A Review on Forensic Cloud Environment

Nilima Nikam[1]
[1.]Professor at YTIET,
Bhivpuri, Karjat, India.

Poorna R. Pimpale[2], Pranali Pawar[2], Anita Shirture[2]
[2.]PG Students,
YTIET, Bhivpuri, Karjat,India

***Abstract*- The concept of "BIG Data" that is emerging rapidly concerns the vast amount of data or information that is being processed, analyzed. The advancement of "BIG data" has also proven to be beneficial in the field of Digital Forensics. Also with the evolution of this technology crimes committed within the digital domain are increasing. The dimensions of digital evidence has grown to large extent which has subsequently affected the digital forensics. The analysis of such large volumes of digital data is a problem. Thus the current analysis tools cannot cope with the increasing demand of digital BIG data analysis. In this paper, a review of a generic Forensic Cloud Environment (FCE) which facilitates the digital analysis of BIG data in forensics is presented**

Keywords*— *BIG Data, Digital Forensics, Forensic Cloud Environment, Hadoop, Digital Forensic as a service.

I. INTRODUCTION

BIG DATA: The term "BIG Data" is relatively new and deals with the large amount of information that is analyzed to reveal patterns and trends relatively to human behavior and interactions.

DIGITAL FORENSICS: It is the process of interpreting electronic data. It deals with the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence which are derived from digital sources for the purpose of reconstruction of the events that are found to be criminal.

The concept of BIG Data deals with the three Vs Velocity, Volume, Variety. The data produced by most of the technologies fall in these three categories. With the explosion of Big Data, forensic investigators need to be prepared to analyze the peta bytes of data stored. BIG Data has affected the field of digital forensics in many ways. The digital forensic investigators have experienced drastic increase in size of forensic collection. In 2012, the Computer Analysis Response Team, (CART—a department that provides assistance to the FBI in the search and seizure of digital evidence) supported around 14,000 investigations, conducted more than 133,000 digital investigations, and analysed more than 10,500 Terabytes of data [23]. Whilst this data illustrates the size of data investigated collectively by an agency, the problem is further evidenced by a number of recent individual investigations which have involved the analysis of very large data sources for instance, in July 2012, the FBI was ordered to copy 150 terabytes of data held on the Mega Uploads server by Kim Dotcom [24]. The increase in forensic collection has increased the turnaround time for digital forensic investigations. This is because current digital forensic tools are struggling to analyze the flood of data presented to them [5]. One of the challenges that digital forensic examiners are facing is the inability to investigate multiple devices that belong to the same case together [6].

In 2010, IBM [8] mentioned that 80% of data generated today is unstructured. Data being unstructured is more complex and thus becomes difficult to analyze it. Current tools provide less automation to analyze such complex data. This paper reviews the design of a Forensic Cloud Environment that facilitates the analysis of a digital forensic case involving Big Data. Cloud provides vast storage capacity and processing power which is used by FCE to address Big Data.

II. BACKGROUND

Technology has changed the way people approach everything from online sales to writing. One such advancement is IOT. Internet Of Things can be applied in numeral ways. Benefits of these upcoming technologies are also taken by the criminals for their activities. National Crime Agency [10] has identified the growth in scale and speed of the Internet as one of the influences to serious and organised crimes. The Internet has made criminal activities easy as criminals can now commit crimes that are beyond their reach with minimal effort [11]. They can also collaborate virtually through these technologies to commit global crimes [11]. It is difficult to discover digital trails once the crime has been committed.

III. REQUIREMENT

The essential requirements to address the Big Data challenge in Digital Forensics are as follows

A. Big Data Technology

Traditional database techniques prove to be difficult to process large amount of data. Technique to be used as a Big Data solution should satisfy the specification like reliability, scalability, availability, fault tolerance.

Hadoop was considered as a suitable platform as it satisfies all the specifications. It is an open source software framework for distributed storage and processing of a very large data sets. Hadoop is comprised of features like: Mapreduce and Hadoop Distributed File System. MapReduce is a programming model for distributed processing of Big Data on large cluster of commodity hardware in a reliable, fault tolerant manner [18].

Working of Hadoop[25]:

Forensic Procedure of HDFS	
Preparation	
Identification	
First Step: Collection & Analysis	Collection
	Live Analysis
First Step: Collection & Analysis	Collection
	Transport
	Static Analysis
Reporting	

Fig 1. Working of Hadoop chart

B. Evidence Correlation

A survey conducted by Naked Security (2013) [7], indicates that the average number of devices per person is increasing. It also [7] shows that there were 3.1 devices per person in Germany, 3.0 in USA and 2.7 in the UK. This trend will continue to grow as the technological landscape progresses.

The increase in the number of devices per person is elevating the problems faced by investigators, as current tools are limited to one evidence source per investigation [6]. Evidence correlation is very useful in establishing a sequence of events in an investigation. Analysing a number of evidence sources together may provide a broader picture of the timing and sequence of events.

This platform will allow multiple evidence sources to be investigated, which will make it possible to extract the patterns and can correlate information which can be assisted in evidence discovery.

C. Collaboration

Usually investigators carryout individual investigation based on the evidence sources and then they try to put all evidences together. This process is very challenging when it deals with Big Data.

Collaboration proves to be beneficial here. Digital Forensic examiners can work independently and can collaborate their work with each other. With the support of collaboration investigation is geared up.

D. Intelligence Sharing

Intelligence comes from various sources that includes communication, browser history, user behaviour and textual analysis. Sharing intelligence is a fundamental feature.

It is reviewed that in the framework all objects are gathered and presented in XML format. The resulting XML will be stored in an object database that can be queried by other LEA's using FCE.

E. Knowledge Sharing

Knowledge sharing allows a group of individuals to share information about a common area of interest. Currently knowl-edge attained from investigating a case is not shared with other investigators [14]. This means that when a similar case is being examined new strategies need to be researched [14].

F. Security

Security is one of the key challenges in digital forensics as it deals with to maintain the atomicity and integrity of the evidences. Therefore it is essential that the FCE avoids any form of evidence tampering. Pringles and Burgess [20].

IV. SYSTEM ARCHITECTURE

The main components of the system include; the ingester, Information Interchange Framework (IIF), Hadoop Cloud, workers and the Intelligence Sharing framework.

A. Ingester

With the help of ingester images are inserted in HDFS and case meta data is created. It is also responsible for inserting data from image in to HBase.

B. Information Interchanging Framework

The IIF is an API that controls and facilitates the data exchange during isertion between ingester and Hadoop Cloud.

C. Hadoop Cloud

Hadoop framework runs on cloud environment therefore the system is implemented. HDFS is used to store forensic collection while MapReduce is used for processing the data.

D. Worker

Examiners or workers are responsible for quering, processing data from Hadoop. Many investigators can be assigned to investigate a case together. Each worker is assigned a specific role which will enable them to access certain areas of evidence that are necessary.

E. Intelligence Sharing Framework

The analy-sis stage identifies objects (such as phone numbers, credit card numbers, email address, people and so on) from the evidence. Identified objects are then compared with other objects from the object database.

The object database contains a combination of watch list and other objects accumulated during previous investigations. The results of co-relational analysis will show many important leads, for example, how many times an object appeared in different cases, which cases did the object appear in. In short, this feature gives investigators the opportunity to detect relationships between objects in different cases.

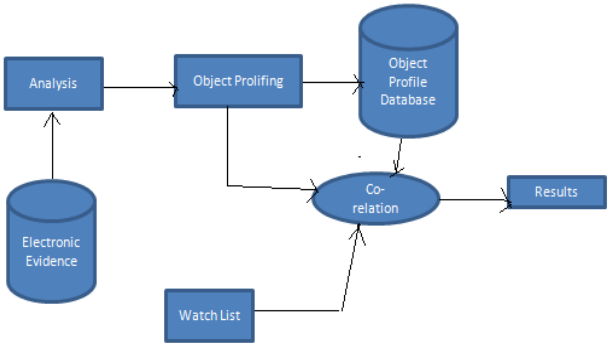

Fig 2.Intelligence sharing framework

V.EXPERIMENT

The purpose of this experiment was to demonstrate that Hadoop can be used for digital forensic analysis. Gather a Windows PC which contained only the OS files, then some enron email datasets are inserted [21]. The evidence machine was then imaged using FTK imager. The md5sum of the evidence was noted before ingesting the evidence into HDFS. The evidence on HDFS was then inserted in HBase using MapReduce program. The md5sum of each file was also calculated for verification after the analysis. We carried out timeline and network analysis on the evidence. These analyses were suitable for the data in the evidence source. Further, analysis such as image analysis (image segmentation, classification), video content analysis and or audio semantic can also be applied.

1) Timeline analysis: Timeline analysis provides a broader picture of the sequence of events and timing of these events. Using the timestamps we were able to organise evidence in a timely manner. With this arrangement we were able to view all events that happened at a specific time span.

In the experiment it is seen that a number of email exchanges has been occurred in January 2001. To achieve this a MapReduce program is written to create a timeline of all files from the evidence source. The output from the MapReduce program was then further processed to produce an html format of the timeline. The timeline output was then visualised using vis.js library [22].

In the experiment it is seen that, the content of any message id.Java-Mail.evans@thyme is revealed. The visualisation library also allow plotted objects to be temporary deleted. This can be a very useful feature as objects that are deemed good can be removed from the plot. This will leave a few objects on display, thereby simplifying the analysis process.

2) Network analysis: We used network analysis technique to identify communication links between individuals. A MapReduce program was written to map communication data from HBase. Some of the code from Metamail program [23] is adopted. The Map program output was a key pair value of the sender's email address and recipient email addresses see, Listing 1. The Reducer program mapped the senders email to all recipients. The output from the MapReduce program was then processed to an html version to visualise it using vis.js library [22].

Arrows in the network graph that can be studied indicates that an email was sent towards an individual. From the network graph that we can obtain from the experiment can identify a number of details such as a group of individuals who exchanged emails amongst each other. We can also see individuals who were responsible for sending many emails. In a real investigation, it will be interesting to carry out further investigation on any of these individuals. Some individual can also be at the centre of investigations to get more information. Overall, network analysis helps in detecting groups that are close to each and also the direction of communication.

MapReduce Algorithm to create communication network
1. Set credit to Metamail
2. Initialize MailRecord with a Session Object and InputStream Object
3. If Mail Record. mail is Mail Record. create (Session,in Stream) then Get mail from the sender Store the information in Array List
4.for (String email to List Array) Concatenate recipients separated with a comma
5.Get output from all the receivers

3)Verification: After completing all the analysis we compared the original md5sum to the final md5sum of all files. The comparison indicated that the evidence was not corrupted. This result proves that Hadoop did not modify the evidence. The ability to maintain the integrity of evidence is a major requirement of all forensic tools. With this outcome we can conclude that Hadoop can be used for forensic analysis.

V. CONCLUSION

This paper briefly reviews the six requirements that can facilitate the analysis of a BIG Data case. It also summarizes the prototype and experiments that are carried to demonstrate that Hadoop can be used for digital forensics.

It is reviewed that the framework is feasible and can be used to improve turnaround time for investigation. Initial stage of implementation was to demonstrate that Hadoop does not corrupt the evidence. The next stage which has not been reported in the papers reviewed which is to design a Big Data case and carry out some experiments will be explored in future research work.

VI. REFERENCES

[1] A. Dragland. (2013) Big data - for better or worse. [Online]. Available: http://www.sintef.no/home/corporate-news/Big-Data–for-better-or-worse/.

[2] P. Zikopoulos, C. Eaton, D. Deroos, T. Deutsch, and G. Lapis, *Under-standing Big Data*. McGraw-Hill, 2012, p. 4.

[3] D. Laney, "The importance of 'big data': A definition," 2012.

[4] RCFL. Regional computer forensic laboratory 2012 annual report. [Online]. Available: http://www.rcfl.gov/downloads/documents/2012-rcfl-national-report

[5] D.Quick and K-K R.Choo,"Impacts of increasing volume of digital forensic data: A survey and future research challenges" Digital Investigation, vol.11, no.4,pp.273-294, 2014. Available: http://www.sciencedirect.com/science/article/pii/S1742287614001066

[6] S.L.L Garfinke"Digital Forensic Research: The next Digital Investigation"vol. 7, the proceedings of Conference.

[7] K.Truong(2013) Infographic: Users weighted down by multiple gadgets- survey reveals the most carried devices.

[8] IBM The enterprise answer for unstructured data

[9] V.Roussev and G.G Richard III, "Breaking The Performance Wall" in proceedings of the 2004 Digital Forensic Research Workshop.

[10] NCA, "National Strategic Assessment of serious and organised crime 2014" National Crime Agency, Tech.Rep.2014.

[11] D.S Wall. Cybercrime Polity Press,2007.

[12] V.Roussev, L.Wang, G.Richard, and L.Marziale, "A cloud computing platform for largescale forensic computing" in Advances in Digital Forensics V, ser. IFTP Advances in Information and Communication Technology, G.Peterson and S. Shenoi, EDS. Springer Berlin Heidel berg,2009.

[13] C. Miller, D.Glendowne, D.Dampier, and K.Blaylock, "Forensic cloud: An Architecture for digital forensic analysis in the cloud" Journal of Cyber Security, vol 3
[14] Baar, H.Beek, E van Eijk, Digital forensics as a service: A game Digit Investigation.
[15] C. Federici,"Almanebula: A computer fornsic framework for cloud," Procedia computer Science vol.19, no.0, 2013, the 4th International Conference on Ambient Systems, Networks and Technologies(ANT 2013), 3rd International Conference on Sustainable energy Information Technology
[16] B. Carrier.(2012) Sleuth kit hadoop framework.
[17] T.White, Hadoop: The Definitive Guide. O"Reilly Media, Inc, 2011.
[18] J.Dean and S.Ghemavat,"MapReduce: Simplified data processing on large clusters" Commun,ACM, vol.51, no.1.
[19] R.A.Best,"" Sharing law enforcement and intelligence information: The congressional role, Foreign Affairs, Defense, and Trade division Tech Rep 2007.
[20] N.Pringle and M.Burgess, "Information assurance in a distributed forensic cluster" Digital Investigation vol.11, Supplement 1, no.0, 2014, in proceedings of DFRWS.
[21] W.W. Cohen. Enron email dataset.
[22] D. Pino. Email analysis tool based on hadoop.
[23] EURIM-ippr(2004). EURIM-IPPRE Crime Study: Partnership Policing for the information Society. Third Discussion paper. Available online: http://www.eurim.org/consult/ecrime/may_04/ECS_DP3_Skills_040505_web.htm (accessed on 31 December 2013.
[24] European Information Society Group. Separating Myth from Reality and Snake Oil from practicality. Partnership Policing for Information Society, London UK, 2010.