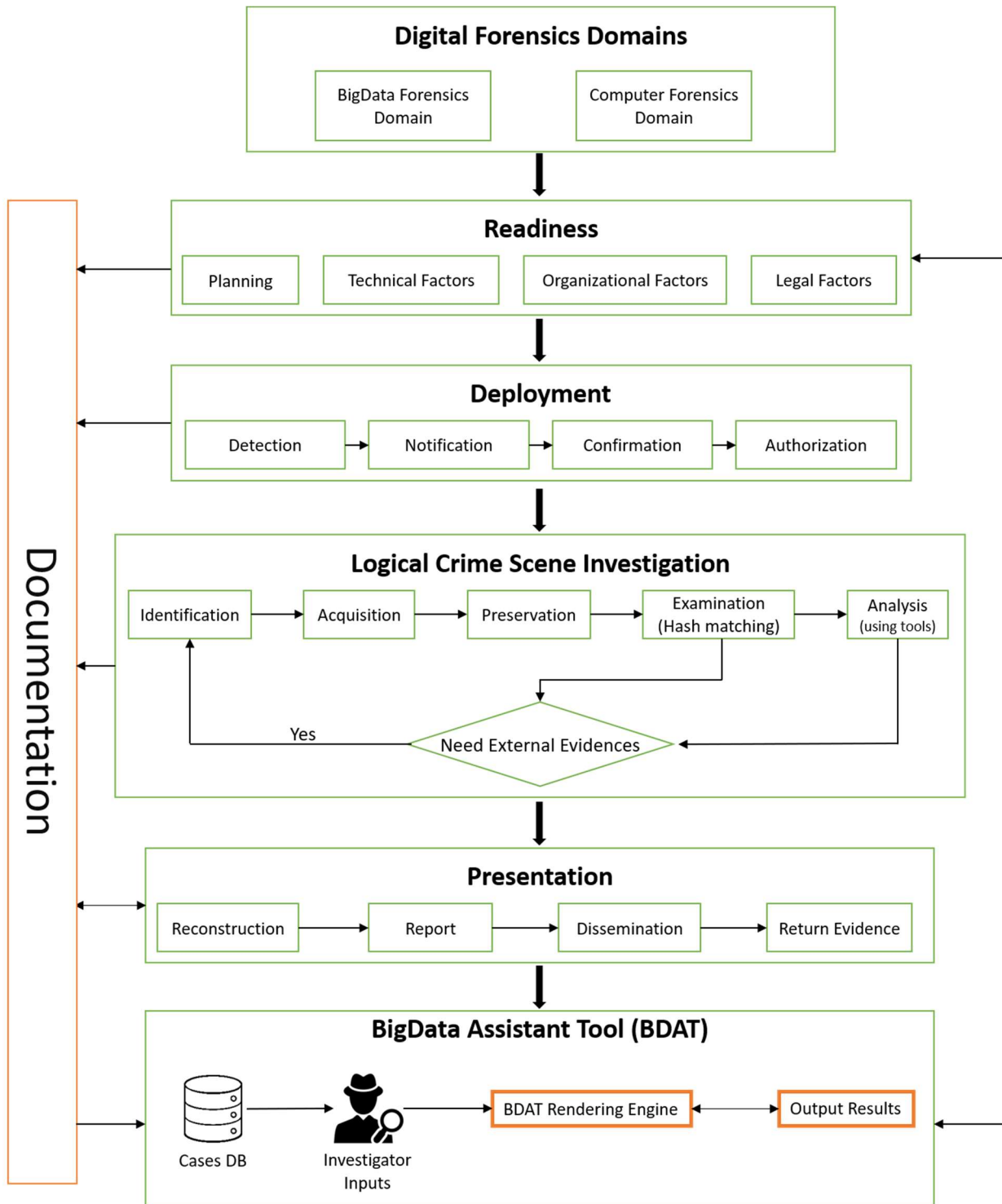


Figure 19. RBDDFF of a BigData and computer attack only

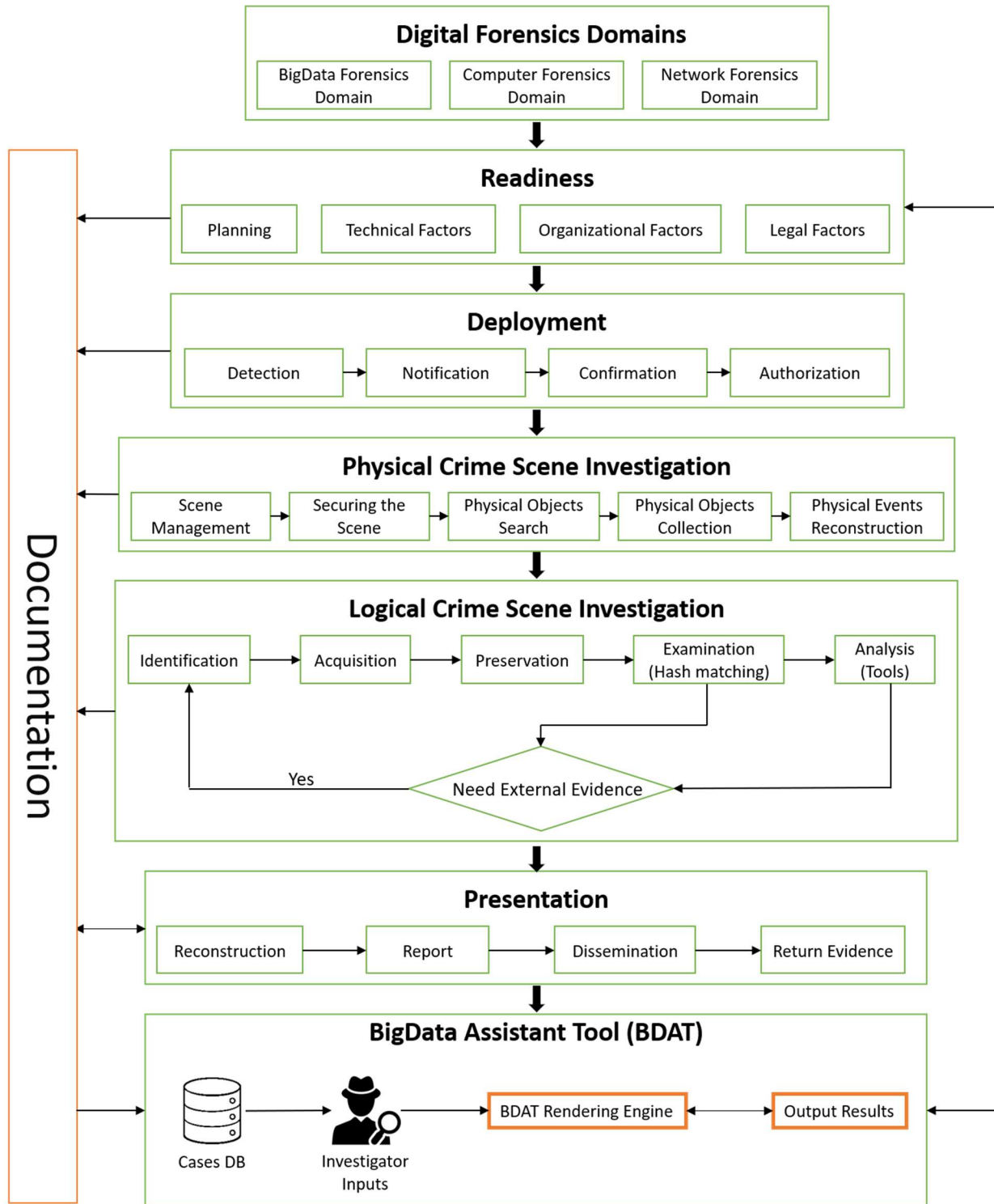


Figure20 . RBDDFF of Attacking All Domains digitally.

4. The Chapter Summary

In this research, the BigData Digital Forensics Framework BDDFF was designed and all its stages were reviewed as to how they cover the domains of the BigData system related to it and the possibilities of their occurrence, and the Crime Types were explained and classified, then this framework was further developed to be responsive to incidents to increase The speed of the criminal investigation process by defining the domains and defining the crime types, as by identifying the elements, the time required for the incident response and evidence collection process is reduced. In addition to explaining a tool that helps the digital forensic investigator in the BigData system decide if it is found in the database and displays the tools necessary for this type of investigation, the tool can also be added to cover more scenarios through the Documentation process.

The next section presents a case with a practical application of the RBDDFF for a fictitious company so that during this scenario its framework will be developed and then explain each stage with the changes that will occur on the overall BDDFF.

Chapter 4: Case, Experiments, and Results

1. Overview

After designing the BDDFF in the previous chapter, and then developing it into RBDDFF, the testing phase comes. In this chapter, a hypothetical attack is built on the BigData system of a fictitious company named Alpha, built using the Hadoop Distributed File System for test. After that, the RBDDFF is applied to this case to test, taking into account that all processes of the framework are applied to it. In addition, all Artifacts are extracted after applying the framework. Moreover, a report is displayed based on Standards Forms to be legally. Finally, the results are discussed.

2. The Company Attack Case

2.1. The company's BigData system and the physical form of the system

A virtual testing company called Alpha works in the field of selling mobile phones and stores their phone numbers, IMEI, accountants, and etc. The company was opened recently and is located in the State of Palestine. It stores its private data on the Hadoop BigData system virtually on computer.

The company's storage system is based on a computer, where the computer has a Virtual Machine installed, inside the Virtual Machine there are 3 Virtual Machines connected with it by a virtual

switch that can access each other if a ping is made between them, each of the VM's has a storage space 20 GB and 2 GB of RAM, as well as the Centos 7 system, installed.

VM 2 and VM3 are Clusters in which data is stored and responsible for them is the Master Node in VM1, which communicates with the Master node and is managed by it, the Blocks are distributed and replications are made through it as shown in figure 21.

The IP addresses of the devices with their names and the name of each host as in the table:

Table 5. Virtual machines on the system

| VM/IP | VM Name | Host | IP address |
|-------|------------|------------|--------------------|
| VM 1 | Masternode | Masternode | 192.168.160.157/24 |
| VM 2 | Cluster2 | Masternode | 192.168.160.158/24 |
| VM 3 | Cluster3 | Masternode | 192.168.160.159/24 |

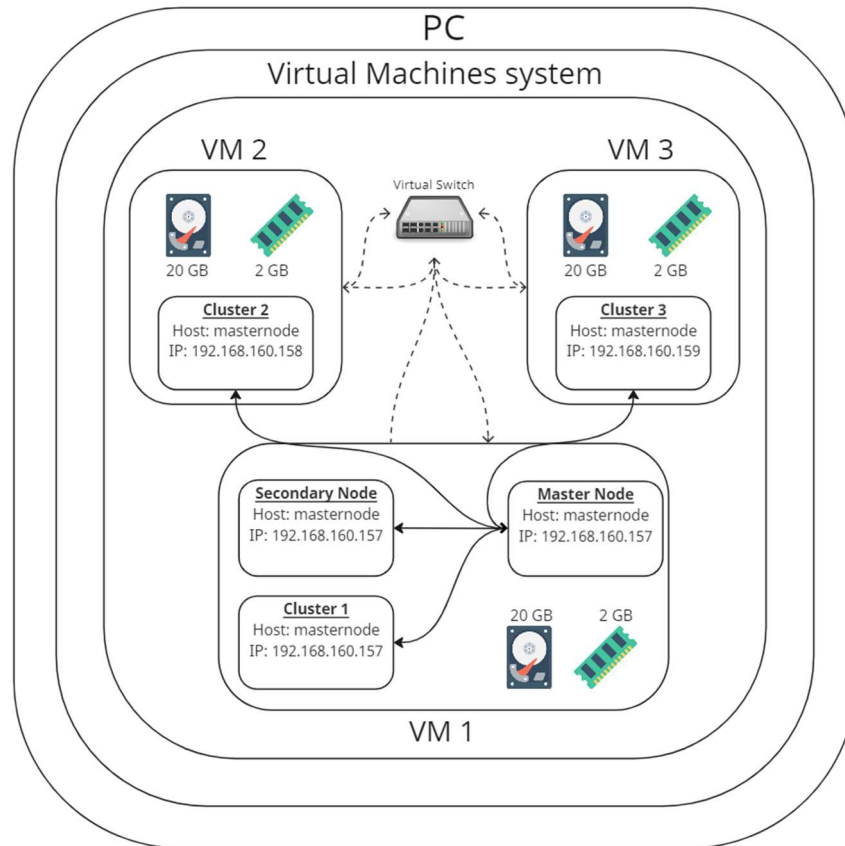


Figure 21. Physical Company System

2.2. Logical BigData system of the company

As shown in figure 22, by analyzing the system in digital terms:

- First: the VM1

This Virtual Machine contains 3 nodes of the BigData system, namely: the Master Node, which is responsible for all operations in the system and management, and the Secondary Node, in which a copy of the FSimage is taken, and in the event of a malfunction in the Master Node, it can receive the Secondary Node Work or retrieve the last backup version of it, and the last one is Cluster 1, in

which the system data is stored in the form of blocks based on the nature of the system work and the work of replications on it.

The master node also contains EditLogs files, which are files that record all operations that take place on the system, including uploading files, deleting, adding, or modifying them.

- Second: VMs 2 and 3

They are also Data nodes in which the company data is stored, and in this system of the company there are 3 Clusters to store data on the 2 data nodes themselves in each VM, and the last one is in the masternode in VM number 1. as shown in figure 21.

- Third: The BigData system

The company uses the Hadoop system from version 2.7.1, as it is an open-source system to store their data.

- Fourth: Protection and communication between the nodes

The system contains Firewall protection from the default CentOS 7, in all data nodes, ports 8020, 9000, and 9001 have been opened, and Port No. 50010 has been opened for the Masternode, and these ports have been opened so that the Clusters can communicate with each other without restrictions or Problems.

The Hadoop system is installed on CentOS without the presence of a superuser, meaning that it works directly from any user on the nodes, so that you can communicate with each other without the presence of powers that prevent communication.

In each node, RSA key generation was made, and then each RSA key was copied for each node in each node so that an SSH connection is made between the clusters without a request from the superuser and the request every time during the connection processes and for the system to work directly in all operations.

The host: for all nodes is the masternode since during the HDFS configuration process, all clusters must be the same as the username of the VM, and the Replication Factor between the clusters should be 3.

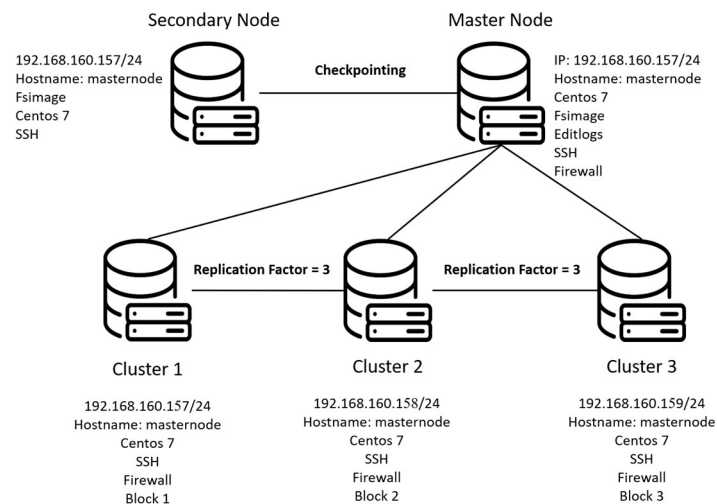


Figure 22 . Company HDFS system

2.3. Attack scenario from the Attacker's point of view

In this scenario and point of view, the attacker first exploits the Hadoop system which is open without permissions where he will upload the ransomware to be installed on a terminal victim machine that has access to the stored data to infect his machine as illustrated in figure 23.

First, the attacker accesses the masternode which accessing it means that you will control all of the file system of BigData, and browses the files of the file system and looks at the most important files that he thinks are loaded heavily, then deletes that file, and replaces it with another file with the same extension and name in its place.

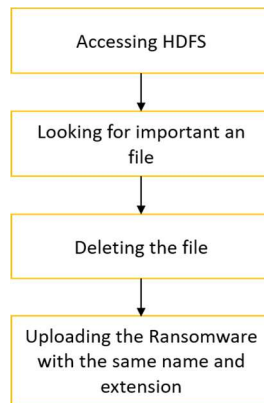


Figure 23. Attacker side

2.4. Attack scenario from the point of view of the Victim

In this scenario, the victim looks at the files, downloads the infected file without knowing that it is a virus, and then runs it to encrypt his files (depending on the nature of the scenario) (see Figure 24).

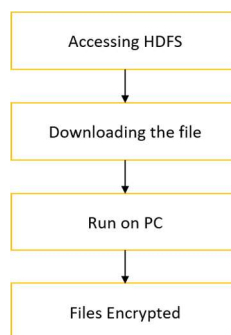


Figure 24. Victim Side

2.5. Attack scenario from the Forensics side

- The attacker gains access to the system initially, and since the HDFS system is open without permissions, he gains access to it.
- The attacker starts by looking for the important file in the system, then deletes it.
- The HDFS system records the operations that take place on the system in the EditLogs file, and the FSimage file takes a Snapshot of the system, in addition to that the Secondary NameNode takes a backup copy of each of the EditLogs and FSimage files every hour or after every 1 million transactions.
- Each of the EditLogs is analyzed to search for new activities that occurred on the HDFS system and the .vmem is analyzed after taking a copy of it and analyzing it and examining the artifacts that can be inferred by analyzing the virtual memory dump
- The attacker uploads the virus to the system with the same name and extension as the original file, which in turn is recorded in the EditLogs. During the upload process, the Network Capture is checked to see which ports have been exploited and where this file (source) was uploaded. In addition to analyzing the .VMEM to check if the commands have been loaded on the virtual memory.
- At some point, the victim accesses the HDFS system and then searches for the file, and downloads it to the device. After this process, a copy of .VMEM which is the virtual memory of the virtual machines that used as the Ram in PCs & .VMDK which is the virtual hard drive of the virtual machines, is taken for analysis.

- The victim runs the virus and encrypts it. The type of encryption depends on the type of virus. During this process, a copy of .VMDK & .VMEM is taken for analysis. The virus used in this scenario is called Revil Ransomware [154].

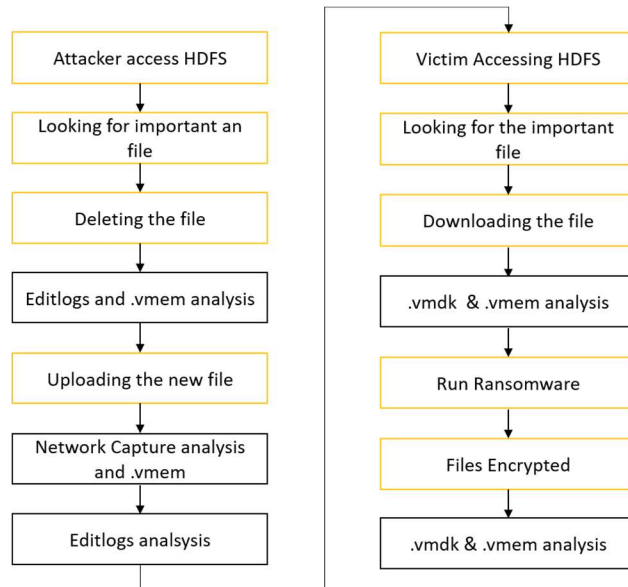


Figure 25. Scenario Forensics Side

3. Explain the scenario and the RBDDFF implementation

Before starting the investigation and work process, the criminal investigator must start building the RBDDFF, because it explains all the processes easily and begins to identify the elements that must be focused on, taking into account mainly that the criminal investigation process must be based on documentation for each step.

3.1. Select the Domains

The attacker will target the 3 domains (Network, Bigdata, and Computer) of the BDDFF in the attack because they used the network to access the system and upload files, and they also specifically targeted the BigData system by exploiting it to attack the victim. The FileSystem of operating system will also be targeted because the virus will affect the operating system and its files when it runs. The extent of the impact on the operating system and its files will depend on the nature of the virus.

Figure 26 illustrates the domains that were targeted during the attack, along with the specific elements that were targeted for each domain.

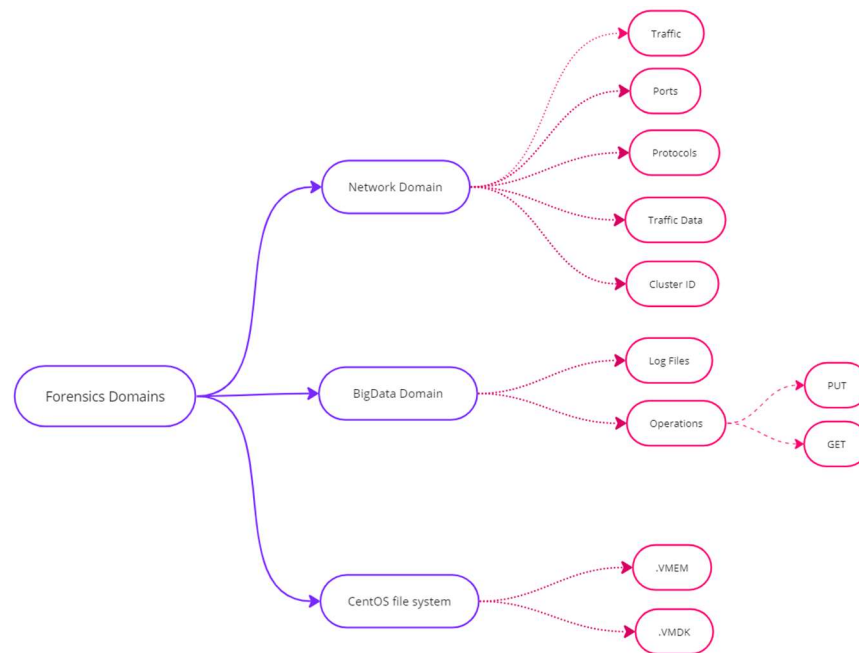


Figure 26. Case Domains possible artifacts and files to analyze

3.2. Readiness Stage

In This stage, we will focus on planning for the investigation of the attack depending on the framework. This stage includes the four necessary operations: planning, addressing technical factors, addressing legal factors, and addressing organizational factors.

The investigators started there planning by looking for any similar cases in the BDAT, if there are nothing then they have to check each domain artifacts, if any domain there hadn't effected they drop it. After that they will check which physical or digital crime scenes happened, and in this scenario only digital crime happened.

3.2.1. Planning

Responding to a ransomware attack can be a complex and time-sensitive process, so it's important to have a plan in place to minimize the impact on the organization and to ensure that it would be able to effectively investigate the attack.

The procedure is as follows:

1. The process will start by disconnecting the internet from external connections so that the hacker cannot cause any other damage to the system before anything else and isolate it from the system.
2. Change the Administrator Password
3. Report the existence of an Attack Notification

4. The Security Department will then confirm through Detection, either by checking the Network traffic or the BigData file system or any evidence in the targeted Domains.
5. Determine the type of attack, and since the attack is only digital and the framework is responsive and time is very important, the Physical crime investigation section will be removed from the Full Framework because there is no human attack.
6. Then the Domains targeted by the attacker will be determined.
7. After confirmation, a backup copy of each exploited Domain will be taken immediately for analysis to make the evidence acceptable in court and legally, and here a Bit-by-Bit image of the Hard Drive, Memory Dump, and Network Capture Backup will be taken for analysis.
8. Start the analysis process and extract the Artifacts to build the Chain of Custody based on the timeline.
9. During every process that is carried out and worked on, everything that happens and everything the department does from the moment the attack is detected until the final report is created with the evidence that occurred on the company's BigData system is documented.
10. All Artifacts from each Domain are collected and based on the timeline, they are linked together to form the final Chain of Custody.
11. After the final Chain of Custody is created, the final report will be created and submitted to the relevant authorities.
12. After completion, the original evidence is returned to the company, and then the artifacts could be extracted, copied, saved, and the chain of custody built, and ready to create the report to give to the court.

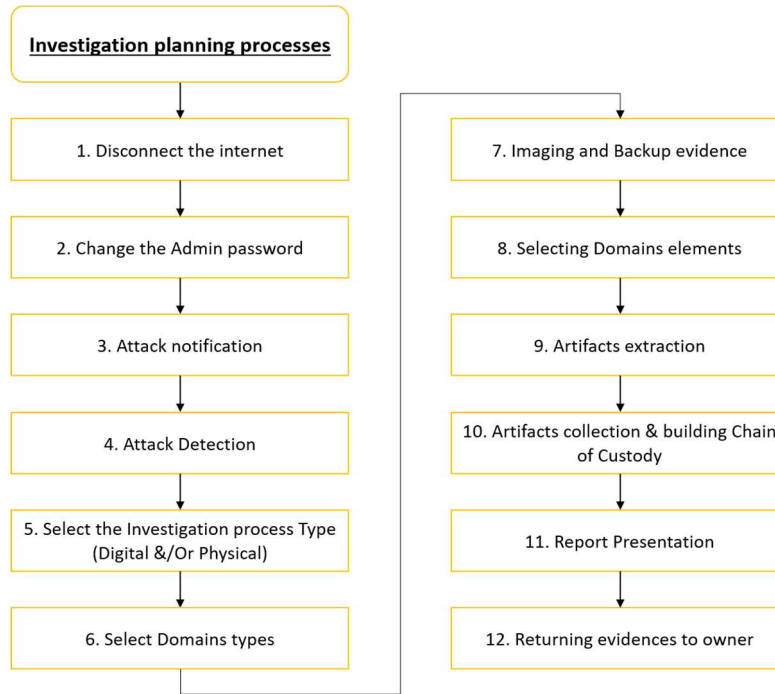


Figure27 . Planning Process

3.2.2. Technical Factors

After identifying the plan and setting the working method in the investigation process, it is necessary to identify the necessary tools to be used during this process, and for this case we will use FTK imager for imaging the .VMEM & .VMDK, HashCalc to check the hash values that they are the same, and FTK image international to analyse them. Table 5 shows the type of BigData system used in the company. The Cyber Security team at the company has been trained to deal with attacks and incidents on the system and methods of dealing with them. The team performs a Backup of the Network traffic Capture every hour, and also takes a Hard Drive bit by bit image and a Memory Dump every 24 hours at 1 am

Table 6. Case technical factors

| Technical factors | |
|--|--|
| Operating System | Linux Centos 7 |
| BigData system | Hadoop 2.7.3 |
| BigData system architecture of the company | 1 NameNode 1 Secondary NameNode 3 Data nodes Replication factor = 3 Secondary node Checkpoint |
| Forensics technologies | Network traffic analyzers .VMDK file system bit-by-bit acquisition tools .VMEM snapshots .VMEM & .VMDK analysis tools |
| BigData Security | Centos 7 Firewall |
| BigData privileges | No superuser privileges on the BigData system No SSH super user do |
| Training | The Cyber Security team is fully trained in dealing with attacks incidents |

3.2.3. Organizational Factors

The Alpha company has established a department specifically to address issues related to cyber security. This department is responsible for managing and protecting the company's online assets and information systems from various types of cyber threats, such as hacking, malware, phishing attacks, and other forms of digital espionage.

To ensure that the department can effectively carry out its tasks, it is divided into several smaller sections or sub-departments. These sections may be specialized in different areas of cyber security. Each section would have its own set of responsibilities and functions and would work together as part of the larger cybersecurity department to protect the company's digital assets:

1. The first department is the Notification section, where monitoring for any anticipated attacks on the systems is carried out.
2. The Detection department, which begins initial investigations into the attack, where it started, how, and when.
3. The Backup and Analysis department, which takes backup copies of the systems and Network Traffics, as well as extracts evidence from Images and Backups.
4. The Documentation department, which documents all incidents on the system in detail, from the moment of the attack until the Report is written.
5. The Report writing department, collects all the extracted evidence and rearranges the scenarios that occurred on the system, taking all Documentation records to build the Chain of Custody.
6. The Law department, which legally prosecutes based on the Report and evidence, and is a team of specialized lawyers in the field of Cyber Security who may also act as consultants.

3.2.4. Legal Factors

The cybersecurity team is taking backup copies of the affected systems or data, based on the specific domain (area or area of expertise) that was targeted in the attack. These backup copies are being made to analyze the attack and gather evidence that may be used in legal proceedings. The

phrase "legally acceptable" likely means that the evidence will meet the standards and requirements for being admissible in a court of law.

3.3. Deployment

In the event of a Hadoop attack on a company, several steps should be taken to detect, confirm, and notify relevant parties of the attack.

Detection: The first step in responding to a Hadoop attack is to detect that an attack is occurring. This can be done through a variety of means, including monitoring network traffic for unusual patterns, reviewing system logs for suspicious activity, and using security tools to scan for vulnerabilities. And for the company scenario this happened when one of the cyber security team noticed that through the FSimage and Editlogs log files.

Notification: Once an attack has been detected, it is important to notify relevant parties, such as the IT security team, management, and potentially law enforcement if the attack is severe enough. This can be done through a variety of means, including email, phone, or in-person communication. And here the employee told the cybersecurity team about.

Confirmation: Once an attack has been detected, the relevant parties need to confirm that it is indeed a Hadoop attack and not a false positive. This can be done by analyzing the attack vector and any available evidence, such as malware samples or network traffic logs.

Authorization: the next step is to determine the appropriate response. This has been described in the planning step. The specific response depends on the nature and severity of the attack, as well as the specific policies and procedures in place at the company.

3.4. Physical Crime Scene Investigation (PCSI)

Alpha Company experienced a cyberattack on its BigData system. In response, the company decided to use the RBDDFF to address the incident. the RBDDFF prioritizes speed and efficiency in its response by condensing or streamlining certain operations.

As part of this response, the physical crime scene investigation (PCSI) process has been excluded. This is because the attack was purely digital and did not involve a physical location or tangible evidence.

However, it's important to note that PCSI can still play a role in a cyberattack investigation. For example, if the attack involved the theft of physical devices or the installation of malicious hardware, physical PCSI could be used to identify and collect evidence related to those aspects of the case.

3.5. Logical Crime Scene Investigation (LCSI)

The cyberattack always begins with an internet connection, where the hacker accesses the devices targeted for attack. Therefore, the digital forensic investigation starts with the process of identifying and monitoring the Network traffic and collecting all possible Artifacts from IP addresses, ports, Traffic, Data, and any other Artifacts that can be collected.

Then a backup is taken after disconnecting the internet for all the Domains that have been identified and the attack has occurred, for analysis.

In this attack, the Computer, BigData, and Network domains were targeted, and Artifacts will be collected from them during the digital forensic investigation.

3.5.1. During the digital forensic investigation of the Hadoop Domain

Many different types of investigation operations can be used to extract evidence from a BigData system. The analysis of a hard drive and memory dump is a common type of operation that involves examining the contents of the storage media and memory to find evidence of criminal activity or other wrongdoing. This type of operation can be conducted manually, or with the use of specialized software tools. The analysis of logs is another type of investigation operation that is used to extract evidence from a BigData system.

3.5.1.1. Hard Drive Imaging

Bit-by-bit hard drive imaging refers to the process of creating an exact copy of a hard drive or partition. This copy, also known as an "image," includes all of the data on the original hard drive, including the operating system, applications, and files.

To create a bit-by-bit hard drive image in our investigation, we will need to use a tool that is capable of reading and copying every single bit of data on the hard drive such as FTK Imager. This process is typically done by connecting the hard drive to a computer and using a bootable imaging tool, such as a bootable CD or USB drive, to access the hard drive and create the image.

Once the image has been created, it can be stored on an external hard drive, a network drive, or other storage media. The image can then be used to restore the hard drive or partition to its original state in the event of a disaster, such as a hardware failure or a virus attack.

To create a virtual hard drive (.VMEM) image using FTK Imager for VMware tool, we will follow these steps:

- Launch FTK Imager.
- In the "File" tab, click the "Add Evidence Item" button.
- In the "Create Disk Image" window, select the "Image File" and click "OK".
- Select the destination where you want to save the image file and enter a name for the image file in the "Evidence Item Name" field.
- Click the "Finish" button to begin the imaging process.
- In the "Create Image" tab, click the "Add" button.
- Select the Image Type (we chose E01).
- Write the "Case Number", "Evidence Number", "Unique Description", "Examiner", and "Notes" then click "Next".
- Select the destination where you want to save the image file and enter a name for the image
- Put the Image Fragment = 0.
- Click the "Finish" button to create the image

The image file that is created will contain a forensic copy of the virtual hard drive, including all of its data and metadata.

3.5.1.2. Memory Dumping

Memory dumping is the process of creating a copy of the contents of a computer's memory at a particular point in time. This copy, also known as a memory dump, can be used for a variety of purposes, such as analyzing the cause of a system crash, troubleshooting software errors, or examining malware.

To perform a memory dump in our investigation, a special utility or debugger is used to access the computer's memory and create a file containing its contents. This file can then be examined using examination tools, such as a hex editor or a memory analysis program, to look for clues about what might be causing a problem.

Memory dumping can be a useful technique for diagnosing problems with a computer, but it should be used with caution. Accessing a computer's memory directly can be risky, as it can potentially expose sensitive information or cause further damage to the system.

To create a virtual hard drive (.VMEM) image using FTK Imager for VMware tool, we will follow the same steps of Hard Drive Imaging. The outcomes of memory dumping of this case as illustrated in figure 54 shows that after investigation, the investigators were able to find the operations happened PUT,GET, also the time stamps, the permissions, the client names, the files paths, and the commands used on the system by the attacker to upload and delete.

3.5.1.3. Hashing

Image hashing is a technique for creating a digital fingerprint, or hash, of an image. This hash is a unique representation of the image's content, and it can be used to identify and locate the image, even if it has been modified or altered in some way.

Several algorithms such as SHA1 and MD5 can be used to create image hashes, each of which produces a unique hash value for a given image.

Image hashing is often used to detect copyright infringement, as well as to locate duplicate images. It can also be used to build image search engines or to identify similar images within a large dataset.

- To compare the hashes of an original hard drive with an image, we follow these steps:
- Create a hash of the original hard drive using a forensic tool such as FTK Imager.
- Create a hash of the image of the hard drive using the same forensic tool.
- Compare the two hashes to determine if they are the same. If the hashes are the same, it indicates that the image is an accurate representation of the original hard drive. If the hashes are different, it indicates that there may be differences between the original hard drive and the image.

It is important to note that comparing hashes is just one step in the process of verifying the integrity of an image. There may be other factors that contribute to the accuracy of the image, and it is important to consider these factors as well. This matching process done to the scenario for .VMEM and .VMDK files after making imaging to work on the copies, and the results were matched.

3.5.1.4. Hadoop Logs Analysis

To perform FSImage and EditLogs analysis to investigate digital crime in a Hadoop cluster, we would need to follow these general steps (the practical analysis is in the next chapter):

- Obtain copies of the FSImage and EditLogs. This typically involves accessing the NameNode, which is the component of Hadoop that manages the file system namespace and file block mapping and extracting the relevant files.
- Parse the FSImage and EditLogs. This typically involves using tools like Hadoop fsck and the Hadoop Offline Image Viewer (OIV) to extract the relevant data from the files.
- Analyze the data: this could involve looking for patterns of behavior or specific file accesses that may indicate wrongdoing, as well as comparing the data to other sources of information (such as system logs or network traffic data) to build a more complete picture of what happened.
- Present the findings: This could involve creating reports or visualizations to share our results with others, or presenting the findings to law enforcement or other authorities.
- It is worth noting that these are general steps and the specifics of how to perform FSImage and EditLogs analysis will depend on the specific tools and techniques you are using, as well as the specifics of your investigation.

Through EditLogs and FSImage we know the file access, modification, deletion, and uploading operations, if any, with the time, date, number of Replications, and the permissions on the files for each of the owners and groups. [Section 5](#) shows briefly all of the steps.

3.5.2. During the digital forensic investigation of the Computer Domain

The hard drive of a computer is a storage device that stores all of the files and data on a computer. When investigating a domain within a BigData framework, it is necessary to analyze the hard drive to gather evidence and access data stored on the computer. This can be done using various forensic tools and techniques.

The memory dump of a computer is a file that contains a snapshot of the computer's memory at a particular point in time. This can be useful in an investigation as it can provide insights into what the computer was doing at the time the memory dump was taken. To analyze a memory dump, forensic analysts can use specialized tools and techniques to extract and interpret the data contained within it.

Both the hard drive and memory dump can provide valuable evidence in an investigation and are important to analyze to fully understand what was happening on the computer at the time in question.

The analysis of both the hard disk and the Memory was explained in the previous section.

3.5.3. During the digital forensic analysis of Network Traffic

Network traffic analysis is the process of monitoring and analyzing network traffic to understand what is happening on a network and identify any issues or security threats.

Network traffic analysis can be used to gather a wide range of artifacts for BigData systems, including:

Packet data: Packets are the basic units of data transmitted over a network. Network traffic analysis can be used to capture and examine individual packets, as well as the data they contain. This can include information such as the source and destination IP addresses, port numbers, and protocol being used.

Traffic patterns: Network traffic analysis can be used to identify patterns in network traffic, such as the volume of traffic at different times of day, the types of protocols being used, and the locations of the devices generating traffic.

Wireshark is a popular tool used for network traffic analysis. Wireshark allows users to capture and analyze network traffic in real time or from a previously recorded file. It provides a range of features for filtering, searching, and analyzing network traffic data, making it a powerful tool for understanding the traffic patterns on a network. To perform network traffic analysis with Wireshark, users can follow these steps:

- **Capture network traffic:** To begin, users can use Wireshark to capture network traffic in real-time by selecting the interface they want to capture from and clicking the "Start" button. Alternatively, users can open a previously recorded traffic file in Wireshark for analysis.
- **Filter traffic:** Wireshark provides a range of options for filtering traffic, such as by protocol, source or destination IP address, or port number. This can help users focus on specific traffic of interest.
- **Analyze traffic:** Once traffic has been captured and filtered, users can use Wireshark's various tools and features to analyze the traffic data. This can include examining the details of individual packets, using the "Follow Stream" feature to view traffic streams in their entirety, and using the "Statistics" menu to generate graphs and reports.

3.6. Artifacts expected to be obtained from analysis processes

By analyzing the Domains that the investigators will do in the company after writing the RBDDFF, Figure 28 displays the Artifacts expected to be found during the investigation process. In computer domain and by analysing the .VMDK image of the company the results were the paths of the files and the ransomware file before and after running, while .VMEM they found the operations happened PUT,GET, also the time stamps, the permissions, the client names, the files paths, and the commands used on the system by the attacker to upload and delete.

In Network domain, the cluster ID, ports used, the operations, the stored Hadoop blocks, time stamps, traffic data, and protocols used the extracted.

Lastly, in the BigData domain they analysed the two log files:

- In Editlogs: OPCODE, Transaction ID, SRC, Client Machine, time stamps, client machine, block size, permissions, INODE ID, and the files paths on Hadoop.
- In FSimage: Files paths on the Hadoop and Metadata.s

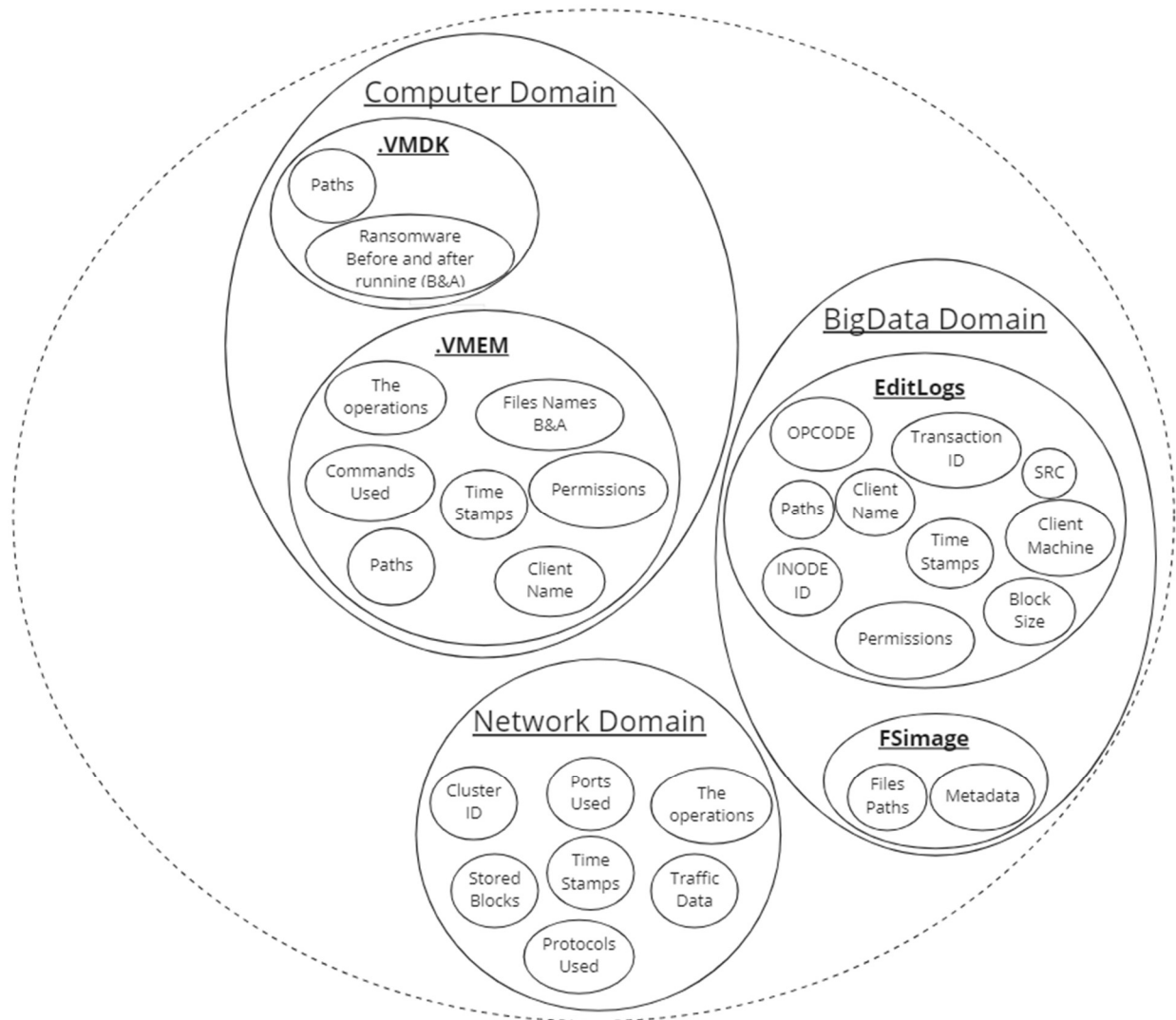


Figure28 . The expected artifacts outcomes

3.7. Presentation

After legally analyzing the evidence through the use of legal programs, taking backups and imaging, and following proper procedures to ensure acceptance in court, the process of collecting evidence from multiple locations to link events and understand them, and writing the Chain of

Custody, then writing the Report to present to the judge is carried out. The process of collecting is based on the chronological sequence of events:

- The evidence is ordered based on time regardless of the location of the evidence, whether it is. VMDK or. VMEM or Network traffic.
- Then the process of building the Chain of Custody and attempting to understand how the attack occurred, where it came from, and who carried it out from the moment the attack began until it was discovered is begun.
- The final report is written after all the complete procedures based on the Documentation and the results of the previous step.
- The original evidence is returned to the company after completing it, once all the steps and procedures and collection of all the evidence are confirmed.

figure 146 presents the final report of the company forensics investigation in which will be shown in the court to take the legal concerns.

4. The RBDDFF of The Case:

Based on all the previous steps that have been reviewed and detailed, we can now build the Responsive Framework for this attack on the BigData system for the company case. The following figure 29 illustrates the final form of the Framework, and the following section explains the practical and applied aspects during the analysis process.

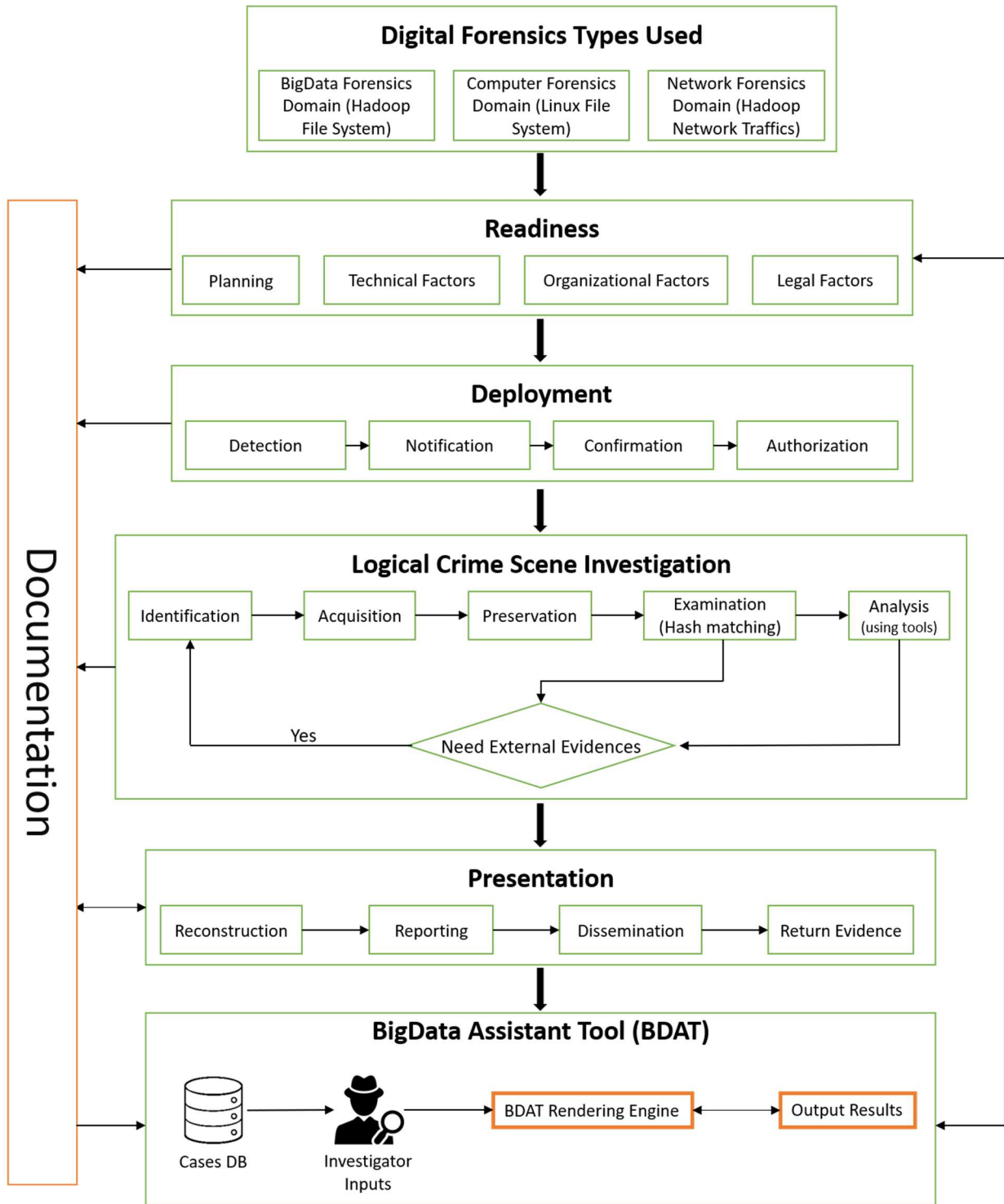


Figure29 . The case RBDDFF

5. Explanation of The Scenario Practically

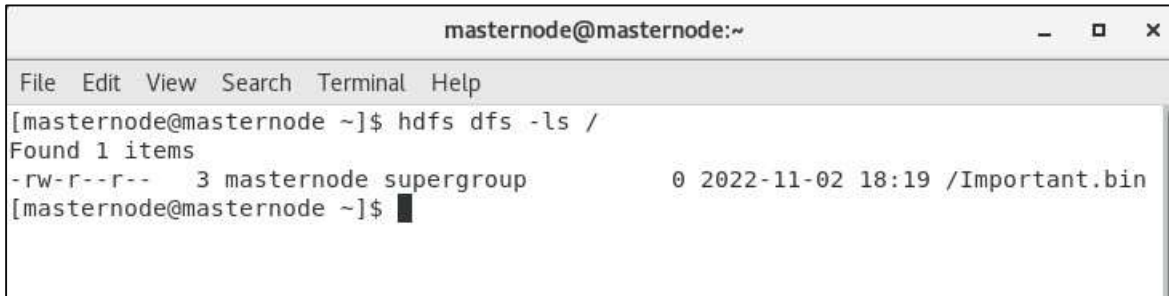
5.1. Digital forensics before the process of deleting the file from the system

figure 30 represents the file on the system through the command "hdfs dfs -ls /", which was uploaded at 18:19 on the system on 2/11/2022, and the file has permission for the owner to read and written, for the group, it is read and for others, it is read We find that the Owner is the masternode and the Group is the supergroup, and the system contains only the Important.bin file in the root directly.

Figure 31 By parsing the FSimage file with the command (the output is saved in an XML file):

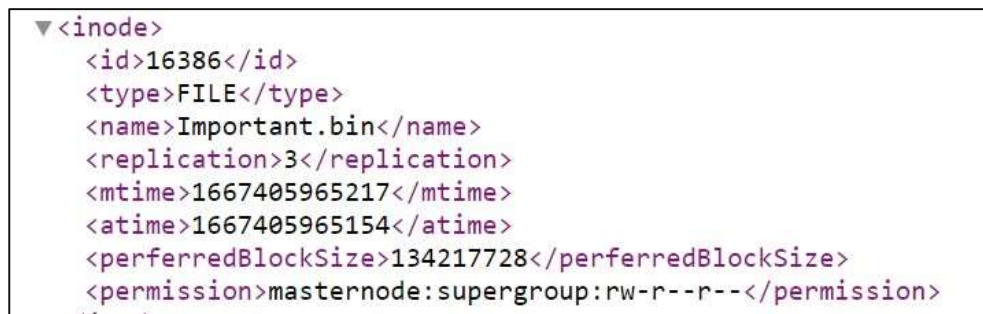
```
$hdfs oiv -p XML -i ~/path of HDFS saving files of the current master node/edits_ number /fsimage_ FSimage file number -o "Path to save"/filename.xml". We see that the ID number is 16386 and the type of data stored during this process is of file type and its name is Important.bin, and the replication occurred for 3 clusters, As for the Access time, it was in 1667405965154 and the modification time is 16674059665217, and they are two close times, and after converting the two times to Human Readable using UNIX time stamp online converter [155].
```

From the Unix system, we see that the hour and minute are the same and that the difference between them is very small, so it is very likely that this is the time the file was created (or uploaded) on the system. We also see that the Owner is the masternode and the Group is the supergroup, and as for the Permissions It is Read and Write for the Owner and Read for both the Group and Others



```
masternode@masternode:~
File Edit View Search Terminal Help
[masternode@masternode ~]$ hdfs dfs -ls /
Found 1 items
-rw-r--r--   3 masternode supergroup          0 2022-11-02 18:19 /Important.bin
[masternode@masternode ~]$
```

Figure30 . listing files on HDFS



```
▼<inode>
  <id>16386</id>
  <type>FILE</type>
  <name>Important.bin</name>
  <replication>3</replication>
  <mtime>1667405965217</mtime>
  <atime>1667405965154</atime>
  <perferredBlockSize>134217728</perferredBlockSize>
  <permission>masternode:supergroup:rw-r--r--</permission>
```

Figure31 . FSImage for HDFS

figure 32 and by analyzing the EditLogs file with the command (the output is saved in a text file):

“ hdfs oev -i ~/”path of HDFS saving files of the current master node”/edits_” number” -o ~/”storing path.txt””, we see that the file Important.bin has been uploaded to HDFS, and this can be seen through the OPCODE, which is OP_ADD, which means add, and this proves that the original file was uploaded at the same time that was mentioned earlier

```

<OPCODE>OP_ADD</OPCODE>
<DATA>
  <TXID>2</TXID>
  <LENGTH>0</LENGTH>
  <INODEID>16386</INODEID>
  <PATH>/Important.bin._COPYING_</PATH>
  <REPLICATION>3</REPLICATION>
  <MTIME>1667405965154</MTIME>
  <ATIME>1667405965154</ATIME>
  <BLOCKSIZE>134217728</BLOCKSIZE>
  <CLIENT_NAME>DFSCClient_NONMAPREDUCE_1461139119_1</CLIENT_NAME>
  <CLIENT_MACHINE>192.168.160.157</CLIENT_MACHINE>
  <OVERWRITE>true</OVERWRITE>
  <PERMISSION_STATUS>
    <USERNAME>masternode</USERNAME>
    <GROUPNAME>supergroup</GROUPNAME>
    <MODE>420</MODE>
  </PERMISSION_STATUS>

```

Figure 32. EditLogs for HDFS

The Current Epoch Unix Timestamp

Enter a Timestamp

1667405965217

Supports Unix timestamps in seconds, milliseconds, microseconds and nanoseconds.

1668878657
SECONDS SINCE JAN 01 1970. (UTC)
7:25:02 PM

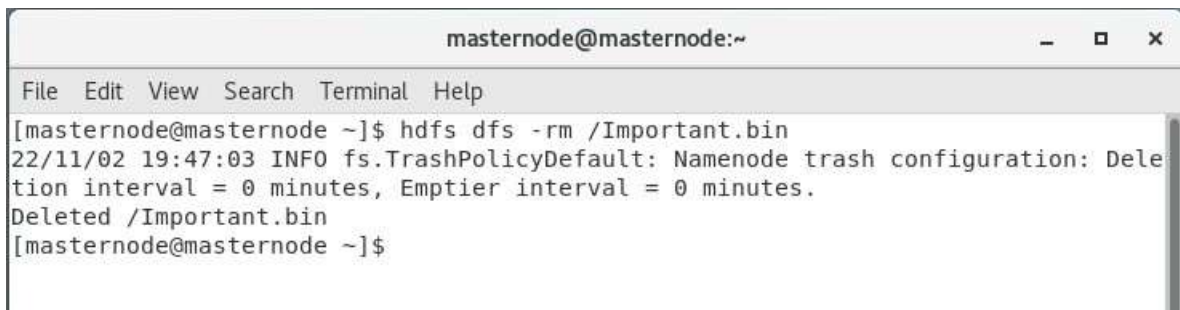
Convert → Copy

| | |
|-----------------------|--|
| Format | Milliseconds (1/1,000 second) |
| GMT | Wed Nov 02 2022 16:19:25 GMT+0000 |
| Your Time Zone | Wed Nov 02 2022 18:19:25 GMT+0200 (Israel Standard Time) |

Figure 33. Timing (Human readable)

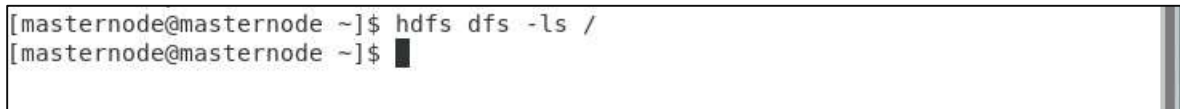
5.2. Digital forensics during the process of deleting the file from the system

Using the HDFS commands, the command "hdfs dfs -rm /Important.bin" was used by the attacker at 19:47:03 on 2/11/2022 as shown in figure 34 and using the command "hdfs dfs -ls /" was Check that the file has been deleted from the root directory.



```
masternode@masternode:~  
File Edit View Search Terminal Help  
[masternode@masternode ~]$ hdfs dfs -rm /Important.bin  
22/11/02 19:47:03 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.  
Deleted /Important.bin  
[masternode@masternode ~]$
```

Figure34 . Removing File from HDFS



```
[masternode@masternode ~]$ hdfs dfs -ls /  
[masternode@masternode ~]$
```

Figure 35. Listing files on HDFS

Through figure 36, we see that the Operation Code is a process of deletion type in the system, the number of the transaction on the system is 40, the path is in the root directly, the file name is Important.bin, and the time of deletion is 1667411223378 and with human readable (see image 20) it is at 19:47:03 19, and the Replication Client ID, as shown in the picture

```

<RECORD>
  <OPCODE>OP_DELETE</OPCODE>
  <DATA>
    <TXID>40</TXID>
    <LENGTH>0</LENGTH>
    <PATH>/Important.bin</PATH>
    <TIMESTAMP>1667411223378</TIMESTAMP>
    <RPC_CLIENTID>db6de0d4-ff21-48bc-a271-9522a1585367</RPC_CLIENTID>
    <RPC_CALLID>3</RPC_CALLID>
  </DATA>
</RECORD>
<RECORD>
  <OPCODE>OP_END_LOG_SEGMENT</OPCODE>
  <DATA>
    <TXID>41</TXID>
  </DATA>
</RECORD>

```

Figure 36. The original old file deletion evidence – EditLogs file Analysis

Enter a Timestamp

1668880898

SECONDS SINCE JAN 01 1970. (UTC)

12:13:14 AM

Supports Unix timestamps in seconds, milliseconds, microseconds and nanoseconds.

Convert →
Copy

| | |
|-----------------------|--|
| Format | Milliseconds (1/1,000 second) |
| GMT | Wed Nov 02 2022 17:47:03 GMT+0000 |
| Your Time Zone | Wed Nov 02 2022 19:47:03 GMT+0200 (Israel Standard Time) |

Figure 37. Time (Human readable)

5.3. Digital forensic analysis during the process of uploading the virus file to the system

The hacker used the command “hdfs dfs -put /”path of the file on masternode directories” /”path of the file on HDFS” and using the view files command, we find that the file was uploaded to the system at 20:10 with the same name as we see in figure 38 on 2/11/2022, the Owner is the

masternode, the Group is the supergroup, and the Permissions are Read and Write for the Owner, and Read for both the Group and the Other.

A terminal window titled 'masternode@masternode:~' with standard window controls. The terminal shows the command 'hdfs dfs -ls /' and its output: 'Found 1 items', '-rw-r--r-- 3 masternode supergroup 0 2022-11-02 20:10 /Important.bin', and the prompt '[masternode@masternode ~]\$'.

```
masternode@masternode:~  
File Edit View Search Terminal Help  
[masternode@masternode ~]$ hdfs dfs -ls /  
Found 1 items  
-rw-r--r-- 3 masternode supergroup 0 2022-11-02 20:10 /Important.bin  
[masternode@masternode ~]$
```

Figure 38 . Listing Ransomware on HDFS

Through figure 39, we can see that the Operation Code is an addition to HDFS, the Transaction number is 53, and the INODE number is 16387, and it was uploaded to the path /Important.bin on the Root directory, and the replication was made between the clusters with a value of 3.

In addition, we see that the Access and Modification times are 1667412631086, and this probably indicates that the file was uploaded at this exact time and the Client Name as shown in the picture. Also, one of the most important Artifacts is the Client Machine and its IP address is 192.168.160.157 It is the same IP address as the Master Node, and this indicates that the file was uploaded from the same Master node and its powers were exploited on HDFS. Also, the Username or Owner is the masternode and the Group is the superuser, and we see that the Replication Client ID as shown in the picture

```

<RECORD>
<OPCODE>OP_ADD</OPCODE>
<DATA>
<TXID>53</TXID>
<LENGTH>0</LENGTH>
<INODEID>16387</INODEID>
<PATH>/Important.bin._COPYING_</PATH>
<REPLICATION>3</REPLICATION>
<MTIME>1667412631086</MTIME>
<ATIME>1667412631086</ATIME>
<BLOCKSIZE>134217728</BLOCKSIZE>
<CLIENT_NAME>DFSClient_NONMAPREDUCE_-628902208_1</CLIENT_NAME>
<CLIENT_MACHINE>192.168.160.157</CLIENT_MACHINE>
<OVERWRITE>>true</OVERWRITE>
<PERMISSION_STATUS>
<USERNAME>masternode</USERNAME>
<GROUPNAME>supergroup</GROUPNAME>
<MODE>420</MODE>
</PERMISSION_STATUS>
<RPC_CLIENTID>c7e59e32-1e87-4327-bbb3-6948301e6004</RPC_CLIENTID>
<RPC_CALLID>3</RPC_CALLID>
</DATA>
</RECORD>

```

Figure 39. Edit Logs analysis

Enter a Timestamp

Supports Unix timestamps in seconds, milliseconds, microseconds and nanoseconds.

Convert →

1668896745

SECONDS SINCE JAN 01 1970. (UTC)

12:25:46 AM

Copy

| | |
|-----------------------|--|
| Format | Milliseconds (1/1,000 second) |
| GMT | Wed Nov 02 2022 18:10:31 GMT+0000 |
| Your Time Zone | Wed Nov 02 2022 20:10:31 GMT+0200 (Israel Standard Time) |

Figure 40. Human readable conversion of the mtime

By analyzing the Network Traffic, we see that the CID is as in figures 41 & 42. Port 5504 was used, see figure 43, and the IP Addresses between the clusters, as well as the Protocol, and the TCP. by search in the. VMEM, the match was found in the CID See figure 44, and this was important congruent evidence to check the search in it, and after confirming the match, the process that took place on the system was accessed, as it was loaded on the RAM first for the MasterNode, then the command was executed, and the results were as in figure 45.

```
.192.168.160.158..cluster2.$3d631138-d67b-49d6-b366-2ac9282160f0 ... (...0...8..8.....
(CID-065a91d5-45ee-4391-9ede-b899f9e6b3a7 .....".2.7.3.1
'DS-acb40ed8-702c-42cf-b1e2-89a2d578e6d0.....0 ... (...0...:-
'DS-acb40ed8-702c-42cf-b1e2-89a2d578e6d0..... (#..... :.0Q.R.'A....[C.|.@.....
4.....".0Q.R.'A....[C.|.(.)
```

Figure 41. CID with the IP address of cluster 2

```
.192.168.160.159..cluster3.$b90da413-f8fa-4d94-833c-1c4ac34f4333 ... (...0...8..8.....
(CID-065a91d5-45ee-4391-9ede-b899f9e6b3a7 .....".2.7.3.1
'DS-0e147761-568d-46f8-8f9e-36a9c2a64a3c.....0 ... (...0...:-
'DS-0e147761-568d-46f8-8f9e-36a9c2a64a3c..... (#..... :.w0..D=E[.;.P...@.....
4.....".w0..D=E[.;.P...(.)
```

Figure 42. CID with the IP address of cluster 3

| | | | | | | | | | | | | | | | | |
|----|--------------|-----------------|-------|-----------------|-----|-----|-------|---|------|------------|----------|---------|---------|---------|---------------|---------------|
| 84 | 10.819703612 | 192.168.160.159 | 55054 | 192.168.160.157 | TCP | 456 | 55054 | → | 8020 | [PSH, ACK] | Seq=1561 | Ack=157 | Win=245 | Len=390 | TsVal=3641000 | TSecr=3640216 |
| 76 | 7.830797013 | 192.168.160.159 | 55054 | 192.168.160.157 | TCP | 66 | 55054 | → | 8020 | [ACK] | Seq=1561 | Ack=157 | Win=245 | Len=0 | TsVal=3638011 | TSecr=3640216 |
| 74 | 7.830259368 | 192.168.160.159 | 55054 | 192.168.160.157 | TCP | 456 | 55054 | → | 8020 | [PSH, ACK] | Seq=1171 | Ack=118 | Win=245 | Len=390 | TsVal=3638011 | TSecr=3637206 |

Figure 43. Source Ports, Network Traffic, and IP addresses

| | | |
|-----------|---|------------------|
| 029d0d440 | 32 30 32 32 0A 6E 61 6D-65 73 70 61 63 65 49 44 | 2022.namespaceID |
| 029d0d450 | 3D 31 31 36 39 30 36 33-39 31 32 0A 63 6C 75 73 | =1169063912.clus |
| 029d0d460 | 74 65 72 49 44 3D 43 49-44 2D 30 36 35 61 39 31 | terID=CID-065a91 |
| 029d0d470 | 64 35 2D 34 35 65 65 2D-34 33 39 31 2D 39 65 64 | d5-45ee-4391-9ed |
| 029d0d480 | 65 2D 62 38 39 39 66 39-65 36 62 33 61 37 0A 63 | e-b899f9e6b3a7.c |
| 029d0d490 | 54 69 6D 65 3D 30 0A 73-74 6F 72 61 67 65 54 79 | Time=0.storageTy |
| 029d0d4a0 | 70 65 3D 4E 41 4D 45 5F-4E 4F 44 45 0A 62 6C 6F | pe=NAME_NODE.blo |
| 029d0d4b0 | 63 6B 70 6F 6F 6C 49 44-3D 42 50 2D 31 37 34 31 | ckpoolID=BP-1741 |
| 029d0d4c0 | 34 35 37 36 30 38 2D 31-39 32 2E 31 36 38 2E 31 | 457608-192.168.1 |
| 029d0d4d0 | 36 30 2E 31 35 37 2D 31-36 36 37 34 30 34 32 38 | 60.157-166740428 |

Figure 44. CID with the IP address of the masternode and the block pool ID

| | | |
|------------|---|------------------|
| 1717656016 | 00 00 00 00 00 00 00 00-31 00 00 00 00 00 00 00 |1..... |
| 1717656032 | 68 64 66 73 20 64 66 73-20 2D 70 75 74 20 7E 2F | hdfs dfs -put ~/ |
| 1717656048 | 44 65 73 6B 74 6F 70 2F-49 6D 70 6F 72 74 61 6E | Desktop/Importan |
| 1717656064 | 74 2E 62 69 6E 20 2F 00-21 00 00 00 00 00 00 00 | t.bin /!..... |

Figure 45. Put Command on. Vmem

5.4. Digital Forensics investigation during virus downloading on the victim's device

The victim user used the command “hdfs dfs -get /Important.bin /”path” to download the file to the device. By analyzing the vmem, we found this command on the RAM and that it was loaded on it during the execution phase.

| | | |
|----------|---|-------------------|
| 1b5edd70 | 68 64 66 73 20 64 66 73-20 2D 67 65 74 20 2F 49 | hdfs dfs -get /I |
| 1b5edd80 | 6D 70 6F 72 74 61 6E 74-2E 62 69 6E 20 7E 2F 44 | important.bin ~/D |
| 1b5edd90 | 65 73 6B 74 6F 70 2F 54-68 65 73 69 73 56 69 72 | esktop/ThesisVir |
| 1b5edda0 | 75 73 2F 0A 68 64 66 73-20 6F 65 76 20 2D 69 20 | us/ -hdfs oev -i |

Figure 46. -get command on. vmem

By analyzing the.VMDK, we see the Path to which the virus was downloaded named Important.bin from HDFS as it is in the following Path on the Masternode: “/masternode/Desktop/ThesisVirus”.

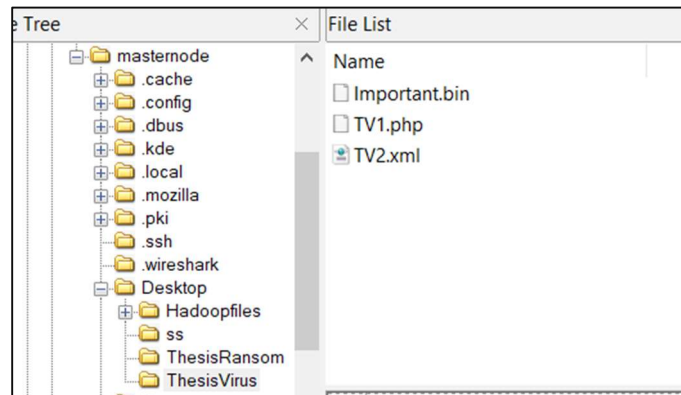


Figure 47. Virus Path

In addition to the presence of other files before running the virus in the same path, which is TV1.php and TV2.xml

5.5. Digital forensic analysis after the process of running the virus file on the system

After running the Revil virus, it was activated and encrypted the files on the same directory, and the files became .rhkrc format (see figure 48) with a Readme file indicating that a sum of money was paid for decryption

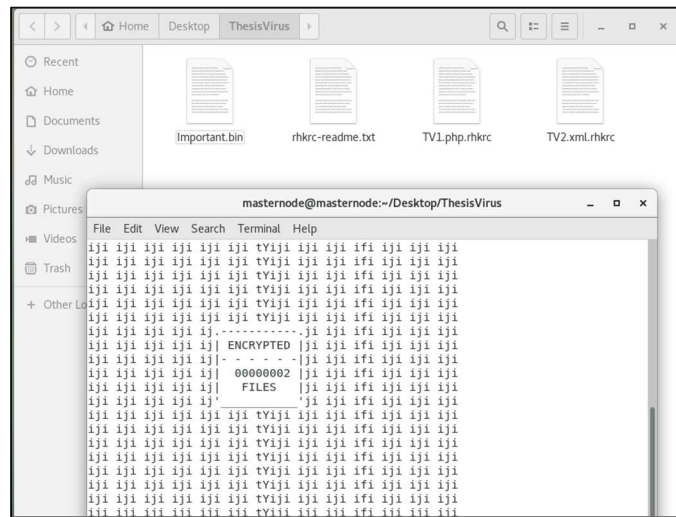


Figure 48. Files Encrypted

After analyzing the .VMDK, the location of the file was analyzed, where the files were encrypted, what they are, and how it happened as shown in figure 49.

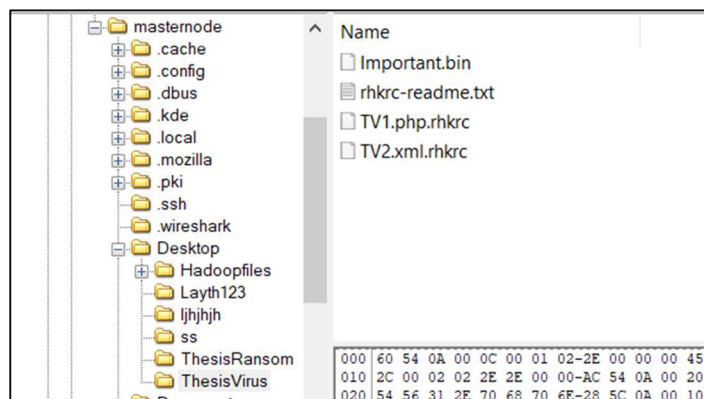


Figure 49. Encrypted files by .VMDK analysis

Then work was done to analyze the VMDK, and here we found that the files were unencrypted with their locations specified, and after the encryption process and the commands that were used to work on the HDFS system

| | | |
|----------|---|------------------|
| lca03b90 | 69 6C 65 3A 2F 2F 2F 68-6F 6D 65 2F 6D 61 73 74 | ile:///home/mast |
| lca03ba0 | 65 72 6E 6F 64 65 2F 44-65 73 6B 74 6F 70 2F 54 | ernode/Desktop/T |
| lca03bb0 | 68 65 73 69 73 56 69 72-75 73 2F 54 56 31 2E 70 | hesisVirus/TV1.g |
| lca03bc0 | 68 70 01 86 A3 05 39 01-86 A3 01 8D 97 01 86 A3 | hp...9...f.....f |

Figure 50. The first file before encryption

| | | |
|----------|---|-------------------|
| lb64f170 | 40 33 19 D8 5D 7F 00 00-00 B8 1A D8 5D 7F 00 00 | @3·0] ····,·0]··· |
| lb64f180 | 74 65 72 6E 6F 64 65 2F-44 65 73 6B 74 6F 70 2F | ternode/Desktop/ |
| lb64f190 | 54 68 65 73 69 73 56 69-72 75 73 2F 54 56 31 2E | ThesisVirus/TV1. |
| lb64f1a0 | 70 68 70 2E 72 68 6B 72-63 00 FF FF 00 00 00 00 | php.rhkrç·ÿÿ···· |

Figure 51. The first file after encryption

| | | |
|----------|---|------------------|
| lc1040a0 | 65 00 3A 00 2F 00 2F 00-2F 00 68 00 6F 00 6D 00 | e-:././././h-o-m |
| lc1040b0 | 65 00 2F 00 6D 00 61 00-73 00 74 00 65 00 72 00 | e-/m-a-s-t-e-r |
| lc1040c0 | 6E 00 6F 00 64 00 65 00-2F 00 44 00 65 00 73 00 | n-o-d-e-/D-e-s |
| lc1040d0 | 6B 00 74 00 6F 00 70 00-2F 00 54 00 68 00 65 00 | k-t-o-p-/T-h-e |
| lc1040e0 | 73 00 69 00 73 00 56 00-69 00 72 00 75 00 73 00 | s-i-s-V-i-r-u-s |
| lc1040f0 | 2F 00 54 00 56 00 32 00-2E 00 78 00 6D 00 6C 00 | /-T-V-2-..x-m-l |

Figure 52. The second file before encryption

| | | |
|----------|---|------------------|
| lcc8a940 | 74 65 72 6E 6F 64 65 2F-44 65 73 6B 74 6F 70 2F | ternode/Desktop/ |
| lcc8a950 | 54 68 65 73 69 73 56 69-72 75 73 2F 54 56 32 2E | ThesisVirus/TV2. |
| lcc8a960 | 78 6D 6C 2E 72 68 6B 72-63 00 00 00 00 00 00 00 | xml.rhkrç····· |
| lcc8a970 | 50 00 00 00 00 00 00 00-30 00 00 00 00 00 00 00 | P·····0····· |

Figure 53. The second file after encryption

6. Results and outcomes from the case investigation

In the beginning, an attack case on a company that operates on the HDFS system was assumed to test the RBDDFF. The targeted Domains were identified and whether the attack was Physical or Logical was determined. The Artifacts that were expected to be obtained during the investigation were identified by analyzing the affected domains. All processes in the Framework were taken into account and the evidence collection was made in a legal manner to ensure acceptance in courts. Through the investigation, it was found that all Domains were targeted through several actions taken by the hacker. The attack was only electronic and there was no physical attack on the company. Evidence was collected from the Domains and it was revealed that the evidence was not as expected, but there were some evidence that could not be accessed. Figure 54 displays the Artifacts expected to be found during the investigation process. In computer domain and by analysing the .VMDK image of the company the results were the paths of the files and the ransomware file before and after running, while .VMEM they found the operations happened PUT,GET, also the time stamps, the permissions, the client names, the files paths, and the commands used on the system by the attacker to upload and delete.

In Network domain, the cluster ID, ports used, the stored Hadoop blocks, time stamps, and protocols used the extracted. While the operations and traffic data didn't find.

Lastly, in the BigData domain they analysed the two log files:

- In Editlogs: OPCODE, Transaction ID, SRC, Client Machine, time stamps, client machine, block size, permissions, INODE ID, and the files paths on Hadoop.
- In FSImage: Files paths on the Hadoop and Metadata.s

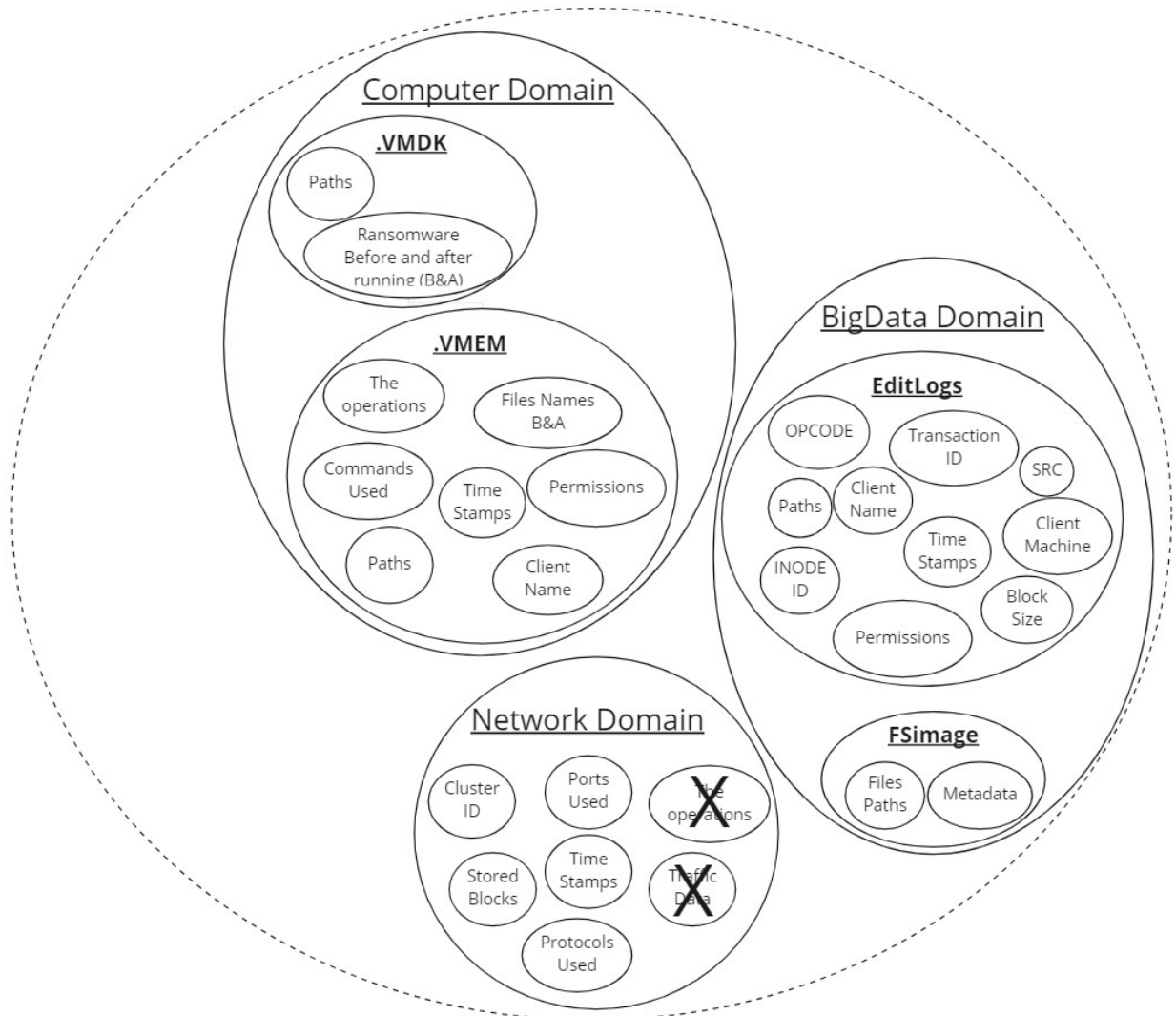


Figure 54. The found Artifacts.

7. The Proposed Report

In the previous sections, work was done on studying the attack on the simulated company Alpha, and the framework was applied to follow the standards and ensure the speed, efficiency, and admissibility of the investigation in court by dropping the physical crime scene investigation from

the framework. Additionally, a practical application of the case was carried out and all possible artifacts were extracted from the investigation processes.

As previously noted, the framework is concerned with ensuring the legality of the investigation process from start to finish. All artifacts were collected and entered into the BDAT tool to obtain the final form of the report that will be presented to the court based on standard forms. Table 7 shows the final report of the investigation. the report displays the most important details which are important to accept in the court depending on Casey & Choen frameworks which follows the standatd forms. the case details of Alpha company shown, Device details, and Framework processes.

Table 7. the proposed Alpha company attack case report

| The Proposed Digital Forensics Report | |
|---|---|
| <i>Case Details</i> | |
| <i>Investigator's name and Association:</i> Layth Abu Arram Islam Amro | <i>Case No.:</i> 9003040601 <i>Report Date:</i> 2/1/2022 <i>Report Time:</i> 10:03 PM (Jerusalem) |
| <i>Location of the Organization:</i> Yatta st. - Yatta – Hebron - Palestine | <i>Name of Suspect(s)/Type of Case:</i> Ali Mohammed / Ransomware exploitation attack |
| <i>Devices Details</i> | |
| <i>Systems Type:</i> Hadoop system installed on different Virtual Machines on one Device | <i>Devices Details:</i> <i>Serial No:</i> 35-6465465 |

| | |
|---|--|
| | <p><i>CPU: Core i7 10500 U</i></p> <p><i>RAM: 16 GB DDR4</i></p> <p><i>Storage: 512 GB M.2</i></p> |
| <p><i>Virtual Machines Details on the organization system:</i></p> | |
| <p><i><u>Virtual Machine Name:</u> <u>Memory Size:</u> <u>Storage Capacity:</u> <u>IP Address:</u></i></p> | |
| Masternode | <p><i>2 GB</i></p> <p><i>20 GB</i></p> <p><i>192.168.160.157</i></p> |
| Cluster 2 | <p><i>2 GB</i></p> <p><i>20 GB</i></p> <p><i>192.168.160.158</i></p> |
| Cluster 3 | <p><i>2 GB</i></p> <p><i>20 GB</i></p> <p><i>192.168.160.159</i></p> |
| <p><i>Framework processes</i></p> | |
| <p><i>Selected Domains:</i></p> <p>Network Domain</p> <p>Computer Domain</p> <p>BigData Domain</p> | <p><i>Type of crime:</i></p> <p>Logical Crime scene only</p> |
| <p><i>Readiness stage:</i></p> <p>A complete plan for the investigation process has been developed.</p> <p>All Organizational Factors have been verified.</p> <p>The evidence was verified to comply with legal methods of collection and analysis.</p> <p>The software required for the investigations of each Domain was used as appropriate.</p> | <p><i>Deployment:</i></p> <p>It was identified by network traffic monitoring.</p> <p>Then the attack was verified.</p> <p>Measures were taken to minimize losses.</p> <p>Then the investigation process began.</p> |
| <p><i>Logical Crime Scene Investigation:</i></p> <p>Identify digital directories from each Domain.</p> | <p><i>Other media:</i></p> <p>Not found</p> |

| | |
|--|--|
| check them out. Make backups, imaging, and dumping. | |
| <i>Findings:</i> <i>On .VMDK:</i> Files paths, ransomware files before and after running. <i>On .VMEM:</i> the operations, files names, commands used, permissions, Timestamps, Client Name, Files Paths. <i>On EditLogs:</i> OPCODE, transaction ID, SRC, Client Machine, Timestamps, InodeID, Client Name, Block Size. <i>On FSimage:</i> Metadata, Files Paths. <i>On Network traffic:</i> Cluster ID, Ports Used, Stored Blocks, Timestamps, Protocols. | |
| <i>Comments:</i> Systems returned to owners with drivers. | |

Chapter 5: Discussion

1. Overview

In this final chapter, the technical and educational challenges faced by researchers during the scientific writing process will be highlighted in the limitations section. The conclusion section will summarize the key outcomes and results of the research. Finally, suggestions for future research in related areas, which are currently considered important topics, will be discussed for researchers interested in pursuing further study.

2. Limitations

During the preparation, research, and preparation of this study, the researchers faced many challenges and problems from the start to the end, including:

The researchers encountered many technical problems in installing the BigData system, despite the existence of explanations for system setup and construction, there were many problems that were not addressed and required a lot of effort to build.

The researchers faced another important problem, which was the lack of available devices for the BigData system. Despite the construction of the system, there had to be a large number of devices for testing. Due to the unavailability, only one laptop was used and a Virtual Machine was installed on it and then the BigData system was installed, which was HDFS.

3. Conclusion:

In this research, various aspects were addressed to solve a crucial issue in digital forensics, which is the forensic analysis of big data. Firstly, the problems encountered by researchers during the forensic analysis of big data were identified, and relevant scientific literature was collected. Then, the strengths and weaknesses of the research were determined and compared with big data, revealing a clear problem in the forensic analysis of big data, which is the absence of a clear framework that deals with the issues as a whole, instead of merely offering theoretical suggestions or steps. Additionally, the framework in this research was built under the name of BigData Digital Forensics Framework (BDDFF) to assist digital criminal investigators in conducting digital forensic investigations of big data and make it easier for them. The framework was built based on relevant research and relying on the Cohen & Casey Frameworks. Furthermore, the framework was developed into the Responsive BigData Digital Forensics Framework (RBDDFF) to speed up digital forensic investigations of big data and reduce the time it takes. The RBDDFF system works by verifying the targeted domains and excluding those that were not targeted. It also checks the targeted crime types and excludes those that were not targeted. Finally, a practical application was carried out as a simulation on the framework, where a BigData system was built using Hadoop on a computer with 3 virtual machines for a hypothetical company called Alpha. At one point, a hacker attacked the system, exploiting privileges and accessing the system, then deleting one of the important files on the system and replacing it with a file with the same name and extension but a Ransomware. After that, one of the users installed and ran the file, causing it to encrypt his files. The digital forensic investigators of Big Data began to investigate the incident, applying the RBDDFF with all its processes legally to ultimately access a set of artifacts extracted from the

system. Finally, the criminal investigators used the BDAT program to automatically generate the final report of the incident.

4. Future Work

The integration of IoT systems and BigData has become a growing area of research, with numerous potential applications and benefits. By linking these systems, researchers can gain a deeper understanding of the data generated by IoT devices and how it can be used in various applications. By analyzing these systems, researchers can gain insights into the structure and workings of IoT systems and develop ways to optimize their performance. One specific area of interest is the use of digital forensic techniques to investigate IoT systems and BigData. By conducting digital forensic investigations, researchers can gather evidence of illegal or malicious activities, identify security threats and vulnerabilities, and recover lost or deleted data. In order to effectively carry out these investigations, researchers can work on developing a BigData and IoT devices Digital Forensics Framework. This framework will provide a systematic and standardized approach to conducting digital forensic investigations of IoT systems and BigData, making it easier for investigators to take action effectively. In addition to investigating IoT systems and BigData, researchers can also work on developing this system to be more responsive to incidents and able to deal with different types of incidents. This can include the development of algorithms and tools to detect and respond to security incidents, and the implementation of security measures to prevent unauthorized access to IoT systems and data. This work will be beneficial for digital forensic investigators, as it will allow them to quickly and effectively deal with different types of incidents, including data breaches, cyber-attacks, and other security threats.

References

- [1] M. Ghobakhloo, "Industry 4.0, digitization, and opportunities for sustainability," *Journal of Cleaner Production*, vol. 252. Elsevier Ltd, Apr. 10, 2020. doi: 10.1016/j.jclepro.2019.119869.
- [2] M. Khan, X. Wu, X. Xu, and W. Dou, "Big data challenges and opportunities in the hype of Industry 4.0," in *IEEE International Conference on Communications*, Institute of Electrical and Electronics Engineers Inc., Jul. 2017. doi: 10.1109/ICC.2017.7996801.
- [3] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues," *Journal of Industrial Information Integration*, vol. 6. Elsevier B.V., pp. 1–10, Jun. 01, 2017. doi: 10.1016/j.jii.2017.04.005.
- [4] H. A. Abbas, "Future SCADA challenges and the promising solution: The agent-based SCADA," *Int. J. Crit. Infrastructures*, vol. 10, no. 3–4, pp. 307–333, 2014, doi: 10.1504/IJCIS.2014.066354.
- [5] O. M. Adedayo, "Big data and digital forensics, Rethinking Digital Forensics," *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*. pp. 1–7, 2016. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7740422&isnumber=7740415>
- [6] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4. King Saud bin Abdulaziz University, pp. 431–448, Oct. 01, 2018. doi: 10.1016/j.jksuci.2017.06.001.

- [7] B. Ramesh, "Big Data Architecture," 2015, pp. 29–59. doi: 10.1007/978-81-322-2494-5_2.
- [8] "Worldometer - real time world statistics." <https://www.worldometers.info/> (accessed Jan. 29, 2023).
- [9] D. Che, M. Safran, and Z. Peng, "From big data to big data mining: Challenges, issues, and opportunities," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7827 LNCS. pp. 1–15, 2013. doi: 10.1007/978-3-642-40270-8_1.
- [10] N. Khan *et al.*, "Big Data: Survey, Technologies, Opportunities, and Challenges," *Sci. World J.*, vol. 2014, no. 8, pp. 1–18, 2014, doi: 10.1155/2014/712826.
- [11] "Statista - The Statistics Portal for Market Data, Market Research and Market Studies." <https://www.statista.com/> (accessed Dec. 02, 2022).
- [12] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big Data Analytics in Intelligent Transportation Systems: A Survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1. Institute of Electrical and Electronics Engineers Inc., pp. 383–398, Jan. 01, 2019. doi: 10.1109/TITS.2018.2815678.
- [13] "Big Data In Business: 9 Examples & Applications | MongoDB." <https://www.mongodb.com/big-data-explained/examples> (accessed Apr. 25, 2023).
- [14] M. Bydon *et al.*, "Big Data Defined: A Practical Review for Neurosurgeons," *World Neurosurg.*, vol. 133, pp. e842–e849, 2020, doi: 10.1016/j.wneu.2019.09.092.
- [15] J. Fan, F. Han, and H. Liu, "Challenges of Big Data analysis," *Natl. Sci. Rev.*, vol. 1, no.

- 2, pp. 293–314, Jun. 2014, doi: 10.1093/nsr/nwt032.
- [16] Darren Quick; Kim-Kwang Raymond Choo, “Data reduction and data mining framework for digital forensic evidence: Storage, intelligence, review and archive.” pp. 1–11. doi: 10.3316/ielapa.615195995666597.
- [17] S. L. Garfinkel, “Digital forensics research: The next 10 years,” *Digit. Investig.*, vol. 7, no. SUPPL., pp. S64–S73, 2010, doi: 10.1016/j.diin.2010.05.009.
- [18] N. Beebe, “Digital forensic research: The good, the bad and the unaddressed,” *IFIP Adv. Inf. Commun. Technol.*, vol. 306, pp. 17–36, 2009, doi: 10.1007/978-3-642-04155-6_2.
- [19] G. Grispos, T. Storer, and W. Glisson, “Calm before the storm: the challenges of cloud,” *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 4, no. 2. pp. 28–48, 2012. [Online]. Available: <https://books.google.com/books?hl=en&lr=&id=U8m6YLhoLesC&oi=fnd&pg=PA211&dq=Calm+Before+the+Storm:+The+Challenges+of+Cloud+Computing+in+Digital+Forensics&ots=EDQ9QKJ0VM&sig=g2rON1mlUaw71Cu2WGmgivXbij8>
- [20] A. Singh and K. Chatterjee, “Cloud security issues and challenges: A survey,” *Journal of Network and Computer Applications*, vol. 79. Academic Press, pp. 88–115, Feb. 01, 2017. doi: 10.1016/j.jnca.2016.11.027.
- [21] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” *Neurocomputing*, vol. 237, pp. 350–361, 2017, doi: 10.1016/j.neucom.2017.01.026.
- [22] J. Wang, Y. Yang, T. Wang, R. Simon Sherratt, and J. Zhang, “Big data service

- architecture: A survey,” *Journal of Internet Technology*, vol. 21, no. 2. pp. 393–405, 2020. doi: 10.3966/160792642020032102008.
- [23] P. Karunaratne, S. Karunasekera, and A. Harwood, “Distributed stream clustering using micro-clusters on Apache Storm,” *J. Parallel Distrib. Comput.*, vol. 108, pp. 74–84, Oct. 2017, doi: 10.1016/j.jpdc.2016.06.004.
- [24] J. M. R. S. Rathore, V. S., Dey, N., Piuri, V., Babo, R., Polkowski, Z., & Tavares, “Rising threats in expert applications and solutions. Advances in intelligent systems and computing, 1187,” in *Advances in Intelligent Systems and Computing*, 2021, pp. 175–186. doi: 10.1007/978-981-15-6014-9_20.
- [25] R. Mukherjee and P. Kar, “A comparative review of data warehousing ETL tools with new trends and industry insight,” in *Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017*, 2017, pp. 943–948. doi: 10.1109/IACC.2017.0192.
- [26] J. Nwokeji, F. Aqlan, A. Anugu, and A. Olagunju, “Big Data ETL Implementation Approaches: A Systematic Literature Review (P),” Jul. 2018, pp. 714–721. doi: 10.18293/SEKE2018-152.
- [27] N. B. Abdul Ghani *et al.*, “Tracking Dengue on Twitter Using Hybrid Filtration-Polarity and Apache Flume,” *Comput. Syst. Sci. Eng.*, vol. 40, no. 3, pp. 913–926, Sep. 2021, doi: 10.32604/CSSE.2022.018467.
- [28] “Welcome to Apache Flume — Apache Flume.” <https://flume.apache.org/> (accessed Dec. 12, 2022).
- [29] H. C. Lu, F. J. Hwang, and Y. H. Huang, “Parallel and distributed architecture of genetic

- algorithm on Apache Hadoop and Spark,” *Appl. Soft Comput. J.*, vol. 95, Oct. 2020, doi: 10.1016/j.asoc.2020.106497.
- [30] B. Shu, H. Chen, and M. Sun, “Dynamic load balancing and channel strategy for apache flume collecting real-time data stream,” in *Proceedings - 15th IEEE International Symposium on Parallel and Distributed Processing with Applications and 16th IEEE International Conference on Ubiquitous Computing and Communications, ISPA/IUCC 2017*, Institute of Electrical and Electronics Engineers Inc., May 2018, pp. 542–549. doi: 10.1109/ISPA/IUCC.2017.00089.
- [31] “Apache Kafka.” <https://kafka.apache.org/> (accessed Dec. 12, 2022).
- [32] H. Jafarpour, R. Desai, and D. Guy, “KSQL: Streaming SQL engine for Apache Kafka,” in *Advances in Database Technology - EDBT*, OpenProceedings.org, 2019, pp. 524–533. doi: 10.5441/002/edbt.2019.48.
- [33] “Taobao.” <https://world.taobao.com/?> (accessed Dec. 12, 2022).
- [34] C. A. D. Deagustini, S. E. Fulladoza Dalibón, S. Gottifredi, M. A. Falappa, C. I. Chesñevar, and G. R. Simari, “Relational databases as a massive information source for defeasible argumentation,” *Knowledge-Based Syst.*, vol. 51, pp. 93–109, Oct. 2013, doi: 10.1016/j.knosys.2013.07.010.
- [35] N. Chaudhry and M. M. Yousaf, “Architectural assessment of NoSQL and NewSQL systems,” *Distrib. Parallel Databases*, vol. 38, no. 4, pp. 881–926, Dec. 2020, doi: 10.1007/s10619-020-07310-1.
- [36] S. Ghemawat, H. Gobioff, and S.-T. Leung Google, “The Google File System,” 2003.

- [37] M. Wang, B. Li, Y. Zhao, and G. Pu, “Formalizing google file system,” in *Proceedings of IEEE Pacific Rim International Symposium on Dependable Computing, PRDC*, IEEE Computer Society, Dec. 2014, pp. 190–191. doi: 10.1109/PRDC.2014.32.
- [38] “Apache Hadoop.” <https://hadoop.apache.org/> (accessed Dec. 24, 2022).
- [39] Z. Zhu, L. Tan, Y. Li, and C. Ji, “PHDFS: Optimizing I/O performance of HDFS in deep learning cloud computing platform,” *J. Syst. Archit.*, vol. 109, Oct. 2020, doi: 10.1016/j.sysarc.2020.101810.
- [40] T. Yeh and T. Chien, “Building a version control system in the Hadoop HDFS,” in *IEEE/IFIP Network Operations and Management Symposium: Cognitive Management in a Cyber World, NOMS 2018*, 2018, pp. 1–5. doi: 10.1109/NOMS.2018.8406190.
- [41] M. M. Shetty and D. H. Manjaiah, “Data security in Hadoop distributed file system,” in *Proceedings of IEEE International Conference on Emerging Technological Trends in Computing, Communications and Electrical Engineering, ICETT 2016*, 2017. doi: 10.1109/ICETT.2016.7873697.
- [42] T. N. Khasawneh, M. H. Al-Sahlee, and A. A. Safia, “SQL, NewSQL, and NOSQL Databases: A Comparative Survey,” in *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, Institute of Electrical and Electronics Engineers Inc., Apr. 2020, pp. 13–21. doi: 10.1109/ICICS49469.2020.239513.
- [43] A. Pavlo and M. Aslett, “What’s really new with NewSQL?,” *SIGMOD Record*, vol. 45, no. 2. pp. 45–55, 2016. doi: 10.1145/3003665.3003674.
- [44] J. Chen, S. Jindal, R. Walzer, R. Sen, N. Jimsheleishvilli, and M. Andrews, “The

- MemSQL query optimizer: A modern optimizer for real-time analytics in a distributed database,” in *Proceedings of the VLDB Endowment*, 2015, pp. 1401–1412. doi: 10.14778/3007263.3007277.
- [45] D. F. Bacon *et al.*, “Spanner: Becoming a SQL system,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Association for Computing Machinery, May 2017, pp. 331–343. doi: 10.1145/3035918.3056103.
- [46] J. Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” in *OSDI 2004 - 6th Symposium on Operating Systems Design and Implementation*, 2004, pp. 137–149. doi: 10.21276/ijre.2018.5.5.4.
- [47] D. Park, J. Wang, and Y. S. Kee, “In-Storage Computing for Hadoop MapReduce Framework: Challenges and Possibilities,” *IEEE Trans. Comput.*, 2021, doi: 10.1109/TC.2016.2595566.
- [48] “Apache Storm.” <https://storm.apache.org/> (accessed Dec. 14, 2022).
- [49] M. Hussain Iqbal and T. Rahim Soomro, “Big Data Analysis: Apache Storm Perspective,” *Int. J. Comput. Trends Technol.*, vol. 19, no. 1, pp. 9–14, 2015, doi: 10.14445/22312803/ijctt-v19p103.
- [50] X. Meng *et al.*, “MLlib: Machine learning in Apache Spark,” *J. Mach. Learn. Res.*, vol. 17, pp. 1–7, 2016.
- [51] Apache, “Apache Spark™ - Unified Engine for large-scale data analytics,” *Apache*, 2022. <https://spark.apache.org/> (accessed Dec. 14, 2022).
- [52] L. Windrim, R. Ramakrishnan, A. Melkumyan, and R. J. Murphy, “A Physics-Based Deep

- Learning Approach to Shadow Invariant Representations of Hyperspectral Images,” *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 665–677, Feb. 2018, doi: 10.1109/TIP.2017.2761542.
- [53] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, “Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility,” *Futur. Gener. Comput. Syst.*, vol. 25, no. 6, pp. 599–616, Jun. 2009, doi: 10.1016/j.future.2008.12.001.
- [54] D. Silver *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.
- [55] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, “DeepFM: A factorization-machine based neural network for CTR prediction,” in *IJCAI International Joint Conference on Artificial Intelligence*, Mar. 2017, pp. 1725–1731. doi: 10.24963/ijcai.2017/239.
- [56] “Cloud Computing Services - Amazon Web Services (AWS).” <https://aws.amazon.com/> (accessed Dec. 15, 2022).
- [57] “Cloud Computing Services | Microsoft Azure.” <https://azure.microsoft.com/en-us/> (accessed Dec. 15, 2022).
- [58] “Alibaba Cloud: Cloud Computing Services.” https://middle-east.alibabacloud.com/en?utm_key=se_1012200289&utm_content=se_1012200289 (accessed Dec. 15, 2022).
- [59] H. Li, W. Li, H. Wang, and J. Wang, “An optimization of virtual machine selection and placement by using memory content similarity for server consolidation in cloud,” *Futur.*

- Gener. Comput. Syst.*, vol. 84, pp. 98–107, 2018, doi: 10.1016/j.future.2018.02.026.
- [60] H. Li, W. Li, Q. Feng, S. Zhang, H. Wang, and J. Wang, “Leveraging content similarity among VMI files to allocate virtual machines in cloud,” *Futur. Gener. Comput. Syst.*, vol. 79, pp. 528–542, Feb. 2018, doi: 10.1016/j.future.2017.09.058.
- [61] M. H. Ghahramani, M. Zhou, and C. T. Hon, “Toward cloud computing QoS architecture: Analysis of cloud systems and cloud services,” *IEEE/CAA J. Autom. Sin.*, vol. 4, no. 1, pp. 6–18, Jan. 2017, doi: 10.1109/JAS.2017.7510313.
- [62] L. Wang, J. Tao, M. Kunze, A. C. Castellanos, D. Kramer, and W. Karl, “Scientific cloud computing: Early definition and experience,” in *Proceedings - 10th IEEE International Conference on High Performance Computing and Communications, HPCC 2008*, 2008, pp. 825–830. doi: 10.1109/HPCC.2008.38.
- [63] S. Costache, D. Dib, N. Parlavantzas, and C. Morin, “Resource management in cloud platform as a service systems: Analysis and opportunities,” *J. Syst. Softw.*, vol. 132, pp. 98–118, Oct. 2017, doi: 10.1016/j.jss.2017.05.035.
- [64] T. (Tom E. . White, *Hadoop: the definitive guide*. O’Reilly, 2012.
- [65] M. R. Ghazi and D. Gangodkar, “Hadoop, mapreduce and HDFS: A developers perspective,” in *Procedia Computer Science*, Elsevier B.V., 2015, pp. 45–50. doi: 10.1016/j.procs.2015.04.108.
- [66] D. Borthakur, “The hadoop distributed file system: Architecture and design,” *Hadoop Proj. Website*, pp. 1–14, 2007, [Online]. Available: <http://hadoop.apache.org/core/>.
- [67] X. Tang, Z. Tao, P. Tang, and J. Li, “Research and improvement of HDFS,” in *2014 11th*

- International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2014*, 2014, pp. 427–429. doi: 10.1109/ICCWAMTIP.2014.7073442.
- [68] A. S. Nagdive, R. M. Tugnayat, and S. Agnihotri, “A Review of Hadoop Ecosystem for BigData,” 2018. [Online]. Available: <http://blog.agro-know.com/?p=3810>
- [69] G. Porter, “Decoupling storage and computation in Hadoop with SuperDataNodes,” in *Operating Systems Review (ACM)*, 2010, pp. 41–46. doi: 10.1145/1773912.1773923.
- [70] C. H. Lin, J. C. Liu, and T. C. Peng, “Performance evaluation of cluster algorithms for Big Data analysis on cloud,” in *Proceedings of the 2017 IEEE International Conference on Applied System Innovation: Applied System Innovation for Modern Technology, ICASI 2017*, 2017, pp. 1434–1437. doi: 10.1109/ICASI.2017.7988182.
- [71] S. Maitrey and C. K. Jha, “MapReduce: Simplified Data Analysis of Big Data,” in *Procedia Computer Science*, Elsevier, 2015, pp. 563–571. doi: 10.1016/j.procs.2015.07.392.
- [72] F. Tekiner and J. A. Keane, “Big Data Framework,” in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE, Oct. 2013, pp. 1494–1499. doi: 10.1109/SMC.2013.258.
- [73] J. Gray and T. C. Bressoud, “Towards a MapReduce Application Performance Model,” *Midstates Conf. Undergrad. Res. Comput. Sci. Math.*, no. 1, 2012, [Online]. Available: <http://math.owu.edu/MCURCSM/papers/paper4.pdf>
- [74] J. Ekanayake, S. Pallickara, and G. Fox, “MapReduce for data intensive scientific

- analyses,” in *Proceedings - 4th IEEE International Conference on eScience, eScience 2008*, 2008, pp. 277–284. doi: 10.1109/eScience.2008.59.
- [75] A. Verma, B. Cho, N. Zea, I. Gupta, and R. H. Campbell, “Breaking the MapReduce stage barrier,” *Cluster Comput.*, vol. 16, no. 1, pp. 191–206, 2013, doi: 10.1007/s10586-011-0182-7.
- [76] N. Ruff, “Windows memory forensics,” *J. Comput. Virol.*, vol. 4, no. 2, pp. 83–100, May 2008, doi: 10.1007/s11416-007-0070-0.
- [77] “FTK Imager.” https://www.pluralsight.com/courses/accessdata-forensic-toolkit-ftk-imager?aid=7010a000002BWqGAAW&promo=&utm_source=non_branded&utm_medium=digital_paid_search_google&utm_campaign=EMEA_Dynamic&utm_content=&gclid=Cj0KCQiA45qdBhD-ARIsAOHbVdF49Kk3ZID-dM_JjD1lzvJx37nKUyFXzDRy3VSdzAXhOwMIHZy3YkAaAt0FEALw_wcB (accessed Dec. 24, 2022).
- [78] “Belkasoft RAM Capturer: Volatile Memory Acquisition Tool.” <https://belkasoft.com/ram-capturer> (accessed Dec. 24, 2022).
- [79] “Acquire Memory with MAGNET RAM Capture.” <https://support.magnetforensics.com/s/article/Acquire-Memory-with-MAGNET-RAM-Capture> (accessed Dec. 24, 2022).
- [80] “DumpIt (Memory Dumper) :: Tools - ToolWar | Information Security (InfoSec) Tools.” <http://www.toolwar.com/2014/01/dumpit-memory-dump-tools.html> (accessed Dec. 24, 2022).

- [81] “Wireshark · Go Deep.” <https://www.wireshark.org/> (accessed Jun. 30, 2022).
- [82] A. V. Hussein Al-Saadawi, “VOLATILE MEMORY ANALYSIS TOOLS FOR VOIP FORENSIC APPLICATIONS: A CLASSIFICATION STUDY.” 2017.
- [83] M. N. Faiz and W. A. Prabowo, “Comparison of Acquisition Software for Digital Forensics Purposes,” *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, pp. 37–44, Nov. 2018, doi: 10.22219/kinetik.v4i1.687.
- [84] V. Ndatinya, Z. Xiao, V. R. Manepalli, K. Meng, and Y. Xiao, “Network forensics analysis using Wireshark,” *Int. J. Secur. Networks*, vol. 10, no. 2, p. 91, 2015, doi: 10.1504/IJSN.2015.070421.
- [85] P. Adluru, S. S. Datla, and X. Zhang, “Hadoop eco system for big data security and privacy,” in *2015 Long Island Systems, Applications and Technology*, IEEE, May 2015, pp. 1–6. doi: 10.1109/LISAT.2015.7160211.
- [86] G. S. Chhabra, V. Singh, and M. Singh, “Hadoop-based analytic framework for cyber forensics,” *Int. J. Commun. Syst.*, vol. 31, no. 15, p. e3772, Oct. 2018, doi: 10.1002/dac.3772.
- [87] M. Asim, D. R. McKinnel, A. Dehghantanha, R. M. Parizi, M. Hammoudeh, and G. Epiphaniou, “Big Data Forensics: Hadoop Distributed File Systems as a Case Study,” in *Handbook of Big Data and IoT Security*, Cham: Springer International Publishing, 2019, pp. 179–210. doi: 10.1007/978-3-030-10543-3_8.
- [88] E. Harshany, R. Benton, D. Bourrie, and W. Glisson, “Big Data Forensics: Hadoop 3.2.0 Reconstruction,” *Forensic Sci. Int. Digit. Investig.*, vol. 32, p. 300909, Apr. 2020, doi:

- 10.1016/j.fsidi.2020.300909.
- [89] J. Baldwin, O. M. K. Alhawi, S. Shaughnessy, A. Akinbi, and A. Dehghantanha, “Emerging from the Cloud: A Bibliometric Analysis of Cloud Forensics Studies,” in *Advances in Information Security*, Springer New York LLC, 2018, pp. 311–331. doi: 10.1007/978-3-319-73951-9_16.
- [90] F. Daryabar, A. Dehghantanha, B. Eterovic-Soric, and K.-K. R. Choo, “Forensic investigation of OneDrive, Box, GoogleDrive and Dropbox applications on Android and iOS devices,” *Aust. J. Forensic Sci.*, vol. 48, no. 6, pp. 615–642, Nov. 2016, doi: 10.1080/00450618.2015.1110620.
- [91] F. Norouzizadeh Dezfouli, A. Dehghantanha, B. Eterovic-Soric, and K.-K. R. Choo, “Investigating Social Networking applications on smartphones detecting Facebook, Twitter, LinkedIn and Google+ artefacts on Android and iOS platforms,” *Aust. J. Forensic Sci.*, vol. 48, no. 4, pp. 469–488, Jul. 2016, doi: 10.1080/00450618.2015.1066854.
- [92] S. H. Mohtasebi *et al.*, *Cloud Storage Forensics: Analysis of Data Remnants on SpiderOak, JustCloud, and pCloud*, no. 13. Elsevier, 2017. [Online]. Available: <https://arxiv.org/ftp/arxiv/papers/1706/1706.08042.pdf>
- [93] A. Dehghantanha and T. Dargahi, “Residual Cloud Forensics: CloudMe and 360Yunpan as Case Studies,” in *Contemporary Digital Forensic Investigations of Cloud and Mobile Applications*, Elsevier Inc., 2017, pp. 247–283. doi: 10.1016/B978-0-12-805303-4.00014-9.
- [94] M. N. Yusoff, A. Dehghantanha, and R. Mahmood, “Network Traffic Forensics on Firefox

- Mobile OS: Facebook, Twitter, and Telegram as Case Studies,” in *Contemporary Digital Forensic Investigations of Cloud and Mobile Applications*, 2017, pp. 63–78. doi: 10.1016/B978-0-12-805303-4.00005-8.
- [95] H. Haughey, G. Epiphaniou, H. Al-Khateeb, and A. Dehghantanha, “Adaptive traffic fingerprinting for darknet threat intelligence,” in *Advances in Information Security*, 2018, pp. 193–217. doi: 10.1007/978-3-319-73951-9_10.
- [96] Y. Y. Teing, A. Dehghantanha, K. K. R. Choo, Z. Muda, and M. T. Abdullah, “Greening Cloud-Enabled Big Data Storage Forensics: Syncany as a Case Study,” *IEEE Trans. Sustain. Comput.*, vol. 4, no. 2, pp. 204–216, Apr. 2019, doi: 10.1109/TSUSC.2017.2687103.
- [97] B. Martini and K. K. R. Choo, “Distributed filesystem forensics: XtreamFS as a case study,” *Digit. Investig.*, vol. 11, no. 4, pp. 295–313, Dec. 2014, doi: 10.1016/j.diin.2014.08.002.
- [98] S. Dinesh, S. Rao, and K. Chandrasekaran, “Traceback: A forensic tool for distributed systems,” in *Smart Innovation, Systems and Technologies*, Springer Science and Business Media Deutschland GmbH, 2016, pp. 17–27. doi: 10.1007/978-81-322-2529-4_2.
- [99] E. Alshammari, G. Al-Naymat, and A. Hadi, “A New Technique for File Carving on Hadoop Ecosystem,” in *Proceedings - 2017 International Conference on New Trends in Computing Sciences, ICTCS 2017*, Institute of Electrical and Electronics Engineers Inc., Jul. 2017, pp. 72–77. doi: 10.1109/ICTCS.2017.16.
- [100] Y.-Y. Teing, A. Dehghantanha, K.-K. R. Choo, T. Dargahi, and M. Conti, “Forensic

- Investigation of Cooperative Storage Cloud Service: Symform as a Case Study,” *J. Forensic Sci.*, vol. 62, no. 3, pp. 641–654, May 2017, doi: 10.1111/1556-4029.13271.
- [101] Y. Y. Teing, A. Dehghantanha, K. K. R. Choo, and L. T. Yang, “Forensic investigation of P2P cloud storage services and backbone for IoT networks: BitTorrent Sync as a case study,” *Comput. Electr. Eng.*, vol. 58, pp. 350–363, Feb. 2017, doi: 10.1016/j.compeleceng.2016.08.020.
- [102] K. Hausknecht, D. Foit, and J. Burić, “RAM data significance in Digital Forensics,” no. May. pp. 1372–1375, 2015.
- [103] M. Al Fahdi, N. L. Clarke, and S. M. Furnell, “Challenges to digital forensics: A survey of researchers & practitioners attitudes and opinions,” in *2013 Information Security for South Africa*, IEEE, Aug. 2013, pp. 1–8. doi: 10.1109/ISSA.2013.6641058.
- [104] G. Horsman, “Tool testing and reliability issues in the field of digital forensics,” *Digit. Investig.*, vol. 28, pp. 163–175, Mar. 2019, doi: 10.1016/j.diin.2019.01.009.
- [105] E. A. Vincze, “Challenges in digital forensics,” *Police Pract. Res.*, vol. 17, no. 2, pp. 183–194, Mar. 2016, doi: 10.1080/15614263.2015.1128163.
- [106] N. Rana, G. Sansanwal, K. Khatter, S. Singh, and M. Rachna, “Taxonomy of Digital Forensics : Investigation Tools and Challenges Department of Computer Science and Engineering Accendere Knowledge Management Services Pvt . Ltd ., India,” *Comput. Soc.*, no. April, 2017, [Online]. Available: <https://arxiv.org/ftp/arxiv/papers/1709/1709.06529.pdf>
- [107] L. Caviglione, S. Wendzel, and W. Mazurczyk, “The Future of Digital Forensics:

- Challenges and the Road Ahead,” *IEEE Secur. Priv.*, vol. 15, no. 6, pp. 12–17, Nov. 2017, doi: 10.1109/MSP.2017.4251117.
- [108] M. Rafique and M. N. A. Khan, “Exploring Static and Live Digital Forensics: Methods, Practices and Tools,” *Int. J. Sci. Eng. Res.*, vol. 4, no. 10, pp. 1048–1056, 2013, [Online]. Available: <http://www.ijser.org/researchpaper%5CExploring-Static-and-Live-Digital-Forensic-Methods-Practices-and-Tools.pdf>
- [109] X. Du *et al.*, “SoK: Exploring the state of the art and the future potential of artificial intelligence in digital forensic investigation,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2020. doi: 10.1145/3407023.3407068.
- [110] A. Sayakkara, N. A. Le-Khac, and M. Scanlon, “A survey of electromagnetic side-channel attacks and discussion on their case-progressing potential for digital forensics,” *Digital Investigation*, vol. 29. Elsevier Ltd, pp. 43–54, Jun. 01, 2019. doi: 10.1016/j.diin.2019.03.002.
- [111] J. Yaacoub, H. Noura, ... O. S. preprint arXiv, and U. 2021, “Digital Forensics vs. Anti-Digital Forensics: Techniques, Limitations and Recommendations,” *arxiv.org*, 2021, [Online]. Available: <https://arxiv.org/abs/2103.17028>
- [112] H. M. A. Van Beek, E. J. Van Eijk, R. B. Van Baar, M. Ugen, J. N. C. Bodde, and A. J. Siemelink, “Digital forensics as a service: Game on,” *Digit. Investig.*, vol. 15, pp. 20–38, Dec. 2015, doi: 10.1016/j.diin.2015.07.004.
- [113] R. Kaur and A. Kaur, “Digital Forensics,” *Int. J. Comput. Appl.*, vol. 50, no. 5, pp. 5–9,

- Jul. 2012, doi: 10.5120/7765-0844.
- [114] Z. H. Yunus Yusoff, Roslan Ismail, “Common Phases of Computer Forensics Investigation,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, 2011, doi: 10.5121/ijcsit.2011.3302.
- [115] M. G. Noblett, M. M. Pollitt, and L. A. Presley, “Recovering and Examining Computer Forensic Evidence,” *Forensic Sci. Commun.*, vol. 2, no. 4, pp. 1–8, 2000.
- [116] F. Solinas and G. Fenu, “COMPUTER FORENSICS INVESTIGATION AN APPROACH TO EVIDENCE IN CYBERSPACE.” [Online]. Available: <https://www.researchgate.net/publication/240291059>
- [117] M. M. Pollitt, “An Ad Hoc Review of Digital Forensic Models,” *Proc. Second Int. Work. Syst. Approaches to Digit. Forensic Eng.*, pp. 43–54, 2007, [Online]. Available: <http://www.digitalevidencepro.com/Resources/Approach.pdf>
- [118] F. Adelstein, “Live forensics: Diagnosing your system without killing it first,” *Communications of the ACM*, vol. 49, no. 2. pp. 63–66, 2006. doi: 10.1145/1113034.1113070.
- [119] L. Wang, R. Zhang, and S. Zhang, “A model of computer live forensics based on physical memory analysis,” *2009 1st International Conference on Information Science and Engineering, ICISE 2009*. pp. 4647–4649, 2009. doi: 10.1109/ICISE.2009.69.
- [120] F. Albanna and I. Riadi, “Forensic Analysis of Frozen Hard Drive Using Static Forensics Method,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 15, no. 1, pp. 173–178, 2017, [Online]. Available: <https://sites.google.com/site/ijcsis/>

- [121] A. Prayogo, I. Riadi, and A. Luthfi, "Mobile Forensics Development of Mobile Banking Application using Static Forensic," *Int. J. Comput. Appl.*, vol. 160, no. 1, pp. 5–10, Feb. 2017, doi: 10.5120/ijca2017912925.
- [122] T. Rochmadi, I. Riadi, and Y. Prayudi, "Live Forensics for Anti-Forensics Analysis on Private Portable Web Browser," *Int. J. Comput. Appl.*, vol. 164, no. 8, pp. 31–37, 2017, doi: 10.5120/ijca2017913717.
- [123] G. L. Garcia, "Forensic physical memory analysis: an overview of tools and techniques," *TKK T-110.5290 Semin. Netw. Secur.*, pp. 305–320, 2007.
- [124] & W. Ligh, M. H., Case, A., Levy, J., "The Art of Memory Forensics Detecting Malware and Threats in Windows, Linux, and Mac Memory," *John Wiley Sons*, vol. 58, no. 12, pp. 7250–7257, 2014, doi: 10.1128/AAC.03728-14.
- [125] "Foremost." <https://foremost.sourceforge.net/> (accessed Dec. 24, 2022).
- [126] "e-fense :: Cyber Security & Computer Forensics Software." <http://www.e-fense.com/products.php> (accessed Dec. 24, 2022).
- [127] "The Volatility Foundation - Open Source Memory Forensics." <https://www.volatilityfoundation.org/> (accessed Dec. 24, 2022).
- [128] "WinHex: Hex Editor & Disk Editor, Computer Forensics & Data Recovery Software." <https://www.x-ways.net/winhex/> (accessed Dec. 24, 2022).
- [129] A. Kazim, F. Almaeeni, S. Al Ali, F. Iqbal, and K. Al-Hussaeni, "Memory Forensics: Recovering Chat Messages and Encryption Master Key," in *2019 10th International Conference on Information and Communication Systems, ICICS 2019*, 2019, pp. 58–64.

- doi: 10.1109/IACS.2019.8809179.
- [130] A. Ghafarian and C. Wood, “Forensics data recovery of skype communication from physical memory,” in *Advances in Intelligent Systems and Computing*, 2019, pp. 995–1009. doi: 10.1007/978-3-030-01177-2_72.
- [131] A. Case and G. G. Richard, “Memory forensics: The path forward,” *Digit. Investig.*, vol. 20, pp. 23–33, Mar. 2017, doi: 10.1016/j.diin.2016.12.004.
- [132] Y. T. Chang, M. J. Chung, C. F. Lee, C. T. Huang, and S. J. Wang, “Memory forensics for key evidence investigations in case illustrations,” *Proceedings - 2013 8th Asia Joint Conference on Information Security, AsiaJCIS 2013*. pp. 96–101, 2013. doi: 10.1109/ASIAJCIS.2013.22.
- [133] S. Rahman and M. N. A. Khan, “Review of Live Forensic Analysis Techniques,” *Int. J. Hybrid Inf. Technol.*, vol. 8, no. 2, pp. 379–388, Feb. 2015, doi: 10.14257/ijhit.2015.8.2.35.
- [134] R. Layton, P. Watters, and R. Dazeley, “Automatically determining phishing campaigns using the USCAP methodology,” in *General Members Meeting and eCrime Researchers Summit, eCrime 2010*, 2010. doi: 10.1109/ecrime.2010.5706698.
- [135] S. Khan, A. Gani, A. W. A. Wahab, M. Shiraz, and I. Ahmad, “Network forensics: Review, taxonomy, and open challenges,” *Journal of Network and Computer Applications*, vol. 66. Academic Press, pp. 214–235, May 01, 2016. doi: 10.1016/j.jnca.2016.03.005.
- [136] Y. Zhu, “Attack Pattern Discovery in Forensic Investigation of Network Attacks,” *IEEE J.*

- Sel. Areas Commun.*, vol. 29, no. 7, pp. 1349–1357, Aug. 2011, doi: 10.1109/JSAC.2011.110802.
- [137] M. Thapliyal, A. Bijalwan, N. Garg, and E. Pilli, “A Generic Process Model for Botnet Forensic Analysis,” *Conf. Adv. Commun. Control Syst. 2013 (CAC2S 2013)*, vol. 2013, no. Cac2s, pp. 98–102, 2013, [Online]. Available: http://www.atlantispress.com/php/download_paper.php?id=6285
- [138] M. M. Alani, “Chapter 2 OSI Model,” *Guid. to OSI TCP/IP Model.*, no. March 1978, pp. 5–17, 2014, doi: 10.1007/978-3-319-05152-9_2.
- [139] L. F. Sikos, “Packet analysis for network forensics: A comprehensive survey,” *Forensic Sci. Int. Digit. Investig.*, vol. 32, p. 200892, Mar. 2020, doi: 10.1016/j.fsidi.2019.200892.
- [140] M. Khanafseh, M. Qataweh, and W. Almobaideen, “A survey of various frameworks and solutions in all branches of digital forensics with a focus on cloud forensics,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, pp. 610–629, 2019, doi: 10.14569/ijacsa.2019.0100880.
- [141] B. Carrier and E. H. Spafford, “COMMONWEALTH OF AUSTRALIA Getting Physical with the Digital Investigation Process,” *Int. J. Digit. Evid. Fall*, vol. 2, no. 2, pp. 1–20, 2003, [Online]. Available: www.ijde.org
- [142] M. M. Houck, “Forensic Science: An Introduction to Scientific and Investigative Techniques,” *J. Forensic Sci.*, vol. 51, no. 1, pp. 205–205, 2006, doi: 10.1111/j.1556-4029.2005.00042.x.
- [143] R. Saferstein, *Criminalistics an Introduction to Forensic Science*, vol. 3. 2018.
- [144] M. M. Haque and S. A. Hossain, “National digital forensics framework for Bangladesh,”

- 3rd International Conference on Electrical Information and Communication Technology, EICT 2017*, vol. 2018-Janua. pp. 1–6, 2018. doi: 10.1109/EICT.2017.8275133.
- [145] J. Kizza and F. Migga Kizza, “Digital Evidence and Computer Crime,” *Securing the Information Infrastructure*. pp. 298–317, 2011. doi: 10.4018/978-1-59904-379-1.ch015.
- [146] X. Du, N. A. Le-Khac, and M. Scanlon, “Evaluation of digital forensic process models with respect to digital forensics as a service,” *European Conference on Information Warfare and Security, ECCWS*. pp. 573–581, 2017.
- [147] N. Kishore, S. Saxena, and P. Raina, “Big data as a challenge and opportunity in digital forensic investigation,” in *2nd International Conference on Telecommunication and Networks, TEL-NET 2017*, 2018, pp. 1–5. doi: 10.1109/TEL-NET.2017.8343573.
- [148] E. Bulgakova, V. Bulgakov, I. Trushchenkov, D. Vasilev, and E. Kravets, “Big data in investigating and preventing crimes,” in *Studies in Systems, Decision and Control*, 2019, pp. 61–69. doi: 10.1007/978-3-030-01358-5_6.
- [149] A. Shalaginov, J. W. Johnsen, and K. Franke, “Cyber crime investigations in the era of big data,” *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, vol. 2018-Janua. pp. 3672–3676, 2017. doi: 10.1109/BigData.2017.8258362.
- [150] S. Zawoad and R. Hasan, “Digital forensics in the age of big data: Challenges, approaches, and opportunities,” in *Proceedings - 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security and 2015 IEEE 12th International Conference on Embedded Software and Systems, H*, 2015, pp. 1320–1325.

doi: 10.1109/HPCC-CSS-ICESS.2015.305.

- [151] S. Satpathy, C. Mallick, and S. K. Pradhan, “Big Data Computing Application in Digital Forensics Investigation and Cyber Security,” *Int. J. Comput. Sci. Mob. Appl.*, vol. 5, no. 515, pp. 129–136, 2019, [Online]. Available:
<https://www.researchgate.net/publication/332572561>
- [152] M. S. Al-kahtani, “Big Data Networking : Requirements, Architecture and Issues,” *Int. J. Wirel. Mob. Networks*, vol. 8, no. 6, pp. 35–49, 2016, doi: 10.5121/ijwmn.2016.8604.
- [153] J. Song and J. Li, “A Framework for Digital Forensic Investigation of Big Data,” in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, May 2020, pp. 96–100. doi: 10.1109/ICAIBD49809.2020.9137498.
- [154] “REvil Ransomware Download | Tutorial Jinni.” <https://www.tutorialjinni.com/revil-ransomware-download.html> (accessed Nov. 25, 2022).
- [155] “Unix Time Stamp - Epoch Converter.” <https://www.unixtimestamp.com/> (accessed Nov. 25, 2022).

الملخص

تعتبر البيانات الضخمة أحد ركائز الثورة الصناعية الرابعة، وقد حظيت باهتمام كبير في السنوات الأخيرة نظرًا لقدرتها على التعامل مع كميات كبيرة من البيانات بسرعة وتنوع. ومع ذلك، يواجه المحققون الجنائيون تحديات عند إجراء الأدلة الجنائية الرقمية على نظام البيانات الضخمة بسبب الحجم الهائل من البيانات المخزنة، واحتمال حدوث سيناريوهات هجوم متنوعة، بالإضافة إلى نقص الأبحاث ذات الصلة التي تتحدث عنها. في هذه الأطروحة، نقدم إطار عمل خاص بالتحقيق الجنائي الرقمي للبيانات الضخمة مستجيب مع الحوادث والهجمات على اختلاف أنواعها، والتكيف مع خصائص الحوادث المتنوعة وطبيعتها، وجمع الأدلة التي يمكن استخدامها لإعادة بناء الأحداث الرقمية. يجمع هذا الإطار في التحقيق في مجال الكمبيوتر ومجال الشبكة ومجال البيانات الضخمة لضمان مراعاة شمولية التحقيق لجميع الجوانب ذات الصلة بالنظام. تتضمن عملية التحقيق الحفاظ على النظام والأدلة الرقمية، والبحث عنها، واستخراجها، وإعادة بناء تسلسل الأحداث. علاوة على ذلك، يقدم هذا البحث أداة تساعد المحققين الجنائيين الرقميين في عمليات اتخاذ القرارات والاستجابة للحوادث من خلال قاعدة بيانات مخزنة لمجموعة من السيناريوهات والأدوات المناسبة التي تساعد في تسريع عملية التحقيق، كذلك يقدم نموذجًا لتقرير الحادثة بشكل نهائي يتبع المعايير يمكن استخدامه لتقديمه للمحاكم بشكل قانوني.

الكلمات المفتاحية: البيانات الضخمة، التحقيق الجنائي الرقمي، التحقيق الجنائي الرقمي للبيانات الضخمة، الجرائم الإلكترونية، إطار عمل للتحقيق الجنائي الرقمي للبيانات الضخمة.