CRIB: Cyber Crime Investigation, Data Archival and Analysis using Big Data Tool

Priyanka Dhaka [#1]
[#] *University School Of Information And Communication Technology,*
G.G.S. Indraprastha University, New Delhi 110078, India
[1] *pridhaka92@gmail.com*

Rahul Johari [*2]
[*] *Computer Science & Engineering ,*
University School Of Information And Communication Technology,
G.G.S. Indraprastha University, New Delhi 110078, India
[2] rahul.johari.in@ieee.org

***Abstract*— In today's world Cyber Crimes are happening at a very rapid rate, the tools and techniques required to handle bulky and complex cyber crimes and attacks, are making organisations to remain coherent with evolving threats. Big data analytics provides a better and focussed way to overcome these threats and shorten their time to remediate with the help of tools like MongoDB. While the crime complete impairment are still unexplored there is need to control this damage by analysing the fast investigation using Big Data tools. This paper puts forth the ideas and strategies of using the cyber crime investigation tool to analyse the data and generate a report and deploying this data in MongoDB in order to manage large data and extracting structured data sets that it could help us to better understand and analyzed how to ensure threats from recurring.**

Keywords— Cyber crime, investigation tool, MongoDB, extracting;

INTRODUCTION

Cyber Crime is an illegal activity that involves committing a crime using computer as its basic level of communication. The U.S. Department of Justice elaborates the Cybercrime as any illegal activity that uses the computer for storage of evidence. The increasing list of cybercrime includes network intrusion and the dissemination of computer viruses as well as thefts like cyber stalking, phishing, blue jacking, key logger, DDoS and terrorism. The U.S. Digital Media Copyright Act (DMCA) of 1998 specifies that exchange of files of copyright material such as music and video is an illegal and punishable by law. As the cyber network impart communication and economy and supply government with different services it is also intruding threats to our privacy, economy and social life.

In order to detect these illegal behaviour we uses cyber security methodology which insures the protection of computer system or any information system from thefts and damage to software and hardware and information contained in them, It controls the hardware device from network access and data access and also from injecting codes containing harmful information, Cyber security involves the use of some cyber crime investigation tools to detect the major attacks and recovery from them, a key fact of using cyber forensics investigation is that, it is subjected to map the events of an incident from distinct source in order to obtain evidence of an incident to be used for other alternative investigations. Computer forensics gives the description of a process that runs on a particular technological device in order to monitor it so to determine whether the product is hacked previously or is being watched [1]. Several times the computer is considered as the way to identify the suspect and sometimes it contains the most inculpatory evidence, but as the crime and their investigation is increasing rapidly it is becoming a necessity to control this bulk amount of data in a structured format as to prevent data losses and to speed up the defect detection and prevention, this structured format also shortens the time of redress when crime takes place because the selective data extraction captures the optimum amount of data.

In this paper the basic approach was to use forensic tool to analyse the upcoming errors and their recurrence and prevent the data by removing such threats on the basis of this work the tool has generated a report that was further imposed to big data tool like MongoDB for more analysis and required extraction of optimum amount of data in order to minimise the threats and attacks after comparing their types from extracted data and also preventing them from occurring again.

This paper is organized as follows: Section 1 specifies about the digital forensic investigation process to be followed, Section 2 specifies about the forensic tools we can use to recover the data, Section 3 describes the motive of using big data in cyber crime, Section 4 describes the related work about big data, Section 5 specifies how the recovered information from forensic tools will be imposed to big data tool like MongoDB in order to extract valuable data and minimise thefts and Section 6 concludes the paper followed by references.

1. Digital forensic investigation process

Forensic investigation should be started by investigating whether or not any illegal process is running on a digital device or if any suspect imports illegal images on a computer. We followed certain steps while investigating illegal files or documents[1].

1.1. Initiating with the chain of custody and securing the digital evidence

Securing the data or evidence in early stage maximises the chances of successful investigation and litigation. The first custodian must be capable of describing the seen more accurately about the initial document.

1.2. Image storage devices

It particularly means collection of the information from a digital document and recording it in some information medium.

1.3. Examination of the information obtained from the digital device

We examined the information by making the copy of original evidence and importing it into a forensic tool for further investigation.

1.4. Data recovery

We recovered the data in order to identify the potential evidences using various methodologies.

1.5. Report generation

We used forensic tools that possessed built-in automatic report generation capability and generated the report that include audit reports and summarised data. All the steps discussed above are detailed as follows in figure 1.

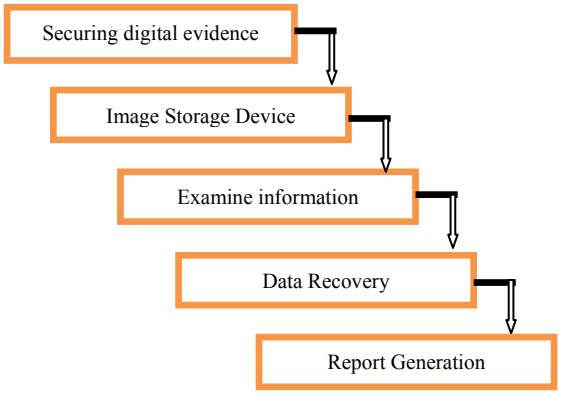

Figure 1: Investigation Process

2. Forensic tools to recover the data

This section discuss various forensic tools and the tool which were used for simulation and analysis:

2.1. FTK Imager [4]

FTK Imager is a disk imaging program. It saves an image of a hard disk in a file or segment which can be reconstructed when required later. It also calculates MD5 hash values and confirms the data integrity before choosing the document. The result is a image file that can be saved in different formats including DD raw.

2.2. Sleuth Kit [5]

The Sleuth kit is a depository and accumulation of command line tools that allows an investigator to integrate additive modules to examine file content and construct automated systems. The command line tools can be directly used to find evidence.

2.3. Award Key Logger [6]

Award Key Logger is a software that tracks key presses on a keyboard, it is capable of finding out what criminal can do when a device owner is away from a device, it records every keystroke to a log file, which will mediate everything that is typed and searched and takes a screen shot whenever a program detects particular keywords. The program can send the log files in an encrypted format by email or FTP to a desired receiver.

2.4. WinHex [7]

WinHex is a universal hexadecimal editor which review and edit all kind of files and recover deleted or lost files from hard drives. WinHex was made by X Ways Software Technology AG of Germany which offers following ability to :

- Read and directly edit RAM.
- Edit boot sectors
- Join and Split Files
- Analyse and compare files
- Search and replace
- Create hash values
- Recover and encrypt data

2.5. ProDiscover Basic [2]

ProDiscover Basic forensic tool unable to locate all the data on the computer disc and on the same time protect evidence and creates automatic reports with all the information needed to be presented as an evidence in legal proceeding, assembling data such as time zone information and driven information including things like volume, Serial number and hidden sectors, evidence of interest, internet activities etc[3]. ProDiscover Basic can recover deleted files, Examine slake space, access windows alternative Data Streams[2] and as it reads the disc at the sector level it is impossible to hide data from it. ProDicover Basic allows a search through entire disc of keywords, phrases with full Boolean search capabilities to find the necessary data. Hash comparison capability can be used to find known illegal files[2].

In this paper ProDiscover Basic was used to recover data from USB flash drive, the data is gathered from local computer and some through internet[9], this tool will also help in examining hard disk security and gives an exact picture of what is going on and what has happened with every cluster of the hard disk being analysed.

2.5.1. Viewing a DOCX File

The lower right pane displays the content in ASCII, as shown below. As it is a DOCX file the content are not suitable to read in this format.

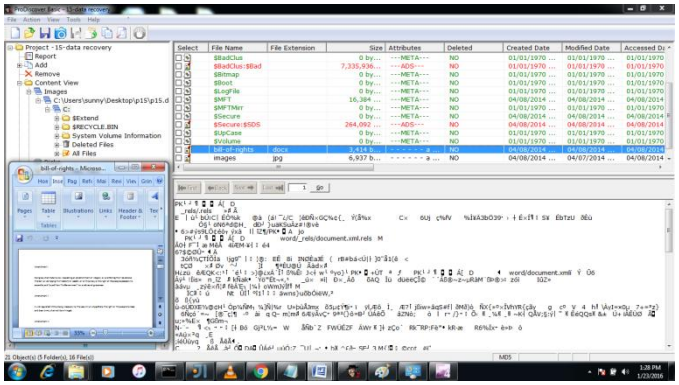

Figure 2 : Viewing a .docx file

2.5.2. Viewing the Physical Drive in Cluster View

In the top right pane the physical drive is shown in cluster view. In order to get down to raw bytes we can use cluster view.

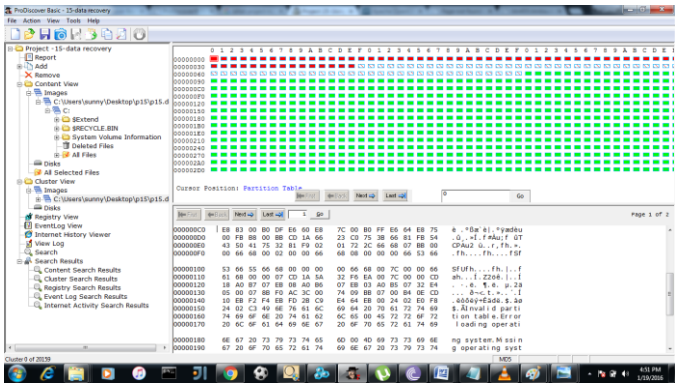

Figure 3 : Tool Snapshot

2.5.3. Viewing Report

Prodiscover Basic is capable of generating an automated reports that contains all the information needed to be presented as an evidence in legal proceeding. The generated report also contain information like serial number, file system used, bytes per sector, total clusters, total capacity and search results.

3. Scope of using Big Data in Cyber Crime

Fraud, Attacks, cyber terrorism and ethical hacking etc all these cyber crimes are happening at phenomenal rate. Big Data has become a key repository for fighting against these cyber crimes. Big Data analytics tools and techniques are helpful in managing the network security by predicting the criminal behaviour and type of attack. As huge amount of structured and unstructured data is present in the network, the security risks have increased so, the enterprises need to be alert due to vast increase in cyber attacks. Big Data along with cyber security is enabling business to fight against cyber attacks by analysing the tremendous data transferred during financial transactions and social communication and the data that is stored securely. Organisations prefer to avoid the complexities that emerges while understanding these cyber attacks, but they concentrate on the facts that how these attacks can break their defence system, in order to gain such information and to protect the defence system, the companies need to invest in Big Data tools and techniques.

Within the Big Data environment the usage of tool like MongoDB which manage a large amount of sensitive data like configuration files, error logs, system logs and more at any given time have gained importance. Big Data analytics provide a huge comparative exposure if the secured data falls into the wrong hands. Big Data analytics is capable of providing the enterprise a way to combine and correlate external and internal information to analyse a large picture of cyber crime and attacks against their enterprise.

Some of the major gains of using Big Data along with cyber forensic tools is[8] :

- Big Data fertilises the current monitoring system by applying logical analytics and cutting down the noise.
- Correlates the priority alerts throughout monitoring systems to identify the forms of thefts and attacks.
- Combining the internal and external data into single logical place and looking for known patterns of frauds.
- Making a record file of users account and internet activity information and image files to look for abnormal transactions against these files.

Big Data analytics is also adopted to prove various forensic tools effectiveness. By imposing the Big Data analytics on cyber investigation tools can investigates the incidents over multiple dimensions within a workspace , finds connections between apparently unrelated events and maps unfriendly events based on source. Big Data analytics tools can also tracks how cyber crimes changes over time and mitigating thefts that have seen before.

4. Related Work

In [10] author(s) discuss about how did the Big Data evolved gradually over period of time, how traditional DBMS couldn't

compete with large data sets and what are the issues and challenges that big data and the tools face today. In [11] author(s) gave the overview of Big data and discuss about it's association with other related technologies like cloud computing, distributed computing, Data Mining etc. In [12] author(s) have shown the understanding and design of large wireless network dataset which possessed actual movement of the nodes in the real time environment. The author(s) have showcased the successful execution of the relational algebra and SQL queries using the Hadoop Software. In [13] author(s) have weighed the merits and demerits of the the two well known RDBMS (Relational database management system): MySQL and MongoDB. Comparison was performed between two databases on basis of database operations such as insertion, deletion, selection, projection et al. In [14] author(s) strived to weigh the benefits and the banes of the Big Data along with the listing of the Prominent Big Data tools for data analytic's and discussion on the applications and challenges big data faces today. In [15] author(s) discuss about ideas and strategy of using data mining techniques such as genetic algorithm that can be applied on the data which was acquired from world statistics of mental health and deployed this data into a big data tool like MongoDB. In [16] author(s) discuss about the issues and problems attached with the Predictive Agriculture Analysis of Data Integration in India.

5. Imposing information obtained from Prodicover Basic to MongoDB in order to Analyse , Extract and Manage valuable information

MongoDB, a Big Data tool is designed to assure data security and also provides role based access control, encrypted communication and robust auditing. It provides easy to use integrated key management and protects the data within a given directory. In this paper the tool was used for importing the content of various reports of criminal investigation generated from Prodiscover Basic as it provided a clear structure of a single object and has no complex joins, it simply used keys and their values to store, run, and extract data in more effective manner whenever it was required. It has also used internal memory for storing the working sets and enabled faster access of data. MongoDB also enables to assign index on any attribute.

In this paper after having obtained the investigation reports from Prodicover Basic ,the key-value pairs of information obtained from each report were written and stored in MongoDB to extract the desired value of pairs whenever an investigation was needed, that can be helpful in comparing the threats and attacks that would be existed in any digital device. The use of MongoDB tool in storing the information from forensic tool is beneficial as it ensures data security and depicts the vulnerability of defects by correlating the priority alerts throughout the monitoring system. To our knowledge and wisdom our work is one of it's first kind approach in this direction. The desired information obtained from forensic reports was inserted into MongoDB as key-value pairs and after inserting the desired information one can find all the investigation reports as detailed in Report 1 and Report 2 and as described in Figure 4.

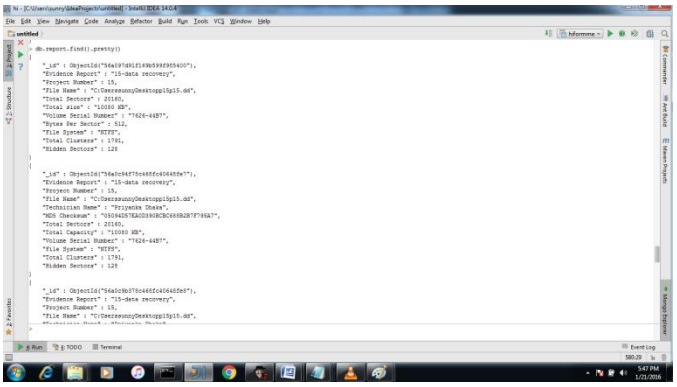

Figure 4 : Tool Snapshot with records

we can find One Investigation information as:

Report 1

```
db.report.findOne();
```

Result obtained is:

```
{
        "_id" : ObjectId("56a097d91f169b599f985400"),
        "Evidence Report" : "15-data recovery",
        "Project Number" : 15,
        "File Name" : "C:UserssunnyDesktopp15p15.dd",
        "Total Sectors" : 20160,
        "Total size" : "10080 KB",
        "Volume Serial Number" : "7626-44B7",
        "Bytes Per Sector" : 512,
        "File System" : "NTFS",
        "Total Clusters" : 1791,
        "Hidden Sectors" : 128
}
```

As depicted in the report above a unique id was generated for a particular report, that was helpful in rapid extraction of information and with the help of unique id information was also updated by using the following command:

Report 2

```
db.report.update({"_id":ObjectId("56a097d91f169b599f985400"){
        "Evidence Report":"18-recovery",
```

"Project Number":15,
"File Name":"C:\Users\sunny\Desktop\p15\p15.dd",
"Technician Name":"Priyanka Dhaka",
"MD5 Checksum":
"05094D57EA0D390BCBC688B2B7F795A7",
"Total Sectors":20160,
"Total Capacity":"10080 KB",
"Volume Serial Number":"7626-44B7",
"File System":"NTFS",
"Total Clusters":1791,
"Hidden Sectors": 128
}
)

Result obtained is:

WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })

Thus, the evidence report name is modified from "15-data recovery" to "18-recovery". Other operations in MongoDB can also be performed like remove, search, index, aggregate values and more. By using these operations the Computer forensics investigation was extracted, refined, updated and managed after collecting desired number of evidences.

6. Conclusion

This paper has discussed about Cyber crime, about forensic tools that can be used to investigate the type of crime and threats by collecting evidences, used Big Data tool such as MongoDB in which the investigation details and technical details were archived and analysed. The approach was simulated with the help of synthetic data set but it would be definitely performed on real time dataset once the funding will be obtained for this effort.

REFERENCES

[1] Ravneet Kaur, Amandeep Kaur, "Digital Forensics", International Journal of Computer Applications, 2012.
[2] http://prodiscover-basic.software.informer.com.
[3] http://searchsecurity.techtarget.com/definition/cybercrime
[4] Schneier, Bruce, "Secure Passwords Keep You Safer", January 2009.
[5] http://www.sleuthkit.org.
[6] Preeti Tuli, Priyanka Sahu, "System, Monitoring and Security Using Keylogger", International Journal of Computer Science and Mobile Computing, 2013.
[7] http://www.techrepublic.com/article/winhex-a-powerful-data-recovery-and-forensics-tool.
[8] https://www.promptcloud.com/blog/big-data-to-fight-cyber-crime.
[9] http://samsclass.info/121/proj/p15.
[10] Vibha Bhardwaj, Rahul Johari "Big Data Analysis: Issues and Challenges", IEEE International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), VIIT, Visakhapatnam, Andhra Pradesh, January 2015.
[11] Purva Grover, Rahul Johari "BCD : Big Data, Cloud Computing and Distributed Computing", IEEE Global Conference on Communication Technologies (GCCT -2015) Kanyakumari,TamilNadu,April2015.
[12] Vibha Bhardwaj, Rahul Johari, Priti Bhardwaj "Query Execution Evaluation in Wireless Network Using MyHadoop ",4th IEEE International Conference on Reliability, Infocom Technologies and Optimization (ICRITO 2015), AMITY University,September2015.
[13] Purva Grover, Rahul Johari "MVM : MySQL Vs MongoDB", Springer's 5th International Conference on Soft Computing for Problem Solving, IIT Roorkee, December 2015.
[14] Smita Bajaj, Rahul Johari "Big Data:A Boon or Bane–the Big Question", IEEE 2nd International Conference on Computational intelligence and communication technology (ICICT-2016),February2016.
[15] Priyanka Dhaka, Rahul Johari "Big Data Application: Study and Archival of Mental Health Data, Using MongoDB", IEEE International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT -2016), March 2016.
[16] Purva Grover, Rahul Johari "PAID: Predictive Agriculture Analysis of Data Integration in India", IEEE International Conference on Computing for Sustainable Global Development(INDIACom), March 2016.